NYPD Shooting Incidents

Ian Ho - Harvard Data Science Professional Certificate Program

4/2/2021 - Vancouver, Canada

Contents

1 Executive Summary

In this Capstone Project, we explore a data set named NYPD Shooting Incident Data. I am particularly interested in this data set because of the amount of news headlines in 2020 & 2021 surrounding racial profiling, shooting incidents, and police brutality around the world. My motivating question in this project was: **Can we predict victim's race using date time, location, and victim related data?** I believe there is value in applying machine learning techniques to predicting a victim's race because it can give a better sense of whether or not a demographic, location, and/or datetime has an effect on a particular race is involved in a shooting incident. It is possible that shooting incidents are racially influenced or there may be influence from unknown confounding variables, our goal is to get a better understanding of it through the lens of data science.

To overview the process of this data science project. I initially explored the data and found a number of missing values in namely 4 columns: Location Description, Perpetrator Age Group, Perpetrator Sex, and Perpetrator Race. Because of their high proportion to the overall data set, I did not feel comfortable replacing the missing the values with the mean/mode nor was I comfortable with removing the entire row. I therefore made the judgment to remove these columns from the data set as I believe the large number of missing values would negatively influence the machine learning algorithms. After cleaning the data, I explored the data by looking at counts, proportions, and proportion of deaths in shooting incidents via various lenses in the data set. I then visualized these insights in the following section before diving further into distributions and probabilities of the data set with a focus on the victim's race to gain a better understanding of how a particular predictor may have had an effect on the victim's race. Finally, I applied machine learning algorithms to try and predict the victim races. After observing their accuracies, I decide to cross validate their tuning parameters to obtain potentially higher accuracies while avoiding overfitting. The final model is trained with the entire data set and then tested against the validation set created at the beginning of the script.

Overall classification accuracy is the most important metric in our models because it is equally important for all races to be correctly identified. While sensitivity and specificity are important qualities to have in various classification problems, our goal is to have a balanced accuracy cross all races as they are all equally important to accurately predict. A baseline goal is to have a better prediction than the naive solution, guessing all victim races to be the mode, Black. In our case, that would mean a better accuracy than **0.71487**. Testing multiple models, and lots of trial and error to fine tune each model, I came up with an accuracy of **0.76515** in training. The final model tested against the validation set came up with an accuracy of **0.77819**.

2 Exploratory Data Analysis

2.1 Preliminary Data Exploration

The overall NYPD Shooting Incident Data set has 21624 rows. For the purposes of mimicking an unknown data set, we split the data into 18378 rows (`dat`) for data exploration, analysis, and machine learning training and testing and 3246 rows (`validation`) for testing our final model.

There are 11 columns in the data:

- OCCUR_DATE `<date>` contains the date of the shooting incident.
- OCCUR_TIME `<time>` contains the time of the shooting incident.
- BOROUGH `<character>` contains the borough for where the shooting incident took place in New York City.
- PRECINCT `<numeric>` contains the NYPD precinct that responded to the shooting incident.
- JURISDICTION_CODE `<numeric>` contains the jurisdiction code with respect to the shooting incident.
- STATISTICAL_MURDER_FLAG `<logical>` contains TRUE for a shooting incident causing death and FALSE fora nonfatal shooting incident.
- VIC_AGE_GROUP `<character>` contains age ranges for which the victim of the shooting incident belongs to.
- VIC_SEX `<character>` contains genders for which the victim of the shooting incident belongs to.
- VIC_RACE `<factor>` contains races for which the victim of the shooting incident belongs to. This is the variable we are interested in predicting.
- Longitude `<numeric>` contains the longitudinal geographic coordinate for the shooting incident.
- Latitude `<numeric>` contains the latitudinal geographic coordinate for the shooting incident.

Inspecting the first 10 rows of the data frame:

OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE
2019-08-23	22:10:00	QUEENS	103	0
2019-11-27	15:54:00	BRONX	40	0
2019-02-02	19:40:00	MANHATTAN	23	0
2019-10-24	00:52:00	STATEN ISLAND	121	0
2019-08-22	18:03:00	BRONX	46	0
2019-06-07	17:50:00	BROOKLYN	73	0
2019-03-11	16:30:00	BROOKLYN	81	0
2019-10-03	01:45:00	BROOKLYN	67	0
2019-07-10	02:56:00	BROOKLYN	69	0
2019-06-03	23:05:00	BRONX	46	0

STATISTICAL_MURDER_FLAG	VIC_AGE_GROUP	VIC_SEX	VIC_RACE	Longitude	Latitude
FALSE	25-44	M	BLACK	-73.808	40.698
FALSE	25-44	F	BLACK	-73.919	40.819
FALSE	18-24	M	BLACK HISPANIC	-73.945	40.792
TRUE	25-44	F	BLACK	-74.166	40.638
FALSE	18-24	M	BLACK	-73.913	40.855
FALSE	25-44	M	BLACK	-73.908	40.680
FALSE	25-44	M	BLACK	-73.939	40.688
TRUE	25-44	M	BLACK	-73.926	40.645
FALSE	25-44	M	BLACK	-73.898	40.649
FALSE	18-24	M	WHITE HISPANIC	-73.895	40.861

Below we explore some of the values we see in some columns:

- 2006-01-01 is the earliest shooting incident date as found in OCCUR_DATE.
- 2019-12-31 is the latest shooting incident date as found in OCCUR_DATE.
- QUEENS, BRONX, MANHATTAN, STATEN ISLAND, BROOKLYN are all the different boroughs in New York City under the BORO column.
- 103, 40, 23, 121, 46, 73, 81, 67, 69, 101, 120, 75, 45, 49, 105, 61, 48, 47, 25, 44, 52, 114, 34, 71, 102, 63, 60, 77, 42, 41, 113, 83, 79, 43, 88, 26, 70, 32, 110, 28, 108, 106, 62, 33, 9, 30, 5, 90, 84, 72, 17, 122, 7, 20, 109, 107, 19, 115, 50, 112, 1, 100, 10, 104, 24, 123, 94, 14, 76, 66, 68, 6, 78, 13, 18, 111 are all the different precincts in New York City under the PRECINCT column.
- 0, 2, 1 are the jurisdiction codes in New York City under JURISDICTION_CODE.
- 25-44, 18-24, 45-64, <18, 65+, UNKNOWN are the different age groups related to victims of shooting incidents in VIC_AGE_GROUP.
- M, F, U are the different genders related to victims of shooting incidents in VIC_SEX.
- BLACK, BLACK HISPANIC, WHITE HISPANIC, WHITE, UNKNOWN, ASIAN / PACIFIC ISLANDER, AMERICAN INDIAN/ALASKAN NATIVE are the different races related to victims of shooting incidents in VIC_RACE.
- 0.19121 is the proportion of deaths caused by shooting incidents in STATISTICAL_MURDER_FLAG.

2.2 Advanced Data Exploration

2.2.1 Shooting Incidents grouped by Borough

We are interested to see if there is a more likely borough to have shooting incidents and whether or not those shooting incidents are more likely to result in death.

BORO	count	prop	prop_death
BROOKLYN	7566	0.41169	0.19284
BRONX	5249	0.28561	0.18880
QUEENS	2781	0.15132	0.20101
MANHATTAN	2233	0.12150	0.17689
STATEN ISLAND	549	0.02987	0.20036

2.2.2 Top 10 Shooting Incidents grouped by Precinct

Which precincts have the most shooting incidents in New York City? Are some precinct shooting incidents more likely to result in death than others? Below we observe the top 10 precincts invovled in shooting incidents.

PRECINCT	count	prop	prop_death
106	152	0.00827	0.33553
109	76	0.00414	0.27632
107	67	0.00365	0.26866
72	64	0.00348	0.29688
122	42	0.00229	0.40476
5	35	0.00190	0.37143
6	20	0.00109	0.30000
14	20	0.00109	0.30000
112	17	0.00093	0.41176
17	5	0.00027	0.40000

2.2.3 Shooting Incidents grouped by Jurisdiction Code

How are shooting incidents related to jurisdiction codes? Are they evenly distributed across codes or is are certain jurisdiction codes more involved in shooting incidents?

JURISDICTION_CODE	count	prop	prop_death
0	15337	0.83453	0.19834
2	2997	0.16308	0.15449
1	44	0.00239	0.20455

2.2.4 Shooting Incidents grouped by Victim Age Group

What age groups are more likely to be involved in shooting incidents? Below we look at the age groups and the number of shooting incidents, proportion to total shooting incidents, and proportion to death.

VIC_AGE_GROUP	count	prop	prop_death
25-44	7879	0.42872	0.22274
18-24	7126	0.38775	0.16250
<18	2017	0.10975	0.12295
45-64	1180	0.06421	0.24915
65+	126	0.00686	0.36508
UNKNOWN	50	0.00272	0.26000

2.2.5 Shooting Incidents grouped by Victim Sex

Is one gender more likely to be involved in shooting incidents? Below we look at all genders and their involvement in shooting incidents.

VIC_SEX	count	prop	prop_death
M	16669	0.90701	0.18963
F	1699	0.09245	0.20718
U	10	0.00054	0.10000

2.2.6 Shooting Incidents grouped by Victim Race

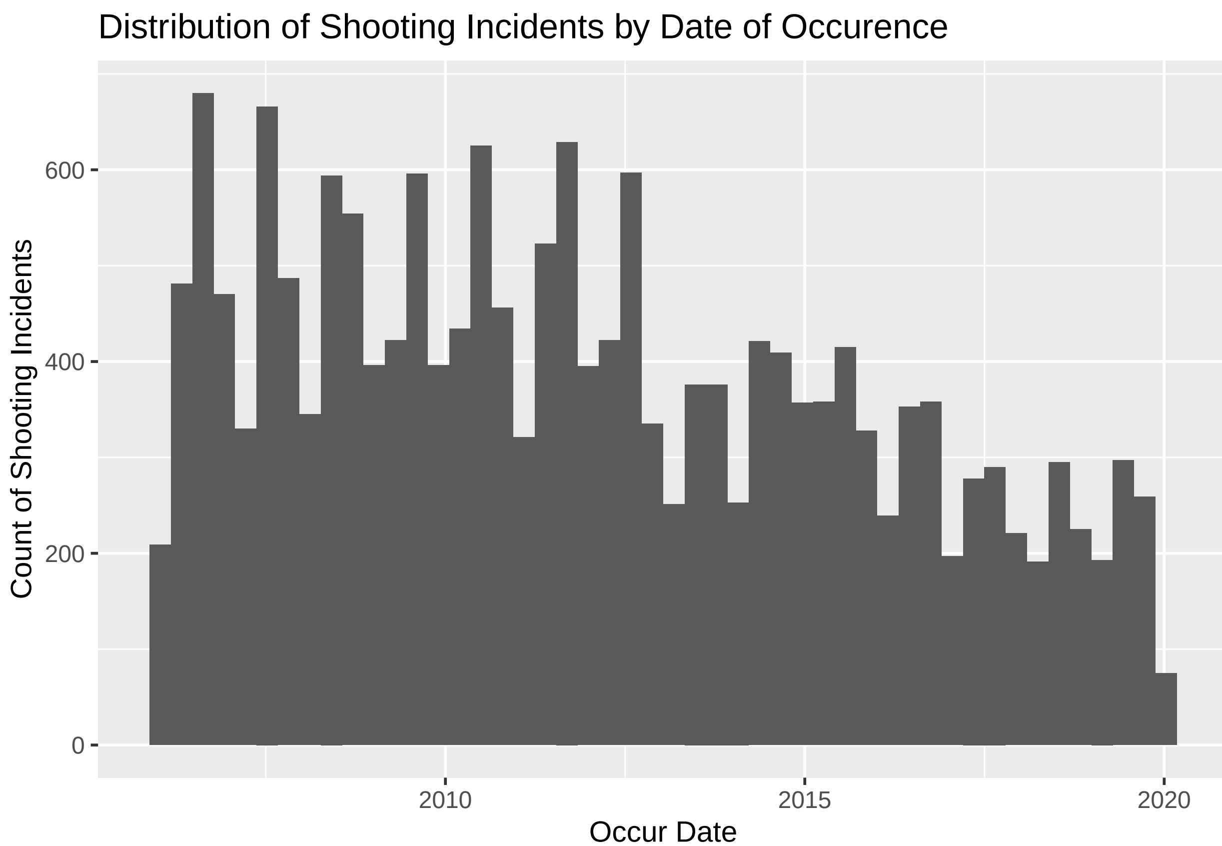
What is the relationship between victim race and shooting incidents? Are some races more likely to be involved in shooting incidents compared to others? Are some races more likely to die?

VIC_RACE	count	prop	prop_death
BLACK	13148	0.71542	0.18550
WHITE HISPANIC	2638	0.14354	0.21418
BLACK HISPANIC	1772	0.09642	0.16479
WHITE	491	0.02672	0.28717
ASIAN / PACIFIC ISLANDER	243	0.01322	0.25926
UNKNOWN	79	0.00430	0.17722
AMERICAN INDIAN/ALASKAN NATIVE	7	0.00038	0.00000

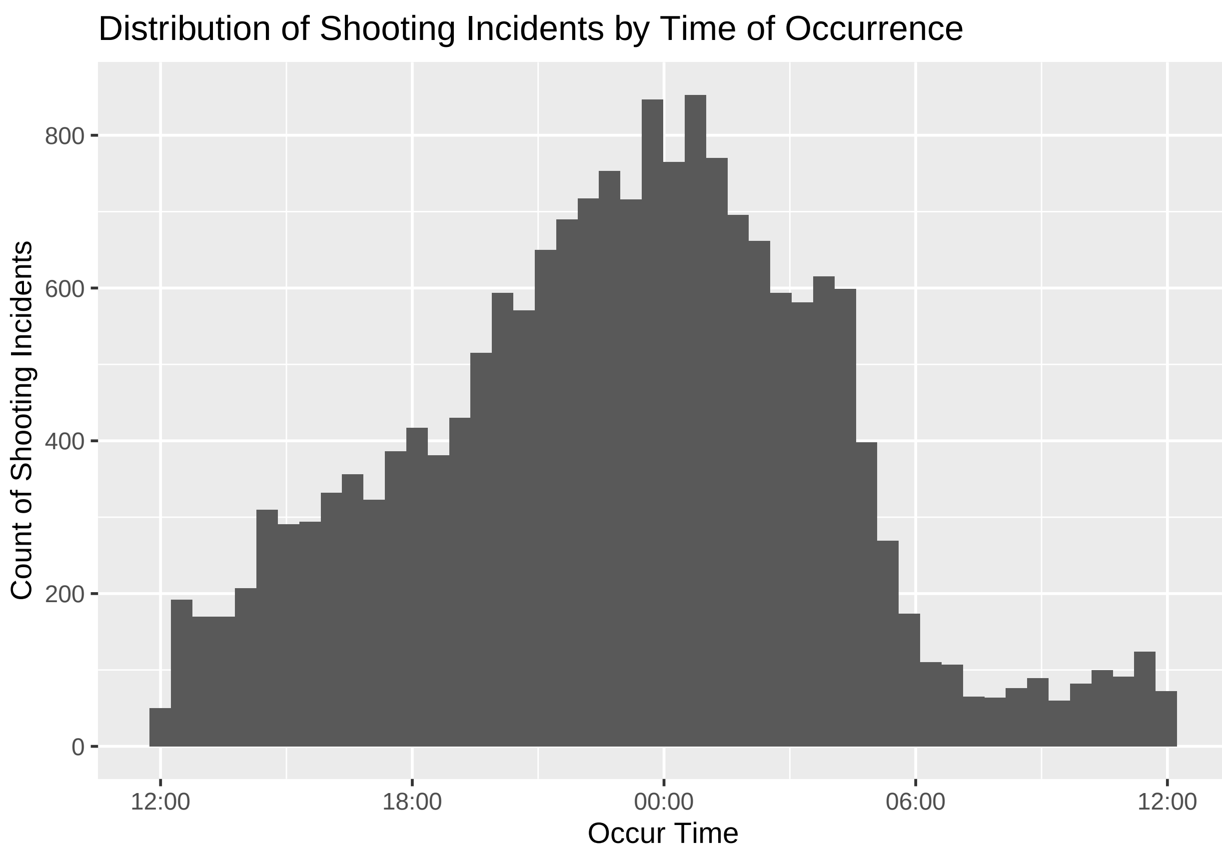
3 Data Visualization

3.1 Distribution Plots

Here we look at the distribution of shooting incidents by occurrence date. This gives us a better idea of when shooting incidents were more likely to occur historically.

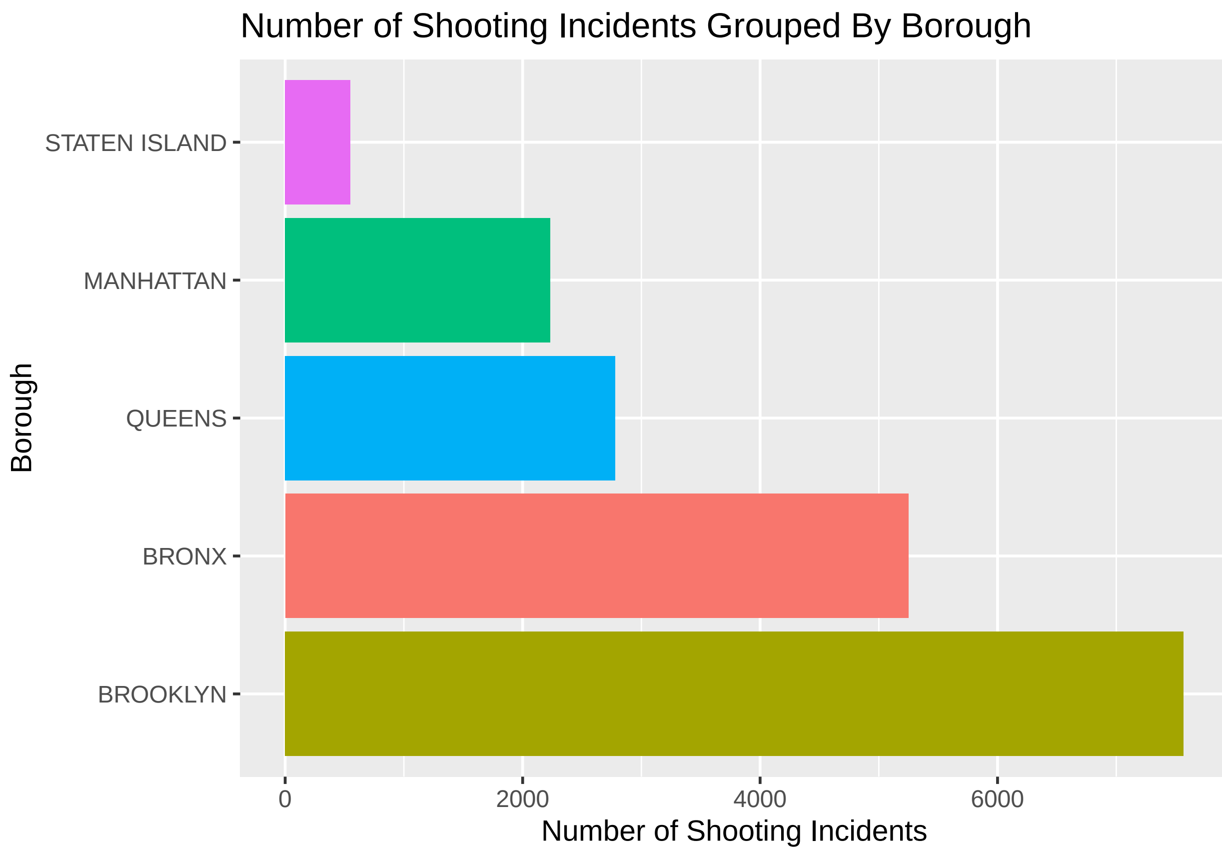


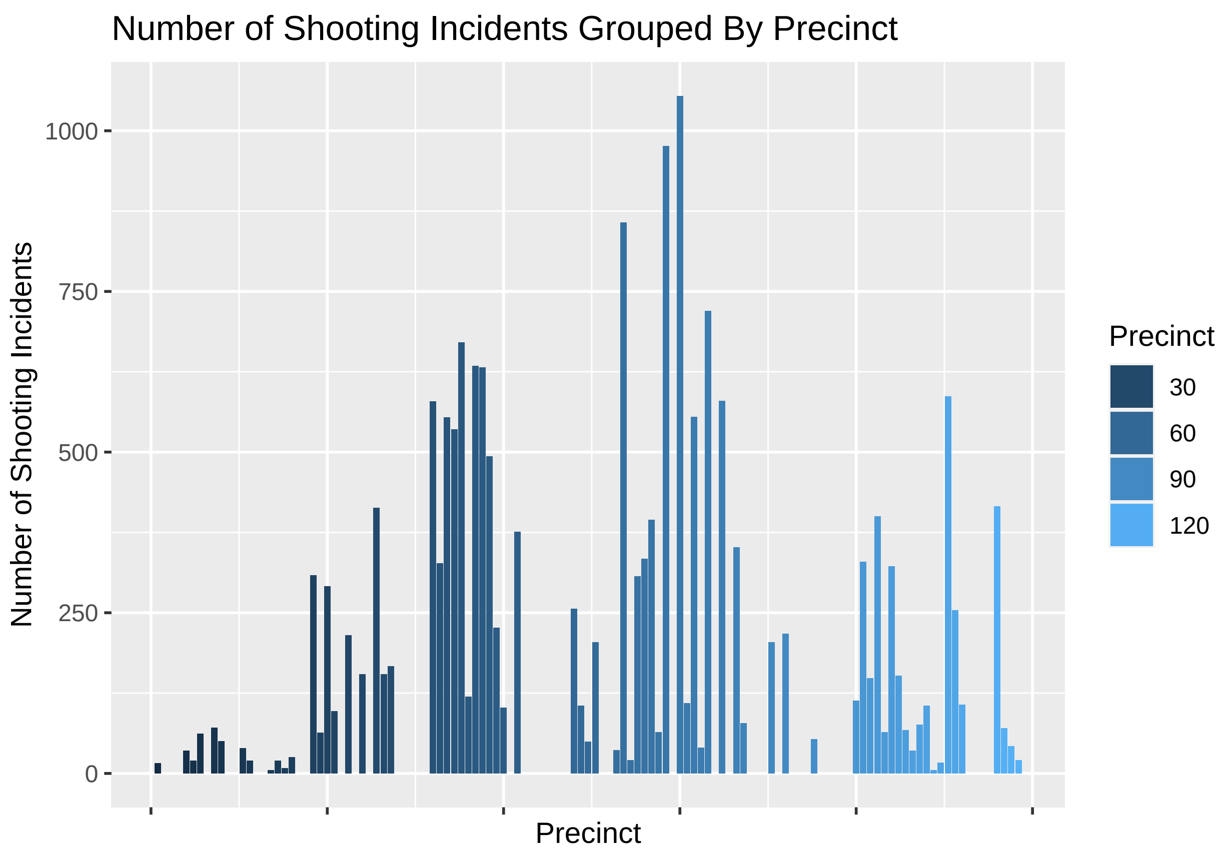
Now we take a look at the distribution of shooting incidents grouped by occurrence time. We originally inspected the data and found that most values tended to center around midnight, below is a modified version of the data to better visualize this finding.


Distribution of Shooting Incidents by Time of Occurrence
Count of Shooting Incidents
800
600
400
200
0
12:00
18:00
00:00
06:00
12:00
Occur Time

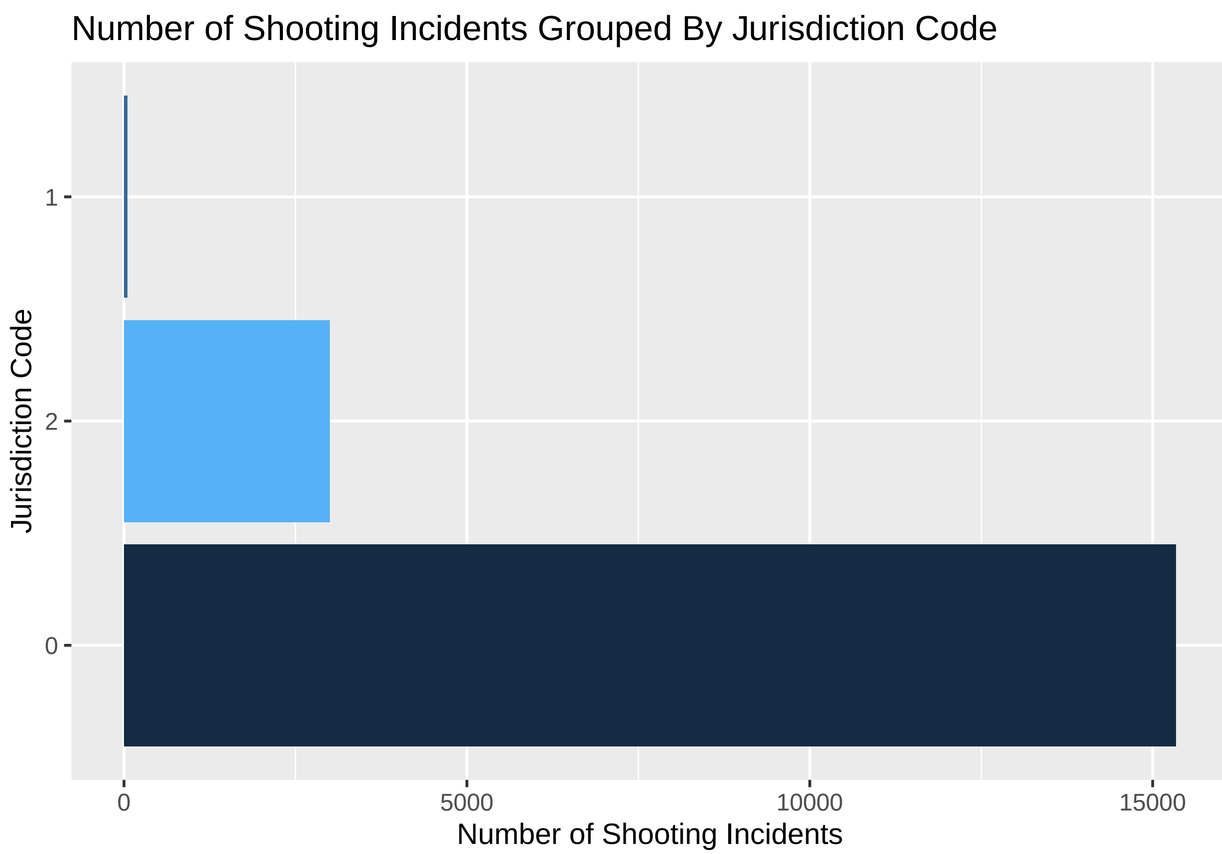
3.2 Barplots

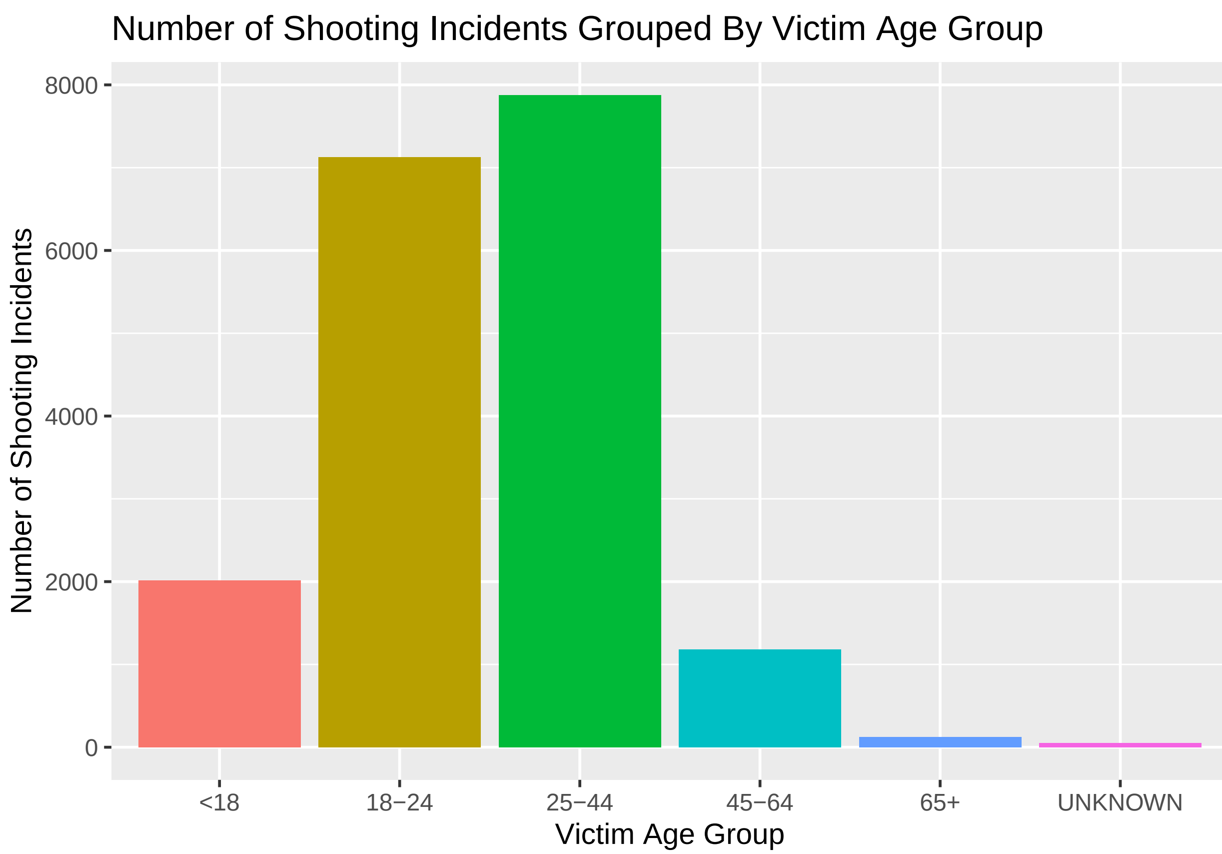
The following barplots illustrate the insights gained in the Advanced Data Exploration section prior. These succinctly show which values in each column are related to the most shooting incidents.



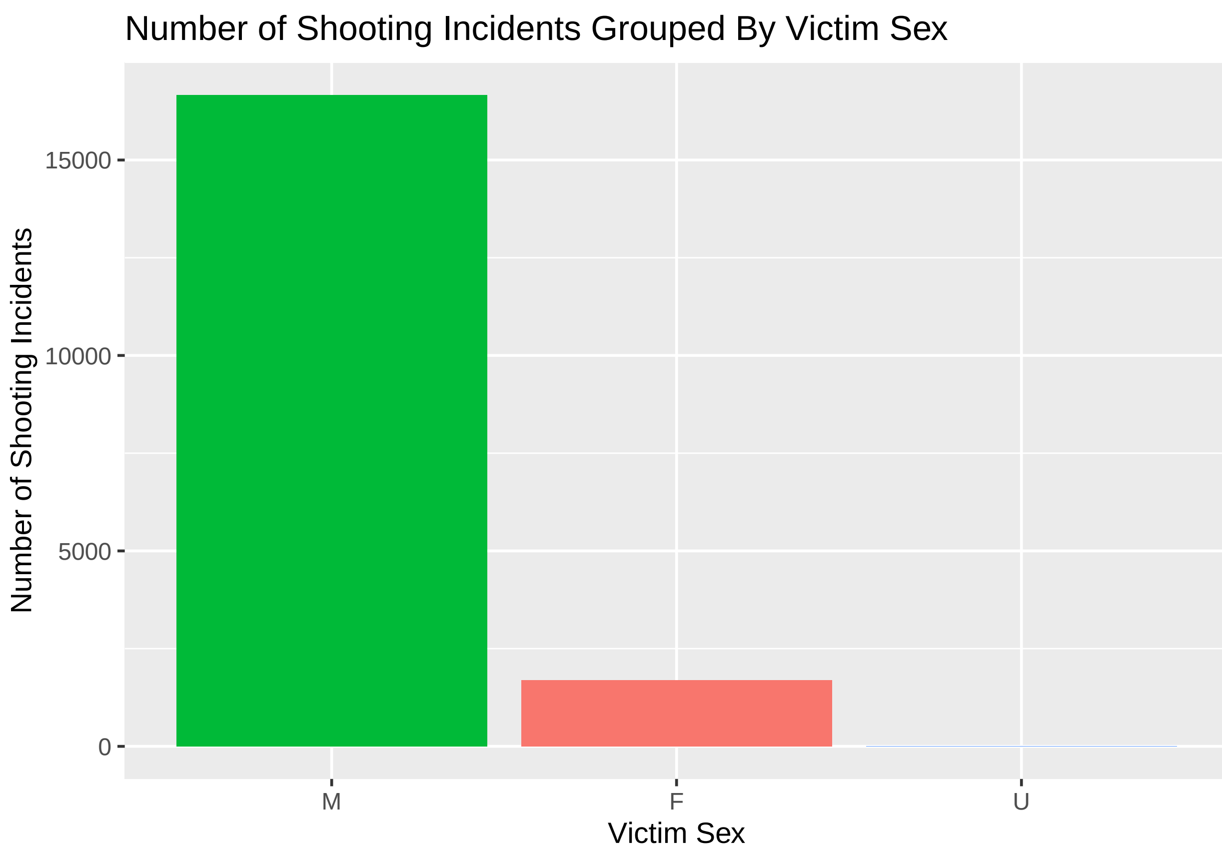


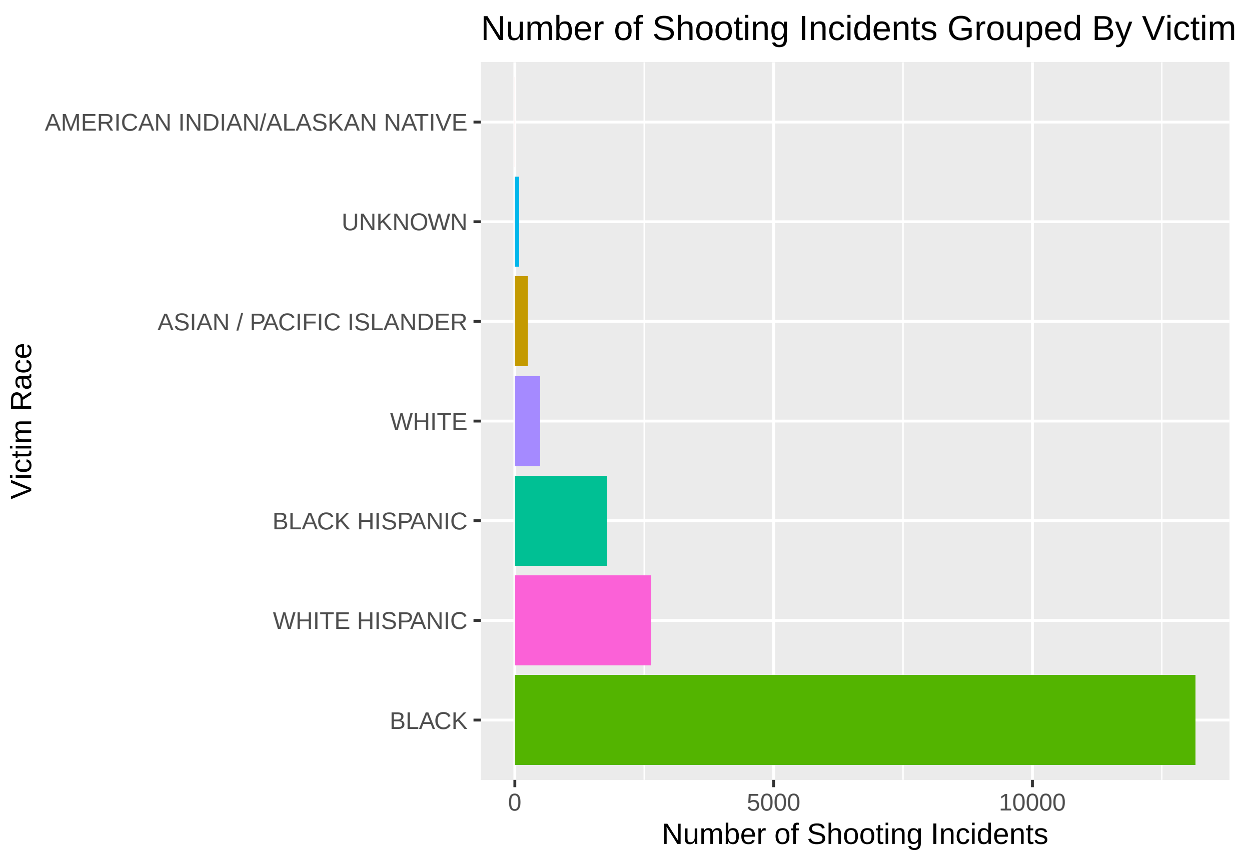
Number of Shooting Incidents Grouped By Precinct
Number of Shooting Incidents
1000
750
500
250
0
Precinct
Precinct
30
60
90
120


Number of Shooting Incidents Grouped By Jurisdiction Code
Jurisdiction Code
1
2
0
0
5000
10000
15000
Number of Shooting Incidents



Number of Shooting Incidents Grouped By Victim Age Group
8000
6000
4000
2000
0
Number of Shooting Incidents
<18
18−24
25−44
45−64
65+
UNKNOWN
Victim Age Group


Number of Shooting Incidents Grouped By Victim Sex
Number of Shooting Incidents
15000
10000
5000
0
M
F
U
Victim Sex


Number of Shooting Incidents Grouped By Victim
Victim Race
AMERICAN INDIAN/ALASKAN NATIVE
UNKNOWN
ASIAN / PACIFIC ISLANDER
WHITE
BLACK HISPANIC
WHITE HISPANIC
BLACK
0
5000
10000
Number of Shooting Incidents

3.3 Geographic Plots

In this cluster plot, we can see that there are longitudinal and latitudinal clusters where more shooting incidents take place. This tells us that the information from `Longitude` and `Latitude` are more specific and useful compared to just using borough information.

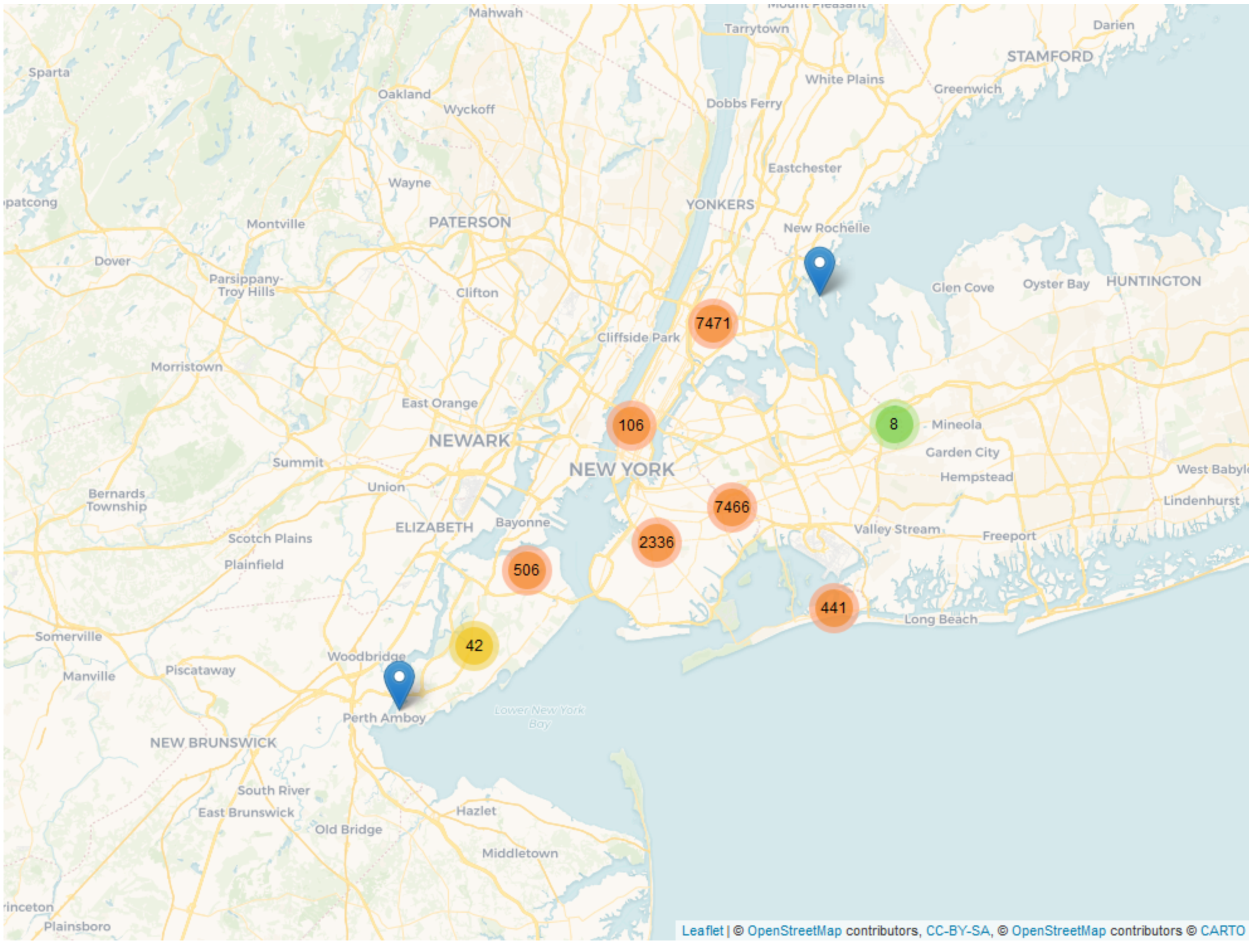


Figure 1: Shooting Incident Clusters in New York City

We can compare cluster plot with the borough plot below. As we can see, shooting incidents in `Queens` is split into multiple clusters due to a higher variability in locations of incidents.

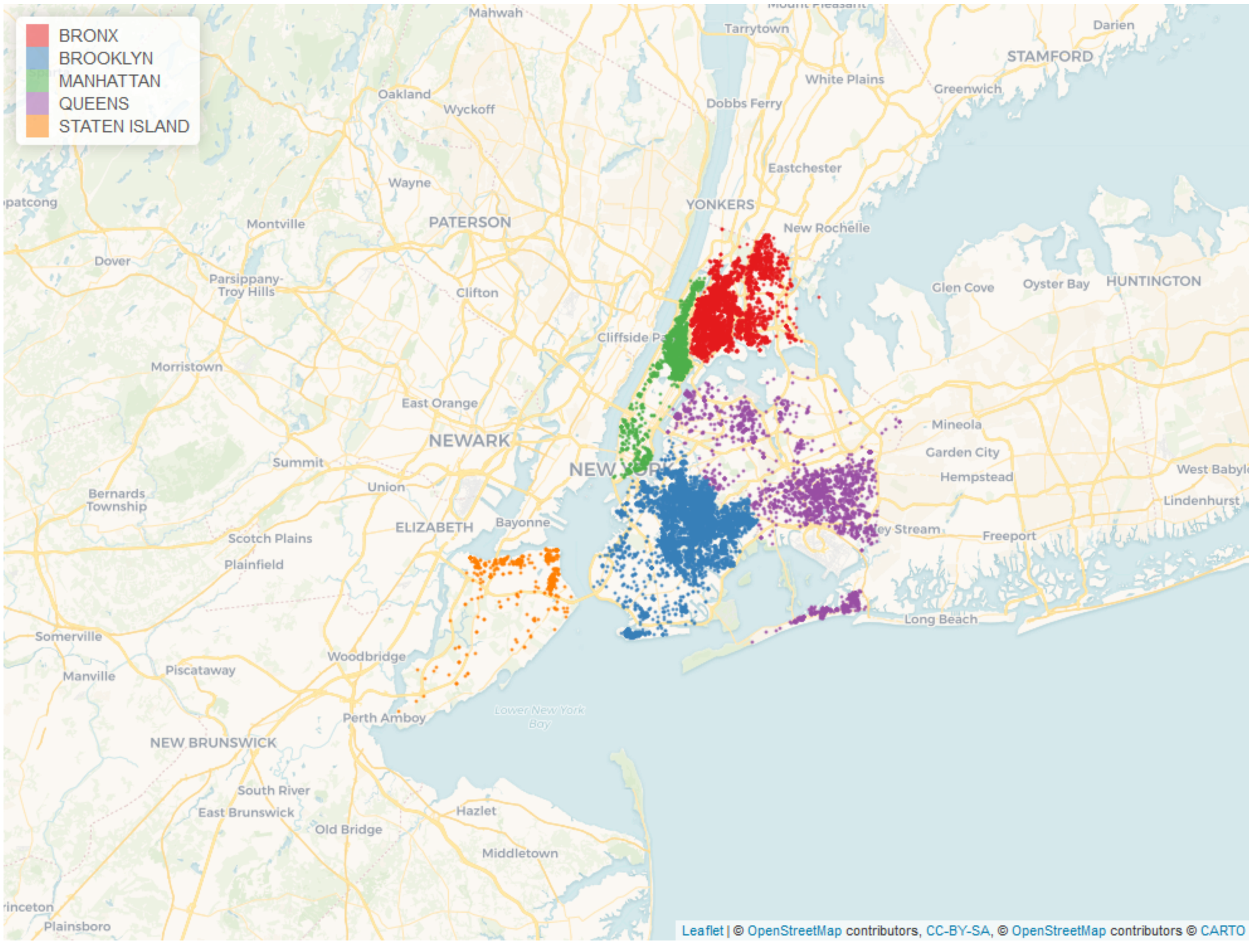


Figure 2: Shooting Incident Coloured by Borough in New York City

By colouring shooting incidents by precinct, we gain a better understanding of the relationship between borough, precinct, and total shooting incidents in New York City.

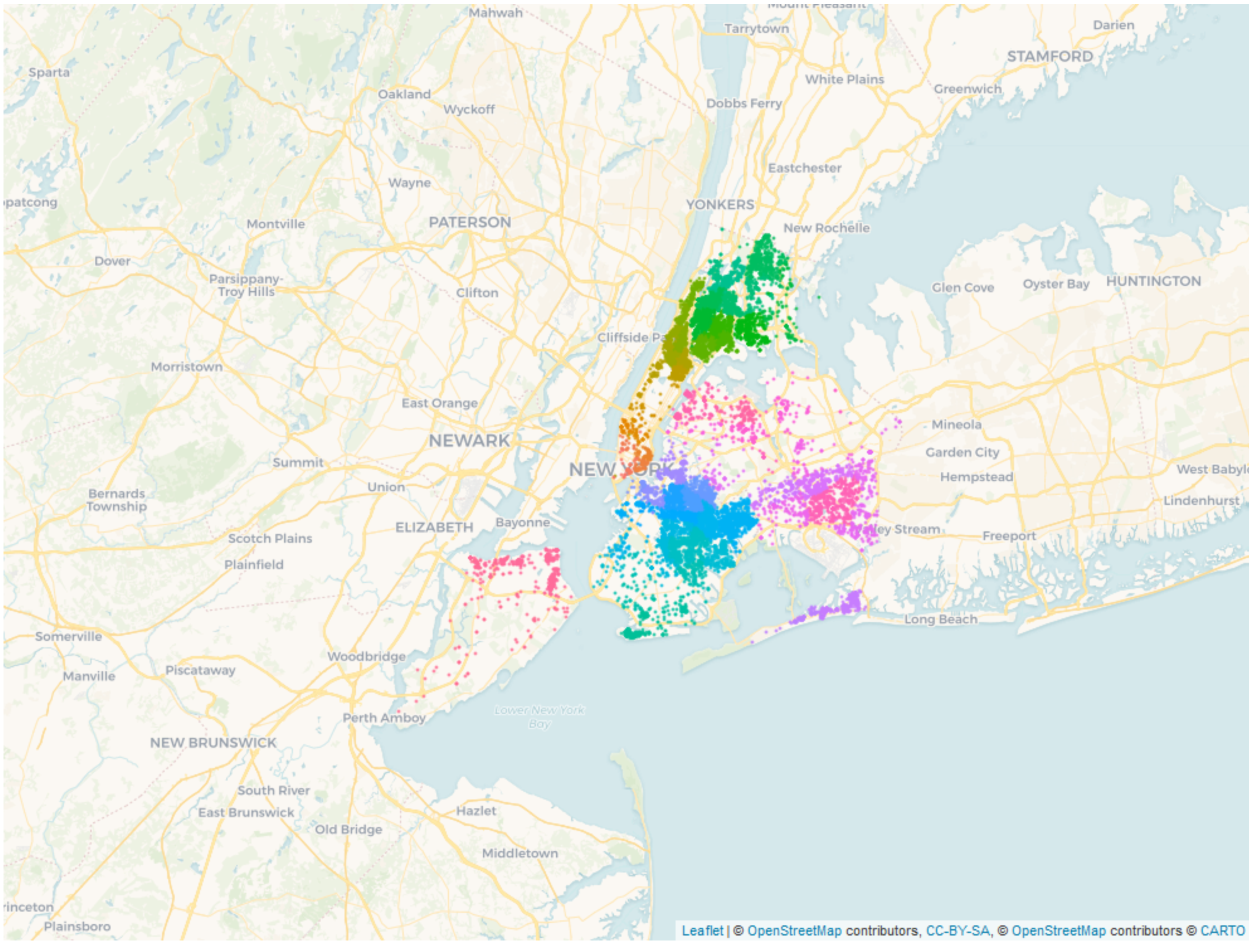


Figure 3: Shooting Incidents Coloured by Precinct in New York City

Are certain age groups more likely to be shot in certain areas? We visually inspect this idea by taking a closer look at a specific borough, Staten Island. Here we can see that there are multiple clusters of shooting incidents and we can note that `<18` tends to be more sparse compared to `18-24` and `25-44` age groups.

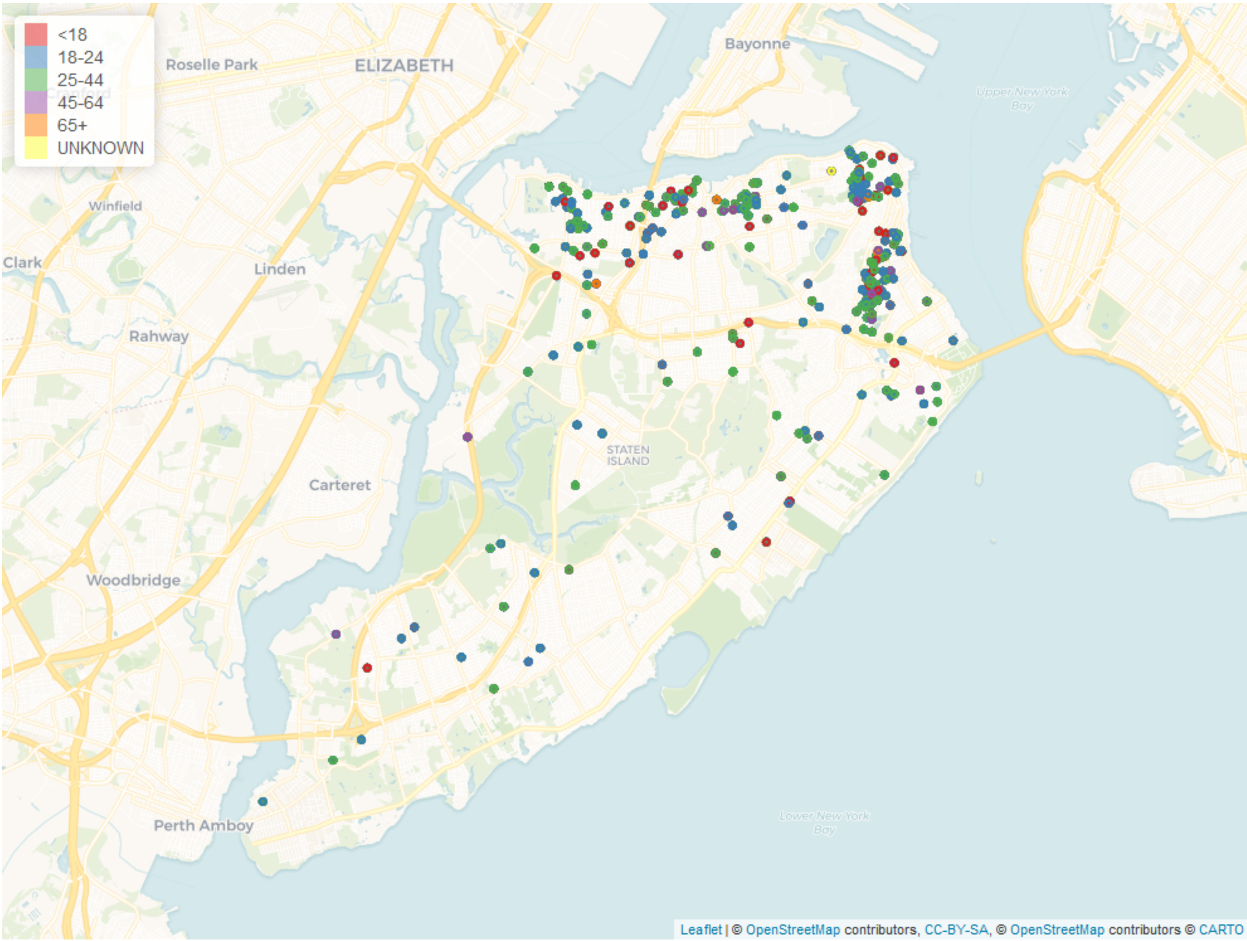


Figure 4: Shooting Incidents in Staten Island grouped by Victim Age

We do the same inspection for victim sex and we find that all shooting incidents are largely involving `males` Shooting incidents involving `females` and `unknown` don't seem to have distinct locations in Staten Island.

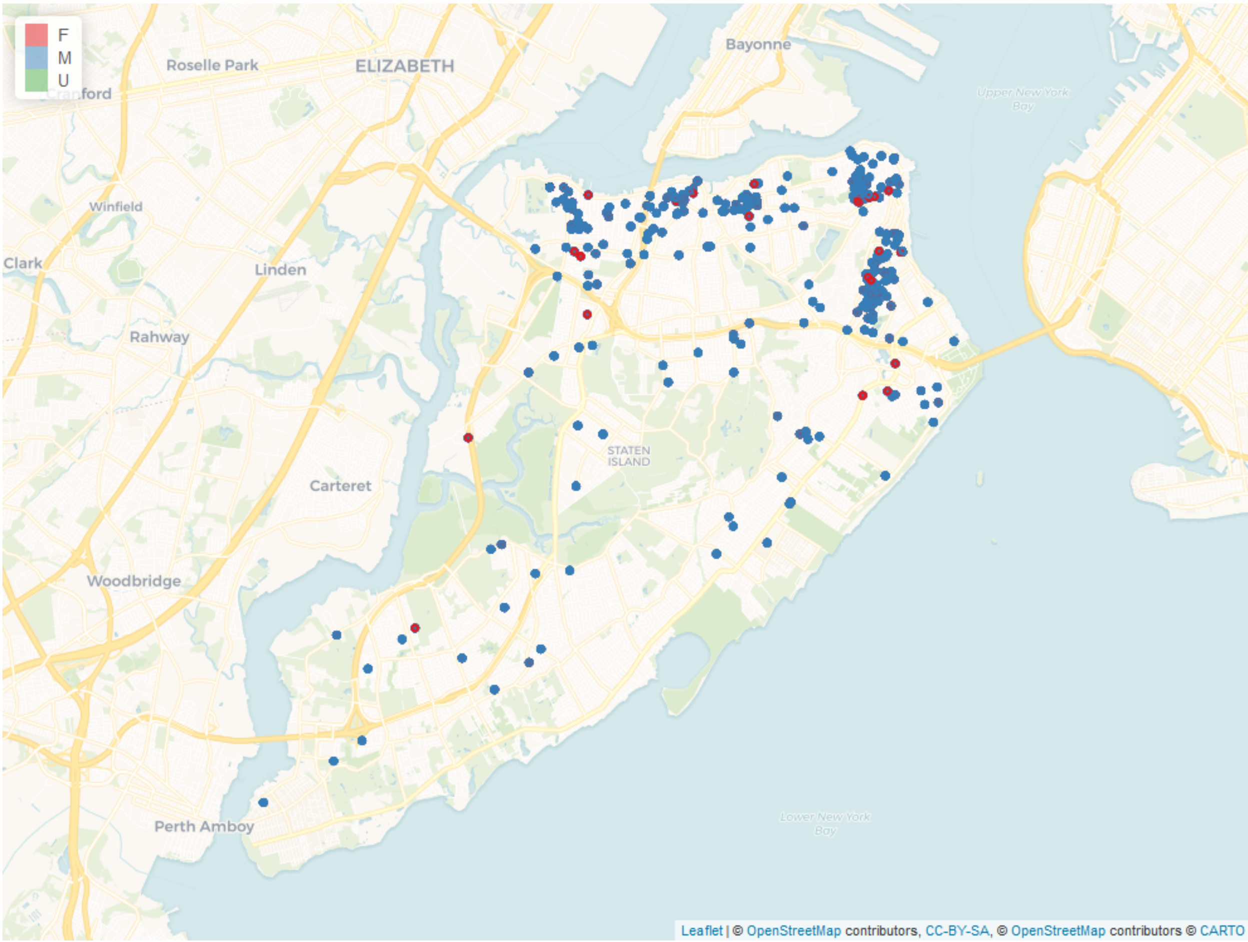


Figure 5: Shooting Incidents in Staten Island grouped by Victim Sex

Lastly we take a look at victim race within Staten Island. We see that most shooting incidents involve a `Black` victim and that clusters tend to include `Black` and `White Hispanic` victims.

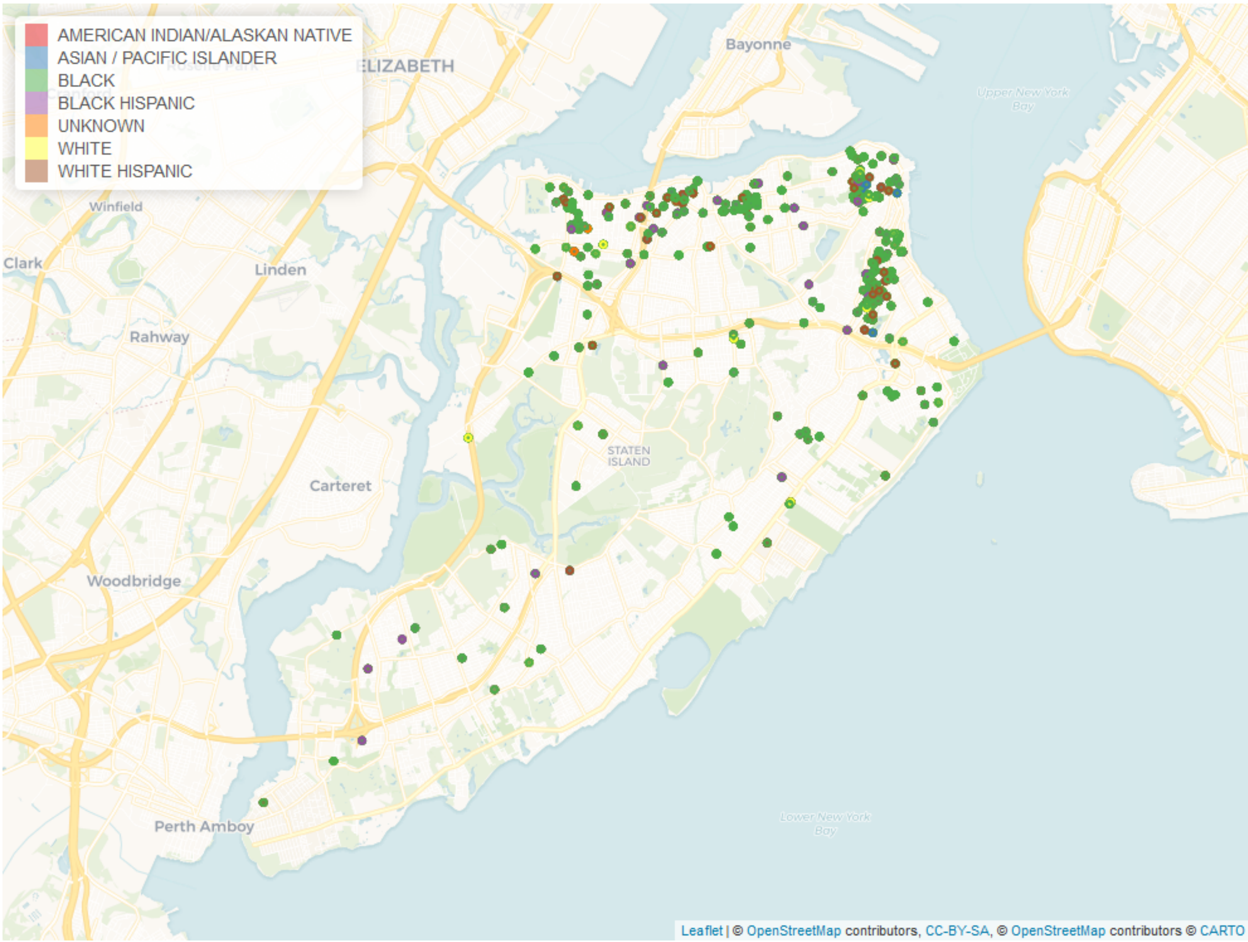


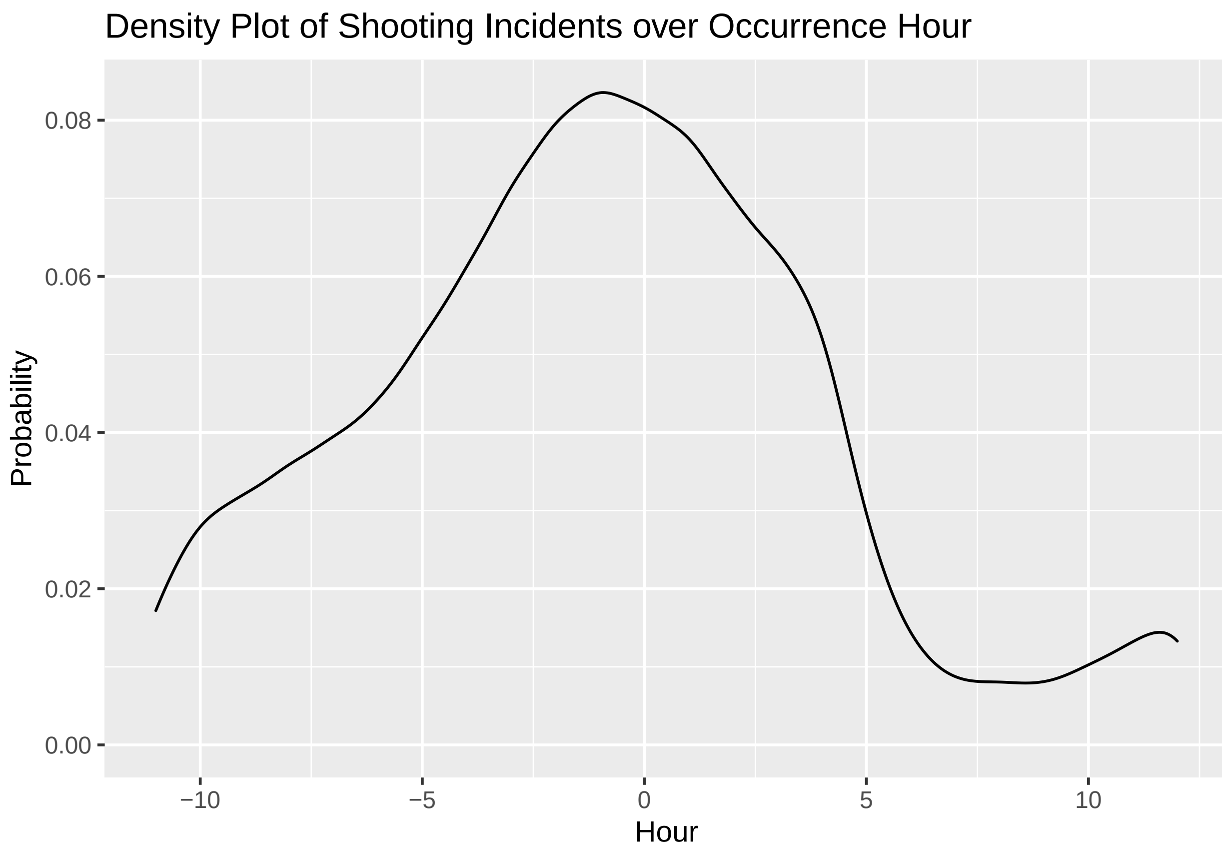
Figure 6: Shooting Incidents in Staten Island grouped by Victim Race

4 Distribution & Probability Analysis

4.1 Density Distributions

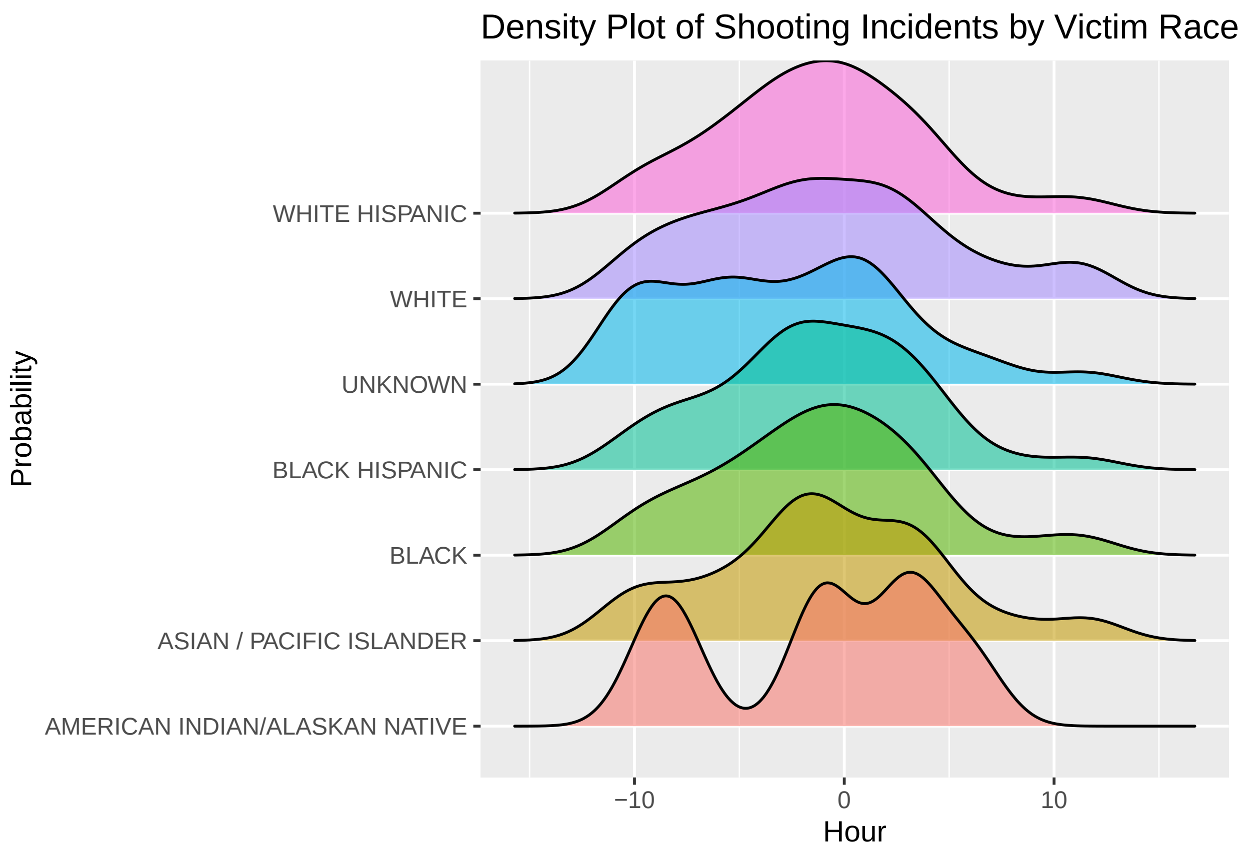
4.1.1 Shooting Incidents over Occurrence Hour

To further visualize the effect of occurrence time, we stratify the occurrence time into occurrence hour centered around midnight and then we plot a density plot to see what times shooting incidents most likely happen. We see that most shootings happen at or before midnight and shootings rarely occur past 5 am.



4.1.2 Shooting Incidents over Occurence Hour split by Victim Race

Here is the same idea from above but split between victim races to see if any one race tends to have a more distinct time for when a shooting incident is to occur. As we can see visually, there tends to be no difference between races, however, we can note that `AMERICAN INDIAN/ALASKAN NATIVE` has a lower likelihood at around evening time.

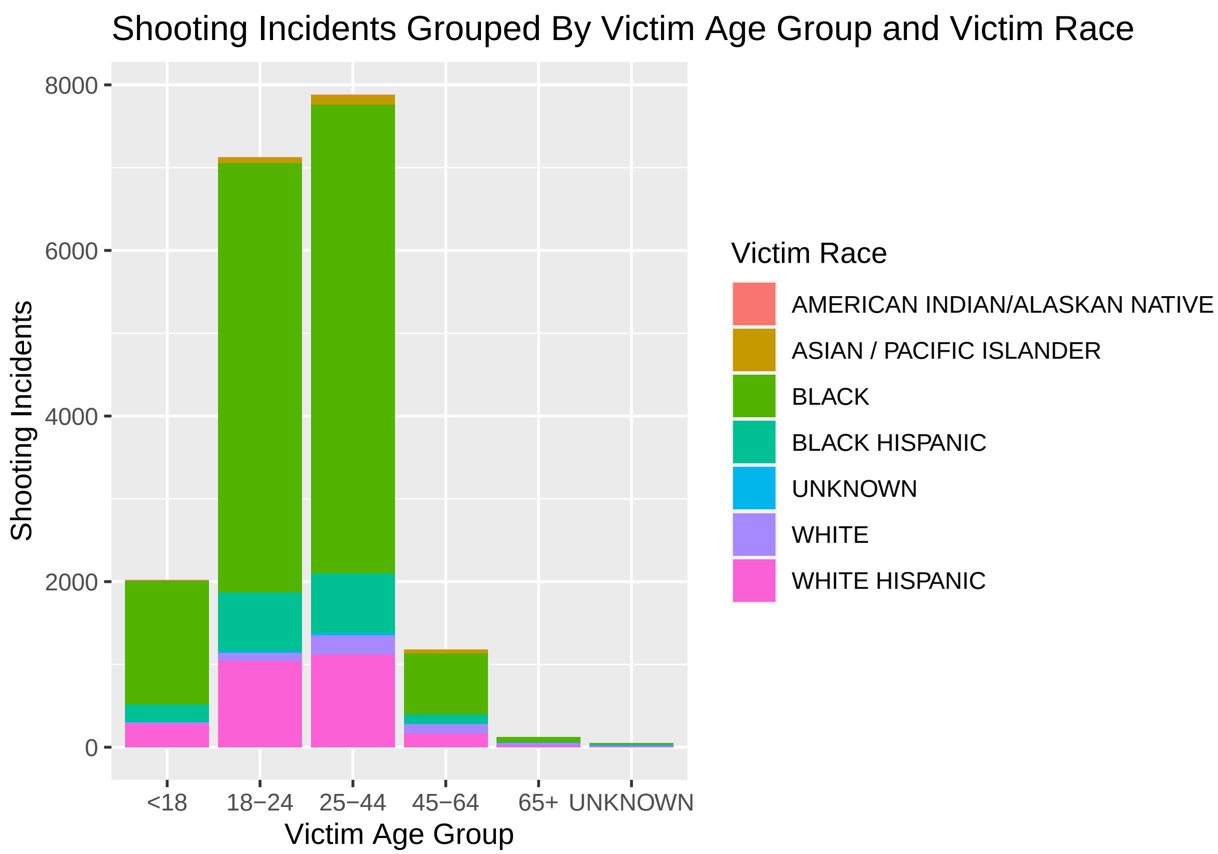


4.2 Probability Distributions

4.2.1 Victim Age Group and Victim Race

Here we take a look at the probabilities of victim races with respect to victim age groups. Given that the victim is of a particular age group, what is the likelihood that they are of a certain race? Here we show the top 10 most probable victim races across all age groups and races and we have a visual plot to show.

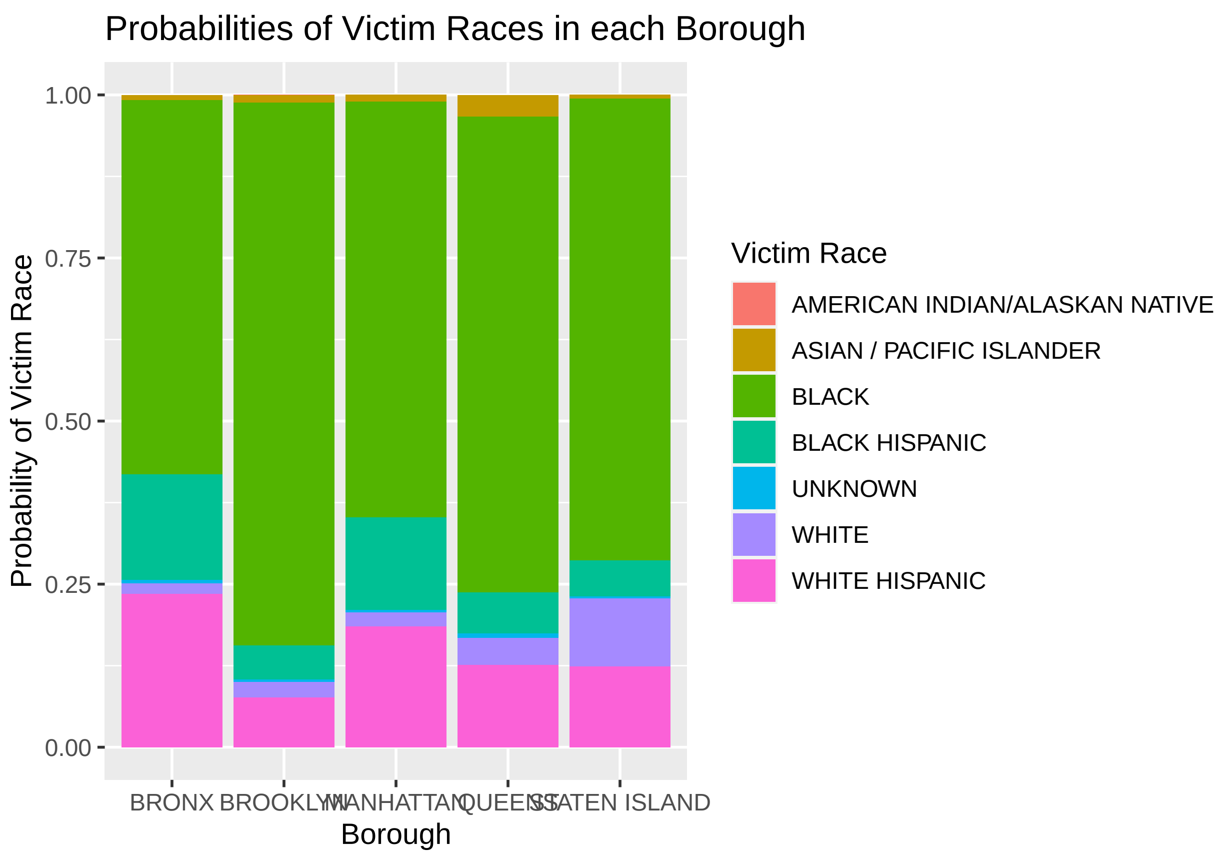
VIC_AGE_GROUP	VIC_RACE	count	prob
25-44	BLACK	5658	0.71811
18-24	BLACK	5183	0.72734
<18	BLACK	1493	0.74021
25-44	WHITE HISPANIC	1122	0.14240
18-24	WHITE HISPANIC	1040	0.14594
45-64	BLACK	737	0.62458
25-44	BLACK HISPANIC	723	0.09176
18-24	BLACK HISPANIC	711	0.09978
<18	WHITE HISPANIC	281	0.13932
25-44	WHITE	226	0.02868
<18	BLACK HISPANIC	212	0.10511
45-64	WHITE HISPANIC	165	0.13983
25-44	ASIAN / PACIFIC ISLANDER	117	0.01485
45-64	WHITE	113	0.09576
45-64	BLACK HISPANIC	112	0.09492
18-24	WHITE	100	0.01403
18-24	ASIAN / PACIFIC ISLANDER	67	0.00940
65+	BLACK	61	0.48413
45-64	ASIAN / PACIFIC ISLANDER	46	0.03898
25-44	UNKNOWN	31	0.00393
65+	WHITE	25	0.19841
65+	WHITE HISPANIC	24	0.19048
18-24	UNKNOWN	21	0.00295
<18	WHITE	18	0.00892
UNKNOWN	UNKNOWN	17	0.34000
UNKNOWN	BLACK	16	0.32000
65+	BLACK HISPANIC	13	0.10317
<18	ASIAN / PACIFIC ISLANDER	9	0.00446
UNKNOWN	WHITE	9	0.18000
45-64	UNKNOWN	7	0.00593
UNKNOWN	WHITE HISPANIC	6	0.12000
18-24	AMERICAN INDIAN/ALASKAN NATIVE	4	0.00056
<18	UNKNOWN	3	0.00149
65+	ASIAN / PACIFIC ISLANDER	3	0.02381
25-44	AMERICAN INDIAN/ALASKAN NATIVE	2	0.00025
<18	AMERICAN INDIAN/ALASKAN NATIVE	1	0.00050
UNKNOWN	ASIAN / PACIFIC ISLANDER	1	0.02000
UNKNOWN	BLACK HISPANIC	1	0.02000


Shooting Incidents Grouped By Victim Age Group and Victim Race
Shooting Incidents
8000
6000
4000
2000
0
<18
18−24
25−44
45−64
65+
UNKNOWN
Victim Age Group
Victim Race
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
BLACK
BLACK HISPANIC
UNKNOWN
WHITE
WHITE HISPANIC

4.2.2 Borough and Victim Race

Furthermore, we want to see which races are most probable depending on the borough. Perhaps a certain borough has a higher likelihood to be a particular race. Across all boroughs, the leading victim race is `black` followed by `white hispanics` and `black hispanics`.

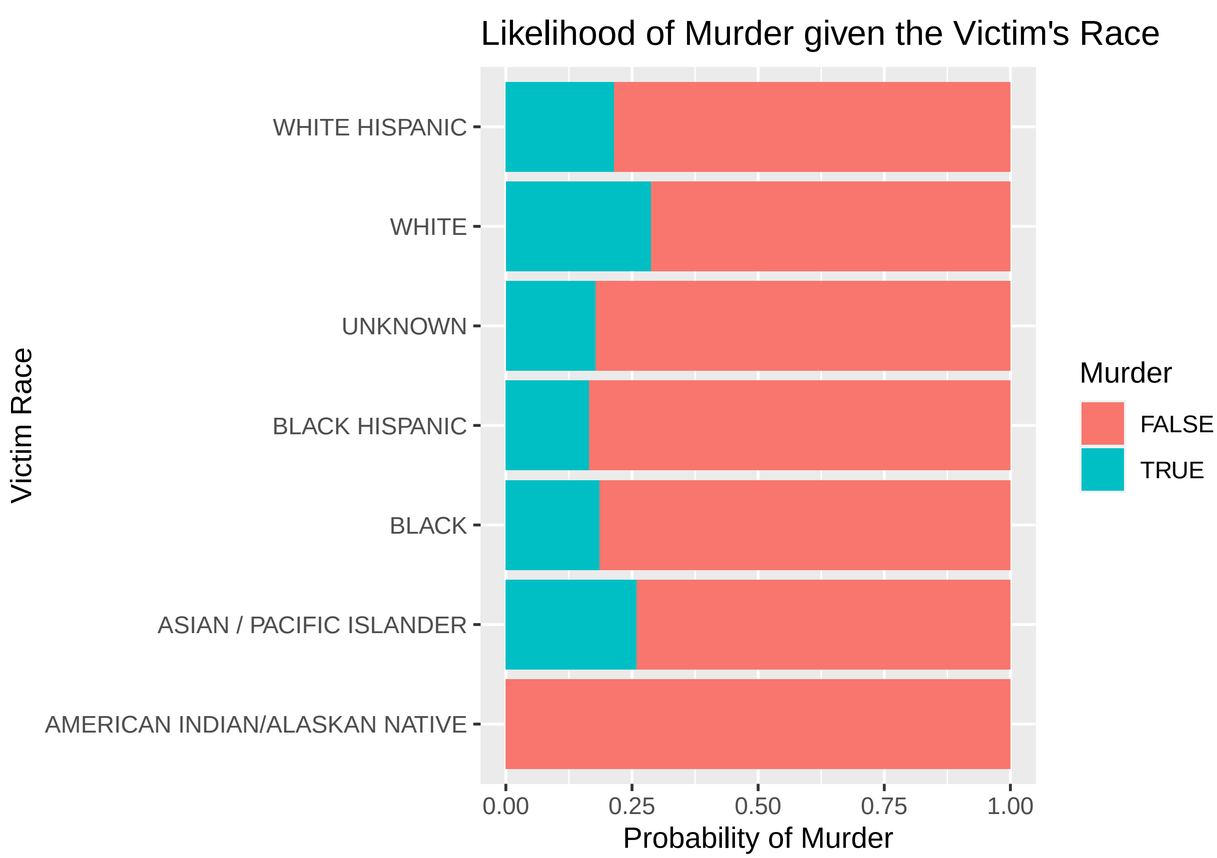
BORO	VIC_RACE	count	prob
BROOKLYN	BLACK	6295	0.83201
QUEENS	BLACK	2029	0.72959
STATEN ISLAND	BLACK	389	0.70856
MANHATTAN	BLACK	1422	0.63681
BRONX	BLACK	3013	0.57401
BRONX	WHITE HISPANIC	1232	0.23471
MANHATTAN	WHITE HISPANIC	413	0.18495
BRONX	BLACK HISPANIC	851	0.16213
MANHATTAN	BLACK HISPANIC	317	0.14196
QUEENS	WHITE HISPANIC	350	0.12585
STATEN ISLAND	WHITE HISPANIC	68	0.12386
STATEN ISLAND	WHITE	57	0.10383
BROOKLYN	WHITE HISPANIC	575	0.07600
QUEENS	BLACK HISPANIC	175	0.06293
STATEN ISLAND	BLACK HISPANIC	30	0.05464
BROOKLYN	BLACK HISPANIC	399	0.05274
QUEENS	WHITE	116	0.04171
QUEENS	ASIAN / PACIFIC ISLANDER	91	0.03272
BROOKLYN	WHITE	183	0.02419
MANHATTAN	WHITE	48	0.02150
BRONX	WHITE	87	0.01657
BROOKLYN	ASIAN / PACIFIC ISLANDER	87	0.01150
MANHATTAN	ASIAN / PACIFIC ISLANDER	24	0.01075
BRONX	ASIAN / PACIFIC ISLANDER	38	0.00724
QUEENS	UNKNOWN	18	0.00647
STATEN ISLAND	ASIAN / PACIFIC ISLANDER	3	0.00546
BRONX	UNKNOWN	25	0.00476
MANHATTAN	UNKNOWN	9	0.00403
STATEN ISLAND	UNKNOWN	2	0.00364
BROOKLYN	UNKNOWN	25	0.00330
QUEENS	AMERICAN INDIAN/ALASKAN NATIVE	2	0.00072
BRONX	AMERICAN INDIAN/ALASKAN NATIVE	3	0.00057
BROOKLYN	AMERICAN INDIAN/ALASKAN NATIVE	2	0.00026


Probabilities of Victim Races in each Borough
Probability of Victim Race
1.00
0.75
0.50
0.25
0.00
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
Borough
Victim Race
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
BLACK
BLACK HISPANIC
UNKNOWN
WHITE
WHITE HISPANIC

4.2.3 Murder and Victim Race

Is one race more likely to be murdered in the even of a shooting? We group the data by victim race to find out. Across all races except `AMERICAN INDIAN/ALASKAN NATIVE`, murder rates tend to be similar. `AMERICAN INDIAN/ALASKAN NATIVE` is the only victim race to have no murders from shootings.

VIC_RACE	STATISTICAL_MURDER_FLAG	count	prob
AMERICAN INDIAN/ALASKAN NATIVE	FALSE	7	1.00000
BLACK HISPANIC	FALSE	1480	0.83521
UNKNOWN	FALSE	65	0.82278
BLACK	FALSE	10709	0.81450
WHITE HISPANIC	FALSE	2073	0.78582
ASIAN / PACIFIC ISLANDER	FALSE	180	0.74074
WHITE	FALSE	350	0.71283
WHITE	TRUE	141	0.28717
ASIAN / PACIFIC ISLANDER	TRUE	63	0.25926
WHITE HISPANIC	TRUE	565	0.21418
BLACK	TRUE	2439	0.18550
UNKNOWN	TRUE	14	0.17722
BLACK HISPANIC	TRUE	292	0.16479



5 Machine Learning Modelling

5.1 Creating Training and Test Sets

```
y <- dat$VIC_RACE
set.seed(718, sample.kind = "Rounding")
test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
train_set <- dat %>% slice(-test_index)
test_set <- dat %>% slice(test_index)
```

5.2 Machine Learning Models

5.2.1 Naive Model

Our baseline model is to simply predict the victim race with the most occurrences in the data set. In this model, we guess `black` for every shooting incident victim.

```
naive_guess <- train_set %>%
  group_by(VIC_RACE) %>%
  summarize(count = n()) %>%
  filter(count == max(count)) %>%
  pull(VIC_RACE)

y_naive <- test_set %>%
  mutate(y_hat = naive_guess) %>%
  pull(y_hat)
```

```
naive_acc <- confusionMatrix(y_naive, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.71487.

Sitting at just above 71% accuracy, this naive model performs poorly to predict victim races. A lot can be improved upon.

5.2.2 Decision Tree Model

Here we use a decision tree to see if it performs better than the naive model. A decision tree was chosen because of insights gained from information about boroughs, precincts, and murder rate.

```
fit_rt <- train(VIC_RACE ~ ., data = train_set, method = "rpart")
```

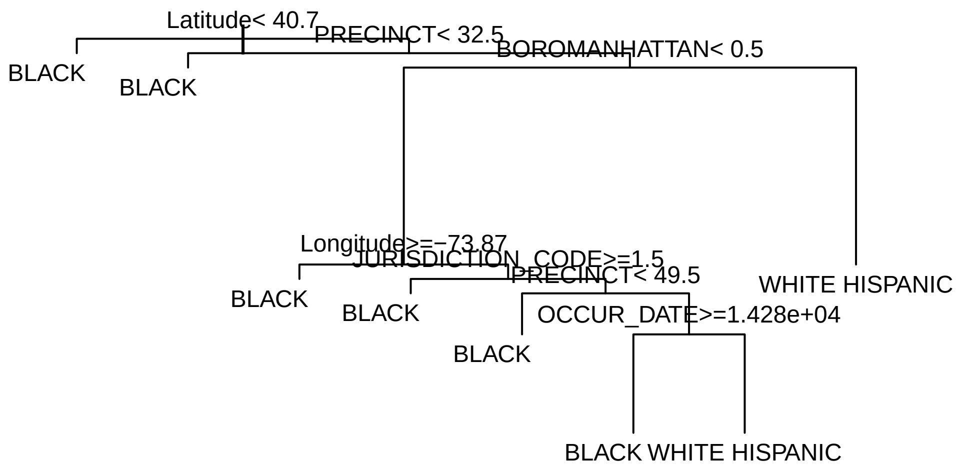
```
y_rt <- predict(fit_rt, newdata = test_set)
```

```
rt_acc <- confusionMatrix(y_rt, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.72329.

At 72% accuracy, the Decision Tree model is not significantly better than the naive model.

Plot of Decision Tree:



5.2.3 Random Forest Model

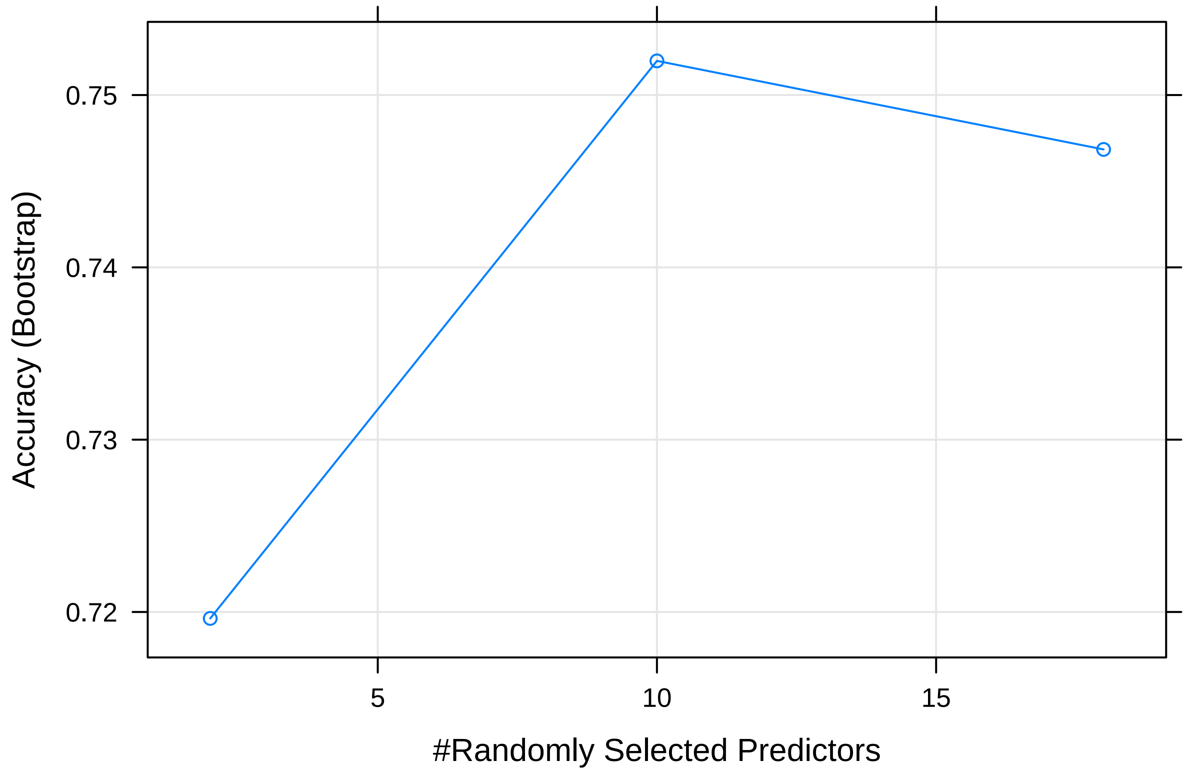
To further improve our decision tree model, I decided to try a random forest in hopes to improve accuracy as not one decision tree can fit all different shooting incidents.

```
fit_rf <- train(VIC_RACE ~ ., data = train_set, method = "rf", allowParallel = TRUE)

y_rf <- predict(fit_rf, newdata = test_set)

rf_acc <- confusionMatrix(y_rf, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.7608.

At 76% accuracy, the Random Forest model is already a much better alternative than the Decision Tree model, but because we're dealing with a dangerous situation (shooting incidents), it would be ideal to achieve a much better accuracy. We can potentially increase the accuracy some more if we tune the number of randomly selected predictors.

Plot of Random Forest model accuracy:



5.2.4 K-Nearest Neighbours Model

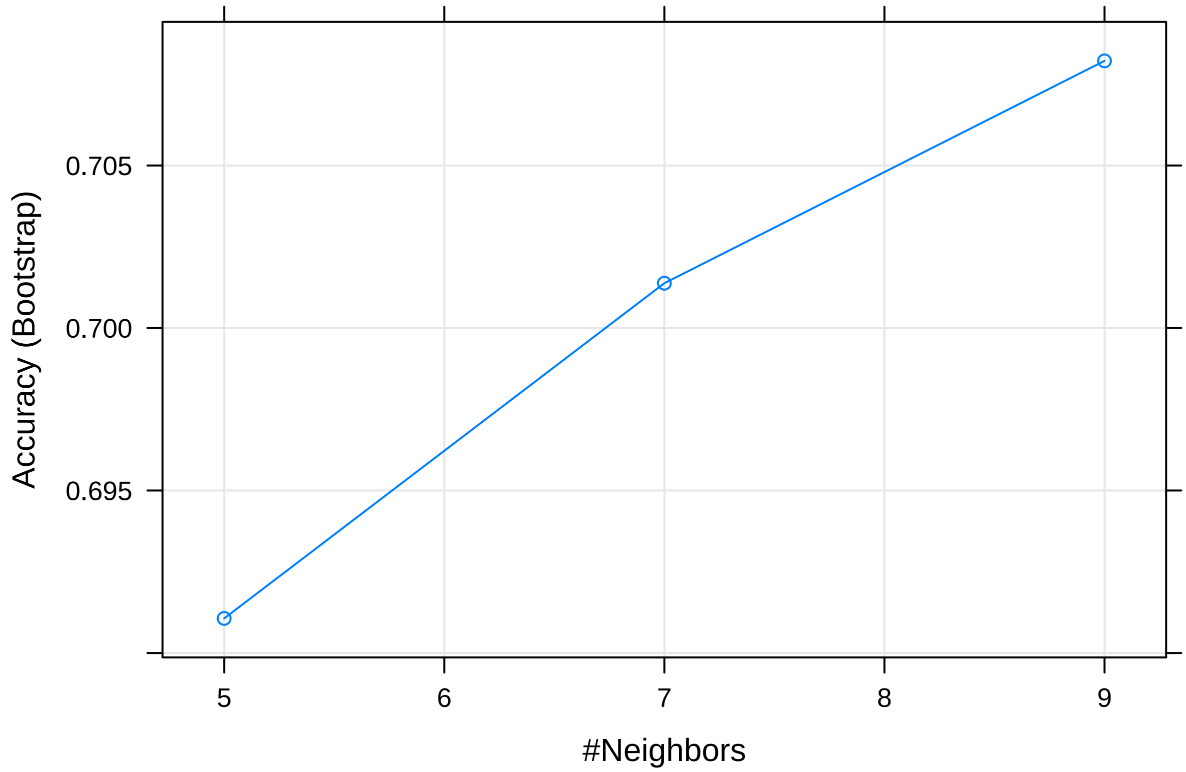
The idea with clusters intrigued me to use a KNN model because we saw that certain victim races were grouped with one another in Staten Island. Originally, I had used all variables to train the model below, however, after trial and error, I found that using only `Latitude` and `Longitude`, the model performed best.

```
fit_knn <- train(VIC_RACE ~ Latitude + Longitude, data = train_set, method = "knn")

y_knn <- predict(fit_knn, newdata = test_set) %>% as.factor()

knn_acc <- confusionMatrix(y_knn, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.729.

The KNN model is surprisingly worse than the Random Forest model at nearly 73% accuracy. However, upon looking at the plot below, we see that there is potentially higher accuracy if we fine tune the model by adjusting the number of neighbours.

Plot of KNN model accuracy:



5.2.5 Naive Bayes Model

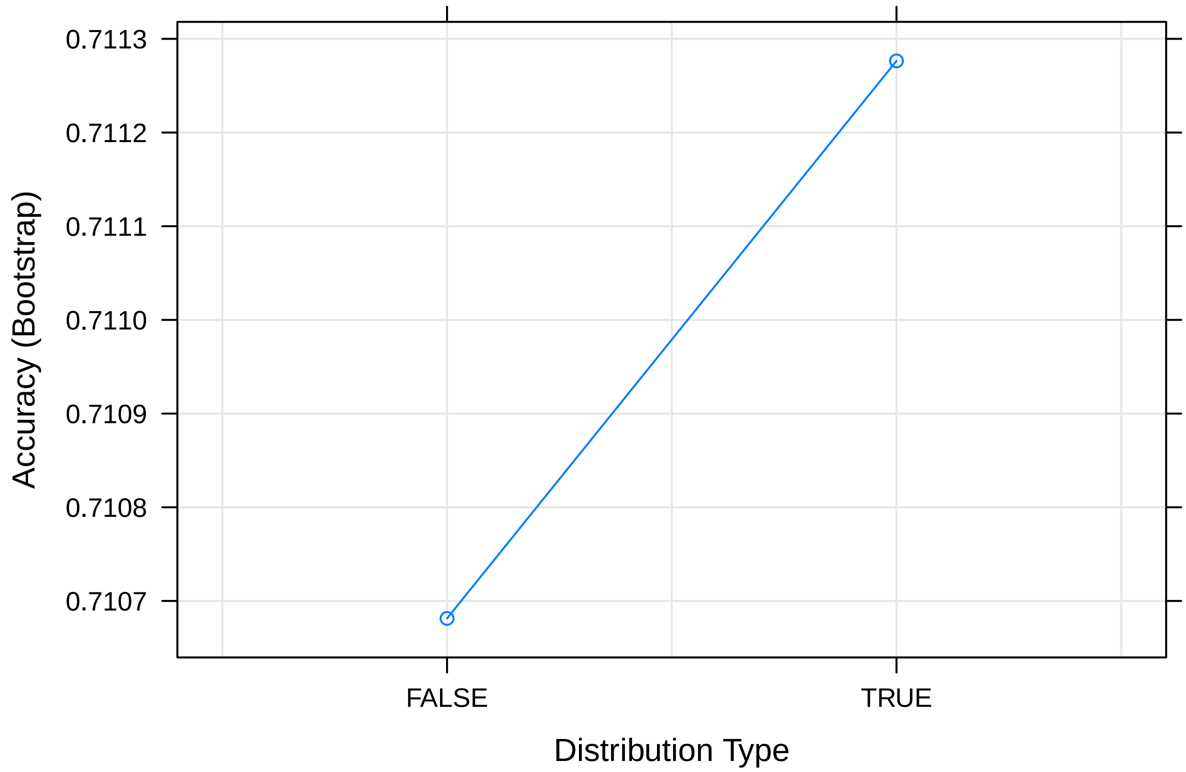
Since we explored the idea of conditional probabilities in the earlier sections, I figured that it would be appropriate to try a Naive Bayes model to see if could predict victim race. Similar to the KNN model, only using `Latitude` and `Longitude` provided the best results after trial and error.

```
fit_nb <- train(VIC_RACE ~ Longitude + Latitude, data = train_set, method = "naive_bayes")

y_nb <- predict(fit_nb, newdata = test_set)

nb_acc <- confusionMatrix(y_nb, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.71596.

The accuracy in the Naive Bayes model is only slightly better than the naive model. Perhaps there is room for improvement if we use cross validation to tune the model, but it appears that it won't get a lot better.

Plot of Naive Bayes accuracy model:



5.2.6 Multinomial Regression Model

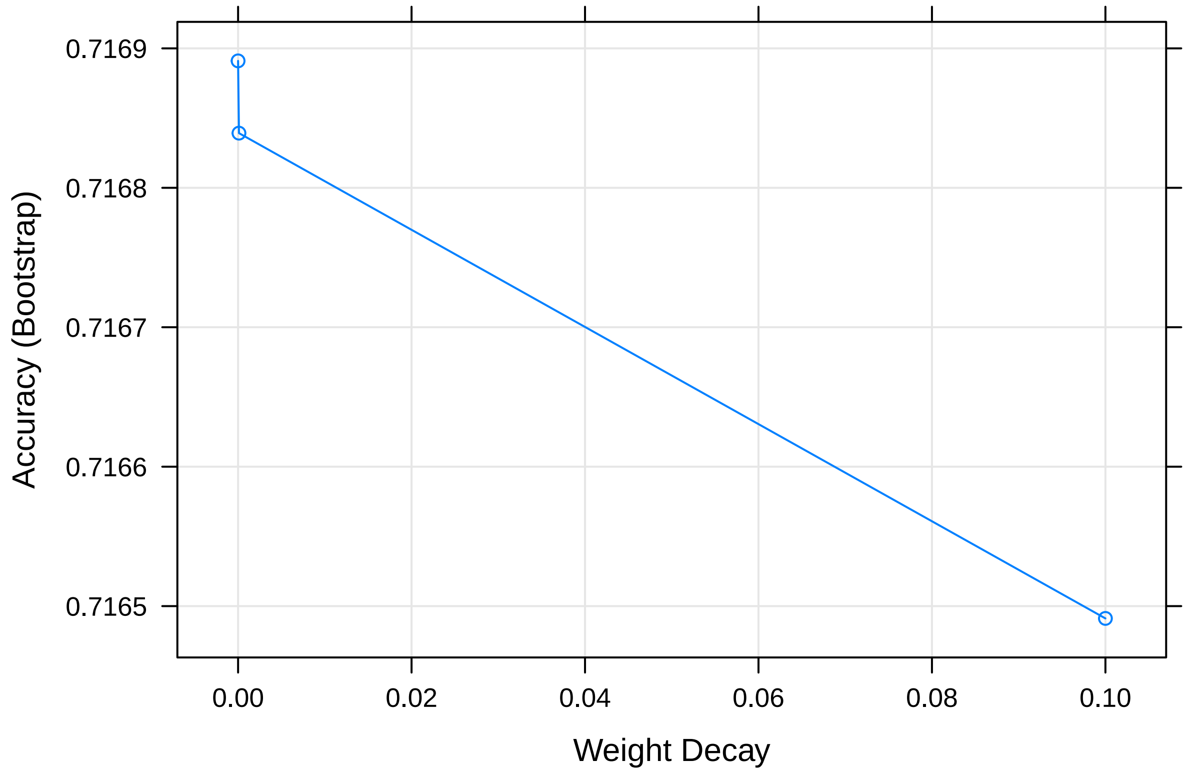
I then thought about a logistic regression model and how that could work on this data set. Upon some research and reading, I remember we learned a bit about multinomial regression and since there are multiple races, I believe it can be useful to try.

```
fit_mln <- train(VIC_RACE ~ ., data = train_set, method = "multinom", MaxNWts = 1000000)

y_mln <- predict(fit_mln, newdata = test_set) %>% as.factor()

mln_acc <- confusionMatrix(y_mln, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.71514.

Similar to the Naive Bayes model. The Multinomial Regression model is performing poorly but we will include them in the next section to see if there is any improvement at all.

Plot of Multinomial Regression accuracy model:



5.3 Model Table of Results

Here is the table of results for the preliminary models. Random Forest performs the best out of all models.

Models	Accuracy
Naive Model	0.71487
Decision Tree	0.72329
Random Forest	0.76080
K-Nearest Neighbours	0.72900
Naive Bayes	0.71596
Multinomial Regression	0.71514

6 Advanced Modelling

6.1 Cross Validation

```
control <- trainControl(method = "cv", number = 10, p = .9)
```

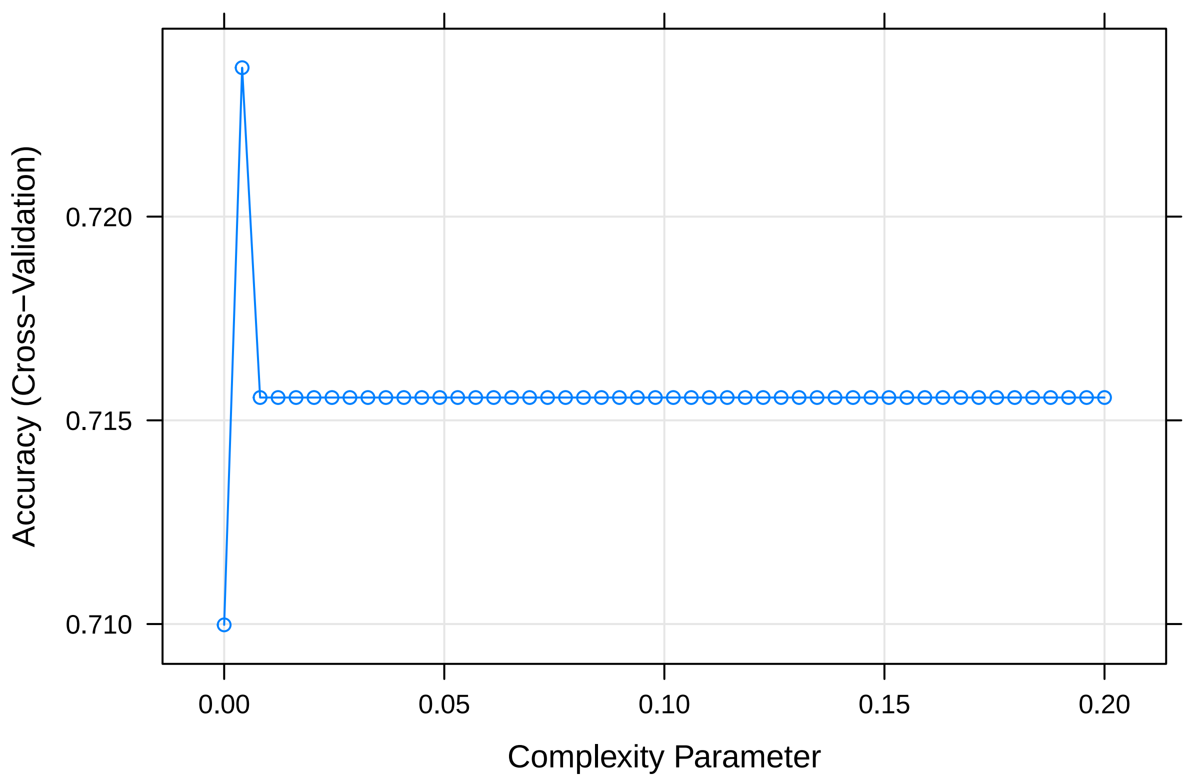
6.1.1 K-Fold Cross Validation Decision Tree Model

```
fit_rt <- train(VIC_RACE ~ .,
                data = train_set,
                method = "rpart",
                tuneGrid = data.frame(cp = seq(0.0, 0.2, len = 50)),
                trControl = control)

y_rt <- predict(fit_rt, newdata = test_set)

rt_acc <- confusionMatrix(y_rt, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.71976.

It appears that the decision tree did not perform better after cross validating. Perhaps it performed slightly better prior to cross validation because of the randomness of training and test splits.

Plot of Decision Tree Tuning Results:



6.1.2 K-Fold Cross Validation Random Forest Model

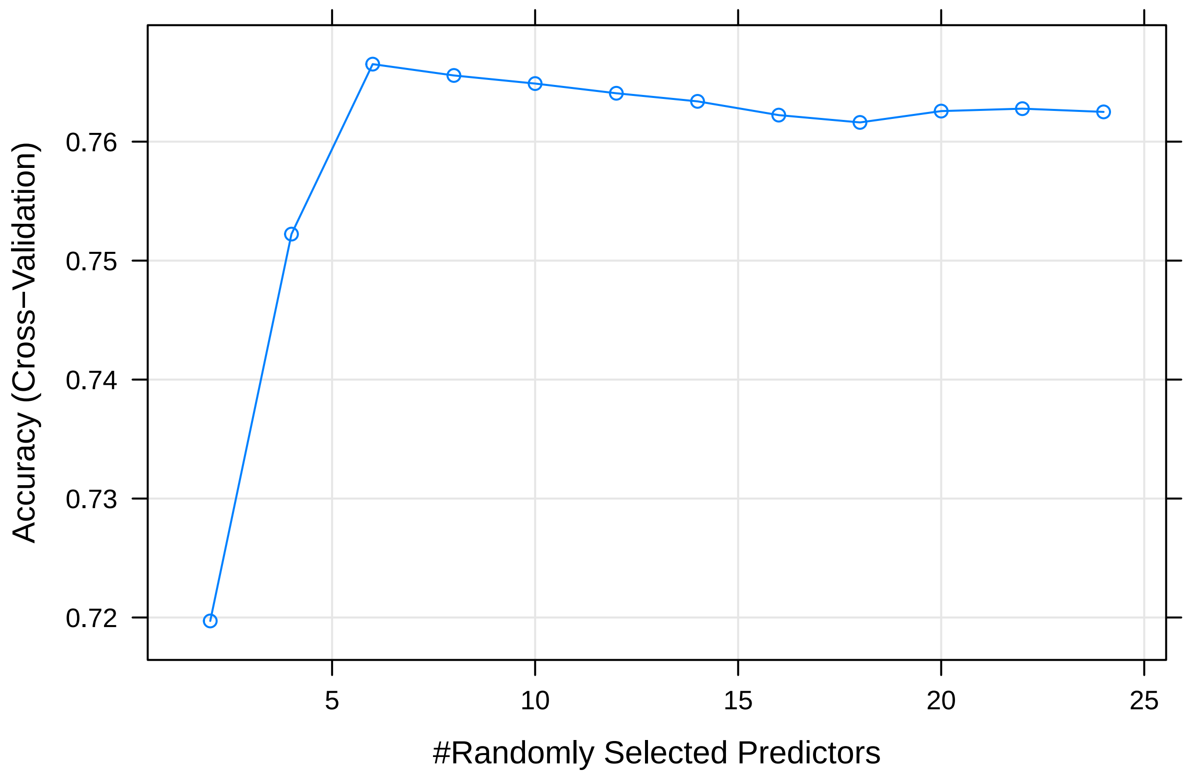
```
fit_rf <- train(VIC_RACE ~ .,
                data = train_set,
```

```
                method = "rf",
                tuneGrid = data.frame(mtry = seq(2,24,2)),
                trControl = control,
                allowParallel = TRUE)

y_rf <- predict(fit_rf, newdata = test_set)

rf_acc <- confusionMatrix(y_rf, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.76597.

The Random Forest model slightly improved after cross validation which is a small success. However, these are not big improvements.

Plot of Random Forest Tuning Results:



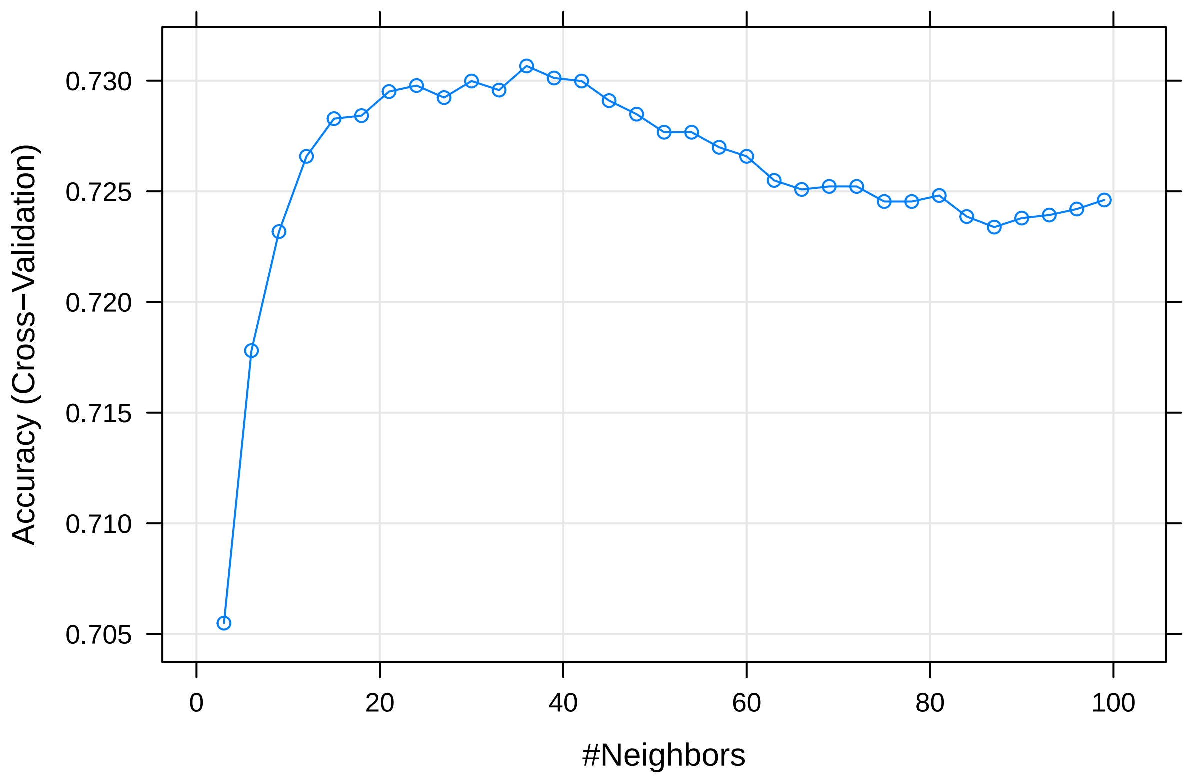
6.1.3 K-Fold Cross Validation K-Nearest Neighbour Model

```
fit_knn <- train(VIC_RACE ~ Latitude + Longitude,
                 data = train_set,
                 method = "knn",
                 tuneGrid = data.frame(k = seq(3,101,3)),
                 trControl = control)

y_knn <- predict(fit_knn, newdata = test_set)

knn_acc <- confusionMatrix(y_knn, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.73036.

Similar to the K-Fold Cross Validation Random Forest model, the KNN model after cross validation performed slightly better than before.

Plot of KNN Tuning Results:



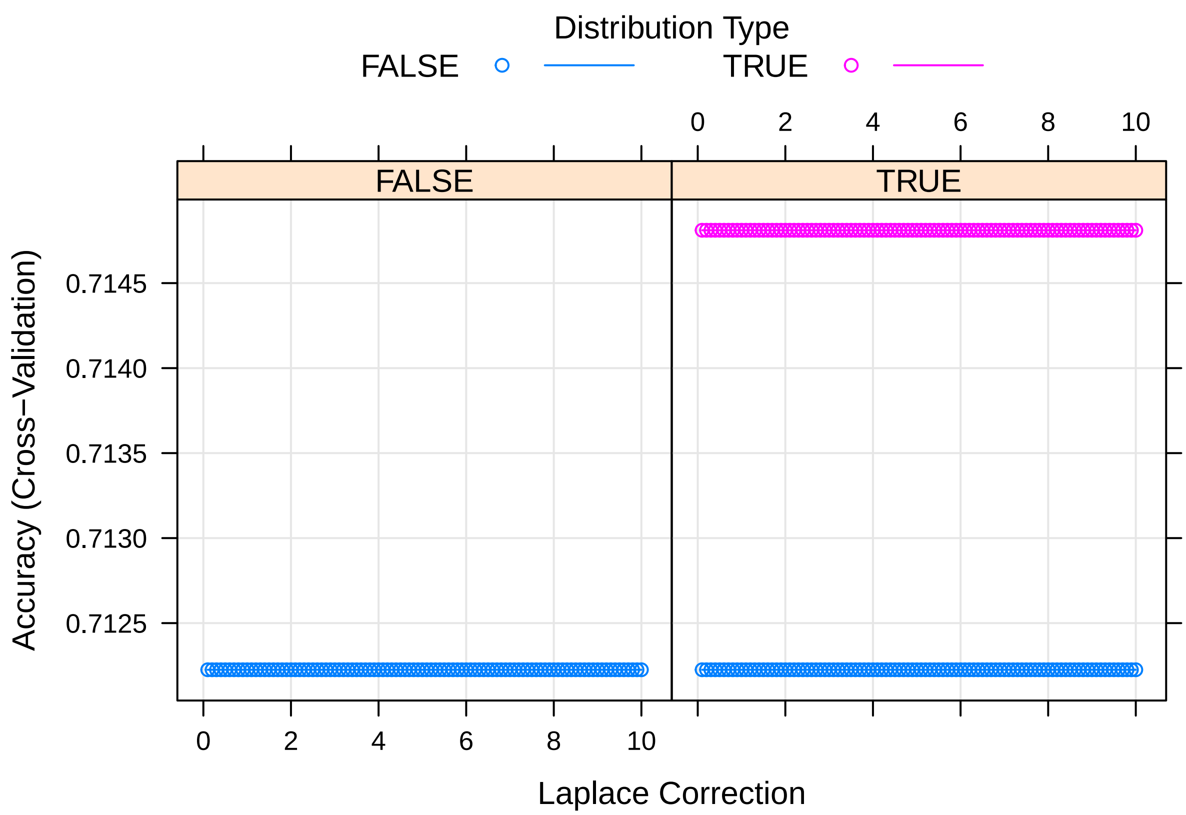
6.1.4 K-Fold Cross Validation Naive Bayes Model

```
fit_nb <- train(VIC_RACE ~ Latitude + Longitude,
                data = train_set,
                method = "naive_bayes",
                tuneGrid = expand.grid(laplace = seq(0.1, 10, 0.1),
                                       usekernel = c(TRUE, FALSE),
                                       adjust = c(TRUE, FALSE)),
                trControl = control)

y_nb <- predict(fit_nb, newdata = test_set)

nb_acc <- confusionMatrix(y_nb, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.71596.

The accuracy of the Naive Bayes model did not improve with cross validation. We can see that the tuning parameters had no effect on the accuracy looking at the plot below.

Plot of Naive Bayes Tuning Results:



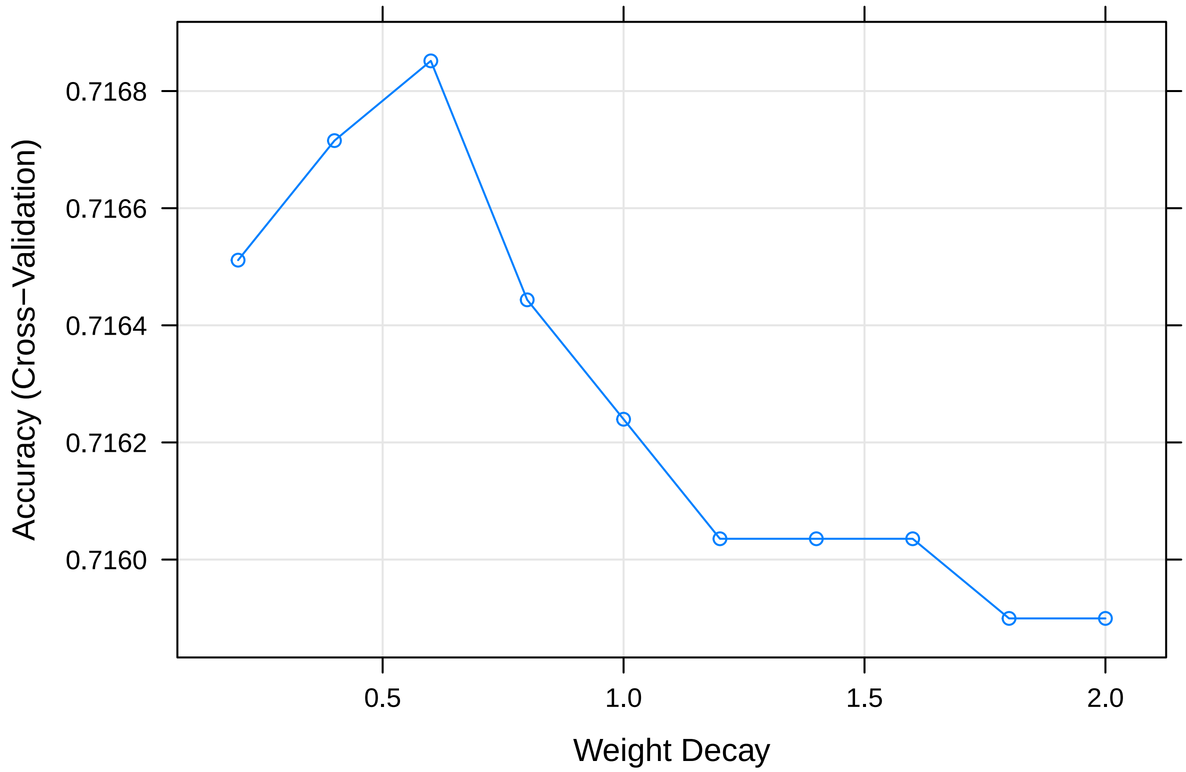
6.1.5 K-Fold Cross Validation Multinomial Regression Model

```
fit_mln <- train(VIC_RACE ~ .,
                 data = train_set,
                 method = "multinom",
                 tuneGrid = data.frame(decay = seq(0.2, 2, 0.2)),
                 trControl = control,
                 MaxNWts = 1000000)

y_mln <- predict(fit_mln, newdata = test_set) %>% as.factor()

mln_acc <- confusionMatrix(y_mln, reference = test_set$VIC_RACE)$overall["Accuracy"]
```

Accuracy: 0.71541.

Like the Naive Bayes model, the Multinomial Regression model did not improve with cross validation. These are disappointing results as I hoped that they could have more predicting power. However, results are results, I cannot force them to be predictable.

Plot of Multinomial Regression Tuning Results:



6.2 Cross Validation Models Table of Results

Here is the table of results for the cross validated models. Random Forest and K-Nearest Neighbours improved with this technique but not by a lot.

Models	Accuracy
K-Fold Cross Validated Decision Tree	0.71976
K-Fold Cross Validated Random Forest	0.76597
K-Fold Cross Validated K-Nearest Neighbours	0.73036
K-Fold Cross Validated Naive Bayes	0.71596
K-Fold Cross Validated Multinomial Regression	0.71541

6.3 Ensemble Model

Perhaps an ensemble model would perform better than all the models individually. An ensemble prediction was formed by looking at the most predicted race in each row and using that as the prediction.

```
races <- levels(dat$VIC_RACE)
y_ensemble <- data.frame(rt = y_rt,
                         rf = y_rf,
                         knn = y_knn,
                         nb = y_nb,
                         mln = y_mln)

ensemble <- apply(y_ensemble, 1, function(pred) {
  prob_race <- sapply(races, function(race) {
    mean(pred == race)
  })
  races[which.max(prob_race)]
})

ensemble <- factor(ensemble, levels = levels(test_set$VIC_RACE))
ens_acc <- confusionMatrix(ensemble, reference = factor(test_set$VIC_RACE))$overall["Accuracy"]
```

Accuracy: 0.72085.

At 72% accuracy, this ensemble model is worse than the Random Forest and KNN model. Therefore, it is not a good choice for our final model.

6.4 Aggregate Table of Results

Below is a table showing the results of applying all the machine learning algorithms to the data set so far. As it turns out, the Random Forest model performs the best and improved via K-Fold Cross Validation, therefore we choose this algorithm for our final model.

Models	Accuracy
Naive Model	0.71487
Decision Tree	0.72329
Random Forest	0.76080
K-Nearest Neighbours	0.72900
Naive Bayes	0.71596
Multinomial Regression	0.71514
K-Fold Cross Validated Decision Tree	0.71976
K-Fold Cross Validated Random Forest	0.76597
K-Fold Cross Validated K-Nearest Neighbours	0.73036
K-Fold Cross Validated Naive Bayes	0.71596
K-Fold Cross Validated Multinomial Regression	0.71541
Ensemble Model	0.72085

7 Final Model

7.1 Random Forest Model

Using the highest accuracy model from all of the training we've done, we are ready to train the final model using the entire data set. As it turned out, the best model was the Random Forest model using a tuning parameter of mtry = 6

```
fit_final <- randomForest(VIC_RACE ~ .,
                          data = dat,
                          allowParallel = TRUE,
                          mtry = fit_rf$bestTune[,"mtry"])

y_final <- predict(fit_final, newdata = validation)

final_acc <- confusionMatrix(y_final, reference = factor(validation$VIC_RACE))$overall["Accuracy"]
```

Accuracy: 0.77819.

Our final accuracy was nearly 78% on the validation set. While this is a decent result overall, compared to the naive method of simply guessing `black` for every victim, this isn't that much of an improvement. Nevertheless, this shows that trial and error and using proper techniques can improve machine learning algorithms even if its just slightly.

7.2 Table of Results

Below is the final table of results, we see that we tested many models and achieved relatively the similar accuracies throughout. However, with each step of the process we slightly improved upon on the accuracy and achieved a final accuracy of nearly **78%**.

Models	Accuracy
Naive Model	0.71487
Decision Tree	0.72329
Random Forest	0.76080
K-Nearest Neighbours	0.72900
Naive Bayes	0.71596
Multinomial Regression	0.71514
K-Fold Cross Validated Decision Tree	0.71976
K-Fold Cross Validated Random Forest	0.76597
K-Fold Cross Validated K-Nearest Neighbours	0.73036
K-Fold Cross Validated Naive Bayes	0.71596
K-Fold Cross Validated Multinomial Regression	0.71541
Ensemble Model	0.72085
Final Model	0.77819

8 Conclusion

This Capstone Project for Harvard Data Science Professional Certificate Program has taught me a lot by giving me the freedom to choose a data set that interested me. It didn't necessarily go as planned because I initially hoped for a high accuracy prediction for a machine learning project. Upon delving into the project; exploring the data and assessing my own hypotheses, I realized not all data sets are very predictable and that is the nature of data science. By going through the extensive work to unpack and understand the data through multiple lenses, we get a better idea of the world around us.

I first obtained, read, and cleaned the data. Then I explored the data, looking various values that could be present and then looked at the number of shooting incidents after grouping variables together. After that I decided to use data visualization to gain a better idea of the proportion of shooting incidents in relation to other variables as well as inspecting the data geographically. Then I looked through probabilities and distributions for various columns in relation to victim race. Finally, I tested multiple machine learning algorithms and used cross validation to see what performed best. The final model was a Random Forest model and had an accuracy of **78%**. The implications of these results could be interpreted that we need more powerful machine learning techniques, require the use of the variables I decided not to include, or that we're missing information that could lead to predictability. One thing is certain, however, it is that shooting incidents involve `black` victims a disproportionate amount because even without any additional insights gained or strategies used, making a naive guess of only `black` victims amounts to a **71%** accuracy. This isn't to imply that other races should be involved in shooting incidents more but rather this alludes to victim races being inexplicable from the known variables.

The potential impact of such an algorithm would be to help law enforcement and/or paramedics. For example, let's say a shooting incident occurred in New York City but there is no information of victim race yet, however, the location, borough, precinct, jurisdiction code, etc. is all known. If we are able to predict the victim's race accurately, it may be useful for law enforcement to send more specialized help based on the victim's race to de-escalate a situation. We know that in 2020 & 2021, the relationship between civilians and law enforcement has grown more tense with ideas of racial profiling. Being able to send out an officer that is more reassuring to a particular race could end up saving lives. Of course this is only idealistic. Realistically, this project has a number of limitations. For example, we might not always know the exact location, borough, precinct, etc. and that it is far more important that any help (not specific help) is sent toward the shooting incident and the victim. There isn't always time to have all the information. Secondly, the nature of shooting incidents is not and will not be fully predictable. Anybody can be a victim of shooting, whether it be a stray bullet, or a targeted shooting, nothing is or certain and therefore predicting may not be useful.

There is potential for this project to be completed further but I am satisfied with where it sits. It may not have a lot of predictability with the current algorithms, however a neural network may have a better time or with more data. There are also a number of columns that I left out such as `PERP_AGE_GROUP`, `PERP_SEX`, and `PERP_RACE`, these columns may have led to higher prediction power but with how many values were missing from these columns, I left them out. New and more innovative approaches could be useful in this project as well, such as transforming the data and adding penalty terms. If you are interested in learning and building out a more powerful model, feel free to do so.