STAT 380 Data Science Through Statistical Reasoning and Computation – Final Project Description

Due: By 11:59 PM on Tuesday, May 6, 2025

Project Groups

Students are required to work in groups. Groups must consist of 3-4 people. (If you want to work with students from a different section of 380 that I teach, please reach out to me and get approval. Groups of 2 students will also need special approval.) While you have the option to choose your own teammates, I will assign you to a group if you do not find your own group. If you are looking for teammates, please add your name and PSU email the Google Doc linked below. Feel free to contact others to try and build your team. **Once you have found a team, remove your name from the list**.

https://docs.google.com/spreadsheets/d/1fL9Wofc4464v3IruBZvO8P_upEKdkOf5vc_6TGsnz14/edit?usp=sharing

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

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. If you did not participate or contribute to the group, I reserve the right to lower your project grade.

At the conclusion of the project, each group member will be asked to rate the contributions of the other group members. Specifically, 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, add to the quality of the overall product, etc.). Your group evaluation should also discuss the contributions of each team member. Finally, I am expecting a thoughtful reflection of the group experience. The reflection should be much more detailed than just saying "All group members were great." or "10/10 for everyone." I want you to take time to actually reflect on the experience and contributions.

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 (i.e., if your code looks nothing like the ideas/code we have covered in class), you should expect a reduction in your project grade. Remember that part of the goal is to demonstrate that you understand the ideas presented in class.

In many ways, the project is like a take-home final exam; however, unlike an exam, you will work with a group. As with an exam, I will not look at your code, debug your code, or tell you if your answers are correct. I am willing to have high level conversations with you about the questions or your approach.

A note on generative AI: The goal of this project is to demonstrate ***your*** understanding of the material. If you are using generative AI (such as chatGPT) in this project, you are required to disclose that, provide the prompt(s) and response provided by the generative AI, ***and*** provide a detailed discussion of what you have added to the solution. If all you do is copy the answer/code from the technology and add a sentence or two, you are failing to demonstrate your understanding of the material. This will result in a grade of 0. If you fail to disclose your use of generative AI, you will receive a grade of 0. The goal of the project is to demonstrate your understanding.

Project Datasets

This problem involves multiple datasets. You can access the files by going to the Files >> Project folder in Canvas.

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 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. To avoid a lot of screenshotting plots, it is likely easiest for you to use R Markdown.
  - Either way, you should have explanations next to relevant plots/figures/tables, etc.
- A submission of R code. (Each group member should submit a copy of the code which includes all group member names.)
  - If the report was created using R Markdown, submit both the .html and .rmd files.
- 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.) Your evaluation is meant to be a thoughtful reflection of the project and should be more extensive than “Everyone was great” or “Everyone did their fair share.”

Goals/Tasks

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 – Complete without Generative AI):

Relevant Information: (Complete without using Generative AI) 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 when they are an option?

Notes: To answer this question, write a paragraph (or more) discussing how you plan to answer this question. Be sure to address the data quality issues mentioned below and discuss how you will do the calculations. Then, write code and answer the question. (If I must answer questions about your approach/decision making process by reading your code rather than your discussion, you will lose points.) As part of your solution, you should calculate the proportion/probability that a map wins the vote given that it was a candidate. To do this, you will have to calculate the number 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 if it was selected since it was `Map1` and the vote was a tie. You should also include a visualization of the results. There might be 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.

Final Emphasis: Again, in Task 1, you are ***not*** to use generative AI. Further, after doing Task 2, do NOT go back and "fix" your answer to Task 1 by implementing the Task 2 code. The point of Tasks 1 and 2 is to take two approaches to the same problem and compare results.

Task 2 (Data Cleaning and Data Visualization – Complete using Generative AI):

Repeat Task 1 using a generative AI of your choice. To answer this question, mention the tool (including version number if appropriate) you have selected. Then, discuss the prompt(s) you have used and provide the solution produced by the generative AI. While it is fine to paste the question into the generative AI as your first prompt, you should also use additional follow-up prompts if it is beneficial to do so. Be sure to discuss all prompts used in your report.

Then, implement the generative AI solution.

Finally, and most importantly, you should compare your solution from Task 1 to the generative AI solution. Discuss similarities/differences, strengths/weaknesses, etc., and provide an overall assessment of which solution is better. The discussion should consider the correctness of the answers and should be substantial. Demonstrate that you have given the comparison considerable thought by making at least 3 substantial points as part of your comparison. Each point should take the form of a well-written paragraph.

Task 3 (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, and so on for the other game types. 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 4 (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 should describe how the procedure works.) 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. If I have to learn about your decisions/analysis by reading your code, you will lose points.

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 24 at 11:59 PM (Eastern), submit the name of all group members. All group members are expected to submit. If you have not found a team, please indicate this so that I know that I can assign you to a team. Text box entry in Canvas. ***It is recommended that you make a submission sooner.***
- (3 Points) Milestone 2: In Canvas, by Wednesday, April 30 at 11:59 PM, submit evidence of progress towards at least 1 of the 4 tasks. This evidence could include your code, plots, part of your paper, etc. Submit both your .rmd and .html files. All group members are expected to submit. ***It is recommended that you make a submission sooner.***
- (71 Points) Final Report: This is due by the end of the day (11:59 PM State College time) on Tuesday, May 6. All group members are expected to submit. Submit in Canvas.
- (0 Points) Group Evaluation: Complete the evaluation of your team members by the end of the day (11:59 PM State College time) on Tuesday, May 6 All group members are expected to submit. Text box entry in Canvas.

Score Breakdown

- 4 Points - Milestones
- 61 Points - Tasks
  - 17 Points – Task 1
  - 10 Points – Task 2
  - 12 Points – Task 3
  - 22 Points – Task 4
  - The point totals for each task are approximate and may change slightly when I am grading the final projects (e.g., Task 1 might be worth 16 and Task 2 might be worth 11)
- 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, included output/plots are explained, 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): 75 Points