DUE DATE: FRIDAY, OCTOBER 14, 2022 by 11:59 PM on Gradescope
No lates accepted.

For this project, you will be given some data regarding a sample of apps from the Google Play Store. You will conduct various types of analysis and summarize your results in a report. You should create your report in RMarkdown. Your report should read like a professional report that could be read by someone without an extensive statistics background. Answers should be given in complete sentences and paragraphs. Graphs should be appropriately labeled. When statistics are computed, include an explanation of what information can be determined from those values. For the report itself, hide chunks of code where appropriate to make your output cleaner. Submit both your Markdown file and the knitted code as a **PDF** document. (You may create an HTML file, and when you open it, then Print to PDF.)

Submit both files to Gradescope.

Dataset Information:
The data set *apps_data.csv* contains data for a number of apps from the Google Play Store.

- **App** - The name of the app.
- **Category** - The category the app falls in. Only a subset of categories was chosen for the data.
- **Rating** - The average rating for the app.
- **Reviews** - The number of user reviews for the app.
- **App_Size** - Size of the app in MB.
- **Content.Rating** - The age rating of the app based on content.
- **Genre** - The genre of the app.

Tasks:

Task 1) Two quantitative variables that are of interest are `Rating` and `Reviews`. Write a paragraph summarizing and describing each variable. This paragraph should include relevant graphs (minimally a histogram and a boxplot), detailed descriptions of the graphs, appropriate descriptive statistics (minimally measuring the center and spread), and explanations what those statistics describe about the data. The explanations should be such that a person with limited statistical knowledge can understand. You should initially check for NA values.

Task 2) Two categorical variables that are of interest are `Category` and `Content.Rating`. Write a paragraph summarizing and describing each variable. This paragraph should include at least one relevant graph, detailed descriptions of the graphs, appropriate descriptive statistics (such as frequency), and explanations what those statistics describe about the data. The explanations should be such that a person with limited statistical knowledge can understand. You should initially make sure each variable is a factor.

Task 3) One would like to know if both `Rating` and `Reviews` vary by `Category`. Write a paragraph explaining whether or not you think these two variables vary by `Category` and detailing how you reached each conclusion. You should include graphs (minimally side by side boxplots) and summary statistics for each group that support your conclusions.

Task 4) One would like to know if `Content.Rating` varies by `Category`. Write a paragraph explaining whether or not you think `Content.Rating` varies by `Category` and detailing how you reached your conclusion. You should include at least one relevant graph and summary statistics (such as frequency) for each group that support your conclusions.

Task 5) Two variables that might be studied more in the future are `App_Size` and `Rating`. It would be helpful to know if these variables are normally distributed. Write a paragraph for each variable explaining why one should or should not assume the variable is normally distributed. Include all necessary graphs (at least one of a histogram or a Q-Q Plot), computations (skew and kurtosis), and tests (Shapiro-Wilk) conducted. Explain the implications of the graphs, computations, and test results. Explain in a way that a person with limited statistical knowledge would understand. For any test conducted, state all hypotheses and use a 1% significance level.

Task 6) Create a data frame that consists of only apps whose `Category` is `GAME`. Write a paragraph summarizing and describing the variable `Genre`. This paragraph should include at least one relevant graph, detailed descriptions of the graphs, appropriate descriptive statistics (such as frequency or relative frequency), and explanations what those statistics describe about the data. The explanations should be such that a person with limited statistical knowledge can understand. You should initially make sure each variable is a factor.

Task 7) Using your game data frame from Task 6), create a new ordered factor called `Size_Tier` that takes on the value of “Small” if the `App_Size` is less than 27, “Moderate” if the `App_Size` is at 27 and less than 63, and “Large” if the `App_Size` is at least 63. Write a paragraph explaining whether or not you think `Size_Tier` varies by `Genre` and detail how you reached your conclusion. You should include at least one relevant graph and summary statistics (such as frequency) for each group that support your conclusions.