Assignment 2: Statistics and trends

Name: Adeel

Student Id :22021887

Git Hub link: https://github.com/muhammadadeel22/ADS2_assognment2.git

Abstract:

World urban population is increasing year by year as it is obvious from the dataset which comprises of data from year 1960 to 2021. The dataset used for this study is taken from world bank website. Statistical parameters like five number summary was used to take an insight into dataset. Furthermore after summing up two six years' urban population data. I have found a positive correlation of 0.88.

Findings:

I have chosen one of the dataset from World Bank data available online. I then written a function in python and read that csv file into it. Since I was getting error in reading that file. So I opened the file and remove unbalanced rows in beginning of the file. After reading the csv in that function it will return 3 dataframes d1 with county name and indicator other d2 with country name and country code.

I also used another function of statistics named "pearsonr" from Scipy.stats library to see whether there is any correlation between each five years i.e. 1980 to 1985 and 1986 to 1991.What I come to know is that there is a positive correlation of 0.88 in both of the datasets

```
PearsonRResult(statistic=0.88
33012212389648,
pvalue=0.01963327016259317)
```

The third data frame named d3 was of my interest and to take further insight in this dataset. Since in original data frame the years are used as columns from 1960 to 2021. So I use melt function to convert the dataset columns to rows values.

```
df3=df.melt(id_vars=['Country
Name', 'Country Code',
'Indicator Name', 'Indicator
Code']

        var_name="Year",

        value_name="Value")
```

Then I have further extracted data from 1980 to 1985 six year data into data frame d4. Similarly in data frame d5 I have extracted the data from year 1985 to 1991.Since in this data frame before extraction year wise data I have to change the year data which was string so I changed the year column as Int using as type.

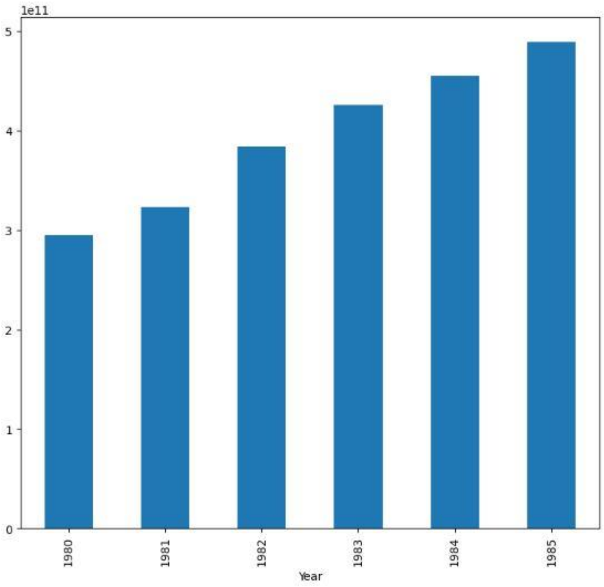
```
d3.Year = d3.Year.astype(int)
```

I then have dealt with missing values of both the data frames using function fillna(0).Then to get an insight in these two dataframes I used describe function which will give me five number summary of every five year data. Then I further summed up year wise data of both these data frames.

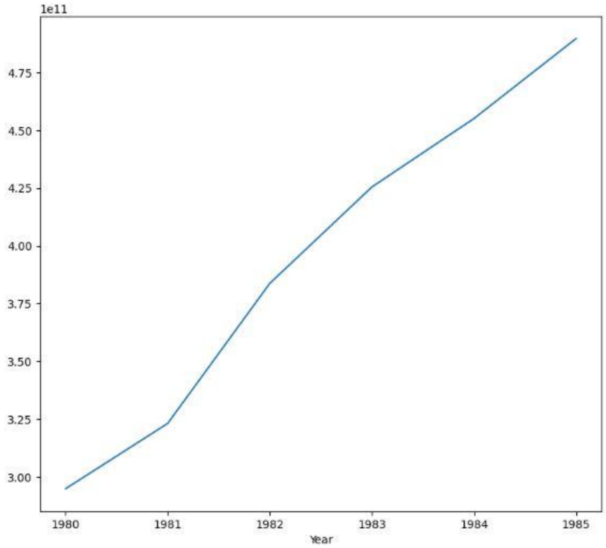
```
                Year          Value
count  121296.000000  1.212960e+05
mean     1982.500000  1.955467e+07
std         1.707832  5.735851e+08
min      1980.000000 -2.096742e+05
25%      1981.000000  0.000000e+00
50%      1982.500000  0.000000e+00
75%      1984.000000  9.643784e-01
max      1985.000000  5.424800e+10
```

```
                Year          Value
count  121296.000000  1.212960e+05
mean     1988.500000  3.967037e+07
std         1.707832  1.515153e+09
min      1986.000000 -1.023697e+06
25%      1987.000000  0.000000e+00
50%      1988.500000  0.000000e+00
75%      1990.000000  6.429411e+00
max      1991.000000  1.560820e+11
```

As I calculated the correlation between these two Data Frames. I used Pyplots to further demonstrate it efficiently. As you can see in the first graph of years ranging from (1981-1985) have a symmetric rate of increase in the population.

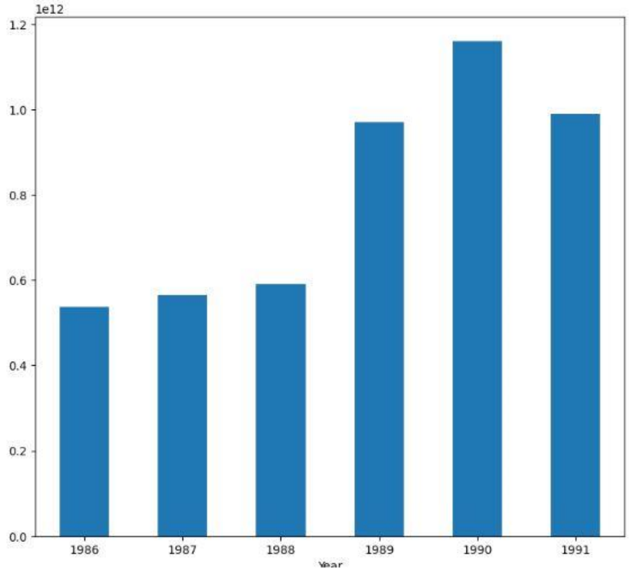


Then I used the time series to see the symmetric rate of increase

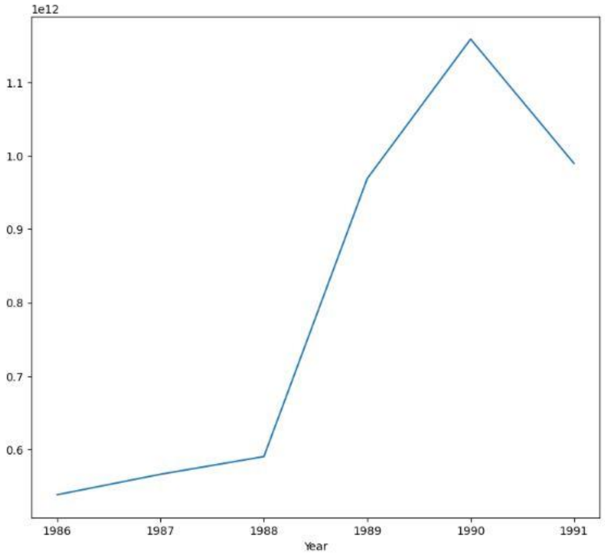


This shows that after the year 1981 the growth in population increase with 2x speed. After this year it start increase gradually.

For Year (1986-1991) we can see that there is a moderate change in population till 1988. After it there is 80% increase in the population in the year 1989 and almost 100% in the year 1990. After 1990 there is a 10% decrease in the population growth.



After the bar plot I try creating a line plot which can show a complete time series of the increase in value w.r.t the years.


1e12
Year