ETL Project

Is there any relation between cancer death rate and GDP for all the countries around the world?

Prepared by Joyce Muthondu and Behnam Firoozfard

Summary:

We were interested in cancer mortality rate among different nations .Cancer is among the leading causes of death worldwide. In 2012, there were 14.1 million new cases and 8.2 million cancer-related deaths worldwide.57% of new cancer cases in 2012 occurred in less developed regions of the world that include Central America and parts of Africa and Asia; 65% of cancer deaths also occurred in these regions.The number of new cancer cases per year is expected to rise to 23.6 million by 2030.

The Gross Domestic Product measures the value of economic activity within a country. Strictly defined, GDP is the sum of the market values, or prices, of all final goods and services produced in an economy during a period of time

In this project we tried to make a Dataset to see whether there is a correlation between cancer rate deaths and the GDP. In less developed countries or regions rather than the number of deaths per year was very huge which made us come with the following questions:

- Is there any relation between cancer death rate and GDP for all the countries around the world?
- Which type of cancer is more common in a certain country?
- How has the GDP rate being changing over the past 16 years from the year 2000 to 2016 per country?

Extract

We extracted our data from the following sources:

- *GDP (Gross Domestic Product) for the whole world from 1960:*

https://data.worldbank.org/indicator/ny.gdp.mktp.cd?view=map

- The Mortality rate in the world based on all kinds of diseases.

Cause-Specific Mortality, 2000–2016: In which we extracted two datasets for both male and female sexes.

https://www.who.int/healthinfo/global_burden_disease/estimates/en/

1:ghe2016_deaths_country_fmle---330 MB, include 91500 Records, 23 columns
2:ghe2016_deaths_country_male---330 MB, include 91500Records,23 columns
The dataset has columns for sex, age group, Cause name and years .

Transform

We transformed our data from a CSV files to Pandas Dataframes using jupyter notebook and did data cleaning of the extracted data.

The steps we followed to transform the data were as follows:

- ghe2016_deaths_country_fmle
  - Dropping unnecessary columns
  - Checking if there are Null values
  - Filtering the column with causename just being cancer
- Ghe2016_deaths_country_male
  - Dropping unnecessary columns
  - Checking if there are Null values
  - Filtering the column with causename just being cancer
- Combined both Dataframes for sex
  - Using Concat, since both data had same columns
- GDP data:
  - We selected the column we needed from the years 2000 to 2016
  - Filled the missing values using interpolation technique and using current data to estimate null values.
- Put all data in one Dataframe
  - Renamed one column in both cancer cases(i.e female and male) with the same name in gdp 'country code'
  - Merged two main Dataframes to get one master DataFrame

One more step that can be helpful for the analysis :

As different years appeared in 16 columns, We tried to make all of them as one column with the year name .We used a for loop and concat in this section.

With this step our data frame has “year”, “age” and “Country” columns to groupby and do some epaltory data analysis and also see the relation between gdp and cancer .

Load.

We exported the pandas dataframe in different formats for future use .

- Saved data as a CSV file
- Saved data as a Json file
- Saved data as Postgres table in ETL_project dataset .We can have other tables in postgres like Age, country and cause to normalize the data if needed.