NYC’s Criminal Summers

This is a quick look at crime in NYC.

Marty Harris

11/15/2020

#### Built in R 4.0.2

# Executive Summary

Domain expert want to understand major trends and insights about NYC crime/violence from several public data source to enable the police officers on patrol. This report analyzes and looks for trends that “beat officers” can use in their daily operations.

The output of this is a model that police officers can use to predict crime in each of their five Boros based on the date and weather. Crime picks a seasonal pattern and weather plays a factor. With client iteration, this tool could be invaluable in managing scarce police department resources.

## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.0  
## v tidyr 1.1.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# Read Data from internet  
  
# df\_arrests\_hist <-read\_csv("https://data.cityofnewyork.us/api/views/8h9b-rp9u/rows.csv?accessType=DOWNLOAD")%>%  
# janitor::clean\_names()  
  
df\_arrests\_hist <-read\_csv("df\_arrests\_hist.csv")%>%  
 janitor::clean\_names()

## Parsed with column specification:  
## cols(  
## arrest\_key = col\_double(),  
## arrest\_date = col\_character(),  
## pd\_cd = col\_double(),  
## pd\_desc = col\_character(),  
## ky\_cd = col\_double(),  
## ofns\_desc = col\_character(),  
## law\_code = col\_character(),  
## law\_cat\_cd = col\_character(),  
## arrest\_boro = col\_character(),  
## arrest\_precinct = col\_double(),  
## jurisdiction\_code = col\_double(),  
## age\_group = col\_character(),  
## perp\_sex = col\_character(),  
## perp\_race = col\_character(),  
## x\_coord\_cd = col\_double(),  
## y\_coord\_cd = col\_double(),  
## latitude = col\_double(),  
## longitude = col\_double(),  
## lon\_lat = col\_character()  
## )

df\_arrests\_ytd <- read\_csv("https://data.cityofnewyork.us/api/views/uip8-fykc/rows.csv?accessType=DOWNLOAD")%>%  
 janitor::clean\_names()

## Parsed with column specification:  
## cols(  
## ARREST\_KEY = col\_double(),  
## ARREST\_DATE = col\_character(),  
## PD\_CD = col\_double(),  
## PD\_DESC = col\_character(),  
## KY\_CD = col\_double(),  
## OFNS\_DESC = col\_character(),  
## LAW\_CODE = col\_character(),  
## LAW\_CAT\_CD = col\_character(),  
## ARREST\_BORO = col\_character(),  
## ARREST\_PRECINCT = col\_double(),  
## JURISDICTION\_CODE = col\_double(),  
## AGE\_GROUP = col\_character(),  
## PERP\_SEX = col\_character(),  
## PERP\_RACE = col\_character(),  
## X\_COORD\_CD = col\_double(),  
## Y\_COORD\_CD = col\_double(),  
## Latitude = col\_double(),  
## Longitude = col\_double(),  
## `New Georeferenced Column` = col\_character()  
## )

df\_shooting <- read\_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD") %>%   
 janitor::clean\_names()

## Parsed with column specification:  
## cols(  
## INCIDENT\_KEY = col\_double(),  
## OCCUR\_DATE = col\_character(),  
## OCCUR\_TIME = col\_time(format = ""),  
## BORO = col\_character(),  
## PRECINCT = col\_double(),  
## JURISDICTION\_CODE = col\_double(),  
## LOCATION\_DESC = col\_character(),  
## STATISTICAL\_MURDER\_FLAG = col\_logical(),  
## PERP\_AGE\_GROUP = col\_character(),  
## PERP\_SEX = col\_character(),  
## PERP\_RACE = col\_character(),  
## VIC\_AGE\_GROUP = col\_character(),  
## VIC\_SEX = col\_character(),  
## VIC\_RACE = col\_character(),  
## X\_COORD\_CD = col\_number(),  
## Y\_COORD\_CD = col\_number(),  
## Latitude = col\_double(),  
## Longitude = col\_double(),  
## Lon\_Lat = col\_character()  
## )

### Let’s begin exploring the data to see any initial insights.

glimpse(df\_arrests\_hist)

## Rows: 5,012,956  
## Columns: 19  
## $ arrest\_key <dbl> 144026181, 144507595, 144565062, 144500188, 14421...  
## $ arrest\_date <chr> "06/26/2015", "07/14/2015", "07/16/2015", "07/14/...  
## $ pd\_cd <dbl> 639, 969, 101, 879, 478, 339, 849, 203, 511, 511,...  
## $ pd\_desc <chr> "AGGRAVATED HARASSMENT 2", "TRAFFIC,UNCLASSIFIED ...  
## $ ky\_cd <dbl> 361, 881, 344, 675, 343, 341, 677, 352, 235, 235,...  
## $ ofns\_desc <chr> "OFF. AGNST PUB ORD SENSBLTY & RGHTS TO PRIV", "O...  
## $ law\_code <chr> "PL 2403002", "VTL051101A", "PL 1200001", "AC 010...  
## $ law\_cat\_cd <chr> "M", "M", "M", "V", "M", "M", "V", "M", "M", "M",...  
## $ arrest\_boro <chr> "Q", "M", "K", "Q", "M", "B", "K", "B", "B", "Q",...  
## $ arrest\_precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115, 111, 3...  
## $ jurisdiction\_code <dbl> 0, 3, 0, 0, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ age\_group <chr> "45-64", "25-44", "18-24", "25-44", "18-24", "18-...  
## $ perp\_sex <chr> "M", "M", "F", "M", "M", "M", "M", "M", "F", "M",...  
## $ perp\_race <chr> "WHITE HISPANIC", "WHITE HISPANIC", "WHITE HISPAN...  
## $ x\_coord\_cd <dbl> 1031076, 984791, 994026, 1037132, 984602, 1030990...  
## $ y\_coord\_cd <dbl> 193779, 209846, 195548, 196129, 210686, 255310, 1...  
## $ latitude <dbl> 40.69844, 40.74266, 40.70341, 40.70486, 40.74497,...  
## $ longitude <dbl> -73.83113, -73.99805, -73.96474, -73.80927, -73.9...  
## $ lon\_lat <chr> "POINT (-73.83112953899997 40.69843969400005)", "...

glimpse(df\_arrests\_ytd)

## Rows: 103,376  
## Columns: 19  
## $ arrest\_key <dbl> 218375425, 216884434, 216348943, 216186332...  
## $ arrest\_date <chr> "09/26/2020", "08/21/2020", "08/09/2020", ...  
## $ pd\_cd <dbl> 157, 175, 109, 157, 397, 157, 244, 792, 26...  
## $ pd\_desc <chr> "RAPE 1", "SEXUAL ABUSE 3,2", "ASSAULT 2,1...  
## $ ky\_cd <dbl> 104, 233, 106, 104, 105, 104, 107, 118, 12...  
## $ ofns\_desc <chr> "RAPE", "SEX CRIMES", "FELONY ASSAULT", "R...  
## $ law\_code <chr> "PL 1303501", "PL 1305500", "PL 1200501", ...  
## $ law\_cat\_cd <chr> "F", "M", "F", "F", "F", "F", "F", "F", "F...  
## $ arrest\_boro <chr> "M", "B", "B", "B", "B", "K", "M", "K", "K...  
## $ arrest\_precinct <dbl> 1, 43, 44, 40, 48, 77, 34, 67, 66, 72, 88,...  
## $ jurisdiction\_code <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ age\_group <chr> "25-44", "18-24", "25-44", "25-44", "25-44...  
## $ perp\_sex <chr> "M", "M", "F", "M", "M", "M", "M", "M", "M...  
## $ perp\_race <chr> "WHITE", "BLACK", "BLACK", "WHITE HISPANIC...  
## $ x\_coord\_cd <dbl> 982285, 1020316, 1007481, 1004926, 1015113...  
## $ y\_coord\_cd <dbl> 201682, 239179, 241648, 234532, 245325, 18...  
## $ latitude <dbl> 40.72026, 40.82310, 40.82992, 40.81040, 40...  
## $ longitude <dbl> -74.00709, -73.86969, -73.91606, -73.92531...  
## $ new\_georeferenced\_column <chr> "POINT (-74.00709027999993 40.720255223000...

glimpse(df\_shooting)

## Rows: 21,626  
## Columns: 19  
## $ incident\_key <dbl> 74146165, 66928846, 29114164, 85180336, 734...  
## $ occur\_date <chr> "08/14/2010", "10/17/2009", "05/18/2007", "...  
## $ occur\_time <time> 03:11:00, 18:03:00, 23:00:00, 17:15:00, 04...  
## $ boro <chr> "QUEENS", "BROOKLYN", "BROOKLYN", "BROOKLYN...  
## $ precinct <dbl> 113, 67, 75, 81, 47, 110, 114, 113, 113, 43...  
## $ jurisdiction\_code <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ location\_desc <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ statistical\_murder\_flag <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FA...  
## $ perp\_age\_group <chr> NA, NA, NA, NA, NA, NA, "25-44", NA, NA, NA...  
## $ perp\_sex <chr> NA, NA, NA, NA, NA, NA, "M", NA, NA, NA, NA...  
## $ perp\_race <chr> NA, NA, NA, NA, NA, NA, "BLACK", NA, NA, NA...  
## $ vic\_age\_group <chr> "25-44", "45-64", "25-44", "25-44", "25-44"...  
## $ vic\_sex <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M"...  
## $ vic\_race <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK"...  
## $ x\_coord\_cd <dbl> 1046573, 1003313, 1016292, 1005597, 1023551...  
## $ y\_coord\_cd <dbl> 183057, 176413, 176228, 188673, 263366, 210...  
## $ latitude <dbl> 40.66891, 40.65088, 40.65033, 40.68452, 40....  
## $ longitude <dbl> -73.77534, -73.93130, -73.88453, -73.92303,...  
## $ lon\_lat <chr> "POINT (-73.77534099699994 40.6689147720000...

### Let’s build a dataset that combines all the arrest data into one while removing duplicates from where they overlap.

df\_arrests <- df\_arrests\_hist %>%   
 bind\_rows(df\_arrests\_ytd)%>%  
 distinct() %>%   
 drop\_na(ofns\_desc)  
glimpse(df\_arrests)

## Rows: 5,107,279  
## Columns: 20  
## $ arrest\_key <dbl> 144026181, 144507595, 144565062, 144500188...  
## $ arrest\_date <chr> "06/26/2015", "07/14/2015", "07/16/2015", ...  
## $ pd\_cd <dbl> 639, 969, 101, 879, 478, 339, 849, 203, 51...  
## $ pd\_desc <chr> "AGGRAVATED HARASSMENT 2", "TRAFFIC,UNCLAS...  
## $ ky\_cd <dbl> 361, 881, 344, 675, 343, 341, 677, 352, 23...  
## $ ofns\_desc <chr> "OFF. AGNST PUB ORD SENSBLTY & RGHTS TO PR...  
## $ law\_code <chr> "PL 2403002", "VTL051101A", "PL 1200001", ...  
## $ law\_cat\_cd <chr> "M", "M", "M", "V", "M", "M", "V", "M", "M...  
## $ arrest\_boro <chr> "Q", "M", "K", "Q", "M", "B", "K", "B", "B...  
## $ arrest\_precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115,...  
## $ jurisdiction\_code <dbl> 0, 3, 0, 0, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, ...  
## $ age\_group <chr> "45-64", "25-44", "18-24", "25-44", "18-24...  
## $ perp\_sex <chr> "M", "M", "F", "M", "M", "M", "M", "M", "F...  
## $ perp\_race <chr> "WHITE HISPANIC", "WHITE HISPANIC", "WHITE...  
## $ x\_coord\_cd <dbl> 1031076, 984791, 994026, 1037132, 984602, ...  
## $ y\_coord\_cd <dbl> 193779, 209846, 195548, 196129, 210686, 25...  
## $ latitude <dbl> 40.69844, 40.74266, 40.70341, 40.70486, 40...  
## $ longitude <dbl> -73.83113, -73.99805, -73.96474, -73.80927...  
## $ lon\_lat <chr> "POINT (-73.83112953899997 40.698439694000...  
## $ new\_georeferenced\_column <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

df <- df\_arrests %>%   
 rename(incident\_key=arrest\_key) %>%  
 rename(occur\_date = arrest\_date) %>%   
 bind\_rows(df\_shooting)  
glimpse(df)

## Rows: 5,128,905  
## Columns: 29  
## $ incident\_key <dbl> 144026181, 144507595, 144565062, 144500188...  
## $ occur\_date <chr> "06/26/2015", "07/14/2015", "07/16/2015", ...  
## $ pd\_cd <dbl> 639, 969, 101, 879, 478, 339, 849, 203, 51...  
## $ pd\_desc <chr> "AGGRAVATED HARASSMENT 2", "TRAFFIC,UNCLAS...  
## $ ky\_cd <dbl> 361, 881, 344, 675, 343, 341, 677, 352, 23...  
## $ ofns\_desc <chr> "OFF. AGNST PUB ORD SENSBLTY & RGHTS TO PR...  
## $ law\_code <chr> "PL 2403002", "VTL051101A", "PL 1200001", ...  
## $ law\_cat\_cd <chr> "M", "M", "M", "V", "M", "M", "V", "M", "M...  
## $ arrest\_boro <chr> "Q", "M", "K", "Q", "M", "B", "K", "B", "B...  
## $ arrest\_precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115,...  
## $ jurisdiction\_code <dbl> 0, 3, 0, 0, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, ...  
## $ age\_group <chr> "45-64", "25-44", "18-24", "25-44", "18-24...  
## $ perp\_sex <chr> "M", "M", "F", "M", "M", "M", "M", "M", "F...  
## $ perp\_race <chr> "WHITE HISPANIC", "WHITE HISPANIC", "WHITE...  
## $ x\_coord\_cd <dbl> 1031076, 984791, 994026, 1037132, 984602, ...  
## $ y\_coord\_cd <dbl> 193779, 209846, 195548, 196129, 210686, 25...  
## $ latitude <dbl> 40.69844, 40.74266, 40.70341, 40.70486, 40...  
## $ longitude <dbl> -73.83113, -73.99805, -73.96474, -73.80927...  
## $ lon\_lat <chr> "POINT (-73.83112953899997 40.698439694000...  
## $ new\_georeferenced\_column <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ occur\_time <time> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ boro <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ precinct <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ location\_desc <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ statistical\_murder\_flag <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ perp\_age\_group <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ vic\_age\_group <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ vic\_sex <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...  
## $ vic\_race <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...

### I’m going to first look at arrest data and see what types of categories exist. I have a feeling that some types of crimes are more frequent than others.

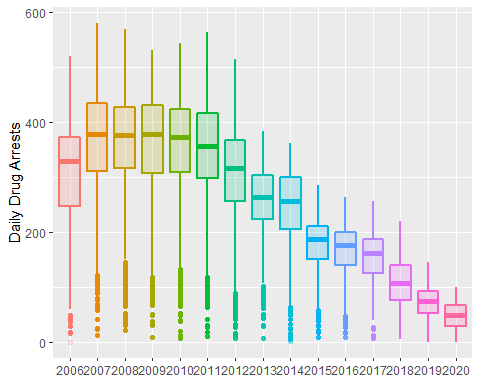
df%>%   
 count(ofns\_desc,sort = T)

## # A tibble: 87 x 2  
## ofns\_desc n  
## <chr> <int>  
## 1 DANGEROUS DRUGS 1085712  
## 2 ASSAULT 3 & RELATED OFFENSES 509465  
## 3 OTHER OFFENSES RELATED TO THEFT 304191  
## 4 OTHER STATE LAWS 241494  
## 5 PETIT LARCENY 218527  
## 6 FELONY ASSAULT 206760  
## 7 VEHICLE AND TRAFFIC LAWS 203825  
## 8 CRIMINAL TRESPASS 200591  
## 9 DANGEROUS WEAPONS 195493  
## 10 MISCELLANEOUS PENAL LAW 183511  
## # ... with 77 more rows

### Drug offenses seem to be the most frequent historical occurrence within the data. I wonder how that frequency changes with time. Let’s take a look at what the day averages look like each year.

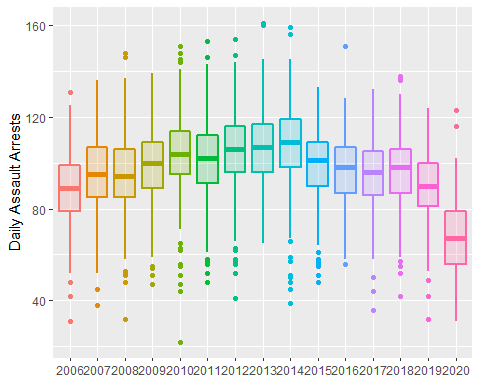
It seems that arrests for dangerous drugs are on a decline.

df %>%  
 filter(ofns\_desc == "DANGEROUS DRUGS") %>%   
 mutate(  
 occur\_date = lubridate::mdy(occur\_date),  
 year = lubridate::year(occur\_date),  
 year = factor(year),  
 month = lubridate:: month(occur\_date),  
 month = factor(month))%>%  
 group\_by(occur\_date)%>%  
 mutate(  
 count\_daily = n())%>%  
 ungroup()%>%  
 ggplot(aes(factor(year), count\_daily,  
 fill = year, color = year  
 )) +  
 geom\_boxplot(alpha = 0.2, size = 1, show.legend = FALSE) +  
 labs(x = NULL, y = "Daily Drug Arrests")



### Perhaps in era of shrinking police budgets, it would be useful to target police officer scheduling to have more police scheduled to be in generally correct locations at the right times to discourage nefarious activities and increase public safety. While Daily Drug arrests might be interesting, they may not be a large danger to society. However, violent crime may be. First, I’ll label data to find violent crime. For this we will need to clean some of the data. Let’s look at the second most common crime assault. In this exploratory data section, I’ll use this as a proxy for overall violent crimes to determine a good timeframe to use for predictions.

# The palette with grey:  
df %>%  
 filter(ofns\_desc == "ASSAULT 3 & RELATED OFFENSES") %>%   
 mutate(  
 occur\_date = lubridate::mdy(occur\_date),  
 year = lubridate::year(occur\_date),  
 year = factor(year),  
 month = lubridate:: month(occur\_date),  
 month = factor(month))%>%  
 group\_by(occur\_date)%>%  
 mutate(  
 count\_daily = n())%>%  
 ungroup()%>%  
 ggplot(aes(year, count\_daily,  
 fill = year, color = year, group =year)) +   
 geom\_boxplot(alpha = 0.2, size = 1, show.legend = FALSE)+  
 labs(x = NULL, y = "Daily Assault Arrests")

 ### It appears that a large reduction in “ASSAULT 3 & RELATED OFFENSES,” occurs around 2015. I’ll assume this is the modern era of policing and use only data from 2015 on.

df1 <- df %>%  
 mutate(  
 occur\_date = lubridate::mdy(occur\_date),  
 year = lubridate::year(occur\_date),  
 month = lubridate:: month(occur\_date),  
 year\_mon = zoo::as.yearmon(occur\_date, "%m/%Y"))%>%   
 filter(year >=2015) %>%  
 mutate(  
 ofns\_desc=ifelse(is.na(ofns\_desc),"shooting",ofns\_desc),  
 statistical\_murder\_flag= ifelse(is.na(statistical\_murder\_flag), FALSE, statistical\_murder\_flag),  
 occur\_date = lubridate::as\_date(occur\_date),  
 occur\_time = ifelse(is.na(occur\_time),12:00,occur\_time),  
 incident\_key = ifelse(is.na(incident\_key),arrest\_key,incident\_key),  
 precinct = ifelse(is.na(precinct),arrest\_precinct,precinct),  
 boro = ifelse(is.na(boro),arrest\_boro,boro),  
 boro = ifelse(boro=="M","MANHATTAN",  
 ifelse(boro=="K", "BROOKLYN",  
 ifelse(boro=="S", "STATEN ISLAND",  
 ifelse(boro=="B", "BRONX",  
 ifelse(boro=="Q","QUEENS",boro)))))) %>%  
 select(-vic\_race,-vic\_sex,-vic\_age\_group,-perp\_age\_group,-location\_desc,  
 -arrest\_boro,-new\_georeferenced\_column)  
   
 glimpse(df1)

## Rows: 1,507,038  
## Columns: 25  
## $ incident\_key <dbl> 144026181, 144507595, 144565062, 144500188,...  
## $ occur\_date <date> 2015-06-26, 2015-07-14, 2015-07-16, 2015-0...  
## $ pd\_cd <dbl> 639, 969, 101, 879, 478, 339, 849, 203, 511...  
## $ pd\_desc <chr> "AGGRAVATED HARASSMENT 2", "TRAFFIC,UNCLASS...  
## $ ky\_cd <dbl> 361, 881, 344, 675, 343, 341, 677, 352, 235...  
## $ ofns\_desc <chr> "OFF. AGNST PUB ORD SENSBLTY & RGHTS TO PRI...  
## $ law\_code <chr> "PL 2403002", "VTL051101A", "PL 1200001", "...  
## $ law\_cat\_cd <chr> "M", "M", "M", "V", "M", "M", "V", "M", "M"...  
## $ arrest\_precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115, ...  
## $ jurisdiction\_code <dbl> 0, 3, 0, 0, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0...  
## $ age\_group <chr> "45-64", "25-44", "18-24", "25-44", "18-24"...  
## $ perp\_sex <chr> "M", "M", "F", "M", "M", "M", "M", "M", "F"...  
## $ perp\_race <chr> "WHITE HISPANIC", "WHITE HISPANIC", "WHITE ...  
## $ x\_coord\_cd <dbl> 1031076, 984791, 994026, 1037132, 984602, 1...  
## $ y\_coord\_cd <dbl> 193779, 209846, 195548, 196129, 210686, 255...  
## $ latitude <dbl> 40.69844, 40.74266, 40.70341, 40.70486, 40....  
## $ longitude <dbl> -73.83113, -73.99805, -73.96474, -73.80927,...  
## $ lon\_lat <chr> "POINT (-73.83112953899997 40.6984396940000...  
## $ occur\_time <dbl> 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 1...  
## $ boro <chr> "QUEENS", "MANHATTAN", "BROOKLYN", "QUEENS"...  
## $ precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115, ...  
## $ statistical\_murder\_flag <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F...  
## $ year <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2...  
## $ month <dbl> 6, 7, 7, 7, 7, 7, 6, 7, 7, 7, 6, 7, 7, 7, 8...  
## $ year\_mon <yearmon> Jun 2015, Jul 2015, Jul 2015, Jul 2015,...

### First, I would like to see how many crime categories there are and simplify a way to bin them. My going assumption is to focus on violent crimes, thefts or damage to property, and petty crimes. We will see what the data yields.

df1 %>%   
 select(ofns\_desc) %>%   
 n\_distinct()

## [1] 83

### Eighty-three crime categories are a lot. I think binning could reduce this dimensionality. In this case, I plan to bin by violent crimes, theft or property crimes, and other crimes.

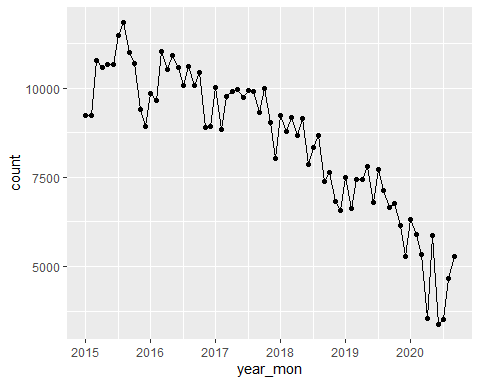
df1 <- df1 %>%   
 mutate(  
 crime\_type = case\_when(  
 str\_detect(ofns\_desc, "ASSAULT")~"violent",  
 str\_detect(ofns\_desc, "shooting")~"violent",  
 str\_detect(ofns\_desc, "HOMICIDE")~"violent",  
 str\_detect(ofns\_desc, "MURDER")~"violent",  
 str\_detect(ofns\_desc, "RAPE")~"violent",  
 str\_detect(ofns\_desc, "FORC")~"violent",  
 str\_detect(ofns\_desc, "SEX")~"violent",  
 str\_detect(ofns\_desc, "KID")~"violent",  
 str\_detect(ofns\_desc, "WEAP")~"violent",  
 str\_detect(ofns\_desc, "DANG")~"violent",  
 str\_detect(ofns\_desc, "HARASSMENT")~"violent",  
 str\_detect(ofns\_desc, "THEFT")~"theft\_property",  
 str\_detect(ofns\_desc, "LARCENY")~"theft\_property",  
 str\_detect(ofns\_desc, "ARSON")~"theft\_property",  
 str\_detect(ofns\_desc, "ENTRY")~"theft\_property",  
 str\_detect(ofns\_desc, "FRAUD")~"theft\_property",  
 str\_detect(ofns\_desc, "VEHICLE")~"theft\_property",  
 str\_detect(ofns\_desc, "ROBBERY")~"theft\_property",  
 str\_detect(ofns\_desc, "BURGLAR")~"theft\_property",  
 str\_detect(ofns\_desc, "FORG")~"theft\_property",  
 str\_detect(ofns\_desc, "PROP")~"theft\_property",  
 str\_detect(ofns\_desc, "PARK")~"theft\_property",  
 str\_detect(ofns\_desc, "DRIVE")~"theft\_property",  
 TRUE~"other"  
 ),  
 crime\_type=factor(crime\_type)  
 )  
  
df1 %>%   
 select(crime\_type) %>%   
 n\_distinct()

## [1] 3

### Three Seems to be more manageable. Now that I have crime data, I can examine any seasonal patterns that might yield insights. For that l will look at scatterplot of the monthly assaults as a proxy measure for violent crimes.

df1 %>%   
 filter(crime\_type == "violent") %>%   
 group\_by(year\_mon) %>%   
 summarize(count = n()) %>%  
 ungroup() %>%   
 ggplot(aes(year\_mon,count))+  
 geom\_point()+  
 geom\_line()

## `summarise()` ungrouping output (override with `.groups` argument)



### There appears to be a seasonal spike in violent crimes during the summer months. It does appear that there is an overall lack of crime reported in 2020 probably due to COVID. With this seasonality in mind, I plan to add NOAA weather data to the overall data set to see how weather might affect the overall daily violent crimes. From previous work with forecasting grocery demand, I know that holidays influence behaviors. I will add a list of national holidays to help tease out any effects that holidays might have on crime.

# Weather data obtained from NOAA  
df\_weather <- read\_csv("https://www.ncei.noaa.gov/orders/cdo/2353296.csv") %>%   
 janitor::clean\_names() %>%   
 select(date,awnd,prcp,snow,snwd,tmax,tmin,wsf2,wt01,wt02,wt03,wt04,wt06,wt08) %>%   
 replace\_na(list(wt01=0,wt02=0,wt03=0,wt04=0,wt06=0,wt08=0))

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## STATION = col\_character(),  
## NAME = col\_character(),  
## DATE = col\_date(format = ""),  
## PGTM = col\_logical(),  
## TAVG = col\_logical(),  
## TSUN = col\_logical(),  
## WT03 = col\_logical()  
## )

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

## Warning: 1 parsing failure.  
## row col expected actual file  
## 1096 PGTM 1/0/T/F/TRUE/FALSE 1509 'https://www.ncei.noaa.gov/orders/cdo/2353296.csv'

# Holidays  
df\_holiday <- read\_csv("usholidays.csv") %>%   
 janitor::clean\_names() %>%   
 mutate(date = lubridate::mdy(date))

## Parsed with column specification:  
## cols(  
## Date = col\_character(),  
## Holiday = col\_character()  
## )

### Now lets add these features to our data set.

df1 <-df1 %>%   
 inner\_join(df\_weather,by = c("occur\_date"="date")) %>%   
 left\_join(df\_holiday, by = c("occur\_date" = "date")) %>%   
 mutate(holiday = ifelse(is.na(holiday),0,1))  
glimpse(df1)

## Rows: 1,507,038  
## Columns: 40  
## $ incident\_key <dbl> 144026181, 144507595, 144565062, 144500188,...  
## $ occur\_date <date> 2015-06-26, 2015-07-14, 2015-07-16, 2015-0...  
## $ pd\_cd <dbl> 639, 969, 101, 879, 478, 339, 849, 203, 511...  
## $ pd\_desc <chr> "AGGRAVATED HARASSMENT 2", "TRAFFIC,UNCLASS...  
## $ ky\_cd <dbl> 361, 881, 344, 675, 343, 341, 677, 352, 235...  
## $ ofns\_desc <chr> "OFF. AGNST PUB ORD SENSBLTY & RGHTS TO PRI...  
## $ law\_code <chr> "PL 2403002", "VTL051101A", "PL 1200001", "...  
## $ law\_cat\_cd <chr> "M", "M", "M", "V", "M", "M", "V", "M", "M"...  
## $ arrest\_precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115, ...  
## $ jurisdiction\_code <dbl> 0, 3, 0, 0, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0...  
## $ age\_group <chr> "45-64", "25-44", "18-24", "25-44", "18-24"...  
## $ perp\_sex <chr> "M", "M", "F", "M", "M", "M", "M", "M", "F"...  
## $ perp\_race <chr> "WHITE HISPANIC", "WHITE HISPANIC", "WHITE ...  
## $ x\_coord\_cd <dbl> 1031076, 984791, 994026, 1037132, 984602, 1...  
## $ y\_coord\_cd <dbl> 193779, 209846, 195548, 196129, 210686, 255...  
## $ latitude <dbl> 40.69844, 40.74266, 40.70341, 40.70486, 40....  
## $ longitude <dbl> -73.83113, -73.99805, -73.96474, -73.80927,...  
## $ lon\_lat <chr> "POINT (-73.83112953899997 40.6984396940000...  
## $ occur\_time <dbl> 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 1...  
## $ boro <chr> "QUEENS", "MANHATTAN", "BROOKLYN", "QUEENS"...  
## $ precinct <dbl> 102, 10, 90, 103, 10, 45, 78, 47, 52, 115, ...  
## $ statistical\_murder\_flag <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F...  
## $ year <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2...  
## $ month <dbl> 6, 7, 7, 7, 7, 7, 6, 7, 7, 7, 6, 7, 7, 7, 8...  
## $ year\_mon <yearmon> Jun 2015, Jul 2015, Jul 2015, Jul 2015,...  
## $ crime\_type <fct> other, other, violent, other, theft\_propert...  
## $ awnd <dbl> 4.47, 2.68, 5.59, 2.68, 2.91, 4.25, 3.13, 4...  
## $ prcp <dbl> 0.00, 0.42, 0.00, 0.42, 0.00, 1.95, 0.00, 0...  
## $ snow <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ snwd <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ tmax <dbl> 81, 82, 80, 82, 82, 87, 83, 81, 94, 81, 71,...  
## $ tmin <dbl> 69, 73, 64, 73, 66, 76, 65, 72, 82, 72, 58,...  
## $ wsf2 <dbl> 16.1, 8.1, 15.0, 8.1, 8.9, 12.1, 13.0, 13.0...  
## $ wt01 <dbl> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0...  
## $ wt02 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ wt03 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ wt04 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ wt06 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ wt08 <dbl> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0...  
## $ holiday <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...

### To do a bit of exploring, I want to see if rain has a visual effect on the number of murders.

boro\_site <- "https://data.cityofnewyork.us/api/geospatial/tqmj-j8zm?method=export&format=GeoJSON" #CAA internet blocked me from downloading this  
# load boros  
nyboros <- geojsonio::geojson\_read("json/Borough Boundaries.geojson", what = "sp")   
#prepare color mapping  
nyboros$boro\_name <- factor(nyboros$boro\_name)  
factpal <- colorFactor("Set1", nyboros$boro\_name)  
# build murders in the rain map  
leaflet() %>%   
 addPolygons(data = nyboros, weight = 2, fillColor = ~factpal(boro\_name), group = "Boros") %>%   
 setView(-74,40.7,zoom=10) %>%   
 addTiles(group = "default") %>%   
 addMarkers(data = df1 %>% filter(statistical\_murder\_flag == TRUE & prcp>0.1),~longitude, ~latitude,popup = ~as.character(boro), label = ~as.character(ofns\_desc), group = "Shooting Murders in the Rain") %>%   
 addMarkers(data = df1 %>% filter(statistical\_murder\_flag == TRUE & prcp==0),~longitude, ~latitude,popup = ~as.character(boro), label = ~as.character(ofns\_desc), group = "Shooting Murders in the Clear") %>%   
 addLayersControl(  
 baseGroups = c("default"),  
 overlayGroups = c("Boros", "Shooting Murders in the Rain", "Shooting Murders in the Clear"),  
 options = layersControlOptions(collapsed = FALSE))

## PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

### The Map above demonstrates why I think weather is a factor. If it’s a rainy day there are way less shooting murders. I think I should model crime. I plan to focus on building a predictive model across each Boro to determine how many violent crimes, theft/property crimes and other crimes the NYPD should expect on a given month and date of week. The parameters to control will include weather and holidays.

df2 <- df1 %>%   
 select(occur\_date,boro,month,awnd,prcp,snow,tmax,tmin,wsf2,wt01,wt02,wt03,wt04,wt06,wt08,holiday,crime\_type) %>%   
 mutate(  
 prcp = ifelse(prcp >0.25, 1,0),  
 snow = ifelse(snow >0.25, 1,0),  
 awnd = ifelse(awnd>mean(awnd,na.rm = T),1,0), # proxy for windy days  
 weekday = lubridate::wday(occur\_date,label = F, abbr = F)  
 ) %>%   
 group\_by(boro,occur\_date, crime\_type) %>%   
 mutate(value = n()) %>%   
 distinct() %>%   
 ungroup() %>%   
 select(-occur\_date) %>%   
 na.omit()  
  
glimpse(df2)

## Rows: 28,400  
## Columns: 18  
## $ boro <chr> "QUEENS", "MANHATTAN", "BROOKLYN", "QUEENS", "MANHATTAN"...  
## $ month <dbl> 6, 7, 7, 7, 7, 7, 6, 7, 7, 7, 6, 7, 7, 7, 8, 7, 7, 6, 7,...  
## $ awnd <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ prcp <dbl> 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0,...  
## $ snow <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ tmax <dbl> 81, 82, 80, 82, 82, 87, 83, 81, 94, 81, 71, 88, 87, 83, ...  
## $ tmin <dbl> 69, 73, 64, 73, 66, 76, 65, 72, 82, 72, 58, 75, 76, 72, ...  
## $ wsf2 <dbl> 16.1, 8.1, 15.0, 8.1, 8.9, 12.1, 13.0, 13.0, 12.1, 13.0,...  
## $ wt01 <dbl> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,...  
## $ wt02 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ wt03 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ wt04 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ wt06 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ wt08 <dbl> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1,...  
## $ holiday <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ crime\_type <fct> other, other, violent, other, theft\_property, theft\_prop...  
## $ weekday <dbl> 6, 3, 5, 3, 6, 5, 5, 4, 2, 4, 7, 3, 5, 2, 4, 7, 2, 4, 4,...  
## $ value <int> 72, 110, 121, 71, 111, 72, 122, 57, 95, 90, 71, 124, 96,...

### Since we are attempting to predict 3 outputs across 5 boros I plan to take a simple approach using a nested linear model.

library(tidymodels)

## Warning: package 'tidymodels' was built under R version 4.0.3

## -- Attaching packages --------------------------------------------------------------------- tidymodels 0.1.1 --

## v broom 0.7.0 v recipes 0.1.13  
## v dials 0.0.9 v rsample 0.0.7   
## v infer 0.5.3 v tune 0.1.1   
## v modeldata 0.1.0 v workflows 0.2.1   
## v parsnip 0.1.4 v yardstick 0.0.7

## Warning: package 'dials' was built under R version 4.0.3

## Warning: package 'infer' was built under R version 4.0.3

## Warning: package 'modeldata' was built under R version 4.0.3

## Warning: package 'parsnip' was built under R version 4.0.3

## Warning: package 'tune' was built under R version 4.0.3

## Warning: package 'workflows' was built under R version 4.0.3

## Warning: package 'yardstick' was built under R version 4.0.3

## -- Conflicts ------------------------------------------------------------------------ tidymodels\_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()

lm\_model <- df2 %>%   
 group\_by(boro,crime\_type) %>%   
 nest() %>%   
 mutate(model= map(data, ~lm(value~.,data=.x)),  
 predicted = map2(model,data,predict))  
saveRDS(lm\_model,"model.RDS") # save model for later use   
  
eval <- lm\_model %>%   
 mutate(glance = map(model,glance)) %>%   
 unnest(glance)  
   
eval

## # A tibble: 15 x 17  
## # Groups: boro, crime\_type [15]  
## boro crime\_type data model predicted r.squared adj.r.squared sigma  
## <chr> <fct> <lis> <lis> <list> <dbl> <dbl> <dbl>  
## 1 QUEE~ other <tib~ <lm> <dbl [1,~ 0.116 0.109 16.6   
## 2 MANH~ other <tib~ <lm> <dbl [1,~ 0.105 0.0978 25.1   
## 3 BROO~ violent <tib~ <lm> <dbl [1,~ 0.148 0.141 25.5   
## 4 MANH~ theft\_pro~ <tib~ <lm> <dbl [1,~ 0.0917 0.0845 33.9   
## 5 BRONX theft\_pro~ <tib~ <lm> <dbl [1,~ 0.0685 0.0611 19.5   
## 6 BROO~ other <tib~ <lm> <dbl [1,~ 0.139 0.132 21.4   
## 7 BRONX other <tib~ <lm> <dbl [1,~ 0.0759 0.0685 14.2   
## 8 BRONX violent <tib~ <lm> <dbl [1,~ 0.149 0.143 24.6   
## 9 QUEE~ violent <tib~ <lm> <dbl [1,~ 0.0962 0.0890 16.8   
## 10 QUEE~ theft\_pro~ <tib~ <lm> <dbl [1,~ 0.0615 0.0541 18.7   
## 11 BROO~ theft\_pro~ <tib~ <lm> <dbl [1,~ 0.0861 0.0789 26.1   
## 12 MANH~ violent <tib~ <lm> <dbl [1,~ 0.177 0.170 23.9   
## 13 STAT~ other <tib~ <lm> <dbl [1,~ 0.0391 0.0314 3.88  
## 14 STAT~ theft\_pro~ <tib~ <lm> <dbl [1,~ 0.0162 0.00819 4.81  
## 15 STAT~ violent <tib~ <lm> <dbl [1,~ 0.0662 0.0588 5.77  
## # ... with 9 more variables: statistic <dbl>, p.value <dbl>, df <dbl>,  
## # logLik <dbl>, AIC <dbl>, BIC <dbl>, deviance <dbl>, df.residual <int>,  
## # nobs <int>

### The models are not great overall, but it is easy to tweak and refine the assumptions to build a more predictive model. In this example, I’m just using a linear model to use as to build a minimum viable product that the client can use in the future. Now that I have a model, I’ll design a method to take user input and make a daily prediction. The NYPD could better allocate their scarce resources to the correct Boros where they should expect higher crime rates on a given day of a month.I’ll use a shiny app for this.

test <- df2 %>%   
 group\_by(boro,crime\_type) %>%   
 sample\_n(1) %>%   
 select(-value) %>%   
 nest()  
  
df\_test <-test %>% select(boro,crime\_type) %>% ungroup()  
#write\_csv(df\_test, "pred\_df.csv")  
preds <- df\_test %>%   
 mutate(pred = map2(lm\_model$model,test$data,predict)) %>%   
 unnest(pred)  
preds

## # A tibble: 15 x 3  
## boro crime\_type pred  
## <chr> <fct> <dbl>  
## 1 BRONX other 45.4   
## 2 BRONX theft\_property 36.6   
## 3 BRONX violent 94.9   
## 4 BROOKLYN other 88.5   
## 5 BROOKLYN theft\_property 55.0   
## 6 BROOKLYN violent 54.6   
## 7 MANHATTAN other 43.4   
## 8 MANHATTAN theft\_property 91.1   
## 9 MANHATTAN violent 46.1   
## 10 QUEENS other 44.0   
## 11 QUEENS theft\_property 57.7   
## 12 QUEENS violent 72.7   
## 13 STATEN ISLAND other 8.83  
## 14 STATEN ISLAND theft\_property 8.87  
## 15 STATEN ISLAND violent 9.76