BOSTON CRIME - TIME SERIES ANALYSIS

PRANAV SAI REDDY GUDIPALLY- 001404028

HARSH SAMANI - 001403994

ANUJ GOLESAR- 001408549

NORTHEASTERN UNIVERSITY

MPS - ANALYTICS

INTERMEDIATE ANALYTICS

INSTRUCTOR NAME- ROY WADA

Date- 31/3/2019



**INTRODUCTION**

Massachusetts stands 9th in the list of safest states in United States of America. The increasing crime rate In Boston has becomes a major part. It costs and affects just about everyone to some degree. In Boston Crime Record Time-series analysis, our aim is to determine the variation of crime rate in coming future. We have selected Boston crime record data set. The inspiration behind this is to answer how has crime changed over the years, in which area most crimes are committed and how we can gather insights on crime rate, to make it safer for residents and provide aid to Police Department.

**Research questions:**

1. **what will be the change(s) in crime rate in coming future?**
2. **Which are the top areas of the city are involved in the crime?**
3. **What are the kind of offenses that stand on the top (top 10)?**
4. **During which hours the crime rate is the highest, the lowest and is moderate?**

**Dataset used:**

We have selected [Boston Crime Records dataset](https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/12cb3883-56f5-47de-afa5-3b1cf61b257b) which contains records from the new crime incident report system , which include reduced set of field focused on capturing the type of incident as well as when it occurred.

**Contents:**

This dataset has 2,60,760 rows and 16 columns.

* INCIDENT\_NUMBER
* OFFENSE\_CODE
* OFFENSE\_CODE\_GROUP
* OFFENSE\_DESCRIPTION
* DISTRICT
* REPORTING\_AREA
* SHOOTING
* OCCURRED\_ON\_DATE
* YEAR
* MONTH
* DAY\_OF\_WEEK
* HOUR
* STREET
* LATITUDE
* LONGITUDE
* LOCATION

**We have considered following variable for analysis part:**

1. OFFENSE\_CODE
2. OCCURRED\_ON\_DATE
3. HOUR
4. STREET
5. LOCATION
6. YEAR
7. DISTRICT

**Data Preparation:** Data preparation is generally used to get an overview of the data set. Like the names of the attributes, first 5 values all these factors are really important to perform any operation on the data set.

getwd()

## [1] "/Users/harshsamani/Desktop/Masters/Intermediate Analytics/Group Project"

setwd("/Users/harshsamani/Desktop/Masters/Intermediate Analytics/Group Project/")  
dir()

## [1] "AGGR.png"   
## [2] "Assignment 6.R"   
## [3] "Dataset.csv"   
## [4] "Final Report.Rmd"   
## [5] "Final\_Report\_files"   
## [6] "Final\_Report.Rmd"   
## [7] "INTERMEDIATE ANALYTICS GROUP PROJECT.docx"  
## [8] "tmpbfk03plv.csv"

pdata<-read.csv("Dataset.csv")  
head(pdata)

## INCIDENT\_NUMBER OFFENSE\_CODE OFFENSE\_CODE\_GROUP  
## 1 I192021064 1847 Drug Violation  
## 2 I192021064 2102 Operating Under the Influence  
## 3 I192021064 2405 Disorderly Conduct  
## 4 I192021061 3301 Verbal Disputes  
## 5 I192021059 3114 Investigate Property  
## 6 I192021056 3207 Property Found  
## OFFENSE\_DESCRIPTION DISTRICT REPORTING\_AREA  
## 1 DRUGS - POSS CLASS C - INTENT TO MFR DIST DISP NA  
## 2 OPERATING UNDER THE INFLUENCE DRUGS NA  
## 3 DISORDERLY CONDUCT NA  
## 4 VERBAL DISPUTE E5 549  
## 5 INVESTIGATE PROPERTY C11 355  
## 6 PROPERTY - FOUND A7 19  
## SHOOTING OCCURRED\_ON\_DATE YEAR MONTH DAY\_OF\_WEEK HOUR UCR\_PART  
## 1 3/21/19 16:34 2019 3 Thursday 16 Part Two  
## 2 3/21/19 16:34 2019 3 Thursday 16 Part Two  
## 3 3/21/19 16:34 2019 3 Thursday 16 Part Two  
## 4 3/21/19 21:35 2019 3 Thursday 21 Part Three  
## 5 3/21/19 22:10 2019 3 Thursday 22 Part Three  
## 6 3/21/19 21:59 2019 3 Thursday 21 Part Three  
## STREET Lat Long Location  
## 1 NA NA (0.00000000, 0.00000000)  
## 2 NA NA (0.00000000, 0.00000000)  
## 3 NA NA (0.00000000, 0.00000000)  
## 4 WASHINGTON ST 42.26181 -71.15641 (42.26181177, -71.15640773)  
## 5 GIBSON ST 42.29756 -71.05971 (42.29755533, -71.05970910)  
## 6 SARATOGA ST 42.37865 -71.03089 (42.37864819, -71.03089238)

names(pdata)

## [1] "INCIDENT\_NUMBER" "OFFENSE\_CODE" "OFFENSE\_CODE\_GROUP"   
## [4] "OFFENSE\_DESCRIPTION" "DISTRICT" "REPORTING\_AREA"   
## [7] "SHOOTING" "OCCURRED\_ON\_DATE" "YEAR"   
## [10] "MONTH" "DAY\_OF\_WEEK" "HOUR"   
## [13] "UCR\_PART" "STREET" "Lat"   
## [16] "Long" "Location"

**Data Cleaning & Checking:** As we all know data cleaning and checking the uniqueness of the variables in the data set is the most crucial step to perform before any operation starts. Data cleaning can be done in many ways like replacing the null value with the 0, omitting the null values. Below is the code for same.

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

## Warning: package 'VIM' was built under R version 3.5.2

## Loading required package: colorspace

## Warning: package 'colorspace' was built under R version 3.5.2

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.5.2

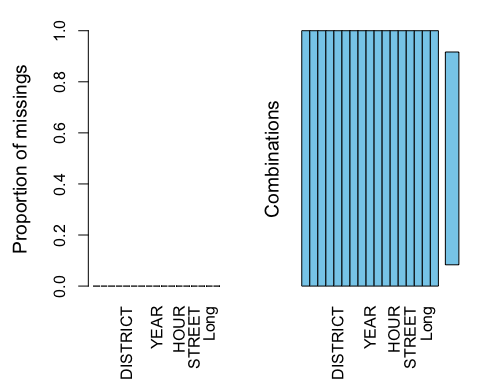
## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

#checking purity of data  
a<-aggr(pdata)



Looking at the above graph we can say that the data is pure as there is no data in the left plot which stands for the proportion of missing data

##   
## Missings in variables:  
## [1] Variable Count   
## <0 rows> (or 0-length row.names)

Here we are looking at the data variables and their uniqueness to get a heads up on what kind of data is in the variables and how many different kinds of data are there that keep repeating, we can see that there are 67 kinds of offenses in the given offense code group column and likely we look the same for different variables

#checking uniqueness of variables in data  
d<-unique(pdata$OFFENSE\_CODE\_GROUP)  
o<-unique(pdata$REPORTING\_AREA)  
  
head(d,10)

## [1] Drug Violation Operating Under the Influence  
## [3] Disorderly Conduct Verbal Disputes   
## [5] Investigate Property Property Found   
## [7] Violations Aggravated Assault   
## [9] Simple Assault Larceny   
## 67 Levels: Aggravated Assault Aircraft ... Warrant Arrests

head(o,10)

## [1] 0 549 355 19 498 342 128 358 121 465

e<-unique(pdata$DISTRICT)  
f<-unique(pdata$OFFENSE\_CODE)  
  
head(e,10)

## [1] E5 C11 A7 E18 D4 A1 B3 E13 C6   
## Levels: A1 A15 A7 B2 B3 C11 C6 D14 D4 E13 E18 E5

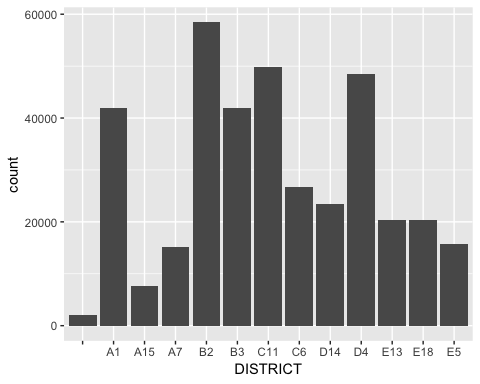
head(f,10)

## [1] 1847 2102 2405 3301 3114 3207 2907 413 802 617

In the above code we have taken unique values from the variables OFFENSE CODE GROUP, REPORTING AREA, DISTRICT, OFFENSE CODE and displayed the top 10 values for it.

**Visualizations:**

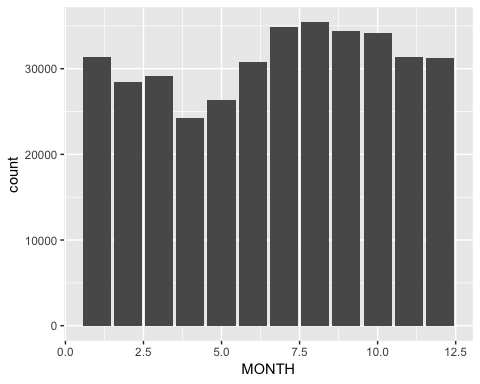
DISTRICTS



Interpretation

We can see from the graph that the district B2 has the highest reported crimes according to a given data set closing around 60,000 reported incidents in one year and A15 to be having the least count around 6000 incidents being reported, looking at this we can say that B2 is the most unsafe district to live in while A15 is the safest

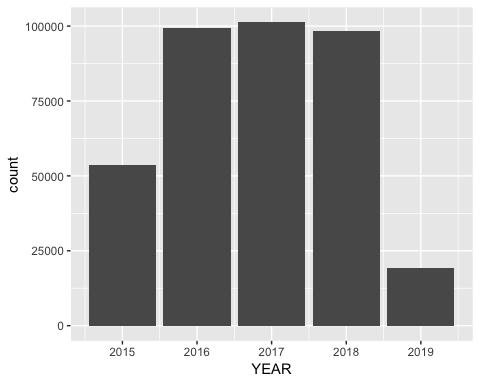
MONTHS



Interpretation

Looking at the above plot we shall infer that most crimes happened have occurred at the time of post summer and early fall, that is around July to October and minimal crimes during the snow and early spring that is December and starting few months of the year

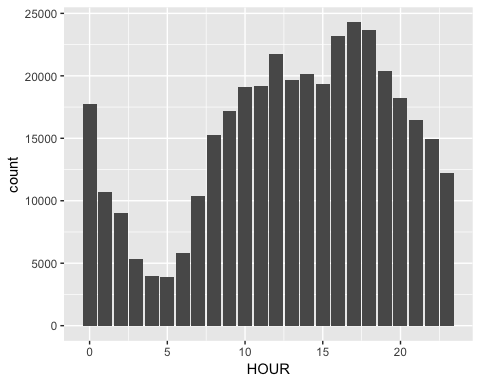
YEARS



Interpretation

Keeping in mind the kind of dataset we have where we have complete data about the years 2016,17 and 18, the graph goes up obviously, we shall not say looking at the graph that 2015 has the least crime rate as we have the data available only after august 2015, considering the 3 years where we have complete data, we shall say that there have been close to 100,000 crimes that happen every year but in the year 2017, it had exceeded 100,000 and stands out as the year for the highest ever reported crimes within the given dataset with complete data throughout the year

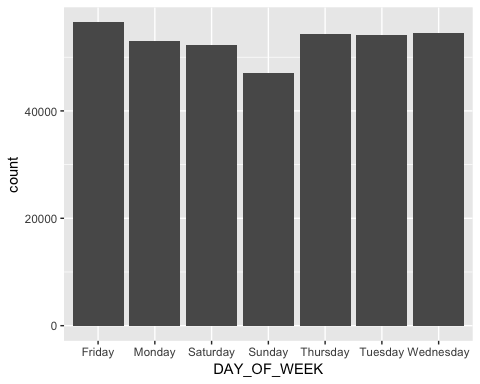
TIME OF THE DAY



Interpretation

We can say looking at the above bar graph that most crimes happening have been reported at around 5PM and 6PM in the evening which we can say is the duration when a usual person is done with their day at work and returning home, and the minimal crimes happen around early in the morning at 4AM and 5AM

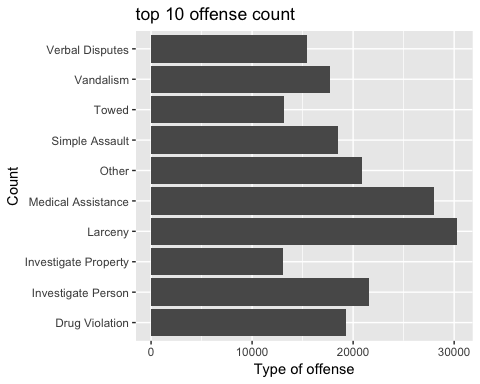
DAY OF THE WEEK



Interpretation

Looking at the bar graph we can say that there are almost equal number of crimes that occur irrespective of the day of the week but the start of the weekend, that is Friday sees the highest ever crime count where crime shall be reported, Fridays alone have the highest count around 56,000 crime scenes that have been reported in the 3-4 years span (duration of data available for us in the data set) and Sunday has the lowest around 45,000

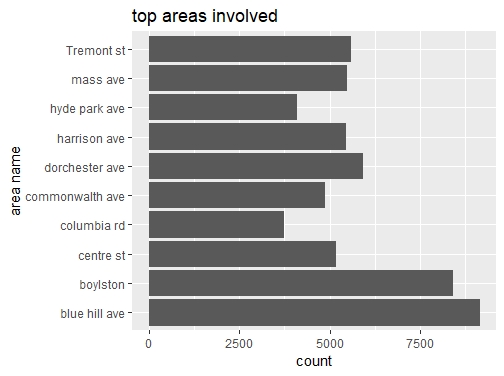
OFFENSES FREQUENCY



Interpretation

The bar graph above depicts the top 10 most occurring(frequent) type of crimes in Boston, out of all the 67 kinds of offenses, Larceny stands out to be the type of crime that happens the most, there have been over 30,000 reported incidents of larceny(personal items theft) and Towing is the least of around 12,500

AREAS INVOLVED



Interpretation

Looking at the above graph, we can say that blue hill avenue street in Boston is the most unsafe area to live in as there have been close to 9000 crime related incidents that have been reported which means close to 7 crimes being reported each and every day, in contrary, Columbia road is the most safest street to live in as it has seen 3750 crimes in a span of 4 years which is close to 3 crimes a day

In the below graph we have plotted bar graph for OFFENSE CODE VARIABLE for three years 2016, 2017, 2018 and we have also used sort() command to filter the data in descending order to get the count of the highest number of offense happened in Boston.

#considering only 2016 data and showing highest offense codes reported  
  
library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ tibble 2.0.1 ✔ purrr 0.3.1   
## ✔ tidyr 0.8.3 ✔ dplyr 0.8.0.1  
## ✔ readr 1.3.1 ✔ stringr 1.4.0   
## ✔ tibble 2.0.1 ✔ forcats 0.4.0

## Warning: package 'tibble' was built under R version 3.5.2

## Warning: package 'tidyr' was built under R version 3.5.2

## Warning: package 'purrr' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

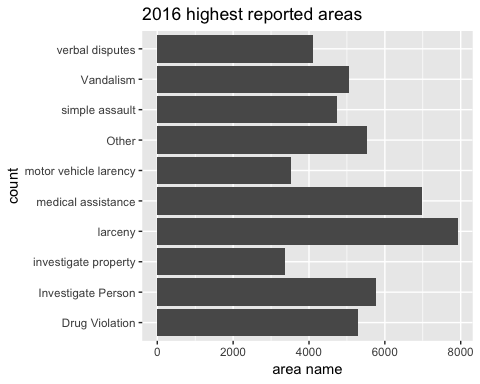
## Warning: package 'forcats' was built under R version 3.5.2

## ── Conflicts ─────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::between() masks data.table::between()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::first() masks data.table::first()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::last() masks data.table::last()  
## ✖ purrr::transpose() masks data.table::transpose()

We are creating a data frame (ggg) of data of just the data in the variabes that map to the year 2016

ggg<-filter(pdata,YEAR==2016)   
table(ggg$OFFENSE\_CODE\_GROUP)

2016 CRIME DATA



sum(headdata$newcolumn4)

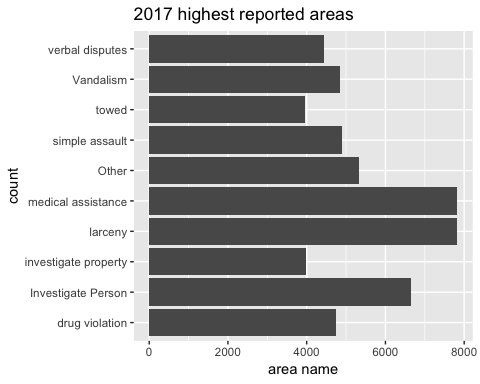
## [1] 52304

#2017 dataframe  
  
aaa<-filter(pdata,YEAR==2017)  
ocg2<-sort(table(aaa$OFFENSE\_CODE\_GROUP),decreasing = TRUE)[2:11]

Interpretation

In the top 10 kinds of offenses happening in Boston, Larceny still stands out to be the most frequently occurring crime even in 2016 but investigation of personal property, not towing stands as the least occurred crime

2017 CRIME DATA



sum(headdata$newcolumn6)

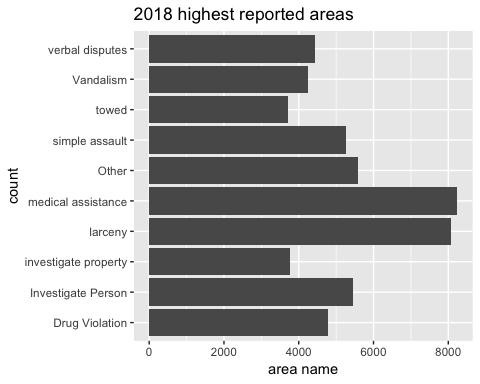
## [1] 54496

#2018 dataframe  
bbb<-filter(pdata,YEAR==2018)  
ocg3<-sort(table(bbb$OFFENSE\_CODE\_GROUP),decreasing = TRUE)[2:11]  
ocg3

Interpretation

In the top 10 kinds of offenses happening in Boston, Larceny and Medical assistance stand out to be the most frequently occurring crimes even in 2017 and towing stands as the least occurred crime as per the year 2017 alone

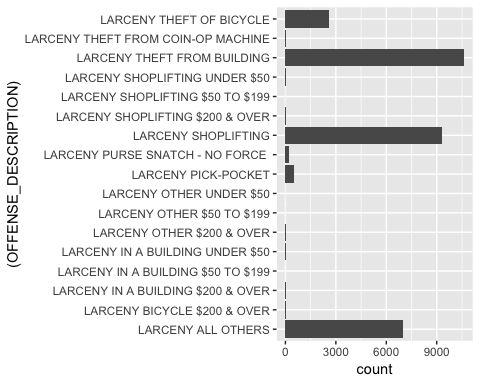
CRIME DATA 2018



Interpretation

In the top 10 kinds of offenses happening in Boston, Medical assistance stands out to be the most frequently occurring crime even in 2018 and towing stands as the least occurred crime

LARCENY TYPES

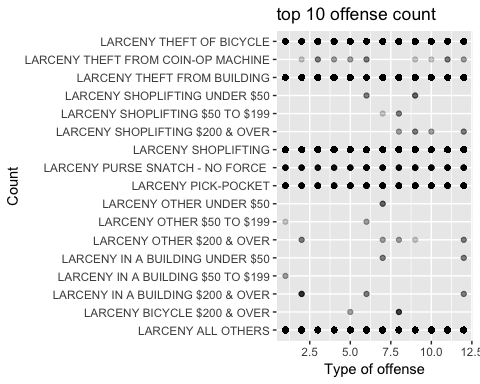


Interpretation

As we can see, larceny from building, that is theft from residences have been happening the highest in Boston and shoplifting is the second highest which happen close to around 10,500 and 9200 time respectively in the span of the 48 months of data that we have in our data set

As we know larceny has happened the most, let’s see at what rate and frequency they occur at different months of the year

LARCENY AND MONTH OF THE YEAR



Interpretation

As we can see that Larceny theft of bicycle, shoplifting, purse snatch, pick pocket happen throughout the year with the same frequency and other kinds like costly worth shoplifting happen around December the most and larceny from building the most frequent in the month of February

**Time Series Analysis**

In this data set, we have used time series analysis to forecast the future, whether the crime will grow or it will get low considering the variable OFFENSE CODE GROUP with the count. After performing time series analysis on the data set what we found is that offense rate is not growing but they are not even getting less it is going to be constant. We have used forecast() function for that, accuracy() function is used to find out the accuracy of the model over the data set. Below is the code for same with the desired output.

offensedata <- ts(tdf,start=c(100))  
  
attributes(offensedata)

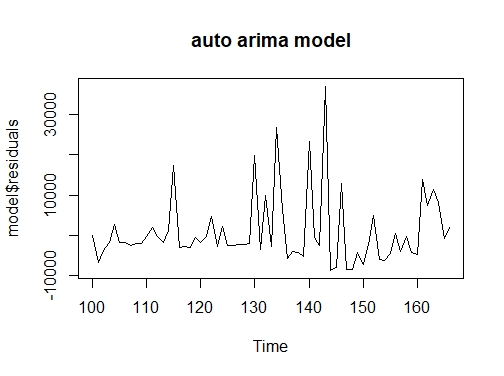
## $dim  
## [1] 67 1  
##   
## $dimnames  
## $dimnames[[1]]  
## NULL  
##   
## $dimnames[[2]]  
## [1] "Freq"  
##   
##   
## $tsp  
## [1] 100 166 1  
##   
## $class  
## [1] "ts"

summary(offensedata)

## Freq   
## Min. : 2.0   
## 1st Qu.: 278.5   
## Median : 1517.0   
## Mean : 5549.0   
## 3rd Qu.: 6663.0   
## Max. :43312.0

cycle(offensedata)

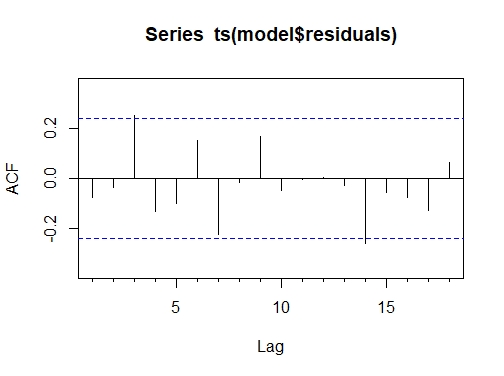
## Time Series:  
## Start = 100   
## End = 166   
## Frequency = 1   
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1



Interpretation

The above graph, From the time plot, it appears that the random fluctuations in the time series are roughly constant in size over time, so an additive model is probably appropriate for describing this time series

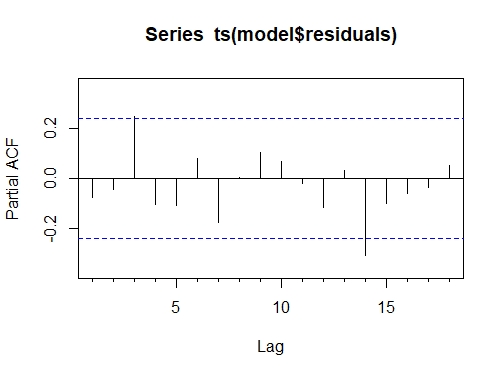
Graph:



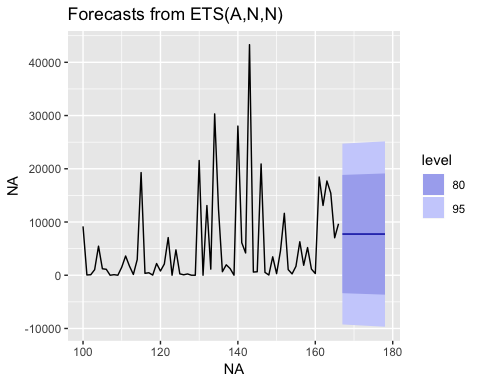
Interpretation

From the above Figure we see from the correlogram that the autocorrelations for lags 14, 3 exceed the significance bounds, and that the autocorrelations tail off to ideal after and before these lags. The autocorrelations for lags 3 are positive around 0.21, and -0.22 for 14.

Graph:



From the partial auto correlogram stated in Figure above, we see that the partial autocorrelation at lag 3 is positive and exceeds the significance bounds (0.21), while the partial autocorrelation at lag 14 is negative and also exceeds the significance bounds (-0.25). The partial autocorrelations tail off to ideal before and after these outliers



The above figure shows the future trend for the birth rate for the coming years. The graph shows constancy in count of values for the frequency of different offenses that may occur

accuracy(offense\_forecast)

## ME RMSE MAE MPE MAPE MASE  
## Training set 1253.646 8534.861 5431.217 -7220.589 7259.556 0.7359871  
## ACF1  
## Training set -0.0840181

Conclusion:

Looking at the graphs above we can say that the count of offenses which matters most in the whole of the data set where we have considered frequency of each of these offenses occurring to do time series forecasting has given us out results that the frequency is going to remain constant in the coming years as well