

**REPORT**

**Boston: Is it safe?**

Course ALY6015 : **Intermediate Analytics**

**CRN:** 80797

Term: Winter 2019 Quarter

**Team Members:** Pranay Reddy, Shraddha Thakkar, Manpreet Kaur

**Submitted to:** Valeriy Shevchenko

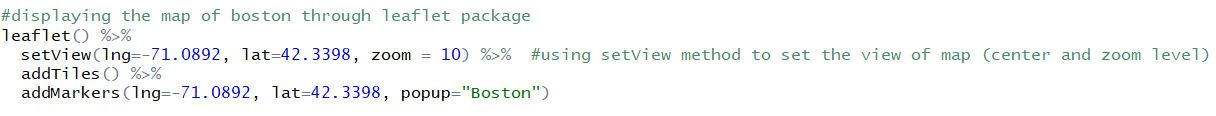
**Dated:** May 14, 2019

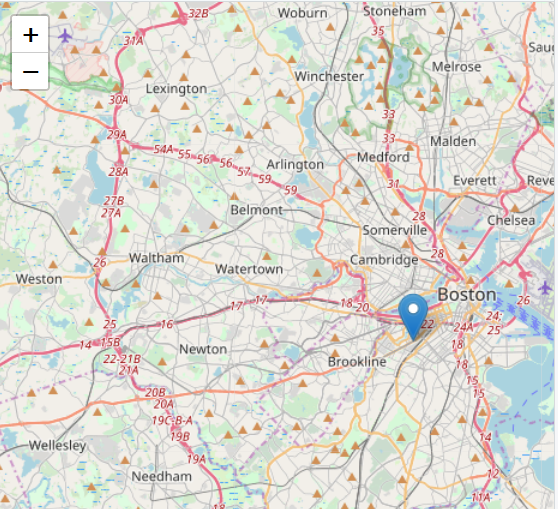
**Abstract**

We started the project wanting to know that this Boston city, although we know is greater than most cities in the US, how safe it was.Is it true to say that Boston was more violent than New York and Seattle, but less violent than Chicago and Las Vegas, according to numbers from the FBI, based on crimes committed back in 2015.As of 12/21/18 Nationally, Boston ranked 14 out of 50 according to Us News. Our goal was to dig into the dataset of Boston Crimes, collected from Kaggle and analyze the data to throw light about the crimes expected in the year 2019. Also, we did our analyses regarding the crime rates during weekends and weekdays. A lot of other interesting analysis will be presented through this report.

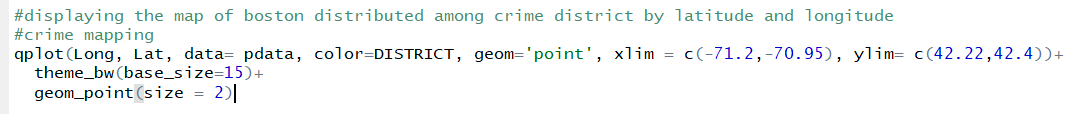
**Exploratory Data Analysis**

Displaying the code of location Boston through leaflet package by using the following code:

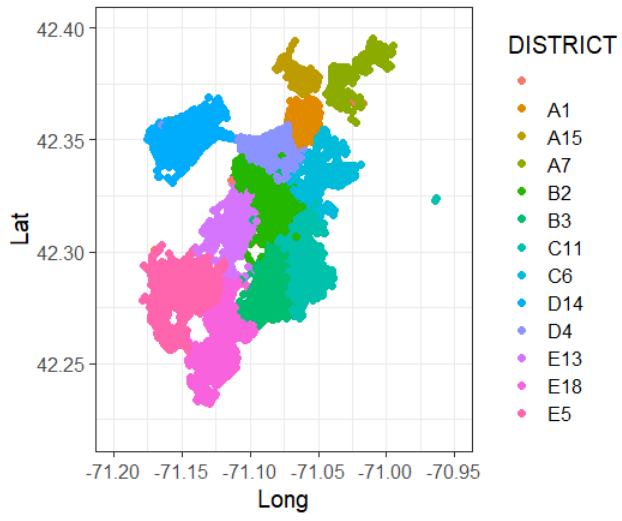




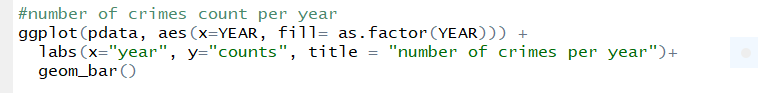
Following is the map of Boston distributed among crime districts. We used the following code snippet for the same.

****

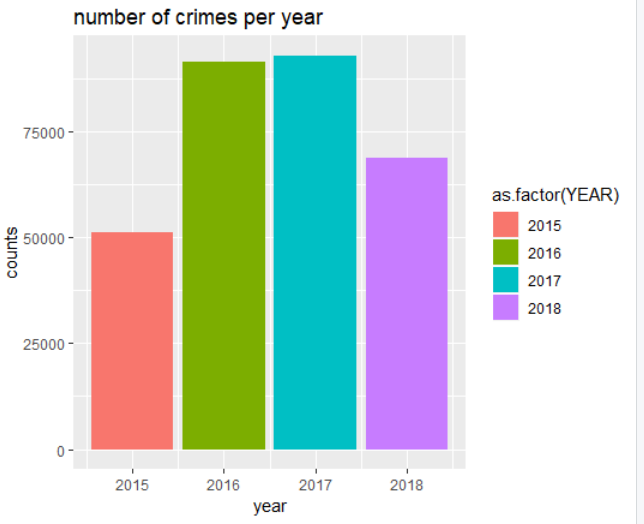
We get the following output:

****

Now, we are displaying the crimes count per year by using the following code:

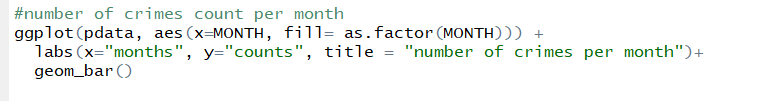


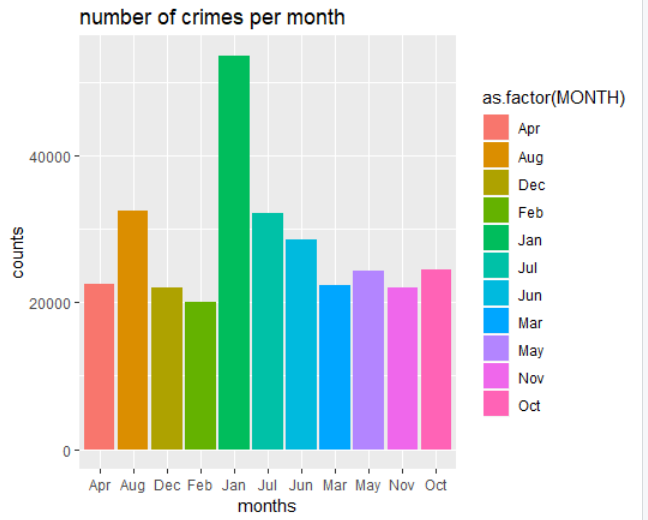
We get the following output:



As we can see the crime rate has increased from 2015 and is highest in 2017. Fortunately, 2018 has experienced less crime rate compared to the previous 2 years.

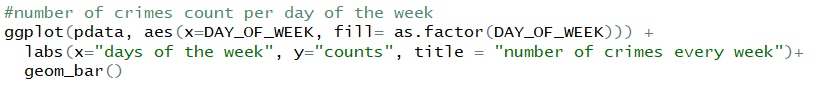
Further, we are displaying the crimes count per month by using the following code:



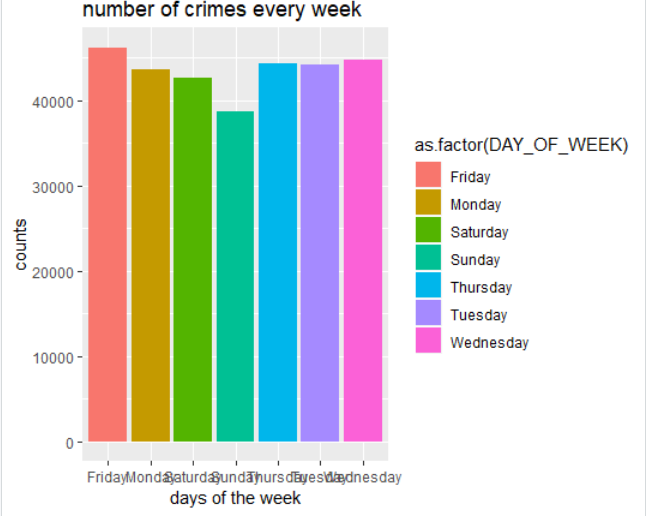


From the output, we can say that the maximum crimes occurred was in the month of January. May be it's because of the new year. And maximum public or tourist come around.

Further, we are displaying the crimes count per week by using the following code:



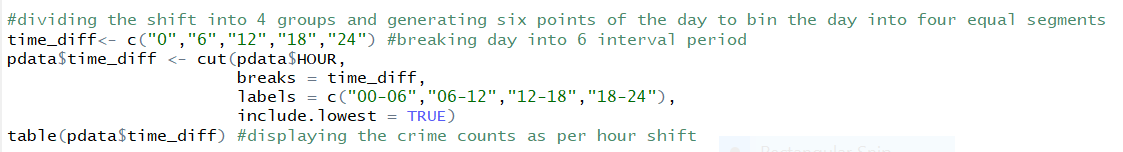
We get the following output:



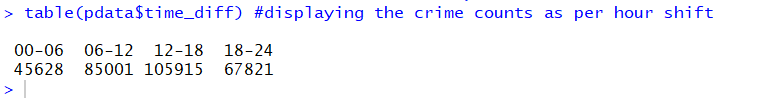
From the graph, it’s obvious to note that on Friday maximum crime occurs. The reason is inevitable that it's weekend and maximum public roam around outside.

Further, displaying the crimes as per hours shift, on which time the maximum crimes occurred.

For that, we divided the 24 hours time slot into 4 parts each of 6 hours by using the following code:

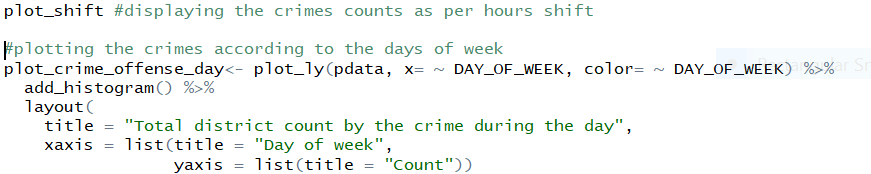


we have got the counts of crimes as per hours shift:

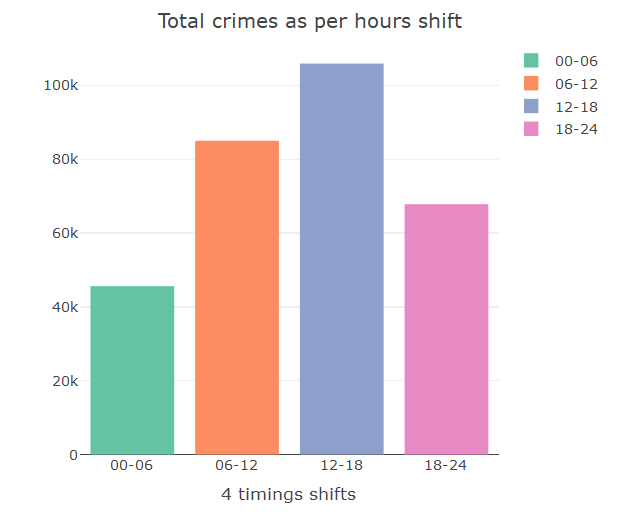


From the above output we can see that ,from 12:00am to 6:00am 45,628 crimes occured. From 6:00am to 12:00pm 85,001 crimes occured. From 12:00pm to 6:00pm the highest number of crimes occured of 105,915. Lastly in evening from 6:00pm to 12:00 am 67,821 crimes occured.

Next, we are plotting the crimes as per hours shift.

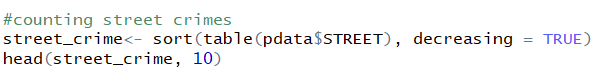


We get the following output:

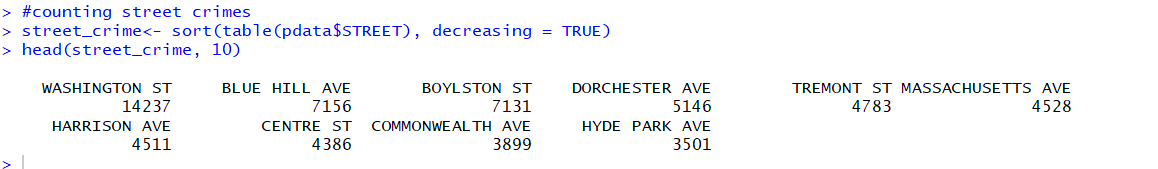


The highest number of crimes has occured in 12:00pm-6:00pm

Next, we are counting the street crimes and displaying top 10 crimes

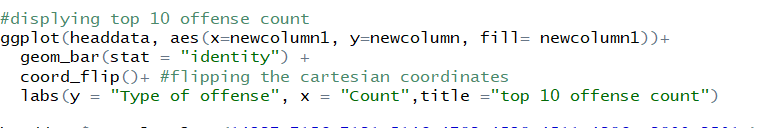


We get the following output:

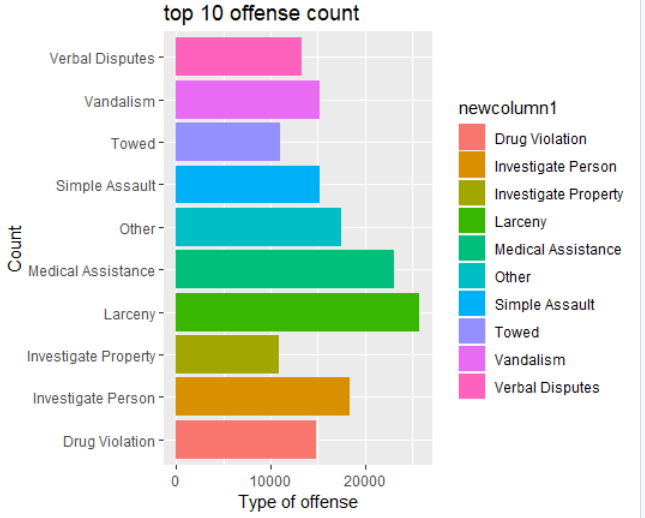


The highest crime has occured on washington crime of 14,237. The least number of crimes has occured on hyde park ave of 3501.

Further, displaying top 10 offense count by using the following code:

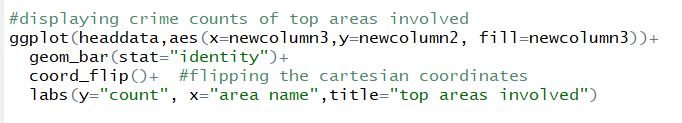


We get the following output:

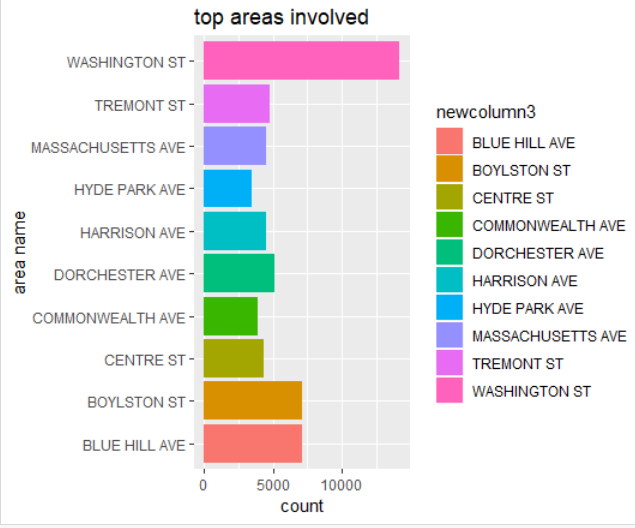


The highest crime that has occured is Larceny of about 21,000. Second highest is Medical Assistance. The least is Vehicle of towing about 11,000.

Further displaying the crime counts of top areas involved by using the following code:



We get the following output:



As we can see, the washington street tops the lists of more than 15,000 crimes.

The Boylston street and Blue Hill avenue are same with rate of more than 5000 crimes.

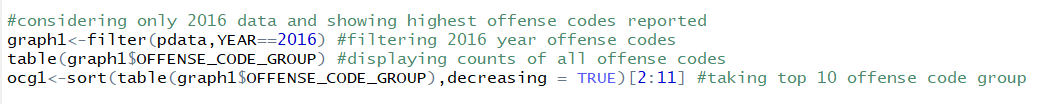
The least is Hyde park ave with count of less than 5000 crimes.

Generating word cloud in order to understand and visualize crimes as per streets.

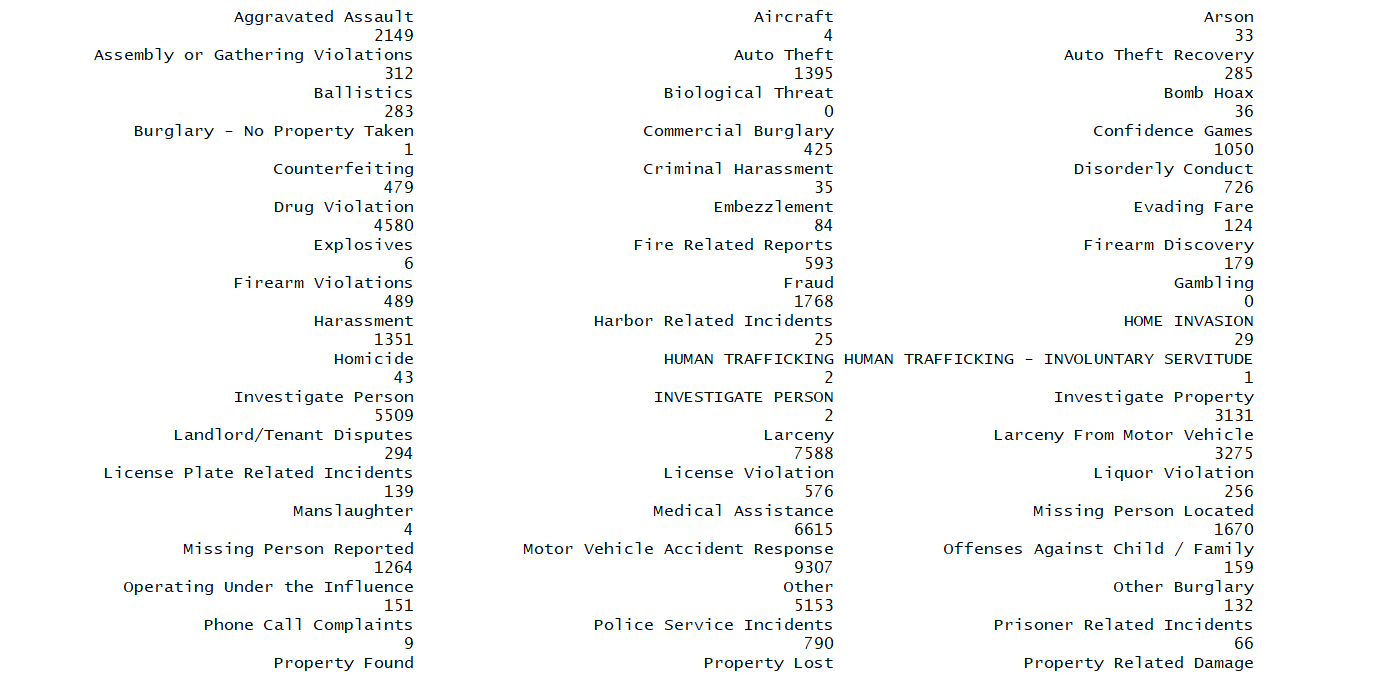


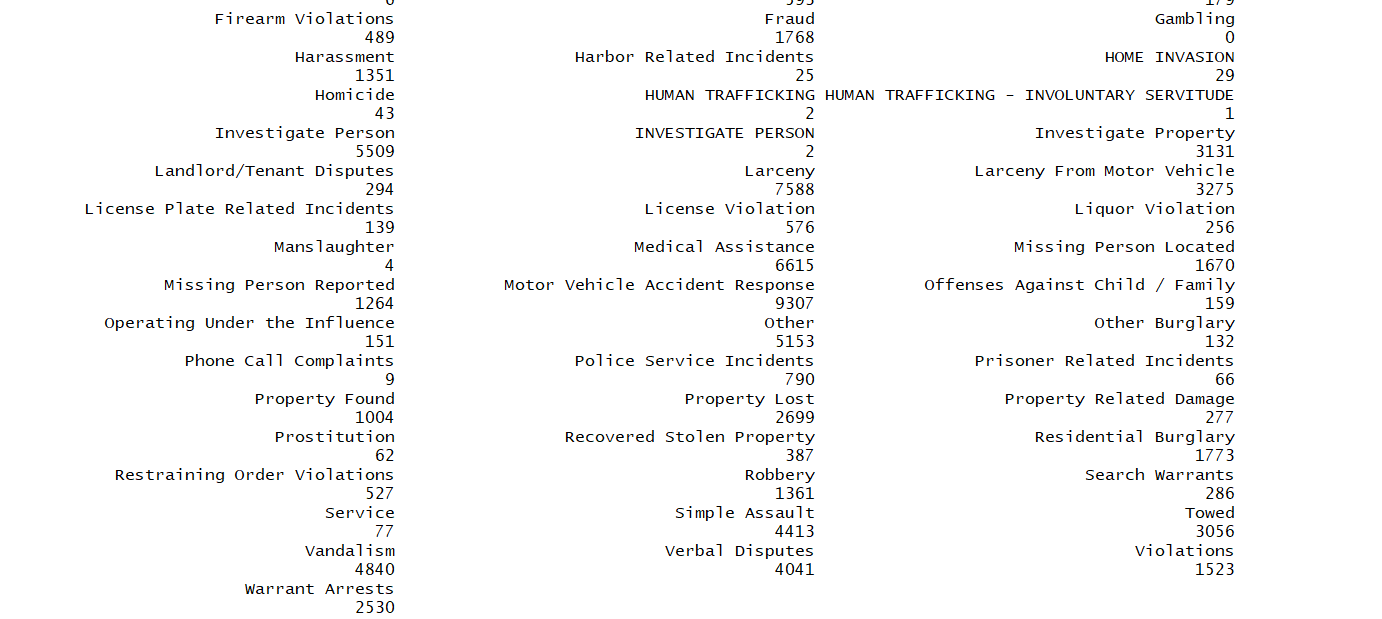
Our EDA supports the fact that the most dangerous street is washington street. Higher the font, more that city is likely to be  in danger of crime.

Now, considering 2016 data in order to analyse crime rates:

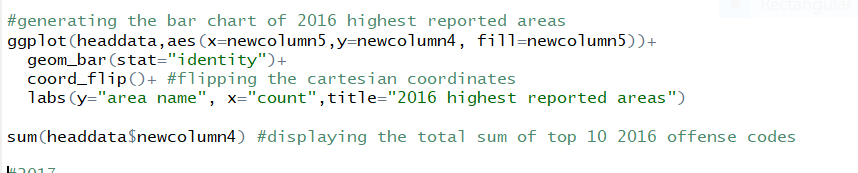


We get the counts of types of crimes occured.

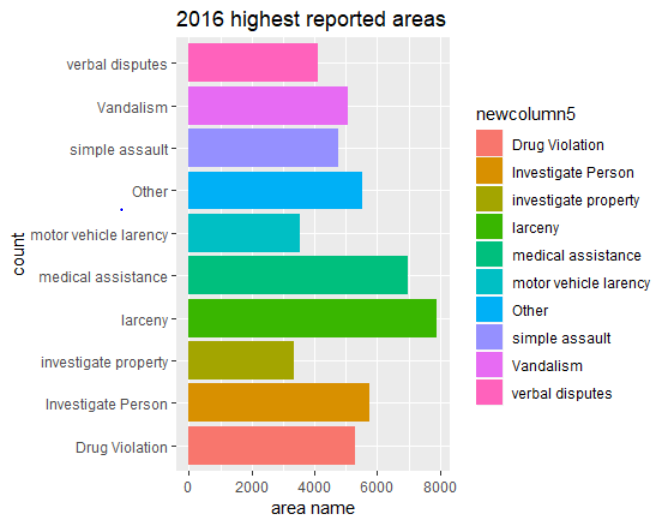


****

Displaying the 2016 crimes of highest reported areas:



We get the following output:

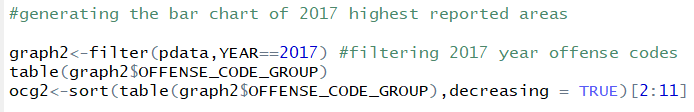


As we can see that larceny tops the list amongst highest crimes in 2016 with count of nearly 8000. The second highest is medical assistance. The least is investigation of property with count of more than 3000.

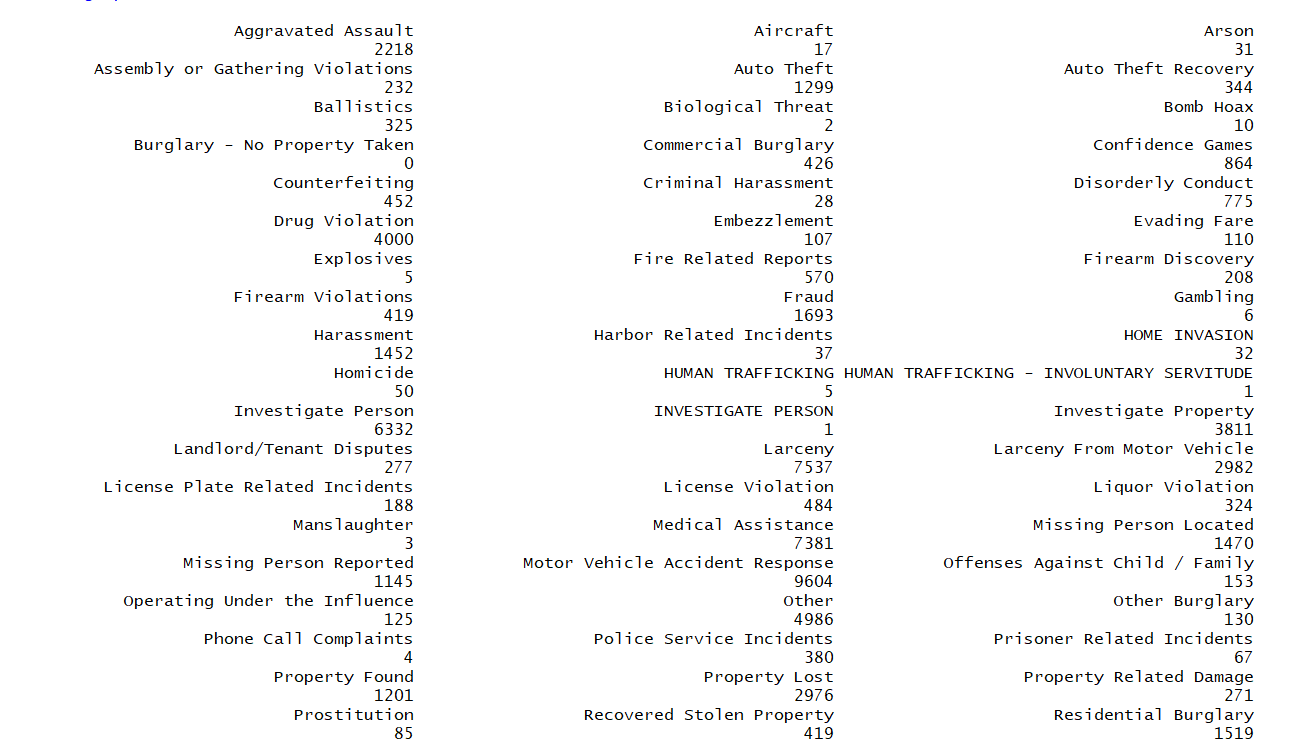
Total 52,256 crimes occured in 2016.

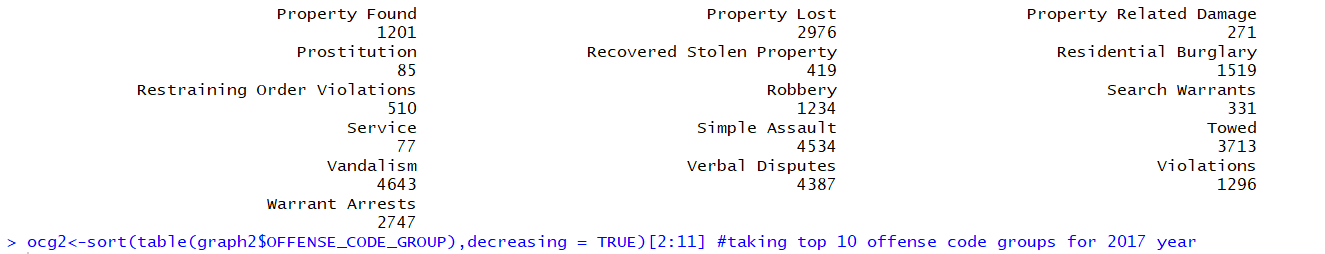


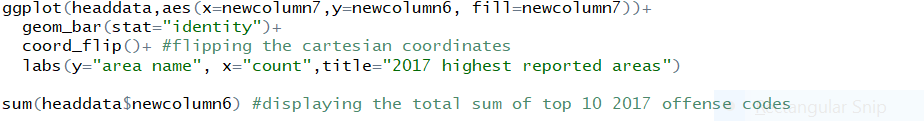
Similarly, amalyzing crime counts and data for 2017 and 2018

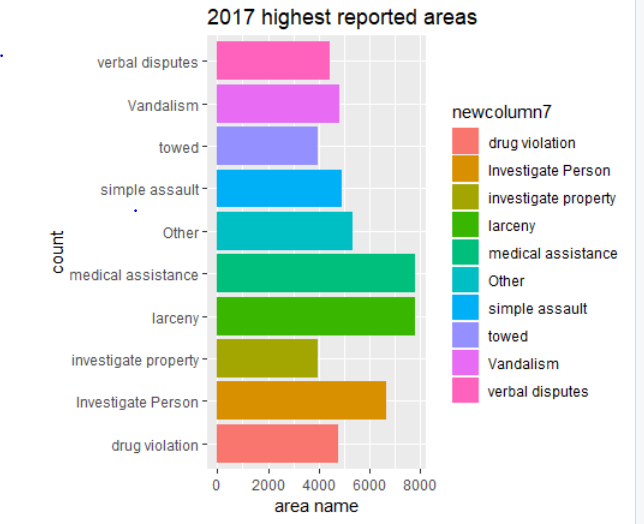


Following are the crime counts of the crimes that occured in 2017.

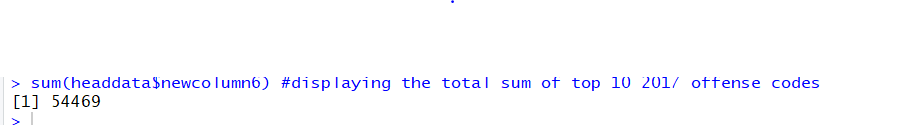
****

****



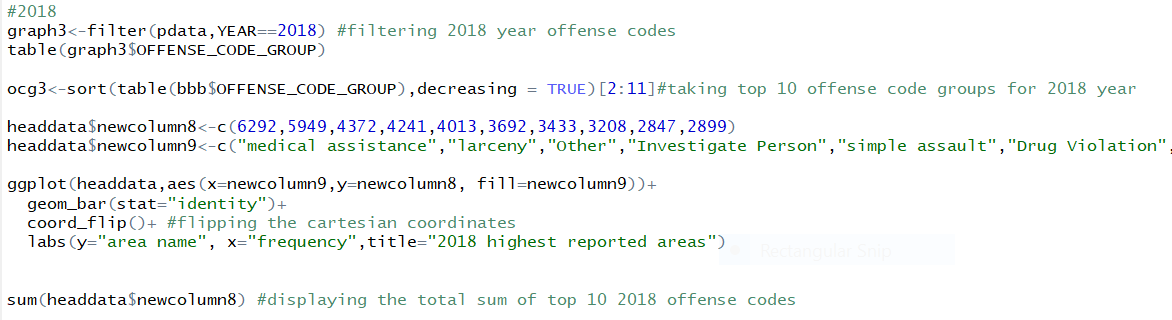


From the above graph we can say that larceny and medical assistance has the same crime count of nearly 8000. The least was towing and the second largest was investigation of the property with count of approx. 6600.

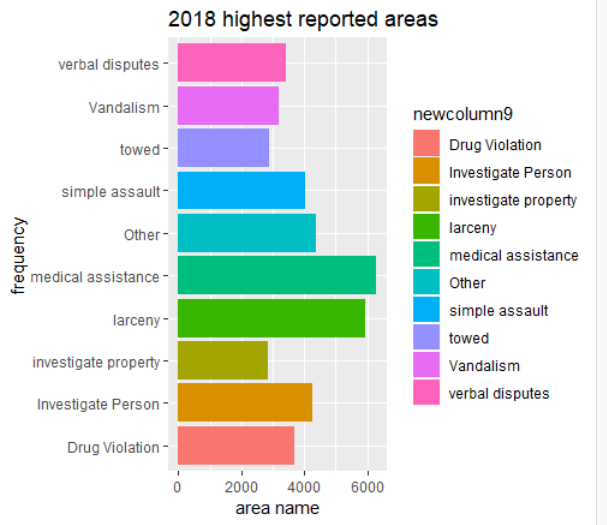
****

Total number of crimes that occured in 2017 were 54,469.

Following is the code snippet for 2018 crime analysis:

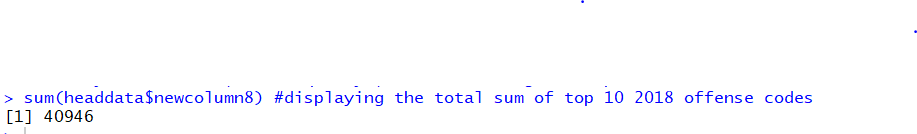


We get the following output:

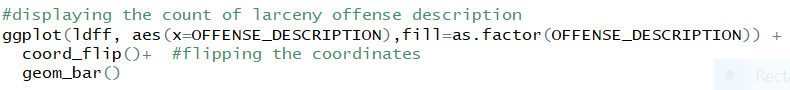
****

We can see that in 2018, the highest crime occured was medical assistance with count of more than 6000 followed by larceny with count of almost 6000. The least was towing with count of nearly 3000.

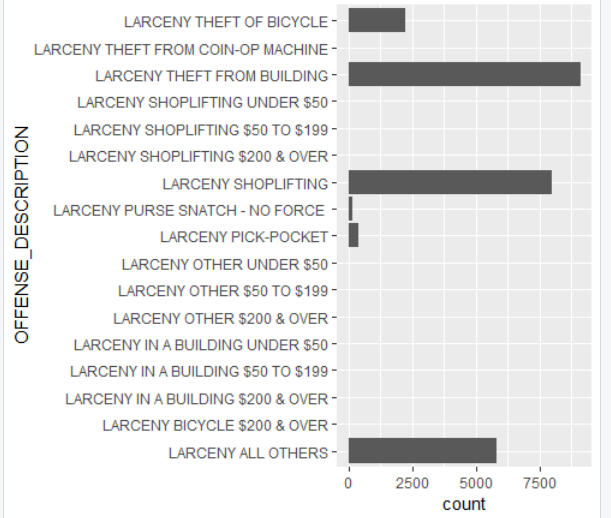
In 2018 there were 40,946 crimes that occured.



From the above analysis of 3 years graph we can conclude that larceny was highest crime that occured. So we digged further into it what type of larceny occurs.

****

We get the following output:



On analyzing we come to know that larceny theft from the building tops the list with count of almost more than 8000. The second crime occured is Larceny shoplifting with count of almost more than 7500.

The least larceny occured is pick pocketing.

**Hypothesis Testing**

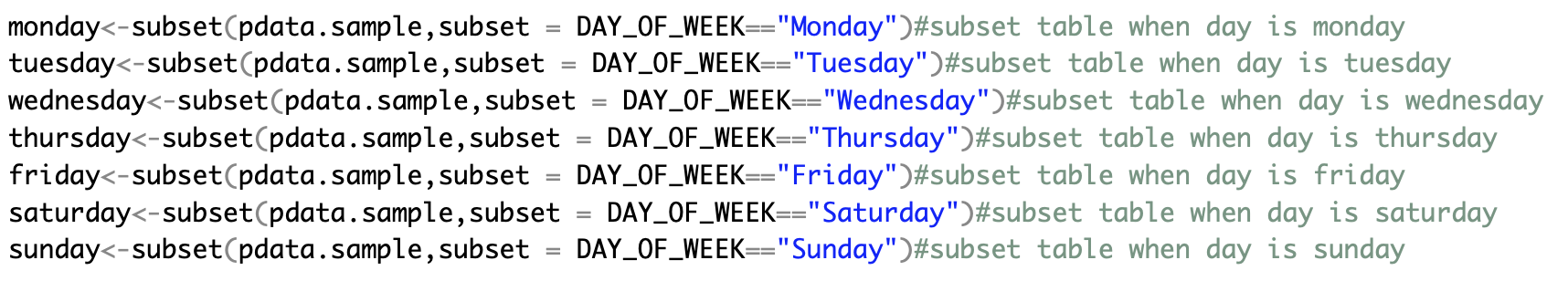
**To find out if the mean of the crimes during weekends and weekdays are similar or not, we set the null and alternate hypothesis as below**

****

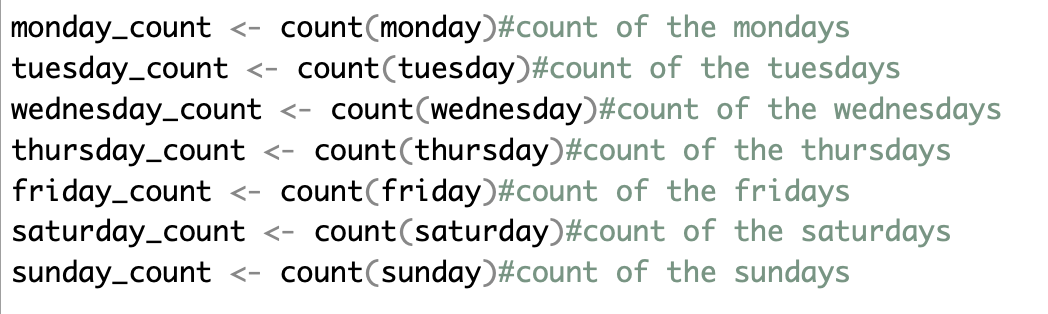
**We set a constant sample selection and selected a sample of size 30**

****

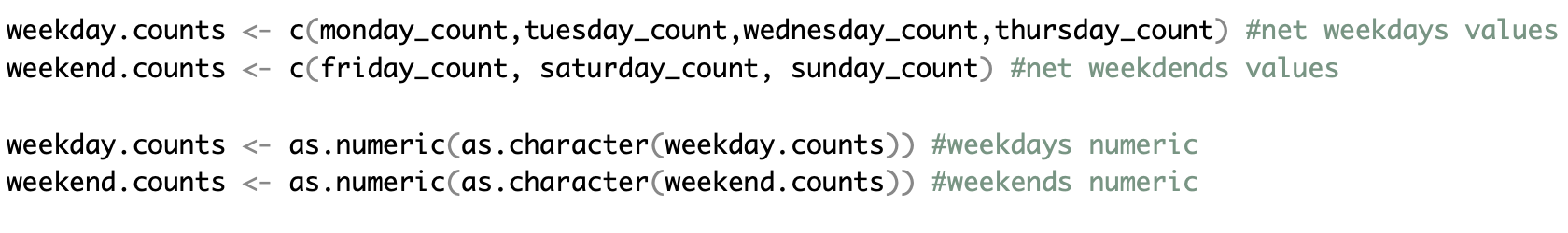
**subset the table according to days**

****

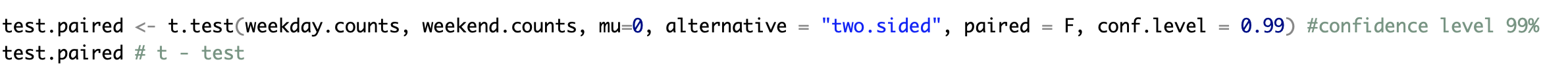
**Count of the number of weekdays**

****

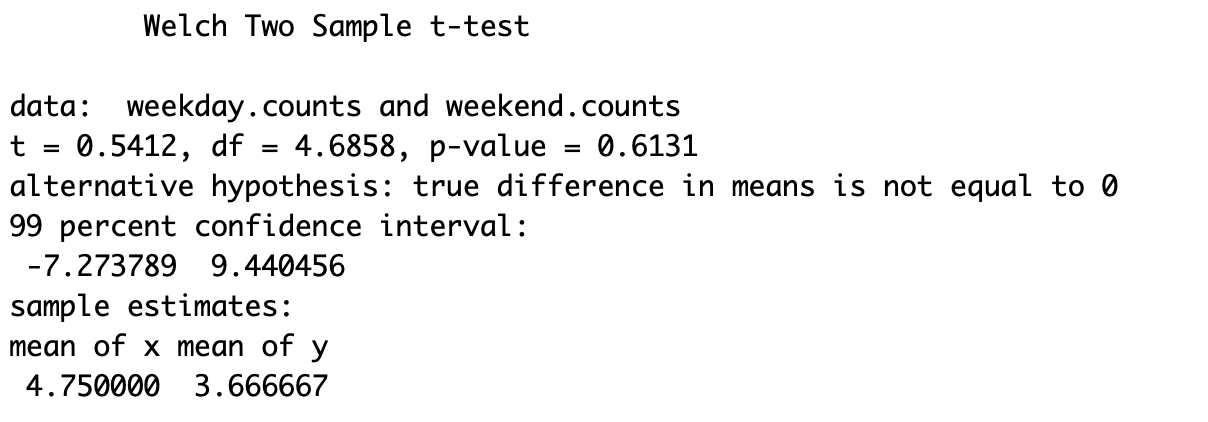
**Separately count the weekdays and weekends and convert it to numeric type**

****

**Carry out the hypothesis 2 tailed t-test**

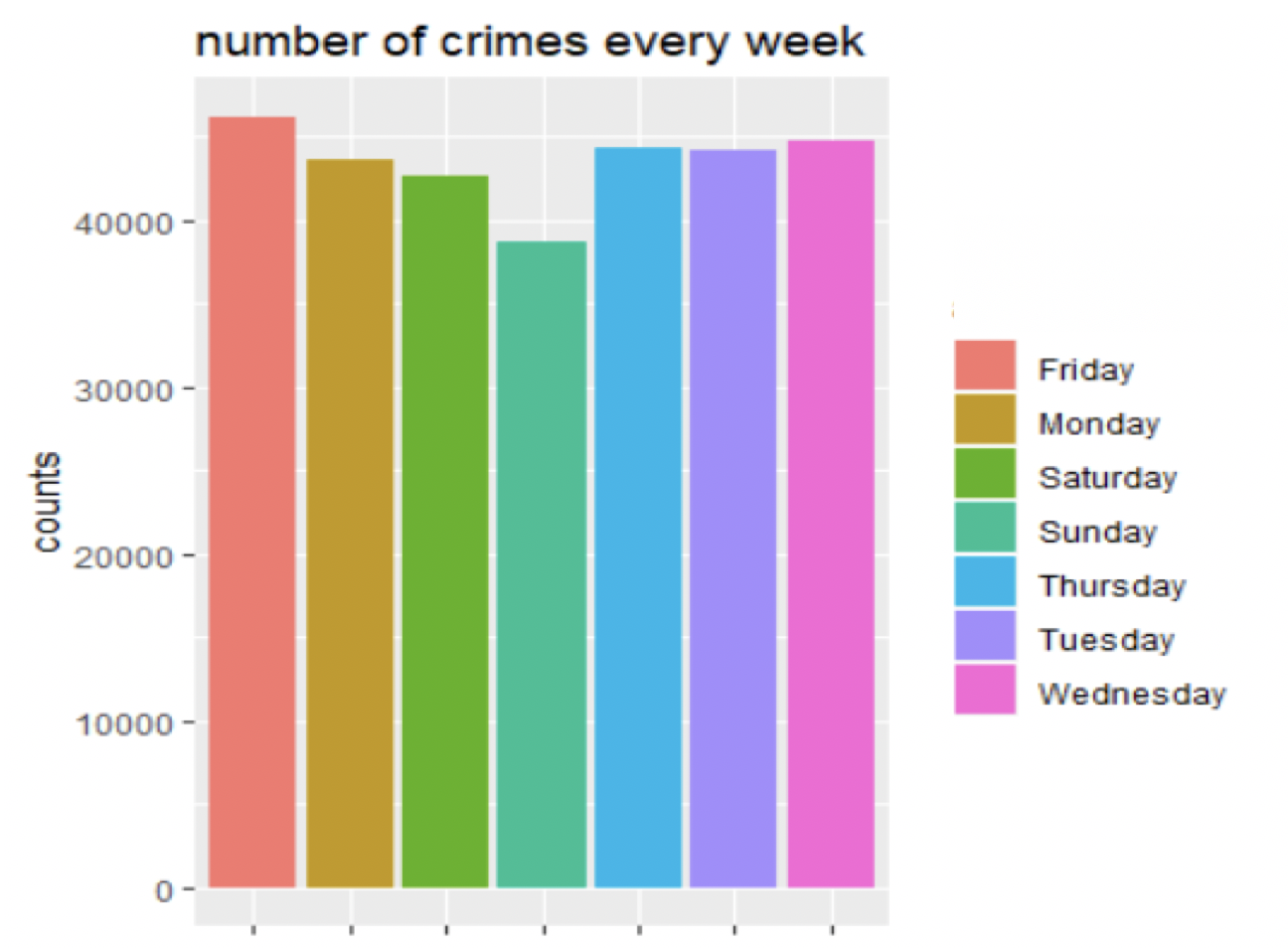
****

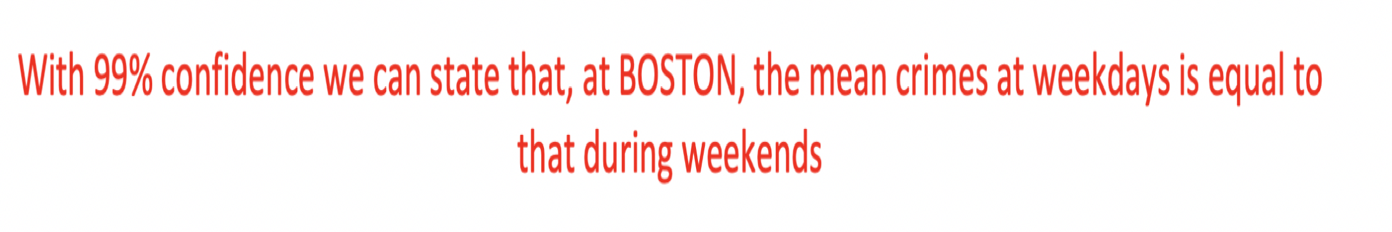
**The result of the t-test is as below**

****

**The p-value is 0.6131 which much higher 0.05. This shows that we refuse to reject the null hypothesis.**

**This is consistent with the following plot where the count of crimes in all the weekdays are almost similar.**

****

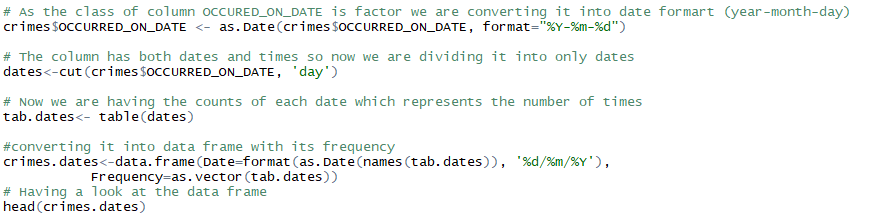
****

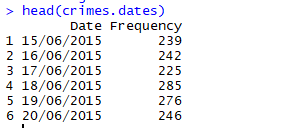
**Time Series Analysis**

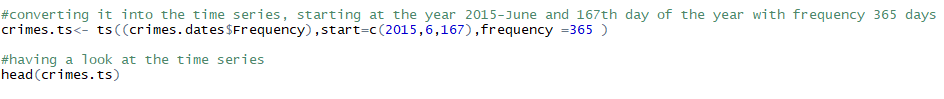
**Converting it to Time series**

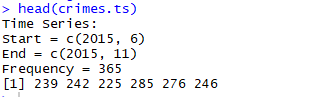






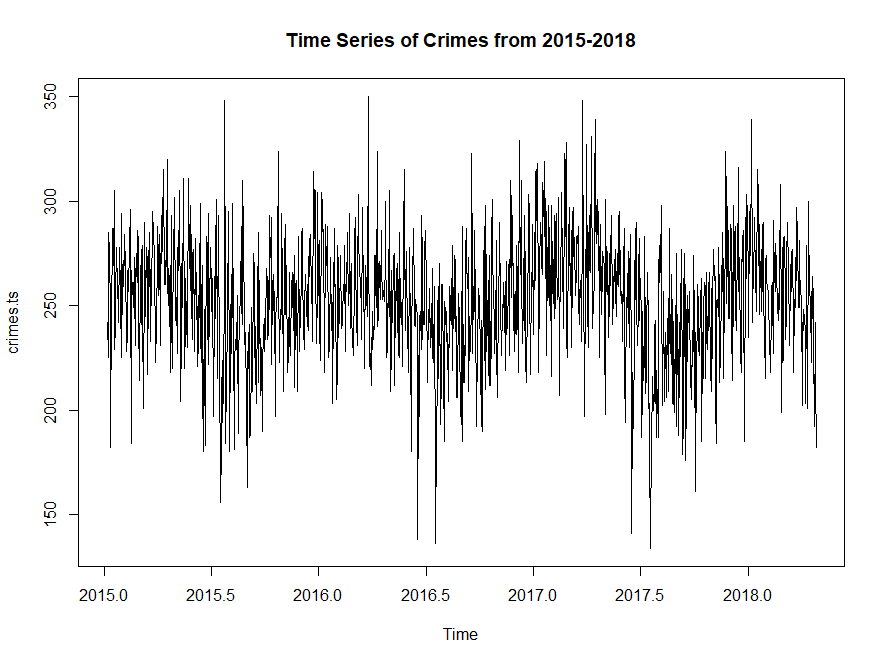






In this time series, The starting date is 15th June 2015 and the frequency is 365.

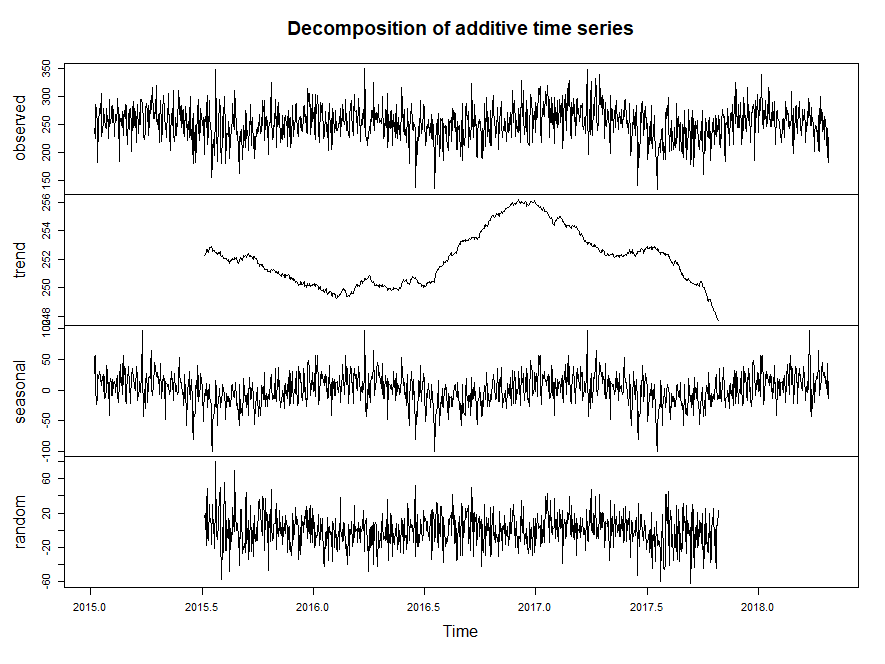




We can see that the above times series fluctuates consistently and the mean and variance do not change over time so it's an additive time series. Now we can decompose this additive time series to find the trend.

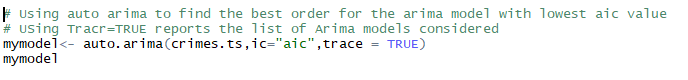
Decomposing of additive time series

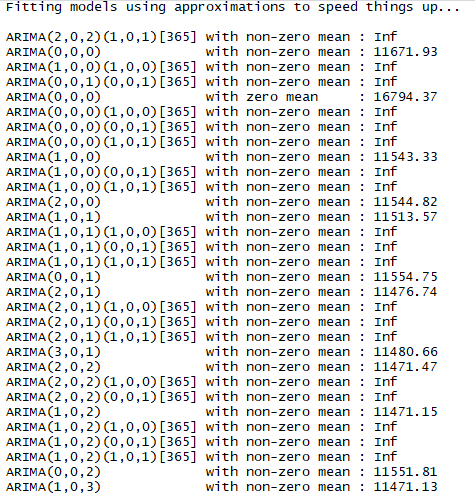


****

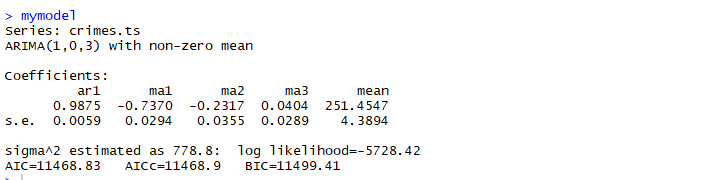
From the trend graph, we can see that we can observe that the Number of crimes per day in Boston has increased significantly from June 2017 and then gradually decreased from the year 2017.

**Finding the best fit model for ARIMA**

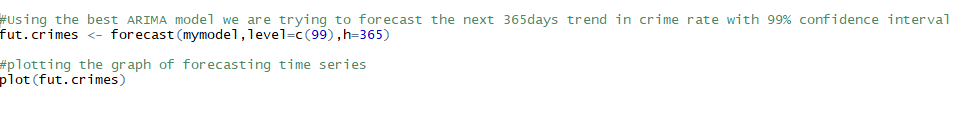
****

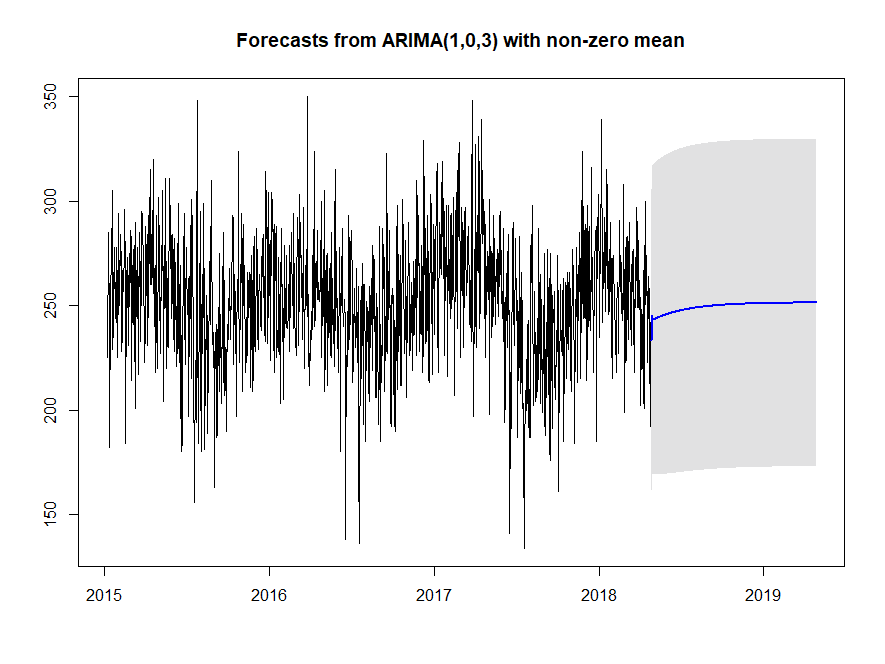
****

Through various combinations, we are trying to find the best fit model which has the lowest AIC value.



From the results, we found that the best-fitted model for ARIMA is of order ARIMA(1,0,3) with non-zero mean which has AIC=11468.83. So using this order we will forecast the crime rate for next 1 year.

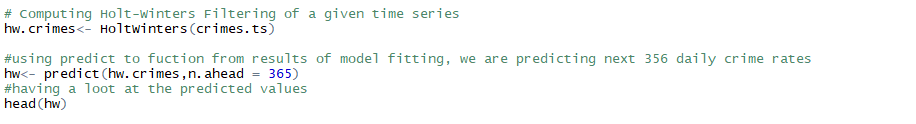
Forecasting 



From the above forecasting graph, we could see that the trend in the number of crimes per day has slightly increased for the next 1 year compared to past year.

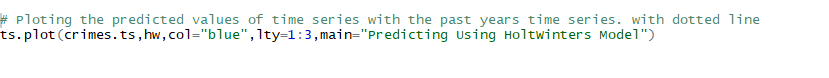
Holt-Winters Model

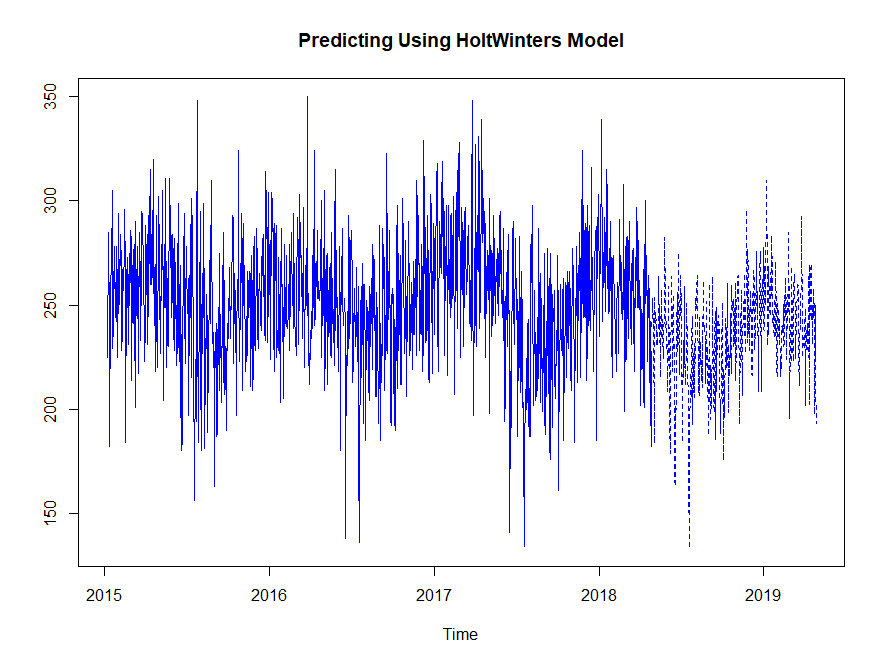
Now we are using Holt-winters model to predict the number of crimes for on a daily basis i n Boston





Plotting the graph of Holt-Winters predicted values





From this plot we can have a predicted value of crimes on each day for the next 365days.This Analysis helps Boston police to act accordingly when the crime rate is high and try to reduce them.

**Conclusion**

**After performing an in-depth analysis of crimes in Boston, we have analyzed the trends and patterns of the different locations at different times of the year, month and day. Some important points are mentioned below**

1. **Washington street had the most crimes in the past 3 years.**
2. **January is the month when most crimes occur in Boston**
3. **Larceny is the most reported crime in Boston, and the building is the most popular among it.**
4. **Weekdays and weekends makes no difference in the crime rates**
5. **There is a chance of a slight increase in crimes in the year 2019, based on the previous year's data.**

Appendix

#package installations and library read

install.packages("ggplot2")

install.packages("dplyr")

install.packages("RColorBrewer")

install.packages("forecast")

install.packages("wordcloud")

install.packages("tidyverse")

install.packages("VIM")

install.packages("leaflet")

install.packages("plotly")

install.packages("lubridate")

install.packages("magrittr")

install.packages("wordcloud")

library(ggplot2)

library(dplyr)

library(RColorBrewer)

library(forecast)

library(wordcloud)

library(tidyverse)

library(VIM)

library(leaflet)

library(plotly)

library(lubridate)

library(magrittr)

library(wordcloud)

#initial steps- data reading and viewing

pdata<-read.csv(file.choose(), header = T)

head(pdata)

pdata[][is.na(pdata[])] <- 0

pal = brewer.pal(9,"Blues")

crimes <- pdata#creating another variable for the crime data

names(crimes)

#Creating nice looking color palettes especially for thematic maps

pal = brewer.pal(9,"Blues")

#building a contingency table of the counts at each combination of factor levels

street\_name <- as.tibble(table(Street\_Name= pdata$STREET))

#naming the columns

colnames(street\_name) <- c("Street\_Name", "Count")

#Integer vector, containing the random number generator

set.seed(21)

#Plotiing the word Cloud, Which has minimum frequency 50, Range of font size

wordcloud(street\_name$Street\_Name, street\_name$Count, min.freq = 50, random.order = F, random.color = T, colors =c("black", "cornflowerblue", "darkred"), scale = c(2.4,.7))

#checking purity of data

aggr(pdata)

#checking uniqueness of variables in data

d<-unique(pdata$OFFENSE\_CODE\_GROUP)

d

o<-unique(pdata$REPORTING\_AREA)

o

head(d,10)

head(o,10)

e<-unique(pdata$DISTRICT)

e

f<-unique(pdata$OFFENSE\_CODE)

f

head(e,10)

head(f,10)

View(pdata) #view the data

#PERFORMING NECESSARY DATA VISUALIZATIONS AND EXPLORATORY DATA ANALYSIS (EDA)

#visualizations and EDA where required

#displaying the map of boston through leaflet package

leaflet() %>%

setView(lng=-71.0892, lat=42.3398, zoom = 10) %>% #using setView method to set the view of map (center and zoom level)

addTiles() %>%

addMarkers(lng=-71.0892, lat=42.3398, popup="Boston")

#displaying the map of boston distributed among crime district by latitude and longitude

#crime mapping

qplot(Long, Lat, data= pdata, color=DISTRICT, geom='point', xlim = c(-71.2,-70.95), ylim= c(42.22,42.4))+

theme\_bw(base\_size=15)+

geom\_point(size = 2)

#number of crimes count per year

ggplot(pdata, aes(x=YEAR, fill= as.factor(YEAR))) +

labs(x="year", y="counts", title = "number of crimes per year")+

geom\_bar()

#number of crimes count per month

ggplot(pdata, aes(x=MONTH, fill= as.factor(MONTH))) +

labs(x="months", y="counts", title = "number of crimes per month")+

geom\_bar()

#number of crimes count per day of the week

ggplot(pdata, aes(x=DAY\_OF\_WEEK, fill= as.factor(DAY\_OF\_WEEK))) +

labs(x="days of the week", y="counts", title = "number of crimes every week")+

geom\_bar()

#dividing the shift into 4 groups and generating six points of the day to bin the day into four equal segments

time\_diff<- c("0","6","12","18","24") #breaking day into 6 interval period

table(pdata$time\_diff) #displaying the crime counts as per hour shift

#plotting the crimes according to the timings

plot\_shift = plot\_ly(pdata, x= ~ time\_diff, color = ~ time\_diff) %>%

add\_histogram() %>%

layout(

title = "Total crimes as per hours shift",

xaxis = list(title = "4 timings shifts",

yaxis = list(title = "Count"))

)

plot\_shift #displaying the crimes counts as per hours shift

#counting street crimes

street\_crime<- sort(table(pdata$STREET), decreasing = TRUE)

head(street\_crime, 10)

#displaying the summary of the dataset

summary(pdata)

#renaming the month names

pdata$MONTH <- as.character(pdata$MONTH)

pdata$MONTH[pdata$MONTH =='1'] <- 'Jan'

pdata$MONTH[pdata$MONTH =='2'] <- 'Feb'

pdata$MONTH[pdata$MONTH =='3'] <- 'Mar'

pdata$MONTH[pdata$MONTH =='4'] <- 'Apr'

pdata$MONTH[pdata$MONTH =='5'] <- 'May'

pdata$MONTH[pdata$MONTH =='6'] <- 'Jun'

pdata$MONTH[pdata$MONTH =='7'] <- 'Jul'

pdata$MONTH[pdata$MONTH =='8'] <- 'Aug'

pdata$MONTH[pdata$MONTH =='9'] <- 'Sep'

pdata$MONTH[pdata$MONTH =='10'] <- 'Oct'

pdata$MONTH[pdata$MONTH =='11'] <- 'Nov'

pdata$MONTH[pdata$MONTH =='12'] <- 'Dec'

pdata$MONTH

l<-sort(table(pdata$STREET),decreasing = TRUE)[1:10] # displaying top 10 streets

l #displaying the streets

m<-sort(table(pdata$OFFENSE\_CODE\_GROUP),decreasing = TRUE)[2:11] # displaying top 10 offense code group

m #displaying the offense code group

n<-sort(table(pdata$REPORTING\_AREA),decreasing=TRUE)[1:10] # displaying top 10 reporting areas

n #displying the reporting areas

headdata<-head(pdata,10)

headdata$newcolumn<-c(25773,22996,18363,17494,15247,15217,14807,13316,11037,10867) #values of highest crimes counts

headdata$newcolumn1<-c("Larceny", "Medical Assistance", "Investigate Person", "Other", "Simple Assault","Vandalism","Drug Violation", "Verbal Disputes", "Towed", "Investigate Property")

#displying top 10 offense count

ggplot(headdata, aes(x=newcolumn1, y=newcolumn, fill= newcolumn1))+

geom\_bar(stat = "identity") +

coord\_flip()+ #flipping the cartesian coordinates

labs(y = "Type of offense", x = "Count",title ="top 10 offense count")

headdata$newcolumn2<-c(14237,7156,7131,5146,4783,4528,4511,4386, 3899,3501 ) #values of crimes counts in streets

headdata$newcolumn3<-c("WASHINGTON ST","BLUE HILL AVE","BOYLSTON ST","DORCHESTER AVE","TREMONT ST", "MASSACHUSETTS AVE", "HARRISON AVE","CENTRE ST","COMMONWEALTH AVE","HYDE PARK AVE")

#displaying crime counts of top areas involved

ggplot(headdata,aes(x=newcolumn3,y=newcolumn2, fill=newcolumn3))+

geom\_bar(stat="identity")+

coord\_flip()+ #flipping the cartesian coordinates

labs(y="count", x="area name",title="top areas involved")

#considering only 2016 data and showing highest offense codes reported

graph1<-filter(pdata,YEAR==2016) #filtering 2016 year offense codes

table(graph1$OFFENSE\_CODE\_GROUP) #displaying counts of all offense codes

ocg1<-sort(table(graph1$OFFENSE\_CODE\_GROUP),decreasing = TRUE)[2:11] #taking top 10 offense code group

headdata$newcolumn4<-c(7903,6978,5765,5538,5284,5063,4744,4100,3521,3360)

headdata$newcolumn5<-c("larceny","medical assistance","Investigate Person","Other","Drug Violation", "Vandalism", "simple assault","verbal disputes","motor vehicle larency","investigate property")

#generating the bar chart of 2016 highest reported areas

ggplot(headdata,aes(x=newcolumn5,y=newcolumn4, fill=newcolumn5))+

geom\_bar(stat="identity")+

coord\_flip()+ #flipping the cartesian coordinates

labs(y="area name", x="count",title="2016 highest reported areas")

sum(headdata$newcolumn4) #displaying the total sum of top 10 2016 offense codes

#generating the bar chart of 2017 highest reported areas

graph2<-filter(pdata,YEAR==2017) #filtering 2017 year offense codes

table(graph2$OFFENSE\_CODE\_GROUP)

ocg2<-sort(table(graph2$OFFENSE\_CODE\_GROUP),decreasing = TRUE)[2:11] #taking top 10 offense code groups for 2017 year

headdata$newcolumn6<-c(7817,7812,6659,5323,4898,4838,4761,4437,3973,3951)

headdata$newcolumn7<-c("medical assistance","larceny","Investigate Person","Other","simple assault", "Vandalism", "drug violation","verbal disputes","investigate property","towed")

ggplot(headdata,aes(x=newcolumn7,y=newcolumn6, fill=newcolumn7))+

geom\_bar(stat="identity")+

coord\_flip()+ #flipping the cartesian coordinates

labs(y="area name", x="count",title="2017 highest reported areas")

sum(headdata$newcolumn6) #displaying the total sum of top 10 2017 offense codes

#2018

graph3<-filter(pdata,YEAR==2018) #filtering 2018 year offense codes

table(graph3$OFFENSE\_CODE\_GROUP)

ocg3<-sort(table(bbb$OFFENSE\_CODE\_GROUP),decreasing = TRUE)[2:11]#taking top 10 offense code groups for 2018 year

headdata$newcolumn8<-c(6292,5949,4372,4241,4013,3692,3433,3208,2847,2899)

headdata$newcolumn9<-c("medical assistance","larceny","Other","Investigate Person","simple assault","Drug Violation","verbal disputes", "Vandalism","investigate property","towed")

ggplot(headdata,aes(x=newcolumn9,y=newcolumn8, fill=newcolumn9))+

geom\_bar(stat="identity")+

coord\_flip()+ #flipping the cartesian coordinates

labs(y="area name", x="frequency",title="2018 highest reported areas")

sum(headdata$newcolumn8) #displaying the total sum of top 10 2018 offense codes

#as we see larceny is occuring most often in all the years, we further dig deeper to see what kind of larceny happens the most for all the 3 years

ldff<-filter(pdata,OFFENSE\_CODE\_GROUP=='Larceny') #filtering the larcency offense group

#displaying the count of larceny offense description

ggplot(ldff, aes(x=OFFENSE\_DESCRIPTION),fill=as.factor(OFFENSE\_DESCRIPTION)) +

coord\_flip()+ #flipping the coordinates

geom\_bar()

#HYPOTHESIS TESTING PAIRED SAMPLE T-TEST

set.seed(7) #to set the sample selection

pdata.sample <- sample\_n(pdata,30, replace = TRUE)#select random 30 samples

monday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Monday")#subset table when day is monday

tuesday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Tuesday")#subset table when day is tuesday

wednesday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Wednesday")#subset table when day is wednesday

thursday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Thursday")#subset table when day is thursday

friday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Friday")#subset table when day is friday

saturday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Saturday")#subset table when day is saturday

sunday<-subset(pdata.sample,subset = DAY\_OF\_WEEK=="Sunday")#subset table when day is sunday

monday\_count <- count(monday)#count of the mondays

tuesday\_count <- count(tuesday)#count of the tuesdays

wednesday\_count <- count(wednesday)#count of the wednesdays

thursday\_count <- count(thursday)#count of the thursdays

friday\_count <- count(friday)#count of the fridays

saturday\_count <- count(saturday)#count of the saturdays

sunday\_count <- count(sunday)#count of the sundays

weekday.counts <- c(monday\_count,tuesday\_count,wednesday\_count,thursday\_count) #net weekdays values

weekend.counts <- c(friday\_count, saturday\_count, sunday\_count) #net weekdends values

weekday.counts <- as.numeric(as.character(weekday.counts)) #weekdays numeric

weekend.counts <- as.numeric(as.character(weekend.counts)) #weekends numeric

test.paired <- t.test(weekday.counts, weekend.counts, mu=0, alternative = "two.sided", paired = F, conf.level = 0.99) #confidence level 99%

test.paired # t - test

#2 sample T-test conclusion

#Ho: weekday crime = weekend crime

#Ha: weekday crime != weekend crime

#p-value = 0.6131 > 0.05

#do not reject null hypothesis

#weekday crime = weekend crime with 99% confidence

#TIME SERIES ANALYSIS

crimes <- pdata#creating another variable for the crime data

names(crimes)

# Finding the class of column OCCURED\_ON\_DATE

class(crimes$OCCURRED\_ON\_DATE)

# As the class of column OCCURED\_ON\_DATE is factor we are converting it into date formart (year-month-day)

crimes$OCCURRED\_ON\_DATE <- as.Date(crimes$OCCURRED\_ON\_DATE, format="%Y-%m-%d")

# The colum has both dates and times so now we are dividing it into only dates

dates<-cut(crimes$OCCURRED\_ON\_DATE, 'day')

# Now we are having the counts of each date which represents the number of times

tab.dates<- table(dates)

#converting it into data frame with its frequency

crimes.dates<-data.frame(Date=format(as.Date(names(tab.dates)), '%d/%m/%Y'),

Frequency=as.vector(tab.dates))

# Having a look at the data frame

head(crimes.dates)

#converting it into the time series, starting at the year 2015-June and 167th day of the year with frequency 365 days

crimes.ts<- ts((crimes.dates$Frequency),start=c(2015,6,167),frequency =365 )

#having a look at the time series

head(crimes.ts)

#ploting the time series graph

plot(crimes.ts,main="Time Series of Crimes from 2015-2018")

# Decomposing the time series into 3 other components trend,seasonal & random. To find how the trend shifts

plot(decompose(crimes.ts))

# Using auto arima to find the best order for the arima model with lowest aic value

# USing Tracr=TRUE reports the list of Arima models considered

mymodel<- auto.arima(crimes.ts,ic="aic",trace = TRUE)

mymodel

#Using the best ARIMA model we are trying to forecast the next 365days trend in crime rate with 99% confidence interval

fut.crimes <- forecast(mymodel,level=c(99),h=365)

#plotting the graph of forecasting time series

plot(fut.crimes)

# Computing Holt-Winters Filtering of a given time series

hw.crimes<- HoltWinters(crimes.ts)

#using predict to fuction from results of model fitting, we are predicting next 356 daily crime rates

hw<- predict(hw.crimes,n.ahead = 365)

#having a loot at the predicted values

head(hw)

# Ploting the predicted values of time series with the past years time series. with dotted line

ts.plot(crimes.ts,hw,col="blue",lty=1:3,main="Predicting Using HoltWinters Model")

#END