A

PROJECT REPORT

ON

Study of Effect of Climatic Factors on the Prevalence and Intensity of Wildfires

DEPARTMENT OF COMPUTER SCIENCE

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY

57 US HIGHWAY 1. NEW BRUNSWICK, NJ 08901

By

Neil C Pillai - NCP67,

Ronit Kumar De - RKD55,

Dheeraj Goli - DG1009,

Lakshmi Priya Gayatri Vutukuri - GLV16,

Shanmukh Aditya Yenikapati - SAY38,

Nikhil P Panikulangara - NPP88,

Vidhya Venkatesan - VV256

ABSTRACT

Wildfires date back to the Silurian period i.e. about 420 million years ago and they are not an uncommon sight in the contemporary world. They have been increasing in their spread and frequency over time, culminating in record-setting wildfire seasons, such as those in 2020 and 2021 in California and Australia which have defaced and devastated wildlife and their habitat almost irreversibly. There are several reasons why forest fires take place. It could be due to natural reasons like global warming, lightning or even dry climatic conditions. It could also be a product of human activities like the campfires left unattended, the burning of debris, equipment use and malfunction such as downed power lines, negligently discarded cigarettes, firearms and fireworks and acts of arson and California gender-reveal forest fires. In this report, we aim to examine the impact of climate change upon forest fires in the USA and determine if any correlations exist. Furthermore, we seek to model, utilizing various machine learning approaches, the frequency and intensity (as defined by area of coverage) of US wildfires based on climatic features.

Moreover, we also intend to explore through relevant visualizations, the characteristics of the fire, it's spatial distribution over the different states in the USA and find out if certain areas are more prone to forest fires than others. As per historical data, about 10% of forest fires are caused due to natural events like lightning strikes and volcanic lava. Climate change has been intrinsically linked by scientists to increased daytime temperatures, reduced surface moisture content and higher frequency of lightning strikes (According to a study by California University in 2015, every 1 degree increase in atmospheric temperature can increase the risk of lightning strikes by 12%, a figure which is predicted to touch 18-20% over the next century). The economic impact of forest fires in California in the year 2018 had cost the state over 100 billion USD i.e. roughly around 0.5% of GDP of the country, which is quite significant. This gives us enough motivation to explore the below mentioned datasets and draw useful insights about climate change and how it may have been linked to the forest fires over the years in the USA.

DATASETS USED AND UNDERSTANDING THEM

We have referenced two datasets publicly available on Kaggle (https://www.kaggle.com), a popular Data Science platform which frequently publishes bonafide datasets from a wide variety of subject areas ranging from education, astrophysics to genome research and cell cancer detection.

The two Kaggle datasets relevant to our project are the following:
https://www.kaggle.com/rtatman/188-million-us-wildfires
This dataset was exported from the Fire Program Analysis fire-occurrence database (FPA FOD) which contains close to 1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned during the 24-year period. The dataset was in the form of tabular data from a relational database management system called SQLite. We had to export the tabular data from the original SQLite database dump file to CSV in order to load it on R (a free software tool for statistical computing). The columns containing dates were transformed from Julian to Gregorian. The rows containing no information(NA) were ignored and removed before the data analysis.

https://www.kaggle.com/econdata/climate-change
This dataset has been accredited to MITx Analytix, (https://www.edx.org/school/mitx) a free online educational platform supported by MIT (Massachusetts Institute of Technology, USA). The dataset contains almost 25 years of climate data starting from the year 1983. The values of total solar irradiance (TSI) was sourced from the SOLARIS-HEPA project website (http://solarisheppa.geomar.de/solarisheppa/cmip5) and MEI(Multivariate El Nino Southern Oscillation) was acquired from ESRL/NOAA Physical Sciences Division(http://www.esrl.noaa.gov/psd/enso/mei/table.html)

ANALYSIS

We start by analyzing the forest fires dataset. We have generated plots of varying intensity and affected areas both on a monthly and annual basis. The

1) **Frequency distribution per year:**

   This graph shows the number of forest fires occurring through the years 1992-2015. Annual distribution of fires peaked in the year 2006 at 114004 and the minimum was 61450 in the year 1993.

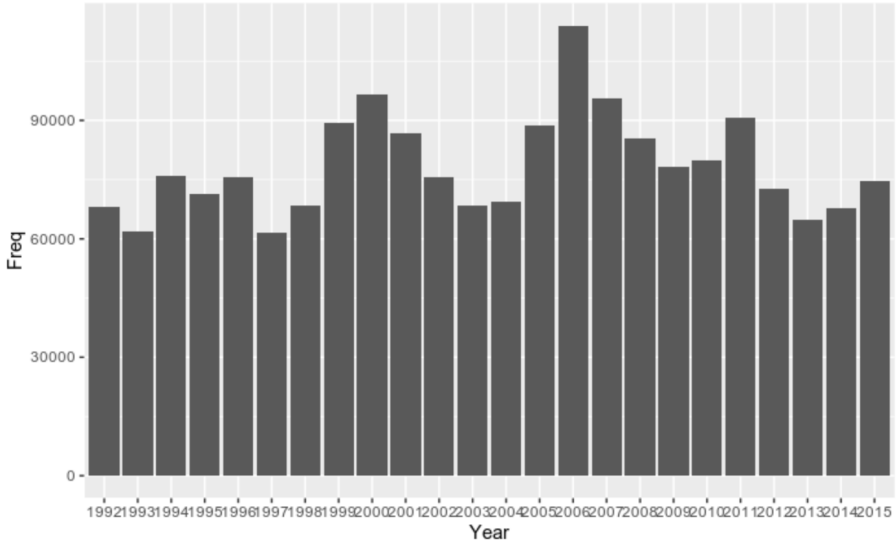


Fig: Frequency distribution per year

2) **Average Size per year:**

   This graph shows the average size (in acres) of the forest fires through the years 1992-2015. The maximum (average size of fires) was 137.17 acres and the minimum was 28.68 acres.

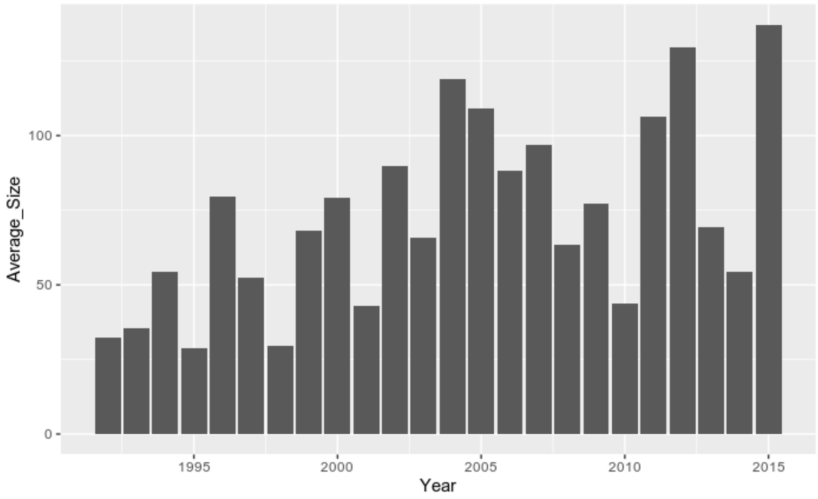


Fig: Average size per year

3) Count of Class per year:

This graph shows the number of forest fires per class (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres) through the years 1992-2015. The maximum number of class A fires were observed in 2006 and minimum were observed in 1993. The maximum number of class B fires were observed in 2006 and minimum were observed in 2013. The maximum number of class C fires were observed in 2006 and a minimum number was observed in 2013.

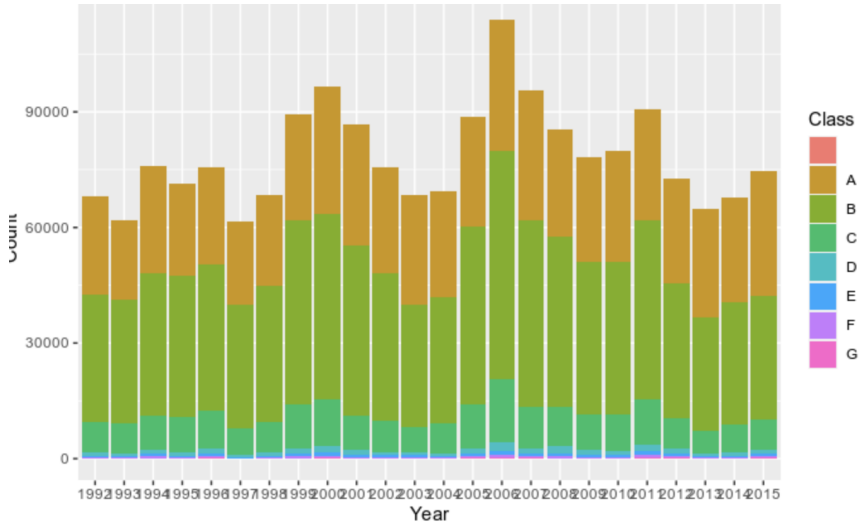


Fig: Count of class per year

4) Annual count by Cause:

This graph shows the count of the fires due to various causes through the years 1992-2015. The major conclusion from this visualization is that 88% of fires are caused by humans.

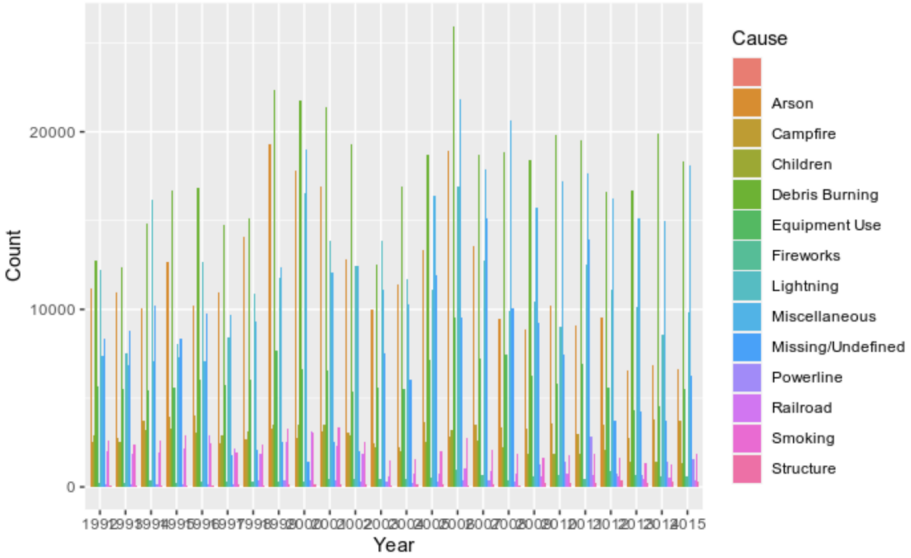


Fig: Annual Count by Cause per year

5) Frequency distribution per month:

This graph shows the number of forest fires occurring every month through the years 1992-2015

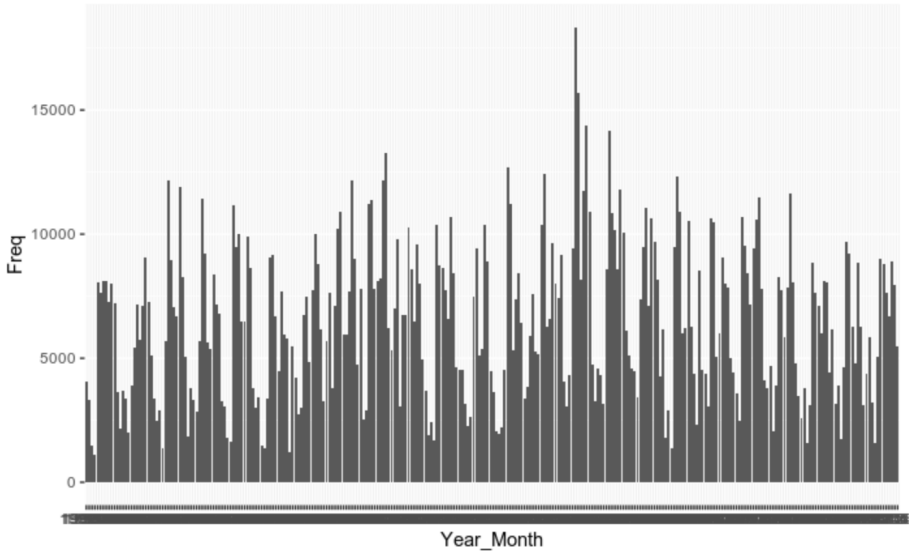


Fig: Frequency distribution per month

6) Average Size per month:

This graph shows the average size of the forest fires (acres) every month through the years 1992-2015

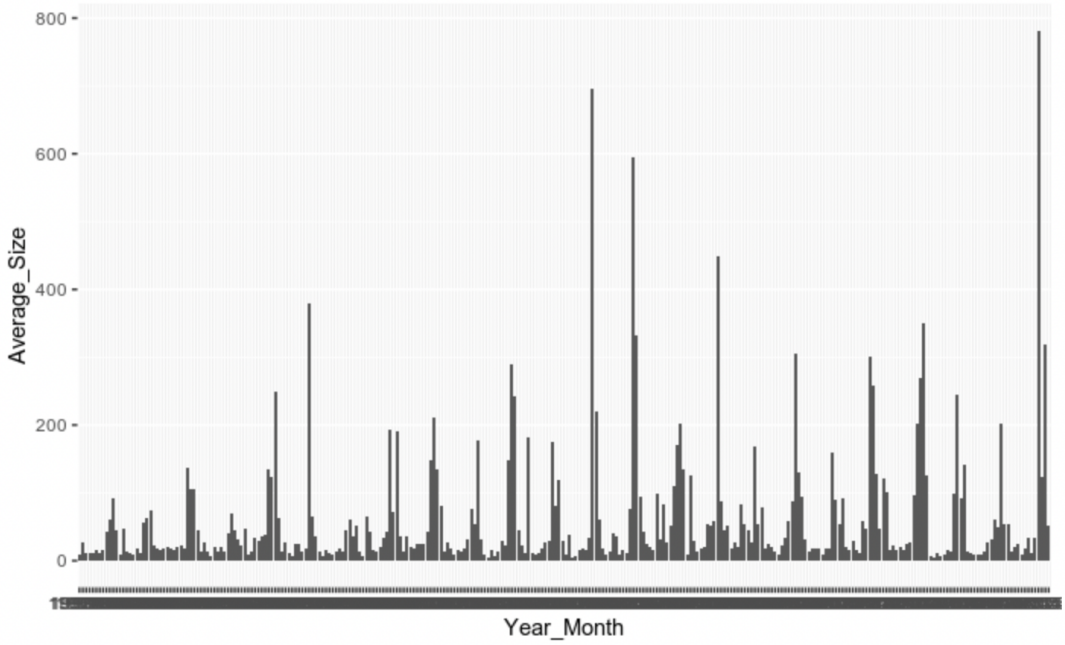


Fig: Average Size per month

7) Count of Class per month:

This graph shows the number of forest fires per class(A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres) every month through the years 1992-2015.

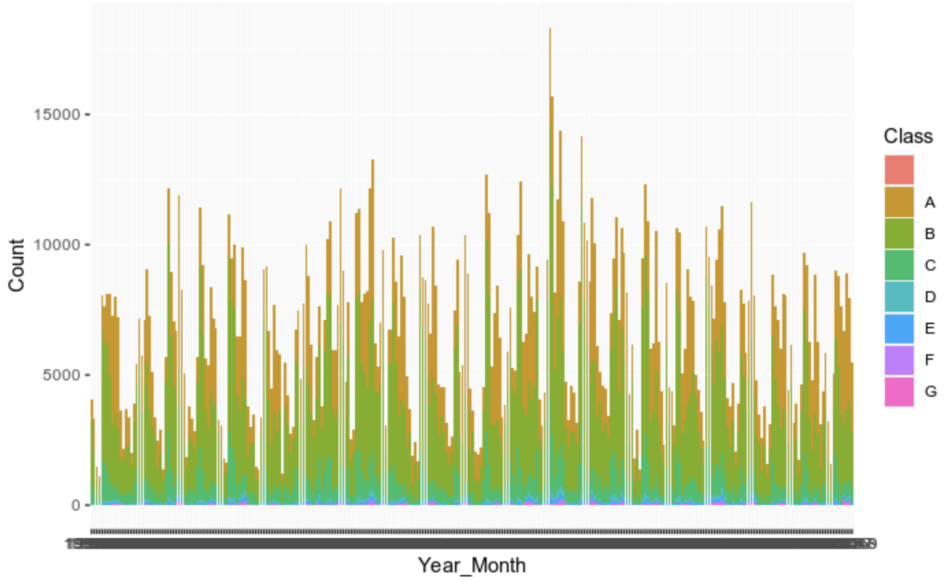


Fig: Count of class per month

8) Annual count by Cause:

This graph shows the count of the fires due to various causes every month through the years 1992-2015.

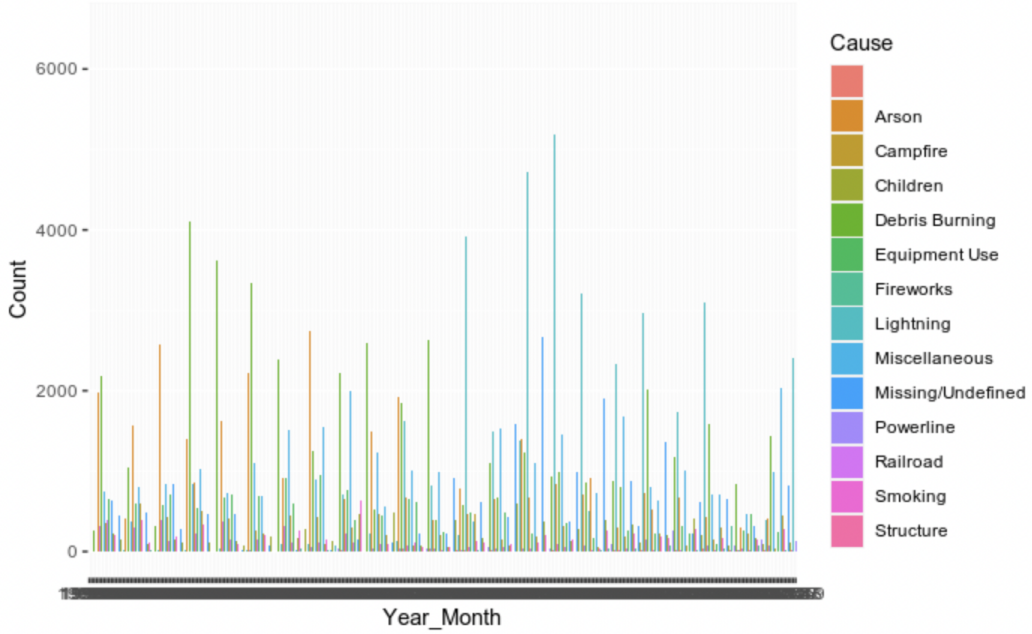


Fig: Annual count by Cause per month

Now, we look into our dataset for climate change:

1) **Temperature Frequency distribution per month:**

   This graph shows the change in temperature every month through the years 1983-2008.

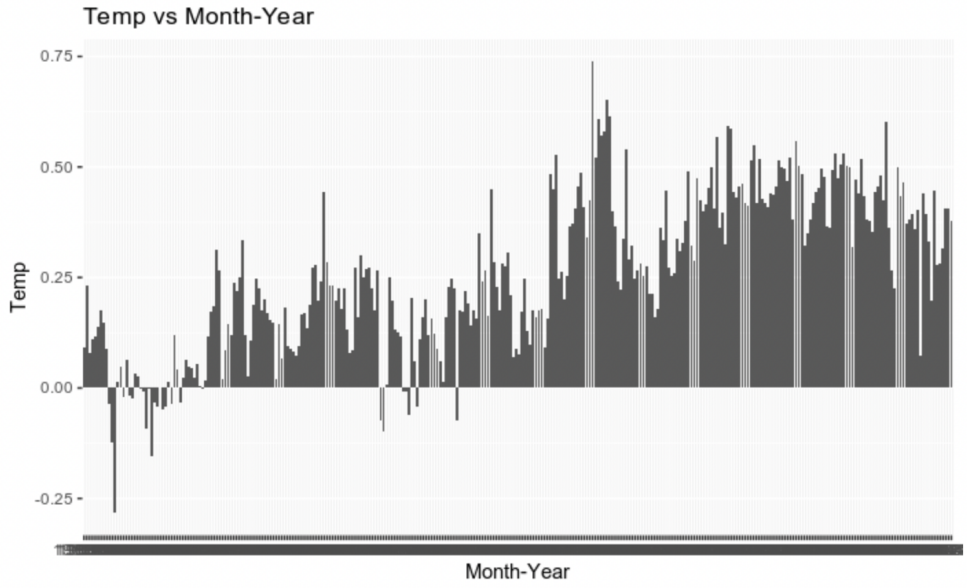


Fig: Temperature Frequency distribution per month

2) **MEI Frequency distribution per month:**

   This graph shows the change in multivariate El Nino Southern Oscillation index (MEI) every month through the years 1983-2008.

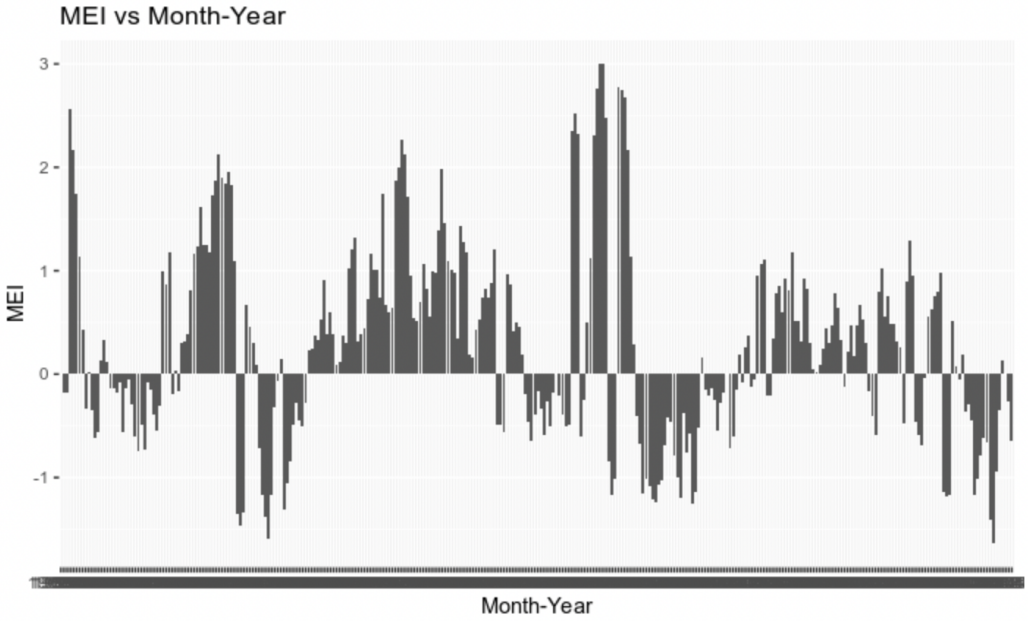


Fig: MEI Frequency distribution per month

3) CO2 Frequency distribution per month:

This graph shows the change in CO2 emissions every month through the years 1983-2008.

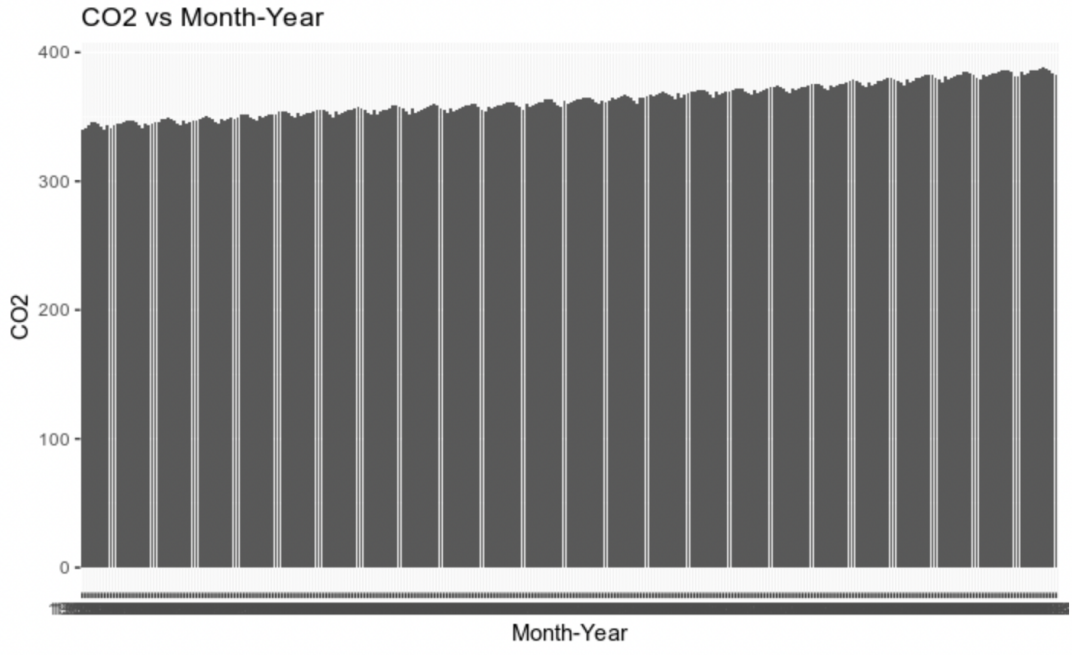


Fig: CO2 Frequency distribution per month

4) CH4 Frequency distribution per month:

This graph shows the change in CH4 emissions every month through the years 1983-2008.

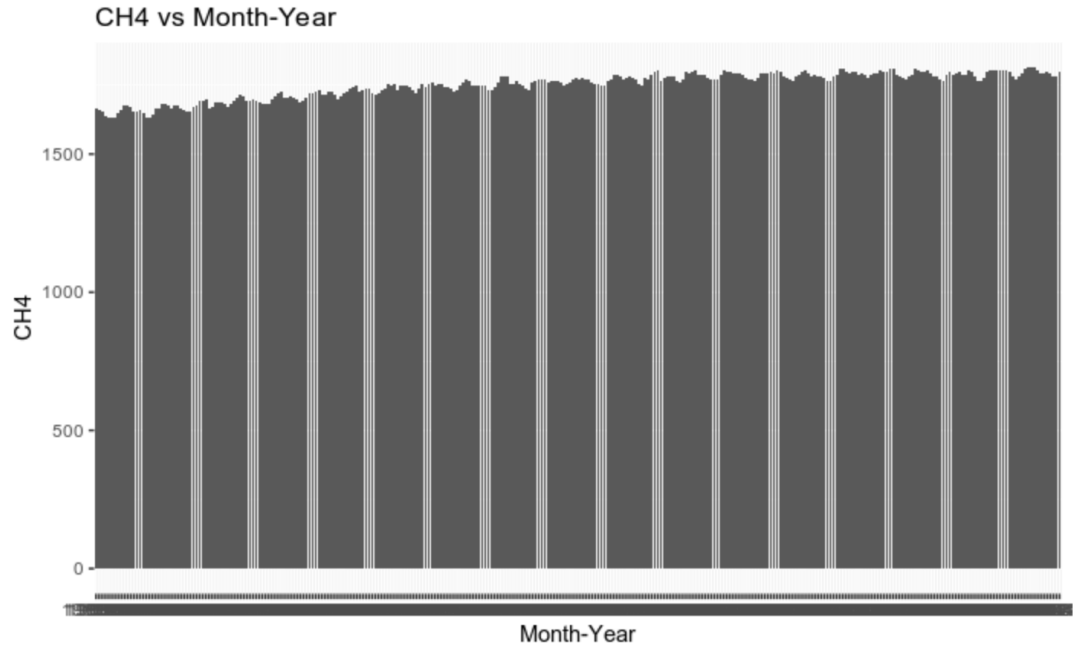


Fig: CH4 Frequency distribution per month

5) N2O Frequency distribution per month:

This graph shows the change in N2O emissions every month through the years 1983-2008.

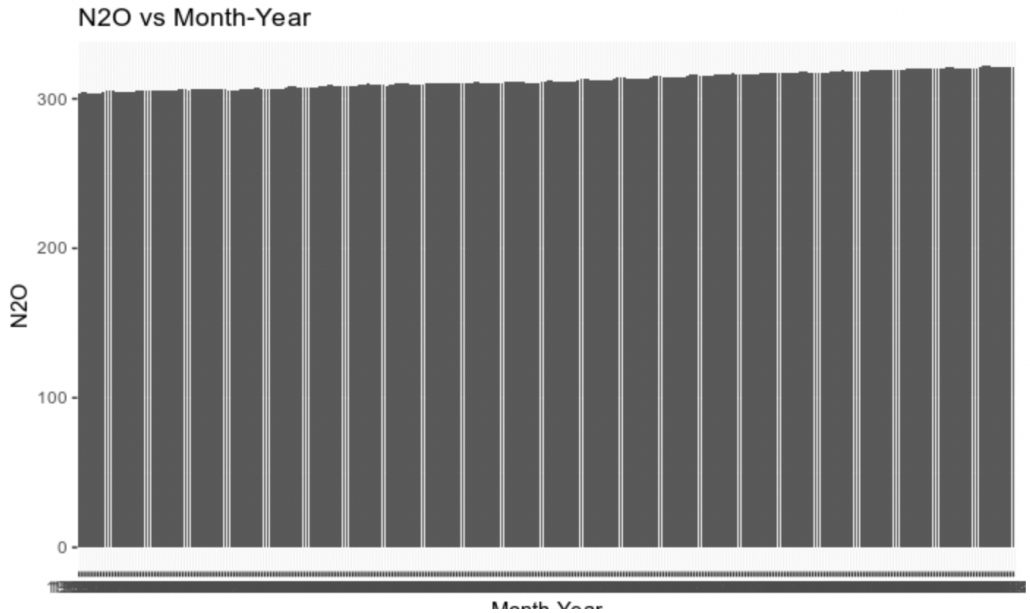


Fig: N2O Frequency distribution per month

6) CFC.11 Frequency distribution per month:

This graph shows the change in CFC.11(Trichlorofluoromethane) emissions every month through the years 1983-2008.

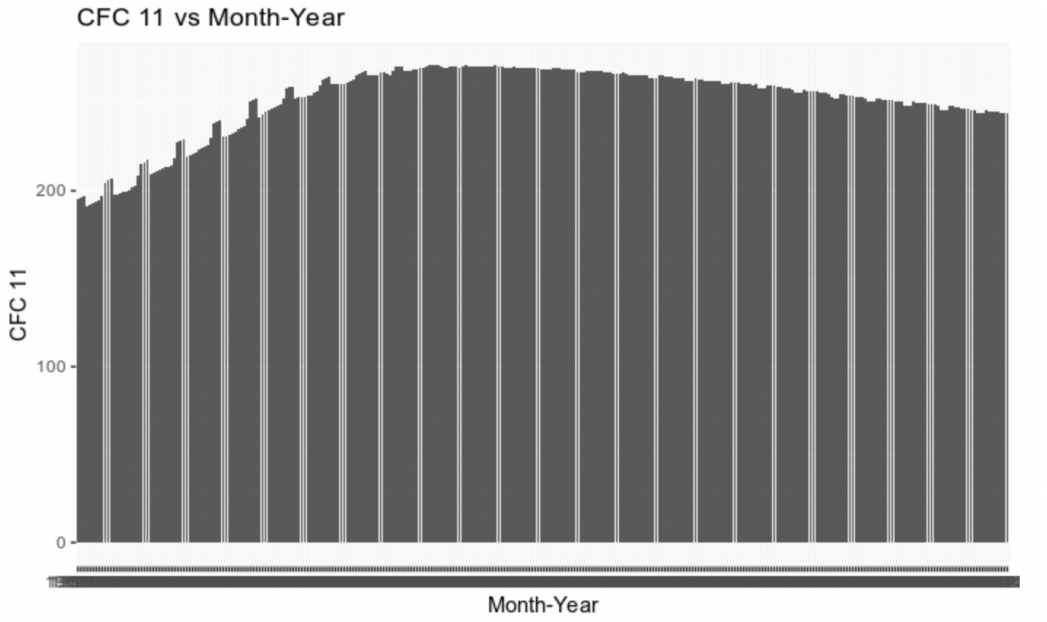


Fig: CFC.11 Frequency distribution per month

7) **CFC.12 Frequency distribution per month:**

This graph shows the change in CFC.12(Dichlorodifluoromethane) emissions every month through the years 1983-2008.

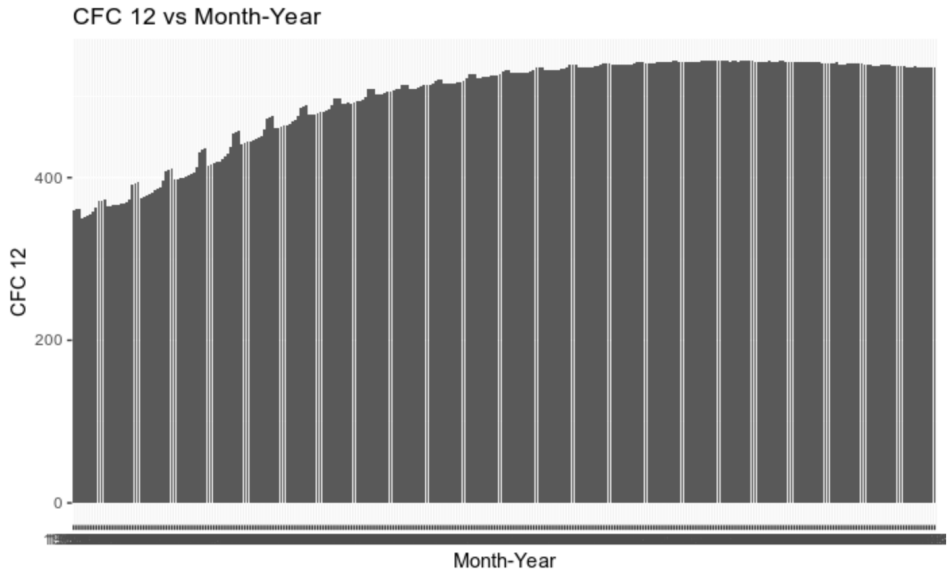


Fig: CFC 12 Frequency distribution per month

8) **TSI Frequency distribution per month:**

This graph shows the change in CFC.12(Dichlorodifluoromethane) emissions every month through the years 1983-2008.

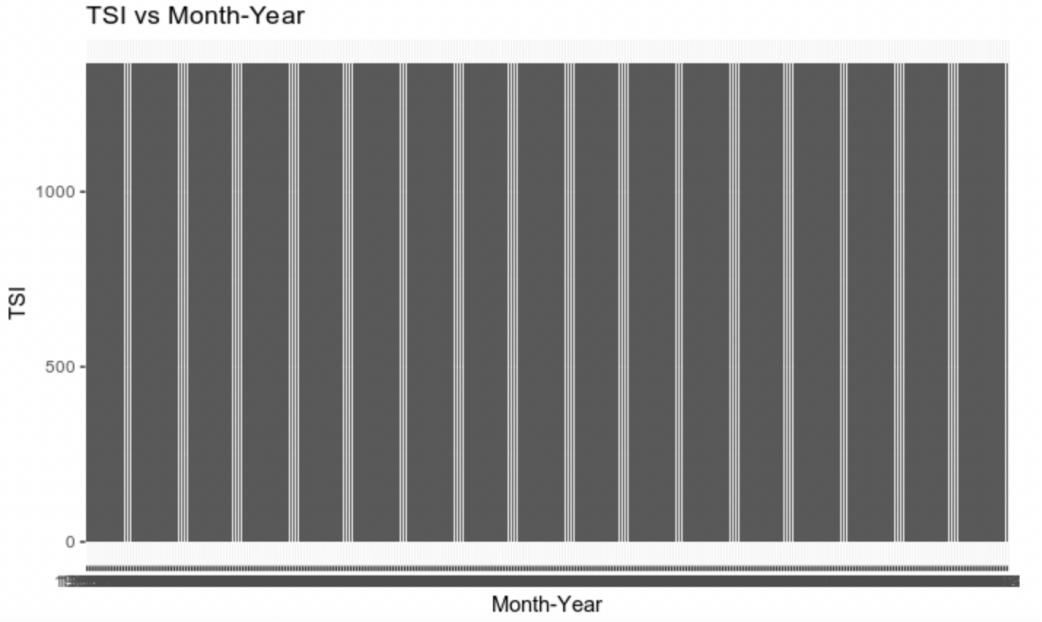


Fig: TSI Frequency distribution per month

9) **Aerosols Frequency distribution per month:**

This graph shows the change in CFC.12(Dichlorodifluoromethane) emissions every month through the years 1983-2008.

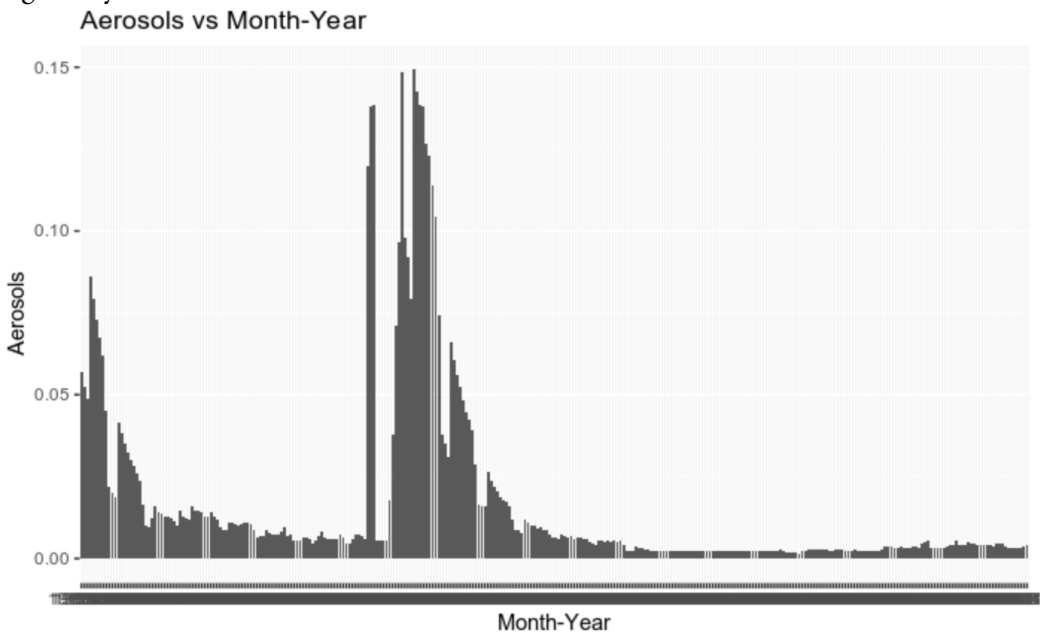


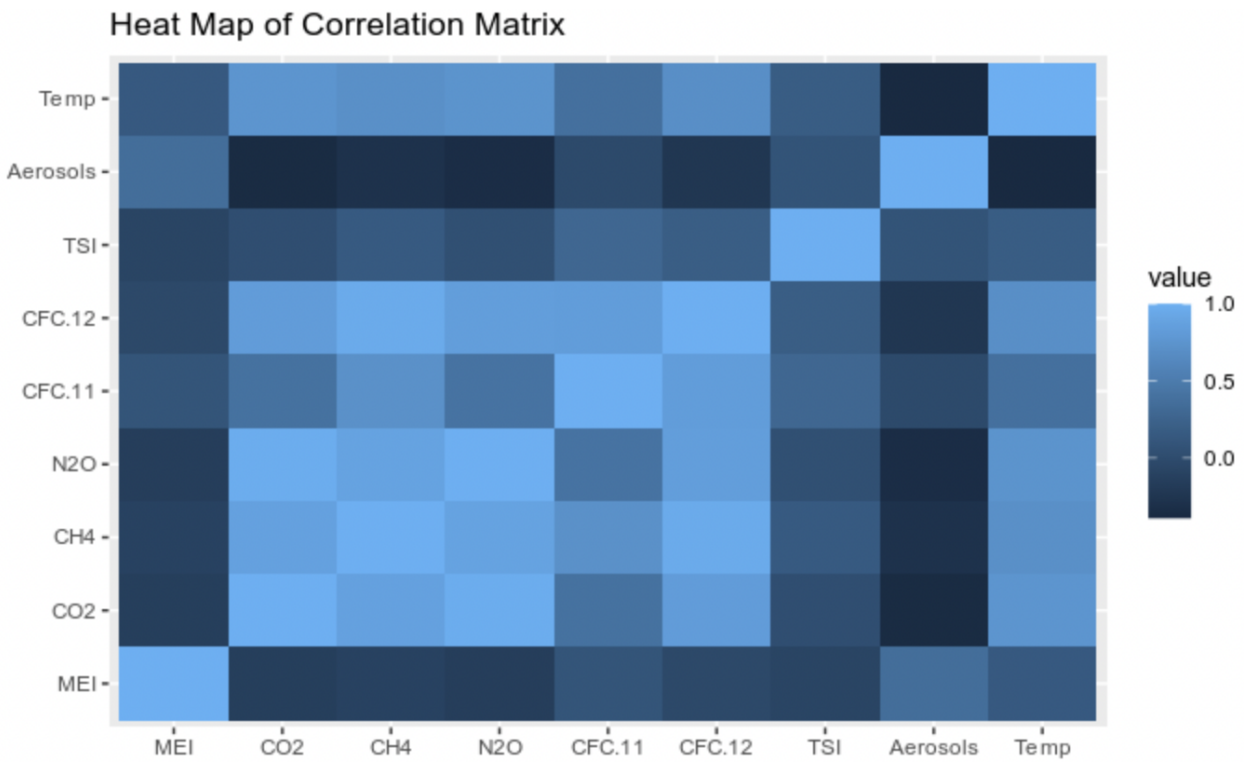
Fig: Aerosols Frequency distribution per month

Correlation matrix

We now take a look at the correlation matrix of the climate change dataset.

The correlation between two features x and y can be defined by the following formula:

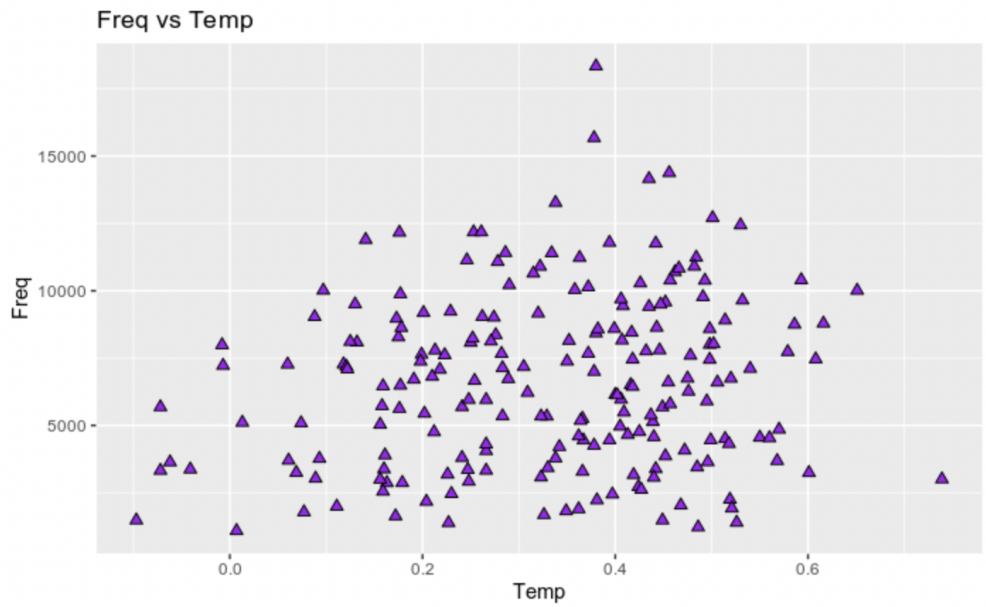
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



It can be observed from the above heat map that there is a positive correlation between temperature and the concentration of gases (like Carbon Dioxide(CO2), Methane(CH4), Dinitrogen Oxide (N2O) and CFC12 in the atmosphere. The values of TSI (Total Solar Irradiation Index i.e. the total solar energy radiated by the sun per unit area) and MEI (Multivariate El-Nino Southern Oscillation Index i.e. a measure of the strength of El-Nino, a weather effect in the Pacific Ocean which affects global temperatures) however do not show a positive correlation with temperature.

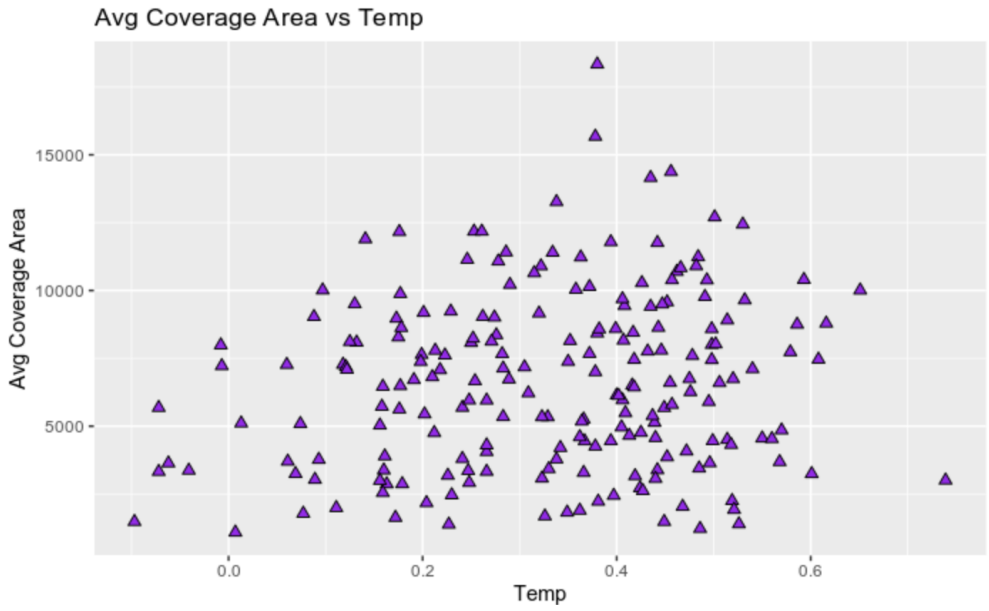
Scatterplots

1) **Frequency vs Temperature Difference(between average global temperature and a reference value)**



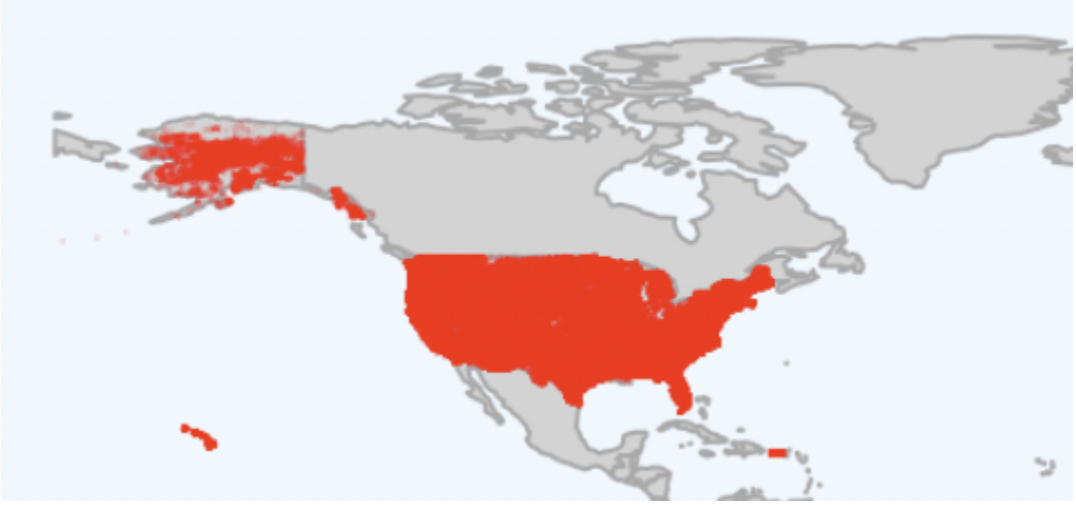
It can be observed that the frequency of forest fires does increase with the increase in the temperature difference.

2)Average Area of Coverage vs Temperature(between average global temperature and a reference value)



It can be observed that the average area of coverage of forest fires also increases with the increase in the global average temperature difference.

Geographical Plot - Plotting fires locations on a map



We used the geospatial information in the forest fires dataset to visualize the concentration of the fires in the country. It can be observed from the above plot that forest fires are distributed evenly throughout the country (USA) including the regions of Hawaii and Alaska.

ALGORITHMS BASED DATA ANALYSIS

We are using regression analysis to predict the frequency of forest fires and average size of forest fires per month. We used two algorithms, linear regression and random forest for this purpose.

The rule of thumb in regression is that there must be a minimum 10 rows for each independent variable, i.e., in this case we need at least 40 observations for reasonably good prediction results. Our dataset contains 204 observations. Hence we could use this data for prediction.

First, to proceed with regression analysis, we narrowed down upon the important independent variables of the climate that actually affect the frequency and average size of forest fires. By observing the correlation between the independent variables we can achieve this.

```
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
[1,]  1.00  0.10  0.60  0.53  0.60 -0.53
[2,]  0.10  1.00 -0.32 -0.40 -0.35  0.29
[3,]  0.60 -0.32  1.00  0.73  0.97 -0.96
[4,]  0.53 -0.40  0.73  1.00  0.77 -0.68
[5,]  0.60 -0.35  0.97  0.77  1.00 -0.98
[6,] -0.53  0.29 -0.96 -0.68 -0.98  1.00
```

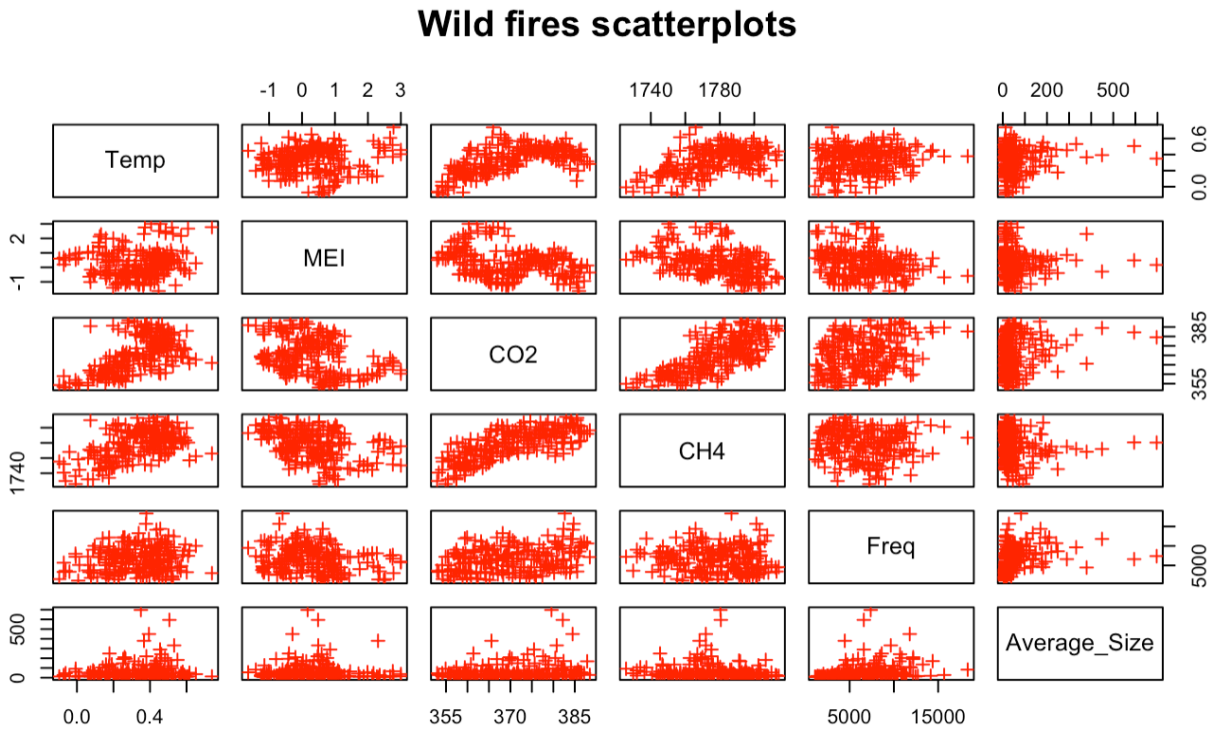
We included 6 independent variables here, temperature, CO2, MEI, CH4, N2O and CFC11, since historically, they have been important climatic factors to global warming and could perhaps affect the frequency and intensity of wildfires.

If the correlation is >0.8 for any two variables, only one among them is enough to include in linear regression. Since CO2, N2O and CFC.11 are strongly related to each other, we can include only one among them. Let's consider only CO2 Hence we are using multiple linear regression on **Temp, CO2, MEI and CH4**

Linear regression

Predicting forest fires per month based on climatic conditions.

Firstly, scatterplots should be produced for each independent variable with the dependent variable to see if the relationship is linear



There are mostly no non-linear patterns between any pair of variables, (some do exist and we can see it's effect in the summary of our model).

```
Call:
lm(formula = Freq ~ Temp + MEI + CO2 + CH4, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4346.6 -1979.5  -273.7  1556.0  9093.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 144706.00   23796.96   6.081 6.01e-09 ***
Temp          4068.89    1675.42   2.429    0.016 *
MEI          -1061.09     240.35  -4.415 1.66e-05 ***
CO2            208.45      32.20   6.474 7.30e-10 ***
CH4           -121.67      15.31  -7.945 1.41e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2683 on 199 degrees of freedom
Multiple R-squared:  0.316, Adjusted R-squared:  0.3022
F-statistic: 22.98 on 4 and 199 DF,  p-value: 1.258e-15
```

The last column contains the p-values for each of the independent variables. A p-value < 0.05, provides evidence that the coefficient is different to 0(*** = highly significant). We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between Freq and independent variables exist. From the summary we can say that “MEI, CO2 & CH4” are all significant in predicting the freq of wildfires per month

The Estimate column in the coefficients table, gives us the coefficients for each independent variable in the regression model. Our model is Freq(y) = 144706.00 + 4068.89(Temp) - 1061.09(MEI) + 208.45(CO2) - 121.67(CH4)

The Multiple R-squared value generally increases with the increase in number of independent variables. Hence it is better to use the adjusted R squared for an understanding of the model. The adjusted R2 indicates that a 30.22% increase in the frequency of fires per month can be explained by the model containing Temp, MEI, CO2 and CH4.l

Since 88% of fires are caused by humans(observation from annual count by cause visualization), this 30.22% variation is quite high hence predictions from the regression equation are fairly reliable.

Hence we can state that climatic conditions do play a significant role in freq of wildfires

Predicting average size of fire per month based on climatic conditions

```
Call:
lm(formula = Average_Size ~ Temp + MEI + CO2 + CH4, data = train_data)

Residuals:
   Min     1Q Median     3Q    Max
-86.85 -39.25 -18.48  11.07 591.35

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3083.8709   706.3280   4.366 2.03e-05 ***
Temp          45.7955    49.7288   0.921    0.358
MEI           -7.0846     7.1339  -0.993    0.322
CO2            5.6544     0.9557   5.916 1.42e-08 ***
CH4           -2.8870     0.4545  -6.352 1.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.64 on 199 degrees of freedom
Multiple R-squared:  0.2127,    Adjusted R-squared:  0.1969
F-statistic: 13.44 on 4 and 199 DF,  p-value: 1.023e-09
```

From the summary we can say that “CO2 & CH4” are significant in predicting the average size of wildfires per month.
The Estimate column in the coefficients table, gives us the coefficients for each independent variable in the regression model. Our model is Freq(y) = 3083.8709 + 45.7955 (Temp) - 7.0846 (MEI) + 5.6544 (CO2) - -2.8870 (CH4)
The adjusted R2 indicates that 19.69% increase in the average size of wildfires per month can be explained by the model containing Temp, MEI, CO2 and CH4. Since this is not a significant increase we can conclude that climatic conditions do not play a significant role in the variation of average size of the wild fires

Random Forest regression

Predicting forest fires per month based on climatic conditions.

Random forest regression predicts output by average by bootstrap algorithm or bagging of independently built decision trees.

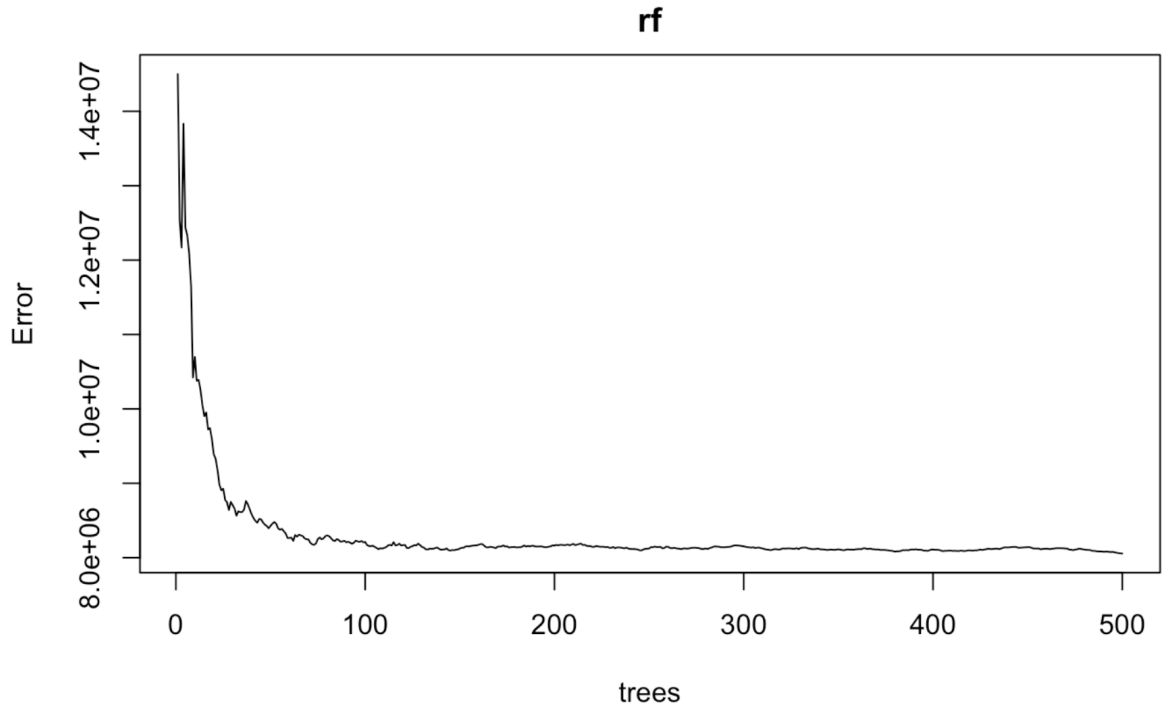
There are a lot of combinations possible between the parameters. Without trying all of them we can use Grid Search in R. With grid search the model will be evaluated over all the combinations you pass in the function, using cross-validation.

We divided the data into 75% - 25% train and test data. Our model results are

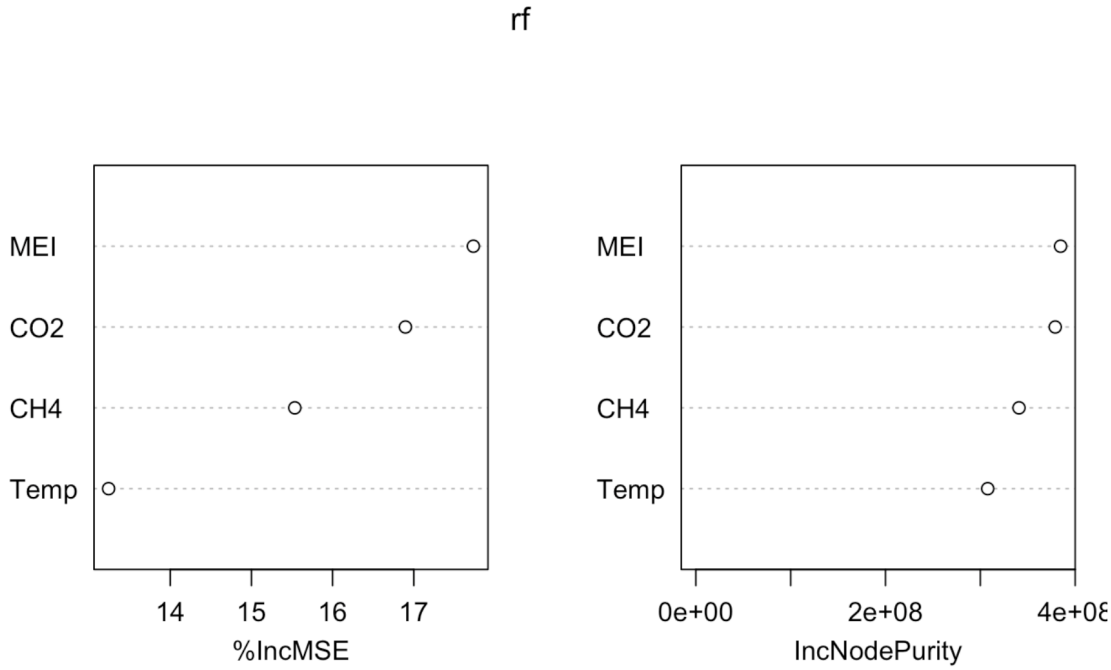
```
Call:
 randomForest(formula = train$Freq ~ ., data = train, importance = T,      trControl = trControl)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 8055783
                    % Var explained: 22.35
```

The Error vs Trees plots of the model is



The important independent variables that affects the frequency of forest fires can be found by the below plots



The x-axis displays the average increase in node purity and increase in mean squared error of the regression trees based on splitting on the various predictors displayed on the y-axis.

We can see MEI is the most important predictor variable and Temp is the least.

The predictions using the test data are

	actual <int>	predicted <dbl>
1	4055	5978.750
10	7987	5895.069
13	2175	6654.761
15	3370	4243.969
16	1991	5841.415
37	1823	6504.811
38	3801	5315.033
45	5347	8103.918
46	8360	8122.341
55	10014	5268.706

1-10 of 51 rows Previous 1 2 3 4 5 6 Next

This model has a mean squared error of :

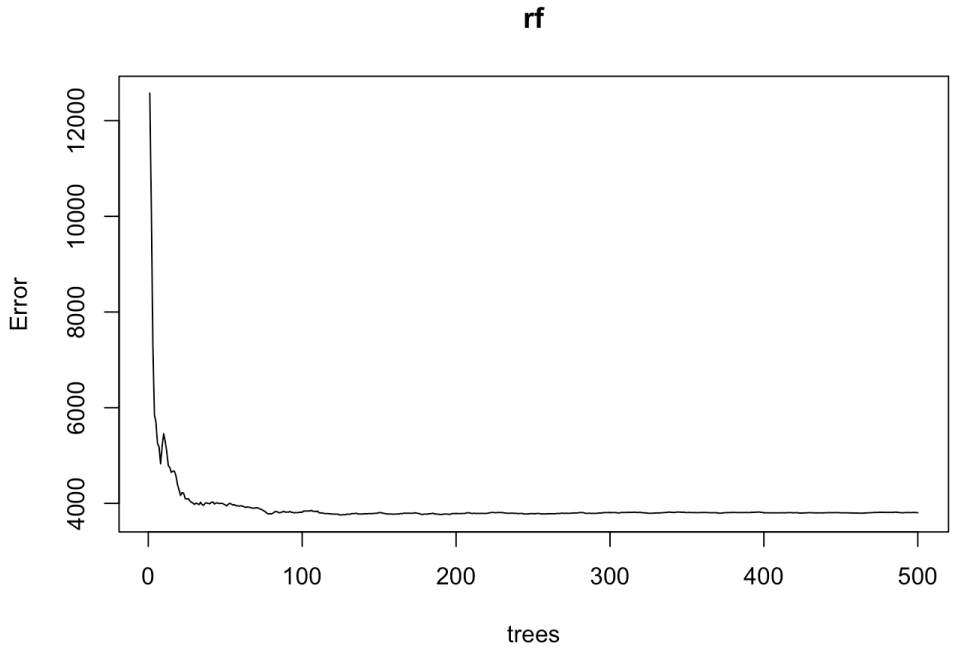
```
## [1] 9450565
```

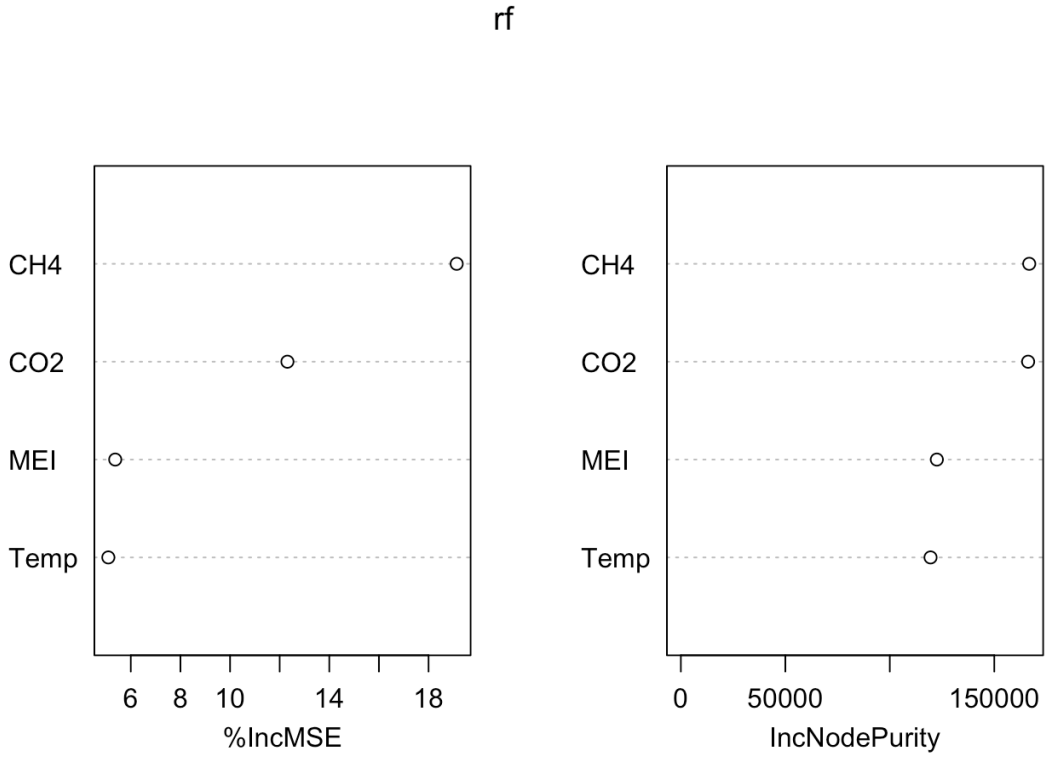
Predicting average size of fire per month based on climatic conditions

Our model results are

```
##
## Call:
##  randomForest(formula = train$Average_Size ~ ., data = train,      importance = T, trControl = trControl)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 3805.199
##                     % Var explained: 13.8
```

The Error vs Trees plots of the model is





Here CH4 is most important and MEI is least important Predicting using test data

The predictions using the test data are

	actual <dbl>	predicted <dbl>
5	12.450995	20.26801
9	41.639427	21.66631
14	46.508833	26.09520
15	14.150368	39.33812
20	56.757407	23.35782
21	63.807381	26.01755
23	21.937584	58.55043
26	17.228388	23.83418
27	21.110917	22.09429
33	137.611446	46.45290

This model has a mean squared error of

```
## [1] 16806.63
```

RESULTS AND FUTURE WORK

The main goal of the project was to test the hypothesis that climatic conditions affect the frequency of forest fires and average size of the forest fires.
We have used two models to test the hypothesis, and both the models results agree that: **Climatic conditions affect the frequency of forest fires whereas they do not affect the average size of it.**

The results of the comparative study of two regression models are,

- From the summary of linear regression and from the variable importance plots of random forest, the results from both the models agree that “Climatic conditions affect the frequency of forest fires whereas they do not affect the average size of it”.
- For predictive results of the models, Linear regression is more accurate than random forest in our case.

REFERENCES

[1] Climate Change, https://www.kaggle.com/econdata/climate-change

[2] Forest Fires, https://www.kaggle.com/rtatman/188-million-us-wildfires