Project on Chicago

## Prince Agyabeng

Loading the Libraries

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(ggcorrplot)  
library(mice)

## Loading required package: lattice

##   
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':  
##   
## complete

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

## Registered S3 methods overwritten by 'car':  
## method from  
## influence.merMod lme4  
## cooks.distance.influence.merMod lme4  
## dfbeta.influence.merMod lme4  
## dfbetas.influence.merMod lme4

## 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

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday, week,  
## yday, year

## The following object is masked from 'package:base':  
##   
## date

library(ggcorrplot)  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(leaps)  
library(caret)

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(rpart)   
library(RColorBrewer)   
library(rattle)

## Rattle: A free graphical interface for data science with R.  
## Version 5.3.0 Copyright (c) 2006-2018 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

##   
## Attaching package: 'rattle'

## The following object is masked from 'package:VIM':  
##   
## wine

library(ranger)

##   
## Attaching package: 'ranger'

## The following object is masked from 'package:rattle':  
##   
## importance

library(nnet)

Reading the data

chicago <- read\_csv("chicago.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## ID = col\_double(),  
## Arrest = col\_logical(),  
## Domestic = col\_logical(),  
## Beat = col\_double(),  
## District = col\_double(),  
## Ward = col\_double(),  
## `Community Area` = col\_double(),  
## `X Coordinate` = col\_double(),  
## `Y Coordinate` = col\_double(),  
## Year = col\_double(),  
## Latitude = col\_double(),  
## Longitude = col\_double()  
## )

## See spec(...) for full column specifications.

Examining the struture and summary of the data

str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 22 variables:  
## $ ID : num 11948745 11945748 11944351 11933173 11924359 ...  
## $ Case Number : chr "JD112418" "JD108838" "JD107176" "JC561174" ...  
## $ Date : chr "1/1/2018 0:00" "1/1/2018 0:00" "1/1/2018 0:00" "1/1/2018 0:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : chr "OFFENSE INVOLVING CHILDREN" "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE" "OFFENSE INVOLVING CHILDREN" ...  
## $ Description : chr "SEX ASSLT OF CHILD BY FAM MBR" "FRAUD OR CONFIDENCE GAME" "FINANCIAL IDENTITY THEFT OVER $ 300" "AGG CRIM SEX ABUSE FAM MEMBER" ...  
## $ Location Description: chr "RESIDENCE-GARAGE" "APARTMENT" "RESIDENCE" "RESIDENCE" ...  
## $ Arrest : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Domestic : logi FALSE FALSE FALSE TRUE FALSE FALSE ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : num 24 24 16 19 9 22 12 7 9 16 ...  
## $ Ward : num 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : num 1 2 10 4 61 73 28 67 61 76 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ X Coordinate : num NA NA NA NA NA ...  
## $ Y Coordinate : num NA NA NA NA NA ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Updated On : chr "1/12/2020 15:48" "1/9/2020 15:57" "1/8/2020 15:49" "1/4/2020 15:44" ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...  
## $ Location : chr NA NA NA NA ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ID = col\_double(),  
## .. `Case Number` = col\_character(),  
## .. Date = col\_character(),  
## .. Block = col\_character(),  
## .. IUCR = col\_character(),  
## .. `Primary Type` = col\_character(),  
## .. Description = col\_character(),  
## .. `Location Description` = col\_character(),  
## .. Arrest = col\_logical(),  
## .. Domestic = col\_logical(),  
## .. Beat = col\_double(),  
## .. District = col\_double(),  
## .. Ward = col\_double(),  
## .. `Community Area` = col\_double(),  
## .. `FBI Code` = col\_character(),  
## .. `X Coordinate` = col\_double(),  
## .. `Y Coordinate` = col\_double(),  
## .. Year = col\_double(),  
## .. `Updated On` = col\_character(),  
## .. Latitude = col\_double(),  
## .. Longitude = col\_double(),  
## .. Location = col\_character()  
## .. )

summary(chicago)

## ID Case Number Date Block   
## Min. : 23757 Length:267185 Length:267185 Length:267185   
## 1st Qu.:11286550 Class :character Class :character Class :character   
## Median :11375979 Mode :character Mode :character Mode :character   
## Mean :11352463   
## 3rd Qu.:11465374   
## Max. :11968996   
##   
## IUCR Primary Type Description Location Description  
## Length:267185 Length:267185 Length:267185 Length:267185   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Arrest Domestic Beat District Ward   
## Mode :logical Mode :logical Min. : 111 Min. : 1.0 Min. : 1.00   
## FALSE:213769 FALSE:223427 1st Qu.: 611 1st Qu.: 6.0 1st Qu.:10.00   
## TRUE :53416 TRUE :43758 Median :1031 Median :10.0 Median :24.00   
## Mean :1143 Mean :11.2 Mean :23.45   
## 3rd Qu.:1723 3rd Qu.:17.0 3rd Qu.:35.00   
## Max. :2535 Max. :31.0 Max. :50.00   
## NA's :4   
## Community Area FBI Code X Coordinate Y Coordinate   
## Min. : 0.00 Length:267185 Min. :1092706 Min. :1813897   
## 1st Qu.:23.00 Class :character 1st Qu.:1153330 1st Qu.:1859425   
## Median :32.00 Mode :character Median :1166911 Median :1894254   
## Mean :36.47 Mean :1165006 Mean :1886600   
## 3rd Qu.:53.00 3rd Qu.:1176455 3rd Qu.:1908739   
## Max. :77.00 Max. :1205119 Max. :1951535   
## NA's :2 NA's :4365 NA's :4365   
## Year Updated On Latitude Longitude   
## Min. :2018 Length:267185 Min. :41.65 Min. :-87.93   
## 1st Qu.:2018 Class :character 1st Qu.:41.77 1st Qu.:-87.71   
## Median :2018 Mode :character Median :41.87 Median :-87.66   
## Mean :2018 Mean :41.84 Mean :-87.67   
## 3rd Qu.:2018 3rd Qu.:41.91 3rd Qu.:-87.63   
## Max. :2018 Max. :42.02 Max. :-87.53   
## NA's :4365 NA's :4365   
## Location   
## Length:267185   
## Class :character   
## Mode :character   
##   
##   
##   
##

Data cleaning and Preparations Delete the columns below

chicago = chicago %>% dplyr::select(-ID)   
chicago = chicago %>% dplyr::select(-`Case Number`)  
chicago = chicago %>% dplyr::select(-`Updated On`)  
chicago = chicago %>% dplyr::select(-`X Coordinate`)  
chicago = chicago %>% dplyr::select(-`Y Coordinate`)  
chicago = chicago %>% dplyr::select(-`Location`)  
str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 16 variables:  
## $ Date : chr "1/1/2018 0:00" "1/1/2018 0:00" "1/1/2018 0:00" "1/1/2018 0:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : chr "OFFENSE INVOLVING CHILDREN" "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE" "OFFENSE INVOLVING CHILDREN" ...  
## $ Description : chr "SEX ASSLT OF CHILD BY FAM MBR" "FRAUD OR CONFIDENCE GAME" "FINANCIAL IDENTITY THEFT OVER $ 300" "AGG CRIM SEX ABUSE FAM MEMBER" ...  
## $ Location Description: chr "RESIDENCE-GARAGE" "APARTMENT" "RESIDENCE" "RESIDENCE" ...  
## $ Arrest : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Domestic : logi FALSE FALSE FALSE TRUE FALSE FALSE ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : num 24 24 16 19 9 22 12 7 9 16 ...  
## $ Ward : num 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : num 1 2 10 4 61 73 28 67 61 76 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ID = col\_double(),  
## .. `Case Number` = col\_character(),  
## .. Date = col\_character(),  
## .. Block = col\_character(),  
## .. IUCR = col\_character(),  
## .. `Primary Type` = col\_character(),  
## .. Description = col\_character(),  
## .. `Location Description` = col\_character(),  
## .. Arrest = col\_logical(),  
## .. Domestic = col\_logical(),  
## .. Beat = col\_double(),  
## .. District = col\_double(),  
## .. Ward = col\_double(),  
## .. `Community Area` = col\_double(),  
## .. `FBI Code` = col\_character(),  
## .. `X Coordinate` = col\_double(),  
## .. `Y Coordinate` = col\_double(),  
## .. Year = col\_double(),  
## .. `Updated On` = col\_character(),  
## .. Latitude = col\_double(),  
## .. Longitude = col\_double(),  
## .. Location = col\_character()  
## .. )

Converting the date

chicago = chicago %>% mutate(Date = mdy\_hms(Date))  
str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 16 variables:  
## $ Date : POSIXct, format: "2020-01-01 18:00:00" "2020-01-01 18:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : chr "OFFENSE INVOLVING CHILDREN" "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE" "OFFENSE INVOLVING CHILDREN" ...  
## $ Description : chr "SEX ASSLT OF CHILD BY FAM MBR" "FRAUD OR CONFIDENCE GAME" "FINANCIAL IDENTITY THEFT OVER $ 300" "AGG CRIM SEX ABUSE FAM MEMBER" ...  
## $ Location Description: chr "RESIDENCE-GARAGE" "APARTMENT" "RESIDENCE" "RESIDENCE" ...  
## $ Arrest : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Domestic : logi FALSE FALSE FALSE TRUE FALSE FALSE ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : num 24 24 16 19 9 22 12 7 9 16 ...  
## $ Ward : num 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : num 1 2 10 4 61 73 28 67 61 76 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...

## Preparing the Data

Converting the Month

chicago = chicago %>% mutate(Month = month(Date))  
str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 17 variables:  
## $ Date : POSIXct, format: "2020-01-01 18:00:00" "2020-01-01 18:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : chr "OFFENSE INVOLVING CHILDREN" "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE" "OFFENSE INVOLVING CHILDREN" ...  
## $ Description : chr "SEX ASSLT OF CHILD BY FAM MBR" "FRAUD OR CONFIDENCE GAME" "FINANCIAL IDENTITY THEFT OVER $ 300" "AGG CRIM SEX ABUSE FAM MEMBER" ...  
## $ Location Description: chr "RESIDENCE-GARAGE" "APARTMENT" "RESIDENCE" "RESIDENCE" ...  
## $ Arrest : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Domestic : logi FALSE FALSE FALSE TRUE FALSE FALSE ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : num 24 24 16 19 9 22 12 7 9 16 ...  
## $ Ward : num 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : num 1 2 10 4 61 73 28 67 61 76 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...  
## $ Month : num 1 1 1 1 1 1 1 1 1 1 ...

Converting the hour

chicago = chicago %>% mutate(Hour = hour(Date))  
str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 18 variables:  
## $ Date : POSIXct, format: "2020-01-01 18:00:00" "2020-01-01 18:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : chr "OFFENSE INVOLVING CHILDREN" "DECEPTIVE PRACTICE" "DECEPTIVE PRACTICE" "OFFENSE INVOLVING CHILDREN" ...  
## $ Description : chr "SEX ASSLT OF CHILD BY FAM MBR" "FRAUD OR CONFIDENCE GAME" "FINANCIAL IDENTITY THEFT OVER $ 300" "AGG CRIM SEX ABUSE FAM MEMBER" ...  
## $ Location Description: chr "RESIDENCE-GARAGE" "APARTMENT" "RESIDENCE" "RESIDENCE" ...  
## $ Arrest : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Domestic : logi FALSE FALSE FALSE TRUE FALSE FALSE ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : num 24 24 16 19 9 22 12 7 9 16 ...  
## $ Ward : num 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : num 1 2 10 4 61 73 28 67 61 76 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...  
## $ Month : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Hour : int 18 18 18 18 18 18 18 18 18 18 ...

Factor conversion and recoding

chicago = chicago %>% mutate(Year = as\_factor(as.character(Year))) %>%  
mutate(Year = fct\_recode(Year))  
  
chicago = chicago %>% mutate(Month = as\_factor(as.character(Month))) %>%  
mutate(Month = fct\_recode(Month,  
 "Jan" = "1",  
"Feb" = "2",  
"March" = "3",  
"April" = "4",  
"May" = "5",  
"June" = "6",  
"July" = "7",  
"Aug" = "8",  
"Sept" = "9",  
"Oct" = "10",  
"Nov" = "11",  
"Dec" = "12"))  
  
chicago = chicago %>% mutate(Hour = as\_factor(as.character(Hour))) %>%  
mutate(Hour = fct\_recode(Hour))  
  
chicago = chicago %>% mutate(Arrest = as\_factor(as.character(Arrest))) %>%  
mutate(Arrest = fct\_recode(Arrest,  
"Yes" = "TRUE",  
"No" = "FALSE"))  
  
chicago = chicago %>% mutate(Domestic = as\_factor(as.character(Domestic))) %>%  
mutate(Domestic = fct\_recode(Domestic,  
"Domestic\_violence" = "TRUE",  
"Non\_Domestic\_violence" = "FALSE"))  
  
chicago = chicago %>% mutate(chicago = as\_factor(as.character(`Primary Type`))) %>%  
mutate(`Primary Type` = fct\_recode(`Primary Type`))  
  
chicago = chicago %>% mutate(Description = as\_factor(as.character(Description))) %>%  
mutate(Description = fct\_recode(Description))  
  
chicago = chicago %>% mutate(`Location Description` = as\_factor(as.character(`Location Description`))) %>%  
mutate(`Location Description` = fct\_recode(`Location Description`))  
  
chicago = chicago %>% mutate(Domestic = as\_factor(as.character(Domestic))) %>%  
mutate(Domestic = fct\_recode(Domestic))  
  
chicago = chicago %>% mutate(District = as\_factor(as.character(District))) %>%  
mutate(District = fct\_recode(District))  
  
chicago = chicago %>% mutate(Ward = as\_factor(as.character(Ward))) %>%  
mutate(Ward = fct\_recode(Ward))  
  
chicago = chicago %>% mutate(`Community Area` = as\_factor(as.character(`Community Area`))) %>%  
mutate(`Community Area` = fct\_recode(`Community Area`))  
str(chicago)

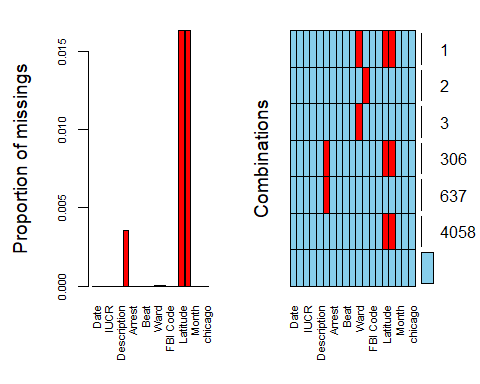
## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 19 variables:  
## $ Date : POSIXct, format: "2020-01-01 18:00:00" "2020-01-01 18:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : Factor w/ 32 levels "ARSON","ASSAULT",..: 22 9 9 22 22 6 6 22 22 24 ...  
## $ Description : Factor w/ 301 levels "SEX ASSLT OF CHILD BY FAM MBR",..: 1 2 3 4 4 5 6 4 4 7 ...  
## $ Location Description: Factor w/ 132 levels "RESIDENCE-GARAGE",..: 1 2 3 3 3 4 4 3 2 3 ...  
## $ Arrest : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...  
## $ Domestic : Factor w/ 2 levels "Non\_Domestic\_violence",..: 1 1 1 2 1 1 1 1 1 1 ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : Factor w/ 23 levels "24","16","19",..: 1 1 2 3 4 5 6 7 4 2 ...  
## $ Ward : Factor w/ 50 levels "49","50","41",..: 1 2 3 4 5 6 7 8 9 3 ...  
## $ Community Area : Factor w/ 78 levels "1","2","10","4",..: 1 2 3 4 5 6 7 8 5 9 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ Year : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...  
## $ Month : Factor w/ 12 levels "Jan","Feb","March",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Hour : Factor w/ 1 level "18": 1 1 1 1 1 1 1 1 1 1 ...  
## $ chicago : Factor w/ 32 levels "OFFENSE INVOLVING CHILDREN",..: 1 2 2 1 1 3 3 1 1 4 ...

summary(chicago)

## Date Block IUCR   
## Min. :2020-01-01 18:00:00 Length:267185 Length:267185   
## 1st Qu.:2020-04-12 18:16:00 Class :character Class :character   
## Median :2020-07-06 18:13:15 Mode :character Mode :character   
## Mean :2020-07-05 08:19:16   
## 3rd Qu.:2020-09-28 18:09:00   
## Max. :2020-12-31 18:00:00   
##   
## Primary Type Description   
## THEFT :65088 SIMPLE : 29660   
## BATTERY :49704 $500 AND UNDER : 24772   
## CRIMINAL DAMAGE :27727 DOMESTIC BATTERY SIMPLE: 24221   
## ASSAULT :20358 OVER $500 : 15251   
## DECEPTIVE PRACTICE:19300 TO VEHICLE : 13949   
## OTHER OFFENSE :17205 TO PROPERTY : 13098   
## (Other) :67803 (Other) :146234   
## Location Description Arrest Domestic   
## STREET :58900 No :213769 Non\_Domestic\_violence:223427   
## RESIDENCE:44814 Yes: 53416 Domestic\_violence : 43758   
## APARTMENT:34559   
## SIDEWALK :21098   
## OTHER :10896   
## (Other) :95975   
## NA's : 943   
## Beat District Ward Community Area   
## Min. : 111 11 : 19146 42 : 18107 25 : 15105   
## 1st Qu.: 611 6 : 16455 24 : 12616 8 : 13061   
## Median :1031 8 : 16337 28 : 11901 32 : 10860   
## Mean :1143 18 : 16172 27 : 11212 28 : 9424   
## 3rd Qu.:1723 1 : 15639 2 : 10072 29 : 9395   
## Max. :2535 7 : 14266 (Other):203273 (Other):209338   
## (Other):169170 NA's : 4 NA's : 2   
## FBI Code Year Latitude Longitude   
## Length:267185 2018:267185 Min. :41.65 Min. :-87.93   
## Class :character 1st Qu.:41.77 1st Qu.:-87.71   
## Mode :character Median :41.87 Median :-87.66   
## Mean :41.84 Mean :-87.67   
## 3rd Qu.:41.91 3rd Qu.:-87.63   
## Max. :42.02 Max. :-87.53   
## NA's :4365 NA's :4365   
## Month Hour chicago   
## Aug : 25356 18:267185 THEFT :65088   
## July : 25201 BATTERY :49704   
## May : 24682 CRIMINAL DAMAGE :27727   
## June : 24184 ASSAULT :20358   
## Sept : 23033 DECEPTIVE PRACTICE:19300   
## Oct : 22789 OTHER OFFENSE :17205   
## (Other):121940 (Other) :67803

Viewing missing valuse

vim\_plot = aggr(chicago, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

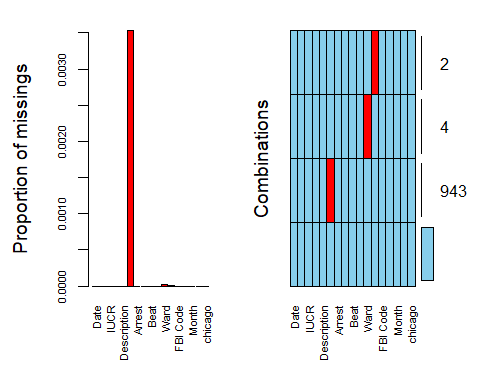
 Looking at the graph, we need to remov the latitude and longtitude variables as we missing many vaules.

chicago = chicago %>% dplyr::select(-Latitude)  
chicago = chicago %>% dplyr::select(-Longitude)  
str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 17 variables:  
## $ Date : POSIXct, format: "2020-01-01 18:00:00" "2020-01-01 18:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : chr "1753" "1130" "1153" "1752" ...  
## $ Primary Type : Factor w/ 32 levels "ARSON","ASSAULT",..: 22 9 9 22 22 6 6 22 22 24 ...  
## $ Description : Factor w/ 301 levels "SEX ASSLT OF CHILD BY FAM MBR",..: 1 2 3 4 4 5 6 4 4 7 ...  
## $ Location Description: Factor w/ 132 levels "RESIDENCE-GARAGE",..: 1 2 3 3 3 4 4 3 2 3 ...  
## $ Arrest : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...  
## $ Domestic : Factor w/ 2 levels "Non\_Domestic\_violence",..: 1 1 1 2 1 1 1 1 1 1 ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : Factor w/ 23 levels "24","16","19",..: 1 1 2 3 4 5 6 7 4 2 ...  
## $ Ward : Factor w/ 50 levels "49","50","41",..: 1 2 3 4 5 6 7 8 9 3 ...  
## $ Community Area : Factor w/ 78 levels "1","2","10","4",..: 1 2 3 4 5 6 7 8 5 9 ...  
## $ FBI Code : chr "2" "11" "11" "17" ...  
## $ Year : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Month : Factor w/ 12 levels "Jan","Feb","March",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Hour : Factor w/ 1 level "18": 1 1 1 1 1 1 1 1 1 1 ...  
## $ chicago : Factor w/ 32 levels "OFFENSE INVOLVING CHILDREN",..: 1 2 2 1 1 3 3 1 1 4 ...

Viewing additional values

vim\_plot = aggr(chicago, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Deleting rows of missingnes

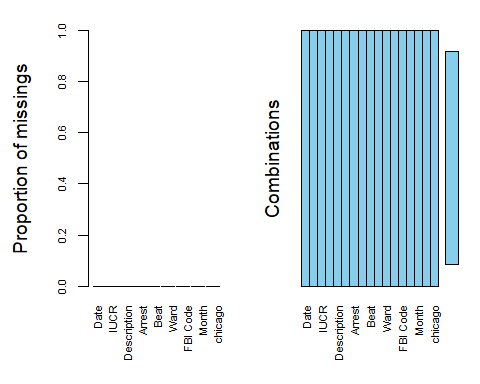
chicago = chicago %>% drop\_na()  
summary(chicago)

## Date Block IUCR   
## Min. :2020-01-01 18:00:00 Length:266236 Length:266236   
## 1st Qu.:2020-04-12 18:16:26 Class :character Class :character   
## Median :2020-07-06 18:13:42 Mode :character Mode :character   
## Mean :2020-07-05 08:22:32   
## 3rd Qu.:2020-09-28 18:08:01   
## Max. :2020-12-31 18:00:00   
##   
## Primary Type Description   
## THEFT :65088 SIMPLE : 29658   
## BATTERY :49703 $500 AND UNDER : 24772   
## CRIMINAL DAMAGE :27727 DOMESTIC BATTERY SIMPLE: 24221   
## ASSAULT :20357 OVER $500 : 15251   
## DECEPTIVE PRACTICE:18357 TO VEHICLE : 13949   
## OTHER OFFENSE :17204 TO PROPERTY : 13098   
## (Other) :67800 (Other) :145287   
## Location Description Arrest   
## STREET :58899 No :212821   
## RESIDENCE :44813 Yes: 53415   
## APARTMENT :34559   
## SIDEWALK :21098   
## OTHER :10896   
## PARKING LOT/GARAGE(NON.RESID.): 7644   
## (Other) :88327   
## Domestic Beat District   
## Non\_Domestic\_violence:222479 Min. : 111 11 : 19122   
## Domestic\_violence : 43757 1st Qu.: 611 6 : 16412   
## Median :1031 8 : 16316   
## Mean :1142 18 : 16099   
## 3rd Qu.:1723 1 : 15566   
## Max. :2535 7 : 14251   
## (Other):168470   
## Ward Community Area FBI Code Year   
## 42 : 18039 25 : 15068 Length:266236 2018:266236   
## 24 : 12599 8 : 13001 Class :character   
## 28 : 11881 32 : 10830 Mode :character   
## 27 : 11177 29 : 9382   
## 2 : 10020 28 : 9375   
## 17 : 8845 43 : 8625   
## (Other):193675 (Other):199955   
## Month Hour chicago   
## Aug : 25296 18:266236 THEFT :65088   
## July : 25118 BATTERY :49703   
## May : 24600 CRIMINAL DAMAGE :27727   
## June : 24107 ASSAULT :20357   
## Sept : 22953 DECEPTIVE PRACTICE:18357   
## Oct : 22694 OTHER OFFENSE :17204   
## (Other):121468 (Other) :67800

viewing the data

vim\_plot = aggr(chicago, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)

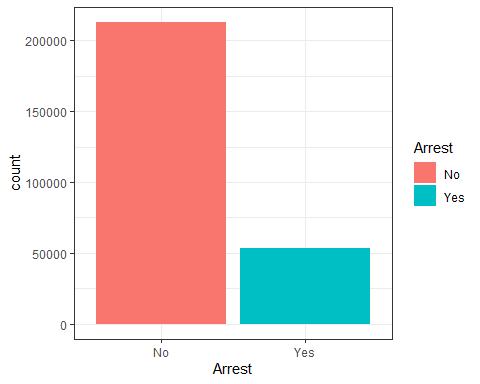
## Warning in plot.aggr(res, ...): not enough horizontal space to display  
## frequencies



### Data Exploration

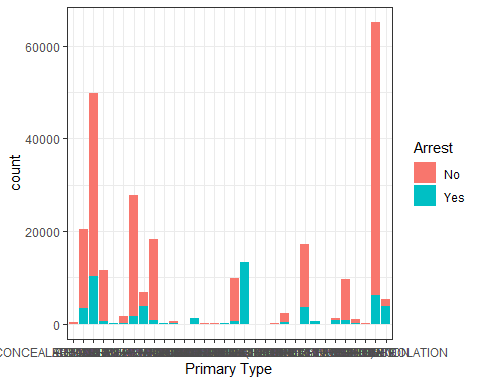
Begin exploring the data by looking at the count of our response variable only

ggplot(chicago, aes(x=Arrest, fill = Arrest)) + geom\_bar() + theme\_bw()



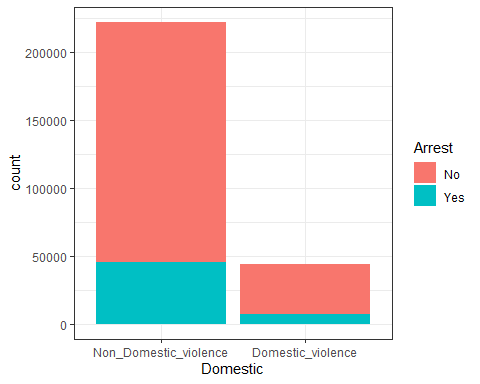
`Data visualization

ggplot(chicago, aes(x=`Primary Type`, fill = Arrest)) + geom\_bar() + theme\_bw()

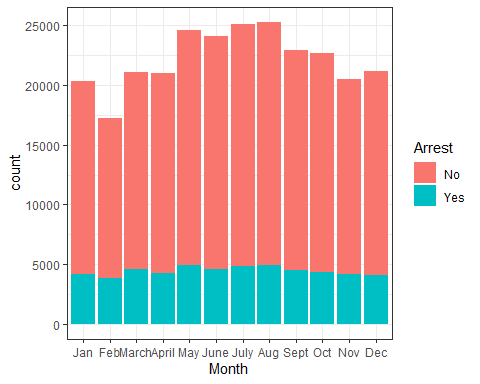


There’s seems to be relationship between Primary Type and Arrest.

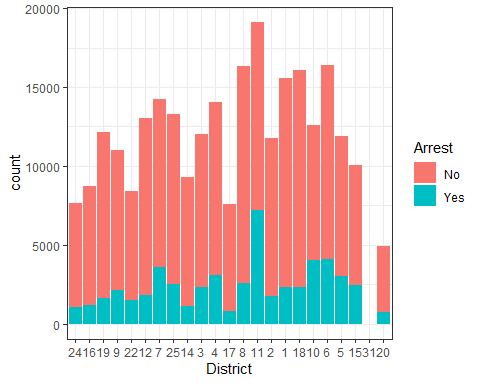
ggplot(chicago, aes(x=Domestic, fill = Arrest)) + geom\_bar() + theme\_bw()

 There’s no relationship between Domestic and Arrest as issues without domestic violence leads to more arrest

ggplot(chicago, aes(x=Month, fill = Arrest)) + geom\_bar() + theme\_bw()

 From the graph there’s a strong relationship between Month and Arrest. In the summer there’s seem to be more arrest than the other months.

ggplot(chicago, aes(x=District, fill = Arrest)) + geom\_bar() + theme\_bw()



## Building Predictive model

Selecting important variables

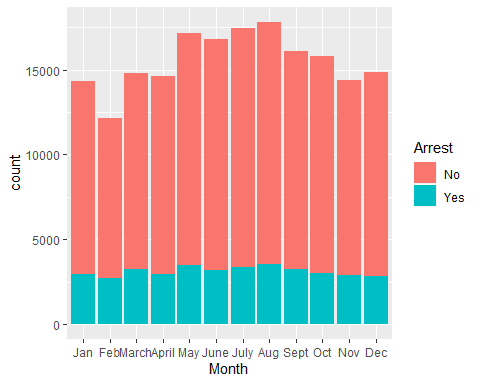
chicago2 = chicago %>% dplyr::select("Arrest", "Description", "Domestic", "Ward", "Month", "Year", "Location Description", "District")

Splitting the data

set.seed(123)  
train.rows = createDataPartition(y = chicago$Arrest, p=0.7, list = FALSE)  
train = chicago2[train.rows,]   
test = chicago2[-train.rows,]

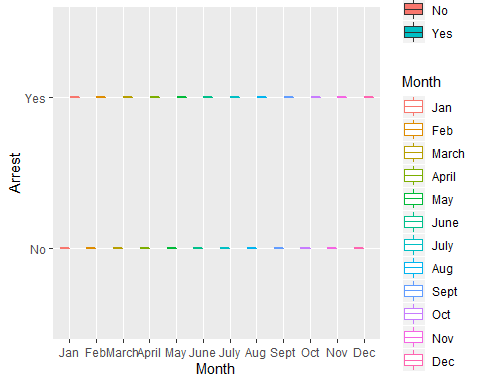
Visualize using the training set (looking at relationship between SeriousDlqin2yrs and the other variables).

ggplot(train,aes(x=Month, fill = Arrest)) + geom\_bar()



Finding more relationhips

ggplot(train, aes(Month,Arrest))+  
 geom\_boxplot(aes(color = Month, fill=Arrest))



Let’s build a model with Month

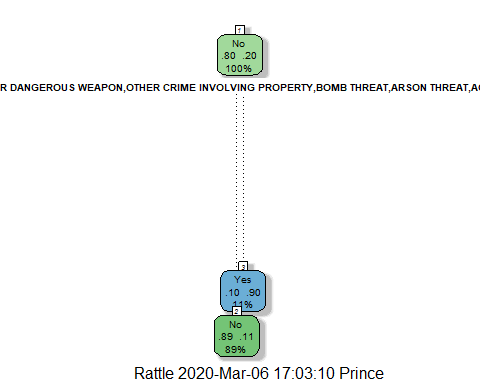
mod1 = glm(Arrest ~ Month , train, family = "binomial")  
summary(mod1)

##   
## Call:  
## glm(formula = Arrest ~ Month, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.7084 -0.6726 -0.6650 -0.6462 1.8271   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.34264 0.02062 -65.123 < 2e-16 \*\*\*  
## MonthFeb 0.08812 0.03004 2.934 0.003350 \*\*   
## MonthMarch 0.07706 0.02862 2.693 0.007085 \*\*   
## MonthApril -0.02844 0.02913 -0.976 0.328924   
## MonthMay -0.02420 0.02801 -0.864 0.387587   
## MonthJune -0.11773 0.02853 -4.126 3.69e-05 \*\*\*  
## MonthJuly -0.10247 0.02821 -3.633 0.000280 \*\*\*  
## MonthAug -0.05379 0.02789 -1.929 0.053783 .   
## MonthSept -0.04498 0.02852 -1.577 0.114725   
## MonthOct -0.10298 0.02889 -3.564 0.000365 \*\*\*  
## MonthNov -0.03308 0.02928 -1.130 0.258581   
## MonthDec -0.09648 0.02930 -3.292 0.000993 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 186842 on 186365 degrees of freedom  
## Residual deviance: 186725 on 186354 degrees of freedom  
## AIC: 186749  
##   
## Number of Fisher Scoring iterations: 4

Note the AIC of this model (a measure of model quality) is 186,749. We can use this value to compare this model to others. Smaller AIC is better.

##Building Classification For Further Analysis, let’s build Classification tree

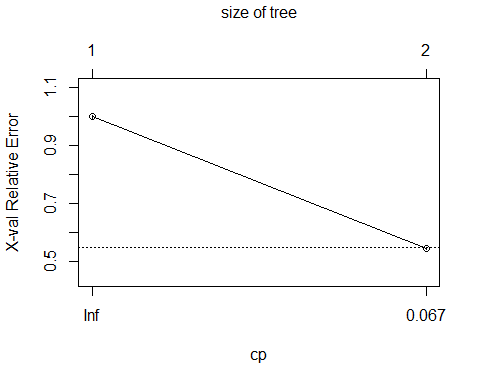
tree1 = rpart(Arrest ~., train, method="class")  
fancyRpartPlot(tree1)



printcp(tree1)

##   
## Classification tree:  
## rpart(formula = Arrest ~ ., data = train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Description  
##   
## Root node error: 37391/186366 = 0.20063  
##   
## n= 186366   
##   
## CP nsplit rel error xerror xstd  
## 1 0.45369 0 1.00000 1.00000 0.0046237  
## 2 0.01000 1 0.54631 0.54708 0.0036091

plotcp(tree1)



Prediction on the training set

treepred = predict(tree1, train, type = "class")  
head(treepred)

## 1 2 3 4 5 6   
## No No No No No No   
## Levels: No Yes

A look at the Accuracy on the training set

confusionMatrix(treepred,train$Arrest,positive="Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 146863 18315  
## Yes 2112 19076  
##   
## Accuracy : 0.8904   
## 95% CI : (0.889, 0.8918)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5921   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.5102   
## Specificity : 0.9858   
## Pos Pred Value : 0.9003   
## Neg Pred Value : 0.8891   
## Prevalence : 0.2006   
## Detection Rate : 0.1024   
## Detection Prevalence : 0.1137   
## Balanced Accuracy : 0.7480   
##   
## 'Positive' Class : Yes   
##

Prediction on the test set

treepred\_test = predict(tree1, test, type = "class")  
head(treepred\_test)

## 1 2 3 4 5 6   
## No No No No No No   
## Levels: No Yes

Confusion Matrix on the test set

confusionMatrix(treepred\_test,test$Arrest,positive="Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 62943 8023  
## Yes 903 8001  
##   
## Accuracy : 0.8882   
## 95% CI : (0.886, 0.8904)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.582   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.4993   
## Specificity : 0.9859   
## Pos Pred Value : 0.8986   
## Neg Pred Value : 0.8869   
## Prevalence : 0.2006   
## Detection Rate : 0.1002   
## Detection Prevalence : 0.1115   
## Balanced Accuracy : 0.7426   
##   
## 'Positive' Class : Yes   
##

###Building Random Forest Model with Caret Building a Random Forest

fit\_control = trainControl(method = "cv",   
 number = 5)   
  
  
set.seed(123)   
rf\_fit = train(Arrest~Month,train,  
 method = "ranger",   
 importance = "permutation",  
 trControl = fit\_control,  
 num.trees = 10)

Checking the Random Forest

varImp(rf\_fit)

## ranger variable importance

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf  
  
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

## Overall  
## MonthApril NaN  
## MonthOct NaN  
## MonthNov NaN  
## MonthJuly NaN  
## MonthDec NaN  
## MonthAug NaN  
## MonthFeb NaN  
## MonthMarch NaN  
## MonthSept NaN  
## MonthJune NaN  
## MonthMay NaN

rf\_fit

## Random Forest   
##   
## 186366 samples  
## 1 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 149093, 149093, 149092, 149093, 149093   
## Resampling results across tuning parameters:  
##   
## mtry splitrule Accuracy Kappa  
## 2 gini 0.7993679 0   
## 2 extratrees 0.7993679 0   
## 6 gini 0.7993679 0   
## 6 extratrees 0.7993679 0   
## 11 gini 0.7993679 0   
## 11 extratrees 0.7993679 0   
##   
## Tuning parameter 'min.node.size' was held constant at a value of 1  
## Accuracy was used to select the optimal model using the largest value.  
## The final values used for the model were mtry = 2, splitrule = gini  
## and min.node.size = 1.

Prediction of Random Forest

predRF = predict.train(rf\_fit, train)  
head(predRF)

## [1] No No No No No No  
## Levels: No Yes

Accuracy for Random Forest

confusionMatrix(predRF, train$Arrest, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 148975 37391  
## Yes 0 0  
##   
## Accuracy : 0.7994   
## 95% CI : (0.7975, 0.8012)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : 0.5014   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.7994   
## Prevalence : 0.2006   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : Yes   
##

### Building Neural Network

start\_time = Sys.time()   
fitControl = trainControl(method = "cv",   
 number = 5)  
  
nnetGrid <- expand.grid(size = 6, decay = 0.1)  
  
set.seed(1234)  
nnetBasic = train(Arrest~Month,train,  
 method = "nnet",  
 tuneGrid = nnetGrid,  
 trControl = fitControl,  
 verbose = FALSE,  
 trace = FALSE)  
  
end\_time = Sys.time()  
end\_time-start\_time

## Time difference of 3.081958 mins

nnetBasic

## Neural Network   
##   
## 186366 samples  
## 1 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 149093, 149093, 149093, 149092, 149093   
## Resampling results:  
##   
## Accuracy Kappa  
## 0.7993679 0   
##   
## Tuning parameter 'size' was held constant at a value of 6  
## Tuning  
## parameter 'decay' was held constant at a value of 0.1

Prediction and Confustion matrix Neural Network

predNetBasic = predict(nnetBasic, train)  
  
confusionMatrix(predNetBasic, train$Arrest, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 148975 37391  
## Yes 0 0  
##   
## Accuracy : 0.7994   
## 95% CI : (0.7975, 0.8012)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : 0.5014   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.7994   
## Prevalence : 0.2006   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : Yes   
##

Making Prediction on the test data

predNetBasic = predict(nnetBasic, test)  
  
confusionMatrix(predNetBasic, test$Arrest, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 63846 16024  
## Yes 0 0  
##   
## Accuracy : 0.7994   
## 95% CI : (0.7966, 0.8021)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : 0.5021   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.0000   
## Specificity : 1.0000   
## Pos Pred Value : NaN   
## Neg Pred Value : 0.7994   
## Prevalence : 0.2006   
## Detection Rate : 0.0000   
## Detection Prevalence : 0.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : Yes   
##