STAT 380 Data Science Through Statistical Reasoning and Computation – Mini-Project 1

Due: By 11:59 PM on Thursday, October 17, 2024

Deadline

Since the mini-projects are a significant portion of your grade, you want to avoid late penalties on these assignments.

Instructions

The purpose of this Mini-Project is to have you explore a dataset. Some of the tasks required to do this project have not been covered in STAT 380. This is typical of a project in which you have some knowledge, but you must develop new skills as well.

# Data

The dataset CODGames2\_mp.xlsx contains information about the results of an online, first-person shooter video game. In the game, the player is part of a team trying to eliminate enemy combatants and/or achieve objectives. Each row represents the results of a single online match. All rows correspond to the same player. Currently, no description of the variables is available.

Submission

Each student should submit a report ***and*** code that addresses the tasks listed below. (The easiest way to create the report is to use an R Markdown document so that your code, plots, and explanations are next to each other as you do in the HW.)

This assignment is meant to be more formal than a typical HW assignment. A portion of the grade (approximately 4-5 points out of 30 points) will be assigned regarding the neatness and quality of your work. Be sure to proofread to correct typos, fix headings that are not displaying properly, correct hanging sentences (sentences/paragraphs that start immediately to the right of a plot but conclude below the plot), ensure plots are readable and use customized axis titles rather than raw variable names, be sure that discussions are not R comments placed in R chunks, and make an effort to write thoughtful discussions.

For this assignment, you have the option to work alone or with a small group. If you know nothing about these types of video games, it may be helpful to work with someone that does. This is a good opportunity to try out teammates for the final project.

Each student should submit the report. If you work with other students, include the names of any students that you worked with in your report ***and*** a discussion of each group member’s contributions. (All group members should submit the same report.) Although students like a divide-and-conquer approach to reduce the amount of work, all group members are expected to read and contribute to the work done by others.

# Frequently Asked Questions

* Question: Is there a limit on the number of students working together? Groups may include up to 3 students; however, all group members should be contributing. If you did the work by yourself, do not allow others to get credit if they have not contributed.
* Question: If I work with others, can we submit the same report? Yes, but **make sure everyone submits a report in Canvas**, make sure that all participants are listed in the report, and make sure that you include a discussion of group member contributions.
* Question: What if I do not know anything about these types of video games? Ask questions, read about the game, work with people who do. Although knowledge of the game should help, I do not assume that you are familiar with it. Your grade is not based on your understanding of the game.

# Tasks

1. Read the dataset into R.

* We have not covered methods for reading an .xlsx file into R.
* Yes, you could save as .csv but that risks potential data loss.
* It would be better to search for a method (R function) that allows you to open .xlsx files in R.

2. (DO NOT DO THIS TASK UNTIL AFTER CLASS ON FRIDAY, OCTOBER 11) Begin to explore the dataset. Explore 5 variables of your choice. (You should explore at least one categorical and at least one quantitative variable.)

* For each, you should identify the variable, the type of variable (categorical, quantitative, etc.), information about the types of values the variable can take, and the amount of missing data within the variable.
* For each variable, report some summary statistics and make a visualization.
  + If a variable is categorical, make a table of possible values and include counts or proportions. Then, make an appropriate plot to show the distribution of values.
  + If a variable is quantitative, include some basic summary stats like the mean, median, standard deviation, min, max, etc. and make an appropriate plot to show the distribution of values.
* For each variable, you should write a few sentences to describe what you have learned about the variable. Incorporate the summary statistics and comment on the amount of missingness into your discussion. Remember, the goal is not to make the plot/calculate the summaries. The goal is to use plots and summaries to gain knowledge from the data.
* Place each discussion next to the relevant plot/summary statistics rather than writing code for all 5 variables first followed by all discussion. If you the discussions are not near the plots, it makes it difficult for your audience to follow your explanations without excessive scrolling.

3. Begin a list of questions that you have about the dataset. Suppose you are working on a project and your manager gives you this dataset. Write 3 questions that you would ask your manager.

* You have not been given any information about what the variables represent. So, if you are unsure about the meaning of any of the variables, write your questions.
* You have not been given any information about how the data were collected. If you have questions about the data collection, write them down.
* If you have any other questions, write them down.
* NOTE: Even if you are familiar with the game and think that you understand the meaning of the variables, you are still expected to write 3 potential questions that you could ask.

4. Asking a question that can be answered using the dataset and data visualization.

* Part 1 – You will answer my question: Is the player’s performance, as quantified by the amount of experience points gained (TotalXP variable) changing over time?
  + To answer this, you should find a way to extract the month from the Date variable (we have not covered working with dates in STAT 380)
  + The lubridate package (part of tidyverse) is useful for working with dates
  + Once you have the months, create side-by-side boxplots showing the TotalXP based on the Month, where the months are named (i.e., month should be June, not the number 6). The months should be in chronological order, not alphabetical order
  + Based on the plot, answer the question. If you have knowledge about these types of games, you can also mention other variables that could affect TotalXP but were not included in the question.
* Part 2 – You will write your own question, create a data visualization that can help answer your question, and write an answer to your question. The question/visualization should explore the relationship between at least 2 of the variables in the dataset. Be sure you include both the question and the answer to your question based on the visualization you have created.