Module 10 - Trend Analysis

Sooraj Shetty

April 17, 2017

## For this module I will be using my research project for the time series analysis using trend analysis to forecast a prediction. Hence the initial part of the assignment is data wrangling for my project. It will be helpful as a base for creating the prediction model of the project.

Loading the required libraries

library(RCurl)

## Loading required package: bitops

library(plyr)

## Warning: package 'plyr' was built under R version 3.3.3

library(forecast)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: timeDate

## This is forecast 7.3

library(tseries)

## Warning: package 'tseries' was built under R version 3.3.3

Loading the training data for the assignment (and project)

train <- read.csv("train.csv",header = TRUE)

Converting the data columns into factor

# Wrangling the data for use  
  
train$Dates <- as.POSIXct(train$Dates, format = "%Y-%m-%d %H:%M:%S")  
train$Date <- (format(train$Dates, "%d"))  
train$Year <- format(train$Dates, "%Y")  
train$months <- (format(train$Dates, "%m"))  
train$Hours <- (format(train$Dates, "%H"))  
  
train$Category <- as.factor(train$Category)  
train$DayOfWeek <- as.factor(train$DayOfWeek)  
train$PdDistrict <- as.factor(train$PdDistrict)  
train$Resolution <- as.factor(train$Resolution)  
train$Address <- as.factor(train$Address)  
  
str(train)

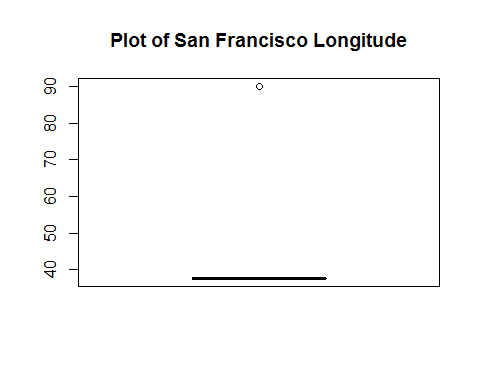
## 'data.frame': 878049 obs. of 13 variables:  
## $ Dates : POSIXct, format: "2015-05-13 23:53:00" "2015-05-13 23:53:00" ...  
## $ Category : Factor w/ 39 levels "ARSON","ASSAULT",..: 38 22 22 17 17 17 37 37 17 17 ...  
## $ Descript : Factor w/ 879 levels "ABANDONMENT OF CHILD",..: 867 811 811 405 405 407 740 740 405 405 ...  
## $ DayOfWeek : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 7 7 7 7 7 7 7 ...  
## $ PdDistrict: Factor w/ 10 levels "BAYVIEW","CENTRAL",..: 5 5 5 5 6 3 3 1 7 2 ...  
## $ Resolution: Factor w/ 17 levels "ARREST, BOOKED",..: 1 1 1 12 12 12 12 12 12 12 ...  
## $ Address : Factor w/ 23228 levels "0 Block of HARRISON ST",..: 19791 19791 22698 4267 1844 1506 13323 18055 11385 17659 ...  
## $ X : num -122 -122 -122 -122 -122 ...  
## $ Y : num 37.8 37.8 37.8 37.8 37.8 ...  
## $ Date : chr "13" "13" "13" "13" ...  
## $ Year : chr "2015" "2015" "2015" "2015" ...  
## $ months : chr "05" "05" "05" "05" ...  
## $ Hours : chr "23" "23" "23" "23" ...

Converting the columns to numeric for the year, date, months and hours

train$Year <- as.numeric(train$Year)  
train$Date <- as.numeric(train$Date)  
train$months <- as.numeric(train$months)  
train$Hours <- as.numeric(train$Hours)

Removing the anomalies/outliers?

boxplot(train$Y, main = "Plot of San Francisco Longitude")



nrow(train)

## [1] 878049

long <- which(train$Y == 90, arr.ind = T)  
length(long)

## [1] 67

train <- train[-long,]  
nrow(train)

## [1] 877982

#removing 47 rows where date is NA  
  
train <- train[-which(is.na(train$Dates)),]  
nrow(train)

## [1] 877935

head(train)

## Dates Category Descript  
## 1 2015-05-13 23:53:00 WARRANTS WARRANT ARREST  
## 2 2015-05-13 23:53:00 OTHER OFFENSES TRAFFIC VIOLATION ARREST  
## 3 2015-05-13 23:33:00 OTHER OFFENSES TRAFFIC VIOLATION ARREST  
## 4 2015-05-13 23:30:00 LARCENY/THEFT GRAND THEFT FROM LOCKED AUTO  
## 5 2015-05-13 23:30:00 LARCENY/THEFT GRAND THEFT FROM LOCKED AUTO  
## 6 2015-05-13 23:30:00 LARCENY/THEFT GRAND THEFT FROM UNLOCKED AUTO  
## DayOfWeek PdDistrict Resolution Address X  
## 1 Wednesday NORTHERN ARREST, BOOKED OAK ST / LAGUNA ST -122.4259  
## 2 Wednesday NORTHERN ARREST, BOOKED OAK ST / LAGUNA ST -122.4259  
## 3 Wednesday NORTHERN ARREST, BOOKED VANNESS AV / GREENWICH ST -122.4244  
## 4 Wednesday NORTHERN NONE 1500 Block of LOMBARD ST -122.4270  
## 5 Wednesday PARK NONE 100 Block of BRODERICK ST -122.4387  
## 6 Wednesday INGLESIDE NONE 0 Block of TEDDY AV -122.4033  
## Y Date Year months Hours  
## 1 37.77460 13 2015 5 23  
## 2 37.77460 13 2015 5 23  
## 3 37.80041 13 2015 5 23  
## 4 37.80087 13 2015 5 23  
## 5 37.77154 13 2015 5 23  
## 6 37.71343 13 2015 5 23

count(train$Year)

## x freq  
## 1 2003 73881  
## 2 2004 73405  
## 3 2005 70741  
## 4 2006 69909  
## 5 2007 68008  
## 6 2008 70168  
## 7 2009 68992  
## 8 2010 66539  
## 9 2011 66616  
## 10 2012 71731  
## 11 2013 75604  
## 12 2014 74760  
## 13 2015 27581

Adding a column as CategoryMap which will have factor values for category

train$CategoryMap <- train$Category  
  
levels(train$CategoryMap) <- gsub("ARSON", 1, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("ASSAULT", 2, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("BAD CHECKS", 3, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("BRIBERY", 4, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("BURGLARY", 5, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("DISORDERLY CONDUCT", 6, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("DRIVING UNDER THE INFLUENCE", 7, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("DRUG/NARCOTIC", 8, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("DRUNKENNESS", 9, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("EMBEZZLEMENT", 10, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("EXTORTION", 11,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("FAMILY OFFENSES", 12,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("FRAUD", 13, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("FORGERY/COUNTERFEITING", 14,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("GAMBLING", 15, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("KIDNAPPING", 16, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("LARCENY/THEFT", 17, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("LIQUOR LAWS", 18,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("LOITERING", 19, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("MISSING PERSON", 20, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("NON-CRIMINAL", 21, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("OTHER OFFENSES", 22, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("PORNOGRAPHY/OBSCENE MAT", 23, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("PROSTITUTION", 24, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("RECOVERED VEHICLE", 25, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("ROBBERY", 26, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("RUNAWAY", 27, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("SECONDARY CODES", 28, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("SEX OFFENSES FORCIBLE", 29, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("SEX OFFENSES NON FORCIBLE", 30,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("STOLEN PROPERTY", 31,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("SUICIDE", 32, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("SUSPICIOUS OCC", 33, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("TREA", 34,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("TRESPASS", 35,levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("VANDALISM", 36, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("VEHICLE THEFT", 37, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("WARRANTS", 38, levels(train$CategoryMap))  
levels(train$CategoryMap) <- gsub("WEAPON LAWS", 39, levels(train$CategoryMap))

# Assingment

*NOTE: It's your choice to submit EITHER Lesson 02 (ARIMA models) OR Lesson 03 (Trend Analysis) for the module Time Series Analysis.*

Answer the following questions:  
\* Can some form of Time series analysis help in your research project to forcast a prediction?  
\* If it can apply Time series analysis to your research project? Does it help?  
\* If (and only if) you can't use some form of Time series analysis help in your research project then apply a form of Time series analysis to a data set you find at the [UC Irvine Machine Learning Repository](http://archive.ics.uci.edu/ml/)

# Trend Analysis

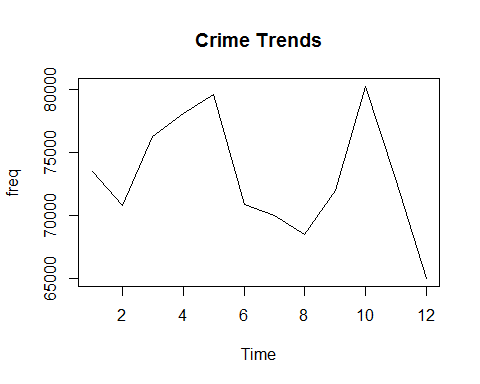
Trend Analysis is the practice of collecting information and attempting to spot a pattern, or trend, in the information. Typically this involves analyzing the variance for a change over time. The null hypothesis: is that there is no trend. Many techniques can be used to identify trends, we'll use an ARMA model again.

# Trend Analysis in R

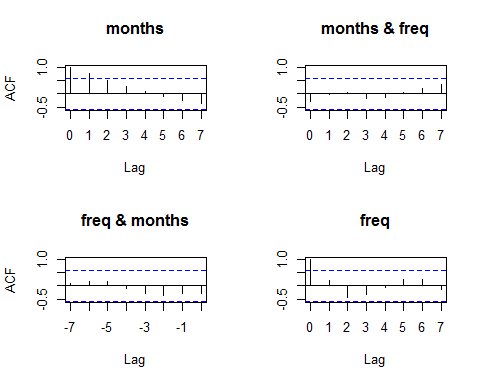
#----------- CRIME TRENDS --------  
crime<- count(train,c("months"))  
head(crime)

## months freq  
## 1 1 73534  
## 2 2 70811  
## 3 3 76279  
## 4 4 78084  
## 5 5 79640  
## 6 6 70882

plot(crime,type='l',xlab='Time',main='Crime Trends')



acf(crime)



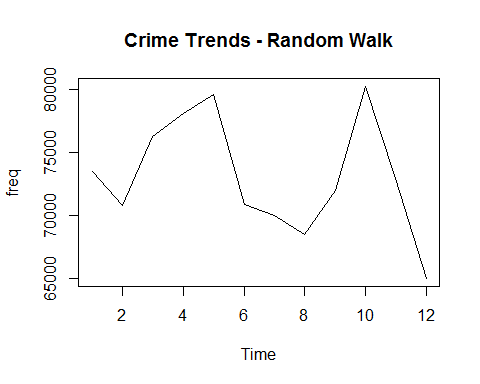
var(crime)

## months freq  
## months 13.000 -5060.955  
## freq -5060.955 21529623.841

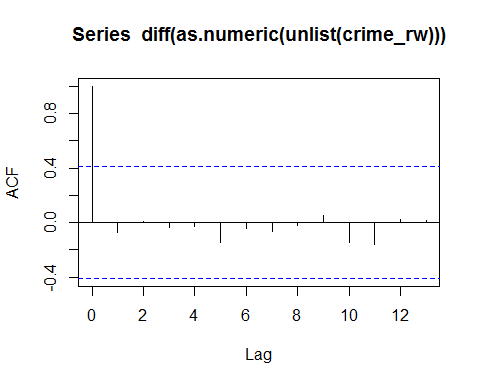
#------------- Random Walk -------------  
crime\_rw<- count(train,c("months"))  
head(crime\_rw)

## months freq  
## 1 1 73534  
## 2 2 70811  
## 3 3 76279  
## 4 4 78084  
## 5 5 79640  
## 6 6 70882

plot(crime\_rw,type='l',xlab='Time',main='Crime Trends - Random Walk')



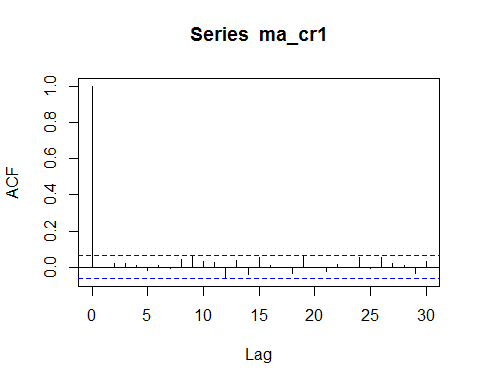
acf(diff(as.numeric(unlist(crime\_rw))))



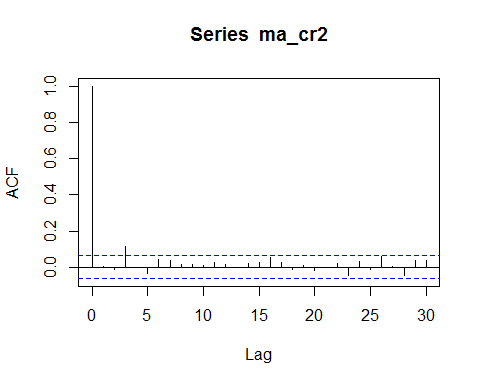
#----------------------------- ARMA model ----------------------  
  
# Moving Average Model  
ma\_cr1 <- arima.sim(model = list(crime\_rw, sd = 1.2), n = 1000)  
head(ma\_cr1, n = 8)

## [1] -1.8247409 -0.2518914 -0.4583773 0.6461076 0.4078227 0.2079539  
## [7] -0.7066349 -1.0907062

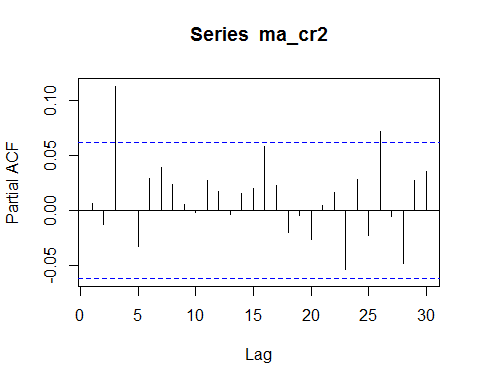
acf(ma\_cr1)



# Autoregressive model  
ma\_cr2 <- arima.sim(model = list(crime\_rw, sd = 1.2), n = 1000)  
acf(ma\_cr2)



pacf(ma\_cr2)



#--------------Dickey-Fuller for stationarity -----------------------  
crime\_rwS<- count(train,c("Year", "months"))  
adf.test(crime\_rwS$months, alternative = "stationary")

## Warning in adf.test(crime\_rwS$months, alternative = "stationary"): p-value  
## smaller than printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: crime\_rwS$months  
## Dickey-Fuller = -8.9938, Lag order = 5, p-value = 0.01  
## alternative hypothesis: stationary

#------------ Another unit root test : Philips-Perron test -------  
PP.test(crime\_rwS$months)

##   
## Phillips-Perron Unit Root Test  
##   
## data: crime\_rwS$months  
## Dickey-Fuller = -6.5431, Truncation lag parameter = 4, p-value =  
## 0.01

# ------------ Seasonal Trend Decomposition in R --------  
  
#The Seasonal Trend Decomposition using Loess (STL) is an algorithm that was developed   
#to help to divide up a time series into three components namely: the trend, seasonality and remainder.  
  
myts <- ts(crime\_rwS[,2], start=c(2009, 1), end=c(2014, 12), frequency=12)   
crime.stl <- stl(myts, s.window="periodic")  
plot(crime.stl)

## Warning in plot.window(xlim, ylim, log, ...): relative range of values = 17  
## \* EPS, is small (axis 2)

