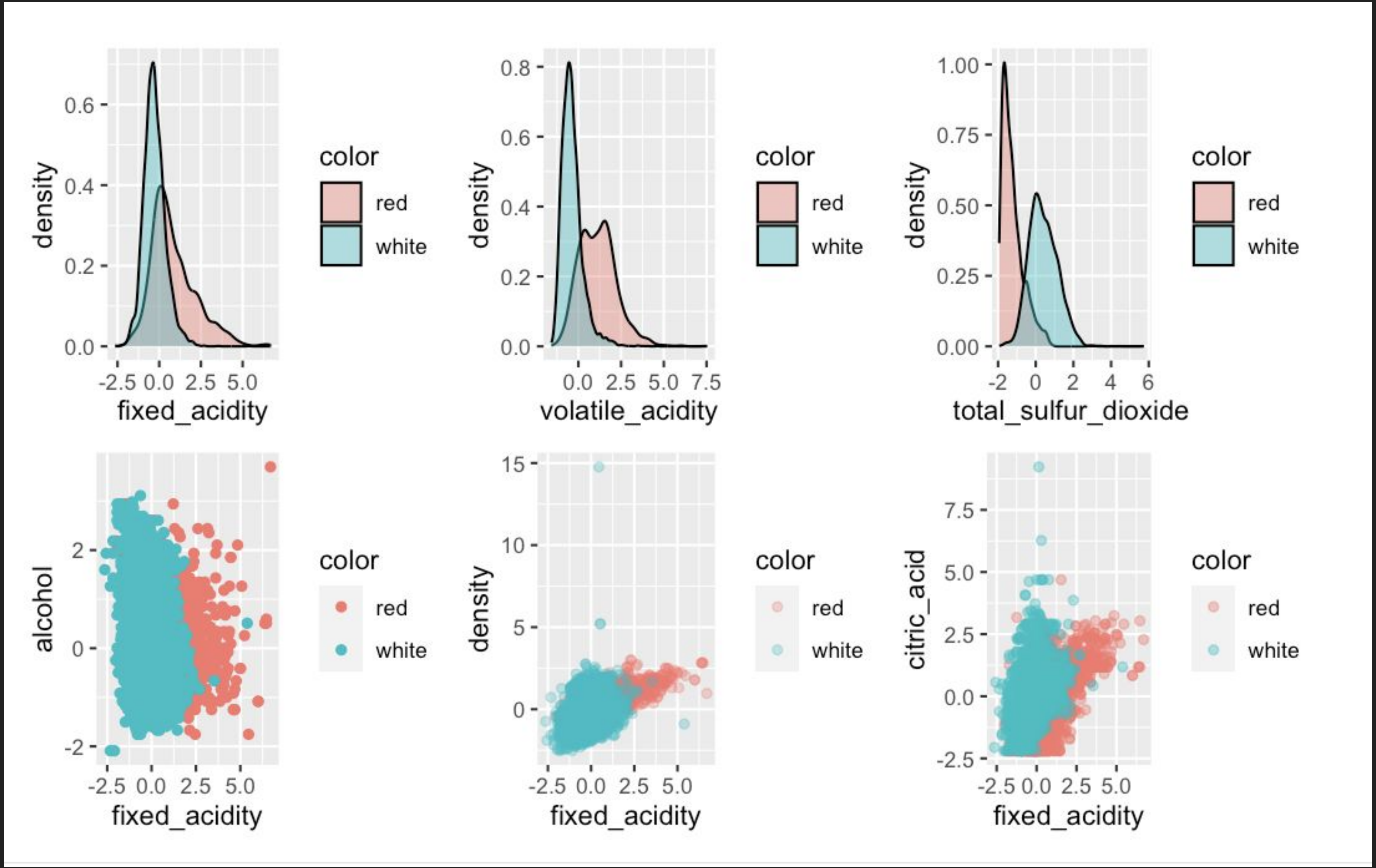
Wine Quality Clustering

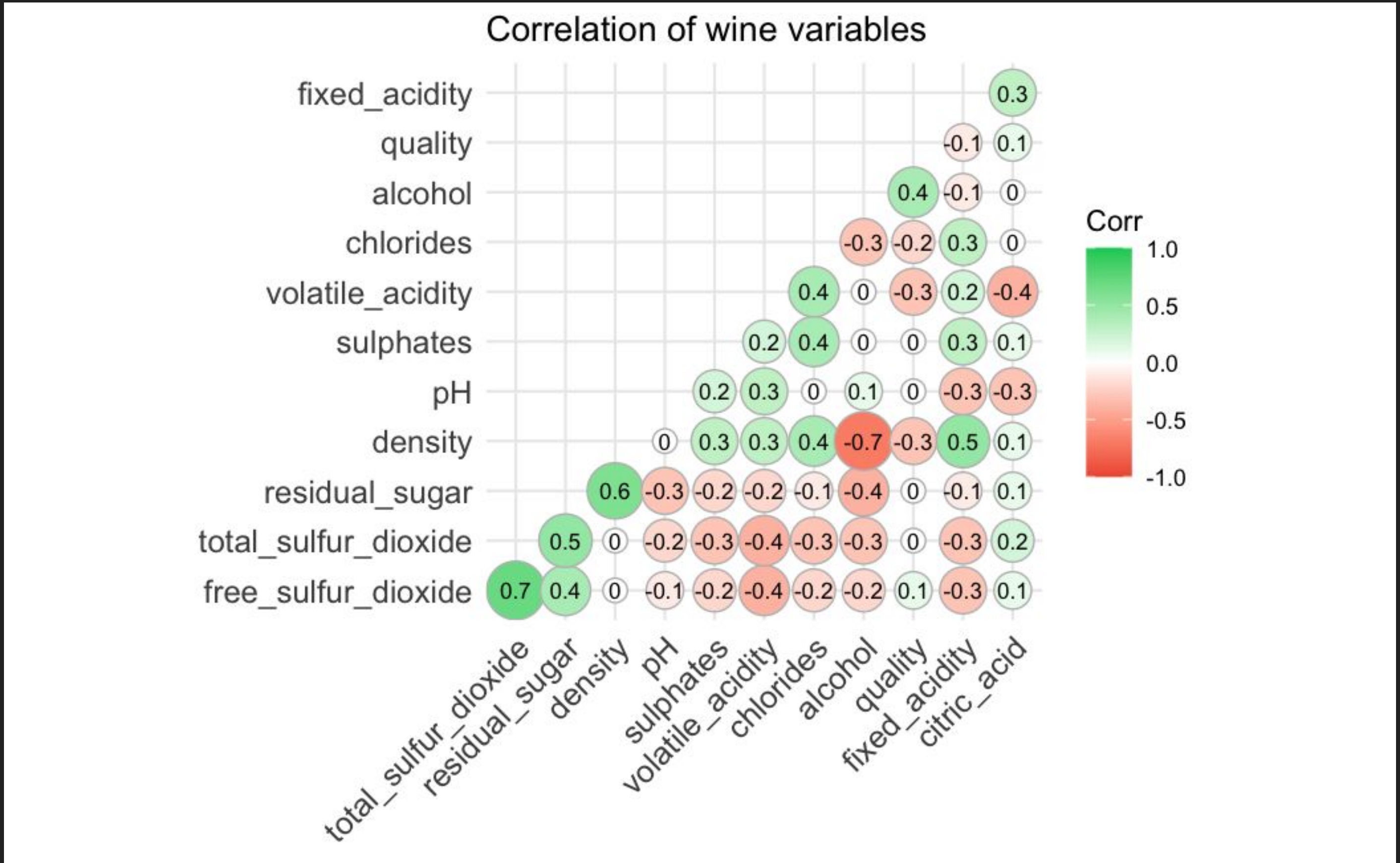
Dataset

- **6497 observations and 13 variables**
- **Fixed_acidity:** corresponds to the set of low volatility organic acids
- **Volatile_acidity:** acids extracted from the sample by means of a distillation process
- **Citric_acid:** amount of acidity to complement a specific flavor
- **Residual_sugar:** sugars (natural juices) that are left in wine after fermentation
- **Chlorides:** relative to overall taste caused by salt
- **Free_sulfur_dioxide:** a preservative to prevent oxidation and microbial spoilage
- **Total_sulfur_dioxide:** the overall amount of sulfites that have reacted
- **Density:** the amount of wine must (fruit juice) used
- **pH:** level of pH (0: most acidic, 14: most basic)
- **Sulphates:** a chemical compound that occurs naturally at low levels during the process of wine fermentation
- **Alcohol:** percentage of Alcohol
- **Quality:** rating from 3-9
- **Color:** red or white

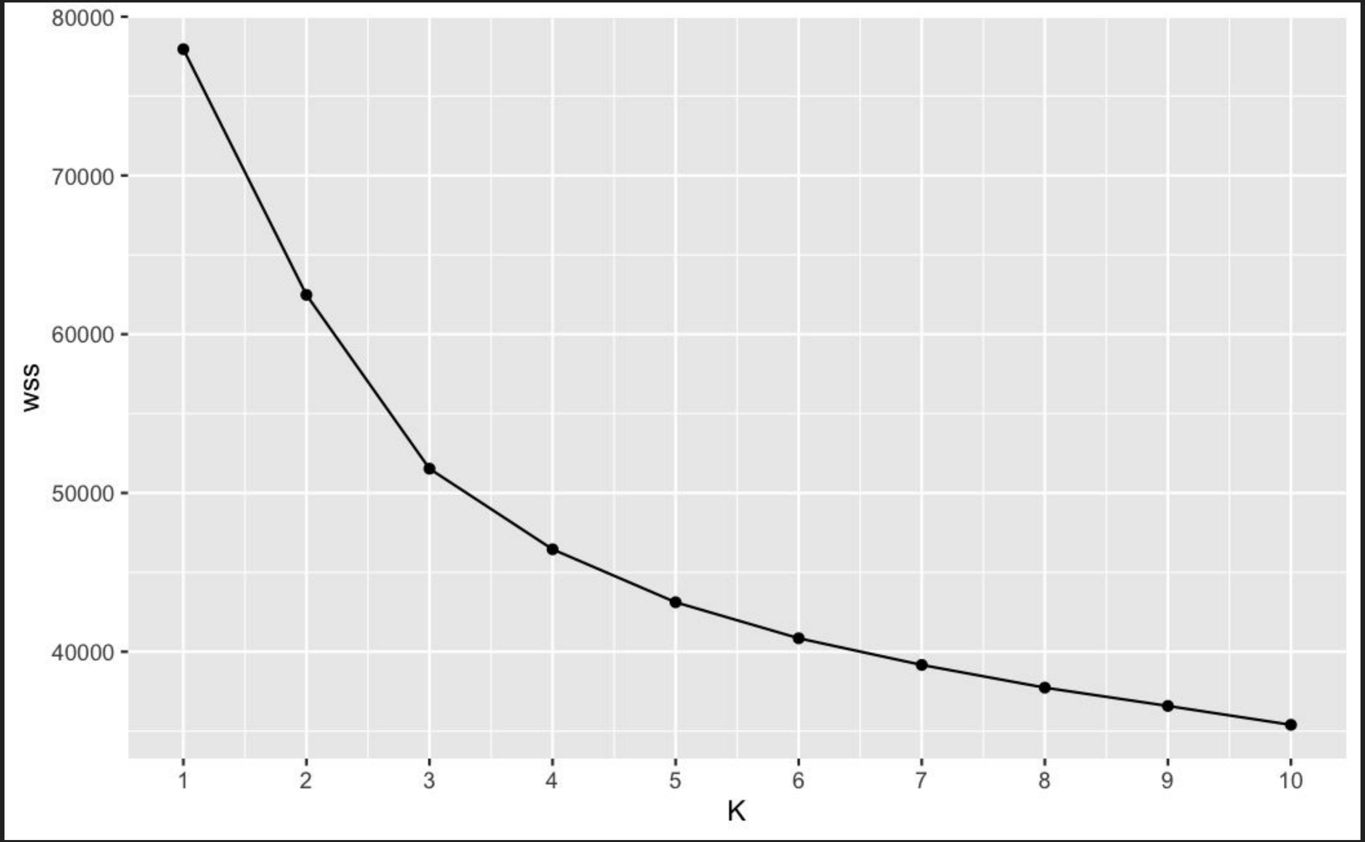
EDA


density
fixed_acidity
color
red
white
density
volatile_acidity
color
red
white
density
total_sulfur_dioxide
color
red
white
alcohol
fixed_acidity
color
red
white
density
fixed_acidity
color
red
white
citric_acid
fixed_acidity
color
red
white

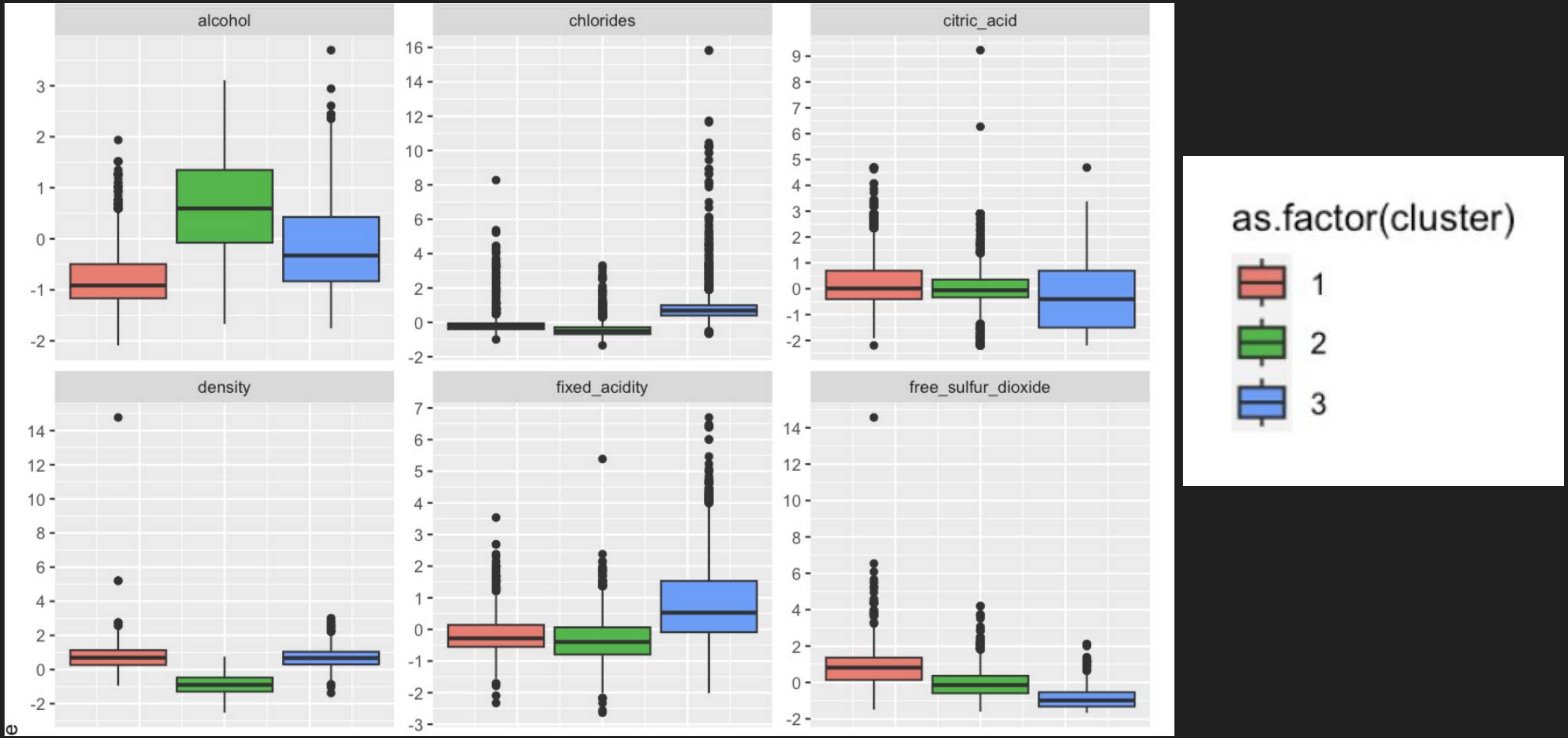
Exploratory Analysis (Correlation Matrix)



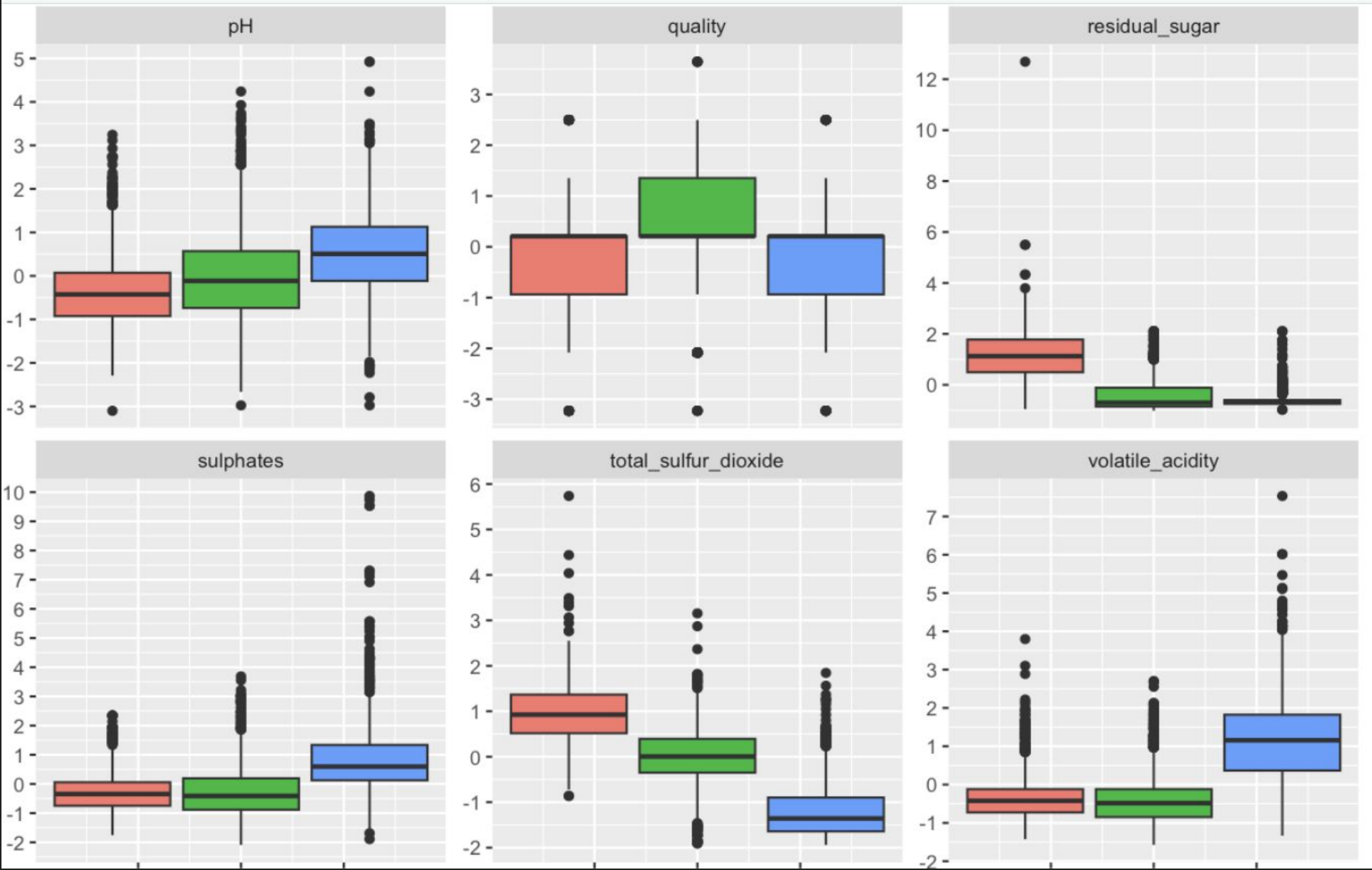
Elbow Method

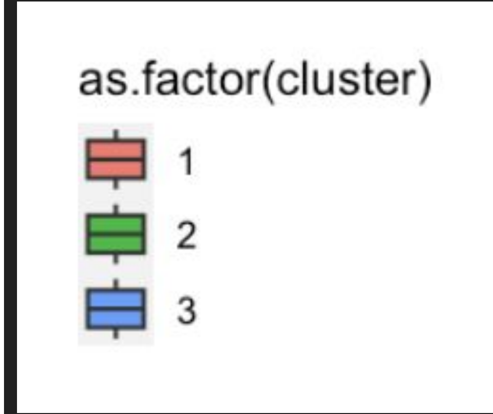


K-Means: 3 Clusters



K-Means with 3 Clusters





K-Means-3 clustering comments

- Cluster 1: Sweet White Wines (Rose)
  - Low alcohol,
  - high in sulfur dioxide
  - high in residual sugar
- Cluster 2: Natural Red Wines (Merlot)
  - High in alcohol
  - low in density
- Cluster 3: Dry Wines (Combining)
  - Slightly higher fixed acidity
  - lowest sulfur dioxide
  - low residual sugar

PCA: 3 Clusters

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	1.7440	1.6278	1.2812	1.03374	0.91679	0.81265	0.75088	0.7183	0.6770	0.54682	0.47706	0.18107
Proportion of Variance	0.2535	0.2208	0.1368	0.08905	0.07004	0.05503	0.04699	0.0430	0.0382	0.02492	0.01897	0.00273
Cumulative Proportion	0.2535	0.4743	0.6111	0.70013	0.77017	0.82520	0.87219	0.9152	0.9534	0.97830	0.99727	1.00000

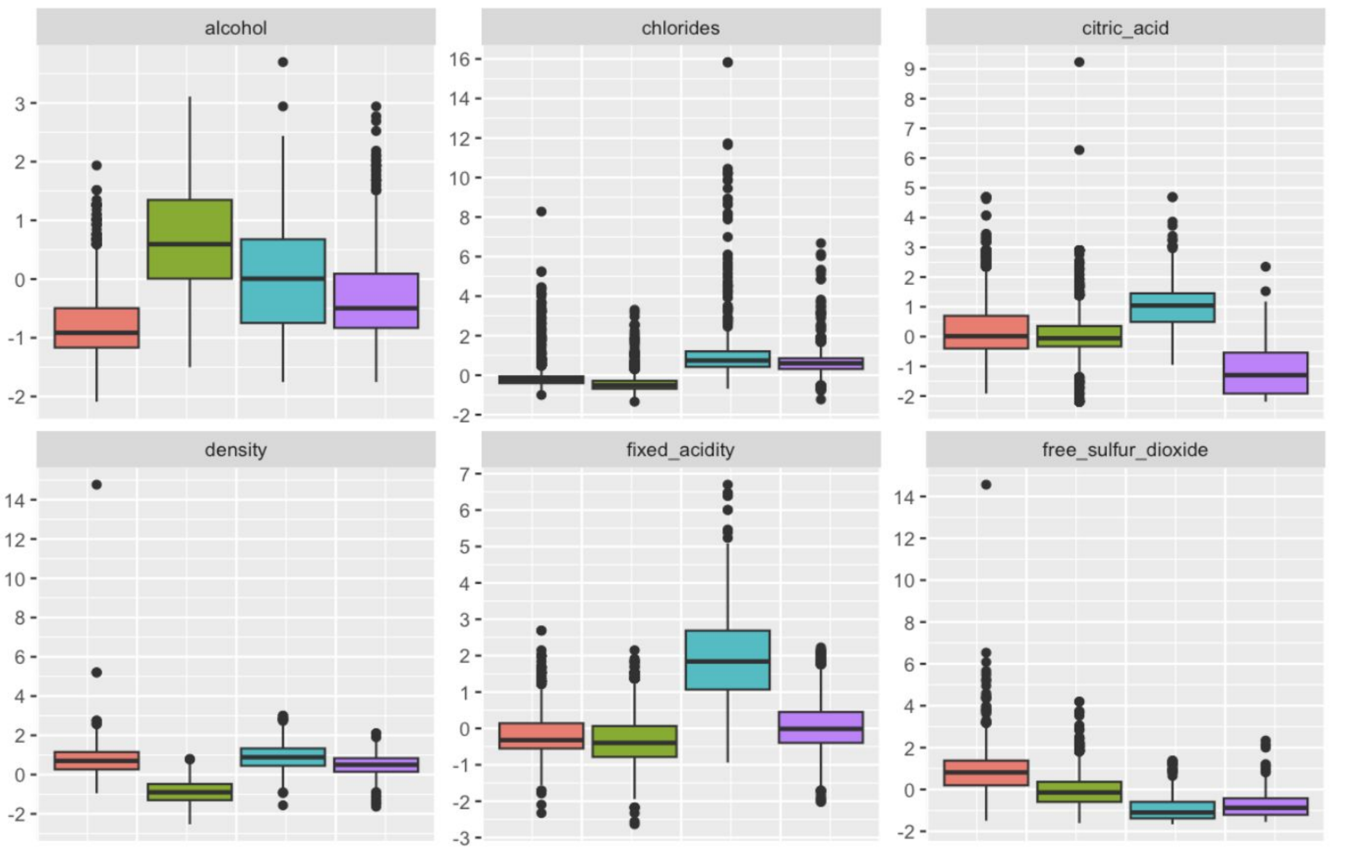
PCA with 3 clusters

	PC1	PC2	PC3	PC4
	Measure of gaseous components	Taste Intensity	Measure of acidity	
fixed_acidity	-0.25692873	0.2618431	-0.46748619	0.14396377
volatile_acidity	-0.39493118	0.1051983	0.27968932	0.08005785
citric_acid	0.14646061	0.1440935	-0.58807557	-0.05551036
residual_sugar	0.31890519	0.3425850	0.07550170	-0.11245623
chlorides	-0.31344994	0.2697701	-0.04676921	-0.16529004
free_sulfur_dioxide	0.42269137	0.1111788	0.09899801	-0.30330631
total_sulfur_dioxide	0.47441968	0.1439475	0.10128143	-0.13223199
density	-0.09243753	0.5549205	0.05156338	-0.15057853
pH	-0.20806957	-0.1529219	0.40678741	-0.47147768
sulphates	-0.29985192	0.1196342	-0.16869128	-0.58801992
alcohol	-0.05892408	-0.4927275	-0.21293142	-0.08003179
quality	0.08747571	-0.2966009	-0.29583773	-0.47243936

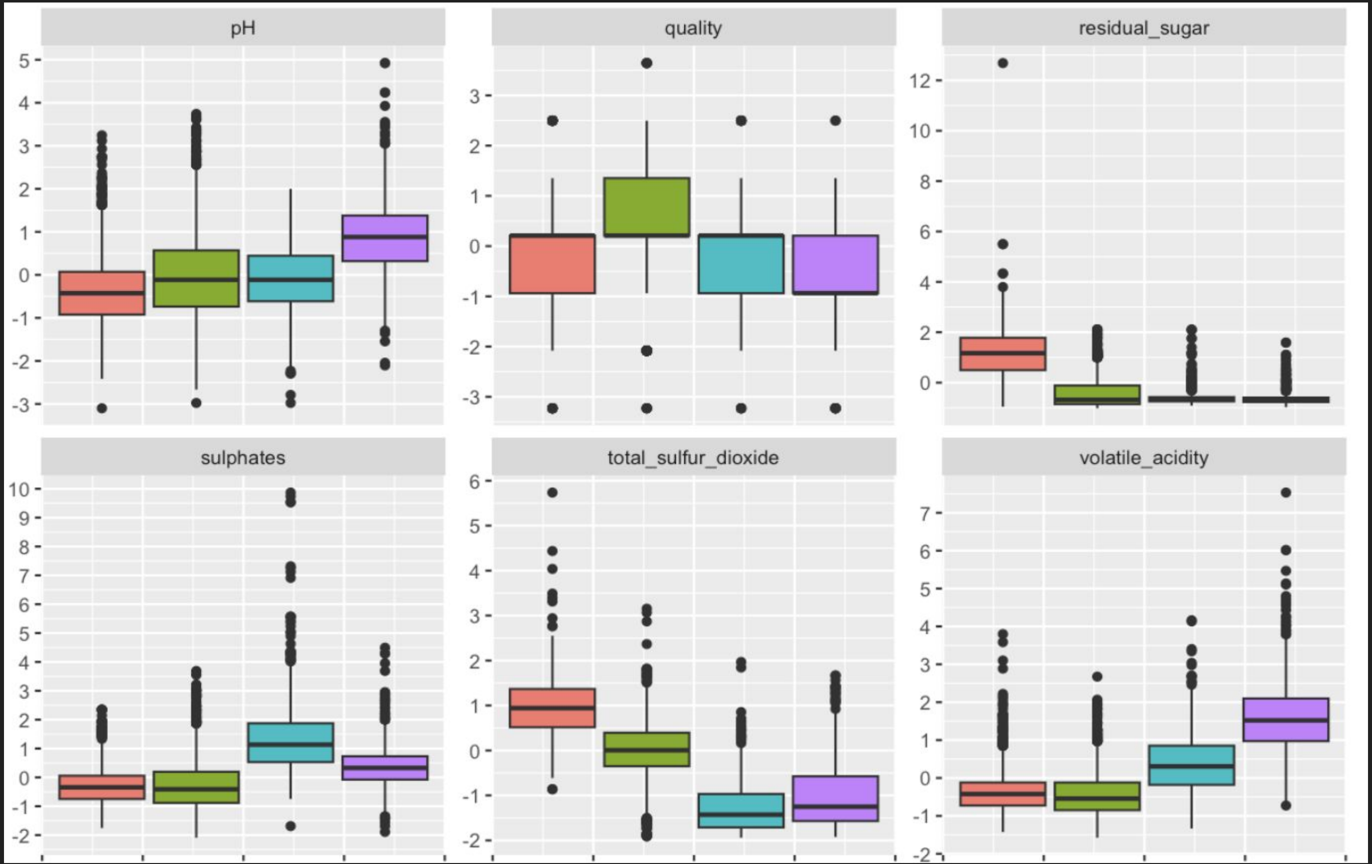
3D Plot of PCA

```
.fittedPC1 .fittedPC2 .fittedPC3
  1.511924   1.453781   0.298750
.fittedPC1 .fittedPC2 .fittedPC3
 0.4063623 -1.3408644 -0.2297762
 .fittedPC1  .fittedPC2  .fittedPC3
-2.56960554  0.63945912  0.04919818
```

K-means Clustering with 4 clusters



K-means Clustering with 4 clusters



KMeans-4 clustering comments

- Cluster 1: Sweet White Wines
  - Low Alcohol percentage, high residual sugar, high total_sulfur dioxide
- Cluster 2: Dry Red Wines
  - Highest in alcohol percentage, lowest density
- Cluster 3: Top Shelf White Wines (Chardonnay or Sauvignon Blanc)
  - highest in citric acid, fixed_acidity is high, sulfates is high
- Cluster 4: Low Shelf Red Wines
  - Lowest in citric acid, highest in volatile acidity, highest pH, low total sulfur dioxide

PCA with 4 clusters

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	1.7440	1.6278	1.2812	1.03374	0.91679	0.81265	0.75088	0.7183	0.6770	0.54682	0.47706	0.18107
Proportion of Variance	0.2535	0.2208	0.1368	0.08905	0.07004	0.05503	0.04699	0.0430	0.0382	0.02492	0.01897	0.00273
Cumulative Proportion	0.2535	0.4743	0.6111	0.70013	0.77017	0.82520	0.87219	0.9152	0.9534	0.97830	0.99727	1.00000

PCA with 4 clusters

	PC1	PC2	PC3	PC4
	Measure of Gaseous components	Taste Intensity	Measure of Acidity	
fixed_acidity	-0.25692873	0.2618431	-0.46748619	0.14396377
volatile_acidity	-0.39493118	0.1051983	0.27968932	0.08005785
citric_acid	0.14646061	0.1440935	-0.58807557	-0.05551036
residual_sugar	0.31890519	0.3425850	0.07550170	-0.11245623
chlorides	-0.31344994	0.2697701	-0.04676921	-0.16529004
free_sulfur_dioxide	0.42269137	0.1111788	0.09899801	-0.30330631
total_sulfur_dioxide	0.47441968	0.1439475	0.10128143	-0.13223199
density	-0.09243753	0.5549205	0.05156338	-0.15057853
pH	-0.20806957	-0.1529219	0.40678741	-0.47147768
sulphates	-0.29985192	0.1196342	0.16869128	-0.58801992
alcohol	-0.05892408	-0.4927275	-0.21293142	-0.08003179
quality	0.08747571	-0.2966009	-0.29583773	-0.47243936

3D Plot of PCA

```
.fittedPC1 .fittedPC2 .fittedPC3
 1.5376534  1.4540786  0.2998627
.fittedPC1 .fittedPC2 .fittedPC3
 0.4355544 -1.3629507 -0.2815429
.fittedPC1 .fittedPC2 .fittedPC3
 -2.518977   1.212844  -1.984068
.fittedPC1 .fittedPC2 .fittedPC3
-2.4033036  0.2238458  1.4197218
```

Differences

3 Clusters	4 clusters
● Cluster 3 was a mix-mash of things ● Generalized red wines more ● Focused more on taste profile	● With four clusters, it addressed the confusion into better-defined clusters ● Made distinguishable separations of red white and white wine ● Added quality into the cluster descriptions

Exploration with GMM

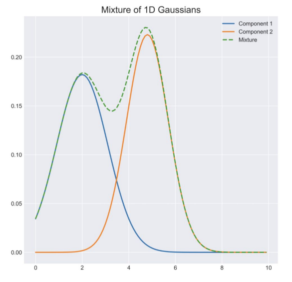
GMM Introduction

- GMM= Gaussian Mixture Model
- Combined multivariate Gaussians
- Sum of weighted Gaussian where the weights are the prior probabilities for the respective Gaussian
- Minimize weighted distances(by Variance.)
  - Note Kmeans tries to minimize just distances
- Probabilistic
  - Note: Kmeans is a direct assignment.
  - Note: GMM will have probabilities that the observation belongs to cluster(s) k
- Noticed bimodal densities
- GMM will allow us to represent the distribution better

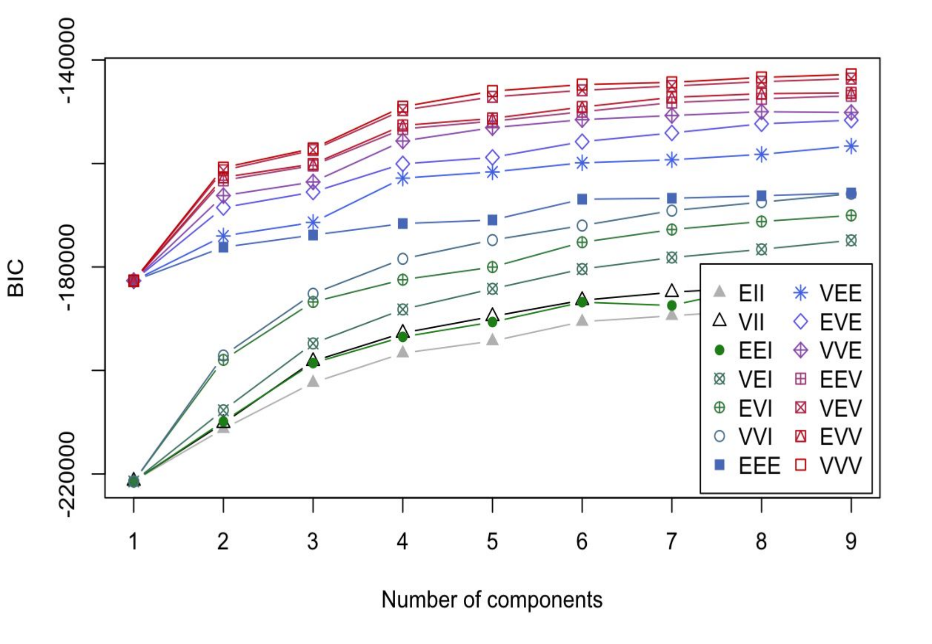
$$p(x) = \pi_0 N(x|\mu_0, \Sigma_0) + \pi_1 N(x|\mu_1, \Sigma_1) + \ldots + \pi_k N(x|\mu_k, \Sigma_k)$$

$$\text{where } \sum_{i=0}^{k} \pi_i = 1$$

$$\boxed{p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k)}$$

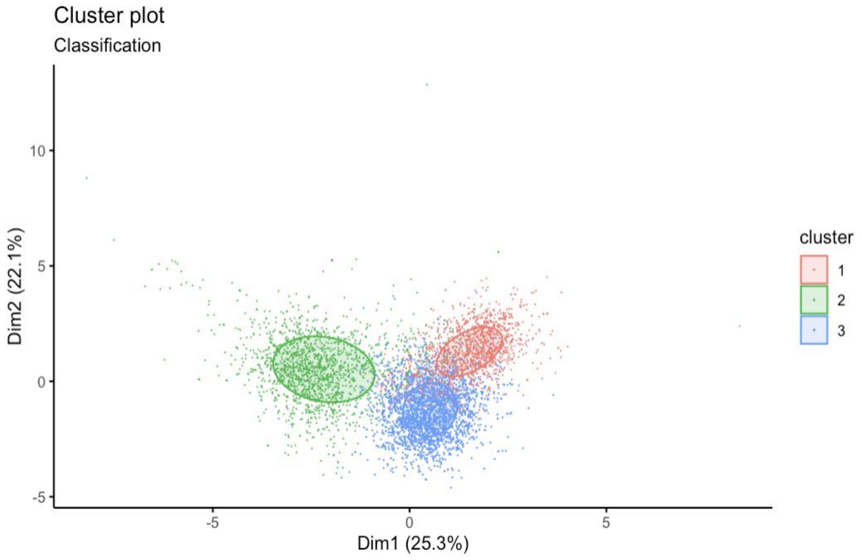


Best Model based on BIC(Bayesian Information Criterion)

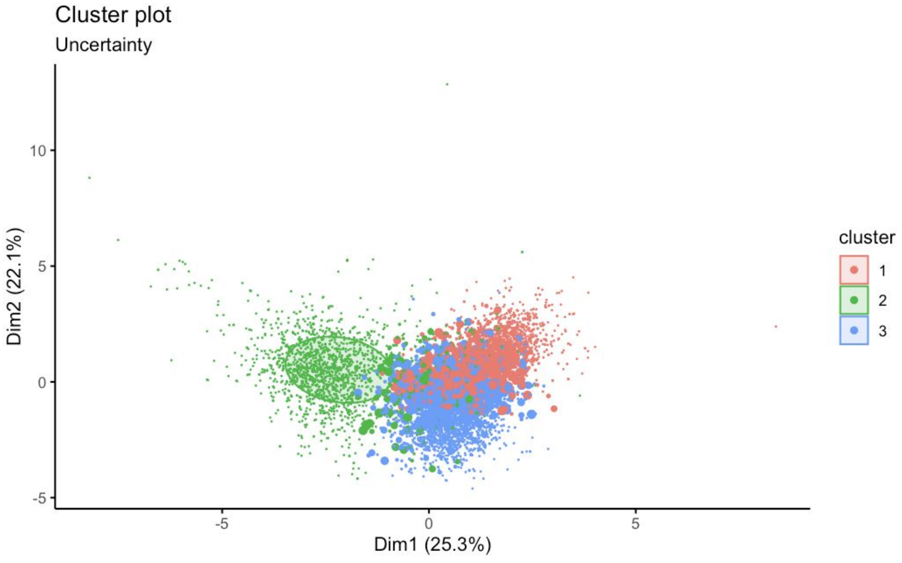


Model	$\boldsymbol{\Sigma}_k$	Distribution	Volume	Shape	Orientation
EII	$\lambda \boldsymbol{I}$	Spherical	Equal	Equal	—
VII	$\lambda_k \boldsymbol{I}$	Spherical	Variable	Equal	—
EEI	$\lambda \boldsymbol{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k \boldsymbol{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda \boldsymbol{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k \boldsymbol{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k \boldsymbol{D}\boldsymbol{A}_k\boldsymbol{D}^\top$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k \boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	Variable	Variable	Variable

3 clusters

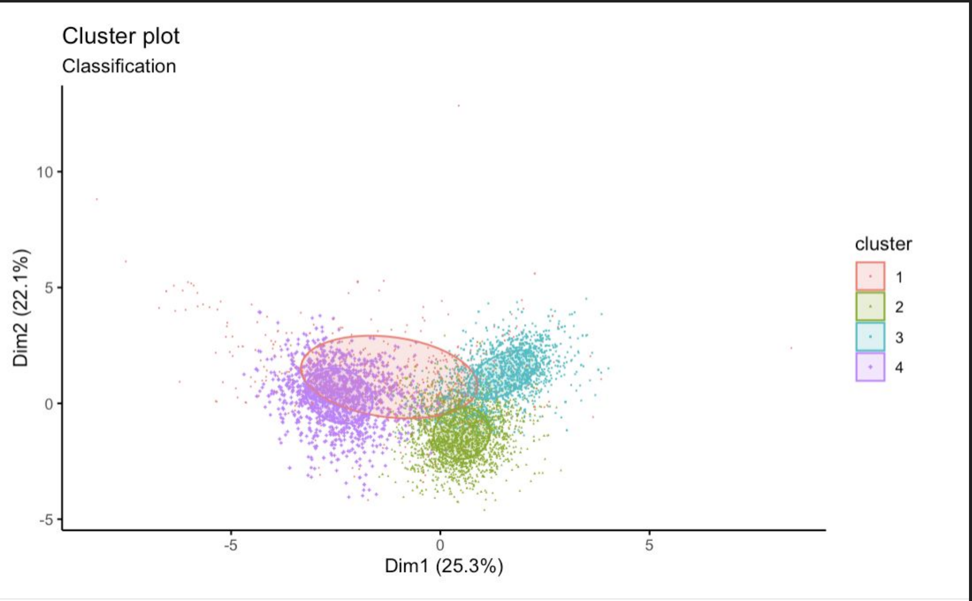


Cluster plot
Classification
Dim2 (22.1%)
Dim1 (25.3%)
10
5
0
-5
-5
0
5
cluster
1
2
3

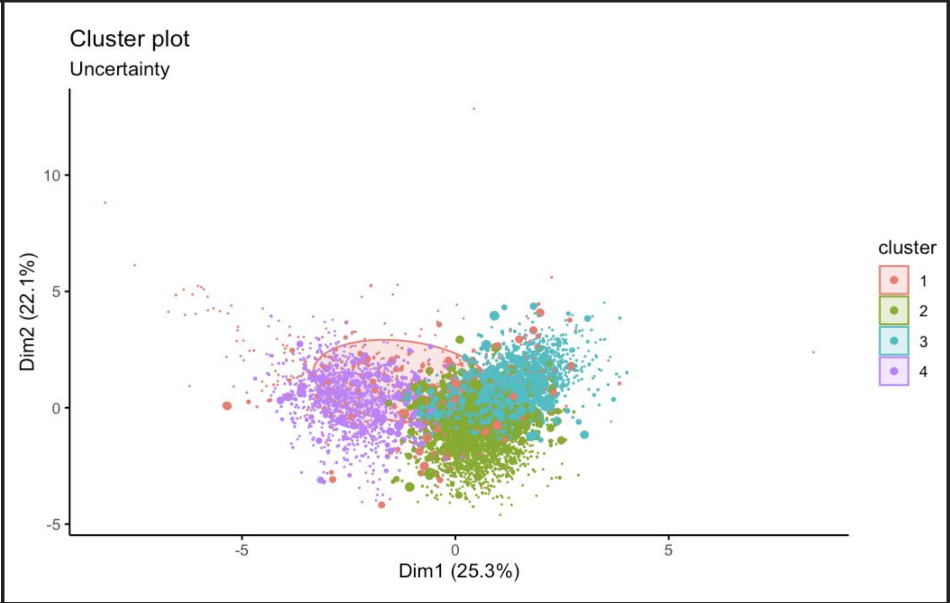


Cluster plot
Uncertainty
Dim2 (22.1%)
Dim1 (25.3%)
10
5
0
-5
-5
0
5
cluster
1
2
3

4 clusters

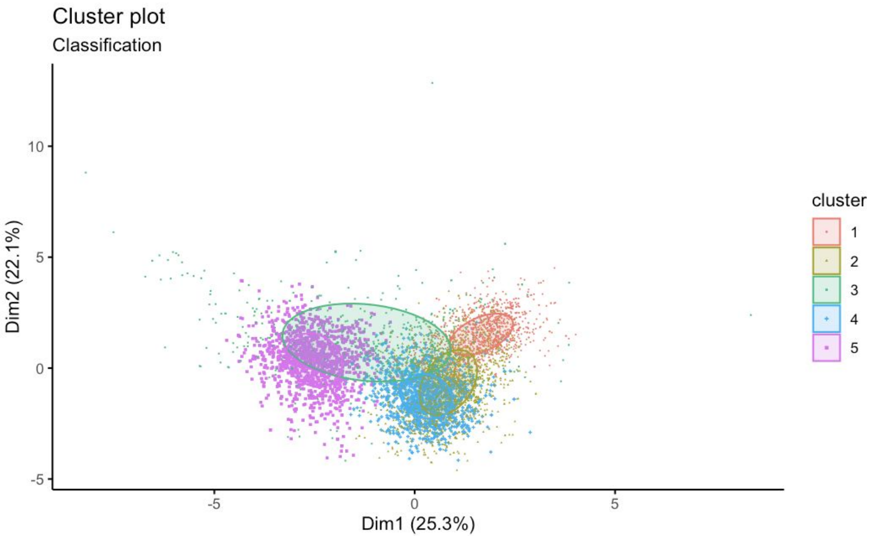


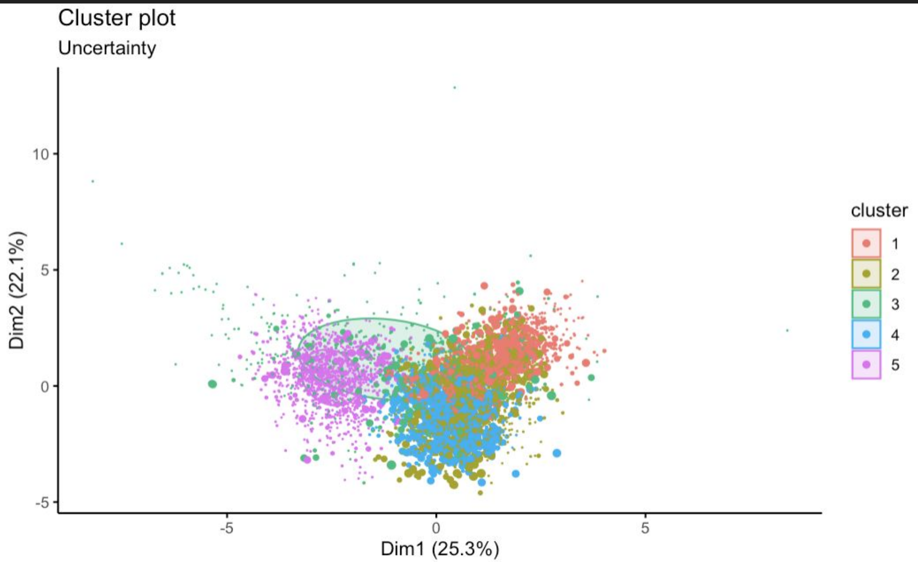
Cluster plot
Classification
Dim2 (22.1%)
Dim1 (25.3%)
10
5
0
-5
-5
0
5
cluster
1
2
3
4



Cluster plot
Uncertainty
Dim2 (22.1%)
Dim1 (25.3%)
10
5
0
-5
-5
0
5
cluster
1
2
3
4

5 clusters


Cluster plot
Classification
Dim2 (22.1%)
Dim1 (25.3%)
10
5
0
-5
-5
0
5
cluster
1
2
3
4
5


Cluster plot
Uncertainty
Dim2 (22.1%)
Dim1 (25.3%)
10
5
0
-5
-5
0
5
cluster
1
2
3
4
5

Conclusions on GMM

- As the number of clusters go up, we found there was more uncertainty among the intersectional points
  - 3 was subjectively considered to be the best at capturing the distinctions
- Performance
  - Not all clusters might have assumed Gaussian Distributions = more uncertainty
    - Density and pH vs volatile_acidity