ALY 6000

INTRO TO ANALYTICS

MODULE 3 PROJECT REPORT

PROFESSOR – ROY WADA

ITI ROHILLA

INTRODUCTION

Using R, a robust and sophisticated statistical programming language, we analyze the "books.csv" dataset and use data cleaning, visualization, and analytic techniques to extract useful information.

This analysis of two categories of primary objectives. Our primary objective is to completely understand the dataset, which includes details on book ratings, publication dates, page counts, publishers, and more. Second, we seek to draw meaningful conclusions from the data by highlighting patterns, discrepancies in book reviews, publisher distributions, release dates, and their connections.

KEY FINDINGS

Our analysis of the "books.csv" dataset yielded the following insights and conclusions:

Data Cleaning and Preparation: The dataset was carefully prepared and cleaned, with issues like missing values and incorrect date formats being fixed. The data preparation stage ensured that the data used in the analysis that followed was accurate and reliable.

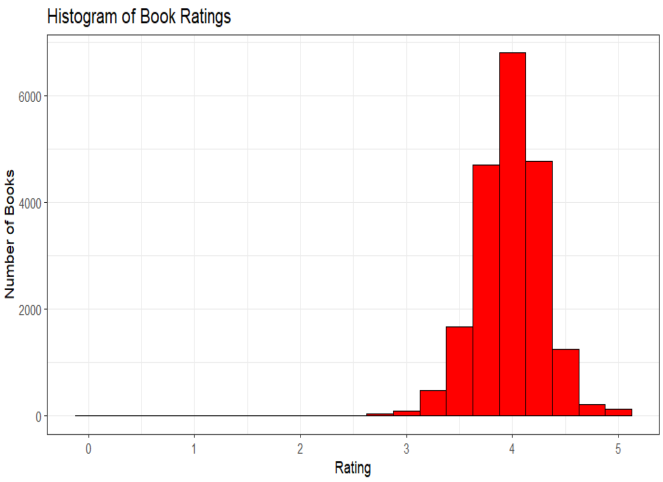
Pareto Chart: cumulative counts is used in a Pareto chart to show the distribution of books among the top publishers. With the help of this striking illustration, you can see how the majority of publications are controlled by a small number of major publishers, highlighting the 80/20 rule.

A scatter plot of pages vs. rating, color-coded by publication year, provides an eye-catching visual representation of how book ratings relate to their page counts and publishing years. Thanks to it, we can detect potential trends or outliers in the dataset.

.

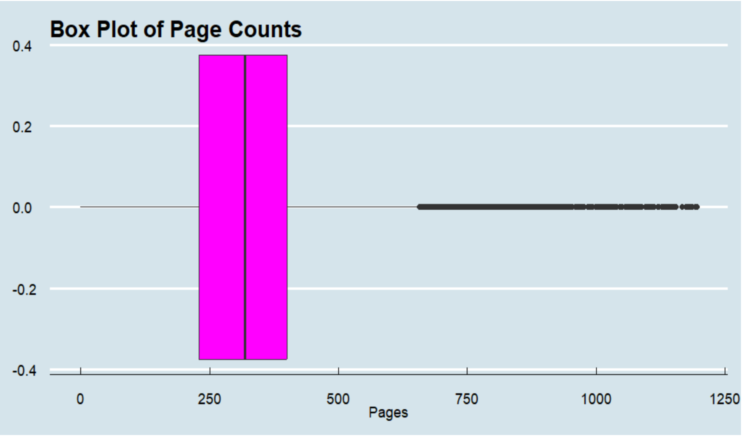
Rating Histogram - Data Analysis – problem number 3

We use the ggplot2 program to generate a histogram to show the distribution of book ratings. The number of books is indicated on the y-axis, and the rating values are shown on the x-axis, in this histogram, which shows the frequency of various rating values. By adjusting the plot's binwidth to 0.25, coloring the bars red, and using theme_bw to apply a simple, minimalistic theme, we can make it our own.



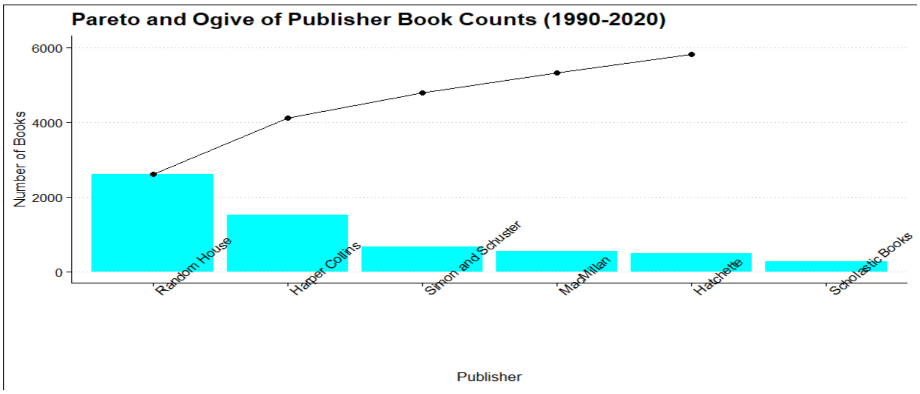
Box Plot of Page Counts – Data Analysis – problem number 4

```
> box_plot <- ggplot(data=book, aes(x=pages))+
+   geom_boxplot(fill="magenta", color="black")+
+   labs(title = "Box Plot of Page Counts",
+        x="Pages")+
+   theme_economist()
> box_plot
```



Pareto chart – Data Analysis – problem number 6

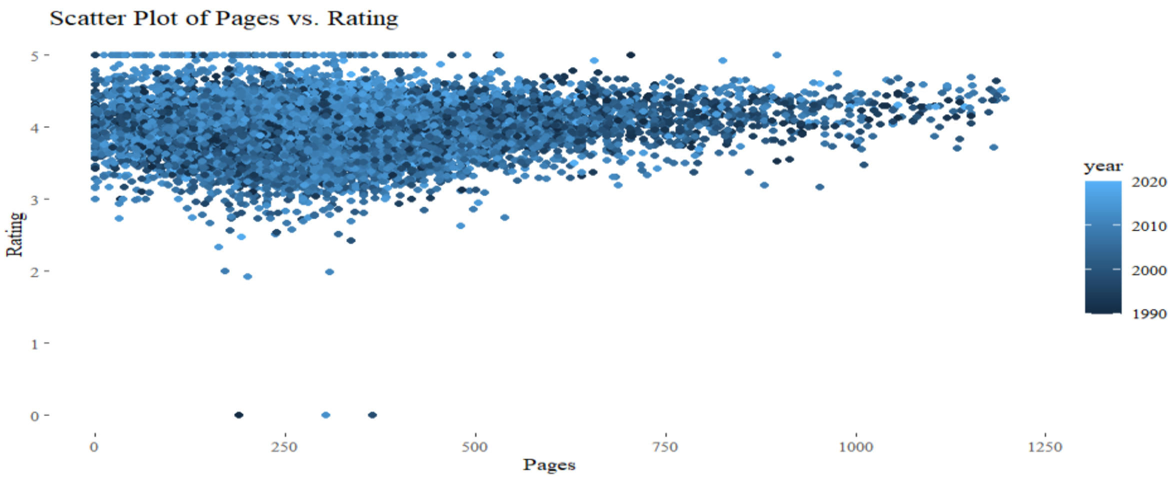
```
> pareto_chart <- ggplot(books_summary,aes(x = publisher,y = total_books))+

+   geom_bar(stat = "identity",fill = "cyan")+

+   geom_line(aes(x = publisher , y = cum_count),group=1,size = 0.5 )+

+   geom_point(aes(y=cum_count))+

+   labs(

+     x="Publisher",

+     y="Number of Books",

+     title = "Pareto and Ogive of Publisher Book Counts (1990-2020)"

+   )+

+   theme_clean()+

+   easy_rotate_x_labels(angle = 45)+

+   ylim(0,6000)

> pareto_chart
```



Scatter Plot – Data Analysis – problem number 7

```
> library(ggplot2)

> library(ggthemes)

>
```

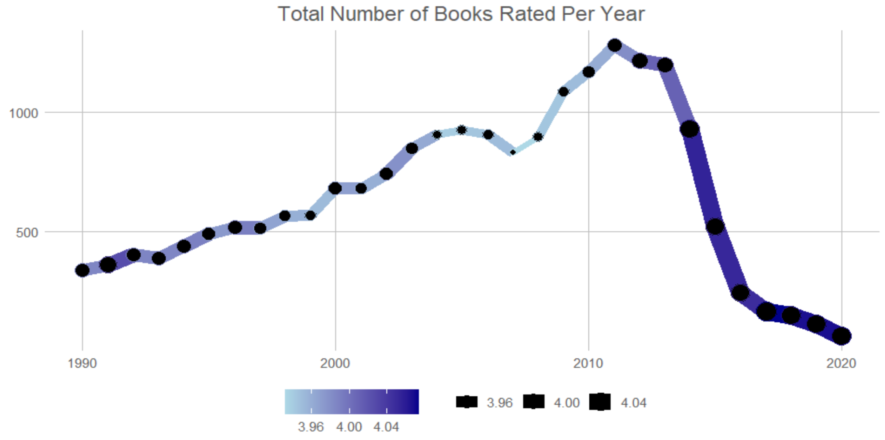
```
> scatter_plot <- ggplot(book, aes(x = pages, y = rating, color = year)) +

+   geom_point() +

+   labs(

+     x = "Pages",

+     y = "Rating",

+     title = "Scatter Plot of Pages vs. Rating"

+   ) +

+   theme_tufte()

> scatter_plot
```



Line plot – Data Analysis – problem number 9

```
> library(dplyr)

> library(ggplot2)

> library(ggthemes)

>

> color_palette <- c("lightblue", "darkblue")

>
```

```
> line_plot <- ggplot(by_year, aes(x = year, y = total_books, color = avg_rating, size = avg_rating)) +
+   geom_line() +
+   geom_point(shape = 21, fill = "black") +
+   labs(
+     x = "Year",
+     y = "Total Number of Books",
+     title = "Total Number of Books Rated Per Year"
+   ) +
+   scale_size_continuous(range = c(2, 6)) +
+   scale_color_gradientn(colors = color_palette) +
+   theme_excel_new()
> line_plot
```



Comparison with the population stats. -problem number 12

In Code 1 Direct calculations from the complete dataset are used to determine the mean, variance, and standard deviation of the book ratings. The book ratings are placed in a tibble with the name book_ratings To save these information, a Tibble called population_stats is created using the summarize() method.

In Code 2, three random samples of size 100 are created from the 'rating' column of the dataset using the sample() function.The mean, variance, and standard deviation are calculated for each of the three samples.

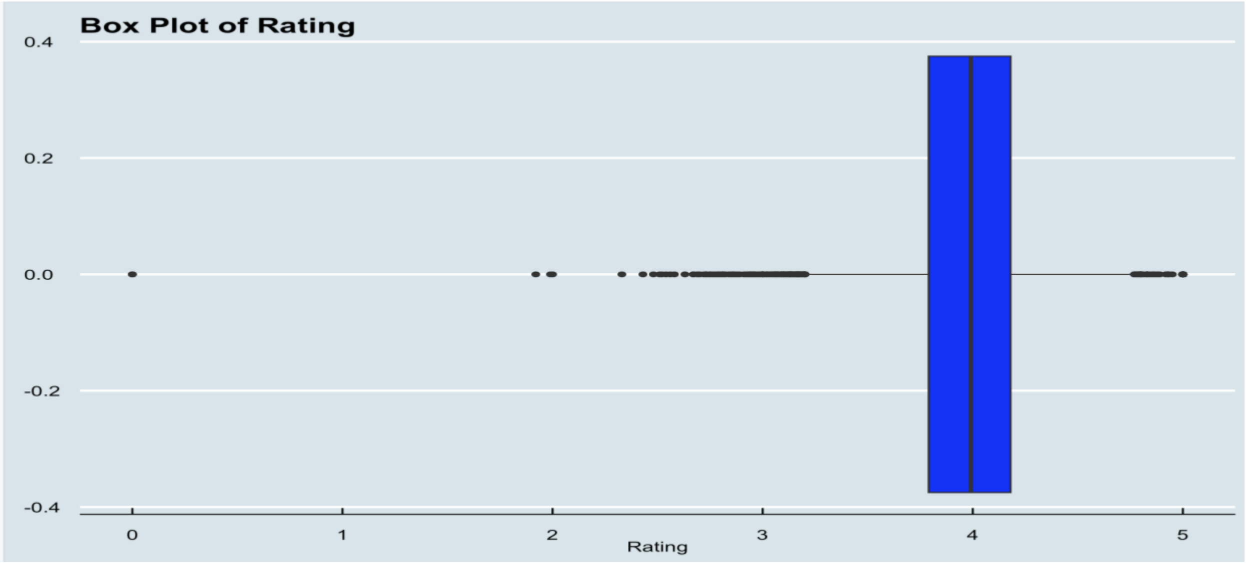
The main difference is that while Code 2 generates random samples and computes sample statistics for these samples, Code 1 derives population statistics directly from the complete dataset. We can evaluate how accurately the samples represent the population and the dependability of the sample statistics as estimators of the population statistics by contrasting the findings between Code 1 (population statistics) and Code 2 (sample statistics). Due to random sampling, it is normal for sample statistics to vary, but if the samples are representative, they should be near to population statistics.

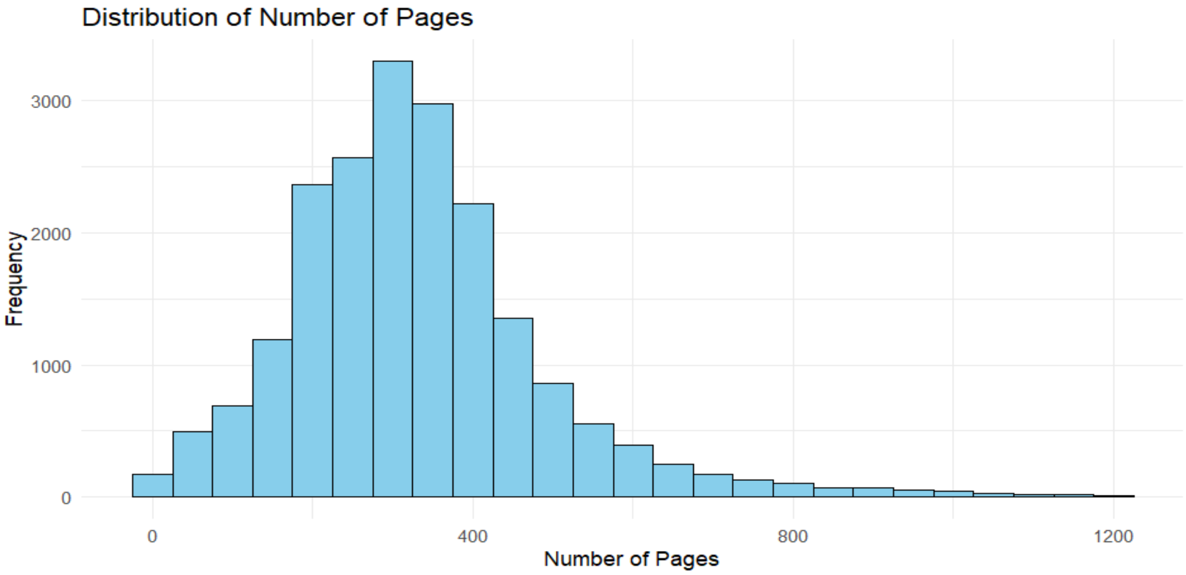
Additional Visualization – Data Analysis – problem number 13

First visualization is to create a box plot of book ratings. Box plots are excellent for visualizing the distribution and summary statistics of a dataset. In this case, it's helpful for understanding how book ratings are distributed. It can reveal information about the central tendency, spread, and the presence of outliers in the rating data.

On the other hand, second visualization part of the code is to create a histogram that shows the distribution of the number of pages in books. Histograms are valuable for understanding the frequency of values within a continuous variable. In this case, it helps to visualize how the number of pages is distributed across the books in the dataset.

Together, these visualizations offer a comprehensive view of the dataset, allowing us to explore relationships between variables, identify trends, and gain a better understanding of the characteristics of the books in the dataset.





SUMMARY

In Conclusion, the method of data cleaning and preparation made it simpler to analyze the dataset. The rating histogram and box plot of page counts, among other visualizations, offered data on book ratings and page length distribution. The Pareto Chart and publisher summary highlighted the various publications' contributions to the dataset. These conclusions lay the groundwork for a deeper analysis and comprehension of the book and publishing industries.

WORKS CITED

- https://www.youtube.com/watch?v=gXwI9W5wqjc
- https://www.youtube.com/watch?v=IX0QvffX9n0
- https://www.geeksforgeeks.org/calculate-the-average-variance-and-standard-deviation-in-r-programming/
- Dataquest https://stackoverflow.com/questions/30610584/writing-functions-in-r