Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 1

Eric Yang, Samir Datta, Carlos Castro

September 30, 2017

Lab Instructions:

- **Due Date: 10/01/2017 11:59 PM PT**
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - SectionNumber_hw01_FirstNameLastNameFirstInitial.fileExtension
    - For example, if you are in Section 1 and have two students named John Smith and Jane Doe, you should name your file the following
      - Section1_hw01_JohnS_JaneD.Rmd
      - Section1_hw01_JohnS_JaneD.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your report.
  - This lab can be completed in a group of up to 3 people. Each group only needs to make one submission. Although you can work by yourself, we encourage you to work in a group.
  - When working in a group, do not use the "division-of-labor" approach to complete the lab. That is, do not divide the lab by having Student 1 completed questions 1 - 3, Student 2 completed questions 4 - 6, etc. Asking your teammates to do the questions for you is asking them take away your own opportunity to learn.
- Other general guidelines:
  - If you use R libraries and/or functions to conduct hypothesis tests not covered in this course, you will have to explain why the functions you use are appropriate for the hypothesis you are asked to test. Lacking explanations will result in a score of zero for the corresponding question.

  - Thoroughly analyze the given dataset. Detect any anomalies, including missing values, potential of top and/or bottom code, etc, in each of the variables.

  - Your report needs to include a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score.

  - Your analysis needs to be accompanied by detailed narrative. Remember, make sure your that when your audience (in this case, the professors and your classmates) can easily understand your your main conclusion and follow your the logic of your analysis. Note that just printing a bunch of graphs and model results, which we call "output dump", will likely receive a very low score.

  - Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence. Remember to use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.

  - All the steps to arrive at your final model need to be shown and explained very clearly.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Description of the Business Problem and the Data

Imagine you work in a data science consulting company. Your client is interested in selling T-shirts to voters who are likely to support politically liberal candidates (such as Bernie Sanders). Your client has data 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.

Your client conducted a t-test and found that younger voters were more likely to support Sanders and is willing to target younger voters/shoppers based on this analysis. He thinks that you can do better.

For reference, the United States is considered a two party system. The Democratic Party tends to be associated with politically liberal polices while the Republican Party tends to be associated with politically conservative ideas. Voters are not required to be associated with these two parties and, as you will see later, a high proportion of voters are not associated with these two parties.

Note: This dataset is modified from the 2016 American National Election Survey.

1. Model the relationship between age and voters' preference for Bernie Sanders over Hillary Clinton. Select the model that you prefer the most and describe why you chose these variables and functional form.

    a. 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*).

    b. Describe the variables your have included in your model and justify why you chose these variables and the model's functional form. (*Hint: you will have to conduct a very careful EDA and use insights generated from the EDA to support your modeling decision. DO NOT USE OUTPUT-DUMP, meaning do not just print a bunch of graphs and let us interpret the graphs for you. Choose your graphs/tables very selectively and present them with narratives to suppor your modeling decisions.*)

    c. 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.

    d. Print the model results of your chosen model, even if you did so earlier.

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

    f. Interpret the impact of age on the dependent variable using odds ratios and be sure to include confidence intervals.

2. For your chosen model, 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.

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

Data loading

Now we load the data and transform the factor variables for more semantically meaningful levels.

```
library(ggplot2)
theme_set(theme_bw())
library(car)
publicopinion <- read.csv('public_opinion.csv')

publicopinion$partyfactor <- ifelse(publicopinion$party==1, 'Democrat',
                                    ifelse(publicopinion$party==2, 'Other',
                                           'Republican'))

publicopinion$age <- 2017 - publicopinion$birthyr
publicopinion$genderfactor <- ifelse(publicopinion$gender==1,'Male', 'Female')
publicopinion$racefactor <- ifelse(publicopinion$race_white==1, 'White', 'Non-White')
publicopinion$spfactor <- ifelse(publicopinion$sanders_preference==1, "Yes", "No")
publicopinion_narm <- publicopinion[!is.na(publicopinion$sanders_preference),]
```

1.a Model

Model Overview

The model looks at the likelihood of supporting bernie sanders based on 4 explanatory variables. The three explantory variables are age as of 2017, race_white (non white vs. white), independent voter (baseline democrat), republican (baseline democrat).

Age

The age variable was caculated as of 2017 based off the birth year variable in the original dataset. Age has a negative coefficient which is statistically significant showing that an increase in age reduces the likelihood of supporting sanders. The odds_age object shows that for every 10 years decrease in age we see a ~1.13 increase in the odds of supporting sanders.

Racefactor

Racefactor = 0 is non white, race_white = 1 is white. Racefator has a positive coefficient in our model that is statistically significant. Looking at the odds ratios we see that being white increases odds of supporting sanders by 2.39 holding all other variables constant.

Partyfactor

Both the independent and republican variables have statistically significant positive coefficients. This shows that either party affiliation versus democrat would increase the probability of supporting sanders. A look at the odds ratio shows an odds increase of 2.04 for independents and 1.81 for republicans.

```
#Section 1a.
model.final <- glm(sanders_preference ~ age + racefactor + partyfactor, data = publicopinion, family = 

summary(model.final)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + racefactor + partyfactor,
##     family = binomial(link = logit), data = publicopinion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7036  -1.1792   0.7907   0.9881   1.6662
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.115017   0.205360  -0.560 0.575428
## age                     -0.012480   0.003666  -3.404 0.000664 ***
## racefactorWhite          0.872782   0.141872   6.152 7.66e-10 ***
## partyfactorOther         0.713501   0.140368   5.083 3.71e-07 ***
## partyfactorRepublican  0.594231   0.162972   3.646 0.000266 ***
## ---
## 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: 1533.2  on 1186  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 1543.2
##
## Number of Fisher Scoring iterations: 4
```

```
odds_age <- exp(model.final$coefficients[2]*-10)
odds_age
```

```
##      age
## 1.132925
```

```
odds_white<- exp(model.final$coefficients[3])
odds_white
```

```
## racefactorWhite
##         2.39356
```

```
odds_ind <- exp(model.final$coefficients[4])
odds_ind
```

```
## partyfactorOther
##         2.041124
```

```
odds_rep <- exp(model.final$coefficients[5])
odds_rep
```

```
## partyfactorRepublican
##              1.811638
```

1.b EDA

Overview

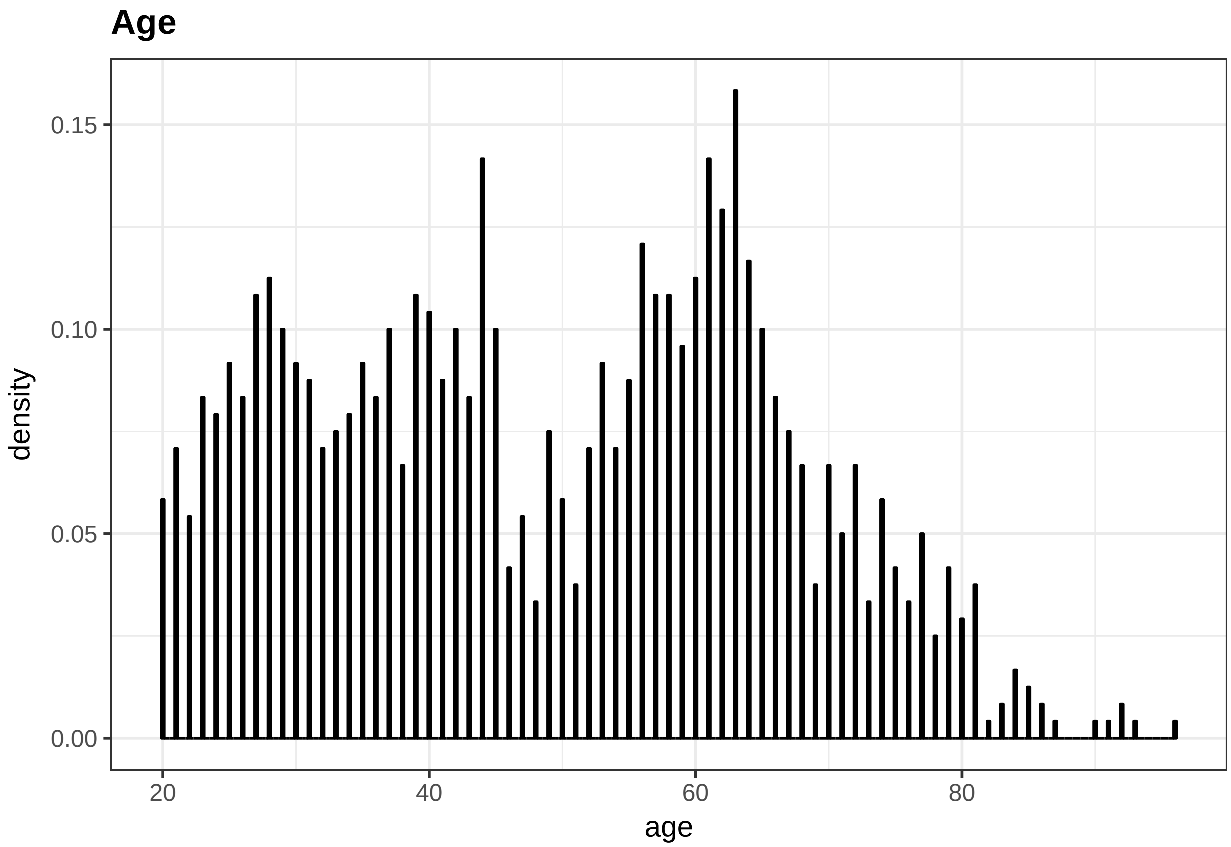
```
summary(publicopinion)
```

```
##  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     partyfactor             age        genderfactor      
##  Min.   :1921   Length:1200        Min.   :20.00   Length:1200       
##  1st Qu.:1955   Class :character   1st Qu.:35.00   Class :character  
##  Median :1968   Mode  :character   Median :49.00   Mode  :character  
##  Mean   :1968                      Mean   :49.06                     
##  3rd Qu.:1982                      3rd Qu.:62.25                     
##  Max.   :1997                      Max.   :96.00                     
##                                                                      
##   racefactor          spfactor        
##  Length:1200        Length:1200       
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
##                                       
##                                       
```

We are missing 9 responses to sanders preference, average preference for sanders is 57.6%, race_white is 72.9%

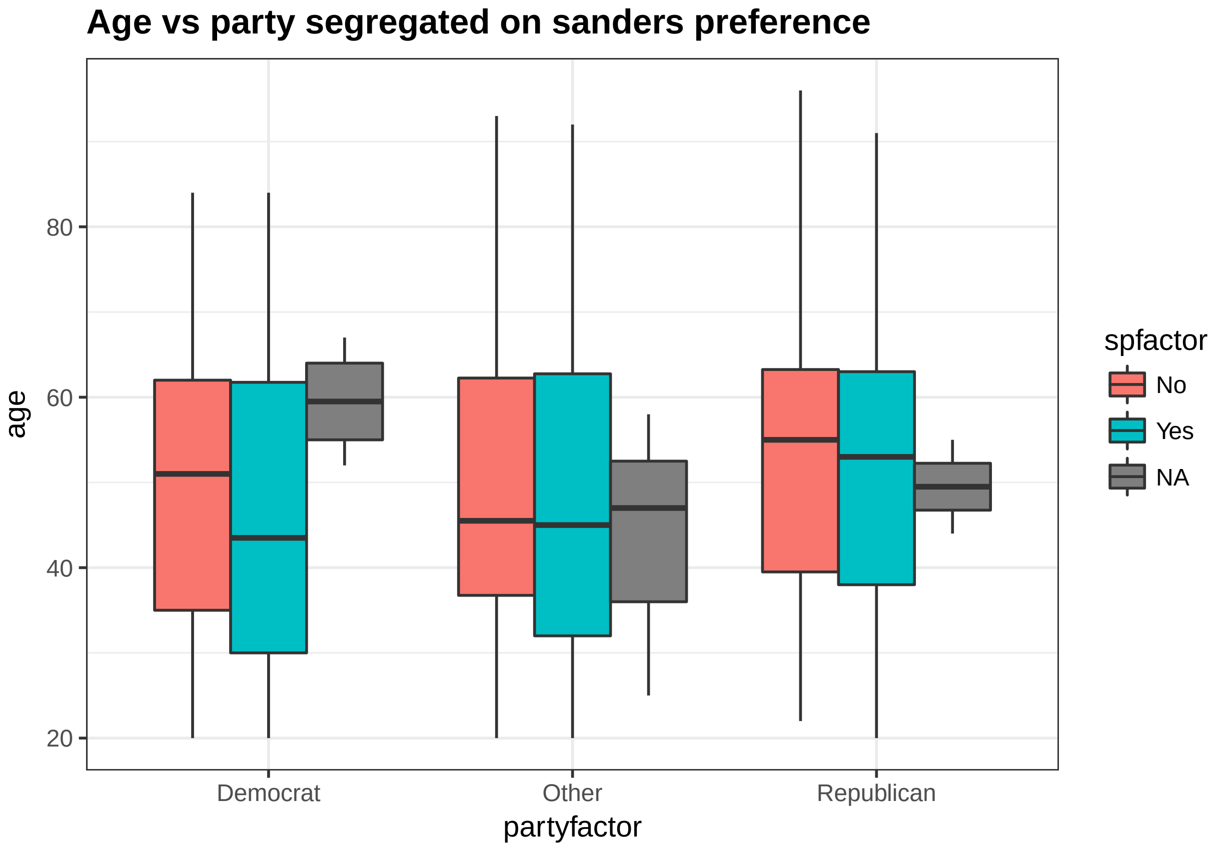
Age

```
# Distribution of Age
ggplot(publicopinion, aes(x = age)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.2, fill="#0072B2", colour="black") +
  ggtitle("Age") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



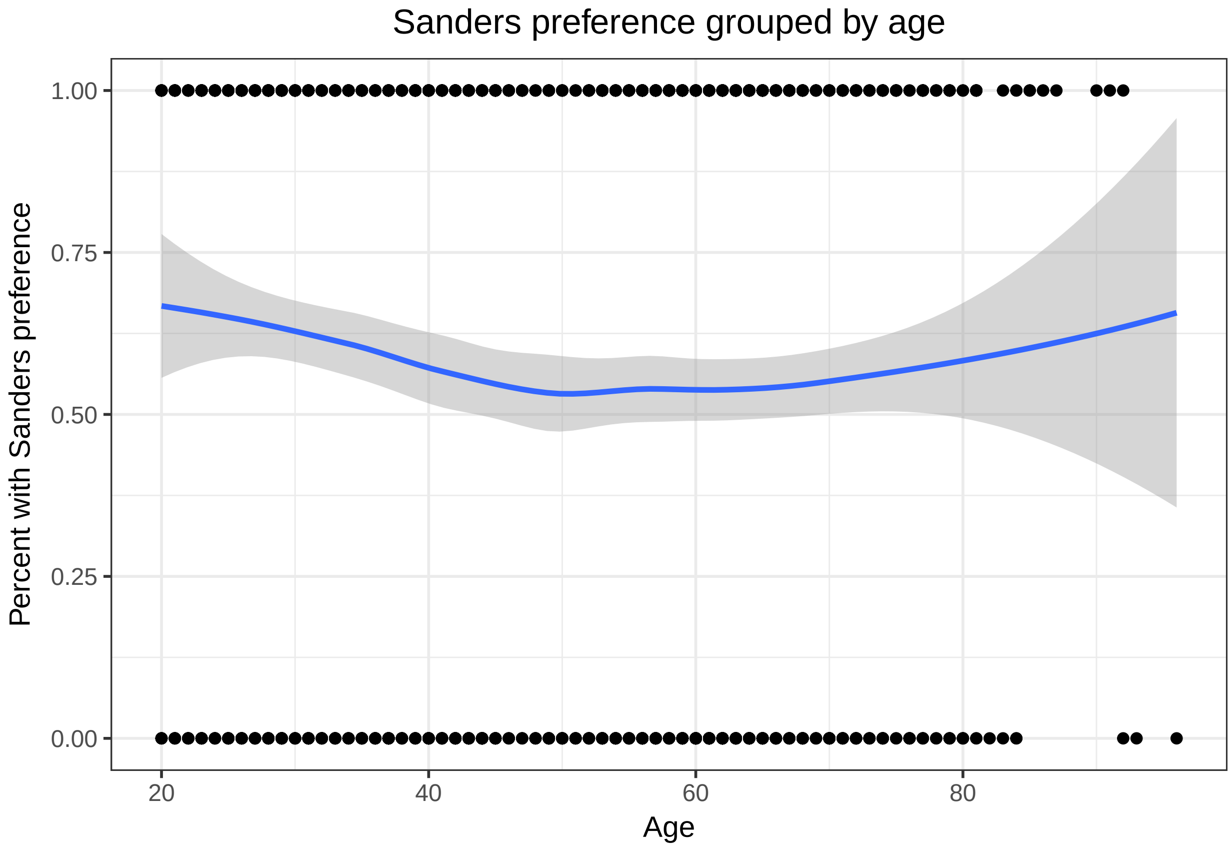
Ages range from 20 to 96. Peaks in the mid 60s and mid 40s and a dip in the 50s.

```r
ggplot(publicopinion, aes(partyfactor, age)) +
  geom_boxplot(aes(fill = spfactor)) +
  #geom_jitter() +
  ggtitle("Age vs party segregated on sanders preference") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



The box plot above shows a difference in age between party affiliation. Republican voters skew older for both sanders preference outcomes with the median age in the mid fifties. The indpendent voters skew younger compared to republicans with median age in the mid 40s. There appears to be the largest gap between democratic voters when taking into account sanders preference. Older democratic voters appear less likely to support sanders than younger demcratic voters.

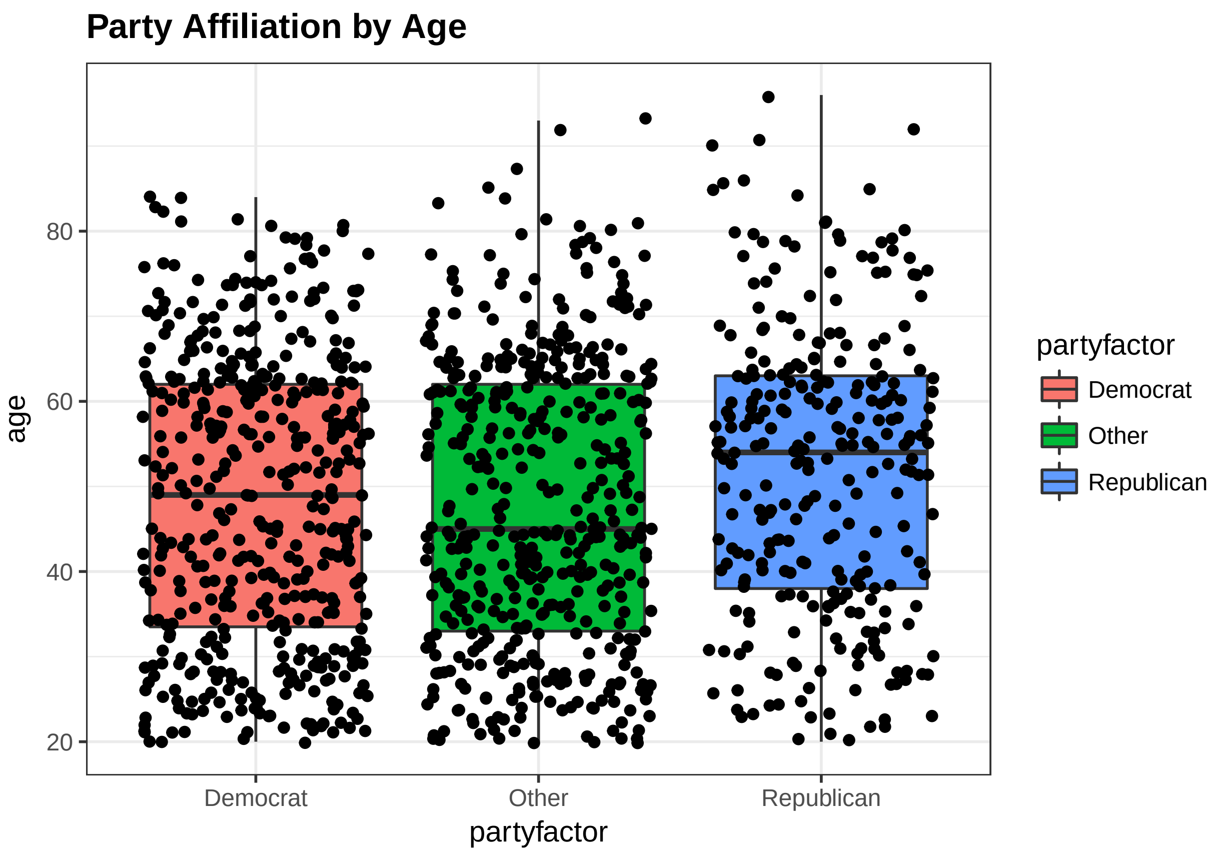
```
ggp <- ggplot(publicopinion_narm, aes(x=age, y=sanders_preference))

ggp + geom_point()+
  geom_smooth(method="loess", se=T)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age")+
  theme(plot.title=element_text(hjust=.5))
```



Above is a scatterplot of all of the points in the set, with age on the x-axis and the binary variable sanders_preference on the y-axis. Displaying the dots like this is not necessarily informative, but the loess smooth curve - and standard error ribbon - reveals an interesting trend. Below the age 50, there seems to be a trend for younger voters to prefer Sanders. However, this trend is also seen in the opposite direction for voters above around 70. This would suggest that including a quadratic term for age might be useful. However, it is important to note that the standard error ribbon is very large towards the older end of the age range, which is indicative of how few voters of that age range we really have. While we should try modeling a quadratic term for age, we should be careful not to over-interpret any result based off insufficient data.

Party

```
ggplot(publicopinion, aes(partyfactor, age)) +
  geom_boxplot(aes(fill = partyfactor)) +
  geom_jitter() +
  ggtitle("Party Affiliation by Age") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



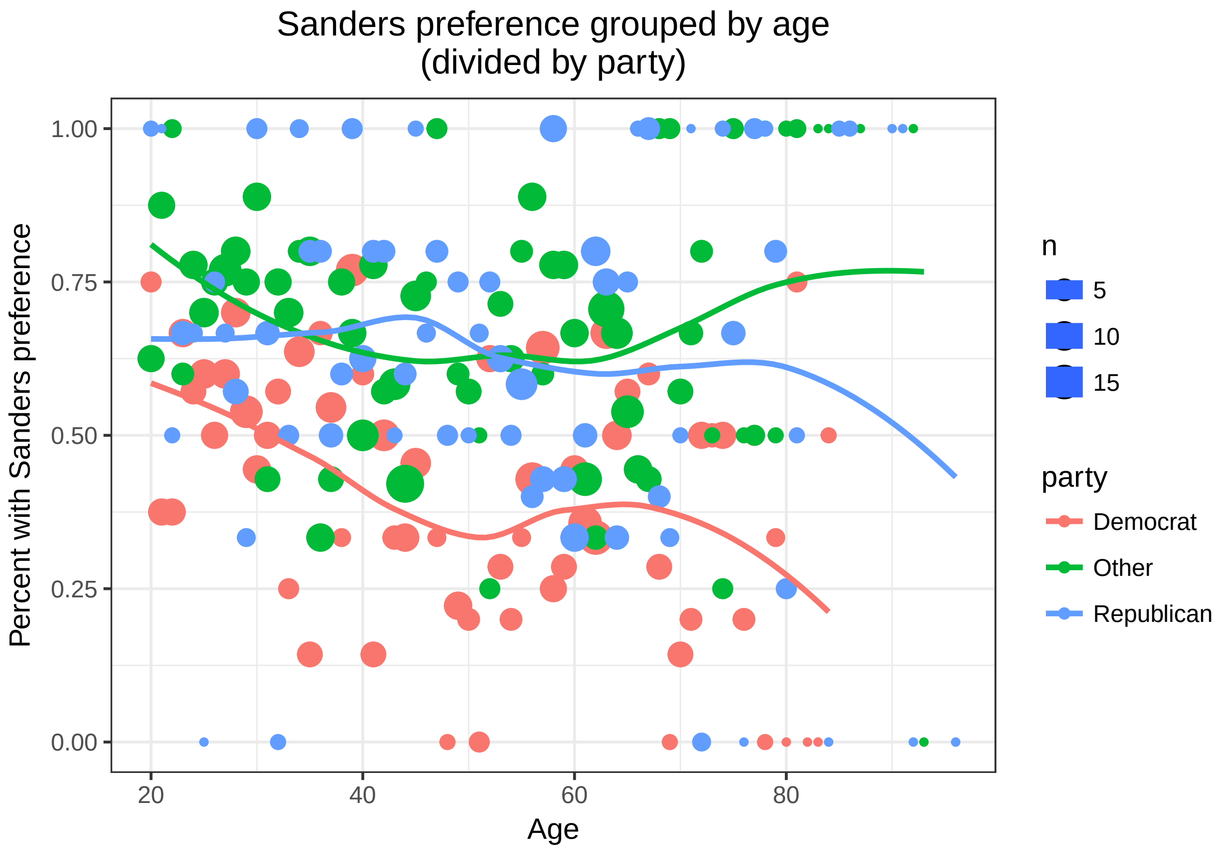
Democrats median age is 49, Independents are 45, Republicans are 54. Republicans have a higher age range and have more observations in the 80+ age range. We should add party to control for party affiliation since there seems to be a difference in age between groups.

```
po_party_agg <- with(publicopinion_narm,
                     aggregate(cbind(100*sanders_preference),
                               list(partyfactor=partyfactor),
                               mean))
po_party_agg$n <- with(publicopinion_narm,
                     aggregate(cbind(100*sanders_preference),
                               list(partyfactor=partyfactor),
                               length))[,2]

po_party_df <- data.frame(sanders_pref_percent=po_party_agg$V1,
                          party=po_party_agg$partyfactor,
                          sample_percent = 100*po_party_agg$n/1191)
po_party_df
```

```
##   sanders_pref_percent      party sample_percent
## 1             45.27473   Democrat       38.20319
## 2             65.93886      Other       38.45508
## 3             64.02878 Republican       23.34173
```

Republicans, who make up 23% of the sample, are slightly underrepresented compared to Democrats and Other, but not to the extent that we should be worried about sampling bias. A slight majority (55%) of Democrats polled preferred Clinton to Sanders, while a majority of Republicans (64%) and Other (66%) preferred Sanders. This difference supports including party as an explanatory variable.

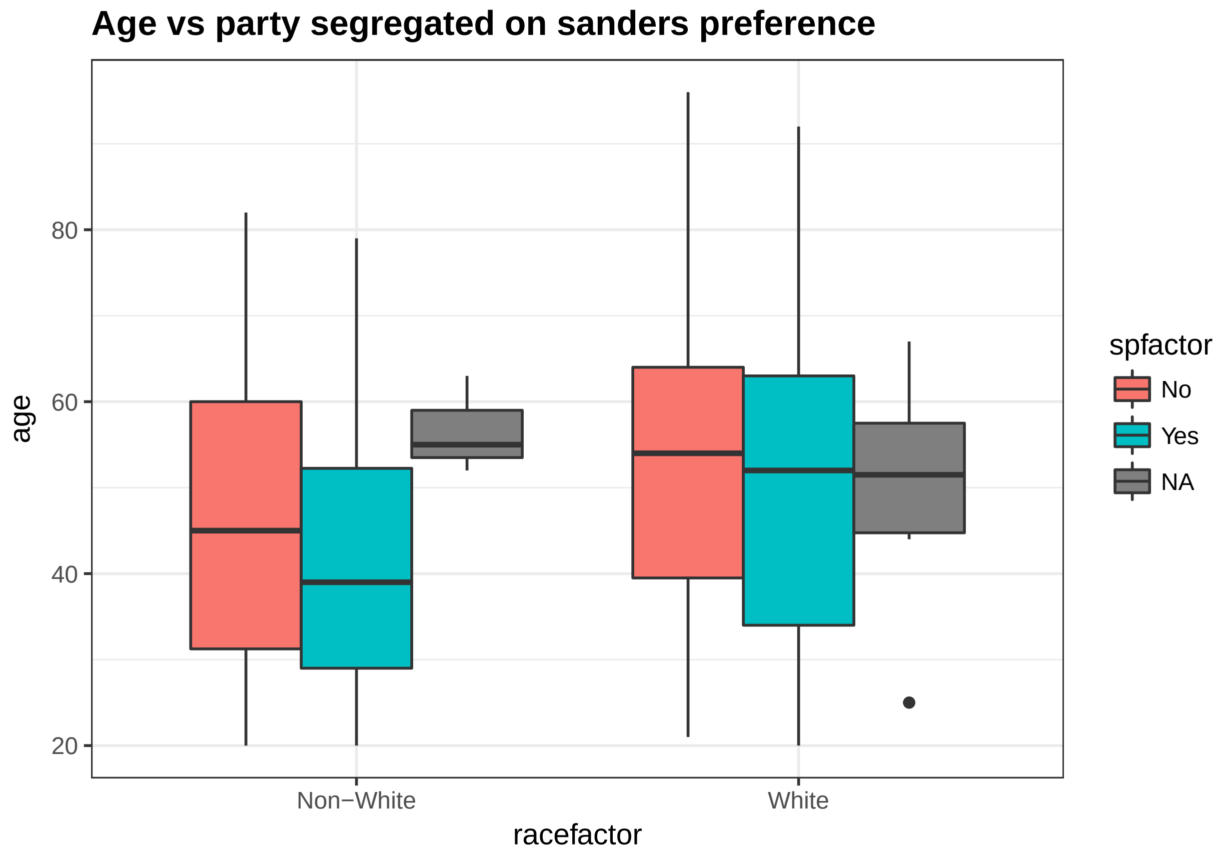
```
age_bin_agg_party <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       party=partyfactor), mean))
age_bin_agg_party$n <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       party=partyfactor), length))[,3]

ggp <- ggplot(age_bin_agg_party, aes(x=agebin, y=sanders_preference,
                                     color=party, size=n))

ggp + geom_point(aes(color=party))+
  geom_smooth(method="loess", se=F)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age\n(divided by party)")+
  theme(plot.title=element_text(hjust=.5))
```



Above is a scatterplot of age on the x-axis, where color represents the political party. Each dot's position on the y-axis represents the percent of people of that specific party and age that preferred Sanders.

The democrats seem to have a fairly clear relationship with age in that younger democrats look more likely to support Sanders than older ones. The relationship within Republicans is less clear, and for independents, it looks almost quadratic (as the smooth curve lifts upwards both for younger and older voters). The curves are loess smoothed curves and not meant to be a perfect representation of overall trends. However, there is still enough evidence to support at least trying to model an age by party interaction, since it looks like different parties may have different relationships with age. ## Race

```
ggplot(publicopinion, aes(racefactor, age)) +
  geom_boxplot(aes(fill = spfactor)) +
  #geom_jitter() +
  ggtitle("Age vs party segregated on sanders preference") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



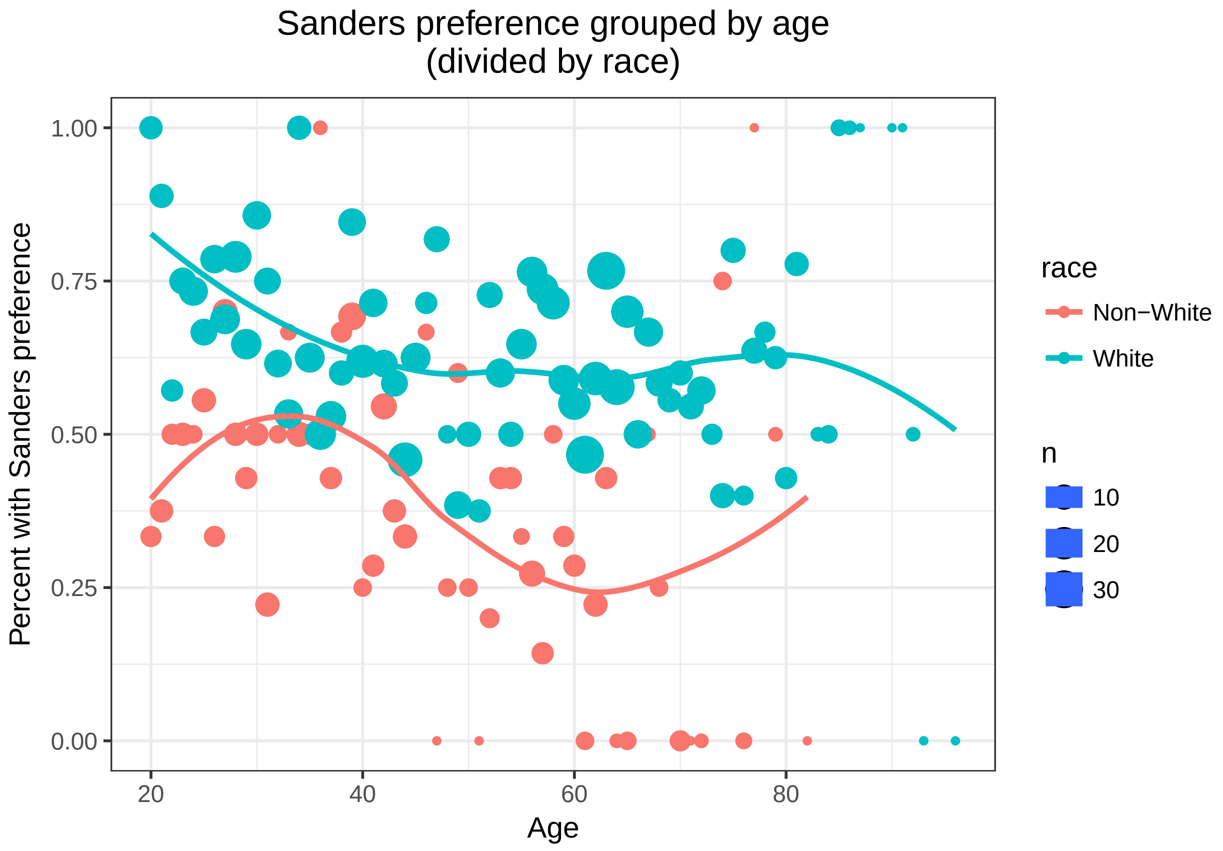
Looking at the race_white by age chart we see the distribution of age with respect to race. Whites skew older with a median age of 53 compared to that of 43 for non whites. This large difference in preference suggests we should add the race_white variable so we can control for the effect of race when evaluating age. Potentially would be interesting to look add an iteraction term the model, age:race_white, to see the effect of age with respect to age and sanders preference.

```
po_race_agg <- with(publicopinion_narm,
                    aggregate(cbind(100*sanders_preference),
                              list(racefactor=racefactor),
                              mean))
po_race_agg$n <- with(publicopinion_narm,
                    aggregate(cbind(100*sanders_preference),
                              list(racefactor=racefactor),
                              length))[,2]

po_race_df <- data.frame(sanders_pref_percent=po_race_agg$V1,
                         party=po_race_agg$racefactor,
                         sample_percent = 100*po_race_agg$n/1191)
po_race_df
```

```
##   sanders_pref_percent     party sample_percent
## 1             40.99379 Non-White        27.0361
## 2             63.75144     White        72.9639
```

White voters make up about 73% of the sample, which is close to the estimated percentage of White people in America, which further supports our sample being representative. There is a large difference between how many white voters (64%) vs. non-whie voters (41%) prefer Sanders.

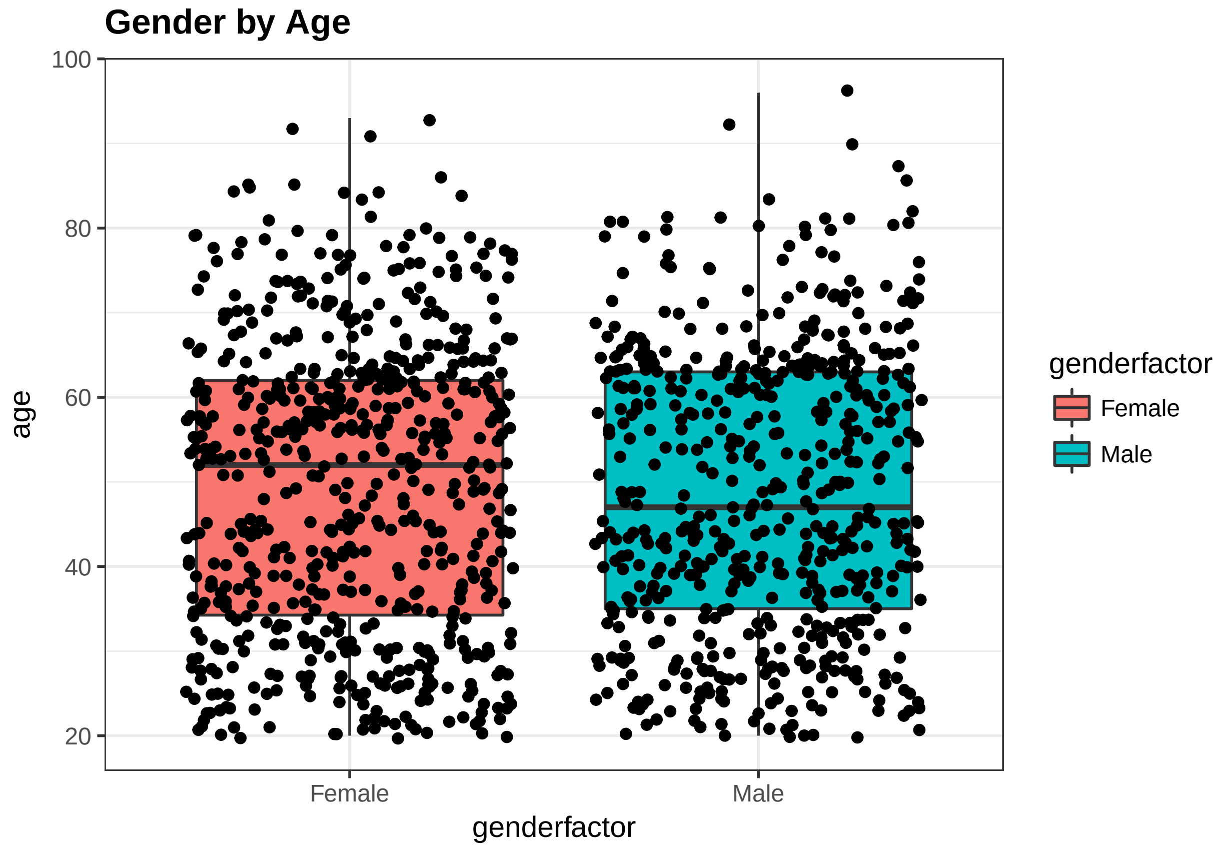
```
age_bin_agg_race <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       race=racefactor), mean))
age_bin_agg_race$n <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       race=racefactor), length))[,3]

ggp <- ggplot(age_bin_agg_race, aes(x=agebin, y=sanders_preference,
                                     color=race, size=n))

ggp + geom_point(aes(color=race))+
  geom_smooth(method="loess", se=F)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age\n(divided by race)")+
  theme(plot.title=element_text(hjust=.5))
```



At first glance the relationship seems very different for non-white voters, but this may be an artifact of the loess smoothing curve attempting to compensate for the data points in the youngest age groups. When looking at the overall distribution of the dots it seems that both white and non-white voters have a negative relationship with age.

Gender

```
ggplot(publicopinion, aes(genderfactor, age)) +
  geom_boxplot(aes(fill = genderfactor)) +
  geom_jitter() +
  ggtitle("Gender by Age") +
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



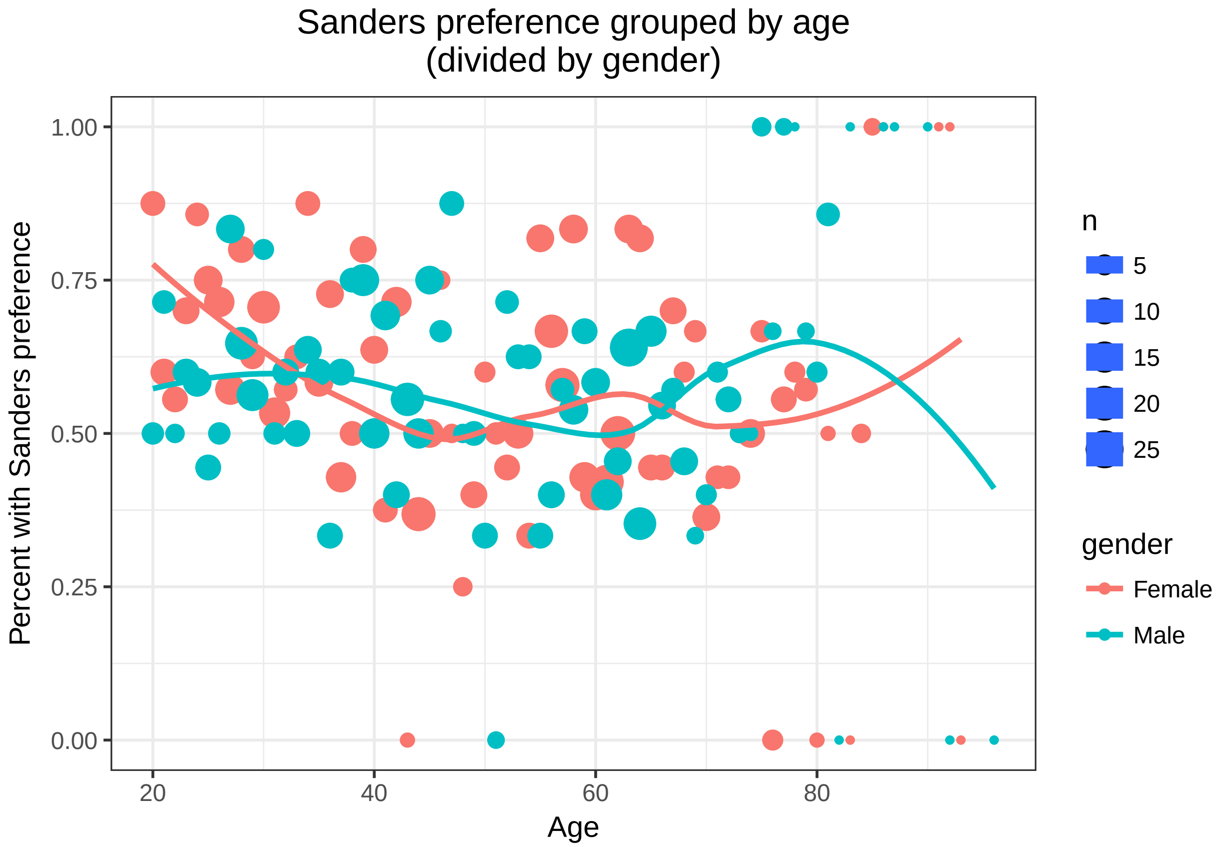
In the dataset women skew towards an older age, median 52 versus 47 for males. We should add age to control for it's effect in our model.

```
po_gender_agg <- with(publicopinion_narm,
                     aggregate(cbind(100*sanders_preference),
                               list(genderfactor=genderfactor),
                               mean))
po_gender_agg$n <- with(publicopinion_narm,
                     aggregate(cbind(100*sanders_preference),
                               list(genderfactor=genderfactor),
                               length))[,2]

po_gender_df <- data.frame(sanders_pref_percent=po_gender_agg$V1,
                          party=po_gender_agg$genderfactor,
                          sample_percent = 100*po_gender_agg$n/1191)
po_gender_df
```

```
##   sanders_pref_percent  party sample_percent
## 1             57.71704 Female       52.22502
## 2             57.46924   Male       47.77498
```

Male and female voters are close to equally represented, both around 50%. Across the sample male and female voters prefer Sanders at almost the same rate (57.5% for male, 57.7% for female)

```
age_bin_agg_gender <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       gender=genderfactor), mean))
age_bin_agg_gender$n <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       gender=genderfactor), length))[,3]

ggp <- ggplot(age_bin_agg_gender, aes(x=agebin, y=sanders_preference,
                                     color=gender, size=n))

ggp + geom_point(aes(color=gender))+
  geom_smooth(method="loess", se=F)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age\n(divided by gender)")+
  theme(plot.title=element_text(hjust=.5))
```



Towards the younger end of the age range, it seems like the negative relationship between sanders preference and age exists mostly for female voters and not so much for male voters. This suggests that investigating a gender by age interaction may be useful.

1.c Alternate models

In our EDA, we determined that race and party had large effects on Sanders preference and would be important to control for. Gender did not seem like it had much explanatory power on its own, although a gender by age interaction seemed plausible. A race by age interacton also looked to be worth testing. Finally, we wanted to test the plausability of a quadratic age term.

Gender by age interaction

```
glm.out.base <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))
glm.out.int <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor + age:genderfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))

summary(glm.out.base)
```

```
##
## Call:
```

```
## glm(formula = sanders_preference ~ age + partyfactor + racefactor + 
##     genderfactor, family = binomial(link = "logit"), data = publicopinion_narm)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7263  -1.1765   0.7857   0.9837   1.7032  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)           -0.058622   0.212382  -0.276 0.782531    
## age                   -0.012602   0.003671  -3.433 0.000598 ***
## partyfactorOther       0.731136   0.141515   5.166 2.39e-07 ***
## partyfactorRepublican  0.601001   0.163208   3.682 0.000231 ***
## racefactorWhite        0.877155   0.142044   6.175 6.61e-10 ***
## genderfactorMale      -0.129182   0.123222  -1.048 0.294472    
## ---
## 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: 1532.1  on 1185  degrees of freedom
## AIC: 1544.1
## 
## Number of Fisher Scoring iterations: 4
```

```r
summary(glm.out.int)
```

```
## 
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor + 
##     genderfactor + age:genderfactor, family = binomial(link = "logit"), 
##     data = publicopinion_narm)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7841  -1.1721   0.8037   0.9526   1.6738  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            0.193790   0.269511   0.719 0.472114    
## age                   -0.017725   0.004981  -3.559 0.000373 ***
## partyfactorOther       0.731612   0.141670   5.164 2.41e-07 ***
## partyfactorRepublican  0.604385   0.163439   3.698 0.000217 ***
## racefactorWhite        0.881578   0.142283   6.196 5.79e-10 ***
## genderfactorMale      -0.675620   0.376790  -1.793 0.072958 .  
## age:genderfactorMale   0.011047   0.007195   1.535 0.124710    
## ---
## 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: 1529.7  on 1184  degrees of freedom
## AIC: 1543.7
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
anova(glm.out.base, glm.out.int, test="LR")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age + partyfactor + racefactor + genderfactor
## Model 2: sanders_preference ~ age + partyfactor + racefactor + genderfactor +
##     age:genderfactor
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1185     1532.1
## 2      1184     1529.7  1   2.3628   0.1243
```

We noticed that the statistical significance of the age term got slightly stronger in the interaction model than in the base model. In the base model, the coefficient for age represented the relationship between age and Sanders preference across the sampl. In the interaction model, it represents the relationship for the base level of gender only, which is female. The coefficient of the interaction term is .011, which means the coefficient for age in males only would be -.018 + .011 = -.007. We can interpret this to mean that the model is suggesting there is a strong negative relationship for female voters but not male voters.

We have decided against this model for a few reasons. First of all, we know from looking at the data and at both models that the effect of gender is very small. Second of all, the p-values given to the interaction term and the likelihood ratio test are not statistically significant even at $p < 0.1$. This tells us that although there may be a hint of a gender interaction term it isn't strong enough to justify including it in our final model.

Party by age interaction

```
glm.out.base <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))
glm.out.int <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor + age:partyfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))

summary(glm.out.base)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor, family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7263  -1.1765   0.7857   0.9837   1.7032
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.058622   0.212382  -0.276 0.782531
## age                   -0.012602   0.003671  -3.433 0.000598 ***
## partyfactorOther       0.731136   0.141515   5.166 2.39e-07 ***
## partyfactorRepublican  0.601001   0.163208   3.682 0.000231 ***
## racefactorWhite        0.877155   0.142044   6.175 6.61e-10 ***
## genderfactorMale      -0.129182   0.123222  -1.048 0.294472
## ---
## 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: 1532.1  on 1185  degrees of freedom
## AIC: 1544.1
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(glm.out.int)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor + age:partyfactor, family = binomial(link = "logit"),
##     data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6971  -1.1605   0.8035   0.9550   1.7720
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.179279   0.300649   0.596  0.55097
## age                       -0.017556   0.005764  -3.046  0.00232 **
## partyfactorOther           0.357773   0.422464   0.847  0.39707
## partyfactorRepublican      0.144621   0.505968   0.286  0.77501
## racefactorWhite            0.878499   0.142220   6.177 6.53e-10 ***
## genderfactorMale          -0.129135   0.123317  -1.047  0.29502
## age:partyfactorOther       0.007719   0.008240   0.937  0.34888
## age:partyfactorRepublican  0.009072   0.009379   0.967  0.33341
## ---
## 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: 1530.8  on 1183  degrees of freedom
## AIC: 1546.8
##
## Number of Fisher Scoring iterations: 4
```

There are a number of reasons that modeling an age by party interaction does not seem like a good idea. First of all, neither of the interacton terms (age:Other, age:Republican) have large effects. Their coefficients are relatively small compared to the original coefficient of the age term, and their p-values are nowhere near statistical significance ($p > .33$). The AIC for the model with the interaction term is larger than for the model without it, suggesting that our additional model complexity is not helping the overall model.

```r
anova(glm.out.base, glm.out.int, test="LR")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age + partyfactor + racefactor + genderfactor
## Model 2: sanders_preference ~ age + partyfactor + racefactor + genderfactor +
##     age:partyfactor
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1185     1532.1
## 2      1183     1530.8  2   1.2789   0.5276
```

The likelihood ratio test, with a p-value of .53, also suggests our interaction model is not more useful than the simpler model. For these reasons we decided to not model an age by party interaction.

Quadratic age term

```
glm.out.base <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))
glm.out.quad <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor + I(age^2), data=publicopinion_narm,
                 family=binomial(link='logit'))

summary(glm.out.base)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor, family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7263  -1.1765   0.7857   0.9837   1.7032
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.058622   0.212382  -0.276 0.782531
## age                    -0.012602   0.003671  -3.433 0.000598 ***
## partyfactorOther        0.731136   0.141515   5.166 2.39e-07 ***
## partyfactorRepublican   0.601001   0.163208   3.682 0.000231 ***
## racefactorWhite         0.877155   0.142044   6.175 6.61e-10 ***
## genderfactorMale       -0.129182   0.123222  -1.048 0.294472
## ---
## 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: 1532.1  on 1185  degrees of freedom
## AIC: 1544.1
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm.out.quad)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor + I(age^2), family = binomial(link = "logit"),
##     data = publicopinion_narm)
##
```

```
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7853  -1.1679   0.7913   0.9462   1.6336  
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            0.8125517  0.5165613   1.573 0.115718    
## age                   -0.0519434  0.0215906  -2.406 0.016136 *  
## partyfactorOther       0.7353441  0.1418181   5.185 2.16e-07 ***
## partyfactorRepublican  0.6031312  0.1633682   3.692 0.000223 ***
## racefactorWhite        0.8722500  0.1425814   6.118 9.50e-10 ***
## genderfactorMale      -0.1209202  0.1234752  -0.979 0.327428    
## I(age^2)               0.0003921  0.0002120   1.849 0.064446 .  
## ---
## 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: 1528.6  on 1184  degrees of freedom
## AIC: 1542.6
## 
## Number of Fisher Scoring iterations: 4
```

```
anova(glm.out.base, glm.out.quad, test="LR")
```

```
## Analysis of Deviance Table
## 
## Model 1: sanders_preference ~ age + partyfactor + racefactor + genderfactor
## Model 2: sanders_preference ~ age + partyfactor + racefactor + genderfactor + 
##     I(age^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1      1185     1532.1                       
## 2      1184     1528.6  1   3.4751   0.0623 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Like the EDA showed, the model is showing that a quadratic age term might be plausible. The significance of the term in the model is slightly larger than .05 (.06), as is the signifiance of the likelihood ratio test (.06). The AIC of the model with the quadratic term is also lower.

However, we decided to not to include the quadratic age term in the end. Aside from the lack of statistical significance - although it is close - the main reason for this is the lack of representation in the older age range that is driving this result. We would not feel comfortable recommending this model and suggesting that older voters be targeted when we have so few older people that are contributing to this trend.

1.d Selected model results

1.e Statistical Tests

In the subsections below we perform two statistical tests on our model: Wald test and Likelihood Ratio Test.

Wald Test

The summary of our model actually displays the statistics for the Wald Test. Recall that the Wald statistic is given by

$$Z_0 = \frac{\beta_r \hat{-} \beta_r}{\sqrt{Var(\hat{\beta_r})}}$$

We use this statistic to test the null hypothesis $H_0 : \beta_r = 0$ versus the alternat hypothesis $H_a : \beta_r \neq 0$.

For large samples, this test statistic has an approximate standard normal distribution if the null hypothesis is true, and we reject the null hypothesis if the statistic value is unexpected for a standard normal distribution.

Now, since the Walrd test statistic is provided automatically for each individual β parameter, we summarize the model again with the purpose of analyzing the Wald statistics.

```r
summary(model.final)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + racefactor + partyfactor,
##     family = binomial(link = logit), data = publicopinion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7036  -1.1792   0.7907   0.9881   1.6662
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.115017   0.205360  -0.560 0.575428
## age                  -0.012480   0.003666  -3.404 0.000664 ***
## racefactorWhite       0.872782   0.141872   6.152 7.66e-10 ***
## partyfactorOther      0.713501   0.140368   5.083 3.71e-07 ***
## partyfactorRepublican  0.594231   0.162972   3.646 0.000266 ***
## ---
## 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: 1533.2  on 1186  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 1543.2
##
## Number of Fisher Scoring iterations: 4
```

We can observe that for each explanatory variable there is a hypothesis test, with its associated p-value.

For example, for the *age* explanatory variable, we observe that the p-value is 0.000664, which is highly statistically significant, so we can reject the null hypothesis H_0 that $\beta_{age} = 0$, which can be interpreted as there is sufficient statistical evidence to indicate that age has an effect on the probability of voters supporting Sanders. For the other explanatory variables we analogously reject the null hypotheses that their corresponding β equal 0, since all the statistics are highly statistically significant.

Likelihood Ratio Test

We now perform likelihood ratio tests on our model. The LRT statistic is defined as

$$\Lambda = \frac{Maximum\ of\ likelihood\ function\ under\ H_0}{Maximum\ of\ likelihood\ function\ under\ H_0\ or\ H_a}$$

The test, as with the Wald statistic, is for $H_0 : \beta_r = 0$ versus $H_a : \beta_r \neq 0$.

We then calculate $-2log(\Lambda)$, and if the null hypothesis is true, then $-2log(\Lambda)$ has an approximate χ^2_1 distribution for a large sample.

In R, we can perform the likelihood ratio test using the Anova function authored by Professor John Fox as part of the Car package. Below are the results of our test.

```
Anova(model.final)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: sanders_preference
##              LR Chisq Df Pr(>Chisq)
## age            11.710  1  0.0006217 ***
## racefactor     38.468  1  5.565e-10 ***
## partyfactor    28.509  2  6.447e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For example, for the *age* explanatory variable, we obtain a statistically significant p-value 0.0006217, rejecting the null hypothesis that $\beta_{age} = 0$. For Race and Party we also obtained statistically significant p-values and we can also reject their null hypothesis, which means that there is evidence that each explanatory variable has an effect on the probability of voters supporting Sanders over Clinton.

1.f Age Interpretation and Odds Ratios

Odds Ratios

First, let's obtain an expression for the effect in the odds of supporting Sanders caused by a *c* year change in age. For this, we calculate the odds ratio:

$$OR = \frac{Odds_{age+c}}{Odds_{age}} = \frac{e^{\beta_0+\beta_{age}(age+c)+\beta_{race.white}race.white+\beta_{party.republican}party.republican+\beta_{party.independent}party.independent}}{e^{\beta_0+\beta_{age}(age+c)+\beta_{race.white}race.white+\beta_{party.republican}party.republican+\beta_{party.independent}party.independent}}$$

Which using the properties of exponentiation can be simplified to:

$$OR = e^{c\beta_{age}}$$

For example, we can calculate the Odds Ratio for a 10 year decrease in age by inverting the formula:

```
1 / exp(model.final$coefficients[2] * 10)
```

```
##      age
## 1.132925
```

This Odd Ratio can be interpreted as: the odds of supporting sanders are 1.132925 times larger for every 10 year decrease in age of the voters.

Odds Ratios and Confidence Intervals

To include confidence intervals in our odd ratios, we use the Wald confidence interval, which comes from the following expression:

$$c\hat{\beta_{age}} \pm cZ_{1-\alpha/2}\sqrt{(}Var(\hat{\beta_{age}}))$$

This means we need the variance of β_{age}, which we can obtain from the variance-covariance matrix for our model:

```
vcov(model.final)
```

```
##                        (Intercept)           age racefactorWhite
## (Intercept)           0.0421728565 -5.850979e-04   -0.0074979329
## age                  -0.0005850979  1.344168e-05   -0.0001025566
## racefactorWhite      -0.0074979329 -1.025566e-04    0.0201277092
## partyfactorOther     -0.0080500132  6.879216e-06   -0.0025705793
## partyfactorRepublican -0.0047833711 -3.831271e-05   -0.0043273448
##                      partyfactorOther partyfactorRepublican
## (Intercept)             -8.050013e-03         -4.783371e-03
## age                      6.879216e-06         -3.831271e-05
## racefactorWhite         -2.570579e-03         -4.327345e-03
## partyfactorOther         1.970331e-02          9.889343e-03
## partyfactorRepublican    9.889343e-03          2.655982e-02
```

Where $var(\hat{\beta_{age}})$ is in the diagonal, with value $1.344168e - 05$.

Now we can calculate the intervals:

```
c = 10
alpha = 0.05
confint = c * qnorm(1 - alpha/2) * sqrt(1.344168e-05)

odds.ratio = 1 / exp(model.final$coefficients[2] * 10)

lower = odds.ratio - confint
upper = odds.ratio + confint

odds.ratio
```

```
##      age
## 1.132925
```

```
lower
```

```
##      age
## 1.061067
```

```
upper
```

```
##      age
## 1.204783
```

So going back to our interpretation, the odds of supporting sanders are 1.132925 times larger for every 10 year decrease in age of the voters. This ratio, for the 95% confidence interval, can be found between 1.204783 and 1.061067.

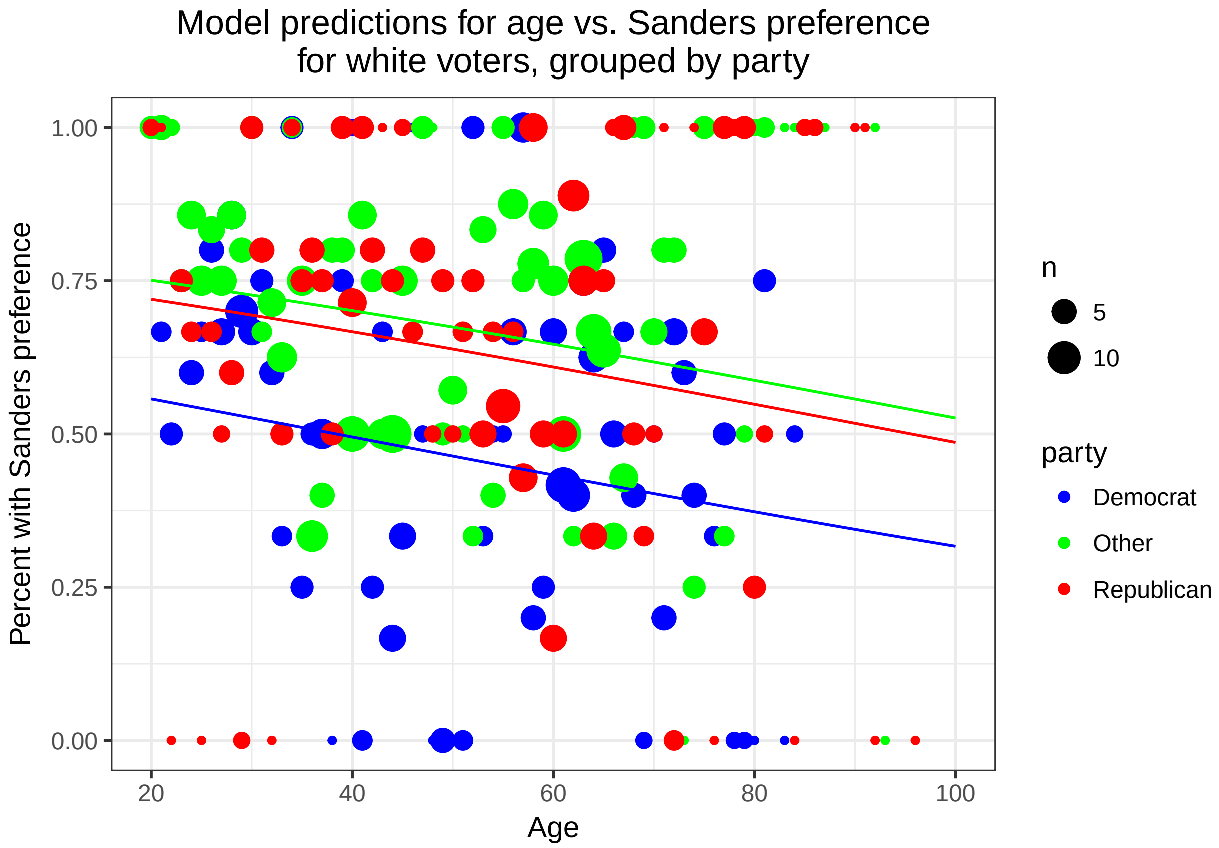
2 Plot: Age vs Predicted probability of supporting Sanders

```
#create variable a to represent ages 20 to 100
#create different y values for each subgroup of party and race - 6 total
#using the model predited probabilities for sanders_preference
a = c(20:100)
y_dem_nw = exp(model.final$coefficients[1] + model.final$coefficients[2]*a)/
       (1+exp(model.final$coefficients[1] + model.final$coefficients[2]*a))

y_oth_nw = exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[3])/
       (1+exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[3]))

y_rep_nw = exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[4])/
       (1+exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[4]))



y_dem_w = exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                model.final$coefficients[5])/
       (1+exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                model.final$coefficients[5]))

y_oth_w = exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[3] +
                model.final$coefficients[5])/
       (1+exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[3] +
                model.final$coefficients[5]))

y_rep_w = exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[4] +
                model.final$coefficients[5])/
       (1+exp(model.final$coefficients[1] + model.final$coefficients[2]*a +
                 model.final$coefficients[4] +
                model.final$coefficients[5]))
```

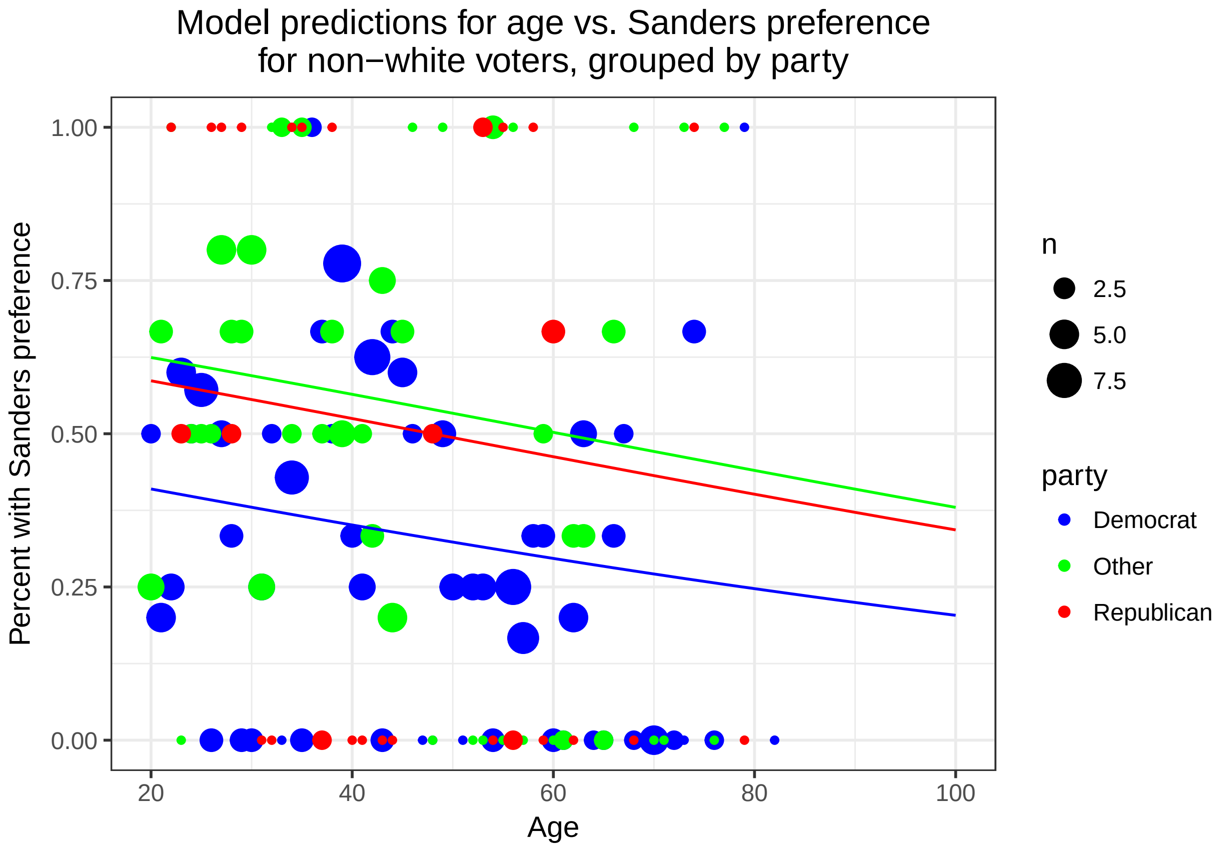
```
age_bin_agg_all <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       party=partyfactor,
                                       race=racefactor), mean))
age_bin_agg_all$n <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       party=partyfactor,
                                       race=racefactor), length))[,4]
```

```
#plot data and predictions for white voters
ggp <- ggplot(age_bin_agg_all[age_bin_agg_all$race=="White",], aes(x=agebin, y=sanders_preference,
                                     color=party, size=n))
```

```
ggp + geom_point()+
  geom_path(inherit.aes=F, data=data.frame(x=a, y=y_dem_w),
            aes(x = a, y = y), color="blue")+
  geom_path(inherit.aes=F, data=data.frame(x=a, y=y_oth_w),
            aes(x = a, y = y), color="green")+
    geom_path(inherit.aes=F, data=data.frame(x=a, y=y_rep_w),
            aes(x = a, y = y), color="red")+
  #facet_grid(party~.)+
  scale_color_manual(values=c("blue", "green", "red"))+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Model predictions for age vs. Sanders preference\nfor white voters, grouped by party")+
  theme(plot.title=element_text(hjust=.5))
```



```
#plot data and prediction for non-white voters
ggp <- ggplot(age_bin_agg_all[age_bin_agg_all$race=="Non-White",], aes(x=agebin, y=sanders_preference,
                                    color=party, size=n))

ggp + geom_point()+
  geom_path(inherit.aes=F, data=data.frame(x=a, y=y_dem_nw),
            aes(x = a, y = y), color="blue")+
  geom_path(inherit.aes=F, data=data.frame(x=a, y=y_oth_nw),
            aes(x = a, y = y), color="green")+
    geom_path(inherit.aes=F, data=data.frame(x=a, y=y_rep_nw),
            aes(x = a, y = y), color="red")+
  #facet_grid(party~.)+
  scale_color_manual(values=c("blue", "green", "red"))+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Model predictions for age vs. Sanders preference\nfor non-white voters, grouped by party")+
  theme(plot.title=element_text(hjust=.5))
```



The above plots show the original data with a bubble for every age group, with each color representing a different political party. For the sake of not having crowded graphs we've separated the white and non-white voters into two different graphs.

Because our final model did not include interaction terms, all 6 lines - i.e. the models' predictions for each subgroup - are parallel. The model predicts the same age-preference relationship for all subgroups, and based on the subgroups boosts or lowers the probability (other > republican > democrat, and white > non-white)

3 Comment on Importance of Age and Client recommendation

The results of the model and the graphs can help inform our client on who to target for this marketing campaign. The model does suggest that younger voters are more likely to support Sanders, as each 10 year decrease in age corresponds to a 1.13 increase in the odds of supporting Sanders.

The significant differences in supporting Sanders when looking at diferent party affiliations and races can help our client arrive at more targeted campaigns. Non-white Democrats appear to be the group least likely to support Sanders, as our model predicts them to have a less than 50% probability of supporting Sanders even at their youngest age range. (A 20 year old non-white Democrat is predicted to have a 41.0% chance of supporting Sanders). Non-white voters of other parties are also less likely to support Sanders than white voters, but our model does predict rates above 50% once you get below a certain age (Non-white Republicans age 39 and younger, as well as Non-white Other/Independents age 48 and younger are predicted to support Sanders with a probability greater than 50%)

Our client should attempt to target white voters who, across the sample, see an increase of 2.39 in odds of supporting Sanders compared to non-white voters. Our model predicts that white Republicans and Other/Independents have a greater than 50% chance of supporting Sanders across all age groups. For white Democrats, it appears that voters of age 61 and less are more than 50% likely to support Sanders.

All that being said, if our client is interested in targeting voters who, broadly speaking, support politically liberal candidates, these recommendations need to be taken with a grain of salt. Our model shows that Republicans are far more likely to prefer Sanders than Democrats. However, given the nature of the question and the typical stance of Republican voters, this is likely more due to being against Clinton than supporting Sanders. It would likely be unwise to target Republicans in a campaign for liberal merchandise despite what these data suggest. Independent voters would likely be a good target. And as the model suggests, younger voters are definitely more likely to prefer Sanders. Our model predictions are able to provide ages within each subgroup where the chance of preferring Sanders rises above 50%, which could be natural cutoffs for a targeted marketing campaign.