w271 Lab1

K Iwasaki

September 29, 2017

Contents

0. Introduction

Objectives of this project are to create statical models that incorporate the relationship between voters' preference and dependent variables including age, gendear, 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) dataset 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. Specificially, we check summary statistics, variable categories (categorical, continous and etc), NA values in each column, and distribution for each column.

Some observations from the intial exploration:

- There are 1200 examples in the dateset.
- 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 indepenent variables have very weak correlation with sanders_preference. Also note that depedent variables don't show strong colleration 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 seeems 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.

```
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
```

```
nrow(df)
```

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

```
# 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
##
```

```
# 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 ...
```

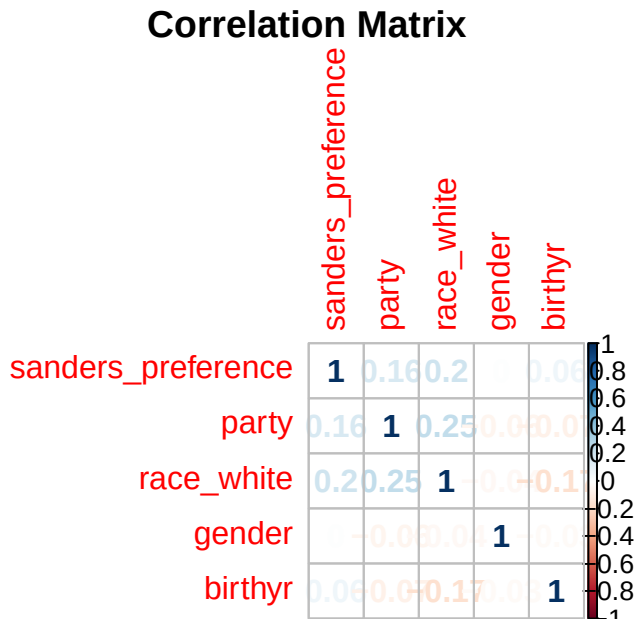
```r
# check NA values
apply(is.na(df), 2, sum)
```

```
## 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))
```



```r
# conver columns into factor
df$race_white = factor(df$race_white)
df$party = factor(df$party)
df$gender = factor(df$gender)

# confirm the change
str(df)
```

```
## 'data.frame':    1191 obs. of  5 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 ...
```

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

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

```r
# 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
```

```r
# expalin this later
df$bi.party = ifelse(df$party == 1, 1, 0)
```

2. Model

a. Description of the model

Describe your chosen model in words, along with a brief description of the variables and the model's functional form (Note: You do not have to justify your choices at this step).

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

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

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
- Race — Included in the model
- Party — Included in the model

- Age — Included in the model
- 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.tset. With the following obesrvations, 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 siginificant evidence to conclude there is a difference in the two proportions. Also the practical significance is small.

```
# 1. male, 2. female

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

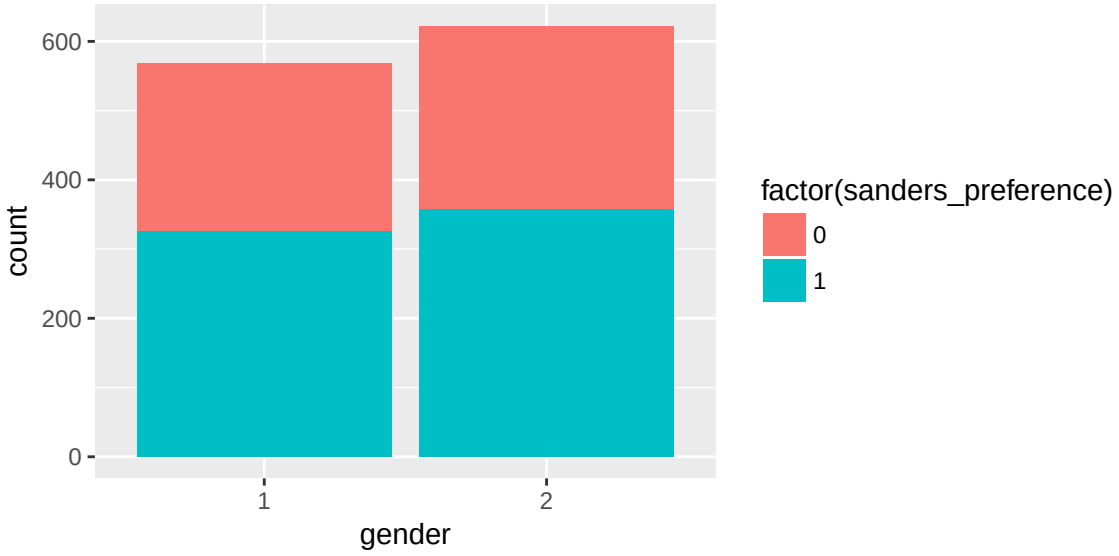
```
##
##       1   2
##   0 242 263
##   1 327 359
```

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

```
##
##             1         2
##   0 0.2031906 0.2208228
##   1 0.2745592 0.3014274
```

```
ggplot(df, aes(x = gender, fill = factor(sanders_preference))) + geom_bar()

# conduct t.test
male = df[df$gender == 1, ]$sanders_preference
female = df[df$gender == 2, ]$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 differece between white and non-white group is statistically and practically significant.
- This validates the previous result of the colleration 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
```

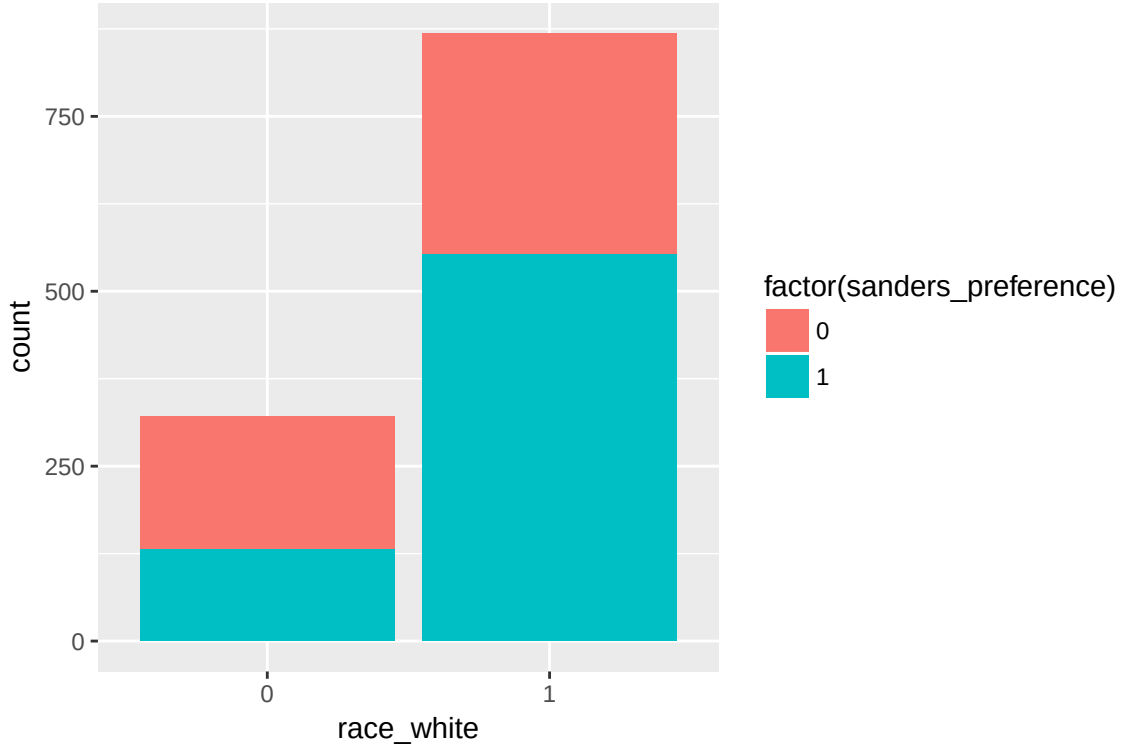
```
prop.table(table(df$sanders_preference, df$race_white))
```

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

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

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

```r
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 (0 as democrat, 1 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 Repulibcan 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)

```r
# 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(sanders_preference))) + geom_bar()

# t.test
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
```

```
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
```

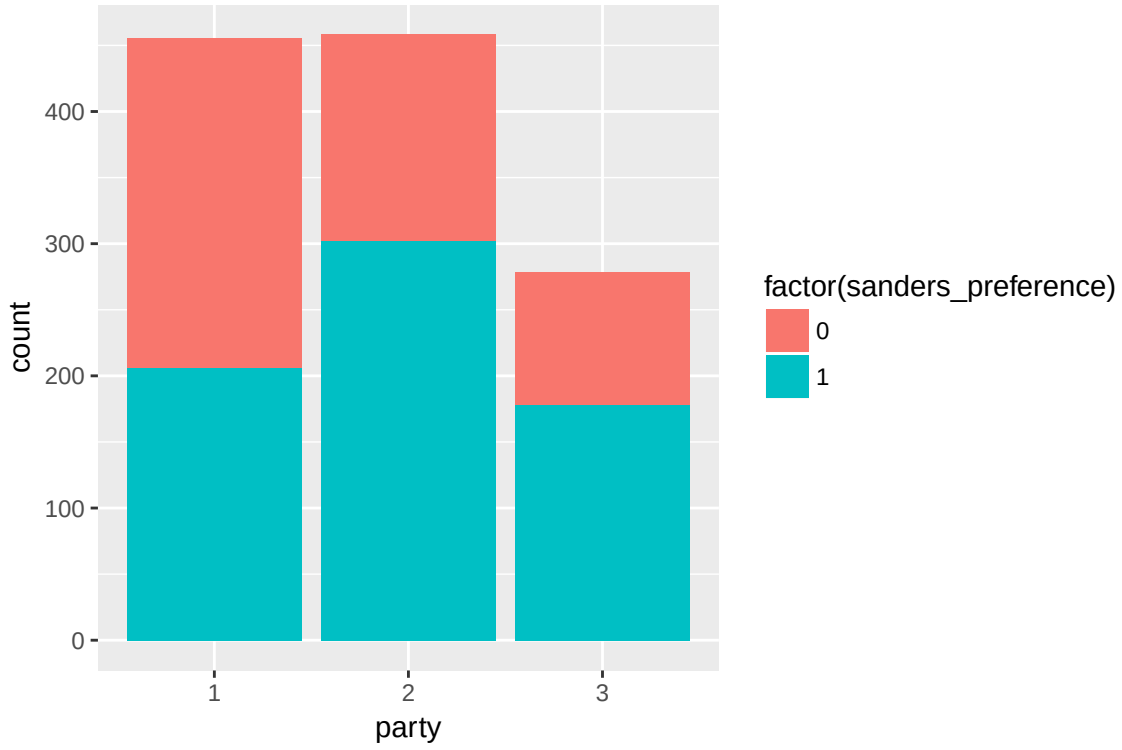
```
t.test(party1, party3)
```

```
##
##  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
```

```
# create new variable
df$bi.party = ifelse(df$party == 1, 0, 1)
table(df$bi.party)
```

```
##
##   0   1
## 455 736
```



Age – Included in the model

Recall that the correlation matrix shows that there is no strong correlation between age and the dependent variable and instead, there is a negative correlation between race_white and age. with that in mind, we observe as followings through the analysis.

- The t.test shows that avereage 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 differrent, might cause different inpretations 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.

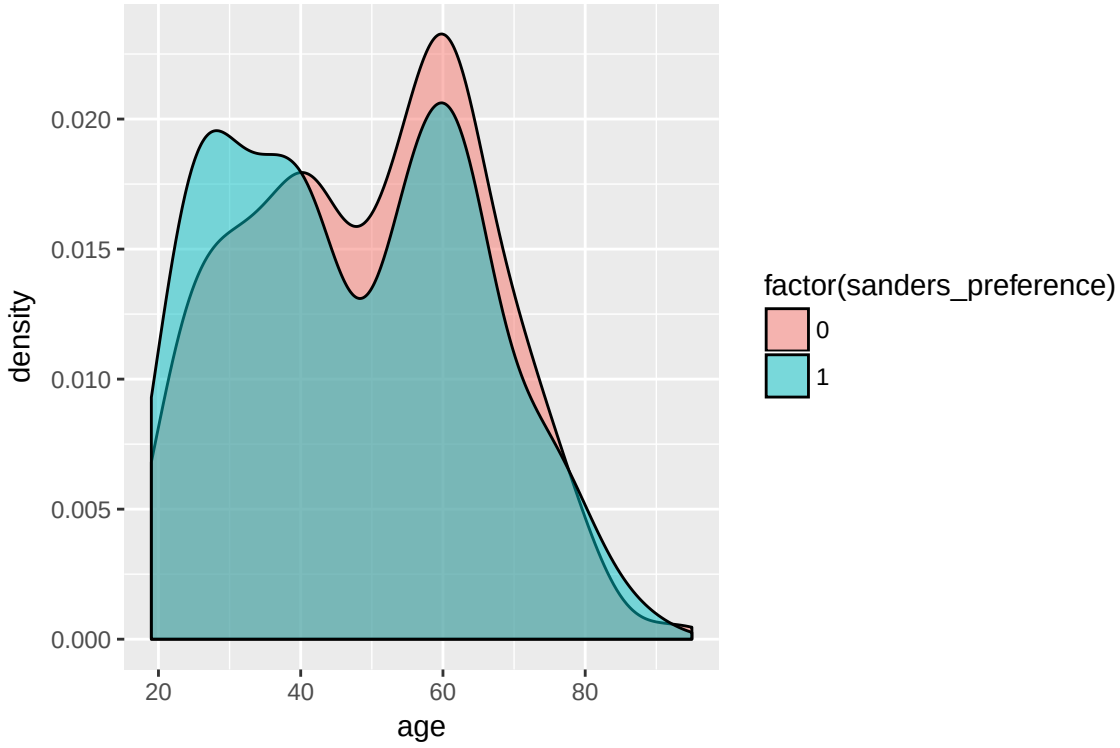
```
ggplot(df, aes(x = age, fill = factor(sanders_preference))) +
  geom_density(alpha = 0.5)

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

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

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

```r
# 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

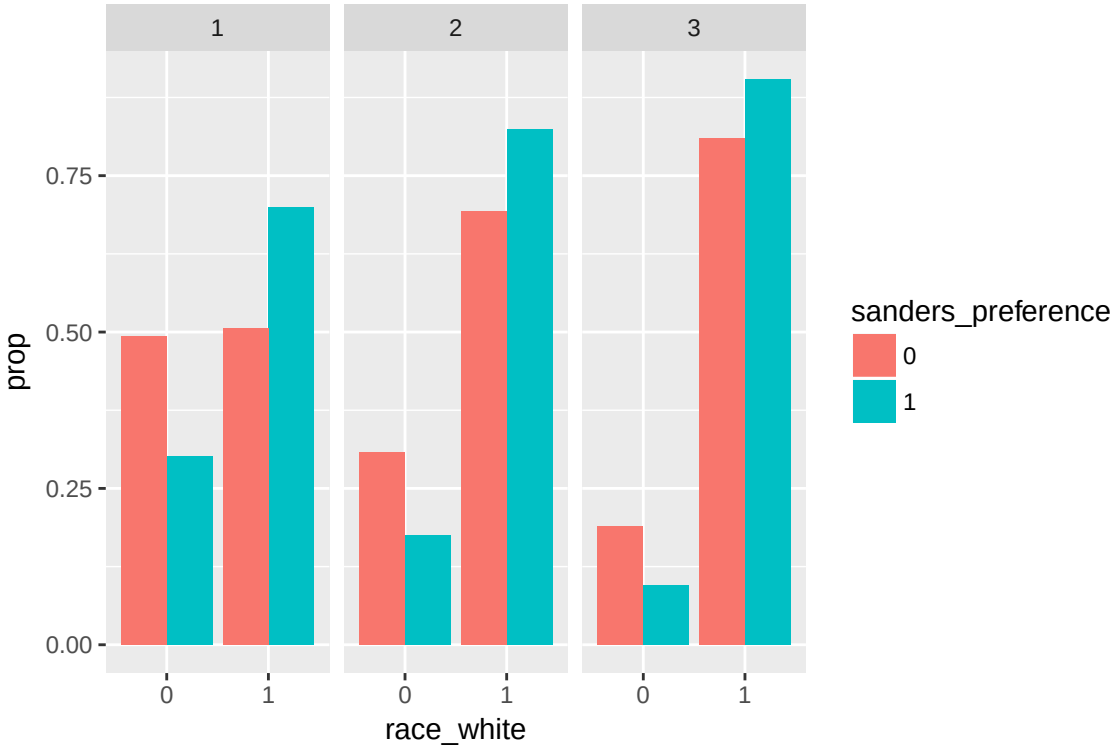
Move onto investigate interaction terms: we focus on look for particular segment of voteres that shows siginificantly difference in terms of the peference.

White voters consistently support Sanders across parties and non-white voters consistenly 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
```

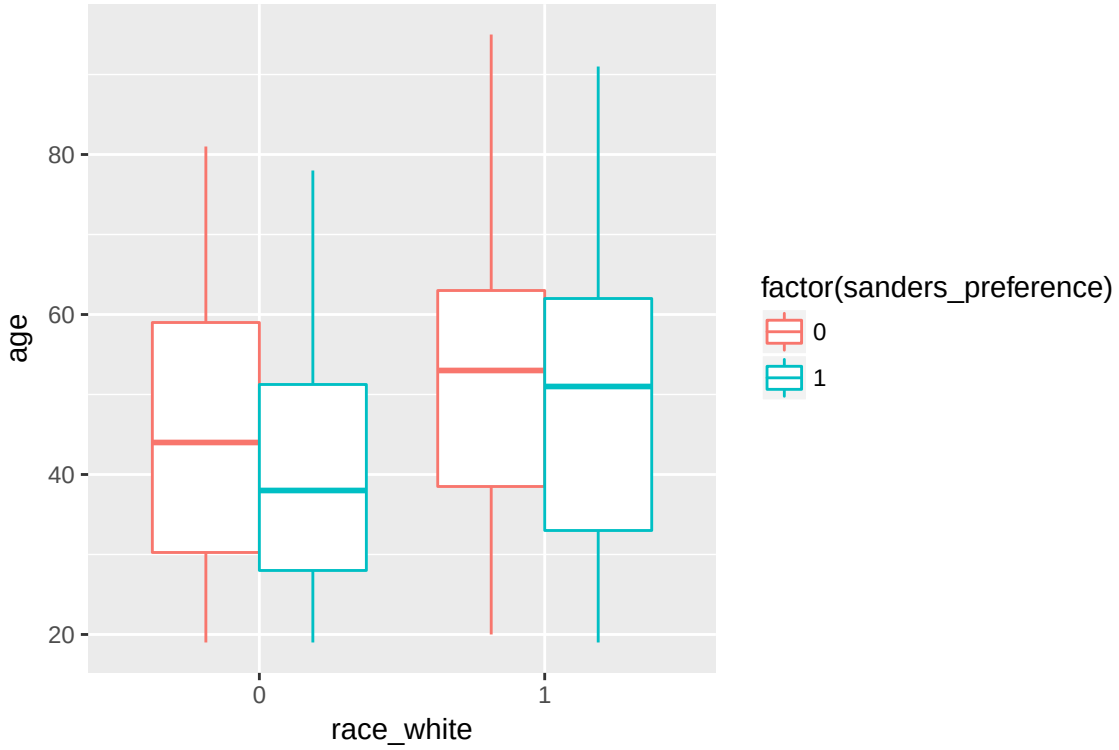
```
ggplot(df, aes(x =race_white, y = ..prop.., group = sanders_preference, fill = sanders_preference)) +
  geom_bar(position = "dodge") +
  facet_grid(.~party)
```



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

It looks Sanders supporters are yourngerthan 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.

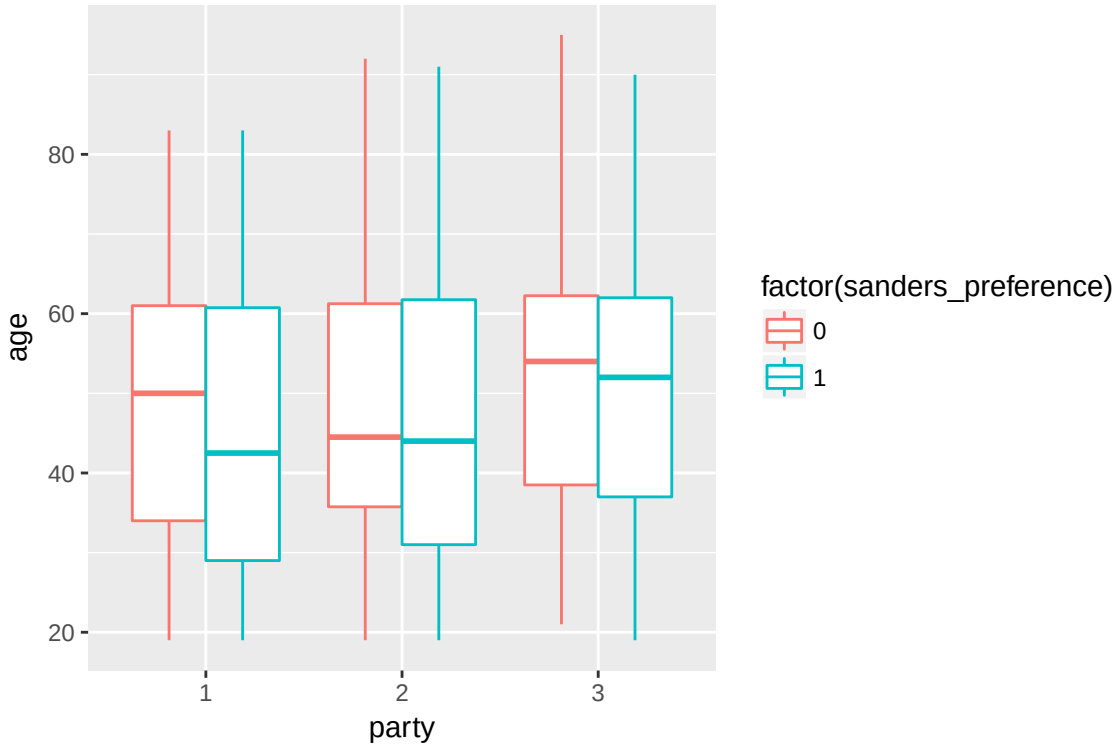
```
ggplot(df, aes(x = race_white, y = age, col = factor(sanders_preference))) + geom_boxplot()
```



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 Cinton supports. This combination is a good candidate for an interaction term.

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



c. Comparison with other candidate models

Based on your EDA, describe other models that you might have considered and why you ended up choosing your final model. Be sure to print each of the model results and any statistical tests you used to choose which model to use.

```
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) -0.113002   0.201460  -0.561 0.574854
## age         -0.012688   0.003655  -3.472 0.000518 ***
## bi.party     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
```

```
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*** -0.017*** -0.021***
##                      (0.003)   (0.004)   (0.006)   (0.008)
##
## bi.party                      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.party                              0.008
##                                          (0.007)
##
## age:race_white1                                     0.010
##                                                    (0.009)
##
## Constant            0.669***   -0.113     0.108     0.228
##                      (0.177)   (0.201)   (0.290)   (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
```

d. Model result

```
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) -0.113002   0.201460  -0.561 0.574854
## age         -0.012688   0.003655  -3.472 0.000518 ***
## bi.party     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

Conduct all of the relevant statistical tests on your chosen model.

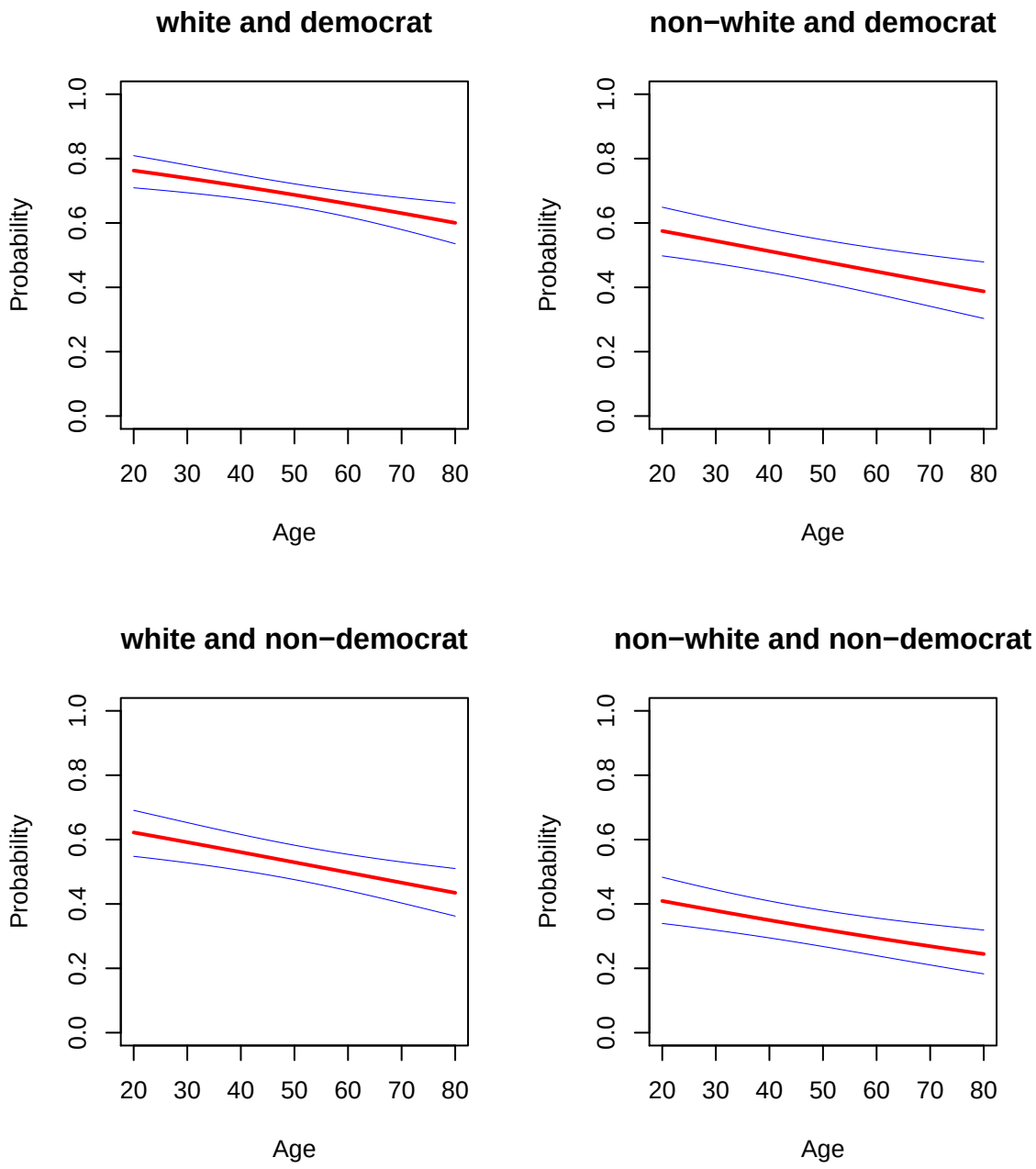
- f-test
- wald -

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

Graph the relationship between age and the predicted probabilty of supporting Sanders. – Be sure to include any graphs that helps you understand how your model can help you answer the question at hand.

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

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

# white and non-democrat
```

```r
newdf3 = data.frame(age = seq(from = 20, to = 80, by = 1),
                    bi.party = 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 = 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))
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")
```



4. Conclusion

Comment on the importance of age and evaluate your client's decision to target younger voters.

5. Appendix