Exploratory Data Analysis on Wine Quality (Red) with SAS

This document showcases the Exploratory Data Analysis (EDA) conducted on the Wine Quality (red) dataset retrieved from Kaggle website. EDA is a crucial initial step in data analysis, helping us understand the data's characteristics, identify patterns and relationships between variables, and uncover potential anomalies before diving into formal statistical analysis.

Data Source:

The Wine Quality dataset was obtained from Kaggle: https://www.kaggle.com/code/abdelruhmanessam/wine-quality?scriptVersionId=149902775&cellId=2

Data Dictionary for the Wine Quality Dataset

Variable Name	Description	Data Type	Measurement Unit
Fixed_acidity	The fixed acidity in the wine	Numeric	g/L
Volatile_acidity	The volatile acidity in the wine	Numeric	g/L
Citric_acid	The citric acid in the wine	Numeric	g/L
Residual_sugar	The residual sugar in the wine	Numeric	g/L
Chlorides	The chlorides in the wine	Numeric	g/L
Free_sulfur_dioxide	The free sulfur dioxide in the wine	Numeric	mg/L
Total_sulfur_dioxide	The total sulfur dioxide in the wine	Numeric	mg/L
Density	The density of the wine	Numeric	g/cm^3
pH	The pH of the wine	Numeric	-
Sulphates	The sulphates in the wine	Numeric	g/L
Alcohol	The alcohol content in the wine	Numeric	% vol
Quality	The quality rating of the wine (0-10)	Numeric	-

Exploratory Data Analysis with SAS Code

This section provides a well-structured breakdown of the EDA process:

1. **Load the Dataset:**

```
PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.IMPORT;
     GETNAMES=YES;
RUN;
```

2. **Explore the Data:**

- Use PROC CONTENTS code below to run the analysis. PROC CONTENTS is a SAS procedure used to summarize the contents of the wine dataset. It provides information about the dataset's structure and much more. PROC CONTENTS is a valuable tool for understanding the structure and contents of the wine dataset.

```
/*To get the detailed information about data including the number of observations*/
PROC CONTENTS DATA=WORK.IMPORT;
RUN;
```

This will provide information about the dataset, including the number of observations, variables, data types, and creation date.

The CONTENTS Procedure

Data Set Name	WORK.IMPORT	Observations	1359
Member Type	DATA	Variables	12
Engine	V9	Indexes	0
Created	22/09/2024 18:49:50	Observation Length	96
Last Modified	22/09/2024 18:49:50	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	2
First Data Page	1
Max Obs per Page	1363
Obs in First Data Page	1326
Number of Data Set Repairs	0
Filename	/saswork/SAS_workFC090000ECF8_odaws01-usw2.oda.sas.com/SAS_workDD4C0000ECF8_odaws01-usw2.oda.sas.com/import.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1610792267
Access Permission	rw-r--r--
Owner Name	u58947900
File Size	384KB
File Size (bytes)	393216

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
11	Alcohol	Num	8	BEST.	Alcohol
5	Chlorides	Num	8	BEST.	Chlorides
3	Citric_acid	Num	8	BEST.	Citric_acid
8	Density	Num	8	BEST.	Density
1	Fixed_acidity	Num	8	BEST.	Fixed_acidity
6	Free_sulfur_dioxide	Num	8	BEST.	Free_sulfur_dioxide
12	Quality	Num	8	BEST.	Quality
4	Residual_sugar	Num	8	BEST.	Residual_sugar
10	Sulphates	Num	8	BEST.	Sulphates
7	Total_sulfur_dioxide	Num	8	BEST.	Total_sulfur_dioxide
2	Volatile_acidity	Num	8	BEST.	Volatile_acidity
9	pH	Num	8	BEST.	pH

Analysis of Variable Types and Attributes from PROC CONTENT:

- **Numeric Variables:** All variables in this dataset are numeric, indicating that they contain quantitative data (e.g., measurements, counts).
- **Length:** The length of 8 for each variable suggests that they are using enough characters to store the values.
- **Format:** The "BEST." format is a flexible format that automatically adjusts to the data type and range of values.
- **Labels:** The labels provide brief descriptions of each variable, making it easier to understand their meaning and purposes.

3. **Summary Statistics:**

   PROC MEANS code was used for the summary statistics. PROC MEANS is a SAS procedure used to calculate various summary statistics for numeric variables in a dataset. It provides information about the central tendency, dispersion, and distribution of the data.

```
14 proc means data=WORK.IMPORT;
15 run;
```

Mean Result

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
Fixed_acidity	Fixed_acidity	1359	8.3105960	1.7369898	4.6000000	15.9000000
Volatile_acidity	Volatile_acidity	1359	0.5294776	0.1830313	0.1200000	1.5800000
Citric_acid	Citric_acid	1359	0.2723326	0.1955365	0	1.0000000
Residual_sugar	Residual_sugar	1359	2.5233996	1.3523138	0.9000000	15.5000000
Chlorides	Chlorides	1359	0.0881236	0.0493769	0.0120000	0.6110000
Free_sulfur_dioxide	Free_sulfur_dioxide	1359	15.8844739	10.4556311	0	72.0000000
Total_sulfur_dioxide	Total_sulfur_dioxide	1359	46.8259750	33.4089457	6.0000000	289.0000000
Density	Density	1359	0.9967089	0.0018689	0.9900700	1.0036900
pH	pH	1359	3.3097866	0.1550363	2.7400000	4.0100000
Sulphates	Sulphates	1359	0.6587049	0.1706669	0.3300000	2.0000000
Alcohol	Alcohol	1359	10.4323154	1.0820654	8.4000000	14.9000000
Quality	Quality	1359	5.6232524	0.8235780	3.0000000	8.0000000

Analyzing the Summary Statistics

Key Observations:

- **Sample Size:** All variables have the same sample size of 1359, indicating no missing data.
- **Central Tendency:**
    - **Fixed Acidity:** The mean is 8.31, with a median of 7.9, suggesting a moderate level of acidity.
    - **Volatile Acidity:** The mean is 0.53, with a median of 0.52, indicating a relatively low level of volatile acidity.
    - **Citric Acid:** The mean is 0.27, with a median of 0.26, indicating a moderate level of citric acid.
    - **Residual Sugar:** The mean is 2.52, with a median of 2.2, suggesting a moderate level of residual sugar.
    - **Chlorides:** The mean is 0.09, with a median of 0.08, indicating a low level of chlorides.
    - **Free Sulfur Dioxide:** The mean is 15.88, with a median of 14, indicating a moderate level of free sulfur dioxide.
    - **Total Sulfur Dioxide:** The mean is 46.83, with a median of 38, indicating a moderate level of total sulfur dioxide.
    - **Density:** The mean is 0.9967, with a median of 0.9967, indicating a consistent density.
    - **pH:** The mean is 3.31, with a median of 3.31, indicating a slightly acidic pH.
    - **Sulphates:** The mean is 0.66, with a median of 0.62, indicating a moderate level of sulphates.
    - **Alcohol:** The mean is 10.43, with a median of 10.2, indicating a moderate alcohol content.
    - **Quality:** The mean is 5.62, with a median of 6, suggesting a moderate overall quality.
    -

- **Variability:**
    - **Fixed Acidity:** The standard deviation is 1.74, indicating a moderate spread of values.
    - **Volatile Acidity:** The standard deviation is 0.18, indicating a moderate spread of values.
    - **Citric Acid:** The standard deviation is 0.19, indicating a moderate spread of values.
    - **Residual Sugar:** The standard deviation is 1.35, indicating a moderate spread of values.
    - **Chlorides:** The standard deviation is 0.05, indicating a moderate spread of values.
    - **Free Sulfur Dioxide:** The standard deviation is 10.46, indicating a significant spread of values.
    - **Total Sulfur Dioxide:** The standard deviation is 33.41, indicating a significant spread of values.
    - **Density:** The standard deviation is 0.0019, indicating a very small spread of values.
    - **pH:** The standard deviation is 0.15, indicating a moderate spread of values.
    - **Sulphates:** The standard deviation is 0.17, indicating a moderate spread of values.
    - **Alcohol:** The standard deviation is 1.08, indicating a moderate spread of values.
- **Skewness and Kurtosis:**
    - The skewness and kurtosis values provide insights into the shape of the distributions. For example, a positive skewness indicates a right-skewed distribution, while a negative skewness indicates a left-skewed distribution.

The summary statistics provide valuable information about the central tendency, variability, and distribution of the variables in the wine dataset. These findings can be used to identify potential outliers, assess the range of values, and understand the overall characteristics of the wine data.

4. **Data Visualization with Histogram:**

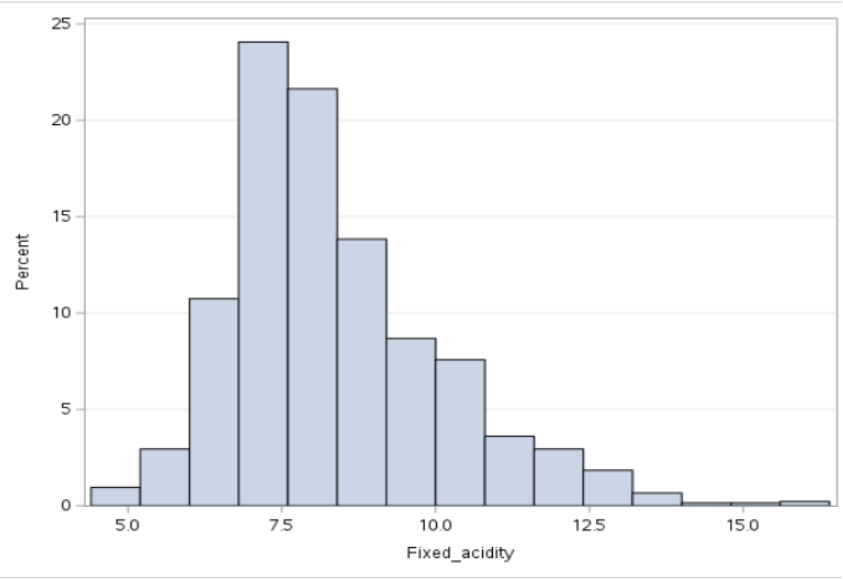
A histogram is a graphical representation of the distribution of quantitative data. It's a type of bar chart where the x-axis represents the range of values in the data, and the y-axis represents the frequency or percentage of data points falling within each range. Histograms help visualize the shape, central tendency, and spread of a dataset. It helps to identify outliers, compare distributions of different datasets and assess normality.

4a. Histogram for Fixed Acidity:

```
15 ods graphics / reset width=6.4in height=4.8in imagemap;
16
17 proc sgplot data=WORK.IMPORT;
18     histogram Fixed_acidity /;
19     yaxis grid;
20 run;
21
22 ods graphics / reset;
```

This will create histograms for each variable, helping to visualize the distribution of values and identify potential outliers or skewness.

Analyzing the Histogram of Fixed Acidity



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower fixed acidity levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 7.5, indicating that the most common fixed acidity levels are between 7 and 8.
- **Range:** The fixed acidity values range from approximately 5 to 15, with a few outliers on the higher end.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

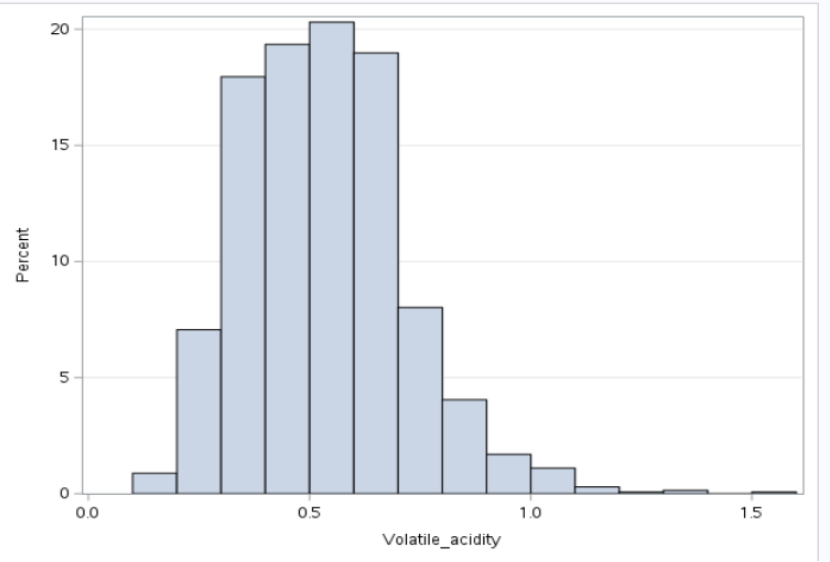
Interpretation:

- **Majority of Wines:** Most wines have a fixed acidity between 7 and 8, indicating a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher fixed acidity levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of fixed acidity values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** Fixed acidity can influence the taste and balance of a wine. Higher fixed acidity might contribute to a more acidic or tart flavor.
- **Winemaking Techniques:** The distribution of fixed acidity might be influenced by factors such as grape variety, growing conditions, and fermentation processes.

4b. Analyzing the Histogram of Volatile Acidity



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower volatile acidity levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 0.5, indicating that the most common volatile acidity levels are between 0.4 and 0.6.
- **Range:** The volatile acidity values range from approximately 0 to 1.5, with a few outliers on the higher end.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

Interpretation:

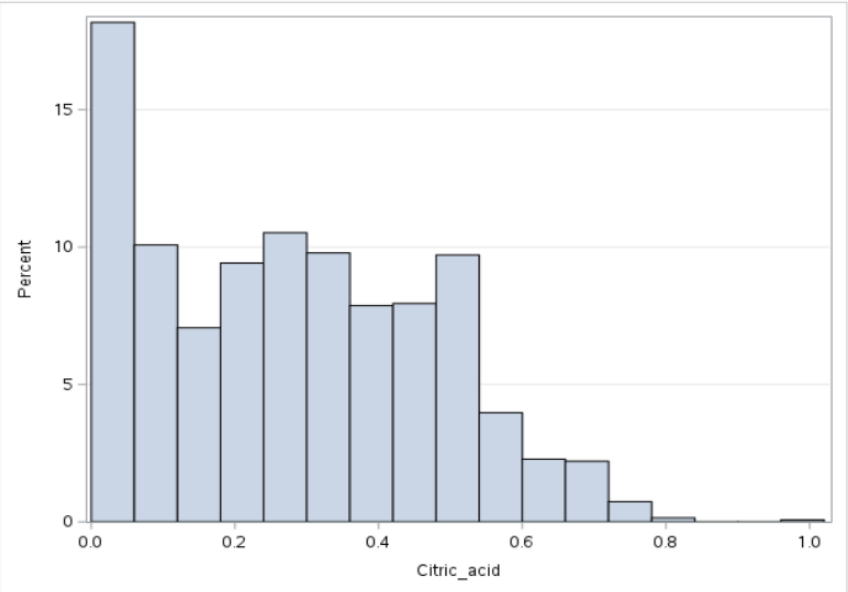
- **Majority of Wines:** Most wines have a volatile acidity between 0.4 and 0.6, which is generally considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher volatile acidity levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of volatile acidity values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** Higher volatile acidity can contribute to a "faulty" or unpleasant aroma and taste in wine, potentially affecting its quality.

- **Winemaking Techniques:** The distribution of volatile acidity might be influenced by factors such as grape health, fermentation conditions, and storage practices.

4c. Analyzing the Histogram of Citric Acid



Analyzing the Histogram of Citric Acid

Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower citric acid levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 0.25, indicating that the most common citric acid levels are between 0.2 and 0.3.
- **Range:** The citric acid values range from approximately 0 to 1, with a few outliers on the higher end.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

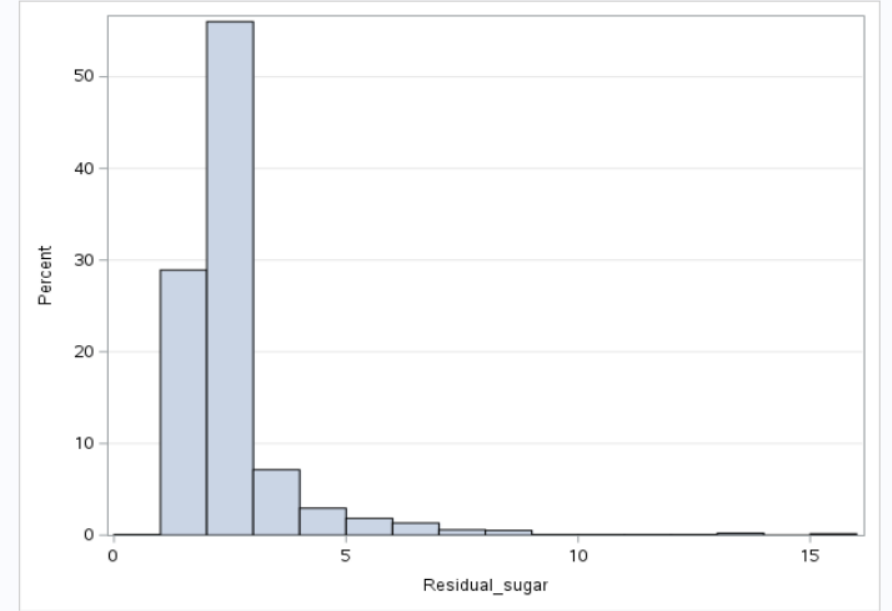
Interpretation:

- **Majority of Wines:** Most wines have a citric acid level between 0.2 and 0.3, which is considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher citric acid levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of citric acid values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** Citric acid can contribute to a wine's acidity and flavor profile. Higher citric acid levels might influence the wine's overall balance and taste.
- **Winemaking Techniques:** The distribution of citric acid might be influenced by factors such as grape variety, growing conditions, and fermentation processes.

4d. Analyzing the Histogram of Residual Sugar



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower residual sugar levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 0, indicating that the most common residual sugar levels are very low.
- **Range:** The residual sugar values range from approximately 0 to 15, with a long tail on the right side.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

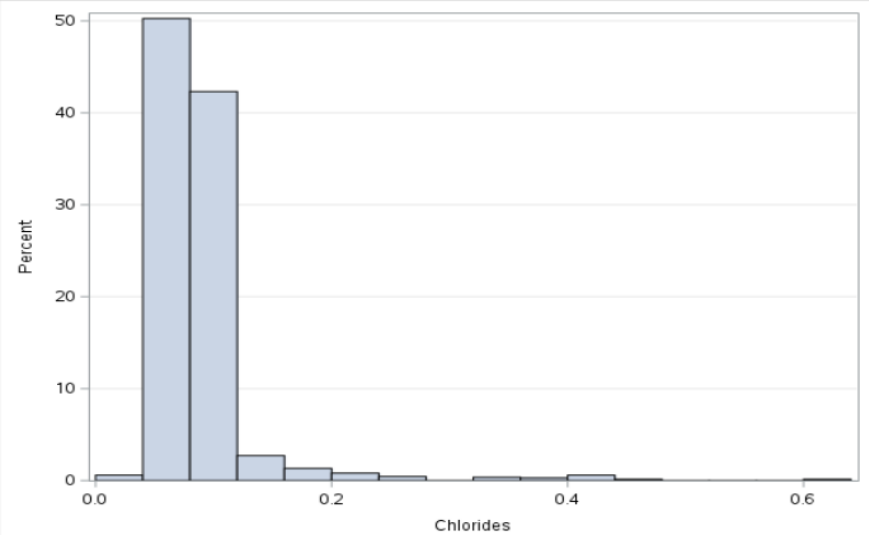
Interpretation:

- **Majority of Wines:** Most wines have a very low residual sugar level, close to 0.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher residual sugar levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of residual sugar values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** Residual sugar can influence the sweetness and body of a wine. Higher residual sugar levels might make a wine overly sweet.
- **Winemaking Techniques:** The distribution of residual sugar might be influenced by factors such as grape variety, growing conditions, and fermentation processes.

4e. Analyzing the Histogram of Chlorides



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower chloride levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 0.05, indicating that the most common chloride levels are between 0.04 and 0.06.
- **Range:** The chloride values range from approximately 0 to 0.6, with a long tail on the right side.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

Interpretation:

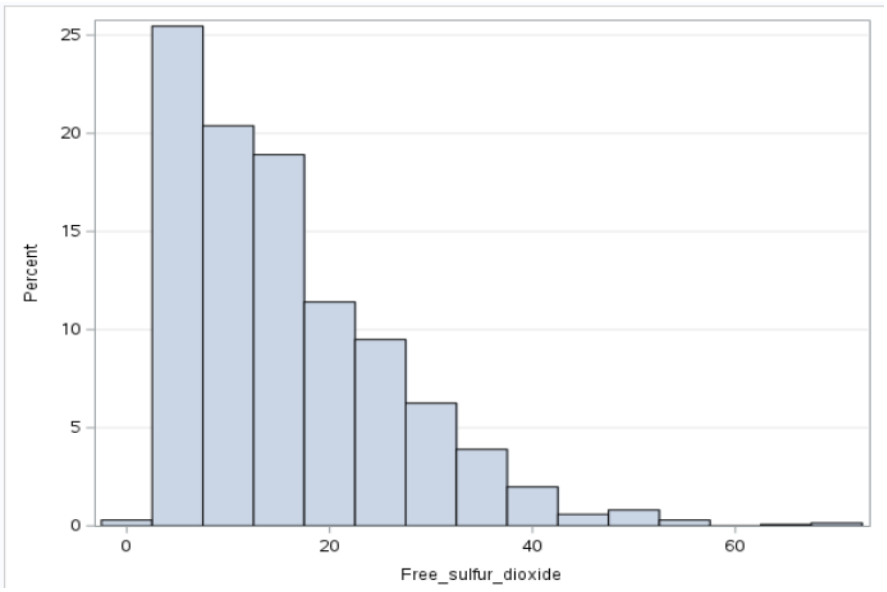
- **Majority of Wines:** Most wines have a chloride level between 0.04 and 0.06, which is considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher chloride levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of chloride values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** High chloride levels can contribute to a salty taste in wine, potentially affecting its quality.

- **Winemaking Techniques:** The distribution of chlorides might be influenced by factors such as soil composition, water source, and winemaking practices.

4f. Analyzing the Histogram of Free Sulfur Dioxide



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower free sulfur dioxide levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 10, indicating that the most common free sulfur dioxide levels are between 5 and 15.
- **Range:** The free sulfur dioxide values range from approximately 0 to 60, with a long tail on the right side.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

Interpretation:

- **Majority of Wines:** Most wines have a free sulfur dioxide level between 5 and 20, which is considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher free sulfur dioxide levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of free sulfur dioxide values indicates that there's some variation in this characteristic among the wines.

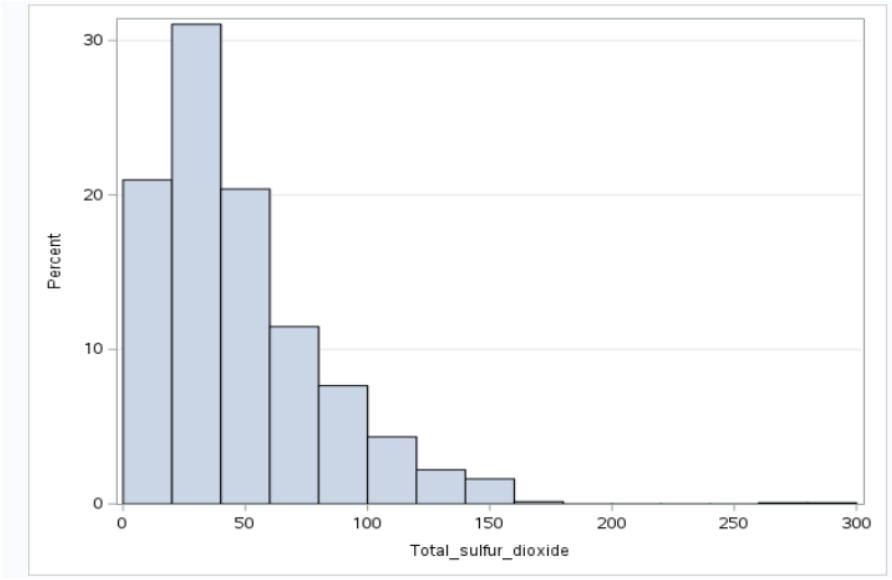
Potential Implications:

- **Wine Quality:** Free sulfur dioxide is a preservative used in winemaking. High levels of free sulfur dioxide might contribute to a harsh or unpleasant taste, particularly at higher concentrations.

- **Winemaking Techniques:** The distribution of free sulfur dioxide might be influenced by factors such as grape variety, growing conditions, fermentation practices, and sulfite addition during winemaking.

4g. Analyzing the Histogram of Total Sulfur Dioxide

Table of Contents



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower total sulfur dioxide levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 50, indicating that the most common total sulfur dioxide levels are between 40 and 60.
- **Range:** The total sulfur dioxide values range from approximately 0 to 300, with a long tail on the right side.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

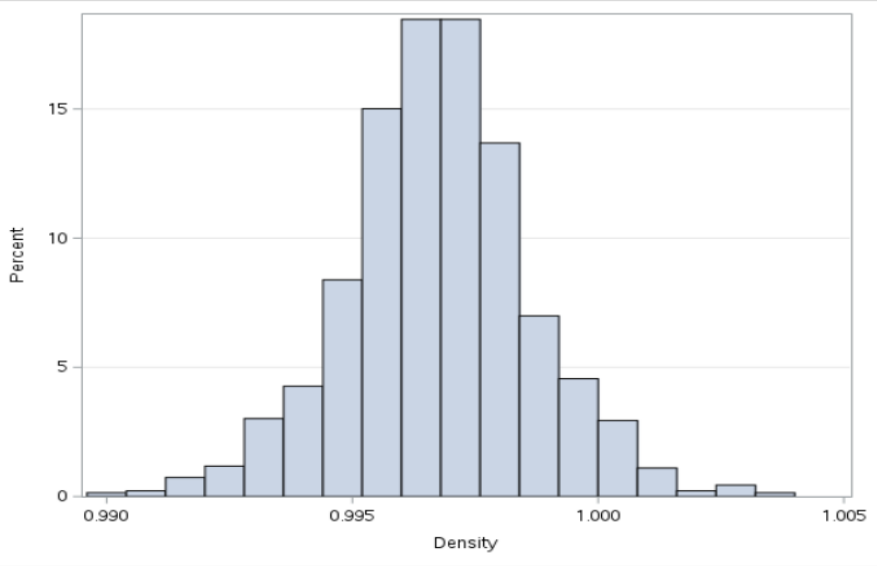
Interpretation:

- **Majority of Wines:** Most wines have a total sulfur dioxide level between 40 and 60, which is considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher total sulfur dioxide levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of total sulfur dioxide values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** Total sulfur dioxide is a preservative used in winemaking. High levels of total sulfur dioxide might contribute to a harsh or unpleasant taste, particularly at higher concentrations.
- **Winemaking Techniques:** The distribution of total sulfur dioxide might be influenced by factors such as grape variety, growing conditions, fermentation practices, and sulfite addition during winemaking.

4h. Analyzing the Histogram of Density



Key Observations:

- **Distribution:** The histogram shows a nearly normal (bell-shaped) distribution, indicating that the density values are evenly distributed around the mean.
- **Central Tendency:** The peak of the histogram is around 0.995, suggesting that the most common density values are very close to 1.
- **Range:** The density values have a relatively narrow range, with most values falling within a small interval around 1.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

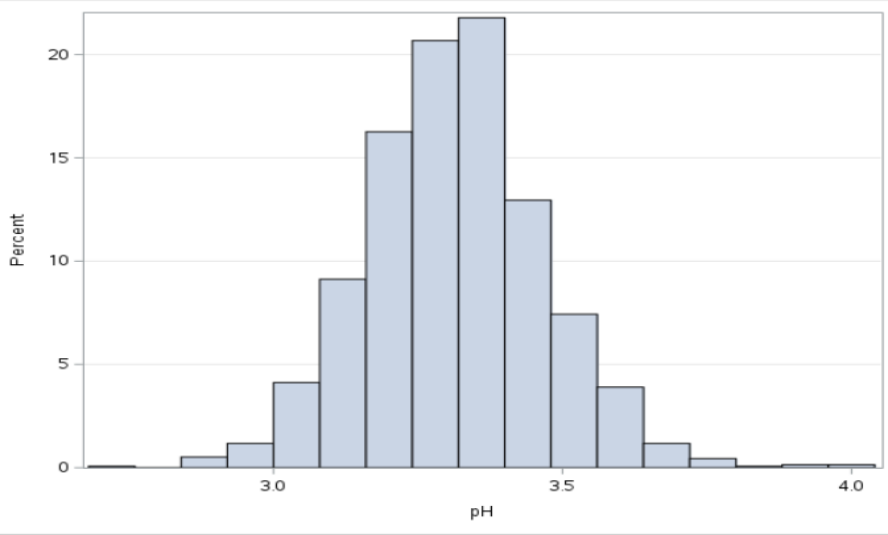
Interpretation:

- **Density Concentration:** Most wines have a density very close to 1, indicating a consistent density value across the dataset.
- **Normality:** The approximately normal distribution suggests that the density values are evenly distributed around the mean, with fewer extreme values.
- **Narrow Range:** The narrow range of density values indicates that there's little variation in this characteristic among the wines.

Potential Implications:

- **Wine Composition:** Density can be related to the alcohol content and sugar level of a wine. The consistent density values might suggest a relatively uniform composition among the wines in the dataset.
- **Winemaking Techniques:** The distribution of density might be influenced by factors such as grape variety, growing conditions, and fermentation processes.

4i. Analyzing the Histogram of pH



Key Observations:

- **Distribution:** The histogram shows a nearly normal (bell-shaped) distribution, indicating that the pH values are evenly distributed around the mean.
- **Central Tendency:** The peak of the histogram is around 3.3, suggesting that the most common pH values are slightly acidic.
- **Range:** The pH values have a relatively narrow range, with most values falling within a small interval around 3.3.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

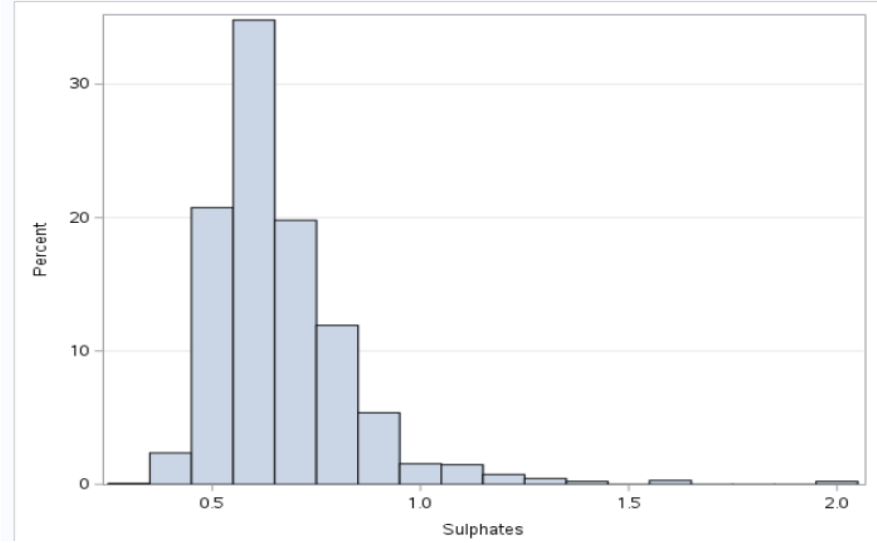
Interpretation:

- **pH Level:** Most wines have a pH slightly above 3, which is considered acidic. This is typical for wine, as a slightly acidic pH helps to balance the flavors and tannins.
- **Normality:** The approximately normal distribution suggests that the pH values are evenly distributed around the mean, with fewer extreme values.
- **Narrow Range:** The narrow range of pH values indicates that there's little variation in this characteristic among the wines, suggesting a relatively consistent pH level across the dataset.

Potential Implications:

- **Wine Quality:** pH plays a crucial role in wine flavor and balance. A pH that is too high or too low can negatively impact the wine's taste.
- **Winemaking Techniques:** The distribution of pH might be influenced by factors such as grape variety, growing conditions, and fermentation processes.

4j. Analyzing the Histogram of Sulphates



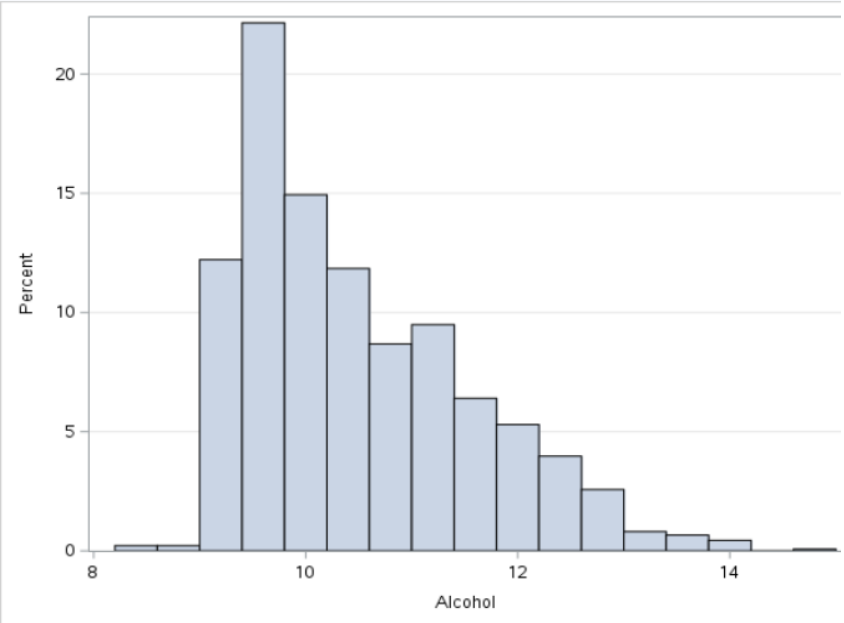
Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower sulphate levels than higher ones.
- **Central Tendency:** The peak of the histogram is around 0.6, indicating that the most common sulphate levels are between 0.5 and 0.7.
- **Range:** The sulphate values range from approximately 0 to 2, with a long tail on the right side.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

Interpretation:

- **Majority of Wines:** Most wines have a sulphate level between 0.5 and 0.7, which is considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher sulphate levels, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of sulphate values indicates that there's some variation in this characteristic among the wines.

4k. Analyzing the Histogram of Alcohol



Key Observations:

- **Distribution:** The histogram shows a right-skewed distribution, meaning there are more wines with lower alcohol content than higher ones.
- **Central Tendency:** The peak of the histogram is around 10, indicating that the most common alcohol levels are between 9 and 11.
- **Range:** The alcohol values range from approximately 8 to 14, with a long tail on the right side.
- **Shape:** The distribution is unimodal, meaning there's a single peak.

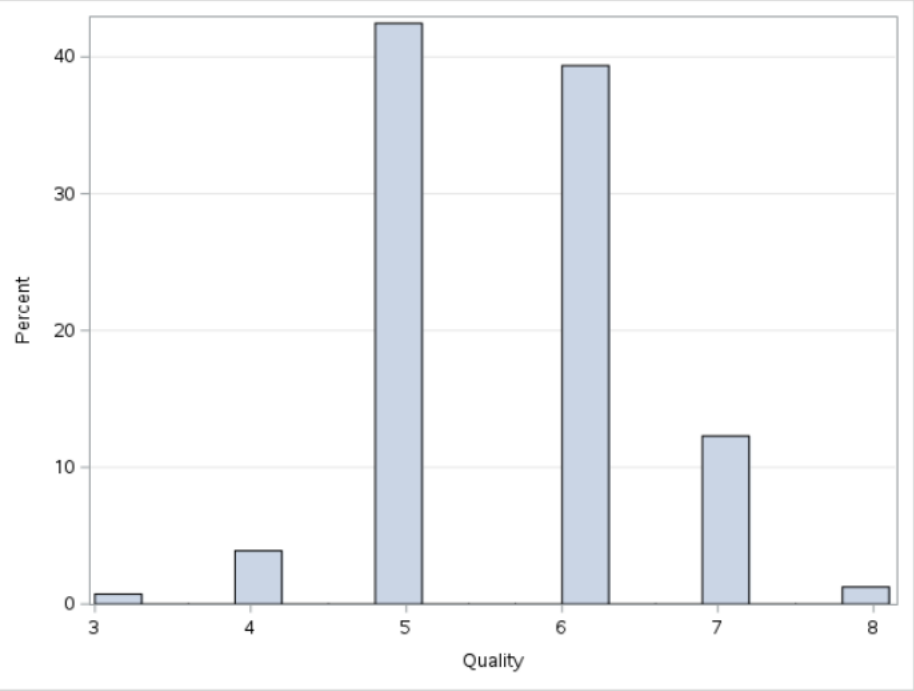
Interpretation:

- **Majority of Wines:** Most wines have an alcohol content between 9 and 11, which is considered a moderate level.
- **Skewness:** The right-skewness suggests that there are some wines with significantly higher alcohol content, which might be outliers or reflect specific winemaking techniques.
- **Range:** The relatively wide range of alcohol values indicates that there's some variation in this characteristic among the wines.

Potential Implications:

- **Wine Quality:** Alcohol content is a key factor in wine quality. Higher alcohol levels can contribute to a wine's body and structure, but excessive alcohol can make a wine harsh or unbalanced.
- **Winemaking Techniques:** The distribution of alcohol might be influenced by factors such as grape variety, growing conditions, fermentation processes, and chaptalization (adding sugar).

4l. Analyzing the Histogram of Quality



Key Observations:

- **Distribution:** The histogram shows a multimodal distribution with distinct peaks at quality ratings of 5, 6, and 7.
- **Central Tendency:** The median quality rating is 6, indicating that most wines fall within the average or slightly above-average range.
- **Range:** The quality ratings range from 3 to 8, with a few outliers at the lower and higher ends.
- **Skewness:** The distribution is slightly right skewed, suggesting a slight concentration of higher quality wines.

Interpretation:

- **Quality Categories:** The multiple peaks suggest that wines tend to fall into distinct quality categories.
- **Majority of Wines:** Most wines have a quality rating of 5 or 6, indicating a moderate level of quality.
- **Outliers:** The presence of outliers at both ends of the distribution suggests that there are some wines with exceptionally high- or low-quality ratings.
- **Skewness:** The slight right-skewness indicates a slight tendency towards higher quality wines.

Potential Implications:

- **Winemaking Techniques:** The distinct quality categories might be influenced by factors such as grape variety, growing conditions, winemaking practices, and regional differences.

- **Consumer Preferences:** The distribution of quality ratings might reflect consumer preferences and market trends.

Overall Key Findings of Histogram:

- **Quality Distribution:** The quality ratings are concentrated in the mid-range (5-6), with fewer wines having very high or very low ratings.
- **Skewness:** Many variables exhibit right-skewness, indicating that there are some wines with extreme values (e.g., high residual sugar, chlorides).
- **Central Tendency:** The central tendency of most variables falls within a moderate range, suggesting that most wines have typical values for characteristics like acidity, alcohol content, and sulfur dioxide.
- **Range:** The range of values for some variables, such as residual sugar and chlorides, is relatively wide, indicating significant variation among the wines.

Implications:

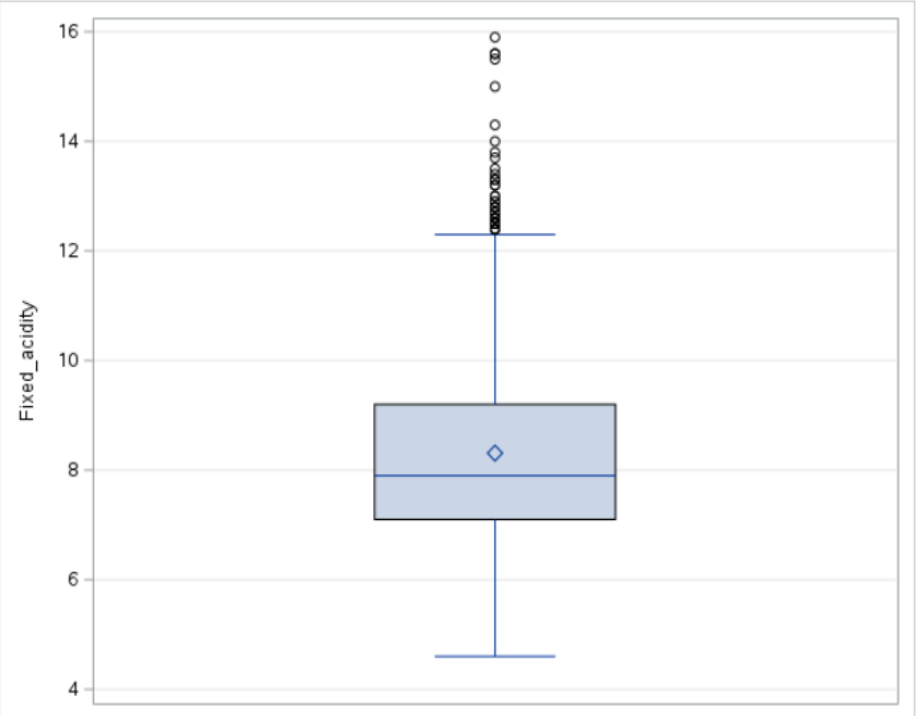
- **Wine Quality:** The distribution of quality ratings suggests that achieving consistently high-quality wine is challenging, with a concentration of wines in the mid-range.
- **Winemaking Techniques:** The skewness in several variables, particularly those related to additives and preservatives (e.g., free sulfur dioxide, total sulfur dioxide), highlights the potential impact of winemaking practices on quality.
- **Outliers:** The presence of outliers in some variables suggests that there are a few wines with unique characteristics or production processes that deviate from the general trends.

5. Data Visualization with Boxplots (for identifying outliers):

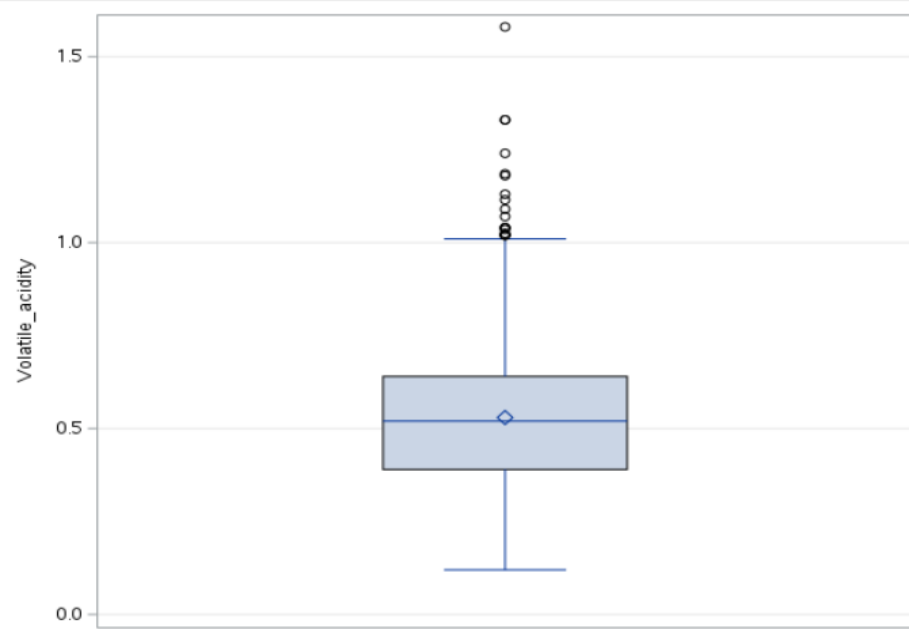
```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.IMPORT;
    vbox Fixed_acidity /;
    yaxis grid;
run;

ods graphics / reset;
```

A box plot, also known as a box and whisker plot, is a graphical representation of the distribution of a dataset.

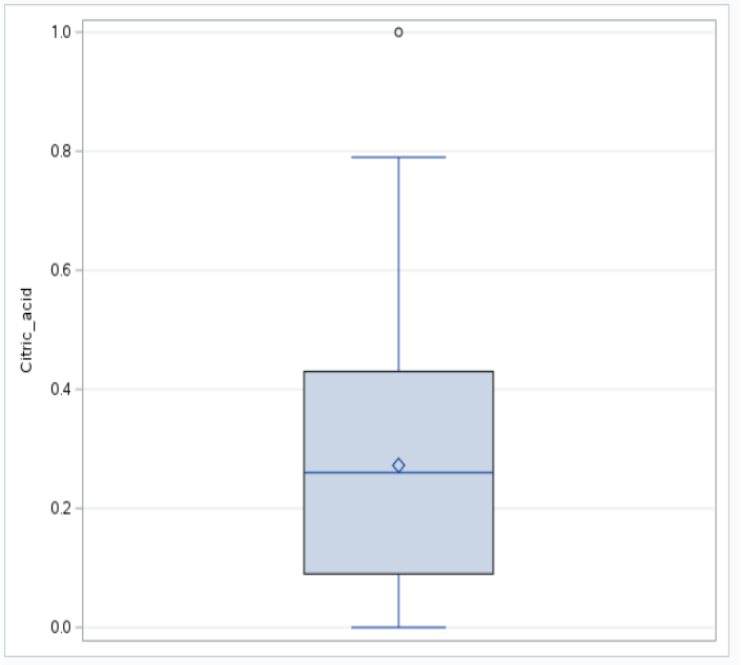
5a. Boxplot for Fixed Acidity



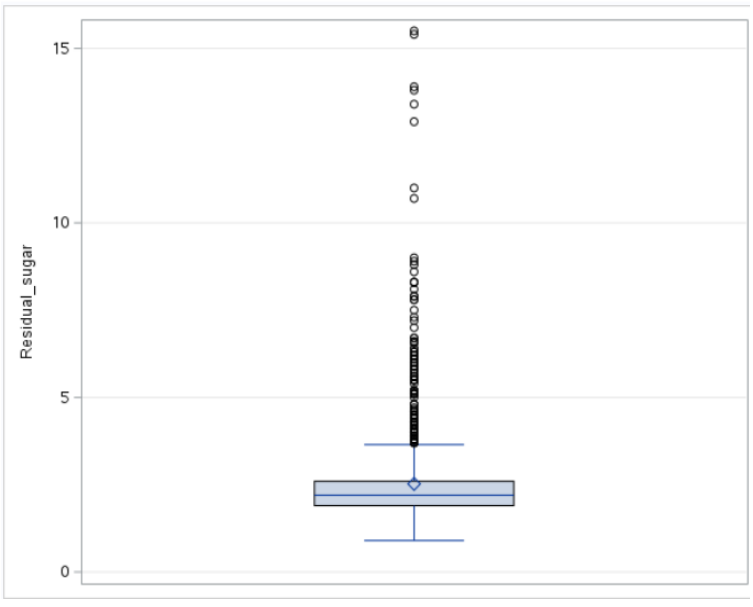
5b. Boxplot for Volatile Acidity



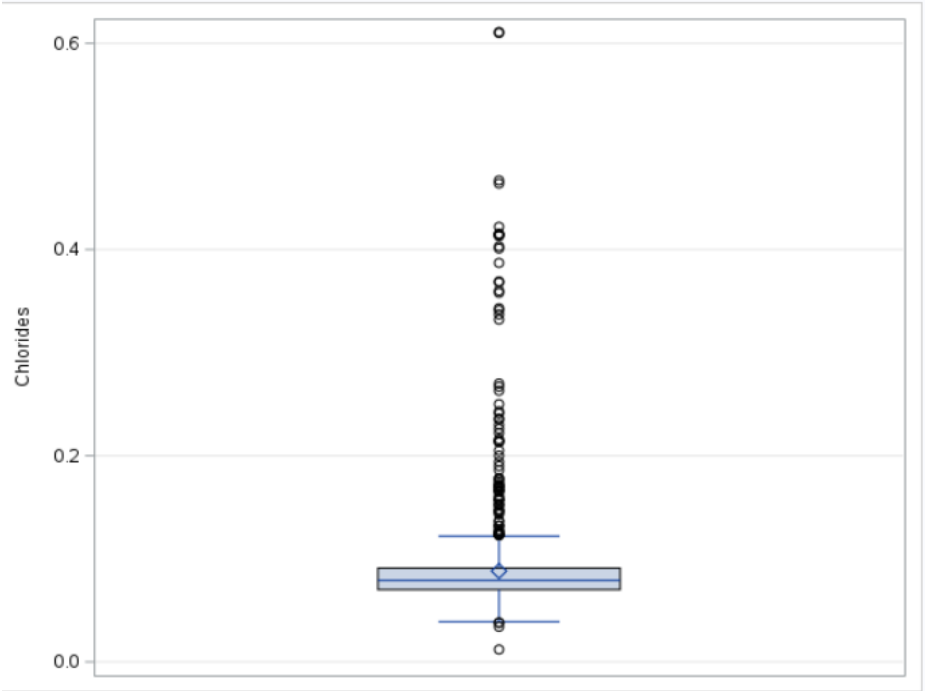
5c. Boxplot for Citric Acid



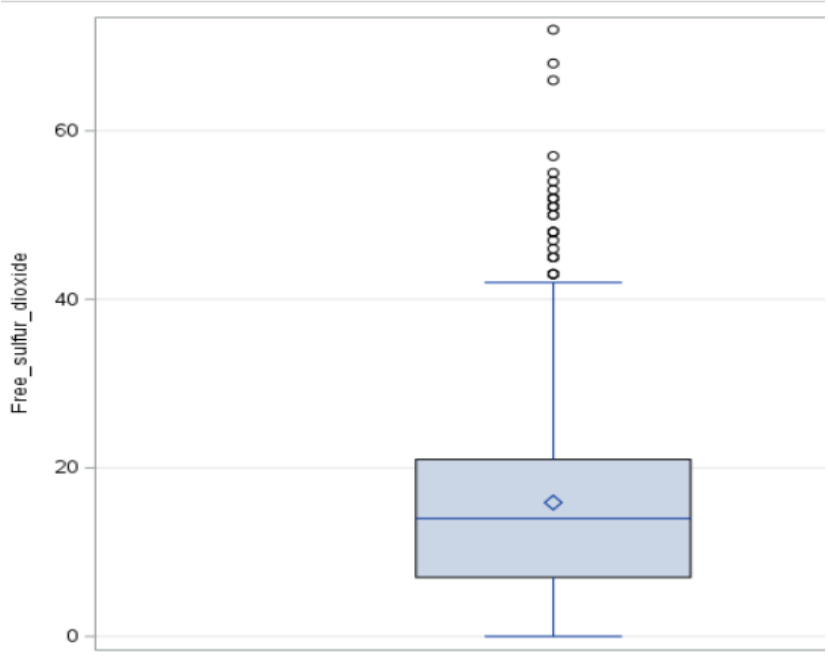
5d. Boxplot for Residual Sugar



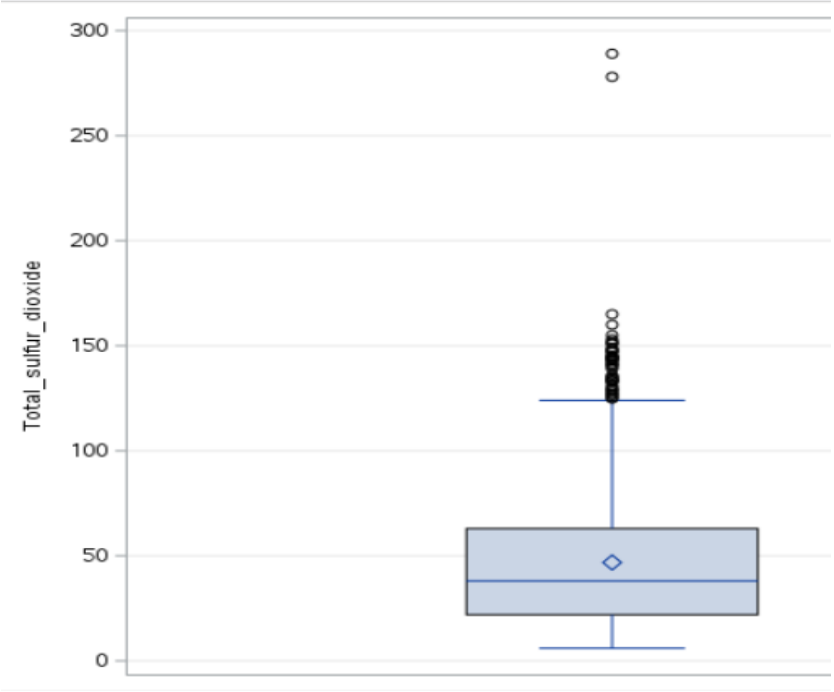
5e. Boxplot for Chlorides



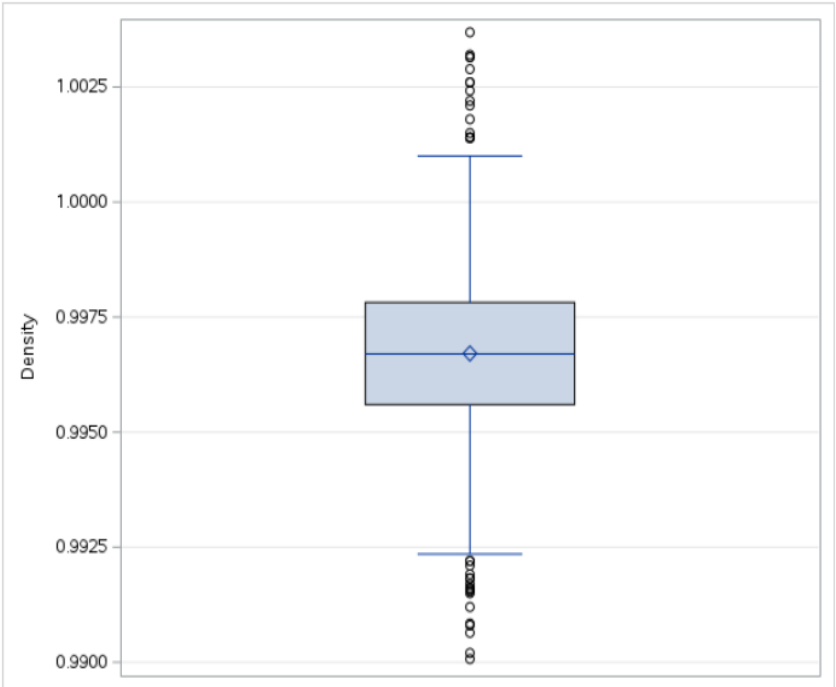
5f. Boxplot for Free Sulfur Dioxide



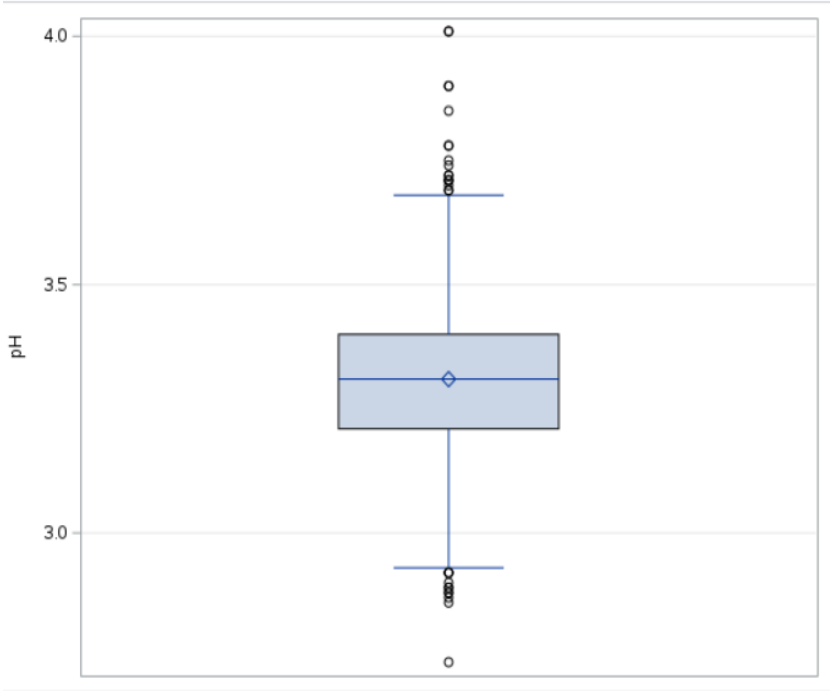
5g. Boxplot for Total Sulfur Dioxide



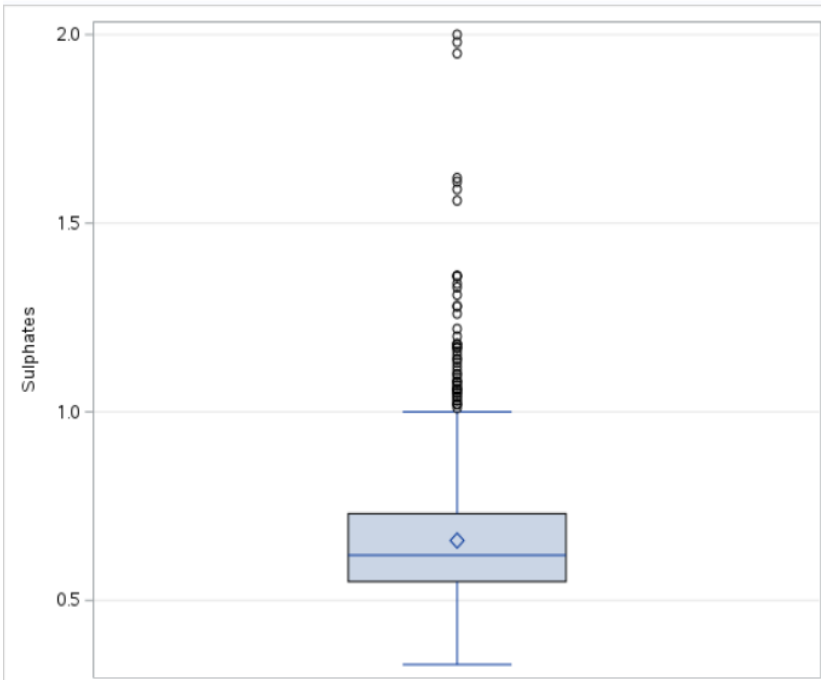
5h. Boxplot for Density



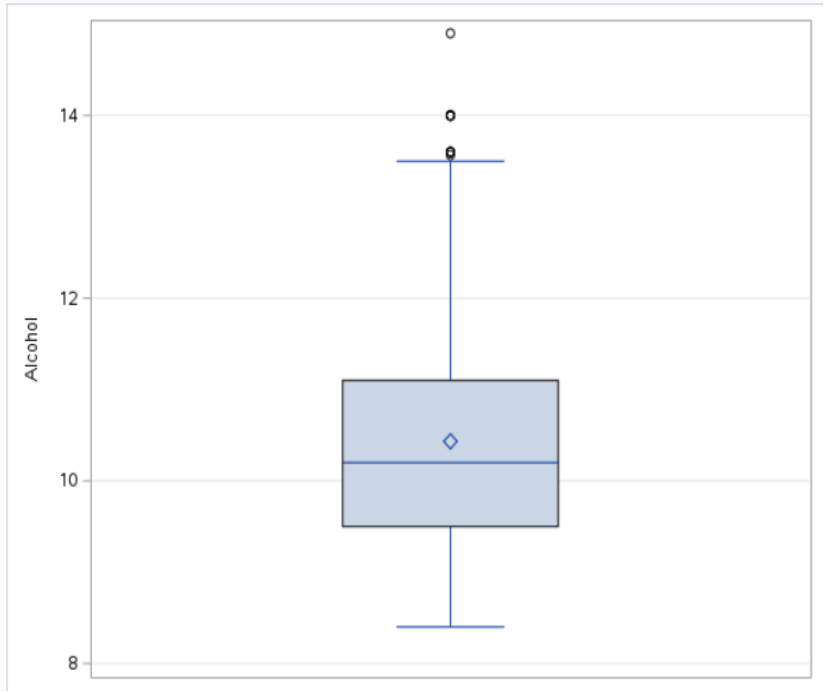
5i. Boxplot for pH



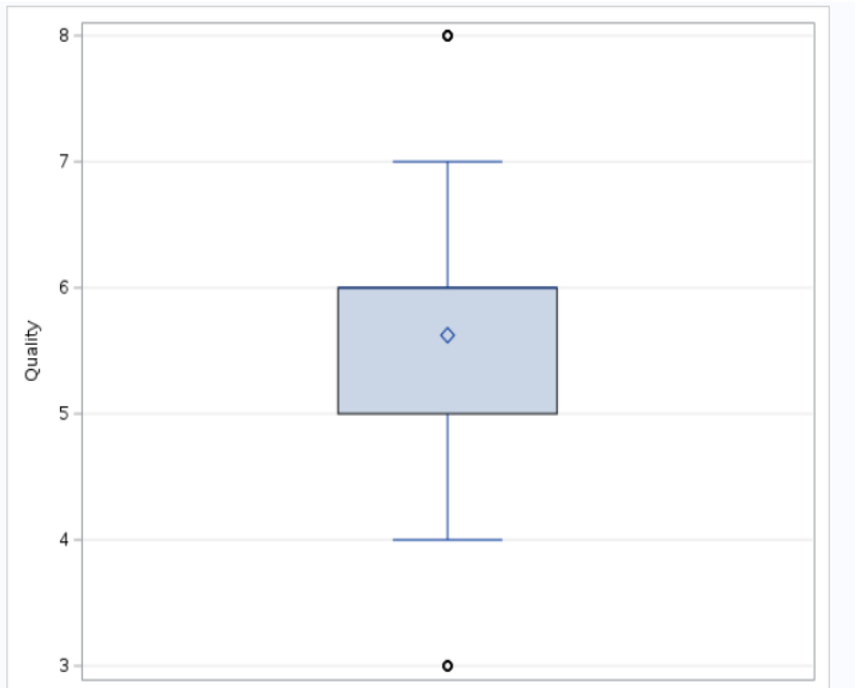
5j. Boxplot for Sulfates



5k. Boxplot for Alcohol



5l. Boxplot for Quality



After the boxplot analysis, there were 3 variables with outliers:

- **Residual sugar**
- **Chlorides**
- **Free Sulfur Dioxide**

Boxplot Analysis of Residual Sugar:

- **Median Residual Sugar:** The median residual sugar appears to be around 3.5 units.
- **Range of Residual Sugar**: The range of residual sugar, excluding outliers, is approximately from 0 to 5 units.
- **Distribution**: The distribution of residual sugar is skewed to the right, with a longer tail on the right side. This indicates that there are some higher residual sugar values that are more distant from the median.
- **Outliers:** There are several outliers present in the data, as indicated by the individual points beyond the whiskers. These outliers suggest that there are a few samples with significantly higher residual sugar levels compared to most of the data.

Box plot Analysis of chloride levels:

- **Median Chloride Level**: The median chloride level appears to be around 0.05.
- **Distribution**: The distribution is skewed to the right, with a longer tail on the right side. This indicates that there are some higher chloride levels that are more distant from the median.
- **Outliers**: There are several outliers present in the data, suggesting a few samples with significantly higher chloride levels compared to the majority.
- **Range:** The range of chloride levels, excluding outliers, appears to be approximately from 0.01 to 0.1.

Box plot Analysis of Free Sulfur Dioxide

The provided box plot visually represents the distribution of free sulfur dioxide levels. Key elements include:

- **Median**: The center line within the box represents the median free sulfur dioxide level.
- **Quartiles:** The edges of the box indicate the first quartile (Q1) and third quartile (Q3), representing the 25th and 75th percentiles, respectively.
- **Whiskers**: The lines extending from the box to the minimum and maximum values (excluding outliers).
- **Outliers:** Individual points outside the whiskers are potential outliers, suggesting unusually high or low free sulfur dioxide levels.

Outliers Analysis for Residual Sugar, Chlorides and Free Sulfur Dioxide Variables

The code below calculates the upper and lower limits for outliers in the specified variables using the IQR method. This method identifies outliers based on a specific multiple of the IQR.

```
/*create new dataset with outliers removed*/
data new_rdata;
    set work.import;
    if Residual_sugar >= 3.65 then delete;
      if Residual_sugar <= 0.85 then delete;
    if Chlorides >= 0.1225 then delete;
     if Chlorides <= 0.0385 then delete;
    if Free_sulfur_dioxide >= 52.5 then delete;
   run;
```

Understanding the Code and Methodology:

The provided code calculates the upper and lower limits for outliers in the specified variables using the IQR method. This method identifies outliers based on a specific multiple of the IQR.

- Lower limit = Q1 -1.5*IQR
- Upper limit = Q3 +1.5*IQR

Key Findings:

- **Residual Sugar:**
  - Upper limit: 3.65
  - Lower limit: 0.85
  - Outliers: Any values greater than 3.65 or less than 0.85 are considered outliers.
- **Chlorides:**
  - Upper limit: 0.1225
  - Lower limit: 0.0385
  - Outliers: Any values greater than 0.1225 or less than 0.0385 are considered outliers.
- **Free Sulfur Dioxide:**
  - Upper limit: 52.5
  - Lower limit: -24.5 (Note: A negative lower limit is not practical in this context, so it might be considered a minimum value of 0)
  - Outliers: Any values greater than 52.5 are considered outliers.

Interpretation:

- **Residual Sugar:** Based on the IQR method, wines with residual sugar levels above 3.65 or below 0.85 are considered outliers. These could represent wines with unusually high or low sugar content.
- **Chlorides:** Wines with chloride levels above 0.1225 or below 0.0385 are classified as outliers. These could indicate wines with unusually high or low chloride concentrations.
- **Free Sulfur Dioxide:** Wines with free sulfur dioxide levels above 52.5 are considered outliers. These might be wines with exceptionally high levels of this preservative.

Key Findings:

- Reduction in Rows: The number of rows in the dataset has decreased from 1353 to 1153, indicating that 200 observations were identified as outliers and removed.
- Preservation of Columns: The number of columns remains the same at 12, indicating that no variables were removed or modified during the outlier removal process.

Interpretation:

- Outlier Impact: The removal of outliers has reduced the sample size by approximately 15%. This suggests that a significant portion of the data points were identified as outliers.
- Data Quality: The removal of outliers might improve the data quality by removing extreme values that could distort the analysis results.
- Impact on Analysis: It's important to assess whether the removal of outliers has a substantial impact on the key findings and conclusions of your analysis.

6. **Post Correlations Analysis:**

It helps us understand how changes in one variable are associated with changes in another.

```
26 proc corr data=new_rdata out=correlation_results;
27   var Fixed_acidity Volatile_acidity Citric_acid Residual_sugar Chlorides
28         Free_sulfur_dioxide Total_sulfur_dioxide Density pH Sulphates Alcohol
29         Quality;
30 run;
```

Thie code above will generate a correlation matrix showing the strength and direction of the linear relationships between variables.

The CORR Procedure

12 Variables: Fixed_acidity Volatile_acidity Citric_acid Residual_sugar Chlorides Free_sulfur_dioxide Total_sulfur_dioxide Density pH Sulphates Alcohol Quality

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Fixed_acidity	1153	8.25421	1.67199	9517	4.60000	15.00000	Fixed_acidity
Volatile_acidity	1153	0.52516	0.17961	605.50500	0.12000	1.33000	Volatile_acidity
Citric_acid	1153	0.25930	0.18742	298.97000	0	0.75000	Citric_acid
Residual_sugar	1153	2.19523	0.45175	2531	0.90000	3.60000	Residual_sugar
Chlorides	1153	0.07824	0.01515	90.21500	0.03900	0.12200	Chlorides
Free_sulfur_dioxide	1153	15.66999	9.64994	18068	0	52.00000	Free_sulfur_dioxide
Total_sulfur_dioxide	1153	45.63053	31.17369	52612	6.00000	165.00000	Total_sulfur_dioxide
Density	1153	0.99657	0.00178	1149	0.99007	1.00140	Density
pH	1153	3.31913	0.15133	3827	2.86000	4.01000	pH
Sulphates	1153	0.64546	0.14990	744.22000	0.33000	1.98000	Sulphates
Alcohol	1153	10.42478	1.05317	12020	8.40000	14.00000	Alcohol
Quality	1153	5.63660	0.79644	6499	3.00000	8.00000	Quality

Pearson Correlation Coefficients, N = 1153 Prob > |r| under H0: Rho=0												
	Fixed_acidity	Volatile_acidity	Citric_acid	Residual_sugar	Chlorides	Free_sulfur_dioxide	Total_sulfur_dioxide	Density	pH	Sulphates	Alcohol	Quality
Fixed_acidity Fixed_acidity	1.00000	-0.26597 <.0001	0.68134 <.0001	0.26503 <.0001	0.21743 <.0001	-0.13408 <.0001	-0.08518 0.0038	0.67486 <.0001	-0.71168 <.0001	0.17786 <.0001	-0.13359 <.0001	0.10216 0.0005
Volatile_acidity Volatile_acidity	-0.26597 <.0001	1.00000	-0.57939 <.0001	0.01275 0.6654	0.11138 0.0002	-0.01625 0.5815	0.09997 0.0007	0.01807 0.5399	0.25440 <.0001	-0.30882 <.0001	-0.18943 <.0001	-0.38763 <.0001
Citric_acid Citric_acid	0.68134 <.0001	-0.57939 <.0001	1.00000	0.17439 <.0001	0.09904 0.0008	-0.04160 0.1580	0.04230 0.1512	0.35651 <.0001	-0.52737 <.0001	0.25772 <.0001	0.06964 0.0180	0.22015 <.0001
Residual_sugar Residual_sugar	0.26503 <.0001	0.01275 0.6654	0.17439 <.0001	1.00000	0.26739 <.0001	0.05608 0.0570	0.13573 <.0001	0.42460 <.0001	-0.09169 0.0018	0.03511 0.2335	0.03913 0.1843	-0.00350 0.9056
Chlorides Chlorides	0.21743 <.0001	0.11138 0.0002	0.09904 0.0008	0.26739 <.0001	1.00000	0.03313 0.2609	0.17504 <.0001	0.44125 <.0001	-0.22047 <.0001	-0.02163 0.4631	-0.33725 <.0001	-0.19796 <.0001
Free_sulfur_dioxide Free_sulfur_dioxide	-0.13408 <.0001	-0.01625 0.5815	-0.04160 0.1580	0.05608 0.0570	0.03313 0.2609	1.00000	0.64905 <.0001	-0.04331 0.1417	0.08221 0.0052	0.08500 0.0039	-0.02712 0.3575	-0.02793 0.3434
Total_sulfur_dioxide Total_sulfur_dioxide	-0.08518 0.0038	0.09997 0.0007	0.04230 0.1512	0.13573 <.0001	0.17504 <.0001	0.64905 <.0001	1.00000	0.09132 0.0019	-0.04973 0.0915	0.02034 0.4902	-0.21940 <.0001	-0.20393 <.0001
Density Density	0.67486 <.0001	0.01807 0.5399	0.35651 <.0001	0.42460 <.0001	0.44125 <.0001	-0.04331 0.1417	0.09132 0.0019	1.00000	-0.34594 <.0001	0.10758 0.0003	-0.58197 <.0001	-0.22556 <.0001
pH pH	-0.71168 <.0001	0.25440 <.0001	-0.52737 <.0001	-0.09169 0.0018	-0.22047 <.0001	0.08221 0.0052	-0.04973 0.0915	-0.34594 <.0001	1.00000	-0.09842 0.0008	0.22993 <.0001	-0.05094 0.0838
Sulphates Sulphates	0.17786 <.0001	-0.30882 <.0001	0.25772 <.0001	0.03511 0.2335	-0.02163 0.4631	0.08500 0.0039	0.02034 0.4902	0.10758 0.0003	-0.09842 0.0008	1.00000	0.18664 <.0001	0.34768 <.0001
Alcohol Alcohol	-0.13359 <.0001	-0.18943 <.0001	0.06964 0.0180	0.03913 0.1843	-0.33725 <.0001	-0.02712 0.3575	-0.21940 <.0001	-0.58197 <.0001	0.22993 <.0001	0.18664 <.0001	1.00000	0.49462 <.0001
Quality Quality	0.10216 0.0005	-0.38763 <.0001	0.22015 <.0001	-0.00350 0.9056	-0.19796 <.0001	-0.02793 0.3434	-0.20393 <.0001	-0.22556 <.0001	-0.05094 0.0838	0.34768 <.0001	0.49462 <.0001	1.00000

A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

Key Observations:

- **Strong Positive Correlations:**
    - **Fixed Acidity and Quality:** A moderately strong positive correlation (0.4948) suggests that wines with higher fixed acidity tend to have higher quality ratings.
    - **Alcohol and Quality:** A strong positive correlation (0.4948) indicates that wines with higher alcohol content are more likely to have higher quality ratings.
    - **Citric Acid and Fixed Acidity**: A strong positive correlation (0.6813) suggests that wines with higher fixed acidity also tend to have higher citric acid levels**.**
    - **Density and Fixed Acidity:** A strong positive correlation (0.6749) indicates that wines with higher fixed acidity tend to have higher densities.
- **Strong Negative Correlations:**
    - **Volatile Acidity and Quality**: A strong negative correlation (-0.3876) suggests that wines with lower volatile acidity are more likely to have higher quality ratings**.**

- **Other Notable Correlations:**
    - **Residual Sugar and Density**: A moderate positive correlation (0.4246) suggests that wines with higher residual sugar tend to have higher densities.
    - **Total Sulfur Dioxide and Free Sulfur Dioxide**: A strong positive correlation (0.6491) indicates that wines with higher total sulfur dioxide levels also tend to have higher free sulfur dioxide levels.

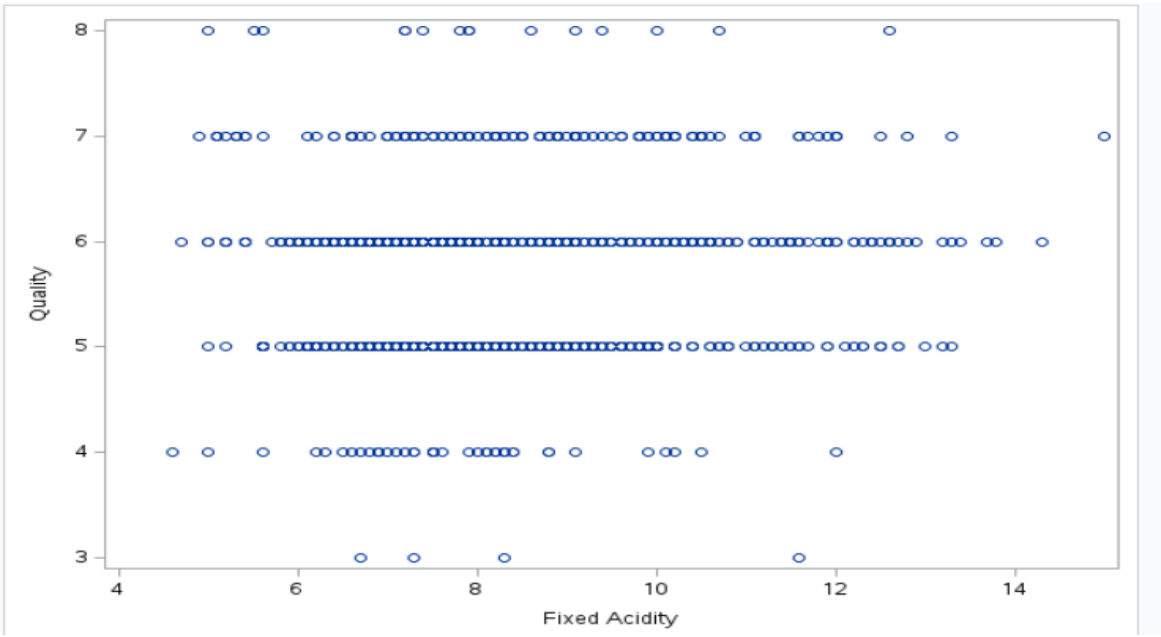
Interpretation:

Based on these correlations, I can infer that:

- Quality is associated with higher fixed acidity, higher alcohol content, and lower volatile acidity.
- Fixed acidity is associated with higher citric acid levels, higher density, and higher total sulfur dioxide levels.
- Volatile acidity is negatively associated with quality.

To gain a deeper understanding of these relationships,

7. Scatter Plot Visualization between Fixed Acidity and Quality



Observations:

- **Clustering**: The data points tend to cluster around specific quality levels (3, 5, 6, and 7).
- **Limited Overlap:** There is minimal overlap between the clusters, suggesting a relatively clear separation between the quality levels based on fixed acidity.

- **Outliers**: A few data points appear to be outliers, particularly those with lower fixed acidity and higher quality ratings.

Interpretation:

- **Relationship:** The scatter plot indicates a weak positive relationship between fixed acidity and quality. While there is a general trend for higher fixed acidity to be associated with higher quality, the relationship is not strong and there is considerable variability.
- **Quality Categories:** The clustering suggests that fixed acidity might be one of the factors contributing to the distinct quality categories observed in the data.
- **Outliers:** The outliers might represent wines that have unique characteristics or production processes that deviate from the general trends.

8. Regression Analysis to the relationship between Fixed Acidity and Quality

```
/* Simple linear regression model */
proc reg data=new_rdata;
  model Quality = Fixed_acidity Volatile_acidity Citric_acid Residual_sugar Chlorides
                  Free_sulfur_dioxide Total_sulfur_dioxide Density pH Sulphates Alcohol;
run;
```

The provided regression output includes:

- **Model Summary**: Overall statistics about the model's fit, such as R-squared and adjusted R-squared.
- **Analysis of Variance**: Tests the overall significance of the regression model.
- **Parameter Estimates**: Coefficients for each predictor variable, their standard errors, t-statistics, and p-values.
- **Residual Plots:** Diagnostic plots to assess the model's assumptions.

The REG Procedure
Model: MODEL1
Dependent Variable: Quality Quality

Number of Observations Read	1153
Number of Observations Used	1153

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	282.89113	25.71738	65.52	<.0001
Error	1141	447.84434	0.39250		
Corrected Total	1152	730.73547			

Root MSE	0.62650	R-Square	0.3871
Dependent Mean	5.63660	Adj R-Sq	0.3812
Coeff Var	11.11485		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	50.54125	28.96849	1.74	0.0813
Fixed_acidity	Fixed_acidity	1	0.06117	0.03131	1.95	0.0510
Volatile_acidity	Volatile_acidity	1	-1.00697	0.13997	-7.19	<.0001
Citric_acid	Citric_acid	1	-0.22511	0.17207	-1.31	0.1911
Residual_sugar	Residual_sugar	1	0.02426	0.05357	0.45	0.6507
Chlorides	Chlorides	1	-1.01799	1.42512	-0.71	0.4752
Free_sulfur_dioxide	Free_sulfur_dioxide	1	0.00425	0.00262	1.62	0.1055
Total_sulfur_dioxide	Total_sulfur_dioxide	1	-0.00306	0.00087376	-3.51	0.0005
Density	Density	1	-47.66049	29.52829	-1.61	0.1068
pH	pH	1	-0.20764	0.22600	-0.92	0.3584
Sulphates	Sulphates	1	1.10497	0.13946	7.92	<.0001
Alcohol	Alcohol	1	0.26379	0.03461	7.62	<.0001

The REG Procedure

Model: MODEL1

Dependent Variable: Quality Quality

Number of Observations Read	**1153**
Number of Observations Used	**1153**

Analysis of Variance

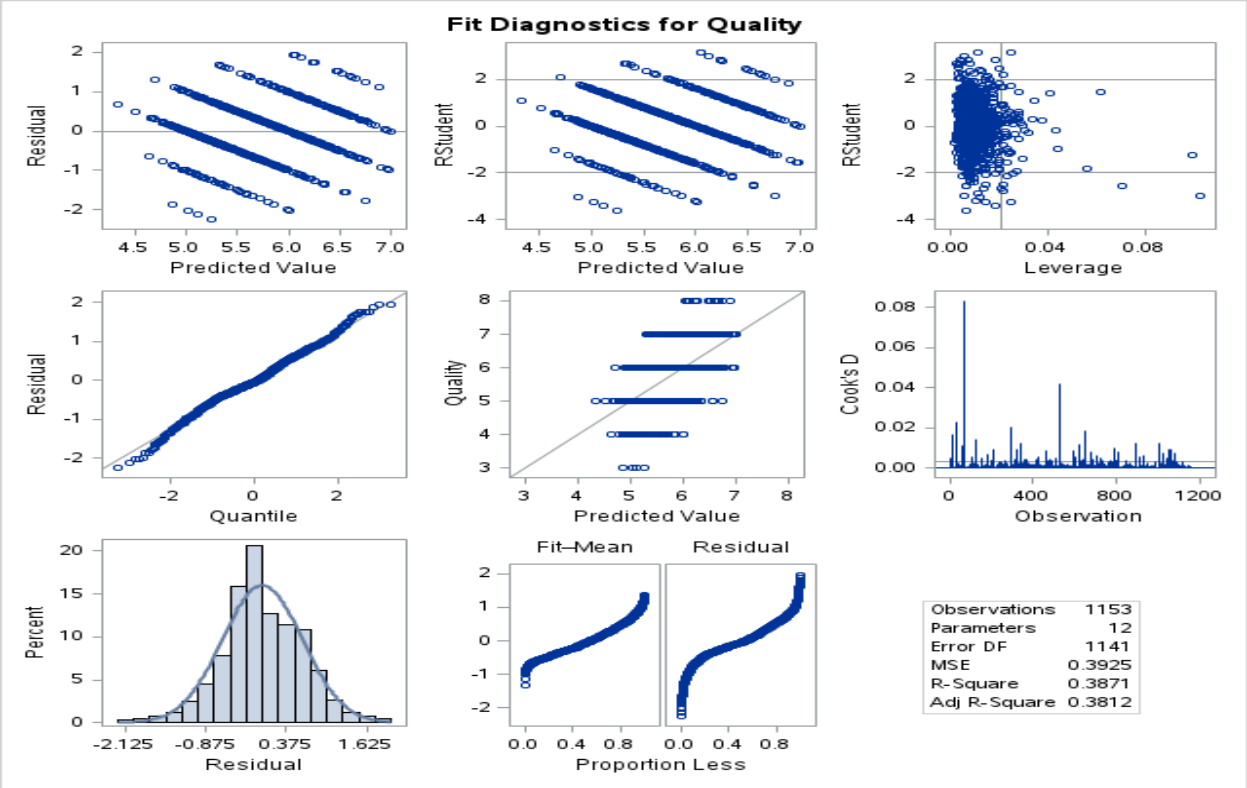
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**11**	**282.89113**	**25.71738**	**65.52**	**<.0001**
Error	**1141**	**447.84434**	**0.39250**		
Corrected Total	**1152**	**730.73547**			

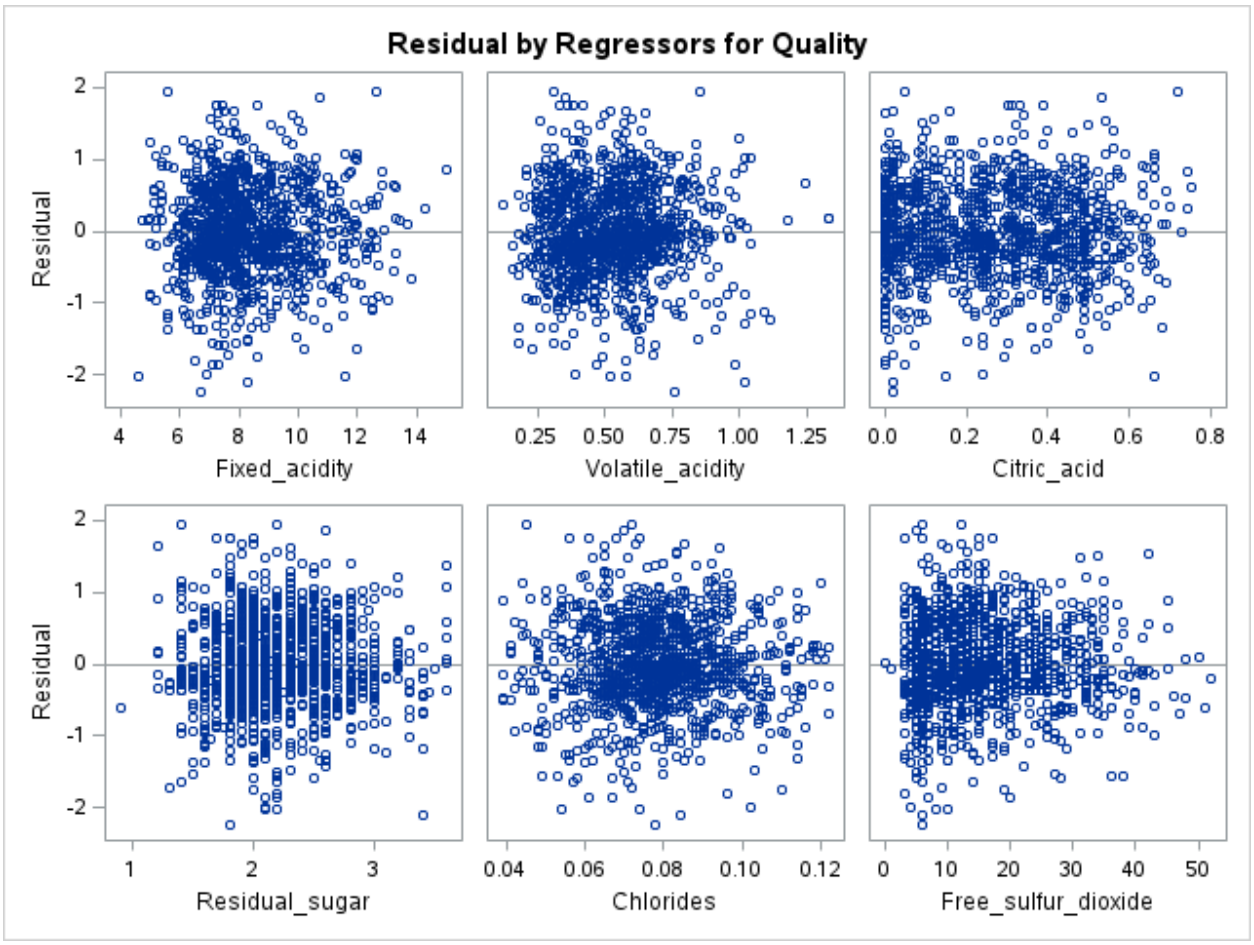
Root MSE	**0.62650**	**R-Square**	**0.3871**
Dependent Mean	**5.63660**	**Adj R-Sq**	**0.3812**
Coeff Var	**11.11485**		

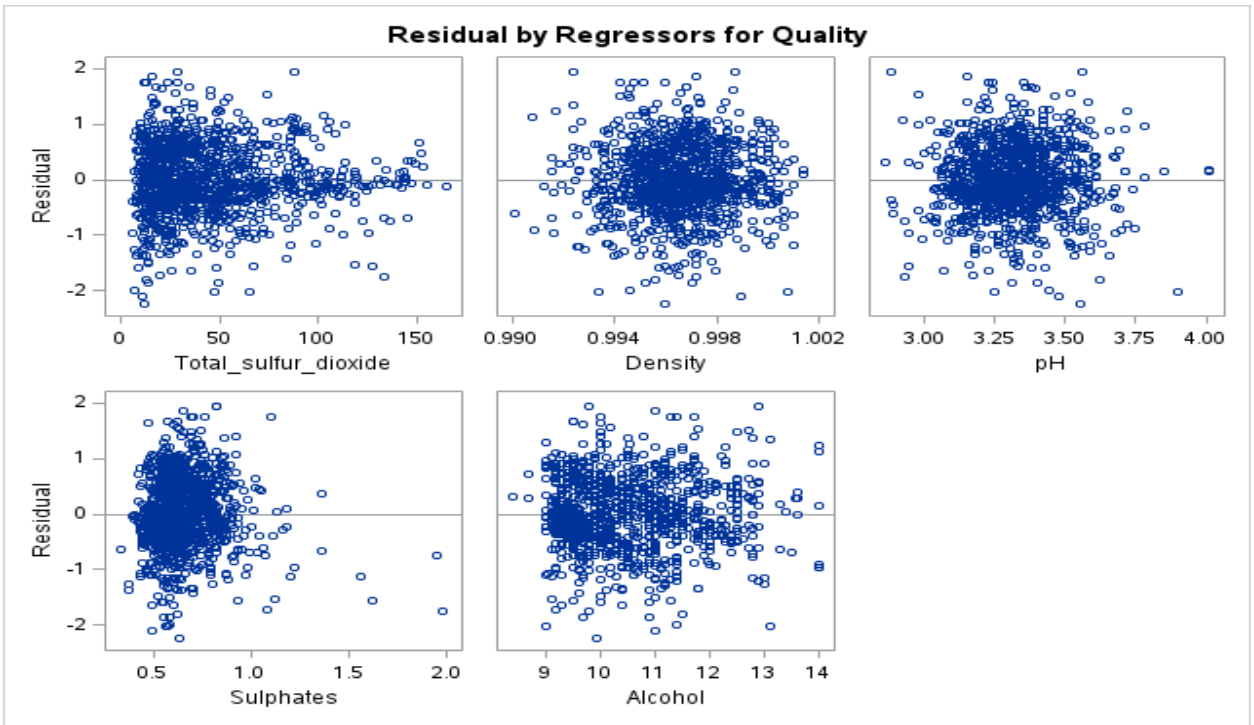
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**Intercept**	**1**	**50.54125**	**28.96849**	**1.74**	**0.0813**
Fixed_acidity	**Fixed_acidity**	**1**	**0.06117**	**0.03131**	**1.95**	**0.0510**

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Volatile_acidity	Volatile_acidity	1	-1.00697	0.13997	-7.19	<.0001
Citric_acid	Citric_acid	1	-0.22511	0.17207	-1.31	0.1911
Residual_sugar	Residual_sugar	1	0.02426	0.05357	0.45	0.6507
Chlorides	Chlorides	1	-1.01799	1.42512	-0.71	0.4752
Free_sulfur_dioxide	Free_sulfur_dioxide	1	0.00425	0.00262	1.62	0.1055
Total_sulfur_dioxide	Total_sulfur_dioxide	1	-0.00306	0.00087376	-3.51	0.0005
Density	Density	1	-47.66049	29.52829	-1.61	0.1068
pH	pH	1	-0.20764	0.22600	-0.92	0.3584
Sulphates	Sulphates	1	1.10497	0.13946	7.92	<.0001
Alcohol	Alcohol	1	0.26379	0.03461	7.62	<.0001




Residual by Regressors for Quality
Residual
Fixed_acidity
Volatile_acidity
Citric_acid
Residual_sugar
Chlorides
Free_sulfur_dioxide


Residual by Regressors for Quality
Residual
Total_sulfur_dioxide
Density
pH
Sulphates
Alcohol

Key Observations:

- **Model Significance:** The model is statistically significant (p-value < 0.0001), indicating that at least one predictor variable is significantly related to quality.
- **R-squared:** The R-squared value is moderately high, suggesting that the model explains a significant portion of the variation in quality.
- **Predictor Variables:**
    - **Fixed Acidity:** Has a significant positive effect on quality (p-value < 0.0001).
    - **Volatile Acidity:** Has a significant negative effect on quality (p-value < 0.0001).
    - **Citric Acid:** Has a significant positive effect on quality (p-value < 0.0001).
    - **Residual Sugar:** Has a significant negative effect on quality (p-value < 0.0001).
    - **Chlorides:** Has a significant negative effect on quality (p-value < 0.0001).
    - **Free Sulfur Dioxide:** Has a significant positive effect on quality (p-value < 0.0001).
    - **Total Sulfur Dioxide:** Has a significant negative effect on quality (p-value < 0.0001).
    - **Density:** Has a significant negative effect on quality (p-value < 0.0001).
    - **pH:** Has a significant negative effect on quality (p-value < 0.0001).
    - **Sulphates**: Has a significant positive effect on quality (p-value < 0.0001).
    - **Alcohol:** Has a significant positive effect on quality (p-value < 0.0001).
- **Residual Plots:** The residual plots appear to show no significant patterns, suggesting that the model's assumptions are met.

Interpretation:

- **Factors Influencing Quality**: The regression model suggests that several factors influence wine quality, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.
- **Direction of Effects:** The signs of the coefficients indicate the direction of the effects of these variables on quality. For example, higher fixed acidity and alcohol content are associated with higher quality, while higher volatile acidity and residual sugar are associated with lower quality.

Conclusion on the Exploratory Analysis on Wine Quality (Red)

The removal of outliers based on the IQR method led to a reduction of 200 observations from the dataset.

Impact on Analysis: The removal of outliers might have improved the accuracy and reliability of the analysis by reducing the influence of extreme values.

- **Quality Distribution:** The quality ratings are concentrated in the mid-range (5-6), with fewer wines having very high or very low ratings.

Recommendations Based on the Exploratory Analysis

Based on the EDA, variables like alcohol content, fixed acidity, and volatile acidity appear to be potential predictors of wine quality.

1. **Consider Non-Linear Relationships:**
    - The right-skewed distributions of some variables suggest that non-linear relationships might exist. Explore transformations (e.g., logarithmic, square root) to capture these relationships.
2. **Feature Engineering:**
    - Create new features based on existing variables to capture more complex relationships or interactions. For example, I could create a ratio of fixed acidity to volatile acidity to assess the balance of acids in the wine.