Project 3: Exploring Data Visualizations in R

Mohammed Saif Wasay (002815958)

Masters of Professional Studies in Informatics, Northeastern University

ALY 6000: Introduction to Analytics

Harpreet Sharma

January 30th, 2024

Introduction and Key Findings

In this project, we will look into different techniques to creating useful and engaging visualizations. We will investigate a Books dataset and draw relevant conclusions from the data and visualization. We will also study statistical concepts such as samples and populations, as well as measures of dispersion and central tendency, using the dataset. Here, Simple visualizations can be easily created in R, but more abilities and knowledge are needed to produce powerful and educational visuals. You should bear the following in mind while you improve your R visualization skills (Albusairi, 2023). This project is divided into two parts: first, we will clean the dataset and perform Data Wrangling to choose only the desired columns, and then we will experiment with different visualizations in R to obtain insights from data.

Part – 1: Cleaning the dataset and extracting desired columns

1) First step for any data analysis will be reading the data, making sure that the names are R friendly. Next, we use date functions from lubridate package in R to create new date columns to support our analysis. Different functions used are mdy() and year() to extract dates/year.

```
Console   Terminal ×   Background Jobs ×
R 4.3.1 · ~/Northeastern University/ALY-6000/Project 3/
> #Cleaning Dataset
> #Problem 1
> p_load(janitor)
> books <- clean_names(books)
>
>
> #Problem 2
> p_load(lubridate)
> books <- mutate(books, first_publish_date = mdy(first_publish_date))
Warning message:
There was 1 warning in `mutate()`.
i In argument: `first_publish_date = mdy(first_publish_date)`.
Caused by warning:
! All formats failed to parse. No formats found.
>
> #Problem 3
> books <- mutate(books, year = year(first_publish_date))
> books <- clean_names(books)
> names(books)
 [1] "title"              "series"             "author"             "rating"
 [5] "description"        "language"           "isbn"               "genres"
 [9] "characters"         "book_format"        "edition"            "pages"
[13] "publisher"          "publish_date"       "first_publish_date" "awards"
[17] "num_ratings"        "ratings_by_stars"   "liked_percent"      "setting"
[21] "bbe_score"          "bbe_votes"          "price"              "year"
>
```

2) We filter out the books to get data between 1990 and 2020 and with pages less than 1200. We then select columns required for our analysis.

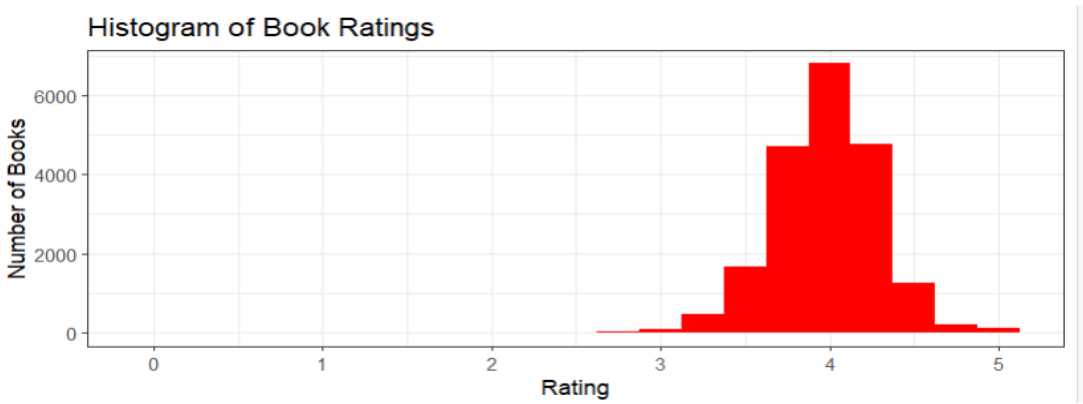
```
> #Problem 4
> books <- filter(books, year >= 1990 & year <= 2020)
>
> #Problem 5
> books <- subset(books, select = -c(publish_date, edition, characters, price, genres, setting, isbn))
>
> #Problem 6
> books <- filter(books, pages < 1200)
> summary(books)
    title              series             author              rating       description
 Length:20116       Length:20116       Length:20116       Min.   :0.000   Length:20116
 Class :character   Class :character   Class :character   1st Qu.:3.790   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :3.990   Mode  :character
                                                          Mean   :3.979
                                                          3rd Qu.:4.180
                                                          Max.   :5.000

   language         book_format           pages          publisher          first_publish_date
 Length:20116       Length:20116       Min.   :   0.0   Length:20116       Min.   :1990-01-01
 Class :character   Class :character   1st Qu.: 229.0   Class :character   1st Qu.:2000-10-30
 Mode  :character   Mode  :character   Median : 319.0   Mode  :character   Median :2006-10-30
```

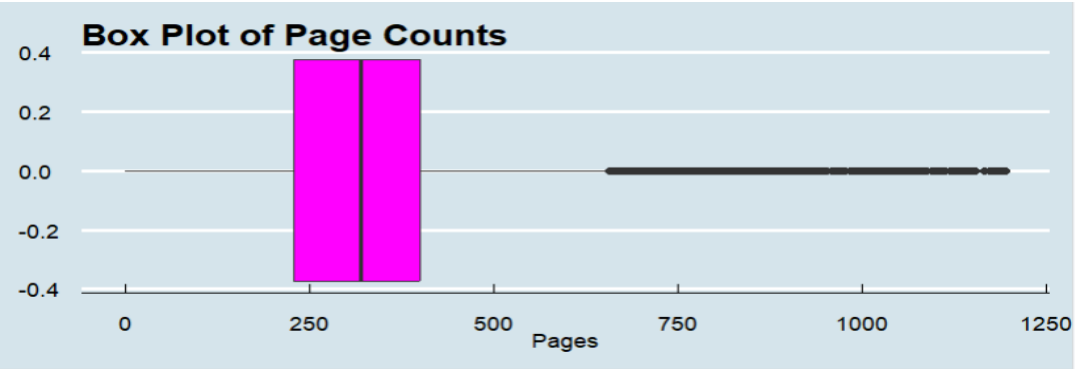
Part – 2: Visualizing Books Dataset using R

This section uses the dataset generated in the previous section to produce visualizations and get insights from data. The distribution of a single continuous variable is shown using a histogram. It separates the data into bins and uses bars to show how frequently observations occur in each bin (Albusairi, 2023).

To begin our research, we created a histogram of the number of books and their ratings. The rating has a typical distribution between 1990 and 2020.



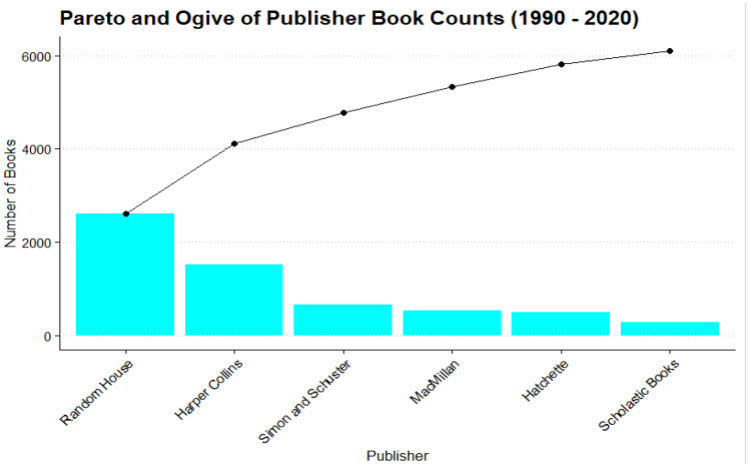
We plotted a box-plot for Number of Pages of books. We could see that 50 percentile of the books has 200-400 pages.



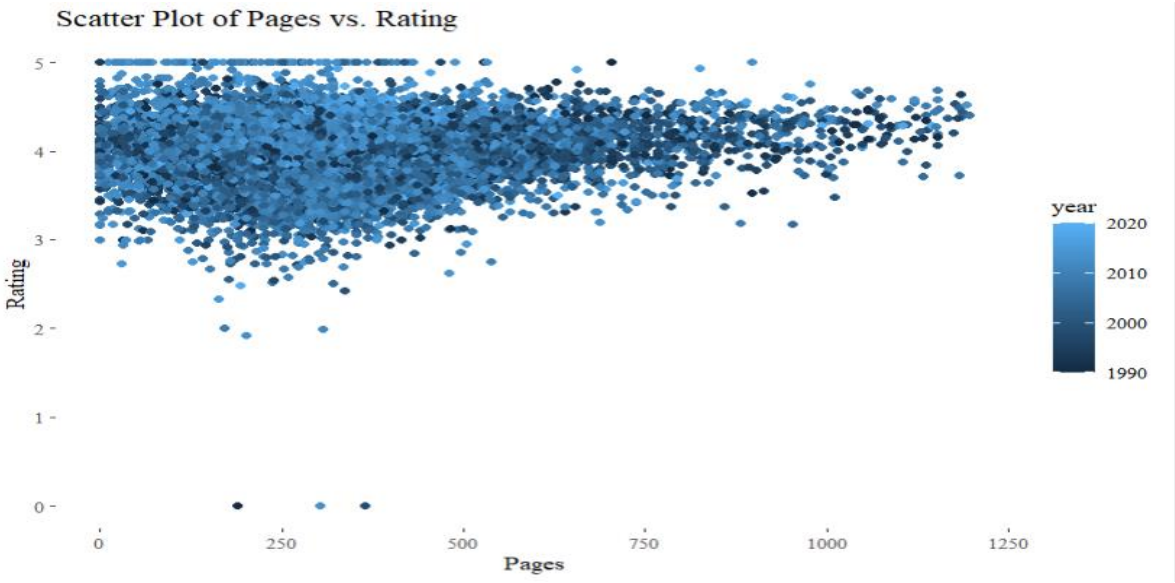
We generated a summary table containing information of publisher with atleast 250 books and calculated cumulative frequency, relative frequency and cumulative frequency.

```
publisher_summary <- books %>% group_by(publisher) %>%
                      summarise(total_books = n()) %>%
                      drop_na(publisher) %>%
                      arrange(desc(total_books)) %>%
                      filter(total_books >= 250) %>%
                      mutate(publisher = as.factor(publisher)) %>%
                      mutate(cum_count = cumsum(total_books)) %>%
                      mutate(rel_freq = total_books/sum(total_books)) %>%
                      mutate(cum_freq = cumsum(rel_freq)) %>%
                      select(publisher, total_books, cum_count, rel_freq, cum_freq)
publisher_summary
A tibble: 6 x 5
publisher          total_books cum_count rel_freq cum_freq
<fct>                    <int>     <int>    <dbl>    <dbl>
Random House              2607      2607    0.428    0.428
Harper Collins            1512      4119    0.248    0.676
Simon and Schuster         663      4782    0.109    0.785
MacMillan                  541      5323    0.0888   0.874
Hatchette                  493      5816    0.0809   0.955
Scholastic Books           277      6093    0.0455   1
```

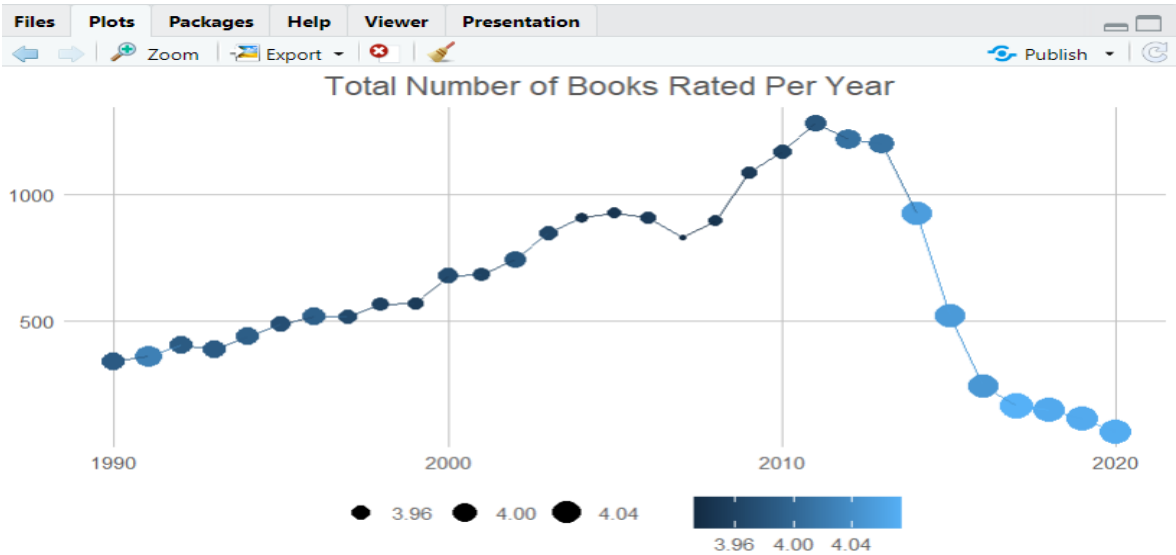
Plotting Pareto and Ogive of Publisher Book Counts.



Visualizing a scatter plot for exploring relation between Number of Pages in a Book and its rating.



We generated a dataset with total number of books and their average rating per year and plotted this information.



We obtained the statistics for the book ratings using new functions we developed to compute the mean, variance, and standard deviation of a population. Additionally, we divided the dataset into 3 samples and examined the standard deviation, variance, and average ratings of each

sample. These values appeared to be close to the values of the entire dataset, as we could observe.

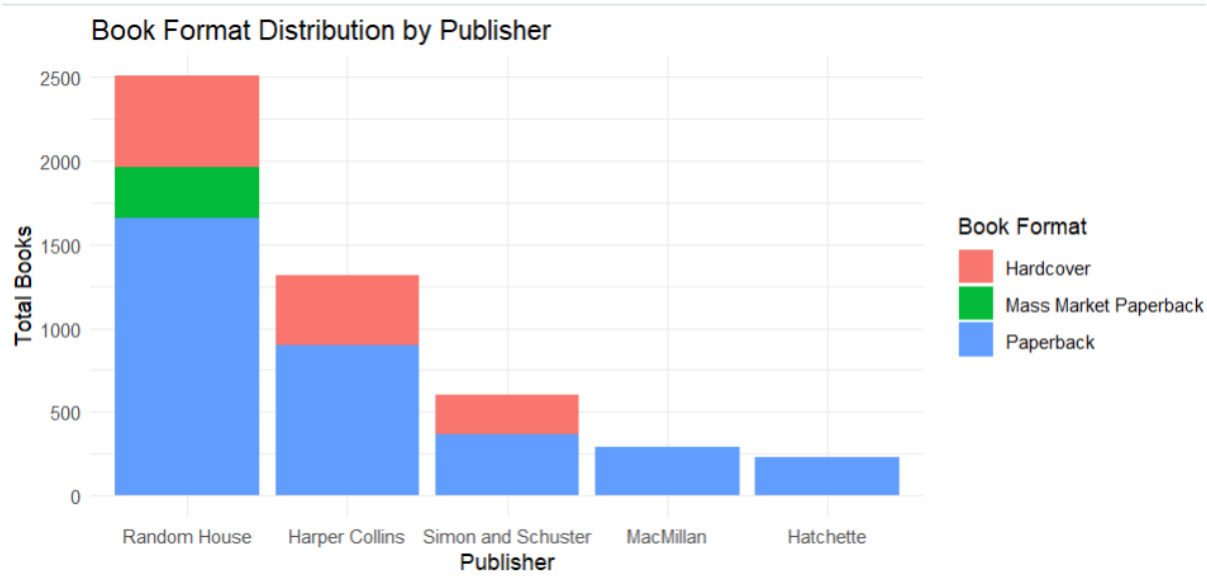
```
> #Problem 10
> average <- function(r){
+   sum(r)/length(r)
+ }
> pop_var <- function(r){
+   out_vector = c()
+   for (i in r){
+     result = (i - average(r))**2
+     out_vector = c(out_vector, result)
+   }
+   sum(out_vector)/length(r)
+ }
> sd_var <- function(r){
+   sqrt(abs(pop_var(r)))
+ }
> #Problem 11
> book_rating <- data.frame(avg_rating = average(books$rating),
+ variance = pop_var(books$rating),
+ sd = sd_var(books$rating))
> book_rating
  avg_rating   variance        sd
1   3.978595 0.09633514 0.310379
```

```
Console   Terminal ×   Background Jobs ×
R 4.3.1 · ~/Northeastern University/ALY-6000/Project 3/
> #Problem 12
> sample_1 <- sample_n(books, 100)
> sample_2 <- sample_n(books, 100)
> sample_3 <- sample_n(books, 100)
> sample_1_rating <- data.frame(avg_rating = average(sample_1$rating),
+                               variance = pop_var(sample_1$rating),
+                               sd = sd_var(sample_1$rating))
> sample_1_rating
  avg_rating  variance        sd
1     4.0389 0.1126018 0.3355619
> sample_2_rating <- data.frame(avg_rating = average(sample_2$rating),
+                               variance = pop_var(sample_2$rating),
+                               sd = sd_var(sample_2$rating))
> sample_2_rating
  avg_rating  variance        sd
1     3.9521 0.1159466 0.3405093
> sample_3_rating <- data.frame(avg_rating = average(sample_3$rating),
+                               variance = pop_var(sample_3$rating),
+                               sd = sd_var(sample_3$rating))
> sample_3_rating
  avg_rating   variance        sd
1     3.9786 0.09087004 0.3014466
> book_rating
  avg_rating   variance        sd
1   3.978595 0.09633514 0.310379
```

Furthermore, we examined Publishers based on their book format and total number of books, as well as their ratings such as average rating, minimum rating, and highest rating for each

format. For the publishers, we also displayed the distribution of total books in each format.

```
> by_publisher <- books %>% group_by(publisher, book_format) %>% summarise(total_books = n(),
+                                                 avg_rating = mean(rating),
+                                                 min_rating = min(rating),
+                                                 max_rating = max(rating)) %>%
+                           drop_na(publisher) %>%
+                           arrange(desc(total_books)) %>%
+                           filter(total_books >= 200)
`summarise()` has grouped output by 'publisher'. You can override using the `.groups` argument.
> by_publisher
# A tibble: 9 x 6
# Groups:   publisher [5]
  publisher          book_format           total_books avg_rating min_rating max_rating
  <chr>              <chr>                       <int>      <dbl>      <dbl>      <dbl>
1 Random House       Paperback                    1661       3.94       2.51       4.59
2 Harper Collins     Paperback                     900       3.95       2.82       4.68
3 Random House       Hardcover                     542       3.92       3.06       4.69
4 Harper Collins     Hardcover                     413       3.95       3.1        4.66
5 Simon and Schuster Paperback                     362       3.94       2.97       4.44
6 Random House       Mass Market Paperback         305       3.98       3.34       4.53
7 MacMillan          Paperback                     287       3.92       2.77       4.53
8 Simon and Schuster Hardcover                     242       3.95       3.02       4.61
9 Hatchette          Paperback                     227       3.96       2.96       4.49
```



Conclusion/Recommendations

Following our investigation, we visulaized the data, which helped us understand the data and draw numerous conclusions. We learned how to use the various date functions to extract dates from data. R is an excellent tool for data analysis and visualization. By understanding its basic yet powerful visualization capabilities, we can maximize the potential of our data and convey our results effectively (Albusairi, 2023). Furthermore, the data is Normally Distributed. After sampling the datasets we could see that the average, variance and standard deviation of the sample values were

near to the values of the entire dataset. We summarized publisher data by book format, generating the number of books, average rating, minimum and maximum rating. We can observe from this data that the majority of the books are paperback, followed by hardcover, and that Random House is the only publisher of mass market paperbacks.

Citations

R Documentation, An introduction to R. Retrieved 30th January 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation

Albusairi, F. (2023, March 26). Mastering Simple R Visualizations: From Scatter Plots to Heat Maps. . Retrieved 30th January 2024 from https://www.linkedin.com/pulse/mastering-simple-r-visualizations-from-scatterplots-heat-albusairi/.

Dataset Reference, Goodreads. Retrieved 30th January 2024 from https://www.goodreads.com/.