Final

Jared Brotamonte

12/11/2023

Introduction

For this project, I decided to look into the topic of law enforcement and whether or not law enforecment truly targets minorities and people of color.

The difference in power when comparing law enforcement to the average citizen has always been hugely drastic. This imbalance of power has recently led to many conflicts arising between law enforcement and US citizens with the bases of these conflicts being that law enforcement has too much ability to abuse their power without enough concequences. In particular, I wanted to focus on claim that law enforcement abuses their power towards minorities or more specifically people of color.

The data set I chose for this project is the "NYPD_Arrest_Data_2023" data set from kaggle.com. This data set provides information on the many arrests the new york police department made in the year 2023. This data set contains 18 differerent variables about each arrest made with a there being a mix of variables with numerical values and variables with character values.

```
data <- read.csv("NYPD_Arrest_Data_2023.csv")
summary(data)
```

```
##    ARREST_KEY        ARREST_DATE            PD_CD         PD_DESC
##  Min.   :261180920   Length:170095      Min.   :  2.0   Length:170095
##  1st Qu.:264955502   Class :character   1st Qu.:114.0   Class :character
##  Median :268503631   Mode  :character   Median :397.0   Mode  :character
##  Mean   :268370489                      Mean   :425.2
##  3rd Qu.:271826052                      3rd Qu.:705.0
##  Max.   :275235005                      Max.   :997.0
##
##      KY_CD        OFNS_DESC           LAW_CODE          LAW_CAT_CD
##  Min.   :101.0   Length:170095      Length:170095      Length:170095
##  1st Qu.:113.0   Class :character   Class :character   Class :character
##  Median :236.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :249.8
##  3rd Qu.:344.0
##  Max.   :995.0
##  NA's   :13
##  ARREST_BORO        ARREST_PRECINCT  JURISDICTION_CODE  AGE_GROUP
##  Length:170095      Min.   :  1.00   Min.   : 0.0000   Length:170095
##  Class :character   1st Qu.: 40.00   1st Qu.: 0.0000   Class :character
##  Mode  :character   Median : 62.00   Median : 0.0000   Mode  :character
##                     Mean   : 63.42   Mean   : 0.9459
##                     3rd Qu.:101.00   3rd Qu.: 0.0000
```

```
##                     Max.   :123.00   Max.   :97.0000
##
##    PERP_SEX          PERP_RACE           X_COORD_CD        Y_COORD_CD
##  Length:170095      Length:170095      Min.   :      0   Min.   :     0
##  Class :character   Class :character   1st Qu.: 991360   1st Qu.:186065
##  Mode  :character   Mode  :character   Median :1005511   Median :206851
##                                        Mean   :1005863   Mean   :208326
##                                        3rd Qu.:1017933   3rd Qu.:236175
##                                        Max.   :1067185   Max.   :271819
##
##     Latitude       Longitude
##  Min.   : 0.00   Min.   :-74.25
##  1st Qu.:40.68   1st Qu.:-73.97
##  Median :40.73   Median :-73.92
##  Mean   :40.74   Mean   :-73.92
##  3rd Qu.:40.81   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :  0.00
##
```

Data evaluation

The data set contains 18 different variables that help describe each arrest. The arrest has a couple variables that are used to help keep track of the arrest like ARREST_KEY, PD_CD, and KY_CD. All of these variables contain the arrests ID's and codes that help keep track of the arrest. These variables will not be necassary for analyzing the data given the variables just contain the ID's and codes the government uses to organize the arrests, they won't be of much use, and even if codes and ID's repeat, it'd be hard to understand the meaning of these numbers.

```
selected_columns <- c('ARREST_KEY', 'PD_CD', 'KY_CD')
selected_data <- data[selected_columns]
head(selected_data)
```

```
##   ARREST_KEY PD_CD KY_CD
## 1  261209118   109   106
## 2  262984267   515   117
## 3  263664549   105   106
## 4  261345231   105   106
## 5  263536618   109   106
## 6  262030390   105   106
```

The data also contains information on the arrest itself. So this includes the ARREST_DATE, The PD_DESC which is the description based on the PD_CD, and the OFNS_DESC which is the description based on the KY_CD. The LAW_CODE which are the charges according to the NYS Penal law. The LAW_CAT_CD which is the level of offense, this has three catagories being felony(F), misdemeanor(M), and violation(V). The LAW_CODE will not be used due to it being confusing to understand, in the data set the LAW_CODE shows the actual number of the law without a description making it harder to understand whats going on. As for the description of the arrest, to keep things more clean I plan to only use one of the description variables thus I will be opting to utilize OFNS_DESC over PD_DESC because it's a more generalized and easier to understand.

```
selected_columns <- c('PD_DESC', 'OFNS_DESC', 'LAW_CODE', 'LAW_CAT_CD')
selected_data <- data[selected_columns]
head(selected_data)
```

```
##                       PD_DESC       OFNS_DESC   LAW_CODE LAW_CAT_CD
## 1    ASSAULT 2,1,UNCLASSIFIED  FELONY ASSAULT PL 1200501          F
## 2 CONTROLLED SUBSTANCE,SALE 3 DANGEROUS DRUGS PL 2203901          F
## 3           STRANGULATION 1ST  FELONY ASSAULT PL 1211200          F
## 4           STRANGULATION 1ST  FELONY ASSAULT PL 1211200          F
## 5    ASSAULT 2,1,UNCLASSIFIED  FELONY ASSAULT PL 12005WX          F
## 6           STRANGULATION 1ST  FELONY ASSAULT PL 1211200          F
```

The data set also has a group of variables on when and where the arrest was as well as who was responsible for the arrest. It has information on the ARREST_DATE. It has the ARREST_BORRO which is the borough of the arrest with the catagories of the Bronx(B), Staten Island(S), Brooklyn(K), Manhattan(M), and Queens(Q), as well as the ARREST_PRECINT which has information on what precint the arrest occured in. The data set also has more specific variables explaining where the arrest occured through the X_COORD_CD and the Y_COORD_CD which are x and y coordinates based on the New York State Plane Coordinate System as well the Latitude and Longitude variables which are the coordinates based on the Global Coordinate System. The data set also has information on who actually performed the arrest with the variable JURISDICTION_CODE which is a code for the jurisdiction that performed the arrest with the catagories being 0-3 codes representing NYPD jurisdictions with Patrol(0), Transit(1), Housing(3), and codes above 3 represent non-NYPD jurisdictions. I will most likely not need the ARREST_PRECINT due to it's dificulty to understand. I will opted to use the Latitude and Longitude variables instead of the X_COORD_CD and Y_COORD_CD to keep things cleaner and because it's easier to pinpoint locations using latitude and longitude.

```
selected_columns <- c('ARREST_DATE', 'ARREST_BORO', 'ARREST_PRECINCT', 'JURISDICTION_CODE',
                      'X_COORD_CD', 'Y_COORD_CD', 'Latitude', 'Longitude')
selected_data <- data[selected_columns]
head(selected_data)
```

```
##   ARREST_DATE ARREST_BORO ARREST_PRECINCT JURISDICTION_CODE X_COORD_CD
## 1  01/01/2023           K              77                 0     999335
## 2  02/03/2023           K              73                 0    1009318
## 3  02/15/2023           K              62                 0     982272
## 4  01/04/2023           M              32                 0     999899
## 5  02/13/2023           K              71                 0    1001437
## 6  01/17/2023           Q             113                 0    1040264
##   Y_COORD_CD Latitude Longitude
## 1     186085 40.67743 -73.94562
## 2     178259 40.65592 -73.90965
## 3     158771 40.60247 -74.00712
## 4     238684 40.82180 -73.94346
## 5     183080 40.66918 -73.93804
## 6     190275 40.68876 -73.79802
```

Lastly, the data set contains information on the perpetraters themselves. This inlcudes the AGE_GROUP the perpatrator is in, the perpatrators sex (PERP_SEX), and the perpatrators race(PERP_RACE). All of these variables will be useful when it comes to analyzing this data set.

```r
selected_columns <- c('AGE_GROUP', 'PERP_SEX', 'PERP_RACE')
selected_data <- data[selected_columns]
head(selected_data)
```

```
##   AGE_GROUP PERP_SEX PERP_RACE
## 1     45-64        F     BLACK
## 2     25-44        M     BLACK
## 3     25-44        M     WHITE
## 4     25-44        M     BLACK
## 5     25-44        M     BLACK
## 6     45-64        F     BLACK
```

```r
########################################################################
# CLEAN THE DATA
########################################################################

# setup to use only the columns I plan to use
selected_columns <- c('ARREST_DATE', 'OFNS_DESC', 'LAW_CAT_CD', 'ARREST_BORO', 'JURISDICTION_CODE', 'AG
data <- data[selected_columns]

# Convert ARREST_DATE to Date format
data$ARREST_DATE <- as.Date(data$ARREST_DATE, format = "%m/%d/%Y")

# Check unique values in categorical variables
sapply(data[, c("OFNS_DESC", "LAW_CAT_CD", "ARREST_BORO", "PERP_SEX", "PERP_RACE")], unique)
```

```
## $OFNS_DESC
##  [1] "FELONY ASSAULT"
##  [2] "DANGEROUS DRUGS"
##  [3] "RAPE"
##  [4] "FORGERY"
##  [5] "BURGLARY"
##  [6] "ARSON"
##  [7] "ASSAULT 3 & RELATED OFFENSES"
##  [8] "ROBBERY"
##  [9] "PETIT LARCENY"
## [10] "DANGEROUS WEAPONS"
## [11] "MISCELLANEOUS PENAL LAW"
## [12] "HARRASSMENT 2"
## [13] "OFF. AGNST PUB ORD SENSBLTY &"
## [14] "JOSTLING"
## [15] "SEX CRIMES"
## [16] "(null)"
## [17] "FRAUDS"
## [18] "ESCAPE 3"
## [19] "BURGLAR'S TOOLS"
## [20] "VEHICLE AND TRAFFIC LAWS"
## [21] "OFFENSES AGAINST THE PERSON"
## [22] "OFFENSES INVOLVING FRAUD"
## [23] "INTOXICATED & IMPAIRED DRIVING"
## [24] "FOR OTHER AUTHORITIES"
## [25] "OTHER OFFENSES RELATED TO THEF"
## [26] "POSSESSION OF STOLEN PROPERTY"
```

```
## [27] "OTHER TRAFFIC INFRACTION"
## [28] "GRAND LARCENY"
## [29] "CRIMINAL MISCHIEF & RELATED OF"
## [30] "OTHER STATE LAWS (NON PENAL LA"
## [31] "PROSTITUTION & RELATED OFFENSES"
## [32] "GAMBLING"
## [33] "CRIMINAL TRESPASS"
## [34] "OFFENSES AGAINST PUBLIC ADMINI"
## [35] "MURDER & NON-NEGL. MANSLAUGHTE"
## [36] "OTHER STATE LAWS"
## [37] "CANNABIS RELATED OFFENSES"
## [38] "NYS LAWS-UNCLASSIFIED FELONY"
## [39] "OFFENSES AGAINST PUBLIC SAFETY"
## [40] "GRAND LARCENY OF MOTOR VEHICLE"
## [41] "UNAUTHORIZED USE OF A VEHICLE"
## [42] "ADMINISTRATIVE CODE"
## [43] "OFFENSES RELATED TO CHILDREN"
## [44] "THEFT-FRAUD"
## [45] "INTOXICATED/IMPAIRED DRIVING"
## [46] "ANTICIPATORY OFFENSES"
## [47] "FRAUDULENT ACCOSTING"
## [48] "THEFT OF SERVICES"
## [49] "ENDAN WELFARE INCOMP"
## [50] "OTHER STATE LAWS (NON PENAL LAW)"
## [51] "ALCOHOLIC BEVERAGE CONTROL LAW"
## [52] "MOVING INFRACTIONS"
## [53] "DISORDERLY CONDUCT"
## [54] "KIDNAPPING & RELATED OFFENSES"
## [55] "AGRICULTURE & MRKTS LAW-UNCLASSIFIED"
## [56] "HOMICIDE-NEGLIGENT,UNCLASSIFIE"
## [57] "PARKING OFFENSES"
## [58] "CHILD ABANDONMENT/NON SUPPORT"
## [59] "KIDNAPPING"
## [60] "UNLAWFUL POSS. WEAP. ON SCHOOL"
## [61] "DISRUPTION OF A RELIGIOUS SERV"
## [62] "HOMICIDE-NEGLIGENT-VEHICLE"
## [63] "FELONY SEX CRIMES"
## [64] "ADMINISTRATIVE CODES"
##
## $LAW_CAT_CD
## [1] "F" "M" "V" ""  "9" "I"
##
## $ARREST_BORO
## [1] "K" "M" "Q" "B" "S"
##
## $PERP_SEX
## [1] "F" "M" "U"
##
## $PERP_RACE
## [1] "BLACK"                          "WHITE"
## [3] "ASIAN / PACIFIC ISLANDER"       "WHITE HISPANIC"
## [5] "BLACK HISPANIC"                 "UNKNOWN"
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
```

```r
# # Filter out invalid values in LAW_CAT_CD
valid_law_cat <- c('F', 'M', 'V')
data <- data[data$LAW_CAT_CD %in% valid_law_cat, ]

# Check unique values in LAW_CAT_CD after filtering
# should only print out 'F' 'M' and 'V'
unique(data$LAW_CAT_CD)
```

```
## [1] "F" "M" "V"
```

```r
# change variables into factors
data$LAW_CAT_CD <- as.factor(data$LAW_CAT_CD)
data$ARREST_BORO <- as.factor(data$ARREST_BORO)
data$AGE_GROUP <- as.factor(data$AGE_GROUP)
data$PERP_SEX <- as.factor(data$PERP_SEX)
data$PERP_RACE <- as.factor(data$PERP_RACE)


# combine a bunch of the factors in OFNS_DESC to more generalized factors for less factors overall
category_mapping <- c(
  "FELONY ASSAULT" = "ASSAULT",
  "DANGEROUS DRUGS" = "DRUGS",
  "RAPE" = "SEX CRIMES",
  "FORGERY" = "FRAUD",
  "BURGLARY" = "BURGLARY",
  "ARSON" = "ARSON",
  "ASSAULT 3 & RELATED OFFENSES" = "ASSAULT",
  "ROBBERY" = "ROBBERY",
  "PETIT LARCENY" = "LARCENY",
  "DANGEROUS WEAPONS" = "WEAPONS",
  "MISCELLANEOUS PENAL LAW" = "OTHER PENAL LAW",
  "HARRASSMENT 2" = "HARRASSMENT",
  "OFF. AGNST PUB ORD SENSBLTY &" = "PUBLIC ORDER OFFENSES",
  "JOSTLING" = "ASSAULT",
  "SEX CRIMES" = "SEX CRIMES",
  "FRAUDS" = "FRAUD",
  "ESCAPE 3" = "ESCAPE",
  "BURGLAR'S TOOLS" = "BURGLARY",
  "VEHICLE AND TRAFFIC LAWS" = "TRAFFIC",
  "OFFENSES AGAINST THE PERSON" = "OTHER OFFENSES",
  "OFFENSES INVOLVING FRAUD" = "FRAUD",
  "INTOXICATED & IMPAIRED DRIVING" = "TRAFFIC",
  "OTHER OFFENSES RELATED TO THEF" = "OTHER OFFENSES",
  "POSSESSION OF STOLEN PROPERTY" = "PROPERTY CRIME",
  "OTHER TRAFFIC INFRACTION" = "TRAFFIC",
  "GRAND LARCENY" = "LARCENY",
  "CRIMINAL MISCHIEF & RELATED OF" = "OTHER OFFENSES",
  "OTHER STATE LAWS (NON PENAL LA" = "OTHER OFFENSES",
  "PROSTITUTION & RELATED OFFENSES" = "PROSTITUTION",
  "GAMBLING" = "OTHER OFFENSES",
  "CRIMINAL TRESPASS" = "TRESPASS",
  "OFFENSES AGAINST PUBLIC ADMINI" = "PUBLIC ORDER OFFENSES",
  "MURDER & NON-NEGL. MANSLAUGHTE" = "HOMICIDE",
```

```
  "OTHER STATE LAWS" = "OTHER OFFENSES",
  "CANNABIS RELATED OFFENSES" = "DRUGS",
  "NYS LAWS-UNCLASSIFIED FELONY" = "OTHER OFFENSES",
  "OFFENSES AGAINST PUBLIC SAFETY" = "PUBLIC ORDER OFFENSES",
  "GRAND LARCENY OF MOTOR VEHICLE" = "LARCENY",
  "UNAUTHORIZED USE OF A VEHICLE" = "TRAFFIC",
  "ADMINISTRATIVE CODE" = "OTHER OFFENSES",
  "OFFENSES RELATED TO CHILDREN" = "OTHER OFFENSES",
  "THEFT-FRAUD" = "FRAUD",
  "INTOXICATED/IMPAIRED DRIVING" = "TRAFFIC",
  "ANTICIPATORY OFFENSES" = "OTHER OFFENSES",
  "FRAUDULENT ACCOSTING" = "FRAUD",
  "THEFT OF SERVICES" = "LARCENY",
  "ENDAN WELFARE INCOMP" = "OTHER OFFENSES",
  "OTHER STATE LAWS (NON PENAL LAW)" = "OTHER OFFENSES",
  "ALCOHOLIC BEVERAGE CONTROL LAW" = "OTHER OFFENSES",
  "DISORDERLY CONDUCT" = "DISORDERLY CONDUCT",
  "KIDNAPPING & RELATED OFFENSES" = "KIDNAPPING",
  "AGRICULTURE & MRKTS LAW-UNCLASSIFIED" = "OTHER OFFENSES",
  "HOMICIDE-NEGLIGENT,UNCLASSIFIE" = "HOMICIDE",
  "CHILD ABANDONMENT/NON SUPPORT" = "OTHER OFFENSES",
  "KIDNAPPING" = "KIDNAPPING",
  "UNLAWFUL POSS. WEAP. ON SCHOOL" = "WEAPONS",
  "DISRUPTION OF A RELIGIOUS SERV" = "OTHER OFFENSES",
  "HOMICIDE-NEGLIGENT-VEHICLE" = "HOMICIDE",
  "FELONY SEX CRIMES" = "SEX CRIMES",
  "ADMINISTRATIVE CODES" = "OTHER OFFENSES"
)
data$OFNS_DESC <- as.factor(category_mapping[as.character(data$OFNS_DESC)])

str(data)
```

```
## 'data.frame':    168186 obs. of  10 variables:
##  $ ARREST_DATE      : Date, format: "2023-01-01" "2023-02-03" ...
##  $ OFNS_DESC        : Factor w/ 21 levels "ARSON","ASSAULT",..: 2 5 2 2 2 2 2 18 7 3 ...
##  $ LAW_CAT_CD       : Factor w/ 3 levels "F","M","V": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ARREST_BORO      : Factor w/ 5 levels "B","K","M","Q",..: 2 2 2 3 2 4 3 2 3 3 ...
##  $ JURISDICTION_CODE: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AGE_GROUP        : Factor w/ 5 levels "<18","18-24",..: 4 3 3 3 3 4 3 1 4 4 ...
##  $ PERP_SEX         : Factor w/ 3 levels "F","M","U": 1 2 2 2 2 1 2 2 2 2 ...
##  $ PERP_RACE        : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 3 3 6 3 3 3 3 3 3 3 ..
##  $ Latitude         : num  40.7 40.7 40.6 40.8 40.7 ...
##  $ Longitude        : num  -73.9 -73.9 -74 -73.9 -73.9 ...
```

```
# check if there is any NULL's in the data and clean them out
sapply(data, function(x) sum(is.na(x)))
```

```
##       ARREST_DATE         OFNS_DESC       LAW_CAT_CD       ARREST_BORO
##                 0                13                0                 0
## JURISDICTION_CODE         AGE_GROUP          PERP_SEX        PERP_RACE
##                 0                 0                 0                 0
##          Latitude         Longitude
##                 0                 0
```

```r
data <- na.omit(data)

summary(data)
```

```
##   ARREST_DATE                    OFNS_DESC     LAW_CAT_CD ARREST_BORO
##  Min.   :2023-01-01   ASSAULT         :41323   F:73853    B:39646
##  1st Qu.:2023-03-11   LARCENY         :27507   M:93347    K:46606
##  Median :2023-05-19   OTHER OFFENSES  :16056   V:  973    M:38902
##  Mean   :2023-05-17   TRAFFIC         :13658              Q:35613
##  3rd Qu.:2023-07-25   DRUGS           :11869              S: 7406
##  Max.   :2023-09-30   OTHER PENAL LAW:11132
##                       (Other)         :46628
##  JURISDICTION_CODE AGE_GROUP     PERP_SEX
##  Min.   : 0.000    <18  : 6242   F: 28827
##  1st Qu.: 0.000    18-24:29720   M:135866
##  Median : 0.000    25-44:96428   U:  3480
##  Mean   : 0.946    45-64:33018
##  3rd Qu.: 0.000    65+  : 2765
##  Max.   :97.000
##
##                            PERP_RACE        Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:  548   Min.   : 0.00   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      : 8907   1st Qu.:40.68   1st Qu.:-73.97
##  BLACK                         :81340   Median :40.73   Median :-73.92
##  BLACK HISPANIC                :16253   Mean   :40.74   Mean   :-73.92
##  UNKNOWN                       : 2424   3rd Qu.:40.82   3rd Qu.:-73.88
##  WHITE                         :16142   Max.   :40.91   Max.   :  0.00
##  WHITE HISPANIC                :42559
```

Modeling Introduction

For this project I decided to implement a total of 4 models. I wanted to implement a Linear Discriminant Analysis(LDA) model, a Random Forest model, a lLogistic Regression model, and a Clustering model.

Applying a Linear Discriminant Analysis (LDA) model to this dataset is advantageous, particularly when investigating the relationship between race and other variables. LDA works well with multivariate data and would work well with highlighting the differences between racial groups due to the various variables like offense type, age group, and arrest location. By using LDA, I can extract insights into how these variables collectively contribute to racial distribution in arrests. The model provides coefficients that signify the importance of each variable in distinguishing between racial categories important information in understanding the relationship between the various variables and the race of the perpatrator.

```r
########################################################################
# LDA MODEL
########################################################################
# Remove last 100 observations for testing
train_data <- data[1:(nrow(data) - 100), ]
test_data <- data[(nrow(data) - 99):nrow(data), ]

# Create a formula for the model
lda_formula <- PERP_RACE ~ OFNS_DESC + LAW_CAT_CD + ARREST_BORO + AGE_GROUP + PERP_SEX + Latitude + Long
```

```r
# Fit the LDA model on the training data
lda_model <- lda(lda_formula, data = train_data)

# Project the data onto LD axes
lda_projection <- predict(lda_model, newdata = test_data)

# Display the projected data
head(lda_projection$posterior)
```

```
##        AMERICAN INDIAN/ALASKAN NATIVE ASIAN / PACIFIC ISLANDER     BLACK
## 169996                     0.002603833               0.02080500 0.5612509
## 169997                     0.007153616               0.27139066 0.3017381
## 169998                     0.002543575               0.02710438 0.6663655
## 169999                     0.006949486               0.06509446 0.5322015
## 170000                     0.002227655               0.02533800 0.5768453
## 170001                     0.007822622               0.21803158 0.3197433
##        BLACK HISPANIC     UNKNOWN      WHITE WHITE HISPANIC
## 169996     0.09126216 0.006393235 0.08706813      0.2306168
## 169997     0.02565579 0.003750032 0.14507988      0.2452319
## 169998     0.06431123 0.004386903 0.02950470      0.2057837
## 169999     0.06562497 0.006085892 0.13865636      0.1853874
## 170000     0.05456122 0.003680149 0.12175446      0.2155932
## 170001     0.04217849 0.004247337 0.08739544      0.3205813
```

```r
# This will show the posterior probabilities for each class based on LD projections

# Make predictions on the test data
lda_predictions <- predict(lda_model, newdata = test_data)$class
# This will give you the predicted class labels

# Compare predictions to actual values
confusion_matrix <- table(lda_predictions, test_data$PERP_RACE)
```

The reason I did a Random Forest model is because a A Random Forest model is well-suited for examining the relationships within this dataset, especially concerning the impact of race on various variables. Its ability to handle diverse datasets and capture complex, nonlinear patterns makes it an effective choice. With factors like offense type, age group, and location influencing arrests, Random Forests can highlight variable importance and interactions. This model is particularly good at handling categorical variables, providing insights into the significance of each factor in predicting racial disparities in arrests.

```r
#########################################################################
# RANDOM FOREST MODEL
#########################################################################

# Set the fraction of data to be held out
# because data set is so large, use only about 100ish data points
test_fraction <- 0.0006

# Assuming your target variable is PERP_RACE
set.seed(123)  # for reproducibility
random_subset_rf <- sample.int(nrow(data), nrow(data) * test_fraction)
data.train_rf <- data[-random_subset_rf, ]
data.val_rf <- data[random_subset_rf, ]
```

```r
# Convert PERP_RACE to a factor
data.train_rf$PERP_RACE <- factor(data.train_rf$PERP_RACE)
data.val_rf$PERP_RACE <- factor(data.val_rf$PERP_RACE)

# Fit random forest model
rf.fit <- randomForest(PERP_RACE ~ ., data = data.train_rf, ntree = 25, importance = TRUE)

# Predict on validation set
rf.preds <- predict(rf.fit, data.val_rf)

# Calculate the confusion matrix
conf_matrix <- table(rf.preds, data.val_rf$PERP_RACE)

# Calculate accuracy
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
```

A Multinomial Logistic Regression model is well-suited for this dataset because it accommodates for multiple categorical outcomes, making it apt for predicting diverse racial categories in arrest incidents. By estimating the probabilities of each race category, considering factors like offense type, age group, and arrest location, the model can explain the impact of each variable on the likelihood of a specific racial outcome. Its ability to examine interactions between predictors provides insights into nuanced relationships, enhancing the understanding of how various factors contribute to the racial distribution of arrests. Here is my Multinominal Logistic Regression model, I had to also implement the nnet package due to race having more than 2 factors. I also perform the cross-validation test in this model as well.

```r
##########################################################################
# MULTINOMINAL LOGISTIC REGRESSION
##########################################################################

# Create vectors to store the accuracy values for each fold
accuracy_list <- c()

# Set the number of folds
k <- 5

# Shuffle the data so that the folds are effectively random
data.shuffle <- data[sample(1:nrow(data)), ]

# Loop through each fold
for (fold in 1:k) {

  # Use the fold size to determine the validation set indexes
  val_indexes <- c(((fold - 1) * round(nrow(data) * test_fraction) + 1):min(fold * round(nrow(data) * t

  # Define the train and validation sets
  data.train <- data.shuffle[-val_indexes, ]
  data.val <- data.shuffle[val_indexes, ]

  # Fit the multinomial logistic regression model
  log_reg_fit <- multinom(PERP_RACE ~ ., data = data.train)

  # Predict on the validation set
  predictions_log_reg <- predict(log_reg_fit, newdata = data.val, type = "probs")
```

```r
  # Convert predicted probabilities to class labels
  predicted_classes <- colnames(predictions_log_reg)[apply(predictions_log_reg, 1, which.max)]

  # Evaluate predictions and save the accuracy value
  accuracy <- sum(predicted_classes == data.val$PERP_RACE) / nrow(data.val)
  accuracy_list <- append(accuracy_list, accuracy)
}
```

```
## # weights:  266 (222 variable)
## initial  value 327053.010572
## iter  10 value 250349.450203
## iter  20 value 246382.367726
## iter  30 value 244205.325768
## iter  40 value 241862.738083
## iter  50 value 236488.413953
## iter  60 value 230732.685733
## iter  70 value 228735.783206
## iter  80 value 226021.779843
## iter  90 value 222533.292562
## iter 100 value 220681.511047
## final  value 220681.511047
## stopped after 100 iterations
## # weights:  266 (222 variable)
## initial  value 327053.010572
## iter  10 value 251142.133017
## iter  20 value 246787.868199
## iter  30 value 244562.235243
## iter  40 value 242000.245699
## iter  50 value 235155.028949
## iter  60 value 230997.372796
## iter  70 value 228673.432855
## iter  80 value 225830.243386
## iter  90 value 222799.432939
## iter 100 value 221405.921750
## final  value 221405.921750
## stopped after 100 iterations
## # weights:  266 (222 variable)
## initial  value 327053.010572
## iter  10 value 250218.942408
## iter  20 value 245800.541650
## iter  30 value 243648.910925
## iter  40 value 241622.203725
## iter  50 value 235781.172193
## iter  60 value 229989.794299
## iter  70 value 228351.566137
## iter  80 value 225853.098736
## iter  90 value 222404.525440
## iter 100 value 220749.331259
## final  value 220749.331259
## stopped after 100 iterations
## # weights:  266 (222 variable)
## initial  value 327053.010572
## iter  10 value 250179.636183
```

```
## iter  20 value 245569.292648
## iter  30 value 243463.015635
## iter  40 value 241320.459071
## iter  50 value 235710.614964
## iter  60 value 230723.156214
## iter  70 value 228447.758888
## iter  80 value 225421.314155
## iter  90 value 223145.548506
## iter 100 value 221199.858086
## final  value 221199.858086
## stopped after 100 iterations
## # weights:  266 (222 variable)
## initial  value 327053.010572
## iter  10 value 251130.820302
## iter  20 value 246773.571064
## iter  30 value 244545.704125
## iter  40 value 241986.198076
## iter  50 value 235263.656837
## iter  60 value 228900.432580
## iter  70 value 227366.127601
## iter  80 value 225032.351834
## iter  90 value 222344.455867
## iter 100 value 221090.820484
## final  value 221090.820484
## stopped after 100 iterations
```

```
# Calculate the mean accuracy across all folds
mean_accuracy <- mean(accuracy_list)
```

I used a K-means clustering model because they are particularly good at understanding the relationship between race and other variables. By identifying natural clusters, this model provides insights into distinct arrest profiles, allowing for a more precise understanding of the potential associations and disparities across different groups in the dataset. In this model I decided to just focus on the clusters for the clustering of the

```
#########################################################################
# CLUSTERING  MODEL
#########################################################################
# Filter data because the outliers can cause errors
clustering_data <- data %>%
  filter(Latitude > quantile(data$Latitude, 0.05) & Latitude < quantile(data$Latitude, 0.95),
         Longitude > quantile(data$Longitude, 0.05) & Longitude < quantile(data$Longitude, 0.95))

# Assuming 'Latitude' and 'Longitude' are the columns representing coordinates
coordinates <- clustering_data[, c('Latitude', 'Longitude')]

# Standardize the data
scaled_coordinates <- scale(coordinates)

# Run K-Means clustering
k <- 4  # Adjust the number of clusters based on your analysis
km <- kmeans(scaled_coordinates, centers = k)

# Add cluster information to your original data
clustering_data$Cluster <- km$cluster
```

```r
# Plot pairs of variables
clusters <- ggplot(clustering_data, aes(Latitude, Longitude, color = as.factor(Cluster))) +
  geom_point(size = 3) +
  theme_bw() +
  theme(legend.title = element_blank())

# Calculate centroids for black arrests
centroids <- clustering_data %>%
  group_by(Cluster) %>%
  summarize(mean_latitude = mean(Latitude),
            mean_longitude = mean(Longitude))

# Plot pairs of variables for black arrests with centroids
clusters_w_centroids <- clusters +
  geom_point(data = centroids, aes(x = mean_latitude, y = mean_longitude),
             color = "black", size = 5, shape = 4) +   # X shape
  geom_label(data = centroids, aes(x = mean_latitude, y = mean_longitude,
                                   label = paste("(", round(mean_latitude, 4), ", ", round(mean_longitu
             color = "black", vjust = 1.4, hjust = 1) +  # Adjust label position
  ggtitle("Clustering and Centroids")
```

The biggest limitation is that the data set as seen earlier mostly has catagorical variables, thus meaning most models are unusable or much harder to implement. Like in the case of Logistic regression, because PERP_RACE is has more than two factors, normal logistic regression couldn't be done and I had to resort to doing Multinominal Logistic Regression. Not only does it make the data harder to model, but also it makes it so that the results of the models are also a bit weirder to analyze and understand.

Analysis Results

The Linear Discriminant Analysis (LDA) model was trained to predict the perpetrator's race based on various features such as offense description, arrest borough, age group, and geographic coordinates. The confusion matrix reveals the model's performance across different racial categories. Notably, the accuracy of the model is 45%, suggesting that it correctly predicted the perpetrator's race in 45% of cases. However, the Kappa value of 0.1149 indicates only slight agreement beyond what would be expected by chance. The model faces challenges in accurately predicting certain racial classes, as evidenced by varying sensitivities and low precision for specific groups. For instance, the sensitivity for the "BLACK" class is high (95.12%), but precision is relatively low (44.32%). Similar patterns are observed for other classes. The overall balanced accuracy is modest, reflecting the trade-off between sensitivity and specificity.

Based on this analysis the variables can be seen as having an significant impact on the ability to determine a perpetrator's race. But because the sensitivity for the model guessing that the perpetrator would be black and only having a 44.32% success rate, one could make the claim that due to the majority of the arrests in the data being black, this could have skewed the results.

```r
# LDA analysis
cat("Summary of LDA Model:", "\n")
```

```
## Summary of LDA Model:
```

```
summary(lda_model)
```

```
##         Length Class  Mode
## prior      7   -none- numeric
## counts     7   -none- numeric
## means    238   -none- numeric
## scaling  204   -none- numeric
## lev        7   -none- character
## svd        6   -none- numeric
## N          1   -none- numeric
## call       3   -none- call
## terms      3   terms  call
## xlevels    5   -none- list
```

```
# print confusion matrix using caret package
confusionMatrix(lda_predictions, test_data$PERP_RACE)
```

```
## Confusion Matrix and Statistics
##
##                                 Reference
## Prediction                       AMERICAN INDIAN/ALASKAN NATIVE
##   AMERICAN INDIAN/ALASKAN NATIVE                              0
##   ASIAN / PACIFIC ISLANDER                                    0
##   BLACK                                                       0
##   BLACK HISPANIC                                              0
##   UNKNOWN                                                     0
##   WHITE                                                       0
##   WHITE HISPANIC                                              0
##                                 Reference
## Prediction                       ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     0              0
##   ASIAN / PACIFIC ISLANDER                              0     0              0
##   BLACK                                                 5    39             10
##   BLACK HISPANIC                                        0     0              0
##   UNKNOWN                                               0     1              1
##   WHITE                                                 0     0              0
##   WHITE HISPANIC                                        1     1              0
##                                 Reference
## Prediction                       UNKNOWN WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0     0              0
##   BLACK                                2    11             21
##   BLACK HISPANIC                       0     0              0
##   UNKNOWN                              1     0              0
##   WHITE                                1     3              1
##   WHITE HISPANIC                       0     0              2
##
## Overall Statistics
##
##                Accuracy : 0.45
##                  95% CI : (0.3503, 0.5527)
##     No Information Rate : 0.41
```

```
##     P-Value [Acc > NIR] : 0.2375
##
##                   Kappa : 0.1149
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: AMERICAN INDIAN/ALASKAN NATIVE
## Sensitivity                                             NA
## Specificity                                              1
## Pos Pred Value                                          NA
## Neg Pred Value                                          NA
## Prevalence                                               0
## Detection Rate                                           0
## Detection Prevalence                                     0
## Balanced Accuracy                                       NA
##                      Class: ASIAN / PACIFIC ISLANDER Class: BLACK
## Sensitivity                                     0.00       0.9512
## Specificity                                     1.00       0.1695
## Pos Pred Value                                   NaN       0.4432
## Neg Pred Value                                  0.94       0.8333
## Prevalence                                      0.06       0.4100
## Detection Rate                                  0.00       0.3900
## Detection Prevalence                            0.00       0.8800
## Balanced Accuracy                               0.50       0.5604
##                      Class: BLACK HISPANIC Class: UNKNOWN Class: WHITE
## Sensitivity                           0.00         0.2500       0.2143
## Specificity                           1.00         0.9792       0.9767
## Pos Pred Value                         NaN         0.3333       0.6000
## Neg Pred Value                        0.89         0.9691       0.8842
## Prevalence                            0.11         0.0400       0.1400
## Detection Rate                        0.00         0.0100       0.0300
## Detection Prevalence                  0.00         0.0300       0.0500
## Balanced Accuracy                     0.50         0.6146       0.5955
##                      Class: WHITE HISPANIC
## Sensitivity                        0.08333
## Specificity                        0.97368
## Pos Pred Value                     0.50000
## Neg Pred Value                     0.77083
## Prevalence                         0.24000
## Detection Rate                     0.02000
## Detection Prevalence               0.04000
## Balanced Accuracy                  0.52851
```

The Random Forest model research sheds light on the elements that influence the prediction of perpetrator race in criminal occurrences. ARREST_DATE, OFNS_DESC (offense description), and geographical information such as Latitude and Longitude can all be considered as significant contributors to the model's accuracy and impurity reduction. The time factor given by ARREST_DATE emphasizes the importance of arrest timing in predicting perpetrator race. Similarly, the type of offense (OFNS_DESC) is important, implying that certain offenses are associated with specific racial groups. Geographical characteristics indicate that the location of the arrest gives useful information for predicting perpetrator race. It's definitely important to note that although the type of offense and the location of the arrests may be helpful in predicting a perpetrator's race. It could also be said that instead its more so that the race of the perpetrator is

the factor that can predict where the arrest was, and type of offense commited. Unfortunatly I did not do enough models to test this but in the future I would hope to do so.

```
cat("Summary of Random Forest Model:", "\n")
```

```
## Summary of Random Forest Model:
```

```
summary(rf.fit)
```

```
##                 Length  Class  Mode
## call                 5 -none- call
## type                 1 -none- character
## predicted       168073 factor numeric
## err.rate           200 -none- numeric
## confusion           56 -none- numeric
## votes           1176511 matrix numeric
## oob.times       168073 -none- numeric
## classes              7 -none- character
## importance          81 -none- numeric
## importanceSD        72 -none- numeric
## localImportance      0 -none- NULL
## proximity            0 -none- NULL
## ntree                1 -none- numeric
## mtry                 1 -none- numeric
## forest              14 -none- list
## y               168073 factor numeric
## test                 0 -none- NULL
## inbag                0 -none- NULL
## terms                3 terms  call
```

```
cat("Importance:", "\n")
```

```
## Importance:
```

```
importance((rf.fit))
```

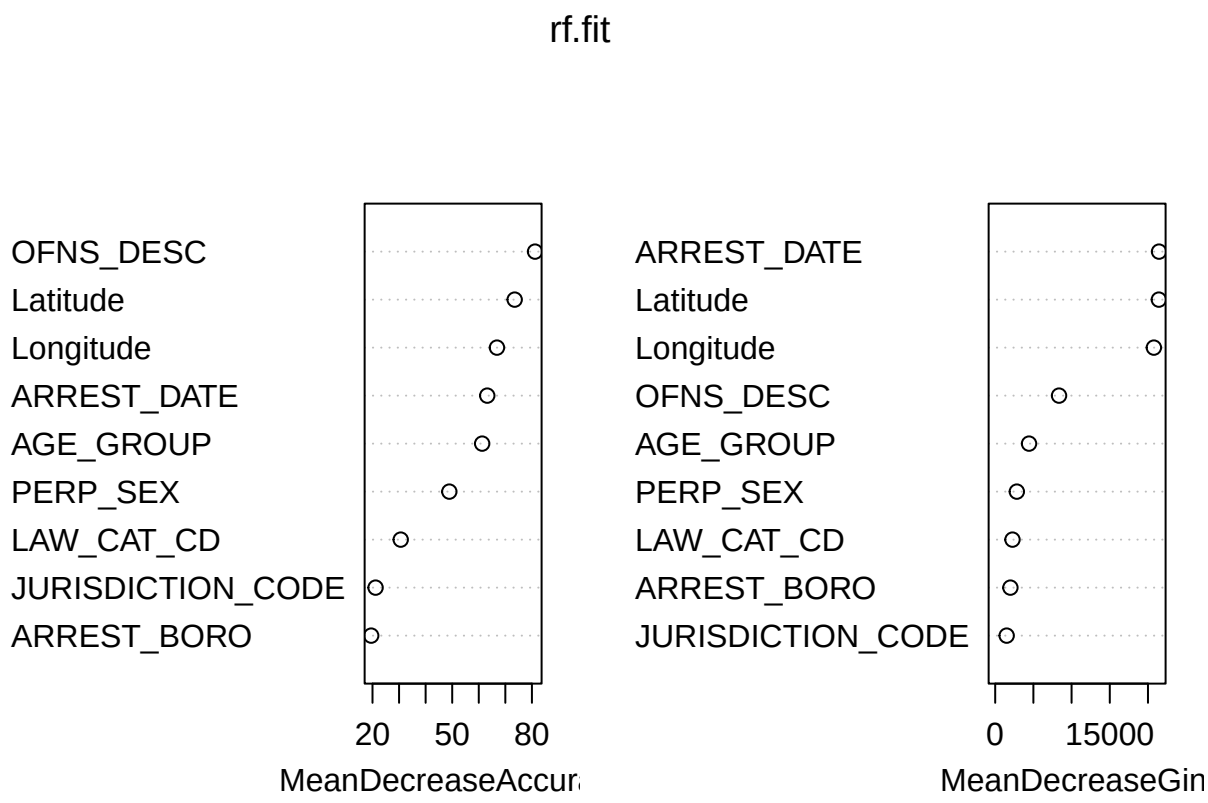
```
##                   AMERICAN INDIAN/ALASKAN NATIVE ASIAN / PACIFIC ISLANDER
## ARREST_DATE                           14.597045                 30.90722
## OFNS_DESC                             12.546890                 47.40216
## LAW_CAT_CD                             9.892974                 14.67879
## ARREST_BORO                           10.010663                 10.46481
## JURISDICTION_CODE                      3.354151                 15.42301
## AGE_GROUP                              9.272701                 28.69527
## PERP_SEX                               5.696261                 17.81220
## Latitude                              19.311700                 38.82224
## Longitude                             17.398298                 43.83145
##                      BLACK BLACK HISPANIC   UNKNOWN    WHITE WHITE HISPANIC
## ARREST_DATE       49.11145      33.191456 18.089595 41.61337      48.584578
## OFNS_DESC         55.71637      36.882891 11.533428 32.40422      48.406003
## LAW_CAT_CD        23.32098      17.672147  5.894095 18.56477      22.424023
## ARREST_BORO       17.41807       5.282252  3.055081  9.29638       7.934707
```

```
## JURISDICTION_CODE 17.06268      15.081603  7.245924 14.23062      17.975228
## AGE_GROUP         40.93281      45.296177  8.560093 44.17330      42.026251
## PERP_SEX          31.19444      27.561072 23.937474 20.92129      24.383479
## Latitude          46.90705      14.043807 15.457104 24.90009      20.099751
## Longitude         50.36865      14.360570  8.206155 24.91526      24.990472
##                   MeanDecreaseAccuracy MeanDecreaseGini
## ARREST_DATE                   63.18650        21450.122
## OFNS_DESC                     81.17889         8357.442
## LAW_CAT_CD                    30.62522         2263.247
## ARREST_BORO                   19.51057         1990.141
## JURISDICTION_CODE             21.15746         1500.829
## AGE_GROUP                     61.27879         4430.113
## PERP_SEX                      48.92162         2831.252
## Latitude                      73.50919        21424.116
## Longitude                     66.84125        20771.833
```

```
cat("VarImpPlot:", "\n")
```

```
## VarImpPlot:
```

```
varImpPlot(rf.fit)
```




```
print("Confusion Matrix:")
```

```
## [1] "Confusion Matrix:"
```

```
print(conf_matrix)
```

```
##
## rf.preds                          ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                         0     0              0
##   ASIAN / PACIFIC ISLANDER                               2     1              0
##   BLACK                                                  2    45              3
##   BLACK HISPANIC                                         0     0              2
##   UNKNOWN                                                0     0              0
##   WHITE                                                  0     0              0
##   WHITE HISPANIC                                         0     8              2
##
## rf.preds                          UNKNOWN WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0     0              0
##   BLACK                                0     7             15
##   BLACK HISPANIC                       0     0              1
##   UNKNOWN                              1     0              0
##   WHITE                                0     2              1
##   WHITE HISPANIC                       0     0              8
```

```
# print accuracy
print(paste("Accuracy: ", accuracy))
```

```
## [1] "Accuracy:  0.445544554455446"
```

The multinomial logistic regression model aimed to predict the perpetrator's race (PERP_RACE) based on various predictor variables. The coefficients provide insights into the impact of each predictor on the log-odds of the response variable for different racial categories. Notably, the significance of coefficients was evaluated through standard errors, indicating the precision of estimates. The results suggest that variables such as the type of offense (OFNS_DESC), arrest date (ARREST_DATE), and jurisdiction code (JURISDICTION_CODE) play significant roles in predicting the perpetrator's race. The model's performance, assessed by the residual deviance and AIC, indicates a reasonably good fit to the data. However, attention should be given to large standard errors, suggesting potential issues of overfitting or multicollinearity. The multinomial logistic regression model was also cross-validated with k=5 folds, yielding a mean accuracy of 51.5%. Like the Random Forest model, this model found both the type of offense and the arrest date to be significant on predicting the perpetrators race, unlike the other model, this model seems to find that the jurisdiction code plays a significant role in predicting the perpetrators race. This could mean that the both the type of arrest and the arrest date do have an impact on the prediction of the perps arrest. Unfortunatly this does not help prove nor disprove the idea of cops unfairly abusing power towards minorities.

```
summary(log_reg_fit)
```

```
## Call:
## multinom(formula = PERP_RACE ~ ., data = data.train)
##
## Coefficients:
##                           (Intercept)   ARREST_DATE OFNS_DESCASSAULT
## ASIAN / PACIFIC ISLANDER   -1.049720 -0.0005641705      -0.16903957
## BLACK                      13.821752 -0.0001177038       0.14865447
## BLACK HISPANIC             -2.137731  0.0001656322      -0.13761197
```

```
## UNKNOWN                    1.026469  0.0011269193       0.62699175
## WHITE                     -1.922736 -0.0010300972      -0.61514448
## WHITE HISPANIC            -9.865109 -0.0005975706      -0.07628827
##                          OFNS_DESCBURGLARY OFNS_DESCDISORDERLY CONDUCT
## ASIAN / PACIFIC ISLANDER       -0.66279517                 0.234741220
## BLACK                           0.12754993                 0.003872296
## BLACK HISPANIC                 -0.07064326                 0.550848653
## UNKNOWN                         0.57305799                 0.032819549
## WHITE                           0.02657709                -0.624213890
## WHITE HISPANIC                  0.04728494                -0.449751510
##                          OFNS_DESCDRUGS OFNS_DESCESCAPE OFNS_DESCFRAUD
## ASIAN / PACIFIC ISLANDER     0.02060226      -0.4412155    -0.22076097
## BLACK                        0.26261240       1.0853308     0.26284691
## BLACK HISPANIC               0.04552504       0.2907681     0.30146091
## UNKNOWN                      0.51768929      -0.2782775     0.65472584
## WHITE                        0.05770966      -0.2058511    -0.63105788
## WHITE HISPANIC              -0.03161790      -0.3952926     0.07621965
##                          OFNS_DESCHARRASSMENT OFNS_DESCHOMICIDE
## ASIAN / PACIFIC ISLANDER           -0.1104358        0.07855553
## BLACK                               0.1808701        0.63088256
## BLACK HISPANIC                      0.7212776        0.17296565
## UNKNOWN                             0.2879209        1.19335531
## WHITE                              -2.9500144       -0.60332696
## WHITE HISPANIC                     -0.4993709       -0.42939856
##                          OFNS_DESCKIDNAPPING OFNS_DESCLARCENY
## ASIAN / PACIFIC ISLANDER         -0.45210560    -0.5469072501
## BLACK                             0.04895750     0.2200742699
## BLACK HISPANIC                    0.36547144     0.0002273251
## UNKNOWN                           0.60836874     0.7159018362
## WHITE                            -0.10684360    -0.0731993441
## WHITE HISPANIC                   -0.06277246     0.0435218986
##                          OFNS_DESCOTHER OFFENSES OFNS_DESCOTHER PENAL LAW
## ASIAN / PACIFIC ISLANDER             -0.25182324               -0.3022205
## BLACK                                 0.22088173                0.1121789
## BLACK HISPANIC                       -0.02833329               -0.1631845
## UNKNOWN                               0.47399886                0.5123288
## WHITE                                -0.20544205               -0.1289075
## WHITE HISPANIC                       -0.18310853               -0.2262512
##                          OFNS_DESCPROPERTY CRIME OFNS_DESCPROSTITUTION
## ASIAN / PACIFIC ISLANDER             -0.18684129              6.544886
## BLACK                                 0.15470348              5.121571
## BLACK HISPANIC                        0.20776923             -4.435622
## UNKNOWN                               0.75413029             -9.952686
## WHITE                                -0.74551343             -3.199820
## WHITE HISPANIC                       -0.07316255              5.001031
##                          OFNS_DESCPUBLIC ORDER OFFENSES OFNS_DESCROBBERY
## ASIAN / PACIFIC ISLANDER                    -0.3287764      -0.88857638
## BLACK                                        0.4064260       0.45705071
## BLACK HISPANIC                              -0.1262340       0.03418664
## UNKNOWN                                      0.8178444       0.72170914
## WHITE                                       -0.3070190      -0.45161489
## WHITE HISPANIC                              -0.2479346      -0.14813669
##                          OFNS_DESCSEX CRIMES OFNS_DESCTRAFFIC OFNS_DESCTRESPASS
## ASIAN / PACIFIC ISLANDER         0.08470663       -0.4483247        -0.1288815
```

```
## BLACK                            0.01605648        0.1176079         0.2317900
## BLACK HISPANIC                  -0.21528391        0.0156126        -0.3299622
## UNKNOWN                          0.60773805        0.6231065         0.5498292
## WHITE                           -0.65397207       -0.7055210        -0.2076773
## WHITE HISPANIC                   0.04290005       -0.2158533        -0.0541822
##                           OFNS_DESCWEAPONS LAW_CAT_CDM LAW_CAT_CDV ARREST_BOROK
## ASIAN / PACIFIC ISLANDER      -0.65532863  0.04769191 -0.40433772   0.32776085
## BLACK                          0.30680900 -0.08292610  0.04904157   0.33568305
## BLACK HISPANIC                -0.01611568  0.00796691 -0.02087490  -0.64901491
## UNKNOWN                        0.81388312 -0.04634401  0.14267009  -0.49119380
## WHITE                         -0.30393559  0.27787200 -0.15512768  -0.68303022
## WHITE HISPANIC                -0.31113829  0.09329454  0.50550206  -0.04063161
##                           ARREST_BOROM ARREST_BOROQ ARREST_BOROS
## ASIAN / PACIFIC ISLANDER   0.27126888  1.503131528   -0.2824934
## BLACK                      0.06132957 -0.361567811    0.1154720
## BLACK HISPANIC            -0.29207150 -1.219870706   -0.3454011
## UNKNOWN                   -0.20996909 -0.619437253   -1.0750633
## WHITE                     -0.08466663 -0.002746735   -1.1040582
## WHITE HISPANIC             0.01027054  0.017508742    0.2018539
##                           JURISDICTION_CODE AGE_GROUP18-24 AGE_GROUP25-44
## ASIAN / PACIFIC ISLANDER      0.0064676798     0.00560854    -0.00963043
## BLACK                        -0.0233176270    -0.18389006    -0.21503160
## BLACK HISPANIC               -0.0084287100    -0.30218946    -0.40980601
## UNKNOWN                       0.0037539791     0.22020330     0.17360588
## WHITE                        -0.0002071838     0.20291270     0.67842631
## WHITE HISPANIC               -0.0060633688     0.04223806     0.01045244
##                           AGE_GROUP45-64 AGE_GROUP65+    PERP_SEXM   PERP_SEXU
## ASIAN / PACIFIC ISLANDER     0.19751079    0.5385988  0.086135496 -0.03142283
## BLACK                       -0.18368258   -0.3842203 -0.031071592 -0.13197600
## BLACK HISPANIC              -0.80530169   -1.2698413  0.148726222 -0.10739127
## UNKNOWN                      0.02908281   -0.0196708  0.108330799  2.04962273
## WHITE                        0.84812074    1.3435852 -0.347011492 -0.57961225
## WHITE HISPANIC              -0.22567193   -0.5465529 -0.006617268 -0.08665460
##                            Latitude  Longitude
## ASIAN / PACIFIC ISLANDER -2.203466 -1.3844034
## BLACK                     1.400968  0.8822246
## BLACK HISPANIC            3.802214  2.0776603
## UNKNOWN                  -1.642780 -0.5904670
## WHITE                    -7.892031 -4.6677341
## WHITE HISPANIC            2.549726  1.0768217
##
## Std. Errors:
##                            (Intercept)  ARREST_DATE OFNS_DESCASSAULT
## ASIAN / PACIFIC ISLANDER 2.132059e-09 1.117556e-06     7.590332e-07
## BLACK                    2.179525e-09 9.783811e-07     5.594807e-07
## BLACK HISPANIC           1.828826e-09 1.053693e-06     7.626387e-08
## UNKNOWN                  2.538087e-09 1.267203e-06     1.062553e-07
## WHITE                    2.001175e-09 1.053568e-06     4.952019e-07
## WHITE HISPANIC           1.371841e-09 9.942332e-07     3.501552e-07
##                           OFNS_DESCBURGLARY OFNS_DESCDISORDERLY CONDUCT
## ASIAN / PACIFIC ISLANDER      1.849497e-07                1.605634e-09
## BLACK                         3.335108e-07                4.807702e-10
## BLACK HISPANIC                2.833870e-07                1.974648e-09
## UNKNOWN                       3.449895e-07                1.899284e-09
```

```
## WHITE                          3.682123e-07               8.363369e-10
## WHITE HISPANIC                 2.254894e-07               1.251175e-09
##                           OFNS_DESCDRUGS OFNS_DESCESCAPE OFNS_DESCFRAUD
## ASIAN / PACIFIC ISLANDER   6.966216e-07    1.523948e-10   4.045591e-08
## BLACK                      3.652900e-07    5.514303e-10   1.952286e-07
## BLACK HISPANIC             3.956090e-07    5.093476e-10   3.318700e-07
## UNKNOWN                    6.919838e-07    2.190518e-10   1.148501e-07
## WHITE                      5.485635e-07    4.149116e-10   4.015195e-08
## WHITE HISPANIC             3.945090e-07    1.810128e-10   1.689891e-07
##                           OFNS_DESCHARRASSMENT OFNS_DESCHOMICIDE
## ASIAN / PACIFIC ISLANDER         2.816013e-09      5.771253e-08
## BLACK                            2.725993e-09      5.660263e-08
## BLACK HISPANIC                   5.347808e-09      3.073045e-08
## UNKNOWN                          3.694145e-09      4.570827e-08
## WHITE                            1.920416e-10      1.834934e-08
## WHITE HISPANIC                   2.524737e-09      1.820277e-08
##                           OFNS_DESCKIDNAPPING OFNS_DESCLARCENY
## ASIAN / PACIFIC ISLANDER        4.717052e-09     6.717546e-07
## BLACK                           6.424470e-09     1.251497e-06
## BLACK HISPANIC                  9.940580e-09     1.286882e-06
## UNKNOWN                         7.622661e-09     1.458904e-06
## WHITE                           5.388039e-09     1.289336e-06
## WHITE HISPANIC                  4.726926e-09     9.081201e-07
##                           OFNS_DESCOTHER OFFENSES OFNS_DESCOTHER PENAL LAW
## ASIAN / PACIFIC ISLANDER             4.051886e-07              1.870950e-07
## BLACK                                8.587041e-07              3.456404e-07
## BLACK HISPANIC                       4.800806e-07              4.605139e-07
## UNKNOWN                              3.297169e-07              4.422714e-07
## WHITE                                4.798819e-07              3.238810e-07
## WHITE HISPANIC                       4.332618e-07              2.517683e-07
##                           OFNS_DESCPROPERTY CRIME OFNS_DESCPROSTITUTION
## ASIAN / PACIFIC ISLANDER             5.036845e-08          2.651455e-08
## BLACK                                9.967885e-08          2.278307e-08
## BLACK HISPANIC                       1.333773e-07          1.784670e-13
## UNKNOWN                              7.973785e-08          1.075775e-15
## WHITE                                8.279204e-09          2.385538e-12
## WHITE HISPANIC                       5.860476e-08          3.846133e-09
##                           OFNS_DESCPUBLIC ORDER OFFENSES OFNS_DESCROBBERY
## ASIAN / PACIFIC ISLANDER                   5.508195e-07     1.992362e-07
## BLACK                                      1.376785e-06     5.466529e-07
## BLACK HISPANIC                             1.573434e-06     5.615465e-07
## UNKNOWN                                    1.555932e-06     5.344874e-07
## WHITE                                      8.017370e-07     2.287082e-07
## WHITE HISPANIC                             8.977508e-07     3.040525e-07
##                           OFNS_DESCSEX CRIMES OFNS_DESCTRAFFIC OFNS_DESCTRESPASS
## ASIAN / PACIFIC ISLANDER        1.967935e-07     1.610453e-07      1.383361e-08
## BLACK                           9.461664e-08     6.213724e-07      8.252761e-08
## BLACK HISPANIC                  4.961797e-08     7.574397e-07      1.831656e-08
## UNKNOWN                         7.265800e-09     6.193132e-07      1.613242e-08
## WHITE                           6.700515e-08     3.149497e-07      2.082822e-08
## WHITE HISPANIC                  8.587204e-08     4.100905e-07      4.393187e-08
##                           OFNS_DESCWEAPONS  LAW_CAT_CDM  LAW_CAT_CDV
## ASIAN / PACIFIC ISLANDER     3.604073e-08 1.341833e-06 2.504193e-08
## BLACK                        1.353768e-07 4.817366e-07 2.434477e-08
```

```
## BLACK HISPANIC               2.908015e-08 6.536601e-07 2.099565e-08
## UNKNOWN                      3.259830e-08 9.780312e-07 4.862473e-08
## WHITE                        1.709527e-07 1.271216e-06 2.224281e-08
## WHITE HISPANIC               6.189473e-08 7.481514e-07 3.493951e-08
##                           ARREST_BOROK ARREST_BOROM ARREST_BOROQ ARREST_BOROS
## ASIAN / PACIFIC ISLANDER 5.126562e-07 1.357008e-06 2.108526e-06 1.185631e-07
## BLACK                    1.404819e-06 1.069803e-06 1.526735e-06 5.190833e-07
## BLACK HISPANIC           8.998282e-07 1.424198e-07 7.461799e-07 1.877827e-07
## UNKNOWN                  1.753661e-06 7.324858e-07 1.177717e-06 2.479754e-07
## WHITE                    1.317832e-06 1.579682e-06 8.693514e-07 6.423917e-07
## WHITE HISPANIC           7.275288e-07 7.278532e-07 8.771283e-07 2.001736e-07
##                           JURISDICTION_CODE AGE_GROUP18-24 AGE_GROUP25-44
## ASIAN / PACIFIC ISLANDER      0.0008104139   3.671293e-07   2.834222e-07
## BLACK                         0.0007568552   5.677303e-07   3.439789e-07
## BLACK HISPANIC                0.0008668791   7.274387e-07   2.886615e-07
## UNKNOWN                       0.0010668836   8.027864e-07   6.043083e-07
## WHITE                         0.0007840267   3.662494e-07   3.667622e-07
## WHITE HISPANIC                0.0005942318   4.437994e-07   2.485230e-07
##                           AGE_GROUP45-64 AGE_GROUP65+    PERP_SEXM    PERP_SEXU
## ASIAN / PACIFIC ISLANDER   8.289634e-08 1.265487e-07 2.232175e-07 7.447009e-08
## BLACK                      3.183197e-07 8.382314e-08 4.048894e-07 5.499352e-07
## BLACK HISPANIC             1.258466e-07 3.667270e-08 5.753584e-08 8.259232e-08
## UNKNOWN                    1.157919e-07 1.621869e-07 5.294221e-07 6.322337e-07
## WHITE                      1.404265e-07 2.297257e-07 2.694249e-07 2.576541e-08
## WHITE HISPANIC             1.404595e-07 6.471485e-08 1.686978e-07 1.062621e-07
##                              Latitude    Longitude
## ASIAN / PACIFIC ISLANDER 3.458885e-07 3.275114e-07
## BLACK                    4.020819e-07 2.182404e-07
## BLACK HISPANIC           2.118977e-07 6.446037e-08
## UNKNOWN                  4.593910e-07 1.390343e-07
## WHITE                    4.193211e-07 1.062145e-07
## WHITE HISPANIC           2.119374e-07 1.377978e-07
##
## Residual Deviance: 442181.6
## AIC: 442625.6
```

```
# Print the mean accuracy
cat("\n", "Mean Accuracy:", mean_accuracy, "\n")
```

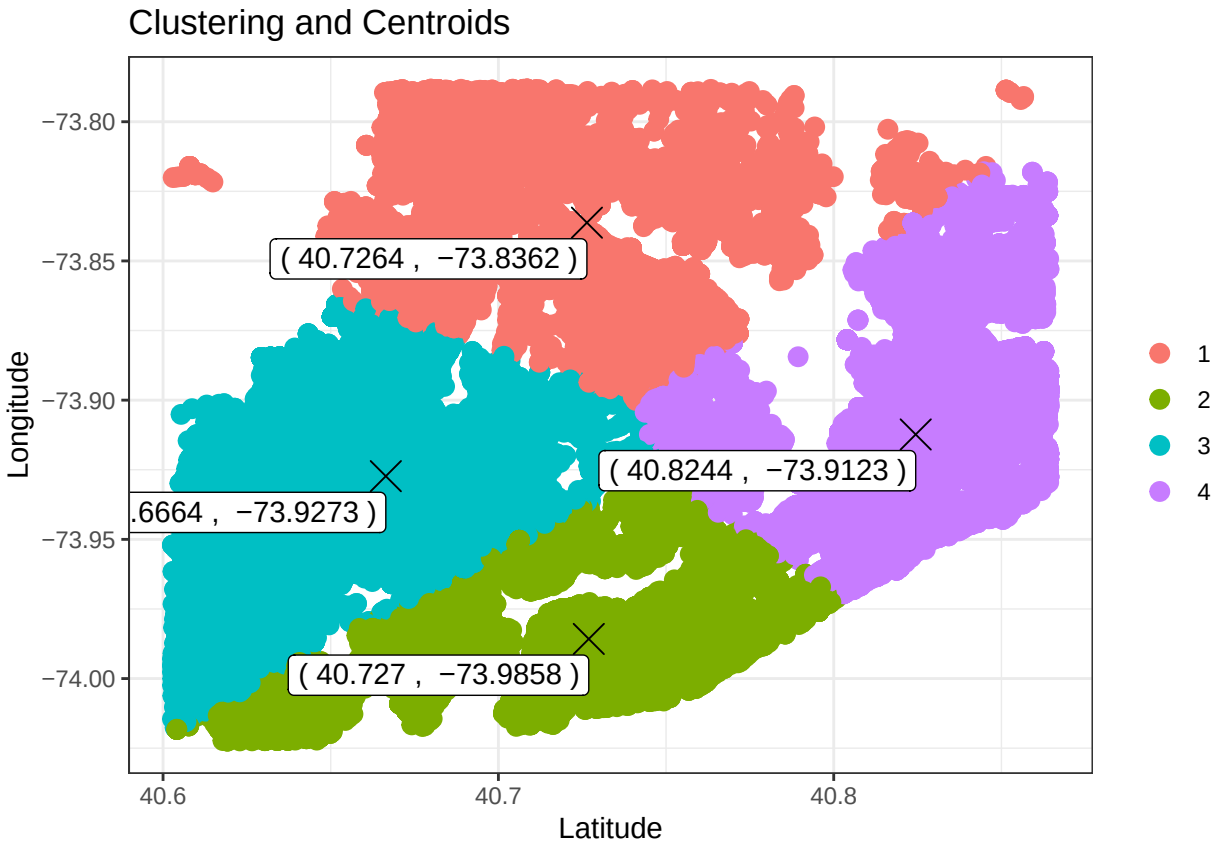
```
##
##  Mean Accuracy: 0.5148515
```

The cluster model, trained using the K-means function, looked into the various groupings that could be formed when looking at the latitudinal and longitudinal coordinates. The goal was to test if there were any clear "hotspots" of crime. This meant looking into the actual coordinates to see if the calculated centroids reflected anything in the data. Unfortunately with my limited knowledge about the Newyork streets and neighborhoods, most of the coordinates meant I knew nothing about. But interestingly, at least for the fourth centroid (40.82445, -73.91233) these coordinates led to a point in between Harlem and the Bronx, two places infamously known for being poorer areas. But because the data does not contain any information surrounding this topic it's harder to make any further assumptions.

```
# Display the centroids for black arrests
print(centroids)
```

```
## # A tibble: 4 x 3
##   Cluster mean_latitude mean_longitude
##     <int>         <dbl>          <dbl>
## 1       1          40.7          -73.8
## 2       2          40.7          -74.0
## 3       3          40.7          -73.9
## 4       4          40.8          -73.9
```

```
# print graph
clusters_w_centroids
```



Conclusions

Overall, I feel that based off the data, it's mostly unknown if police do have tendencies to unfairly treat minorities. Unfortunately this data set seemed to not have specific enough information to answer such a question. This data set was probable more likely suited to analyze the future tendencies of crime and not the tendencies of the police themselves. Answering such a question through this data set alone also seems rather more difficult than simply searching out stories and examples of police brutality and hate crimes. But nonetheless this project was a real eyeopening experience to the actual capabilities data science and statics could have. It was eyeopening to realize that actual possibilities that data science could pursue, and in my eyes, I could see how through the use of data science people have the capability to help the world in a powerful way. In the future though, I will try to pick a data set that is more suited to answer my desired questions, and I will also pick a data set with variables that have mostly numerical values because the categorical variables were a pain to work around during this project, but now that I am more knowledged about categorical variables maybe it won't be as much of a problem for the future.

Citations Justin Pakzad. (2023, December). NYPD Arrests Dataset (2023), Version 1. Retrieved December 11, 2023 from https://www.kaggle.com/datasets/justinpakzad/nypd-arrests-2023-dataset/data.