Data Project on The Countries of The World

Information on population, region, area size, infant mortality and more from the World Factbook obtained via https://www.kaggle.com/fernandol/countries-of-the-world.

The data contained within consists of the different countries of the world separated into their regions and shows the following attributes per region:

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 227 entries, 0 to 226
Data columns (total 20 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Country                             227 non-null    object
 1   Region                              227 non-null    object
 2   Population                          227 non-null    int64
 3   Area (sq. mi.)                      227 non-null    int64
 4   Pop. Density (per sq. mi.)          227 non-null    object
 5   Coastline (coast/area ratio)        227 non-null    object
 6   Net migration                       224 non-null    object
 7   Infant mortality (per 1000 births)  224 non-null    object
 8   GDP ($ per capita)                  226 non-null    float64
 9   Literacy (%)                        209 non-null    object
 10  Phones (per 1000)                   223 non-null    object
 11  Arable (%)                          225 non-null    object
 12  Crops (%)                           225 non-null    object
 13  Other (%)                           225 non-null    object
 14  Climate                             205 non-null    object
 15  Birthrate                           224 non-null    object
 16  Deathrate                           223 non-null    object
 17  Agriculture                         212 non-null    object
 18  Industry                            211 non-null    object
 19  Service                             212 non-null    object
dtypes: float64(1), int64(2), object(17)
memory usage: 35.6+ KB
```

I chose this dataset because it contains numerical values and categorical values and has a usability rating of 8.2 on www.kaggle.com which means the data is generally reliable.

Initial Plan for Data Exploration

The first step is pre-processing and cleaning of the data. This data contains some null values and because it is relatively small, it was decided to fill the null values with their mean values so as to preserve data integrity:

```
In [30]: df = df.fillna(df.mean())
```

Actions taken for data cleaning and feature engineering:

The next step was to rename the columns to make the data easier to work with when using the python programming language as well as to change data types and remove trailing and leading spaces:

```
In [9]: df.columns = (["country","region","population","area","density","coastline","migration","infant_mortality","gdp","literacy","phon
```

```
In [10]: df.country = df.country.astype('category')
         df.region = df.region.astype('category')
```

remove trailing and leading spaces

```
In [11]: df = df.applymap(lambda x: x.strip() if isinstance(x, str) else x)
```

It is necessary to explore this data by region and so we look for the count of the values in this column and assign it as variable “region”:

```
In [12]: df['region'].value_counts()
```

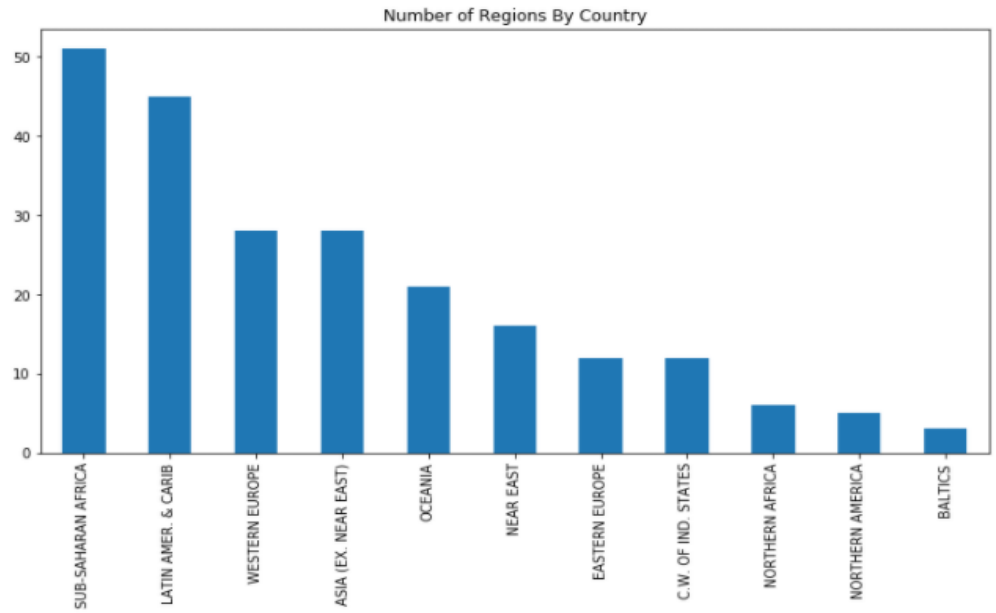
```
Out[12]: SUB-SAHARAN AFRICA      51
         LATIN AMER. & CARIB     45
         WESTERN EUROPE          28
         ASIA (EX. NEAR EAST)    28
         OCEANIA                 21
         NEAR EAST               16
         EASTERN EUROPE          12
         C.W. OF IND. STATES     12
         NORTHERN AFRICA          6
         NORTHERN AMERICA         5
         BALTICS                  3
         Name: region, dtype: int64
```

group country and number of regions

```
In [13]: region = df.region.value_counts()
```

Next, we plot this using matplotlib:

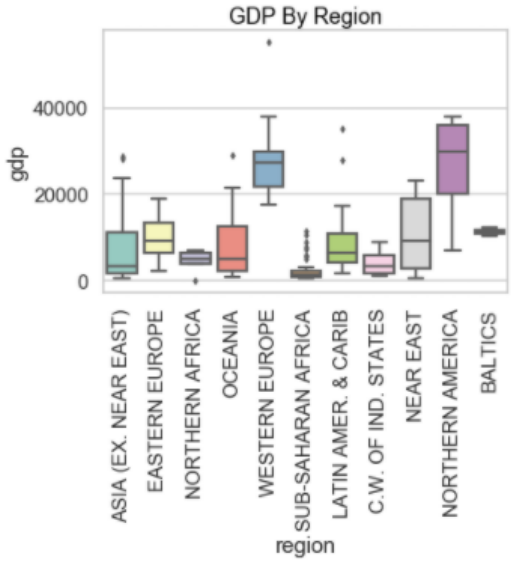
```
In [14]: ax = region.plot(kind='bar',figsize = (12,6))
         plt.title('Number of Regions By Country')

         ax.tick_params(axis='both', which='major', labelsize=10)
         ax.tick_params(axis='both', which='minor', labelsize=20)
```



From this data we can see that the regions are more numerous in Sub-Saharan Africa and Latin America. Does this mean that GDP is related to number of regions? Is bigger always better? Let us examine with the Seaborn python library:

```
In [33]: sns.boxplot(x="region",y="gdp",data=df,width=0.8,palette="Set3",fliersize=4)
         plt.xticks(rotation=90)
         plt.title("GDP By Region",color="black")
```

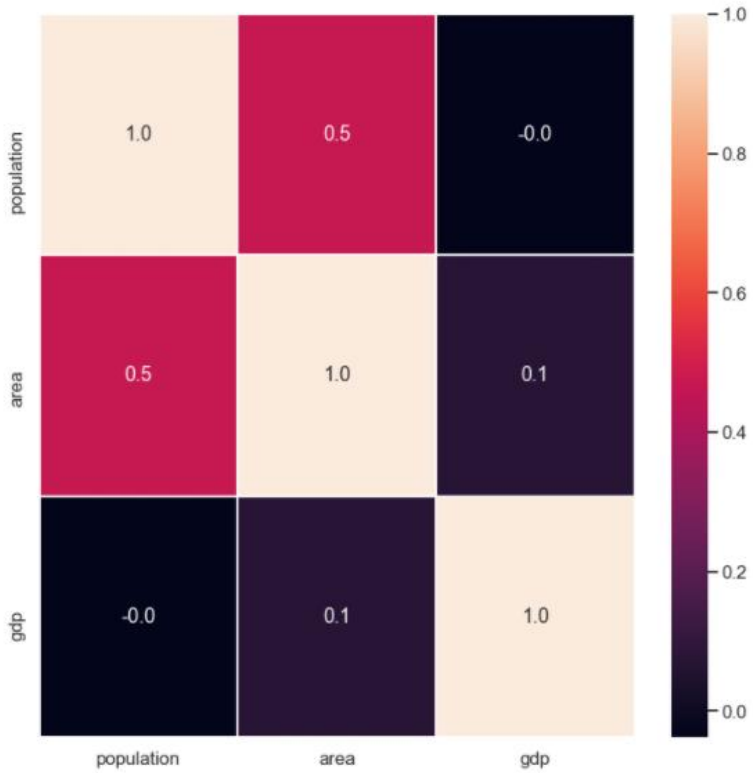
```
Out[33]: Text(0.5, 1.0, 'GDP By Region')
```



It appears that GDP is highest in Northern America and Western Europe. Correlations need to be explored for us to select the correct features; we can do this with Seaborn heatmaps:

```
In [37]: sns.set_style("dark")
         corr = df.corr()
         f,ax = plt.subplots(figsize=(12, 12))
         sns.heatmap(df.corr(), annot=True, linewidths=.8, fmt= '.1f',ax=ax)
```

```
Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x28377d08460>
```



Key finding and Insights

It is clear from the above findings that GDP has some sort of correlation with area and this can be demonstrated with the below graph using a pair plot from Seaborn:

```
In [42]: sns.set_style("dark")
         sns.pairplot(df, hue='region',height = 5);
```

Thus, we decide to choose area as our target:

```
In [43]: #Separate our features from our target
         X = df.loc[:,['gdp','population','region']]
         y = df['area']

In [44]: mean = df['area'].mean()
```

Hypothesis Testing

Null Hypothesis: There is no relationship between area and GDP.

Alternative Hypothesis: There is a relationship between area and GDP.

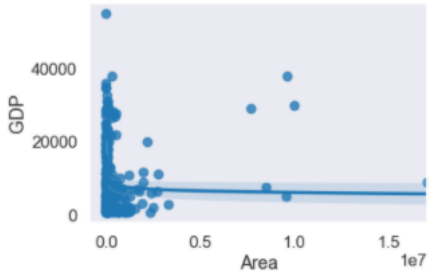
Third Hypothesis: There is a relationship between another variable and GDP.

Formulating a significance test for the Null Hypothesis

The Pearson Correlation test is used to analyse the strength of a relationship between two provided variables, both quantitative in nature. The value, or strength of the Pearson correlation, will be between +1 and -1.

A correlation of 1 indicates a perfect association between the variables, and the correlation is either positive or negative. The data needs to be transformed into logarithmical data using logx=True in the Seaborn regplot:

```
In [106]: def plt_regression(x, y, data, label_1, label_2):

              reg_plot = regplot(x=x, y=y, fit_reg=True, data=data,logx=True)
              plt.xlabel(label_1)
              plt.ylabel(label_2)
              plt.show()

          plt_regression('area', 'gdp', df, 'Area', 'GDP')
          print(pearsonr(df['area'], df['gdp']))
```




```
(0.07217948209438932, 0.2788514652326703)

In [ ]:
```

The first value is the direction and strength of the correlation, while the second is the P-value.

The data shows that our null hypothesis should be rejected as we attain a p-value of ~0.27 which indicates a correlation does indeed exist between area and GDP as shown by the guidelines for the Pearson Correlation below:

Below are the proposed guidelines for the Pearson coefficient correlation interpretation:

	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

Suggestions for next steps in analysing this data:

This dataset could be merged with an additional dataset of similar or the same features to create a larger dataset as the sample size for this dataset is relatively small. Increasing the sample size allows for greater insights and stronger relationships to be observed in data.

Summary of Data Quality and Request for Additional Data:

The data set did not require much cleaning, there were very few null values and columns were well labelled. The data itself used an acceptable file format as it was a comma separated values file (.csv) and contained both categorical and numerical data as well as a well known and respect source, the Central Intelligence Agency (CIA) in the United States of America.

A request for additional data could be made from various organisations such as the World Bank or even from the CIA itself via https://www.cia.gov/library/publications/the-world-factbook/docs/history.html