Green House Gas Emission Data clustering & Data Fitting

Poster(7PAM2000)
Muhammad Saeed
22015028

Abstract: Posters are the typical source of the data visualization and being used in story telling. In this task, Green House Gas (GHG) emission data is chosen from 1990 to 2019. the data is collected from World Bank (https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE) having information about all countries in kilo tons of carbon emission. The dataset is downloaded, cleaned (dropping the unnecessary data, dealing with null values etc.), preprocessed, analyzed (IDA, EDA) for inspection of data patterns. Simple as possible models are created to show the correlation in between years, cluster plots, and curve fitting.

Introduction: Green house gases (GHG) gases are Carbon dioxide, Methane gas, Nitrous oxide, and other industrial gases i.e. HFCs, PFCs, HF6, NF3 etc. are serious concern in environmental reduction and increasing the CO2 emission (https://doi.org/10.1016/j.renene.2021.12.118). GHG have significant contribution in the CO2 emission which is proportional to the climate change and has hazardous affect on environment and living being. From literature (https://doi.org/10.1016/j.worlddev.2020.105317) it is observed that GHG can be controlled by improving the products efficiency and economic complexity.

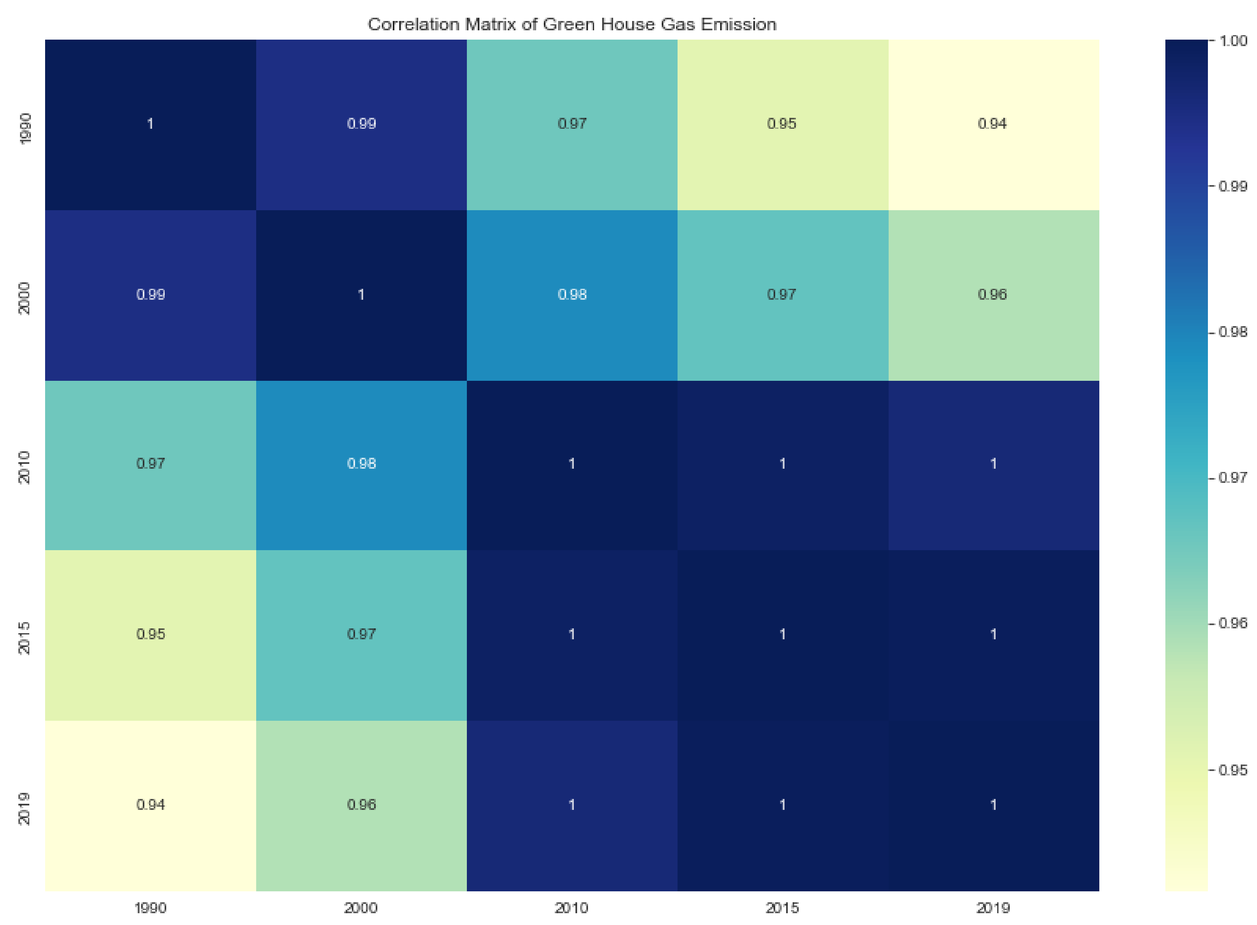
Methodology: This poster design followed this method;

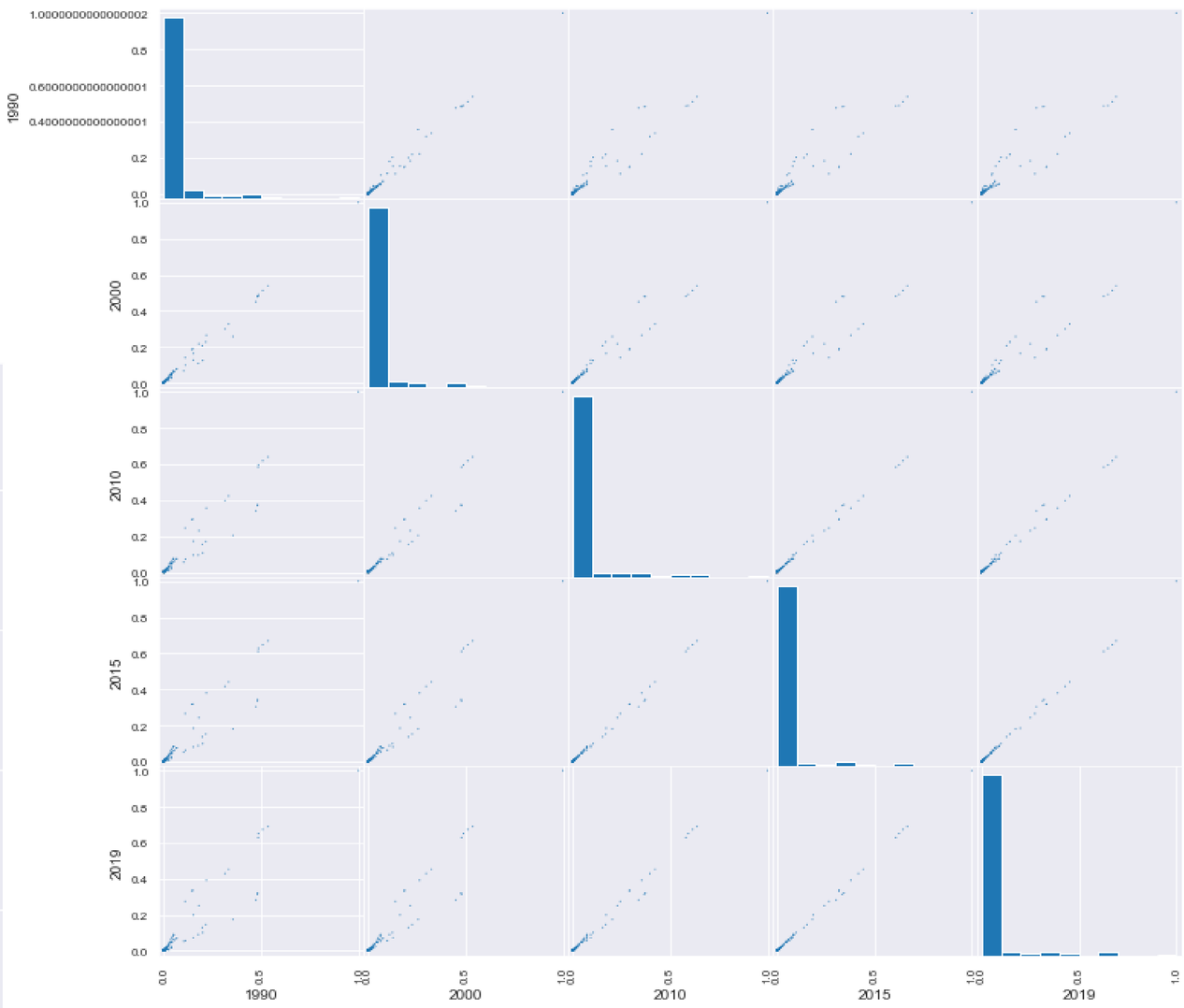
- Data selection – GHG data
- Tool Selection – Anaconda (Python programming language)
- Importing libraries – Numpy, Pandas, Sklearn, matplot, cluster tool, etc.
- Data file reading
- Data Description – count, mean, minimum, maximum, standard deviation, percentiles (25%, 50%, & 75%)
- Data Correlation – in between various years i.e. 1990, 2000, 2010, 2015, & 2019
- Modeling Cluster – finding centers, sizing, K-mean, silhouette score, etc.
- Modeling curve fitting – curve fit optimization, linear functions defining, X, Y axis data assigning, extraction of fitting parameters.

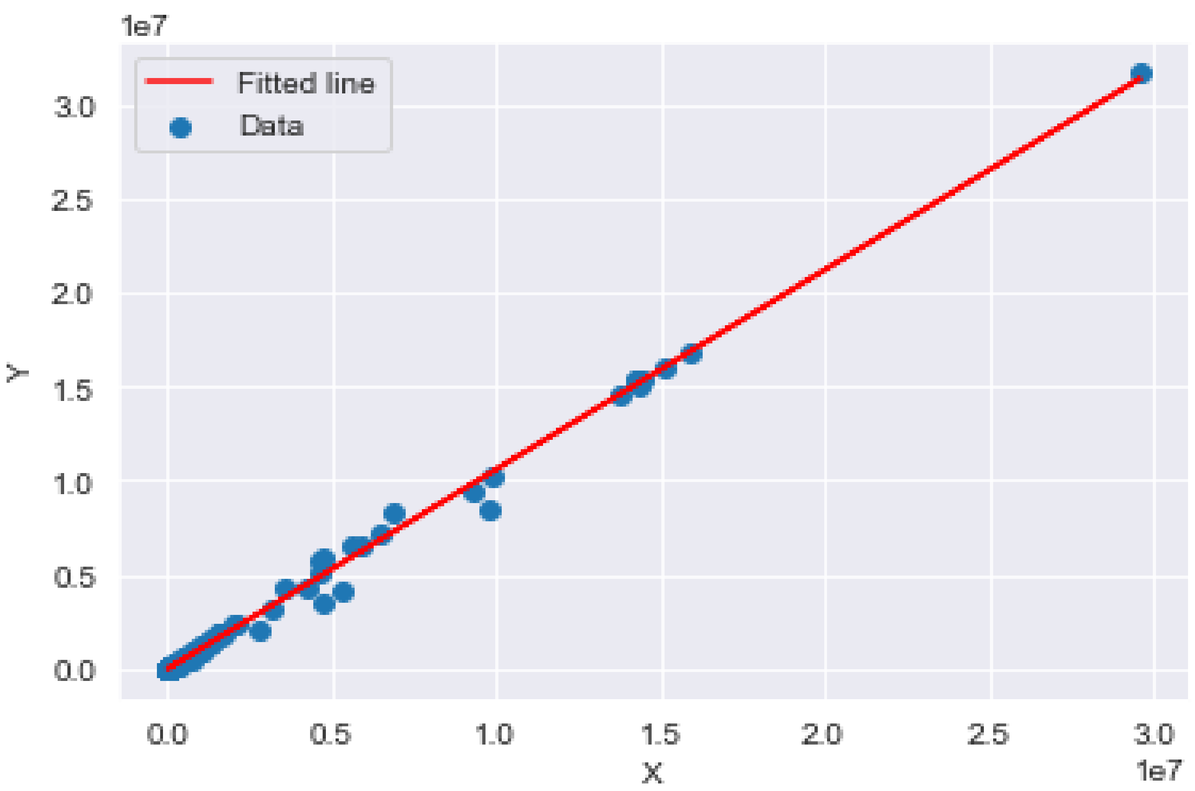
```
df_green = df[['1990', '2000', '2010', '2015', '2019']]
print(df_green.describe())
```

	1990	2000	2010	2015	2019
count	238.000000	239.000000	239.000000	239.000000	239.000000
mean	0.038121	0.037878	0.039761	0.040258	0.040715
std	0.114625	0.113145	0.120793	0.123469	0.125510
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000286	0.000296	0.000290	0.000275	0.000264
50%	0.001409	0.001404	0.001407	0.001224	0.001188
75%	0.012966	0.013403	0.012187	0.010474	0.010138
max	1.000000	1.000000	1.000000	1.000000	1.000000

	1990	2000	2010	2015	2019
1990	1.000000	0.994675	0.965507	0.950840	0.941662
2000	0.994675	1.000000	0.978944	0.966803	0.958492
2010	0.965507	0.978944	1.000000	0.998524	0.996275
2015	0.950840	0.966803	0.998524	1.000000	0.999443
2019	0.941662	0.958492	0.996275	0.999443	1.000000







Summary: In this poster green house gas emission data is selected for analysis and to find meaningful and interesting clusters in the data. The assessment goals i.e. finding the interesting and valuable cluster are found in the GHG data set using cluster plots. The correlation matrix is also printed and plotted which shows the relationship of years. The cluster plot is molded and plotted by defining their parameters. At last the data fitting method is used to plot the data on a fitted line as shown in the last plot. The code file is available at https://github.com/saeed356/ADSAssignment3.git

References:
[1] - https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE
[2] – ttps://doi.org/10.1016/j.renene.2021.12.118
[3] – Lecture 8
[4] - https://doi.org/10.1016/j.worlddev.2020.105317
[5] - https://github.com/saeed356/ADSAssignment3.git

University of Hertfordshire UH