APPLIED DATA SCIENCE

Table of Contents

Introduction

Programming for data analytics of a given dataset is used in this paper to better comprehend data mining and data analysis. As a result, a collection of public data from the World Bank is provided, with a focus on country-by-country indicators connected to climate change, where it is required to perform several data analytic operations in the Jupyter notebook. The dataset was imported, and there is required for utilizing appropriate Python modules. As a result, numerous operations on the dataset are necessary to comprehend and detect variable relationships, among the other variables.

Background

In this report, a dataset is provided where world bank information is contained. Therefore, in the dataset, there are 266 rows and 36 columns where a range of indicators are present relevant to climate change, for example access to electricity, agricultural activity, urban population, etc. Thus, it is required to understand the relationship between different countries' climate changes. Using python language, it is easy to make various comparisons to understand the dependency and relationship between the variables as per the provided dataset for this report.

Data visualization

Data visualization is a common and effective process in Jupyter Notebook to present the dataset in graphically. Therefore, to visualize there is required to perform data preprocessing and data processing on the dataset to make the best fit the dataset in the algorithm. Thus, the dataset is imported and there is required to check the presence of null values in the dataset. Using the ***"df.isnull().sum()"*** code it can be shown the total null value in the dataset. Therefore, in the dataset, there are 266 null values that need b dropped from the dataset. Thus, to drop the null value from the dataset ***"df1 = df.drop"*** is applied to the dataset (Sahoo *et al.* 2019). This whole process is called data preprocessing or data cleaning. After cleaning the dataset there is required to perform the data visualization the dataset.

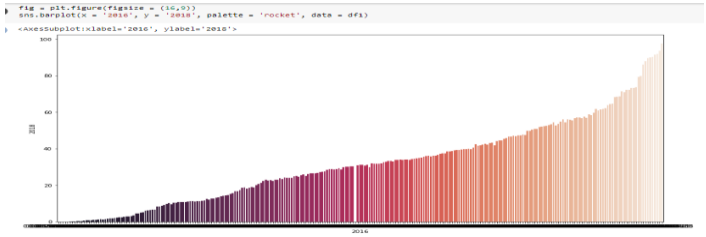

Figure 1: Bar plot between 2016 and 2018
(Source: Self-created)

At first, ***"import seaborn as sns"*** is imported in Jupyter which is a python library for visualizing various plots as per the provided dataset. Therefore, a bar plot is applied to the dataset wherein the x-axis 2016 column is taken and the y-axis is taken as the 2018 column. After that, another bar plot is used between 1990 and 1999 from the dataset as data visualization. However, many other plots are there which are performed in Jupyter on the dataset to understand the relationship and data comparison between various variables as per the provided dataset.

```
plt.figure(figsize=(16,9))
df1['2000'].value_counts().plot.pie()
```

```
<AxesSubplot:ylabel='2000'>
```

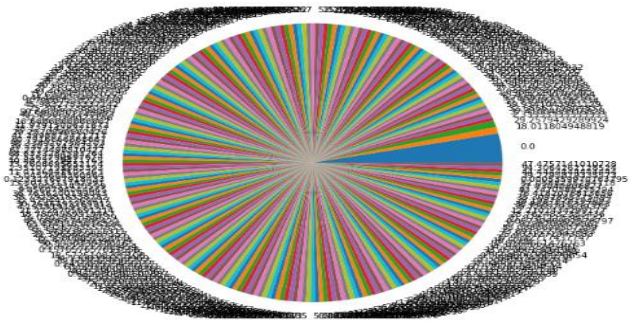

Figure 2: Bar plot

(Source: Self-created)

Data analysis

In this part, data analysis is performed as the correlation between the variables. Using the correlation code as ***"data.corr()"*** in Jupyter it shows the relation between the variables in the dataset (Wette, 2020). Therefore, in the dataset, various information about the energy efficiency and CO2 production, and other parameter are provided and it is noticed from the operations that the amount of CO2 production is increasing day by day.

```
[ ] daf = pd.concat([Mexico_data, UAE_data], axis=0)
```

```
[ ] daf
```

	Country Name	Country Code	Indicator Name	Indicator Code	1990	1991	1992	1993	1994	1995	...	2012	2013	2014	2015	
154	Mexico	MEX	Forest area (% of land area)	AG.LND.FRST.ZS	36.313511	36.199810	36.086108	35.972407	35.858705	35.745004	...	34.310733	34.247728	34.184722	34.121716	34
8	United Arab Emirates	ARE	Forest area (% of land area)	AG.LND.FRST.ZS	3.449732	3.540453	3.631174	3.721895	3.812616	3.903337	...	4.467756	4.467756	4.467756	4.467756	4

2 rows × 36 columns

Figure3: Country based CO2 production

(Source: Self-created)

Therefore, various information about the dataset is recognized through the data analysis such as deforestation in various countries and others.

Conclusion

Several procedures are conducted on the dataset in this report to better understand the variables and their relationships. Data preprocessing is used to remove and replace null values from the dataset in order to achieve the best fit. The data is then processed to change the data type and summary statistics to match the dataset. Then, to illustrate the relationships between the nations and other variables, data visualization is used.

References

Sahoo, K., Samal, A.K., Pramanik, J. and Pani, S.K., 2019. Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(12), p.2019.

Wette, K., 2020. SWIGLAL: Python and Octave interfaces to the LALSuite gravitational-wave data analysis libraries. SoftwareX, 12, p.100634.