1. Linear Regression:
   Data set Used: White Wines

```
In [2]: File = 'C:\\Users\\yandr\\OneDrive\\Desktop\\IDA\\Assignments\\four\\winequality-white.csv'

        df = pd.read_csv(File,sep=';')
        print("Column headings:")
        print(df.columns)

        df.shape
```

```
Column headings:
Index(['fixed_acidity', 'volatile_acidity', 'citric_acid', 'residual_sugar',
       'chlorides', 'free_sulfur_dioxide', 'total_sulfur_dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

```
Out[2]: (4898, 12)
```

1a)**Do linear regression to learn the single-feature regression models, one model for each of the 11 features. Find the R^2 and AIC values for each of these models. Report these values for the models.**

Functions Used from the **statsmodels.regression.linear_model ::**

%%Generating the linear regression Model:

lmwhite_i = smf.ols(formula='quality~'+ i,data=df).fit()

%% Calculating the R2 value:

r2_value_i=(lmwhite_i.rsquared)

%% Calculating the AIC value:

Aic_value_i=(lmwhite.aic)

Output Values:

```
Single Feature Regression Models

Model::  Feature: fixed_acidity R2_Value: 0.0129 AIC_Value: 12649.542675

Model::  Feature: volatile_acidity R2_Value: 0.0379 AIC_Value: 12523.903238

Model::  Feature: citric_acid R2_Value: 0.0001 AIC_Value: 12712.818013

Model::  Feature: residual_sugar R2_Value: 0.0095 AIC_Value: 12666.374965

Model::  Feature: chlorides R2_Value: 0.0441 AIC_Value: 12492.465077

Model::  Feature: free_sulfur_dioxide R2_Value: 0.0001 AIC_Value: 12712.907425

Model::  Feature: total_sulfur_dioxide R2_Value: 0.0305 AIC_Value: 12561.351623

Model::  Feature: density R2_Value: 0.0943 AIC_Value: 12227.966722

Model::  Feature: pH R2_Value: 0.0099 AIC_Value: 12664.571954

Model::  Feature: sulphates R2_Value: 0.0029 AIC_Value: 12699.100368

Model::  Feature: alcohol R2_Value: 0.1897 AIC_Value: 11682.782414

Best Feature with highest r2 value: alcohol R2_Value :  0.1897 AIC_Value: 11682.782414
```

Hence the best model obtained:

```
Best Feature with highest r2 value: alcohol R2_Value : 0.1897
AIC_Value: 11682.782414
```

1b) Select the model with the highest R^2 value, combine with its feature other features, one at a time, and thus generate all bivariate regression models (models containing two features). One of these two features is from the selected single-feature model and the other is from one of the remaining 10 features. Report the R^2 and AIC values for all the bivariate regression models.

- Combining all the features one at a time with the Alcohol (best feature) from the single feature regression model to obtain the bivariate regression model.

%%Generating the linear regression Model:

lmwhite_i = smf.ols(formula='quality~'+alcohol+i ,data=df).fit()

Output Values:

```
Bivariate Regression Models

Model::  Feature: alcohol+fixed_acidity R2_Value: 0.1935 AIC_Value: 11661.895001

Model::  Feature: alcohol+volatile_acidity R2_Value: 0.2402 AIC_Value: 11369.551596

Model::  Feature: alcohol+citric_acid R2_Value: 0.1903 AIC_Value: 11681.344223

Model::  Feature: alcohol+residual_sugar R2_Value: 0.202 AIC_Value: 11610.316711

Model::  Feature: alcohol+chlorides R2_Value: 0.193 AIC_Value: 11665.198591

Model::  Feature: alcohol+free_sulfur_dioxide R2_Value: 0.2044 AIC_Value: 11595.558629

Model::  Feature: alcohol+total_sulfur_dioxide R2_Value: 0.1903 AIC_Value: 11681.509616

Model::  Feature: alcohol+density R2_Value: 0.1925 AIC_Value: 11668.254839

Model::  Feature: alcohol+pH R2_Value: 0.1919 AIC_Value: 11671.478547

Model::  Feature: alcohol+sulphates R2_Value: 0.1935 AIC_Value: 11662.029399

Best Feature with highest r2 value: alcohol+volatile_acidity R2_Value :  0.2402 AIC_Value: 11369.551596
```

Hence the best model in the Bivariate regression model obtained is:

```
Best Feature with highest r2 value: alcohol+volatile_acidity R2_Value:
0.2402 AIC_Value: 11369.551596
```

1c) Select the bivariate model with the highest R^2 value as the Best model at this stage. Combine a third feature from the remaining nine features with this selected bivariate model to build (and then select the best) 3-feature regression models. Report the R^2 and AIC values of all these models.

- Combining all the remaining features one at a time with the Alcohol + Volatile_acidity (best feature) from the single feature regression model to obtain the 3-feature regression model.

%%Generating the linear regression Model:

lmwhite_i = smf.ols(formula='quality~'+alcohol+ volatile_acidity+i *,data=df).fit()*

```
3 - feauture Regression Models

Model::  Feature: alcohol+volatile_acidity+fixed_acidity R2_Value: 0.2444 AIC_Value: 11344.439663

Model::  Feature: alcohol+volatile_acidity+citric_acid R2_Value: 0.2403 AIC_Value: 11371.045945

Model::  Feature: alcohol+volatile_acidity+residual_sugar R2_Value: 0.2585 AIC_Value: 11252.166211

Model::  Feature: alcohol+volatile_acidity+chlorides R2_Value: 0.2414 AIC_Value: 11364.038715

Model::  Feature: alcohol+volatile_acidity+free_sulfur_dioxide R2_Value: 0.2508 AIC_Value: 11303.127343

Model::  Feature: alcohol+volatile_acidity+total_sulfur_dioxide R2_Value: 0.2431 AIC_Value: 11352.775905

Model::  Feature: alcohol+volatile_acidity+density R2_Value: 0.2469 AIC_Value: 11328.328629

Model::  Feature: alcohol+volatile_acidity+pH R2_Value: 0.2417 AIC_Value: 11362.329766

Model::  Feature: alcohol+volatile_acidity+sulphates R2_Value: 0.2431 AIC_Value: 11353.044283

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar R2_Value :  0.2585 AIC_Value: 11252.166211
```

Hence the best model in the 3-Feature regression model obtained is:

Best Feature with highest r2 value:**alcohol+volatile_acidity+residual_sugar** R2_Value : 0.2585 AIC_Value: 11252.166211

1d) Repeat the steps above to generate (k+1)-feature models from the k-feature models until the following situation arises: all the (k+1)-feature models have an AIC value higher than the AIC value of the k-feature model from which they are being generated. Stop the process and report the k-feature model found as being the best regression model for this data. Report the features included, their coefficients, and p-values for the coefficients. Comment on the magnitudes of the p-values.

```
# Iterating through the combination of features

while(Flag='False'):
    print(count+1,"- feature Regression Models","\n")
    count=count+1
    counter=0
    i_vector=[]
    aic_value=[]
    lmwhite=[]
    for i in df.columns:
        if i=="quality" or feature.find(i)>0:
            continue;
# Building the model and calculating the R2 and AIC value
        if feature=="":
            lmwhite_dummy=smf.ols(formula='quality~'+ i,data=df).fit()
        else :
            lmwhite_dummy=smf.ols(formula='quality~ '+feature+"+"+ i
,data=df).fit()
        i_vector.append(feature+ "+"+i)
```

```
        lmwhite.append(lmwhite_dummy)
        r2_value_dummy=(lmwhite_dummy.rsquared)
        r2_value.append(r2_value_dummy)
        aic_value_dummy=(lmwhite_dummy.aic)
        aic_value.append(aic_value_dummy)
        print("Model::","Feature:",(feature+"+"+i).strip("+"),"R2_Value:",
round(r2_value_dummy,4),"AIC_Value:",round(aic_value_dummy,6),"\n")

# Finding the Best model in the iteration

    i_vec.append(i_vector)
    aic.append(aic_value)
    r2.append(r2_value)
    feature_num = r2_value.index(max(r2_value))
    feature=i_vector[feature_num]

    r2_value=[]
    best_aic.append(aic_value[feature_num])
    bestlm.append(lmwhite[feature_num])

    # Stopping Condition
    for j in aic_value:
        if j>=best_aic[-2]:
            counter=counter+1
        if counter==len(aic_value)-1:
            flag = 'True'
```

Results Obtained:

```
4 - feature Regression Models

Model::  Feature: alcohol+volatile_acidity+residual_sugar+fixed_acidity R2_Value: 0.2635 AIC_Value: 11221.377758

Model::  Feature: alcohol+volatile_acidity+residual_sugar+citric_acid R2_Value: 0.2589 AIC_Value: 11251.434603

Model::  Feature: alcohol+volatile_acidity+residual_sugar+chlorides R2_Value: 0.259 AIC_Value: 11251.347159

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide R2_Value: 0.264 AIC_Value: 11217.91164

Model::  Feature: alcohol+volatile_acidity+residual_sugar+total_sulfur_dioxide R2_Value: 0.259 AIC_Value: 11250.895046

Model::  Feature: alcohol+volatile_acidity+residual_sugar+density R2_Value: 0.2639 AIC_Value: 11218.251013

Model::  Feature: alcohol+volatile_acidity+residual_sugar+pH R2_Value: 0.262 AIC_Value: 11230.846986

Model::  Feature: alcohol+volatile_acidity+residual_sugar+sulphates R2_Value: 0.2619 AIC_Value: 11231.640638

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide R2_Value :  0.264 AIC_Value: 11
217.91164
```

```
5 - feature Regression Models

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+fixed_acidity R2_Value: 0.268 AIC_Value: 11192.98
9069

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+citric_acid R2_Value: 0.2646 AIC_Value: 11216.079
98

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+chlorides R2_Value: 0.2646 AIC_Value: 11216.06779
1

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+total_sulfur_dioxide R2_Value: 0.2646 AIC_Value:
11215.734957

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density R2_Value: 0.269 AIC_Value: 11186.809713

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+pH R2_Value: 0.267 AIC_Value: 11199.996083

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+sulphates R2_Value: 0.2669 AIC_Value: 11200.66343
7

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density R2_Value :  0.269 AIC_V
alue: 11186.809713
```

```
6 - feature Regression Models

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+fixed_acidity R2_Value: 0.27 AIC_Value: 1
1181.541293

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+citric_acid R2_Value: 0.2691 AIC_Value: 1
1188.057845

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+chlorides R2_Value: 0.2693 AIC_Value: 111
86.343059

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+total_sulfur_dioxide R2_Value: 0.269 AIC_
Value: 11188.662708

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH R2_Value: 0.2752 AIC_Value: 11146.8863
28

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+sulphates R2_Value: 0.2747 AIC_Value: 111
49.890655

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH R2_Value :  0.2752 A
IC_Value: 11146.886328
```

```
7 - feature Regression Models

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+fixed_acidity R2_Value: 0.2759 AIC_Val
ue: 11143.725576

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+citric_acid R2_Value: 0.2752 AIC_Valu
e: 11148.493578

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+chlorides R2_Value: 0.2753 AIC_Value:
11147.904806

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+total_sulfur_dioxide R2_Value: 0.2752
AIC_Value: 11148.71186

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates R2_Value: 0.2801 AIC_Value:
11115.406935

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates R2_Value :
0.2801 AIC_Value: 11115.406935
```

```
8 - feature Regression Models

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acidity R2_Value: 0.28
18 AIC_Value: 11106.287754

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+citric_acid R2_Value: 0.2802
AIC_Value: 11117.156191

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+chlorides R2_Value: 0.2803 A
IC_Value: 11116.466068

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+total_sulfur_dioxide R2_Valu
e: 0.2802 AIC_Value: 11116.640217

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acid
ity R2_Value :  0.2818 AIC_Value: 11106.287754
```

```
9 - feature Regression Models

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acidity+citric_acid R2
_Value: 0.2818 AIC_Value: 11108.264712

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acidity+chlorides R2_V
alue: 0.2818 AIC_Value: 11108.093048

Model::  Feature: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acidity+total_sulfur_d
ioxide R2_Value: 0.2818 AIC_Value: 11107.719164

Best Feature with highest r2 value: alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acid
ity+total_sulfur_dioxide R2_Value :  0.2818 AIC_Value: 11107.719164
```

In this Iteration the AIC values obtained are 11108.26,11108.09,11107.71 which are all greater than the best combination from 8 feature regression model 11106.28. Hence, we stop the iterations and get the best model as:

8- Feature Regression model best fitting to the data with optimal R2 and AIC value is:
`Best Feature with highest r2 value:` **`alcohol+volatile_acidity+residual_sugar+free_sulfur_dioxide+density+pH+sulphates+fixed_acidity`** `R2_Value: 0.2818 AIC_Value: 11106.287754`

Feature	Co-efficient	P Value
Intercept	154.1062	2.2E-17
Alcohol	0.1932	1.31E-15
volatile_acidity	-1.8881	1.02E-64
residual_sugar	0.0828	1.39E-29
free_sulfur_dioxide	0.0033	7.67E-07
Density	-154.2913	5.28E-17
pH	0.6942	2.07E-11
sulphates	0.6285	3.52E-10
fixed_acidity	0.0681	8.64E-04

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.282
Model:	OLS	Adj. R-squared:	0.281
Method:	Least Squares	F-statistic:	239.7
Date:	Thu, 29 Nov 2018	Prob (F-statistic):	0.00
Time:	20:46:59	Log-Likelihood:	-5544.1
No. Observations:	4898	AIC:	1.111e+04
Df Residuals:	4889	BIC:	1.116e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	154.1062	18.100	8.514	0.000	118.622	189.591
alcohol	0.1932	0.024	8.021	0.000	0.146	0.240
volatile_acidity	-1.8881	0.110	-17.242	0.000	-2.103	-1.673
residual_sugar	0.0828	0.007	11.370	0.000	0.069	0.097
free_sulfur_dioxide	0.0033	0.001	4.950	0.000	0.002	0.005
density	-154.2913	18.344	-8.411	0.000	-190.254	-118.329
pH	0.6942	0.103	6.717	0.000	0.492	0.897
sulphates	0.6285	0.100	6.287	0.000	0.433	0.824
fixed_acidity	0.0681	0.020	3.333	0.001	0.028	0.108

Omnibus:	114.194	Durbin-Watson:	1.621
Prob(Omnibus):	0.000	Jarque-Bera (JB):	251.255
Skew:	0.075	Prob(JB):	2.76e-55
Kurtosis:	4.099	Cond. No.	9.95e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.95e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

P_Value: The p-values for the coefficients indicate whether these relationships are statistically significant. **Along with the coefficients, p values provide enough evidence to reject the otherwise taken null hypothesis instead of the regression line obtained**.

A higher Pvalue indicate that the variable is not significant for the regression model, whereas a lower magnitude provides enough evidence that the inclusion of the variable is significant for the regression model and that the target value is dependent on the variable.

In our model all the pvalues are almost equal to 0.00 which means that they are all very much significant in regression model and contribute in the prediction.

1e) Find the five wines that have the largest magnitudes of difference between the predicted and the actual wine-quality values. Look at the regression model, the rest of the data, and comment on why you think these wines are outliers.

#Fitting the obtained Regression line on the data to predict the value:

```
pred=bestlm[7].predict(X_Val)
error=abs(pred)
```

Top 5 wines with largest magnitude of error: (Index started from 0)

Index in Data	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	Predicted_Value	Errror
253		0.24	0.44	3.5	0.029	5	109	0.9913	3.53	0.43	11.7	3	6.3867023	3.3867023
445	7.1	0.32	0.32	11	0.038	16	66	0.9937	3.24	0.4	11.5	3	6.353270875	3.353270875
3307	9.4	0.24	0.29	8.5	0.037	124	208	0.99395	2.9	0.38	11	3	6.431779787	3.431779787
3810	6.8	0.26	0.34	15.1	0.06	42	162	0.99705	3.24	0.52	10.5	3	6.238258625	3.238258625
4745	6.1	0.26	0.25	2.9	0.047	289	440	0.99314	3.44	0.64	10.5	3	6.824598916	3.824598916

The quality groups of 3,9 are only 25 among the 4898. The very little proportion of these stay as the outliers providing very little learning data for the regression models. If the model is fitted to include these as well, it would lead to overfitting of the data.

2. **Clustering**:

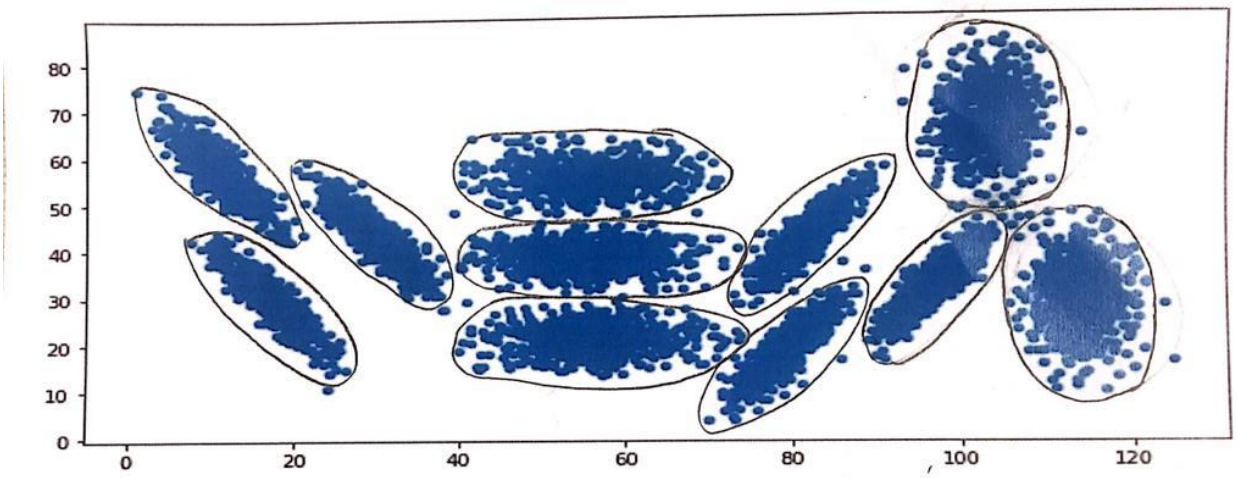
Data set Used: White Wines

In [2]:
```
File = 'C:\\Users\\yandr\\OneDrive\\Desktop\\IDA\\Assignments\\four\\HW4GaussianClustersData.csv'

df1 = pd.read_csv(File)

print("Column headings:")
print(df1.columns)
df1.shape
```

```
Column headings:
Index(['X', 'Y'], dtype='object')
```

Out[2]: (6600, 2)

2a) Plot the data on a 2-D scatter plot and mark by hand the boundaries of the ideal clusters that you would like discovered in this dataset.



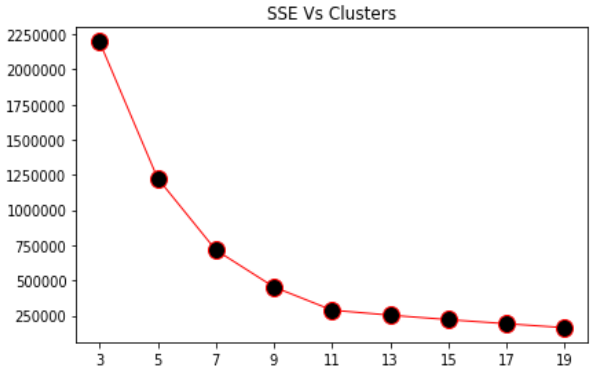
2b) Run the k-means algorithm for k = 3, 5, 7, 9, 11, 13, 15, 17 and 19. Plot the total SSE and BIC values for the above values of k. What is the best number of clusters for this dataset? How did you find the best number of clusters, briefly explain.

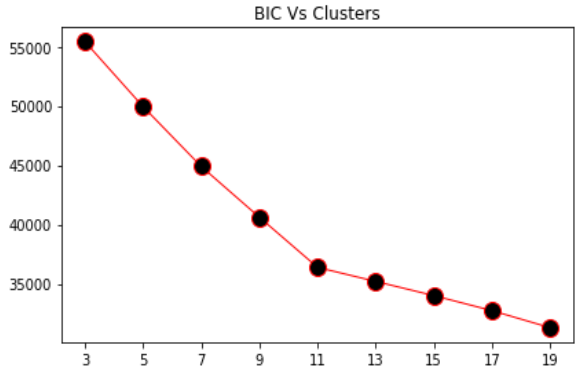
```
Function used : k=KMeans(n_clusters=n, random_state=0).fit(X_set)
sse_1=k.inertia_
```

SSE and BIC values plotted:

SSE – Sum of Squares Error
BIC = n*log(SSE/n) + log(n)*c*(d+1)

SSE : [2197916.389830343, 1226034.9648687756, 715958.7767559565, 452886.12803022546, 288182.23747907684, 252374.8302387761, 220838.09934794006, 191738.1678454749, 163550.34196059694]
BIC : [55418.59571144362, 49936.62508592988, 44890.813398306156, 40606.13696001127, 36377.96469481235, 35190.765884857836, 33995.874346819684, 32726.58369880081, 31288.6548795303]

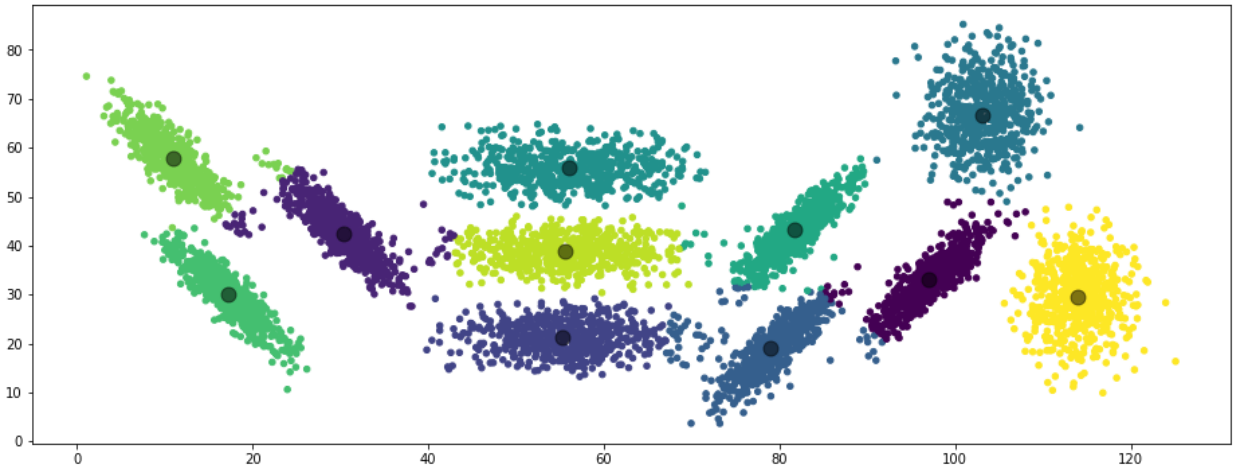




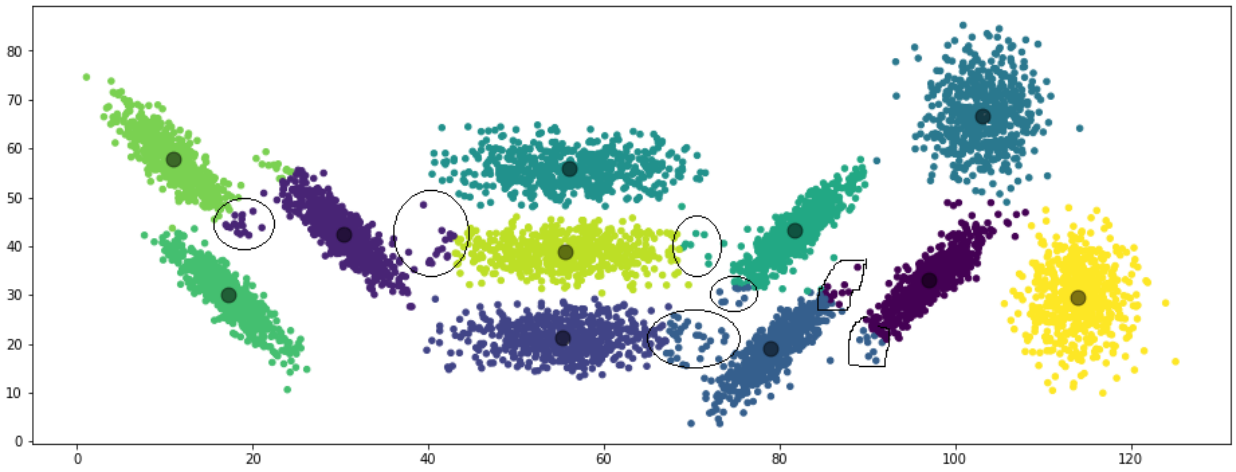
As the above two plots infer, the knee point is obtained at the k =11 ie., the change in SSE or BIC is not very significant or almost remains same when compared to the other k values earlier in the curve.
Though the errors of the higher k value is low it is deceptive in the sense that it breaks the original clusters into smaller clusters and so the error decreases.
Hence k =11 is chosen as the best number of clusters.

2c) For the best number of clusters selected above, plot the scatter plot of the data showing the points of each cluster with a different color/symbol. Mark the points on the scatter plot that belong to clusters other than what your intuition says. Why did k-means algorithm place them in these different clusters – explain very briefly.

Scatter Plot of the Clusters Formed:



Kmeans differed from my intuition for the marked points:



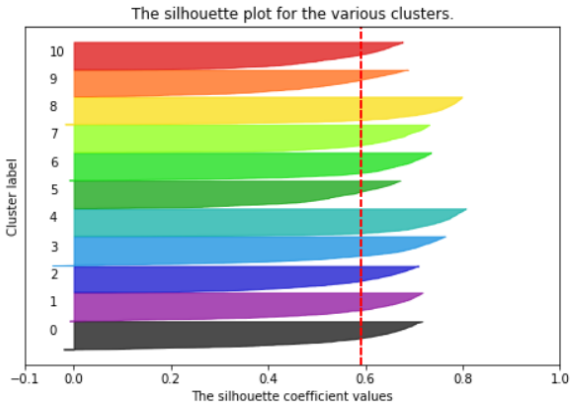
K means clustering is **a center based clustering technique**. Hence, to a greater extent depends on the initial centroids chosen. Now once the centroids are chosen initially, iteratively the distances are calculated from all the other data points to the centroid and all the data points nearest to the centroid are made into a cluster.

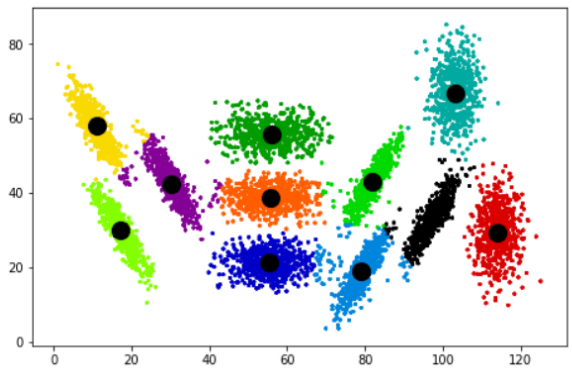
Hence, K means tend to give the globular shaped structures and may not be very effective for others.

In our example as well, the points marked are all points that are marked are nearer to the centroids that the k means chose and so are placed in that way, which is different from the intuition.

This is one of those situations where we can see that the K means may fail with the non-globular structures.

2d) Plot the silhouette diagram for the best clustering you have selected. Comment on the characteristics of the silhouette diagram that you think are informative about this clustering. Comment using the cluster numbers and their plots on the silhouette diagram.





The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

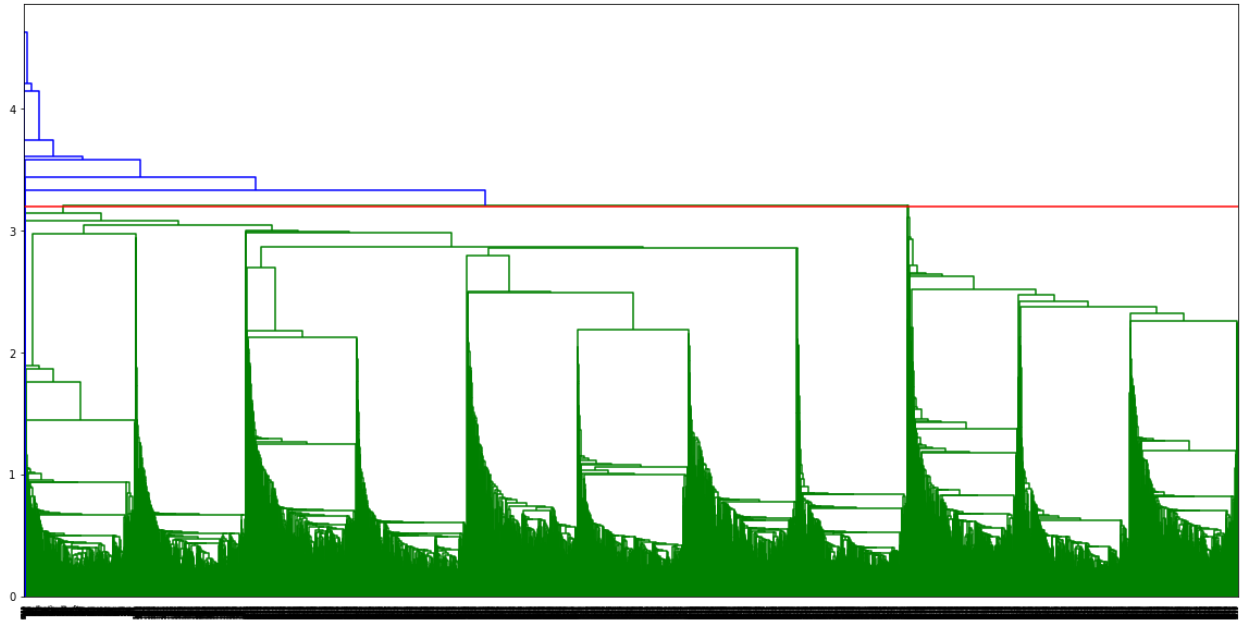
So in the above figure, the clusters (8)yellow,(4) and (3)blue are having data points with higher values which means that they are closely packed and are nearer to their centroids and far away from other centroids. And the width of the clusters depict the number of data points.
However, the clusters (5) green,(9) orange,(10)red are loosely packed as they have points that are away from the centroid of its own cluster.
The cluster (4) is a good cluster with greater number of points closely bounded followed by (8).
On the whole the coefficient is very much nearer to 0.6 which means the points are well in cohesion and the clusters are well separated to some extent. Though this is not very great, they are not bad as well.

2e) Perform single-linkage hierarchical clustering for this data and cut the dendrogram to obtain 11 clusters. There are options/parameters in most toolboxes to generate a given number of clusters. Plot the 2-D scatter plot of the dataset showing data points of each of the 11 clusters with different color/symbol.

Function Used:

```
z=sc.cluster.hierarchy.linkage(X_set,method='single',metric='euclidean)
r=sc.cluster.hierarchy.dendrogram(z,no_plot=False)
```

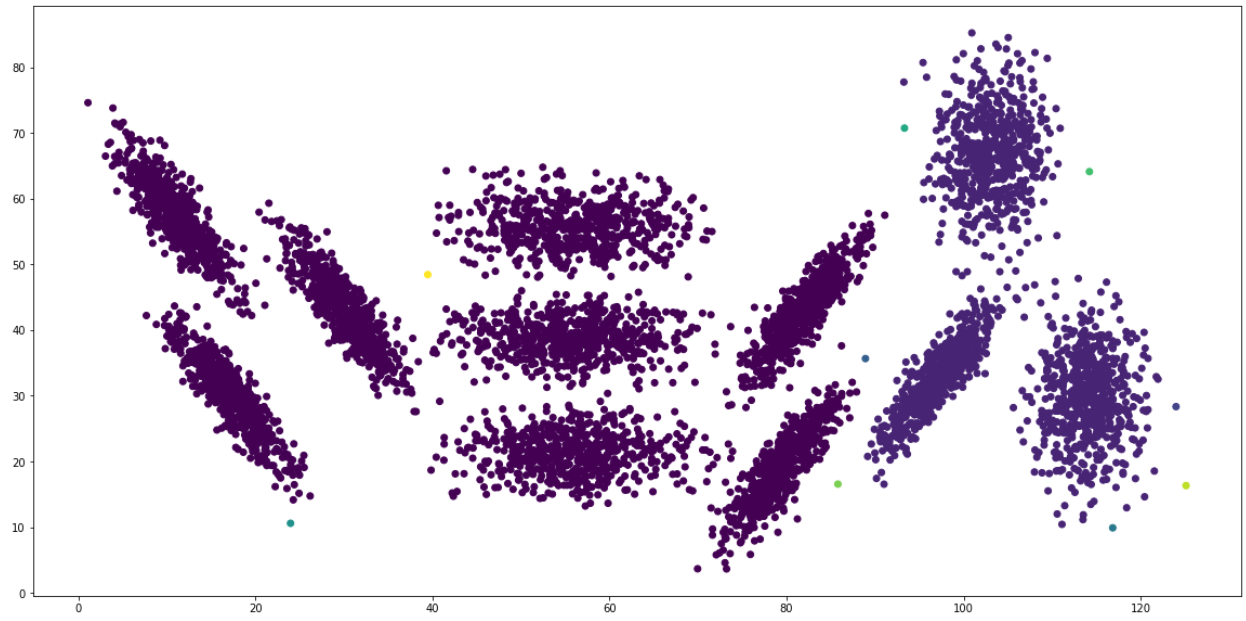
The red line in the below graph shows the vertical distance that would give 11 clusters.




```
%% Cut the dendogram at the 3.2 distance
fc=fcluster(z,3.2,criterion='distance')

values, counts = np.unique(fc, return_counts=True)
print("Number of Data points in each cluster ",counts)
print("Cluster Number",values)
```

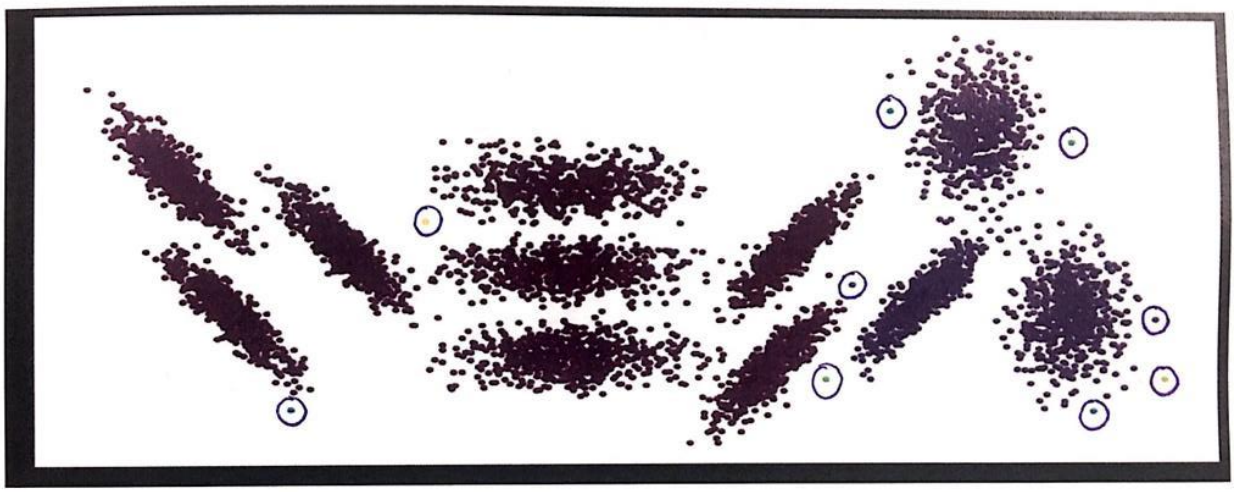
Cluster Number	Number of Data Points
1	4797
2	1794
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1

Scatter Plot:



2f) Mark any data points on this scatter plot that are clustered differently from your intuitive view of the correct clusters. Explain why Single-linkage clustering may have placed them in counter-intuitive clusters.

Single Linkage clustering is a type of Agglomerative Hierarchical clustering where the clusters are merged when the nearest points between clusters have a minimum distance. The clustering starts from the individual points starting at random and iteratively combines the clusters that have minimum distances between its nearest points.



Except the points that are marked all the other are intuitively placed into separate clusters but the single linkage placed them into 2 clusters. This happened because it forms contingency clusters which is one of the major drawbacks of single linkage clusters.
Single linkage cannot identify the clusters that have data points between them, even if they are scarce it combines them.

All the clusters here have at least few data points between them , hence they are formed as one cluster.