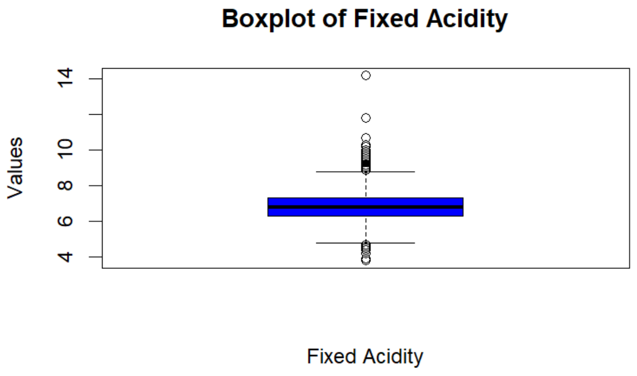
1 Understanding the dataset

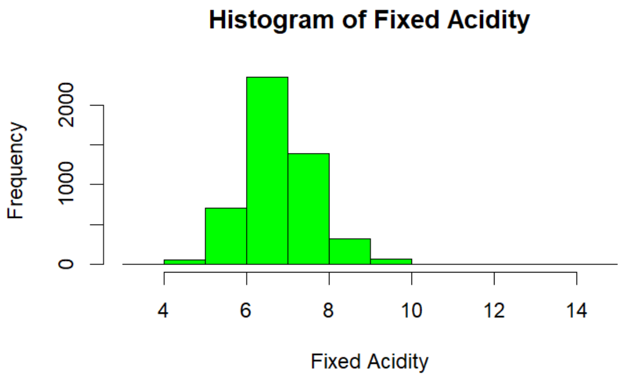
1.1 Statistical summary

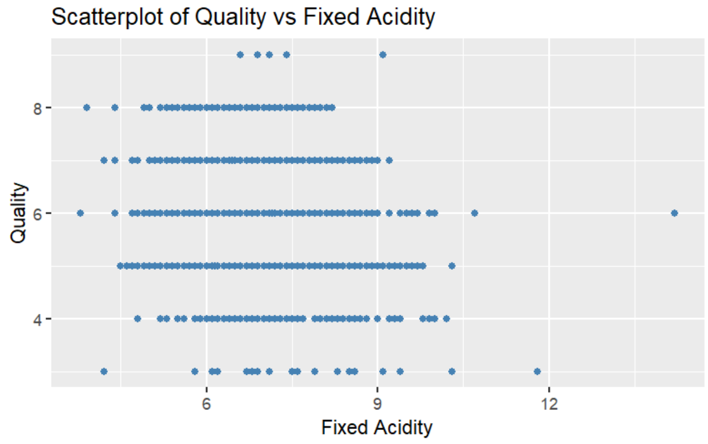
Variable	Type	#	Min	Q1	Median	Mean	Q3	Max	Outlier	Corr with Target
fixed acidity	num	4898	3.800	6.300	6.800	6.855	7.300	14.200	119	-0.114
volatile acidity	num	4898	0.080	0.210	0.260	0.2782	0.320	1.100	186	-0.195
citric acid	num	4898	0.000	0.270	0.320	0.3342	0.390	1.660	270	-0.009
residual sugar	num	4898	0.600	1.700	5.200	6.391	9.900	65.800	07	-0.097
chlorides	num	4898	0.009	0.036	0.043	0.0458	0.050	0.346	208	-0.210
free sulfur dioxide	num	4898	2.000	23.00	34.00	35.31	46.00	289.00	50	0.008
total sulfur dioxide	num	4898	9.000	108.0	134.0	138.4	167.0	440.0	19	-0.175
density	num	4898	0.987	0.991	0.994	0.9940	0.996	1.039	05	-0.307
pH	num	4898	2.720	3.090	3.180	3.188	3.280	3.820	75	0.09
sulphates	num	4898	0.220	0.410	0.470	0.4898	0.550	1.080	124	0.054
alcohol	num	4898	8.000	9.500	10.40	10.51	11.40	14.20	0	0.435

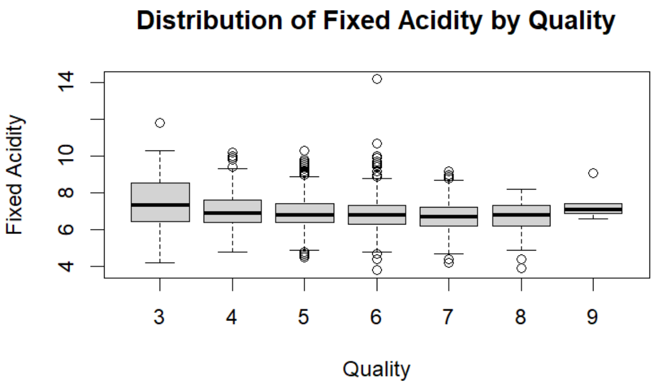
1.2 Fixed Acidity

The variable "Fixed Acidity" has a mean value of around 6.855, with values ranging from 3.800 to 14.200. There have been a total of 119 outliers detected, indicating the existence of extremely high or low values. The target variable and fixed acidity have a negative association (-0.114), meaning that an increase in fixed acidity is linked to a little decrease in wine quality.


Boxplot of Fixed Acidity

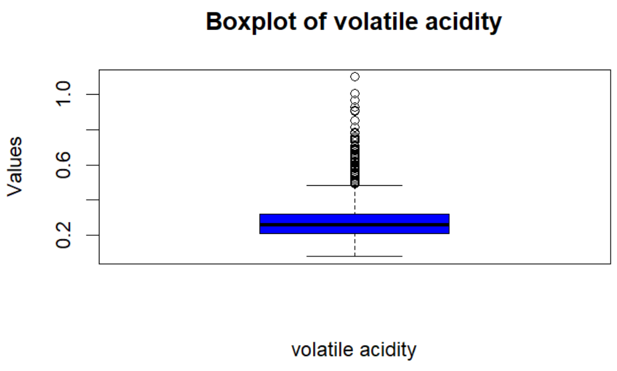

Histogram of Fixed Acidity

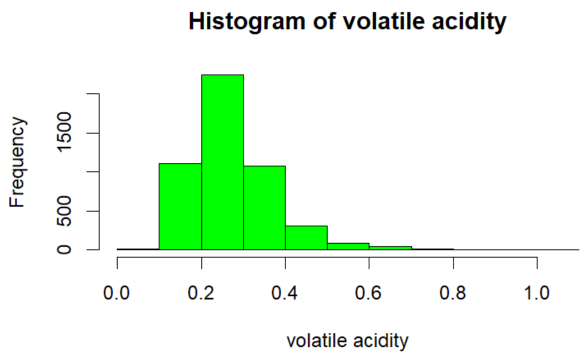

Scatterplot of Quality vs Fixed Acidity

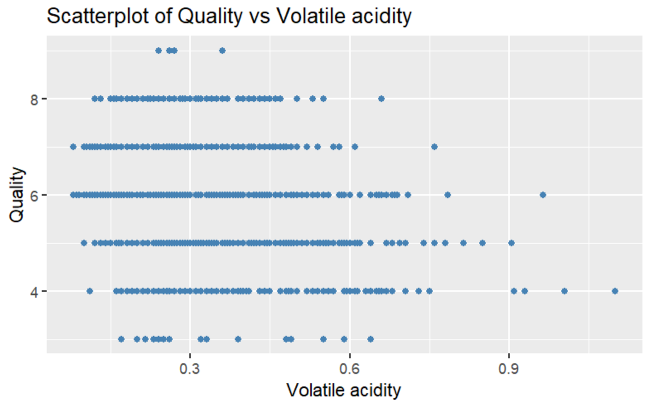

Distribution of Fixed Acidity by Quality

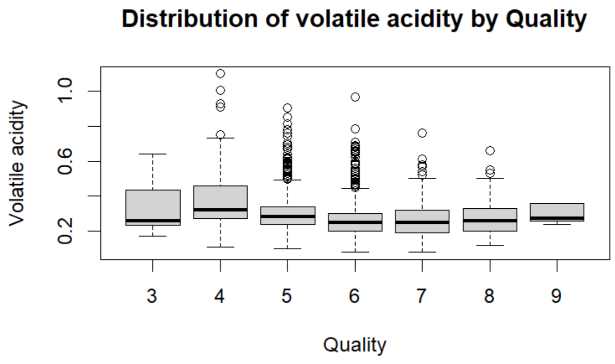
1.3 Volatile acidity

The mean volatile acidity is around 0.2782, with a range of values between 0.0800 and 1.1000. There have been a total of 186 outliers detected, which are extremely high or low values in the dataset. The negative correlation coefficient of -0.195 suggests a little decline in wine quality as volatile acidity increases.



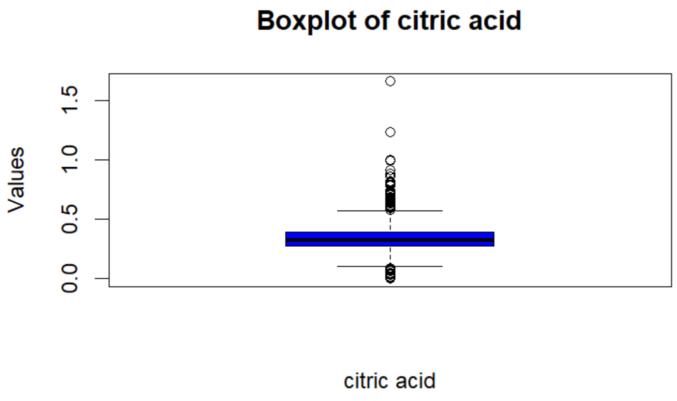


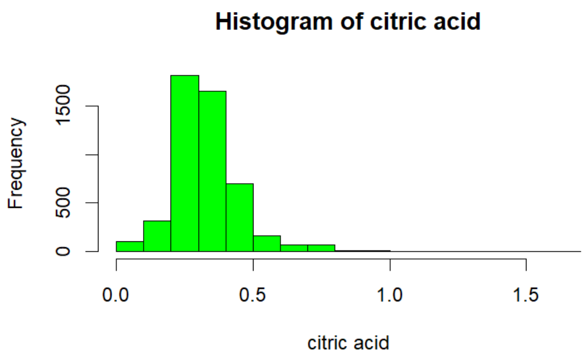


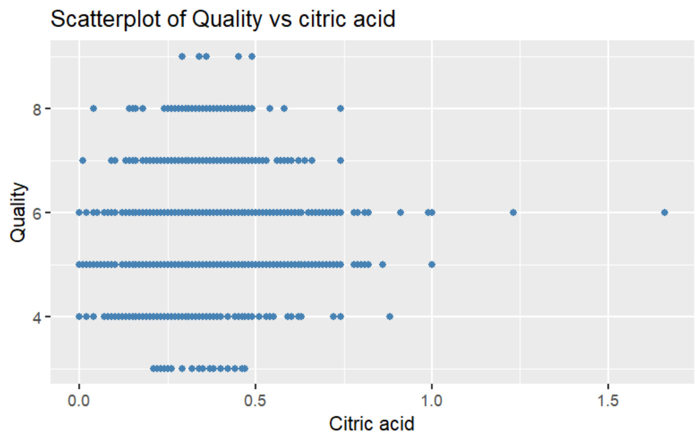


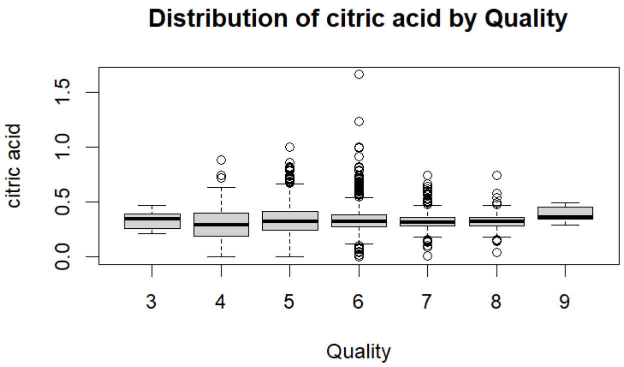
1.4 Citric acid

The mean value of citric acid is around 0.3342, with a range of values spanning from 0.0000 to 1.6600. There have been a total of 270 outliers detected, indicating the existence of extremely high or low values. The correlation coefficient between the target variable and this variable is close to zero (-0.009), suggesting a weak linear relationship.



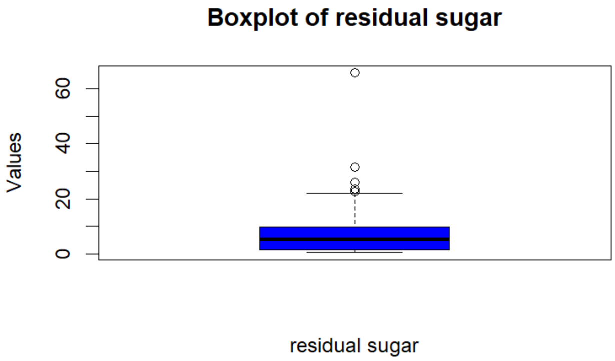


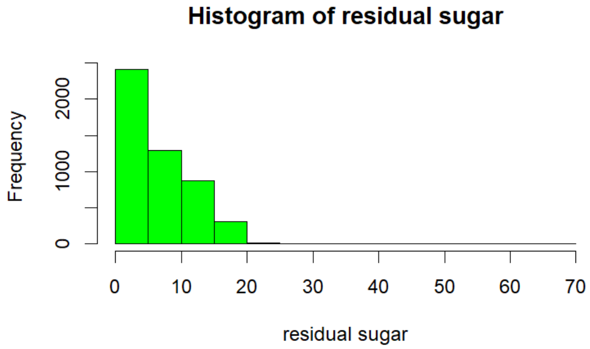


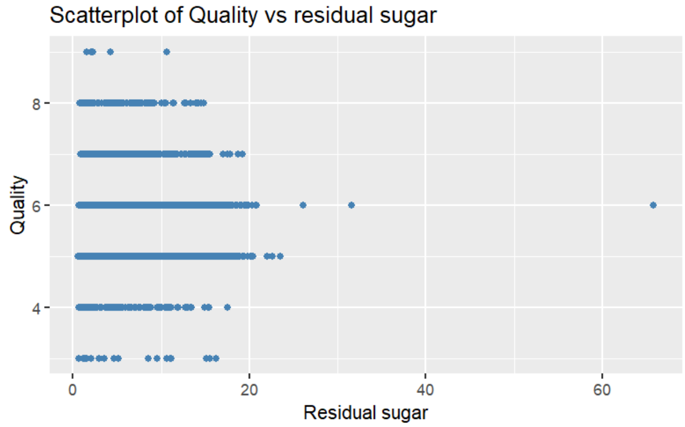


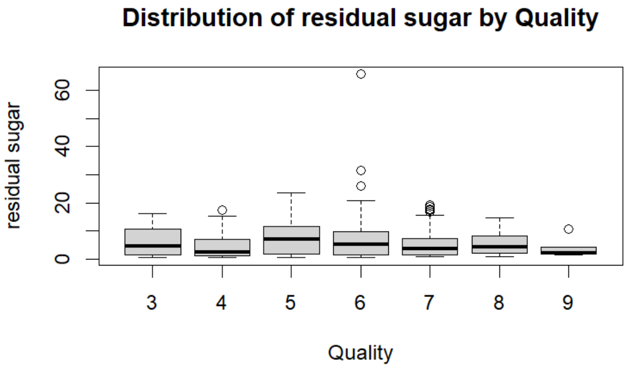
1.5 Residual sugar

The variable "Residual Sugar" has an approximate mean value of 6.391, with values ranging from 0.600 to 65.800. There have been a total of 7 outliers detected, which suggests the existence of values that are abnormally high or low. The correlation coefficient between the target variable and the variable in question is -0.097, indicating a modest negative link.



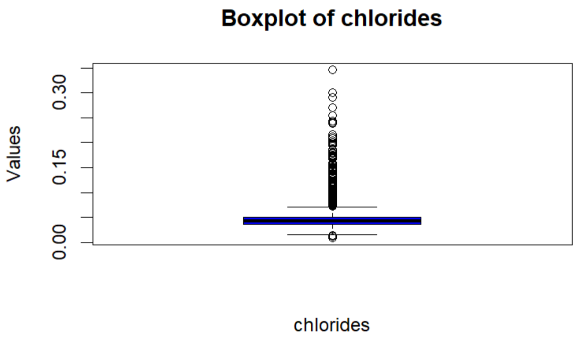


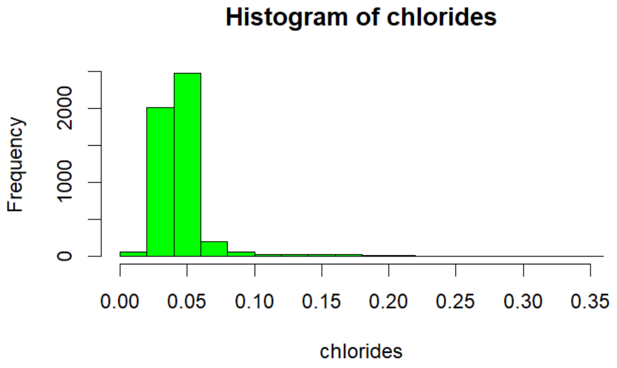


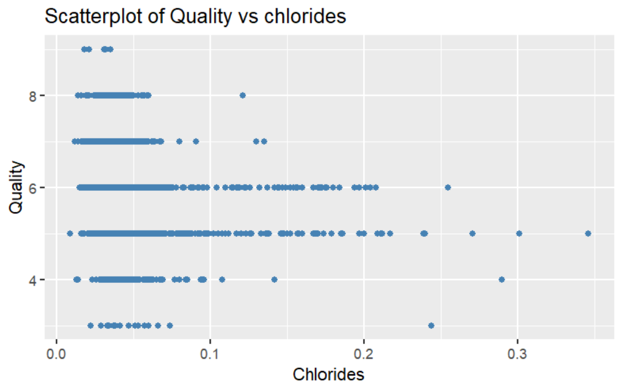


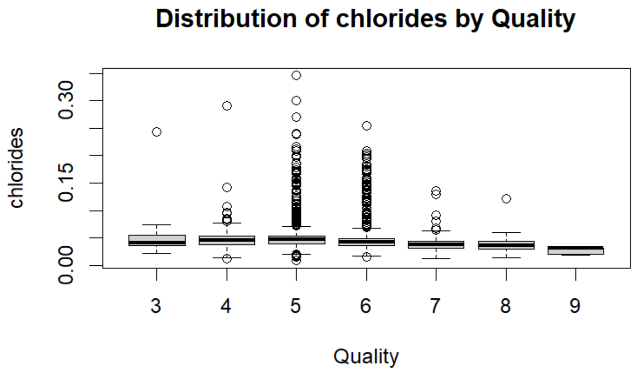
1.6 Chlorides

The mean concentration of chlorides is around 0.04577, with a range of values spanning from 0.00900 to 0.34600. There have been 208 outliers detected, indicating the existence of values that are either abnormally high or low. The negative correlation coefficient of -0.210 suggests that there is a substantial inverse association between the chloride concentration and the quality of the wine. This means that as the chloride level grows, the wine quality tends to deteriorate.



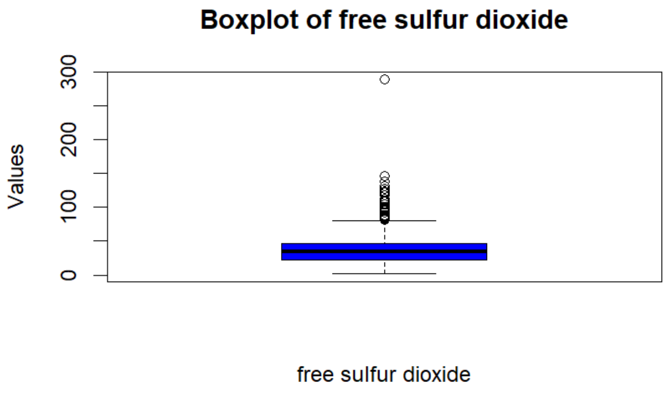


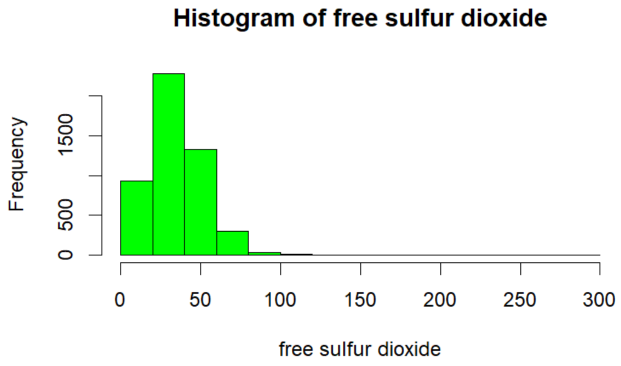


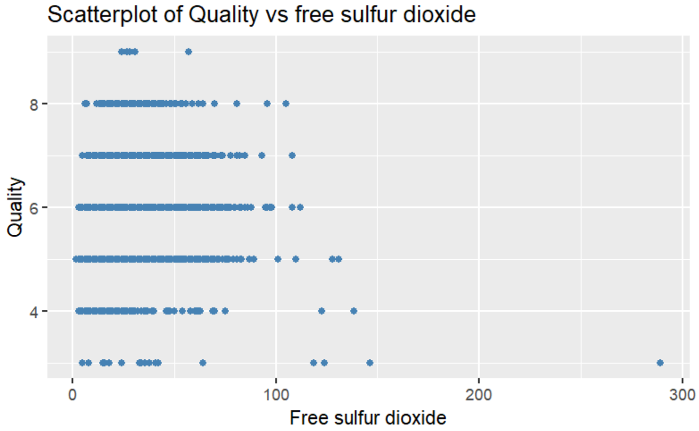


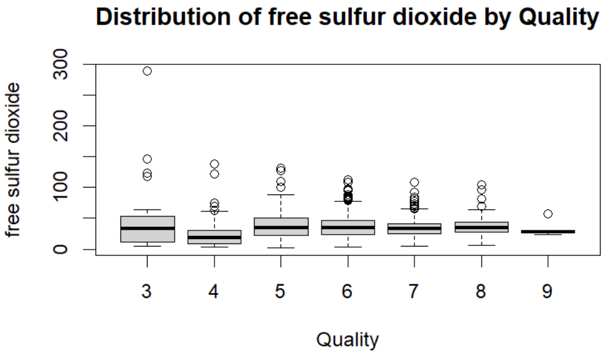
1.7 Free Sulfur dioxide

The variable "Free Sulphur Dioxide" has an average value of around 35.31, with values ranging from 2.00 to 289.00. There have been a total of 50 outliers detected, which suggests the existence of values that are either abnormally high or low. The correlation coefficient between the target variable and the variable in question is around zero (0.008), indicating a weak linear relationship.



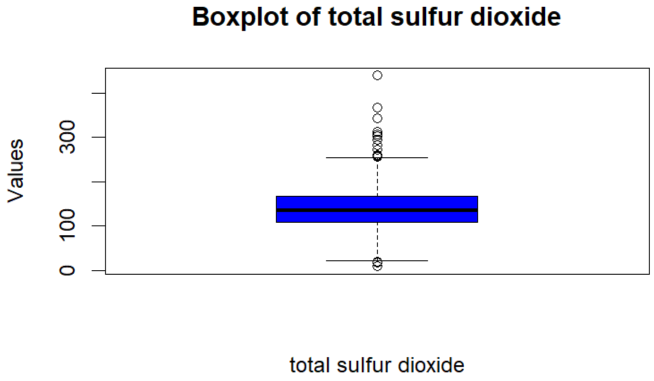


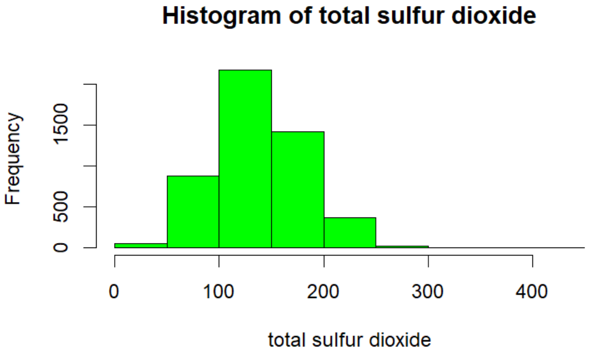


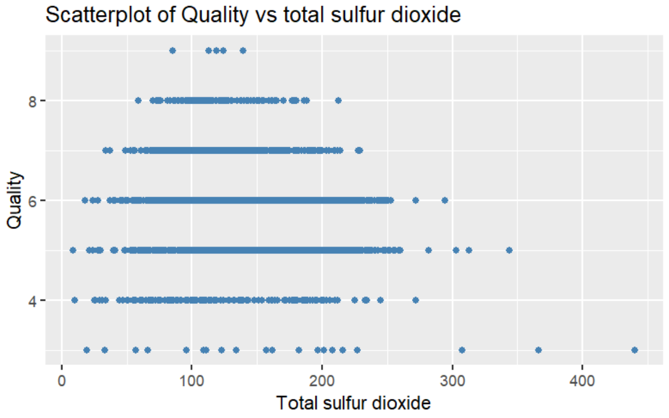


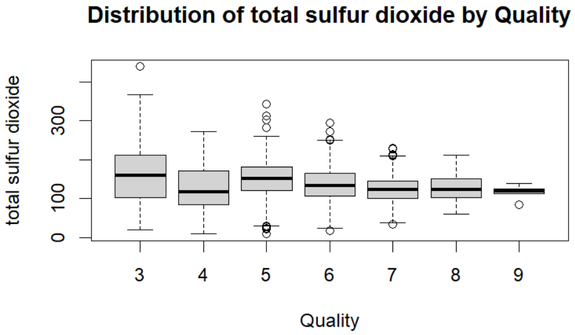
1.8 Total Sulfur dioxide

The mean value of total sulphur dioxide is around 138.4, with a range of values spanning from 9.0 to 440.0. There have been a total of 19 outliers detected, which are extremely high or low numbers. The negative correlation value of -0.175 suggests a weak relationship between higher levels of total sulphur dioxide and worse wine quality.



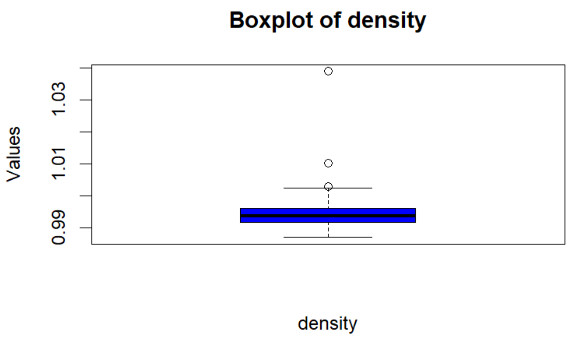


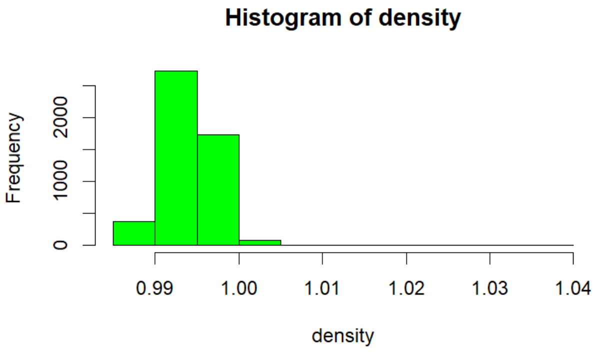


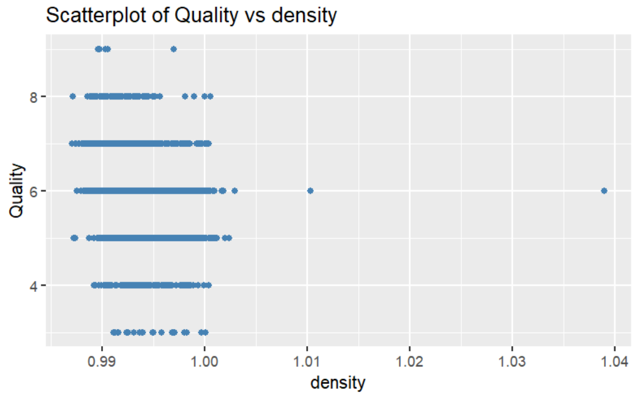


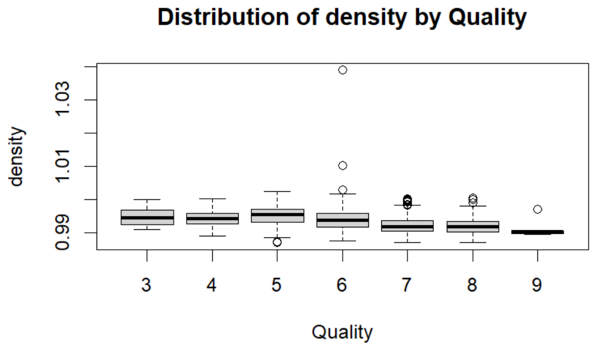
1.9 Density

The variable has an approximate mean of 0.9940, with values ranging from 0.9871 to 1.0390. Five data points that deviate significantly from the rest of the dataset have been detected, indicating the existence of extreme values. The presence of a negative correlation (-0.307) suggests an inverse association between density and wine quality. This implies that as density rises, wine quality generally declines.



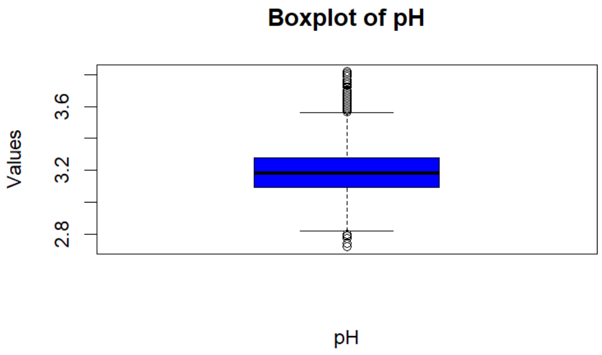


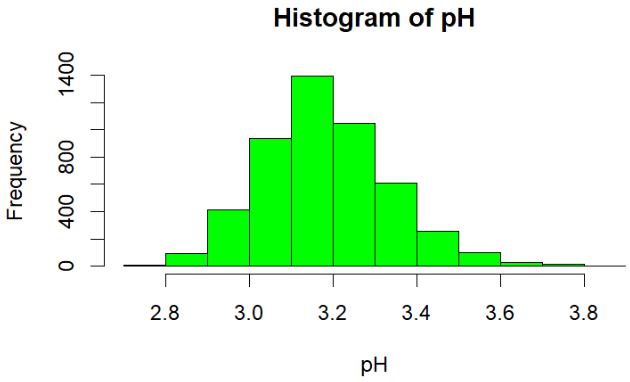


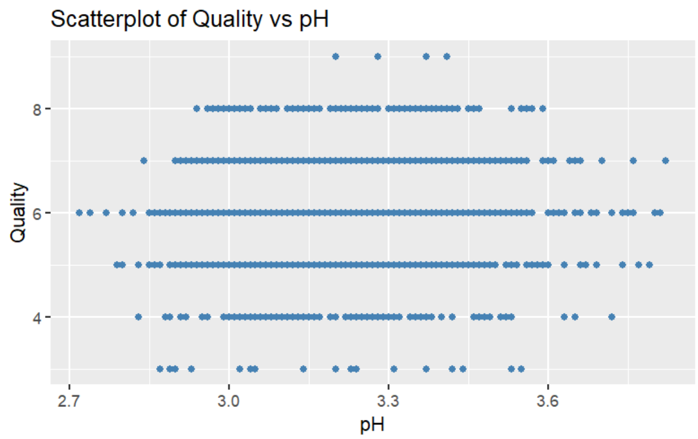


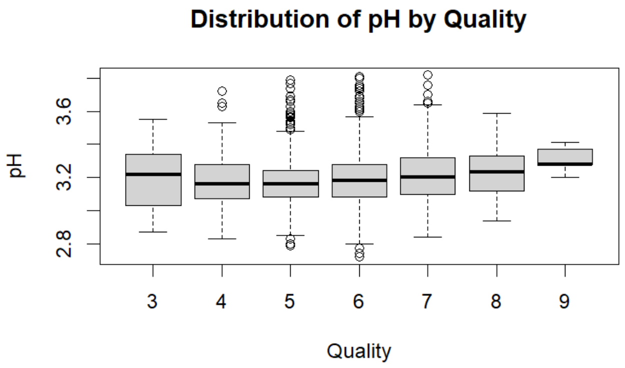
1.10 pH

The mean pH value is around 3.188, with a variation between 2.720 and 3.820. A total of 75 anomalies have been detected. The correlation coefficient between the target variable and pH levels is 0.099, suggesting a favourable association. These findings indicate a potential correlation between elevated pH levels and a minor improvement in wine quality.



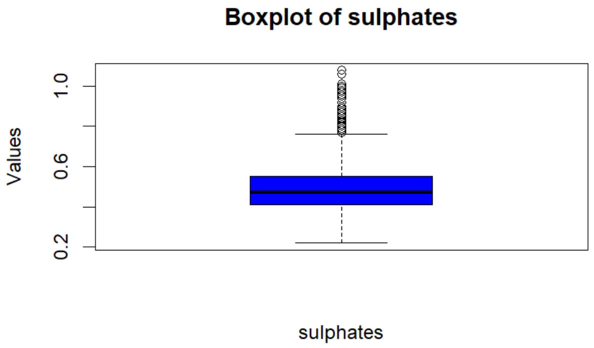


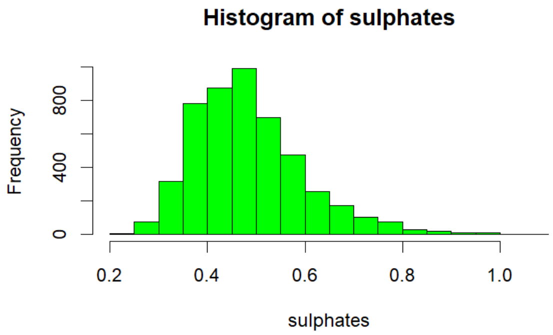


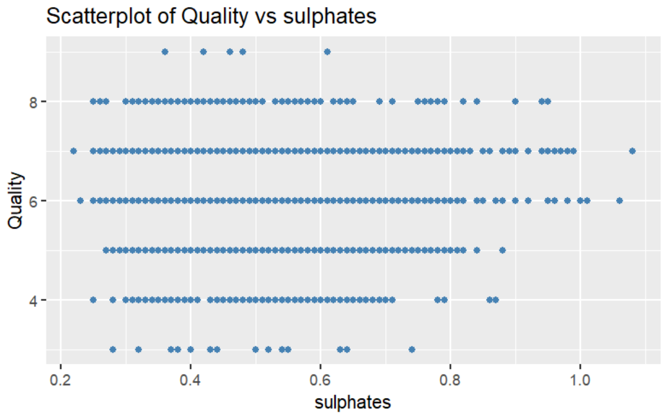


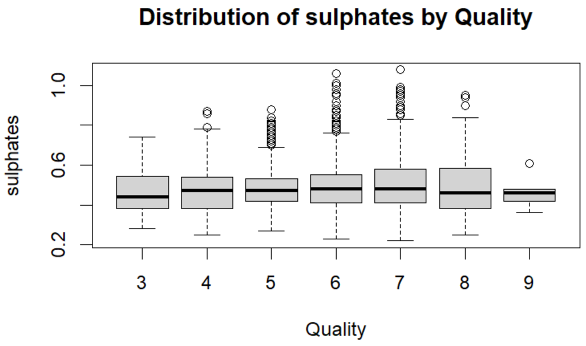
1.11 Sulphates

The mean value of this variable is around 0.4898, with values ranging from 0.2200 to 1.0800. There have been a total of 124 outliers detected, indicating the existence of extremely high or low values. The correlation coefficient between the target variable and sulphate levels is positive (0.054), suggesting a potential link between increased sulphate levels and somewhat improved wine quality.



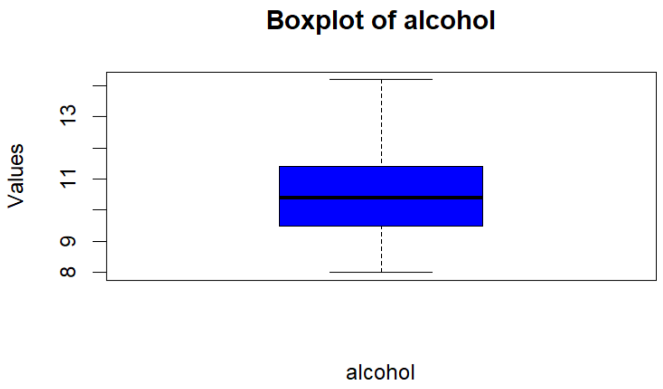


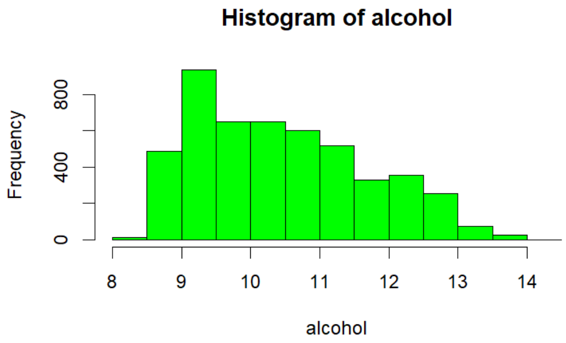


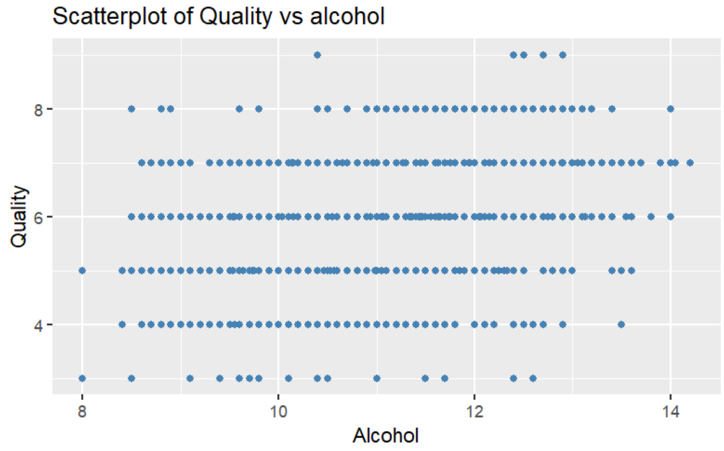


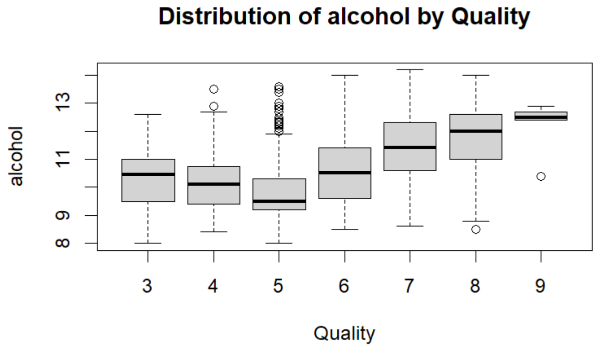
1.12 Alcohol

The mean alcohol content is around 10.51, with measurements varying from 8.00 to 14.20. There have been no instances of outliers detected. The target variable and alcohol level exhibit a robust positive connection (0.435), suggesting that when the alcohol concentration rises, the quality of the wine likewise tends to improve.



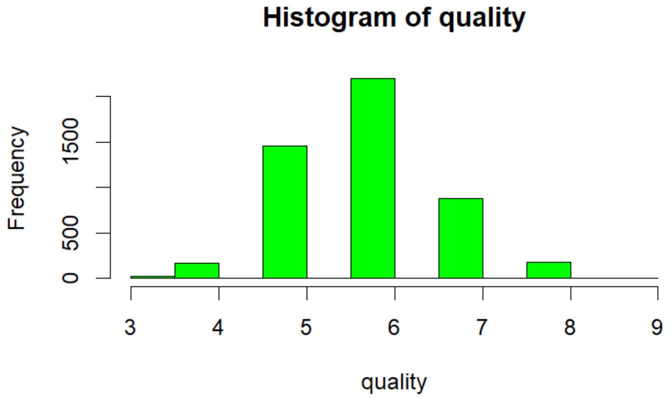


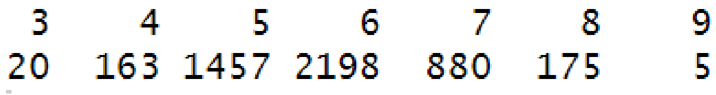




1.13 Quality

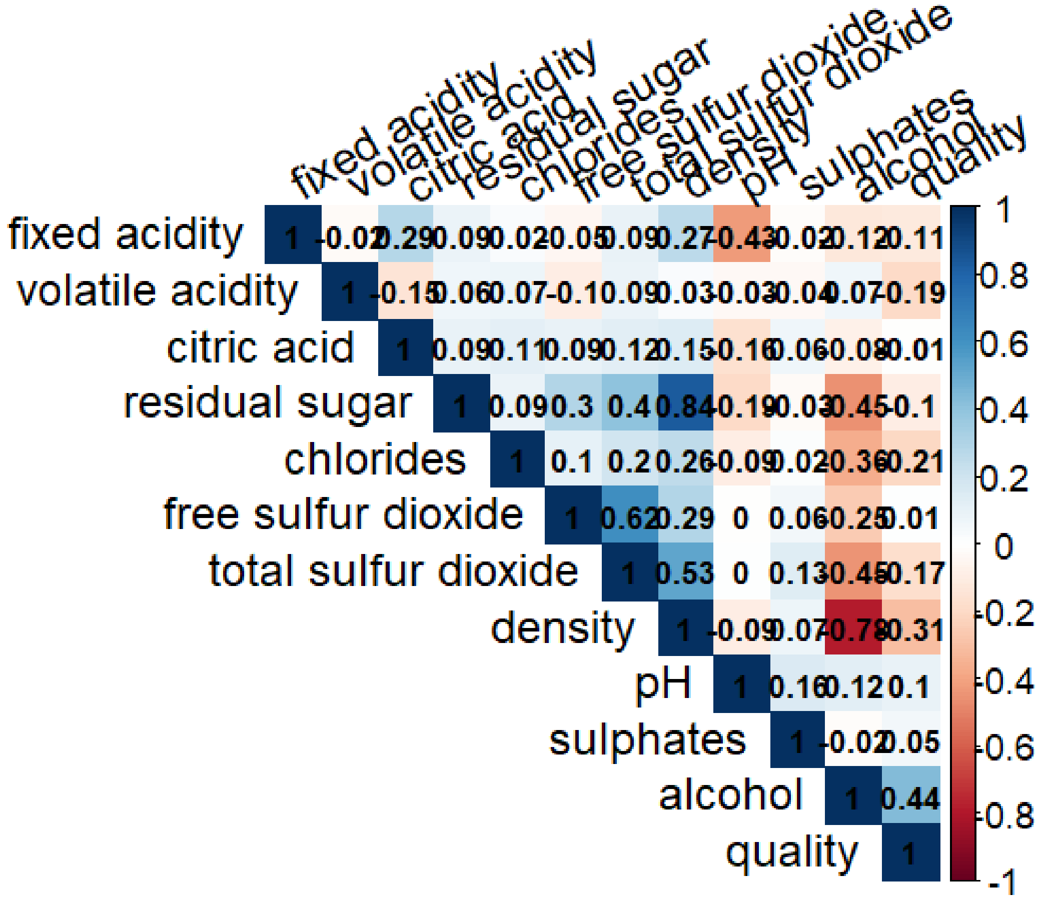
This variable is a discrete property that signifies the excellence of wine. This property has a range of values from 3 to 9. Because this data is categorical, it lacks descriptive statistics like the mean or quartiles. However, it is frequently used as the dependent variable for analysis.





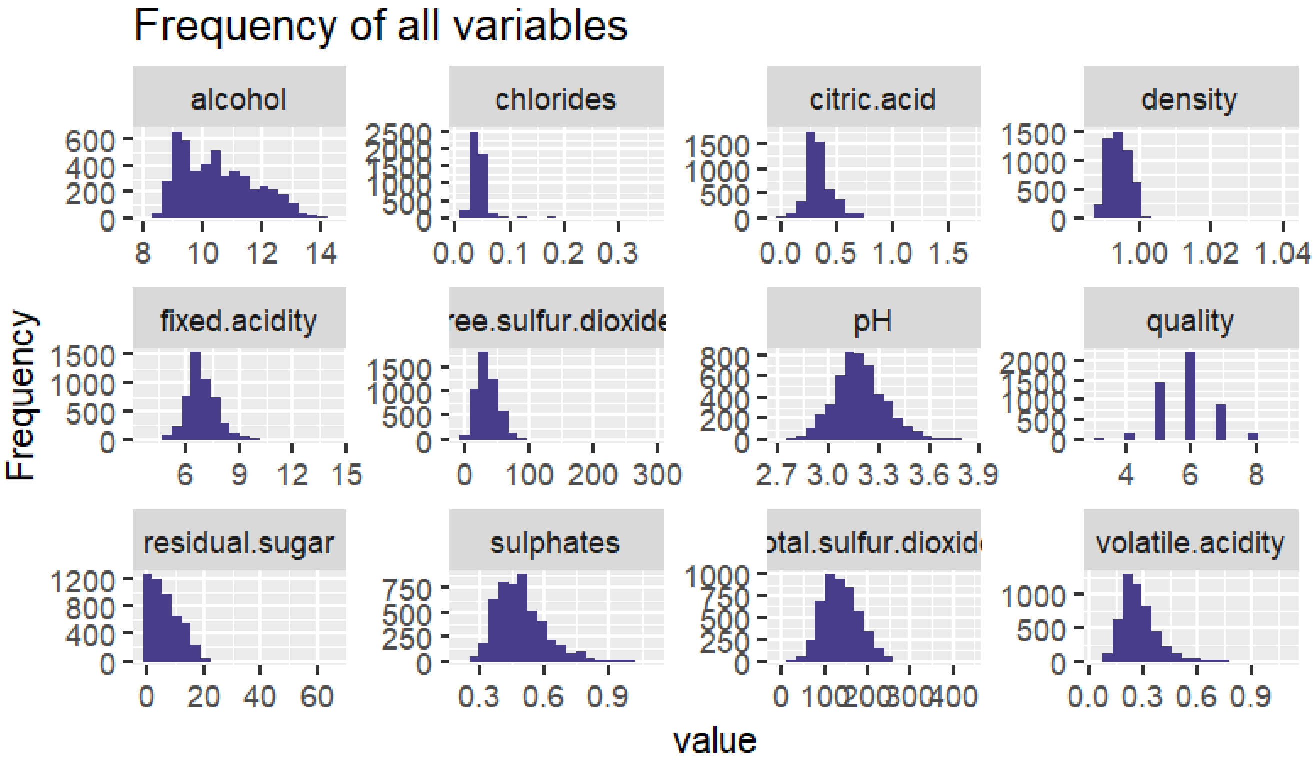
```
  3    4    5    6    7    8    9
 20  163 1457 2198  880  175    5
```

1.14 Correlation analysis

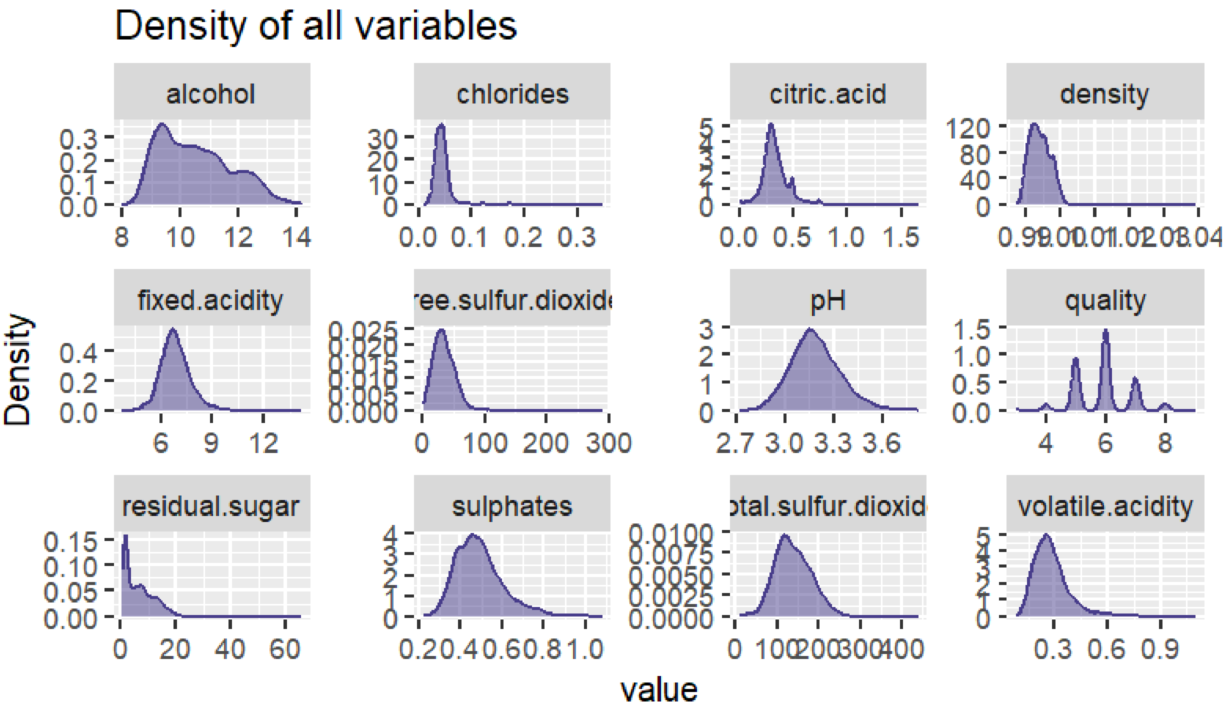


Variable	Correlation	Absolute Correlation	Rank
Alcohol	0.435	0.435	1
Density	-0.307	0.307	2
Chlorides	-0.209	0.209	3
Volatile acidity	-0.194	0.194	4
Total sulfur dioxide	-0.174	0.174	5
Fixed acidity	-0.113	0.113	6
pH	0.099	0.099	7
Residual sugar	-0.097	0.097	8
Sulphates	0.053	0.053	9
Citric acid	-0.009	0.009	10
Free sulfur dioxide	0.0081	0.008	11

1.15 Frequency of all variables



1.16 Density of all variables



2 Data preprocessing

Data preparation is a crucial step in the data mining process that involves eliminating mistakes, transforming data into a usable format, and arranging it for further analysis or modelling. This section will explore several data preparation techniques employed to preprocess a dataset for clustering analysis.

2.1 Removing target variable form data set

Clustering is unsupervised learning without labelled data and a target variable. Before clustering, the target variable is often removed for specific reasons. Clustering seeks patterns or groupings in data without explicit guidance. Unsupervised learning would fail if the method relied on the target variable for clustering. Prevent data leaking by excluding the target variable during clustering. Data leaking occurs when target variable information affects the model during training. Using feature space, the method may find natural groupings in the data without bias by removing the target variable before clustering.

2.2 Removing outliers

The dataset, after excluding outliers, consists of 4484 rows and 12 columns.

2.3 Min-Max scaling

One of the pre-processing methods employed is max-min standardisation of data. In this study, there are two instances of max-min scaling used: (1) with the original data before eliminating outliers, and (2) after removing outliers. Both of these data are utilised in various configurations of k-means clustering.

(1) Max-min scaling with original data before eliminating outliers

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed acidity	0.0000	0.2404	0.2885	0.2937	0.3365	1.0000
volatile acidity	0.0000	0.1275	0.1765	0.1944	0.2353	1.0000
citric acid	0.0000	0.1627	0.1928	0.2013	0.2349	1.0000
residual sugar	0.0000	0.01687	0.07055	0.08883	0.14264	1.0000
chlorides	0.0000	0.08012	0.10089	0.10912	0.12166	1.0000
free sulfur dioxide	0.0000	0.07317	0.11150	0.11606	0.15331	1.0000
total sulfur dioxide	0.0000	0.2297	0.2900	0.3001	0.3666	1.0000
density	0.0000	0.08892	0.12782	0.13336	0.17332	1.0000
pH	0.0000	0.3364	0.4182	0.4257	0.5091	1.0000
sulphates	0.0000	0.2209	0.2907	0.3138	0.3837	1.0000
alcohol	0.0000	0.2419	0.3871	0.4055	0.5484	1.0000

(2) Max-min scaling after eliminating outliers

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed acidity	0.0000	0.3878	0.4898	0.4984	0.5918	1.0000
volatile acidity	0.0000	0.2653	0.3673	0.3892	0.4898	1.0000
citric acid	0.0000	0.3913	0.4493	0.4736	0.5507	1.0000
residual sugar	0.0000	0.05446	0.23267	0.28781	0.46040	1.0000
chlorides	0.0000	0.2473	0.3226	0.3343	0.4086	1.0000
free sulfur dioxide	0.0000	0.2530	0.3855	0.3953	0.5181	1.0000
total sulfur dioxide	0.0000	0.3693	0.4730	0.4918	0.6100	1.0000
density	0.0000	0.3064	0.4438	0.4616	0.6054	1.0000

pH	0.0000	0.3571	0.4643	0.4741	0.5833	1.0000
sulphates	0.0000	0.3167	0.4167	0.4422	0.5333	1.0000
alcohol	0.0000	0.1897	0.3448	0.3689	0.5172	1.0000

2.4 Apply Z-score scaling

The applied pre-processing approach is z-score standardisation of the data. This study used two instances of max-min scaling: (1) on the original data before to outlier elimination, and (2) after the removal of outliers. Both of these datasets are used in different variations of k-means clustering.

(1) Z-score standardisation with original data before eliminating outliers

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed acidity	-3.61998	-0.65743	-0.06492	0.00000	0.52758	8.70422
volatile acidity	-1.9668	-0.6770	-0.1810	0.0000	0.4143	8.1528
citric acid	-2.7615	-0.5304	-0.1173	0.0000	0.4612	10.9553
residual sugar	-1.1418	-0.9250	-0.2349	0.0000	0.6917	11.7129
chlorides	-1.6831	-0.4473	-0.1269	0.0000	0.1935	13.7417
free sulfur dioxide	-1.95848	-0.72370	-0.07691	0.00000	0.62867	14.91679
total sulfur dioxide	-3.0439	-0.7144	-0.1026	0.0000	0.6739	7.0977
density	-2.31280	-0.77063	-0.09608	0.00000	0.69298	15.02976
pH	-3.10109	-0.65077	-0.05475	0.00000	0.60750	4.18365
sulphates	-2.3645	-0.6996	-0.1739	0.0000	0.5271	5.1711
alcohol	-2.04309	-0.82419	-0.09285	0.00000	0.71974	2.99502

(2) Z- score standardisation after eliminating outliers

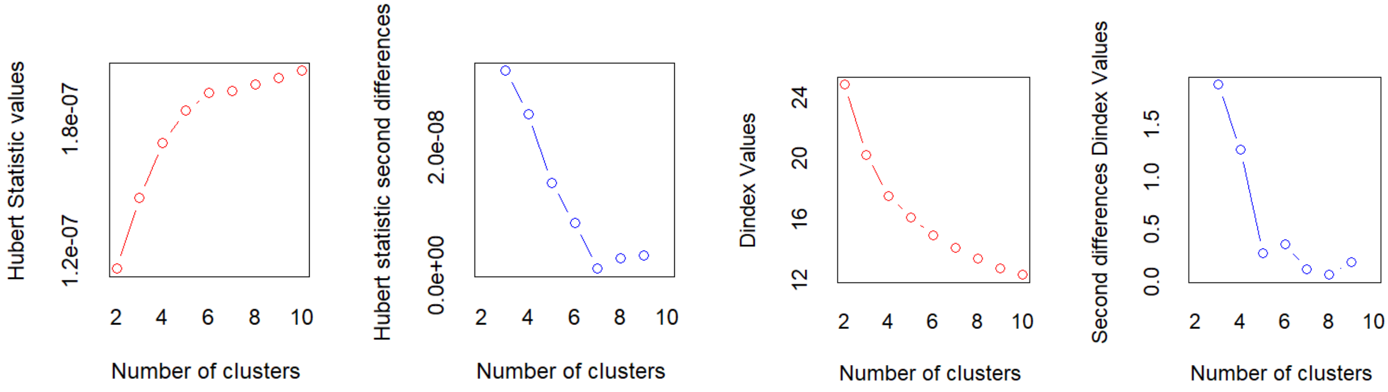
Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fixed acidity	-3.10820	-0.69002	-0.05366	0.00000	0.58270	3.12816
volatile acidity	-2.2408	-0.7135	-0.1260	0.0000	0.5790	3.5162
citric acid	-3.2416	-0.5634	-0.1666	0.0000	0.5277	3.6027
residual sugar	-1.1730	-0.9510	-0.2247	0.0000	0.7034	2.9026
chlorides	-2.69753	-0.70181	-0.09441	0.00000	0.59975	5.37213
free sulfur dioxide	-2.12921	-0.76647	-0.05266	0.00000	0.66116	3.25685
total sulfur dioxide	-2.8670	-0.7140	-0.1093	0.0000	0.6890	2.9629
density	-2.35792	-0.79288	-0.09119	0.00000	0.73433	2.74996
pH	-2.77980	-0.68557	-0.05731	0.00000	0.64077	3.08403
sulphates	-2.5110	-0.7127	-0.1448	0.0000	0.5177	3.1679
alcohol	-1.7481	-0.8494	-0.1142	0.0000	0.7028	2.9904

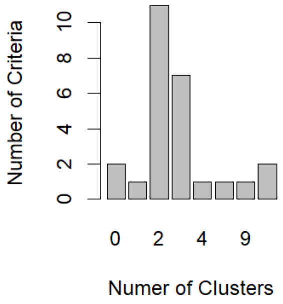
3 Partitioning Clustering (K-Mean)

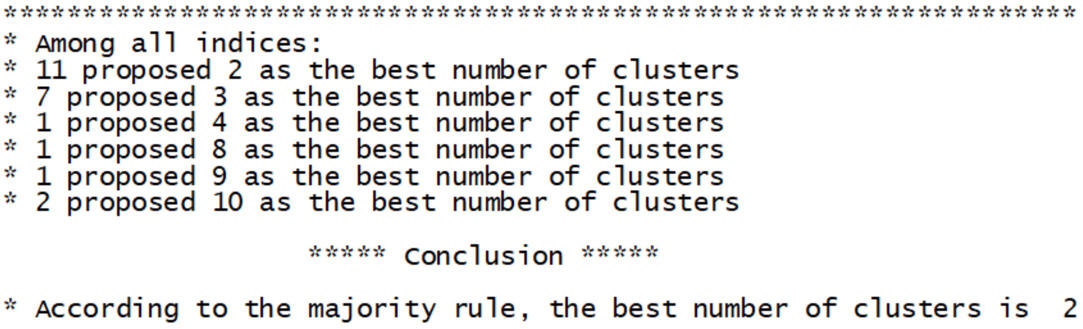
3.1 Find the ideal number of clusters

3.1.1 NbClust method

NbClust is a specialised algorithm designed to identify the ideal amount of clusters in a data set. The NbClust contains 30 indices that measure the number of clusters. Furthermore, it has the capability to execute k-means and hierarchical clustering using many distance metrics and aggregation approaches. This facilitates the determination of the most suitable number of clusters. This study employed the "euclidean" distance measure.







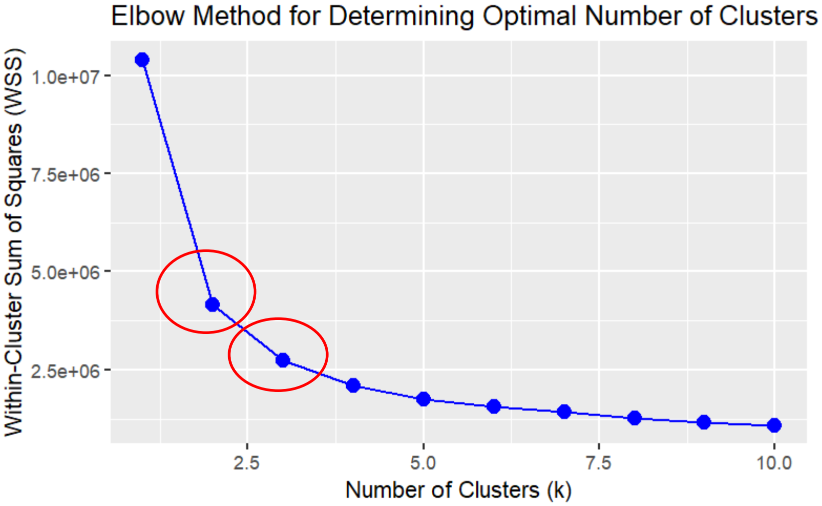
```
*******************************************************************
* Among all indices:
* 11 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 2 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

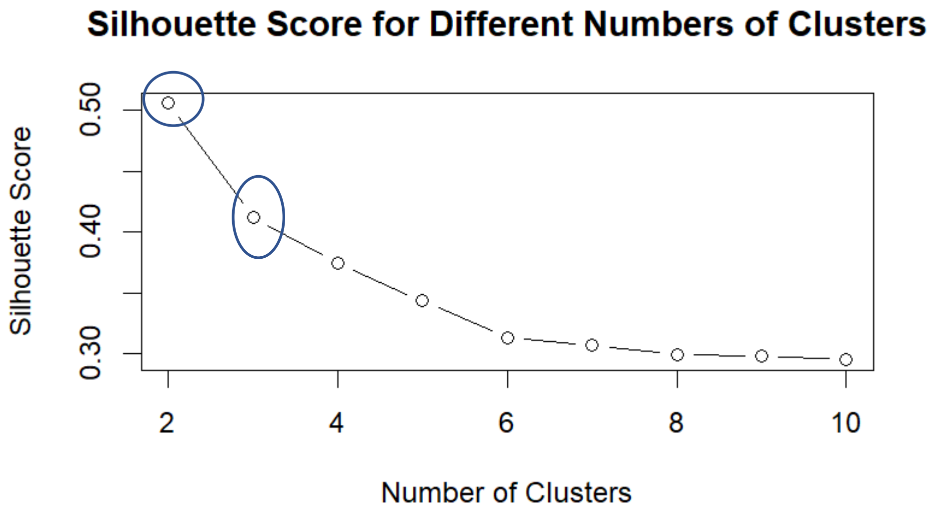
3.1.2 Elbow method

The "elbow method" is a commonly used strategy to determine the optimal number of clusters by calculating the within-cluster sum of squares (withinss) for various cluster numbers.



3.1.3 silhouette scores

The silhouette score is a technique used to determine the most suitable number of clusters, similar to the "elbow method," but with a specific emphasis on silhouette analysis. The silhouette score quantifies the degree of similarity between an item and its corresponding cluster (cohesion) relative to other clusters (separation). The silhouette score is a numerical measure that varies between -1 and 1.



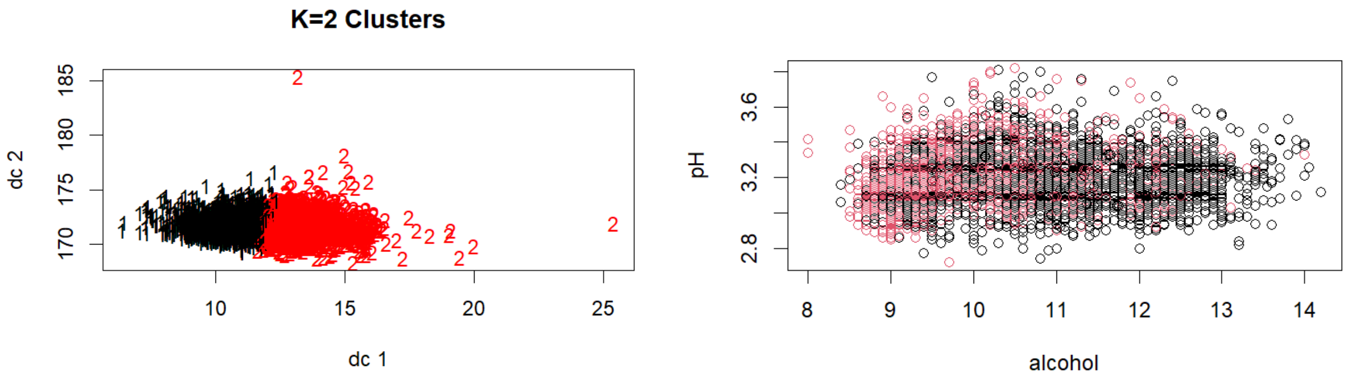
3.1.4 Summary

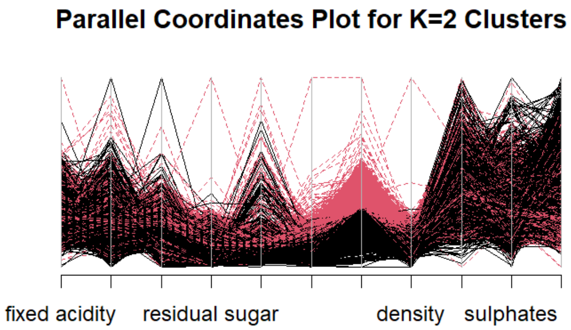
According to the Nbclust analysis, the ideal number of clusters is determined to be K=2, with K=3 being the second most favourable choice based on the chart. Based on the elbow method, the most suitable number of clusters is K=2, while the second best choice is K=3, as seen by the chart. According to the Silhouette analysis, the value of K=2 is deemed to be the most ideal, with the K=3 option being the next best choice. Therefore, K=2 is the best number of clusters for the whitewine dataset, while K=3 is the next ideal number of clusters.

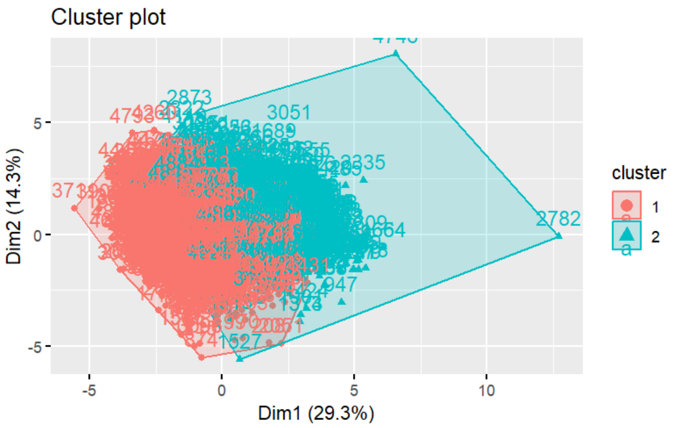
3.2 Perform K-means algorithms and evaluation

3.2.1 K=2

There are two clusters, consisting of 2798 events in cluster 1 and 2100 events in cluster 2. The clustering method was executed using 25 distinct initial stages configurations (nstart).





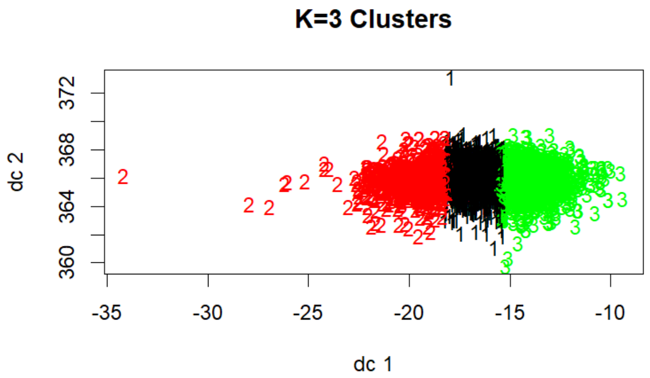


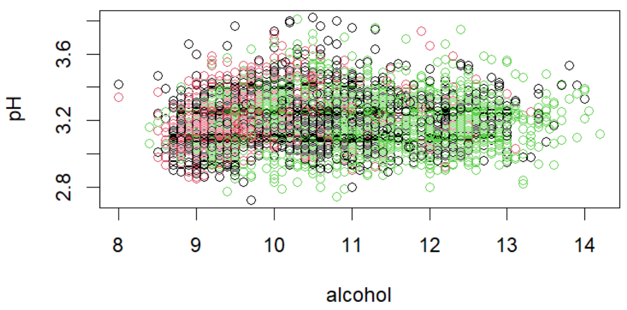
View final data for k=2

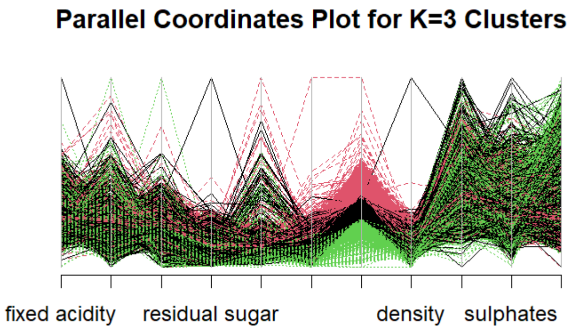
Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol	Quality	Cluster
7.0	0.27	0.36	20.7	0.045	45	170	1.0010	3.00	0.45	8.8	6	2
6.3	0.30	0.34	1.6	0.049	14	132	0.9940	3.30	0.49	9.5	6	1
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	6	1
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.40	9.9	6	2
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.40	9.9	6	2
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	6	1

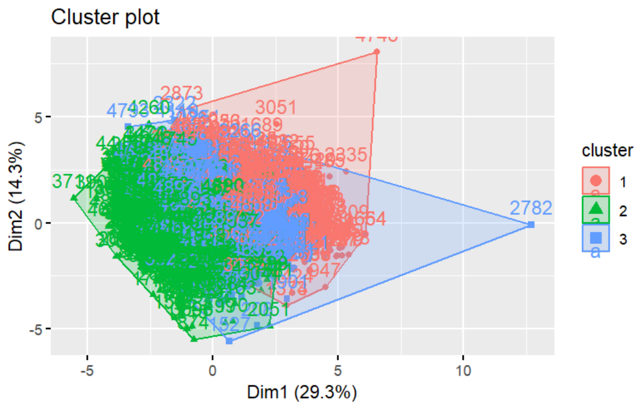
3.2.2 K = 3

There are three clusters, each containing a different number of observations. Cluster 1 has 1125 observations, cluster 2 has 1797 observations, and cluster 3 has 1976 observations. The clustering method was executed using 25 distinct startup configurations (nstart).









View final data for k=3

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	cluster
7.0	0.27	0.36	20.7	0.045	45	170	1.0010	3.00	0.45	8.8	6	2
6.3	0.30	0.34	1.6	0.049	14	132	0.9940	3.30	0.49	9.5	6	1
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	6	3
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.40	9.9	6	2
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.40	9.9	6	2
8.1	0.28	0.40	6.9	0.050	30	97	0.9951	3.26	0.44	10.1	6	3

3.3 Evaluating clustering

3.3.1 Silhouette analysis

		K=2		**K=3**		
		Cluster 1	**Cluster 2**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Cluster sizes		2798	2100	1125	1797	1976
average silhouette widths		0.5477614	0.4508931	0.3943549	0.4648050	0.3750042
Individual silhouette widths	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
K= 2	-0.04368	0.41426	0.56828	0.50623	0.63540	0.70346
K = 3	-0.04118	0.28761	0.45950	0.41240	0.56264	0.64426

K = 2: Cluster 1 has a mean silhouette width of around 0.548. Cluster 2 has an average silhouette width of around 0.451. The range of individual silhouette widths spans from around -0.044 to 0.703, with a mean value of around 0.506. Cluster 1 has a higher average silhouette width than Cluster 2, indicating that the data points within Cluster 1 are more comparable to each other than they are to the data points in Cluster 2. The silhouette score for K=2 suggests that the clustering is reasonable, as indicated by a mean silhouette width of around 0.506. This shows a moderate amount of separation across the clusters.

K = 3: Cluster 1 has a mean silhouette width of around 0.394. Cluster 2 has an approximate average profile width of 0.465. Cluster 3 has an average silhouette width of around 0.375. The range of individual silhouette widths is around -0.041 to 0.644, with a mean of roughly 0.412. Cluster 2 has the highest average silhouette width, indicating that the data points within Cluster 2 are more similar to each other compared to the data points in another cluster. Cluster 1 and Cluster 3 have much lower average silhouette widths compared to Cluster 2, suggesting that the differentiation between data points within these clusters may not be as apparent. The silhouette score for K=3 is lower than K=2, with an average silhouette width of around 0.412. This indicates that the clustering may not be as optimal as when K=2.

Conclusion: The clustering with a value of K=2 demonstrates a more pronounced differentiation across clusters, as seen by the greater average silhouette width and total silhouette score. Therefore, the most favourable outcome for K-means clustering, taking into account K=2 and K=3, is K=2.

3.3.2 Cluster center analysis

To assess both K=2 and K=3, let will investigate the mean of each cluster by evaluating the cluster centres for each sample.

3.3.2.1 Cluster Centers for K=2

Variable	Cluster 1	Cluster 2
Fixed Acidity	6.785150	6.947571
Volatile Acidity	0.2728681	0.2854000
Citric Acid	0.3218799	0.3505952
Residual Sugar	4.793352	8.520643
Chlorides	0.04217691	0.05056286
Free Sulfur Dioxide	27.16976	46.15143
Total Sulfur Dioxide	108.5960	178.0186
Density	0.9927779	0.9956921
pH	3.188867	3.187467
Sulphates	0.4793531	0.5038286
Alcohol	10.962258	9.917373

Cluster 1 has lower values for most characteristics in comparison to Cluster 2. On the other hand, Cluster 2 demonstrates higher values for residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, and density. According on these criteria, it seems that these clusters may be classified as two distinct categories.

3.3.2.2 Cluster Centers for K=3

Variable	Cluster 1	Cluster 2	Cluster 3
Fixed Acidity	6.858882	6.970133	6.778075
Volatile Acidity	0.2742713	0.2934800	0.2730662
Citric Acid	0.3384565	0.3533956	0.3174791
Residual Sugar	6.722874	9.356089	4.170924
Chlorides	0.04677328	0.05179556	0.04090095
Free Sulfur Dioxide	36.88790	50.69689	23.93684
Total Sulfur Dioxide	143.48634	196.93911	96.05175
Density	0.9942740	0.9963027	0.9923317
pH	3.191624	3.183013	3.187863
Sulphates	0.4846761	0.5149511	0.4798164
Alcohol	10.432463	9.707156	11.109507

Cluster 1 exhibited moderate levels for most variables in comparison to the other clusters. In Cluster 2, the levels of residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, and density are higher, similar to Cluster 2 in K=2. Cluster 3 often exhibits lower values for most qualities, similar to Cluster 1 in the case of K=2. By including Cluster 3, we are able to include more data points that possess distinct characteristics.

Conclusion: The selection of K=2 or K=3 depends on the desired level of clustering granularity. When K=2 is used, the process of separating into two groups becomes simpler. A segmentation with K=3 is more extensive as it contains an intermediate group. The selection should be made by considering the specific analytic objectives and the distinctive properties of the clusters that are relevant to the dataset.

3.4 Consistency with quality

	K = 2	K = 3
Confusion tables	1 2 Total 3 9 11 20 4 107 56 163 5 631 826 1457 6 1282 916 2198 7 644 236 880 8 120 55 175 9 5 0 5 4898	1 2 3 Total 3 3 9 8 20 4 36 36 91 163 5 575 514 368 1457 6 889 473 836 2198 7 395 73 412 880 8 75 20 80 175 9 3 0 2 5 4898
Accuracy	0.0132707227439771	0.0830951408738261
Precision	0.0149416249702168	0.0794346575526612
Recall	0.422205762589453	0.481223398339446
F1-score	0.0209592439236021	0.0998130588998533
ARI	0.01522975	0.01297025

Confusion Tables displays the counts of actual labels compared to the predicted labels for different degrees of quality. Each row displays the current quality level, while each column represents the anticipated level. There are no missing values in Confusion tables.

Precision refers to the proportion of correctly predicted positive cases among all the predicted positive occurrences. **The aforementioned models exhibit a notable frequency of false positives, as seen by their low accuracy ratings.**

Recall, often referred to as sensitivity, is a measure of the proportion of true positive situations in relation to all positive occurrences. **The higher memory scores of the aforementioned models suggest a superior ability to detect positive occurrences in comparison to accuracy and precision.**

The F1-score is a metric that balances accuracy and memory by using the harmonic mean. **The F1-score values above suggest a lack of precision and recall in the model, leading to subpar performance**.

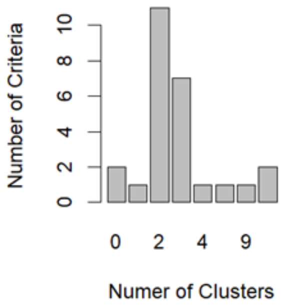
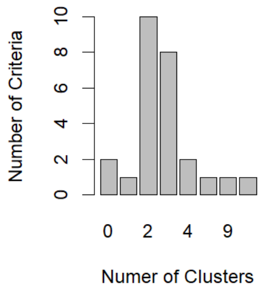
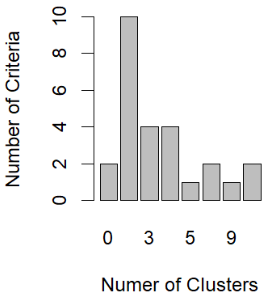
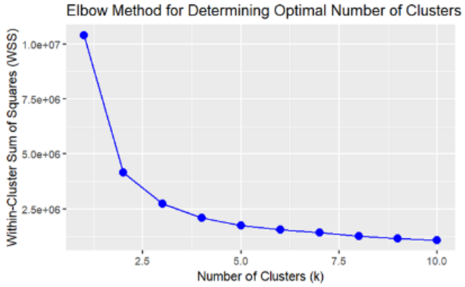
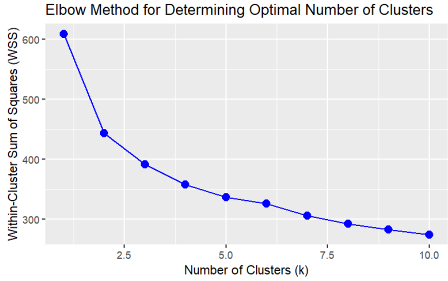
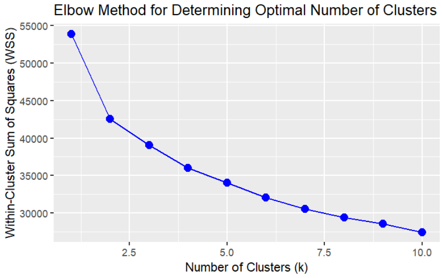
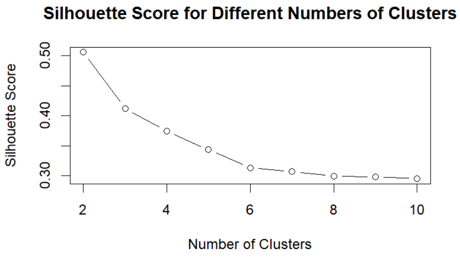
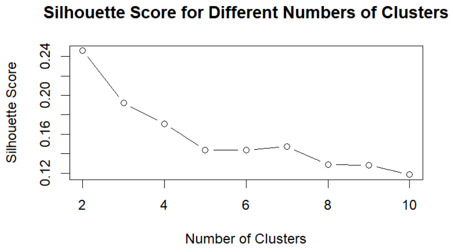
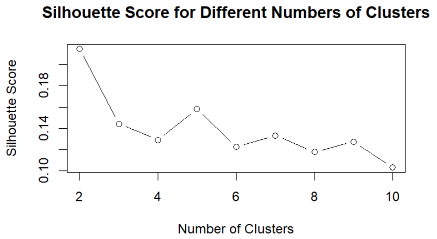
The Adjusted Rand Index (ARI) is a measure that evaluates the similarity of clustering findings by considering all pairs of samples and counting those allocated to the same or different clusters in anticipated and true labels. **Low ARI values signify inadequate concordance between predicted and actual labels**.

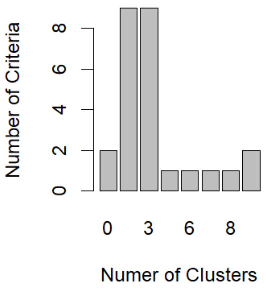
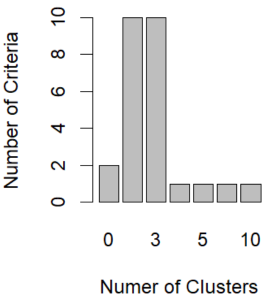
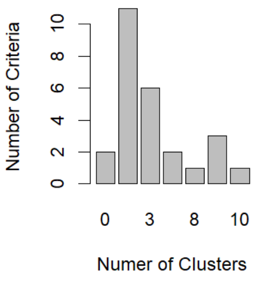
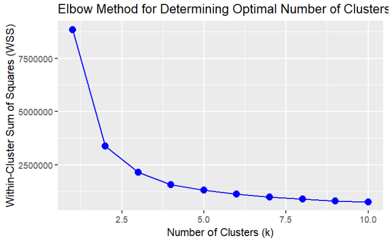
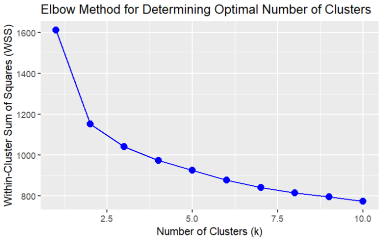
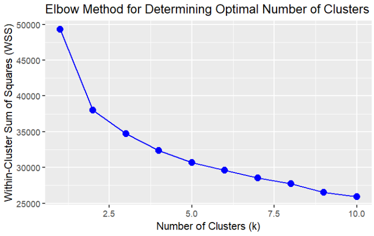
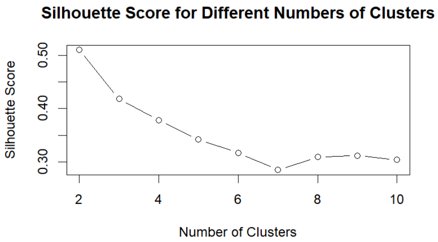
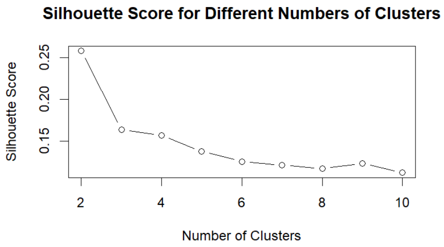
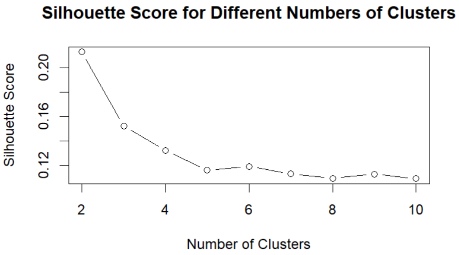
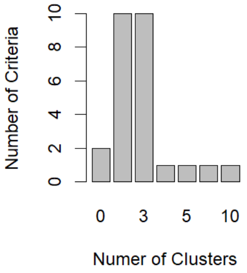
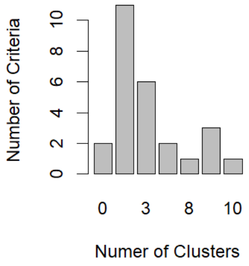
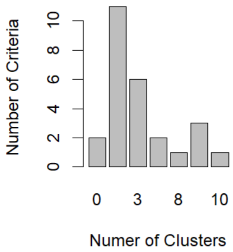
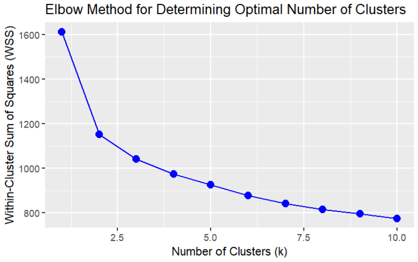
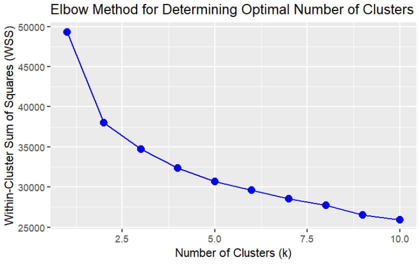
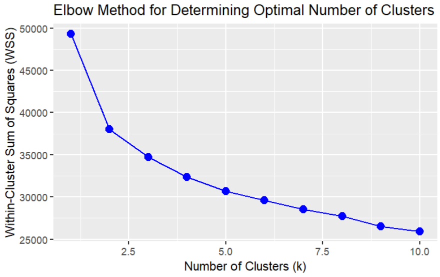
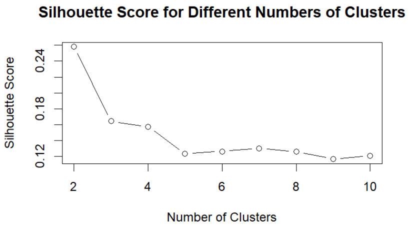
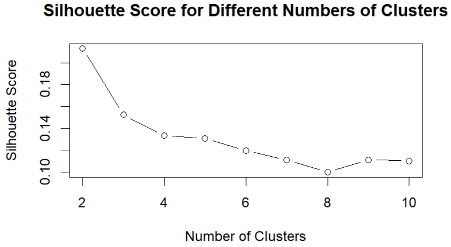
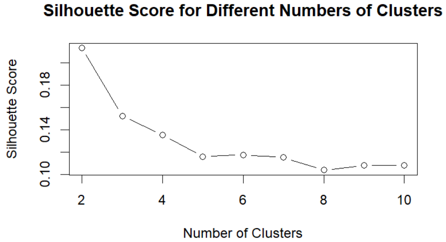
In conclusion, the models frequently make incorrect classifications of wine quality. Low accuracy, precision, and F1-score imply high misclassification rates. The recall rate is higher, but, it is insufficient to compensate for the poor level of accuracy. The ARI values suggest that the clustering results do not align with the given labels. These findings indicate that improving the characteristics of the features, adjusting the model, or employing different algorithms might potentially improve the accuracy of predicting wine quality.

3.5 Alternative input configurations

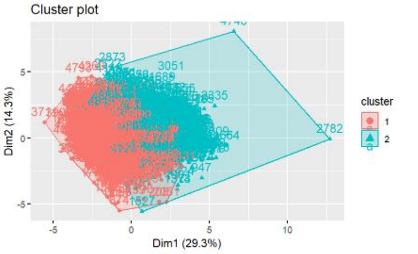
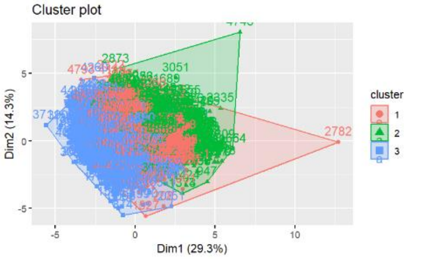
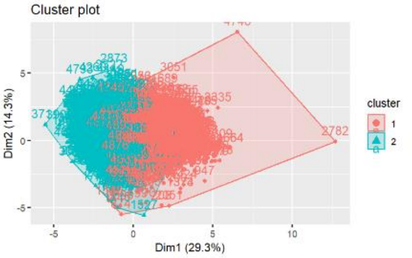
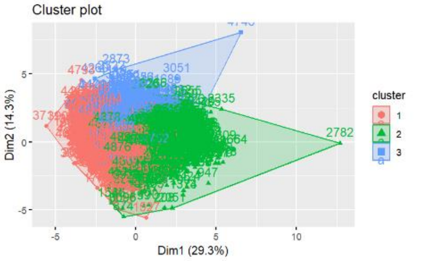
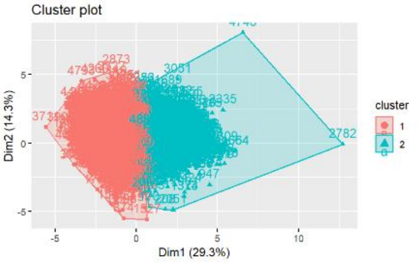
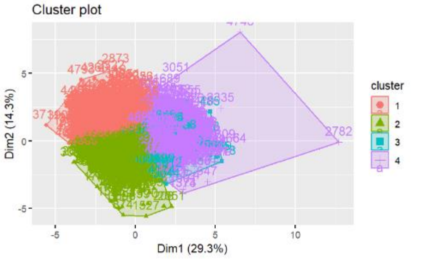
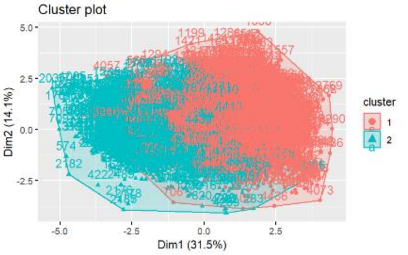
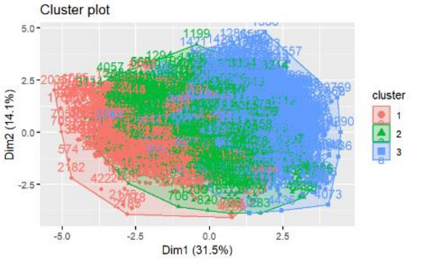
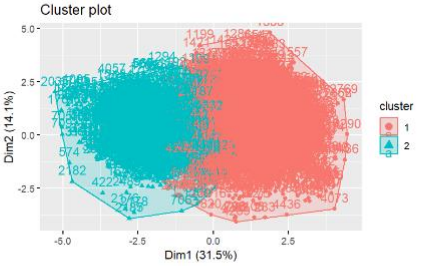
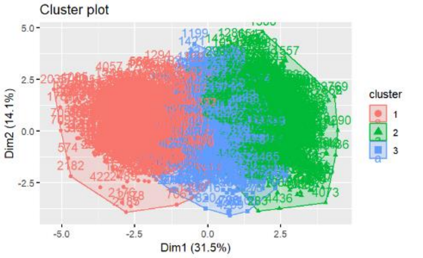
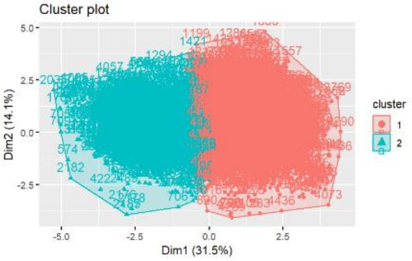
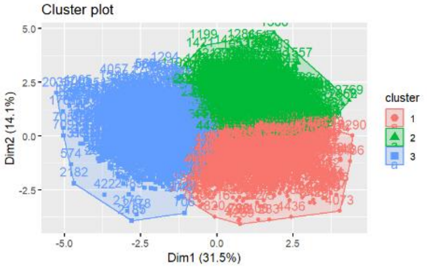
Evaluate the performance of the k-means method using various combinations of starting centroids (nstart = 25, 50, 10), alternate scaling procedures such as z-score standardisation and min-max normalisation, and different preprocessing tactics, including the inclusion of outliers.

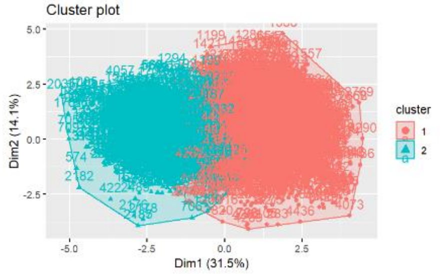
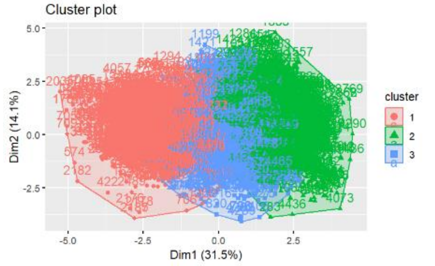
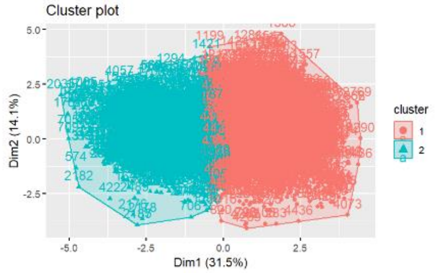
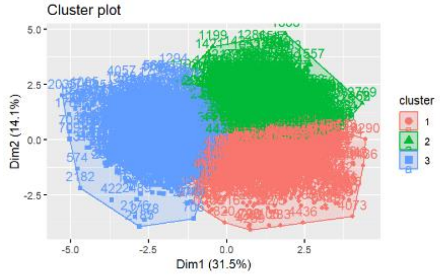
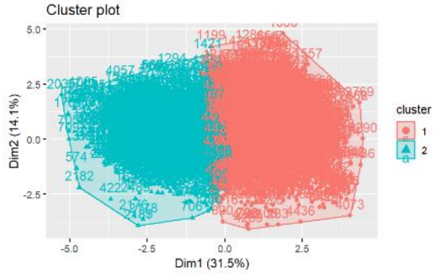
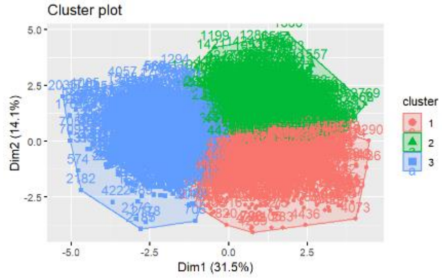
Config	nstart	Outlier	Scaling
1	25	Not removed	Not done
2	25	Not removed	Min-max
3	25	Not removed	Z-score
4	25	Removed	Not done
5	25	Removed	Min-max
6	25	Removed	Z-score
7	50	Removed	Min-max
8	50	Removed	Z-score
9	10	Removed	Z-score

Configuration 1 K= 2 and K= 3	Configuration 2 K = 2 and K = 3	Configuration 3 k = 2 and K = 4
		
		
		

Configuration 4 K = 2 and K = 3	Configuration 5 : K = 2 and K = 3	Configuration 6: K = 2 and K = 3
		
		
		
Configuration 7 k = 2 and K = 3	**Configuration 8: K = 2 and K = 3**	**Configuration 9 k = 2 and K = 3**
		
		
		

The provided visualization presents the cluster demarcation of several setups (Config1 to Config9) utilising varying values of K (the number of clusters) in a clustering algorithm.

Config	**K = 2**		**K = 3/4**	
Config1 (base)	2798 2100		1976 1125 1797	
Config2	2654 2244		1619 2023 1256	
Config3	2941 1957		1632 1462 107 1697	
Config4	2558 1926		1059 1790 1635	
Config5	2814 1670		1516 1355 1613	
Config6	2677 1807		1494 1325 1665	

Config7	2814 1670		1516 1355 1613	
Config8	2677 1807		1494 1325 1665	
Config9	2677 1807		1494 1325 1665	

The provided data presents the performance metrics of several setups (Config1 to Config9) utilising varying values of K (the number of clusters) in a clustering algorithm.

Configuration	K	Precision	Recall	F1-score	ARI
Config1(base)	K=2	0.014	0.422	0.020	0.015
	K=3	0.079	0.481	0.099	0.012
Config2	K=2	0.015	0.603	0.022	0.078
	K=3	0.093	0.500	0.111	0.067
Config3	K=2	0.014	0.380	0.020	0.025
	K=4	0.238	1.608	0.215	0.028
Config4	K=2	0.012	0.390	0.017	0.014
	K=3	0.071	0.442	0.090	0.013
Config5	K=2	0.012	0.378	0.016	0.016
	K=3	0.111	0.667	0.140	0.042
Config6	K=2	0.012	0.343	0.016	0.026
	K=3	0.158	0.874	0.200	0.034
Config7	K=2	0.012	0.378	0.016	0.016
	K=3	0.111	0.667	0.140	0.042
Config8	K=2	0.012	0.343	0.016	0.026
	K=3	0.158	0.874	0.200	0.034
Config9	K=2	0.012	0.343	0.016	0.026
	K=3	0.158	0.874	0.200	0.034

Here is a brief summary of the findings

Configurations Overview

Each setup is tested with different values of K. The value of K remains constant at 2 throughout all configurations, but in other configurations, K varies but the majority has a value of 3.

Assessment of Metrics

Precision is a quantitative measure that determines the proportion of accurately detected positive cases out of the total number of anticipated positive occurrences. The findings range from 0.012 to 0.238 across different configurations and K values.Recall is a quantitative measure that represents the percentage of true positive cases that have been properly recognised. The recall values range from 0.343 to 1.608. The F1-score is a statistical measure that quantifies the harmonic mean of accuracy and recall. It provides an equitable balance between these two metrics. The range extends from 0.016 to 0.215. The Adjusted Rand Index (ARI) measures the level of similarity between the observed grouping and the anticipated clustering. The range varies between 0.012 and 0.078.

Findings

Experiments show that configurations with greater values of K tend to have better accuracy, recall, and F1-scores. This indicates that as the number of clusters increases, the clustering capacity improves. There are differences in performance measures among various configurations, suggesting that the selection of configuration significantly impacts the quality of clustering.

Configurations with K=3 or K=4 often demonstrate better performance metrics than K=2 in most cases, however there could be a few exceptions.

For configurations 6, 8, and 9, when K is set to 2 and 3, there is a noticeable separation between clusters that may be easily observed visually.

Recommendation

Further investigation is necessary to determine the optimal configuration for the particular dataset and clustering goal. This may involve comparing more metrics, including visual analysis of clusterings, and maybe exploring a wider range of K values.

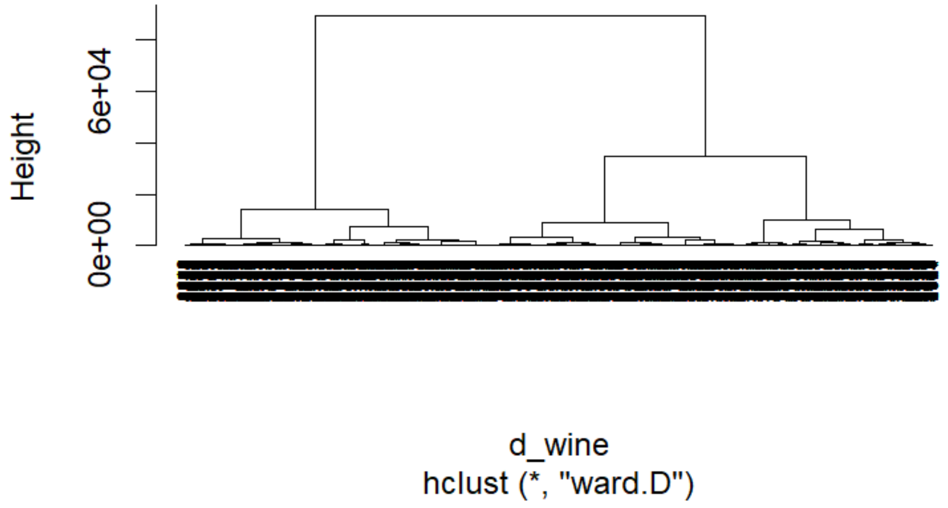
4 Hierarchical Clustering

Agglomerative clustering is also known as AGNES (Agglomerative Nesting). It functions in a hierarchical manner, commencing at the lowest level and advancing higher. Initially, every object is considered as an own cluster, referred to as a leaf. Within each iteration of the process, the two clusters that demonstrate the greatest level of similarity are combined to create a bigger cluster. This procedure is iterated until all data points are assigned to a single, cohesive cluster, known as the root. The result is a hierarchical arrangement that may be visually shown as a dendrogram. I do agglomerative hierarchical clustering using the **hclust** function. Firstly, I compute the dissimilarity values by utilising the **dist** function. Afterwards, I enter these values into the **hclust** function and clearly specify the preferred agglomeration strategy. Afterwards, I create a **dendrogram** and display it graphically.

4.1 Hierarchical clustering with different methods

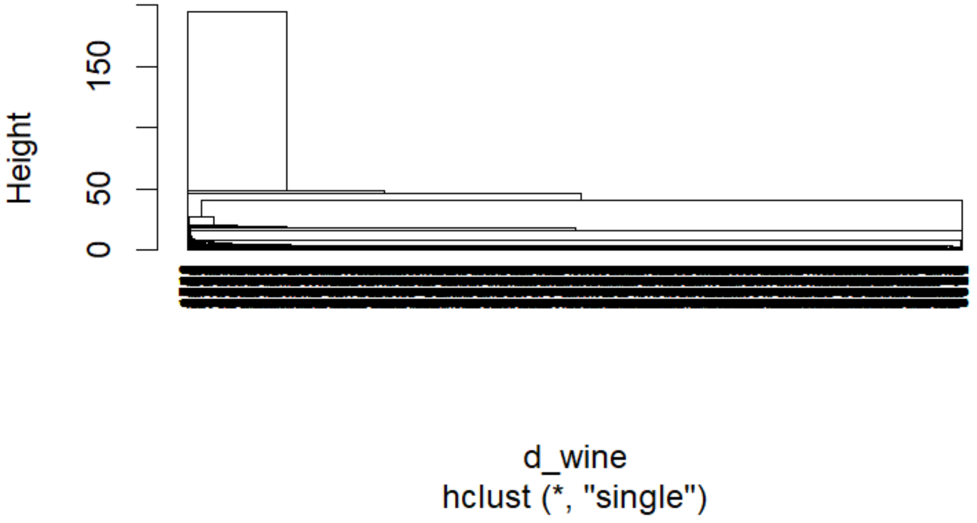
4.1.1 ward.D

2 branches and 4898 members total, at height 89388.34



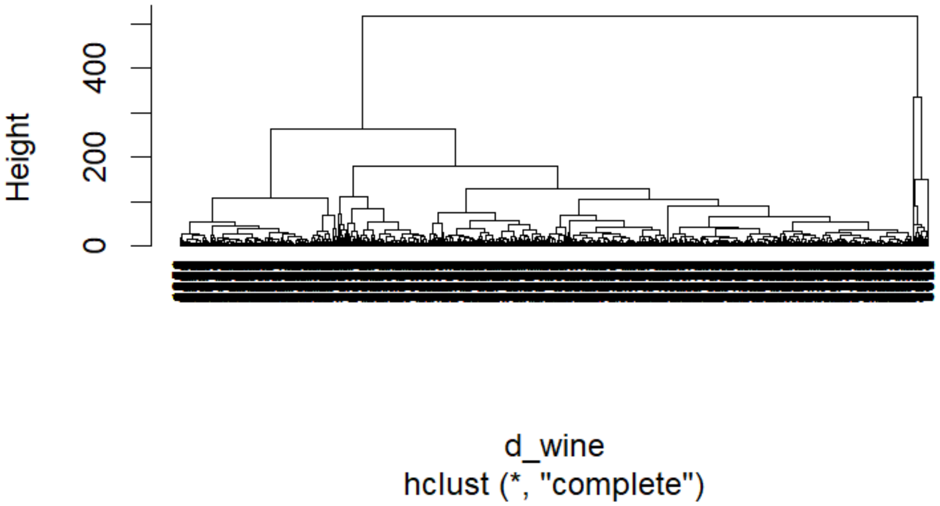
4.1.2 single

2 branches and 4898 members total, at height 194.5886



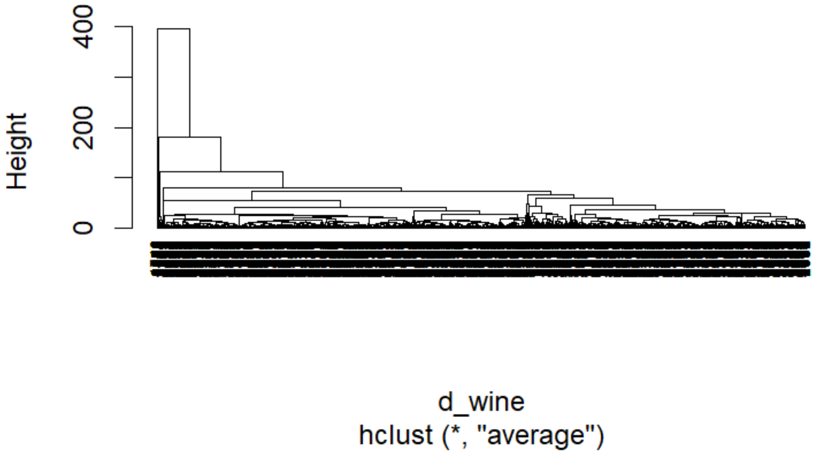
4.1.3 complete

2 branches and 4898 members total, at height 516.171



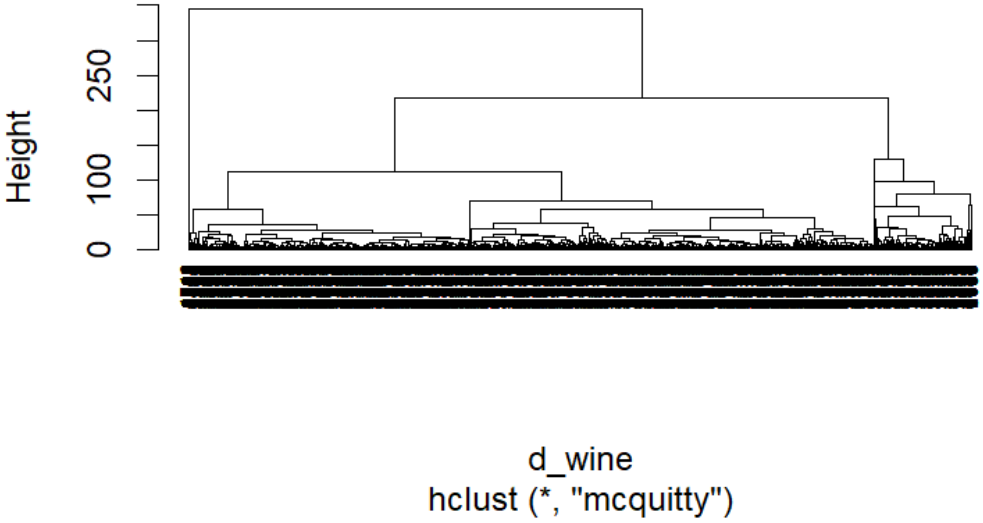
4.1.4 average

2 branches and 4898 members total, at height 394.9189



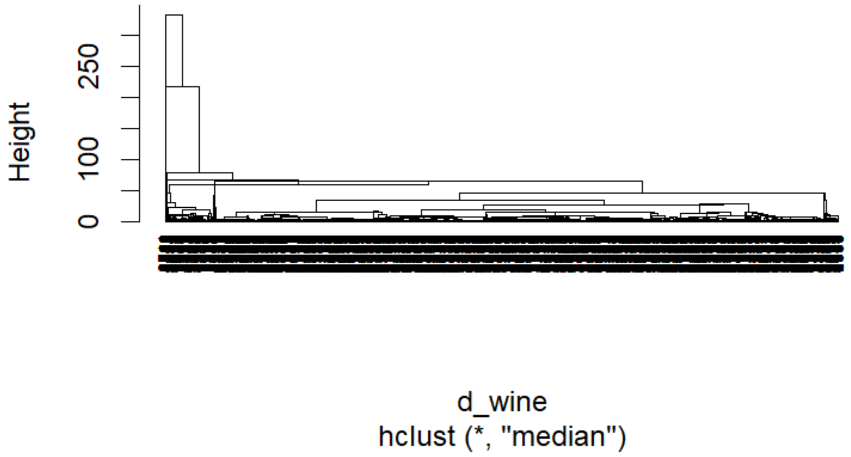
4.1.5 mcquitty

2 branches and 4898 members total, at height 344.6953



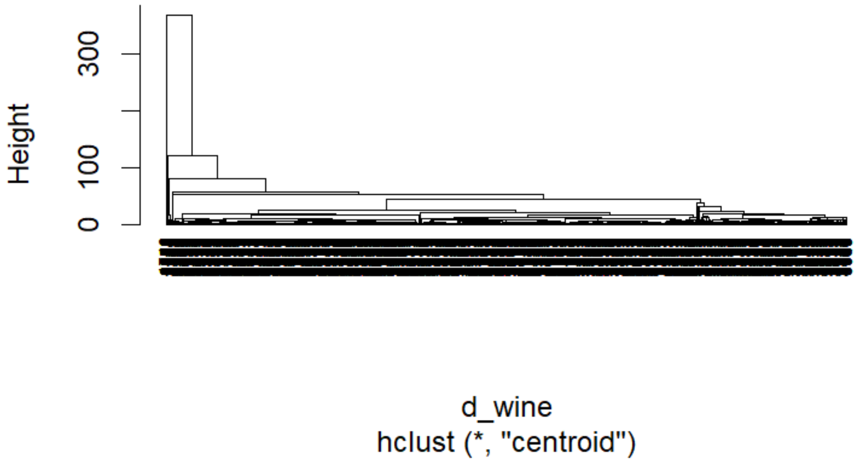
4.1.6 median

2 branches and 4898 members total, at height 333.1029



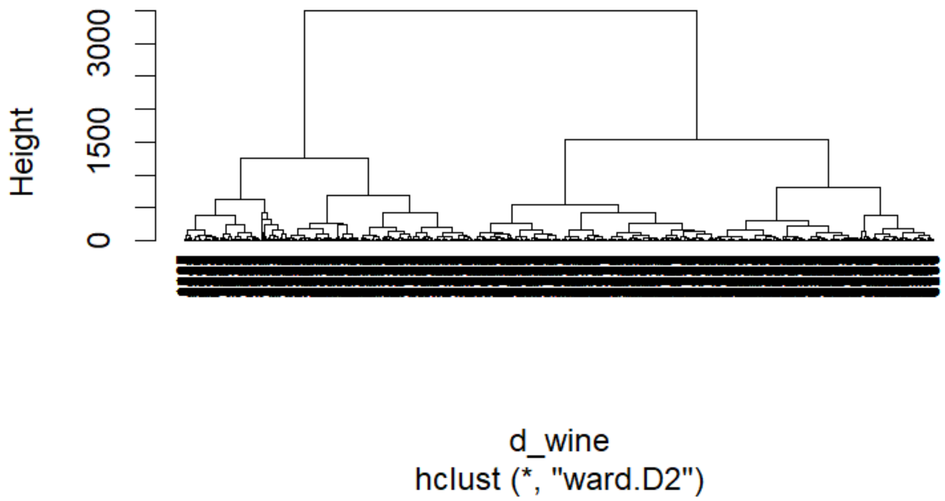
4.1.7 centroid

2 branches and 4898 members total, at height 367.7703

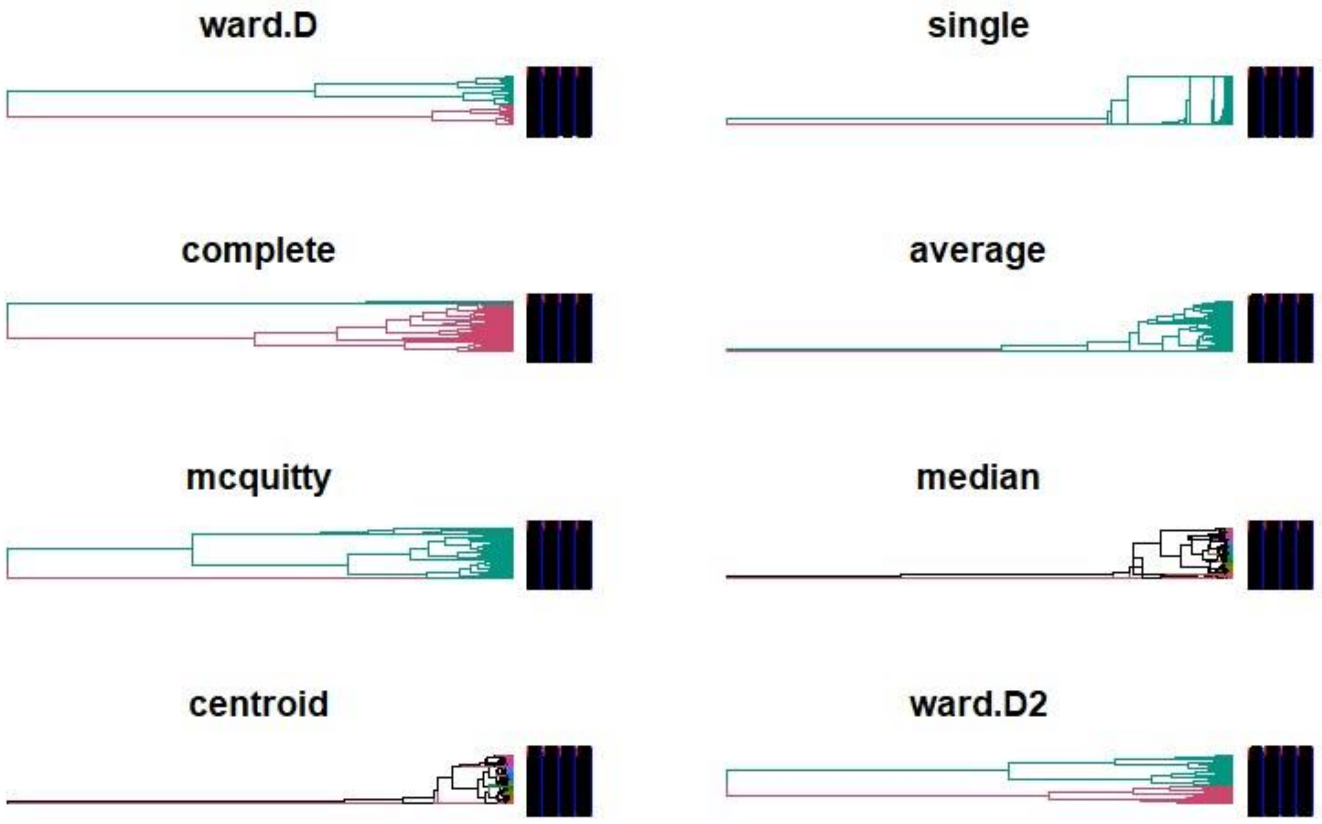


4.1.8 ward.D2

2 branches and 4898 members total, at height 3492.987



4.1.9 Dendrogram visualization with two clusters



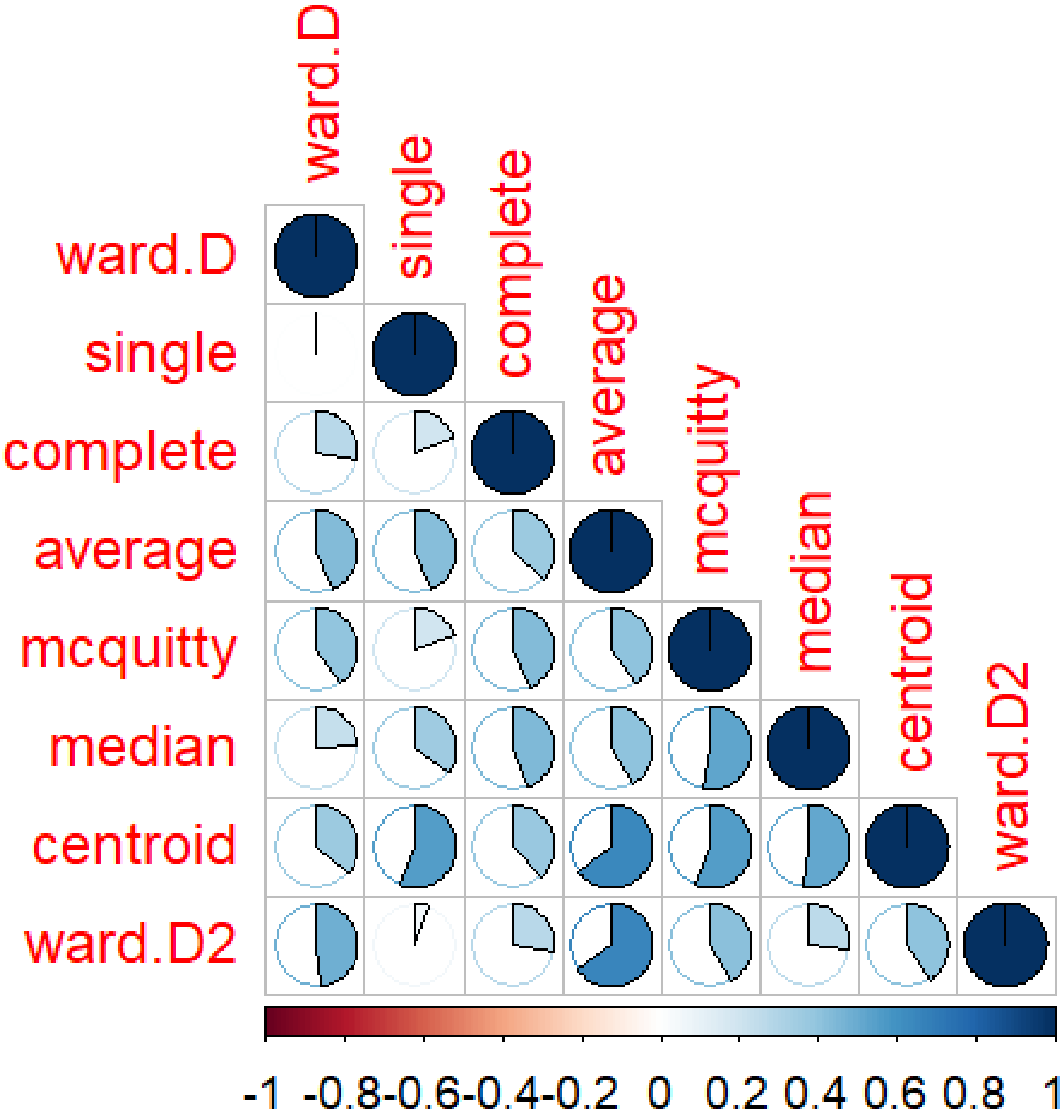
4.2 Cophenetic correlation

Cophenetic correlation coefficients are utilised to evaluate clustering solutions by measuring the extent to which a particular dendrogram maintains the pairwise distances of the original distance matrix. The cophenetic correlation coefficients are computed by comparing the original distance matrix with the cophenetic distance matrix. The cophenetic distance matrix computes the distances between clusters by taking into account the distances between the individual elements inside the clusters after they have been combined.

The cophenetic correlation coefficients are computed for each individual clustering solution, where higher values indicate better clustering performance. The cophenetic correlation coefficients may be used to easily and effectively compare different clustering algorithms. Nevertheless, this method is vulnerable to the impact of outliers. A higher score indicates a stronger likeness, whereas a lower value indicates a weaker similarity.

	ward.D	single	complete	average	mcquitty	median	centroid	ward.D2
ward.D	1.000	0.006	0.273	0.430	0.399	0.239	0.363	0.481
single	0.006	1.000	0.195	0.429	0.198	0.352	0.554	0.056
complete	0.273	0.195	1.000	0.368	0.437	0.444	0.378	0.275
average	0.430	0.429	0.368	1.000	0.404	0.408	0.645	0.654
mcquitty	0.399	0.198	0.437	0.404	1.000	0.527	0.558	0.411
median	0.239	0.352	0.444	0.408	0.527	1.000	0.514	0.264
centroid	0.363	0.554	0.378	0.645	0.558	0.514	1.000	0.407
ward.D2	0.481	0.056	0.275	0.654	0.411	0.264	0.407	1.000

The correlation matrix offers insights into the degree of similarity or dissimilarity between the outcomes of various hierarchical clustering approaches. A stronger correlation implies a higher degree of similarity among the generated clusters, whereas a weaker correlation indicates a larger level of dissimilarity. The corrplot function generates a circular correlation plot, with the size and colour of the circles indicating the magnitude and direction of the correlation coefficients.



The connection between "ward.D" and "average" is strong, as is the correlation between "centroid" and "average", showing a significant positive association. Consequently, the outcomes of these two approaches exhibit a significant correlation. The correlation between the variables "Centroid" and "single" is strong, as is the correlation between "median" and "mcquity", showing a significant positive connection. Consequently, there is a substantial correlation between the outcomes of these two approaches. The correlation between the variables "ward.D" and "single" is similar to the correlation between "ward.D2" and "single", suggesting a low level of correlation. These findings indicate that the outcomes of these two approaches are unrelated and not strongly connected.