w271 Lab1

K Iwasaki

September 29, 2017

Contents

0. Introduction

Objectives of this project are to create statistically models that incorporate the relationship between voters' preference and dependent variables including age, gender, and race, and to extract insights from the modeling exercise for the client who is interested in selling T-shirts to voters who are likely to support politically liberal candidates.

We are given the data-set from a political survey conducted in January of 2016 and is able to identify voters who preferred Bernie Sanders over Hillary Clinton (1 = Likes Bernie more than Clinton; 0 = Likes Clinton more than Bernie). In addition, this (extremely simple) data-set contains information on respondents':

- Party affiliation (1 if Democrat , 2 if Independent or Other, and 3 if Republican);
- Race (1 if white, 0 otherwise);
- Gender (2 if female, 1 if male);
- and Birthyear.

1. Set-up

Before diving into the analysis, we look at the data-set at high level. Specifically, we check summary statistics, variable categories (categorical, continuous and etc), NA values in each column, and distribution for each column.

Some observations from the intial exploration:

- There are 1200 examples in the date-set.
- There are 9 NA values in the preference columns.Removed rows with NA values

- Correlation matrix shows that race_white and party are associated with sanders_preference while other independent variables have very weak correlation with sanders_preference. Also note that dependent variables don't show strong correlation among them except race_white and party, and race_white and birthyr
- 57.2% of them prefer Bernie Sanders over Hillary Clinton.
- 72.9% of them are white.
- 52.5% of them are male and the rest are female.
- Their age are median 48 and mean 48.Since Min 19 and Max 95, it seems there is no outliers.
- It's important to make sure that the 1200 examples are representative of the population our client is interested in and that they are randomly sampled. Otherwise, the inference we make in the following sections are invalid.

```r
df = read.csv("public_opinion.csv")

head(df)
```

```
##   sanders_preference party race_white gender birthyr
## 1                  1     1          1      1    1960
## 2                  0     2          1      2    1957
## 3                  1     3          1      1    1963
## 4                  1     1          1      1    1980
## 5                  1     2          1      1    1974
## 6                  1     2          1      1    1958
```

```r
nrow(df)
```

```
## [1] 1200
```

```r
# summary stats
summary(df)
```

```
##  sanders_preference     party         race_white         gender     
##  Min.   :0.000      Min.   :1.000   Min.   :0.0000   Min.   :1.000  
##  1st Qu.:0.000      1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000  
##  Median :1.000      Median :2.000   Median :1.0000   Median :2.000  
##  Mean   :0.576      Mean   :1.851   Mean   :0.7292   Mean   :1.525  
##  3rd Qu.:1.000      3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:2.000  
##  Max.   :1.000      Max.   :3.000   Max.   :1.0000   Max.   :2.000  
##  NA's   :9                                                          
##     birthyr    
##  Min.   :1921  
##  1st Qu.:1955  
##  Median :1968  
##  Mean   :1968  
##  3rd Qu.:1982  
##  Max.   :1997  
## 
```

```r
# variable categories
str(df) # notifce they are not factors
```

```
## 'data.frame':    1200 obs. of  5 variables:
##  $ sanders_preference: int  1 0 1 1 1 1 1 1 1 1 ...
##  $ party             : int  1 2 3 1 2 2 1 3 2 1 ...
##  $ race_white        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ gender            : int  1 2 1 1 1 1 1 1 1 1 ...
##  $ birthyr           : int  1960 1957 1963 1980 1974 1958 1978 1951 1973 1936 ...
```

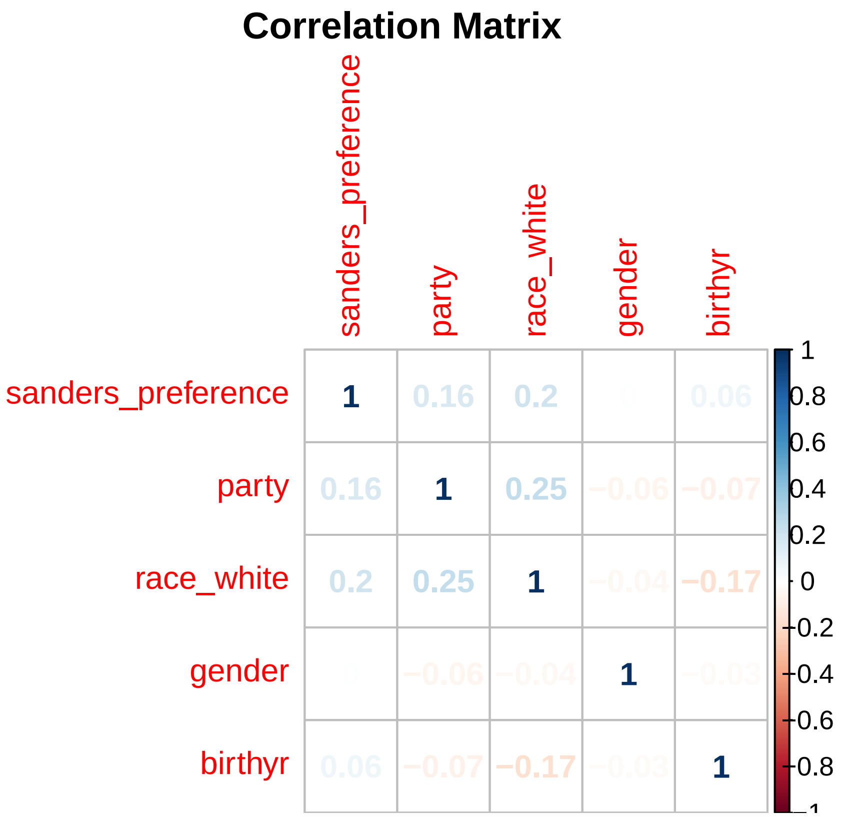
```r
# check NA values
apply(is.na(df), 2, sum)
```

```
## sanders_preference              party         race_white 
##                  9                  0                  0 
##             gender            birthyr 
##                  0                  0
```

```r
# investigate NA values further
df[is.na(df$sanders_preference),]
```

```
##      sanders_preference party race_white gender birthyr
## 448                  NA     1          1      2    1961
## 601                  NA     1          1      2    1950
## 844                  NA     1          0      2    1954
## 887                  NA     2          1      2    1959
## 989                  NA     2          1      2    1970
## 1011                 NA     1          0      1    1965
## 1026                 NA     3          1      2    1973
## 1098                 NA     2          1      2    1992
## 1162                 NA     3          0      2    1962
```

```r
# drop rows with NA
df = df %>%
  filter(!is.na(sanders_preference))

# plot correlation matrix
par(oma=c(0,0,2,0))
corrplot(cor(df), method = "number", title = "Correlation Matrix", mar = c(2, 0, 1, 0))
```



```
# conver columns into factor and new columns
df$race_white = factor(df$race_white)
df$party = factor(df$party)
df$gender = factor(df$gender)
df$male = factor(ifelse(df$gender == 1, 1, 0)) # male == 1, female == 0
df$preference = factor(ifelse(df$sanders_preference == 1, "Sanders", "Clinton"))
df$bi.party = factor(ifelse(df$party == 1, 1, 0)) # democrat == 1, non-democrat == 0

# confirm the change
str(df)
```

```
## 'data.frame':    1191 obs. of  8 variables:
##  $ sanders_preference: int  1 0 1 1 1 1 1 1 1 1 ...
##  $ party             : Factor w/ 3 levels "1","2","3": 1 2 3 1 2 2 1 3 2 1 ...
##  $ race_white        : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ gender            : Factor w/ 2 levels "1","2": 1 2 1 1 1 1 1 1 1 1 ...
##  $ birthyr           : int  1960 1957 1963 1980 1974 1958 1978 1951 1973 1936 ...
##  $ male              : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 2 ...
##  $ preference        : Factor w/ 2 levels "Clinton","Sanders": 2 1 2 2 2 2 2 2 2 2 ...
##  $ bi.party          : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 2 1 1 2 ...
```

```
# check
table(df$party)
```

```
##
##   1   2   3
## 455 458 278
```

```
# create column age
df$age = 2016 - df$birthyr # since the poll was conducted in January 2016

# get stats for the age column
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   33.50   48.00   48.04   61.50   95.00
```

2. Model

a. Description of the model

We build a model to estimate probability π of a respondent being a Sanders supporter. Logit transformation is used to fit the binary outcome variable as a dependent variable (1 being a Sanders supporter and 0 being a Clinton supporter) in the model.

$$log(odds) = 0.557 - 0.013 age + 0.670 bi.party + 0.865 race\_white$$

As an independent variable, we have age, bi.party and race_white in the model. Each variable has a coefficient and its sign indicates an association with the dependent variable. Lastly, we have 0.557 as an intercept in the model.

b. Description of variables in the model

We examine variables that are included and are not included in the model one by one. Below is a quick summary.

- Gender — Not included in the model because there is no evidence that gender is associated with the preference.
- Race — Included in the model since white respondents prefer Sanders than non-white respondents do.
- Party — Included in the model. Democratic respondents less prefer Sanders than non-democratic respondents.
- Age — Included in the model because Sanders supporters are younger than Clinton supporters by 2 years with statistical significance.
- Interaction term: party:race_white — Not included in the model
- Interaction term: age:race_white — Not included in the model
- Interaction term: party: age — Not included in the model

Gender — Not included in the model

We inspected the variable by visualization and t.test. With the following observations, we decided NOT to include the variable age into the model.

- Previous correlation matrix shows that there is little correlation between gender and sanders_preference.
- 57.4% of males prefer Sanders while 57.7% of females prefer Sanders. There is no significant evidence to conclude there is a difference in the two proportions. Also the practical significance is small.

```
table(df$sanders_preference, df$male)
```

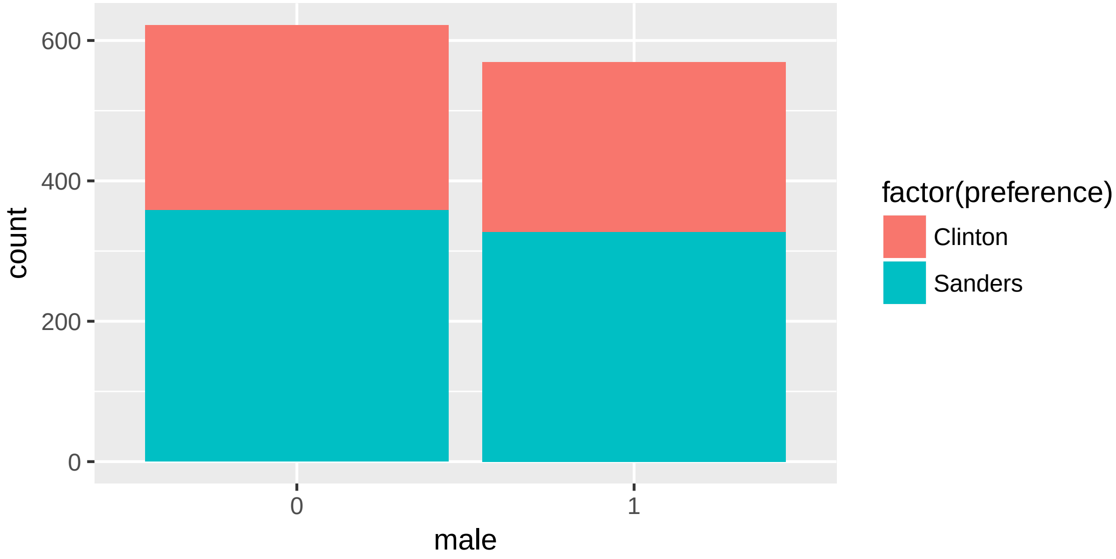
```
##
##       0   1
##   0 263 242
##   1 359 327
```

```
prop.table(table(df$sanders_preference, df$male))
```

```
##
##             0         1
##   0 0.2208228 0.2031906
##   1 0.3014274 0.2745592
```

```
ggplot(df, aes(x = male, fill = factor(preference))) + geom_bar()

# conduct t.test
male = df[df$male == 1, ]$sanders_preference
female = df[df$male == 0, ]$sanders_preference

t.test(male, female)
```

```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = -0.086361, df = 1179.4, p-value = 0.9312
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05877358  0.05381763
## sample estimates:
## mean of x mean of y
## 0.5746924 0.5771704
```



Race – Included in the model

We decide to include race_white variable as a result of the inspection as below. There are some notes:

- 63.7% of White respondents prefer Sanders while 41% of non-white respondents prefer Sanders. T.test results show that the difference between white and non-white group is statistically and practically significant.
- This validates the previous result of the correlation matrix.
- Will follow up on potential interaction effect of the variable with age and party variable.

```
# 1. White 0. otherwise

table(df$sanders_preference)
```

```
##
##   0   1
## 505 686
```

```
table(df$sanders_preference, df$race_white)
```

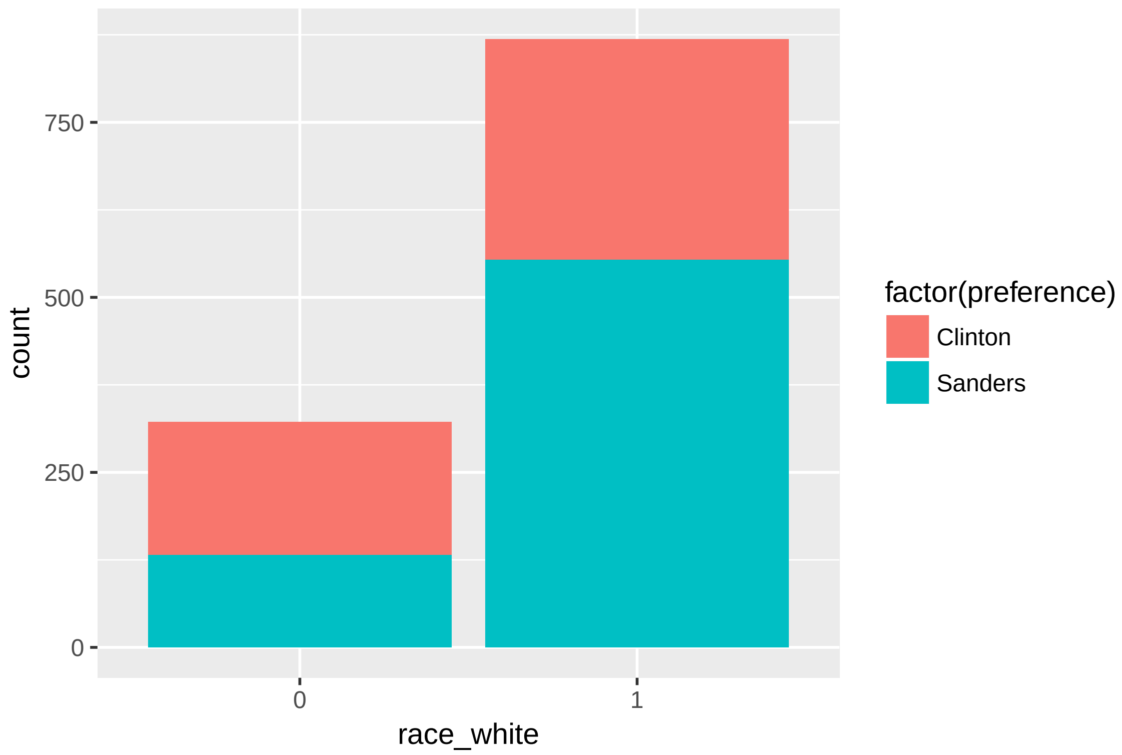
```
##
##       0   1
##   0 190 315
##   1 132 554
```

```
prop.table(table(df$sanders_preference, df$race_white))
```

```
##
##             0         1
##   0 0.1595298 0.2644836
##   1 0.1108312 0.4651553
```

```
ggplot(df, aes(x = race_white, fill = factor(preference))) + geom_bar()

# conduct t.test
white = df[df$race_white == 1, ]$sanders_preference
non_white = df[df$race_white == 0, ]$sanders_preference

t.test(white, non_white)
```

```
##
##  Welch Two Sample t-test
##
## data:  white and non_white
## t = 7.1265, df = 561.95, p-value = 3.174e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1648519 0.2903011
## sample estimates:
## mean of x mean of y
## 0.6375144 0.4099379
```



Party – Included in the model as bi.party varible (1 as democrat, 0 as non-democrat)

We decide to include party variable in the model.

- Democratic voters (party1) shows clearly lower preference for Sanders compared to Independent(party2) and Republican(party3) voters. Average 45% of democratic voters prefer Sanders while about 65% of Independent and Republican voters prefer Sanders respectively. The differences are statistically significant as well according to the t.test below.
- To simplify the model and its interpretation later, create variables with binary values: democrat(1) or non-democrat(0)

```
# 1. Democrat, 2. Independent or other 3. Republican

table(df$sanders_preference, df$party)
```

```
##
##       1   2   3
##   0 249 156 100
##   1 206 302 178
```

```
prop.table(table(df$sanders_preference, df$party))
```

```
##
##              1          2          3
##   0 0.20906801 0.13098237 0.08396306
##   1 0.17296390 0.25356843 0.14945424
```

```
ggplot(df, aes(x = party, fill = factor(preference))) + geom_bar()

# t.test
```

```r
party1 = df[df$party == 1,]$sanders_preference
party2 = df[df$party == 2,]$sanders_preference
party3 = df[df$party == 3,]$sanders_preference

t.test(party1, party2)
```

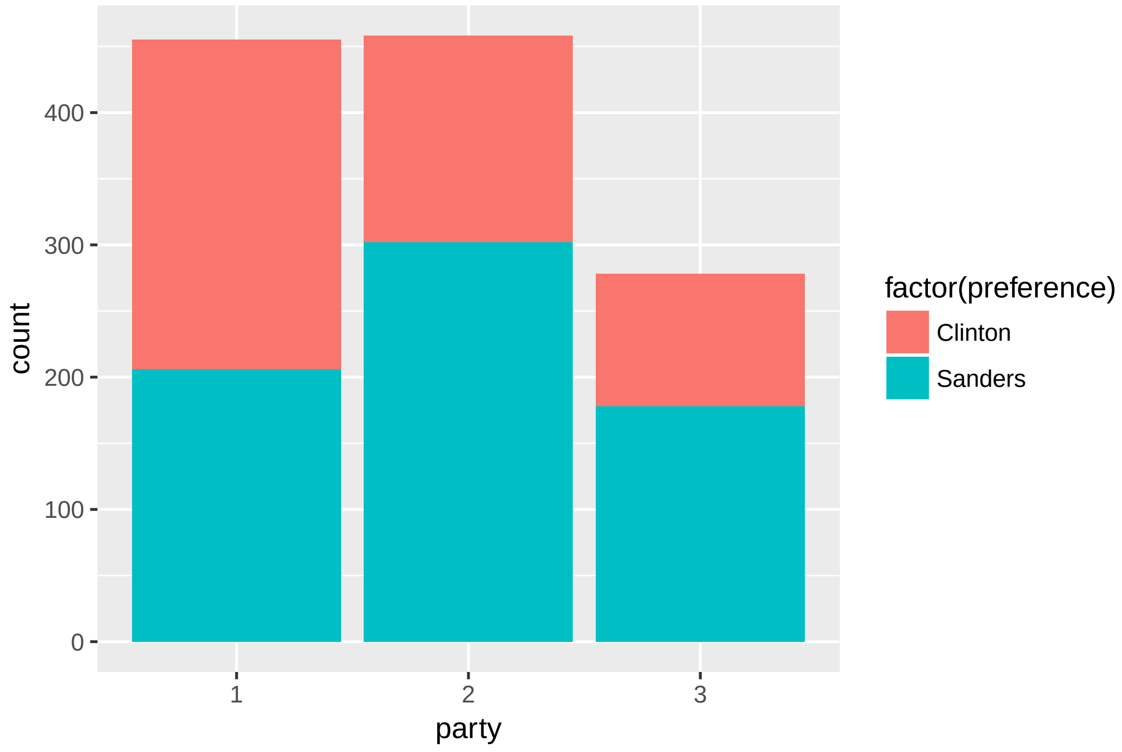
```
##
##  Welch Two Sample t-test
##
## data:  party1 and party2
## t = -6.4163, df = 908.19, p-value = 2.245e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2698474 -0.1434354
## sample estimates:
## mean of x mean of y
## 0.4527473 0.6593886
```

```r
t.test(party2, party3)
```

```
##
##  Welch Two Sample t-test
##
## data:  party2 and party3
## t = 0.52515, df = 578.68, p-value = 0.5997
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05233664  0.09053840
## sample estimates:
## mean of x mean of y
## 0.6593886 0.6402878
```

```r
t.test(party1, party3)
```

```
##
##  Welch Two Sample t-test
##
## data:  party1 and party3
## t = -5.0535, df = 601.78, p-value = 5.762e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2604232 -0.1146579
## sample estimates:
## mean of x mean of y
## 0.4527473 0.6402878
```



Age – Included in the model

Recall that the correlation matrix shows that there is no strong correlation between age and the dependent variable and there is a negative correlation between race_white and age. We observe as followings through the analysis.

- The t.test shows that average age of Sanders supports is Sanders supporters are on average younger than Clinton supporters by two years with statistical significance.
- Effect size, the two-year different, might cause different interpretations that this is large or small. I would argue this is small because the survey respondents distribute from age 19 to 95. Two-year difference is no significant.
- I keep age variable in the model because this variable is particular interest of the client.

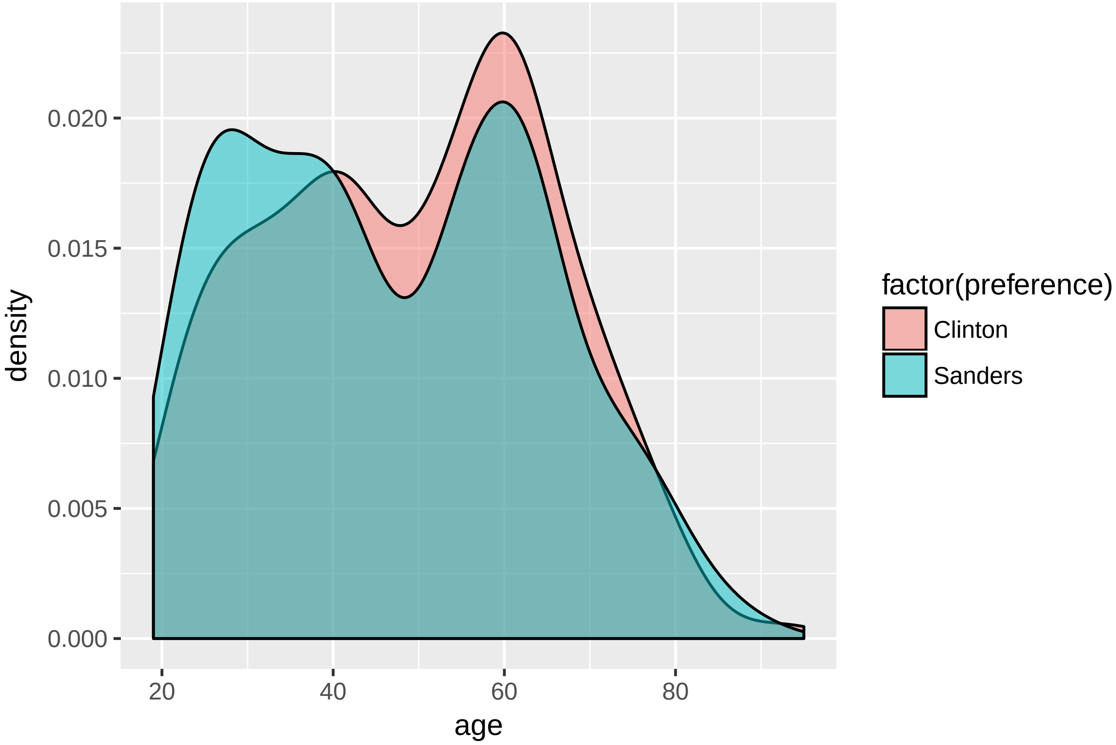
```
ggplot(df, aes(x = age, fill = factor(preference))) +
  geom_density(alpha = 0.5)

# binning age
df$bin_age = .bincode(df$age, c(18, 30, 40, 50, 60, 70, 100), TRUE)

# check the distribution of age across the bins
table(df$bin_age)
```

```
##
##   1   2   3   4   5   6
## 244 208 172 238 210 119
```

```
# indepedent t-test

sanders_age = df[df$sanders_preference == 1,]$age
clinton_age = df[df$sanders_preference == 0,]$age
```

```
t.test(sanders_age, clinton_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  sanders_age and clinton_age
## t = -2.1991, df = 1116.5, p-value = 0.02808
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.1128247 -0.2342468
## sample estimates:
## mean of x mean of y
##  47.11953  49.29307
```



Interaction term: party x race_white — Not included in the model

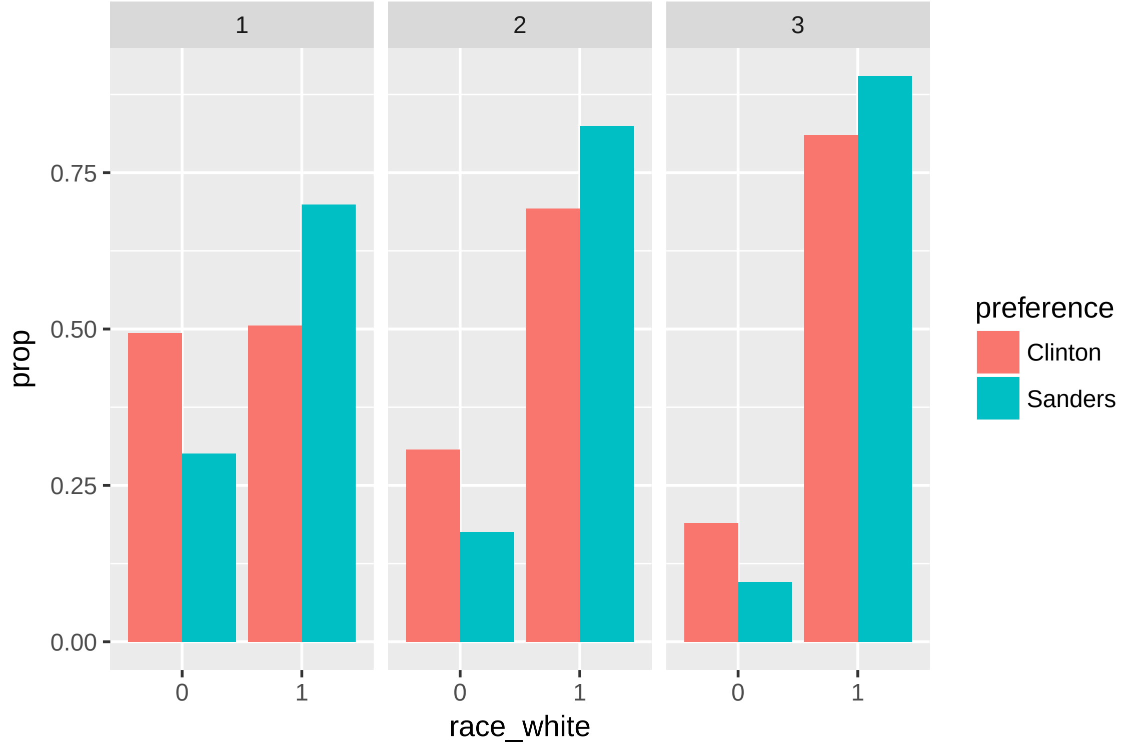
Move onto investigate interaction terms: we focus on look for particular segment of voters that shows significantly difference in terms of the preference.

White voters consistently support Sanders across parties and non-white voters consistently support Clinton. There is no particular segment of voters show difference. Thus we don't observe interaction effect here.

```
df$sanders_preference = factor(df$sanders_preference)

ftable(df %>% select(race_white, party, sanders_preference))
```

```
##                    sanders_preference   0   1
## race_white party
## 0          1                          123  62
##             2                         48  53
##             3                         19  17
## 1           1                        126 144
##             2                        108 249
##             3                         81 161
```

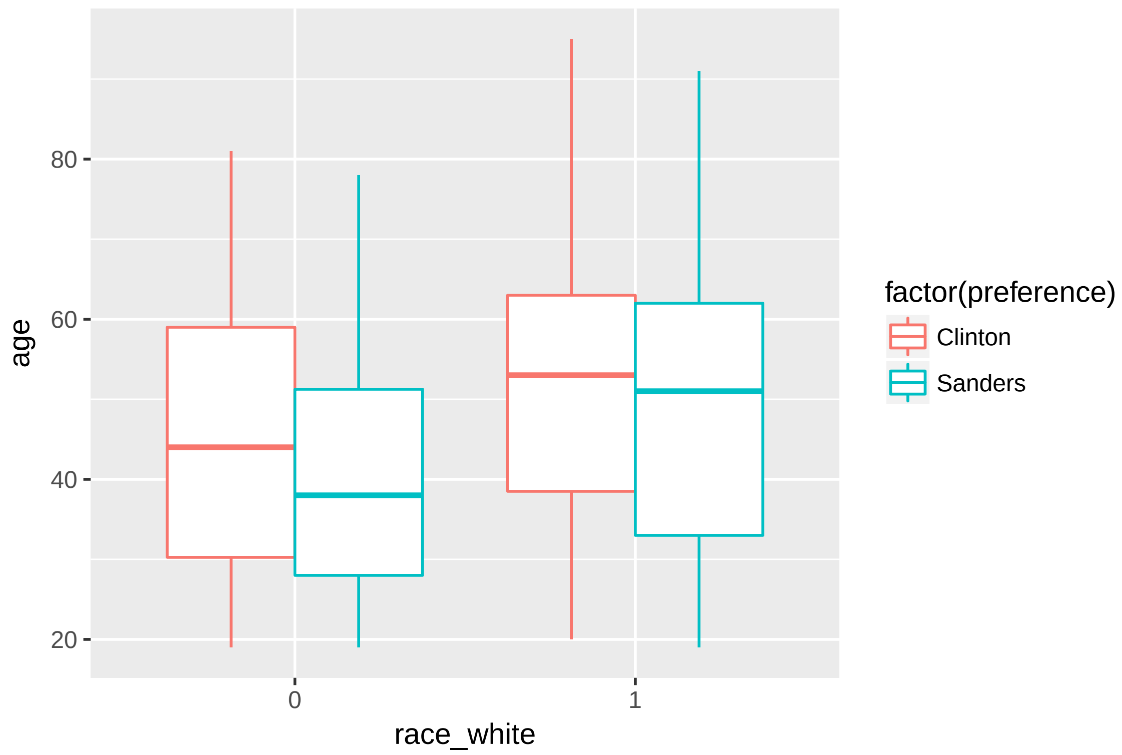
```r
ggplot(df, aes(x =race_white, y = ..prop.., group = preference, fill = preference)) +
  geom_bar(position = "dodge") +
  facet_grid(.~party)
```



Interaction term: age x race_white — Not included in the model

It looks Sanders supporters are younger than their opponents in each race group. The age gap between Sanders supporters and Clinton supporters in the non-white race group is larger than the one in the white race group. This combination might be a candidate for an interaction term.

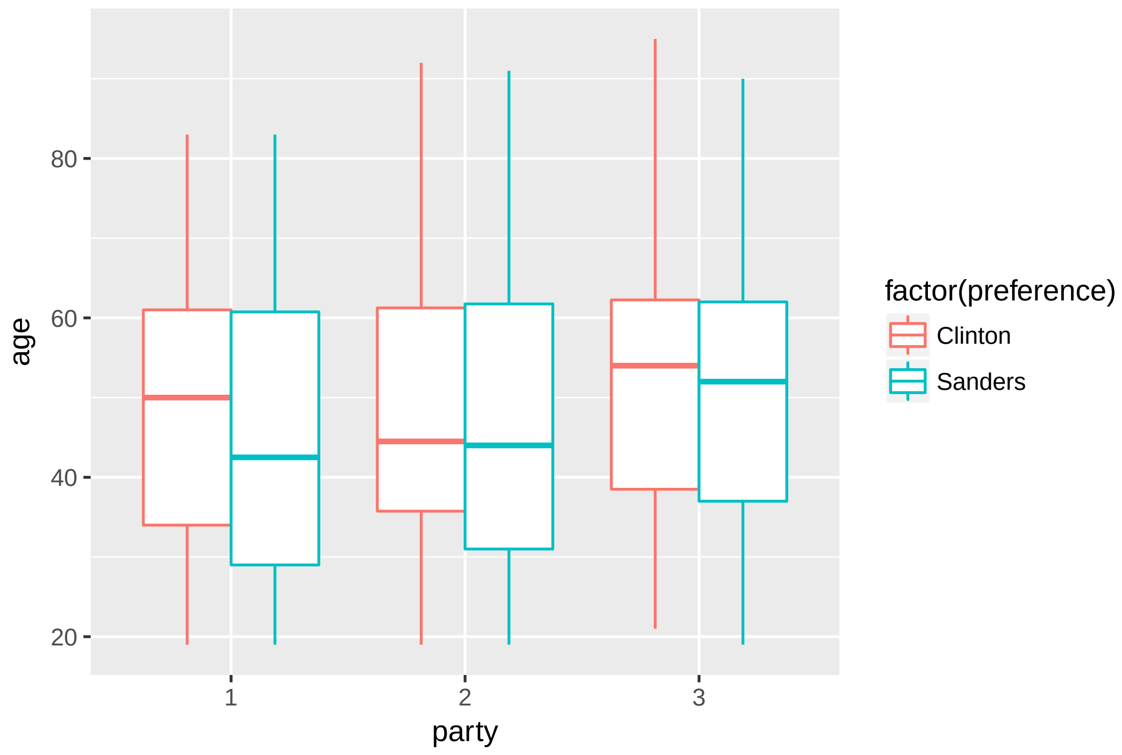
```r
ggplot(df, aes(x = race_white, y = age, col = factor(preference))) + geom_boxplot()
```



Interaction term: party x age — Included in the model

It is interesting to observe that in the Democratic voters (party1) shows the largest age gap between Sanders supporters and Clinton supports. This combination is a good candidate for an interaction term.

```
ggplot(df, aes(x = party, y = age, col = factor(preference))) + geom_boxplot()
```



c. Comparison with other candidate models

We consider four models based on the exploratory data analysis that was shown in the previous section. Below chart shows the summary.

```r
base = glm(sanders_preference ~ age, family = "binomial", data = df)

mod.glm = glm(sanders_preference ~ age + bi.party + race_white, family = "binomial", data = df)

mod.glm.interaction1 = glm(sanders_preference ~ age + bi.party + race_white + age:bi.party,
                           family = "binomial", data = df)

mod.glm.interaction2 = glm(sanders_preference ~ age + bi.party + race_white + age:race_white,
                           family = "binomial", data = df)

summary(mod.glm)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + bi.party + race_white,
##     family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6907  -1.1821   0.7904   0.9892   1.6669
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.556729   0.207245   2.686 0.007224 **
## age         -0.012688   0.003655  -3.472 0.000518 ***
## bi.party1   -0.669732   0.126733  -5.285 1.26e-07 ***
## race_white1  0.865223   0.141409   6.119 9.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1533.7  on 1187  degrees of freedom
## AIC: 1541.7
##
## Number of Fisher Scoring iterations: 4
```

```r
stargazer(base, mod.glm, mod.glm.interaction1, mod.glm.interaction2, type = "text")
```

```
##
## =========================================================
##                             Dependent variable:
##                   ---------------------------------------
##                             sanders_preference
##                      (1)       (2)       (3)       (4)
## ---------------------------------------------------------
## age               -0.008**  -0.013*** -0.010**  -0.021***
##                    (0.003)   (0.004)   (0.005)   (0.008)
##
## bi.party1                   -0.670***  -0.297   -0.664***
##                              (0.127)   (0.374)   (0.127)
##
## race_white1                 0.865***  0.865***    0.405
##                              (0.141)   (0.141)   (0.409)
##
## age:bi.party1                          -0.008
##                                        (0.007)
##
## age:race_white1                                   0.010
##                                                  (0.009)
##
## Constant          0.669***  0.557***    0.405    0.892**
##                    (0.177)   (0.207)   (0.251)   (0.349)
##
## ---------------------------------------------------------
## Observations        1,191     1,191     1,191     1,191
## Log Likelihood    -809.356  -766.869  -766.306  -766.148
## Akaike Inf. Crit. 1,622.712 1,541.737 1,542.611 1,542.296
## =========================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

** Model Selection Process **

We follow the principle of parsimony that the simpler model is better. Also we gradually add variables by testing if the new variable improve the model. We use Anova()/anova() function for the testing.

- First, we come up the simplest model which is "base" model that incorporates only age variable.
- Second, we come up with the "mod.glm" which add bi.party and race_white because our EDA shows these two variables have clear association with the dependent variable. Look at the output of Anova()

function on the mod.glm. There is statistical evidence that they have coefficients that are not 0.

```
Anova(mod.glm, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: sanders_preference
##             LR Chisq Df Pr(>Chisq)
## age           12.184  1  0.0004821 ***
## bi.party      27.973  1  1.230e-07 ***
## race_white    38.026  1  6.982e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Third, we test if we should add an interaction term. We consider two options: age:bi.party and age:race_white. Look at the two test results below. They show that their coefficient might be just by chance and not statistically significant. Thus, we decide to incorporate these interaction terms in the model.

```
anova(mod.glm, mod.glm.interaction1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age + bi.party + race_white
## Model 2: sanders_preference ~ age + bi.party + race_white + age:bi.party
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1187     1533.7
## 2      1186     1532.6  1   1.1257   0.2887
```

```
anova(mod.glm, mod.glm.interaction2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age + bi.party + race_white
## Model 2: sanders_preference ~ age + bi.party + race_white + age:race_white
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1187     1533.7
## 2      1186     1532.3  1   1.4412     0.23
```

- Lastly, we check AIC for all the models to see the fit. The smaller AIC means better fit. The second model “mod.glm” has the best AIC score and this result aligns with the anova/Anova tests so far. Thus we decide to pick “mod.glm”.

d. Model result

Here is our selected model.

$$log(odds) = 0.557 - 0.013age + 0.670bi.party + 0.865race\_white$$

Here is the model output

```
summary(mod.glm)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + bi.party + race_white,
##     family = "binomial", data = df)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6907  -1.1821   0.7904   0.9892   1.6669
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.556729   0.207245   2.686 0.007224 **
## age         -0.012688   0.003655  -3.472 0.000518 ***
## bi.party1   -0.669732   0.126733  -5.285 1.26e-07 ***
## race_white1  0.865223   0.141409   6.119 9.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1533.7  on 1187  degrees of freedom
## AIC: 1541.7
##
## Number of Fisher Scoring iterations: 4
```

e. Statistical tests for the model

Let's interpret the model result above one by one.

Deviance Residuals: A perfect fit of point gives a deviance of zero while a poorly fitting point has a large residual deviance. Our residuals have median 0.7904, max 1.6669, and -1.6907.

p-value for coefficients: p-value for all the coefficients are very small thus each coefficient is statistically significant.

Confidence interval for coefficients:

```
# CIs using profiled log-likelihood
confint(mod.glm)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %       97.5 %
## (Intercept)  0.15155141  0.964533512
## age         -0.01988578 -0.005549915
## bi.party1   -0.91859033 -0.421596641
## race_white1  0.58912267  1.143790653
```

Confidence interval for pi: We construct CI using both Wald CI and LRT and compare them. As shown below, they are reasonably close. So we are going to use Wald CI later for convenience.

```
# set-up dataframe and matrix
age = c(20, 30, 40, 50, 60, 70, 80)
newdf = data.frame(age = age,
                   bi.party = factor(1), # democrat
                   race_white = factor(1) # white
                   )

matrix.PLR = data.matrix(data.frame(col = 1, newdf))
```

```
### Wald Condidence Internval
lp.hat = predict.glm(mod.glm, newdata = newdf, type = "link", se.fit = TRUE)

# calcualte ci
lp.hat.mean = lp.hat$fit
lp.hat.lci = lp.hat$fit - 1.96 * lp.hat$se.fit
lp.hat.uci = lp.hat$fit + 1.96 * lp.hat$se.fit

# convert to probability
pi.hat = exp(lp.hat.mean) / (1 + exp(lp.hat.mean))
pi.hat.lci = exp(lp.hat.lci) / (1 + exp(lp.hat.lci))
pi.hat.uci = exp(lp.hat.uci) / (1 + exp(lp.hat.uci))

### Profile Likelihood Ratio Interval
# calculate ci
linear.combo = mcprofile(object = mod.glm, CM = matrix.PLR)
ci.logit.profile = confint(object = linear.combo, level = 0.95)
ci = exp(ci.logit.profile$confint)/(1 + exp(ci.logit.profile$confint))

### store the result in the df
result = data.frame(age, pi.hat.lci, pi.hat.uci, ci)
colnames(result) <- c("age","wald-lower", "wald-higher", "profile-lower", "profile-higher")
result
```

```
##   age wald-lower wald-higher profile-lower profile-higher
## 1  20  0.5480335   0.6908694     0.5335123      0.7041895
## 2  30  0.5278910   0.6528232     0.5154365      0.6648474
## 3  40  0.5042277   0.6159700     0.4931871      0.6269013
## 4  50  0.4755699   0.5825922     0.4650009      0.5930975
## 5  60  0.4413501   0.5541732     0.4302332      0.5652354
## 6  70  0.4028036   0.5304363     0.3903754      0.5430288
## 7  80  0.3621726   0.5100693     0.3481375      0.5248656
```

f. Interpret the dependent variable using odds ratios

Race: Holding other variables constant, the odds of being a Sanders supporter for white people over the the odds of being a Sanders supporter for non-white people is exp(0.865) = 2.375. In terms of percent change, the odds of being a Sanders supporter for white are 137% (2.375 - 1) higher than the odds for non-white. Notice that CI is wide from 1.8 to 3.1.

Party: Similarly, holding other variables constant, NOT being a democrat increases the odds of being a Sanders supporter vs. being a democrat by 95% (1/ exp(0.670) - 1). Notice that CI is 0.399 to 0.656.

Age: The coefficient for the age says one we will see 1.3% (1 / exp(-0.013) - 1) increases in the odds of being Sanders supporter for a one-unit decrease in age (one year younger). CI is narrow.

```
# odds ratio
exp(coef(mod.glm))
```

```
## (Intercept)         age   bi.party1 race_white1
##   1.7449561   0.9873920   0.5118459   2.3755367
```

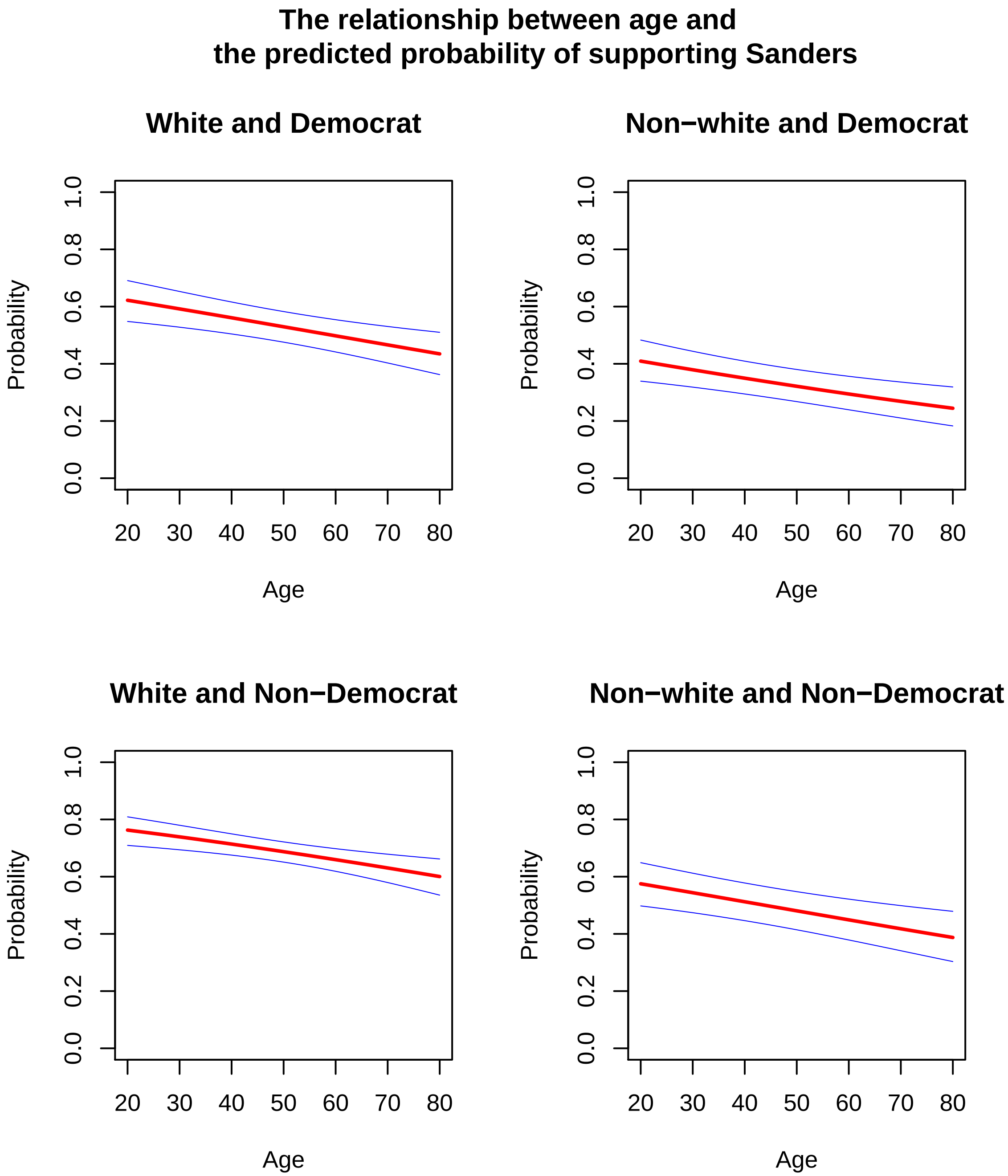
```
# odds ratios and 95% CI using profiled log-likelihood
exp(cbind(OR = coef(mod.glm), confint(mod.glm)))
```

```
## Waiting for profiling to be done...
##                     OR     2.5 %    97.5 %
## (Intercept) 1.7449561 1.1636381 2.6235635
## age         0.9873920 0.9803106 0.9944655
## bi.party1   0.5118459 0.3990812 0.6559986
## race_white1 2.3755367 1.8024064 3.1386434
```

3. Relationship between age and the predicted probabilty of supporting Sanders

Recall some findings from the exploration data analysis in the previous section. White people are more likely to support Sanders than non-white people are. Non-democratic voters are more likely to support Sanders than democratic voters are. Younger people are more likely to support Sanders. Let's check these characteristics are reflected in our selected model. First we construct dataframe to represent four types of demographics. Then we plot the relationship between age and the predicted probability of supporting Sanders for each demography.

```
### create dfs
# white and democrat
newdf = data.frame(age = seq(from = 20, to = 80, by = 1),
                    bi.party = factor(1), # democrat
                    race_white = factor(1) # white
                    )

# non-white and democrat
newdf2 = data.frame(age = seq(from = 20, to = 80, by = 1),
                    bi.party = factor(1), # democrat
                    race_white = factor(0) # non-white
                    )

# white and non-democrat
newdf3 = data.frame(age = seq(from = 20, to = 80, by = 1),
                    bi.party = factor(0), # non-democrat
                    race_white = factor(1) # white
                    )

# non-white and non-democrat
newdf4 = data.frame(age = seq(from = 20, to = 80, by = 1),
                    bi.party = factor(0), # non-democrat
                    race_white = factor(0) # non-white
                    )


### function to plot ci
plot_ci = function(newdf, title) {
  # predict
  lp.hat = predict.glm(mod.glm, newdata = newdf, type = "link", se.fit = TRUE)

  # calcualte ci
  lp.hat.mean = lp.hat$fit
  lp.hat.lci = lp.hat$fit - 1.96 * lp.hat$se.fit
  lp.hat.uci = lp.hat$fit + 1.96 * lp.hat$se.fit
  # convert to probability
  pi.hat = exp(lp.hat.mean) / (1 + exp(lp.hat.mean))
  pi.hat.lci = exp(lp.hat.lci) / (1 + exp(lp.hat.lci))
  pi.hat.uci = exp(lp.hat.uci) / (1 + exp(lp.hat.uci))

  # plot
  age = newdf$age # x axis
  plot(age, pi.hat, ylim = range(c(0, 1)),
     xlab = "Age", ylab = "Probability", main = title,type = 'l', col = 'red', lwd = 2 )
  lines(age, pi.hat.lci, col = 'blue', lwd = 0.5)
  lines(age, pi.hat.uci, col = 'blue', lwd = 0.5)

}

### plot
par(mfrow=c(2,2), oma=c(0,0,2,0))
plot_ci(newdf, "White and Democrat")
plot_ci(newdf2, "Non-white and Democrat")
plot_ci(newdf3, "White and Non-Democrat")
plot_ci(newdf4, "Non-white and Non-Democrat")
title("The relationship between age and
      the predicted probability of supporting Sanders", outer=TRUE)
```



The plots below align with our data exploration results that discussed in the previous paragraph. White and Non-Democrat plot (bottom left) show the highest probability of supporting Sanders. In the same group, the younger the higher probability of supporting Sanders. If it changes from Non-Democrat to Democrat or from White to Non-white, the probability curve goes down (See top left and bottom right). Non-white and Democrat plot (top right) reflect both of these effects and as a result has the lowest probability curve among the four plots.

4. Conclusion

There is a statistical evidence to believe that there is a relationship between age and Sanders supporters. Sanders supporters tend to be younger. This would be an interest for the client who wants to sell goods targeting Sanders supporters. Our analysis shows that there are better way of targeting Sanders supporters by considering race and party affiliation. *Target non-democrat:* Look at the plots above: Non-democrat voters are more likely to be a Sander supporter than democrat voters are. *Target caucasian:* Caucasians are more likely to support Sanders.

Taking advantage of these findings, we propose to create a marketing plan to reach Sanders supporters better. For example, if the client have a channel to reach out to college students who are mostly Caucasian and non-democrat, it is worth investing on this channel. While it is wise not to invest in Democrat dominant, non-white neighbors.

Please note that the suggestions are derived from the inference of the data-set and they are not related to my own brief or my thinking.