IST 652- Scripting for Data Analysis

Final Project Report

Climate Change – Identifying the major cause for the same

Team Members: Harshit Joshi and Yashaswini Kulkarni

Data and its source:

The data sets have been gathered from the Kaggle compilation of data sets for machine learning and data science projects. For the initial analysis of the Earth’s temperature rise, we are making use of the following data sets:

This data has been collected from [Berkeley Earth](http://berkeleyearth.org/about/), which is affiliated with Lawrence Berkeley National Laboratory. The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives. It is nicely packaged and allows for slicing into interesting subsets (for example by country). It includes the following csv files which have been actively used in the jupyter notebook:

* GlobalTemperature.csv – Gives the global land temperature including the average, minimum and maximum starting from the year 1750.
* GlobalLandTemperatureByCity.csv - This gives the global land temperatures grouped by all cities.
* Historical\_emissions.csv - Historical emissions of carbon dioxide for close to 3 decades for all countries across the globe.
* World\_population.csv – This csv contains world population in terms of countries. This is used to determine per capita CO2 emissions for every country.

The most important csv used throughout the analysis purpose is the GlobalTemperature.csv, which has the following attributes:

* Date: starts in 1750 for average land temperature and 1850 for max and min land temperatures and global ocean and land temperatures
* LandAverageTemperature: global average land temperature in Celsius
* LandAverageTemperatureUncertainty: the 95% confidence interval around the average
* LandMaxTemperature: global average maximum land temperature in Celsius
* LandMaxTemperatureUncertainty: the 95% confidence interval around the maximum land temperature
* LandMinTemperature: global average minimum land temperature in Celsius
* LandMinTemperatureUncertainty: the 95% confidence interval around the minimum land temperature
* LandAndOceanAverageTemperature: global average land and ocean temperature in Celsius
* LandAndOceanAverageTemperatureUncertainty: the 95% confidence interval around the global average land and ocean temperature

Further to identify and to narrow down the reason for the rise in Earth’s temperature as the emission of Carbon dioxide, we use make use of the Historical\_emissions.csv data set. The data is gathered from the Climate Watch Data portal using CAIT data source. The data is collected for all the countries around the world, covering a span of close to 3 decades(1990-2018).

Additionally, to answer a few questions during the analysis, we are making use of the world\_population.csv data set which contains the population grouped by each country. This is used to merge the emissions dataset on the country column to answer one of the questions.

Pre – processing of Data:

Data cleaning is one of the most important steps before the analysis is started. In this particular project, we are employing the following data munging steps:

* NA values in the dataset causes hinderance in any kind of computation during the analysis of data. Hence, they have to be fixed as a pre-requisite. Here, we are using the dropna() method from the pandas library to remove the rows with NA values.
* Further, it is observed that the Date column in one of csv files used for the data set is not of the type ‘dateTime’. Hence, this attribute is converted into the DateTime type so that it can be used in its correct format.
* Once the column has been casted, we add four additional columns in the dataset using the newly converted Date column. We add ‘day’,’week’,’month’,’year’ columns in the dataset.
* Based on the initial analysis, we find that 1975 is the timeframe around which there is spike in the Earth’s temperature. Hence, we create a new column called ‘turnpoint’ in the dataset which will indicate if the year is either ‘before’ or ‘after’ 1975 so that the year is easily identifiable for analysis.
* During the analysis, we also remove the unnecessary columns from the dataset so that it is convenient for the processing of data.
* For answering a few questions, we have also merged the datasets using inner join on the common columns like ‘country’ column.

Data Analysis:

Post the data pre-processing, the data is analyzed to answer the following questions:

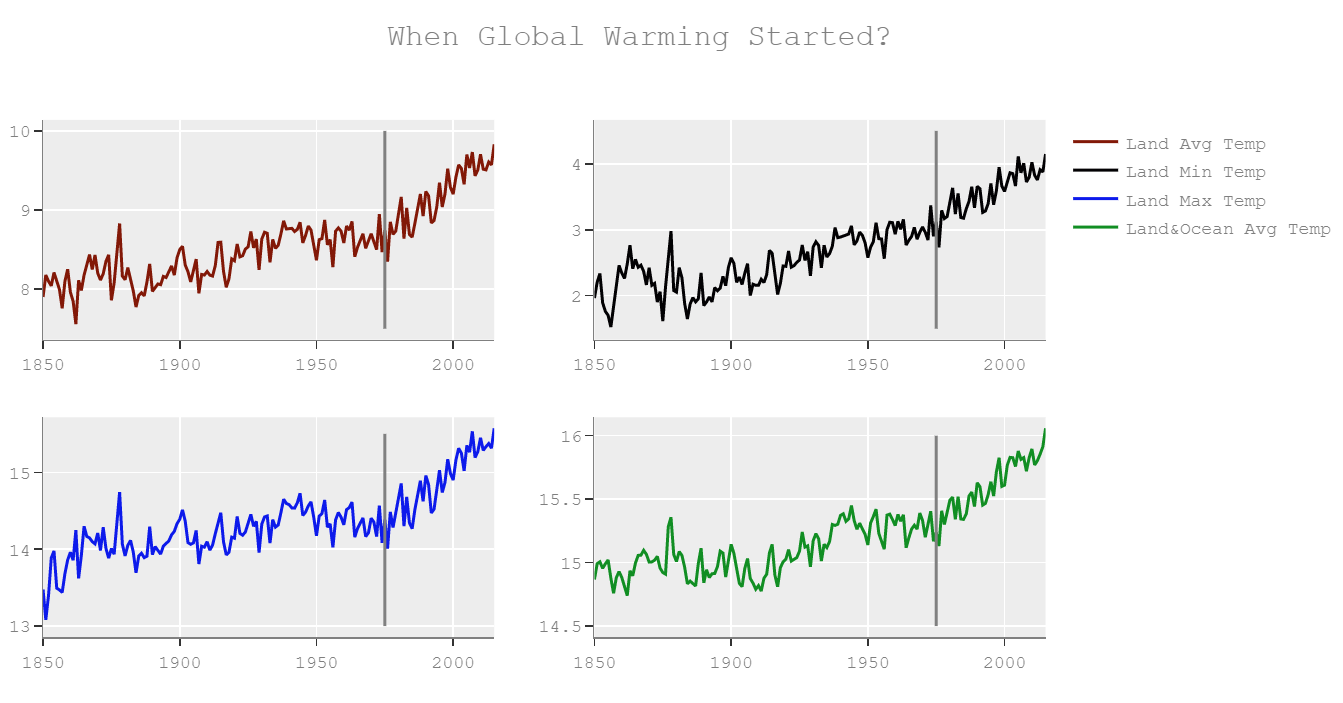
* At what point in the world history, did global warming start increasing?
* Is CO2 emission one of the major contributing factor for the increase?
* Which country has the highest per capita CO2 emissions?
* Which countries are among the highest contributing countries towards the global warming?

Initially, we start the data analysis by comparing the temperature records and error in recording them. This is done by using the LandAvgTemperatureUncertainity column. This helps us understanding that because of a few constraints in the initial years, there are a few errors in the temperature recording system. With the years progressing, we see that the errors and anomalies keep decreasing.

Chart

Description automatically generated

With the use of scatter plots and further plots using the ggplot, we identify that the Earth’s temperature has seen a sudden spike in 1975.



With the help of few more plots for the carbon dioxide data set we see that there is an increase in the carbon dioxide emission concentrations in the environment from the year 1975 itself. The major contributing factor for this is the industrial revolution that took place around the same time. Industrial revolution gave rise to building of new factories, companies and industrialization in full swing. This in-turn encouraged burning of fuels which increased the amount of carbon dioxide in the Earth’s environment. This is easily seen in the plot made in the script.

Chart, line chart

Description automatically generated

For the 3rd and 4th question, we make use of the emissions.csv and world\_population.csv data sets to narrow down the answers. Firstly, we start by removing all the unnecessary columns from the dataset so that it is convenient to compute the required measure. Post this step, we find the mean value of the emissions from the dataset and group that based on the country column. We then form a subset of the dataframe using only the mean emissions and country column. This is done to merge it with the world population dataset to find the emission per capita for each country. Further from the world population data, we again pick up only the population column and the country column. Then both the subset dataframes are merged based on the country column. The emission per capita is then calculated as mean\_emission/mean\_population for each country. Based on this value, it is concluded that Qatar has the highest carbon dioxide emission per capita.

For the last question, by employing plotly we plot a graph showing the emissions for each country in a particular year. In this graph, we can change the year which is being considered for comparison, and the trend for the countries vs the emission amount is observed.

Graphical user interface, application

Description automatically generated

Description of the Python Program:

We start by importing the required Python libraries for the program like, pandas, seaborn, matplotlib, datetime, plotly, pycountry and geopandas. We further do data preprocessing by the above mentioned methods. Post that,we perform data analysis with the help of plots using seaborn and plotly. Throughout the data analysis, we answer the proposed questions.

Output and Conclusions:

* During our analysis, we discover that the world temperature is increasing at alarming rates in the past 30 years.
* We identify that in 1975, there was a sharp increase in the global temperature. This suggests that this was the point of sharp jump in the impacts of global warming.
* Since the time frame of industrialization which was between the late 19th century and the beginning of the 20th century coincides with our previous conclusion, it is concluded that, industrialization might have had a negative impact on the global temperature and hence intensified global warming.
* Through our analysis, we conclude that ‘Qatar’ has the highest per capita CO2 emissions.
* Additionally, from the graphs, we observe that usually irrespective of the year, United states has always been the highest contributor to the carbon dioxide emissions.