Executive Summary Report

Yash Singh

ALY6000 “Introduction to Analysis”

Module 1 Project 3

Prof. John Wilder

2024 – 10 – 05

Introduction

The dataset analysed for this project was obtained from Goodreads and archived on Kaggle, containing information on over 52,000 books. It includes details such as book ratings, number of pages, publication years, and associated publishers. The project focused on cleaning, analyzing, and visualizing the data to uncover trends and relationships within the book publishing industry between the years 1990 and 2020.

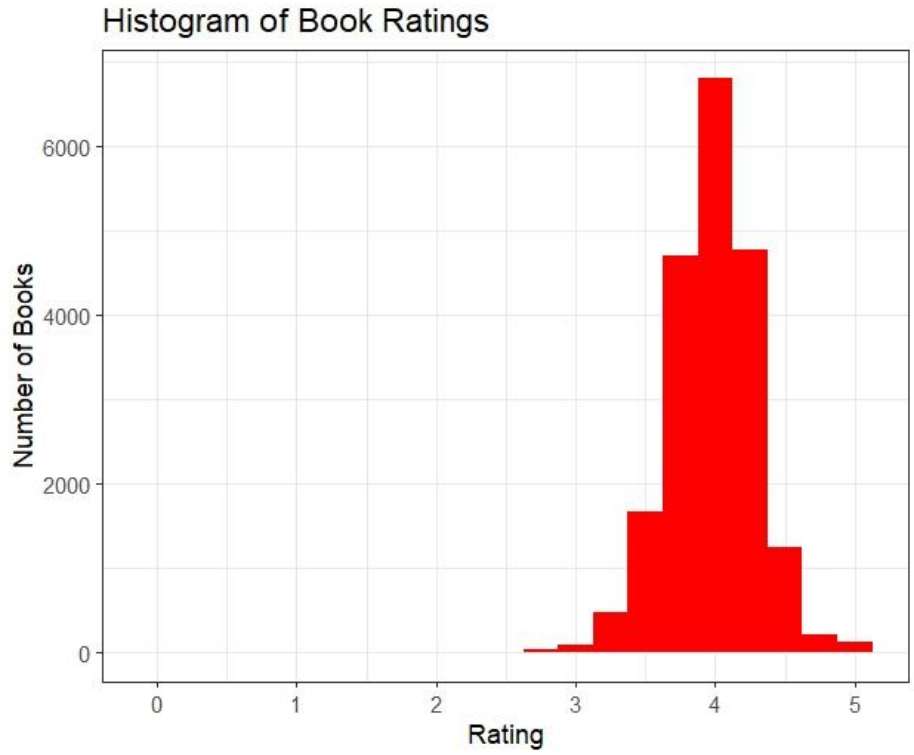
The goals of this project were to:

1. Explore the statistical properties of the data, focusing on measures such as mean, variance, and standard deviation.
2. Create visual representations that illustrate book ratings, the relationship between book length and user ratings, and publisher dominance in the market.
3. Compare population statistics with sample statistics derived from random samples within the dataset.
4. Derive actionable insights from the visualizations and statistical analysis to understand user preferences and publisher trends.

Through cleaning and filtering the dataset, we analysed books that were published between 1990 and 2020, focusing on those with fewer than 1200 pages. Various forms of visualizations were created, such as histograms, scatter plots, box plots, and Pareto charts, to illustrate trends in book ratings and the dominance of major publishers.

Key Findings

1. **Histogram of Book Ratings**
    - **Description**: This histogram illustrates the distribution of book ratings from the dataset, with most books having ratings between 3.5 and 4.5 stars.
    - **Key Takeaways**: The distribution is slightly skewed towards higher ratings, with a peak around 4 stars, indicating that users tend to rate books positively.

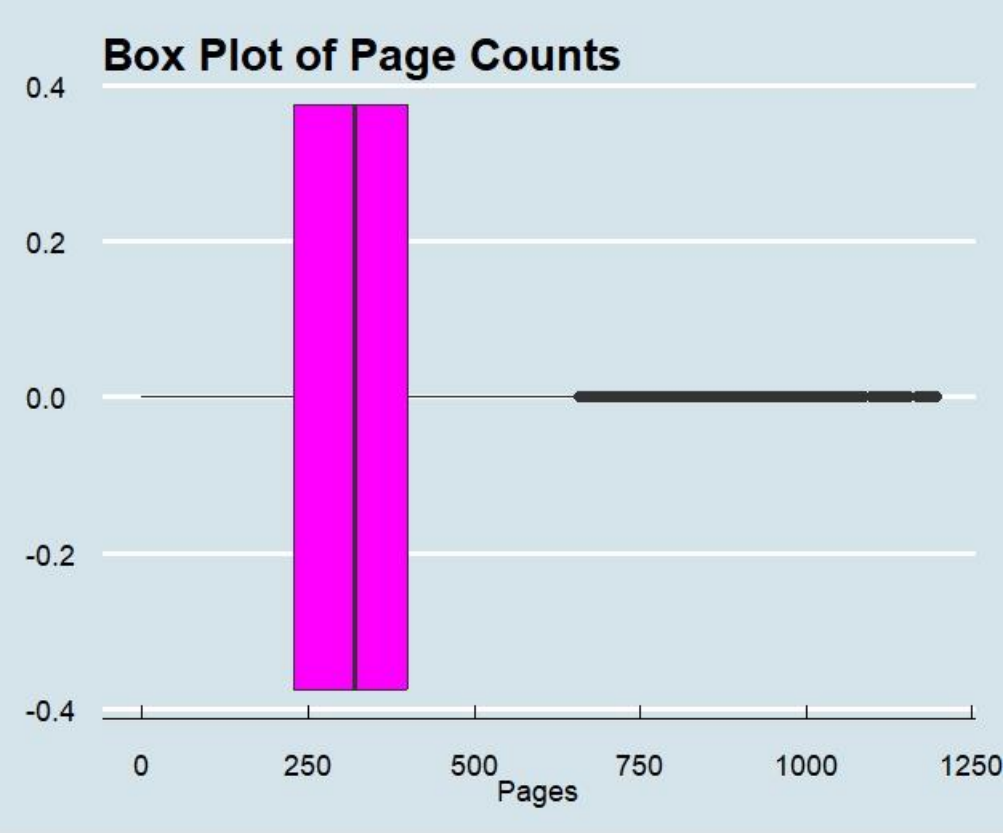


Histogram of Book Rating: Number of Books vs Rating

Visualization:

- The x-axis represents book ratings.
- The y-axis represents the number of books in each rating range.
- The histogram uses red coloring and a binwidth of 0.25 for clarity.

2. **Box Plot of Page Counts**
    - **Description**: A horizontal box plot was created to represent the distribution of page counts among the books.
    - **Key Takeaways**: The majority of books have fewer than 500 pages, with a few outliers reaching close to 1200 pages. This shows that shorter books are more common, but longer books also exist as outliers.

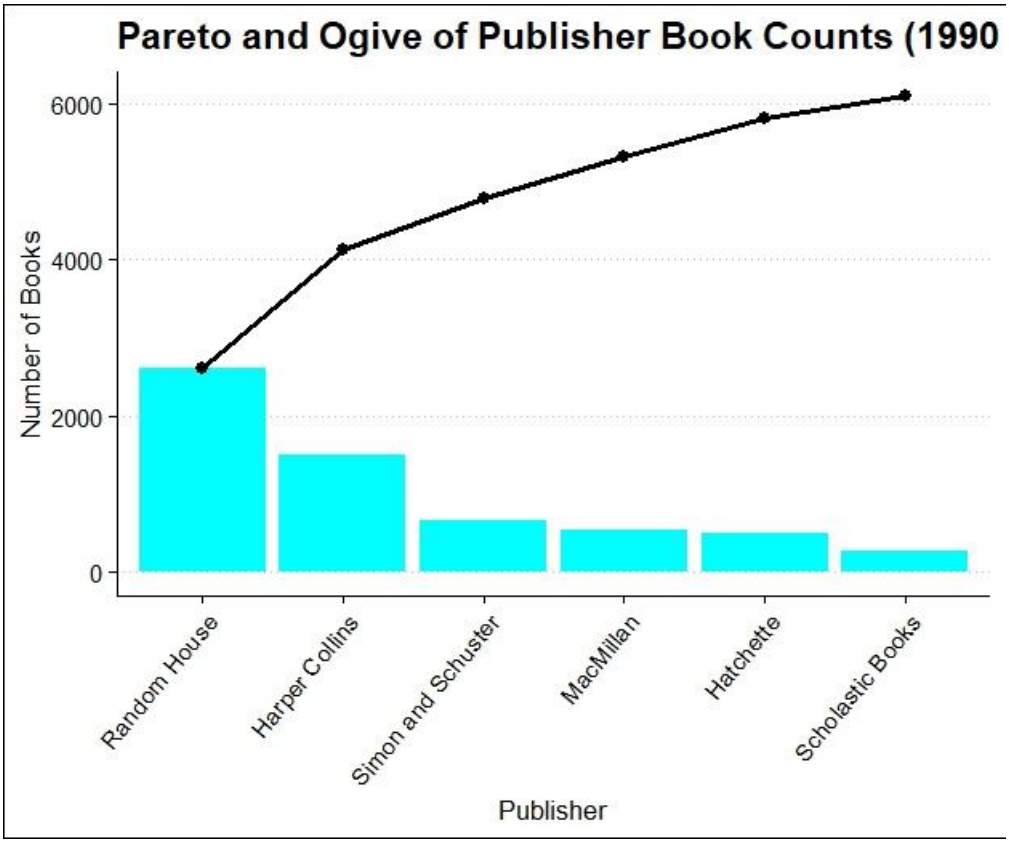


Box Plot of Page Counts

Visualization:

- The x-axis shows the number of pages.
- The box plot was filled with magenta color and follows the theme from the ggthemes package.

3. **Pareto Chart of Publishers**
    - **Description**: This Pareto chart visualizes the cumulative book counts of the top publishers between 1990 and 2020.
    - **Key Takeaways**: A small number of publishers, such as Random House and Harper Collins, dominate the industry, with these two publishers alone accounting for nearly 70% of all published books.

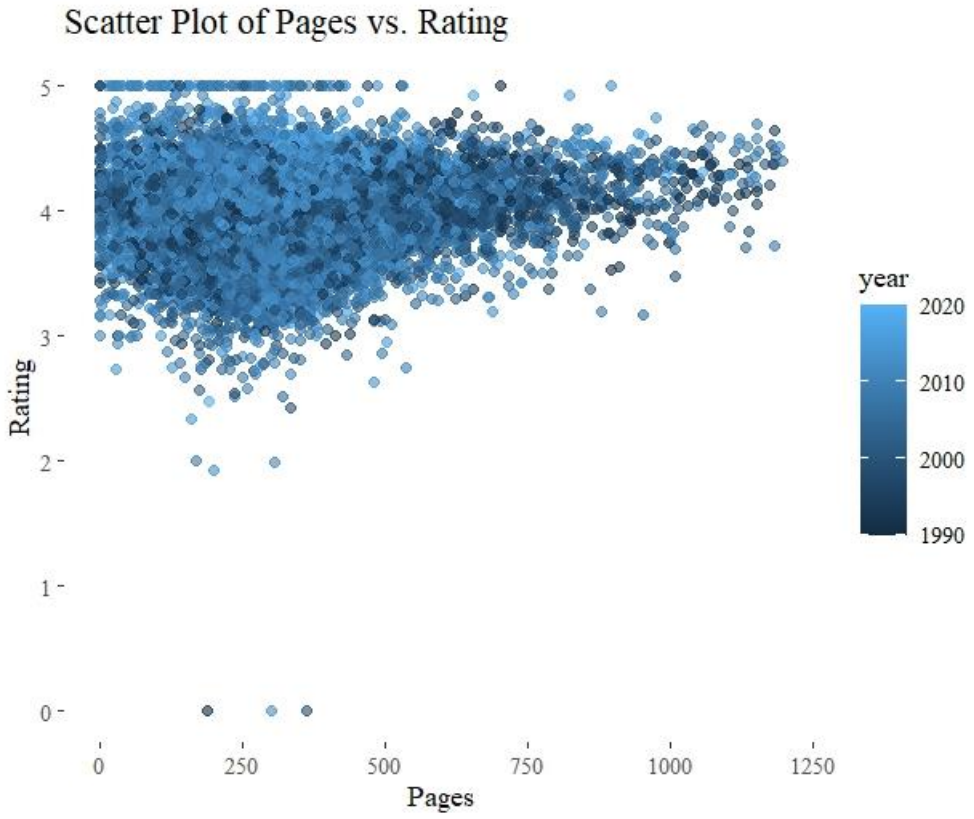


Pareto and Ogive of Publisher Book Counts (1990 - 2020)

Visualization:

- The x-axis lists publishers, rotated by 45 degrees for readability.
- The chart was filled with cyan and utilizes the theme_clean().

4. **Scatter Plot of Pages vs. Rating**
    - **Description**: A scatter plot was created to investigate the relationship between the number of pages and the book rating, with colors representing the year of publication.
    - **Key Takeaways**: There is no clear correlation between page length and book rating, but more recent books (published after 2010) tend to have higher ratings on average.

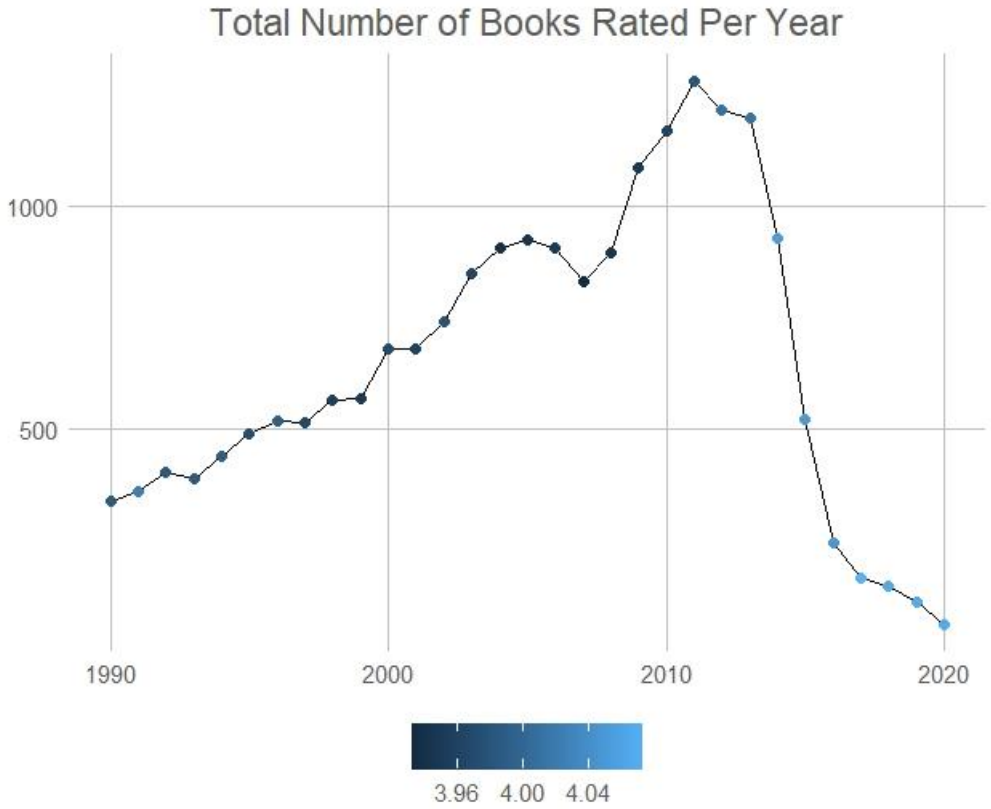


Scatter Plot of Pages vs. Rating

Visualization:

- The x-axis shows the number of pages.
- The y-axis shows the book ratings, with the year represented by color.

5. **Line Plot of Total Books Rated Per Year**
    - **Description**: This line plot tracks the number of books rated per year, with points colored based on the average rating for each year.
    - **Key Takeaways**: The number of books rated each year has grown steadily, particularly in the last decade, while the average ratings have remained stable around 4 stars.



Total Number of Books Rated Per Year

Visualization:

- The plot uses the theme_excel_new() theme and emphasizes the steady increase in book ratings over time.

6. **Population vs. Sample Statistics**
    - **Description**: The average rating, variance, and standard deviation were calculated for the entire dataset, and these values were compared with those derived from three random samples of 100 books each.
    - **Key Takeaways**: The sample statistics closely match the population statistics, confirming that the samples are representative of the overall dataset.

```
> sample_stats
  mean variance   sd        sample
1 3.98     0.10 0.31 Books Ratings
2 4.01     0.08 0.28      Sample 1
3 3.95     0.09 0.30      Sample 2
4 4.02     0.09 0.29      Sample 3
```

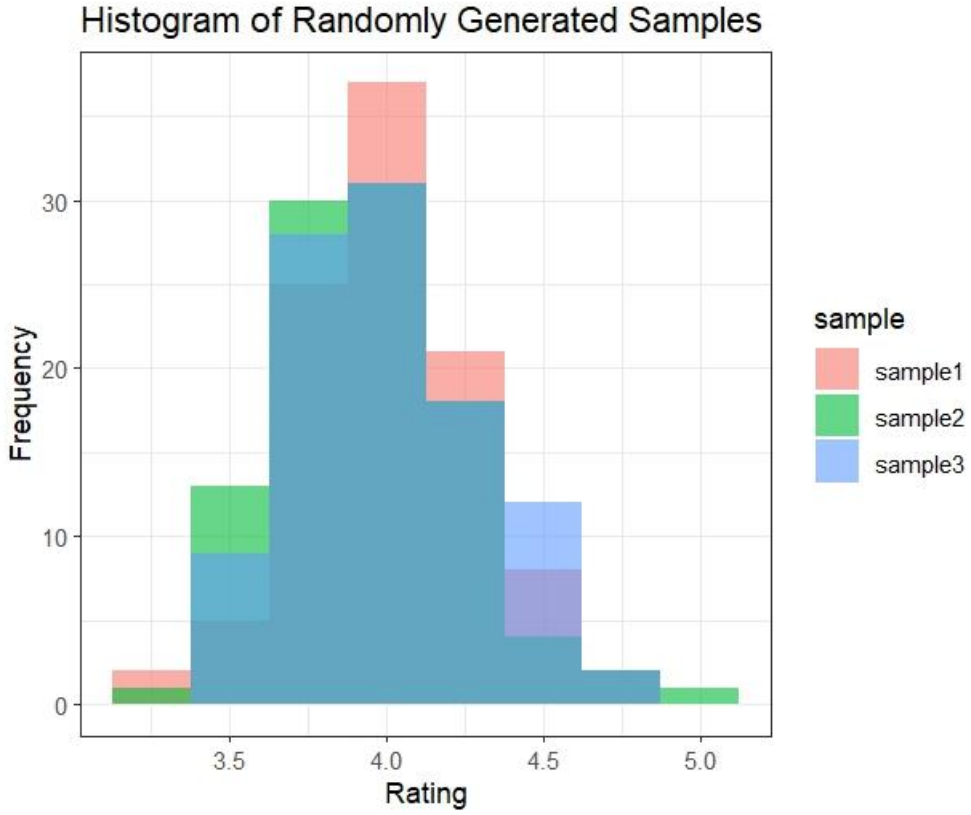
Population vs Sample Comparison

Population vs. Sample Comparison:

- **Population Mean**: 3.98
- **Sample Mean (Sample 1)**: 4.01
- **Sample Mean (Sample 2)**: 3.95
- **Sample Mean (Sample 3)**: 4.02

7. **Additional Visualization**

   - **Description**: A heatmap was created to further explore the relationship between book pages and ratings. The heatmap highlights the density of books in specific page and rating ranges, with a gradient color scale representing density.
   - **Key Takeaways**: The heatmap confirms that books tend to have ratings between 3.5 and 4.5, regardless of page length. This visualization provides additional evidence that page length does not significantly affect book ratings.



Histogram of Randomly Generated Samples

Conclusion

From the analysis, several important trends were identified. Most notably, book ratings on Goodreads are skewed towards positive reviews, with an average rating of 4.0. Additionally, while page length does not appear to influence ratings, certain publishers dominate the book market, particularly Random House and Harper Collins.

Recommendations:

1. **Focus on maintaining high user ratings**: Publishers should prioritize maintaining high user engagement and favorable reviews, as ratings are a crucial driver of visibility and user interest on platforms like Goodreads.
2. **Consider strategic publishing partnerships**: Smaller publishers may find it advantageous to form partnerships or focus on niche genres to compete with industry giants like Random House.
3. **Further explore genre-specific trends**: Future analysis could explore the relationship between book genres and ratings or examine how user preferences have evolved in terms of book length and subject matter over time.

Works Cited

- Goodreads. (n.d.). *Books dataset*. Retrieved from http://www.goodreads.com

Appendix

```
#This is the script for Project 3

#Author: Yash S

#Created On: 2024-09-30

#Last Edited: 2024-10-

#Class: ALY6000


cat("\014") # clears console

rm(list = ls()) # clears global environment

try(dev.off(dev.list()["RStudioGD"]), silent = TRUE) # clears plots

try(p_unload(p_loaded(), character.only = TRUE), silent = TRUE) #clears
packages

options(scipen = 100) # disables scientific notation for entire R session


library(pacman)

p_load(tidyverse)

p_load(janitor)

p_load(lubridate)

p_load(ggthemes)

p_load(ggeasy)

p_load(tibble)

p_load(testthat)


#1

#Loads the data set

books <- read_csv("books.csv")
```

```
#Cleaning the data set

#1
#Cleans the name of the columns
books <-clean_names(books)

#2
books <- books %>%
  mutate(first_publish_date = mdy(first_publish_date))

#3
books <- books %>%
  mutate(year = year(first_publish_date))

#4
books <- books %>%
  filter(year >= 1990 & year <= 2020)

#5
books <- books %>%
  select(-publish_date, -edition, -characters, -price, -genres, -setting, -isbn)


#6
books <- books %>%
```

```
  filter(pages<1200)

#Data Analysis

#1
glimpse(books)

#2
summary(books)

#3
ggplot(books, aes(x = rating)) +
  geom_histogram(binwidth = 0.25, fill = "Red") +
  labs(x= "Rating", y="Number of Books", title = "Histogram of Book Ratings") +
  theme_bw()

#4
ggplot(books, aes(x = pages)) +
  geom_boxplot(fill = "Magenta") +
  labs(x = "Pages", title = "Box Plot of Page Counts") +
  theme_economist()

#5
book_publishers <- books %>%
  group_by(publisher) %>%
  summarize(total_books = n())%>%
```

```
  filter(!is.na(publisher) & total_books>250) %>%
  arrange(desc(total_books)) %>%
  mutate(
    publisher = factor(publisher, levels = publisher),
    cum_count = cumsum(total_books),
    rel_freq = total_books/ sum(total_books),
    cum_freq = cumsum(rel_freq)
    )

#6
ggplot(book_publishers, aes(x = publisher, y = total_books)) +
  geom_bar(fill = "Cyan", stat = "identity") +
  geom_line(aes(y=cum_count), color = "black",group=1,  size = 1) +
  geom_point(aes(y=cum_count), color = "black", size = 2)+
  labs(x="Publisher", y="Number of Books", title = "Pareto and Ogive of
Publisher Book Counts (1990 - 2020)")+
  theme_clean()+
  theme(axis.text.x = element_text(angle = 50, hjust = 1))

#7
ggplot(books, aes(x=pages, y=rating, colour = year))+
  geom_point(alpha = 0.6)+
  labs(x="Pages", y="Rating", title = "Scatter Plot of Pages vs. Rating")+
  theme_tufte()

#8
by_year <- books %>%
```

```
  group_by(year) %>%
  summarise(total_books = n(), avg_rating = mean(rating, na.rm = TRUE))

#9
ggplot(by_year, aes(x = year, y = total_books)) +
  geom_line() +
  geom_point(aes(color = avg_rating)) +
  labs(title = "Total Number of Books Rated Per Year", x = "Year", y = "Number of
Books") +
  theme_excel_new()

#10
average <- function(x){
  n <- length(x[!is.na(x)])
  return(sum(x, na.rm = TRUE)/n)
}

pop_var <- function(x){
  avg <- average(x)
  n <- length(x[!is.na(x)])
  return(sum((x[!is.na(x)] - avg)^2) / n)
}

sd_var <- function(x){
  return(sqrt(pop_var(x)))
}
```

```
#11
n <- length(books$rating)
books_rating <- tibble(
  avg_rating = average(books$rating),
  variance = (pop_var(books$rating) * (n - 1) / n),
  sd = sd_var(books$rating)
)

#12
#Population statistics from the actual data
population_mean <- mean(books$rating, na.rm = TRUE)
population_sd <- sd(books$rating, na.rm = TRUE)

#Create three samples of size 100 using rnorm with population statistics
set.seed(123)  # For reproducibility
sample1 <- rnorm(100, mean = population_mean, sd = population_sd)
sample2 <- rnorm(100, mean = population_mean, sd = population_sd)
sample3 <- rnorm(100, mean = population_mean, sd = population_sd)

#Create a tibble to store the sample data
samples <- tibble(
  sample1 = sample1,
  sample2 = sample2,
  sample3 = sample3
)
samples
```

```
#Function to compute sample statistics (mean, variance, sd)
sample_statistics <- function(sample) {
  avg <- round(mean(sample), 2)
  var <- round(var(sample), 2)
  sd <- round(sd(sample), 2)
  return(data.frame(mean = avg, variance = var, sd = sd))
}

#Compute statistics for each sample
stats_sample1 <- sample_statistics(sample1)
stats_sample2 <- sample_statistics(sample2)
stats_sample3 <- sample_statistics(sample3)
og_sample <- sample_statistics(books$rating)

#Combine the statistics into a single data frame
sample_stats <- bind_rows(
  og_sample %>% mutate(sample = "Books Ratings"),
  stats_sample1 %>% mutate(sample = "Sample 1"),
  stats_sample2 %>% mutate(sample = "Sample 2"),
  stats_sample3 %>% mutate(sample = "Sample 3")
)
#Print sample statistics
sample_stats

#13
```

```
#Convert samples to long format for ggplot
samples_long <- samples %>%
  pivot_longer(cols = everything(), names_to = "sample", values_to = "rating")

# Create histogram for each sample
ggplot(samples_long, aes(x = rating, fill = sample)) +
  geom_histogram(binwidth = 0.25, alpha = 0.6, position = "identity") +
  labs(title = "Histogram of Randomly Generated Samples", x = "Rating", y =
"Frequency") +
  theme_bw()

testthat::test_file("project3_tests.R")
```