Student: PinHo Wang
Instructor: Prof. Nik Bear Brown

Countries of the World EDA

Assignment 1

Abstract

In this assignment, we investigate the different kinds of facts in countries of the world including GDP, literacy, phones and so on, and in the end trying to training a model which can predict countries GDP. First of all, we will take a glance to the original data which provide by US government and category them into some subtopics in order to help us understand these data and let us easier to process them for the later on prediction. The tables of these data will be first presented, then, we will use some charts trying to understand the relationship between them.

We will investigate some countries which have higher GDP, then, research the relation between region versus GDP, and the rate of literacy and agriculture. The correlation of the facts will also be researched, which will be plotted as color table. The scatterplot of GDP versus others facts are shown latter on.

Investigation

We first take a peek of the data we read in as listing first five observations. It shows that it seems list by the alpha (first five are 'A'). And the features are showing too.

[2]:

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)
0	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	48.0	0.00	23.06	163.07	700.0	36.0	3.2	12.13	0.22
1	Albania	EASTERN EUROPE	3581655	28748	124.6	1.26	-4.93	21.52	4500.0	86.5	71.2	21.09	4.42
2	Algeria	NORTHERN AFRICA	32930091	2381740	13.8	0.04	-0.39	31.00	6000.0	70.0	78.1	3.22	0.25
3	American Samoa	OCEANIA	57794	199	290.4	58.29	-20.71	9.27	8000.0	97.0	259.5	10.00	15.00
4	Andorra	WESTERN EUROPE	71201	468	152.1	0.00	6.60	4.05	19000.0	100.0	497.2	2.22	0.00

The above shows the data by row, now we take a look into column data, on the other words, the statistical relationship, maximum, minimum, mean, standard, and so on, between features in every countries.

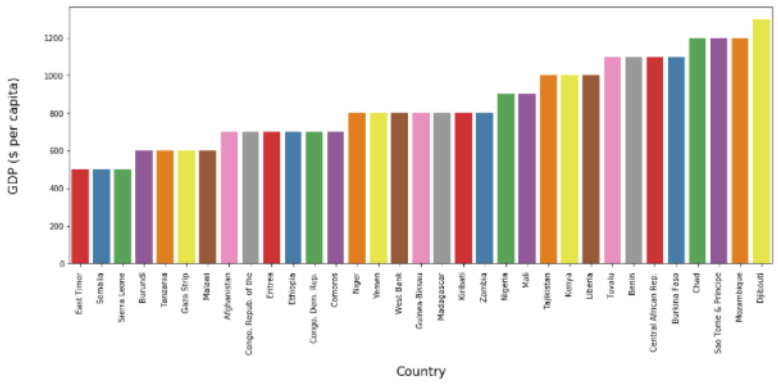
[3]:

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Ar
count	2.270000e+02	2.270000e+02	227.000000	227.000000	224.000000	224.000000	226.000000	209.000000	223.000000	225
mean	2.874028e+07	5.982270e+05	379.047137	21.165330	0.038125	35.506964	9689.823009	82.838278	236.061435	1
std	1.178913e+08	1.790282e+06	1660.185825	72.286863	4.889269	35.389899	10049.138513	19.722173	227.991829	13
min	7.026000e+03	2.000000e+00	0.000000	0.000000	-20.990000	2.290000	500.000000	17.600000	0.200000	0
25%	4.376240e+05	4.647500e+03	29.150000	0.100000	-0.927500	8.150000	1900.000000	70.600000	37.800000	3
50%	4.786994e+06	8.660000e+04	78.800000	0.730000	0.000000	21.000000	5550.000000	92.500000	176.200000	10
75%	1.749777e+07	4.418110e+05	190.150000	10.345000	0.997500	55.705000	15700.000000	98.000000	389.650000	20
max	1.313974e+09	1.707520e+07	16271.500000	870.660000	23.060000	191.190000	55100.000000	100.000000	1035.600000	6

Then, because we want to predict GDP, so it is worthy to know the top 20 GDP countries in these data. So, we sort them by GDP descending. As you can see, the highest GDP country is Luxembourg, then is Norway. Then, we show the sorted result as the histogram.

[5]:

	GDP ($ per capita)	Country
58	500.0	East Timor
188	500.0	Somalia
183	500.0	Sierra Leone
33	600.0	Burundi
200	600.0	Tanzania
74	600.0	Gaza Strip
125	600.0	Malawi
0	700.0	Afghanistan
46	700.0	Congo, Repub. of the
63	700.0	Eritrea
65	700.0	Ethiopia
45	700.0	Congo, Dem. Rep.
44	700.0	Comoros
151	800.0	Niger
224	800.0	Yemen
222	800.0	West Bank
87	800.0	Guinea-Bissau
124	800.0	Madagascar
108	800.0	Kiribati
225	800.0	Zambia



It's important that find out whether or not there is a null value in our data because it may cause difficulty of our numerical analysis. So, we calculate how many null value in each column and also find out how many percentage in all of our observations.

[7]:

	total_missing	percent
Climate	22	9.69
Literacy (%)	18	7.93
Industry	16	7.05
Service	15	6.61
Agriculture	15	6.61
Deathrate	4	1.76
Phones (per 1000)	4	1.76
Birthrate	3	1.32
Infant mortality (per 1000 births)	3	1.32
Net migration	3	1.32
Other (%)	2	0.88
Crops (%)	2	0.88
Arable (%)	2	0.88
GDP ($ per capita)	1	0.44

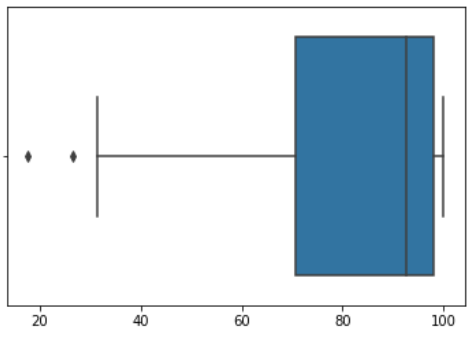
The countries' GDP may also cause by the region it locates, so we gather these countries by region groups, there are eleven groups out there. Then, we investigate the median of GDP, literacy and agriculture of these countries.

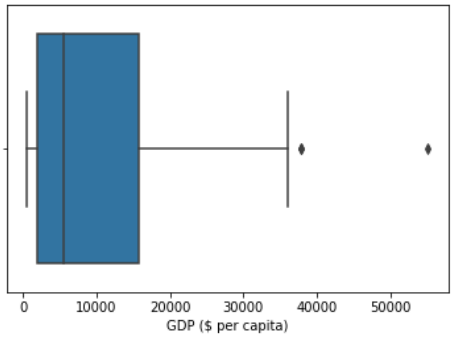
[4]:

Region	GDP ($ per capita)	Literacy (%)	Agriculture
ASIA (EX. NEAR EAST)	3450.0	90.60	0.1610
BALTICS	11400.0	99.80	0.0400
C.W. OF IND. STATES	3450.0	99.05	0.1980
EASTERN EUROPE	9100.0	98.60	0.0815
LATIN AMER. & CARIB	6300.0	94.05	0.0700
NEAR EAST	9250.0	83.00	0.0350
NORTHERN AFRICA	6000.0	70.00	0.1320
NORTHERN AMERICA	29800.0	97.50	0.0100
OCEANIA	5000.0	95.00	0.1505
SUB-SAHARAN AFRICA	1300.0	62.95	0.2760
WESTERN EUROPE	27200.0	99.00	0.0220

Then, we plot the box digram of GDP and literacy rate to show the distribution in all range.

```
[5]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1b14ca5[6]: <matplotlib.axes._subplots.AxesSubplot at 0x107f7d9b0>
```

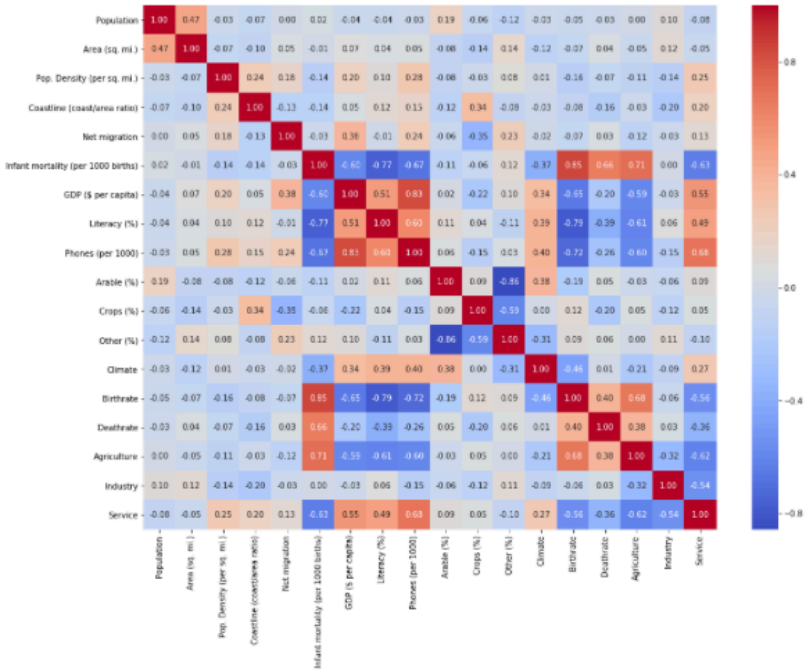




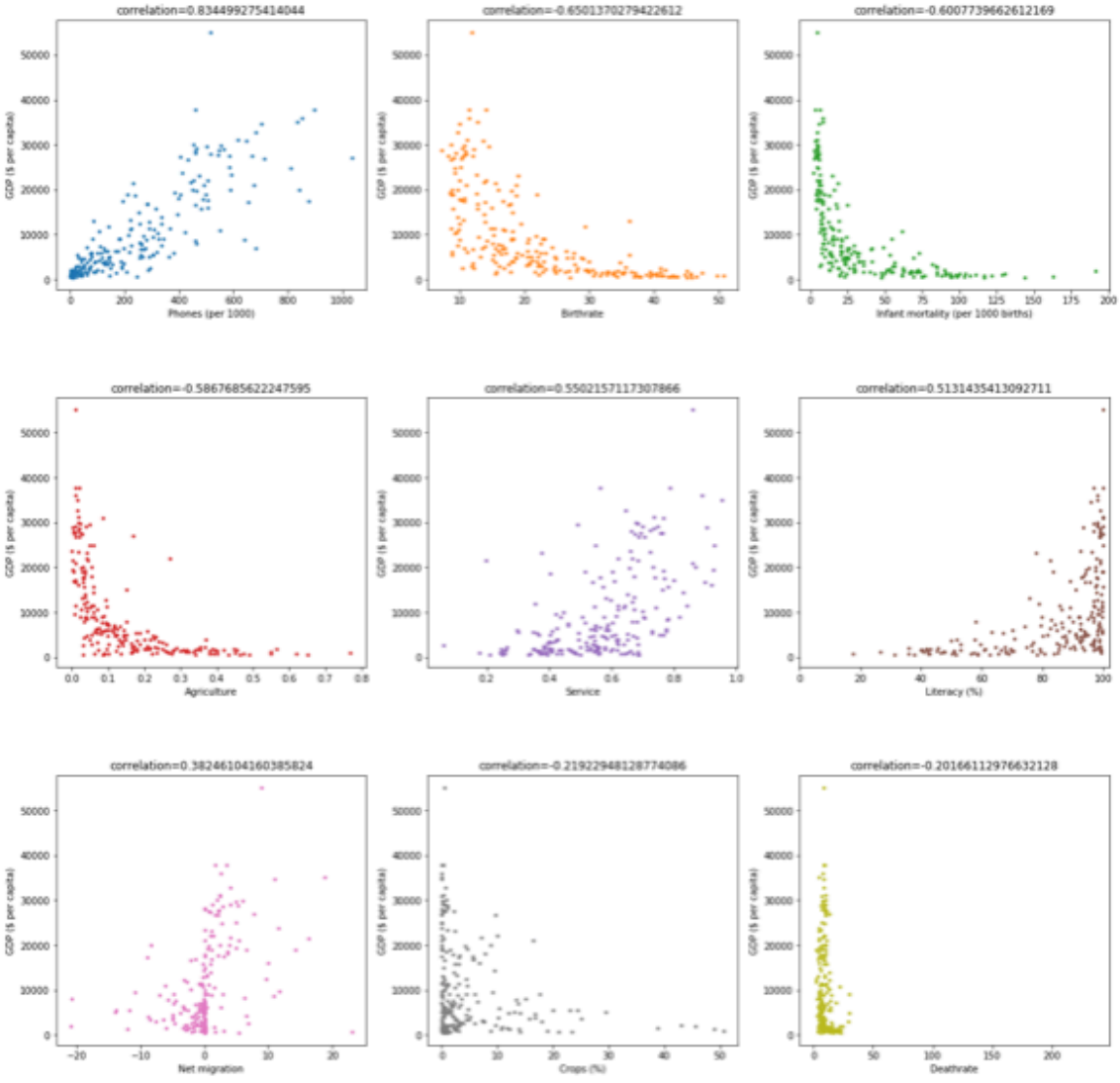
We calculate the correlation between each features. It is interested that GDP is relatively independent on infant mortality and birth rate, but is relatively dependent on phone rate. We can say that if a country's economic is in a good condition, people there may afford to buy cell phone.

[7]:

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP p capi
Population	1.000000	0.469985	-0.028190	-0.068199	0.000011	0.023003	-0.0393
Area (sq. mi.)	0.469985	1.000000	-0.066917	-0.095627	0.047470	-0.007163	0.0721
Pop. Density (per sq. mi.)	-0.028190	-0.066917	1.000000	0.241690	0.178038	-0.144238	0.1957
Coastline (coast/area ratio)	-0.068199	-0.095627	0.241690	1.000000	-0.134145	-0.136577	0.0491
Net migration	0.000011	0.047470	0.178038	-0.134145	1.000000	-0.025015	0.3824
Infant mortality (per 1000 births)	0.023003	-0.007163	-0.144238	-0.136577	-0.025015	1.000000	-0.6007
GDP ($ per capita)	-0.039324	0.072185	0.195773	0.049118	0.382461	-0.600774	1.0000
Literacy (%)	-0.043481	0.035994	0.097230	0.117095	-0.008761	-0.765536	0.5131
Phones (per 1000)	-0.030819	0.053403	0.280172	0.152129	0.239601	-0.672312	0.8344
Arable (%)	0.188251	-0.079714	-0.081015	-0.120772	-0.058170	-0.113859	0.0197
Crops (%)	-0.060757	-0.143916	-0.031661	0.341820	-0.347968	-0.062366	-0.2192
Other (%)	-0.120606	0.138974	0.081859	-0.079494	0.225299	0.124100	0.0978
Climate	-0.026546	-0.117134	0.009064	-0.032579	-0.015170	-0.374406	0.3407
Birthrate	-0.045090	-0.066423	-0.159192	-0.075765	-0.066537	0.845764	-0.6501
Deathrate	-0.028367	0.040079	-0.070045	-0.161866	0.033611	0.656338	-0.2016
Agriculture	0.000376	-0.050640	-0.107142	-0.030250	-0.116719	0.709475	-0.5867
Industry	0.099333	0.121565	-0.139803	-0.199995	-0.026324	0.003322	-0.0289
Service	-0.080852	-0.054369	0.248774	0.196528	0.127134	-0.633982	0.5502



Then, we plot the scatterplot for the GDP versus others parameters for all the countries and also indicate the correlation between GDP and the parameter. We can clear see that, for most of the countries, the phone rate is proportional to GDP. The net migration, crops, and death rate are relatively independent to GDP.



Conclusions

Contributions

Citations

License