Visualizing the Wine Quality Red Dataset

This document outlines the data visualization techniques used to explore the Wine Quality dataset, and the insights derived from these visualizations. The primary goal is to provide a comprehensive understanding of the key factors influencing wine quality**.**

Data Exploration and Cleaning

- **Data Loading:** The Wine Quality dataset was loaded into SAS for analysis.
- **Data Cleaning**: Necessary data cleaning steps were performed, including handling missing values and outliers.

Before diving into visualizations, let's see the key variables in the Wine Quality dataset:

- **Fixed Acidity:** The fixed acidity in the wine
- **Volatile Acidity:** The volatile acidity in the wine
- **Citric Acid:** The citric acid in the wine
- **Residual Sugar:** The residual sugar in the wine
- **Chlorides:** The chlorides in the wine
- **Free Sulfur Dioxide:** The free sulfur dioxide in the wine
- **Total Sulfur Dioxide:** The total sulfur dioxide in the wine
- **Density:** The density of the wine
- **pH:** The pH of the wine
- **Sulphates:** The sulphates in the wine
- **Alcohol:** The alcohol content in the wine
- **Quality:** The quality rating of the wine (0-10)

Visualization Techniques

1. Histogram of Quality Ratings

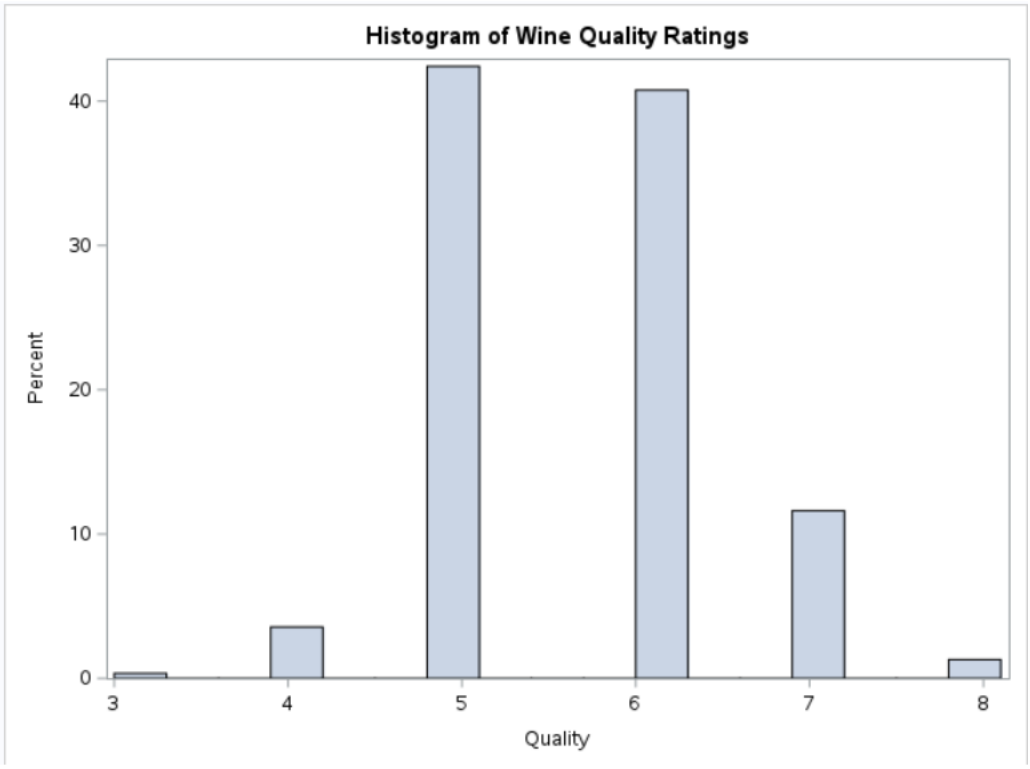
A histogram was created to visualize the distribution of quality ratings.

Creating the Histogram:

Here's the SAS code to create a histogram of the Quality variable:

```
175 /* Data Visualization*/
176
177 proc sgplot data=new_rdata;
178   histogram Quality;
179   title "Histogram of Wine Quality Ratings";
180 run;
181
```

Histogram Visualization of Wine Quality Ratings



Analyzing the Histogram of Wine Quality Ratings

Key Observations:

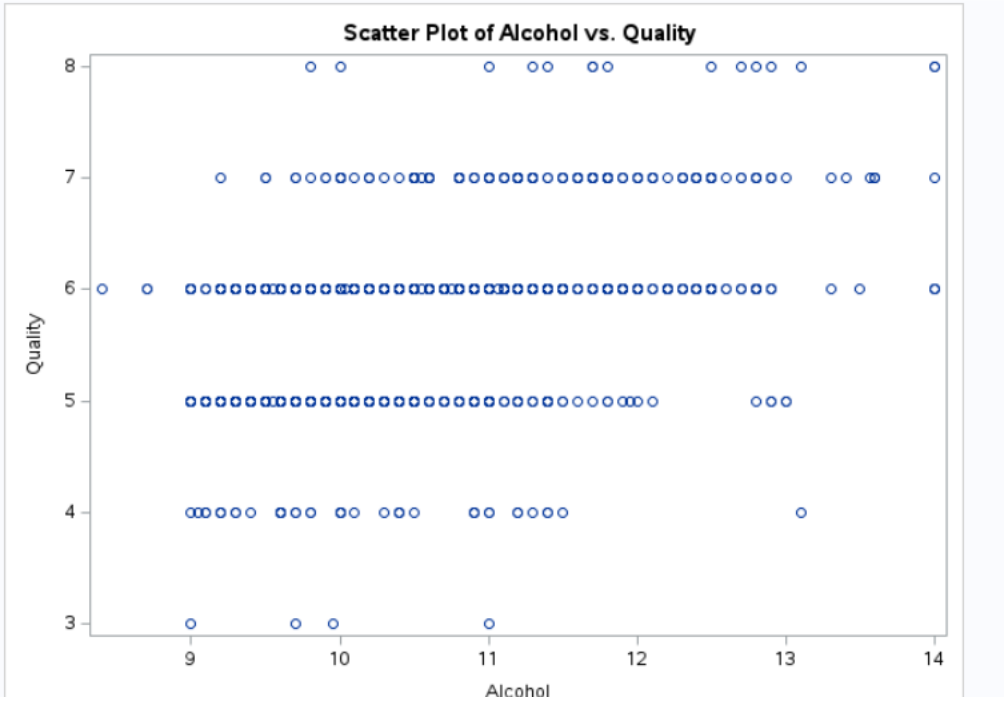
- **Distribution:** The histogram maintains its multimodal shape, with distinct peaks at quality ratings of 5, 6, and 7.
- **Central Tendency:** The median quality rating is likely still around 6, indicating that most wines fall within the average or slightly above-average range.
- **Range:** The overall range of quality ratings might have changed slightly due to outlier removal.
- **Skewness:** The distribution might have become less skewed or more symmetric depending on the specific outliers removed.

2. Scatter Plots Visualizations:

2a. Alcohol vs. Quality: Explore the relationship between alcohol content and quality.

2b. Scatter plot of Alcohol, Quality, and Fixed Acidity: Analyze the relationship between alcohol, fixed acidity and quality.

2a. Analyzing the Scatter Plot of Alcohol vs. Quality



Key Observations:

- **Clustering:** The data points tend to cluster around specific quality levels, particularly 5, 6, and 7.
- **Limited Overlap:** There is limited overlap between the clusters for different quality levels, suggesting that alcohol content might be a differentiating factor.
- **Outliers:** A few outliers are visible, especially at lower quality levels and higher alcohol values.
- **Overall Trend:** While there is a general trend of higher quality wines having slightly higher alcohol levels, the relationship is not perfectly linear.

Interpretation:

- **Quality Bands:** The clustering around specific quality levels indicates that alcohol content might be one of the factors influencing wine quality, but it's not the sole determinant.

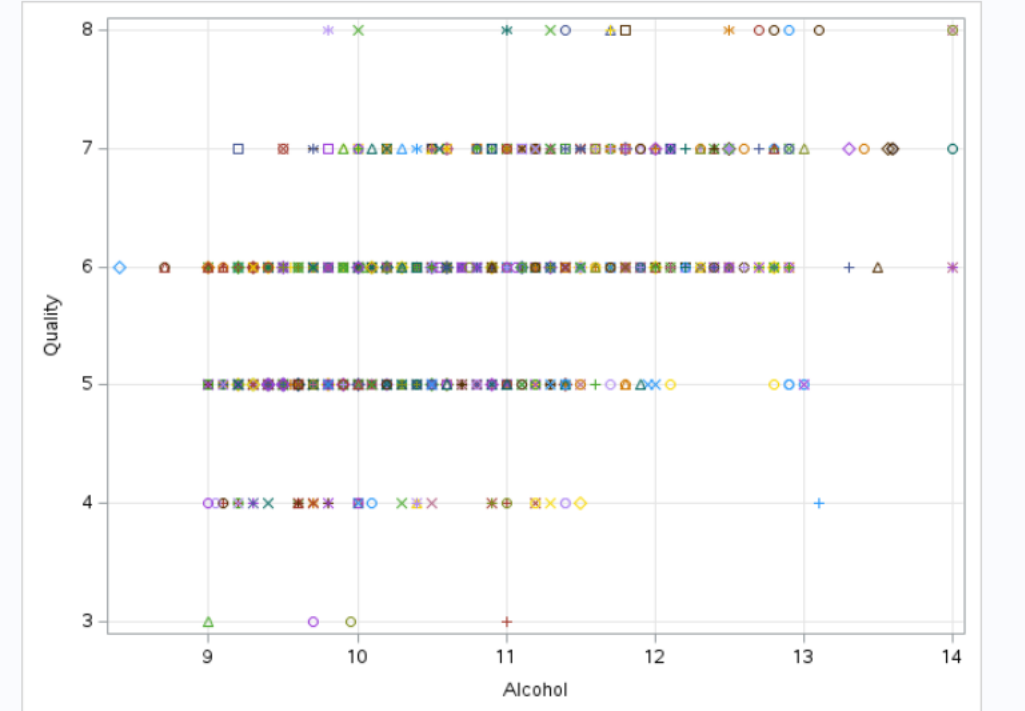
- **Outliers:** The outliers suggest that there might be some wines with unique characteristics or production processes that deviate from the general trends.
- **Non-Linear Relationship:** The scatter plot hints at a potential non-linear relationship between alcohol and quality. A more complex model might be needed to capture this relationship accurately.

2b. Scatter plot of Alcohol, Quality, and Fixed Acidity: Analyze the relationship between fixed acidity, alcohol and quality.

Example of the code below:

```
223 ods graphics / reset width=6.4in height=4.8in imagemap;
224
225 proc sgplot data=new_rdata;
226     scatter x=Alcohol y=Quality / group=Fixed_acidity;
227     xaxis grid;
228     yaxis grid;
229 run;
230
231 ods graphics / reset;
```

Scatter plot of Alcohol, Quality, and Fixed Acidity



Key Observations:

- **Clustering:** The data points tend to cluster around specific quality levels, particularly 5, 6, and 7, regardless of fixed acidity.
- **Limited Overlap:** There is limited overlap between the clusters for different quality levels, suggesting that fixed acidity might not be a strong differentiating factor for overall quality.
- **Outliers:** A few outliers are visible, especially at lower quality levels and higher alcohol values.
- **No Clear Trend:** There doesn't seem to be a strong linear relationship between alcohol and quality when considering fixed acidity as a grouping factor.

Interpretation:

- **Quality is Influenced by Multiple Factors:** The scatter plot suggests that wine quality is influenced by multiple factors beyond just alcohol and fixed acidity. Other variables might play a more significant role.
- **Fixed Acidity's Limited Effect:** While fixed acidity might contribute to quality, it doesn't seem to be a primary driver of the observed quality differences.
- **Outliers:** The outliers might represent wines with unique characteristics or production processes that deviate from the general trends.

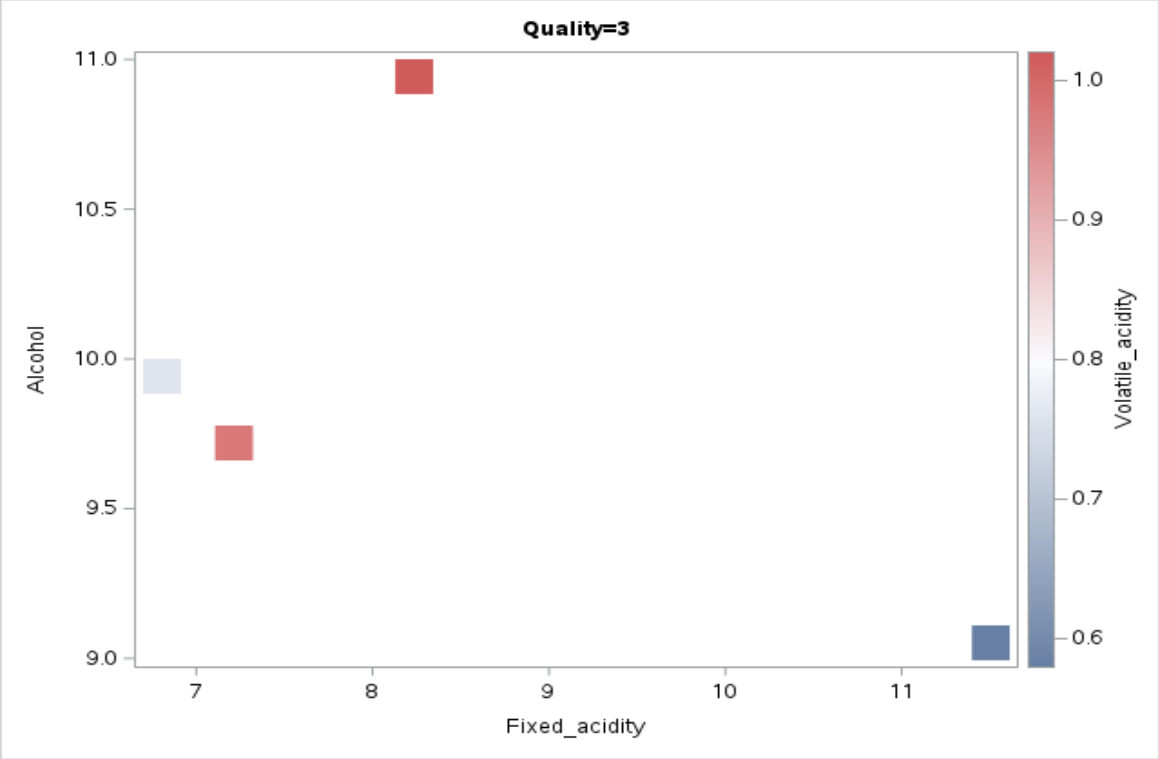
3. Heatmap code is shown below to show interactions effects:

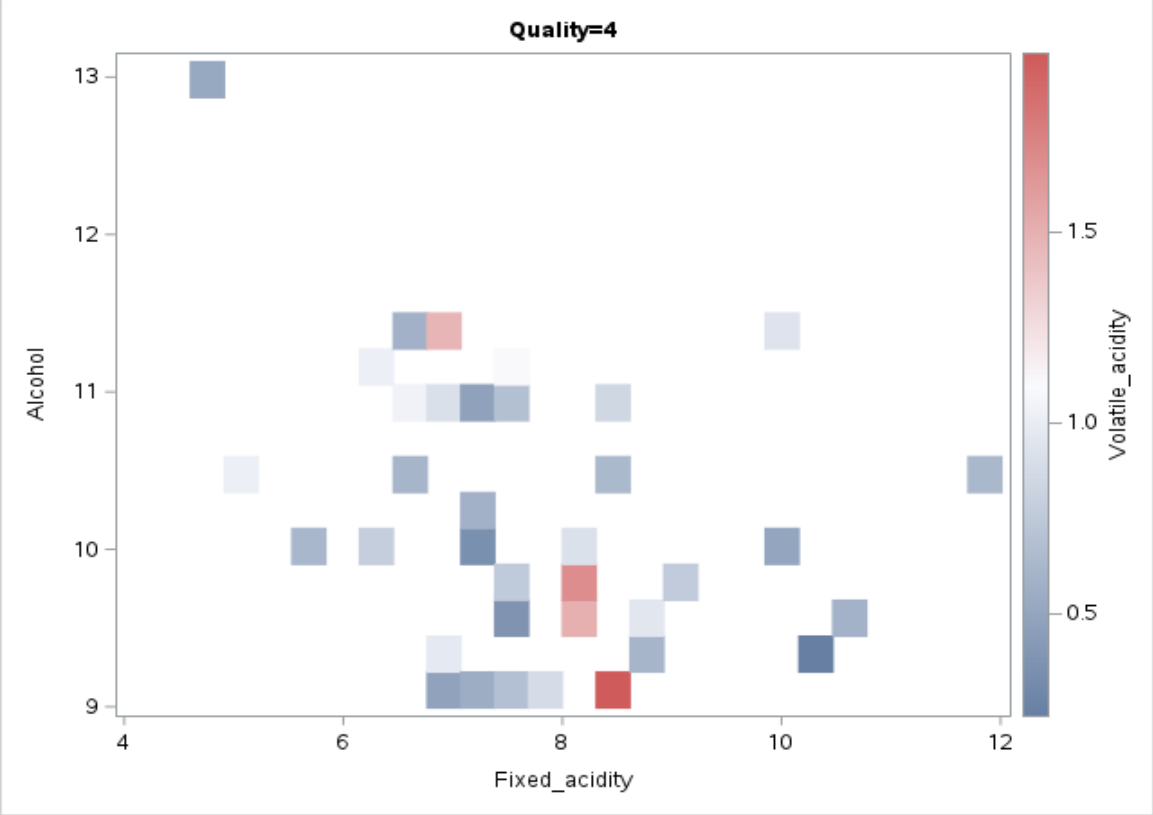
A correlation heatmap was generated to visualize the relationships between all variables.

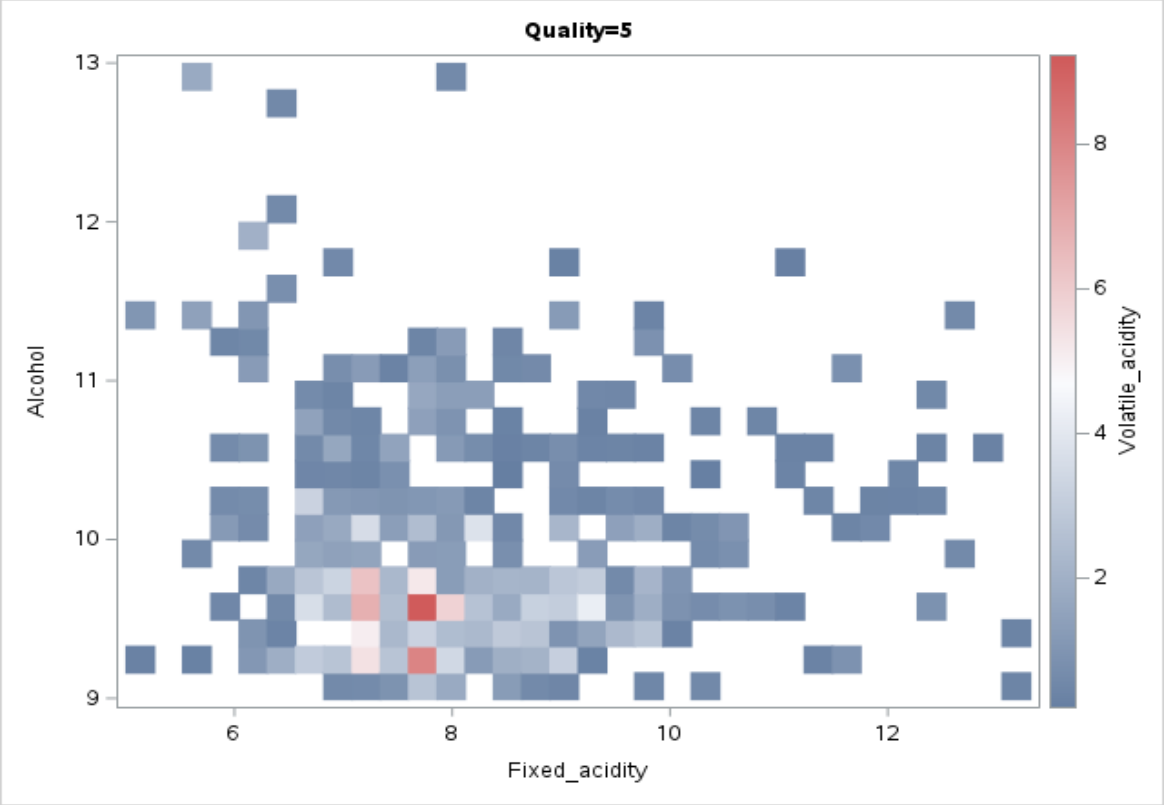
See the code generated in SAS below:

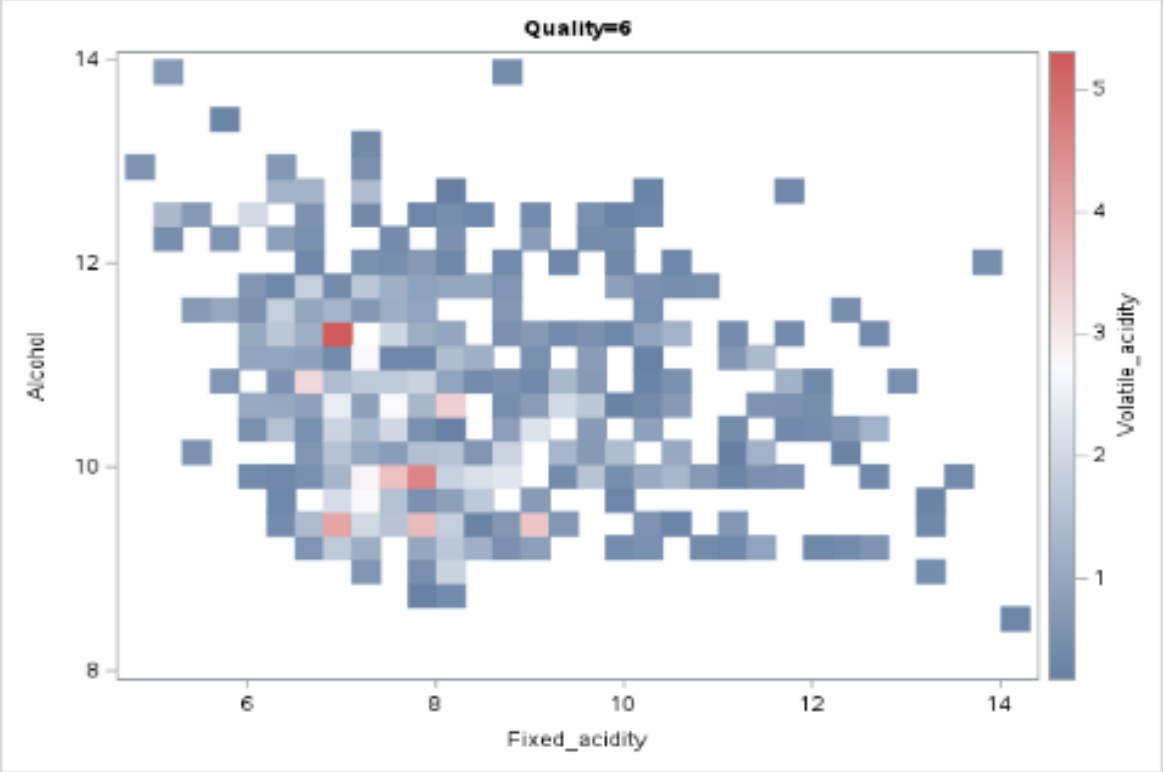
```
/*heatmap*/
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sort data=new_rdata out=_HeatMapTaskData;
    by Quality;
run;

proc sgplot data=_HeatMapTaskData;
    by Quality;
    heatmap x=Fixed_acidity y=Alcohol / name='HeatMap'
        colorresponse=Volatile_acidity;
    gradlegend 'HeatMap';
run;

ods graphics / reset;

proc datasets library=WORK noprint;
    delete _HeatMapTaskData;
    run;
```

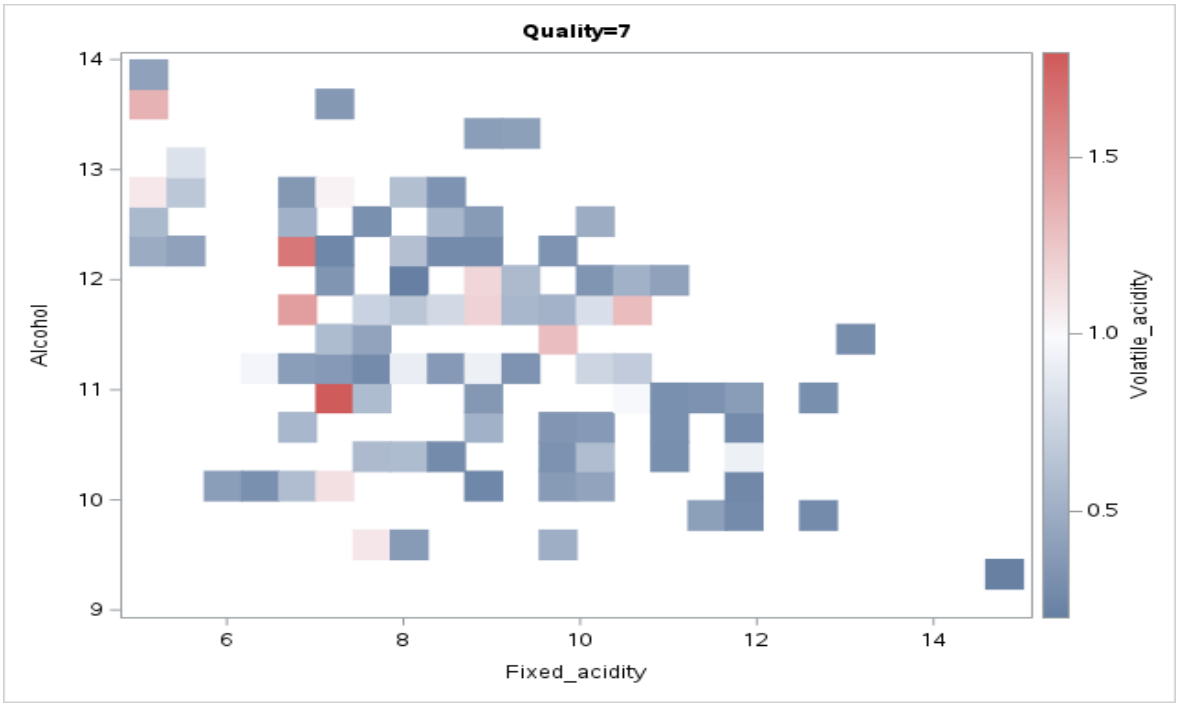
Result of Heatmap Visualization

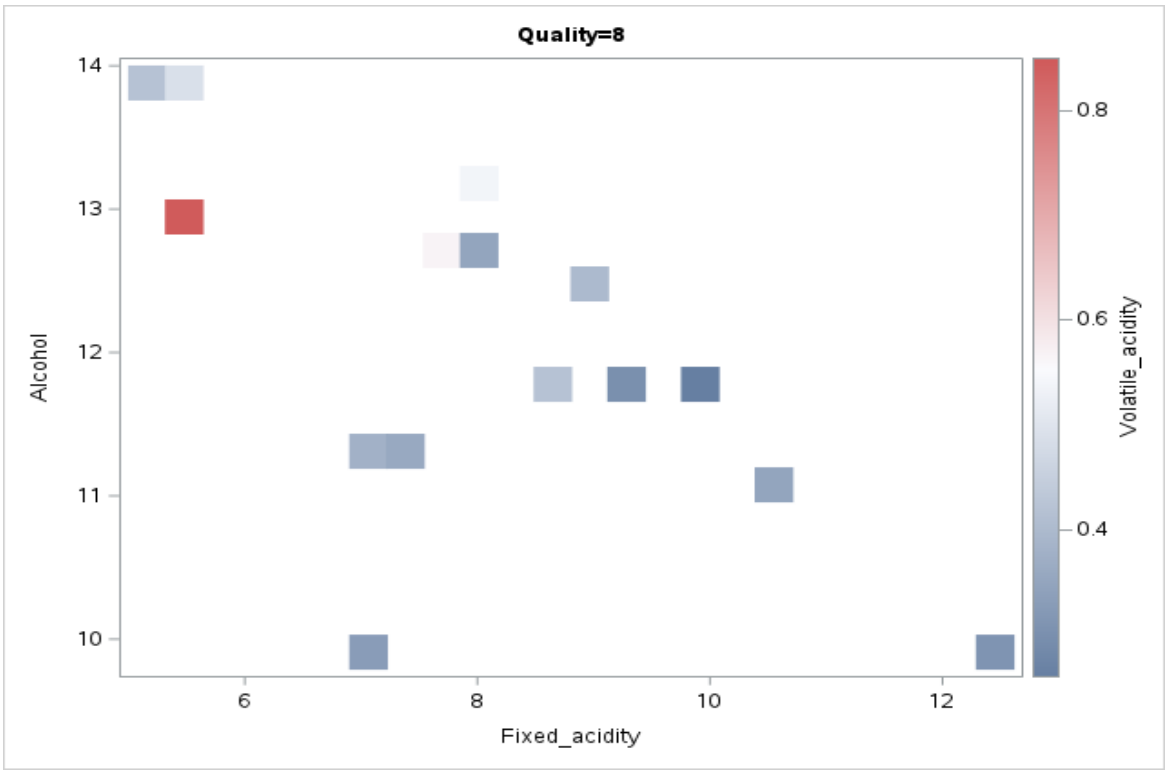





Quality=5
Alcohol
Fixed_acidity
Volatile_acidity
13
12
11
10
9
6
8
10
12
8
6
4
2


Quality=6
Alcohol
Fixed_acidity
Volatile_acidity
14
12
10
8
6
8
10
12
14
5
4
3
2
1


Quality=7
Alcohol
Fixed_acidity
Volatile_acidity
14
13
12
11
10
9
6
8
10
12
14
1.5
1.0
0.5


Quality=8
Alcohol
Fixed_acidity
Volatile_acidity
14
13
12
11
10
6
8
10
12
0.8
0.6
0.4

Analyzing the Correlation Plots

Key Observations:

- **Alcohol vs. Quality:** The scatter plot shows a general positive correlation between alcohol content and quality, with higher alcohol levels associated with higher quality ratings. However, there are some outliers, and the relationship is not perfectly linear.
- **Fixed Acidity vs. Quality:** The scatter plot reveals a weak negative correlation between fixed acidity and quality. This suggests that wines with slightly lower fixed acidity might have a slight advantage in terms of quality.
- **Other Variable Relationships:** The other scatter plots show various relationships between the variables. Some variables might have stronger or weaker correlations with quality than others.

Interpretation:

- **Alcohol as a Key Factor:** Alcohol content appears to be a significant factor influencing wine quality, with higher alcohol levels generally associated with higher ratings.
- **Fixed Acidity's Limited Effect:** Fixed acidity has a limited negative impact on quality, suggesting that it's not a primary driver of quality.
- **Other Variables:** The relationships between other variables and quality might be more complex and require further exploration.

4. **Multivariate Analysis:** we are going to use multivariate analysis techniques; specifically; **principal component analysis (PCA)** to identify underlying patterns and relationships among variables.

Principal Component Analysis) PCA is statistical techniques used to reduce the dimensionality of data by combining correlated variables into a smaller set of uncorrelated components. SAS provides several procedures to perform these analyses.

```
320
321 /*principal component analysis */
322 ods noproctitle;
323 ods graphics / imagemap=on;
324
325 proc princomp data=new_rdata plots(only)=(scree);
326     var Fixed_acidity Volatile_acidity Citric_acid Residual_sugar Chlorides
327         Free_sulfur_dioxide Total_sulfur_dioxide Density pH Sulphates Alcohol;
328     partial Quality;
329 run;
```

Result of the Principal Component Analysis

Observations	1153
Variables	11
Partial Variable	1

Simple Statistics

	Fixed_acidity	Volatile_acidity	Citric_acid	Residual_sugar	Chlorides	Free_sulfur_dioxide	Total_sulfur_dioxide	Density	pH	Sulphates	Alcohol	Quality
Mean	8.254206418	0.5251561145	0.2592974848	2.195229835	0.0782437121	15.66999133	45.63052905	0.9965666522	3.319132697	0.6454640069	10.42477595	5.636600173
StD	1.671986847	0.1796149235	0.1874192673	0.451750803	0.0151509911	9.64993531	31.17369302	0.0017773755	0.151334723	0.1499044162	1.05316937	0.796441452

Correlation Matrix

		Fixed_acidity	Volatile_acidity	Citric_acid	Residual_sugar	Chlorides	Free_sulfur_dioxide	Total_sulfur_dioxide	Density	pH	Sulphates	Alcohol	Quality
Fixed_acidity	Fixed_acidity	1.0000	-.2660	0.6813	0.2650	0.2174	-.1341	-.0852	0.6749	-.7117	0.1779	-.1336	0.1022
Volatile_acidity	Volatile_acidity	-.2660	1.0000	-.5794	0.0127	0.1114	-.0162	0.1000	0.0181	0.2544	-.3088	-.1894	-.3876
Citric_acid	Citric_acid	0.6813	-.5794	1.0000	0.1744	0.0990	-.0416	0.0423	0.3565	-.5274	0.2577	0.0696	0.2201
Residual_sugar	Residual_sugar	0.2650	0.0127	0.1744	1.0000	0.2674	0.0561	0.1357	0.4246	-.0917	0.0351	0.0391	-.0035
Chlorides	Chlorides	0.2174	0.1114	0.0990	0.2674	1.0000	0.0331	0.1750	0.4413	-.2205	-.0216	-.3373	-.1980
Free_sulfur_dioxide	Free_sulfur_dioxide	-.1341	-.0162	-.0416	0.0561	0.0331	1.0000	0.6490	-.0433	0.0822	0.0850	-.0271	-.0279
Total_sulfur_dioxide	Total_sulfur_dioxide	-.0852	0.1000	0.0423	0.1357	0.1750	0.6490	1.0000	0.0913	-.0497	0.0203	-.2194	-.2039
Density	Density	0.6749	0.0181	0.3565	0.4246	0.4413	-.0433	0.0913	1.0000	-.3459	0.1076	-.5820	-.2256
pH	pH	-.7117	0.2544	-.5274	-.0917	-.2205	0.0822	-.0497	-.3459	1.0000	-.0984	0.2299	-.0509
Sulphates	Sulphates	0.1779	-.3088	0.2577	0.0351	-.0216	0.0850	0.0203	0.1076	-.0984	1.0000	0.1866	0.3477
Alcohol	Alcohol	-.1336	-.1894	0.0696	0.0391	-.3373	-.0271	-.2194	-.5820	0.2299	0.1866	1.0000	0.4946
Quality	Quality	0.1022	-.3876	0.2201	-.0035	-.1980	-.0279	-.2039	-.2256	-.0509	0.3477	0.4946	1.0000

Regression Statistics

	Fixed_acidity	Volatile_acidity	Citric_acid	Residual_sugar	Chlorides	Free_sulfur_dioxide	Total_sulfur_dioxide	Density	pH	Sulphates	Alcohol
R-Square	0.0104358742	0.1502542305	0.0484641038	0.0000122165	0.0391891365	0.0007799992	0.0415878273	0.0508751904	0.0025951745	0.1208835108	0.244653711
RMSE	1.6639620063	0.1656439205	0.1829007143	0.4519442428	0.014857597	9.6503605383	30.531841043	0.0017323253	0.1512038668	0.1406132409	0.9157139821

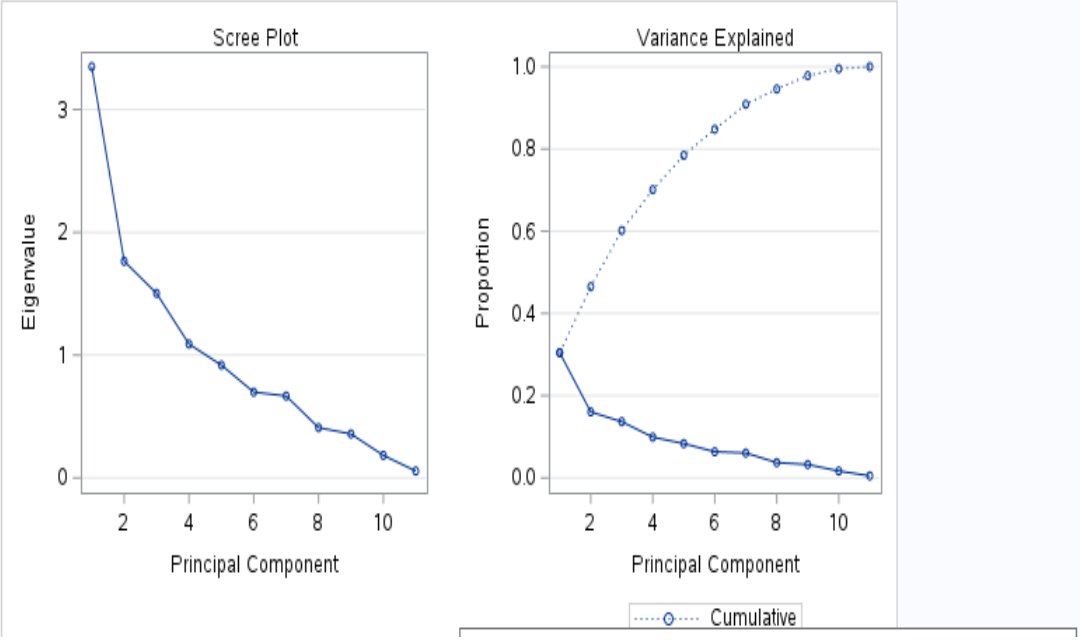
Standardized Regression Coefficients

	Fixed_acidity	Volatile_acidity	Citric_acid	Residual_sugar	Chlorides	Free_sulfur_dioxide	Total_sulfur_dioxide	Density	pH	Sulphates	Alcohol
Quality	0.1021561266	-.3876264058	0.2201456423	-.0034952070	-.1979624624	-.0279284655	-.2039309377	-.2255552934	-.0509428556	0.3476830609	0.4946248184

Partial Correlation Matrix												
		Fixed_acidity	Volatile_acidity	Citric_acid	Residual_sugar	Chlorides	Free_sulfur_dioxide	Total_sulfur_dioxide	Density	pH	Sulphates	Alcohol
Fixed_acidity	Fixed_acidity	1.0000	-.2469	0.6790	0.2668	0.2437	-.1320	-.0661	0.7201	-.7111	0.1526	-.2130
Volatile_acidity	Volatile_acidity	-.2469	1.0000	-.5494	0.0124	0.0383	-.0284	0.0232	-.0772	0.2549	-.2014	0.0029
Citric_acid	Citric_acid	0.6790	-.5494	1.0000	0.1796	0.1492	-.0364	0.0913	0.4274	-.5298	0.1981	-.0463
Residual_sugar	Residual_sugar	0.2668	0.0124	0.1796	1.0000	0.2721	0.0560	0.1379	0.4350	-.0920	0.0387	0.0470
Chlorides	Chlorides	0.2437	0.0383	0.1492	0.2721	1.0000	0.0282	0.1403	0.4153	-.2355	0.0514	-.2809
Free_sulfur_dioxide	Free_sulfur_dioxide	-.1320	-.0284	-.0364	0.0560	0.0282	1.0000	0.6574	-.0509	0.0809	0.1011	-.0153
Total_sulfur_dioxide	Total_sulfur_dioxide	-.0661	0.0232	0.0913	0.1379	0.1403	0.6574	1.0000	0.0475	-.0615	0.0994	-.1393
Density	Density	0.7201	-.0772	0.4274	0.4350	0.4153	-.0509	0.0475	1.0000	-.3674	0.2036	-.5556
pH	pH	-.7111	0.2549	-.5298	-.0920	-.2355	0.0809	-.0615	-.3674	1.0000	-.0862	0.2939
Sulphates	Sulphates	0.1526	-.2014	0.1981	0.0387	0.0514	0.1011	0.0994	0.2036	-.0862	1.0000	0.0180
Alcohol	Alcohol	-.2130	0.0029	-.0463	0.0470	-.2809	-.0153	-.1393	-.5556	0.2939	0.0180	1.0000

Eigenvalues of the Partial Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.34930549	1.58434673	0.3045	0.3045
2	1.76495877	0.26143615	0.1605	0.4649
3	1.50352262	0.41269606	0.1367	0.6016
4	1.09082656	0.17214379	0.0992	0.7008
5	0.91868277	0.22009812	0.0835	0.7843
6	0.69858265	0.03143927	0.0635	0.8478
7	0.66714338	0.25799531	0.0606	0.9085
8	0.40914807	0.05092885	0.0372	0.9457
9	0.35821922	0.17470249	0.0326	0.9782
10	0.18351673	0.12741896	0.0167	0.9949
11	0.05609777		0.0051	1.0000

Eigenvectors												
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
Fixed_acidity	Fixed_acidity	0.481356	-.158545	0.007467	0.036924	-.117353	0.313440	0.114575	0.273312	-.188770	-.357099	0.615422
Volatile_acidity	Volatile_acidity	-.212309	0.146358	-.542366	0.098352	0.012192	0.542258	0.397527	0.208354	0.170168	0.324135	0.004327
Citric_acid	Citric_acid	0.412983	-.116451	0.358481	0.048010	-.174553	-.065459	-.009398	0.352586	0.420718	0.589029	-.044196
Residual_sugar	Residual_sugar	0.213632	0.178424	-.173237	0.725768	-.152264	0.027896	-.306178	-.440405	-.030933	0.181266	0.142467
Chlorides	Chlorides	0.251747	0.207119	-.348823	0.100409	0.066054	-.706206	0.486891	0.121888	-.058420	-.009412	0.058775
Free_sulfur_dioxide	Free_sulfur_dioxide	-.021049	0.628573	0.264324	-.070881	-.110267	0.082870	-.044023	0.239720	-.619971	0.257377	-.022877
Total_sulfur_dioxide	Total_sulfur_dioxide	0.061912	0.650233	0.176572	-.095964	-.162922	0.088843	0.074330	-.113896	0.571983	-.389646	0.032484
Density	Density	0.447016	0.068407	-.291524	0.059126	0.192399	0.139621	-.319165	0.314666	-.036719	-.268014	-.613549
pH	pH	-.402679	0.104128	-.066831	0.249577	0.252715	-.213630	-.437036	0.552347	0.190248	-.073869	0.336170
Sulphates	Sulphates	0.146739	0.093062	0.306195	0.143981	0.869074	0.148106	0.206024	-.156023	0.023063	0.076874	0.069753
Alcohol	Alcohol	-.236366	-.151851	0.375943	0.590837	-.182179	0.045761	0.393921	0.223327	-.066575	-.293685	-.315651



- **Scree Plot**: The scree plot shows a rapid decrease in eigenvalues with the first few principal components, indicating that a significant amount of variance can be explained by a small number of components.
- **Variance Explained**: The variance explained plot shows that the first two principal components capture a substantial portion of the total variance in the data, suggesting that they are effective in summarizing the important information.
- **Eigenvectors**: The eigenvectors (loadings) for each principal component indicate the contribution of each variable to that component. You can use these loadings to interpret the meaning of the principal components.

Interpreting the Principal Components:

- **PC1:** Examine the loadings for PC1 to understand the variables that contribute most to this component. For example, if variables related to acidity have high loadings on PC1, it might suggest that PC1 represents an acidity factor.
- **PC2:** Analyze the loadings for PC2 to identify the variables that contribute most to this component. This might reveal another underlying factor or pattern in the data.
- **Subsequent Components**: Examine the loadings for subsequent principal components to understand the additional factors captured by the analysis**.**

Analyzing the Scree Plot

Key Observations:

- **Steep Descent:** The eigenvalues decrease rapidly with the first few principal components.
- **Elbow Point:** There appears to be an elbow point around the 3rd or 4th principal component.
- **Diminishing Returns:** Beyond the elbow point, the eigenvalues decrease more gradually.

Interpretation:

- **Dominant Factors:** The steep descent in the scree plot suggests that a few principal components capture a significant portion of the total variance in the data.
- **Optimal Number of Components**: The elbow point indicates that retaining the first 3 or 4 principal components might be sufficient to capture most of the important information.
- **Diminishing Returns**: Beyond the elbow point, adding more principal components would provide diminishing returns in terms of explained variance.

Based on the scree plot, it seems that 3 or 4 principal components would be sufficient to capture most of the important variation in the Wine Quality dataset. This suggests that the complexity of the data can be effectively reduced to a smaller number of dimensions.

Conclusion

- **Quality is influenced by multiple factors:** Wine quality is not determined by a single variable but rather by a combination of factors.
- **Alcohol and fixed acidity:** While alcohol content and fixed acidity play a role, they are not the sole determinants of quality.
- **Other variables:** Other variables, such as residual sugar, volatile acidity, and pH, also contribute to wine quality.
- **Dimensionality reduction:** PCA can be used to reduce the dimensionality of the data and identify the most important factors influencing quality.