Brian Choi

Visualizing Reader Preferences on Goodreads

Abstract:

The goal of this project is apply concepts learned in DSE241 to build beautiful and insightful visualizations of a public data set. The visualizations for this project were built using data from Goodreads, a social cataloging website for books, and the Bokeh library, with the majority of the coding done on Jupyter notebooks and then exported in to HTML. Goodreads has a detailed data set including a nearly complete catalogue of every book ever printed and whether or not users enjoy those books. Through visualization, this projects aims to show trends in reader behavior and preferences in ways that are intuitive, surprising, and interesting.

Although additional data and statistical analysis was necessary both as preprocessing and exploratory stages, the visualizations themselves are intended to provide all the insight the project has to offer. In particular, the project will cover what types of books users like to read, what period of literature is the most popular amongst contemporary readers, which authors tend to be the most popular (including which period they tend to be from), and whether or not author popularity tends to be dominated by "one-hit wonders".

Some of the immediate findings from the analysis and visualizations show that contemporary books are far more popular than classics, with the vast majority of books being read and reviewed today being published after 2000. Similarly, contemporary authors are far more popular than older authors, with authors of young adult novels such as Suzanne Collins and J.K. Rowling having a dominant hold on readership amongst Goodread users. Furthermore,

Introduction:

Goodreads is a social cataloging website where users can document, tag, and rate books they have read. By linking via Facebook, friends on the network can view each others' books to read and read books, providing a way for users to share their favorite books with others. With a simple rating system (out of 5), a user can also quickly look up a book and its full-text reviews to see whether or not a book is worth reading or purchasing in addition to detailed information including number of pages, publisher, ISBN, language, and number of editions. Because Goodreads has explicit links to Amazon, it provides an easy way for readers to purchase a book if reviews have piqued that users' interests.

The main motivation for this project is to analyze user preferences between "classics" and contemporary novels. Although series like Harry Potter and Hunger Games have motivated millions of young readers to pick up books, classics still hold a very important part in literature history. Whereas popular contemporary books tend to either young adult, erotic fiction, or thrillers, older popular novels tend to be "serious" literature that is often read in school (Animal Farm, To Kill a Mockingbird, etc.). The true fundamental differences between "serious" literature and contemporary popular literature is of course up for debate but will not be a major part of this project. Genres are inherently tricky and the data set does not provide enough depth to make a true in-depth project of genres, although it will be touched on a little bit in the form of tags. This project will instead focus on the relationship between date of publication and popularity, largely ignoring some of the more intricate and subjective factors in what defines a "classic"

as opposed to contemporary popular fiction. By only looking at a select few features, the project will hope to maintain a tighter focus on its stated goals and avoid spreading the overall purpose too thin.

Dataset:

The dataset was acquired from https://github.com/zygmuntz/goodbooks-10k. Github user Zygmuntz scrapped the data as part of a recommender system project he/she was working on and graciously made the data set public. The data set consists of the top 10,000 most popular books on Goodreads as of 2017 as well as six million ratings sampled from these books. The data set comes in the form of five csv files: book_tags.csv, books.csv, ratings.csv, tags.csv, to_read.csv that can be joined together by user_id and book_id primary and foreign keys. For the purposes of this project, only books was used. While the original intention was to use tags in one of the visualizations, in particular to plot the relationship between time periods and genre prevalence, tags were too messy of a table and required too much clean up. With more time, the data would have been very interesting to explore but with the constraints of the project, the scope was intentionally limited to focus on a select few topics.

As mentioned above, the books table is the only table used in the visualizations for this project. Five out of the six visualizations built for this project use only the books table as the attributes it contains are enough to draw very insightful conclusions from. The books table contains the following fields to be used in analysis and visualization: authors, title, original_publication_year, ratings_count, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5. Ratings_1 to ratings_5 denote the number of 1 to 5 star ratings given by users to the book, given as a total count. Other features such as ISBN, book_id, image_url, editions were not included as part of this project. Note that rating count will stand in for read count. Because the read count is not measured in the books table, rating counts will be used instead to measure how popularity a book is. The assumption is made that rating count is proportional to read count.

Several additional features were created from the nine features mentioned above for the analysis: average_rating, main_author, and nth_release. Average_rating is calculated for each book by taking the weighted average of ratings_1 to ratings_5. Main_author is necessary to extract from the authors field because the authors field often contains the main author plus the book's illustrator and publisher who are irrelevant in this study. Main_author is extracted by converting author in to a list and taking the first element as the primary author of a book, much like in a research publication, will be listed first. Nth_release denotes the chronological order in which a book has been published. Because the dataset only contains the 10,000 most popular books, not every book written by a given author will necessarily be listed. Nevertheless, nth_release provides a useful overview of an author's career and how popular they are in different stages of their career. It is calculated by sorting books by their publication date, grouping by author, and using a ranking function. Other group features are created for specific visualizations that will be discussed in the later sections.

All data preprocessing was done using the Pandas and Numpy libraries which seamlessly integrate with Bokeh. Dataframes and Numpy arrays are easy to manipulate and allow the user to massage datasets for manipulation very quickly. Exploratory data analysis was also necessary to find key trends in the data that could be effectively visualized. As a result, much iterative work was done using Pandas and Numpy for discovery and preparation and the large amount of the code base will comprise of dataframe manipulation and analysis.

Tasks:

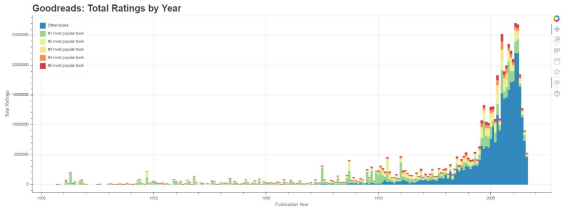
The project can be evenly divided into two main areas of analysis: individual books and authors. The data set is very rich and has opportunities for exploration in many different directions but the focus here will largely be based on the temporal effect on book/author popularity. The questions that this project seek to answer are:

- Among today's readers, how popular are classics compared to contemporary books?
- What are the most popular books of all time and is there a relationship between popularity and average rating on Goodreads?
- Are certain periods of literature better regarded than others? How do 19th century books stack up against contemporary 21st century books and 20th century books?
- Who are the most popular authors? Are the most popular authors also the most prolific or highly rated?
- How do authors' careers evolve over time? Which authors tend to hit their "peak" at certain points and which authors tend to have steady popularity throughout their entire career?

This project will be mainly targeted at book publishers, advertisers, and users of Goodreads who are interested in finding out more about what types of books readers enjoy. Just as targeted advertisement via machine learning has become a huge part of retail industries over the last 5-10 years, analysis of reader behavior will be an important part of both increasing sales on the part of the book retailers and helping users pick books they are most likely to enjoy. Although machine learning is not a part of this project, statistical analysis and visualization of the features will hopefully shine a similar light on trends not seen before.

Solution:

Viz 1, How Popular are Classics Compared to Contemporary Books?:

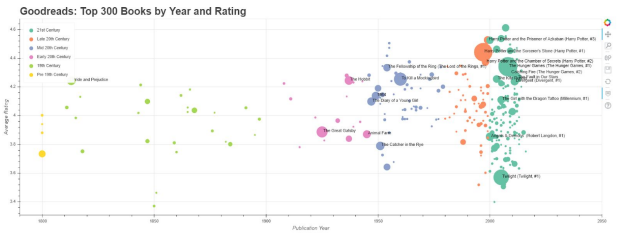


The first visualization is a stacked bar plot representing the number of ratings given for books published in a given year. A bar plot was used here to visualize two quantitative attributes, one discrete (year) and another almost continuous (number of ratings). Because bar plots work well with categorical or at least discrete attributes, they are a natural match here. An early attempt was made to use line or area plots but because of the wild changes in total ratings per year and the need to also plot the top five most popular books, bar plots seemed like a better fit. Line plots had the problem having far too extreme peaks and valleys and although it was able to convey the same information, the overall aesthetic of the plot was simply not very appealing.

The y-axis goes from 0 to 30,000,000 while the x-axis goes from year 1800 to 2020 (with latest published book in 2017). While the data set contains books from before 1800, the choice was made to exclude those books because of the effect they would have on x-axis. From a quick glance, it is clear that the vast majority of ratings are for books from year 2000 and beyond. Including very old books such by books by Plato (dating back to 400 B.C.) and Shakespeare (1600 A.D.) will squish the largest peaks on the right side of the bar chart. In order to preserve the temporal expressiveness of the visualization, filtering out books from before 1800 was necessary.

The stacked bars represent 6 different categories, the number of ratings for the #1-5 most popular books and the number of ratings for all other books summed up. The height of the bars represents the total number of ratings for all books that year. The #1-5 most popular books are colored on a green to red spectrum while the other books are colored solid blue, giving a very clear distinction between the the top 5 books and the "other books". Hovering over each bar gives a tooltip that displays the year, number of total ratings for the year, and top 5 most popular books.

Viz 2, What are the Most Popular Books of All Time and is There a Relationship Between Popularity and Average Rating on Goodreads?

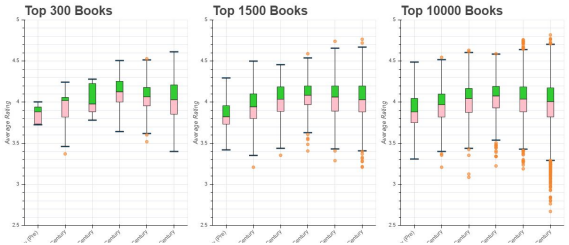


The second visualization is a scatter plot that similarly uses publication year as the x-axis. However, instead of using the y-axis as total count, average rating is used instead to visualize the relationship between publication year and average rating. Total count is encoded by the radius of the points with larger circles indicating more popular (read) books. Different scales were used for mapping count to radius but ultimately, a linear scale proved to be most effective again, showing the enormous difference in readership between books like Harry Potter and books like Anna Karenina. An additional color channel was used to indicate which period of literature (self-defined) each book belonged to. This is a redundant channel that amplifies the temporal element of the visualization. More will be discussed later on the choice of period bucketing. The goal of this visualization is to simply plot the top 300 most popular books in a way that users can clearly identify books that are popular and how they tend to be rated by other users, with the design choices mentioned above intending to demonstrate this as clearly as possible.

Again, the x-axis starts at 1800 but instead this time, books that were published before 1800 were simply changed to the year 1800. The same was not done in visualization 1 to avoid a large spike from the glut of pre-19th century books on the far left side which would have been both misleading and visually unappealing. In this case, there are only four books from pre 19th century as opposed to full aggregation, preventing the far left side from being too cluttered. The use of a color map here clearly distinguishes pre-19th century books from 19th century books despite appearing right on top of the year 1800 point.

Interactions here play an important role in making the most out of the visualization. Because radius is linearly scaled, books that were otherwise not widely read tend to be hidden behind books that are extremely widely read such as Angels & Demons and Twilight. The zoom interaction gives the ability to move closer to the x/y coordinate of choice to see books that would have otherwise been hidden. The tooltip interaction also gives the ability to lookup books by simply hovering over them. With over 300 points, not every point can be labeled (only the top 20 are in this case) so the tooltip in this case is extremely important for looking up and discovering books that are of interest.

Viz 3, Are Certain Periods of Literature Better Regarded than Others? How Do 19th Century Books Stack Up Against 21st Century Books and 20th Century Books?:

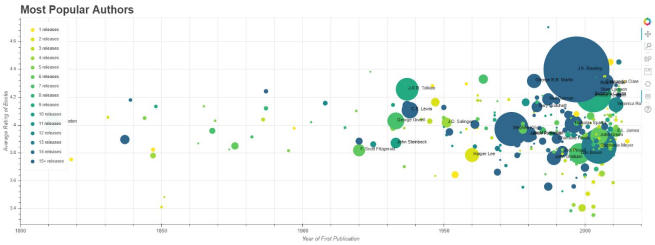


The third plot is a box and whisker plot, designed to show the statistical distribution of ratings by period.

The x-axis represents time periods, same as the ones from visualization 2, while the y-axis represents the non-weighted average rating of books in that time period. As with all box and whisker plots, the middle section represents the median. 25th quartile, and 75th quartile. The whiskers represent extremes, represented by 1.5 times the interquartile range on both sides. However, if no outliers exist, represented by the orange circles, then the whiskers are shrunk to the respective min/max of each period. Outliers are defined as points that exist outside of the whiskers. Although it would have been preferable to squeeze the y-axis a little tighter to further highlight the changes, a wider y-axis was necessary to show all the outliers in the top 10k. Box and whisker plots are an excellent way of viewing statistical distributions, in particular percentiles and outliers. Directly relating to the previous two visualizations, this visualization attempts to tie them all together and look at periods as a whole.

Three box and whisker plots are shown here to show the difference in distribution between the top 300 books, top 1500 books, and top 10,000 books. Because the distribution can be very different depending on how many top books were selected, showing all gives a clear picture of how the very most popular books stack up with less popular books. The same y-axis was used for every plot so that comparisons of the same period can be made across different levels of selection. For example, by looking at only late 20th century books across all three plots, the user the distribution of the most popular late 20th century books against all late 20th century books, including their medians, outliers, and 75th and 25th percentiles.

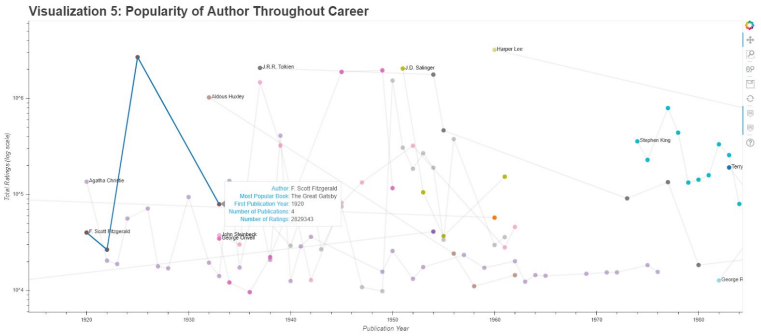
Viz 4: Who are the Most Popular Authors? Are the Most Popular Authors Also the Most Prolific or Highly Rated?



Visualization 4 follows many of the same principles as visualization 2, using a scatter plot along a publication year (denoting year of author's first publication) and average rating to obtain a sense of popularity for each author. Visualization 4 mainly differs from visualization 2 in that color is now used encode the number of books that an author has published that has appeared in the Goodreads top 10k. Whereas color was used in visualization 2 to used to provide redundancy, color is used here to encode another quantitative attribute, giving the plot a total of four quantitative attributes. Because the two channels are used here are two of the most visible (hue and size), the visualization hopefully does not contribute too much to cognitive load. In this case, the average rating of an author is not weighted by book popularity.

As in visualization 2, interaction, in particular zoom and hover, plays an important role in using the plot. Because certain authors (J.K. Rowling for example) dominate the plot, zooming allows a degree of exploration to view less popular authors who have written in the same time span and are similarly rated.

Viz 5: How do Authors' Careers Evolve Over Time? Which Authors Tend to Hit Their "Peak" at Certain points and Which Authors Tend to Have Steady Popularity Throughout Their Entire Career?



Visualization 5 is a line and scatter plot charting the progress of authors by their book releases. The points represent when a book is released by an author and the line follows the points to create a career progression for the author. The y-axis is defined as the publication year and the x-axis the total ratings by log scale. Only the top 5-7 authors per time period are used in this visualization to avoid excessive clutter. Because each author can author 10 or more books in their career, having too many authors can easily cause cognitive overload as the plot is filled with circles and overlapping lines. Each author is additionally defined by the color of the points and a label on the plot at the point when his/her first book was released. This gives the user an ability to scan through the visualization and then highlight the author of their interest.

Interactions are fundamentally important with this visualization as hovering over a line will highlight it, showing the progression of an author's career. All other lines are by default very lightly colored and only by hovering over a line will the user be able to clearly see how popular each of their books are. Thi allows a user to look one of his/her favorite authors, highlight the line, and see if their career followed a relatively consistent path or they were simply a one hit wonder.

Implementation and Usage:

All work was done using Python with the Pandas, Numpy, and Bokeh libraries. Bokeh was used to generate all plots and all expressive features and limitations that go along with the library. While interactivity could have been further enhanced with custom Javascript, most of the interactions were kept within the limitations of built-in Bokeh functions. The visualizations were exported to HTML to be displayed as a top-down website.

Results (refer to snapshots from Solution section):

Visualization 1:

In terms of expressiveness and effectiveness, the visualization accomplishes its goals by answering the very specific question of classics vs contemporary popularity. In addition to giving quantitative information about the totals throughout the x-axis, a rough estimate of the total area occupied from time ranges shows how each era compares to one another. The use of a discrete color map to differentiate other books with top 5 popular books also gives the user a quick and obvious way to tell the difference between how much the top 5 books dominate the total standings for the year. As mentioned in the solutions section, a decision was made to start the plot from 1800.

The visualization shows a clear trend that newer books, particularly ones published from 1995 to 2017 are significantly more popular than books published before then. Because the y-axis is fully linear, the graph shows the enormity of the difference between contemporary books and classics in terms of readership. Furthermore, by looking at the amount of area the top 5 books occupy as opposed to the "other books", noted in blue, it is clear that classics are largely dominated by 1 or 2 books a year as opposed. For example, To Kill a Mockingbird alone makes up nearly 75% of the total popularity in 1960.

Comparing to 2008, The Hunger Games, one of the most popularity books on all of Goodreads, merely makes up about 7% of the plot. It can therefore be seen that while the most well known classics are still widely read, many of the older less popular novels that were published in the same era from 1800-1980 have largely been forgotten. On the other hand, contemporary books are filled with all kinds of books ranging from highly popular Young Adult fiction to biographies like Steve Jobs. The variety of books being published and read today are highly diverse.

Visualization 2 and 4:

In terms of expressiveness and effectiveness, visualizations 2 and 4 follow many of the same pros and cons. While a scatter plot is excellent at encoding data between 2 or more quantitative attributes plus a categorical attribute, it does sometimes take a while to query the points of choice. In viz 2's case, color was used to map each point to a period. By doing so, the visualization adds extra cues for the user to see how books from one time period stack up with another. Furthermore, the addition of a size channel gives the user the immediate ability to query some of the most popular books of all time such as Catcher in the Rye and Harry Potter. Color is used a little differently in viz 4 as it encodes a third quantitative attribute. In this case, it is interesting to see the relationship between overall popularity and the number of releases an author has had. While it adds cognitive load to process a mix of colors along with size, the combination of the two creates a visualization that covers multiple bases at once. In both cases, the top 300 books/authors are used to reduce clutter, a very important step in ensuring an effective scatter plot. The time periods chosen here were subjectively chosen. Early 20th Century represents the pre-war period of 1900-1945, mid 20th century covers 1945-1980, and late 20th century covers 1980-2000. There is no set definition on what constitutes early, mid, and late 20th century literature so a decision was made to roughly follow historical events such as the end of World War 2 and the fall of Communism to define periods.

From visualization 2, it can be seen that while popular books are in general fairly well rated, mainly referring to the Harry Potter series, there is no directly correlation between book rating and book popularity nor is there a correlation between period and rating given. While certain books like Catcher in the Rye and To Kill a Mockingbird remain perennial favorites, the number of books from 19th century and before are sparse, indicating a falling out of favor among modern readers. Viz 4 shows an obvious but still interesting finding that prolific authors are generally more read than less prolific authors but there exists many profilic authors such as Kathy Reichs who are not that well read compared to someone like Harper Lee who only released two books in their lifetime. While Viz 2 already shows the absolute dominance in readership of the Harry Potter series, Viz 4 further stresses upon that by showing how much bigger J.K. Rowling's point is compared to almost every other author.

Visualization 3:

In terms of expressiveness and effectiveness, this visualization accomplishes the task of a typical box and whisker plot, showing the distribution of ratings for different periods of literature. With only six periods compared in each mini-plot, the differences between each period is immediately clear. As mentioned in the solutions section, the splitting of the plot by top 300, top 1500, and top 10k gives very different pictures of how the distribution looks when only comparing popular books vs comparing all books. The downfall of the top 300 books is the small sample size for which the top 1500 books are meant to provide a middle buffer between it and the top 10k books. The inclusion of outliers in the visualizations, sometimes just a luxury but very useful here, give further insight on how many books are

tend to be categorized as “really great” or “really bad”. As talked about in visualization 2, the periods are indeed subjective which is a downside of this visualization.

The box and whisker plots show that older books, particularly from the 19th century and before and generally regarded worse than books from newer periods. While the sample size may be to low for the top 300 books to make a strong conclusion about the distribution, it can be seen that the majority of books from before the 19th century specifically score significantly lower than books from the 21st century. This visualization validates some of the trends that were seen in visualization 2 and 4 where the users can get a good sense of how books and authors of various periods fare purely on a visual level. Interestingly, in the top 10k books section, the 21st century section shows a huge extremity range (shown by whiskers) and also shows a huge number of outliers on both sides, in particular the lower outliers. This perhaps indicates that while 21st century books are generally well regarded, it is also filled with lower quality pulp fiction that while decently popular, are generally not well written books.

Visualization 5:

In terms of expressiveness and effectiveness, visualization 5 tries to find a fine balance between being informative and staying as expressive as possible. This mainly lies in the choice of authors to represent on the plot and how many is too many. Obviously not every author can be included in the visualization. After playing around with the numbers a bit, a decision was made to select the top 5-8 authors per period so the visualization maintains a balance across the entire timeline. Even with about 30 authors, there is still a significant amount of clutter which takes away from the overall effectiveness of the visualization. Additionally, because only 30 authors are given as opposed to 300 in viz 2 and 4, the amount of information gained from this visualization is not as strong. An alternative option that would have worked better is to use multiple line charts with each author occupying their own row in order to best visualize this data. While the temporal element would be lost, it would be much easier to query an author.

The visualization itself shows that for some of the oldest authors such as Dostoyevsky, their popularity is mainly covered by one or two books, namely Crime and Punishment and Brothers Karamazov.. On the other hand, contemporary prolific authors such as Terry Pratchett enjoy a consistent career with strong books every couple releases. This perhaps highlights the point made in visualization 1 that older authors/periods tend to be dominated by a few select books every year and that lesser known books have fallen to the wayside in popularity. Many of the greatest authors of the past were also mainly known for one or two books, F Scott Fitzgerald and Aldous Huxley being prime examples

Challenges and Lessons Learned:

While Bokeh is an intuitive and highly powerful library, it still has a significant learning curve. Combined with the preprocessing needed for the various visualizations, the two provided a significant work and time challenge. After carefully considering the options from the beginning, Tableau may have been a better choice to develop these visualizations. The final did provide a very strong learning opportunity for the beautiful Bokeh library which will undoubtedly prove useful in the future.

The huge amount of features in the data set provided a major initial challenge as it increased the amount of time needed to explore the data and find trends that were actually fitting to visualize. The initial plan was to include tags as well but after further exploring and realizing the enormity of the tags data set and how much cleaning was required, the decision was made to not include it. As learned in the capstone class and this class, EDA is a vital part of visualizations and data science in general and having a strong plan for EDA is paramount to producing a good product.