DATA CLUSTTERING AND FITTING FOR DATASETS

ABSTRACT

- A study regarding two datasets taken in a certain period for selected countries to find and plot clustering
- A study for curve fitting is done using two datasets in a certain period for the countries selected to plot the curve in the data fitted.

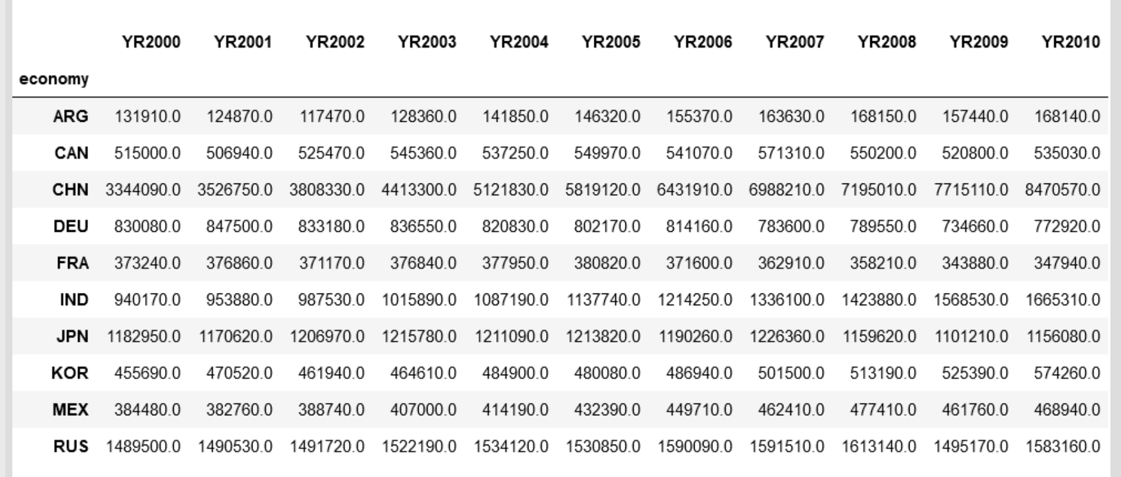
CONTACT

SREEJITH ANILKUMAR
Student ID: 21032113

INTRODUCTION

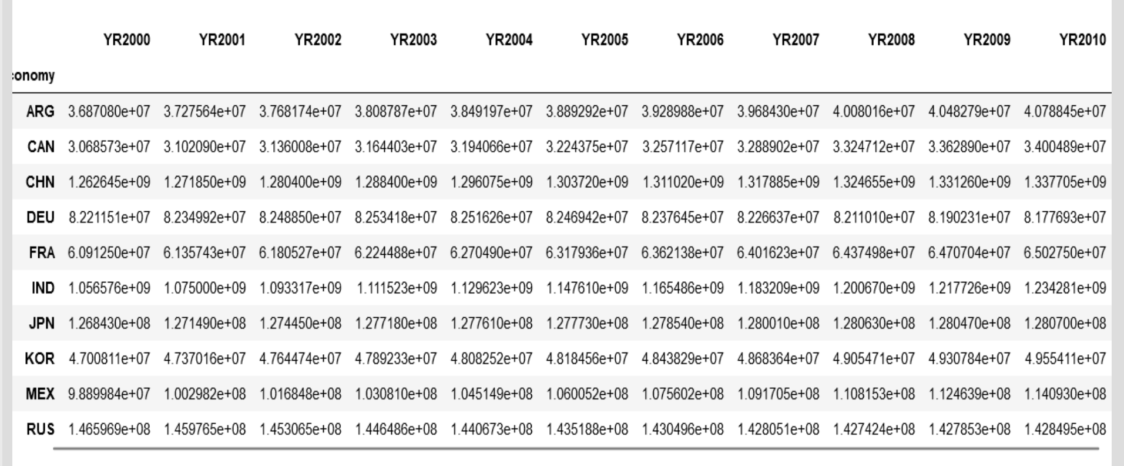
Through this poster we are going through various levels of statistical study using datasets taken from Worldbank data.

At first we have taken the data regarding the total CO2 emissions from a period of 2000-2010 for 10 countries.



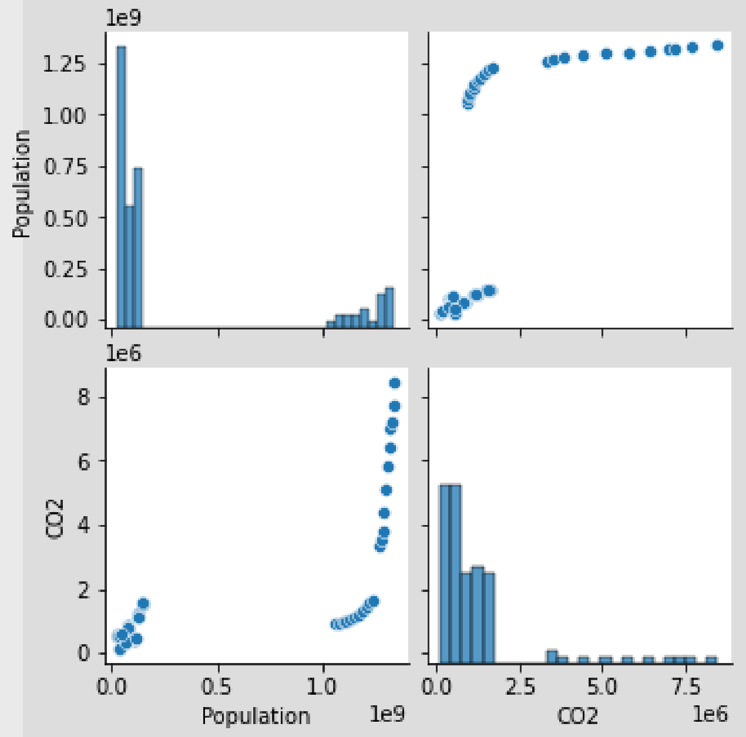
economy	YR2000	YR2001	YR2002	YR2003	YR2004	YR2005	YR2006	YR2007	YR2008	YR2009	YR2010
ARG	131910.0	124870.0	117470.0	128360.0	141850.0	146320.0	155370.0	163630.0	168150.0	157440.0	168140.0
CAN	515000.0	506940.0	525470.0	545360.0	537250.0	549970.0	541070.0	571310.0	550200.0	520800.0	535030.0
CHN	3344090.0	3526750.0	3808330.0	4413300.0	5121830.0	5819120.0	6431910.0	6988210.0	7195010.0	7715110.0	8470570.0
DEU	830080.0	847500.0	833180.0	836550.0	820830.0	802170.0	814160.0	783600.0	789550.0	734660.0	772920.0
FRA	373240.0	376860.0	371170.0	376840.0	377950.0	380820.0	371600.0	362910.0	358210.0	343880.0	347940.0
IND	940170.0	953880.0	987530.0	1015890.0	1087190.0	1137740.0	1214250.0	1336100.0	1423880.0	1568530.0	1665310.0
JPN	1182950.0	1170620.0	1206970.0	1215780.0	1211090.0	1213820.0	1190260.0	1226360.0	1159620.0	1101210.0	1156080.0
KOR	455690.0	470520.0	461940.0	464610.0	484900.0	480080.0	486940.0	501500.0	513190.0	525390.0	574260.0
MEX	384480.0	382760.0	388740.0	407000.0	414190.0	432390.0	449710.0	462410.0	477410.0	461760.0	468940.0
RUS	1489500.0	1490530.0	1491720.0	1522190.0	1534120.0	1530850.0	1590090.0	1591510.0	1613140.0	1495170.0	1583160.0

On the other hand we have to choose the dataset which contains the data of total population over the period 2000-2010 for 10 countries.



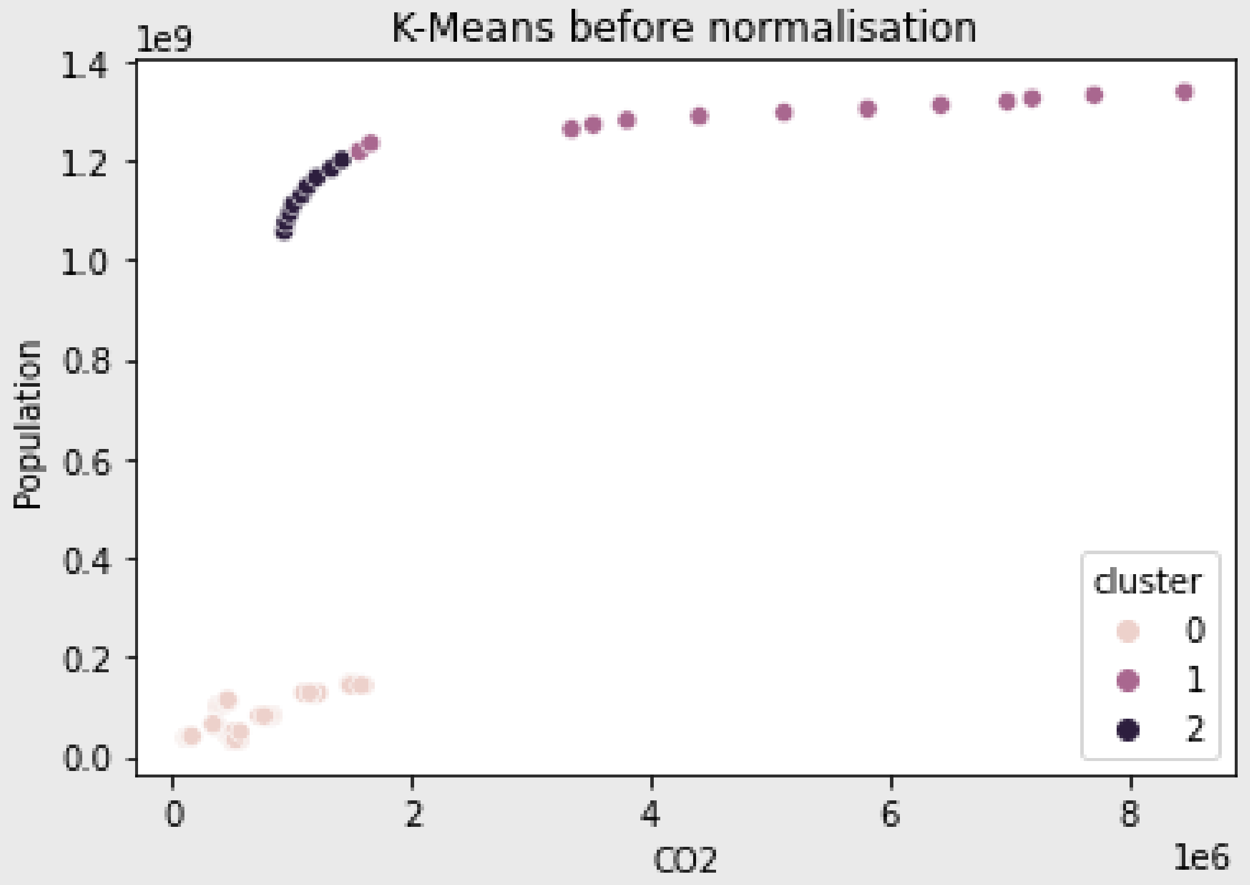
economy	YR2000	YR2001	YR2002	YR2003	YR2004	YR2005	YR2006	YR2007	YR2008	YR2009	YR2010
ARG	3.687080e+07	3.727564e+07	3.768174e+07	3.808787e+07	3.849197e+07	3.889292e+07	3.928988e+07	3.968430e+07	4.008016e+07	4.048279e+07	4.078845e+07
CAN	3.068573e+07	3.102090e+07	3.136008e+07	3.164403e+07	3.194066e+07	3.224375e+07	3.257117e+07	3.288902e+07	3.324712e+07	3.362890e+07	3.400489e+07
CHN	1.262645e+09	1.271850e+09	1.280400e+09	1.288400e+09	1.296075e+09	1.303720e+09	1.311020e+09	1.317885e+09	1.324655e+09	1.331260e+09	1.337705e+09
DEU	8.221151e+07	8.234992e+07	8.248850e+07	8.253418e+07	8.251626e+07	8.246942e+07	8.237645e+07	8.226637e+07	8.211010e+07	8.190231e+07	8.177693e+07
FRA	6.091250e+07	6.135743e+07	6.180527e+07	6.224488e+07	6.270490e+07	6.317936e+07	6.362138e+07	6.401623e+07	6.437498e+07	6.470704e+07	6.502750e+07
IND	1.056576e+09	1.075000e+09	1.093317e+09	1.111523e+09	1.129623e+09	1.147610e+09	1.165486e+09	1.183209e+09	1.200670e+09	1.217726e+09	1.234281e+09
JPN	1.268430e+08	1.271490e+08	1.274450e+08	1.277180e+08	1.277610e+08	1.277730e+08	1.278540e+08	1.280010e+08	1.280630e+08	1.280470e+08	1.280700e+08
KOR	4.700811e+07	4.737016e+07	4.764474e+07	4.789233e+07	4.808252e+07	4.818456e+07	4.843829e+07	4.868364e+07	4.905471e+07	4.930784e+07	4.955411e+07
MEX	9.889984e+07	1.002982e+08	1.016848e+08	1.030810e+08	1.045149e+08	1.060052e+08	1.075602e+08	1.091705e+08	1.108153e+08	1.124639e+08	1.140930e+08
RUS	1.465969e+08	1.459765e+08	1.453065e+08	1.446486e+08	1.440673e+08	1.435188e+08	1.430496e+08	1.428051e+08	1.427424e+08	1.427853e+08	1.428495e+08

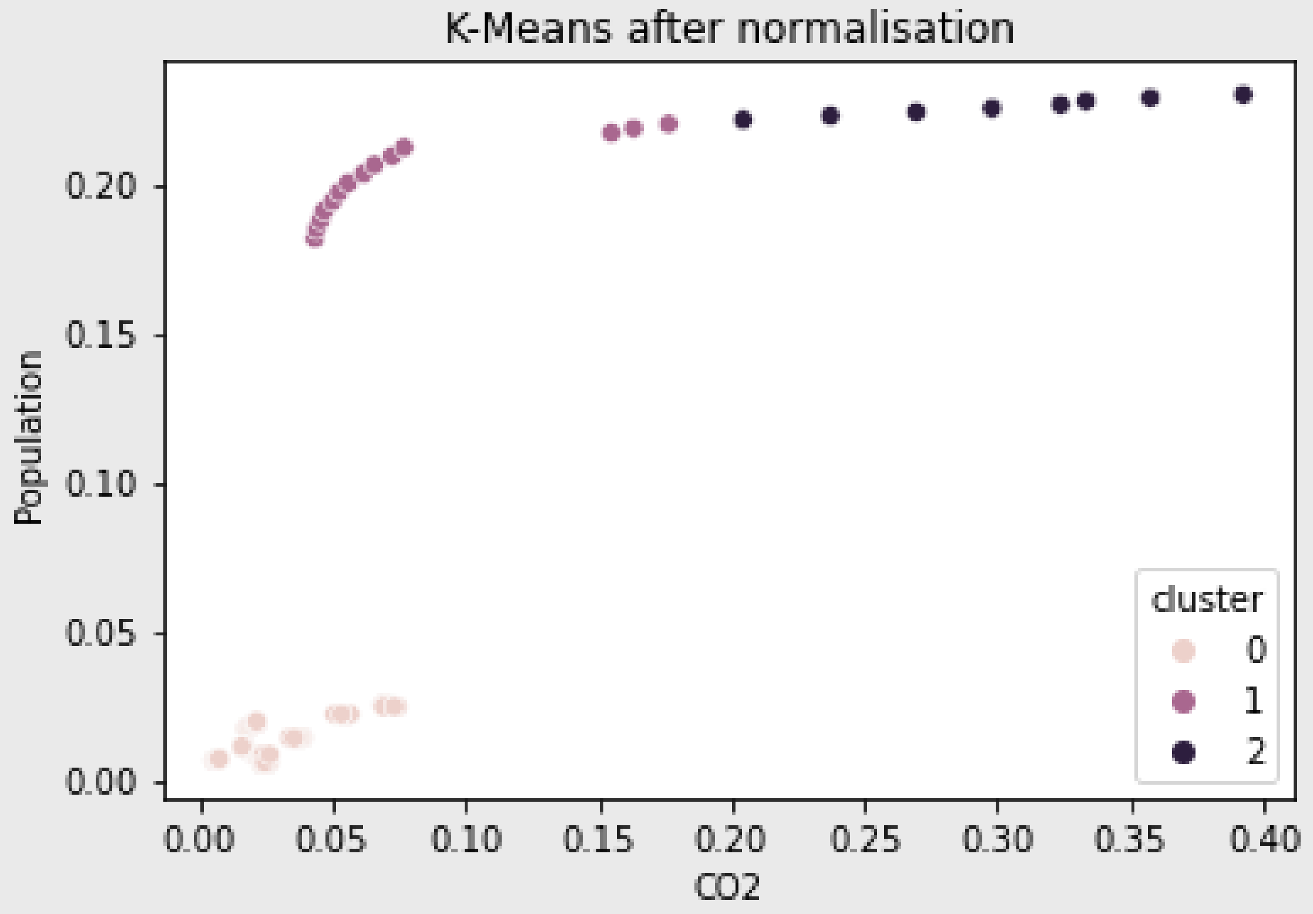
We have simply plotted a pair plot comparing both datasets and given as follow:



Primarly,we are looking onto the clustering using K means.Before that we have to know what is clusttering and what is the use of clustering in statistics.

- **process of making a group of abstract objects into classes of similar objects**.
- **Collection of an array of data points with similar traits and arrange them in a same cluster**.
- Here we have done the clusttering by K means.
- We have imported the libraries and done the whole clusttering process.
- The K means clustering graph before and normalization is given below for further study:

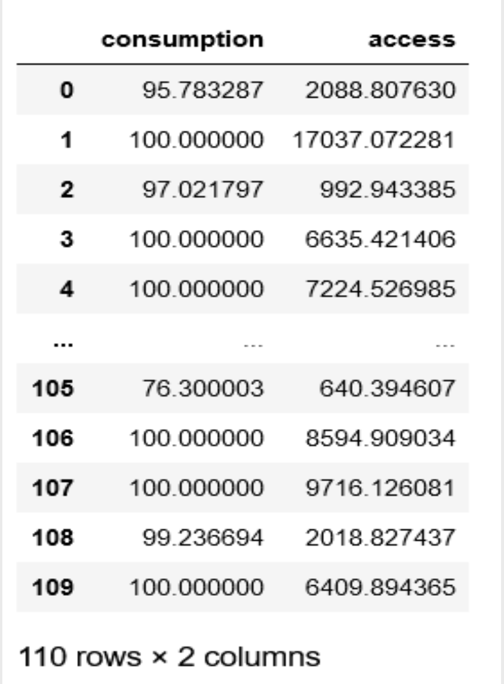




Another interesting feature is data curve fitting.

What is curve fitting?

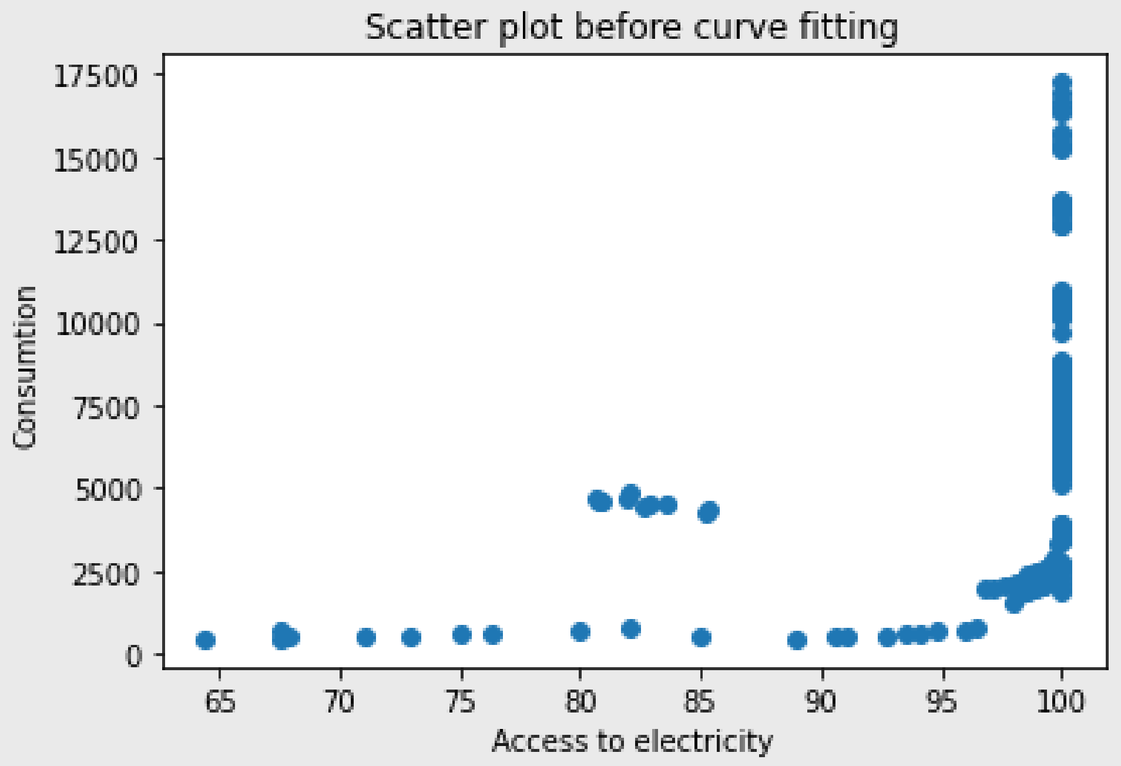
- A process of finding a function that fits perfectly for a data.
- For this fitting we have selected two different datasets.
- One having data containing the total access to electricity and another having data regarding the total energy consumption.
- Clustered data by both the datasets are tabled below:



	consumption	access
0	95.783287	2088.807630
1	100.000000	17037.072281
2	97.021797	992.943385
3	100.000000	6635.421406
4	100.000000	7224.526985
...	...	...
105	76.300003	640.394607
106	100.000000	8594.909034
107	100.000000	9716.126081
108	99.236694	2018.827437
109	100.000000	6409.894365

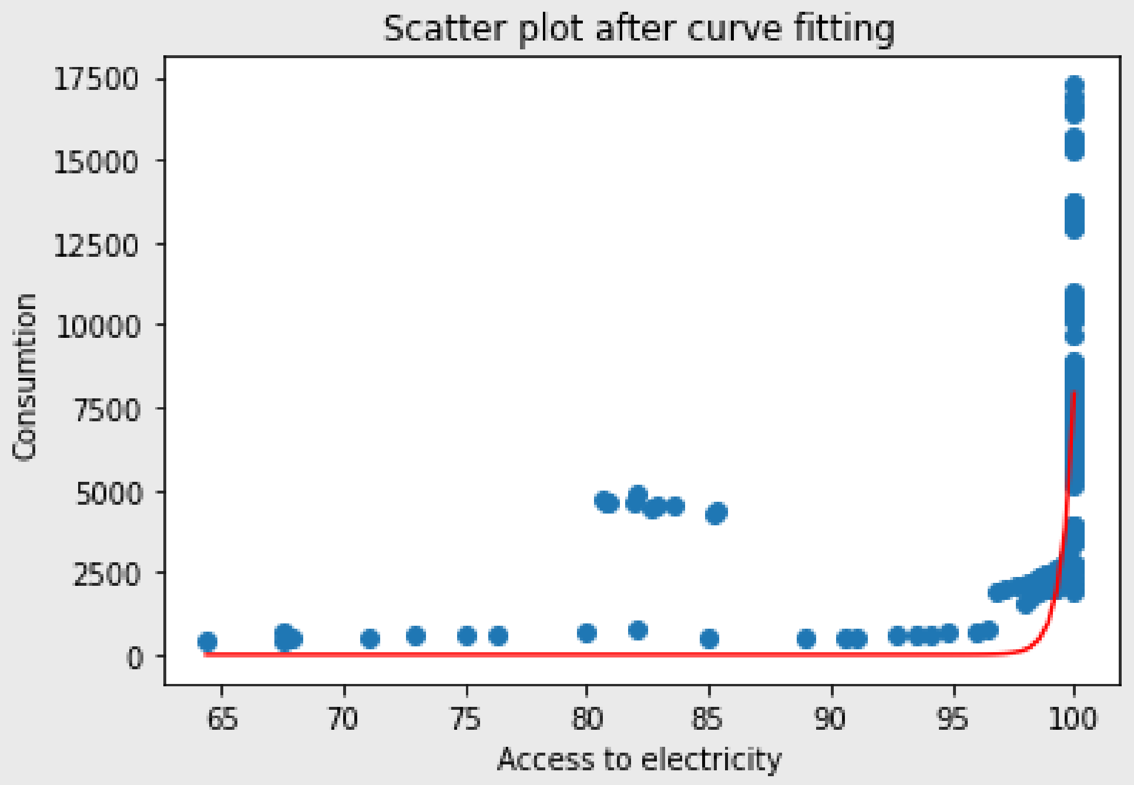
110 rows × 2 columns

The graphs formed shown below:



The process of finding the curve fitting is mentioned below:

- The data is selected for 10 countries.
- Plotted a scattered plot
- An exponential function generated and runned.
- Thus the curve is plotted in the scattered plot.



CONCLUSION

Thus we have gone through total four types of data for total clustering and fitting.The findings are plotted and shown. A research regarding these findings are done and studied.