To: Dr Stephan Poikonen
From: Chinmayi Mahadik

A report analysing the world economic factors which influence GDP of a country.

EXECUTIVE SUMMARY

Major Findings.

- GDP is positively impacted by factors like Population Density, Immigration, Literacy rate, Technology. Countries which have a coastline may have a benefit of delivering services in a more efficient way leading to increase in GDP.
- Most of the Countries have a literate population. Educated countries could make smarter and better decisions for the economy. I observed that most of the literate countries have higher GDP.
- Countries which have a better GDP have low infant mortality rate. This may be because these countries have better facilities and health care systems.
- Healthy migration can cause surge in the GDP.
- Temperate Climates act as a favourable influence thus enhancing the GDP of Western European regions like Luxembourg, UK.
- 90% of variability in GDP can be attributed to Net Migration, Literacy, Technology (Phones), Service, Coastline existence and climatic conditions.
- 92% of the variability in energy consumption can be explained by a factor of population and GDP.
- 79% of the variability in literacy can be explained infant mortality, GDP, phones, and birth rate.
- 70% of the variability in literacy can be explained infant mortality, GDP, phones, and birth rate.

Recommendations for Action.

The analysis was conducted just on few factors which affect GDP. We would suggest taking factors suggested by the World Bank, to predict better GDP.

Since literacy has a positive impact on GDP, increasing the number of educational institutions, motivating below poverty line people to study, providing free education in a country, might be helpful for future GDP.

Concentrating on improving health and medical facilities might decrease the infant mortality. Countries which don't have a coastline, may build cargo airports to enable better trade and services.

Analytical Overview.

Elementary data analysis was performed on the countries of the world data set. After analysing each factor contributing to GDP, Energy Consumption, Literacy, and Infant Mortality, multiple regression models were built to attribute its variability to the factors.

APPENDIX

A. Data cleaning.

First step in the data analysis was cleaning the data. There were formatting issues related with the decimal separator for all numerical variables. Data set had commas as a decimal separator and it was replaced for a decimal separator. This was done previously in Excel before exporting the data to Python.

The following step was to deal with missing values. Since it was a known the data set source, most of the missing values were found, however for some countries that were missing important data that was not finally found, they were eliminated. These countries were mainly territories which belonged to other countries and in terms of population and GDP were not representative. The total of countries eliminated were 9.

B. Data summarizing and Inference from univariant charts.

The data set of countries in the world includes 20 columns and 218 rows.

```
countryData = pd.read_csv("countriesoftheworld.csv")
print(countryData)
```

Table 1. Countries of the world.

	Country	Region	...	Industry	Service
0	Afghanistan	ASIA (EX. NEAR EAST)	...	0.24	0.38
1	Albania	EASTERN EUROPE	...	0.19	0.58
2	Algeria	NORTHERN AFRICA	...	0.60	0.30
3	American Samoa	OCEANIA	...	0.12	0.60
4	Andorra	WESTERN EUROPE	...	0.34	0.55
..	...	...	...	...	...
213	Virgin Islands	LATIN AMER. & CARIB	...	0.19	0.80
214	West Bank	NEAR EAST	...	0.28	0.63
215	Yemen	NEAR EAST	...	0.47	0.39
216	Zambia	SUB-SAHARAN AFRICA	...	0.29	0.49
217	Zimbabwe	SUB-SAHARAN AFRICA	...	0.24	0.58

Data set has been already clean and does not have missing values.

```
print(countryData.isnull().sum())
```

```
In [375]: print(countryData.isnull().sum())
Country              0
Region               0
Population           0
Area                 0
Pop. Density         0
Coastline            0
Net migration        0
Infant mortality     0
GDP                  0
Literacy             0
Phones               0
Arable               0
Crops                0
Other                0
Climate              0
Birthrate            0
Deathrate            0
Agriculture          0
Industry             0
Service              0
dtype: int64
```

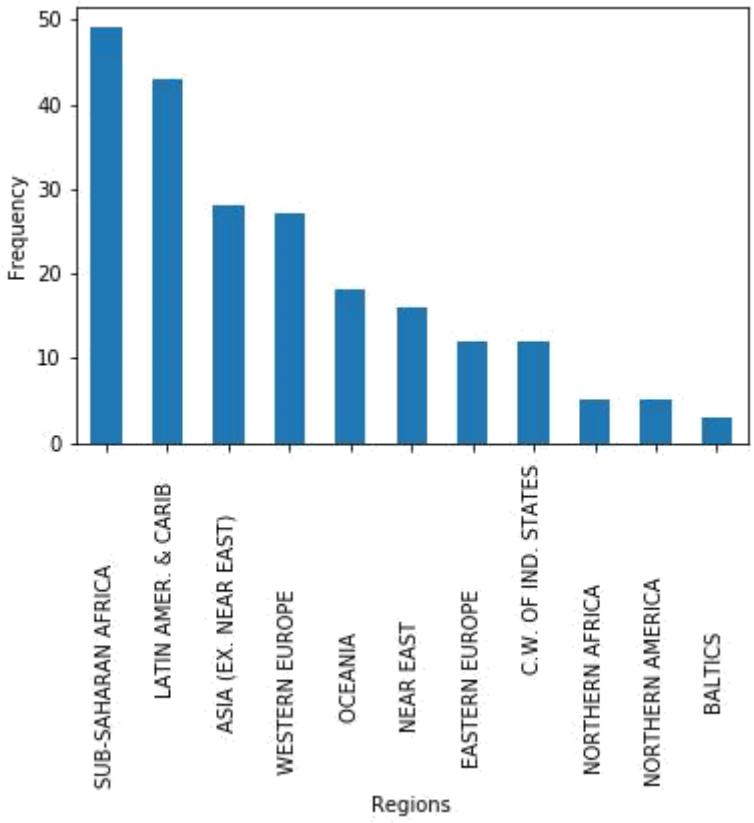
Region.

This variable groups the countries based on its location on the globe by 11 categories.

```
countryData['Region'].value_counts()
```

Table 2. Countries per Region.

SUB-SAHARAN AFRICA	49
LATIN AMER. & CARIB	43
ASIA (EX. NEAR EAST)	28
WESTERN EUROPE	27
OCEANIA	18
NEAR EAST	16
EASTERN EUROPE	12
C.W. OF IND. STATES	12
NORTHERN AFRICA	5
NORTHERN AMERICA	5
BALTICS	3



Sub-Saharan Africa and Latin America and the Caribbean are the most representative categories including the majority of the countries on this data set.

Population.

```
countryData['Population'].describe()
```

```
count    2.180000e+02
mean     2.992282e+07
std      1.201639e+08
min      7.026000e+03
25%      5.870655e+05
50%      5.445054e+06
75%      1.880989e+07
max      1.313974e+09
```

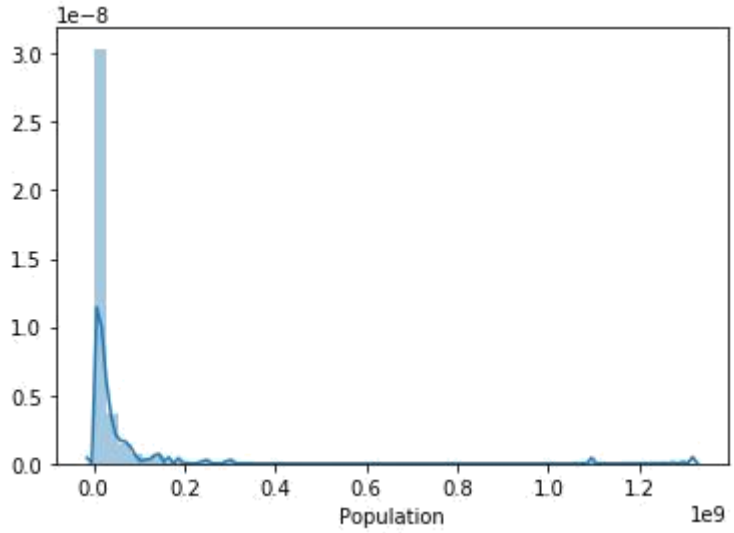
```
countryData['Country'].iloc[countryData.index[countryData['Population']==countryData['Population'].max()].tolist()]
```

```
42    China
```

```
countryData['Country'].iloc[countryData.index[countryData['Population']==countryData['Population'].min()].tolist()]
```

```
168    St Pierre & Miquelon
```

From above I can conclude that population mean is 2.992282e+07 and median 5445054.5. The data is right skewed, mean is greater than the median. The country with the largest population is China with 1.313974e+09 people and the country with the smallest population is St Pierre & Miquelon with 7026 people.



Area.

```
countryData['Area '].describe()
```

```
count    2.180000e+02
mean     6.216903e+05
std      1.823130e+06
min      2.000000e+00
25%      1.106825e+04
50%      9.234550e+04
75%      4.493230e+05
max      1.707520e+07
```

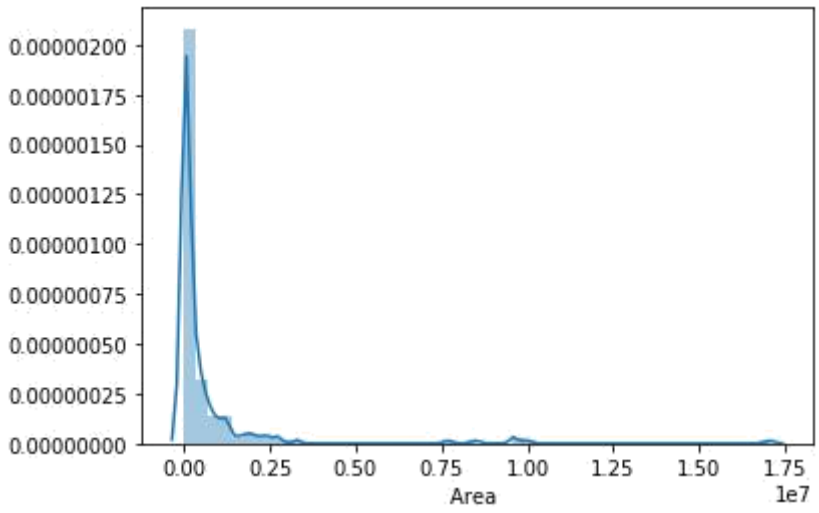
```
countryData['Country'].iloc[countryData.index[countryData['Area ']==countryData['Area '].max()].tolist()]
```

```
164    Russia
```

```
countryData['Country'].iloc[countryData.index[countryData['Area ']==countryData['Area '].min()].tolist()]
```

```
134    Monaco
```

From above I can conclude that area mean is 6.217e+05 Km^2 and median 92345.5 Km^2. The data is right skewed, mean is greater than the median. The country with the largest area is Russia with 1.71e+07 Km^2 and the country with the smallest area is Monaco with 2 Km^2.



Population Density. Number of people per square kilometre.

```
countryData['Pop. Density'].describe()
```

```
count      218.000000
mean       385.327982
std       1693.069600
min          0.000000
25%         29.075000
50%         78.100000
75%        187.250000
max      16271.500000
```

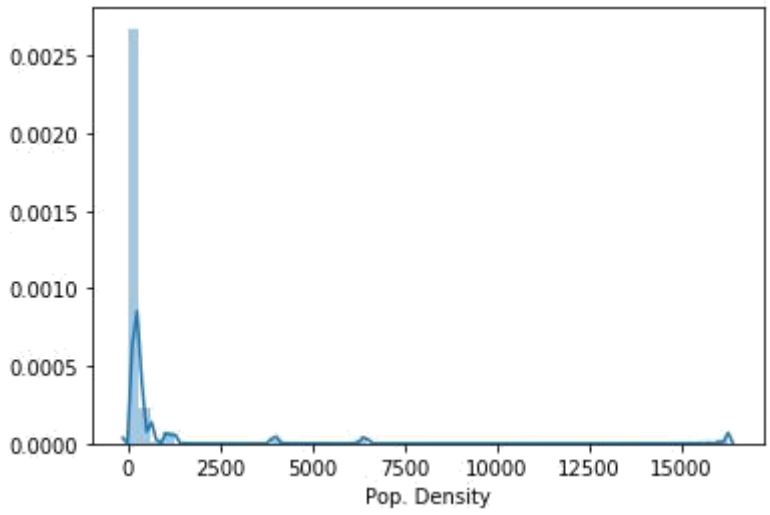
```
countryData['Country'].iloc[countryData.index[countryData['Pop. Density']==countryData['Pop. Density'].max()].tolist()]
```

```
134    Monaco
```

```
countryData['Country'].iloc[countryData.index[countryData['Pop. Density']==countryData['Pop. Density'].min()].tolist()]
```

```
80    Greenland
```

From above I can conclude that population density mean is 385.38 people per Km^2 and median 78.1 people per Km^2. The data is right skewed, mean is greater than the median. The country with the highest population density is Monaco with 16271.5 people per Km^2 and the country with the smallest population density is Greenland with 0 people per Km^2.



Coastline. Coastline /area. A coastline of zero indicates that the country is landlocked.

```
countryData['Coastline '].describe()
```

```
count    218.000000
mean      19.885550
std       73.028896
min        0.000000
25%        0.090000
50%        0.700000
75%        8.132500
max      870.660000
```

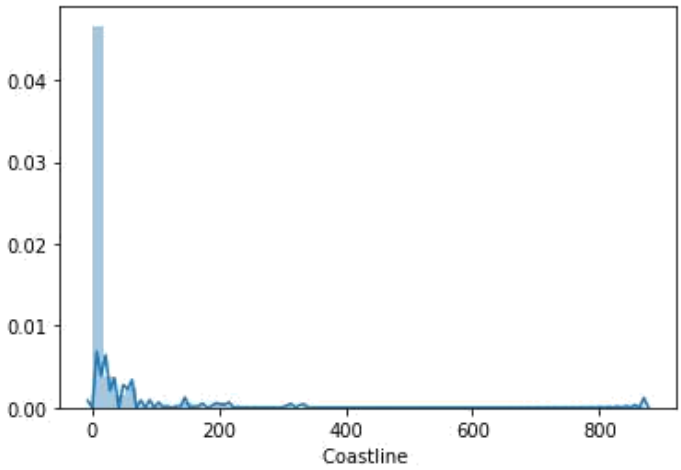
```
countryData['Country'].iloc[countryData.index[countryData['Coastline ']==countryData['Coastline '].max()].tolist()]
```

```
132    Micronesia Fed. St.
```

```
Countrieswithzerocoastline=0
for i in range(len(countryData)):
    if (countryData['Coastline '].iloc[i]==0):
        Countrieswithzerocoastline+=1
    else:
        pass
print(Countrieswithzerocoastline)
```

```
44
```

From above I can conclude that coastline mean is 19.88 km and median 0.7 km. The data is right skewed, mean is greater than the median. The country with the largest coastline is Federated States of Micronesia with 870.66 km and there are 44 countries with zero coastline.



Net migration. Difference between the number of immigrants and the number of emigrants throughout the year (per 1000 people).

```
countryData['Net migration'].describe()
```

```
count    218.000000
mean      -0.159908
std        5.068030
min      -21.100000
25%       -1.177500
50%        0.000000
75%        0.965000
max       23.060000
```

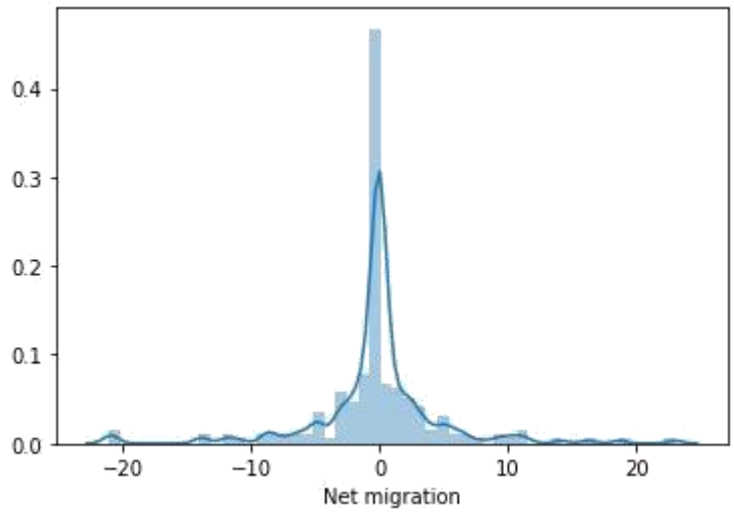
```
countryData['Country'].iloc[countryData.index[countryData['Net migration']==countryData['Net migration'].max()].tolist()]
```

```
0    Afghanistan
```

```
countryData['Country'].iloc[countryData.index[countryData['Net migration']==countryData['Net migration'].min()].tolist()]
```

```
47    Cook Islands
```

From above I can conclude that net migration mean is -0.16 and median 0. The data is approximately normal, mean and median are almost equal. The country with the largest net migration rate is Afghanistan with 23.06 and country with the smallest net migration rate is Cook Islands.



Infant mortality rate. Compares the number of deaths of infants under one year old each year per 1,000 live births in the same year.

```
countryData['Infant mortality '].describe()
```

```
count    218.000000
mean      35.787018
std       35.643327
min        2.290000
25%        8.290000
50%       21.040000
75%       56.095000
max      191.190000
```

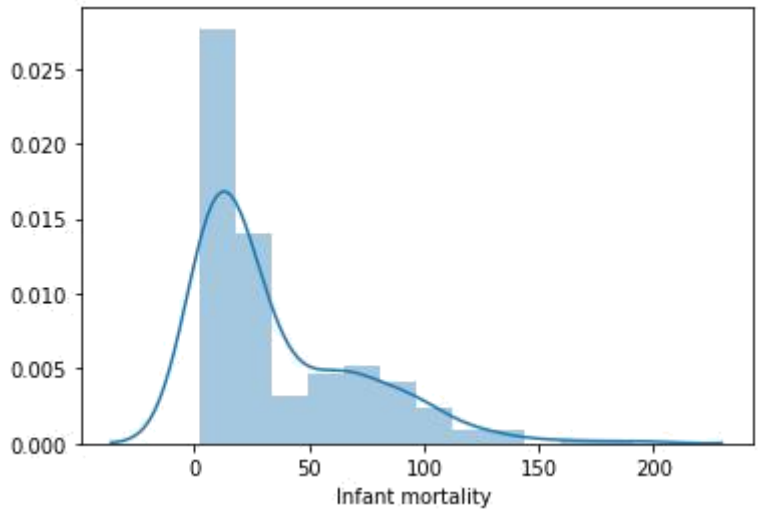
```
countryData['Country'].iloc[countryData.index[countryData['Infant mortality ']==countryData['Infant mortality '].max()].tolist()]
```

```
5    Angola
```

```
countryData['Country'].iloc[countryData.index[countryData['Infant mortality ']==countryData['Infant mortality '].min()].tolist()]
```

```
178    Singapore
```

From above I can conclude that the infant mortality rate mean is 35.79 and median 21.04. The data is right skewed, mean is greater than the median. The country with the largest infant mortality rate is Angola with 191.19 and the country with the smallest infant mortality rate is Singapore with 2.29.



Gross Domestic Product per capita (GDP).

```
countryData['GDP '].describe()
```

```
count      218.000000
mean      9753.669725
std      10121.410911
min        500.000000
25%       1900.000000
50%       5700.000000
75%      15700.000000
max      55100.000000
```

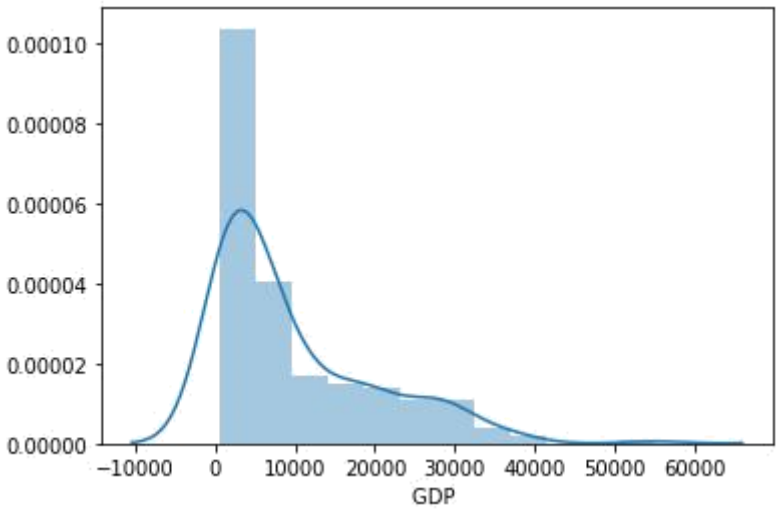
```
countryData['Country'].iloc[countryData.index[countryData['GDP ']==countryData['GDP '].max()].tolist()]
```

```
118    Luxembourg
```

```
countryData['Country'].iloc[countryData.index[countryData['GDP ']==countryData['GDP '].min()].tolist()]
```

```
58       East Timor
177    Sierra Leone
182         Somalia
```

From above I can conclude that GDP mean is 9753.67 USD and median 5700 USD. The data is right skewed, mean is greater than the median. The country with the largest GDP is Luxembourg with 55,100 USD and the countries with the smallest GDP rate are East Timor, Sierra Leone, and Somalia with 500 USD each.



Literacy. Measures literacy among persons aged 15 years and older. Expressed as a percentage of the total population in that age group.

```
countryData['Literacy '].describe()
```

```
count    218.000000
mean      83.645046
std       19.430718
min       17.600000
25%       75.900000
50%       92.600000
75%       98.000000
max      100.000000
```

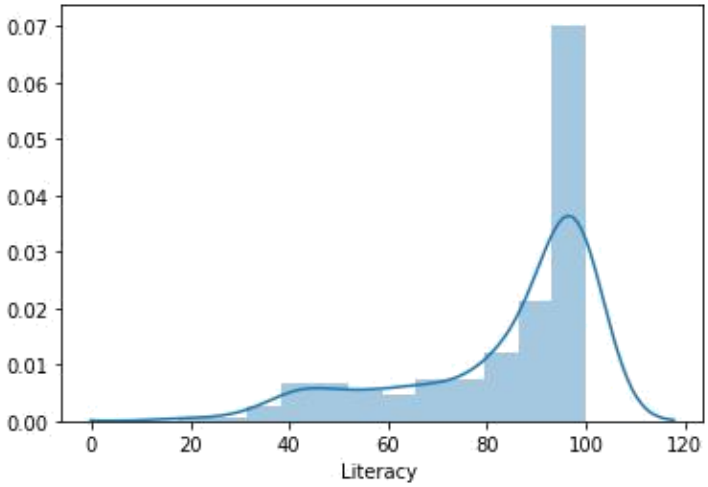
```
countryData['Country'].iloc[countryData.index[countryData['Literacy ']==countryData['Literacy '].max()].tolist()]
```

```
4            Andorra
11         Australia
54           Denmark
68           Finland
116    Liechtenstein
118       Luxembourg
149           Norway
```

```
countryData['Country'].iloc[countryData.index[countryData['Literacy ']==countryData['Literacy '].min()].tolist()]
```

```
146    Niger
```

From above I can conclude that the literacy rate mean is 83.65 and median is 92.6. The data is left skewed, median is greater than the mean. The countries with the largest literacy rate are Andorra, Australia, Denmark, Finland, Liechtenstein, Luxemburg, and Norway with 100 each. The country with the smallest literacy rate is Niger with 17.6.



Phones. Number of phones per 1000 people.

```
countryData['Phones'].describe()
```

```
count     218.000000
mean      233.387615
std       225.885794
min         0.200000
25%        34.175000
50%       172.900000
75%       383.900000
max      1035.600000
```

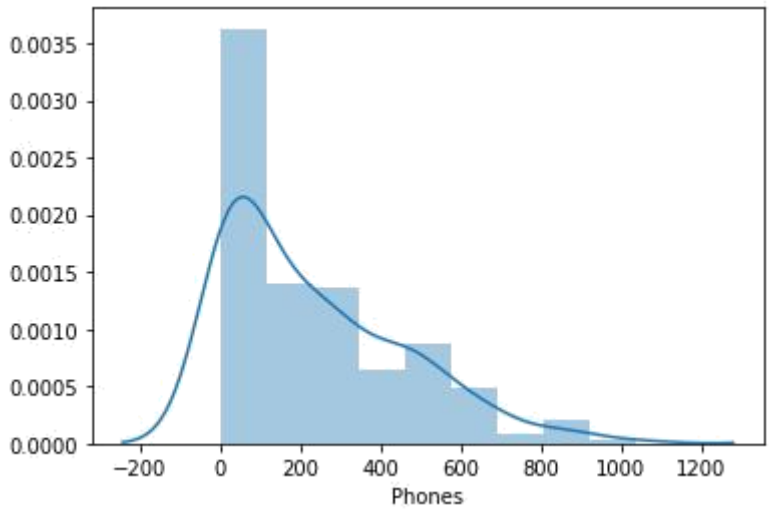
```
countryData['Country'].iloc[countryData.index[countryData['Phones']==countryData['Phones'].max()].tolist()]
```

```
134    Monaco
```

```
countryData['Country'].iloc[countryData.index[countryData['Phones']==countryData['Phones'].min()].tolist()]
```

```
45    Congo Dem. Rep.
```

From above I can conclude that phones mean is 233.39 per 1000 people and median is 172.90 phones per 1000 people. The data is right skewed, mean is greater than the median. The country with the largest number of phones per 1000 people is Monaco with 1035.6 phones. The country with the smallest number of phones per 1000 people is Democratic Republic of the Congo with 0.2 phones per 1000 people.



Arable. Percentage of the land that is arable.

```
countryData['Arable'].describe()
```

```
count    218.000000
mean      14.001239
std       13.145796
min        0.000000
25%        3.562500
50%       10.600000
75%       20.570000
max       62.110000
```

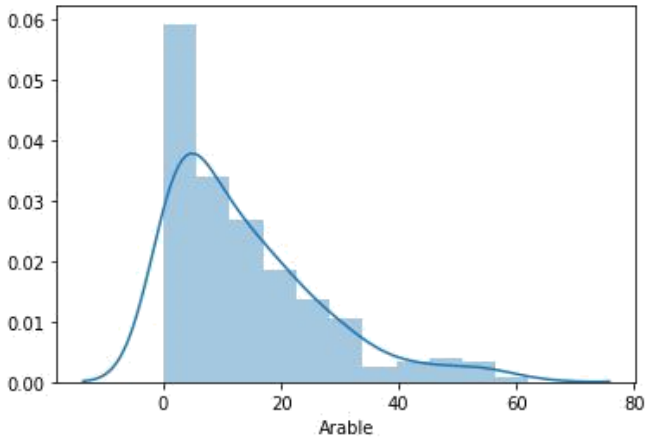
```
countryData['Country'].iloc[countryData.index[countryData['Arable']==countryData['Arable'].max()].tolist()]
```

```
16    Bangladesh
```

```
countryData['Country'].iloc[countryData.index[countryData['Arable']==countryData['Arable'].min()].tolist()]
```

```
6       Anguilla
78     Gibraltar
80     Greenland
102       Jersey
119        Macau
134       Monaco
139        Nauru
150         Oman
202       Tuvalu
```

From above I can conclude that arable mean is 14% and median is 10.6%. The data is right skewed, mean is greater than the median. The country with the largest proportion of arable land is Bangladesh with 62.11%. The countries with the smallest proportion of arable land are Anguilla, Gibraltar, Greenland, Jersey, Macau, Monaco, Nauru, Oman, and Tuvalu with 0% each.



Crops. Percentage of the land used for crops.

```
countryData['Crops '].describe()
```

```
count    218.000000
mean       4.288578
std        7.703449
min        0.000000
25%        0.210000
50%        1.055000
75%        4.427500
max       48.960000
```

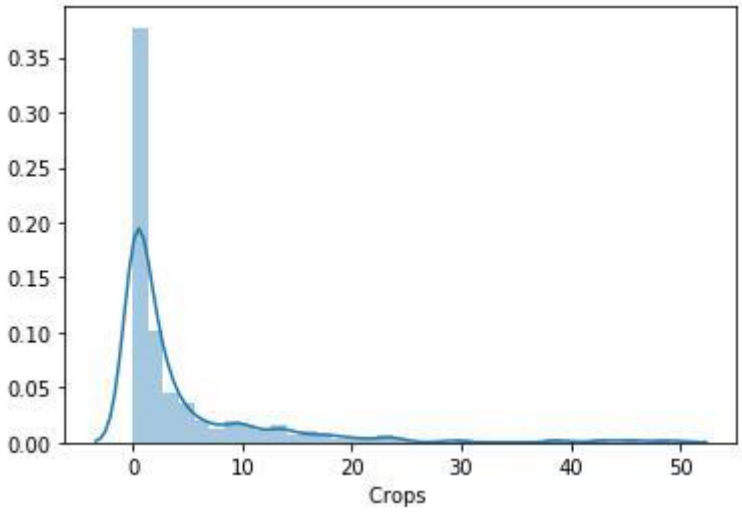
```
countryData['Country'].iloc[countryData.index[countryData['Crops ']==countryData['Crops '].max()].tolist()]
```

```
172    Sao Tome & Principe
```

```
countryData['Country'].iloc[countryData.index[countryData['Crops ']==countryData['Crops '].min()].tolist()]
```

```
4                   Andorra
6                  Anguilla
10                    Aruba
22                  Bermuda
38           Cayman Islands
55                 Djibouti
66           Faroe Islands
78               Gibraltar
80               Greenland
91                 Iceland
97             Isle of Man
102                 Jersey
116          Liechtenstein
119                  Macau
134                 Monaco
135               Mongolia
138                Namibia
139                  Nauru
142    Netherlands Antilles
149                 Norway
168    St Pierre & Miquelon
171              San Marino
178               Singapore
202                  Tuvalu
```

From above I can conclude that crops mean is 4.29% and median is 1.055%. The data is right skewed, mean is greater than the median. The country with the largest proportion of crops land is São Tomé and Príncipe with 48.96%. There are 24 countries with a proportion of land crops of zero.



Other. Percentage of permanent meadows and pastures.

```
countryData['Other '].describe()
```

```
count    218.000000
mean      81.709817
std       16.109662
min       33.330000
25%       71.817500
50%       85.885000
75%       95.335000
max      100.000000
```

```
countryData['Country'].iloc[countryData.index[countryData['Other ']==countryData['Other '].max()].tolist()]
```

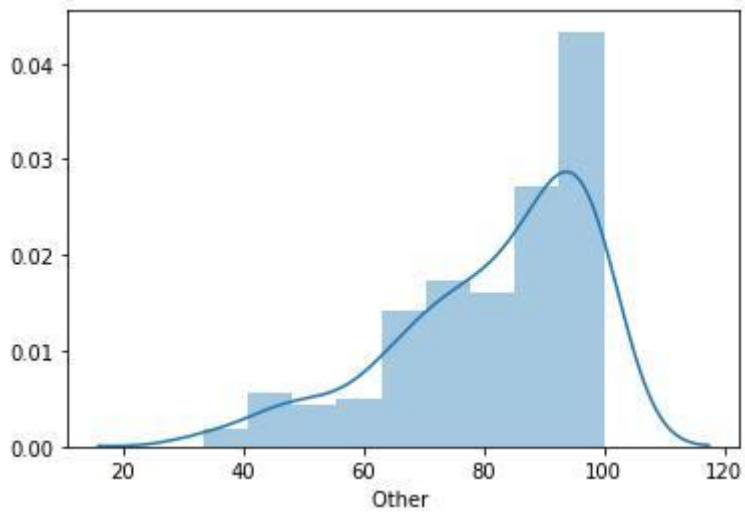
```
6       Anguilla
78     Gibraltar
80     Greenland
102       Jersey
119        Macau
134       Monaco
139        Nauru
202       Tuvalu
```

```
countryData['Country'].iloc[countryData.index[countryData['Other ']==countryData['Other '].min()].tolist()]
```

```
197    Tonga
```

From the previous page I can conclude that the proportion mean of permanent meadows and pasture land is 81.71% and median is 85.88%. The data is left skewed, median is greater than the mean. The countries with the largest proportion of permanent meadows and pasture land are Anguilla, Gibraltar, Greenland, Jersey, Macau, Monaco, Nauru, and Tuvalu

with 100% each. The country with the smallest proportion of permanent meadows and pasture land is Tonga with 33.33%.

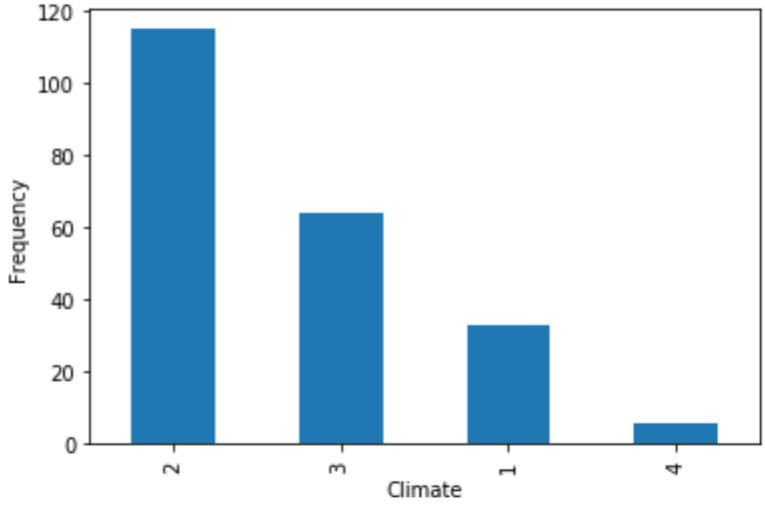


Climate. Categorical variable with 4 categories:

- Category 1. Dry tropical or tundra and ice.
- Category 2. Wet tropical.
- Category 3. Temperate humid subtropical and temperate continental.
- Category 4. Dry hot summers and wet winters.

```
countryData['Climate'].value_counts()
```

```
2    115
3     64
1     33
4      6
```



From the plot above can be conclude that 52.7% on the countries have a wet tropical weather, 29.35% of the countries have a category 1 weather (Temperate humid subtropical and temperate continental), 15.13% of the countries have a dry tropical or tundra and ice

weather, finally 2.75% of the countries have a category 4 weather (Dry hot summers and wet winters).

Birth rate. Compares the average annual number of births during a year per 1,000 people.

```
countryData['Birthrate'].describe()
```

```
count    218.000000
mean      22.074037
std       11.213765
min        7.290000
25%       12.597500
50%       18.750000
75%       29.785000
max       50.730000
```

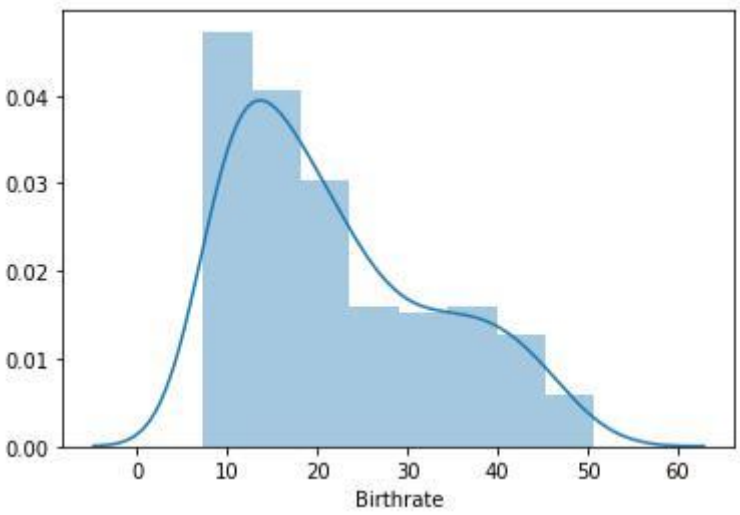
```
countryData['Country'].iloc[countryData.index[countryData['Birthrate']==countryData['Birthrate'].max()].tolist()
```

```
146    Niger
```

```
countryData['Country'].iloc[countryData.index[countryData['Birthrate']==countryData['Birthrate'].min()].tolist()]
```

```
89    Hong Kong
```

From the above can be concluded that the birth rate mean is 22.07 number of births per 1000 people and the median is 18.75 births per 1000 people. The data is right skewed, mean is greater than the median. The country with the largest birth rate is Niger with 50.73 births per 1000 people and the country with the smallest birth rate is Hong Kong with 7.29 births per 1000 people.



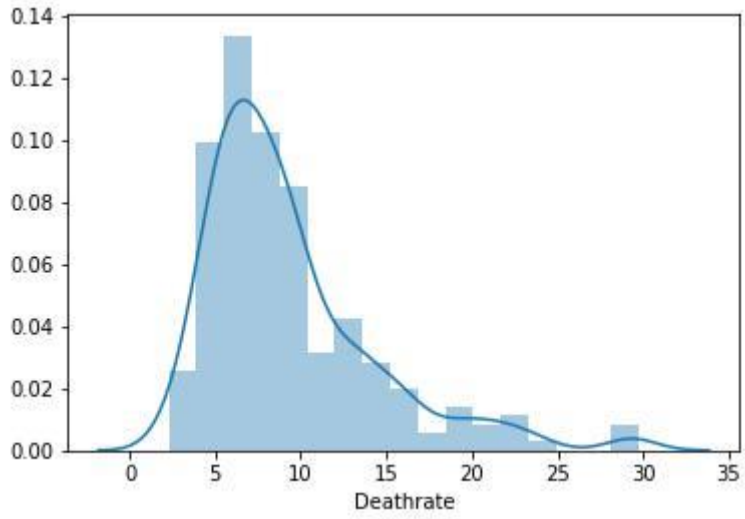
Death rate. Compares the average annual number of deaths during a year per 1,000 people.

```
countryData['Deathrate'].describe()
```

```
count    218.000000
mean       9.339128
std        5.031355
min        2.290000
25%        5.972500
50%        8.175000
75%       11.077500
max       29.740000
```

```
countryData['Country'].iloc[countryData.index[countryData['Deathrate']==countryData['Deathrate'].max()].tolist()]
188    Swaziland
countryData['Country'].iloc[countryData.index[countryData['Deathrate']==countryData['Deathrate'].min()].tolist()]
148    N. Mariana Islands
```

From the above can be concluded that the death rate mean is 9.34 number of deaths per 1000 people and the median is 8.175 deaths per 1000 people. The data is right skewed, mean is greater than the median. The country with the largest death rate is Swaziland with 29.74 deaths per 1000 people and the country with the smallest death rate is Northern Mariana Islands with 2.29 deaths per 1000 people.



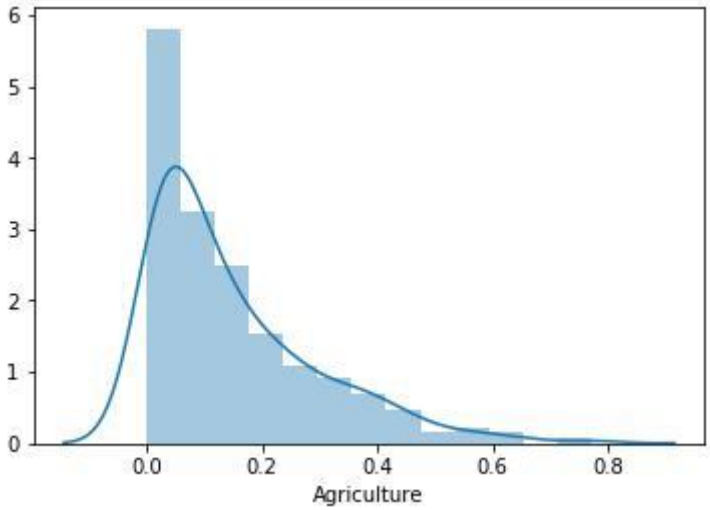
Agriculture. Proportion contribution of this sector in the GDP.

```
countryData['Agriculture'].describe()
count    218.000000
mean       0.148761
std        0.146517
min        0.000000
25%        0.040000
50%        0.100000
75%        0.220000
max        0.770000
```

```
countryData['Country'].iloc[countryData.index[countryData['Agriculture']==countryData['Agriculture'].max()].tolist()]
114    Liberia
countryData['Country'].iloc[countryData.index[countryData['Agriculture']==countryData['Agriculture'].min()].tolist()]
```

```
10          Aruba
78      Gibraltar
89      Hong Kong
108        Kuwait
119         Macau
134        Monaco
161         Qatar
171    San Marino
178     Singapore
```

From the above can be concluded that the mean proportion of agriculture is 14.88% and the median is 10%. The data is right skewed, mean is greater than the median. The country with the largest proportion of agriculture in its GDP is Liberia with 77%. The countries with the smallest proportion of agriculture in their GDP are Aruba, Gibraltar, Hong Kong, Kuwait, Macau, Monaco, Qatar, San Marino, and Singapore with 0% each.



Industry. Proportion contribution of this sector in the GDP.

```
countryData['Industry'].describe()
```

```
count    218.000000
mean       0.282156
std        0.140320
min        0.000000
25%        0.190000
50%        0.270000
75%        0.340000
max        0.910000
```

```
countryData['Country'].iloc[countryData.index[countryData['Industry']==countryData['Industry'].max()].tolist()]
```

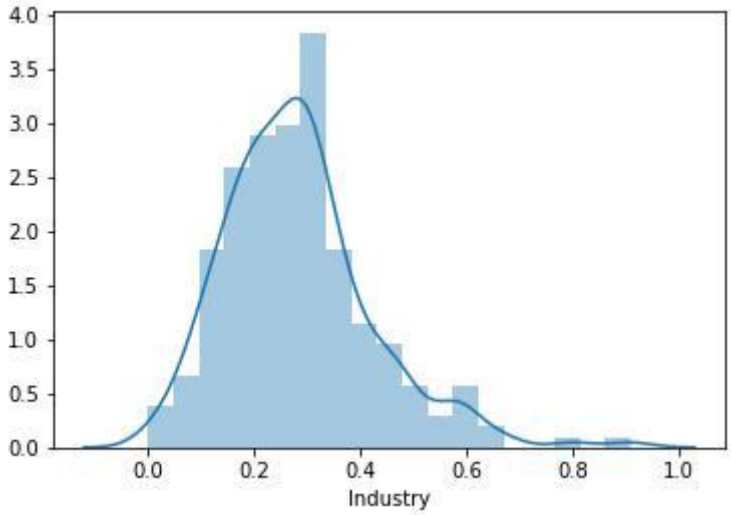
```
62    Equatorial Guinea
```

```
countryData['Country'].iloc[countryData.index[countryData['Industry']==countryData['Industry'].min()].tolist()]
```

```
78    Gibraltar
```

From the above can be concluded that the mean proportion of industry is 28.22% and the median is 27%. The data is right skewed, mean is greater than the median. The country with

the largest proportion of industry in its GDP is Equatorial Guinea 91%. The country with the smallest proportion of industry in its GDP is Gibraltar with 0%.



Service. Proportion contribution of this sector in the GDP.

```
countryData['Service'].describe()
```

```
count    218.000000
mean       0.569679
std        0.167460
min        0.060000
25%        0.440000
50%        0.580000
75%        0.680000
max        1.000000
```

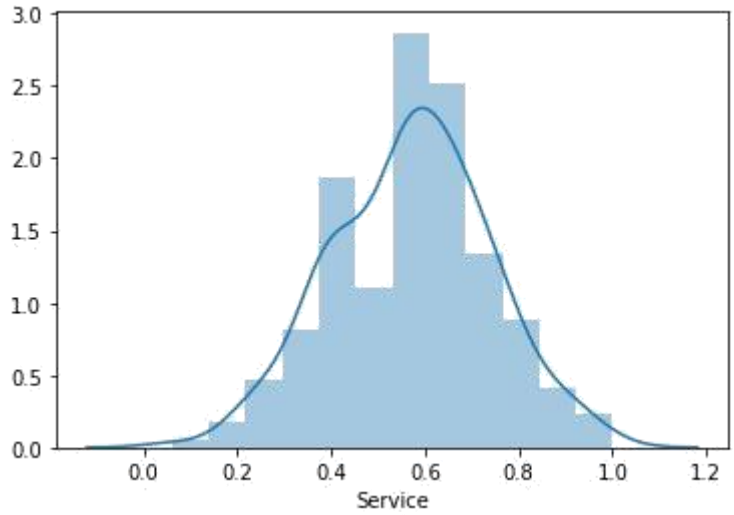
```
countryData['Country'].iloc[countryData.index[countryData['Service']==countryData['Service'].max()].tolist()]
```

```
78    Gibraltar
```

```
countryData['Country'].iloc[countryData.index[countryData['Service']==countryData['Service'].min()].tolist()]
```

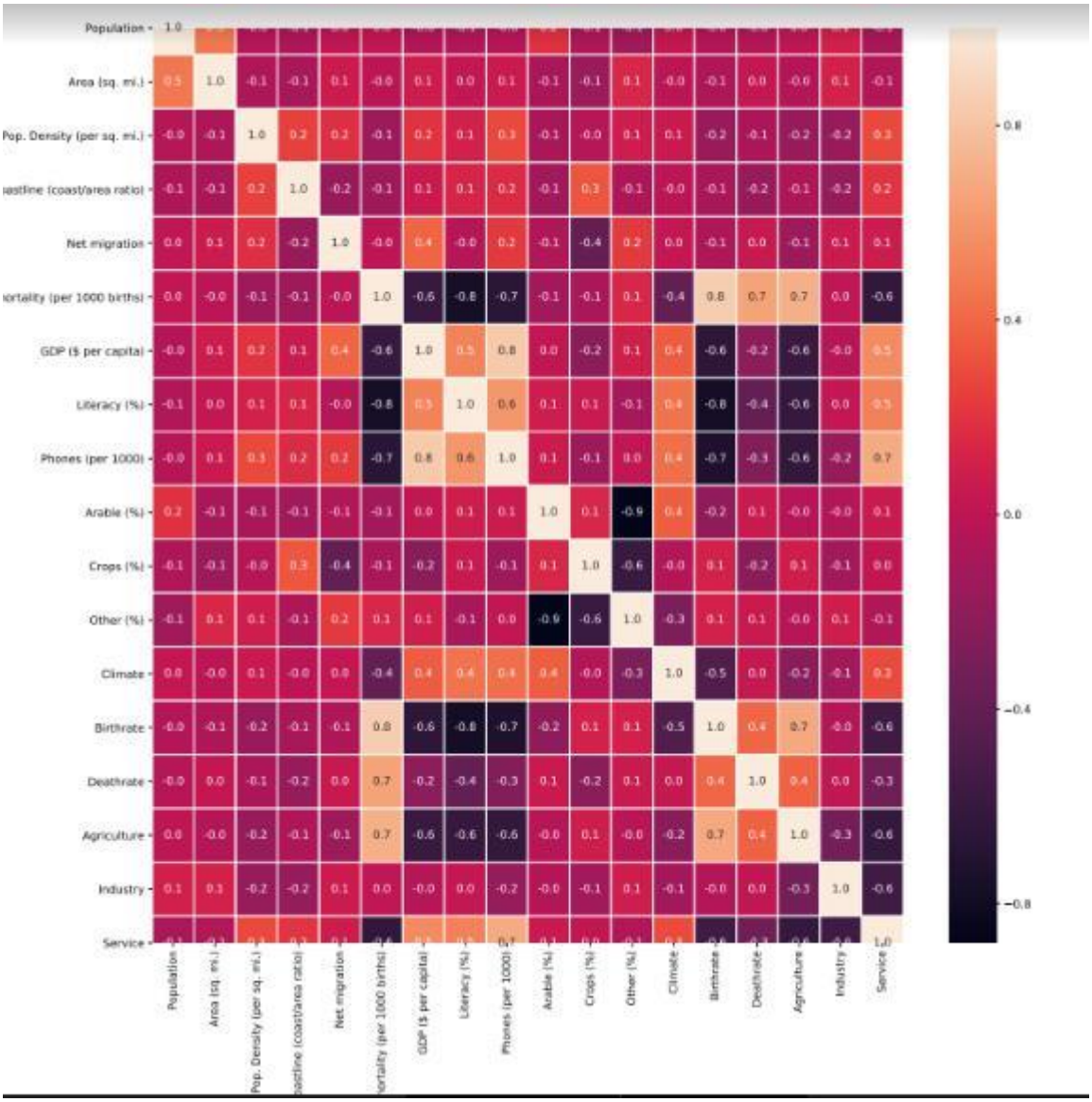
```
62    Equatorial Guinea
```

From the above can be concluded that the mean proportion of service is 56.97% and the median is 58%. The data is approximately normal distributed mean and median are almost equal. The country with the largest proportion of service in its GDP is Gibraltar with 100%. The country with the smallest proportion of Service in its GDP is Equatorial Guinea with 6%.


3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Service

C. Inference from bivariant charts.

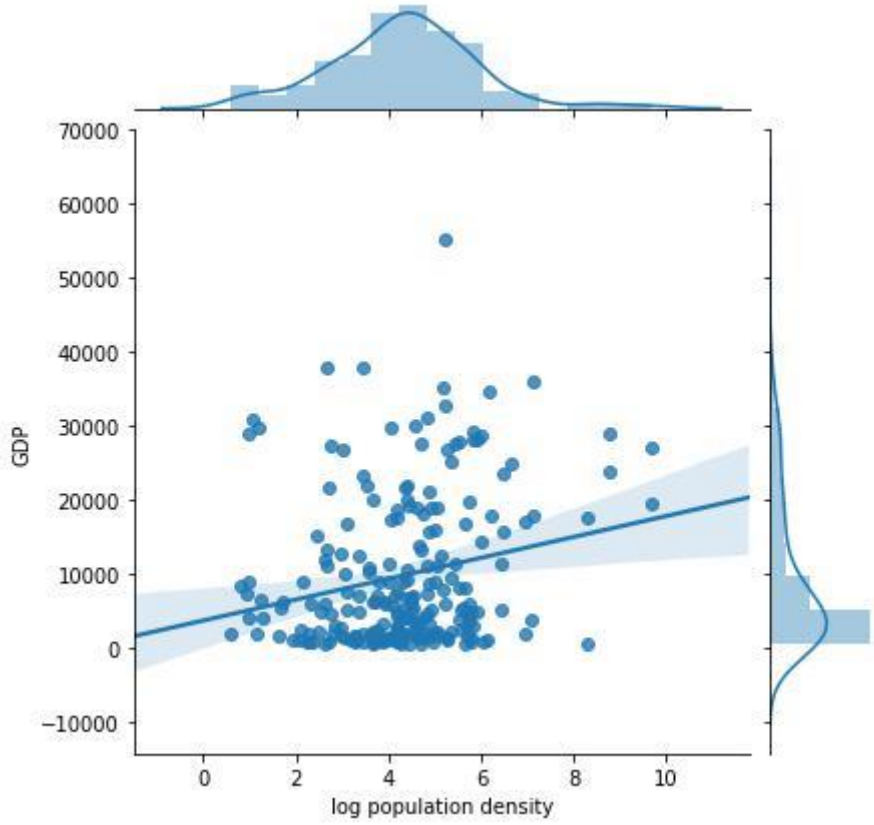
The below is the initial raw data correlation matrix giving me an overview on how each of the factors are correlated.



1) Population Density:

We have normalized the population density using log to meet the normality assumptions. I see that there is some positive correlation between GDP and log population density. From this I infer that GDP increases by some factor as population density increases.

sns.jointplot(x="log population density", y="GDP", data=countryData, kind="reg")



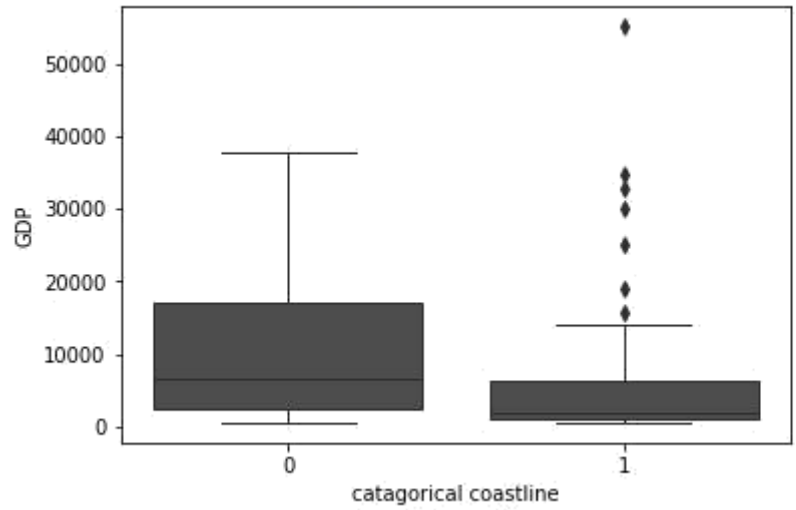
2) Coastline:

sns.boxplot(x="catagorical coastline", y="GDP", data=countryData, color=".3", linewidth=1)

After ignoring a few exceptions, I infer that countries which have a coastline have higher GDP and countries without coastline have lesser GDP.

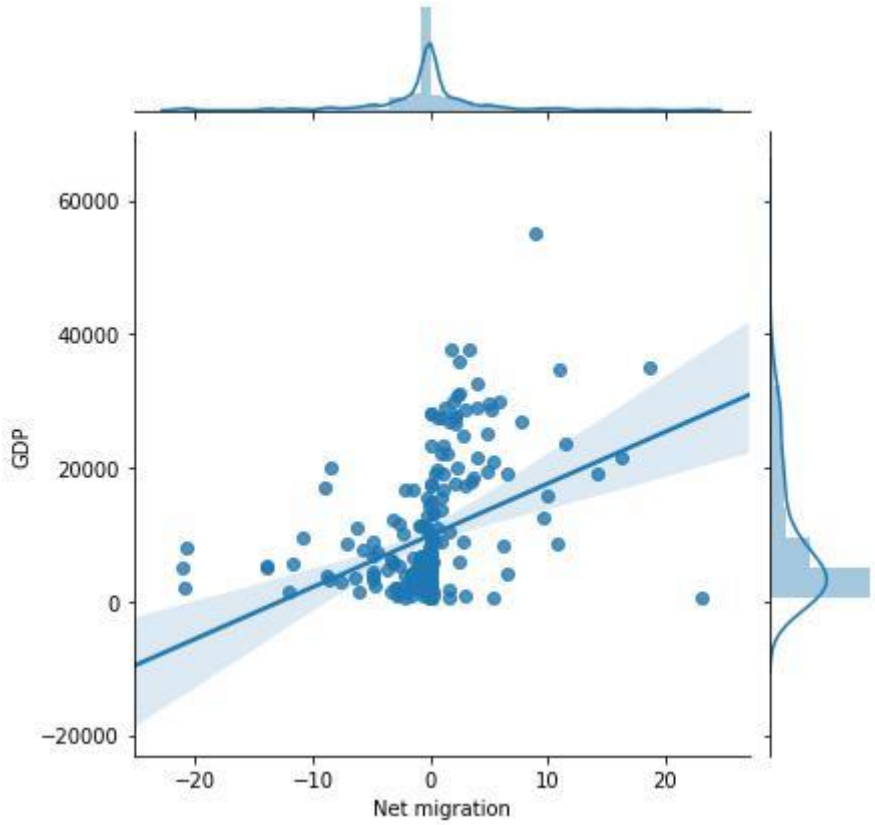
Some countries like Luxemburg are an exception to this. Their high literacy, low infant mortality, fair use of technology let them come up with more innovative ideas to boost their economy rather than relying on coastline.

```
Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x1b1f05d35f8>
```



3) Net Migration:

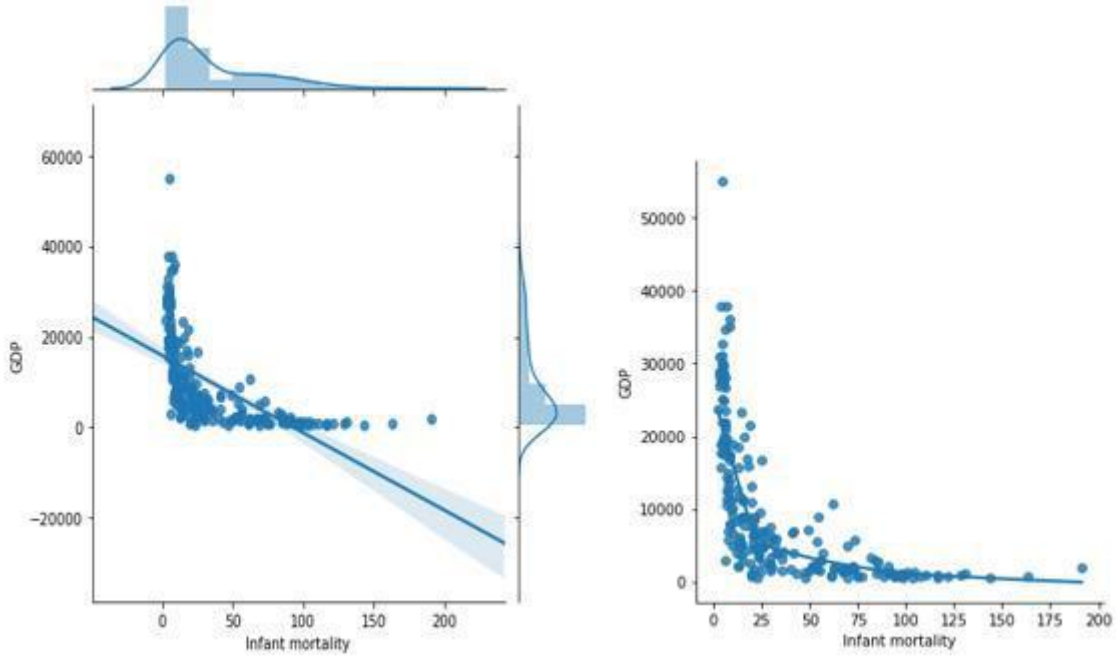
Net Migration and GDP are positively correlated. With larger immigration there is higher per capita GDP.



4) Infant Mortality:

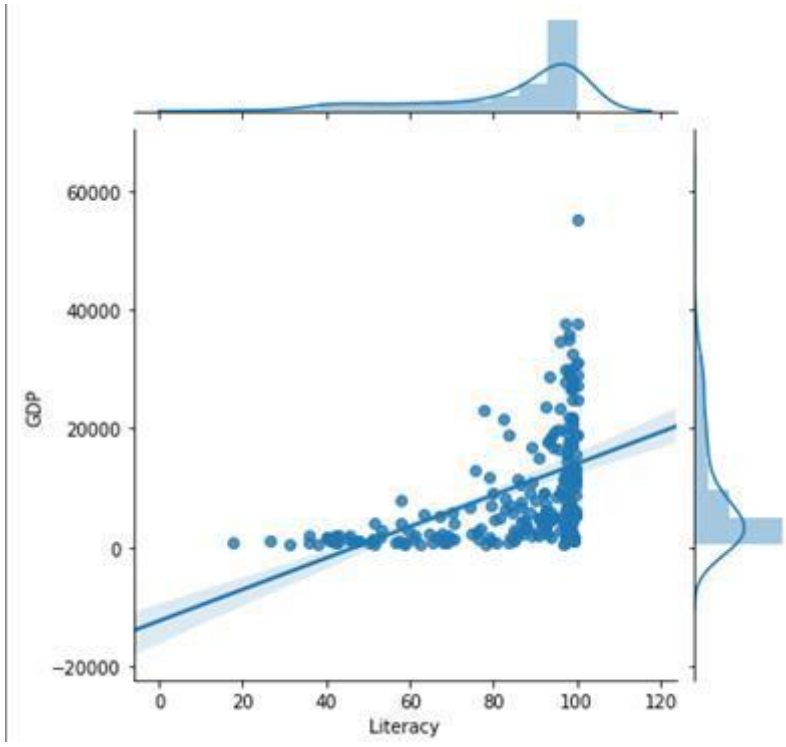
Infant mortality rate (IMR) is the number of deaths per 1,000 live births of children under one year of age.

From below graphs I infer that most of the countries have less infant mortality rate. This data cant normalized as I can't take random samples of infant mortality. Its definitive measure.



5) Literacy:

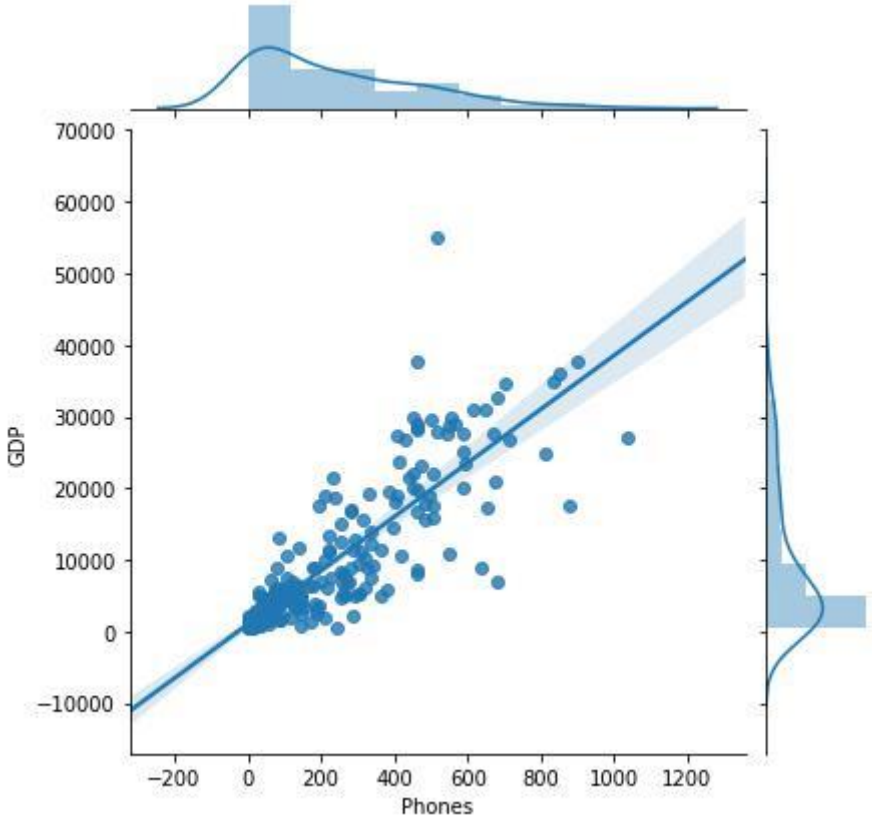
Most countries have high literacy rates. I see a positive correlation between literacy and GDP. In some places where the literacy rate is low, the GDP is too low.



6) Phone (Technology)

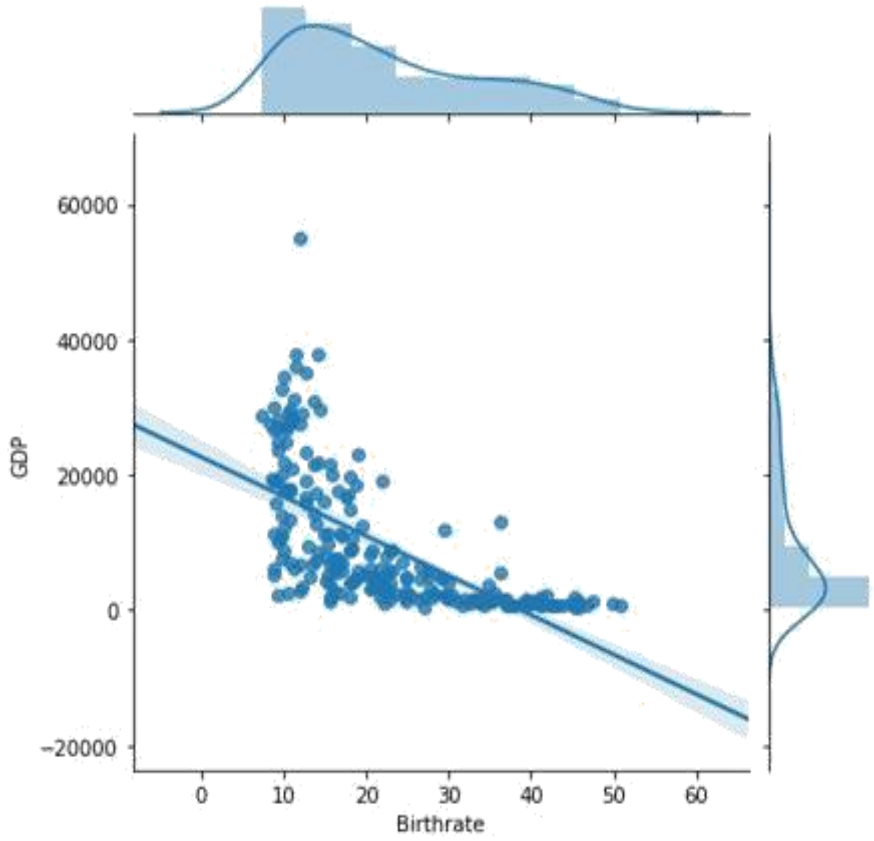
Below graph indicates that the increase in the use of technology indicates the high rate of GDP. Since it's a discrete variable I cannot explain the randomness of the observations.

Countries with high GDP has access to good education system, making them easy to understand and use high tech machines, therefore, it explains why countries with high GDP rates have high use of technology.



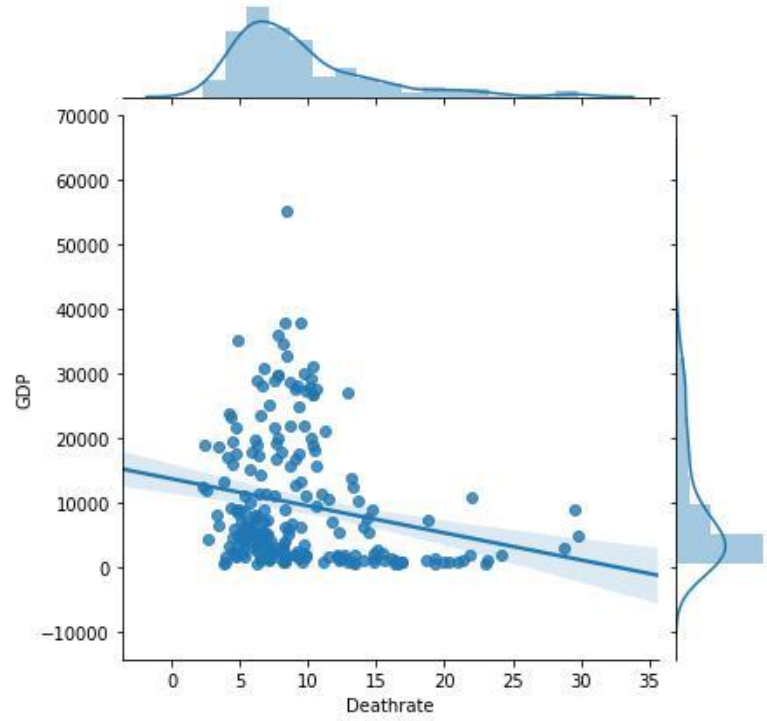
7) Birth Rate:

Birth Rate and GDP are inversely related. A highly educated country with better facilities can keep the birthrate in control.



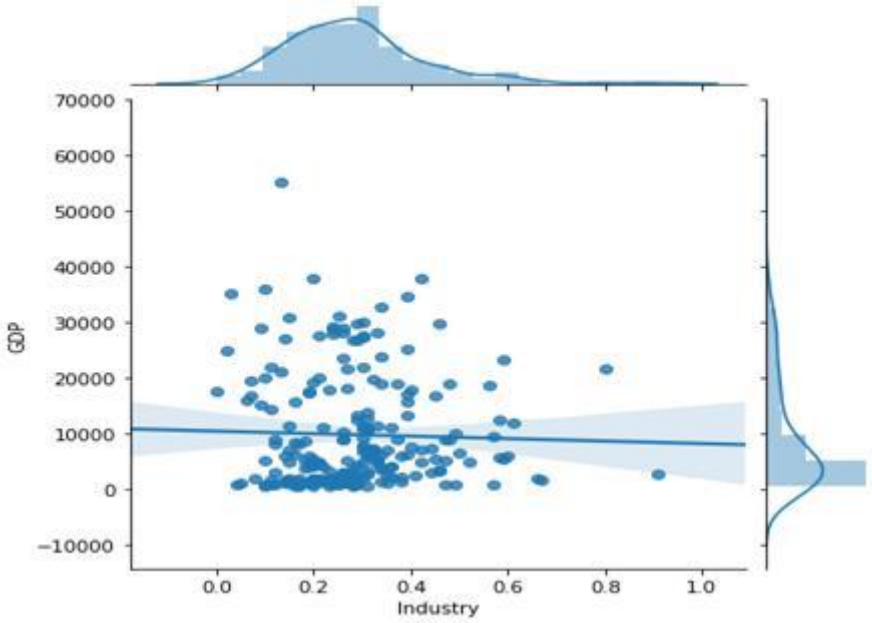
8) Death Rate:

There is very less negative correlation between Death rate and GDP.



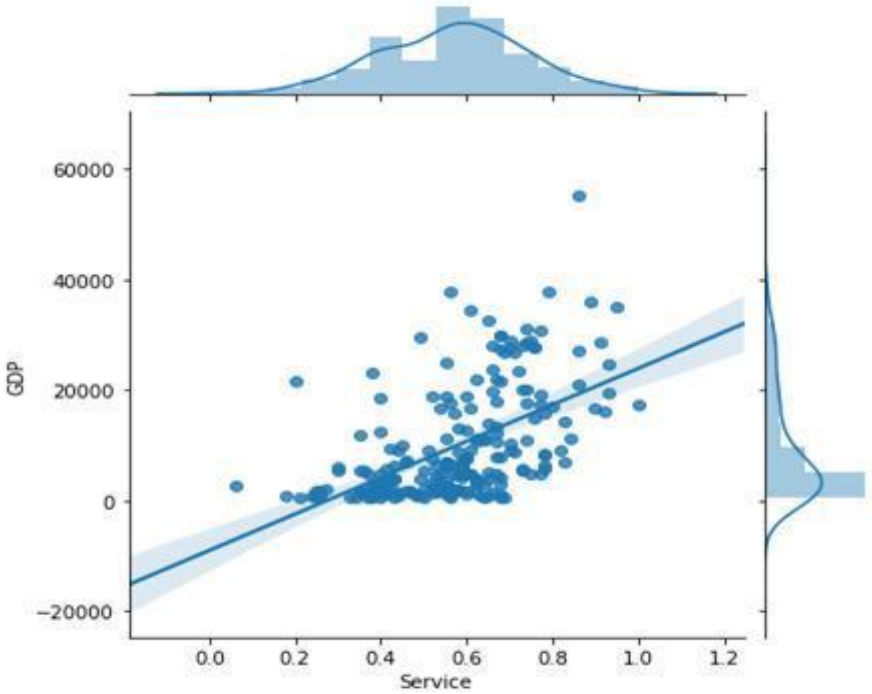
9) Industry:

We don’t see any relationship between Industry and GDP.



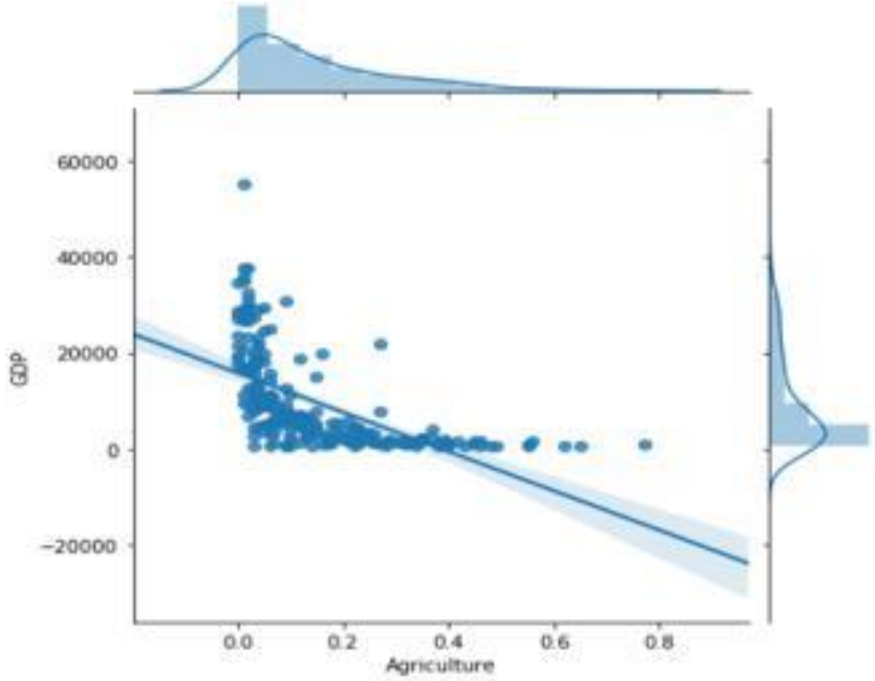
10) Service:

Service and GDP are positively correlated. In developed countries where there is a huge network of services generate higher GDP.



11) Agriculture:

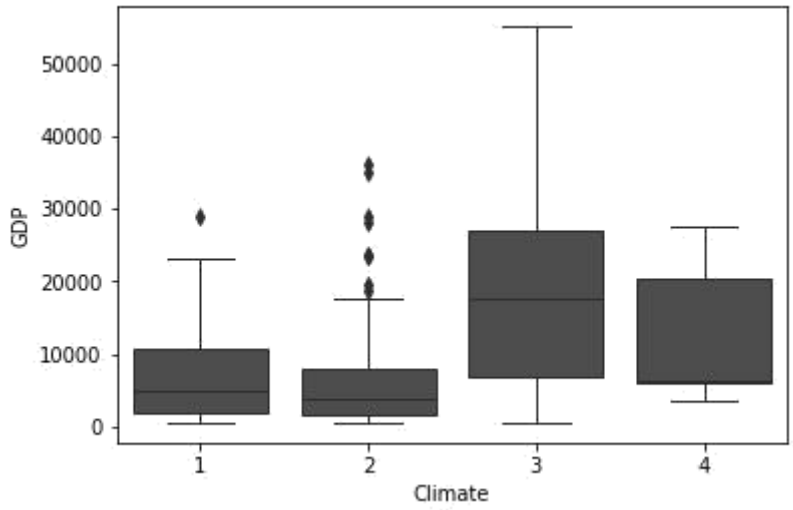
Agriculture is significant only in developing countries where GDP is moderate. Countries with higher GDP may generate GDP through services, technology and other factors more rather than agriculture.



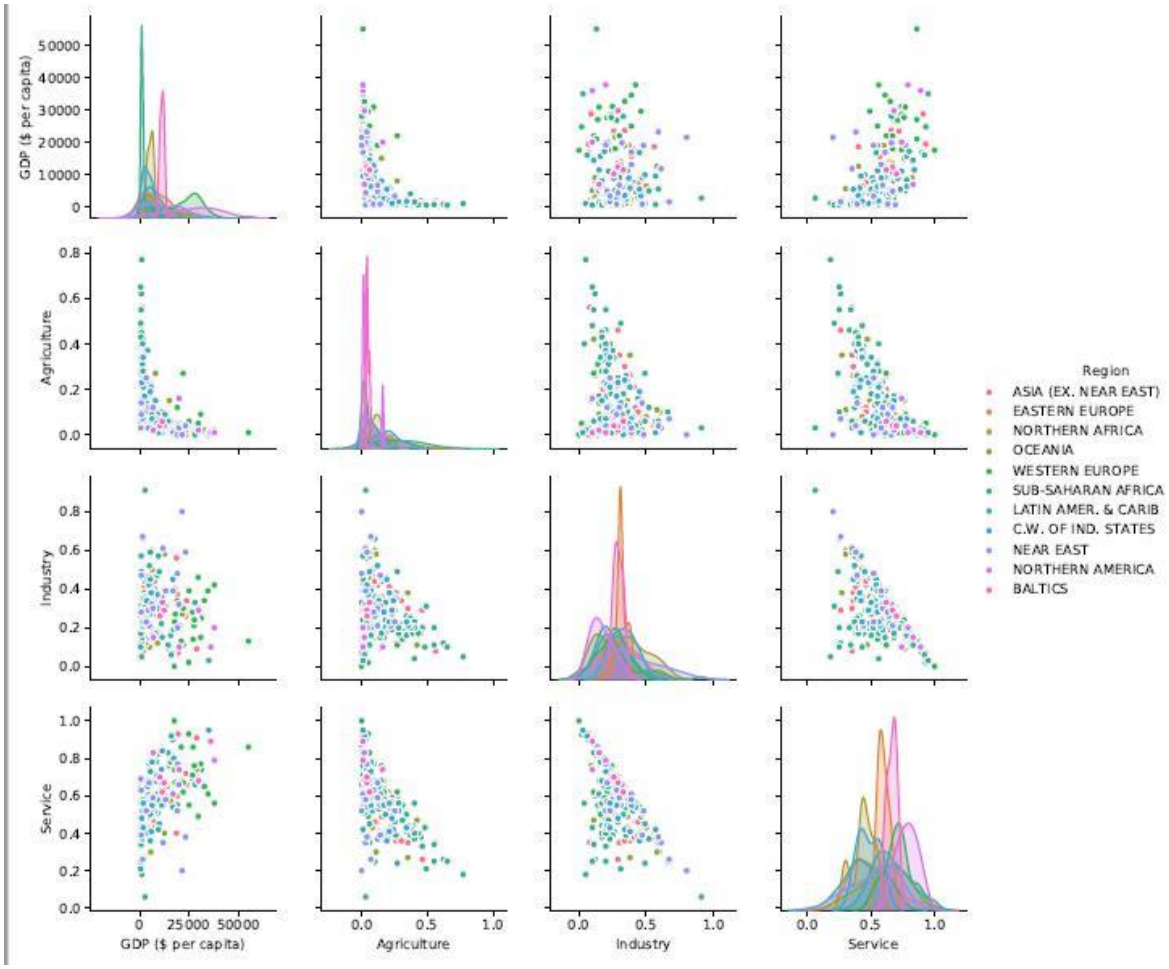
12) Climate:

Countries with Temperate /humid/tropical / subtropical climates have higher GDP.

1. Dry tropical/tundra/ice
2. Wet tropical
3. Temperate /humid/tropical / subtropical
4. Dry hot summers and wet winters



D. Inference from multivariate charts.



The above multivariate graph shows that Western Europe is a more developed region and has higher GDP. Next comes Asia. Sub-Saharan Africa, C.W of Ind. States, Baltics are regions which are developing and have a low GDP.

E. Predictions and models

Predicting Energy Consumption:

The following regression analysis is based on a study conducted by Gokhan Aydin (2014). Energy consumption is modelled by regression analysis based on population and gross domestic product. Because the data set used for this project did not include the energy consumption variable, data scraping was used to obtain this information.

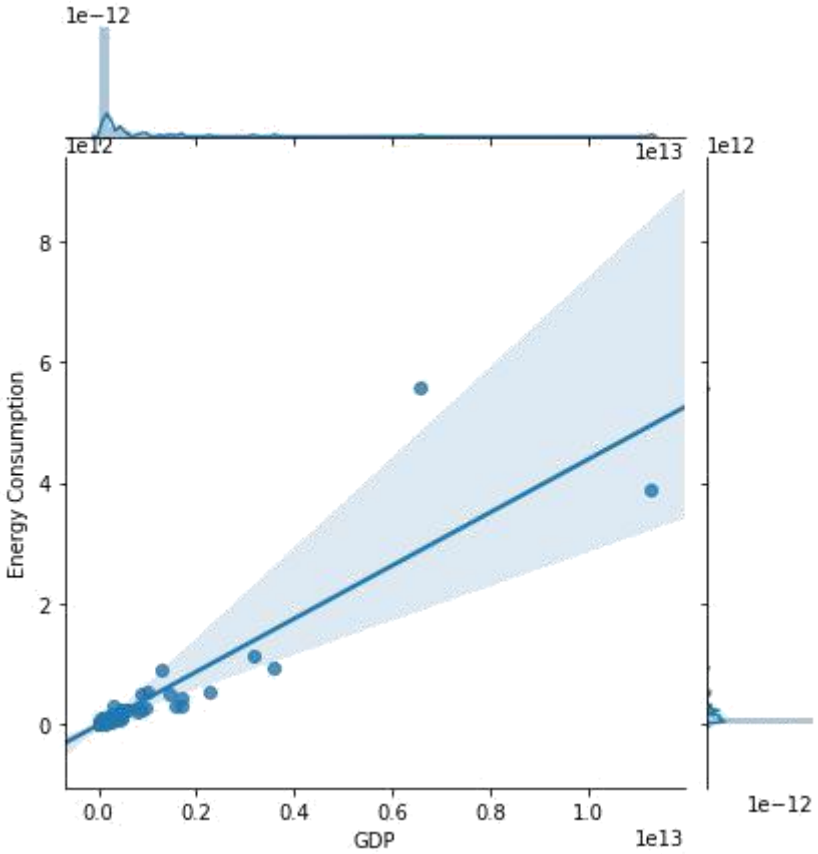
Table 2. Energy consumption, population, and GDP per country.

	Country	Energy Consumption	Population	GDP
0	China	5.564000e+12	1313973713	6569868565000
1	United States	3.902000e+12	298444215	11281191327000
2	India	1.137000e+12	1095351995	3176520785500
3	Japan	9.437000e+11	127463611	3594473830200
4	Russia	9.096000e+11	142893540	1271752506000
..	...	...	...	...
186	Comoros	3.906000e+07	690948	483663600
187	Guinea-Bissau	3.627000e+07	1442029	1153623200
188	Cook Islands	3.162000e+07	21388	106940000
189	Nauru	2.232000e+07	13287	66435000
190	Gaza Strip	2.020000e+05	1428757	857254200

For the regression model energy consumption is the dependent variable, population and GDP independents variables.

Correlation between energy consumption and GDP.

```
import seaborn as sns

sns.jointplot(x='GDP', y='Energy Consumption', data=newcountries3, kind='reg')

newcountries3['Energy Consumption'].corr(newcountries3['GDP'])
0.8955018627290472
```

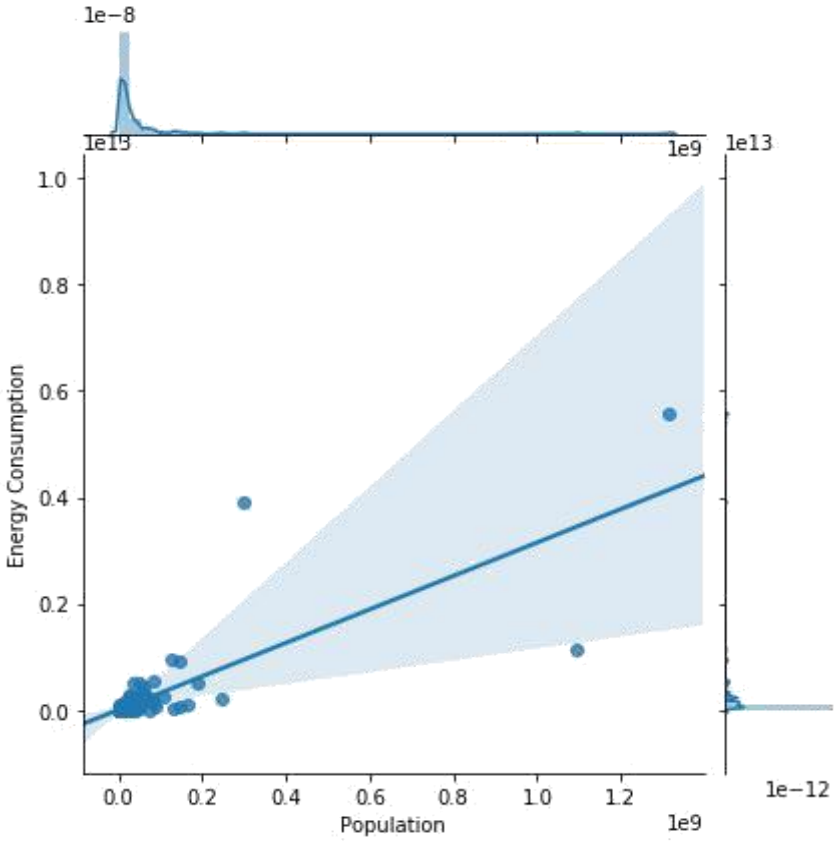


From above can be conclude that there is a strong positive correlation between energy consumption and GDP.

Correlation between energy consumption and population.

```
sns.jointplot(x='Population', y='Energy Consumption', data=newcountries3, kind='reg')
```

```
newcountries3['Energy Consumption'].corr(newcountries3['Population'])
0.790528233245963
```



From above can be conclude that there is a strong positive correlation between energy consumption and population.

A regression model is built to predict energy consumption on the basis of population and GDP.

Energy Consumption = 1579176268.743164 +188.77366296*Population+0.33089406*GDP

R^2 = 0.6758822131021183

R^2 adjusted = 0.6724341515393748

67% of the variability in energy consumption can be explained by population and GDP. The other portion must be explained by other factors.

Linear regression assumptions were checked for this model and all of them were violated. A transformation log was applied to the data.

Table 3. Transformation log of energy consumption, population, and GDP per country.

	Country	Energy Consumption log	Population log	GDP log
0	China	29.347338	20.996322	29.513515
1	United States	28.992510	19.514094	30.054158
2	India	27.759414	20.814342	28.786808
3	Japan	27.573074	18.663341	28.910419
4	Russia	27.536271	18.777610	27.871417
..	...	...	...	...
186	Comoros	17.480609	13.445820	19.996900
187	Guinea-Bissau	17.406502	14.181562	20.866173
188	Cook Islands	17.269300	9.970585	18.487778
189	Nauru	16.920994	9.494541	18.011735
190	Gaza Strip	12.216023	14.172315	20.569245

Energy Consumption log = -3.2226894937679873 - 0.25961265*Population log + 1.25123995*GDP log

R^2 = 0.9245368241376731

R^2 adjusted = 0.9237340243944568

92% of the variability in energy consumption log can be explained by population log and GDP log. The other portion must be explained by other factors.

To build other regression models, the correlation among all numerical variables is calculated.

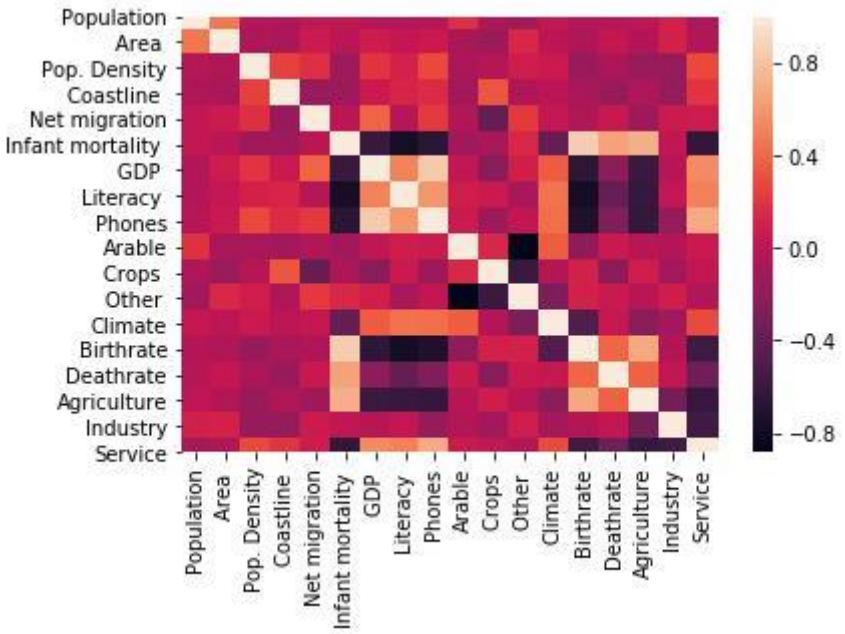
```
countries=pd.read_csv('countriesoftheworld.csv')

correlationdf=countries.corr()
print(correlationdf)
```

Table 4. Correlation among numerical variables.

	Population	Area	...	Industry	Service
Population	1.000000	0.468358	...	0.097473	-0.085165
Area	0.468358	1.000000	...	0.109688	-0.053565
Pop. Density	-0.029171	-0.068227	...	-0.164344	0.281215
Coastline	-0.064602	-0.090616	...	-0.171213	0.200203
Net migration	0.009768	0.059876	...	0.063854	0.068846
Infant mortality	0.021207	-0.009912	...	0.012838	-0.629663
GDP	-0.041379	0.070970	...	-0.030871	0.544785
Literacy	-0.053059	0.026239	...	0.028204	0.504064
Phones	-0.028556	0.058619	...	-0.173899	0.686750
Arable	0.186045	-0.085083	...	-0.042055	0.055806
Crops	-0.058144	-0.146327	...	-0.111255	0.017747
Other	-0.124003	0.139421	...	0.087493	-0.054029
Climate	0.035032	-0.001836	...	-0.115511	0.290589
Birthrate	-0.045117	-0.066237	...	-0.044979	-0.558597
Deathrate	-0.032931	0.033853	...	0.016972	-0.339375
Agriculture	0.007003	-0.044197	...	-0.315715	-0.609607
Industry	0.097473	0.109688	...	1.000000	-0.558151
Service	-0.085165	-0.053565	...	-0.558151	1.000000

```
import seaborn as sns
sns.heatmap(correlationdf)
```

Correlation heatmap.



From above, literacy can be predicted on the basis of GDP, infant mortality, phones, and birth rate. Also, infant mortality can be predicted on the basis of GDP, literacy, agriculture, birth rate, and phones.

Predicting Literacy.

It was concluded from the correlation heatmap that literacy can be predicted on the basis of GDP, infant mortality, phones, and birth rate.

```
countries2=pd.concat([countries['Literacy '],countries['Infant mortality '],countries['GDP '],countries['Phones'],countries['Birthrate']],axis=1)

print(countries2)
```

Table5. Literacy, infant mortality, GDP, phones, and birthrate per country.

	Literacy	Infant mortality	GDP	Phones	Birthrate
0	36.00	163.07	700	3.2	46.60
1	86.50	21.52	4500	71.2	15.11
2	70.00	31.00	6000	78.1	17.14
3	97.00	9.27	8000	259.5	22.46
4	100.00	4.05	19000	497.2	8.71
..	...	...	...	...	...
213	99.00	8.03	17200	652.8	13.96
214	97.22	19.62	800	145.2	31.67
215	50.20	61.50	800	37.2	42.89
216	80.60	88.29	800	8.2	41.00
217	90.70	67.69	1900	26.8	28.01

Literacy = 106.70373858929594 -2.14326759e-01*Infant mortality - 1.76185927e-04*GDP + 9.44272674e-03*Phones - 7.16700770e-01*Birthrate

R- square = 0.7910148235286492

Adjusted R – square = 0.787090219275666

79% of the variability in literacy can be explained infant mortality, GDP, phones, and birth rate. The other portion must be explained by other factors.

Linear regression assumptions were checked and met by the model.

Predicting Infant Mortality.

It was concluded from the correlation heatmap that infant mortality can be predicted on the basis of GDP, literacy, agriculture, birth rate, and phones.

```
countries3=pd.concat([countries['Infant mortality '],countries['Literacy '],countries['GDP '],countries['Phones'],countries['Birthrate'],countries['Agriculture']],axis=1)
print(countries3)
```

Table 6. Infant mortality, literacy, GDP, phones, birth-rate, agriculture per country.

	Infant mortality	Literacy	GDP	Phones	Birthrate	Agriculture
0	163.07	36.00	700	3.2	46.60	0.38
1	21.52	86.50	4500	71.2	15.11	0.23
2	31.00	70.00	6000	78.1	17.14	0.10
3	9.27	97.00	8000	259.5	22.46	0.27
4	4.05	100.00	19000	497.2	8.71	0.12
..	...	...	...	...	...	...
213	8.03	99.00	17200	652.8	13.96	0.01
214	19.62	97.22	800	145.2	31.67	0.09
215	61.50	50.20	800	37.2	42.89	0.14
216	88.29	80.60	800	8.2	41.00	0.22
217	67.69	90.70	1900	26.8	28.01	0.18

Infant Mortality = 37.46180406923113 - 4.74253812e-01*Literacy + 3.48845353e-05*GDP - 1.33317006e-02*Phones + 1.59955122e+00*Birthrate +4.36334373e+01*Agriculture

R^2 = 0.7093735706529352

R^2 adjusted = 0.7025191737343723

70% of the variability in literacy can be explained infant mortality, GDP, phones, and birth rate. The other portion must be explained by other factors.

Linear regression assumptions were checked and met by the model.

Regression Analysis to predict GDP and factors affecting GDP.

This regression analysis could help find the GDP of the country based on some factors which are very significant to shape the GDP of the country. To start with, I took the reference of the EDA and made sure which columns are correlated with my response variable “GDP”. There I saw that columns like Population, Area, Arable, Crops, Other and Death rate either made less sense or had no significance to the GDP variable. Hence they were not considered.

Based on the variables I had I made a linear regression model and computed its coefficients and intercept.

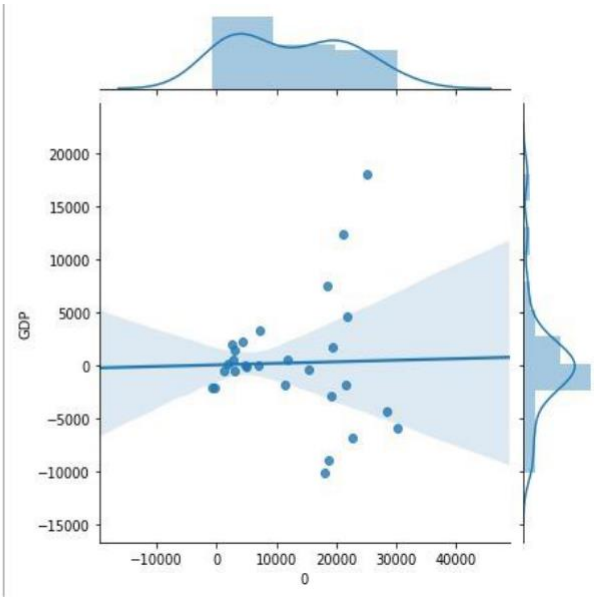
GDP=11661.585593486609+ 389.31206068 (Net Migration)-40.62119415 (Infant Mortality) -32.35065671 (Literacy) +31.61864295 (Phones) -28.2403482 (Birth rate) -7317.89149118 (Agriculture) -4278.94471318 (Service) -383.41104039 (log population density) + 1196.71700527 (categorical coastline) + 906.45777607 (Dry tropical) + 138.4195533 (Wet tropical) + 821.62072849 (Temperate) -1866.49805785 (Dry hot summers and wet winters)

The variation of GDP with respect to factors like Net Migration, infant mortality, literacy, phones,

services, population, coastline, climate is around 80.15%.

After checking for regression assumptions, p-values some of the predictor variables in the model was omitted by making an educated guess. This step was taken to get rid of multi-collinearity and variables with less significance.

```
    ...: print(p_value)
3.1443774014634358e-09
8.644728708991333e-23
1.0306963855399722e-15
9.457113578628043e-59
2.8471372302727404e-27
4.297584063240643e-22
3.0095445502416146e-18
0.0011496060241279538
0.08010840857867521
0.19401385016941705
1.2705292881708723e-08
5.635247881001433e-13
0.5120356521610565
```



Model 2.

Variables removed: log(population density), Infant Mortality, Birth rate, Agriculture

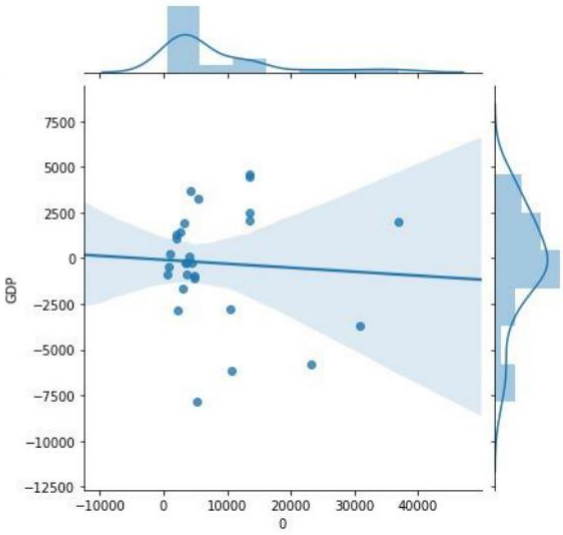
```
...: print(r2)
Net migration
Literacy
Phones
Service
catagorical coastline
isClimate1
isClimate2
isClimate3
isClimate4
GDP
[  432.88852888    35.01271072    32.56431735 -1340.54097876
  -535.86429181   435.86785249  -542.583237    1160.33772294
 -1053.62233843]
41.488953886389936
0.8992972293502963
```

GDP= 41.488953886389936 + 432.88852888 (Net Migration) + 35.01271072 (Literacy) + 32.56431735 (Phones) -1340.54097876 (Service) - 535.86429181 (categorical Coastline) + 435.86785249 (Dry tropical) -542.583237 (Wet tropical)+ 1160.33772294 (Temperate) - 1053.62233843 (Dry hot summers and wet winters)

89.99 % of variability in the model an be attributed to Net Migration, Literacy, Technology (Phones), Service, Coastline existence and climate.

From the residual analysis of errors vs prediction, I don't see heteroskedasticity. Also, the P-values of the factors support the fact that they are more significant than the previous model.

```
...: print(p_value)
3.1443774014634358e-09
1.0306963855399722e-15
9.457113578628043e-59
3.0095445502416146e-18
0.08010840857867521
0.19401385016941705
1.2705292881708723e-08
5.635247881001433e-13
0.5120356521610565
```



F. Group member Contribution

Chinmayi Mahadik: Presentation, EDA, Data preprocessing

Julian Munoz-Ramirez: Web scrapping, model building, EDA, Data preprocessing

Prathamesh Dhapodkar: Model building, EDA, Data preprocessing

Prerana Ramesh: Report, EDA, Data preprocessing