Project Final Report (Individual)

- Due Monday by 11:59p.m.
- Points 100

Data Science Project

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, 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/165752/files/39494591?wrap=1) **(https://canvas.ubc.ca/courses/165752/files/39494591/download?download_frd=1)** : A list of all unique players, including data about each player.

sessions.csv (https://canvas.ubc.ca/courses/165752/files/39494592?wrap=1) **(https://canvas.ubc.ca/courses/165752/files/39494592/download?download_frd=1)** : A list of individual play sessions by each player, including data about the session.

Grade Breakdown

The project is worth 3% of your final grade overall.

Individual Report

Each student is expected to prepare an electronic report in English with a maximum of 2000 words (excluding citations) using Jupyter. The report should include the posed question, conducted analysis, and derived conclusion. If needed, consult your TA and Instructor for further guidance.

On **Gradescope**, you must submit the following**:**

- a .pdf file which includes a link to your project's GitHub repository.
- a .ipynb file. **This file must be fully reproducible. It must run completely from top to bottom without any additional files.**

Each report should include the following sections:

- **Title**
- **Introduction**:
  - Background: provide some relevant background information on the topic so that someone unfamiliar with it will be prepared to understand the rest of your report
  - Question(s): clearly state the question you tried to answer with your project. 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.
  - Data Description: identify and fully describe the dataset that was used to answer the question. Provide a full descriptive summary of the dataset, including information such as the number of observations, summary statistics, 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.
- **Methods & Results**:
  - describe the methods you used to perform your analysis from beginning to end that narrates the analysis code.
  - your report should include code which:
    - loads data
    - wrangles and cleans the data to the format necessary for the planned analysis
    - performs a summary of the data set that is relevant for exploratory data analysis related to the planned analysis
    - creates a visualization of the dataset that is relevant for exploratory data analysis related to the planned analysis
      - 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
    - performs the data analysis. For your analysis, you should think about and provide a brief explanation of 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 did you compare and select the model?
      - Note: you should also think about the following:
        - 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?
    - creates a visualization of the analysis
    - *note: all figures should have a figure number and a legend*
- **Discussion**:
  - summarize what you found
  - discuss whether this is what you expected to find?
  - discuss what impact could such findings have?

- discuss what future questions could this lead to?

- **References**
  - You may include references if necessary, as long as they all have a consistent citation style.

GitHub Repository

On your GitHub repository, you must have at least **five commits** with a description of the work that has been done.

report_rubric (1)

Criteria	Ratings							Pts
Mechanics	**10 pts** **Excellent** The submission is 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 as well as the .ipynb source. The report is 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, as well as the .ipynb source.	**5 pts** **Unsatisfactory** The submission was an HTML rendering of an .ipynb notebook, as well as the .ipynb source. The .ipynb does not run all the way through, or the evaluator noted obvious flaws in the code or text.	**2 pts** **Poor** The submission was not an HTML rendering of an .ipynb notebook or its source. 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