CS 6220 Data Mining — Assignment 2

Exploring Data with Pandas

Prior to beginning your work on this assignment, download and run this notebook file (`https://goo.gl/BFprVd`), which will cover some basics on data exploration, loading data, extracting basic statistics from the various features, and generating visualizations.

Assignment Description:
This assignment will require that you implement and interpret some of the data understanding concepts that were introduced in class, such as summary statistics and data visualizations. Further, you will be working with real-world data retrieved from an online repository, and while you will be asked to utilize a variety of modules and functions, these have all been covered in the notebook files that were shared. Keep in mind that the main objective of this assignment is to highlight the insights that we can derive from the data understanding process – the coding aspect is secondary. Accordingly, you are welcome to consult any online documentation and/or code so long as all references and sources are properly cited. You are also encouraged to use code libraries, but be sure to acknowledge any source code that was not written by you by mentioning the original author(s) directly in your source code (comment or header).

Submission:
Through the assignment submission portal on Canvas, submit your ipynb with a pdf of your assignment solution; no need to zip the files.

1 Wine Quality Dataset [60 Points]

The Wine Quality dataset is available in the UCI Machine Learning Repository, download the Wine Quality dataset HERE (`https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality`. The dataset consists of two separate datasets that we will combine to create a new combined dataset. One dataset contains red wine samples and the other contains white wine samples. Each sample consists of several physicochemical features (e.g., pH, alcohol content) and a quality rating between 0 and 10.
Use pandas to load both datasets into separate DataFrames. Add a new column to each DataFrame to identify the wine type: red or white. For example, add a column named wineType with a value of "red" to the red wine DataFrame, and add a column named wineType with a value of "white" to the white wine DataFrame.

Concatenate the two DataFrames into a single DataFrame using the pandas. Make sure to set the ignore index parameter to True to reset the index of the combined DataFrame. You now have a single DataFrame containing both the red and white wine datasets combined.

1.1 Summary Statistics [10 Points]

Compute and display summary statistics for each feature available in the dataset. These must include 1) minimum value, 2) maximum value, 3) mean, 4) range, 5) standard deviation, 6) variance, 7) count, 8) 25:50:75 percentiles.

1.2 Data Visualization [25 Points]

Histograms: To illustrate the feature distributions, create a histogram for each feature in the dataset. You may plot each histogram individually or combine them all into a single plot. When generating histograms for this assignment, use the default number of bins. Recall that a histogram provides a graphical representation of the distribution of the data.

Box Plots: To further assess the data, create a boxplot for each feature in the dataset. All of the boxplots will be combined into a single plot. Recall that a boxplot provides a graphical representation of the location and variation of the data through their quartiles; they are especially useful for comparing distributions and identifying outliers.

Pairwise Plot: To understand the relationship between the features, create scatter plot for each pair of the features. If there are n features in the dataset, there should be $nC2$ plots.

Class-wise Visualization: Create pairwise plots for each pair of features in a similar way for each of the different classes present in the data.

1.3 Conceptual Questions [25 Points]

Answer the following questions about the analysis you just performed. Include the answers to these questions as text content (using markdown or text cells on Jupyter notebook) in the same notebook file used for visualization.

1. How many features are there? What are the types of the features (e.g., numeric, nominal, discrete, continuous)?

2. What can you conclude from the histograms about the distribution of the features in the dataset? Are there any features that are approximately normally distributed? Are there any features that are highly skewed?

3. Based on the box plots, are there any features that appear to have many outliers? Are there any features that appear to have a similar spread of values across different quality ratings? Are there any features that appear to have different spreads of values across different quality ratings?

4. Based on the pairwise plots, which features appear to be highly correlated? Are there any features that do not appear to be correlated with any other features?

5. Based on the class-wise visualizations, are there any pairs of features that appear to be more correlated for certain wine types than for others?

2 Forest Fires Dataset [40 Points]

The Forest Fires dataset is a dataset of meteorological and other data from Portugal that is used to predict the size of forest fires. You can download the dataset from the UCI Machine Learning Repository here (http://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv).
The dataset includes information about the location and date of the fire, as well as variables such as temperature, humidity, wind speed, and rain, among others.
A description of this dataset can be found here (http://archive.ics.uci.edu/ml/datasets/Forest+Fires).

2.1 Summary Statistics [5 Points]

Similarly as in Section 1, Compute and display summary statistics for each feature available in the dataset. These must include 1) minimum value, 2) maximum value, 3) mean, 4) range, 5) standard deviation, 6) variance, 7) count, 8) 25:50:75 percentiles.

2.2 Data Visualization [15 Points]

As done in Section 1, create histograms and boxplots for the dataset. Now, create another boxplot without the outliers. You can use `showfliers=False` to remove outliers from the boxplots. You are expected to present two Boxplots in total.

2.3 Conceptual Questions [20 Points]

Answer the following questions about the analysis you just performed. Include the answers to this questions as text content (using markdown or text cells on Jupyter notebook) in the same notebook file used for visualization.

1. From the boxplot without outliers, which features has a significantly different distribution from others? Why?

2. What does the outlier in the features mean? If you remove the outliers from the dataset, what problems might arise?

3. Create a histogram for only FFMC after removing all the values outside of range [88, 96].

4. What distribution does the new histogram follow?