Final Project: New York City Stop-Question-Frisk Analysis

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

Due December 5, 2016

Question 1 Code

```
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

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

Question 2 Chunk 1 Code

```
#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
```

Question 2 Chunk 2 Code

```
#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
```

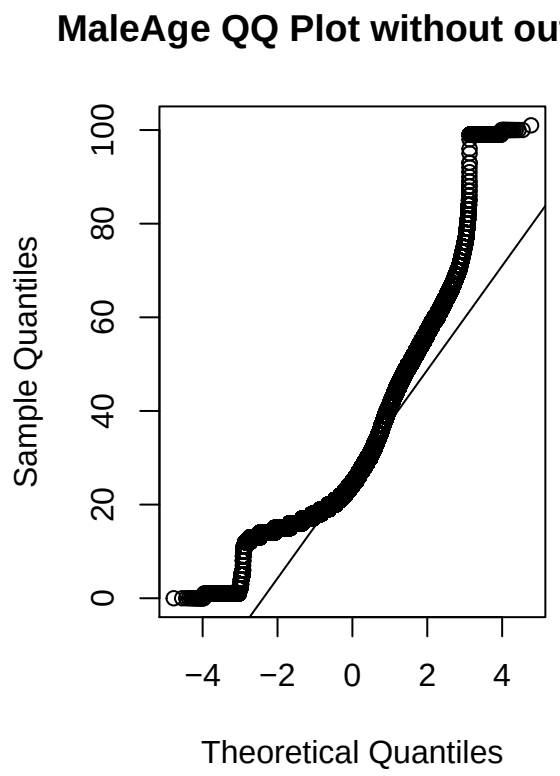
```
## 10     BRONX MANHATTAN    0
## 12     BRONX STATEN IS    0
## 14  BROOKLYN     BRONX    0
## 17  BROOKLYN    QUEENS    0
## 20 MANHATTAN     BRONX    0
## 23 MANHATTAN    QUEENS    0
## 27    QUEENS  BROOKLYN    0
## 28    QUEENS MANHATTAN    0
## 30    QUEENS STATEN IS    0
## 32 STATEN IS  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

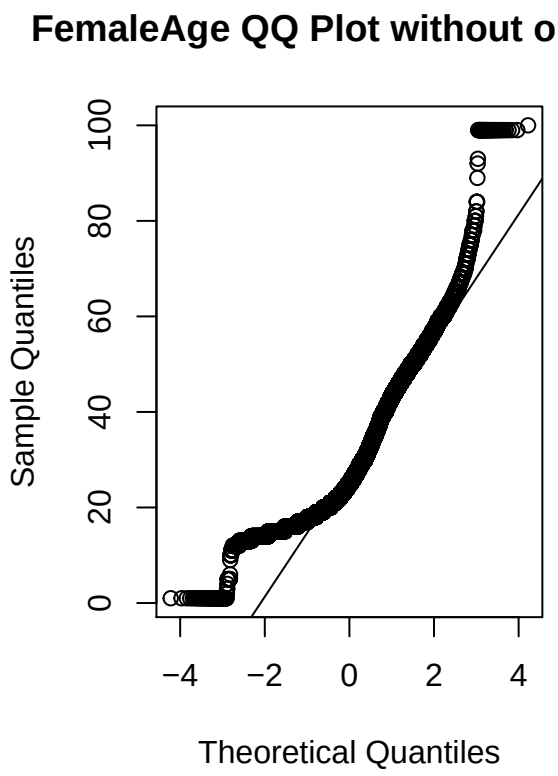
```
#find the groups with the smalled differences (top 60% of pvals excluding identical co)
results_sub=results[results$pval<1,]
results_sub[results_sub$pval>0.4*max(results_sub$pval),]
```

```
##         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
```

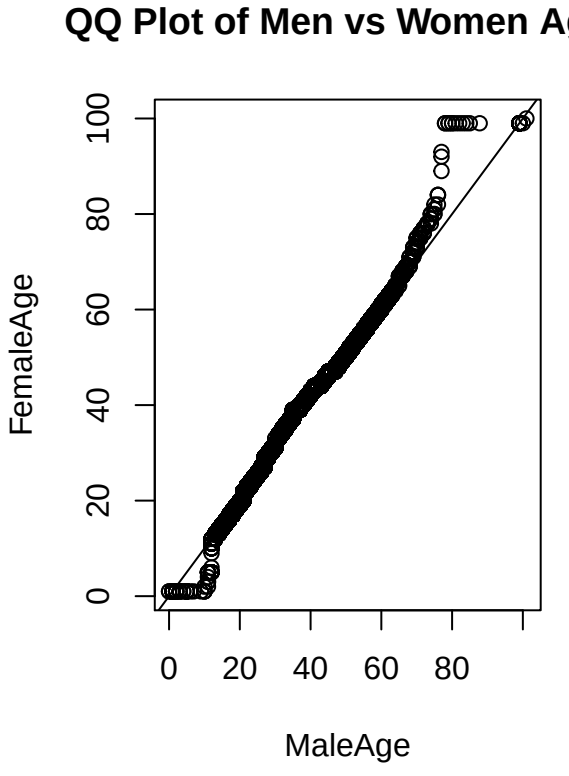
Question 3 Code

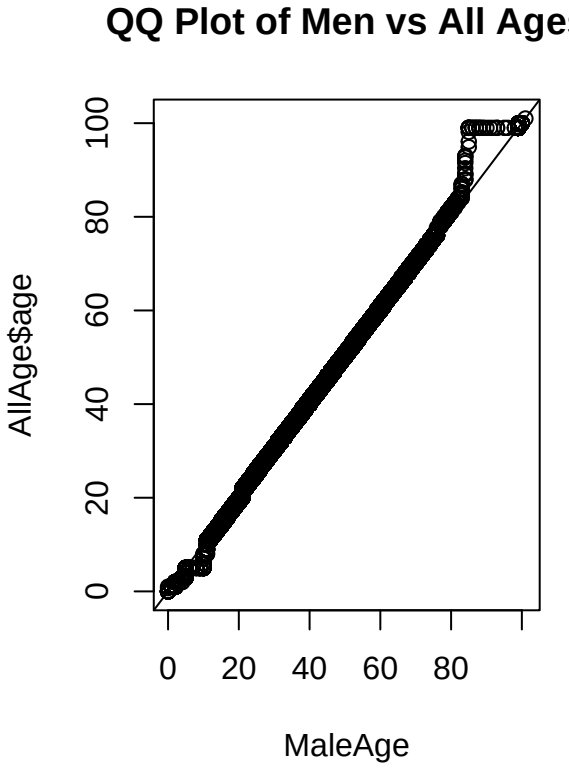
```
#throw out outliers
AllAge <- sqf2010[sqf2010$age <110,]

#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)
```





Question 4 Chunk 1 Code

```
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.

Question 4 Chunk 2 Code

```
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.

Question 4 Chunk 3 Code

```
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
```

Question 5 Code

```
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
```

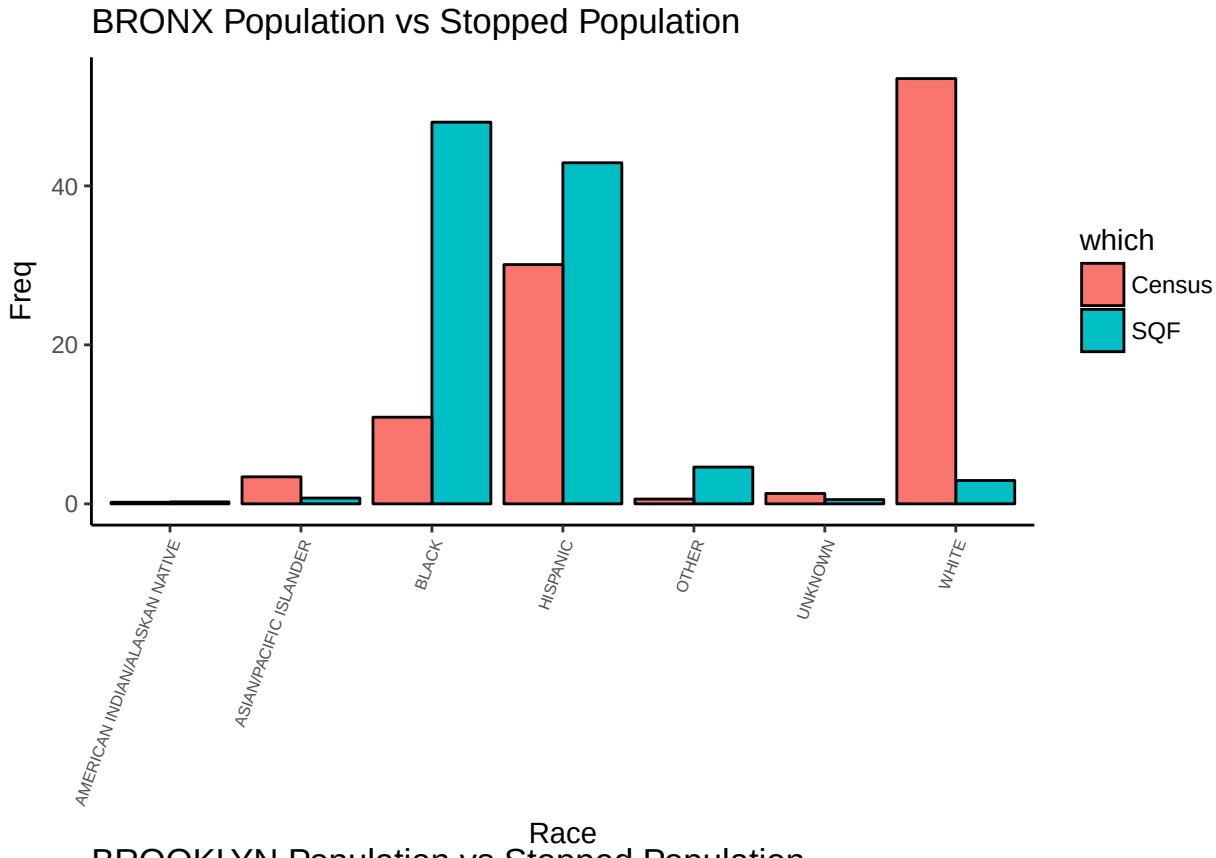
Team-Chosen Questions - 30% of Grade

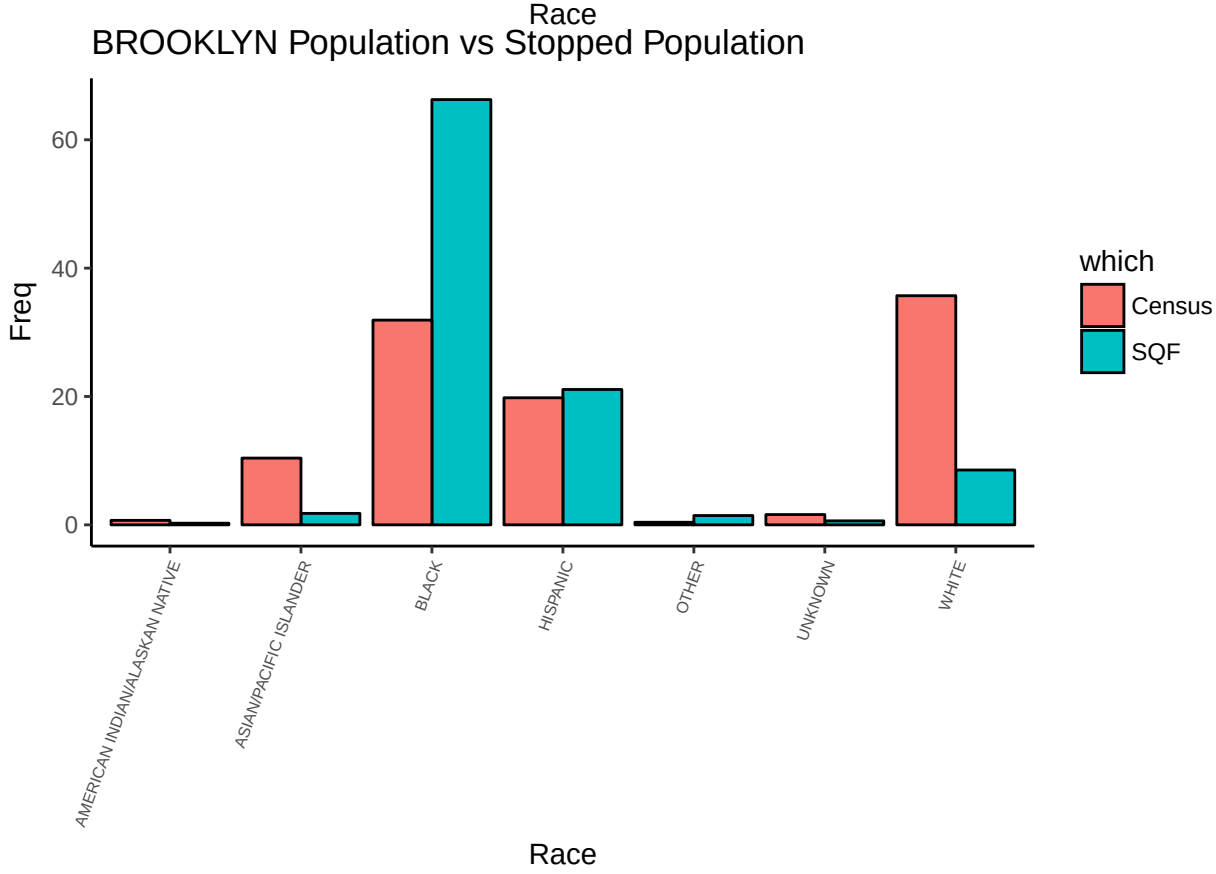
Question 1: Are people stopped in proportion to the demographic characteristics of the area?

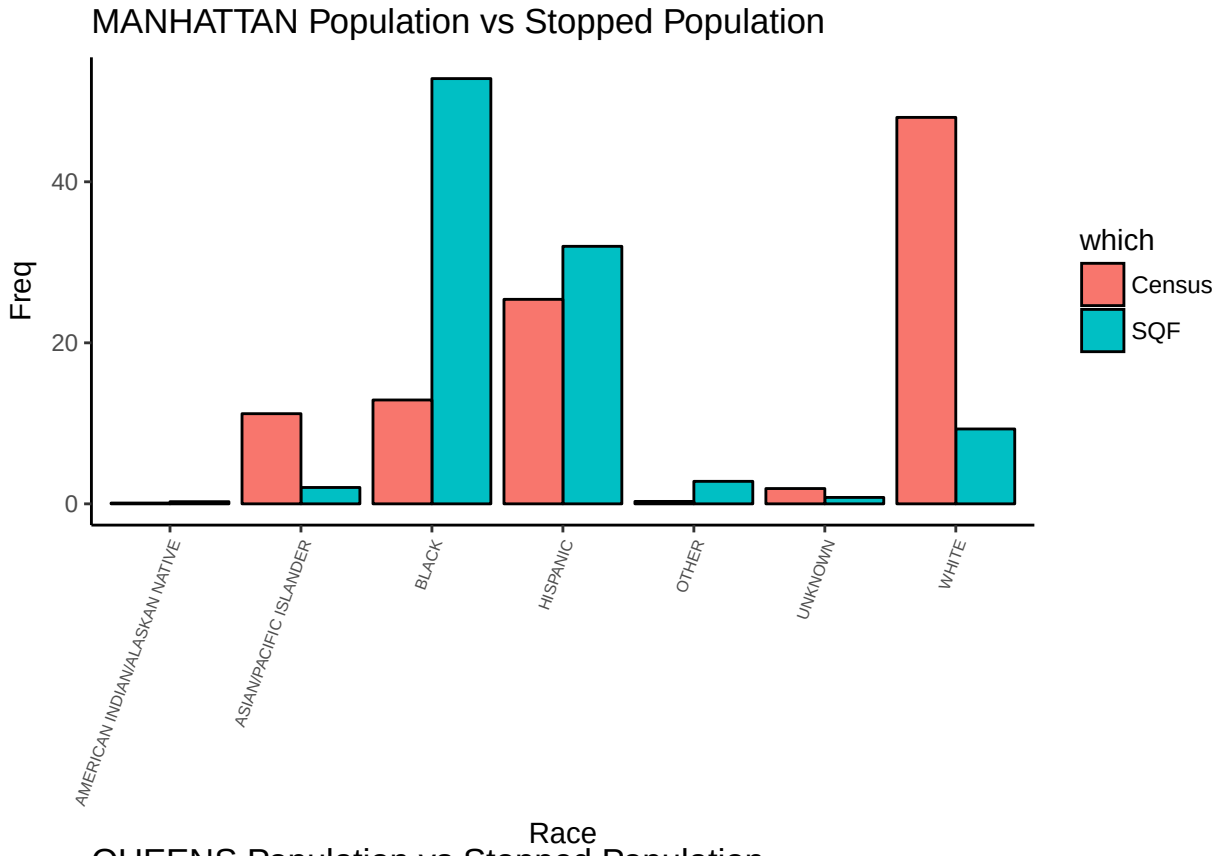
Team-Chosen Question 1 Chunk 1 Code

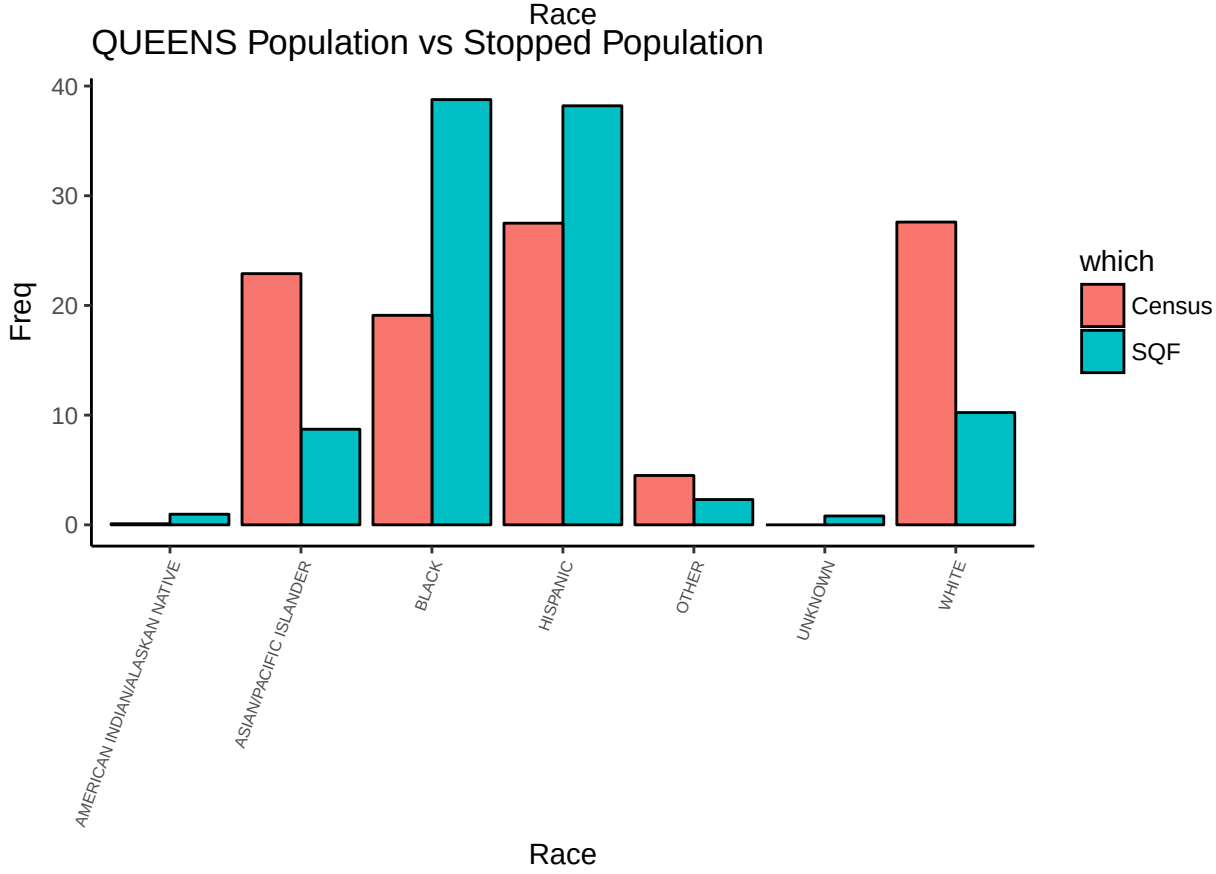
```
sqf2010$race=gsub("BLACK-","",sqf2010$race)
sqf2010$race=gsub("WHITE-","",sqf2010$race)
sqf2010_by_borough=split(sqf2010, sqf2010$city)
races=unique(sqf2010$race)

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)
```

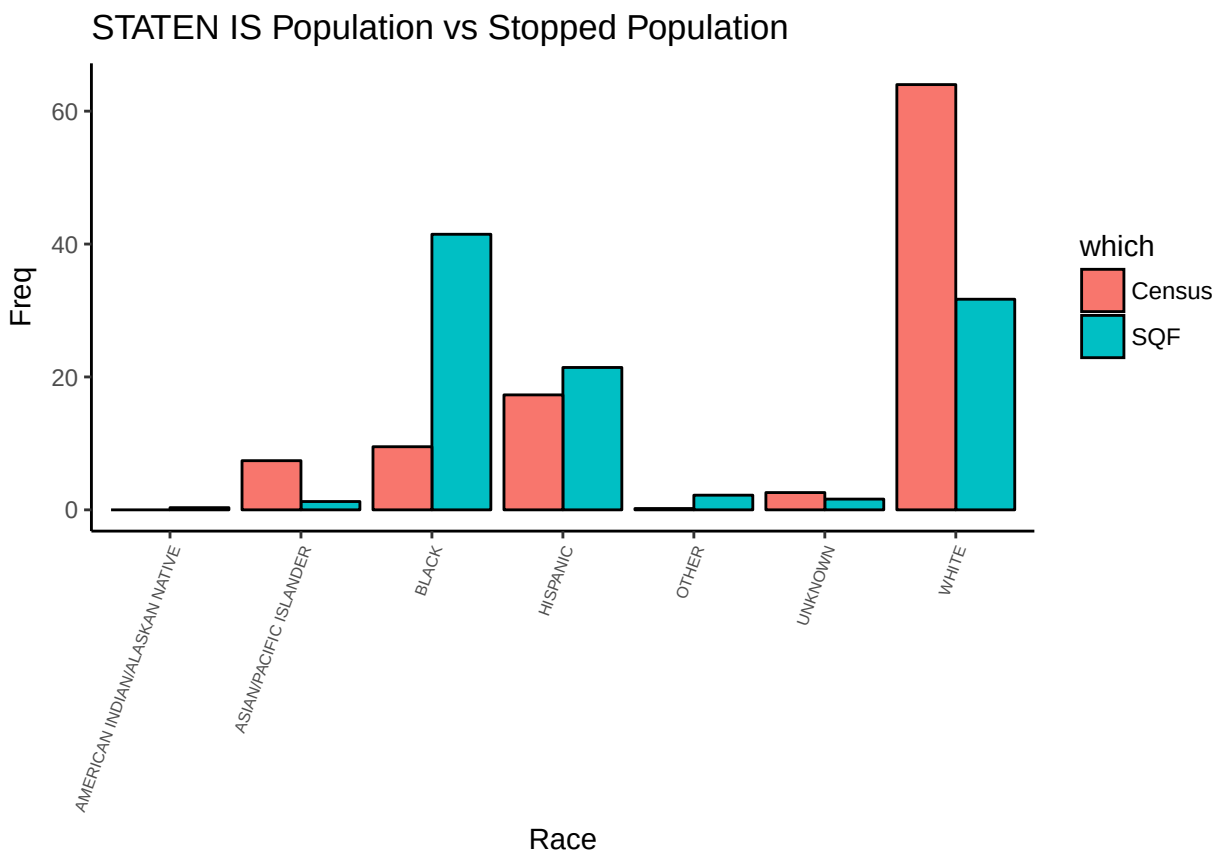
```r
statenData=data.frame(Race=races, Freq=staten, which="Census")

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
  }
```


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


BROOKLYN Population vs Stopped Population
Freq
0
20
40
60
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


QUEENS Population vs Stopped Population
Freq
40
30
20
10
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
HISPANIC
OTHER
UNKNOWN
WHITE
Race
which
Census
SQF



Team-Chosen Question 1 Chunk 2 Code

```
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
```

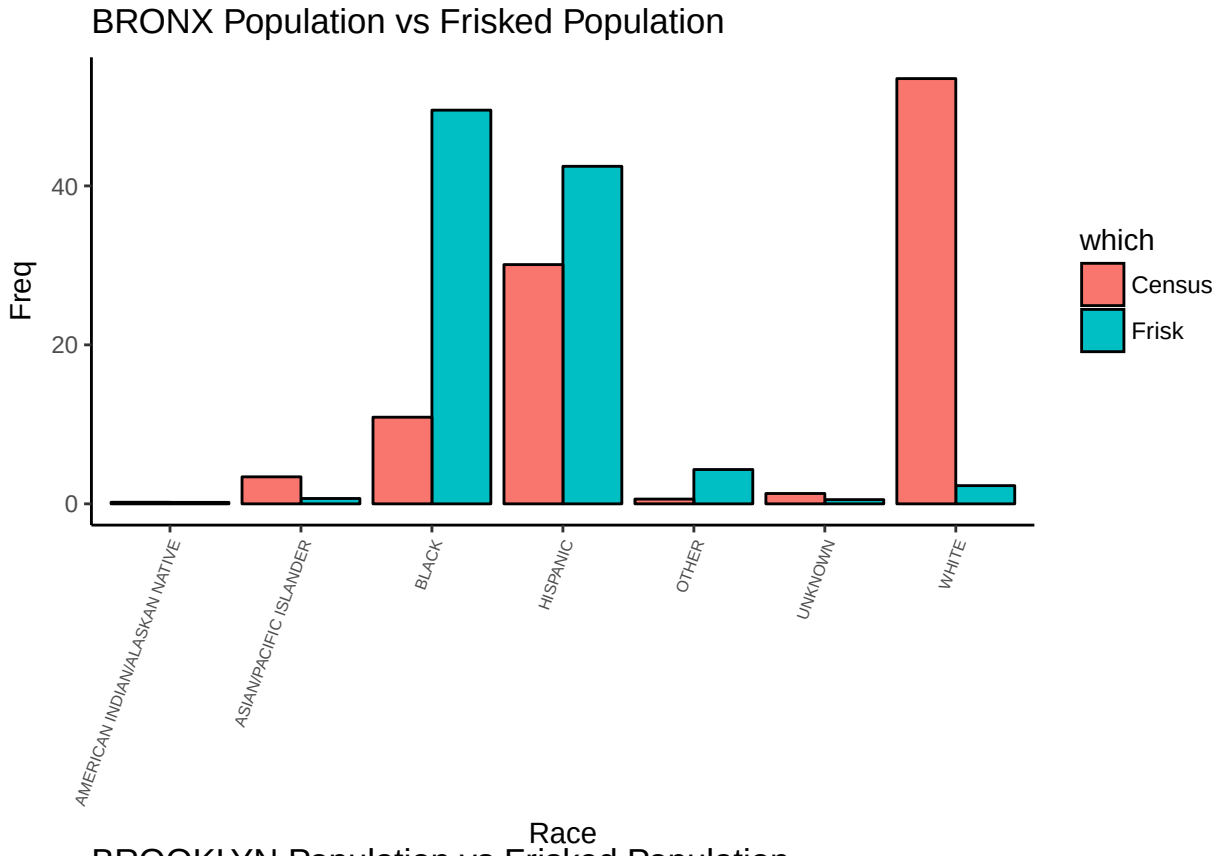
```
#hypothesis test to test whether the the proportion of stopped who are black is greater than the propor
pData_white
```

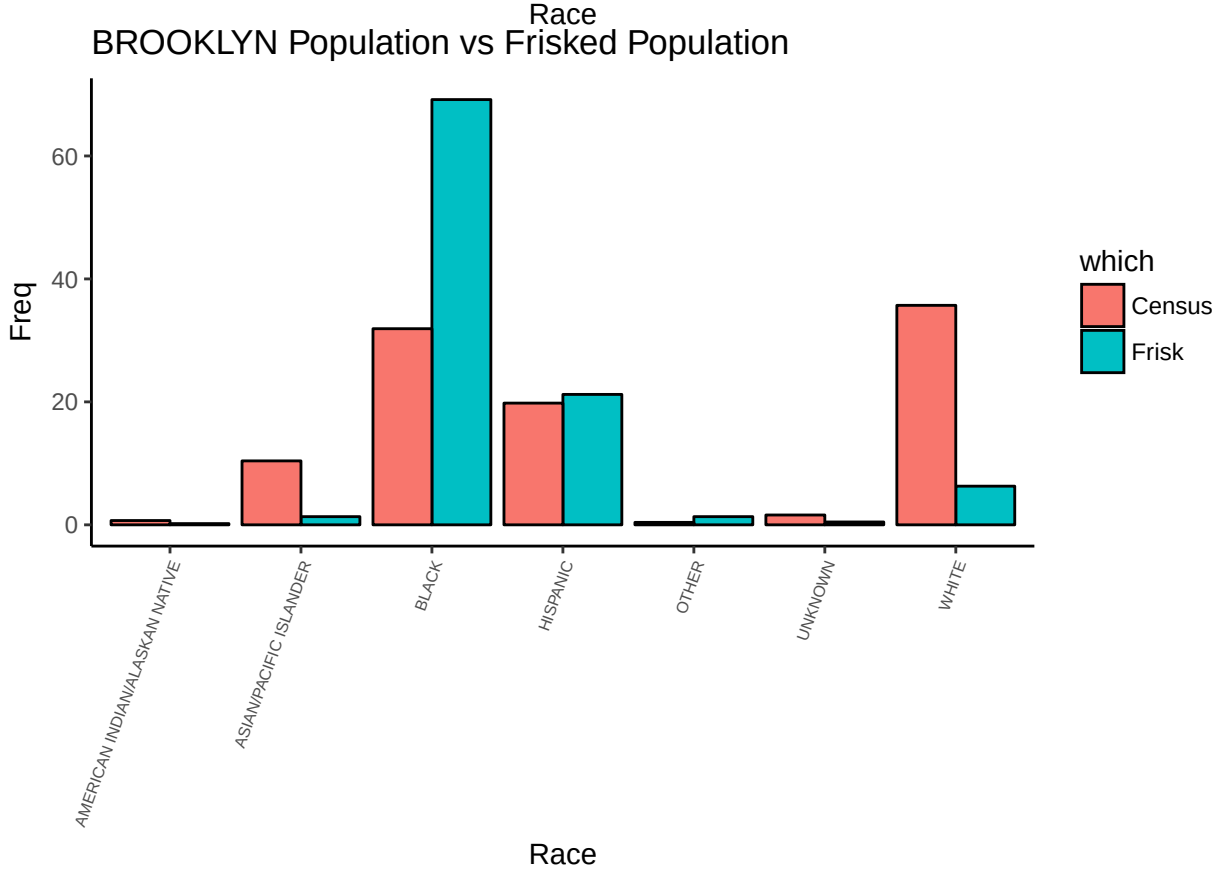
```
##        city pval
## 1     BRONX    0
## 2  BROOKLYN    0
## 3 MANHATTAN    0
## 4    QUEENS    0
## 5 STATEN IS    0
```

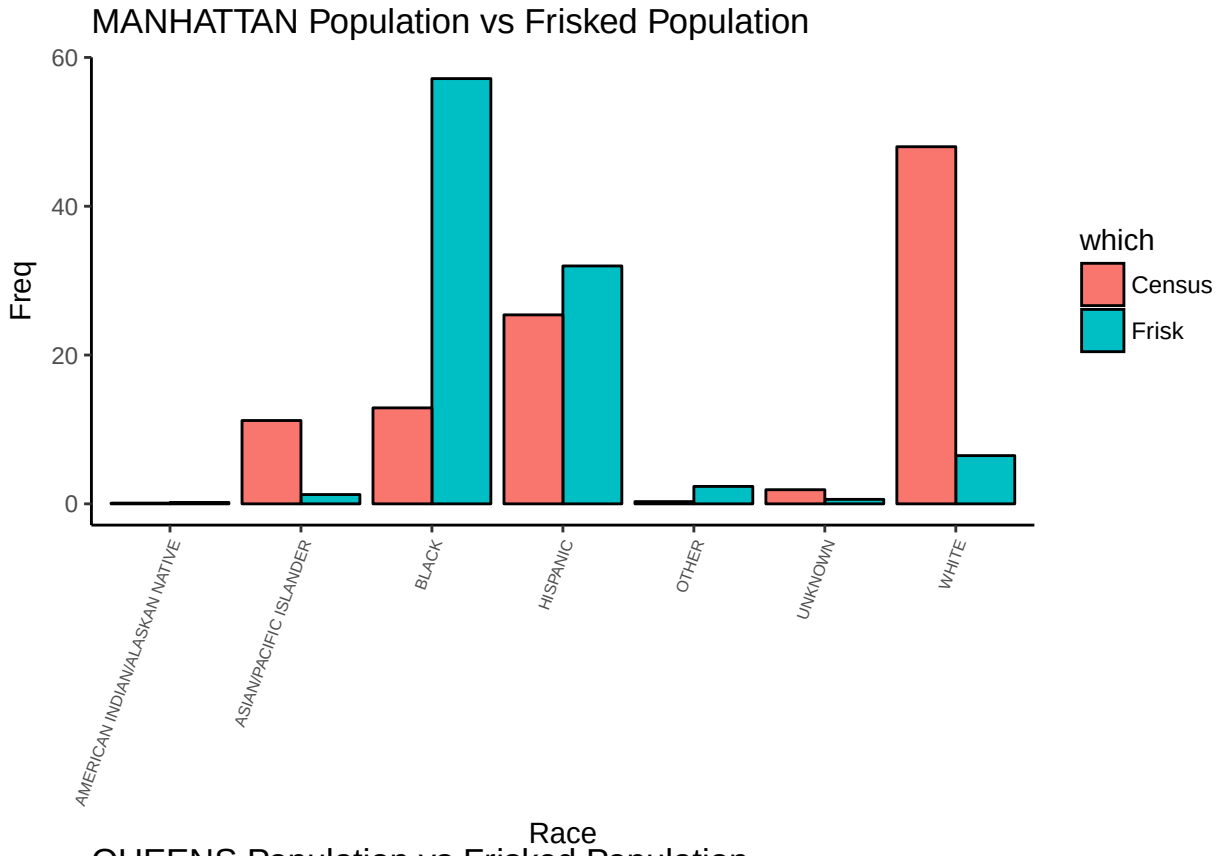
Team-Chosen Question 2 Chunk 1 Code

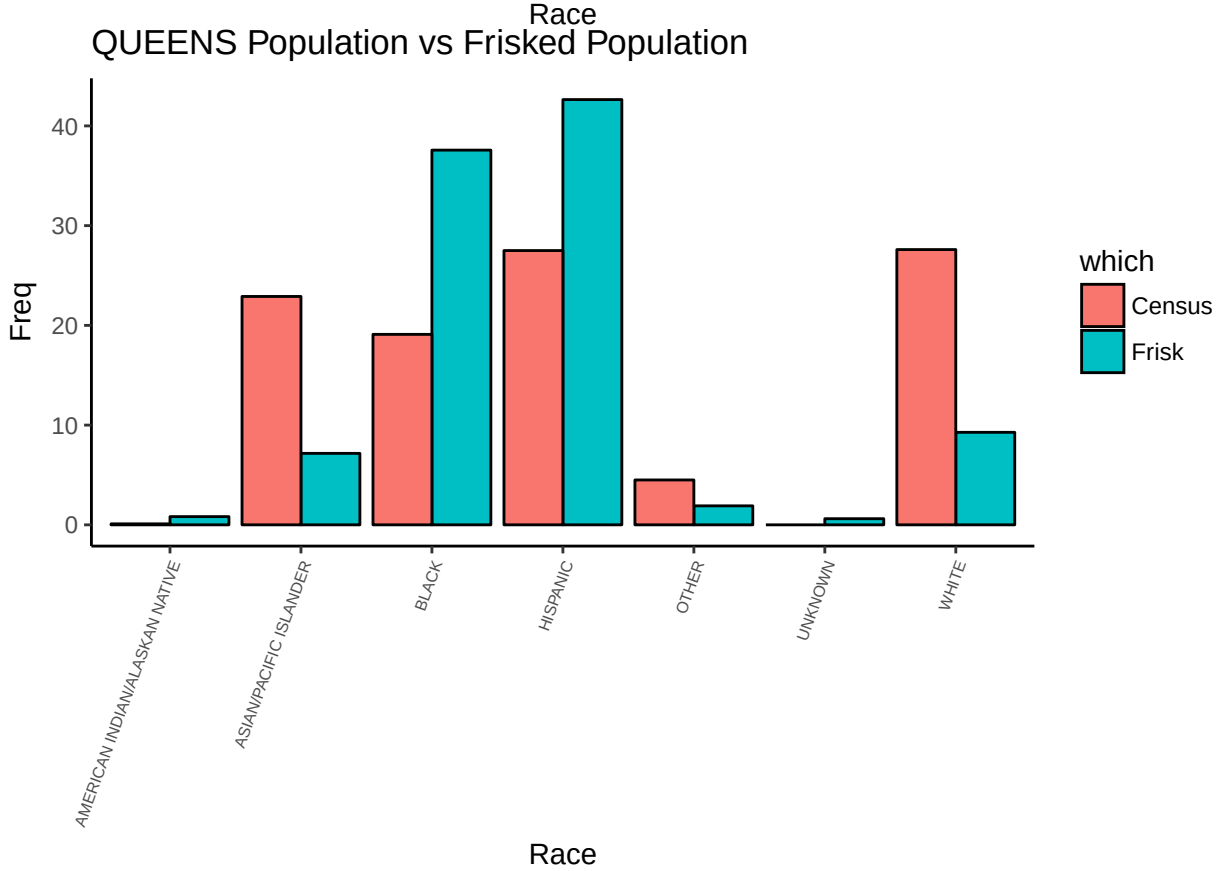
```
Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "BRONX"])*100/sum(sqf2010$frisked[sqf2010
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_bronx=rbind(friskData, bronxData)
```

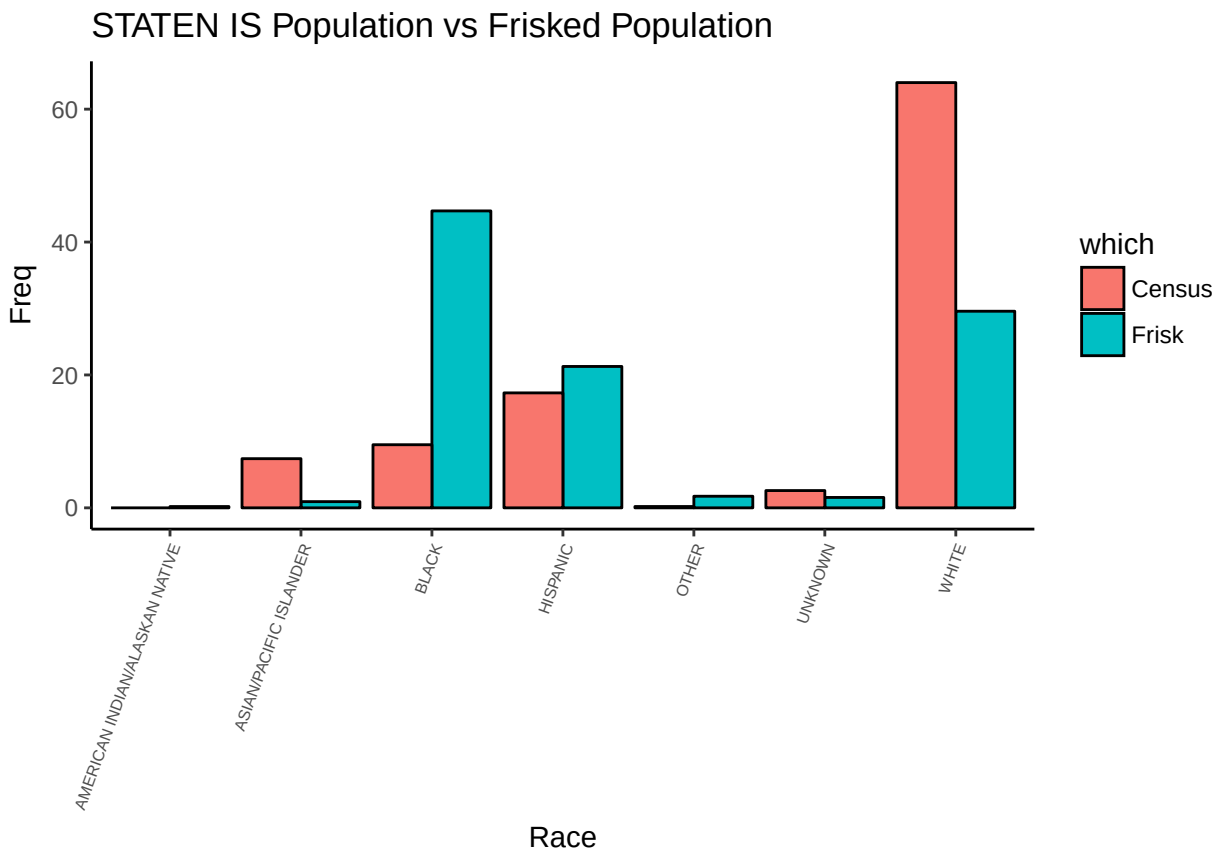
```
Freq <- (table(sqf2010$race[sqf2010$frisked & sqf2010$city == "STATEN IS"])*100/sum(sqf2010$frisked[sqf
friskData <- data.frame(Freq)
friskData$which <- "Frisk"
names(friskData)=names(bronxData)
frisk_vs_staten=rbind(friskData, statenData)

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
}
```


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


BROOKLYN Population vs Frisked Population
Freq
0
20
40
60
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


QUEENS Population vs Frisked Population
Freq
40
30
20
10
0
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
HISPANIC
OTHER
UNKNOWN
WHITE
Race
which
Census
Frisk



Team-Chosen Question 2 Chunk 2 Code

```
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
```

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