project

Alan- Kevin

11/1/2019

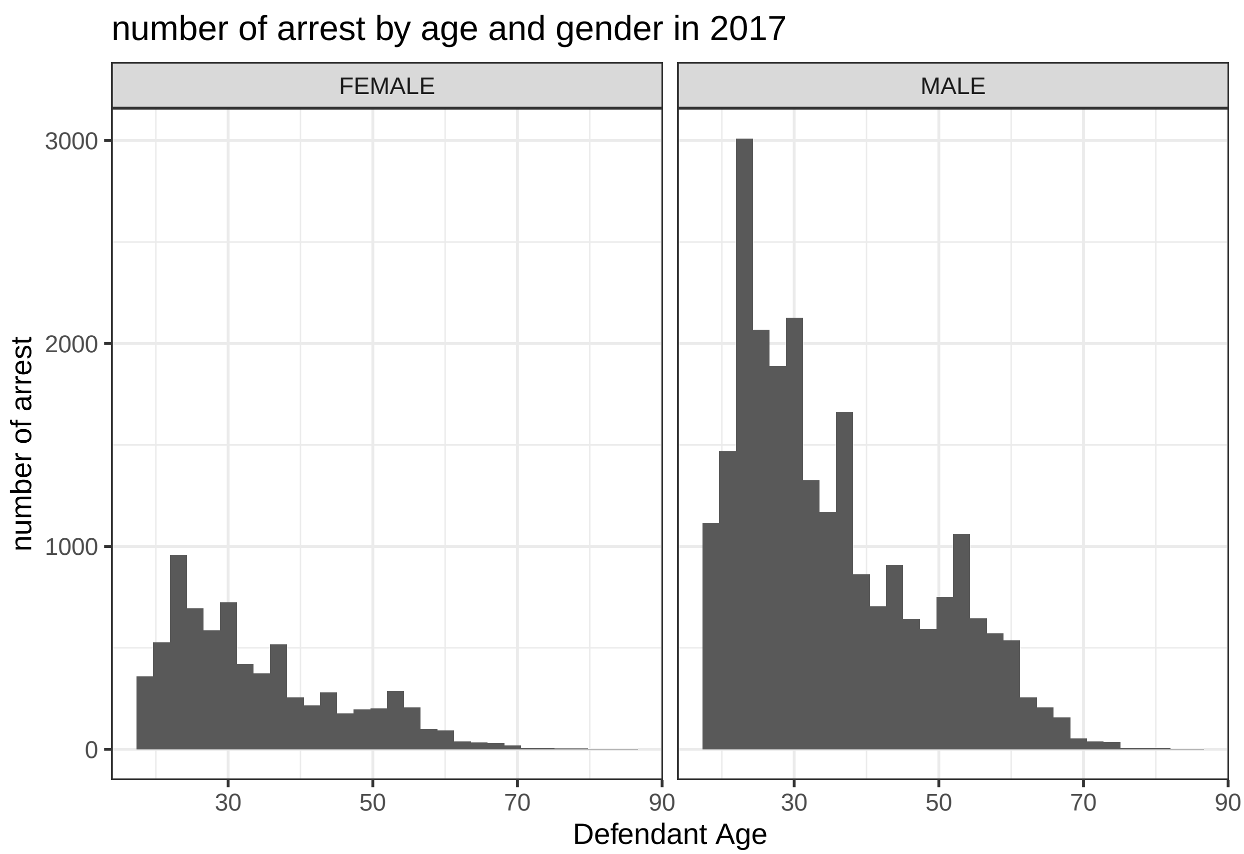
1. There is a association between defendant profile (age, race, ethnicity and/or sex, state address) and type of arrest category.

- 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.
- It were also the most occuring offense for all races (table i and table ii)
- We create new crime category income motivated crime and non income motivated crime, there are no difference in distribution with overall crime. (Graph d)

Conclusion: 1.From crime distribution, here is a association between defendant profile (age, race and sex) and number of crime. 2. No association between between defendant profile (age, race, ethnicity and/or sex) and type of arrest category.

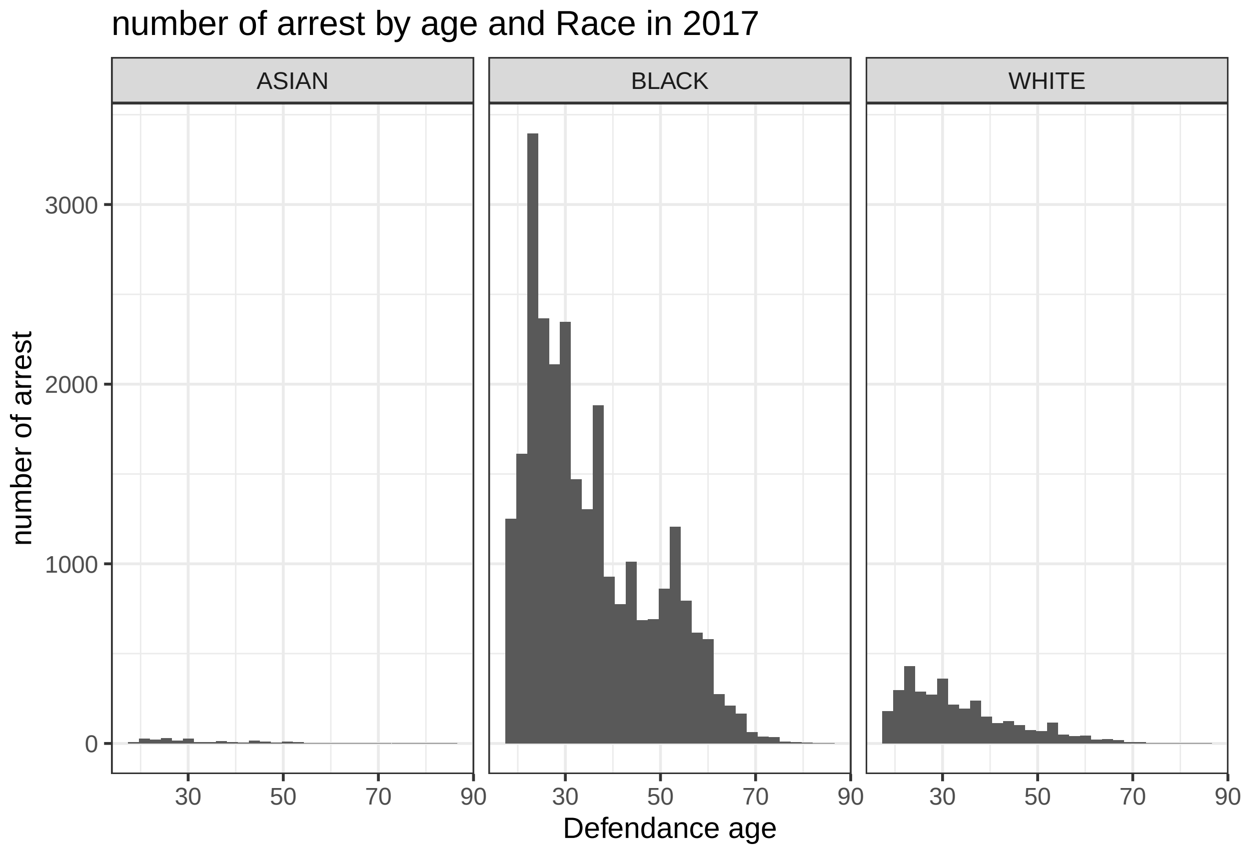
Graph a

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



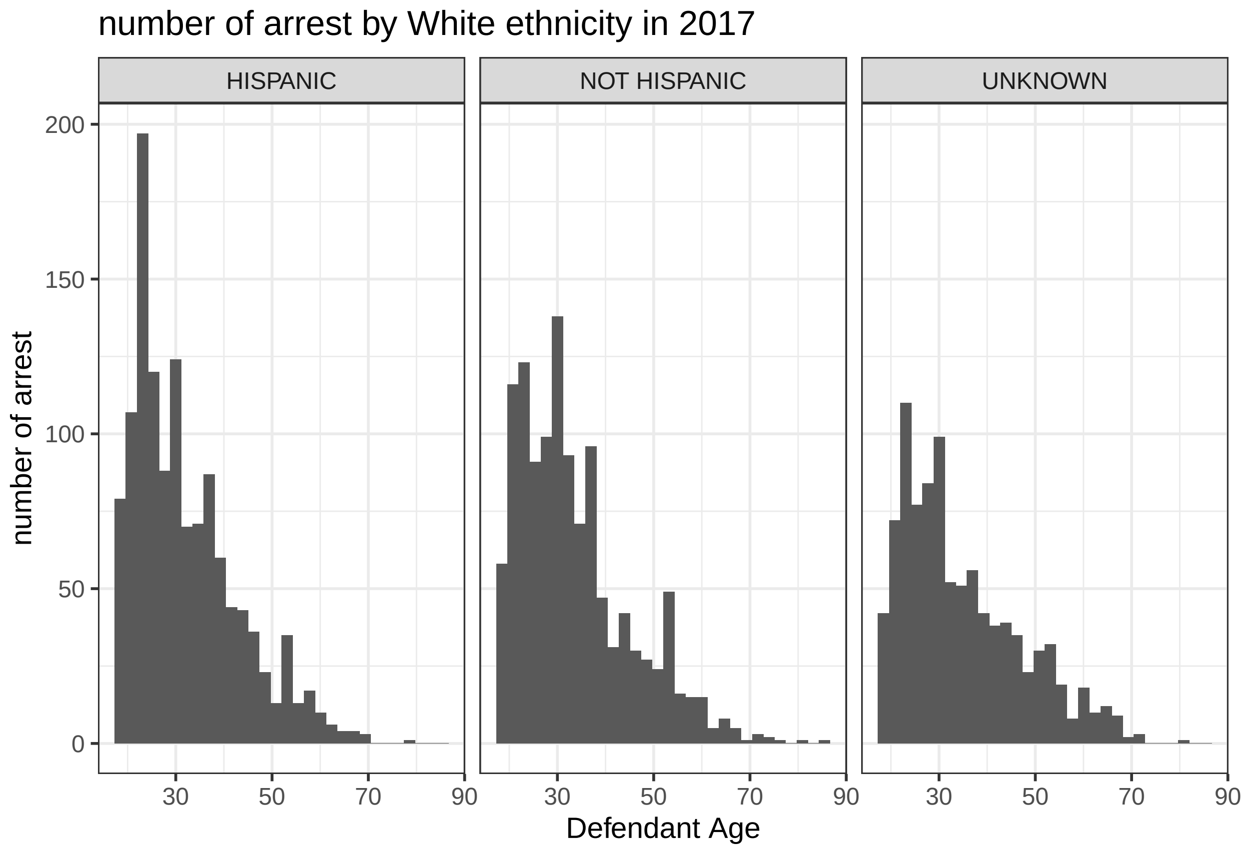
Graph b

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



Graph c

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



##table i

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

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

##table ii

```
## # A tibble: 29 x 4
##    `Arrest Category`                  ASIAN BLACK WHITE
##    <chr>                              <int> <int> <int>
##  1 Aggravated Assault                     3   142    18
##  2 Arson                                 NA     4    NA
##  3 Assault on a Police Officer            1   358    59
##  4 Assault with a Dangerous Weapon        2   768    69
```

```
##  5 Burglary                              3   212    21
##  6 Damage to Property                    4   701    91
##  7 Disorderly Conduct                    9   430   274
##  8 Driving/Boating While Intoxicated    15   972   256
##  9 Fraud and Financial Crimes            7   102    26
## 10 Gambling                             NA    94    NA
## # ... with 19 more rows
```

```
## # A tibble: 3 x 2
##   `Arrest Category`  ASIAN
##   <chr>              <int>
## 1 Simple Assault        41
## 2 Theft                 29
## 3 Traffic Violations    23
```

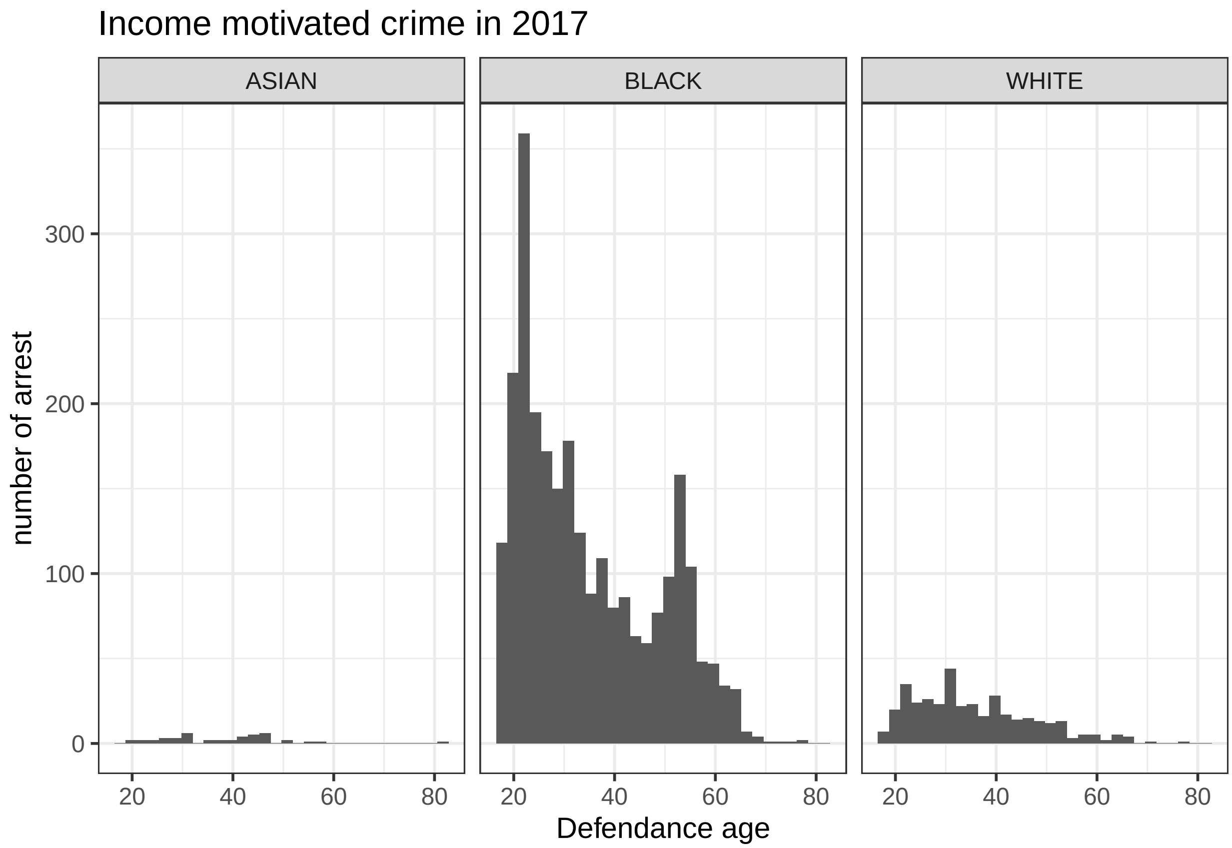
```
## # A tibble: 3 x 2
##   `Arrest Category`           BLACK
##   <chr>                       <int>
## 1 Simple Assault               5312
## 2 Release Violations/Fugitive  4071
## 3 Traffic Violations           3879
```

```
## # A tibble: 3 x 2
##   `Arrest Category`           WHITE
##   <chr>                       <int>
## 1 Simple Assault                640
## 2 Traffic Violations            394
## 3 Release Violations/Fugitive   359
```

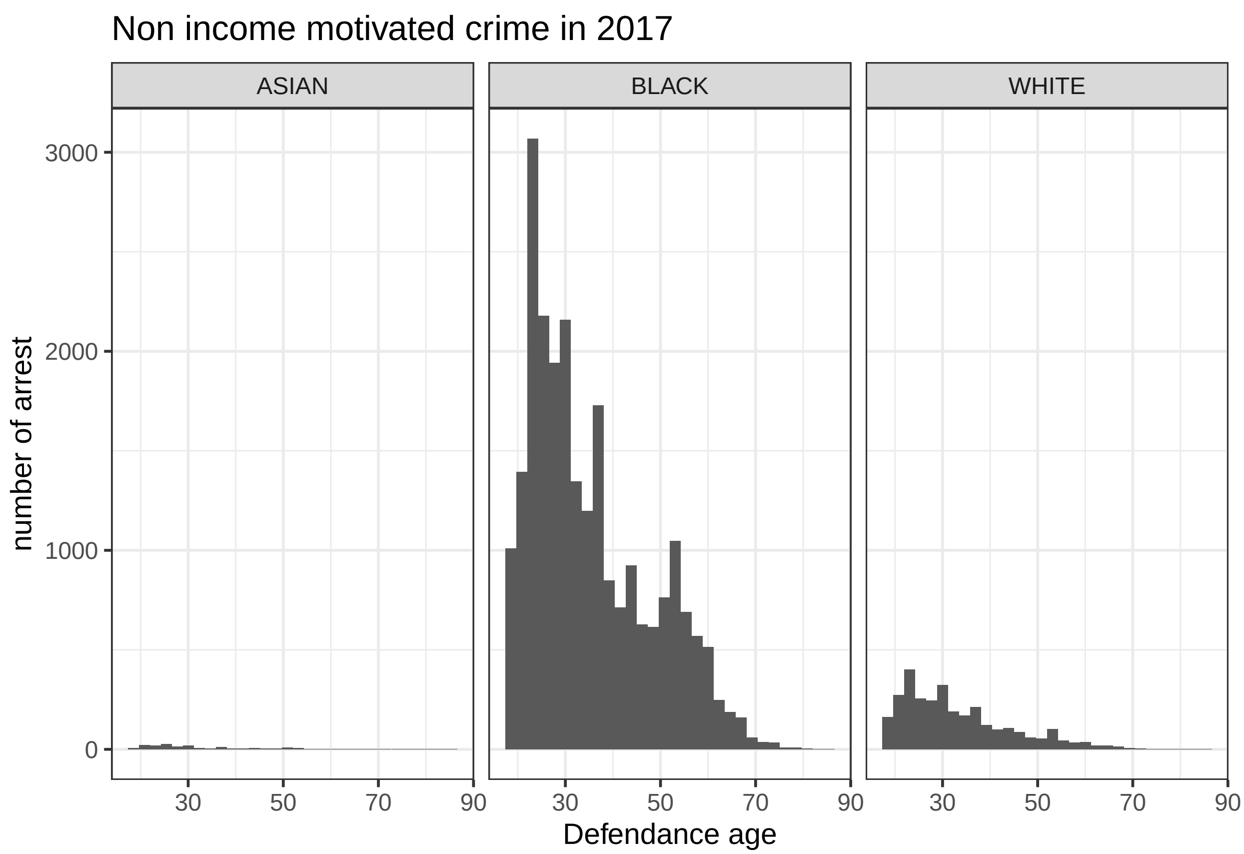
Graph D

income related crime

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




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

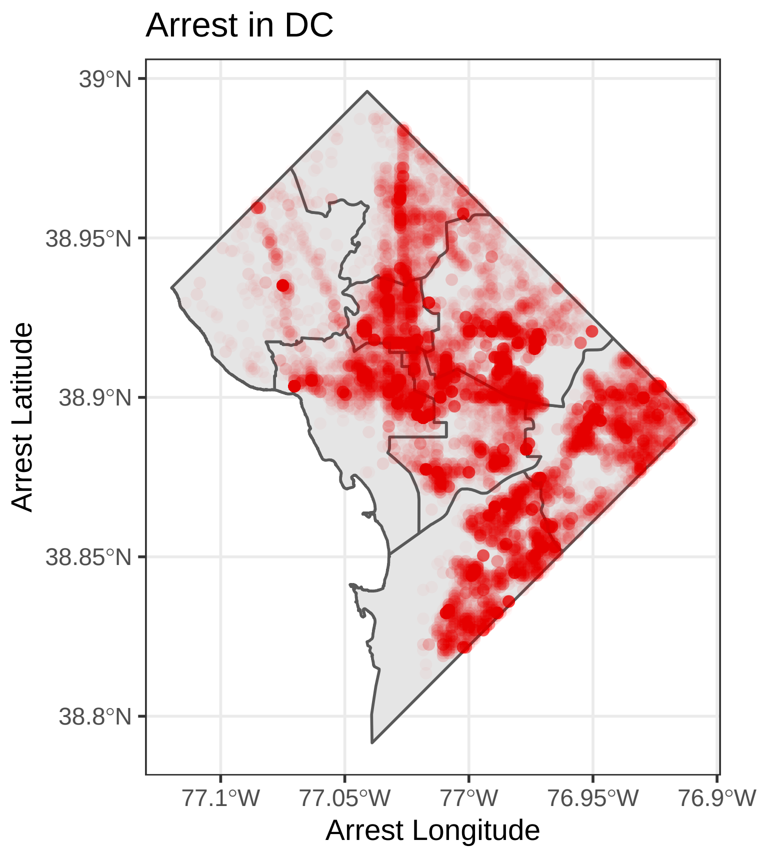


2. There are area which more offence and/or arrest than other. (*I think since PSA is defined area, we can make 2 and 3 as one hypotesis)

- Most offense occured in center and south part of DC (map a)
- there was no difference between offense location and arrest location (map b)
- 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 PSA in DC. Notably, few crimes occured in north part of DC (map c)

Conclusion : Most offense occured in center and south part of DC. - There are area which more offence and/or arrest than other.

Map a


Arrest in DC
39°N
38.95°N
38.9°N
38.85°N
38.8°N
Arrest Latitude
77.1°W
77.05°W
77°W
76.95°W
76.9°W
Arrest Longitude

map b

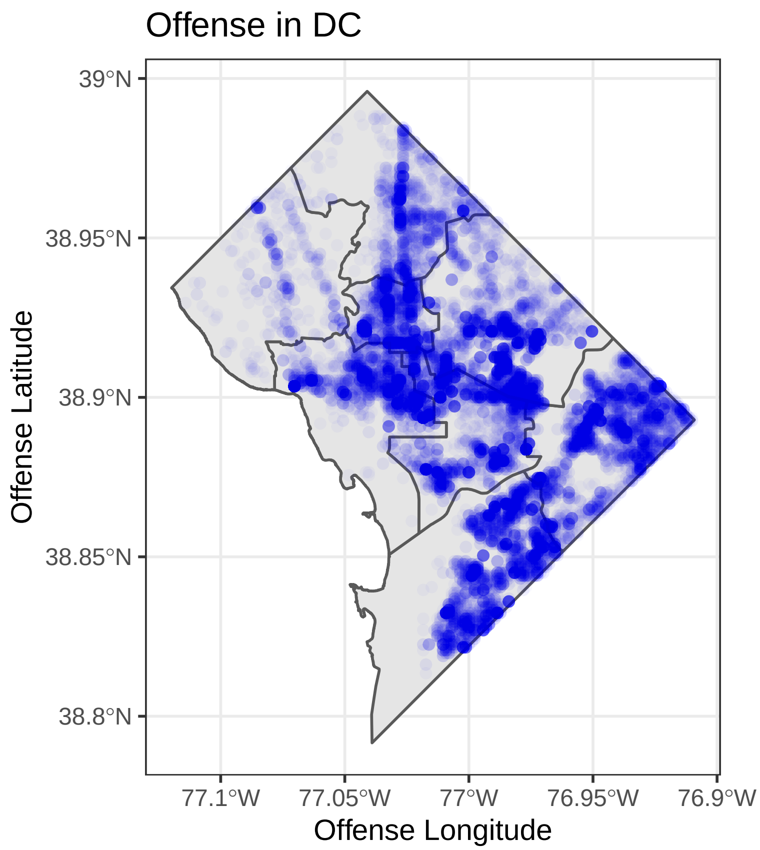
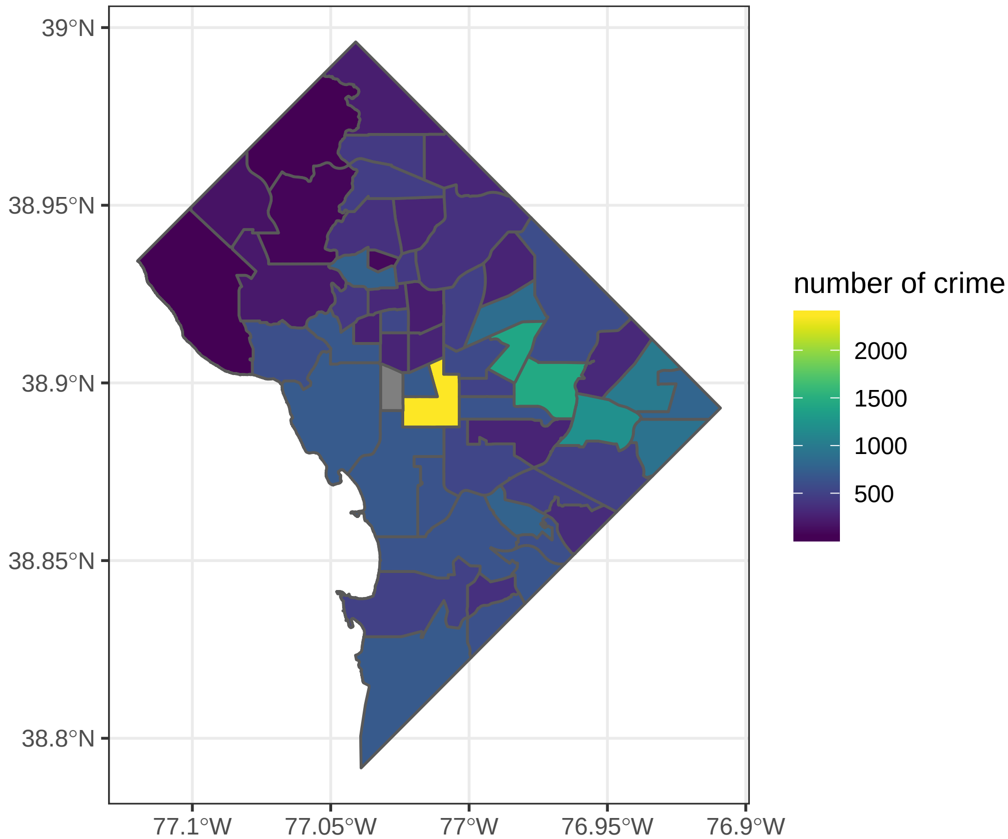


table iii

```
## # 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

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

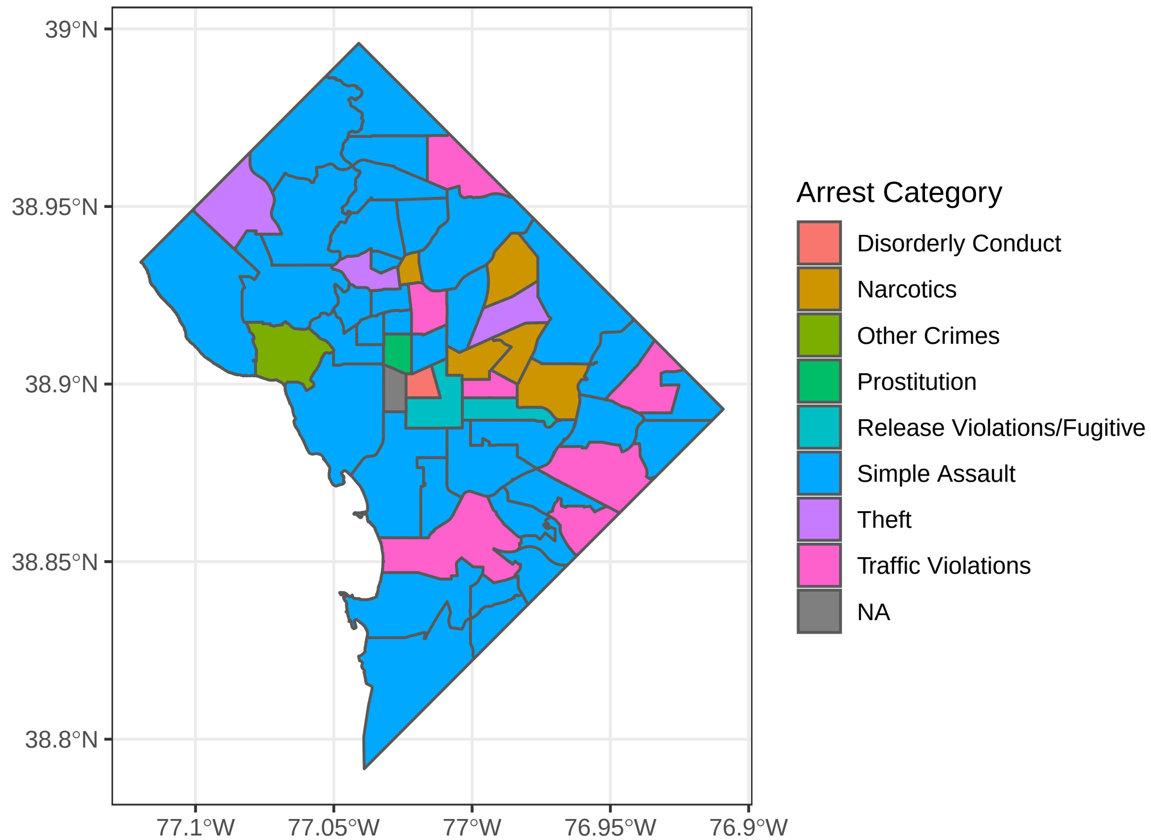


3. There is an association between type of offense and location.

- 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.
- Prostitution was the most occured offense in PSA 307, which cover much of the area surrounding Logan Circle. https://www.borderstan.com/2012/02/01/closer-look-at-psa-307s-new-boundaries-personnel/
- Release Violation/Fugitive was was the most occured offense in PSA 102 and 108. Probably because there is a court in those area, Most Release Violation/Fugitive occurs when people did not show at court.

Conclusion : There is an association between type of offense and location. Certain crime occured more frequently in some areas than in other areas.

map d



Income related crime, with addition to ACS data Hypothesis: 1. The type of housing within an area contributed to the location that crime occurs, suggesting that most crime occurs within the neighborhood the perpetrator lives (John Hipp, Young-An Kim, and Kevin Kane) 2. There is a relationship between number of crime and income.

- Only 34.7% ffense committed by defendant happen within their own PSA, where for income related crime, only 13.93% offense committed by defendant happen within their own PSA.(table iv)
- There is a negative relationship between median income in a ward (from ACS data) with the number of crime. The higher median income in a ward, the lower number of offense occured.

table iv

```
## # A tibble: 3 x 2
##   same_location   sum
##   <chr>         <int>
## 1 N             23123
## 2 Y              8028
## 3 <NA>             58
```

```
## # A tibble: 3 x 2
##   same_location   sum
##   <chr>         <int>
## 1 N              2735
```

```
## 2 Y               381
## 3 <NA>              6
```

#graph e

```r
## if any code should be displayed in power point, we think this is the most challenging code since ACS
ACS <- read_csv("612 project/economic.csv")
```

```
## Warning: Duplicated column names deduplicated: 'Estimate' =>
## 'Estimate_1' [4], 'Percent' => 'Percent_1' [5], 'Estimate' =>
## 'Estimate_2' [6], 'Percent' => 'Percent_2' [7], 'Estimate' =>
## 'Estimate_3' [8], 'Percent' => 'Percent_3' [9], 'Estimate' =>
## 'Estimate_4' [10], 'Percent' => 'Percent_4' [11], 'Estimate' =>
## 'Estimate_5' [12], 'Percent' => 'Percent_5' [13], 'Estimate' =>
## 'Estimate_6' [14], 'Percent' => 'Percent_6' [15], 'Estimate' =>
## 'Estimate_7' [16], 'Percent' => 'Percent_7' [17]
```

```
## Parsed with column specification:
## cols(
##   desc = col_character(),
##   Estimate = col_character(),
##   Percent = col_character(),
##   Estimate_1 = col_character(),
##   Percent_1 = col_character(),
##   Estimate_2 = col_character(),
##   Percent_2 = col_character(),
##   Estimate_3 = col_character(),
##   Percent_3 = col_character(),
##   Estimate_4 = col_character(),
##   Percent_4 = col_character(),
##   Estimate_5 = col_character(),
##   Percent_5 = col_character(),
##   Estimate_6 = col_character(),
##   Percent_6 = col_character(),
##   Estimate_7 = col_character(),
##   Percent_7 = col_character()
## )
```

```r
ACS <- na.omit(ACS)
ACS <- ACS %>%
  gather(-desc, key = "key", value = "value") %>%
  separate(key, into = c("key","ward"), sep = "_") %>%
  mutate(ward = recode(ward,
                      `1` = "2",
                      `2` = "3",
                      `3` = "4",
                      `4` = "5",
                      `5` = "6",
                      `6` = "7",
                      `7` = "8", .missing = "1"))
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 274 rows [1,
## 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```r
ACS <- ACS %>%
  filter(str_detect(desc,"Median household income"),
         key == "Estimate") %>%
  mutate(value = str_replace_all(value,",",""),
         value = parse_number(value))
point <- as.data.frame(cbind(crime$`Offense Longitude`,crime$`Offense Latitude`))
point <-  na.omit(point)
point <- st_as_sf(point, coords = c("V1", "V2"))
st_crs(map)
```

```
## Coordinate Reference System:
##   EPSG: 4326
##   proj4string: "+proj=longlat +datum=WGS84 +no_defs"
```

```r
point <- st_sf(point, crs = "+init=epsg:4326")

intersect <- st_intersection(x = map, y = point)
```

```
## although coordinates are longitude/latitude, st_intersection assumes that they are planar
```

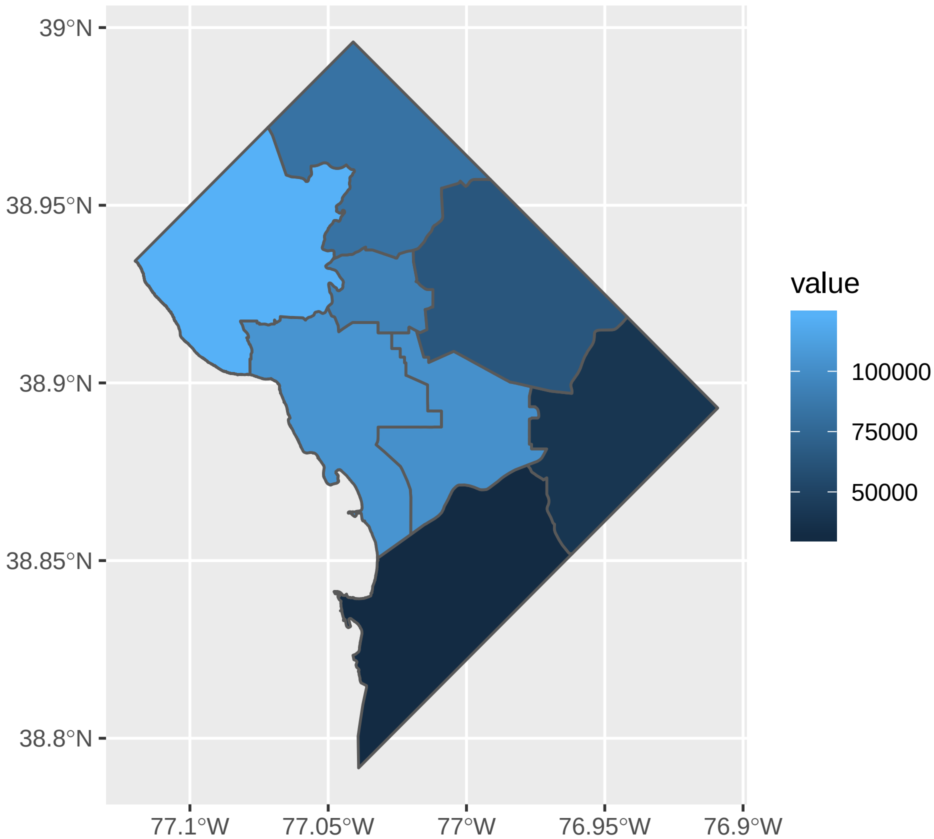
```
## Warning: attribute variables are assumed to be spatially constant
## throughout all geometries
```

```r
intersect <- intersect %>%
  group_by(NAME) %>%
  tally()
intersect <- intersect %>%
  separate(NAME, into= c("ward1", "ward"))
map <- map %>%
  separate(NAME, into= c("ward1", "ward"), sep = " ") %>%
  left_join(ACS, by = "ward") %>%
  st_join(intersect, by = "ward")
```

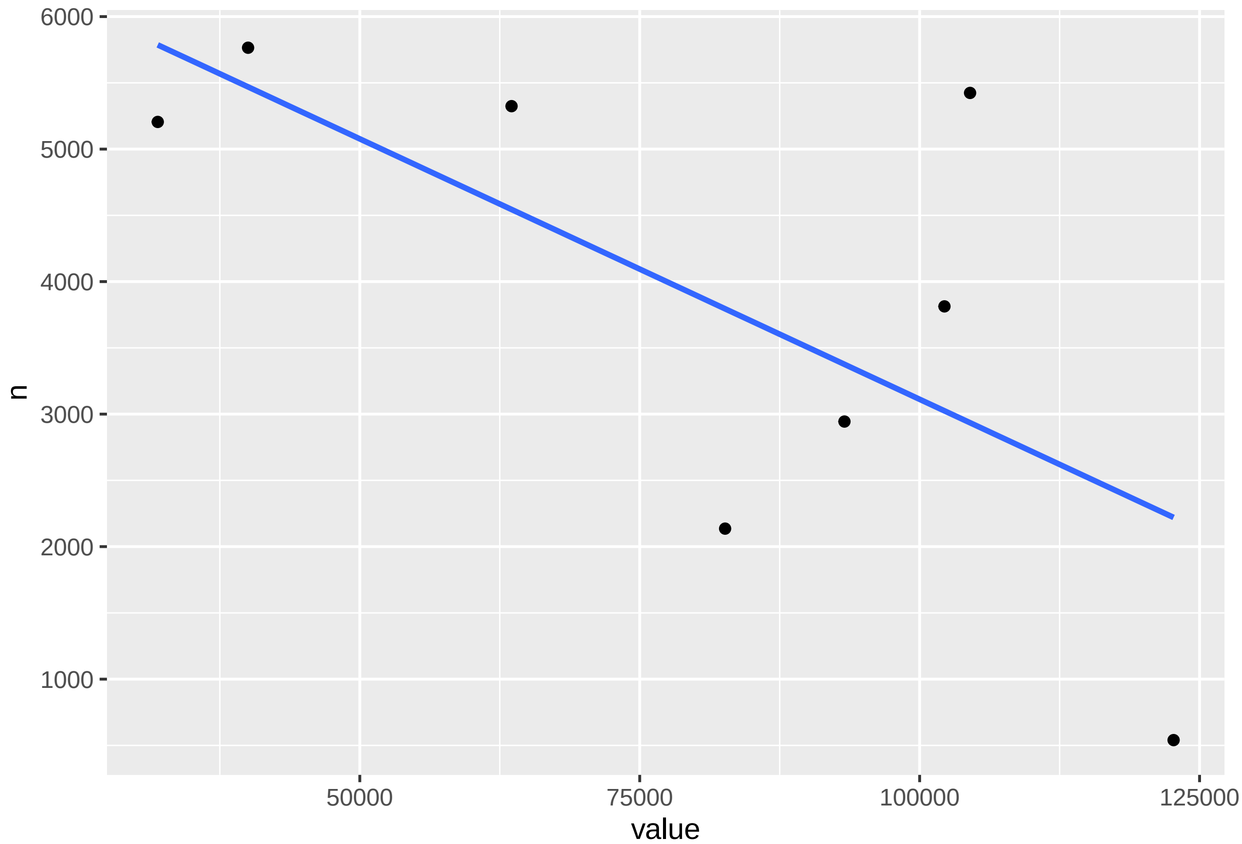
```
## although coordinates are longitude/latitude, st_intersects assumes that they are planar
```

```r
##south area in DC had lower median income

ggplot(map) + geom_sf(aes(fill = value))
```




```
crime_income_pw <- map %>%
  select(ward.x, value,n) %>%
  arrange(desc(value,n))
ggplot(crime_income_pw, aes(value,n)) + geom_point()+ geom_smooth(method= lm, se=FALSE)
```




```
##negative relatinship between median income and number of crime.
```