Boston Crime Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.4.0     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.1.3
```

```
library(forcats)
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.1.3

##
## Attaching package: 'scales'

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

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(dplyr)
library(sf)
```

```
## Warning: package 'sf' was built under R version 4.1.3
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.1, PROJ 7.2.1; sf_use_s2() is TRUE
```

```
library(mapview)
```

```
## Warning: package 'mapview' was built under R version 4.1.3
```

```
library(ggmap)
```

```
## Warning: package 'ggmap' was built under R version 4.1.3
```

```
## i Google's Terms of Service: <https://mapsplatform.google.com>
```

```
## i Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.1.3
```

```
library(deforestable)
```

```
## Warning: package 'deforestable' was built under R version 4.1.3
```

```
#importing the data
boston_crime_df <- read.csv("C:\\Users\\17579\\Desktop\\Boston_Crime_Data.csv")
```

```
str(boston_crime_df)
```

```
## 'data.frame':    476655 obs. of  17 variables:
##  $ INCIDENT_NUMBER    : chr  "I182061268" "I172040657" "I162013546" "I152067251" ...
##  $ OFFENSE_CODE       : int  3201 2629 3201 1102 2647 1106 3130 3115 3201 1874 ...
##  $ OFFENSE_CODE_GROUP : chr  "Property Lost" "Harassment" "Property Lost" "Fraud" ...
##  $ OFFENSE_DESCRIPTION: chr  "PROPERTY - LOST" "HARASSMENT" "PROPERTY - LOST" "FRAUD - FALSE PRETENSE
##  $ DISTRICT           : chr  "" "C11" "B3" "A1" ...
##  $ REPORTING_AREA     : int  NA 397 433 93 359 456 20 20 282 289 ...
##  $ SHOOTING           : chr  "" "" "" "" ...
##  $ OCCURRED_ON_DATE   : chr  "6/15/2015 0:00" "6/15/2015 0:00" "6/15/2015 0:00" "6/15/2015 0:00" ...
##  $ YEAR               : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ MONTH              : int  6 6 6 6 6 6 6 6 6 6 ...
##  $ DAY_OF_WEEK        : chr  "Monday" "Monday" "Monday" "Monday" ...
##  $ HOUR               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ UCR_PART           : chr  "Part Three" "Part Two" "Part Three" "Part Two" ...
##  $ STREET             : chr  "BERNARD" "MELBOURNE ST" "NORFOLK ST" "FANEUIL HALL SQ" ...
##  $ Lat                : num  -1 42.3 42.3 42.4 42.3 ...
##  $ Long               : num  -1 -71.1 -71.1 -71.1 -71.1 ...
##  $ Location           : chr  "(-1.00000000, -1.00000000)" "(42.29109287, -71.06594539)" "(42.2836343
```

```
summary(boston_crime_df)
```

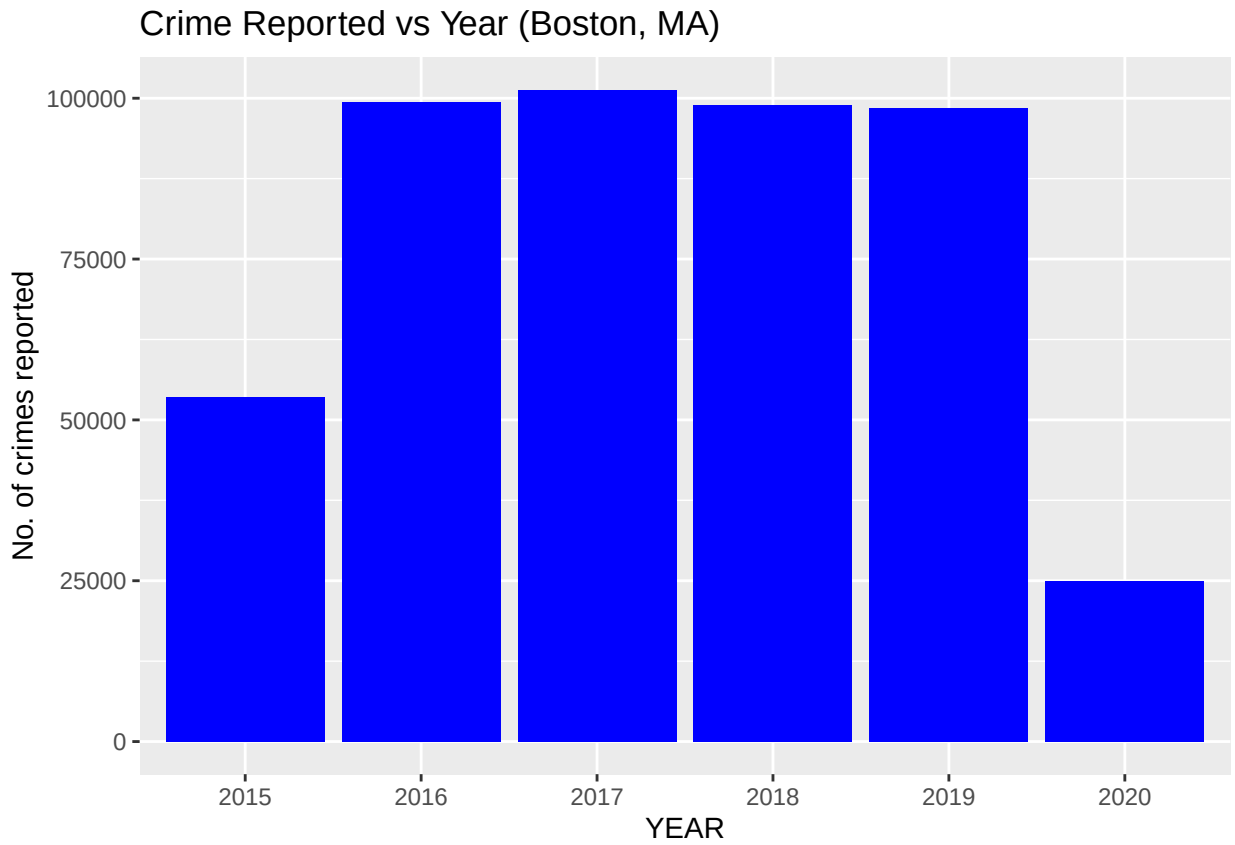
```
##  INCIDENT_NUMBER     OFFENSE_CODE  OFFENSE_CODE_GROUP OFFENSE_DESCRIPTION
##  Length:476655      Min.   : 111   Length:476655      Length:476655
##  Class :character   1st Qu.:1102   Class :character   Class :character
##  Mode  :character   Median :3005   Mode  :character   Mode  :character
##                     Mean   :2333
##                     3rd Qu.:3201
##                     Max.   :3831
##
##    DISTRICT          REPORTING_AREA    SHOOTING          OCCURRED_ON_DATE
##  Length:476655      Min.   :  0.0   Length:476655      Length:476655
##  Class :character   1st Qu.:178.0   Class :character   Class :character
##  Mode  :character   Median :345.0   Mode  :character   Mode  :character
##                     Mean   :384.6
##                     3rd Qu.:542.0
##                     Max.   :962.0
##                     NA's   :32293
##       YEAR          MONTH        DAY_OF_WEEK             HOUR
##  Min.   :2015   Min.   : 1.000   Length:476655      Min.   : 0.00
##  1st Qu.:2016   1st Qu.: 4.000   Class :character   1st Qu.: 9.00
##  Median :2017   Median : 7.000   Mode  :character   Median :14.00
##  Mean   :2017   Mean   : 6.634                      Mean   :13.09
##  3rd Qu.:2019   3rd Qu.:10.000                      3rd Qu.:18.00
##  Max.   :2020   Max.   :12.000                      Max.   :23.00
##
##    UCR_PART            STREET               Lat             Long
##  Length:476655      Length:476655      Min.   :-1.00   Min.   :-71.24
##  Class :character   Class :character   1st Qu.:42.30   1st Qu.:-71.10
##  Mode  :character   Mode  :character   Median :42.33   Median :-71.08
##                                        Mean   :42.23   Mean   :-70.94
##                                        3rd Qu.:42.35   3rd Qu.:-71.06
##                                        Max.   :42.45   Max.   :  0.00
##                                        NA's   :28183   NA's   :28183
##    Location
##  Length:476655
##  Class :character
##  Mode  :character
##
##
##
##
```

```
boston_crime_df$INCIDENT_NUMBER <- as.factor(boston_crime_df$INCIDENT_NUMBER)
boston_crime_df$OFFENSE_CODE_GROUP <- as.factor(boston_crime_df$OFFENSE_CODE_GROUP)
boston_crime_df$OFFENSE_DESCRIPTION <- as.factor(boston_crime_df$OFFENSE_DESCRIPTION)
boston_crime_df$DISTRICT <- as.factor(boston_crime_df$DISTRICT)
boston_crime_df$SHOOTING <- as.factor(boston_crime_df$SHOOTING)
boston_crime_df$OCCURRED_ON_DATE <- as.Date(boston_crime_df$OCCURRED_ON_DATE, format = "%m/%d/%y")
boston_crime_df$DAY_OF_WEEK <- as.factor(boston_crime_df$DAY_OF_WEEK)
boston_crime_df$UCR_PART <- as.factor(boston_crime_df$UCR_PART)
boston_crime_df$STREET <- as.factor(boston_crime_df$STREET)
boston_crime_df$Location <- as.factor(boston_crime_df$Location)
```

```
boston_crime_df$YEAR <- as.factor(boston_crime_df$YEAR)
```

Boston's Crime Trend

```
# Yearly crime trend in Boston, MA.

boston_crime_df %>%
  select(YEAR) %>%
  group_by(YEAR) %>%
  summarise(count_k = n()) %>%
  ggplot()+
  geom_bar(fill = 'blue', stat = 'identity', aes(x = YEAR, y = count_k))+
  ggtitle("Crime Reported vs Year (Boston, MA)")+
  ylab("No. of crimes reported")
```



Fewer records exist for years 2015 and 2020 as compared to rest of the years.

```
boston_crime_df %>%
  group_by(YEAR, MONTH) %>%
  summarise(
    count_crimes = n()
  ) %>%
  summarise(
```

```
    average_monthly_crime = mean(count_crimes)
  )
```

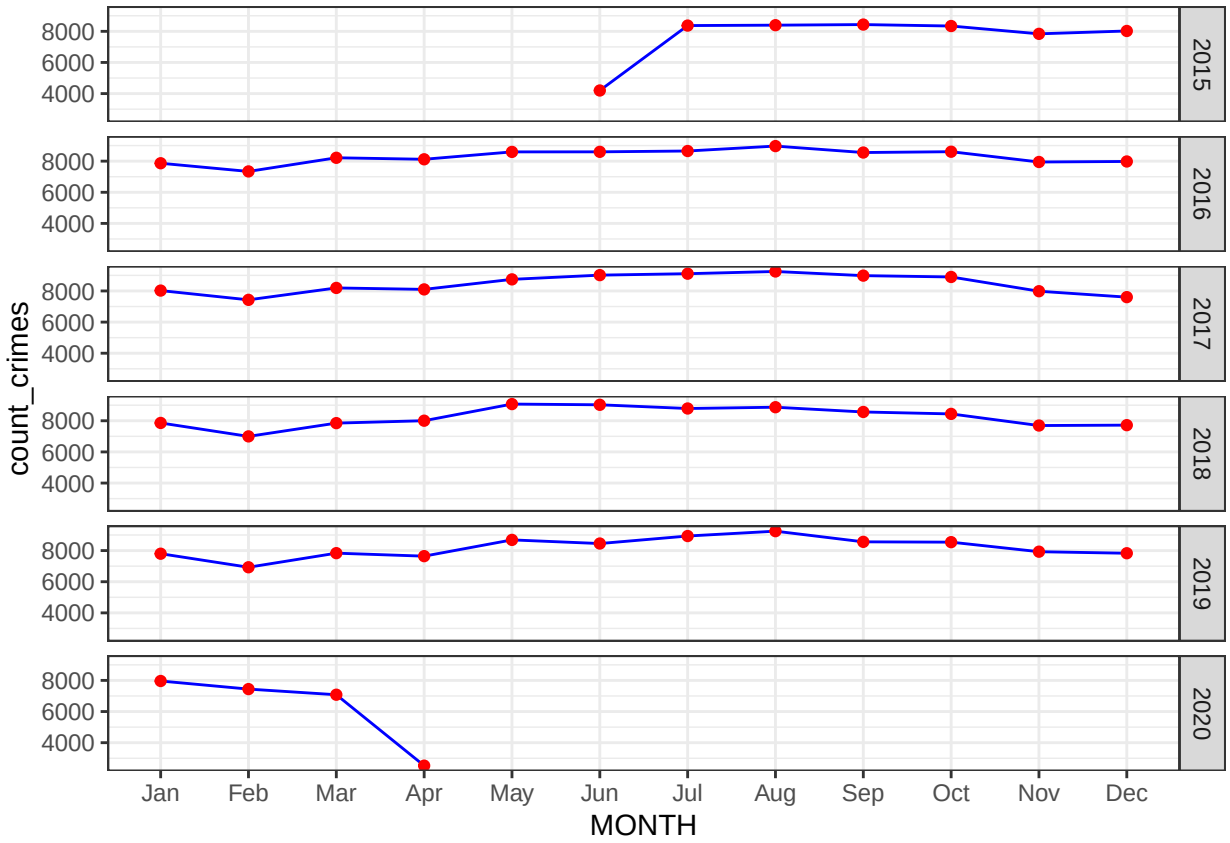
```
## `summarise()` has grouped output by 'YEAR'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 6 x 2
##   YEAR  average_monthly_crime
##   <fct>                 <dbl>
## 1 2015                  7657.
## 2 2016                  8286.
## 3 2017                  8445.
## 4 2018                  8241.
## 5 2019                  8200.
## 6 2020                  6252
```

Similar monthly crime rate over the years.

```
boston_crime_df %>%
  group_by(YEAR, MONTH) %>%
  summarise(
    count_crimes = n()
  ) %>%
  ggplot(aes(x = MONTH, y = count_crimes))+
  geom_line(color = "blue")+
  geom_point(color = "red") +
  scale_x_discrete(limits = month.abb)+
  facet_grid(rows = vars(YEAR))+theme_bw()
```

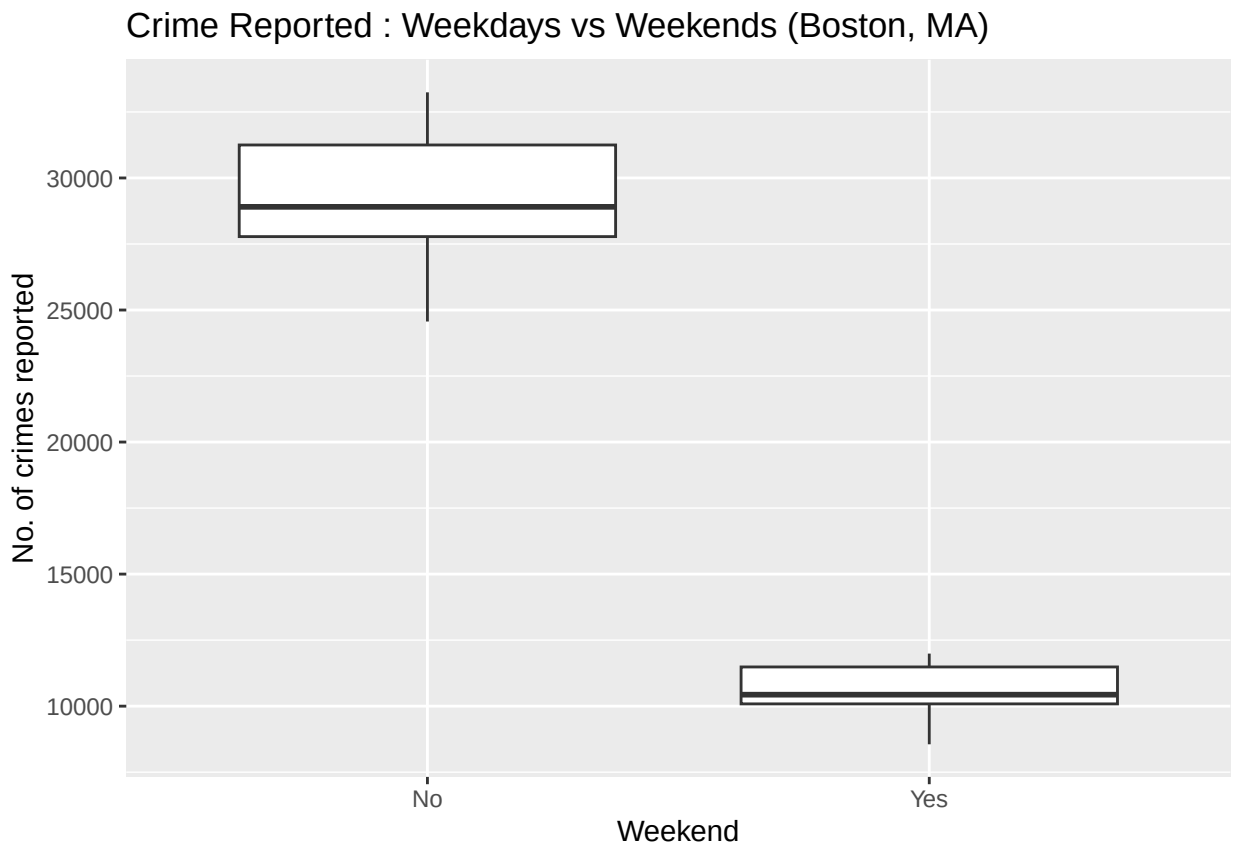
```
## `summarise()` has grouped output by 'YEAR'. You can override using the `.groups`
## argument.
```



Recording starts from June 2015 and ends at April 2020. Monthly crimes follows similar pattern each year.

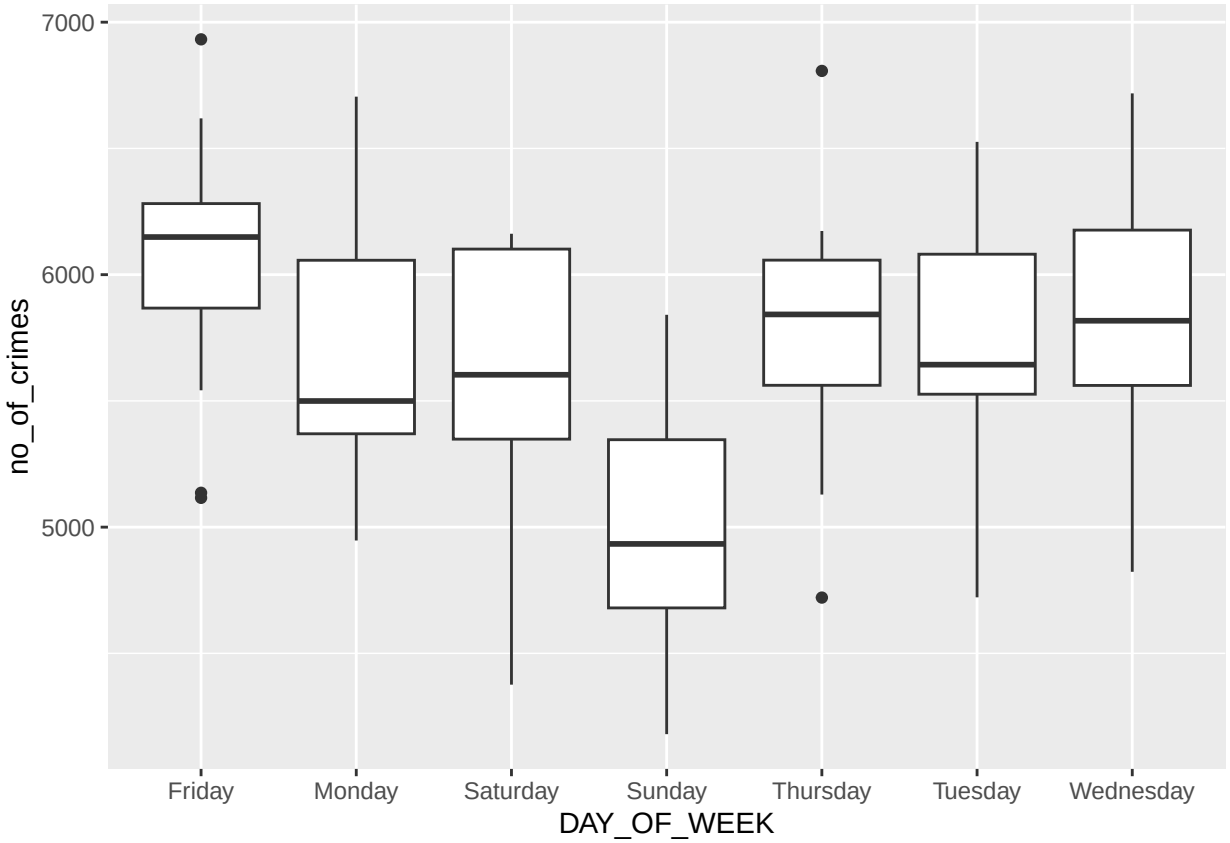
```
boston_crime_df %>%
  mutate(weekend = if_else(DAY_OF_WEEK == "Saturday" | DAY_OF_WEEK == "Sunday", "Yes", "No")) %>%
  group_by(MONTH, weekend) %>%
  summarise(
    no_of_days = n()
  ) %>%
  ggplot(aes(x = weekend, y = no_of_days))+
  geom_boxplot()+
  ggtitle("Crime Reported : Weekdays vs Weekends (Boston, MA)")+
  ylab("No. of crimes reported")+
  xlab("Weekend")
```

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



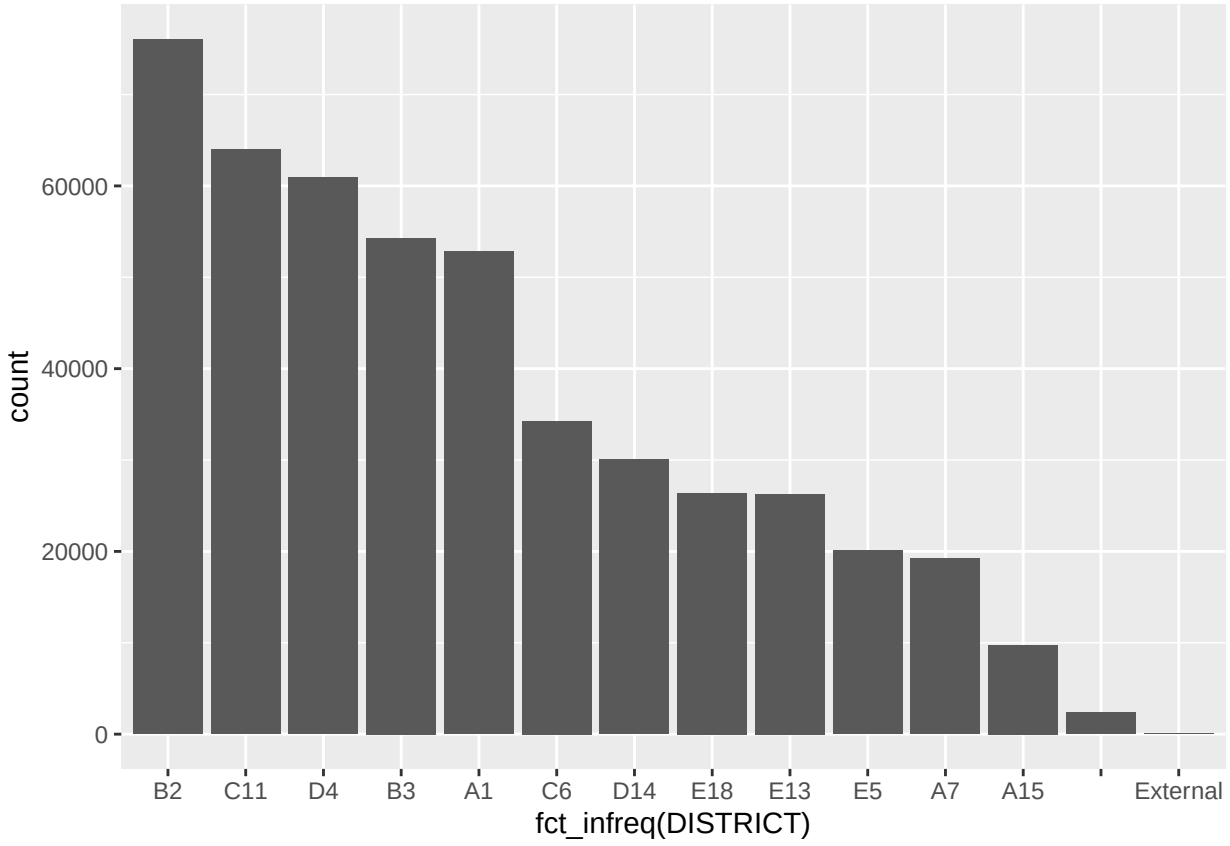

```
boston_crime_df %>%
  group_by(MONTH, DAY_OF_WEEK) %>%
  summarise(
    no_of_crimes = n()
  ) %>%
  ggplot(aes(x = DAY_OF_WEEK, y = no_of_crimes))+
  geom_boxplot()
```

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



Crime Pattern in Boston's Districts and Streets.

```
ggplot(boston_crime_df, aes(x=fct_infreq(DISTRICT)))+
  geom_bar(stat = "count")
```




```
b1 <- boston_crime_df %>%
  filter(DISTRICT != "") %>%
  group_by(DISTRICT, STREET) %>%
  summarise(
    countk = n()
  ) %>%
  arrange(DISTRICT, desc(countk))
```

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

```
b1
```

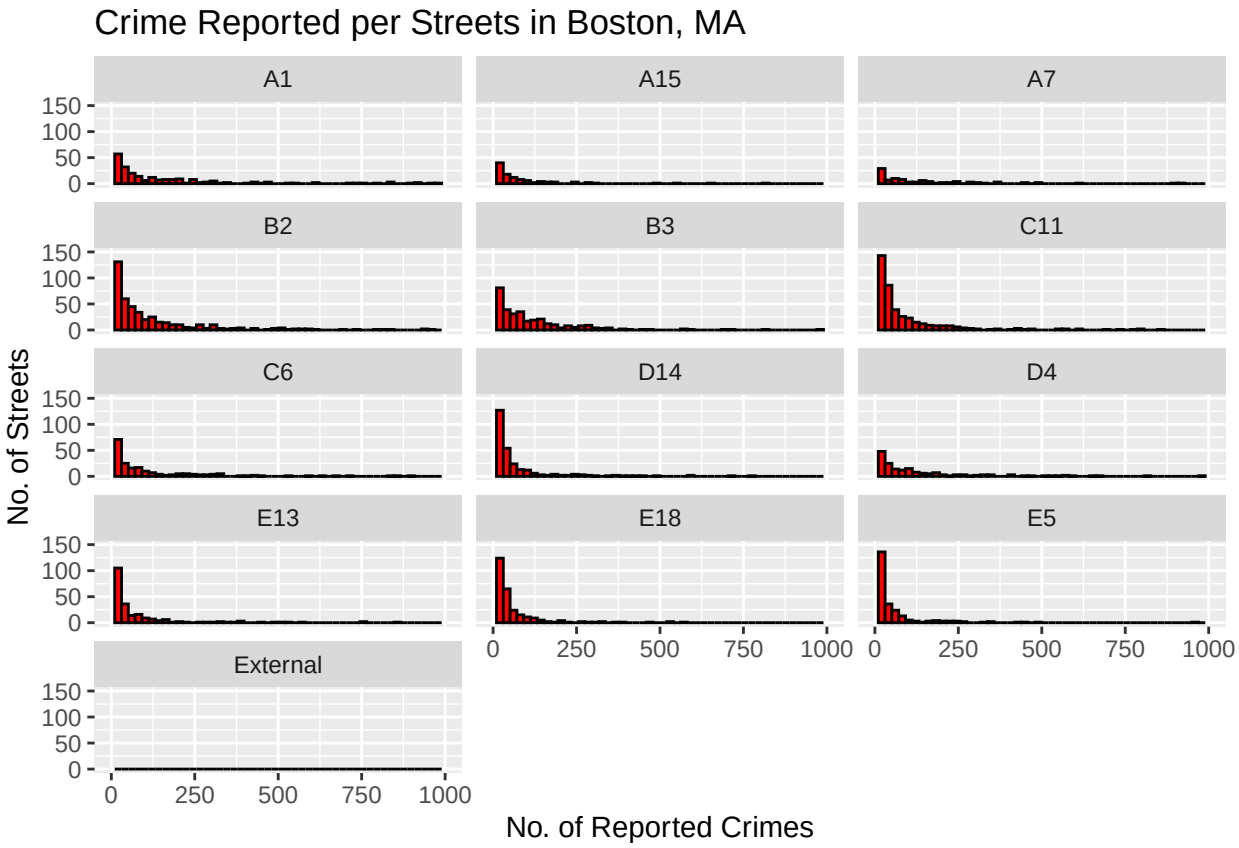
```
## # A tibble: 5,705 x 3
## # Groups:   DISTRICT [13]
##    DISTRICT STREET           countk
##    <fct>    <fct>             <int>
##  1 A1       "WASHINGTON ST"    4034
##  2 A1       "TREMONT ST"       3260
##  3 A1       ""                 2440
##  4 A1       "BOYLSTON ST"      2303
##  5 A1       "NEW SUDBURY ST"   2082
##  6 A1       "ATLANTIC AVE"     1269
##  7 A1       "STUART ST"        1067
##  8 A1       "STATE ST"          990
```

```
##  9 A1       "SUMMER ST"          972
## 10 A1       "CHARLES ST"         969
## # ... with 5,695 more rows
```

```
b1 %>%
  ggplot(aes(x = countk))+
  geom_histogram(bins = 50, fill = "red", color = "black")+
  scale_x_continuous(limits = c(0,1000))+
  scale_y_continuous(limits = c(0,150))+
  facet_wrap(~ DISTRICT, ncol = 3)+
  theme_get()+
  ggtitle("Crime Reported per Streets in Boston, MA")+
  ylab("No. of Streets")+
  xlab("No. of Reported Crimes")
```

```
## Warning: Removed 64 rows containing non-finite values (`stat_bin()`).
```

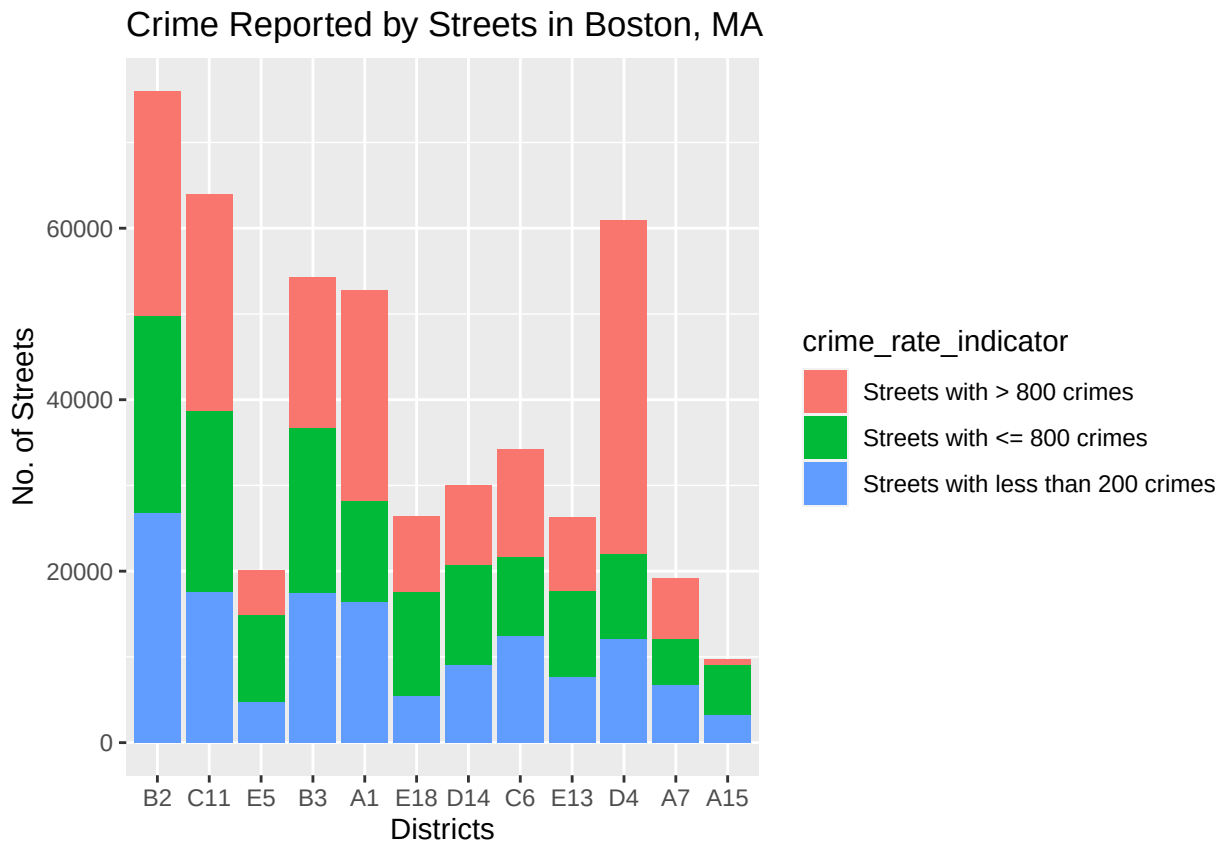
```
## Warning: Removed 26 rows containing missing values (`geom_bar()`).
```



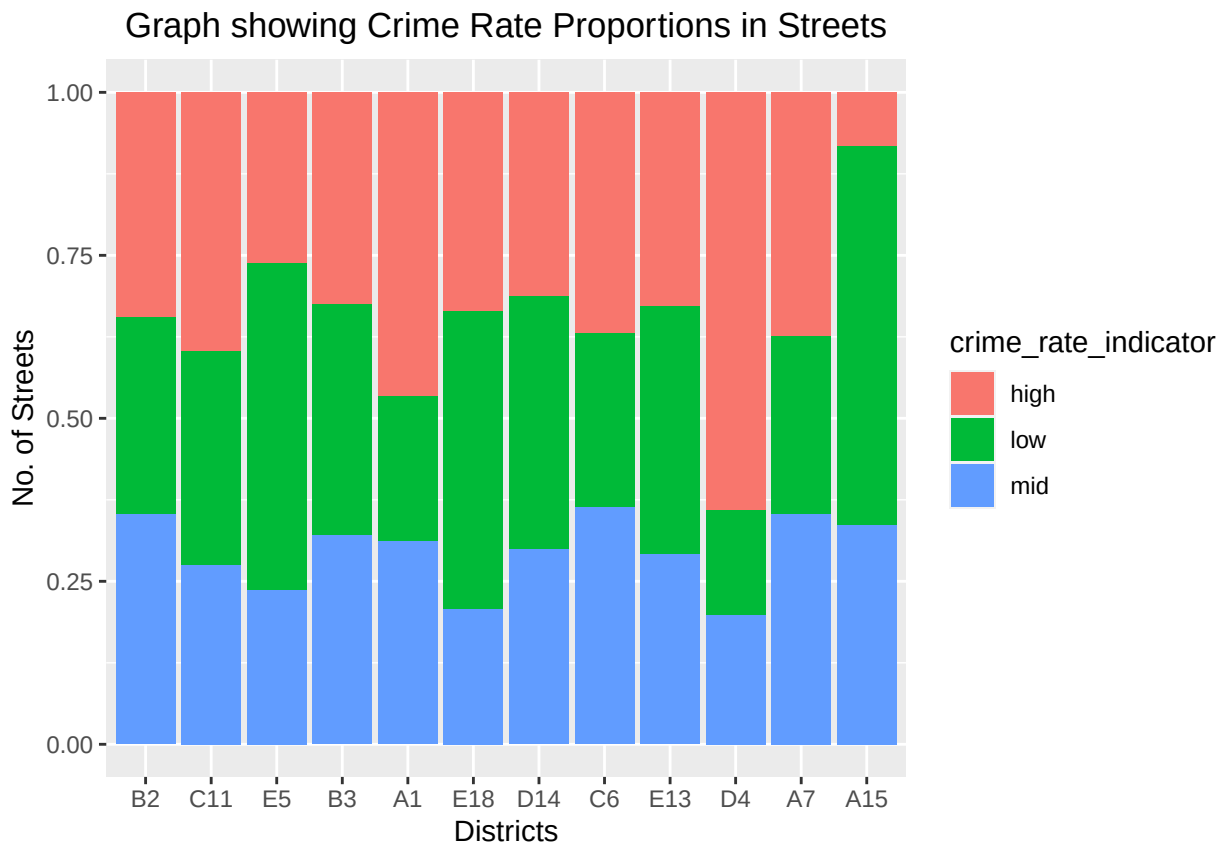
Majority of the streets had reported less than 250 crimes during the period 2015 - 2020.

```
b2 <- b1 %>%
  filter(DISTRICT != "External") %>%
  mutate(crime_rate_indicator = case_when(
    countk > 800 ~ "high",
```

```
    countk > 200 & countk <= 800 ~ "mid",
    countk >0 & countk <= 200 ~ "low"
  ))

b2 %>%
  ggplot(aes(x=fct_infreq(DISTRICT), y = countk, fill = crime_rate_indicator))+
  geom_bar(stat = "identity")+
  scale_fill_discrete(labels = c("Streets with > 800 crimes", "Streets with <= 800 crimes", "Streets wi
  ggtitle("Crime Reported by Streets in Boston, MA")+
  ylab("No. of Streets")+
  xlab("Districts")
```




```
ggplot(b2, (aes(x=fct_infreq(DISTRICT), y = countk, fill = crime_rate_indicator)))+
geom_bar(position = "fill", stat = "identity")+
ggtitle("Graph showing Crime Rate Proportions in Streets")+
theme(plot.title = element_text(hjust = 0.5))+
  ylab("No. of Streets")+
  xlab("Districts")
```



District D4 has the most number of streets with more than 800 reported crimes over the course of four years (2015 - 2020).

UCR Part One Crimes in Boston, MA

```
boston_crime_df %>%
  filter(UCR_PART == 'Part One') %>%
  group_by(DISTRICT) %>%
  summarise(
    count_of_partone = n()
  ) %>%
  arrange(desc(count_of_partone))
```

```
## # A tibble: 13 x 2
##    DISTRICT count_of_partone
##    <fct>               <int>
##  1 "D4"                15906
##  2 "B2"                11401
##  3 "A1"                11236
##  4 "C11"                9171
##  5 "B3"                 6756
##  6 "C6"                 6005
##  7 "D14"                5020
##  8 "E13"                4674
##  9 "E18"                3404
```

```
## 10 "A7"                 2729
## 11 "E5"                 2680
## 12 "A15"                1557
## 13 ""                    244
```

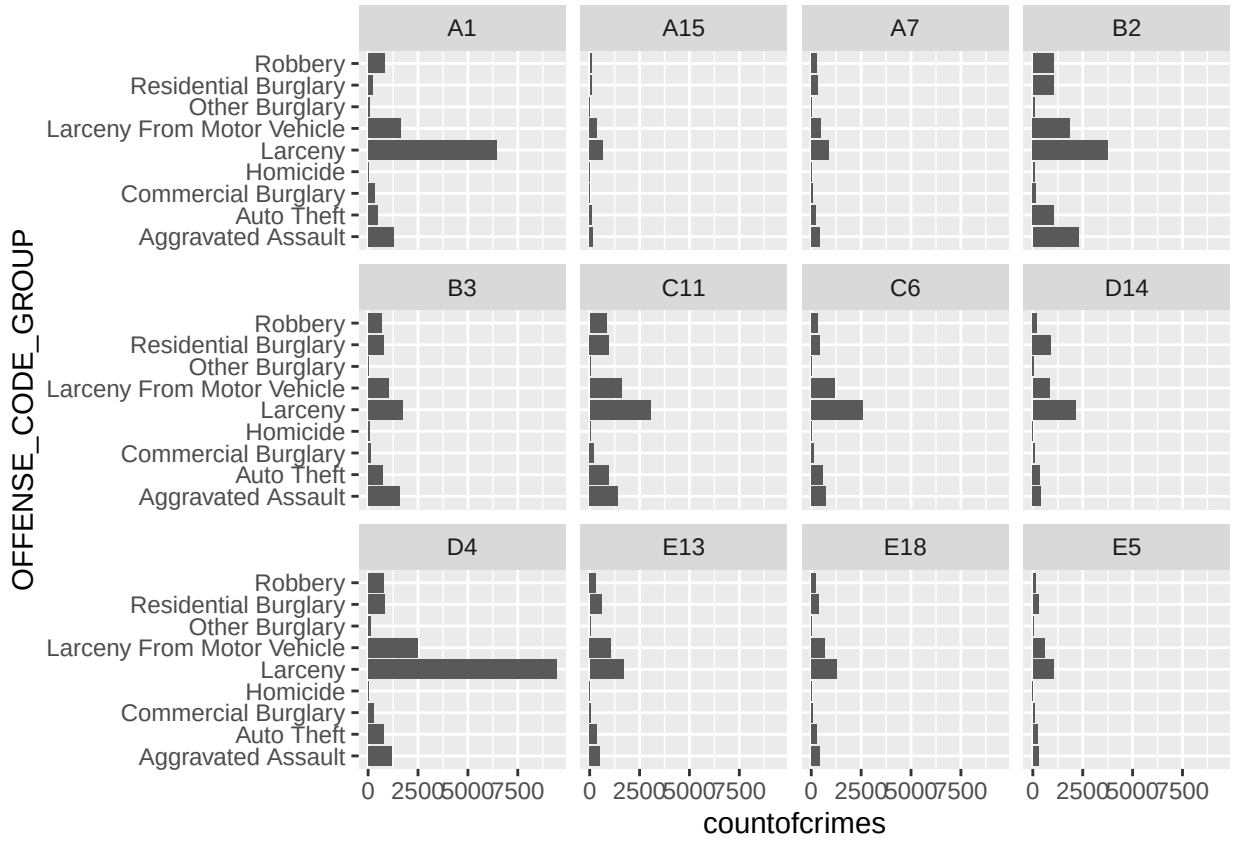
```
#most frequent Part One crimes by street

b3 <- boston_crime_df %>%
  filter(UCR_PART == "Part One") %>%
  filter(DISTRICT != "") %>%
  group_by(DISTRICT, OFFENSE_CODE_GROUP) %>%
  summarise(
    countofcrimes = n()
  )%>%
  arrange(DISTRICT, desc(countofcrimes))
```

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

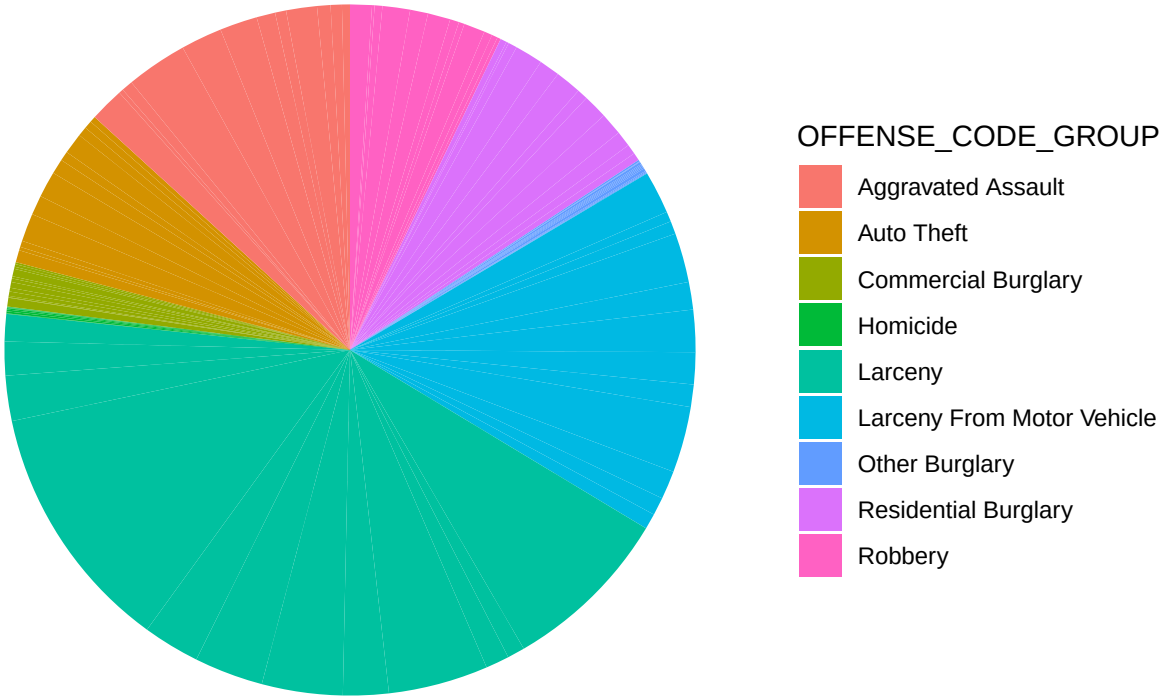
```
b3
```

```
## # A tibble: 108 x 3
## # Groups:   DISTRICT [12]
##    DISTRICT OFFENSE_CODE_GROUP         countofcrimes
##    <fct>    <fct>                              <int>
##  1 A1       Larceny                             6430
##  2 A1       Larceny From Motor Vehicle          1607
##  3 A1       Aggravated Assault                  1287
##  4 A1       Robbery                              820
##  5 A1       Auto Theft                           459
##  6 A1       Commercial Burglary                  328
##  7 A1       Residential Burglary                 238
##  8 A1       Other Burglary                        60
##  9 A1       Homicide                               7
## 10 A15      Larceny                              656
## # ... with 98 more rows
```

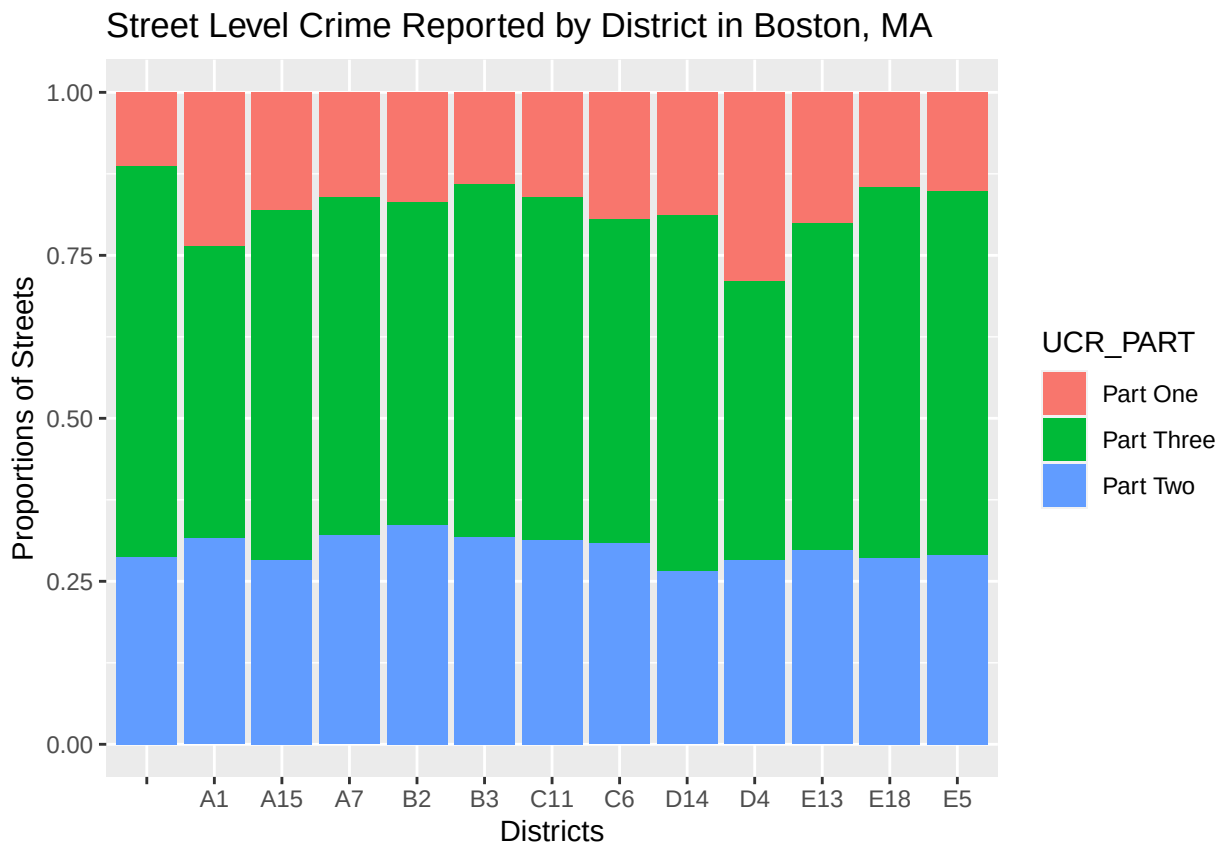
```
b3 %>%
  ggplot(aes(x = OFFENSE_CODE_GROUP, y = countofcrimes))+
  geom_bar(stat = "identity")+
  coord_flip()+
  facet_wrap(~DISTRICT)
```




```
# pie chart showing the proportion of part one crimes in Boston from 2016 - 2020.

b4 <- b3 %>%
  group_by(OFFENSE_CODE_GROUP) %>%
  arrange(desc(OFFENSE_CODE_GROUP)) %>%
  mutate(prop = countofcrimes / sum(b3$countofcrimes) *100) %>%
  mutate(ypos = cumsum(prop)- 1*prop )

ggplot(b4, aes(x="", y=countofcrimes, fill=OFFENSE_CODE_GROUP)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)+
  theme_void()
```




```
boston_crime_df %>%
  filter(UCR_PART != "" & UCR_PART != "Other")%>%
  ggplot(aes(x = DISTRICT, fill = UCR_PART))+
  geom_bar(position = "fill", stat = "count")+
  ggtitle("Street Level Crime Reported by District in Boston, MA")+
  ylab("Proportions of Streets")+
  xlab("Districts")
```



Finding most dangeorus streets in each district.

```
street_dangerous <-boston_crime_df %>%
  group_by(DISTRICT, STREET, UCR_PART) %>%
  summarise(
    street_crime = n()
  ) %>%
  filter(street_crime > 800 & UCR_PART == "Part One")
```

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

```
street_dangerous
```

```
## # A tibble: 9 x 4
## # Groups:   DISTRICT, STREET [9]
##   DISTRICT STREET            UCR_PART street_crime
##   <fct>    <fct>             <fct>           <int>
## 1 A1       WASHINGTON ST     Part One         1493
## 2 B2       BLUE HILL AVE     Part One          840
## 3 B3       BLUE HILL AVE     Part One         1000
## 4 C11      DORCHESTER AVE    Part One         1128
## 5 D4       BOYLSTON ST       Part One         3112
```

```
## 6 D4       HARRISON AVE      Part One          891
## 7 D4       HUNTINGTON AVE    Part One          857
## 8 D4       MASSACHUSETTS AVE Part One          867
## 9 D4       NEWBURY ST        Part One         1412
```

```
boston_crime_df %>%
  group_by(DISTRICT, STREET, UCR_PART) %>%
  summarise(
    street_crime = n()
  ) %>%
  filter(street_crime > 800 & UCR_PART == "Part Two")
```

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

```
## # A tibble: 17 x 4
## # Groups:   DISTRICT, STREET [17]
##    DISTRICT STREET            UCR_PART street_crime
##    <fct>    <fct>             <fct>           <int>
##  1 A1       BOYLSTON ST       Part Two          999
##  2 A1       TREMONT ST        Part Two         1222
##  3 A1       WASHINGTON ST     Part Two         1144
##  4 B2       BLUE HILL AVE     Part Two         1161
##  5 B2       DUDLEY ST         Part Two         1224
##  6 B2       WARREN ST         Part Two          812
##  7 B2       WASHINGTON ST     Part Two         1358
##  8 B3       BLUE HILL AVE     Part Two         2566
##  9 C11      DORCHESTER AVE    Part Two         1867
## 10 C11      WASHINGTON ST     Part Two         1076
## 11 D14      COMMONWEALTH AVE  Part Two          886
## 12 D4       BOYLSTON ST       Part Two         1575
## 13 D4       HARRISON AVE      Part Two         1642
## 14 D4       MASSACHUSETTS AVE Part Two         1028
## 15 E13      CENTRE ST         Part Two          929
## 16 E13      WASHINGTON ST     Part Two          873
## 17 E18      HYDE PARK AVE     Part Two         1151
```

```
shoot <- boston_crime_df %>%
  filter(SHOOTING == "1" | SHOOTING == "Y") %>%
  group_by(DISTRICT, STREET) %>%
  summarise(
    shoot_count = n()
  ) %>%
  filter(shoot_count > 20) %>%
  arrange(DISTRICT, desc(shoot_count))
```

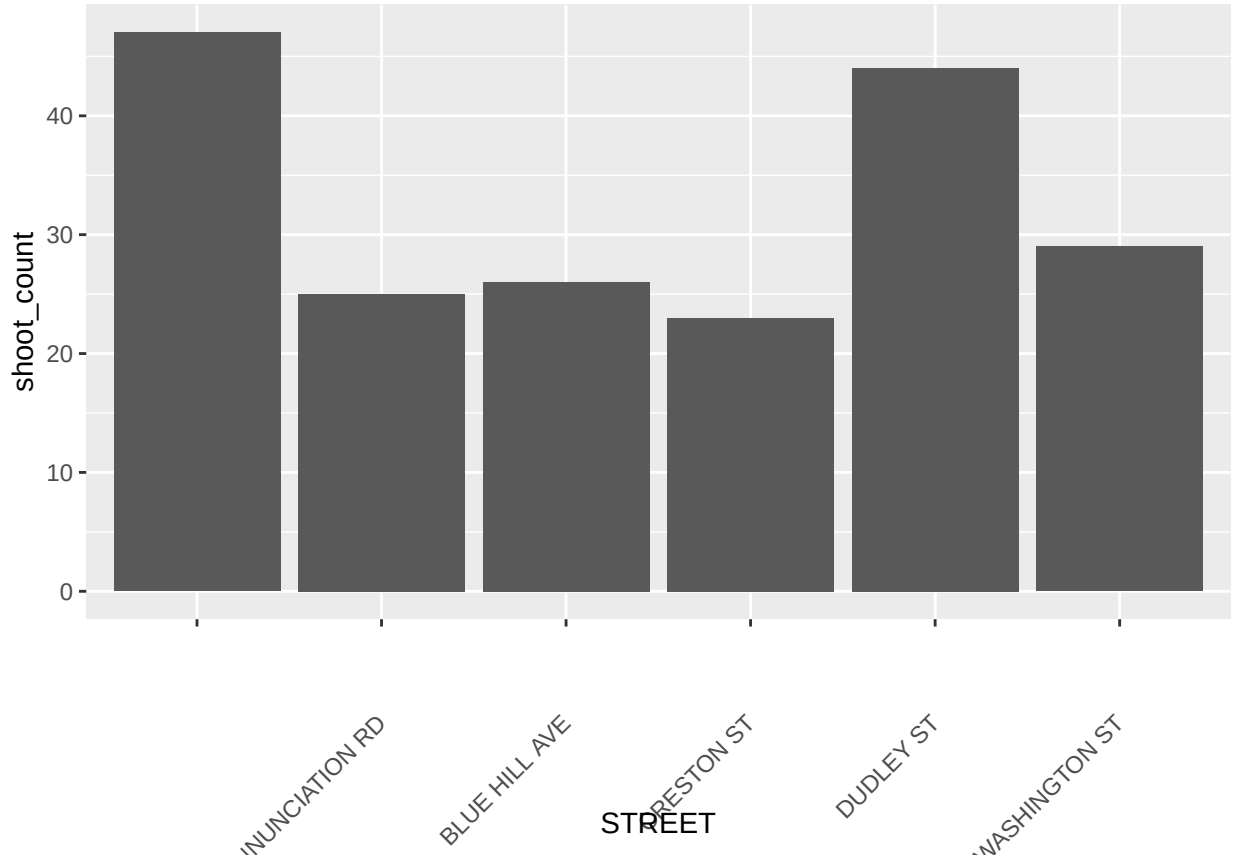
```
## `summarise()` has grouped output by 'DISTRICT'. You can override using the
## `.groups` argument.
```

```
shoot
```

```
## # A tibble: 11 x 3
```

```
## # Groups:   DISTRICT [4]
##    DISTRICT STREET            shoot_count
##    <fct>    <fct>                   <int>
##  1 B2       ""                         47
##  2 B2       "DUDLEY ST"                44
##  3 B2       "WASHINGTON ST"            29
##  4 B2       "BLUE HILL AVE"            26
##  5 B2       "ANNUNCIATION RD"          25
##  6 B2       "CRESTON ST"               23
##  7 B3       "BLUE HILL AVE"            33
##  8 B3       ""                         28
##  9 C11      "WASHINGTON ST"            36
## 10 C11      "DORCHESTER AVE"           23
## 11 E13      "CENTRE ST"                27
```

```
shoot %>%
  filter(DISTRICT == "B2") %>%
  filter(shoot_count > 20) %>%
  ggplot(aes(x = STREET, y = shoot_count))+
  geom_bar(stat = "identity")+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```

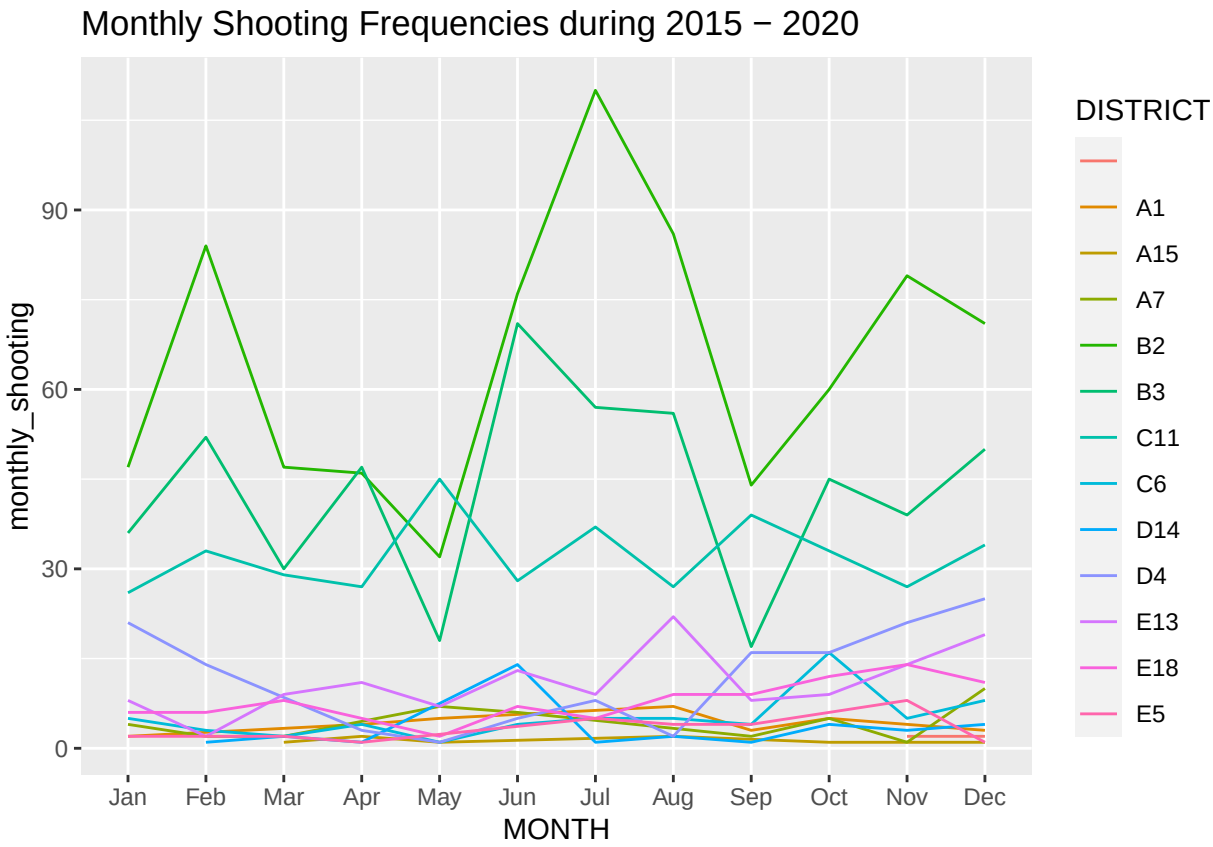



```
shoot %>%
  filter(DISTRICT == "C11") %>%
  filter(shoot_count > 20) %>%
```

```
  ggplot(aes(x = STREET, y = shoot_count))+
  geom_bar(stat = "identity")
```


```
boston_crime_df %>%
  filter(SHOOTING == "1" | SHOOTING == "Y") %>%
  group_by(DISTRICT, MONTH) %>%
  summarise(
    monthly_shooting = n()
  ) %>%
  ggplot(aes(x = MONTH, y = monthly_shooting, col = DISTRICT))+
  geom_line()+
  scale_x_discrete(limits = month.abb)+
  ggtitle("Monthly Shooting Frequencies during 2015 - 2020")
```

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


Monthly Shooting Frequencies during 2015 − 2020
monthly_shooting
90
60
30
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
MONTH
DISTRICT
A1
A15
A7
B2
B3
C11
C6
D14
D4
E13
E18
E5