CSE 545: Big Data Analytics
Climate Action

Team Members:
Heena Agarwal
Kajal Dalvi
Pulkit Kalia
Shruti Singh

Introduction:

"Climate change is real. It is happening right now; it is the most urgent threat facing our entire species and we need to work collectively together and stop procrastinating." [1] - Leonardo Di Caprio.
This is one of the many voices that have been raised against the changing climate and its impact. We have seen Amazon and Australian wildfires, anomalies in weather pattern and global warming because of human activities and ignorance. Climate change phenomenon is slow (may take hundreds of years) but its effects are seen in the longer run. Thus, it is important for mankind to study the factors which leads to the climate change and access the situations we would face if no appropriate measures are taken.

In this project, firstly, we aim to analyze the correlation between human caused factors like CO_2 emission, population, energy consumption etc. that might affect climate attributes like temperature and precipitation. Secondly, we predicted temperature and precipitation till the year 2050 to see what changes we can expect. Lastly, we conducted country wise analysis for the factors affecting the climate. We have focused on local and global level to get better insights and analysis. Thus, our project helps to understand the gravity of climate change and to figure out the factors that can help in reducing its effects.

Background:

Several research papers and similar work conducted in the past were taken as a reference to understand the topic and the approaches taken. In one of the studies conducted by NCAR [2], the rate of change of temperature given a time-series was studied and student's t-test was considered to check statistical significance. Research titled "Statistical significance of seasonal warming/cooling trends" [3] showed that temperature data cannot be treated as independent and characterized by short-term memory only and combined Monte Carlo simulations with the Holm–Bonferroni method to demonstrate how to obtain reliable estimates of the statistical significance.

Another study focused on "The Long-Run Effects of Climate Change on Conflict" [4] examines the long-run effects of climate change using fixed effects regression and why it is helpful to use the same which removes the time-invariant independent variables replacing them with the unique value α_i(fixed effect) which is the unique value for each individual entity in the data panel.

Data:

We have used 4 datasets, our primary dataset consisted of daily sensor readings of temperature, precipitation, wind speed, snow depth along with latitude and longitude. Since more than 50% of the data for snow depth and wind speed was missing, it was dropped in further analysis. Our secondary data had 3 datasets each containing attributes like country code, country name, year, and an indicator (CO_2, energy, population). Overall, there was 8% of the data missing from our secondary data. Further characteristics of the datasets have been tabulated in Table 1.

For one of our goals (time series analysis), our primary dataset for years 1929-2020 was used while for the other tasks we had to take intersection of the years of data from secondary and primary datasets and hence used 55 years of data. Missing values in secondary datasets was handled using KNNImputer library from sklearn with parameter k (neighbors) set to 2. KNNImputer works on the principle of taking the values from its closest neighbors.

Dataset	**No. of Features**	**Years of Data Available**	**% of missing values**	**Size**
Global climate summary of the day[5]	28	1929-2020	2%	30GB
Human activities – CO_2 emission[6]	5	1960-2014	3%	505 KB
Human activities – energy use[6]	5	1960-2014	2%	480 KB
Human activities – population[6]	5	1960-2014	3%	302KB

Table 1: Summary of datasets

Methods:

Data Pipelines: We have used Spark for efficient and parallel computations as well as data cleaning and manipulations. While TensorFlow[10] was used for performing fixed effects linear regression[7].
Algorithms: We did hypothesis testing to validate the correlation between different parameters pertaining to climate and human actions whereas time-series analysis was used to predict temperature and precipitation till year 2050 using SARIMA model [9].

- Data merging and cleaning:
  Climate dataset has latitude and longitude attributes while secondary dataset has country code. We have used ReverseGeocoder (Google API) [12] and pycountry library to find alpha3 country code and then merged both the datasets. All precipitation values were not on the same level, so we updated them to a common scale. For one of the tasks we had to aggregate data on a monthly basis. For other 2 tasks we aggregated the data on per year basis so as to merge it with secondary dataset.

- Time Series Analysis:
  In this task we aimed to predict temperature and precipitation over the next 30 years. SARIMA model was preferred since it handles seasonal nature of the data well. Since SARIMA model has many parameters upon which the model depends, it was necessary to tune its parameters to get optimal model. We used grid search approach for the same. Our dataset was split into train data from years 1930 to 2018 and test data from 2019 onwards. To check the accuracy of the model, we have used RMSE as a cost function.

- Country wise analysis:
  The goal for this task is to check the impact of CO_2 emission on temperature controlling for population and energy per country. We aggregated the average CO_2 emission for all the years available and filtered top 10 and bottom 10 CO_2 emitting countries. For these countries, we calculated beta values from multivariate linear regression and analyzed them to study the impact of CO_2 on temperature locally rather on global scale.

- Hypothesis Testing:
  Our aim in this task was to find the correlation of human related activities like CO_2 emission, energy use and population on climate attributes like temperature, precipitation. We defined null hypothesis (H_0) as 'the above human factors have no significant effect on climate'. For this, fixed effects linear regression was used to add control for geographical attributes and also handle time dependent variables.

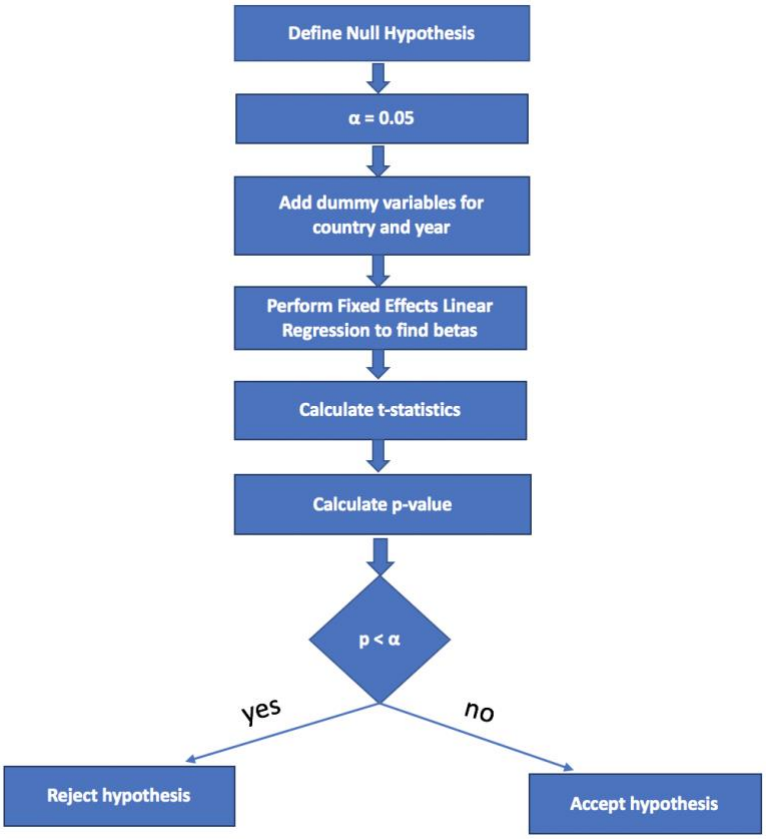

Figure 1: Flowchart for hypothesis testing

Evaluation/Results:

- Time Series Analysis:
  When evaluating for accuracy for our temperature and precipitation model, we were able to achieve RMSE of 1.19 and 0.48 respectively, on the test data. This indicates that the model was able to generate predictions close to actual values, as shown in figure 2(1 unit on y-axis=2.5° Celsius) and figure 3(1 unit on y-axis=0.04 inches). Also, using the same models we saw increase

of 1.9° Celsius and decrease in precipitation level by 0.08 inches, in the time period of 2020-2050, as shown in figure 4 and figure 5. These results align with previous findings [8].

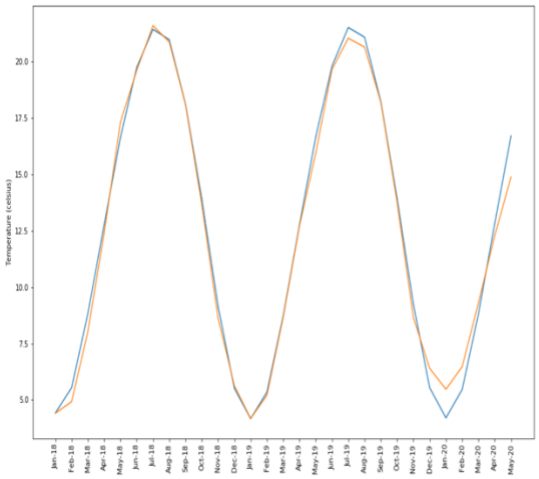

Figure 2: Temperature model test

Figure 3: Precipitation model test

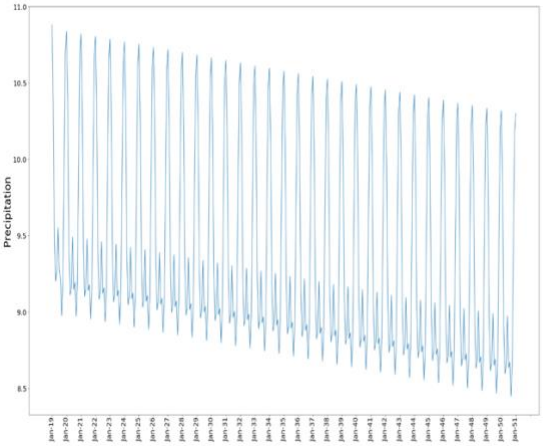

Figure 4: Temperature model predictions

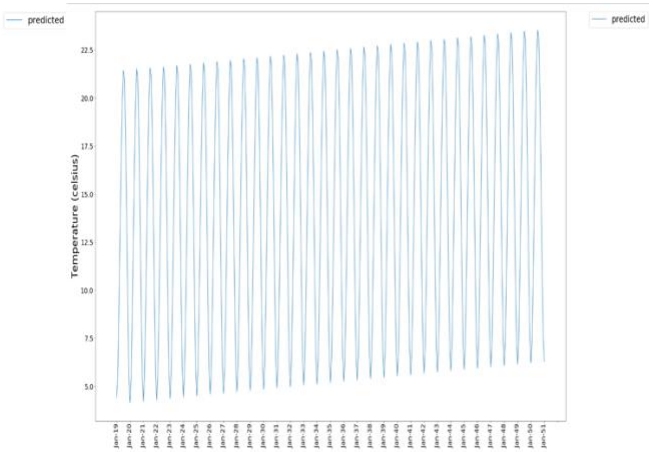

Figure 5: Precipitation model predictions

- Country wise analysis:
  Our country wise analysis from Figure 6 and Figure 7 shows that top 10 and bottom 10 CO_2 emitting countries appear in cluster on the map [11] suggesting that CO_2 emission have far wider reach than the local country. We cannot conclude a certain relationship between CO_2 emission and temperature as beta values were ranging from positive to negative in both the cases.

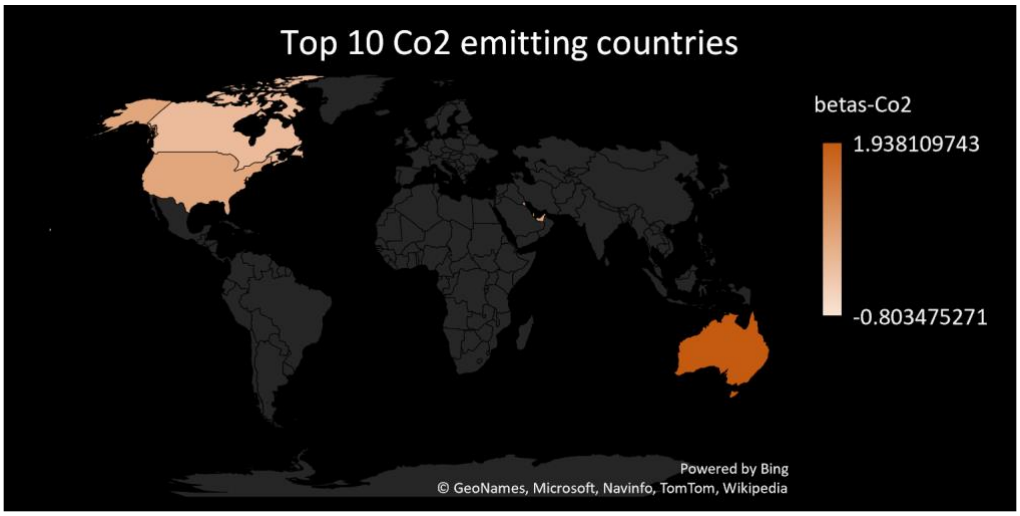

Figure 6: Top 10 CO_2 emitting countries

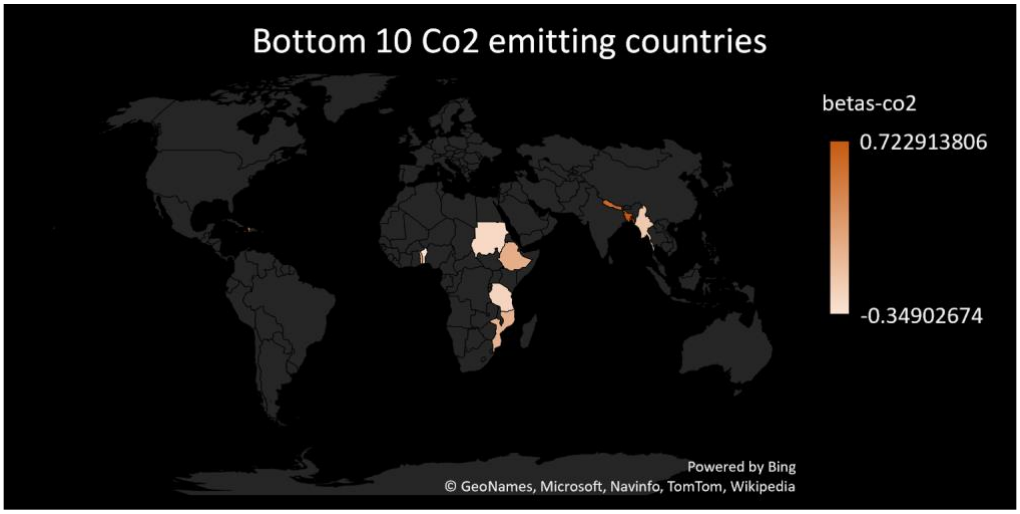

Figure 7: Bottom 10 CO_2 emitting countries

- Hypothesis Testing:
  In the case of temperature and CO_2 we did not find any significant correlation and between every other pairs we rejected null hypothesis as probability of happening by chance(p-value) was less than 0.05. Hence, we would need more data to comment on definite correlation between CO_2 emission and precipitation, energy use and population on climate attributes. The summarized results can be seen in Table 2.

  The below discussed results are based on 55 years of data. It would be better to have data for hundreds of years to have more accurate results and better insights.

Attribute Y	Attribute X	Correlation	T-stats	P-value	Hypothesis
Temperature	CO_2	-2.45e-04	-0.877	0.38	Accept
Temperature	Energy	9.76e-03	4.83	1.4e-6	Reject
Temperature	Population	1.90e-02	13.92	0.0	Reject
Precipitation	CO_2	5.58e-02	3.6	3.2e-4	Reject
Precipitation	Energy	-4.98e-02	-4.2	2.7e-5	Reject
Precipitation	Population	1.34e-02	8.78	2.22e-18	Reject

Table 2: Results for hypothesis testing

Conclusion:

Going at the current pace, we can see a definite rise in temperature and fall in precipitation level in coming years. This could lead to serious issues like wildfires, lost crops, increased floods and shrinking glaciers. Hence, it is high time we take the climate change seriously. From our results, it can be seen that there is variation in effect of CO_2 per country and on global level. It would be helpful to study and implement climate change solutions both regionally and globally. From our hypothesis testing, we did not see any significant correlation between CO_2 emission and temperature. This could be due to the fact that climate change is slow and it takes hundreds of years of data to see the effects. Also, other greenhouse gases like Nitrous Oxide, Methane are also said to contribute towards temperature change. Therefore, it is important to combat climate change and its impacts.

References:

[1] https://www.scientificamerican.com/article/leonardo-dicaprio-uses-oscar-speech-to-urge-action-on-climate-change/
[2] https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/trendanalysis
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5393220/
[4] https://economics.stanford.edu/sites/g/files/sbiybj9386/f/climate_20160208.pdf
[5] https://www.ncei.noaa.gov/data/global-summary-of-the-day/archive/
[6] https://data.worldbank.org/indicator/
[7] https://are.berkeley.edu/courses/EEP118/current/handouts/eep118_panel_data_fixed_effects.pdf
[8] https://www.co2.earth/global-warming-update
[9] https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/
[10] https://www3.cs.stonybrook.edu/~has/CSE545/Slides/tf_demo_2020.py
[11] https://www.empowersuite.com/en/blog/creating-a-world-map-in-powerpoint-this-is-how-it-works
[12] https://www.geeksforgeeks.org/python-reverse-geocoding-to-get-location-on-a-map-using-geographic-coordinates/