Analysis on Climate Change

|  |
| --- |
| Skanda Abhishek Venkatesh |
| PES1201700987 PES1201701563 PES1201701626 |
| Department of CSE Department of CSE Department of CSE |
| PES University PES University PES University |
| Bengaluru, India Bengaluru, India Bengaluru, India |
| [skandavc18@gmail.com](mailto:skandavc18@gmail.com) [abhishek574abhi@gmail.com](mailto:abhishek574abhi@gmail.com) Venkatesh.kumar.4188@gmail.com |

Abstract :

Climate Change has become a hotly contested debate nowadays. Many people believe that Climate Change is real and is severely affecting the world. But at the same time , there are few people who believe that it is a hoax. The project aims to an analysis on temperature change which has been happening throughout the world and how various geographic regions of the world are being affected. Also, the project aims to predict future temperatures so that timely remedies are taken to prevent Kerala and Kodagu flood like events.

Introduction :

The aim is to be able to predict the future yearly temperatures and know the trends.

The data-set was cleaned to fill any voids. After that the data was organized so that it could be effectively used for the future part of the analysis.

In descriptive analysis part, the distribution of data , the various statistics like mean, median, upper and lower quartile were found to know more about the data. A histogram and box plot were plotted to know more about the distribution of data.

After that correlation analysis was carried to find out whether there is really any relation between time and temperature. Also the relation between latitude and increase in temperature was found out.

Lastly the regression model was developed which gave satisfactory results.

Data Set :

The data set was taken from Kaggle. It consists of 7 columns : date, average temperature, average temperature uncertainty , city , country ,latitude and longitude.

It has 8.6 million rows. There are totally 3448 cities spanning across 159 countries. The latitudes too vary vastly between -40 degrees or 40 degrees South to 60 degrees North.

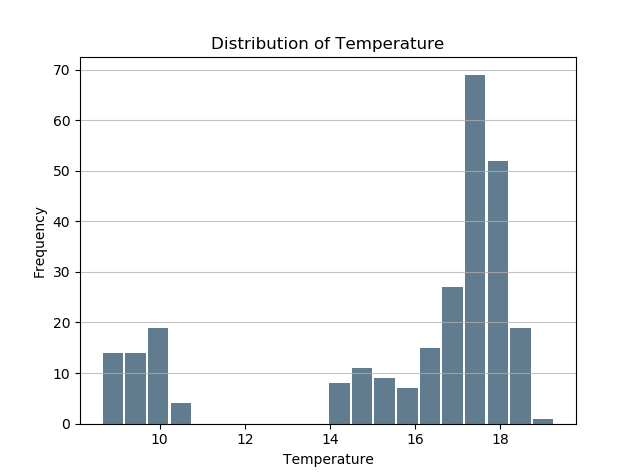
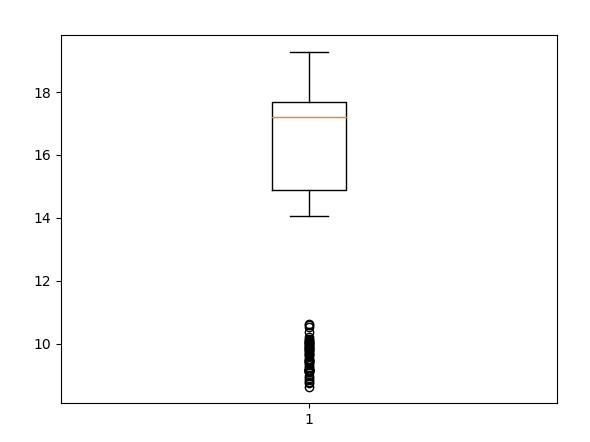
The cities have values from years 1744 to 2011 month wise.

Link to Data Set : https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

Data Cleaning :

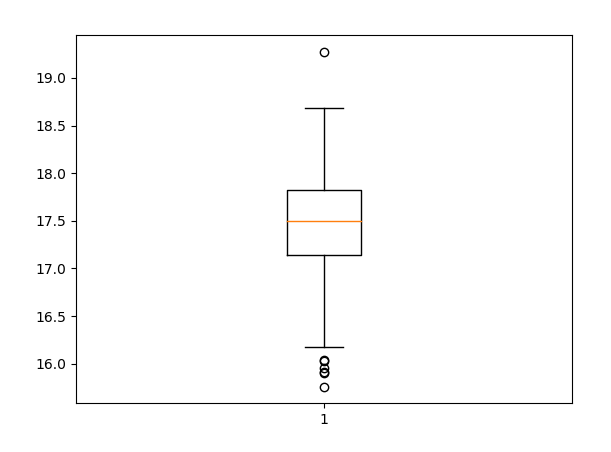
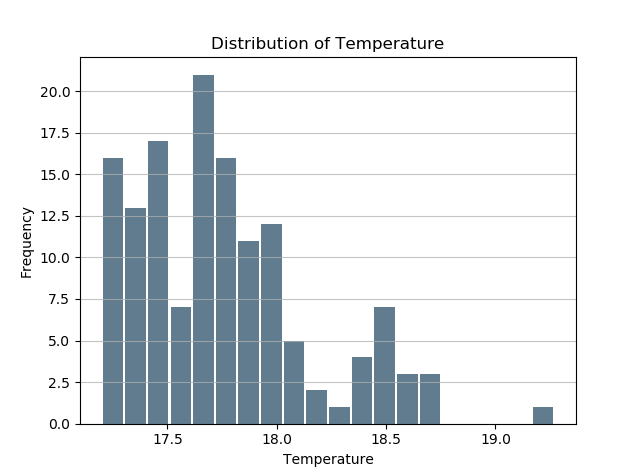
Data Cleaning is the first part in any analysis. There are bound to be many inconsistencies in such large data sets. There were many missing values in our data set. So we filled those missing values with the mean temperature for that month from 1744 to 2011.

Since Longitude is not being considered , we created a new cleaned data set which consists of only time, city , date, latitude and country.The average temperatures were found out for all years and a box plot was plotted to check for any outliers. The data had a lot of outliers.Mean Yearly Temperatures before removing Outliers



Since the data had a lot of outliers , to make predictions more accurate , it had to be removed. The Regression models are generally sensitive to outliers.On inspection of data , it was found that temperature data for all years bellow 1875 could be considered as outliers. The data was also highly erratic. It was later found out that it was from 1870’s that reliable temperature data for cities are available. So those values were before 1875 were not not considered.

After Outliers were removed, the distribution of data looked like this.



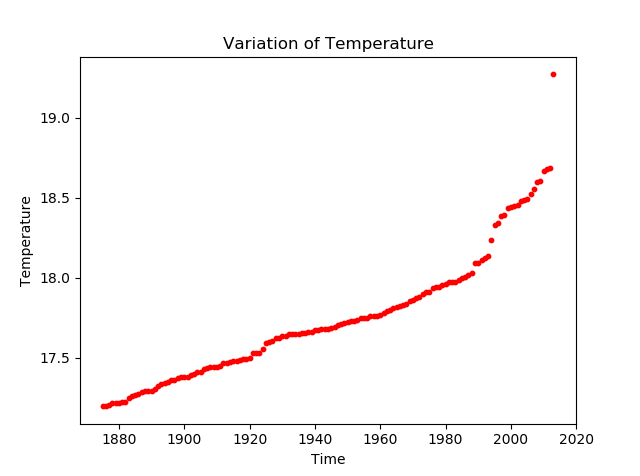
Descriptive Analysis :

Before correlation analysis was done, the descriptive analysis part was done.

Looking at the histogram, the distribution is right skewed. The following is the summary of data

1. Mean Temperature 17.75355448096384
2. Standard Deviation of Temperature 0.40089795866950517
3. Median Temperature 17.683528846153845
4. Lower Quartile Temperature 17.443350083930724
5. Upper Quartile Temperature 17.950036847103515

After that , a scatter plot of yearly mean temperature vs time to get a brief idea if yearly temperature and time are related or not.



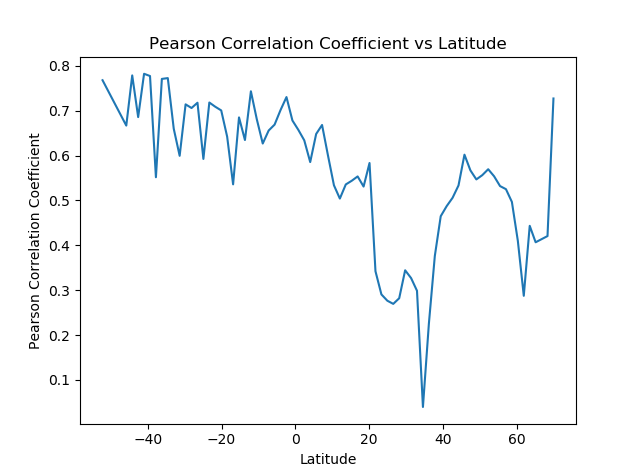
So there may be a positive correlation between time and temperature. But scatter plots can be misleading. A slight change in scale can lead to an altogether different conclusion. So Correlation analysis was done.

Correlation Analysis :

Pearson’s correlation coefficient was found out to find the strength of the relationship between Time and Temperature.

The result was in tune with the scatter plot analysis. The Coefficient was found to be 0.88 which means there is a very strong relation between the 2 variables.

To know how the strength of relationship between temperature and time has been varying across various regions, Pearson’s correlation coefficient between temperature and time was found out for all cities. All these coefficients were plotted against latitude.

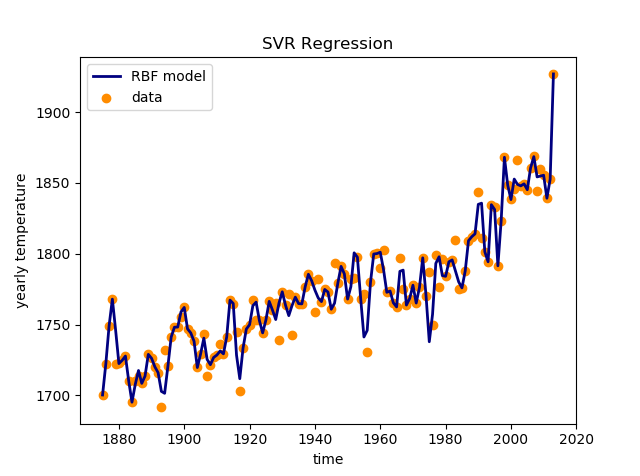
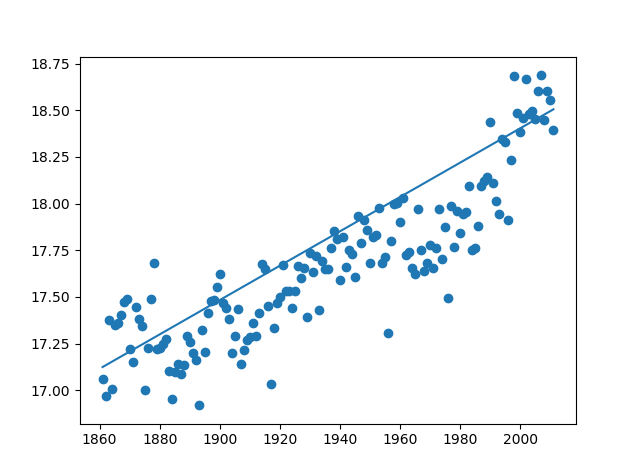


Regression Analysis

Since there was a strong positive correlation, linear regression was done. But it was not fit a best-fit line. And even with fine tuning , it could not be obtained.

Though the data points seem to lie on a straight line with a positive slope, in reality, the data is non-linear where there are slight rise and fall in temperatures with an overall rise in temperature over greater temperature intervals.

So non-linear regression was used. SVR or Support Vector Regression was used with RBF (Radial Basis Function) as Kernel. The parameters C and gamma were optimized to get accurate predictions. The y axis data which is the temperature data was scaled so that accuracy could be increased. The accuracy of the model is 97.7%. The model fits most of the points.



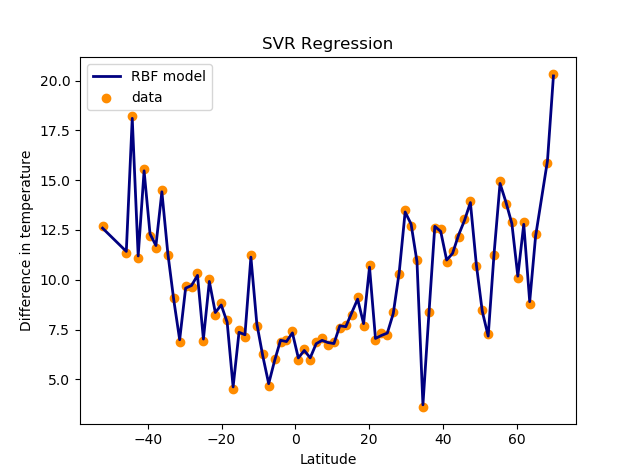
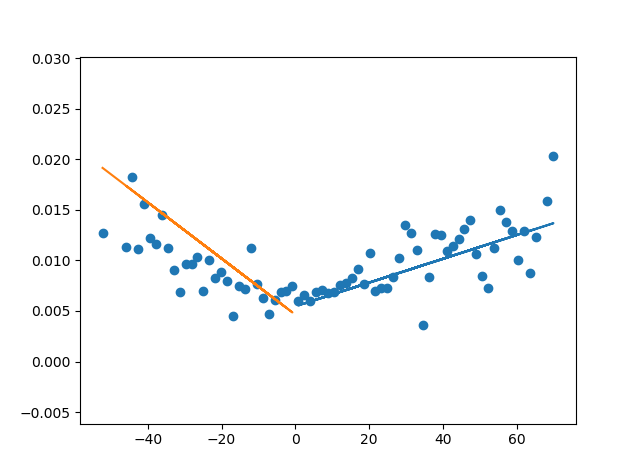
With Linear Regression With SVR Regression

The temperature values in the data set are till 2013. The value predicted for 2017 is 19.23 C. Generally Yearly temperatures are expressed as rise in temperature with respect to 20th century average.

The 20th Century average is 17.73 C. So almost 1.5 C increase in 17 years.

According to the organizations maintaining temperatures, there was an increase of 1.3C in land temperatures with variation of 0.15 C [1] which is very close to predicted value. Also our data is just a sample of 3000 cities in the world concentrated between 40S to 60N. So the predicted value is in very close agreement with the expected value.

Also the average rate of increase in temperature was also found out for each city and plotted against latitudes to find out how temperature change is across regions. In Linear regression , a best fit line could not be obtained since the variation was not linear. So SVR was again chosen as the regression technique.The data was scale d so that a curve could be fitted.



With Linear Regression With SVR Regression

[1] : https://www.ncdc.noaa.gov/sotc/global/201713

Conclusion

The key takeaways of the Regression model which were constructed are the there has been a long term increase in temperature irrespective of the region. But the trends are not uniform across all regions. According to the model, the regions away from the equator are going to be affected the most since the rate of increase in temperature keeps on increasing as one moves away from the equator. So coastal regions will be affected the most since the increase in temperature is highest near the poles, the polar ice caps will at a faster rate thereby increasing the global mean sea level and thereby flooding most costal areas. Since most of the major cities are in coastal areas , it can cause a lot of problems.

So precautions must be taken before it becomes too late.