# BAN 502 - Course Project

## Phase 1

### Runge, Laura

library(tidyverse)  
library(lubridate)  
library(VIM)  
library(gridExtra)  
library(GGally)  
library(MASS)  
library(car)  
library(ranger)  
library(mice)  
library(caret)  
library(forcats)  
library(rcompanion)

**CLEANING DATA**

Load Dataset

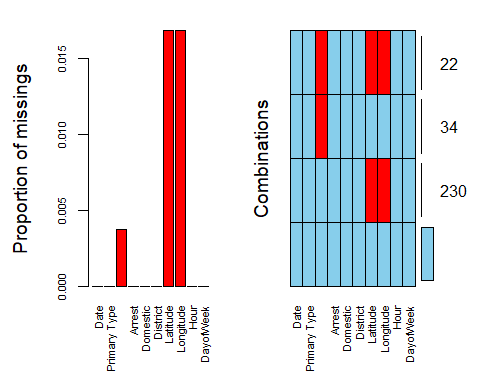
chicago <- read\_csv("chicago.csv")

Select Variables of Interest  
Mutate Variables to categorized factors

chicago2=chicago %>%  
 dplyr::select(-c("ID", "Block", "Description", "IUCR","Case Number","Updated On", "X Coordinate","Y Coordinate","Location", "FBI Code","Year","X1", "Beat", "Ward", "Community Area")) %>%  
 mutate(Date = mdy\_hm(Date)) %>%  
 mutate(Hour = hour(Date)) %>%  
 mutate(DayofWeek = wday(Date)) %>%  
 mutate(Hour=as.factor(Hour))  
  
  
chicago2$`Primary Type`=as.factor(chicago2$`Primary Type`)  
chicago2$District=as.factor(chicago2$District)  
  
  
chicago2 = chicago2 %>% mutate(Arrest = as.factor(Arrest)) %>%   
 mutate(Arrest = fct\_recode(Arrest, "No" = "FALSE", "Yes" = "TRUE" ))  
  
chicago2 = chicago2 %>% mutate(Domestic = as.factor(Domestic)) %>%   
 mutate(Domestic = fct\_recode(Domestic, "No" = "FALSE", "Yes" = "TRUE" ))  
  
chicago2 = chicago2 %>% mutate(DayofWeek= as.factor(DayofWeek)) %>%   
 mutate(DayofWeek = fct\_recode(DayofWeek, "Sunday" = "1", "Monday" = "2", "Tuesday"="3", "Wednesday"="4", "Thursday"="5", "Friday"="6", "Saturday"="7" ))

Review missing observations

vim\_plot = aggr(chicago2, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



Remove missing observations

chicago2=chicago2 %>%  
 dplyr::select(-c("Latitude","Longitude","Date")) %>%  
 drop\_na()

Consolidate factor levels

Location Description:

chicago2$`Location Description` = fct\_collapse(chicago2$`Location Description`,   
 "Airport" = c("AIRCRAFT","AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA","AIRPORT BUILDING NON-TERMINAL - SECURE AREA","AIRPORT EXTERIOR - NON-SECURE AREA","AIRPORT EXTERIOR - SECURE AREA","AIRPORT PARKING LOT","AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA","AIRPORT TERMINAL LOWER LEVEL - SECURE AREA","AIRPORT TERMINAL UPPER LEVEL - NON-SECURE AREA","AIRPORT TERMINAL UPPER LEVEL - SECURE AREA","AIRPORT TRANSPORTATION SYSTEM (ATS)","AIRPORT VENDING ESTABLISHMENT"),   
 "Business/Retail/Resturant"=c("APPLIANCE STORE","ATHLETIC CLUB","AUTO / BOAT / RV DEALERSHIP","BAR OR TAVERN","BARBERSHOP","BOWLING ALLEY","CAR WASH","CLEANING STORE","COMMERCIAL / BUSINESS OFFICE","CONVENIENCE STORE","DEPARTMENT STORE","DRUG STORE","GAS STATION","GROCERY FOOD STORE","MOVIE HOUSE/THEATER","PAWN SHOP","RESTAURANT","SMALL RETAIL STORE","TAVERN/LIQUOR STORE","FACTORY/MANUFACTURING BUILDING","WAREHOUSE","ATM (AUTOMATIC TELLER MACHINE)","BANK","CURRENCY EXCHANGE","SAVINGS AND LOAN","HOTEL","HOTEL/MOTEL"),   
 "Government Building/Land"=c("FEDERAL BUILDING","FIRE STATION","GOVERNMENT BUILDING","GOVERNMENT BUILDING/PROPERTY","LIBRARY","FOREST PRESERVE","LAKEFRONT/WATERFRONT/RIVERBANK","PARK PROPERTY","JAIL / LOCK-UP FACILITY","POLICE FACILITY/VEH PARKING LOT"),   
 "Medical/Hospital"=c("ANIMAL HOSPITAL","HOSPITAL BUILDING/GROUNDS","MEDICAL/DENTAL OFFICE","NURSING HOME/RETIREMENT HOME"),  
 "Other"=c("CEMETARY","CONSTRUCTION SITE","OTHER","POOL ROOM","SPORTS ARENA/STADIUM","ABANDONED BUILDING","VACANT LOT", "VACANT LOT/LAND","CHURCH/SYNAGOGUE/PLACE OF WORSHIP"),  
 "Public Transportation"=c("CTA BUS","CTA BUS STOP","CTA GARAGE / OTHER PROPERTY","CTA PLATFORM","CTA STATION","CTA TRAIN","OTHER COMMERCIAL TRANSPORTATION","OTHER RAILROAD PROP / TRAIN DEPOT"),  
 "Residential"=c("APARTMENT","CHA APARTMENT","CHA HALLWAY/STAIRWELL/ELEVATOR","CHA PARKING LOT/GROUNDS","DRIVEWAY - RESIDENTIAL","HALLWAY","HOUSE","PORCH","RESIDENCE","RESIDENCE PORCH/HALLWAY","RESIDENCE-GARAGE","RESIDENTIAL YARD (FRONT/BACK)","YARD"),  
 "School/College"=c("COLLEGE/UNIVERSITY GROUNDS","COLLEGE/UNIVERSITY RESIDENCE HALL","DAY CARE CENTER","SCHOOL, PRIVATE, BUILDING","SCHOOL, PRIVATE, GROUNDS","SCHOOL, PUBLIC, BUILDING","SCHOOL, PUBLIC, GROUNDS"),  
 "Street"=c("ALLEY","BRIDGE","HIGHWAY/EXPRESSWAY","PARKING LOT","PARKING LOT/GARAGE(NON.RESID.)","SIDEWALK","STREET"),  
 "Vehicle/Boat"=c("TAXICAB","VEHICLE - DELIVERY TRUCK","VEHICLE - OTHER RIDE SHARE SERVICE (E.G., UBER, LYFT)","VEHICLE NON-COMMERCIAL","VEHICLE-COMMERCIAL","VEHICLE-COMMERCIAL - ENTERTAINMENT/PARTY BUS","AUTO","BOAT/WATERCRAFT"))

Primary Type:

chicago2$`Primary Type` = fct\_collapse(chicago2$`Primary Type`,  
 "Assault"="ASSAULT",  
 "Battery"="BATTERY",  
 "Burglary"="BURGLARY",  
 "Damage"="CRIMINAL DAMAGE",  
 "Disturbing the Peace"=c("PUBLIC PEACE VIOLATION","INTERFERENCE WITH PUBLIC OFFICER","OBSCENITY","INTIMIDATION"),  
 "Drugs/Alcohol"=c("NARCOTICS","LIQUOR LAW VIOLATION"),  
 "Fraud"="DECEPTIVE PRACTICE",  
 "Homicide"="HOMICIDE",  
 "Other"=c("KIDNAPPING","OTHER OFFENSE","ARSON","STALKING","OFFENSE INVOLVING CHILDREN","GAMBLING","NON-CRIMINAL"),  
 "Robbery"="ROBBERY",  
 "Sexual"=c("SEX OFFENSE","CRIM SEXUAL ASSAULT","PROSTITUTION","PUBLIC INDECENCY"),  
 "Theft"="THEFT",  
 "Auto Theft"="MOTOR VEHICLE THEFT",  
 "Trespass"="CRIMINAL TRESPASS",  
 "Weapons"=c("WEAPONS VIOLATION","CONCEALED CARRY LICENSE VIOLATION"))

**REVIEWING DATA**

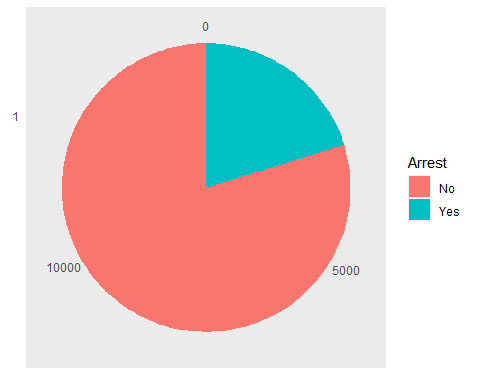
fit\_control = trainControl(method = "cv", number = 10)  
  
set.seed(1234)   
rf\_fit = train(x=chicago2[,-3], y=chicago2$Arrest,  
 method = "ranger",   
 importance = "permutation",  
 trControl = fit\_control)  
  
varImp(rf\_fit)

## ranger variable importance  
##   
## Overall  
## Primary Type 100.000  
## Location Description 22.010  
## Domestic 11.714  
## District 8.443  
## Hour 2.427  
## DayofWeek 0.000

rf\_fit

## Random Forest   
##   
## 14944 samples  
## 6 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 13449, 13450, 13449, 13449, 13449, 13450, ...   
## Resampling results across tuning parameters:  
##   
## mtry splitrule Accuracy Kappa   
## 2 gini 0.8694452 0.4959947  
## 2 extratrees 0.8722558 0.5013754  
## 4 gini 0.8563302 0.4962328  
## 4 extratrees 0.8672370 0.5186779  
## 6 gini 0.8503079 0.4887614  
## 6 extratrees 0.8683744 0.5289181  
##   
## 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 = extratrees  
## and min.node.size = 1.

ggplot(chicago2, aes(x=factor(1), fill=Arrest))+  
 geom\_bar(width = 1)+  
 coord\_polar("y")+  
 theme(  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 panel.border = element\_blank(),  
 panel.grid=element\_blank(),  
 axis.ticks = element\_blank(),  
 plot.title=element\_text(size=14, face="bold")  
 )

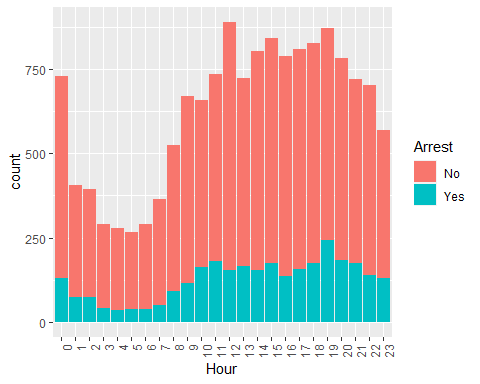


pietable=table(chicago2$Arrest)

**Time Variables**

Hour

p1=ggplot(chicago2,aes(x=Hour, fill = Arrest)) + geom\_bar()+ theme(axis.text.x = element\_text(angle=90, hjust=1))  
p1



t1 = table(chicago2$Arrest,chicago2$Hour)  
t1

##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17  
## No 600 333 322 248 243 231 254 315 432 553 495 553 735 558 650 668 653 652  
## Yes 129 74 73 42 35 37 37 50 92 117 164 181 155 165 153 175 137 158  
##   
## 18 19 20 21 22 23  
## No 653 630 601 543 563 439  
## Yes 174 242 183 176 140 131

prop.table(t1, margin = 2)

##   
## 0 1 2 3 4 5 6  
## No 0.8230453 0.8181818 0.8151899 0.8551724 0.8741007 0.8619403 0.8728522  
## Yes 0.1769547 0.1818182 0.1848101 0.1448276 0.1258993 0.1380597 0.1271478  
##   
## 7 8 9 10 11 12 13  
## No 0.8630137 0.8244275 0.8253731 0.7511381 0.7534060 0.8258427 0.7717842  
## Yes 0.1369863 0.1755725 0.1746269 0.2488619 0.2465940 0.1741573 0.2282158  
##   
## 14 15 16 17 18 19 20  
## No 0.8094645 0.7924081 0.8265823 0.8049383 0.7896010 0.7224771 0.7665816  
## Yes 0.1905355 0.2075919 0.1734177 0.1950617 0.2103990 0.2775229 0.2334184  
##   
## 21 22 23  
## No 0.7552156 0.8008535 0.7701754  
## Yes 0.2447844 0.1991465 0.2298246

kruskal.test(x=as.factor(chicago2$Hour),g=as.factor(chicago2$Arrest))

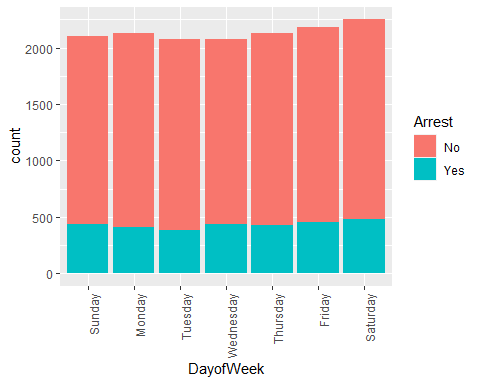
##   
## Kruskal-Wallis rank sum test  
##   
## data: as.factor(chicago2$Hour) and as.factor(chicago2$Arrest)  
## Kruskal-Wallis chi-squared = 43.046, df = 1, p-value = 5.346e-11

cramerV(chicago2$Arrest, chicago2$Hour, bias.correct = FALSE)

## Cramer V   
## 0.09316

Day of the week

p2=ggplot(chicago2,aes(x=DayofWeek, fill = Arrest)) + geom\_bar()+ theme(axis.text.x = element\_text(angle=90, hjust=1))  
p2



t2 = table(chicago2$Arrest,chicago2$DayofWeek)  
t2

##   
## Sunday Monday Tuesday Wednesday Thursday Friday Saturday  
## No 1665 1717 1696 1645 1702 1727 1772  
## Yes 438 408 382 433 428 454 477

prop.table(t2, margin = 2)

##   
## Sunday Monday Tuesday Wednesday Thursday Friday Saturday  
## No 0.7917261 0.8080000 0.8161694 0.7916266 0.7990610 0.7918386 0.7879057  
## Yes 0.2082739 0.1920000 0.1838306 0.2083734 0.2009390 0.2081614 0.2120943

kruskal.test(x=as.factor(chicago2$DayofWeek),g=as.factor(chicago2$Arrest))

##   
## Kruskal-Wallis rank sum test  
##   
## data: as.factor(chicago2$DayofWeek) and as.factor(chicago2$Arrest)  
## Kruskal-Wallis chi-squared = 1.83, df = 1, p-value = 0.1761

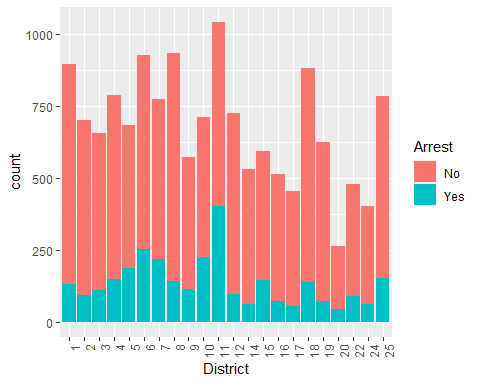
cramerV(chicago2$Arrest, chicago2$DayofWeek, bias.correct = FALSE)

## Cramer V   
## 0.02393

**Location Variables**

Distict

p3=ggplot(chicago2,aes(x=District, fill = Arrest)) + geom\_bar()+ theme(axis.text.x = element\_text(angle=90, hjust=1))  
p3



t3 = table(chicago2$Arrest,chicago2$District)  
t3

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 17 18 19  
## No 765 608 546 640 495 674 556 791 460 486 638 629 471 449 440 401 743 552  
## Yes 132 93 111 148 188 254 219 141 113 227 404 96 61 145 72 55 137 73  
##   
## 20 22 24 25  
## No 220 389 341 630  
## Yes 45 90 63 153

prop.table(t3, margin = 2)

##   
## 1 2 3 4 5 6 7  
## No 0.8528428 0.8673324 0.8310502 0.8121827 0.7247438 0.7262931 0.7174194  
## Yes 0.1471572 0.1326676 0.1689498 0.1878173 0.2752562 0.2737069 0.2825806  
##   
## 8 9 10 11 12 14 15  
## No 0.8487124 0.8027923 0.6816269 0.6122841 0.8675862 0.8853383 0.7558923  
## Yes 0.1512876 0.1972077 0.3183731 0.3877159 0.1324138 0.1146617 0.2441077  
##   
## 16 17 18 19 20 22 24  
## No 0.8593750 0.8793860 0.8443182 0.8832000 0.8301887 0.8121086 0.8440594  
## Yes 0.1406250 0.1206140 0.1556818 0.1168000 0.1698113 0.1878914 0.1559406  
##   
## 25  
## No 0.8045977  
## Yes 0.1954023

kruskal.test(x=as.factor(chicago2$District),g=as.factor(chicago2$Arrest))

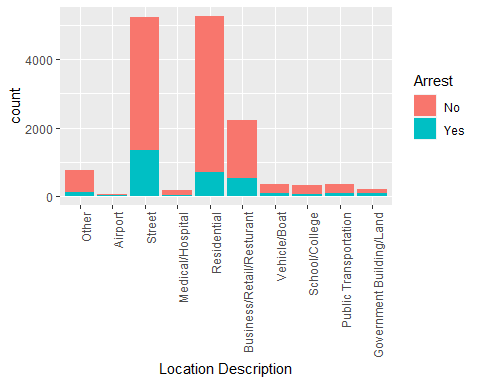
##   
## Kruskal-Wallis rank sum test  
##   
## data: as.factor(chicago2$District) and as.factor(chicago2$Arrest)  
## Kruskal-Wallis chi-squared = 8.6502, df = 1, p-value = 0.00327

cramerV(chicago2$Arrest, chicago2$District, bias.correct = FALSE)

## Cramer V   
## 0.1929

Location Description

p4=ggplot(chicago2,aes(x=`Location Description`, fill = Arrest)) + geom\_bar() + theme(axis.text.x = element\_text(angle=90, hjust=1))  
p4



t4 = table(chicago2$Arrest,chicago2$`Location Description`)  
t4

##   
## Other Airport Street Medical/Hospital Residential  
## No 647 54 3905 130 4574  
## Yes 102 17 1340 30 697  
##   
## Business/Retail/Resturant Vehicle/Boat School/College  
## No 1681 268 273  
## Yes 531 85 53  
##   
## Public Transportation Government Building/Land  
## No 263 129  
## Yes 86 79

prop.table(t4, margin = 2)

##   
## Other Airport Street Medical/Hospital Residential  
## No 0.8638184 0.7605634 0.7445186 0.8125000 0.8677670  
## Yes 0.1361816 0.2394366 0.2554814 0.1875000 0.1322330  
##   
## Business/Retail/Resturant Vehicle/Boat School/College  
## No 0.7599458 0.7592068 0.8374233  
## Yes 0.2400542 0.2407932 0.1625767  
##   
## Public Transportation Government Building/Land  
## No 0.7535817 0.6201923  
## Yes 0.2464183 0.3798077

kruskal.test(x=as.factor(chicago2$`Location Description`),g=as.factor(chicago2$Arrest))

##   
## Kruskal-Wallis rank sum test  
##   
## data: as.factor(chicago2$`Location Description`) and as.factor(chicago2$Arrest)  
## Kruskal-Wallis chi-squared = 5.8978, df = 1, p-value = 0.01516

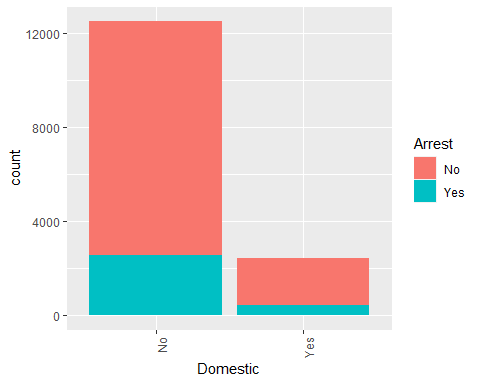
cramerV(chicago2$Arrest, chicago2$`Location Description`, bias.correct = FALSE)

## Cramer V   
## 0.1518

**Incident Type**

Domestic

p5=ggplot(chicago2,aes(x=Domestic, fill = Arrest)) + geom\_bar()+ theme(axis.text.x = element\_text(angle=90, hjust=1))  
p5



t5 = table(chicago2$Arrest,chicago2$Domestic)  
t5

##   
## No Yes  
## No 9910 2014  
## Yes 2584 436

prop.table(t5, margin = 2)

##   
## No Yes  
## No 0.7931807 0.8220408  
## Yes 0.2068193 0.1779592

kruskal.test(x=as.factor(chicago2$Domestic),g=as.factor(chicago2$Arrest))

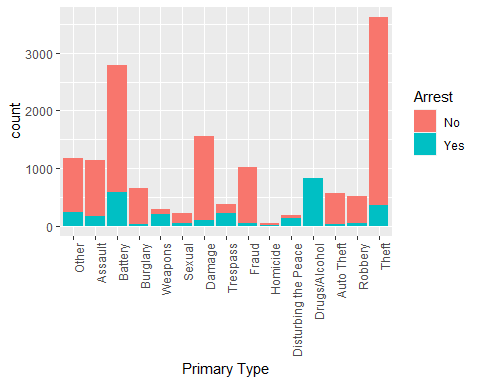
##   
## Kruskal-Wallis rank sum test  
##   
## data: as.factor(chicago2$Domestic) and as.factor(chicago2$Arrest)  
## Kruskal-Wallis chi-squared = 10.58, df = 1, p-value = 0.001143

cramerV(chicago2$Arrest, chicago2$Domestic, bias.correct = FALSE)

## Cramer V   
## 0.02661

Primary Type

p6=ggplot(chicago2,aes(x=`Primary Type`, fill = Arrest)) + geom\_bar()+ theme(axis.text.x = element\_text(angle=90, hjust=1))  
p6



t6 = table(chicago2$Arrest,chicago2$`Primary Type`)  
t6

##   
## Other Assault Battery Burglary Weapons Sexual Damage Trespass Fraud  
## No 934 959 2208 623 88 170 1459 164 972  
## Yes 232 172 576 34 194 52 101 217 49  
##   
## Homicide Disturbing the Peace Drugs/Alcohol Auto Theft Robbery Theft  
## No 27 40 0 547 478 3255  
## Yes 14 139 817 24 38 361

prop.table(t6, margin = 2)

##   
## Other Assault Battery Burglary Weapons Sexual  
## No 0.80102916 0.84792219 0.79310345 0.94824962 0.31205674 0.76576577  
## Yes 0.19897084 0.15207781 0.20689655 0.05175038 0.68794326 0.23423423  
##   
## Damage Trespass Fraud Homicide Disturbing the Peace  
## No 0.93525641 0.43044619 0.95200784 0.65853659 0.22346369  
## Yes 0.06474359 0.56955381 0.04799216 0.34146341 0.77653631  
##   
## Drugs/Alcohol Auto Theft Robbery Theft  
## No 0.00000000 0.95796848 0.92635659 0.90016593  
## Yes 1.00000000 0.04203152 0.07364341 0.09983407

kruskal.test(x=as.factor(chicago2$`Primary Type`),g=as.factor(chicago2$Arrest))

##   
## Kruskal-Wallis rank sum test  
##   
## data: as.factor(chicago2$`Primary Type`) and as.factor(chicago2$Arrest)  
## Kruskal-Wallis chi-squared = 8.7616, df = 1, p-value = 0.003076

cramerV(chicago2$Arrest, chicago2$`Primary Type`, bias.correct = FALSE)

## Cramer V   
## 0.5871