DATA621 Final Project

Misha Kollontai

12/14/2020

Abstract

The past two presidential elections in the USA have gone hand-in-hand with critiques of polling techniques. Due to various factors pre-election polling results were at times drastically different from the actual results of individual races. This paper aims to take a simpler look at predicting the 2020 election using a combination of census and socio-economic data broken down by US County. By taking objective data from 2016 and studying correlations with the election results, we will attempt to predict the 2020 election. The variables used will pull from US census data as well as information from the US Department of Agriculture on education, poverty and unemployment. The idea is to use a combination of census data and detailed vote counts from the past two elections to see just how well demographics alone could have been used to predict the outcome of the 2020 election. Using 2016 census/socio-economic data as the training dataset with the 2016 election results as the target variable, we will attempt to create a model that predicts the results of the 2020 election broken down by US county. The poverty and child mortality data did not show strong statistical significance to predicting the percentage of votes for Trump, though some of the education data proved to be a fairly strong predictor. Demographic breakdowns when converted to percentages also proved to be dairly strong predictors of the vote breakdown. While the models resulted in decent R^2 values, the margins between the prediction and the actual vote percentages in 2020 were simply too high. It was, however, instructive to identify the variables that prove to be the strongest predictors. The percentage of the population that is White/Black/Hispanic, the gender divide, and the percentage of a population that has a Graduate Degree seem to be the strongest predictors of votes for Trump. Of these, the percentage of the population that identifies as Hispanic as well as the percentage iwth a Graduate Degree having negative coefficients, suggesting Trump is unpopular with the hispanic population and with Post-Grads. Finally, the percentage Male variable had a strong positive relationship with the predictor, suggesting that percent Female would have a negative relationship - agreeing with Trump's perceived unpopularity with women.

Key Words

Election, Demographics, Votes, USA

Introduction

The 2020 Presidential election in the United States has been a turbulent one. One recurring theme seems to be how the polls predicting the outcome of the election were "off" in so many places The Polls Weren't Great. But That's Pretty Normal. Polling leading up to elections has garnered more attention than ever before - due at least partially to the ease with which this information can be shared in social media. Polling has inherent errors associated with it - whether that be with regard to the coverage of respondents, the honesty of the answers or other factors.

One way to simplify such an analysis and remove some of these sources of error would be to create predictions based on objective measurable data as opposed to questionnaires. This will doubtless lead to a less precise

picture, but it would be valuable to see just how close to the truth we can get by looking at certain socio-economic factors.

How well could we have predicted the 2020 election using only objective socio-economic data? If we were to simply look at some of this type of data of a state, could we have come close to predicting the 2020 election results?

The polling in 2016 was criticized especially heavily due to the high stakes of the election. Some studies looked into the reasons behind results not agreeing with polls. Some of the

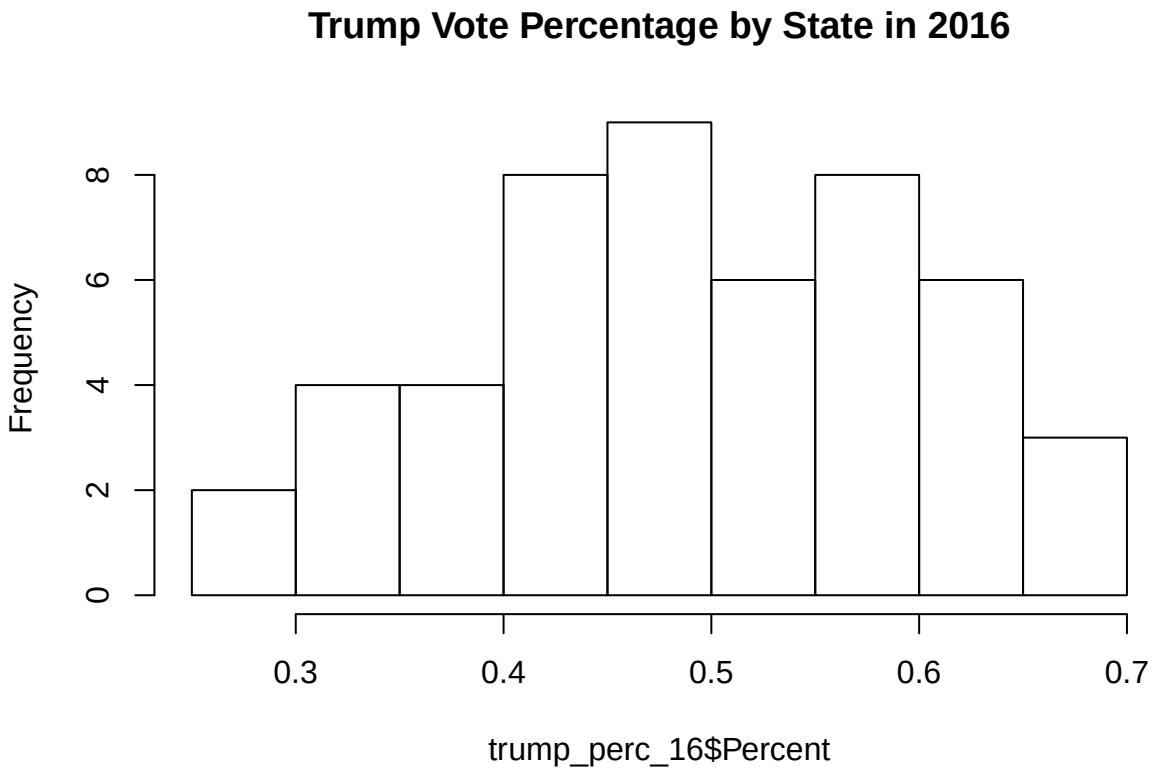
Methodology

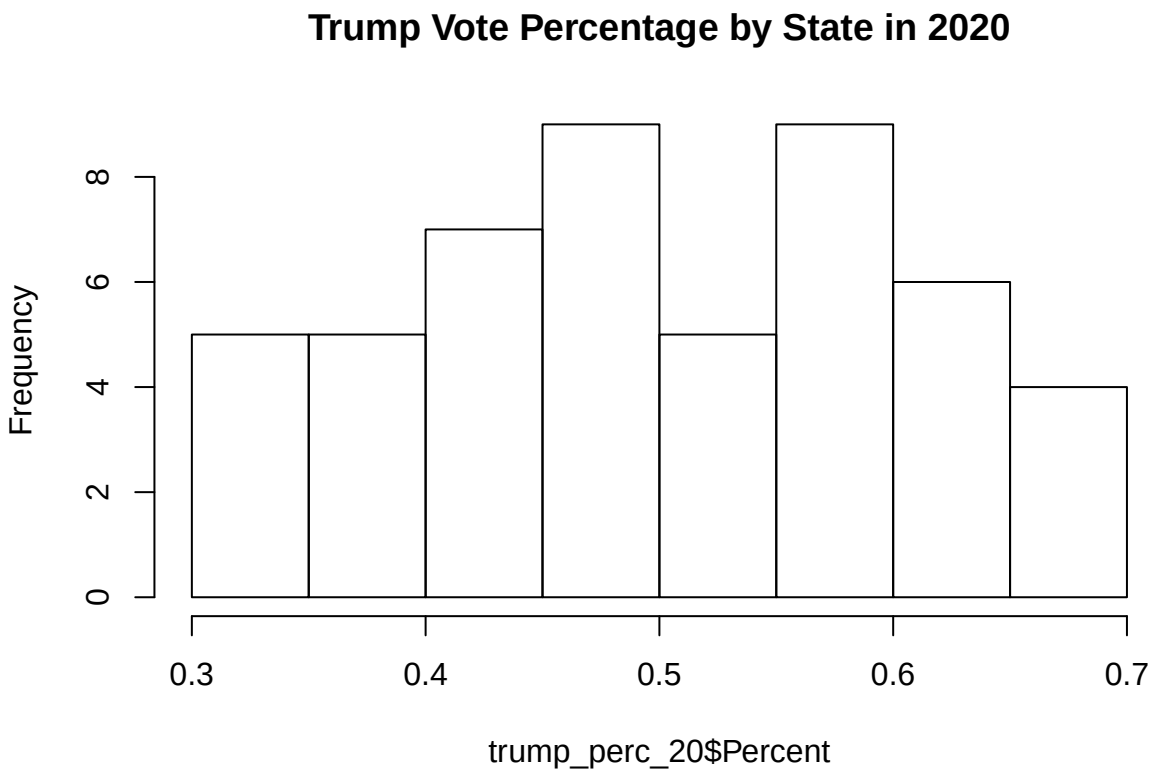
In order to perform this analysis we needed to combine and prepare data from several sources into a cohesive dataset. the first step was to determine our target variable. 2016 Vote Results were combined with 2020 Presidential Vote data broken down by County from Kaggle found on Kaggle.com. The variable of particular interest here was the percent of the vote won by the Republican party in each county both in 2016 and 2020.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```





The predictor variables were also collected from various sources and combined into a single dataframe. Information from the Kids Count Data Center. From here we were able to collect data on education attainment, poverty levels and infant mortality levels on a US state level. County level data was the original goal of this effort, but since that granularity is not available for recent years as of yet, a broader scope was adopted - the state level. Where available calculations were performed to show trends in certain variables. As an example, unemployment data was available for years leading up to 2016, so a calculation was performed to see the change over the last 4 years both for 2016 and 2019 (2020 data is not yet available, so 2019 was used as a stand-in). This was then used as an additional predictive variable. This data was supplemented with Census Data on each county. This data was used for overall population estimates as well as gender, age and race breakdowns. In order to prepare the dataset for model creation, each of the dataframes was organized and reshuffled into the same format. Different time ranges were provided based on the dataset, so inly the relevant years were pulled out. Since it is too early in 2020, no data was available for the current year. Though not ideal, especially considering the degree to which Coronavirus will potentially affect some of these numbers more dramatically than in other years. Numbers from a year or two before then election are obviously not going to be as accurate in terms of predictors, but they will work as stand-ins for this case study.

In effect the data consists of two groupings: population information covering breakdowns of age/gender/race and more general information on education, poverty and child mortality statistics. The two groups are used both separately and in conjunction with one another to provide as much valuable input for the model from the data as possible. The census data provided was all in terms of total counts of subsets of the population. This was converted to percentages of population to account for the relative sizes of the states.

Experimentation and results

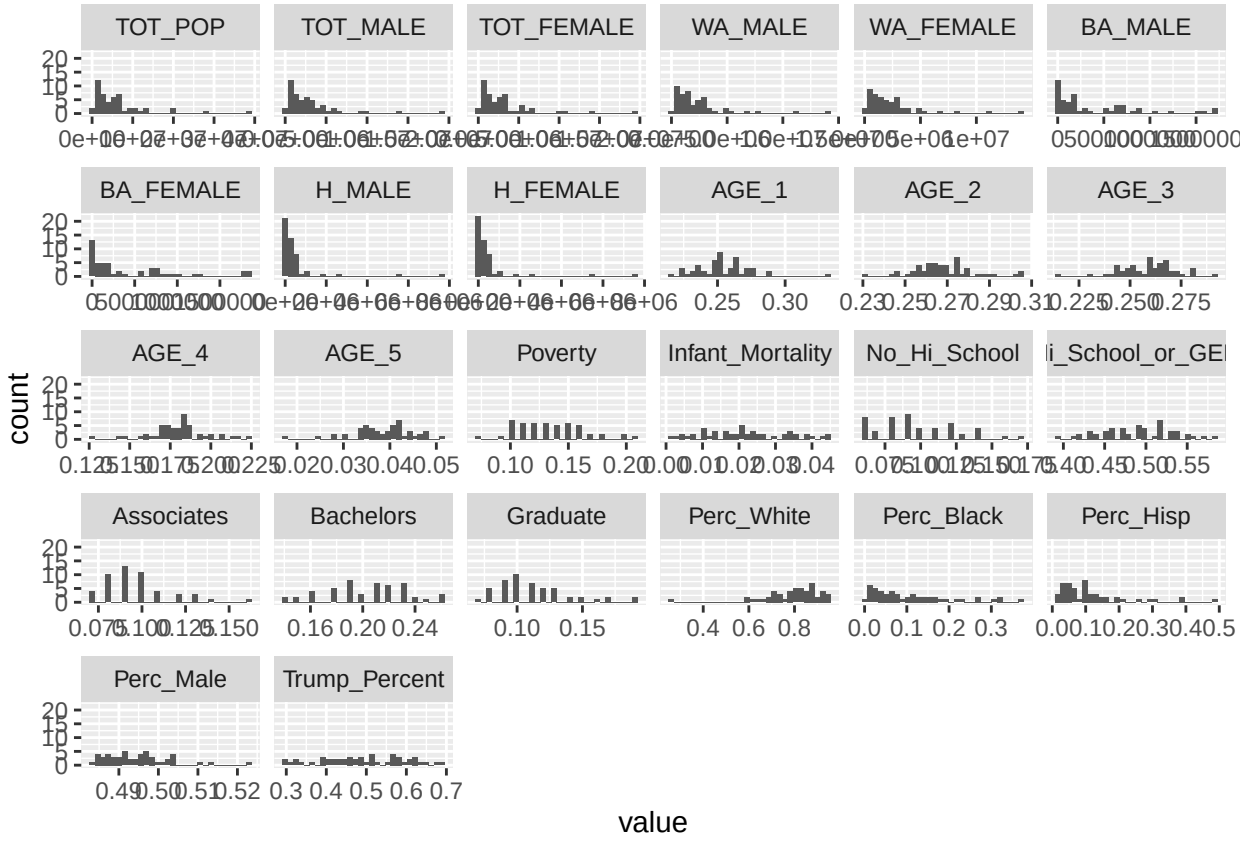
The distinction between the training and test datasets was ready-made. We used data from 2016 as our training dataset, with the election results from that year used to train the model. For the test dataset, we were forced to use data from 2019 (and 2018 for one variable). Predictions generated using the models were then compared against the results of the 2020 election.

In order to get a better idea of the distribution within the range of predictive variables, we looked at the histograms for each.

In order to supplement this information, we will pull the Census Data on each county, again narrowing our scope to the data after 2016. This data will provide a more detailed demographic breakdown in terms of gender as well as race.

```
## No id variables; using all as measure variables
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



It's interesting to note that few of the variables seem normally distributed. Nearly all of the population graphics are right-skewed with a majority of the data falling closer to the left end of the distribution. The population data used contained breakdowns in terms of Male and Female representation in:

- Total
- White Only
- Black Only
- Hispanic

Each of these was also available for a multitude of age ranges, which were combined slightly to reduce the number of bins. By combining all of the bins and using only the total data for the gender/race variables, we were able to create new variables tracking how much of the population of a state fell within a certaiin age range. The final binning used in our analysis was as follows:

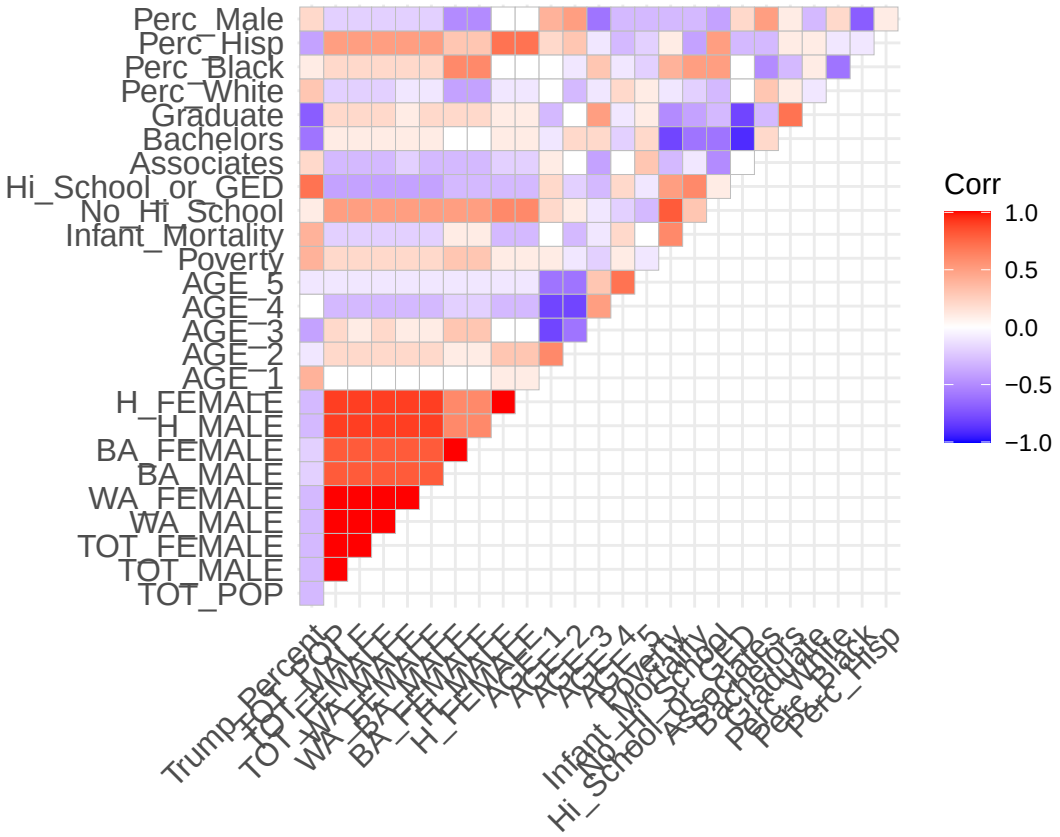
- Bin 1: Ages 0 to 19 years
- Bin 2: Ages 20 to 39 years
- Bin 3: Ages 40 to 59 years
- Bin 4: Ages 60 to 79 years
- Bin 5: Ages 80+ years

The race and gender data was combined into total percentage breakdowns by dividing the total number associated with each category by the total population value. Only one of the two options needed to be calculated as each of the categories was binary (e.g. Male OR Female). In this way we were able to create variables for:

- Percent Male
- Percent White
- Percent Black
- Percent Hispanic

The education data was available for 5 different designations within the population: people who don't have a high school education, those with only a high scool education or a GED, those with an Associates degree, those with a Bachelors and finally people who have a Gradute degree. In every model, the "No high School Education" variable seemed to show very low correlation with the target variable. A likely reason for this is that many of the people below the age of 20 would fall into that category. It's interesting to see that the percent of the population with a Graduate degree seems to have one of the highest levels of correlation (negative) with the percentage of votes garnered by Donald Trump.

A correlation graph for the variables available gave us an idea of those that may be more impactful for any model we were to create.



To start, we created a model that looked at all of the available variables. This led to some redundant variables, since the presence of both total population and total male population covers the data contained in the total female population (only two of these 3 variables would be used moving forward). The same was true for the Age Bin variables, with only 4 of the 5 necessary due to the bins always adding up to 100% (Age Bin 5 was removed for all models). Within this first model it was immediately evident that many of the variables provided very little predictive value. The stronger variables within the model were the four age bin variables, suggesting that the age distribution within a state is a somewhat strong predictor of the percentage of it that voted for Trump in 2016. Bins 2 and 3 in particular (ages 20 to 60) showed some of the lowest p-values in the first model and suggest they should be included in any model moving forward. The percentage of the population that is male proves to be a strong predictor, though a likely reason is the fact that it inherently contains the percentage of female population, with Trump being notoriously less popular among women. The other strong predictors in the first model were *White Population Percentage*, *Hispanic Population Percentage* and *Percentage with a Graduate Degree*. This model provided us with a baseline of $R^2 = 0.8627$, Adjusted $R^2 = 0.7412$ and an $Fstatistic = 10.55$ on 17 and 32 Degrees of Freedom.

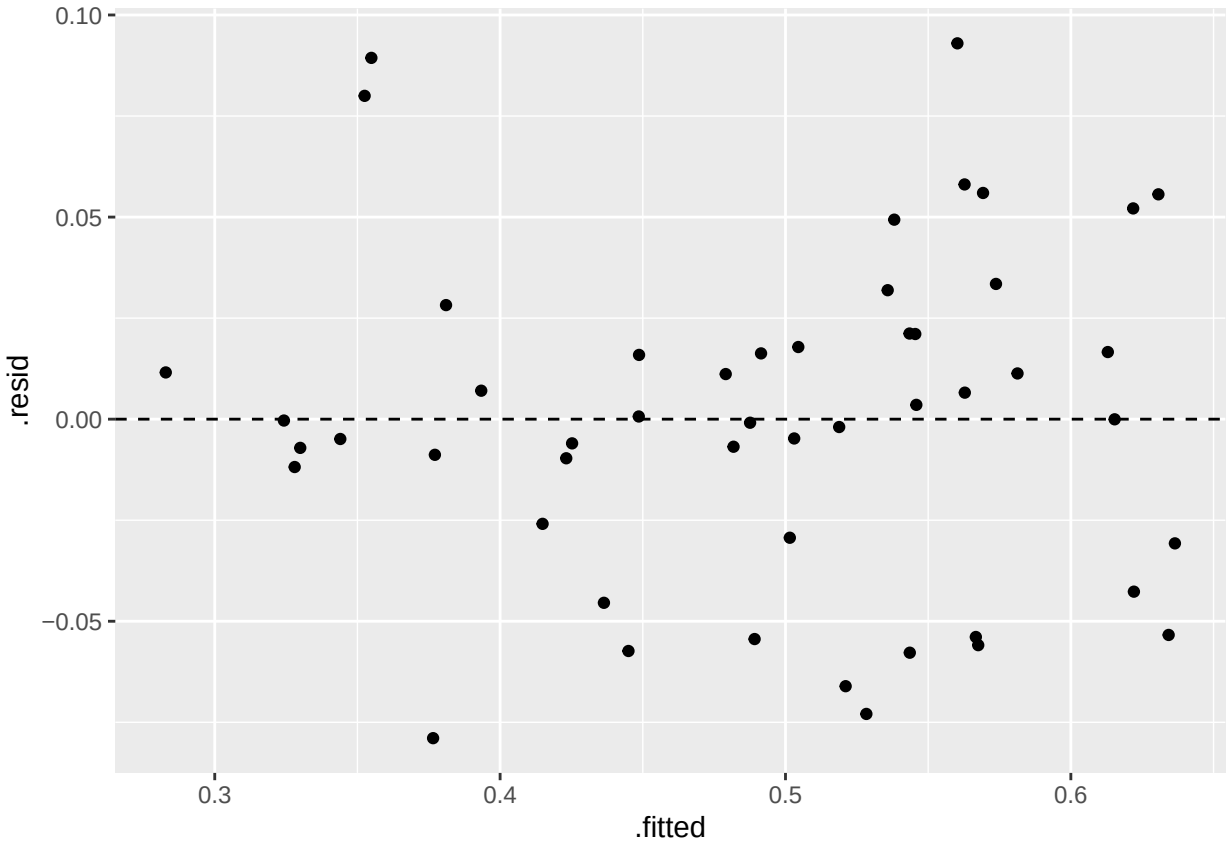
An additional model was created using only the data from the Kids Count Data Center, covering population information on education level, poverty and infant mortality. As in the first model, neither the poverty nor infant mortality information seem to be strong predictors for the percentage of votes for Trump from a given state. This model suggests that the population data provides a lot of value, since removing it resulted in $R^2 = 0.5681$, Adjusted $R^2 = 0.4962$ and an $Fstatistic = 7.894$ on 7 and 42 Degrees of Freedom.

Model 3 attempted to look at the population data alone - total numbers, percentage breakdowns and age distributions. This involved ignoring the fairly strong predictor of the Graduate degree percentage from model 1 and we therefore saw a noticeable dropoff in the quality of the model - $R^2 = 0.6783$, Adjusted $R^2 = 0.5223$ and an $Fstatistic = 4.348$ on 16 and 33 Degrees of Freedom. Focusing on only these variables did help identify which were less valuable in terms of a predictive model - we see that from the population data, Total numbers and numbers designated as Black Alone appear to be superflous. Moving forward, only

the numbers for White Only and Hispanic counts will be utilized.

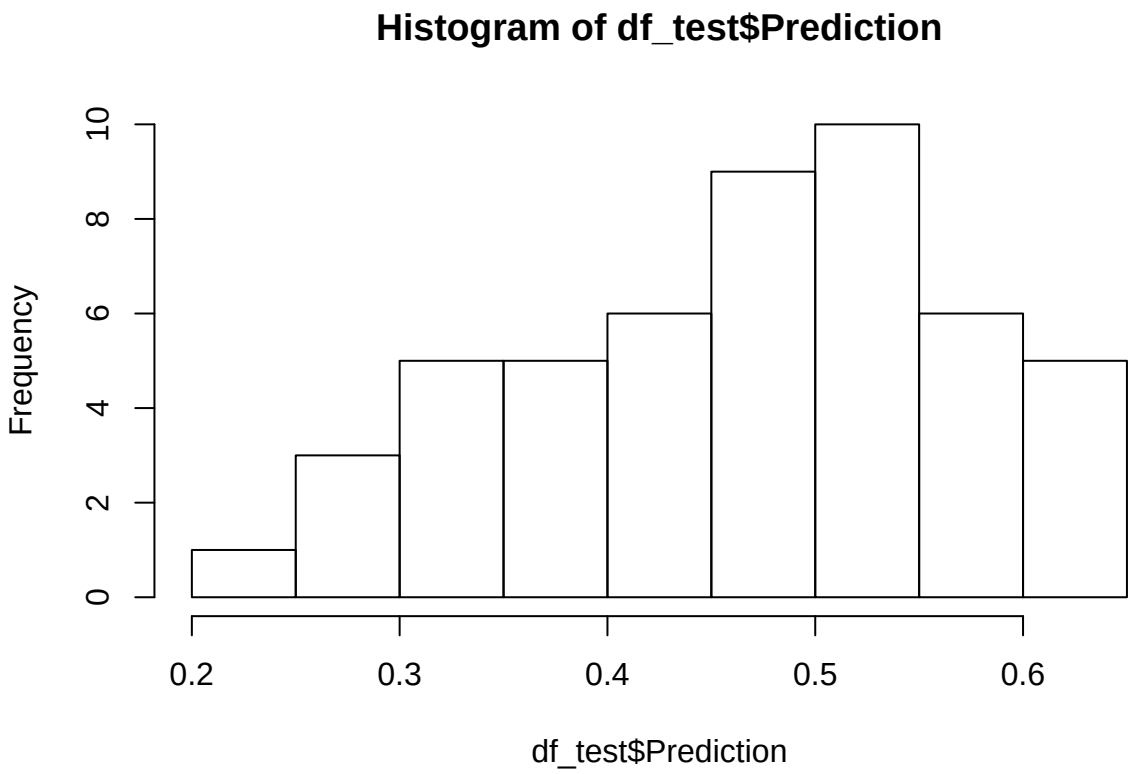
```
##
## Call:
## lm(formula = Trump_Percent ~ . - STNAME - TOT_FEMALE - AGE_5 -
##     No_Hi_School - Hi_School_or_GED - Poverty - BA_FEMALE - BA_MALE -
##     TOT_POP - TOT_MALE, data = df_train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.078925 -0.028479 -0.000189  0.020267  0.092984
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.242e+00  2.230e+00   1.006  0.32190
## WA_MALE          -9.308e-07  3.416e-07  -2.725  0.01021 *
## WA_FEMALE         8.938e-07  3.285e-07   2.721  0.01031 *
## H_MALE            1.620e-06  7.076e-07   2.289  0.02859 *
## H_FEMALE         -1.568e-06  6.967e-07  -2.250  0.03122 *
## AGE_1            -4.326e+00  2.552e+00  -1.695  0.09940 .
## AGE_2            -5.829e+00  2.586e+00  -2.254  0.03095 *
## AGE_3            -5.476e+00  2.526e+00  -2.167  0.03752 *
## AGE_4            -6.145e+00  3.038e+00  -2.023  0.05127 .
## Infant_Mortality -1.345e+00  1.132e+00  -1.189  0.24296
## Associates       -1.065e+00  6.660e-01  -1.599  0.11931
## Bachelors        -6.376e-01  5.439e-01  -1.172  0.24948
## Graduate         -1.980e+00  5.543e-01  -3.571  0.00112 **
## Perc_White        3.840e-01  8.921e-02   4.305  0.00014 ***
## Perc_Black        6.115e-01  1.872e-01   3.266  0.00254 **
## Perc_Hisp        -3.389e-01  1.196e-01  -2.834  0.00779 **
## Perc_Male         7.269e+00  2.407e+00   3.020  0.00485 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05059 on 33 degrees of freedom
## Multiple R-squared:  0.8411, Adjusted R-squared:  0.7641
## F-statistic: 10.92 on 16 and 33 DF,  p-value: 6.169e-09
```

Model 4 attempted to combine the most useful variables identified above while removing as many as possible to simplify the model. The 4 gender/race variables identified in model 3 (WA_MALE, WA_FEMALE, H_MALE, H_FEMALE), the 4 age bin variables as well as the percentage breakdowns and education data combined to create our fourth model. In it we find the best of each grouping of variables and a fiarly low dropoff in R^2 compared to the initial model while reducing the number of variables by 7. This model resulted in an $R^2 = 0.8411$, Adjusted $R^2 = 0.7641$ and an $Fstatistic = 10.92$ on 16 and 33 Degrees of Freedom. The residuals associated with the model do appear to be randomly distributed and all lower than 0.1. This would be our final model and the one used to generate final predictions.

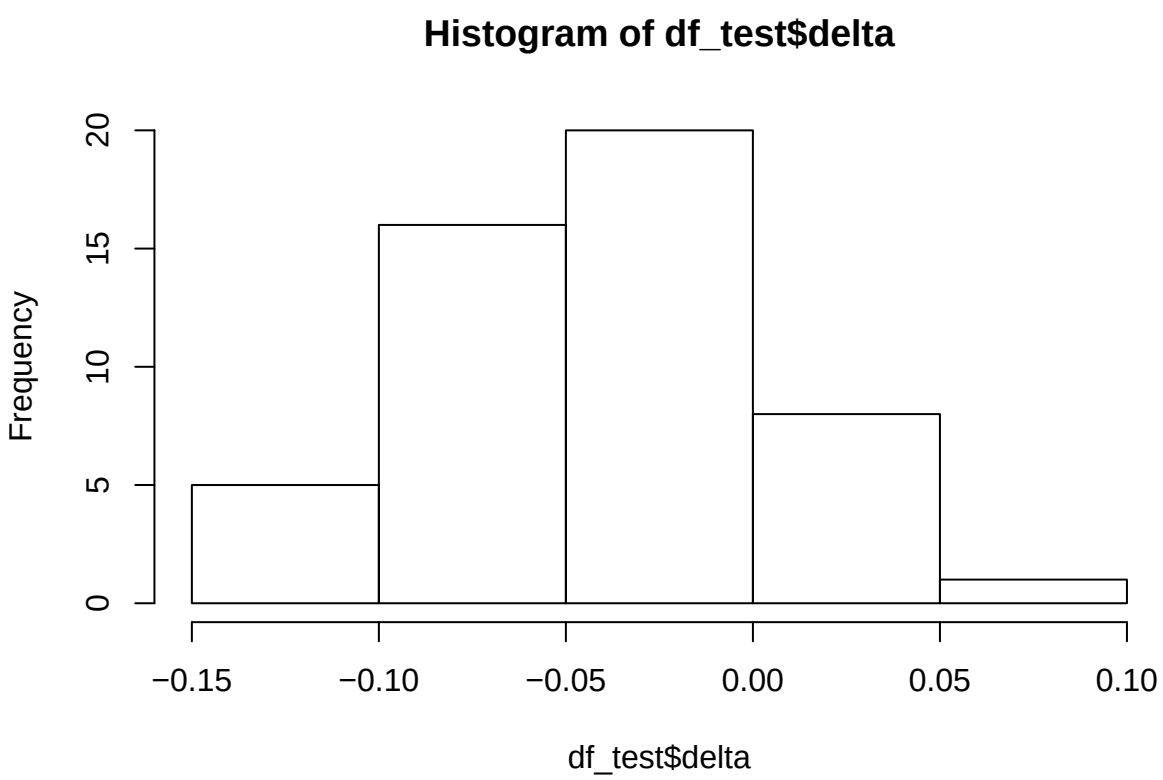


Discussion and Conclusions

Predictions were generated using *model4* and the demographic information available from 2019 (2018 for one variable). In our prediction we see a distribution that is centered around 50%, much like the histograms above odf both the 2016 and 2020 elections. Most states aren't won by huge margins, so one would expect many of the predictions to fall around this area. There does appear to be a projection with a percentage below 25, something that wasn't seen in 2016.



We compared the prediction to the actual values and see that the model predicts vote percentages as far off as 15%. In an election this is an unacceptable margin for a prediction as the stakes are so high.



Based on these results it appears that the demographic data we selected is not enough to create a viable model predicting election results. We pursued this effort due to the recent lack of reliability in poll projections, assuming that the somewhat subjective nature inherent in polling is responsible for some of these discrepancies. The theory was that a more objective set of data used as predictors would result in a more standardized model unaffected by the mood or disposition of respondents. While the results of this study cannot definitevly reject tis hypothesis, we can see that the information used here is simply not sufficient for the type of projection we were attempting. While demographic information may play a large role in the way the population may lean in any given election, it cannot be forgotten that the votes need to be cast - and not everyone votes. What would perfhaps be a valuable addition to this study would be demographic information on those who actually vote.

Possible future work would include more detailed information, including county-level demographic data (the initial goal of this study), additional population characteristics and potentially using older data used for the training dataset. This last point would be complicated as the parties running in Presidential Elections change fairly significantly over time and certain demographics that may be associated with one today may not have also done so in a prior election.

References

1. Silver, N. (2020, November 11). The Polls Weren't Great. But That's Pretty Normal. Retrieved December 17, 2020, from https://fivethirtyeight.com/features/the-polls-werent-great-but-thats-pretty-normal/

2. Kennedy, C., Blumenthal, M., Clement, S., Clinton, J., Durand, C., Franklin, C., . . . Wlezien, C. (2018, February 03). Evaluation of the 2016 Election Polls in the United States. Retrieved December 17, 2020, from https://academic.oup.com/poq/article/82/1/1/4837043

Appendix with R Code

```
library(dplyr)
library(ggplot2)
library(ggcorrplot)
library(reshape2)
library(broom)
library(jtools)
```

```
vote_data_2020 <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/pr
vote_data_2016 <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/Ra
```

```
vote_data_2016$Percent <- vote_data_2016$candidatevotes/vote_data_2016$totalvotes
trump_perc_16 <- vote_data_2016[,c(2,7,8)]%>%
  filter(state!='District of Columbia')
hist(trump_perc_16$Percent, main = "Trump Vote Percentage by State in 2016", breaks = 10)
```

```
Total20 <- vote_data_2020 %>%
  group_by(state, candidate) %>%
  summarise(
    total = sum(total_votes)
  )
trump_perc_20 <- group_by(Total20, state) %>%
  mutate(Percent = total/sum(total)) %>%
  filter(candidate=='Donald Trump') %>%
  select(-candidate) %>%
  filter(state!='District of Columbia')

hist(trump_perc_20$Percent, main = "Trump Vote Percentage by State in 2020", breaks = 10)
```

```
poverty_df <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/Raw%20
poverty_df <- poverty_df[which(poverty_df$TimeFrame == 2016 | poverty_df$TimeFrame == 2019),]
poverty_df$Location <- as.character(poverty_df$Location)
```

```
education_df <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/Raw
education_df <- education_df[which(education_df$TimeFrame == 2016 | education_df$TimeFrame == 2019),]
education_df$Location <- as.character(education_df$Location)
education_df$TimeFrame <- as.numeric(education_df$TimeFrame)
```

```
inf_mort_df <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/Raw%
inf_mort_df <- inf_mort_df[which(inf_mort_df$TimeFrame == 2016 | inf_mort_df$TimeFrame == 2018),]
inf_mort_df$Location <- as.character(inf_mort_df$Location)
inf_mort_df$Data <- as.numeric(inf_mort_df$Data)
inf_mort_df$Data <- inf_mort_df$Data / 1000
```

```
census_2016 <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/Raw%
census_2019 <- read.csv('https://raw.githubusercontent.com/mkollontai/DATA621/main/Final%20Project/Raw%

percent_pop <- function(df){
  df$PERC_POP <- df$TOT_POP
```

```
  st = ""
  tot_pop = 0
  for (i in 1:nrow(df)){
    if (df[i,'STNAME'] != st){
      st = df[i,'STNAME']
      tot_pop = df[i,'TOT_POP']
    }
    df[i,'PERC_POP'] = df[i,'TOT_POP']/tot_pop
  }
  return (df)
}

census_2016$TOT_MALE = census_2016$TOT_POP - census_2016$TOT_FEMALE
census_2019$TOT_MALE = census_2019$TOT_POP - census_2019$TOT_FEMALE

census_2016 <- percent_pop(census_2016)
census_2019 <- percent_pop(census_2019)
```

```
find_ages<- function(df_to, df_from){
  df_to$AGE_1 <- NA
  df_to$AGE_2 <- NA
  df_to$AGE_3 <- NA
  df_to$AGE_4 <- NA
  df_to$AGE_5 <- NA

  for (i in 1:nrow(df_to)){
    df_to[i,'AGE_1']=
      df_from[
        which(
          df_from$AGEGRP==1 & df_from$STNAME == df_to[i,'STNAME']),'PERC_POP']
    df_to[i,'AGE_2']=
      df_from[
        which(
          df_from$AGEGRP==2 & df_from$STNAME == df_to[i,'STNAME']),'PERC_POP']
    df_to[i,'AGE_3']=
      df_from[
        which(
          df_from$AGEGRP==3 & df_from$STNAME == df_to[i,'STNAME']),'PERC_POP']
    df_to[i,'AGE_4']=
      df_from[
        which(
          df_from$AGEGRP==4 & df_from$STNAME == df_to[i,'STNAME']),'PERC_POP']
    df_to[i,'AGE_5']=
      df_from[
        which(
          df_from$AGEGRP==5 & df_from$STNAME == df_to[i,'STNAME']),'PERC_POP']
  }
  return (df_to)
}

add_demo_info <- function(df, year){
  df$Poverty <- NA
  df$Infant_Mortality <- NA
```

```
df$No_Hi_School <- NA
df$Hi_School_or_GED <- NA
df$Associates <- NA
df$Bachelors <- NA
df$Graduate <- NA

for (i in 1:nrow(df)){
  df[i,'Poverty'] =
    poverty_df[
      which(
        poverty_df$TimeFrame == year & poverty_df$Location == df[i,'STNAME']),'Data']
  df[i,'No_Hi_School'] =
    education_df[
      which(
        education_df$TimeFrame == year &
        education_df$Location == df[i,'STNAME'] &
        education_df$Education == 'Not a high school graduate'
      )
      ,'Data']
  df[i,'Hi_School_or_GED'] =
    education_df[
      which(
        education_df$TimeFrame == year &
        education_df$Location == df[i,'STNAME'] &
        education_df$Education == 'High school diploma or GED'
      )
      ,'Data']
  df[i,'Associates'] =
    education_df[
      which(
        education_df$TimeFrame == year &
        education_df$Location == df[i,'STNAME'] &
        education_df$Education == 'Associate\'s Degree'
      )
      ,'Data']
  df[i,'Bachelors'] =
    education_df[
      which(
        education_df$TimeFrame == year &
        education_df$Location == df[i,'STNAME'] &
        education_df$Education == 'Bachelor\'s Degree'
      )
      ,'Data']
  df[i,'Graduate'] =
    education_df[
      which(
        education_df$TimeFrame == year &
        education_df$Location == df[i,'STNAME'] &
        education_df$Education == 'Graduate degree'
      )
      ,'Data']
  if (year == 2019){
    year2 <- 2018
```

```
    }else{
      year2 <- year
    }
    df[i,'Infant_Mortality'] =
      inf_mort_df[
        which(
          inf_mort_df$TimeFrame == year2 & inf_mort_df$Location == df[i,'STNAME']),'Data']

  }
  return(df)
}

age_race_breakdown <- function(df){
  df$Perc_White <- (df$WA_MALE + df$WA_FEMALE) / df$TOT_POP
  df$Perc_Black <- (df$BA_MALE + df$BA_FEMALE) / df$TOT_POP
  df$Perc_Hisp <- (df$H_MALE + df$H_FEMALE) / df$TOT_POP
  df$Perc_Male <- df$TOT_MALE / df$TOT_POP

  return(df)
}
```

```
df_train <- census_2016 %>%
  filter(AGEGRP==0 & STNAME!="District of Columbia" ) %>%
  select(-AGEGRP & -X)
df_test <- census_2019 %>%
  filter(AGEGRP==0 & STNAME!="District of Columbia") %>%
  select(-AGEGRP & -X )

df_train <- find_ages(df_train,census_2016) %>%
  select(-PERC_POP)
df_test <- find_ages(df_test,census_2019) %>%
  select(-PERC_POP)

df_train <- add_demo_info(df_train,2016)
df_test <- add_demo_info(df_test,2019)

df_train <- age_race_breakdown(df_train)
df_test <- age_race_breakdown(df_test)

colnames(trump_perc_16)[1] <- 'STNAME'
df_train <- merge(df_train,trump_perc_16[,-c(2)], by ='STNAME')
names(df_train)[names(df_train) == 'Percent'] <- 'Trump_Percent'
```

```
d <- melt(df_train[,-c(1)])
ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram()
```

```
corr <- round(cor(df_train[,-c(1)]),1)
ggcorrplot(corr, type = 'upper')
```

```
model1 <- lm(Trump_Percent ~ . - STNAME - TOT_FEMALE - AGE_5,df_train)
summary(model1)
```

```
model2 <- lm(Trump_Percent ~ Poverty + Infant_Mortality + No_Hi_School + Hi_School_or_GED + Associates +
summary(model2)
```

```
model3 <- lm(Trump_Percent ~ . -STNAME - TOT_FEMALE -Poverty - Infant_Mortality - No_Hi_School - Hi_Scho
summary(model3)
```

```
model4 <- lm(Trump_Percent ~ . -STNAME -TOT_FEMALE -AGE_5 -No_Hi_School -Hi_School_or_GED -Poverty -BA_
summary(model4)
```

```
ggplot(model4, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype='dashed')
```

```
#predict2020 <- predict(model4,df_test)
df_test$Prediction <- predict(model4,df_test)
df_test$delta <- df_test$Prediction - trump_perc_20$Percent
hist(df_test$Prediction)
```

```
hist(df_test$delta)
```