STAT 380 Data Science Through Statistical Reasoning and Computation – Final Project Description (4/17/23)

*Due: By 11:59 PM on Tuesday, May 2nd, 2023*

***Project Groups***

Students may work alone or in a group of up to four people. If you choose to work in a group, your group members will be asked to evaluate your performance/contribution(s) to the project. If you are a person who joins the group, but does not contribute, you should expect your final project grade to be reflective of this.

This project is a lot of work. If you choose to work alone, you are agreeing to do all this work by yourself and should not be upset that the project is a lot of work.

Groups should ***not*** collaborate with other groups or other people about this project.

***Project Overview***

The final project for this class involves an analysis of video game data that is related to the datasets used for the mini-projects. The final project will involve a combination of using data wrangling, data visualization, and modeling to answer a variety of research questions. In addition to applying modeling techniques that we have learned in class, you will be responsible for picking, researching, and implement one modeling technique that we have not covered.

The goal is to demonstrate that you are proficient at asking meaningful questions and answering them with results of an appropriate data analysis, that you are proficient in using the R tools we have covered, and that you are able independently acquire knowledge/implement a method that you were not directly taught. ***I expect to see you demonstrate an understanding of the concepts and code presented in class.*** If your code is not convincing in terms of proficiency with the ideas we have covered, you should expect a reduction in your project grade.

***Project Datasets***

This problem involves multiple datasets.

1. CODGames\_p1\_380.csv contains information about the results of an online, first-person shooter video game. (Games such as Call of Duty, Fortnite, Apex Legends, Battlefield has been immensely popular in the last 5 years.) The dataset contains information about a single player, who we will call Player 1. In the game, the player is part of a team trying to win an online match against other online competitors. Points are earned for various tasks such as eliminating enemy combatants, collecting items, capturing a location, etc. Each row represents the results of a single, randomly selected, online match in which Player 1 participated.
   * NOTE: When reading in this dataset, make sure the variables names are Map1, Map2, Choice, etc. instead of V1, V2, …
   * We have seen this issue before. To solve it, click the box for Yes for Heading in the Import Dataset interface.
   * This is not the same dataset used in the mini-projects and may require additional cleaning.
2. CODGames\_p2\_380.csv is similar to CODGames\_p1.csv but is for a different player, who we will call Player 2.
   * You will want to combine the information in this dataset with the information in CODGames\_p1\_380.csv.
3. CODMaps.csv contains information about the various maps available in the game. The dataset contains 3 variables:
   * Name – The name of the map/battleground
   * FirstAvailable – The event during which the map become available (Launch, Season 1, Season 2, etc.)
   * Date – The date on which the map become available
4. CODGameModes.csv contains information about various types of games available in the game. The dataset contains 3 variables:
   * Mode – The game type
   * ScoreLimit – The maximum number of points that a team can score in a game mode. If a team reaches this limit, the game ends regardless of the time remaining.
   * TimeLimit – The maximum length of the game. If the time limit is reached before either team reaches the score limit, the game ends.

***Project Deliverables***

There are several components to the final project:

* A written report that includes the details of your analysis, data visualizations, model comparisons, etc., and answers to the research questions. (Each group member should submit a copy of the report which includes all group member names. All members in the group should submit the same report.)
  + You have the option to do this in a word processing program (such as Microsoft Word, Google Docs, etc.) or as a Markdown document.
  + Either way, you should have explanations next to relevant plots/figures/tables, etc.
* A submission of R code. (Each group member whould submit a copy of the code which includes all group member names.)
* An evaluation for the performance of each teammate. You should evaluate your teammates in terms of participation (did they come to meetings, were the responsive to group discussions, etc.) and contribution (did they understand the material, offer insights, and add to the quality of the overall product). Your group evaluation should also discuss the contributions of each team member. (Each person in a group should submit individually. There will be a separate place to add this information in Canvas.)

***Goals/Tasks*** NOTE: In the next several days, I will be completing the project. As I do so, I may make adjusted to this section. Although the research questions will not change, I may modify the information that I ask you to provide or the hints.

The goal of this project is to demonstrate proficiency in a) the ability to ask a question and answer it with data, b) the techniques we have covered in class, and c) and that you are able independently acquire knowledge/implement a method that you were not directly taught. To do this, you will complete the tasks listed below.

***Task 1*** (Data Cleaning and Data Visualization):

Relevant Information: Prior to each online match, players in the game lobby are presented with two options for the battlefield of the upcoming game (`Map1` and `Map2`). The players have the option to vote and the resulting vote is recorded in the `MapVote` column. The winning map is listed in the `Choice` column. In the event of a tie vote, the map listed in `Map1` is chosen. (Games for which the player entered the lobby after the vote has taken place have no information in `Map1` and `Map2` but have the winning map presented in `Choice`.)

Research Question: Which maps are the most likely to ***win*** the map vote?

Notes: To answer this question, write a paragraph discussing how you plan to answer this question. Be sure to address the data quality issues mentioned below. Then, write code and answer the question. As part of your solution, you should calculate the proportion of times that each map was listed as a candidate (Map1 or Map2) ***and*** earned more votes than the other candidate. As part of this, you should consider whether a given map won the vote by getting more votes than the other option or was selected since it was `Map1` and the vote was a tie. You should also include a visualization of the results. There are some data quality issues (such as misspelled map names and extra (trailing) blanks in some entries) to solve for this problem. You can find the proper names/spellings in the CODMaps.csv file. To full receive full credit, you must write code to solve these issues rather than editing the .csv files.

***Task 2*** (Inference):

Relevant Information: There are a variety games types (GameType variable) within this dataset. The difference between the game types is that players have different objectives for the game. For instance, in the game type “Hardpoint”, teams earn points by capturing and defending a location. In “TDM” teams earn points by eliminating enemy opponents. As these game types have different objectives and may last for different amounts of time, the game type might affect the TotalXP earned.

Research Question: How does the game type affect TotalXP after accounting for the Score?

Notes: Score refers to the player’s score, not the “score” of the match (i.e., not the Result column). This answer requires some data wrangling that may require knowledge that we have not covered. (Again, part of the skillset you are working to develop is learning how to answer questions you have not seen previously.) In particular, there is no distinction between HC – TDM and TDM, no difference between HC – Hardpoint and Hardpoint. Write code to clean the values in the GameType column to reflect this information. Then, perform an exploratory data analysis by create appropriate visualizations/summary statistics that explore the distribution of the variables and show the relationship between TotalXP, Score, and GameType. (You decide on the type/number of visualizations, but the analysis should be complete.) Finally, build an appropriate model for TotalXP based on Score and GameType. You should use the model to then answer the research question.

***Task 3*** (Prediction):

Relevant Information: In this task, your goal is to compare a variety of classification methods. In particular, you should write your own research question that can be answered by comparing the effectiveness of various classification methodologies. To demonstrate your understanding of these methods, you should implement two classification methods from class, one of which must be random forest, and a third method that we ***will not*** cover in class. (The purpose of the using a method we did not cover is I want you to practice learning about a method and its implementation on your own. Basically, find a tutorial that explains the method and how to implement it.) You will then have to compare the results and decide which method was the most effective.

Research Question: Write your own question and be sure that the question and answer are clearly written in your report.

Notes: Since you will be using random forest, do ***not*** use a decision/classification tree as one of your other methods. For this problem, you should provide a brief description of the methods that you will use. (A description is more than listing the name of the procedure.) You will implement and compare the effectiveness of these methods. As part of this process, you will have to make a number of decisions such as whether you will do any data wrangling (maybe you remove partial matches, maybe you create new variables, etc.), which methods will you use, how will you fairly compare the results between methods, which method is best etc. All of these decisions should be included in your report.

NOTE1: You will make your life easier if you pick a response with a small number of levels.

NOTE2: As you are picking a method that we did not cover, one way to find techniques is by looking at the textbook that we have been covering: An Introduction to Statistical Learning with Applications in R by Gareth James et al. You can find a pdf of the textbook on the author’s site: <https://www.statlearning.com/>.

***Important Dates*:**

* (1 Points) Milestone 1: In Canvas, by Thursday, April 20th at 11:59 PM, submit the name of all group members.
* (3 Points) Milestone 2: In Canvas, by Tuesday, April 25th at 11:59 PM, submit evidence of progress towards at least 1 of the 3 tasks. This evidence could include your code, plots, part of your paper, etc.  ***It is recommended that you make a submission sooner***
* (1 Point) Final Report: This is due by the end of the day (11:59 PM State College time) on Tuesday, May 2nd. NOTE: I will consider slight (24 hour) extensions provided you can provide documentation of conflicts with multiple other finals on the due date May 2nd. This is mainly for people who are working individually.

Score Breakdown:

5 Points - Milestones

15 Points – Task 1

15 Points – Task 2

25 Points – Task 3

10 Points – Overall quality of the report. This is not meant to be a comprehensive list, but be sure the document is typo free, uses proper grammar, is well organized, plots are located close to narrative, you have demonstrated an understanding of class topics, etc. If it looks like you did the bare minimum and you have one sentence explanations, expect a low quality grade.

Total (including points from Milestones): 70 Points