Exploratory Data Analysis on Wine Quality

Jennifer Nguyen

COSC 3337 - Data Science I

Jingchao Ni

22 September 2024

Introduction

The dataset used in this exploratory data analysis consists of a variety of red wine, and their attributes that are a range of physicochemical and sensory measurements. Scores are given based on the taster's indicators of perceived quality. The wine quality dataset consists of 1,599 observations of red wine with 12 attributes, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content, quality.

The primary objective of this analysis is to explore the relationships between different wine attributes and their given quality score to a wine. By understanding these relationships, we aim to identify the key factors that influence wine quality. This report will cover an examination of the dataset of the wine quality dataset, including correlation analysis between attributes, distribution analysis of variables. and insights of these relationships impacts.

The overall goal of this Exploratory Data Analysis (EDA) is to uncover patterns that can guide further analysis to better understand what makes some wines more preferable based on attributes that positively or negatively affects their quality.

Data Overview

The wine quality dataset contains data on 1,599 different red wines, with each wine having 12 attributes that describe their chemical property and given rating. The table below contains an in-depth description of the dataset's structure.

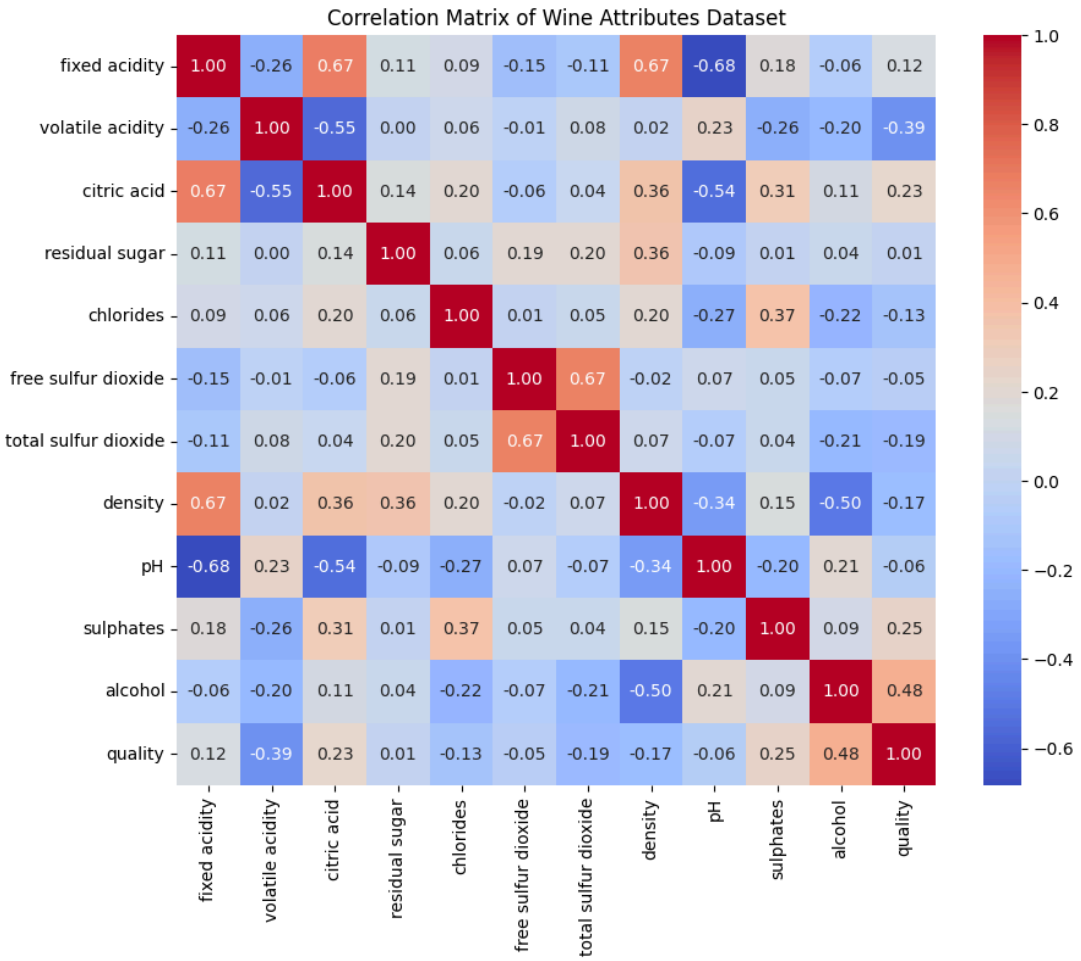
Attribute	Description	Type
Fixed Acidity	The concentration of fixed acids (like tartaric acid) in g/dm^3	Numerical
Volatile Acidity	The concentration of acetic acids (affecting flavor) in g/dm^3	Numerical
Citric Acid	A natural preservative	Numerical

	and acid found in wines (g/dm^3)	
Residual Sugar	Amount of sugar left after fermentation (g/dm^3)	Numerical
Chlorides	Salt content in wine (g/dm^3)	Numerical
Free Sulfur Dioxide	Sulfur dioxide not bound to other molecules (mg/dm^3)	Numerical
Total Sulfur Dioxide	Total amount of sulfur dioxide in the wine (mg/dm^3)	Numerical
Density	Density of the wine (g/dm^3), related to sugar and alcohol levels	Numerical
pH	Acidity level of the wine, from 0 to 14	Numerical
Sulphates	Potassium sulphate levels, acting as a wine preservative (g/dm^3)	Numerical
Alcohol	Alcohol content by volume (%)	Numerical
Quality	Wine quality score (0 to 10)	Ordinal

Data Visualization

Step 1 (Correlation)

The correlation matrix provided below allows better understanding of the relationships between the chemical attributes of the wines and their quality ratings. A heatmap was generated to visually represent these correlations.



Correlation Matrix Analysis:

Below are pulled key observations for each type of correlation.

Strong Positive Correlations:

- Fixed Acidity and Citric Acid (0.67): This suggests that wines with a higher fixed acidity tend to have a higher citric acid content.
- Fixed Acidity and Density (0.67): This suggests that wines with higher acidity are also denser.
- Free Sulfur Dioxide and Total Sulfur Dioxide (0.67): This is an expected high correlation since total sulfur dioxide includes free sulfur dioxide.

Strong Negative Correlations:

- Fixed Acidity and pH (-0.68): This makes sense as higher acidity corresponds to lower pH.
- Volatile Acidity and Citric Acid (-0.55): Wines with higher volatile acidity tend to have lower citric acid content.

- Density and Alcohol (-0.50): This indicated wines with higher alcohol levels are associated with lower density levels.

Moderate Correlations with Wine Quality:

- Alcohol (0.48): Alcohol content shows the strongest correlation with quality. This suggests that alcohol content of wines potentially plays a key role in determining overall wine quality.
- Volatile Acidity (-0.39): There is a negative correlation with wine quality, suggesting higher volatile acidity results in lower wine quality.
- Sulphates (0.25): There is a weak positive correlation with quality, suggesting that higher sulphate level may slightly improve wine preservation. This means that this may slightly improve wine quality scores.

Weak or No Correlations:

- Residual Sugar (0.01): Residual sugar shows almost no correlation with quality, suggesting that sweetness in wine does not significantly impact overall wine quality score.
- Free Sulfur Dioxide (-0.05) and pH (-0.06): Both of these attributes also show very weak negative correlations, suggesting little to no impact on overall wine quality.

Interrelated Features:

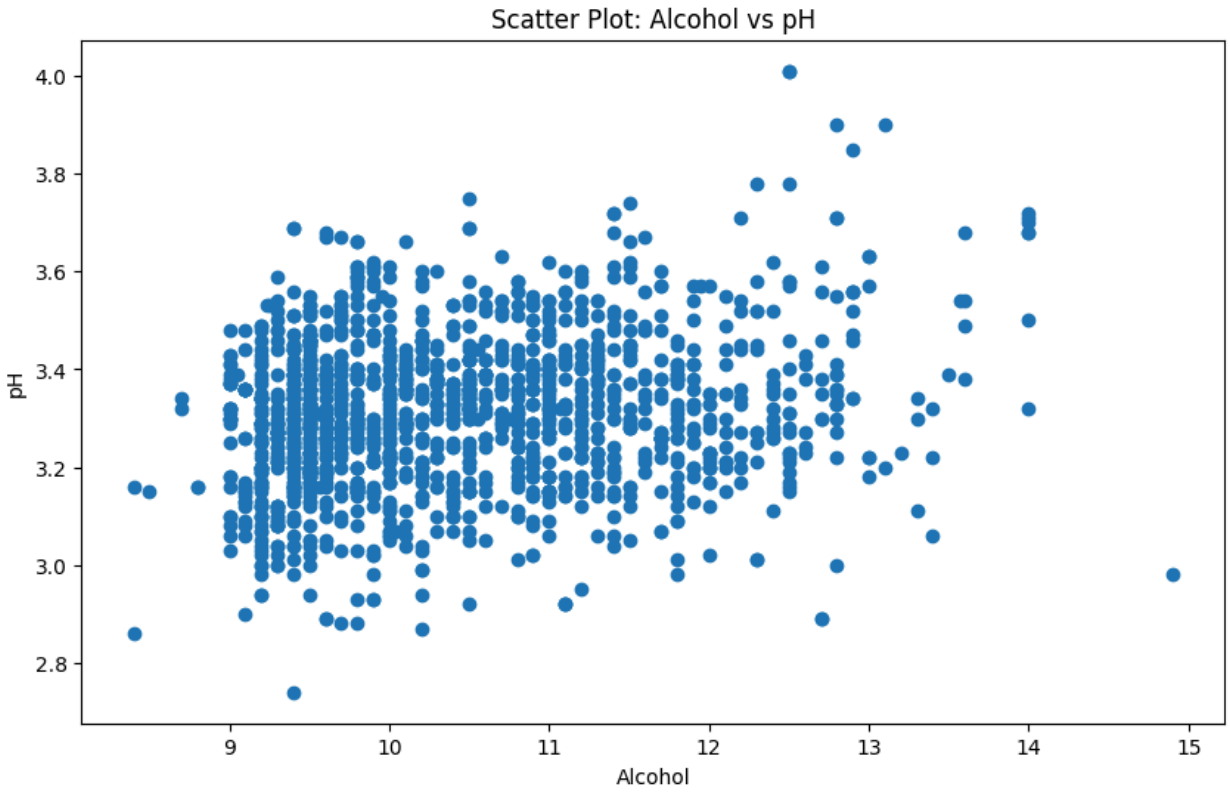
- Acidity Measures: Fixed acidity, volatile acidity, citric acid, and pH are all interrelated, as expected, since they collectively contribute to the wine's acidity profile
- Sulfur Dioxide Measures: Free and total sulfur dioxide are strongly correlated but show little relationship with other attributes or quality.

Conclusion

Based on these statistical findings for this correlation analysis, the alcohol content attribute appears to be the strongest positive single predictor of quality, followed by volatile acidity and sulphate attributes being the strongest negative predictor of quality. With the alcohol content being the strongest positive predictor of wine quality, an assumption can be made that increasing alcohol content could possibly improve wine quality. With volatile and sulphates being the strongest negative predictor of wine quality, an assumption can be made that reducing both these attributes could possibly improve wine quality.

Step 2-4 (Scatterplot Analysis)

Attributes: Alcohol and pH



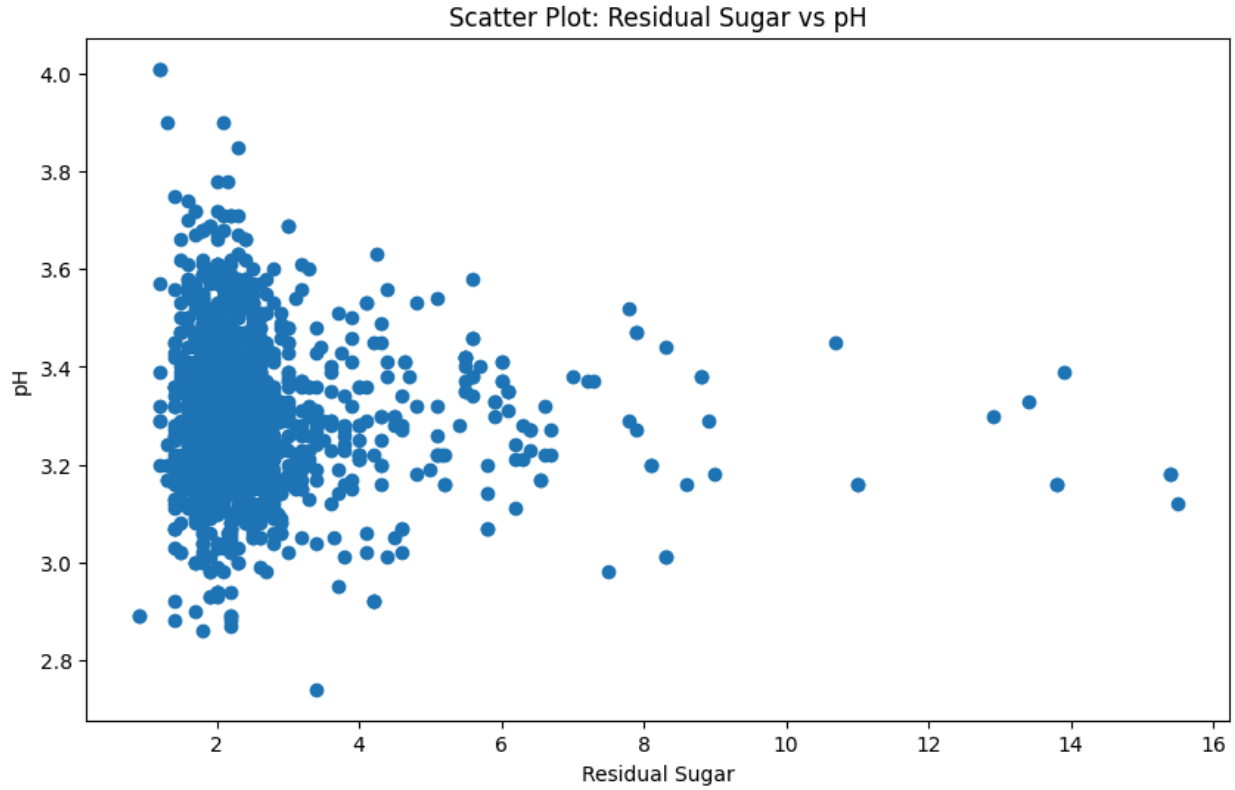
Key Observations:

This scatter plot contains 'Alcohol' percentages for the x-values ranging from 0% to 15% and 'pH' values for the y-values ranging from 1.0 to 4.0. Examining the plot, there seems to be a slight negative correlation visible - as alcohol content increases, there's a weak tendency for pH to decrease. However, the relationship is not strong, as there's considerable scatter in data. The majority of the data points are clustered in the alcohol range of 9% to 12%, and pH range of 3.0 to 3.6. There exists some outliers with higher pH values of around 4.0 at various alcohol levels.

Interpretation:

While there's a weak negative relationship between alcohol content and pH values, other factors likely influence pH more strongly. The wide scatter suggests that red wines can have various pH levels across different alcohol contents.

Attributes: Residual Sugar and pH



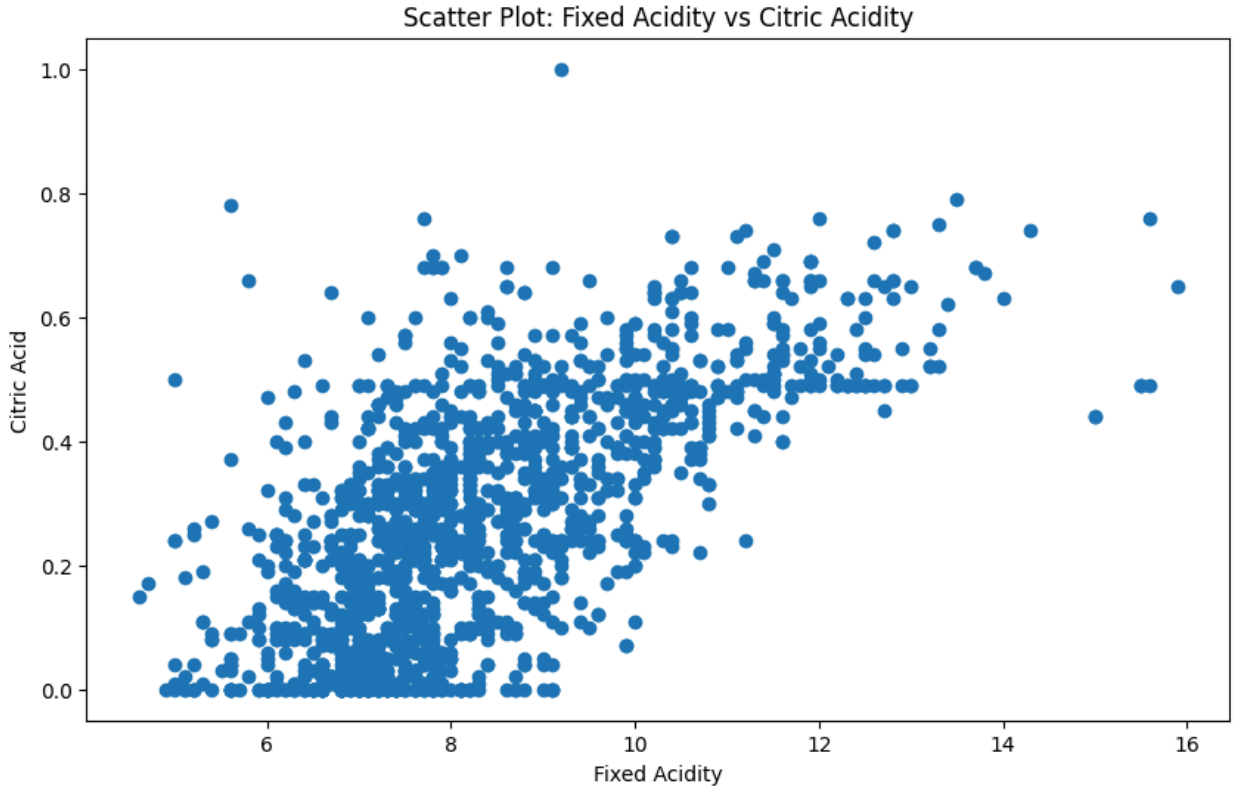
Key Observations:

This scatter plot has 'Residual Sugar' for its x-values that range from 0 to 16 g/L, and has 'pH' values for its y-values that range from 0 to 4.0. Inspecting the plot graph, most wines appear to have residual sugar levels below 5 g/L, with a dense cluster below 2.0 g/L. There are fewer wines with higher residual sugar levels, creating a right-skewed distribution. There's only a slight negative correlation - as residual sugar increases, there's a weak tendency for pH to decrease. The relationship between 'Residual Sugar' and 'pH' is not strong as there is considerab;e variation in pH at all sugar levels.

Interpretation:

While there's a slight trend to lower pH with higher residual sugar, the relationship is weak. Other factors likely have more influence on pH. The plot also indicates that most wines in this dataset have relatively low residual sugar content.

Attributes: Fixed Acidity and Citric Acidity



Key Observations:

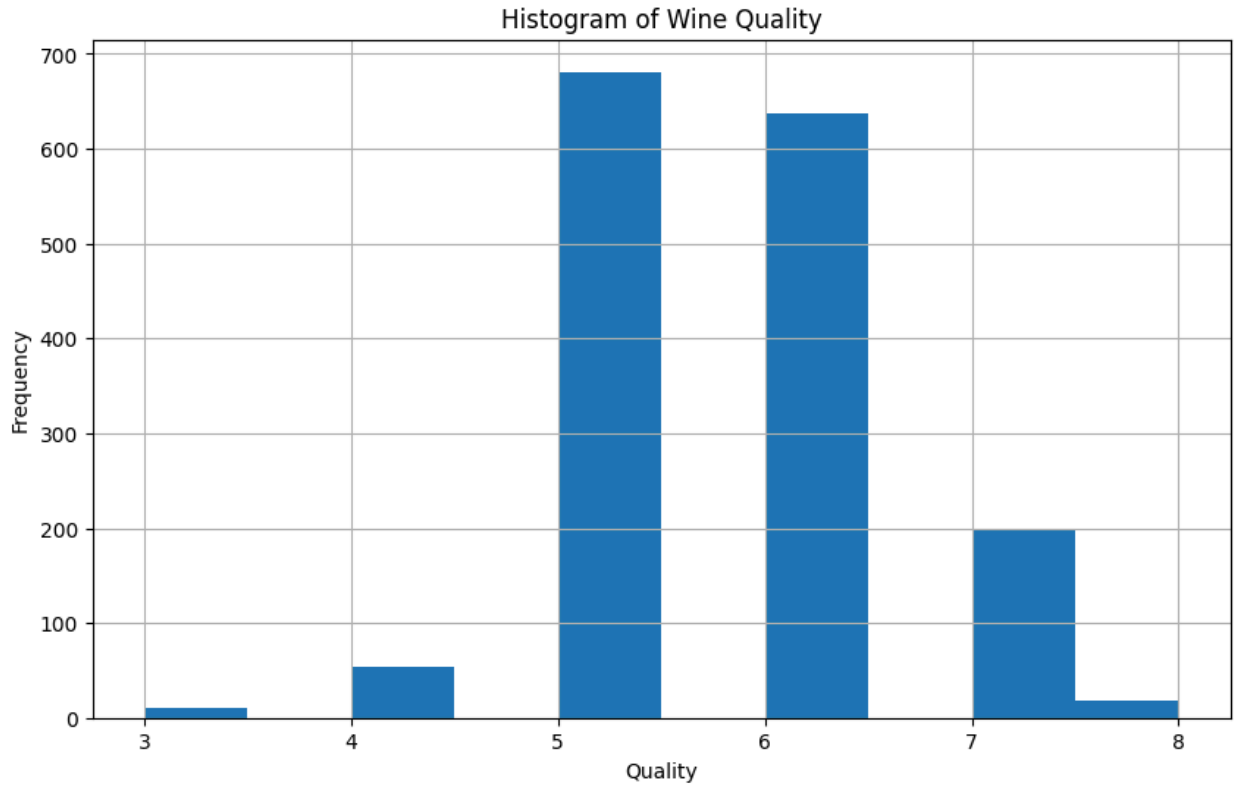
This scatter plot has 'Fixed Acidity' for its x-values that ranges from 0 to 16 g/L, and has 'Citric Acid' for its y-values that ranges from 0 to 1.0 g/L. Examining the plot graph, there appears to be a positive correlation between fixed acidity and citric acid content. Additionally, the correlation seems strong in the middle range of the fixed acidity - about 6 to 12 g/L. The relationship appears to be moderately strong, with citric acid generally increasing as fixed acidity increases. There exists a notable cluster of wines with very little to none citric acid content across various fixed acidity levels.

Interpretation:

This plot indicates a moderately strong positive relationship between fixed acidity and citric acid content in wines. As fixed acidity increases, so does citric acid content. However, there exists a notable presence of wines with very low citric acid across the acidity spectrum, indicating that other acids contribute to the fixed acidity in some wines.

Step 5 (Histogram)

Attributes: Wine Quality



Key Observations:

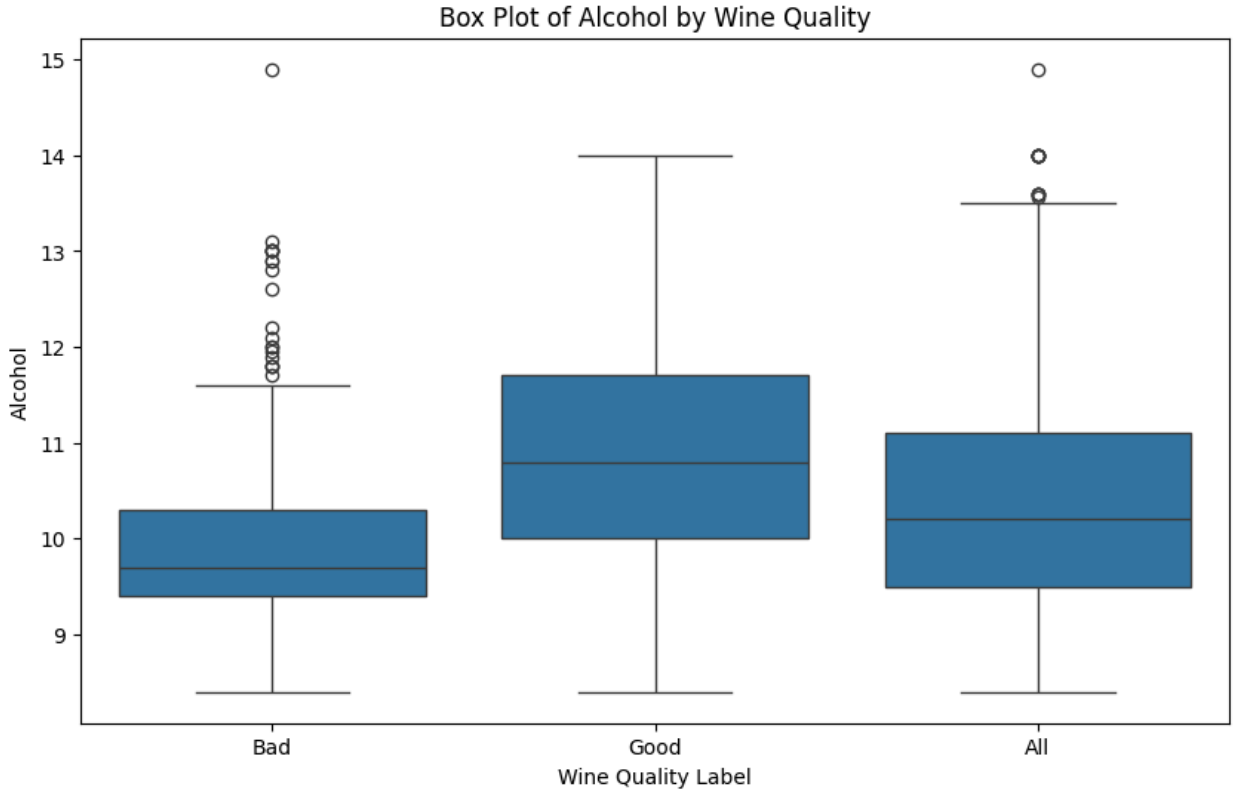
This histogram shows the distribution of wine quality ratings. It has 'Quality' for its x-values that range from 3 to 8, and 'Frequency' for its y-values that range from 0 to 700. The distribution is roughly bell-shaped, but slightly skewed to the right. The most common scores for quality ratings are 5 and 6, with 5 being slightly more frequent.

Interpretation:

The majority of wines in this dataset are of average quality - ranging from 5 to 6. Examining the histogram, there's a tendency towards slightly above-average ratings, with more wines rated 7 than rated 4. Exceptional wines, with either very low or very high quality, are on the rarer side. The distribution of this histogram indicates that it's challenging to produce wines of the highest quality as most wines fall within the average rating.

Step 6-7 (Box Plot)

Attributes: Alcohol by Wine Quality



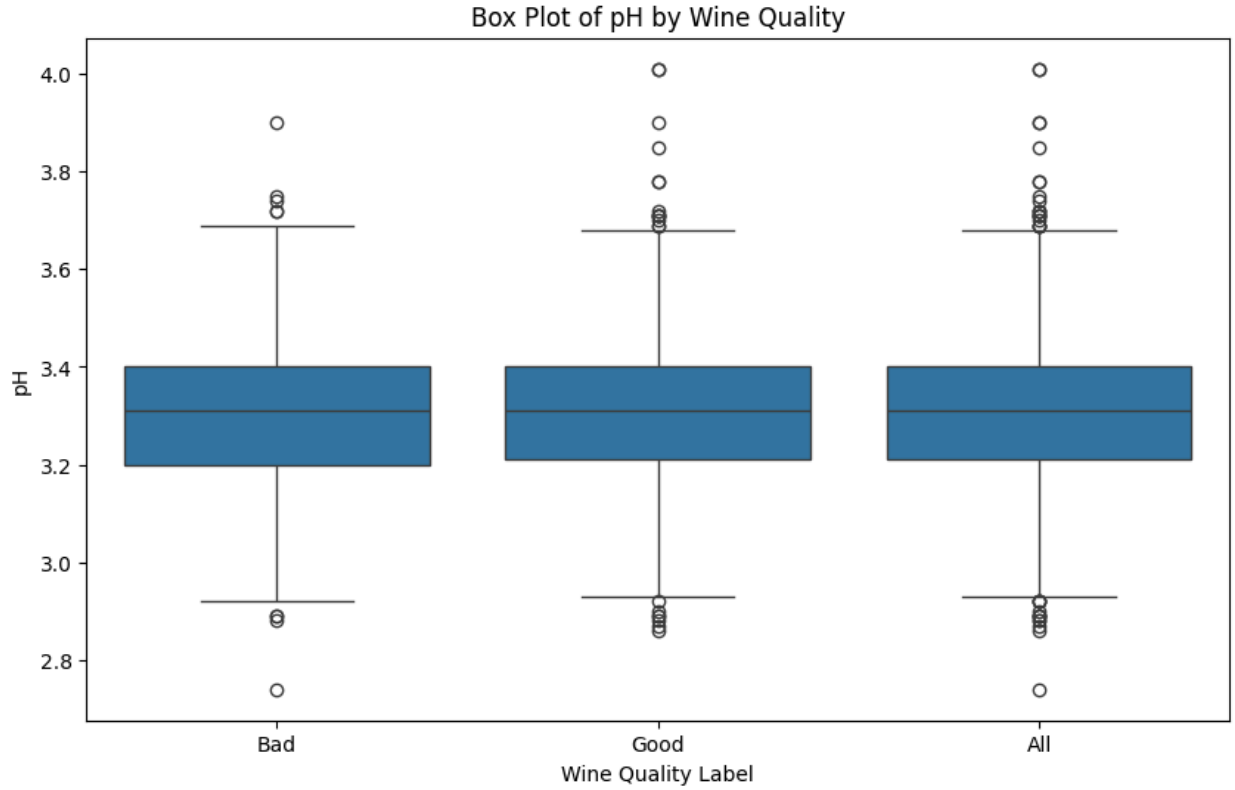
Key Observations:

This box plot compares alcohol content (y-value) across different wine quality categories (x-values). These categories are 'Bad', 'Good', and 'All'. Examining the box plot, 'Good' wines appear to have higher alcohol content than 'Bad' wines, with the median alcohol content being higher as well for 'Good' wines. Additionally, there exists more variability in alcohol content for 'Good' wines than 'Bad' wines. Only 'Bad' wines contain outliers, with the outliers going up towards higher alcohol content. The 'All' category shows the overall distribution, which is closer to the 'Good' category.'

Interpretation:

There appears to be a positive relationship between alcohol content and wine quality. Better-rated wines tend to have higher alcohol levels, though there's a significant overlap. The higher variability in 'Good' Wines indicate that high quality wines can be achieved across a wider range of alcohol levels, while lower-quality wines cluster more tightly around lower alcohol content.

Attributes: pH by Wine Quality



Key Observations:

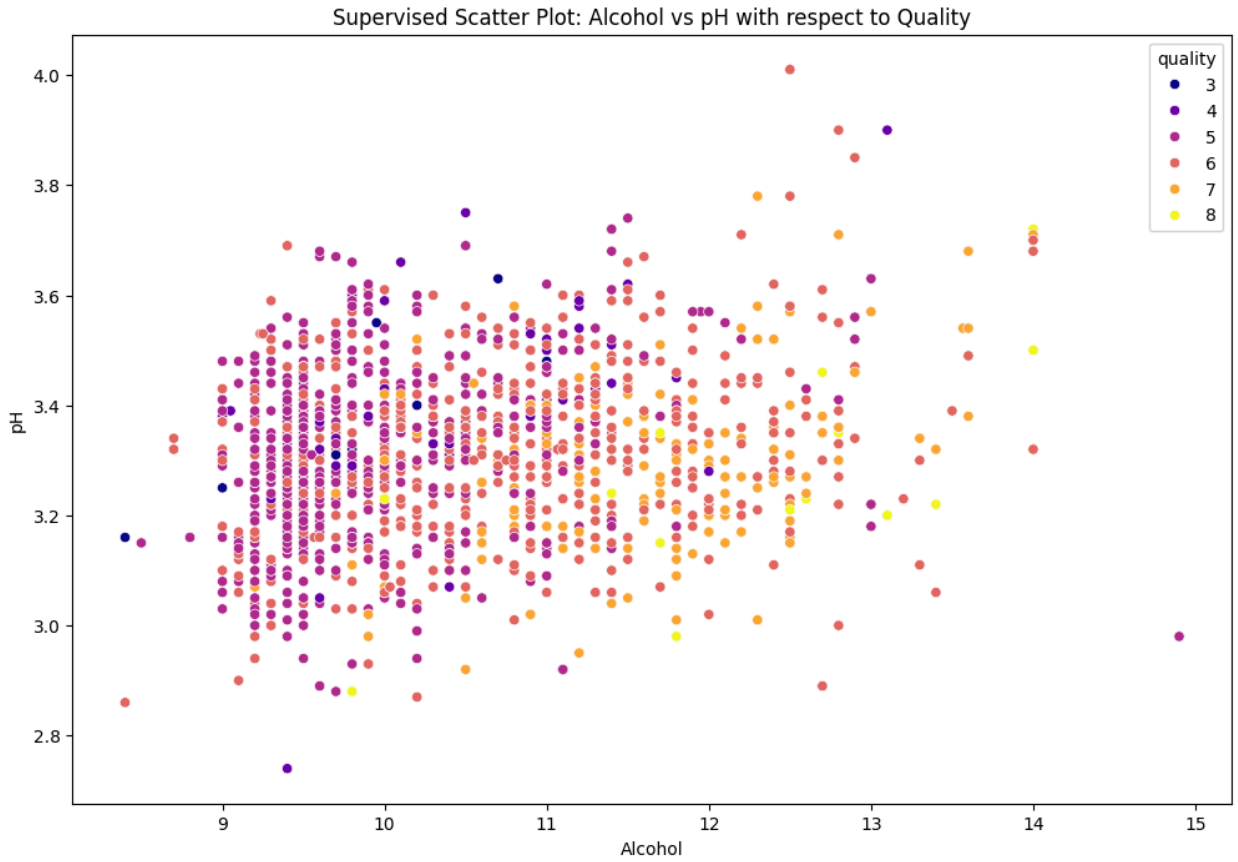
This box plot compares pH values (y-value) across different wine quality categories (x-values). These categories are 'Bad', 'Good', and 'All'. Examining the box plot, median pH levels are nearly identical across all the wine quality categories. There exists some outliers in both high and low pH for all categories. The overall pH range is approximately 2.8 to 4.0, with most of the wines falling between 3.2 and 3.4.

Interpretation:

Unlike alcohol content, pH doesn't seem to have a strong relationship with wine quality. Both 'Good' and 'Bad' wines have similar pH distribution, indicating that pH is not the sole determining factor in wine quality. The similarity across all categories of wine quality suggests that wine generally fall within a narrow pH range, regardless of their quality rating.

Step 8 (Supervised Scatter Plots)

Attributes: Alcohol vs pH with respect to Quality



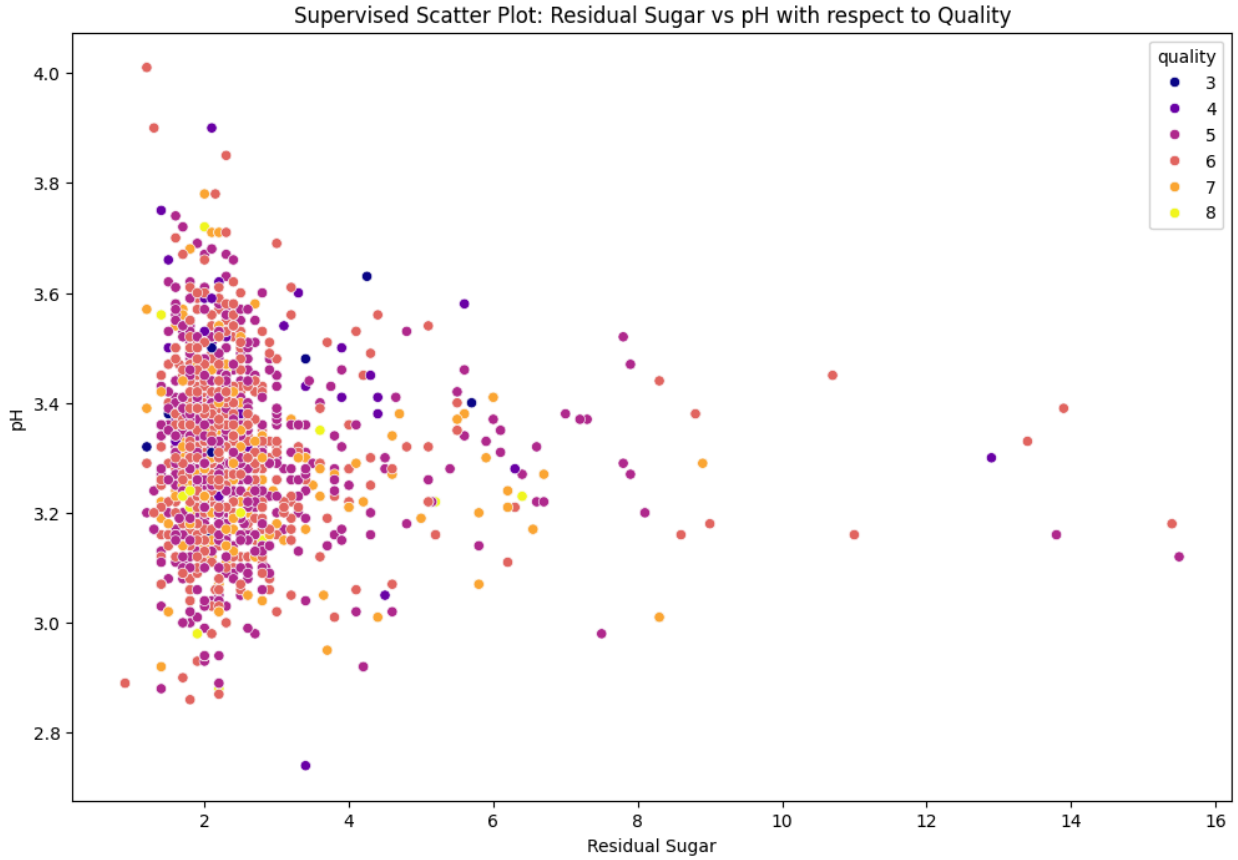
Key Observations:

This supervised scatter plot compares 'Alcohol' to 'pH' values with respect to 'Quality' scores of wines. Examining the plot graph, there's a slight negative correlation between alcohol content and pH levels. Higher quality wines that range from 6 to 8 in score tend to cluster more towards higher alcohol content that range from 10% to 14%. On the other hand, lower quality wines that range from 3 to 5 in score are more spread out, but overall tend to have lower alcohol content. The pH values are clustered between 3.0 to 3.6

Interpretation:

While there's a trend of higher quality wines having higher alcohol content, the relationship isn't strong enough to clearly separate quality levels. As there is significant overlap in the distribution of different quality levels, a strong assumption can be made that pH doesn't seem to have a strong correlation with quality.

Attributes: Residual Sugar vs pH with respect to Quality



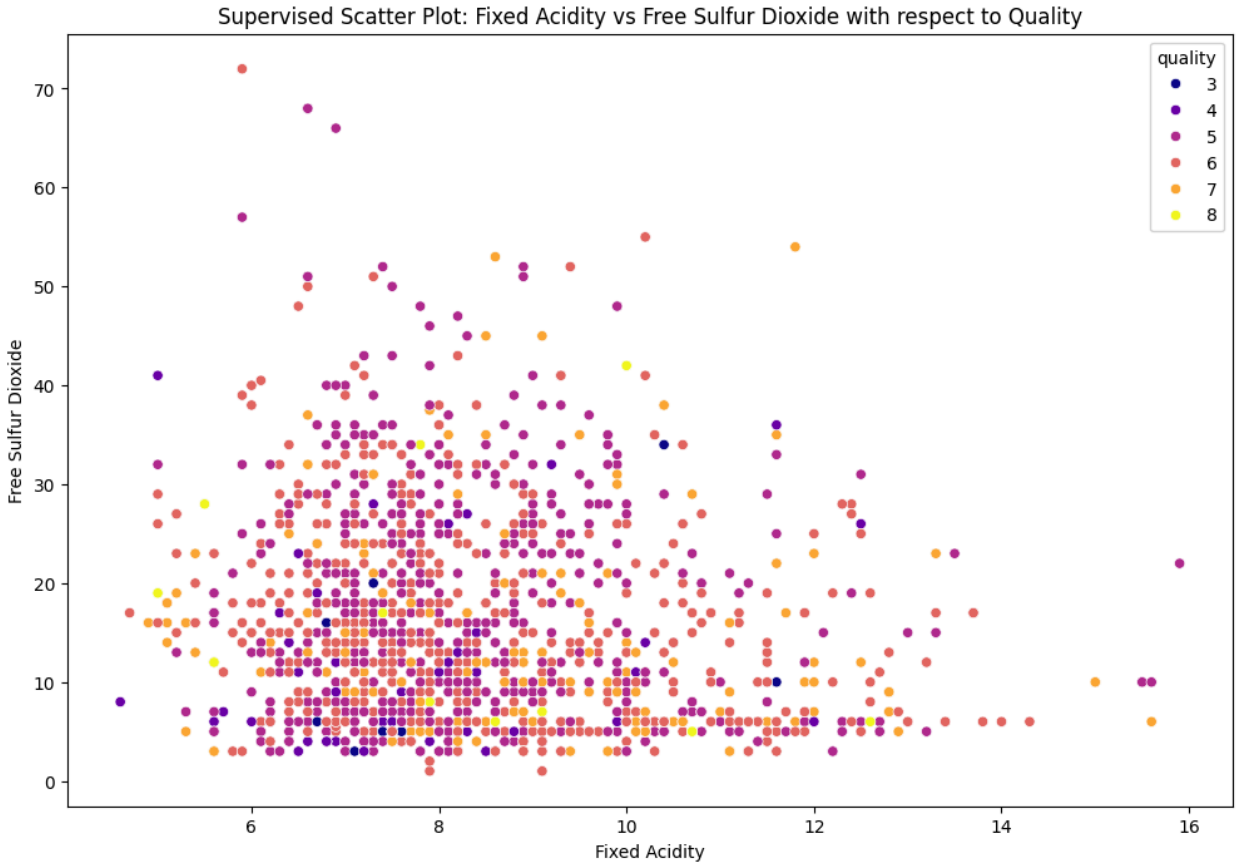
Key Observations:

This supervised scatter plot compares 'Residual Sugar' to 'pH' values with respect to 'Quality' scores of wines. Examining the plot graph, there is a slight negative correlation between 'Residual Sugar' and 'pH'. Most of the wines cluster at lower residual sugar levels, ranging from 0 to 4 g/L. Furthermore, higher quality wines, with scores ranging from 6 to 8, tend to have lower residual sugar. The pH levels are also fairly consistent across quality levels. There exists some outliers with high residual sugar levels.

Interpretation:

Lower residual sugar seems to be associated with higher quality, but the relationship isn't definitive. Additionally, pH doesn't have a strong correlation with quality scores. The clustering of data points indicates that most wines that were used in this dataset are dry (low sugar).

Attributes: Fixed Acidity vs Free Sulfur Dioxide with respect to Quality



Key Observations:

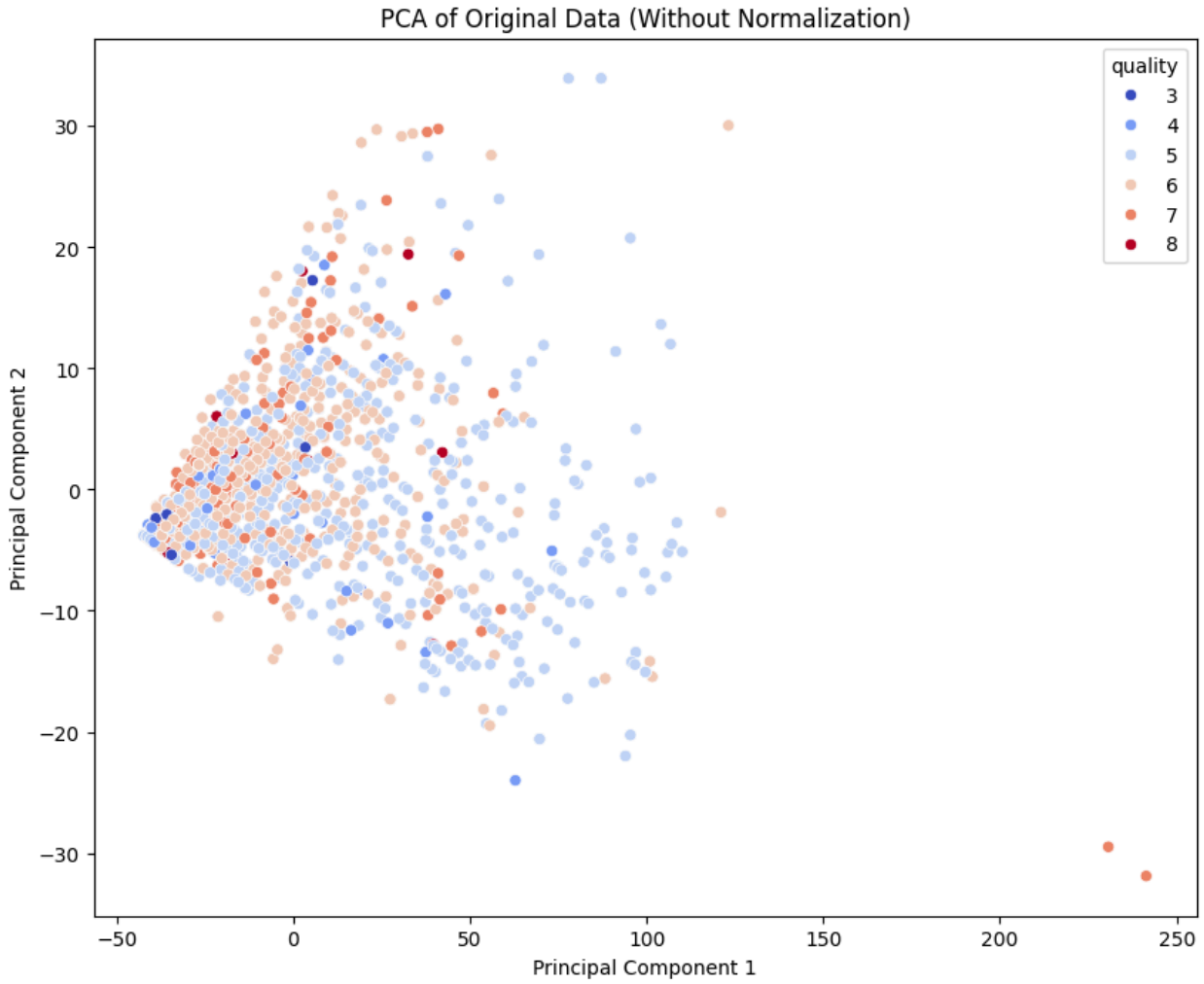
This supervised scatter plot compares 'Fixed Acidity' to 'Free Sulfur Dioxide' with respect to 'Quality' scores of wines. Examining this plot graph, there appears to be a weak negative correlation between fixed acidity and free sulfur dioxide. Although quality scores are quite mixed throughout the plot, higher quality wine, that ranges from 6 to 8 in score, seem to have a slight tendency towards lower free sulfur dioxide levels.

Interpretation:

This plot illustrates the least clear separation between quality levels as both fixed acidity and free sulfur dioxide do not seem to have a strong individual correlation with wine quality.

Step 9 (PCAs)

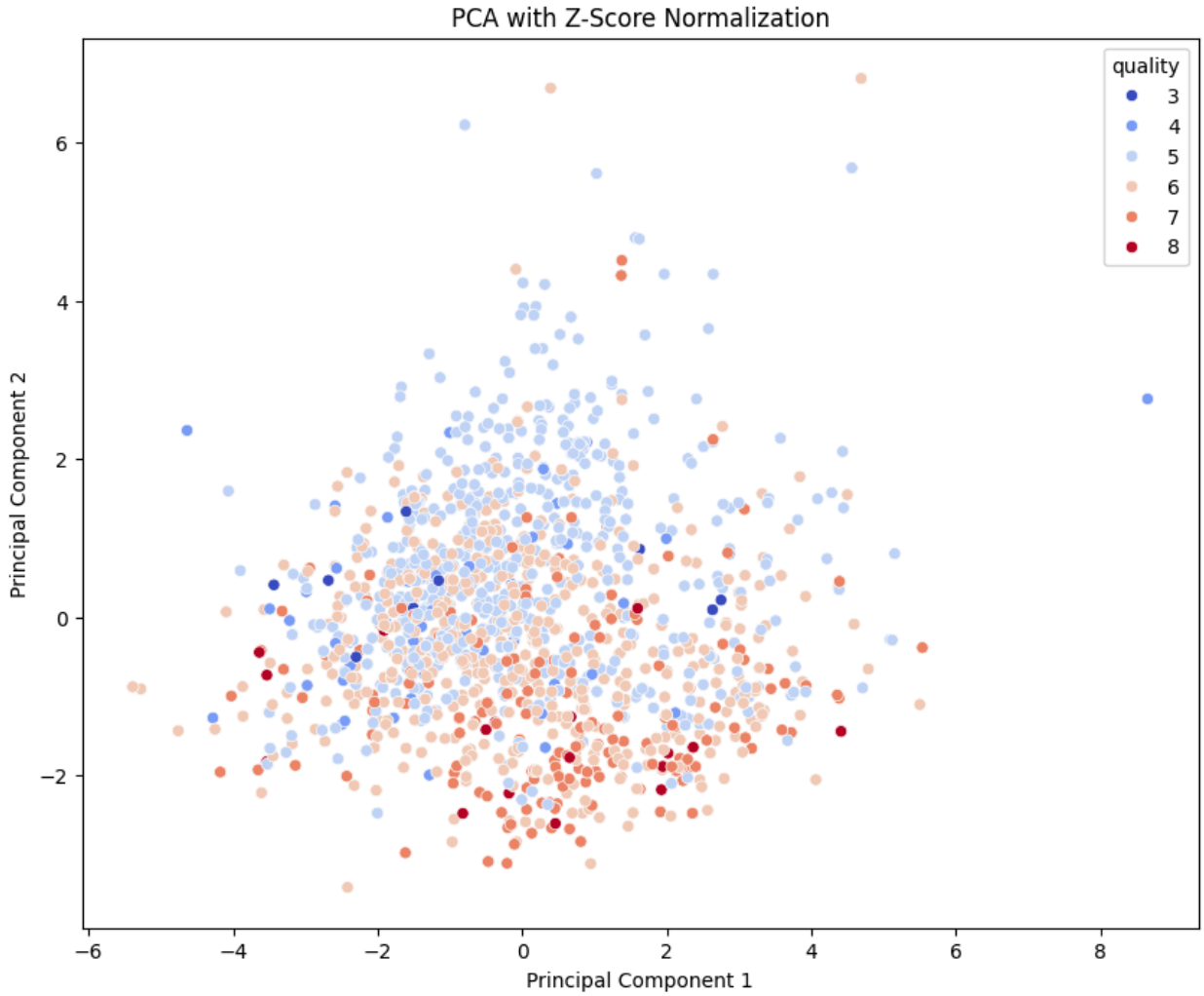
PCA: Original Data without Normalization



Key Observations:

In this PCA of original data without normalization, the data points are widely spread, with most clustered between -50 and 100 on 'Principal Component 1'. Examining the data points, there's a noticeable skew towards the left side of the plot. Some outliers are visible, especially the bottom 2 data points on the far right of 'Principal Component 1'. The quality classes are mixed, with slight trends of higher quality wines being more on the right.

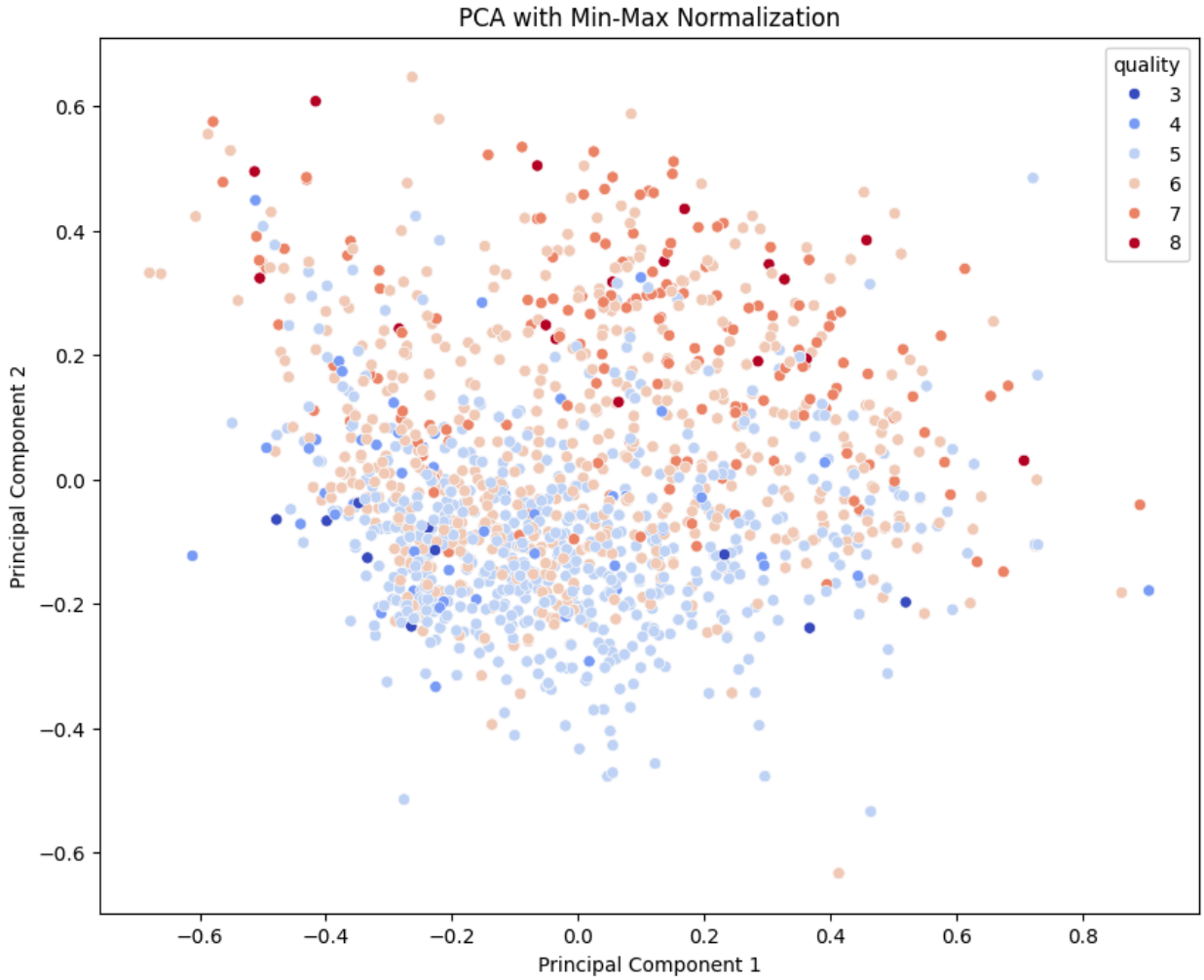
PCA: Z-Score Normalization



Key Observations:

In this PCA with Z-score normalization, the data points are more evenly distributed around the origin. The overall shape of the data cloud is more circular. Additionally, the quality classes are still mixed, but with slightly clearer trends.

PCA: Min-Max Normalization



Key Observations:

In this PCA with min-max normalization, the data points are more evenly distributed around the origin, just like the Z-score normalization. However, this plot is more compact.

Difference Before and After Normalization

The unnormalized PCA shows a much wider spread of data points and is heavily influenced by the scale or original variables. Both normalized PCAs result in a more compact and centered distribution. Normalization reduces the impact of variables with larger scales, allowing all variables to contribute more equally to the principal components.

Reason for Differences
The original data likely had variables with very different scales such as alcohol percentages versus total sulfur dioxide in mg/L. Normalization standardizes these scales, which prevents variables with larger magnitudes from dominating the PCA.

Which Normalization is Better
Looking at the Z-score normalization, it results in a slightly more spread out distribution. Looking at the min-max normalization, it constrains all data to a fixed range. Overall, the Z-score appears better because it preserves more of the data's structure while standardizing scales, and is less affected by outliers, which is important when extreme values might appear.

How PCA is Useful and Benefits in this Data Analysis
Principal Component Analysis (PCA) aids in dimensionality reduction by condensing numerous variables into just two, which simplifies visualization and analysis. Additionally, PCA facilitates feature extraction, as the principal components are combinations of the original variables that capture the most variance in data. THis method also helps in noise reduction by concentrating on the primary sources of variation. Furthermore, PCA can reveal subtle patterns in wine quality, despite overlap. Overall, this analysis allows us to see additional features beyond the chemical properties provided in the dataset.

Conclusion

Based on these analyses, alcohol content, volatile acidity, and sulphates seem to be the most useful attributes for predicting wine quality. In general, higher alcohol content and lower volatile acidity are indicative of better wines. This analysis teaches us that while certain physicochemical properties are correlated with quality, the relationships are not always linear or straightforward.

This dataset reveals that certain attributes like sweetness or acidity levels (pH) do not heavily influence quality, and it may be worth exploring other sensory factors or expert ratings in future analyses. Through the use of various graphs and plots in this analysis, I have gained valuable insights into how to visually explore and interpret data. The correlation matrix revealed strong and weak relationships between wine attributes, helping me identify which features, like alcohol content and volatile acidity, are most relevant to wine quality. Scatter

plots allowed me to visualize trends and distributions, showing how higher alcohol content tends to be associated with higher quality wines, while attributes like residual sugar and pH showed weaker correlations. Histograms and box plots highlighted the distribution of quality ratings, emphasizing that most wines are rated as average, while box plots helped me understand the variability of alcohol content across different quality levels. Principal Component Analysis (PCA) taught me the importance of normalization when dealing with variables on different scales, as it allowed for a clearer visualization of the dataset's variance. Overall, I've learned that while visualizations are powerful for detecting trends and relationships, they also reveal the complexity of predicting wine quality, suggesting the need for more advanced analytical techniques.