Exploratory Data Analysis

ON DESCRIPTIVE STATISTICS OF CERTAIN SPECIFIC PARAMETERS (Infant Mortality Rate, Birth Rate, Death Rate, GDP per Capita) FOR THE WORLD IN COMPARISION TO INDIA FOLLOWING A DETAILED DRILL DOWN ON INDIAN STATES

- Manas Kumar Sarkar(A22022)
- Ashutosh Pandey (A22010)
- Atish Kumar Majee (A22012)
- Soumya Chatterjee (A22046)

Under the guidance of

Dr. Sayantani Roy Choudhury

STS-I Project

DS Fall Batch 2022

Motivation Behind the Project:

The aim of this project study is to explore the distribution and relation of infant mortality rate, death rate & birth rate with GDP per capita. With the help of exploratory data analysis, we intend to gain insights on the aforementioned parameters for the world, and compare them with India to see where it stands in the developing world. We further take a dig onto these data of Indian states and compare the data to get a holistic view of these measures across different geographies of our country.

Definitions:

- The infant mortality rate is **the number of infant deaths for every 1,000 live births**. In addition to giving us key information about maternal and infant health, the infant mortality rate is an important marker of the overall health of a society.

- Birth rate is **the number of individuals born in a population in a given amount of time**. Human birth rate is stated as the number of individuals born per year per 1000 in the population. For example, if 35 births occur per year per 1000 individuals, the birth rate is 35.

- Mortality rate, or death rate, is **a measure of the number of deaths (in general, or due to a specific cause) in a particular population, scaled to the size of that population, per unit of time**. Human death rate is stated as the number of individuals born per year per 1000 in the population.

- Gross Domestic Product (GDP) per capita **shows a country's GDP divided by its total population**

Data Collection:

The Data used for the project is for the year 2019. The data is collected in the form of secondary data from web sources which are listed below:

1) Infant Mortality Rate: https://data.worldbank.org/indicator/SP.DYN.IMRT.IN
2) Birth Rate:
3) Death Rate:
4) GDP per capita:
5) Birth Rate, Death Rate, Infant Mortality rates for Indian States: Economic Survey 2021-22
6) GPD per capita for Indian States: Wikipedia

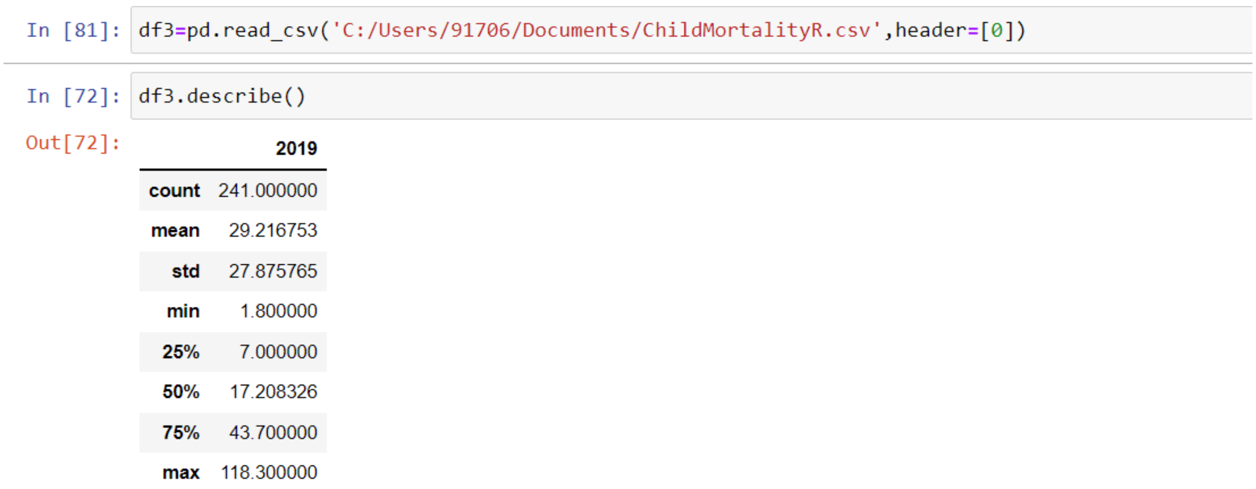
Analysis:

TOOLS USED:

- Microsoft Excel
- Python- Jupyter Notebook
- Power BI
- Tableau

The World with respect to India:

1) Infant Mortality rate:

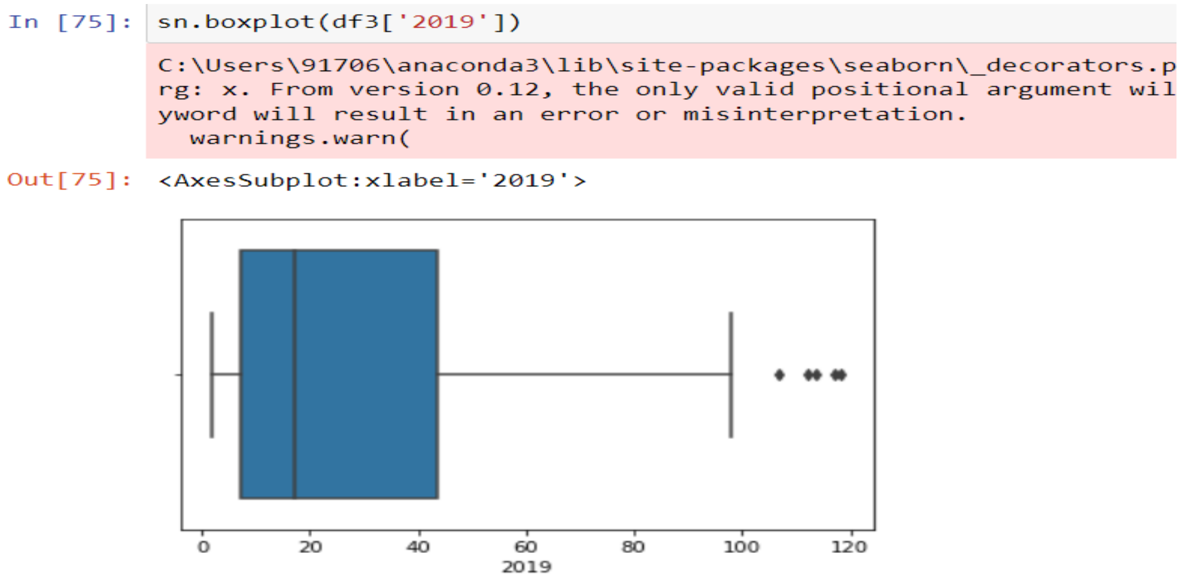


```
In [81]: df3=pd.read_csv('C:/Users/91706/Documents/ChildMortalityR.csv',header=[0])
```

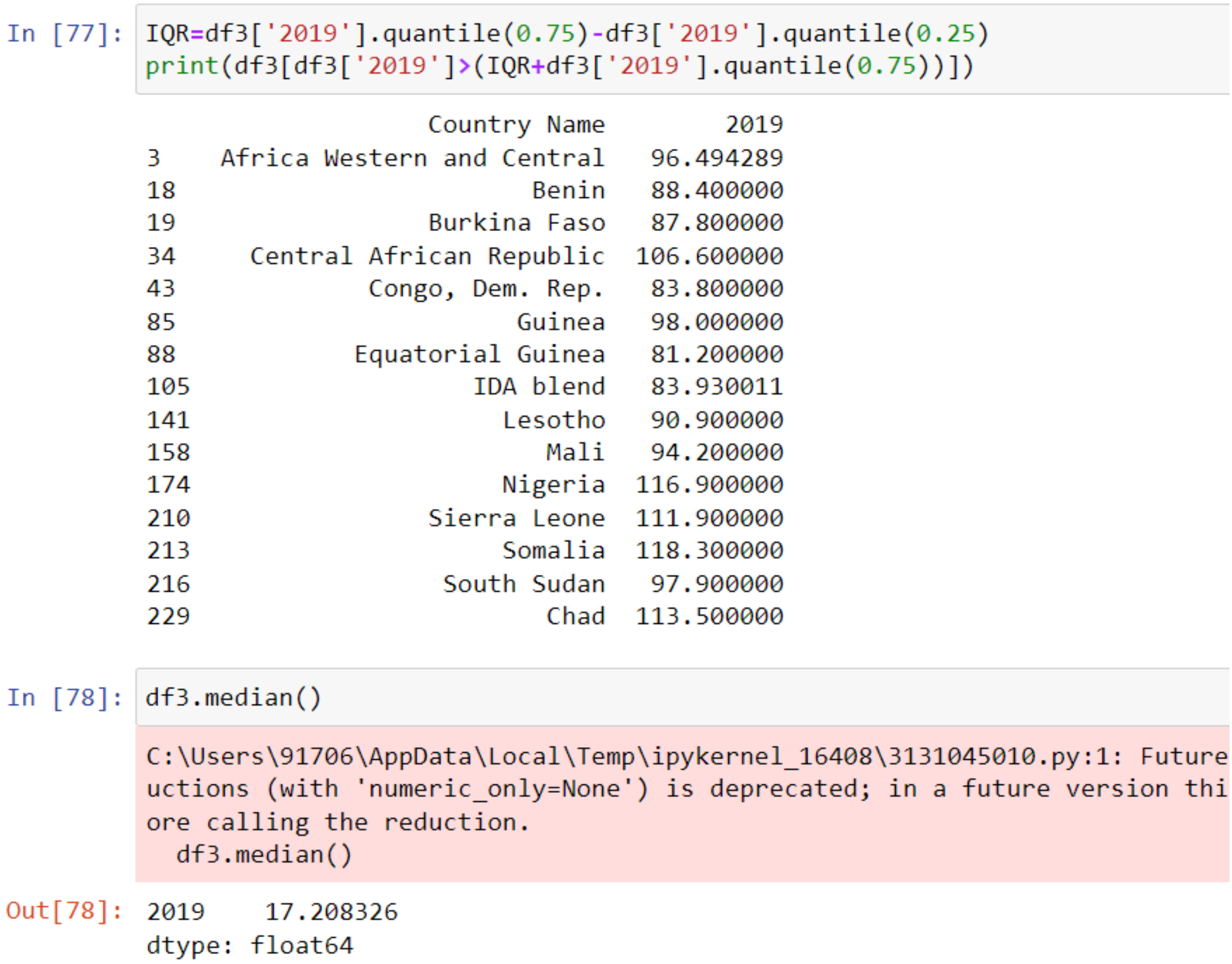
```
In [72]: df3.describe()
```

Out[72]:

	2019
count	241.000000
mean	29.216753
std	27.875765
min	1.800000
25%	7.000000
50%	17.208326
75%	43.700000
max	118.300000



```
In [75]: sn.boxplot(df3['2019'])

C:\Users\91706\anaconda3\lib\site-packages\seaborn\_decorators.p
rg: x. From version 0.12, the only valid positional argument wil
yword will result in an error or misinterpretation.
  warnings.warn(

Out[75]: <AxesSubplot:xlabel='2019'>
```



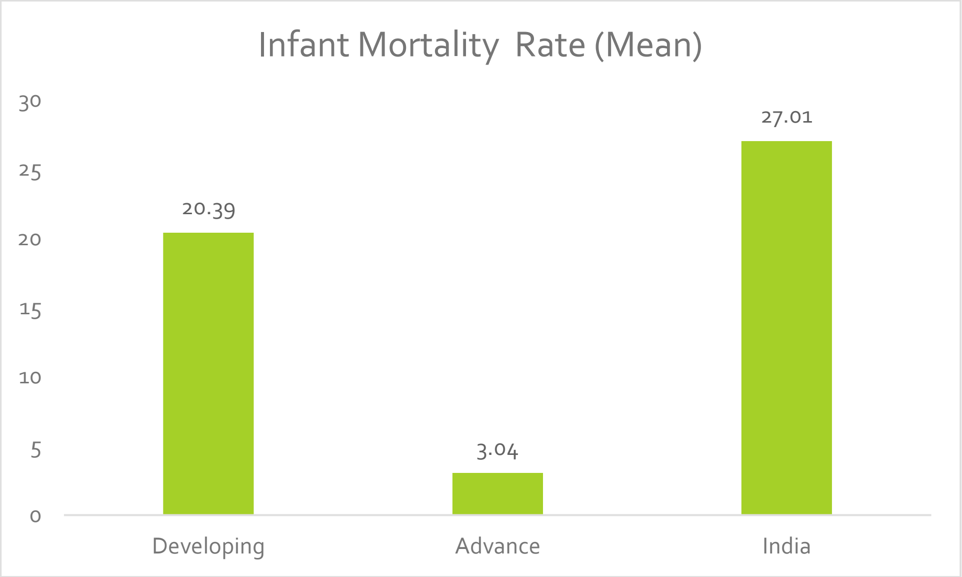
```
In [77]: IQR=df3['2019'].quantile(0.75)-df3['2019'].quantile(0.25)
         print(df3[df3['2019']>(IQR+df3['2019'].quantile(0.75))])
```

```
                   Country Name        2019
3     Africa Western and Central   96.494289
18                         Benin   88.400000
19                  Burkina Faso   87.800000
34      Central African Republic  106.600000
43              Congo, Dem. Rep.   83.800000
85                        Guinea   98.000000
88             Equatorial Guinea   81.200000
105                    IDA blend   83.930011
141                      Lesotho   90.900000
158                         Mali   94.200000
174                      Nigeria  116.900000
210                 Sierra Leone  111.900000
213                      Somalia  118.300000
216                  South Sudan   97.900000
229                         Chad  113.500000
```

```
In [78]: df3.median()
```

```
C:\Users\91706\AppData\Local\Temp\ipykernel_16408\3131045010.py:1: Future
uctions (with 'numeric_only=None') is deprecated; in a future version thi
ore calling the reduction.
  df3.median()
```

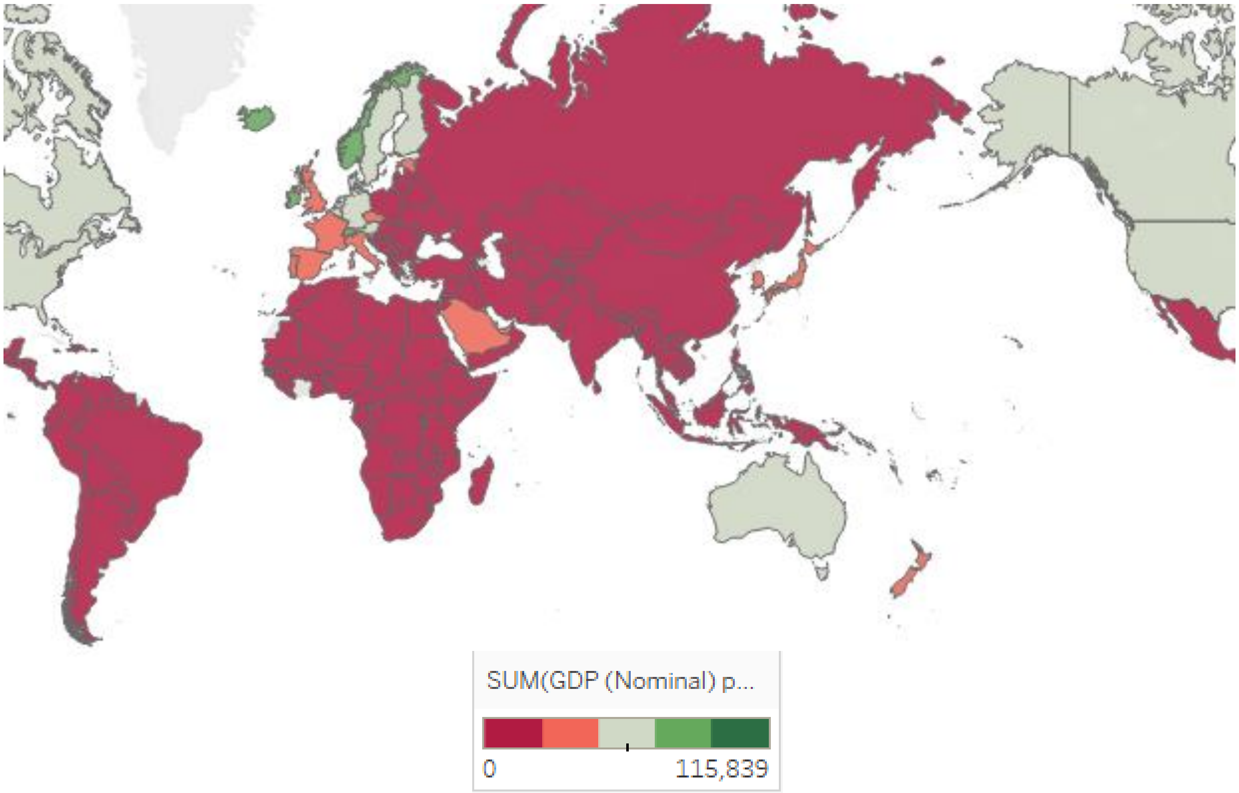
```
Out[78]: 2019    17.208326
         dtype: float64
```

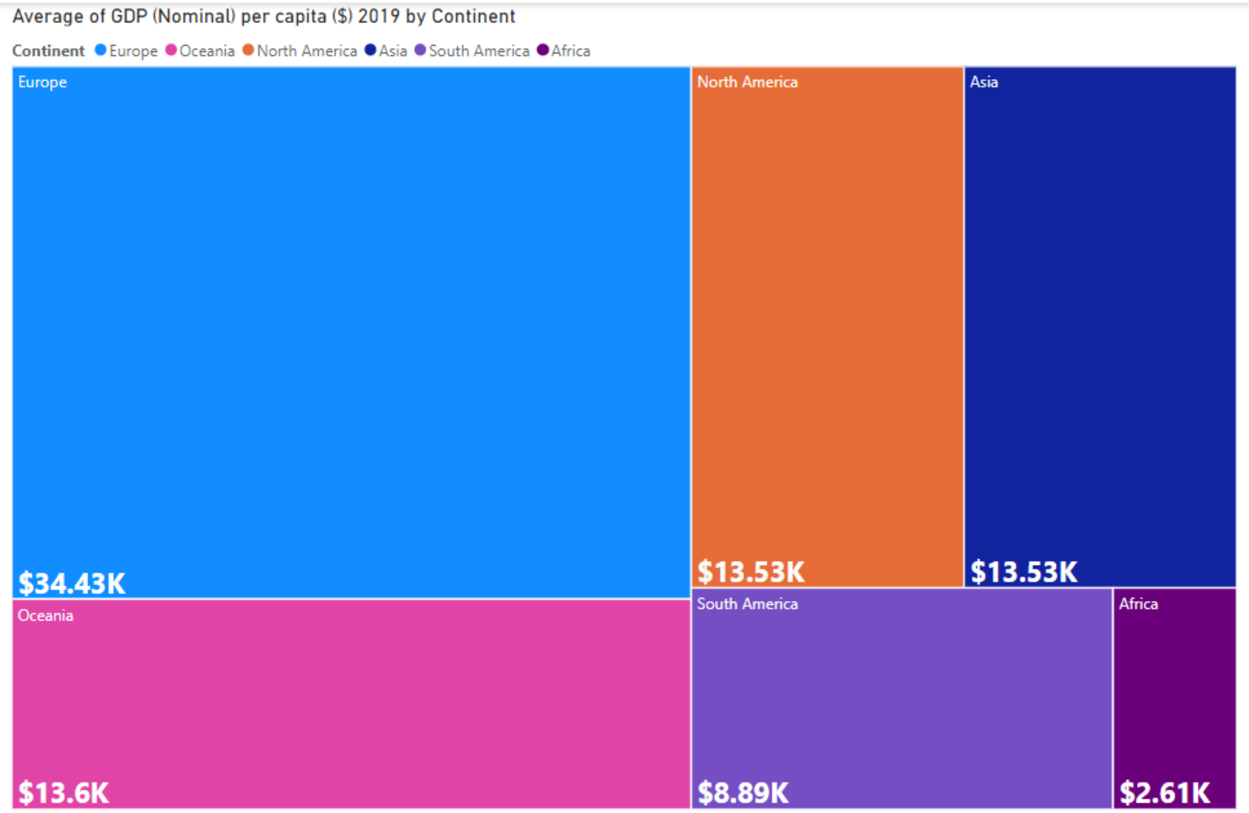


The infant mortality rate is rightly skewed with many outliers. The infant mortality rate for the world is 29.1 and median is 17.20. Average infant mortality rate for India is 27.01 which is lesser than the world average but still higher than the median infant mortality rate of the world.

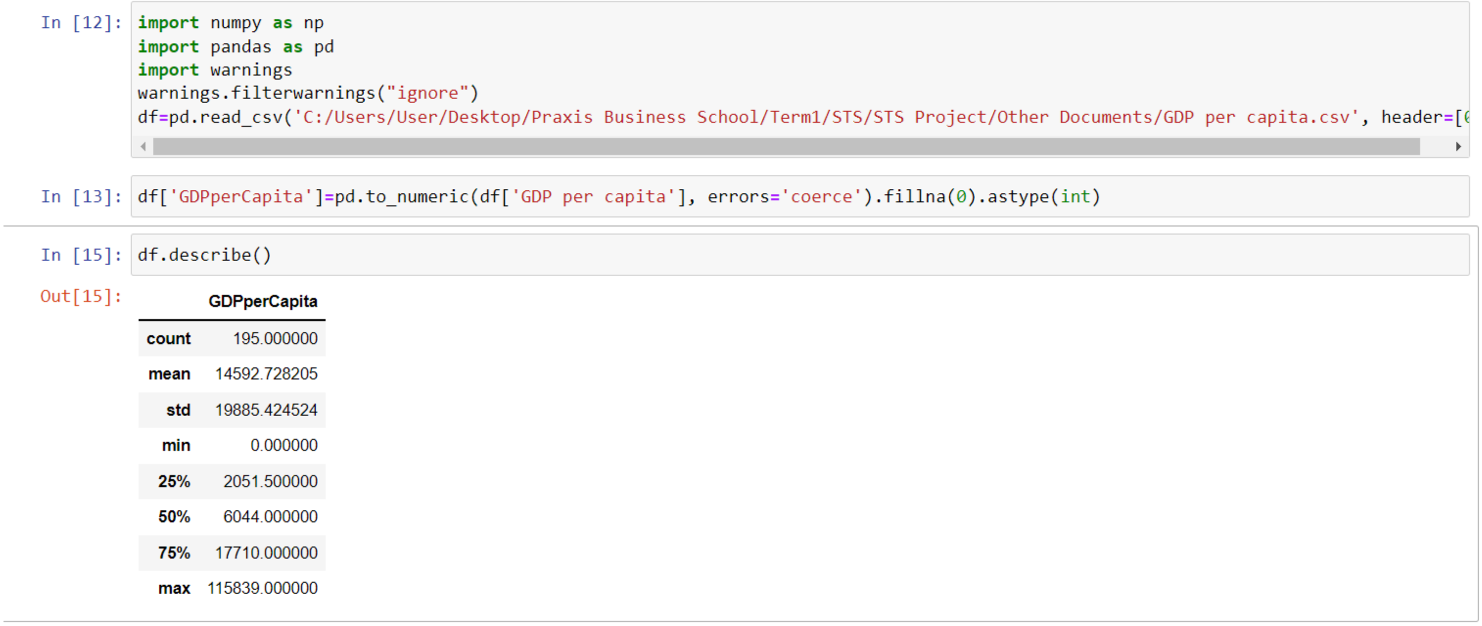
Among the developing countries, the average infant mortality rate is much higher than the average of the developing countries. It is a matter of concern & we need to emphasize on infant healthcare.

2) GDP per capita





The above plots give the distribution of GDP per capita for the world and the average GDP across the continents.



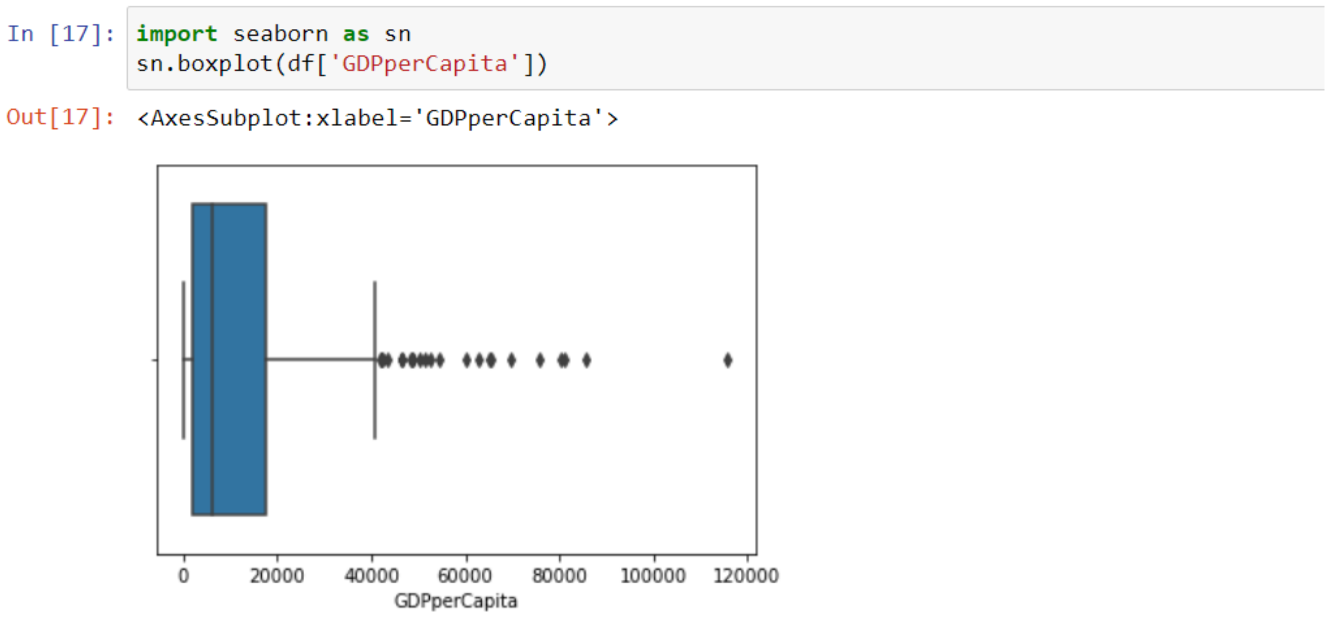
```
In [12]: import numpy as np
         import pandas as pd
         import warnings
         warnings.filterwarnings("ignore")
         df=pd.read_csv('C:/Users/User/Desktop/Praxis Business School/Term1/STS/STS Project/Other Documents/GDP per capita.csv', header=[
```

```
In [13]: df['GDPperCapita']=pd.to_numeric(df['GDP per capita'], errors='coerce').fillna(0).astype(int)
```

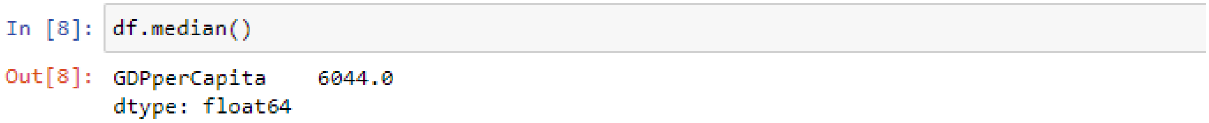
```
In [15]: df.describe()
```

Out[15]:

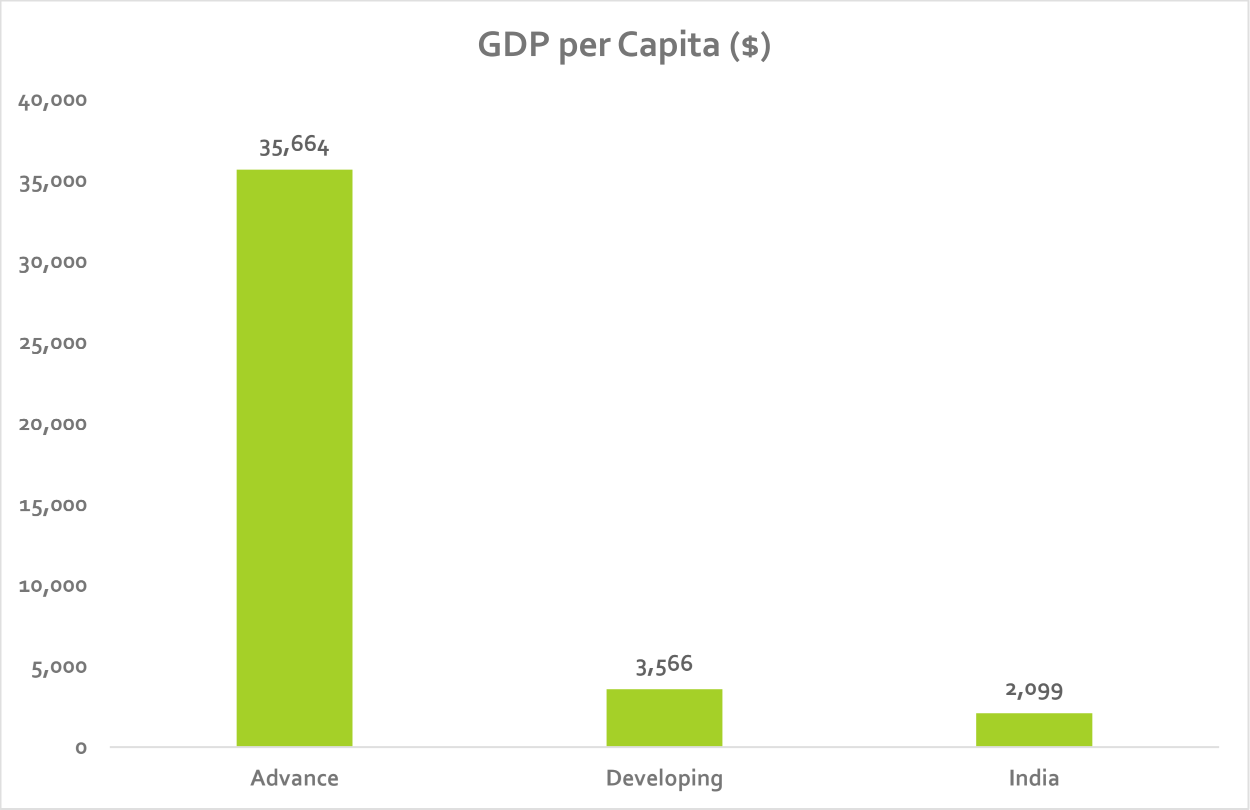
	GDPperCapita
count	195.000000
mean	14592.728205
std	19885.424524
min	0.000000
25%	2051.500000
50%	6044.000000
75%	17710.000000
max	115839.000000



```
In [17]: import seaborn as sn
         sn.boxplot(df['GDPperCapita'])

Out[17]: <AxesSubplot:xlabel='GDPperCapita'>
```

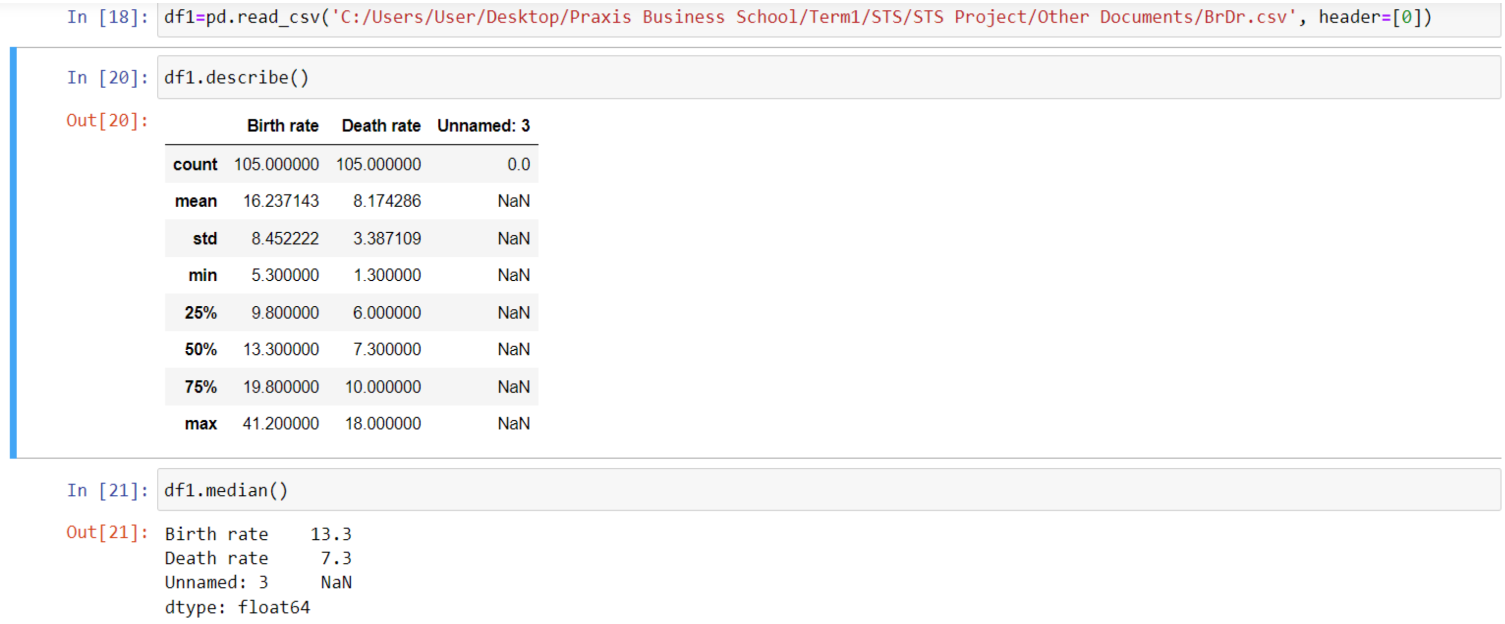


```
In [8]: df.median()

Out[8]: GDPperCapita    6044.0
        dtype: float64
```



The distribution of GDP per capita across the world is rightly skewed & has several outliers which are mostly the advanced countries. The mean GDP per capita for the world is 14592.72 dollars whereas the median GDP is 6044 dollars. Per capita GDP for India is 2099 dollars which is very low even in the developing countries. It is even less than the average GDP per capita of Asia.

3) Birth Rate & 4) Death Rate



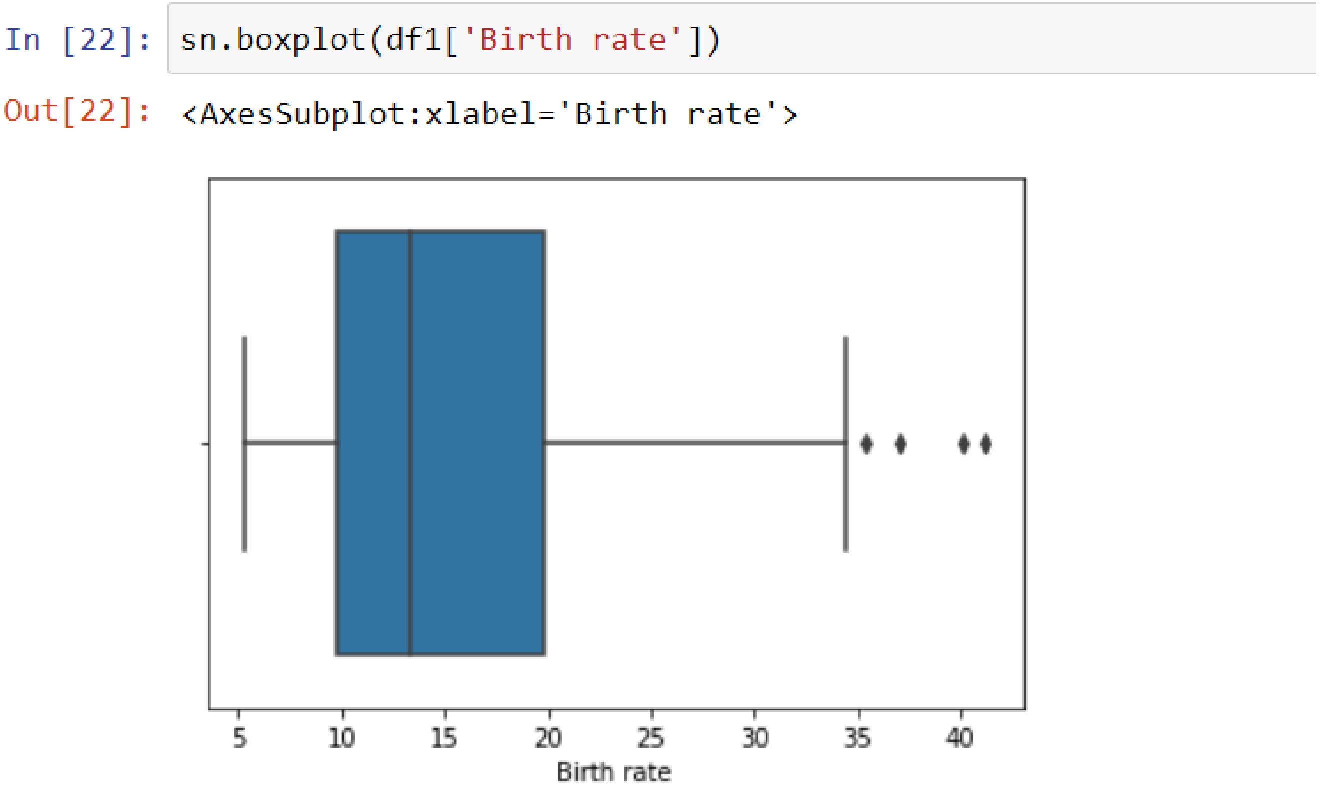
```
In [18]: df1=pd.read_csv('C:/Users/User/Desktop/Praxis Business School/Term1/STS/STS Project/Other Documents/BrDr.csv', header=[0])
```

```
In [20]: df1.describe()
```

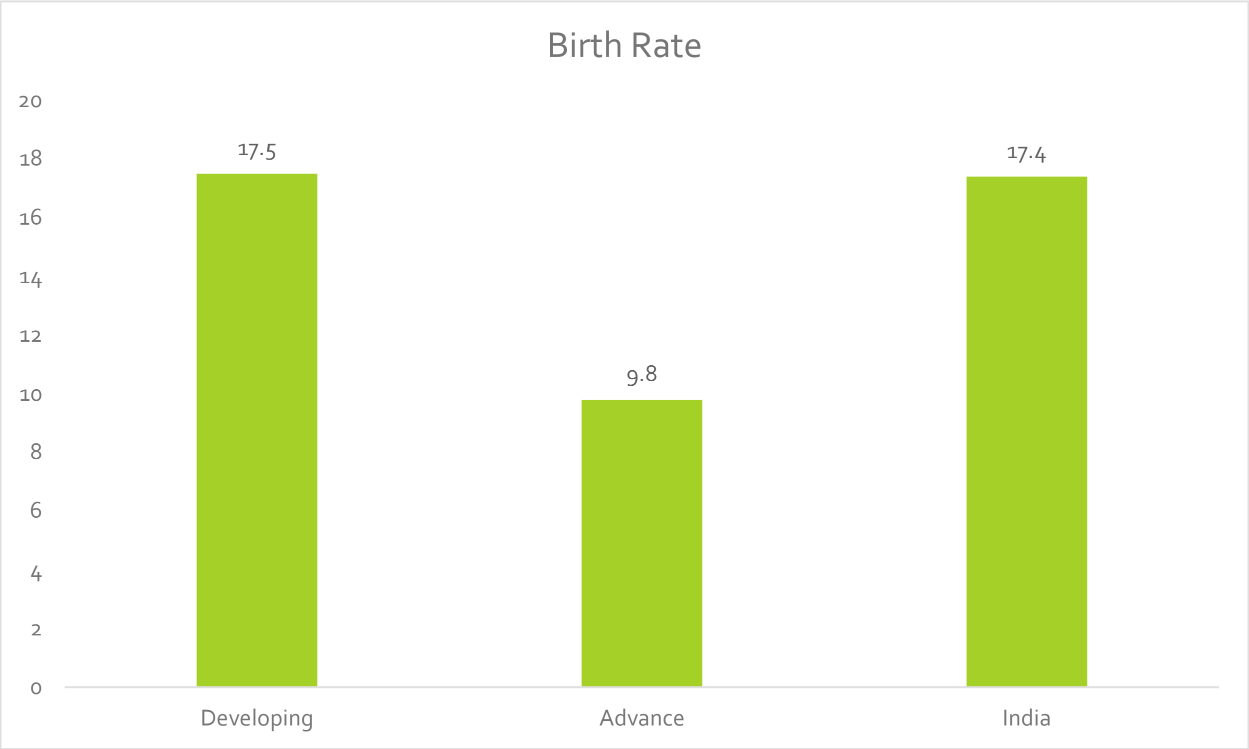
Out[20]:

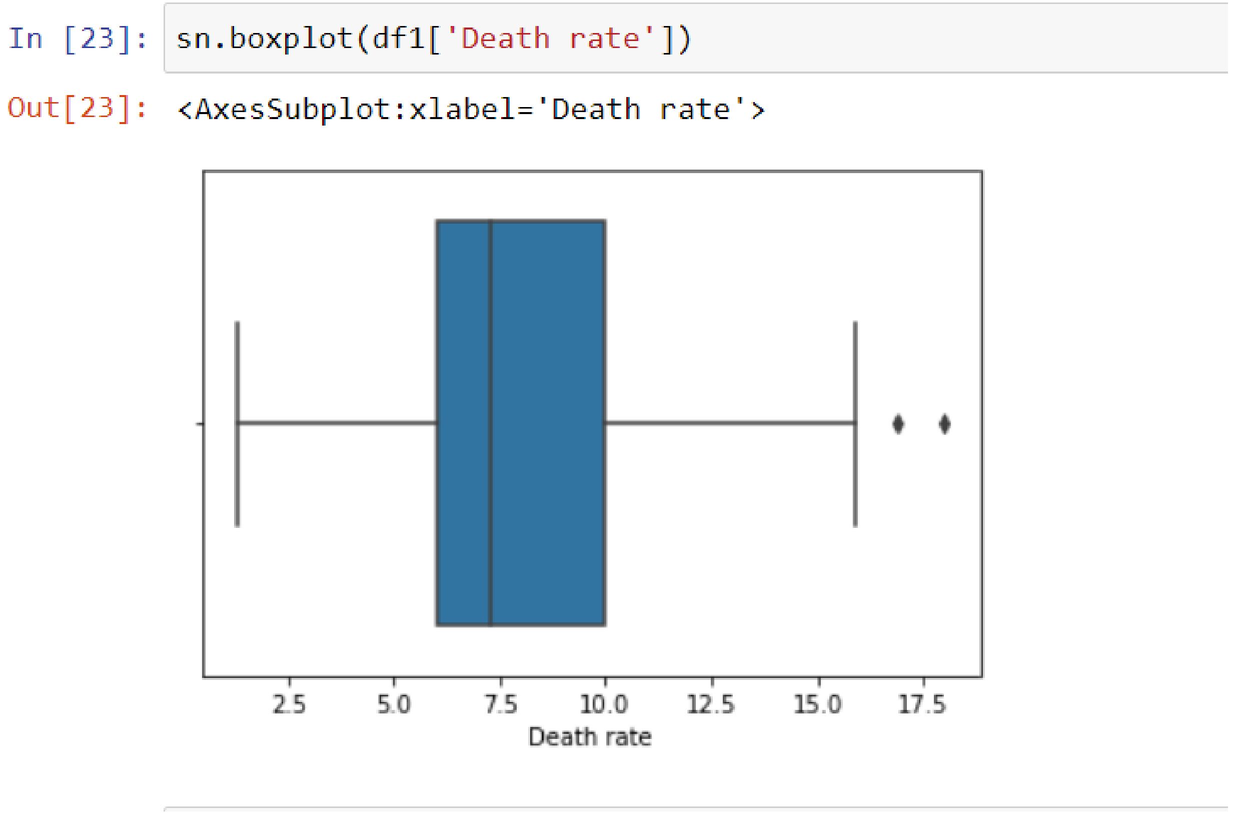
	Birth rate	Death rate	Unnamed: 3
count	105.000000	105.000000	0.0
mean	16.237143	8.174286	NaN
std	8.452222	3.387109	NaN
min	5.300000	1.300000	NaN
25%	9.800000	6.000000	NaN
50%	13.300000	7.300000	NaN
75%	19.800000	10.000000	NaN
max	41.200000	18.000000	NaN

```
In [21]: df1.median()

Out[21]: Birth rate     13.3
         Death rate      7.3
         Unnamed: 3      NaN
         dtype: float64
```



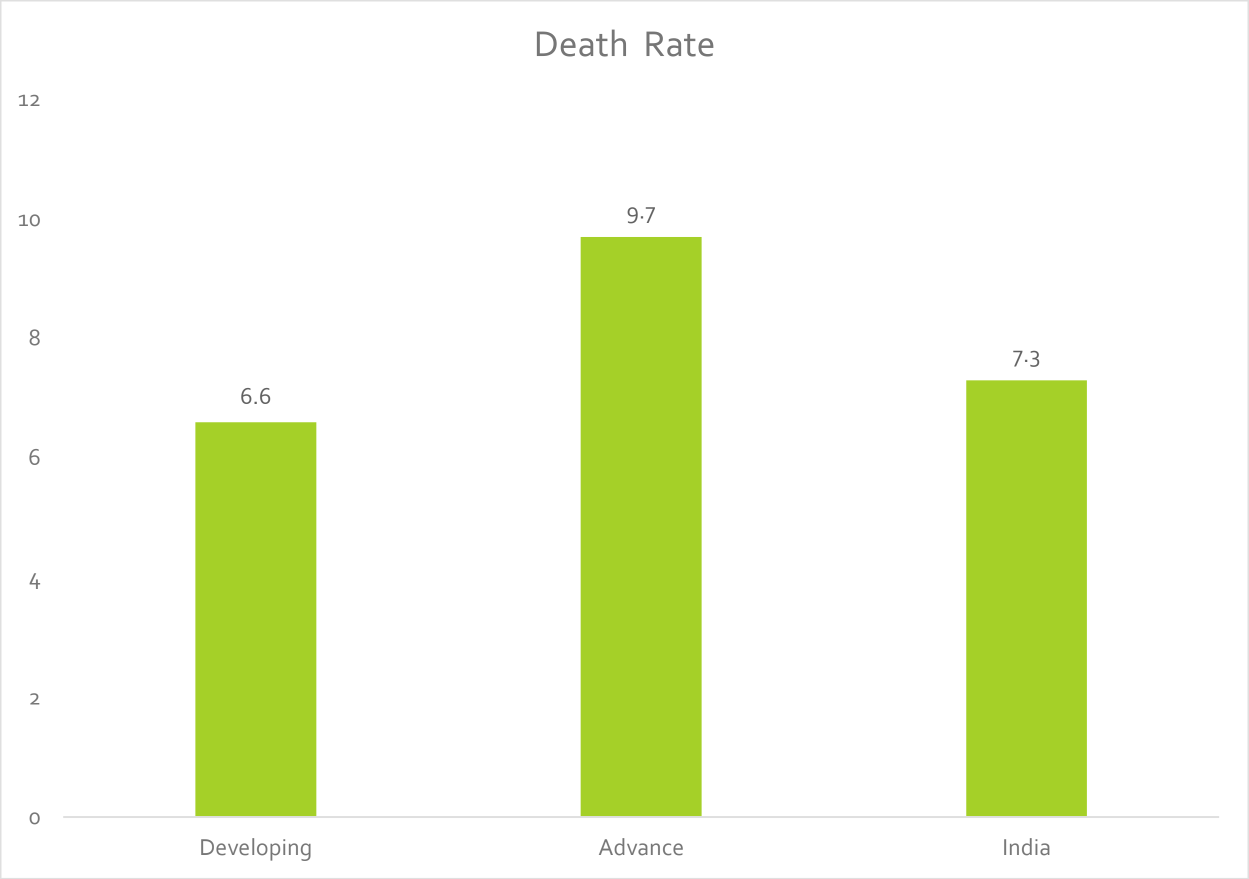
```
In [22]: sn.boxplot(df1['Birth rate'])

Out[22]: <AxesSubplot:xlabel='Birth rate'>
```





```
In [23]: sn.boxplot(df1['Death rate'])
```

```
Out[23]: <AxesSubplot:xlabel='Death rate'>
```



The birth rate is rightly skewed with a few outliers. The average birth rate for the world is 16.23 and median is 13.30. Average birth rate for India is 17.4 which is greater as well as the birth rate of the world.

Among the developing countries, the average birth rate of India is much higher than the average of the developing countries. That explains the more rise in population of India as compared to other countries.

The death rate is rightly skewed with a few outliers. The average death rate for the world is 8.17 and median is 7.30. Average birth rate for India is 7.3 which is same as the median death rate of the world.

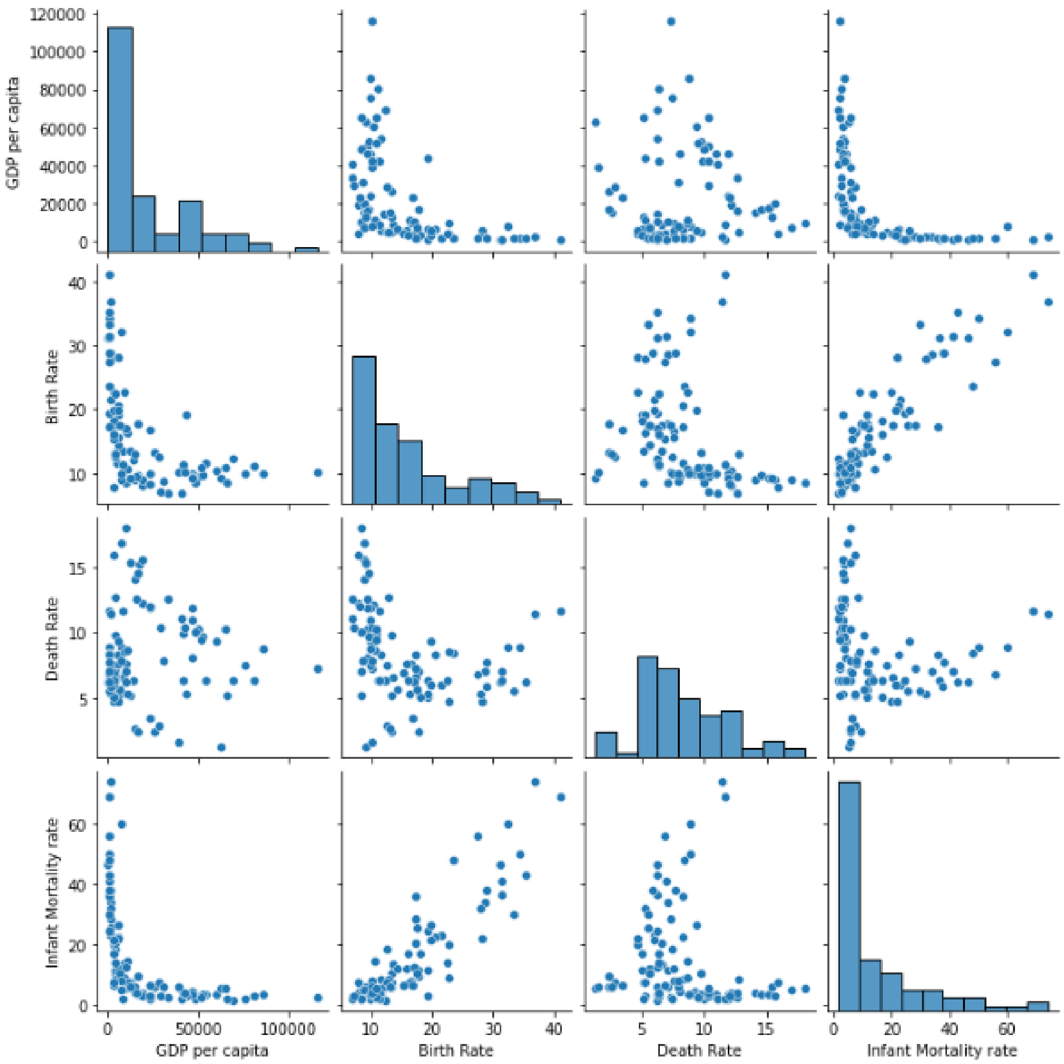
Among the developing countries, the average death rate of India is much higher than the average of the developing countries.

Pair Plots

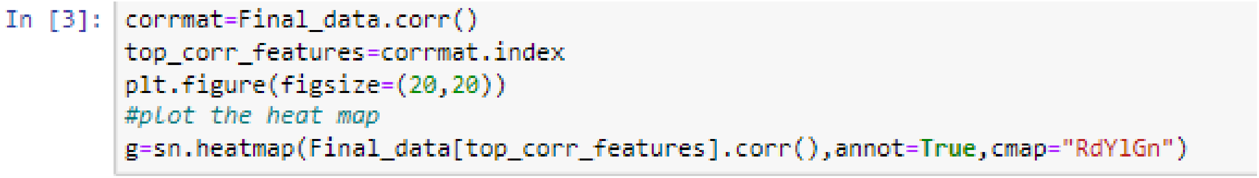
In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sn
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
%matplotlib inline

Final_data=pd.read_csv('C:/Users/User/Desktop/proj_data.csv')
```

In [2]:
```
sn.pairplot(Final_data)
```

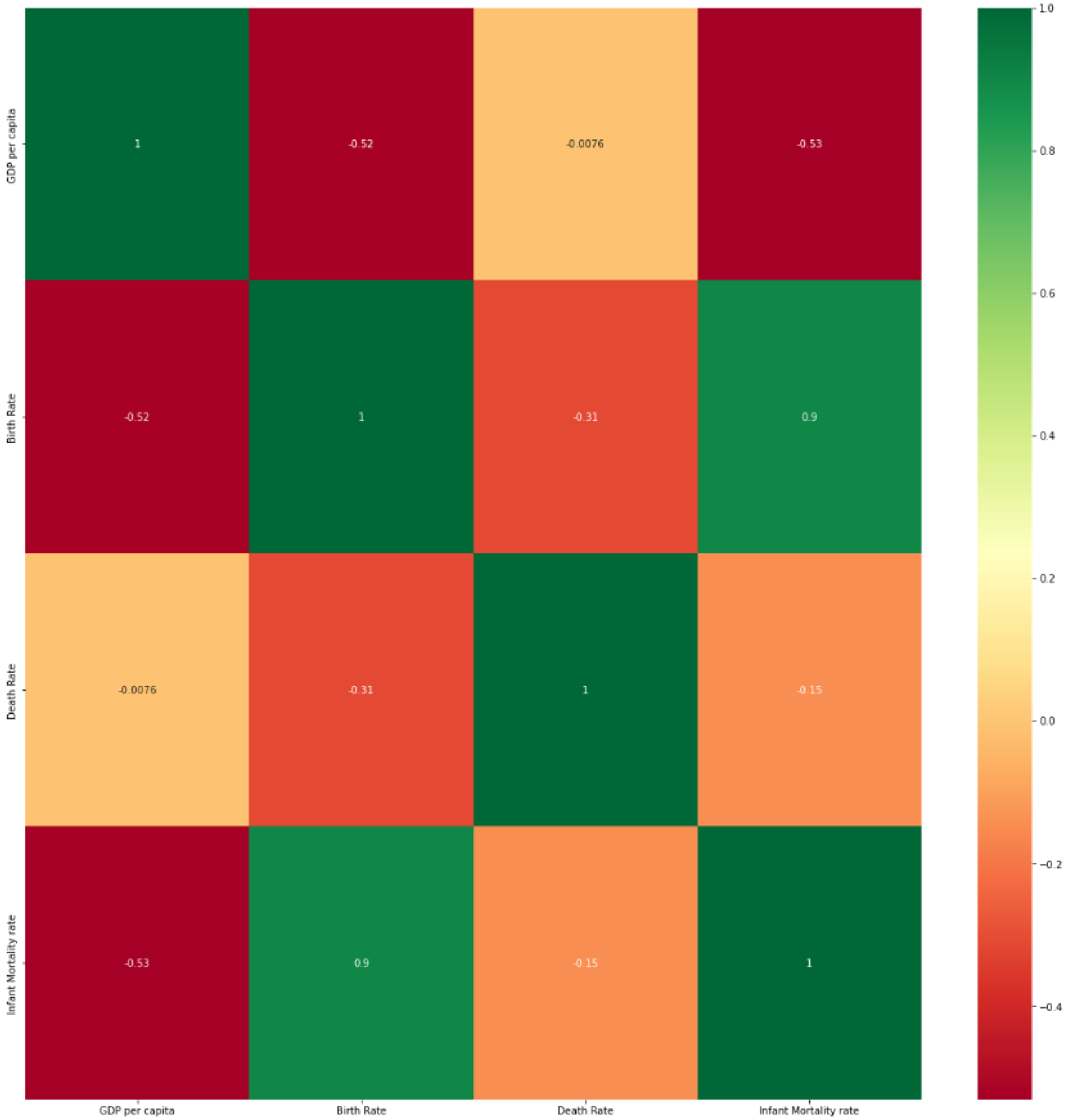
Out[2]:
```
<seaborn.axisgrid.PairGrid at 0x27fa31d28b0>
```



Heat map



In [3]:
```
corrmat=Final_data.corr()
top_corr_features=corrmat.index
plt.figure(figsize=(20,20))
#plot the heat map
g=sn.heatmap(Final_data[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```



From the **PAIR PLOT** and **HEAT MAP** we observe that:

1. **Birth Rate** has a **Moderate Negative** correlation with **GSDP** with r= -0.52
2. **Death Rate** has a **Weak Negative** correlation with **GSDP** with r= -0.0076
3. **Infant Mortality Rate** has a **Moderate Negative** correlation with **GSDP** with r= -0.53

Inference:

From the results we can infer that for the countries in the world, with the increase in GDP per capita the Birth rate & the infant mortality rate tend to decrease significantly whereas the Death Rate seems to be very weak negatively correlated with the GDP.

Regression (best fit Lines):

$$Y_{\text{(Infant mortality rate)}} = -0.00036X_{\text{(GDP per capita)}} + 22.48$$

$$Y_{\text{(Death rate)}} = -1.09*10^{-6}\,X_{\text{(GDP per capita)}} + 8.25$$

$$Y_{\text{(Birth Rate)}} = -0.00018X_{\text{(GDP per capita)}} +19.852$$

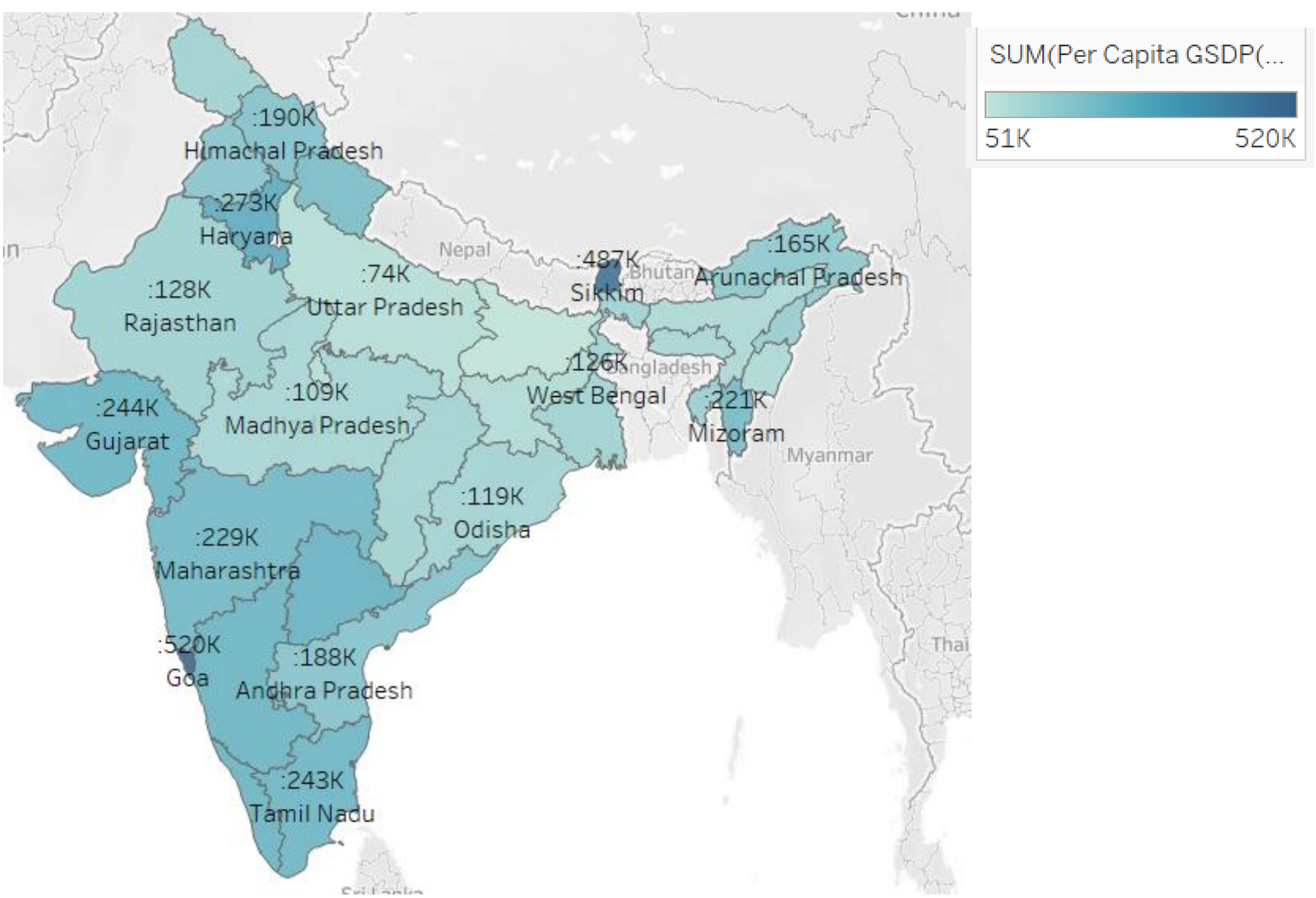
I

INDIA

We have collected these data for the year 2019 for different states in INDIA.

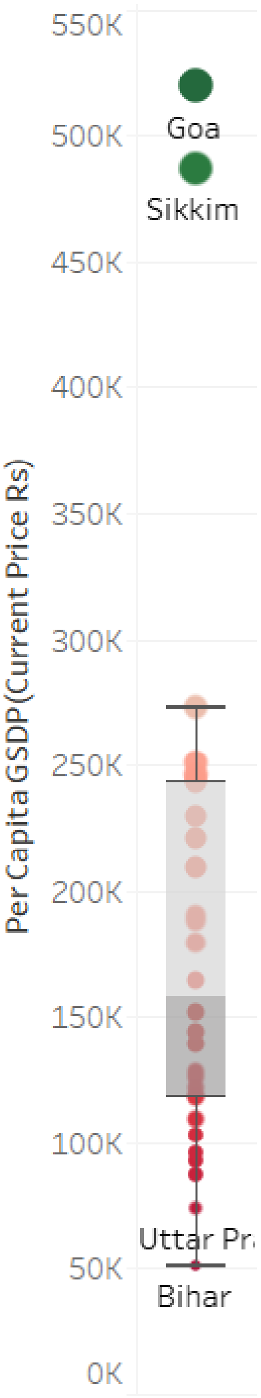
STATE	Per Capita GSDP (Current Price Rs)	Birth Rate	Death Rate	Infant Mortality rate
Andhra Pradesh	188069	15.9	6.4	25
Arunachal Pradesh	164557	17.6	5.8	29
Assam	96224	21	6.3	40
Bihar	50735	25.8	5.5	29
Chhattisgarh	117700	22.2	7.3	40
Goa	520030	12.3	5.9	8
Gujarat	243761	19.5	5.6	25
Haryana	272884	20.1	5.9	27
Himachal Pradesh	190407	15.4	6.9	19
Jharkhand	87126	22.3	5.3	27
Jammu & Kashmir	121971	14.9	4.6	20
Karnataka	246419	16.9	6.2	21
Kerala	245323	13.5	7.1	6
Madhya Pradesh	109372	24.5	6.6	46
Maharashtra	229489	15.3	5.4	17
Manipur	92427	13.6	4.3	10
Meghalaya	102672	23.2	5.6	33
Mizoram	221384	14.5	4	3
Nagaland	144138	12.7	3.5	3
Odisha	119075	18	7.1	38
Punjab	179163	14.5	6.6	19
Rajasthan	128318	23.7	5.7	35
Sikkim	487201	16.5	4.2	5
Tamil Nadu	243189	14.2	6.1	15
Telangana	250920	16.7	6.1	23
Tripura	139540	12.8	5.5	21
Uttar Pradesh	74141	25.4	6.5	41
Uttarakhand	209116	17.1	6	27
West Bengal	126121	14.9	5.3	20

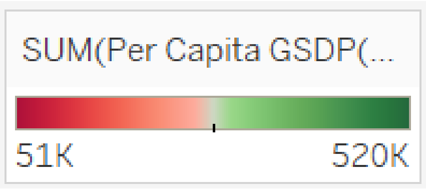
1) GSDP per capita:

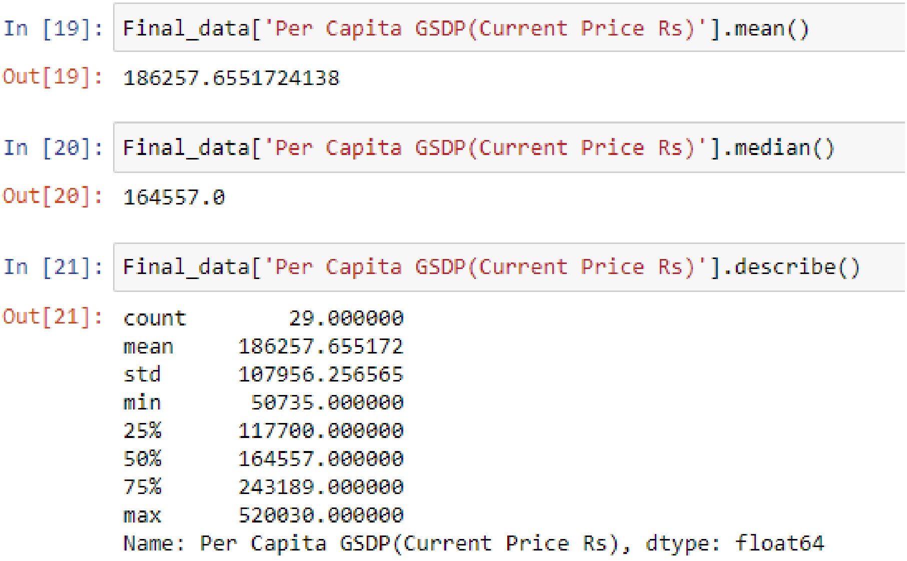


1.From the Geolocation chart it can be seen that states located in the northern or the south-western part of the country are having higher GSDP(Per-capita).

2.Goa is having the highest GSDP (per capita) whereas Sikkim has the 2nd highest in terms of GSDP(per capita).



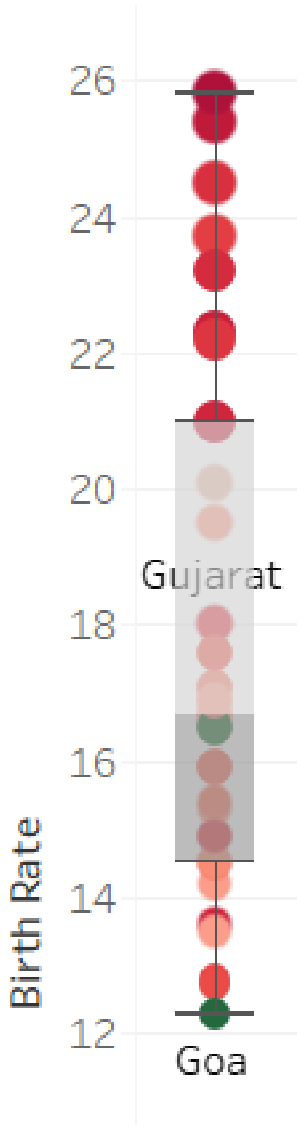


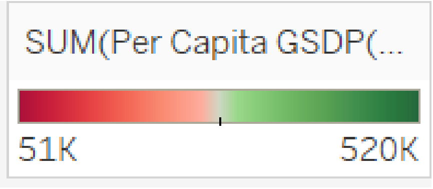


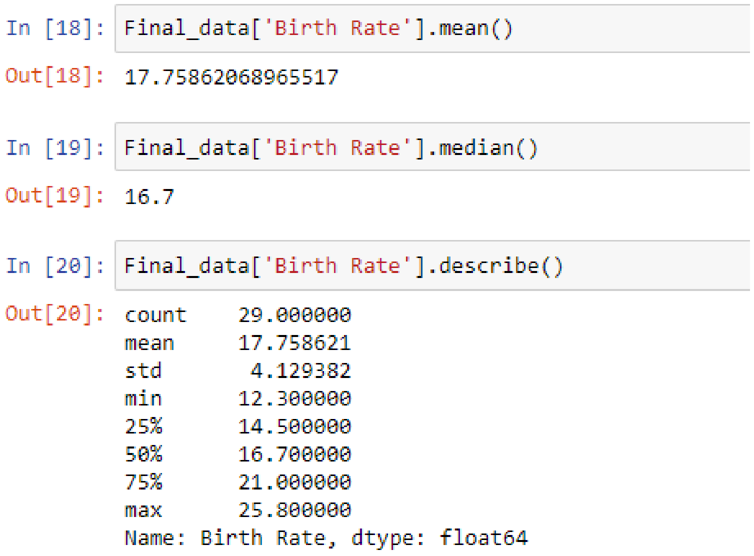
```
In [19]: Final_data['Per Capita GSDP(Current Price Rs)'].mean()

Out[19]: 186257.6551724138

In [20]: Final_data['Per Capita GSDP(Current Price Rs)'].median()

Out[20]: 164557.0

In [21]: Final_data['Per Capita GSDP(Current Price Rs)'].describe()

Out[21]: count        29.000000
         mean     186257.655172
         std      107956.256565
         min       50735.000000
         25%      117700.000000
         50%      164557.000000
         75%      243189.000000
         max      520030.000000
         Name: Per Capita GSDP(Current Price Rs), dtype: float64
```

The GSDP distribution is positively skewed where Goa and Sikkim are outliers in the positive side. Also, eastern part of India seems to be having lesser GSDP per Capita as compared to the Northern and Western part of India. The mean GSDP per capita is 186257.65 rupees and median is 164557 rupees.

2) Birth Rate



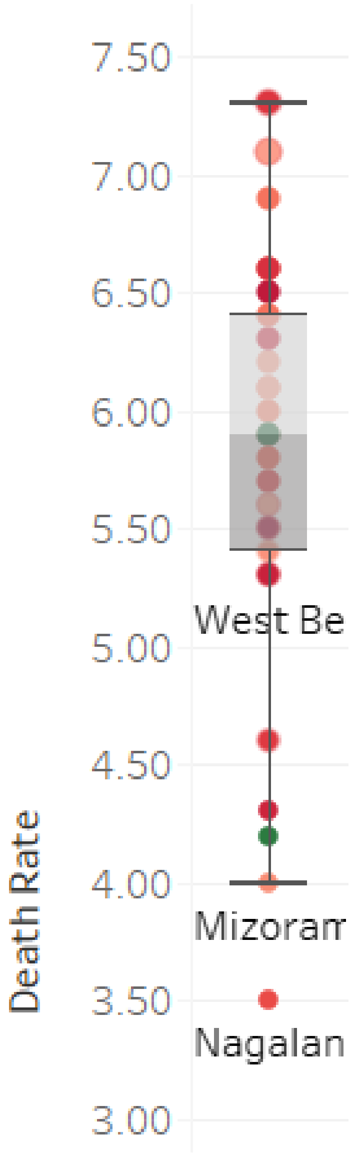


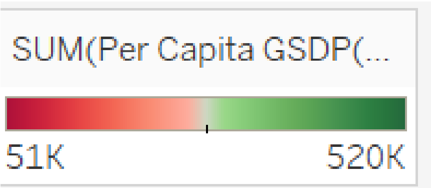


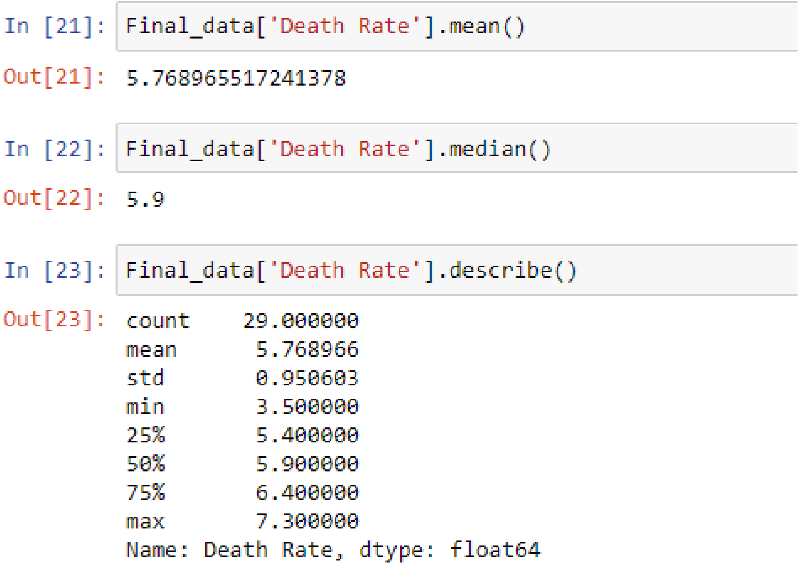
```
In [18]: Final_data['Birth Rate'].mean()

Out[18]: 17.75862068965517

In [19]: Final_data['Birth Rate'].median()

Out[19]: 16.7

In [20]: Final_data['Birth Rate'].describe()

Out[20]: count    29.000000
         mean     17.758621
         std       4.129382
         min      12.300000
         25%      14.500000
         50%      16.700000
         75%      21.000000
         max      25.800000
         Name: Birth Rate, dtype: float64
```

The Birth Rate distribution is positively skewed. The states with higher GSDP (Goa, Sikkim, Kerala etc) per capita are having lesser Birth Rate, while states with lower GSDP (Bihar, Uttarpradesh, Madhyapradesh etc) have higher Birth rates. The mean birth rate is 17.75 and the median values is 16.7.

3) Death Rate



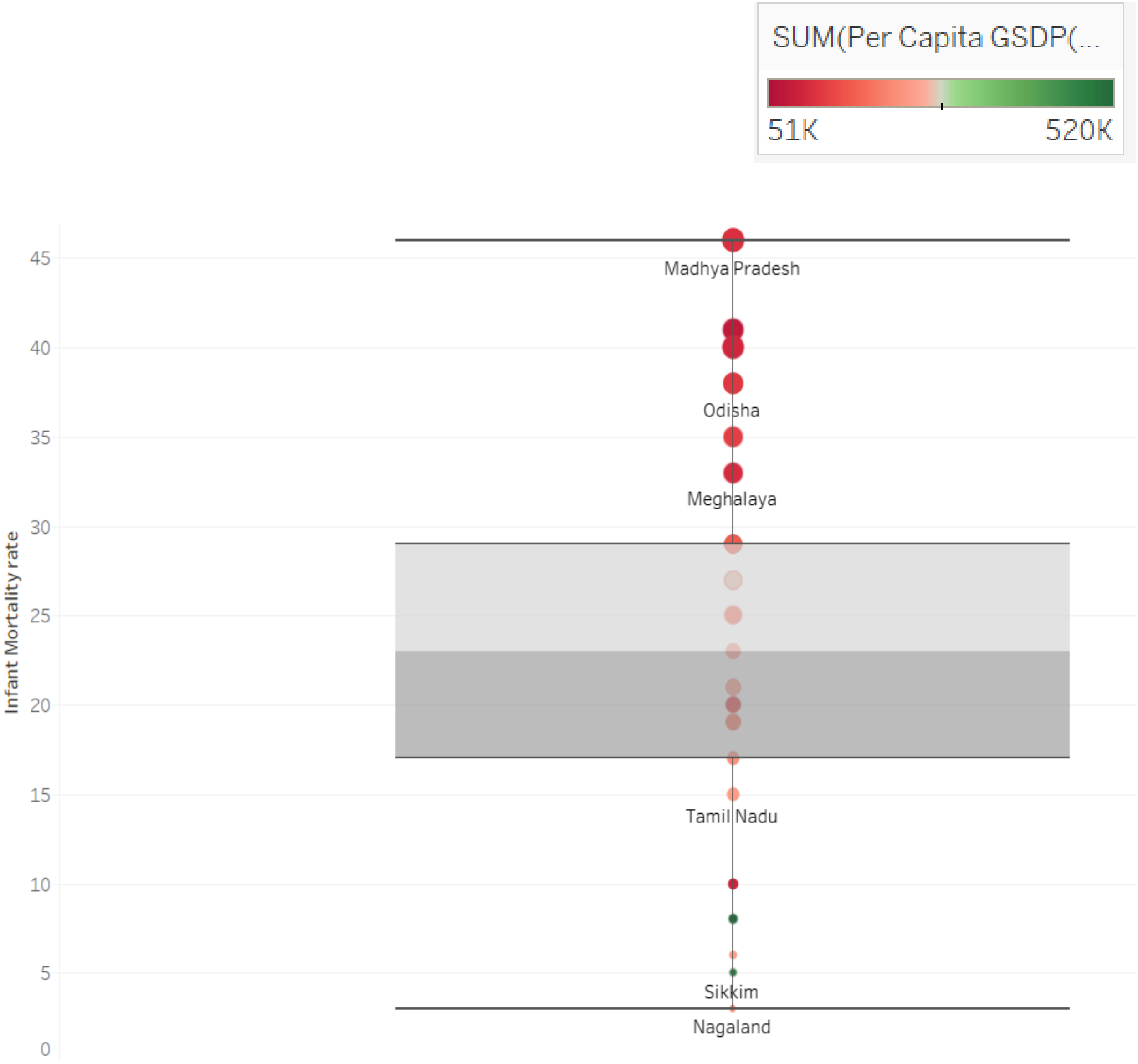


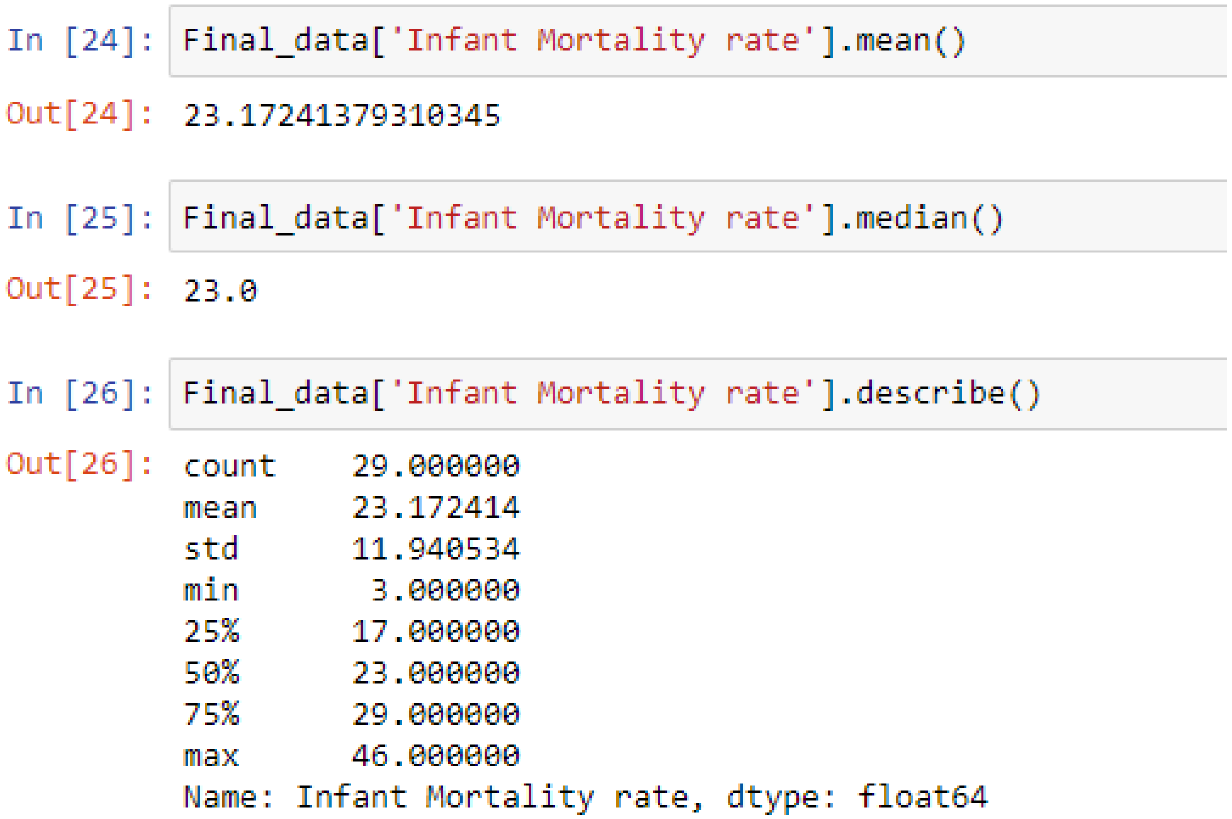


```
In [21]: Final_data['Death Rate'].mean()
Out[21]: 5.768965517241378

In [22]: Final_data['Death Rate'].median()
Out[22]: 5.9

In [23]: Final_data['Death Rate'].describe()
Out[23]: count    29.000000
         mean      5.768966
         std       0.950603
         min       3.500000
         25%       5.400000
         50%       5.900000
         75%       6.400000
         max       7.300000
         Name: Death Rate, dtype: float64
```

Death rate distribution is approximately normal. Only outlier is the state of Nagaland. The mean death rate is 5.76 and the median is 5.9

4) Infant Mortality Rate





```
In [24]: Final_data['Infant Mortality rate'].mean()

Out[24]: 23.17241379310345

In [25]: Final_data['Infant Mortality rate'].median()

Out[25]: 23.0

In [26]: Final_data['Infant Mortality rate'].describe()

Out[26]: count    29.000000
         mean     23.172414
         std      11.940534
         min       3.000000
         25%      17.000000
         50%      23.000000
         75%      29.000000
         max      46.000000
         Name: Infant Mortality rate, dtype: float64
```

It can be seen that the states with low GSDP have high infant mortality rates and states with high GSDP has low infant mortality rates. The mean infant mortality rate across the states of India is 23.17 and median is 23.0

Pair Plot

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sn
        import warnings
        warnings.filterwarnings('ignore')
        import matplotlib.pyplot as plt
        %matplotlib inline
```

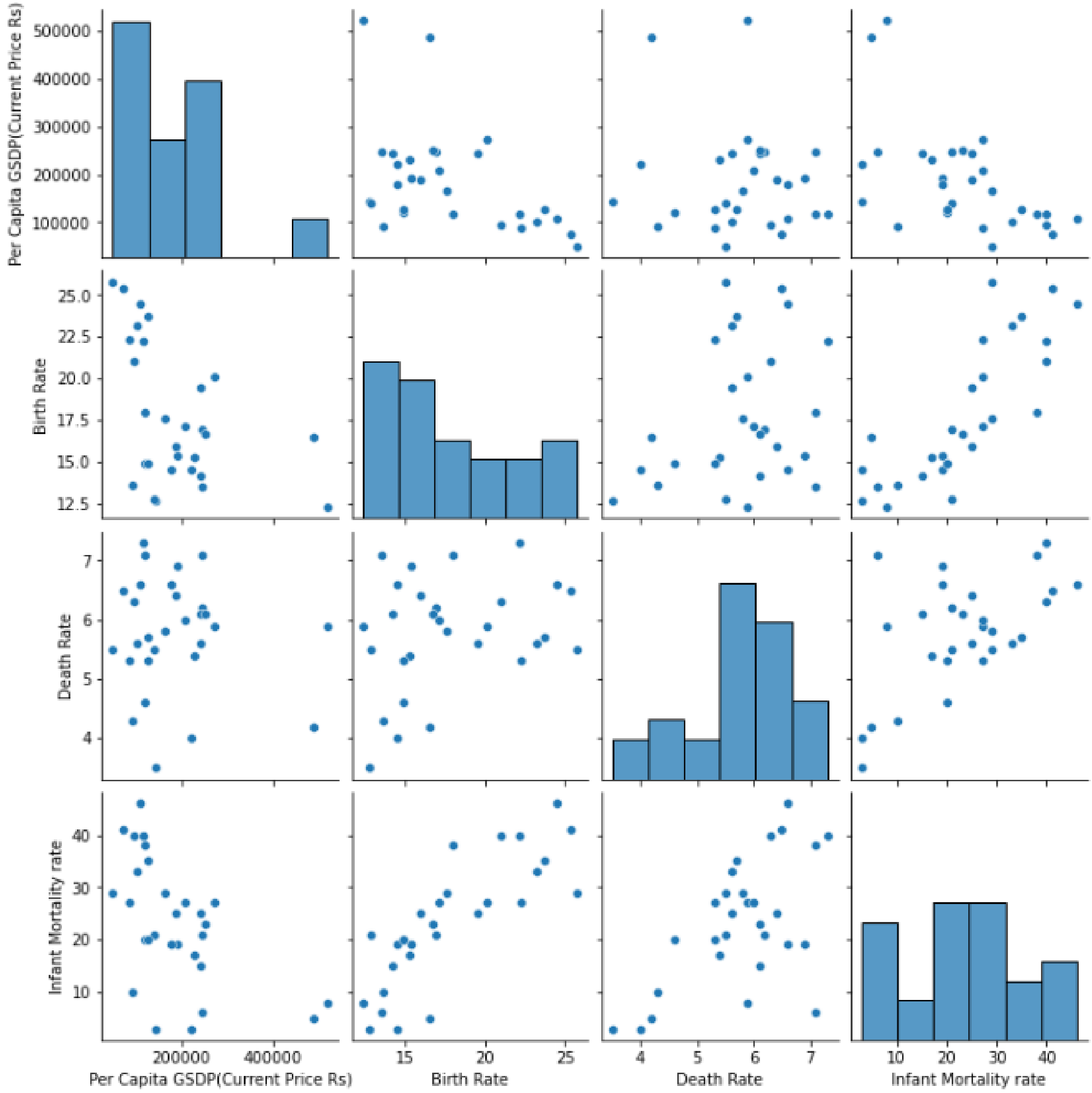
```
In [2]: Final_data=pd.read_csv('C:/Users/User/Desktop/Praxis Business School/Term1/STS/STS Project/Manas STS/STS Final Dataset 3.csv')
```

```
In [3]: Final_data.shape
```

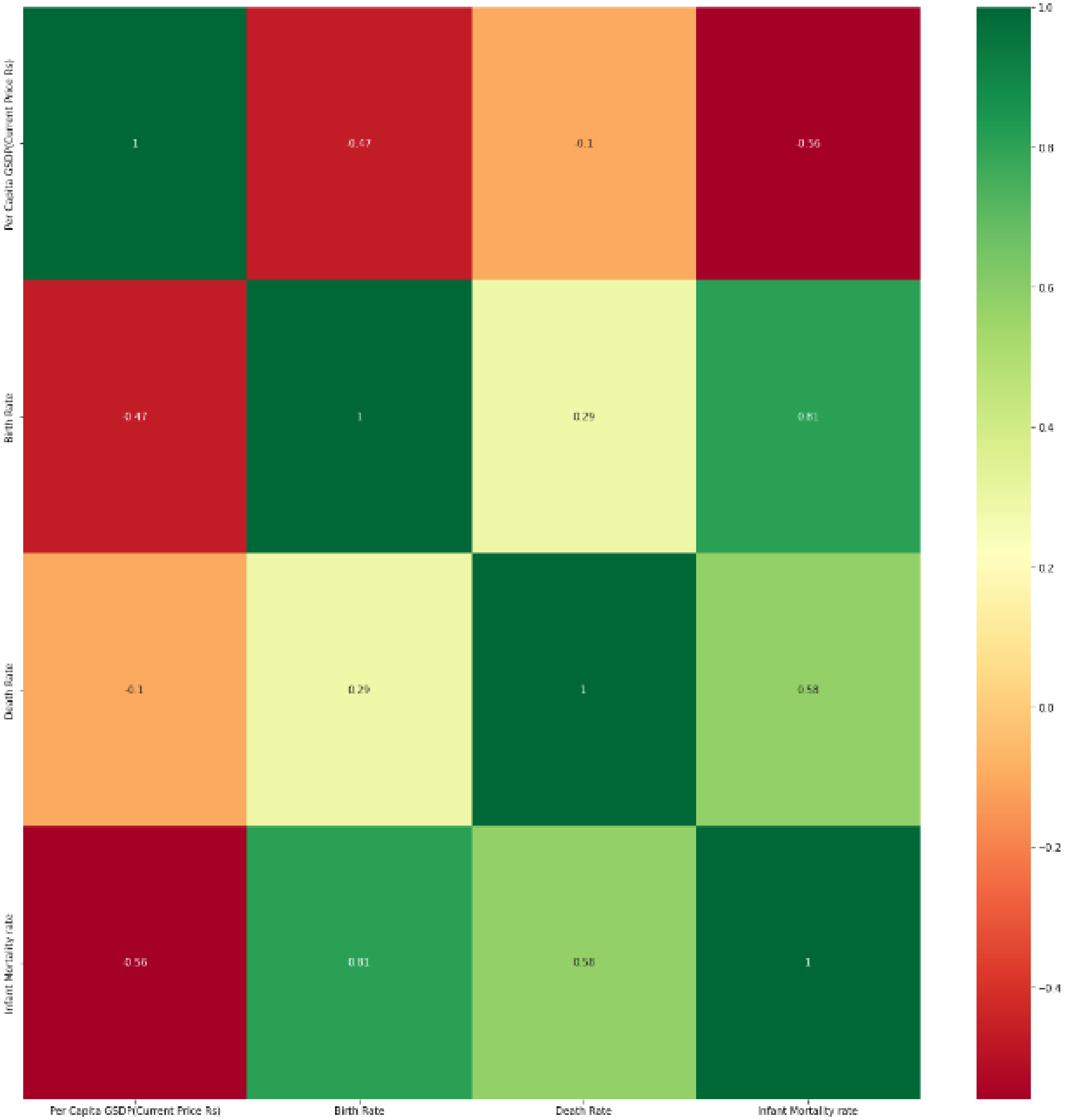
```
Out[3]: (29, 5)
```

```
In [4]: sn.pairplot(Final_data)
```

```
Out[4]: <seaborn.axisgrid.PairGrid at 0x1dd6a5119d0>
```



Heat Map


Per Capita GSDP(Current Price Rs)
Birth Rate
Death Rate
Infant Mortality rate
1
-0.47
-0.1
-0.56
-0.47
1
0.29
0.81
-0.1
0.29
1
0.58
-0.56
0.81
0.58
1
Per Capita GSDP(Current Price Rs)
Birth Rate
Death Rate
Infant Mortality rate
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

From the **PAIR PLOT** and **HEAT MAP** we observe that:

4. **Birth Rate** has a **Moderate Negative** correlation with **GSDP** with r= -0.47
5. **Death Rate** has a **Weak Negative** correlation with **GSDP** with r= -0.1
6. **Infant Mortality Rate** has a **Moderate Negative** correlation with **GSDP** with r= -0.52

Inference:

From the results we can infer that for the states in India with the increase in GSDP per capita the Birth rate & the infant mortality rate tend to decrease significantly whereas the Death Rate seems to be very weak negatively correlated with the GSDP

Regression (best fit Lines):

$\mathbf{Y}_{\text{(Birth Rate)}} \mathbf{= -1.9*10^{-5}X}_{\text{(GDP per capita)}} \mathbf{+ 21.385}$

$\mathbf{Y}_{\text{(Death Rate)}} \mathbf{= -1.2*10^{-6}X}_{\text{(GDP per capita)}} \mathbf{+ 6.03}$

$\mathbf{Y}_{\text{(Infant mortality rate)}} \mathbf{= -6.4*10^{-5}X}_{\text{(GDP per capita)}} \mathbf{+ 35.25}$

CONCLUSION:

In this project we can see how this Exploratory Data Analysis is playing an Essential part to get some incredibly valuable insights about the World Health Care as well as Indian Health care by considering Infant mortality Rate, Birth Rate, Death Rate with respect to GDP per capita. The rising cost of Health Care is considered a real concern and is directly linked to our income which affects our financial capability pertaining to various ailments and daily well-being & hence, the data of Infant mortality Rate, Birth Rate and Death Rate.

For our use case here, we used the basics statistical knowledge to generate various charts and investigated relationship of different variables and the spread of a certain variable across different countries, region & the states of India. The data set collected from web helped us find how much these parameters are linked to GDP with the help of correlation coefficient. Also, we found the best fit line to predict these parameters given the GDP per capita.

Finally, we hope that the entire effort given to draw a picture of the World Health Care with few of its parameters may be of some use to understand the need to uplift the living standards of the general public. We would like to thank Praxis Business School for providing us such an opportunity to present our work.

Thank you!