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

Part I: Individual Analysis

Purpose:

This project will give you the opportunity to apply the techniques learned in Math 35 to a *real* and *sizeable* data set. Moreover, you will use these techniques to draw conclusions about the efficacy of a controversial crime policy: the stop-question-frisk (SQF), also known as stop-and-frisk, program, in New York City.

For this first part of the project, you will work *individually* to familiarize yourselves with the data. This portion of the project is due **Tuesday, December 6, in class.**

Background:

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. Since the data from the form has over 100 columns, we will be giving you a version of the data that has been simplified to 22 of the most interesting fields. Every SQF stop from 2010 and 2015 is still represented in the data. If you are interested, you can find the full data set at: http://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_report.shtml. However, for this project, please **only use the version of the SQF data that has been provided for the course**.

Instructions: Individual Portion - Due in class on Tuesday December 6 - 15% of Grade

The purpose of this individual portion is to help you get familiar with the data prior to class on Tuesday. Class time on Tuesday, December 6 and Thursday, December 8 will be allocated for assigning teams and to work on the project in your groups. Therefore, **attendance on these days will be mandatory and count as additional 10% of your project grade**.

Complete the following tasks on your own, using R. You do not need a formal writeup, but please submit your R code, either as an RMarkdown file, an R script file, or as a printout of your input and output from the R console.

1. **Loading the Data**

First, download the necessary files for the project by running

```
source("https://www.math.hmc.edu/m35f/FinalLoader.R")
```

This will download the `2010_sqf_m35.csv` and `2015_sqf_m35.csv` files, which contain the SQF data from 2010 and 2015, respectively, and `NewYorkSQF.R`, which contains the initial commands for getting setup.

To load the data from the files into R, run the commands

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

These will load the 2010 and 2015 SQF data into the data frames `sqf2010` and `sqf2015`, respectively. If you close R and want to reload the data, you will first want to set your working directory to the `Math35` folder within your `Documents` folder, which is where the `2010_sqf_m35.csv` and `2015_sqf_m35.csv` files are located. To do so without having to download the .csv files again by re-running the `source()` command above, you can run the following code, which is at the top of the `NewYorkSQF.R` file:

```
setwd('~')
if(regexpr('Documents',getwd())==-1) {
  if(!dir.exists('~/Documents')) dir.create('~/Documents')
  setwd('~/Documents')
}
if(!dir.exists('Math35')) dir.create('Math35')
setwd('Math35')
```

Below is a key for the fields (columns) in the data frame.

Column name	Description
datestop	Date of the stop in MMDDYYYY format
timestop	Time of the stop in HHMM format, where HH is in 24 hour format
sex	Sex of the suspect
race	Race of the suspect
age	Age of the suspect
weight	Weight of the suspect
haircolr	Hair color of the suspect
eyecolor	Eye color of the suspect
build	Suspect's build
city	City (borough) of New York City in which stop occurred
inside	Stop was made inside (1) or outside (0)
location	Was location of the stop at a transit authority or housing authority?
typeofid	Type of ID provided by suspect
perobs	Period of observation, in minutes
perstop	Period of stop, in minutes
arstmade	Was an arrest made? (1=yes, 0=no)
sumissue	Was a summons issued? (1=yes, 0=no)
frisked	Was the suspect frisked? (1=yes, 0=no)
searched	Was the suspect searched? (1=yes, 0=no)
contrabn	Was contraband found on the suspect? (1=yes, 0=no)
radio	Was there a radio run? (1=yes, 0=no)
height	Height of the suspect, in inches
pf	Was physical force used by the officer (including hands, weapons drawn, baton, handcuffs, pepper spray)? (1=yes, 0=no)
weap	Was a weapon (pistol, knife, assualt rifle, machine gun, or other) found on the suspect? (1=yes, 0=no)

2. **Summarize the Data**

Use the `head`, `summary`, and `dim` commands to get an initial look at the contents of the data. How many entries are in the 2010 data? How many entries are in the 2015 data?

For the rest of the problems in this preliminary section, choose either the 2010 or 2015 data. You may want to refer to Lecture 7 on Exploratory Data Analysis for the necessary commands to complete this section.

```
sqf2010 = read.csv("2010_sqf_m35.csv")
head(sqf2010)
```

```
##   datestop timestop sex           race age weight        haircolr eyecolor
## 1  1012010      340   M          BLACK  17    180           BLACK    BROWN
## 2  1042010     1548   M BLACK-HISPANIC  20    160           BLACK    BROWN
## 3  1092010     1550   M WHITE-HISPANIC  55    145 SALT AND PEPPER    BROWN
## 4  1112010     1120   M WHITE-HISPANIC  17    165           BROWN    BROWN
## 5  1222010     1620   M          BLACK  55    250           BLACK    BROWN
## 6  1282010     1342   M WHITE-HISPANIC  57    160           BLACK    BROWN
##    build      city inside          location typeofid perobs perstop
## 1 MEDIUM  BROOKLYN      0           NEITHER   VERBAL      1      13
## 2 MEDIUM MANHATTAN      0           NEITHER    PHOTO      2       5
## 3 MEDIUM MANHATTAN      1 TRANSIT AUTHORITY    PHOTO      1       6
## 4 MEDIUM    QUEENS      0           NEITHER    PHOTO      5      10
## 5  HEAVY MANHATTAN      0 HOUSING AUTHORITY    PHOTO      1       5
## 6 MEDIUM MANHATTAN      1 TRANSIT AUTHORITY    PHOTO      5       6
##   arstmade sumissue frisked searched contrabn radio height pf weap
## 1        0        1       1        0        0     0     70  1    0
## 2        0        1       1        0        0     0     66  0    0
## 3        0        0       1        1        0     0     65  1    0
## 4        0        0       1        0        0     0     69  0    0
## 5        0        0       0        0        0     0     73  0    0
## 6        1        0       0        0        0     0     69  0    0
```

```
summary(sqf2010)
```

```
##     datestop           timestop    sex
##  Min.   : 1012010   Min.   :   0   F: 41609
##  1st Qu.: 4012010   1st Qu.: 922   M:549917
##  Median : 6182010   Median :1602   Z:  9759
##  Mean   : 6480570   Mean   :1401
##  3rd Qu.: 9272010   3rd Qu.:2040
##  Max.   :12312010   Max.   :2359
##
##                      race             age             weight
##  BLACK                 :315083   Min.   :  0.00   Min.   :  0.0
##  WHITE-HISPANIC        :150637   1st Qu.: 19.00   1st Qu.:150.0
##  WHITE                 : 54810   Median : 25.00   Median :170.0
##  BLACK-HISPANIC        : 38689   Mean   : 29.01   Mean   :169.9
##  ASIAN/PACIFIC ISLANDER: 19732   3rd Qu.: 34.00   3rd Qu.:180.0
##  OTHER                 : 15360   Max.   :999.00   Max.   :999.0
##  (Other)               :  6974
##     haircolr         eyecolor          build               city
##  BLACK  :452646   BROWN  :538596   HEAVY   : 51245            :    60
##  BROWN  :116088   BLACK  : 39411   MEDIUM  :359304   BRONX    :112415
##  BALD   : 10124   BLUE   : 10035   MUSCULAR:  1969   BROOKLYN :195155
##  BLOND  :  7134   GREEN  :  4044   THIN    :183713   MANHATTAN:122082
##  GRAY   :  5609   HAZEL  :  3759   UNKNOWN :  5054   QUEENS   :144072
##  UNKNOWN:  3047   UNKNOWN:  2322                     STATEN IS: 27501
##  (Other):  6637   (Other):  3118
##      inside                    location         typeofid
##  Min.   :0.0000   HOUSING AUTHORITY: 84679   OTHER  :  8383
##  1st Qu.:0.0000   NEITHER          :470090   PHOTO  :329735
##  Median :0.0000   TRANSIT AUTHORITY: 46516   REFUSED: 13560
##  Mean   :0.2207                              VERBAL :249607
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

```
##
##      perobs            perstop           arstmade          sumissue
##  Min.   :  0.000   Min.   :  0.000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:  1.000   1st Qu.:  3.000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :  1.000   Median :  5.000   Median :0.00000   Median :0.00000
##  Mean   :  2.475   Mean   :  5.742   Mean   :0.06833   Mean   :0.07077
##  3rd Qu.:  2.000   3rd Qu.:  5.000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :930.000   Max.   :999.000   Max.   :1.00000   Max.   :1.00000
##
##     frisked           searched          contrabn           radio
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :1.0000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.5611   Mean   :0.09246   Mean   :0.01948   Mean   :0.2512
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##      height            pf              weap
##  Min.   :36.0   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:67.0   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :69.0   Median :0.0000   Median :0.00000
##  Mean   :68.6   Mean   :0.2324   Mean   :0.01264
##  3rd Qu.:71.0   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :95.0   Max.   :1.0000   Max.   :1.00000
##
```

```
dim(sqf2010)
```

```
## [1] 601285     24
```

```
sqf2015 = read.csv("2015_sqf_m35.csv")
head(sqf2015)
```

```
##   datestop timestop sex  race age weight haircolr eyecolor  build
## 1  1012015      315   M WHITE  33    190    BROWN    BROWN MEDIUM
## 2  1152015     1747   M BLACK  14    140    BLACK    BROWN   THIN
## 3  1292015     1745   M BLACK  14    140    BLACK    BROWN   THIN
## 4  1292015     1745   M BLACK  14    180    BLACK    BROWN MEDIUM
## 5  1292015     1745   M BLACK  13    160    BLACK    BROWN MEDIUM
## 6  1292015     1745   M WHITE  13    130    BLACK    BROWN MEDIUM
##        city inside location typeofid perobs perstop arstmade sumissue
## 1  BROOKLYN      0  NEITHER   VERBAL      2      10        0        0
## 2 MANHATTAN      0  NEITHER   VERBAL      1       4        0        0
## 3 MANHATTAN      0  NEITHER   VERBAL      1      16        0        0
## 4 MANHATTAN      0  NEITHER   VERBAL      1      16        0        0
## 5 MANHATTAN      0  NEITHER   VERBAL      1      16        0        0
## 6 MANHATTAN      0  NEITHER   VERBAL      1      16        0        0
##   frisked searched contrabn radio height pf weap
## 1       1        0        0     0     71  0    0
## 2       1        0        0     1     68  1    0
## 3       0        0        0     1     63  1    0
## 4       0        0        0     1     69  0    0
## 5       0        0        0     1     70  1    0
## 6       0        0        0     1     62  1    0
```

```
summary(sqf2015)
```

```
##     datestop           timestop    sex
##  Min.   : 1012015   Min.   :   0   F: 1515
##  1st Qu.: 3112015   1st Qu.: 505   M:20853
##  Median : 5282015   Median :1623   Z:  195
##  Mean   : 5957353   Mean   :1377
##  3rd Qu.: 9012015   3rd Qu.:2052
##  Max.   :12312015   Max.   :2359
##
##                      race            age             weight
##  BLACK                 :11950   Min.   :  0.00   Min.   :  1.0
##  WHITE-HISPANIC        : 5090   1st Qu.: 19.00   1st Qu.:150.0
##  WHITE                 : 2514   Median : 24.00   Median :170.0
##  BLACK-HISPANIC        : 1409   Mean   : 28.96   Mean   :171.4
##  ASIAN/PACIFIC ISLANDER: 1103   3rd Qu.: 33.00   3rd Qu.:185.0
##  OTHER                 :  298   Max.   :999.00   Max.   :999.0
##  (Other)               :  199
##     haircolr        eyecolor          build              city
##  BLACK  :16221   BROWN  :20023   HEAVY   : 2167   BRONX    :4754
##  BROWN  : 4742   BLACK  : 1458   MEDIUM  :10986   BROOKLYN :6354
##  BALD   :  549   BLUE   :  419   MUSCULAR:  194   MANHATTAN:3941
##  BLOND  :  320   HAZEL  :  216   THIN    : 8880   QUEENS   :5718
##  UNKNOWN:  240   GREEN  :  162   UNKNOWN :  336   STATEN IS:1796
##  GRAY   :  183   UNKNOWN:  156
##  (Other):  308   (Other):  129
##      inside                    location        typeofid
##  Min.   :0.0000   HOUSING AUTHORITY: 3371   OTHER  :  433
##  1st Qu.:0.0000   NEITHER          :18291   PHOTO  :12978
##  Median :0.0000   TRANSIT AUTHORITY:  901   REFUSED:  639
##  Mean   :0.1872                             VERBAL : 8513
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
##      perobs            perstop        arstmade         sumissue
##  Min.   :  0.000   5      :8003   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:  1.000   10     :4792   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :  1.000   3      :1638   Median :0.0000   Median :0.00000
##  Mean   :  2.639   15     :1518   Mean   :0.1759   Mean   :0.02606
##  3rd Qu.:  2.000   2      :1505   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :535.000   1      : 925   Max.   :1.0000   Max.   :1.00000
##                    (Other):4182
##     frisked          searched         contrabn           radio
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.6762   Mean   :0.1863   Mean   :0.04986   Mean   :0.4161
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##
##      height            pf              weap
##  Min.   :36.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:67.00   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :69.00   Median :0.0000   Median :0.00000
```

```
##  Mean   :68.83   Mean   :0.3317   Mean   :0.04822
##  3rd Qu.:71.00   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :95.00   Max.   :1.0000   Max.   :1.00000
##
```

```
dim(sqf2015)
```

```
## [1] 22563    24
```

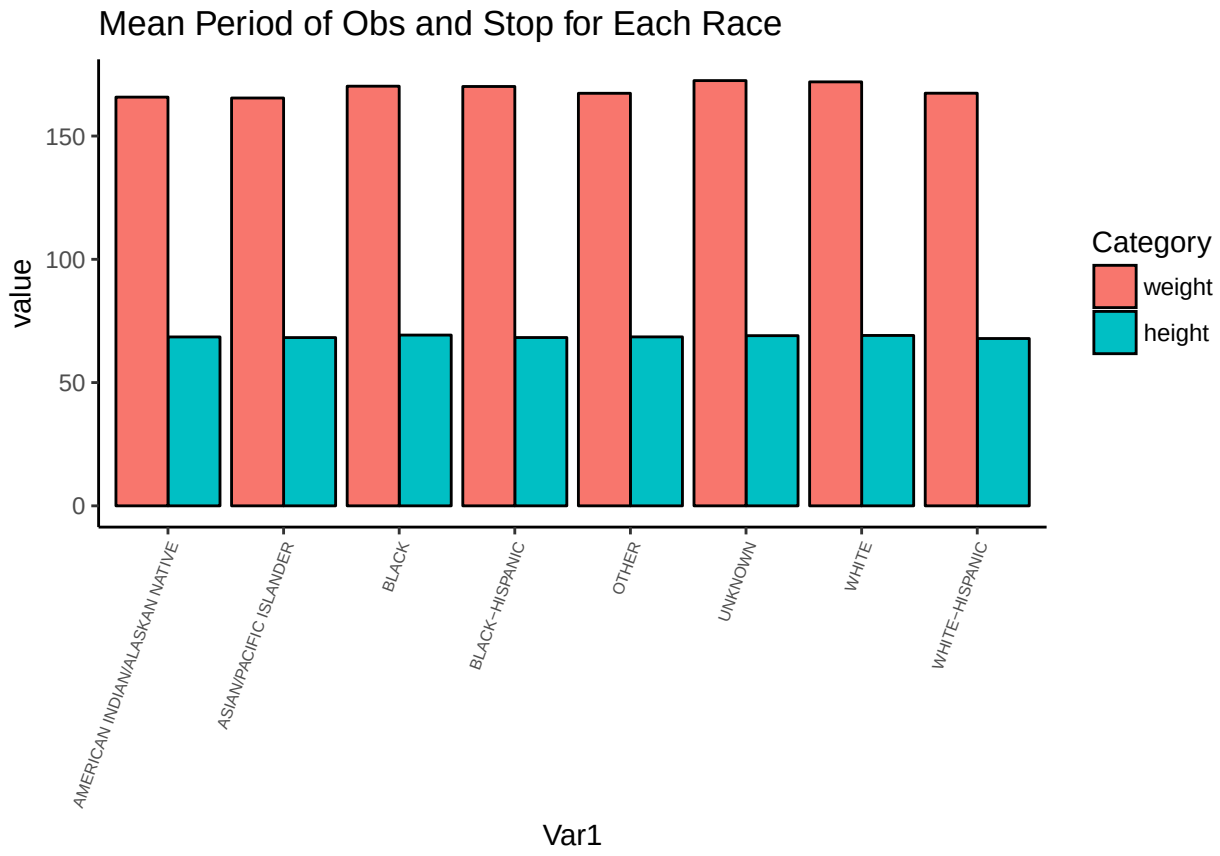
---

3. **Bar Plots and Box Plots**

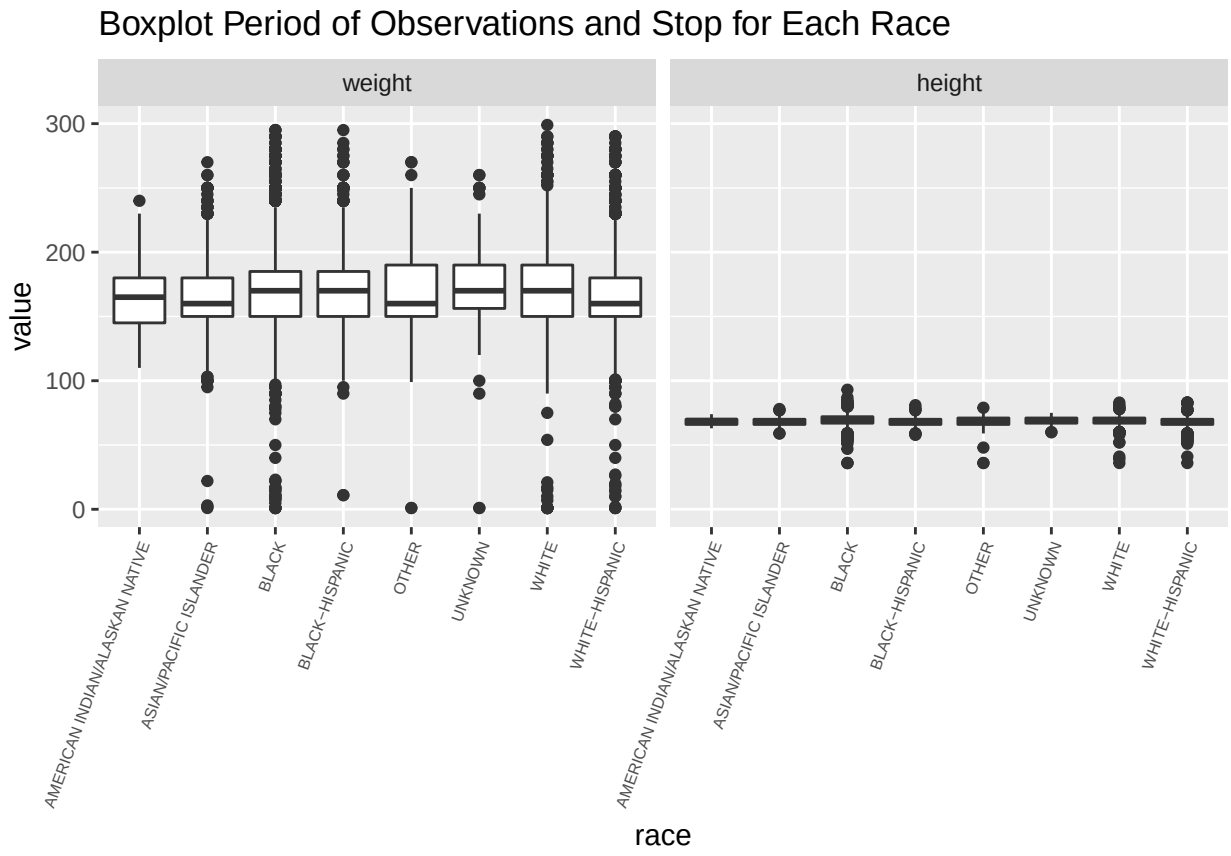
Choose at least two of the quantitative columns and make a comparison bar plot over the categories in another column. Repeat with box plots instead of bar plots.

```
data = sqf2015[,c('race','weight','height')]
data = data[data$weight<300,]
melted = melt(data,id.vars='race')
names(melted)[2] = 'Category'
data = melt(acast(melted, race ~ Category, mean))
names(data)[2] = 'Category'

ggplot(data= data, aes(x=Var1, y=value, fill=Category))+
  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("Mean Period of Obs and Stop for Each Race")
```



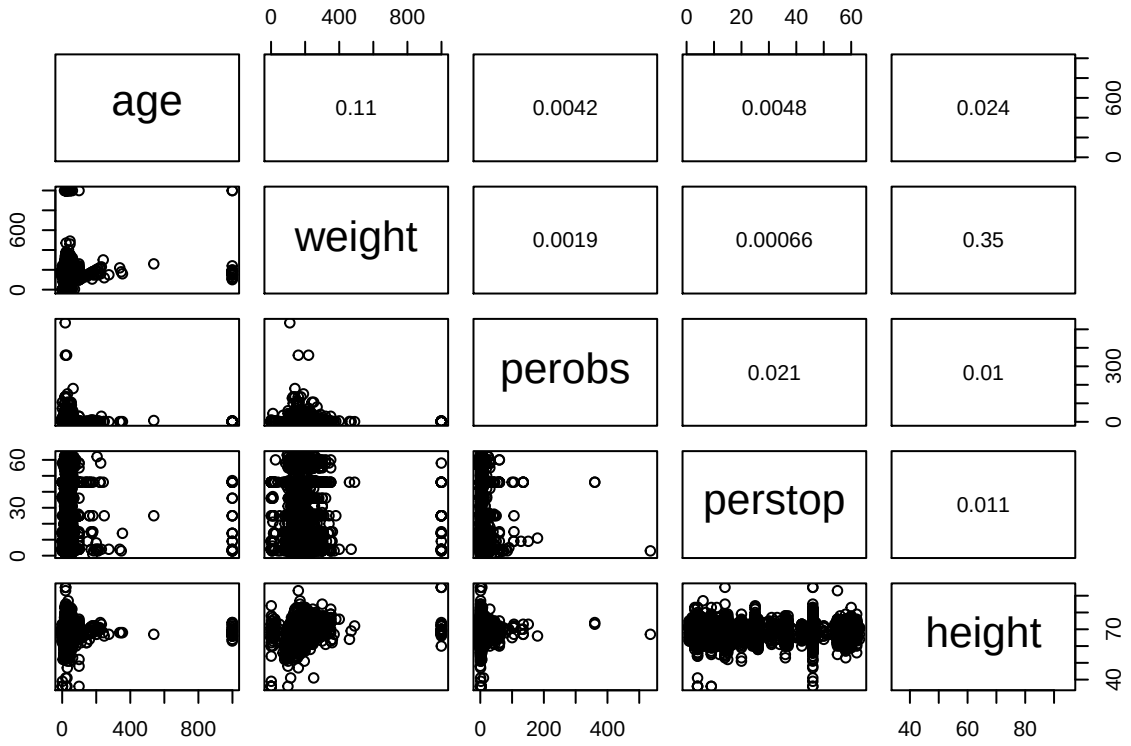

```
ggplot(data=melted, aes(x=race, y=value))+
  geom_boxplot()+
  facet_wrap(~Category)+
  theme(axis.text.x = element_text(angle = 70, hjust = 1, size = rel(0.7) ))+
  ggtitle("Boxplot Period of Observations and Stop for Each Race")
```



4. **Pairwise Plots**

Make pairwise plots, with correlations calculated, of the quantitative columns and determine which columns (if any) seem linearly related. Use linear regression to check the strength of the relationship (see Lecture 11-12 for linear regressions).

```
sqf2015$perstop <- as.numeric(sqf2015$perstop)
panel.pearson <- function(x, y, ...) {
horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
vertical <- (par("usr")[3] + par("usr")[4]) / 2;
text(horizontal, vertical, format(abs(cor(x,y)), digits=2))
}
pairs(as.matrix(sqf2015[,c('age','weight','perobs','perstop','height')])
      , upper.panel=panel.pearson)
```




```
summary(step(lm(formula = weight ~ age + perobs + perstop + height, data = sqf2015)))
```

```
## Start:  AIC=168796.4
## weight ~ age + perobs + perstop + height
##
##           Df Sum of Sq      RSS    AIC
## - perobs   1       205 40014055 168794
## - perstop  1      1139 40014989 168795
## <none>                 40013850 168796
## - age      1    441883 40455732 169042
## - height   1   5454898 45468748 171678
##
## Step:  AIC=168794.5
## weight ~ age + perstop + height
##
##           Df Sum of Sq      RSS    AIC
## - perstop  1      1120 40015175 168793
## <none>                 40014055 168794
## - age      1    441813 40455868 169040
## - height   1   5454784 45468839 171676
##
## Step:  AIC=168793.1
## weight ~ age + height
##
##          Df Sum of Sq      RSS    AIC
## <none>                40015175 168793
## - age     1    441620 40456795 169039
## - height  1   5453727 45468902 171674

##
## Call:
```

```
## lm(formula = weight ~ age + height, data = sqf2015)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -239.66  -17.51   -5.14   12.43  870.64
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.526e+02  5.781e+00  -26.39   <2e-16 ***
## age          1.278e-01  8.101e-03   15.78   <2e-16 ***
## height       4.653e+00  8.391e-02   55.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.12 on 22560 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1298
## F-statistic:  1684 on 2 and 22560 DF,  p-value: < 2.2e-16
```

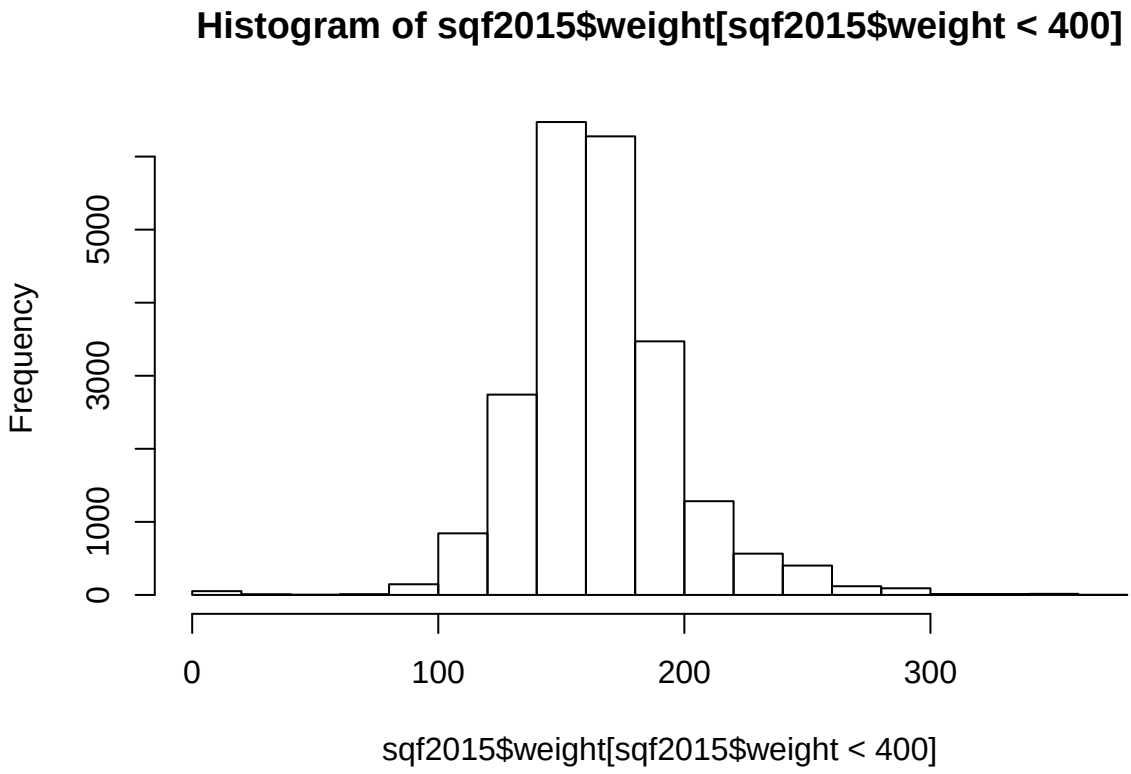
---

5. **Histograms and QQ Plots**

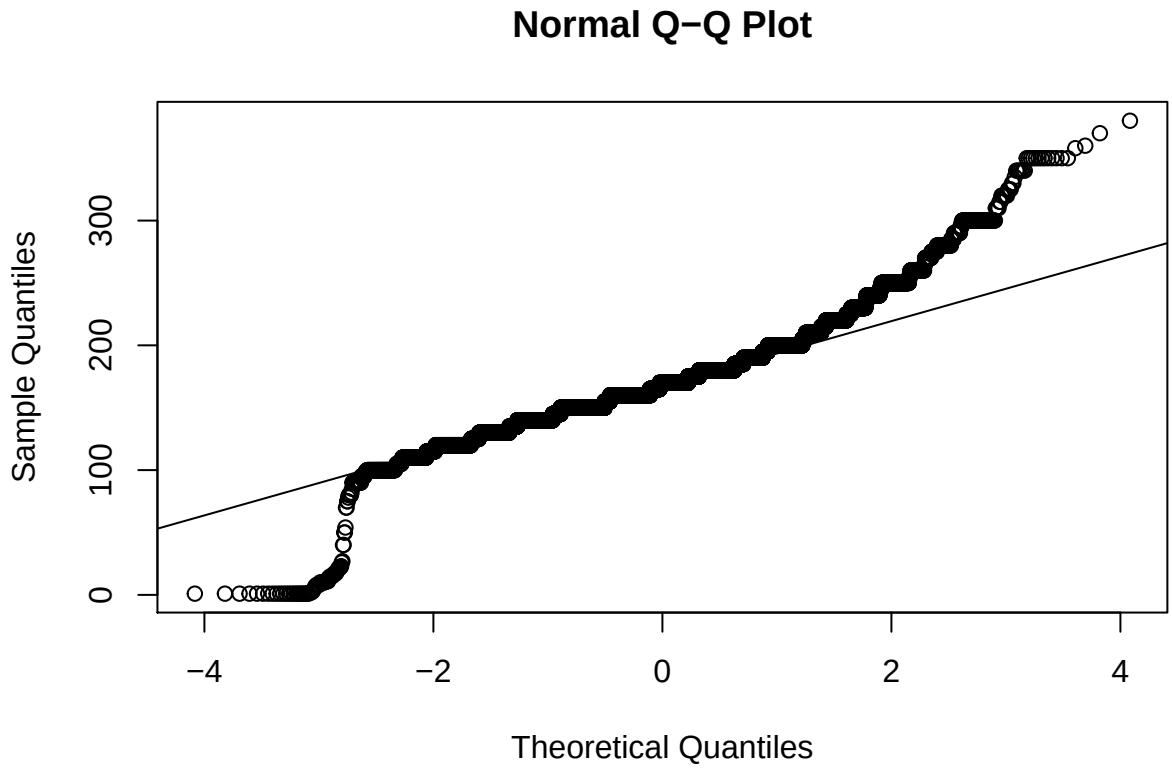
Use histograms of the quantitative columns to determine which, if any, are normally distributed. Use QQ-plots to verify. You may notice that several of the columns have outlier points which you may want to ignore to obtain a histogram that is easier to read. You can use a Boolean expression to filter data so you only keep the reasonable data points. For example, to look at only the ages in the 2010 data that are below 100, we can use:

```
sqf2010$age[sqf2010$age<100]
```

```
hist(sqf2015$weight[sqf2015$weight<400])
```




```
qqnorm(sqf2015$weight[sqf2015$weight<400])
qqline(sqf2015$weight[sqf2015$weight<400])
```



6. **Categorical Data**

For non-quantitative (categorical) data, we will need another way to dissect the data based on categories. We can use the `table()` command to compare one column with another. Type `?table()` to see how to use the command. To compare the number of arrests in 2010 for each race, type:

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

Try some other comparisons with the non-quantitative data to see if you can find any interesting relationships.

```
table(sqf2015$searched,sqf2015$race)
```

```
##
##     AMERICAN INDIAN/ALASKAN NATIVE ASIAN/PACIFIC ISLANDER BLACK
##   0                              65                    921  9806
##   1                              12                    182  2144
##
##     BLACK-HISPANIC OTHER UNKNOWN WHITE WHITE-HISPANIC
##   0           1044   249      97  2138           4039
##   1            365    49      25   376           1051
```

---

7. **Binary Variables**

Several of the non-quantitative binary variables have been expressed as 1 for "Yes" and "0" for no. We will see that this is convenient if we want to compare the results for two subgroups of the SQF data to see if they are different. For example, if we want to compare the arrest rates for people with gray hair versus people with salt and pepper hair, we can first make vectors for each subgroup that tell us whether each person was arrested or not.

```
arrestsGrayHair <- sqf2015$arstmade[sqf2015$haircolr=="GRAY"]
arrestsSPHair <- sqf2015$arstmade[sqf2015$haircolr=="SALT AND PEPPER"]
```

The means for each subgroup now represents the arrest rate for the group (why?). We can now use a two-sample T-test to determine if their means (= arrest rates) are different. Try this on the data with variables of your choice and find at least one pair of (binary variable and two categories) where the rates for the two groups are different with 95% confidence. You may want to refer to Lecture 10 for how to conduct a two-sample T-test.

```
arrestsGrayHair <- sqf2015$arstmade[sqf2015$race=="BLACK"]
arrestsSPHair <- sqf2015$arstmade[sqf2015$race=="WHITE"]
t.test(x=arrestsGrayHair, y=arrestsSPHair, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  arrestsGrayHair and arrestsSPHair
## t = 3.9107, df = 3860.8, p-value = 9.359e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01493239 0.04495663
## sample estimates:
## mean of x mean of y
```

```
## 0.1667782 0.1368337
```

Bibliography

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