Team Introduction

- Rahul Sinha - Full-Time MBA Class of 2020
- Sweta Sharma - Full-Time MBA Class of 2021

Problem Statement

As per NASA's study on Global Warming[1], average temperature has increased by over 2 degrees fahrenheit since 1880. We wanted to explore climate data to identify the relationship between temperature levels and concentrations of emissions in the atmosphere and other factors.

In this project, we will study the relationship between average global temperature and various other factors to see if those affect the average temperature. We will use linear models to understand the advantages and shortcomings of such models in explaining the variation in the independent variable.

Dataset

The dataset is obtained from MIT Open Courseware[2]. The data was compiled from sources such as the Climatic Research Unit at the University of East Anglia, ESRL/NOAA Global Monitoring Division, etc. The dataset contains climate data from May 1983 to December 2008. The variables in the dataset are listed below.

Year: The observation year from 1983 to 2008.

Month: The observation month in the particular year.

Temp: (Degree Celsius) The average temperature difference between the average global temperature in that period and a reference value. This variable is our target or the dependent variable.

CO2: (Parts per million by volume) Concentration of carbon dioxide during the period

N2O: (Parts per million by volume) Concentration of nitrous oxide during the period

CH4: (Parts per million by volume) Concentration of methane during the period

CFC.11: (Parts per billion by volume) Concentration of trichlorofluoromethane during the period

CFC.12: (Parts per billion by volume) Concentration of dichlorodifluoromethane

[1] World of Change: Global Temperatures. (n.d.). Retrieved from https://earthobservatory.nasa.gov/world-of-change/global-temperatures

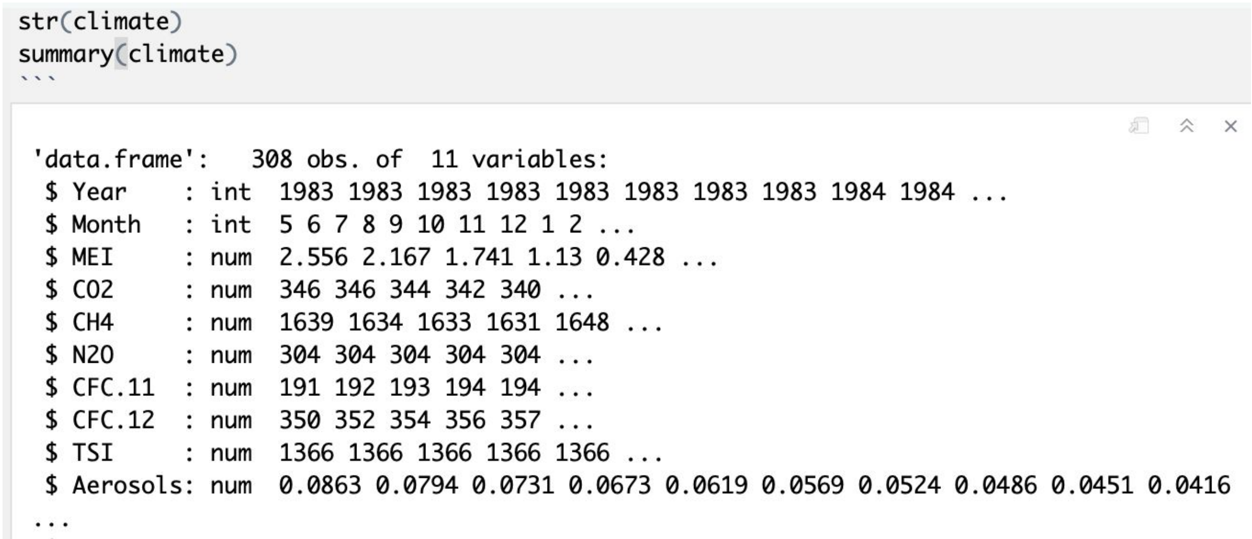
[2] Bertsimas, Dimitris. (n.d.). 2.5 Assignment 2. Retrieved from https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/linear-regression/assignment-2/

Aerosols: The mean stratospheric aerosol optical depth at 550 nm. Volcanic eruptions add particles to the atmosphere and particles affect how the sun's energy is reflected into space.

TSI: The total solar irradiance (TSI) in W/m2

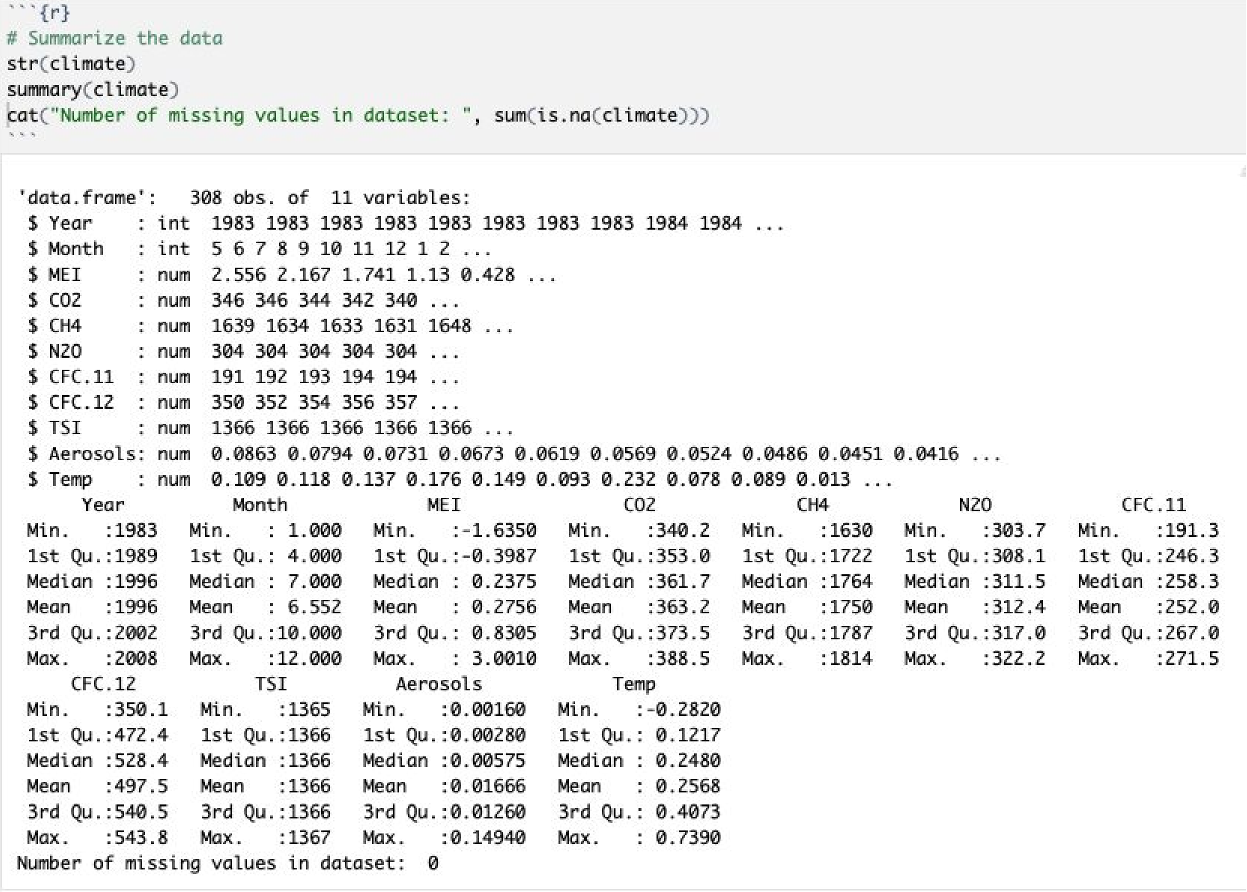
MEI: Multivariate El Nino Southern Oscillation Index (MEI)

Variables Values & Data Types



```
str(climate)
summary(climate)
```
```

```
'data.frame':   308 obs. of  11 variables:
 $ Year    : int  1983 1983 1983 1983 1983 1983 1983 1983 1984 1984 ...
 $ Month   : int  5 6 7 8 9 10 11 12 1 2 ...
 $ MEI     : num  2.556 2.167 1.741 1.13 0.428 ...
 $ CO2     : num  346 346 344 342 340 ...
 $ CH4     : num  1639 1634 1633 1631 1648 ...
 $ N2O     : num  304 304 304 304 304 ...
 $ CFC.11  : num  191 192 193 194 194 ...
 $ CFC.12  : num  350 352 354 356 357 ...
 $ TSI     : num  1366 1366 1366 1366 1366 ...
 $ Aerosols: num  0.0863 0.0794 0.0731 0.0673 0.0619 0.0569 0.0524 0.0486 0.0451 0.0416
...
```

Methodology

We followed the Sample, Explore, Modify, Model, and Assess (SEMMA) steps in our project.

1. **Purpose**: Predict our target output variable - Temp and understand the relation between Temp and other variables in the dataset.
2. **Sample**: We will use the data from climatre_change.csv. The dataset is small with 308 rows and hence we do not need to sample from the data.
```

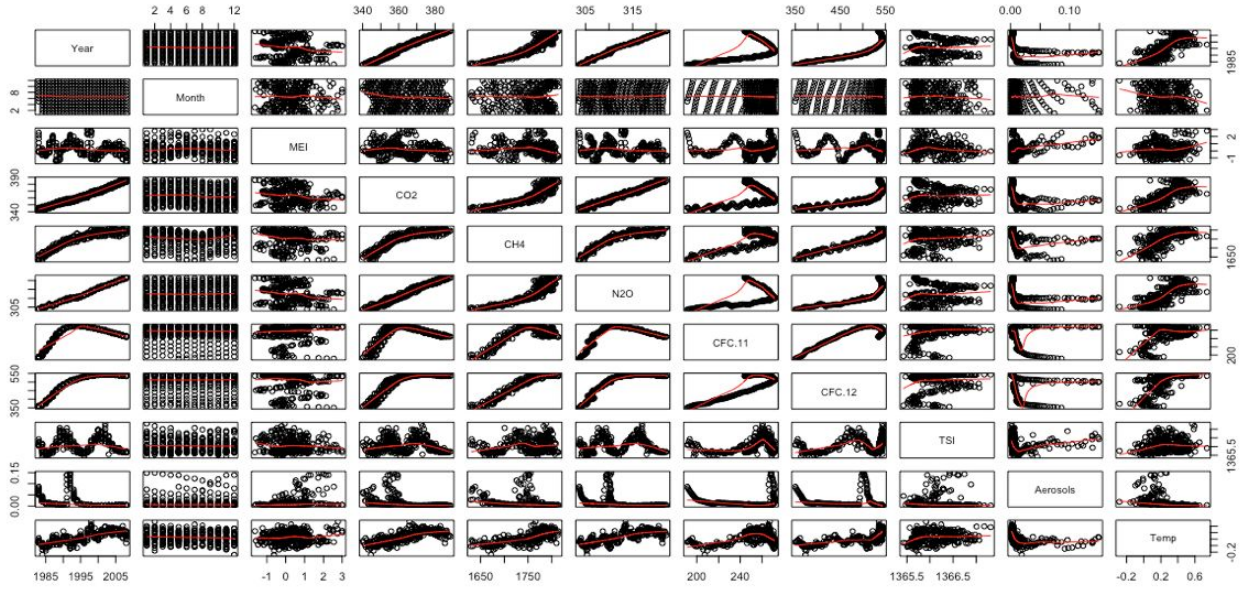
3. **Explore**: In this step, we will explore, clean, and preprocess the data. We find that the dataset is clean without any missing values.



~~~
```{r}
# Summarize the data
str(climate)
summary(climate)
cat("Number of missing values in dataset: ", sum(is.na(climate)))
```
~~~

```
'data.frame':	308 obs. of  11 variables:
 $ Year     : int  1983 1983 1983 1983 1983 1983 1983 1983 1984 1984 ...
 $ Month    : int  5 6 7 8 9 10 11 12 1 2 ...
 $ MEI      : num  2.556 2.167 1.741 1.13 0.428 ...
 $ CO2      : num  346 346 344 342 340 ...
 $ CH4      : num  1639 1634 1633 1631 1648 ...
 $ N2O      : num  304 304 304 304 304 ...
 $ CFC.11   : num  191 192 193 194 194 ...
 $ CFC.12   : num  350 352 354 356 357 ...
 $ TSI      : num  1366 1366 1366 1366 1366 ...
 $ Aerosols: num  0.0863 0.0794 0.0731 0.0673 0.0619 0.0569 0.0524 0.0486 0.0451 0.0416 ...
 $ Temp     : num  0.109 0.118 0.137 0.176 0.149 0.093 0.232 0.078 0.089 0.013 ...
      Year          Month             MEI               CO2             CH4             N2O            CFC.11     
 Min.   :1983   Min.   : 1.000   Min.   :-1.6350   Min.   :340.2   Min.   :1630   Min.   :303.7   Min.   :191.3  
 1st Qu.:1989   1st Qu.: 4.000   1st Qu.:-0.3987   1st Qu.:353.0   1st Qu.:1722   1st Qu.:308.1   1st Qu.:246.3  
 Median :1996   Median : 7.000   Median : 0.2375   Median :361.7   Median :1764   Median :311.5   Median :258.3  
 Mean   :1996   Mean   : 6.552   Mean   : 0.2756   Mean   :363.2   Mean   :1750   Mean   :312.4   Mean   :252.0  
 3rd Qu.:2002   3rd Qu.:10.000   3rd Qu.: 0.8305   3rd Qu.:373.5   3rd Qu.:1787   3rd Qu.:317.0   3rd Qu.:267.0  
 Max.   :2008   Max.   :12.000   Max.   : 3.0010   Max.   :388.5   Max.   :1814   Max.   :322.2   Max.   :271.5  
     CFC.12           TSI          Aerosols             Temp        
 Min.   :350.1   Min.   :1365   Min.   :0.00160   Min.   :-0.2820  
 1st Qu.:472.4   1st Qu.:1366   1st Qu.:0.00280   1st Qu.: 0.1217  
 Median :528.4   Median :1366   Median :0.00575   Median : 0.2480  
 Mean   :497.5   Mean   :1366   Mean   :0.01666   Mean   : 0.2568  
 3rd Qu.:540.5   3rd Qu.:1366   3rd Qu.:0.01260   3rd Qu.: 0.4073  
 Max.   :543.8   Max.   :1367   Max.   :0.14940   Max.   : 0.7390  
Number of missing values in dataset:  0
```

Additionally, we looked at all the correlations in the data using the pairs function:

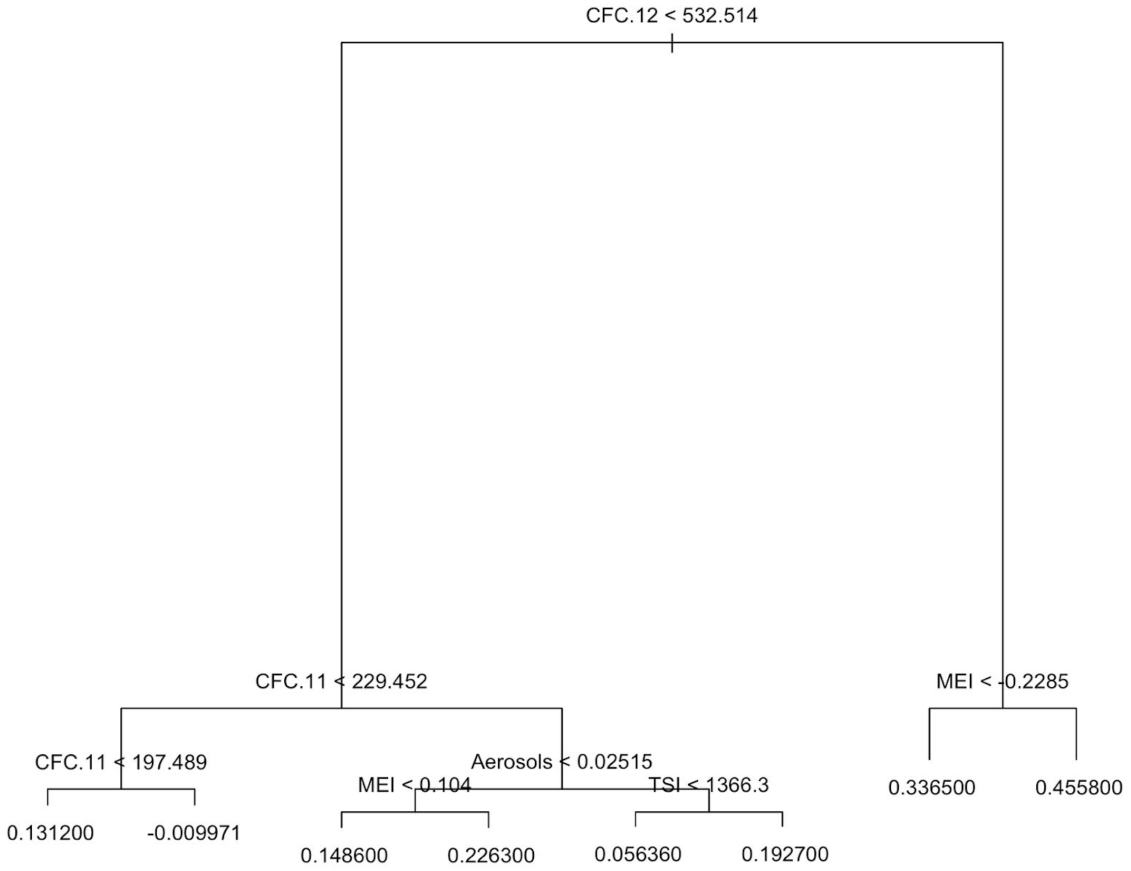


There is a pronounced positive correlation between Temp and the pollutant variables except for Aerosols.

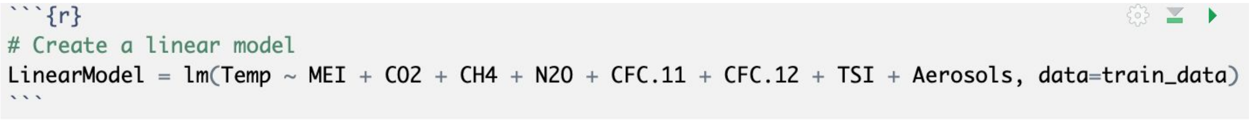
We looked at the histograms of the different columns to confirm that the variables were unimodal, but observed that some of the variable distributions were skewed.

4. **Modify**: In this step, we partition the data into train and test data. We will ignore the Year and Month data as we only want to understand the effect of the concentration of different gases on the variable Temp. The Temp and pollutant concentration variables are all continuous, hence, there is no need to recode any variable. Additionally, the Temp variable is a difference of temperatures and is normally distributed. We partitioned the dataset into training and test data. Training data is for the years less than or equal to 2005 and the test data is for years greater than 2005. Thus, we used 272 rows or 88% of the dataset for training and 36 rows or 12% of the dataset for testing.

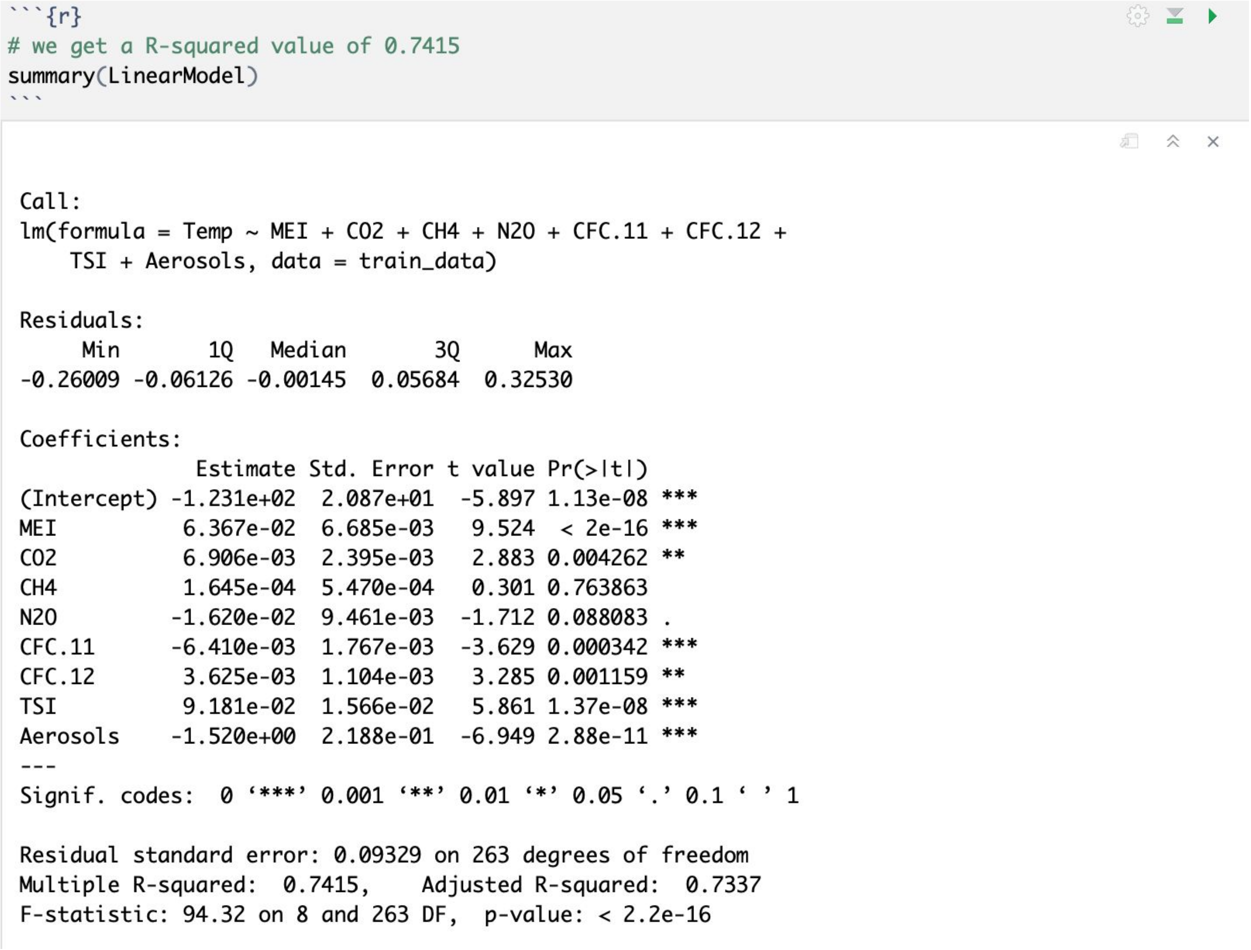
5. **Model**: After partitioning the data, we wanted to apply linear regression on the training data to build a model. We started with a tree model to get an idea of the interactions between the variables. The tree model shows us that the interactions are not too complex. As per the tree, MEI and TSI are the most important factors affecting Temp. MEI is important at both low and high levels of CFC.12. At low CFC.12 levels, TSI matters when Aerosols and CFC.11 concentrations are high. However, at high CFC.12 levels, higher MEI concentrations result in higher temperatures. In both cases, higher MEI is associated with a higher temperature difference.



Now that we have an understanding of the interactions and the relationships between variables, we start our linear modeling. Our target variable, Temp, is continuous and we start with a linear model consisting of all the variables.



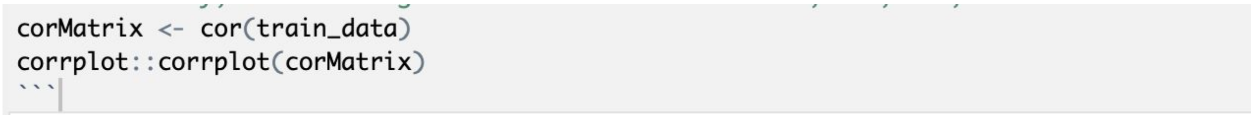
```{r}
# Create a linear model
LinearModel = lm(Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + TSI + Aerosols, data=train_data)
```

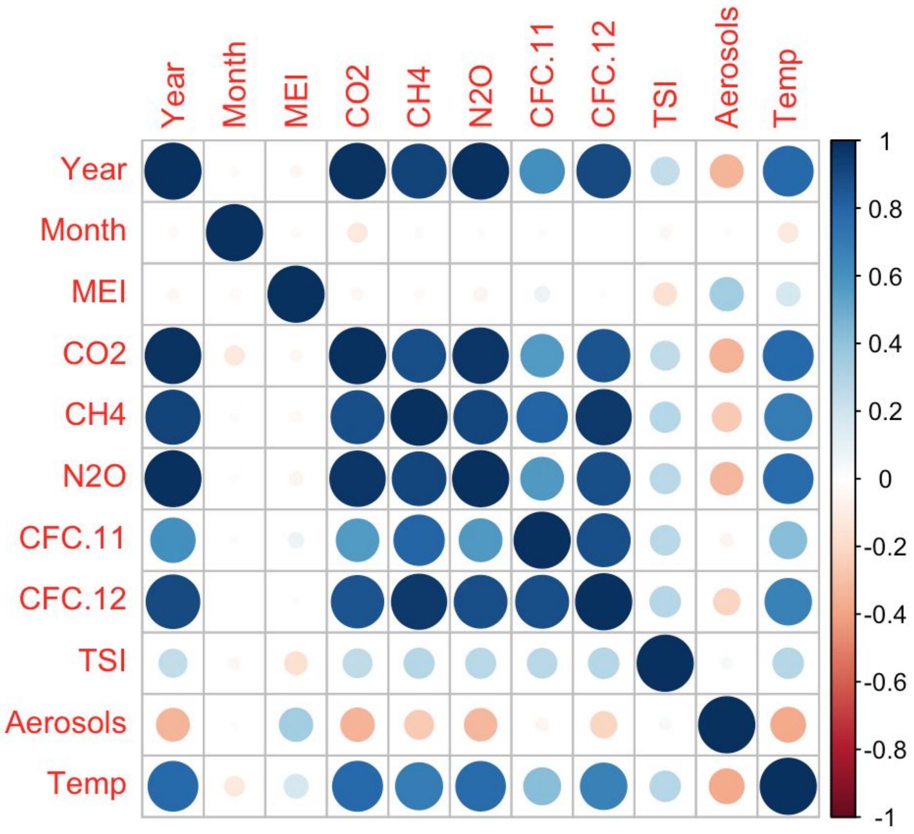
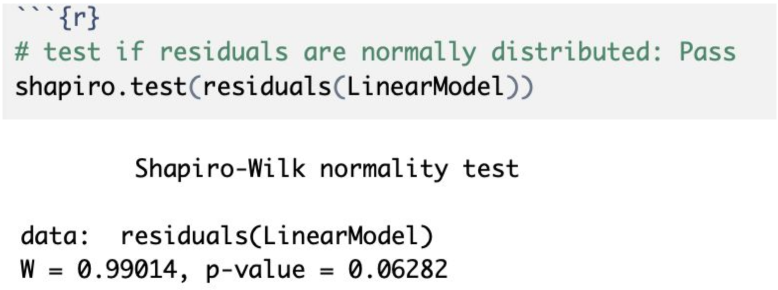


```{r}
# we get a R-squared value of 0.7415
summary(LinearModel)
```

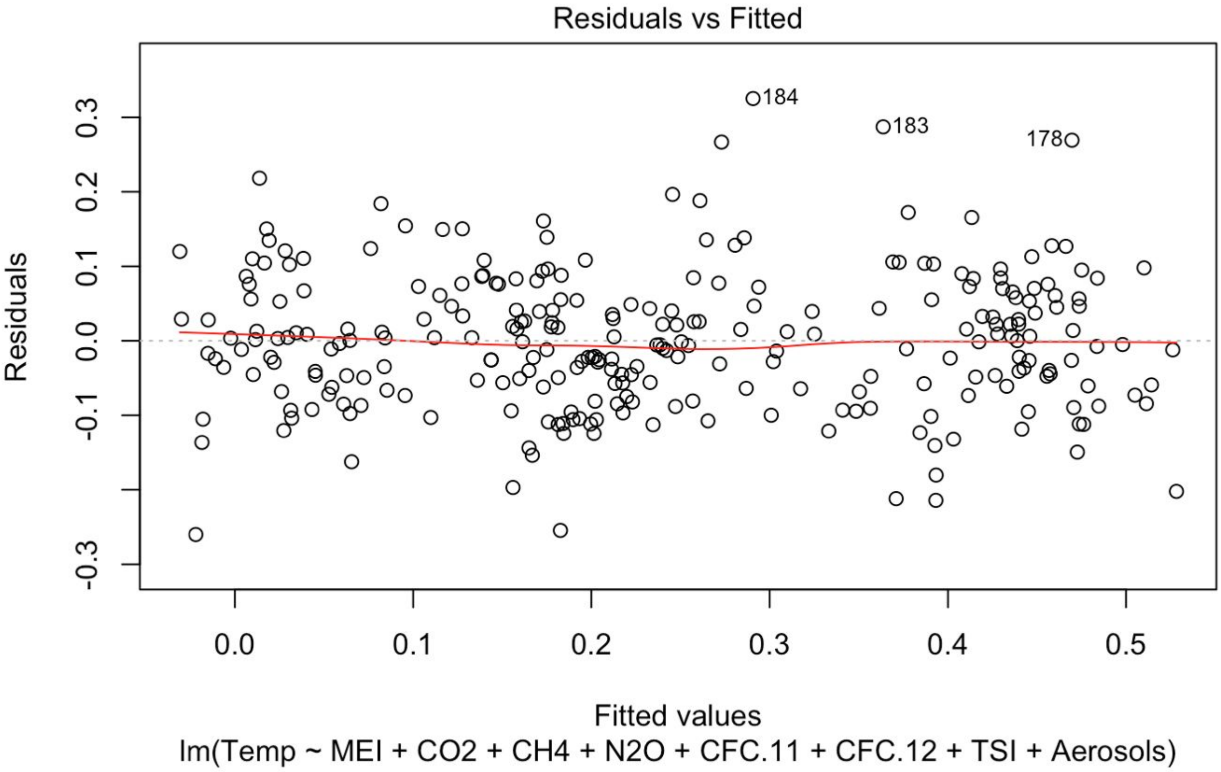
```
Call:
lm(formula = Temp ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + 
    TSI + Aerosols, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26009 -0.06126 -0.00145  0.05684  0.32530 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.231e+02  2.087e+01  -5.897 1.13e-08 ***
MEI          6.367e-02  6.685e-03   9.524  < 2e-16 ***
CO2          6.906e-03  2.395e-03   2.883 0.004262 ** 
CH4          1.645e-04  5.470e-04   0.301 0.763863    
N2O         -1.620e-02  9.461e-03  -1.712 0.088083 .  
CFC.11      -6.410e-03  1.767e-03  -3.629 0.000342 ***
CFC.12       3.625e-03  1.104e-03   3.285 0.001159 ** 
TSI          9.181e-02  1.566e-02   5.861 1.37e-08 ***
Aerosols    -1.520e+00  2.188e-01  -6.949 2.88e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09329 on 263 degrees of freedom
Multiple R-squared:  0.7415,	Adjusted R-squared:  0.7337 
F-statistic: 94.32 on 8 and 263 DF,  p-value: < 2.2e-16
```

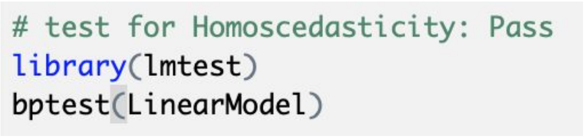
We get an R-squared value of 0.7415 and the model suggests that CH4 and N20 are not significant. Importantly, the model gives negative correlations of Temp with N20 and CFC.11, but we know that is contrary to scientific understanding. We suspect that this may be due to the correlation between the independent variables. We thus plot a correlation matrix to understand the correlation between variables.



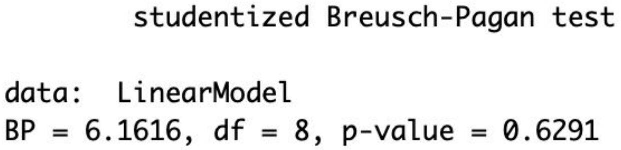
```
corMatrix <- cor(train_data)
corrplot::corrplot(corMatrix)
```
```


Year
Month
MEI
CO2
CH4
N2O
CFC.11
CFC.12
TSI
Aerosols
Temp
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

We notice that N2O is correlated with CO2, CH4, and CFC.12. Also, CFC.11 is correlated with CH4 and CFC.12. Additionally, there is significant correlation between CO2, CH4, N2O, and CFC.12

Once we have our linear model, we test whether the residuals are normally distributed and whether the residuals show homoscedasticity. We find that both conditions are met.



```
```{r}
# test if residuals are normally distributed: Pass
shapiro.test(residuals(LinearModel))
```

```

	Shapiro-Wilk normality test

data:  residuals(LinearModel)
W = 0.99014, p-value = 0.06282
```



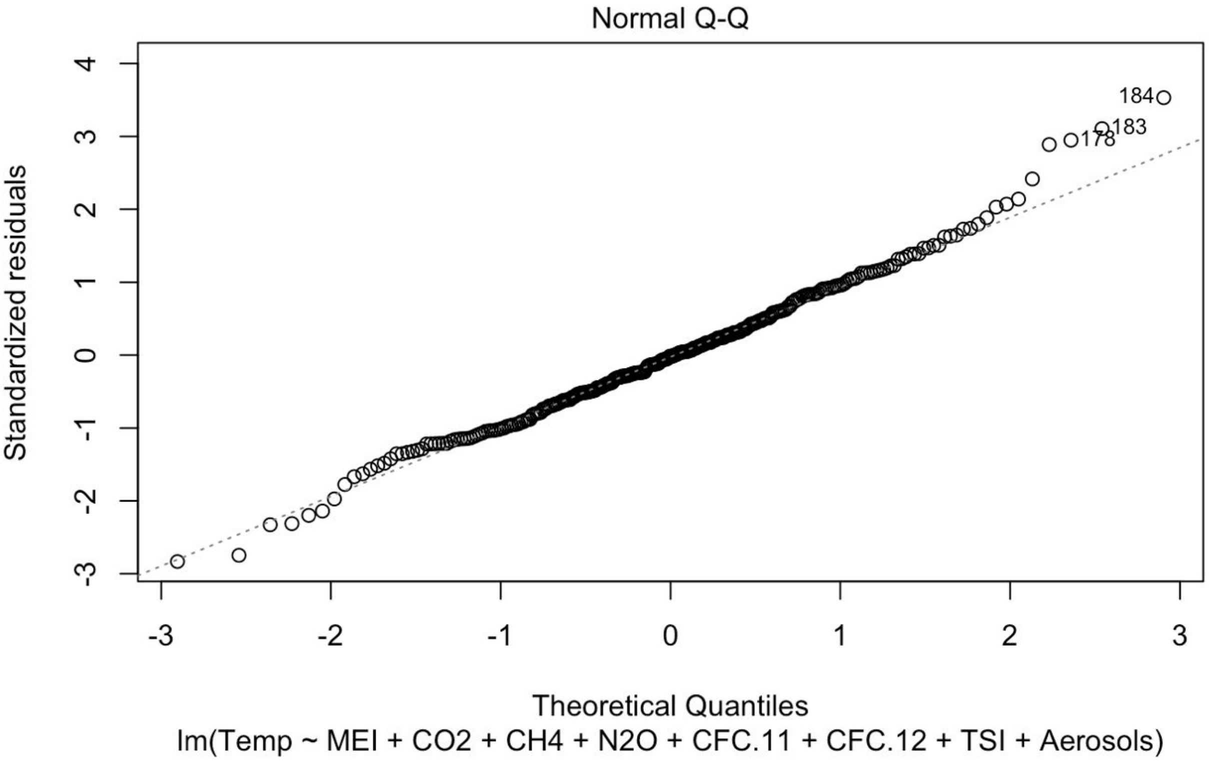


```
# test for Homoscedasticity: Pass
library(lmtest)
bptest(LinearModel)
```

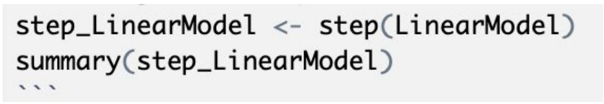


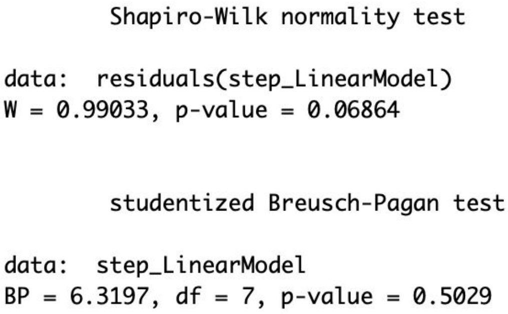
```
	studentized Breusch-Pagan test

data:  LinearModel
BP = 6.1616, df = 8, p-value = 0.6291
```

We confirm our findings by plotting Residuals vs. Fitted chart and a Q-Q plot of the residuals. We notice the constant variance or homoscedasticity of the residuals with the increasing fitted mean. We can also see the nearly straight line formed by the residuals on the Q-Q plot.



As CO2, CH4, N2O, and CFC.12 are correlated we may try building a model with MEI, TSI, Aerosols, and one of the correlated variables, say CH4 from the group of CO2, CH4, N2O, and CFC.12. We could remove the variables one by one removing the least significant variable first from the 'LinearModel' that uses all the variables. We use the step function provided by R to achieve this result:

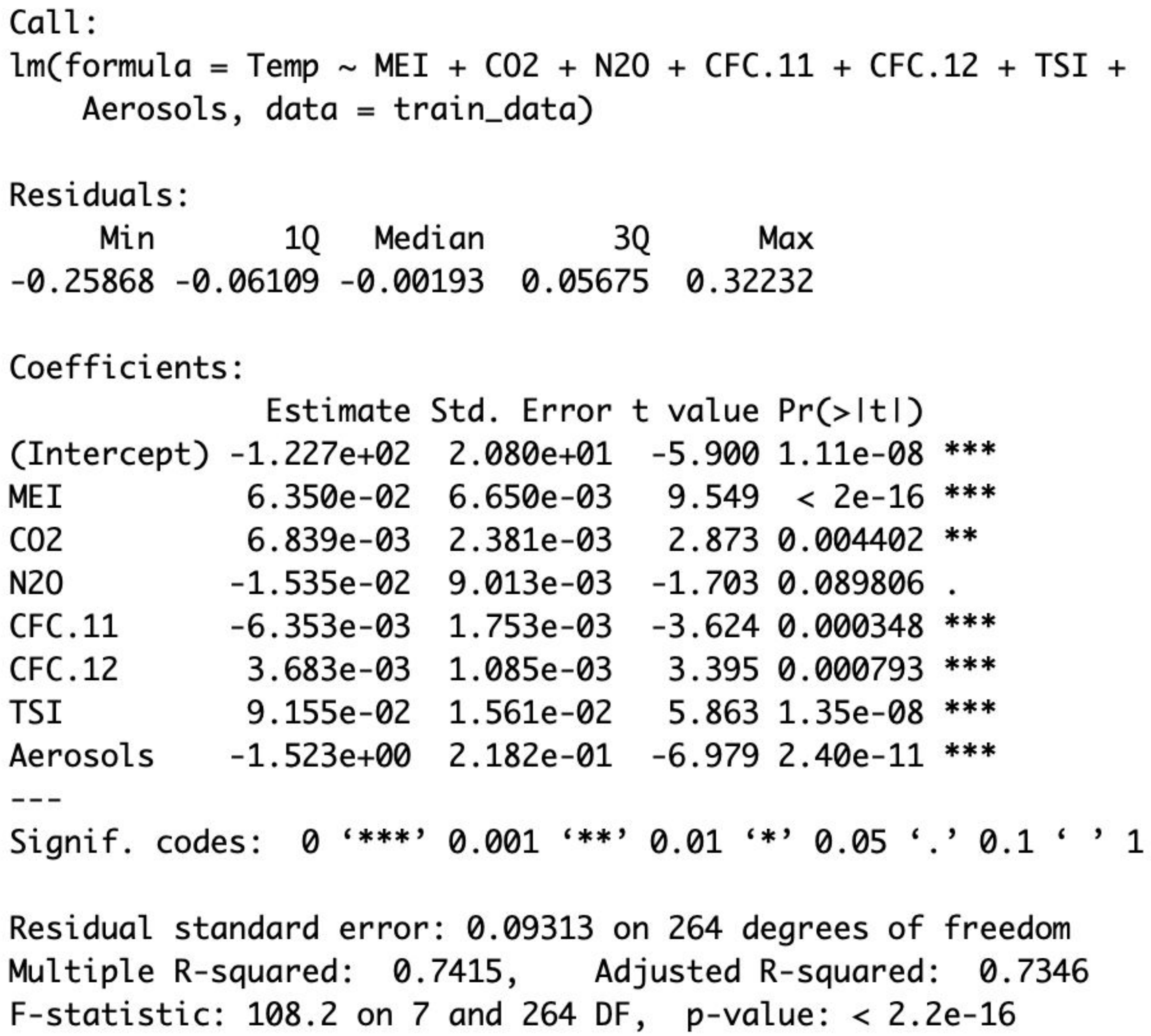


```
step_LinearModel <- step(LinearModel)
summary(step_LinearModel)
```
```

Our final model is given by :
Temp ~ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI + Aerosols

We note that the new model has the same R-squared value of 0.7415 as the full model, but has a lower Adjusted R-squared of 0.7346

Once we have our linear model, we test whether the residuals are normally distributed and whether the residuals show homoscedasticity. We find that both conditions are met.



```
                Shapiro-Wilk normality test

data: residuals(step_LinearModel)
W = 0.99033, p-value = 0.06864

                studentized Breusch-Pagan test

data: step_LinearModel
BP = 6.3197, df = 7, p-value = 0.5029
```
```



```
Call:
lm(formula = Temp ~ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI +
    Aerosols, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25868 -0.06109 -0.00193  0.05675  0.32232

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.227e+02  2.080e+01  -5.900 1.11e-08 ***
MEI          6.350e-02  6.650e-03   9.549  < 2e-16 ***
CO2          6.839e-03  2.381e-03   2.873 0.004402 **
N2O         -1.535e-02  9.013e-03  -1.703 0.089806 .
CFC.11      -6.353e-03  1.753e-03  -3.624 0.000348 ***
CFC.12       3.683e-03  1.085e-03   3.395 0.000793 ***
TSI          9.155e-02  1.561e-02   5.863 1.35e-08 ***
Aerosols    -1.523e+00  2.182e-01  -6.979 2.40e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09313 on 264 degrees of freedom
Multiple R-squared:  0.7415,    Adjusted R-squared:  0.7346
F-statistic: 108.2 on 7 and 264 DF,  p-value: < 2.2e-16
```

6. **Assess:** Next, we assess our model by running it on the test data, predicting the Temp target variable for the data, and calculating the R-Squared value. We get an R-squared value of 0.7630, which indicates that our model explains 76.30% of the variance of our target variable Temp.

Conclusion

First, our analysis reveals how the concentrations of gases affect the change in temperatures. We note that these results are good at explaining 73.46% of the variance in Temp at 95% confidence levels. Second, we saw how a regression tree model was much easier to understand compared to the linear model. Third, even though removing the correlated independent variables did not affect our R-squared significantly, the step function optimizes with AIC and gives a model with improved R-squared but still consisting of numerous correlated variables. The negative coefficient of N20 and CFC.11 indicates that although the model explains most of the variance in our target variable, we may want to look for better models that are easier to understand and do not give results contrary to established scientific understanding.