ALY_6000 Project 3 Report

Junchen Yi

College Professional Studies in Analytics, Northeastern University

ALY_6000: Introduction to Analytics

Zhi He

Submission Date: 10/10/2023

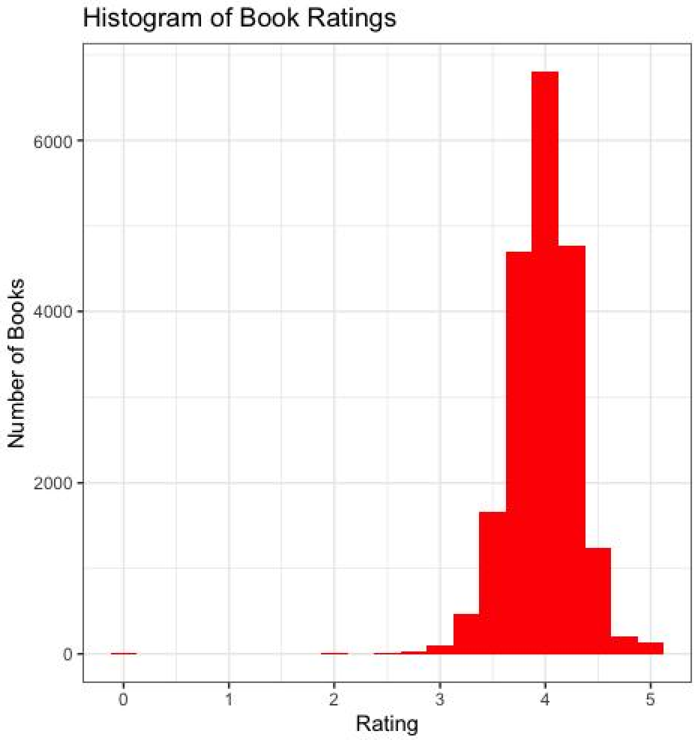
Introduction

For this assignment, I was asked to learn how to clear a specific data set and learn to visualize the data using different methods.

Analysis and Key Findings

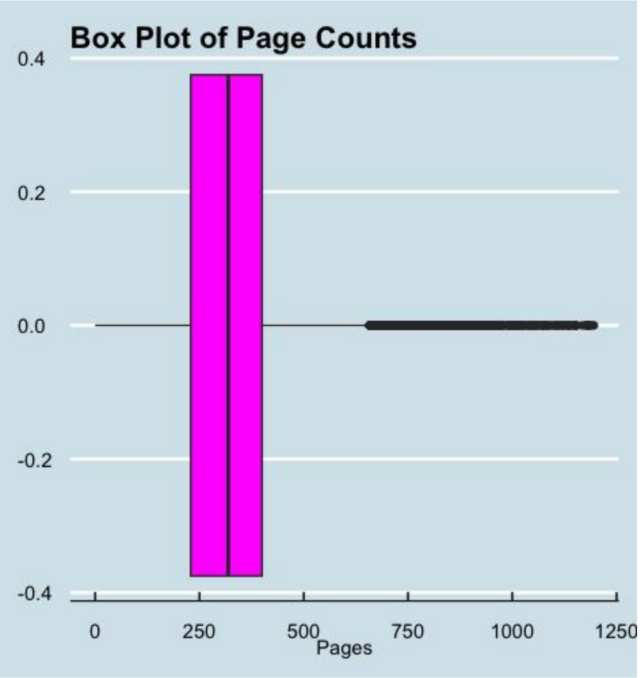
Question 3:

The first graph I drew was a histogram, using the geom_histogram function from the ggplot2 package, with rating as the horizontal coordinate, number of books as the vertical coordinate, a group spacing of 0.25, and the histogram filled in red. As the graph shows, almost all the books have ratings between four and six, and more than 6,000 books received a rating of four, which indicates that most of the books are rated highly.



Question 4:

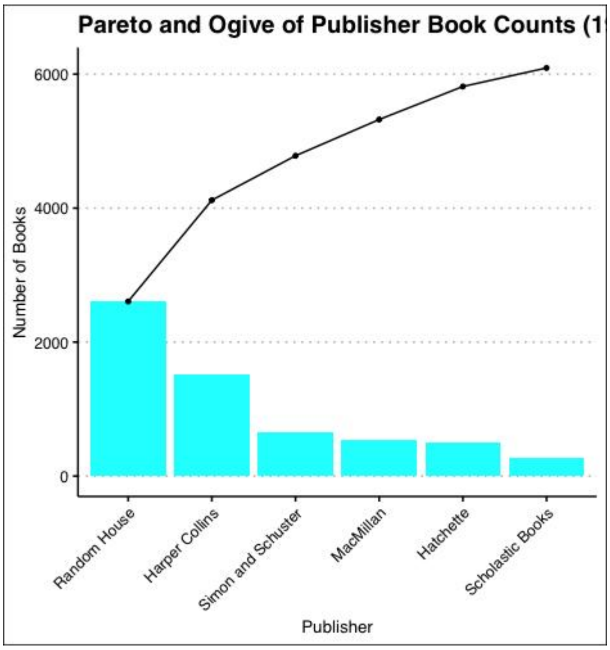
This problem uses the geom_histogram method to draw a box plot, using ggplot() to initialize the plot with the data frame subset_data and specify that the x-axis represents the "page" column. Use geom_boxplot() to add a horizontal boxplot layer and fill the boxes with magenta. Use labs() to add labels to the x-axis and set a title. Apply the "theme_economist()" theme to the plot.



Question 6:

This problem uses three methods to draw a Pareto Chart, using geom_bar to draw a bar chart, geom_line to draw a line chart, and geom_point to draw each data point. Also, use ggplot() to initialize the plot with the data frame summary_data and specify the x-axis as the "publisher" column (reordered by number of books). Use geom_bar() to add a bar layer and fill the bars with cyan. Use labs() to add labels for the x-axis, y-axis and set titles. Apply the "theme_clean()" theme for a neat look. When rotating the labels on the x-axis, I found that using ‘element_text(angle = 45, hjust = 1)’ can achieves requirement。

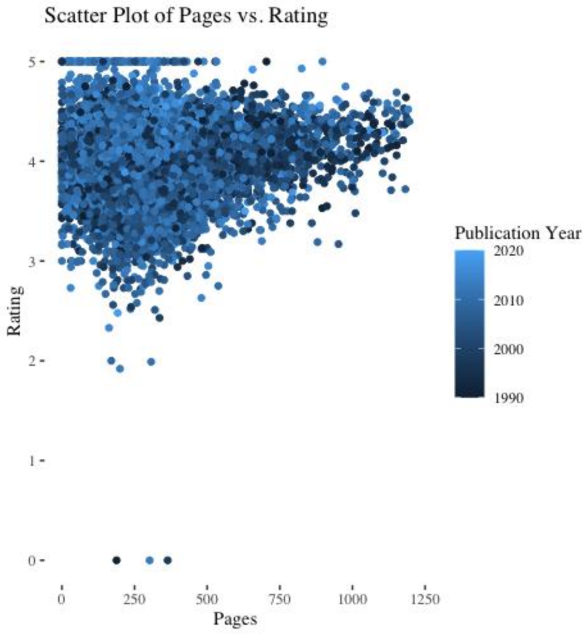
In the Pareto Chart, we can find that the publisher with the most books published is Random House, the only publisher with more than 2,000 books. In contrast, the publisher with the fewest books published is Scholastic Books with only 277 books.



Question 7:

In this code, I used geom_point() to create a scatter plot and mapped "pages" to the x-axis and "ratings" to the y-axis. Next, use color = year to color the points according to the year of publication, use 'na.rm = TRUE' to remove nulls, and provide appropriate labels and titles. Finally, apply the 'theme_tufte()' theme for a cleaner look.

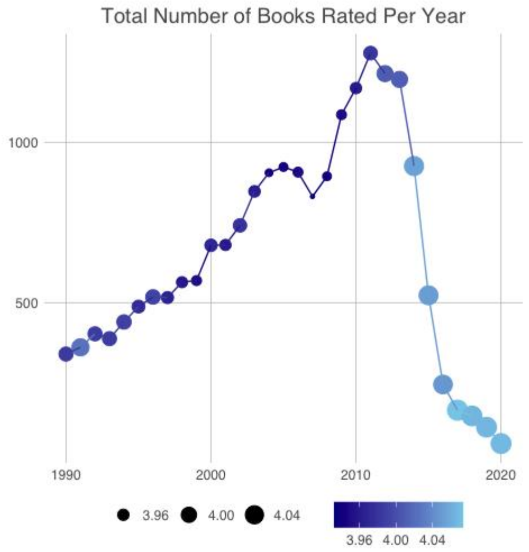
This scatter plot shows that most of the books have a page count of less than 750. by looking at the rating, a small number of the ratings are able to reach a score of 5 or more, and the majority of the books are rated between 3 and 5. I think the more pages a book has, the less likely it is to have a low rating



Question 9:

In this code, we use ggplot to create a line plot, mapping "Year" to the x-axis, "Count" to the y-axis, and "AvgRating" to the color as well as the size of the points. Use geom_point() and geom_line() to add points and lines, respectively.

A graph is drawn showing the total number of books rated each year from 1990 to 2020, and the points are colored according to the average rating for each year. We can see an upward trend in the number of books rated from 1990 to 2013, followed by a sharp decline in that number to 61 in 2020. Although the number keeps decreasing, the average rating of the books keeps increasing. From 2014 onwards, the average rating of the books is above 4.04.



Question 12:

Sample Result:

```
  average    variance        sd    name
1  4.0096 0.08770489 0.2961501 sampleA
2  3.9292 0.11329632 0.3365952 sampleB
3  3.9634 0.11906105 0.3450522 sampleC
```

Original Result:

```
avg_rating    variance        sd
1    3.978595 0.09633514 0.310379
```

I created three data frames, each selected with a data size of 100, and merged them. We can find that each attribute of the sampling result is different from the original data frame, when the mean is smaller than the original data, the variance and standard deviation will be larger than the original data (as can be seen by Sample B and Sample C); when the mean is larger than the original data is, the variance and standard deviation will be lower than the original data (as can be seen by Sample A).

Question 13:

For the visualization exercise, I want to find the number of books that have a ‘linked percent’ greater than ninety percent and show it as a bar graph.
Process:
1. created a new data frame add_data, used 'group_by' to create the grouping based on the variable "linked_percent" and then summarize calculated the number of books with this percent.
2. Filter the data with "linked_percent" greater than 90 and sort in descending order.
3. use 'geom_bar' to draw a bar graph with the x-axis representing the linked percent and the y-axis representing the number of books at that percent.

Conclusion:
These are my data processing for this assignment and my analysis of the visualization results. Through this assignment, I learned how to clean the data and learned more in-depth how to visualize the data and learned how to analyze the visualized data through Scatter Plot, Line Chart, Bar Chart, Box Plot and Pareto Chart.