Project 1
STAT 207 - Data Science Exploration

Due: Tuesday, February 20 by 11:59 pm on GitHub

Main Goal of Analysis

The main goal of this project is to prepare a dataset, perform an exploratory analysis, and to communicate your results clearly.

You are required to perform two main analytical tasks:

1. Data Cleaning
2. Summarizing Data Research Question

Additional descriptions for these tasks can be found later in this document. If you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is required in this document.

Project Format

Your project will be submitted as a written report to GitHub. This report is worth 40 points. Additional details for the report can be found below.

Group Structure

You can work in groups of up to 3 people, or you can work individually.

- If you work with a group of 3, you must do at least 25% of the work in order to get full credit.
- If you work with a group of 2, you must do at least 33% of the work in order to get full credit.

Dataset Options

You can choose your own dataset, or you can choose from one of the three supplied datasets below. There are several places you can go to find interesting datasets, but here are some places to start:

- https://www.kaggle.com/datasets
- https://corgis-edu.github.io/corgis/csv
- https://archive.ics.uci.edu/ml/datasets.php
- https://github.com/fivethirtyeight/data
- For sports data, you may choose to explore (nflfastr.com for NFL, billpetti.github.io/baseballr for MLB, cfbfastr.sportsdataverse.org/index.html for CFB, and sportsdataverse.org for more sports data).

Choosing your own data:

If you choose your own data, it must meet the following specifications.

1. It must have at least three variables total (five for later projects)
    a. Variables that have uninformative information don't count and won't be useful. Examples of uninformative variables include those that provide the observational unit (a row name or row id) or are a linear combination of other variables in the datasets. If you aren't sure, come ask!
2. It must have at least one categorical variable
3. It must have at least one quantitative variable
4. It must have at least 50 rows

Provided dataset:

1. Video Games Data (video_games.csv)
    a. 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"
    b. This was originally collected and curated by Dr. Joe Cox
    c. This data was originally downloaded on 2/8/2024 from here: https://researchportal.port.ac.uk/en/datasets/video-games-dataset
    d. Read more about this data here: https://researchportal.port.ac.uk/en/publications/what-makes-a-blockbuster-video-game-an-empirical-analysis-of-us-s

Project Report Specifications

Deadline: Tuesday, February 20 by 11:59 pm on GitHub
Format:

- Jupyter notebook
- This should be a clean data analysis report that you could submit to an employer or client (not a homework assignment). At the very least, your report should have a title, headings for each section, and be written in paragraphs and with complete sentences.
- You can use and modify the attached project_01_template.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.

1. **Introduction [4 points]**
    a. Title: Give your research report a title
    b. Goal: in your introduction, you should orient the reader to what they are about to read. This will help to prepare the reader and figure out how to connect the different components that they will read.
    c. Motivation: Describe the motivation for why you or someone else would want to explore your dataset (or a dataset of this type). Describe one person or the type of person who would find these insights useful. You can give background research (with citations) if this would help back up your motivation.
    d. Research Questions: You will answer at least one **set** of research questions. In your introduction, you should briefly describe why you (or someone else) would be interested in answering these research questions. How could the answers to these research questions be used?

2. **Dataset Discussion [10 points]**
    a. Goal: In your dataset discussion, you will introduce your reader to your data and discuss some of the limitations. You should write a paragraph to describe these ideas, including built-in code cells and output.
    b. Source: You should record where your data came from, including a link or reference, how you accessed the data, when you accessed the data, and if a codebook is available.
    c. Dataset: Display at least 6 observations from your dataframe in this section and show how many rows and columns it has.
    d. Unit of Observation: Define what the unit of observation is for your data. While it could be recorded as a column in your dataset, it also may not appear explicitly in your data.
    e. Type of Data: Is this data a sample or a census? Is it representative of the population of interest (as far as you can tell). Explain.
    f. Data Collection: How was this dataset collected?
    g. Variables: Describe the three variables you intend to explore in this analysis.
    h. Think Critically: What limitations exist within your data? What information is included in your data? What information (observations or variables) is not included in your data? What might you want to add? Are there any practical or

ethical considerations for this data? You don't need to answer all of these questions, but you should answer some.

3. **Dataset Cleaning [10 points]**
    a. Goal: You should prepare your data for your future analyses and communicate these steps clearly.
    b. Missing values: Does your data have any missing values automatically detected by Python? Does it have any implicit missing values? Be sure to check for your variables of interest in the data. Instruct Python to recognize all missing values as such.
        i. If you choose to drop missing values, do so. How many rows did you drop?
        ii. Describe the advantages and drawbacks of treating missing values as you have.
    c. Uncommon Values: Inspect your data to determine if there are uncommon values for any of your variables.
        i. For categorical variables, observe if you have any uncommon values. Attempt to combine these uncommon levels with other, similar levels, if possible.
        ii. For quantitative variables, observe the distributions. Are there observations that you are concerned are not accurate? Explain.
        iii. If you choose to drop observations, do so. Record how many observations you dropped.
        iv. Explain why you made your decision and any limitations that will be associated with your later analyses.
    d. Other Data Cleaning: Did you encounter any additional data cleaning needs as you pursued your next research question? If so, explain what you discovered, any decisions that you made, and any important limitations or clarifications that these decisions introduced for your results.

4. **Summarizing Data Research Question Set [10 points]**
    a. Goal: You should pick at least two variables (at least one categorical) and explore the relationship between these variables *in the dataset* (descriptive analytics).
        i. For instance, you could ask "What is the relationship between x and y in this dataset?"
        ii. As a challenge, you may ask "How does the relationship between x and y change based on different values of z in this dataset?"
    b. State your research question that you will answer with your analysis. Remember, descriptive analytics only involves describing relationships in the dataset that you have, so your first research question should be *just* about the data.
    c. Generate at least one appropriate visualization to observe the behavior of your variable(s) of interest in the data. The visualization should be clear and easy to read. Describe what you see in your visualization, what it tells you, and how it

helps answer your research question. Be sure to completely describe any distributions.
   d. Calculate one appropriate set of summary statistics that can be used to supplement your visualization.
   e. Finally, discuss how your conclusion answers your research question.

5. **Conclusion [4 points]**
   a. Summarization: Summarize your data basics, data cleaning, and research question in the conclusion. Provide at least a paragraph. (This will likely be a restatement of what you have already included in your report).
   b. Limitations: What limitations did you face in your analysis, results, or interpretations? What challenges did you face in your data analysis? What contextual information is important before you make strong claims from these results? How might these affect how the person you described in the introductory motivation paragraph uses these results?
   c. Future work: If you (or someone else) were to conduct future work based on these analyses, what kind of research questions or analyses might that entail?

The remaining 2 points of this project will be graded on writing quality, clarity, conciseness, and professional and neat formatting of the report.

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 STAT 207 classmates. **Theoretically, you should be able to send your report to one of your classmates (who is not on your team), and they should be able to understand everything that you did and the claims that you are making.**

Grading

In addition to being graded for correctness and completion (as noted), 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 a 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.
        - Know why you made the choices that you did in your analyses.
- **Clarity about Motivation (i.e. the "so what?") of your analyses**
    - Beginning of the Report
        - Someone who is **about to** read the body of your report should be able to clearly answer the questions:
            - Why should I (or someone else) care about the report that I am about to read?
            - What research questions do they intend to answer?
            - How do these research questions relate to their motivation?
        - Therefore, in the introduction of your report you should make this clear.
    - Middle of the Report:
        - While **in the middle of** your report, 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, 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 should be able to clearly answer the questions:
            - Why should I (or someone else) care about the analysis that I just read?
            - Did their analyses and conclusions answer the research questions that they stated at the beginning of the report? 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 you should make this clear.