Project 1: Boston Crime Data Analysis and Visualization

Dipti Bhattad & Shruthi Subbaiah Machimada

February 19, 2019

Reading the csv file and subsetting the data

crime <- read.csv("~/Desktop/Visualization/BostonCrime.csv")  
  
crime\_subset <- subset(crime, year >2015 & year <2019)  
  
crime\_shooting <- subset(crime\_subset, shooting == 'Y')

(Since data for full years is available only for 2016-2018, disregarding other years- 2015 and 2019)

Loading the libraries

library('ggplot2')  
library('dplyr')

## Warning: package 'dplyr' was built under R version 3.4.4

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

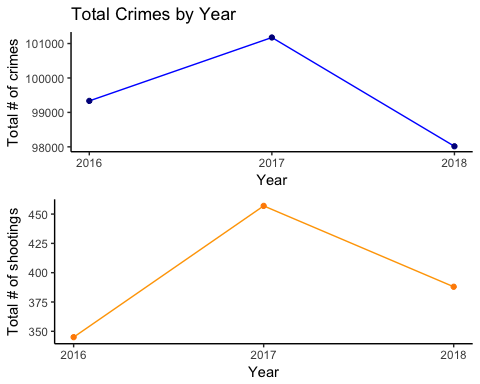
library('gridExtra')

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

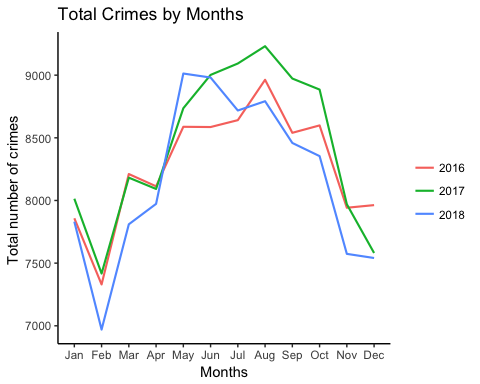
Year-wise trend in Crimes and Shootings

p1<-ggplot(data=summarise(group\_by(crime\_subset,year),incidents=n()),  
 aes(year, incidents)) + scale\_x\_continuous(breaks=seq(2016,2018,1))+  
 geom\_line(color="blue")+ geom\_point(size=1.5,color="dark blue") + ylab("Total # of crimes") + xlab("Year") + theme\_classic() + ggtitle("Total Crimes by Year")  
  
p2<-ggplot(data=summarise(group\_by(crime\_shooting,year),shooting=n()),  
 aes(year, shooting)) + scale\_x\_continuous(breaks=seq(2016,2018,1))+scale\_y\_continuous(breaks=seq(50,500,25))+  
 geom\_line(color="orange") + geom\_point(size=1.5,color="dark orange") + ylab("Total # of shootings") + xlab("Year") + theme\_classic()   
  
grid.arrange(p1,p2)

 2017 had a huge increase in crimes. This was observed in all major cities across the US.

Month-wise trend in Crimes

month\_crimes<-crime\_subset %>% group\_by(year,month) %>% count()  
month\_crimes$month<-month.abb[month\_crimes$month]  
month\_crimes$id<-rep(1:12,3)  
ggplot(data=month\_crimes,aes(reorder(month,id),n)) + geom\_line(aes(group=as.factor(year),color=as.factor(year)), lwd = 0.75) + ylab("Total number of crimes ")+ xlab("Months") + theme\_classic() + theme(legend.title=element\_blank()) + ggtitle("Total Crimes by Months")

 Feburary seems to be the month with least crimes. Looking into the mean crimes per day will help us test that.

Mean crimes per day

x<-tidyr::separate(crime\_subset,occurred\_on\_date,c("date","time"),sep=" ")  
daily\_crimes\_cnt<-summarise(group\_by(x,month,date,year),crime\_count=n())  
mean(daily\_crimes\_cnt$crime\_count)

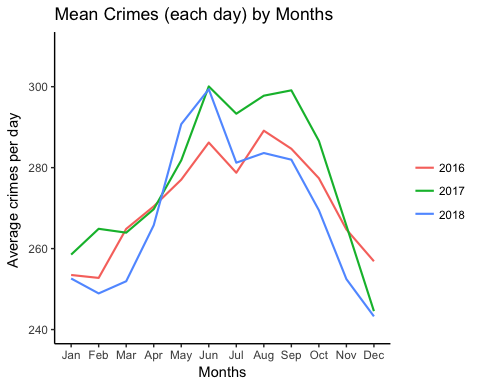
## [1] 272.3777

Mean crimes per day for each month

daily\_avg\_by\_month<-summarise(group\_by(daily\_crimes\_cnt,year,month),daily\_avg=mean(crime\_count))  
daily\_avg\_by\_month$month<-month.abb[daily\_avg\_by\_month$month]  
daily\_avg\_by\_month$id<-rep(1:12,3)

Plot of daily means for each month

ggplot(data=daily\_avg\_by\_month,  
 aes(reorder(month,id), daily\_avg)) +  
 geom\_line(aes(group=as.factor(year),color=as.factor(year)), lwd = 0.75) + ylab("Average crimes per day ") + xlab("Months") +ylim(240,310) + theme\_classic() + theme(legend.title=element\_blank()) + ggtitle("Mean Crimes (each day) by Months")

 Feburary does not have the least mean crimes per day for every year. The drop in crimes in Feburary may be because Feburary has lesser number of days. Most crimes occur between May to September. This may be because the weather is warmer in summer, or because the days are longer and most crimes occur during the day.

Hour-wise trend in Crimes

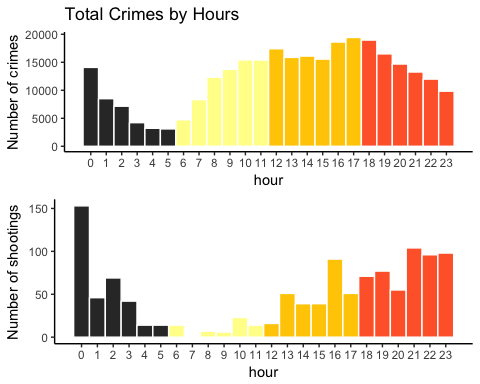
p1 <- ggplot(data = crime\_subset) + geom\_histogram(mapping = aes(x=hour),fill=c(rep("#333333",6),rep("#FFFF99",6),rep("#FFCC00",6),rep("#FF6633",6)),width=1,stat="count", color = "white") +ylab("Number of crimes")+theme\_classic()+scale\_x\_continuous(breaks=seq(0,23,1)) + ggtitle("Total Crimes by Hours")

## Warning: Ignoring unknown parameters: binwidth, bins, pad

p2 <- ggplot(data = crime\_shooting) + geom\_histogram(mapping = aes(x=hour),fill=c(rep("#333333",6),rep("#FFFF99",5),rep("#FFCC00",6),rep("#FF6633",6)),width=1,stat="count", color = "white") +ylab("Number of shootings")+theme\_classic()+scale\_x\_continuous(breaks=seq(0,23,1))

## Warning: Ignoring unknown parameters: binwidth, bins, pad

grid.arrange(p1,p2)

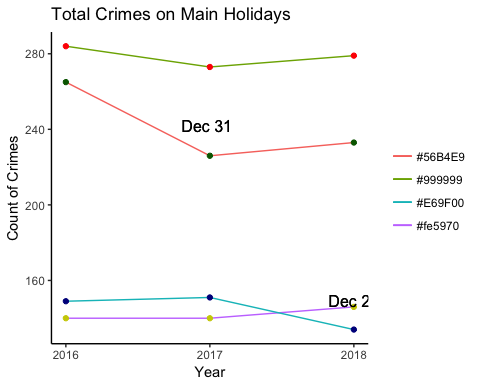
 The colours depict the 4 times in a day -Night, Morning, Noon and Evening. The crimes decrease after night and are least in the early hours of morning. The crimes increase again during the day. Most crimes occur during the day. Most shootings occur in the evening and night.

Crimes on few Main Holidays- 4th of July, Thanksgiving Day, Christmas and New Years Eve

x<-tidyr::separate(crime\_subset,occurred\_on\_date,c("date","time"),sep=" ")  
dec25 <- subset(x,date %in% c("2018-12-25","2017-12-25","2016-12-25"))  
dec31 <- subset(x,date %in% c("2018-12-31","2017-12-31","2016-12-31"))  
thanksgiving <- subset(x,date %in% c("2018-11-22","2017-11-23","2016-11-24"))  
independence <- subset(x,date %in% c("2018-07-04","2017-07-04","2016-07-04"))  
  
#install.packages("directlabels")  
library('directlabels')

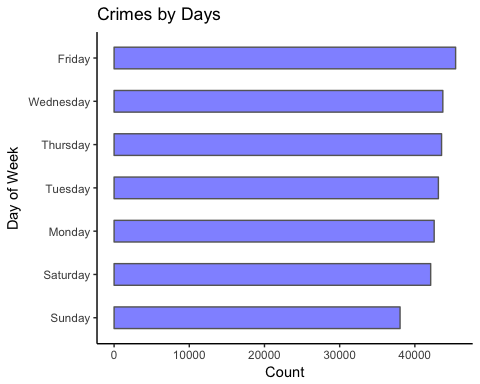
## Warning: package 'directlabels' was built under R version 3.4.4

ggplot()+  
 geom\_line(data=summarise(group\_by(dec25,year),incidents=n()), mapping = aes(x = year, y = incidents, color = "#fe5970")) + geom\_point(data=summarise(group\_by(dec25,year),incidents=n()), mapping = aes(x = year, y = incidents),size=1.5, color="yellow3") +  
geom\_dl(data=summarise(group\_by(dec25,year),incidents=n()), mapping = aes(x = year, y = incidents,label = "Dec 25"), method = "top.qp") +  
 geom\_line(data=summarise(group\_by(dec31,year),incidents=n()), mapping = aes(x = year, y = incidents, color = "#56B4E9")) + geom\_point(data=summarise(group\_by(dec31,year),incidents=n()), mapping = aes(x = year, y = incidents),size=1.5, color="dark green") +  
geom\_dl(data=summarise(group\_by(dec31,year),incidents=n()), mapping = aes(x = year, y = incidents,label = "Dec 31"), method = "smart.grid") +  
 geom\_line(data=summarise(group\_by(thanksgiving,year),incidents=n()), mapping = aes(x = year, y = incidents, color = "#E69F00")) + geom\_point(data=summarise(group\_by(thanksgiving,year),incidents=n()), mapping = aes(x = year, y = incidents),size=1.5, color="dark blue") +  
geom\_dl(data=summarise(group\_by(thanksgiving,year),incidents=n()), mapping = aes(x = year, y = incidents,label = "Thanksgiving"), method = "smart.grid") +  
 geom\_line(data=summarise(group\_by(independence,year),incidents=n()), mapping = aes(x = year, y = incidents, color = "#999999")) + geom\_point(data=summarise(group\_by(independence,year),incidents=n()), mapping = aes(x = year, y = incidents),size=1.5, color="red") +  
geom\_dl(data=summarise(group\_by(independence,year),incidents=n()), mapping = aes(x = year, y = incidents,label = "4th of July"), method = "top.qp") + scale\_x\_continuous(breaks=seq(2016,2018,1)) + theme\_classic() + theme(legend.title=element\_blank()) + ggtitle("Total Crimes on Main Holidays") + xlab("Year") + ylab("Count of Crimes")

 On Thanksgiving and Christmas, the number of crimes is a lot less that the mean number of crimes per day for those months.

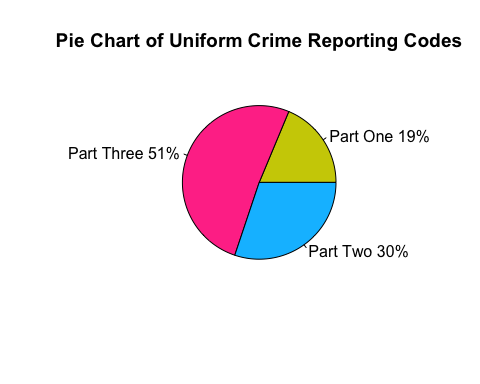
Day-wise trend in Crimes

d <- summarise(group\_by(crime\_subset,day\_of\_week),incidents=n()) %>% arrange(desc(incidents))  
ggplot(d, aes(x = reorder(day\_of\_week,incidents), y = incidents)) + geom\_bar(stat = "identity", fill = "blue", color = "grey40", alpha = .5, width = 0.5, size = 0.5) + coord\_flip() + ggtitle("Crimes by Days") + theme\_classic() + xlab("Day of Week") + ylab("Count")

 Most crime occur in Friday and the least occur on Saturday and Sunday. Weekends in Boston may be safer than weekdays.

Percentage of UCR crimes

temp <- summarise(group\_by(crime\_subset,ucr\_part), total\_count = n())  
data <- filter(temp, ucr\_part %in% c("Part One", "Part Two", "Part Three"))  
pct <- round(data$total\_count/sum(data$total\_count)\*100)  
lbls <- paste(data$ucr\_part, pct) # add percents to labels  
lbls <- paste(lbls,"%",sep="") # ad % to labels   
pie(data$total\_count,labels = lbls, main="Pie Chart of Uniform Crime Reporting Codes", col = c("yellow3", "violetred1", "deepskyblue"))

 The most common crimes are Ucr Part3 crimes, which are not very violent.

Top 5 UCR Part 1 crimes

### UCR Part 1 crimes

ucr\_part1 <- subset(crime\_subset, ucr\_part == "Part One")

### Top crimes in ucr part 1

top\_crimes\_ucrpart1 <- ucr\_part1 %>% count(offense\_code\_group) %>% arrange(desc(n)) %>% top\_n(5)

## Selecting by n

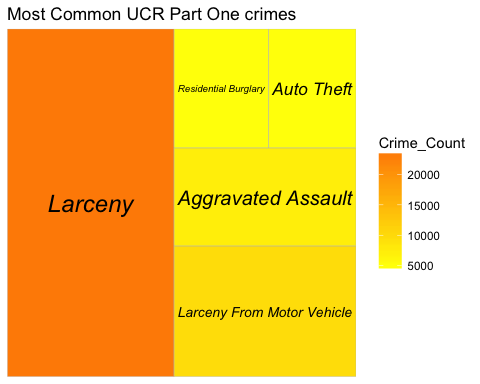
top\_crimes\_ucrpart1\_v <- as.character(top\_crimes\_ucrpart1$offense\_code\_group)  
top\_crimes\_ucrpart1$Crime\_Count <- top\_crimes\_ucrpart1$n

### Treemaps for most common ucr part 1 crimes

library(treemapify)

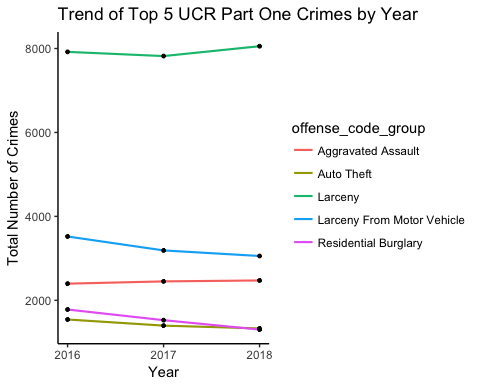
## Warning: package 'treemapify' was built under R version 3.4.4

ggplot(top\_crimes\_ucrpart1, aes(area=Crime\_Count,label=offense\_code\_group,fill=Crime\_Count))+geom\_treemap()+ geom\_treemap\_text(fontface = "italic", colour = "black", place = "centre",grow = FALSE)+ggtitle("Most Common UCR Part One crimes") + scale\_fill\_gradient(low = "yellow", high = "dark orange")



Year-wise Trend in Top 5 UCR Part One Crimes

df1 <- subset(ucr\_part1, offense\_code\_group %in% top\_crimes\_ucrpart1\_v)  
ggplot(data = summarise(group\_by(df1,offense\_code\_group,year), crime\_count\_ucr1=n())) + geom\_line(aes(year,crime\_count\_ucr1, color = offense\_code\_group), lwd = 0.75) + geom\_point(aes(year,crime\_count\_ucr1),size=1,color="black") + xlab("Year") + ylab("Total Number of Crimes") + scale\_x\_continuous(breaks=seq(2016,2018,1)) + theme\_classic() + ggtitle("Trend of Top 5 UCR Part One Crimes by Year")



Distribution of Top 5 UCR Part 1 Crimes Across Top 5 Unsafest Districts

### Unsafe districts in ucr part 1

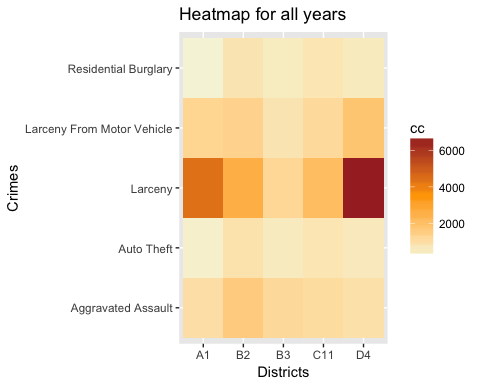
unsafe\_dist\_ucr1 <- ucr\_part1 %>% count(district) %>% arrange(desc(n)) %>% top\_n(5)

## Selecting by n

unsafe\_dist\_ucr1\_v <- as.character(unsafe\_dist\_ucr1$district)

### Heatmap for all years together

heatmap\_data = subset(ucr\_part1, offense\_code\_group %in% top\_crimes\_ucrpart1\_v & district %in% unsafe\_dist\_ucr1\_v)  
ggplot(data = summarise(group\_by(heatmap\_data,district,offense\_code\_group), cc = n())) + geom\_tile(aes(x=district,y=offense\_code\_group, fill=cc)) + scale\_fill\_gradient2(low = "beige", mid = "orange", high = "brown", midpoint = 3500) + ggtitle("Heatmap for all years") + xlab("Districts") + ylab("Crimes")

 Larceny, the most common UCR Part 1 crime is highly distributed across all districts. The next common crime is Larceny from Motor Vehicle, which is not that high. The distribution of these crimes were similar for all years, so we aggregated all years in one heatmap.

Map of Boston districts and the distribution of crime across them

The shapefile could not be read from the Rmd file, so we read it and converted into a dataframe and exported it into a csv. We then read the csv through Rmd.

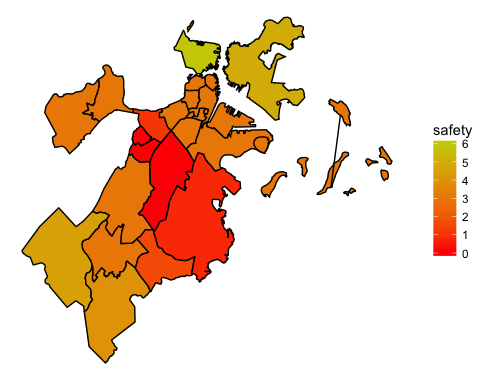
library(maptools)

## Warning: package 'maptools' was built under R version 3.4.4

## Loading required package: sp

## Checking rgeos availability: FALSE  
## Note: when rgeos is not available, polygon geometry computations in maptools depend on gpclib,  
## which has a restricted licence. It is disabled by default;  
## to enable gpclib, type gpclibPermit()

#area<-readShapePoly("Boston\_Neighborhoods.shp")   
#area\_df<-fortify(area)  
area\_df<-read.csv("~/Desktop/Visualization/Area\_Maps.csv")  
area\_df$safety<-3  
#Most crimes  
#Roxbury  
area\_df <- within(area\_df, safety[group == 8.1] <- 0) #B2  
area\_df <- within(area\_df, safety[group == 2.1] <- 0) #B2  
area\_df <- within(area\_df, safety[group == 3.1] <- 0) #B2  
#Dorchester  
area\_df <- within(area\_df, safety[group == 21.1] <- 0.7) #C11  
#Fenway,Backbay  
area\_df<-within(area\_df, safety[group == 16.1] <- 1) #D4  
#Mattapan  
area\_df<-within(area\_df, safety[group == 20.1] <- 1.5) #B3  
#Least crimes  
area\_df<-within(area\_df, safety[group == 12.1] <- 6) #A15  
area\_df<-within(area\_df, safety[group == 11.1] <- 5) #A7  
area\_df<-within(area\_df, safety[group == 18.1] <- 4.5) #E5  
area\_df<-within(area\_df, safety[group == 19.1] <- 4) #E18  
ggplot(area\_df,aes(x=long,y=lat,group=group,fill=safety))+geom\_polygon(color="black")+theme\_void()+scale\_fill\_gradient(low = "red", high = "yellow3")

 Most populated- Dorchester(Red) and Roxbury (Dark Red) Most number of crimes- Roxbury (Dark Red) and Dorchester(Red) Most number of shootings - Roxbury (Dark Red) and Mattapan(light red) Most number of UCR Part 1 crimes - Fenway, BackBay and Downtown (Orange) Safest areas (Least crimes) - East Boston, Charleston (Light green)

Has the safety of districts changed over the years

unsafe\_dist <- crime\_subset %>% count(district) %>% arrange(desc(n)) %>% top\_n(5)

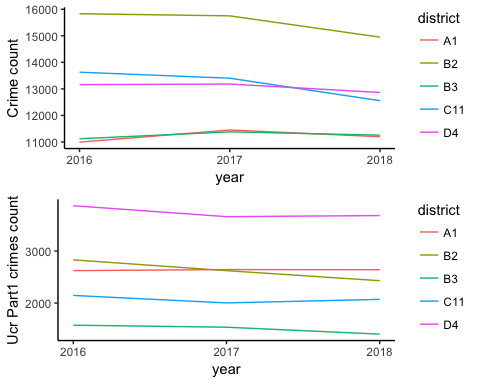
## Selecting by n

unsafe\_dist

## # A tibble: 5 x 2  
## district n  
## <fct> <int>  
## 1 B2 46534  
## 2 C11 39589  
## 3 D4 39210  
## 4 B3 33762  
## 5 A1 33646

unsafe\_dist\_v <- as.character(unsafe\_dist$district)

df2 <- subset(crime\_subset, district %in% unsafe\_dist\_v)  
p1<-ggplot(data = summarise(group\_by(df2,district,year), crime\_count\_dist=n())) + geom\_line(aes(year,crime\_count\_dist, color = district)) + scale\_x\_continuous(breaks=seq(2016,2018,1))+ylab("Crime count")+theme\_classic()  
  
df\_ucr<-subset(ucr\_part1,district %in% unsafe\_dist\_v)  
p2<-ggplot(data = summarise(group\_by(df\_ucr,district,year), crime\_count\_dist=n())) + geom\_line(aes(year,crime\_count\_dist, color = district)) + scale\_x\_continuous(breaks=seq(2016,2018,1))+ylab("Ucr Part1 crimes count")+theme\_classic()  
  
grid.arrange(p1,p2)



The crimes in Roxbury and Dorchester have decreased in 2018, when compared to 2016 and 2017. But the UCR Part 1 crimes in these districts have remained the same. Given more historic data we would be able to find trends.

Unsafe Streets in Boston

unsafe\_streets = crime\_subset %>% count(street) %>% arrange(desc(n)) %>% top\_n(11)

## Selecting by n

unsafe\_streets <- unsafe\_streets[c(1,3:11),]  
unsafe\_streets$Crime\_Count <- unsafe\_streets$n  
#unsafe\_streets

Unsafe streets in Dorchester and Roxbury

unsafe\_street\_RoxDorch = filter(crime\_subset,district=="B2" |district=="C11" ) %>% count(street) %>% arrange(desc(n)) %>% top\_n(11)

## Selecting by n

unsafe\_street\_RoxDorch <- unsafe\_street\_RoxDorch[c(1:3,5:11),]  
unsafe\_street\_RoxDorch$Crime\_Count <- unsafe\_street\_RoxDorch$n  
#unsafe\_street\_RoxDorch

Unsafe Streets(Shooting) in Boston

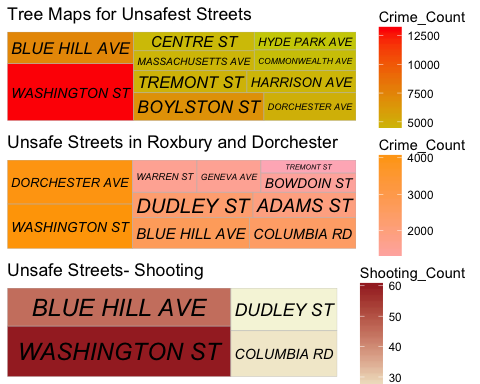
unsafe\_street\_shooting = subset(crime\_shooting) %>% count(street) %>% arrange(desc(n)) %>% top\_n(11)

## Selecting by n

unsafe\_streets\_shooting <- unsafe\_street\_shooting[c(1:4),]  
unsafe\_streets\_shooting$Shooting\_Count <- unsafe\_streets\_shooting$n

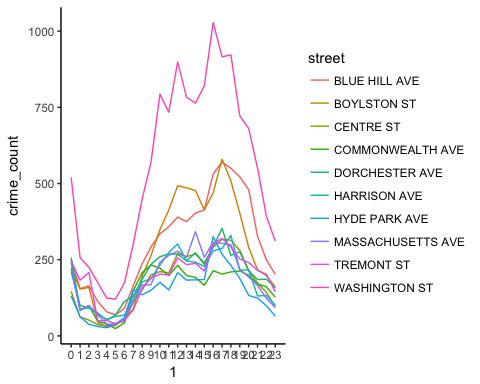
Unsafe streets in boston, and Unsafe streets in Roxbury and Dorchester

t1 <- ggplot(unsafe\_streets, aes(area=Crime\_Count,label=street,fill=Crime\_Count))+geom\_treemap()+ geom\_treemap\_text(fontface = "italic", colour = "black", place = "centre",grow = FALSE)+ggtitle("Tree Maps for Unsafest Streets") + scale\_fill\_gradient(low = "yellow3", high = "red")  
  
t2 <- ggplot(unsafe\_street\_RoxDorch, aes(area=Crime\_Count,label=street,fill=Crime\_Count))+geom\_treemap()+ geom\_treemap\_text(fontface = "italic", colour = "black", place = "centre",grow = FALSE)+ggtitle("Unsafe Streets in Roxbury and Dorchester") + scale\_fill\_gradient(low = "light pink", high = "orange")  
  
t3 <- ggplot(unsafe\_streets\_shooting, aes(area=Shooting\_Count,label=street,fill=Shooting\_Count))+geom\_treemap()+ geom\_treemap\_text(fontface = "italic", colour = "black", place = "centre",grow = FALSE)+ggtitle("Unsafe Streets- Shooting") + scale\_fill\_gradient(low = "beige", high = "brown")  
  
grid.arrange(t1,t2,t3)



Daily pattern of crimes on the unsafe streets

unsafe\_streets\_trend<- crime\_subset[crime\_subset$street %in% unsafe\_streets$street,]  
ggplot(data=summarise(group\_by(unsafe\_streets\_trend,hour,street),crime\_count=n()))+geom\_line(aes(hour,crime\_count,color=street))+scale\_x\_continuous(breaks=0:23,1)+xlab("Hour of day")+theme\_classic()

 Crimes increase in all these streets during the day, with most crimes accuring from 5 PM to 8 PM. Crimes in Boylston street and Washington street follow a very unpredictable pattern.

Crimes which involved shooting

shooting\_offenses<-crime\_shooting %>% count(offense\_code\_group) %>% arrange(desc(n)) %>% top\_n(10)

## Selecting by n

ggplot(data=shooting\_offenses,aes(x=reorder(offense\_code\_group,n),y=n))+geom\_col(aes(fill=offense\_code\_group))+coord\_flip()+scale\_x\_discrete(breaks=seq(0,600,100))+  
 xlab("Offense")+ylab("Number of shootings")+theme\_classic()

