## Course Project Phase 1 and 2

## Bridgett Davis

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(lubridate)

##   
## Attaching package: 'lubridate'

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

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:lubridate':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday,  
## week, yday, year

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

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

## 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(MASS)

##   
## Attaching package: 'MASS'

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

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

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(caret)   
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

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.

library(rpart.plot)

Read in file

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.

chicago

## # A tibble: 267,185 x 22  
## ID `Case Number` Date Block IUCR `Primary Type` Description  
## <dbl> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 1.19e7 JD112418 1/1/~ 069X~ 1753 OFFENSE INVOL~ SEX ASSLT ~  
## 2 1.19e7 JD108838 1/1/~ 070X~ 1130 DECEPTIVE PRA~ FRAUD OR C~  
## 3 1.19e7 JD107176 1/1/~ 072X~ 1153 DECEPTIVE PRA~ FINANCIAL ~  
## 4 1.19e7 JC561174 1/1/~ 047X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 5 1.19e7 JC547965 1/1/~ 051X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 6 1.19e7 JC532311 1/1/~ 013X~ 265 CRIM SEXUAL A~ AGGRAVATED~  
## 7 1.12e7 JB101037 1/1/~ 017X~ 281 CRIM SEXUAL A~ NON-AGGRAV~  
## 8 1.12e7 JB147547 1/1/~ 056X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 9 1.19e7 JC518051 1/1/~ 009X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 10 1.19e7 JC513241 1/1/~ 052X~ 2826 OTHER OFFENSE HARASSMENT~  
## # ... with 267,175 more rows, and 15 more variables: `Location  
## # Description` <chr>, Arrest <lgl>, Domestic <lgl>, Beat <dbl>,  
## # District <dbl>, Ward <dbl>, `Community Area` <dbl>, `FBI Code` <chr>,  
## # `X Coordinate` <dbl>, `Y Coordinate` <dbl>, Year <dbl>, `Updated  
## # On` <chr>, Latitude <dbl>, Longitude <dbl>, Location <chr>

## Course Project Phase 1

## Data Cleansing

Delete Columns ID, Case Number, Updated On, X Coordinate, Y Coordinate, Location

chicago = chicago %>%   
 dplyr::select(-1, -2, -16, -17, -19, -22)  
chicago

## # A tibble: 267,185 x 16  
## Date Block IUCR `Primary Type` Description `Location Descr~ Arrest  
## <chr> <chr> <chr> <chr> <chr> <chr> <lgl>   
## 1 1/1/~ 069X~ 1753 OFFENSE INVOL~ SEX ASSLT ~ RESIDENCE-GARAGE FALSE   
## 2 1/1/~ 070X~ 1130 DECEPTIVE PRA~ FRAUD OR C~ APARTMENT FALSE   
## 3 1/1/~ 072X~ 1153 DECEPTIVE PRA~ FINANCIAL ~ RESIDENCE FALSE   
## 4 1/1/~ 047X~ 1752 OFFENSE INVOL~ AGG CRIM S~ RESIDENCE FALSE   
## 5 1/1/~ 051X~ 1752 OFFENSE INVOL~ AGG CRIM S~ RESIDENCE FALSE   
## 6 1/1/~ 013X~ 265 CRIM SEXUAL A~ AGGRAVATED~ OTHER FALSE   
## 7 1/1/~ 017X~ 281 CRIM SEXUAL A~ NON-AGGRAV~ OTHER FALSE   
## 8 1/1/~ 056X~ 1752 OFFENSE INVOL~ AGG CRIM S~ RESIDENCE TRUE   
## 9 1/1/~ 009X~ 1752 OFFENSE INVOL~ AGG CRIM S~ APARTMENT FALSE   
## 10 1/1/~ 052X~ 2826 OTHER OFFENSE HARASSMENT~ RESIDENCE FALSE   
## # ... with 267,175 more rows, and 9 more variables: Domestic <lgl>,  
## # Beat <dbl>, District <dbl>, Ward <dbl>, `Community Area` <dbl>, `FBI  
## # Code` <chr>, Year <dbl>, Latitude <dbl>, Longitude <dbl>

Convert Date Object

chicago = chicago %>% mutate(Date = mdy\_hm(Date))   
chicago = chicago %>% mutate(Hour = hour(Date))  
summary(chicago$Hour)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 9.00 14.00 13.08 18.00 23.00

chicago

## # A tibble: 267,185 x 17  
## Date Block IUCR `Primary Type` Description  
## <dttm> <chr> <chr> <chr> <chr>   
## 1 2018-01-01 00:00:00 069X~ 1753 OFFENSE INVOL~ SEX ASSLT ~  
## 2 2018-01-01 00:00:00 070X~ 1130 DECEPTIVE PRA~ FRAUD OR C~  
## 3 2018-01-01 00:00:00 072X~ 1153 DECEPTIVE PRA~ FINANCIAL ~  
## 4 2018-01-01 00:00:00 047X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 5 2018-01-01 00:00:00 051X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 6 2018-01-01 00:00:00 013X~ 265 CRIM SEXUAL A~ AGGRAVATED~  
## 7 2018-01-01 00:00:00 017X~ 281 CRIM SEXUAL A~ NON-AGGRAV~  
## 8 2018-01-01 00:00:00 056X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 9 2018-01-01 00:00:00 009X~ 1752 OFFENSE INVOL~ AGG CRIM S~  
## 10 2018-01-01 00:00:00 052X~ 2826 OTHER OFFENSE HARASSMENT~  
## # ... with 267,175 more rows, and 12 more variables: `Location  
## # Description` <chr>, Arrest <lgl>, Domestic <lgl>, Beat <dbl>,  
## # District <dbl>, Ward <dbl>, `Community Area` <dbl>, `FBI Code` <chr>,  
## # Year <dbl>, Latitude <dbl>, Longitude <dbl>, Hour <int>

Consider which variables should be factors (categorical) and then convert them

chicago1 = chicago %>%   
 mutate(Hour=as.factor(Hour)) %>%  
 mutate(IUCR=as.factor(IUCR)) %>%  
 mutate(`Primary Type` = as\_factor(as.character(`Primary Type`))) %>%  
 mutate(Description=as.factor(Description)) %>%  
 mutate(Arrest=as.factor(Arrest)) %>%  
 mutate(Arrest = fct\_recode(Arrest, "No" = "FALSE", "Yes" = "TRUE" )) %>%  
 mutate(Domestic=as\_factor(as.character(Domestic))) %>%  
 mutate(Arrest = fct\_recode(Arrest, "No" = "FALSE", "Yes" = "TRUE" )) %>%  
 mutate(District=as.factor(District)) %>%  
 mutate(Ward=as.factor(Ward)) %>%  
 mutate(`FBI Code` = as\_factor(as.character(`FBI Code`))) %>%  
 mutate(`Location Description` = as\_factor(as.character(`Location Description`))) %>%  
 mutate(`Community Area` = as\_factor(as.character(`Community Area`)))

## Warning: Unknown levels in `f`: FALSE, TRUE

str(chicago1)

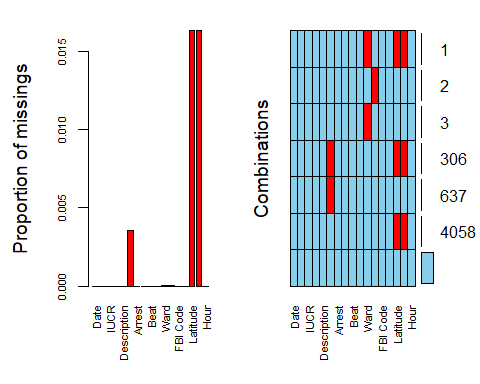
## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 17 variables:  
## $ Date : POSIXct, format: "2018-01-01 00:00:00" "2018-01-01 00:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : Factor w/ 322 levels "031A","031B",..: 113 21 27 112 112 169 175 112 112 178 ...  
## $ Primary Type : Factor w/ 32 levels "OFFENSE INVOLVING CHILDREN",..: 1 2 2 1 1 3 3 1 1 4 ...  
## $ Description : Factor w/ 301 levels "$500 AND UNDER",..: 241 127 119 3 3 33 178 3 3 135 ...  
## $ 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 "FALSE","TRUE": 1 1 1 2 1 1 1 1 1 1 ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : Factor w/ 23 levels "1","2","3","4",..: 21 21 15 18 9 20 12 7 9 15 ...  
## $ Ward : Factor w/ 50 levels "1","2","3","4",..: 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : Factor w/ 78 levels "1","2","10","4",..: 1 2 3 4 5 6 7 8 5 9 ...  
## $ FBI Code : Factor w/ 26 levels "2","11","17",..: 1 2 2 3 3 1 1 3 3 4 ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...  
## $ Hour : Factor w/ 24 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(chicago1)

## Date Block IUCR   
## Min. :2018-01-01 00:00:00 Length:267185 820 : 24772   
## 1st Qu.:2018-04-12 16:00:00 Class :character 486 : 24221   
## Median :2018-07-06 13:15:00 Mode :character 460 : 16079   
## Mean :2018-07-05 06:54:08 810 : 15251   
## 3rd Qu.:2018-09-28 09:00:00 560 : 13428   
## Max. :2018-12-31 00:00:00 1310 : 13098   
## (Other):160336   
## 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 Beat   
## STREET :58900 No :213769 FALSE:223427 Min. : 111   
## RESIDENCE:44814 Yes: 53416 TRUE : 43758 1st Qu.: 611   
## APARTMENT:34559 Median :1031   
## SIDEWALK :21098 Mean :1143   
## OTHER :10896 3rd Qu.:1723   
## (Other) :95975 Max. :2535   
## NA's : 943   
## District Ward Community Area FBI Code   
## 11 : 19146 42 : 18107 25 : 15105 6 :65088   
## 6 : 16455 24 : 12616 8 : 13061 08B :42047   
## 8 : 16337 28 : 11901 32 : 10860 14 :27727   
## 18 : 16172 27 : 11212 28 : 9424 26 :24758   
## 1 : 15639 2 : 10072 29 : 9395 11 :17639   
## 7 : 14266 (Other):203273 (Other):209338 08A :14622   
## (Other):169170 NA's : 4 NA's : 2 (Other):75304   
## Year Latitude Longitude Hour   
## Min. :2018 Min. :41.65 Min. :-87.93 12 : 16303   
## 1st Qu.:2018 1st Qu.:41.77 1st Qu.:-87.71 18 : 15205   
## Median :2018 Median :41.87 Median :-87.66 19 : 15193   
## Mean :2018 Mean :41.84 Mean :-87.67 15 : 14959   
## 3rd Qu.:2018 3rd Qu.:41.91 3rd Qu.:-87.63 17 : 14788   
## Max. :2018 Max. :42.02 Max. :-87.53 16 : 14337   
## NA's :4365 NA's :4365 (Other):176400

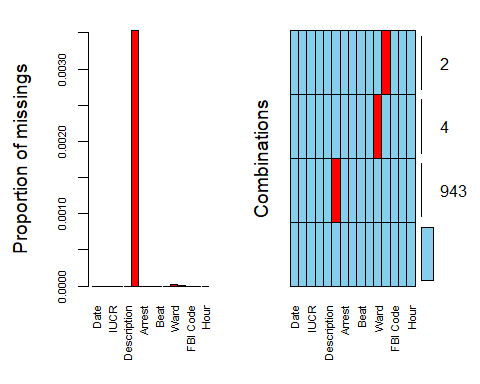
Missing Values

vim\_plot = aggr(chicago1, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Latitude and Logitude missing most of data - using columnwise to remove

chicago1 = chicago1 %>% dplyr::select(-Latitude, -Longitude)   
vim\_plot = aggr(chicago1, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Using row-rise deltion for remaining missingness since data is mostly whole

chicago1 = chicago1 %>% drop\_na()   
summary(chicago1)

## Date Block IUCR   
## Min. :2018-01-01 00:00:00 Length:266236 820 : 24772   
## 1st Qu.:2018-04-12 16:26:30 Class :character 486 : 24221   
## Median :2018-07-06 13:42:30 Mode :character 460 : 16078   
## Mean :2018-07-05 06:57:39 810 : 15251   
## 3rd Qu.:2018-09-28 08:01:00 560 : 13427   
## Max. :2018-12-31 00:00:00 1310 : 13098   
## (Other):159389   
## 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 Domestic   
## STREET :58899 No :212821 FALSE:222479   
## RESIDENCE :44813 Yes: 53415 TRUE : 43757   
## APARTMENT :34559   
## SIDEWALK :21098   
## OTHER :10896   
## PARKING LOT/GARAGE(NON.RESID.): 7644   
## (Other) :88327   
## Beat District Ward Community Area   
## Min. : 111 11 : 19122 42 : 18039 25 : 15068   
## 1st Qu.: 611 6 : 16412 24 : 12599 8 : 13001   
## Median :1031 8 : 16316 28 : 11881 32 : 10830   
## Mean :1142 18 : 16099 27 : 11177 29 : 9382   
## 3rd Qu.:1723 1 : 15566 2 : 10020 28 : 9375   
## Max. :2535 7 : 14251 17 : 8845 43 : 8625   
## (Other):168470 (Other):193675 (Other):199955   
## FBI Code Year Hour   
## 6 :65088 Min. :2018 12 : 16210   
## 08B :42046 1st Qu.:2018 18 : 15168   
## 14 :27727 Median :2018 19 : 15163   
## 26 :24757 Mean :2018 15 : 14900   
## 11 :16697 3rd Qu.:2018 17 : 14755   
## 08A :14621 Max. :2018 16 : 14279   
## (Other):75300 (Other):175761

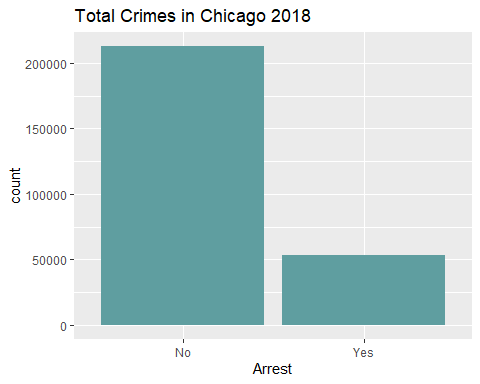
Create appropriate visualizations (charts, tables, etc.) to examine the relationship between variables and the “Arrest” variable.

Total Number of Arrests Yes/No

t0 = table(chicago1$Arrest)  
summary(chicago$Arrest)

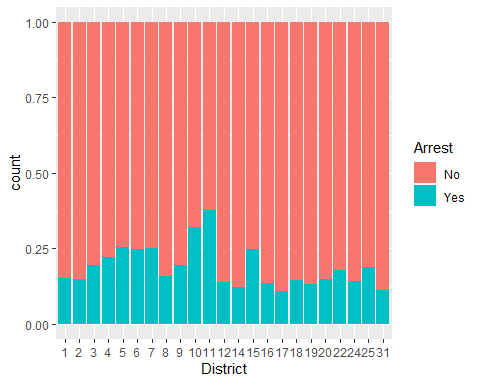
## Mode FALSE TRUE   
## logical 213769 53416

ggplot(chicago1, aes(x=Arrest)) + geom\_bar(fill="cadetblue")+  
 ggtitle("Total Crimes in Chicago 2018")



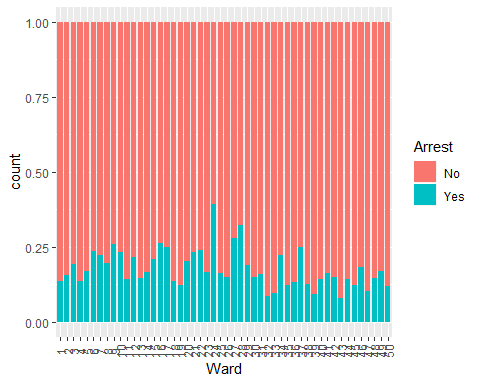
District

ggplot(chicago1, aes(x=District, fill=Arrest)) + geom\_bar(position="fill")



Ward: City Council District where crime occurred

ggplot(chicago1, aes(x=Ward, fill=Arrest)) +   
 geom\_bar(position="fill",width=.80) +  
 theme(axis.text.x=element\_text(angle=90,hjust=.5,vjust=0.5))

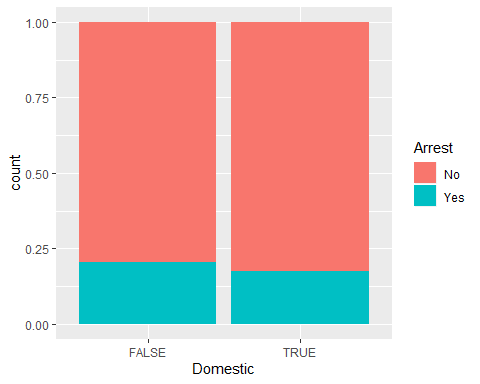


t1 = table(chicago1$Ward,chicago1$Arrest)  
prop.table(t1,margin = 2)

##   
## No Yes  
## 1 0.019467064 0.012131424  
## 2 0.039723524 0.029317607  
## 3 0.026303795 0.024824488  
## 4 0.022239347 0.014097164  
## 5 0.024494763 0.020050548  
## 6 0.031698000 0.038753159  
## 7 0.025030425 0.028905738  
## 8 0.026632710 0.025760554  
## 9 0.025354641 0.035645418  
## 10 0.014707195 0.017710381  
## 11 0.012949850 0.008705420  
## 12 0.011465034 0.012562014  
## 13 0.011709371 0.008106337  
## 14 0.012038286 0.009622765  
## 15 0.021224409 0.022259665  
## 16 0.023381151 0.033511186  
## 17 0.031176435 0.041374146  
## 18 0.016760564 0.010502668  
## 19 0.009125039 0.005073481  
## 20 0.028192707 0.028755967  
## 21 0.029461378 0.035739025  
## 22 0.011892623 0.014808574  
## 23 0.012315514 0.009866142  
## 24 0.035987990 0.092483385  
## 25 0.015036110 0.011588505  
## 26 0.014157437 0.009866142  
## 27 0.037811118 0.058597772  
## 28 0.037872202 0.071534213  
## 29 0.021506336 0.020219040  
## 30 0.012052382 0.008518206  
## 31 0.012296719 0.009416830  
## 32 0.017808393 0.006777123  
## 33 0.010999854 0.004736497  
## 34 0.025199581 0.028737246  
## 35 0.013231777 0.007376205  
## 36 0.010736722 0.006608631  
## 37 0.023883921 0.031545446  
## 38 0.010652144 0.006140597  
## 39 0.010572265 0.004399513  
## 40 0.011808045 0.007881681  
## 41 0.011535516 0.009042404  
## 42 0.072060558 0.050603763  
## 43 0.018607186 0.006365253  
## 44 0.017056587 0.011382570  
## 45 0.011333468 0.006365253  
## 46 0.012536357 0.011270242  
## 47 0.011056240 0.005036039  
## 48 0.011206601 0.007563419  
## 49 0.014143341 0.011551062  
## 50 0.011507323 0.006309089

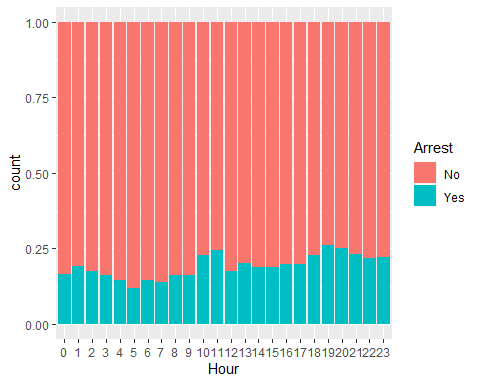
Domestic: Indicates whether or not incident involved domestic violence

ggplot(chicago1, aes(x=Domestic, fill=Arrest)) +   
 geom\_bar(position="fill")



Hour

ggplot(chicago1, aes(x=Hour, fill=Arrest)) +   
 geom\_bar(position="fill")

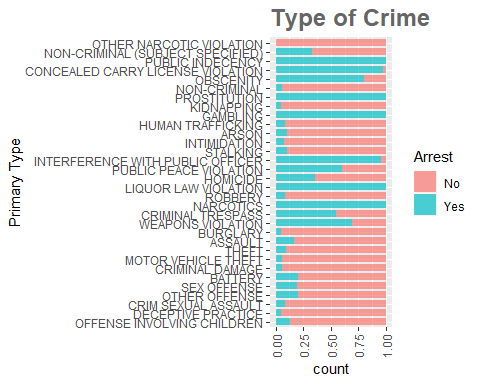


t2 = table(chicago1$Hour,chicago1$Arrest)  
prop.table(t2,margin = 2)

##   
## No Yes  
## 0 0.049699043 0.039614341  
## 1 0.030166196 0.028755967  
## 2 0.025481508 0.021548254  
## 3 0.022182961 0.016905364  
## 4 0.018085621 0.012187588  
## 5 0.017065985 0.009304502  
## 6 0.019316703 0.012917720  
## 7 0.026877047 0.017148741  
## 8 0.037186180 0.028306655  
## 9 0.048181335 0.037218010  
## 10 0.043416768 0.051352616  
## 11 0.044224959 0.057474492  
## 12 0.062803953 0.053243471  
## 13 0.048477359 0.048918843  
## 14 0.052612289 0.048338482  
## 15 0.056855292 0.052419732  
## 16 0.053721202 0.053280914  
## 17 0.055502042 0.055096883  
## 18 0.055050958 0.064626041  
## 19 0.052570000 0.074417299  
## 20 0.049243261 0.065674436  
## 21 0.046320617 0.055733408  
## 22 0.046292424 0.051764486  
## 23 0.038666297 0.043751755

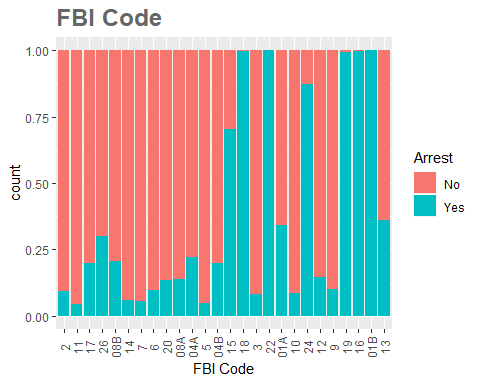
Type of Crime

ggplot(chicago1, aes(x=`Primary Type`, fill=Arrest)) +   
 geom\_bar(position="fill",alpha=.7,width=.75) +   
 theme(axis.text.x=element\_text(angle=90,hjust=2,vjust=0.5))+  
 ggtitle("Type of Crime")+  
 theme(plot.title = element\_text(color="#666666", face="bold", size=18, hjust=0))+  
 coord\_flip()



FBI COde

ggplot(chicago1, aes(x=`FBI Code`, fill=Arrest)) +   
 geom\_bar(position="fill") +   
 theme(axis.text.x=element\_text(angle=90,hjust=.5,vjust=0.5))+  
 ggtitle("FBI Code")+  
 theme(plot.title = element\_text(color="#666666", face="bold", size=18, hjust=0))



labs(y="Count")

## $y  
## [1] "Count"  
##   
## attr(,"class")  
## [1] "labels"

## Course Project Phase 2

chicago1 = chicago1 %>%   
 group\_by(`Primary Type`)  
str(chicago1)

## Classes 'grouped\_df', 'tbl\_df', 'tbl' and 'data.frame': 266236 obs. of 15 variables:  
## $ Date : POSIXct, format: "2018-01-01 00:00:00" "2018-01-01 00:00:00" ...  
## $ Block : chr "069XX N CLARK ST" "070XX N KEDZIE AVE" "072XX W BALMORAL AVE" "047XX N ARTESIAN AVE" ...  
## $ IUCR : Factor w/ 322 levels "031A","031B",..: 113 21 27 112 112 169 175 112 112 178 ...  
## $ Primary Type : Factor w/ 32 levels "OFFENSE INVOLVING CHILDREN",..: 1 2 2 1 1 3 3 1 1 4 ...  
## $ Description : Factor w/ 301 levels "$500 AND UNDER",..: 241 127 119 3 3 33 178 3 3 135 ...  
## $ 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 "FALSE","TRUE": 1 1 1 2 1 1 1 1 1 1 ...  
## $ Beat : num 2431 2411 1613 1911 932 ...  
## $ District : Factor w/ 23 levels "1","2","3","4",..: 21 21 15 18 9 20 12 7 9 15 ...  
## $ Ward : Factor w/ 50 levels "1","2","3","4",..: 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : Factor w/ 78 levels "1","2","10","4",..: 1 2 3 4 5 6 7 8 5 9 ...  
## $ FBI Code : Factor w/ 26 levels "2","11","17",..: 1 2 2 3 3 1 1 3 3 4 ...  
## $ Year : num 2018 2018 2018 2018 2018 ...  
## $ Hour : Factor w/ 24 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...  
## - attr(\*, "groups")=Classes 'tbl\_df', 'tbl' and 'data.frame': 32 obs. of 2 variables:  
## ..$ Primary Type: Factor w/ 32 levels "OFFENSE INVOLVING CHILDREN",..: 1 2 3 4 5 6 7 8 9 10 ...  
## ..$ .rows :List of 32  
## .. ..$ : int 1 4 5 8 9 12 13 18 19 23 ...  
## .. ..$ : int 2 3 11 28 33 38 44 45 61 63 ...  
## .. ..$ : int 6 7 15 47 51 68 80 94 120 186 ...  
## .. ..$ : int 10 32 52 58 66 72 81 112 136 211 ...  
## .. ..$ : int 14 16 17 20 21 46 55 59 74 85 ...  
## .. ..$ : int 22 27 30 69 142 185 198 203 206 208 ...  
## .. ..$ : int 25 34 37 121 122 125 126 127 200 202 ...  
## .. ..$ : int 26 35 209 221 243 244 286 326 328 347 ...  
## .. ..$ : int 29 31 36 39 40 62 90 91 108 117 ...  
## .. ..$ : int 60 71 250 256 270 279 305 310 317 324 ...  
## .. ..$ : int 82 123 124 239 267 289 330 337 345 358 ...  
## .. ..$ : int 199 204 205 207 210 213 214 253 257 263 ...  
## .. ..$ : int 215 318 533 545 627 717 758 766 780 800 ...  
## .. ..$ : int 234 251 306 598 608 661 664 707 763 781 ...  
## .. ..$ : int 280 297 307 341 343 352 390 485 595 609 ...  
## .. ..$ : int 299 1227 1809 3200 3762 3765 5405 6785 7522 7990 ...  
## .. ..$ : int 322 538 739 854 1599 3099 3355 3495 3508 4102 ...  
## .. ..$ : int 594 833 1092 1135 1369 1488 1607 1815 2062 2320 ...  
## .. ..$ : int 663 848 1387 1628 1878 1969 2238 2425 2457 2592 ...  
## .. ..$ : int 721 893 2581 3323 3377 4549 5760 6125 10520 12304 ...  
## .. ..$ : int 815 18160 19689 19932 20080 25479 28019 30190 30371 30978 ...  
## .. ..$ : int 1462 2058 2468 5172 5603 6848 8328 8557 9492 10510 ...  
## .. ..$ : int 1678 18710 32963 39423 57944 165442 179291 185845 199888 207738 ...  
## .. ..$ : int 1875 1899 2503 3825 17533 28873 40206 44299 49662 49721 ...  
## .. ..$ : int 1888 9807 10414 11826 13610 13650 13890 14329 14344 15065 ...  
## .. ..$ : int 1961 1970 1978 3978 3990 4007 4305 6255 6280 6291 ...  
## .. ..$ : int 2133 7884 12431 16749 26908 30885 34536 47734 56975 59605 ...  
## .. ..$ : int 3848 5964 8939 10368 12252 13152 14131 15286 16666 18427 ...  
## .. ..$ : int 12336 12835 17681 18659 23762 25653 27408 28514 30594 34836 ...  
## .. ..$ : int 60933 85152 85594 115104 115623 116955 120794 139407 156374 191988 ...  
## .. ..$ : int 69202 143046 171757  
## .. ..$ : int 230906  
## ..- attr(\*, ".drop")= logi TRUE

Split data into training and testing sets

set.seed(1234)   
train.rows = createDataPartition(y = chicago1$Arrest, p=0.7, list = FALSE)   
train = chicago1[train.rows,]   
test = chicago1[-train.rows,]

Logistic regression model

model4 = glm(Arrest ~ `Primary Type`,train, family="binomial")   
summary(model4)

##   
## Call:  
## glm(formula = Arrest ~ `Primary Type`, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.8833 -0.6070 -0.4543 -0.3084 2.5435   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -1.95520 0.07574  
## `Primary Type`DECEPTIVE PRACTICE -1.06661 0.08651  
## `Primary Type`CRIM SEXUAL ASSAULT -0.41225 0.13043  
## `Primary Type`OTHER OFFENSE 0.62332 0.07897  
## `Primary Type`SEX OFFENSE 0.52573 0.11763  
## `Primary Type`BATTERY 0.59980 0.07690  
## `Primary Type`CRIMINAL DAMAGE -0.78311 0.08150  
## `Primary Type`MOTOR VEHICLE THEFT -0.91307 0.09265  
## `Primary Type`THEFT -0.26368 0.07736  
## `Primary Type`ASSAULT 0.35715 0.07900  
## `Primary Type`BURGLARY -0.99673 0.09102  
## `Primary Type`WEAPONS VIOLATION 2.79235 0.08359  
## `Primary Type`CRIMINAL TRESPASS 2.14338 0.08108  
## `Primary Type`NARCOTICS 9.49465 0.45370  
## `Primary Type`ROBBERY -0.46300 0.08776  
## `Primary Type`LIQUOR LAW VIOLATION 17.52127 106.71495  
## `Primary Type`HOMICIDE 1.31224 0.12809  
## `Primary Type`PUBLIC PEACE VIOLATION 2.40617 0.10153  
## `Primary Type`INTERFERENCE WITH PUBLIC OFFICER 5.07156 0.18032  
## `Primary Type`STALKING -0.39618 0.31146  
## `Primary Type`INTIMIDATION -0.37429 0.32480  
## `Primary Type`ARSON -0.28125 0.22363  
## `Primary Type`HUMAN TRAFFICKING -0.12424 1.06336  
## `Primary Type`GAMBLING 17.52127 119.63296  
## `Primary Type`KIDNAPPING -1.23938 0.46253  
## `Primary Type`PROSTITUTION 17.52127 64.32014  
## `Primary Type`NON-CRIMINAL -0.57053 0.73874  
## `Primary Type`OBSCENITY 3.29893 0.33288  
## `Primary Type`CONCEALED CARRY LICENSE VIOLATION 6.55032 1.00789  
## `Primary Type`PUBLIC INDECENCY 17.52127 514.56074  
## `Primary Type`NON-CRIMINAL (SUBJECT SPECIFIED) -13.61087 1029.12147  
## `Primary Type`OTHER NARCOTIC VIOLATION -13.61087 1455.39753  
## z value Pr(>|z|)   
## (Intercept) -25.815 < 2e-16 \*\*\*  
## `Primary Type`DECEPTIVE PRACTICE -12.329 < 2e-16 \*\*\*  
## `Primary Type`CRIM SEXUAL ASSAULT -3.161 0.001574 \*\*   
## `Primary Type`OTHER OFFENSE 7.893 2.94e-15 \*\*\*  
## `Primary Type`SEX OFFENSE 4.469 7.85e-06 \*\*\*  
## `Primary Type`BATTERY 7.800 6.19e-15 \*\*\*  
## `Primary Type`CRIMINAL DAMAGE -9.609 < 2e-16 \*\*\*  
## `Primary Type`MOTOR VEHICLE THEFT -9.855 < 2e-16 \*\*\*  
## `Primary Type`THEFT -3.409 0.000653 \*\*\*  
## `Primary Type`ASSAULT 4.521 6.15e-06 \*\*\*  
## `Primary Type`BURGLARY -10.951 < 2e-16 \*\*\*  
## `Primary Type`WEAPONS VIOLATION 33.407 < 2e-16 \*\*\*  
## `Primary Type`CRIMINAL TRESPASS 26.435 < 2e-16 \*\*\*  
## `Primary Type`NARCOTICS 20.927 < 2e-16 \*\*\*  
## `Primary Type`ROBBERY -5.276 1.32e-07 \*\*\*  
## `Primary Type`LIQUOR LAW VIOLATION 0.164 0.869583   
## `Primary Type`HOMICIDE 10.245 < 2e-16 \*\*\*  
## `Primary Type`PUBLIC PEACE VIOLATION 23.698 < 2e-16 \*\*\*  
## `Primary Type`INTERFERENCE WITH PUBLIC OFFICER 28.126 < 2e-16 \*\*\*  
## `Primary Type`STALKING -1.272 0.203372   
## `Primary Type`INTIMIDATION -1.152 0.249165   
## `Primary Type`ARSON -1.258 0.208520   
## `Primary Type`HUMAN TRAFFICKING -0.117 0.906987   
## `Primary Type`GAMBLING 0.146 0.883559   
## `Primary Type`KIDNAPPING -2.680 0.007371 \*\*   
## `Primary Type`PROSTITUTION 0.272 0.785309   
## `Primary Type`NON-CRIMINAL -0.772 0.439936   
## `Primary Type`OBSCENITY 9.910 < 2e-16 \*\*\*  
## `Primary Type`CONCEALED CARRY LICENSE VIOLATION 6.499 8.08e-11 \*\*\*  
## `Primary Type`PUBLIC INDECENCY 0.034 0.972837   
## `Primary Type`NON-CRIMINAL (SUBJECT SPECIFIED) -0.013 0.989448   
## `Primary Type`OTHER NARCOTIC VIOLATION -0.009 0.992538   
## ---  
## 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: 130017 on 186334 degrees of freedom  
## AIC: 130081  
##   
## Number of Fisher Scoring iterations: 14

K-Fold

ctrl = trainControl(method = "cv",number = 10)   
  
set.seed(1234)   
modkFold = train(Arrest ~ `Primary Type`, train, method = "glm", trControl = ctrl)

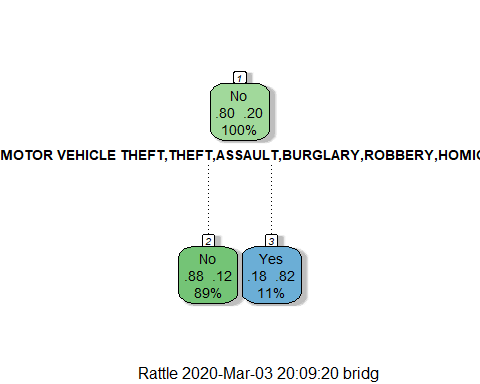
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type  
## == : prediction from a rank-deficient fit may be misleading

summary(modkFold)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.8833 -0.6070 -0.4543 -0.3084 2.5435   
##   
## Coefficients:  
## Estimate  
## (Intercept) -1.95520  
## `\\`Primary Type\\`DECEPTIVE PRACTICE` -1.06661  
## `\\`Primary Type\\`CRIM SEXUAL ASSAULT` -0.41225  
## `\\`Primary Type\\`OTHER OFFENSE` 0.62332  
## `\\`Primary Type\\`SEX OFFENSE` 0.52573  
## `\\`Primary Type\\`BATTERY` 0.59980  
## `\\`Primary Type\\`CRIMINAL DAMAGE` -0.78311  
## `\\`Primary Type\\`MOTOR VEHICLE THEFT` -0.91307  
## `\\`Primary Type\\`THEFT` -0.26368  
## `\\`Primary Type\\`ASSAULT` 0.35715  
## `\\`Primary Type\\`BURGLARY` -0.99673  
## `\\`Primary Type\\`WEAPONS VIOLATION` 2.79235  
## `\\`Primary Type\\`CRIMINAL TRESPASS` 2.14338  
## `\\`Primary Type\\`NARCOTICS` 9.49465  
## `\\`Primary Type\\`ROBBERY` -0.46300  
## `\\`Primary Type\\`LIQUOR LAW VIOLATION` 17.52127  
## `\\`Primary Type\\`HOMICIDE` 1.31224  
## `\\`Primary Type\\`PUBLIC PEACE VIOLATION` 2.40617  
## `\\`Primary Type\\`INTERFERENCE WITH PUBLIC OFFICER` 5.07156  
## `\\`Primary Type\\`STALKING` -0.39618  
## `\\`Primary Type\\`INTIMIDATION` -0.37429  
## `\\`Primary Type\\`ARSON` -0.28125  
## `\\`Primary Type\\`HUMAN TRAFFICKING` -0.12424  
## `\\`Primary Type\\`GAMBLING` 17.52127  
## `\\`Primary Type\\`KIDNAPPING` -1.23938  
## `\\`Primary Type\\`PROSTITUTION` 17.52127  
## `\\`Primary Type\\`NON-CRIMINAL` -0.57053  
## `\\`Primary Type\\`OBSCENITY` 3.29893  
## `\\`Primary Type\\`CONCEALED CARRY LICENSE VIOLATION` 6.55032  
## `\\`Primary Type\\`PUBLIC INDECENCY` 17.52127  
## `\\`Primary Type\\`NON-CRIMINAL (SUBJECT SPECIFIED)` -13.61087  
## `\\`Primary Type\\`OTHER NARCOTIC VIOLATION` -13.61087  
## Std. Error z value  
## (Intercept) 0.07574 -25.815  
## `\\`Primary Type\\`DECEPTIVE PRACTICE` 0.08651 -12.329  
## `\\`Primary Type\\`CRIM SEXUAL ASSAULT` 0.13043 -3.161  
## `\\`Primary Type\\`OTHER OFFENSE` 0.07897 7.893  
## `\\`Primary Type\\`SEX OFFENSE` 0.11763 4.469  
## `\\`Primary Type\\`BATTERY` 0.07690 7.800  
## `\\`Primary Type\\`CRIMINAL DAMAGE` 0.08150 -9.609  
## `\\`Primary Type\\`MOTOR VEHICLE THEFT` 0.09265 -9.855  
## `\\`Primary Type\\`THEFT` 0.07736 -3.409  
## `\\`Primary Type\\`ASSAULT` 0.07900 4.521  
## `\\`Primary Type\\`BURGLARY` 0.09102 -10.951  
## `\\`Primary Type\\`WEAPONS VIOLATION` 0.08359 33.407  
## `\\`Primary Type\\`CRIMINAL TRESPASS` 0.08108 26.435  
## `\\`Primary Type\\`NARCOTICS` 0.45370 20.927  
## `\\`Primary Type\\`ROBBERY` 0.08776 -5.276  
## `\\`Primary Type\\`LIQUOR LAW VIOLATION` 106.71495 0.164  
## `\\`Primary Type\\`HOMICIDE` 0.12809 10.245  
## `\\`Primary Type\\`PUBLIC PEACE VIOLATION` 0.10153 23.698  
## `\\`Primary Type\\`INTERFERENCE WITH PUBLIC OFFICER` 0.18032 28.126  
## `\\`Primary Type\\`STALKING` 0.31146 -1.272  
## `\\`Primary Type\\`INTIMIDATION` 0.32480 -1.152  
## `\\`Primary Type\\`ARSON` 0.22363 -1.258  
## `\\`Primary Type\\`HUMAN TRAFFICKING` 1.06336 -0.117  
## `\\`Primary Type\\`GAMBLING` 119.63296 0.146  
## `\\`Primary Type\\`KIDNAPPING` 0.46253 -2.680  
## `\\`Primary Type\\`PROSTITUTION` 64.32014 0.272  
## `\\`Primary Type\\`NON-CRIMINAL` 0.73874 -0.772  
## `\\`Primary Type\\`OBSCENITY` 0.33288 9.910  
## `\\`Primary Type\\`CONCEALED CARRY LICENSE VIOLATION` 1.00789 6.499  
## `\\`Primary Type\\`PUBLIC INDECENCY` 514.56074 0.034  
## `\\`Primary Type\\`NON-CRIMINAL (SUBJECT SPECIFIED)` 1029.12147 -0.013  
## `\\`Primary Type\\`OTHER NARCOTIC VIOLATION` 1455.39753 -0.009  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## `\\`Primary Type\\`DECEPTIVE PRACTICE` < 2e-16 \*\*\*  
## `\\`Primary Type\\`CRIM SEXUAL ASSAULT` 0.001574 \*\*   
## `\\`Primary Type\\`OTHER OFFENSE` 2.94e-15 \*\*\*  
## `\\`Primary Type\\`SEX OFFENSE` 7.85e-06 \*\*\*  
## `\\`Primary Type\\`BATTERY` 6.19e-15 \*\*\*  
## `\\`Primary Type\\`CRIMINAL DAMAGE` < 2e-16 \*\*\*  
## `\\`Primary Type\\`MOTOR VEHICLE THEFT` < 2e-16 \*\*\*  
## `\\`Primary Type\\`THEFT` 0.000653 \*\*\*  
## `\\`Primary Type\\`ASSAULT` 6.15e-06 \*\*\*  
## `\\`Primary Type\\`BURGLARY` < 2e-16 \*\*\*  
## `\\`Primary Type\\`WEAPONS VIOLATION` < 2e-16 \*\*\*  
## `\\`Primary Type\\`CRIMINAL TRESPASS` < 2e-16 \*\*\*  
## `\\`Primary Type\\`NARCOTICS` < 2e-16 \*\*\*  
## `\\`Primary Type\\`ROBBERY` 1.32e-07 \*\*\*  
## `\\`Primary Type\\`LIQUOR LAW VIOLATION` 0.869583   
## `\\`Primary Type\\`HOMICIDE` < 2e-16 \*\*\*  
## `\\`Primary Type\\`PUBLIC PEACE VIOLATION` < 2e-16 \*\*\*  
## `\\`Primary Type\\`INTERFERENCE WITH PUBLIC OFFICER` < 2e-16 \*\*\*  
## `\\`Primary Type\\`STALKING` 0.203372   
## `\\`Primary Type\\`INTIMIDATION` 0.249165   
## `\\`Primary Type\\`ARSON` 0.208520   
## `\\`Primary Type\\`HUMAN TRAFFICKING` 0.906987   
## `\\`Primary Type\\`GAMBLING` 0.883559   
## `\\`Primary Type\\`KIDNAPPING` 0.007371 \*\*   
## `\\`Primary Type\\`PROSTITUTION` 0.785309   
## `\\`Primary Type\\`NON-CRIMINAL` 0.439936   
## `\\`Primary Type\\`OBSCENITY` < 2e-16 \*\*\*  
## `\\`Primary Type\\`CONCEALED CARRY LICENSE VIOLATION` 8.08e-11 \*\*\*  
## `\\`Primary Type\\`PUBLIC INDECENCY` 0.972837   
## `\\`Primary Type\\`NON-CRIMINAL (SUBJECT SPECIFIED)` 0.989448   
## `\\`Primary Type\\`OTHER NARCOTIC VIOLATION` 0.992538   
## ---  
## 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: 130017 on 186334 degrees of freedom  
## AIC: 130081  
##   
## Number of Fisher Scoring iterations: 14

Classification Tree

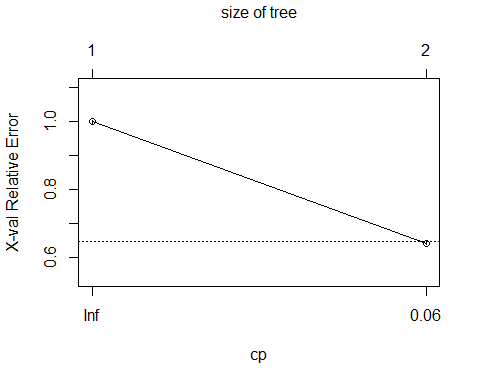
tree1 = rpart(Arrest~ `Primary Type`,train, method="class")  
fancyRpartPlot(tree1)



printcp(tree1)

##   
## Classification tree:  
## rpart(formula = Arrest ~ `Primary Type`, data = train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Primary Type  
##   
## Root node error: 37391/186366 = 0.20063  
##   
## n= 186366   
##   
## CP nsplit rel error xerror xstd  
## 1 0.35792 0 1.00000 1.00000 0.0046237  
## 2 0.01000 1 0.64208 0.64208 0.0038678

plotcp(tree1)



Predictions - 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

Confusion Matrix - Training

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

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 145231 20264  
## Yes 3744 17127  
##   
## Accuracy : 0.8712   
## 95% CI : (0.8696, 0.8727)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5188   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.4581   
## Specificity : 0.9749   
## Pos Pred Value : 0.8206   
## Neg Pred Value : 0.8776   
## Prevalence : 0.2006   
## Detection Rate : 0.0919   
## Detection Prevalence : 0.1120   
## Balanced Accuracy : 0.7165   
##   
## 'Positive' Class : Yes   
##

Predictions - testing 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

Confustion Matrix-test

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

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 62205 8665  
## Yes 1641 7359  
##   
## Accuracy : 0.871   
## 95% CI : (0.8686, 0.8733)  
## No Information Rate : 0.7994   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5187   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.45925   
## Specificity : 0.97430   
## Pos Pred Value : 0.81767   
## Neg Pred Value : 0.87773   
## Prevalence : 0.20063   
## Detection Rate : 0.09214   
## Detection Prevalence : 0.11268   
## Balanced Accuracy : 0.71677   
##   
## 'Positive' Class : Yes   
##

`