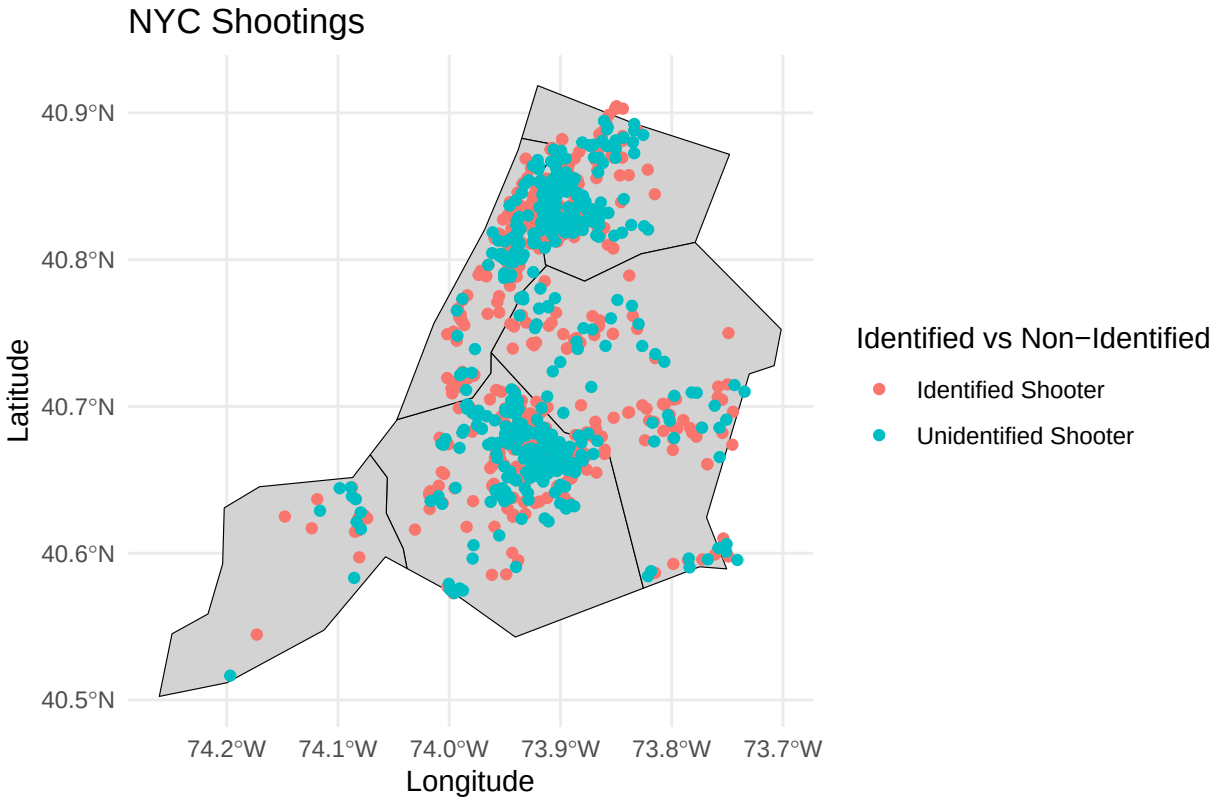
```
##   |                                                                      |
```

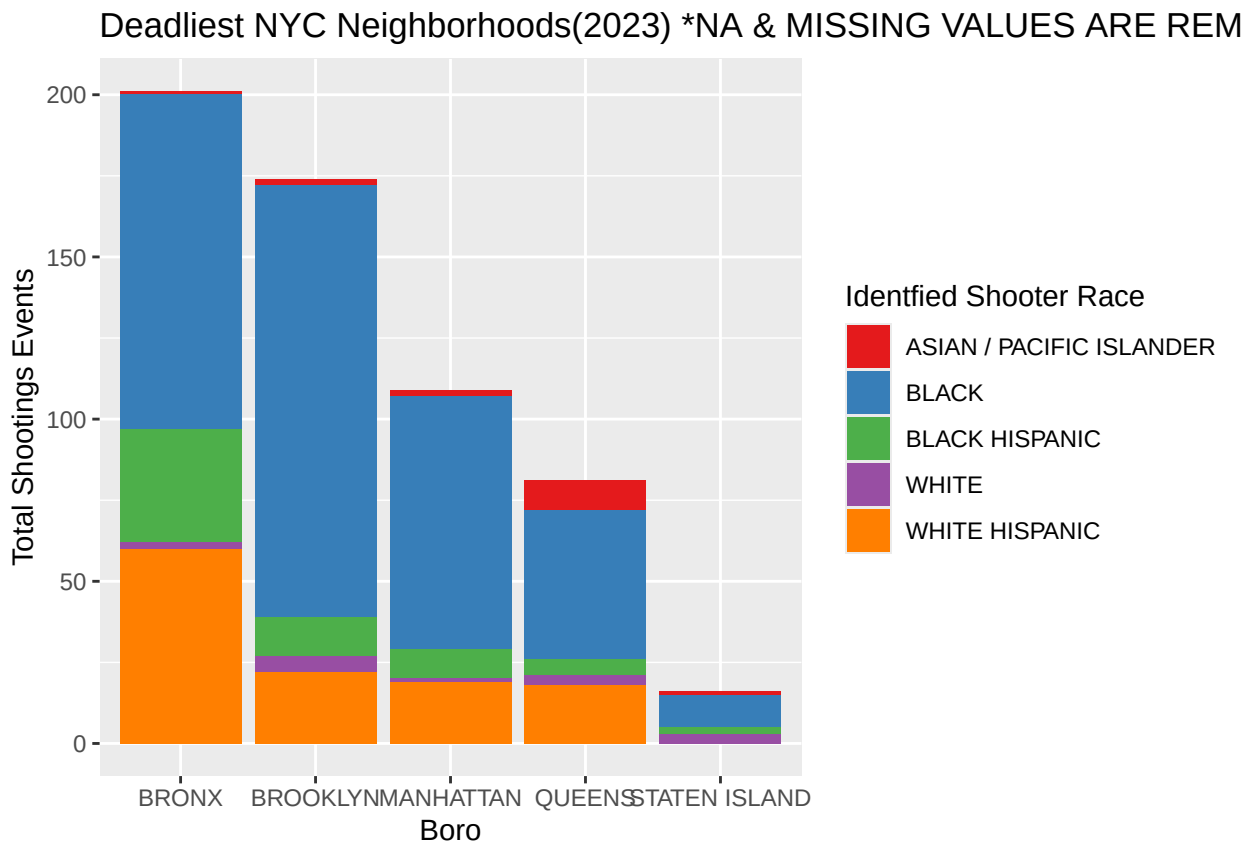



```
##############################################################################################
# The above map displays shooting incidents with unidentified perpetrators forming
# clusters when plotted via (Long,Lat) coordinates.
# Please notice the denselt packed clusters forming in the neighborhoods of Bronx and -
# - Brooklyn particularly where they border Manhattan
# This might indicate some sort of correlation of  inter-communal sections
# that cross-border might perhaps have higher rate of unidentified shooters per total
# recorded shooting events.
#
# The map also indicates that we may have a high number of total shooting events
# that have unidentified shooters which may indicate a possible bias.
#
# **IMPORTANT NOTE: Unidentfied shooters are grouped from cells of col Perp_Race containing
# blank or NA Values
```

```
result <- nyc_data %>%
   filter(str_detect(OCCUR_DATE, "2023") & !str_detect(PERP_RACE, "(null)" ) & !str_detect(PERP_RACE, "
  group_by(BORO, PERP_RACE) %>%
  summarise(unique_count = n_distinct(INCIDENT_KEY))
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```
ggplot(result, aes(x = BORO, y = unique_count, fill=PERP_RACE))  +
  geom_bar(stat = "identity") +
  labs(title = "Deadliest NYC Neighborhoods(2023) *NA & MISSING VALUES ARE REMOVED ",
       x = "Boro",
       y = "Total Shootings Events",
       fill = "Identfied Shooter Race")  + scale_fill_brewer(palette = "Set1")
```

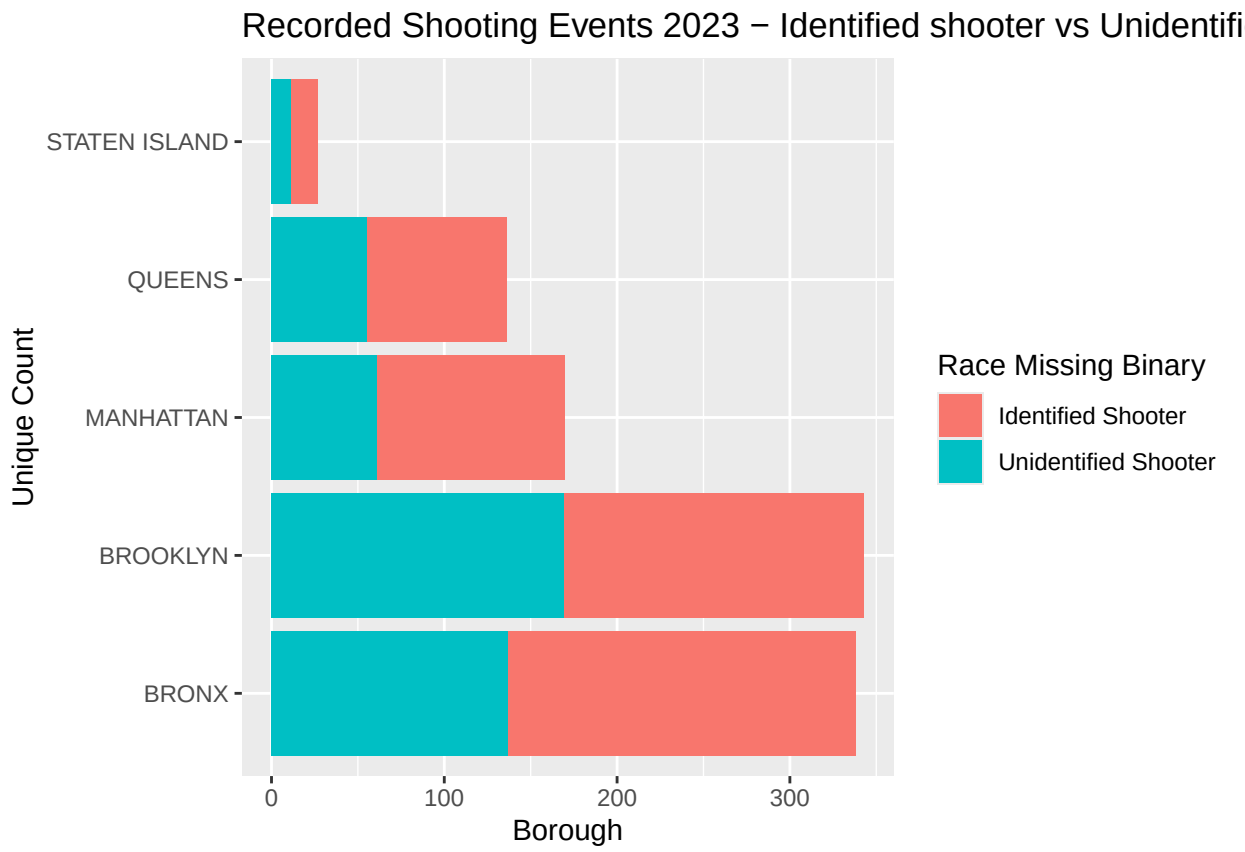



```
##############################################################################################
#
# In the above table we can see that if we remove the NA and Missing values
# this tremendously skews are data when grouping by the shooter's race.
# Let us continue analyzing NA or missing values to asses any potential bias
# or capture any negative impact that would relate to our data integrity.
##############################################################################################
```

```
identified_vs_unidentified_2023 <- nyc_data %>%
   filter(      str_detect(OCCUR_DATE, "2023"))   %>%
  group_by(race_missing_binary, INCIDENT_KEY, Latitude,Longitude,  BORO, PERP_RACE) %>%
   summarise(unique_count = n_distinct(INCIDENT_KEY))
```

```
## `summarise()` has grouped output by 'race_missing_binary', 'INCIDENT_KEY',
## 'Latitude', 'Longitude', 'BORO'. You can override using the `.groups` argument.
```

```
ggplot(identified_vs_unidentified_2023, aes(x = unique_count, y = BORO, fill = race_missing_binary)) +
 geom_col(position = "stack") +
  labs(title = "Recorded Shooting Events 2023 - Identified shooter vs Unidentified Shooter - ",
       x = "Borough",
       y = "Unique Count",
       fill = "Race Missing Binary")
```

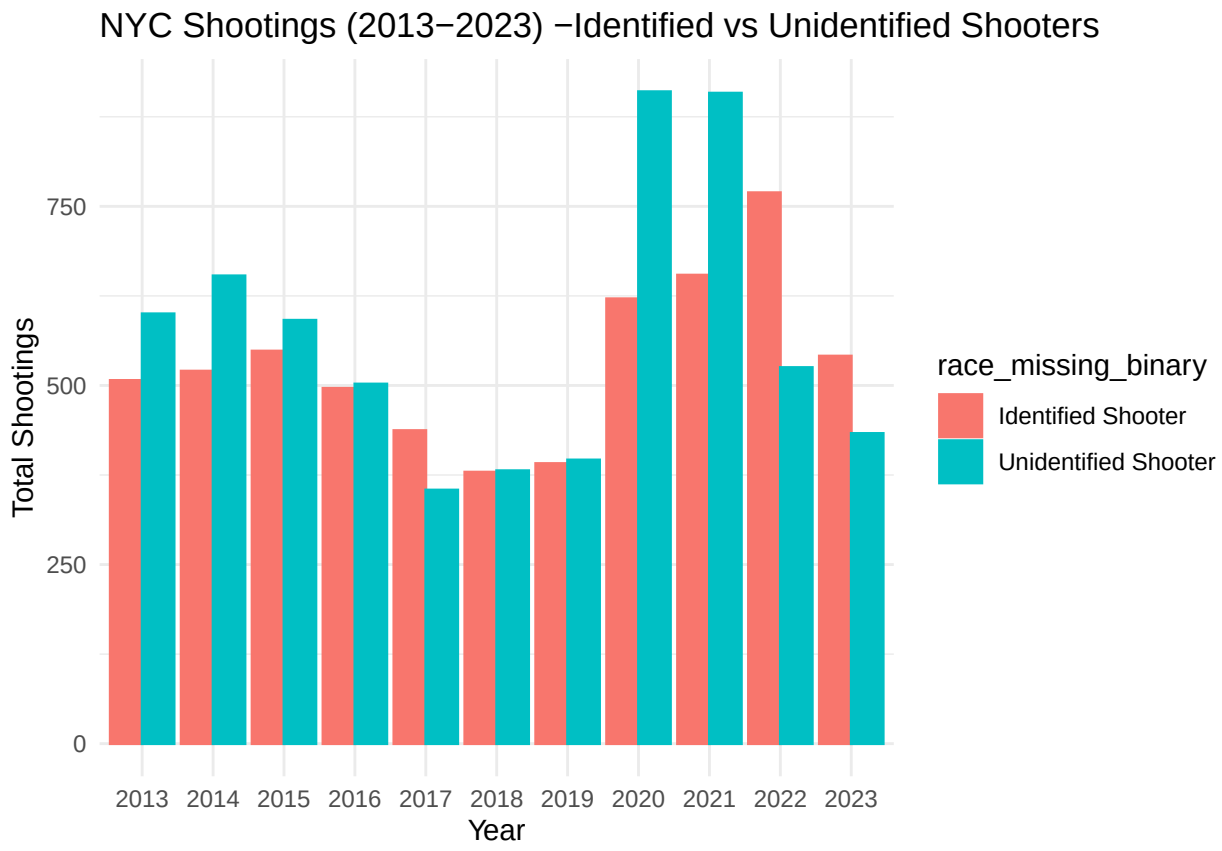



```
##############################################################################################
# The chart indicates that when grouping data for all identified shooters, its number
# is lower than the individuals who remain unidentified.
# This presents a huge data gap but this data view shows us only 2023 data.
# Lets continue to analyze and tidy to get data for a longer date range
##############################################################################################
```

```
nyc_10_year <- nyc_data %>%
  filter(year %in% c(  "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021", "2022",
   group_by( race_missing_binary, year) %>%
   summarise(unique_count = n_distinct(INCIDENT_KEY))
```

```
## `summarise()` has grouped output by 'race_missing_binary'. You can override
## using the `.groups` argument.
```

```r
nyc_10_year$year <- as.factor(nyc_10_year$year)
ggplot(nyc_10_year, aes(x=year, y= unique_count, col=race_missing_binary, fill = race_missing_binary ))
  geom_bar(stat = "identity", position = "dodge")  +
  labs(title = "NYC Shootings (2013-2023) -Identified vs Unidentified Shooters",
       x = "Year",
       y = "Total Shootings") +
  theme_minimal()
```



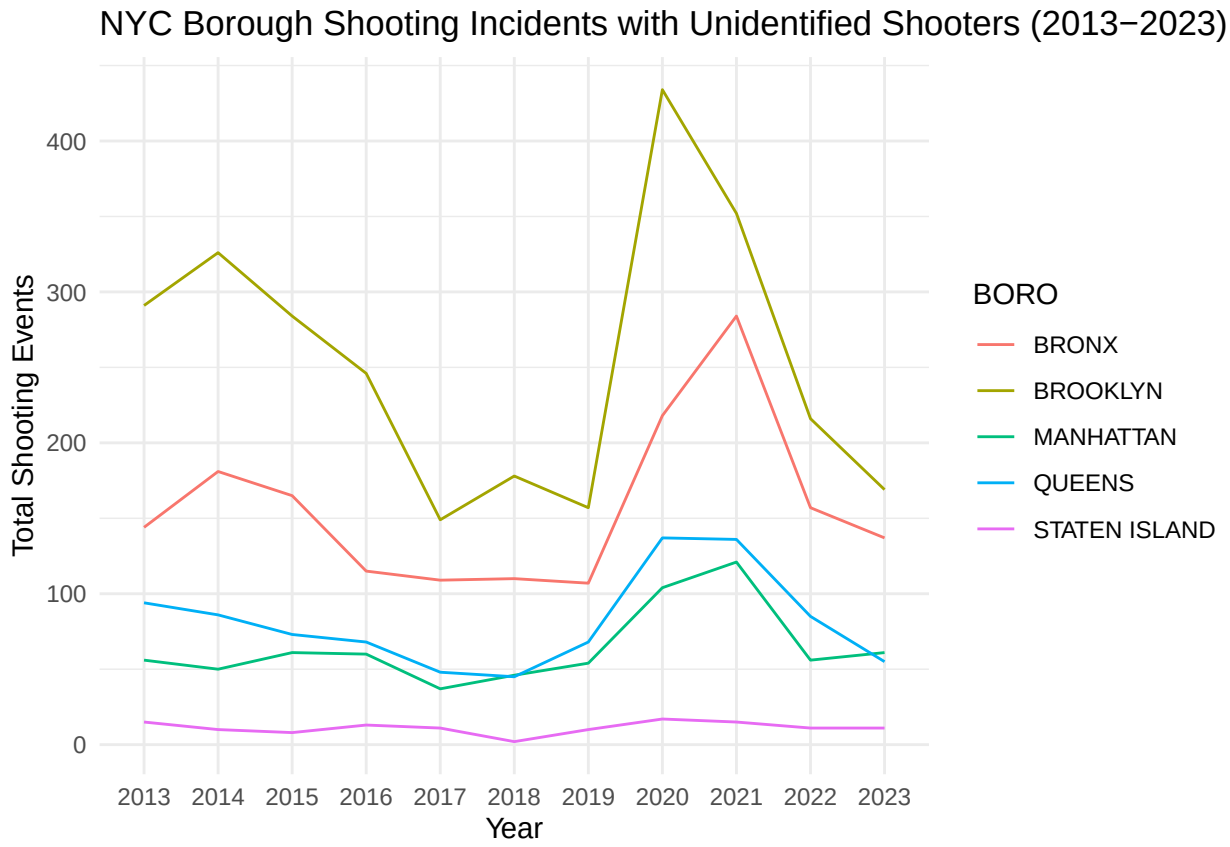

```r
##########################################################################################
# In the above chart we can observe an S shaped curve with levels of
# Unidentified Shooters surpassing Identified Shooters in the 8 out of 10 years
# for the period of (2013-2014)
# We can interpret this as a possible bias in our data as we know that the majority of
# shooters missing identification are from the traditionally lower income neighborhoods
# of Bronx and Brooklyn, particularly in the intersections with Manhattan.
# It is also possible for this to be a result of technical issues on the NYPD side,
# Or perhaps an inability of law enforcement to successfully resolve cases in
# communities with traditionally higher crime rates and higher rates of social exclusion
```

```r
nyc_10_year <- nyc_data %>%
  filter(year %in% c(  "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021", "2022",
  filter(str_detect(race_missing_binary, "Unidentified Shooter")) %>%
   group_by( race_missing_binary, year, BORO) %>%
   summarise(unique_count = n_distinct(INCIDENT_KEY))
```

```
## `summarise()` has grouped output by 'race_missing_binary', 'year'. You can
## override using the `.groups` argument.
```

```
nyc_10_year$year <- as.factor(nyc_10_year$year)


# Plot the data with the correct grouping
ggplot(nyc_10_year, aes(x=year, y= unique_count, col=BORO, group = BORO, fill = BORO )) +
  geom_line(stat = "identity" )  +
  labs(title = "NYC Borough Shooting Incidents with Unidentified Shooters (2013-2023)",
       x = "Year",
       y = "Total Shooting Events") +
  theme_minimal()
```




```
##############################################################################################
# The above line plot suggests that the neighborhoods of Bronx and Brooklyn lead in shootings with
# the shooter to remain permanently unidentified. We can see that the line curves are
# having a rebound level with shooting events in 2020-2021 surpassing the previous
# level high shootings recorded in 2014 period, proceeding to dramaticaly decline accross
# all neighborhoods after year 2021.
##############################################################################################
```

```
# output_file_path <- "C:/Users/kursh/Downloads/NYPD_Shooting_Incidents_QA.csv"
# write.csv(nyc_10_year, file = output_file_path, row.names = FALSE)
```

```r
logi_data <- nyc_data %>%
  filter(year %in% c(  "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021", "2022",
  # filter(str_detect(race_missing_binary, "Unidentified Shooter")) %>%
   group_by( race_missing_binary, year, BORO) %>%
   summarise(unique_count = n_distinct(INCIDENT_KEY))
```

```
## `summarise()` has grouped output by 'race_missing_binary', 'year'. You can
## override using the `.groups` argument.
```

```r
logi_data$race_missing_binary <- ifelse(logi_data$race_missing_binary == "Unidentified Shooter", 1, 0)
logistic_model <- glm(race_missing_binary ~ BORO + year + unique_count, data = logi_data, family = bino

summary(logistic_model)
```

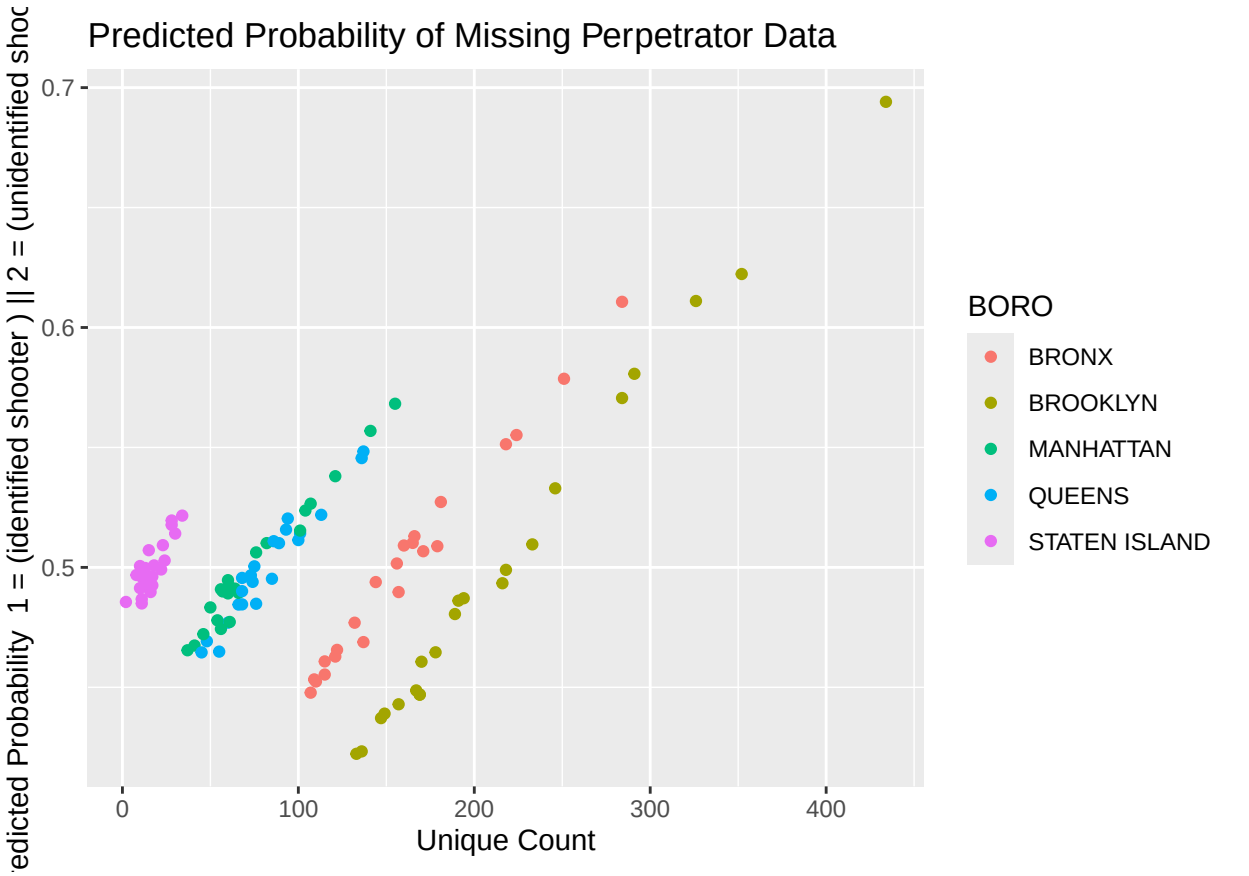
```
##
## Call:
## glm(formula = race_missing_binary ~ BORO + year + unique_count,
##     family = binomial, data = logi_data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        14.235067 123.117201   0.116    0.908
## BOROBROOKLYN       -0.210174   0.654257  -0.321    0.748
## BOROMANHATTAN       0.323381   0.714738   0.452    0.651
## BOROQUEENS          0.296613   0.698157   0.425    0.671
## BOROSTATEN ISLAND   0.544879   0.881976   0.618    0.537
## year               -0.007356   0.061058  -0.120    0.904
## unique_count        0.003812   0.004501   0.847    0.397
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 152.49  on 109  degrees of freedom
## Residual deviance: 151.76  on 103  degrees of freedom
## AIC: 165.76
##
## Number of Fisher Scoring iterations: 4
```

```r
tidy_logistic_model <- tidy(logistic_model)
print(tidy_logistic_model)
```

```
## # A tibble: 7 x 5
##   term              estimate std.error statistic p.value
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)       14.2     123.          0.116   0.908
## 2 BOROBROOKLYN      -0.210     0.654      -0.321   0.748
## 3 BOROMANHATTAN      0.323     0.715       0.452   0.651
## 4 BOROQUEENS         0.297     0.698       0.425   0.671
## 5 BOROSTATEN ISLAND  0.545     0.882       0.618   0.537
```

```
## 6 year              -0.00736   0.0611     -0.120   0.904
## 7 unique_count       0.00381   0.00450     0.847   0.397
```

```
logi_data$predicted_prob <- predict(logistic_model, type = "response")
ggplot(logi_data, aes(x = unique_count, y = predicted_prob, color = BORO)) +
  geom_point() +
  labs(title = "Predicted Probability of Missing Perpetrator Data",
       x = "Unique Count",
       y = "Predicted Probability  1 = (identified shooter ) || 2 = (unidentified shooter)")
```




```
##############################################################################################
#                                Logistic Regression Table
#
# Predicted Variable Breakout: Unidetentified Shooter = 1 vs Identified shooters = 0
#
##############################################################################################
```