Global GDP and socio-economic factors

Dataset

The dataset summarizes 214 countries’ region code, country code, GDP, population, Population compound growth rate (CGR), internet users (per 100), percent of population in the largest city, 2014 life expectancy from birth, female literacy rate and exportation rate. Outside of country classifications, there were subsets that included various countries or similar cultures based on the categories mentioned above.

<Snapshot of raw data set>

Objective

From this dataset, the objective was to see whether there was a correlation between GDP and internet penetration, life expectancy and population size. The analysis factors all reported countries, however, the information was analyzed from a regional standpoint.

Target Audience

The target audience for this analysis would be students, researchers, development initiatives, and institutions interested in global development or socio-economic trends. This analysis would prove useful in talks about global development and the role of socio-economic factors in spurring GDP and overall economic growth.

Limitations

The limitations are:

* The dataset provides a snapshot of the countries in 2015 rather than an analysis over several years. Had the analysis covered a wider scope of years there would be greater potential to identify trends or better identify the impact of each factor impact over time.
* Whilst the analysis details the impact of internet penetration, life expectancy and population size, other socio-economic factors play impactful roles in the development of an economy i.e. political stability, infrastructure, education and investment, overall technological advancement, among others.
* Missing values resulted in some columns being deleted. The most notable columns that would have been important to also explore for their effects on GDP were women’s literacy and exportation rates. These, among other, missing values restricted a more fulsome.analysis (58 countries did not provide women’s literacy rates and 69 countries did not provide information on exportation rates).

Data Cleaning

Data cleaning focuses on formatting and ensuring data consistency to properly represent the various regions and their GDP. This involved:

* File identification

Identifying the file type (chardet) that would best help us to read the source csv file i.e. “latin-1”

* Deletion

Removing the additional subsets that would inflate the information i.e. remove data without “Region Codes” e.g. Arab World, middle income, etc.

* Renaming

Renaming and mapping columns to data to have clearer categories

Renaming “NA” region code to “North America” as the region code as it was being read as NaN

* Formatting

Identifying and converting the columns that did not have the correct data types i.e. object to float64

Resetting the index given the deletion of certain rows

Removing characters such as commas, spaces, percent symbol, etc.

Adding commas and decimal to float, in addition to, removing scientific notation for numbers

Analysis

Finding mean, median, variance, standard deviation, sems of GDP, internet penetration, life expectancy, population size.

<Provide table> X 4

Findings

GDP vs Internet

<Stats>

<visualization>

<Explanation>

GDP vs Life Expectancy

<Stats>

<visualization>

<Explanation>

GDP vs Population size

<Stats>

<visualization>

<Explanation>

Conclusion