CptS 475/575: Data Science, Fall 2022

Assignment 2: R Basics and Exploratory Data Analysis

Release Date: September 2, 2022 **Due Date**: September 9, 2022 (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 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 data set (*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 redwine. 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 a relationship between two numeric (not factor or boolean) features of your choice in the dataset. Ensure it has appropriate axis labels and a title.

(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 10.2 or not (alcohol greater than 10.2 is considered High otherwise it is considered Medium).

Now produce side-by-side boxplots of the ratio of sulplates to chorides (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 total_sulfur_dioxide 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 forestfires.csv dataset which can be found under the Datasets modules in Canvas. The features of the dataset are:

- X: x-axis spatial coordinate
- Y: y-axis spatial coordinate
- month: month of the year (‘jan’ to ‘dec’)
- day: day of the week (‘mon’ to ‘sun’)
- FFMC: Fine Fuel Moisture Code index
- DMC: Duff Moisture Code index
- DC: Drought code index
- ISI: Initial spread index
- temp: Temperature in degrees Celsius
- RH: Relative Humidity in %
- wind: Wind speed (km/h)
- rain: Amount of rainfall (mm/m2)
- area: area that got burnt in the forest fire

(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 color, appearance, type etc.), if any? Keep in mind that a qualitative variable may be represented as a quantitative type in the dataset, or the reverse. You may wish to adjust the types of your variables based on your findings.

(b, **8 points**) What is the range, mean and standard deviation of each quantitative predictor? Which day of the week has the highest number of fires?

(c, **8 points**) Now remove the 40th through 80th (inclusive) observations from the dataset. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(d, **10 points**) Produce a bar plot to show the count of forest fires in each month for which wind is greater than 4. During which months are high wind forest fires most common? (Hint: filter data by wind, group data by month and calculate count)

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

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