What Lies Behind Voters' Choice? — A Data-Question Story on the Pre-Election Survey

Yutong Yang yy3116

2/3/2021

Contents

1. Introduction

The *American National Election Studies* (ANES) are surveys of voters in the U.S. on the national scale. For each presidential election since 1948, ANES collects responses from respondents both before and after the election. The goal of ANES is to understand political behaviors using systematic surveys. ANES's data and results have been routinely used by news outlets, election campaigns and political researchers.

ANES 2020 Exploratory Testing Survey Questionnaire Specifications

The *ANES 2020 Exploratory Testing Survey* was conducted for the purpose of **testing new questions and carrying out methodological research** to inform the design of the ANES 2020 Time Series study. As distinct from many ANES pilot surveys, the primary aim of the study was to allow for more targeted experimentation and testing of longer batteries of questions, with **less concern for estimation of population characteristics**. In line with these goals, the study relies on non-probability samples, and no sampling weights are provided. Therefore, the data might not be an ideal choice for making inferences about the distribution of opinions in the American electorate. Accesss more details here.

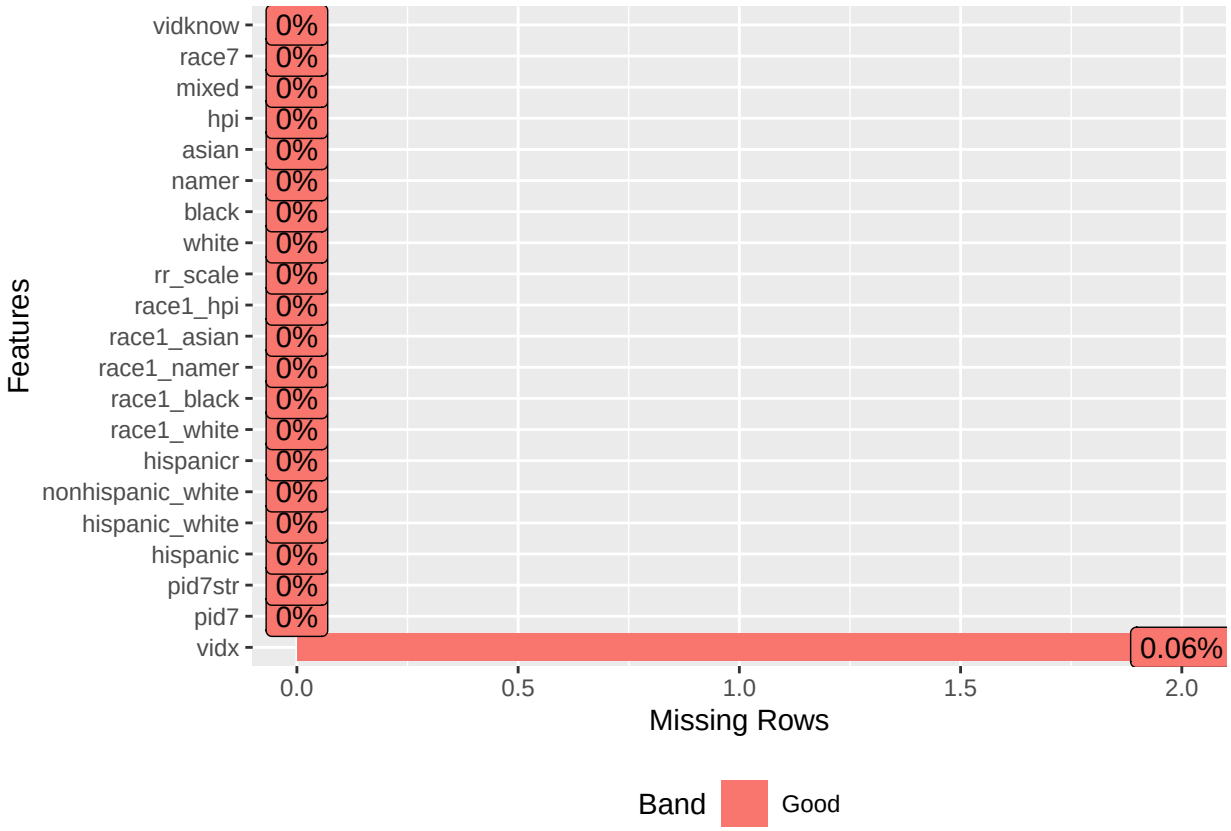
Nevertheless, the data set still reveals valuable information as it reflects what American voters think when they are provided with a long series of questions. Besides, this being the only available ANES data for 2020 presidential election makes it relatively appropriate for understanding the voters' opinion and choice.

Taking the factors above into account, the data is chosen as the primary target of analysis. The data collection was conducted between **April 10, 2020 and April 18, 2020**. The combined final sample includes responses from *3,080* adult citizens from across the United States.

2. Data Preprocessing

2.1: Missing Values and Data Types

The first step of data cleaning is checking out on the missing values, after plotting out the missing values, only the variable *ethnic2* have missing values. Since the percentage of missing values is indeed small, only accounting for a tiny fraction of the data set, I directly dropped the missing values.



Secondly, the data types can also be viewed using R. However, considering the huge amount of variables included in this data set, it would be wiser to directly reference the codebook.

2.2: Varaible selection with initial linear regression

Question1: What factors have the highest correlation with the voters' choice?

This is actually the most fundamental as well as the most centered question for the whole study. To answer this core question and in the meantime conducting a variable selection process, I ran a linear regression on the data set to get a quick view of the key variables, using the *vote20jb* as the independent variable and omitting the text variables which contain description on the users' devices. The data set was split into a training set and a test set, with 80% of the original data as the training set.

ps. the variable vote20jb reflects the result of the question: "If the 2020 presidential election were between Donald Trump for the Republicans and Joe Biden for the Democrats, would you vote for Donald Trump, Joe Biden, someone else, or probably not vote?" Thus, it can be used to identify supporters of Donald Trump and Joe Biden.

```
# screen the variables
df_reg<-df_init[,c(11:470)]
# Partitioning; get 80% train set
splitPercent <- round(nrow(df_reg) %*% .8)
set.seed(1234)
idx      <- sample(1:nrow(df_reg), splitPercent)
trainSet <- df_reg[idx, ]
testSet  <- df_reg[-idx, ]
```

```
# fit the model
# lmfit <- lm(vote20jb ~ ., trainSet)
```

The regression results shows that the following variables have the highest significance: (Since the regression was ran on too may variables, I stored the results in a file.)

	coe	res	pvalue
pk_cjusin kongers	-0.3084254	-2.80E-16	1.63E-27
vote20bs	0.264909629	0.2245958	2.43E-08
pk_cjusno,idea	NA	-1.64E-16	2.43E-08
ethnic2l dont know	-0.667522853	0.486276321	1.61E-07
asians_4	0.061086914	-0.082184165	1.61E-07
maugarepresenative	-0.575518394	1.12E-16	0.000531939
expconvert	0.021820862	0.317318853	0.000721533
maugacongress person	0.981211398	-9.02E-17	0.000769959
pk_cjusch justg	-0.711551948	1.91E-16	0.002327883
turnout16b	-0.461967103	0.011231063	0.002741292
pk_cjusjudiciary committee	NA	-0.08078035	0.002741292
pk_cjusjejejek	NA	3.08E-16	0.004516397
fthaley1	0.000369319	2.06E-16	0.005124629
pk_cjusn.a.	NA	-3.18E-17	0.005124629
pk_cjuschief justice if the supreme court	-0.375687163	1.29E-16	0.005693585
mis_covid2	-0.228208253	1.71E-16	0.006793317
pk_cjussupream court	-0.112536191	9.71E-19	0.006793317
vote16	0.162491973	3.33E-16	0.007174773
pk_cjusjulia roberts chauffeur	NA	4.06E-16	0.007174773
dtleader1	0.136592708	-1.37E-16	0.008248475
pk_cjusnot sure	-0.123291677	-3.41E-16	0.008248475
relig1_11_TEXTlutheran	-1.934145808	4.71E-17	0.008637937
pk_cjuswhite house counsel	0.498992035	-0.206546696	0.008637937
ethnic2Italian American	0.848180426	-0.045646305	0.013220876
vote20turnoutjb	-1.671599283	5.51E-16	0.013220876
pk_cjuscheif justice of america	NA	0.214615082	0.013685743
maugaw	NA	1.68E-16	0.014267773
impact8	0.097212532	0.130977698	0.014321239
pk_germa dviser	1.7919683	-2.41E-16	0.014321239
maugasecretaria	NA	0.104528909	0.015530811

	coe	res	pvalue
turnout16a	-0.409022733	6.30E-16	0.0178768
pk_cjusjudicial assistant	NA	-0.143846398	0.0178768
maugaunsure	0.094756205	-4.39E-17	0.017882249
stress3	0.011041948	-1.81E-16	0.018015963
maugapresident of liberia	0.231959387	0.079671486	0.018268688
ethnic2White Caucasian	1.276204578	-8.28E-16	0.018605882
1_TEXTButtist beliefs but don鈥檛 concider mys	-1.55025527	-1.77E-16	0.019694282
pk_cjussurpreme court	-1.022491675	-1.63E-16	0.019694282
persuade	0.189319309	0.186291424	0.019914026
pk_cjuskmldfber erb e	NA	0.080984504	0.019914026
ethnic2American	0.357749119	-5.13E-17	0.020464355
pk_germprime minister	-0.179258661	6.04E-16	0.020464355
maugafdgdgdfgg	NA	3.65E-16	0.021432201
pk_cjusit manager	NA	0.028500257	0.022168969
ethnic2ldkkk	0.361674625	8.49E-16	0.023740578
asians_6	-0.000574858	-0.042976913	0.023740578
talk3	-0.046212333	1.25E-16	0.025442138
pk_cjuslieutenant governor of american samoa	-3.350718334	-0.08635088	0.025442138
pk_cjusexotic dancer	NA	2.73E-16	0.026441407
ethnic2Cubans	-0.611420363	-0.279078436	0.026953193

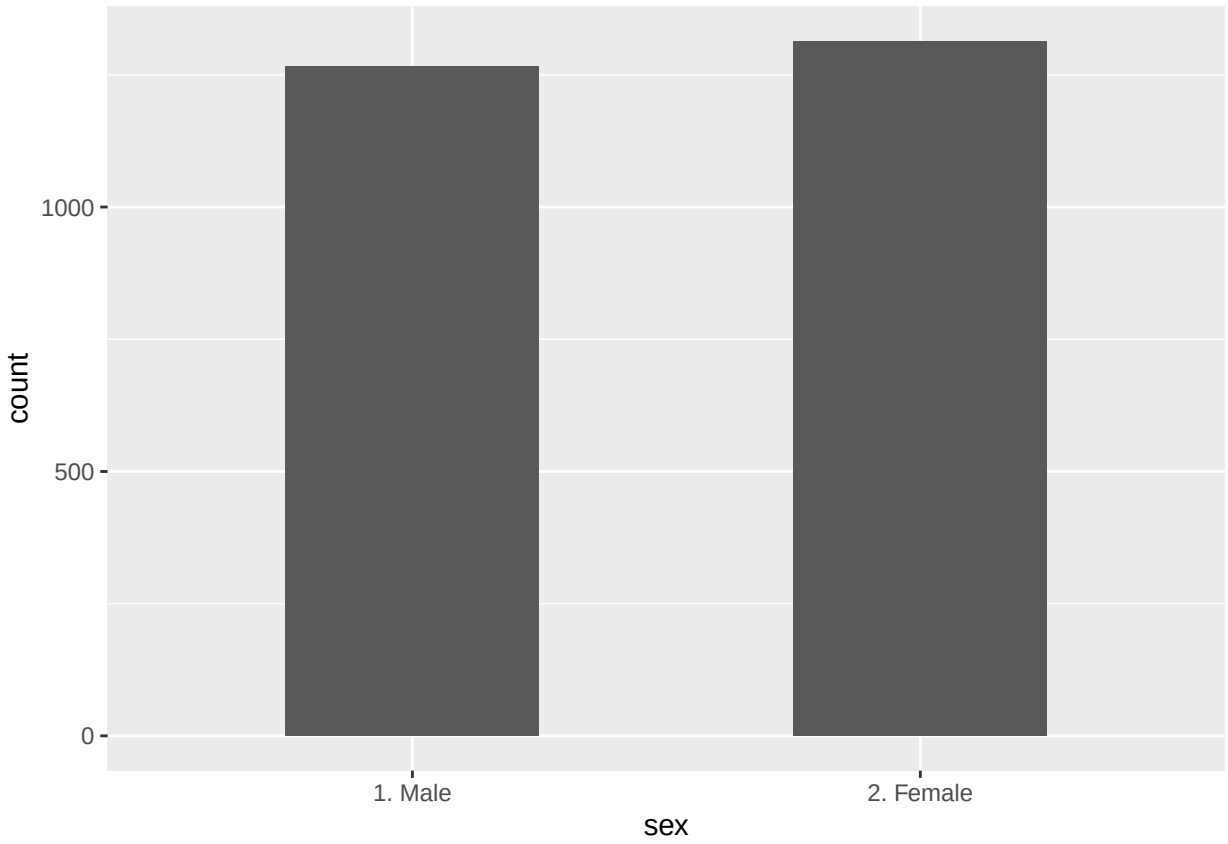
Taking a quick glance at the table of variables, I found that variables whose name start with "pk" tend to be highly significant among all the variables. For example, the variable **"pk__cjusin kongers"** has the smallest p-value of 1.63E-27, **"pk__cjusno,idea"**'s p-value of 2.43E-08 ranks the third smallest among all variables. This indicates that the voters' political knowledge affect their choices on a relatively large scale.

Besides, judging from the regression results, it was found that significant factors also lie in voters' emotions, previous voting choices, ethnics and religions. The following questions for data analysis are therefore drawn out based on these observations.

2.3: Checking whether the data is balanced

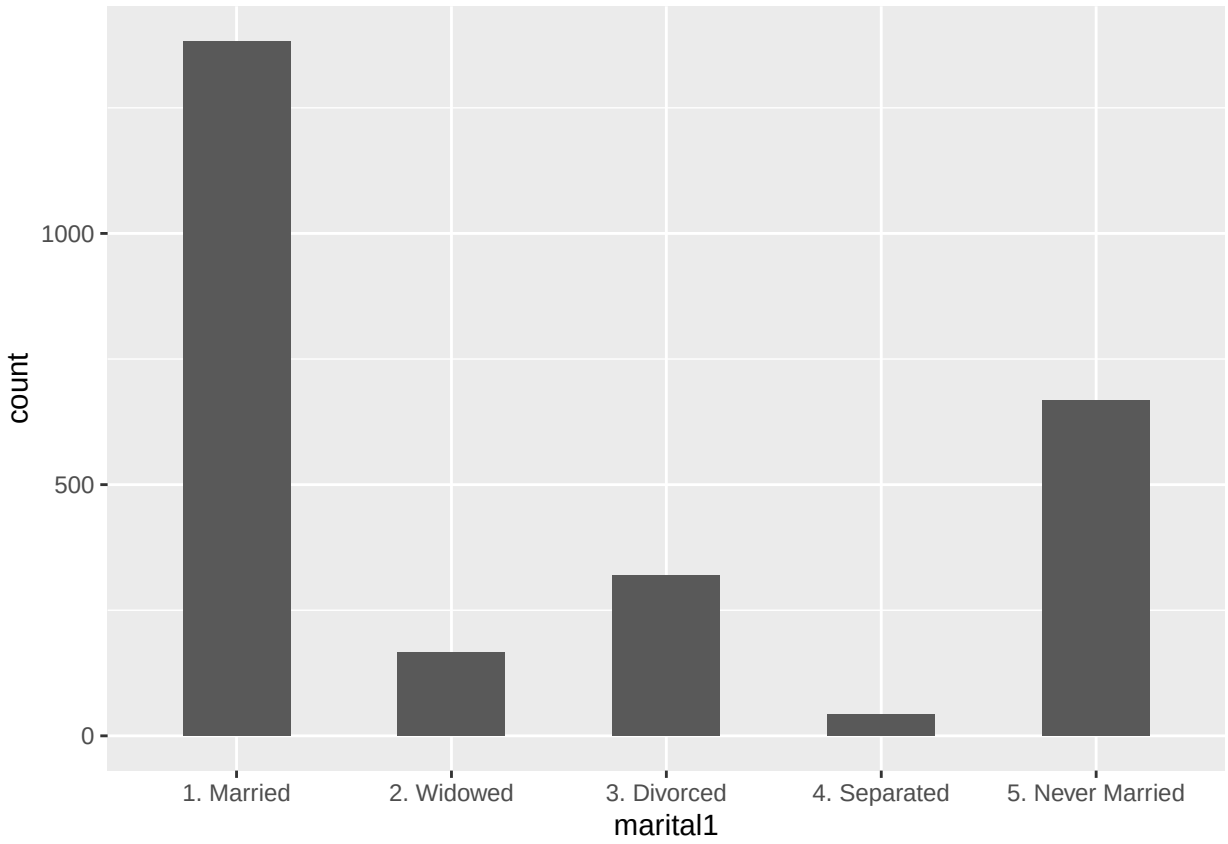
Before we go on with looking further into the significant variables we retrieved from linear regression, we should first check whether the data is distributed in a balanced scale with respect to the key variables. Therefore, let's go over the sex, education and marital variables.

sex represents result of the question: "What is your sex?" 1. Male 2. Female



The plot indicates that the sample distribution of gender is overall balanced, with the number of females going a little higher than that of the males.

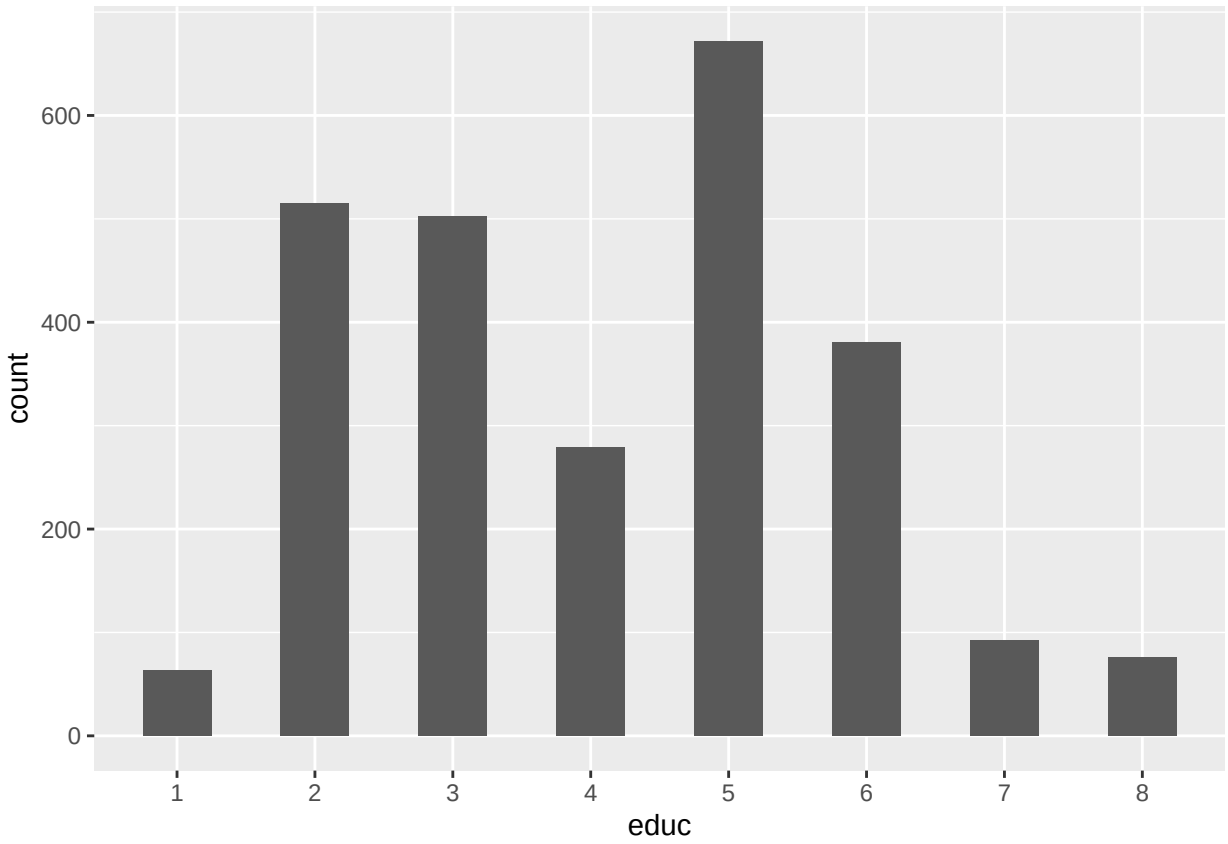
marital1 represents result of the question: "Are you now married, widowed, divorced, separated, or never married?" 1. Married 2. Widowed 3 Divorced 4. Separated 5. Never Married




```
##       1. Married      2. Widowed     3. Divorced     4. Separated
##             1383             166             320               43
## 5. Never Married      9. Missing
##              668               0
```

The plot indicates that the a large proportion of the sample is married, the percentage of married sample is 1383/3080=44.9%, which is biased from the marital statistics accessed here.

educ represents result of the question: "What is the highest level of school you have completed or the highest degree you have received?" 1. 12th grade or below, no high school diploma 2. High school graduate/diploma or equivalent 3. Some college but no degree 4. Associate degree 5. Bachelor's degree 6. Master's degree 7. Professional degree (e.g., MD, DDS, JD) 8. Doctorate



The plot indicates that the the highest proportion of the sample has achieved a Bachelor's degree, while a large proportion of the sample has the education level of High school diploma or attended college but didn't achieved degree. The sample's opinion and view therefore reflects the attitude of a relatively well-educated section of people.

Based on the analysis of the general information of the data set, we may come to the conclusion that the sample of the data is imperfect, having bias in certain features. Therefore, we should always keep in mind that the following analysis might have flaws in accuracy and precision. However, since the results for the 2020 election has come out, we may also compare what the data reflects with the final fact. In this way, we may also take as a reference how this section of voters reflect the actual results.

3. Data Analysis

Having taken a look at the basic distribution of the data set, we may dig further into how the two sides of supporters differ in their basic personal backgrounds and status.

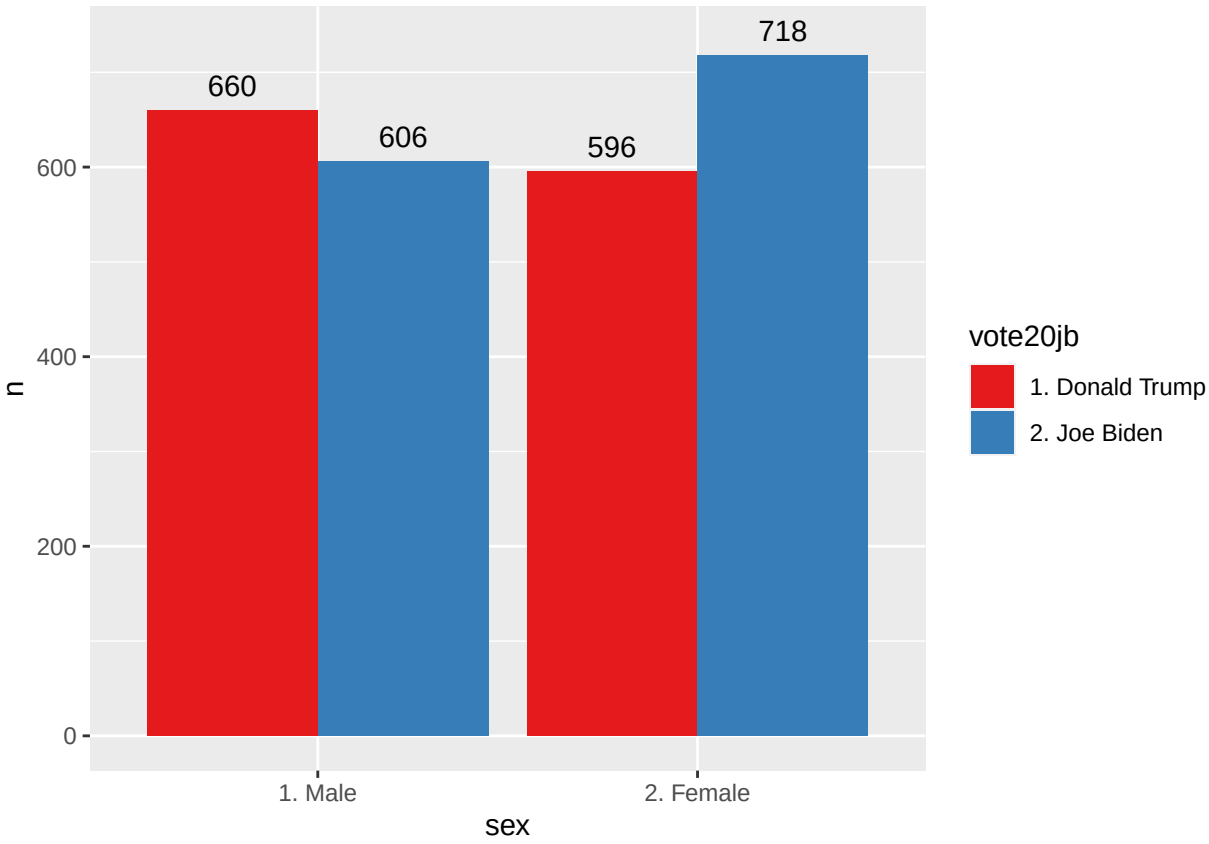
Besides, combining the results of the initial linear regression, we may analyze factors such as race, emotion, political position and knowledge.

The following data analysis part will focus on how the two sides of voters differ in their gender, marital status, educational level, emotions, race and political backgrounds.

3.1 Question2: How does gender affect voters' choice?

First let's plot out how different gender of people differ in their voting choices.

```
## `summarise()` regrouping output by 'vote20jb' (override with `.groups` argument)
```



The plot between gender and voting choice shows that more females display a supportive attitude towards Joe Biden, while the males shows the opposite tendency.

Considering the actual result that Biden has won the election, female voters might served as an important force for his success.

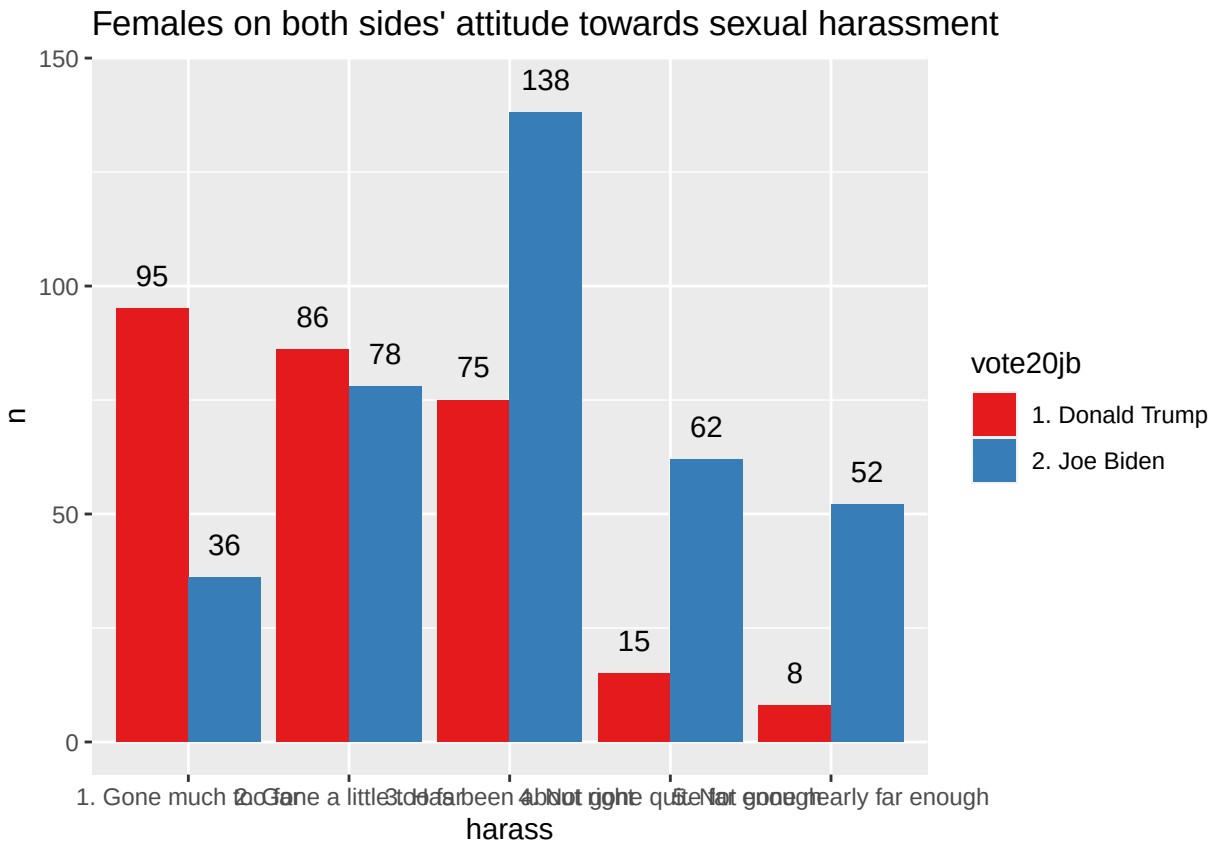
Since the questionnaire also included questions testing the voters' attitude on other political issues, we may wonder how the females on both sides hold their attitudes on the policies and the society.

###3.1.1: How females care about sexual harassment.

harass represents result of the question: "Do you think attention to sexual harassment from the #MeToo movement has gone much too far, has gone a little too far, has been about right, has not gone quite far enough, or has not gone nearly far enough?"

The answers are: –Gone much too far [1] –Gone a little too far [2] –Has been about right [3] –Not gone quite far enough [4] –Not gone nearly far enough [5]

```
## `summarise()` regrouping output by 'vote20jb' (override with `.groups` argument)
```



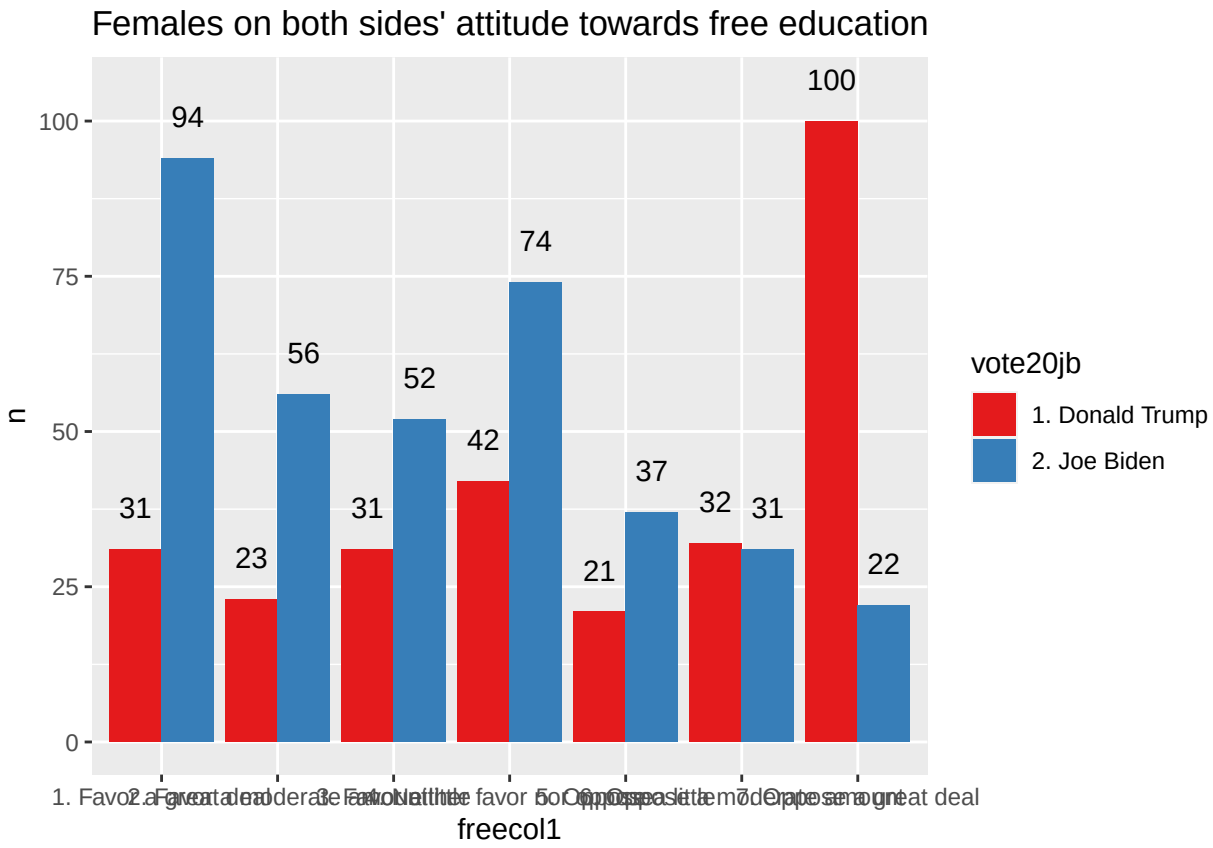
The result shows that Biden's female supporters are significantly more concerned about sexual harassment issues! In contrast, only a tiny fraction of females who support Trump show concern for this heated topic.

###3.1.2: How females care about free educaiton.

harass represents result of the question: "Do you favor, oppose, or neither favor nor oppose guaranteeing free tuition at public colleges or universities for anyone admitted? The $79 billion per year cost would be paid for with higher taxes."

The answers are: – Favor a great deal [1] – Favor a moderate amount [2] – Favor a little [3] – Neither favor nor oppose [4] – Oppose a little [5] – Oppose a moderate amount [6] – Oppose a great deal [7]

```
## `summarise()` regrouping output by 'vote20jb' (override with `.groups` argument)
```



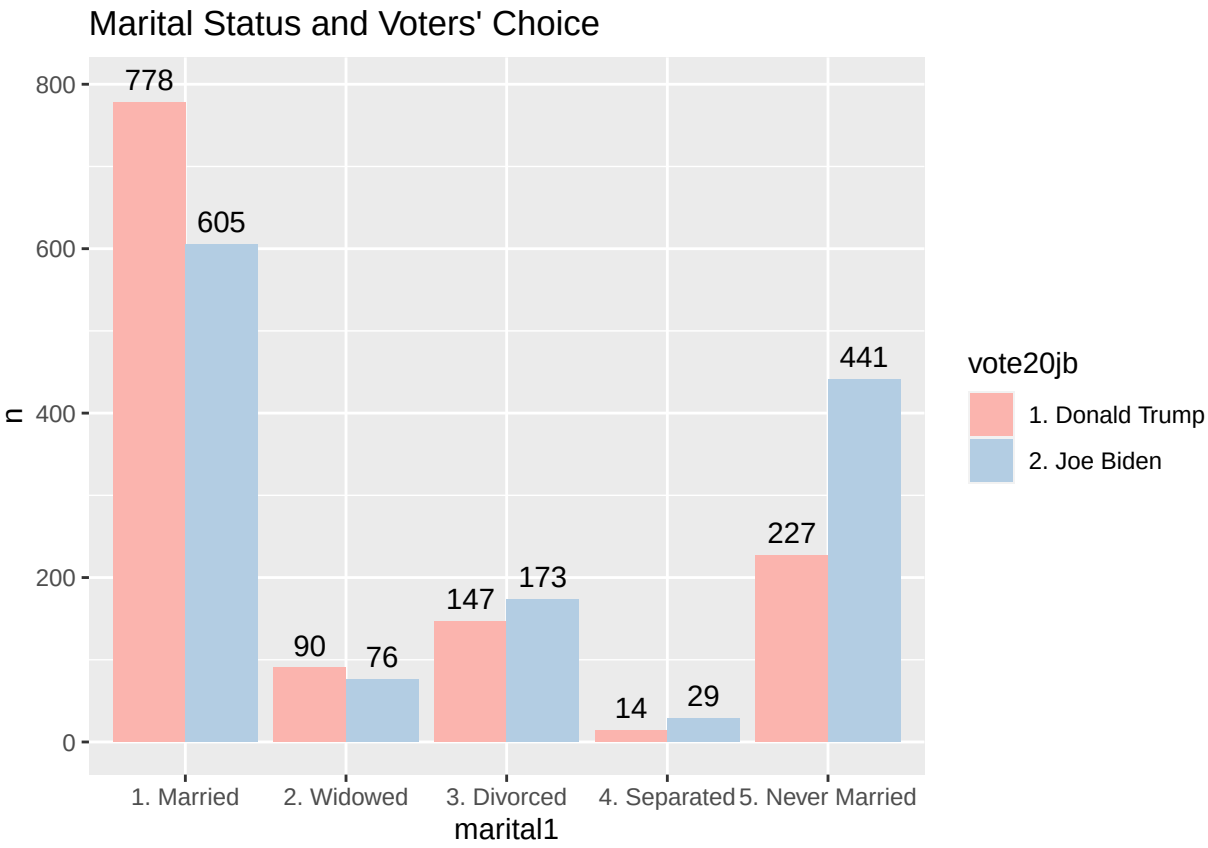
The result again shows a huge distinction between the two group of females! Trump's supporters are strongly opposed to sacrifice their taxes for free college education.

The above analysis indicates that females who support Trump seem to care more about their personal benefits, but tend to be conservative when it comes to guarding females' actual rights. This might not be helpful to Trump, since this group of people would place more attention on their personal economic benefits. Under the pandemic situation, Trump's policy undoubtedly has added to the financial loss of general public. Thus, it is likely that he would be losing support from this group of people.

3.2 Question3: How does marital status affect voters' choice?

Since a large proportion of the sample are married, let's go on to see how marital status would affect voters' voting choices.

```
## `summarise()` regrouping output by 'vote20jb' (override with `.groups` argument)
```

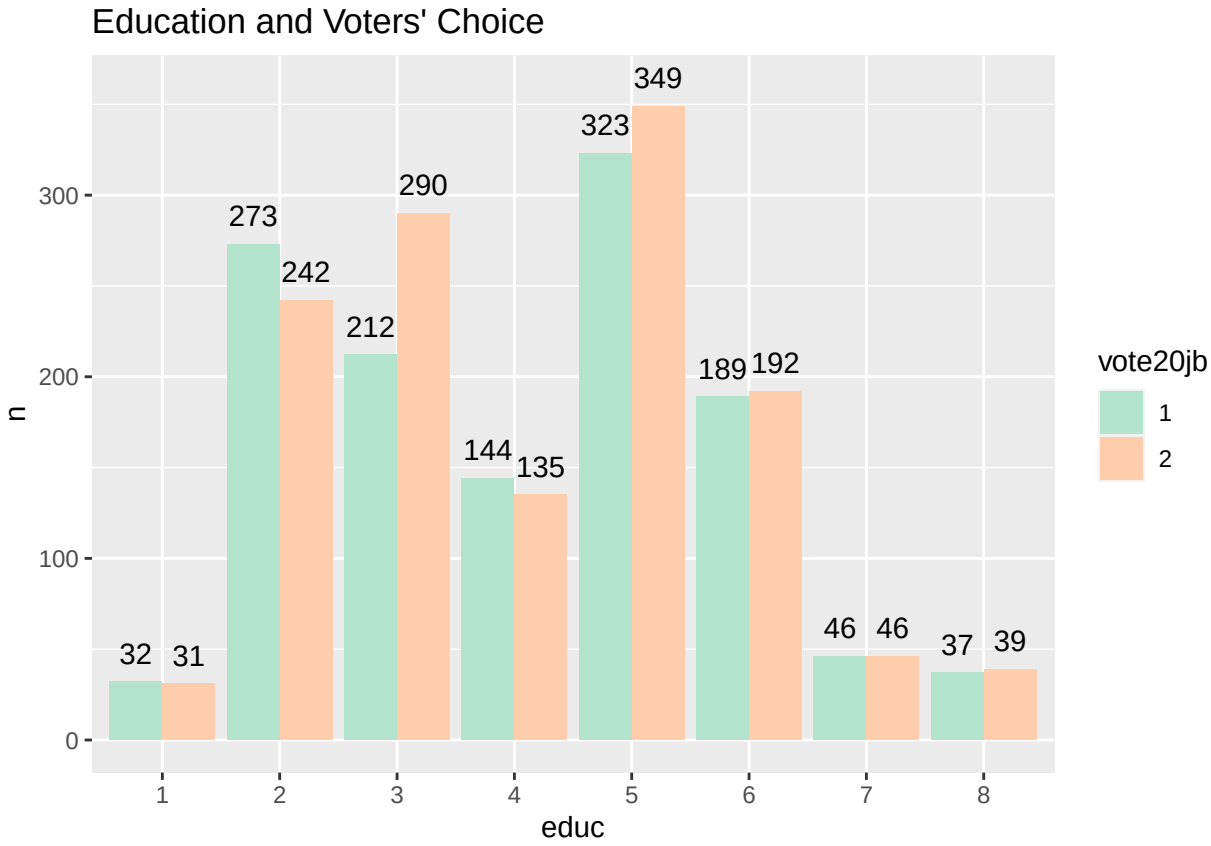


The result shows that a larger proportion of married people would choose to support Trump, while people who divorced or have never been married tend to support Joe Biden. This is explainable since Biden's supporters are younger on average.

3.3 Question4: How does educational level affect voters' choice?

The relationship between educational level and people's political stances has always been an interesting topic. Here we can also take a quick glance from this prospective.

```
## `summarise()` regrouping output by 'vote20jb' (override with `.groups` argument)
```

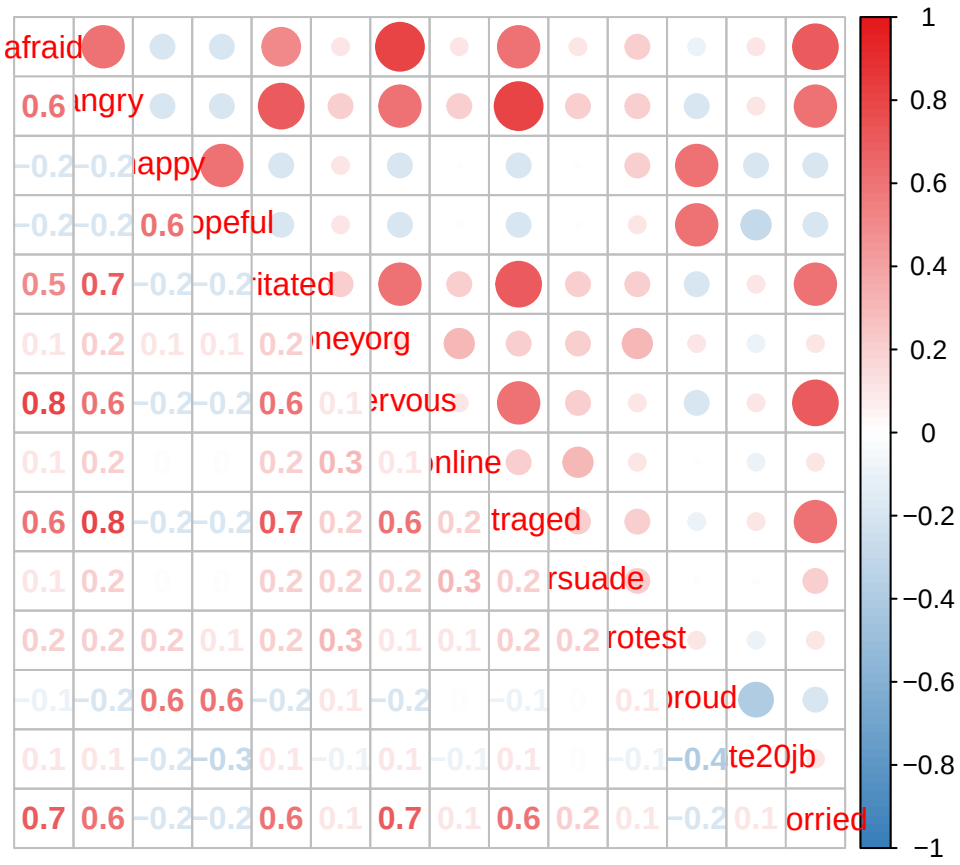


Here category "1" represents supporters for Trump, "2" for the other. We can see that both sides of voters have a relatively balanced distribution across the educational groups, with Biden's supporters outperforms a little.

3.4 Question5: What emotions are held by supporters for Trump and Biden?

Taking the fact that variables depicting voters' emotions are relatively significant in the regression result, I went on to explore what different emotions the two group of people might have.

First of all, a correlation plot is drawn to visualize how different kinds of emotions are correlated with the voters' choices.



From the plot, we can observe that the "Hopeful" and "Proud" emotions have the highest absolute value of correlation coefficients with voters' choices. To examine the relationship in detail, let's see the distribution of emotions within the two groups. The following chart is plotted based on voters' response in the question: "Generally speaking, how do you feel about the way things are going in the country these days?"

From the distribution of the emotions, we can see that Trump's supporters display an overall positive emotion, while Biden's supporters show negative emotional status. Biden's supporters tend to be more **angry, worried, afraid, irritated and outraged** towards the country's current situation, which is a reasonable phenomenon since it indicates their dissatisfaction on Trump's political "feat", thus they are more likely to vote for a new candidate of president.

3.5 Question6: How do supporters of Trump and Biden differ in their ethnicity?

The regression result shows that variables such as **'asians_4'** and **'ethnic2Italian American'** are significant to the voters' choice. Does ethnicity really contribute a lot to the variations of the voters' choices?

Let's plot out the result of the questions:"Please choose one or more races that you consider yourself to be." to test this assumption.

The plot shows difference races' different electoral choices. For instance, it is clear that a larger proportion of whites tend to vote for Donald Trump, while Joe Biden received more support from voters of other race groups.

Given the difference of both parties of voters' emotions and ethnicity, one may further consider how different race of people hold their emotions. Let's see the radar chart of the emotional score of different race of people. The emotional scores are standardized using the scale() function to avoid the effect of the size of race samples. Since the sample sizes for American Indians, Alaskan Natives and Hawaiians and Pacific Islanders are too small, we will mainly focus on the group of whites, blacks, asians and hispanics.

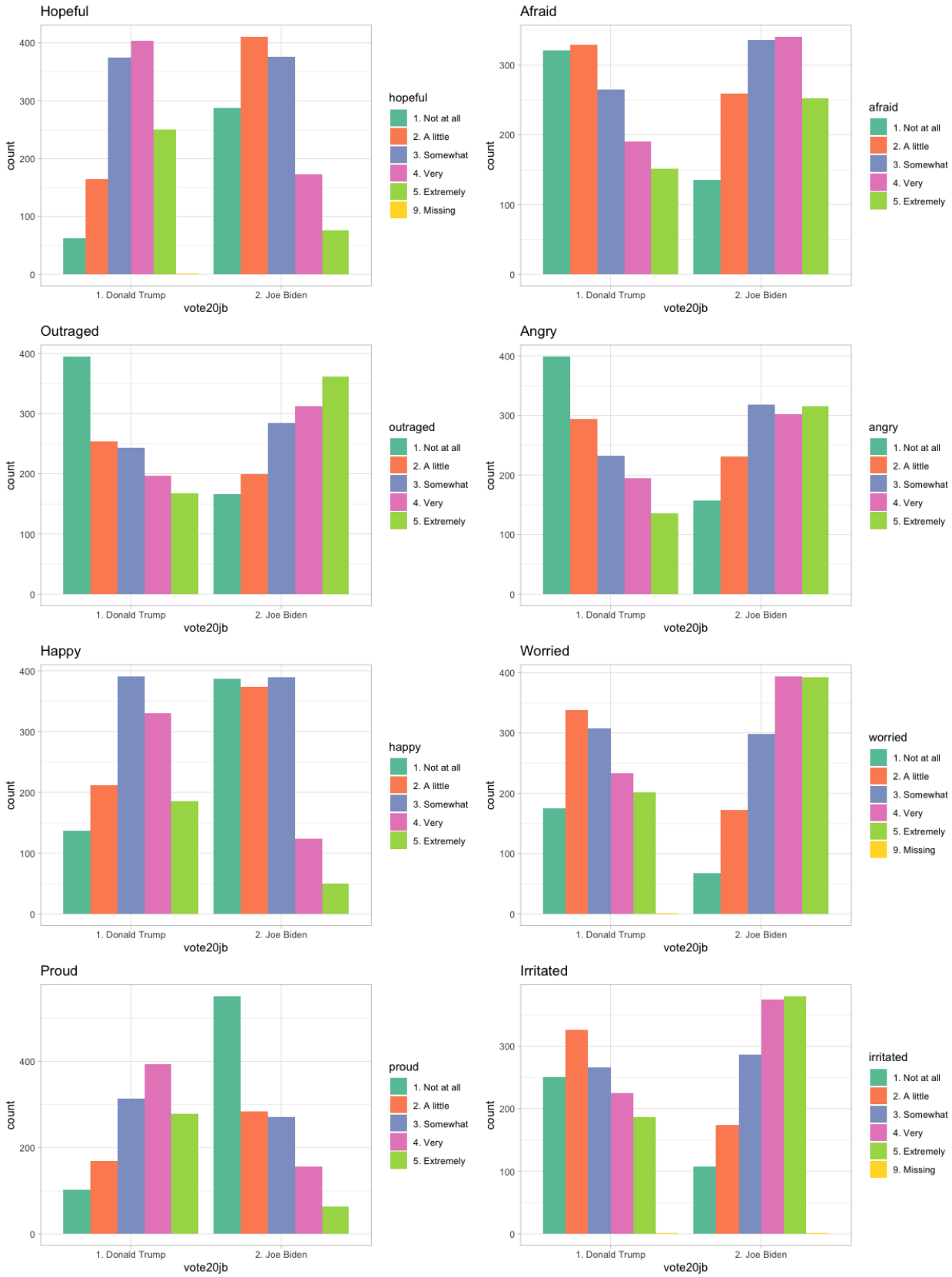


Figure 1: Distribution Plot of Emotions

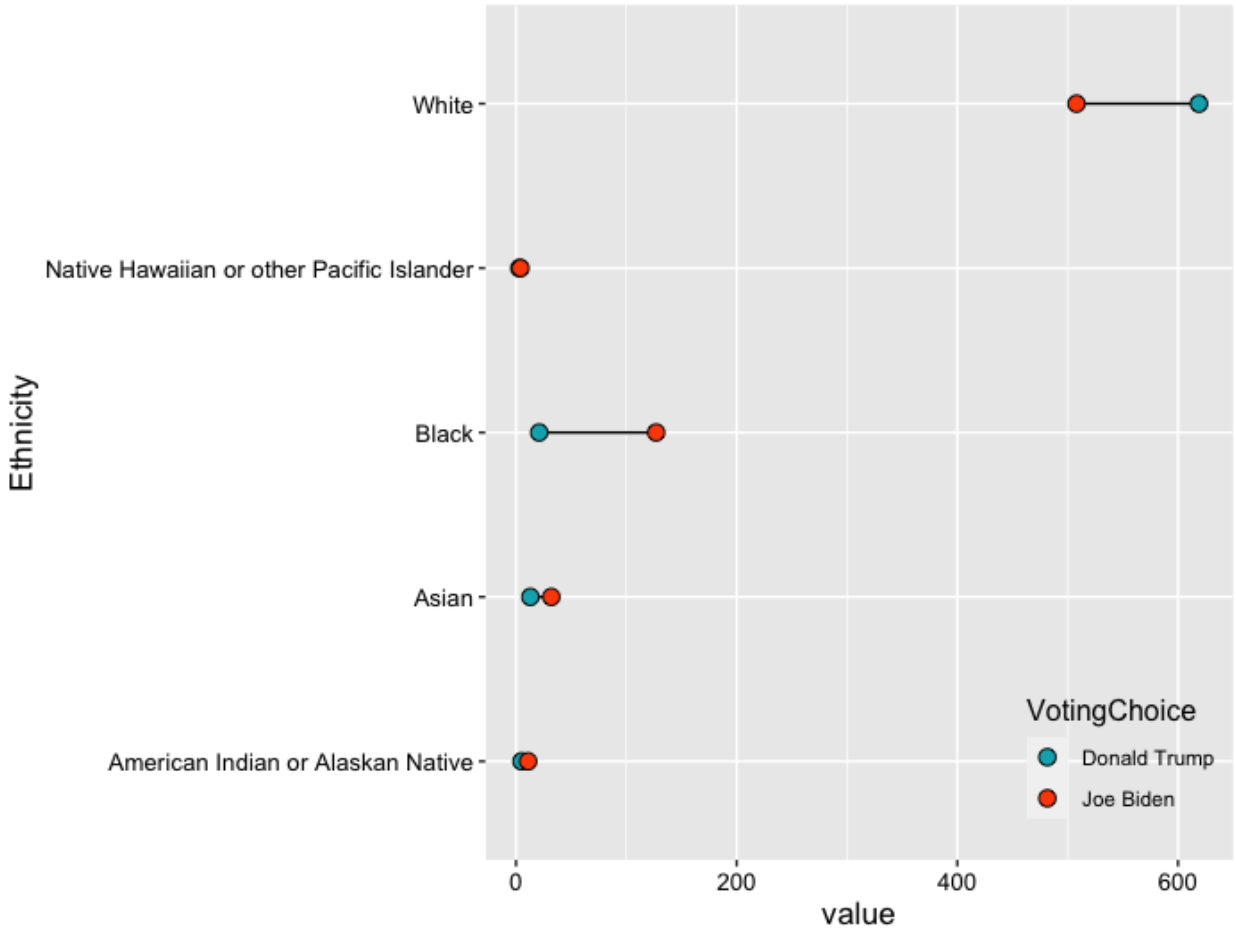


Figure 2: Race Plot of Two Gruop of Supporters

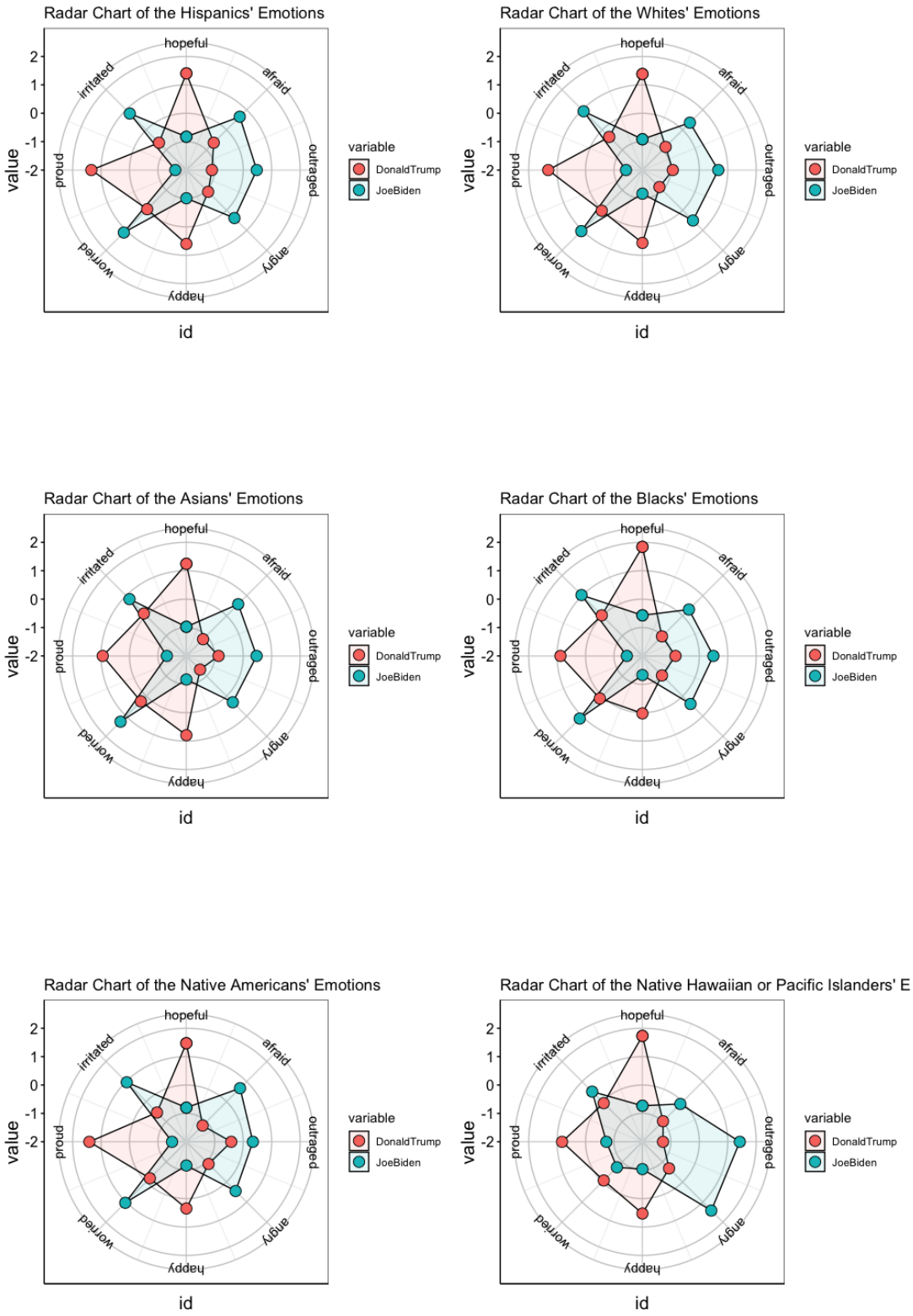


Figure 3: Emotions of Different Races

From the plot we observe some interesting facts:

1. Among Trump's supporters, the Asians shows the lowest score of being hopeful, while the blacks show the lowest score of being happy. Overall, this section of Asians and the Blacks, though supporting Trump, still display a higher level of negative emotions.

2. Among Biden's supporters, the emotions seem to be distributed in a relatively fair proportion among different races.

3.6 Question7: How do supporters of Trump and Biden differ in their voting history or other choices?

How voters act under different circumstances reveals the voters' political attitude. Therefore, by drawing a Sankey plot which shows how supporters of Trump and Biden act differently in other voting situations, some insights on how the two group of people can be retrieved.

Turnout16a represents result of the question: "In 2016, the major candidates for president were Donald Trump for the Republicans and Hillary Clinton for the Democrats. In that election, did you definitely vote, definitely not vote, or are you not completely sure whether you voted?"; – Definitely voted [1] – Definitely did not vote [2] – Not completely sure [3]

Vote20bs represents result of the question: "If the 2020 presidential election were between Donald Trump for the Republicans and Bernie Sanders for the Democrats, would you vote for Donald Trump, Bernie Sanders, someone else, or probably not vote?"; – Donald Trump [1] – Bernie Sanders [2] – Someone else [3] – Probably not vote [4]

Cvote2020 represents result of the question: "If the election for the U.S. House of Representatives were being held today, and you had to make a choice, would you be voting for the Republican candidate or the Democrat candidate in your district?" – Democrat [1] – Republican [2] – Other [3] – Won't vote [4] – Don't know [5]

From the Sankey plot, it can be seen that supporters from the two groups do show some distinctions. For example, it seems that nearly all of Trump's supporters definitely voted in the 2016 election, while more than half of Biden's supporters didn't. When faced with the selection between Sanders and Trump, a small fraction of Biden's supporters would rather choose not to vote for someone else, but Trump's supporters are more firm-minded.

3.7 Question8: What do supporters of Trump and Biden differ in their economic circumstances?

Many previous study have researched on the voters' economic background and attitudes have influence on their political choices. For example, Likhitha Butchireddygari put forward that Voters Who Think The Economy Is The Country's Biggest Problem Are Pretty Trumpy. That Might Not Help Him Much.. How are voters' choices correlated with their economic choices and what are other factors that contributes to the variations in economic stances? Let's take a look at some variables that can be used for analysis.

finworry represents result of the question: "So far as you and your family are concerned, how worried are you about your current financial situation?"
Thus, it can be used to evaluate the voter's economic status as a reference. To be specific, we can consider people who are "Not at all worried" about their financial situation as relatively affluent.

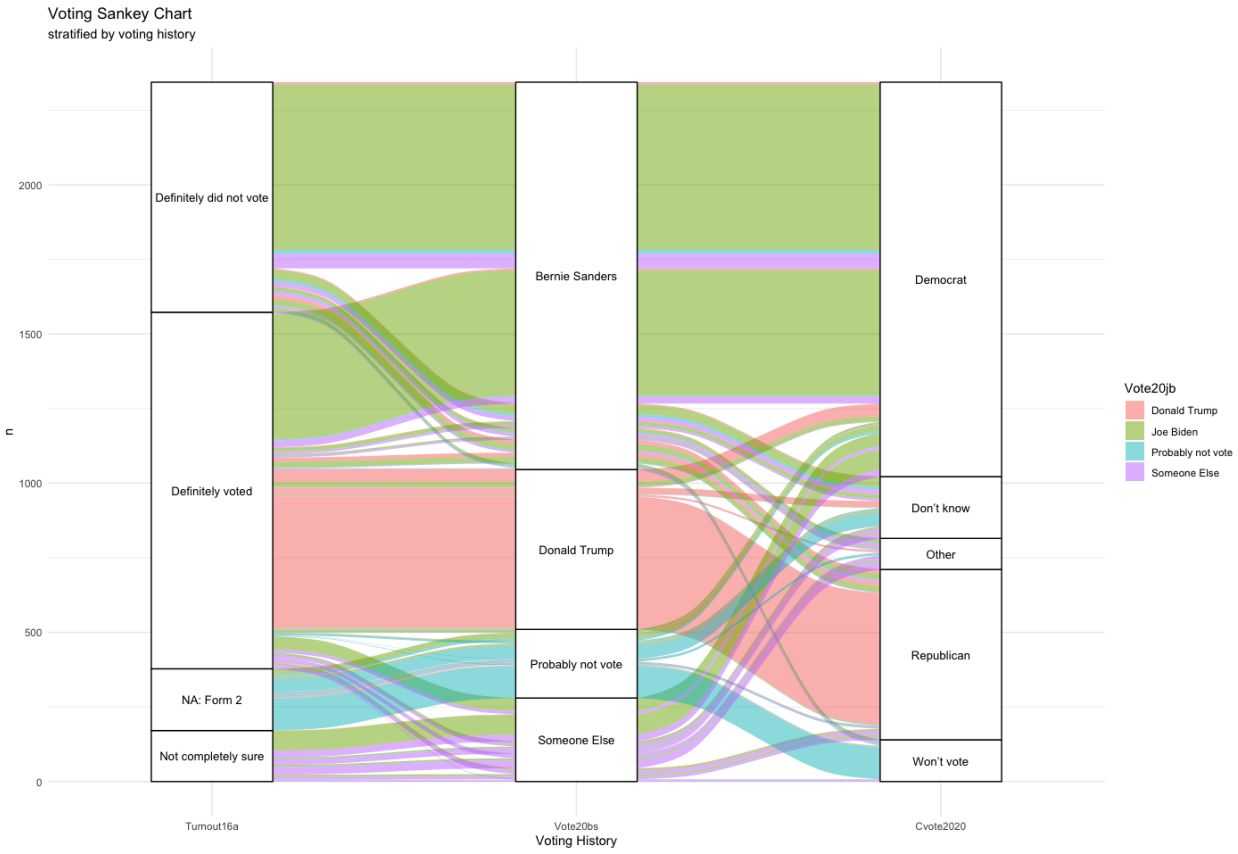
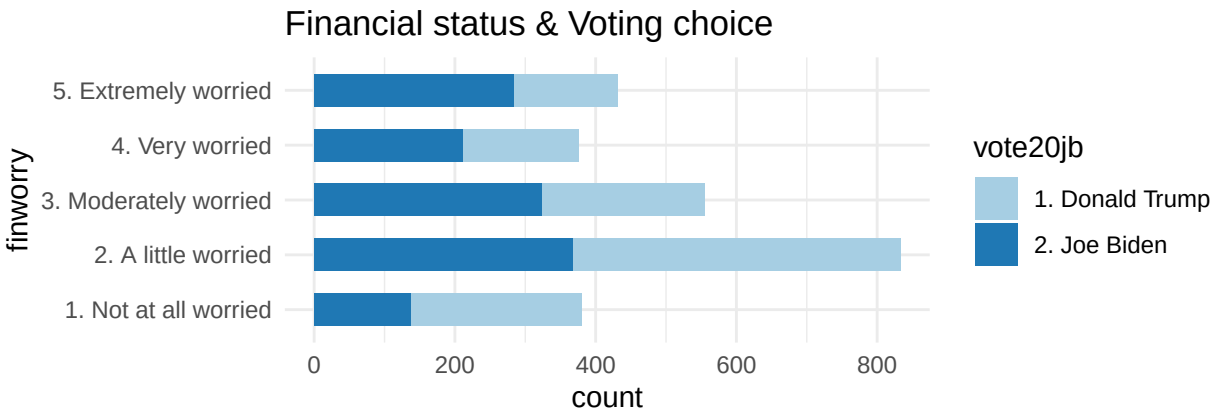
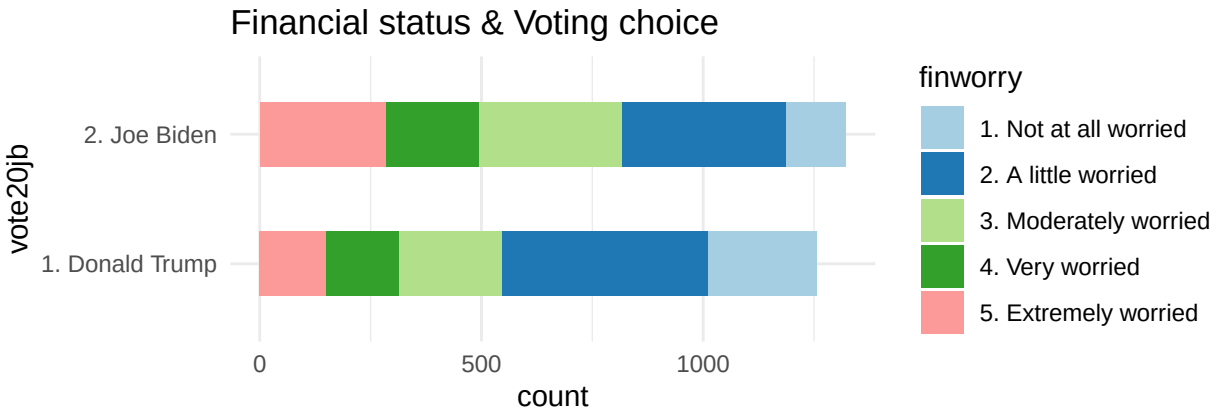


Figure 4: Sankey Plot of Voting Behaviour





If we plot out the voting choice of people from different financial backgrounds, many previous studies' assumption that one's political choice is correlated with his/her economic status can be validated. It is obvious that a greater proportion of wealthy people have a tendency towards voting for Trump, while Biden's supporters have more diverdified finantial status.

Now let's go on to see how people from different economic backgrounds differ in their political choices and attitudes.

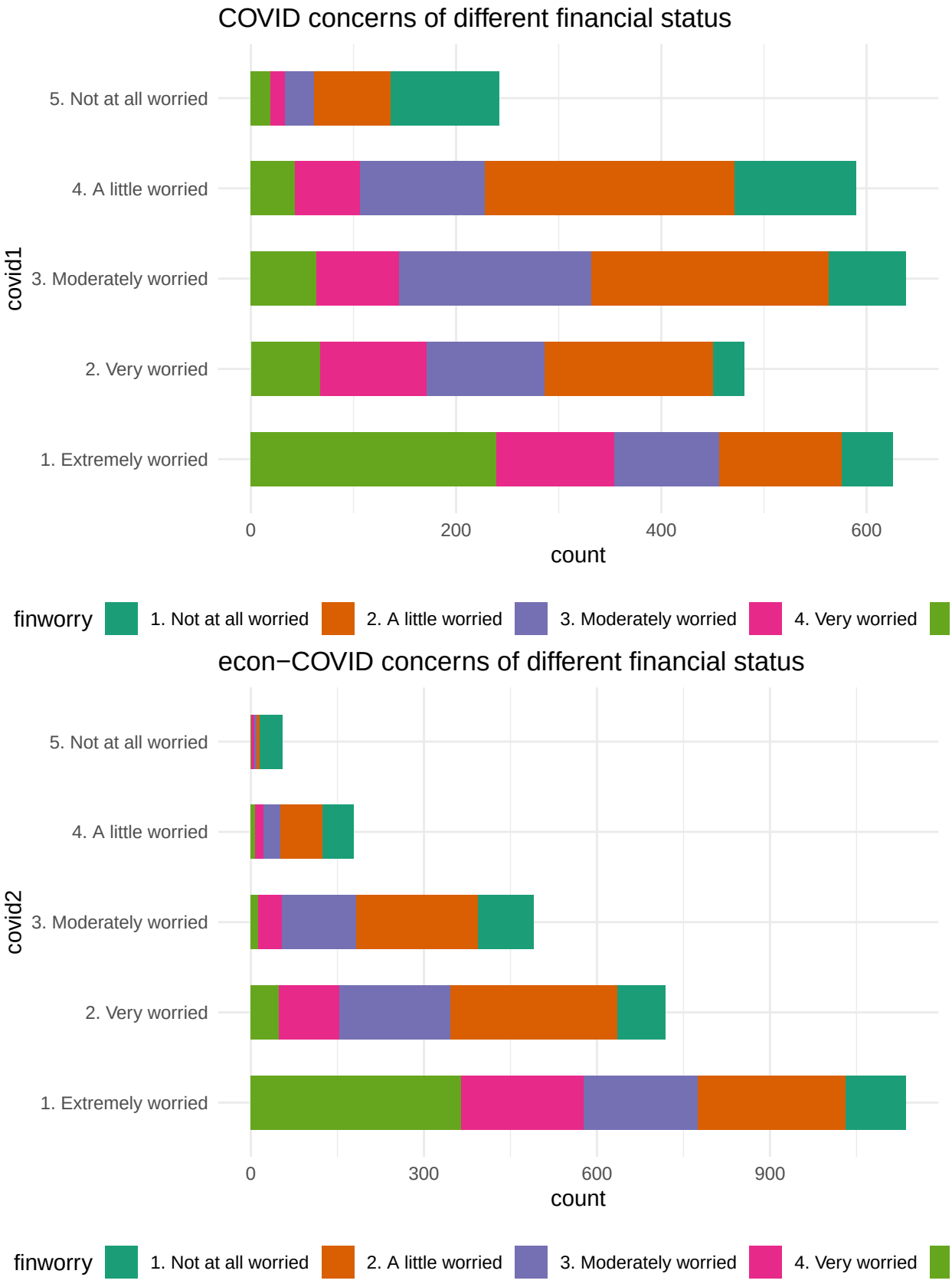
First let's take a glance at the variables that can be used for this part of analysis.

wall7 represents result of the question: "Do you favor, oppose, or neither favor nor oppose building a wall on the U.S. border with Mexico?"
Thus, this variable can evaluate a voter's attitude on Trump's immigration policies.

covid1 represents result of the question: "How worried are you personally about getting the coronavirus (COVID-19)?"
Therefore, this variable reveals the extent to which a voter is concerned about the COVID situation.

covid2 represents result of the question: "How worried are you about the economic impact of the coronavirus?"
Therefore, this variable is correlated with a voter's economic attitude on the COVID situation.

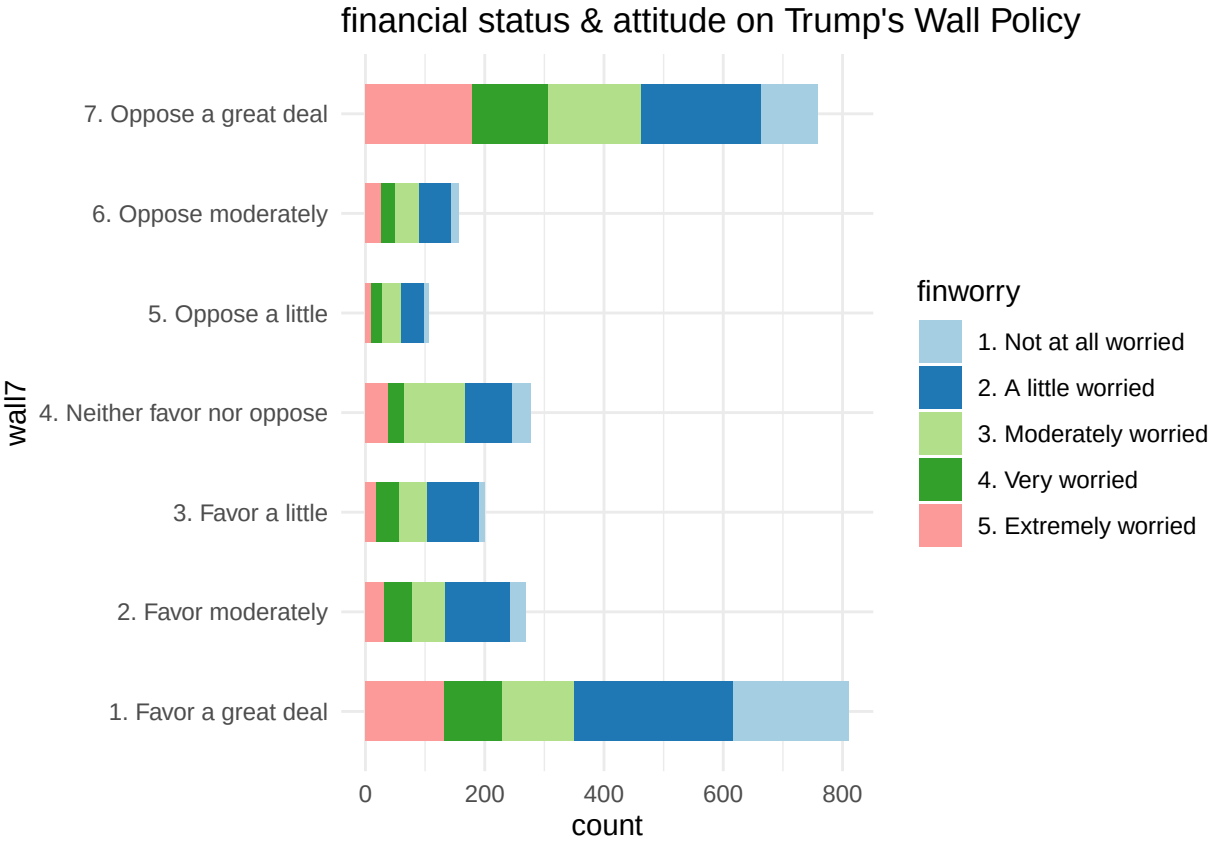
Having the above knowledge, we may go on to discover the relationship of voters' finanical status and the factors.



Now let's take a closer look on how the rich and poor differ on their concern with the COVID situation. If

we place our focus on the diagnal of the graph, we can easily discover that rich people tend to worry less about the pandemic situation, with poor people holding the opposite view.

As to the concern on the economic influence brought by COVID, we can see that people who worry the least about their financial status correspondingly have the least worry on the economic issues brought by the COVID situation.



At last, let's take a glance at how people react to Trump's immigration policies. An interesting phenomenon appears at the graph above! The sample piles at the two polar end of attitudes towards the immigration policy. Whatever financial status people are in, their attitude towards the policy of building a wall tend to be firm and absolute.

What is also worth noting is people who are in a relatively better financial status tend to be in support of the wall policy.

4. Conclusion

After conducting the data analysis above, we may identify some features of both group of voters and derive some insights into the general American public. Although due to the limitations of the data set, the data is still biased to some extent, the relations it reveals still can be referenced and do have reasonability. For example, we have found that factors such as economic status, marital status, gender, ethnicity indeed affect people's voting choices. Up till now, the election has come to an end and Joe Biden has won. With a review on the ANES Exploratory Testing Survey Questionnaire data, we did have a better understanding as to why Biden has won.