AC50002R

Programming Assignment

SMITA SANJAY PABLE (2543320)

**PART 1: Most popular names**

This assignment assumes you have worked through the module material on R using the tidyverse.

PART 1: Most popular namesLoad the babynames tibble. Use dplyr to find out which were the three most popular names given tobabies of male sex during the 20th century (1900-1999) in the USA. Use ggplot2 to produce a singlegraph that shows how the popularity these three names changed over the course of the century.You should submit a one-page document containing your R code and your graph. Include commentsin your code explaining what each line does.

**R Code:**

# Install necessary packages if not already installed

install.packages("tidyverse")

install.packages("babynames")

install.packages("ggplot2")

install.packages("dplyr")

install.packages("tibble")

# Load required libraries

library(ggplot2)

library(tibble)

library(babynames)

library(dplyr)

# Filter babynames dataset for popular male names born between 1900 and 1999

popular\_names <- dplyr::filter(babynames, sex == "M" & year >= 1900 & year <= 1999)

View(popular\_names)

# Calculate the total occurrences of each name

name\_counts <- popular\_names %>%

group\_by(name) %>%

summarize(total = sum(n)) %>%

arrange(desc(total)) %>%

head(3)

View(name\_counts)

# Filter babynames dataset for the three most popular male names

filtered\_data <- dplyr::filter(babynames, sex == "M" & name %in% name\_counts$name & year >= 1900 & year <= 1999)

# Create a line plot to show the popularity of the three most popular male names over time

ggplot(filtered\_data, aes(x = year, y = n, color = name)) +

geom\_line() +

labs(

title = "Popularity of the Three Most Popular Male Names (1900-1999)",

x = "Year",

y = "Number of Births",

color = "Name"

) +

theme\_minimal()

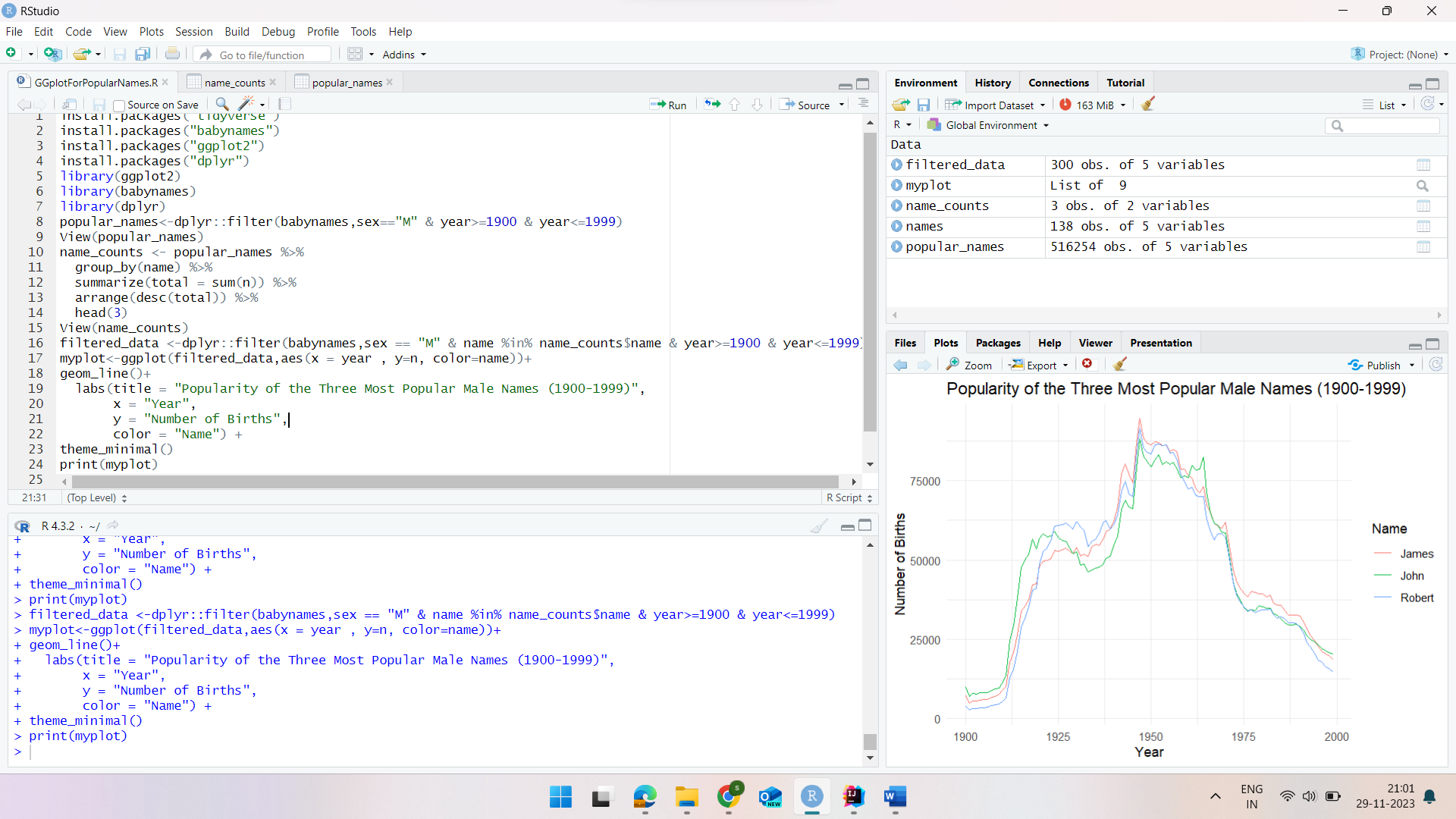


Figure :Execusion of Code and Output

A white background with a black border

Description automatically generated with medium confidence

Figure : Extracted Most Popular 3 Male names(1900-1999)

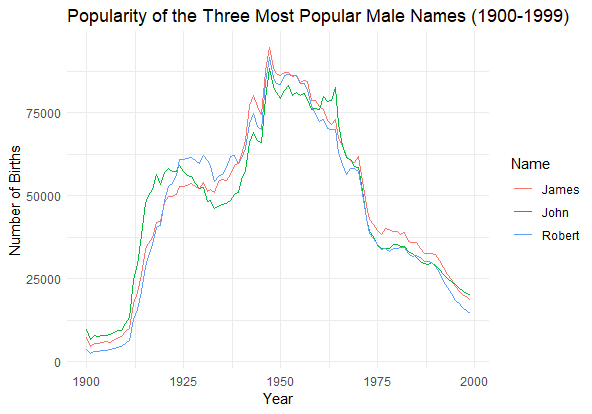


Figure : popularity of three most popular boys' names born between 1900 and 1999.

**Explanation:**

In order to identify and plot the trend in popularity of three most popular boys' names born between 1900 and 1999, provided R code uses a “babynames” dataset. *Figure 2*. Shows filtered out top three names and the total count for each name, based on popularity. In conclusion, it filters the dataset for these three names and creates a line graph as shown in *Figure 3*. to display each selected name's birth-rate every year. Where, each line indicates a different name(those are **James, John and Robert**), with x-axis showing years and y-axis indicating the number of births.

**PART 2: Data exploration**

**Amazon Kindle Books Dataset 2023 (130K Books):**

In this assignment, I choose a dataset that contains information on **Amazon Kindle Books** and their parameters, such as book title, authors, ratings, pricing, and sales statistics. Its goal is to give insights into the e-book industry's dynamics and to assist business choices or marketing tactics. This dataset was downloaded from Kaggle.com, the URL to which is provided below,

<https://www.kaggle.com/datasets/asaniczka/amazon-kindle-books-dataset-2023-130k-books/>

**About Dataset:**

A huge dataset including 130K Amazon Kindle books. Book titles, authors, ratings, price, and sales information are provided through October 2023. The dataset's columns are listed below.

The column names in the `kindle\_data` dataset:

1. `asin`: Product ID from Amazon. (type: str)
2. `title`: Book title. (type: str)
3. `author`: Author of the book (type: str)
4. `soldBy`: Publisher or seller of the book (type: str)
5. `imgURL`: URL of the book cover image. (type: str)
6. `productURL`: URL of the book. (type: str)
7. `stars`: Average star rating of the book (type: float)
8. `reviews`: Number of reviews the book has received (type: int)
9. `price`: Price of the book (type: float)
10. `isKindleUnlimited`: Indicates if the book is available on Kindle Unlimited ("True" or "False") (type: bool)
11. `category\_id`: Serial id assigned to the category this book belong to. (type: int)
12. `isBestSeller`: Indicates if the book is a bestseller ("True" or "False") (type: bool)
13. `isEditorsPick`: Indicates if the book is an editor's pick ("True" or "False") (type: bool)
14. `isGoodReadsChoice`: Indicates if the book is a Goodreads Choice winner ("True" or "False") (type: bool)
15. `publishedDate`: Date when the book was published (type: str, format: YYYY-MM-DD)
16. `category\_name`: Category or genre of the book (type: str)

To visualize this data, I have used the **RStudio** and the libraries I have used to do that are,

tidyverse: A collection of packages for data manipulation and visualization.

ggplot2: A popular package for creating visualizations in R.

dplyr: A package for data manipulation and summarization.

Scales: Provides functions for formatting and scaling data

Tibble: Uses for creating and working with data frames

**1.Simple Distribution of Stars**

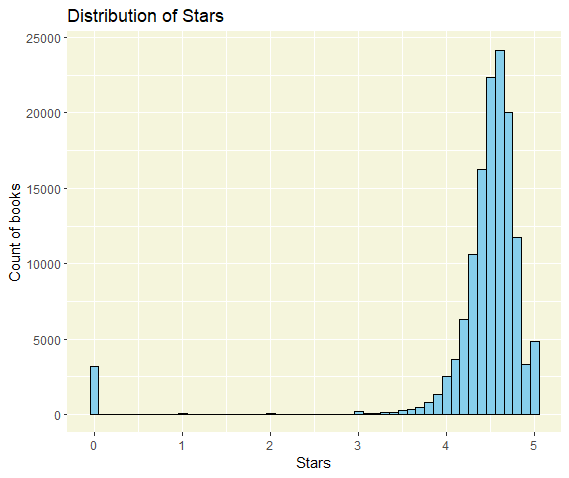


Figure : Simple Distribution of Stars.

**Explanation:**

The *Figure 4* titled "Distribution of Stars" histogram visualizes the spread of star ratings across all books. Where, x-axis represents the number of stars and y-axis represent count of the books. With a bin width of 0.1, the graph displays the count of books within each star rating range. Interestingly, most of the books available on **Amazon Kindle** are rated between **4 to 5 Stars**. That means Amazon Kindle is a great place to read a book.

**2. Top 5 Most Expensive Books and Their Publishers**

A screenshot of a computer

Description automatically generated

Figure : Extracted 5 Most Expensive Books and Their Publishers

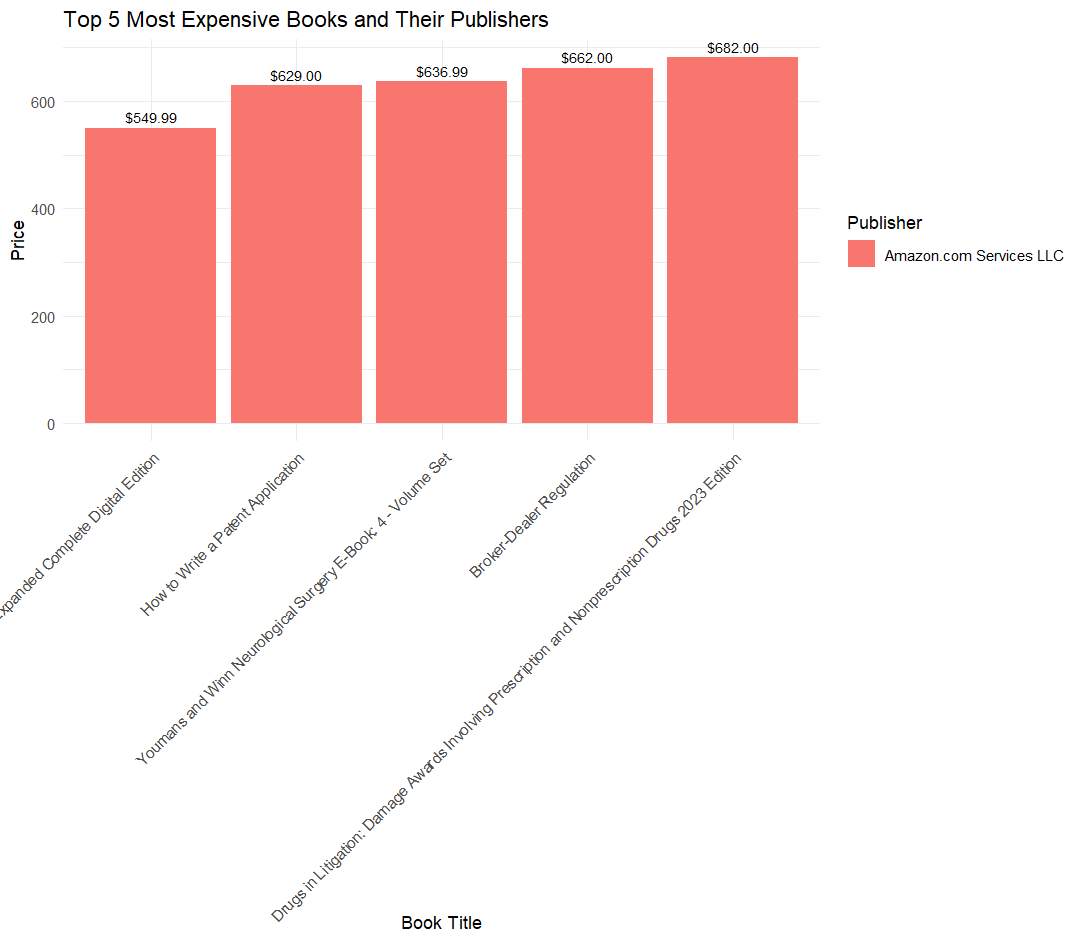


Figure : Plot to display top 5 expensive books and publisher.

**Explanation:**

Figure 2. shows the top five most expensive books and their publisher (soldBy) of all time from “kindle\_data”, ranking them by their price in descending order , Then the Figure 3. represents a bar chart based on ggplot2 will be used to illustrate the chosen books. The x-axis shows the book titles to be reordered by price, with the y-axis showing the prices of those books. Each bar is filled with the publisher (`soldBy`), and text annotations display the price above each bar. The plot is given a title and axis labels for clarity, with minimal theme adjustments, including rotating x-axis labels for improved readability. A visual representation of the top five highest priced books and their publishers is given in this graph. It is visual that top 5 expensive books are published by “**Amazon.com Services LLC**” , and most expensive book is “**Drugs in Litigation: Damage Awards Involving Prescription and Nonprescription Drugs 2023 Edition**”.

**3.Top 10 Bestselling Categories and Their Average Prices:**

A screenshot of a computer

Description automatically generated

Figure : Extracted Top 10 Bestselling Categories and Their Average Prices

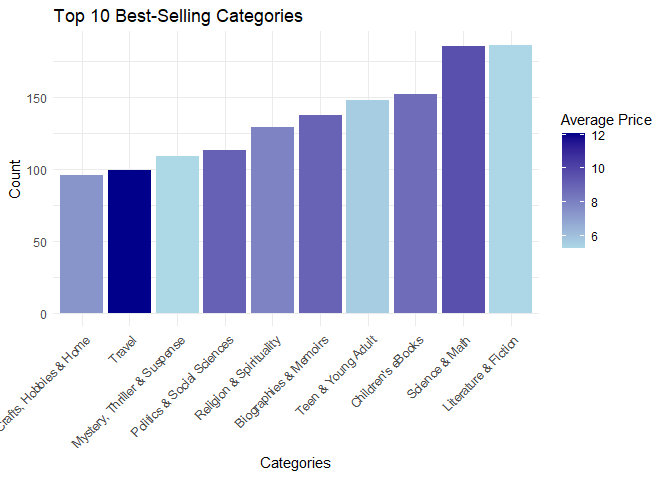


Figure : Top 10 Best-Selling Categories and its avg Price

**Explanation:**

*Figure 7* displays the top 10 Best Selling Categories, its average price and its total book count extracted from the dataset. As shown in the *figure 8* , the resultant bar plot use ggplot, with categories(bestselling = True) on the x-axis, book count on the y-axis, and bars coloured based on average price. The narrative digs into the popularity and pricing of many e-book genres. The graph states that most expensive bestselling category is “**Travel**” whose **average price is 12** with **book count 99**. Where**, Science & Math and Literature & Fiction** are **best selling categories** with **highest book count** whose price is also **not that expensive.**

**4.Kindle Unlimited Impact:**

Kindle Unlimited’s impact on book’s popularity.

A screenshot of a computer

Description automatically generated

Figure : Extracted` isKindleUnlimited` status and its avg reviews.

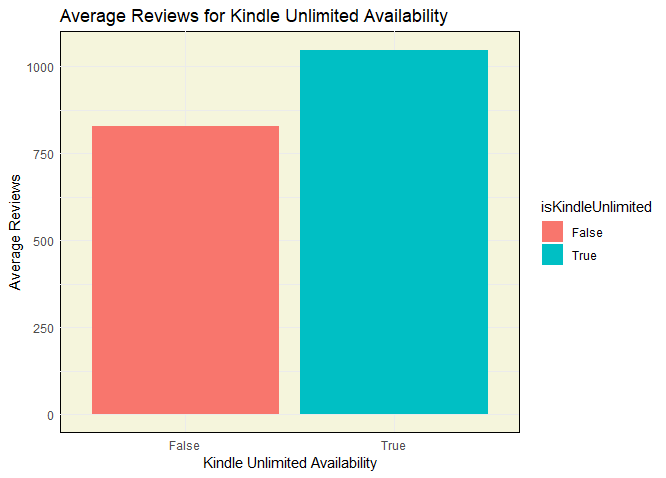


Figure : Average Reviews for Kindle Unlimited Availability

**Explanation:**

*Figure 9* shows the extracted data gives Kindle Unlimited availability status on average book reviews. then (as shown in *figure 10*) using ggplot to generate a bar plot comparing the average reviews for books with and without Kindle Unlimited access. In which x-axis shows the status of Kindle Unlimited availability that is True or False and y-axis represents the average count of reviews. Here the title, axis labels, and legend provide context to interpret the results. From above graph it is pretty evident that **Kindle Unlimited available** books are more **popular** and has a **greater number of average reviews**.

**5.Identify the most popular authors based on their number of published books and ratings.**

A screenshot of a computer

Description automatically generated

Figure : Top Authors with their total books and average rating

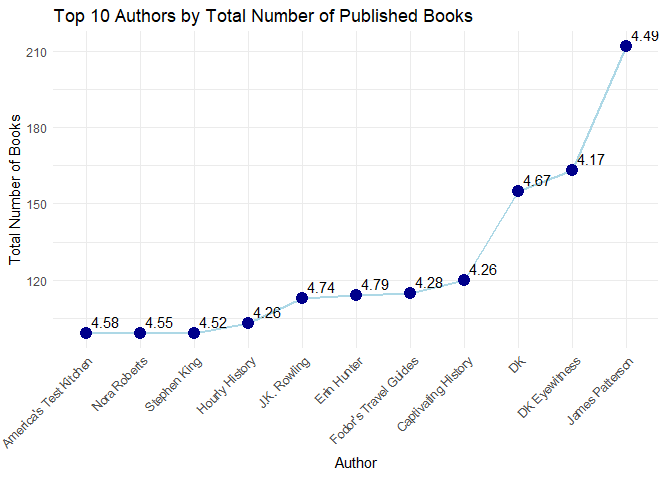


Figure : Top 10 Authors with book counts and average stars

**Explanation:**

*Figure 11* shows the exacteddata in descending order based on the total number of books and its average rating(stars) along with the author’s name. *Figure 12* represents The top 10 authors , and a line graph with points is created using ggplot. The line graph represents the total number of published books for each author, and points are annotated with the average rating. For clarity, the title, axis labels, and extra stylistic elements are used. The generated graph gives insight into author popularity based on their publication activity and average ratings. From analysis, it is clear that “**James Patterson**” is the most popular Author with **highest number of books** and **rating 4.49**. Although, “**Erin Hunter**” the **highest stars (4.79)** among all of them but the count of her books are less.

**6. Plot the horizontal bar plot for all categories based on Reviews and GoodReads Choice Status**

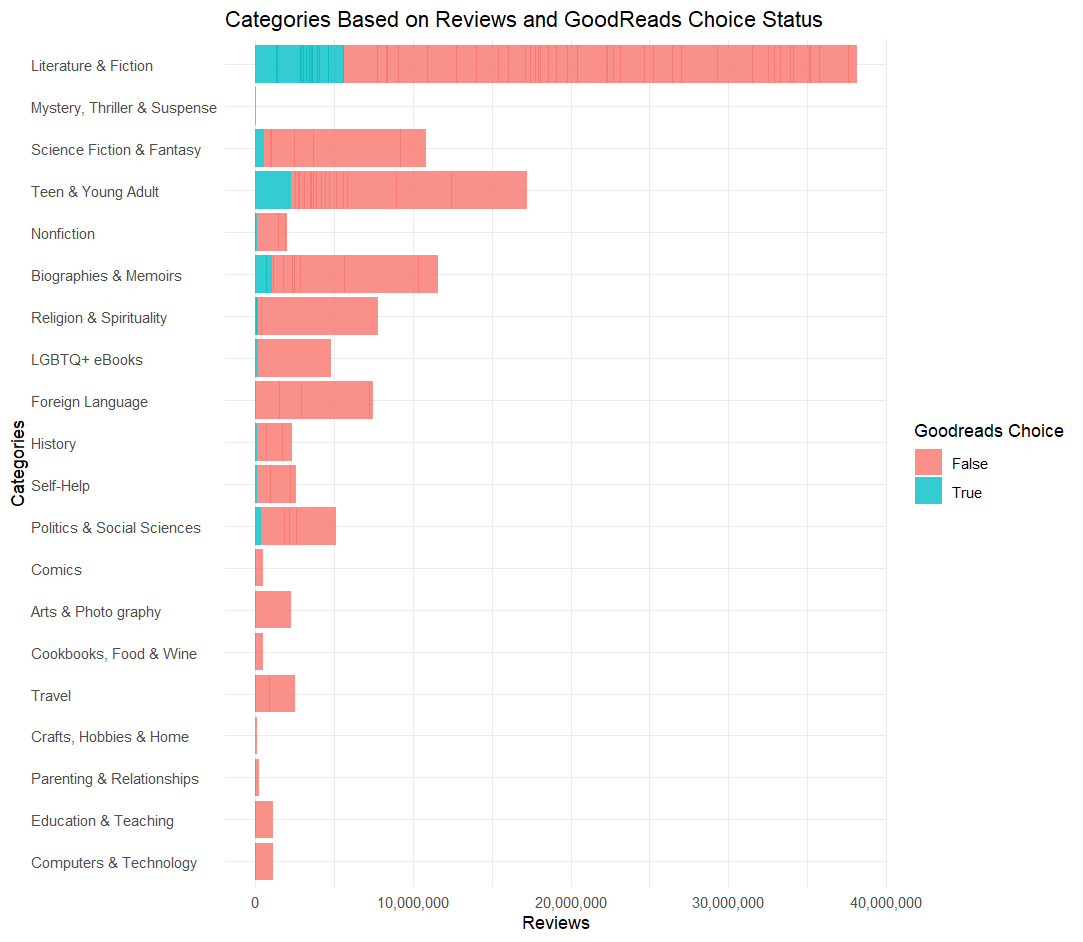


Figure : Categories based on Reviews and GoodReads Choice Status

**Explanation:**

As Shown in *Figure* 13, ggplot is used to create a bar plot that illustrates the distribution of reviews across different book categories. It eliminates categories with no reviews, offers a horizontal bar plot for improved visibility. The x-axis labels are formatted with commas, and the plot title, x-axis label, y-axis label, and fill legend title are all set to be clear. The final figure indicates the distribution of reviews across several categories, considering the GoodReads Choice status. The axes have been inverted for easier reading, and the legend title has been added to the fill legend. Graph states that the **Literature and Fiction** category has most number books which has reviews and Good reads choice is True. And the least one is **Mystery , Thrill and Suspense** which has least reviewed books and no Good reads choice.

**7. Display Top 10 Free Books(based on count of books) Categories with Stars >= 4**

A screenshot of a computer

Description automatically generated

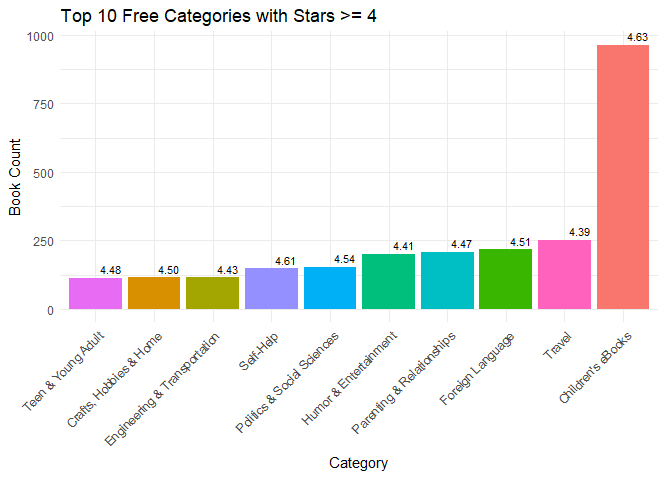
Figure : Extracted Top 10 free categories with stars >=4 and its book count

Figure : Top 10 free categories with highest stars

**Explanation:**

*Figure 14* displays the extracted free books with more than four stars, classifies them by category, and summarises the number and average stars for each. The top ten categories are then chosen based on book count, and *Figure 15.* Represents a ggplot bar graph is created. The graph shows the number of books in each category, with the top ten categories listed by count. For clarity, average stars are marked on the graph, and extraneous legends are deleted. The plot depicts the distribution of highly rated free books across different categories in a visual manner. Here, its is evident that **Children’s eBook** category has most number of free books with highest rating. One more little confusing category is **Travel,** as we saw above in *figure 8* , Travel is one of the top 10 bestselling category with highest price. That means Travel having most numbers of books some are costly, and some are free.

**8. Distribution of Top 10 Publishers (soldBy based on book counts) throughout the data**

**A screenshot of a computer

Description automatically generated**

Figure : Top 10 Publishers and their count

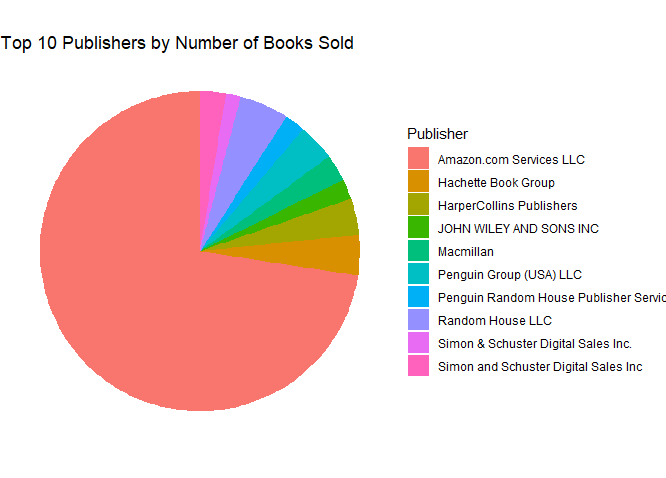
****

Figure : Top 10 Publishers by number of books sold

**Explanation:**

*Figure 16* shows the top 10 publishers and the count of the books sold. The pie chart (*Figure 17)* illustrates the distribution of the top ten publishers by number of books sold. Each slice symbolises a publisher correlate to how many books the publisher has sold. The chart illustrates the distribution of book sales among the top publishers in a simple yet insightful manner and it’s Clear that **Amazon.com Services LLC** has most number of books which makes it most popular publisher.

**Conclusion :**

Numerous aspects of the book data have been thoroughly and perceptively explored through the examination of the Kindle dataset. We have developed a comprehensive grasp of the dataset by exploring price information, sales trends, the impact of Kindle Unlimited availability, and the popularity of authors, categories and publishers(soldBy).

In conclusion, the integration of data exploration and visualisation has improved our understanding of the Kindle dataset and made it easier to extract insightful information. This information may be extremely useful in making educated decisions about publishing, marketing, and reader interaction in the digital book environment.