STAT207 – Data Science Exploration – Mini-Project #1 – 35 Points

Due: Friday, February 21 11:59pm CST on Canvas.

Main Goal of Analysis

The main goal of this project, is to tell a compelling story based on the data science analyses you will perform on a dataset. **This is an individual project.**

To receive full credit, you should follow the steps and answer the questions given in this document for your project.

In addition to being graded for correctness and completion, this project will be graded on a qualitative basis. Qualitatively, we will be looking for the following things.

- **Clarity about Analyses, Algorithms, and Data Choices**
    - Someone who has taken STAT207-level class should be able to read through your report and easily be able to do the following.
        - Replicate what you did in your analyses, *without looking at the code!*
        - Know why you made the choices that you did in your analyses.
- **Clarity about Motivation (ie. the "so what?") of your Analyses**
    - *Beginning of the Report:*
        - Someone who is **about to** read your report and watch your presentation should be able to clearly answer the questions.
            - *"Why should I (or someone else) care about the report that I am about to read/listen to?"*
            - *"What research questions do they intend to answer?"*
            - *"How do these research questions relate to their motivation?"*
        - Therefore, in the introduction of your report and presentation you should make this clear.
    - *Middle of the Report:*
        - While **in the middle of** your report and presentation, your audience should be able to clearly answer the question.
            - *"How do each of these analyses/algorithms/data choices that they're making/using tie back into the overarching motivation of this whole analysis?"*
        - Therefore, for each new analysis/model/algorithm/data choice that you make, you should explain this and make it clear to your audience.
    - *End of the Report:*
        - Someone who has **just finished** reading your report and watching your presentation should be able to clearly answer the questions:
            - *"Why should I (or someone else) care about the analysis that I just read/listened to?"*
            - *"Did their analyses and conclusions answer the research questions that they stated at the beginning of the report/presentation? If so, how? What were the answers to these research questions?"*

        - *"How would the results/answers to these research questions be useful to someone?"*
    - Therefore, in the conclusion of your report and presentation you should make this clear.
- **Professionalism**
    - Your report and findings should be well-explained and written in **paragraphs** and **complete sentences** and in the **markdown cells (not in code blocks or in comments).**
    - Do not just spit out code and expect your reader to automatically know:
        - Why you chose to use this code, what its purpose is, what you're doing in the code block, and what you want them to notice in the result.
        - Why the output of your code is important.
        - How your code answers any relevant questions.
    - Any paragraphs, sentences, and explanations that you write should be considered satisfactory to, say, your high school writing teacher.
    - **Code or text that exhibits the linguistic style or structure of AI tools like ChatGPT will lose significant points in the corresponding professionalism sections in the report rubric.**

Intended Audience/Reader of your Project

The intended audience of your report/presentations should be someone who has the same level statistical/python knowledge as you and your STAT207 classmates. **Theoretically, you should be able to send your report to one of your classmates and they should be able to understand everything that you did and the claims that you are making.**

Project AI Tools Policy

To reiterate, **code or text that exhibits the linguistic style or structure of AI tools like ChatGPT will lose significant points in the corresponding professionalism sections in the report rubric.** The reason for this is because, in addition to running the risk of the AI tools making false claims as well as writing incorrect code (in which you would also lose points for correctness), using AI tools to generate technical content or code can make the reader of your report/code more skeptical that you understand the claims/code that has been written.

Other reasons why ChatGPT generated content in a report can lead to a decline in report professionalism are the following.

- The report is written in a style/to an audience that is not our intended target audience (ie. a boss/client/researcher that has the same STAT207 knowledge as you). For instance, your report should not read as if you are educating the reader on a new topic that they are not familiar with.
- The report injects superfluous, off-topic sentences/code that are not relevant to the target reader and your overarching research motivation. This can make your report less concise, readable, and clear.
- The report discusses broad potential generalities, rather than address what you know about the actual dataset that you are exploring.

What's not ok

- GENERATING code/content with AI tools will lose points.

Project Format

Project Report [27.5 pt]

Deadline: Friday, February 21 11:59pm CST on Canvas.

Should contain: Everything stipulated in the **Project Report Specifications** discussed below.

Format:

- Jupyter notebook.
- This should look like a clean data analysis report that you would theoretically submit to an employer (not a homework assignment). Thus, at the very least, your report should have:
  - a title
  - headings for each of your sections
  - You should write paragraphs and in complete sentences.
- You can use and modify the attached project **Mini_Project_1_YOURNAMEHERE.ipynb** file as a template for this report if you'd like. You can add and delete as many cells as you'd like in this file.

Graded:

- See "Project Report Specifications" section below for point breakdown.

Peer Evaluation [7.5 pts]

Deadline: Friday, February 28 11:59pm CST on Canvas.

- **Purpose**:
  - For report writers:
    - The purpose of this final part of the project **report writers** is to give constructive feedback on:
      - how **clearly** you were able to communicate and answer your research questions with your analyses
      - how well you were able to **motivate** your research to a peer, and
      - how **reproducible** your analysis was.
  - For readers:
    - The purpose of this final part of the project **for report readers** is to **get ideas** as to how to make your own report delivery better.
- **Steps**:
  - Ater you submit your report, you will be randomly assigned to read another person's report.
  - After reading their report you will fill out a survey form on **Canvas**, which will ask you the following questions (see last page of this document).
  - **The person that you evaluated will see the constructive feedback that you give.**
  - If you are unclear about how to answer the questions in this document, you are encouraged to reach out to the person that you were assigned to for clarification.
- **Graded**:
  - For completeness

Dataset Options

You can choose your own dataset or you can use the supplied dataset discussed in the next page. The csv for this dataset is located in the same folder that this document is in. There is more information about each of these datasets below.

There are several places you can go to to find interesting datasets, but here are some places you can start.

https://www.kaggle.com/datasets

https://corgis-edu.github.io/corgis/csv/

https://archive.ics.uci.edu/ml/datasets.php

https://data.world/datasets/regression

https://github.com/fivethirtyeight/data

For students interested in sports data:

- NFL: https://www.nflfastr.com/
- MLB and other baseball: https://billpetti.github.io/baseballr/
- CFB: https://saiemgilani.github.io/cfbfastR/index.html
- More sports stuff: https://sportsdataverse.org/

Choosing your Own Dataset

If you decide to choose your own dataset, it must meet the following specifications.

1. It must have **at least 50 rows.**
2. It must have at least **3 meaningful variables**. That is, three variables that are either categorical or numerical.
3. You will choose to explore three variables in this dataset.
    a. **Y**: One of your variables must be **numerical**.
    b. **Z**: One of your variables must be **categorical**.
    c. X: The other variable can be **categorical or numerical.**
4. Your categorical variables should not be something with a distinct value for every row (like name/userid/etc).

Pre-Selected Dataset Option

1. **Video Games Dataset** Originally collected by Dr. Joe Cox, this dataset has information about the sales and playtime of over a thousand video games released between 2004 and 2010. The playtime information was collected from crowd-sourced data on “How Long to Beat”.
   Some more information can be found here.

   This is where Dr. Ellison downloaded this csv file from on 9/8/2023.

Project Report Specifications

Your report should include the analyses, code, and explanations detailed in each of the following sections.

1. Introduction You should write an introduction (about a paragraph) for your report. Your introduction paragraph should include/incorporate the following things.	
Professionalism * Paragraph, written in complete sentences. * Written in a markdown cell, not a code cell * Well explained	2.5
Research Question Clearly state the research question that you intend to answer in the report. * "How does the nature of the relationship between `x` and `y` change for different values of `z` in the dataset?" (Fill in the x, y, z with the variables that you selected).	0.5
Research Motivation * Clearly state the motivation for why you are seeking to answer this particular research question. * Describe at least one person (or type of person) who may find the answers to this research question useful. * How might this person USE the answers to this research question?	1.5
2. Dataset Discussion You should write a paragraph in your report discussing your dataset(s) that you will be using to answer these research questions. This paragraph should include/incorporate the following things.	
Professionalism * Paragraph, written in complete sentences. * Written in a markdown cell, not a code cell. * Well explained	0.5
Dataset Display * Read your csv file and display the first 5 rows of your dataframe. * How many rows are in your dataframe (originally before any data cleaning)?	0.5
Dataset Source * State where YOU got this csv file (dataset) from. * Provide a link/reference to where it came from. * State when you downloaded this csv file.	1

Original Dataset Information In the place where you found this dataset, try to answer the following questions. If the source does not give the answer to these questions, say so. * What do the rows (ie. observations) represent in this dataset? * How was this dataset collected? * Is this dataset inclusive of ALL possible types of observations that could have been considered in this dataset? If not, what types of observations might be left out? * How does your answer to the question above impact the types of actions that the person in your research motivation might take based on the answer to your research questions? * Describe the three variables you intend to explore in this analysis. What do they represent?	2.5
3. Dataset Cleaning You should show and discuss any dataset cleaning decisions that you made in this section.	
Professionalism * Your written discussion in this section should be written in complete sentences. * Written in a markdown cell, not a code cell. * Well explained	0.5
Missing Value Detection and Cleaning * Does your dataset have any IMPLICIT or EXPLICIT missing values? Demonstrate in the code whether it does or does not. * If it does have IMPLICIT missing values, what strings represent these missing values? * Deal with these missing values (implicit and explicit) using one of the techniques we discussed in class. If you dropped rows, how many rows did you drop? * Evaluate the pros and cons of using this missing values cleaning technique that you just used.	2
Sample Size Cleaning * If JUST ONE of your three variables was categorical (Z), then you should make sure that every DISTINCT VALUE of this categorical variable has AT LEAST 10 observations in the dataset that correspond to it. If not you should drop all rows that correspond to DISTINCT VALUES that have less than 10 observations. * * If TWO of your three variables were categorical (X and Z), then you should make sure that every COMBINATION OF DISTINCT VALUES of X and Z has AT LEAST 10 observations in the dataset that correspond to it. If not you should drop all rows that correspond to these groups that have less than 10 observations.	1
Outlier Cleaning - Single Variable Outlier Inspection * Create a boxplot for every numerical variable that you have selected to explore. * If you detect any outliers in these boxplot(s), evaluate the pros and cons of dropping these outliers, when it comes to answering your research questions. * If you chose to drop the outliers, do so. * How many rows did you drop?	2

Outlier Cleaning - Two Variable Outlier Inspection * If you have two numerical variables that you're exploring, create a scatterplot of these two numerical variables. * If you detect any outliers in this scatterplot, evaluate the pros and cons of dropping these outliers, when it comes to answering your research questions. * If you chose to drop the outliers, do so. * If you dropped rows, how many did you drop?	2
Other Data Cleaning * When you tried to answer your research questions below, did you discover any other data cleaning ideas that might help make the answer to your research question more clear? What were they? Why did you choose to perform this additional data cleaning? * If there are, do so here. * If you dropped rows, how many did you drop?	0.5

4. Research Question

You should answer your research question B here. "How does the Relationship between `x` and `y` Change based on Different Values of `z` in the Dataset?"

Professionalism * Your written discussion in this section should be written in complete sentences. * Written in a markdown cell, not a code cell. * Well explained	0.5
Research Question Statement * Clearly state your research question that you intend to answer at the beginning of this section.	0.5
Visualization * Plot the appropriate visualization that will help you answer this research question. * Visualizations should have an appropriate title. * Visualizations should be readable.	2.5
Summary Statistics * Use the appropriate summary statistic(s) that will help you answer this research question.	1.5
Hint: If your `x` and `y` are numerical and `z` is categorical * There are five things that we discussed in class that you should be prepared to discuss when answering this question. You should discuss and calculate them here.	
Hint: If your `x` and `z` are categorical and `y` is numerical * There are four things that we discussed in class that you should be prepared to discuss when answering this question. You should discuss and calculate them here.	
Research Question Answer * Clearly state the answer to your research question at the end of this section.	0.5

5. Conclusion

You should answer your research question B here. "How does the Relationship between `x` and `y` Change based on Different Values of `z` in the Dataset?"

Professionalism * Your written discussion in this section should be a paragraph written in complete sentences. * Written in a markdown cell, not a code cell.	0.5
Summarization * Summarize your research answer here.	1

Shortcomings/Caveats * Discuss any shortcomings to your analysis here (all analyses have SOME shortcomings). * In particular, how might these shortcomings impact how the person in our research motivation might USE the answer to our research question?	1.5
Future Work * Based on what you observed in your analysis, what is one idea you might have for future work?	1.5
6. Peer Evaluation Feedback After submitting your report, you will be randomly assigned to another student (and vice versa) to provide feedback on their report.	
See the Canvas quiz questions.	7.5
Total	**35**

STAT207 Peer Evaluation Questions [7.5 points]

Deadline: **Friday, February 28 11:59pm CST** on **Canvas**.

These questions will be posted on a Canvas quiz for you to submit.

Motivation

1. What is the **motivation** for the analysis in this report? Or in other words, why should you (or someone else) care about the analysis that you just read?

Research Question/Answer

2. What was their **research question**, and what was the **answer** to their research question?

Usefulness

3. For the person/type of person that they talked about in their motivation, what do you think would be their **"next steps"** upon learning the answer to their research question?

Correctness

4. Did you catch any **code or interpretation errors** in this analysis? If so, what did you catch?

5. **How confident** are you that you have caught all the code/interpretation errors in this analysis (if any)? Explain your answer. **Describe your process for how you went about looking for errors.**

Robustness

6. Name at least one **step/decision/interpretation** that this person made in their report in which you could envision another data scientist doing something different. Why do you think that this other data scientist might have done something different?

Transparency

7. Were there any analysis decisions made in this report, in which they made a particular decision, but **did not explain** either:
    - that they **made a decision** or
    - **why they made that particular decision?**

   If so, what were they?

8. Are there any **shortcomings** that you can think of (the report should have mentioned some) in which their analyses may not have provided *perfect* answers to their research questions.

Clarity/Readability

9. Roughly how long did it take for you to come up with an answer to each of questions (1)-(8) above?
10. What are some **tips** that you would give to this researcher for how you might have been able to come up with answers (1)-(8) more easily/more quickly?