Project Planning Stage (Individual)

Start Assignment

- Due Nov 12 by 6:01a.m.
- Points 100
- Submitting a file upload
- File Types ipynb and html
- Available Sep 29 at 12a.m. - Dec 30 at 11:59p.m.

Data Science Project: Planning Stage (Individual)

In this class, you will complete a full Data Science project from beginning to end, and produce a report communicating your methods and conclusions in a Jupyter Notebook. The Jupyter Notebook will perform the entire analysis: the code cells will download a dataset, reproducibly and sensibly wrangle and clean, summarize and visualize the data, as well as appropriately answer a predictive question. Markdown cells will be used throughout the document to narrate the analysis and communicate the question asked, methods used and the conclusion reached.

Problem: Predicting Usage of a Video Game Research Server

This year we have a unique opportunity: we have a **real data science project with real stakeholders** who are looking for answers to a few questions about their data.

In particular, **a research group in Computer Science at UBC (https://plai.cs.ubc.ca/)** , led by **Frank Wood (https://www.cs.ubc.ca/~fwood/)** , is collecting data about how people play video games. They have set up a **MineCraft server (https://plaicraft.ai)** , and players' actions are recorded as they navigate through the world. But running this project is not simple: they need to target their recruitment efforts, and make sure they have enough resources (e.g., software licenses, server hardware) to handle the number of players they attract. There are three broad questions of interest.

Question 1: What player characteristics and behaviours are most predictive of subscribing to a game-related newsletter, and how do these features differ between various player types?

Question 2: We would like to know which "kinds" of players are most likely to contribute a large amount of data so that we can target those players in our recruiting efforts.

Question 3: We are interested in demand forecasting, namely, what time windows are most likely to have large number of simultaneous players. This is because we need to ensure that the number of licenses on hand is sufficiently large to accommodate all parallel players with high probability.

In your project, you will select one of these broad questions and use it to formulate a specific question using some of the variables in the dataset. Your project should answer your specific question.

The Data

The data consist of two files:

players.csv (https://canvas.ubc.ca/courses/171896/files/41488716?wrap=1) **(https://canvas.ubc.ca/courses/171896/files/41488716/download?download_frd=1)** : A list of all unique players, including data about each player.

sessions.csv (https://canvas.ubc.ca/courses/171896/files/40880732?wrap=1) **(https://canvas.ubc.ca/courses/171896/files/40880732/download?download_frd=1)** : A list of individual play sessions by each player, including data about the session.

Grade Breakdown

The group project is worth 10% of your final grade overall, with the following breakdown:

- Team Contract: 0% (you will receive 0% on the *whole project* if you do not complete this with your group)
- Individual Planning Report: 3%
- Final Project Report: 7%

Individual Planning Report

Each student is expected to prepare a 1 page (max 500 words, where code does not count toward the word count) written proposal that describes the data they are working on, demonstrates an understanding of all variables and potential issues in the data, and identifies both the broad question they would like to address and the specific question they have formulated. The proposal should be done in a Jupyter notebook, and then submitted in **two formats:**

- as an .html file (File -> Download As -> HTML)
- as an .ipynb file. **This file must be fully reproducible. It must run completely from top to bottom without any additional files.**

It's important to note that this first step in the project will be completed individually. Every student needs to write and submit their own assignment. We aim to ensure that all students in the group are well-prepared and able to contribute effectively to the final report.

In your planning report you need to cover the following:

(1) Data Description:

Provide a full descriptive summary of the dataset, including information such as the number of observations, summary statistics (report values to 2 decimal places), number of variables, name and type of variables, what the variables mean, any issues you see in the data, any other potential issues related to things you cannot directly see, how the data were collected, etc. Make sure to use bullet point lists or tables to summarize the variables in an easy-to-understand format.

Note that the selected dataset(s) will probably contain more variables than you need. In fact, exploring how the different variables in the dataset affect your model may be a crucial part of the project. You need to summarize the full data regardless of which variables you may choose to use later on.

(2) Questions:

Clearly state one broad question that you will address, and the specific question that you have formulated. Your question should involve one response variable of interest and one or more explanatory variables, and should be stated as a question. One common question format is: “Can [explanatory variable(s)] predict [response variable] in [dataset]?”, but you are free to format your question as you choose so long as it is clear. Describe clearly how the data will help you address the question of interest. You may need to describe how you plan to wrangle your data to get it into a form where you can apply one of the predictive methods from this class.

(3) Exploratory Data Analysis and Visualization

In this assignment, you will:

- Demonstrate that the dataset can be loaded into R.
- Do the **minimum necessary** wrangling to turn your data into a tidy format. Do not do any additional wrangling here; that will happen later during the group project phase.
- Compute the mean value for each quantitative variable in the players.csv data set. Report the mean values in a table format.
- Make a few exploratory visualizations of the data to help you understand it.
    - Use our visualization best practices to make high-quality plots (make sure to include labels, titles, units of measurement, etc)

- Explain any insights you gain from these plots that are relevant to address your question

Note: **do not** perform any predictive analysis here. We are asking for an exploration of the relevant variables to demonstrate that you understand them well *before* performing any additional modelling, and to identify potential problems you anticipate encountering.

(4) Methods and Plan

Propose one method to address your question of interest using the selected dataset and explain why it was chosen. **Do not** perform any modelling or present results at this stage. We are looking for high-level planning regarding model choice and justifying that choice.

In your explanation, respond to the following questions:

- Why is this method appropriate?
- Which assumptions are required, if any, to apply the method selected?
- What are the potential limitations or weaknesses of the method selected?
- How are you going to compare and select the model?
- How are you going to process the data to apply the model? For example: Are you splitting the data? How? How many splits? What proportions will you use for the splits? At what stage will you split? Will there be a validation set? Will you use cross validation?

(5) GitHub Repository

Provide the link to your GitHub repository for the project. You must have at least **five commits** with a description of the work that has been done towards completion of the individual report in the commit history of this repository.

proposal_rubric (3)

Criteria	Ratings							Pts
Mechanics	**10 pts** **Excellent** The submission appears to be self-contained and work flawlessly; any necessary libraries to install are made obvious that that the evaluator must install them. Student submitted an HTML rendering of an .ipynb notebook. The submission was a single file with all figures included.		**7 pts** **Good** The submission had minor errors in style but works. The submission was an HTML rendering of an .ipynb notebook containing all text and figures.	**5 pts** **Unsatisfactory** The submission was an HTML rendering of an .ipynb notebook, but the evaluator noted obvious flaws in the code or text.	**2 pts** **Poor** The submission was not an HTML rendering of an .ipynb notebook. The evaluator was unable to open the submission, or noted many significant flaws in code or text.		**0 pts** **No Marks** No attempt/submission	10 pts
Reasoning	**70 pts** **Excellent** Mastery of the learning material is demonstrated, original ideas may be presented. The scientific question is well posed, creative and interesting. The correct method is proposed. Thesis is clear and the arguments that support it are flawless and very well-reasoned, leaving no obvious gaps. Structure of argument is very clear and straightforward; the reader almost never has to jump back and forth unless clearly instructed to do so by references.	**63 pts** **very good** Between excellent and good.	**56 pts** **Good** There is a clear purpose to the submission, understanding of the learning material is demonstrated. The scientific question is well posed. The proposed methodology is sound. Thesis is clear and the arguments and reasoning presented back up the thesis well. Structure of argument is clear and delineated sensibly into paragraphs. Included figures are labelled clearly and sensibly.	**42 pts** **Satisfactory** There is purpose to the submission, some understanding of the learning material is demonstrated. The scientific question is well-posed. Reasonable methods are proposed. Thesis, arguments and reasoning are present but do not strongly back up the thesis. Structure of argument is somewhat clear. Paragraph delineation could be improved. Included figures labels need more clarity/could be better.	**35 pts** **Unsatisfactory** Proposed project lacks a purpose, little to no understanding of the learning material is displayed, important information is lacking. Scientific question is unclear. The text may contradict itself, or obvious gaps in the argument are present. Reasoning is flawed or insufficient, does not accurately back up claim, or no clear thesis established. Structure of argument is confusing and poorly laid-out; the reader may have to jump back and forth	**14 pts** **Poor** Submission does not propose a reasonable project or makes no sense.	**0 pts** **No Marks** No attempt/submission	70 pts

Criteria	Ratings							Pts
					through the text. Any figures included are not labelled clearly or sensibly.			
Writing	**20 pts** **Excellent** No grammar or spelling errors are present. The submission is concise and to the point; the page, word or sentence count was respected.	**16 pts** **Good** Fewer than 5 grammatical or spelling errors are present. The submisison is not too long or too short; if there was a word or sentence count given, then it was not exceeded by any significant margin.	**12 pts** **Satisfactory** Fewer than 10 grammatical or spelling errors are present. The submisison is not too long or too short; if there was a word or sentence count given, then it was not exceeded by any significant margin.	**10 pts** **Unsatisfactory** Many (> 10) grammatical and/or spelling errors are present but the meaning of text is not significantly obscured by grammar or spelling errors. The submisison is not too long or too short; if there was a word or sentence count given, then it was not exceeded by any significant margin.	**4 pts** **Poor** Meaning of text is obscured due to significant grammar and spelling errors. The submission is far too long (or far too short); word or sentence count significantly exceeded if one was given. The submisison is not too long or too short; if there was a word or sentence count given, then it was not exceeded by any significant margin.	**0 pts** **No Marks** No attempt/submission		20 pts
								Total Points: 100