# 2020 US presidential election prediction

Taojun Wang

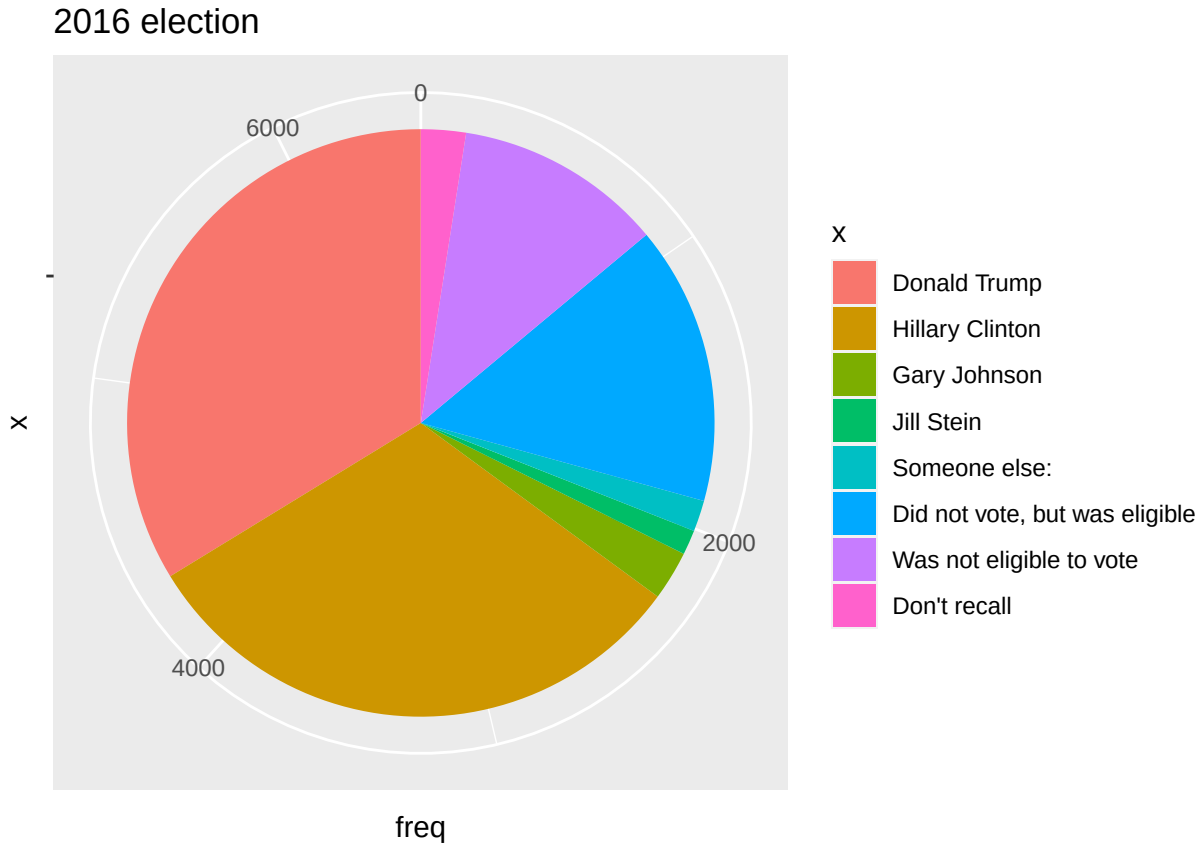
2020/11/1

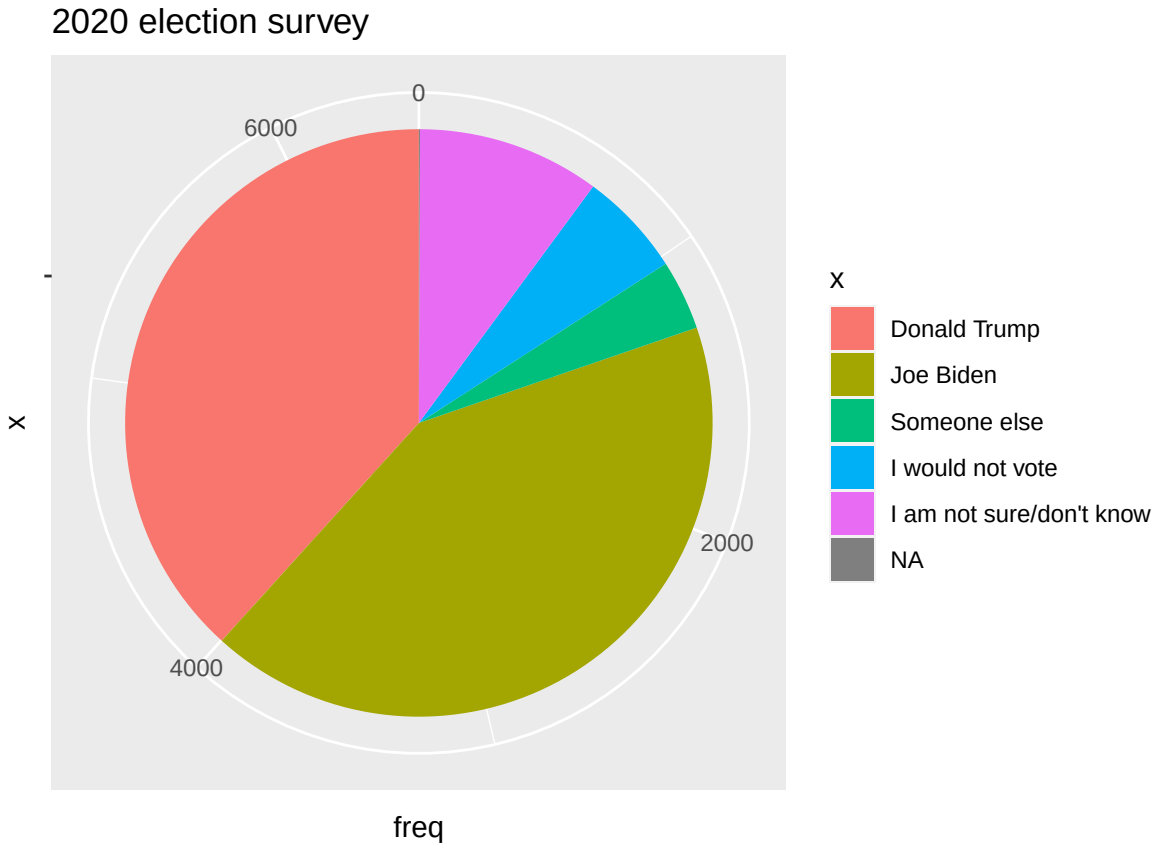
## Abstract

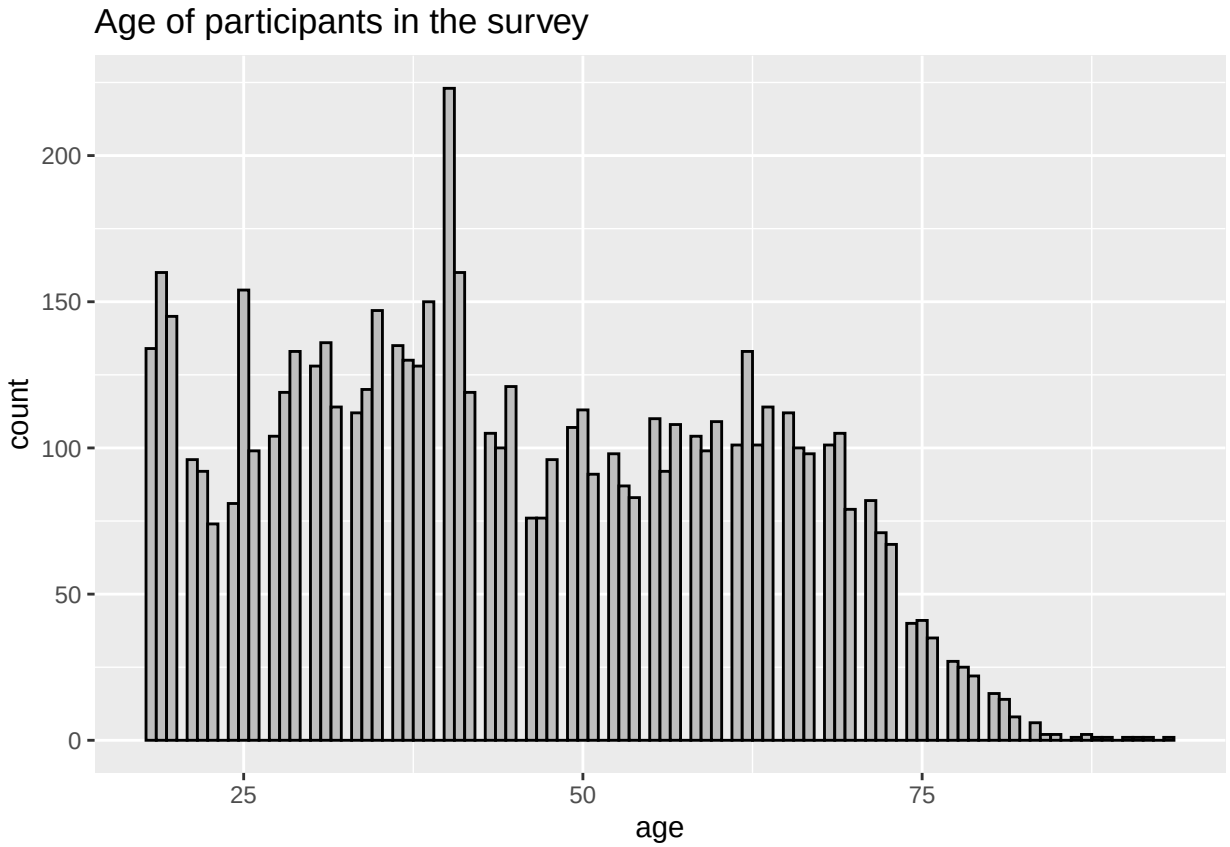
This report is aimed to predict the result of 2020 US presidential election. We use multilevel regression of post-stratification to process the survey day from nearly 7000 participants and build a logistic model based on it. We use the model to deal with the US census data, and find that the election is very close (Trump is expected to get 49%-52% of all votes).

## Background Introduction:

Forecasting the US presidential elections is always a meaningful endeavor owing to America's important status in the world. Therefore, every election period in the country does not only attracts domestic but also international attention. In 2016, Donald Trump made an inconceivable triumph in the election when most polls and statistical analysis considered him as an underdog, and people want to know whether Joe Biden could beat him this time. In this case, through understanding the performance and opinions of the general public, it would be possible to determine the candidate that is most likely to win the elections. This report is based on evaluating and presenting a forecast on the candidate that will win the 2020 elections.


2016 election
0
2000
4000
6000
freq
x
Donald Trump
Hillary Clinton
Gary Johnson
Jill Stein
Someone else:
Did not vote, but was eligible
Was not eligible to vote
Don't recall

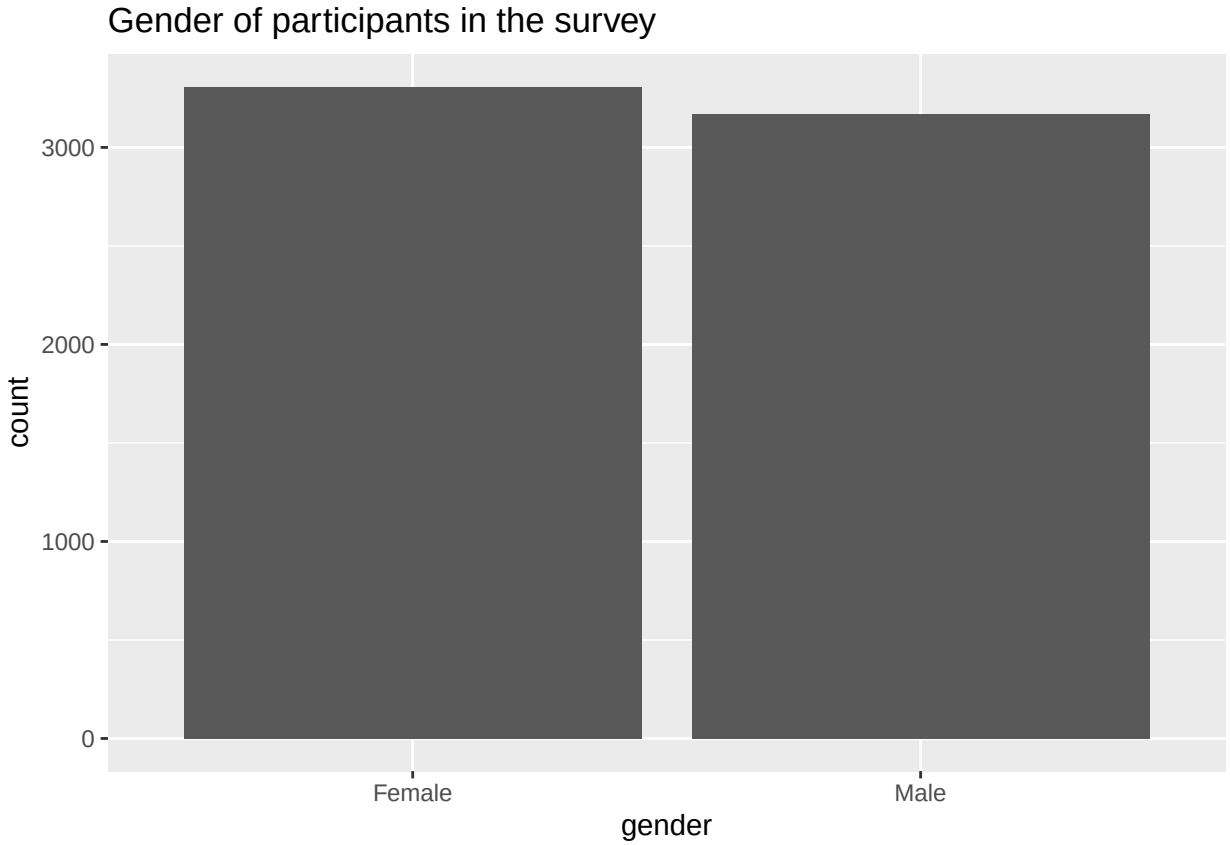

2020 election survey
0
2000
4000
6000
freq
x
-
x
Donald Trump
Joe Biden
Someone else
I would not vote
I am not sure/don't know
NA




```
##
##                             White         Black, or African American
##                              4816                                774
##   American Indian or Alaska Native               Asian (Asian Indian)
##                                90                                102
##                     Asian (Chinese)                   Asian (Filipino)
##                                84                                 46
##                   Asian (Japanese)                    Asian (Korean)
##                                21                                 14
##                  Asian (Vietnamese)                    Asian (Other)
##                                13                                 37
## Pacific Islander (Native Hawaiian)       Pacific Islander (Guamanian)
##                                10                                  1
##          Pacific Islander (Samoan)           Pacific Islander (Other)
##                                 3                                  8
##                   Some other race
##                               460
```

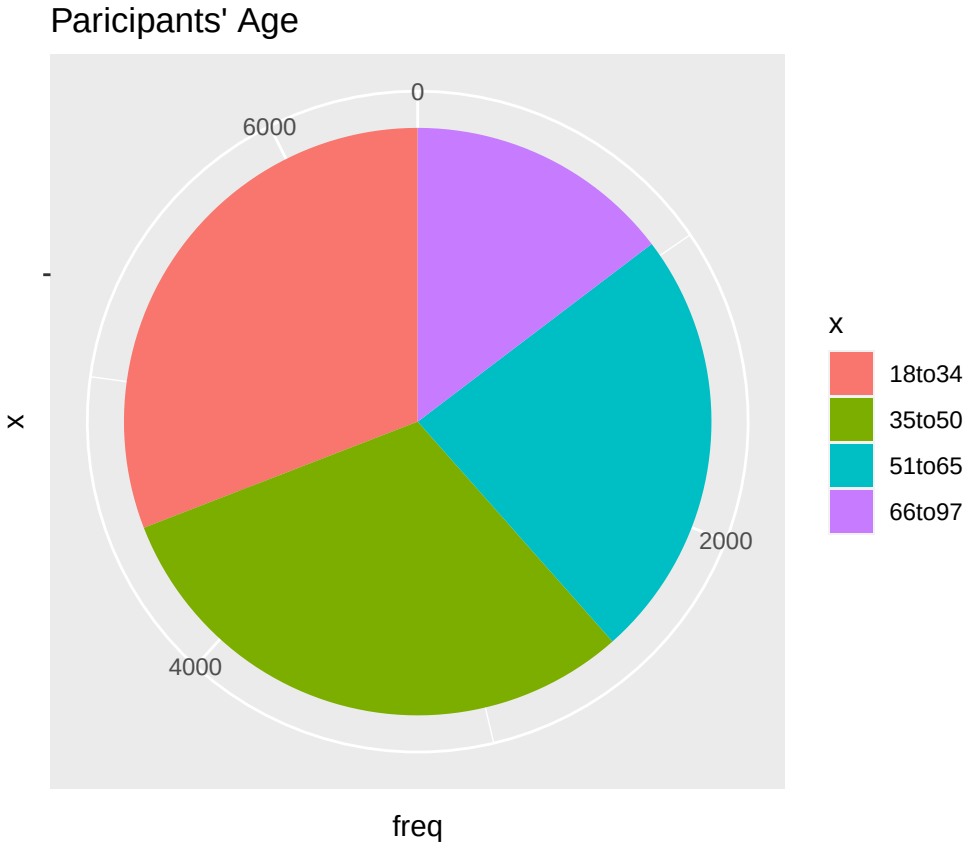
```
##
##                         3rd Grade or less
##                                        11
##               Middle School - Grades 4 - 8
##                                        26
##                 Completed some high school
##                                       638
```

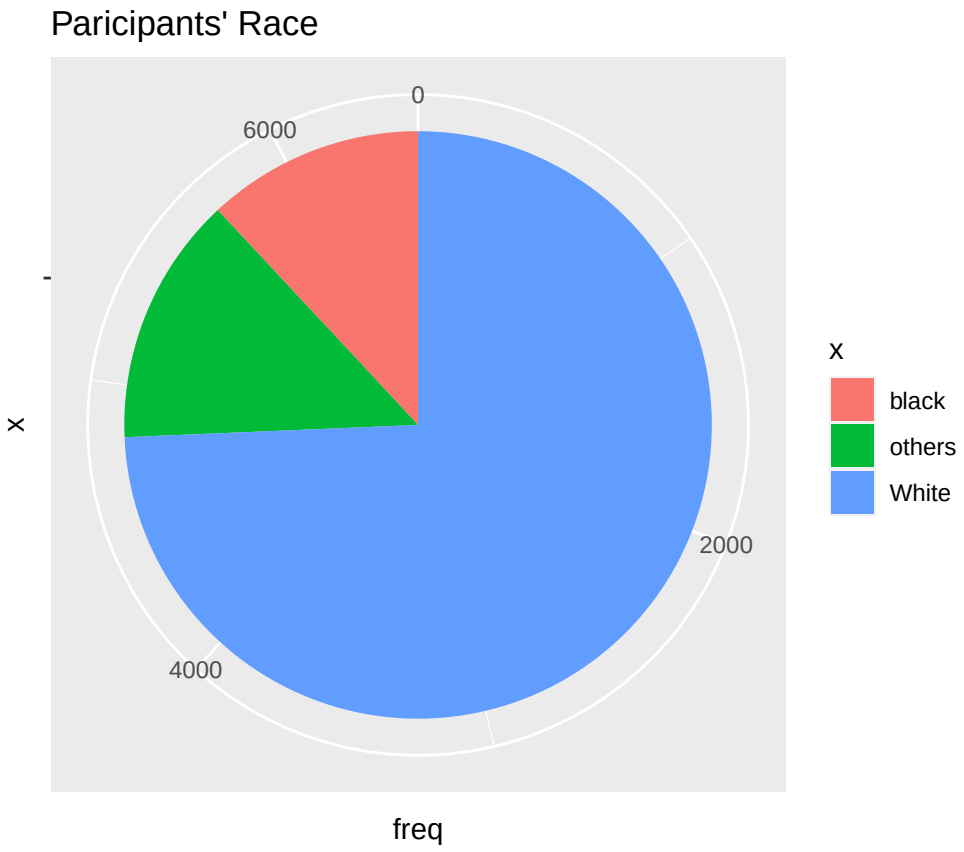
```
##                       High school graduate
##                                       1079
## Other post high school vocational training
##                                        324
##      Completed some college, but no degree
##                                       1327
##                            Associate Degree
##                                        570
##        College Degree (such as B.A., B.S.)
##                                       1477
##     Completed some graduate, but no degree
##                                        238
##                              Masters degree
##                                        643
##                            Doctorate degree
##                                        146
```

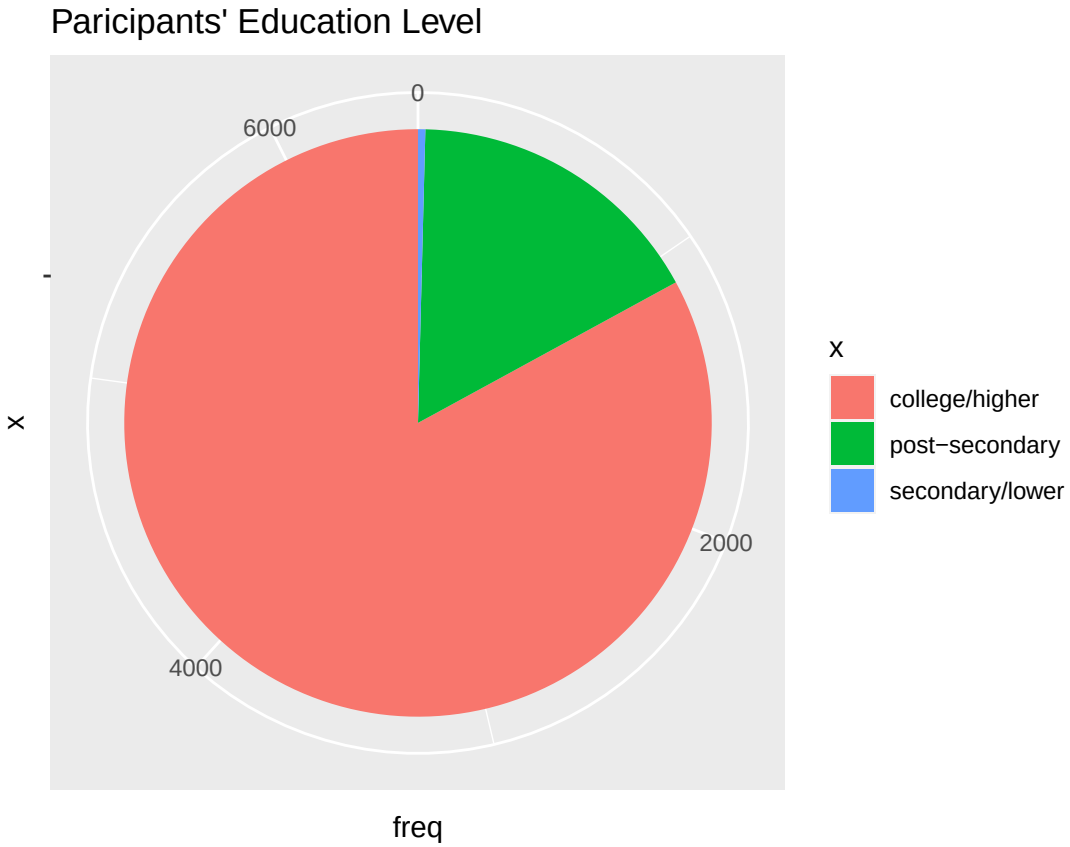



# Survey Data The data analysis of this report will be based on the survey data collected by Democracy Fund and UCLA Nationscape from 6479 participants. The data was collected through 24 weeks in the first half of 2020 fielded by a market research platform called Lucid (Nationscape Data Set, September 2020). The full data set includes 265 variables, and five of them will be taken into consideration in this report to simplify our analysis . The variable vote_2020 will be used as the response variable, which is the candidates participants claim to vote for in the survey. The predictor variables in use will include age, gender, race ethnicity and education level. It is obvious to find that among the participants people who plan to vote for Biden are more than that who plan to Trump. In terms of participants' age, the plot show a bi-modal characteristics, whether most people are around 40 or 60. Also, most participants have a relatively high education level, half of whom have completed college degree (or higher), which might not be

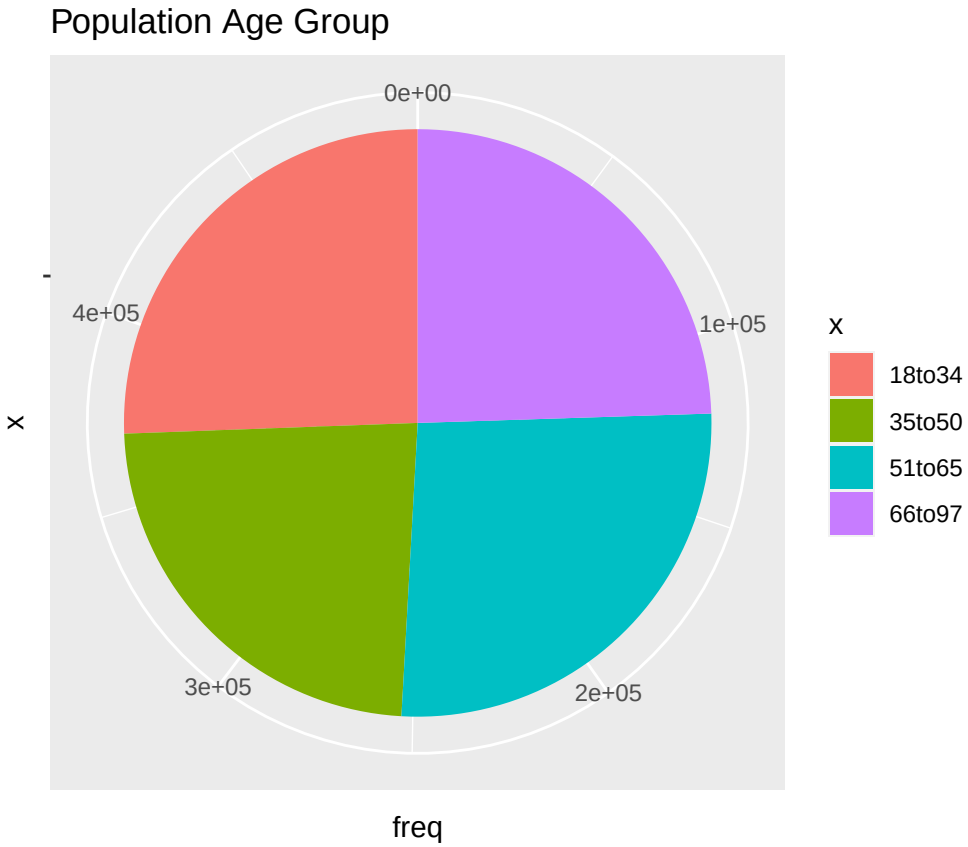
very representative for the entire population. Therefore, this report plans to use multilevel regression with post-stratification (MRP) to predict the final election.

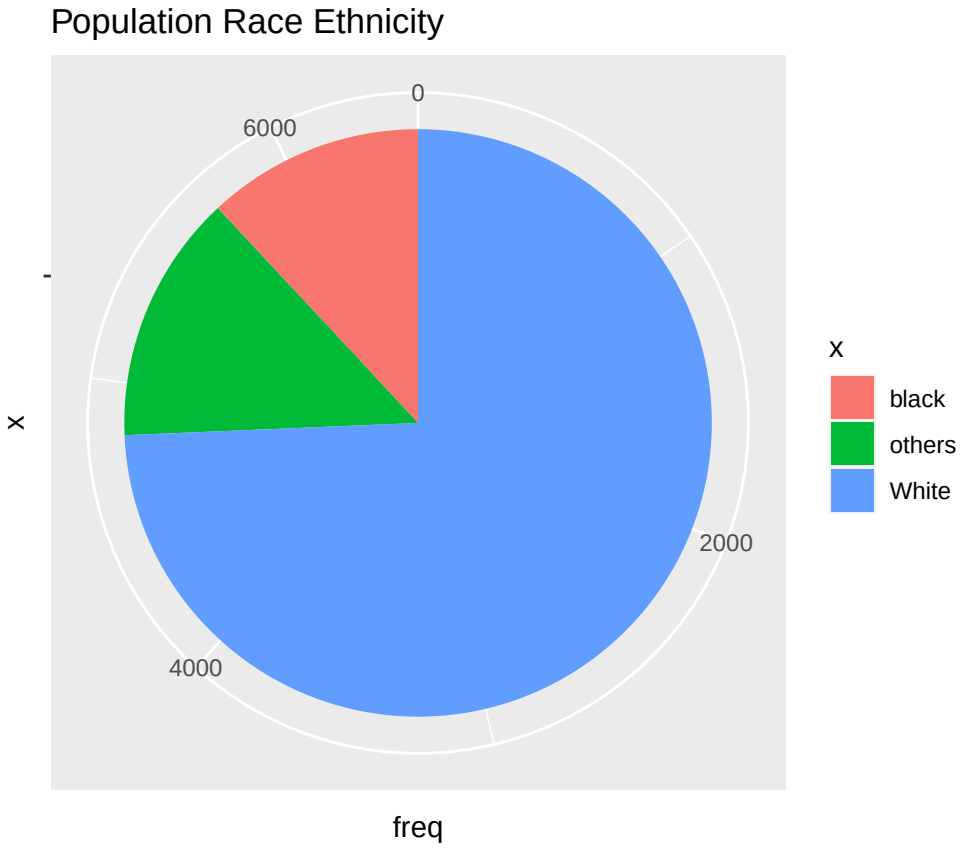


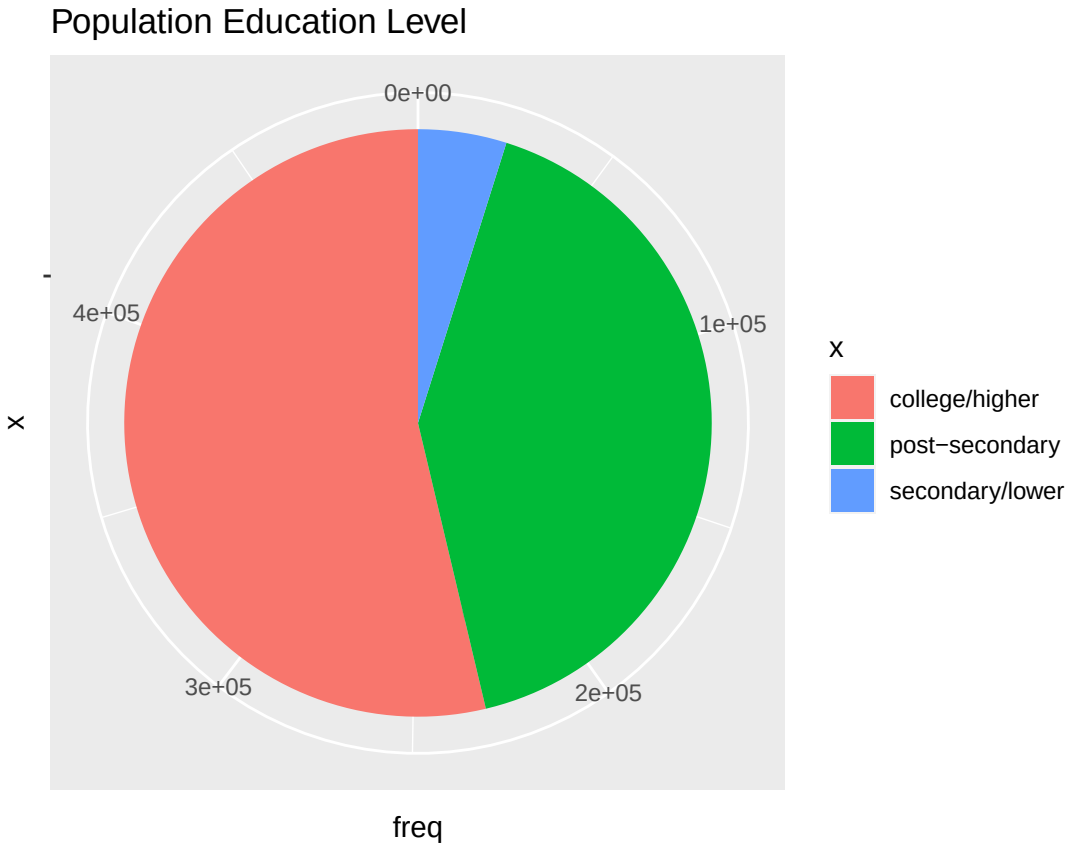

Paricipants' Race
0
6000
4000
2000
x
freq
x
black
others
White




# Post-stratification Dataset In order to use MRP, each variable is divided into many levels in order to form small cells which are favorable for model construction. Age(above 18, eligible for voting) is divided in to four age groups. For race ethnicity, all the minorities other than African Americans are simply included into one category, “others”. That is to say, there are three classifications in the “race” variable, which respectively “white”, “black”, and “others”. Education level is classified into three categories, “college/higher”,“post-secondary”, and “secondary/lower”. The same method is used to process both the survey and the census data, and we get the post-stratification data set on which we will build models. From the plot above(survey) and below, we can find that the there are too many highly-educated people in the survey sample compared with the entire population. Finally, since it is obvious that no one else other than Biden or Trump would win the election, the answers such as “not sure whom to vote for/not vote” will add no explanatory information to the model, we can simply take them away and make our response variable binary (Trump/Biden) We set an new variable “vote_Trump_or_not” to denote participants who vote for Trump as 1 and denote the Biden supports as 0


Population Age Group
0e+00
1e+05
2e+05
3e+05
4e+05
x
freq
x
18to34
35to50
51to65
66to97


Population Race Ethnicity
0
6000
2000
4000
x
freq
x
black
others
White




```
##
## Call:
## glm(formula = vote_Trump_or_not ~ age_group + education + gender +
##     race, family = "binomial", data = survey)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5379  -1.1822  -0.4546   1.0488   2.2780
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.51708    0.14024 -17.948  < 2e-16 ***
## age_group35to50            0.50613    0.07887   6.418 1.38e-10 ***
## age_group51to65            0.42863    0.08269   5.183 2.18e-07 ***
## age_group66to97            0.35059    0.09301   3.769 0.000164 ***
## educationpost-secondary    0.29920    0.08434   3.547 0.000389 ***
## educationsecondary/lower   0.02218    0.53462   0.041 0.966909
## genderMale                 0.41015    0.05920   6.928 4.28e-12 ***
## raceothers                 1.41033    0.15435   9.137  < 2e-16 ***
## raceWhite                  2.11807    0.13361  15.852  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7197.8  on 5199  degrees of freedom
```

```
## Residual deviance: 6609.6  on 5191  degrees of freedom
## AIC: 6627.6
##
## Number of Fisher Scoring iterations: 4
```

# Model

We estimate the data to follow a logistic regression like:

$$Pr(y_i = 1) = \text{logit}^{-1}\left(\alpha^{age}_{a[i]} + \alpha^{educ}_{e[i]} + \alpha^{gender}_{g[i]} + \alpha^{race}_{r[i]}\right)$$

where

$$y_i = 1$$

when the individual i claims to vote for Trump. The

$$\alpha$$

s respectively refer to age group, education level group, gender and race group. The notation

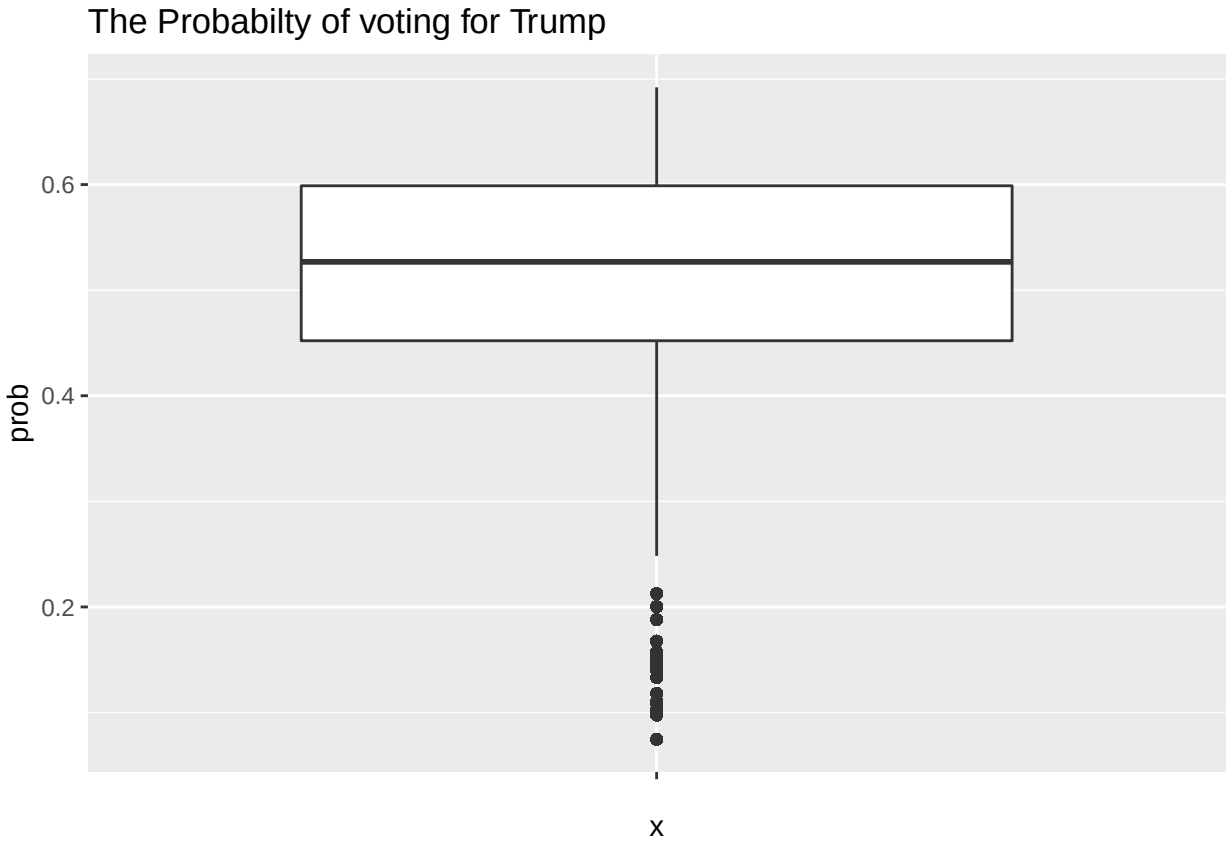
$$a[i]$$

refers to the age group an participant belongs to in the post-stratification data set. We assume that for each individual, his response to each variable is normally distributed with constant variance. From the data we find that the category of education level "secondary/lower"contributes very little to the explanatory results but has a higher estimation error in the same time, so we decide to combine it into the category "high_school/lower".

```
##
## Call:
## glm(formula = vote_Trump_or_not ~ age_group + education + gender +
##     race, family = "binomial", data = survey)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5349  -1.1819  -0.4531   1.0514   2.2784
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.51807    0.14024 -17.956  < 2e-16 ***
## age_group35to50             0.50615    0.07886   6.418 1.38e-10 ***
## age_group51to65             0.42966    0.08267   5.197 2.02e-07 ***
## age_group66to97             0.35147    0.09299   3.780 0.000157 ***
## educationhigh_school/lower  0.29338    0.08357   3.511 0.000447 ***
## genderMale                  0.40921    0.05917   6.915 4.66e-12 ***
## raceothers                  1.41053    0.15434   9.139  < 2e-16 ***
## raceWhite                   2.11932    0.13360  15.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7197.8  on 5199  degrees of freedom
## Residual deviance: 6609.8  on 5192  degrees of freedom
```

```
## AIC: 6625.8
##
## Number of Fisher Scoring iterations: 4
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0746  0.4521  0.5268  0.4999  0.5989  0.6921
```




## Diagonis

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07247 0.38159 0.49965 0.47466 0.61284 0.70547
```

```
## [1] 0.4711538
```

To diagnose the effectiveness of our model, we randomly slice the survey data into two parts: 20% of the data is the test data, and the other 80% is the train model. We build a training model based on the training data using the same method as above, and use this model to predict the test data. The prediction result of the test data is: the average probability of voting for Trump is 47.466%. In the actual test data, 47.12% of people claim to vote for Trump, which is quite close to the predicted value. It shows that our model is valid.

## Discussion:

Among all age groups, young people (less than 35) are most unlikely to vote for Trump. In contrast, middle-aged people (35-50) are most likely to vote for Trump. If other conditions are kept the same (this assumption

will be always used in the following discussion), a middle-aged person is 62.5% more likely to vote for Trump than a young person. The number for people from 51 to 65 and people from 66 to 97 are 60.5% and 58.5% respectively. People who receive a lower level of education is more likely to vote for Trump (57.3% higher than that of people who have a college degree or higher). Men are more likely to vote for Trump (60.1% higher than that of women). African American are most unlikely to vote for Trump among all races. If a person is white, it will increase his probability of voting for Trump by 89.3% compared with a black person).

The prediction result shows that among all 497338 observations in the census data, the mean probability of voting for Trump is 49.99%, while the median is 52.68%. This is because even though some people are very unlikely to vote for Trump (very low probability), but most people show an ambitious attitude to him, so the election is very close and it it hard to forecast the final result. In the survey sample, the proportion of poorly-educated people is much lower than that in the entire population, so the supporting rate for Trump is under-estimated.

# Weakness

The result of a presidential election is not directly determined by the number of national votes, but is determined by the result in each state. That is to say, a candidate winning more than half votes could still lose the election. For simplicity, this report does not consider the situation in each state, which will negatively influence the effectiveness of the final prediction. Also, the classification of "cells" in the post-stratification part is not fine enough. For example, all white people, regardless of Hispanic or not, are simple included into one category. However, Hispans can have a very different voting tendency compared with European whites. Finally, the predictor variables are too few (only four are used). The nature of voters is very complex, and cannot be included in only four variables.

References: Democracy Fund and Voter Group Study(2020 September). "Nationscape Data Set". https://www.voterstudygroup.org/publication/nationscape-data-set Ruggles, S. et al. (2020) 'IPSUM USA data extract'. Minneapolis, MN: IPUMS USA, (PUMS USA: Version 10.0 [dataset]). doi: https://doi.org/10.18128/D010.V10.0.

```
##
## To cite R in publications use:
##
##   R Core Team (2020). R: A language and environment for statistical
##   computing. R Foundation for Statistical Computing, Vienna, Austria.
##   URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2020},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

# Codes

Code supporting this analysis can be found at https://github.com/TaojunWang/US-election-prediction/blob/main/final%20work.Rmd