CSCI 347 Project 1: Exploratory Data Analysis

Partner work is allowed on this project.

Browse through the UCI Machine Learning Repository to find a data set that is interesting to you, and has both categorical and numerical data (but is small enough to work with).

Note: If your chosen data set is not small enough to download, take a sample of the data set that contains at least 200 entities (instances/rows).

Part 1: Write an introduction.
In a well-written paragraph, answer the following questions about the data:

- [4 points] What was the data used for?
- [2 points] Who (or what organization) uploaded the data?
- [5 points] How many attributes and how many entities are represented in the data?
  - How many numerical attributes?
  - How many categorical attributes?
    * Would you suggest that each categorical attribute be label-encoded or one-hot-encoded? Why?
- [4 points] Are there missing values in the data? If so, what proportion of the data is missing overall? What proportion of data is missing per attribute (you may use a plot or table to summarize this information)?
- [7 points] Why is this data set interesting to you?
- [6 points] Of the attributes used to describe this data, which do you think are the most descriptive of the data and why (before doing any data analysis) ?

Part 2: Write Python code for data analysis.
Use Python to write the following functions, without using any functions with the same purpose in sklearn, pandas, numpy, or any other library (though you may want to use these libraries to check your answers):

- [5 points] A function that will compute the mean of a numerical, multidimensional data set input as a 2-dimensional numpy array
- [5 points] A function that will compute the sample covariance between two attributes that are input as one-dimensional numpy vectors
- [5 points] A function that will compute the correlation between two attributes that are input as two numpy vectors.
- [5 points] A function that will normalize the attributes in a two-dimensional numpy array using range normalization.
- [5 points] A function that will normalize the attributes in a two-dimensional numpy array using standard normalization.
- [5 points] A function that will compute the covariance matrix of a data set.
- [5 points] A function that will label-encode a two-dimensional categorical data array that is passed in as input.

Part 3: Analyze the data with your code and write up the results.
Use your code from Part 2 to answer the following questions in a well-written paragraph, and create the following plots from the numerical portion of the data:

Use your functions to compute the multi-variate mean and covariance matrix of the **numerical portion** of your data set.

- **Before answering the questions:**
  - [5 points] Convert all categorical attributes using label encoding or one-hot-encoding
  - [2 points] If your data has missing values, fill in those values with the attribute mean.

Questions to answer:
- [2 points] What is the multivariate mean of the numerical data matrix (where categorical data have been converted to numerical values)?
- [4 points] What is the covariance matrix of the numerical data matrix (where categorical data have been converted to numerical values)?
- [5 points] Choose 5 pairs of attributes that you think could be related. Create scatter plots of all 5 pairs and include these in your report, along with a description and analysis that summarizes why these pairs of attributes might be related, and how the scatter plots do or do not support this intuition.
- [3 points] Which range-normalized numerical attributes have the greatest sample covariance? What is their sample covariance? Create a scatter plot of these range-normalized attributes.
- [3 points] Which Z-score-normalized numerical attributes have the greatest correlation? What is their correlation? Create a scatter plot of these Z-score-normalized attributes.
- [3 points] Which Z-score-normalized numerical attributes have the smallest correlation? What is their correlation? Create a scatter plot of these Z-score-normalized attributes.
- [3 points] How many pairs of features have correlation greater than or equal to 0.5?
- [3 points] How many pairs of features have negative sample covariance?
- [2 points] What is the total variance of the data?
- [2 points] What is the total variance of the data, restricted to the five features that have the greatest sample variance?

- Your submission must include the complete written report and all associated code. Ensure that your code is well-commented to facilitate understanding of the methods used to generate your results.
- The full names of all members should be included in the first page of the report as well as name of the file (it can include only last names).
- The team size must be between 2-4 members.
- Each team is required to submit their report and code only once per team, and the same team member should submit on both D2L and GradeScope platforms. Multiple submission by same person is allowed.
- The submission on GradeScope will be the primary document for grading purposes.
- To clarify individual contributions, please prefix your initials before each question (or part) to indicate who worked on each problem.
- Please be aware of the following points, as failure to comply will result in a deduction of points:

    teams with fewer than 2 or more than 4 members.
    Missing partners name on either the report or file name
    Multiple submissions by the same team, even if identical.
    Different versions of the report submitted by various team members
    If your report does not include the necessary code, or if the code provided is not appropriately attached as a supplementary document or as inline code following each question.
    Reports that are disorganized, unreadable, or contain excessive unrelated code
    Missing initials for each question.