DUE DATE: WEDNESDAY, NOVEMBER 23, 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 sets *apps_data.csv* and *paid_apps.csv* contain data for a number of apps from the Google Play Store. *apps_data.csv* consists of free apps and *paid_apps.csv* consists of paid apps.

- **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
- **Price** - Price of the App. (*paid_apps.csv* only).
- **Content.Rating** - The age rating of the app based on content.
- **Genre** - The genre of the app. Have not used

Tasks:

Task 1) Convert `Category`, `Content.Rating`, and `Genre` to factors in both data sets. Please include the code in your project, but you do not need to comment on it.

Task 2) Use *apps_data.csv*. One might be concerned that the population mean `App_Size` is greater than 25 MB.

- Conduct a significance test at an 8% significance level to determine if this is the case and describe your results in a paragraph. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge.
- Include an appropriate confidence interval or confidence bound and explain how that supports your conclusion.
- In Project 1, it was determined that `App_Size` was probably not normally distributed. Explain why you could still conduct the tests you conducted even when the population was not normal.

Task 3) Use *apps_data.csv* to determine if the mean number of reviews is different when an app is a game or not a game.

- Conduct a significance test at a 7% significance level to determine if the mean of `Review` for non games is different than the mean for games. Describe your results in a paragraph. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge. If you are making any assumptions, be sure to include any tests to validate those assumptions, also at a 7% significance level.
- Include an appropriate confidence interval or confidence bound and explain how that supports your conclusion.

Hint: Create two vectors, one where `Category == "GAME"` and one where it does not.

Task 4) Use *apps_data.csv*. One would like to know if `Content_Rating` varies by `Category`. Conduct a significance test at a 4% significance level to determine if the two variables are independent. Describe your results in a paragraph. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge.

Task 5) Use *apps_data.csv*. Create a simple linear regression model to predict the app's `Rating` using `App_Size`.

- Your response should include a scatterplot of the data, and a computation of the Pearson correlation coefficient.
- Your response should include checking the assumptions for linear regression (linearity, normality, and equal variance). Continue with the analysis even if the assumptions are not met.
- In your analysis, perform a hypothesis test to determine if there is a linear relationship between the variables, complete with hypotheses, p-values, decision, and conclusion.
- Include the equation of the regression line, and the value of R^2.
- Include an explanation of what the values of r and R^2 tell you, whether you believe there is a linear association, how well you believe the model fits the data, and why you believe those things.

Conduct any hypothesis tests at a 3% significance level.

Task 6) Using the data set *paid_apps.csv*, create a multiple linear regression model without interactions to predict `Price` as predicted by `Rating` and `App_Size`.

- Your response should include checking that the assumptions for linear regression are met (linearity, normality, and equal variance). Continue with the analysis even if the assumptions are not met.
- In your analysis, perform a hypothesis test to determine if the independent variables explain some of the variation in the dependent variable (complete with hypotheses, p-values, decision, and conclusion).
- Include the values of R^2 and R^2_{adj} in your report.
- If any of the independent variables are involved, conduct a hypothesis test to determine which ones are important (complete with hypotheses, p-values, decision, and conclusion). Explain how you decided which were important.

Conduct any hypothesis tests at a 10% significance level.

Task 7) Use *apps_data.csv*. In Project 1, we considered if `Reviews` vary by `Category`. Conduct a One-Way ANOVA test to see if the mean number of review varies by category at a 4% significance level.

- Check that the assumptions for ANOVA are reasonably met (normality and equal variance). Continue with the analysis even if the assumptions are not met. Describe your results in a paragraph, including any relevant graphs.
- In a second paragraph, describe the results of the test. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge.
- If there is an effect, conduct a Tukey Test and describe what those results tell you.

Task 8) Use *apps_data.csv*. In Project 1, we considered if `Rating` varies by `Category`. Conduct a One-Way ANOVA test to see if the mean rating varies by category at a 3% significance level.

- Check that the assumptions for ANOVA are reasonably met (normality and equal variance). Continue with the analysis even if the assumptions are not met. Describe your results in a paragraph, including any relevant graphs.
- In a second paragraph, describe the results of the test. Your paragraph should include hypotheses tested, the p-value, the decision you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge.
- If there is an effect, conduct a Tukey Test and describe what those results tell you.

Task 9) Using the data set *paid_apps.csv*, conduct a Two-Way ANOVA test with interactions to test the effects of `Content.Rating` and `Category` on the variable `Price`.

- Check that the assumptions for ANOVA are reasonably met (normality and equal variance). Continue with the analysis even if the assumptions are not met. Describe your results in a paragraph, including a bar chart of average price for each predictor.
- In a second paragraph, describe the results of your test. Your paragraph should include hypotheses tested, p-values, the decisions you made, and your conclusion. Your conclusion should be understandable to a person with limited statistical knowledge.
- You do NOT need to create an interaction plot or conduct a Tukey test.

Conduct any hypothesis tests at a 2.5% significance level.