APPLIED DATA SCIENCE – CLUSTERING AND FITTING

STUDENT NAME:

SUPERVISOR:

Aim

The main goal is to analyse on the fitting and clustering applied to World Bank Dataset. Population growth statistics and data on electricity usage by country were among the datasets used in this study.

FIND INTERESTING CLUSTERS OF DATA.

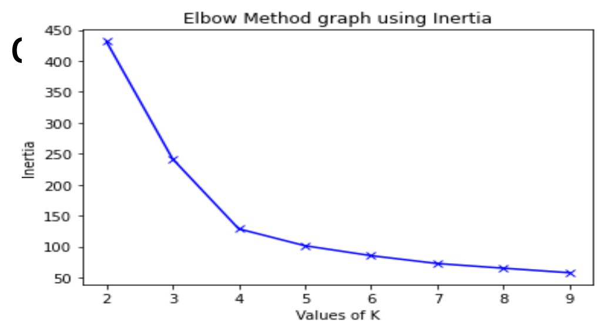
Data Collection

The dataset used here is Population growth data from world bank dataset. The dataset shape is given below.

Number of Rows	Number of Columns
266	65

Plotting Elbow Method graph using Inertia

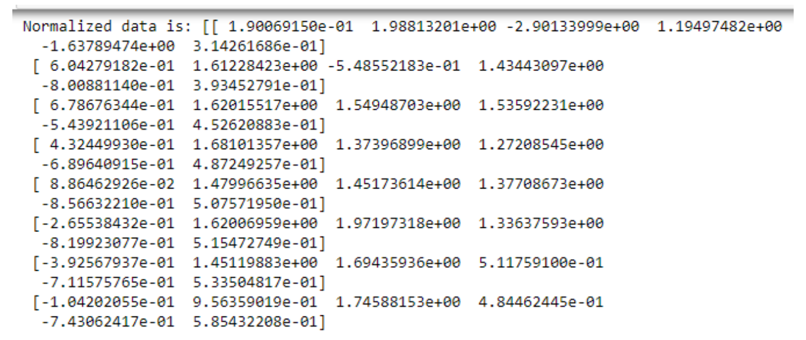
Elbow Method is used to calculate the number of c means algorithm (Kane, A. and Nagar, J., 2012).



Looking at the Elbow method graph above, we can see that the number of clusters for population growth data is four.

Scaling the input data to fit the k-means algorithm

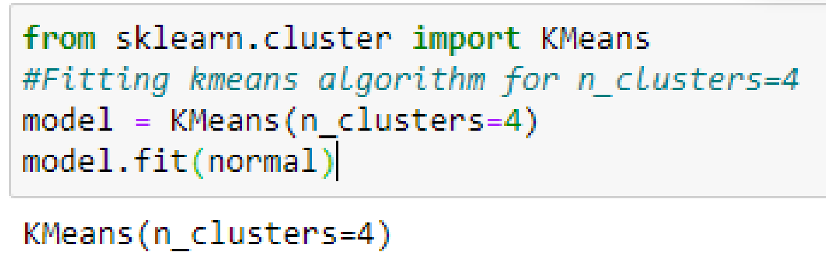
When the data is normalised, clustering works best. As a result, in this assignment, normalisation is performed using the scikit learn package's standard scalar technique.



```
Normalized data is: [[ 1.90069150e-01  1.98813201e+00 -2.90133999e+00  1.19497482e+00
  -1.63789474e+00  3.14261686e-01]
 [ 6.04279182e-01  1.61228423e+00 -5.48552183e-01  1.43443097e+00
  -8.00881140e-01  3.93452791e-01]
 [ 6.78676344e-01  1.62015517e+00  1.54948703e+00  1.53592231e+00
  -5.43921106e-01  4.52620883e-01]
 [ 4.32449930e-01  1.68101357e+00  1.37396899e+00  1.27208545e+00
  -6.89640915e-01  4.87249257e-01]
 [ 8.86462926e-02  1.47996635e+00  1.45173614e+00  1.37708673e+00
  -8.56632210e-01  5.07571950e-01]
 [-2.65538432e-01  1.62006959e+00  1.97197318e+00  1.33637593e+00
  -8.19923077e-01  5.15472749e-01]
 [-3.92567937e-01  1.45119883e+00  1.69435936e+00  5.11759100e-01
  -7.11575765e-01  5.33504817e-01]
 [-1.04202055e-01  9.56359019e-01  1.74588153e+00  4.84462445e-01
  -7.43062417e-01  5.85432208e-01]
```

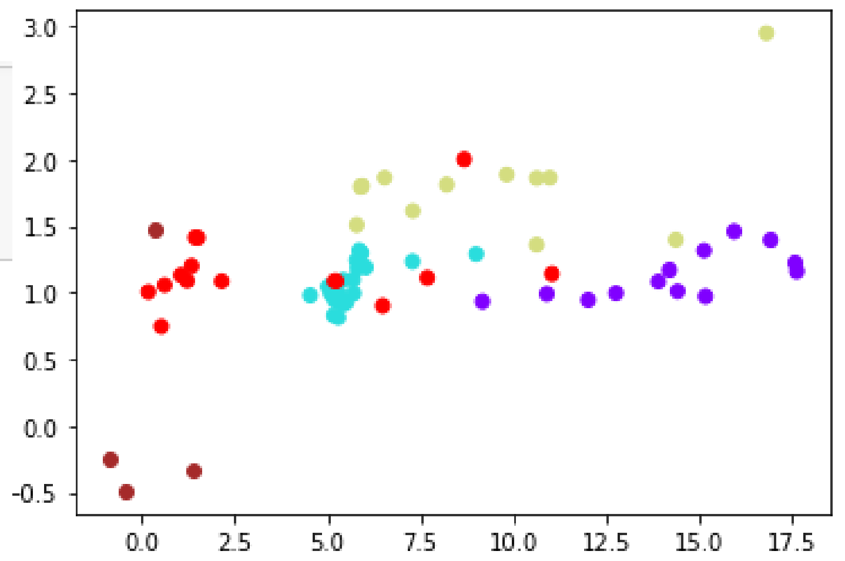
Fitting K-Means Clustering Algorithm

The K-means method is used, and the outcomes are obtained. Clusters are constructed from the results (Singh, A., Yadav, A. and Rana, A., 2013). The following is a graphic representation of Centroid:



```
from sklearn.cluster import KMeans
#Fitting kmeans algorithm for n_clusters=4
model = KMeans(n_clusters=4)
model.fit(normal)
```

```
KMeans(n_clusters=4)
```



Inference:

From the graph, it is understood that four clusters are formed, which are relatively separated with few outliers.

CREATE SIMPLE MODEL(S) FITTING DATA SETS WITH CURVE_FIT

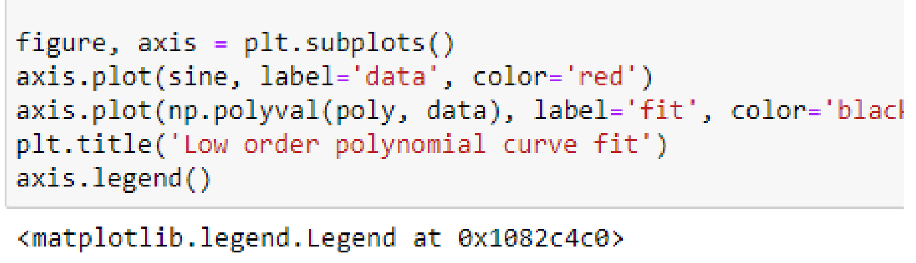
Data Used for Curve Fit

Ireland Country Population Growth was selected as the suitable dataset for curve fit.

Making low order polynomial for country Ireland

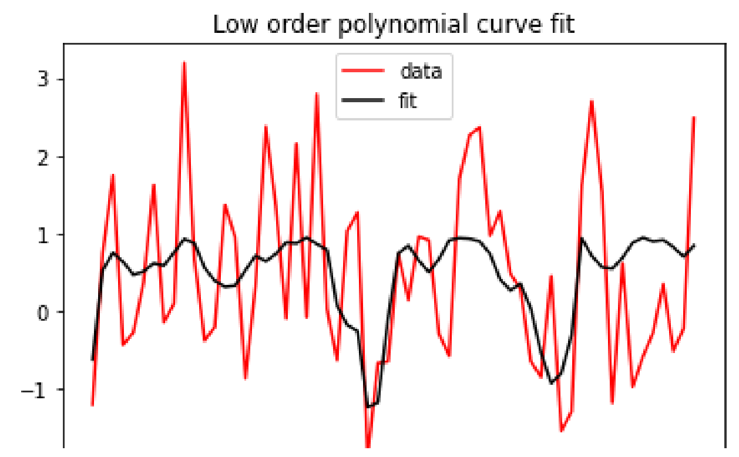
The coefficients of a polynomial p(x) with degree n that provides the greatest fit for the data in y are returned by the polynomial function (Makarenkov, V. and Legendre, P., 2002). The coefficients in p are written in a descending power format, and the length of p is one more than n. The degree of polynomial that is being utilised here is 5, and the colours red and black are being specified. Following the curve is an implementation of a low order polynomial that has been provided for your convenience.

As can be seen in the graphic to the right, the outcome of applying a low-order polynomial revealed that the curve that was formed had a greater chance of falling in the middle of a data flow.



```
figure, axis = plt.subplots()
axis.plot(sine, label='data', color='red')
axis.plot(np.polyval(poly, data), label='fit', color='black
plt.title('Low order polynomial curve fit')
axis.legend()
```

```
<matplotlib.legend.Legend at 0x1082c4c0>
```



References

Kane, A. and Nagar, J., 2012. Determining the number of clusters for a k-means clustering algorithm. Indian Journal of Computer Science and Engineering, 3(5), pp.670-672.

Singh, A., Yadav, A. and Rana, A., 2013. K-means with Three different Distance Metrics. International Journal of Computer Applications, 67(10).

Makarenkov, V. and Legendre, P., 2002. Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. Ecology, 83(4), pp.1146-1161.

Data.worldbank.org. 2022. Population growth (annual %) | Data. [online] Available at: https://data.worldbank.org/indicator/SP.POP.GROW?view=chart.

Data.worldbank.org. 2022. Electric power consumption (kWh per capita) | Data. [online] Available at: https://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC?end=2019&start=1960&view=chart.

CLUSTERING AND FITTING

Dataset

The dataset used for clustering and fitting is Electric Power consumption data from the world bank.

The total rows and columns are:

Number of rows: 266

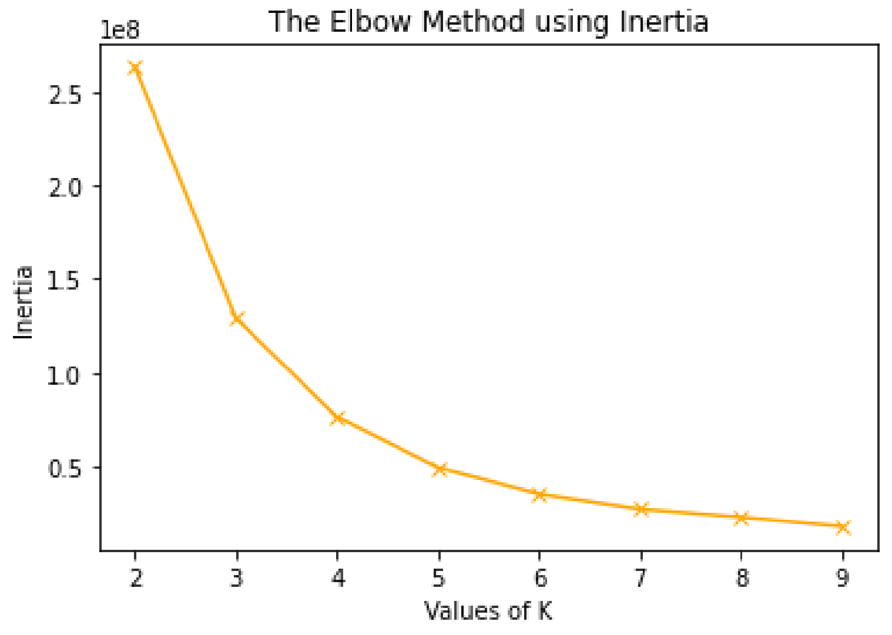
Number of Columns: 65

Curve fitting using Low order polynomial for country Ireland

The dataset is processed by the Curve fit technique, and the degree of polynomial that is used is 5.

Finding the Number of clu

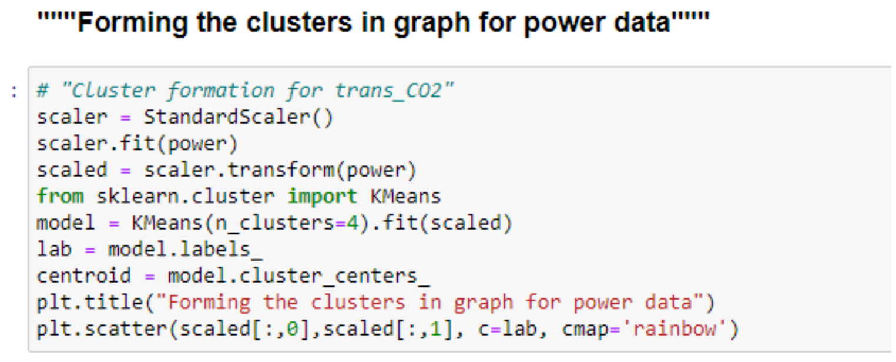
When the elbow Method is applied, it is understood that the number of right clusters for this variables is four, as gleaned from the graph that is produced. The curve takes a sharp turn to the right at cluster 4, forming an elbow bend.



Forming k-means clusters in graph for power data:

Inference

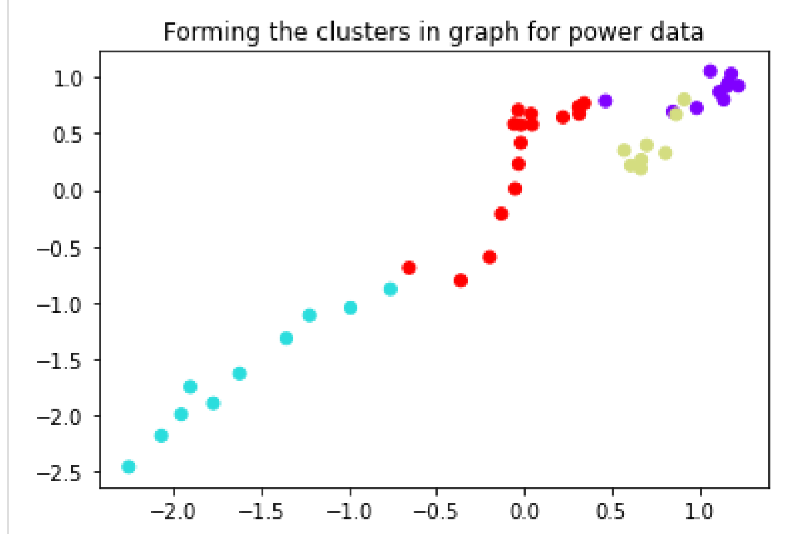
The K-means algorithm is put into action in this part of the work, and the results are presented thereafter. As a direct result of the findings, clusters have been constructed. Based on the findings, it is possible to draw the conclusion that these clusters are situated in one of four distinct geographic areas. This indicates that there have been major shifts in the amount of electrical power that has been consumed in Ireland over the course of the past 45 years.



"""Forming the clusters in graph for power data"""

```
# "Cluster formation for trans_CO2"
scaler = StandardScaler()
scaler.fit(power)
scaled = scaler.transform(power)
from sklearn.cluster import KMeans
model = KMeans(n_clusters=4).fit(scaled)
lab = model.labels_
centroid = model.cluster_centers_
plt.title("Forming the clusters in graph for power data")
plt.scatter(scaled[:,0],scaled[:,1], c=lab, cmap='rainbow')
```

```
<matplotlib.collections.PathCollection at 0x11b6713(
```


Forming the clusters in graph for power data