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

RUTGERS

THE STATE UNIVERSITY OF NEW JERSEY

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

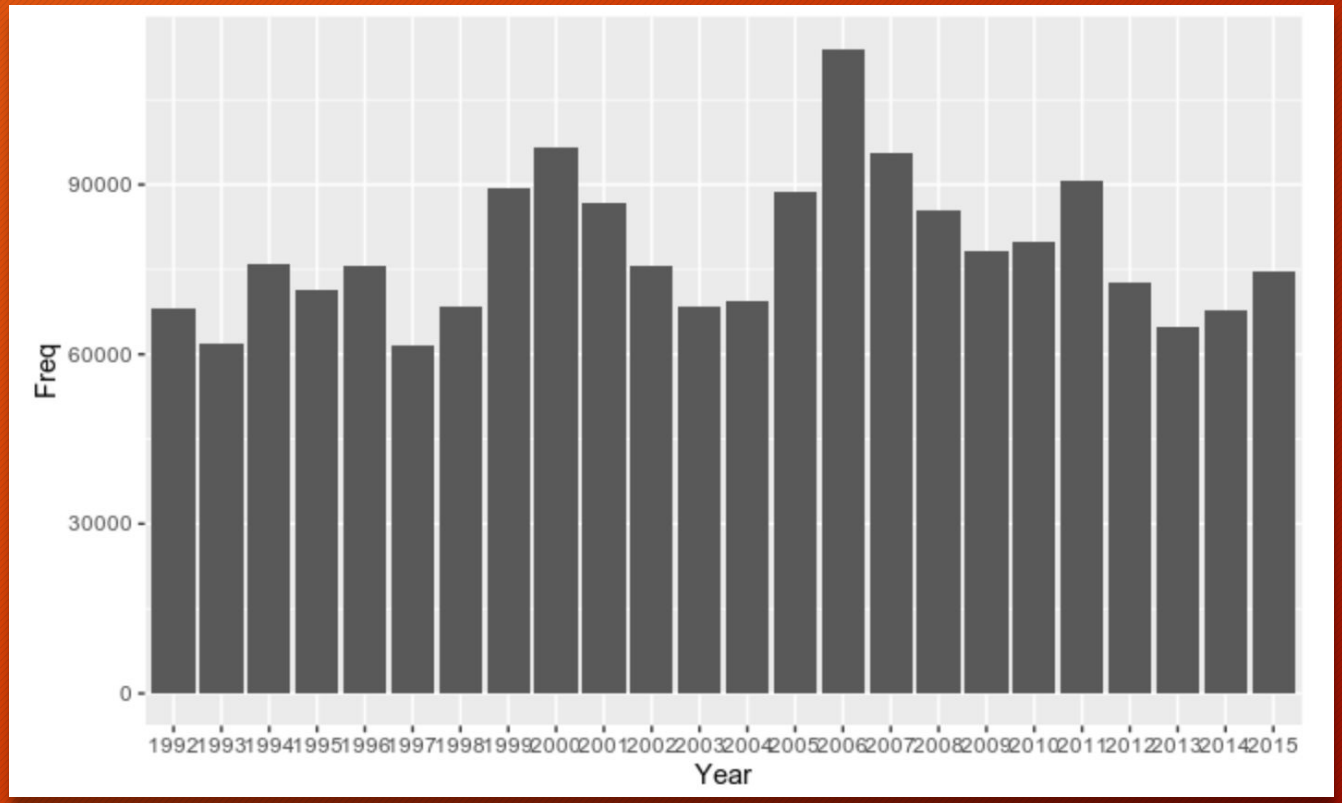
Abstract:

- 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 climate. It could also be a product of human activities like the 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.
- 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.

Forest Fires Dataset Analysis

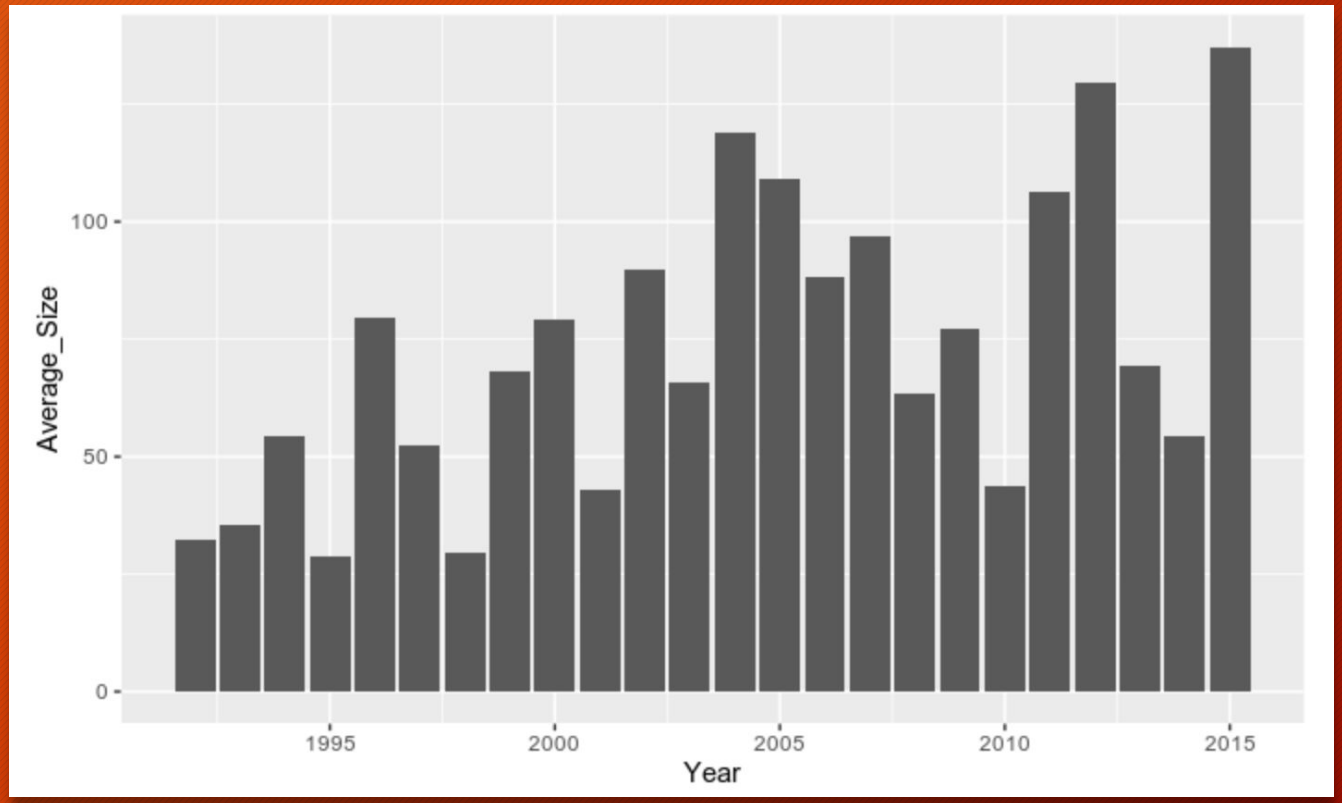
1) **Frequency distribution per year:**

- 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.



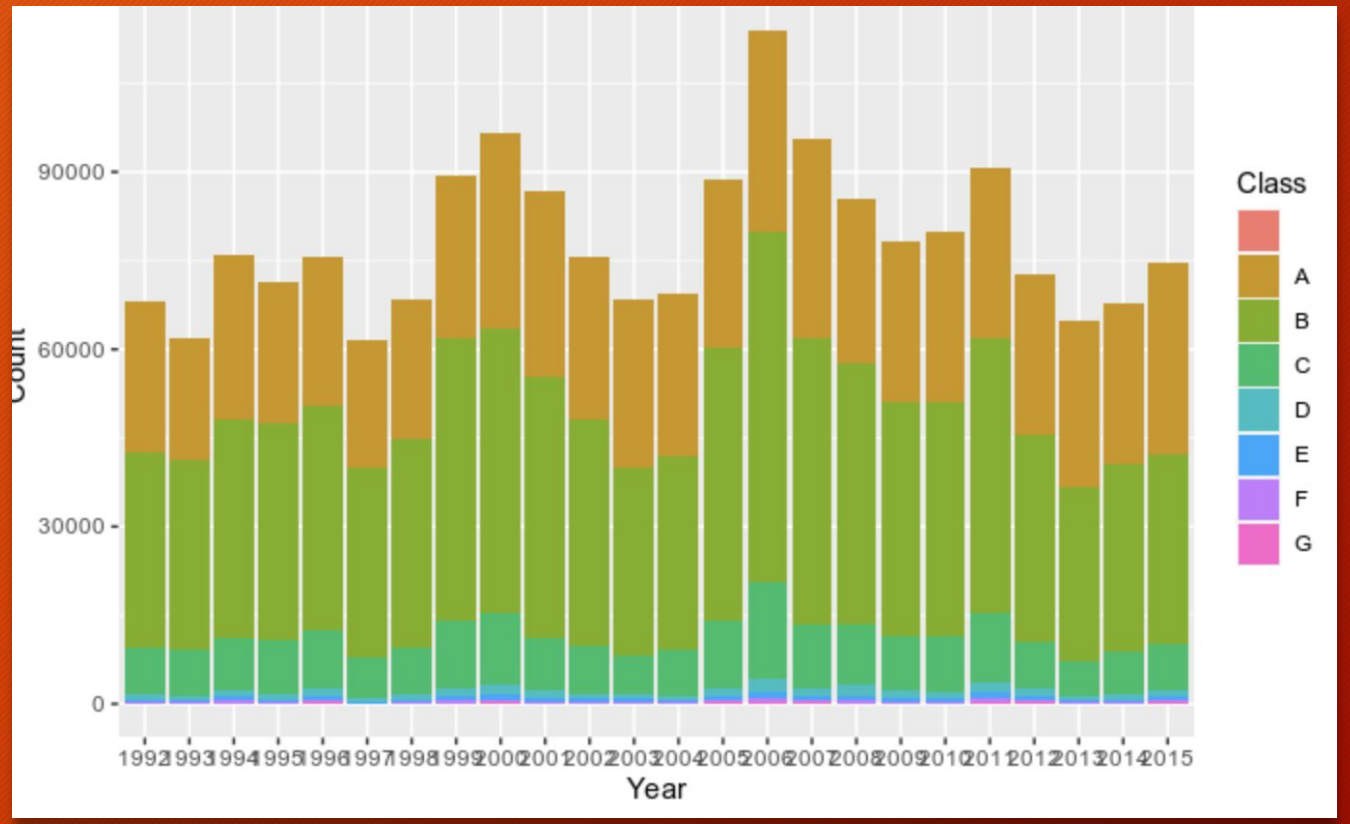
2) Average Size per year:

- 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.



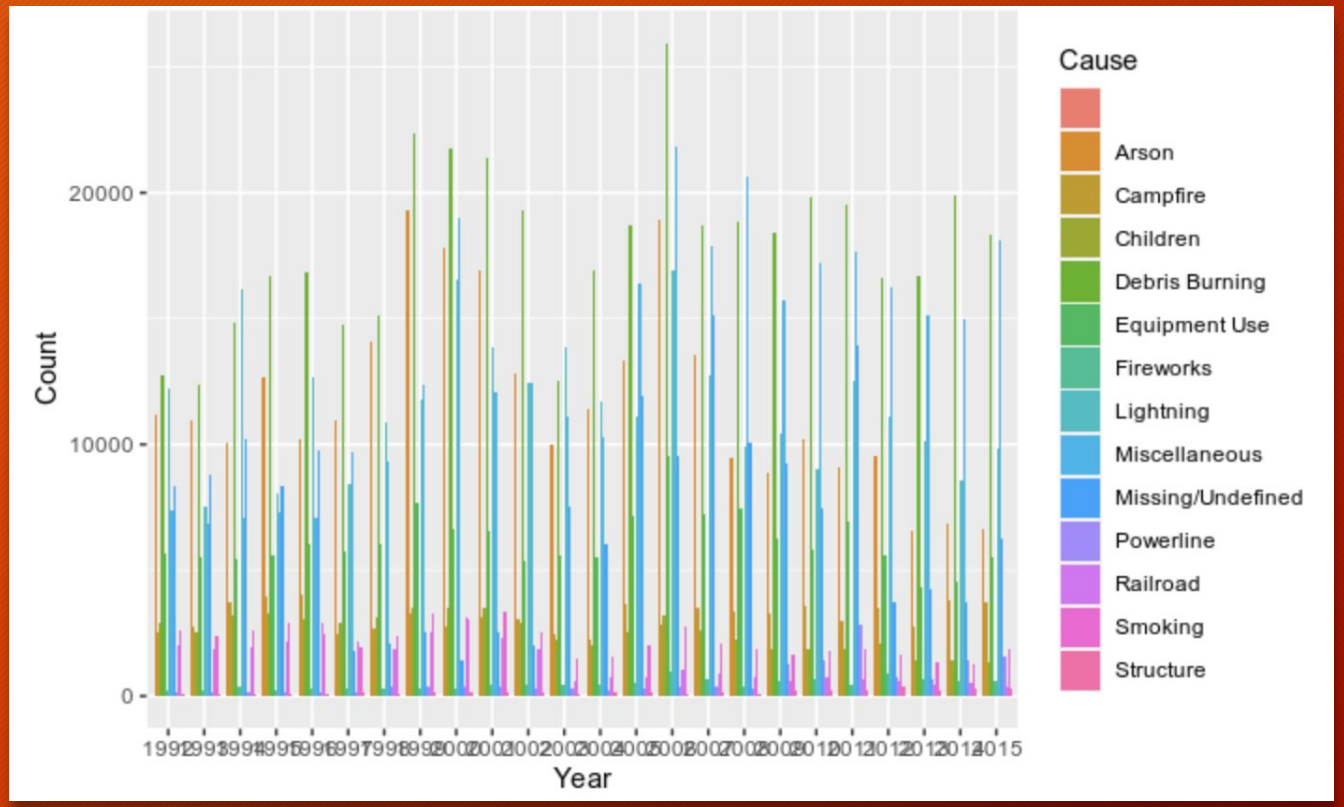
3) Count of Class per year:

- 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.



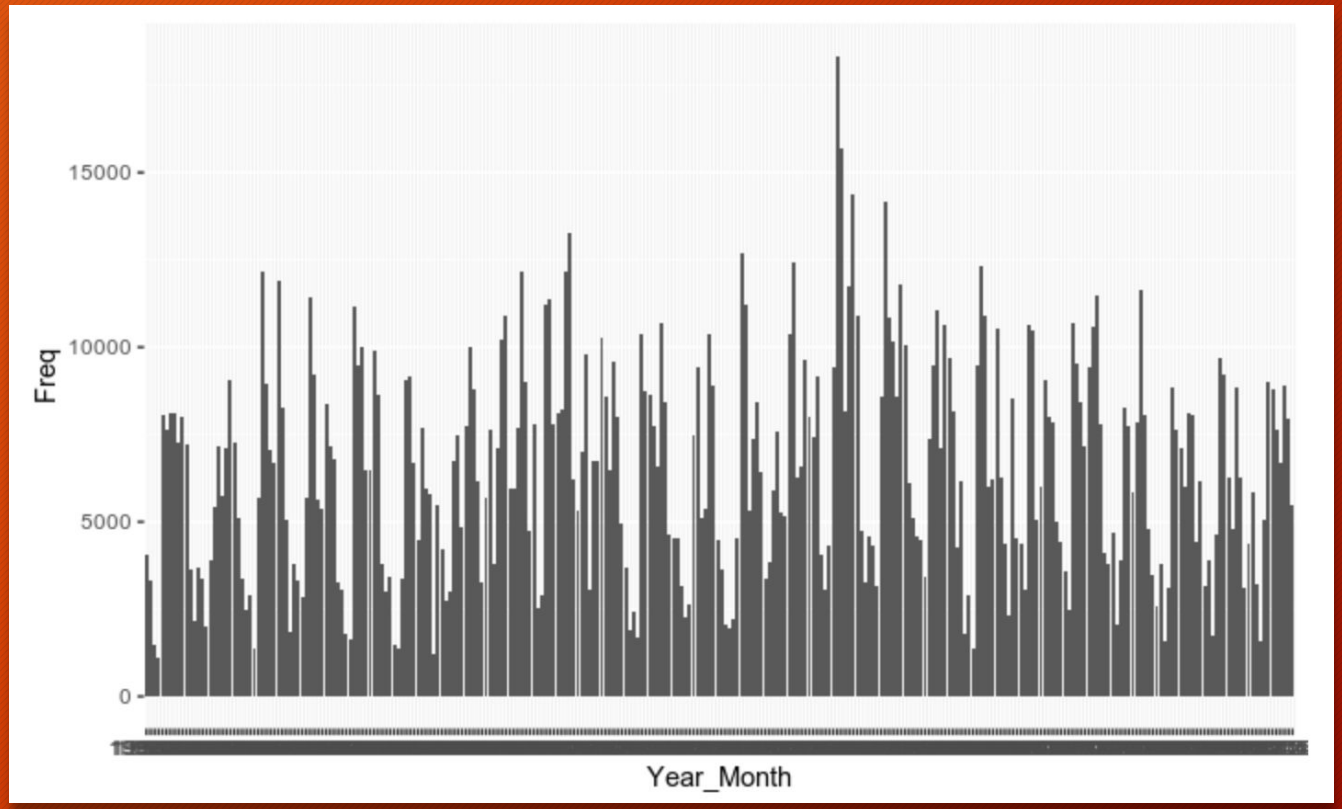
4) Annual count by Cause:

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



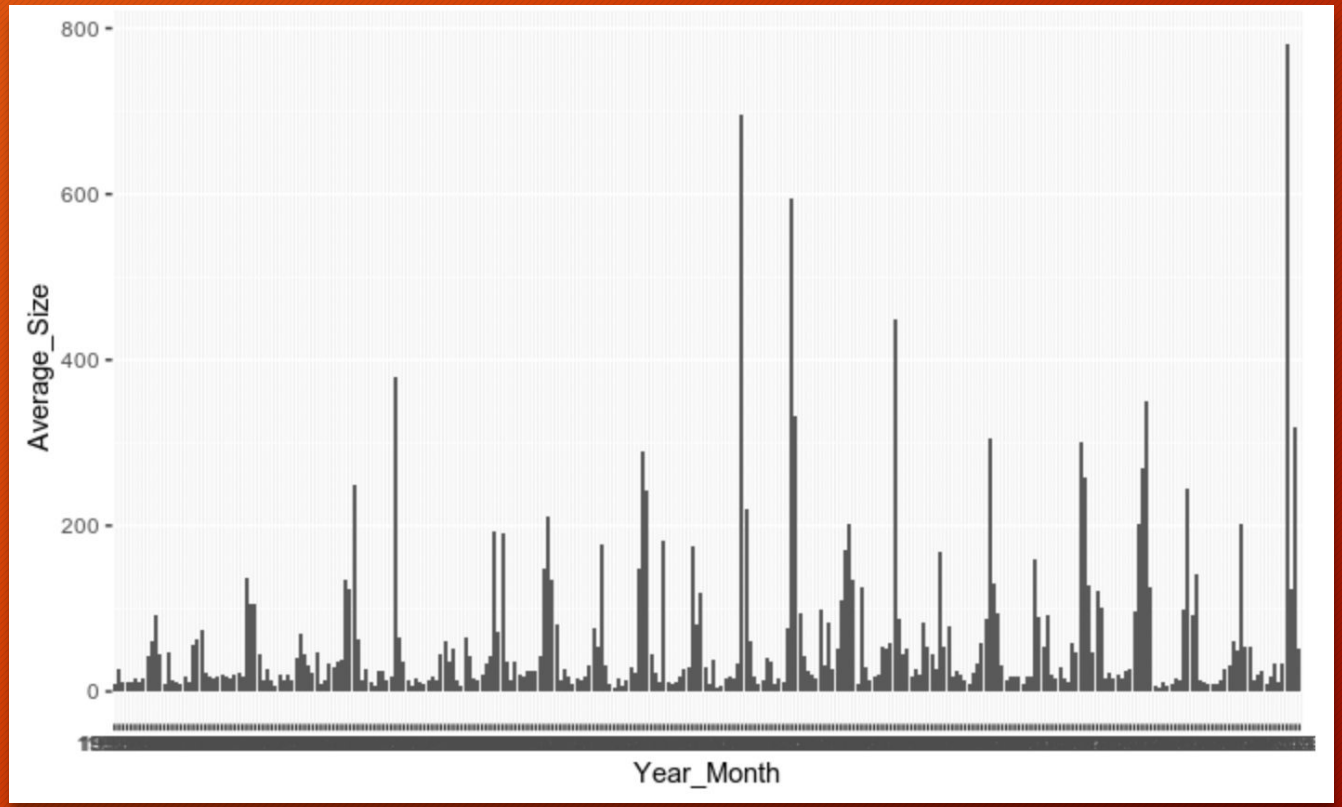
5) **Frequency distribution per month:**

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



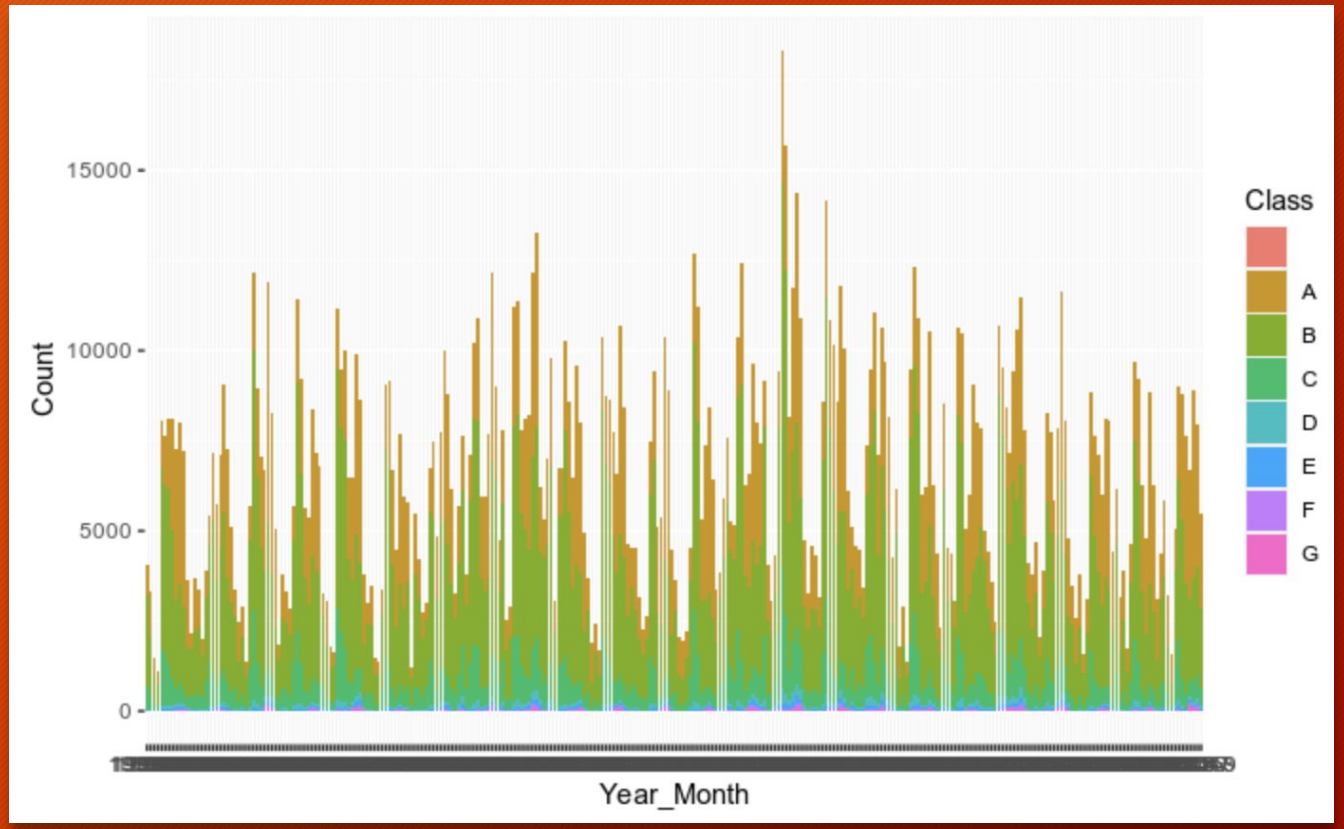
6) Average Size per month:

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



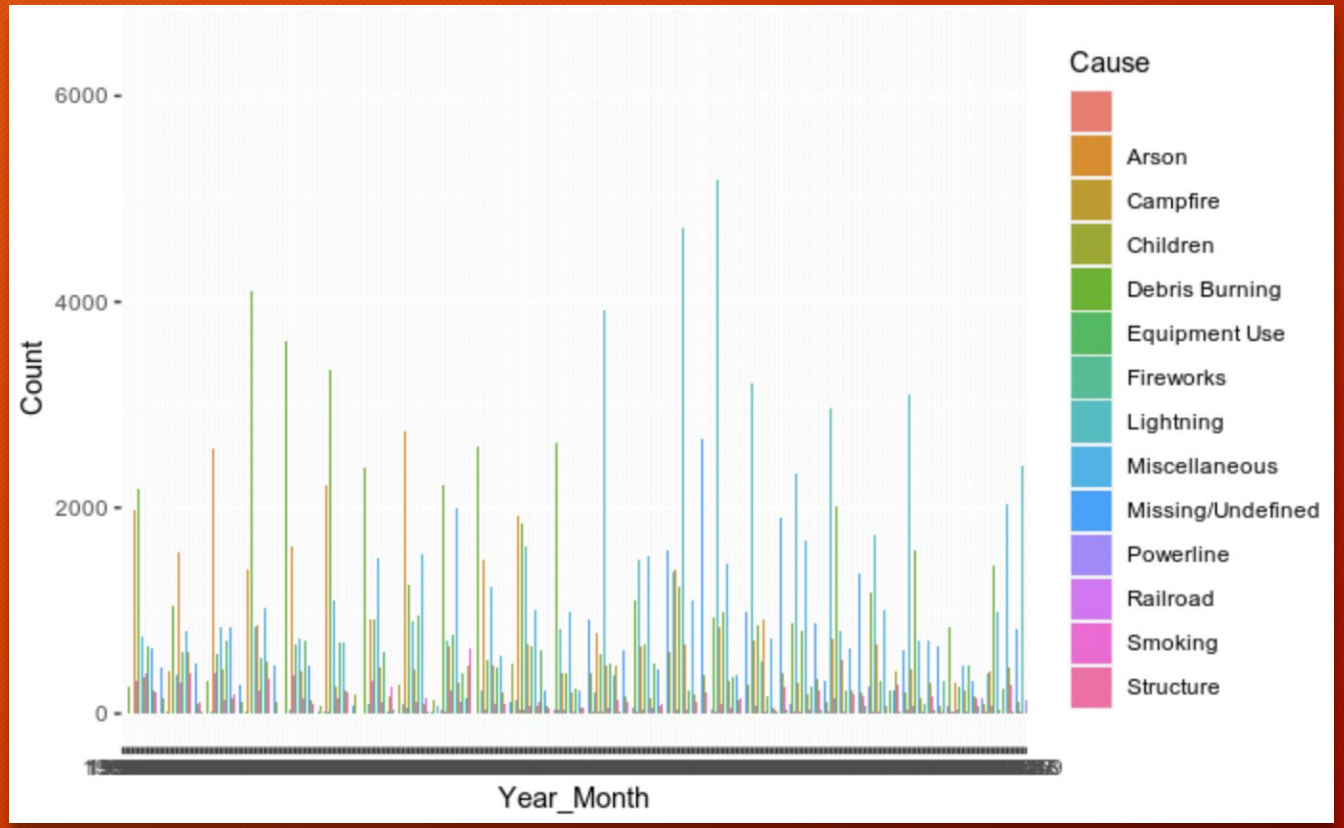
7) **Count of Class per month:**

- 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.



8) Annual count by Cause:

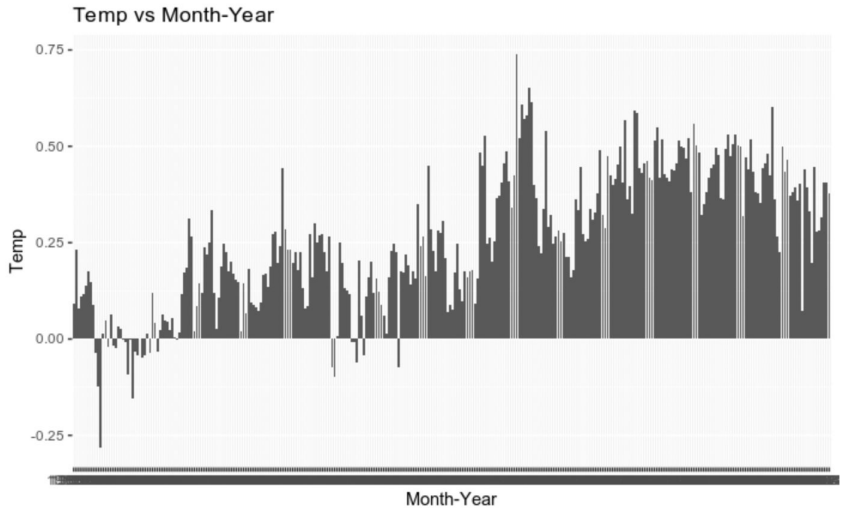
- This graph shows the count of the fires due to various causes every month through the years1992-2015.



Climate Change Dataset Analysis

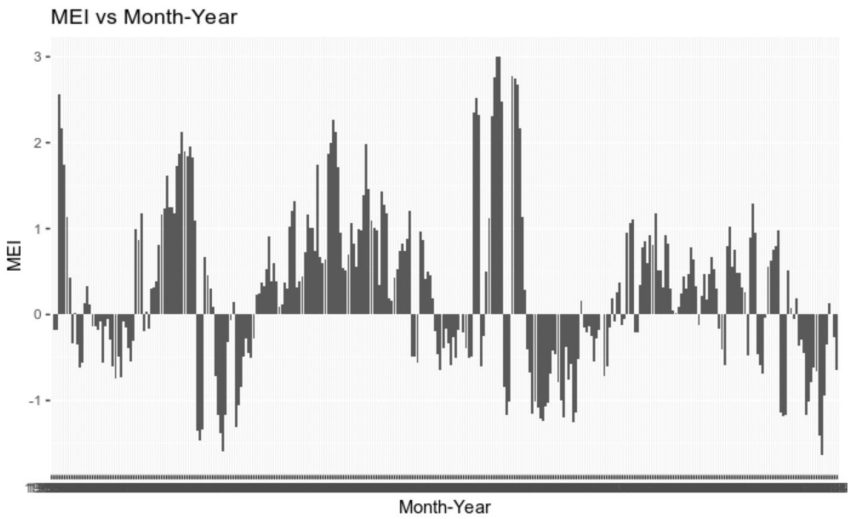
1) **Temperature Frequency distribution per month:**

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



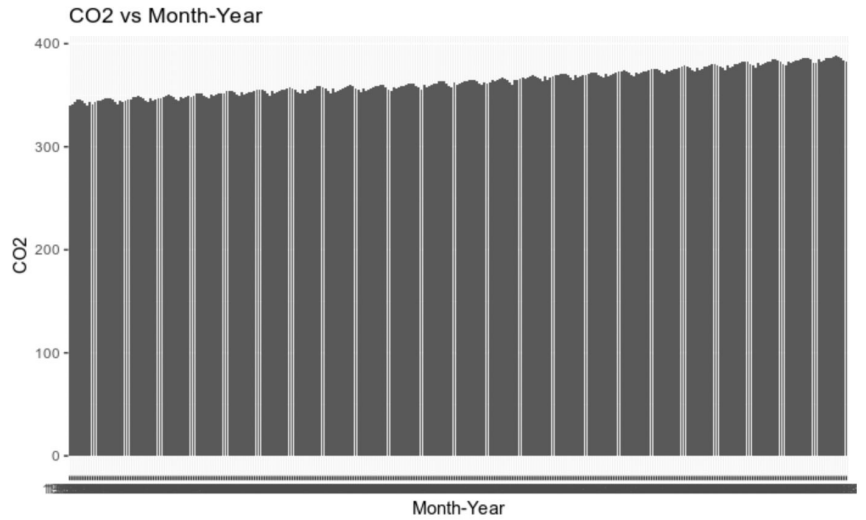
2) **MEI Frequency distribution per month:**

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



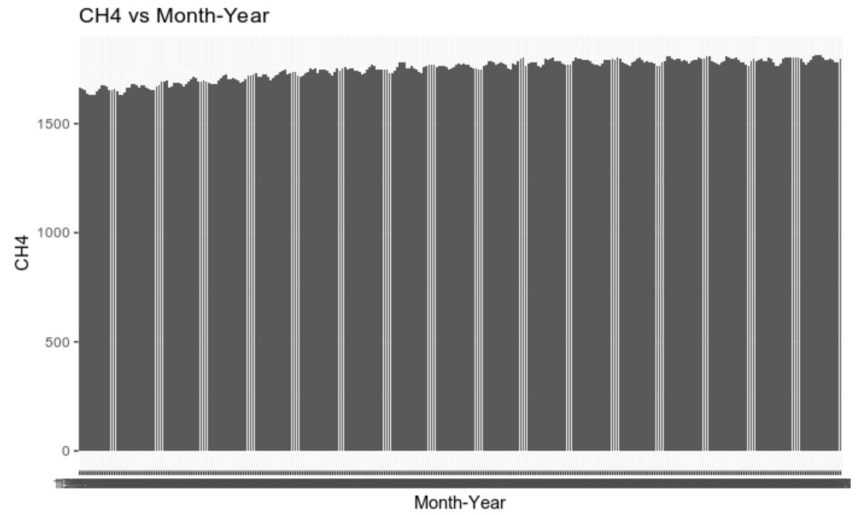
3) CO2 Frequency distribution per month:

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



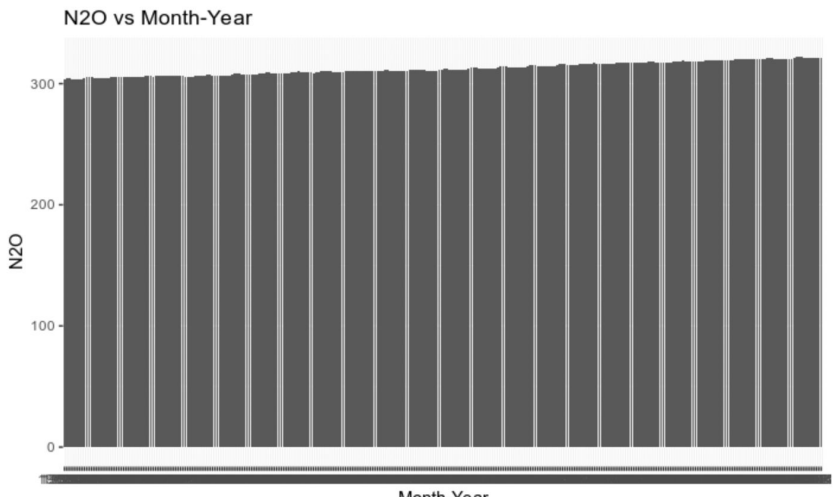
4) CH4 Frequency distribution per month:

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



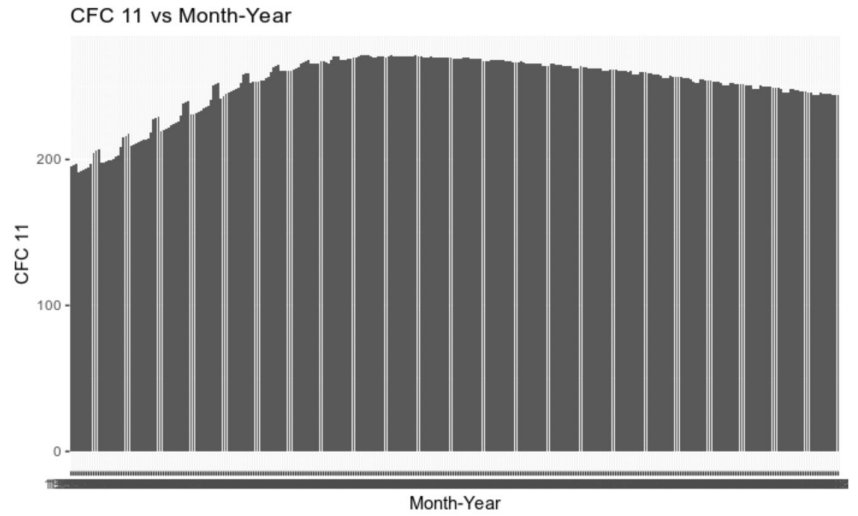
5) N2O Frequency distribution per month:

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



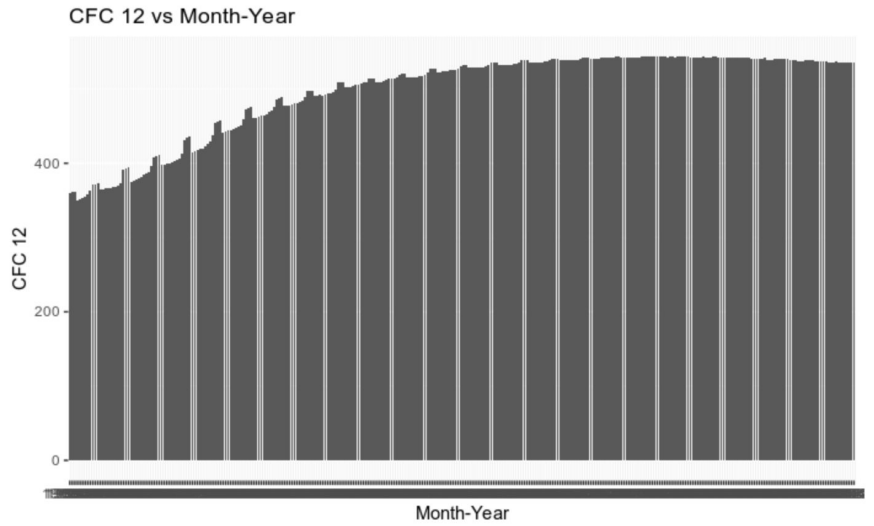
6) CFC.11 Frequency distribution per month:

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



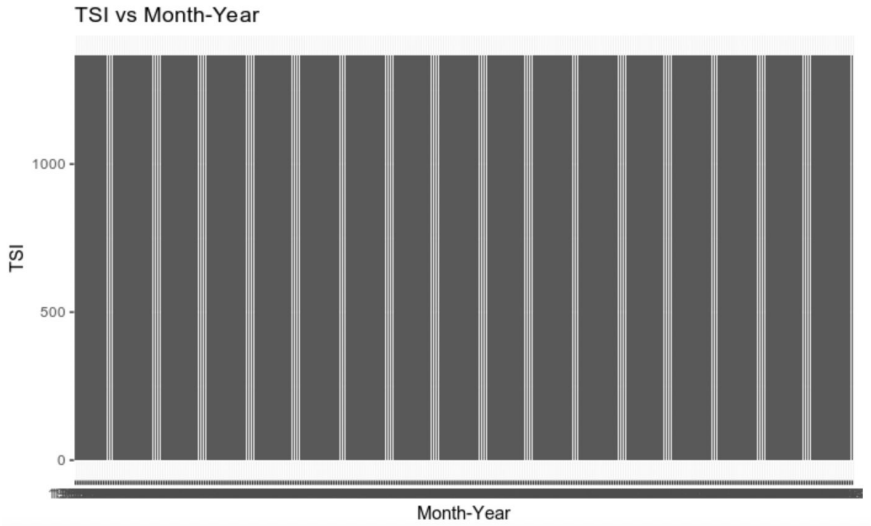
7) **CFC.12 Frequency distribution per month:**

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



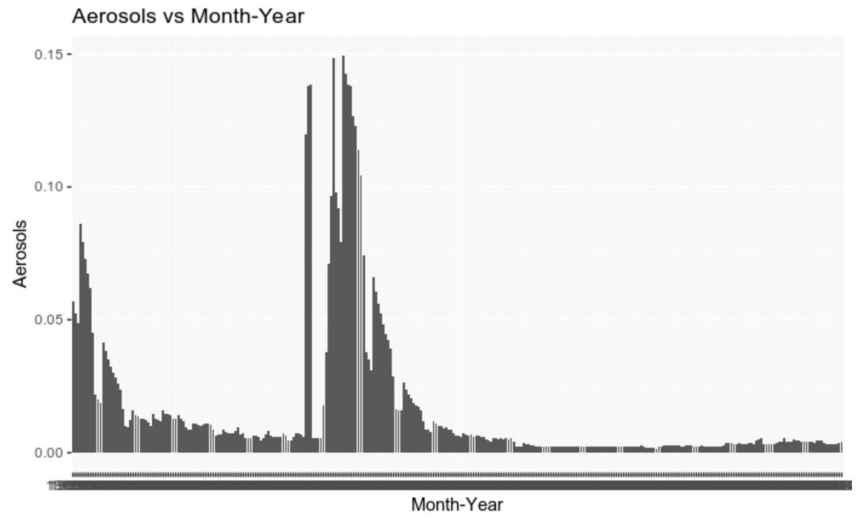
8) TSI Frequency distribution per month:

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



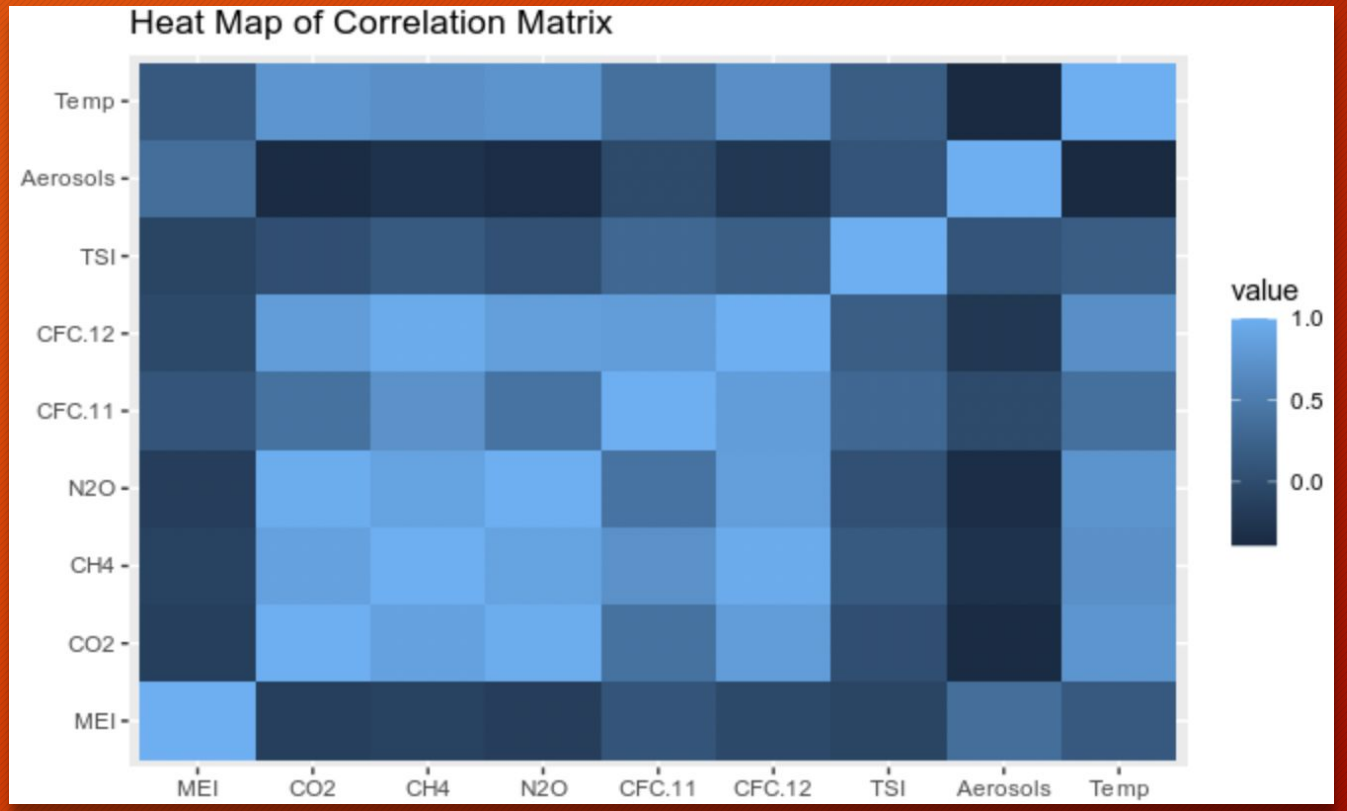
9) Aerosols Frequency distribution per month:

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

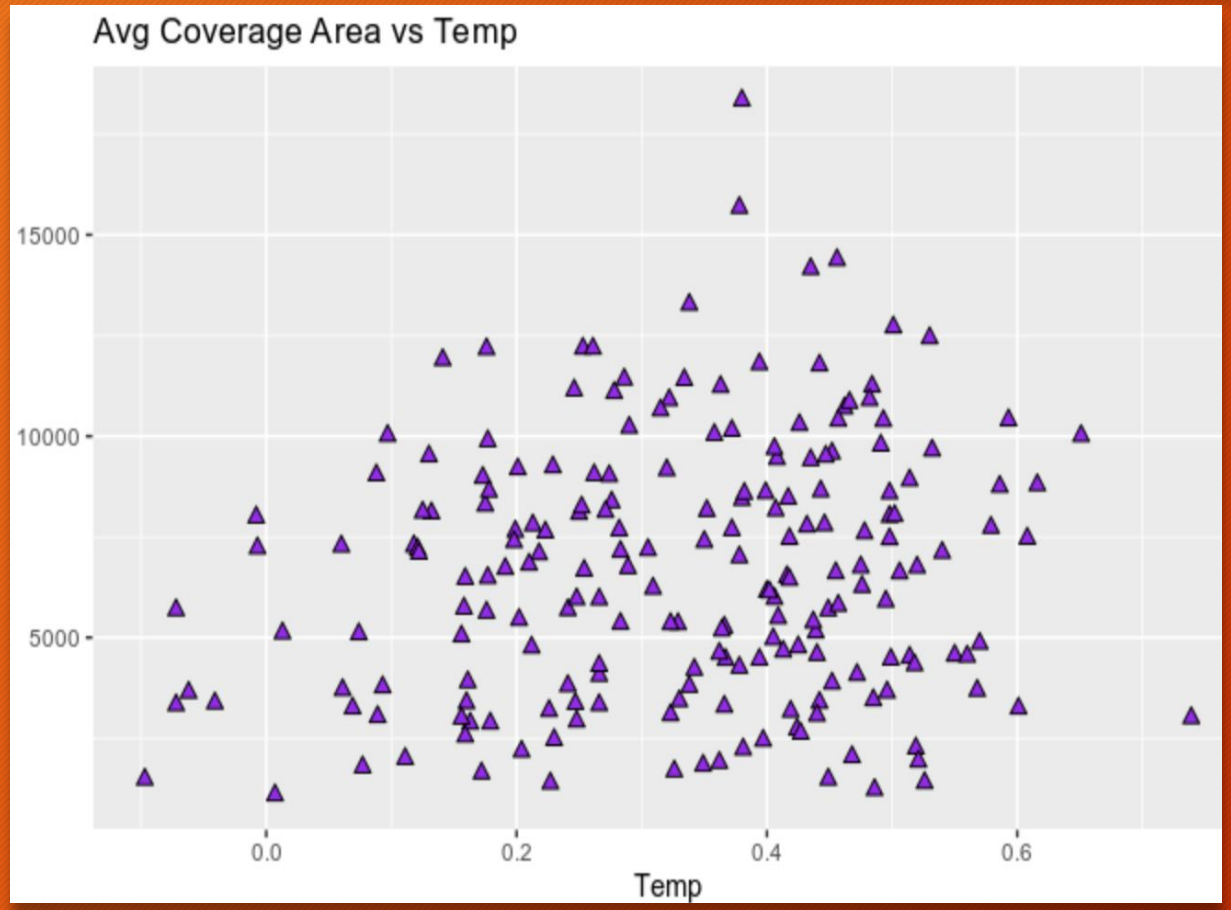


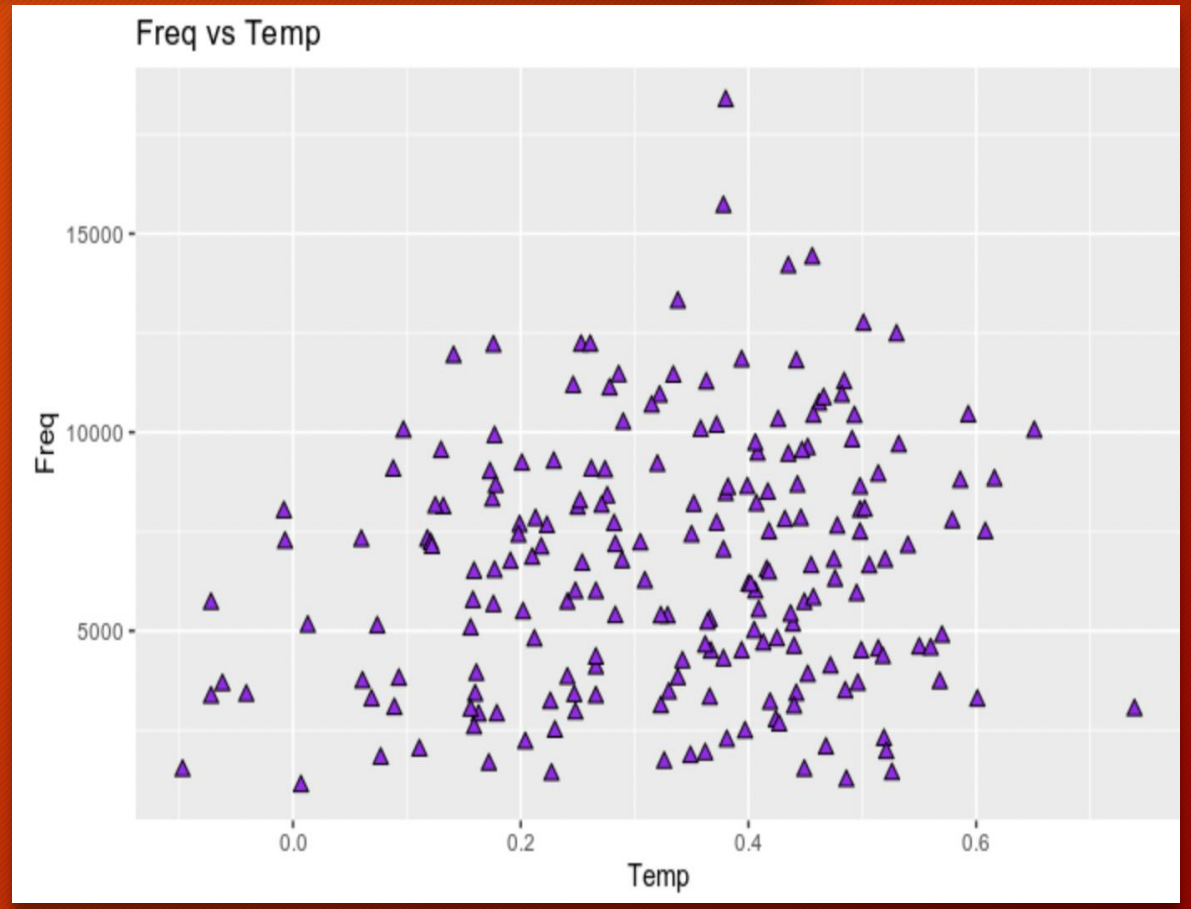
Correlation matrix

- 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, Methane, Dinitrogen Oxide, CFC 12) in the atmosphere.

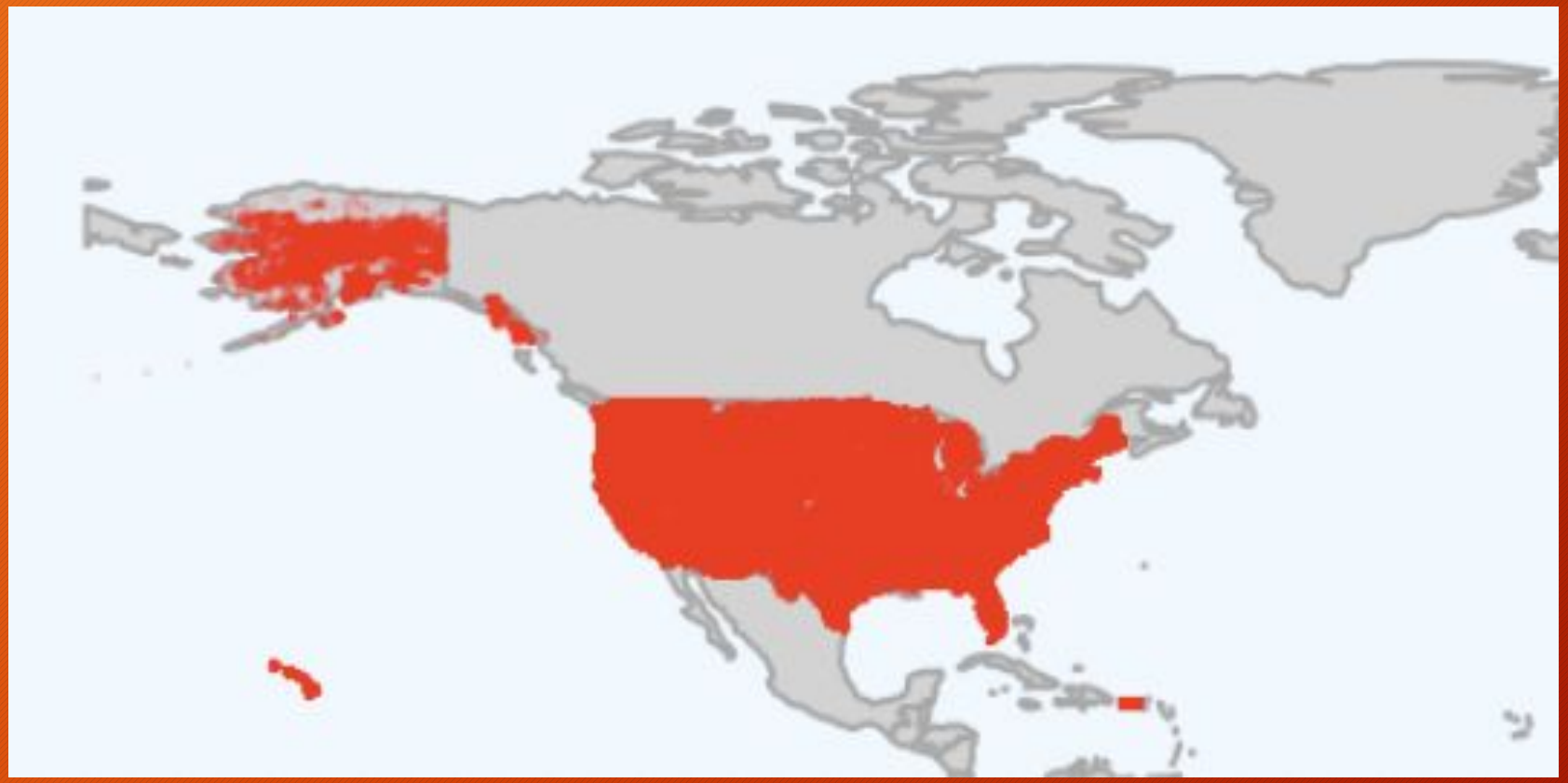


Scatterplots





Geometric Plot – Fire locations on a map



Algorithms

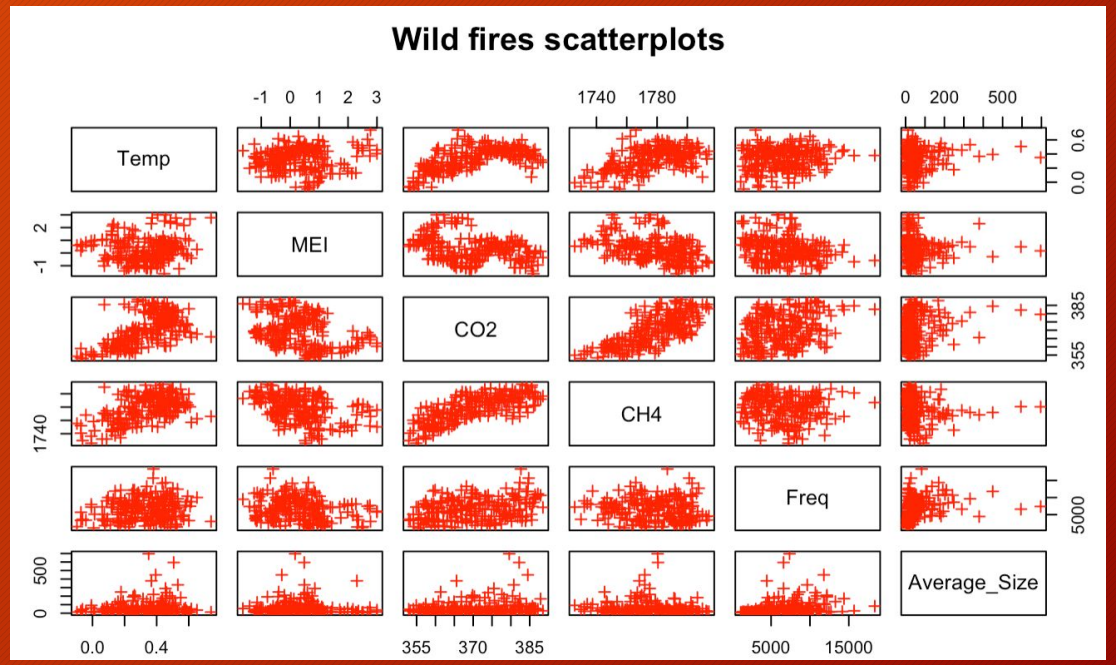
Regression Analysis

- Narrowed down upon the important independent variables of the climate that actually affect the frequency and average size of forest fires
- 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 highly correlated, hence we considered only CO2
- The final independent variables are Temp, CO2, CH4 and MEI

	[,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

Linear regression

- Scatterplots are 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



Linear regression

- 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.

```
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
```

Linear regression(frequency)

- 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.

- 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**

```
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
```

Linear regression(average_size)

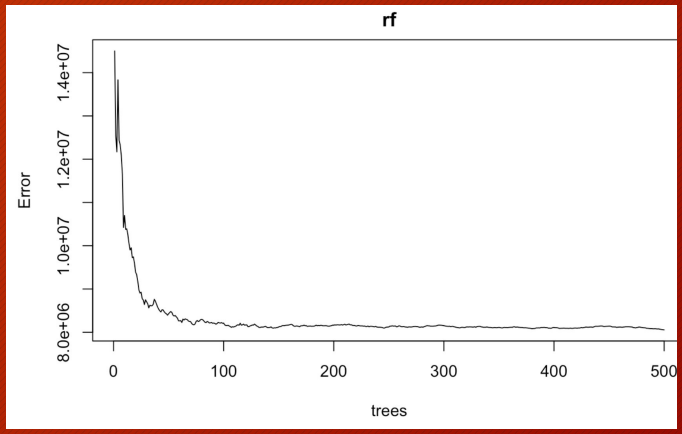
- From the summary we can say that “CO2 & CH4” are significant in predicting the average size of wildfires per month.
- The Estimate column in the coefficients table, gives us the coefficients for each independent variable in the regression model. Our model is Freq(y) = 3083.8709 + 45.7955 (Temp) - 7.0846 (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

```
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
```

Random Forest regression(Freq)

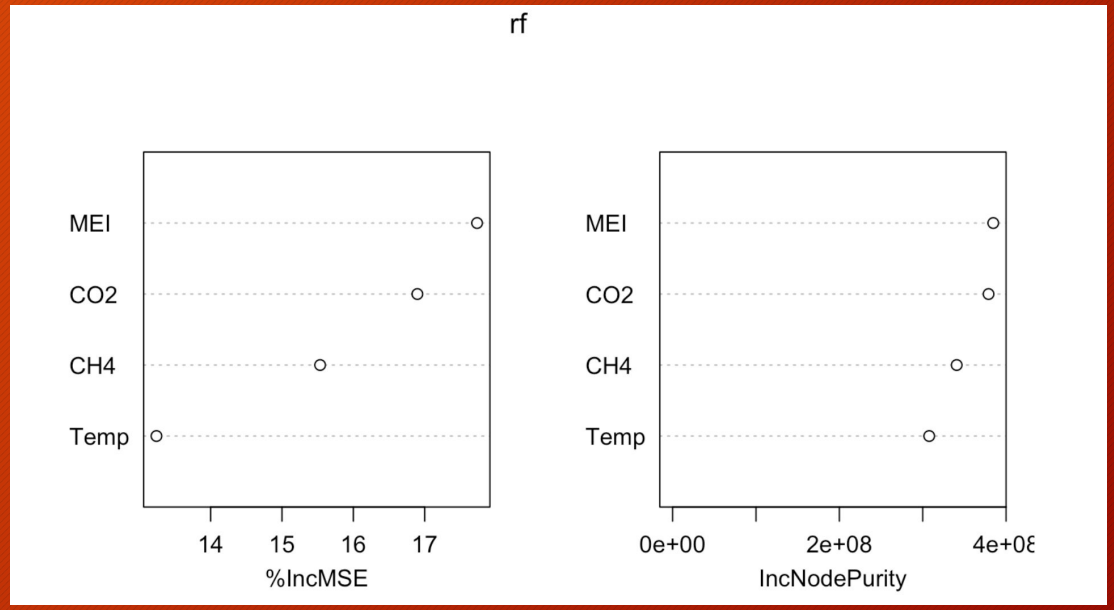
- Random forest regression predicts output by average by bootstrap algorithm or bagging of independently built decision trees.
- There are a lot of combinations possible between the parameters. Without trying all of them we can use Grid Search in R. With grid search the model will be evaluated over all the combinations you pass in the function, using cross-validation.

```
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
```



Random Forest regression(Freq)

- 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.



Random Forest regression(Freq)

- These are the prediction results and mean squared error

	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

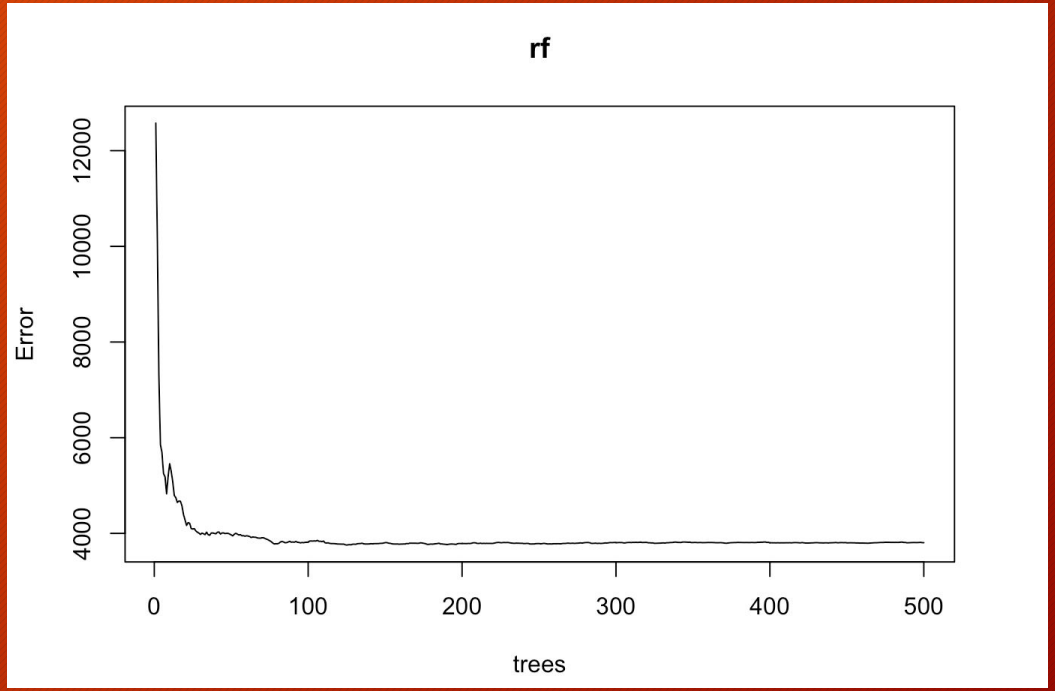
1-10 of 51 rows Previous 1 2 3 4 5 6 Next

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

Random Forest regression(Average_size)

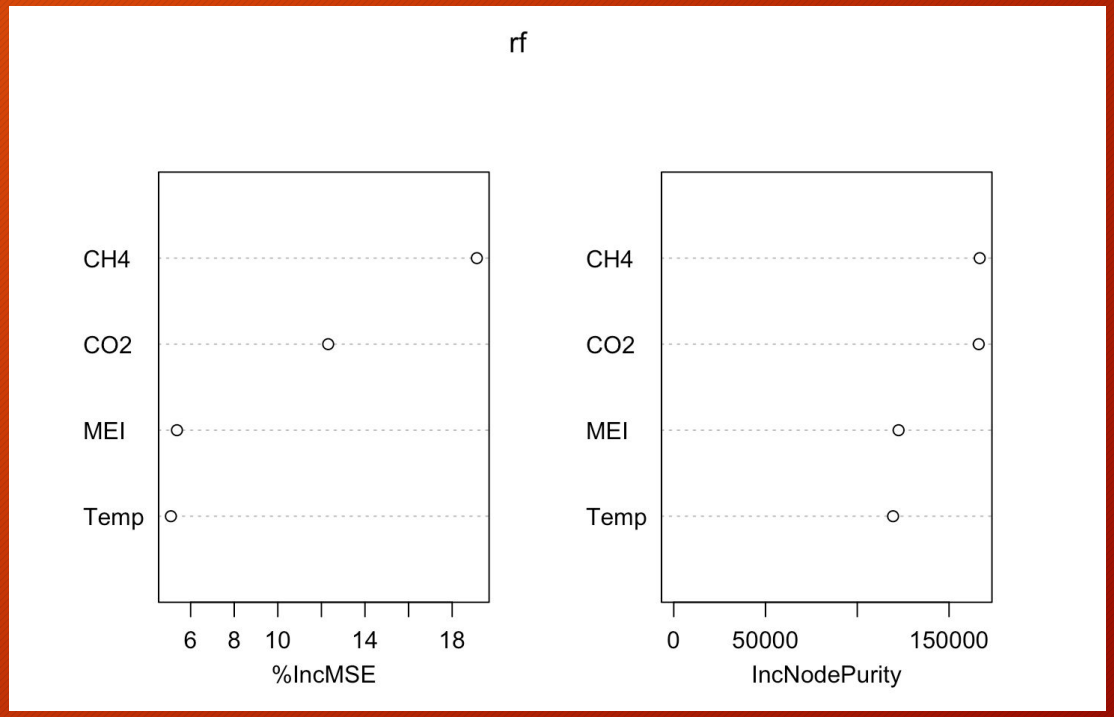
- The model summary for average_size

```
##
## 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
```



Random Forest regression(Average_size)

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



Random Forest regression(Average_size)

- The predictions and the MSE 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

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

Results & Future Work

Results

- 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.

Future Work

- When would the next fire is probably going to happen? Doing Time Series Analysis
- Prediction based on other important climatic conditions like precipitation, humidity, atmospheric pressure
- Instead of predicting the collective affect of the climatic conditions, future work can be done to find how each climatic condition can affect forest fires

Thank you!