Crimes <- read.csv("C:/Users/potdarjs/Downloads/Crimes\_-\_2001\_to\_present.csv", header=FALSE)

View(Crimes)

names(Crimes) <- c("Case", "Number", "Date", "Block", "IUCR", "Primary Type", "Description",

"Location Desc", "Arrest", "Domestic", "Beat", "District", "Ward", "Community Area",

"FBI Code", "X Coordinate", "Y Coordinate", "Year", "Updated On",

"Latitude", "Longitude", "Location")

head(Crimes)

str(Crimes)

# -------------------------------------------------------------------------------------

#a. Find out top 5 attributes having highest correlation (select only Numeric features).

Crimes <- na.omit(Crimes)

names(Crimes)

c <- cor(Crimes[c(11,12,13,14,18,20,21)])

c

library(reshape2)

m <- melt(c)

library(dplyr)

m

top <- m%>%select(X1, X2, value)%>%filter(value != 1)

top[order(top$value, decreasing = T)[1:10],]

# District~Beat, Ward~District, Ward~Beat, Latitude ~Ward, Latitude~District are top5 attributes with highest correclation

# -------------------------------------------------------------------------------------

# Find out top 3 reasons for having more crime in a city.

x <- as.data.frame(table(Crimes$Description))

x[order(x$Freq, decreasing = T)[1:3],]

# Simple, $500 and Under and Domestic Battery Simple are the top 3 reasons for having more crime

# -------------------------------------------------------------------------------------

# Which all attributes have correlation with crime rate?

# The data we have are at a crime incident level, that is, for each recorded each crime incident

# Hence we need to transform it to workable model

crime <- Crimes

head(crime)

table(is.na(crime))

crime$Date <- as.POSIXlt(crime$Date, format= "%m/%d/%Y %H:%M:%S")

crime$`Updated On` <- as.POSIXlt(crime$`Updated On`, format= "%m/%d/%Y %H:%M:%S")

library(chron)

crime$Time <- times(format(crime$Date,"%H:%M:%S"))

crime$Date <- as.POSIXct(crime$Date)

crime$`Updated On` <- as.POSIXct(crime$`Updated On`)

# There could be certain time intervals of the day where criminal activity is more prevalent

time.tag <- chron(times=c("00:00:00", "06:00:00", "12:00:00", "18:00:00","23:59:00"))

time.tag

crime$time.tag <- cut(crime$Time, breaks= time.tag,

labels= c("00-06","06-12", "12-18", "18-00"), include.lowest =TRUE)

table(crime$time.tag)

# date variable to contain just the date part

crime$date <- as.POSIXlt(strptime(crime$Date, format = "%Y-%m-%d"))

crime$date <- as.POSIXct(crime$date)

# days and months could be predicatble variable

crime$day <- as.factor(weekdays(crime$Date, abbreviate = TRUE))

crime$month <- as.factor(months(crime$Date, abbreviate = TRUE))

str(crime$day)

str(crime$month)

# converting Arrest yes / no to binary varibale

crime$Arrest <- ifelse(as.character(crime$Arrest) == "true",1,0)

# The data contain about 31 crime types, not all of which are mutually exclusive. We can combine

# two or more similar categories into one to reduce this number and make the analysis a bit easier.7

crime$crime <- as.character(crime$`Primary Type`)

crime$crime <- ifelse(crime$crime %in% c("CRIM SEXUAL ASSAULT","PROSTITUTION", "SEX OFFENSE","HUMAN TRAFFICKING"), 'SEX', crime$crime)

crime$crime <- ifelse(crime$crime %in% c("MOTOR VEHICLE THEFT"), "MVT", crime$crime)

crime$crime <- ifelse(crime$crime %in% c("GAMBLING", "INTERFEREWITH PUBLIC OFFICER", "INTERFERENCE WITH PUBLIC OFFICER", "INTIMIDATION",

"LIQUOR LAW VIOLATION", "OBSCENITY", "NON-CRIMINAL", "PUBLIC PEACE VIOLATION",

"PUBLIC INDECENCY", "STALKING", "NON-CRIMINAL (SUBJECT SPECIFIED)","NON - CRIMINAL"),

"NONVIO", crime$crime)

crime$crime <- ifelse(crime$crime == "CRIMINAL DAMAGE", "DAMAGE",crime$crime)

crime$crime <- ifelse(crime$crime == "CRIMINAL TRESPASS","TRESPASS", crime$crime)

crime$crime <- ifelse(crime$crime %in% c("NARCOTICS", "OTHER NARCOTIC VIOLATION", "OTHER NARCOTIC VIOLATION"), "DRUG", crime$crime)

crime$crime <- ifelse(crime$crime == "DECEPTIVE PRACTICE","FRAUD", crime$crime)

crime$crime <- ifelse(crime$crime %in% c("OTHER OFFENSE", "OTHEROFFENSE"), "OTHER", crime$crime)

crime$crime <- ifelse(crime$crime %in% c("KIDNAPPING", "WEAPONS VIOLATION", "CONCEALED CARRY LICENSE VIOLATION","OFFENSE INVOLVING CHILDREN"), "VIO", crime$crime)

table(crime$crime)

# A potential important indicator of criminal activity in a particular area could be the history of

# criminal activities in the past.

temp <- aggregate(crime$crime, by=list(crime$crime, crime$time.tag), FUN=length)

names(temp) <- c("crime", "time.tag", "count")

library(plyr)

temp <- ddply(crime, .(crime, day), summarise, count = length(date))

library(doBy)

temp <- summaryBy(Case ~ crime + month, data = crime, FUN= length)

names(temp)[3] <- 'count'

crime.agg <- ddply(crime, .(crime, Arrest, Beat, date, `X Coordinate`, `Y Coordinate`, time.tag, day, month),

summarise, count=length(date), .progress='text')

beats <- sort(unique(crime.agg$Beat))

dates <- sort(as.character(unique(crime.agg$date)))

temp <- expand.grid(beats, dates)

names(temp) <- c("Beat", "date")

model.data <- aggregate(crime.agg[, c('count', 'Arrest')], by=

list(crime.agg$Beat, as.character(crime.agg$date)), FUN=sum)

names(model.data) <- c("Beat", "date", "count", "Arrest")

model.data <- merge(temp, model.data, by= c('Beat', 'date'), all.x= TRUE)

View(model.data)

model.data$count[is.na(model.data$count)] <- 0

model.data$Arrest[is.na(model.data$Arrest)] <- 0

model.data$day <- weekdays(as.Date(model.data$date), abbreviate= TRUE)

model.data$month <- months(as.Date(model.data$date), abbreviate= TRUE)

pastDays <- function(x) {c(0, rep(1, x))}

model.data$past.crime.1 <- ave(model.data$count, model.data$Beat,

FUN=function(x) filter(x, pastDays(1), sides= 1))

model.data$past.crime.7 <- ave(model.data$count, model.data$Beat,

FUN=function(x) filter(x, pastDays(7), sides= 1))

model.data$past.crime.30 <- ave(model.data$count, model.data$Beat,

FUN=function(x) filter(x, pastDays(30), sides= 1))

meanNA <- function(x){mean(x, na.rm= TRUE)}

model.data$past.crime.1 <- ifelse(is.na(model.data$past.crime.1),

meanNA(model.data$past.crime.1), model.data$past.crime.1)

model.data$past.crime.7 <- ifelse(is.na(model.data$past.crime.7),

meanNA(model.data$past.crime.7), model.data$past.crime.7)

model.data$past.crime.30 <- ifelse(is.na(model.data$past.crime.30),

meanNA(model.data$past.crime.30), model.data$past.crime.30)

# past variables for arrests

model.data$past.arrest.30 <- ave(model.data$Arrest, model.data$Beat,

FUN= function(x) filter(x, pastDays(30), sides= 1))

model.data$past.arrest.30 <- ifelse(is.na(model.data$past.arrest.30),

meanNA(model.data$past.arrest.30), model.data$past.arrest.30)

# arrests per crime

model.data$policing <- ifelse(model.data$past.crime.30 == 0, 0,

model.data$past.arrest.30/model.data$past.crime.30)

# trend

model.data$crime.trend <- ifelse(model.data$past.crime.30 == 0, 0,

model.data$past.crime.7/model.data$past.crime.30)

# season could be another reason

model.data$season <- as.factor(ifelse(model.data$month %in% c("Mar", "Apr", "May"), "spring",

ifelse(model.data$month %in% c("Jun", "Jul", "Aug"), "summer",

ifelse(model.data$month %in% c("Sep", "Oct","Nov"), "fall", "winter"))))

model.cor <- cor(model.data[, c("count", "past.crime.1", "past.crime.7",

"past.crime.30","policing", "crime.trend")])

model.cor

psych::cor.plot(model.cor)

mean(model.data$count)

var(model.data$count)

# The variance is much greater than the mean indicating that the distribution is overdispersed.

# Asuitable way to model for such overdispersion is using the negative binomial distribution

library(MASS)

model <- glm.nb(count ~past.crime.1 + past.crime.7 + past.crime.30 +

+ policing + crime.trend + factor(day) + season, data= model.data)

summary(model)

# All the varibales considered in the model have significant relation with the crime.

#---------------------------

write.csv(model.data, "E:\\Data Analytics with RET\\Assignment\\crimemodel.csv")