File-Name: demographics and crime stats_v1

Version: 1

Date: 03/04/17

Author: Stephanie Langeland

Purpose: Find demographic and crime stats data/clean the data

Input Files: 2015_stopandfrisk_CLEAN_w_counties.csv, crime data_combined_raw2.csv

Output Files: cleaned_2015_crime_data.csv

Data Output: None

Dependencies: None

Required by: Final Prject

Status: In progress

Machine: Stephanie's 2011 MacBook Pro

R version: 3.3.1

Stop and Frisk Data: General:

```
rm(list = ls(all = TRUE))   # cleans everything in the workspace

sfd <- read.csv("/Users/StephanieLangeland/Desktop/Columbia/Applied Data Science/Git/QMSS_G5069_Applied_D_S/Data+Code Book/Cleaned/2015_stopandfrisk_CLEAN_w_counties.csv") # cleaned stop and frisk data

# SPACE FOR SOMEONE ELSE TO COMMENT OUT MY PATH ABOVE AND PUT THEIR PATH IN HERE
```

Stop and Frisk Data: Time period of the data:

```
# convert to date format:
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
summary(sfd$datestop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   101.0   311.0   528.0   595.4   901.0  1231.0
```

```
summary(is.na(sfd$datestop))
```

```
##    Mode   FALSE    NA's
## logical   21747       0
```

```
typeof(sfd$datestop)
```

```
## [1] "integer"
```

```
sfd$datestop <- mdy(sfd[ , 5]) # column 5 is the DATE OF STOP (M-D-YYYY) <- the code book is wrong, it says that the format is (MM-DD-YYY)
```

```
## Warning: All formats failed to parse. No formats found.
```

```
summary(is.na(sfd$datestop))
```

```
##    Mode    TRUE    NA's
## logical   21747       0
```

```
min(sfd$datestop) #confirmed that the data are complete for 2015
```

```
## [1] NA
```

```
max(sfd$datestop)
```

```
## [1] NA
```

```
summary((sfd$dob)) # SUSPECT'S DATE OF BIRTH (CCYY-MM-DD) is all missing -> use `age` va
riable instead
```

```
##    Mode    NA's
## logical   21747
```

Demographic data for the Bronx: General

data source: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF (https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF)

```
#the raw dataset was such a mess that I had to clean it in excel - the follwing is the c
leaned version:


# this dataest is in a really strange format
pop <- read.csv("/Users/StephanieLangeland/Desktop/Columbia/Applied Data Science/Git/QMS
S_G5069_Applied_D_S/Data+Code Book/Cleaned/CLEAN_ACS_15_5YR_DP05.csv")

#pop <- pop[-22, ] # delete duplicate rows
#pop <- pop[-21, ] # delete duplicate rows

# Using the "tp_Race alone or in combination with one or more other races" category:
library(tibble)

demographic <- data.frame(subset(pop[53:58, 1:2]))

typeof(demographic$Estimate)
```

```
## [1] "integer"
```

```
str(demographic)
```

```
## 'data.frame':    6 obs. of  2 variables:
##  $ Subject : Factor w/ 77 levels "Total housing units",..: 75 73 71 68 66 67
##  $ Estimate: Factor w/ 76 levels "0","1,060,732",..: 23 22 33 46 16 57
```

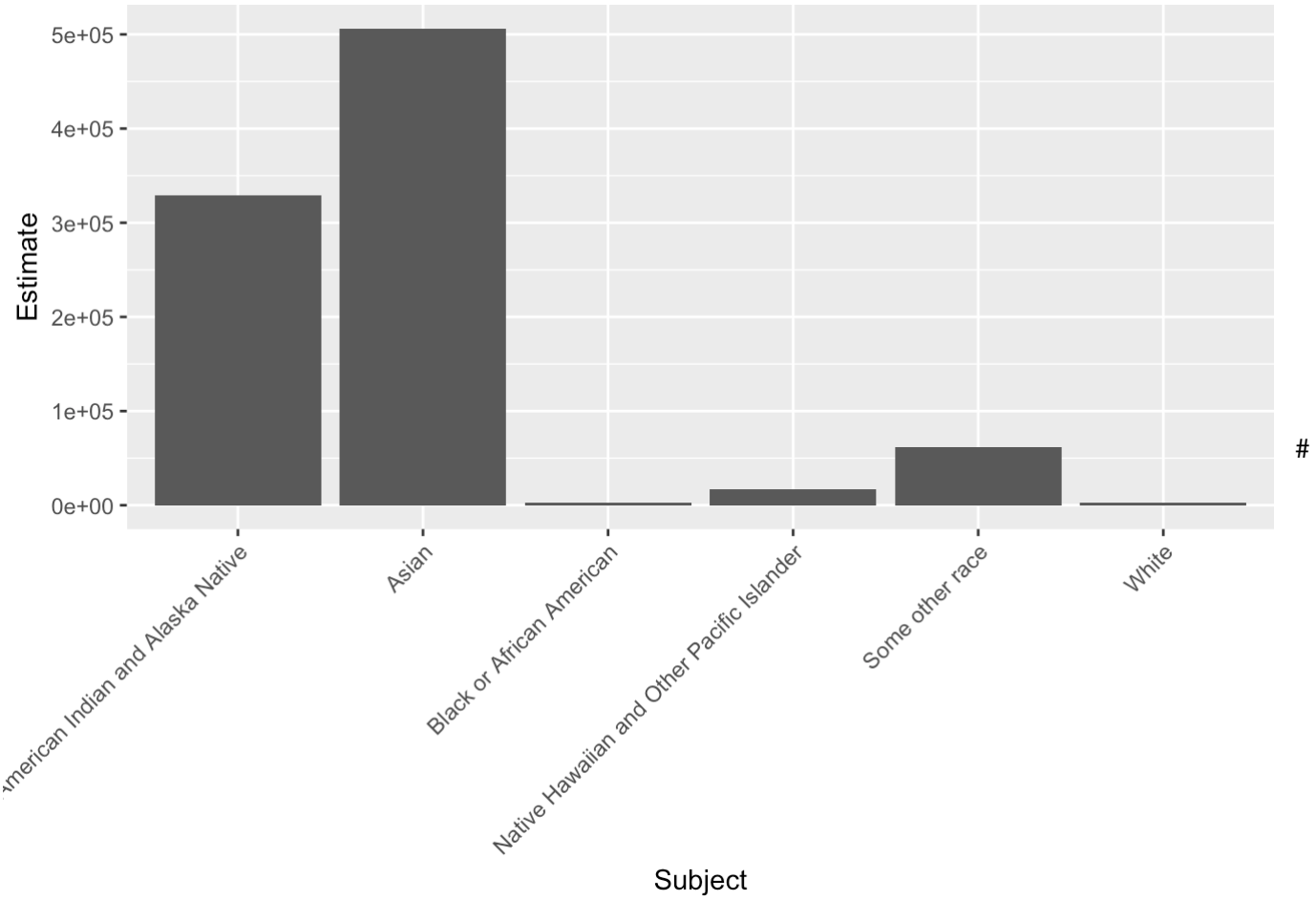
```
demographic$Estimate <- as.numeric(gsub(",", "", as.character(demographic$Estimate)))
demographic
```

```
##
     Subject
## 53                                                          tp_Two or more races_Whi
te and Asian
## 54           tp_Two or more races_Black or African American and American Indian and A
laska Native
## 55                              tp_Race alone or in combination with one or more other
races_White
## 56         tp_Race alone or in combination with one or more other races_Black or Afri
can American
## 57 tp_Race alone or in combination with one or more other races_American Indian and A
laska Native
## 58                              tp_Race alone or in combination with one or more other
races_Asian
##    Estimate
## 53     2884
## 54     2602
## 55   329118
## 56   506514
## 57    16731
## 58    61276
```

```
demographic[ , 1]
```

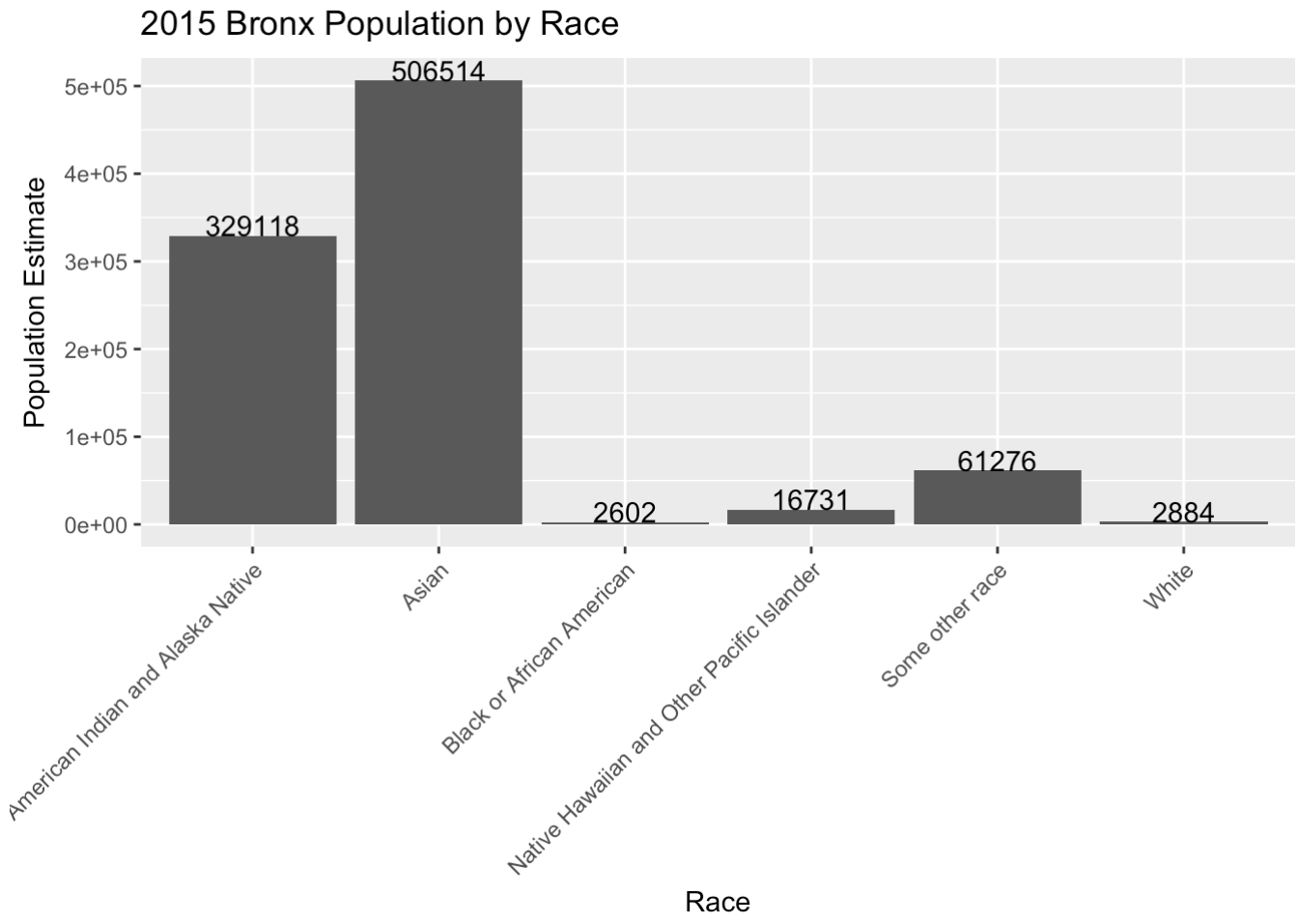
```
## [1] tp_Two or more races_White and Asian

## [2] tp_Two or more races_Black or African American and American Indian and Alaska Nat
ive
## [3] tp_Race alone or in combination with one or more other races_White

## [4] tp_Race alone or in combination with one or more other races_Black or African Ame
rican
## [5] tp_Race alone or in combination with one or more other races_American Indian and
 Alaska Native
## [6] tp_Race alone or in combination with one or more other races_Asian

## 77 Levels: Total housing units Total population ... tp_Under 5 years
```

```
demographic$Subject <- c("White",
              "Black or African American",
              "American Indian and Alaska Native",
              "Asian",
              "Native Hawaiian and Other Pacific Islander",
              "Some other race")

library(ggplot2)
ggplot(demographic, aes(x = Subject, y = Estimate)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Compare stop and frisk events to demographic data:

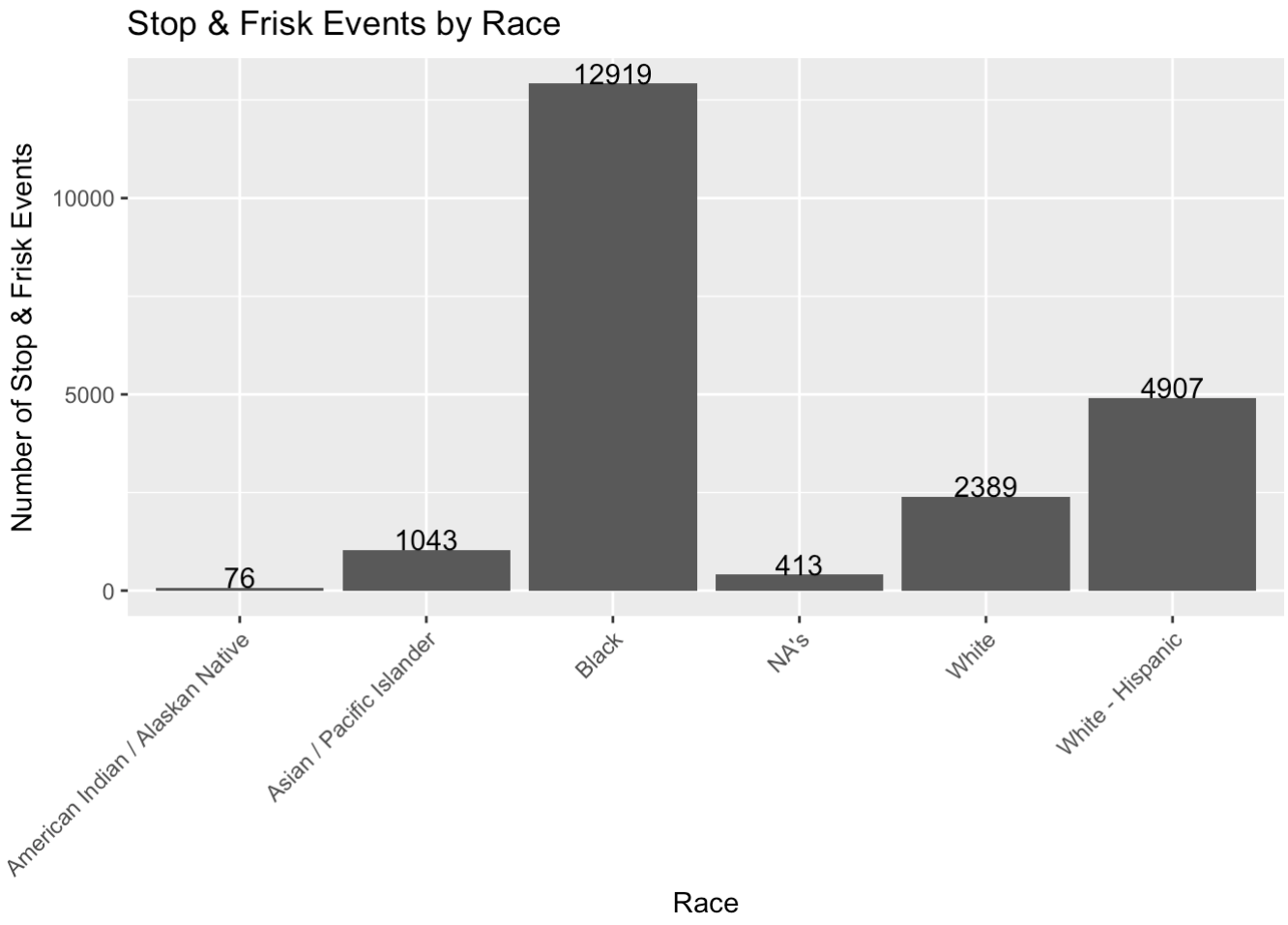
```
ggplot(demographic, aes(x = Subject, y = Estimate)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  xlab("Race") +
  ylab("Population Estimate") +
  ggtitle("2015 Bronx Population by Race") +
  geom_text(aes(label = Estimate), vjust = 0, colour = "black")
```




```
stops_by_race <- sfd[ , "race"]

stops_by_race2 <- summary(stops_by_race)
stops_by_race2 <- as.data.frame(stops_by_race2)
stops_by_race2
```

```
##                                 stops_by_race2
## American Indian / Alaskan Native             76
## Asian / Pacific Islander                   1043
## Black                                     12919
## White                                      2389
## White - Hispanic                           4907
## NA's                                        413
```

```
stops_by_race2$race <- rownames(stops_by_race2)
stops_by_race2$count <- stops_by_race2$stops_by_race2

ggplot(stops_by_race2, aes(x = race, y = count)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ylab("Number of Stop & Frisk Events") +
  xlab("Race") +
  ggtitle("Stop & Frisk Events by Race") +
  geom_text(aes(label = count), vjust = 0, colour = "black")
```



FINAL GRAPHS for 03/08/17 class:

```
#stop and frisk events by race:
stops_by_race_graph <- ggplot(stops_by_race2, aes(x = race, y = count, fill = race)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1)) +
  ylab("Number of Stop & Frisk Events") +
  xlab("Race") +
  ggtitle("Stop & Frisk Events by Race") +
  geom_text(aes(label = count), vjust = 0, colour = "black") +
  scale_color_brewer(palette = "Greens") +
  theme(axis.line = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        #axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        legend.position = "none",
        panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.background = element_blank())

stops_by_race2
```

```
##                                   stops_by_race2
## American Indian / Alaskan Native             76
## Asian / Pacific Islander                   1043
## Black                                     12919
## White                                      2389
## White - Hispanic                           4907
## NA's                                        413
##                                                                    race count
## American Indian / Alaskan Native American Indian / Alaskan Native    76
## Asian / Pacific Islander                 Asian / Pacific Islander  1043
## Black                                                       Black 12919
## White                                                       White  2389
## White - Hispanic                                 White - Hispanic  4907
## NA's                                                         NA's   413
```

```
# new race categories in pop file:
# group pop races as such to match the stop and frisk race categories:
## "tp_One race_American Indian and Alaska Native" - row 31
## "tp_One race_Asian" (row 36) + "tp_One race_Native Hawaiian and Other Pacific Islande
r" (row 44) = one category
## "tp_One race_Black or African American" row 30
## "tp_One race_White" row 29
## "tp_HISPANIC OR LATINO AND RACE_Not Hispanic or Latino_White alone" (row 67) footnot
e: (b) Hispanics may be of any race, so also are included in applicable race categories


demographic2 <- data.frame(subset(pop[c(31, 36, 44, 30, 29, 67), 1:2]))

typeof(demographic2$Estimate)
```

```
## [1] "integer"
```

```
str(demographic2)
```

```
## 'data.frame':    6 obs. of  2 variables:
##  $ Subject : Factor w/ 77 levels "Total housing units",..: 45 50 59 58 65 41
##  $ Estimate: Factor w/ 76 levels "0","1,060,732",..: 63 50 60 41 28 12
```

```
demographic2$Estimate <- as.numeric(gsub(",", "", as.character(demographic2$Estimate)))
demographic2
```

```
##                                                                 Subject
## 31                    tp_One race_American Indian and Alaska Native
## 36                                                 tp_One race_Asian
## 44            tp_One race_Native Hawaiian and Other Pacific Islander
## 30                              tp_One race_Black or African American
## 29                                                 tp_One race_White
## 67 tp_HISPANIC OR LATINO AND RACE_Not Hispanic or Latino_White alone
##    Estimate
## 31     7980
## 36    52457
## 44      666
## 30   475378
## 29   299869
## 67   146928
```

```
demographic2[ , 1]
```

```
## [1] tp_One race_American Indian and Alaska Native
## [2] tp_One race_Asian
## [3] tp_One race_Native Hawaiian and Other Pacific Islander
## [4] tp_One race_Black or African American
## [5] tp_One race_White
## [6] tp_HISPANIC OR LATINO AND RACE_Not Hispanic or Latino_White alone
## 77 Levels: Total housing units Total population ... tp_Under 5 years
```

```
demographic2$Subject <- c("American Indian and Alaska Native",
              "Asian",
              "Native Hawaiian and Other Pacific Islander",
              "Black or African American",
              "White",
              "White - Hispanic")

#combine "Asian" and "Native Hawaiian and Other Pacific Islander" to match races in sfd
 file:
52457 + 666
```

```
## [1] 53123
```

```
demographic3 <- data.frame(Subject = c("Asian/Native Hawaiian and Other Pacific Islande
r"),
                           Estimate = 53123)
demographic4 <- demographic2[-c(2:3) , 1:2]
demographic4
```

```
##                              Subject Estimate
## 31 American Indian and Alaska Native     7980
## 30         Black or African American   475378
## 29                             White   299869
## 67                 White - Hispanic   146928
```

```
demographic4[5, 1:2] <- demographic3
demographic4
```

```
##                              Subject Estimate
## 31 American Indian and Alaska Native     7980
## 30         Black or African American   475378
## 29                             White   299869
## 67                 White - Hispanic   146928
## 1                                 1    53123
```

```
# demographic4[5, 2] <- demographic3[c("Asian/Native Hawaiian and Other Pacific Islande
r"), c("53123")]
demographic4[5, 1] <- "Asian/Native Hawaiian and Other Pacific Islander"
demographic4
```

```
##                                          Subject Estimate
## 31               American Indian and Alaska Native     7980
## 30                       Black or African American   475378
## 29                                           White   299869
## 67                               White - Hispanic   146928
## 1  Asian/Native Hawaiian and Other Pacific Islander    53123
```

```
population_race_graph <- ggplot(demographic4, aes(x = Subject, y = Estimate, fill = Subj
ect)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1)) +
  xlab("Race") +
  ylab("Population Estimate") +
  ggtitle("2015 Bronx Population by Race") +
  geom_text(aes(label = Estimate), vjust = 0.05, colour = "black") +
  scale_color_brewer(palette = "Greens") +
  theme(axis.line = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        #axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        legend.position = "none",
        panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.background = element_blank())

#percentage comparison stop and frisk events / pop race cat:
stops_by_race2
```

```
##                                  stops_by_race2
## American Indian / Alaskan Native             76
## Asian / Pacific Islander                   1043
## Black                                     12919
## White                                      2389
## White - Hispanic                           4907
## NA's                                        413
##                                                                    race count
## American Indian / Alaskan Native American Indian / Alaskan Native    76
## Asian / Pacific Islander                 Asian / Pacific Islander  1043
## Black                                                       Black 12919
## White                                                       White  2389
## White - Hispanic                                   White - Hispanic  4907
## NA's                                                         NA's   413
```

```
demographic4
```

```
##                                           Subject Estimate
## 31                American Indian and Alaska Native     7980
## 30                        Black or African American   475378
## 29                                            White   299869
## 67                                  White - Hispanic   146928
## 1  Asian/Native Hawaiian and Other Pacific Islander    53123
```

```
AIAN <- round(((76 / 7980) * 100), digits = 2) # "American Indian / Alaskan Native"
API <- round(((1043 /   53123) * 100), digits = 2) # "Asian / Pacific Islander"
BLK <- round(((12919    / 475378) * 100), digits = 2) # "Black"
WHT <- round(((2389 / 299869) * 100), digits = 2) # "White"
WH <- round(((4907 / 146928) * 100), digits = 2) # "White - Hispanic"

#MAKE THIS INTO A MATRIX AND WRITE UP A QUICK SUMMARY FOR WEDNESDAY
comparison <- matrix(c("Race", "% of each race who were stopped and frisked",
         "American Indian / Alaskan Native", AIAN,
         "Asian / Pacific Islander", API,
         "Black", BLK,
         "White", WHT,
         "White - Hispanic", WH),
       nrow = ,
       ncol = 2,
       byrow = TRUE)

comparison2 <- as.data.frame(comparison)

comparison2
```

```
##                                 V1
## 1                             Race
## 2 American Indian / Alaskan Native
## 3         Asian / Pacific Islander
## 4                            Black
## 5                            White
## 6                 White - Hispanic
##                                          V2
## 1 % of each race who were stopped and frisked
## 2                                        0.95
## 3                                        1.96
## 4                                        2.72
## 5                                         0.8
## 6                                        3.34
```

```
comparison2 <- comparison2[-1, 1:2]

comparison2
```

```
##                                 V1   V2
## 2 American Indian / Alaskan Native 0.95
## 3         Asian / Pacific Islander 1.96
## 4                            Black 2.72
## 5                            White  0.8
## 6                 White - Hispanic 3.34
```

```
comparison_graph <- ggplot(comparison2, aes(x = V1, y = V2, fill = V1)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1)) +
  xlab("Race") +
  ylab("Percentage") +
  ggtitle("Percentage of Each Race of the Bronx Population Being Stopped & Frisked") +
  geom_text(aes(label = V2), vjust = 0, colour = "black") +
  theme(axis.line = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        #axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        legend.position = "none",
        panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.background = element_blank())
```

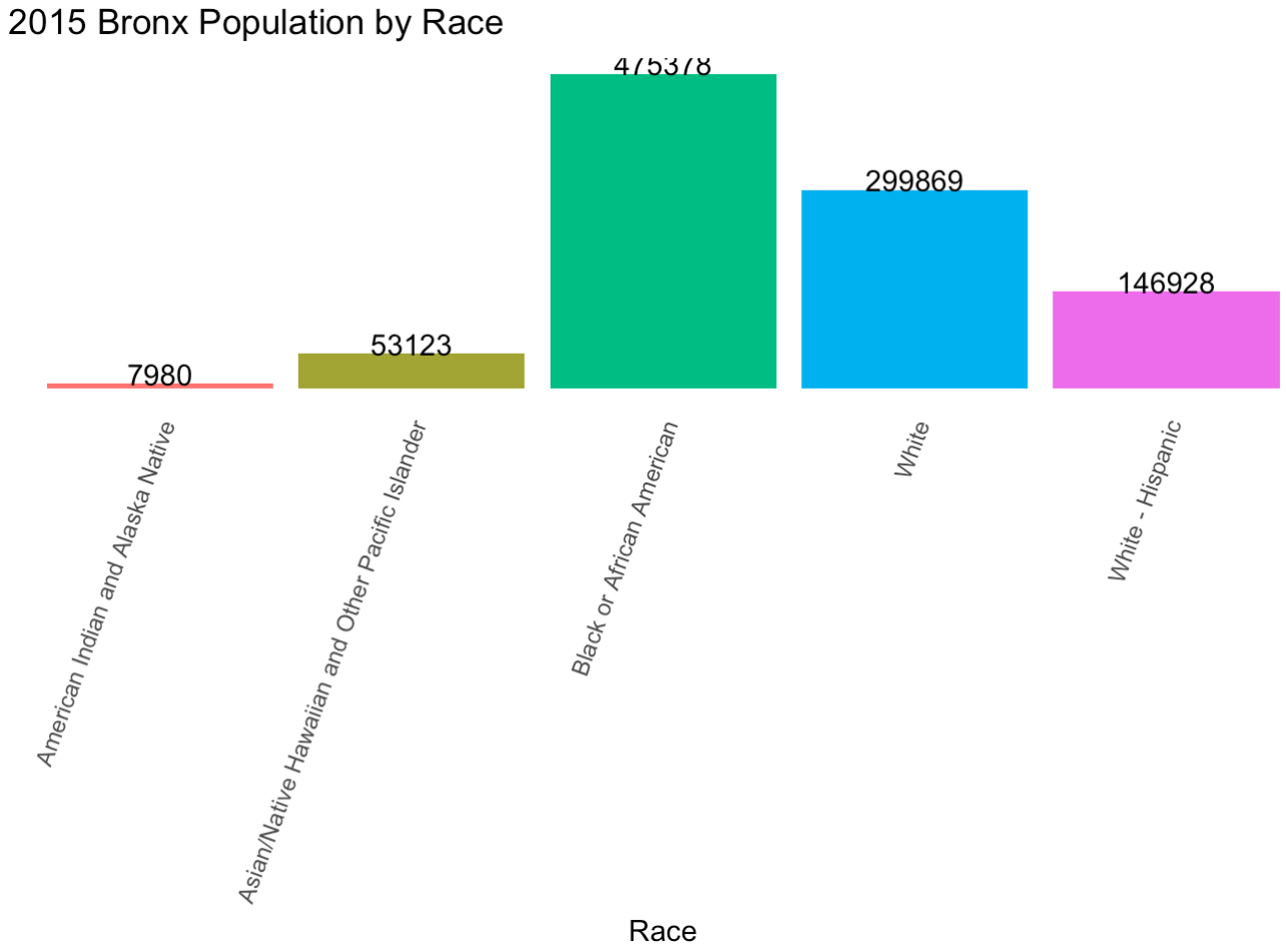
03/08/17 Class Update for Team 3:

1. Gathered demographic data for the Bronx and compared it to stop and frisk events by race. NYPD crime data were also gathered and will be cleaned within the coming weeks to understand the relationship between the reasons police officers

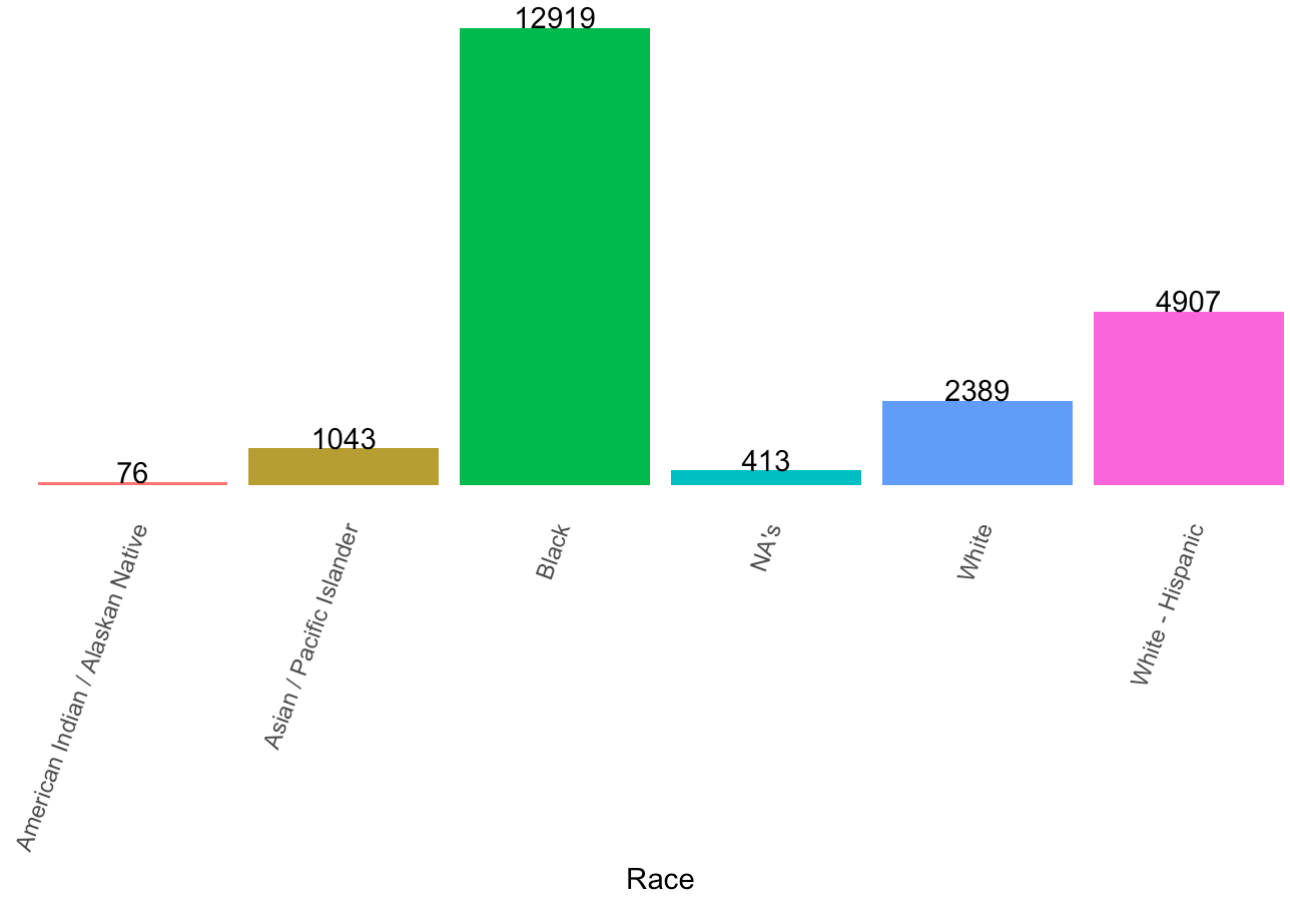
stop and frisk versus population crimes in the Bronx by precinct (to pinpoint location). Next, we will start to build our forecasting models for probability of being stopped and frisked in the Bronx by race, use of force, gender, and other potential variables.

2. The output (graphs) for the week show the breakdown of 5 races by population in the Bronx, stop and frisk events by race, and stop and frisk events by race as a percentage of each race within the Bronx population, respectively. All data are for the Bronx in 2015.

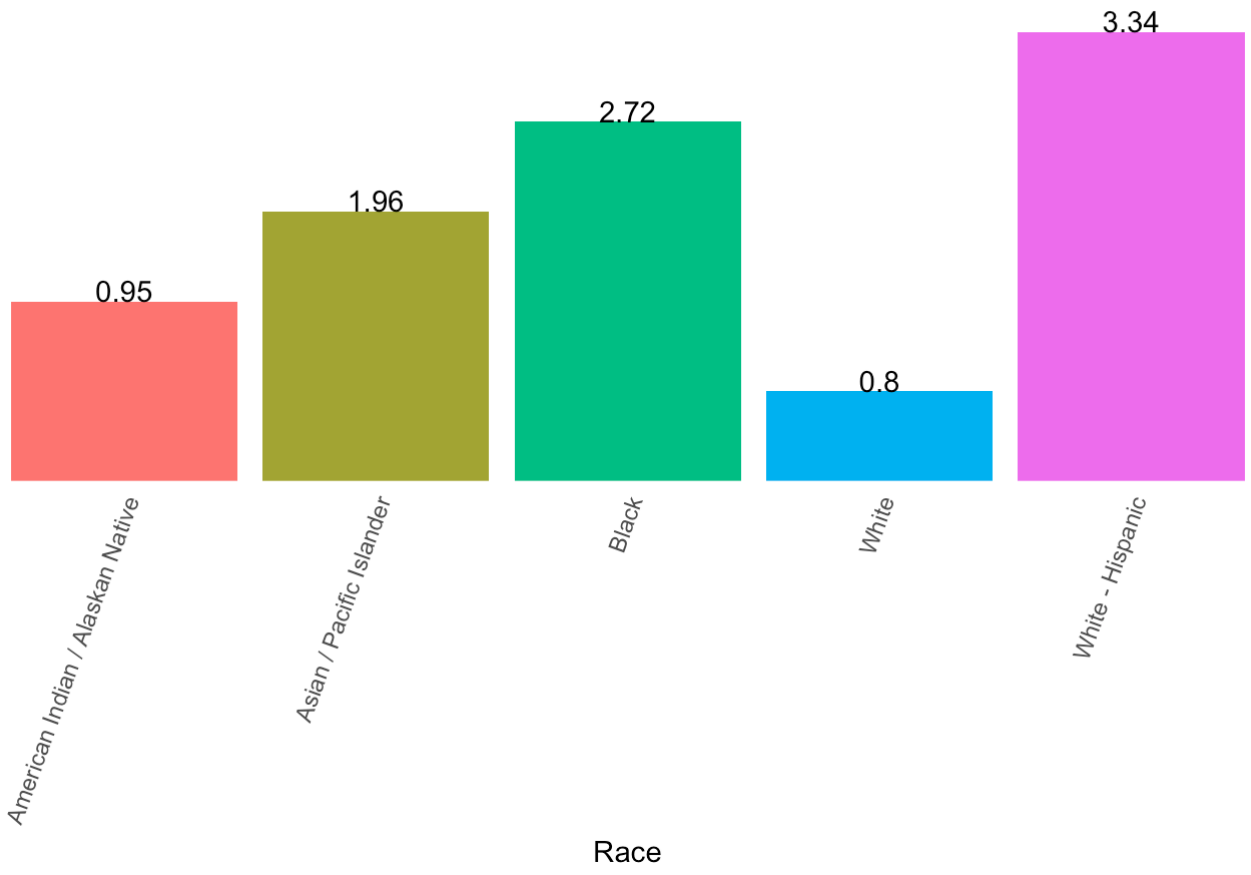
```
population_race_graph
```




```
stops_by_race_graph
```




```
comparison_graph
```



Crime Stats for the Bronx: general:

Clean the data:

https://stackoverflow.com/questions/37509886/how-to-read-merged-excel-cells-with-r
(https://stackoverflow.com/questions/37509886/how-to-read-merged-excel-cells-with-r)

```
rm(list = ls(all = TRUE))   # cleans everything in the workspace

path <- "/Users/StephanieLangeland/Desktop/Columbia/Applied Data Science/Git/QMSS_G5069_
Applied_D_S/Data+Code Book/Raw:Outdated/crime data_combined_raw2.csv"

raw_crime <- read.csv(path)
raw_crime1 <- raw_crime ## create raw_crime1 to alter the data so that raw_crime stays i
n its original form

raw_crime1 <- raw_crime1[!grepl("TOTAL", raw_crime1$CRIME),] ## remove the rows with tot
als

raw_crime_2015 <- raw_crime1 ## create a dataset to store only 2015 data

raw_crime_2015 <- raw_crime_2015[ , -c(4:18)] ## remove all years that are not 2015

colnames(raw_crime_2015)[4] <- "crimes_count" ## since all data are form 2015, rename ra
w_crime_2015$X2015

colnames(raw_crime_2015)[3] <- "crime" ## rename raw_crime_2015$CRIME with lowercase col
umn heading for consistency

colnames(raw_crime_2015)[2] <- "pct" ## rename raw_crime_2015$PCT with lowercase column
 heading for consistency

cleaned_2015_crime_data <- write.csv(raw_crime_2015, file = "cleaned_2015_crime_data.cs
v")
```