Level: 6000

Git Repo: https://github.com/jj960708/Spring2020_DataAnalytics/tree/master/dsproject

Abstraction:

Australia suffered from massive bushfires in the early of this year, and experts estimated that Australia's bush fires will become more frequent and more intense as climate warming worsens. Global warming is the rise of the average temperature of Earth's climate system. It is believed that the average global temperature has increased at the fastest rate in the past 50 years, but there are some opinions argue that the global warming is not as serious as the public believes. Therefore, I want to explore if the global average temperature increased in the past 200 years. In addition, I want to explore what will our future temperature be based on the historical data to know if we will experience global warming. There are some existing climate models that can predict the temperature in the next few days accurately (e.g. Weather forecasting), but it is difficult to get accurate predictions of the future temperature in long-term. However, we need to have long-term predictions of the average temperature in order to study the impact of global warming. In this project, I want to explore two different models to have reasonable predictions of the temperature in the next few years.

Data Description:

One of the purposes of this project is to see whether the global average temperature is increasing, so I need to find a dataset that contains the historical data of average temperature. The first dataset I use in this project is land surface temperature dataset from Berkeley Earth. This dataset contains monthly average temperature of major cities across the world since 1750. Since according to the scientists, global warming happens when carbon dioxide and other greenhouses gases gathered in the atmosphere and absorbed solar radiation that have bounced off the earth's temperature[1], I also use the historical total solar irradiance reconstruction dataset and three major greenhouse gases emission dataset in this project. The source of historical total solar irradiance reconstruction dataset is LASP interactive solar irradiance datacenter of University of Colorado[2]. The land surface temperature is shown as figure 1.1, the solar irradiance dataset is shown as figure 1.2 and greenhouse gas emission dataset is shown as 1.3

	dt	AverageTemperature	AverageTemperatureUncertainty	Country
1	1743-11-01	4.384	2.294	Ã…land
2	1743-12-01	NA	NA	Ã…land
3	1744-01-01	NA	NA	Ã…land
4	1744-02-01	NA	NA	Ã…land
5	1744-03-01	NA	NA	Ã…land

	time..yyyy.	Irradiance..W.m.2.
1	-1610	-1360.186
2	-1611	-1360.470
3	-1612	-1360.680
4	-1613	-1360.957
5	-1614	-1361.088
6	-1615	-1361.021
7	-1616	-1360.674

Figure 1.1 Figure 1.2

	Entity	Code	Year	SFâ...gases..tonnes.	PFC.gases..tonnes.	HFC.gases..tonnes.	Nitrous.oxide..Nâ..O...tonnes.	Methane..CHâ.....tonnes.	Carbon.Dioxide..COâ.....tonnes.
1	Afghanistan	AFG	1960	NA	NA	NA	NA	NA	414371
2	Afghanistan	AFG	1961	NA	NA	NA	NA	NA	491378
3	Afghanistan	AFG	1962	NA	NA	NA	NA	NA	689396
4	Afghanistan	AFG	1963	NA	NA	NA	NA	NA	707731
5	Afghanistan	AFG	1964	NA	NA	NA	NA	NA	839743
6	Afghanistan	AFG	1965	NA	NA	NA	NA	NA	1008425

Figure 1.3

There are some missing data in the greenhouse gases emission data, and greenhouse gases emission dataset only contains the data since 1960. So, if we use greenhouse emission dataset in our model, we cannot use the temperature data before 1960 which will lose a lot of data and our model may potentially lack of enough data .In addition, I also mutate both the original land average temperature dataset and greenhouse gases emission datasets in order to merge these two datasets. After the merge and mutation, the new dataset looks like as the Figure 1.4 shows.

	Entity	Year	MeanT	MeanUnT	Code	SFA	PFC	HFC	NO2	CH4	CO2
1	Afghanistan	1960	13.98542	0.4418333	AFG	NA	NA	NA	NA	NA	414371
2	Afghanistan	1961	14.06492	0.3980833	AFG	NA	NA	NA	NA	NA	491378
3	Afghanistan	1962	13.76867	0.4061667	AFG	NA	NA	NA	NA	NA	689396
4	Afghanistan	1963	15.03342	0.4012500	AFG	NA	NA	NA	NA	NA	707731

Figure 1.4

Analysis:

First, a line chart of the solar irradiance is made as shown as Figure 2.1 and a line chart of global yearly average temperature is shown as Figure 2.2.

We can see that the solar irradiance is almost constant since 1700(around 1361.5 W/m2). The solar irradiance fluctuates, but there is not a clear trend shows that the solar irradiance level increased since 1700. However, we can observe an upward trend of yearly average temperature since 1750 from Figure 2.2. We cannot see the relationship between solar irradiance and global warming from these two plots. Since this project mainly focused on the temperature change since 1850, it is unnecessary to include the solar irradiance dataset in our analysis.

From the Figure 2.2, it is clear that the global temperature has increased for last 200 years and it increased rapidly in the past 50 years. The global average temperature rises to approximately 9.7 degree in 2012 from 8.7 degree in 1960. Therefore, global warming is real and is becoming more and more severe.

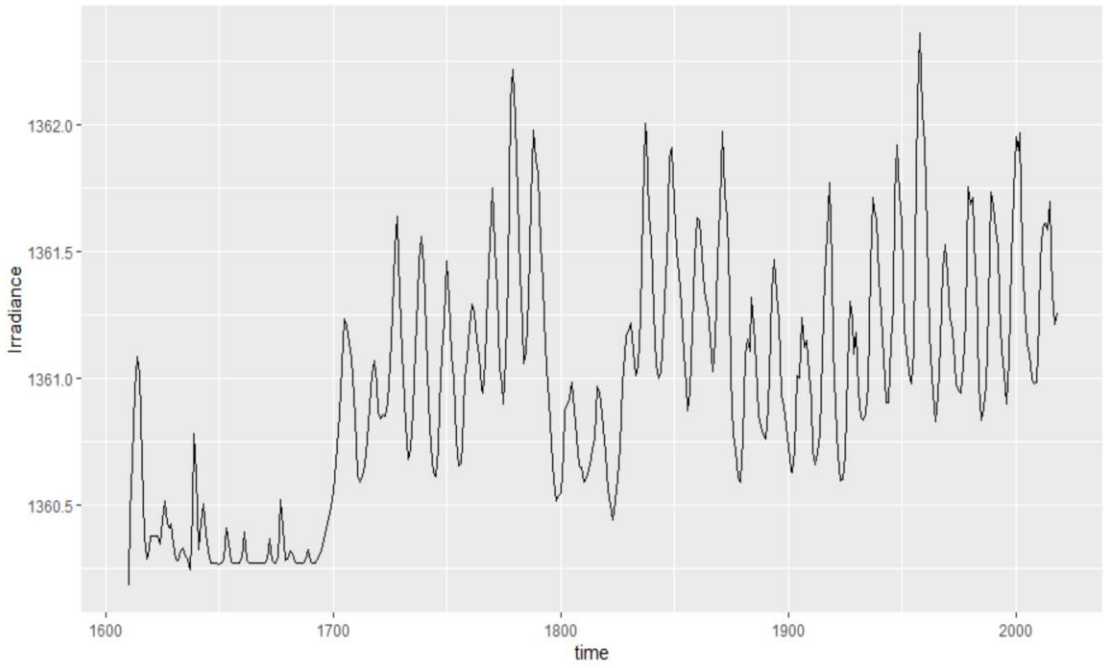

Figure 2.1

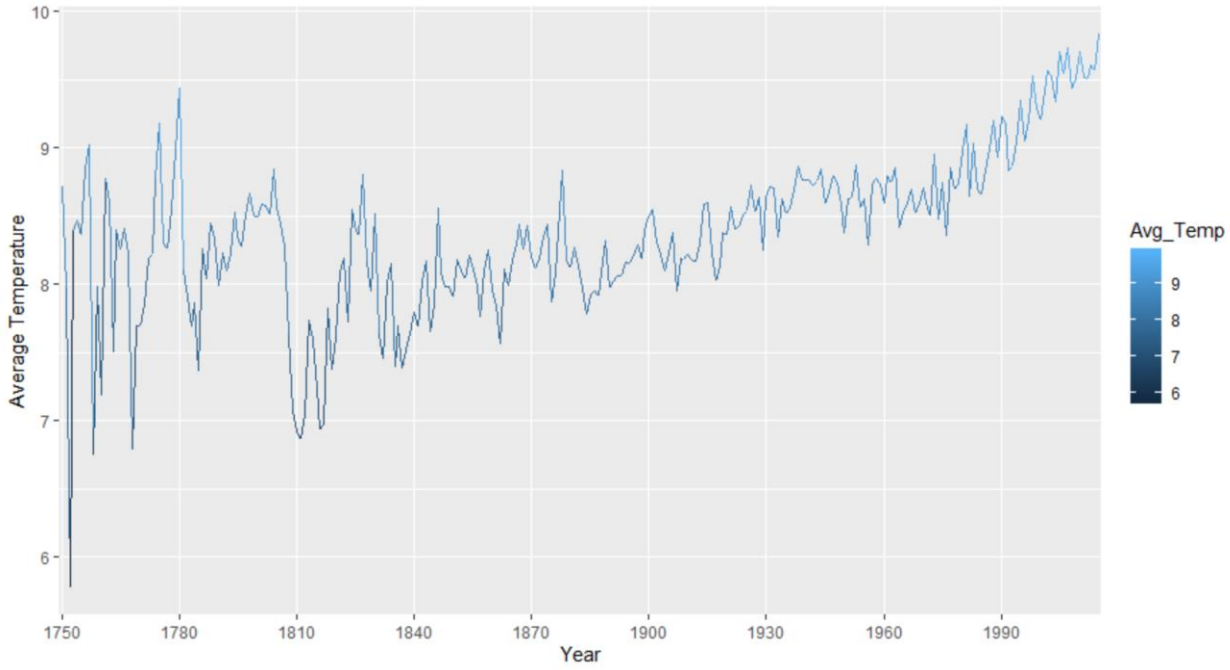

Figure 2.2

The original land average temperature dataset records the monthly average temperature of different countries, I applied some data transformation (Figure 2.3) to calculate the yearly average temperature to generate the Figure 2.2

```
attach(globalTemp)
globalTemp$Year <- format(as.Date(globalTemp$dt),"%Y")
globalTemp <- globalTemp%>%group_by(Year)%>%summarize(Avg_Temp=mean(LandAverageTemperature,na.rm=T))
ggplot(globalTemp, aes(x=globalTemp$Year, y=globalTemp$Avg_Temp,
                       color=Avg_Temp,group= 1))+geom_line()+
  labs(y= "Average Temperature", x = "Year")+scale_x_discrete(breaks = seq(1750, 2013, by = 30))
```

Figure 2.3

Next, I create plots that display the three major greenhouse gases emission (Figure 2.4, Figure 2.5, Figure 2.6) of United States, China, India, United Kingdom, France, Australia, Japan and Germany. In order to observe the relationship of greenhouse gases emission and temperature, I also plot the temperature of these countries since 1960 (Figure 2.7). The reason I prefer using emission data by country to the total emission by all countries is that some countries may increase their greenhouse gases emission while other countries may keep the emission constant, so we can see if there is difference in how the temperature changes for these countries from the plots.

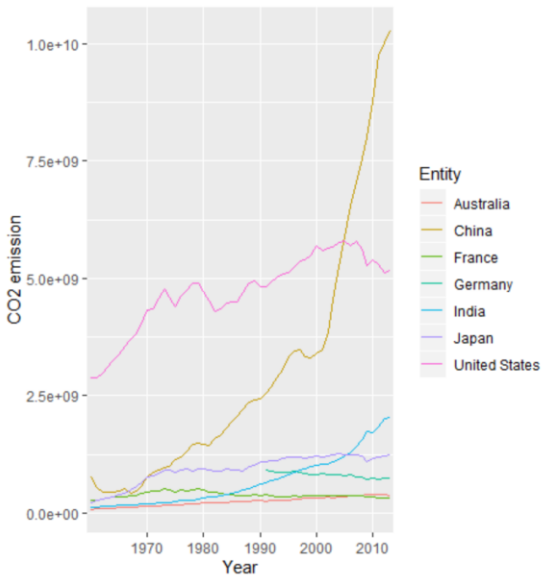

Figure 2.4

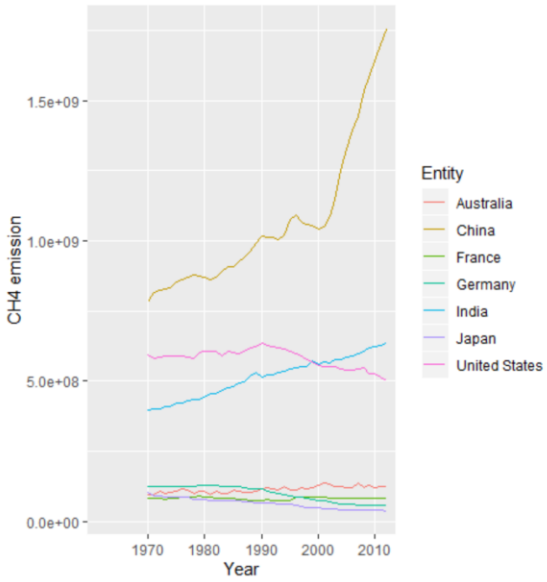

Figure 2.5

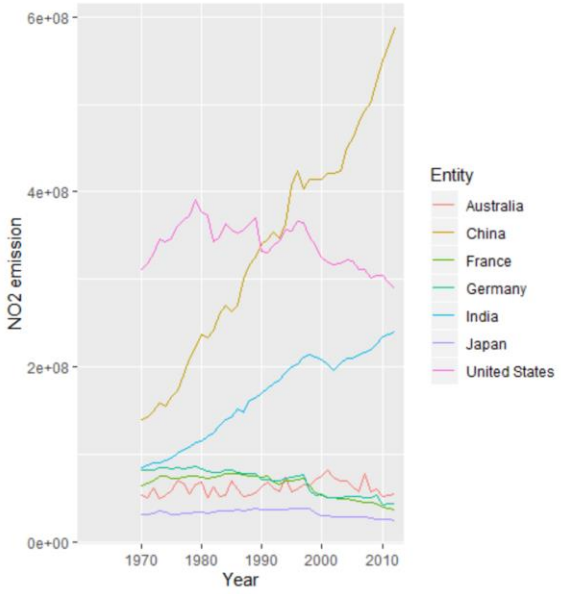

Figure 2.6

Figure 2.7

Models:

ARIMA:

The first model used for this dataset is ARIMA model. ARIMA model stands for autoregressive integrated moving average model, and it is a popular forecasting model that utilize historical data to make predictions. Our land temperature dataset is a time series dataset and we will only use this single dataset for the ARIMA model. We will use the yearly average temperature data for United States for this model.

For the ARIMA model, there are three major parameters[3]:

- p: The number of lag observations included in the model, also called the lag order.
- d: The number of times that the raw observations are differenced, also called the degree of differencing.
- q: The size of the moving average window, also called the order of moving average.

There are some missing data and outliers in this dataset, so we can use the teclean() function to replace the missing value and outliers using series smoothing and decomposition. Figure 3.1 plots the dataset before data cleaning and figure 3.2 shows the dataset after the data cleaning.

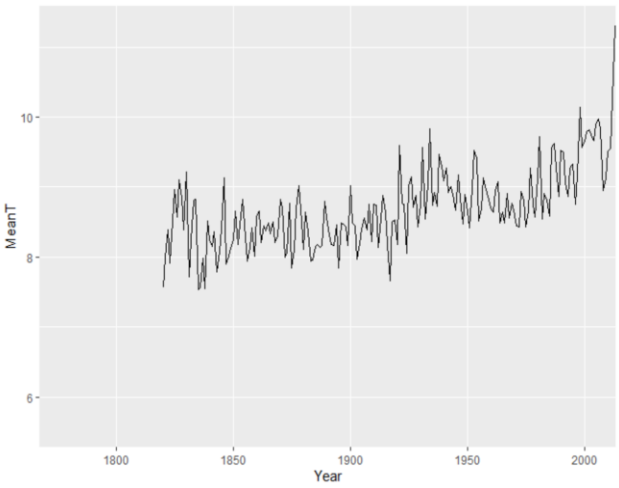

Figure 3.1

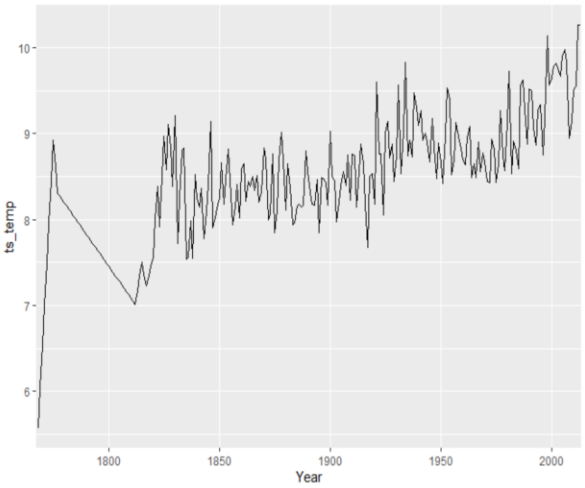

Figure 3.2

Data are required to be stationary to fit an ARIMA model, and stationary means that the mean, variance and autocovariance of a series are time invariant. We can use ADF test to test whether a series is stationary. From Figure 3.3, we can see that the data is stationary.

```
        Augmented Dickey-Fuller Test

data:  ts_temp
Dickey-Fuller = -3.4994, Lag order = 6, p-value = 0.04341
alternative hypothesis: stationary
```

Figure 3.3

Next, we need to find three parameters p, d and q for the model. The R package we used provides us with the function auto.arima() to generate the optimal combination of p, d and q for the model. Figure 3.4 shows the parameters of the fitted model generated by auto.arima() function and Figure 3.5 plots the model diagnostics.

```
Series: deseasonal_tmp[c(1:236)]
ARIMA(4,1,3)

Coefficients:
         ar1      ar2     ar3      ar4      ma1     ma2      ma3
      1.7019  -0.7867  0.3275  -0.2617  -2.1663  1.3736  -0.2035
s.e.  0.1735   0.2868  0.1414   0.0753   0.1762  0.3530   0.1777

sigma^2 estimated as 0.1375:  log likelihood=-97.55
AIC=211.1   AICc=211.73   BIC=238.77
> 
```

Figure 3.4

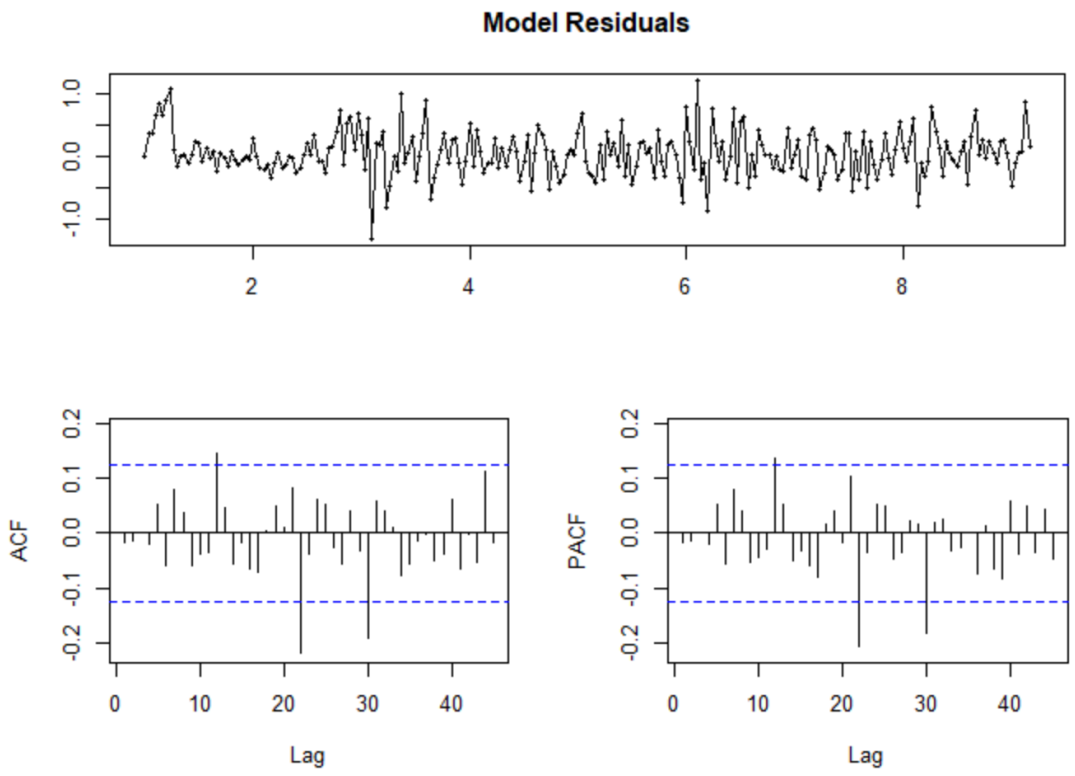

Figure 3.5

I use the first 236 data frames to fit the model, so I can use the model to make the predictions and compare the predictions to the real value. Figure 3.6 shows the statistical predictions. Figure 3.7 shows the predictions and real values. Predictions is the blue line in the end, and predictions are provided with confidence bounds: 80% confidence limits shaded in darker grey, and 95% in lighter grey.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
237	9.599839	9.124654	10.075024	8.873106	10.32657
238	9.569420	9.030352	10.108488	8.744986	10.39385
239	9.528752	8.967026	10.090477	8.669667	10.38784
240	9.459273	8.853794	10.064751	8.533273	10.38527
241	9.400547	8.755933	10.045161	8.414695	10.38640
242	9.349899	8.679102	10.020695	8.324004	10.37579
243	9.297782	8.604195	9.991369	8.237032	10.35853
244	9.247877	8.533942	9.961811	8.156007	10.33975
245	9.202720	8.472363	9.933076	8.085736	10.31970
246	9.161311	8.417663	9.904958	8.023999	10.29862

Figure 3.6

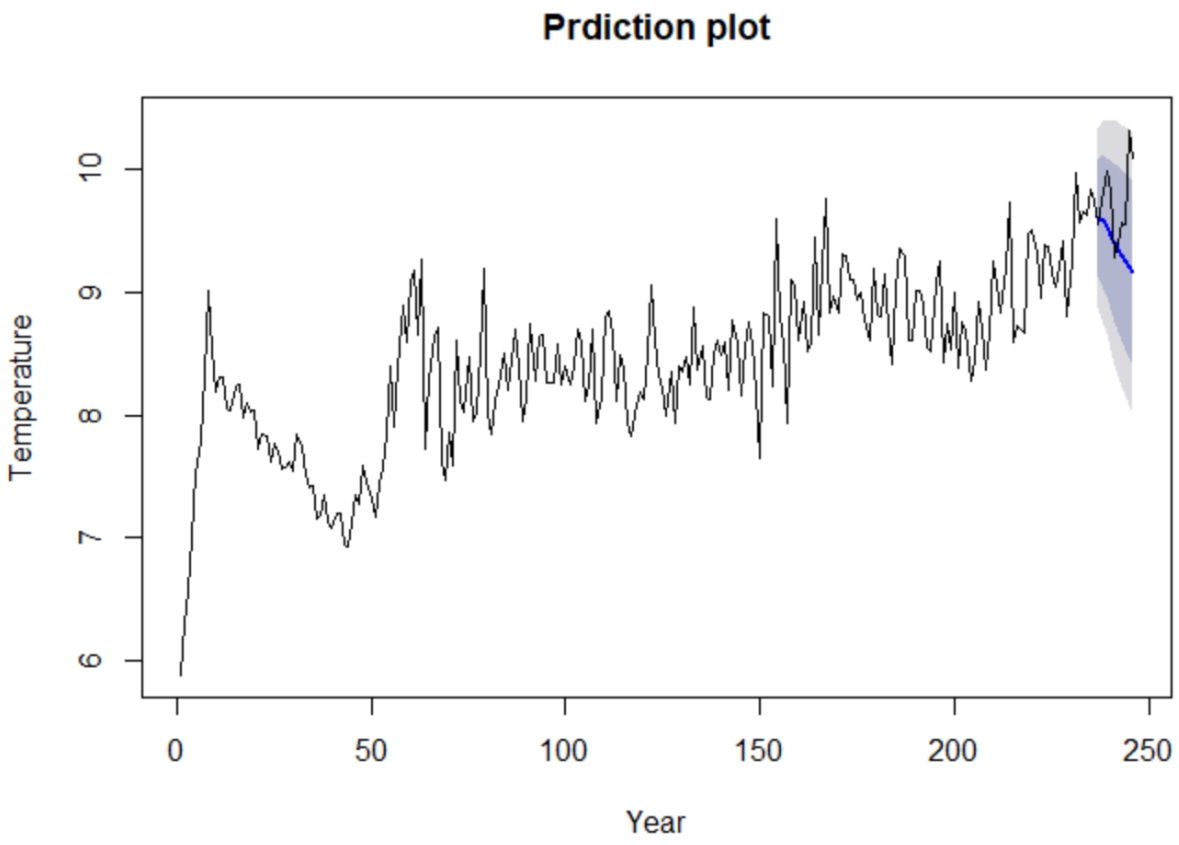

Figure 3.7

Regression model:

The new dataset after the merge of greenhouse gas emission dataset and land temperature dataset that mentioned above will be used to fit the linear regression model.

We cannot use all the data frames in our datasets for the regression model. Since the temperature is also influenced by other factors like the latitude and longitude of a country. Therefore, two countries which have the same greenhouse gases emission may have completely different temperatures. Since our independent variable for the regression model will only be the CO2, CH4 and NO2, we need to use temperature data of countries that have similar temperature.

From Figure 2.7 in the Analysis part above, we can see that Germany, China and US share similar temperatures. So we will use the land temperature data of Germany, China and US. Since there are some missing values in the dataset, we replaced all the NA in the dataset and smooth the data. Figure 3.8 shows the diagram of CO2, CH4 and NO2.

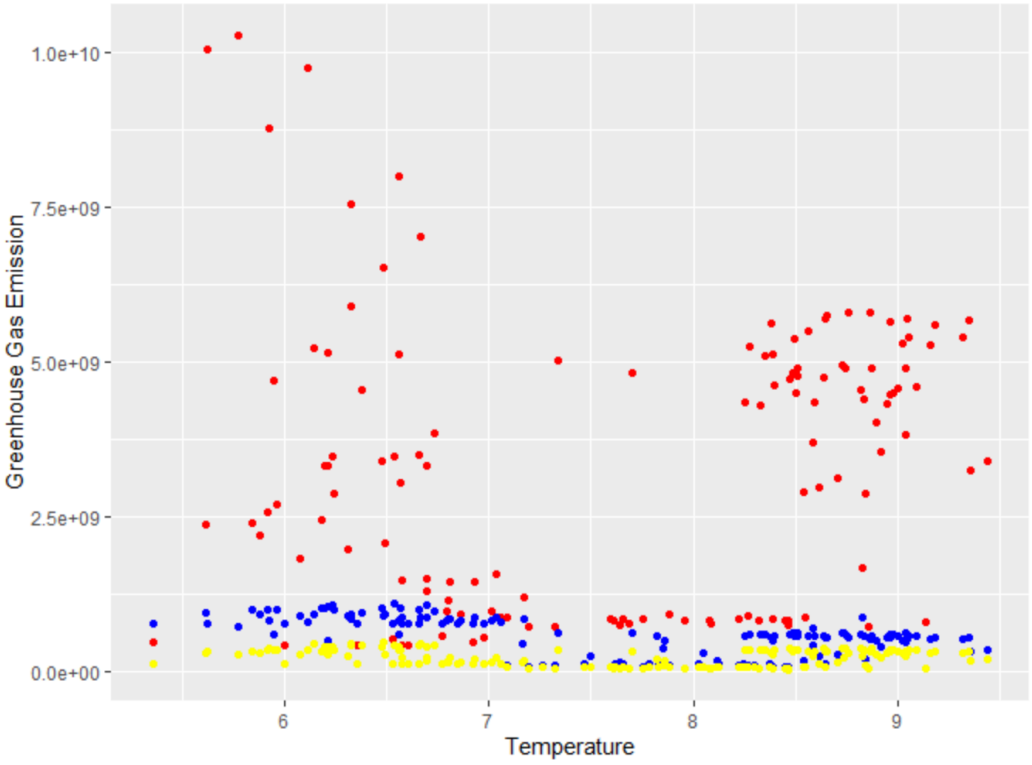

Figure 3.8 (red: CO2, blue: CH4, yellow: NO2)

I split the data into training dataset and testing dataset. The statistical coefficients of the trained regression model is showed as Figure 3.9 and plots of the model are shown as Figure 3.10, Figure 3.11, Figure 3.12 and Figure 3.13

```
Call:
lm(formula = MeanT ~ NO2 + CH4 + CO2, data = trainingData)

Residuals:
   Min     1Q Median     3Q    Max
-9.190 -4.336 -1.795  5.394 14.175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.420e+01  1.737e+00  13.932  < 2e-16 ***
NO2          2.596e-08  9.924e-09   2.616  0.00987 **
CH4         -1.543e-08  2.769e-09  -5.571 1.24e-07 ***
CO2         -2.544e-09  3.640e-10  -6.988 1.02e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.621 on 141 degrees of freedom
Multiple R-squared:  0.4888,    Adjusted R-squared:  0.4779
F-statistic: 44.94 on 3 and 141 DF,  p-value: < 2.2e-16
```

Figure 3.9

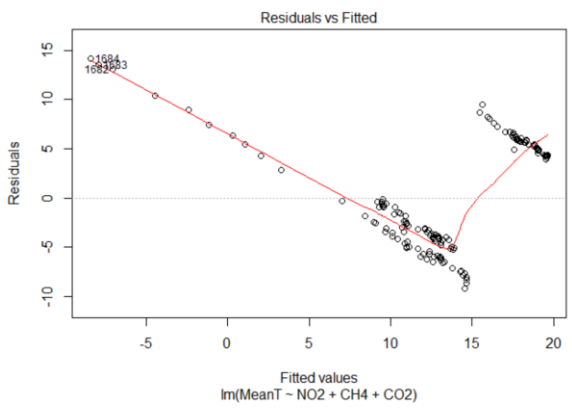

Figure 3.10

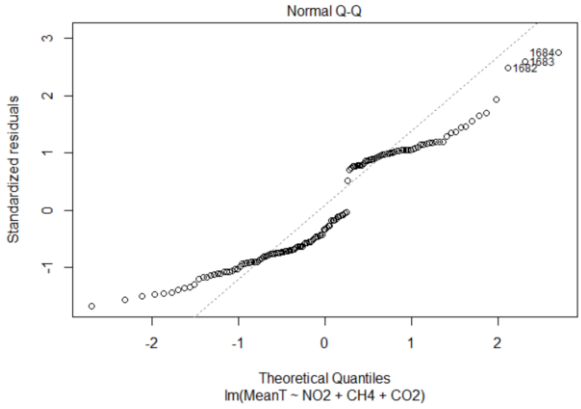

Figure 3.11

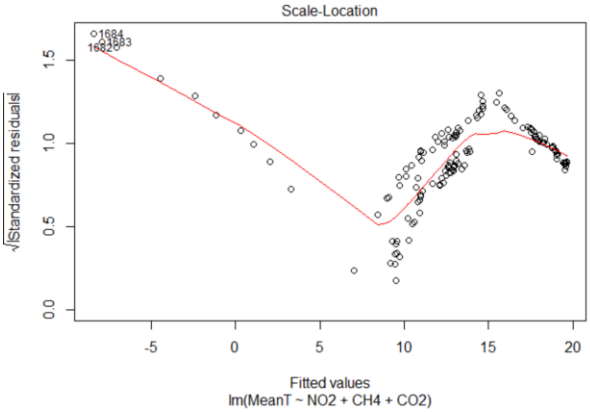

Figure 3.12

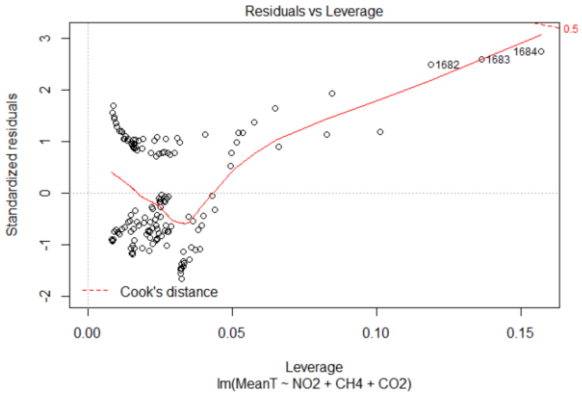

Figure 3.13

From the Figure 3.9, we can see that the P-value of all my variables are less than 0.05. Therefore, my linear regression model is statistically significant.

Next, I use the regression to predict the values in the Testing dataset. The plot of the predictions and actual value are shown as Figure 3.14. The prediction value is blue line and actual value is red line.

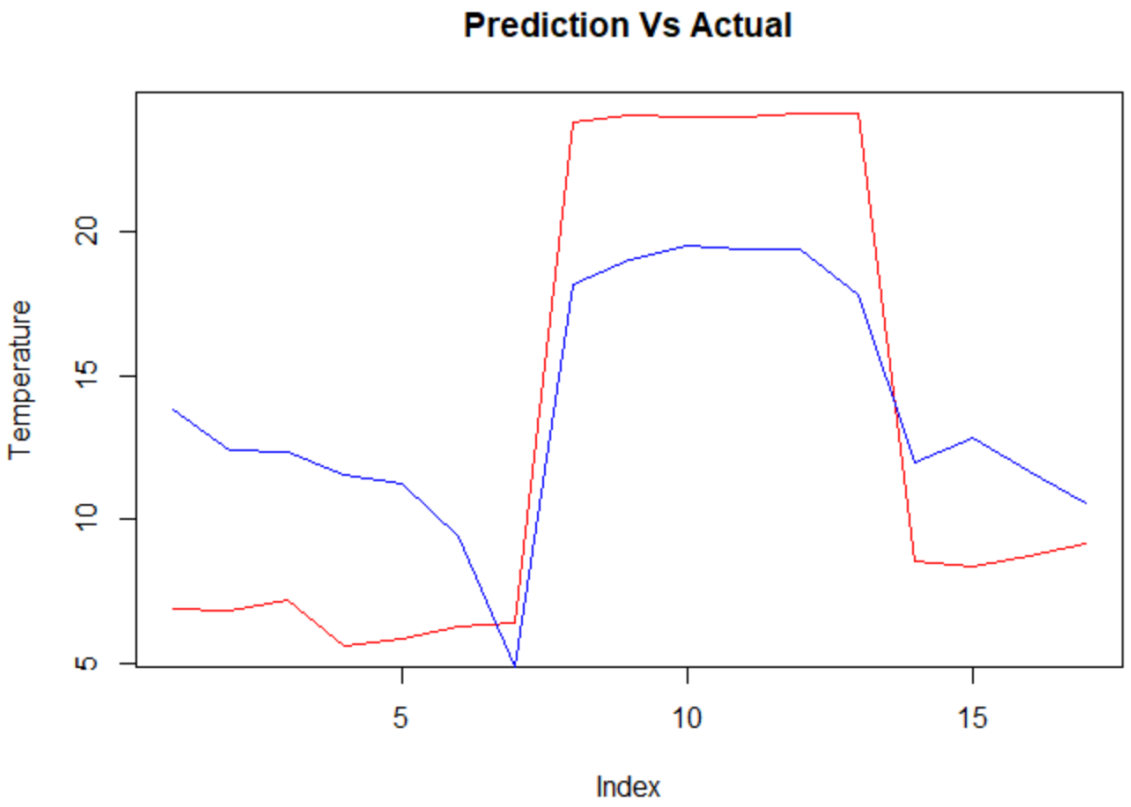

Figure 3.14

After we have the regression model, I want to predict the future temperature to see if the global warming will worsen in the future. I find the projected greenhouse gas emission of 2025 in the 2016 second biennial report of the united states(CO2: 5305000000, CH4:674000000,NO2: 335000000)[4]. I will use the projected emission to predict the temperature of United States in 2025.

The predicted temperature of 2025 is 9.00289 degree, which is higher than the temperature in 2012 and 2013 according to Figure 3.15

7649	United States	2011	9.023250
7650	United States	2012	8.390542
7651	United States	2013	6.213722

Figure 3.15

Conclusion:

For the ARIMA model, we can see that the prediction is not very accurate and the trend of predicted temperature changes (blue line) goes downward from figure 4.1. One of the possible reasons is that it is difficult to predict the temperature purely based on the historical data. There does not exists a clear pattern of the temperature changes. Temperature fluctuated throughout the whole period and rapidly increases in the past 50 years. The model tried to find trend component and cycle component (through decomposition) of this dataset, but both trend component and cycle component of this dataset is not very clear. So the prediction is not very accurate.

In addition, time series dataset typically has seasonality, and our dataset also have monthly temperature data which shows seasonality. In the future, we can include a seasonal component in our model to improve the model. We can also try some other parameters of the ARIMA model to see if there is improvement in the accuracy.

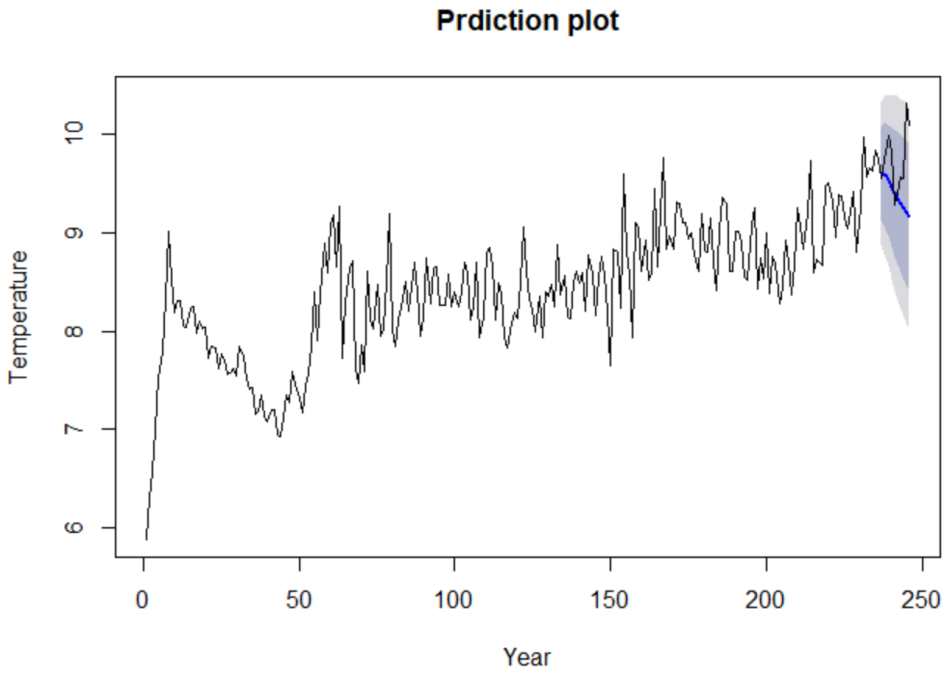

Figure 4.1

For the linear regression model, although there are differences between predicted values and actual values, the trend of the temperature changes is similar for predicted values and actual values. The possible reason is that there are a little difference between predicted values and actual values are the there are not sufficient training data for the model. The data is insufficient since I only use the countries that have similar temperature for the training data and the

greenhouse gas emission dataset only includes the emission data after 1960. In addition, I initially want to include three other greenhouse gases SFA, PFC and HFC in my model, but the dataset I found for these three gases are very incomplete (missing a lot of data). Therefore, I cannot build a regression model that uses more input variables. If I can find datasets that have more data, the model accuracy may become higher.

Reference:

1. *Amanda MacMillan,* Global Warming 101, https://www.nrdc.org/stories/global-warming-101
2. Historical Total Solar Irradiance Reconstruction, Time Series, https://lasp.colorado.edu/lisird/data/historical_tsi/
3. *Jason Brownlee, How to Create an ARIMA Model for Time Series Forecasting in Python,* https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
4. 2016 SECOND BIENNIAL REPORT of the United States of America, https://unfccc.int/files/national_reports/biennial_reports_and_iar/submitted_biennial_reports/application/pdf/2016_second_biennial_report_of_the_united_states_.pdf

R scripts:

```
library("plyr")
library(dplyr)
library(ggplot2)

solar <- read.csv("C:/Users/dingj/Desktop/dsproject/historical_tsi.csv") ##the data for solar irradiance
temp<-read.csv("C:/Users/dingj/Desktop/dsproject/GlobalLandTemperaturesByCountry.csv") ## data for land monthly average tem
globalTemp <- read.csv("C:/Users/dingj/Desktop/dsproject/GlobalTemperatures.csv")
ghg <- read.csv("C:/Users/dingj/Desktop/dsproject/greenhouse-gas-emissions-by-gas.csv") ## data for greenhouse gas emission

## visualize the solar irradiance plot
head(solar)
attach(solar)
ggplot(solar, aes(x=time, y=Irradiance))+geom_line()


## dataset that after merge of greenhouse gases emission dataset and average land temperature

dates<- format(as.Date(temp$dt), format = "%Y")

temp2 <- ddply(temp, .( Country,dates), summarise,
               mean_TSO4 = mean(AverageTemperature),
               mean_TNO3 = mean(AverageTemperatureUncertainty))
colnames(temp2) <- c("Entity", "Year","MeanT","MeanUnT")
total <- merge(temp2,ghg,by=c("Entity","Year"))
colnames(total) <- c("Entity", "Year","MeanT","MeanUnT",
                     "Code","SFA","PFC","HFC","NO2","CH4","CO2")
# plot the global average temperature

attach(globalTemp)
globalTemp$Year <- format(as.Date(globalTemp$dt),"%Y")

globalTemp <- globalTemp%>%group_by(globalTemp$Year)%>%summarize(Avg_Temp=mean(globalTemp$LandAverageTemperature,na.rm=T))
ggplot(globalTemp, aes(x=globalTemp$Year, y=globalTemp$Avg_Temp,
                       color=Avg_Temp,group= 1))+geom_line()+
  labs(y= "Average Temperature", x = "Year")+scale_x_discrete(breaks = seq(1750, 2013, by = 30))
```

```r
##plot the temperature for different countries
colnames(temp2) <- c("Entity", "Year","MeanT","MeanUnT")

attach(temp2)
temp3 <- temp2 %>% filter(Year>=1850 & Year<= 2012)
## explore the temperature dataset
library(scales)



## exploration on greenhouse gas
attach(total)
x <- subset(total,total$Entity=="United States")

plot(x$CO2,type = "l",col = "red", xlab = "Month", ylab = "Rain fall",
     main = "Rain fall chart")

x <- subset(total,total$Entity=="United States"|
              total$Entity== "China"|total$Entity=="India"|
              total$Entity== "United Kindom"|total$Entity== "France"|
              total$Entity== "Australia"|total$Entity== "Japan"|
               total$Entity== "Germany")
## plot of the temperature
ggplot(aes(x = Year, y = MeanT,group =Entity), data = x)+geom_line(aes(color=Entity))+
  labs(y= "Average Temperature", x = "Year")+scale_x_discrete(breaks=seq(1960, 2012, 10))

## plots of six major greenhouse gases
ggplot(aes(x = Year, y = CO2,group =Entity  ), data = x)+geom_line(aes(color=Entity))+
  labs(y= "CO2 emission", x = "Year")+scale_x_discrete(breaks=seq(1970, 2012, 10))

ggplot(aes(x = Year, y = CH4,group =Entity  ), data = x)+geom_line(aes(color=Entity))+
  labs(y= "CH4 emission", x = "Year")+scale_x_discrete(breaks=seq(1970, 2012, 10))

ggplot(aes(x = Year, y = NO2,group =Entity  ), data = x)+geom_line(aes(color=Entity))+
  labs(y= "NO2 emission", x = "Year")+scale_x_discrete(breaks=seq(1970, 2012, 10))
```

```r
### ARIMA Modelling
install.packages("tseries")
install.packages("forecast")
library(tseries)
library(forecast)
## use models on the temperature for United States
US_temp <- subset(temp,temp$Country=="United States")
US_temp <- subset(temp2,temp2$Entity=="United States")
attach(US_temp)
## clearn the dataset e.g outliers missing NA
US_temp$clean_temp <- tsclean(MeanT)
ts_temp = ts(na.omit(US_temp$clean_temp), frequency=30)
ggplot(US_temp, aes(x=Year, y=MeanT,group=1))+geom_line()+scale_x_discrete(breaks=seq(1750, 2012, 50))

ggplot() +
  geom_line(data = US_temp, aes(x = Year, y = MeanT, group=1)) +scale_x_discrete(breaks=seq(1750, 2012, 50))
ggplot( ) +
  geom_line(data = US_temp, aes(x = Year, y = ts_temp,group=1))+scale_x_discrete(breaks=seq(1750, 2012, 50))
decomp = stl(ts_temp, s.window="periodic")
plot(decomp)
deseasonal_tmp <- seasadj(decomp)
## test if stationary
adf.test(ts_temp, alternative = "stationary")
## find parameters using auo arima
deseasonal_tmp
## use the first 236 as training dataset and last 10 as test
fit<-auto.arima(deseasonal_tmp[c(1:236)], seasonal=FALSE)
tsdisplay(residuals(fit), lag.max=45, main='Model Residuals')
fit
fcast <- forecast(fit, h=10)
fcast
plot(fcast, main="Prdiction plot",ylab="Temperature",xlab="Year")
lines(ts(deseasonal_tmp))
```

```r
## Regression model
## using the dataset after mnerge
attach(total)
regression_temp <- subset(total,total$Entity=="United States"|
                    total$Entity== "China"|
                    total$Entity== "Germany")
## clean the dataset
attach(regression_temp)
## plot before data cleaning

regression_temp$NO2 <- tsclean(NO2)
regression_temp$CH4 <- tsclean(CH4)
ggplot() +
  geom_point(data = regression_temp, aes(x = MeanT, y = CO2, group=1),colour="red") +
  geom_point(data = regression_temp, aes(x = MeanT, y = CH4, group=1),colour="blue")+
  geom_point(data = regression_temp, aes(x = MeanT, y = NO2, group=1),colour="yellow")+
  labs(y= "Greenhouse Gas Emission", x = "Temperature")

trainingRowIndex <- sample(1:nrow(regression_temp), 0.9*nrow(regression_temp))
trainingRowIndex
trainingData <- regression_temp[trainingRowIndex, ]
testData  <- regression_temp[-trainingRowIndex, ]
regression <- lm(MeanT~NO2+CH4+CO2,data= trainingData)
summary(regression)
plot(regression)

myvars <- c("CO2", "CH4", "NO2")

pred <- predict.lm(regression,newdata = testData[myvars])
plot(pred)
actuals_preds <- data.frame(cbind(actuals=testData$MeanT, predicteds=pred))
pred
actual_temp <- testData$MeanT
plot(actual_temp,type='l',col = "red",main = "Prediction Vs Actual",ylab="Temperature")
lines(pred,col="blue")
abline(a=0,b=1)
## predict the temperature in 2025
data2025 <- data.frame(CO2=5305000000,CH4=674000000,NO2=335000000)
temp2025 <-predict.lm(regression,newdata = data2025)
temp2025
```