project

Alan- Kevin

11/1/2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
##load df
crime <- read_csv(file = "Arrests 2017 Public.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Arrest Year` = col_double(),
##   `Arrest Date` = col_date(format = ""),
##   `Arrest Hour` = col_double(),
##   Age = col_double(),
##   `Offense GEOY` = col_double(),
##   `Offense GEOX` = col_double(),
##   `Offense PSA` = col_double(),
##   `Arrest Latitude` = col_double(),
##   `Arrest Longitude` = col_double(),
##   `Offense Latitude` = col_double(),
##   `Offense Longitude` = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 58 parsing failures.
##  row         col expected  actual                      file
## 1581 Offense PSA a double #N/A    'Arrests 2017 Public.csv'
## 2505 Offense PSA a double UNKNOWN 'Arrests 2017 Public.csv'
## 2654 Offense PSA a double UNKNOWN 'Arrests 2017 Public.csv'
## 3763 Offense PSA a double #N/A    'Arrests 2017 Public.csv'
## 3818 Offense PSA a double #N/A    'Arrests 2017 Public.csv'
## .... ........... ........ ....... ........................
## See problems(...) for more details.
```

```r
head(crime)
```

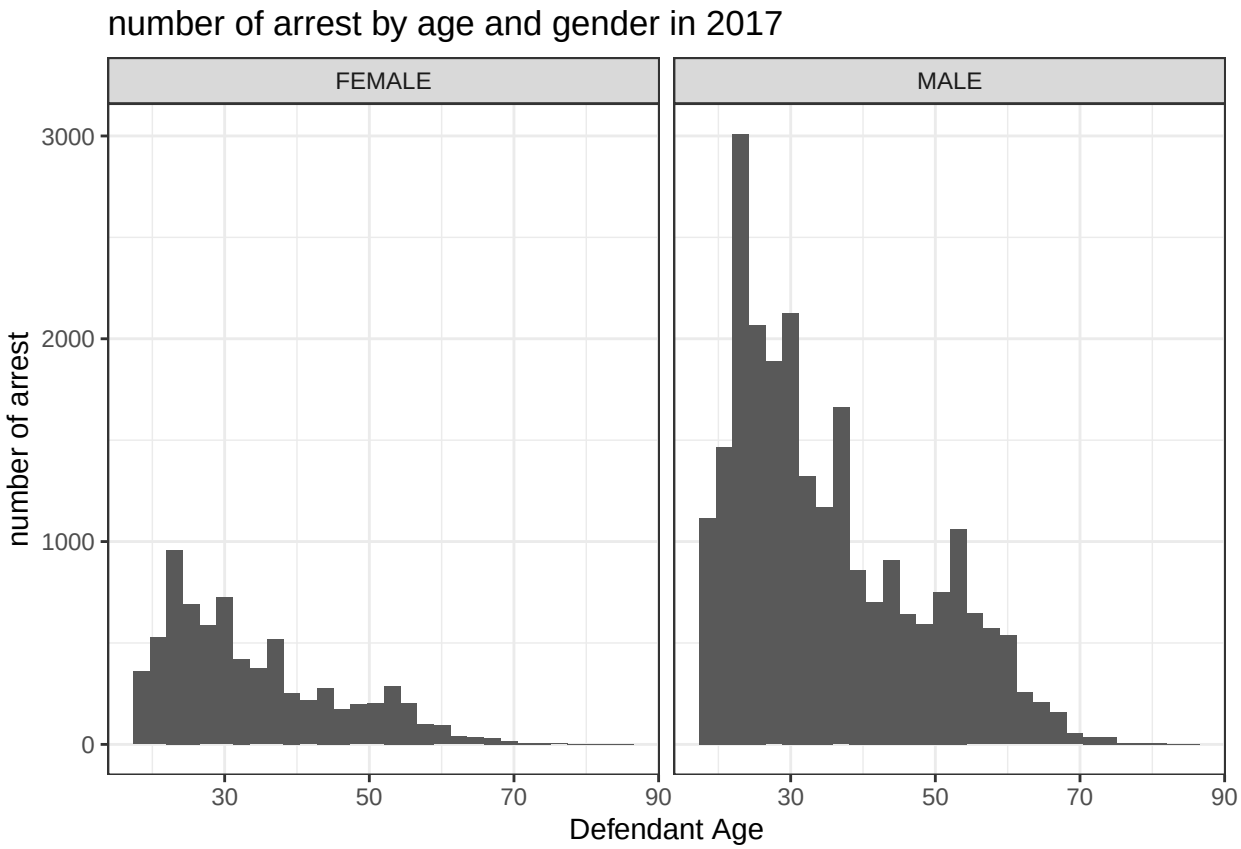
```
## # A tibble: 6 x 26
##   `Arrestee Type` `Arrest Year` `Arrest Date` `Arrest Hour` CCN
##   <chr>                   <dbl> <date>                <dbl> <chr>
## 1 Adult Arrest             2017 2017-01-01                0 ce59~
## 2 Adult Arrest             2017 2017-01-01                0 bbe5~
## 3 Adult Arrest             2017 2017-01-01                0 1a6a~
## 4 Adult Arrest             2017 2017-01-01                0 bbe5~
## 5 Adult Arrest             2017 2017-01-01                0 bbe5~
## 6 Adult Arrest             2017 2017-01-01                0 dc33~
## # ... with 21 more variables: `Arrest Number#` <chr>, Age <dbl>,
## #   `Defendant PSA` <chr>, `Defendant District` <chr>, `Defendant
## #   Race` <chr>, `Defendant Ethnicity` <chr>, `Defendant Sex` <chr>,
## #   `Arrest Category` <chr>, `Charge Description` <chr>, `Arrest Location
## #   PSA` <chr>, `Arrest Location District` <chr>, `Arrest Location Block
## #   GeoX` <chr>, `Arrest Location Block GeoY` <chr>, `Offense GEOY` <dbl>,
## #   `Offense GEOX` <dbl>, `Offense PSA` <dbl>, `Offense District` <chr>,
## #   `Arrest Latitude` <dbl>, `Arrest Longitude` <dbl>, `Offense
## #   Latitude` <dbl>, `Offense Longitude` <dbl>
```

1. There is a association between defendant profile (age, race, ethnicity and/or sex, state address) and type of arrest category.

- There was more offense by male than by female; most defendant age is between 20-30 (graph a)
- There was more offense by Black than white and Asian. (graph b)
- Between White, there was not much difference for hispanic and non hispanic (Graph c)
- Simple Assault, Release Violations/Fugitive and Traffic Violations were the most occuring offense, either for male and female (table i and table ii) ## Graph a

```r
crime %>%
  filter(`Defendant Sex` != "UNK") %>%
  ggplot(aes(Age)) +
  geom_histogram() +
  theme_bw() +
  facet_wrap(~`Defendant Sex`) +
  xlab("Defendant Age") + ylab("number of arrest") + ggtitle("number of arrest by age and gender in 201
```

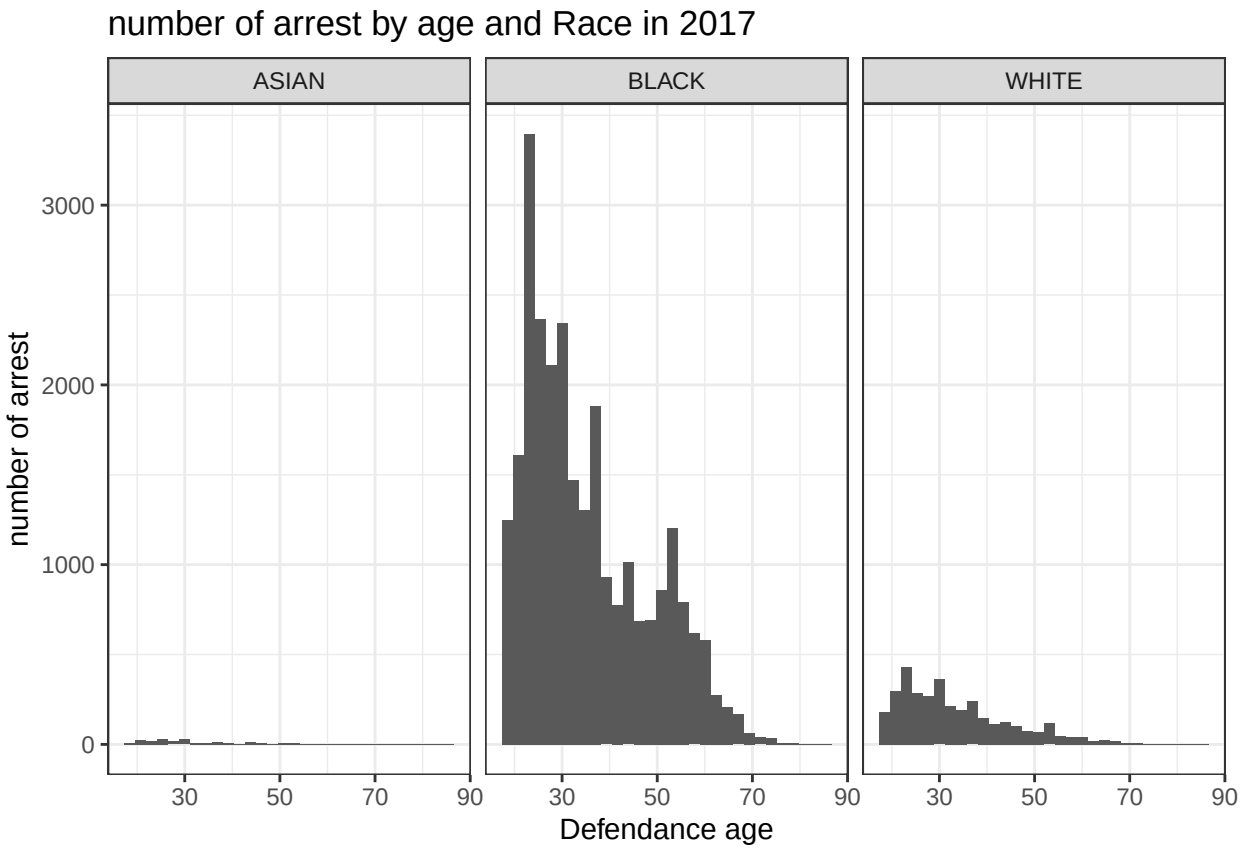
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Graph b

```
crime %>%
  filter(`Defendant Race` != "UNK") %>%
  ggplot(aes(Age)) +
  geom_histogram() +
  theme_bw() +
  facet_wrap(~`Defendant Race`) +
  xlab("Defendance age") + ylab("number of arrest") + ggtitle("number of arrest by age and Race in 2017
```

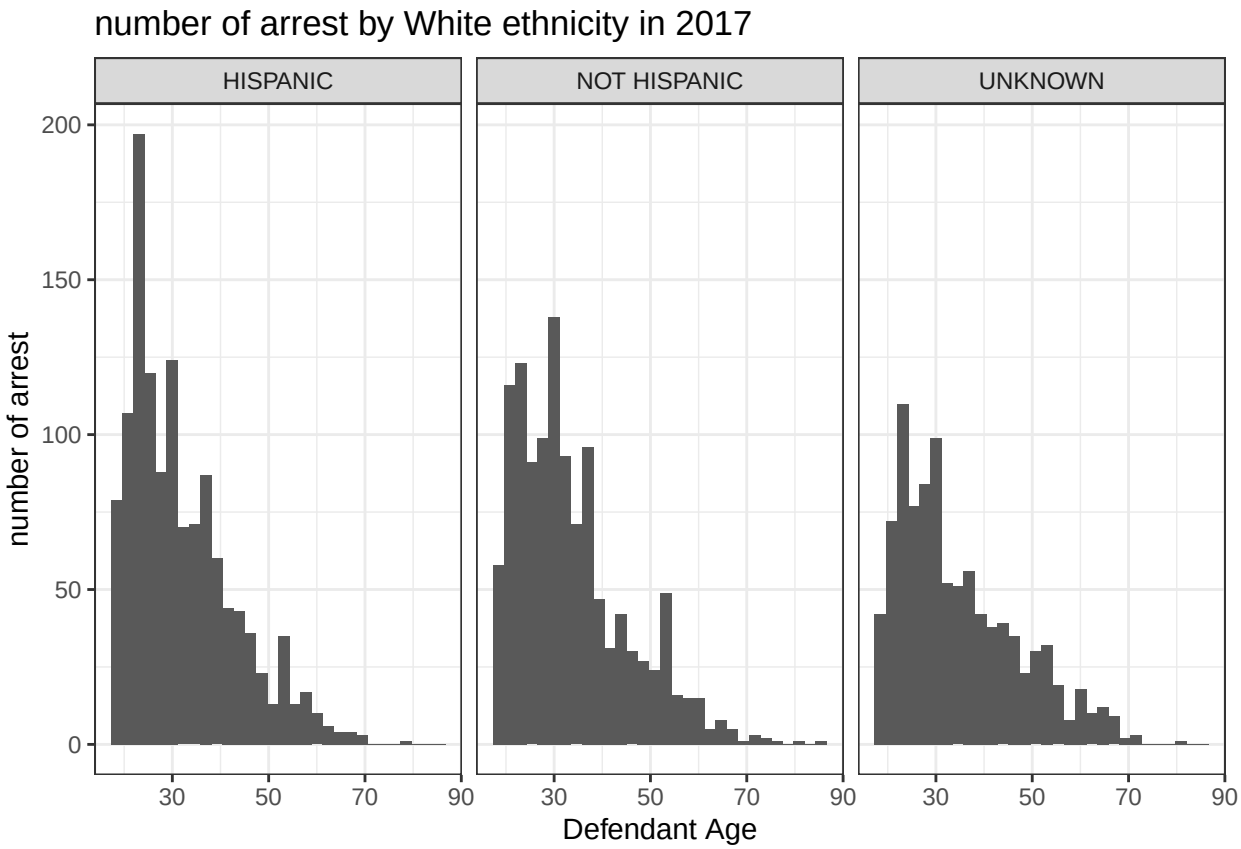
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Graph c

```
crime %>%
  filter(`Defendant Race` == "WHITE") %>%
  ggplot(aes(Age)) +
  geom_histogram() +
  theme_bw() +
  facet_wrap(~`Defendant Ethnicity`) +
  xlab("Defendant Age") + ylab("number of arrest") + ggtitle("number of arrest by White ethnicity in 20
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



##table i and table ii

```
crime_by_sex <- crime %>%
  filter(`Defendant Sex` != "UNK") %>%
  group_by(`Arrest Category`,`Defendant Sex`) %>%
  tally () %>%
  ungroup() %>%
  spread(key = `Defendant Sex`, value = n)


##table i - top 3 female offense
crime_by_sex %>%
  select(1,2) %>%
  arrange(desc(FEMALE)) %>%
  slice(1:3)
```

```
## # A tibble: 3 x 2
##   `Arrest Category`            FEMALE
##   <chr>                         <int>
## 1 Simple Assault                 2124
## 2 Traffic Violations              959
## 3 Release Violations/Fugitive     882
```

```
##table ii - top 3 male offense
crime_by_sex %>%
  filter( `Arrest Category` != "Other Crimes") %>%
```

```
  select(1,3) %>%
  arrange(desc(MALE)) %>%
  slice(1:3)
```

```
## # A tibble: 3 x 2
##   `Arrest Category`            MALE
##   <chr>                       <int>
## 1 Simple Assault               4072
## 2 Release Violations/Fugitive  3615
## 3 Traffic Violations           3509
```

2. There are area which more offence and/or arrest than other.

- Most offense occured in center and south part of DC (map a)
- there was no difference between offense location and arrest location (map b)

```
library(sf) #package to handling shape file
```

```
## Linking to GEOS 3.7.2, GDAL 2.4.2, PROJ 5.2.0
```

```
library(viridis) ##color pallete for fill
```

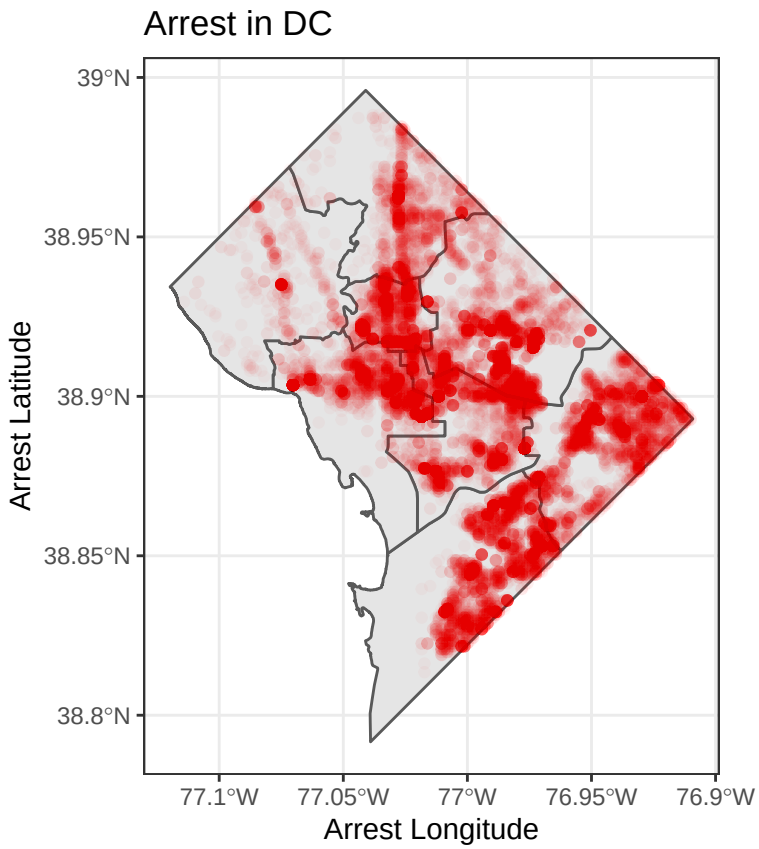
```
## Loading required package: viridisLite
```

```
map <- read_sf("Ward_from_2012.shp") #dc map by ward from dc open data
class(map) #check type file
```

```
## [1] "sf"         "tbl_df"     "tbl"        "data.frame"
```

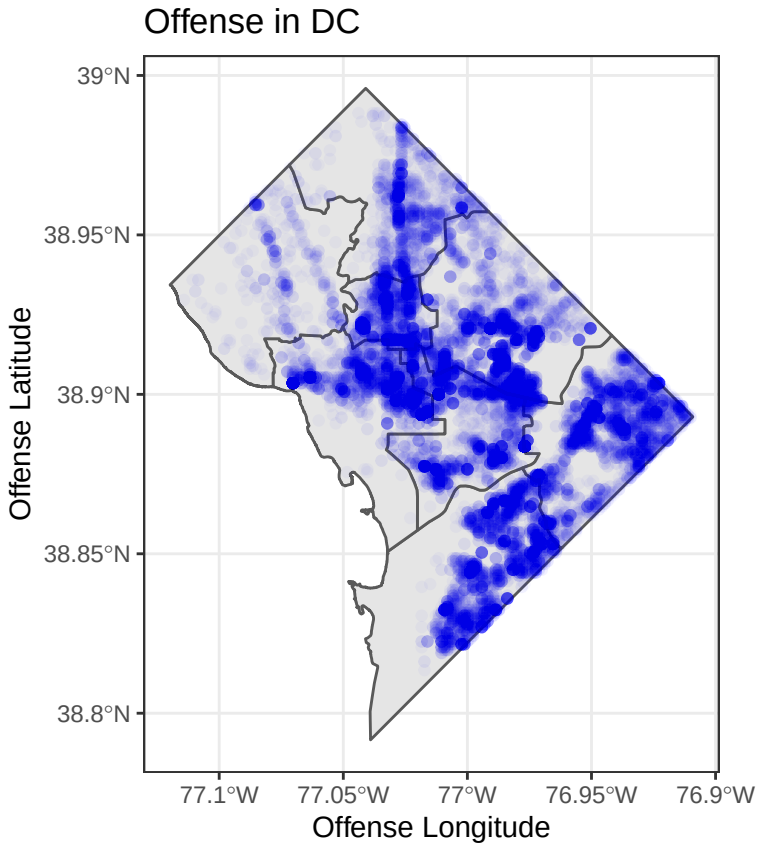
Map a

```
#map laying
plotmap <- ggplot(map) +
  geom_sf(aes()) +
  theme_bw()
crime2 <- na.omit(crime)
plotmap +
  geom_point(data = crime2, aes(`Arrest Longitude`, `Arrest Latitude`), alpha = 0.03, colour= "red") +
  ggtitle("Arrest in DC") ##arrest location as xy point
```



map b

```
plotmap +
  geom_point(data = crime2, aes(`Offense Longitude`, `Offense Latitude`), alpha = 0.03, colour= "blue")
  ggtitle("Offense in DC")  ## offense location as xy point
```



3. There are PSA had more arrest than other.

- PSA number 102, 507, 506, 603, and 602 were the PSA with the most number of offenses (table iii)
- Most offense occured in center and south east PSA in DC. Notably, few crimes occured in north part of DC (map c)
-

table iii

```
psa <- read_sf("Police_Service_Areas.shp") ## data from dc open data
psa_crime <- crime %>%
  group_by(`Offense PSA`) %>%
  tally() %>%
  arrange(desc(n)) %>%
  rename(PSA = "Offense PSA") ##count offense by PSA

psa_crime %>%
  slice(1:5) %>%
  rename(Top_5_offense_location = PSA)
```

```
## # A tibble: 5 x 2
```

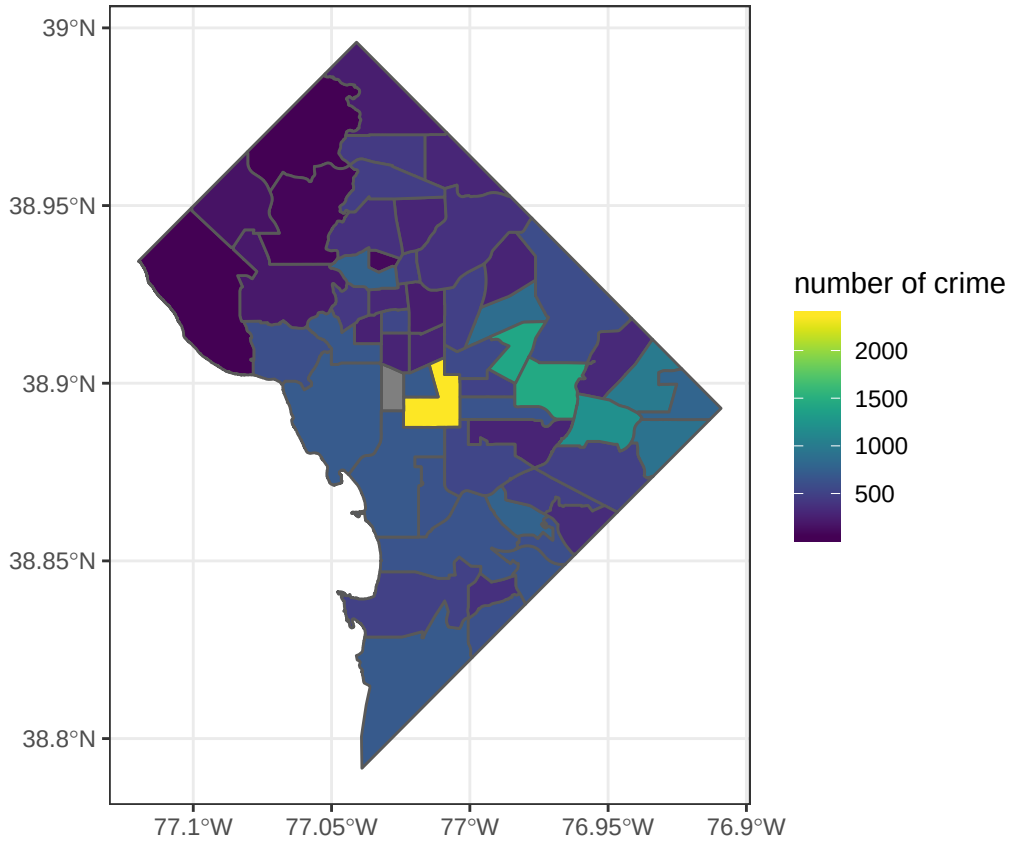
```
##   Top_5_offense_location     n
##                    <dbl> <int>
## 1                    102  2353
## 2                    507  1442
## 3                    506  1420
## 4                    603  1203
## 5                    602   993
```

map c

```
psa <- left_join(psa,psa_crime)
```

```
## Joining, by = "PSA"
```

```
plotmap2 <- ggplot(psa) +
  geom_sf(aes(fill = n)) +
  scale_fill_viridis("number of crime") +
  theme_bw()
plotmap2
```



4. There is association between type of arrest/offense and location.

- by mapping most number of offense in each PSA, we know that Simple assault are the most offense occured in most of PSA location
- Narcotic related offense was "popular" in center part of DC, while traffic offense was "most popular" in south part of DC

map d

```
df <- crime %>%
  group_by(`Offense PSA`,`Arrest Category`) %>%
  tally() %>%
  mutate(the_rank  = rank(-n, ties.method = "random")) %>%
  filter(the_rank == 1) %>%
  rename(PSA = `Offense PSA`)
crime_type <- left_join(psa,df, by = "PSA")

ggplot(crime_type) +
  geom_sf(aes(fill = `Arrest Category`)) +
  scale_fill_viridis("top arrest category by PSA", discrete = TRUE) +
  theme_bw()
```