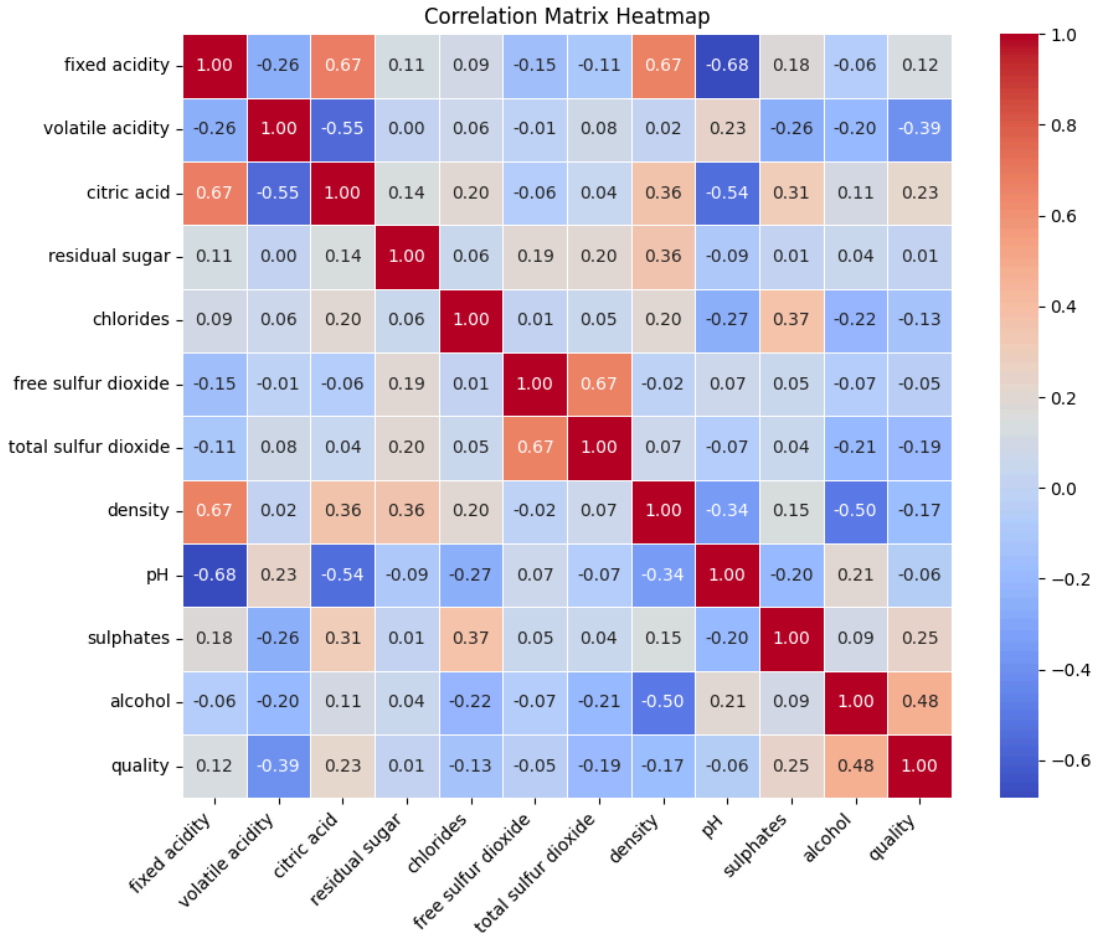
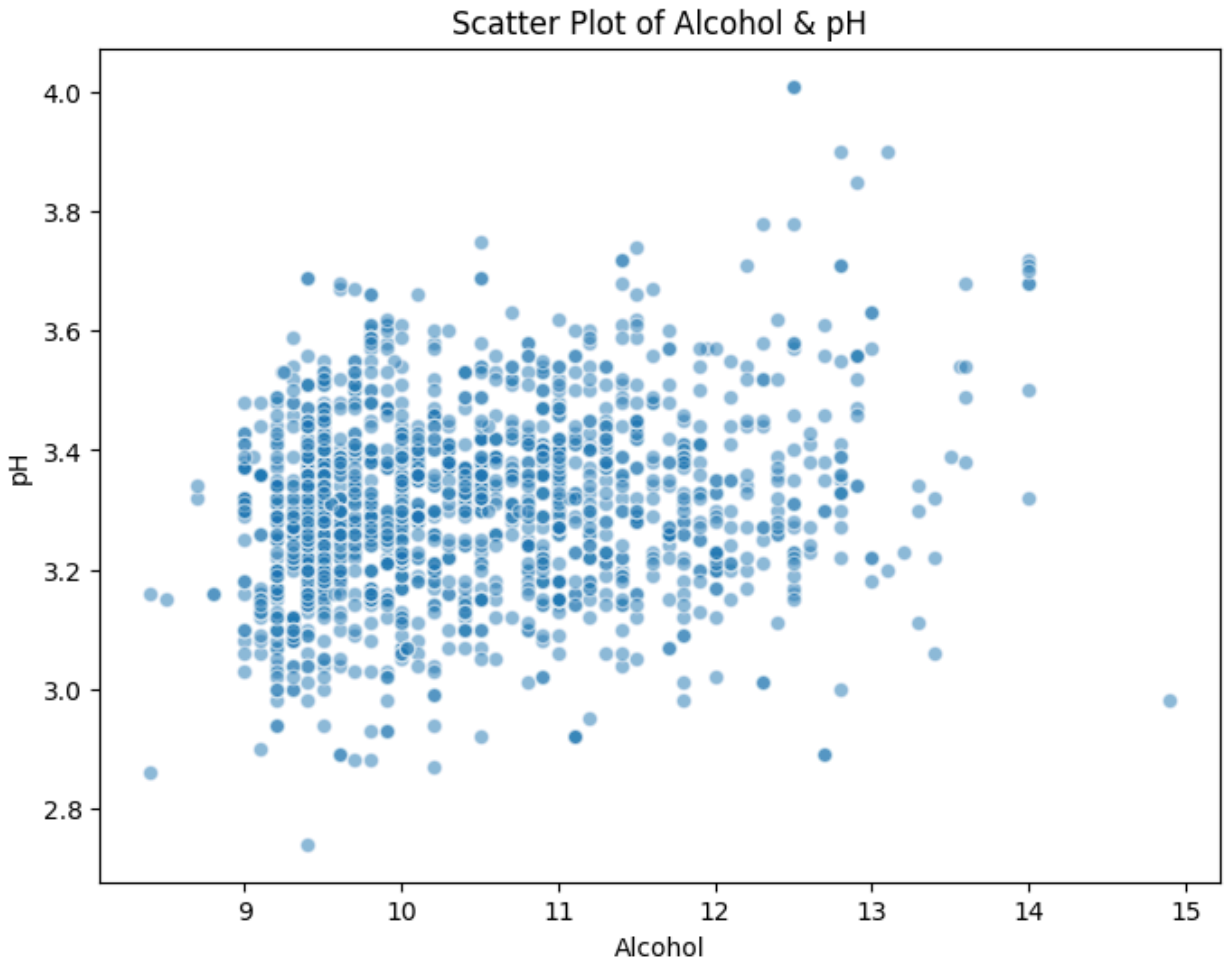
Cosc3337 - hw1 - Report - Minh Nam Nguyen

1) Compute the correlations for each of the pairs of attributes available in the dataset. Interpret the statistical findings.



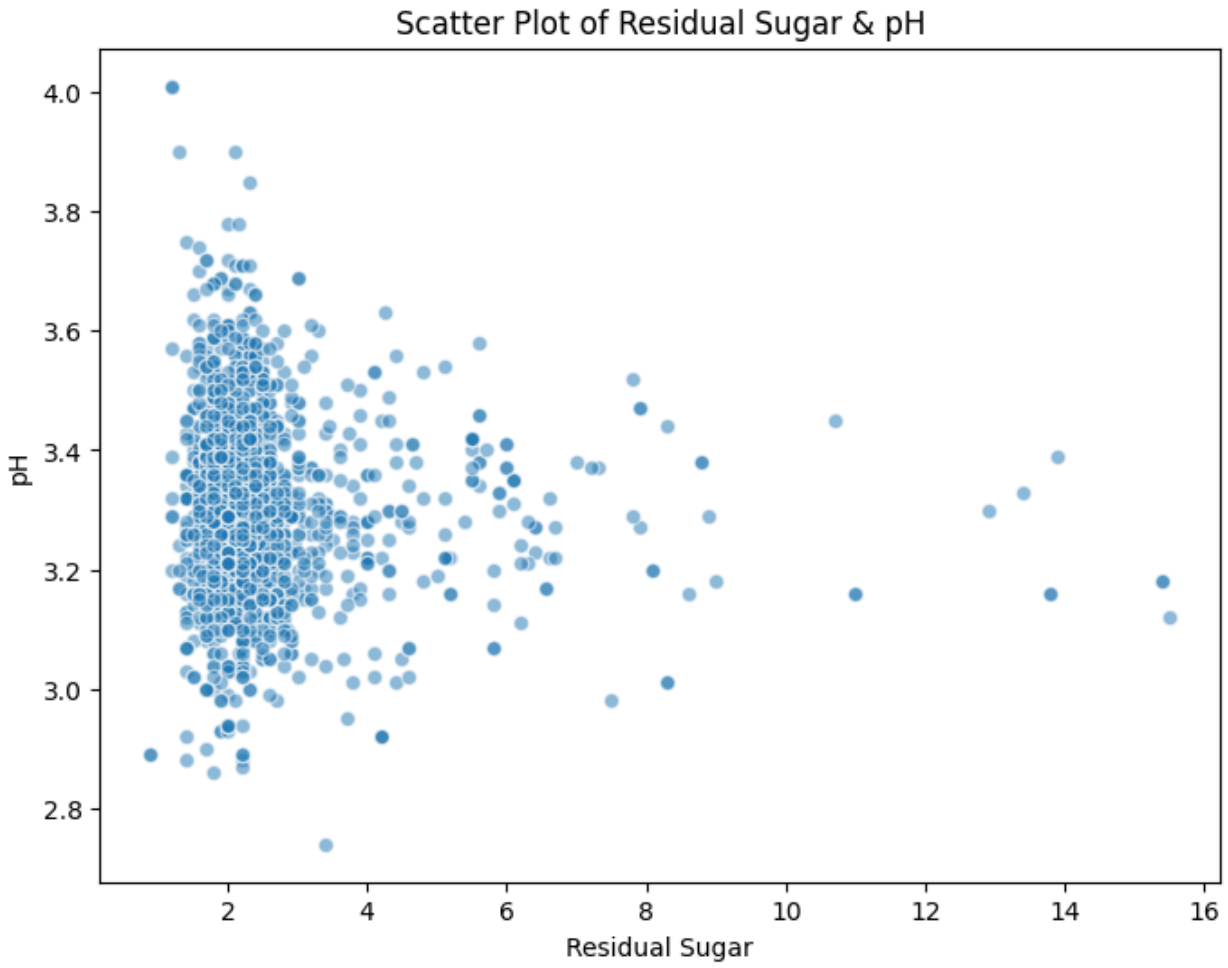
```
* A correlation matrix provides insight to relationships between pairs of attributes. A high correlation means the attribute will tend to move together. A low correlation means the attributes will tend to move in the opposite way of each other.

* Using the colors on the correlation matrix, we can quickly identify which attribute pairs have a low or high correlation.
    * RED   =   HIGH
    * BLUE =   LOW
```

2) Create a scatter plot for the attributes **alcohol** and **pH** and interpret the plot.



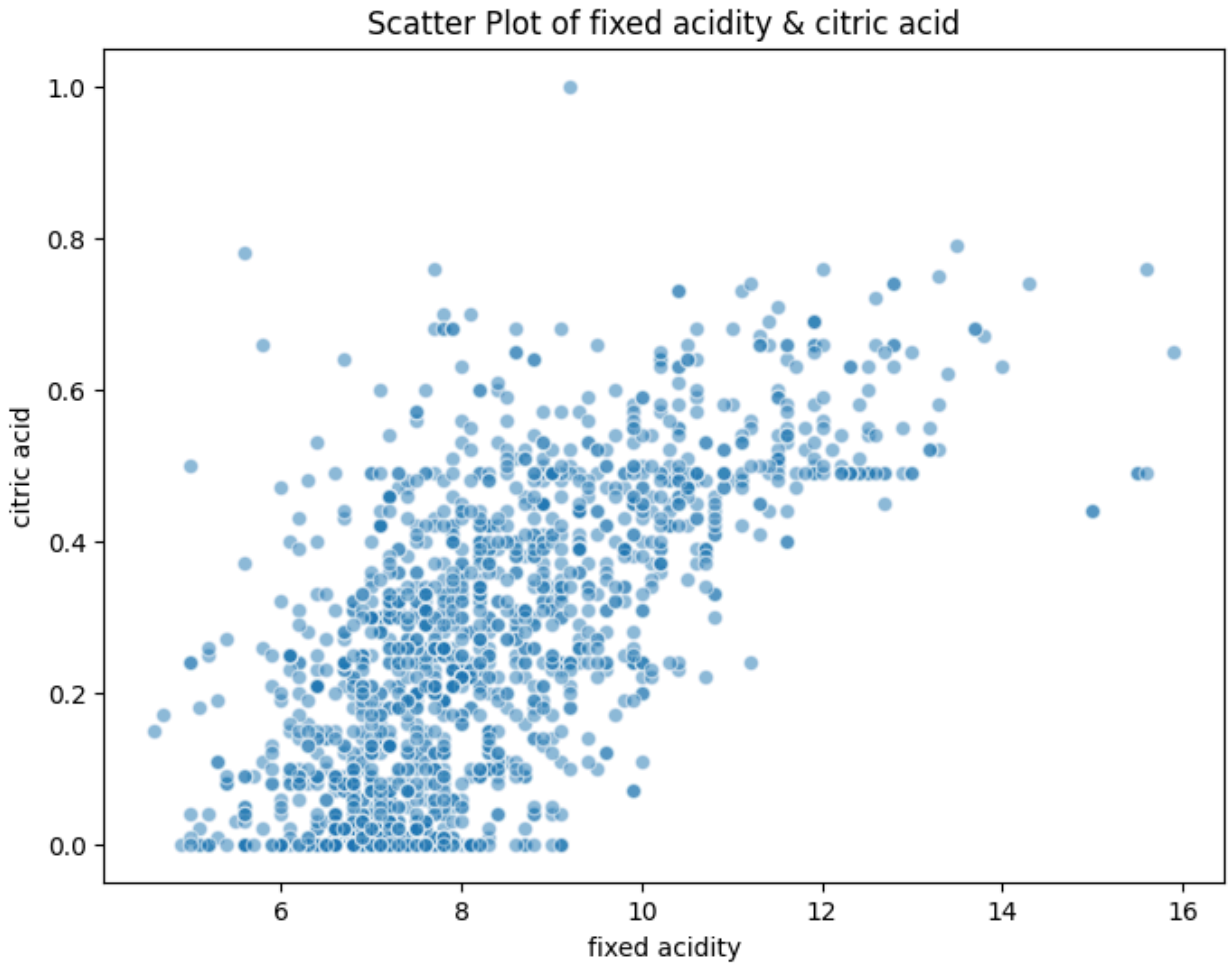
```
* From what I can tell from the scatter plot, alcohol and pH has a slightly positive
correlation. We can see that as the values for both attributes increases, there is a
slight positive trend.


* Correlation Matrix Value = 0.205633
```

3) Create a scatter plot for the attributes **residual sugar** and **pH** and interpret the plot.



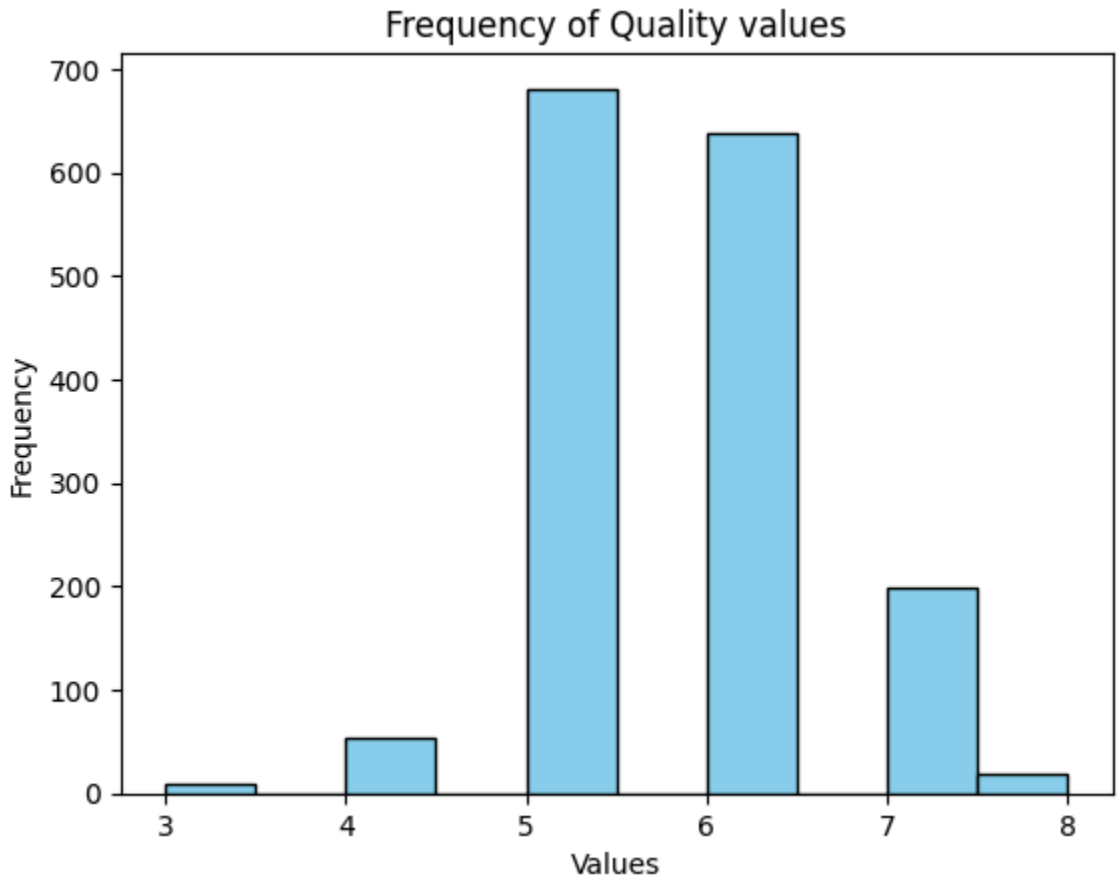
```
* According to the scatter plot for Residual Sugar & pH, the relationship seems weak. The values tend to be opposite from each other but this is very miniscule. From the data, it seems to be a slightly negative correlation

* Correlation Matrix Value = -0.085652
```

4) Create a scatter plot for the attributes **fixed acidity** and **citric acid** and interpret the plot.



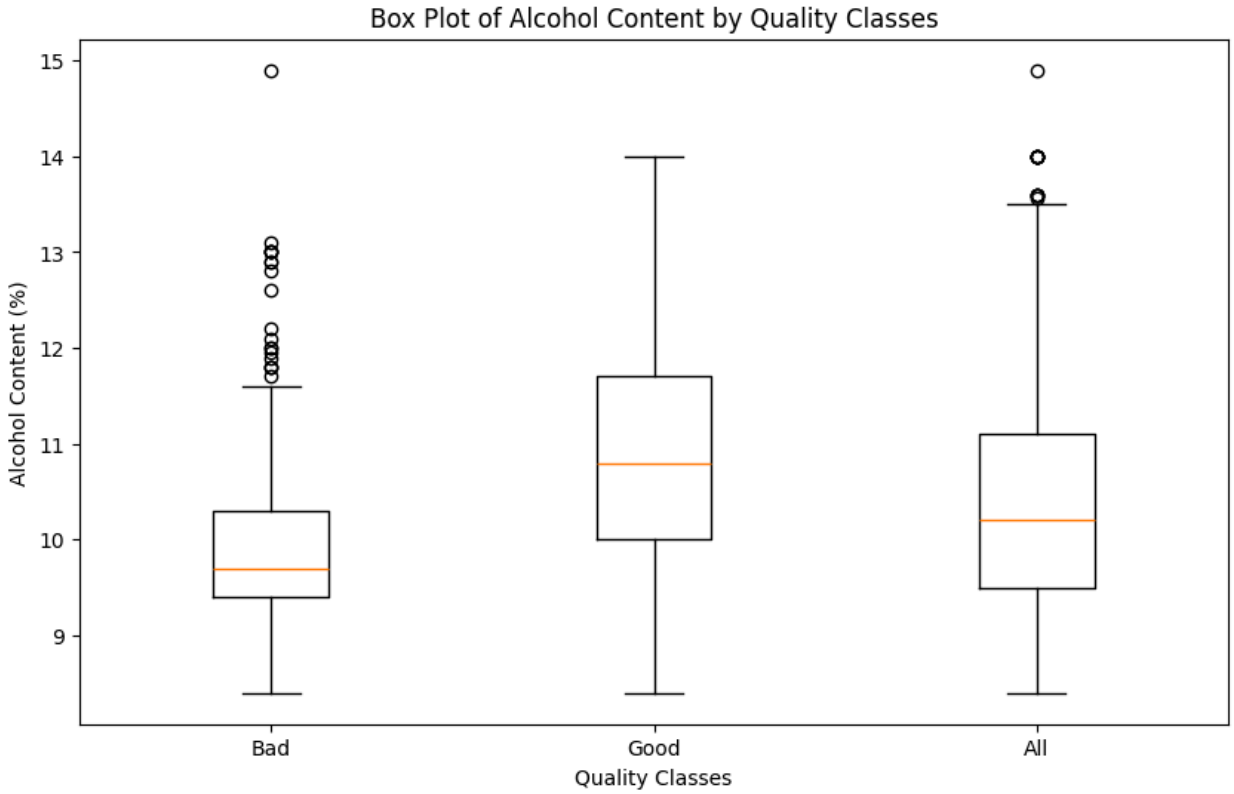
```
* There seems to be a positive trend with the correlation of acidity & citric acid. So
far, the relationship between these two attributes has been the strongest when compared
to the previous attributes.


* Correlation Matrix Value = 0.671703
```

5) Create a histogram of the **quality** attribute and interpret the resulting histogram.



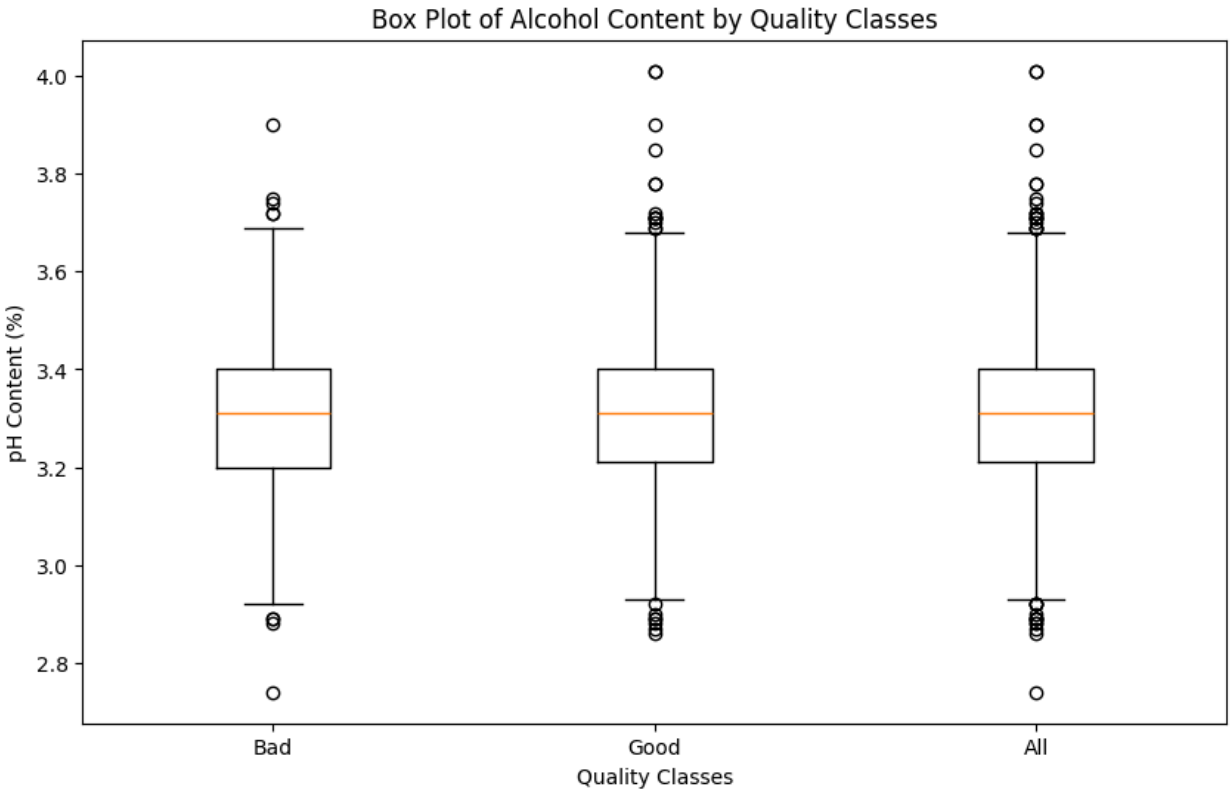
```
* Looking at the histogram, we can tell the frequency of each value for quality. We can
see that 5 and 6 are the most common when it comes to quality in this dataset. Even
though 5, which is considered bad, is the most prominent value -- when considering the
whole dataset, we will get more and have a slightly higher chance of getting a good
quality wine.
```

6) Create a box plot for **alcohol** attribute for the instances of the **quality** classes (Bad, Good), and all instances in the dataset (three boxes in the same plot). Interpret the resulting box plots.



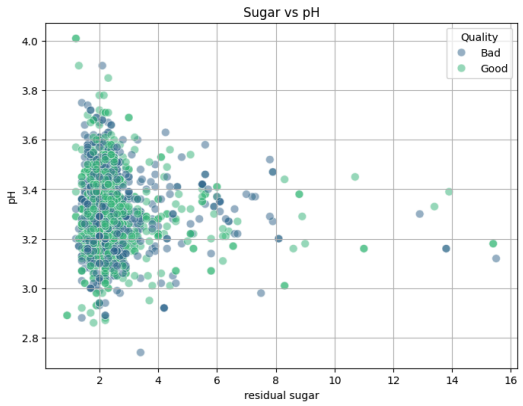
```
* From looking at the box plot of alcohol content by quality classes, we can determine
that wine with a higher alcohol content will tend to have a higher quality. The
difference is not significant but wine quality will tend to have a higher quality when it
contains at least 10% alcohol content.
```

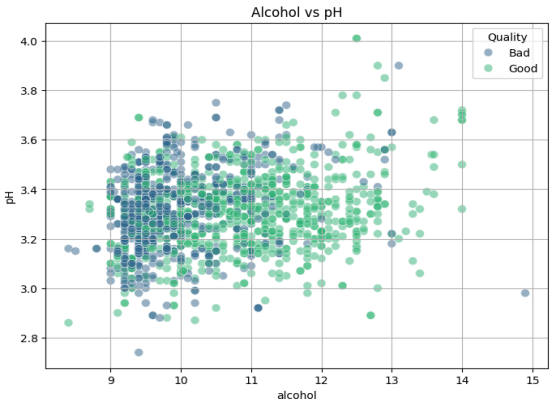
7) Create a box plot for **pH** attribute for the instances of the **quality** classes (Bad, Good), and all instances in the dataset (three boxes in the same plot). Interpret the resulting box plots.

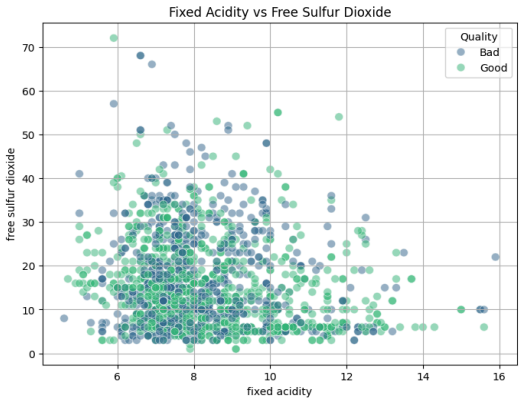


```
* From the data of the box plot -- there is not much of a difference when comparing
quality to pH. The pH levels of the wine seem to not correlate to the quality of wine. As
we can see from the data, the box plot location seems nearly identical. This tells us
that pH levels do not have a strong relationship with wine quality levels.
```

8) Create supervised scatter plots for the following 3 pairs of attributes using **quality** as a class variable: **alcohol** / **pH**, **residual sugar** / **pH**, and **fixed acidity** / **free sulfur dioxide**. Use different colors for the class variable. Interpret the obtained plots and address what can be said about the difficulty in predicting the **quality** and the distribution of the instances of the three classes. Identify the best pair of attributes based on the generated supervised scatter plots.

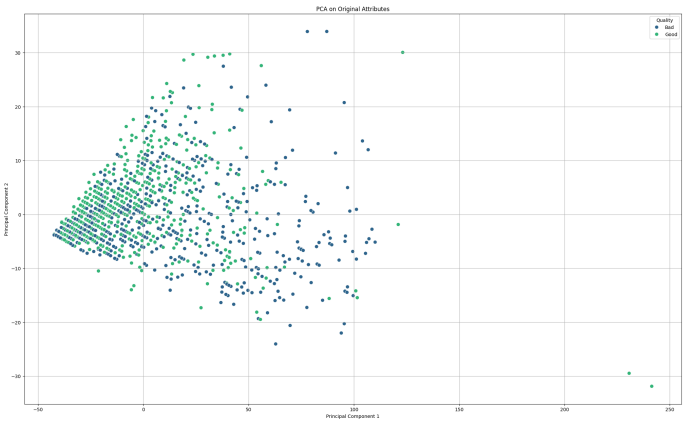


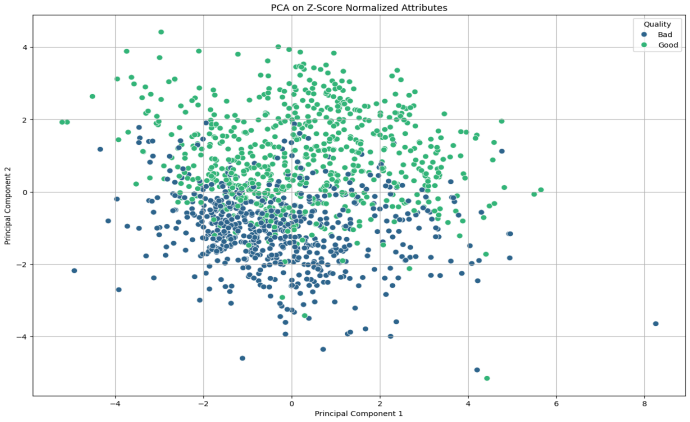


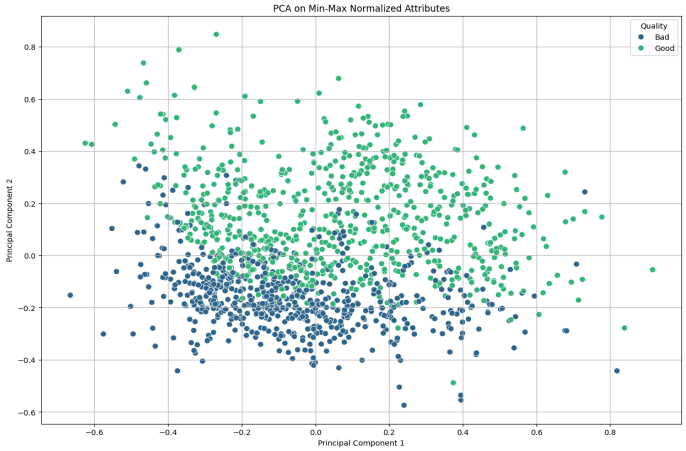


```
* According to the 3 scatter plots above, we can say overall that it is difficult to
predict the quality of wine. This is due to the significant amount of overlap in the dots
between good and bad wine quality. This tells us that there is minimal correlation
between the pairs of attributes when trying to predict wine quality. Out of the 3 pairs
of attributes -- I would say alcohol and pH is the best attribute pair when trying to
predict wine quality. This is because we can see the slight distinction between good and
bad wine quality. Even though there is overlap, we can see a bad wine quality cluster
forming on the left and a good wine quality cluster forming to the right. From the
scatter plot data, it seems alcohol has a higher significance when compared to pH, this
is because we can see the clusters separating along the alcohol axis.
```

9) Perform PCA on all of the attributes, reduce the dimension to **2 principal components** (2D PCA), and create a supervised scatter plot on the 2 principal components. Then normalize all of the attributes with **Z-Score** and **Min-Max** normalization, perform 2D PCA on the normalized attributes, and create two supervised scatter plots, one for the 2 principal components of the Z-Score normalized attributes, another for the 2 principal components of the Min-Max normalized attributes. Tell the difference before and after normalization, infer the reason of any difference, tell which normalization method is better and why, interpret how PCA is useful and what are its benefits in this data analysis.







---

* Min-Max normalization scales data to the range [0,1], which preserves the relationship between features while reducing the impact of outliers.

* From the data above, the separation seems to be an improvement from the original data and also the Z-Score data. There is significant improvement when compared to the original data, but minimal improvement when compared to the Z-Score data. This tells us that Min-Max normalization is the superior method of normalization for this dataset. This is probably due to the dataset's characteristic of having bounded values such as ratings and percentages. Min-Max normalization keeps these values within expected range -- which can help with predicting a wine's quality.

* **Benefits of PCA:**

  * **Noise Reduction:** Removes less important variables -- which in hand focuses on the most significant patterns

  * **Improved Visualization:** Allows for a clear visual representation of complex data. This helps us to detect patterns and clusters

  * **Dimension Reduction:** Reduces the number of variables while maintaining similar details/data. This helps data easier to visualize.

10) Write a brief conclusion summarizing the most important findings of this task; in particular, address the findings obtained related to predicting the quality of red wine. If possible, write about which attributes seem useful for predicting wine quality and what you as an individual can learn from this dataset.

- To conclude this data visualization of quality of wine – we can determine overall it is challenging to accurately predict the quality of red wine. We can see the slight hint when we look at the correlation matrix heatmap generated on question 1, due to the fact that the heat map is mostly blue, which tells us the relationship/correlation between the attributes is minimal. This can lead to future conflicts as contradicting movements between the attributes can be quite frustrating when trying to predict an outcome of one certain attribute. The low correlation was visualized through a couple attributes using scatter plots. From the data shown on the scatter plots, we can see that the dots are generally spread out and only show small signs of correlation. The box plots also showed us that our chances of predicting red wine quality will be low because when separated by quality classes (Good and Bad) we see that the boxplots are generally equivalent in terms of location on the graph. Working with the two attributes given from the assignment, we can conclude that changes in pH or alcohol percentages do not typically affect the quality in red wine. Even when we visualize: alcohol vs pH, sugar vs pH, and fixed acidity vs free sulfur dioxide, through supervised scatter plots using quality as a class variable. When trying to determine which attributes are the most useful, trying to predict wine quality: alcohol, sulphates, and citric acid stands out. This is because of their correlation level and when visualized through the supervised scatter plots. Throughout this process, PCA (Principal Component Analysis) was introduced in attempts to reduce the dimensionality of the dataset. Through creating a supervised scatter plot on the 2 principal components – we can see the significant overlap between the ‘Good’ and ‘Bad’ classes. Normalization of all attributes with Z-Score and Min-Max normalization was introduced, performing 2D PCA on the normalized attributes – 2 supervised scatter plots were visualized. With the new normalization method, there was a huge difference in separating between the ‘Good’ and ‘Bad’ clusters. Between the 2 methods, Min-Max produced the best results – showing that keeping bounded values (percentages and ratings) within expected ranges will yield best separation between the quality classes. To wrap this report up nicely, I was able to learn how to use PCA in my data analysis, which will help me in the future whenever I need to reduce the dimensions of my dataset for easier interpretation. This dataset allowed me to hone in my abilities to interpret various statistical graphs, which is great for future projects/ambitions. I also got to learn how to use Min-Max and Z-Score normalization. It was interesting to see how the data changed from such statistical analysis, it was very insightful. Overall, this homework is a great introduction for this class and I had a pleasant time.