Team-Chosen Questions

Keizo Morgan, Max Holloway, Bowen Jiang, and Jorge Canedo

Introduction

Our group is concerned with discrimination in the Stop-Question-Frisk policy in 2015. In particular we would like to see what neighborhoods arrest more people of a certain race or gender. We explored multiple factors that may point to discrimination. In one test, we examine the rates of arrest for different races in each of the neighborhoods. Coupled with our knowledge of the population demographics of those neighborhoods in 2015, we can determine whether or not certain racial groups are subject to arrest more than others.

We will also explore how these rates changed between 2010 and 2015, testing to see if the changes are significant. While we hope that these statistical tests can help to identify racism in the NYPD, we acknowledge that the results of this study must be taken in context.

Initialize Data

Here we get the data necessary for the statistical analysis.

```
if (!file.exists("2015_sqf_m35.csv")) {
  download.file("http://math.hmc.edu/m35f/2015_sqf_m35.csv", "2015_sqf_m35.csv")
}
if(!file.exists("2015_sqf_m35.csv")){
  download.file('http://math.hmc.edu/m35f/2010_sqf_m35.csv',
                '2010_sqf_m35.csv')
}

sqf2010 <- read.csv("2010_sqf_m35.csv")
sqf2010 <- subset(sqf2010, sqf2010$weight > 50 & sqf2010$weight < 400)
sqf2010 <- subset(sqf2010, sqf2010$age < 100)

sqf2015 <- read.csv("2015_sqf_m35.csv")
sqf2015 = sqf2015[!sqf2015$perstop=="**"
                  & !sqf2015$perstop==" ",]
sqf2015$perstop = as.numeric(as.character(sqf2015$perstop))
```

Exploratory Graphical Analysis

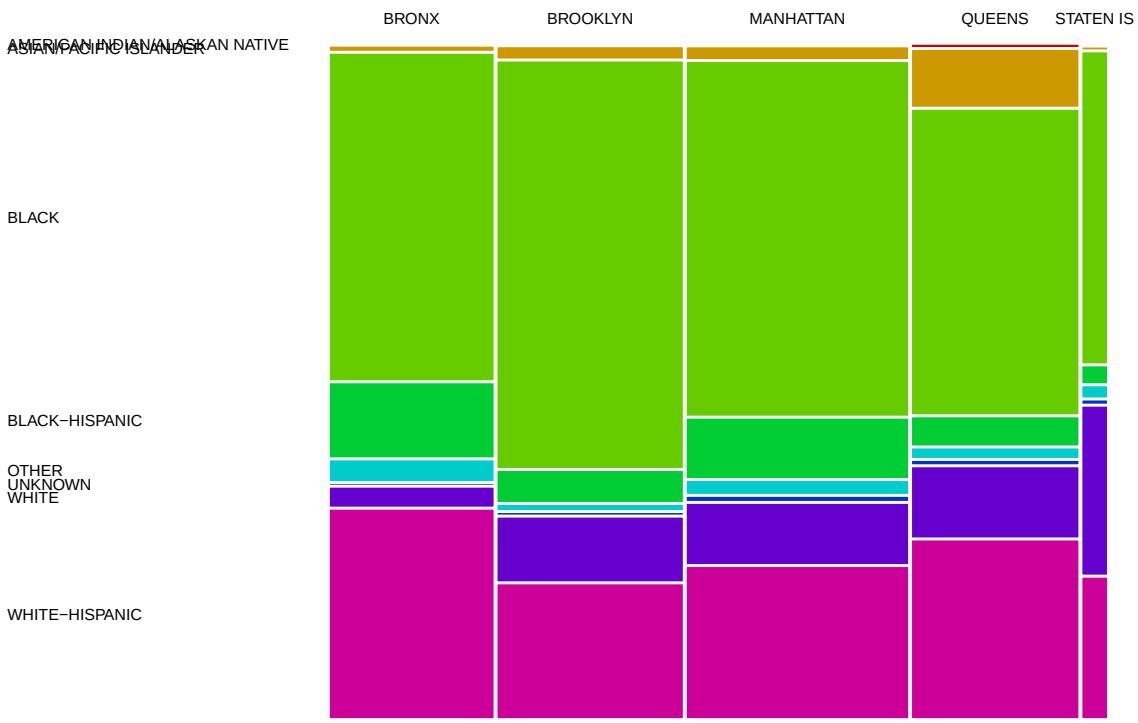
In order to get a summary of our data, we will first show some plots that shed light on the variables that we will analyze later.

Mosaic plot

```
sqf2010.arrested <- sqf2010[sqf2010$arstmade == 1,]
sqf2010.arrested <- sqf2010.arrested[!sqf2010.arrested$city == " ",]
sqf2010.arrested$city <- droplevels(sqf2010.arrested$city)
```

```
city_labels <- c("Bronx",
                 "Brooklyn",
                 "Manhatten",
                 "Queens",
                 "Staten Is.")
race_labels <- c("Amer. Indian",
            "Asian / Pac.",
            "Black",
            "Black-Hispanic",
            "Other",
            "Unknown",
            "White",
            "White-Hispanic")
city_colors = rainbow(5, v = 0.8)
race_colors = rainbow(8, v = 0.8)
```

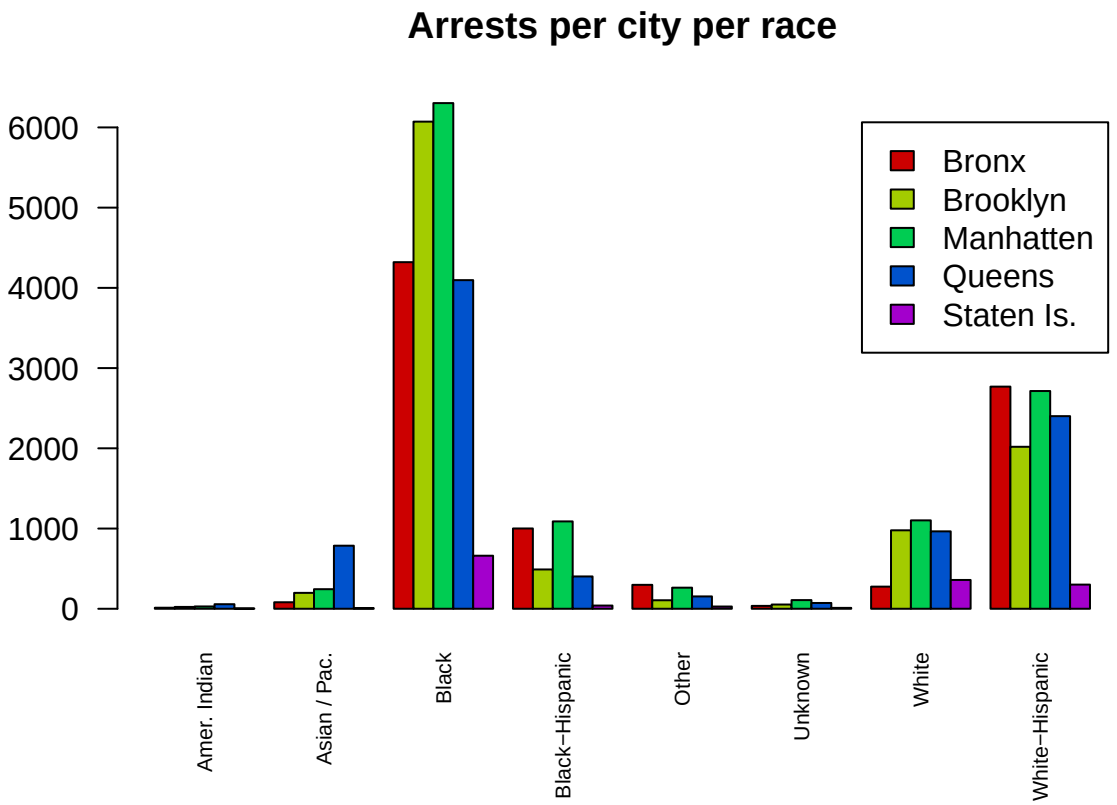
```
par(mar = c(1, 1, 3, 1))
mosaicplot(table(sqf2010.arrested$city, sqf2010.arrested$race),
           main = "Proportion of races per city of arrested",
           cex.axis = 0.5,
           las = 1,
           off = 1,
           color = race_colors,
           border = "white")
```



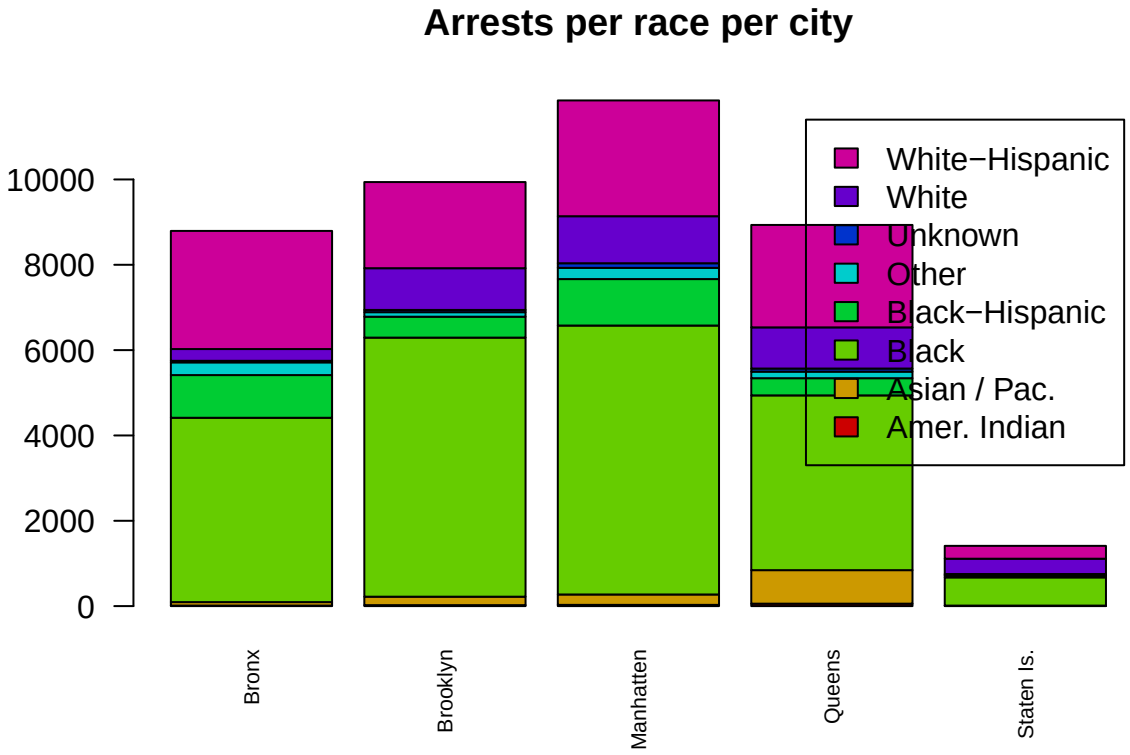
We note that compared to the average, there is a higher proportion of black arrests in Brooklyn. Furthermore, there is a higher proportion of White-Hispanic arrests in the Bronx. However, it is possible that this difference in proportion is caused by different demographics of each neighborhood.

Paired plot

```
barplot(table(sqf2010.arrested$city, sqf2010.arrested$race),
        main = "Arrests per city per race",
        beside = TRUE,
        names.arg = race_labels,
        col = city_colors,
        cex.names = 0.7,
        las = 2,
        legend.text = city_labels)
```




```
barplot(table(sqf2010.arrested$race, sqf2010.arrested$city),
        main = "Arrests per race per city",
        #beside = TRUE,
        names.arg = city_labels,
        col = race_colors,
        cex.names = 0.7,
        las = 2,
        legend.text = race_labels)
```



Now looking at the division of arrests numerically as opposed to proportionally, we can see that regardless of city, the most arrests were black arrests. Furthermore, the most arrests were made in Manhattan.

Proportion Testing for Races in different Neighborhoods

In this section, we will explore the rates of arrest for people of each race in each neighborhood. Did police in some neighborhoods arrest people of certain races at higher rates in 2015? In order to answer this question, we perform a multiple proportion test over the rate of arrest for each race, and we test to see if there are higher rates of arrest in certain neighborhood. Thus, our null hypothesis for each test is that people of each race are arrested at the same rate in each neighborhood, and the alternate hypothesis is that the proportion of people arrested is different.

The results of all of these tests for 2010 and 2015 are in Appendix A, however the general synopsis of the results is simple: people of almost all races are arrested more in some neighborhoods than others. The proportion test for blacks' arrest rate in 2015 is shown below.

Proportion Test for Blacks arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numBlackArrested out of numBlack
## X-squared = 500.66, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.29474990 0.11192791 0.20984215 0.12511211 0.06599553
```

This test shows that the rate of arrest for black people is not constant across the neighborhoods with a p-value of $p < 2.2 * 10^{-16}$. This is an extremely low number, implying that the different neighborhoods' police arrest black people at higher rates than others. Further, the rate of arrest for black people is highest in Manhattan, where it is 9.82%. This is only one example. See Appendix A for the proportion test for all races' arrest rates in the different cities.

Comparison between 2010 and 2015

Overview

We are mainly interested in how the composition of people arrested changes between 2010 and 2015, espeically in terms of races.

We can first look at some basic information of the two years.

The number of people who got arrested in 2010 is,

```
arst2010=sqf2010$arstmade
sum(arst2010)
```

```
## [1] 40937
```

Out of a population size of

```
length(arst2010)
```

```
## [1] 598645
```

Therefore, the percentage of arrest is around 6.83%

```
sum(arst2010)/length(arst2010)
```

```
## [1] 0.06838276
```

Similarly, for 2015,

```
arst2015=sqf2015$arstmade
sum(arst2015)
```

```
## [1] 3936
```

```
length(arst2015)
```

```
## [1] 22502
```

```
sum(arst2015)/length(arst2015)
```

```
## [1] 0.1749178
```

We have 3936 people who got arrested out of 22502, the arrest rate is around 17.49%

Arrested population for each race

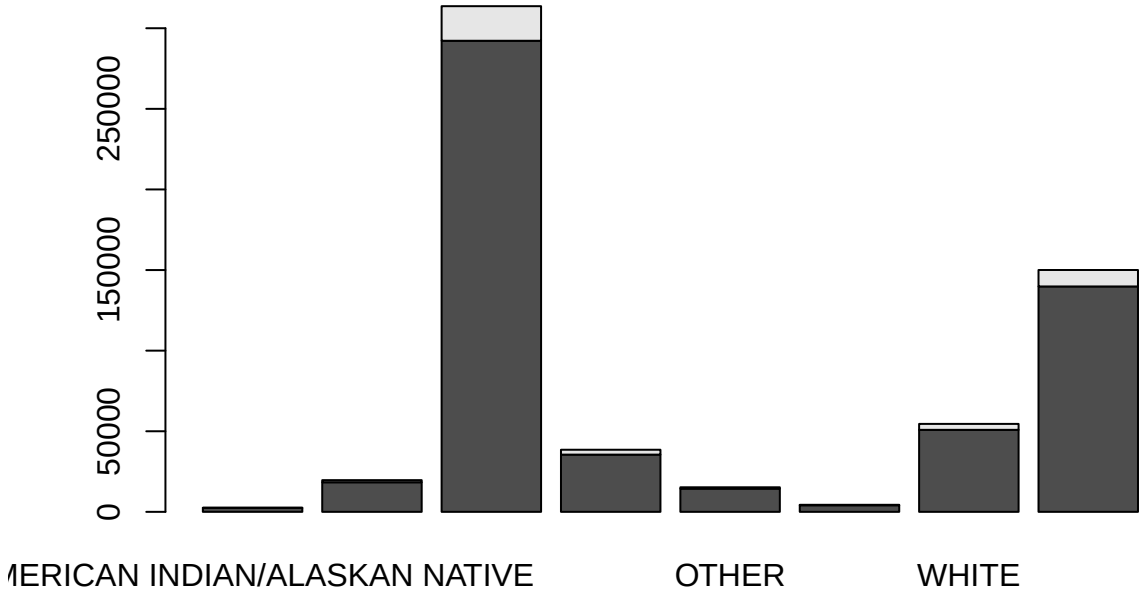
We can now look at the breakdown of arrested population in race.

In 2010, we have,

```
counts2010=table(sqf2010$arstmade,sqf2010$race)
counts2010
```

```
##
##     AMERICAN INDIAN/ALASKAN NATIVE ASIAN/PACIFIC ISLANDER  BLACK
##   0                           2443                  18350 292231
##   1                            126                   1317  21456
##
##     BLACK-HISPANIC  OTHER UNKNOWN  WHITE WHITE-HISPANIC
##   0          35511  14400    4087  50868          139818
##   1           3023    849     280   3681           10205
```

```
barplot(counts2010)
```



we can also convert the data into percentage

```
perc2010=prop.table(counts2010,2)
perc2010
```

```
##
##     AMERICAN INDIAN/ALASKAN NATIVE ASIAN/PACIFIC ISLANDER      BLACK
##   0                     0.95095368             0.93303503 0.93160061
##   1                     0.04904632             0.06696497 0.06839939
##
##     BLACK-HISPANIC      OTHER    UNKNOWN      WHITE WHITE-HISPANIC
##   0     0.92154980 0.94432422 0.93588276 0.93251939     0.93197710
##   1     0.07845020 0.05567578 0.06411724 0.06748061     0.06802290
```

then we have a percentage of arrested for each race,

American Indian/Alaskan Native: 4.89%

Asian/Parcific Islander: 6.69%

Black: 6.83%

Black-hispanic: 7.83%

Other: 5.57%

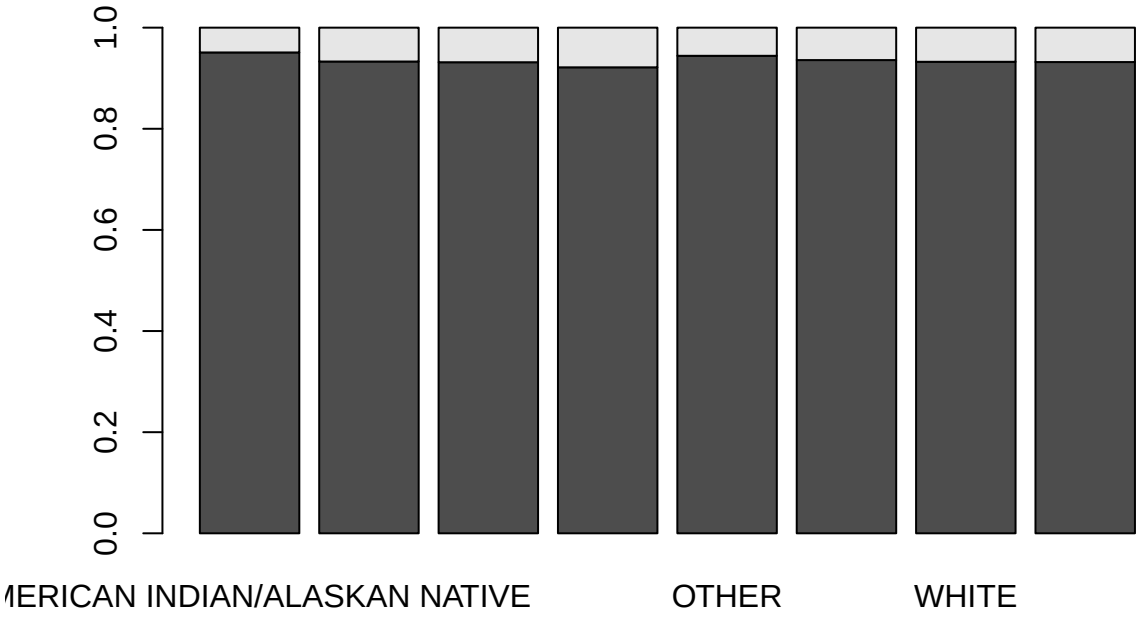
Unknown: 6.39%

White: 6.75%

White-hispanic: 6.80%

which we can also plot on a barplot,

```
barplot(perc2010)
```

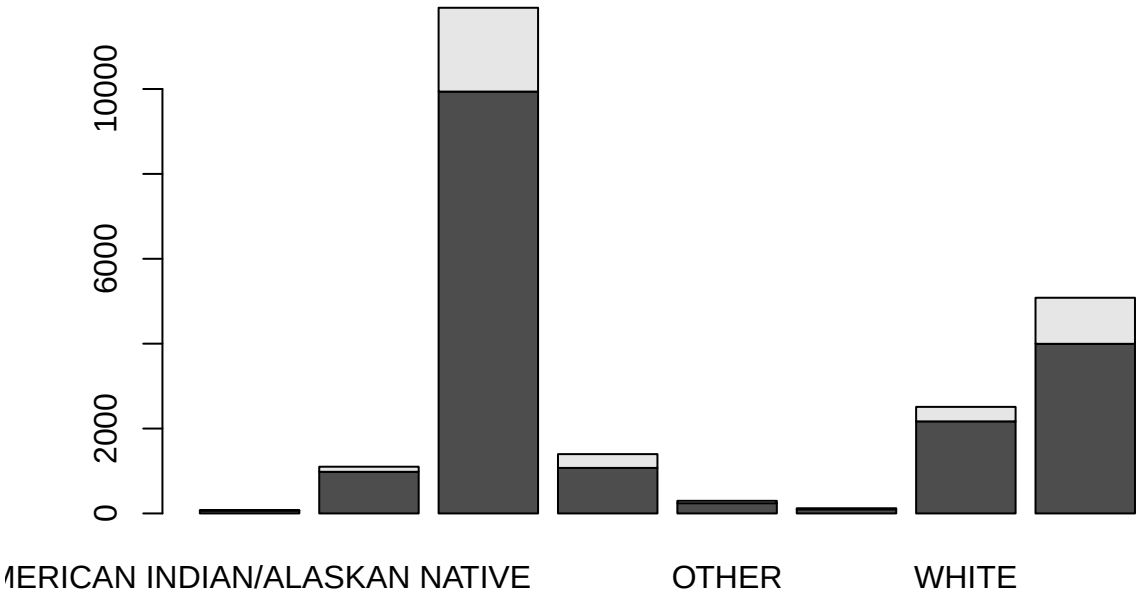


similarly, we calculate the statistics of 2015,

```
counts2015=table(sqf2015$arstmade,sqf2015$race)
counts2015
```

```
##
##     AMERICAN INDIAN/ALASKAN NATIVE ASIAN/PACIFIC ISLANDER BLACK
##   0                             69                    981  9939
##   1                              8                    120  1975
##
##     BLACK-HISPANIC OTHER UNKNOWN WHITE WHITE-HISPANIC
##   0           1073   242      97  2169           3996
##   1            325    55      25   342           1086
```

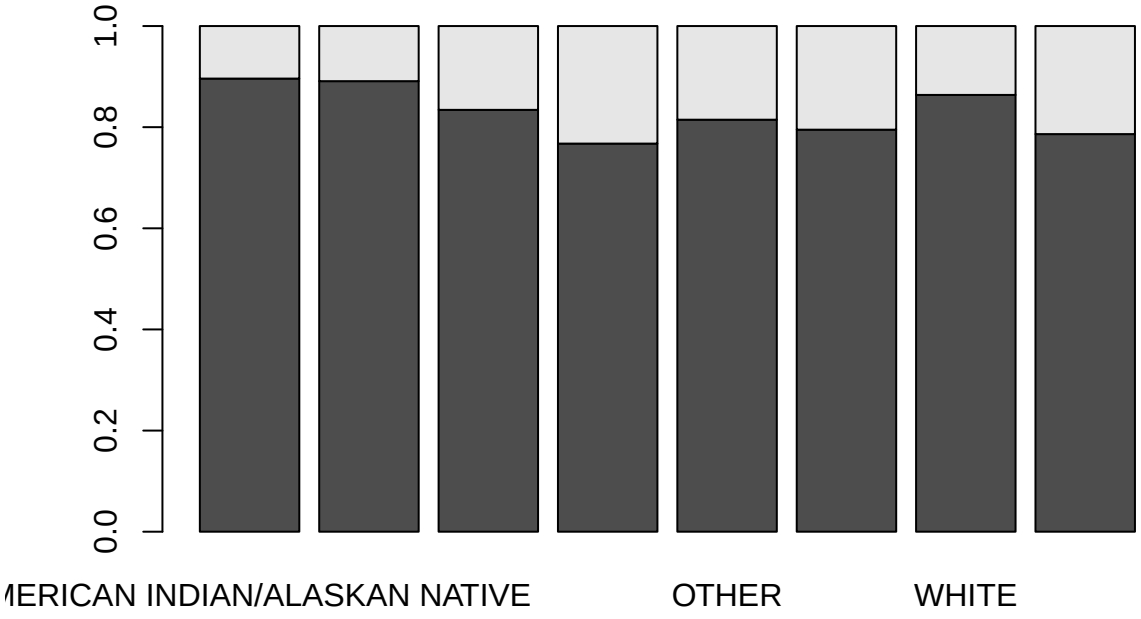
```
barplot(counts2015)
```




```
perc2015=prop.table(counts2015,2)
perc2015
```

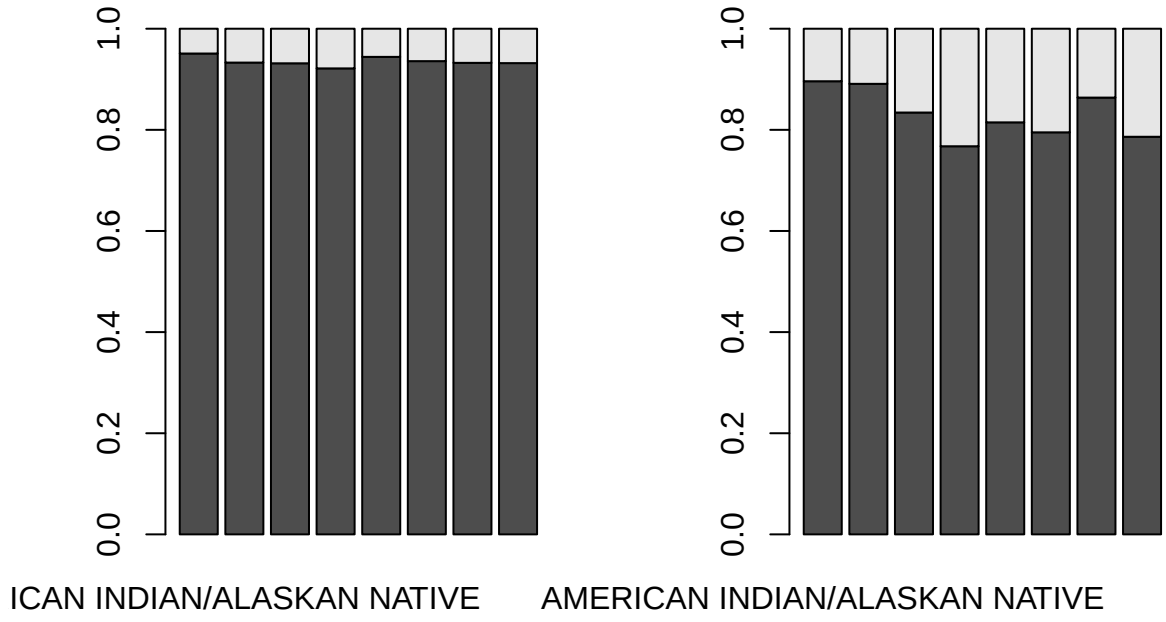
```
##
##     AMERICAN INDIAN/ALASKAN NATIVE ASIAN/PACIFIC ISLANDER     BLACK
##   0                      0.8961039              0.8910082 0.8342286
##   1                      0.1038961              0.1089918 0.1657714
##
##     BLACK-HISPANIC     OTHER   UNKNOWN     WHITE WHITE-HISPANIC
##   0      0.7675250 0.8148148 0.7950820 0.8637993      0.7863046
##   1      0.2324750 0.1851852 0.2049180 0.1362007      0.2136954
```

```
barplot(perc2015)
```



To visualize the difference, we put the two bar plots of percentage together,

```
par(mfrow=c(1,2))
barplot(perc2010)
barplot(perc2015)
```




```
par(mfrow=c(1,2))
```

For year 2010, the race that is most likely to be arrested is black-hispanic, and it stays the same for year 2015.

Conclusion

In this study we have explored the different arrest rates for different races in different neighborhoods. We were able to find that with 95% confidence, for any given race, that race is not arrested at the same rate in each neighborhood (except for American Indians, for whom there was not enough data).

Additionally, we explored the changes that took place for arrest rates between 2010 and 2015. This showed that there was an almost 3x increase in the arrest rate between 2010 and 2015. Also, we performed proportion tests (in Appendix A) that showed the p-values for the proportions of people arrested by race in each neighborhood were far lower than the p-values for the same tests in 2015. A higher p-value implies that there is less certainty that races are being discriminated against in different neighborhoods.

Overall, it is almost certain that people are not being arrested at the same rate in each neighborhood, no matter their race. Some neighborhoods have higher arrest rates, which may point to the inefficacy of the SQF policies in the NYPD between 2010 and 2015.

Appendix A: More Proportion Tests

Proportion Tests for 2010

Below are the proportion tests for the difference in arrest rates for each race in each neighborhood in 2010. The null hypothesis being tested is that $p_1 = p_2 = p_3 = p_4 = p_5$.

Proportion Test for Asians arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numAsianArrested out of numAsian
## X-squared = 70.68, df = 4, p-value = 1.631e-14
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.09902200 0.05696203 0.09760000 0.06273446 0.02325581
```

Proportion Test for American Indians arrested in different neighborhoods

```
## Warning in prop.test(x = numAmIndianArrested, n = numAmIndian): Chi-squared
## approximation may be incorrect
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numAmIndianArrested out of numAmIndian
## X-squared = 13.123, df = 4, p-value = 0.01069
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.04761905 0.04655870 0.08895706 0.04109589 0.04494382
```

Proportion Test for Blacks arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numBlackArrested out of numBlack
## X-squared = 1965.4, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.08028994 0.04719526 0.09821735 0.07360155 0.05810478
```

Proportion Test for Black-Hispanics arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numBlackHispanicArrested out of numBlackHispanic
## X-squared = 154.49, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
```

```
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.08069327 0.06541183 0.10147223 0.05786073 0.04273504
```

Proportion Test for Others arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numOtherArrested out of numOther
## X-squared = 54.936, df = 4, p-value = 3.351e-11
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.05761794 0.03799283 0.07792593 0.04661017 0.04620462
```

Proportion Test for Unknowns arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numUnknownArrested out of numUnknown
## X-squared = 55.534, df = 4, p-value = 2.511e-11
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.06020067 0.04416667 0.11134021 0.06233766 0.02488688
```

Proportion Test for Whites arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numWhiteArrested out of numWhite
## X-squared = 288.9, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.08368708 0.05897962 0.09748762 0.06569989 0.04141671
```

Proportion Test for White-Hispanics arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numWhiteHispArrested out of numWhiteHisp
## X-squared = 689.8, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.07770014 0.06050345 0.09651837 0.05008448 0.06093117
```

Proportion Tests for 2015

Below are the proportion tests for the difference in arrest rates for each race in each neighborhood in 2015. The null hypothesis being tested is that $p_1 = p_2 = p_3 = p_4 = p_5$.

Proportion Test for Asians arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numAsianArrested out of numAsian
## X-squared = 70.68, df = 4, p-value = 1.631e-14
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.09902200 0.05696203 0.09760000 0.06273446 0.02325581
```

Proportion Test for American Indians arrested in different neighborhoods

```
## Warning in prop.test(x = numAmIndianArrested, n = numAmIndian): Chi-squared
## approximation may be incorrect
```

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numAmIndianArrested out of numAmIndian
## X-squared = 2.855, df = 4, p-value = 0.5824
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.08333333 0.00000000 0.18181818 0.10000000 0.25000000
```

Proportion Test for Blacks arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numBlackArrested out of numBlack
## X-squared = 500.66, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.29474990 0.11192791 0.20984215 0.12511211 0.06599553
```

Proportion Test for Black-Hispanics arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numBlackHispanicArrested out of numBlackHispanic
## X-squared = 33.184, df = 4, p-value = 1.095e-06
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5
## 0.2976827 0.1799163 0.2457912 0.1265306 0.1964286
```

Proportion Test for Others arrested in different neighborhoods

```
## Warning in prop.test(x = numOtherArrested, n = numOther): Chi-squared
## approximation may be incorrect
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numOtherArrested out of numOther
## X-squared = 32.424, df = 4, p-value = 1.567e-06
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.37974684 0.11956522 0.21276596 0.06060606 0.00000000
```

Proportion Test for Unknown Race arrested in different neighborhoods

```
## Warning in prop.test(x = numUnknownArrested, n = numUnknown): Chi-squared
## approximation may be incorrect
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numUnknownArrested out of numUnknown
## X-squared = 7.3658, df = 4, p-value = 0.1178
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5
## 0.3913043 0.1153846 0.2093023 0.1428571 0.0000000
```

Proportion Test for Whites arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numWhiteArrested out of numWhite
## X-squared = 54.371, df = 4, p-value = 4.4e-11
```

```
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.23267327 0.14102564 0.20854271 0.10984848 0.07474747
```

Proportion Test for White-Hispanics arrested in different neighborhoods

```
##
##  5-sample test for equality of proportions without continuity
##  correction
##
## data:  numWhiteHispArrested out of numWhiteHisp
## X-squared = 210.47, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5
## 0.32751397 0.18597561 0.23977273 0.13728129 0.06333333
```

Appendix B: Neural Net for Predicting