BOSTON CRIME: Data Analysis Project IV

James Harrington

11/29/2021

Math 365: Introduction to Data Science

Instructor: Dr. Mostafa



Introduction:

The City of Boston’s police districts have been having some issues with crime for from 2015-2018. This project will mainly be observing offenses that are deemed violent or dangerous such as homicide, drug violations, shootings, robberies, and burglaries. There will be an analysis on how these crimes show up in each district as well as the statistical significance of each crime.

**Importing the data set and loading libraries:**

crime <- read.csv("crime.csv")

library(tidyverse)

library(ggplot2)

library(maps)

library(sf)

library(rgdal)

**How many cases (instances/rows) are in your dataset?**

In the crime data set there are 319,703 rows and 17 columns/variables in the data set. The data set contains 391028 missing values.

nrow(crime)

## [1] 319073

ncol(crime)

## [1] 17

names1 <- names(table(crime$OFFENSE\_CODE\_GROUP))

cbind(names1, 1:67)

keeplist <- c(-2,-6,-17,-18,-23,-33,-37,-43,-44,-47,-34,-53,-55,-61)  
keeplist

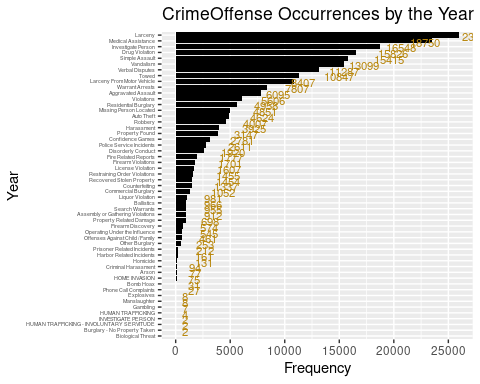
## [1] -2 -6 -17 -18 -23 -33 -37 -43 -44 -47 -34 -53 -55 -61

names2 <- names1[keeplist]

CriminalOffenses <- crime %>% filter(OFFENSE\_CODE\_GROUP %in% names2)

**Here in these two graphs, the amounts of each offense will be displayed. In the Second graph it is reordered from greatest to least.**

ggplot(CriminalOffenses, aes(x=reorder(OFFENSE\_CODE\_GROUP,OFFENSE\_CODE\_GROUP, function(x)length(x))))+  
 geom\_bar(position = "stack", fill= "black")+  
 geom\_text(stat='count', aes(label = ..count..), hjust= -1,vjust=0,size=3, color = "darkgoldenrod")+  
 labs(title= "CrimeOffense Occurrences by the Year",  
 x="Year",  
 y= "Frequency")+   
 scale\_y\_continuous(breaks=seq(0,40000,5000))+  
 coord\_flip()+  
 theme(axis.text.y= element\_text(angle=0, size=4))



Vandalism<-length(CriminalOffenses$OFFENSE\_CODE\_GROUP)  
sum(Vandalism)#How do I find out how many of the individual offenses there are? The way Im doing it here is just spitting out the integer 1.

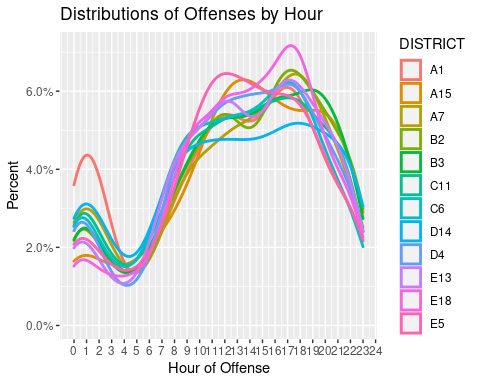
## [1] 229555

table(CriminalOffenses$OFFENSE\_CODE\_GROUP[CriminalOffenses$OFFENSE\_CODE\_GROUP=="Vandalism"])

##   
## Vandalism   
## 15415

CriminalOffenses %>%   
filter(DISTRICT== c("A1","A15","A7", "B2","B3","C11","C6","D14","D4","E13","E18","E5","E18")) %>%   
 ggplot(aes(x=HOUR, y=..density.., color= DISTRICT))+  
 geom\_density(size=1)+  
 scale\_y\_continuous(labels=scales::percent)+ #I'm not sure how to fix the percentage numbers  
 scale\_x\_continuous(breaks=seq(0, 24,1))+  
 labs(title="Distributions of Offenses by Hour",  
 y="Percent",  
 x= "Hour of Offense")

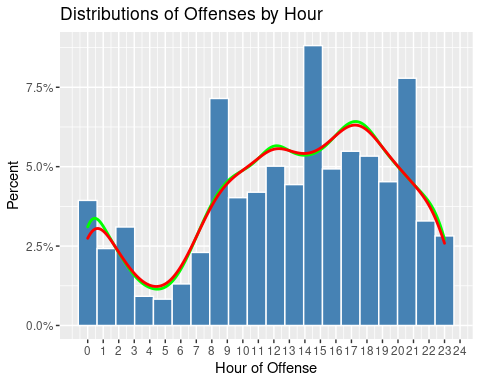
## Warning in DISTRICT == c("A1", "A15", "A7", "B2", "B3", "C11", "C6", "D14", :  
## longer object length is not a multiple of shorter object length



**According to the graph below most offenses happen at 2 pm**

CriminalOffenses%>%  
 filter(DISTRICT== c("A1","A15","A7", "B2","B3","C11","C6","D14","D4","E13","E18","E5","E18")) %>%   
 ggplot(aes(x=HOUR, y=..density..))+  
 geom\_histogram(bins = 20, color= "white", fill= "steelblue")+  
 geom\_density(color= "green", size=1, alpha=.3)+  
 geom\_density(color="red", size=1, bw=1)+ #binwidth makes the line smoother  
 scale\_y\_continuous(labels=scales::percent)+ #I'm not sure how to fix the percentage numbers  
 scale\_x\_continuous(breaks=seq(0, 24,1))+  
 labs(title="Distributions of Offenses by Hour",  
 y="Percent",  
 x= "Hour of Offense")

## Warning in DISTRICT == c("A1", "A15", "A7", "B2", "B3", "C11", "C6", "D14", :  
## longer object length is not a multiple of shorter object length



This is Boston’s B2’S distribution of homicides by hours a day within a 4-year span. The B2 District has the most total homicides from 2015-2018 out of all other police districts. The highest number of crimes happen during the night and 4pm-6pm. The least amount of crime happens during a few hours before daybreak as well as the middle of the the day (around noon time). It seems that most homicides are happening when most people are out of school or out of the normal workday hours. Maybe more police patrol should be in the B2 district during.

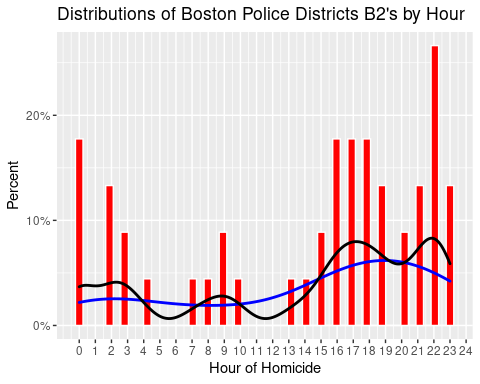
Offenses<-CriminalOffenses %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n()) %>%   
 group\_by(DISTRICT) %>%   
 summarize(Average=mean(INCIDENTS),  
 Std\_dev=sd(INCIDENTS))

## `summarise()` has grouped output by 'DISTRICT', 'YEAR'. You can override using the `.groups` argument.

The B2 District had the most incidents of average per month with a standard deviation about 165 offenses. A1 seems to be the safest district with less offenses reported. aIt also has the smallest standard deviation of offenses. From From this, I would say District A1 is the safest district to live in. However we need the exact population for these districts to have a more accurate answer on the safest areas in the city.

Homicide1 <- filter(CriminalOffenses, OFFENSE\_CODE\_GROUP == "Homicide")

Homicide1%>%  
 filter(DISTRICT== "B2") %>%   
 ggplot(aes(x=HOUR, y=..density..))+  
 geom\_histogram(bins = 50, color= "white", fill= "red")+  
 geom\_density(color= "blue", size=1, alpha=.3)+  
 geom\_density(color="black", size=1, bw=1)+ #binwidth makes the line smoother  
 scale\_y\_continuous(labels=scales::percent)+   
 scale\_x\_continuous(breaks=seq(0, 24,1))+  
 labs(title="Distributions of Boston Police Districts B2's by Hour",  
 y="Percent",  
 x= "Hour of Homicide")



table(CriminalOffenses$OFFENSE\_CODE\_GROUP[CriminalOffenses$OFFENSE\_CODE\_GROUP=="Homicide"])

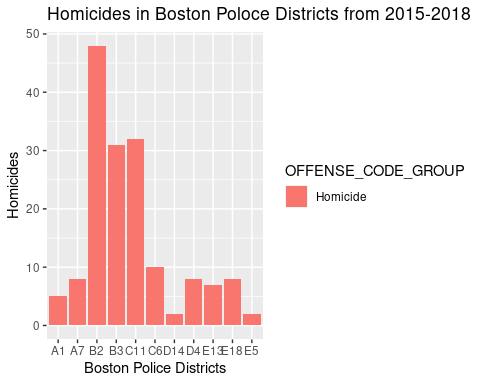
##   
## Homicide   
## 161

Homicide<-table(CriminalOffenses$OFFENSE\_CODE\_GROUP[CriminalOffenses$OFFENSE\_CODE\_GROUP=="Homicide"])  
Homicide

##   
## Homicide   
## 161

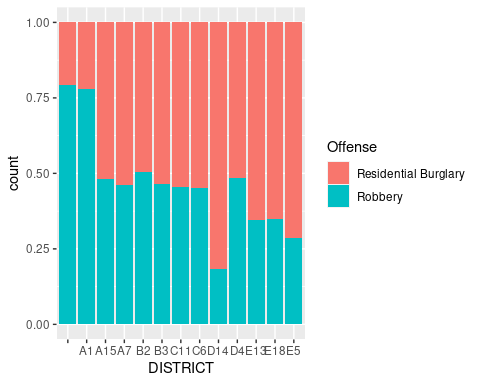
**Homicides**

ggplot(Homicide1, aes(DISTRICT, fill=OFFENSE\_CODE\_GROUP))+  
 geom\_bar()+  
 labs(title= "Homicides in Boston Poloce Districts from 2015-2018",  
 x="Boston Police Districts",  
 y="Homicides")



Moving from Homicides to theft, it is observed that districts have different percentages of residential burglary and robbery. However, most districts have a higher percentage of residential burglary than robbery. This tells us that most residents in Boston should probably look into more home security rather than means of security while out, since the districts have higher percentages of residential burglaries

crime2 <- crime %>%   
 filter(OFFENSE\_CODE\_GROUP %in% c("Robbery", "Residential Burglary")) #Use this set up for burglary  
  
crime2 %>%   
 filter(!is.na(DISTRICT), DISTRICT!="NA") %>%   
ggplot(aes(DISTRICT, fill=OFFENSE\_CODE\_GROUP))+  
 geom\_bar(position = "fill", show.legend = T)+  
 scale\_fill\_discrete(name="Offense")



**What district has the most homicides?**

B2 Has the most homicides. It also has the highest average per month in drug violations. The question on the significance of drug violations by year, district, and month is answered. The null hypothesis was rejected most of the p values were less than .05, proving that the amount of drug violations is dependent on these variables.

Deaths<-CriminalOffenses %>%   
 filter(OFFENSE\_CODE\_GROUP == "Homicide") %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n()) %>%   
 group\_by(DISTRICT) %>%   
 summarize(Average=mean(INCIDENTS),  
 Std\_dev=sd(INCIDENTS))

## `summarise()` has grouped output by 'DISTRICT', 'YEAR'. You can override using the `.groups` argument.

drugviolations<-CriminalOffenses %>%   
 filter(OFFENSE\_CODE\_GROUP == "Drug Violation") %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n()) %>%   
 group\_by(DISTRICT) %>%   
 summarize(Average=mean(INCIDENTS),  
 Std\_dev=sd(INCIDENTS))

## `summarise()` has grouped output by 'DISTRICT', 'YEAR'. You can override using the `.groups` argument.

**What district has the most robberies?**

District B2 had the highest average of 21.5 robberies per month with a standard deviation of about 8 robberies

CriminalOffenses %>%   
 filter(OFFENSE\_CODE\_GROUP == "Robbery") %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n()) %>%   
 group\_by(DISTRICT) %>%   
 summarize(Average=mean(INCIDENTS),  
 Std\_dev=sd(INCIDENTS))

## `summarise()` has grouped output by 'DISTRICT', 'YEAR'. You can override using the `.groups` argument.

## # A tibble: 13 × 3  
## DISTRICT Average Std\_dev  
## <chr> <dbl> <dbl>  
## 1 "" 1.73 1.10  
## 2 "A1" 16.5 6.54  
## 3 "A15" 2.4 1.70  
## 4 "A7" 5.8 2.86  
## 5 "B2" 21.5 7.82  
## 6 "B3" 13.6 5.32  
## 7 "C11" 16.5 6.97  
## 8 "C6" 6.58 3.37  
## 9 "D14" 4.38 2.38  
## 10 "D4" 15.6 5.20  
## 11 "E13" 6.58 3.09  
## 12 "E18" 4.25 2.10  
## 13 "E5" 2.67 1.57

**Does the district with the highest average of burglaries also have the highest average per month or is it a different district?**

Yes, B2 has the highest average of 21.2 residential burglaries per month as well as robberies and a standard deviation of about 8 burglaries. If there was some economic data on income levels, regressions and correlations based on economic levels and robberies/ burglaries could be done.

CriminalOffenses %>%   
 filter(OFFENSE\_CODE\_GROUP == "Residential Burglary") %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n()) %>%   
 group\_by(DISTRICT) %>%   
 summarize(Average=mean(INCIDENTS),  
 Std\_dev=sd(INCIDENTS))

## `summarise()` has grouped output by 'DISTRICT', 'YEAR'. You can override using the `.groups` argument.

## # A tibble: 13 × 3  
## DISTRICT Average Std\_dev  
## <chr> <dbl> <dbl>  
## 1 "" 1 0   
## 2 "A1" 4.77 2.34  
## 3 "A15" 2.68 1.84  
## 4 "A7" 6.97 3.53  
## 5 "B2" 21.2 8.09  
## 6 "B3" 15.7 7.99  
## 7 "C11" 19.7 7.69  
## 8 "C6" 8.18 3.55  
## 9 "D14" 19.7 11.8   
## 10 "D4" 16.7 8.73  
## 11 "E13" 12.1 5.66  
## 12 "E18" 8.15 4.34  
## 13 "E5" 6.29 3.42

There is a weak correlation between robberies and shootings robberies, but a moderate correlation between homicides and shootings.  
  
```r  
cor(crime$OFFENSE\_CODE\_GROUP=="Homicide", crime$SHOOTING=="Y")

## [1] 0.2980164

cor(crime$OFFENSE\_CODE\_GROUP=="Robbery", crime$SHOOTING=="Y")

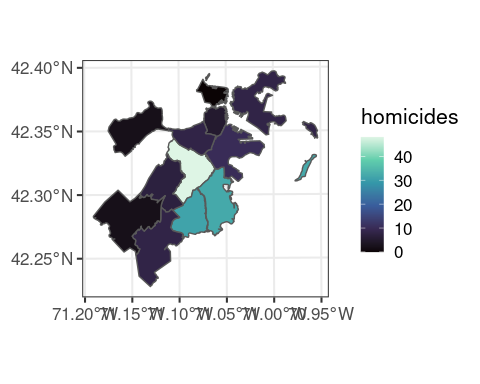
## [1] 0.001967356

boston <- st\_read(dsn=".",layer="Police\_Districts")

## Reading layer `Police\_Districts' from data source `/cloud/project' using driver `ESRI Shapefile'  
## Simple feature collection with 12 features and 8 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: 739826.9 ymin: 2908285 xmax: 804052.5 ymax: 2970073  
## Projected CRS: NAD83 / Massachusetts Mainland (ftUS)

**Here is a map of the Boston Police Districts color coated by amounts of homicides.**

homicideby.district<- Homicide1 %>%   
 group\_by(DISTRICT) %>%   
 summarize(homicides= n())  
  
boston<- merge(boston, homicideby.district, by="DISTRICT",all.x= TRUE ) #all.x is another way to do left\_join  
boston<- boston %>%   
 replace\_na(list(homicides=0))  
  
#ggplot(boston) +geom\_sf()  
  
ggplot(boston, aes()) +  
 geom\_sf(aes(fill = homicides)) +  
 theme\_bw(base\_size = 16) +  
 scale\_fill\_viridis\_c(option = "G")



Map

Description automatically generated

Conclusion:

During the analysis of this project, it was found the offenses that more homicides, burglaries, robberies, and drug violations occurred in District B2. The trends in times of crime were similar among districts even if the amount and distribution of crimes are different. There is a correlation between shootings and homicides. As far as drug violations, there is a dependency with districts and year.

This analysis supports the takeaway that more consideration for crime remediation is needed in the B2 section of Boston. Diminishing overall crime in the B2 District will make the largest difference on Boston’s overall crime rate.

Personal Solutions:

Assuming that economic earnings may be why there’s an issue with crime in District B2. It is believed that the city of Boston should increase police presence and security surveillance cameras in the areas most dangerous. Boston also could create educational programs for children in District B12 while creating tax business incentives to bring jobs back to the district. There are tradeoffs between these two options. Cameras are cheap, but the latter option is more money and builds up the skill and opportunity for the residents.