CptS 475/575: Data Science, Fall 2024

Assignment 2: R Basics and Exploratory Data Analysis

Release Date: September 4, 2024   **Due Date**: September 11, 2024 (11:59 pm)

This assignment has **two exercises.** For questions that ask you to produce a specific plot, include that plot along with the code you used to generate it. You are strongly encouraged to use R Markdown (or Quarto) to prepare your solution. Be sure to clearly number each response in line with the questions and give each plot appropriate axis labels and title.

1 (**50 points**). This exercise relates to the Red Wine Quality dataset (*winequality-red.csv*), which can be found under the Datasets modules in Canvas. The dataset contains a number of physicochemical test variables for 1599 different red wine variants of the Portuguese "Vinho Verde" wine. The variables are

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- pH
- sulphates
- alcohol (output variable based on sensory data)
- quality (score between 0 and 10)

Before reading the data into R or Python, you can view it in Excel or a text editor. For each of the following questions, include the code you used to complete the task as your response, along with any plots or numeric outputs produced. You may omit outputs that are not relevant (such as dataframe contents), but still include all of your code.

(a, **6 points**) Use the read.csv() function to read the data into R, or the csv library to read in the data with python. In R you will load the data into a dataframe. In python you may store it as a list of lists or use the pandas dataframe to store your data. Call the loaded data red_wine_data. Ensure that your column headers are not treated as a row of data.

(b, **8 points**) Find the median quality of all the wine samples. Then find the mean alcohol level for all the wine samples.

(c, **8 points**) Produce a scatterplot that shows the relationship between wine density and volatile_acidity. Ensure it has appropriate axis labels and a title. Briefly state if you see any effect of volatile_acidity on density.

(d, **10 points**) Create a new qualitative variable, called ALevel, by binning the alcohol variable into two categories (High and Medium). Specifically, divide the data into two groups based on whether the alcohol level exceeds 10.5 or not (alcohol greater than 10.5 is considered High otherwise it is considered Medium).

Now produce side-by-side boxplots of the ratio of sulphates to chlorides (hint: create a new variable that calculates sulphates / chlorides) for each of the two ALevel categories. There should be two boxes on your figure, one for High and one for Medium. How many samples are in the High category?

(e, **8 points**) Produce a histogram showing the citric_acid numbers for both High and Medium (ALevel) wine samples. You may choose to show both on a single plot (using side by side bars) or produce one plot for High samples and one for Medium samples. Ensure whatever figures you produce have appropriate axis labels and a title.

(f, **10 points**) Continue exploring the data, producing two new plots of any type, and provide a brief (one to two sentence) summary of your hypotheses and what you discover. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

2 (**50 points**). This exercise involves the Bike Sharing dataset (*bikes.csv*) dataset which can be found under the Datasets modules in Canvas. The features of the dataset are:

- date: Date of the observation
- season: Season (1: winter, 2: spring, 3: summer, 4: fall)
- holiday: Whether the day is a holiday (1: yes, 0: no)
- workingday: Whether the day is a working day (1: yes, 0: no)
- weather: Weather situation (1: clear, 2: misty/cloudy, 3: light snow/rain, 4: heavy rain/snow)
- temp: Temperature in degrees Celsius
- atemp: "Feels like" temperature in degrees Celsius
- humidity: Relative humidity in %
- windspeed: Wind speed (km/h)
- count: Count of total rental bikes

(a, **6 points**) Specify which of the predictors are quantitative (measuring numeric properties such as size or quantity) and which are qualitative (measuring non-numeric properties such as type, category, boolean variable, etc.). Keep in mind that a qualitative variable may be represented as a quantitative type in the dataset, or the reverse. Adjust the types of your variables based on your findings if necessary.

(b, **8 points**) What is the range, mean, and standard deviation of each quantitative predictor? Which season has the highest average bike rental count?

(c, **8 points**) Produce boxplots of bike rental counts by weather condition. Your figure should have a boxplot for each weather condition (1 through 4). Which weather condition has the highest median bike rental count?

(d, **10 points**) Produce a bar plot showing the count of rentals for each month of the year. (Hint: You can extract the month from the date variable using the format function in R.) Which month has the highest rentals?

(e, **10 points**) Using the full dataset, investigate the relationships between predictors graphically, using scatterplots, correlation scores, or other tools of your choice. Create a correlation matrix for the relevant quantitative variables.

(f, **8 points**) Suppose that we wish to predict the total count of bike rentals based on the other variables. Which, if any, of the other variables might be useful in predicting the bike rental count? Justify your answer based on the prior correlations.