Air Quality Prediction

#Installing or importing all packages  
library(ggplot2)  
library(expss)

##   
## Use 'expss\_output\_viewer()' to display tables in the RStudio Viewer.  
## To return to the console output, use 'expss\_output\_default()'.

##   
## Attaching package: 'expss'

## The following object is masked from 'package:ggplot2':  
##   
## vars

library(VIF)  
library(forecast)

## Registered S3 method overwritten by 'xts':  
## method from  
## as.zoo.xts zoo

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

## Registered S3 methods overwritten by 'forecast':  
## method from   
## fitted.fracdiff fracdiff  
## residuals.fracdiff fracdiff

library(Metrics)

##   
## Attaching package: 'Metrics'

## The following object is masked from 'package:forecast':  
##   
## accuracy

library(DMwR)

## Loading required package: lattice

## Loading required package: grid

#Read the airquality dataset obtained from AQS Data Mart  
air.data<-read.csv(file='ProjectDataset.csv')  
#Remove all missing values  
airqualitys.data<-na.omit(air.data)  
#Create a subset of the dataset by selecting only the required columns for analysis  
myvar<-c(1,4,5,9,11,14,16,19,20,21,22,23,24,30,31,32,33,34,35,42,51)  
airqualitysubset.data<-airqualitys.data[myvar]  
  
#Summary Statistics for Average Pollution  
summary(airqualitysubset.data$X1st.Max.Value)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.200 0.900 1.200 1.466 1.700 12.800

#Summary Statistics for % of observations  
summary(airqualitysubset.data$Observation.Percent)

## Length Class Mode   
## 0 NULL NULL

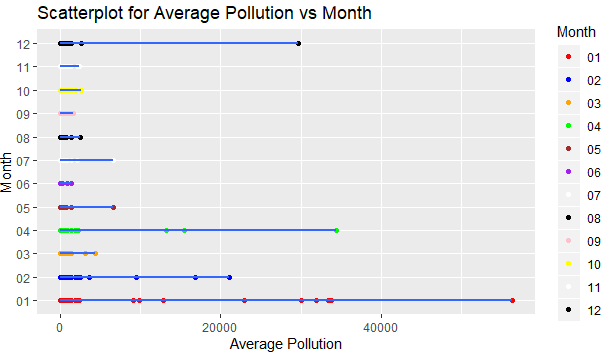
#Summary Statistics for Primary Exceedance Count  
summary(airqualitysubset.data$Primary.Exceedance.Count)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.04422 0.00000 7.00000

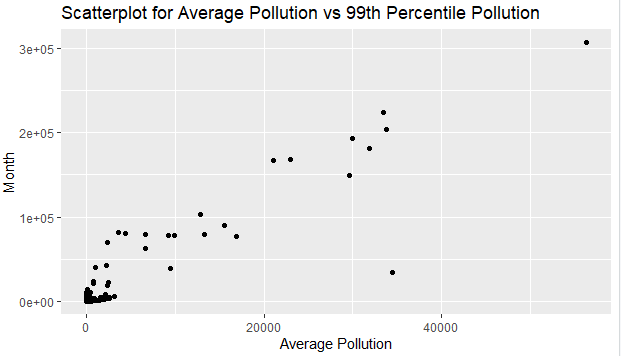
#Extracting month number from DateTime field - The 1st Max DateTime field gives the date when the highest value of pollutant is recorded  
myDate = as.POSIXct(air.data$X1st.Max.DateTime)  
maxmonthnumber<-format(myDate,"%m")  
airqualitysubset.data<-transform(air.data, Month =maxmonthnumber)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Scatter Plots\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

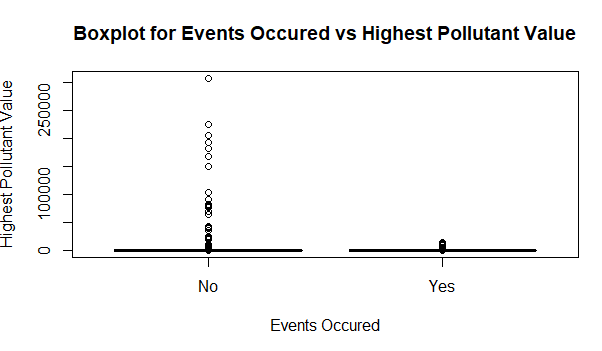
#Scatterplot for Average Pollution vs Month  
ggplot(airqualitysubset.data, aes(x = airqualitysubset.data$Arithmetic.Mean , y = Month)) +  
 geom\_point(aes(color=Month))+  
 geom\_smooth(method="loess",se=TRUE,fullrange=FALSE,level=0.95) +   
 scale\_color\_manual(values=c("red","blue","orange","green","brown","purple","white","black","pink","yellow","white","black"))+  
 xlab("Average Pollution")+ylab("Month")+  
 labs(title="Scatterplot for Average Pollution vs Month ")



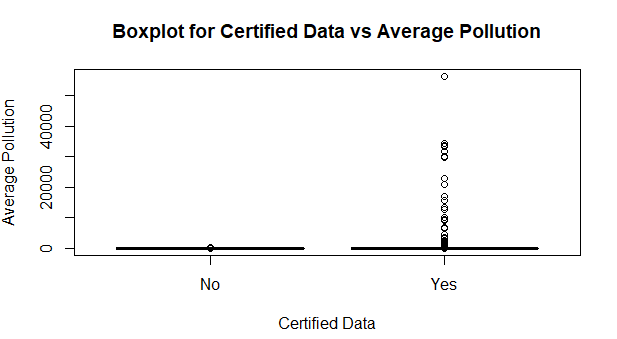
#Scatterplot for Average Pollution vs 99th Percentile value of Pollutant  
ggplot(airqualitysubset.data, aes(x = airqualitysubset.data$Arithmetic.Mean , y = airqualitysubset.data$X1st.Max.Value)) +  
 geom\_point()+  
 geom\_smooth(method="loess",se=TRUE,fullrange=FALSE,level=0.95) +   
 xlab("Average Pollution")+ylab("Month")+  
 labs(title="Scatterplot for Average Pollution vs 99th Percentile Pollution")



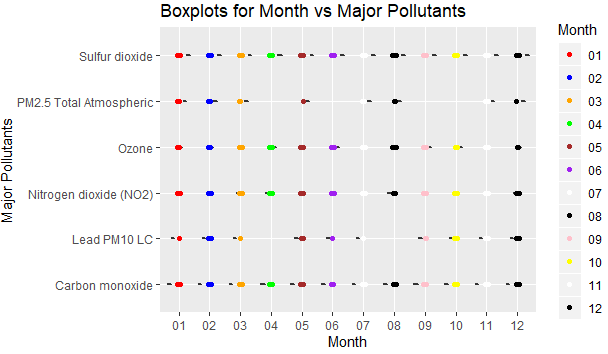
#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Box Plots\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#  
  
#Boxplot for Events Occurred vs Highest pollutant Value  
boxplot(airqualitysubset.data$X1st.Max.Value~airqualitysubset.data$Events.Occurred,xlab="Events Occured",ylab="Highest Pollutant Value",main="Boxplot for Events Occured vs Highest Pollutant Value")



#Boxplot for Average pollution vs Certified Pollution Data  
boxplot(airqualitysubset.data$Arithmetic.Mean~airqualitysubset.data$Certified,xlab="Certified Data",ylab="Average Pollution",main="Boxplot for Certified Data vs Average Pollution")



#Boxplot for Month vs Major Pollutants  
airqualitymonth.data<-subset(airqualitysubset.data, Parameter.Name %in% c("Ozone","Carbon monoxide","Nitrogen dioxide (NO2)","Lead PM10 LC","Sulfur dioxide","PM2.5 Total Atmospheric"))  
ggplot(airqualitymonth.data, aes(x=Month, y=Parameter.Name)) +  
 geom\_boxplot(fill="white")+  
 geom\_jitter(aes(color=Month),  
 position=position\_jitter(w=0.05, h=0))+  
 scale\_color\_manual(values = c("red","blue","orange","green","brown","purple","white","black","pink","yellow","white","black"))+  
 xlab("Month")+ylab("Major Pollutants")+  
 labs(title="Boxplots for Month vs Major Pollutants")



#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Correlation Analysis\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#  
  
#Correlation analysis between average pollution and highest pollutant value  
cor.test(airqualitysubset.data$Arithmetic.Mean, airqualitysubset.data$X1st.Max.Value, method = "pearson", use = "complete.obs")

##   
## Pearson's product-moment correlation  
##   
## data: airqualitysubset.data$Arithmetic.Mean and airqualitysubset.data$X1st.Max.Value  
## t = 505.9, df = 66358, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8895502 0.8926834  
## sample estimates:  
## cor   
## 0.8911274

#Correlation analysis between 99th percentile value of pollutant and highest pollutant value  
cor.test(airqualitysubset.data$X99th.Percentile, airqualitysubset.data$X1st.Max.Value, method = "pearson", use = "complete.obs")

##   
## Pearson's product-moment correlation  
##   
## data: airqualitysubset.data$X99th.Percentile and airqualitysubset.data$X1st.Max.Value  
## t = 1059.7, df = 66358, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9712739 0.9721230  
## sample estimates:  
## cor   
## 0.9717016

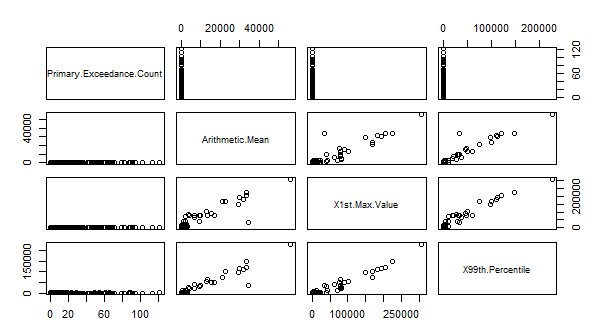
#Correlation analysis between the number of days daily monitoring criteria were met(Valid Day Count) and number of samples that exceeded the air quality standard(Primary Exceedence Count)  
cor(airqualitysubset.data$Valid.Day.Count, airqualitysubset.data$Primary.Exceedance.Count, method = "pearson", use = "complete.obs")

## [1] 0.1349807

#Correlation analysis between Longitude and number of samples that exceeded the air quality standard(Primary Exceedence Count)  
cor.test(airqualitysubset.data$Longitude, airqualitysubset.data$Primary.Exceedance.Count, method = "pearson", use = "complete.obs")

##   
## Pearson's product-moment correlation  
##   
## data: airqualitysubset.data$Longitude and airqualitysubset.data$Primary.Exceedance.Count  
## t = -27.129, df = 11996, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2572138 -0.2234944  
## sample estimates:  
## cor   
## -0.2404266

#Identify relationships between Primary Exceedance Count, Arithmetic Mean, 1st Max Value, 99th Percentile  
myvar<-c(24,28,30,42)  
airqualitypairs.data<-airqualitysubset.data[myvar]  
pairs(airqualitypairs.data)



#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Regression Analysis\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#  
  
  
#Regression Analysis using single predictor(independent) variable  
lmodel1<-lm(airqualitysubset.data$X1st.Max.Value~airqualitysubset.data$X99th.Percentile)  
summary(lmodel1)

##   
## Call:  
## lm(formula = airqualitysubset.data$X1st.Max.Value ~ airqualitysubset.data$X99th.Percentile)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50807 0 5 5 52939   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -5.063552 2.311477 -2.191  
## airqualitysubset.data$X99th.Percentile 1.572227 0.001484 1059.688  
## Pr(>|t|)   
## (Intercept) 0.0285 \*   
## airqualitysubset.data$X99th.Percentile <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 594.8 on 66358 degrees of freedom  
## Multiple R-squared: 0.9442, Adjusted R-squared: 0.9442   
## F-statistic: 1.123e+06 on 1 and 66358 DF, p-value: < 2.2e-16

distPred\_lmodel1 <- predict(lmodel1,airqualitysubset.data)

#Root Mean Sqaure Error  
rmse(airqualitysubset.data$X1st.Max.Value,predict(lmodel1,airqualitysubset.data))

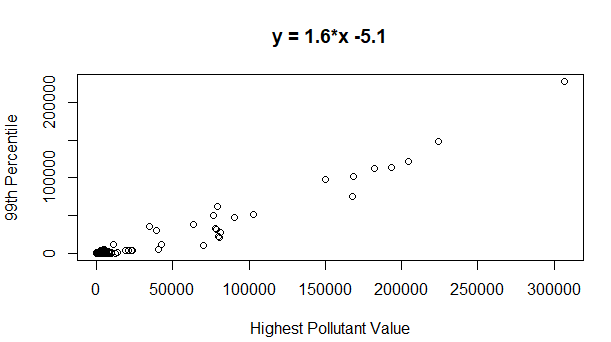
## [1] 594.757

#Coefficients of the linear model  
lmodel1coef=coef(lmodel1)  
print(lmodel1coef)

## (Intercept)   
## -5.063552   
## airqualitysubset.data$X99th.Percentile   
## 1.572227

#Equation of the linear regression  
eq = paste0("y = ", round(lmodel1coef[2],1), "\*x ", round(lmodel1coef[1],1))

#Plotting the graph  
plot(airqualitysubset.data$X1st.Max.Value,airqualitysubset.data$X99th.Percentile, main=eq, xlab="Highest Pollutant Value",ylab="99th Percentile”)



#Confidence Interval of the model  
confint(lmodel1)

## 2.5 % 97.5 %  
## (Intercept) -9.594046 -0.5330578  
## airqualitysubset.data$X99th.Percentile 1.569319 1.5751353

#Analysis of Variance  
anova(lmodel1)

## Analysis of Variance Table  
##   
## Response: airqualitysubset.data$X1st.Max.Value  
## Df Sum Sq Mean Sq F value  
## airqualitysubset.data$X99th.Percentile 1 3.9724e+11 3.9724e+11 1122940  
## Residuals 66358 2.3474e+10 3.5375e+05   
## Pr(>F)   
## airqualitysubset.data$X99th.Percentile < 2.2e-16 \*\*\*  
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Regression Evaluation  
actuals\_pred\_lmodel1 <- data.frame(cbind(lmodel1\_actuals=airqualitysubset.data$X1st.Max.Value, lmodel1\_predicteds=distPred\_lmodel1))  
regr.eval(actuals\_pred\_lmodel1$lmodel1\_actuals, actuals\_pred\_lmodel1$lmodel1\_predicteds)

## mae mse rmse mape   
## 51.07865 353735.85300 594.75697 Inf

#Regression Analysis using multiple predictor(independent)variables  
lmodel2<-lm(airqualitysubset.data$X1st.Max.Value~airqualitysubset.data$X99th.Percentile+airqualitysubset.data$Arithmetic.Mean+airqualitysubset.data$State.Code+airqualitysubset.data$Observation.Count+airqualitysubset.data$Parameter.Code+airqualitysubset.data$Valid.Day.Count+airqualitysubset.data$Longitude)  
summary(lmodel2)

##   
## Call:  
## lm(formula = airqualitysubset.data$X1st.Max.Value ~ airqualitysubset.data$X99th.Percentile +   
## airqualitysubset.data$Arithmetic.Mean + airqualitysubset.data$State.Code +   
## airqualitysubset.data$Observation.Count + airqualitysubset.data$Parameter.Code +   
## airqualitysubset.data$Valid.Day.Count + airqualitysubset.data$Longitude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53747 -30 -10 15 52525   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -2.122e+01 1.584e+01 -1.340  
## airqualitysubset.data$X99th.Percentile 1.732e+00 4.021e-03 430.713  
## airqualitysubset.data$Arithmetic.Mean -5.932e-01 1.388e-02 -42.736  
## airqualitysubset.data$State.Code -5.939e-01 1.428e-01 -4.158  
## airqualitysubset.data$Observation.Count 3.231e-03 3.973e-04 8.133  
## airqualitysubset.data$Parameter.Code 6.682e-04 1.064e-04 6.282  
## airqualitysubset.data$Valid.Day.Count 1.123e-01 2.008e-02 5.593  
## airqualitysubset.data$Longitude 2.254e-01 1.347e-01 1.673  
## Pr(>|t|)   
## (Intercept) 0.1803   
## airqualitysubset.data$X99th.Percentile < 2e-16 \*\*\*  
## airqualitysubset.data$Arithmetic.Mean < 2e-16 \*\*\*  
## airqualitysubset.data$State.Code 3.22e-05 \*\*\*  
## airqualitysubset.data$Observation.Count 4.27e-16 \*\*\*  
## airqualitysubset.data$Parameter.Code 3.37e-10 \*\*\*  
## airqualitysubset.data$Valid.Day.Count 2.24e-08 \*\*\*  
## airqualitysubset.data$Longitude 0.0943 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 585.9 on 66352 degrees of freedom  
## Multiple R-squared: 0.9459, Adjusted R-squared: 0.9459   
## F-statistic: 1.656e+05 on 7 and 66352 DF, p-value: < 2.2e-16

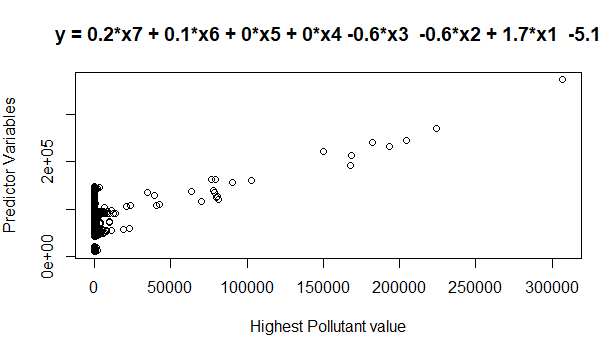
distPred\_lmodel2 <- predict(lmodel2,airqualitysubset.data)

#Coefficients of the model  
lmodel2coef=coef(lmodel2)  
print(lmodel2coef)

## (Intercept)   
## -2.122397e+01   
## airqualitysubset.data$X99th.Percentile   
## 1.731945e+00   
## airqualitysubset.data$Arithmetic.Mean   
## -5.932299e-01   
## airqualitysubset.data$State.Code   
## -5.939120e-01   
## airqualitysubset.data$Observation.Count   
## 3.231363e-03   
## airqualitysubset.data$Parameter.Code   
## 6.682045e-04   
## airqualitysubset.data$Valid.Day.Count   
## 1.122995e-01   
## airqualitysubset.data$Longitude   
## 2.253777e-01

#Equation of the linear regression  
eq = paste0("y = ", round(lmodel2coef[8],1), "\*x7 + ",round(lmodel2coef[7],1), "\*x6 + ",round(lmodel2coef[6],2), "\*x5 + ",round(lmodel2coef[5],2), "\*x4 " ,round(lmodel2coef[4],1), "\*x3 " ,round(lmodel2coef[3],1), "\*x2 + " ,round(lmodel2coef[2],1), "\*x1 ",round(lmodel1coef[1],1))

#Plotting the graph  
plot(airqualitysubset.data$X1st.Max.Value,airqualitysubset.data$X99th.Percentile+airqualitysubset.data$Arithmetic.Mean+airqualitysubset.data$State.Code+airqualitysubset.data$Observation.Count+airqualitysubset.data$Parameter.Code+airqualitysubset.data$Valid.Day.Count+airqualitysubset.data$Longitude,main=eq,xlab="Highest Pollutant Value",ylab="Predictor Variables")



#Confidence Intervals of the Model  
confint(lmodel2)

## 2.5 % 97.5 %  
## (Intercept) -5.227211e+01 9.8241661881  
## airqualitysubset.data$X99th.Percentile 1.724063e+00 1.7398259883  
## airqualitysubset.data$Arithmetic.Mean -6.204372e-01 -0.5660225990  
## airqualitysubset.data$State.Code -8.738749e-01 -0.3139491082  
## airqualitysubset.data$Observation.Count 2.452590e-03 0.0040101363  
## airqualitysubset.data$Parameter.Code 4.597152e-04 0.0008766937  
## airqualitysubset.data$Valid.Day.Count 7.294850e-02 0.1516504819  
## airqualitysubset.data$Longitude -3.866127e-02 0.4894166156

#Analysis of Variance  
anova(lmodel2)

## Analysis of Variance Table  
##   
## Response: airqualitysubset.data$X1st.Max.Value  
## Df Sum Sq Mean Sq  
## airqualitysubset.data$X99th.Percentile 1 3.9724e+11 3.9724e+11  
## airqualitysubset.data$Arithmetic.Mean 1 6.2451e+08 6.2451e+08  
## airqualitysubset.data$State.Code 1 5.2884e+06 5.2884e+06  
## airqualitysubset.data$Observation.Count 1 4.2951e+07 4.2951e+07  
## airqualitysubset.data$Parameter.Code 1 1.4322e+07 1.4322e+07  
## airqualitysubset.data$Valid.Day.Count 1 1.0132e+07 1.0132e+07  
## airqualitysubset.data$Longitude 1 9.6076e+05 9.6076e+05  
## Residuals 66352 2.2776e+10 3.4326e+05  
## F value Pr(>F)   
## airqualitysubset.data$X99th.Percentile 1.1573e+06 < 2.2e-16 \*\*\*  
## airqualitysubset.data$Arithmetic.Mean 1.8194e+03 < 2.2e-16 \*\*\*  
## airqualitysubset.data$State.Code 1.5406e+01 8.678e-05 \*\*\*  
## airqualitysubset.data$Observation.Count 1.2513e+02 < 2.2e-16 \*\*\*  
## airqualitysubset.data$Parameter.Code 4.1724e+01 1.058e-10 \*\*\*  
## airqualitysubset.data$Valid.Day.Count 2.9518e+01 5.559e-08 \*\*\*  
## airqualitysubset.data$Longitude 2.7990e+00 0.09433 .   
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Regression Evaluation  
actuals\_pred\_lmodel2 <- data.frame(cbind(lmodel2\_actuals=airqualitysubset.data$X1st.Max.Value, lmodel2\_predicteds=distPred\_lmodel2))

regr.eval(actuals\_pred\_lmodel2$lmodel2\_actuals, actuals\_pred\_lmodel2$lmodel2\_predicteds)

## mae mse rmse mape   
## 52.22632 343214.91890 585.84547 Inf

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Hypothesis Testing\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

#Hypothesis Test 1  
# H0(Null Hypothesis): Mean of highest pollution when no events occured=Mean of highest pollution when events occured  
# H1(Alternative Hypothesis: True difference in means is not equal to 0)

#checking variances of the two groups

var(airqualitysubset.data$X1st.Max.Value[airqualitysubset.data$Events.Occurred=="No"])

## [1] 6655878

var(airqualitysubset.data$X1st.Max.Value[airqualitysubset.data$Events.Occurred=="Yes"])

## [1] 552654.8

#The variances are not equal. Hence considering non equal variances  
#Hypothesis Testing using Two Sample t-test  
t.test(airqualitysubset.data$X1st.Max.Value~airqualitysubset.data$Events.Occurred,mu=0,alt="two.sided",conf=0.95,var.eq=F,paired=F)

##   
## Welch Two Sample t-test  
##   
## data: airqualitysubset.data$X1st.Max.Value by airqualitysubset.data$Events.Occurred  
## t = -9.0268, df = 9273, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -179.2342 -115.2788  
## sample estimates:  
## mean in group No mean in group Yes   
## 104.3652 251.6217

#Hypothesis Testing using Wilcox Test  
wilcox.test(airqualitysubset.data$X1st.Max.Value~airqualitysubset.data$Events.Occurred,data=airqualitysubset.data)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: airqualitysubset.data$X1st.Max.Value by airqualitysubset.data$Events.Occurred  
## W = 44146423, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0

#Hypothesis Test 2  
# H0(Null Hypothesis): Mean of certified pollution data=Mean of uncertified pollution data  
# H1(Alternative Hypothesis: True difference in means is not equal to 0)  
#checking variances of the two groups  
var(airqualitysubset.data$Arithmetic.Mean[airqualitysubset.data$Certified=="No"])

## [1] 36.2195

var(airqualitysubset.data$Arithmetic.Mean[airqualitysubset.data$Certified=="Yes"])

## [1] 225575.7

#The variances are not equal. Hence considering non equal variances  
#Hypothesis Testing using Two Sample t-test  
t.test(airqualitysubset.data$Arithmetic.Mean~airqualitysubset.data$Certified,mu=0,alt="two.sided",conf=0.95,var.eq=F,paired=F)

##   
## Welch Two Sample t-test  
##   
## data: airqualitysubset.data$Arithmetic.Mean by airqualitysubset.data$Certified  
## t = -21.886, df = 59919, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -46.36718 -38.74488  
## sample estimates:  
## mean in group No mean in group Yes   
## 5.36853 47.92456

#Hypothesis Testing using Wilcox Test  
wilcox.test(airqualitysubset.data$Arithmetic.Mean~airqualitysubset.data$Certified,data=airqualitysubset.data)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: airqualitysubset.data$Arithmetic.Mean by airqualitysubset.data$Certified  
## W = 238672542, p-value < 2.2e-16  
## alternative hypothesis: true location shift is not equal to 0