P5: Exploring Amazon's Best-Sellers List Data

Margaret Baxter, Sukriti Bhardwaj, Justin Hu, Philip Huynh, William Sheppard

December 6, 2020
CS 4460

Dataset:

Name Name of the Book	Author The author of the Book	User Rating Amazon User Rating	Reviews Number of written reviews on amazon	Price The price of the book at 13/10/2020)	Year The Year(s) it ranked on the bestseller	Genre Whether fiction or non-fiction
Publication Manual ... 2% StrengthsFinder 2.0 2% Other (531) 97%	Jeff Kinney 2% Suzanne Collins 2% Other (527) 96%	3.3 4.9	37 87.8k	0	2009 2019	Non Fiction 56% Fiction 44%
10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
5,000 Awesome Facts (About Everything!)	National Geographic Kids	4.8	7665	12	2019	Non Fiction

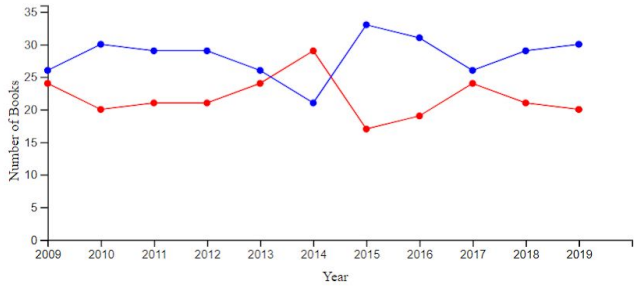
Visualization:

Number of Best-Selling Books on Amazon

Book Price Cutoff:

Filter Data

Reset Filter



Highest rated non-fiction books from 2017
4.9 - Obama: An Intimate Portrait
4.8 - The 5 Love Languages: The Secret to Love that Lasts
4.7 - First 100 Words
4.7 - How to Win Friends & Influence People
4.7 - The 7 Habits of Highly Effective People: Powerful Lessons in Personal Change

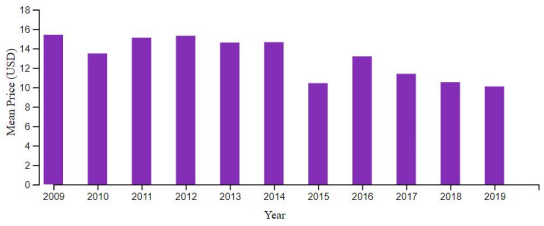
Non-Fiction
Fiction

Mean Prices of Best-Selling Books on Amazon

Mean Price Cutoff:

Filter Data

Reset Filter



For our final project, we examined book related data that showed the name, author, user rating, number of reviews, price, publish year, and genre of books that made it to the best-sellers list of Amazon shown above. The full dataset can be found here: https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019. We decided to encode various variables in two different visualizations; a line graph and a bar chart. Our audience is anyone who is interested in learning more about Amazon's best-sellers data. We made our visualization simple and intuitive so that anyone could interact with it and learn something without having much prior knowledge on information visualization.

Line graphs are especially effective at showing relationships over time, so in our line graph we plotted time on the x-axis (each year spanning from 2009-2019) and frequency of novels on the y-axis. Genre is encoded using color; non-fiction is a blue line and fiction is a red line. We chose these specific variables so that users would be able to easily look at our visualization and compare the number of best-selling fiction vs. non-fiction books by year. For instance, in 2015, it is clear that there were significantly more non-fiction books that became best-sellers rather than fiction. Additionally, we created a bar chart that shows the average price of best sellers per year (of both fiction and non-fiction combined). Like the line graph, we plotted each year on the x-axis ranging from 2009-2019; however, on the y-axis we plotted the average price per book. Using a bar chart for this allows readers to very quickly and easily interpret the average price per book per year as well as any trends in terms of the average price of a best-seller. We used purple to fill the bars of the bar chart since the chart showed fiction and non-fiction data and the two genres were encoded with blue and red in the aforementioned line chart.

Another feature we included on the line graph was the ability to hover over dots on the chart and see the highest rated books for the specified year and genre. The line graph enabled us to do this because each data value was aggregated into a discrete point onto the graph, which allows easy hoverability for the user. The data associated with the hovering appears on the right side of the line graph above the legend. We implemented this hover functionality so that users would be able to get a deeper look into the data - rather than just frequencies, they are also able to glance at which book titles were the highest rated best-sellers by genre and year. Another piece of functionality we included was the ability to filter both of the graphs by price according to user input. For the line graph, this is shown by adjusting the data points on the graph to show the frequency of books that fall under the specified price for each year and genre. This capability to incorporate price data allows additional user features, such as visualizing varied frequencies and bringing out different trends. The highest rated books list shown when hovering also updates when users opt to filter the chart. On the bar chart, we chose to include a filter so that users would be able to see how the average book price changes for each year based on their input. The bar chart filters the data by removing books in which the actual price is higher than or equal to the user input. This way, users can see the average price of a book in a given year while only taking into consideration the books that fell below their inputted price. There is also a reset button which we included so that users could easily start over and choose new prices to filter by and not have to refresh the page each time.

Contribution Summary: Justin worked on preprocessing the data as well as the bar chart. He also managed the GitHub repository and ensured there were no conflicts. Sukriti worked on the line graph and data preprocessing. Maggie worked on adding the titles and legend in addition to leading the write up. Philip worked on the filter functionality and the video component. William worked on the tooltip which shows highest priced books when users hover over dots on the line graph as well as the video component.

Video Link: https://youtu.be/Kk64iQmm4F8