Week3

Willis Banks

2023-11-18

Data Source

The data below comes from the NYPD historical shooting incidents, as publicly available in the link below.

```
dataURL <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
rawData <- read_csv(dataURL)
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Tidying

The following was done to the data prior to analysis:

- Occurrence date and time were cast from character/numeric to a single timestamp variable called OCCUR_TIMESTAMP
- All descriptive encoding were cast as factor
  - The existing data set is sufficiently clean to cast the assorted character descriptions directly (eg LOCATION_DESC)
- Redundant columns were removed
  - This includes occurrence date/time as well as the extra latitude/longitude column

After tidying, there is missing data several columns. Predominantly, the perpetrator description columns, as the missing data would be associated with a perpetrator that was never caught at the time of the data's publishing. For any perpetrator based analysis, these rows would need to be removed. Alternatively, one could use these rows to analyze the difference between crimes in which the perpatrator is un/known. In addition to this, location description data is also sparse. If specific building types are part of the analysis, one would have to remove those rows of data.

```
tidyData <- rawData
factorCols = c("BORO","LOC_OF_OCCUR_DESC","PRECINCT","JURISDICTION_CODE","LOC_CLASSFCTN_DESC","LOCATION_
dropCols = c("OCCUR_DATE","OCCUR_TIME","Lon_Lat")
tidyData$OCCUR_TIMESTAMP <- mdy_hms(paste(tidyData$OCCUR_DATE,tidyData$OCCUR_TIME))
tidyData = tidyData %>% mutate_at(factorCols,factor)
tidyData = subset(tidyData,select = !(names(tidyData) %in% dropCols))
summary(tidyData)
```

```
##   INCIDENT_KEY                  BORO       LOC_OF_OCCUR_DESC    PRECINCT
##  Min.   :  9953245   BRONX        : 7937   INSIDE :  242     75     : 1557
##  1st Qu.: 63860880   BROOKLYN     :10933   OUTSIDE: 1474     73     : 1452
##  Median : 90372218   MANHATTAN    : 3572   NA's   :25596     67     : 1216
##  Mean   :120860536   QUEENS       : 4094                     44     : 1020
##  3rd Qu.:188810230   STATEN ISLAND:  776                     79     : 1012
##  Max.   :261190187                                           47     :  953
##                                                              (Other):20102
##  JURISDICTION_CODE  LOC_CLASSFCTN_DESC                  LOCATION_DESC
##  0   :22809        STREET    : 1103    MULTI DWELL - PUBLIC HOUS: 4832
##  1   :   74        HOUSING   :  280    MULTI DWELL - APT BUILD  : 2835
##  2   : 4427        DWELLING  :  127    (null)                   :  977
##  NA's:    2        COMMERCIAL:  100    PVT HOUSE                :  951
##                    OTHER     :   31    GROCERY/BODEGA           :  694
##                    (Other)   :   75    (Other)                  : 2046
##                    NA's      :25596    NA's                     :14977
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX              PERP_RACE
##  Mode :logical           18-24  :6222   (null):  640   BLACK         :11432
##  FALSE:22046             25-44  :5687   F     :  424   WHITE HISPANIC: 2341
##  TRUE :5266              UNKNOWN:3148   M     :15439   UNKNOWN       : 1836
##                          <18    :1591   U     : 1499   BLACK HISPANIC: 1314
##                          (null) : 640   NA's  : 9310   (null)        :  640
##                          (Other): 680                  (Other)       :  439
##                          NA's   :9344                  NA's          : 9310
##  VIC_AGE_GROUP   VIC_SEX                            VIC_RACE
##  <18    : 2839   F: 2615   AMERICAN INDIAN/ALASKAN NATIVE:   10
##  1022   :    1   M:24686   ASIAN / PACIFIC ISLANDER      :  404
##  18-24  :10086   U:   11   BLACK                         :19439
##  25-44  :12281             BLACK HISPANIC                : 2646
##  45-64  : 1863             UNKNOWN                       :   66
##  65+    :  181             WHITE                         :  698
##  UNKNOWN:   61             WHITE HISPANIC                : 4049
##    X_COORD_CD        Y_COORD_CD        Latitude       Longitude
##  Min.   : 914928   Min.   :125757   Min.   :40.51   Min.   :-74.25
##  1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67   1st Qu.:-73.94
##  Median :1007731   Median :194487   Median :40.70   Median :-73.92
##  Mean   :1009449   Mean   :208127   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :1066815   Max.   :271128   Max.   :40.91   Max.   :-73.70
##                                     NA's   :10      NA's   :10
##  OCCUR_TIMESTAMP
##  Min.   :2006-01-01 02:00:00.00
##  1st Qu.:2009-07-18 04:20:00.00
##  Median :2013-04-29 15:35:00.00
##  Mean   :2014-01-07 11:55:45.83
```

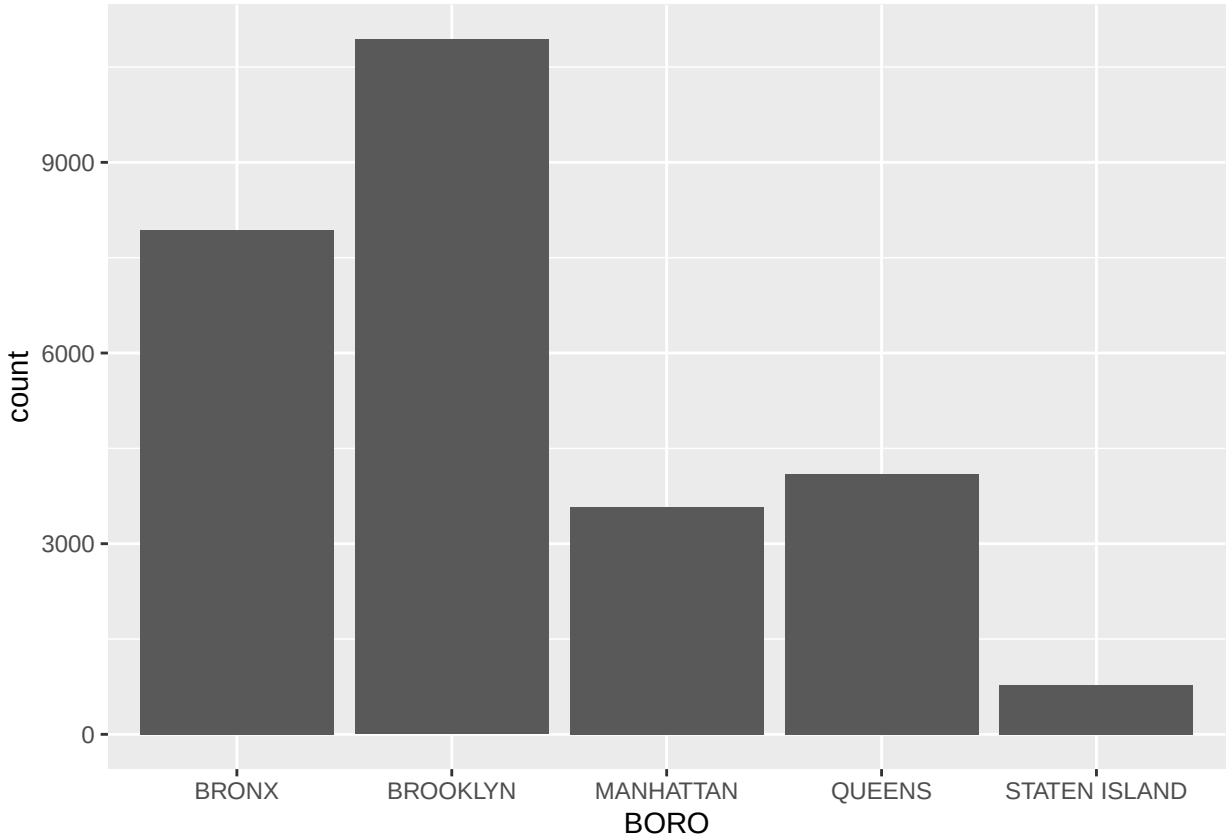
```
##  3rd Qu.:2018-10-15 16:56:30.00
##  Max.   :2022-12-31 23:41:00.00
##
```

Simple Analysis

Below are a few simple plots for analysis.

The first is a bar plot of which boroughs the crimes were committed in. At a glance, Brooklyn and The Bronx dominate the other three boroughs, while Staten Island appears to be significantly safer than the rest by a good margin. The first question to ask is what factors may contribute to these boroughs being so much more/less safe compared to Manhattan or Queens. Or is there some reason that The Bronx and Brooklyn are potentially overrepresented? Or Staten Island underrepresented?

```
p1 <- ggplot(tidyData, aes(x=BORO)) + geom_bar()
print(p1)
```

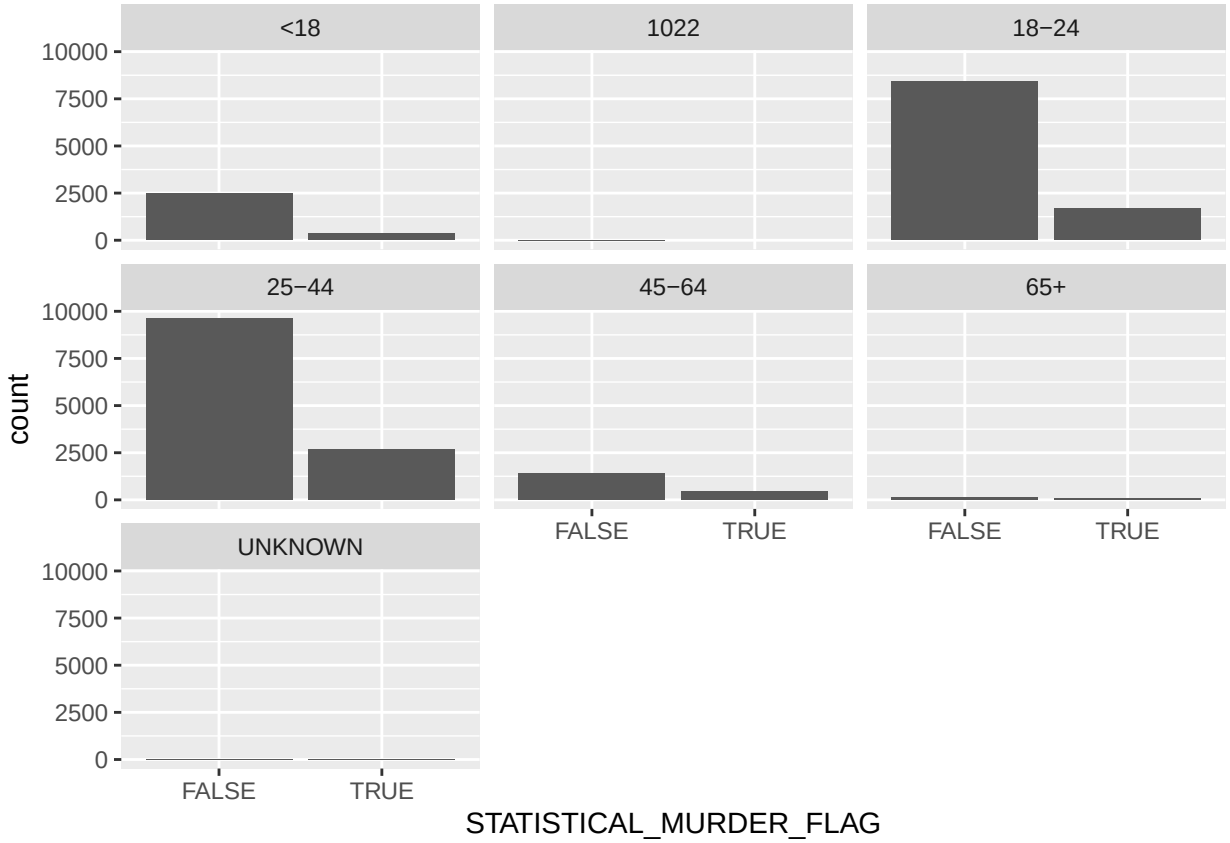


The second is victim age vs homicide. This shows both the relative proportions of victims were in a given age bracket against how many died from their injuries. The data shows a couple things.

One, an almost unused category (1022) was present in the data that would need to be removed in further reports. A single data point on what seems to be an erroneously categorized incident only makes the plots harder to read.

Two, the data runs with the general idea of younger adult age groups, specifically 18-24 and 25-44, are proportionally in more incidents and subsequently are more deaths. The first question to ask is what factors would contribute to these categories having a higher incidence count than the others.

```
p2 <- ggplot(tidyData, aes(x=STATISTICAL_MURDER_FLAG)) + geom_bar() + facet_wrap( ~ VIC_AGE_GROUP)
print(p2)
```



Bias

All of the above plots are subject to the biases of the reporters (in this case, the NYPD). While larger items, such as location and age, are unlikely to be directly biased (eg, an officer in The Bronx is unlikely to report an incident to be elsewhere) they would be subject to systemic biases. If an area has an above average police presence, one would expect higher incident numbers as a results of those locations, and subsequently more would go unaccounted for in locations where there are fewer officers. Similarly, different victim groups are more/less likely to report an incident or for an incident to be noticed (eg an individual in the 45-64 range may have fewer social ties that would notice if they were to go missing).

None of these address controlling for populations. While it could be done with the counts in the data, that yields a proportion of incidents without greater context. Better analysis would obtain population counts for New York City associated with the assorted groups and control that way. From there, it would be easier to determine if Brooklyn is really more dangerous than the other Boroughs, or if it simply has a higher population than the others.

As for personal biases, there are none worth noting. This is being done as obliged by my coursework vs particular interest and the specific analyses were functionally a roll of the dice and very simple. The only significant bias was avoiding discussing any racial components. This is due to my own discomfort in analyzing a very complex issue with a limited, toy data set that I suspect is significantly biased in that regard.