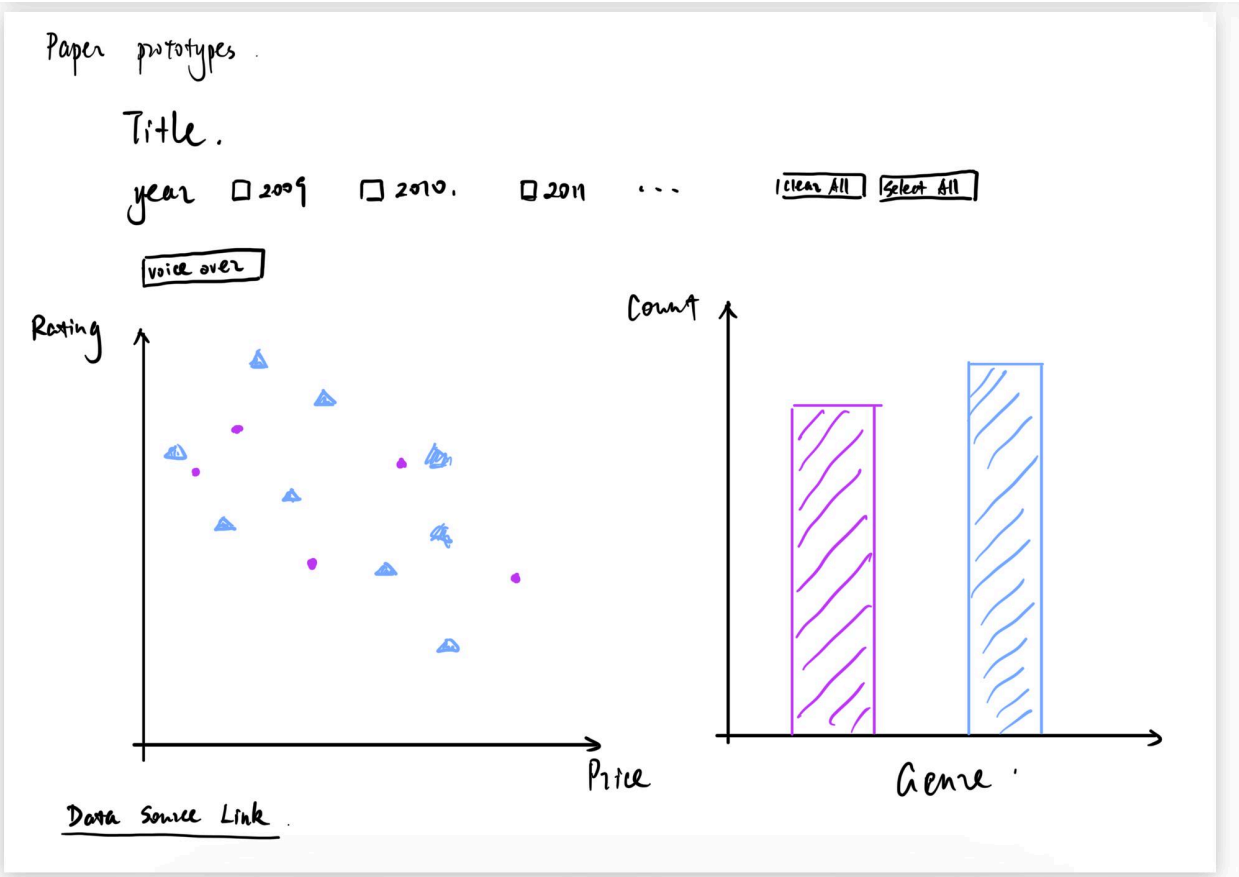
Final Project Writeup

1. Dataset Write-up:

- Our dataset is titled "Amazon Top 50 Bestselling Books (2009-2019)."
- The dataset includes each year's top 50 best-selling books on Amazon.com every year from 2009 to 2019.
- This dataset is useful for exploring commercial literary trends over time.
- It allows users to explore the insights of each top50 best selling books such as book name, author, price, rating on Amazon, and number of reviews.
- Each entry in the dataset is an individual book, with their attributes being price, author, Amazon user rating, number of user reviews, genre, and the year it was ranked within the top 50.
    - There are 550 books in this dataset
    - There are duplicates present because certain books remained in the top 50 for several years, some consecutive, some not.
    - The dataset does not specify the currency for the Price column.
        - Given that Amazon is a US-based company and this dataset was compiled using the U.S. version of Amazon, we assumed that all prices are in US dollars.
        - This assumption is explicitly noted in our documentation for transparency purposes
    - The dataset attribute ranges are as follows:
        - No. of Authors: 529 (quantitative)
        - Ratings: 3.3 – 4.9 (ordinal)
        - No. of Reviews: 37 to 87.8k (quantitative)
        - Price: $0 to $105 USD (quantitative)
        - Year: 2009 to 2019 (ordinal)
        - Genres: 2, non-fiction and fiction (categorical)
    - We found this dataset on Kaggle, and it was an cleaned dataset, therefore, no further dataset cleaning is needed
        - The dataset author scraped the data from Amazon, using Goodreads to categorise the books by genre.
        - We double-checked the "cleaned" status of the dataset, regardless, using the Pandas library to look for:
            - Null values
            - Values that incorrectly fall outside the ranges of the data
        - Our conclusions from this search were as follows:
            - There are no null values within the dataset.
            - No data entry mishaps (i.e. negative values for "price", ratings outside the range of 3.3-4.9, years outside of 2009-2019, more genre entries than non-fiction/fiction, and number of reviews.

2. Paper Prototype:



Annotations:

- **Interactions**
  - **Filter by Year**: Users filter the data by selecting one or more years using the check boxes. All check boxes are keyboard accessible (tab to navigate and space for selection).
  - **Bar Selection**: The users can click (or using keyboard) to select the genres they want to display. And the scatterplot will change accordingly. Multiple genres can be selected.
  - **Hover:** For the barchart, the users can hover on a bar to show the genre and book count. And for the scatterplot, the users can hover on a symbol (dot) to display information including book name, author, year, price, and rating.
- **Alt Text / VoiceOver**
  - All interactive elements (bars and symbols) have descriptive aria-labels to provide meaningful screen reader output.
- **Color Choices**
  - **Color Scale:** Uses a colorblind-friendly ordinal color palette. These colors should be easily distinguishable in both hue and saturation and were chosen to ensure visual clarity across genres.

3. Task Analysis:

3.1. Compare Values

Task: ***In a given year (e.g. 2015), which genre—Fiction or Nonfiction—had more books in the Top 50 list?***

How the user performs this task:

The user begins by filtering the dataset using the year checkboxes at the top of the dashboard, selecting only the year 2015. Upon selection, the bar chart dynamically updates to show the count of books in each genre for that year. The user compares the **length** of the two **bars** labeled "Fiction" and "Non Fiction." The taller bar indicates the genre with more bestselling books during the selected year. The user can further explore more details.

Marks/Channels Used:

- **Marks:** Bars (in bar chart)
- **Channels:**
    - **Height** (length of bar along y-axis) – encodes count,
    - **Color hue** (hue differentiates Fiction vs Non Fiction)

3.2. Range

Task: ***What is the range of user ratings for bestselling books across all years?***

How the user performs this task:

With all years selected, the user focuses on the scatterplot, where each book is encoded with its price on the x-axis and user rating on the y-axis. To determine the range of ratings, the user identifies the **points** that have the highest **vertical position** and lowest vertical position. For more precise values, the user may hover over the points or navigate to them using keyboard to display the exact ratings. And similarly, the user can explore the range of ratings or price in any given year.

Marks/Channels Used:

- **Marks:** Points (in scatter plot)
- **Channels:**
    - **Vertical position** – encodes the ordinal attribute of ratings

3.3. Order

Task: ***Among all the books in 2019, which three had the highest prices?***

How the user performs this task:

The user filters the dataset by selecting only the year 2019 check box. In the updated scatterplot, the user locates the **points** with rightmost **horizontal positions**. The three rightmost data correspond to the highest-priced books. By hovering over each point or via keyboard, the user accesses tooltip information that reveals the book title and price, allowing for a direct comparison and ordering.

Marks/Channels Used:

- **Marks: Points** (in scatter plot)
- **Channels:**
    - **Horizontal position** – encodes the quantitative attribute price,
    - **Shape** (circle or triangle) & **color** (hue) – encodes the categorical attribute genre.

4. Accessibility Analysis:

4.1. Compare Values (Non-Visual)

Task: ***In a given year (e.g. 2015), which genre—Fiction or Nonfiction—had more books in the Top 50 list?***

How the user performs this task:

A screen reader user navigates to the bar chart using keyboard tabbing. Each bar has an aria-label attribute that announces its genre and corresponding count (e.g., “Fiction: 22 books”). The user can tab between the two bars and listen to the respective counts.

Features:

- Tooltip information for each bar
- Keyboard focus on each point
- Voice output provided by screen reader

Information Gained/Lost:

- The exact count for each genre is retained through tooltip and voiceover.
- Visual comparison via bar height is lost, requiring more cognitive effort to compare values sequentially.
- Order of magnitude and visual gestalt are less immediate.

4.2. Range (Non-Visual)

Task: ***What is the range of user ratings for bestselling books across all years?***

How the user performs this task:

Users tab through each data point in the scatterplot. Each point is focusable and has an aria-label that describes the book's rating, author, title, and year. By noting the lowest and highest announced ratings, the user can infer the range.

Features:

- Tooltip information for each dot Keyboard focus on each point

Information Gained/Lost:

- The user can access every individual rating precisely, but must mentally keep track of min/max.
- Range estimation requires more memory and attention compared to visual scanning.

4.3. Order (Non-Visual)

Task: ***Among all the books in 2019, which three had the highest prices?***
How the user performs this task:
After filtering the dataset to 2019, the user tabs through all points. Each tooltip provides the price. The user can make notes or use screen reader memory features to track and sort the top three prices.
Features Needs:

- Checkbox filters accessible via keyboard
- Descriptive tooltips for each symbol

Information Gained/Lost:

- Exact pricing is accessible.
- Ordering requires note-taking or temporary memory retention since the scatterplot does not verbally sort data.
- The user loses the benefit of spatial arrangement that can be directly captured by vision.

5. Pilot Testing

5.1. Script and Steps:

Speaker: "We are evaluating our visualization and are asking you, the participant, to complete some tasks using the visualization and then provide feedback about the visualization and experience. As a reminder, we are evaluating the visualization, not you as a participant, so you don't need to worry about being "right" as you complete these tasks. There are three tasks, followed by a brief feedback session. The whole pilot session should take under 5 minutes. Do you consent to participate?"

[Wait for yes]

Speaker: "Thank you for agreeing to participate. We will start with the three tasks. Please 'think aloud' as you complete the task, meaning voice what you are thinking as you work through the task. Your first task is: ***In a given year (e.g. 2015), which genre—Fiction or Nonfiction—had more books in the Top 50 list?***

[Participant complete task 1]

Speaker: Thanks for completing the first task, the second task is**: *What is the range of user ratings for bestselling books across all years?***

[Participant complete task 2]

Speaker: Thanks for completing the first task, the third task is: ***Among all the books in 2019, which three had the highest prices?***

[Participant complete task 3]

Speaker: "That is the end of the third task. For this last bit, we welcome any feedback you may have about the visualization or about your process for completing the tasks."

[Allow participant to speak first, then informal discussion]

5.2. Notes

- Task 1
    - Participant 4 said, “So I need to look at just 2015” After clicking “Clear All,” he checked the box for 2015. “Non Fiction is definitely taller—more books in that genre.”
- Task 2
    - He clicked “Select All” and looked at the scatterplot. “The highest rating I see is 4.9,” he said while hovering over a topmost dot, then added, “and this one at the bottom is 3.3, so the full rating range is between those two.”
- Task 3
    - He clicked “clear all” and chose only 2019. He said, “Most expensive, so that’s on the right,” and hovered on three dots with the highest prices. “These three should be the most expensive.”

5.3. Feedback from participant

The participant thinks this visualization is generally very effective and clean. He also gave one suggestion of adding tooltip information that will display when hovering over the bar to show the number of books in each genre in selected years.

5.4. Changes to make

Adding tooltip information that will display when hovering over the bar to show the number of books in each genre in selected years.

6. Final Data Visualisation:

6.1 github.io pages and repo

- **Christian’s github.io page and repo:** https://theredplanetsings.github.io./ & https://github.com/theredplanetsings/theredplanetsings.github.io
- This is Russell’s github.io page and repo.

6.2 WAVE Checks Summary

- Task 1
    - This task can be achieved by tabbing. After the user used tabbing to clear all years and select 2015, the user can tab to the two bars, and screen reader accessible tooltip information will be displayed, telling the user the count of books in each genre in 2015. Then the user can know which genre has more books ranked top-50 in 2015.
- Task 2

    - This task can also be achieved by the hidden texts in the webpage. The user uses the keyboard to move through the structured HTML elements on the page, then reach the hidden heading, from which the screen reader will reads the alt text aloud: *most books are highly rated, typically between 4.3 and 4.9 stars.* The user then knows the range of ratings of books through all years.
- Task 3
    - This task can be achieved by tabbing. After the user used tabbing to clear all years and select 2019, the user can tab through all dots in the scatterplot and remember the highest priced books. The limitation in this is it requires attention and memory, therefore not very effective.

7. User Testing:

7.1. Participant 1

- Major/Minor: Math
- Notes
    - Task 1
        - Participant 1 read the first task, then had a glimpse at the screen, “so first clear all years” he clicked the “clear all” button. Then he said, “2015”, then clicked the checkbox of 2015, then the data of 2015 displayed. He said “okay, so the bar of nonfiction is significantly higher, I guess that means there are more non-fictions books in the top-50 this year.”
    - Task 2:
        - Participant 1 read the second task, then clicked the button “select all”, then the data for all years displayed. He hovered the mouse to the dot with the highest vertical position, then hovered it to other dots with the same vertical position. “So there are no books with ratings higher than 4.9.” Then he hovered the mouse to the dot with the lowest vertical position (on x-axis). When the text with information showed up, he said “so the lowest is 3.3.” Then I think the range is from 3.3 to 4.9.”
    - Task 3:
        - Participant 1 read the third task, then clicked the button “clear all”, then selected the checkbox of 2019. Then he started by hovering the mouse to the point with the rightmost horizontal position. Then check the second rightmost horizontal position dot, then the third. He read the book names that are displayed by hovering on the dot, and ranked from the rightmost to the left for the three most expensive books.
- Participant feedback
    - Participant 1 thought it was an interesting visualization, and what he liked the most was the buttons of “select all” and “clear all.” He thought this visualization is very effective and easy to read.

- Summary
    - Participant 1 was able to complete all three tasks easily. He found the buttons “Select All” and “Clear All” intuitive and effective in simplifying the filtering process.

7.2. Participant 2

- Major/Minor: Computer Science
- Notes
    - Task 1:
        - Participant 2 read the task aloud, then said “let’s start by selecting 2015.” He clicked “Clear All,” then selected the checkbox for 2015. Observing the bar chart, they immediately said, “Non-fiction is clearly higher, so there were more top-selling nonfiction books this year”
    - Task 2:
        - Participant 2 clicked “Select All” to display all years. He hovered over the dots with the highest vertical position on the scatterplot and noted: “Okay, there are many 4.9, but no 5.0, so 4.9 is the highest rating” Then he hovered toward the bottom of the chart and said, “Here’s 3.3—so the range is 3.3 to 4.9.”
    - Task 3:
        - After clearing the checkboxes and selecting only 2019, Participant 2 examined the scatterplot by moving from right to left along the x-axis. He hovered over each dot to read the price and book title, then said, “So those three books are the most expensive”
- Participant feedback
    - Participant 2 thought this visualization is clean and effective, and he thought it will be better if the two plots can take all of the screen’s length (so maximizing horizontal length of the screen).
- Summary
    - Participant 2 said the interaction was “really smooth and clean,” and suggested possibly increasing the size of the charts to make it take up the full length of the screen.

7.3. Participant 3

- Major/Minor: Psychology
- Notes
    - Task 1:
        - Participant 3 began by reading the task aloud and said, “Okay so the year is 2015.” he pressed the “Clear All” button, then selected only the checkbox for 2015. Looking at the bar chart, they hovered over each bar and commented, “Non Fiction has a higher count than Fiction, so Non Fiction wins this year.”
    - Task 2:

        - Participant 3 clicked “Select All” to include all years. He hovered to the points at the top of the scatterplot and said, “Looks like the highest is 4.9,” then moved toward the lower dots and added, “and this one down here is 3.3. So the range is from 3.3 to 4.9.”
    - Task 3:
        - Participant 3 reset the filters by clearing all checkboxes and selecting only 2019. He moved across the scatterplot from right to left, pausing on three dots with the highest x-values. He read the tooltip contents and wrote down the book names as they hovered over them. “I think these are the top three most expensive.”
- Participant feedback
    - Participant 3 found the tooltip information “super helpful.”
- Summary
    - Participant 3 also completed all tasks easily and really loved the tooltip information displayed when hovering over bars and dots.

8. Personal Reflections:

- Russell:
    - The original prototype presented in presentation #4 is a stacked barchart with a dropdown menu that allows users to select which attribute to be displayed. However, we are inspired from class that we explore interaction and multiple views more. Therefore, we decided to do something similar to lab 7, to control the data shown in the scatterplot by selecting bars in the bar chart. This visualization allows the users to explore detailed information of individual books, which is one aspect that may not be achieved in our original prototype and questioned by one of our classmates. We decided to use price and ratings as attributes to be shown in the two axes because the number of reviews varies in a too large range, causing the graph to look clustered for books with few reviews and to lose valuable insights. With this general idea, we try to expand on our lab 7, and one thing we noticed is that if we show data from all years, it will cause the scatter plot to be too dense and not provide meaningful insights. To solve this, we add check boxes that allow users to select none to multiple years of books they want to display. When the user wants to explore the detail of a specific entry, the user can hover to that dot, and the information of this book will be displayed, providing more insights. In terms of accessibility, we are very careful in designing the visualizations to make it accessible. For example, we not only used color hue to distinguish books of different genres, we also used shapes (triangle and dots) to encode the genre attribute, making sure it is accessible. In addition, we make sure that each selectable checkbox, bar, and hoverable dots and bars can be accessed through tab. In the testing process, we have positive feedback and all the participants can access those tasks with ease.
- Christian:

- Looking back, I feel satisfaction with the evolution of our work and what we delivered. Initially, our dataset seemed fairly straightforward—bestselling books, ranked by year with common attributes such price, rating, and genre. However, as we began designing the visualisation, we realised that in order to convey meaningful insights, we needed to actually consider not just what we were showing, but how users would interact with and interpret that information. One of the most important design decisions we made was moving away from a single, static chart to a more interactive & dynamic experience. Our goal here was to allow users to explore trends across multiple years and between genres in a way that was visually intuitive and accessible alike. I'm particularly pleased with the scatterplot & bar chart coordination, and how filters allow users to have direct control over the dataset displayed on them. In early brainstorming sessions, it was emphasised that we needed something more engaging than a basic filter or dropdown, and I think we achieved that by implementing multiple interaction modes. Another important aspect of this project was accessibility. Assigning meaningful aria-labels, using shape and colour to distinguish genre, and ensuring that all interactive elements were focusable via keyboard were not just extra steps, we saw them as essential. On the technical side, testing and iterating with pilot users helped us uncover usability issues we hadn't initially anticipated. We received feedback suggesting tooltips on the bar chart, which we then implemented. This improved the experience for both visual and non-visual users. Overall, I learned a great deal not only about visualisation tools and techniques, but also about how thoughtful interaction design makes what are simple datasets come alive to become a great data visualisation.

References

Saalu, Sooter. “Amazon Top 50 Bestselling Books 2009 - 2019.” *Kaggle*, 13 Oct. 2020, www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019/data.