Clustering

1. *Load and study the dataset and comment on the dataset's characteristics / structure. Do not scale data for this part.*
    - The dataset has 178 observations of 14 variables in the form of number and integer
    - The dataset has 7 NA values. A step to remove the NA values is conducted, and the final dataset has 171 observations of 14 variables.
    -

2. *Conduct exploratory data analysis for all variables to get a better understanding about each of them (numerically and visually). Comment on the results and findings.*

```
> summary(Wine)
      TYPE          ALCOHOL          MALIC            ASH          ALCALINITY
 Min.   :1.000   Min.   :11.41   Min.   :0.740   Min.   :1.360   Min.   :10.60
 1st Qu.:1.000   1st Qu.:12.35   1st Qu.:1.610   1st Qu.:2.210   1st Qu.:17.20
 Median :2.000   Median :13.05   Median :1.880   Median :2.360   Median :19.50
 Mean   :1.942   Mean   :13.00   Mean   :2.343   Mean   :2.364   Mean   :19.48
 3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.110   3rd Qu.:2.555   3rd Qu.:21.50
 Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
   MAGNESIUM         PHENOLS        FLAVANOIDS     NONFLAVANOIDS    PROANTHOCYANINS
 Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300   Min.   :0.410
 1st Qu.: 88.00   1st Qu.:1.710   1st Qu.:1.210   1st Qu.:0.2650   1st Qu.:1.245
 Median : 98.00   Median :2.300   Median :2.110   Median :0.3400   Median :1.540
 Mean   : 99.39   Mean   :2.279   Mean   :2.015   Mean   :0.3599   Mean   :1.571
 3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.825   3rd Qu.:0.4300   3rd Qu.:1.910
 Max.   :151.00   Max.   :3.880   Max.   :5.080   Max.   :0.6600   Max.   :3.580
     COLOR             HUE            DILUTION        PROLINE
 Min.   : 1.280   Min.   :0.480   Min.   :1.270   Min.   : 278.0
 1st Qu.: 3.230   1st Qu.:0.780   1st Qu.:1.945   1st Qu.: 500.0
 Median : 4.600   Median :0.960   Median :2.780   Median : 672.0
 Mean   : 5.010   Mean   :0.952   Mean   :2.616   Mean   : 735.8
 3rd Qu.: 6.115   3rd Qu.:1.120   3rd Qu.:3.185   3rd Qu.: 977.5
 Max.   :13.000   Max.   :1.450   Max.   :4.000   Max.   :1547.0
```

– TYPE: The type of wine, into one of three categories 1, 2, and 3. Type 2 has most observations with about 70, following by type 1 and type 3.

– ALCOHOL: Alcohol in %, from 11% to 14.83%, mean is 13%.

– MALIC: Malic acid, from 0.7 to 5.8, mean is 2.3.

– ASH: Ash, from 1.4 to 3.2, mean is 2.4.

– ALCALINITY: Alcalinity of ash, from 10.6 to 30, mean is 19.5.

– MAGNESIUM: Magnesium, from 62.6 to 151, mean is 99.2.

– PHENOLS: Total phenols, from 0.98 to 3.9, mean is 2.3.

– FLAVANOIDS: Flavanoids, from 0.3 to 5.1, mean is 2.0.

– NONFLAVANOIDS: Nonflavanoid phenols, from 0.1 to 0.7, mean is 3.6.

– PROANTHOCYANINS: Proanthocyanins, from 0.4 to 3.6, mean is 1.6.

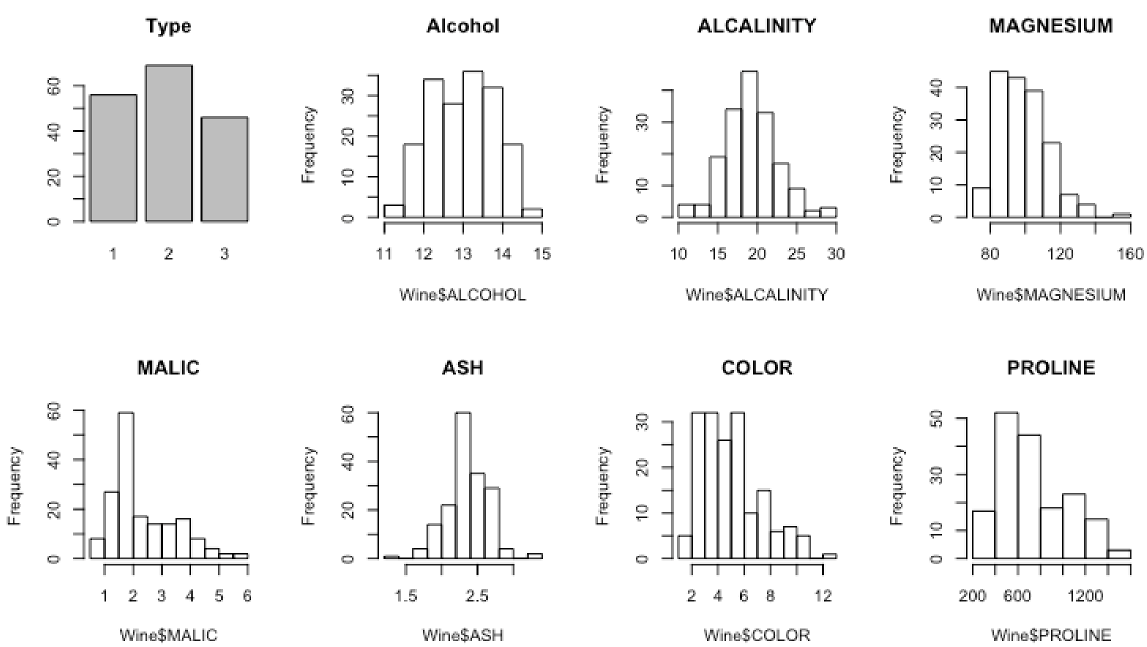
– COLOR: Color intensity, from 1.3 to 13, mean is 5.0.

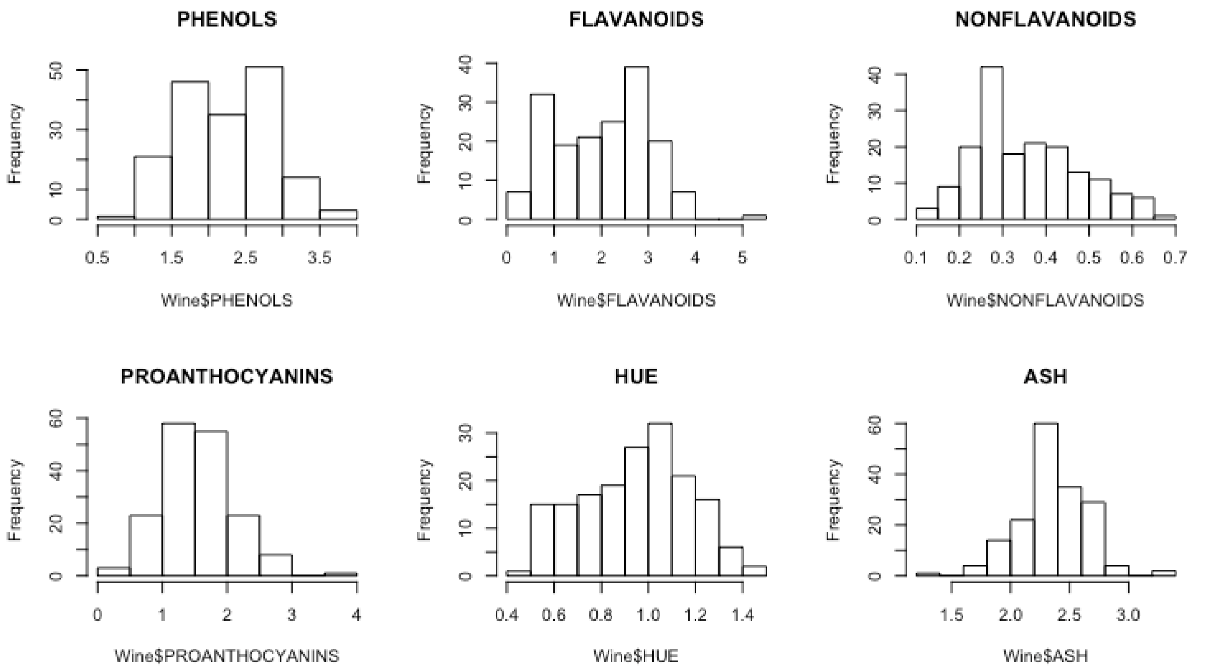
– HUE: Hue, from 0.5 to 1.5, mean is 0.95.

– DILUTION: D280/OD315 of diluted wines, from 1.3 to 4, mean is 2.6.

– PROLINE: Proline, from 278 to 1547, mean is 735.8.

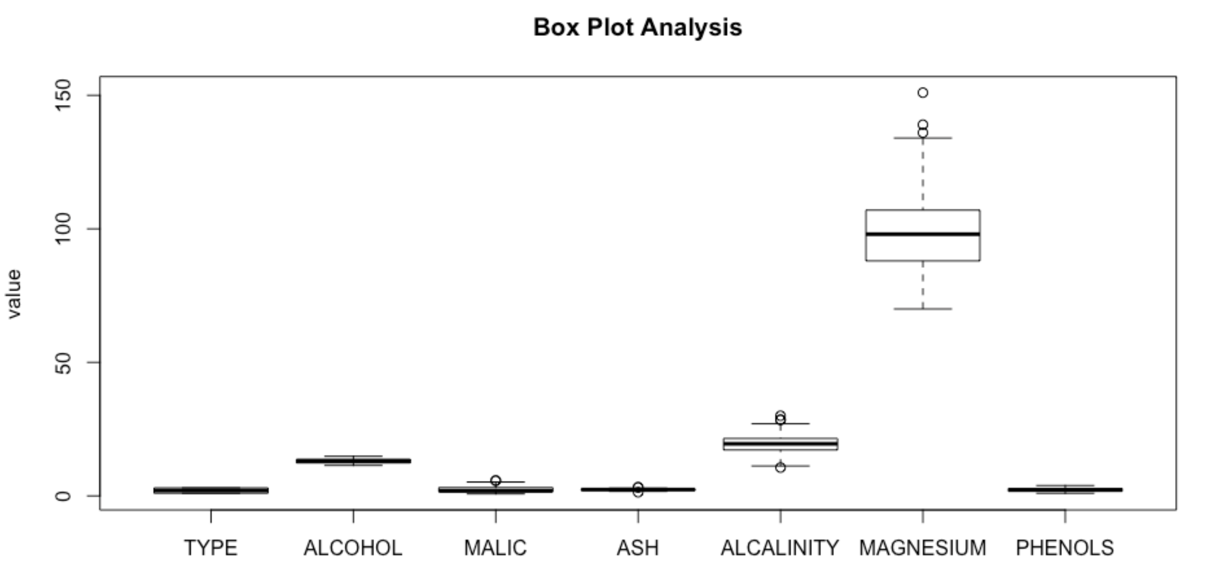
- Basic plotting of all variables:

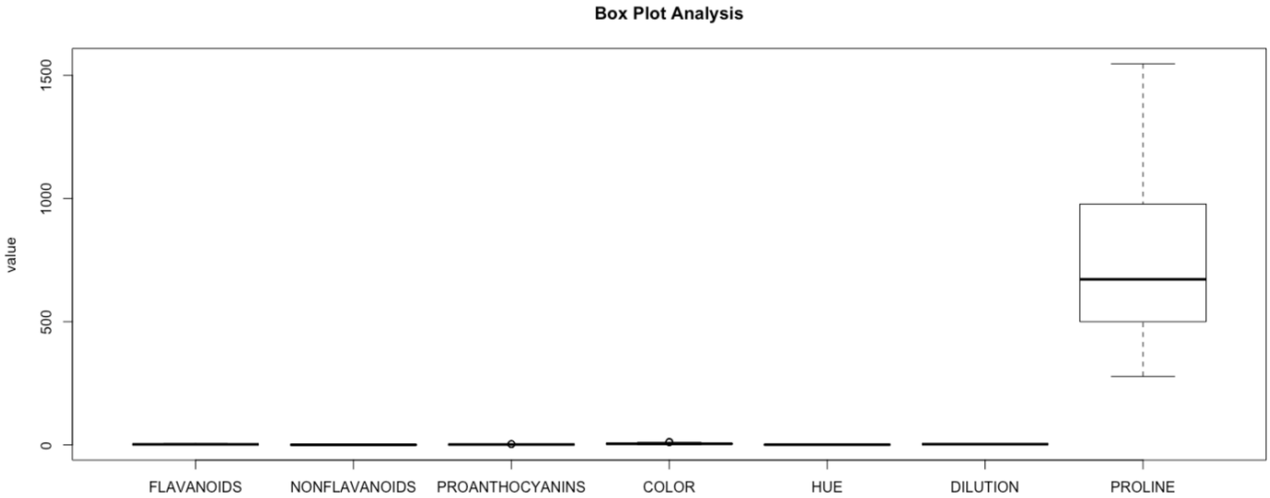




- <u>Box Plot Analysis for all variables</u>

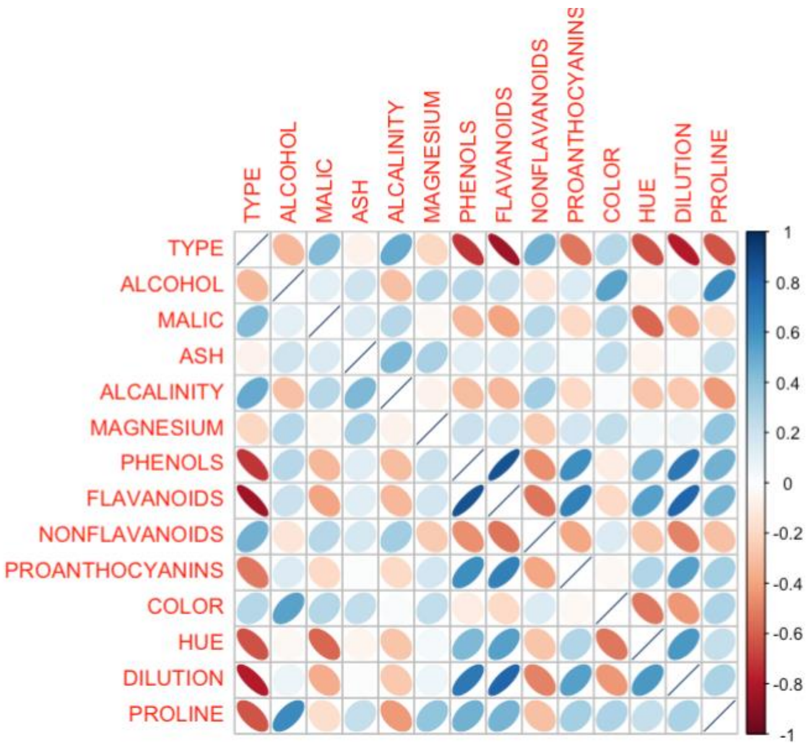
    - There are 7 variables have outliers which are MALIC, ASH, ALCALINITY, MAGNESIUM, PROANTHOCYANINS, COLOR.



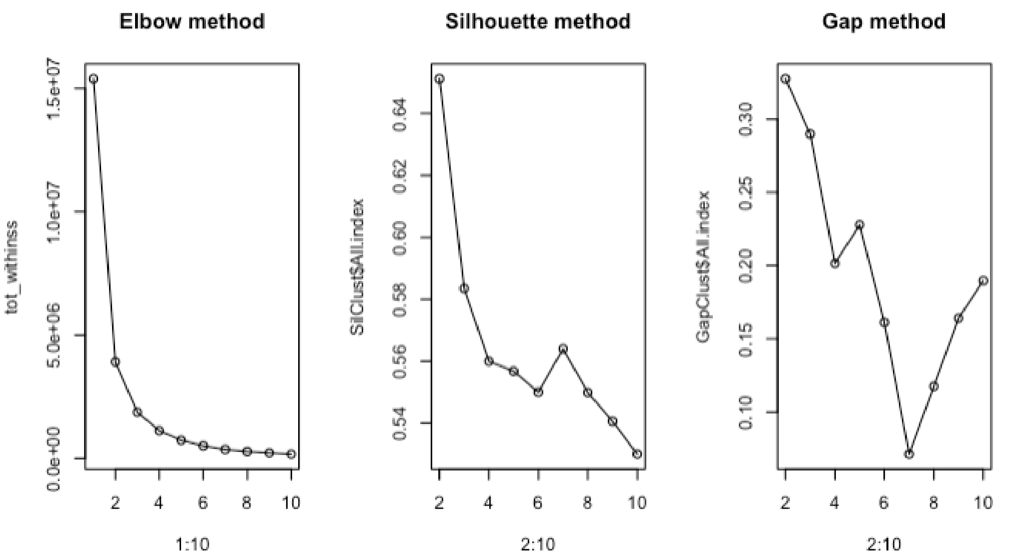


- Correlation Analysis for all variables

  - PHENOLS and FLAVANOIDS have strongest positive (absolute) relationship.
  - PHENOLS, FLAVANOIDS, DELUTION have strong negative relationship with TYPE.



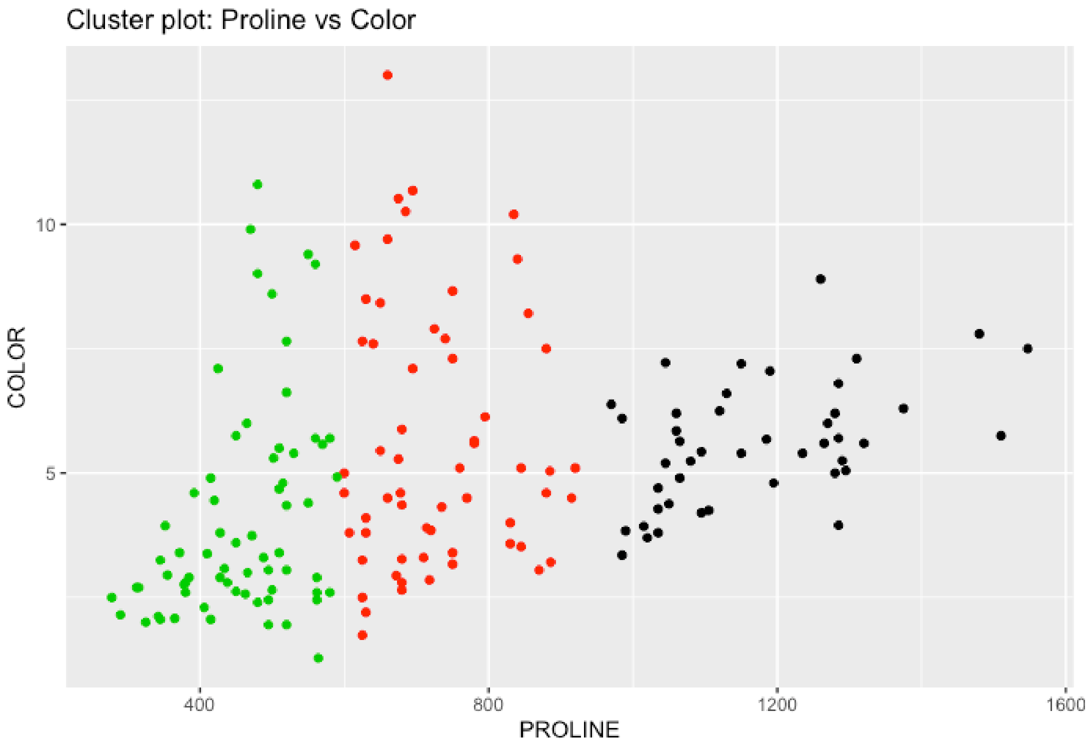
3. *Using three different methods: Elbow method, Silhouette method and Gap statistic method, how many numbers of clusters for this dataset you suggest? Visualize and interpret your findings and reasoning in detail. [Notice: If the optimal number of clusters is not clear, then make your own choice and present the reasoning for your choice].*
   - Based on Elbow method, the optimal number of clusters are 3. However, based on Silhouette method and Gap method, the optimal number of clusters are 2.
   - When we compare the total within sum of square, the value decreases significantly from over 3.92 million in 2 cluster case to 1.88 million in 3 cluster case, so in this model will apply 3 clusters as the optimal result.



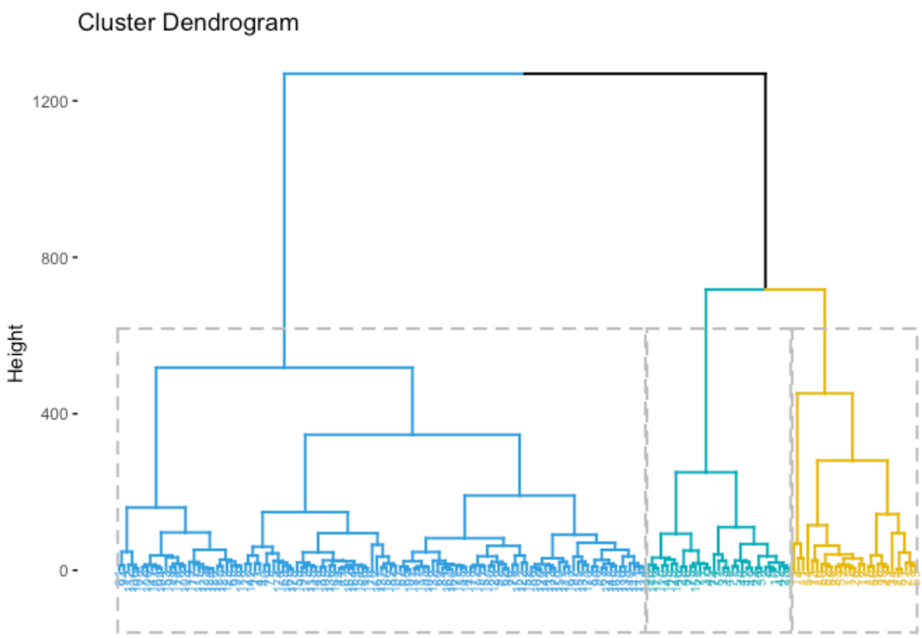
4. *Use now the optimal number of clusters for the clustering (with at least 25 random initializations [nstart]). Evaluate and discuss in detail the clustering results and performance of your model regarding the wine dataset. This includes analyzing the clustering results with respect to the characteristics of the groups and suitable visualization(s).*
   - <u>K-means result:</u> 3 clusters of sizes 44, 60, 67

- Within cluster sum of square by cluster: 933,029(1), 507,447 (2), 437,225 (3)
- Total within sum of square is 1,877,701
- Ratio of Between_SS and Total_SS is 87.8 %

- The plots of all variables are constructed to visualize the clusters. It is observed from the plots that there are clear cluster groups between PROLINE and other variables. One example of PROLINE and COLOR are plotted in order to see the distribution of the clusters.



- Hierachecal clustering: The dendrogram of 3 clusters is also conducted using “complete” method to visualize the distribution of clusters.

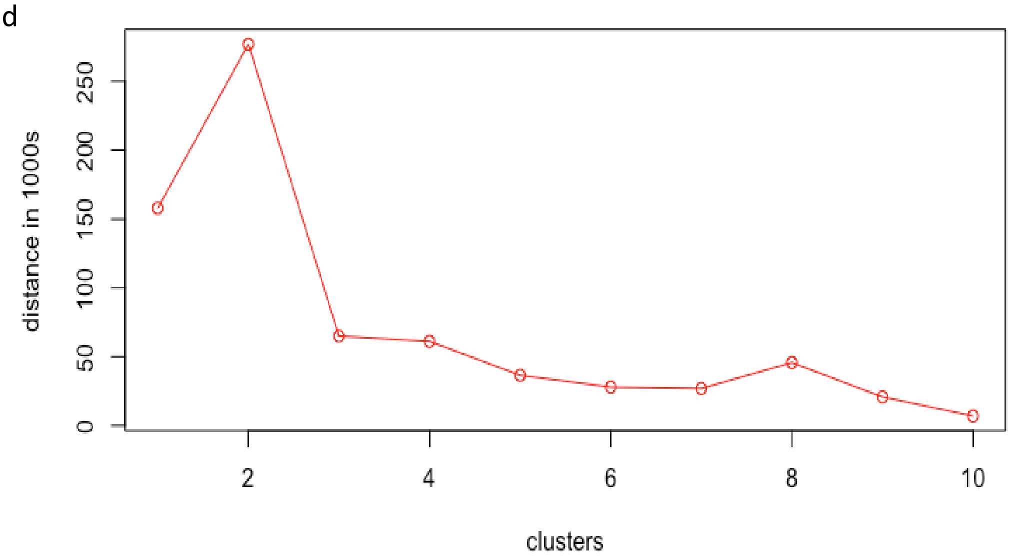


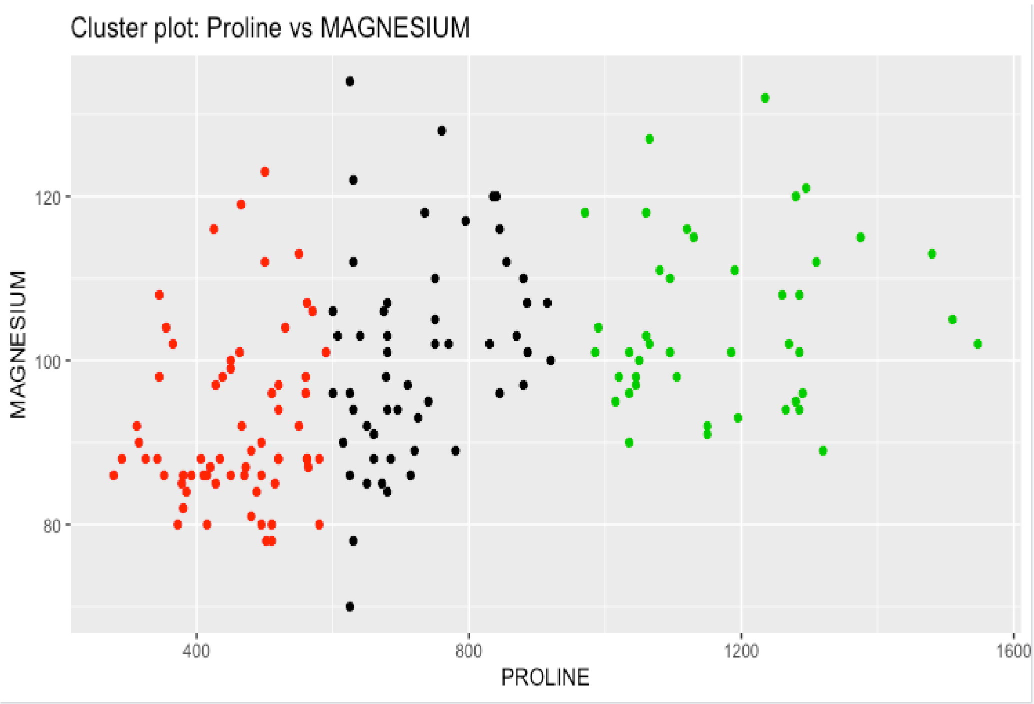
5. *Is your clustering model sensitive to outliers? Does removing the "extreme" outliers help to improve the quality of your clustering model? Justify your answer.*

There are 20 outliers in the dataset, however the model is only considered to remove 9 extreme outliers in MAGNESIUM and COLOR variables. New data sets is called "Wine_new" with 162 observations of 14 variables.

K-means model of Wine_new is constructed and then we compare the total within sum of square (TWSS) of the original model and the new model without the extreme outliers. It can be seen in the graph below that the different of TWSS is highest in 2 clusters case (above 270,000) and decreases dramatically in 3 cluster case (appx. 60,000). From 4 clusters the different of TWSS moves moderately around 20,000 and decreases compare with the number of clusters.

Finally, as the current model has 3 clustering, therefore it is not highly sensitive to outliers. However, removing the "extreme" outliers could help to improve the ratio of Between_SS and Total_SS increase from 87.8% to 88.1%. On the other hand, it should also be considered when removing outliers because it could affect to the observation size, in this case being d





6. *Give a detailed conclusion on your findings and results. Consider and explain how these findings that can help the wine company in the future.*

In conclusion, the clustering algorithms using in this report are hierarchical method and k-means methods. In order to choose the optimal k, Elbow method, Silhouette method and Gap method are used. Since the results of three methods are different, the total within sum of square method is used in order to reach the final decision. Finally, the model with 3 clusters is built to see the similarity of each group, in which PROLINE variable can be used to distinguish the distribution of the clusters.

Summary of standard deviation and average of all variables in 3 cluster members are provided in below tables:

- *Standard deviation between cluster group:*

	member	TYPE_sd	ALCOHOL_sd	MALIC_sd	ASH_sd	ALCALINITY_sd	MAGNESIUM_sd	PHENOLS_sd	FLAVANOIDS_sd	NONFLAVANOIDS_sd	PROANTHOCYANINS_sd	COLOR_sd	HUE_sd	DILUTION_sd	PROLINE_sd
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	0.447	0.651	1.22	0.302	3.17	10.5	0.567	0.850	0.120	0.565	2.20	0.238	0.675	80.6
2	2	0.795	0.703	1.19	0.283	2.77	14.9	0.586	0.971	0.137	0.536	2.59	0.227	0.774	91.4
3	3	0.151	0.471	0.563	0.196	3.07	11.5	0.368	0.406	0.0680	0.459	1.23	0.120	0.363	147.

- Average of all cluster group

	member	TYPE_avg	ALCOHOL_avg	MALIC_avg	ASH_avg	ALCALINITY_avg	MAGNESIUM_avg	PHENOLS_avg	FLAVANOIDS_avg	NONFLAVANOIDS_avg	PROANTHOCYANINS_avg	COLOR_avg	HUE_avg	DILUTION_avg	PROLINE_avg
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	2.27	12.5	2.46	2.28	20.8	92.4	2.07	1.77	0.385	1.45	4.11	0.931	2.50	458.
2	2	2.25	12.9	2.53	2.41	19.9	103.	2.10	1.58	0.387	1.46	5.60	0.885	2.37	727.
3	3	1.02	13.8	1.90	2.43	17.0	105.	2.85	2.98	0.285	1.91	5.58	1.08	3.13	1171.