Lead Poisoning – Income, County   
and Community

Sai Sri Sindhuja Pragadaraju

Bhanu Tejaswi Kidambi

Johnson and Wales University

Providence, RI.

Table of Contents

[Abstract 3](#_Toc38054284)

[Introduction 4](#_Toc38054285)

[Data Collection 5](#_Toc38054286)

[Data cleaning and wrangling 6](#_Toc38054287)

[Data Exploration 7](#_Toc38054288)

[Exploratory Analysis and Pattern finding 7](#_Toc38054289)

[Model Development 9](#_Toc38054290)

[Feature selection 9](#_Toc38054291)

[Decision Tree Model 10](#_Toc38054292)

[Results Interpretation 12](#_Toc38054293)

[Validating results 12](#_Toc38054294)

[Conclusion & Future scope 13](#_Toc38054295)

[References 14](#_Toc38054296)

[Tables 15](#_Toc38054297)

[Appendix 16](#_Toc38054298)

Abstract

Lead poisoning has been an area of study for researchers for long but saw new rise since the Flint water crisis that happened between April 25, 2014 to October 15, 2015, when approximately 99,000 residents got exposed to toxic amount of lead through drinking water when the source was changed to Flint Water System from the already existing Detroit Water Authority for the city of Flint, MI (Flint Lead Exposure Registry, 2019). Lead poisoning and its effects on children has been much talked about but some studies suggest that only half of the children effected by lead poisoning are actually tested and treated. Many researchers have stated the causes for lead poisoning as paint, pipes, exported items, soil, etc. but factors such as race, income and demographics may need deeper study. Massachusetts state data has been collected to understand the factors that could cause lead poisoning in children. For this regression algorithms are used to find the importance of each variable in predicting. MS Excel, R Studio and Tableau tools were used in this study.

# Introduction

More than 1.2 million American children are said to have lead poisoning but only half of them are tested. According to CDC the threshold for lead poisoning is 5µg/dL. In early 1980s, 17% of preschoolers had lead content in blood more than 15µg/dL which has since seen a decline of 0.53% of testes children with more than 10µg/dL. Studies state that more than 2µg/dL of lead in children causes decrease in their IQ levels. 5µg/dL and above result in severe neurological damages. Since testing of lead poison is not required for most students and schools in USA, many cases go untested. This is more prominent in Southern and Western USA (Frostenson, 2017). An applied study is done where the data has been collected from various sources mainly from Bureau of Environmental Health and Census. The Bureau of environmental Health works on studies and enforcing laws and regulation in food pollution, childhood lead poisoning, standards of housing etc.

Lead poisoning is a serious concern in adults and children, with infants and children being more susceptible to it and leading to developmental issues, neurological damages, premature births, seizures etc. Pregnant women can pass these issues to unborn children if exposed to lead poisoning during pregnancies. Causes for lead poisoning are listed below for better understanding. (Lead Poisoning, 2019)

Paint: Although houses in USA have stopped using lead based paints, toys and furniture, already existing houses and furniture are a cause.

Pipes: Pipes of lead or copper with lead soldering can release lead into the water which can be a source of drinking water like taps.

Imported goods: Mexican candies, herbal remedies and cosmetics and other imported good from India, Mexico, and Nigeria are considered to have lead in them, including the traditional Chinese medicines.

Occupation: This can also be a source for lead poisoning when working in auto repairs, mining, battery and paint manufacturing companies and others.

# Data Collection

Massachusetts State, county and community level data has been collected and joined to create one aggregated dataset. A 6-year data that is 2012-2017 for children between 9 to 47 months of age with blood lead level count greater than 5µg/dL was selected along with children screened data. Not all children between the ages 9 months to 47 months were tested, the percentage of confirmed cases can only be taken based on the total number of children tested. For this purpose the population data from Children screened has been taken for all three levels .i.e., State, county and community. Household Income data was taken from List of List of Massachusetts locations by per capita income, which was collected from Wikipedia which had its sources from census (List of Massachusetts locations by per capita income, 2020). The aggregated dataset had 2206 observations and 9 variables for community data. This was mainly used for the regression algorithms and predictions.

The State and county data was mainly used for understanding the overall location verses confirmed lead poisoning cases. The below charts shows the drop in overall confirmed lead poisoning cases in children from 4508 in 2012 to 2709 in 2017.

County data has also show similar results with 2017 year having lowest confirmed case counts for all the counties. Nantucket County confirmed cases for the years 2013-2017 were missing which have been taken as 0.

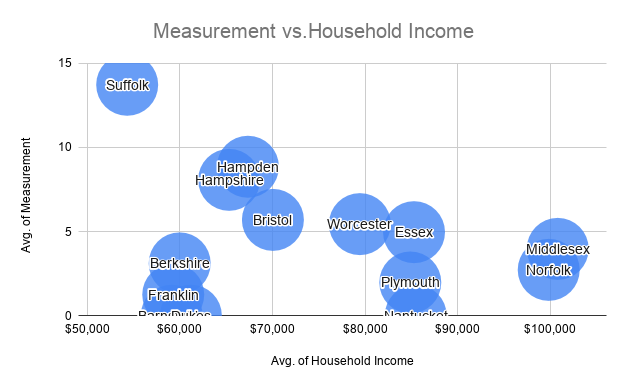
## Data cleaning and wrangling

To understand the actual trend of confirmed cases, a variable called Measurement is present in the dataset which is the ratio of Confirmed cases to Tested. The denominator is ‘Tested’ here because not all the children in the population were screened. The result was then multiplied by 1000 for scaling purposes. The dataset had few missing data where the number of confirmed were not provided, hence the measurement was NA. These values were replaced with 0 in both confirmed cases and measurement variable data. Community data was grouped into types such as Inner Core (IC), Regional Urban Centers (RUC), Maturing Suburbs (MSU), Developing Suburbs (DSU) and Rural Towns (RT). (Council, 2008)

# Data Exploration

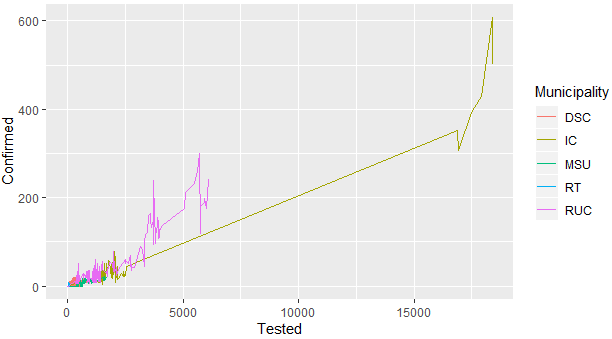
## Exploratory Analysis and Pattern finding

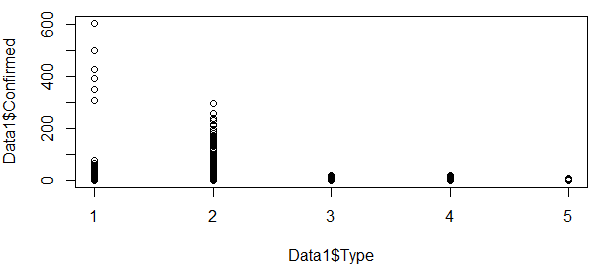
To further understand the influence of various factors on lead poisoning, the income data was combined with Community data with Geography, County and Measurement.



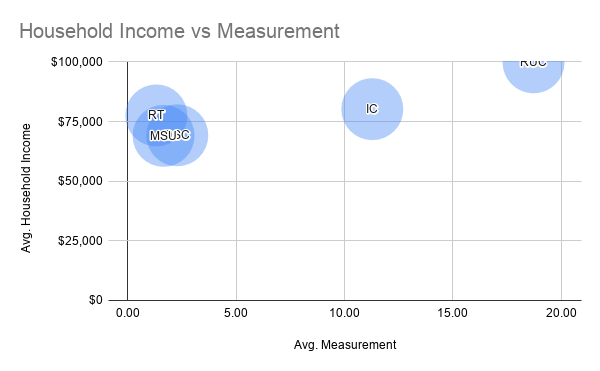
The above graph gives away an evident picture of increase in Household income means decrease in measurement, i.e., Counties which have an average household income of $80,000 or more have less than 5 confirmed for 1000 tested. Suffolk County has an average household income of $54,326 and it also has the highest Measurement. Below chart shows that Cities tend to have higher values of Measurement compared to towns irrespective of the household income. With this it can be said that the main focus needs to be more on developed or developing cities for better results.

Now, let us consider the community wise data which we have divided into community types such as IC (1), RUC (2), MSU (3), DSU (4) and RT (5) as Municipality. The regions under IC, RUC and MSU have comparatively highest confirmed cases.





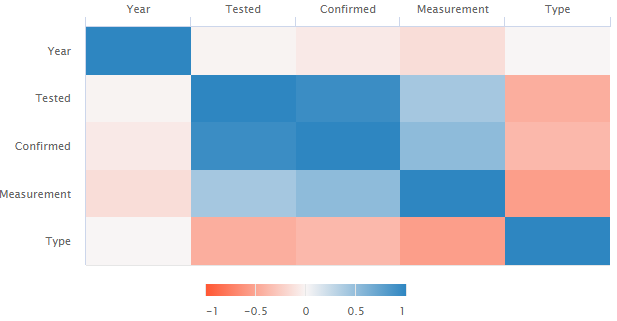
Further, the Household income to Measurement based on Municipality is showing the IC and RUC having high measurement despite of having an average Household income greater than $75,000. Hence, a safe conclusion can be made with the data at hand that household income may not be as big of a factor in lead poisoning case counts in children.



# Model Development

## Feature selection

A correlation heatmap is created to understand the variables which are important for the regression analysis. Here type represents Municipality in numeric form for creating correlation calculations. The variables Tested and Measurement have a positive effect on Confirmed cases where as year and type have a negative effect. This means, as years increase the confirmed cases count goes down and as tested cases increase the confirmed cases may increase resulting in upward trend of measurement. Also as the value of types of community (IC (1), RUC (2), MSU (3), DSU (4) and RT (5)) increase, the confirmed cases decrease. Which is true as RT has lowest confirmed cases compared to IC and RUC. Since measurement is dependent on confirmed and tested it cannot be considered an independent variable.



A general formula for this regression model can be written as below:

**Y(confirmed) = ßo + ß1\*Tested - ß2\*Year – ß3\*Type + Error**

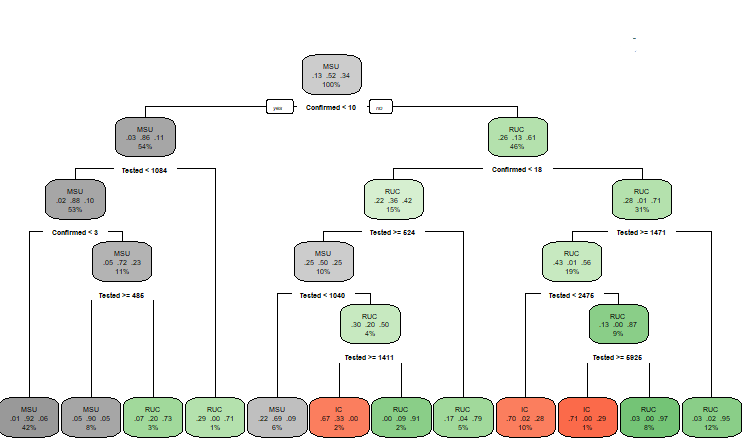
## Decision Tree Model

Since, we have concluded the communities under Inner core, Regional Urban Centers and Maturing Suburbs have more confirmed cases compared to other types. A decision tree classifier is used to build our model.

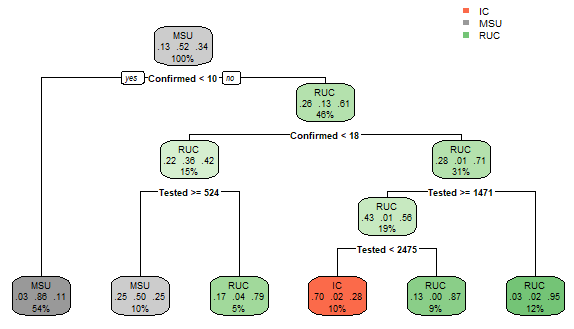
The Municipality variable is factorized to 3 levels for this model. The generated decision tree can be seen below. The prior probability of the types of municipality are

IC: 0.1344, MSU: 0.5237 and RUC: 0.3419.

The tree generated shows the confirmed cases and tested cases have high variable importance compared to Year.



Pruning has been performed to reduce the complexity and remove branches than provide less power to classify the instances.



## Results Interpretation

Step1. If confirmed cases are less than 10, then there is 54% chances of the region falling under MSU which is 86% of that data.

Step2. If confirmed cases are greater than 10 there is 46% chances of the region being RUC which is 61% of that data.

Step3. If confirmed cases are below 18 then 42% of RUC has 15% probability of having confirmed cased lesser than 18.

Step4. It is further checked if the tested cases are greater than or equal to 524. This also can be understood as higher percentage of IC and RUC communities have confirmed cases greater than 18.

Step5. 70% of IC data has 10% probability of having confirmed cases above 18 if the tested cases are between 1471 and 2475.

As the tested cases increase the chances of falling under the regions RUC and IC also increases.

### Validating results

The accuracy for the prunes tree was 78.5%. Variable importance for Confirmed is 133.4, Tested is 119 and for Year is 1.4. Further pruning did not help improve the accuracy other than reducing the tree size to 1 node and 2 leaves. A Naïve Bayes classification was performed on the same data which provided similar results with accuracy 78.9%. For pruning, the complexity parameter (cp) is set to 0.03 which resulted in a simpler tree and good accuracy.

Decision Tree Confusion Matrix Naïve Bayes Confusion Matrix

# Conclusion & Future scope

The study can be concluded stating the communities under Inner Core and Regional Urban Centers have greater lead poisoning cases in children between the ages of 9 to 47 months compared to Rural Towns and Developing and Maturing Suburbs. Increase in household incomes may result in low lead poisoning cases but this could be due to people working in the cities but living in suburban places. This could also be one of the explanations for the cities in counties having higher measurement value compared to the towns of the same counties irrespective of the household income. Children in cities have a greater threat of falling prey to lead poisoning. The household income data taken is a median household income which is a limitation for this study.

For future studies, we may consider to extensively study areas that fall under Inner Core and Regional Urban Centers including offices and schools. Having an estimate for the people working/studying and living in these areas compared to people working/studying in urban cities but living in suburban locations can help narrow down the factors for lead poisoning in children.

Similarly, race, ethnicity and income distribution can be included in the study to understand what cultural, ethnical or household habits could be the deciding factors.

# References

Council, M. A. (2008). *Massachusetts Community Types.* Boston: Metropolitan Area Planning Council.

*Flint Lead Exposure Registry*. (2019, July 30). Retrieved from cdc.gov: https://www.cdc.gov/nceh/lead/programs/flint-registry.htm

Frostenson, S. (2017, April 24). *1.2 million children in the US have lead poisoning. We’re only treating half of them.* Retrieved from vox.com: https://www.vox.com/science-and-health/2017/4/27/15424050/us-underreports-lead-poisoning-cases-map-community

*List of Massachusetts locations by per capita income*. (2020, March 20). Retrieved from wikipedia: https://en.wikipedia.org/wiki/List\_of\_Massachusetts\_locations\_by\_per\_capita\_income

Natasha Grzincic, A. R. (2019, July 23). *These Maps Reveal the Worst Neighborhoods for Lead Poisoning in NYC, Boston, and Chicago*. Retrieved from Vice: https://www.vice.com/en\_us/article/kzmvd9/these-maps-reveal-the-worst-neighborhoods-for-lead-poisoning-in-nyc-boston-and-chicago

Staff, M. C. (2019, December 20). *Lead Poisoning*. Retrieved from mayoclinic: https://www.mayoclinic.org/diseases-conditions/lead-poisoning/symptoms-causes/syc-20354717

*Decision Trees in R using rpart*. (2014, August 24). Retrieved from gormanalysis.com: https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/

Khan, R. (2017, December 20). *Naive Bayes Classifier: theory and R example*. Retrieved from RPubs-by RStudio: https://rpubs.com/riazakhan94/naive\_bayes\_classifier\_e1071

Tables

|  |  |  |
| --- | --- | --- |
| County | Avg. of Measurement | Avg. of Household Income |
| Middlesex | 3.97 | $100,805 |
| Norfolk | 2.73 | $99,836 |
| Nantucket | 0.00 | $85,478 |
| Essex | 4.96 | $85,294 |
| Plymouth | 1.97 | $84,891 |
| Worcester | 5.44 | $79,448 |
| Bristol | 5.70 | $70,058 |
| Hampden | 8.85 | $67,360 |
| Hampshire | 8.08 | $65,330 |
| Dukes | 0.00 | $61,201 |
| Berkshire | 3.11 | $59,992 |
| Franklin | 1.25 | $59,299 |
| Barnstable | 0.00 | $59,133 |
| Suffolk | 13.73 | $54,326 |

|  |  |  |
| --- | --- | --- |
| Municipality | Avg. of Measurement | Avg. of Household Income |
| RUC | 18.72 | $99,845 |
| IC | 11.28 | $80,221 |
| RT | 1.31 | $77,592 |
| DSC | 2.28 | $69,267 |
| MSU | 1.65 | $69,043 |

# Appendix

#Libraries needed

ipak <- function(pkg){

new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]

if (length(new.pkg))

install.packages(new.pkg, dependencies = TRUE)

sapply(pkg, require, character.only = TRUE)

}

# usage

packages <- c("ggplot2", "dplyr", "reshape2","readr","funModeling","data.table","Hmisc","pastecs","psych","foreign",

"factoMineR","factoextra","randomForest","caret","quantmod","MASS","kernlab","mlogit","caTools","e1071","leaps","missForest",

"RColorBrewer", "scales", "grid","readxl","Rfast","lattice","VGAM","ggplot2","GGally","lme4","compiler","parallel","boot","Information","survival","survminer","lubridate","lintr","zoo","odbc","DBI","RODBC","xlsx","skimr","fastDummies","tidyverse")

ipak(packages)

#Loading data into R

Data1 <- read\_excel("C:/Users/Sindhu/Desktop/Final\_Term/capestoneproject/checkpoint1/Dataanalysis.xlsx")

View(Data1)

str(Data1) #structure of the dataframe

dim(Data1) #dimensions of the dataframe

summary(Data1)

write.csv(Data1,file = "C:/Users/Sindhu/Desktop/Dataanalysis.csv",row.names = FALSE)

Data1 %>% ggplot(aes(x = `Confirmed Cases`)) +

geom\_histogram(color = "black", alpha = 0.5,bins=10) +

facet\_wrap(~Municipality, scales = "free")

boxplot(x = `Confirmed Cases`, y= Year, data=Data1)

##Check for null

sapply(Data1, function(x) sum(is.na(x)))

# Exploring the data

basic\_eda <- function(Data1)

{

glimpse(Data1)

df\_status(Data1)

freq(Data1)

profiling\_num(Data1)

plot\_num(Data1)

describe(Data1)

}

basic\_eda(Data1)

#Changing column names

colnames(Data1)[colnames(Data1) == "Denominator"] <- "Population"

colnames(Data1)[colnames(Data1) == "Cases Tested"] <- "Tested"

colnames(Data1)[colnames(Data1) == "Confirmed Cases"] <- "Confirmed"

names(Data1)

#Plots for patterns

plot(Data1$Confirmed ~ Data1$Type, data = Data1)

ggplot(data = Data1, mapping = aes(x = Year, y = Confirmed)) +

geom\_line() + facet\_grid(. ~ Municipality)

ggplot(data = Data1, mapping = aes(x = Tested, y = Confirmed, color=Municipality)) +

geom\_line() + facet\_wrap(facets =vars(Year))

#Data wrangling

IC\_data <- Data1 %>%

filter(Municipality == "IC")

IC\_data\_plot <- IC\_data %>%

ggplot(., aes(x = Tested, y = Confirmed)) + geom\_line() +

facet\_wrap(facets =vars(Year))

IC\_data\_plot

RUC\_data <- Data1 %>%

filter(Municipality == "RUC")

RUC\_data\_plot <- RUC\_data %>%

ggplot(., aes(x = Tested, y = Confirmed, fill =

Year)) + geom\_line() + facet\_wrap(facets =vars(Geography))

RUC\_data\_plot

list1<-c("RUC","IC","MSU")

my\_data <- Data1 %>%

filter(Municipality %in% list1)

my\_data\_plot <- my\_data %>%

ggplot(., aes(x = Measurement, y = Year)) + geom\_line() + facet\_wrap(facets =vars(Municipality))

my\_data\_plot

names(my\_data)

#Top 10 places for confirmed and measurement

topcases <- my\_data %>%

select(Geography, Municipality, Year, Measurement,Confirmed,Type) %>%

group\_by(Geography) %>%

summarise(total = sum(Confirmed)) %>%

top\_n(10, total)

topcases

topmeasure <- my\_data %>%

select(Geography, Municipality, Year, Measurement,Confirmed,Type) %>%

group\_by(Geography) %>%

summarise(total = sum(Measurement)) %>%

top\_n(10, total)

topmeasure

#you then need to filter your original data to be only the data for the top cities

my\_data\_plot2 <- filter(my\_data, Geography %in% topcases$Geography) %in% group\_by(Year)

#correlation heatmap

library(highcharter)

str(my\_data)

hchart(cor(my\_data[c(2,4,5,6,8,9)]))

mydata11<-my\_data[,c(2,8,9)]

mydata12<-my\_data[,c(2,3:6)]

mydata3<-my\_data[c(2,4,5,6,8,9)]

ggplot(mydata11,aes(x=Measurement,y= Type)) +

geom\_line() + facet\_wrap(facets=vars(Year))

hchart(cor(mydata11))

#####Decision Tree###############

library(rpart)

library(rpart.plot)

mydata12$Municipality<-as.factor(mydata12$Municipality)

str(mydata12)

set.seed(12)

sample <- createDataPartition(mydata12$Municipality, p = 0.7, list = F) #split by 70%

Train\_dt <- mydata12[sample,] #70%

Test\_dt <- mydata12[-sample,] #30%

dt <- rpart(Municipality ~.-Population, data = Train\_dt, method = 'class')

summary(dt)

rpart.plot (dt, extra = 104)

p\_dt <- predict(dt, Test\_dt, type = 'class') # prediction

table\_mat <- table(Test\_dt$Municipality, p\_dt) # confusion matrix

accuracy\_Test <- sum(diag(table\_mat)) / sum(table\_mat) # accuracy test

printcp(dt)

dt$variable.importance # variable importance

#prune the tree

dt1 <- prune(dt, cp = .03)

rpart.plot(dt1,extra=100)

p\_dt1 <- predict(dt1, Test\_dt, type = 'class')

table\_mat1 <- table(Test\_dt$Municipality, p\_dt1)

table\_mat1

accuracy\_Test1 <- sum(diag(table\_mat1)) / sum(table\_mat1)

printcp(dt1)

dt1$variable.importance

dt2 <- prune(dt, cp = 0.04)

rpart.plot(dt2,extra=100)

p\_dt2 <- predict(dt2, Test\_dt, type = 'class')

table\_mat2 <- table(Test\_dt$Municipality, p\_dt1)

table\_mat2

accuracy\_Test2 <- sum(diag(table\_mat2)) / sum(table\_mat2)

printcp(dt2)

dt2$variable.importance

#######Naive Bayes###################

set.seed(123)

library(naivebayes)

NB\_model<- naive\_bayes(as.factor(Municipality) ~. -Population,

usekernel = T, data=Train\_dt)

printALL=function(NB\_model){

trainPred=predict(NB\_model, newdata = Train\_dt, type = "class")

trainTable=table(Train\_dt$Municipality, trainPred)

testPred=predict(NB\_model, newdata=Test\_dt, type="class")

testTable=table(Test\_dt$Municipality, testPred)

trainAcc=(trainTable[1,1]+trainTable[2,2]+trainTable[3,3])/sum(trainTable)

testAcc=(testTable[1,1]+testTable[2,2]+testTable[3,3])/sum(testTable)

message("Contingency Table for Training Data")

print(trainTable)

message("Contingency Table for Test Data")

print(testTable)

message("Accuracy")

print(round(cbind(trainAccuracy=trainAcc, testAccuracy=testAcc),3))}

printALL(NB\_model) # check training and testing data accuracies

summary(NB\_model)

plot(NB\_model)

#make predictions

p\_nb <- predict(NB\_model, Test\_dt) #predictions

p\_nb <- table(Test\_dt$Municipality, p\_nb) # confusion matrix

p\_nb

accuracy\_Test\_nb <- sum(diag(p\_nb)) / sum(p\_nb) # accuracy test

#########Linear regression#############

set.seed(1)

# load libraries

library(caret)

library(mlbench)

# prepare sampling method

control <- trainControl(method="cv", number=5)

set.seed(7)

fit <- train(Confirmed~.-Population, data=mydata12, method="glm", trControl=control)

# display results

print(fit)

summary(fit)

#method 2

sample <- createDataPartition(mydata11$Measurement, p = 0.7, list = F) #split by 70%

Train <- mydata11[sample,] #70%

Test <- mydata11[-sample,] #30%

lead\_lm <- lm(Measurement~., data =Train)

summary(lead\_lm)

sample3 <- createDataPartition(mydata3$Confirmed, p = 0.7, list = F) #split by 70%

Train3 <- mydata3[sample,] #70%

Test3 <- mydata3[-sample,] #30%

lead\_lm3<-lm(Confirmed ~.-Municipality-Measurement,data=Train3)

summary(lead\_lm3)

predicted <- predict(lead\_lm3, Test3,type="response") # predicted scores

#Predict the ratings

predictModel<-predict(lead\_lm3, data = Test3)

predictModel

#comparing the predicted values with actual values for rating

plot(Test3$Confirmed, type="l", lty=1.8, col="red")

lines(predictModel,type="l", lty=1.8, col="blue")

#finding accuracy

max(mydata3$Confirmed) #range of the DV

min(mydata3$Confirmed) #range of the DV

RMSE\_C<-sqrt(mean((predictModel-Train3$Confirmed)^2))

RMSE\_C

ggplot(data=Train, aes(lead\_lm[["residuals"]])) +

geom\_histogram(binwidth = 1, color = "black", fill = "purple4") +

theme(panel.background = element\_rect(fill = "white"),axis.line.x=element\_line(),axis.line.y=element\_line())+

ggtitle("Histogram for Model Residuals")

###############END############