w271_2016Fall_Lab1_LaneJ

September 18, 2016

Part 1: Ideology and Candidate Support

1: As in any data science project, check your data, such as examining the structure and the integrity of the data (including the number of observations, number of variables, types of the variables, number of missing values (or oddly coded values) in each of the variables, descriptive statistics of each of the variables, etc). Do not simply print tables and summary statistics without providing context and discussing what conclusions you drew from the initial exploration

```
summary(us_public_opinion)
```

```
##      fthrc          ftsanders          ideo5            pid3
##  Min.   :  0.00   Min.   :  0.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:  3.00   1st Qu.: 19.00   1st Qu.:2.000   1st Qu.:1.000
##  Median : 44.00   Median : 51.00   Median :3.000   Median :2.000
##  Mean   : 43.79   Mean   : 56.73   Mean   :3.234   Mean   :2.072
##  3rd Qu.: 76.00   3rd Qu.: 82.00   3rd Qu.:4.000   3rd Qu.:3.000
##  Max.   :998.00   Max.   :998.00   Max.   :6.000   Max.   :5.000
##    race_white         gender         birthyr
##  Min.   :0.0000   Min.   :1.000   Min.   :1921
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1955
##  Median :1.0000   Median :2.000   Median :1968
##  Mean   :0.7292   Mean   :1.525   Mean   :1968
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:1982
##  Max.   :1.0000   Max.   :2.000   Max.   :1997
```

```
d = describe(us_public_opinion)

print("Both ftsanders and ftclinton are arguably categorical ordinal variables since they measure senti
```

```
## [1] "Both ftsanders and ftclinton are arguably categorical ordinal variables since they measure senti
```

```
print(d[1:2])
```

```
## us_public_opinion
##
##  2  Variables      1200  Observations
## ---------------------------------------------------------------------------
## fthrc
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1200       0     102       1   43.79       0       0       3      44
##     .75     .90     .95
##      76      95     100
##
```

```
## lowest :   0   1   2   3   4, highest:  97  98  99 100 998
## ---------------------------------------------------------------------------
## ftsanders
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1200       0     102       1   56.73       0       2      19      51
##     .75     .90     .95
##      82      97     100
##
## lowest :   0   1   2   3   4, highest:  97  98  99 100 998
## ---------------------------------------------------------------------------
```

```
print("It looks like fthrc and ftsanders have a few values of 998.  This is either a data error or some
```

```
## [1] "It looks like fthrc and ftsanders have a few values of 998.  This is either a data error or som
```

```
data = us_public_opinion[us_public_opinion$fthrc <= 100 & us_public_opinion$ftsanders <= 100,]
print("")
```

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

```
print("ideo5 is ordinal on a 5 point scale, with 6 representing an NA value, so I will convert it to a
```

```
## [1] "ideo5 is ordinal on a 5 point scale, with 6 representing an NA value, so I will convert it to a
```

```
print(d[3])
```

```
## us_public_opinion
##
##  1  Variables      1200  Observations
## ---------------------------------------------------------------------------
## ideo5
##       n missing  unique    Info    Mean
##    1200       0       6    0.95   3.234
##
##             1   2   3   4   5  6
## Frequency 142 208 378 261 121 90
## %          12  17  32  22  10  8
## ---------------------------------------------------------------------------
```

```
data$ideo5 = as.factor(data$ideo5)
levels(data$ideo5) = c("Very Liberal","Liberal","Independent","Conservative","Very Conservative","Not S
print("")
```

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

```
print("pid3 is a simple categorical, so convert it to a factor")
```

```
## [1] "pid3 is a simple categorical, so convert it to a factor"
```

```
print(d[4])
```

```
## us_public_opinion
##
##  1  Variables      1200  Observations
## ---------------------------------------------------------------------------
## pid3
##       n missing  unique    Info    Mean
##    1200       0       5     0.9   2.072
##
##             1   2   3  4 5
## Frequency 459 280 380 77 4
## %          38  23  32  6 0
## ---------------------------------------------------------------------------
```

```
data$pid3 = as.factor(data$pid3)
levels(data$pid3) = c("Democrat","Republican","Independent","Other Party","Not sure")
print("")
```

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

```
print("race_white and gender are both binary variables.  I'm going to subtract one from gender to make i
```

```
## [1] "race_white and gender are both binary variables.  I'm going to subtract one from gender to make
```

```
print(d[5:6])
```

```
## us_public_opinion
##
##  2  Variables      1200  Observations
## ---------------------------------------------------------------------------
## race_white
##       n missing  unique    Info     Sum    Mean
##    1200       0       2    0.59     875  0.7292
## ---------------------------------------------------------------------------
## gender
##       n missing  unique    Info    Mean
##    1200       0       2    0.75   1.525
##
## 1 (570, 48%), 2 (630, 52%)
## ---------------------------------------------------------------------------
```

```
data$race_white = as.factor(data$race_white)
levels(data$race_white) = c("Non-white","White")
data$gender = as.factor(data$gender -1)
levels(data$gender) = c("Male","Female")
print("")
```

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

```r
print("birthyear is a metric variable, but it will severly throw off the intercept, which already needs
```

```
## [1] "birthyear is a metric variable, but it will severly throw off the intercept, which already needs
```

```r
print(d[7])
```

```
## us_public_opinion
##
##  1  Variables      1200  Observations
## ---------------------------------------------------------------------------
## birthyr
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1200       0      73       1    1968    1940    1946    1955    1968
##     .75     .90     .95
##    1982    1991    1994
##
## lowest : 1921 1924 1925 1926 1927, highest: 1993 1994 1995 1996 1997
## ---------------------------------------------------------------------------
```

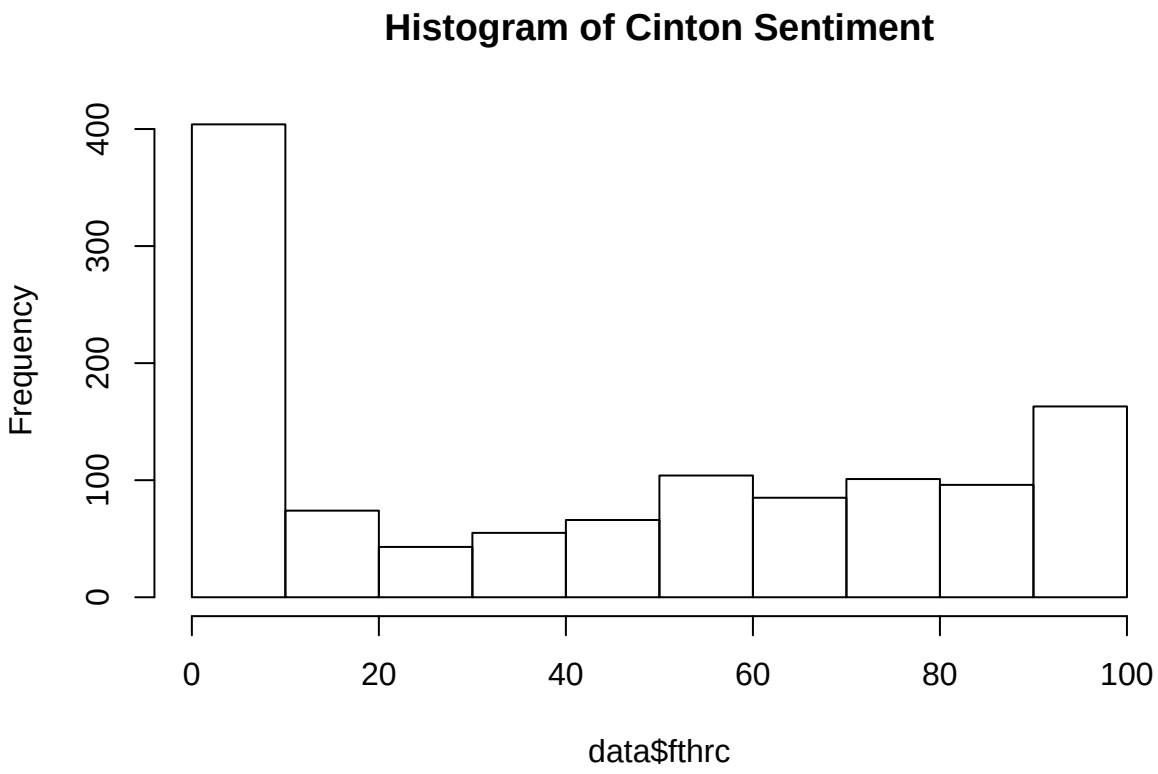
```r
data$age = 2016 - data$birthyr
data$birthyr = NULL
```

2: As emphasized throughout the course, we do not start from building statistical models right away. Instead, we first thoroughly examine the data: Conduct Exploratory Data Analysis (EDA) on the Dependent Variable

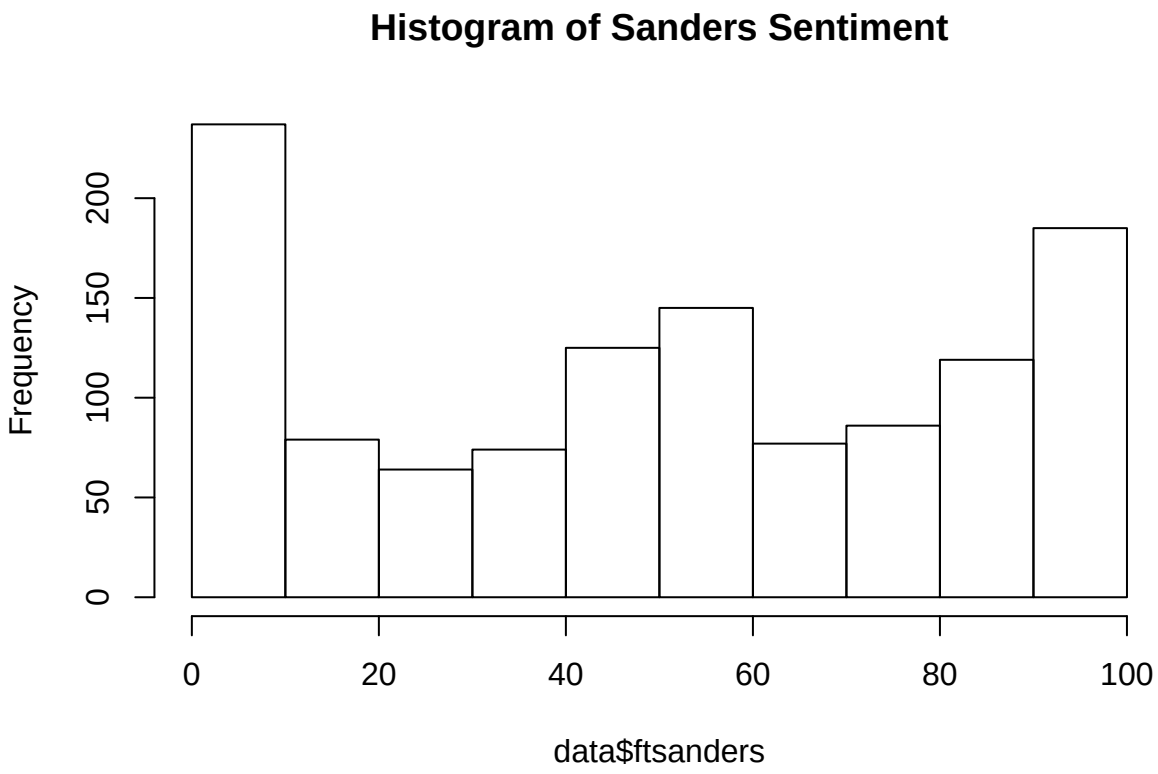
a. Examine the distribution of the variables fthrc and ftsanders. Comment on their distributions.

Both variables are skewed towards the extremes, which is understandable given that they aren't really metric variables. The Sanders variable is less extreme than the Clinton variable.

```r
hist(data$fthrc,main="Histogram of Cinton Sentiment")
```



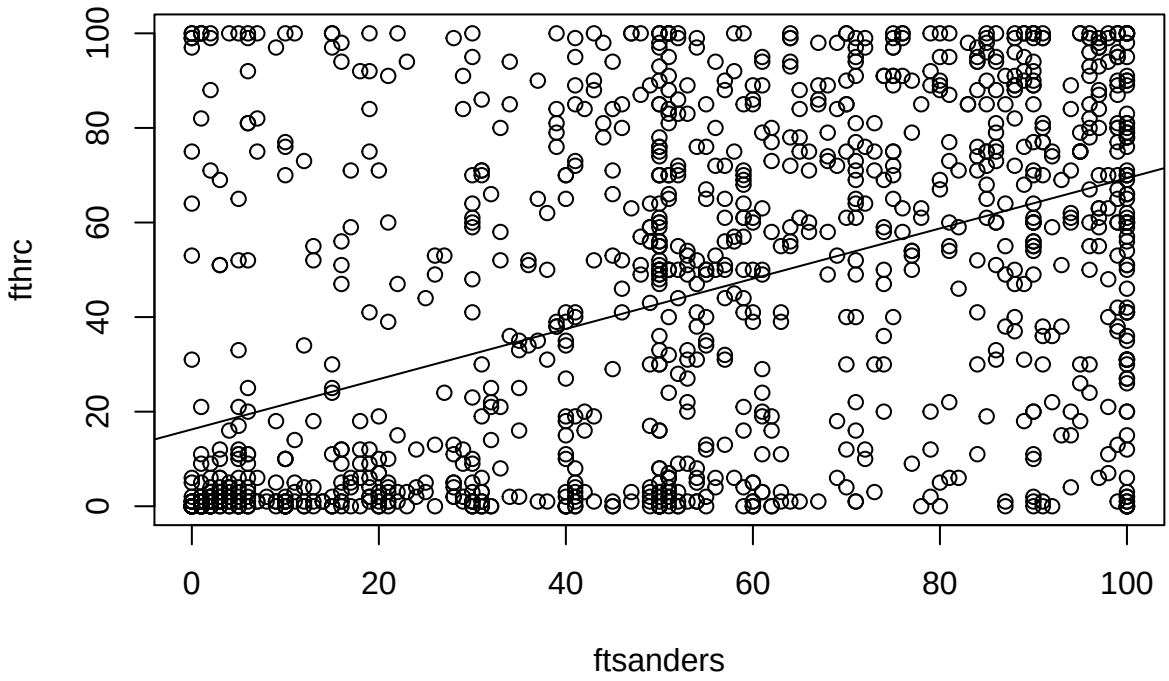

```
hist(data$ftsanders,main="Histogram of Sanders Sentiment")
```



b. Examine the relationship between fthrc and ftsanders and comment on their relationship.

There is a positive relationship between Clinton and Sanders sentiments. It's not a very strong relationship, as there are plenty of points that don't fit the trend line.

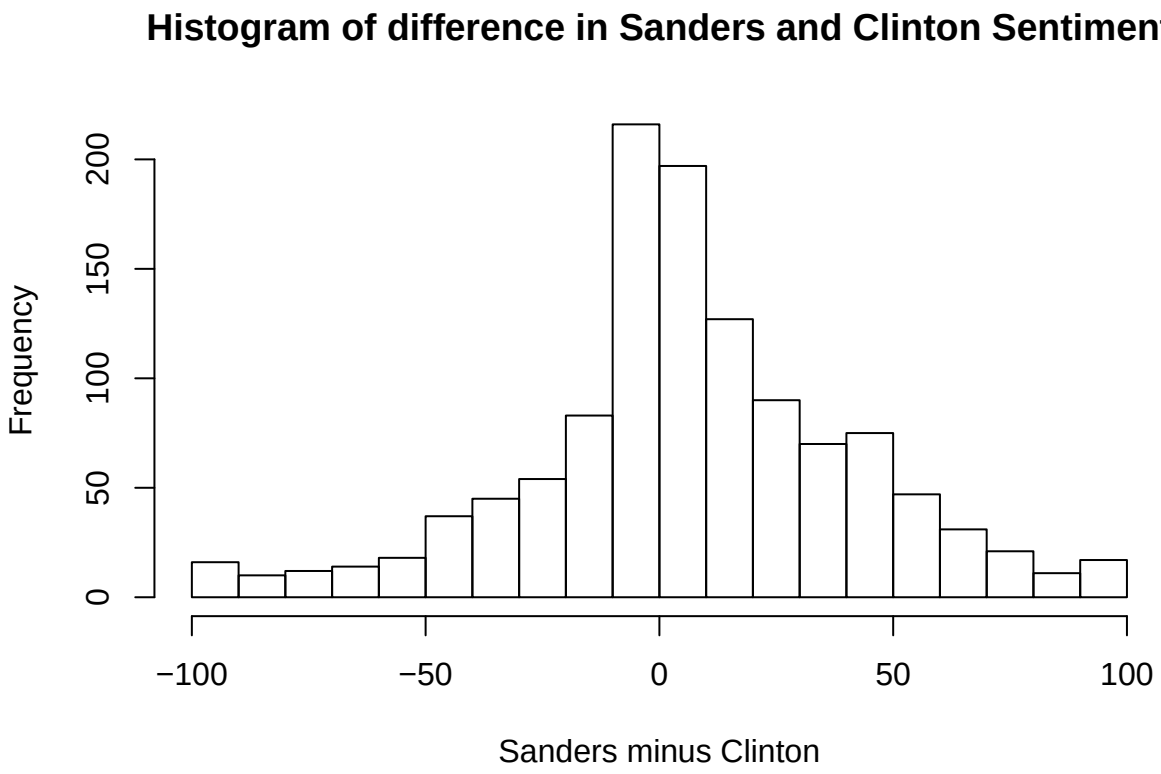
```
plot(fthrc~ftsanders,data=data)
abline(lm(fthrc~ftsanders,data))
```



c. Create a variable called diff, which is the difference between ftsanders and fthrc. Examine this new variable. What does this variable mean? Is there anything noteworthy about its distribution?

This distribution looks more normal, although definitely skewed positive (towards Sanders) and still slightly weighted towards the extremes. This variable shows the difference in opinion people have between and Sanders and Clinton

```
data$diff = data$ftsanders - data$fthrc
hist(data$diff,main="Histogram of difference in Sanders and Clinton Sentiment", xlab="Sanders minus Cli
```



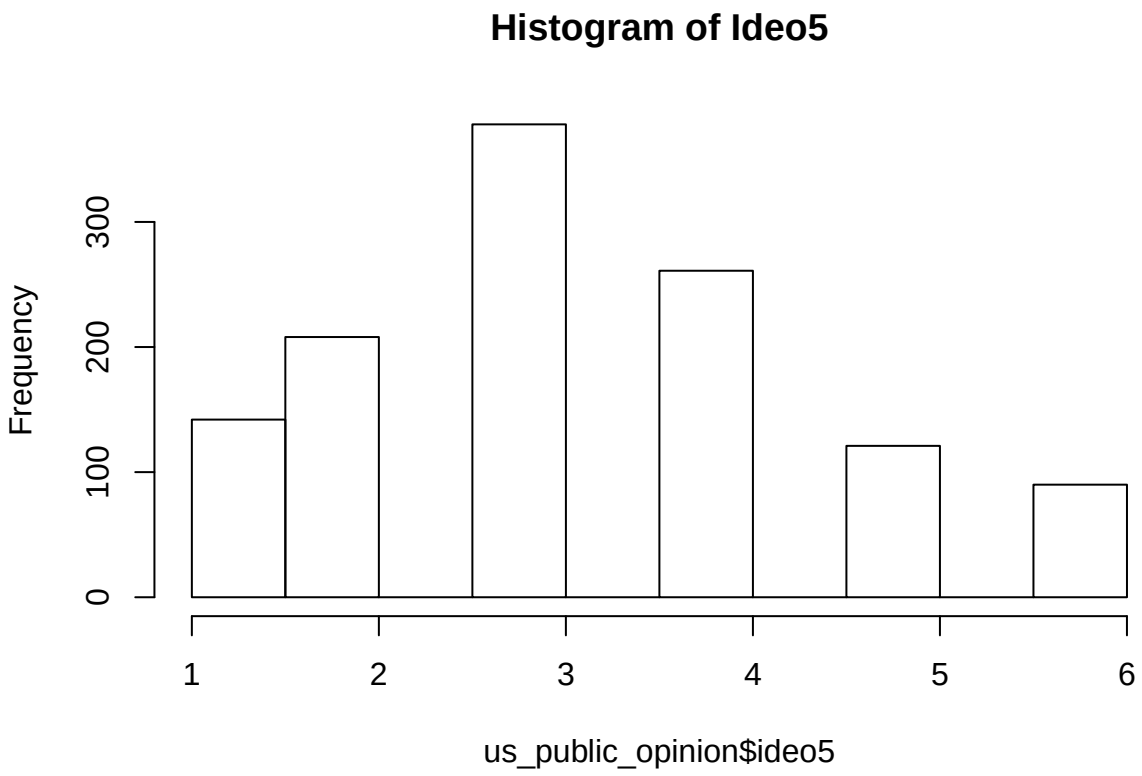
d. We could actually answer this question without creating this variable. Please describe either an alternative coding or alternative modeling strategy.

Instead of recording the sentiment towards each candidate, you could just record the sentiment towards Sanders and then the difference in sentiment towards Clinton.

3. Conduct EDA on our explanatory variable of interest: Ideology

a. Visually inspect and comment on the distribution of the variable, ideo5. Would you include this variable "as is" in a model or does it require any sort of transformation?

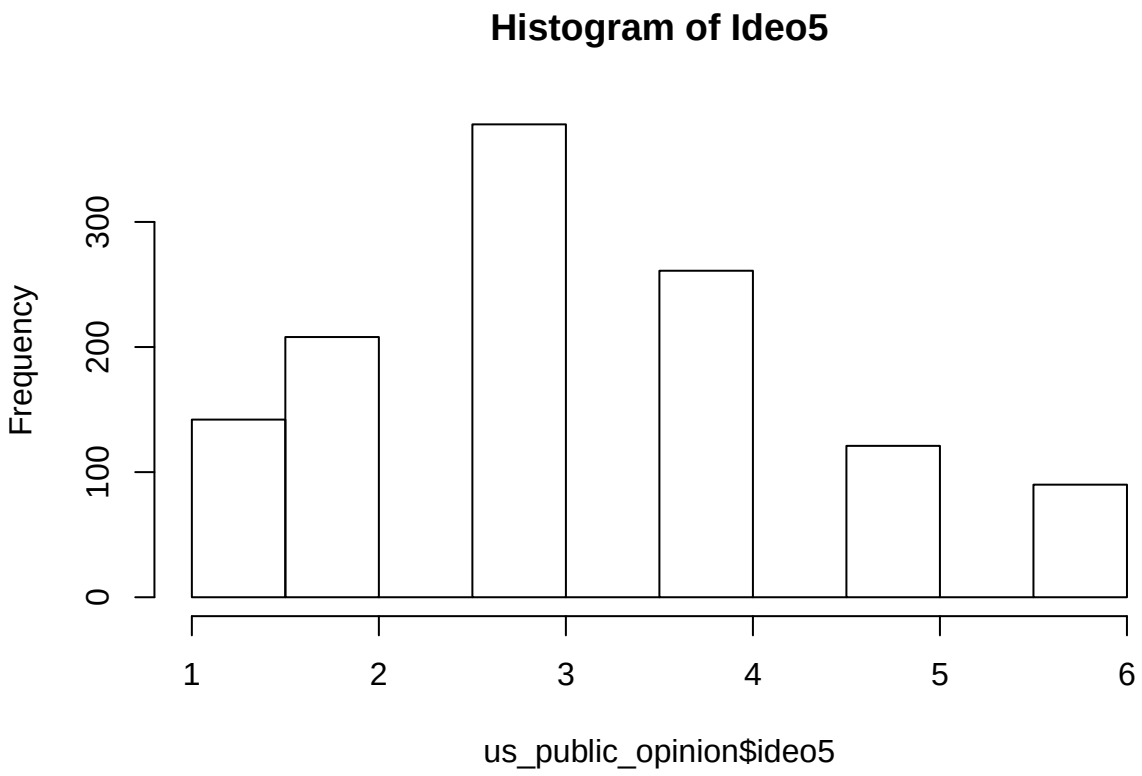
```
hist(us_public_opinion$ideo5,main="Histogram of Ideo5")
```



Definitely can't use the variable as-is. It's categorical and has an NA value encoded as one of the numbers. See problem 1 for the transformation

b. How would you describe the ideological distribution of American voters?

```
hist(us_public_opinion$ideo5,main="Histogram of Ideo5")
```



It looks like there were much more conservatives surveyed than liberals It also looks like at least half of the respondents said they were independent. A small, but still noticeable percentage of people are not sure. See problem 1 for the transformation.

4. Conduct EDA on other explanatory variables

a. Create a variable for age.

See problem 1 for transformation

```
head(data$age)
```

```
## [1] 56 59 53 36 42 58
```

b. Create a dummy variable for gender that takes a value of one if the respondent is female and is zero otherwise.

See problem 1 for transformations.

```
head(data$gender)
```

```
## [1] Male   Female Male   Male   Male   Male
## Levels: Male Female
```

c. Examine each of these explanatory variables. Do you think they require transformation, including binning the variable or creating an additional indicator variable to capture a mass of values, if applicable?

No. The 0-1 encoded *gender* variable we created is already an indicator variable, and there's no further transformation that can be done. While age is often binned or converted to ordinal in other studies, there's nothing to suggest that this type of transformation is needed here, so I'm going to leave *age* as a metric variable for now.

d. What is the average age of respondents?

```
mean(data$age)
```

```
## [1] 48.04114
```

e. What proportion of respondents are white?

```
mean(as.numeric(data$race_white)-1)
```

```
## [1] 0.729639
```

f. What proportion of respondents are female?

```
mean(as.numeric(data$gender)-1)
```

```
## [1] 0.5222502
```

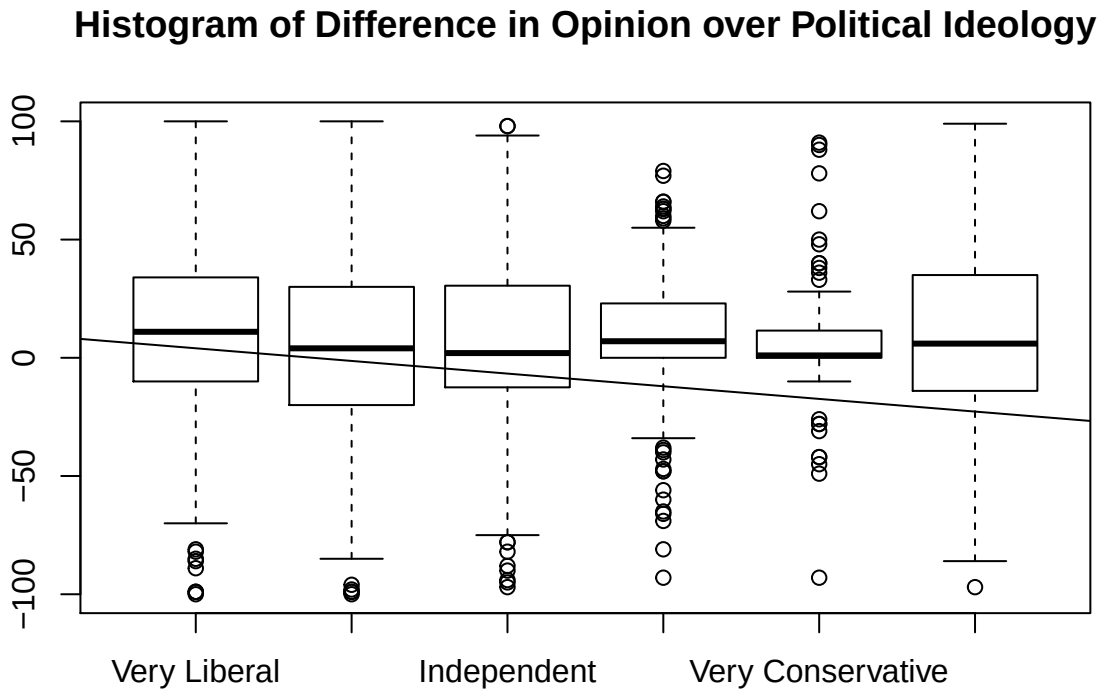
5. Examine Bivariate and Multivariate Relationships, including, but not limited to, the following:

a. Examine the bivariate relationship between the dependent variable and ideology. Based on this initial exploration, do liberal voters have a higher level of support for Sanders over Clinton?

We need to regress *diff* on *ideo5*. The null hypothesis is that the coefficient on *ideo5* is 0. The resulting trend line is above 0 all the way through, meaning that all voters on average tend to have higher levels of support for Sanders over Clinton. The trend is also positive, meaning that liberal voters have more similar opinions of Clinton and Sanders.

```
plot(data$ideo5,data$diff, main="Histogram of Difference in Opinion over Political Ideology")
abline(lm(diff~ideo5,data=data))
```

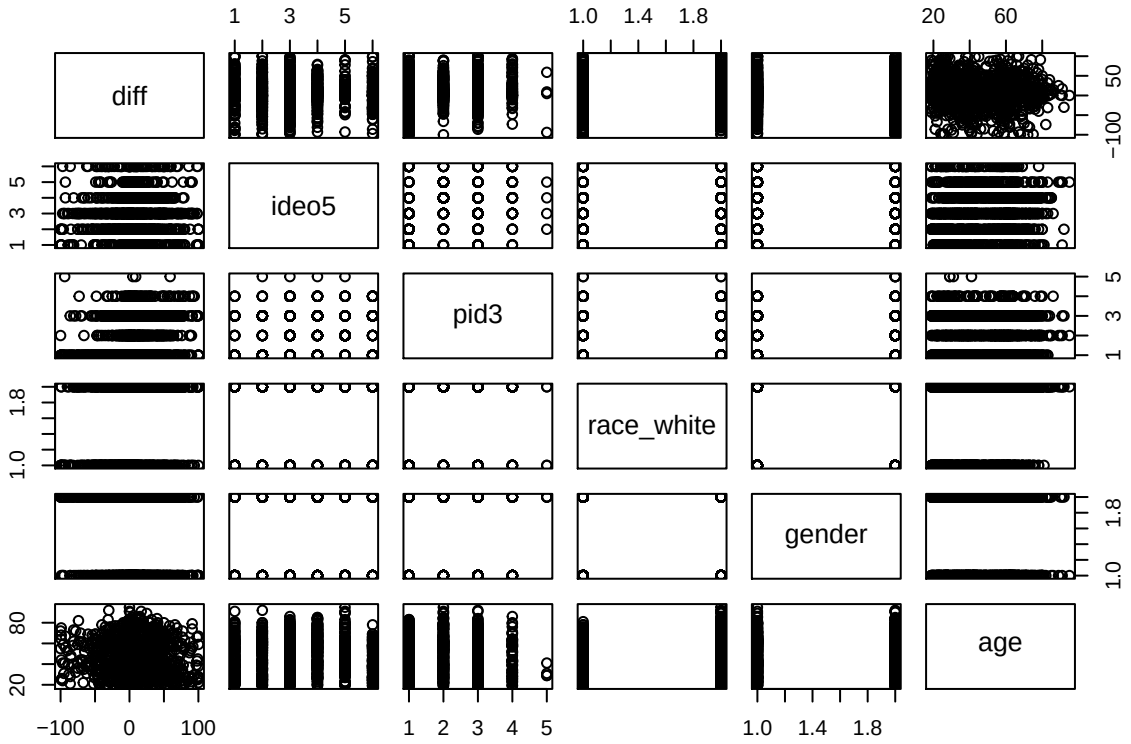
```
## Warning in abline(lm(diff ~ ideo5, data = data)): only using the first two
## of 6 regression coefficients
```



b. Examine (and comment) on the bivariate relationships between the dependent variable and each of the explanatory variables as well as among the explanatory variables themselves. Comment on each of these relationships. Are there any transformations and/or creation of additional variables that you think maybe useful? Might multicolleniarity be a problem? If so, how would it impact your model's results? (Note that the classical linear regression model does not specify that each and everyone of these relationships has to be linear. The most obvious case is binary explanatory variable; it clearly is not related to the dependent variable linearly. The CLM just assumes that the conditional expectation function is a linear, conditional on the set of explanatory variables.)

First, here is just a general scatterplot matrix. Generally, they are pretty useless since most of the explanatory variables are categorical or ordinal, but R doesn't adjust its plot type if you run the command at the data frame level. However, each scatter plot has at least points in at least four different places, so we know there is no bivariate multicolleniarity.

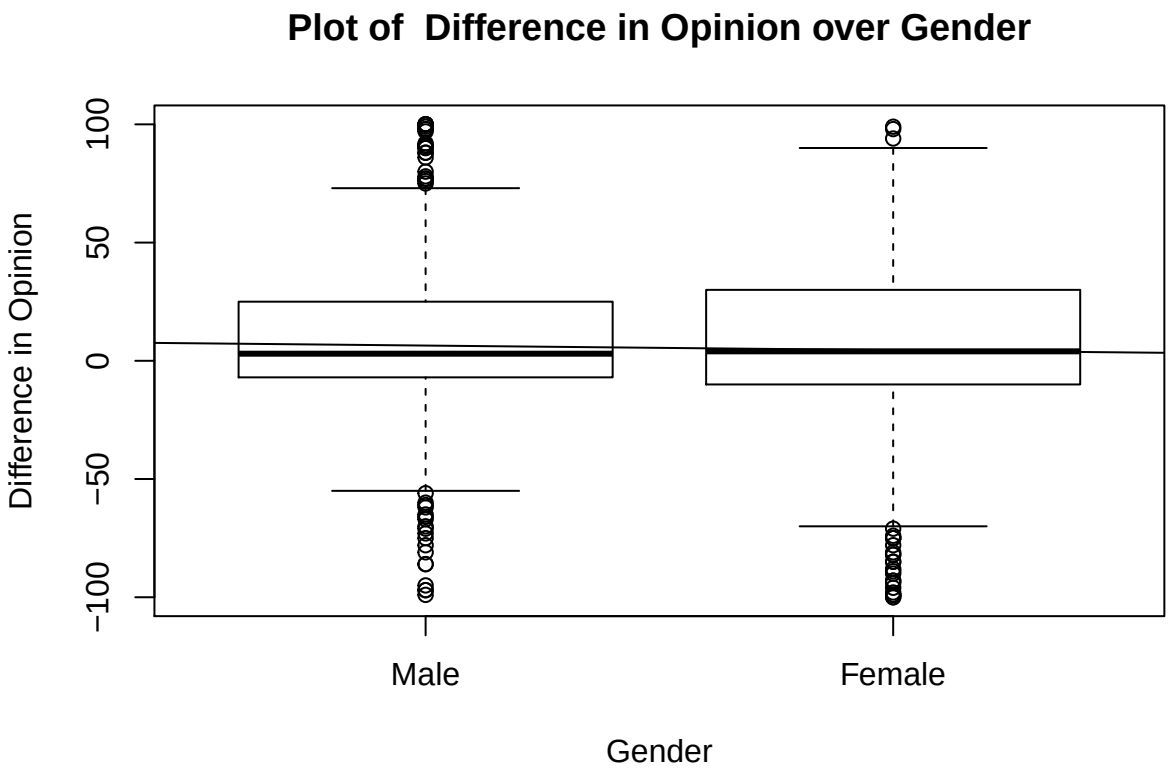
```
plot(data[,c("diff","ideo5","pid3","race_white","gender","age")])
```



More detailed analysis

Between *diff* and *gender*, there doesn't seem to be much of a relationship. The trend line stays above 0, but it has a slight negative slope, meaning that females tend to have more similiar opinions of Sanders and Clinton

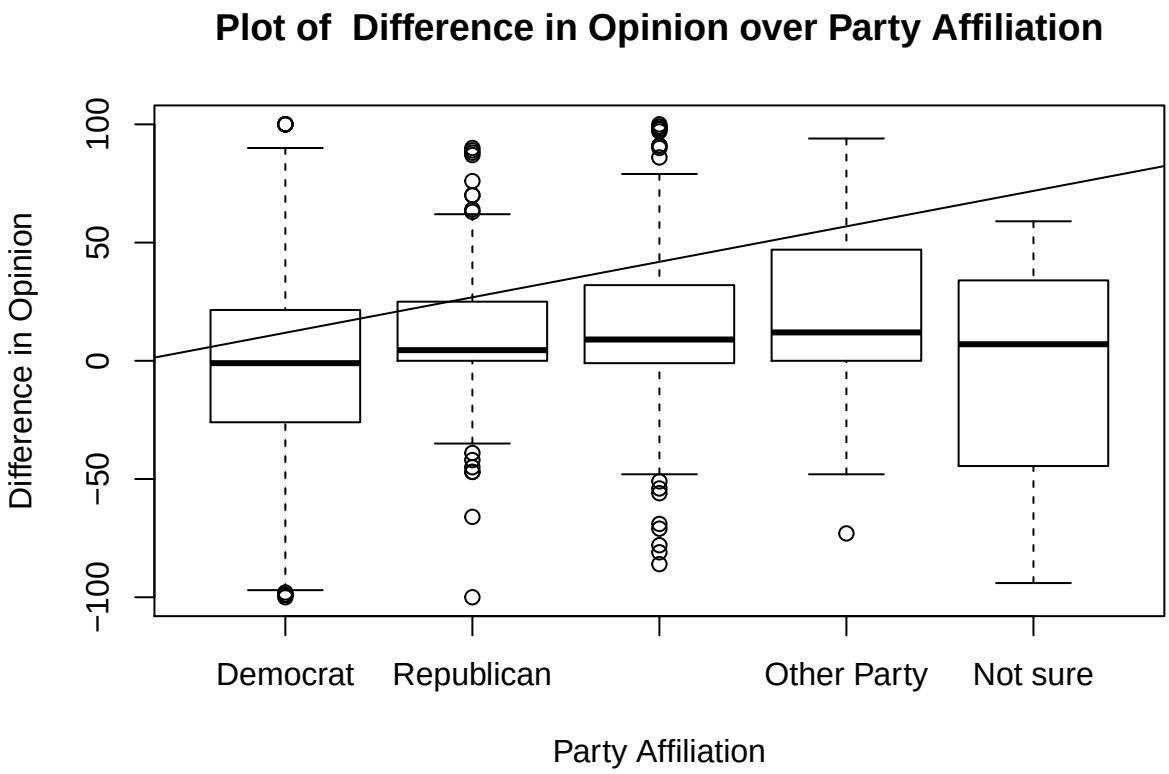
```
plotandtrend = function(x,y,xlab,ylab){
  plot(x,y,xlab = xlab,ylab = ylab, main = paste("Plot of ", ylab,"over",xlab))
  abline(lm(y~x))
}


plotandtrend(data$gender,data$diff,"Gender","Difference in Opinion")
```



Comparing *diff* and *pid3*, it looks like people who identify as Democrats tend to have the least difference in opinion between Clinton and Sanders, while people who identify as Other or Independent have the greatest difference in opinion.

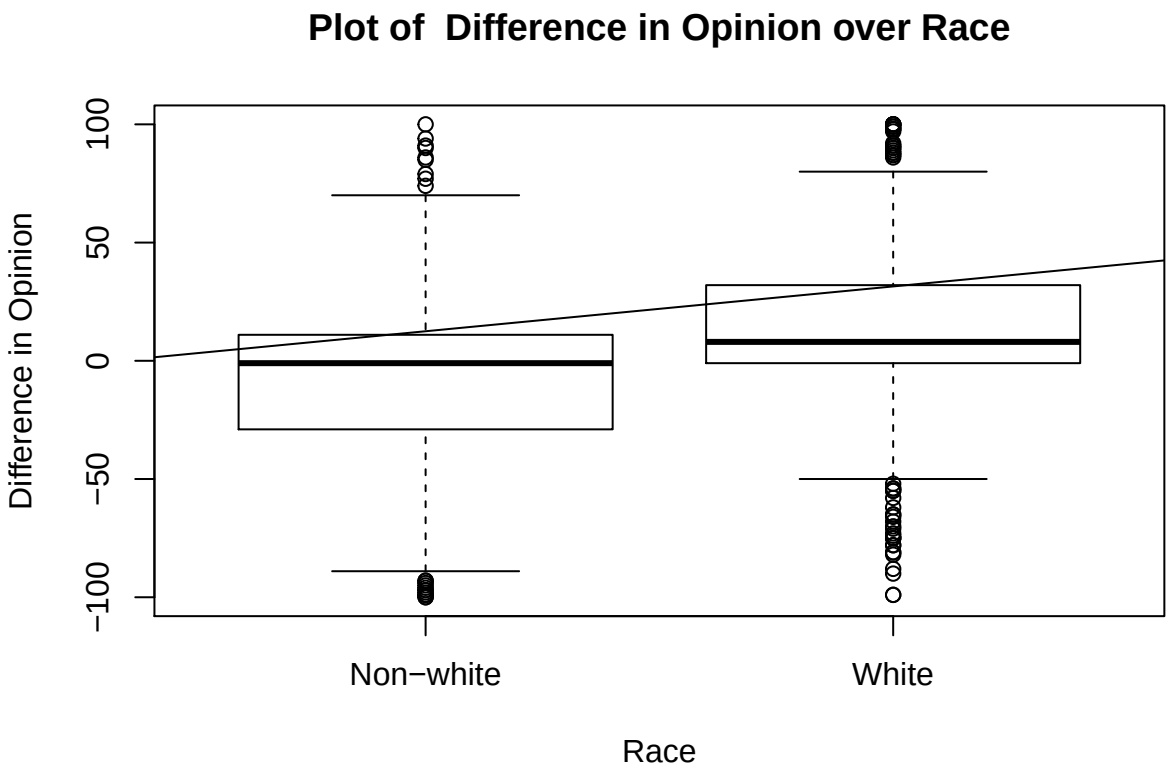
```
plotandtrend(data$pid3,data$diff,"Party Affiliation","Difference in Opinion")
```

```
## Warning in abline(lm(y ~ x)): only using the first two of 5 regression
## coefficients
```



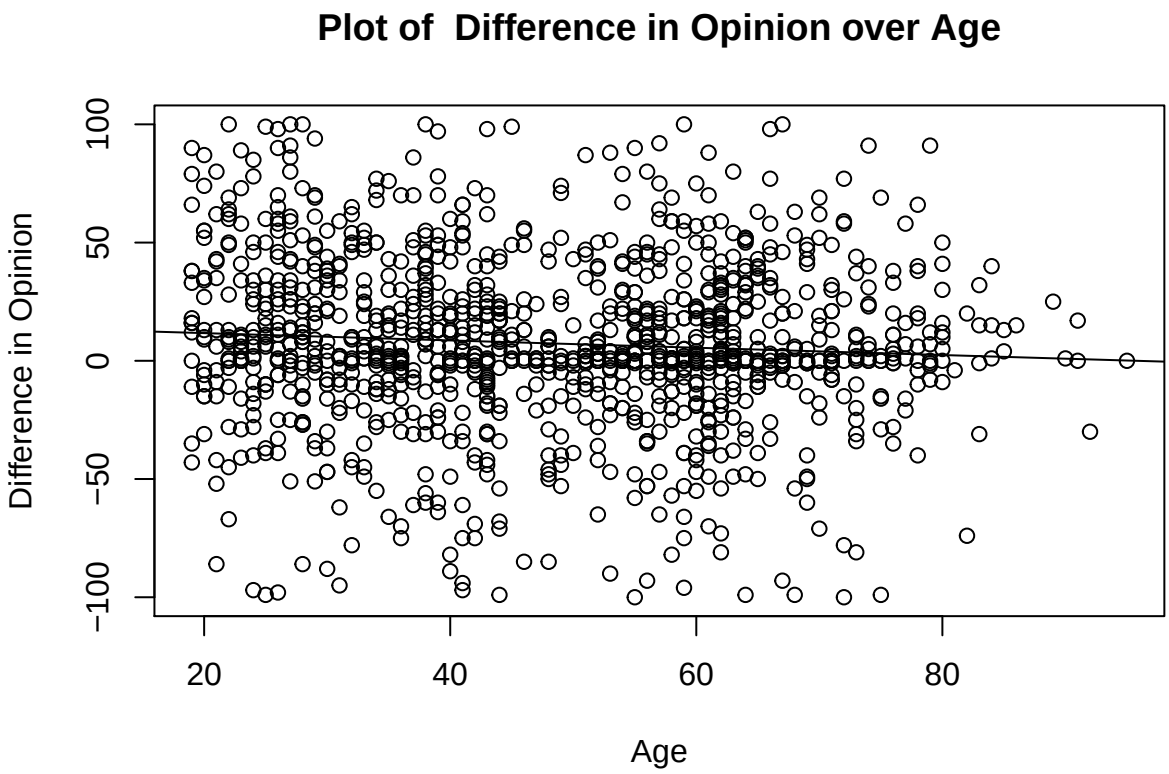
Comparing *diff* with *race_white*: This plot is interesting because there's a change in sign. Non-white people tend to have a higher opinion of Clinton than Sanders, while white people tend to have a higher opinion of Sanders than Clinton. However, the observations about non-white people are not as credible because there are much fewer of them in the sample.

```
plotandtrend(data$race_white,data$diff,"Race","Difference in Opinion")
```



Comparing *diff* with *age*: There appears to be a negative relationship between age and difference in opinion of Sanders and Clinton. Younger people tend to have a higher opinion of Sanders than they do of Clinton, while the difference is smaller as age increases.

```
plotandtrend(data$age,data$diff,"Age","Difference in Opinion")
```



Comparing *pid3* with *ideo5*: Party affiliation and Political Ideology are well correlated in some areas (if someone identifies as a Democraft, they are more likely to be liberal or very liberal), but it's not perfect multicolleniarity as there is some variation of ideology within each political affiliation.

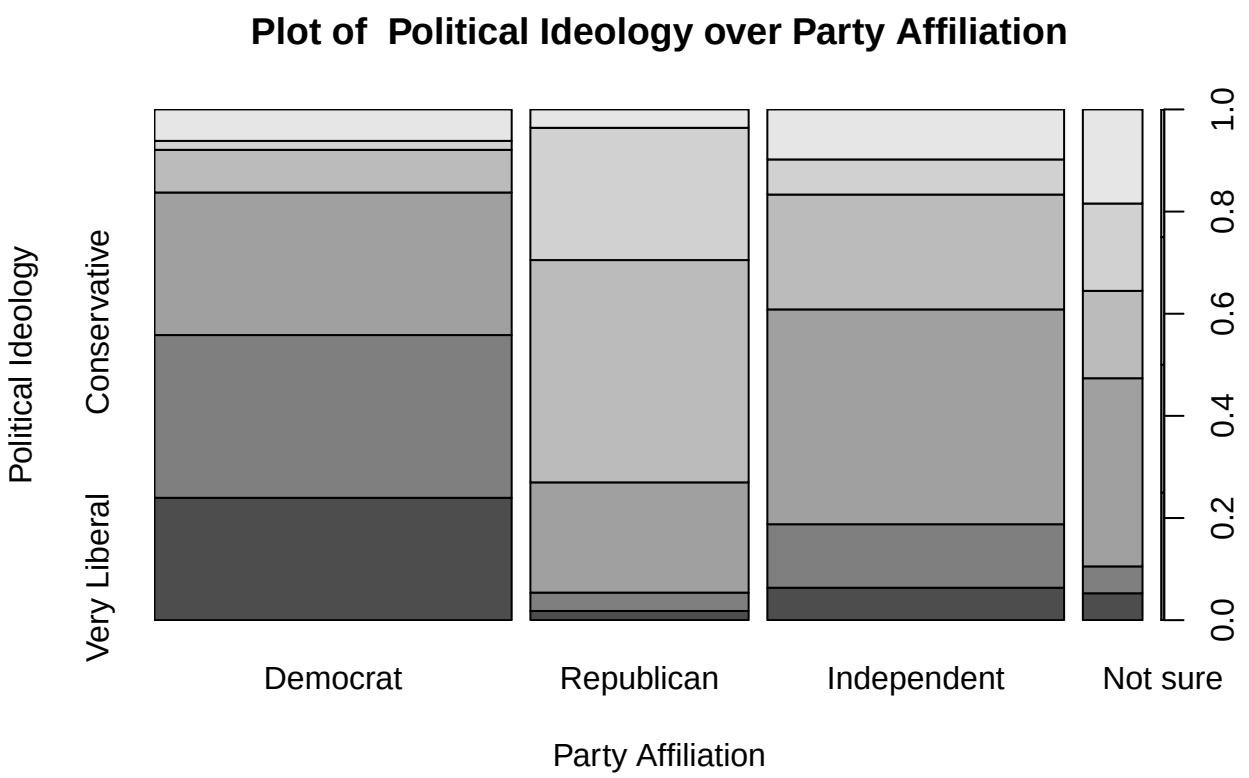
Both these variables should be transformed into multiple indicator variables. These transformations are dsecribed in problem 1. Note that the incidence of multicolleniarity dramatically increases with this transformation if we include all the new indicator variables in the regression. It won't technically be perfect multicolleniarity, but it will be near perfect. To mitigate this only some, but not all, of the new indicators should be included in the new regression.

```
plotandtrend(data$pid3,data$ideo5,"Party Affiliation","Political Ideology")
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

```
## Warning in abline(lm(y ~ x)): only using the first two of 5 regression
## coefficients
```

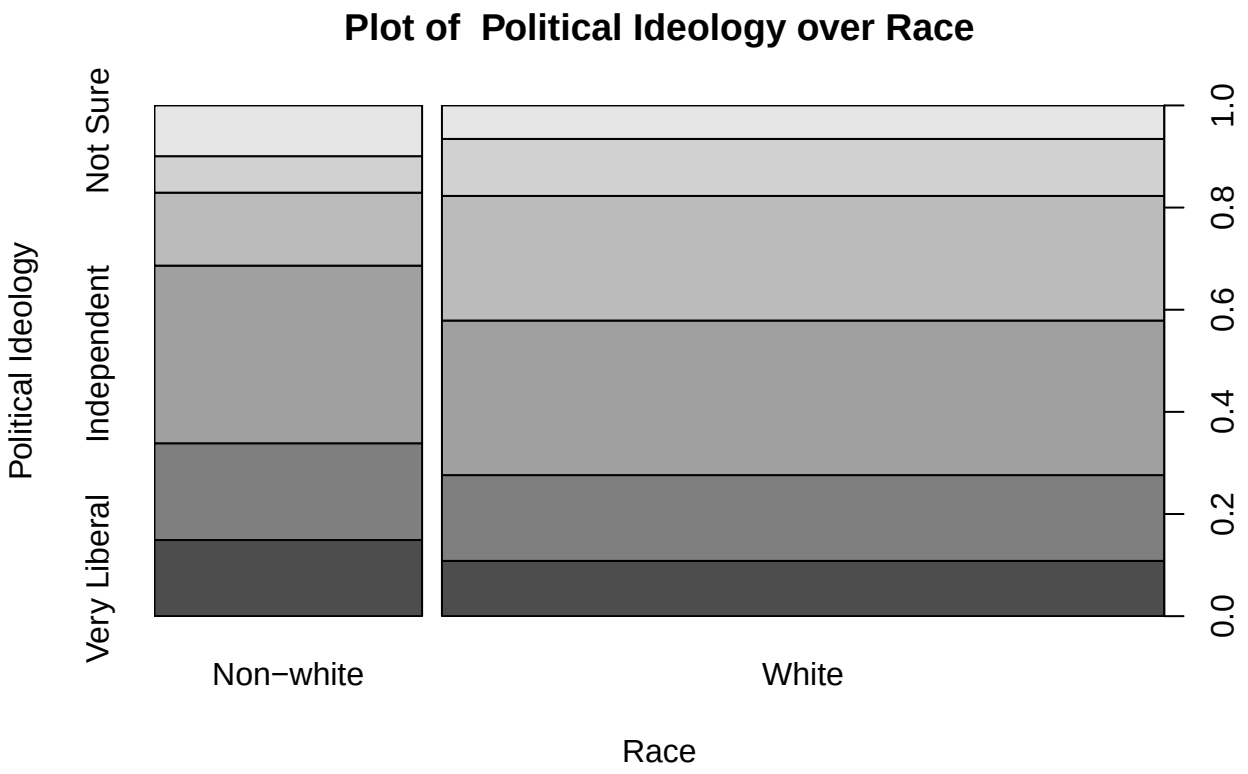


Comparing *race_white* and *ideo5*: The graph shows that the non-whites tend to be slightly more liberal than whites.

```
plotandtrend(data$race_white,data$ideo5,"Race","Political Ideology")
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

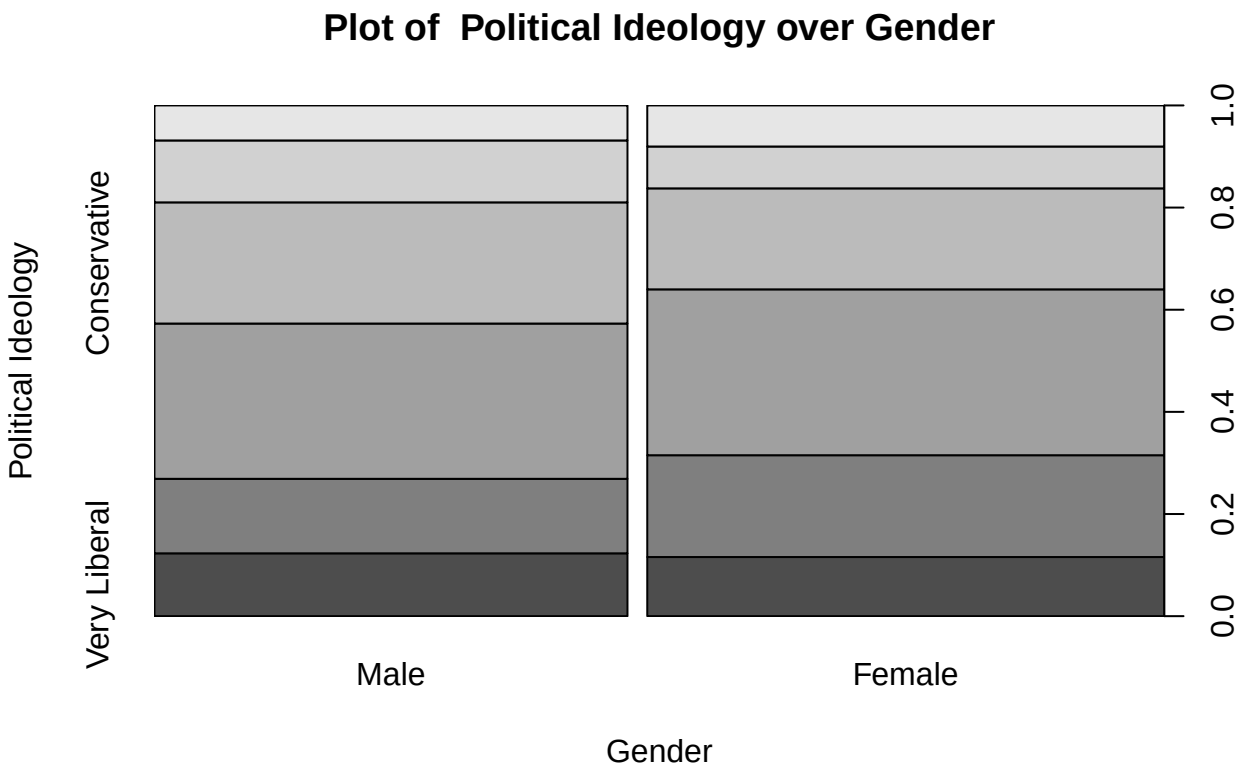


Comparing *gender* and *ideo5*: The graph shows that the females are more likely than males to identify as Democrafts

```
plotandtrend(data$gender,data$ideo5,"Gender","Political Ideology")
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

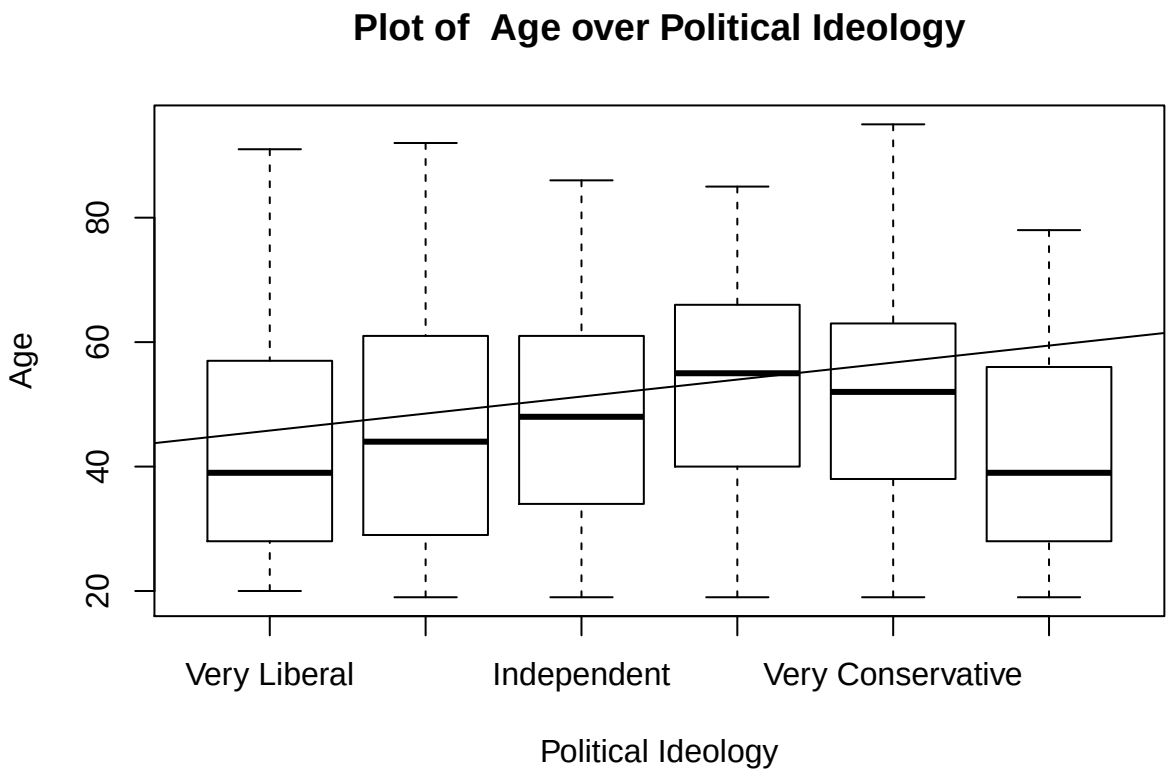
```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```



Comparing *age* and *ideo5*: The graph shows that the younger people tend to be slightly more liberal than older people

```
plotandtrend(data$ideo5,data$age,"Political Ideology","Age")
```

```
## Warning in abline(lm(y ~ x)): only using the first two of 6 regression
## coefficients
```



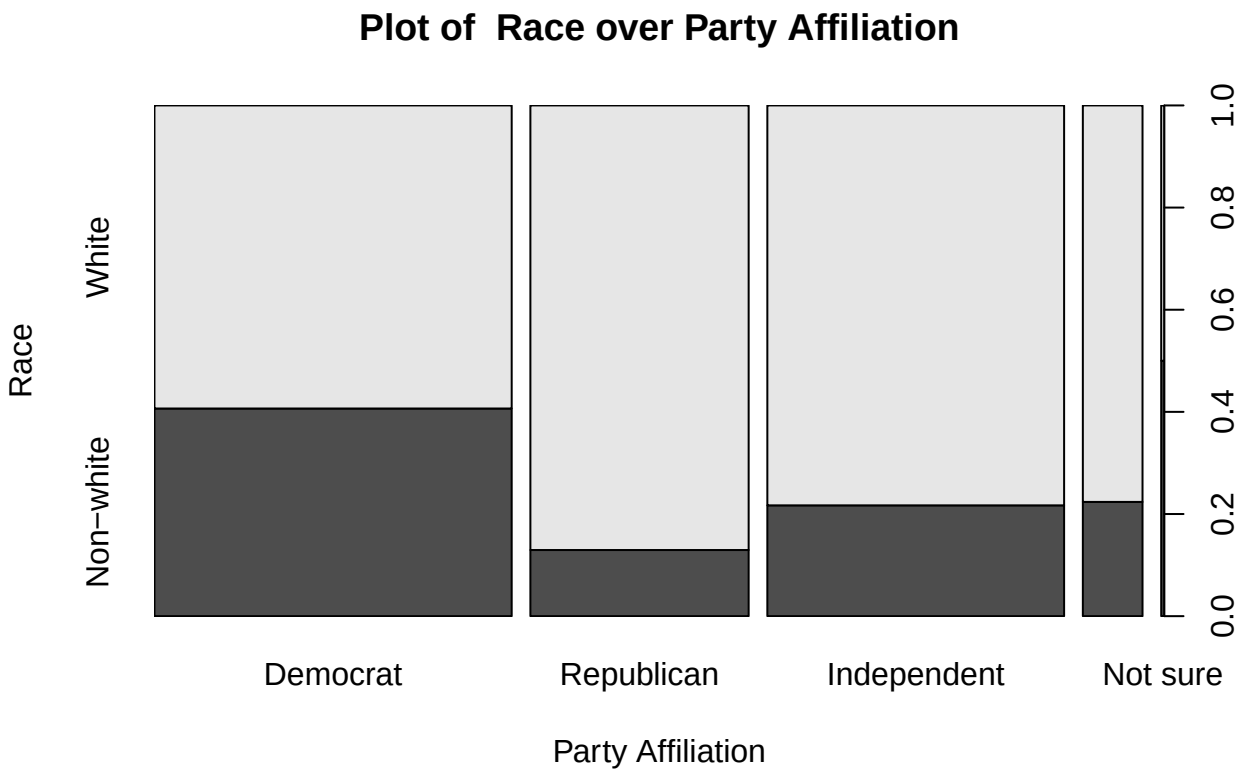
Comparing *race_white* and *pid3*: This graph shows that non-whites are more likely to identify as Democrats than as other parties

```
plotandtrend(data$pid3,data$race_white,"Party Affiliation","Race")
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

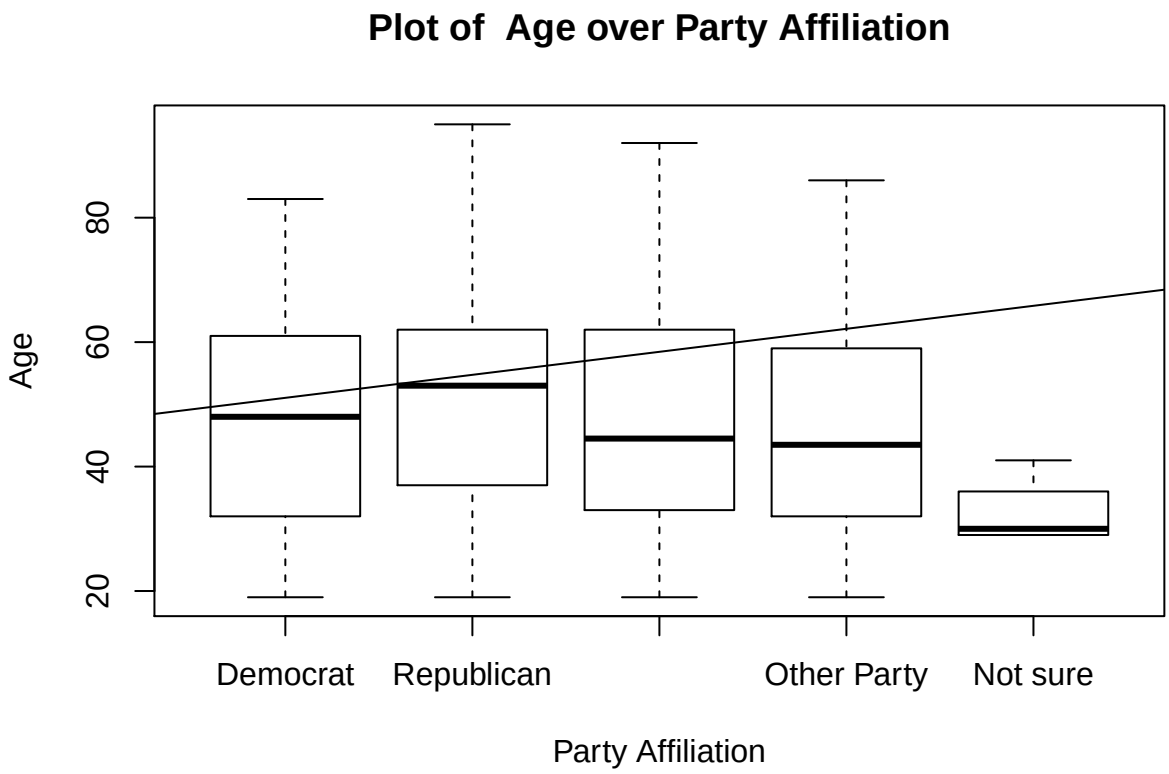
```
## Warning in abline(lm(y ~ x)): only using the first two of 5 regression
## coefficients
```



Comparing *age* and *pid3*: This graph shows that people who identify with the Republican party are generally older than people in other parties.

```
plotandtrend(data$pid3,data$age,"Party Affiliation","Age")
```

```
## Warning in abline(lm(y ~ x)): only using the first two of 5 regression
## coefficients
```



c. In some cases, multivariate analyses could be useful before the modeling stage. Do they apply in this case? If so, conduct some multivariate analyses.

Multivariate analyses are useful to find multicolleniarity between two or more different variables. We've already performed this type of analysis with each pairing of variables in the previous part.

While multivariate analyses with three or more variables is possible, its not neccesary here because the cardinality of the explanatory variables makes it unlikely for there to be any perfect multicollinearity. In the 5 explanatory variables, there are two variables with domain size 2, one variable with domain size 5, one variable with domain size 6, and one variable with domain > 70. Given the fact that none of these variables accept decimal values, it's very unlikely that one of these variables is a linear combination of the other variables.

6. Model Building Part 1: Treat ideo5 as a continuous variable. Consider the following when building your model.

```
#Convert ideo5 back to numeric
data$ideo5 = as.numeric(data$ideo5)
```

a. How does your EDA influence the inclusion of the variables in your model?

The EDA suggests that political ideology and gender probably will not be helpful in this model, as they do not produce trend lines with large slopes.

b. Start from a parsimonious model (with just the independent variable of interest) and gradually build it up. (Note that this is not the only way to build a regression model.) Does the inclusion of additional explanatory variables improve your model? Does the direction of the estimated coefficients of explanatory variables make sense? What, after all, does "improve" your model mean (in the context of answer the data science question at hand)?

In the context of this question, to "improve" means to significantly improve the fit as measured by an F test.

The base model with just the variable of interest. The model is not significant.

```
model1 = lm(diff~1,data)
summary(model1)
```

```
##
## Call:
## lm(formula = diff ~ 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.373  -16.373   -3.373   20.127   92.627
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.373      1.029   7.164 1.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.52 on 1190 degrees of freedom
```

Adding *ideo5* a doesn't significantly improve the fit of the model, so leave it out

```
model2 = lm(diff~ideo5,data)
anova(model2,model1)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ ideo5
## Model 2: diff ~ 1
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1189 1501173
## 2   1190 1501292 -1   -119.14 0.0944 0.7588
```

Adding *gender* also does not seem to significantly improve the model, so leave it out

```
model3 = lm(diff~ gender,data)
anova(model3,model1)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ gender
## Model 2: diff ~ 1
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1189 1500176
## 2   1190 1501292 -1     -1116 0.8845 0.3472
```

Adding *race_white* seems to signficantly improve the model

```
model4 = lm(diff~race_white,data)
anova(model4,model1)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white
## Model 2: diff ~ 1
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   1189 1416853
## 2   1190 1501292 -1    -84439 70.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding *age* further improves the model

```
model5 = lm(diff~race_white+age,data=data)
anova(model5,model4)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age
## Model 2: diff ~ race_white
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1188 1396757
## 2   1189 1416853 -1    -20097 17.093 3.809e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding *pid3* also improves the model

```
#pid3 should be numeric here
model6 = lm(diff~race_white+age+as.numeric(pid3),data=data)
anova(model6,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + as.numeric(pid3)
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1187 1348896
## 2   1188 1396757 -1    -47860 42.116 1.261e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examine the coefficients of the final model:

```
coef(model6)
```

```
##      (Intercept)  race_whiteWhite              age as.numeric(pid3)
##       -8.4506483       18.0172346       -0.2249017        6.5034200
```

The directions of the coefficients make sense and fit what we saw during the EDA.

c. Based on the best model thus far, do liberal voters have a higher level of support for Sanders over Clinton?

The best model does not include *ideo5*, but we can answer the question by looking at model2:

```
summary(model2)
```

```
##
## Call:
## lm(formula = diff ~ ideo5, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107.09  -16.32   -3.55   20.18   93.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6297     2.6291   2.522   0.0118 *
## ideo5         0.2301     0.7489   0.307   0.7588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.53 on 1189 degrees of freedom
## Multiple R-squared:  7.936e-05,  Adjusted R-squared:  -0.0007616
## F-statistic: 0.09436 on 1 and 1189 DF,  p-value: 0.7588
```

So both the interecept and the coefficient on *ideo5*. This means that liberals in this sample tended to have a higher level of support for Sanders over Clinton, but that difference in opinion grows as you get more conservative.

But since the coefficient on ideo5 is not significant, we don't have enough evidence to generalize this statement to the "liberal population."

d. To what extent does this relationship have any practical significance?

Outside of their face value, there's not a lot of practical significance here. Being born a particular race or in a particular year does not directly cause you to prefer Sanders over Hilliary or vice versa. There are some other factors here not captured in the regression. And due to the way *pid3* is encoded, it's hard to come up with any meaningful interpretation of the coefficient.

e. Note: Do not just use some sort of automatic selection methods, such as forward- or backward selection. If you use them, please provide your rationale.

Some forward or backward selection is neccesary to use the F test to measure model fit, as the F test is only appropriate if one model is nested within the other model. While I used forward selection to build my final model, I had a methodology to the ordering of my variables: I first tested the least promising variables (*ideo5* and *gender*) to verify that I could exclude them, then added in the most promising variables in order of potential practical signficance: *race_white* had the highest since its a clear binary variable. *age* is lower because it showed bimodality during EDA. And *pid3* is last because the numeric encoding of political parties seems arbitrary.

f. Remember to conduct regression diagnostics.

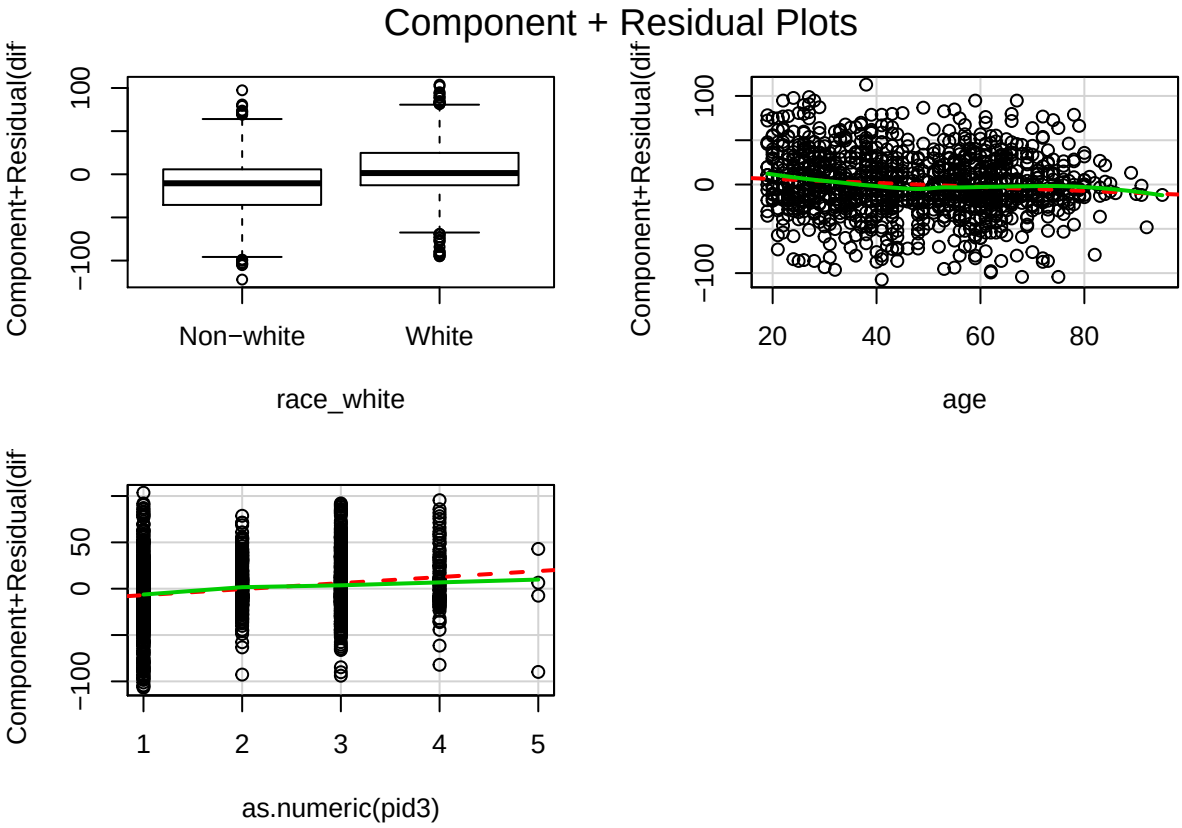
++ Remember to conduct formal statistical tests to test all of the underlying assumptions of the CLM.

We are already given that the data is random, and we know that that the there isn't multicollinearity from the previous problem. Also, since this is not time series data, we don't need to worry about autocorrelation.

It looks like this regression passes all the testable assumption. It doesn't look like the regression has zero conditional mean, but it is exogeneous though.

```
regressionDiagnostics = function (model){
f = summary(model)$fstatistic
p = pf(f[1],f[2],f[3],lower.tail = F)
result = data.frame(Test = "Linearity",Passed=p<.05,stringsAsFactors = F)

print("Plotting Error versus each X")
print(crPlots(model))
#While we can't test for zero conditional mean of erros, we can test for exogeneity
errorModel = lm(model$residuals~.-1,model$model[,-1])
f = summary(errorModel)$fstatistic
p = pf(f[1],f[2],f[3],lower.tail = F)
result = rbind(result,c("Exogeneous Errors",p>.05))
result = rbind(result,c("Homoskedastic Errors",ncvTest(model)$p<.05))

result = rbind(result,c("Normality of Errors",shapiro.test(model$residuals)$p<.05))

return(result)

}
regressionDiagnostics(model6)
```

```
## [1] "Plotting Error versus each X"
```




```
## [1] 0
```

```
##                       Test Passed
## value            Linearity   TRUE
## 2        Exogeneous Errors   TRUE
## 3     Homoskedastic Errors   TRUE
## 4      Normality of Errors   TRUE
```

h. When assumptions are not satisfied, describe its potential impacts on the estimates and statistical inference.

It depends on the type of assumption being violated. If the assumption violated involves linearity of model, random sampling, or exogeneity, then the model's accuracy will not be consistent and will be better or worse for certain values of the explantory variables. If on the other hand one of the BLUE assumptions aren't met, then the model may be consistently accurate, but it might not be the most accurate model possible.

7: Model Building Part 2: Treat ideo5 as a categorical variable

```
print("First, I'm going to create some additional indicator variables so that I can regress on them indi
```

```
## [1] "First, I'm going to create some additional indicator variables so that I can regress on them ind
```

```
data$ideo5 = as.numeric(data$ideo5)
data$VeryLib = data$ideo5 == 1
data$Lib = data$ideo5 == 2
data$Ind = data$ideo5 == 3
data$NotSure = data$ideo5 == 6
data$Con = data$ideo5 ==4
data$VeryCon = data$ideo5 == 5

#Now convert ideo5 to categorical
data$ideo5 = as.factor(data$ideo5)
levels(data$ideo5) = c("Very Liberal","Liberal","Independent","Conservative","Very Conservative","Not S
```

a. All the suggestions in Step 6 apply to this step.

The base model with just the variable of interest. The model is not significant.

```
model1 = lm(diff~1,data)
summary(model1)
```

```
##
## Call:
## lm(formula = diff ~ 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.373  -16.373   -3.373   20.127   92.627
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.373      1.029   7.164 1.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.52 on 1190 degrees of freedom
```

Adding *ideo5* a doesn't significantly improve the fit of the model, so leave it out

```
model2 = lm(diff~ideo5,data)
anova(model2,model1)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ ideo5
## Model 2: diff ~ 1
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1185 1496731
## 2   1190 1501292 -5   -4561.6 0.7223 0.6067
```

Adding *gender* also does not seem to significantly improve the model, so leave it out

```
model3 = lm(diff~ gender,data)
anova(model3,model1)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ gender
## Model 2: diff ~ 1
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1189 1500176
## 2   1190 1501292 -1     -1116 0.8845 0.3472
```

Adding *race_white* seems to signficantly improve the model

```
model4 = lm(diff~race_white,data)
anova(model4,model1)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white
## Model 2: diff ~ 1
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   1189 1416853
## 2   1190 1501292 -1    -84439 70.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding *age* further improves the model

```
model5 = lm(diff~race_white+age,data=data)
anova(model5,model4)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age
## Model 2: diff ~ race_white
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1188 1396757
## 2   1189 1416853 -1    -20097 17.093 3.809e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the last problem, we added *pid3* here to improve the model. Instead, I'm going to try adding in *ideo5* again to see if it improves the model now with some more data in it:

```
model6 = lm(diff~race_white+age+ideo5,data=data)
anova(model6,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + ideo5
## Model 2: diff ~ race_white + age
```

```
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1183 1392232
## 2   1188 1396757 -5   -4524.6 0.7689 0.5722
```

Now I'm going to try just adding in some, but not all, of the new indicator variables I created:

```
model7 = lm(diff~race_white+age+Lib,data=data)
anova(model7,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Lib
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1187 1393839
## 2   1188 1396757 -1   -2917.7 2.4848 0.1152
```

```
model8 = lm(diff~race_white+age+VeryLib,data=data)
anova(model8,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + VeryLib
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1187 1395988
## 2   1188 1396757 -1   -768.81 0.6537  0.419
```

```
model9 = lm(diff~race_white+age+Lib+Ind,data=data)
anova(model9,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Lib + Ind
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1186 1393757
## 2   1188 1396757 -2   -2999.6 1.2762 0.2795
```

```
model10 = lm(diff~race_white+age+Ind+NotSure,data=data)
anova(model10,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Ind + NotSure
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1186 1396647
## 2   1188 1396757 -2   -109.56 0.0465 0.9545
```

```r
model11 = lm(diff~race_white+age+VeryLib+VeryCon,data=data)
anova(model11,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + VeryLib + VeryCon
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1186 1395549
## 2   1188 1396757 -2   -1207.9 0.5133 0.5987
```

```r
model12 = lm(diff~race_white+age+Con,data=data)
anova(model12,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Con
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1187 1395618
## 2   1188 1396757 -1   -1138.8 0.9686 0.3252
```

```r
model13 = lm(diff~race_white+age+VeryLib+Con,data=data)
anova(model13,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + VeryLib + Con
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1186 1394446
## 2   1188 1396757 -2   -2310.5 0.9826 0.3747
```

```r
model14 = lm(diff~race_white+age+VeryCon,data=data)
anova(model14,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + VeryCon
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1187 1396178
## 2   1188 1396757 -1   -578.51 0.4918 0.4832
```

```r
model15 = lm(diff~race_white+age+Lib+Con,data=data)
anova(model15,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Lib + Con
```

```
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1   1186 1393368
## 2   1188 1396757 -2   -3388.2 1.442 0.2369
```

```
model16 = lm(diff~race_white+age+Lib+Con+NotSure,data=data)
anova(model16,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Lib + Con + NotSure
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1185 1393368
## 2   1188 1396757 -3     -3389 0.9607 0.4105
```

```
model17 = lm(diff~race_white+age+VeryLib+Con+Ind,data=data)
anova(model17,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + VeryLib + Con + Ind
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1185 1393160
## 2   1188 1396757 -3   -3596.9 1.0198 0.3829
```

```
model18 = lm(diff~race_white+age+Ind,data=data)
anova(model18,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + Ind
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1187 1396688
## 2   1188 1396757 -1   -68.457 0.0582 0.8094
```

None of the new models are significant. So it seems that the indicator variables are also not useful. The best model is achieved by adding *pid3*

```
#pid3 should still be numeric here
model19 = lm(diff~race_white+age+as.numeric(pid3),data=data)
anova(model19,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + as.numeric(pid3)
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1187 1348896
## 2   1188 1396757 -1    -47860 42.116 1.261e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since it may be significant let's try adding *ideo5* here again:

```
#pid3 should still be numeric here
model20 = lm(diff~race_white+age+as.numeric(pid3)+ideo5,data=data)
anova(model20,model19)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + as.numeric(pid3) + ideo5
## Model 2: diff ~ race_white + age + as.numeric(pid3)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   1182 1340438
## 2   1187 1348896 -5   -8458.2 1.4917 0.1896
```

b. Based on this model, do liberal voters have a higher level of support for Sanders over Clinton?

ideo5 is still not included in the final model here. But let's take a look now at the second model.

The intercept (represents Very Liberal) is 9.37, and the coefficient on Liberal is -5.35. So people in the sample who identified as Very Liberal had on an average higher opinion of Sanders by about 9 points, while people who identified as Liberal had on average a higher opinion of Sanders by about 4 points.

```
summary(model2)
```

```
##
## Call:
## lm(formula = diff ~ ideo5, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.37  -16.12   -3.30   20.10   95.98
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               9.3732     2.9824   3.143  0.00171 **
## ideo5Liberal             -5.3539     3.8725  -1.383  0.16707
## ideo5Independent         -2.2479     3.5019  -0.642  0.52105
## ideo5Conservative         0.4136     3.7135   0.111  0.91134
## ideo5Very Conservative   -3.0732     4.4069  -0.697  0.48570
## ideo5Not Sure            -1.9013     4.8048  -0.396  0.69239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.54 on 1185 degrees of freedom
## Multiple R-squared:  0.003038,   Adjusted R-squared:  -0.001168
## F-statistic: 0.7223 on 5 and 1185 DF,  p-value: 0.6067
```

But again this model is not significant, so we can't generalize the observations to all people who identify as Liberal or Very Liberal.

Finally, let's also use the new indicator variables to perform some basic t tests:

```r
t.test(diff~NotSure,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff by NotSure
## t = -0.025386, df = 100.29, p-value = 0.9798
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.478543  8.264306
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            7.364791            7.471910
```

```r
t.test(diff~VeryLib,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff by VeryLib
## t = -0.59925, df = 165.79, p-value = 0.5498
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.754375  5.211900
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            7.102002            9.373239
```

```r
t.test(diff~Lib,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff by Lib
## t = 1.3626, df = 274.84, p-value = 0.1741
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.805454  9.923311
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            8.078252            4.019324
```

```r
t.test(diff~Con,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff by Con
## t = -1.4812, df = 561.51, p-value = 0.1391
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -7.167956  1.004816
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            6.705252            9.786822
```

```
t.test(diff~VeryCon,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff by VeryCon
## t = 0.48164, df = 186.26, p-value = 0.6306
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.693477  6.079471
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            7.492997            6.300000
```

```
t.test(diff~NotSure,data)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff by NotSure
## t = -0.025386, df = 100.29, p-value = 0.9798
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.478543  8.264306
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            7.364791            7.471910
```

No significant difference

c. Compare and contrast your final model in step 6 and your final model in step 7. Do your answers differ? Is one model more appropriate than the other? Is one model easier to interpret than the other?

Both models are the same. *ideo5* did not help explain a significant amount of the variance in differenes of opinion.

8: The Final Model

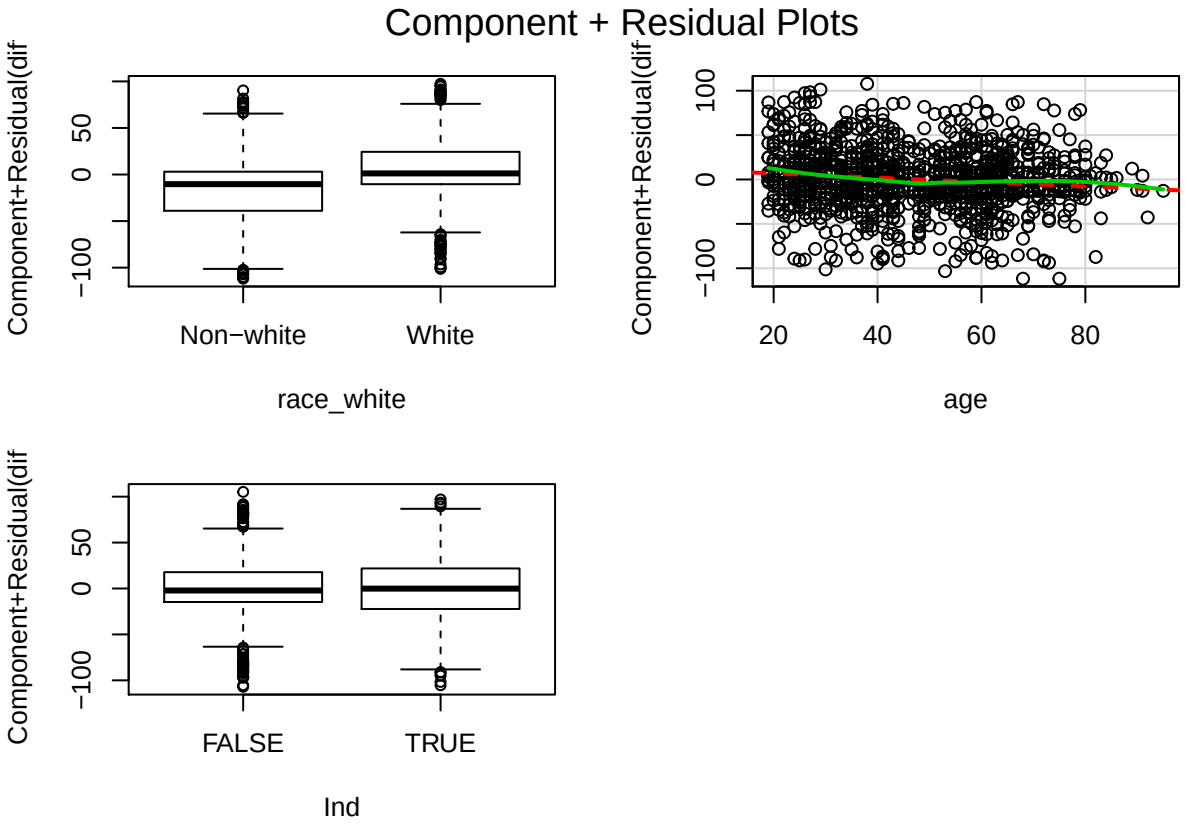
a. Choose your final model.

Both models ended up the same

b. Does this model satisfy all of the CLM assumptions?

```
regressionDiagnostics(model18)
```

```
## [1] "Plotting Error versus each X"
```




```
## [1] 0
```

```
##                       Test Passed
## value            Linearity   TRUE
## 2        Exogeneous Errors   TRUE
## 3     Homoskedastic Errors   TRUE
## 4      Normality of Errors   TRUE
```

c. What changes would you make to your existing models based on these results?

I'd like to try regressing on *pid3* as a categorical instead of as a numeric.

d. What changes are possible to make with the existing data?

You can create interaction terms, bin the age variable, and convert *pid3* to a categorical.

e. What changes would require more data?

It might be nice to get opinions on Donald Trump as well, as it could be an interesting explanatory variable.

f. What changes cannot be plausibly made with additional data only?

g. Most importantly, do not forget to answer the original question posted by the company that hires you. Based on your model, how would you formulate the strategy to ask voters for donations for causes championed by Bernie Sanders.

So far, *ideo5* is not a significant factor in predicting difference in opinion of Sanders and Clinton. So we cannot reject the null hypothesis that Sanders and Clinton supporters are equally liberal.

While I observed other variables that were significant, I can't recommend incorporating them into a strategy because they were not part of the original hypothesis.

However, I would recommend gathering a new set of data to formally test if the *age*, *race_white*, or *pid3* can explain difference in opinion of Sanders and Clinton.

Part 2: Including partisanship

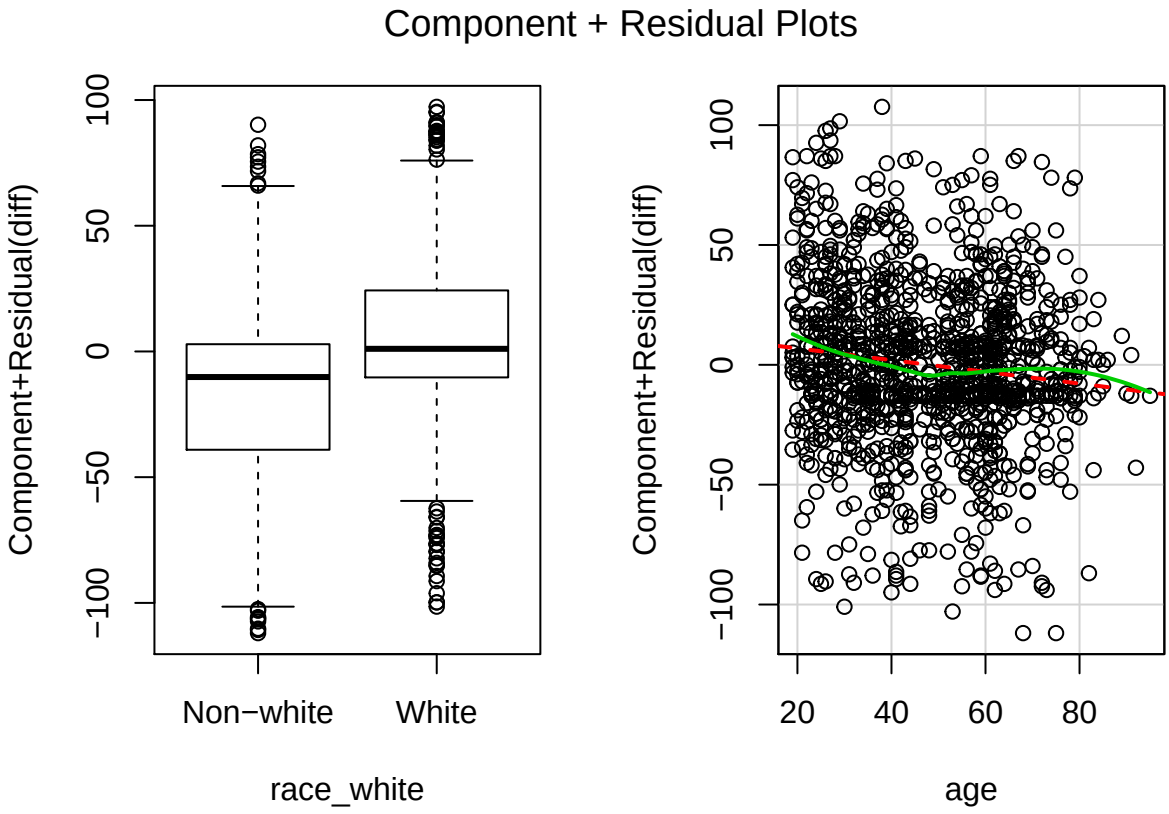
The above analysis does not take into account the partisanship of the survey respondents. The conventional wisdom in American Politics is that liberal voters are members of the Democratic Party while conservative voters are members of the Republican Party. The exclusion of partisanship could lead to biased estimates.

1. By excluding this variable, which of the 5 CLM assumptions did you violate? Explain why.

It doesn't appear that we violate any assumptions by excluding *pid3*. We don't meet zero conditional mean, but including *pid3* doesn't fix this problem and the model still hs exogenous errors.

```
regressionDiagnostics(model5)
```

```
## [1] "Plotting Error versus each X"
```




```
## [1] 0
```

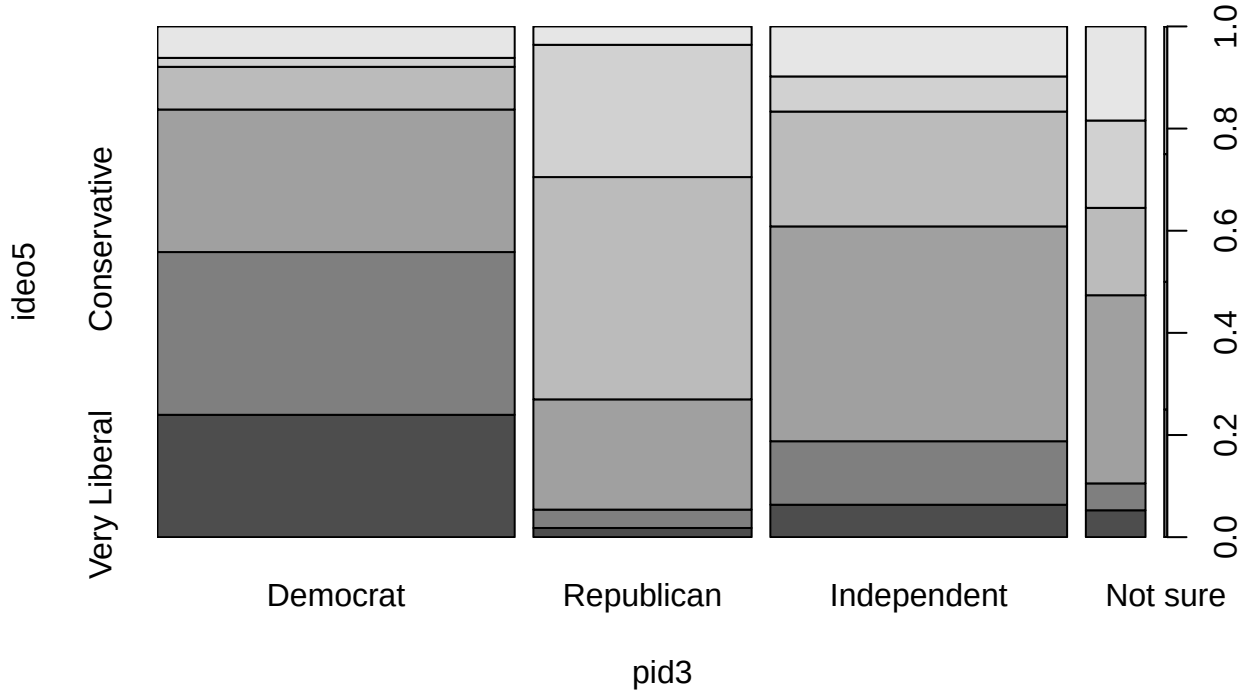
```
##                       Test Passed
## value            Linearity   TRUE
## 2        Exogeneous Errors   TRUE
## 3     Homoskedastic Errors   TRUE
## 4      Normality of Errors   TRUE
```

2. Based on your residuals analysis in Part 1, were you able to diagnose this problem? Why or why not?

3. How closely are partisanship and ideology related? Should you be worried about multi-collinearity?

They are pretty closely related. Enough so that you could regress one on the other.

```
plot(ideo5~pid3,data)
```




```
summary(lm(VeryLib~pid3,data))
```

```
##
## Call:
## lm(formula = VeryLib ~ pid3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23956 -0.23956 -0.06349 -0.01799  0.98201
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.23956    0.01454  16.477  < 2e-16 ***
## pid3Republican  -0.22157    0.02361  -9.386  < 2e-16 ***
## pid3Independent -0.17607    0.02158  -8.158 8.62e-16 ***
## pid3Other Party -0.18693    0.03843  -4.864 1.30e-06 ***
## pid3Not sure    -0.23956    0.15574  -1.538    0.124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3101 on 1186 degrees of freedom
## Multiple R-squared:  0.088,  Adjusted R-squared:  0.08492
## F-statistic: 28.61 on 4 and 1186 DF,  p-value: < 2.2e-16
```

4. Use your final model from Part 1 and include partisanship in your model.

```
summary(model19)
```

```
##
## Call:
## lm(formula = diff ~ race_white + age + as.numeric(pid3), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.845  -18.340   -1.951   19.562  110.493
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -8.45065    3.69721  -2.286 0.022448 *
## race_whiteWhite  18.01723    2.26497   7.955 4.16e-15 ***
## age              -0.22490    0.05833  -3.856 0.000122 ***
## as.numeric(pid3)  6.50342    1.00212   6.490 1.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.71 on 1187 degrees of freedom
## Multiple R-squared:  0.1015, Adjusted R-squared:  0.09924
## F-statistic:  44.7 on 3 and 1187 DF,  p-value: < 2.2e-16
```

5. Does the addition of this variable improve the model? (If so, in what sense?) How do your answers change?

Yes. It improves the fit:

```
anova(model19,model5)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + as.numeric(pid3)
## Model 2: diff ~ race_white + age
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1187 1348896
## 2   1188 1396757 -1    -47860 42.116 1.261e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Suppose you think that the relationship between ideology and the DV has a different slope for each party. How would you account for this in your model?

I would create an interaction term between *ideo5* and *pid3*. I'd need to make *pid3* categorical, but it's debatable whether *ideo5* should be numeric or categorical. I'm going to go with categorical because *ideo5* is sentiment data.

7. Make this adjustment and comment on the relationship between ideology, partisanship, and the DV.

```
model21 = lm(diff~race_white+age+ideo5*pid3+as.numeric(pid3),data)
anova(model21,model19)
```

```
## Analysis of Variance Table
##
## Model 1: diff ~ race_white + age + ideo5 * pid3 + as.numeric(pid3)
## Model 2: diff ~ race_white + age + as.numeric(pid3)
##   Res.Df     RSS  Df Sum of Sq      F   Pr(>F)
## 1   1161 1290115
## 2   1187 1348896 -26    -58781 2.0346 0.001672 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model is a significantly better fit than the best model from part 1.

This implies that the relationship between ideology and the DV has a different slope for each party.

8. Do you think that taking account of respondents' party affiliation is necessary for your purposes? As a political consultant, how would you use this information when deciding to whom solicitations should be made?

The fact that *pid3* improves the fit suggests that taking account of respondents' party affiliation is indeed necessary. Furthermore, the fact that the interaction term between *pid3* and *ideo5* further imrpoves the fit suggests that solicitations the effect of ideologoy differs for each party and that soliciations should be sent based on both party affiliation and ideology: