Final Project: New York City Stop-Question-Frisk Group Write Up

Xingyao Chen, Sonia Sehra, Dave Makhervaks, Jenna Kahn Section 5

Due December 13, 2016

Background provided by instructors:

In stop-question-frisk, a police officer is authorized to stop a pedestrian, question them, and then frisk their body searching for contraband items such as weapons or drugs. The motivation for the policy was to prevent crimes from happening in the first place, though recent studies by the New York Civil Liberties Union suggest the policy did not achieve noticeable reductions in crime. For example, NYCLU estimates that guns were found in fewer than 0.2% of stops. Moreover, the policy was found to be discriminatory because Blacks and Latinos were stopped disproportionately more than their participation in crime would suggest. (Since 2014, new restrictions have limited the use of SQF, and the numbers of SQF incidents have decreased precipitously.)

Data from New York City's stop-question-frisk (SQF) program is publicly available online. Whenever a person is stopped under SQF, the officer is required to fill out a form with information about the stop. Each year, the data is compiled and released by the city.

Required Questions:

i. Because all of the data provided are reported by the police officers who stopped the suspects, it is possible that there is misreported data. For example, an officer may misidentify an Asian/Pacific Islander as Hispanic, or vice-versa. We would like to explore what effect such misidentifications can have on the collected data. To simplify our analysis, we will only consider the subset of the data for which suspects were identified as Asian/Pacific Island or Hispanic. Suppose that a Hispanic person has a 95% chance of correctly being identified as Hispanic, and otherwise is misidentified as Asian/Pacific Islander, and an Asian/Pacific Islander has a 95% chance of correctly being identified as Asian/Pacific Islander, and otherwise is misidentified as Hispanic. Using this subset of the 2010 data, determine the probability that a stopped suspect is Hispanic and the probability that a stopped suspect is Asian/Pacific Islander. Use this to determine the probability that someone who is identified by the officer as Hispanic is actually Hispanic and the probability that someone who is identified as Asian/Pacific Islander is actually Asian/Pacific Islander. What does this say about how we utilize the `race` column of the data?

```
sqf2010=read.csv("2010_sqf_m35.csv")
sqf2015=read.csv("2015_sqf_m35.csv")

sqf2010_as=sqf2010[c(sqf2010$race=="ASIAN/PACIFIC ISLANDER"),]
sqf2010_wh=sqf2010[c(sqf2010$race=="WHITE-HISPANIC"),]
sqf2010_bh=sqf2010[c(sqf2010$race=="BLACK-HISPANIC"),]

sqf2010_sub=rbind(sqf2010_as,sqf2010_wh , sqf2010_bh)
races=sqf2010_sub$race

LH=(38377+149532)/length(races)
LA=(19630)/length(races)
```

LH=suspect is labled as Hispanic

LA=suspect is labled as Asian

A=suspect is Asian

H=suspect is Hispanic

$$
\begin{aligned}
P(LH) &= P(LH|H)P(H) + P(LH|A)P(A) \\
P(LA) &= P(LA|H)P(H) + P(LA|A)P(A) \\
0.9 &= 0.95P(H) + 0.05P(A) \\
0.09 &= 0.05P(H) + 0.95 * P(A) \\
P(H) &= 0.95067 \\
P(A) &= 0.04933
\end{aligned}
$$

The conditional probabiliies

$$
\begin{aligned}
P(H|LH) &= \frac{P(LH|H)P(H)}{P(LH)} \\
P(A|LA) &= \frac{P(LA|A)P(A)}{P(LA)} \\
P(H|LH) &= 0.997 \\
P(A|LA) &= 0.4964
\end{aligned}
$$

ii. Compare the rates at which suspects stopped in 2010 were frisked, broken down by race. Are the differences in rates between the various groups statistically significant? Which borough(s) had the largest differences? The smallest?

Solution

```
#split the frisked data by race
splittedByRace=split(sqf2010$frisked, sqf2010$race)
#forloop to find all the p-values
results=data.frame()
for (i in 1:length(splittedByRace)){
  for(j in 1:length(splittedByRace)){
    var1=names(splittedByRace)[i]
    var2=names(splittedByRace)[j]
    pval=t.test(splittedByRace[[i]], splittedByRace[[j]])$p.value
    row=data.frame(var1, var2, pval)
    results=rbind(results, row)
  }
}

#find the groups with the largest differences
results[results$pval==min(results$pval),]
```

```
##              var1           var2 pval
## 23          BLACK          WHITE    0
## 31 BLACK-HISPANIC          WHITE    0
## 51          WHITE          BLACK    0
## 52          WHITE BLACK-HISPANIC    0
## 56          WHITE WHITE-HISPANIC    0
## 63 WHITE-HISPANIC          WHITE    0
```

```
#find the groups with the smalled differences (top 60% of pvals excluding identical co)
results_sub=results[results$pval<1,]
results_sub[results_sub$pval>0.4*max(results_sub$pval),]
```

```
##                              var1                            var2      pval
## 2  AMERICAN INDIAN/ALASKAN NATIVE         ASIAN/PACIFIC ISLANDER 0.4618376
## 9           ASIAN/PACIFIC ISLANDER AMERICAN INDIAN/ALASKAN NATIVE 0.4618376
## 47                         UNKNOWN                           WHITE 0.2230251
## 54                           WHITE                         UNKNOWN 0.2230251
```

The largest difference is between WHITE and WHITE-HISPANIC, WHITE vs BLACK, and WHITE vs BLACK-HISPANIC. The smallest difference is AMERICAN INDIAN/ALASKAN NATIVE vs ASIAN/PACIFIC ISLANDER, and UNKNOWN vs WHITE.

```
#split the frisked data by city
splittedByCity=split(sqf2010$frisked, sqf2010$city)
#forloop to find all the p-values
results=data.frame()
for (i in 1:length(splittedByCity)){
  for(j in 1:length(splittedByCity)){
    var1=names(splittedByCity)[i]
    var2=names(splittedByCity)[j]
    pval=t.test(splittedByCity[[i]], splittedByCity[[j]])$p.value
    row=data.frame(var1, var2, pval)
    results=rbind(results, row)
  }
}

#find the groups with the largest differences
results[results$pval==min(results$pval),]
```

```
##          var1      var2 pval
## 9       BRONX  BROOKLYN    0
## 10      BRONX MANHATTAN    0
## 12      BRONX STATEN IS    0
## 14  BROOKLYN     BRONX    0
## 17  BROOKLYN    QUEENS    0
## 20 MANHATTAN     BRONX    0
## 23 MANHATTAN    QUEENS    0
## 27    QUEENS  BROOKLYN    0
## 28    QUEENS MANHATTAN    0
## 30    QUEENS STATEN IS    0
## 32 STATEN IS     BRONX    0
## 35 STATEN IS    QUEENS    0
```

```
#find the groups with the smalled differences (top 60% of pvals excluding identical co)
results_sub=results[results$pval<1,]
results_sub[results_sub$pval>0.4*max(results_sub$pval),]
```

```
##          var1      var2      pval
## 3              BROOKLYN 0.7979995
## 4             MANHATTAN 0.7924557
## 6             STATEN IS 0.5704495
## 13  BROOKLYN            0.7979995
## 16  BROOKLYN MANHATTAN 0.7958508
## 19 MANHATTAN            0.7924557
## 21 MANHATTAN  BROOKLYN 0.7958508
## 31 STATEN IS            0.5704495
```

The largest difference is between BRONX and BROOKLYN, BRONX and BROOKLYN, BRONX and

MANHATTAN, BRONX and STATEN IS, BROOKLYN and QUEENS, MANHATTAN and QUEENS, and QUEENS and STATEN IS.

The smallest difference is between BROOKLYN and MANHATTAN, UNKNOW and BROOKLYN, and UNKNOW and STATEN IS, UNKNOW and MANHATTAN.

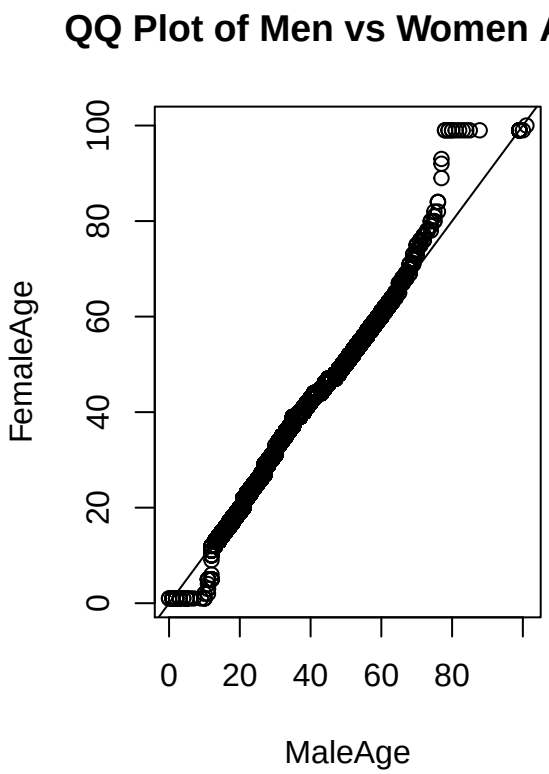
iii. Compare the distribution of ages for male suspects in 2010 with the distribution of ages for female suspects in 2010. Use `qqnorm` to determine if they are normally distributed, and compare them with each other by using `qqplot`. Are the age distributions the same? Compare the age distribution for male suspects to that of the entire population of suspects that were stopped in 2010. Type `?qqplot` for help on how to use the command. Note that some ages are reported as 0 or 999 if the officer did not know the age, so you may want to throw out the extraneous data first.

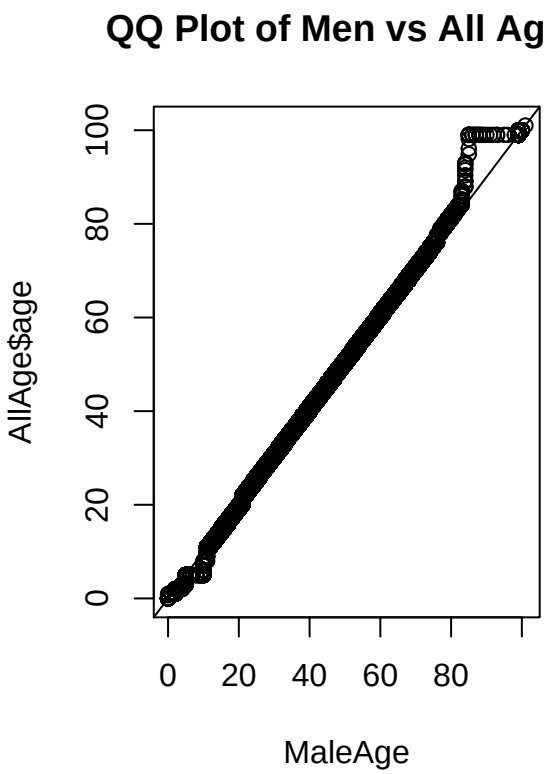
Solution

```
#throw out outliers
AllAge <- sqf2010[sqf2010$age <110,]

#subset F and M
MaleAge <- AllAge$age[AllAge$sex == "M"]
FemaleAge <- AllAge$age[AllAge$sex == "F"]

par(mfrow=c(1,2))
#qqnorm plots for normality
qqnorm(MaleAge, main = "MaleAge QQ Plot without outliers")
qqline(MaleAge)
qqnorm(FemaleAge, main = "FemaleAge QQ Plot without outliers")
qqline(FemaleAge)
```


```
qqplot(MaleAge,FemaleAge, main = "QQ Plot of Men vs Women Ages")
abline(0,1)
```

```
qqplot(MaleAge,AllAge$age , main= "QQ Plot of Men vs All Ages")
abline(0,1)
```





The distribution of ages for male suspects with the outliers included is somewhat normal in the middle, but it is pretty messy and is not a very good QQ plot (not a straight line). The same follows for the distribution of ages for female suspects with the outliers included in the data. However, when the outliers are removed, then the QQ-Plot for both the distribution of both male and female ages are reasonably normal. We know this, because the QQ plots without the outliers are reasonably straight linear excluding the sides of the graph. Creating a QQ Plot of Male ages vs Female ages, we get a very linear line, showing that the male and female distributions are in fact, very similar! Comparing the male distribution of ages against the age distribution of the entire population also yields a linear plot, showing that the distributions are very similar!

The distributions between Females and Males are approximately the same, because the `qqplot` results yielded points that fall pretty closely on the $y = x$ line.

iv. Find the probability, with a 95% confidence interval, that a suspect was frisked (a) for the entire population in 2015, and (b) for suspects in 2015 who refused to provide identification, and determine whether suspects who refused to provide identification had a different probability of being frisked than the population at large.

```
frisked=sqf2015$frisked
mu=mean(frisked)
SEM=1.96*sd(frisked)/sqrt(length(frisked))
mu;SEM
```

```
## [1] 0.6761955
```

```
## [1] 0.006105832
```

(a) The probability, with a 95% confidence interval, that a suspect was frisked for the entire population in 2015 is 0.6761955 ± 0.0061058.

```
frisked_id=frisked[sqf2015$typeofid=="REFUSED"]
mu1=mean(frisked_id)
SEM1=1.96*sd(frisked_id)/sqrt(length(frisked_id))
mu1;SEM1
```

```
## [1] 0.6103286
```

```
## [1] 0.03784224
```

(b) The probability, with a 95% confidence interval, that a suspect who refused to provide ID was frisked is 0.6103286 ± 0.0378422.

A two sample t-test can be used to determine whether these rates significantly differ.

```
t.test(x = sqf2015$frisked, y = sqf2015$frisked[sqf2015$typeofid == "REFUSED"], conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  sqf2015$frisked and sqf2015$frisked[sqf2015$typeofid == "REFUSED"]
## t = 3.368, df = 671.64, p-value = 0.0008005
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02746674 0.10426706
## sample estimates:
## mean of x mean of y
## 0.6761955 0.6103286
```

The t-test suggests there is a significant difference in the probability a suspect in general will be frisked vs. the probability a suspect who refuses to provide ID will be frisked ($p < .05$).

v. For the 2010 data, decide which of the following binary factors: `arstmade`, `searched`, `inside`, `sumissue`, `frisked`, `weap`, `contrabn` `radio`, `pf` had a significant effect on the length of the stop (`perstop`) by using linear regression. Make sure to check your residuals for normality, and apply an appropriate transformation to `perstop` or remove outlier points if it does not look normal (see your notes from Lecture 11 to review how to do this). Note that `perstop` is a discrete variable, so you are looking for an approximately normal distribution for the residuals. Consider the p-values for the coefficients and the R^2 value for your regression model. What do they indicate about how the factors affect the length of the stop? Recall that $R^2 = \frac{SSR}{SST}$. How much of the variability in `perstop` is due to the explanatory variables you have selected? Why does this make sense?

```
sqf2010=sqf2010[sqf2010$perstop<60,]

lmdata=sqf2010[, c("arstmade", "searched", "inside", "sumissue", "frisked", "weap",
                   "contrabn", "radio", "pf")]
mod=list()
lmout=data.frame()
for (i in names(lmdata)){
  mod[[i]]=summary(lm(sqf2010$perstop~as.factor(lmdata[,i])))
  row=data.frame(i, mod[[i]]$r.squared)
  lmout=rbind(lmout, row)
}
lmout
```

```
##          i mod..i...r.squared
## 1 arstmade       0.0109950884
## 2 searched       0.0116435327
## 3   inside       0.0005851448
```

```
## 4 sumissue       0.0142917592
## 5  frisked       0.0034338813
## 6     weap       0.0010134042
## 7 contrabn       0.0021368759
## 8    radio       0.0192165505
## 9       pf       0.0071013775
```

The R^2 values indicate that the linear models are a poor fit for the data. All of the p-values are below 0.05. Not much variability is explained by the selected variables.

To answer this question, it is useful to understand how to interpret a regression model involving indicator variables. Suppose you have an indicator variable X that equals 1 when a condition is true and 0 otherwise, and you fit the following regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Note that when $X = 0$, meaning that the condition is false, $Y = \beta_0 + \epsilon$, and when $X = 1$, meaning that the condition is true, $Y = \beta_0 + \beta_1 + \epsilon$. Therefore, β_1, the coefficient on X in the regression model, gives the *average effect* that the condition has on the outcome. That is β_1 equals the mean difference in the response when the condition is true compared to when it is false. Our hypothesis is constructed in the same way as always: $H_0 : \beta_1 = 0$ is the null hypothesis asserting that the condition has no effect on Y. $H_a : \beta_1 \neq 0$ is the alternative hypothesis asserting that the condition has some effect on Y. If the p-value on β_1 after we fit the regression model is smaller than our significance level α, then we reject H_0 and conclude that the condition being true has a statistically significant effect on the value of Y.

Team-Chosen Questions:

Introduction:

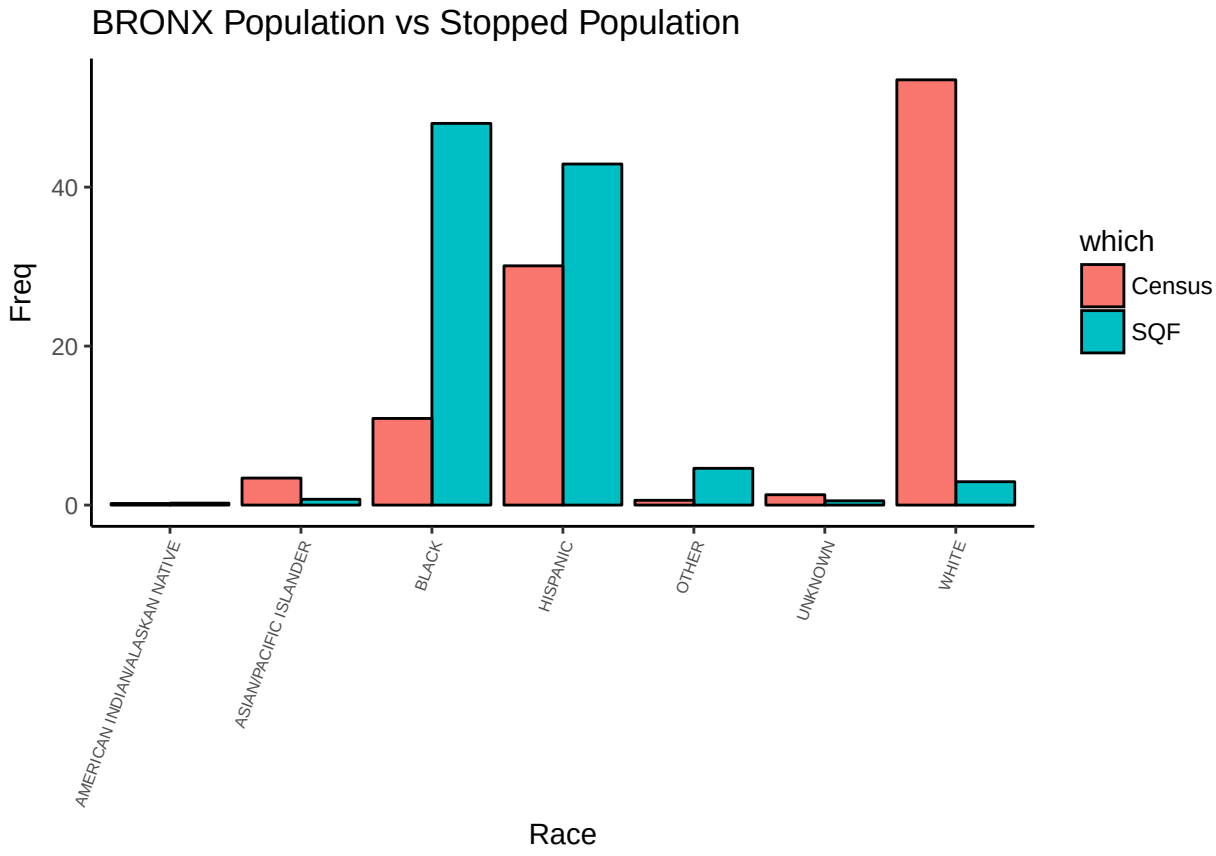
We examined whether people of certain races are more likely to be stopped and/or frisked than would be expected based on the demographic makeup of each borough. This question is important because stop and frisk has faced numerous accusations of racial profiling. If we find that certain racial groups are disproportionately likely to be stopped that, then our findings give support to the claim that stop and frisk unfairly targets those groups. To answer our question, we examined two sets of data. One is the 2010 US government census data for each of the five New York boroughs. We used this find the racial breakdown of population of each borough. The other data set is a collection of police filings from all of the stop-and-frisk incidents in New York in 2010. Each filing contains information on the race of the stopped individual, where the stop occurred, and whether the individual was frisked during the encounter. We used this data set to the relative frequency with which people of each race were likely to be stopped and frisked. We conducted graphical and statistical analyses to determine whether black, Hispanic, and white people were disproportionately likely to be stopped and frisked

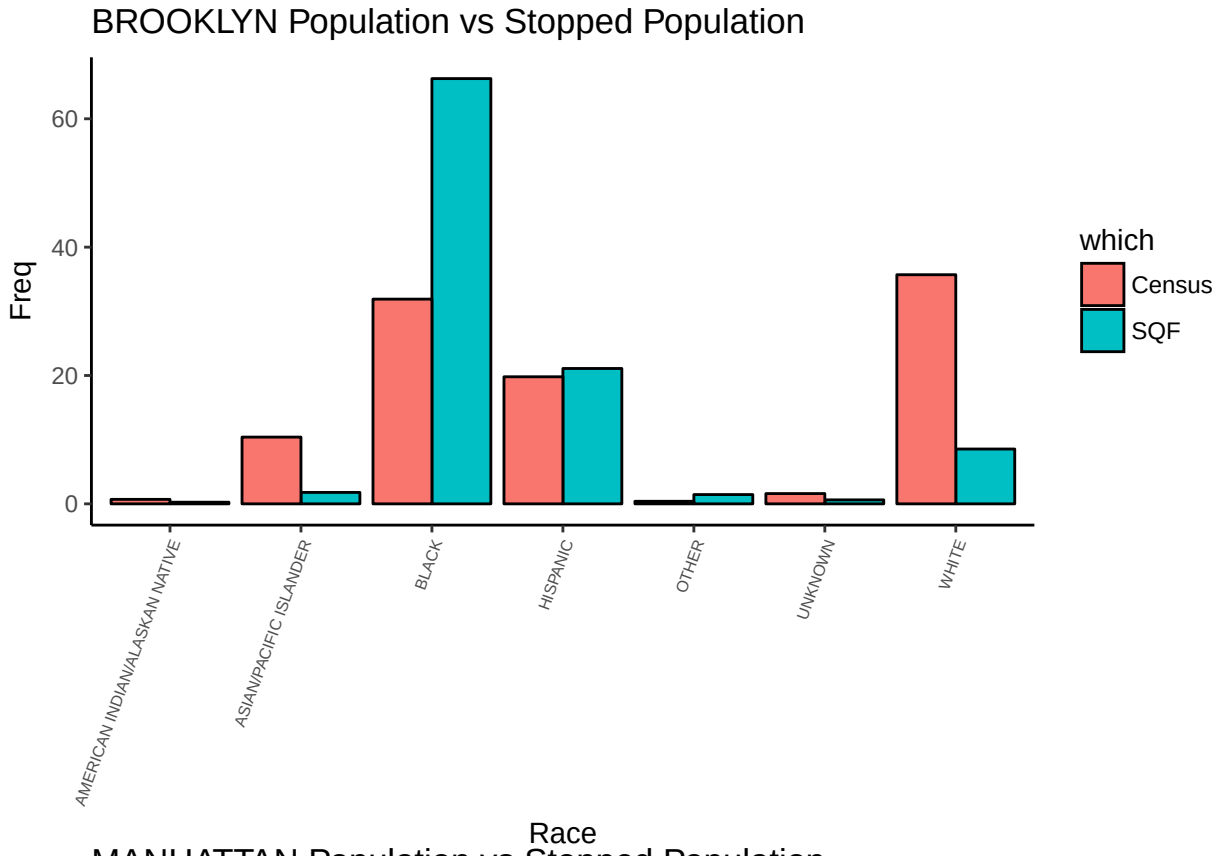
Question 1: Are people stopped in proportion to the demographic characteristics of the area?

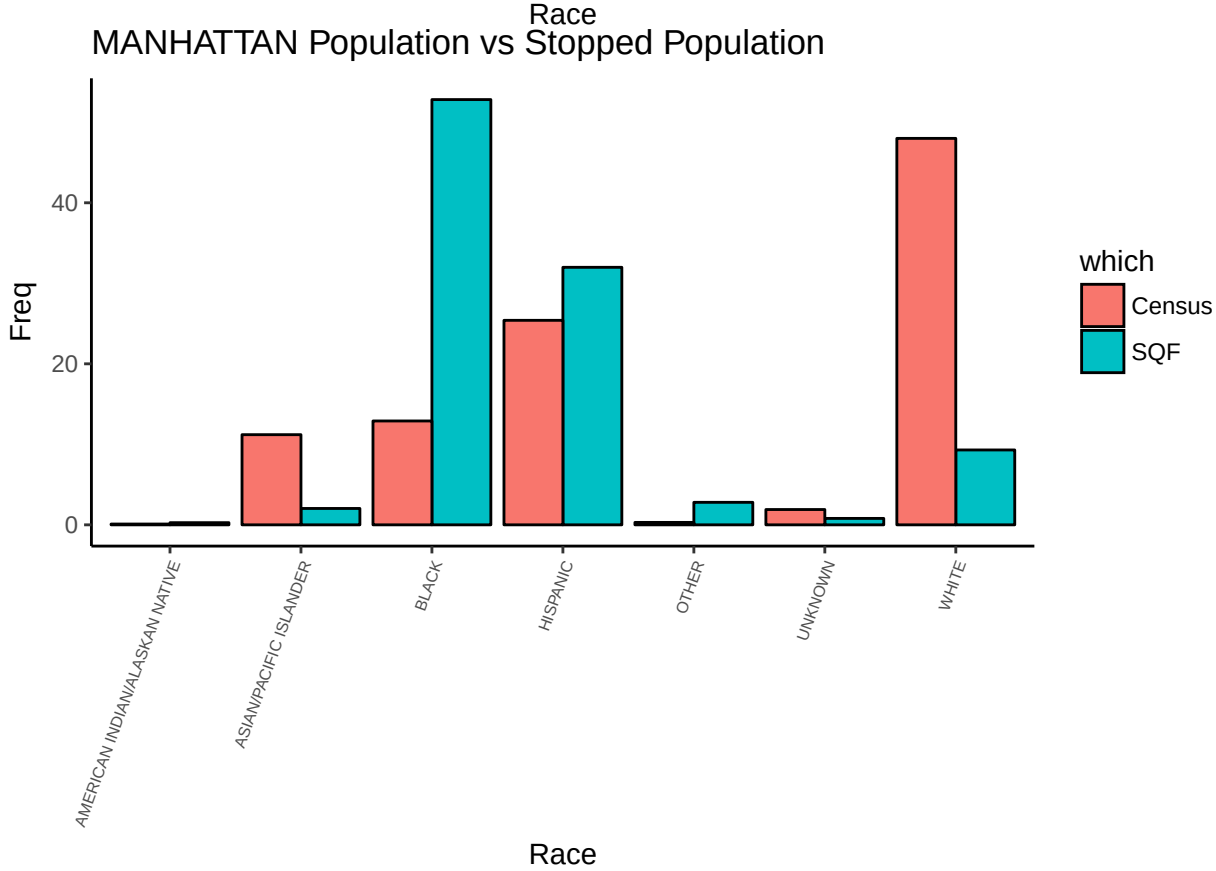
```
sqf2010$race=gsub("BLACK-","",sqf2010$race)
sqf2010$race=gsub("WHITE-","",sqf2010$race)
sqf2010_by_borough=split(sqf2010, sqf2010$city)
races=unique(sqf2010$race)

count=table(sqf2010_by_borough$BRONX$race)
sqfBronx=100*count/sum(count)
sqfBronx=data.frame(sqfBronx)
sqfBronx$which="SQF"
names(sqfBronx)[1]="Race"

count=table(sqf2010_by_borough$BROOKLYN$race)
sqfBrooklyn=100*count/sum(count)
sqfBrooklyn=data.frame(sqfBrooklyn)
```

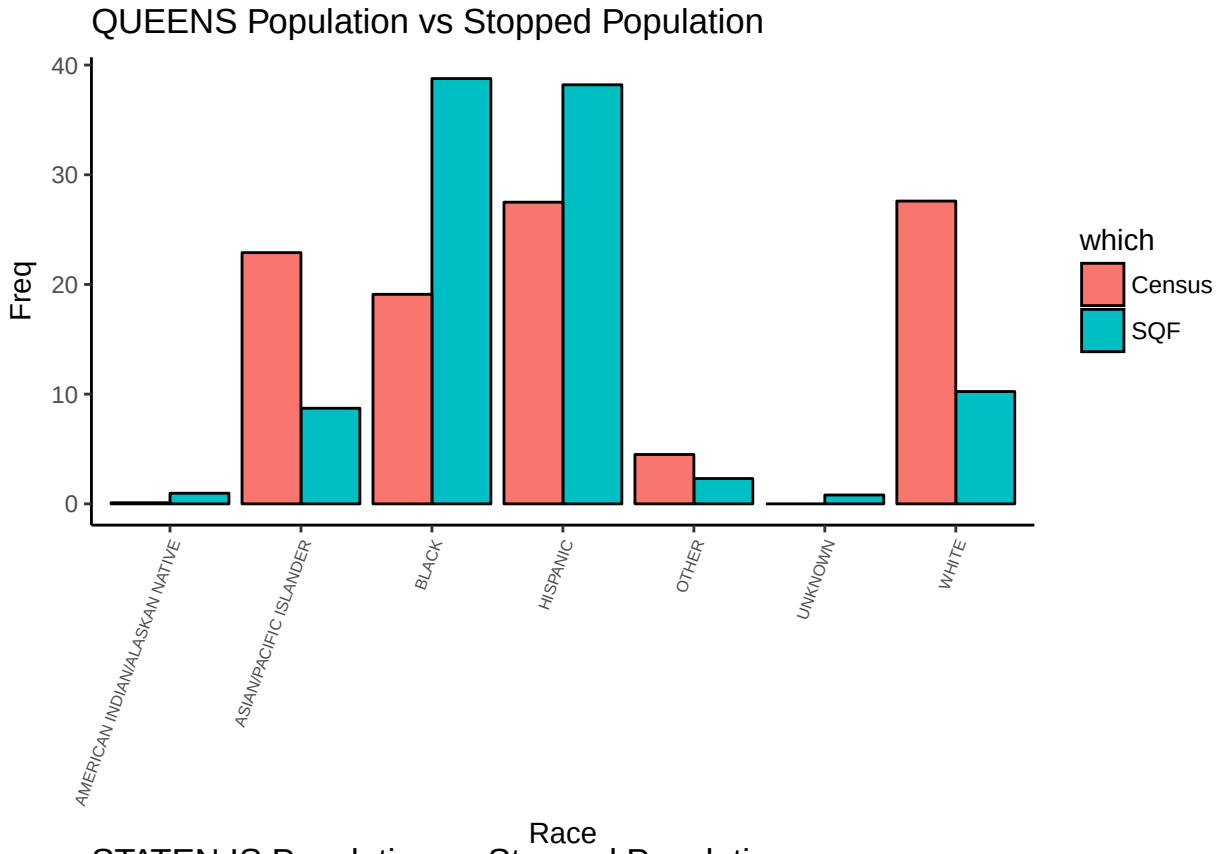
```
sqfBrooklyn$which="SQF"
names(sqfBrooklyn)[1]="Race"

count=table(sqf2010_by_borough$MANHATTAN$race)
sqfMan=100*count/sum(count)
sqfMan=data.frame(sqfMan)
sqfMan$which="SQF"
names(sqfMan)[1]="Race"

count=table(sqf2010_by_borough$QUEENS$race)
sqfQueens=100*count/sum(count)
sqfQueens=data.frame(sqfQueens)
sqfQueens$which="SQF"
names(sqfQueens)[1]="Race"

count=table(sqf2010_by_borough$`STATEN IS`$race)
sqfStanten=100*count/sum(count)
sqfStanten=data.frame(sqfStanten)
sqfStanten$which="SQF"
names(sqfStanten)[1]="Race"



bronx=c(10.9,30.1, 53.5, 0.6, 3.4,  1.3,0.2)
bronxData=data.frame(Race=races, Freq=bronx, which="Census")

staten=c(9.5, 17.3, 64.0, 0.2, 7.4, 2.6 ,0)
statenData=data.frame(Race=races, Freq=staten, which="Census")

queens=c(19.1, 27.5, 27.6, 4.5, 22.9, 0, 0.1)
queensData=data.frame(Race=races, Freq=queens, which="Census")

man=c(12.9, 25.4, 48.0, 0.3, 11.2, 1.9, 0.1)
manData=data.frame(Race=races, Freq=man, which="Census")


brook=c(31.9, 19.8, 35.7, 0.4, 10.4, 1.6, 0.7)
brookData=data.frame(Race=races, Freq=brook, which="Census")

sqf_vs_staten=rbind(sqfStanten, statenData)
sqf_vs_bronx=rbind(sqfBronx, bronxData)
sqf_vs_queens=rbind(sqfQueens, queensData)
sqf_vs_man=rbind(sqfMan, manData)
sqf_vs_brook=rbind(sqfBrooklyn, brookData)


bigL=list(Bronx=sqf_vs_bronx,Brooklyn=sqf_vs_brook,Manhattan=sqf_vs_man, Queens=sqf_vs_queens,StatenIsl
cities=names(sqf2010_by_borough)[-1]
c=1
for (i in bigL){
  city=cities[c]
  p=ggplot(i, aes(x=Race, y=Freq, fill=which))+
    geom_bar(stat="identity", position=position_dodge(), color="Black")+
```

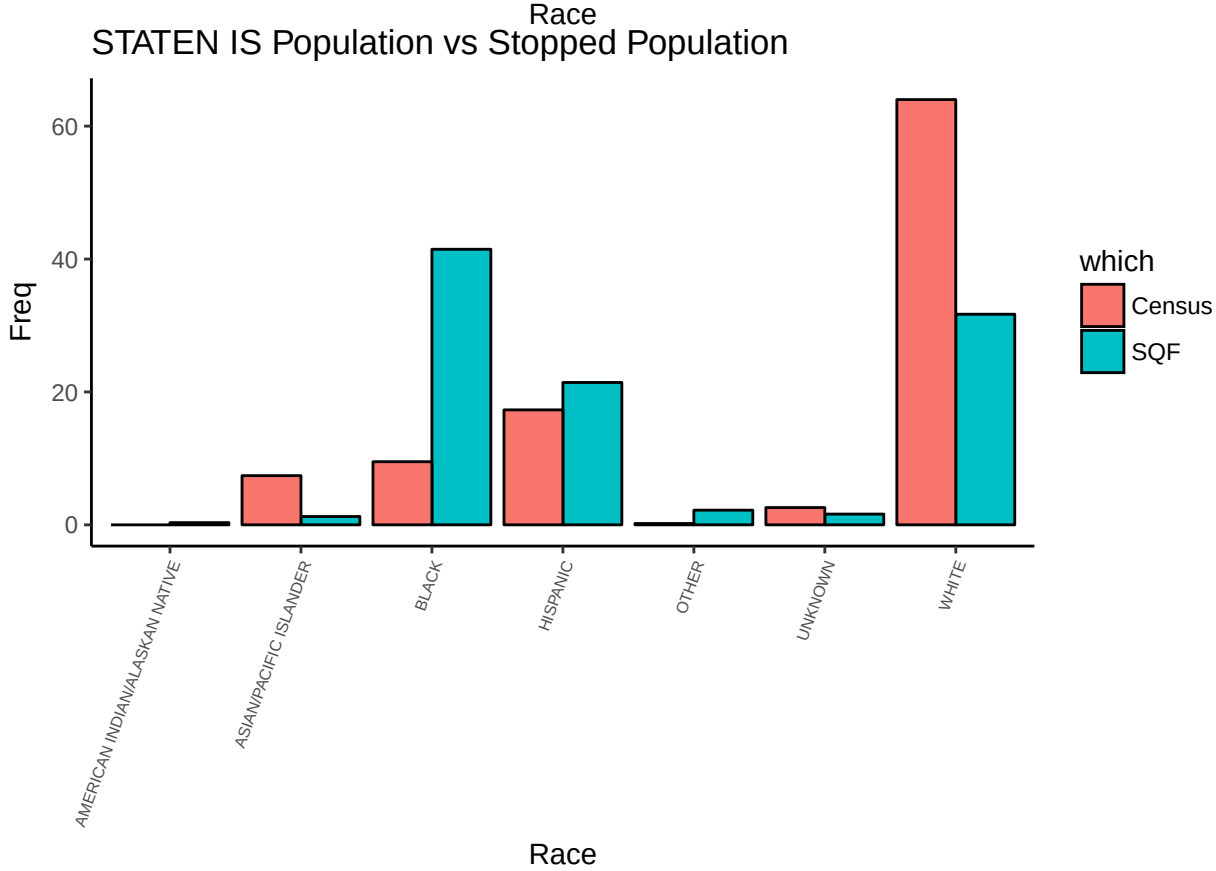
```
  theme_classic()+
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = rel(0.7) ))+
  ggtitle(paste(city,"Population vs Stopped Population"))
plot(p)
c=c+1
}
```




BROOKLYN Population vs Stopped Population
Freq
60
40
20
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
HISPANIC
OTHER
UNKNOWN
WHITE
Race
which
Census
SQF


MANHATTAN Population vs Stopped Population
Freq
40
20
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
HISPANIC
OTHER
UNKNOWN
WHITE
Race
which
Census
SQF





We can use `prop.test` to test for the significance of the difference.

```r
pop_staten=c(nrow(sqf2010_by_borough$`STATEN IS`), 468730)
pop_bronx=c(nrow(sqf2010_by_borough$BRONX),1385108)
pop_man=c(nrow(sqf2010_by_borough$MANHATTAN),1585873)
pop_queen=c(nrow(sqf2010_by_borough$QUEENS),223072)
pop_brook=c(nrow(sqf2010_by_borough$BROOKLYN),2504710)
pop_vectors=rbind(pop_bronx,pop_brook, pop_man, pop_queen , pop_staten)

pData_black=data.frame()
for(i in 1:length(bigL)){
  city_data=bigL[[i]]
  blk=city_data[city_data$Race=="BLACK",][,2]*pop_vectors[i,]/100
  table=cbind(blk, pop_vectors[i,])
  rownames(table)=c("SQF", "Census")
  pval=prop.test(table, alternative="greater",correct=F)
  pval=pval$p.value
  pData_black=rbind(pData_black, data.frame(city=cities[i], pval=pval))
}



pData_his=data.frame()
for(i in 1:length(bigL)){
  city_data=bigL[[i]]
  blk=city_data[city_data$Race=="HISPANIC",][,2]*pop_vectors[i,]/100
  table=cbind(blk, pop_vectors[i,])
  rownames(table)=c("SQF", "Census")
  pval=prop.test(table, alternative="greater",correct=F)
  pval=pval$p.value
  pData_his=rbind(pData_his, data.frame(city=cities[i], pval=pval))
}


pData_white=data.frame()
for(i in 1:length(bigL)){
  city_data=bigL[[i]]
  blk=city_data[city_data$Race=="WHITE",][,2]*pop_vectors[i,]/100
  table=cbind(blk, pop_vectors[i,])
  rownames(table)=c("SQF", "Census")
  pval=prop.test(table, alternative="less", correct=F)
  pval=pval$p.value
  pData_white=rbind(pData_white, data.frame(city=cities[i], pval=pval))
}

#hypothesis test to test whether the the proportion of stopped who are black is greater than the propor
pData_black
```

```
##        city pval
## 1     BRONX    0
## 2  BROOKLYN    0
## 3 MANHATTAN    0
## 4    QUEENS    0
## 5 STATEN IS    0
```

```
#hypothesis test to test whether the the proportion of stopped who are hispanic is greater than the pro
pData_his
```

```
##        city         pval
## 1     BRONX 0.000000e+00
## 2  BROOKLYN 2.115799e-29
## 3 MANHATTAN 0.000000e+00
## 4    QUEENS 0.000000e+00
## 5 STATEN IS 1.869752e-47
```

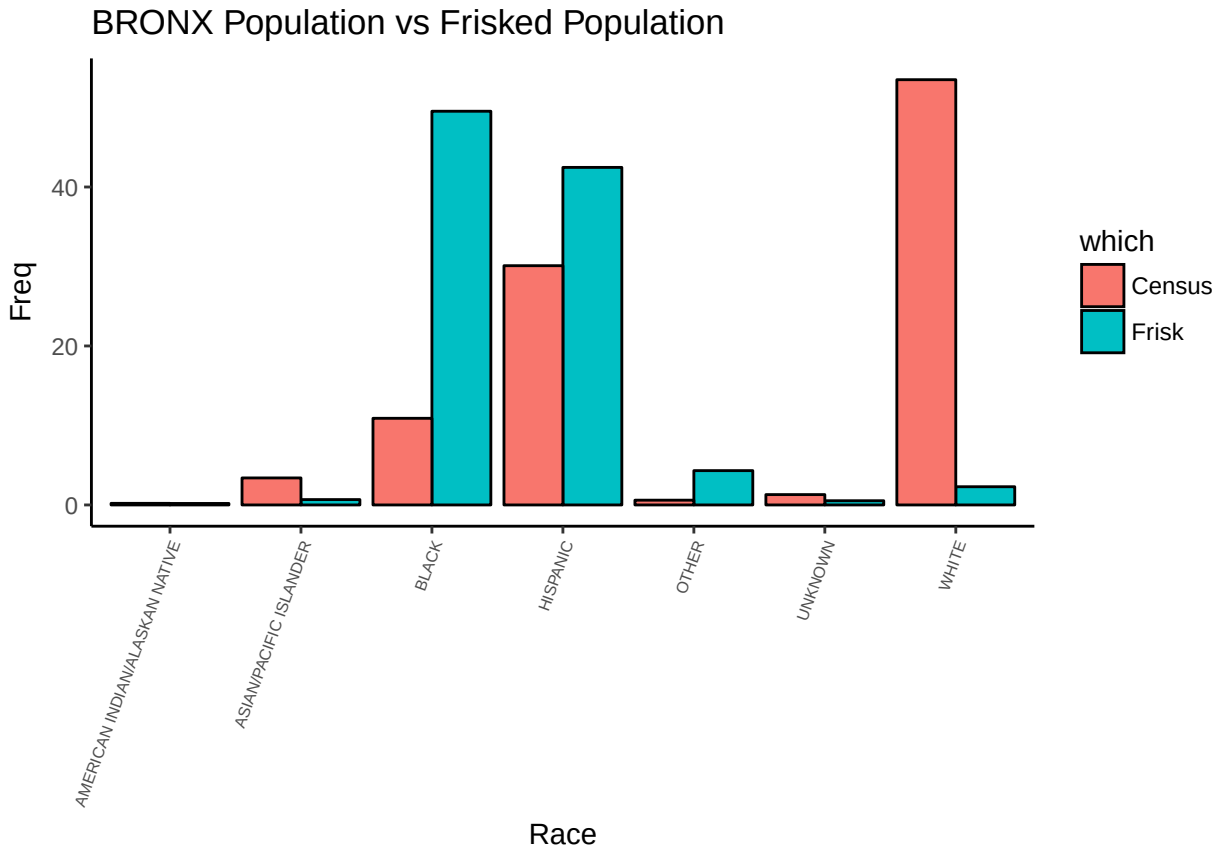
```
#hypothesis test to test whether the the proportion of stopped who are black is greater than the propor
pData_white
```

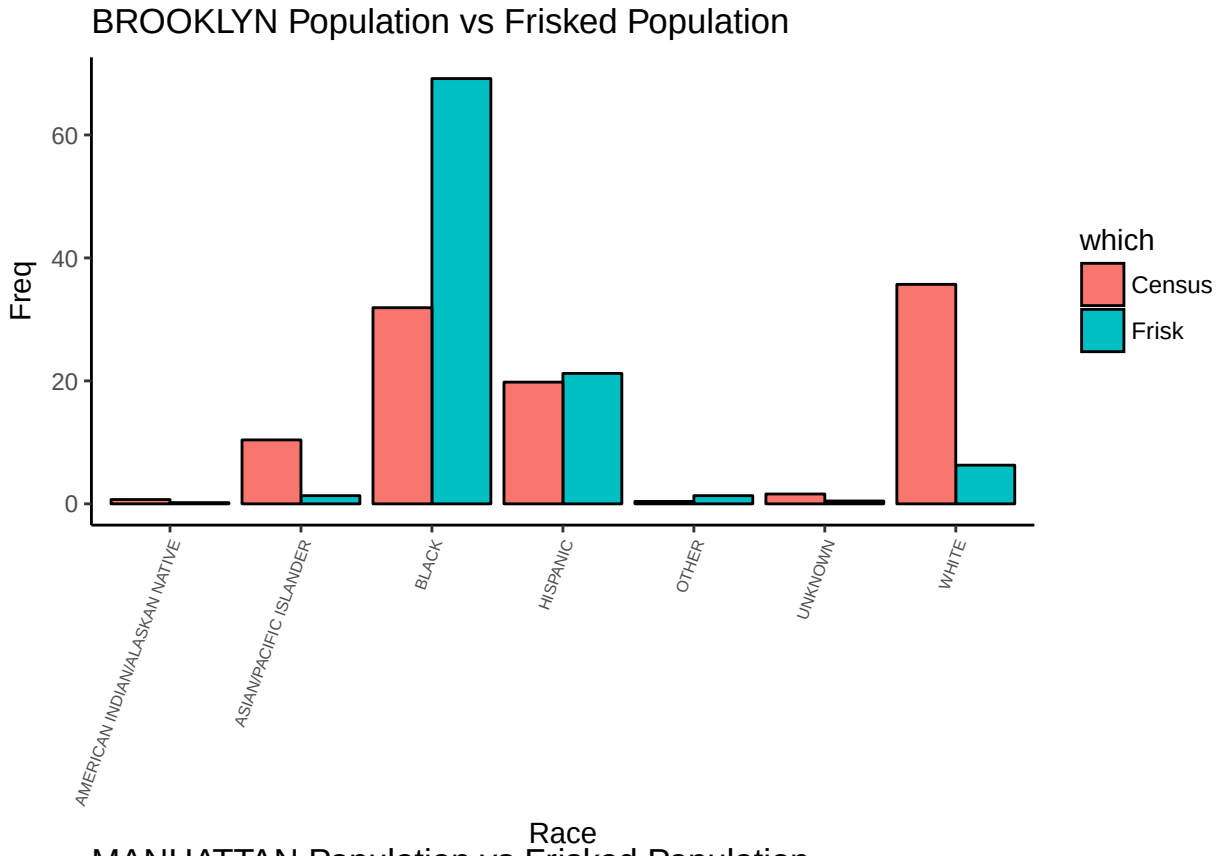
```
##        city pval
## 1     BRONX    0
## 2  BROOKLYN    0
## 3 MANHATTAN    0
## 4    QUEENS    0
## 5 STATEN IS    0
```

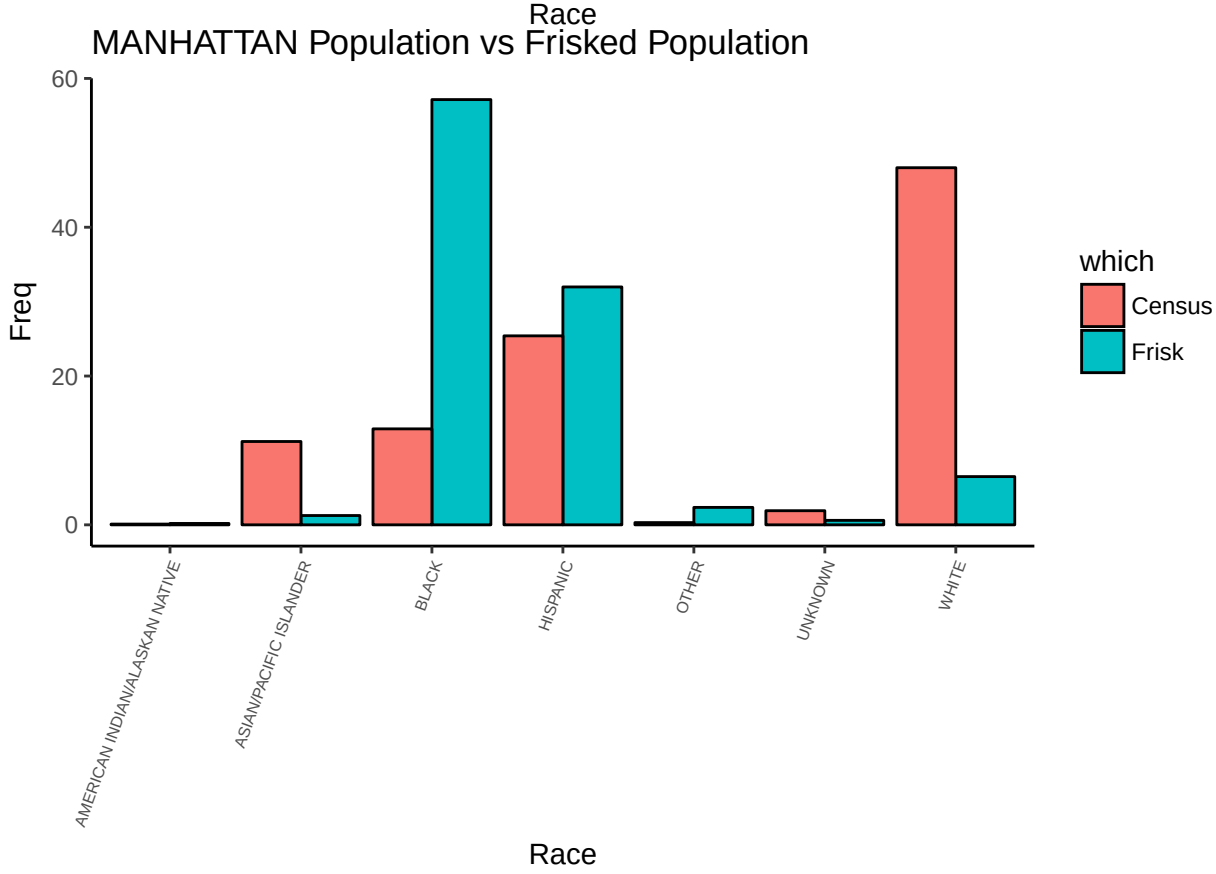
Question 2: Are they frisked in proportion to the demographic characteristics of the area?

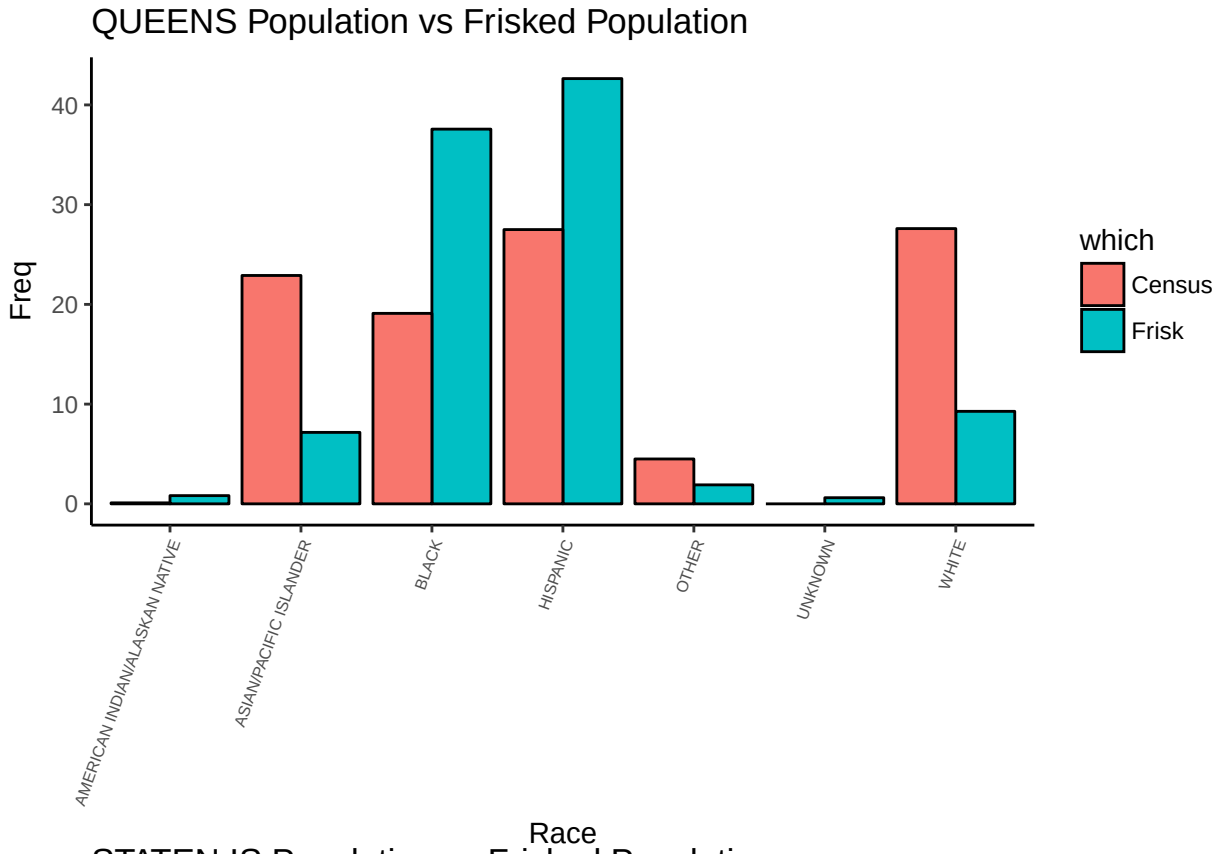
```
Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "BRONX"])*100/sum(sqf2010$frisked[sqf2010
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_bronx=rbind(friskData, bronxData)

Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "STATEN IS"])*100/sum(sqf2010$frisked[sqf
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_staten=rbind(friskData, statenData)

Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "QUEENS"])*100/sum(sqf2010$frisked[sqf201
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_queens=rbind(friskData, queensData)

Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "MANHATTAN"])*100/sum(sqf2010$frisked[sqf
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_man=rbind(friskData, manData)

Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "BROOKLYN"])*100/sum(sqf2010$frisked[sqf2
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_brook=rbind(friskData, brookData)


bigL=list(Bronx=frisk_vs_bronx,Brooklyn=frisk_vs_brook,Manhattan=frisk_vs_man, Queens=frisk_vs_queens,S
cites=names(sqf2010_by_borough)[-1]
```

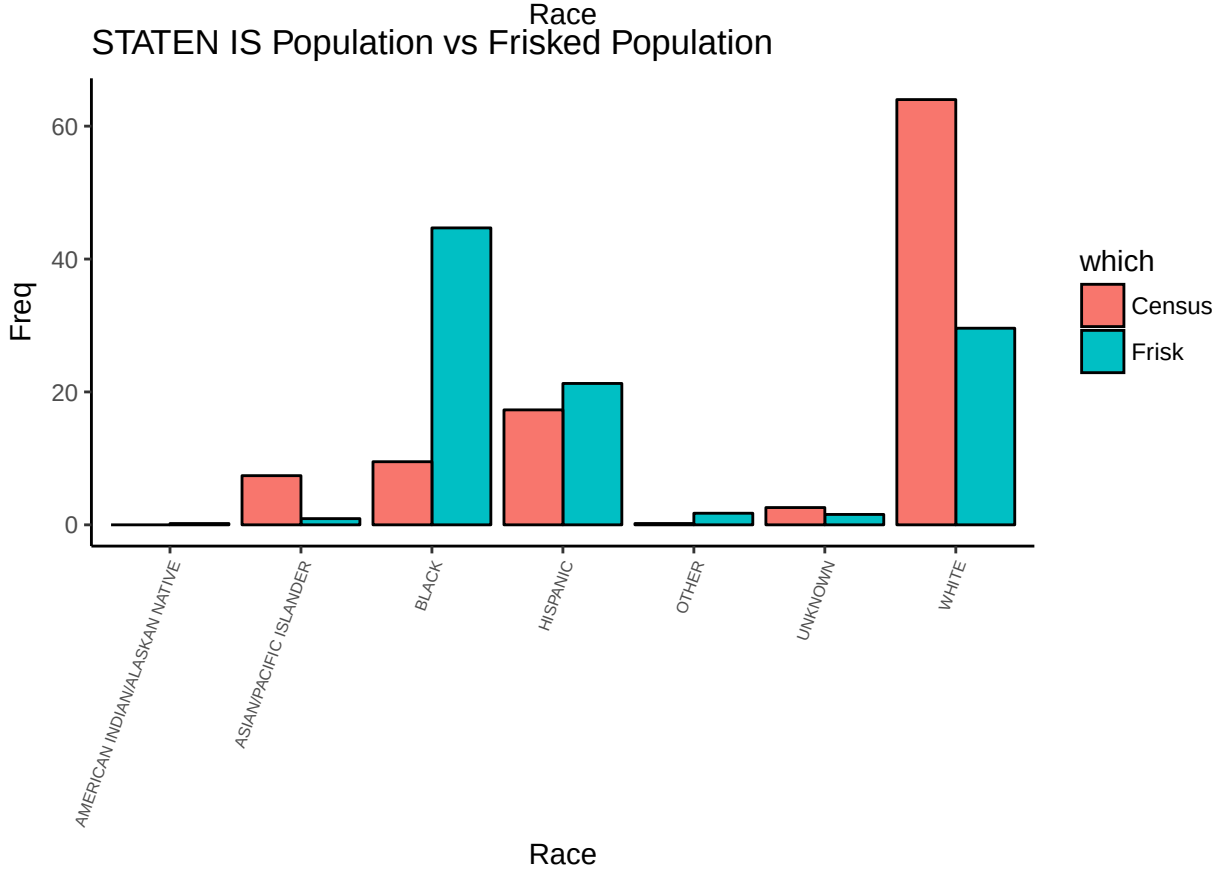
```
c=1
for (i in bigL){
  city=cites[c]
  p=ggplot(i, aes(x=Race, y=Freq, fill=which))+
    geom_bar(stat="identity", position=position_dodge(), color="Black")+
    theme_classic()+
    theme(axis.text.x = element_text(angle = 70, hjust = 1, size = rel(0.7) ))+
    ggtitle(paste(city,"Population vs Frisked Population"))
  plot(p)
  c=c+1
}
```




BROOKLYN Population vs Frisked Population
Freq
60
40
20
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
HISPANIC
OTHER
UNKNOWN
WHITE
Race
which
Census
Frisk


MANHATTAN Population vs Frisked Population
Freq
60
40
20
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
HISPANIC
OTHER
UNKNOWN
WHITE
Race
which
Census
Frisk





We can use `prop.test` to test for the significance of the difference.

```r
pop_staten=c(nrow(sqf2010_by_borough$`STATEN IS`), 468730)
pop_bronx=c(nrow(sqf2010_by_borough$BRONX),1385108)
pop_man=c(nrow(sqf2010_by_borough$MANHATTAN),1585873)
pop_queen=c(nrow(sqf2010_by_borough$QUEENS),223072)
pop_brook=c(nrow(sqf2010_by_borough$BROOKLYN),2504710)
pop_vectors=rbind(pop_bronx,pop_brook, pop_man, pop_queen , pop_staten)

pData_black=data.frame()
for(i in 1:length(bigL)){
  city_data=bigL[[i]]
  blk=city_data[city_data$Race=="BLACK",][,2]*pop_vectors[i,]/100
  table=cbind(blk, pop_vectors[i,])
  rownames(table)=c("SQF", "Census")
  pval=prop.test(table, alternative="greater",correct=F)
  pval=pval$p.value
  pData_black=rbind(pData_black, data.frame(city=cities[i], pval=pval))
}



pData_his=data.frame()
for(i in 1:length(bigL)){
  city_data=bigL[[i]]
  blk=city_data[city_data$Race=="HISPANIC",][,2]*pop_vectors[i,]/100
  table=cbind(blk, pop_vectors[i,])
  rownames(table)=c("SQF", "Census")
  pval=prop.test(table, alternative="greater",correct=F)
  pval=pval$p.value
  pData_his=rbind(pData_his, data.frame(city=cities[i], pval=pval))
}


pData_white=data.frame()
for(i in 1:length(bigL)){
  city_data=bigL[[i]]
  blk=city_data[city_data$Race=="WHITE",][,2]*pop_vectors[i,]/100
  table=cbind(blk, pop_vectors[i,])
  rownames(table)=c("SQF", "Census")
  pval=prop.test(table, alternative="less", correct=F)
  pval=pval$p.value
  pData_white=rbind(pData_white, data.frame(city=cities[i], pval=pval))
}

#hypothesis test to test whether the the proportion of frisked who are black is greater than the propor
pData_black
```

```
##        city pval
## 1     BRONX    0
## 2  BROOKLYN    0
## 3 MANHATTAN    0
## 4    QUEENS    0
## 5 STATEN IS    0
```

```
#hypothesis test to test whether the the proportion of frisked who are hispanic is greater than the pro
pData_his
```

```
##        city         pval
## 1     BRONX 0.000000e+00
## 2  BROOKLYN 9.690730e-35
## 3 MANHATTAN 0.000000e+00
## 4    QUEENS 0.000000e+00
## 5 STATEN IS 2.416089e-44
```

```
#hypothesis test to test whether the the proportion of frisked who are black is greater than the propor
pData_white
```

```
##        city pval
## 1     BRONX    0
## 2  BROOKLYN    0
## 3 MANHATTAN    0
## 4    QUEENS    0
## 5 STATEN IS    0
```

Analysis

In order to test if the difference between the percentage of people of a certain race stopped and the percentage of people of a certain race who live in that borough are drastically different, we used the `prop.test` function in R. For both the black and hispanic populations we said that our alternative hyopthesis was that the percentage of stopped people is greater than the percentage of people living in that borough. We found that the p-value calculated using this function gave us roughly zero for the percentage of stopped people who are black versus the percentage of black people in that borough. Since the p-value is less than 0.05 than we can confidently say that these two percentages are significantly different (which is a problem). The H_0 is rejected. Performing the same prop.test for the hispanic population of each other five boroughs, we get practically identical results, all the p-values are practically 0 (which is less than 0.05) so we can confidently say that these two percentages are significantly different for hispanics as well. As for the white population, we said the alternative hypothesis was that the percentage of stopped people who are white is less than the perentage of white people in that borough. The p-values we got for each of the five boroughs was approximately 0, rejecting our H_0. This means that we can say that the percentage of white people in the borough is significantly greater than the percentage of stopped people that are white.

We also did the same prop.test for the percentage of people frisked in the borough versus percentage of people living in the borough. For the African American population we found that the p-value also approximated to 0 for each of the black boroughs for the some hypothesis we had above. Our results were also identical to the ones above for both Hispanics and Whites.

Concluding Remarks

After looking at all this data, we can make a few important conclusions. First, in New York City and most likely the rest of the United States, our police departments do racial profiling and stop and frisk an unproportionate amount of African American and Hispanic people. Second, there must be changes made to this problem. Some proposed changes could be a decrease in the amount of stop and frisk that police departments around the nation and more particulary, in NYC, do. In addition, racial profiling must be addressed in the education and training of police officers around the United States in order to ensure equal treatement of all citizens and residents of this country.

Bibliography

Barone, M., "Stop-and-Frisk Protects Minorities", http://www.nationalreview.com/article/356481/stop-and-frisk-protects-minorities-michael-barone, 2013.

New York Civil Liberties Union website: http://www.nyclu.org/node/1598

New York Civil Liberties Union, "Stop and Frisk During the Bloomberg Administration", http://www.nyclu.org/files/publications/stopandfrisk_briefer_2002-2013_final.pdf

Geller, A., Fagan, J., Tyler, T., Link, B., "Aggressive Policing and the Mental Health of Young Urban Men", *Am J Public Health.* 2014 December; 104(12): 2321-2327. Published online 2014 December. doi: 10.2105/AJPH.2014.302046