# Assignment 04: Analysis of crime in Washington, DC

AUTHOR
Jacob Ausubel

PUBLISHED
October 8, 2023

# Link to dataset

https://opendata.dc.gov/datasets/DCGIS::crime-incidents-in-2021/about

# Topic and motivation

In this analysis, I demonstrate that violent crime is much more common in some parts of Washington, DC than in others. The motivation for the assignment is an uptick in crime in the city in recent years that has received substantial media attention. Lawmakers and policymakers need to keep variation in homicides, sex abuses, assaults with dangerous weapons, and robberies between wards and neighborhood clusters in mind when designing programs aimed at curbing violent crime.

The dataset I'm using comes from Open Data DC and includes all of the reported crimes that happened in Washington, DC in 2021. I'm focusing on just the crimes that are categorized as violent crimes. (An explanation of the definition is provided later on.)

# Initial data setup, cleaning, and manipulation

```
#Reading in packages
library(tidyverse)
library(readxl)
library(ggthemes)
```

```
#Removing datasets from working directory
rm(list = ls())
```

```
#Reading in crime dataset
crime21 <- read.csv("Crime_Incidents_in_2021.csv")

#Creating violent crime variable
#1=Yes, 0=No
crime21$violent_crime <-
  ifelse(crime21$OFFENSE == "HOMICIDE", 1,
    ifelse(crime21$OFFENSE == "SEX ABUSE", 1,
      ifelse(crime21$OFFENSE == "ASSAULT W/DANGEROUS WEAPON", 1
        ifelse(crime21$OFFENSE == "ROBBERY", 1, 0))))
```

```
#Creating a data frame with 8 rows (one for each ward)
#Column for number of violent crimes
violent_crime_by_ward <- crime21 %>%
  filter(violent_crime == 1) %>%
  group_by(WARD) %>%
  summarise(count = n()) %>%
  filter(!is.na(WARD))
```

```
#Creating a data frame with 32 rows
#Want the counts for violent crimes in each ward broken
  #down by offense type
#8 wards x 4 offense types = 32
violent_crime_by_ward2 <- crime21 %>%
  filter(violent_crime == 1) %>%
  group_by(WARD, OFFENSE) %>%
  summarise(count = n()) %>%
  filter(!is.na(WARD))

#Recoding offense variable
violent_crime_by_ward2$OFFENSE[
  violent_crime_by_ward2$OFFENSE == "ASSAULT W/DANGEROUS WEAPON
  ] <- "Assault w/dangerous weapon"
violent_crime_by_ward2$OFFENSE[
  violent_crime_by_ward2$OFFENSE == "HOMICIDE"
] <- "Homicide"
violent_crime_by_ward2$OFFENSE[
  violent_crime_by_ward2$OFFENSE == "ROBBERY"
] <- "Robbery"
violent_crime_by_ward2$OFFENSE[
  violent_crime_by_ward2$OFFENSE == "SEX ABUSE"
] <- "Sex Abuse"
```

```
#Creating a data frame with 32 rows
#Want the counts for violent crimes in each
  #neighborhood cluster broken
  #down by offense type
#Note that there are 44 neighborhood clusters.
  #Why 154 rows, not 176?
  #After all, 44 x 4 = 176
  #Reason: in some neighborhood clusters, there were
    #0 cases of a specific offense.
violent_crime_by_neigh_cluster <- crime21 %>%
  filter(violent_crime == 1) %>%
  group_by(NEIGHBORHOOD_CLUSTER, OFFENSE) %>%
  summarise(count = n()) %>%
  filter(!is.na(NEIGHBORHOOD_CLUSTER))
```

```
#Creating a new data frame with 44 rows
#Reason: I want a row for each neighborhood
  #cluster and different columns for counts
  #for each offense type
violent_crime_by_neigh_cluster2 <-
  pivot_wider(violent_crime_by_neigh_cluster,
  names_from = OFFENSE,
  values_from = count)

#Replacing NAs with 0s
#Explanation: in some neighborhood clusters,
  #there were 0 cases of a specific offense.
violent_crime_by_neigh_cluster2$`ASSAULT W/DANGEROUS WEAPON`[
  is.na(violent_crime_by_neigh_cluster2$`ASSAULT W/DANGEROUS WE
violent_crime_by_neigh_cluster2$ROBBERY[
  is.na(violent_crime_by_neigh_cluster2$ROBBERY)] <- 0
violent_crime_by_neigh_cluster2$`SEX ABUSE`[
  is.na(violent_crime_by_neigh_cluster2$`SEX ABUSE`)] <- 0
violent_crime_by_neigh_cluster2$HOMICIDE[
  is.na(violent_crime_by_neigh_cluster2$HOMICIDE)] <- 0

#Want a column name with just the number of a
  #cluster (e.g., 1 instead of Cluster 1)
violent_crime_by_neigh_cluster2[
  c('Cluster_word', 'Cluster_num')] <-
  str_split_fixed(violent_crime_by_neigh_cluster2$NEIGHBORHOOD_
violent_crime_by_neigh_cluster2$Cluster_word <- NULL
```

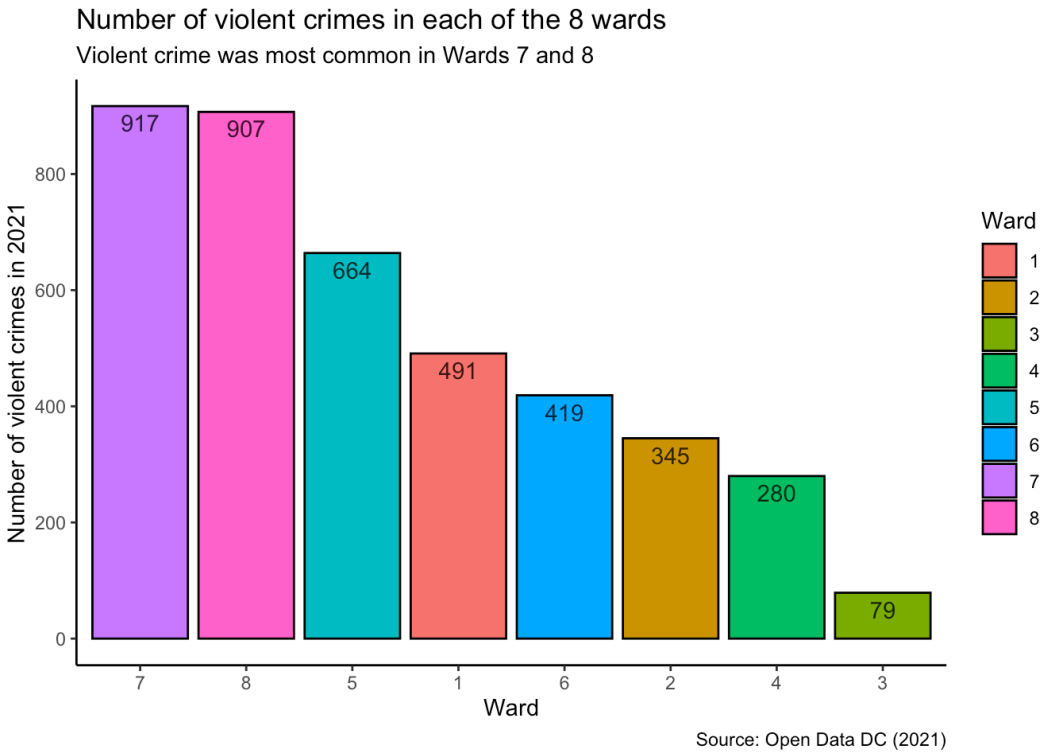
```
#Creating new number of robberies factor variable
  #that contains ranges
violent_crime_by_neigh_cluster2$`Number of robberies` <- NA
violent_crime_by_neigh_cluster2$`Number of robberies`[
  violent_crime_by_neigh_cluster2$ROBBERY >= 0 &
    violent_crime_by_neigh_cluster2$ROBBERY < 40
  ] <- "0-39"
violent_crime_by_neigh_cluster2$`Number of robberies`[
  violent_crime_by_neigh_cluster2$ROBBERY >= 40 &
    violent_crime_by_neigh_cluster2$ROBBERY < 80
  ] <- "40-79"
violent_crime_by_neigh_cluster2$`Number of robberies`[
  violent_crime_by_neigh_cluster2$ROBBERY >= 80 &
    violent_crime_by_neigh_cluster2$ROBBERY < 120
  ] <- "80-119"
violent_crime_by_neigh_cluster2$`Number of robberies`[
  violent_crime_by_neigh_cluster2$ROBBERY >= 120 &
    violent_crime_by_neigh_cluster2$ROBBERY < 160
  ] <- "120+"

violent_crime_by_neigh_cluster2$`Number of robberies` <-
  factor(violent_crime_by_neigh_cluster2$`Number of robberies`,
  levels=c("0-39","40-79", "80-119", "120+"))
```

# First graph

```
#Adding first graph
graph1 <- ggplot(data = violent_crime_by_ward,
#Aesthetic options: x, y, fill (multiple colors)
  aes(x=factor(WARD), y=count,
    fill = factor(WARD))) +
#Non-aesthetic options: stat, color
  geom_bar(stat = "identity", color = "black") +
#Aesthetic option: label
#Non-aesthetic options: vjust, color, alpha
  geom_text(aes(label = count), vjust = 1.5,
    colour = "black", alpha = 0.8) +
  xlab("Ward") +
  scale_x_discrete(
    limits = factor(c(7,8,5,1,6,2,4,3))) +
  ylab("Number of violent crimes in 2021") +
  scale_y_continuous(breaks=seq(0,800,by=200)) +
```

```
  labs(
    title = "Number of violent crimes in each of the 8 wards",
    subtitle = "Violent crime was most common in Wards 7 and 8"
    caption = "Source: Open Data DC (2021)",
    fill = "Ward") +
  theme_classic()
graph1
```




This bar graph shows the number of violent crimes that happened in each of the eight wards of Washington, DC in 2021. The ward where the most violent crimes happened was Ward 7 (917 cases) and the ward where the fewest violent crimes happened was Ward 3 (79). In total, there were 4,106 violent crimes in 2021.

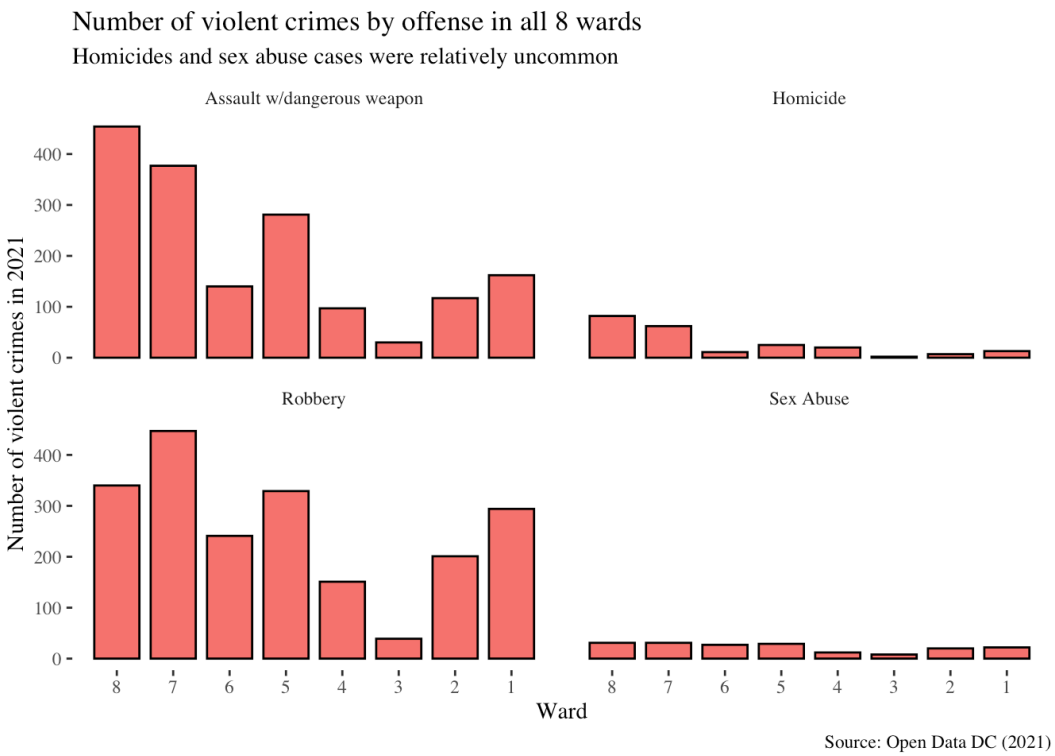
Wards 7 and 8 are high-poverty and have large populations of African Americans. The bar graph helps explain why residents of those wards are especially worried about violent crime.

Note that Open Data DC categorizes homicides, sex abuses, assaults with dangerous weapons, and robberies as violent crimes. Meanwhile, Open Data DC categorizes arson, burglaries, theft with auto, and theft/other as property crimes. This may be a different way of defining violent crime than is used by other researchers but it is the conceptualization that I am using.

(There were 24,208 property crimes in 2021, which is not shown in the graph.)

# Second graph

```
#Adding second graph
graph2 <- violent_crime_by_ward2 %>%
#Aesthetic options: x, y, fill (one color), width
  ggplot(aes(x=WARD, y=count, fill="red", width = 0.8)) +
#Non-aesthetic options: stat, color
  geom_bar(stat = "identity", color = "black") +
  xlab("Ward") +
  scale_x_reverse(
    breaks = seq(8, 1, by = -1),
    labels = c("8", "7", "6", "5", "4", "3", "2", "1")) +
  ylab("Number of violent crimes in 2021") +
  scale_y_continuous(breaks=seq(0,800,by=100)) +
  labs(subtitle = "Homicides and sex abuse cases were relativel
    caption = "Source: Open Data DC (2021)",
    fill = "Ward") +
  ggtitle("Number of violent crimes by offense in all 8 wards")
  theme_tufte() +
  facet_wrap(~OFFENSE) +
  theme(legend.position="none")
graph2
```



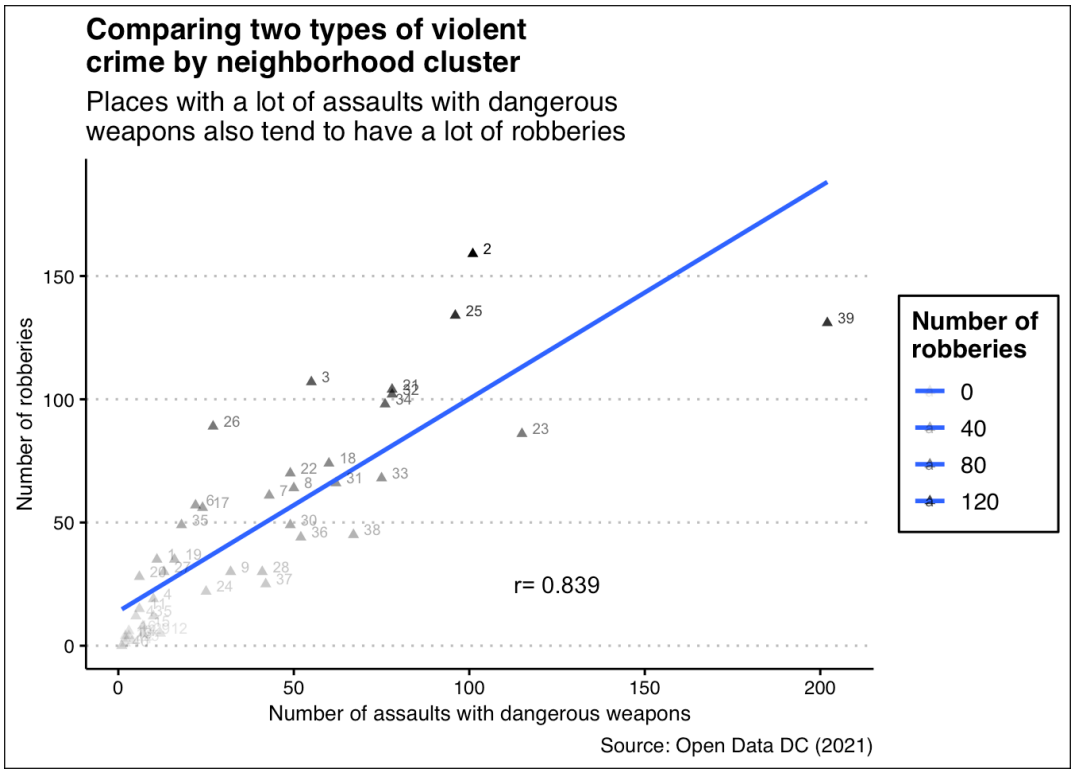

This faceted bar graph shows the number of assaults with dangerous weapons, homicides, robberies, and sex abuses that happened in each of the eight wards of Washington, DC in 2021. The graph suggests that homicides and sex abuses were relatively rare compared to the other two types of violent crime. In addition, the graph shows that assaults with dangerous weapons and robberies were especially likely to happen in Wards 7 and 8 but were rare in Ward 3. Variation in the frequency of homicides and sex abuses between wards were much smaller.

The bar graph suggests that simply counting the number of violent crimes in each ward, as Graph 1 does, is not telling the full story. Rather, we need to break down the numbers by offense type in order to fully understand crime in Washington, DC.

# Third graph

```
#Adding third graph
cors <- cor(violent_crime_by_neigh_cluster2$`ASSAULT W/DANGEROU
  violent_crime_by_neigh_cluster2$ROBBERY)
graph3 <- violent_crime_by_neigh_cluster2 %>%
#Aesthetic options: x, y, label, alpha
  ggplot(aes(x = `ASSAULT W/DANGEROUS WEAPON`,
```

```
    y = ROBBERY,
    label = Cluster_num,
    alpha = ROBBERY)) +
#Non-aesthetic option: pch
  geom_point(pch = 17) +
  geom_smooth(method = "lm", se = FALSE) +
#Non-aesthetic options: hjust, vjust, size, nudge_x
  geom_text(hjust=0, vjust=0, size = 2.5, nudge_x = 3) +
  xlab("Number of assaults with dangerous weapons") +
  ylab("Number of robberies") +
  labs(subtitle = "Places with a lot of assaults with dangerous
  caption = "Source: Open Data DC (2021)",
  alpha = "Number of \nrobberies") +
  ggtitle("Comparing two types of violent \ncrime by neighborhc
  theme_clean()
#Non-aesthetic options: x, y, label
graph3 + annotate("text",x=125, y=25, label=paste("r=",round(cc
```




This scatterplot shows that there is a positive relationship between the number of assaults with dangerous weapons (x variable) and the number of robberies (y variable) in a neighborhood cluster. In other words, parts of Washington, DC with high levels of one type of crime also tend to have high levels of the other type of crime. The Pearson correlation coefficient of 0.839 is very high.

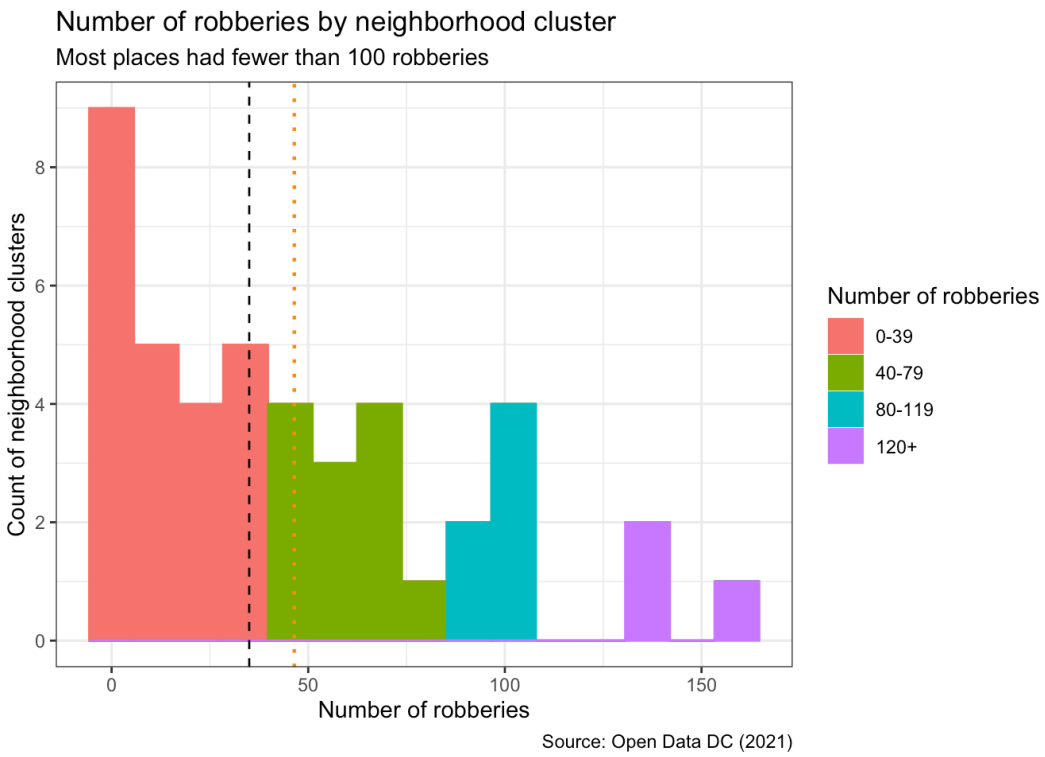
Neighborhood clusters that had especially high levels of both assaults with dangerous weapons and robberies in 2021 include Cluster 2 (Columbia Heights, Mt. Pleasant, Pleasant Plains, Park View), Cluster 25 (NoMa, Union Station, Stanton Park, Kingman Park), and Cluster 39 (Congress Heights, Bellevue, Washington Highlands). Cluster 39 is in Ward 8, which I already showed has a high prevalence of violent crime. However, Cluster 2 is in Ward 1 and Cluster 25 is split between Wards 6 and 7. This shows that even wards with lower overall levels of overall violent crime still have subsections that are very affected by it. The nuance is missed by Graphs 1 and 2 and so the inclusion of Graph 3 helps to tell a more complete story of crime in Washington, DC.

One aspect of the scatterplot is slightly misleading: it shows *counts* of violent crimes, regardless of how many people live in a neighborhood cluster. If I conducted a more in-depth analysis in the future, I would want to account for the fact that some neighborhood clusters have much larger populations than others. If more people live in a neighborhood cluster, there are more opportunities for violent crime to happen, at least in theory.

# Fourth graph

```
#Added fourth graph
graph4 <- violent_crime_by_neigh_cluster2 %>%
#Aesthetic features: x, fill (multiple colors),
    #color (multiple colors)
  ggplot(aes(x=ROBBERY, fill = `Number of robberies`, color=`Nu
  geom_histogram(bins = 15) +
#Non-aesthetic features: x, linetype, color, linewidth
  geom_vline(xintercept =
    median(violent_crime_by_neigh_cluster2$ROBBERY),
    linetype = "dashed", color = "black") +
  geom_vline(xintercept =
    mean(violent_crime_by_neigh_cluster2$ROBBERY),
    linetype = "dotted", color = "darkorange",
    linewidth = 0.8) +
  xlab("Number of robberies") +
  scale_y_continuous(breaks=seq(0,10,by=2)) +
  ylab ("Count of neighborhood clusters") +
  labs(title = "Number of robberies by neighborhood cluster",
       subtitle = "Most places had fewer than 100 robberies",
       caption = "Source: Open Data DC (2021)") +
```

```
  theme_bw()
graph4
```




This histogram shows that most neighborhood clusters had fewer than 100 robberies in 2021. Neighborhood clusters that experienced 0-to-39 robberies are shaded in red, those that experienced 40-to-79 robberies are shaded in green, those that experienced 80-to-119 robberies are shaded in cyan, and those that experienced 120 or more robberies are shaded in purple. The dashed black line shows that the median neighborhood cluster experienced only 35 robberies, while the dotted orange line shows that the mean neighborhood cluster experienced about 46 robberies.

The histogram further illustrates that violent crime (in this case, robberies) was concentrated in a few neighborhood clusters in Washington, DC. It conveys similar information to Graph 3 but hones in one just one variable, as opposed to two.