NYPD Shootings Analysis

Courtney Singer

18/11/2021

In this analysis we look at data on shootings in New York City from 2006 to 2021, whish is cleaned, transformed, visualized and modeled using the tidyverse package in R, specifically ggplot2 and lubridate, .

Introducing the data

The data being used for this analysis includes all reported shootings in the city of New York from 2006 to 2021, as well as some addition information including the location of the incident, some "perp" and victim information if available, and the time and date of the incident. Only some of this data will be used in the analysis, but what that will be decided in the data cleaning section.

The data is retrieved from the website https://catalog.data.gov/dataset/nypd-shooting-incident-data-year-to-date , and read in as a csv. A summary of the data can be seen below, which includes all variables, their type, and some information on the data each contains.

```
#The data we are using is available online through the link below in csv format
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read.csv(url_in)
# Display the stucture of the data so we can see what were dealing with
str(nypd_data)
```

```
## 'data.frame':    23585 obs. of  19 variables:
##  $ INCIDENT_KEY           : int  24050482 77673979 203350417 80584527 90843766 92393427 73057167 2113
##  $ OCCUR_DATE             : chr  "08/27/2006" "03/11/2011" "10/06/2019" "09/04/2011" ...
##  $ OCCUR_TIME             : chr  "05:35:00" "12:03:00" "01:09:00" "03:35:00" ...
##  $ BORO                   : chr  "BRONX" "QUEENS" "BROOKLYN" "BRONX" ...
##  $ PRECINCT               : int  52 106 77 40 100 67 77 81 101 106 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LOCATION_DESC          : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "true" "false" "false" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "65+" "18-24" "<18" ...
##  $ VIC_SEX                : chr  "F" "M" "F" "M" ...
##  $ VIC_RACE               : chr  "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  1017542 1027543 995325 1007453 1041267 ...
##  $ Y_COORD_CD             : num  255919 186095 185155 233952 157134 ...
##  $ Latitude               : num  40.9 40.7 40.7 40.8 40.6 ...
##  $ Longitude              : num  -73.9 -73.8 -74 -73.9 -73.8 ...
##  $ Lon_Lat                : chr  "POINT (-73.87963173099996 40.86905819000003)" "POINT (-73.84392019
```

Some of the variables do not have obvious meanings and may need further description to fully understand. For that I refered to the NYPD website from which the data was retrieved, and have provided a description of each of the columns below.

- **Incident Key**: Randomly generated persistent ID for each arrest
- **Jurisdiction Code**: Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions
- **Statistical Murder Flag**: Shooting resulted in the victim's death
- **X coor CD / X coord CD**: Midblock X and Y-coordinate for New York State Plane Coordinate System

Information about the data set was found on the following website: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Year-To-Date-/5ucz-vwe8

Now we can begin to explore the data.

Cleaning the Data

Not all the data will be used for this analysis, so at this stage the data will be examined to determine which sections are needed and which can be removed.

As a first step were going to look at missing data, and determine if there's any variables with too much missing data to allow for meaningful analysis. We can calculate the percentage of missing data for each variable using a loop, which prints out the variable name and it's percentage of missing data.

```
## [1] "INCIDENT_KEY" "0"
## [1] "OCCUR_DATE" "0"
## [1] "OCCUR_TIME" "0"
## [1] "BORO" "0"
## [1] "PRECINCT" "0"
## [1] "JURISDICTION_CODE" NA
## [1] "LOCATION_DESC" "57.6"
## [1] "STATISTICAL_MURDER_FLAG" "0"
## [1] "PERP_AGE_GROUP" "35.2"
## [1] "PERP_SEX" "35"
## [1] "PERP_RACE" "35"
## [1] "VIC_AGE_GROUP" "0"
## [1] "VIC_SEX" "0"
## [1] "VIC_RACE" "0"
## [1] "X_COORD_CD" "0"
## [1] "Y_COORD_CD" "0"
## [1] "Latitude" "0"
## [1] "Longitude" "0"
## [1] "Lon_Lat" "0"
```

This shows there are 4 variables that are missing data LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, and PERP_RACE. The category with the highest percent of missing data, the location description, is missing for over 50% of the incidents, and so this may not be the most useful variable to do analysis on. The other three, perpetrator sex, race and age group are all missing almost the same amount of data. Taking a step back, a possible reason for this could be that the perpetrator in these incidents may not have been caught or identified. These three categories can still be useful to look at, but it will be important to keep in mind that there are unknowns values so conclusions may not be definite. This analysis does not deal with these variables.

Now that we know which variables have sufficient data, we can begin to look at the data set as a whole and determine what steps need to be taken to make the data easier to work with. The first obvious step is to convert the date and time columns into date and time classes respectively. This is done using the lubricate package.

Now we can look at the classes of each variable. For some of the variables that are initially characters it makes sense to classify them as factors, as there is a distinct number of options that that variable could hold, for example borough can only hold the value Manhattan, Brooklyn, Queens, Bronx, or Staten Island. The variables that are converted into factors in this step are **Borough, Perp age group, Perp race, Victim age group, Victim race, Statistical Murder Flag**.

The final step in this section of cleaning the data is to remove the data that is determined not to be useful. For this analysis, the data removed includes **Incident key and Location description** as they will not be needed for the analysis, as well as **longitude, latitude, X coor CD, and Y coor CD** because we are keeping the combined long lat category to mark the locations of the incidents. An additional step is taken to rename some of the variables.

```
nypd_data <- nypd_data %>%

  # Rename columns
  rename(Borough = BORO, Time = OCCUR_TIME, Date = OCCUR_DATE, Perp_age = PERP_AGE_GROUP,
         Perp_sex = PERP_SEX, Perp_race = PERP_RACE, Vic_age = VIC_AGE_GROUP,
         Vic_sex = VIC_SEX, Vic_race = VIC_RACE, Murder_flag = STATISTICAL_MURDER_FLAG,
         Precinct = PRECINCT) %>%

  # Convert the dates from character class to MDY and times to HMS
  mutate(Date = mdy(Date)) %>%
  mutate(Time = as.POSIXct(Time, format='%H:%M:%S')) %>%


  # Turn some of the variables into factors so they are easier to visualize?
  mutate(Borough = factor(Borough), Perp_age = factor(Perp_age),
         Murder_flag = factor(Murder_flag), Perp_race = factor(Perp_race),
         Vic_race = factor(Vic_race), Perp_age = factor(Perp_age),
         Vic_age = factor(Vic_age), Perp_sex = factor(Perp_sex),
         Vic_sex = factor(Vic_sex)) %>%

  # Remove of columns we wont be using
  select(-c(INCIDENT_KEY, LOCATION_DESC, Longitude, Latitude, Y_COORD_CD,
            X_COORD_CD, JURISDICTION_CODE))

# Display structure of formatted data
summary(nypd_data)
```

```
##       Date                 Time                              Borough
##  Min.   :2006-01-01   Min.   :2021-11-22 00:00:00   BRONX        :6701
##  1st Qu.:2008-12-31   1st Qu.:2021-11-22 03:20:00   BROOKLYN     :9734
##  Median :2012-02-27   Median :2021-11-22 15:00:00   MANHATTAN    :2922
##  Mean   :2012-10-05   Mean   :2021-11-22 12:33:07   QUEENS       :3532
##  3rd Qu.:2016-03-02   3rd Qu.:2021-11-22 20:45:00   STATEN ISLAND: 696
##  Max.   :2020-12-31   Max.   :2021-11-22 23:59:00
##
##     Precinct      Murder_flag      Perp_age    Perp_sex           Perp_race
##  Min.   :  1.00   false:19085          :8295    : 8261   BLACK         :10025
##  1st Qu.: 44.00   true : 4500   18-24  :5508   F:  335                 : 8261
```

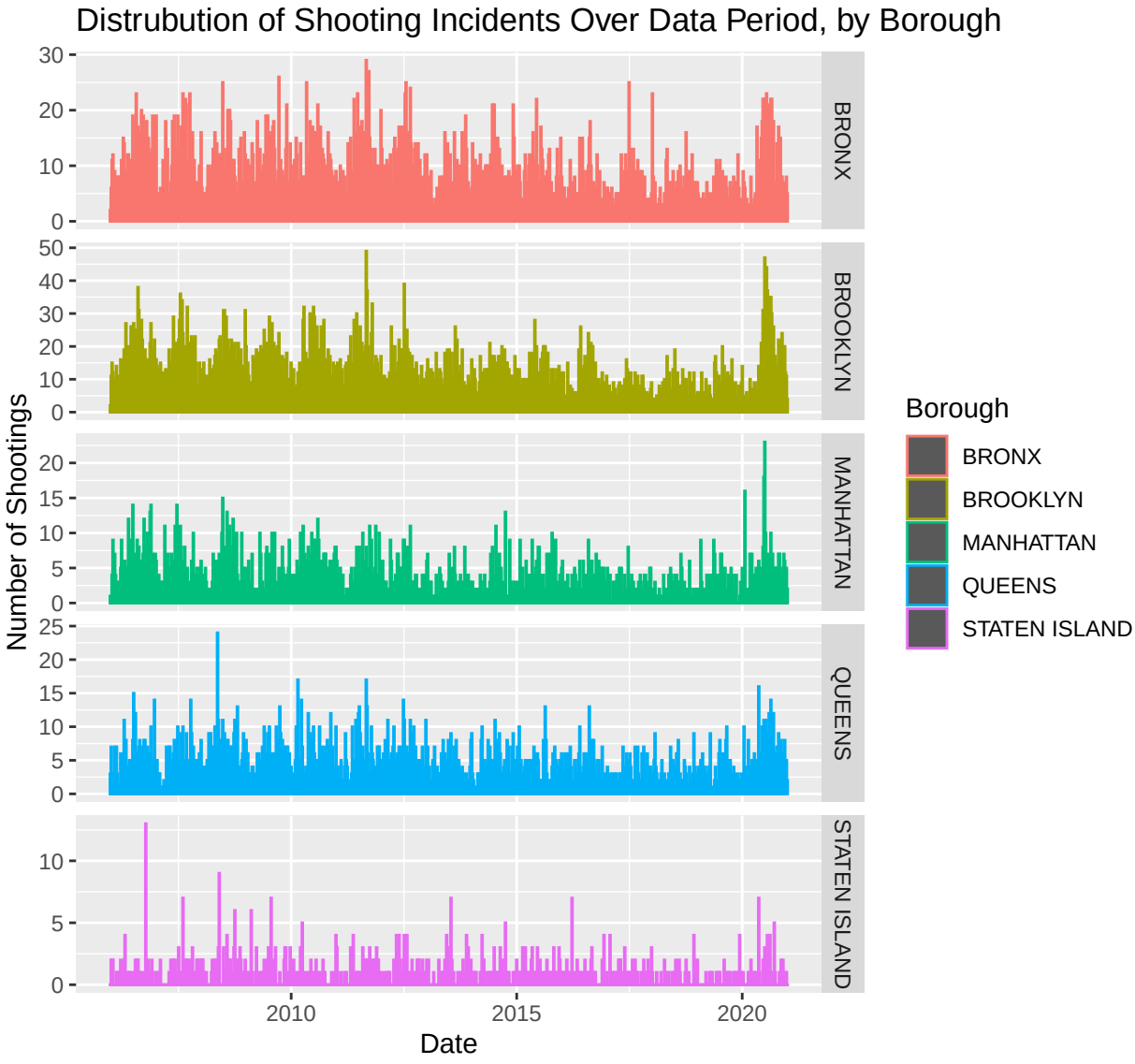
```
##  Median : 69.00                25-44  :4714   M:13490   WHITE HISPANIC: 1988
##  Mean   : 66.21                UNKNOWN:3148   U: 1499   UNKNOWN       : 1836
##  3rd Qu.: 81.00                <18    :1368             BLACK HISPANIC: 1096
##  Max.   :123.00                45-64  : 495             WHITE         :  255
##                                (Other):  57             (Other)       :  124
##     Vic_age      Vic_sex                              Vic_race
##  <18    : 2525   F: 2204   AMERICAN INDIAN/ALASKAN NATIVE:    9
##  18-24  : 9003   M:21370   ASIAN / PACIFIC ISLANDER      :  327
##  25-44  :10303   U:   11   BLACK                         :16869
##  45-64  : 1541             BLACK HISPANIC                : 2245
##  65+    :  154             UNKNOWN                       :   65
##  UNKNOWN:   59             WHITE                         :  620
##                            WHITE HISPANIC                : 3450
##    Lon_Lat
##  Length:23585
##  Class :character
##  Mode  :character
##
##
##
##
```

Now that the data has been cleaned we can begin to transform and visualize it.

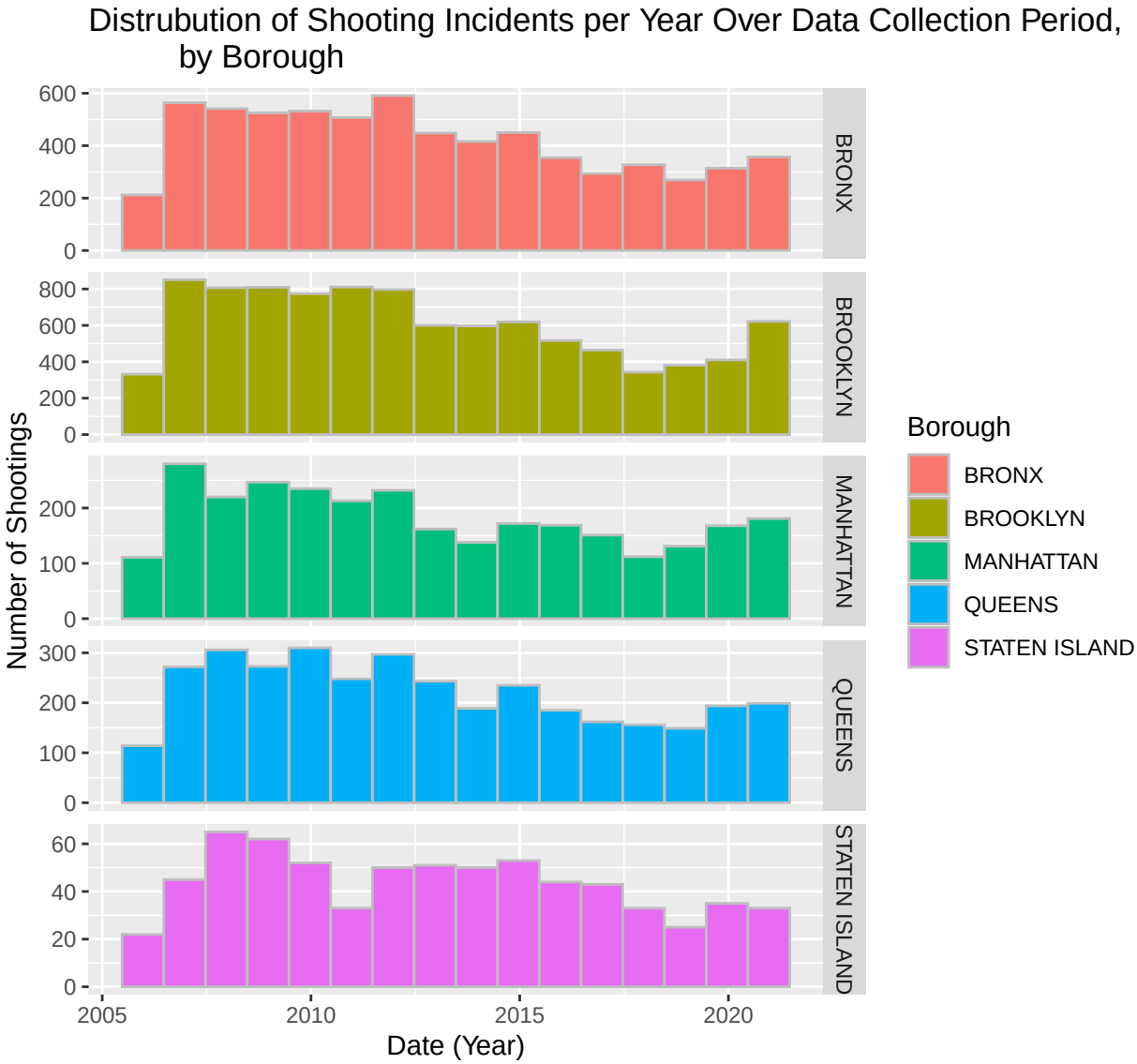
Visualize and transform the data

Frequency of Events over Time Period of Data

The first thing that is looked at is the when the events took place over the time period that the data was collected. This plot is also broken down by Borough to see what localities saw the greatest number of incidents. Each bar on this graph represents one week or 7 days.

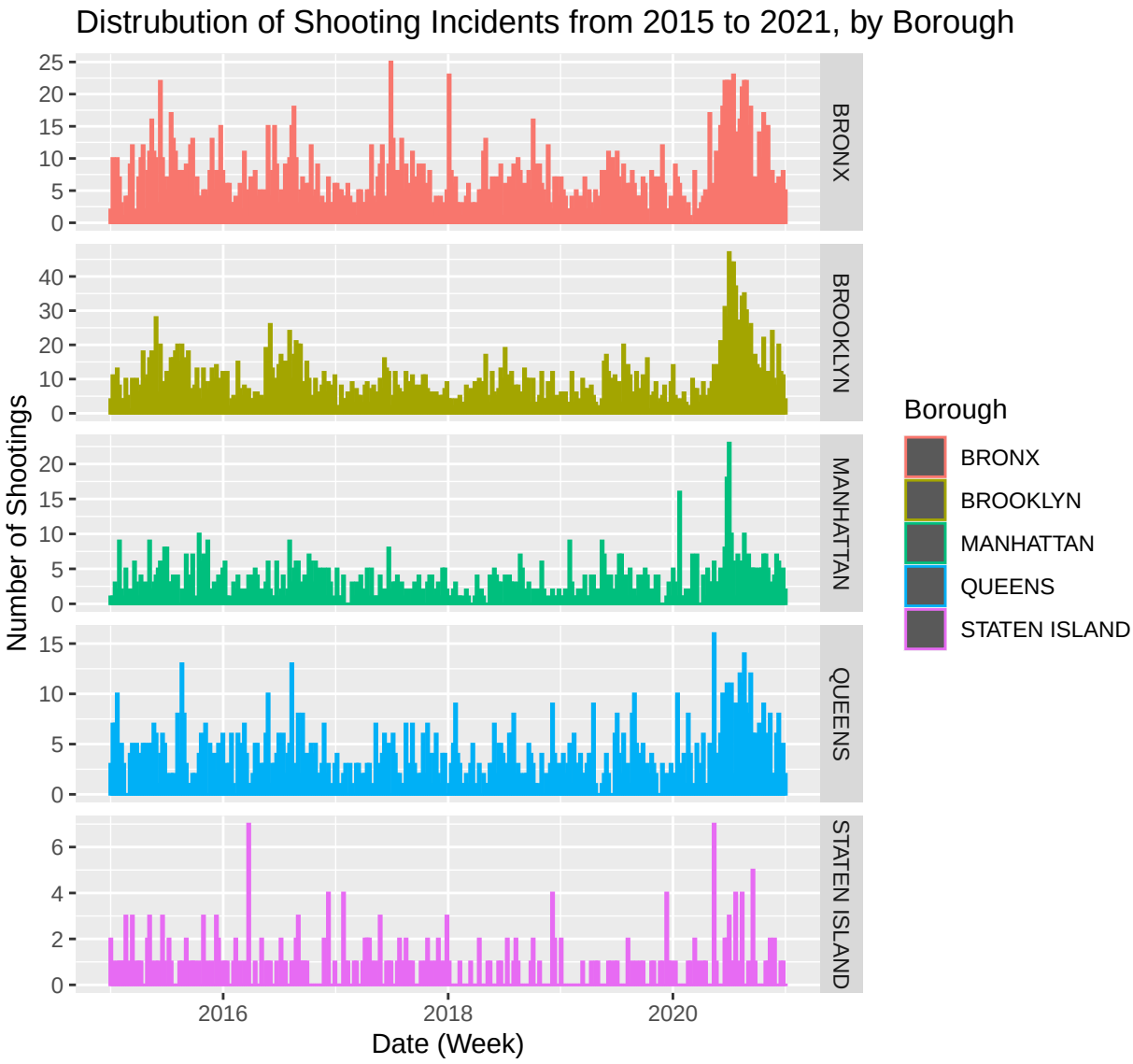


It's clear that Brooklyn sees the greatest number of shooting incidents, with a max of around 50 per week, and Staten Island has the least, with a max of only about 13.You can also see that there was a period of fewer incidents between 2014 ish to 2019. This trend looks strongest in Brooklyn, and is seen the least in the Bronx. It also clear across all Boroughs that there was an increase in incidents in 2020. These trends can be seen more clearly in the next graph which groups the data by year. The increase in the last couple years could potentially be related to the COVID19 pandemic, but further analysis would be needed to say for sure.

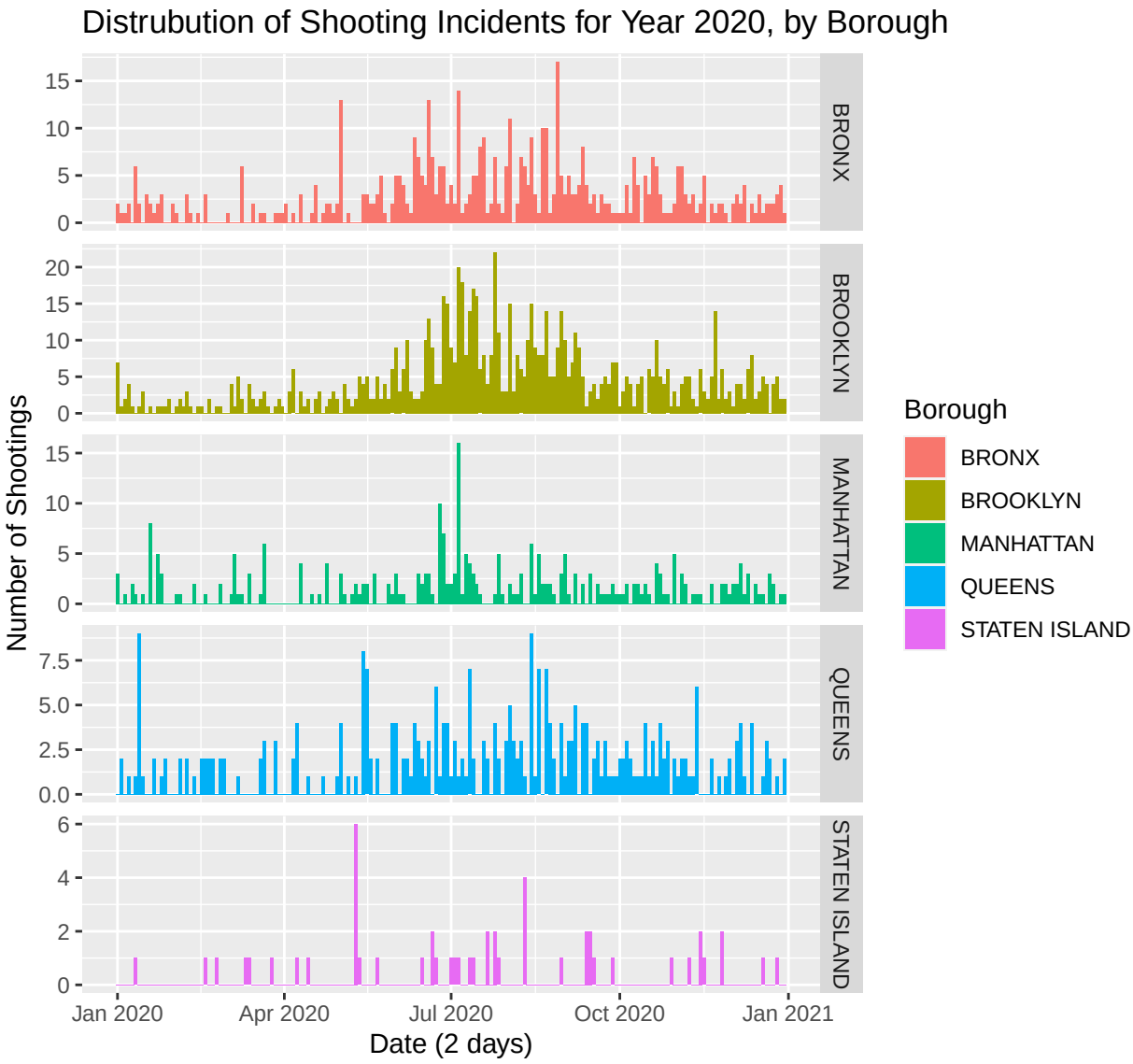


In this plot it is clear that the number of events happening each year did in fact dip between 2015 and 2020, with all the boroughs seeing their lowest anual incidents since 2006 in 2018 or 2019.

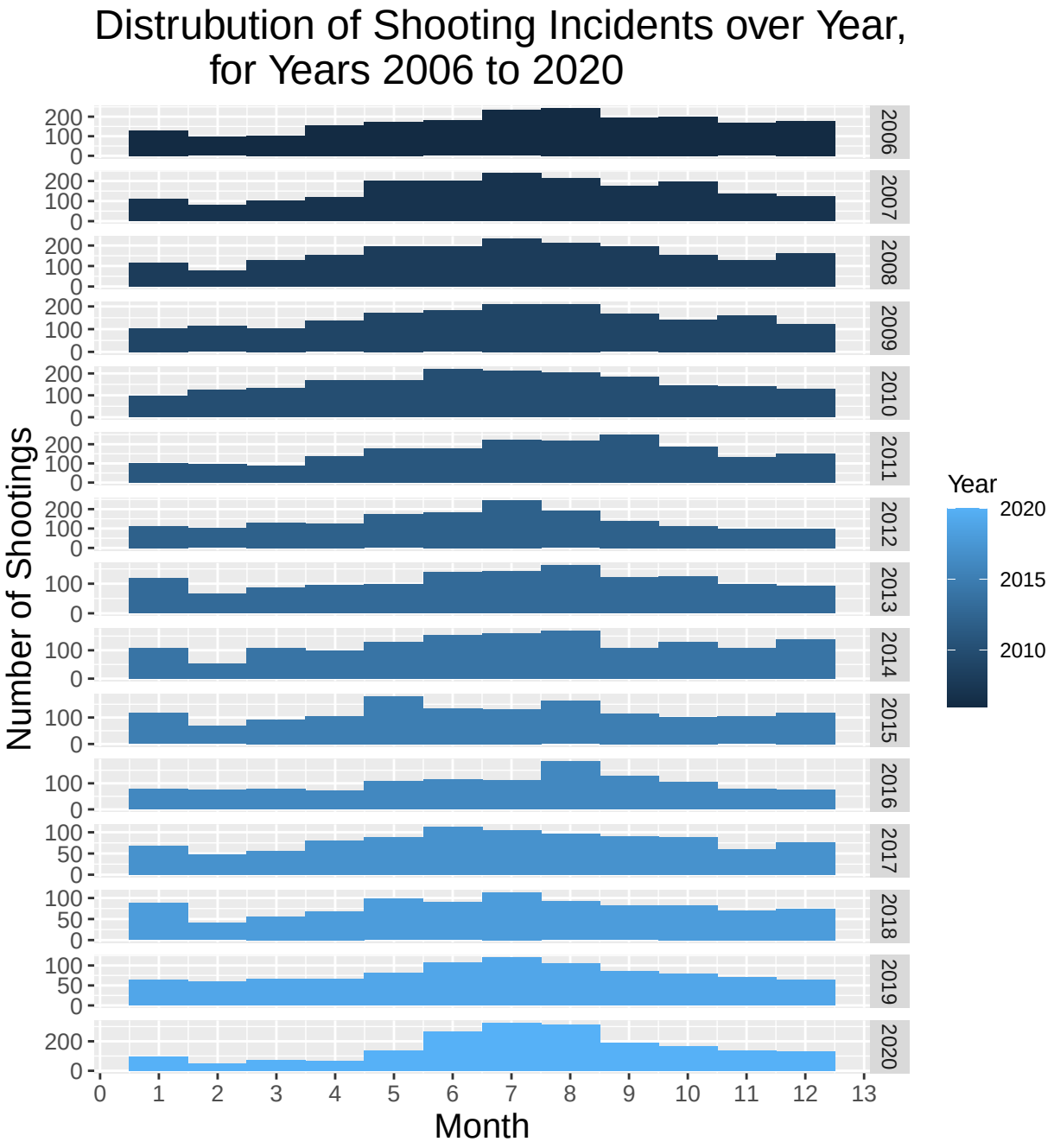
Another thing to note about the first graph is that the number of incidents seems to cycle between periods of higher and lowers numbers incidents within the year. To examine this further a smaller period of time can be looked at. For this a new data frame is created containing data from 2015 to today.



This trend of cycles within each year can still be seen, and it looks like there are certian months that routinley have more incidents, but this might be clearer if we look at an even smaller sample of the data. The next plot shows data from 2020 onwards, and from this you can see there does seem to be a yearly cycle in the number of shooting incidents.



Each bar on this plot represents two days. From this plot you can see that there is substantially more incidents in the summer and into the fall months, at least for the year 2020. It would be interesting to see if this trend is the same for every year. To find out we can use a similar plot, but instead of faceting by borough we will facet by year. In this next plot there is no need to look at borough as those trends have already been examined.

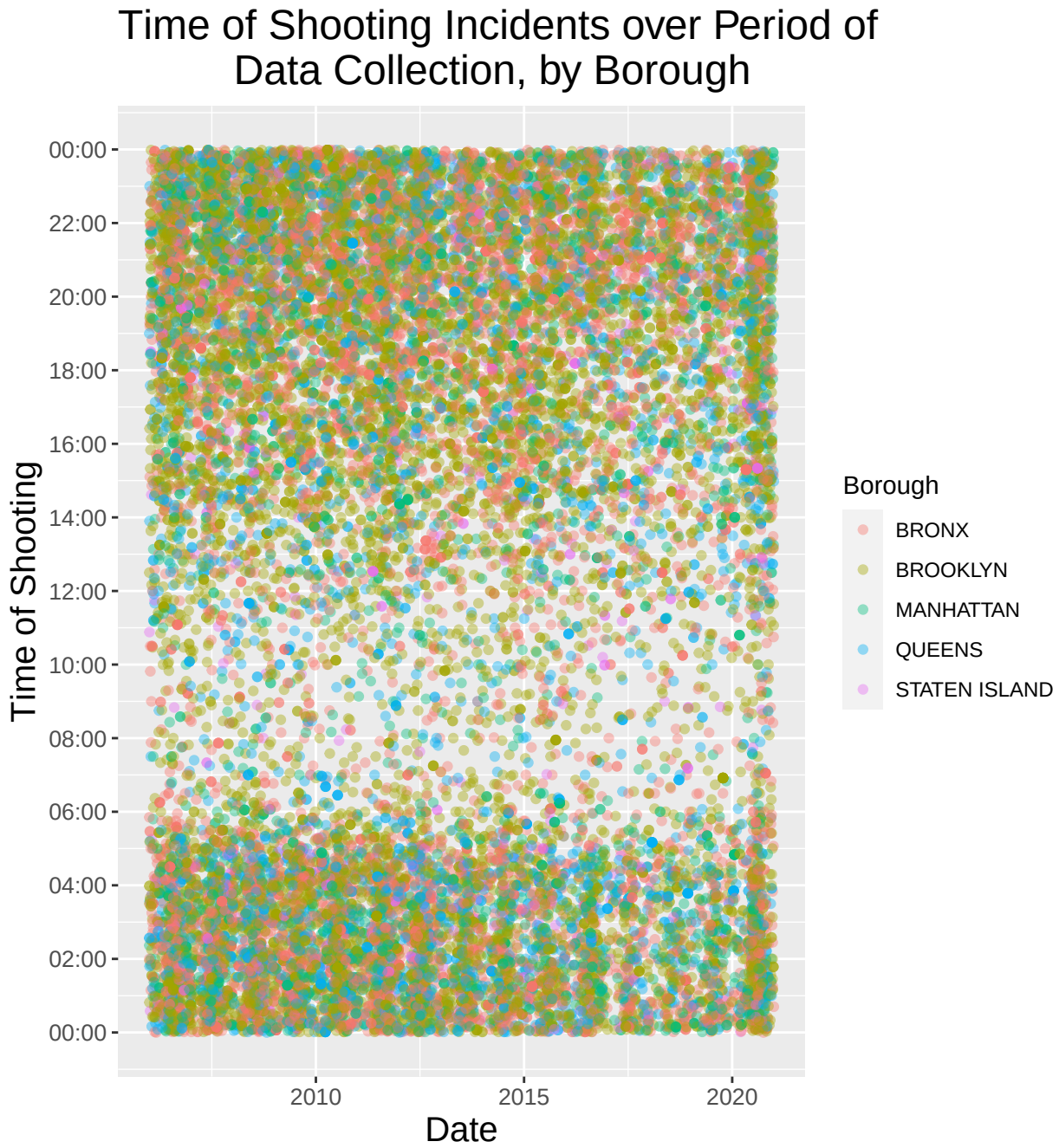


This plot confirms that year over year there does seem to be more incidents in the summer. In each of the years that data was collected for, the month with highest number of shootings is either July or Agust, and the lowest is most commonly February, but somethime January or march. It would be interesting to average monthly totals by day to see if the lowest month is often Feburary as there are less days in that month, or if shootings are in fact less likely at that time of year. The other trends may be related to weather, tourism, or a combination of many other factors, but to say for sure you would need to do further analysis.

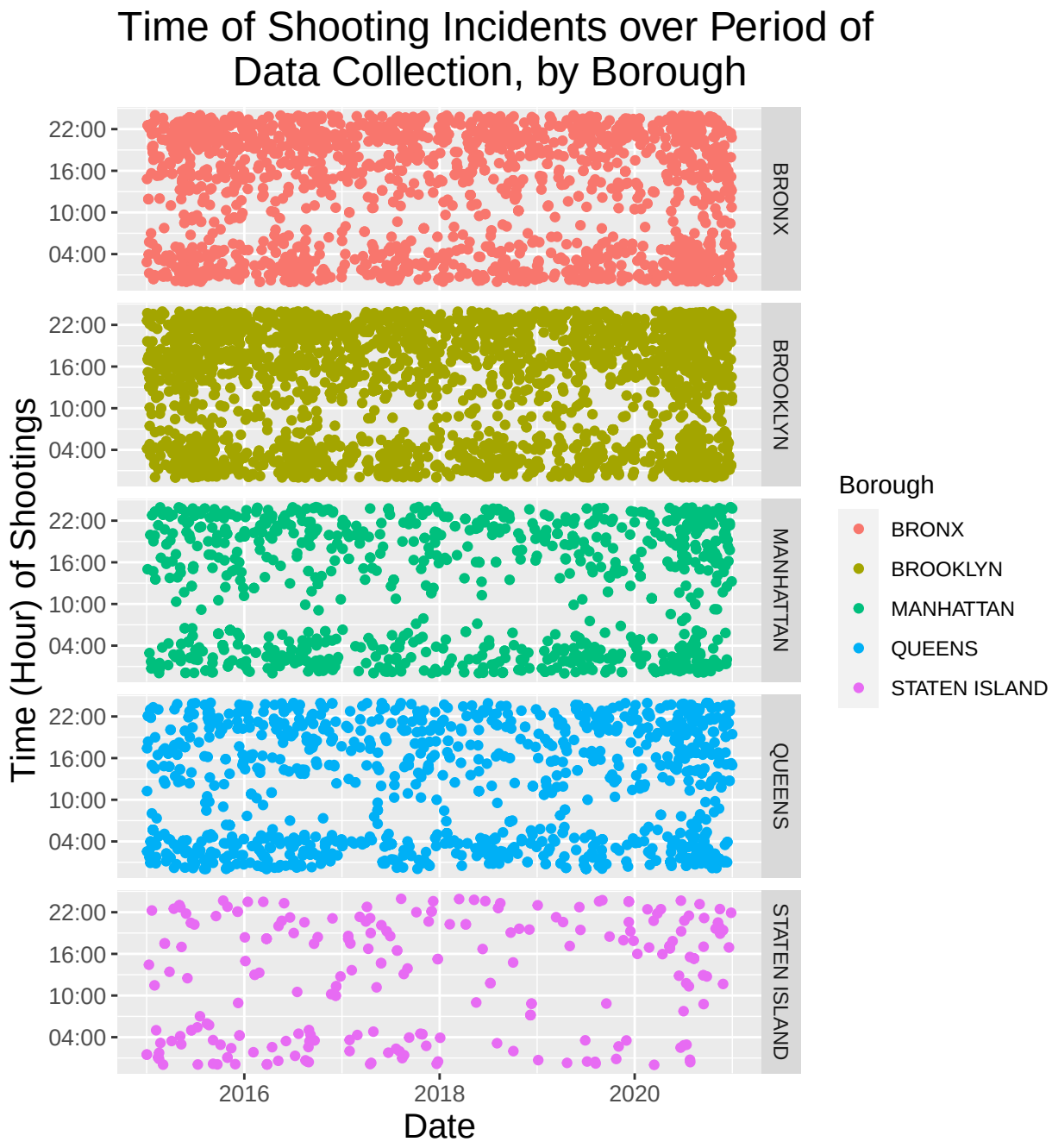
Frequency of Events over Time of Day

The next part of the data looked at is the time of day that shooting incidents took place. The first plot is going to look at the time of day that each incident took place over the entire time period the data covers. It

will also show the borough in which each incident took place to determine if there is a tendency for incidents to happen at a certain time in each localtion.



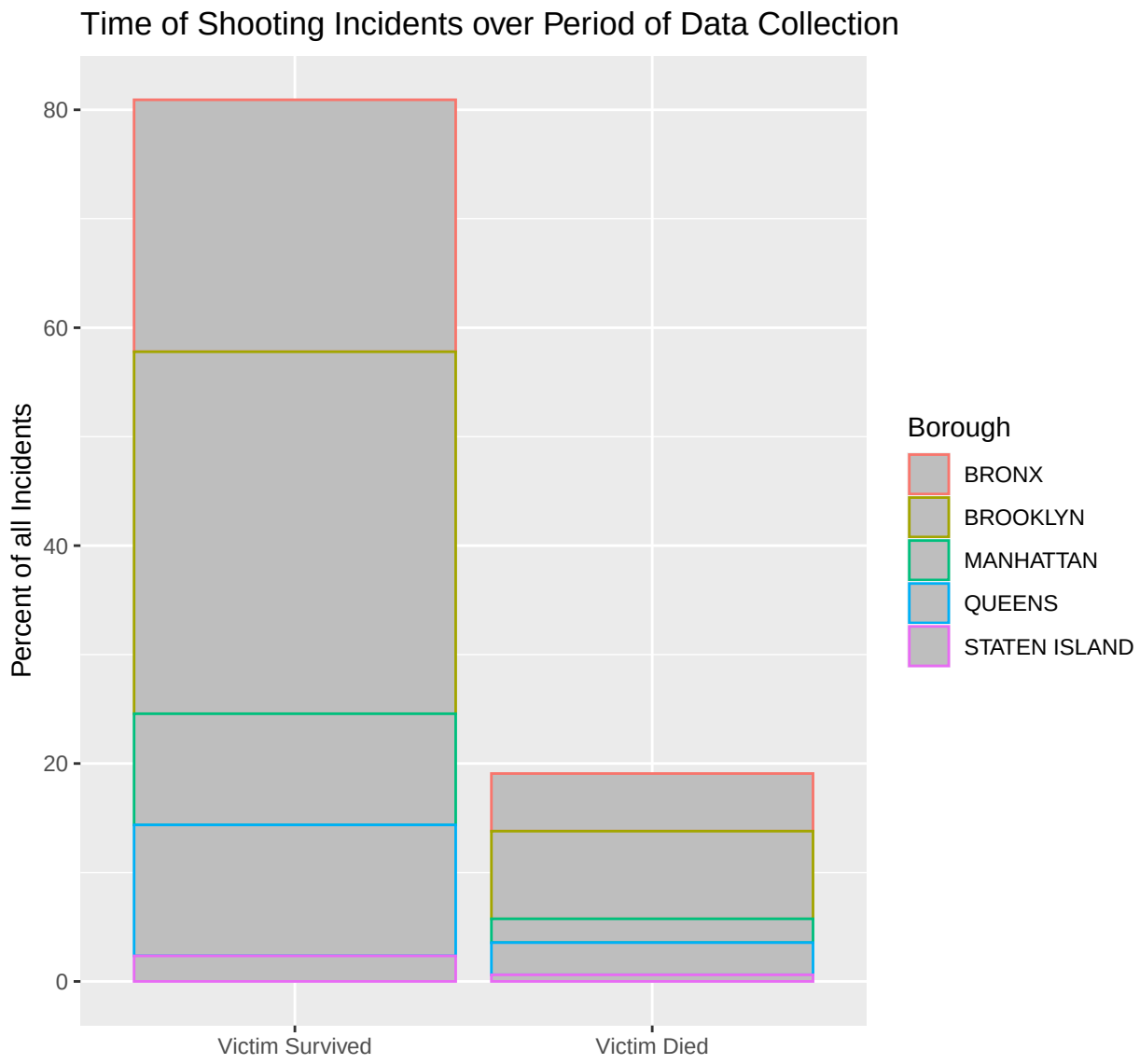
This plot shows, unsurprisingly, that most of the incidents take place late in the evening or very early in the morning. It is hard with this much data to see any potential trends, but we can look at the smaller subset of data from the 2015 on wards to determine if the timing of incidents differs, year to year, month to month or between boroughs.



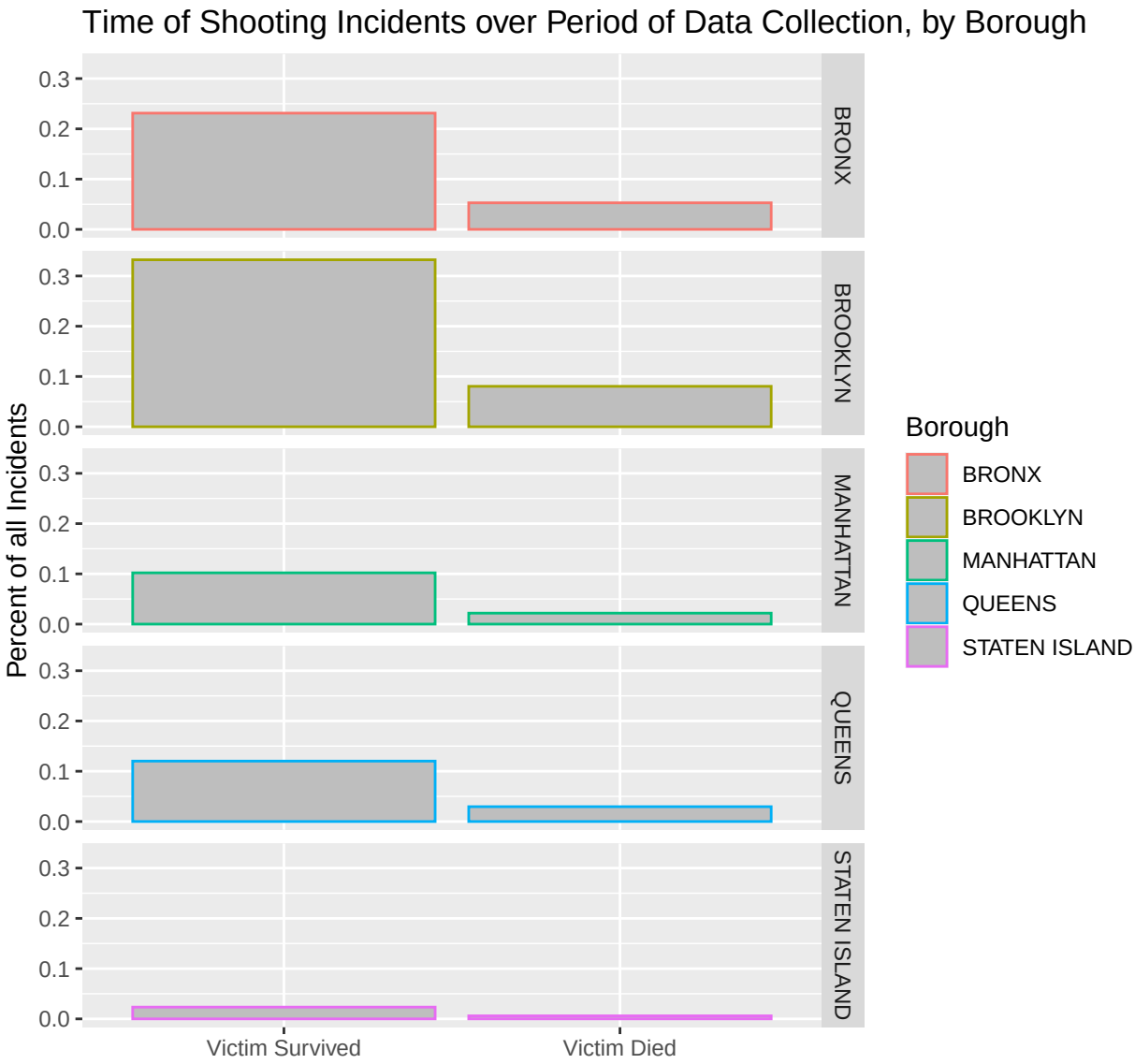
It seems that all the Boroughs exhibit similar trends in the time that shootings take place, however in Brooklyn due to the larger number of incidents more of them occur during the day in off times than in other Boroughs. The time of year seems to have no effect on the time the incidents take place.

Statistical Murder Flag

One interring variable the data set contained was the statistical murder flag, which indicates whether the victim of the shooting was killed, resulting in the incident being classified as a murder. When this variable is true the victim died as a result of the shooting,and when false they survived. The plot below shows this data with the colors representing the borough in which the incident took pack.



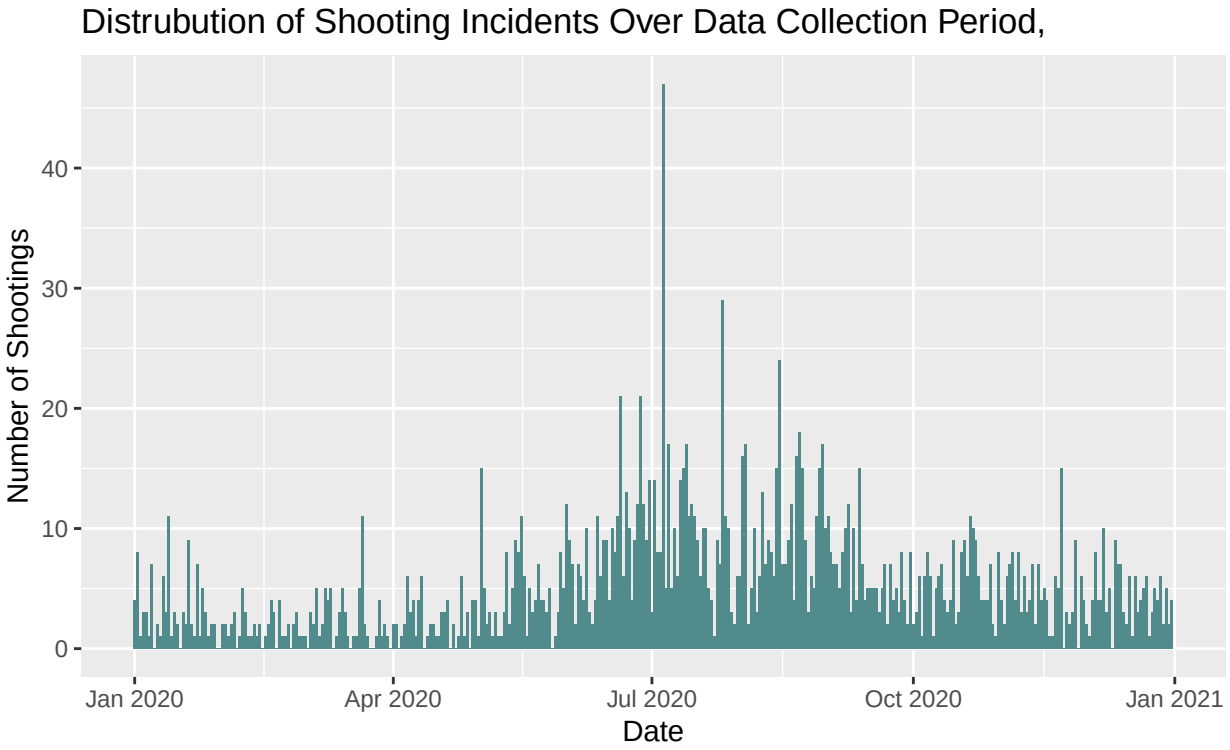
This shows that just over 80% percent of shooting victim do in fact survive. Brooklyn obviously has the most total murders, but surprisingly it looks like there is a relatively greater number shootings from the other boroughs that result in death, to the total numbers in those boroughs. If we split this up by borough it may be easier to observe.



The relative death rate in Brooklyn does seem to be very slightly smaller than the rest of the boroughs, though its hard to tell by exactly how much. This is something that could be further analyzed to be able to say for sure.

Model

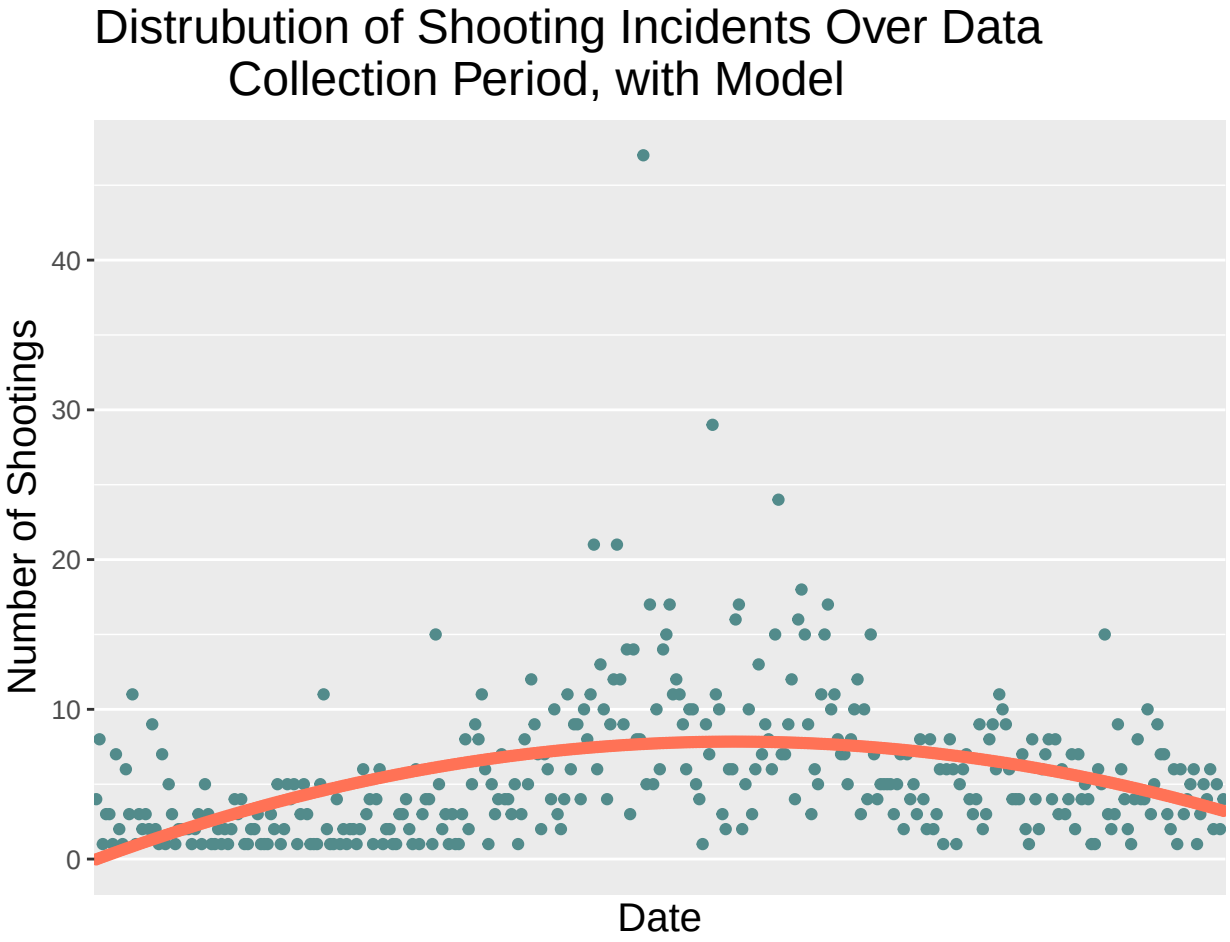
For creating a model, the section of data that we are going to look at one of the plots looked at previously, and is shown again below. The plot looks at the number of shootings that took place from January 1st 2020 to January 1st 2021. As this is the most recent data this, a good model from this data could potentially give an prediction for the number of shootings that will take place in 2021, and help people plan travel or help police and hospitals staff and prepare accordingly.



This representation of the number of incidents looks roughly parabolic, with the greatest number of shootings happening between June and September, so we can try and fit a quadratic model to the data. This is done in the following code block, and the summary of the model created can be seen following it.

```
##
## Call:
## lm(formula = Shootings ~ Day + DaySquared, data = date_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.825 -2.771 -0.748  1.641 39.316
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.005e-01  7.169e-01  -0.140    0.889
## Day          8.167e-02  9.623e-03   8.486 6.69e-16 ***
## DaySquared  -2.099e-04  2.709e-05  -7.747 1.09e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.4 on 340 degrees of freedom
## Multiple R-squared:  0.1819, Adjusted R-squared:  0.1771
## F-statistic: 37.79 on 2 and 340 DF,  p-value: 1.51e-15
```

Now that we have the model created, we can plot it against the actual data to determine how well it fits. The plot described is created below



This model seems like a decent fit to the data that we have, especially in the middle and towards the end. It slightly over predicts around February/March and there is some very high days in the middle of the year that it doesn't catch, but these seem to be outlines and difficult to account for. All around this seems like a sufficiently good model for the data and could provide insight for future years.

Conclusion

There was a lot of information provided to us by this data, and through this analysis we were able to recognize a couple different trends. My investigation of the data was focused on the timing of each of the events throughout the data collection period, year to year, month to month, and then hour to hour as well. I also did some comparisons between the boroughs. It was clear that in the last year, 2020, there was a significant increase in the number of shootings taking place, which may be correlated to the international COVID-19 pandemic. As well I noticed that there was a period of lower total manual incidents from 2015 to 2016, but a potential reason for this would require further investigation. The summer months were also responsible for the greatest number of shootings, with the data following an almost cyclic pattern over the years. Looking at the timing of day each shooting took place, there was a very high percentage of the incidents which occurred during the night. This was not surprising but it was interesting to see the density of incidents between 9pm and 6am.

Looking at the data broken down into Boroughs, Brooklyn had by far the highest number of shootings, with the Bronx in second. Staten Island had the least. Something interesting to look into would be to see if there was any correlation between the populations of each of the boroughs and the number of shootings which

took place in them. When looking into the time of day that each of the events took place I found that the trends were similar across all the boroughs.

Recognition of Bias

There are many areas for potential bias in this data set. The fact that the data is coming from such a large city, and data recorded by individual precincts across the city means there is a possibility that the data is not being recorded in the same manner. Some people may have different ideas about what neighborhoods make up each borough and so that data make have been recorded differently. Timing of this incident could differ as some could record it as when the incident actually took place vs when it was reported. These are just a few of the potential sources of bias in recording and collecting the data.

In this analysis there were some variables that had missing or insufficient data and the decision was made to remove them so that the missing data would not affect our conclusions or introduce bias. Instead of doing this the missing values could have been replaced by the median or most common value for that variable. This may have lead to some interesting discoveries but doing so has the potential to increase bias that may already be in the data.

Every person performing an analysis will also bring some of their bias into their work. I personally am from a country with stricter gun laws, and that has had an affect on my opinions surrounding guns. It's important that our personal bias is not reflected in our work, and to ensure that its not, its good to look at the data from all angles and consider other views as you are working.