Analysis on Temperature Change through the Years

Suraj Peddhibhotla - 19BCE1044 , Girish S - 19BCE1268

Professor Name : Dr. Tulasi Prasad Sariki

Subject : Foundations of Data Analytics CSE3505

Introduction

Climate change is caused by rapidly increasing greenhouse gases in the Earth's atmosphere primarily due to burning fossil fuels such as coal, oil, and natural gas. These heat-trapping gases are warming the Earth and the Oceans resulting in rising sea levels, changes in storm patterns, changes in rainfall, melting snow and ice, more extreme fires and drought.

In the agriculturally important regions of central Maharashtra, the Indo-Gangetic plains and southern coastal zones, rising temperatures and increased extent and incidence of droughts have caused declines in rice and wheat yields.

An estimated 12.6 million people live directly on land that is at risk from sea level rise and nearly 171 million people depend on coastal ecosystems vulnerable to sea level rise, cyclones and storm surges.

Table of Contents

3. Exploratory Data Analysis

- World Temperature Change
- India as a Case Study
- Which Countries have the most and least Change?
- India vs World
- Heatmap & Box Plots
- Correlation Matrix and Analysis

4. Predictive Model Development

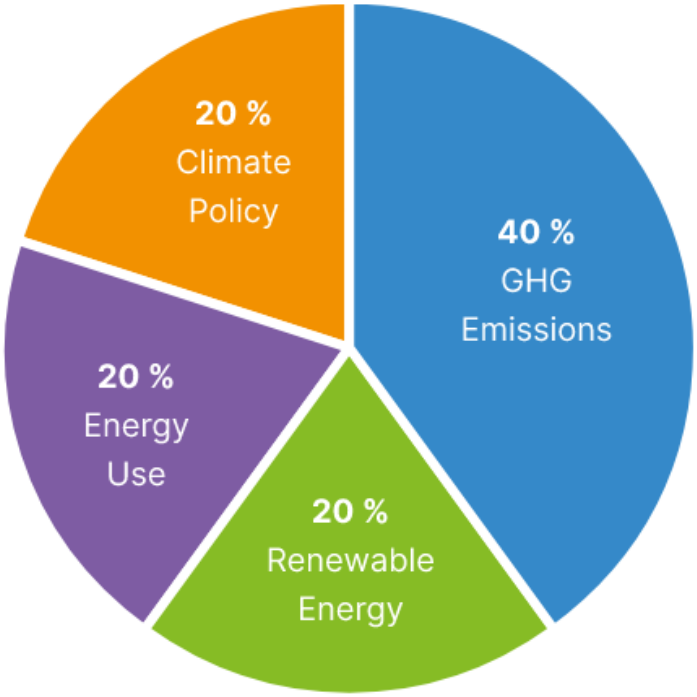
- Multiple Linear and Polynomial Regression
- Validation
- Conclusion
- Recommendations

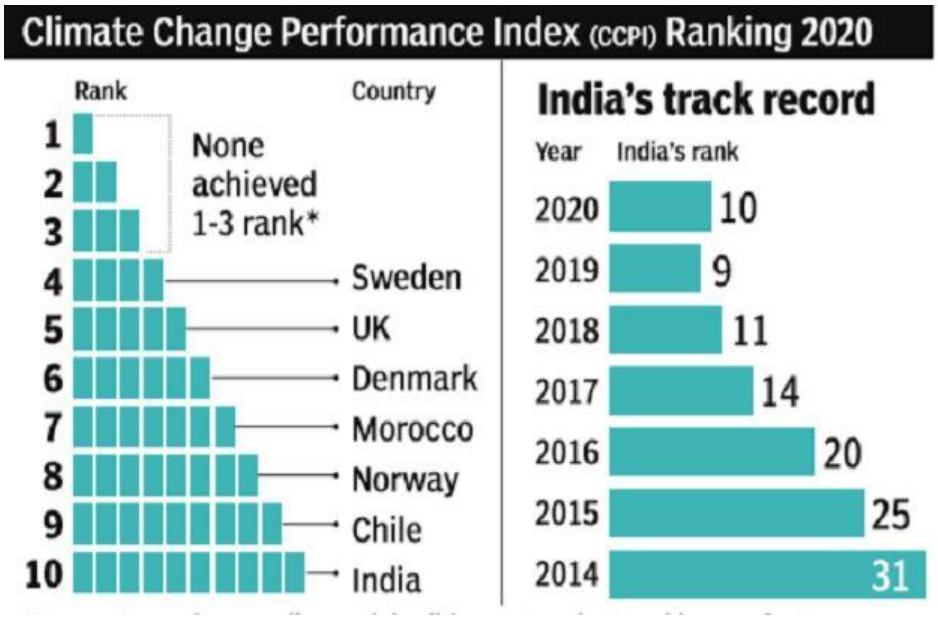
1. Overview

The rate at which climate change has grown across India and the world is alarming. Climate change has major health impacts in India- increasing malnutrition and related health disorders such as child stunting - with the poor affected most severely. Vector-borne diseases are likely to spread into areas where colder temperatures had previously limited transmission. Heat waves have resulted in a very substantial rise in mortality and death with about 2500 deaths have occurred in the year 2015 alone.

The Climate Change Performance Index (CCPI) is a scoring system designed by the German environmental and development organization Germanwatch eV. to enhance transparency in international climate politics.

- The national performances are assessed based on 14 indicators in the following four categories:
    1. GHG emissions (weighting 40%)
    2. Renewable energy (weighting 20%)
    3. Energy use (weighting 20%)
    4. Climate policy (weighting 20%)





- According to the CCPI, none of the countries has yet achieved a performance across all indicators that can be qualified as very high, because no country fulfills the requirements to limit global warming to well below 2°C, as agreed in the Paris Agreement. This is why the first three places in the final ranking remain unoccupied.

- From the recent study, India ranks at 10th position with score of 63.97.

Through this study, we hope to develop some insights that can help organizations (State / Central Pollution Control Boards & NGOs) to advocate more stringent policies to control temperature change.

1.1 Problem Statement

Changing climate is leading to more occurrences of extreme events such as droughts (moisture deficits) and floods (moisture surpluses), which have a negative impact on crop growth and yields.

The number of Indians exposed to heat waves increased by 200% from 2010 to 2016. Heat waves also affect farm labor productivity. The heat waves affect central and northwestern India the most, and the eastern coast and Telangana have also been affected.

Hence, we decided to understand, analyze and try to predict various trends in climate change with emphasis on temperature change across the years. In this project, we use linear regression models and polynomial regression models to predict world temperatures in the future, which tells us the productive measures to be taken for countering this climate change.

We can compare the features of different countries. We will also focus on India as a use case, to analyze monthly data, seasonal data and the yearly data.

1.2 Objective and Scope of the Project

Objective :

- Study Global temperature change distribution from 1960 till the present
- Study the temperature change continent wise
- Study the temperature change for India as a separate case study
- Season wise analysis can be compared with yearly average temperature change.
- Explore the possibility of developing a model to predict how the temperatures will change in the future
- Perform a test-train split for training the model in order to verify the predictability of our model
- Design both a linear model and polynomial regression models to predict world temperatures in the future.

Scope :

- The scope of this study covers temperature change across the world and India as a separate case study.
- The study covers temperature change with respect to a baseline climatology, corresponding to the period 1951–1980.
- The study's focus is on factors for which authentic secondary data are available that can be used for Statistical Analysis
- Continent and season wise data can also be analyzed

1.3 Data Sources

The Dataset has been extracted from FAOSTAT Temperature Change domain
The FAOSTAT Temperature Change domain disseminates statistics of mean surface temperature change by country, with annual updates. It currently covers the period 1961–2019.
Statistics are available for monthly, seasonal and annual mean temperature anomalies, i.e., temperature change with respect to a baseline climatology, corresponding to the period 1951–1980.
It contains details such as:

- Area(Country/Continent)
- Time Period(Month/Year/Seasonal)
- Temperature Change values from 1961 to 2019

1.4 Tools and Techniques

We have used the following Analytical techniques / methodology for analyzing the Data :

1. Data extraction from CSV File into a Data frame
2. Data Cleaning and Preparation
3. Exploratory Data Analysis using Line Charts, Box Plots, Histograms, Pie Charts etc
4. Splitting data into testing and training
5. Using Multiple Linear Regression & Time Series Analysis
6. Tools used: RStudio for R Programming Language
7. Techniques: Box Plot, Histogram, Bar Chart, Line Chart, Infographics, Correlation Matrix, Multiple Linear Regression

1.5 Limitations

There are few limitations that this study has w.r.t data and the methodology

- Due to world politics being in constant flux many countries have come into existence which makes our dataset incomplete.
- Since the dataset obtained from FAOSTAT contains temperature change and standard deviation with respect to a baseline period of 1950-1980 and not the actual temperature, we can predict the future temperature change and not the future temperature.
- The data contains only a few factors such as country, time period and temperature change. Additional features that may affect temperature change are not available to perform analysis

2. Data Description and Preparation

2.1 Data Management

- We were able to obtain temperature changes for all nations from 1950 to 2019. The FAOSTAT Temperature Change domain provided the data for this dataset. The FAOSTAT Temperature Change domain disseminates annual updates to statistics on mean surface temperature change by country.
- The present distribution spans the years 1950 to 2020. Monthly, seasonal, and yearly mean temperature anomalies, i.e. temperature change relative to a baseline climatology, for the period 1951–1980, are provided as statistics. The baseline methodology's standard deviation of temperature change is also accessible.
- We were able to forecast future temperatures using this method.

2.2 Data Tables

Table : List of Variables and their Types			
Variable Abbreviation	**Variable**	**Variable Type**	**Data Type**
Area Code	Country Code	Meterological	Categorical
Area	Country Name	Meterological	Categorical
Months Code	Months Code	Meterological	Categorical
Months	Month Name	Meterological	Categorical
Element Code	Element Code	Meterological	Categorical
Element	Temperature Change / Standard Deviation	Meterological	Categorical
Unit	Unit Temperature	Meterological	Categorical
Y1961	Year 1961	Temperature Change	Continous
Y1962	Year 1962	Temperature Change	Continous
Y1963	Year 1963	Temperature Change	Continous
.	.	.	.
.	.	.	.
.	.	.	.
Y2017	Year 2017	Temperature Change	Continous
Y2018	Year 2018	Temperature Change	Continous
Y2019	Year 2019	Temperature Change	Continous

2.3 Data Quality

- Temperature Change and Standard Deviation for certain countries were not available, so they were empty. Moreover, some countries did not have temperature Change and Standard Deviation, so those countries were removed.
- Various codes like Month code and Country code were not used for our prediction and they are not necessary. So, we removed them from our analysis.
- Our dataset is of dimensions 6760x62 and there were 2896 rows which had empty values. Thus, these rows were removed to get the best result.

2.4 Data Preparations

Variable Transformation :

- We renamed the Area column to country.
- We converted the season wise-grouped columns to seasons.

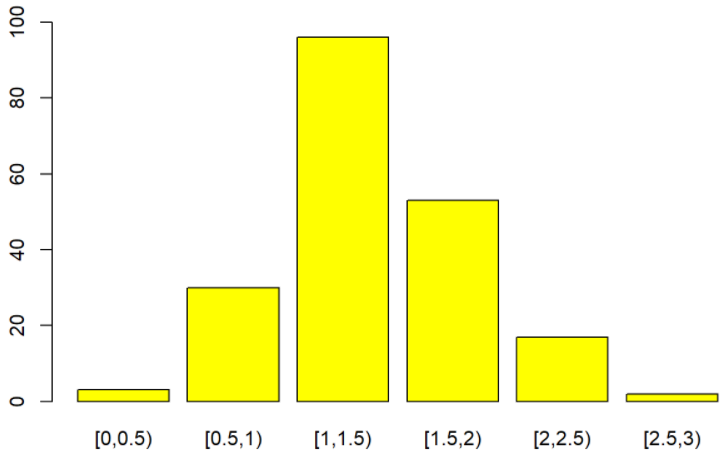
Missing values and Outliers :

1. No specific missing value treatment was used.
2. Rows For which no data was available for the key variables, then that year's record was removed
3. Only rows where observations were recorded for the key variables were included in the analysis.
4. Rows in which outliers were present, the year's record was removed from the data.
5. Deleted the columns which had codes.
6. We also removed the unit column as all temperatures are in Celsius.
7. Renamed Column Area to Country for ease of understanding.
8. Converting group of 3 months to Seasons.

3. Exploratory Data Analysis

(a) World Temperature Change - Histogram & Density Plot

To analyze the data across the world, the Temperature Change in the year 2019 is chosen due it being the last year in the dataset. We begin by splitting the temperature changes into intervals each of range 0.5. We then find how many Countries belong to each interval. A Histogram is made for the same

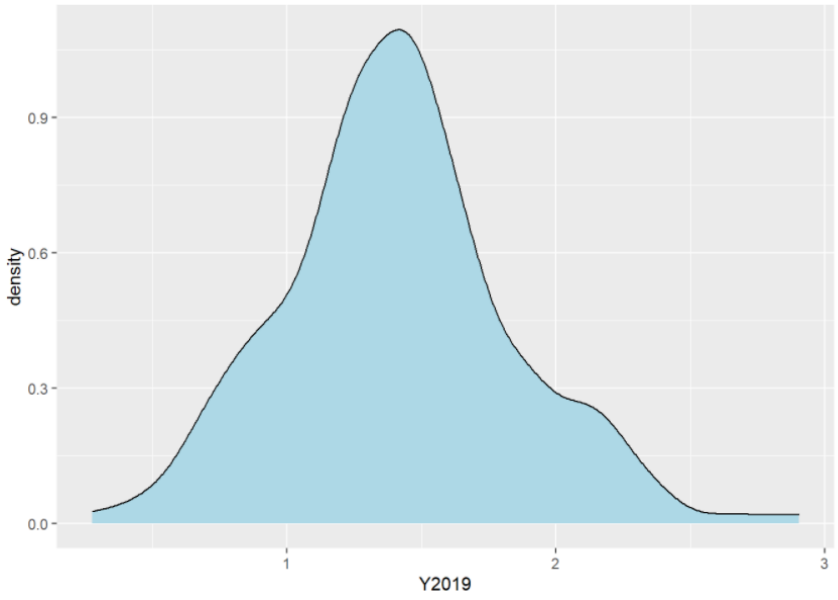


[0,0.5)	3
[0.5,1)	30
[1,1.5)	96
[1.5,2)	53
[2,2.5)	17
[2.5,3)	2

Insight :

- We can see that most of the countries are present in the range [1, 1.5) (96 countries) . The main cause of worry is that 53 countries belong to [1.5,2) range indicating that increasing global temperature values

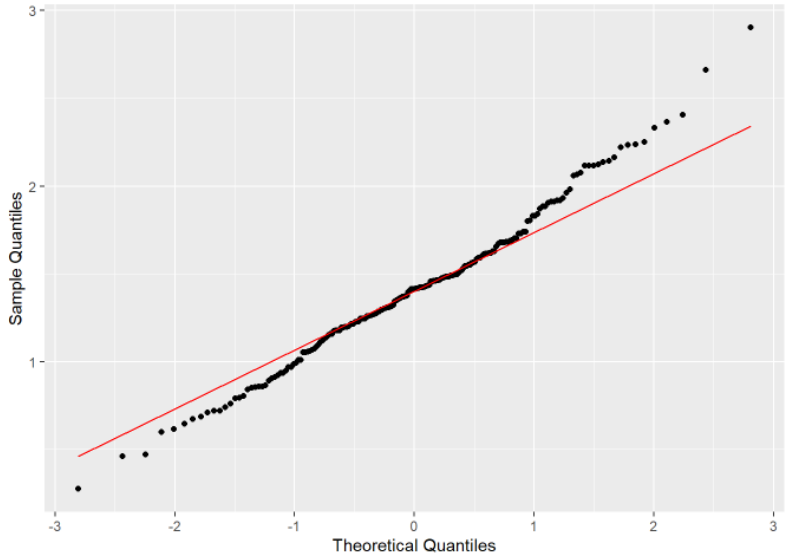
We then proceed to develop a density plot for the year 2019



Insight : We can see from the density plot similar insights can be drawn.

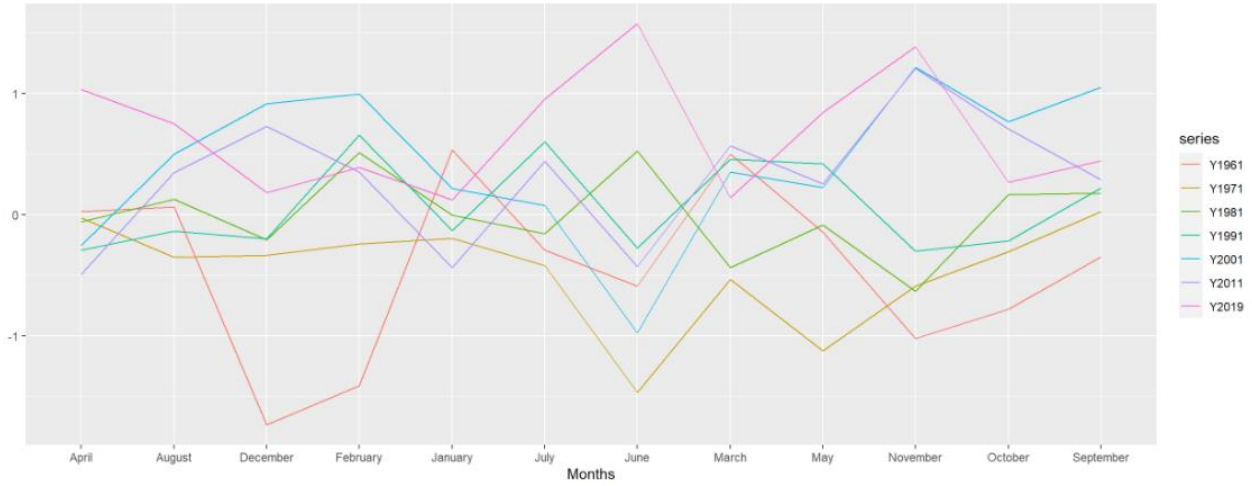
- The peak of the density plot is around 1.5 and most of the data is in the range [1.2]
- The plot is slightly positively skewed with skewness close 0.32
- The plot is also Leptokurtic as we can notice a rather sharp peak in the distribution

We can notice from the plot that the distribution is nearly normal. To view this statistically we can plot a qq-plot. We notice that not all points coincide with the line indicating that it's not completely normally distributed and has slight skewness and kurtosis as mentioned above.

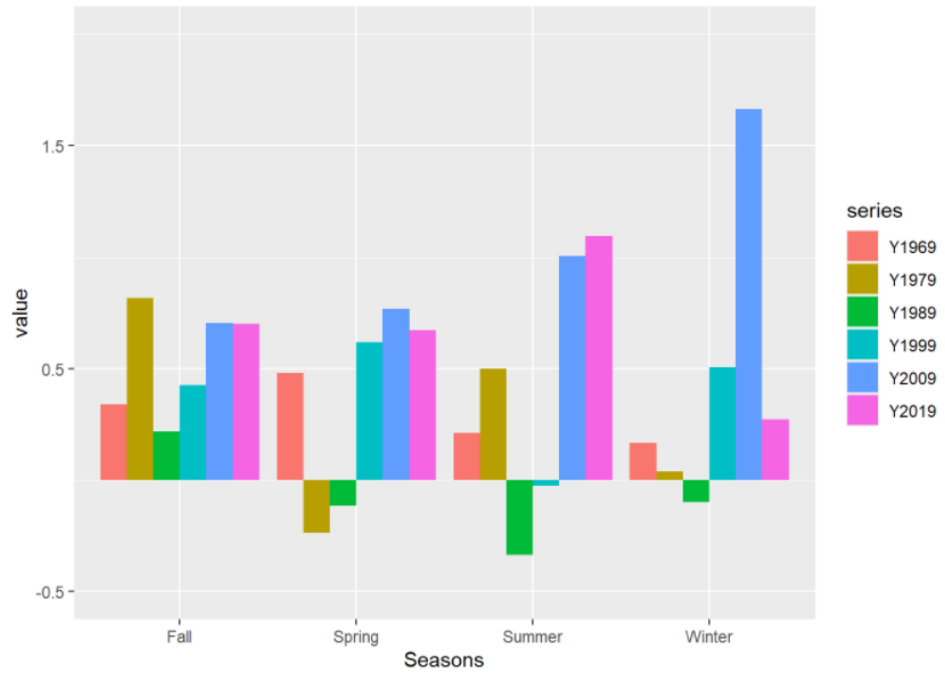


(b) India as a Case Study

We first begin with analyzing the Temperature Change across the decades starting from 1961 to 2019 across each Month of the year.



Insights : While month wise spread of the Temperature Change doesn't seem to follow any specific trend, in most of the years the value remains between -1 to 1. However, we can see that the highest peaks occur in the Year 2019 with values crossing 1.5 in June and nearly the same in November.
We then plot the season wise Temperature Change in India in each decade and try to identify the trends.



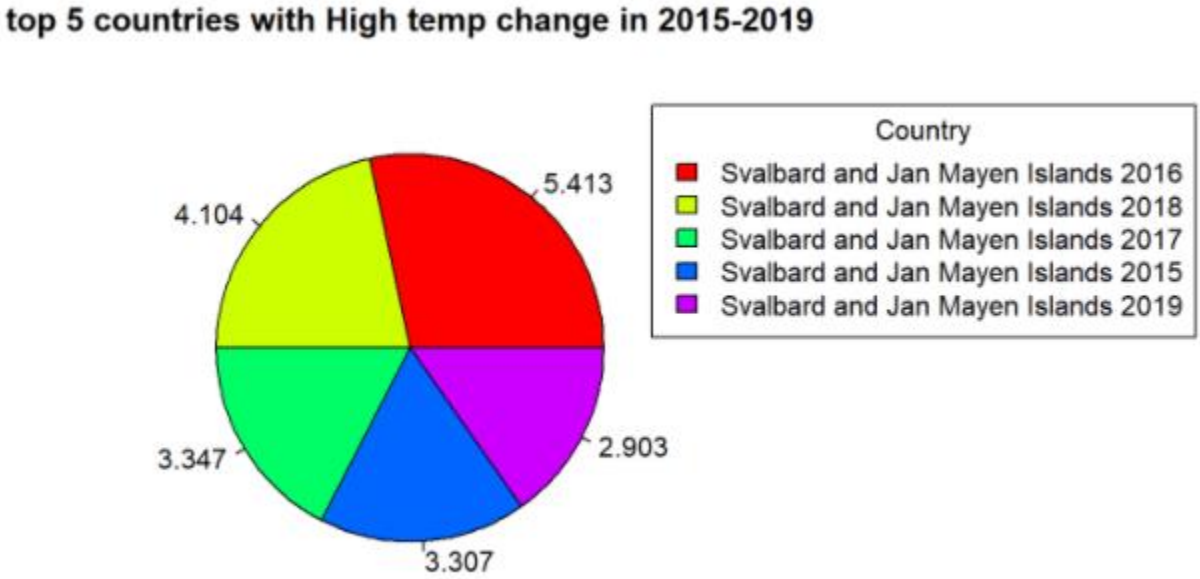
Insights : Through this seasonality analysis, the following insights could be drawn

- During the Fall season the temperature change tends to be uniform (approximately 0.5).However in Summer and Winter the values tend to deviate much more.
- We can also notice that the values in the Years 2009 and Year 2019 which belong to this decade, the values are the highest indicating the more serious nature of Climate change

(c) Which Countries have the most and least Change?

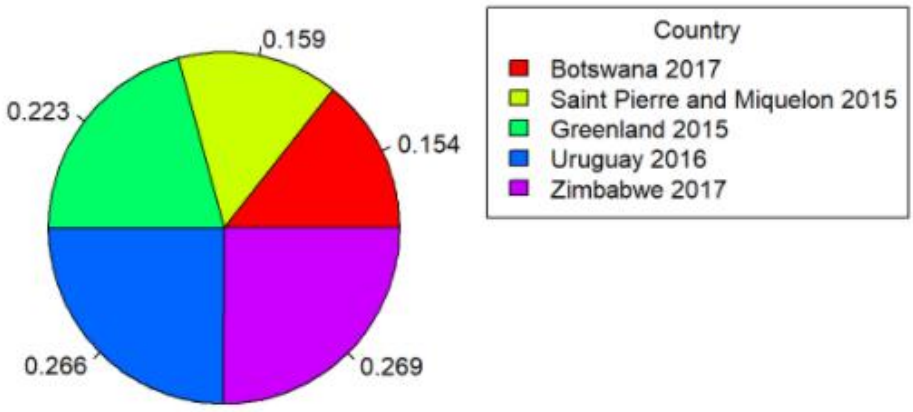
This was a question we were most interested in finding out. We tried to calculate the Top 5 Countries with the highest and lowest temperature change in the past 5 years(2015-2019). So we filtered the top 5 values from each year and merged them into a single list. We then sorted the list in descending order and chose the top 5 Countries.

(1) Highest Temperature Change



Here is a pie chart showing the values of Temperature change. We were able to identify that highest all belonged to the exact same country, Svalbard is an island region which is the northernmost place in the world. It turns out that it is the worst affected region with change nearly 5 times the global average.

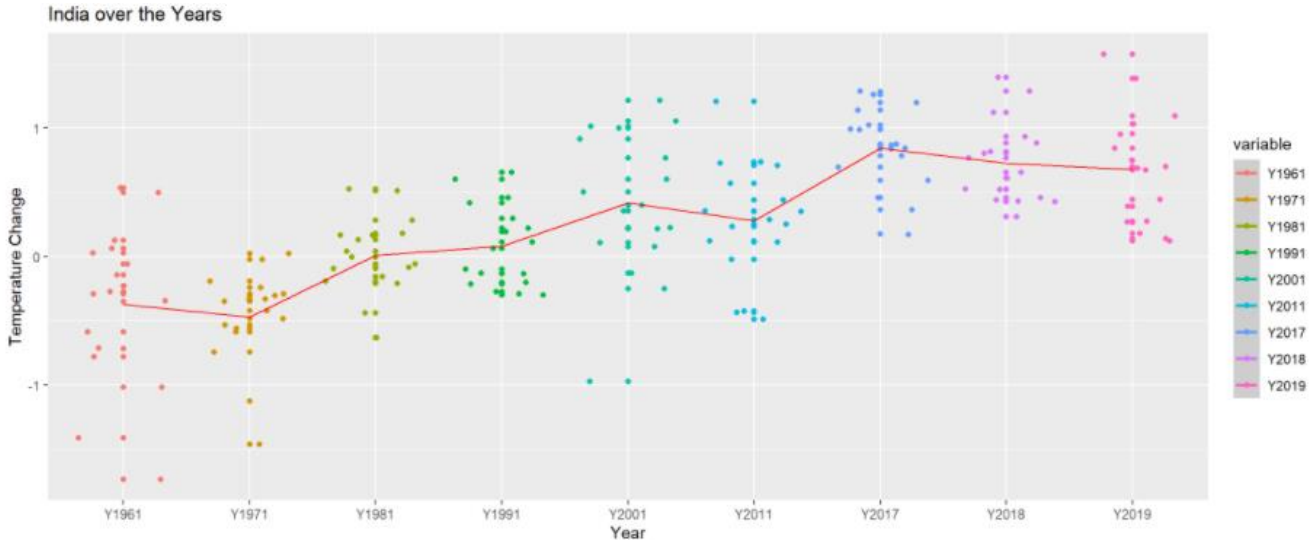
(2) Lowest Temperature Change

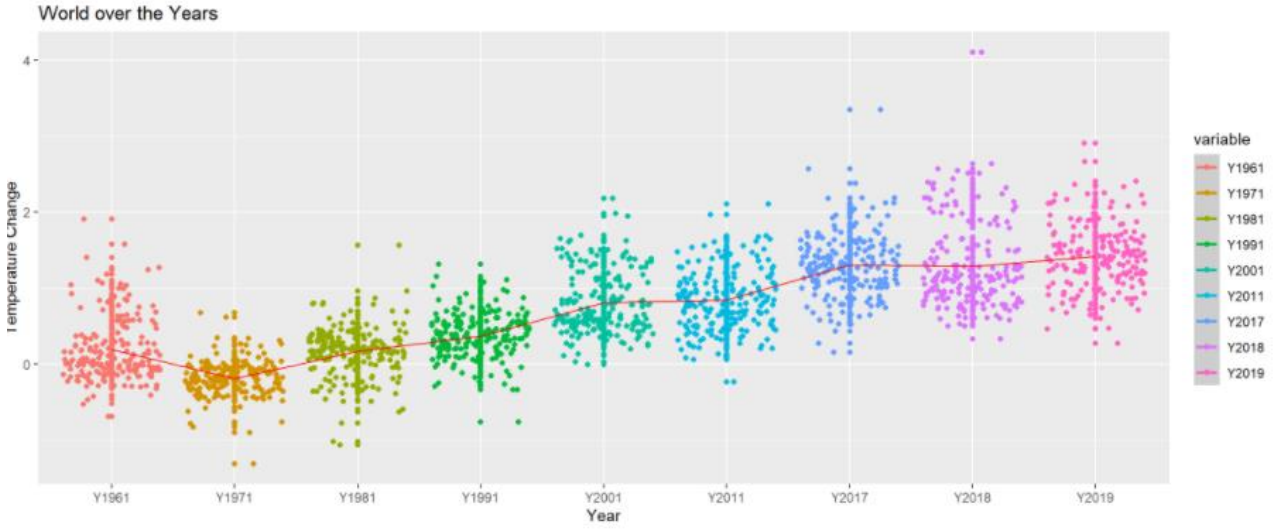


However for the lowest temperature change we were able to identify from different regions from different continents across the world.With lowest at 0.159. We can also notice that the lowest temperature change does not belong to the years 2018 and 2019 and happened before them.

(d) India vs World

We then try to compare the temperature change in India and rest of the world and try to see the differences and fluctuations over the years. We plotted a scatterplot of temperature change in India and another for the world.



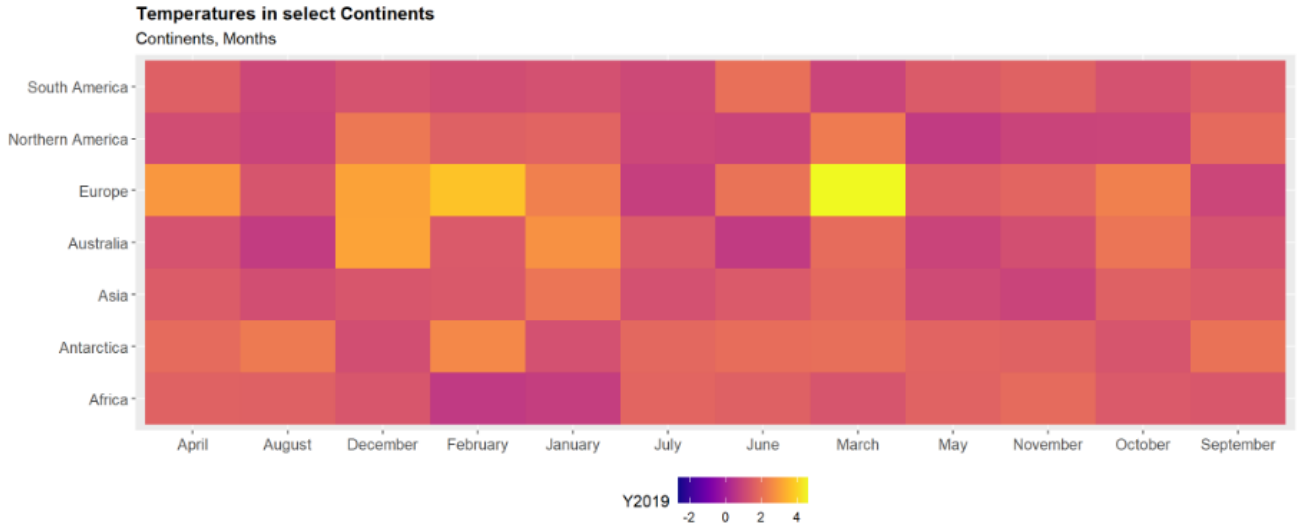


Insights :

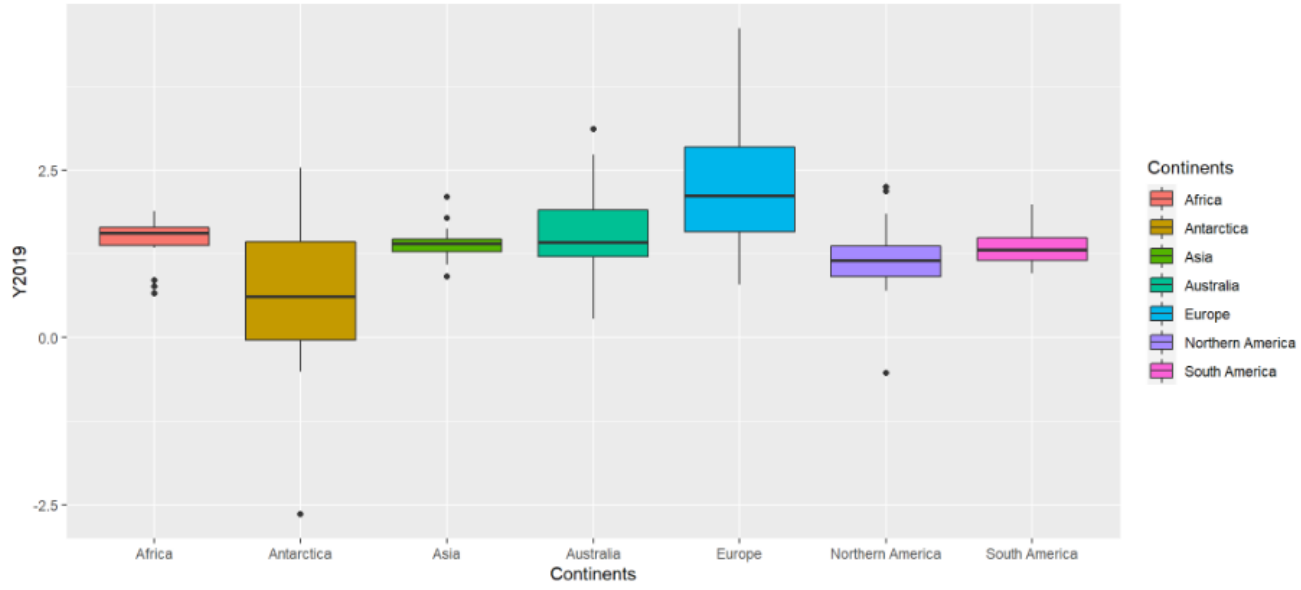
- We can notice that both follow similar trends with value increasing yearly.
- However in the world the value is usually closer to 1 and above however in India the value always stays below 1
- Also in India during the years 2017 to 2019 the value is decreasing which is a positive

(e) Heatmaps and Boxplot

To analyze the continent wise temperature change, we plotted a Heatmap for the values in each of the 7 continents for each Month in the Year 2019. This has been made using the Viridis Colour Map and the scale displayed with the graph with colour range from Yellow (highest) to Dark Blue(lowest)



We then plotted a Box Plot for each of the continents.

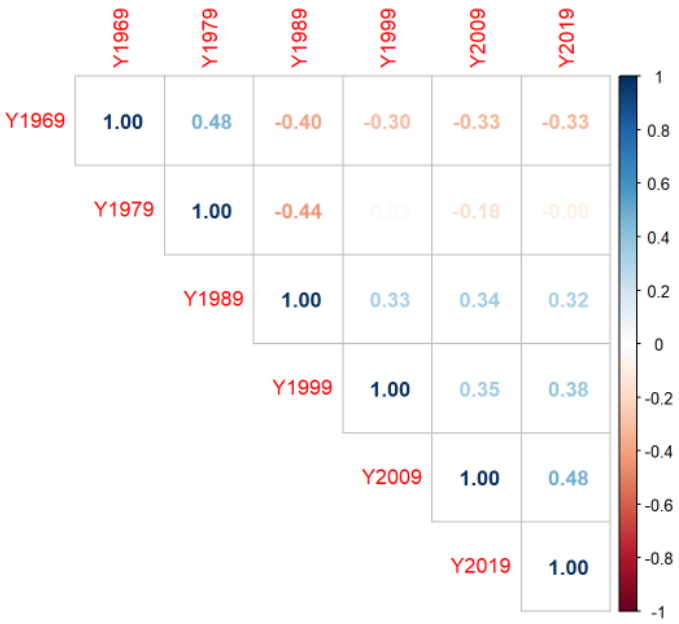


Insights :

- From the heatmap we can see that Europe has the highest Temperature change with the brightest point in March.
- From the Boxplot we can find the Median values for each continent. Just like heatmap we can see that Europe has the highest value.
- Antarctica, surprisingly has the lowest median despite being most affected by climate change.
- We can also see that there are zero or very few outliers in the dataset for each continent

(f) Correlation Matrix & Analysis

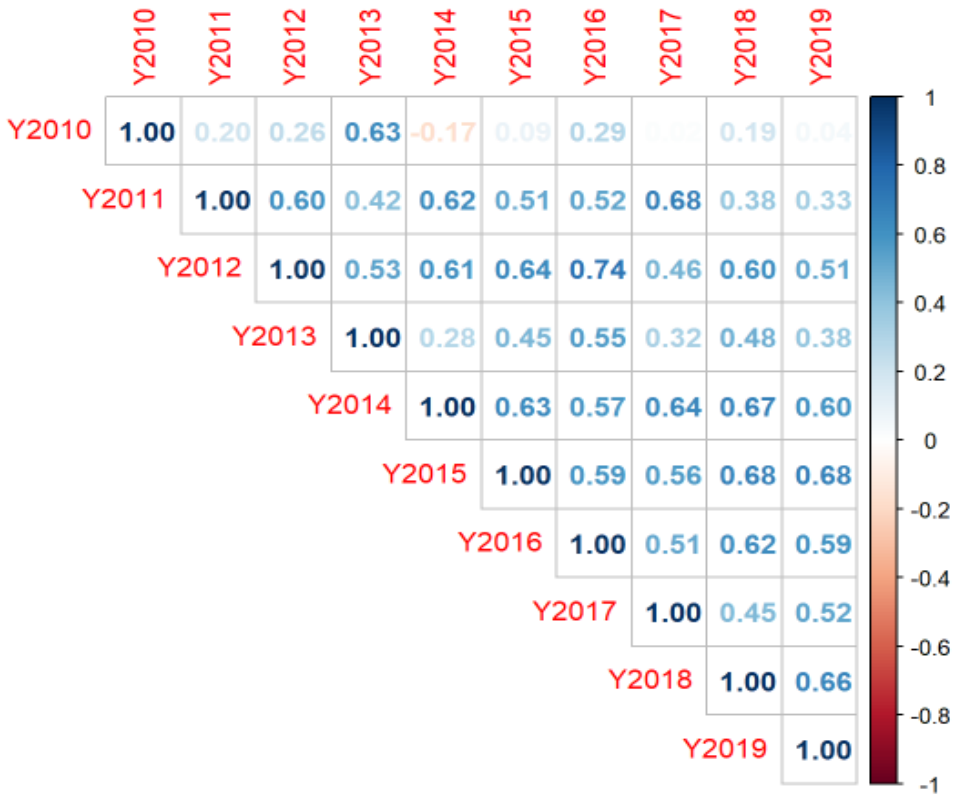
A correlation Matrix was developed for each year of every decade from 1969 to 2019.



Insights :

- We can see that Year 1969 and Year 1979 have negative correlation with almost all other years.
- All the years from 1989 to 2019 have only positive correlation with each other.
- On the whole the correlation between the years is low with value remaining less than 0.5

We now try to find a relationship between the Temperature change in this decade from 2010-2019 to identify how linearly dependent 2019 is on the previous few years.



Insights :

- Here we can see that most of the years have much stronger correlation and are usually positive correlated with each other
- The Highest correlation exists between the year 2016 and 2017 with value 0.74
- The Year 2019 has fairly strong correlation with most of the years especially year 2015 with 0.68 and 2018 with 0.66

4. Predictive Model Development

4.1 Multiple Linear and Polynomial Regression

For predicting the Temperature Change in year 2019 we used Multiple Linear Regression

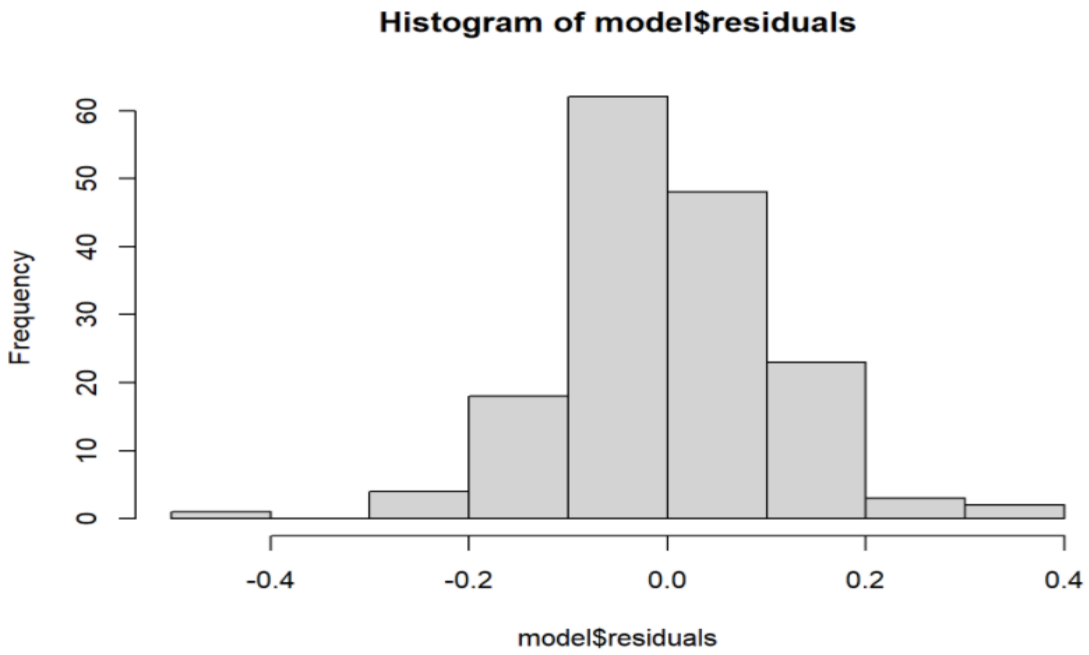
Test-Train Split :

- Before we can make a prediction we need to train our model.
- We will split the data into a test and train a dataset with an 80:20 split.
- This can be used to verify the predictability of our model.

Linear Regression:

- We use the test-train data to train the model and compare the predictions to actual data. The value of RMSE obtained is 0.1729397 and the value of R-square is 0.8249301.
- Since the value of RMSE is close to 0.2, it shows that our model can relatively predict the data accurately.
- Also, the R-square is 0.8, which means 80% of the variation in the output variable is explained by the input variables.

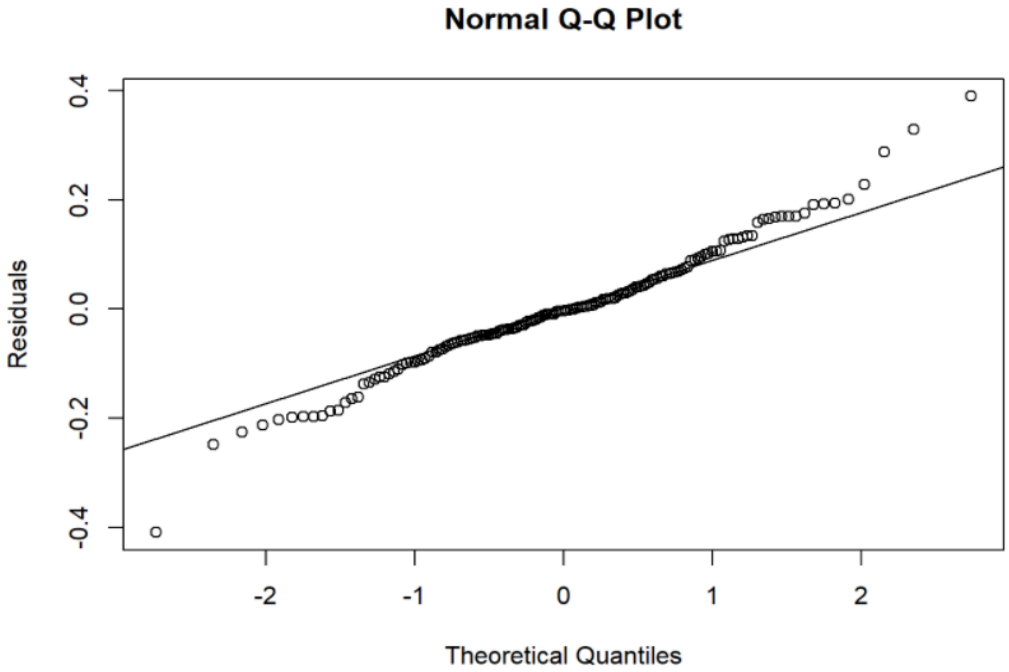
(a) Normality Assumption :



Insights :

- If we join all the peaks of the histogram, we get a normal curve.
- This gives us an idea of the normal distribution of residuals.

(b) Normal Q-Q Plot :



Insights :

- From the graph, we can say that the points are closer to the Q-Q line.
- Since we get a linear line plot, we can say that the residual is normally distributed.

Thus, our linear regression model predictions of existing years are very close to the actual results.

Predicting Temperature Change in the Future

So, we built a multiple linear regression model for the same. But the RMSE and R-Square values obtained were 0.4034857 and 0.5595437. These values indicate that data is not linear, thus we moved onto Polynomial Regression.

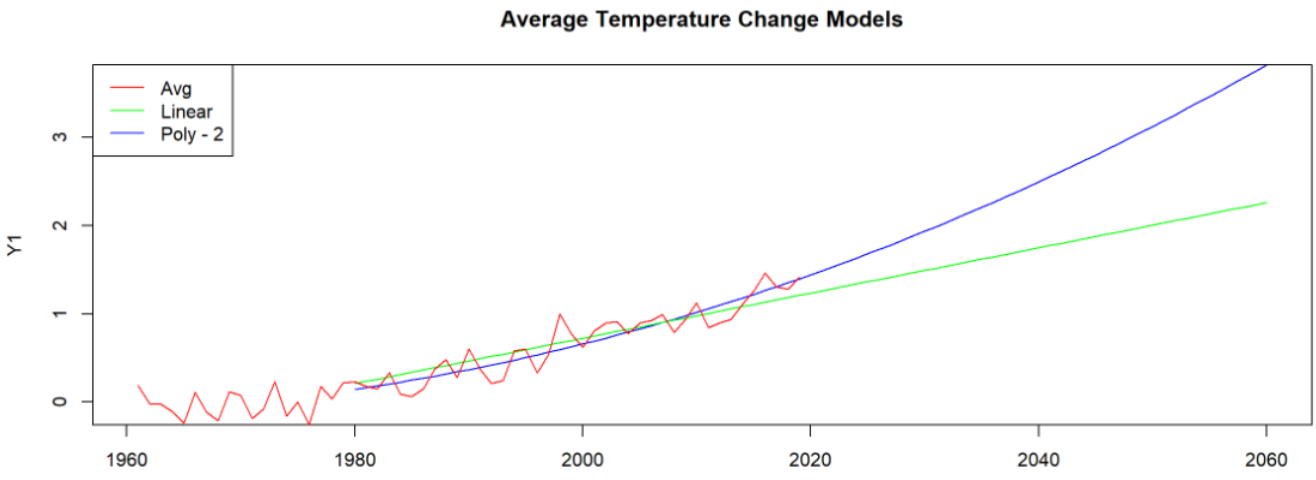
Polynomial Regression :

- As seen in different plots the data is clearly not linear so we experimented with some nonlinear features to see if we can get a more accurate prediction.
- We created test data to validate so that we can use it to test what the model predicts for the future(2020-2060).
- The range obtained through polynomial regression was [0.1458148, 3.8215878].

4.2 Validation

We plotted the results for the linear and polynomial models side-by-side to see how they fair.

(c) **Comparison of Linear and Polynomial Regression Models**



- As this is a linear model the prediction is pretty simple without any big fluctuation.
- The model is still predicting that the temperatures will rise in the future.
- From the model, it is evident that the polynomial model with degree 2 is predicting more accurately than the linear model.

4.3 Conclusion

Polynomial regression model predicts a much higher temperature change and there does not seem to be much change in adding more factors indicating that currently the data does not have high order. At the edge of the available data it seems that the polynomial regression is more accurate than the linear model.
So unless the world does something to mitigate the rise the increase in temperature will continue at a much higher rate.

4.4 Recommendations

1. To reduce the usage of greenhouse emitting machines and vehicles.
2. Invest in energy-efficient appliances
3. Reduce water waste
4. Reduce, reuse, repair & recycle
5. Generate Clean Green Electricity.