Final Project: Crime in Boston

Li

12/15/2017

Introduction

In this project, I wish to investigate the crime dataset from City of Boston database, which documents all the crime incidents reported to police that dated from 2015 to present. The main idea is to provide visualizations to many aspects of serious crime incidents happened within the boston area, and if give suggestions to general public regarding the public safety around city of boston.

Data Import and Clean

First I shall extract all the crime reports from two separate csv files, in which crime0.csv has all the records from July 2012 to Aug 2015, and crime.csv contains all the crime reports from Aug 2015 to Dec 2017.

Note that many columns of the data are not quite statistically meaningful, therefore a subset of columns are selected from the original raw dataset for easier computation

Since the old dataset (2012-2015) is from the legacy system, thus some extra data cleanings are required.

After all the data cleaning, we combine these two data sets to a complete set

```
# prepare packages
if(!require("pacman")) install.packages("pacman")
pacman::p_load("dplyr", "ggmap", "shiny", "tidyverse", "tidyr", "ggplot2", "stringi")

#Contains crime records from July 2012 to Aug 2015
Raw_BosCrime1215 <- read.csv("crime0.csv")
#Contains crime records from Aug 2015 to Dec 2017
Raw_BosCrime1517 <- read.csv("crime.csv")

BosCrime1215 <- subset(Raw_BosCrime1215, select=c("INCIDENT_TYPE_DESCRIPTION", "REPTDISTRICT", "Year", "
BosCrime1517 <- subset(Raw_BosCrime1517, select=c("OFFENSE_CODE_GROUP", "DISTRICT",  "YEAR", "MONTH", "

#Standardize data from two data set
##Convert all character columns to lower case
BosCrime1215 <- mutate_all(BosCrime1215, .funs=tolower)
BosCrime1517 <- mutate_all(BosCrime1517, .funs=tolower)

##Rename columns to same
BosCrime1215 <- dplyr::rename(BosCrime1215, Offense_Type = INCIDENT_TYPE_DESCRIPTION, District = REPTDIS

BosCrime1517 <- dplyr::rename(BosCrime1517, Offense_Type = OFFENSE_CODE_GROUP, District = DISTRICT, Day_


##For BosCrime 2012 - 2015, we need to transform Location to two separate columns, Lat and Long
BosCrime1215 <- BosCrime1215 %>% separate(Location, into = c("Lat", "Long"), sep = ",")
BosCrime1215$Lat <- substring(BosCrime1215$Lat, 2)
BosCrime1215$Long <- stringi::stri_sub(BosCrime1215$Long, 1, -2)

head(BosCrime1215)
```

```
##           Offense_Type District Year Month Day_Week      Street
## 1 residential burglary       d4 2012     7   sunday aberdeen st
## 2   aggravated assault       b2 2012     7   sunday   howard av
## 3              robbery       d4 2012     7   sunday   jersey st
## 4  commercial burglary       b2 2012     7   sunday columbia rd
## 5              robbery      e18 2012     7   sunday  collins st
## 6              robbery      c11 2012     7   sunday   sydney st
##           Lat          Long
## 1 42.34638135  -71.10379454
## 2 42.31684135  -71.07458456
## 3 42.34284135  -71.09698955
## 4  42.3164411  -71.06582908
## 5 42.27051636  -71.11989955
## 6 42.31328183   -71.0530059
```

```
head(BosCrime1517)
```

```
##                     Offense_Type District Year Month Day_Week
## 1              medical assistance      e13 2017    12 thursday
## 2                 verbal disputes      e13 2017    12 thursday
## 3      larceny from motor vehicle       b2 2017    12 thursday
## 4                  counterfeiting      d14 2017    12 thursday
## 5                 verbal disputes       b3 2017    12 thursday
## 6 motor vehicle accident response      e18 2017    12 thursday
##                Street         Lat          Long
## 1           bourne st 42.28875724 -71.11293214
## 2            heath st 42.32624204 -71.10317575
## 3           cheney st 42.30858147  -71.0835504
## 4        n harvard st 42.36402978  -71.1286334
## 5             ames st 42.28956988 -71.08510501
## 6 american legion hwy 42.28314714 -71.11374152
```

```
##Combine the two
BosCrime1217 <- rbind(BosCrime1215, BosCrime1517)
```

Data Filtering and Combining

With the complete data frame obtained, we should notice that the crime reports do contain a wide range of crimes, and they vary greatly in their significances. Thus, for this project we will focus particularly on the serious crime incidents, which include 4 categories: homicide, robbery, aggravated assault, and residential burglary. Also, since there are crime incidents labeled with (0,0) location, we will disgard all those data as irrelevant.

```
#Focus only on violent crimes
#Restrict to relevant Longitude and Latitude data in Boston (Performed only once, and can be commented

bos_violent_crimes <- filter(BosCrime1217,
                             Offense_Type == 'homicide' |
                             Offense_Type == 'robbery' |
                             Offense_Type == 'aggravated assault' |
                             Offense_Type == 'residential burglary',
                             -71.19292 <= as.numeric(Long) & as.numeric(Long) <= -70.89677,
                             42.13287 <= as.numeric(Lat) & as.numeric(Lat) <=  42.49504
                             )
```

```
head(bos_violent_crimes)
```

```
##           Offense_Type District Year Month Day_Week      Street
## 1 residential burglary       d4 2012     7   sunday aberdeen st
## 2   aggravated assault       b2 2012     7   sunday   howard av
## 3              robbery       d4 2012     7   sunday   jersey st
## 4              robbery      e18 2012     7   sunday  collins st
## 5              robbery      c11 2012     7   sunday   sydney st
## 6              robbery       b2 2012     7   sunday   regent st
##           Lat          Long
## 1 42.34638135  -71.10379454
## 2 42.31684135  -71.07458456
## 3 42.34284135  -71.09698955
## 4 42.27051636  -71.11989955
## 5 42.31328183   -71.0530059
## 6 42.32425136  -71.08620956
```

```
#Write all the serious crimes into a csv for later use.
write.csv(bos_violent_crimes, "vio_crime.csv")

#Save the Violent Crime Data, so need to read huge original csv file.
bos_violent_crimes <- read.csv("vio_crime.csv")
```

The result after combining is following

```
head(bos_violent_crimes)
```

```
##   X         Offense_Type District Year Month Day_Week      Street      Lat
## 1 1 residential burglary       d4 2012     7   sunday aberdeen st 42.34638
## 2 2   aggravated assault       b2 2012     7   sunday   howard av 42.31684
## 3 3              robbery       d4 2012     7   sunday   jersey st 42.34284
## 4 4              robbery      e18 2012     7   sunday  collins st 42.27052
## 5 5              robbery      c11 2012     7   sunday   sydney st 42.31328
## 6 6              robbery       b2 2012     7   sunday   regent st 42.32425
##        Long
## 1 -71.10379
## 2 -71.07458
## 3 -71.09699
## 4 -71.11990
## 5 -71.05301
## 6 -71.08621
```

Analysis

General summary

In the following section, I will present some elementary summaries from the `bos_violent_crimes`.

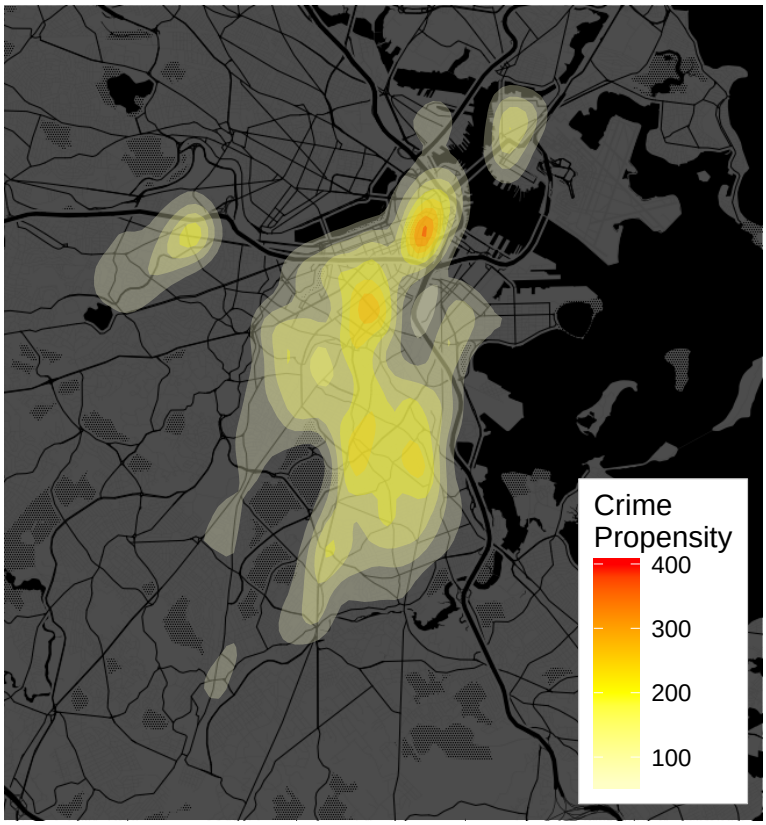
```
##Output general summary of bos_violent_crimes
summary(bos_violent_crimes)
```

```
##        X                       Offense_Type      District
##  Min.   :    1   aggravated assault  :11340   b2     :5660
##  1st Qu.: 7905   homicide            :  258   c11    :4663
##  Median :15809   residential burglary:11634   b3     :3802
```

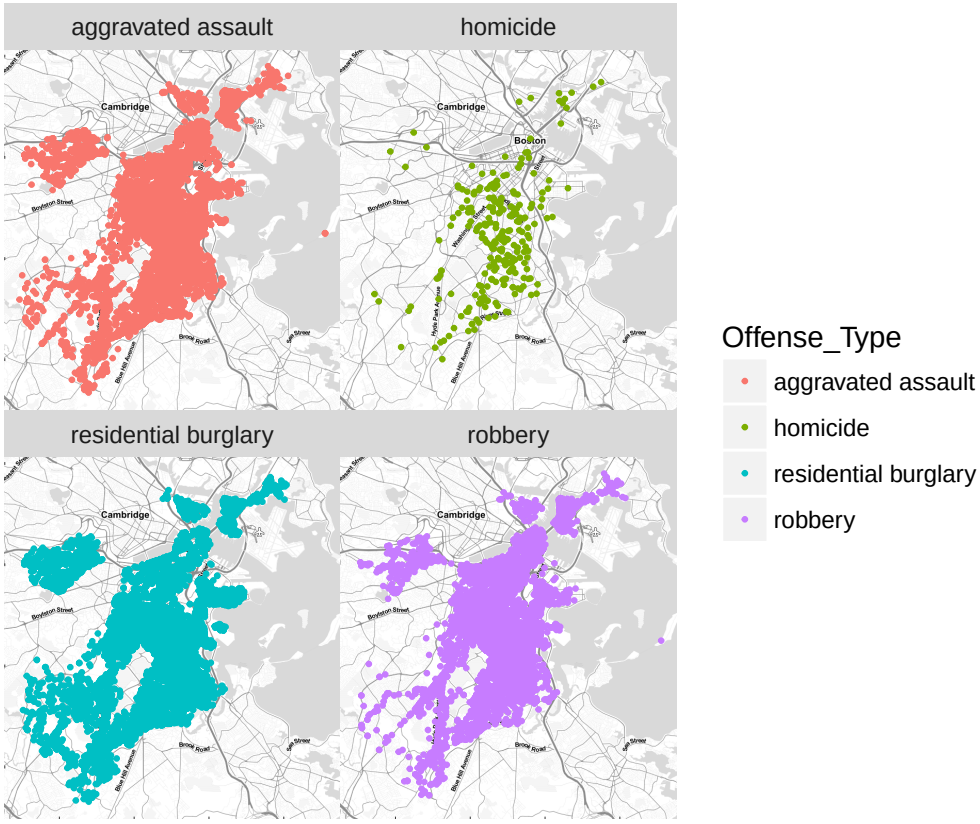
```
##  Mean   :15809   robbery            : 8385   d4     :3769
##  3rd Qu.:23713                               a1     :2881
##  Max.   :31617                               d14    :2569
##                                              (Other):8273
##       Year          Month            Day_Week              Street
##  Min.   :2012   Min.   : 1.000   friday   :4819   washington st: 1283
##  1st Qu.:2013   1st Qu.: 4.000   monday   :4439   tremont st   :  587
##  Median :2015   Median : 7.000   saturday :4529   boylston st  :  553
##  Mean   :2015   Mean   : 6.997   sunday   :4420   blue hill av :  387
##  3rd Qu.:2016   3rd Qu.:10.000   thursday :4484   blue hill ave:  385
##  Max.   :2017   Max.   :12.000   tuesday  :4520   centre st    :  346
##                                  wednesday:4406   (Other)      :28076
##       Lat             Long
##  Min.   :42.23   Min.   :-71.18
##  1st Qu.:42.30   1st Qu.:-71.10
##  Median :42.32   Median :-71.08
##  Mean   :42.32   Mean   :-71.08
##  3rd Qu.:42.35   3rd Qu.:-71.06
##  Max.   :42.40   Max.   :-70.97
##
```

```
# Plot the Lat and Long on the Boston Map, facet based on Offense Type
cimre_summary_map <- qmplot(Long, Lat, data = bos_violent_crimes, maptype = "toner-lite", color = Offens

# Plot the Density (Propensity) of crime incidents on map
crime_propensity_plot <- qmplot(Long, Lat, data = bos_violent_crimes, geom = "blank", zoom = 12, maptype
  stat_density_2d(aes(fill = ..level..), geom = "polygon", alpha = .3, color = NA) +
  scale_fill_gradient2("Crime\nPropensity", low = "white", mid = "yellow", high = "red", midpoint = 200

crime_propensity_plot
```

Crime Density Graph (June 2012 − Dec 2017)




```
cimre_summary_map
```



Noting there are actually some interesting in the output of summary(bos_violent_crimes): - The predominant offense types are the aggravated assault and residential burglary - Out of all the polic districts, the b2 has the most occurances of serious crimes 5660, follows by c11 and b3 - Washington Street, may due to its extended length, is one of the most dangerous streets in Boston

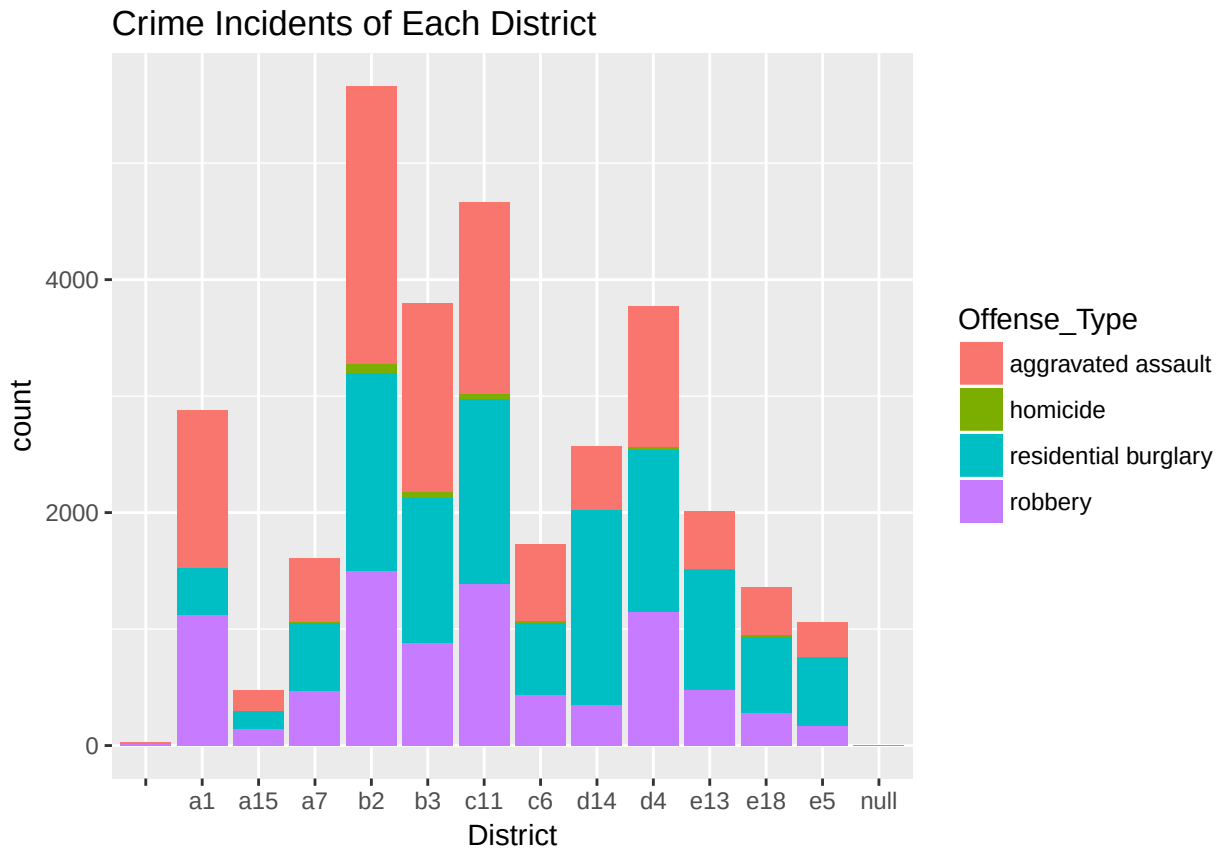
Also, by mapping each Longitude and Latitude to the Crime Summary Map, we see a general trend that as we move toward the down town area, the crime incidents get more frequent.

With the density graph, we can further more identify that the two center of the serious crimes are located near the downtown Boston area and near the Boston Medical School/North Eastern Univeristy.

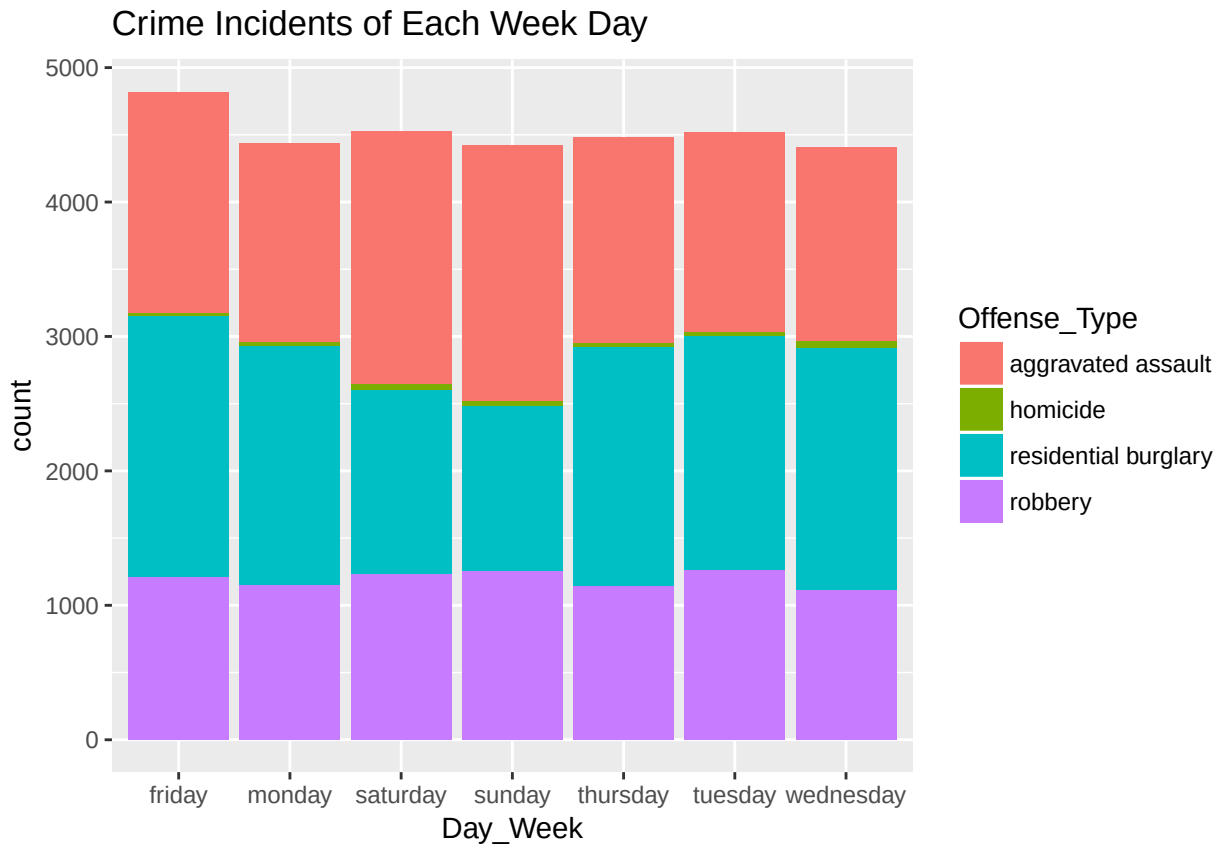
More General plots

In this section, we will investigate more graphs from the `bos_violent_crimes`

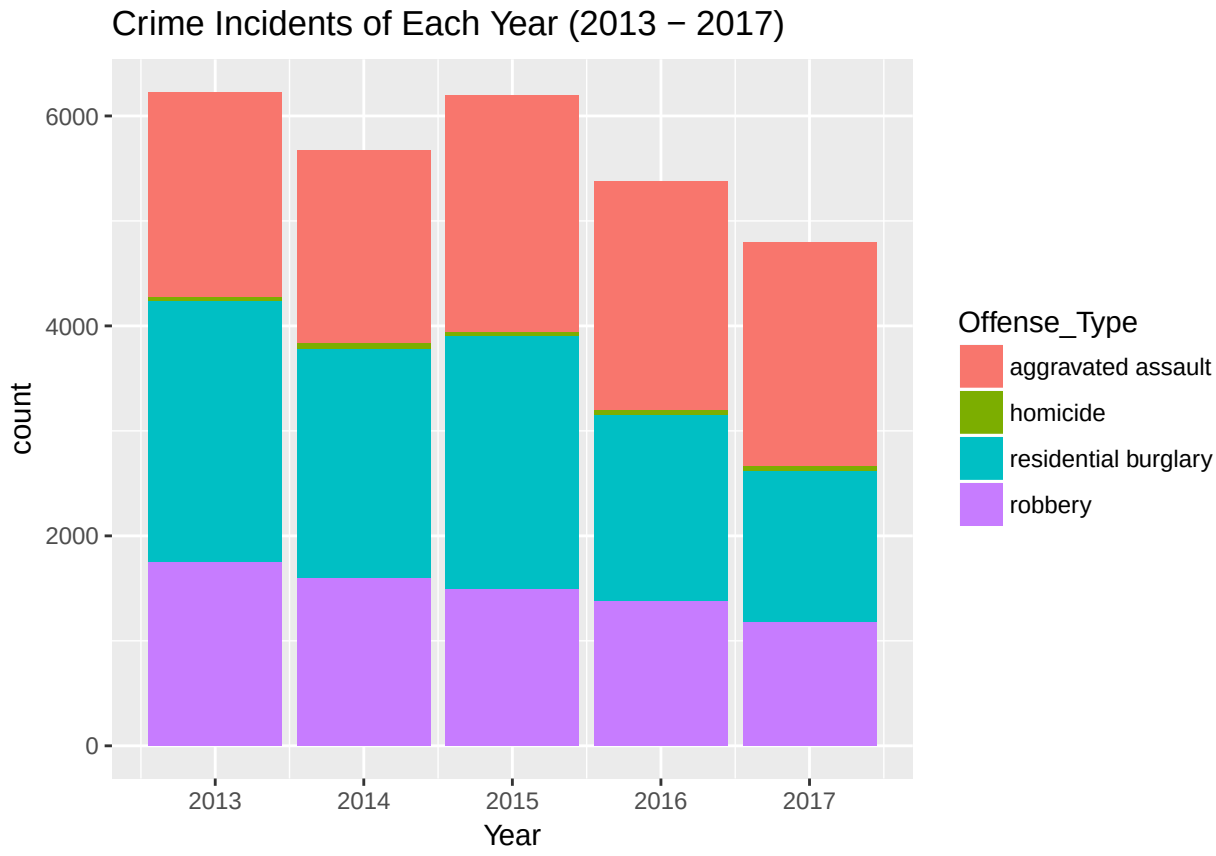
```
#Since we do not have complete 2012 data, therefore, we will exclude 2012 from some analysis
bos_violent_crimes_1317 <- filter(bos_violent_crimes, Year != '2012')
#Plot the crime based on districts
district_summary_plot <- ggplot(data = bos_violent_crimes, aes(District)) + geom_bar(aes(fill=Offense_Ty
#Plot the crime based on day of week
week_day_summary_plot <- ggplot(data = bos_violent_crimes, aes(Day_Week)) + geom_bar(aes(fill=Offense_Ty
#Plot the crime based on Year (Since we have an incomplete data from 2012, thus we exclude 2012)
year_summary_plot <- ggplot(data = bos_violent_crimes_1317, aes(Year)) + geom_bar(aes(fill=Offense_Type
#Plot the crime based on Month
month_summary_plot <- ggplot(data = bos_violent_crimes, aes(Month)) + geom_bar(aes(fill=Offense_Type)) +

district_summary_plot
```



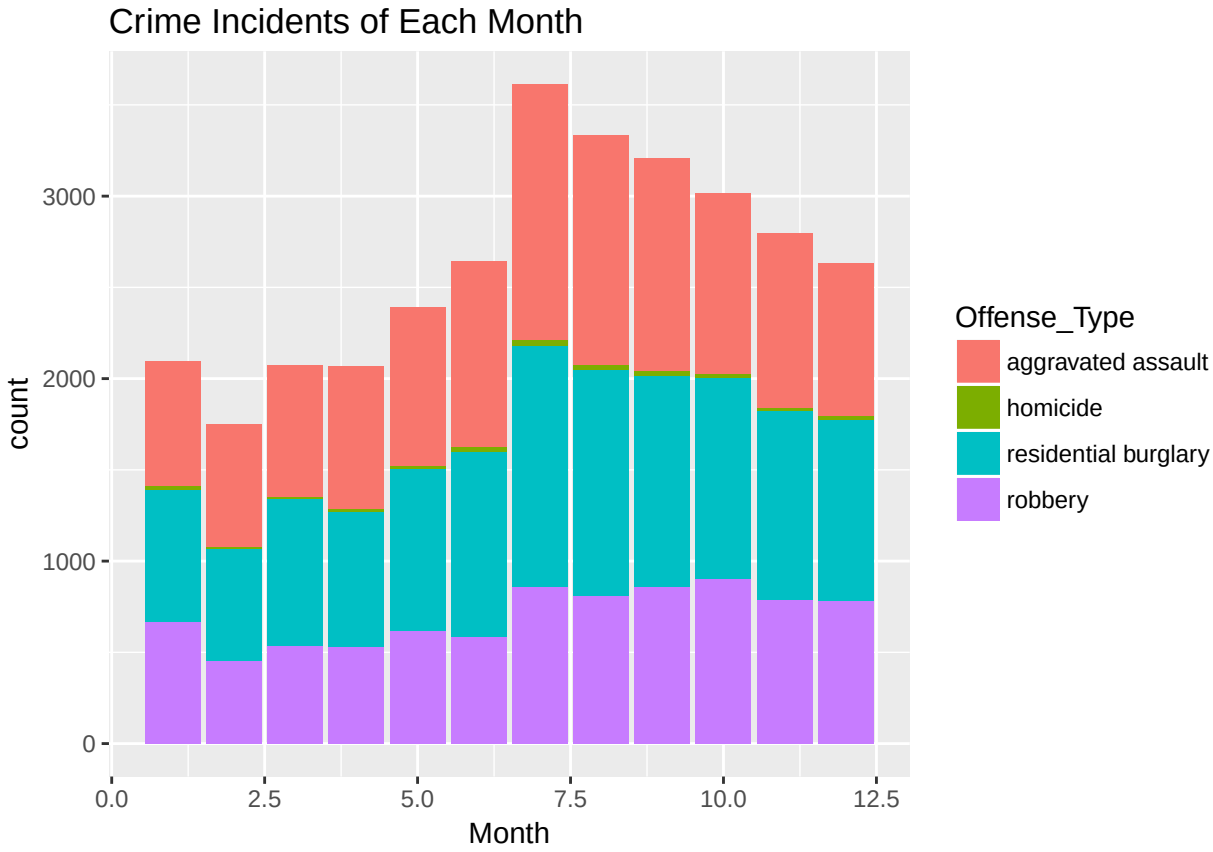

```
week_day_summary_plot
```




```
year_summary_plot
```




```
month_summary_plot
```



From the plots above, we can observe following: - Indeed that b2 district (around North Eastern University) has the worst public safety - Also the occurances of serious crime do not have strong relationship with the day of the week (a slight increase in the Friday) - The overall occurances of serious crime is in a decreasing trend as year moving from 2013 to 2017. - Lastly, there is definitely a seasonal fluctuaton to the occurances of serious crimes. As the temperature increases in the summer, there are more crimes almost in all categories of serious crimes, and as the temperature decreases in the winter, the data shows the occurances decreases to its lowest point at Feburary.

Crime Shifting Pattern Through Years

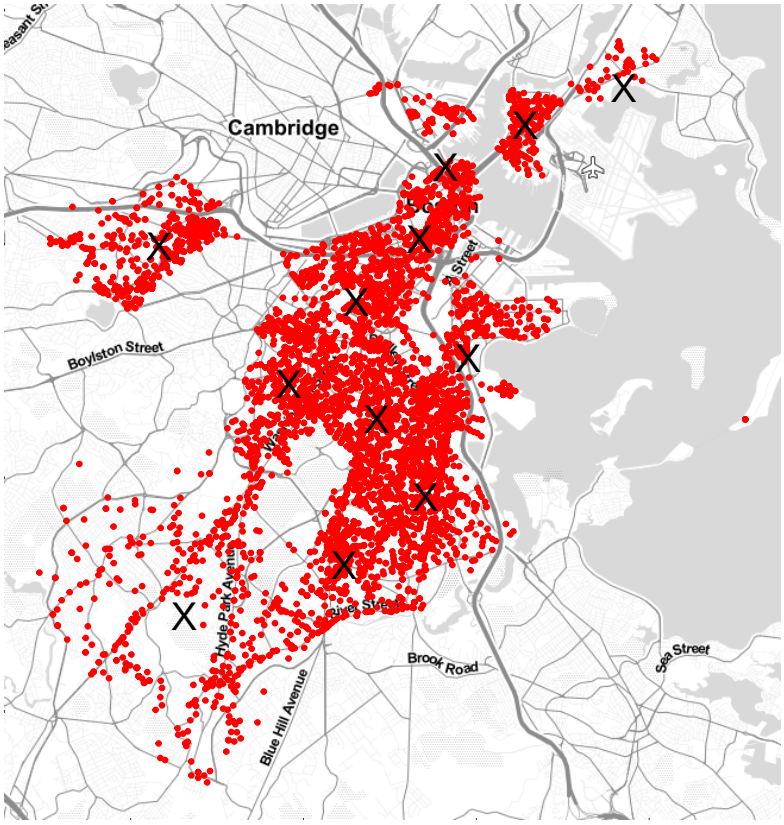
At following section, I will separate all the crime records base on the years and then inspect how the crime center shifts throughout the year.

```
#Get Data of each year
bos_violent_crimes2013 <- filter(bos_violent_crimes, Year == '2013')

bos_violent_crimes2014 <- filter(bos_violent_crimes, Year == '2014')

bos_violent_crimes2015 <- filter(bos_violent_crimes, Year == '2015')

bos_violent_crimes2016 <- filter(bos_violent_crimes, Year == '2016')

bos_violent_crimes2017 <- filter(bos_violent_crimes, Year == '2017')

## Crime clustering behaviors in 2013
bos_violent_crimes2013_loc <- as.data.frame(subset(bos_violent_crimes2013, select=c("Long", "Lat")))
crime_cluster_2013 <- kmeans(bos_violent_crimes2013_loc, 12)
```

```
centers_2013 <- as.data.frame(crime_cluster_2013$centers)

bos_violent_crimes2013_loc$cluster = crime_cluster_2013$cluster

crime_cluster_plot_2013 <- qmplot(Long, Lat, data = bos_violent_crimes2013_loc, maptype = "toner-lite",

crime_propensity_plot_2013 <- qmplot(Long, Lat, data = bos_violent_crimes2013_loc, geom = "blank", zoom
  stat_density_2d(aes(fill = ..level..), geom = "polygon", alpha = .3, color = NA) +
  scale_fill_gradient2("2013\nCrime\nPropensity", low = "white", mid = "yellow", high = "red", midpoint


## Crime clustering behaviors in 2014
bos_violent_crimes2014_loc <- as.data.frame(subset(bos_violent_crimes2014, select=c("Long", "Lat")))
crime_cluster_2014 <- kmeans(bos_violent_crimes2014_loc, 12)
centers_2014 <- as.data.frame(crime_cluster_2014$centers)

bos_violent_crimes2014_loc$cluster = crime_cluster_2014$cluster

crime_cluster_plot_2014 <- qmplot(Long, Lat, data = bos_violent_crimes2014_loc, maptype = "toner-lite",

crime_propensity_plot_2014 <- qmplot(Long, Lat, data = bos_violent_crimes2014_loc, geom = "blank", zoom
  stat_density_2d(aes(fill = ..level..), geom = "polygon", alpha = .3, color = NA) +
  scale_fill_gradient2("2014\nCrime\nPropensity", low = "white", mid = "yellow", high = "red", midpoint



## Crime clustering behaviors in 2015
bos_violent_crimes2015_loc <- as.data.frame(subset(bos_violent_crimes2015, select=c("Long", "Lat")))
crime_cluster_2015 <- kmeans(bos_violent_crimes2015_loc, 12)
centers_2015 <- as.data.frame(crime_cluster_2015$centers)

bos_violent_crimes2015_loc$cluster = crime_cluster_2015$cluster

crime_cluster_plot_2015 <- qmplot(Long, Lat, data = bos_violent_crimes2015_loc, maptype = "toner-lite",

crime_propensity_plot_2015 <- qmplot(Long, Lat, data = bos_violent_crimes2015_loc, geom = "blank", zoom
  stat_density_2d(aes(fill = ..level..), geom = "polygon", alpha = .3, color = NA) +
  scale_fill_gradient2("2015\nCrime\nPropensity", low = "white", mid = "yellow", high = "red", midpoint


## Crime clustering behaviors in 2016
bos_violent_crimes2016_loc <- as.data.frame(subset(bos_violent_crimes2016, select=c("Long", "Lat")))
crime_cluster_2016 <- kmeans(bos_violent_crimes2016_loc, 12)
centers_2016 <- as.data.frame(crime_cluster_2016$centers)

bos_violent_crimes2016_loc$cluster = crime_cluster_2016$cluster

crime_cluster_plot_2016 <- qmplot(Long, Lat, data = bos_violent_crimes2016_loc, maptype = "toner-lite",

crime_propensity_plot_2016 <- qmplot(Long, Lat, data = bos_violent_crimes2016_loc, geom = "blank", zoom
  stat_density_2d(aes(fill = ..level..), geom = "polygon", alpha = .3, color = NA) +
  scale_fill_gradient2("2016\nCrime\nPropensity", low = "white", mid = "yellow", high = "red", midpoint
```

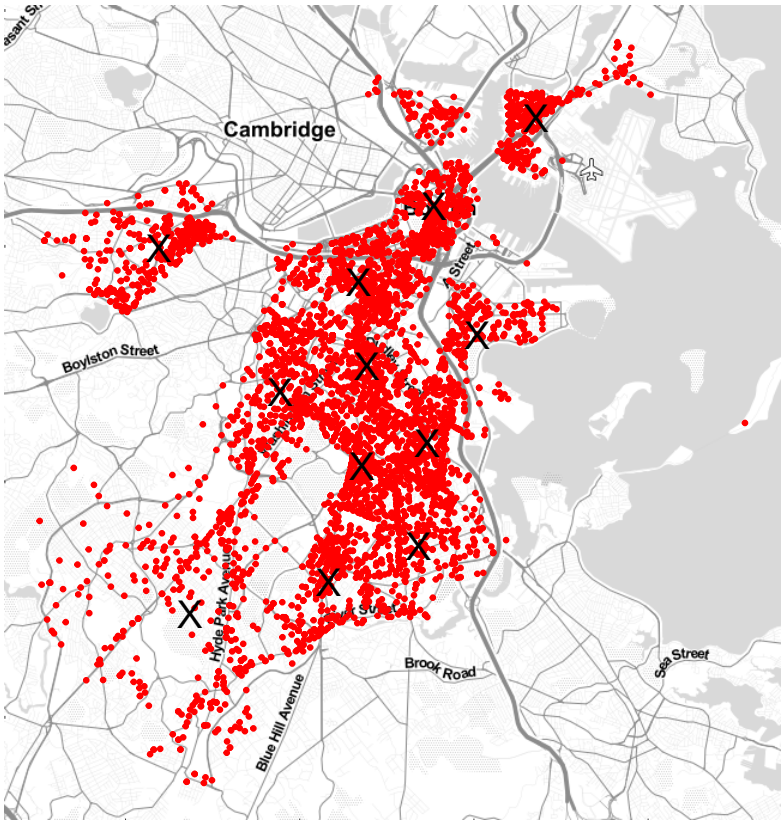
```
## Crime clustering behaviors in 2017
bos_violent_crimes2017_loc <- as.data.frame(subset(bos_violent_crimes2017, select=c("Long", "Lat")))
crime_cluster_2017 <- kmeans(bos_violent_crimes2017_loc, 12)
centers_2017 <- as.data.frame(crime_cluster_2017$centers)
bos_violent_crimes2017_loc$cluster = crime_cluster_2017$cluster

crime_cluster_plot_2017 <- qmplot(Long, Lat, data = bos_violent_crimes2017_loc, maptype = "toner-lite",

crime_propensity_plot_2017 <- qmplot(Long, Lat, data = bos_violent_crimes2017_loc, geom = "blank", zoom
  stat_density_2d(aes(fill = ..level..), geom = "polygon", alpha = .3, color = NA) +
  scale_fill_gradient2("2017\nCrime\nPropensity", low = "white", mid = "yellow", high = "red", midpoint


crime_cluster_plot_2013
```

Crime Cluster Graph 2013



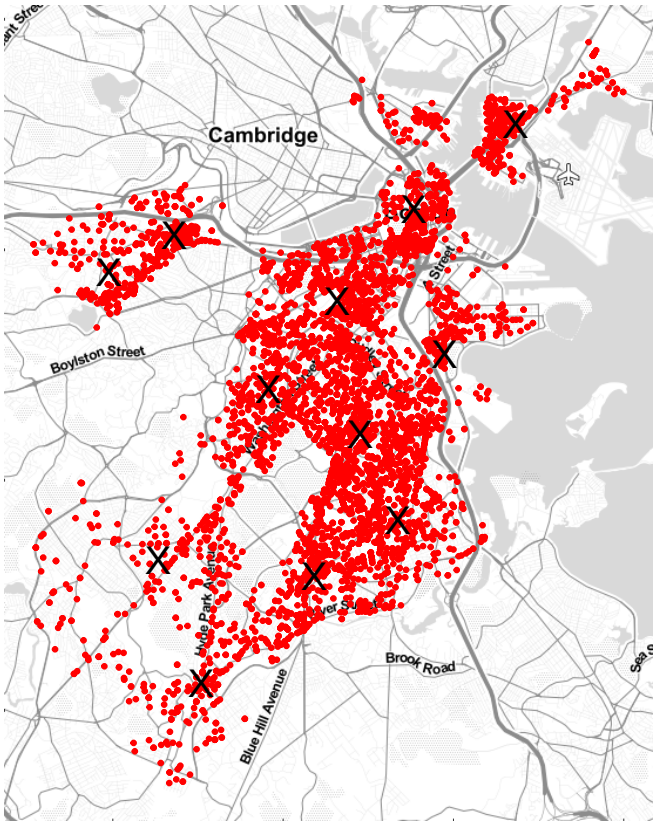

```
crime_cluster_plot_2014
```

Crime Cluster Graph 2014



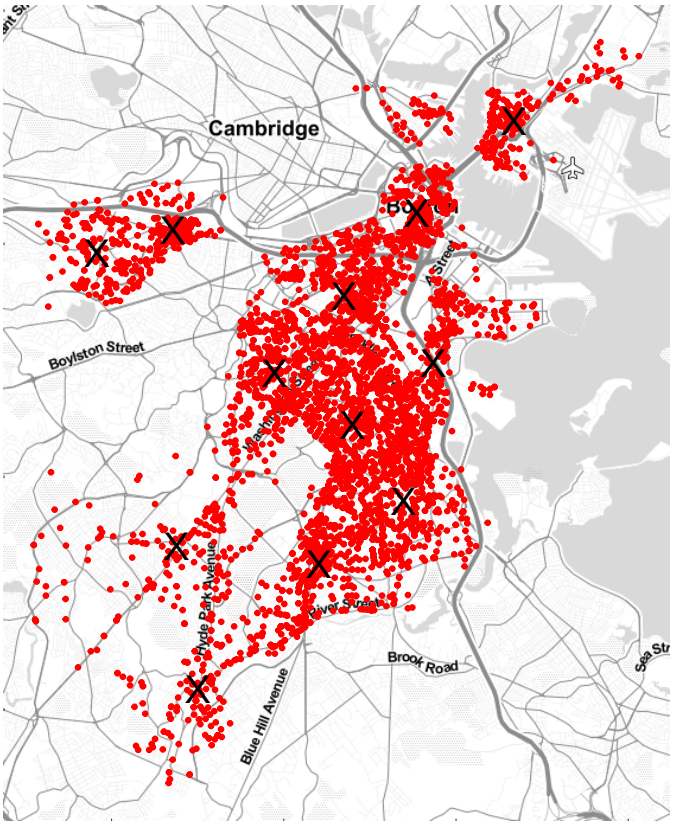

```
crime_cluster_plot_2015
```

Crime Cluster Graph 2015



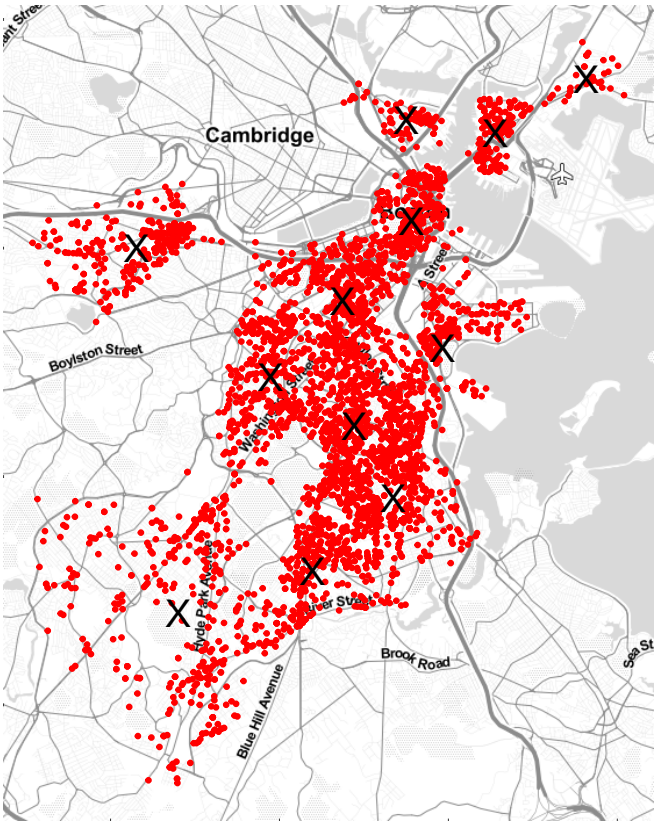

```
crime_cluster_plot_2016
```

Crime Cluster Graph 2016



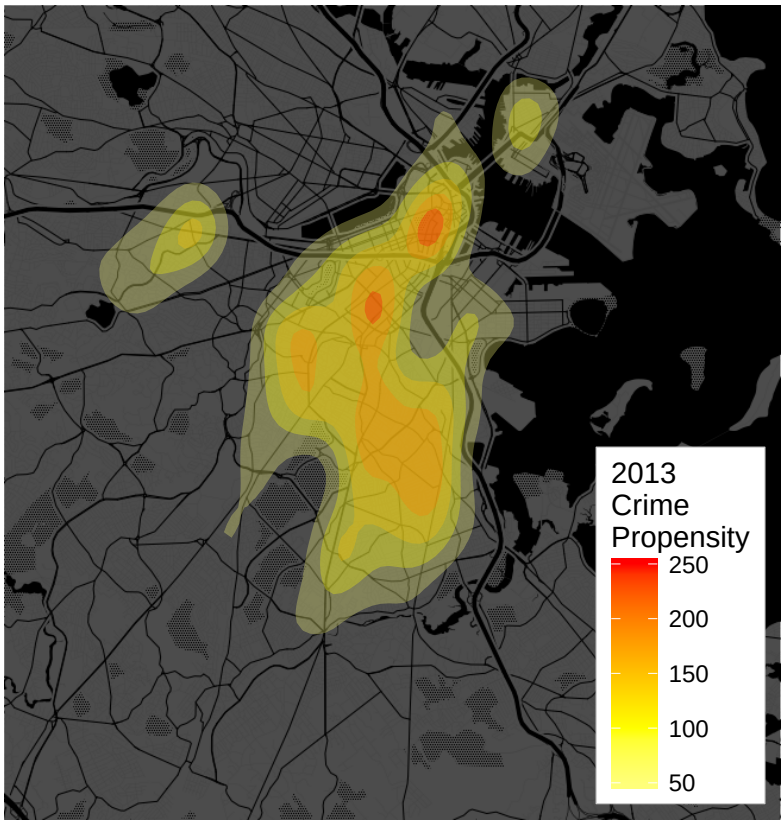

```
crime_cluster_plot_2017
```

Crime Cluster Graph 2017



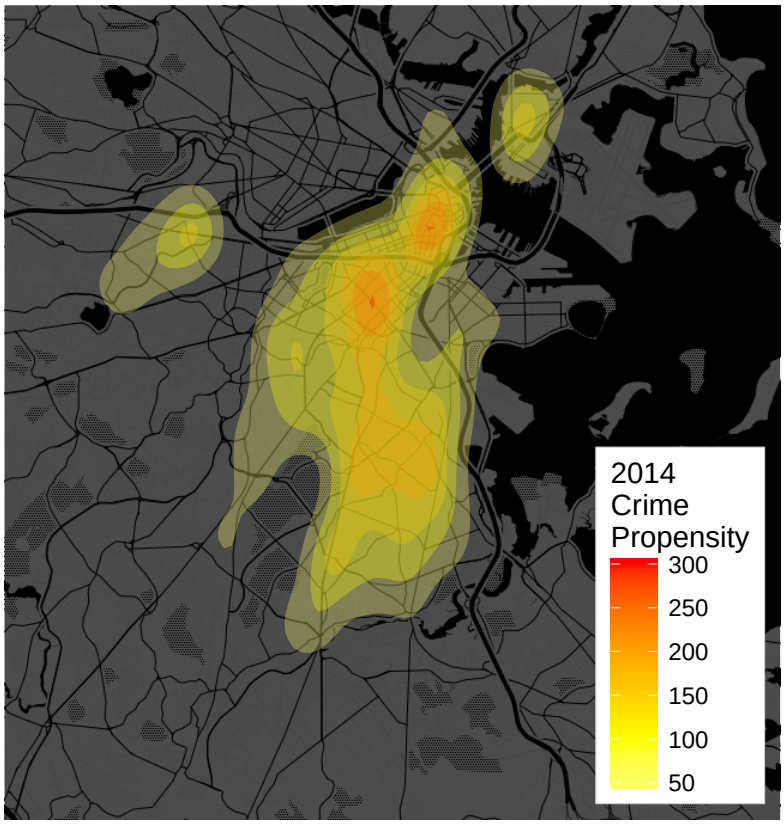

```
crime_propensity_plot_2013
```

Crime Density Graph 2013



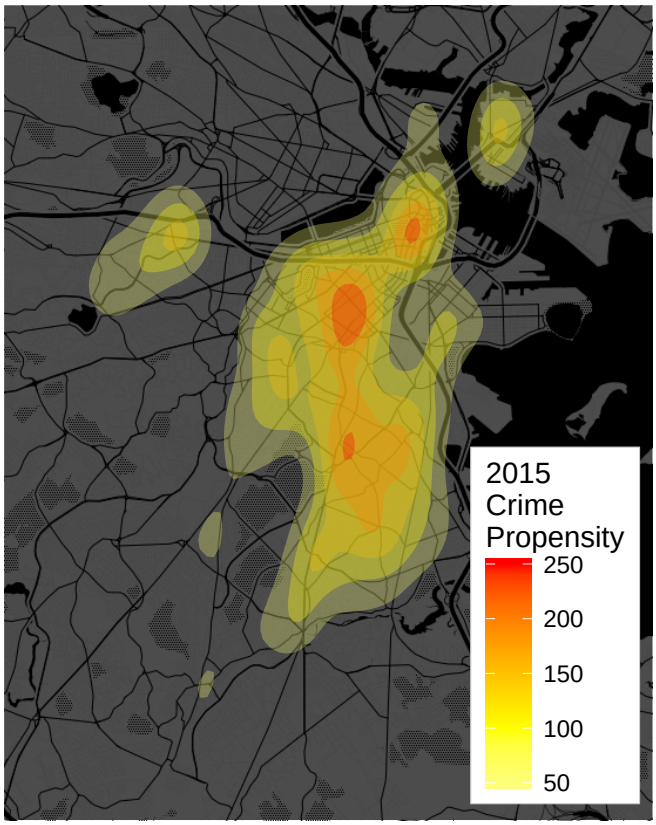

```
crime_propensity_plot_2014
```

Crime Density Graph 2014



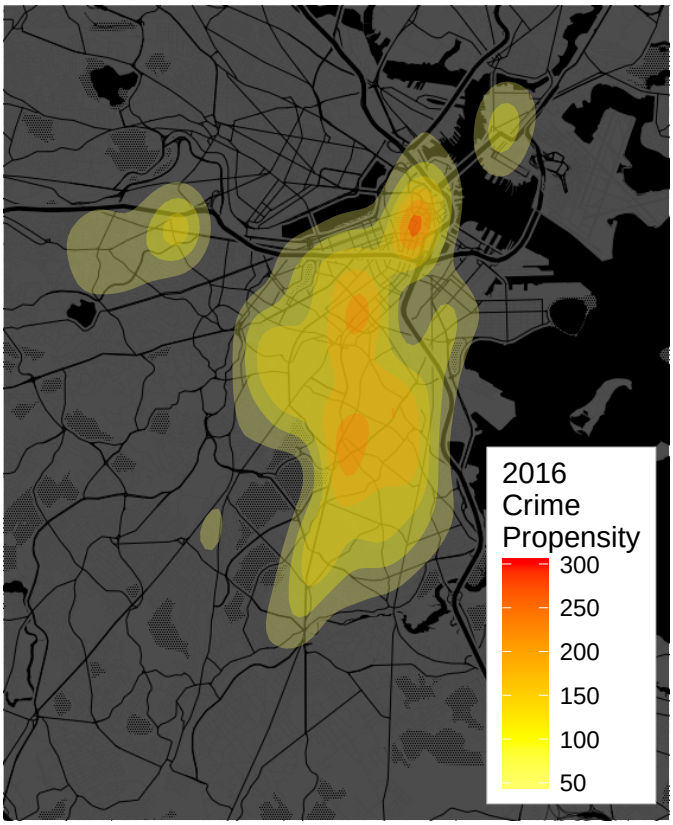

```
crime_propensity_plot_2015
```

Crime Density Graph 2015



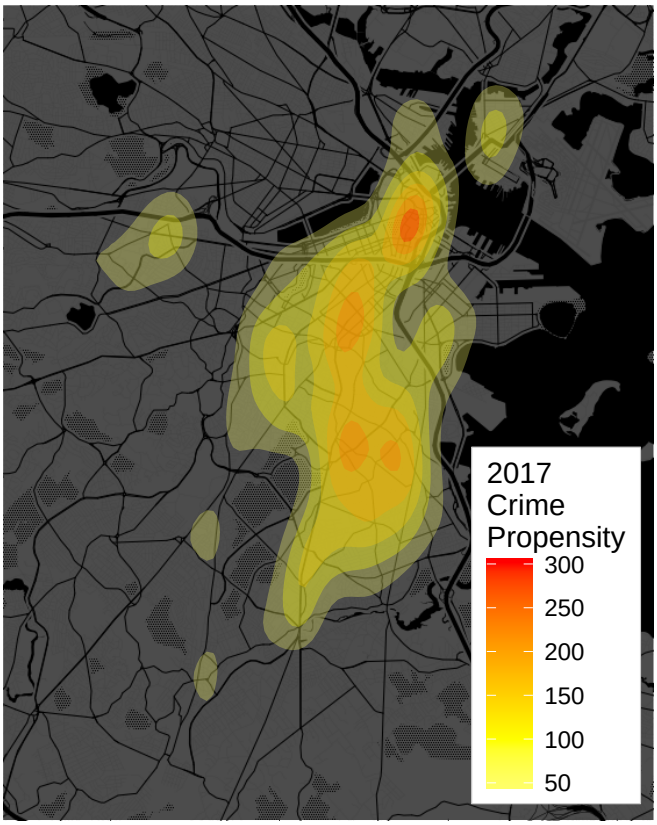

```
crime_propensity_plot_2016
```

Crime Density Graph 2016




```
crime_propensity_plot_2017
```

Crime Density Graph 2017



There are two ways we could depict the clustering of the crime locations: k-means and density. Here I did plot all the incidents from 2013-2017 respectively using both methods.

For the k-means clustering graph, we choose k = 12 as there are 12 different Police Districst in Boston. Furthermore, by comparaing the k-means clustering from each other, though there are some minor shifting in the clustering centers, we do not observe apparent change in clustering from year to year.

From the desity graphs, the trend becomes much more clear, as we see that there are some huge areas of high density around the South Boston and downtown Boston, and as we move forward, the South Boston crime density decreases to some extend. However, the Boston downtwon remains to be a “popular spots” for serious crimes.