

Judging a Book By Its Cover: Bringing the Digital Humanities into Reader's Advisory

by

Laura Michelle Gerlitz

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Arts and Master of Library and Information Studies

Digital Humanities and Library and Information Studies
University of Alberta

© Laura Michelle Gerlitz, 2019

Abstract

This study sets out to examine recurring themes found on book wrappers published by Harlequin in their first seventeen years as a form of marketing strategy. Through the use of specific image and text patterns that correspond to common themes found in paperback genres, Harlequin was able to appeal to targeted audiences, competing with other early reprint companies and eventually become the colossal modern publishing company of today. A unique approach using several digital humanities methods, namely text and image analyses, and data visualizations, will be employed to examine a special collection of several hundred wrappers. An interdisciplinary approach to research in book history will be taken, utilizing a blend of methods from digital humanities, and theories from library and information studies, humanities, and communication studies. The resulting patterns will be connected to reader's advisory as appeal factors in successful book selection by readers.

Preface

This thesis is an original work by Laura Gerlitz. Portions of this thesis will be published in a future volume of the Bibliographical Society of Canada's journal, *Papers of the Bibliographical Society of Canada*, as part of the Society's Emerging Scholar Award for 2018.

Acknowledgements

I would like to thank a number of librarians, professors and other staff and individuals who supported me throughout these three years of research, without whom this thesis would not have been possible. Robert Desmerais, for suggesting the Peel Harlequin collection in the first place, as well as Linda Quirk for allowing me the opportunity to work with the books and working with me on the online exhibit. Thank you to Jeff Papineau for taking the time to deliver the collection to the digitization room in Cameron library when I needed them. My two thesis supervisors, Tami Oliphant and Harvey Quamen, deserve immense credit for assisting me with my research and providing extremely valuable guidance. Cecily Devereaux and Danielle Allard, my first and second readers, for taking the time to be a part of my supervisory committee and in Cecily's case for allowing me to audit her course on Harlequin and the romance genre. I would also like to thank Peggy Sue Ewanyshyn, as well as the University of Alberta Library's Digital Initiatives, for providing me with the resources and space to digitize the collection. Last but certainly not least is my family, for encouraging me to apply for graduate school, and for supporting me throughout my entire program.

Table of Contents

| | |
|---|-----|
| Abstract | ii |
| Preface..... | iii |
| Acknowledgements | iv |
| Chapter 1: Introduction | 1 |
| 1.1: My research journey..... | 1 |
| 1.2: Research Goal and Questions..... | 2 |
| 1.3: Research Outline | 4 |
| Chapter 2: Literature Review | 5 |
| 2.1: Histories | 5 |
| 2.1.1: History of Harlequin | 5 |
| 2.1.2: Reader's advisory history..... | 7 |
| 2.2: Theories..... | 10 |
| 2.2.1: Reader's advisory | 10 |
| 2.2.2: Semiotics | 14 |
| 2.2.3: Colour theory | 18 |
| 2.2.4: Paratext and peritext | 23 |
| 2.2.5: Distant reading..... | 26 |
| 2.3: Literature reviews..... | 28 |
| 2.3.1: Image analysis studies | 28 |
| 2.4: Conclusion..... | 31 |
| Chapter 3: Research Design..... | 33 |
| 3.1: Data collection | 33 |
| 3.1.1: Collection description..... | 34 |
| 3.1.2: Digitization of collection..... | 38 |
| 3.2: Analysis methods | 43 |
| 3.2.1: Text analysis | 43 |
| 3.2.2: Image analysis | 46 |
| 3.3: Conclusion..... | 54 |
| Chapter 4: Analysis Results and Discussion..... | 56 |
| 4.1: Results | 56 |
| 4.1.1: Text Analysis Results and Discussion..... | 57 |
| 4.1.2: Image Analysis Results and Discussion..... | 64 |
| 4.2: Conclusion..... | 93 |
| Bibliography | 97 |

| | |
|---|-----|
| Appendices..... | 107 |
| Appendix A: PHP Scripts to Run Color Summarizer | 107 |
| Autoload | 107 |
| Running different image measurement options in Color Summarizer..... | 107 |
| Color Summarizer Class..... | 107 |
| Appendix B: K-means Clustering Scripts | 110 |
| Hex code extractor | 110 |
| Silhouette Method for Determining number of K-Clusters in R..... | 110 |
| Appendix C: Visualization Scripts..... | 112 |
| Scatterplot of brightness median over time in R..... | 112 |
| Scatterplot of brightness median across genres in R | 112 |
| Comparison of hue, saturation and value within genre in R..... | 112 |
| Palette creator in R..... | 112 |
| Appendix D: Color Summarizer Statistics | 114 |
| “The Manatee” original image vs. resized image aggregate statistics from Color Summarizer | 114 |
| Appendix E: Full size median HSV scatterplots | 116 |
| Canadian fiction..... | 116 |
| Detective/mystery fiction | 117 |
| Doctor/nurse romance..... | 118 |
| General fiction | 119 |
| General romance..... | 120 |
| Historical fiction..... | 121 |
| International location | 122 |
| Nonfiction..... | 123 |
| Western/ranch fiction..... | 124 |
| Appendix F: Colour Categorizations..... | 125 |
| Steinvall’s (2007) colour categories (p. 357) | 125 |
| Krygier and Wood’s (2016) colour categories (p. 266) | 125 |
| Aslam’s (2006) colour categories (p. 19)..... | 126 |
| Allan’s (2008) colour categories (p. 636-637) | 126 |

List of Tables

| | |
|---|-----|
| Table 1 Colours as symbolism for emotions and concepts..... | 23 |
| Table 2 Genre categories for Harlequin collection | 37 |
| Table 3 Female, male and unknown gender author count by genre. | 37 |
| Table 4 Summary of methods and tools used in this thesis. | 38 |
| Table 5 Most frequent words according to genre. | 59 |
| Table 6 Author gender and term frequency for Figure 10 | 61 |
| Table 7 Legend for Figure 10 | 61 |
| Table 8 Author gender and term frequency for Figure 11 | 62 |
| Table 9 Legend for Figure 11 | 62 |
| Table 10 Comparison of genres across median HSV scatter plots | 78 |
| Table 11 The Manatee statistics..... | 115 |
| Table 12 Steinvall's colour categories | 125 |
| Table 13 Krygier and Wood's colour categories..... | 126 |
| Table 14 Aslam's colour categories | 126 |
| Table 15 Allan's colour categories..... | 127 |

List of Figures

| | |
|--|-----|
| Figure 1 Unedited scan of "Gambling on Love" by Gail Jordon, 1949..... | 40 |
| Figure 2 Edited scan of "Gambling on Love" by Gail Jordon, 1949..... | 41 |
| Figure 3 Cirrus with the Works of Jane Austen. From "Cirrus" by G. Rockwell, & S. Sinclair, 2016, https://voyant-tools.org/docs/#!/guide/cirrus | 44 |
| Figure 4 Bubblelines with the Works of Jane Austen. From "Bubblelines" by G. Rockwell, & S. Sinclair, 2016, https://voyant-tools.org/docs/#!/guide/bubblelines | 45 |
| Figure 5 Example of ImageMeasure data used by ImagePlot for visualization of an image collection by Software Studies Initiative, 2011, https://www.flickr.com/photos/culturevis/6026021275/in/photostream | 47 |
| Figure 6 Step 1 example of Auto-blend by G. Ornbo, 2007, https://shapedshed.com/auto-blend-photographs-in-photoshop/ | 51 |
| Figure 7 Step 2 example of Auto-blend by G. Ornbo, 2007, https://shapedshed.com/auto-blend-photographs-in-photoshop/ | 52 |
| Figure 8 Step 3 example of Auto-blend by G. Ornbo, 2007, https://shapedshed.com/auto-blend-photographs-in-photoshop/ | 53 |
| Figure 9 Word cloud of corpus created using Voyant..... | 58 |
| Figure 10 Bubble line chart of most frequent doctor/nurse terms created in Voyant..... | 61 |
| Figure 11 Bubble line chart of most frequent detective and mystery terms created in Voyant.... | 62 |
| Figure 12 Auto-blend of Harlequin covers not using seamless tones and colours option..... | 65 |
| Figure 13 Auto-blend of Harlequin covers using seamless tones and colours option..... | 66 |
| Figure 14 Brightness vs. saturation scatter plot diagram created using ImageMeasure and ImagePlot..... | 69 |
| Figure 15 Brightness median over time created using ImageMeasure and R..... | 71 |
| Figure 16 Value median according to genre, created using R..... | 72 |
| Figure 17 Biographical colour palette..... | 80 |
| Figure 18 Canadian fiction colour palette..... | 81 |
| Figure 19 "Royce of the Royal Mounted" by Amos Moore, 1950 and 1957..... | 82 |
| Figure 20 Detective/mystery fiction colour palette..... | 83 |
| Figure 21 Doctor/nurse romance colour palette..... | 85 |
| Figure 22 General fiction colour palette..... | 86 |
| Figure 23 General romance colour palette..... | 87 |
| Figure 24 Historical fiction colour palette..... | 88 |
| Figure 25 International location colour palette..... | 89 |
| Figure 26 Nonfiction colour palette..... | 90 |
| Figure 27 Science fiction colour palette..... | 91 |
| Figure 28 Sports fiction colour palette..... | 91 |
| Figure 29 Western/ranch fiction colour palette..... | 92 |
| Figure 30 Canadian fiction HSV scatterplot..... | 116 |
| Figure 31 Detective/mystery fiction HSV scatterplot..... | 117 |
| Figure 32 Doctor/nurse romance HSV scatterplot..... | 118 |
| Figure 33 General fiction HSV scatterplot..... | 119 |
| Figure 34 General romance HSV scatterplot..... | 120 |

| | |
|--|-----|
| Figure 35 Historical fiction HSV scatterplot | 121 |
| Figure 36 International location HSV scatterplot | 122 |
| Figure 37 Nonfiction HSV scatterplot | 123 |
| Figure 38 Western/ranch fiction HSV scatterplot..... | 124 |

Chapter 1: Introduction

1.1: My research journey

This thesis took root in an antiquarian bookstore located in Calgary, through the exceedingly lucky connection of disparate threads built from the assistance of university librarians, the scope of the store's inventory, and my own experience. It began with my exposure to working with special collections for retail purposes, preparing descriptions and images and handling transactions for a number of buyers, including the University of Alberta. The more I worked with antiquarian books and ephemera the more my interest in them grew, and the stronger my connections became with not only the library world, but specifically that of special collections. I was able to network with the librarians at the University of Alberta, something that had an unexpectedly profound impact on the creation of my thesis topic. Taking that fortuitous start, my thesis topic was developed and further refined during my time in the Digital Humanities/Library and Information Studies (DH/LIS) program, eventually taking the form of this final study which I hope can support both librarians and non-librarian scholars in furthering their own research.

My first real breakthrough in my thesis work began with my application to the DH/LIS program. Requiring a research proposal as part of the application, I immediately knew I wanted to combine rare books with technology, with the clear idea of creating an online exhibit for one of the collections found at the University of Alberta's Bruce Peel Special Collections and Archives. Upon contacting the head of the library, he suggested a collection that had both been purchased from my book store and fit within my research interests (Canadian history, women's history, and a collection that hadn't been given much attention by the university): a number of early paperbacks published by Harlequin Enterprises. Intrigued by his suggestion I dove into the

history of the company, and thus began my first foray into combining a physical collection with digital research methods. I completed my research proposal, and gradually it was expanded upon as I learned more about the digital humanities. I wanted to keep working with the collection and was given the opportunity to scan the exteriors of the collection as digitizing the interiors was not possible due to their fragility. As I continued on in the program the question of interest became *How can I combine library studies and digital humanities into original research using a limited scanned collection?*

Using text and image analysis were two clear options. Despite not having the interior contents, text from the front cover quotations and the back cover summaries were still available to me via optical character recognition (OCR), and images from the front covers were colourful, eye-catching and, frankly, interesting. Through further research I discovered theories such as paratext, and after discussion with my thesis committee I settled upon my topic. I took the idea of applying the text and image analysis to reader's advisory, knowing that the scans of the book wrappers (that is, the front cover, back cover, and spine) contained the first characteristics examined by a potential reader in deciding their interest in the book. Through my knowledge of the history of the company, I was also aware that the collection represented the beginning of Harlequin's transition from general reprints to focusing on romances, which further supported this angle of research. I settled on my topic, and pursued it enthusiastically.

1.2: Research Goal and Questions

The goal of this research, using digital humanities methods and tools for an interdisciplinary approach, is to analyse elements found in reader's advisory to determine consistent patterns and relationships across the genres of a small paperback collection. These

elements are used by readers during the process of choosing new books for pleasure-reading, and it will be argued that consistency across these elements allows the publisher to target readers with the message that that consistency also equates to reader experience. In other words, *this book* was an enjoyable read, and because *that book* appears similarly to *this book*, *that book* will also be enjoyable. Focusing on this Harlequin collection allows the advantage of historical context as the rise of the company is part of a larger shift in mindset towards books and bookselling. This collection is particularly useful in that it shows a shift in publishing trends, as the company became aware of a change in the genre preferences of their readers, and thus changed their publishing tactics accordingly over time. Examining this time period would not be possible without the use of Digital Humanities (DH) methods and tools, as this computational assistance will allow me to analyse the collection, consisting of several hundred books, rather than a few individual titles. This will allow the collection to be treated in a manner similar to big data; while the collection itself is not large enough to actually *be* big data, methods and techniques will be borrowed for processing and analysis. This thesis will also incorporate theories such as distant reading with the more traditional reader's advisory to give new insights that would otherwise not be possible. Aspects of the wrappers will be measured and analyzed on large scale through methods such as k-clustering of colours and examining themes through word frequency, which would not be possible by a single researcher without computer assistance. This big data approach allows me to track Harlequin's genre shift over the course of nearly two decades.

This study may assist librarians and booksellers in improving reader's advisory services, particularly within the realm of recommended reading lists in online library catalogues. Using these DH methods and tools, librarians will be better equipped to identify common elements

across genres (particularly formula fiction), leading to better discovery systems and ultimately, through both direct and indirect methods, to better assist patrons in choosing books. This thesis intends to answer the following questions:

- What patterns or relationships can be found between the art in the front covers and the genres of the books?
- How is the language in the front cover quotations and back cover summaries used to appeal to potential readers?
- How can taking a DH approach assist in reader's advisory?

1.3: Research Outline

This thesis is structured as follows. Chapter one will explain how my topic developed over time, as well as a summary of my research questions and outline. Chapter two consists of historical background of paperback novels and Harlequin Enterprises, theories that will be used in this thesis, and literature reviews of other academic research involving image analysis. The background covers paperback novels in the early-to-mid-twentieth century and the growth of Harlequin Enterprises until the 1960s. The literature reviews will cover reader's advisory, text and image analysis, and similar published case studies that involve book cover and image analysis, and paratext. These reviews will provide definitions and context for my study. Chapter three will cover the research design of this thesis, specifically the methods and theories used in the text and image analysis. Finally, chapter four will discuss the findings of my analysis, connecting it to reader's advisory, as well as my final thoughts and conclusion.

Chapter 2: Literature Review

This chapter provides an overview of the theories that will be used in this thesis in the form of a literature review. Reader's advisory, semiotics, colour theory, paratext and peritext, and distant reading will all be defined and their relevance to the research topics will be discussed. Further background through a historical examination will also be provided for Harlequin Enterprises, and reader's advisory. Finally, this chapter will wrap up with a discussion on current image analysis research and projects that have been completed in academia.

2.1: Histories

This section explores the history and development of Harlequin Enterprises, as well as reader's advisory. Beginning with the history of Harlequin Enterprises, this chapter will explain how the change in marketing and selling of books first led to the creation and development of the company. This, in turn, directly influenced the genres of books sold within the time period of the collection (1949-1965). The creation and growth of reader's advisory within libraries will follow, examining its various evolutions from the late 1800s to the present day, to supply a thorough understanding of how its resulting modern definition, which will be used in this thesis.

2.1.1: History of Harlequin

The rise of the paperback novel was an indicator of a change in the publication, marketing and selling of books. A widespread growth in literacy through the increased prevalence of public high schools and improvements in printing technology from the Industrial Revolution onwards, such as improvements in wood-pulp paper production and the invention of typesetting machinery, helped push North American booksellers towards advertisement and

other methods of promoting their wares (West, 2011, p. 782; Kaestle, 2007, p. 11 & 30).

Newspapers and other kinds of periodicals began to be advertised in news ways, as media that could be consumed quickly and consistently, and in the 1920s and 1930s books followed suit (Kaestle, p. 11). Beginning with book-of-the-month clubs books became widely available through new retail avenues-mail-order catalogues, department stores, and grocery stores-to a new, diverse customer base (West, 1990, p. 124-125; Kaestle, 2007, p. 24-26). This reach to a wider audience, along with the use of modern printing technology and cheap materials becoming more common allowed for the growth of new reprint and paperback companies such as Harlequin Enterprises (West, 1990, p. 127-128; Radway, 2009, p. 16-17).

The seeds of Harlequin Enterprises were sown well before 1949, and involved three major players: Doug Weld, Jack Palmer, and Richard Bonnycastle. Bonnycastle, a Canadian socialite and previously an employee of the Hudson's Bay Company joined, up with Weld to assist in re-energizing a branch of Weld's family printing firm. This branch, Advocate, specialized in paperback novels of varying genres of fiction, and as Bonnycastle worked with Weld he met the sales manager of one of New York publisher Grosset & Dunlap's distributors, Jack Palmer. The three founded Harlequin Enterprises with the intention of having a small publishing company that specialized in mass-market paperbacks and catered to Canadians (Grescoe, 1997, p. 26-29; Jensen, 1984, p. 32; Radway, 2009, p. 48). *The Manatee* was their first publication in 1949, with approximately two dozen more novels following suit that first year. Despite this large number of publications returns were common, and the company struggled for several years to keep afloat (Grescoe, p. 32). After the death of Jack Palmer, Richard Bonnycastle's wife Mary and his personal secretary Ruth Palmour began to take on major roles in the company. Palmour took on a managerial role, "[corresponding] with publishers across the

continent [to] ask them to submit books for possible reprinting” while Mary held final approval and editing responsibilities (Grescoe, p. 39). The two women were shrewd enough to recognize Harlequin’s early successes in the 1950s took the form of romance novels: from *Beyond the Blue Mountains*, which had only 48 copies returned out of 30,000 sales, to the first medical romances written by Lucy Agnes Hancock (Grescoe, p. 36).

These women were the two key players in creating the partnership that undoubtedly led Harlequin to the booming business that it is today, the partnership with Mills & Boon, an English publishing company that sold pocket-books and the romance genre that Harlequin would quickly become known for. Anne Vinton’s *The Hospital in Buwambo* became the first Mills & Boon novel published by Harlequin in 1957, with Mary Burchell’s *Hospital Corridors* and several other books following in 1958. In that same year Harlequin began to regularly import romance genre novels from Mills & Boon (Jensen, 1984, p. 33). Per Grescoe (1997), “from 1955 on, Harlequin published one romance title almost every month, nearly all of them with nurse-and-doctor themes” (p. 53). This shift from a mixture of genres towards medical romances is plainly seen, both in the contents of the books and their paratext.

Shortly before or around the same time as the introduction of the paperback novel and the formation of Harlequin Enterprises, libraries were also undergoing changes to their patron services. The following section will detail this history of the introduction and development of reader’s advisory.

2.1.2: Reader’s advisory history

The date of creation of formal reader’s advisory services is contested, ranging from the late 1800s to the 1920s. Crowley asserts the exact time period in which it was established is

difficult to pinpoint, as early reader's advisory offered varying services that changed according to local regions. However, he argues that it has existed in various forms for as long as library staff have discussed books and reading with their patrons (Crowley, 2005, p. 37-38). At any rate, there is agreement among scholars that structured services began in American public libraries during the 1920s, with varying opinions as to whether nonfiction was privileged, or if fiction and nonfiction were equally advised upon (Crowley, 2005, p. 38; Saaricks, 2005a, p. 4). As this structure developed through the 1920s and 1930s, reader's advisory appeared within more academic circles. This refers to societal contexts, such as the Adult Education Roundtable at the year American Library Association (ALA) conference, and within professional Library and Information Studies research, in which studies were carried out on the topic of "adult reading and readability" to enhance the effectiveness of public library service (Saaricks, 2005a, p. 5). This breach into the pursuit of research and scholarship quickly led to the publication of monographs and journal articles, though Saaricks is careful to note that, despite this increase in importance, historical reader's advisory remained of a "moralistic, didactic tone," and that the aim was to "move readers toward classic works, to outline a plan of reading that would be educational, not recreational" (Saaricks, p. 6). The encouragement of reading for leisure was not seen as a priority for public libraries until more modern eras.

Flexner and Hopkins' analysis of the New York Public Library's reader's advisory program in the 1920s and 1930s is of particular note in better understanding factors to reader's advisory that are applicable even today. Using over 1200 survey responses from patrons, the two came to several major conclusions in the development of a strong program. One of the main findings was that reader interests were strong in both fiction and nonfiction. A second finding was based on the success of the program, and the importance of connecting the service to the

larger organization as well as promoting its advocacy through training future administrators (Crowley, 2005, p. 40). Finally, they found it is important to ensure an adequate amount of human resources are devoted to meeting patrons' expectations or exceeding them, and that in order for a reader's advisory program to be successful discussions must take place and ideas exchanged (Crowley, p. 41).

The events of World War II led to an overall reduction in leisure time, and a subsequent decrease in requests for reader's advisory services (Crowley, p. 38; Saaricks, 2005a, p. 6). Additionally, librarians began to see the work as too difficult to maintain, and recommended reading lists originally created in mind with patron personality characteristic became more standardized (Saaricks, p. 6). Reader's advisory was "lost" in adult services until the 1980s and 1990s, when there was a revival with a greater focus on leisure over education as well as interest in regards to library staff training, and research towards improving the library experience of the patron (Crowley, 2005, p. 38; Saaricks, 2005a, p. 7-8). Modern reader's advisory can be defined as a "patron-centered library service for adult leisure readers", in which one central tenant is that any reading has an attached intrinsic value (Saaricks, p. 1). The goal is not only to satisfy patron needs but to "advance a culture's goal of a literate population" (Crowley, 2005, p. 37), so while there is a renewed focus on leisure and pleasure, education is still an important aspect of the service. Readers advisors and forms of advisory are a "vital link" between the library's collections and its patrons, particularly as the organization of said collections can be intimidating and a barrier to readers (Saaricks, 2005a, p. 4). Reader's advisory as a modern theory and its relation to the research topic will be further elaborated on in the following section.

2.2: Theories

This section will explore the theories of reader's advisory, semiotics, colour theory, paratext and peritext, and distant reading. The theory of distant reading provides a foundational argument that analyzing the collection metadata rather than the content of the books is a legitimate form of research. Semiotics, paratext and peritext all work in conjunction with one another: paratext and peritext posit that there is meaning found in the structures surrounding the contents of the books, which is supported by semiotics and the deconstruction and understanding of underlying messages represented by different aspects of the book wrappers. One of these aspects, the use of colour, uses colour theory first to explain colour models that allow for the measuring of colour, as well as connecting colour to cultural contexts. Finally, reader's advisory will be utilized to explain the relevance of this thesis' research for libraries.

2.2.1: Reader's advisory

Case studies consisting of in-depth patron interviews and surveys about book selection strategies are common research methods used for uncovering successful reader's advisory, such as Catherine Ross's interviews of nearly 200 heavy readers (Ross, 2000, p. 7). Saarinen and Vakkari (2013) base their studies on two larger research traditions: "information searching in general" and "research on book reading and library use" (p. 737). Ross's study finds that the ability to select books in childhood is important, and having that freedom will lead to a more voracious reader later in life (Ross, 2000, p. 8). In adulthood readers have to balance the choice to spend their leisure time reading, the selection of one book over others, and "the set of choices that gets a particular book home from the library or bookstore and makes it available for reading" (Ross, p. 6). This sounds like many choices, but the freedom to make these decisions is a crucial

aspect in the enjoyment of reading. The inherent pleasure of reading allows a reader to choose this activity over others, and this “pleasure is enhanced when readers are reading something that they have selected” (Ross, p. 8). Further, successful choices are also critically important; this is a skill that readers teach themselves over time, and successful choices “contribute to the bulk of reading experience that enhances the reader’s ability to choose another satisfying book” (Ross, p. 9). The downside is that unsuccessful choices will lessen that desire and result in an individual who is more reluctant to read for pleasure (Ross, p. 12).

In order to increase their success, researchers have found that heavy readers develop complex and elaborate strategies to select a book, while average readers have fewer strategies (Ross, p. 9). While there is disagreement in regards to the importance of the factors that contribute to book selection, the factors themselves appear to be consistent across case studies. Events in daily life as well as the world is the first factor that impact the amount of time a reader has to browse, select and read a book. In addition, their mood affects the type of book they will be interested in selecting (Ross, p. 13; Ooi, & Li Liew, 2011, p. 756-758). More avid readers use systems drawn from “previous experience, knowledge of authors, publishers, cover art, and conventions for promoting books and sometimes depended on a social network of family or friends who recommended and lent books” (Ross, 2000, p. 11), leading the process of selection to eventually become intuitive to them (Ross, p. 10; Ooi, & Li Liew, 2011, p. 740). Less experienced readers, meanwhile, are more reliant on the elements within the book itself, (Ooi, & Li Liew, 2011, p. 740) though avid readers also make use of these elements. Author name and genre are considered the most important, and often used in combination with other elements to give a better indication of a book’s nature (Ross. 2000, p. 12 & 14; Saricks, 2005b, p. 42). Recommendations can factor into book selection, as long as it comes from a trusted source that

has compatible tastes to the reader (Ooi, & Li Liew, 2011, p. 758-759; Ross, 2000, p. 11-12).

Finally, genre is an element of books that helps readers to narrow down their selection. While readers are shown to have discerning tastes when it comes to genre, they draw on their knowledge of genre as a cue to select books as well as looking to the appropriate genre shelving when browsing (Ooi, & Li Liew, 2011, p. 747 & 757). Genre is also used in removing books from selection: for example, a regular reader of romance may also enjoy fantasy, but have no interest in suspense novels (Ooi, & Li Liew, p. 750).

This thesis will focus on the elements used by readers to select books, ones that indicate genre, type, or “feel” that evoke an interest in the reader and can help to explain a reader’s search strategies (Ooi, & Li Liew, p. 737-738). These elements are known as appeal factors or quick identifiers, and relate to a novel’s “pacing, characterization, story line, and atmosphere as well as style” (Saricks, 2005b, p. 40-41; Ross, 2000, p. 14). These appeal factors move beyond subject headings, conveying more meaning “beyond mere subjects and plotlines,” (Saricks, 2005b, p. 42) and acting as “access points” to successful books (Ooi, & Li Liew, p. 737). One major case study by Spiller categorized these appeal factors as text blurbs, parts of text within the book, covers, and titles (Ooi, & Li Liew, p. 740). Ooi and Li Liew add the author’s name to that list (p. 748). Readers of all skill levels often select books based on appeal factors as opposed to subject (Ooi, & Li Liew, p. 757; Saaricks, 2005b, p. 43; Ross, 2000, p. 14). For the purpose of this thesis, appeal and appeal factor are considered interchangeable. Of these appeal factors, the cover and back text blurb will be the most relevant to this thesis. The latter assists in informing readers of the subject of the novel and is one of the most popular factors used in book selection, while the cover is useful when methods of scanning books are used by readers. Oois specifies that “the external appearance of the book... likely hint[s] that the book may be of interest,

whereas the text on the back cover... [is a] criteria of borrowing the novel” (Ooi, & Li Liew, 2011, p. 748). Book covers may also be of greater potential interest to less experienced readers, though there has not been extensive research in this area (Ooi, & Li Liew, p. 751). Regardless, they are said to function in a similar manner to genre, in that they can attract some readers while put off others, particularly in the case of genre covers. It is necessary for readers to be able to read the cover adeptly and use it in conjunction with other identifiers to determine possible deception (Ross, 2000, p. 15-16).

A major component of appeal factors is the idea that they use affect linking to catch a reader’s interest. Affect linking, a theory posited by Gelernter “draws on the idea that we tag our stored memories by their affective qualities (so that our first inkling of a memory is the *feeling* of it). The associative leaps that our minds make in moments of relaxation, he posits, can often be explained by affect links, by the fact that dissimilar events may arouse similar nuances of emotion” (Mackey, 2011, p. 86). Mackey discusses affect linking within the context of picture books, but several times expands her ideas to that of fiction in general. Readers use their own personal experiences and the emotions that have been connected to those events to relate to books, and that fiction specifically is created “to draw out these links between our own memories and the abstraction on the page” (p. 87). Good fiction will evoke these memories and emotions; good appeal factors will therefore function as a mnemonic, or at the very least indicate the novel itself will function in this manner. As Mackey states, the book “represents an interface between an object and two sets of associated experiences: those of the reading event... and those of the experiences evoked by the reader to bring the story to emotional life” (p. 87). By working as an indicator that the book will be this interface, the importance of appeal factors in book selection are further strengthened.

In regard to this thesis, quick indicators and appeal factors fit neatly into the ideas of the theory of semiotics. As it will be discussed in the next section, semiotics play an important role in placing these reader indicators in meaning-making roles. Quick indicators such as colour function as a semiotic sign that conveys a specific message, which to a potential reader may be an appeal factor such as genre. The use of reader's advisory combined with semiotics, paratext and peritext will answer the second research question of this thesis: How is the language in the front cover quotations and back cover summaries used to appeal to potential readers?

2.2.2: Semiotics

In order to argue that intended themes and messages can be found within the cover images and back text blurbs of the collection, semiotics will be employed. Signs and codes, which are respectively defined as objects or concepts “(words, images, objects, etc.) that [refer] to something else” and the systems through which they are conveyed, fall under the study of meaning-making, semiotics (Moriarty, 2005, p. 227-228; Curtin, 2006, p. 52). Codes may not be immediately visible, but regardless they provide further meaning to a sign (Moriarty, p. 235-236). Due to semiotics being entrenched in structuralism, a form of analysis that examines underlying structures that cause human behaviour and events (Curtin, 2006, p. 52), semiotic interpretation involves deconstruction of sign systems and codes (Moriarty, p. 238). Related to semiotics is semiology; also structuralist in nature, semiology is a theory founded by Ferdinand de Saussure, a Swiss linguist. Semiology incorporates the examination of society's role in semiotics, “the hidden rules which organize anything from how people interact in particular social contexts to how stories are written or told” (Curtin, p. 52-53). Saussure focused on the

relationship of signs: there is the signifier, in the context of this thesis the book covers, and signified, the concept for which the signifier stands for (Moriarty, 2005, p. 228).

Saussure also developed semiosis (Moriarty, p. 228), which philosopher Charles Sanders Peirce builds upon. In Peirce's model of semiosis he states there is the sign, which stands for something, the interpretant, the interpretation an individual has of sign, as well as the object, the thing for which sign stands (Curtin, 2006, p. 53). According to Peirce's idea of semiotics "reality (and thoughts) can only be known through representation via signs, further that this signifying activity can best be explained through a three-part model of sign, interpretant, and object" (Moriarty, 2005, p. 228). Peirce also defines sign in a similar manner to Saussure's signifier, and object similar to signified. He does, however, include a third aspect, the interpretant, which is "the idea evoked in a person's mind by the sign" such as a personal experience (Moriarty, p. 228). In other words, personal experience, emotions and so on "affect our interpretation of the sign and its object and lead to individualized interpretations" (Moriarty, p. 229). The multitude of ways to interpret signs led Peirce to classify them into three main categories: icons, indexes, and symbols. Iconic signs imitate whatever they are representing, such as a photograph. Indexes are signs that have some direct relation to their object – for example, a puddle of water is an index of rain, as it was caused by the rain. Finally, symbols are signs that have no logical relationship with their object, their connection has been decided upon by society at large (Kauppinen-Räsänen & Jauffret, 2017, p. 106-107). Charles Morris further builds upon Saussure and Peirce, but introduces a fourth factor, the interpreter (sometimes blended together with interpretant), and introduces three levels of semiosis: syntactics, wherein the relations of signs to one another are important, semantics, "where the relations between signs and denoted objects are

studied” (and under which icon, index and symbol fall), and pragmatics, which focuses on the relations of signs to their interpreters (Caivano, 1998, p. 391).

Roland Barthes is the last theorist useful within the context of this thesis. Like Saussure he is also a structuralist, having written seminal literature on visual semiotics. Barthes introduces connotation, meaning implied through cultural context, and denotation, literal meaning, into semiotics; this is important in visual communication in order to communicate messages to the individual viewing the image (Moriarty, 2005, p. 231-232, 52). Denoted meaning of an image refers to its immediate visual impact on the viewer, while connoted refers to the cultural meaning the viewer, creator and so on attaches to the image (Curtin, 2006, p. 55). Determining meaning is what occurs during the analysis of semiotics; it is identifying the signified “based on the cues given by the signifier, the sign” (Moriarty, p. 232). Semiotics can also delve into the relationships between individual elements within an image (Curtin, 2006, p. 51). Signs themselves define objects by way of opposite relationships-what something is as well as what something is not (Moriarty, 2005, p. 230-231). This is useful in examining images, as the significance of an image is the result of complex interrelationships between an individual, the image itself, and other factors such as culture, where the cultural meaning an individual gives a sign can be both conscious and subconscious (Curtin, 2006, p. 51-52). Images are unique, presenting challenges that cannot be found in words. Unlike words, their individual elements do not need to be combined in a specific way to form a sign, nor are they as intertwined in their meanings as words (Curtin, p. 56). Images are polysemous; that is, they can have many meanings that can be attached according to an individual’s interpretation (Barthes, 1977, p. 39; Moriarty, 2005, p. 239). In many cases, in order to communicate a specific meaning to a viewer, linguistic messages are often used to ‘fix’ these uncertain signs (Barthes, 1977, p. 39). In some cases, words are used

to relay information that may not be found in the image at all, and words and images are complimentary in communicating one or more messages (Barthes, p. 41).

Colour is a visual cue in which semiotics can be applied. Research has been done on semiotics and colour within a marketing context, though these studies focused on branding and its perception by buyers (Kauppinen-Räsänen & Jauffret, 2017, p. 103). However, their use of sensation and perception can be applied to library patrons. The two stages of processing visual stimuli, sensation has a biochemical nature, and occurs “when the stimulus impinges upon the receptor cells of a sensory organ”. Understanding that sensory information is how perception is defined (Krishna, 2012, p. 334). Researchers in marketing have noted differences between sensation and perception “visual cues evoke sensation before they affect perception, while perception captures consumers’ understanding of sensory information” and that further, the understanding of these visual cues are influenced by past experience (Kauppinen-Räsänen & Jauffret, 2017, p. 102). Caivano argues that colour falls easily within Peirce’s three categories of icons, indexes and symbols (1998, p. 391). Colour works as an iconic sign through association, with Caivano giving an example of “warm” vs. “cold” colours. The warm ones—reds, yellows, oranges—can be found in warm environments, or thermal interactions involving heat such as fire (Caivano, p. 395). Colour as an indexical sign, meanwhile, refers to the idea that the “image” we see of it is directly caused by wavelengths of light and the physical connection of physiological and neurological sensory processes (Caivano, p. 396). Lastly, within cultural contexts colours have many connotations, making them symbols (Caivano, p. 397).

The relationships that colour as signs have with their signified messages will be the main examination in this thesis to answer the following research question: What patterns or relationships can be found between the art in the front covers, the text on the wrappers, and the

genres of the books? In the analysis, recurring colours used in covers of a specific genre will be pointed out and argued that the relationship between the two constitutes association for potential readers, particularly those interested in certain genres, proving that colour is an iconic sign as well as a symbol. Colour as a symbol will be further detailed in the next section on colour theory.

2.2.3: Colour theory

Colour is a fundamental feature of an image, and due to its ability to be measured as well as the vivid use of colour in this specific collection it presents itself as one of the properties to focus on for this thesis. Tools were used to convert colour into measurable data for statistical analysis, and these measurements can be used in different ways that can then be compared to the book genres within the collection or other metadata, or arranged on a chronological scale to view changes in aesthetics over time. Knowledge of colour theory is necessary as it provides a strong framework in which to apply image analysis methodologies.

Theories of colour have been posited since Ancient Greece, including the ideas of primary colours and colour mixing (Shyu & Parkkinen, 2013, p. 4). Seventeenth-century experiments with prisms and light led to the understanding of mechanics of light; that is, its nature as a spectrum of wavelengths. The first models of colour in a circular form were also developed (Shyu & Parkkinen, p. 4-5; Crone, 1999, p. 60-61). Colour science followed in the nineteenth century, with scientists developing theories of human colour vision. Of note are the two theories that today explain the basics of human colour vision. The Young-Helmholz theory of colour vision (or the trichromatic theory) argues that human retinas contain three kinds of colour-sensitive cells that respond to red, green and violet, as well as mixtures of these principle three (Crone, p. 156-158; Shyu & Parkkinen, 2013, p. 5). The Hering theory, which modeled

basic colours in two pairings, red-green and blue-yellow, and that these pairings excluded each other in mixed colours. For example, orange occurs between red and yellow, two absolute (by Hering's standards) colours that are adjacent but not opposed to each other. Meanwhile, there is no colour that is considered a mix of red and green because they are in opposition (Shyu & Parkkinen, p. 5; Crone, 1999, p. 166-167). Hering's theory later became the foundation for the opponent colour theory, and together these theories explain the basics of human colour vision, in that the "Young—Helmholz theory explained color vision on retinal color-sensitive cells level, and the Hering's opponent color theory was explaining color processes later in visual pathway" (Shyu & Parkkinen, 2013, p. 5).

These theories led to the creation of modern representations of colour, as well as attempts to specify and standardize it by the Commission Internationale de L'Eclairage, or CIE (Shyu & Parkkinen, p. 9; Crone, 1999, p. 201). Colour matching, in which two objects that appear to be the same colour are measured, was first used in order to measure human perception, due to the inability to directly measure "the observer's sensation of color" (Shyu & Parkkinen, 2013, p. 8-9). This eventually led to more formal forms of standardization using Colorimetry and the "spectral weighting in the color-matching functions" (Shyu & Parkkinen, p. 12). Using James Clerk Maxwell's arrangement of the trichromatic theory's three primary colours and their additives into a colour triangle model (Crone, 1999, p. 153-154), the colour model RGB was formed (Loesdau, Chabrier & Gabillon, 2013, p. 203). This model is typically used for digital image scanning and processing, and in the case of this thesis the book cover collection was scanned using this setting. RGB is an additive model, meaning that the three primary colours are mixed optically in varying amounts, resulting in one new colour (Plataniotis & Venetsanopoulos, 2013, p. 10). In computational terms, every pixel is given a quantifiable red, green, and blue

value, which combine to form a specific colour, and due to these three values it is best represented in a three-dimensional space, a cube. However, as the RGB model “does not seem to correlate with the human perceptual differentiation between colors” (Loesdau, Chabrier & Gabillon, p. 203) several new colour models were created to better reflect human perception of colour, the HSV and HSL models. Standing for hue, saturation and value, and hue, saturation and lightness, these are the characteristics of luminance (lightness) and chrominance (saturation and value), which are used by humans to interpret colour (Plataniotis & Venetsanopoulos, 2013, p. 2).

Hue is reliant on the dominant wavelength of light that is either produced or reflected off of a surface (Sherin, 2012, p. 11). In these models hue is “measured by the angle around the vertical axis and has a range of values between 0 and 360 degrees beginning with red at 0°. It gives a measure of the spectral composition of a color” (Plataniotis & Venetsanopoulos, 2013, p. 24). Saturation is “relative purity or the amount of white light mixed with a hue” (p. 2). For example, pink is a less saturated form of red, as it is a combination of red and white light. Lightness is the perception in which a surface gives off more or less light, and is determined by another quantity called luminance, “which is radiant power weighted by a spectral sensitivity function that is characteristic of human vision” (p. 2). In the HSV model, value is merely the amount of white or black in a colour. To summarize, “hue and saturation together describe the chrominance. The perception of color is basically determined by luminance and chrominance” (p. 2). These models are represented by a cone rather than a cube, and allows for the “separation of chromatic (Hue and Saturation) and achromatic (Value) information,” so that colour and value information may be independent of each other (Loesdau, Chabrier & Gabillon, 2013, p. 203). A final color space, LAB, represents hue by maintaining the opponent colour theory as it relates to human colour

vision: L refers to visual lightness, A is the green-red colour component, and B refers to the blue-yellow colour component (Shyu & Parkkinen, 2013, p. 14).

In addition to the above models used to explain the physiological relationships between colour and people, research has been done to categorize the psychological connections and how colour relates to affect, cognition and behaviour (Elliot and Maier, 2012, p. 62). Tying strongly into semiotics, specifically the idea of colour as a symbolic sign, researchers across different academic fields have categorized colour according to cultural contexts. Elliot and Maier (2012) have proposed the colour-in-context theory, in which they posit that colour and all of its properties carries psychological meaning beyond simple aesthetics, and therefore the perception of it “influences psychological functioning in a manner consistent with the meaning of the color” (p. 67-68). This meaning is caused by both societal learning and biology, and that is context-specific (p. 69-70). Within this context Elliot and Maier (2012) define psychological functioning as consisting of “affect, cognition and behaviour” (p. 62). The context-specific meaning of a colour is evaluated by someone; depending upon that evaluation (i.e. whether the appraisal is hostile or non-hostile), the individual reaction through their emotions, their behaviour, and specific cognitive processes (Elliot and Maier, p. 67-68). For example, within the context of a reader selecting a book, non-hostile reactions to colour may lead to a book being selected, while hostile reactions would lead to a book being rejected.

For the purposes of this thesis a focus will be on Western/Anglo-Saxon culture, specifically the United Kingdom, the United States, and Canada. The novels within the Harlequin collection were all published for this audience in mind, and therefore the covers were specifically targeted towards a Western mindset. Within linguistics Steinvall analyzes colour-emotion collocations within a large English corpus, seeking to answer both questions of “what

color or colors does emotion X call to mind?” (p. 351) and “what emotion(s) does this color make you think of?” (p. 357). For the purposes of this thesis the categorization assigned for Steinvall’s second question will be examined, as in perusing book covers it is the colour that is first sensed rather than the emotion. Allan continues research within linguistics, examining figurative uses and colour-based metaphors that are then classified as orthophemisms (“straight-talking”, a formal expression used as an alternative to an expression that is not preferred), dyphemisms (“offensive language”, a word or phrase whose connotations are offensive to the listening party or directed object), and euphemisms (“sweet-talking”, a more figurative version of an orthophemism) in addition to cultural contexts (2009, p. 626-627). In marketing communications Aslam “reviews the socio-cultural and psychological associations and meanings of colour(s) in a cross-cultural marketing perspective and outlines their role as an intrinsic or extrinsic cue to the product, package, brand or environment or as a symbol of personality and self-image” (2006, p. 16). Krygier and Wood take a more geographical focus that is nonetheless broad enough to be considered for this thesis in *Making Maps: A Visual Guide to Map Design for GIS*, as they not only include map-related features but cultural concepts (p. 266). The various categorizations given by these researchers can be found in Appendix E, and for this thesis these categories have been combined to create the following table for reference when examining cover colours:

| Colour | Concept |
|--------|--|
| Blue | Joy, sadness, coolness, masculine. |
| Red | Anger, love, passion, masculine, heat. |
| Yellow | Joy, cowardice, heat |
| Black | Sadness, fear, death. |
| Green | Envy, vegetation |

| | |
|--------|---------------------------|
| Grey | Dullness |
| White | Fear, purity |
| Brown | Anger, sadness, dirtiness |
| Purple | Authority, anger |

Table 1 Colours as symbolism for emotions and concepts.

Colour as one aspect of the wrappers that sends a specific message to the potential reader about the entirety of the book will be discussed in the next section, paratext and peritext.

2.2.4: Paratext and peritext

Colour is one of many book wrapper aspects that noted literary theorist Genette (1997) describes as “a certain number of verbal or other productions” (p. 1) in relation to the contents of a book. That is, the “productions” that “surround [the text] and extend it, precisely in order to *present* it” (Genette, p. 1). Genette goes into further detail as he defines these as a book’s paratext: additional information related to a book’s text that are “a threshold, or... a ‘vestibule’ that offers the world at large the possibility of either stepping inside or turning back” (p. 2). Genette focuses on the relations between paratext, and between paratext and text, a change from “classical narratology as the science of texts, of their conventions and their relations—a science which tends to assume a text dissociated from its institutional conditions of production and reception... Rather, he sets out to explore reading as an institutionalized practice, manifested in and consolidated by various paratextual layers” (Seidlmeier, 2018, p. 66-67). Paratext itself may not necessarily be textual, and can include other sources of information such as illustrations (Genette, 1997, p. 7).

Genette’s rubric of defining paratext is fairly simplistic, consisting of “determining its location (the question *where?*); the date of its appearance and, if need be, its disappearance

(*when?*); its mode of existence, verbal or other (*how?*); the characteristics of its situation of communication-its sender and addressee (*from whom? To whom?*); and the functions that its message aims to fulfill (*to do what?*)” (Genette, p. 4). Further, he subdivides paratext into *peritext* and *epitext*, the latter of which are categorized as messages made outside of the book such as interviews (Genette, p. 5). Epitext will not be touched on for this thesis, the main focus will stay on peritext, the messages that are situated “in relation to the location of the text itself: around the text and either within the same volume or at a more respectful (or more prudent) distance” (Genette, p. 4).

As space is significant in defining peritext, it will be discussed in more detail. It is used in two ways within *Paratexts*. First, there is the proximity of paratext to its text and the removal of traditional literary boundaries. Genette (1997) explains this as “an ‘undefined zone’ between the inside and the outside, a zone without any hard and fast boundary on either the inward side (turned toward the text) or the outward side (turned toward the world’s discourse about the text)” (p. 2). Additionally, is the idea of space serving as a connection between the two in the form of their relationship in connection to each other. Genette further states that this zone is “not only of transition but of *transaction*: a privileged place of a pragmatics and a strategy, of an influence on the public, an influence that... is at the service of a better reception for the text and a more pertinent reading of it” (p. 2). Prince (2010) uses this idea of space to explicitly tie it into Barthes’ semiotics “as a kind of space between signifier and signified... [and] two different semiotic positions or even two different ontological levels” (p. 4), and that it isn’t the individual symbol that matters but “the ways in which the signs and texts function within and are generated by describable systems, codes, cultural practices and rituals” (Allen, 2011, p. 93).

Examples of paratext are diverse, including everything from the book title, to author name, to external reviews and commentary by critics and scholars. However, this diversity also means that paratext is always changing, “depending on period, culture, genre, author, work, and edition, with varying degrees of pressure” (Genette, 1997, p. 3). Genette argues that due to this constant change this diversity is less important than their “convergence of effects” (Genette, p. 2). That is, the paratext’s function for and relation to its text, as well as its relation with readers (Genette, 11-12; Birke & Christ, 2013, p. 67). ‘Function’ as defined by Genette is vague, whereas Birke and Christ provide a more detailed explanation of it as “an interplay of three different aspects” (p. 67)—interpretive, commercial, and navigational functionality. Interpretive functionality offers readers further understanding of the text, while commercial functionality advertises or promotes the book. Finally, less relevant to this thesis, is navigational function, a guide for readers to approach and orient themselves within the text (Birke & Christ, p. 67-68). Using these three aspects of function, Birke and Christ summarize that paratext “manages the reader’s purchase, navigation, and interpretation of the text in its specific mediation. Individual elements serve one or more of these functions, which, moreover, closely interact and impact on one another” (Birke & Christ, p. 68).

Genette considers himself a structuralist, with paratext being firmly entrenched in the realm of structuralism based on the emphasis in the above-mentioned relations between it and its text. Sedlmeier and Allen argue that though the texts are considered units of analysis, the importance of that text relies on it also having a relational nature to connected to other “texts and generic convention” (Sedlmeier, 2018, p. 65; Allen, 2011, p. 93-94) within a complex network.

The combined metadata gathered from the collection will form the peritext for the research within this thesis. All of the text-based peritext (such as the back text blurbs) will be

used to analyze the collection in order to answer the research question, how is the language in the front cover quotations and back cover summaries used to appeal to potential readers?

Reoccurring, unique words within genres will be determined in order to show that specific language is used to indicate to potential readers of their genre. Due to the large amount of metadata used for this thesis distant reading rather than close reading will be used as the form of literary analysis.

2.2.5: Distant reading

Falling within the realm of the digital humanities is distant reading, a term first coined by Franco Moretti in his essay “Conjectures on World Literature” and later in his book *Graphs, Maps, Trees* as a new form of literary analysis (Jänicke, Franzini, Cheema, & Scheuermann, 2015 p. 1). As he discusses the limits and issues surrounding *world literature*—namely what it is and how to study it—he concludes that “if you want to look beyond the canon... close reading will not do it. It’s not designed to do it, it’s designed to do the opposite” (Moretti, 2000, p. 57). Distant reading, however “allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes-or genres and systems” (Moretti, p. 57). Distant reading, which does not actually involve reading texts at all, provides a stark contrast to close reading. In using distant reading, one examines a system rather than the text, in the process losing the text but gaining the opportunity to understand relationships, networks and so on between the literature and its greater culture or environment. Moretti later refines his theory in *Graphs, Maps, Trees*, and proposes to bring quantitative methods into literary analysis using charts, graphs and maps to seek out patterns, paths and relationships within literature (Moretti, 2005, p. 9 & 26).

Using quantitative methods, scholars can examine Big Data, such as entire genres, as opposed to a canon of only a few hundred texts, the latter of which he states may not provide an accurate view of literary history. He gives an example of these methods and their shortcomings in his case study of Victorian literature and genre in “Style, Inc.” (2009). Using quantitative stylistics, he examines “a major metamorphosis of eighteenth-century titles and try to explain its causes” followed by “how a new type of title that emerged around 1800 may have changed what readers expected of novels” and lastly an “attempt at quantitative stylistics, examining some strategies by which titles point to specific genres” (Moretti, p. 135-136). Further studies can be found in *Graphs, Maps, Trees*, summarized by Jänicke et al. (2015) as

- graphs to analyze genre change of historical novels,
- maps to illustrate geographical aspects of novels, and
- trees to classify different types of detective stories (p. 3)

Following Moretti’s writings, Erlin and Tatlock (2014) note that distant reading has caused debate over “the correct approach to globalizing [comparative literature]” (p. 14). Erlin and Tatlock praise it as a much needed conceptual framework for literary and cultural analysis that seeks to understand “broader categories... and larger corpora in order to pose questions about the functioning of the literary field as a whole” (p. 15). While the Harlequin collection isn’t quite as large a dataset as Moretti’s book describes, using distant reading to analyse the book covers still applies. Metadata such as the year published, the author of the book itself and, if possible, the artist name is all information that can be quantified and visualized. That the texts themselves won’t be read lends to the credence that distant reading is a method that should be utilized.

2.3: Literature reviews

The final section of this chapter provides a summary of the development of image processing and analysis, as well as explain current image analysis research. As image analysis will form the bulk of the research completed in this thesis, it is important to understand its place within academia.

2.3.1: Image analysis studies

Image processing and analysis have been used in the sciences since before the advent of the personal computer, beginning as early as the 1920s with the Bartlane cable picture transmission system as well as perforated tapes and telegrams (Gonzalez & Woods, 2002, p. 3-4). The need to improve image quality spurred the beginnings of scientific image processing, in which images are prepared through digital manipulation “for the measurement and analysis of the features and structures that they reveal” (Russ & Neal, 2016, p. xiii). During the 1960s and 1970s with the development of computers came one of the earliest examples of image processing and analysis software and the predecessor to ImageJ, NIH Image. This software was developed as “a low-cost image-analysis system that the average bench scientist could afford and deploy” (Schneider, Rasband & Eliceiri, 2012 p. 671) in the 1970s by Wayne Rasband who strove to create a “general-purpose extensible image-analysis program that could be used by anyone who wanted to capture, display and enhance images” (Schneider et al, p. 672). Though his intention was not to target a specific field, it quickly found a place in medical research “because autoradiographs, computed axial tomography or positron emission tomography scans and X-rays could be viewed, analyzed and notated” (Schneider et al, p. 672). Much like ImageJ, NIH Image

was open source, distributed by its creator Wayne Rasband via free floppy disks, and today has expanded to include microscopic imaging, archaeology, security, earth sciences, and other scientific fields (Gonzalez & Wood, 2002, p. 5-6). Since NIH Image, Rasband has also encouraged users to build on the original tool through the creation of plugins and scripts, and community discussion through forums and mailing lists (Schneider et al, 2012, p. 672). ImageJ is one of the tools chosen for the image analysis because it is open source and its prevalence in the history of image processing/analysis tools.

Though they are not as common as its use in scientific research, there have been arts and humanities studies published that utilize ImageJ and other image analysis tools. For example, the complexities found in abstract art are quantified and measured in *Topological invariants can be used to quantify complexity in abstract paintings* by Elsa and Zenit (2017), who used these measurements to compare Jackson Pollock to other artists. They found that despite all of the artists falling under the abstract art style, Pollock's paintings offer far more complexity than anyone else. Manovich uses an ImageJ plugin, ImagePlot, created by his Software Studies Initiative in his analysis of several different image collections. The first study, co-authored with Douglass and Huber (2011), examined one million scanned and fan-translated manga pages, comparing special pages inserted by fans as well as the original images to their translated counterparts. This comparison allowed the researchers to argue for the legitimacy of fan-translated scans as important cultural objects that are part of that series or title's lifespan (Manovich, Douglass, & Huber, p. 220-221).

Meanwhile, along with Ushizima, Margolis and Douglass, Manovich (2012) proposes a "framework for identifying clusters [of images] based on visual and/or semantic similarity that uses digital image analysis to describe the content" (p. 31) of 340,000 images from Flickr.

Manovich's influence continues as three researchers make use of his methodology of cultural analytics for their own studies into image analysis. Hristova, in *Images as Data: Cultural Analytics and Aby Warburg's Mnemosyne* (2016), may only examine one image, but through image analysis is able to contrast the use of colour and violence, tying together those elements with cultural and historical themes. Skarpelos, in *Towards a Quantitative Visual Semiotics?* (2015) has a similar method of reasoning, arguing that image features reflect culture and history, though his corpus consists of 24,066 album covers. Martinez and Goodall similarly use colour, though within the context of pigment identification of paintings using clustering algorithms in *Colour cluster analysis for pigment identification* (2008). Finally, in their study *Visualizing Instagram: Tracing Cultural Visual Rhythms*, Hochman and Schwartz (2012) create visualizations with ImageJ based on measurements gathered from another processing tool that contained "the means values of brightness and red, green and blue use in each image" (p. 7) from a corpus of 550,000 images downloaded from Instagram. Through these visualizations they are able to "identify reoccurring spatio-temporal visual deviations in a specific time period and a set place... that provide insights into the study of different cultural practices on a local and global scale" (p. 9). Finally, less analysis-based but still within in the realm of image visualizations created from other images is the artwork created by Jason Salavon entitled *Every Playboy Centerfold, The Decades (normalized)*. Created in 1999, these prints are "the result of mean averaging every Playboy centerfold foldout for the four decades beginning Jan. 1960 through Dec. 1999" (*Every Playboy Centerfold, The Decades (normalized)*, 2002, para. 1). Split into four images according to decade, Salavon shows the development of this specific type of portraiture across time.

2.4: Conclusion

Shortly after the creation of a formal reader's advisory services the rise of the mass market paperback began, and with this new format of book came Harlequin Enterprises, the publisher of the collection used in this thesis. Focusing on the surrounding metadata rather than the content of the books (defined by Genette as peritext), distant reading will be utilized due to the large amount of information that needs to be organized and analyzed. The theory of paratext (and within that, peritext), argues that this metadata provides potential readers with messages about the contents of the books in order to manage the reader's expectations of the story. Paratext is rooted in the theory of semiotics, in which objects refer to something else defined by an underlying system. Paratext refers specifically to information surrounding a book, and in the case of the textual information gathering and analyzing this text en masse will answer two questions of this thesis: What patterns or relationships can be found between the art in the front covers, the text on the wrappers, and the genres of the books? How is the language in the front cover quotations and back cover summaries used to appeal to potential readers? Semiotics, meanwhile, is generalized to any objects within a society, though theorists have further explored and categorized semiotics in forms that allow specific kinds of paratext, mainly colour within a cover, to be better defined. This refers to colour as an icon, index and symbol: it is an indexical sign via its wavelengths and the physiological processes we use to perceive it, an iconic sign through direct association, and finally a symbol due to the application of cultural contexts. Colour as an icon and as a symbol has been examined by researchers in different forms, and this thesis will make use of previous studies in which researchers have determined direct associations with colours (such as heat with red), and cultural connotations (such as black with sadness) in order to identify what messages are being sent by the publisher to the potential reader through

the use of specific colour. Researchers of reader's advisory refers to these aspects being used to indicate certain messages as, unsurprisingly, quick indicators. These quick indicators will be tied to genre in particular, as it is to be expected that books within a genre would contain similarities within their text and imagery to better appeal to readers. Essentially, these theories are combined to provide a foundation in which this thesis will take certain metadata of the book collection, understand reoccurring patterns in the text and images, and explain how they send specific messages to readers. Finally, it will be suggested how this metadata could be applied in a practical sense to reader's advisory, to enhance library services.

The next chapter will consist of the analysis. The collection will be described in further detail as well as how the relevant data was gathered. Text and image analysis software and methods will be described and applied in order to provide background for the final chapter, an in-depth discussion of the analysis results.

Chapter 3: Research Design

3.1: Data collection

The collection and data drawn from it are the focus of this chapter. The collection will be discussed in further detail, including information about the history behind the books and publisher, and metadata statistics including information about the genres of the publications being analyzed. The process of digitization will also be described, from decisions about file formats and adjustments to the images such as cropping, to information about the scanning software and equipment. Following this are the text and image analyses portion of this thesis, which explain what tools and methods have been used to investigate the collection data, how these tools and methods work, and why they were chosen for their specific roles. This chapter will conclude with a summary which will lead into the interpretation of the data in chapter four.

As Manovich states, images themselves are dependent on the kind of research one wants to accomplish, as the quality of the images tie into and can greatly affect one's research questions (2011, p. 39). While the collection being analyzed for this thesis may not fall under the purview of Big Data compared to other studies using thousands or millions of images, the use of big data methods is necessary as it is not possible to closely analyze almost three hundred images to the same depth as one may with only a handful. Therefore, the digitized covers are intended to be as closely representative of the physical collection as possible, as it is the physical collection that is situated within publication history, not the digital copies. The digitized copies of book covers were created with digital humanities tools and are also to be used by the library in an online exhibit and other computational-based endeavours that the physical collection could not sustain. These kind of uses of digital images present strong arguments as to why accuracy in humanities and arts-based image collections is important, and why researchers should take care

to evaluate their image quality needs in relation to their research. The methodology used in this thesis follows the five steps of big data process: acquisition of the data by way of the agreement with the Peel library; the use of OCR, PHP, and scanning for data extraction and cleaning; aggregation and representation of the images as data through recording to spreadsheets; modeling and analysis of the images completed through clustering methods, and the creation of visualizations using *R* and *Voyant*; finally, interpretation by examining the visualizations (Gandomi & Haider, 2015, p. 141). The focus on colour measurements and text was purposeful, as other forms of analysis such as object recognition proved too complex for a non-computer science project, and broadened the research scope too much to be considered reasonable for a master's thesis.

3.1.1: Collection description

The corpus that is used for this analysis consists of two hundred and eighty-six novels published by Harlequin Enterprises from 1949 to 1968, numbered from 1 to 965. The majority of the collection consists of books that were published by Harlequin during the early 1950s. The collection currently resides at the Bruce Peel Special Collections and Archives at the University of Alberta, originally purchased from an antiquarian bookstore in Calgary, Alberta. It is a partial collection, and is in the process of being filled out; the intention of the Peel library is to slowly purchase first editions of the first thousand books published by Harlequin Enterprises, due to the company's importance in the history of publication, marketing and selling of books in Western Canada. For the most part the books are in fine condition for their age and quality of materials; there are only several that have any notable damage (under one dozen), such as seven with writing on the cover. There has been minimal fading of the covers, which is important to note due to the focus of this thesis on the covers and their colour.

The corpus used for this thesis is comprised of peritextual elements rather than the interior text. Specifically, this analysis will use the front cover images including titles, front cover quotations (when they are available), and back cover summaries of each book. In addition, publication date has been recorded from the interior of each book, and author information has been gathered from the collections as well as several external resources detailed below in Table 3. The nature of the data I am able to collect is due to the fragility of the books. Because they were made with poor quality paper and glue the digitization of the interior pages has been deemed not feasible by the Peel library. Publication dates are the only data to be taken from interior pages, and have been transcribed by hand. The full wrappers—that is, the front cover, back cover and spine—have been digitized by the author of this thesis. Notably, the decision was made not to include the spines in any analysis as they do not carry any information that cannot already be found on the front or back covers, and due to the bending of the spines during reading their quality was poor compared to the covers.

The information from the texts has been entered into several spreadsheets for ease of organization, and the texts are split up into separate text documents based upon publication year and author gender. Author names have been taken from the wrappers and supplemented with further information found on the internet. Full names, pseudonyms and author gender were determined using a mixture of Wikipedia entries, online obituaries and the online bibliographic database FantasticFiction (*Fantastic Fiction*, n.d.).

Following this data collection, each book has been categorized into the following genres as outlined below in Table 2. These genres were determined according to Library of Congress' online catalogue, Trove (the National Library of Australia's online catalogue), and when necessary, records from the Online Computer Library Center's (OCLC) Worldcat. Subject

headings found in each library for each book were recorded, and from this information a smaller list of genres was created. The history of Harlequin Enterprises' publications were taken into consideration when creating the genres: for example, general romance was separated from doctor/nurse fiction due to the importance of the doctor/nurse genre in the company's success and growth into the romance novel industry. Up to three genres were assigned per book, though efforts were made to keep the number between one and two; the majority only received one genre classification.

| Genre | Number of books | Books with multiple genres |
|-------------------------------|------------------------|--|
| Detective and mystery fiction | 42 | 0 overlapping |
| Western/ranch fiction | 32 | 2 overlapping with doctor/nurse |
| Biographical fiction | 8 | 1 overlapping with historical fiction 2 overlapping with international location |
| Historical fiction | 12 | 2 overlapping with international location 1 overlapping with biographical fiction |
| Science fiction | 5 | 0 overlapping |
| Sports fiction | 3 | 0 overlapping |
| General fiction | 65 | 6 overlapping with international location 3 overlapping with general romance |
| Doctor/nurse romance | 77 | 2 overlapping with western/ranch fiction |
| General romance | 15 | 4 overlapping with international location 3 overlapping with general fiction 2 overlapping with Canadian fiction |
| Nonfiction | 19 | 0 overlapping |
| Canadian fiction | 16 | 2 overlapping with general romance |
| International location | 15 | 8 overlapping with general fiction 4 overlapping with general romance 2 overlapping with historical fiction 2 overlapping with biographical fiction |

Table 2 Genre categories for Harlequin collection

Similarly, in Table 3 below, the number of female, male, and unknown gender authors have also been counted. It should be noted these numbers took the multiple genres into account (for example, an author who has written a book that falls under “general fiction” and “international location” is counted under both), and that in the cases of three books there were two authors. One book listed both an author and editor named on the front cover and in that instance, both were counted as authors. In the case of one novel listed under science fiction, scans were not available and therefore not included in the image analysis.

| Genre | Female | Male | Unknown |
|-------------------------------|---------------|-------------|----------------|
| Detective and mystery stories | 6 | 36 | 0 |
| Western/ranch stories | 2 | 32 | 1 |
| Biographical fiction | 5 | 3 | 0 |
| Historical fiction | 6 | 6 | 0 |
| Science fiction | 0 | 5 | 0 |
| Sports stories | 0 | 2 | 1 |
| General fiction | 20 | 47 | 0 |
| Doctor/nurse fiction | 68 | 8 | 1 |
| General romance | 11 | 4 | 0 |
| Nonfiction | 4 | 13 | 4 |
| Canadian fiction | 6 | 10 | 0 |
| International location | 2 | 12 | 0 |

Table 3 Female, male and unknown gender author count by genre.

Table 4 provides an overview of the various tools and methods that will be used in the data collection, analysis and visualization for this thesis. As there are many methods and tools being used, some of which overlap and many of which do not, this table can be used as a reference for chapters three and four.

| Type of data | Tools/methods used for data collection | Tools/methods used for data analysis | Tools/methods used for data visualization |
|------------------------|---|---|--|
| Front cover art | Book2Net scanner and BookExpert scanning software: TIF images Color Summarizer: LAB, HSV, hex code measurements, aggregate statistics ImageMeasure: standard deviations and medians of brightness, saturation and hue | <i>R</i> library <i>cluster</i> 's silhouette method: number of clusters Color Summarizer's k-means clustering: average 5 colours in every image | R libraries: <i>quickpalette</i> , <i>ggplot2</i> , <i>xml2</i> Adobe Photoshop CS5's Auto-Blend ImagePlot |
| Front cover quotations | Book2Net scanner and BookExpert scanning software: TIF images OneNote OCR: text | Relative word frequency | Voyant tools: Cirrus, bubbleline chart |
| Back cover text | Book2Net scanner and BookExpert scanning software: TIF images OneNote OCR: text | Relative word frequency | Voyant tools: Cirrus, bubbleline chart |

Table 4 Summary of methods and tools used in this thesis.

The remainder of this chapter will further detail each method and tool listed in Table 4, explaining how and why each were used in this research.

3.1.2: Digitization of collection

The very first digital process involved in this thesis was scanning every individual wrapper. Scanning resources were kindly provided for free by the University of Alberta's Digital Initiatives, and in line with support from a research-intensive university, best practices for digitization were followed. This collection was scanned at 300 DPI as 24-bit depth TIFFs, and a Book2Net Kiosk A2 book scanner along with the scanning software BookExpert were both used

in the scanning process. Additional professional lighting on either side of the scanner were used to ensure greater accuracy of the scans. Aside from having Book2Net remotely calibrate the scanner, there was no cleaning done in BookExpert. This is done using a colour sheet that contains a standard colour profile from the International Colour Consortium. The images use an RGB colour model as that is the baseline for the TIF file format. Adobe Photoshop was used to rotate and crop the images as it has a decent automatic crop tool. A macro was created that ran through the directory of images: the macro would open, then crop, and finally save each scan as a new TIF file. After running the macro it was necessary to manually check the cropped images to ensure the cropping was successful. The macro was useful, though there were a number of images that needed to be re-cropped. Often this was the case due to the original scans not being straight. After rotating the images slightly (often by only one or two degrees), the covers were re-cropped manually and saved as new TIF files. Overall, Photoshop was good at detecting the edges of the covers when the images were straight, and using the macro saved hours of work.

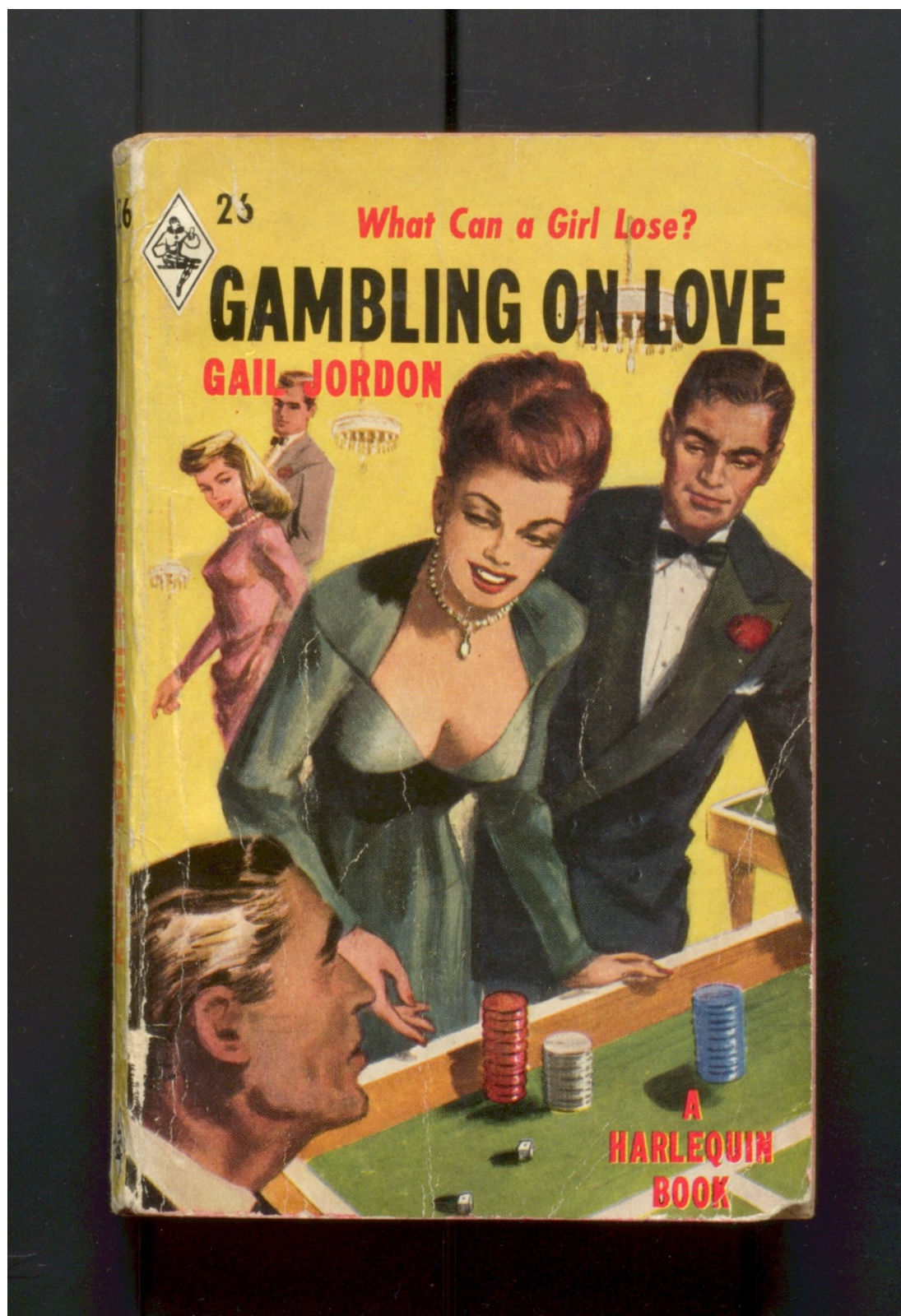


Figure 1 Unedited scan of "Gambling on Love" by Gail Jordon, 1949.



Figure 2 Edited scan of "Gambling on Love" by Gail Jordon, 1949.

The result was two directories: one containing the original unaltered scans, and one containing the rotated and cropped covers. Both directories were backed up twice. The scans of each spine were not edited as the spines themselves were only scanned in order to have the

complete wrapper of every book digitized. The back cover scans were not rotated or cropped in any way as it was not necessary. Unlike the front covers, the back cover art was sparse, often consisting only of the Harlequin logo, two borders, and possibly an abstract shape in a solid colour. In some cases there was a single object related to the story, also in a solid colour. In all cases the use of colour was extremely limited, particularly in comparison to the front covers. Therefore, the information considered relevant to this thesis on the back-covers was limited to the text.

After digitization, a mixture of hand transcription and Microsoft OneNote's optical character recognition (OCR) tool was used to acquire the text from the TIF images. The back cover summaries for the most part did not have this issue due to being black, blue or pink text on a light cream background, and the use of OCR was successful. There were some instances in which a dark-coloured image was printed underneath the text and required hand transcription, and all text needed cleaning by hand to remove spelling and other errors. OneNote was originally used only to test the efficacy of an OCR tool on the scans; since it proved to be such a success it was used on the entire collection. The text was then placed into a spreadsheet. It was not necessary to use OCR on the spines as they only contained titles and authors. In the case of the front cover quotations, the more complex colours and art that served as background behind the front cover quotations caused this text to be difficult for the OCR tool to parse, so it was necessary to hand-transcribe each cover containing a front cover quotation, which was then entered into the same spreadsheet containing the back cover summaries.

3.2: Analysis methods

3.2.1: Text analysis

After the text was converted into machine-encoded text and recorded it was entered into the analysis tool, Voyant, for some preliminary analysis. Voyant, a common tool used in the digital humanities to examine text from hundreds or thousands of books and other materials, was chosen due to its user-friendly interface, its flexibility in analyzing the text and because it offers a variety of visualization outputs for interpretation. Most importantly, it contains several powerful features that are used in the processing of the text. It contains a list of common function words that are then auto-detected and removed to prevent their skewing of the text analysis (Rockwell & Sinclair, 2016d, para. 1-4), and offers tokenization; that is, “the process of identifying words, or sequences of Unicode letter characters that should be considered as a unit” (Rockwell & Sinclair, 2016c, para. 41). Finally, it allows for visualizations based on both relative frequency and raw frequency, a necessity in in-depth text analysis. In the case of the first research question for this thesis, these visualizations are a way to discover any notable patterns or relationships within the text. Voyant’s relative frequency options account for the comparison between two or more uneven lengths of text. In the case of this thesis relative frequencies are used rather than raw frequency for the following comparisons: the text has been split into smaller corpora based upon publication year, genres, as well as male and female authors. The size of these corpora varies, so the use of relative frequency allows these different sizes to be normalized for more accurate comparison. This comparison across genres and authors is another way to determine any notable patterns or relationships within the full corpora of text. This in fact can also answer the second question of this thesis: any consistent language that is found within

the text, particularly within a genre, would have been a deliberate choice by the authors and publisher to appeal to potential readers.

Voyant’s word cloud tool Cirrus (see Figure 3 below as an example) has been used, and word cloud was created using the 105 most frequent words in the back cover text and front cover quotations. The creation of Cirrus was influenced by two other word cloud tools, “Jason Davies’ D3-based Word Cloud library, which in turn is inspired by Wordle” (Rockwell & Sinclair, 2016b, para. 9). The most frequent words found in the corpus are arranged centrally and sized to be the largest, with less frequent words being smaller and less centralized. In Cirrus, the absolute position and colour of the words is not significant, with the latter being chosen by Voyant for aesthetic purposes (Rockwell & Sinclair, 2016b, para. 3). A provided Stopword list is used to remove common words that are not relevant, such as “the” and “and” (Rockwell & Sinclair, 2016b, para. 6).

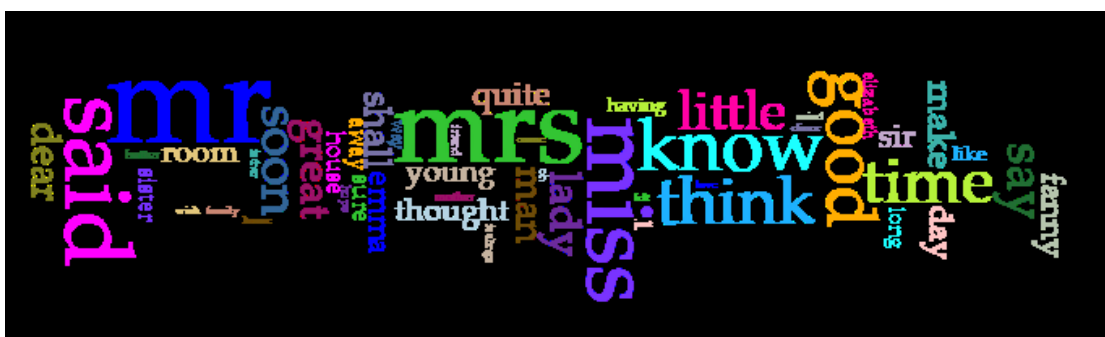


Figure 3 Cirrus with the Works of Jane Austen. From “Cirrus” by G. Rockwell, & S. Sinclair, 2016, <https://voyant-tools.org/docs/#!/guide/cirrus>.

While word clouds are useful in making themes and subjects in a corpus visible, Harris states that their usage may be misleading, that removing the words from their context requires readers to provide their own narrative (2011, para. 10-12). However, Rockwell and Sinclair

argue that in conjunction with other, more in-depth methods of text analysis word clouds can be useful (n.d., para. 10). Following Rockwell and Sinclair’s argument, this thesis will be using a word cloud in addition to other forms of text analysis offered by Voyant, in order to provide a more thorough analysis of the back cover text.

Bubbleline charts (see Figure 2 below as an example) will also be used in this text analysis, in which “each document in the corpus is represented as a horizontal line and divided into segments of equal length (50 segments by default). Each selected word is represented as a bubble with the size of the bubble indicating the word’s frequency in the corresponding segment of text. The larger the bubble the more frequently the word occurs” (Rockwell & Sinclair, 2016a, para. 3).

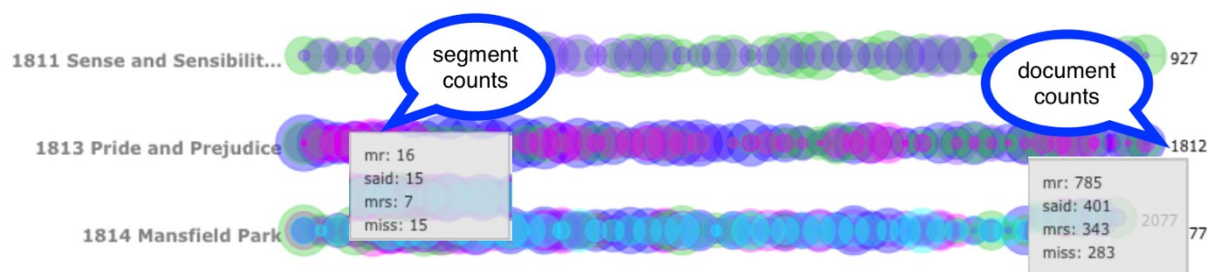


Figure 4 Bubblelines with the Works of Jane Austen. From “Bubblelines” by G. Rockwell, & S. Sinclair, 2016, <https://voyant-tools.org/docs/#!/guide/bubblelines>.

This thesis uses bubblelines as a method of comparison of a small number of words between male and female authors over time. Word clouds and bubbleline charts are examples of methods that use computational technology to analyze massive amounts of text. Importantly, they are a form of distant reading used in the digital humanities, and an alternate form of textual interpretation not found in traditional literary criticism, as they are counter to close reading. The same can be said for image analysis, though as computer technology has only recently become

powerful enough for the average researcher to allow for the processing of massive amounts of images, it is not as common.

3.2.2: Image analysis

As previously discussed during Chapter 2, ImageJ is an example of user-friendly computer technology that allows for mass image processing. ImagePlot and ImageMeasure are two tools run within ImageJ and created by Lev Manovich, a media theorist and director of the Cultural Analytics Lab at the California Institute for Telecommunication and Information and City University of New York (Manovich, n.d., para. 1; Software Studies Initiative and Cultural Analytics Lab, 2016, para. 2) that have been used in several of Manovich's research projects involving image processing and analysis. ImageMeasure is a macro that measures the standard deviations and medians of brightness, saturation and hue of an image. Images are first converted into grayscale for brightness measurements using the following formula

“gray=(red+green+blue)/3” (Manovich, 2011, p.31). This “[eliminates] the hue and saturation information while retaining the luminance”; in other words, the image aspect that is being measured (The Mathworks, Inc., 2018, para. 2). Measurements of saturation and hue are on a 0-255 scale; in the case of hue the degree is mapped from 0-360° to the 0-255 scale (Manovich, 2011, p.31). Combining the red, green and blue channels with the brightness, saturation and hue measurements provides a holistic view of an image's colour, and is ideal for the statistical analysis used in this thesis. ImagePlot is a tool that visualizes images “as a 2D line graph or scatter plot, with the images superimposed over data points” (Manovich, 2011, p. 2). The resulting values outputted by ImageMeasure are used by ImagePlot for the X and Y axes in creating the scatter plots. Multiple image sets may also be compared (Manovich, 2011, p. 34).

Pre-existing metadata can also be entered into ImagePlot for use in visualizing an image collection; in the case of this thesis publication dates and genres have been used. This is particularly useful in viewing changes in the covers over time or comparing genres within the collection. This macro created some minor problems as ImagePlot requires 64-bit Windows, and in the process of upgrading the computer used for the process and analysis of this thesis from 32-bit to 64-bit the Windows backup was corrupted. Thankfully, all the data for this thesis is backed up on an external hard drive, so re-scanning the wrappers and re-collecting the data was not necessary. Only tools such as Color Summarizer, *R*, and Photoshop needed to be downloaded again.

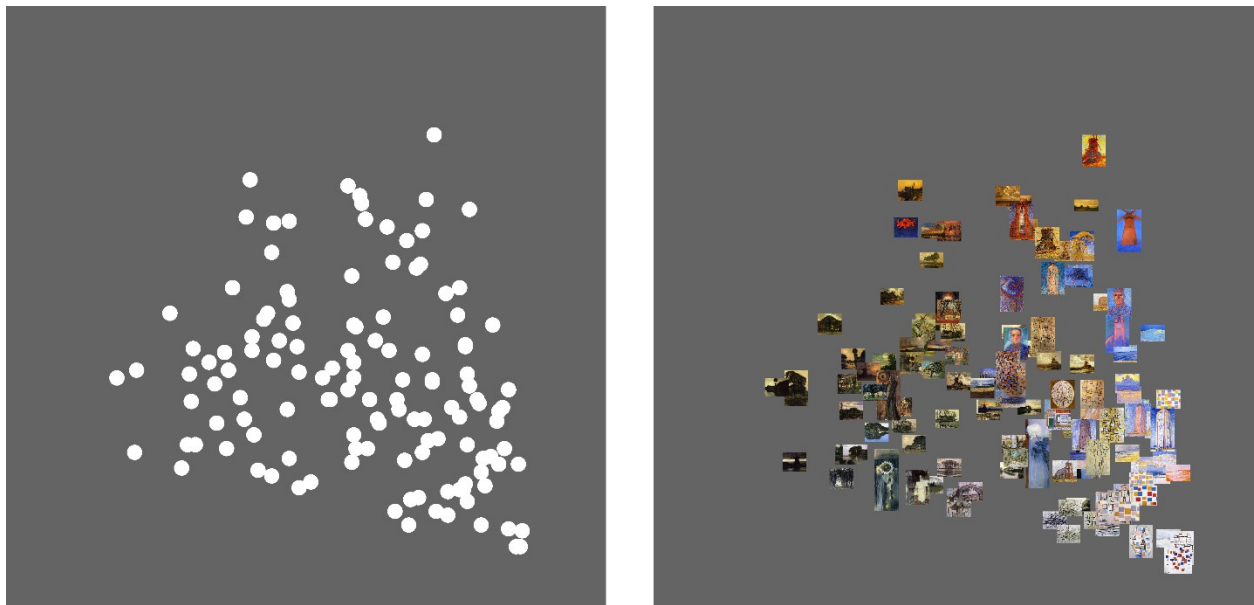


Figure 5 Example of ImageMeasure data used by ImagePlot for visualization of an image collection by Software Studies Initiative, 2011, <https://www.flickr.com/photos/culturevis/6026021275/in/photostream>.

Color Summarizer (version 0.76) is an image analysis and visualization tool that offers a variety of different measurements of an image through an API, an image upload form, and a command line tool. For this thesis the command line tool is utilized as it can be easily integrated into a PHP script to automatically run on a directory of images, with the results then saved

automatically in an XML format. This ability to automate some of the process was necessary in working with a large number of images, as it was not feasible to manually run the command line tool on a single image, and find the resulting data in the created XML file. Further PHP scripts can parse the XML for specific colour measurements. Color Summarizer, created by genomics scientist Martin Krzywinski, breaks an image down by its representative measurements, referred to as “histograms of color components” (Krzywinski, 2018, para. 1). This refers to pixel values, specifically the average, median, minimum and maximum from four color models: RGB (red, green, blue), HSV (hue, saturation, value), LCH (lightness, relative saturation, hue), and LAB (lightness, red-green component, blue-green component) (Krzywinski, 2018, para. 6). RGB, HSV, LCH and LAB components for each pixel are rounded to the nearest integer. Ranges of each component are given on the Color Summarizer website. Hue specifically is represented by either a name (such as red or green) or a single digit, an angle from 0-360° on a standard colour wheel (Wu & Yang, 2017, p.30). Due to this measurement Color Summarizer uses mean of circular quantities to calculate the average hue of an image’s colours (Krzywinski, 2018, para. 8). Value, brightness, lightness, and intensity are measured as follows:

Value and brightness are both defined as the maximum R,G,B component.

Lightness is the midpoint between the maximum and minimum R,G,B values.

Intensity is the average R,G,B value (Krzywinski, 2018, para. 57-59).

This tool is also unique in that it also provides “descriptive statistics for components in each of the color spaces” (Krzywinski, 2018, para. 1), referring to colour names drawn from a large database maintained by the creator that he has compiled from different sources. These names are used in the final measurement Color Summarizer provides that will be used in this thesis, clusters based on similar colours or closest named colour. These are produced using the

LAB model, which approximates human vision, and k-means clustering to determine the average colour of each cluster or its nearest neighbour (Krzywinski, 2018, para. 17-18).

K-means clustering is one of the more well-known clustering algorithms, in which a number of clusters (k) is chosen, and that number becomes the number of ‘centroids’ used by the algorithm (Piech, para. 3). The algorithm then measures data points according to their proximity to these centroids, and assigns centroids to the average middle point within the cluster. The means of the cluster is then used to create a new centroid, and the process repeats until the centroids stop moving. The resulting clusters contain data points that are similar to each other (Mansoor, para. 4; Piech, para. 4; Krzywinski, 2018, para.16-18). In Color Summarizer, the average colour of each cluster is given as a hex code in addition to the number of pixels, the percentage of the image of which the cluster is comprised, the colour that represents the cluster as expressed by a number of colour spaces, and the nearest named colours. This k-clustering method is used in this thesis to produce average palettes for every cover.

Prior to using Color Summarizer’s k-means clustering, R is used to determine the number of clusters used in creating the palette visualizations. The R libraries *XML2* and *cluster* first extract the LAB components from a random selection of the Color Summarizer XML files that contain raw pixel data. LAB values were chosen over others such as RGB or hex code in order to stay consistent with the values used by Color Summarizer’s k-means clustering, which is in turn used as “it is perceptually uniform” (Krzywinski, 2018, para. 17) compared to the other available values. The silhouette method has been chosen to determine the number of clusters, in which every object is assigned a calculated silhouette value between -1 and +1. A value closer to 1 is preferred as this “indicates that the sample is far away from its neighboring cluster and very close to the cluster its assigned”, while -1 means that point is closer to its neighbour, and a 0

value means that the data point is at a boundary between two or more clusters (Alwani, 2015, para. 2). With this in mind, *R*'s *cluster* library can calculate the average silhouette of an image, in which the “optimal number of clusters... is the one that maximize the average silhouette over a range of possible values for *k*” (Kassambara, 2017, para. 15). The library's algorithm computes the average from the minimum to the maximum number of clusters given, and the curve of averages is plotted on a line graph. The cluster with an average closest to 1 is considered the most appropriate number of clusters (Kassambara, para. 17). As this number varied between covers from two to seven, five clusters were ultimately chosen.

Adobe Photoshop CS5.1 (version 12.1)'s Auto-Blend tool has been briefly used to provide an overall look at the front covers in the collection. Using the cropped TIF files, each cover was added as a separate layer to a new image. Layers were not auto-aligned as they had previously been cropped and the step was not necessary. Photoshop has automated the auto-blend process by the simple selection of the tool under the Edit menu, then Auto-Blend, and has three further options available: panorama (which stitches the edges of images together), stack images (for blending images on top of each other), and finally seamless tones and colours for either panorama or stack. Stack images was used for Figures 12 and 13 in the following chapter in order to combine the covers. Photoshop tool is useful in that it allows for a quick and easy way to view reoccurring trends across images. A downside is that as the software is proprietary, the algorithms used for blending (particularly the one used for the seamless tone and colours option) are not able to be viewed and understood by users. The algorithms create a “black box”, so to speak, in which a user simply enters data and receives a result without understanding how that result was reached.



Figure 6 Step 1 example of Auto-blend by G. Ormbo, 2007, <https://shapeded.com/auto-blend-photographs-in-photoshop/>.



Figure 7 Step 2 example of Auto-blend by G. Ornbo, 2007, <https://shapedshed.com/auto-blend-photographs-in-photoshop/>.



Figure 8 Step 3 example of Auto-blend by G. Ornbo, 2007, <https://shapedshed.com/auto-blend-photographs-in-photoshop/>.

The final tool used in this thesis is the statistical computing environment, *R*. In addition to being an environment for statistical analysis and creation of visualizations *R* is a programming language, allowing a user to organize, investigate and interpret their data in many different ways. *R* is also open source and contains a large variety of user-created libraries, making it even more

extensible. Several libraries are utilized for this thesis: *chroma*, *quickpalette*, *ggplot2*, *xml2*, and *cluster*. The library *xml2* allows for a fast and easy way to remove and trim data within an XML file, in this case pixel colour values in the files created by Color Summarizer (Wickham, Hester, & Ooms, 2018, para. 1). Using these values, the library *cluster* can then determine an appropriate number of clusters for an individual cover using the silhouette method (Kassambara, 2017, para. 9). A *quickpalette* function utilizes regular expressions to arrange hex codes (determined by Color Summarizer's use of k-means clustering) into an *R* object (Hvitfeldt, 2018, para. 3), which *chroma* takes to create a colour palette visualization. *chroma* can create multiple rows of palettes that represent a group of covers, in this case a single genre. Finally, *ggplot2* is a robust library that offers a multitude of options in creating more aesthetically pleasing graphs and charts than the *R* default (Tidyverse, 2018, para. 1-3). Scatter plots involving the genres were created using *ggplot2*.

3.3: Conclusion

To summarize, a number of digital humanities methods and tools are being combined to answer the research questions of this thesis. Firstly, colour measurements are acquired using ColorSummarizer and ImageMeasure. The medians, aggregate statistics, and averaged clusters are then determined and visualized for the first research question: the front cover art, as measurements, is categorized according to genre and patterns and relationships within and between genres are then interpreted. Similar methods are used for the front cover quotations and back cover text, and though it is word frequency that is visualized in several different ways, the aim to determine patterns and relationships relating to the genres is still the goal of the text analysis. Additionally, highlighting common words used within genres answers the second

research question; that is, how is the language in the front cover quotations and back cover summaries used to appeal to potential readers? These reoccurring words that stand out within genres typify their genre, and are quick indicators to potential readers who are either looking for that specific genre, or looking to avoid it. Lastly, using these digital humanities methods and tools allows the aggregation and expression of this data in a way that is not possible through traditional literary analysis methods such as close reading; these methods also provide a strong foundation to tie the interpretation of the data into reader's advisory, specifically in pointing out the elements and indicators used by readers in selecting books.

Chapter 4: Analysis Results and Discussion

4.1: Results

The final chapter of this thesis involves the interpretation and discussion of the analysis results from chapter three. The research questions will be reiterated and tied into the image and text analysis. Major themes from the front cover quotations and back cover text will be presented and connected to the collection's genres and author genders, and the argument will be made that this data and the subsequent connections drawn from the analysis function as indicators to potential readers – a return to connecting peritext to reader's advisory. A similar argument will be made for the image analysis portion of this chapter, which follows the text analysis, though the data that will be interpreted involves placement of text and the publisher logo, specific colour values, as well as the colours themselves. Finally, the colours will also be examined in terms of concept and emotion, and what these colours communicate to readers as further indicators as described in reader's advisory. This chapter will finish on a practical note, providing suggestions for how this data could be utilized by a library in order to strengthen web-based reader's advisory services.

With the data organized and methodology described, the next step is to utilize the DH tools in the previous chapter to begin answering the main research questions of this thesis. To reiterate, they are:

- What patterns or relationships can be found between the art in the front covers, the text on the wrappers, and the genres of the books?
- How is the language in the front cover quotations and back cover summaries used to appeal to potential readers?
- How can taking a DH approach assist in reader's advisory?

These questions are answered by examining the text and images in-depth, using a combination of text and image analysis and visualizations. Resultant patterns and relationships brought out by these analyses are then combined with the theories discussed in chapter two (i.e. state the theories here) to form compelling arguments that tie the data to indirect reader's advisory, in which librarians do not supply direct recommendations but resources that readers combine with their own strategies to select books. The final question will take an additional step in offering a hypothetical but practical example of DH methods strengthening pre-existing reader's advisory services.

4.1.1: Text Analysis Results and Discussion

The textual data will be discussed first, beginning with several visualizations that were created from the text gathered from the back cover text and front cover quotations to reveal any noticeable patterns in the collection. The first, Figure 1, is a word cloud representation of the entire corpus, two hundred and eighty-six novels published from 1949 to 1968, created using one of the tools in Voyant.

“woman”, and “men”. Character descriptors (“young”, “nurse”, “doctor”, “beautiful”) are also common, reinforcing the idea of Harlequin’s early publications focusing on character-driven stories. The most common phrases in the corpus support this, consisting of “story of the...” occurring 16 times, and “is a tale of...” occurring 4 times. Several locations are also among the most frequent words, such as “London”, “town”, “hospital”, “island”, and “world”.

Voyant has also been used to determine the frequent words according to genre as shown in Table 4. The top five words for each genre were counted, and in some cases additional words that had the same frequency as the fifth word were included in the analysis. For example, in the genre biographical fiction, “history”, “king”, “life”, and “Rome” were all included as they had the same frequency as “Catherine” and “court”.

| Genre | # of Books | Most frequent words and their wordcount |
|---------------------------|-------------------|---|
| Biographical fiction | 8 | woman (9); henry (7); story (6); catherine (5); court (5); history (5); king (5); life (5); rome (5) |
| Canadian fiction | 16 | story (12); man (9); angus (8); came (8); canadian (8); woman (8); world (8) |
| Detective/mystery fiction | 32 | story (21); murder (18); man (15); london (14); mystery (13) |
| Doctor/nurse romance | 77 | doctor (62); nurse (49); life (47); hospital (46); dr (41) |
| General fiction | 65 | life (38); story (35); love (32); man (28); new (19) |
| General romance | 15 | life (12); love (9); strong (6); gina (5); thought (5) |
| Historical fiction | 12 | story (9); death (8); love (8); life (6); beautiful (5); england (5) |
| International location | 15 | love (9); life (8); island (7); man (6); story (6); strong (6) |
| Nonfiction | 19 | canadian (19); book (16); story (11); canada's (9); life (9); new (9); way (9) |
| Science fiction | 5 | world (9); men (8); women (5); black (4); beautiful (3); civilization (3); house (3); knew (3); stephens (3); woman (3) |
| Sports fiction | 3 | story (5); ball (4); football (4); feel (3); novel (3) |
| Western/ranch fiction | 32 | man (21); gun (20); men (14); life (13); range (13); valley (13) |

Table 5 Most frequent words according to genre.

Word counts are included in brackets for greater clarity, as there is clear difference in frequency of most common terms when comparing a genre with a low number of books (such as historical fiction) to a genre with a high number of books (doctor/nurse, as an example). These most frequent terms have been extracted for further analysis in differentiating genres using front cover quotations and back text blurbs. Names such as “Catherine” and “Gina” occur in such cases where the book title is named after the main character. It is notable that most of these genres contain words that strongly tie to their themes, for example “ball” and “football” being two of the most frequent words found in sports fiction, and “range”, “gun”, and “valley” for western/ranch fiction. A reader can conclude from these individual words the genre of the book, and is evidence of this thesis’ second research question, how language in the form of strongly themed words reoccur within a specific genre to draw in readers who find said genre appealing. This connection between genre and use of frequent themed words can further be extended to the author, specifically their gender. The following figures clearly show that frequently occurring words (ones that are likely going to indicate to reader the genre, or at least the themes, of the book) are more and less common depending on the gender of the author. Figures 2a and 3a represent an example of genre terms compared between male and female authors. Figure 2a uses the five most frequent terms found in doctor/nurse romance novels, while Figure 3a uses the five most frequent terms found in detective and mystery novels.

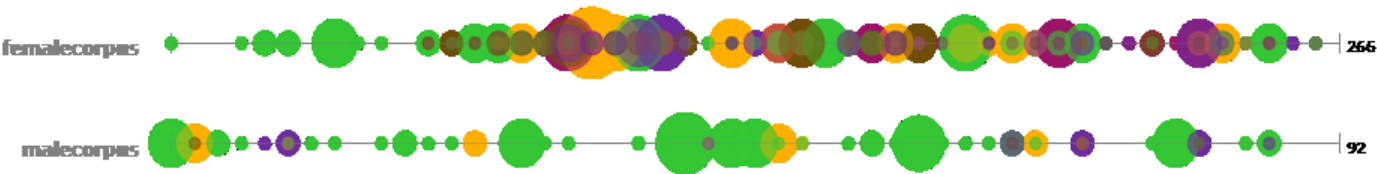


Figure 10 Bubble line chart of most frequent doctor/nurse terms created in Voyant

| Gender | Frequency of doctor/nurse terms |
|----------------|---------------------------------|
| Female authors | 266 |
| Male authors | 92 |

Table 6 Author gender and term frequency for Figure 10






| Term | Colour |
|----------|---|
| Doctor |  |
| Nurse |  |
| Life |  |
| Hospital |  |
| Dr |  |

Table 7 Legend for Figure 10

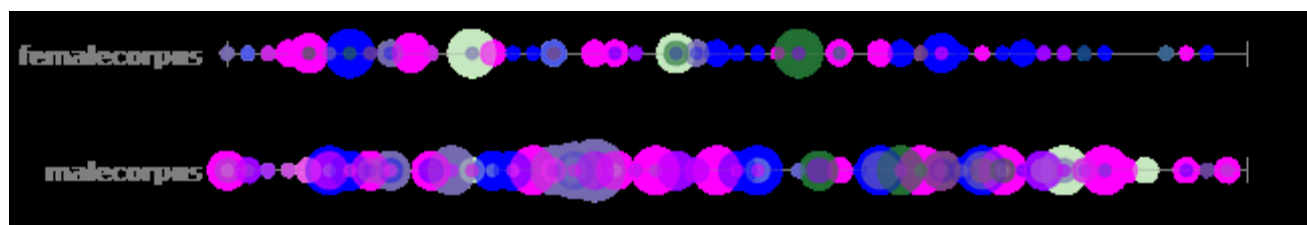


Figure 11 Bubble line chart of most frequent detective and mystery terms created in Voyant

| Gender | Frequency of detective/mystery terms |
|----------------|--------------------------------------|
| Female authors | 91 |
| Male authors | 217 |

Table 8 Author gender and term frequency for Figure 11


| Term | Colour |
|---------|---|
| Story |  |
| Murder |  |
| Man |  |
| London |  |
| Myster* |  |

Table 9 Legend for Figure 11

As it can be plainly seen, there is a clear distinction between female authors using doctor/nurse terms, and male authors gravitating towards the terms commonly found in detective and mystery novels. Regarding women authors, the doctor/nurse terms all become increasingly common over time, while use of these words by men only occur in a few, small, discrete groupings. The exact opposite pattern can be seen when entering terms commonly used in the

detective and mystery genre, with male authors using the terms far more than female authors. Again, when examining the use of the detective and mystery terms by female authors, “story” is fairly common due to it being a frequent occurrence in other genres. Due to the era in which the collection was published it was unlikely back summaries that portrayed violence would have been targeted towards women, and just as with the doctor/nurse romance terms there is a clear distinction in which authors are writing more violent novels. Men vastly outnumber women in this instance, with these terms only appearing in a few places on the female author chart. Additionally, unlike the doctor/nurse romance bubble line chart there is no increase in the use of these terms over time, they instead occur fairly consistently. The sudden increase in use of the doctor/nurse romance terms can be attributed to Harlequin’s partnership with Mills & Boon in the 1950s, and their change in direction towards more romantic-themed novels. While creating these charts it should be noted that along with the change in genres, Harlequin began to publish far more female authors, as this is due to female authors being the dominant force in writing the doctor/nurse romances. For further contrast, male authors dominate the detective and mystery genre.

By examining the Figure 9 world cloud combined with the genre word frequencies, it is clear the Harlequin collection most frequently tried to appeal to its customer base through the prominence of specific characters, giving some detail to them through traits such as occupation and their placement within specific locales. Despite many differing keywords in Table 5, “life” and “story” are extremely common, with one or both appearing in every genre except science fiction. In Figure 9 medical-related terms are of note due to their high frequency and the important historical context in which the medical romance genre is placed. In Figures 10 and 11 it is shown that there appears to be a connection between the gender of the author, the language

they use, and the genre of the book. This is interesting in that, in the cases where the author's full name is a clear indicator of their gender, it can be argued that this in turn gives the patron an idea of what the language used will be, and the language then reinforces the genre as indicated on the cover and in the title. Table 5, which shows word frequency according to genre, shows that even within the small size of the back text blurb keywords are used to clearly communicate genre, so that the reader's expectations of the overall book will be accurate. Similar indicators of genre will be seen in the following results and discussion of this thesis' image analysis.

4.1.2: Image Analysis Results and Discussion

The first visualizations that will be discussed provide an overview of the entire collection of covers. From there, the following figures will provide more in-depth examinations of the covers according to specific measurements such as value medians, and the covers according to genre. The two images that follow, Figures 12 and 13, are an amalgamation of every front cover created using Adobe Photoshop CS5.1's Auto-Blend tool. Figure 13 uses the option for seamless tones and colours, which adjust the colour and tonality for blending. Figure 12, meanwhile, was created without this option, and therefore retains some of the sharp edges and original colours found on the covers. It should be noted along the right and bottom sides are lines in Figure 12 and solid black in Figure 13; this is caused by some of the books having cocked spines, which causes the covers to become unaligned. This condition is caused by the low-quality materials used to produce the books, their age, and the use they have seen over their lifetime. Some standardization of the covers can be seen in both figures.



Figure 12 Auto-blend of Harlequin covers not using seamless tones and colours option.

The placement of the title generally occurs along the top quarter of the cover; in Figure 12 the title appears as mostly white, in generally very light-coloured, while in Figure 13 yellow, white, black and brown appear consistently. These colours reoccur for the title suggests that the colours, much like the title placement, was chosen as deliberately by the publisher in order to remain consistent for readers. Taking these observations to their conclusion, when both visualizations are considered the norm of Harlequin was to standardize the title by placing light-coloured text centrally in the top third of the book. Additionally, the words “A Harlequin Book” can be seen in various places along the bottom of both figures. Much like the title text, “A Harlequin Book” appears white (in some cases yellow) and outlined by black in Figure 12, and light-coloured in Figure 13. Finally, there is similar standardization found in the number of the book, which appears black or white and occurs in the top left or right corners, and the Harlequin Enterprises logo, which is black and white and is located in the top right corner. Standardization of these texts and symbols using colour and placement provide quick indicators of the publisher for potential readers, which became particularly important as Harlequin grew into a monolithic publisher and readers purposely sought out their publications. For some contrast, the covers provide a lot of variation in author name and tagline placement, as they occur essentially everywhere on the covers. Figures 12 and 13 provide a good entry into visualizing the collection, though they rely on viewing similarities and differences rather than using measurements to allow for more in-depth discussion.

Figure 14 below is a comparison of brightness vs. saturation using all the covers created with ImagePlot. ImageMeasure was used to measure the average brightness and average saturation of each cover, the values of which are then used by ImagePlot to arrange the covers as data points on a scatterplot diagram. The X-axis consists of brightness, with lower values

indicating a darker brightness, and higher values indicating a lighter brightness value. Saturation is then found on Y-axis, with lower values specifying duller colours, and higher values specifying more intense colours.

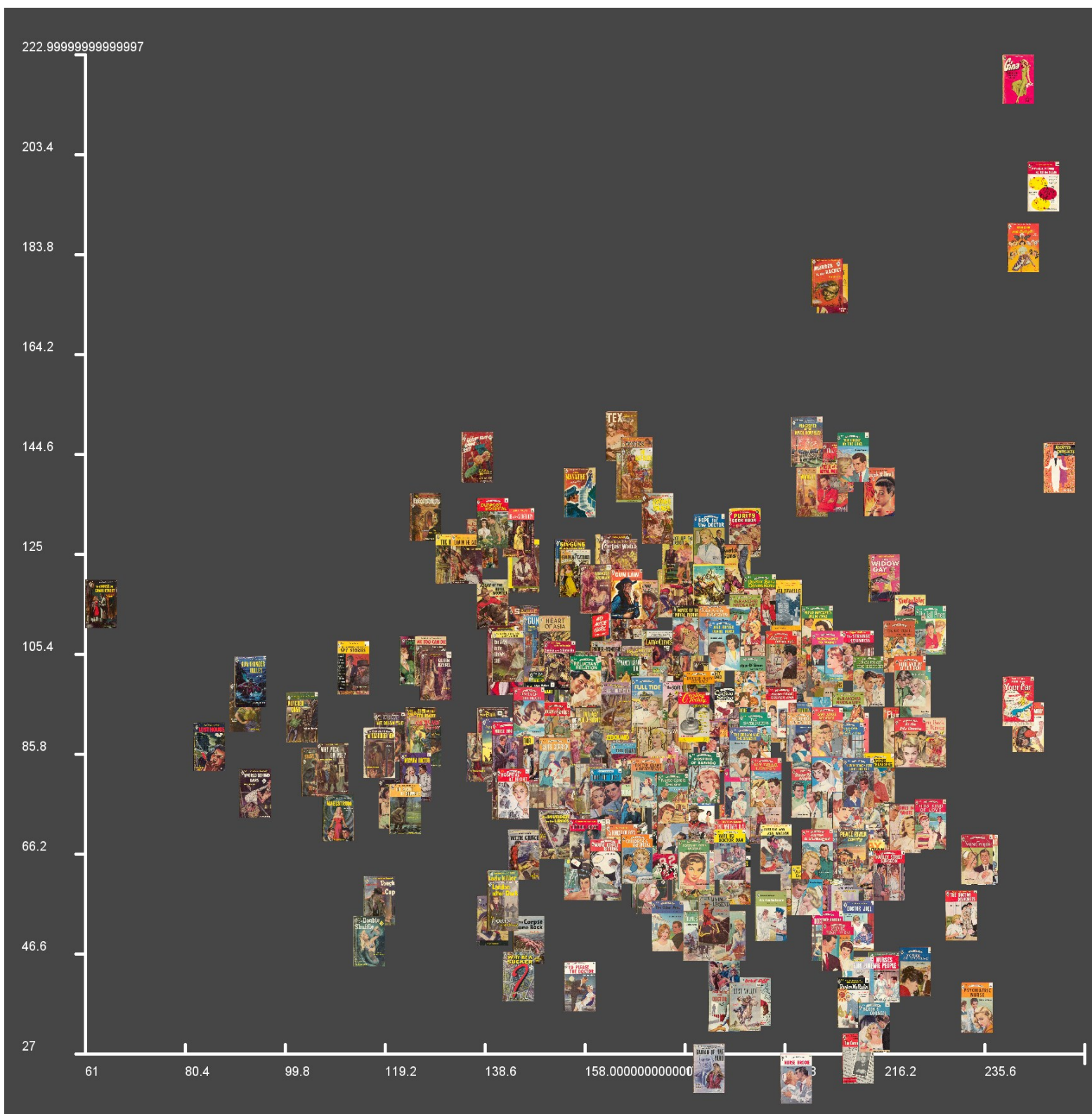


Figure 14 Brightness vs. saturation scatter plot diagram created using ImageMeasure and ImagePlot.

Darker, duller covers are found close to the bottom left corner and brighter, more intense covers towards the top right. While there are some outliers, for the most part, the collection has a similar combined average brightness and saturation, with most of the covers grouped together on the graph. This signifies consistency through the covers, that while there may be great variation in hues there is generally a certain range in lightness and intensity that is chosen by the publisher. Brighter, more intense colours stand out more than darker, duller ones, so the argument can be made the bright and intense covers would catch the eye of a reader. This is another example of using a visualization to view the entire collection in a new way, though this scatterplot utilizes more specific measurements than Figures 12 and 13.

Figure 15 below also uses the brightness medians of the covers, though it adds a chronological element of the collection's publication period determine if there was any kind of change over time. The brightness median (measured on a scale from 0-255, with 0 representing pure black and 255 representing pure white) was again compiled by ImageMeasure, and in this case *R* was used to create the scatterplot graph. Each point represents a single cover, and a line of best fit was added to better represent the two different kinds of data. There is a clear pattern that can be seen in regards to the brightness medians, that the covers generally start out darker and over time they become lighter.

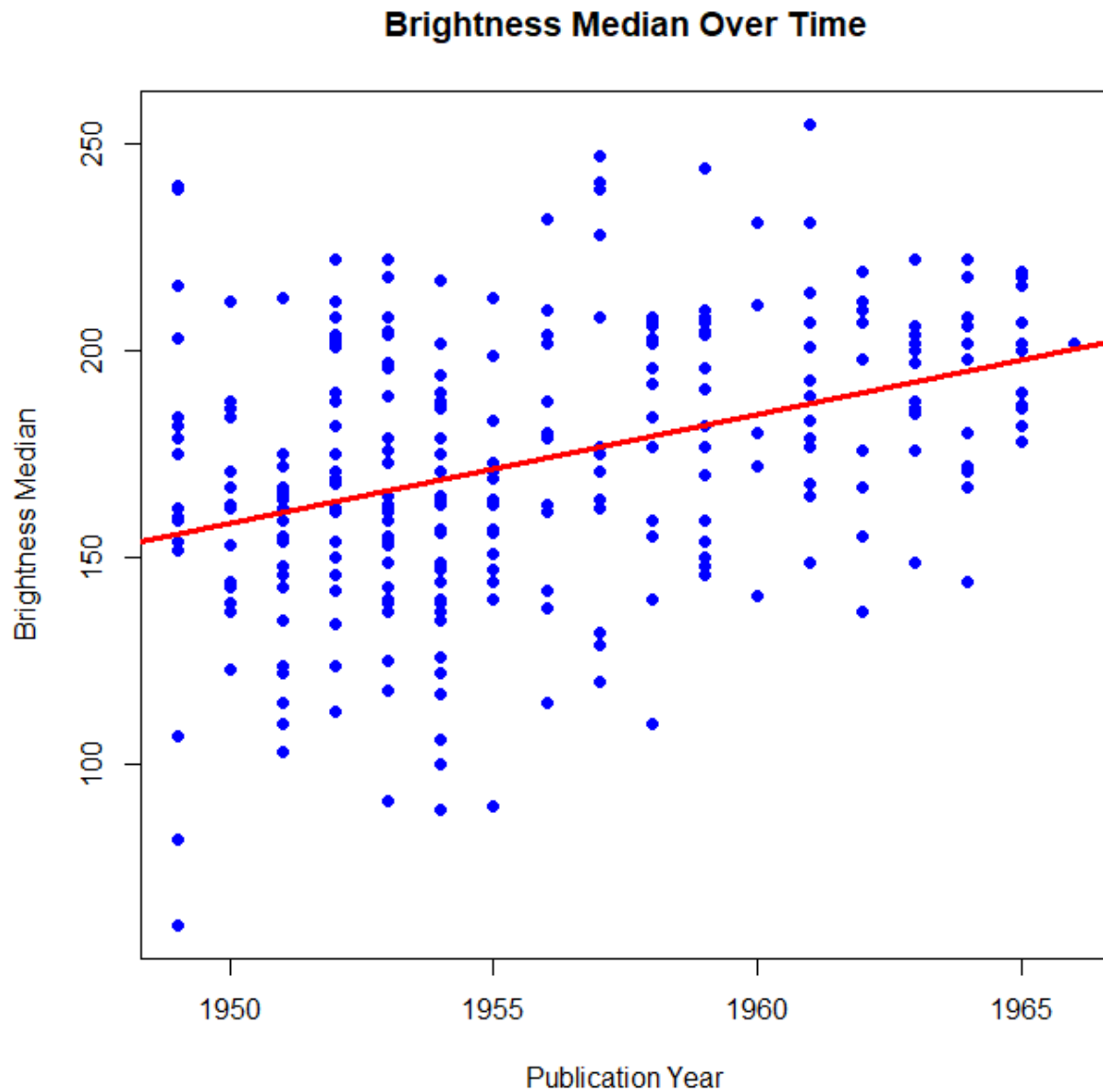


Figure 15 Brightness median over time created using ImageMeasure and R.

Continuing with examining the brightness medians of the covers, Figure 16 groups the collection according to value median (the equivalent to brightness median) to genre. In this case each cover was assigned only one genre; in cases where books had multiple genre the first (considered the most appropriate classification) was used. Within a genre, there tends to be at

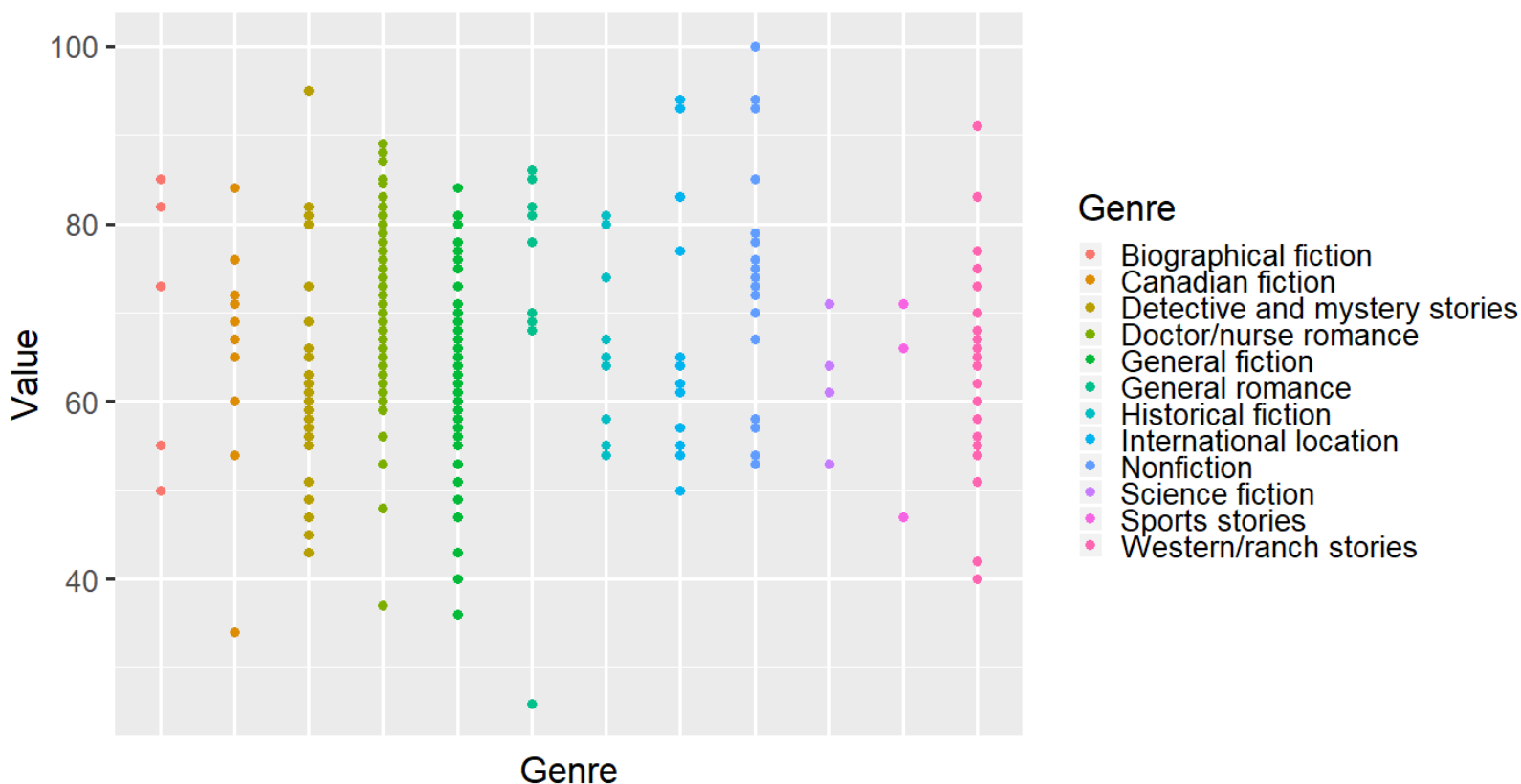


Figure 16 Value median according to genre, created using R

least one major clustering of covers within a specific measured range. However, that range is not consistent when compared across genres: for example, detective/mystery is typically less bright than doctor/nurse fiction. This is a very good indicator that brightness median can contribute to a book's aesthetic quality as an indicator of genre. The downside to this diagram is that the number of covers is not spread evenly across genres, so the genres with only a few covers will not reveal any information as compared to one with a high number of covers. This also means that comparison between genres is not necessarily possible.

Further scatter plots of individual genres were created to compare covers based on their median hue, saturation and values. Again, only one genre was assigned, and in creating these graphs only genres containing ten or more covers were used. Genres with fewer than ten covers suffered from lack of data, meaning that few results could be drawn when compared to genres

with more than ten covers. These scatter plots were created with the intention of identifying any consistencies in colour usage within a genre. Colour palettes were also created using a combination of Color Summarizer's k-means clustering to determine the five average colours per cover, and *R*'s ability to visualize this data. In this case palettes were created for each genre. One line represents one cover, and from left to right is the representation of the most common average colour cluster to the least common average colour cluster.

4.1.2.1: Image Analysis Themes

Consistency is the theme among these two Figures 12 and 13: namely, the title colours and placement of specific text and the Harlequin logo. These placements are quick visual indicators — signs in themselves — of the book's publisher, which in itself is an indication to the patron the kind of book at which they are glancing. This is especially true of the later books, when Harlequin's focus began to narrow away from general reprints and towards romances. One only has to look at either the top left corner or along the bottom to be able to identify that the book is published by Harlequin, and it is easy to then narrow down the book further according to the title (always in the same place), and the book number (in one of the top corners). Figures 12 and 13 also show a lack of placement consistency in regards to the author name and front cover quotations, which is interesting given that, as stated in chapter two, author name is one indicator a reader uses when selecting a book to read. The reason for this lack of consistency in the case of this collection may be that Harlequin, particularly in their later years, began to place more emphasis on their brand as a way of appealing to readers rather than the author name.

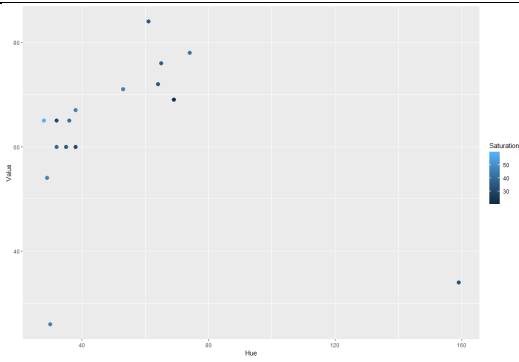
Moving past the text of the front covers and onto the next major sign for patrons (as well as the major focus for this thesis) is the colour. Tying into the examination of the front cover text are the colours chosen. The Harlequin logo is straightforward, being a simple black and white so

as to remain distinct from the rest of the colour. As already discussed, the title text shows a trend towards brighter, warmer colours as well as black and brown. Based on chapter 2's colour code (Table 1), all of these arouse feelings of high energy — fear, joy, anger — while white and yellow are extremely bright colours that stand out. In King's examination of the colour red within human culture, he notes that bright colour and contrast evolved as warning patterns (such as an indicator of poison or venomous defenses) within nature, and include colours such as yellow, white, and alternating colours that strongly contrast each other. He connects this to culture in that "humans have independently developed a similar pattern of color and contrast to attract attention" in cases such as warning labels and signs (p.237). Regarding this collection, it is likely these colours were chosen for two reasons: to elicit a specific emotion, and to attract attention. Based off of Figures 12 and 13 the title text takes up a significant portion of the cover, so using a colour that is either very bright (white, yellow) or bright contrasting with dark (black, brown), would be an effective way to grab the attention of a browsing reader. Given that the covers also tend towards a more intense spectrum, dark text would also be one way for the title to contrast against the rest of the cover and also garner attention.

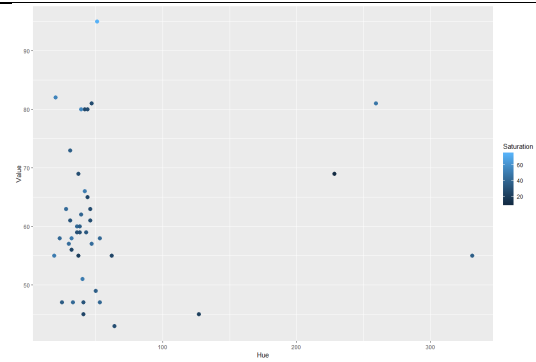
Interestingly, Figure 15 shows that over time there is a gradual increase in brightness across the covers. As brightness on its own does not give much of an indication towards specific colours, additional measurements were examined. Overall Figure 14 shows a trend in average brightness and saturation of the covers. Most are grouped together, with a few outliers. This continues in Figure 16, when the covers are arranged according to genre. Most genres have one or two major clusters of brightness (value on that specific scatter plot graph), followed by some outliers. Genres that contained more than ten covers were also examined on an individual basis according to their median hue, saturation and value, in order to find any trends that occur within

that category. Clearly, the more books within a genre the greater the clarity of the results, but regardless several patterns can be seen.

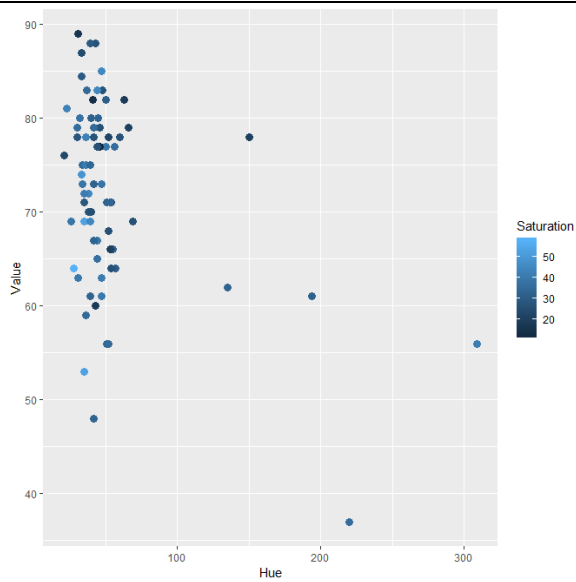
The following section presents several scatter plot graphs showing median HSV measurements within individual genres. The graphs have been resized so that they may be compared against one another, with full size graphs available in Appendix E.



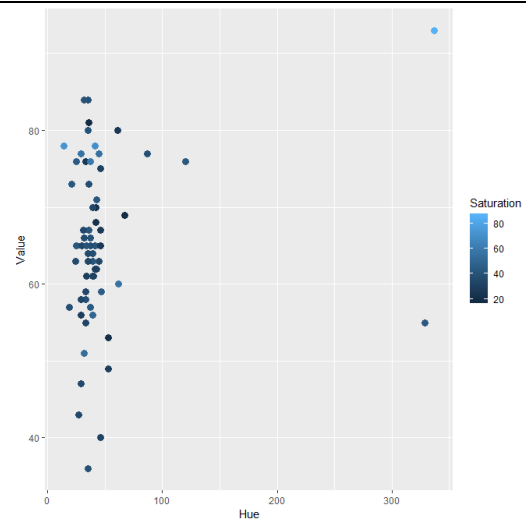
Canadian fiction



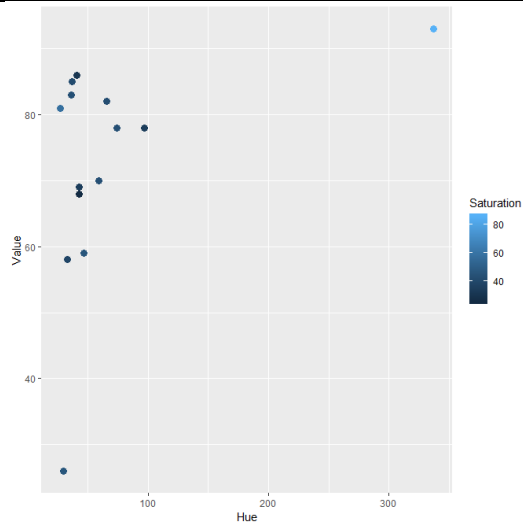
Detective/mystery fiction



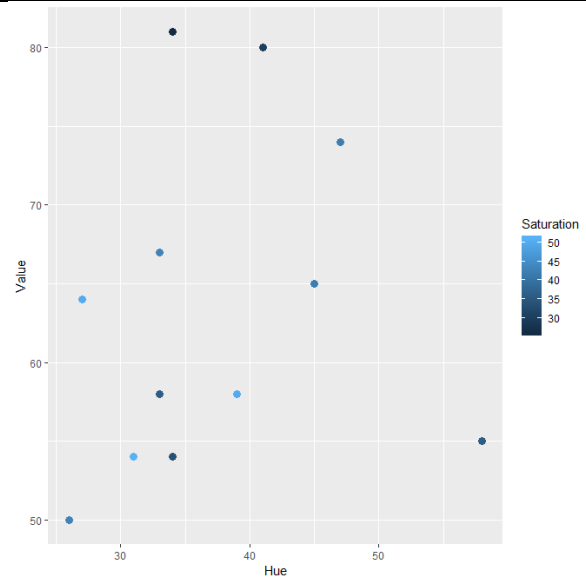
Doctor/nurse romance



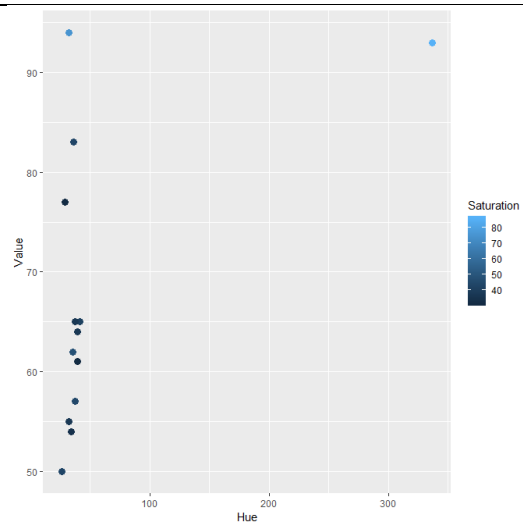
General fiction



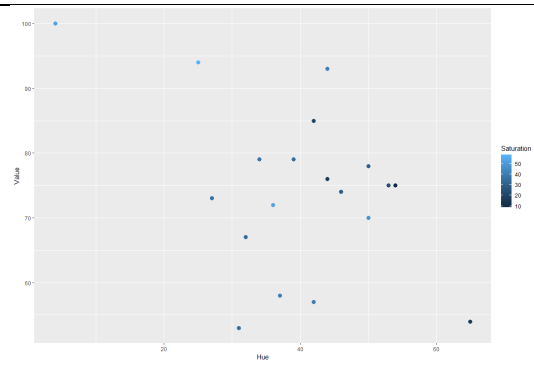
General Romance



Historical fiction



International location



Nonfiction

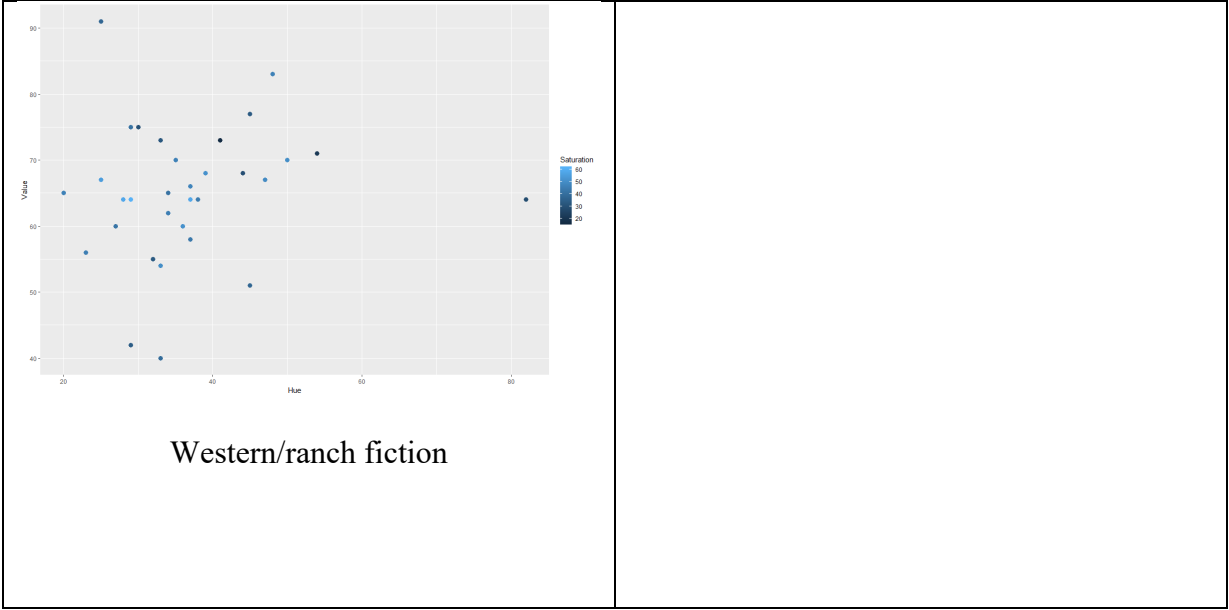


Table 10 Comparison of genres across median HSV scatter plots

Detective/mystery, doctor/nurse romance, and general fiction show obvious clustering of the covers, meaning that their average HSV values are similar. This can also be seen to a lesser extent in the western/ranch, general romance, and Canadian fiction genres; no doubt having more books in these genres would yield clearer results. Additionally, this clustering occurs within the red-orange-yellow hue range. Detective/mystery is notable for being particularly low in value and saturation, while doctor/nurse romances have a high value and mixed saturation. General fiction's value ranges, but saturation appears low, and Canadian fiction has a high value, with mixed saturation. International location has surprisingly dark values and low saturation, western/ranch has a medium value, but high saturation, and finally general romance has mixed values and lower saturation. That these genres have clear clustering, particularly in the cases where there are a large number of books, shows that there is a pattern being considered when creating these covers. There has been some research into the effectiveness of brightness in images, such as Lakens, Fockenberg, Lemmens, Ham and Midden's (2013) five studies into two brightness and valence of images within two databases, the International Affective Picture System and the Geneva Affective Picture Database (p. 1227). Across their studies they found that, overall, brighter images were evaluated more positively than darker ones. This included the same images that had their brightness adjusted to different levels. Another notable conclusion was that images depicting positive concepts or objects were brighter than images depicting negative concepts or objects (p. 1237-1238). While the images in these studies were naturalistic photographs, the human perception of brightness and its association with either positivity or negativity could be applied to this collection. Namely, stories containing darker themes, such as the detective/mystery genre, are darker specifically to associate with negative connotations. In contrast, genres such as the doctor/nurse romances would have brighter covers in order to better

associate with more positive themes such as love. The colour palettes created from the five average colours of each cover offer a deeper look at these colours, which will be discussed in alphabetical order. The colour code system will also be considered.



Figure 17 Biographical colour palette

Biographical, having only eight books, is predominantly browns of varying brightness, along with red and yellow. Given the nature of this genre, it is likely the browns and beiges communicate a more realistic theme or refer to the characters; their story is the central theme to the biographical genre, and taking the time period the collection was published into consideration (1949-1968), these characters were predominantly Caucasian. Red and yellow, meanwhile,

provide a slight amount of excitement and attention through joy or anger, and additionally function as a contrast to the duller browns, effectively working as attention-grabbers.

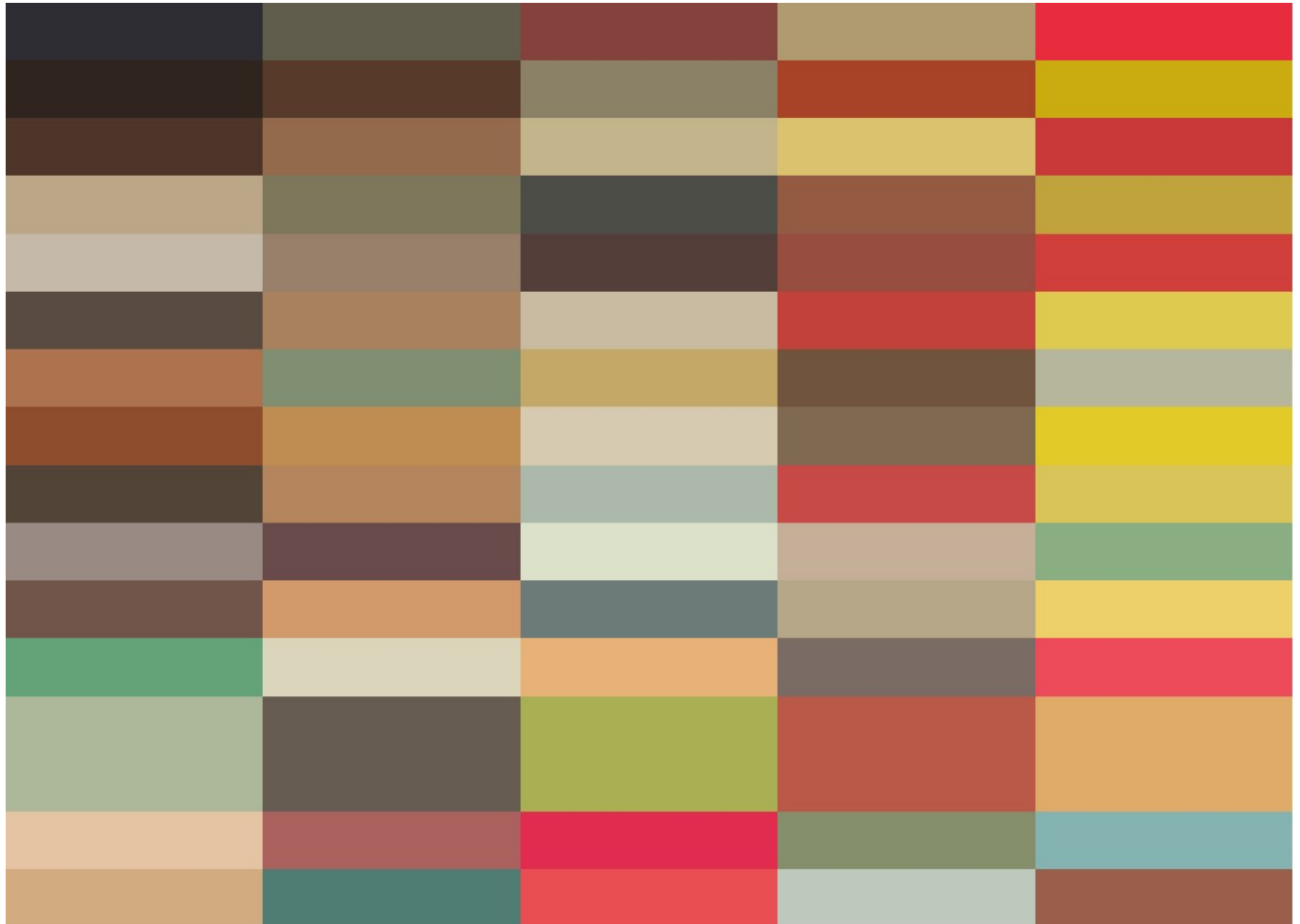


Figure 18 Canadian fiction colour palette

Canadian fiction contains a large number of browns, but also reds and greens. This inclusion of green would suggest the browns all refer to outdoor settings, and of course the red would likely be drawn from the red of the Mountie uniform, as Mountie protagonists are common in this genre. The use of bright red would also play into themes of excitement, given that anger and passion are dominant emotions related to that colour. The brightness of this colour

and the yellows are in line with the HSV scatter plot chart. The darker browns against the bright red would also provide eye-catching contrast to a potential reader.



Figure 19 "Royce of the Royal Mounted" by Amos Moore, 1950 and 1957.

The above example of the Canadian fiction genre shows two different covers of the same title, published seven years apart. In both covers the bright red of the Mountie uniform is extremely prominent. Brown is another visible cover, in the hat and character being manhandled

by (presumably) the Mountie protagonist. It is also worth noting that the title, author name, book number and publisher name and logo are located in consistent places on both book covers.



Figure 20 Detective/mystery fiction colour palette

Detective/mystery, also mostly brown, leans towards darker hues as pointed out in the detective/mystery scatter plot chart. The bright yellows at the end of the palette are stark contrast to the dark browns at the beginning, so it unlikely there is enough yellow in these covers to convey a sense of joy. Detective/mystery likely leans on the more negative emotions and concepts behind brown-anger, sadness, and dirtiness are a few reoccurring concepts supported by the colour category research listed in Appendix E-to convey tough, dark and gritty themes of crime, criminals and murder.



Figure 21 Doctor/nurse romance colour palette

The doctor/nurse genre is interesting in that pale colours dominate the most frequent average colours, and that there is an inclusion of a high number of bright blues, reds and yellows. The pale colours, reds and blues are likely indicators of the medical theming, as these are common colours found within hospitals (with red, of course, being more of a shorthand for blood

and its connotations in a medical context). The combination of blue with yellow would suggest a more positive theming, referring to joy or other positive emotions and concepts.



Figure 22 General fiction colour palette

General fiction is also predominantly brown, but also contains a lot of yellows, and in later covers begin including green and blue. As general fiction is a broad category it is difficult to come to any major conclusions, though like the biographical genre it is likely that the large

amount of beige and brown are related to the characters themselves, who again were predominantly Caucasian. With no strong themes to showcase (unlike others such as the westerns, or detective/mystery fiction), the argument can be made that this collection's general fiction therefore used the main characters to draw in readers.



Figure 23 General romance colour palette



Figure 24 Historical fiction colour palette



Figure 25 International location colour palette

General romance, meanwhile, is a bright mixture of intense colours, including reddish-pink. That colour is likely to suggest passion, desire, possibly femininity to appeal to women readers. The historical genre, in comparison, is not nearly as intense, and browns dominate. The reason for this may be similar to biographical fiction, in that the intent is to convey a realistic setting for the reader. Browns, relating to the earth, in this case would also be connected to a sense of history that typifies this genre. International location is shocking in that there are not

additional brighter colours included to indicate exotic locales, though the presence of orange, an intense and joyous colour, is notable.



Figure 26 Nonfiction colour palette



Figure 27 Science fiction colour palette



Figure 28 Sports fiction colour palette

Nonfiction is a mixture, but contains little green and a large amount of grey. This would communicate to the reader a more moderate theme (and given one of the nonfiction books is a cookbook, unsurprising). Unfortunately, both science fiction and sports contain very few covers, and little to go off of aside from the large amount of browns (and in science fiction's case, reds). More data is needed in order to draw out substantial results.



Figure 29 Western/ranch fiction colour palette

Lastly, it is unsurprising that brown dominate in the western/ranch genre. This type of fiction would obviously utilize settings, characters and objects that contain these colours (cowboy and rancher clothing, cattle, horses, old Western town buildings, farm buildings and structures such as wooden fencing). The inclusion of yellow likely also refers to setting, with western pop culture always occurring in dust-ridden, sometimes desert (hardly ever lush with vegetation) locales. Similar to detective/mystery, the use of browns would communicate tough and gritty themes to the potential reader. In taking an overall look at the palettes, a large amount of earth tones can be seen. This may be related to the backgrounds, which are often either outside or in browns, beiges and yellows, or characters typically being white (due to the time period in which these were published). In keeping the background more neutral tones, this would allow the characters and title text to take the forefront in catching the reader's attention.

4.2: Conclusion

What is the result of all this analysis into the peritext of the collection? This research shows that text and image analysis, digital humanities methods, of non-traditional aspects of books are viable methods of further understanding, describing and categorizing said books. In the case of this Harlequin collection, it has been shown that text such as author names and the publisher titles are consistently displayed on specific sections of the cover, which allows readers to easily find books from their favourite authors (or publisher). In further tying into previously stated reader's advisory selection methods, this text analysis has proven the back cover blurbs use plain, straightforward language to indicate genre. Consistency within genre is further highlighted in the image analysis section through the examination of colour measurements. Patterns of colour can serve as both quick indicators of genre, as well as the mood the book

intends to convey to the reader. As posited by the colour models discussed in this thesis, it can be argued that the use of specific colour, in tying to a specific genre, can then go further and provide an affective link to a pleasant memory a reader may have of successfully choosing and reading a previous book in said genre. All of these appeal factors as well as the affective link are aspects of reader's advisory that allow a reader to not only select a book to read, but to make that selection successful.

Libraries can draw upon this knowledge to improve reader's advisory services in several ways. These text and image measurements, possibly even other digital humanities methods, may be used in traditional discovery and cataloguing tools found in libraries to make their collections more available to patrons. Recommended lists and previously read lists are forms of indirect reader's advisory, and they stand out as a possibility in combining their data with text and image analysis. By using text analysis, an automated tool could use text blurbs to pull out common and uncommon words related to genres. These words can then be compared to the blurbs of previously read books to help build a recommended list. Comparisons may also be made against other forms of metadata, such as author names, publishers, and subject keywords. Particularly in the case of genre fiction, cover colours can indicate the kind of book and comparisons between covers can determine whether two books are similar. These forms of text and image analysis may essentially supplement and fill in "gaps" that may be left by other pre-existing metadata. The end result would be a tool that compares previously read titles to titles in the online catalogue that have not yet been checked out by the patron to discover and build a recommended reading list. This recommended reading list would provide a method of discovery for unread titles that does not involve searching specific keywords. Patrons may not have experience in conducting successful searches, may not be comfortable searching in this way for a variety of reasons, or

may simply not know what they want to search for when seeking new reading. For these reasons this kind of tool could be a valuable alternative way patrons could discover new books within a library.

There are issues, however. This would require either hand transcription or, more likely, an OCR process for older books in order to make their peritextual blurbs, quotations and so on machine readable. What OCR may have been done may not be accurate, and in general this process would require extra effort on the part of the publishers or libraries. Another problem is that raw pixel data may not be easy to acquire due to the quality of the images supplied by producers or publishers. However, high quality images, such as the ones used in this thesis, may not necessarily be required, depending on the image analysis that is being done. Appendix D lists aggregate statistics taken from one image in this collection, “The Manatee”, and compares the measurements to a version of that image that has been resized and saved as a 129x200 pixel JPG, similar to what the Edmonton Public Library uses for cover images in their catalogue (Edmonton Public Library’s BiblioCommons, 2018). The differences are minor, lending to the argument that resized, lower quality images may still be useful in the case of aggregate statistics of colour. Finally, a tool would need to be created that works with many different kinds of software used in library catalogues, which is not an easy accomplishment. Regardless, as online library catalogues become more powerful, utilizing unusual data should be considered in assisting indirect forms of reader’s advisory. Doing so would allow readers to empower themselves in the selection of reading material by allowing them to rely more strongly on their own past choices. Using data such as this may also allow for readers who have limited accessibility to a physical library to better utilize online services, as said online services would be, so to speak, “smarter” and more able to meet their needs. Finally, this may also entice readers who are technologically proficient

and used to similar services from online e-commerce companies such as Amazon but are not regular users of their library system. By offering a service with which they are already familiar and comfortable using, the reader would be more likely to make the leap to become a heavy user of their library.

Book wrappers contain a number of signs that deliberately communicate specific messages to the potential reader. Combining these peritextual elements together, the reader is able to connect this book to a past title containing similar characteristics, as well as receive an indication as to the book's contents. This then allows them to judge whether or not reading the book will be a worthwhile venture. This combination of semiotics, paratext, and reader's advisory plot out a common pathway used by many readers in book selection. This thesis takes that path and focuses on the less traditional aspects – peritextual elements and big data methods – to propose their usefulness in assisting libraries with further developing reader's advisory services. First, by examining front cover quotations and back cover text blurbs on a mass scale, distant reading showed that a book's most frequent words and therefore the language that is employed are key indicators of genre. This can also be discerned through the clustering of the most frequent colours used in a cover. Finally, placement of text and other objects such as publisher logo are typically consistent, providing a quick message to the reader that, when combined with all other characteristics discussed, form a holistic description of the book's contents that can be understood by the reader. By taking these aspects and viewing them as additional forms of metadata, libraries could use them in automated tools to supplement their reader's advisory services through recommended reading lists. While these methods would not replace others, they show that peritextual elements can now be examined through computational analysis in new ways that would be a boon to both libraries and their patrons.

Bibliography

- Adobe. (2017). *Combine images with Auto-Blend Layers*. Retrieved from
<https://helpx.adobe.com/photoshop/using/combine-images-auto-blend-layers.html>
- Allan, K. (2009). The Connotations of English Colour Terms: Colour-Based X-Phemisms. *Journal of Pragmatics: An Interdisciplinary Journal of Language Studies*, 41(3), 626–637.
<https://doi.org/10.1016/j.pragma.2008.06.004>
- Allen, G. (2011). Structuralist Approaches: Genette and Riffaterre. In *Intertextuality* (2nd ed, pp. 92–129). Abingdon, Oxon ; New York: Routledge.
- Alwani, K. (2015, November 10). *Using Silhouette Analysis for Selecting the Number of Cluster for K-Means Clustering (Part 2)*. Retrieved from
<https://kapilddatascience.wordpress.com/2015/11/10/using-silhouette-analysis-for-selecting-the-number-of-cluster-for-k-means-clustering/>
- Aslam, M. M. (2006). Are You Selling the Right Colour? A Cross-cultural Review of Colour as a Marketing Cue. *Journal of Marketing Communications*, 12(1), 15–30.
<https://doi.org/10.1080/13527260500247827>
- Barthes, R. (1977). Rhetoric of the Image. In S. Heath (Trans.), *Image-Music-Text: Essays Selected and Translated by Stephen Heath* (pp. 32-51). London: FontanaPress.
- Birke, D., & Christ, B. (2013). Paratext and Digitized Narrative: Mapping the Field. *Narrative*, 21(1), 65–87. <https://doi.org/10.1353/nar.2013.0003>
- Caivano, J. L. (1998). Color and semiotics: A two-way street. *Color Research & Application*, 23(6), 390–401. [https://doi.org/10.1002/\(SICI\)1520-6378\(199812\)23:6<390::AID-COL7>3.0.CO;2-#](https://doi.org/10.1002/(SICI)1520-6378(199812)23:6<390::AID-COL7>3.0.CO;2-#)

- Crone, R. A. (1999). The trichromatic theory. In *A History of Color: The Evolution of Theories of Light and Color*. Springer Netherlands. Retrieved from www.springer.com/gp/book/9789401539418
- Crowley, B. (2005). Rediscovering the History of Reader's advisory Service. *Public Libraries*, 44(1), 37–41.
- Curtin, B. (2006). Semiotics and visual representation. *International Program in Design and Architecture*, 51–62.
- Edmonton Public Library's BiblioCommons. (2018). *Caddyshack: The Making of A Hollywood Cinderella Story*. Retrieved from <https://epl.bibliocommons.com/item/show/2084665005>
- Elliot, A. J., & Maier, M. A. (2012). Color-in-context theory. In *Advances in experimental social psychology* (Vol. 45, pp. 61-125). Academic Press. <https://doi.org/10.1016/B978-0-12-394286-9.00002-0>
- Elsa, M., & Zenit, R. (2017). Topological invariants can be used to quantify complexity in abstract paintings. *Knowledge-Based Systems*, 126, 48-55. <https://doi.org/10.1016/j.knosys.2017.03.030>
- Erlin, M., & Tatlock, L. (2014). Introduction: "Distant Reading" and the Historiography of Nineteenth-Century German Literature. In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century* (pp. 10–52). Suffolk, UNITED KINGDOM: Boydell & Brewer. Retrieved from <http://ebookcentral.proquest.com/lib/ualberta/detail.action?docID=1641449>
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>

- Genette, G. (1997). *Paratexts thresholds of interpretation*. Cambridge Cambridge University Press. Retrieved from <https://trove.nla.gov.au/work/23121488>
- Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.
- Grescoe, P. (1997). *The Merchants of Venus: Inside Harlequin and the empire of romance*. Raincoast Books.
- Harris, J. (2011, October 3). *Word clouds considered harmful*. Retrieved from <http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/>
- Hochman, N., & Schwartz, R. (2012, June). Visualizing instagram: Tracing cultural visual rhythms. In *Proceedings of the workshop on Social Media Visualization (SocMedVis) in conjunction with the sixth international AAAI conference on Weblogs and Social Media (ICWSM-12)*, 6-9. Retrieved from <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4782/5091>
- Hope, C., & Ryan, J. C. (2014). *Digital Arts: An Introduction to New Media*. Bloomsbury Publishing USA.
- Hristova, S. (2016). Images as Data: Cultural Analytics and Aby Warburg's Mnemosyne. *International Journal for Digital Art History*, 2, 117-132. <http://nbn-resolving.de/urn:nbn:de:bsz:16-dah-234897>
- Hvitfeldt, E. (2018). *quickpalette: R package for quick extraction of color palettes from text and images*. R. Retrieved from <https://github.com/EmilHvitfeldt/quickpalette>
- Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. The Eurographics Association. <https://doi.org/10.2312/eurovisstar.20151113>
- Jensen, M. A. (1984). *Love's Sweet Return: The Harlequin Story*. Popular Press.
- Jordon, G. (1949). *Gambling on Love*. Winnipeg, Manitoba: Harlequin Enterprises.

- Kaestle, C. F. (2007). Seeing the Sites: Readers, Publishers and Local Print Cultures in 1880. In C. F. Kaestle & J. A. Radway (Eds.), *A history of the book in America*. Chapel Hill: Published in association with the American Antiquarian Society by the University of North Carolina Press.
- Kassambara, A. (2017). *Determining the optimal number of clusters: 3 must known methods - Unsupervised Machine Learning*. Retrieved from <http://www.sthda.com/english/wiki/print.php?id=239#average-silhouette-method>
- Kauppinen-Räsänen, H., & Jauffret, M.-N. (2017). Using colour semiotics to explore colour meanings. *Qualitative Market Research: An International Journal*, 21(1), 101–117. <https://doi.org/10.1108/QMR-03-2016-0033>
- King, T. D. (2005). Human color perception, cognition, and culture: why red is always red. In R. Eschbach & G. G. Marcu (eds.), *Proceedings Volume 5667, Color Imaging X: Processing, Hardcopy, and Applications* (p.234-242). California: Electronic Imaging 2005. <https://doi.org/10.1117/12.597146>
- Krishna, A. (2011). An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *Journal of Consumer Psychology*, 22(3), 332–351. <https://doi.org/10.1016/j.jcps.2011.08.003>
- Krygier, J., & Wood, D. (2016). Color on Maps. In *Making Maps, Third Edition: A Visual Guide to Map Design for GIS* (pp. 252–282). New York, United States: Guilford Publications. Retrieved from <http://ebookcentral.proquest.com/lib/ualberta/detail.action?docID=4571765>

- Kryzwinski, M. (2018). *Image Color Summarizer: RGB, HSV, LCH & Lab image color statistics and clustering - simply and easy FAQ*. Retrieved from <http://mkweb.bcgsc.ca/color-summarizer/?faq>
- Lakens, D., Fockenberg, D. A., Lemmens, K. P., Ham, J., & Midden, C. J. (2013). Brightness differences influence the evaluation of affective pictures. *Cognition & emotion*, 27(7), 1225-1246. <https://doi.org/10.1080/02699931.2013.781501>
- Loesdau, M., Chabrier, S., & Gabillon, A. (2014). Hue and Saturation in the RGB Color Space. In A. Elmoataz, O. Lezoray, F. Nouboud, & D. Mamass (Eds.), *Image and Signal Processing* (Vol. 8509, pp. 203–212). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-07998-1_23
- Mackey, M. (2011). Readers Remember: Text, Residue, and Periphery. In C. Mitchell, T. Strong-Wilson, K. Pithouse, & S. Allnutt (Eds.), *Memory and pedagogy* (pp. 81–97). New York: Routledge.
- Manovich, Lev. (2011). *ImagePlot Documentation*. Retrieved from https://docs.google.com/document/d/1zkeik0v2LJmi1TOK4OxT7dVKJO7oCmx_fNP8SYdTG-U/edit?hl=en_US#
- Manovich, L., Douglass, J., & Huber, W. (2011). Understanding scanlation: how to read one million fan-translated manga pages. *Image & Narrative*, 12(1), 206-228. Retrieved from <http://www.imageandnarrative.be/index.php/imagenarrative/article/view/133>
- Manovich, L. (n.d.). *About Lev Manovich*. Retrieved from <http://manovich.net/index.php/about>
- Mansoor, U. (2017, February 19). *Cluster Analysis Using K-means Explained*. Retrieved from <https://codeahoy.com/2017/02/19/cluster-analysis-using-k-means-explained/>
- Moore, A. (1950). *Royce of the Royal Mounted*. Winnipeg, Manitoba: Harlequin Enterprises.

- Moore, A. (1957). *Royce of the Royal Mounted*. Winnipeg, Manitoba: Harlequin Enterprises.
- Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, (1), 54–68.
- Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
- Moretti, F. (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). *Critical Inquiry*, 36(1), 134–158. <https://doi.org/10.1086/606125>
- Moriarty, S. (2005). Visual Semiotics Theory. In K. Smith, S. Moriarty, G. Barbatsis, & K. Kenney (Eds.), *Handbook of visual communication research: theory, methods, and media* (pp. 227–242). Mahwah, N.J: L. Erlbaum.
- National Library of Australia. (n.d.). *Trove*. Retrieved July 5, 2018, from <https://trove.nla.gov.au/>
- Online Computer Library Center, Inc. (n.d.). *WorldCat.org: The World's Largest Library Catalog*. Retrieved July 5, 2018, from <http://www.worldcat.org/>
- Ooi, K., & Li Liew, C. (2011). Selecting fiction as part of everyday life information seeking. *Journal of Documentation*, 67(5), 748–772. <https://doi.org/10.1108/00220411111164655>
- Ornbo, G. (2007). After the Rain by Julie La France and St. Paul's Kirche by Joachim S. Muller combined. [Digital image]. Retrieved from <https://shapeshed.com/auto-blend-photographs-in-photoshop/>
- Ornbo, G. (2007). After the Rain by Julie La France and St. Paul's Kirche by Joachim S. Muller cropped and edited. [Digital image]. Retrieved from <https://shapeshed.com/auto-blend-photographs-in-photoshop/>
- Ornbo, G. (2007). After the Rain by Julie La France and St. Paul's Kirche by Joachim S. Muller auto-blended. [Digital image]. Retrieved from <https://shapeshed.com/auto-blend-photographs-in-photoshop/>

Piech, C. (2013). *CS221: K Means*. Retrieved from

<http://stanford.edu/~cpiech/cs221/handouts/kmeans.html>

Plataniotis, K., & Venetsanopoulos, A. N. (2013). Color Spaces. In *Color Image Processing and Applications* (p. 384). Springer Science & Business Media.

Prince, G. (2010). Gérard Genette and the Pleasures of Poetics. *Narrative*, 18(1), 1-5.

<https://doi.org/10.1353/nar.0.0033>

Radway, J. A. (2009). *Reading the Romance: Women, Patriarchy, and Popular Literature*.

University of North Carolina Press. Retrieved from

<https://www.scholars.northwestern.edu/en/publications/reading-the-romance-women-patriarchy-and-popular-literature-5>

Rockwell, G., & Sinclair, S. (2016a). *Bubblelines*. Retrieved from <http://voyant->

[tools.org/docs/#!/guide/bubblelines](http://voyant-tools.org/docs/#!/guide/bubblelines)

Rockwell, G., & Sinclair, S. (2016b). *Cirrus*. Retrieved from <https://voyant->

[tools.org/docs/#!/guide/cirrus](https://voyant-tools.org/docs/#!/guide/cirrus)

Rockwell, G., & Sinclair, S. (2016c). *Creating a Corpus*. Retrieved from <http://voyant->

[tools.org/docs/#!/guide/corpuscreator](http://voyant-tools.org/docs/#!/guide/corpuscreator)

Rockwell, G., & Sinclair, S. (2016d). *Stopwords*. Retrieved from <https://voyant->

[tools.org/docs/#!/guide/stopwords](https://voyant-tools.org/docs/#!/guide/stopwords)

Rockwell, G., & Sinclair, S. (2016e). *Trends*. Retrieved from <http://voyant->

[tools.org/docs/#!/guide/trends](http://voyant-tools.org/docs/#!/guide/trends)

Rockwell, G., & Sinclair, S. (2016f). *Cirrus with the Works of Jane Austen*. [Digital image].

Retrieved from <https://voyant-tools.org/docs/#!/guide/cirrus>

- Rockwell, G., & Sinclair, S. (2016g). Bubblelines with the Works of Jane Austen. [Digital image]. Retrieved from <https://voyant-tools.org/docs/#!/guide/bubblelines>
- Ross, C. S. (2000). Making Choices: What Readers Say About Choosing Books to Read for Pleasure. *The Acquisitions Librarian*, 13(25), 5–21. https://doi.org/10.1300/J101v13n25_02
- Russ, J. C., & Neal, F. B. (2016). *The image processing handbook* (7th ed.). Boca Raton, FL: CRC Press.
- Saarinen, K., & Vakkari, P. (2013). A sign of a good book: readers' methods of accessing fiction in the public library. *Journal of Documentation*, 69(5), 736–754. <https://doi.org/10.1108/JD-04-2012-0041>
- Salavon, J. (2002). Jason Salavon | Every Playboy Centerfold, The Decades (normalized). Retrieved May 20, 2018, from <http://salavon.com/work/EveryPlayboyCenterfoldDecades/>
- Saricks, J. G. (2005a). A History and Introduction. In *Readers' advisory service in the public library* (3rd ed). Chicago: American Library Association.
- Saricks, J. G. (2005b). Articulating a Book's Appeal. In *Readers' advisory service in the public library* (3rd ed). Chicago: American Library Association.
- Schindelin, J., Arganda-Carreras, I., & Frise, E. et al. (2012). Fiji: an open-source platform for biological-image analysis, *Nature methods* 9(7), 676-682. <https://doi.org/10.1038/nmeth.2019>
- Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 9(7), 671-675. <https://doi.org/10.1038/nmeth.2089>
- Sedlmeier, F. (2018). The Paratext and Literary Narration: Authorship, Institutions, Historiographies. *Narrative*, 26(1), 63–80. <https://doi.org/10.1353/nar.2018.0003>

- Sherin, A. (2012). Communicating with Color. In *Design Elements, Color Fundamentals: A Graphic Style Manual for Understanding How Color Affects Design* (pp. 9-48). Rockport.
- Shyu, M. J., & Parkkinen, J. (2013). Fundamentals of color. In *Advanced Color Image Processing and Analysis* (pp. 1-18). Springer New York
- Skarpelos, Y. (2015). Towards A Quantitative Visual Semiotics? *New Semiotics: Between Tradition and Innovation*, 413-423. <https://doi.org/10.24308/iass-2014-042>
- Software Studies Initiative and Cultural Analytics Lab. (2016). *About Cultural Analytics Lab*. Retrieved from <http://lab.culturalanalytics.info/p/about.html>
- Software Studies Initiative. (2011). *ImagePlot_points_images.Mondrian.1905_1917.X_saturation_median.Y_hue_median.c2500.back_100.b210.ponts64.im100*. [Digital image]. <https://www.flickr.com/photos/culturevis/6026021275/in/photostream>
- Steinvall, A. (2007). Colors and Emotions in English. In, R. E. MacLaury, G. V. Paramei, D. Dedrick (Eds.), *Anthropology of Color: Interdisciplinary Multilevel Modeling* (pp. 347-362). John Benjamins Publishing.
- The Fantastic Fiction Team. (n.d.). *Fantastic Fiction*. Retrieved from <https://www.fantasticfiction.com>
- The Mathworks, Inc. (2018). *Convert RGB image or colormap to grayscale - MATLAB rgb2gray*. Retrieved from <https://www.mathworks.com/help/matlab/ref/rgb2gray.html>
- The R Foundation. (n.d.). *R: What is R?* Retrieved July 27, 2018, from <https://www.r-project.org/about.html>

Tidyverse. (2018). *ggplot2: An implementation of the Grammar of Graphics in R*. R, tidyverse.

Retrieved from <https://github.com/tidyverse/ggplot2>

U.S. Congress. (n.d.). *Home | Library of Congress*. Retrieved from <https://www.loc.gov/>

Ushizima, D., Manovich, L., Margolis, T., & Douglass, J. (2012, May). Cultural analytics of large datasets from Flickr. In *Workshop on Social Media Visualization, Dublin, Ireland*.

Retrieved from

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/download/4781/5097>

West III, J. L. W. (2011). Twentieth-century publishing and the rise of the paperback. In L.

Cassuto, C. V. Eby, & B. Reiss (Eds.), *The Cambridge History of the American Novel* (pp.

781–797). Cambridge: Cambridge University Press.

<https://doi.org/10.1017/CHOL9780521899079.052>

West, J. L. W. (1990). *American Authors and the Literary Marketplace Since 1900*. University of Pennsylvania Press.

Wickham, H., Hester, J., & Ooms, J. (2018). *Bindings to libxml2*. C, R infrastructure. Retrieved

from <https://github.com/r-lib/xml2> (Original work published 2015)

Wu, Q., & Yang, W. (2017). Feature Extraction Algorithms to Color Image. In J. Lu & Q. Xu

(Eds.), *Examining Information Retrieval and Image Processing Paradigms in*

Multidisciplinary Contexts (pp. 27-50). IGI Global.

Appendices

Appendix A: PHP Scripts to Run Color Summarizer

Autoload

```
<?php

function __autoload($class) {

    $filename = $class . '.php';
    require_once $filename;
}

?>
```

Running different image measurement options in Color Summarizer

```
<?php

require_once 'autoload.php';

/*$analysis1 = new ColourAnalysis('C:\\filepath');
//$analysis1->csraw();
$analysis1->cshistogram();
$analysis1->cscluster();
$analysis1->csuniformity();*/

$analysis1 = new ColourAnalysisXML('C:\\filepath');
//$analysis1->csraw();
$analysis1->csstats();
//$analysis1->cshistogram();
//$analysis1->cscluster();
//$analysis1->csuniformity();
?>
```

Color Summarizer Class

```
<?php
//create a variable with the path to images
//create separate functions that iterate through dir of images,
run different statistics, dump measurements into csvs in
specific dir
```

```

Class ColourAnalysisXML {
    protected $filepath;

    public function __construct($new_filepath){
        $this->filepath = $new_filepath;
    }

    public function csraw() {
        $dir = new FileSystemIterator($this->filepath);
        echo "Running raw data analysis.";
        foreach ($dir as $fileinfo) {
            $cmd = "bin\\colorsummarizer -image " . $this->filepath . $fileinfo->getBasename() . " -xml -pixel > " . "C:\\\\filepath" . $fileinfo->getBasename(".tif") . ".xml" . "\n";
            //run the above command line tool
            exec($cmd);
        }
        echo "Finished raw analysis!";
    }

    public function csstats() {
        $dir = new FileSystemIterator($this->filepath);
        echo "Running stat data analysis.";
        foreach ($dir as $fileinfo) {
            $cmd = "bin\\colorsummarizer -image " . $this->filepath . $fileinfo->getBasename() . " -xml -stats > " . "C:\\\\filepath" . $fileinfo->getBasename(".tif") . ".xml" . "\n";
            //run the above command line tool
            exec($cmd);
        }
        echo "Finished stat analysis!";
    }

    public function cshistogram() {
        $dir = new FileSystemIterator($this->filepath);
        echo "Running histogram analysis.";
        foreach ($dir as $fileinfo) {
            $cmd = "bin\\colorsummarizer -image " . $this->filepath . $fileinfo->getBasename() . " -xml -histogram > " . "C:\\\\filepath" . $fileinfo->getBasename(".tif") . ".xml" . "\n";
            //run the above command line tool
            exec($cmd);
        }
        echo "Finished the histograms!";
    }

    public function cscluster() {
        $dir = new FileSystemIterator($this->filepath);
    }

```

```

        echo "Running cluster analysis.";
        foreach ($dir as $fileinfo) {
            $cmd = "bin\\colorsummarizer -image " . $this->filepath . $fileinfo->getBasename() . " -xml -clusters 5 > " . "C:\\\\filepath" . $fileinfo->getBasename(".tif") . ".xml" . "\n";
            //run the above command line tool
            exec($cmd);
        }
        echo "Finished the clusters!";
    }

    public function csuniformity() {
        $dir = new FileSystemIterator($this->filepath);
        echo "Running uniformity analysis.";
        foreach ($dir as $fileinfo) {
            $cmd = "bin\\colorsummarizer -image " . $this->filepath . $fileinfo->getBasename() . " -xml -uniformity RADIUS > " . "C:\\\\filepath" . $fileinfo->getBasename(".tif") . ".xml" . "\n";
            //run the above command line tool
            exec($cmd);
        }
        echo "Finished uniformity!";
    }
}
?>

```

Appendix B: K-means Clustering Scripts

Hex code extractor

```
<?php
//Uses the XML outputs from Color Summarizer to extract the hex
codes and put them into a CSV
function getfile() {

    $fp = fopen('testxml.csv', 'w');
    $dir = 'C:\\\\filepath';
    foreach (glob("$dir\\*") as $file) {
        $content = file_get_contents("$file");
        preg_match_all("/#(.+)(?=\")/", $content, $matches1);
        print_r($matches1[0]);
        foreach ($matches1 as $fields) {
            fputcsv($fp, $fields);
        }
    }
    fclose($fp);
}

getfile();
?>
```

Silhouette Method for Determining number of K-Clusters in R

```
library(xml2)
library(cluster)

pg <- read_xml("C:\\\\filepath")

labs <- xml_find_all(pg, "//lab")

vals <- trimws(xml_text(labs))

vals <- strsplit(vals, ",")

df <- data.frame(matrix(unlist(vals), nrow=15700, byrow=T))

#for some reason I get a warning:
#"Warning message: did not converge in 10 iterations"
#when inputting 10 iterations (even 30!) however, I don't get
the warning with 12?

k.max <- 12
```

```
data <- df
sil <- rep(0, k.max)

for(i in 3:k.max){
  km.res <- kmeans(data, centers = i, nstart = 25)
  ss <- silhouette(km.res$cluster, dist(data))
  sil[i] <- mean(ss[, 3])
}

plot(1:k.max, sil, type = "b", pch = 19,
     frame = FALSE, xlab = "Number of clusters k")
abline(v = which.max(sil), lty = 2)
```


Appendix C: Visualization Scripts

Scatterplot of brightness median over time in R

```
#making a scatterplot of brightness median
library(ggplot2)
genredata = read.csv("C:\\filepath")
require(ggplot2)
attach(genredata)
qplot(Gcode, Brightness_Median)

attach(genredata)
plot(Year, Brightness_Median, main="Brightness Median Over
Time",
      xlab="Genre", ylab="Brightness Median", pch=19, col="blue")
```

Scatterplot of brightness median across genres in R

```
library(ggplot2)
genredata = read.csv("C:\\filepath")
require(ggplot2)
attach(genredata)
ggplot(genredata, aes(Genre, Value, colour = Genre)) +
  geom_point(size = 3) +
  theme_grey(base_size = 25) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

Comparison of hue, saturation and value within genre in R

```
library(ggplot2)
genredata = read.csv("C:\\filepath")
require(ggplot2)
attach(genredata)
ggplot(subset(genredata, Genre %in% c("Genre category")) +
      geom_point(aes(Hue, Value, group=Genre,
colour=Saturation), size = 3)
```

Palette creator in R

```
/*****
Quickpalette: https://github.com/EmilHvitfeldt/quickpalette
Chroma: https://github.com/jiho/chroma
```

```

Determining cluster number:
http://www.sthda.com/english/wiki/print.php?id=239#average-
silhouette-method
*****/

library(quickpalette)
library(chroma)

#dump the hex codes taken from ColorSummarizer's k-clustering
XML #from a cover into an object
#the number of the object revers to the Harlequin number of the
#cover
cover001 <- "#2E4E54 #B7BAA4 #AB5238 #CD9F5D #EBC419"

#regex_palette function from quickpalette library formats the
#above hex codes properly
regex001 <- regex_palette(cover001)

#repeat for every cover
#cover002 <- "#2E2D34 #615D4D #85413B #AF9B6F #E92C3D"
regex002 <- regex_palette(cover002)

#show_col from the chroma library then arranges the above hex
#code objects into palettes with each row representing one cover
show_col(regex1, regex2)

```

Appendix D: Color Summarizer Statistics

“The Manatee” original image vs. resized image aggregate statistics from Color Summarizer

| Original | Resized |
|--|--|
| <pre> <stats> <hsv> <avg>82.000000</avg> <median>62.000000</median> <min>0.000000</min> <max>360.000000</max> <avg>50.000000</avg> <median>56.000000</median> <min>0.000000</min> <max>100.000000</max> <avg>59.000000</avg> <median>60.000000</median> <min>10.000000</min> <max>100.000000</max> </hsv> <lab> <avg>3.000000</avg> <median>-4.000000</median> <min>-24.000000</min> <max>69.000000</max> <avg>15.000000</avg> <median>12.000000</median> <min>-32.000000</min> <max>90.000000</max> <avg>51.000000</avg> <median>50.000000</median> <min>6.000000</min> <max>93.000000</max> </lab> <lch> <avg>28.000000</avg> <median>19.000000</median> <min>0.000000</min> <max>90.000000</max> <avg>118.000000</avg> <median>110.000000</median> <min>2.000000</min> </pre> | <pre> <stats> <hsv> <avg>81.000000</avg> <median>66.000000</median> <min>0.000000</min> <max>360.000000</max> <avg>51.000000</avg> <median>56.000000</median> <min>1.000000</min> <max>100.000000</max> <avg>59.000000</avg> <median>60.000000</median> <min>10.000000</min> <max>100.000000</max> </hsv> <lab> <avg>3.000000</avg> <median>-4.000000</median> <min>-30.000000</min> <max>77.000000</max> <avg>15.000000</avg> <median>12.000000</median> <min>-42.000000</min> <max>91.000000</max> <avg>51.000000</avg> <median>49.000000</median> <min>6.000000</min> <max>96.000000</max> </lab> <lch> <avg>28.000000</avg> <median>20.000000</median> <min>1.000000</min> <max>92.000000</max> <avg>118.000000</avg> <median>111.000000</median> <min>1.000000</min> </pre> |

| | |
|---|---|
| <pre> <max>358.000000</max> <avg>51.000000</avg> <median>50.000000</median> <min>6.000000</min> <max>93.000000</max> </lch> <rgb> <avg>96.000000</avg> <median>83.000000</median> <min>0.000000</min> <max>212.000000</max> <avg>120.000000</avg> <median>107.000000</median> <min>18.000000</min> <max>236.000000</max> <avg>130.000000</avg> <median>145.000000</median> <min>0.000000</min> <max>255.000000</max> </rgb> </stats> </pre> | <pre> <max>356.000000</max> <avg>51.000000</avg> <median>49.000000</median> <min>6.000000</min> <max>96.000000</max> </lch> <rgb> <avg>96.000000</avg> <median>84.000000</median> <min>0.000000</min> <max>232.000000</max> <avg>120.000000</avg> <median>105.000000</median> <min>12.000000</min> <max>246.000000</max> <avg>131.000000</avg> <median>145.000000</median> <min>0.000000</min> <max>255.000000</max> </rgb> </stats> </pre> |
|---|---|

Table 11 The Manatee statistics

Appendix E: Full size median HSV scatterplots

Canadian fiction

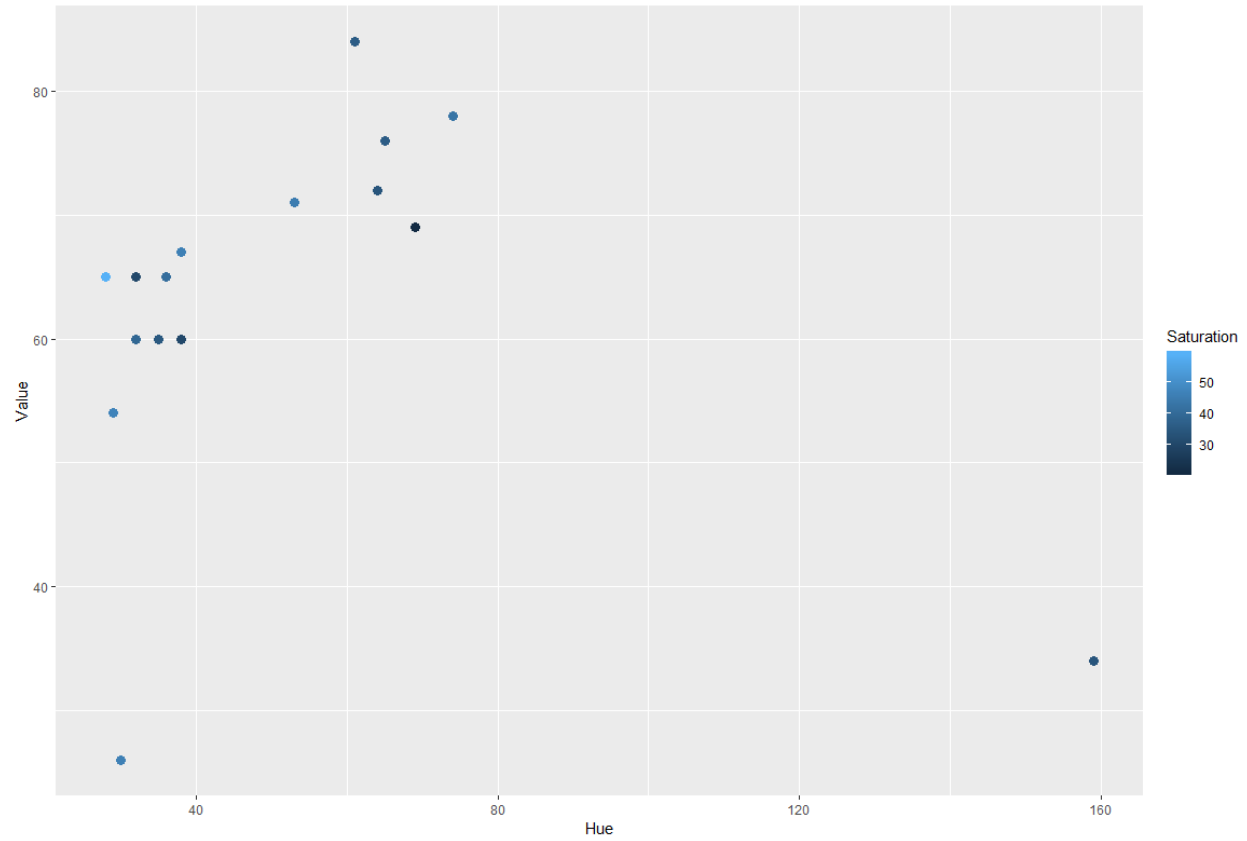


Figure 30 Canadian fiction HSV scatterplot

Detective/mystery fiction

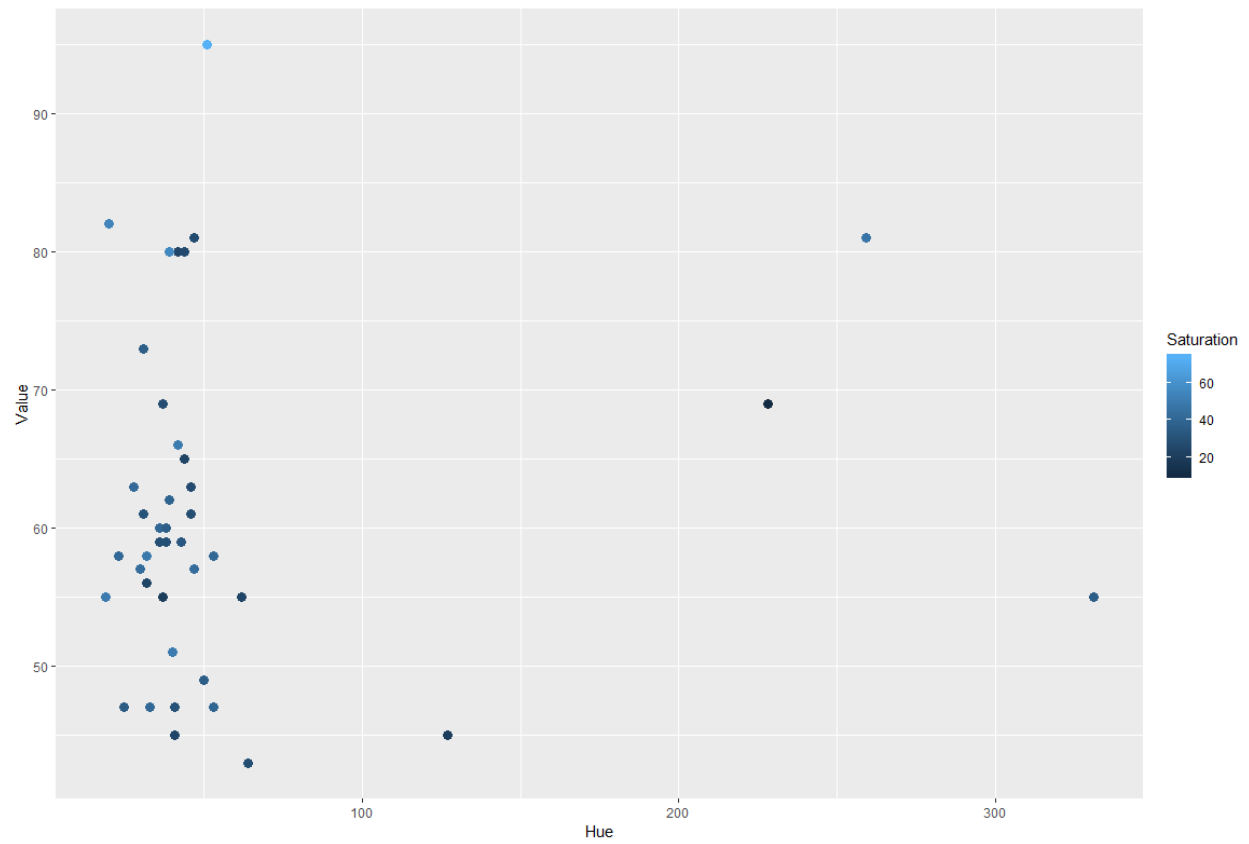


Figure 31 Detective/mystery fiction HSV scatterplot

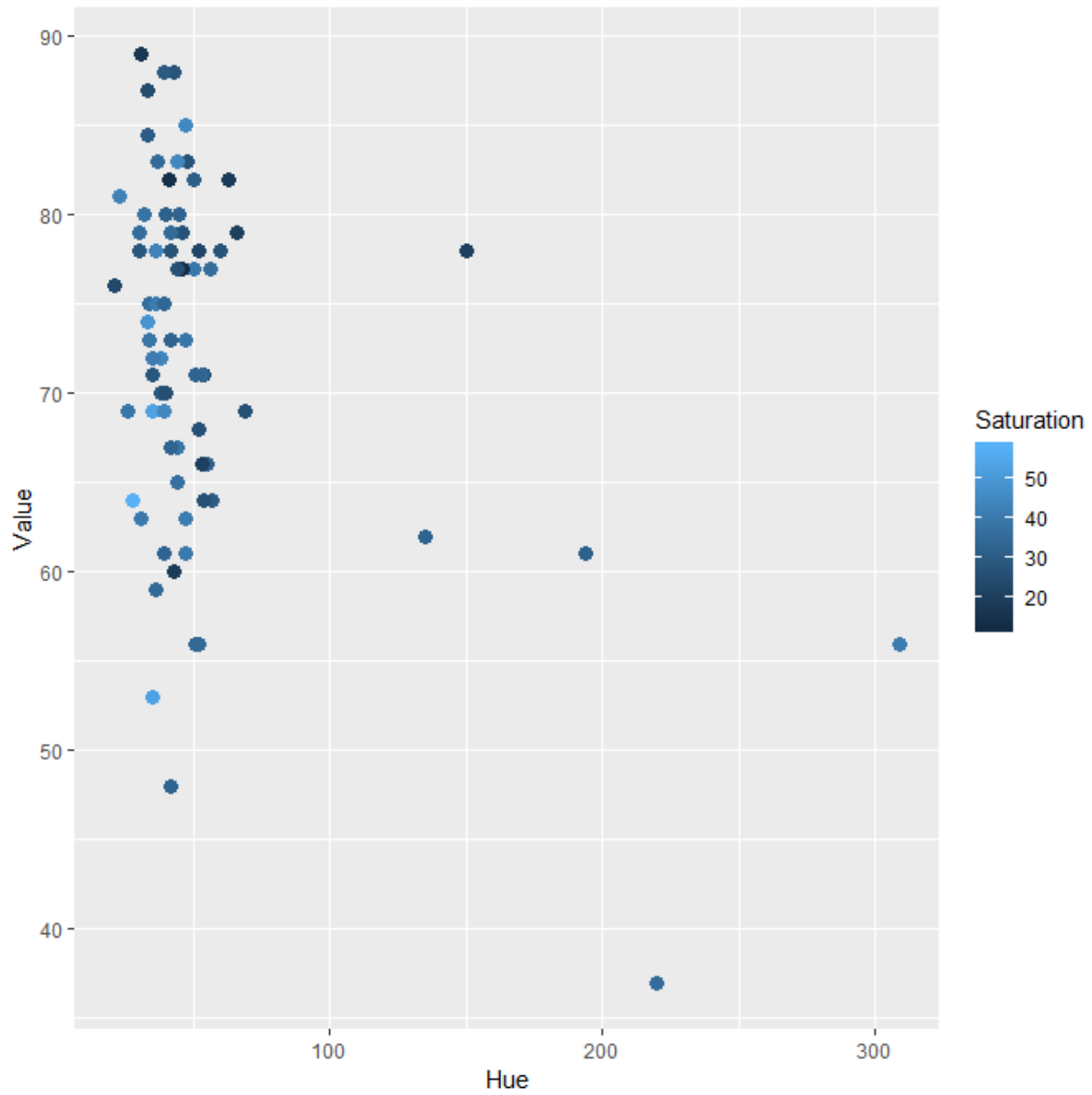
Doctor/nurse romance

Figure 32 Doctor/nurse romance HSV scatterplot

General fiction

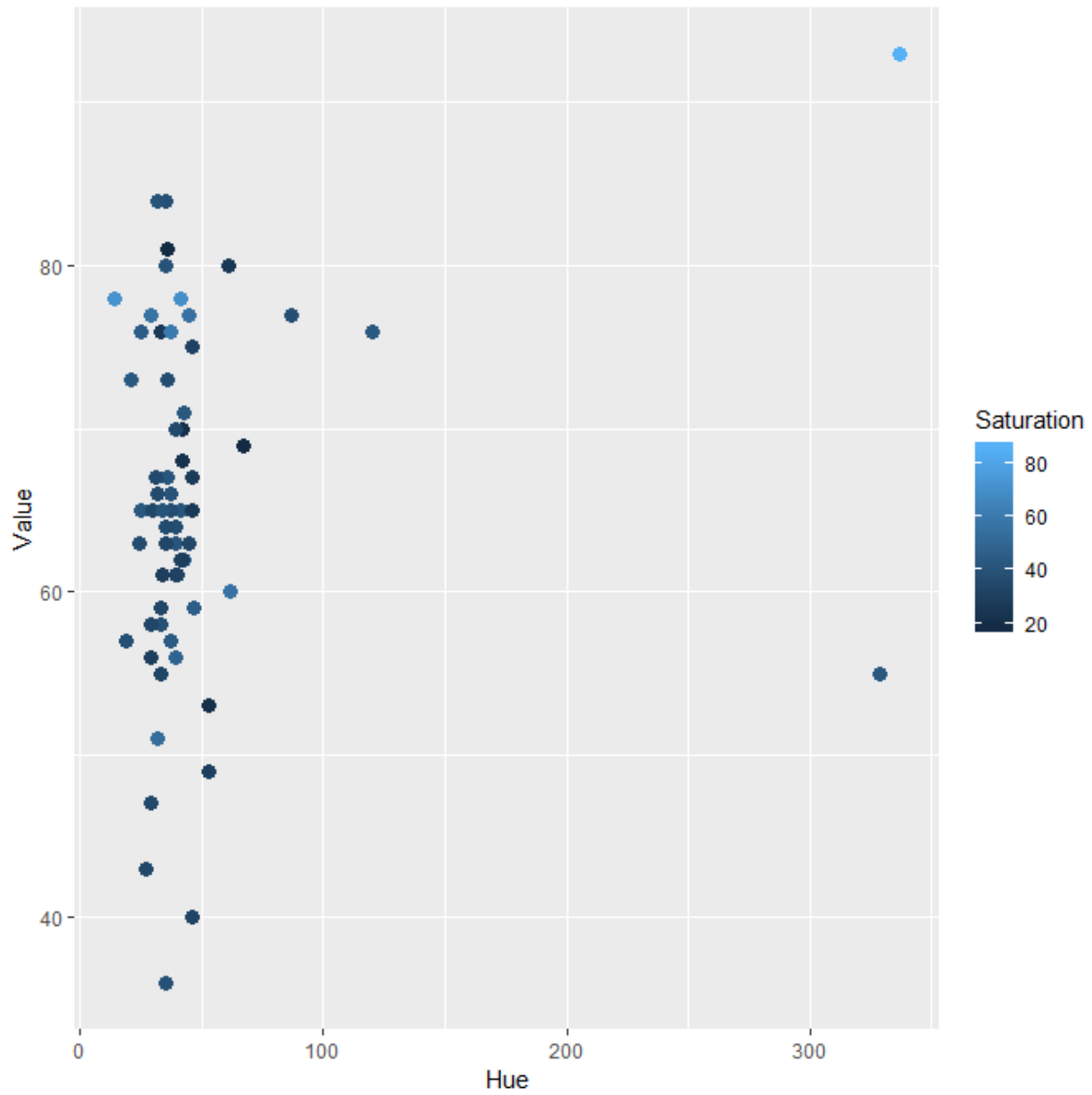


Figure 33 General fiction HSV scatterplot

General romance

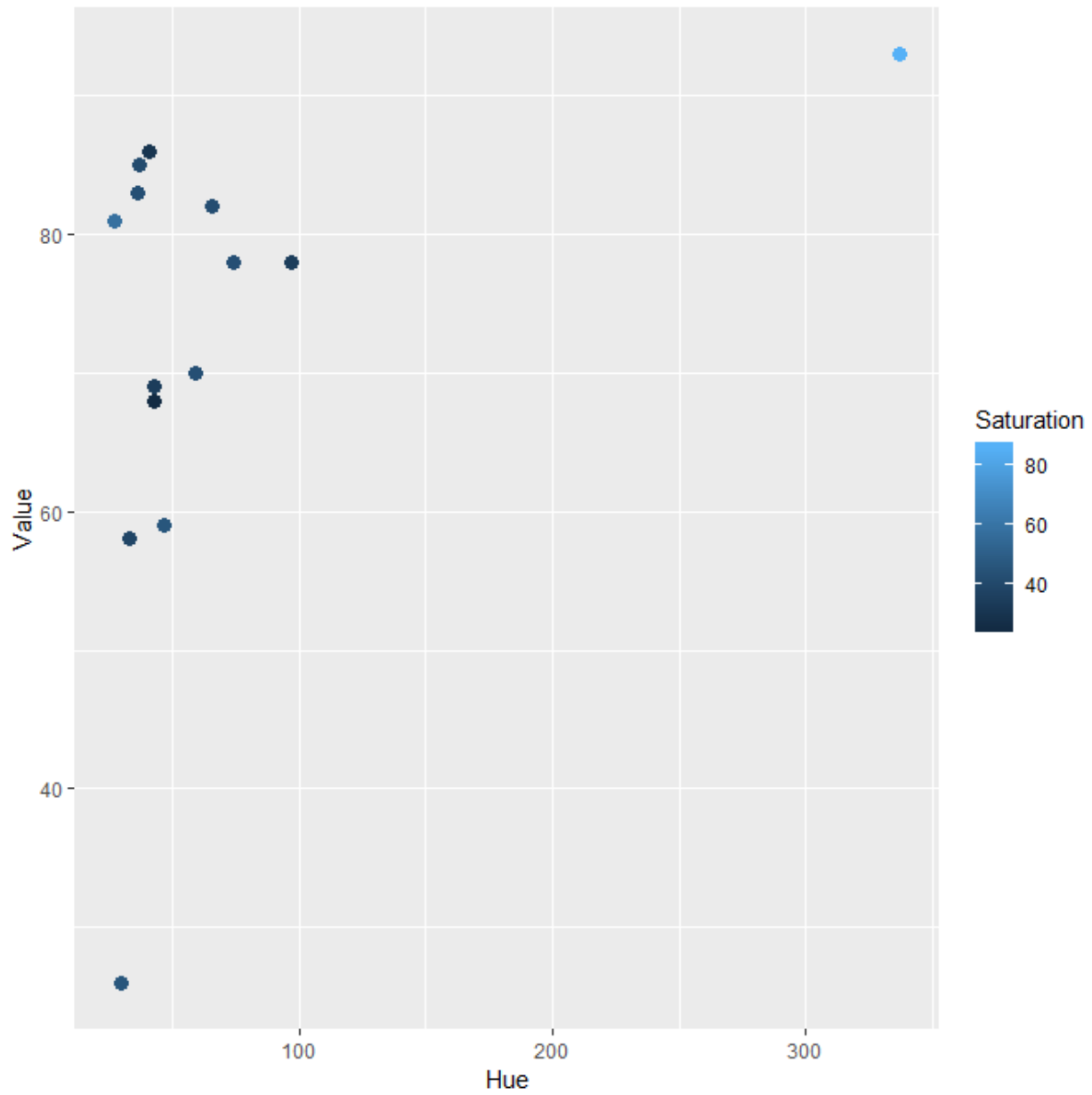


Figure 34 General romance HSV scatterplot

Historical fiction

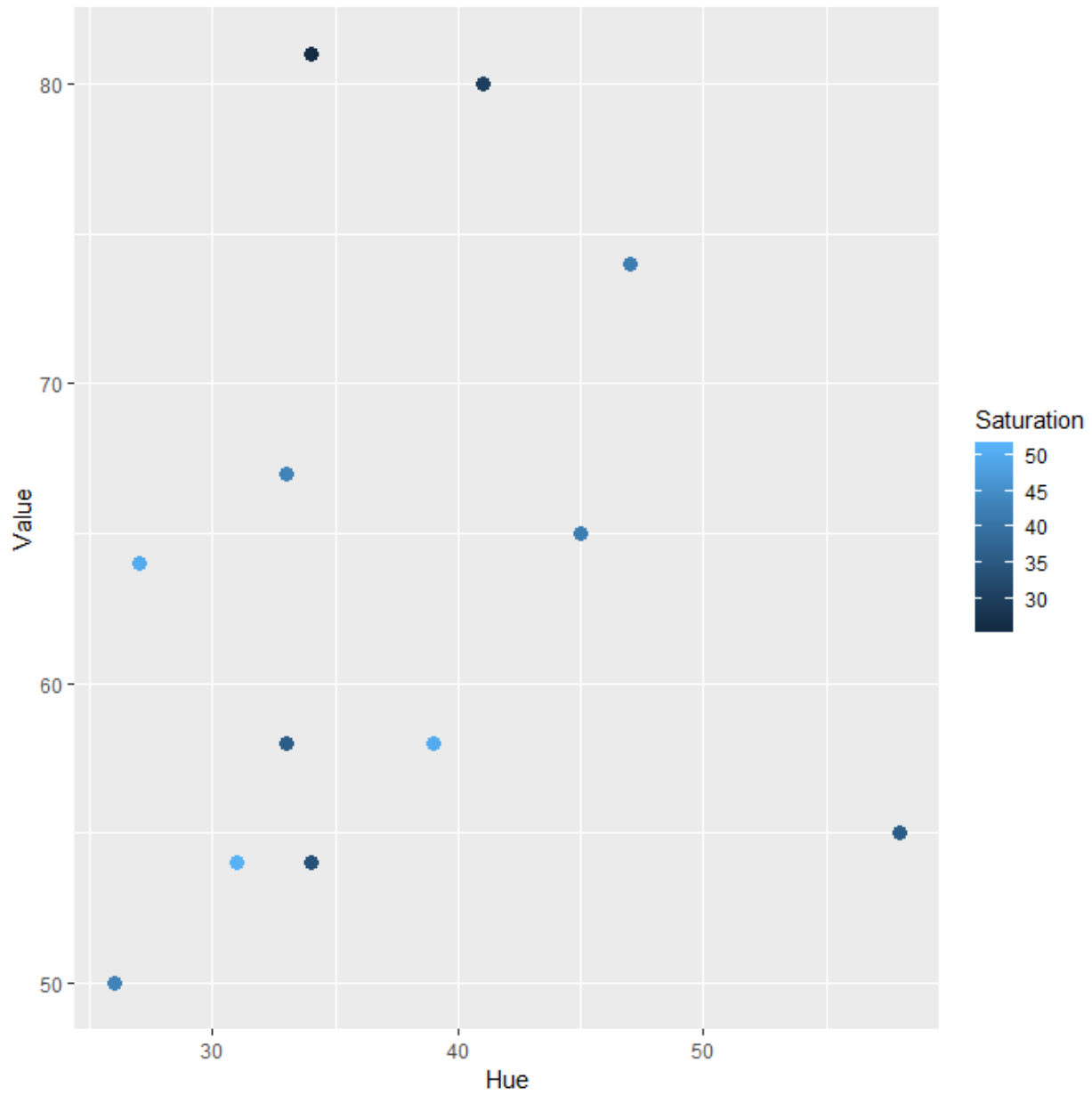


Figure 35 Historical fiction HSV scatterplot

International location

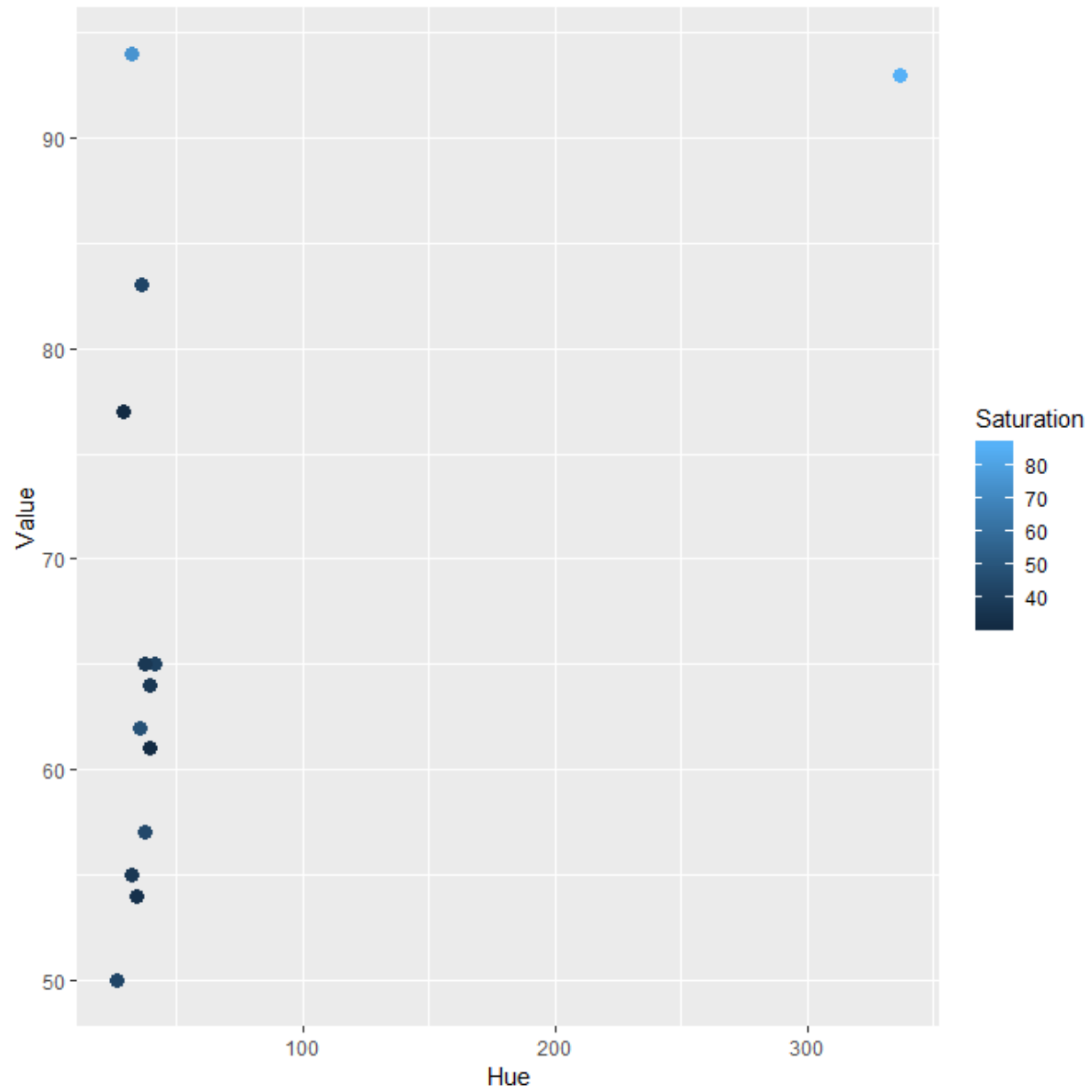


Figure 36 International location HSV scatterplot

Nonfiction

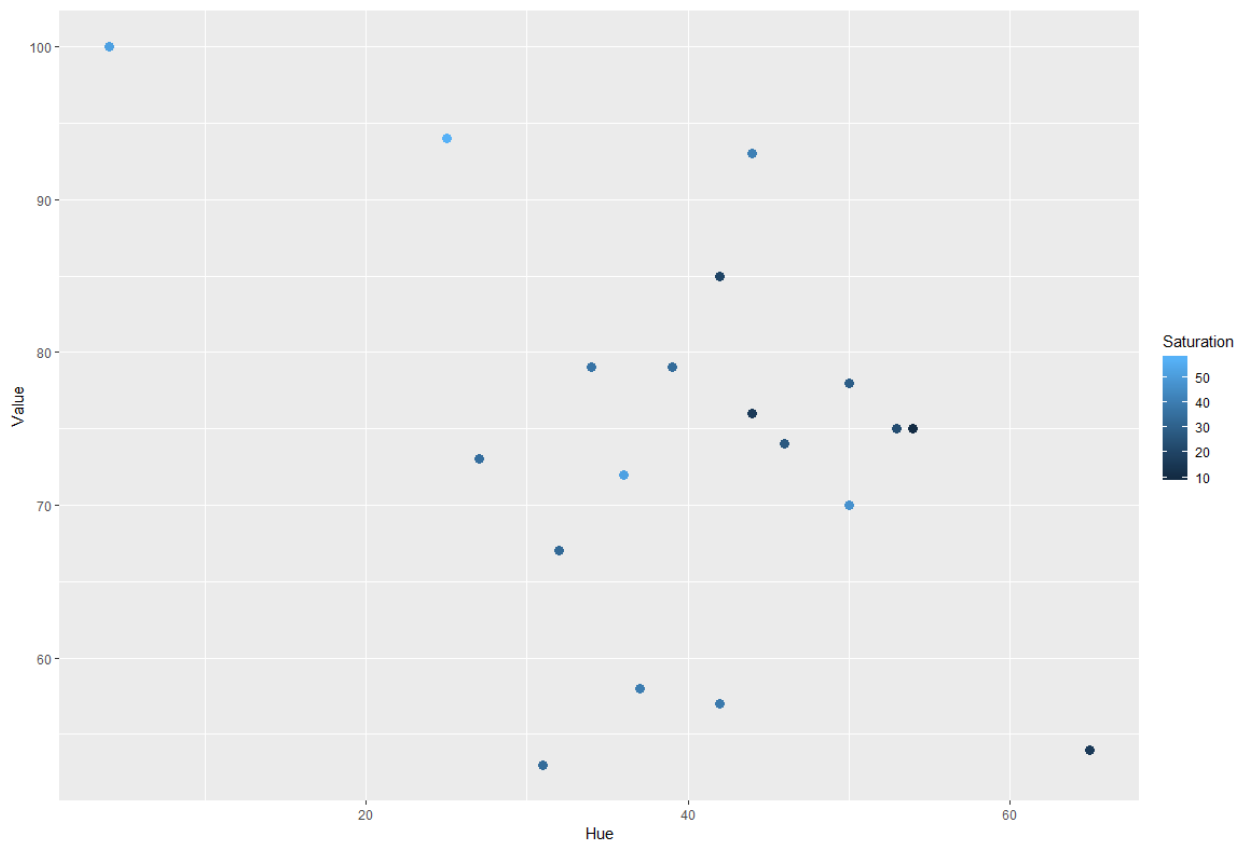


Figure 37 Nonfiction HSV scatterplot

Western/ranch fiction

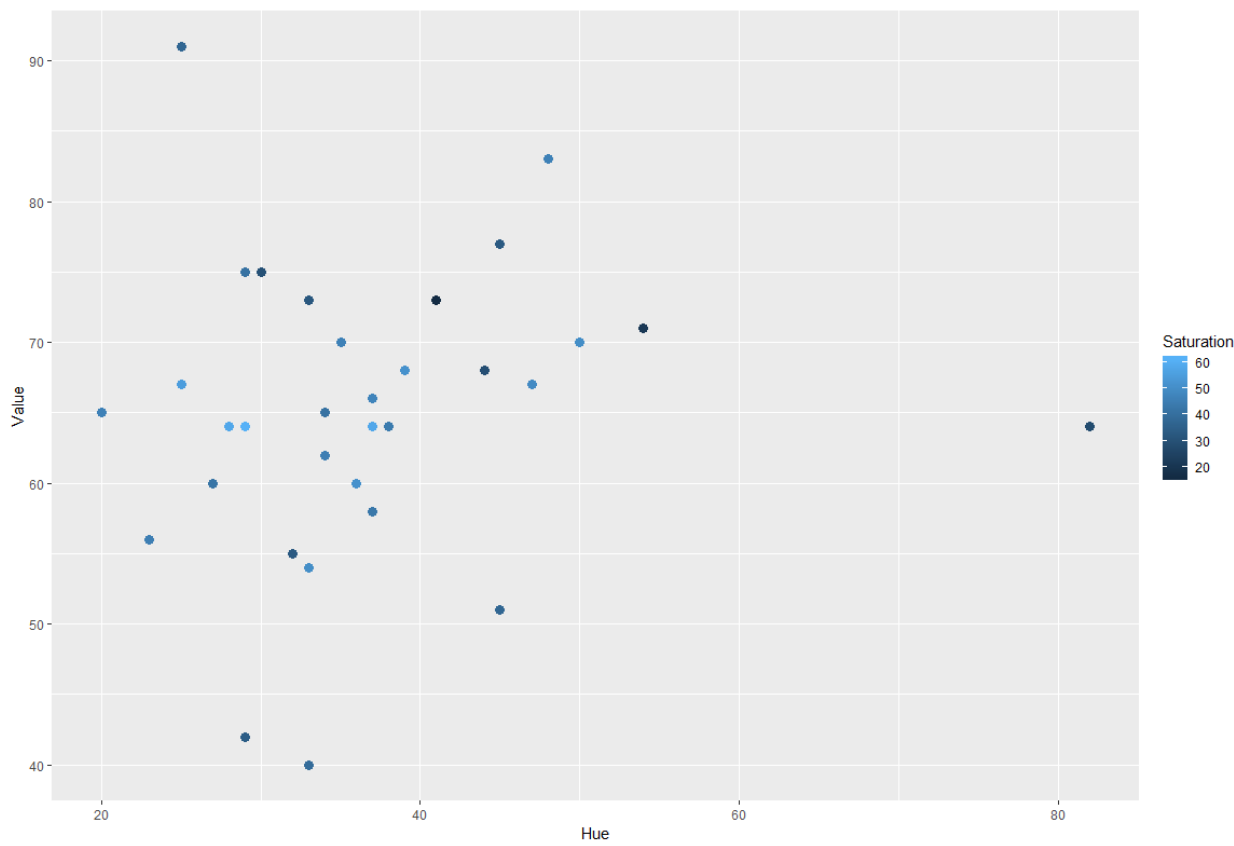


Figure 38 Western/ranch fiction HSV scatterplot

Appendix F: Colour Categorizations

Steinvall's (2007) colour categories (p. 357)

| Colour name | Emotion(s) |
|-------------|----------------|
| Black | Sadness |
| White | Fear |
| Red | Anger |
| Yellow | Joy |
| Green | Anger |
| Blue | Joy, sadness |
| Brown | Anger, sadness |
| Grey | Sadness |
| Pink | Love, joy |
| Orange | Joy |
| Purple | Anger |

Table 12 Steinvall's colour categories

Krygier and Wood's (2016) colour categories (p. 266)

| Colour | Concept |
|------------|--|
| Blue | Water, cool, positive numbers, serenity, purity, depth |
| Green | Vegetation, lowlands, forests, youth, spring, nature, peace |
| Red | Warm, important, negative numbers, action, anger, danger, power, warning |
| Yellow/tan | Dry, lack of vegetation, intermediate elevation, heat |
| Orange | Harvest, fall, abundance, fire, attention, action, warning |

| | |
|--------|--|
| Brown | Landforms (mountains, hills), contours, earthy, dirty, warm |
| Purple | Dignity, royalty, sorrow, despair, richness, elegant |
| White | Purity, clean, faith, illness, life, clarity, absence, light |
| Black | Mystery, strength, heaviness, death, nighttime, presence |
| Grey | Quiet, reserved, sophisticated, controlled, light, bland, dull |

Table 13 Krygier and Wood's colour categories

Aslam's (2006) colour categories (p. 19)

| Colour | Concept |
|---------------|------------------------------------|
| White | Purity, happiness |
| Blue | High quality, corporate, masculine |
| Green | Envy, good taste |
| Yellow | Happy, jealousy |
| Red | Masculine, love, lust, fear, anger |
| Purple | Authority, power |
| Black | Expensive, fear, grief |

Table 14 Aslam's colour categories

Allan's (2008) colour categories (p. 636-637)

| Colour | Concept |
|---------------|---|
| Black | “Used orthophemistically but not euphemistically; it has dysphemistic connotations more often than other colours do. It is often connected to darkness (the night), death, decay, and evil deeds. Black has often been used dysphemistically of human skin colour” (636). |
| White | “Contrast to black and, as such, linked to light and purity; it mostly has positive connotations, though it is rarely used euphemistically. Dysphemistic uses depict cowardice and |

| | |
|--------|--|
| | fear.” |
| Grey | “Used for indeterminability and dullness.” |
| Brown | “Faecal associations of brown lead to several dysphemisms; brown is found in no euphemisms and few orthophemisms in figurative speech.” |
| Yellow | “Dysphemistically used of cowards and cheap paper, and sometimes of East Asiatic people; but it is orthophemistic and positively used of light-coloured African Americans.” |
| Red | “Both positive and negative figurative expressions, links it with blood—life-blood, the blood of the slain, or menstrual blood.” |
| Green | “Living vegetation; negative connotations arise when it is the colour of illness or jealousy (perhaps seen as illness).” |
| Blue | “Connected with the virtuousness and chastity of a bride. The negative aspects of figurative uses of blue arise from fear, fighting, despondency, and tabooed language and behaviour. It is arguable that the use of blue to speak about these topics is euphemistic and that uses of blue are rarely dysphemistic.” |

Table 15 Allan's colour categories