# Fields, Alex

## BAN 502

### Phase 1 Course Project

#### Part 1 - Data Exploration

options(tidyverse.quiet = TRUE)  
library(tidyverse)#Data Cleaning/Wrangling  
library(MASS)#Statistics  
library(GGally)#Correlation  
library(caret)#ML  
library(VIM)#Missing data  
library(mice)#Missing data\_2  
library(flexdashboard)#Dashboarding  
library(lubridate)#Date and Time Functionality  
library(ggpubr)  
library(Hmisc)  
library(DataExplorer)#Density Plot  
theme\_set(theme\_pubr())

## Import/Viewing Dataset

chicago2 <- read\_csv("chicago2.csv")

## Warning: Missing column names filled in: 'X1' [1]

chicago2 = chicago2[-1]#drops first column

### Comments from *Import/View*

We can see when importing the dataset that there is a “Key” Variable that is not needed for our purposes. It is not labeled and R defaults it “X1”. I used bracket notation to remove this for cleaner data.

When looking at the data, I can see that this a large dataset. 15000 rows and 22 variables/columns. Anything computationally heavy would not be ideal on this dataset.

### Refactoring Data

chicago2 = chicago2 %>% mutate(Date = mdy\_hms(Date))  
chicago2 = chicago2 %>% mutate(Hour = hour(Date))  
chicago2 = chicago2 %>% mutate(Months = month(Date))  
  
  
chicago2 = chicago2 %>% mutate(Hour = as.factor(Hour))  
chicago2 = chicago2 %>% mutate(Months = as.factor(Months))  
  
  
chicago2 = chicago2 %>% mutate(Arrest = as\_factor(as.character(Arrest))) %>%  
mutate(Arrest = fct\_recode(Arrest,  
"Arrested" = "TRUE",  
"Not Arrested" = "FALSE"))  
  
chicago2 = chicago2 %>% mutate(Domestic = as\_factor(as.character(Domestic))) %>%  
mutate(Domestic = fct\_recode(Domestic,  
"Domestic Violence" = "TRUE",  
"No Domestic Violence" = "FALSE"))  
  
  
chicago2 = chicago2 %>% mutate(`FBI Code` = as\_factor(as.character(`FBI Code`))) %>%  
mutate(`FBI Code` = fct\_recode(`FBI Code`,  
"Homicide" = "01A",  
"Invol. Manslaughter" = "01B",  
"Sexual Assault" = "02",  
"Robbery" = "03",  
"Aggravated Assault" = "04A",  
"Agravated Battery" = "04B",  
"Buglary" = "05",  
"Larceny" = "06",  
"Motor Vehicle Theft" = "07",  
"Simple Assault" = "08A",  
"Simple Battery" = "08B",  
"Arson" = "09",  
"Forgery & Conterfeiting" = "10",  
"Fraud" = "11",  
"Embezzlement" = "12",  
"Stolen Property" = "13",  
"Vandalism" = "14",  
"Weapons Violation" = "15",  
"Prostitution" = "16",  
"Criminal Sexual Abuse" = "17",  
"Drug Abuse" = "18",  
"Gambling" = "19",  
"Offenses Against Family" = "20",  
"Liquor License" = "22",  
"Disorderly Conduct" = "24",  
"Misc. Offenses" = "26"))

## Warning: Unknown levels in `f`: 01B

#Reorder data  
chicago2 <- chicago2 %>% group\_by(Arrest)  
chicago2 %>% arrange(desc(Hour))

## # A tibble: 15,000 x 24  
## # Groups: Arrest [2]  
## ID `Case Number` Date Block IUCR `Primary Type`  
## <dbl> <chr> <dttm> <chr> <chr> <chr>   
## 1 1.14e7 JB338867 2018-07-06 23:00:00 028X~ 0460 BATTERY   
## 2 1.14e7 JB358529 2018-07-20 23:30:00 050X~ 0486 BATTERY   
## 3 1.12e7 JB139406 2018-02-02 23:40:00 018X~ 0486 BATTERY   
## 4 1.12e7 JB172858 2018-03-03 23:33:00 038X~ 2024 NARCOTICS   
## 5 1.15e7 JB481097 2018-10-18 23:16:00 005X~ 0320 ROBBERY   
## 6 1.15e7 JB497684 2018-10-30 23:00:00 040X~ 0420 BATTERY   
## 7 1.14e7 JB426750 2018-09-07 23:19:00 081X~ 0820 THEFT   
## 8 1.13e7 JB250822 2018-05-03 23:00:00 017X~ 0910 MOTOR VEHICLE~  
## 9 1.12e7 JB132391 2018-01-27 23:30:00 005X~ 1479 CONCEALED CAR~  
## 10 1.13e7 JB296195 2018-06-06 23:02:00 008X~ 041A BATTERY   
## # ... with 14,990 more rows, and 18 more variables: Description <chr>,  
## # `Location Description` <chr>, Arrest <fct>, Domestic <fct>, Beat <chr>,  
## # District <chr>, Ward <dbl>, `Community Area` <dbl>, `FBI Code` <fct>, `X  
## # Coordinate` <dbl>, `Y Coordinate` <dbl>, Year <dbl>, `Updated On` <chr>,  
## # Latitude <dbl>, Longitude <dbl>, Location <chr>, Hour <fct>, Months <fct>

chicago2 <- chicago2 %>% group\_by(Year)  
chicago2 %>% arrange(desc(Months))

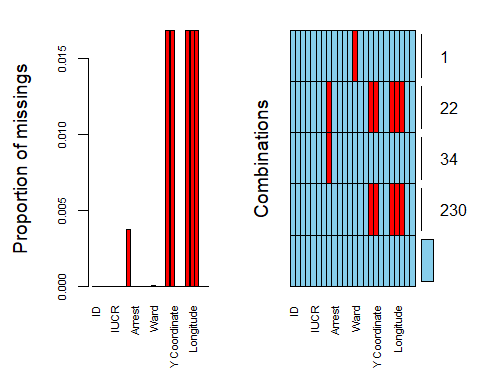
## # A tibble: 15,000 x 24  
## # Groups: Year [1]  
## ID `Case Number` Date Block IUCR `Primary Type`  
## <dbl> <chr> <dttm> <chr> <chr> <chr>   
## 1 1.15e7 JB562732 2018-12-18 15:36:00 006X~ 0860 THEFT   
## 2 1.15e7 JB560399 2018-12-19 18:21:00 047X~ 0460 BATTERY   
## 3 1.15e7 JB547841 2018-12-05 14:00:00 085X~ 0620 BURGLARY   
## 4 1.15e7 JB550645 2018-12-12 18:00:00 0000~ 0890 THEFT   
## 5 1.15e7 JB565310 2018-12-23 17:00:00 001X~ 0860 THEFT   
## 6 1.15e7 JB541194 2018-12-05 13:13:00 019X~ 0860 THEFT   
## 7 1.15e7 JB544601 2018-12-08 00:42:00 049X~ 1710 OFFENSE INVOL~  
## 8 1.15e7 JB567510 2018-12-26 09:23:00 047X~ 0486 BATTERY   
## 9 1.15e7 JB551528 2018-12-13 12:45:00 007X~ 0530 ASSAULT   
## 10 1.16e7 JC161712 2018-12-01 09:00:00 034X~ 1153 DECEPTIVE PRA~  
## # ... with 14,990 more rows, and 18 more variables: Description <chr>,  
## # `Location Description` <chr>, Arrest <fct>, Domestic <fct>, Beat <chr>,  
## # District <chr>, Ward <dbl>, `Community Area` <dbl>, `FBI Code` <fct>, `X  
## # Coordinate` <dbl>, `Y Coordinate` <dbl>, Year <dbl>, `Updated On` <chr>,  
## # Latitude <dbl>, Longitude <dbl>, Location <chr>, Hour <fct>, Months <fct>

### Comments from *Refactoring Data*

We can see that we have to do some clean up of our data set. We need to convert FBI Code, Arrest and Domestic into factors and recode these variables. We also had to create two new variables, Hour and Months. These were also factored. This will be easier to use in the future.

### View missing data

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

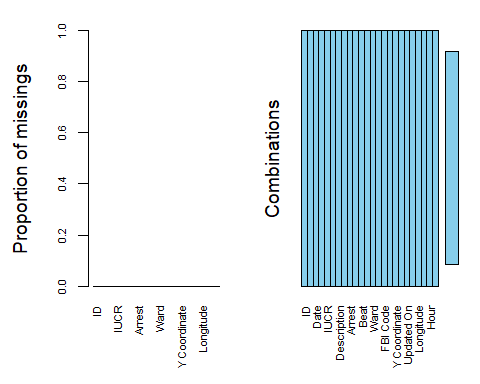


countNA(chicago2)

## [1] 1317

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

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



### Comments from *Missing Data*

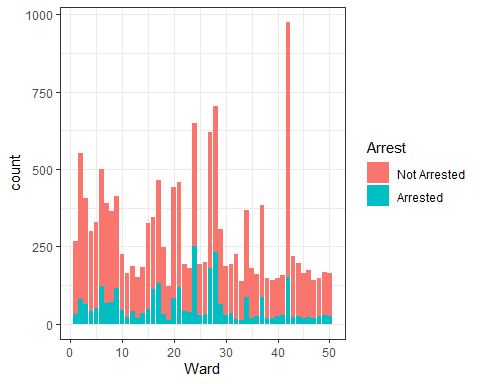
We are showing missing data on Hour, Longitiude, Updated On, Y Coordinate and Description. These are a small percentage of missing data so we will run the Tidyverse drop\_na() function. This will give us the cleasnest possible data.

## Data Cleansing

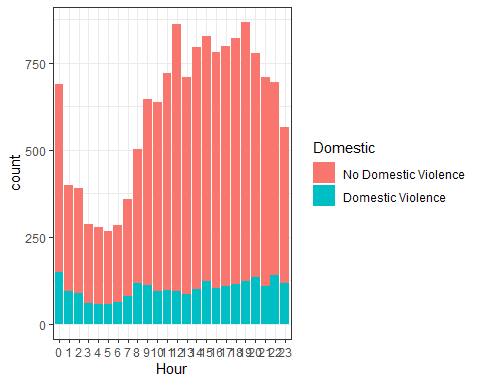
drop <- c("Y Coordinate","Year", "ID", "Case Number", "Updated On", "X Coordinate", "Location") # Drop these variables for insignificance  
new\_chicago = chicago2[,!(names(chicago2) %in% drop)]  
  
new\_chicago = filter(new\_chicago, `FBI Code` != 'Embezzlement' & `FBI Code` != 'Gambling' & `FBI Code` != 'Liquor License' & `FBI Code` != 'Arson' & `FBI Code` != 'Stolen Property')

## Visualizing from Importing

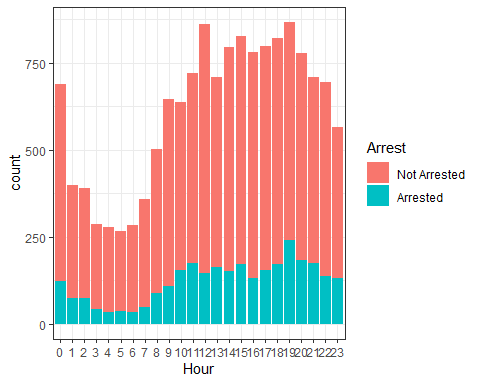
#plotting  
ggplot(new\_chicago, aes(x=Ward, fill = Arrest)) + geom\_bar() + theme\_bw()



ggplot(new\_chicago, aes(x=Hour, fill = Domestic)) + geom\_bar() + theme\_bw()

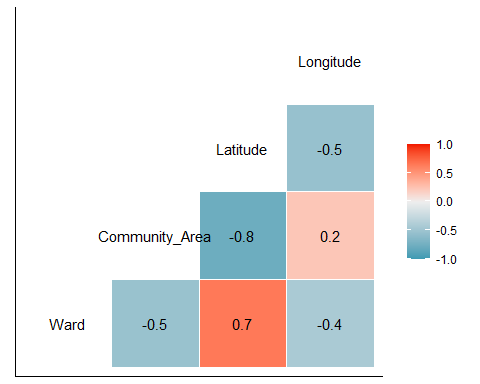


ggplot(new\_chicago, aes(x=Hour, fill = Arrest)) + geom\_bar() + theme\_bw()

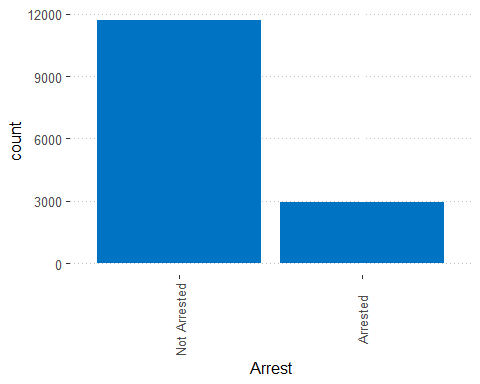


#Correlation  
ggcorr(new\_chicago, label = TRUE)

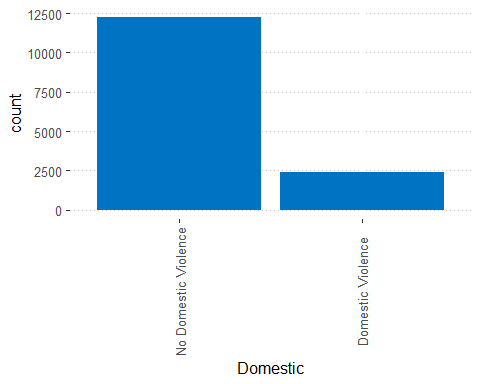
## Warning in ggcorr(new\_chicago, label = TRUE): data in column(s) 'Date', 'Block',  
## 'IUCR', 'Primary Type', 'Description', 'Location Description', 'Arrest',  
## 'Domestic', 'Beat', 'District', 'FBI Code', 'Hour', 'Months' are not numeric and  
## were ignored



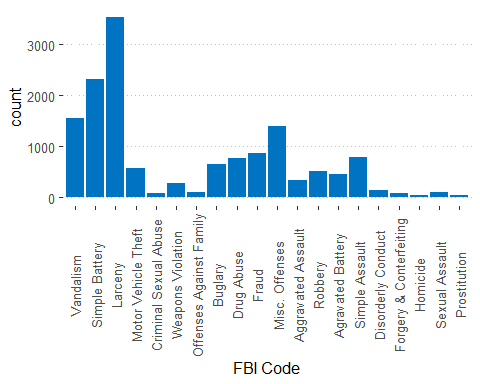
#Mapping  
ggplot(new\_chicago, aes(Arrest)) +  
 geom\_bar(fill = "#0073C2FF") +  
 theme\_pubclean() + theme(axis.text.x = element\_text(angle = 90, vjust= 0.5))



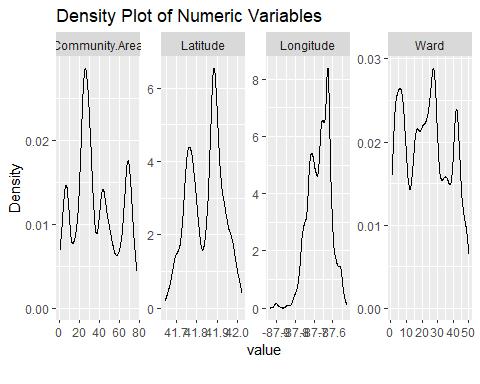
ggplot(new\_chicago, aes(Domestic)) +  
 geom\_bar(fill = "#0073C2FF") +  
 theme\_pubclean() + theme(axis.text.x = element\_text(angle = 90, vjust= 0.5))



ggplot(new\_chicago, aes(`FBI Code`)) +  
 geom\_bar(fill = "#0073C2FF") +  
 theme\_pubclean() + theme(axis.text.x = element\_text(angle = 90, vjust= 0.5))



plot\_density(new\_chicago, title = "Density Plot of Numeric Variables")



### Comments from *Visualizing and Importing*

When graphing this information, I chose different variations of Bar graphs since we are looking at categorical variables. one item that caught my eye was looking at most committed crime by FBI Code was “Larceny”. Another interesting find was that there was almost *4x* the amount of non-domestic violence and non arrest vs Domestic Violence and Arrested. Lastly, I was able to view correlation matrix to see which variables are correlated to one another. I noticed that Community Area and Latitude are highly correlated. Also, X\_Coordinate and Longitude are also 100% correlated but this is due these being the same variable but named differntly. I was also able to see that the least used data was Gambling, Embezzlement, Liquor License, Arson and Stolen Property. These will be taken out for Predictive modeling.

### Random Forest Generation for variable importance

#smaller data pool for CPU latency  
new\_chicago = sample\_frac(new\_chicago, 0.1)  
  
fit\_control = trainControl(method = "cv",   
 number = 10) #set up 10 fold cross-validation  
  
  
set.seed(1234)   
rf\_fit = train(Arrest ~.,  
 data = new\_chicago,   
 method = "ranger",   
 importance = "permutation",  
 trControl = fit\_control,  
 num.trees = 10)

### Validating variable importance

varImp(rf\_fit)

## ranger variable importance  
##   
## only 20 most important variables shown (out of 2048)  
##   
## Overall  
## `FBI Code`Drug Abuse 100.000  
## `Primary Type`NARCOTICS 63.342  
## IUCR143A 26.709  
## `Primary Type`CRIMINAL TRESPASS 25.796  
## DescriptionUNLAWFUL POSS OF HANDGUN 23.402  
## `FBI Code`Weapons Violation 21.096  
## DescriptionDOMESTIC BATTERY SIMPLE 20.649  
## DescriptionTO LAND 17.404  
## `FBI Code`Simple Battery 16.955  
## `Primary Type`WEAPONS VIOLATION 10.047  
## IUCR0860 9.773  
## IUCR0820 8.788  
## `Primary Type`INTERFERENCE WITH PUBLIC OFFICER 8.728  
## `FBI Code`Larceny 8.704  
## Latitude 8.694  
## IUCR0460 8.426  
## IUCR1330 8.226  
## IUCR0810 6.813  
## DescriptionPRO EMP HANDS NO/MIN INJURY 6.346  
## Beat0811 6.328

#### Comments from *variable importance*

We can see a breakdown of most important variables. The top variables are FBI CodeDrug Abuse at 100%, Primary TypeNARCOTICS at 59% and District015 on the low end of 2.6%.

### Data’s structure/statistics

summary(new\_chicago)

## Date Block IUCR   
## Min. :2018-01-01 00:01:00 Length:1467 Length:1467   
## 1st Qu.:2018-04-07 13:24:30 Class :character Class :character   
## Median :2018-06-30 10:00:00 Mode :character Mode :character   
## Mean :2018-07-01 17:14:28   
## 3rd Qu.:2018-09-24 17:53:00   
## Max. :2018-12-30 20:00:00   
##   
## Primary Type Description Location Description Arrest   
## Length:1467 Length:1467 Length:1467 Not Arrested:1177   
## Class :character Class :character Class :character Arrested : 290   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Domestic Beat District   
## No Domestic Violence:1218 Length:1467 Length:1467   
## Domestic Violence : 249 Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##   
## Ward Community Area FBI Code Latitude   
## Min. : 1.00 Min. : 1.00 Larceny :358 Min. :41.65   
## 1st Qu.:10.00 1st Qu.:23.00 Simple Battery:239 1st Qu.:41.77   
## Median :24.00 Median :32.00 Vandalism :145 Median :41.86   
## Mean :23.42 Mean :36.73 Misc. Offenses:131 Mean :41.84   
## 3rd Qu.:34.00 3rd Qu.:54.00 Simple Assault: 96 3rd Qu.:41.91   
## Max. :50.00 Max. :77.00 Fraud : 88 Max. :42.02   
## (Other) :410   
## Longitude Hour Months   
## Min. :-87.91 17 : 86 5 :147   
## 1st Qu.:-87.71 19 : 85 7 :142   
## Median :-87.67 12 : 84 8 :133   
## Mean :-87.67 18 : 82 6 :128   
## 3rd Qu.:-87.63 0 : 78 11 :123   
## Max. :-87.53 11 : 77 3 :120   
## (Other):975 (Other):674

str(new\_chicago)

## tibble [1,467 x 17] (S3: tbl\_df/tbl/data.frame)  
## $ Date : POSIXct[1:1467], format: "2018-11-03 14:30:00" "2018-08-01 14:03:00" ...  
## $ Block : chr [1:1467] "001XX E DELAWARE PL" "031XX W PERSHING RD" "042XX N CAMPBELL AVE" "016XX E 55TH ST" ...  
## $ IUCR : chr [1:1467] "0890" "0560" "0820" "0560" ...  
## $ Primary Type : chr [1:1467] "THEFT" "ASSAULT" "THEFT" "ASSAULT" ...  
## $ Description : chr [1:1467] "FROM BUILDING" "SIMPLE" "$500 AND UNDER" "SIMPLE" ...  
## $ Location Description: chr [1:1467] "RESTAURANT" "RESIDENCE-GARAGE" "ALLEY" "CONVENIENCE STORE" ...  
## $ Arrest : Factor w/ 2 levels "Not Arrested",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Domestic : Factor w/ 2 levels "No Domestic Violence",..: 1 1 1 1 1 2 1 1 1 1 ...  
## $ Beat : chr [1:1467] "1833" "0911" "1911" "0234" ...  
## $ District : chr [1:1467] "018" "009" "019" "002" ...  
## $ Ward : num [1:1467] 2 12 47 5 34 5 28 35 15 5 ...  
## $ Community Area : num [1:1467] 8 58 5 41 75 43 26 21 67 42 ...  
## $ FBI Code : Factor w/ 25 levels "Vandalism","Simple Battery",..: 3 15 3 15 2 11 1 11 3 11 ...  
## $ Latitude : num [1:1467] 41.9 41.8 42 41.8 41.7 ...  
## $ Longitude : num [1:1467] -87.6 -87.7 -87.7 -87.6 -87.7 ...  
## $ Hour : Factor w/ 24 levels "0","1","2","3",..: 15 15 14 12 12 9 21 19 15 23 ...  
## $ Months : Factor w/ 12 levels "1","2","3","4",..: 11 8 5 3 10 8 8 3 12 10 ...

describe(new\_chicago)

## new\_chicago   
##   
## 17 Variables 1467 Observations  
## --------------------------------------------------------------------------------  
## Date   
## n missing distinct Info   
## 1467 0 1453 1   
## Mean Gmd .05 .10   
## 2018-07-01 17:14:28 10251969 2018-01-23 16:05:00 2018-02-09 22:18:00   
## .25 .50 .75 .90   
## 2018-04-07 13:24:30 2018-06-30 10:00:00 2018-09-24 17:53:00 2018-11-26 05:42:00   
## .95   
## 2018-12-12 23:11:12   
##   
## lowest : 2018-01-01 00:01:00 2018-01-01 17:00:00 2018-01-01 18:20:00 2018-01-02 00:45:00 2018-01-02 07:40:00  
## highest: 2018-12-29 11:52:00 2018-12-29 12:30:00 2018-12-29 19:32:00 2018-12-29 21:45:00 2018-12-30 20:00:00  
## --------------------------------------------------------------------------------  
## Block   
## n missing distinct   
## 1467 0 1352   
##   
## lowest : 0000X E 102ND ST 0000X E 103RD ST 0000X E 118TH ST 0000X E 75TH ST 0000X E ADAMS ST   
## highest: 131XX S BALTIMORE AVE 132XX S BURLEY AVE 133XX S CORLISS AVE 133XX S PRAIRIE AVE 134XX S HOUSTON AVE   
## --------------------------------------------------------------------------------  
## IUCR   
## n missing distinct   
## 1467 0 134   
##   
## lowest : 0110 0264 0265 0281 0291, highest: 502R 5110 5111 5112 5131  
## --------------------------------------------------------------------------------  
## Primary Type   
## n missing distinct   
## 1467 0 23   
##   
## lowest : ASSAULT BATTERY BURGLARY CONCEALED CARRY LICENSE VIOLATION CRIM SEXUAL ASSAULT   
## highest: PUBLIC PEACE VIOLATION ROBBERY SEX OFFENSE THEFT WEAPONS VIOLATION   
## --------------------------------------------------------------------------------  
## Description   
## n missing distinct   
## 1467 0 127   
##   
## lowest : $500 AND UNDER AGG CRIM SEX ABUSE FAM MEMBER AGG PO HANDS NO/MIN INJURY AGG PRO.EMP: OTHER DANG WEAPON AGG SEX ASSLT OF CHILD FAM MBR   
## highest: UNLAWFUL USE HANDGUN VEHICLE TITLE/REG OFFENSE VEHICULAR HIJACKING VIOLATE ORDER OF PROTECTION VIOLENT OFFENDER: ANNUAL REGISTRATION  
## --------------------------------------------------------------------------------  
## Location Description   
## n missing distinct   
## 1467 0 65   
##   
## lowest : ABANDONED BUILDING AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA AIRPORT EXTERIOR - NON-SECURE AREA AIRPORT PARKING LOT AIRPORT TERMINAL LOWER LEVEL - SECURE AREA   
## highest: VACANT LOT/LAND VEHICLE-COMMERCIAL VEHICLE - DELIVERY TRUCK VEHICLE - OTHER RIDE SHARE SERVICE (E.G., UBER, LYFT) VEHICLE NON-COMMERCIAL   
## --------------------------------------------------------------------------------  
## Arrest   
## n missing distinct   
## 1467 0 2   
##   
## Value Not Arrested Arrested  
## Frequency 1177 290  
## Proportion 0.802 0.198  
## --------------------------------------------------------------------------------  
## Domestic   
## n missing distinct   
## 1467 0 2   
##   
## Value No Domestic Violence Domestic Violence  
## Frequency 1218 249  
## Proportion 0.83 0.17  
## --------------------------------------------------------------------------------  
## Beat   
## n missing distinct   
## 1467 0 268   
##   
## lowest : 0111 0112 0113 0114 0121, highest: 2531 2532 2533 2534 2535  
## --------------------------------------------------------------------------------  
## District   
## n missing distinct   
## 1467 0 22   
##   
## lowest : 001 002 003 004 005, highest: 019 020 022 024 025  
## --------------------------------------------------------------------------------  
## Ward   
## n missing distinct Info Mean Gmd .05 .10   
## 1467 0 50 0.999 23.42 16.07 2 5   
## .25 .50 .75 .90 .95   
## 10 24 34 43 46   
##   
## lowest : 1 2 3 4 5, highest: 46 47 48 49 50  
## --------------------------------------------------------------------------------  
## Community Area   
## n missing distinct Info Mean Gmd .05 .10   
## 1467 0 75 0.999 36.73 24.55 5 8   
## .25 .50 .75 .90 .95   
## 23 32 54 69 71   
##   
## lowest : 1 2 3 4 5, highest: 73 74 75 76 77  
## --------------------------------------------------------------------------------  
## FBI Code   
## n missing distinct   
## 1467 0 20   
##   
## lowest : Vandalism Simple Battery Larceny Motor Vehicle Theft Criminal Sexual Abuse   
## highest: Disorderly Conduct Forgery & Conterfeiting Homicide Sexual Assault Prostitution   
## --------------------------------------------------------------------------------  
## Latitude   
## n missing distinct Info Mean Gmd .05 .10   
## 1467 0 1424 1 41.84 0.1002 41.70 41.72   
## .25 .50 .75 .90 .95   
## 41.77 41.86 41.91 41.96 41.98   
##   
## lowest : 41.64820 41.65117 41.65143 41.65183 41.65293  
## highest: 42.01830 42.01839 42.01850 42.01855 42.02161  
## --------------------------------------------------------------------------------  
## Longitude   
## n missing distinct Info Mean Gmd .05 .10   
## 1467 0 1424 1 -87.67 0.06618 -87.77 -87.75   
## .25 .50 .75 .90 .95   
## -87.71 -87.67 -87.63 -87.60 -87.58   
##   
## lowest : -87.90523 -87.90412 -87.90350 -87.90098 -87.88361  
## highest: -87.54134 -87.53710 -87.53649 -87.53529 -87.52726  
## --------------------------------------------------------------------------------  
## Hour   
## n missing distinct   
## 1467 0 24   
##   
## lowest : 0 1 2 3 4 , highest: 19 20 21 22 23  
## --------------------------------------------------------------------------------  
## Months   
## n missing distinct   
## 1467 0 12   
##   
## lowest : 1 2 3 4 5 , highest: 8 9 10 11 12  
##   
## Value 1 2 3 4 5 6 7 8 9 10 11  
## Frequency 110 114 120 115 147 128 142 133 111 106 123  
## Proportion 0.075 0.078 0.082 0.078 0.100 0.087 0.097 0.091 0.076 0.072 0.084  
##   
## Value 12  
## Frequency 118  
## Proportion 0.080  
## --------------------------------------------------------------------------------

#### Comments from *Structure and Statistics*

When looking at three functions above, “Summary”, “Str” and “Describe”, I found *Summary* to be the most informative for the numeric data. *Str* for seeing what the data types are and the levels of the factored variables. *Describe* was the most useful for showing frequency and proportion of each variable and data in that variable.