APPLIED DATA SCIENCE

Introduction

The project is based on exploring the statistics of the World Bank's dataset. The data that is explored is regarding public data from the World Bank. Moreover, apart from World Bank data there is also data regarding climate change which varies in each country. The data is manipulated using certain libraries. Exploring information regarding the dataset is performed. Graphical representation is done for better viewing of the data.

Implementation

Various libraries such as numpy, pandas, seaborn as well as kurtosis are imported for analyzing and graphical representation (Bilogur *et al.* 2018). The dataset that is used in this project is "API_19_DS2_en_csv_v2_3931355.csv" which consists of several information regarding country name, country code, name of indicator related to population, health, climate, etc and indicator code. The dataset is loaded as well as arranged according to the CSV file. The null values are removed with certain specified values as this method is meant to return a new data frame object (Manzini *et al.* 2021). The rows along with columns are dropped using the drop function. The missing values are figured out and using a code it returns the missing values number within the dataset. The entire information regarding the data frame is printed and it is found that there are all float values. The dataset contains 8 rows as well as 61 columns.

Statistical analysis

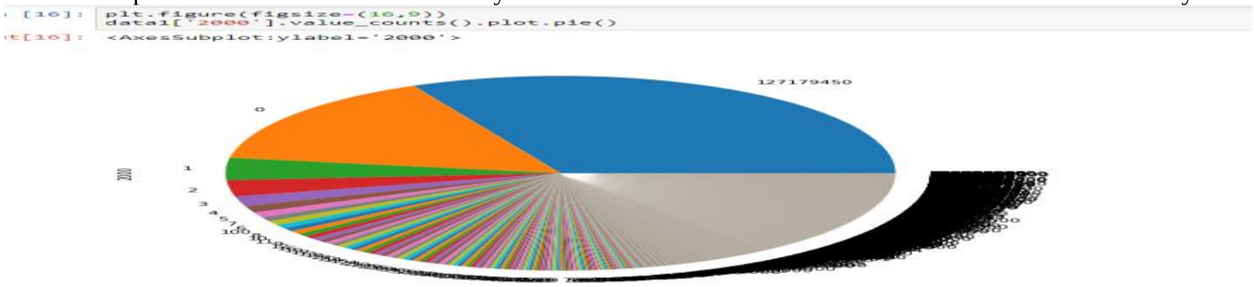
The analysis is done by changing the data type from categorical value to numeric value as machine learning is only possible with numerical values. Therefore, object type of data has been converted into numeric data. Cat codes are applied on country name, country code, and indicator name as well as indicator code. Finally all the data types are now converted to numerical values. The total number of rows as well as columns is 20216 and 65 respectively. Correlation among the indicators is performed and no variation is observed between any of the trends within the time. The relationship of correlation is neutral as there is no relationship within the change of variable. Moreover, the performance regarding some algorithms also reduces two variables that are related barely. This is termed as multicollinearity. The library of kurtosis is imported, which is the 4th central moment that is divided by its square variance. Pylab kutosis type of library functions have been used to perform the required statistical analysis upon the World Bank dataset.

Exploring of correlation

The correlation in this task is performed by three possibilities which state the strength as well as direction of linear association amongst two variables that are quantitative (Narechania *et al.* 2020). The variables are denoted by r which measures between -1 to +1. The positive value denotes the value from +1 to 0.

Visualization

Historical representation is done on the country name which states the details about all the names of the country.



```
[16]: plt.figure(figsize=(16,9))
      data1['2000'].value_counts().plot.pie()
t[16]: <AxesSubplot:ylabel='2000'>
```

Figure 1- Pie chart representation

(Source: Acquired from Jupyter Notebook)

Pie plot is represented on the country name. Similarly the representation of indicators is also performed.

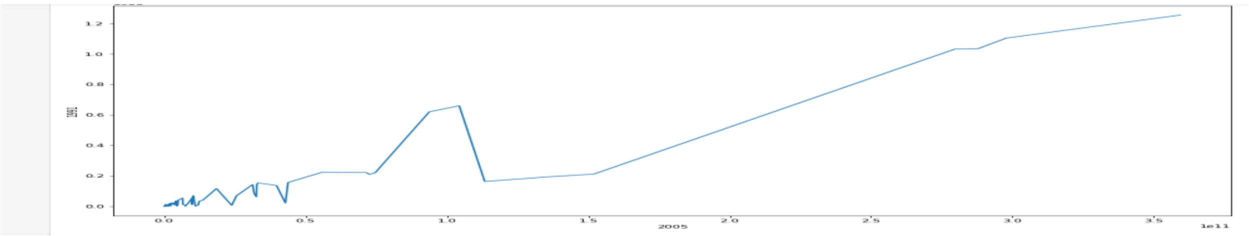
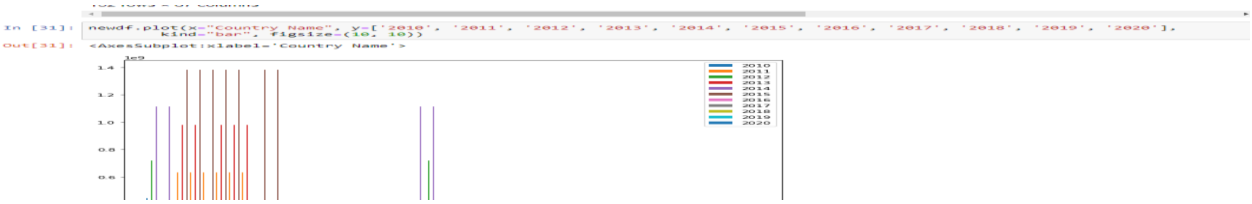


Figure 2- Line plot representation

(Source: Acquired from Jupyter Notebook)

Visualization through heat map is done for the dataset. Line plot is performed in the year of 2017, 2018, 2019 as well as 1965.



```
In [31]: newdf.plot(x="Country Name", y=['2010', '2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019', '2020'],
                    kind="bar", figsize=(10, 10))
Out[31]: <AxesSubplot:xlabel='Country Name'>
```

Figure 3- Bar plot representation

(Source: Acquired from Jupyter Notebook)

The ARuba country and UAE country data have been used to represent the land that is covered by the forest areas. Arabia country comprises of more urban population and thus forest area is less. The forest land is high in Aruba country.

Conclusion

The project concludes that various representations of the World Bank are performed by using the dataset. No spaghetti code is applied in the project. Adherence to "the guidelines of PEP-8" are done which states in utilization of "inline comments sparingly". The usage of inline comments over the same line as the statement is referred to. The inline comments are separated by using two or sometimes more than two spaces within the statement. The inline comments are started with # as well as a single space similar to block comments.

Reference List

Bilogur, A., 2018. Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), p.547.

Manzini, S., Busnelli, M., Colombo, A., Franchi, E., Grossano, P. and Chiesa, G., 2021. reString: an open-source Python software to perform automatic functional enrichment retrieval, results aggregation and data visualization. Scientific reports, 11(1), pp.1-15.

Narechania, A., Srinivasan, A. and Stasko, J., 2020. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. IEEE Transactions on Visualization and Computer Graphics, 27(2), pp.369-379.

SRIVASTAVA, S., GARG, A., SEHGAL, A. and KUMAR, A., 2018. Analysis and Comparison of Loan Sanction Prediction Model using Python. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), 8, pp.1-8.