Neeraj Ankam
Ne9891078

STA-5104 REPORT

Dataset Description:

1. The dataset was not a pre-existing dataset and had to be constructed manually.
2. The dataset consists of nine indicators of development for the three most populous countries-China, India and the USA for the years 2001,2006,2011,2016; they were selected from 'data.worldbank.org'.
3. The indicators are the following: Gross Domestic Product(GDP)(current), GDP per capita, GDP per capita (annual growth), Unemployment rate, Military expenditure, Population growth, Central government debt and Life expectancy.
4. The dataset had some missing values; the missing values had to be obtained from two sources namely 'www.statistica.com' &'www.tradingeconomics.com'.
5. The snapshot below gives an idea of the dataset.

Country N	Country C	Series Nar	2001 [YR2001]	2006 [YR2006]	2011 [YR2011]	2016 [YR2016]
China	CHN	GDP (curre	1.3394E+12	2.75213E+12	7.57255E+12	1.11991E+13
China	CHN	GDP per ca	1053.11	2099.23	5633.80	8123.18
China	CHN	GDP per ca	7.56	12.09	9.01	6.12
China	CHN	Unemploy	3.60	4.10	4.10	4.04
China	CHN	Military e:	2.08	2.01	1.82	1.92
China	CHN	Populatio	0.73	0.56	0.48	0.54
China	CHN	Populatio	1271850000.00	1311020000.00	1344130000.00	1378665000.00
China	CHN	Central go	23.00	25.00	33.60	46.20
China	CHN	Life expec	72.44	74.32	75.45	75.50
India	IND	GDP (curre	4.78965E+11	9.20317E+11	1.82305E+12	2.26352E+12
India	IND	GDP per ca	447.01	792.03	1461.67	1709.39
India	IND	GDP per ca	3.02	7.58	5.25	5.88
India	IND	Unemploy	4.24	4.41	3.56	3.42
India	IND	Military e:	3.02	2.61	2.65	2.47
India	IND	Populatio	1.73	1.55	1.31	1.15
India	IND	Populatio	1071477855.00	1161977719.00	1247236029.00	1324171354.00
India	IND	Central go	59.82	60.96	51.56	69.50
India	IND	Life expec	62.99	64.97	67.03	68.35
United Sta	USA	GDP (curre	1.06218E+13	1.38559E+13	1.55179E+13	1.85691E+13
United Sta	USA	GDP per ca	37273.62	46437.07	49790.67	57466.79
United Sta	USA	GDP per ca	-0.02	1.68	0.85	0.91
United Sta	USA	Unemploy	4.73	4.62	8.95	4.87
United Sta	USA	Military e:	2.94	3.81	4.58	3.29
United Sta	USA	Populatio	0.99	0.96	0.75	0.69
United Sta	USA	Populatio	284968955.00	298379912.00	311663358.00	323127513.00
United Sta	USA	Central go	51.99	55.29	90.18	99.77

Importance:

The dataset gives an idea of how China, India and the United States of America have grown since the beginning of the 21st century. It also gives an idea about the relative growth of the countries.

Tools used for Analysis:

1. Python- Data Loading, Analysis and Plots

   I have used Python to load the data into a Python data-frame and conduct statistical analysis on it. The following code snippet shows how I have imported the data into Python:

```
import pandas as pd
import numpy as np
analysis = pd.read_excel('Countries.xlsx',index_col=0)
```

```
china = analysis.iloc[0:9,:]
#print(china)
china_gdp_pc=china.iloc[1,2:6]
#print(china_gdp)
#print(type(china))

india = analysis.iloc[9:18,:]
india_gdp_pc = india.iloc[1,2:6]
#print(india)

usa=analysis.iloc[18:27,:]
usa_gdp_pc=usa.iloc[1,2:6]
```

2. Excel- Calculations and Visualizations.

   I had to manually calculate some of the values. For example, Unemployment rate was given as a percentage of the total population in the dataset and had to be calculated using the functions of Excel to do further analysis. Military Expenditure was also calculated as a percentage of the GDP from the dataset.

   The bar chart for Military Expenditure and the Unemployment statistics table were done using Excel.

Neeraj Ankam
Ne9891078

Plots:

The two GDP and the correlation matrix visualization have been prepared using Python. The code snippet below shows the implementation:

Code for GDP plots:

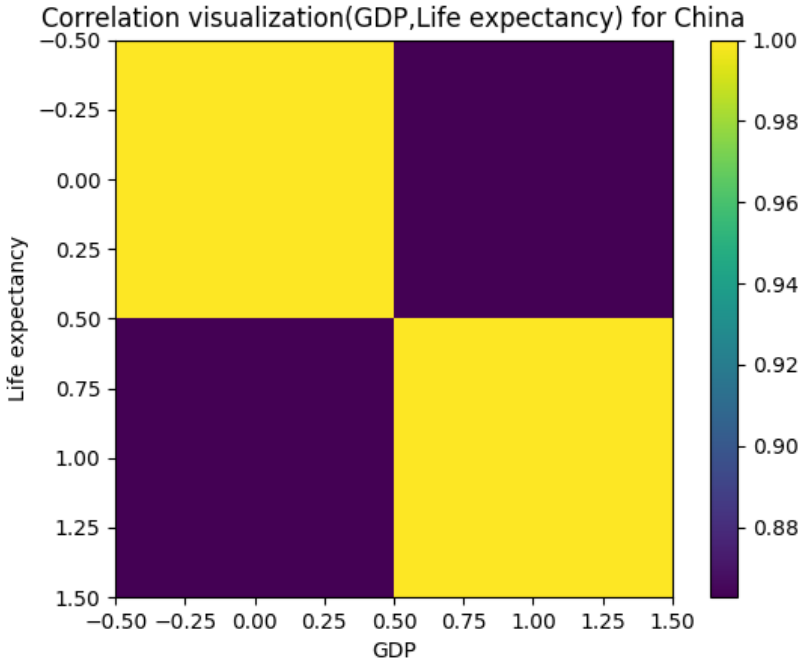
```
import matplotlib.pyplot as plt

fig=plt.figure()
ax=fig.add_subplot(111)

gdp_ticks=[2001,2006,2011,2016]
us, =plt.scatter(usa_gdp_pc,gdp_ticks,label='USA',marker='*')
for xy in zip(usa_gdp_pc,gdp_ticks):
    ax.annotate('%s%s'%xy,xy=xy,textcoords='data')

ch, =plt.plot(china_gdp_pc,gdp_ticks,label='CHINA',marker='o')
#for xy in zip(china_gdp_pc,gdp_ticks):
#   ax.annotate('%s%s'%xy,xy=xy,xytext=(30,40),textcoords='data')
ind, =plt.plot(india_gdp_pc,gdp_ticks,label='INDIA',marker='>')
#for xy in zip(india_gdp_pc,gdp_ticks):
#   ax.annotate('%s%s'%xy,xy=xy,xytext=(0,2),textcoords='data')


#plt.xscale('log',basex=10)
plt.legend([us,ch,ind],['USA','CHINA','INDIA'],loc='best')
plt.xlabel('GDP per capita(in USD)')
plt.ylabel('Year')
plt.title('Countries GDP per capita WRT YEAR')
plt.grid(True)
plt.show()
```

Some of the average values and tables, visualizations have been done using the pre-built functions and graphs of Excel.

Inferences:

- The data from the dataset shows that China and India's GDP respectively have been growing at a rapid pace when compared to the GDP of the US. After a thorough research, it was found that one of the crucial factors contributing to the rapid development was public investment. Also, the tax cuts and the recession of 2007 contributed to the slow growth of the US.

- The military expenditure of the US seems high, which is misleading if only the numbers are taken into consideration. Military benefits for the US soldiers are a lot in comparison to the soldiers from India and China. Also, the wages and the cost of maintenance of each soldier is significantly higher.
- The unemployment rate in the US seems relatively high when compared to China and India because the US has distinct categories of unemployment rates such as U1, U2, U3 et cetera. The figure is misleading because it does not consider some people, such as the set of people who are not looking for a job and hence unemployed.
- The Central Government debt of the US is surprisingly high and in 2016, it is 104 % of the GDP of previous year.
- The correlation between GDP per capita and Life expectancy was calculated using numpy module of Python. It shows that there is a strong correlation between both.
- **Numpy:** It is the fundamental package for scientific computing with Python



The color of the scale gives us the value of the correlation coefficient.

```
china_LE_arr=[china_LE[0],china_LE[1],china_LE[2],china_LE[3]]
china_GDP_arr=[1.3394E+12,2.75213E+12,7.57255E+12,1.11991E+13]

china_core=np.corrcoef(china_GDP_arr,china_LE_arr)
```

Code for Correlation Coefficient Visualization:

```
import matplotlib.pyplot as plt
plt.imshow(china_core,interpolation='nearest')
plt.xlabel('GDP')
plt.ylabel('Life expectancy')
plt.title('Correlation visualization(GDP,Life expectancy) for China')
plt.colorbar()
plt.show()
```

Future Work:

- More countries can be added to the dataset and can be compared to give a sense of the relative growth.
- More indicators of development can be added and correlation between some of them can be calculated to see if one affects the other.