MilestoneII: Report

Group Member: Camilla Ren, Mina Yang, Xiwen Wei

Introduction

In this report, we focus on analyzing the Amazon book dataset (with several merged book dataset), which contains information on books sold on Amazon's platform. This dataset includes a wide range of information, including book titles, authors, genres, prices, ratings, reviews etc. By studying the data, we are able to gain valuable insights into consumer behavior, identify popular genres and authors, and make informed decisions about pricing and marketing strategies. This will help Amazon's CEO and publishers better understand the market and make data-driven decisions to drive growth and profitability.

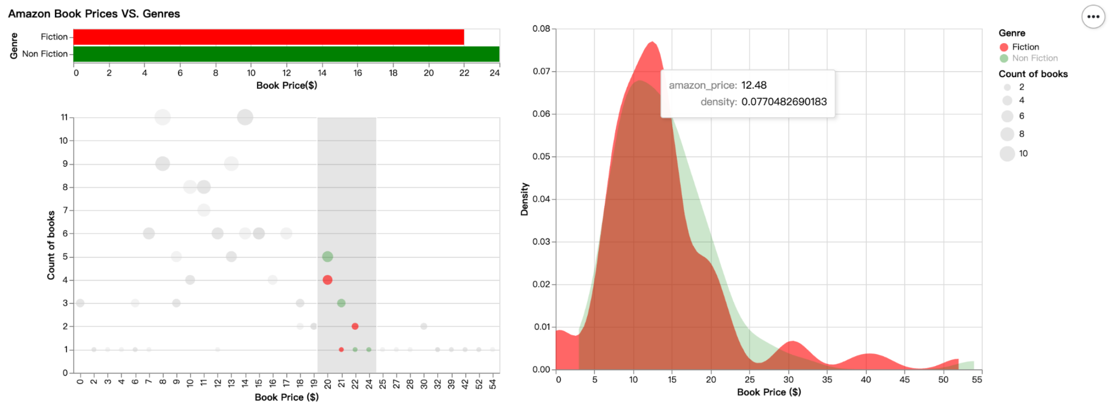
To achieve our goal, we present three distinct views, each utilizing various visualization techniques such as boxplot, scatterplots, bar charts, density plot, and pie chart. We also incorporate interactive features that allow users to explore the data in greater detail. We have identified five tasks below that may be of particular interest to Amazon's CEO and publishers.

Summary of the tasks

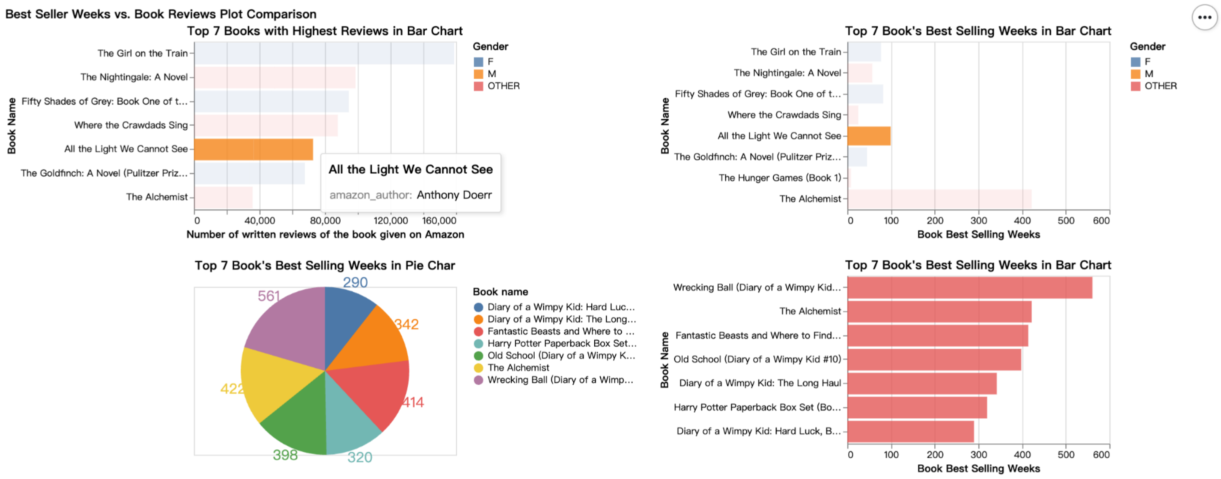
1. Task 1: Browse the prices between fiction and nonfiction books.
2. Task 2: Who is the most popular (female/male) author in the last few years (with highest reviews)?
3. Task 3: Which genre of book has the highest rating?
4. Task 4: Which book is the best seller?
5. Task 5: Is there a big price difference between science fiction and non-science fiction books?

Final prototype

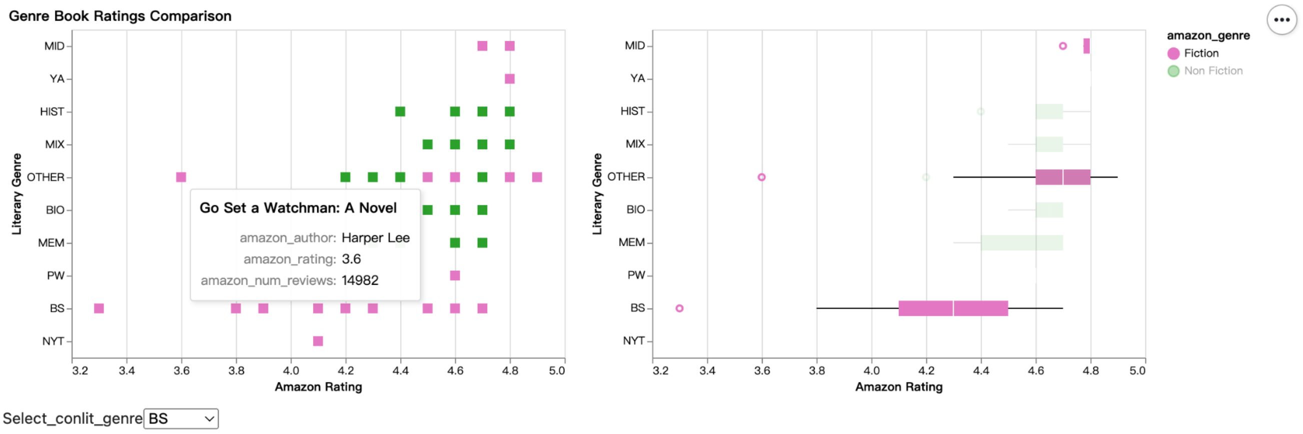
View 1:



View 2:



View 3:



Justification of visualization choices

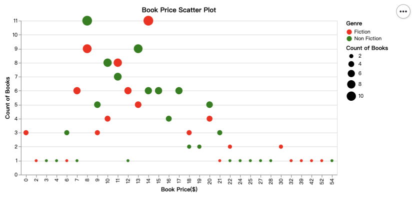
View 1

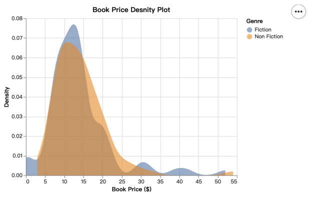
Task 1 + Task 5:

- Task 1: Browse the prices between fiction and nonfiction books.
- Task 5: Is there a big price difference between science fiction and non-science fiction books?

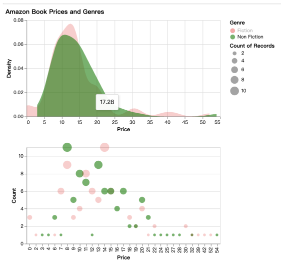
Preliminary sketches:

- Iteration 1:

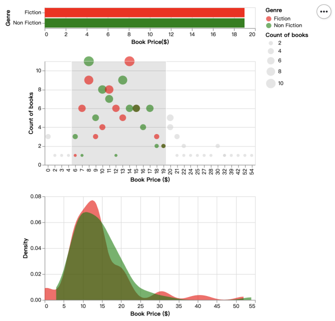




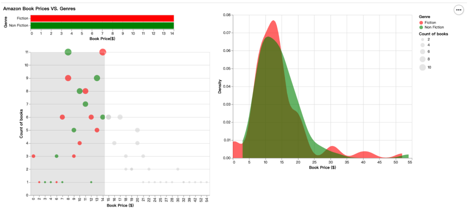
- Iteration 2:



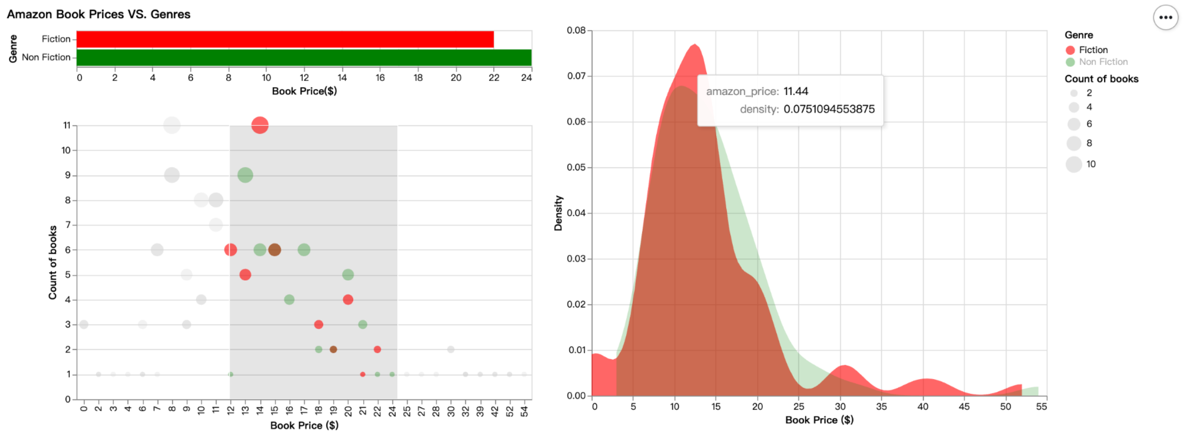
- Iteration 3:



- Iteration 4:



Final Prototype:



Justification:

- **Marks**: point, area, line
- **Channels**:
  - Position: horizontal (x-axis), vertical (y-axis)
  - Color: hue
  - Size: length, area
- **Characteristics of Channels that were exploited**
  - Discriminability: 2 colors and 5 circle sizes
  - Separability: color, size and position
  - Popout: color, tooltip
  - Grouping: color
  - Accuracy: Size, color and position
- **Describe the interaction**
  - Selections (tooltip, mouseover, click):
    - Tooltip is used to provide additional information when the mouse pointer hovers over a data point. In plot_a and plot_b, the tooltip displays the book price and count/density of the selected data points, respectively.
    - Mouseover helps the bushing and bidirectional linking interactions as described below. When the mouse pointer hovers over a legend item in the scatterplot, the corresponding points in the plot are highlighted and their density is shown in the density plot.
    - When the user clicks on a genre in the legend of the scatterplot, the corresponding points of that genre are highlighted in the scatterplot and their density is displayed in the density plot.
  - Brushing:

        - Brushing allows users to select a range of prices on the scatterplot (plot_a) by clicking and dragging the cursor. This selection is then highlighted in the density plot (plot_b) and filtered in the attached bar chart (plot_c) above, enabling users to see the distribution of book prices by genre for the selected range as well as the density of books within the selection.
    - Bidirectional linking:
        - Bidirectional linking enables users to mouseover on a point in the scatterplot (plot_a) to highlight the corresponding points (same genre) and show their density in the density plot (plot_b). Similarly, mouse overing on the area in the density plot (plot_b)) highlights the corresponding points in the scatterplot (plot_a).
    - Ease of Use:
        - The chart is user-friendly and has interactive features that allow users to easily explore the data and understand the connection between different variables.
- **Characteristics of Interaction and Interactivity**
    - There are different facets/views:
        - Horizontal bar chart
        - Scatter plot
        - Density Plot
    - Interactional coupling
        - Bi-drectional (scatter plot and density plot): Hovering over one point/area in one graph leads to changes in other graphs.
        - Unidirectional (scatter plot and bar chart): If points reside within any brush in the scatterplot, it will also be shown in the bar chart. The opposite is not true.
    - Interactivity - Action:
        - Focus
            - Direct: Users can interact directly with the view through hovering over a data point, which allows them to explore specific data points within the same genre.
        - Presence
            - Implicit: User needs to explore to know that we can drag and hover over the data to see more information.
        - Granularity

            - Composite: there are more than one action (hover over, brush, click)
    - Interactivity - Reaction
        - Spread
            - Propagated form
                - Changes in one viz changes other vizzes
        - Activation
            - Immediate: The vis reacts to the user’s action instantaneously. For example, when a user hovers over a data point on the scatter plot, the visualization immediately displays additional information about that data point.
        - Flow
            - Discrete flow: The reaction occurs instantaneously. For example, users can hover over a specific data point, and the visualization would update the display to highlight that data point.
            - Continuous flow: When users brush points in the scatterplot, the bar chart above would immediately update. As the user continues to interact, the updates to the bar chart would occur seamlessly and without interruption.
- **Critique the view**
    - Data represented: The visualization represents data on the price as of October 13, 2020 for books in different genres. The data includes categorical variables such as the genres (fiction and non-fiction), and quantitative variables such as the price and the number of books.
    - Data encoded: The data is encoded in the visualization using 3 types of marks: point, area and line. The size of points represents the number of books within the same price. The channels used in the visualization include the y-axis for density and counts of book, the x-axis for the book price, the color channel for the genre, and the tooltip channel for additional information on individual books.
    - Questions answered: Are non-fiction books generally more expensive than fiction books? Is there a significant price difference between the two genres?
    - Discriminability: Only two colors and five circle sizes are used in the plots, which are sufficient.
    - Separability: Colour and size are perceived separately in the scatterplot with some interference when the circle size is really small. Position is fully

separable from color and size because the use of position channels is not as affected by use of color and size channels.

- Popout: The popouts are the contrasting color scheme (light/dark) and the interactive tooltips, which serve as the primary visual cues.
- Grouping: Color is used to differentiate between genres, making it easy for the user to discern information within the same genre.
- Accuracy: The accuracy is acceptable because the visulization uses size, color and position as channels. Length is accurate as it is linear while color and size are not linear and therefore hard to make comparisons. Accurate information is also provided by tooltip.
- Effectiveness:
  - The use of interactions, such as "mouse hover", “click” and "brush" selections, allows users to filter and explore specific genres or price ranges, making it easy to investigate specific trends and patterns in the data. The tooltip also provides additional information about specific data points as the user hovers over them.
  - Informative: The tooltip provides additional information about the data, giving users context and background information about the chart.
- Limitations and ineffectiveness: A potential weakness of this view is that it only shows data from Amazon, so the results may not be representative of book prices and genres on other platforms. The dataset is also small. In addition, the visualisation does not take into account other factors that may affect book prices, such as publication date, author popularity, or publisher.

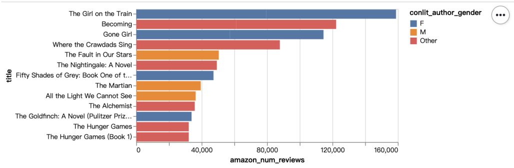
View 2

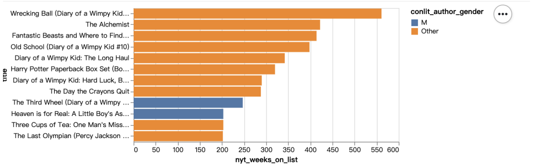
Task 2 + Task 4:

- Task 2: Who is the most popular (female/male) author in the last few years (with highest reviews)?
- Task 4: Which book is the best seller?

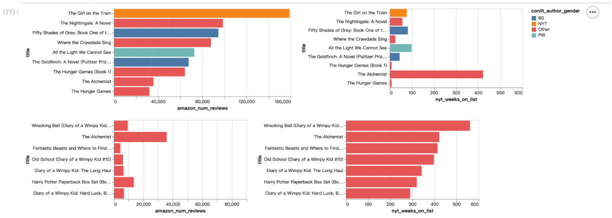
Preliminary sketches:

- Iteration 1:

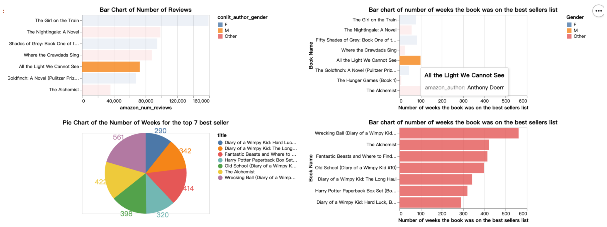




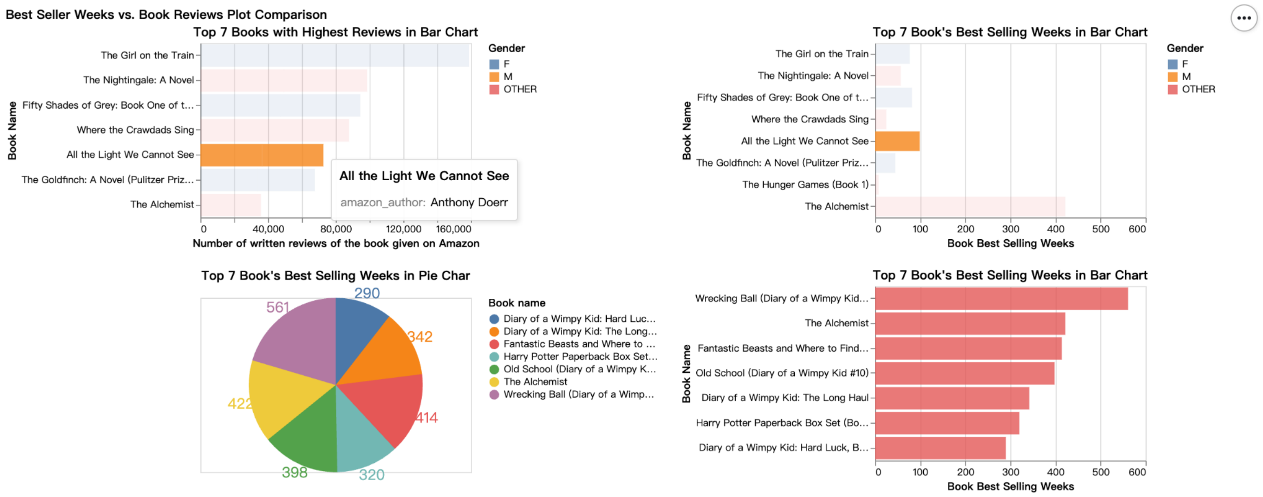
- Iteration 2:



- Iteration 3:



Final Prototype:



Justification:

- **Marks**: line (bar charts), area (pie chart)
- **Channels**:
  - Position: x-axis (horizontal), y-axis (vertical)
  - Color: hue
  - Size: length, area
- **Characteristics of Channels that were exploited**
  - Discriminability:
    - Bar charts: 3 colors
    - Pie chart: 7 colors
  - Separability: color
  - Popout: color (dark/light), tooltip
  - Grouping: color (separate the gender in bar charts and book name in pie chart)
  - Accuracy: number labelled, position
- **Describe the interaction**
  - Linked Data: The chart is designed to show the relationship between the different data points displayed on the three bar charts and the pie chart. The linking of data helps users to understand how the different data points are related and how they interact with each other.
  - Bi-directional linking interaction: The 2X2 side by side chart contains three bar charts and one pie chart, and they are linked together. All charts use the title attribute. The top left bar chart displays data showing the number of reviews while the other three charts show the number of weeks.
  - Mouse-based interaction: When a user hovers over a bar in any of the three bar charts, a tooltip appears, displaying the relevant data point for that bar. The tooltip also shows the name of the author and the title of the book that the data point represents.
  - Mouse-based interaction: If a user clicks on the bar chart on the first row, the corresponding book title is highlighted, showing the author name and book title in the tooltip. The highlighting of the bar chart makes it easier for the user to see which book the data point represents.
  - Tooltip: The chart uses a tooltip to provide additional information about the data points. The tooltip shows the author name and book title, making it easy for users to identify the book associated with each data point.

    - Ease of Use: The chart is designed to be user-friendly and easy to use. The interactive features of the chart make it easy for users to explore the data and gain insights into the relationship between different variables.
- **Characteristics of Interaction and interactivity**
    - There are different facets/views:
        - Horizontal bar chart
        - Pie chart
    - Interactional coupling
        - Bi-drectional (upper two bar charts): Hovering over one bar in one graph leads to changes in other graphs.
    - Interactivity - Action:
        - Focus
            - Direct: Users can interact directly with the view through clicking a bar, which allows them to explore the selected book in another graph.
        - Presence
            - Implicit: User needs to explore to know that we can hover over the bar to see the book title and author name, as well as clicking a bar to highlight the selected book in another graph.
        - Granularity
            - Composite: there are more than one action (hover over, click)
    - Interactivity - Reaction
        - Spread
            - Propagated form
                - Changes in one viz changes other vizzes
        - Activation
            - Immediate: The vis reacts to the user’s action instantaneously. For example, when a user hovers over a bar, the visualization immediately displays additional information about that book.
        - Flow
            - Discrete flow: The reaction occurs instantaneously. For example, users can hover over a bar, and the visualization would update the display to highlight that bar in another graph. When users click a bar, the bar chart beside it would immediately update. As the user continues to interact, the visulization would update the display in a stepwise fashion.

- **Critique the view**
    - Data represented: The visualization represents data on the number of weeks the book was on the bestseller list for books and number of written reviews of the book given on Amazon in different genres. The data includes categorical variables such as the author genders and the names of books, and quantitative variables such as the number of reviews and the number of weeks.
    - Data encoded: The data is encoded in the visualization using two types of marks: line and area. The upper left graph shows the number of reviews the book had. The rest of three charts represent the number of weeks the books are on the bestseller list. The upper twos are for the top 7 books with highest review while the bottom twos are for the top 7 best selling books.
    - Questions answered: Is the most popular book (with the highest number of reviews) the best seller? Alternatively, is the best-selling book the most popular book?
    - Discriminability: Only 3 colors are used in the plots, which are sufficient. However, more than 5 colors are used in the pie chart, which is not recommended because using more colors can create difficulties when perceiving information.
    - Separability: Color and position are fully separable because the use of position channel is not as affected by use of color channel.
    - Popout: The popouts are the contrasting color scheme (light/dark) and the interactive tooltips, which serve as the primary visual cues.
    - Grouping: Color is used to differentiate between genders, making it easy for the user to discern information within the same gender.
    - Accuracy: The accuracy is relatively low in the pie chart because it uses area as channel and it is hard to tell the difference between the different sizes. However, bar charts are accurate as the length is linear. Accurate information is also provided by tooltip.
    - Effectiveness: This type of visualization can be useful for comparing and contrasting data across different categories or groups. The side by side layout allows for easy comparison of the data, and the mouse-based interaction and tooltip can provide additional information and context about the data points, such as the author name and title. This can be particularly helpful in situations where the data is complex or there are many different variables to consider.
    - Limitations and ineffectiveness:

- Lack of Interactivity: While the chart allows users to interact with the data through mouse-based interaction, the actions available are limited to hovering and clicking. It may be beneficial to consider additional interactive features such as drag-and-drop functionality, filtering options, or data sorting capabilities.
- Overcrowding: With three bar charts and a pie chart all displayed in a single 2X2 chart, there is a risk of overcrowding and making the chart difficult to read. It would be helpful to consider reducing the number of charts or finding alternative ways to display the data to avoid overwhelming the user.
- Lack of Context: The chart's interactivity features are dependent on the user's ability to recognize and understand the different data points displayed. It may be beneficial to provide additional context or explanations for the data to help users better understand the chart and interpret the data accurately.

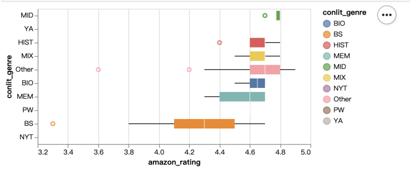
View 3 (UI):

Task 3:

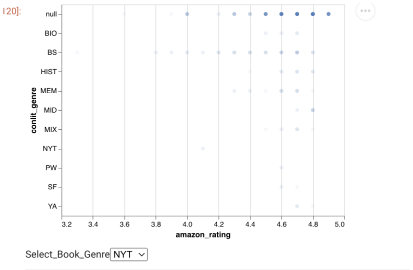
- Task 3: Which genre of book has the highest rating?

Preliminary sketches:

- Iteration 1:

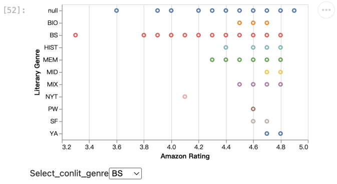


- Iteration 2:



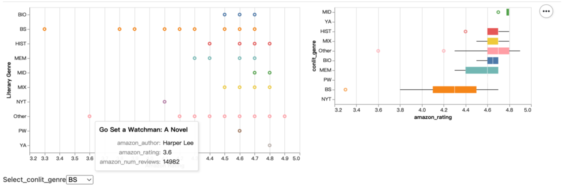
Select_Book_Genre NYT

- Iteration 3:



Select_conlit_genre BS

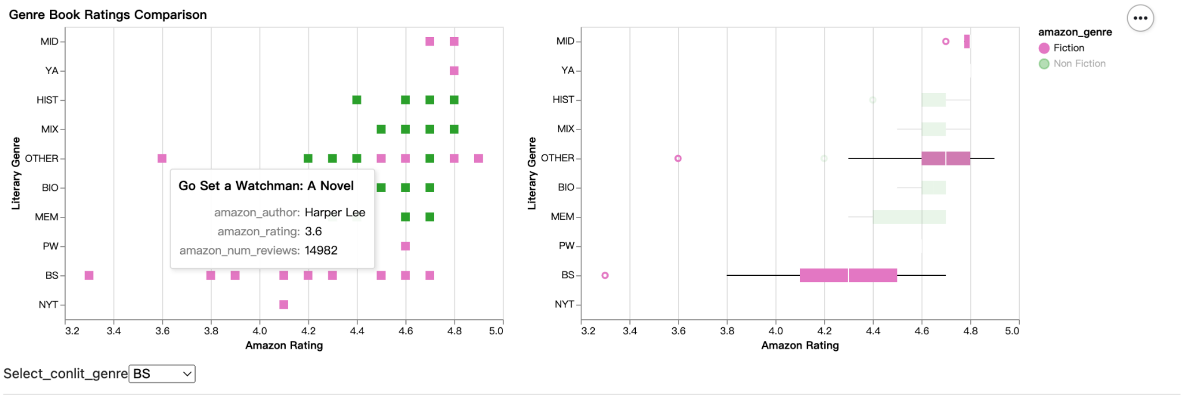
- Iteration 4:



Go Set a Watchman: A Novel
amazon_author: Harper Lee
amazon_rating: 3.6
amazon_num_reviews: 14982

Select_conlit_genre BS

Final Prototype:



Go Set a Watchman: A Novel
amazon_author: Harper Lee
amazon_rating: 3.6
amazon_num_reviews: 14982

Select_conlit_genre BS

Justification:

- **Marks**: boxplot, square point, circle point
- **Channels**:
  - Position: x-axis (horizontal), y-axis (vertical)
  - Color: hue
  - Size: length
- **Characteristics of Channels that were exploited**
  - Discriminability: 2 colors
  - Separability: color
  - Popout: color (dark/light), tooltip
  - Grouping: color (differentiate between the different genres)
  - Accuracy: position
- **Describe the interaction**
  - Dynamic queries (UI widget): The right hand side chart allows viewers to interactively select a literary genre using the selection widget, which filters the data and highlights the selected genre while fading out the others.
  - Panning and Zoom interaction: The right hand side chart allows viewers to inspect dense regions more closely
  - Mouse-based interaction(tooltips): Both charts use tooltip to provide additional information
- **Characteristics of Interaction and interactivity**
  - There are different facets/views:
    - Scatter plot
    - Box plot
  - Interactional coupling
    - Unidirectional: Hovering over one point in scatter plot will highlight the points within the same genre in boxplot. The opposite is not true.
  - Interactivity - Action:
    - Focus
      - Direct: Users can interact directly with the view through hovering over a data point / boxplot, which allows them to explore specific data points within the same genre.
      - Indirect: Users can interact with the view through controls to filter the data and highlight the selected genre while fading out the others, which helps to provide a more focused view of the data.
    - Presence

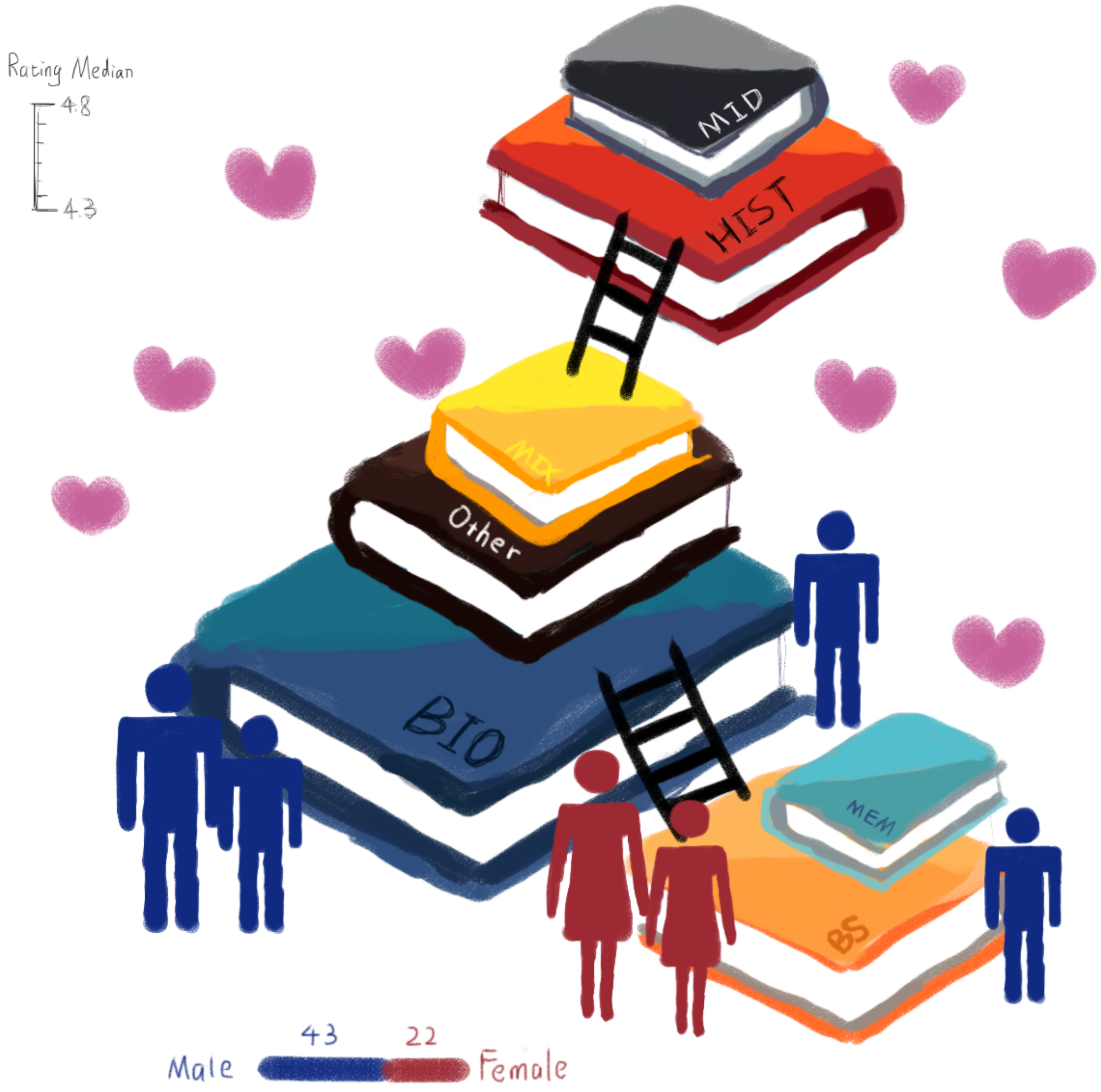
            - Implicit: User needs to explore to know that we can hover over the data to see more information.
        - Granularity
            - Composite: there are more than one action (hover over, selection widget)
    - Interactivity - Reaction
        - Spread
            - Propagated form
                - Changes in one viz changes other vizzes
        - Activation
            - Immediate: The vis reacts to the user’s action instantaneously. For example, when a user hovers over a data point/boxplot, the visualization immediately displays additional information about that data point.
            - On-demand: Users can select an option in a dropdown menu to filter the data within the selected genre.
        - Flow
            - Discrete flow: The reaction occurs instantaneously. For example, users can hover over a point in the scatterplot, and the visualization would update the display to highlight that boxplot. Users can also select options in the dropdown menu to choose a genre. As the user continues to interact, the visulization would update the display in a stepwise fashion.
- **Critique the view**
    - Data represented: The visualization represents data on the Amazon rating for books in different genres. The data includes categorical variables such as the literary genre and the Amazon genre, and quantitative variables such as the Amazon rating and the number of reviews.
    - Data encoded: The data is encoded in the visualization using two types of marks: boxplots and squares. The boxplots represent the median rating for each literary genre, and the squares represent individual books. The channels used in the visualization include the y-axis for the literary genre, the x-axis for the Amazon rating, the color channel for the Amazon genre, and the tooltip channel for additional information on individual books.
    - Questions answered: The visualization answers the question of how the Amazon rating varies across different literary genres and genres on Amazon.

It also allows for filtering by selecting a specific literary genre, as well as zooming in on a specific range of ratings.

- Discriminability: Only 2 colors are used in the plots, which are sufficient.
- Separability: Color and position are fully separable because the use of position channel is not as affected by use of color channel.
- Popout: The popouts are the contrasting color scheme (light/dark) and the interactive tooltips, which serve as the primary visual cues.
- Grouping: Color is used to differentiate between genres, making it easy for the user to discern information within the same genre.
- Accuracy: The linear length of the visualization contributes to its accuracy, and additional information is provided through the use of tooltips.
- Effectiveness:
    - The visualization is effective in showing the distribution of ratings across different genres and in allowing for comparisons between genres. The use of box plots helps to highlight the median rating and the range of ratings for each genre, while the squares allow for exploration of individual books. The selection and zooming features also add interactivity and allow for deeper analysis.
    - Highlight: The interactivity is enabled through the alt.selection_single() function that creates a selection object that can be used to filter the data and update the chart accordingly. When a genre is selected, the opacity of the points that don't belong to the selected genre is reduced, while the selected genre's points are displayed at full opacity. This allows the viewer to focus on the selected genre while still being able to see the other genres' data.
    - Informative: The tooltip provides additional information about the data, giving users context and background information about the chart.
    - User-friendly: The zoom and panning, drop-down menu interaction and tooltip make the chart user-friendly and easy to navigate, enabling users to explore the data without requiring any special technical expertise.
- Limitations and ineffectiveness: One potential limitation of the visualization is that it does not provide information on the number of books within each genre, which may influence the distribution of ratings. Additionally, while the visualization allows for exploration of individual books, it may become cluttered if there are many books in a specific genre.

Novel Vis

For amazon_book dataset:



Height represents rating: the higher the height is the higher the rating

The text on the book represents the genres

The ratio of male author to female author is 2:1