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:

A picture containing graphical user interface

Description automatically generated

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

**Initial Plan for Data Exploration**

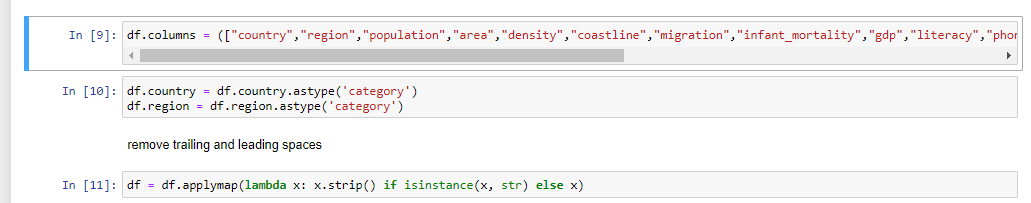
The first step is pre-processing and cleaning of the data. This data contains some null values and because it is relatively small, it was decided to fill the null values with their mean values so as to preserve data integrity:

Chart

Description automatically generated

**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:



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”:

Graphical user interface, application, Word

Description automatically generated

Chart

Description automatically generated Next, we plot this using matplotlib:

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:

Chart, box and whisker chart

Description automatically generated

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:

Chart

Description automatically generated

**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:

A picture containing scatter chart

Description automatically generated

Thus, we decide to choose area as our target:

Graphical user interface, application

Description automatically generated

**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](https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pearsonr.html) 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:

A picture containing graphical user interface

Description automatically generated

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:

Table

Description automatically generated

**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