Final Project: Wine Data Unsupervised Learning

Pavlenko A.

Goal:

Provide an Unsupervised Learning problem resolution to Wine Data set to perform EDA and model analysis.

Methods:

- The clustering performed with k-means approach
- Dimension Reduction by using PCA.

Table of Contents:

Data:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1
6	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6
7	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8
8	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5
9	8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0

```
1 #Size of Dataset:
2 df.shape
```

```
(4898, 11)
```

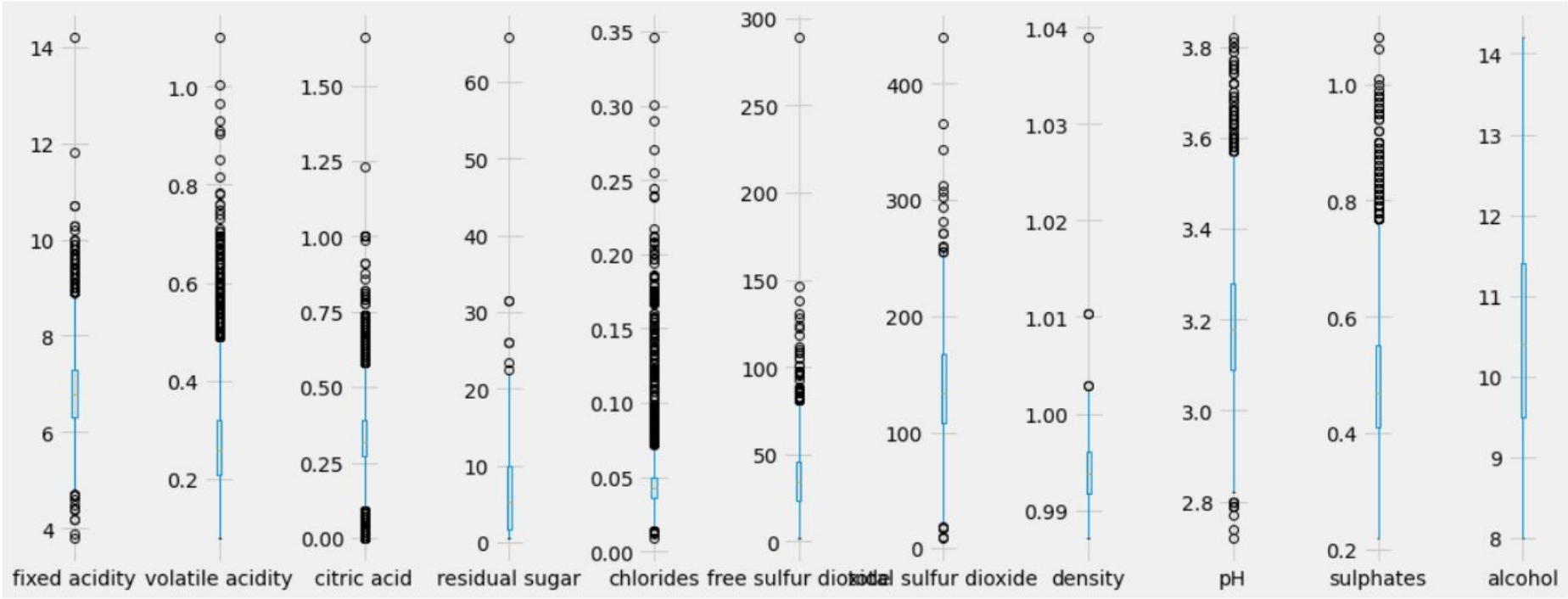
General Statistic:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

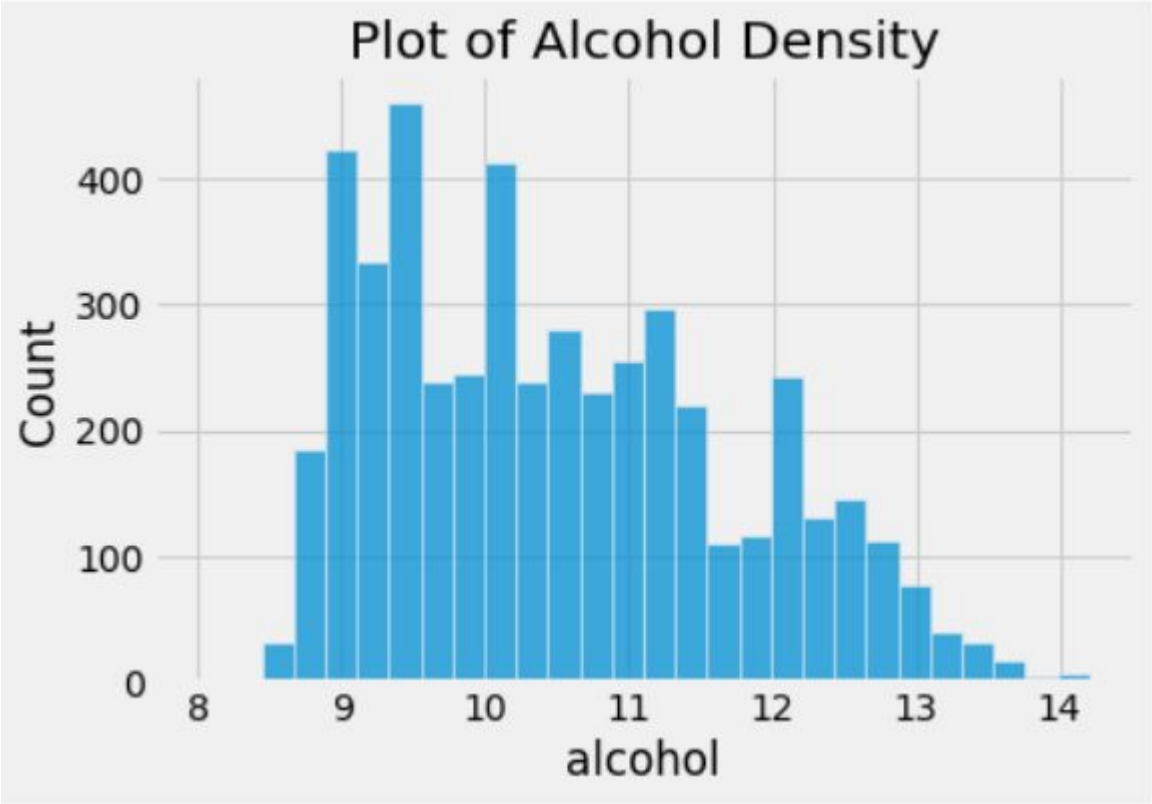
1. EDA - Exploratory Data Analysis:

Variable	Data type	Null values	Outliers
fixed acidity	float64	0	119
volatile acidity	float64	0	186
citric acid	float64	0	270
residual sugar	float64	0	7
chlorides	float64	0	208
free sulfur dioxide	float64	0	50
total sulfur dioxide	float64	0	19
density	float64	0	5
pH	float64	0	75
sulphates	float64	0	124
alcohol	float64	0	0

plot_box_plots()



Only alcohol column has no outliers. The least outliers are visible for: residual sugar and density. The most errors are in: chlorides and volatile acidity.

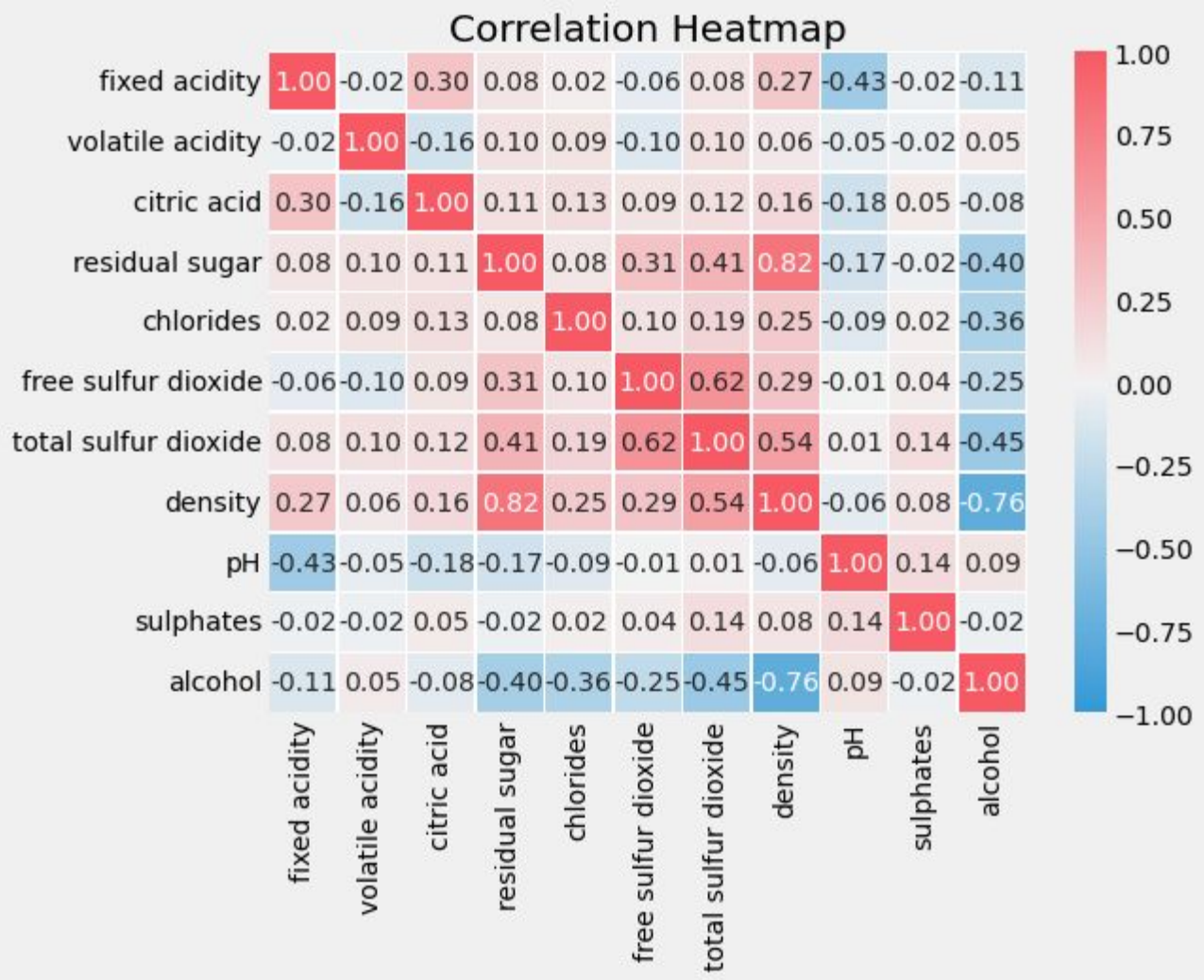

Plot of Alcohol Density
Count
400
300
200
100
0
8
9
10
11
12
13
14
alcohol

Removing duplicate rows

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	3961.00	3961.00	3961.00	3961.00	3961.00	3961.00	3961.00	3961.00	3961.00	3961.00	3961.00
mean	6.84	0.28	0.33	5.91	0.05	34.89	137.19	0.99	3.20	0.49	10.59
std	0.87	0.10	0.12	4.86	0.02	17.21	43.13	0.00	0.15	0.11	1.22
min	3.80	0.08	0.00	0.60	0.01	2.00	9.00	0.99	2.72	0.22	8.00
25%	6.30	0.21	0.27	1.60	0.04	23.00	106.00	0.99	3.09	0.41	9.50
50%	6.80	0.26	0.32	4.70	0.04	33.00	133.00	0.99	3.18	0.48	10.40
75%	7.30	0.33	0.39	8.90	0.05	45.00	166.00	1.00	3.29	0.55	11.40
max	14.20	1.10	1.66	65.80	0.35	289.00	440.00	1.04	3.82	1.08	14.20

```
1 df.shape
```

```
(3961, 11)
```



There are some strong, interesting co-dependencies between some of the features:

- alcohol vs. density;
- total sulfur dioxide (TSO2) vs. free sulfur dioxide (FSO2);
- density vs. residual sugar.

2. Normalizing Data:

$$X_i' = \frac{X_i - min(X)}{max(X) - min(X)},$$

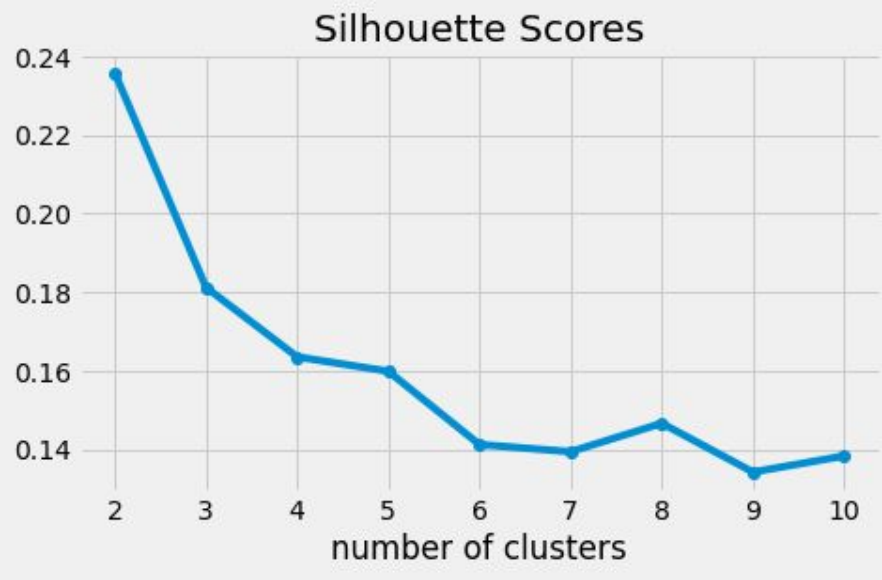
where X_i is the original value, $min(X)$ the minimum value in feature range, and $max(X)$ the maximum value.

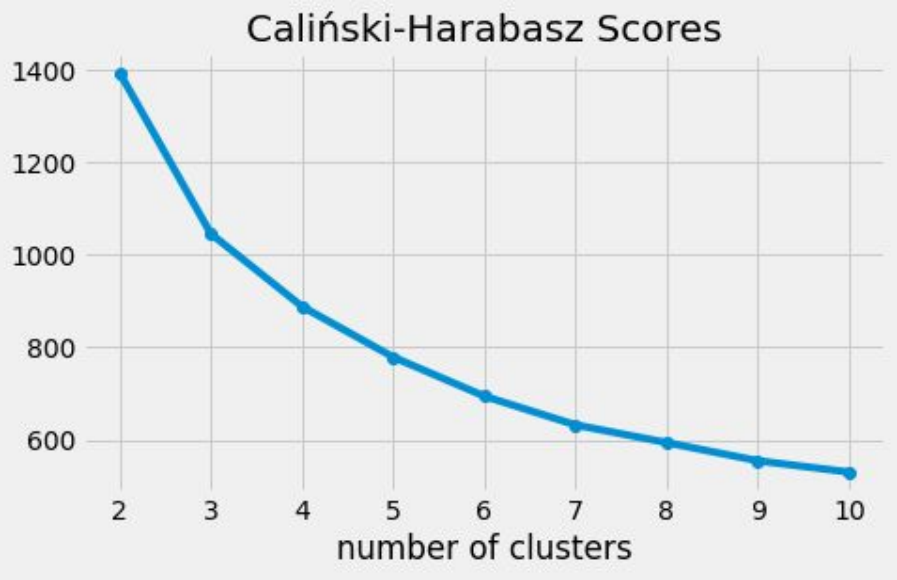
```
array([[0.30769231, 0.18627451, 0.21686747, ..., 0.25454545, 0.26744186,
        0.12903226],
       [0.24038462, 0.21568627, 0.20481928, ..., 0.52727273, 0.31395349,
        0.24193548],
       [0.41346154, 0.19607843, 0.24096386, ..., 0.49090909, 0.25581395,
        0.33870968],
       ...,
       [0.25961538, 0.15686275, 0.11445783, ..., 0.24545455, 0.27906977,
        0.22580645],
       [0.16346154, 0.20588235, 0.18072289, ..., 0.56363636, 0.18604651,
        0.77419355],
       [0.21153846, 0.12745098, 0.22891566, ..., 0.49090909, 0.11627907,
        0.61290323]])
```

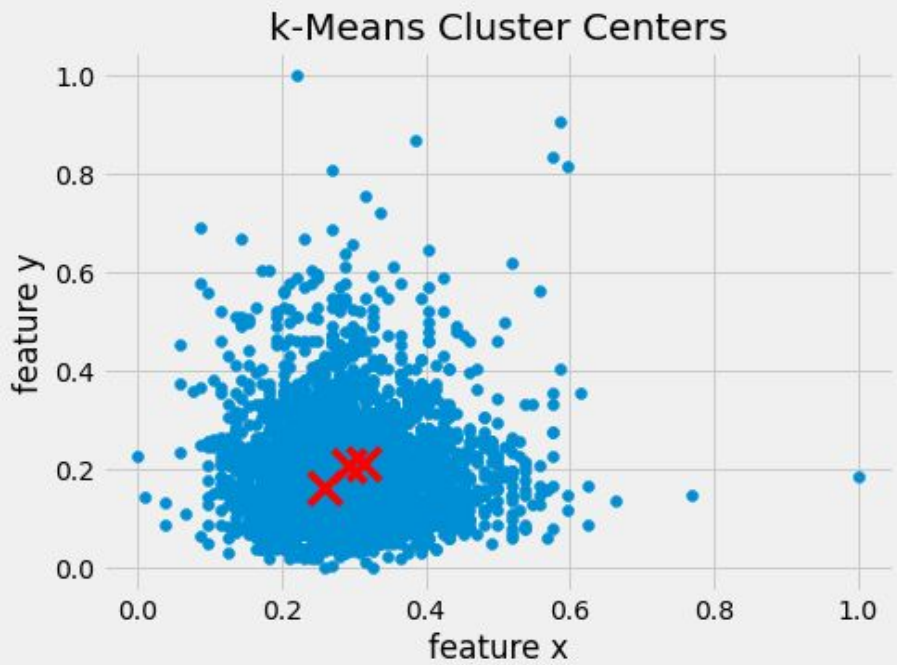
3. Evaluating Clustering Algorithm:

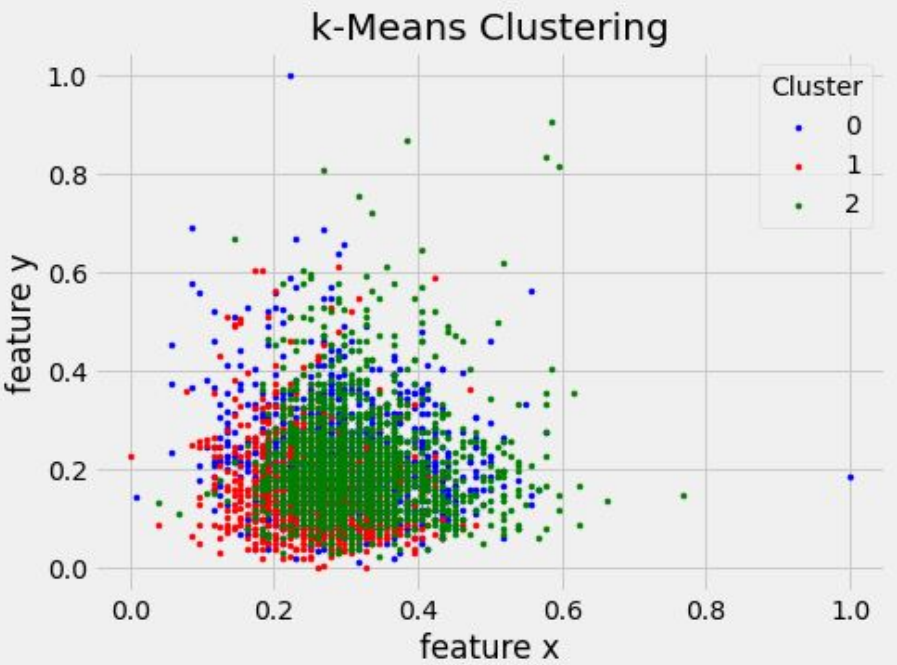
- Visually
- With Measures

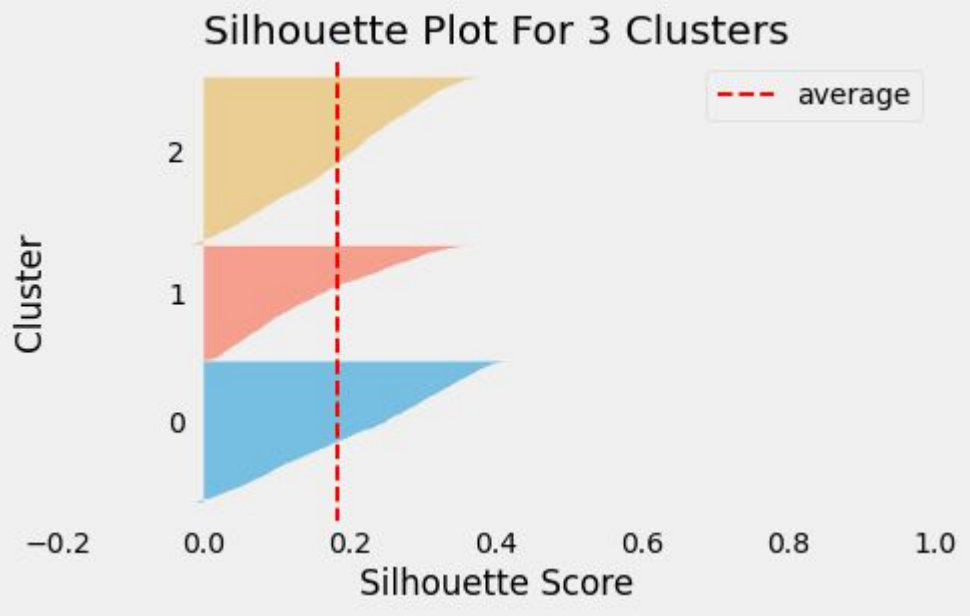
4. k-Means Clustering:


Silhouette Scores
0.24
0.22
0.20
0.18
0.16
0.14
2
3
4
5
6
7
8
9
10
number of clusters


Caliński-Harabasz Scores
1400
1200
1000
800
600
2
3
4
5
6
7
8
9
10
number of clusters

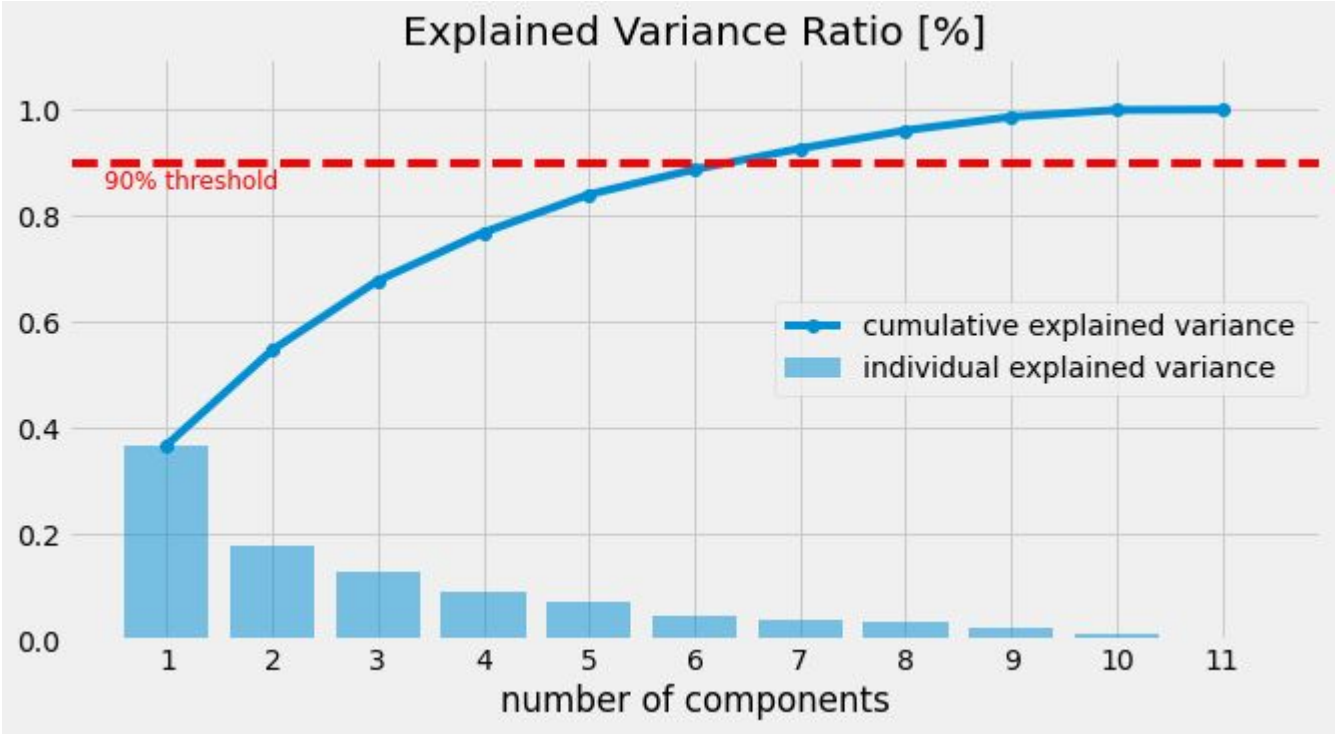

k-Means Cluster Centers
feature y
feature x
0.0
0.2
0.4
0.6
0.8
1.0


k-Means Clustering
Cluster
0
1
2
feature y
feature x
0.0
0.2
0.4
0.6
0.8
1.0




```
#### Model Validation ####
Average Silhouette Score: 0.18126118990661816
Caliński-Harabasz Score: 1045.6321448733058
Davies-Bouldin Index: 1.7659861783764155
```

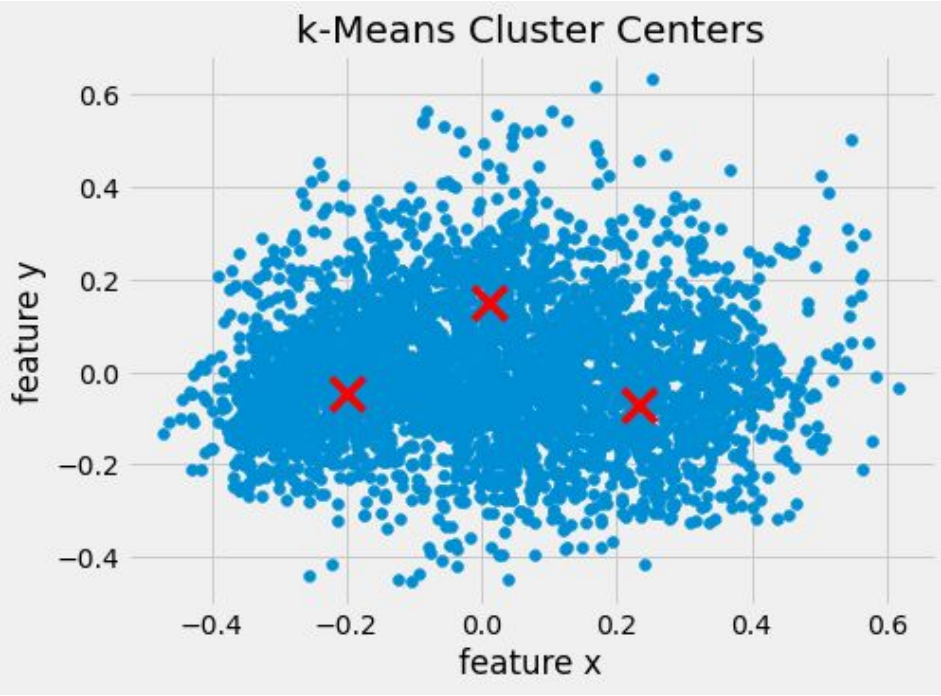
5. Dimensionality Reduction Using Principal Component Analysis (PCA)

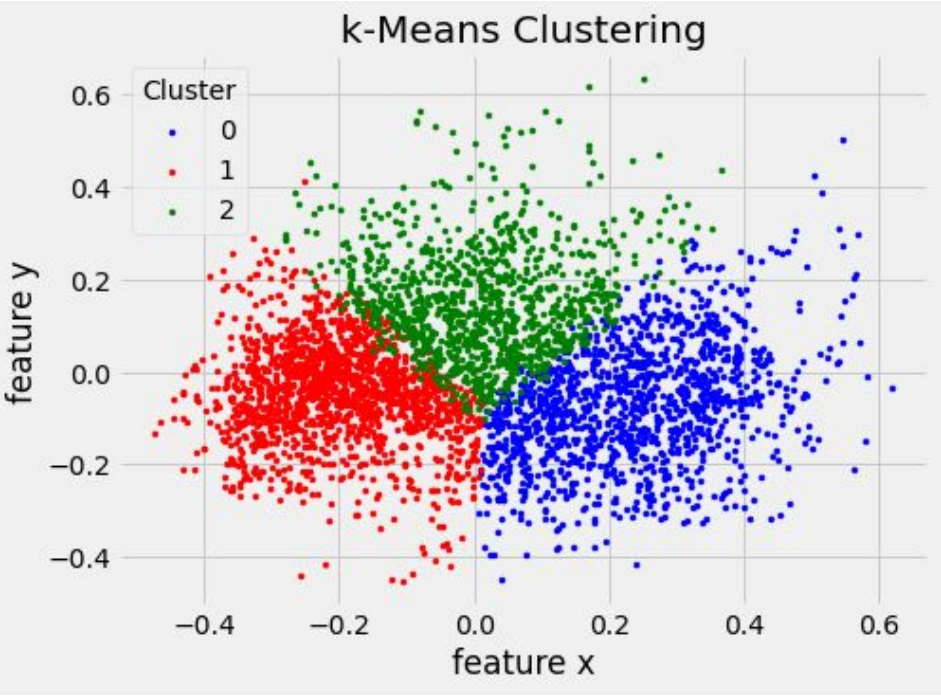


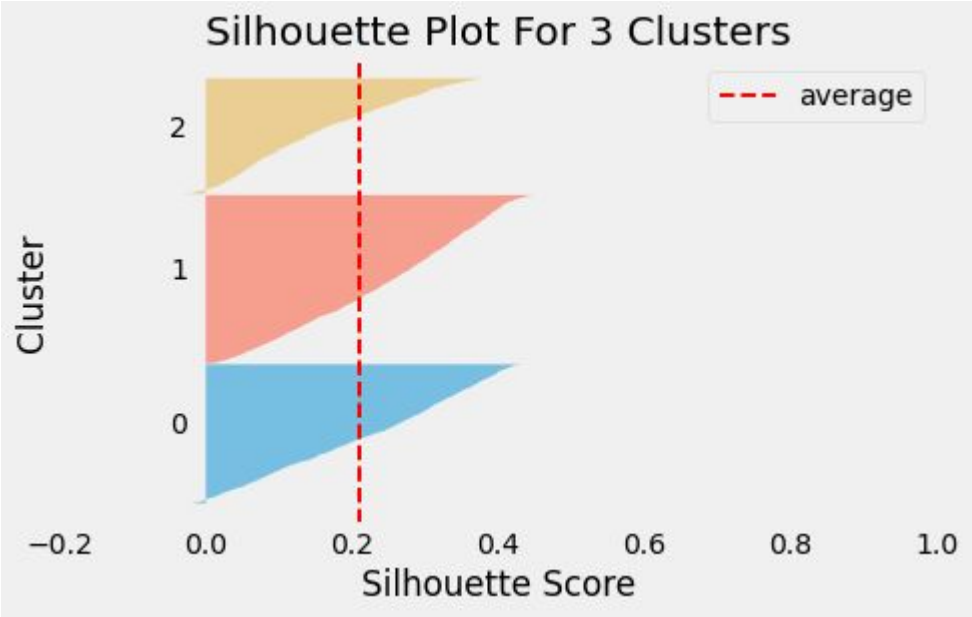
6. After Dimensionality Reduction

```
array([[-0.37376295, -0.1461661 ,  0.01459762,  0.0286885 ,  0.04700836,
        -0.05763429],
       [-0.12220269,  0.1014675 , -0.08277814, -0.02970811, -0.11097741,
         0.02710799],
       [-0.05556597, -0.02265431, -0.06342425, -0.05313953, -0.04707583,
         0.1466882 ],
       ...,
       [-0.14777124, -0.13990322, -0.008534  , -0.0900996 , -0.14365979,
        -0.1279622 ],
       [ 0.41409829,  0.04268417, -0.1525949 ,  0.03149197,  0.03856393,
        -0.03502136],
       [ 0.26234653, -0.05374422, -0.19160334, -0.0793331 ,  0.02902432,
        -0.01109314]])
```

7. Clustering After Dimensionality Reduction








```
#### Model Validation ####
Average Silhouette Score: 0.211239685844038
Caliński-Harabasz Score: 1261.6981910329434
Davies-Bouldin Index: 1.6032056261222243
```

8. Summary / Conclusion:

We can see final results after dimensionality reduction improved:

Method	Silhouette	Caliński-Harabasz	Davies-Bouldin	Cluster 0	Cluster 1	Cluster 2
k-Means	0.2116	1261.7120	1.6024	1075	1308	1578

The scores are higher and clustering process looks more cleaner.

Before:

```
#### Model Validation ####
Average Silhouette Score: 0.18126118990661816
Caliński-Harabasz Score: 1045.6321448733058
Davies-Bouldin Index: 1.7659861783764155
```

After:

```
#### Model Validation ####
Average Silhouette Score: 0.211239685844038
Caliński-Harabasz Score: 1261.6981910329434
Davies-Bouldin Index: 1.6032056261222243
```

Conclusion:

Conclusion:

This project introduces the k-means and PCA unsupervised algorithms that can be applied for exploratory data analysis and preprocessing on white wine dataset.

Right representation of data is crucial for Unsupervised Learning. Important parts of this are Preprocessing and Decomposition methods.

The dimensionality reduction of initial dataset is an essential tool to make sense of the data in the absence of supervision information. Applying PCA method improved the clustering

process. Any further enhancing should be in removing possible outliers in the dataset. Overall, clustering can be a useful exploration tool for identifying

homogeneous groups and pattern recognition within the data. This approach could help us understand more about the data before performing supervised tasks and

develop more refine models.

Thank you!!!