APPLIED DATA SCIENCE – CLUSTERING AND FITTING

STUDENT NAME: Sanga Mounika

Aim

The major objective is to comprehend and evaluate comprehensively the fitting and clustering used to the World Bank dataset. The datasets used in this study include GDP growth statistics and CO2 emission data broken down by country.

FIND INTERESTING CLUSTERS OF DATA.

Data Collection and Data Pre processing

The dataset used here is Country wise GDP growth data. The dataset shape is given below.

Number of Rows	Number of Columns
266	66

Using Clustering Method

Why Normalization

Clustering works best when the data are normalised. Therefore in this assignment, normalization is done using standard scalar method in scikit learn package.

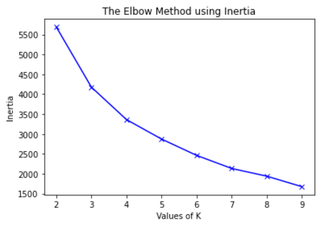
Normalisation

```
from sklearn.preprocessing import StandardScaler
# Scaling the input data to fit the kmeans algorithm
scaler = StandardScaler()
scaler.fit(df)
scaled = scaler.transform(df)
scaled
```

```
array([[-0.50909408,  0.46663196, -5.24001471,  1.42375497, -0.28976269,
         0.4639894 ],
       [-1.18643348,  0.30389072, -2.02387271,  0.78899953,  1.45725467,
        -0.17849517],
       [ 1.61709037, -1.28523241,  0.33077621,  1.26604904,  0.67569427,
         0.45453803],
       [ 2.05215898,  1.7829739 ,  1.499204  ,  1.93248551,  1.31933224,
        -0.51287218],
       [ 1.4837329 , -0.73064522,  1.31682251,  0.59372336,  1.59517709,
         0.27899846],
       [-0.56683189, -0.47231767,  0.38267338, -0.17868048,  1.64115123,
```

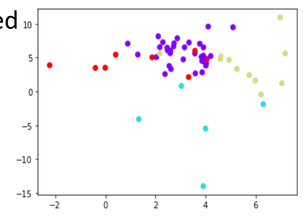
Finding Clusters using Elbow Method

The cumber of clusters suitable for the K-Means algorithm is calculated using Elbow method.



Implementation of K-Means

The K-means algorithm is implemented and the results are obtained. From The results, clusters are formed. The visualization is given below with Centroid is given below:



CREATE SIMPLE MODEL(S) FITTING DATA SETS WITH CURVE_FIT

Data Collection

The dataset chosen for the fitting is Australia Country GDP.

Low order polynomial for country Australia

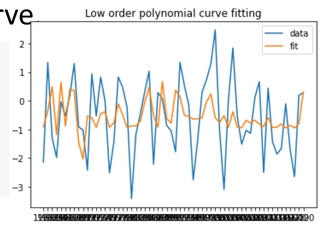
Polynomial returns the coefficients for a polynomial p(x) of degree n that is a best fit for the data in y. The coefficients in p are in descending powers, and the length of p is n+1. **The degree of polynomial used here is 5.** The implementation od low order polynomial is given below followed by the curve

```
"""Making low order polynomial for country Australia"""
x_values=df['Australia']
f = 1/4

sine = np.sin(2*np.pi*f*x_values) + np.random.normal(size=len(x_values))

poly = np.polyfit(x_values, sine, deg=5) #Fitting the polynomial method

fig, ax = plt.subplots()
ax.plot(sine, label='data')
ax.plot(np.polyval(poly, x_values), label='fit')
plt.title('Low order polynomial curve fitting')
ax.legend()
```



Using a low-order polynomial, the result showed that the curve produced is more likely to fall in the centre of a data flow, as seen in the picture above.

Data Source

Country wise GDP growth dataset is taken from https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?view=chart

Country wise CO2 Emission dataset is taken from https://data.worldbank.org/indicator/EN.ATM.CO2E.KT?view=chart

CLUSTERING AND FITTING

Data Collection

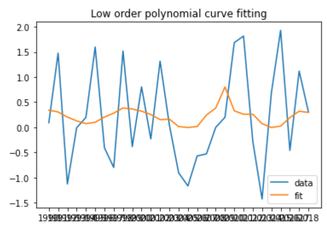
The dataset used here is Country wise CO2 Emission. The dataset shape is given below. The selected data for this Australia.

Number of Rows	Number of Columns
26	1 (CO2 of Australia)

Curve Fitting using Polynomial Curve

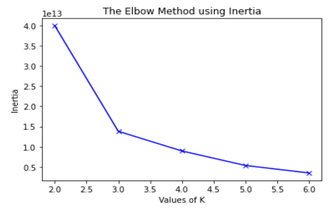
Degree of Polynomial is 5

The dataset is applied to Curve fit method, with the Degree of polynomial as 5.



Elbow Method to find number of Clusters

When applied to elbow Method, it is understood that the number of right cluster for this data is 3, from the graph generated. There is a deep elbow bend in the curve at cluster 3.



Applying K-Means Clustering

In this section, the K-means algorithm is implemented and the results are shown. Clusters are produced as a consequence of the findings. It may be inferred from the results that these clusters are located in three unique locations. The distance between clusters 1 and 3 is rather large, as is the distance between them. This demonstrates that there have been significant fluctuations in CO2 emissions in Australia during the last several years.