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 we deem 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.

Loading necessary packages:

library(readr)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ dplyr 1.0.7  
## ✓ tibble 3.1.5 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.4 ✓ forcats 0.5.1  
## ✓ purrr 0.3.4

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

library(ggplot2)  
library(maps)

##   
## Attaching package: 'maps'

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

library(sf)

## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1

library(rgdal)

## Loading required package: sp

## Please note that rgdal will be retired by the end of 2023,  
## plan transition to sf/stars/terra functions using GDAL and PROJ  
## at your earliest convenience.  
##   
## rgdal: version: 1.5-27, (SVN revision 1148)  
## Geospatial Data Abstraction Library extensions to R successfully loaded  
## Loaded GDAL runtime: GDAL 3.0.4, released 2020/01/28  
## Path to GDAL shared files: /usr/share/gdal  
## GDAL binary built with GEOS: TRUE   
## Loaded PROJ runtime: Rel. 6.3.1, February 10th, 2020, [PJ\_VERSION: 631]  
## Path to PROJ shared files: /usr/share/proj  
## Linking to sp version:1.4-5  
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,  
## use options("rgdal\_show\_exportToProj4\_warnings"="none") before loading sp or rgdal.

library(dplyr)

Importing the data set:

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

Basic Data Set Questions “Crime”

1. 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

#c. Use the function head(dataset, 10) to report the first 10 instances in your data.  
 head(crime, 10)

## INCIDENT\_NUMBER OFFENSE\_CODE OFFENSE\_CODE\_GROUP  
## 1 I182070945 619 Larceny  
## 2 I182070943 1402 Vandalism  
## 3 I182070941 3410 Towed  
## 4 I182070940 3114 Investigate Property  
## 5 I182070938 3114 Investigate Property  
## 6 I182070936 3820 Motor Vehicle Accident Response  
## 7 I182070933 724 Auto Theft  
## 8 I182070932 3301 Verbal Disputes  
## 9 I182070931 301 Robbery  
## 10 I182070929 3301 Verbal Disputes  
## OFFENSE\_DESCRIPTION DISTRICT REPORTING\_AREA SHOOTING  
## 1 LARCENY ALL OTHERS D14 808   
## 2 VANDALISM C11 347   
## 3 TOWED MOTOR VEHICLE D4 151   
## 4 INVESTIGATE PROPERTY D4 272   
## 5 INVESTIGATE PROPERTY B3 421   
## 6 M/V ACCIDENT INVOLVING PEDESTRIAN - INJURY C11 398   
## 7 AUTO THEFT B2 330   
## 8 VERBAL DISPUTE B2 584   
## 9 ROBBERY - STREET C6 177   
## 10 VERBAL DISPUTE C11 364   
## OCCURRED\_ON\_DATE YEAR MONTH DAY\_OF\_WEEK HOUR UCR\_PART STREET  
## 1 9/2/2018 13:00 2018 9 Sunday 13 Part One LINCOLN ST  
## 2 8/21/2018 0:00 2018 8 Tuesday 0 Part Two HECLA ST  
## 3 9/3/2018 19:27 2018 9 Monday 19 Part Three CAZENOVE ST  
## 4 9/3/2018 21:16 2018 9 Monday 21 Part Three NEWCOMB ST  
## 5 9/3/2018 21:05 2018 9 Monday 21 Part Three DELHI ST  
## 6 9/3/2018 21:09 2018 9 Monday 21 Part Three TALBOT AVE  
## 7 9/3/2018 21:25 2018 9 Monday 21 Part One NORMANDY ST  
## 8 9/3/2018 20:39 2018 9 Monday 20 Part Three LAWN ST  
## 9 9/3/2018 20:48 2018 9 Monday 20 Part One MASSACHUSETTS AVE  
## 10 9/3/2018 20:38 2018 9 Monday 20 Part Three LESLIE ST  
## Lat Long Location  
## 1 42.35779 -71.13937 (42.35779134, -71.13937053)  
## 2 42.30682 -71.06030 (42.30682138, -71.06030035)  
## 3 42.34659 -71.07243 (42.34658879, -71.07242943)  
## 4 42.33418 -71.07866 (42.33418175, -71.07866441)  
## 5 42.27537 -71.09036 (42.27536542, -71.09036101)  
## 6 42.29020 -71.07159 (42.29019621, -71.07159012)  
## 7 42.30607 -71.08273 (42.30607218, -71.08273260)  
## 8 42.32702 -71.10555 (42.32701648, -71.10555088)  
## 9 42.33152 -71.07085 (42.33152148, -71.07085307)  
## 10 42.29515 -71.05861 (42.29514664, -71.05860832)

sum(is.na(crime))

## [1] 60248

colSums(is.na(crime))

## INCIDENT\_NUMBER OFFENSE\_CODE OFFENSE\_CODE\_GROUP OFFENSE\_DESCRIPTION   
## 0 0 0 0   
## DISTRICT REPORTING\_AREA SHOOTING OCCURRED\_ON\_DATE   
## 0 20250 0 0   
## YEAR MONTH DAY\_OF\_WEEK HOUR   
## 0 0 0 0   
## UCR\_PART STREET Lat Long   
## 0 0 19999 19999   
## Location   
## 0

1. Identify at least two potential research questions that you plan to answer using your dataset. It is strongly recommended that you define at least 1 question that can be answered using data visualization and correlations, and at least 1 question that needs some sort of predictive modeling (regression/classification)\_:

#What district had the most homicides? “B2” #The independent variable is district and the dependent variable is homicide . #Does that district have the largest amount of homicides every year in the data set #what is the statistical significance of the day the week for drug violations?

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

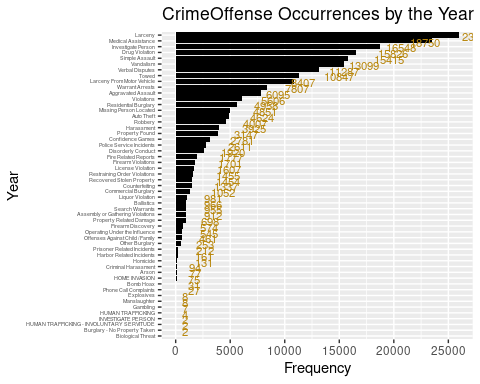
## [1] -2 -6 -17 -18 -23 -33 -37 -43 -44 -47 -34 -53 -55 -61

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.

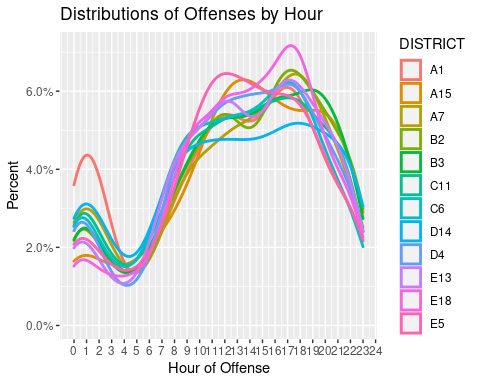
## [1] 229555

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")

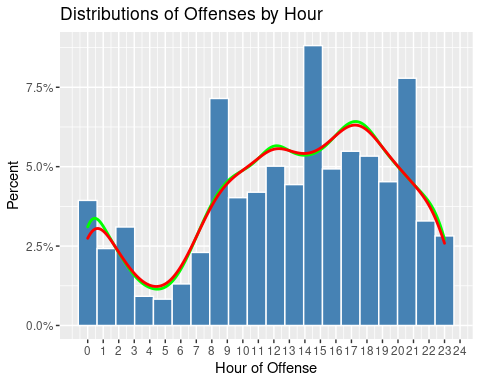
## Warning in DISTRICT == c("A1", "A15", "A7", "B2", "B3", "C11", "C6", "D14", :  
## longer object length is not a multiple of shorter object length



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 Police District 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 Boston Police Districts. We see from the visualization that the highest number of crimes happen in the hours of the night and 4pm-6pm the least amount of crime happens during a few hours before day break as well as the middle of the the day (around noon time). It seems that most homicides are happening when the majority of people are out of school or out of the normal workday hours. Maybe more police patrol should be in the B2 district during.

Since our City of Boston has pushed for safer living environments we will propose that more survelience cameras and street lights be put on the streets of the B2 District with the highest numbers of homicides (view second graph).

Statistics Lets use some statistical methods to find out the standard deviation of offences a month in different district by the hour

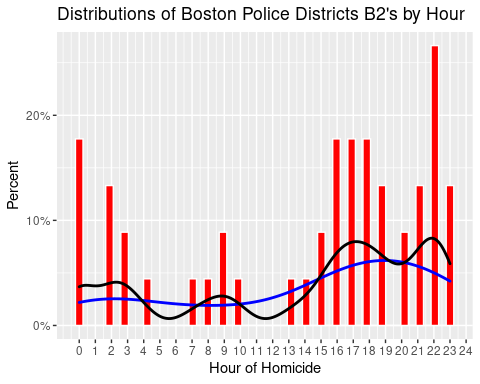
Offenses<-CriminalOffenses %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n()) %>%   
 group\_by(DISTRICT) %>%   
 summarize(Average=mean(INCIDENTS),  
 Std\_dev=sd(INCIDENTS))

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

The B2 District had the most incidents o 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"])

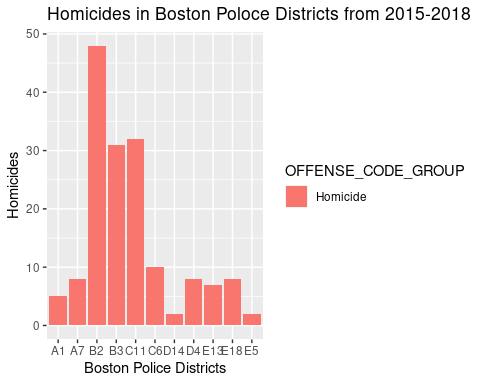
##   
## Homicide   
## 161

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

##   
## Homicide   
## 161

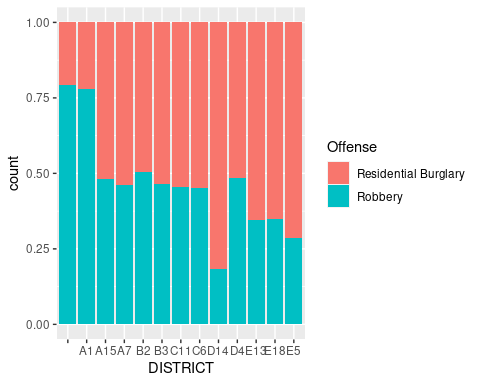
Homicides

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. We observe 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")



Statistics: As we know B2 from previous charts that B2 Has the most homicides. It also has the highest average per moth 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 ammount of drug violations are 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.

Linear Regression of Drug Violations

lr<-CriminalOffenses %>%   
 filter(OFFENSE\_CODE\_GROUP=="Drug Violation") %>%   
 group\_by(DISTRICT, YEAR, MONTH) %>%   
 summarize(INCIDENTS= n())

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

model1<-lm(INCIDENTS ~DISTRICT+YEAR+MONTH, data=lr)   
summary(model1)

##   
## Call:  
## lm(formula = INCIDENTS ~ DISTRICT + YEAR + MONTH, data = lr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -46.390 -7.047 -0.313 6.025 47.883   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7664.2990 1198.7437 6.394 3.79e-10 \*\*\*  
## DISTRICTA1 44.4316 3.1557 14.080 < 2e-16 \*\*\*  
## DISTRICTA15 0.7282 3.1739 0.229 0.818638   
## DISTRICTA7 15.0566 3.1557 4.771 2.42e-06 \*\*\*  
## DISTRICTB2 49.2566 3.1557 15.609 < 2e-16 \*\*\*  
## DISTRICTB3 30.2316 3.1557 9.580 < 2e-16 \*\*\*  
## DISTRICTC11 48.9846 3.1739 15.434 < 2e-16 \*\*\*  
## DISTRICTC6 37.7566 3.1557 11.964 < 2e-16 \*\*\*  
## DISTRICTD14 9.8066 3.1557 3.108 0.001996 \*\*   
## DISTRICTD4 44.3179 3.1739 13.963 < 2e-16 \*\*\*  
## DISTRICTE13 20.1128 3.1739 6.337 5.34e-10 \*\*\*  
## DISTRICTE18 12.0316 3.1557 3.813 0.000155 \*\*\*  
## DISTRICTE5 10.1566 3.1557 3.218 0.001375 \*\*   
## YEAR -3.7943 0.5941 -6.386 3.96e-10 \*\*\*  
## MONTH -0.8034 0.1830 -4.391 1.38e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.64 on 488 degrees of freedom  
## Multiple R-squared: 0.6753, Adjusted R-squared: 0.666   
## F-statistic: 72.49 on 14 and 488 DF, p-value: < 2.2e-16

District B2 had the highest average of 21.5 robberies per month with a standard deviation 0f 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, 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 we could do some regressions and correlations based on economic levels and robberies/ burglaries. I

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

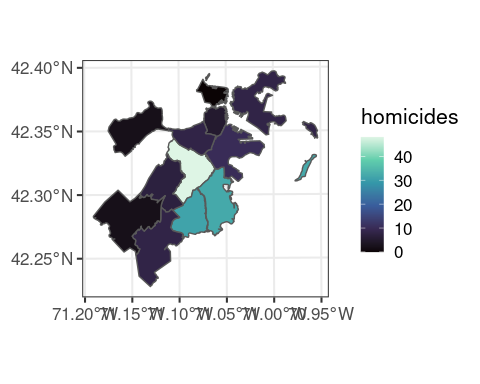
Homicide<-table(CriminalOffensesOFFENSE\_CODE\_GROUP==“Homicide”]) Homicide

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.

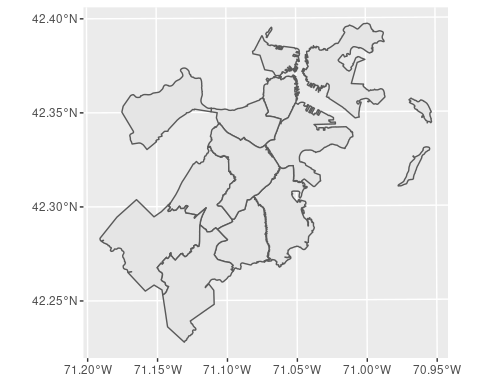
homicideby.district<- Homicide1 %>%   
 group\_by(DISTRICT) %>%   
 summarize(homicides= n())  
  
boston<- merge(boston, homicideby.district, by="DISTRICT",all.x= TRUE ) #all.x is another way to do left\_join  
boston<- boston %>%   
 replace\_na(list(homicides=0))  
  
#ggplot(boston) +geom\_sf() 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



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)

ggplot(boston) +geom\_sf()



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. We learned that there is a correlation between shootings and homicides. As far as drug violations, there is a dependency with districts and year.

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 camera in the areas most dangerous. Boston should create educational programs for children in District B12 while creating tax business incentives to bring jobs back to the district. Diminishing overall crime in the worst District will make the largest difference in Boston’s overall crime rate.