V S N Sai Krishna Mohan Kocherlakota
Exploring Visualizations
10th October 2023

INTRODUCTION

In this project, I worked with two dataset. Primarily, I cleaned the data using the janitor package. I also used the lubridate package to change data types and create new columns. Secondly, I used basic to advanced functions that I have used in previous projects for data pre-processing, data cleaning, and analysis, such as glimpse and filter. I created multiple data frames and visualizations based on the book dataset, including box plots, scatter plots, and Pareto charts. This was a valuable learning experience in data analysis.

KEY-FINDINGS

Part-1: - Data Cleaning

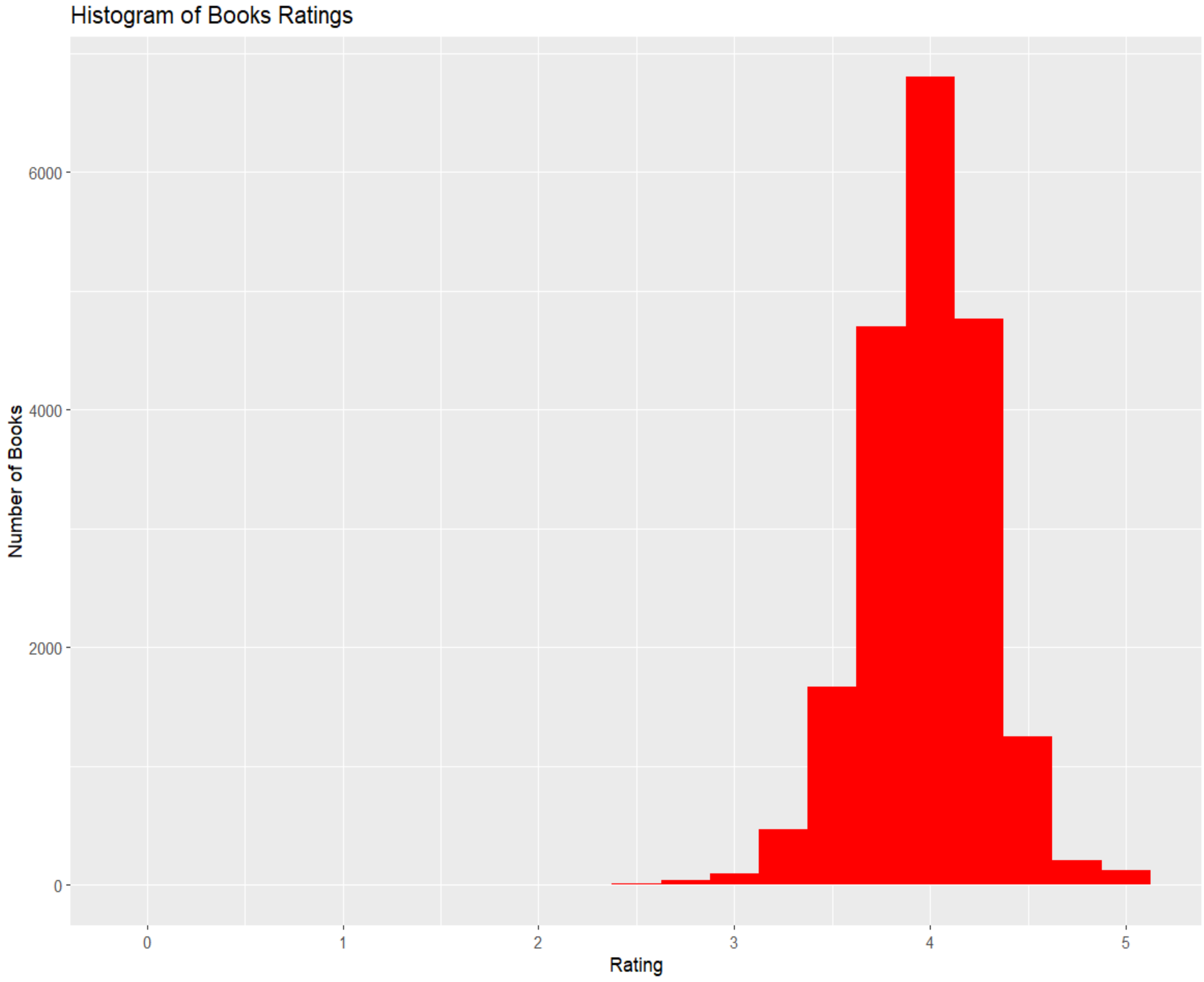
1.Problem 2 &3: -

In this code, Loaded the lubridate package, in which used the 'mdy' and 'year' functions. The lubridate package is used for data type conversion to dates and the creation of new columns individually, such as month, date, or year.

Part-2: - Data Analysis

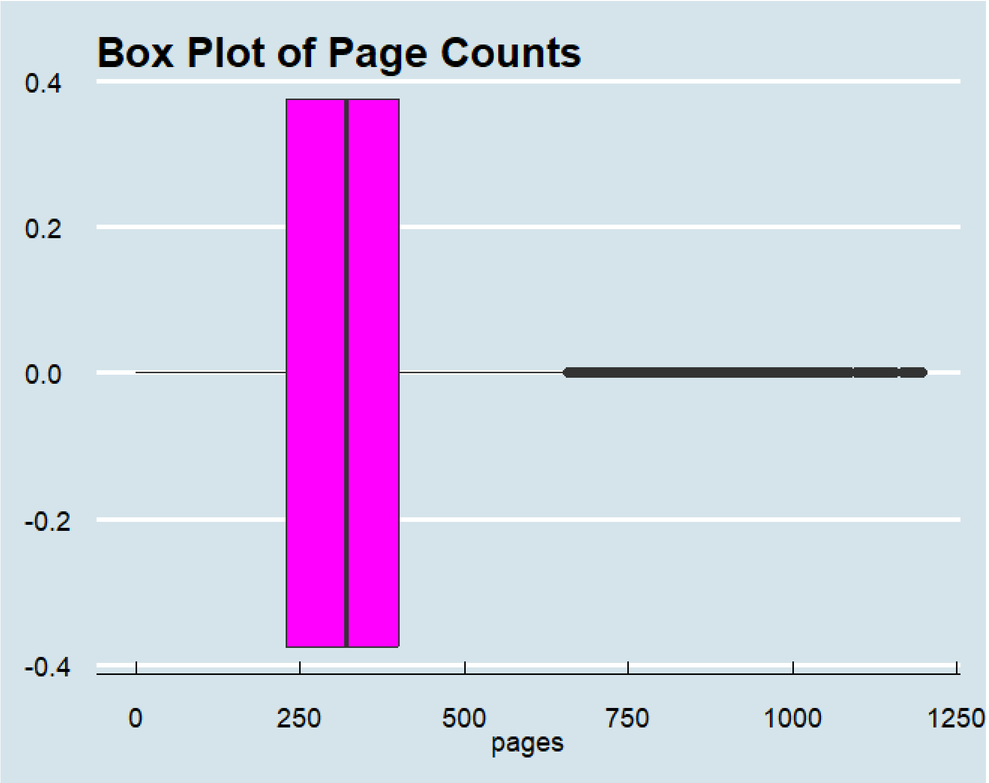
2.Problem 3: -

Based on the data frame 'books,' created a histogram plot. To enhance its visual appearance, applied a pre-defined function called 'theme_bw(),' which provides a black and white frame and background style to the histogram, resulting in a clean and monochromatic appearance.



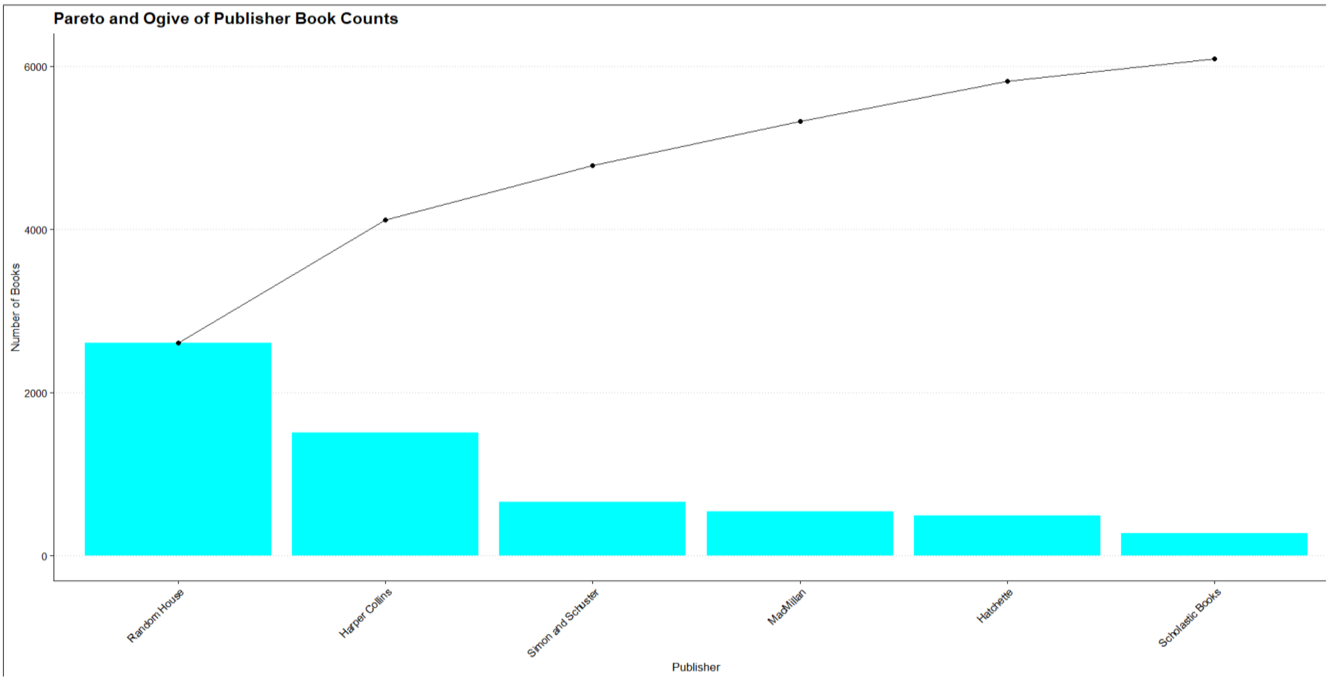
3.Problem 4: -

Created boxplot using ggplot and used theme_economist package. it helps create visualizations that resemble the aesthetics of The Economist's charts and graphs, making them suitable for conveying data-driven information in a clear and polished manner.



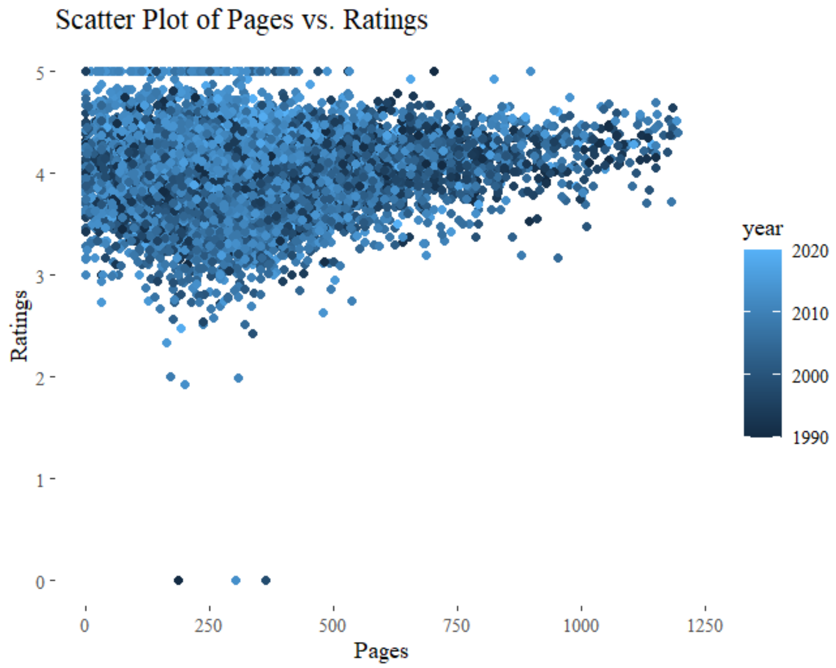
4. Problem 6: -

Created a pareto chart. Used ggplot and made a bar chart, line graph. After which combined and implemented theme_clean() for cleaning the background of the plot.

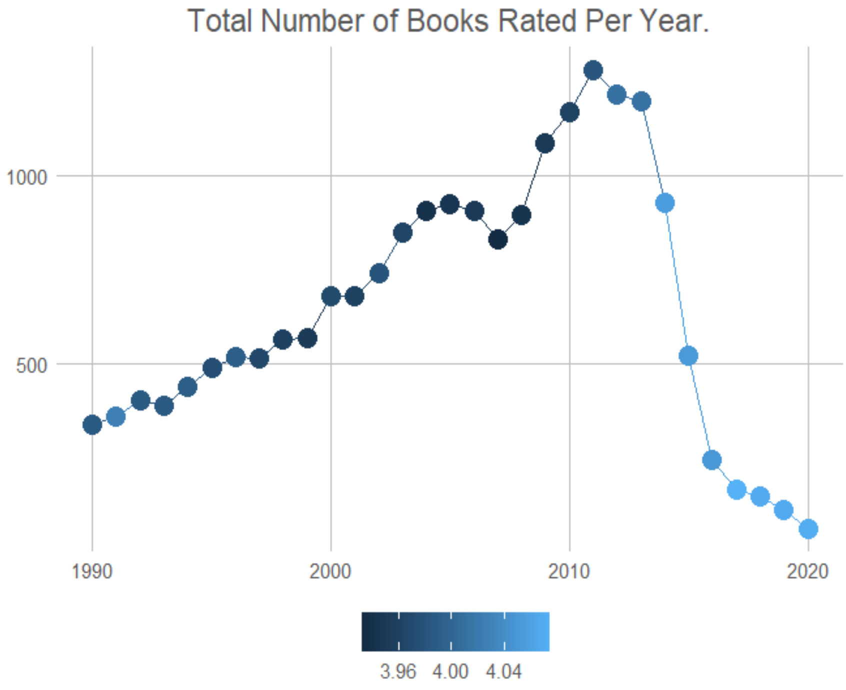


5. Problem 7: -

Created a scatter plot using the ggplot2 library and the 'geom_point()' function. Applied the 'theme_tufte()' package to enhance the understanding of the plot.

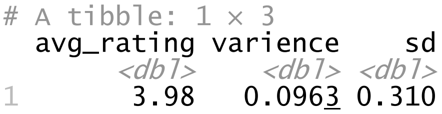


6. Problem 9: -

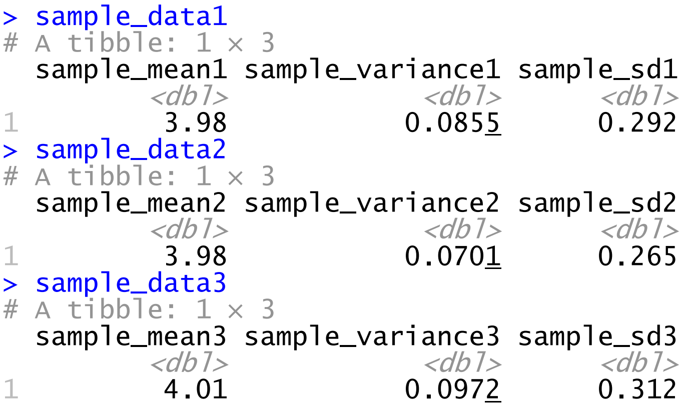


7. Problem 12: -

By checking the summary statistics, such as mean, variance, and standard deviation, of the 'books' data frame and the sample data we created, we observed that the means were the same in the first two cases. However, for the third sample data, there was a slight difference when compared to the 'books' data frame. On the other hand, when it comes to variance, it exhibited the opposite trend compared to the mean.



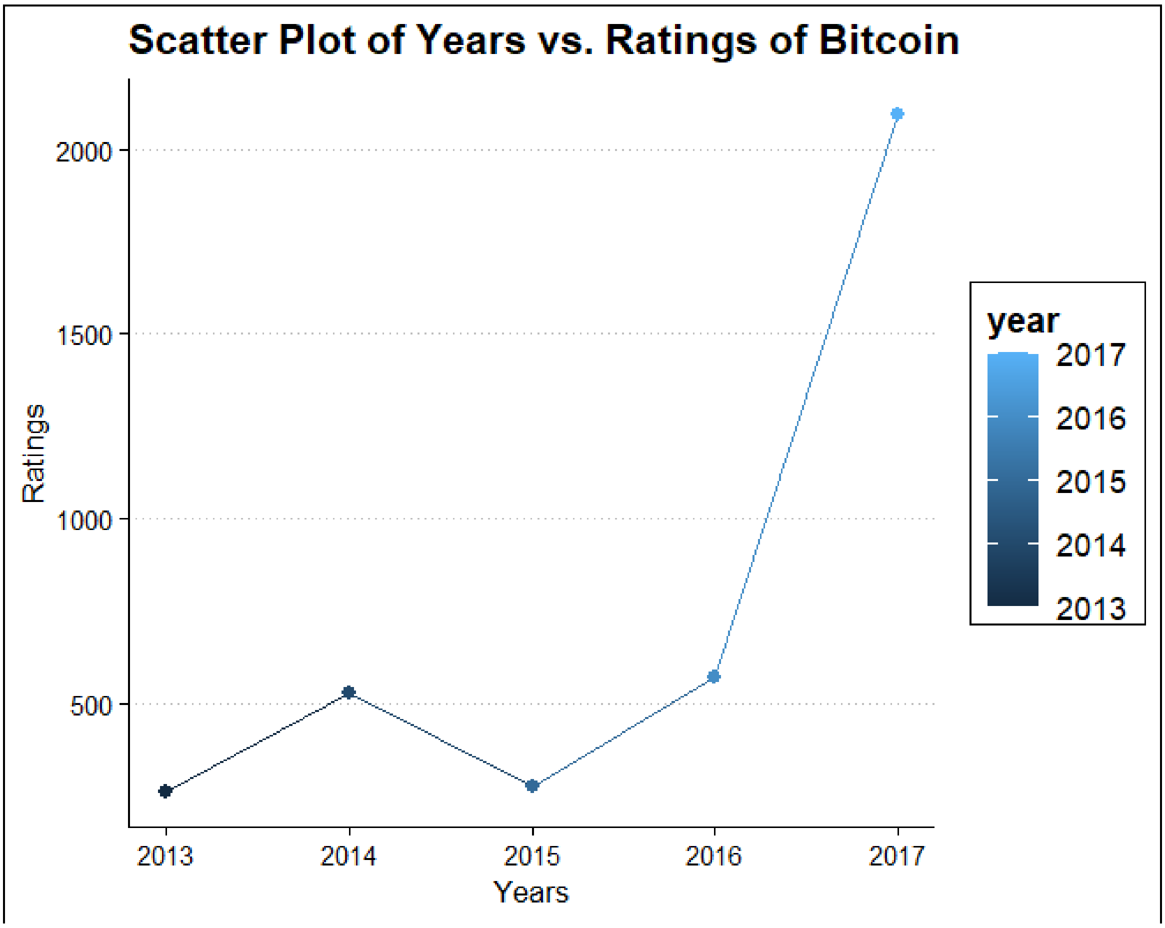
```
# A tibble: 1 × 3
  avg_rating varience    sd
       <dbl>    <dbl> <dbl>
1       3.98   0.0963 0.310
```



```
> sample_data1
# A tibble: 1 × 3
  sample_mean1 sample_variance1 sample_sd1
         <dbl>            <dbl>      <dbl>
1         3.98           0.0855      0.292
> sample_data2
# A tibble: 1 × 3
  sample_mean2 sample_variance2 sample_sd2
         <dbl>            <dbl>      <dbl>
1         3.98           0.0701      0.265
> sample_data3
# A tibble: 1 × 3
  sample_mean3 sample_variance3 sample_sd3
         <dbl>            <dbl>      <dbl>
1         4.01           0.0972      0.312
```

8. Problem 13: -

I have taken a new data frame and created a plot of bitcoin valuation per year.



CONCLUSION(Problem 14)

This analysis provides valuable insights into the 'Books' dataset, including price distributions, rating trends, and genre variations. The data visualizations and key statistics presented here serve as a foundation for decision-making in areas such as pricing strategies, genre selection, and marketing efforts. The findings emphasize the importance of considering book ratings and reviews as influential factors in book sales.

In conclusion, this project has been helpful for my journey of learning in the data world. I faced challenges that increased my enthusiasm and led me through various resources like Stack Overflow, online tutorials, and valuable lectures taught by my professor. The main learning from this project was related to different methods of data cleaning and the various types of charts or graphs that can be created using a single dataset. It was an eye-opening experience to realize what can be achieved with much larger datasets and how results can be summarized with the help of different functions and visualizations. I successfully passed all test cases in one go.

```
> testthat::test_file("project3_tests.R")
[ FAIL 0 | WARN 0 | SKIP 0 | PASS 14]
```

CITATIONS

1.**Error in spec : (df) : inherits(x,"tbl_df") is not true. (n.d.). Stack Overflow**
https://stackoverflow.com/questions/72695516/error-in-spec-df-inheritsx-tbl-df-is-not-true
2. **Using Lubridate in R Studio to create year, month, day columns gives unexpected results. (n.d.). Stack Overflow**
https://stackoverflow.com/questions/50965929/using-lubridate-in-r-studio-to-create-year-month-daycolumns-gives-unexpected
3. GeeksforGeeks. (2023). R Pareto chart. GeeksforGeeks
https://www.geeksforgeeks.org/r-pareto-chart/