Project 3 ALY 6000

Project Instructions

In this project, you will examine data collected from the website www.goodreads.com, archived on www.kaggle.com, and modified for this project. Specifically, you will create different forms of informative and compelling visualizations in R. You will also draw conclusions from the data and report on them in written form, exploring the statistical ideas of samples and populations and the measures of dispersion and central tendency.

Note: Utilize the file **project3_tests.R** with the code below to run a series of tests (not comprehensive) on your code. Any failed test signals that something is wrong with the results or that you have not utilized the specified variable names.

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

Questions not checked by the test file will be graded manually after the due date.

When completed you will submit your work as **LastName-FirstName-Project3.Rmd** and **Lastname_Project3_Report.pdf.**

Project Setup

1. Download and open the R Markdown file "**LastName-FirstName-Project3.Rmd**." Replace LastName and FirstName with your own last name and first name.

2. Run the following code at the very top of the file. This will clear out the environment each time you run your entire code and prevent past actions from interfering with current work.

```
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
```

3. Install the *pacman* package. This is a simple, user-friendly package that makes installing and loading other packages a one-line process.

```
# You should do this line only once in the entire
course. install.packages("pacman")
```

```
# Once you have done the install line, the following line is what
you will always need to do to utilize the pacman package and other
libraries in R
library(pacman)
p_load(testthat)
p_load(tidyverse)
p_load(lubridate)
p_load(ggplot2)
```

4. For each question, write the code to answer it within its own cell in the file. All outputs and visualizations should appear underneath the cell. Note that there are some questions that require you to explain more in the report.

Project 3 Instructions

1. Download the file **books.csv** from Canvas and read the dataset into R.

```
Rows: 52448 Columns: 23
── Column specification
────────────────────────────────────────────────────────
Delimiter: ","
chr (17): title, series, author, description, language, isbn,
genres, charac...
dbl  (6): rating, pages, numRatings, likedPercent, bbeScore,
bbeVotes
```

Cleaning the data set

1. The **janitor** package contains helpful functions that perform basic maintenance of your data frame. Use the **clean_name** function to standardize the names in your data frame.

2. The **lubridate** package contains helpful functions to convert dates represented as strings to dates represented as dates. Convert the **first_publish_date** column to a type date using the **mdy** function.

```
Warning: There was 1 warning in `mutate()`. ℹ In argument:
`first_publish_date = mdy(first_publish_date)`. Caused by warning:
! 1186 failed to parse.
```

3. Using the **year** function in **lubridate,** extract the year from the **first_publish_date** column place it in a new column named **year.**

4. Reduce your dataset to only include books published between 1990 and 2020 (inclusive).

5. Remove the following columns from the data set: publish_date, edition, characters, price, genres, setting, and isbn.

6. Keep only books that are fewer than 1200 pages.

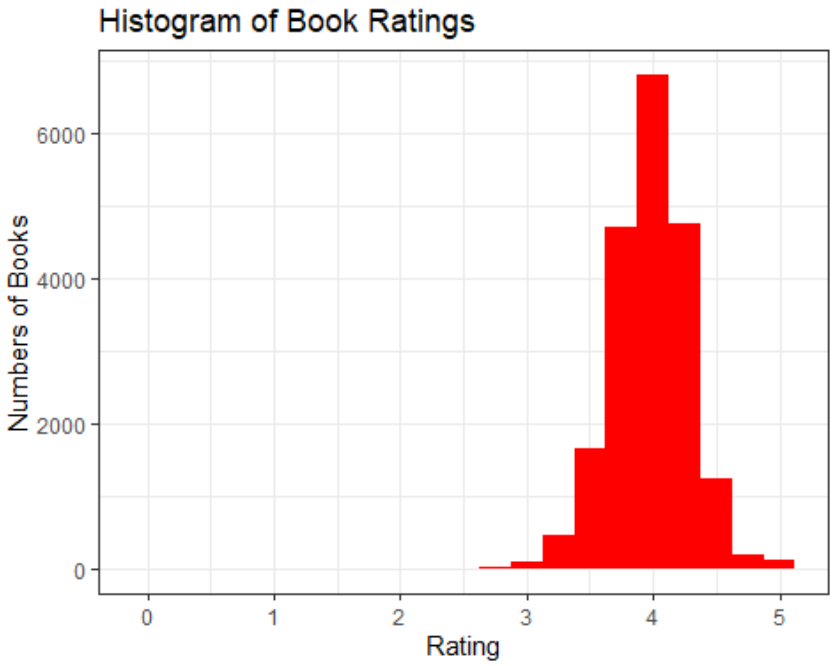
Data Analysis (Use the result from Q6 above for the following questions.)

1. Use the **glimpse** function to produce a long view of the dataset.

2. Use the **summary** function to produce a breakdown of the statistics of the dataset.

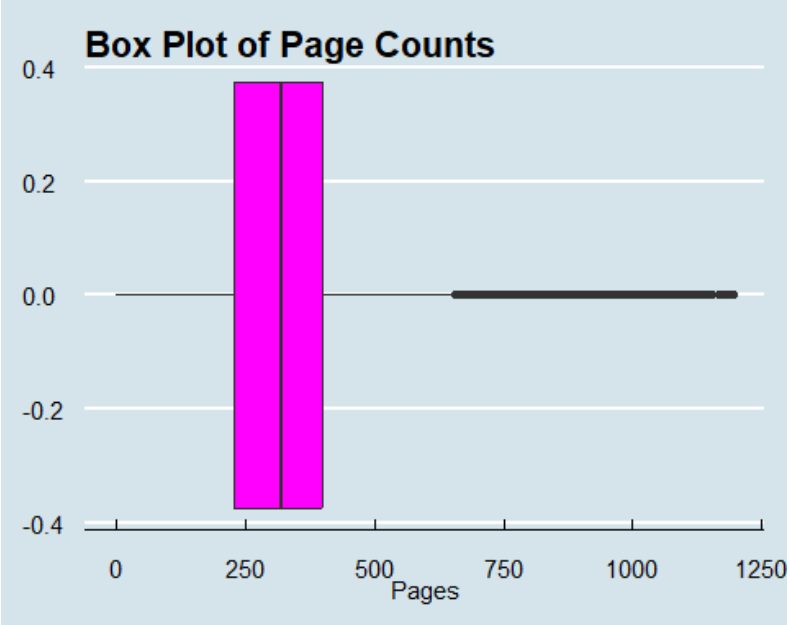
3. Create a rating histogram with the following criteria.

   - The y-axis is labeled "Number of Books."
   - The x-axis is labeled "Rating."
   - The title of the graph "Histogram of Book Ratings."
   - The graph is filled with the color "red."
   - Set a binwidth of .25.
   - Use theme_bw().



4. Create a boxplot of the number pages per book in the dataset with the following requirements.

   - The boxplot is horizontal.
   - The x-axis is labeled "Pages."
   - The title is "Box Plot of Page Counts."
   - Fill the boxplot with the color magenta.
   - Use the theme theme_economist from the ggthemes package.



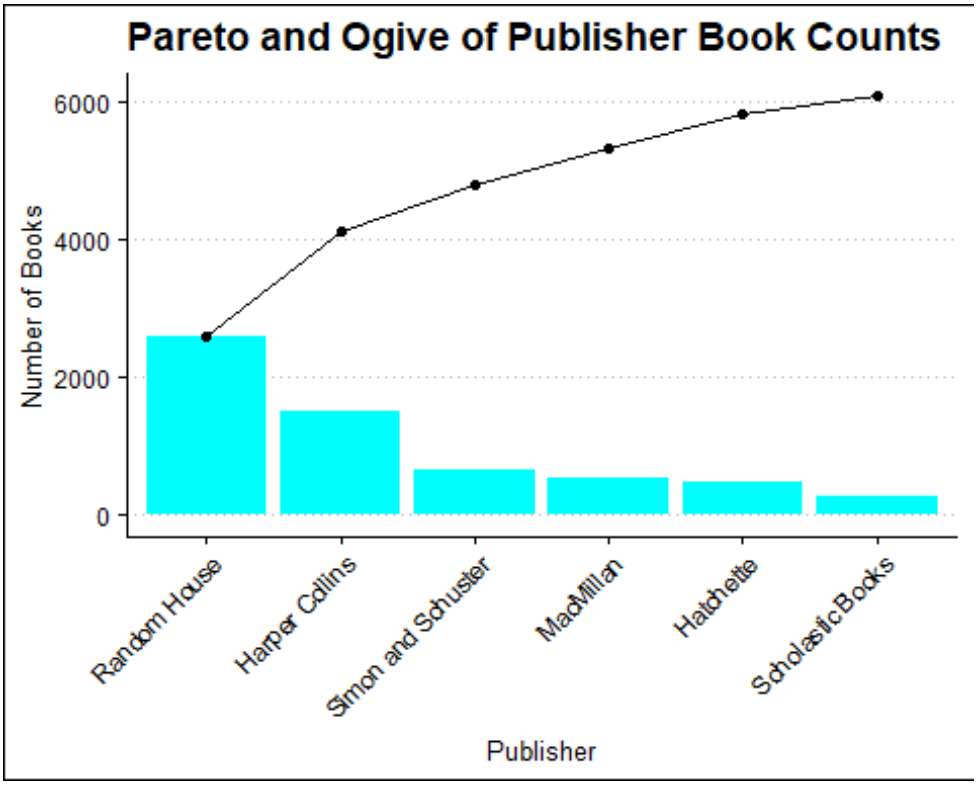
5. Group the data by publisher and produce a summary data frame containing each publisher and their associated number of books in the dataset. With that data frame, make the following refinements:

   - Remove any rows that contain NAs.
   - Remove any publishers with fewer than 250 books.
   - Order the data frame by the total number of books in descending order.
   - Make the publisher into a factor with the levels defined by the current ordering of the publisher.
   - Add a column to the data frame with cumulative count of books.
   - Add a column to the data frame with the relative frequency of books.
   - Add a column to the data frame with the cumulative relative frequency of books.

```
# A tibble: 6 × 5
  publisher          total_books cum_count rel_freq cum_freq
  <fct>                    <int>     <int>    <dbl>    <dbl>
1 Random House              2607      2607   0.428     0.428
2 Harper Collins            1512      4119   0.248     0.676
3 Simon and Schuster         663      4782   0.109     0.785
4 MacMillan                  541      5323   0.0888    0.874
5 Hatchette                  493      5816   0.0809    0.955
6 Scholastic Books           277      6093   0.0455    1
```

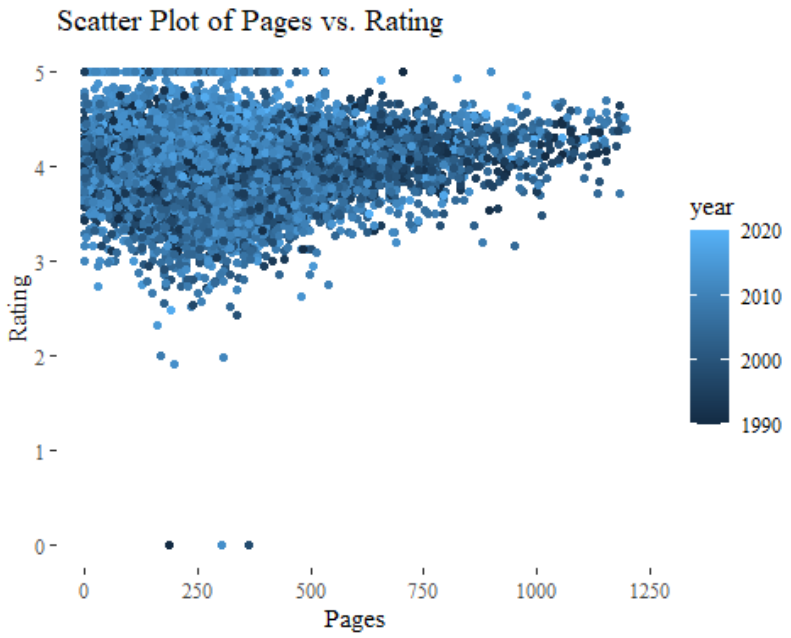
6. Using the data frame constructed in the prior problem, create a Pareto Chart with an ogive of cumulative counts formatted with the following additional criteria:

- The bars are filled with the color cyan.
- The x-axis label is “Publisher.”
- The y-axis label is “Number of Books.”
- The title is “Pareto and Ogive of Publisher Book Counts (1990 - 2020).”
- Use the theme theme_clean().
- Rotate the x-axis labels by 45 degrees (consider the **ggeasy** package).



7. Create a scatter plot of pages vs. rating for the books data frame with the following requirements:

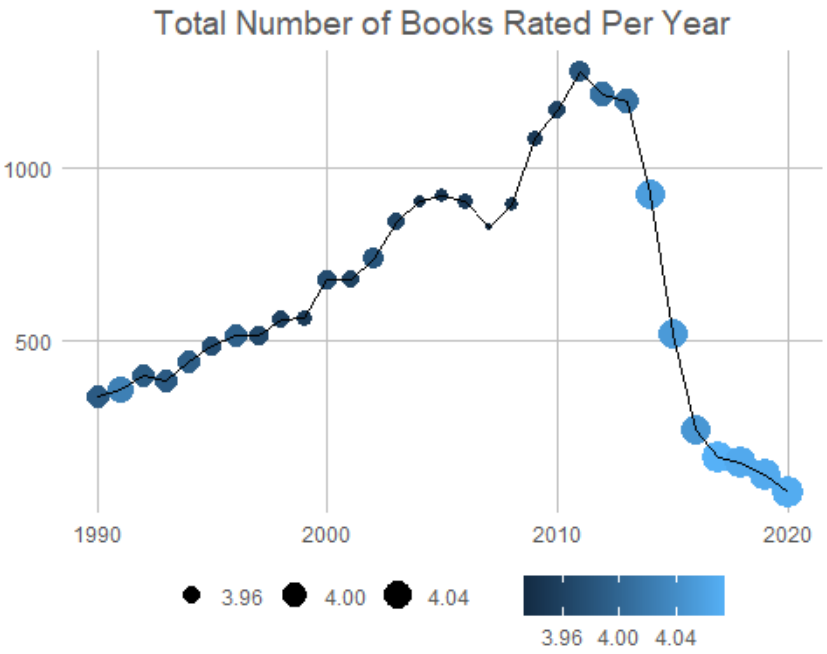
- Color the points based on the year of publication.
- The x-axis is labeled “Pages.”
- The y-axis is labeled “Rating.”
- The graph is titled “Scatter Plot of Pages vs. Rating.”
- Use the theme theme_tufte().



8. Create a data frame from the books data frame that contains a count of the number of books by year and the average rating for each year.

9. Create a line plot with from this data frame with points representing the counts per year from 1990 - 2020. Color the points for each year with the average rating. Format with the following specifications:

   - The graph is titled "Total Number of Books Rated Per Year."
   - The theme is theme_excel_new().



10. R has built-in functions to computer the sample mean (mean), sample variance (var), and sample standard deviation (sd). The function 'calc_mean below takes in a vector of values and returns the average. Using the function as a template, create two more functions; one to compute the population variance (pop_var) and the other one to compute the population standard deviation (sd_var). See module 4's slides for the difference between sample and population variance and standard deviation. You may not use the three built-in functions listed above, but may use other built-in functions. All three functions should accept a single vector of values and return the corresponding computed result.

```
calc_mean = function(x) {
  result = sum(x)/length(x)
  return(result)
}

x = seq(1, 10, by = 1)
calc_mean(x)
```

11. Consider the complete dataset of books to be the population you are analyzing. Compute population stats for the average, variance, and standard deviation of the book rating.

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

12. Create three samples of size 100 from the books data frame using the function **sample(df$rating, size = 100, replace = FALSE).** For each sample, compute sample statistics for mean, variance and standard deviation of the book rating. Compare these results with the population stats in your report.

13. Create one or more additional visualizations based on the existing data or additional analysis that you perform.

14. Write an executive summary report that contains an overview of your analysis, the visualizations you created with textual descriptions of key takeaways, and any key statistics that were computed in your analysis.

Submitting to Canvas

When you are satisfied with your solution.

1. **Remove** any lines in your code that have "install.packages."
2. **Remove** any lines in your code that use the **view** function.
3. Submit 2 files under the assignment in Canvas.
    1. Your R Markdown file named **LastName-FirstName-Project3.Rmd**.
    2. A PDF file of your four-page report titled **Lastname_Project3_Report.pdf**.

Your report should contain the following information formatted as specified:

Title Page
Include your name, assignment title, and submission date.

Introduction and Key Findings
Include an overview of the assignment and any findings.

Conclusion/Recommendations
Include evidence-based recommendations and visualizations or direct presentation of tabular data.

Works Cited
Include all sources, including YouTube videos, instruction materials, Google search results, and texts that informed your study of statistics and R.

Your report should be as concise as possible while maintaining fluency. Your key findings will be strongest if supported by visualizations or direct presentation of tabular data.

Your summary must adhere to APA guidelines, including page numbers on each page (including the title page) in the upper right corner. See the following examples for title pages, citations, and general APA formatting.

Congratulations on completing Project 3!