Exploratory Data Analysis

An Analysis on Red Wine Quality and Mushroom Classification Datasets

Maxie Machado
University of Wyoming

1 Introduction

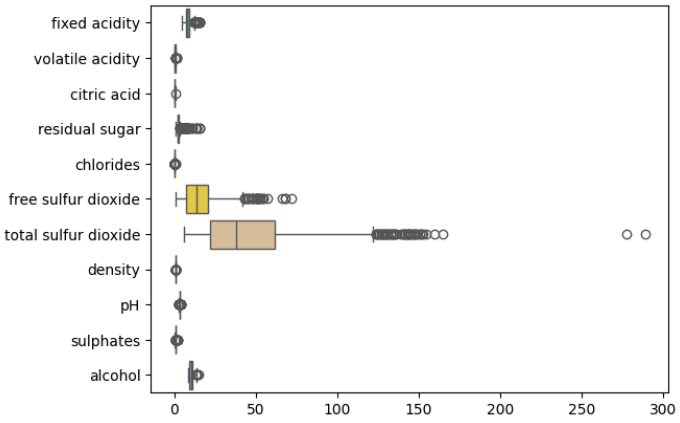
For this exercise the goal was to take two datasets that ideally have different characteristics. Once these datasets are chosen a small portion of preprocessing methods will be chosen to explore the applications of them on the datasets. Both datasets will have the same preprocessing methods used on them to be able to explore the differences of the applications.The first task to be done, once the csv files of the datasets have been imported and read, will be looking at the outliers of the two datasets and the effects of then on the accuracy. Then the task of normalizing the datasets will be done using StandardScaler. Lastly the two datasets will have feature selection done on them using the filter method.

2 Exploratory Data Analysis on Red Wine Quality

The first dataset that was chosen to be explored is the quality of red wine. Specifically the analysis will be on samples from the vinho verde red wine collection, which is sourced from north portugal. This dataset came from UC Irvine Machine Learning Repository.

2.1 Information on Red Wine Quality Dataset

This dataset contains 12 columns and 1599 rows. The characteristics we are looking at to determine the quality of the red wine is the following:

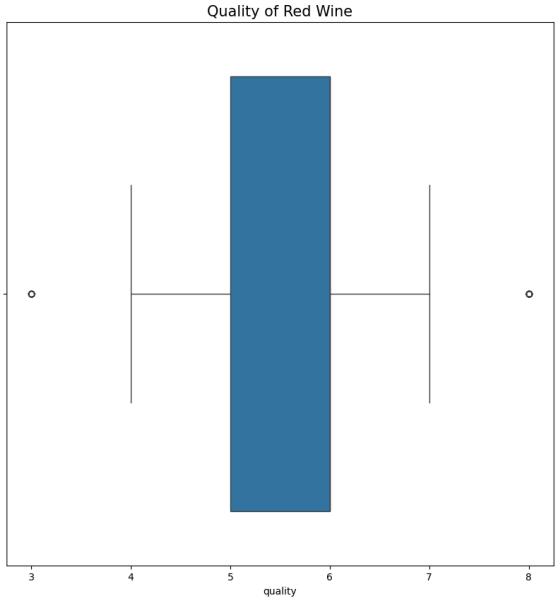


- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- pH
- Sulphates
- alcohol

All of these factors will have been calculated to help determine the quality of vinho verde's collection of red wine. Now before applying any methods to the dataset, viewing it untouched to see what it looks like is desirable so in [fig 1] there will be a simple boxplot visualization of the red wine dataset. As shown the dataset of red wine is messy. Having lots of outliers and too much data to really understand what is going on. With such unclear messy data it is impossible for an onlooker to evaluate this data, let alone a person conducting long strenuous research. Data like this can be cleaned up for easier understanding and a more well-balanced look to it. This is why preprocessing methods must be conducted on the datasets.

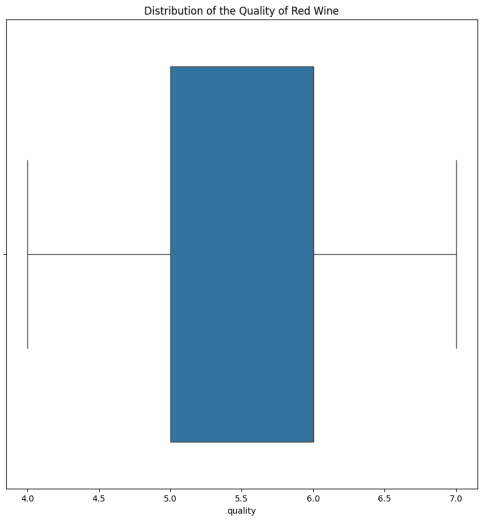
2.2 Processes on Red Wine Quality Dataset

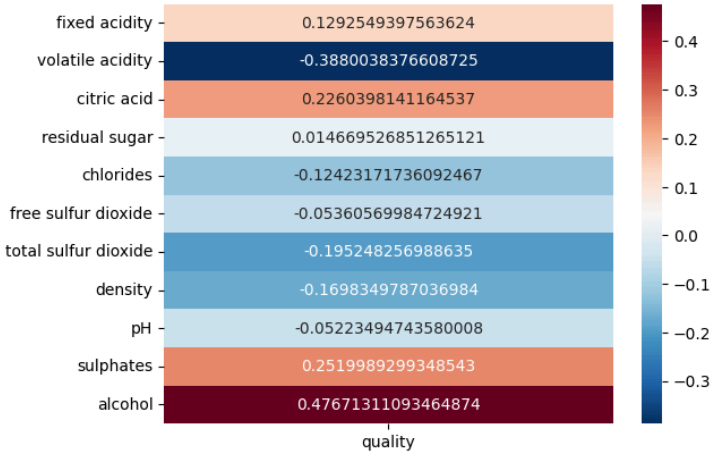
The first preprocessing method that will be done on the red wine quality dataset is analyzing and dealing with the outliers. To do this the z-score method will be used. This task will start by calculating the z-score of the column “quality”.



count	1599.000000
mean	5.636023
std	0.807569
min	3.000000
25%	5.000000
50%	6.000000
75%	6.000000
max	8.000000

Once this is completed we will be identifying the outliers using a z-score greater than three or less than negative three. Again, using a simple boxplot the outliers of quality will be visualized [fig. 2]. Once visualized you can see that the dataset has outliers both less than four and greater than seven. To see it in a more textual way there is a provided statistical summary of the boxplot [fig. 3]. Analyzing the statistical summary of the boxplot shows that the dataset's maximum is eight and the minimum is three, which is including the outliers within the dataset. The mean of the quality of vinho verde’s collection of red wine is 5.636. This opens the question of the accuracy our dataset has having these outliers included within our analysis. With that the next step after identifying the outliers in the red wine dataset is to see the total percentage the outliers take within the dataset. Due to the dataset having outliers on both ends of the dataset the calculations will be done separately. With that in mind, the first step is filtering the data frame to only include outliers with a quality over seven. Once that is completed we will calculate the percentage of the total of the outliers over seven. The calculations demonstrate that 1.126% of total quality is over seven. Which is significant enough to make the accuracy of the data less. Now we will be filtering the data frame to only include outliers with a quality under four. Once that is complete, we will be calculating the percentage of the total of the outliers under four, like how we did with outliers over seven. The calculation once completed demonstrates that 0.625% of total quality is under four. This is a significant enough percentage to question the accuracy of the dataset. With all this in mind now the next step is to handle the outliers to be able to get a more accurate dataset of the quality of red wine. To do this I will be using the winsorize method, applying this method will take the outliers and replace them with the closest non-outlier within the dataset. We will





visualize this using a simple boxplot [fig. 4]. After dealing with the outliers, now the dataset will be normalized using standard scaler. What can be analyzed is the distribution of the dataset with a standard deviation equaling to one. This will help fix the issue with the dataset containing different data types. Lastly but certainly not least the performance of feature selection will be done using the filter method. To visualize this a heatmap will be used [fig.5]. As shown on the heatmap it is clear that certain features positively affect the quality of the red wine and other features negatively affect the quality of the red wine. The features that negatively affect the red wine quality include:

- Volatile Acidity
- Chlorides
- Free Sulfur Dioxide
- pH
- Density
- Total Sulfur Dioxide

Specifically Total Sulfur Dioxide is the feature that mainly negatively affects the quality of red wine, being -0.1925. Now, the features that positively affect the red wine quality include:

- Fixed Acidity
- Citric Acid
- Residual Sugar
- Sulphates
- Alcohol

With this as seen on the heatmap it's clear that the amount of alcohol in the red wine affects the quality of the wine positively the most, being 0.4767.

3 Exploratory Data Analysis on Concrete Compressive Strength Dataset

The first dataset that was chosen to be explored is the concrete compressive strength. Specifically the analysis will be on samples from different types of popular concrete. This dataset came from UC Irvine Machine Learning Repository.

3.1 Information on Concrete Compressive Strength Dataset

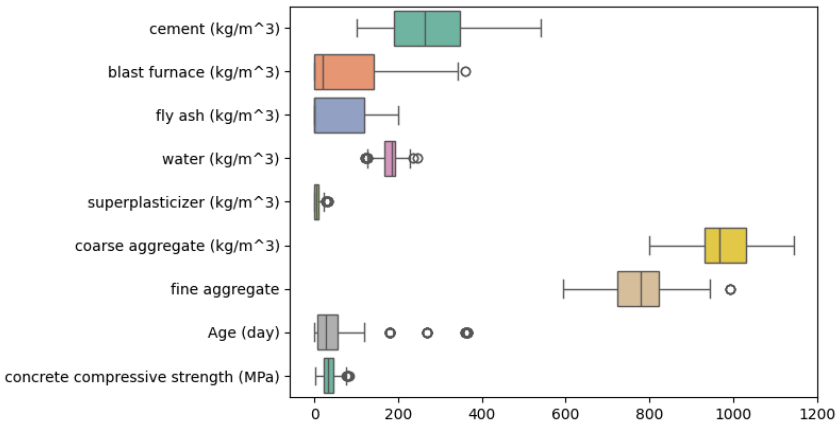
This dataset contains 9 columns and 1039 rows. The different variables that will be analyzed to determine the compressive strength of the concrete is the following:

- Blast Furnace Slag (component 2)(kg in a m^3 mixture)
- Superplasticizer (component 5)(kg in a m^3 mixture)
- Coarse Aggregate (component 6)(kg in a m^3 mixture)
- Fine Aggregate (component 7)(kg in a m^3 mixture)
- Cement (component 1)(kg in a m^3 mixture)
- Fly Ash (component 3)(kg in a m^3 mixture)
- Water (component 4)(kg in a m^3 mixture)
- Age (day)

All of these different variables will have been calculated to help determine the compressive strength of concrete. Although these column names are long and complicated to look at. So before starting any preprocessing methods or even visualizing the untouched data, we will rename the columns. Now what we will be looking at is the following:

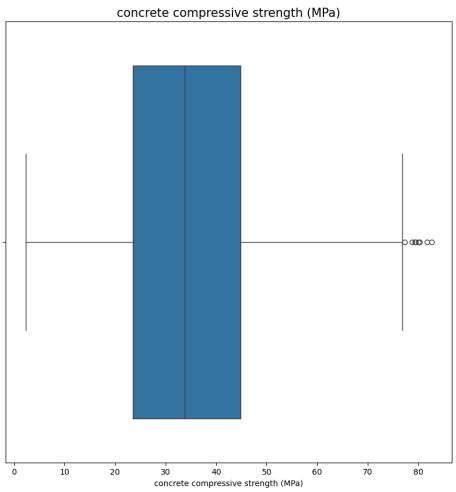
- blast furnace (kg/m^3)
- superplasticizer (kg/m^3)
- coarse aggregate (kg/m^3)
- fine aggregate
- cement (kg/m^3)
- fly ash (kg/m^3)
- water (kg/m^3)
- Age (day)

Now before applying any methods to the dataset, viewing it untouched to see what it looks like is desirable so there will be a simple boxplot visualization of the concrete dataset [fig 6]. As shown the dataset of concrete is slightly messy. Having lots of outliers and too much data to really understand what is going on. With such unclear messy data it is impossible for an onlooker to evaluate this data, let alone a person conducting long strenuous research. Data like this can be cleaned up for easier understanding and a more well-balanced look to it. This is why preprocessing methods must be conducted on the datasets. With this the start of the exploratory data analysis of concrete compressive strength starts now.

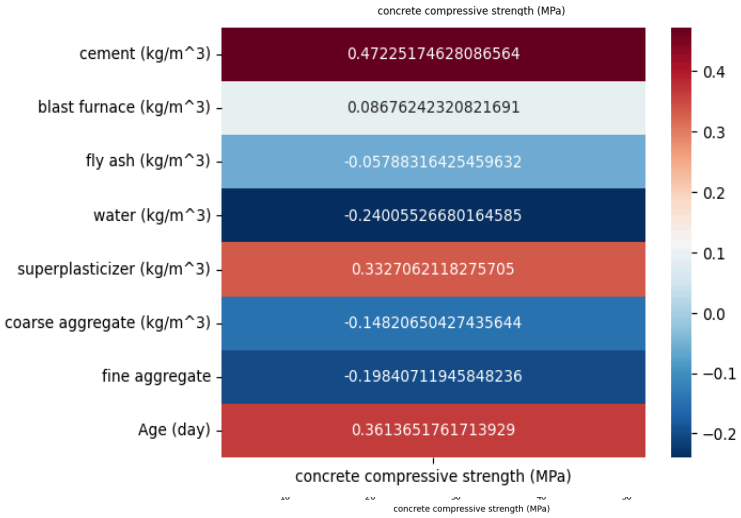


3.2 Processes on Concrete Compressive Strength Dataset

The first preprocessing method that will be done on the concrete compressive strength dataset is analyzing and dealing with the outliers. To do this the z-score method will be used. This task will start by calculating the z-score of the column “concrete compressive strength (MPa)”. Once this is completed we will be identifying the outliers using a z-score greater than three or less than negative three. Again, using a simple boxplot the outliers of quality will be visualized [fig. 7]. Once visualized you can see that the dataset has outliers greater than seventy-eight. To see it in a more textual way there is a provided statistical summary of the boxplot [fig. 8]. Analyzing the statistical summary of the boxplot shows that the dataset's maximum is eighty-two and the minimum is about two, which is including the outliers within the dataset. The mean of the quality of concrete compressive strength is 35.25. This opens the question of the accuracy our dataset has having these outliers included within our analysis. With that the next step after identifying the outliers in the concrete dataset is to see the

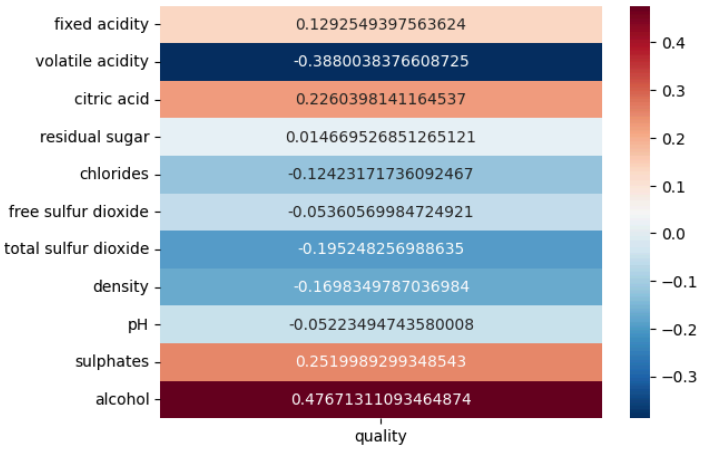


count	1005.000000
mean	35.250378
std	16.284815
min	2.330000
25%	23.520000
50%	33.800000
75%	44.870000
max	82.600000



count	1005.000000
mean	34.322080
std	14.456535
min	6.880000
25%	23.520000
50%	33.800000
75%	44.870000
max	56.810000

total percentage the outliers take within the dataset. In this dataset there are only outliers on one side of the boxplot. Meaning we will only need to do this calculation once. With that in mind, the first step is filtering the data frame to only include outliers with a quality over seventy-eight. Once that is completed we will calculate the percentage of the total of the outliers over seventy-eight. The calculations demonstrate that 0.6965% of total concrete compressive strength is over seventy-eight. Which is significant enough to make the accuracy of the data less. With all this in mind now the next step is to handle the outliers to be able to get a more accurate dataset of the concrete compressive strength. To do this I will be using the winsorize method, applying this method will take the outliers and replace them with the closest non-outlier within the dataset. We will visualize this using a simple boxplot [fig. 9]. After dealing with the outliers, now the dataset will be normalized using standard scaler [fig.10]. What can be analyzed is the distribution of the dataset with a standard deviation equaling to one. This will help fix the issue with the dataset containing different data types. Lastly but certainly not least the performance of feature selection will be done using the filter method. To visualize this a heatmap will be used [fig.11]. As shown on the heatmap it is clear that certain features positively affect the compressive strength of the concrete and other features negatively affect the compressive strength of the concrete. The features that negatively affect the compressive strength of the concrete include:



- Fly Ash (kg/m^3)
- Water (kg/m^3)
- Coarse Aggregate (kg/m^3)
- Fine Aggregate (kg/m^3)

Specifically water is the feature that mainly negatively affects the compressive strength of the concrete, being -0.24. Now, the features that positively affect the compressive strength of the concrete include:

- Cement (kg/m^3)
- Blast Furnace (kg/m^3)
- Superplasticizer (kg/m^3)
- Age (day)

With this as seen on the heatmap it's clear that the cement in the concrete affects the compressive strength of the concrete positively the most, being 0.47.