CISC 3225: Final Project
100 Points
Due May 15, 2025 at 11:59 PM

Introduction

The purpose of this project is to produce a portfolio-quality analysis of a dataset using techniques from class. Aside from a few fundamental requirements shown below, you have significant latitude to ask whatever questions you feel are appropriate about your data, and to answer them using any technique you want.

1. Dataset

Identify a dataset of suitable complexity for an in-depth analysis. Your dataset should have approximately the following characteristics, with some room for variation depending on interest, available data, and your plans for the project:

- At least six major variables, including:
  - 3 or more continuous variables (price, population, age, dimensions, rating, etc.)
  - 3 or more categorical variables (species, product type, political party, home state, etc.)
- Ideally, you should have some domain knowledge about the dataset. If not, you can familiarize yourself with the domain where necessary to explain any observations or insights.

Dataset sources:

- Kaggle
- Datasets for Data Cleaning Practice
- Social Security data
- NYC OpenData
- UCI Machine Learning Repository

2. Exploratory Analysis

Conduct an exploratory analysis of the data. The analysis should include:

- If your dataset has missing values, identify and explain them. If your analysis requires you to handle the missing values, describe your strategy for doing so.
- Numeric variables:
    - Mean, min, max, median
    - Correlations between all continuous variables
    - Visualize data distribution, noting outlier values
- Categorical variables: Value counts with bar charts

3. High-level analysis

Perform at least 6 higher-level analyses of your data. You are free to use any techniques we discussed in class, including but not limited to:

- Use Pandas features to answer specific questions about the data
- Perform a cluster analysis to identify groups within your data
- Identify and motivate a machine learning problem in your data (classification or regression). Create a train/test/validation split and evaluate how well an appropriate model performs
- Perform a linear regression to show the relationship between two variables

If applicable to an analysis, you **must** include:

- Appropriate statistical test(s)
- An appropriate visualization.

Please take advantage of the check-ins or office hours if you are unsure whether a visualization or statistical test is necessary for an analysis.

4. Final Report

Compile your results into a written report submitted separately as a PDF, Word document, or other appropriate text format. Do not include code in the report unless absolutely necessary. Your report should use the following structure:

1. Introduction: Describe your dataset. What is its purpose and what kind of data does it contain? What do you hope to discover in your analysis?
2. Exploratory analysis. Describe the characteristics of the data you observe, with visualization to support your observations. Use domain knowledge to explain interesting observations, citing external sources if necessary.

3. High-level analysis. Introduce each of your analyses and present them, with relevant visualizations, in their own sections.
4. Conclusions. What did you learn from this project? End with a thoughtful discussion of the data and insights you obtained from your analysis, and draw conclusions.

Check-ins

Three project check-ins will be conducted before the end of the semester and are intended to give you feedback on work you have completed so far. **Each check-in is worth 10 points of your final project grade**. More details will be posted on Blackboard.

Submission

Submit your work before 11:59 on the deadline. **Late submissions will not be accepted**. Your submission must include your written report in a suitable format (PDF, Word, LibreOffice, etc.) and all notebooks (in .ipynb format) used to produce results used in the paper. **All code must be executable.** You may include code in your written report if you feel it is useful to do so, but it must be excerpted from the notebook included with your submission.