

**TASK**

**Exploratory Data Analysis on the Countries of the world Data Set**

By Siphamandla Sithole

[](http://www.hyperiondev.com/portal/)

**Introduction**

The data set chosen is on the countries of the world which has been taken from Kaggle: <https://www.kaggle.com/fernandol/countries-of-the-world/data>. The data was originally compiled by the US Government and contains data from 1970 – 2017. The data contains information about 227 countries including population, net migration, mortality rates, GDP etc. There are 20 columns in the data set.

1. Description of data

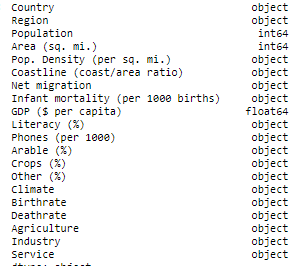
|  |  |  |
| --- | --- | --- |
| COLUMN | DESCRIPTION | DATA TYPE |
| Country | Name of a country | String |
| Region | Geographical region | String |
| Population | Number of people in the country | Integer |
| Area (sq. mi.) | The size of the area | Integer |
| Pop. Density (per sq. mi.) | How many people per square-mile. | Float |
| Coastline (coast/area ratio) | Ratio of coastline to country | Float |
| Net migration | Difference of people coming in vs out | Float |
| Infant mortality (per 1000 births) | Death of children for every 1000 births | Float |
| GDP ($ per capita) | Economic output per individual | Float |
| Literacy (%) | % of people able to read & write | Float |
| Phones (per 1000) | # of people with phones per 1000 | Float |
| Arable (%) | % of country that is arable | Float |
| Crops (%) | % of crops produced from land | Float |
| Other (%) | % of other produced from land | Float |
| Climate | No information | Integer |
| Birthrate | # of live births per thousand per year | Float |
| Death rate | # of deaths per thousand per year | Float |
| Agriculture | % contribution to GDP | Float |
| Industry | % contribution to GDP | Float |
| Service | % contribution to GDP | Float |
|  |  |  |

**DATA CLEANING**

1. Data types

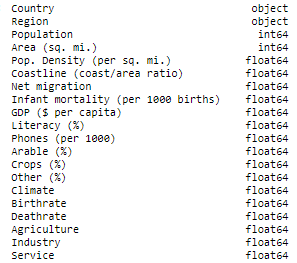
Let’s start by looking at whether we require each column to do the analysis. This will help us to work with data that is relevant and not waste time cleaning and analyzing data that will eventually be deleted. I am happy that all the columns are relevant, and I can continue cleaning the data.

The other factor we must do is check the data types of each column to ensure that the data is in the correct type.

We can immediately see that most of the data has been defaulted to object value even though it is either a float or integer. We need to change this so that we do not get any errors when trying to clean our data.

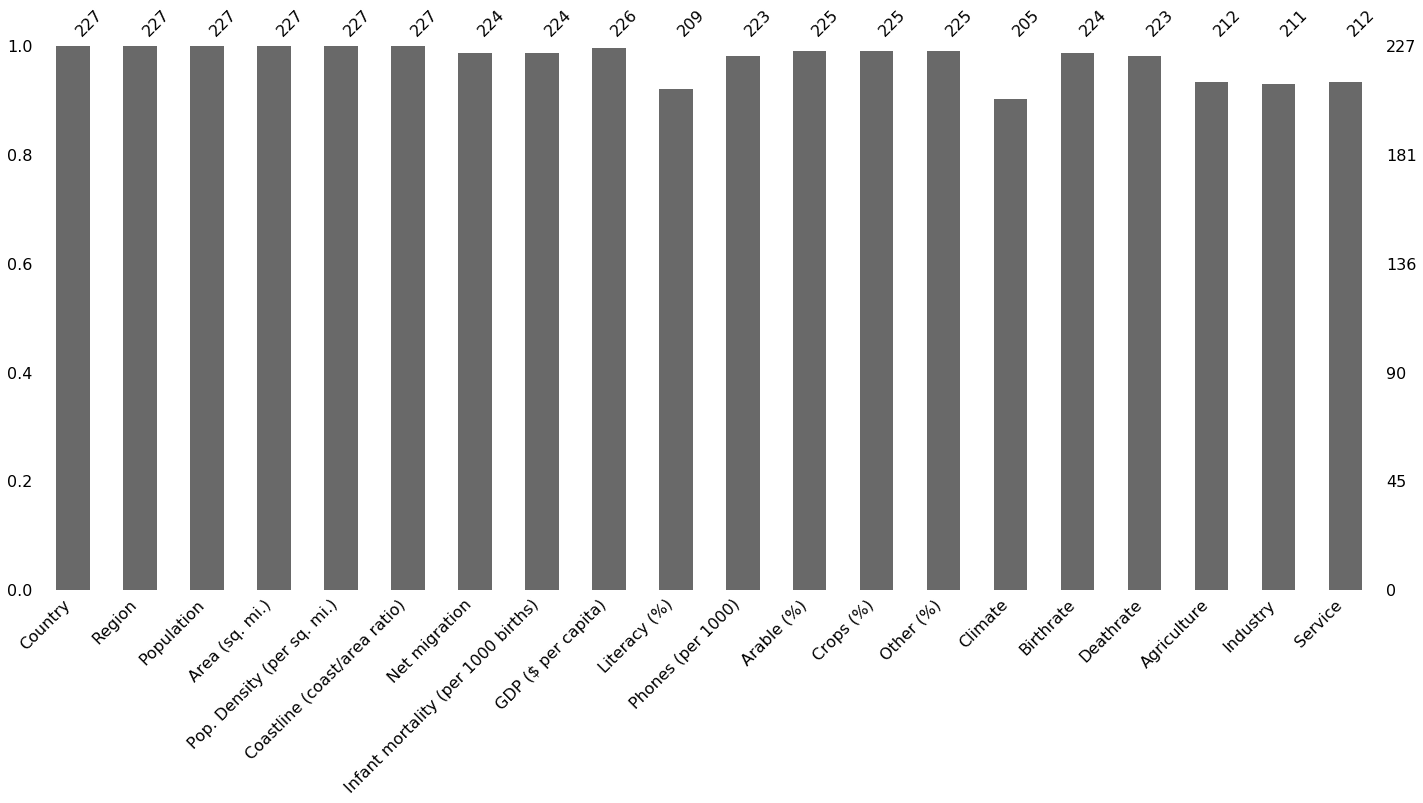
We have seen in the introduction of this document that the only strings in this data frame is ‘Country’ and ‘Region’ column.

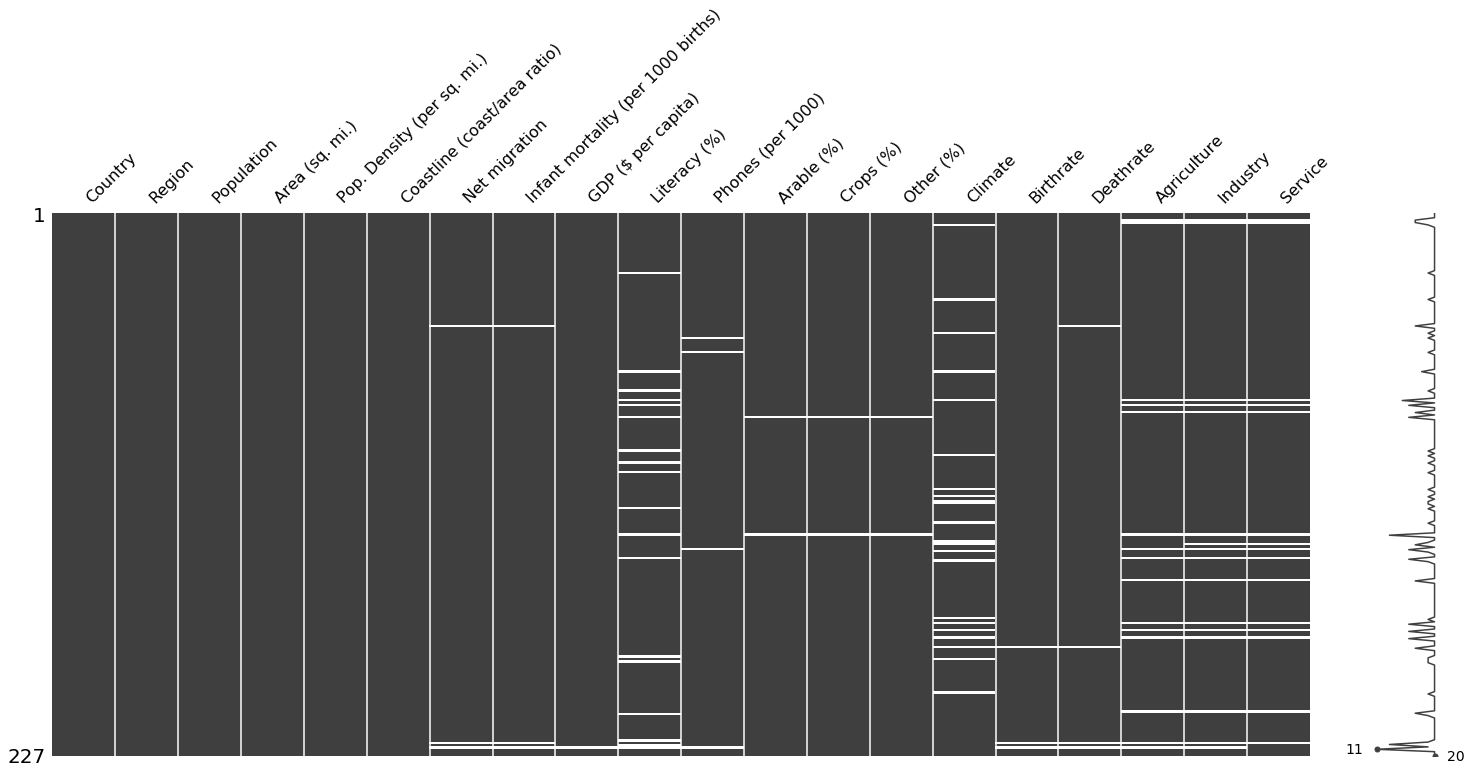
We need to ensure that the other columns are converted to the correct data types.



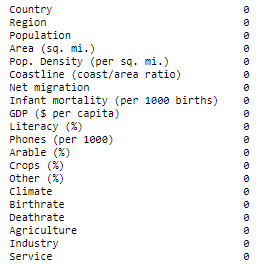
MISSING DATA

We will also look at the missing data and how many values are missing per column. The data in this file cannot be deleted as we will be deleting important data from the country and region column. We will need none conventional ways to either fill the data or represent it in a correct way.

****

****

We will group by ‘Region’ and take the mean of the grouped values to fill the missing information. Looking at the columns with missing values, it highly likely that the countries will have the similar data in the same region.



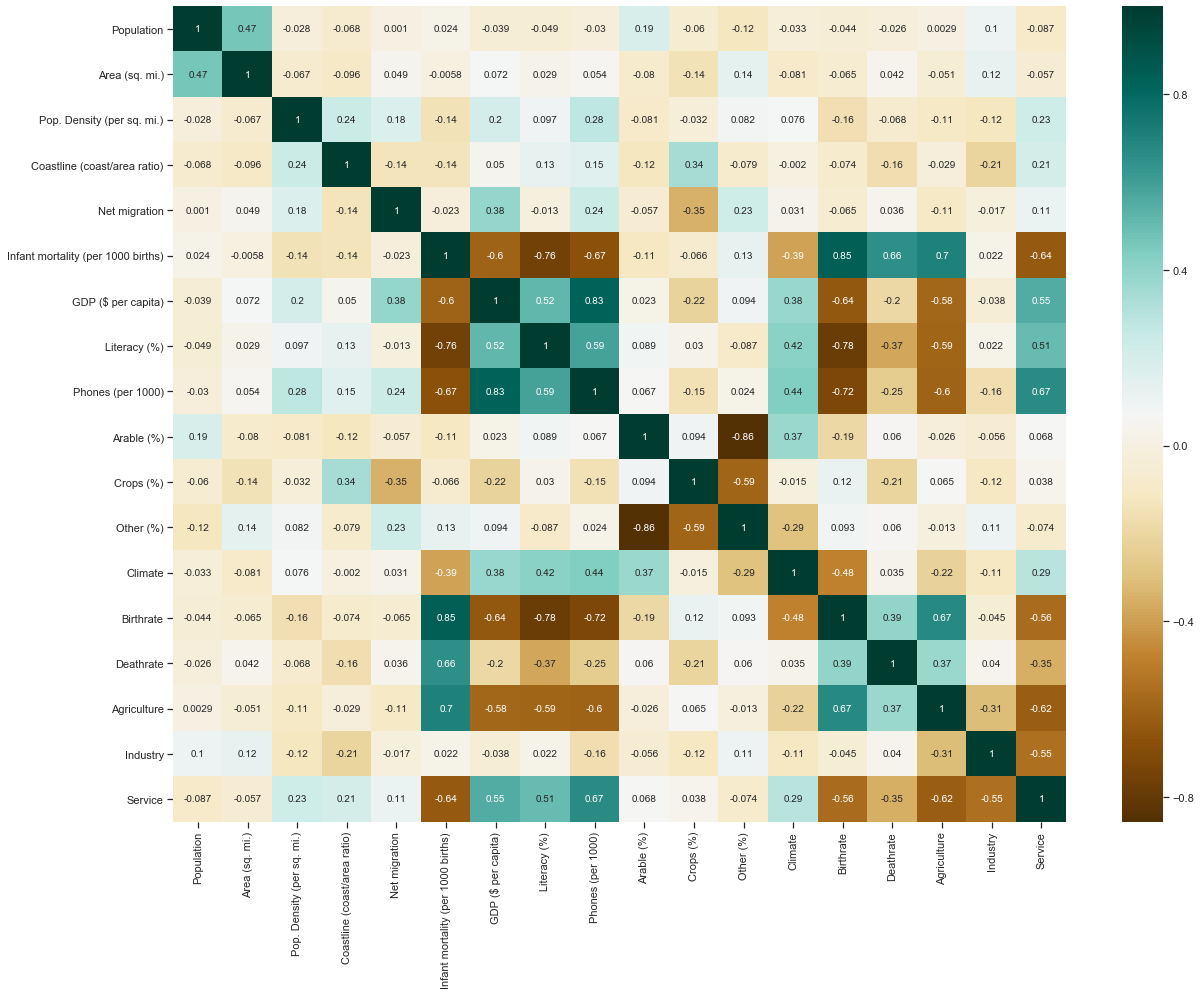
DATA STORIES AND VISUALIZATIONS

Now that we have cleaned the data, we can start analyzing the data and bring out some insights. We first look at the data and make a few hypotheses regarding the data and use visualization to prove and disprove the hypothesis.

a) Hypotheses

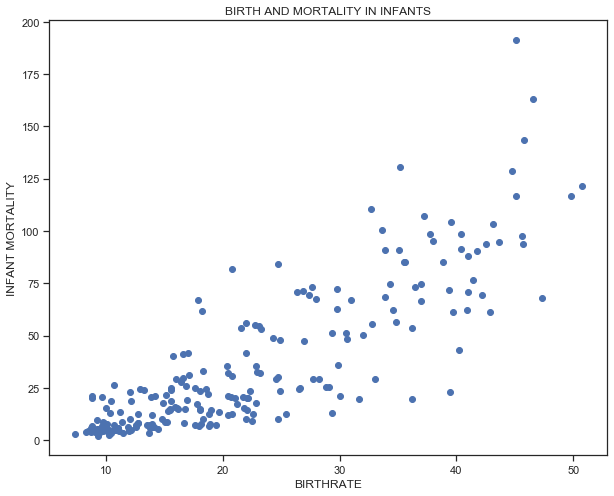
1. Western countries have the highest GDP per capita than any other region.
2. African countries have the lowest literacy rate in the whole world.
3. Asia has the highest literacy rate.
4. The middle eastern regions have the lowest arable lands.
5. The higher the literacy rate, the higher the GDP per capita.
6. The higher the literacy rate, the higher the service component as distribution to the economy of the country.
7. The climate will have a high relationship with how arable the land is.

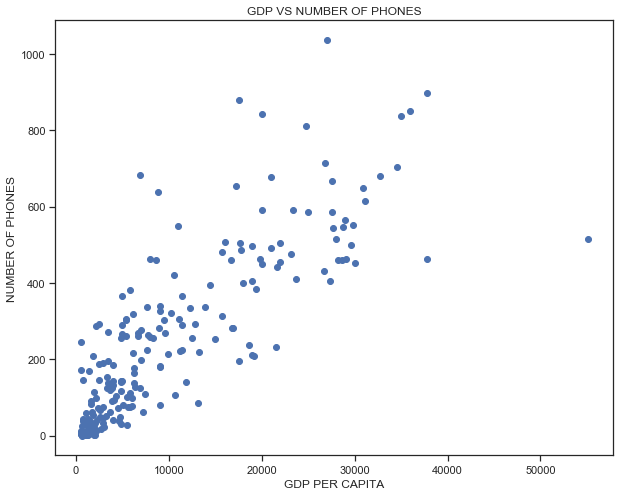
Firstly, we will explore the relationship between the data variables.

****

There are variables that seem to share strong correlations with each other. While we cannot discuss each individual variable here, we will mention a few that share over 80% correlation such as the” birthrate” and “infant mortality rate” & “GDP” and “Phones per 1000”. Let’s have a look at these variables in more detail.

It seems that where the birthrate is high, so is the mortality rate in infants. The data is not definitive if one is the cause of the other. By making that assumption, we could easily come to an incorrect conclusion that the more babies are born at a time, the more chances of increased mortality amongst infants. It maybe that poorer countries still value larger family sets even though access to healthcare may not be immediately available.

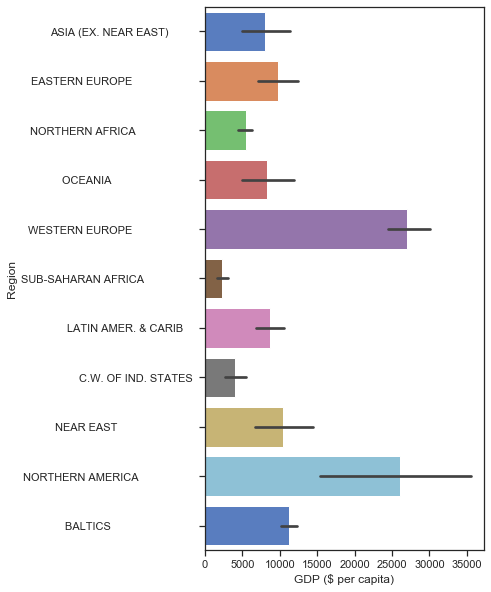




An interesting data correlation is the GDP per capita to phones per 1000. It is clearly evident that the higher the GDP per capita, there more people own phones. This makes logical sense as the more people earn, the higher the disposable income to purchase phones and other electronics.

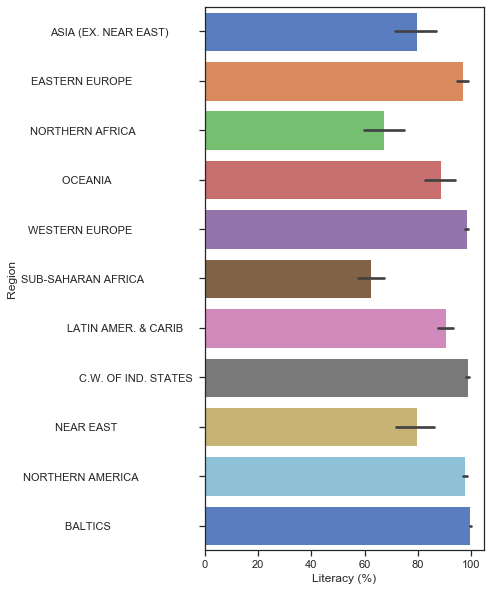
Hypothesis testing

### Western countries have the highest GDP per capita than any other region.

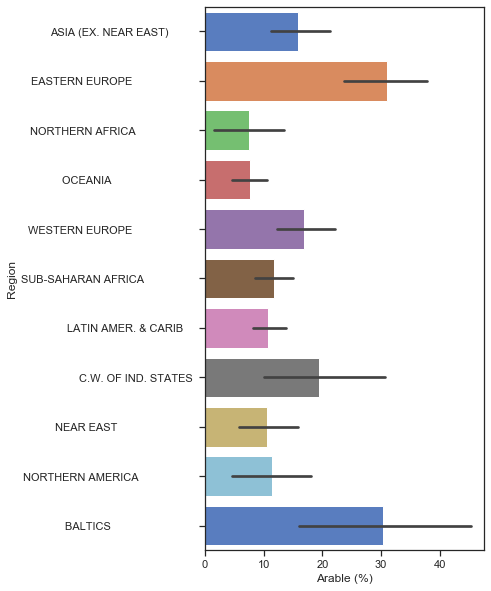


It goes without saying that the initial assumption is correct. Western countries which include North America have the highest GDP per capita in the whole world

#### **African countries have the lowest literacy rate in the whole world and Asia has the highest literacy rate**



While the initial assumption about the literacy rate in Africa is true, it is incorrect that Asians have the highest literacy rate. Western countries show the highest literacy rate in the world.



Here, our assumption is not very far off. Oceana and Northern Africa has the least Arable lands but the Near East follows at a close 3rd.

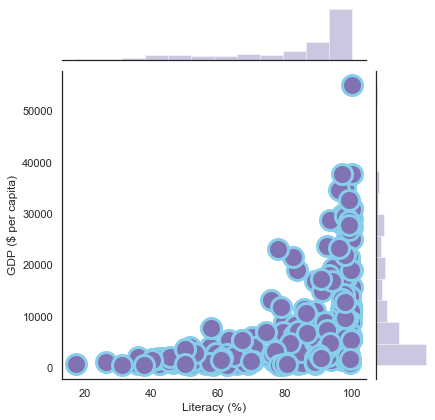
#### **The higher the literacy rate, the higher the GDP per capita**

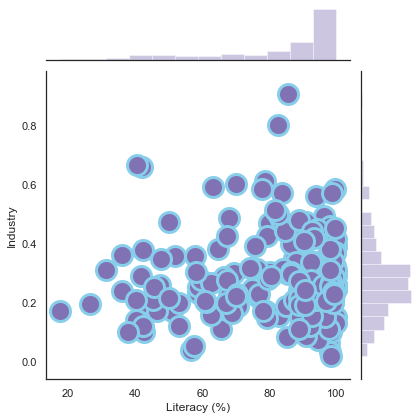
The graph shows that a person’s literacy does not determine the GDP per capita definitively.

We can see that there are instances where the literacy rate is high but the GDP per capita is very low.

The inverse isn’t true though as the only countries with high GDP per capita have high literacy rates.

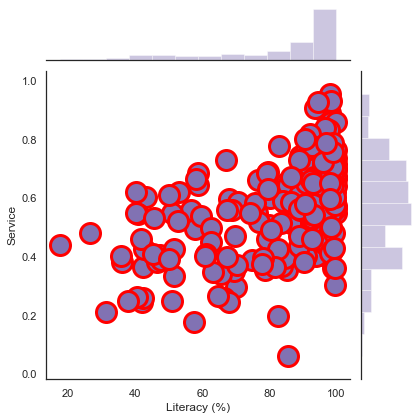
#### **The higher the literacy rate, the higher the service component as distribution to the economy of the country.**

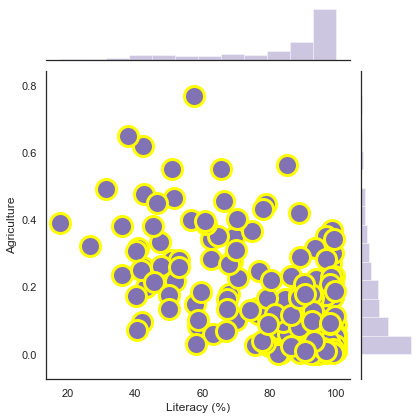


****

When we look at the industry as a contribution to the economy, we can see that it does not correlate or have any relationship with the Literacy rates. The variables are spread across the cartesian graph without a clear specific pattern.

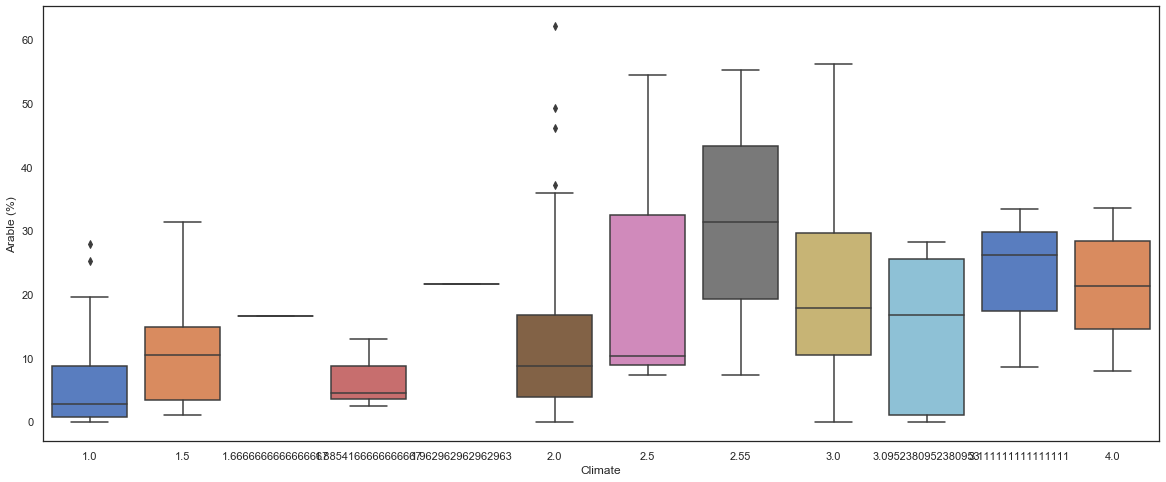
The data changes when we look at the Service component. It seems to react to Literacy rates. There seems to be a marginal shift towards the top as literacy rates increase, This is very clear when looking the far end of the literacy scale.

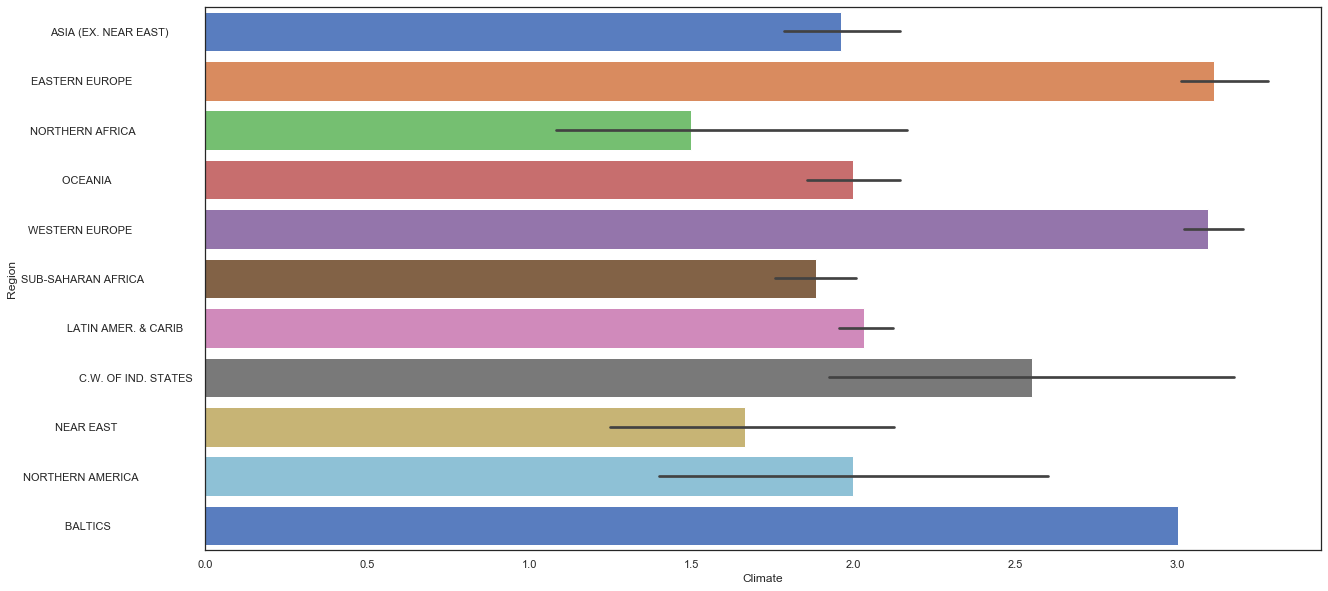




Agriculture reacts in an opposite direction to service. Agriculture declines in countries where Literacy rates are very high. This makes sense as more and more people look towards service and industry, the low skill and minimum wage jobs of the Agricultural sector will decline.

#### **The climate will have a high relationship with how arable the land is.**





From the data we can see that region that have a climate ‘code’ between 2.5 and 3.0 seem to enjoy greater portions of arable land. The regions enjoying this land is Western Europe, the Baltics and Indonesian states.

**THIS REPORT WAS WRITTEN BY : Siphamandla Sithole**

