MINI PROJECT 2 REPORT

AAKASH AHUJA | AYUSH BHATIA

RESEARCH QUESTIONS:

1. How do Switch to D and Switch to R voters differ on the issue variables?
2. How do swing voters differ from loyal Democrats and loyal Republicans on the issue variables?
3. What predicts being a swing voter?

INTRODUCTION:

We require survey data for years 2016 and 2018 to perform the analysis. The survey data was collected by 'thinktank Data for Progress' which represents the number of people registered to vote in 2018 midterm elections. There are a number of variables in the survey data, but we are concerned with the study of swing voters only. The voters are categorized using the following groups:

- **Loyal Democrats**: People who voted for Hillary Clinton in 2016 and a Democratic House candidate in 2018.
- **Loyal Republicans**: People who voted for Donald Trump in 2016 and a Republican House.
- **Swing voters**: All other people who voted in 2018.

Further, there are two subsets of Swing voters as follows:

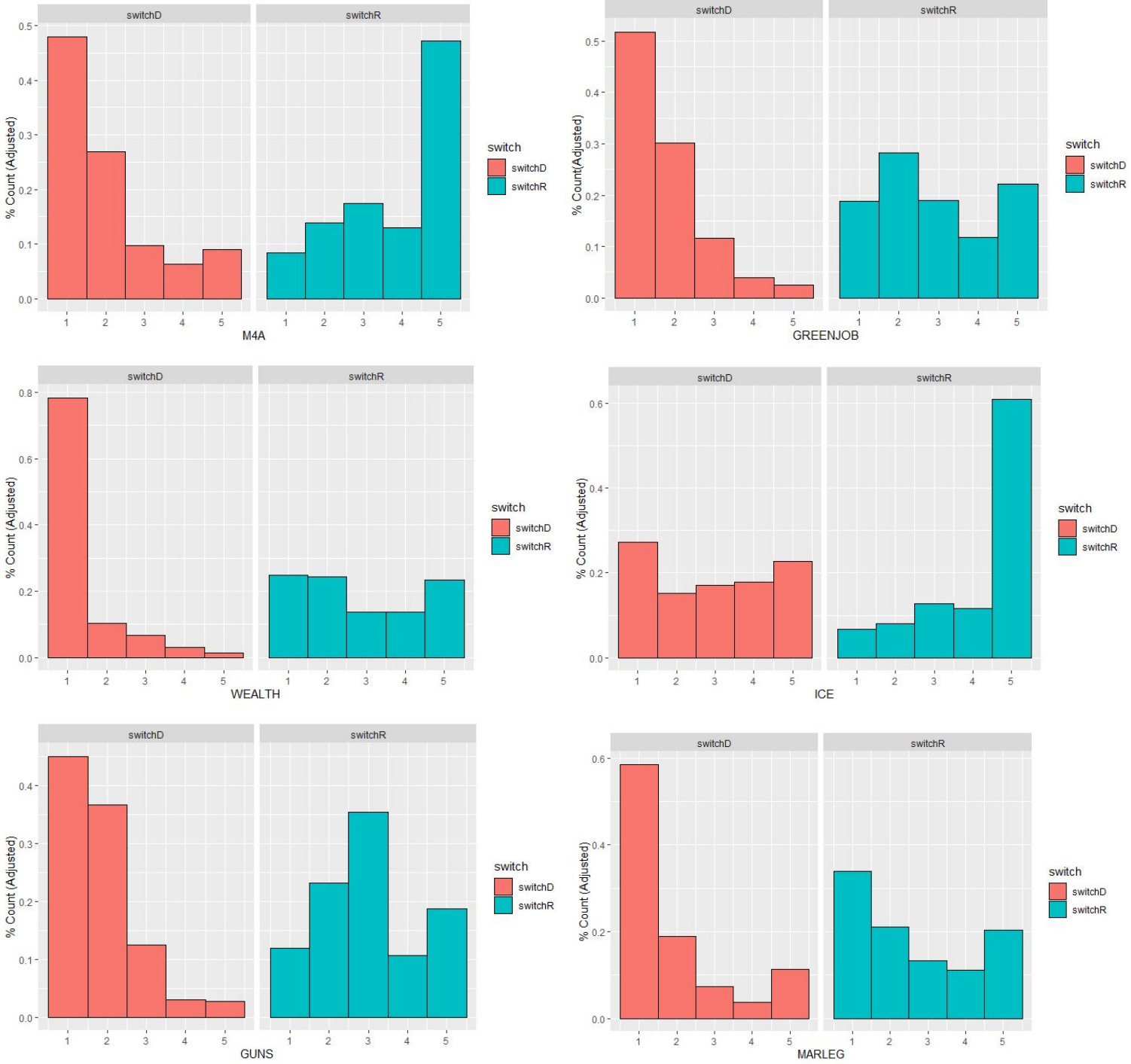
- **Switch to D**: People who didn't vote for Hillary Clinton in 2016 but voted for a Democratic House candidate in 2018.
- **Switch to R**: People who didn't vote for Donald Trump in 2016 but voted for a Republican House candidate in 2018.

RESEARCH QUESTION 1:

For the first research question, we need to study the differences of swing voters on the following issue variables:

- M4A: Medicare for All
- GREENJOB: A Green Jobs program
- WEALTH: A tax on wealth over $100 million
- MARLEG: Legalizing marijuana
- ICE: Defunding Immigration and Customs Enforcement
- GUNS: Gun control

The above variables were measured on a scale of 1-5, where 1 means that the respondent "strongly supports" that issue variable and 5 represents "Strongly opposes". The variable with value of 6 denotes "Not Sure".
We need to study only "Switch to D" and "Switch to R" voters. So we created these two subsets in the form of data frames from the swing voters data. While plotting, we eradicated the null values for that particular attribute and removed the voters who were "Not sure" - 6. The following are the ggplots for each of the Issue variable faceted by "Voter type" :



M4A

The “switchD” and “switchR” seems to follow entirely different distributions One of them is right-skewed whereas the other is left-skewed for the ‘M4A’ issue variable. So, it can be said that “switchD” voters mostly supported M4A, whereas “switchR” voters mostly opposed it.

GREENJOB

Here, the “switchD” voters follow a right-skewed distribution in terms of the ‘GREENJOB’ variable and mostly support it. On the other hand, the “switchR” voters have a mixed opinion about ‘GREENJOB’.

WEALTH

Again, it is observed that “switchD” voters follow a right-skewed distribution in terms of the ‘WEALTH’ variable and mostly support it. On the other hand, the “switchR” voters have a mixed opinion about ‘WEALTH’

ICE

In contrast to the previous two variables, here, “switchD” voters have mixed opinions about the variable “ICE”. Although, “switchR” seems to follow a left-skewed distribution, with most of the voters opposing it.

GUNS

Here, “switchD” voters follow a right-skewed distribution, with most of the respondents supporting the issue of “GUNS”. “switchR” voters mostly have a moderate opinion about “GUNS”.

MARLEG

For the “MARLEG” issue variable, both the swing voters follow a similar distribution, which is sort of right-skewed. And, it can be said that both the swing voters support the issue of “MARLEG”.

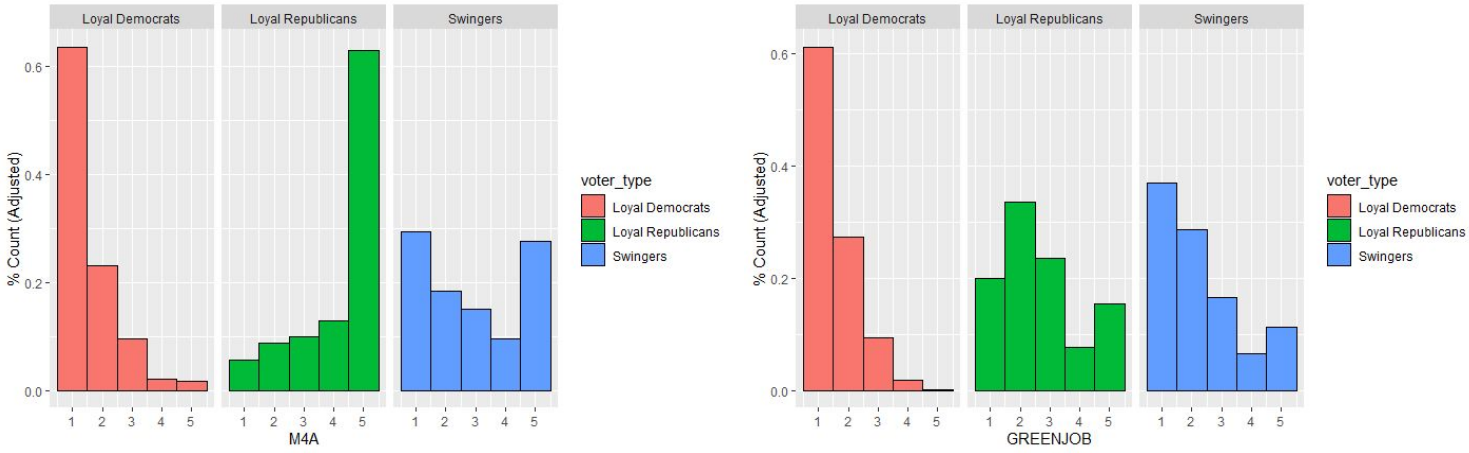
To sum it up, the switchD and switchR voters differ a lot on the issue variable “M4A” and seem to be reasonably similar on the issue variable “MARLEG”.

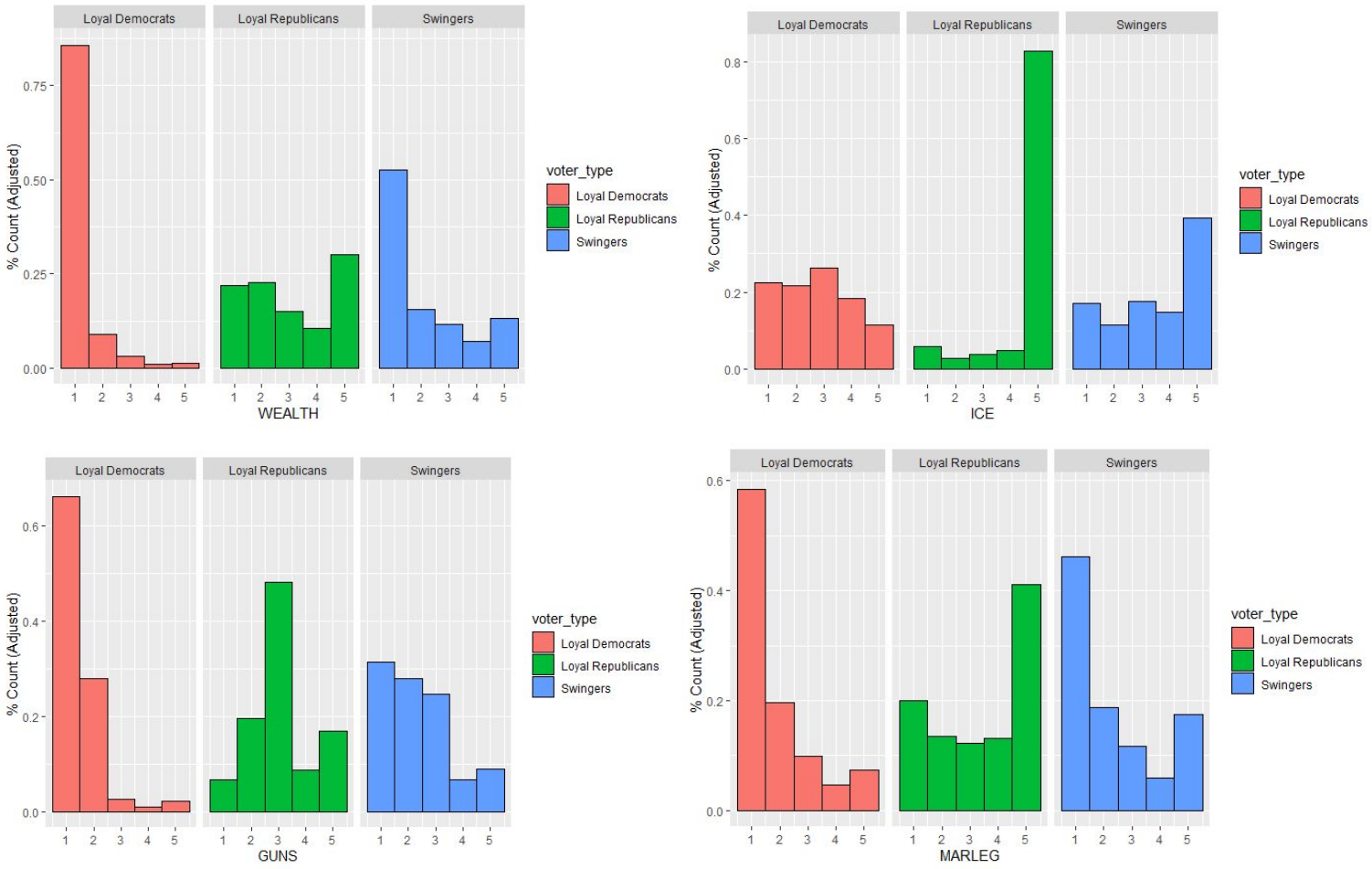
RESEARCH QUESTION 2:

In this research question, we see the responses of all the swing voters(including “switchD” and “switchR”) on the Issue variables and compare them with both “Loyal democrats” and “Loyal republicans”. We use the following hypothesis for the comparison purposes:

Hypothesis 1: Swing voters are moderates, and tend to the in the middle of the distribution when Democrats are on one side and Republicans are on the other.
Hypothesis 2: On most issues, swing voters are split, with some of them acting more like Democrats and others acting more like Republicans.
Hypothesis 3: Swing voters think more like Democrats on some issues and more like Republicans on other issues.
Hypothesis 4: Swing voters are ideologically incoherent and don't have consistent patterns in their issue positions.

We created three subsets - “Loyal Democrats”, “Loyal Republicans” and “Swing Voters” for all the Issue variables. Then, we plotted the combined dataset for each issue variable faceted by the “voter type”, similar to the previous research question, as follows:





M4A

For the issue variable “M4A”, some of the voters act more like Democrats(supporting M4A) and some of the voters more like Republicans(opposing). Hence, the swing voters can be described “split” for this case.

GREENJOB

In this issue variable, it can be observed that swing voters act more like Democrats and provides support for the issue variable “GREENJOB”

WEALTH

Again, it is evident that the swingers follow a similar right-skewed distribution of Loyal Democrats and supports the issue variable “WEALTH”

ICE

Here, the swing voters act more like the Loyal Republicans and follow a left-skewed distribution, with most of the swingers opposing the issue variable “ICE”.

GUNS

In the issue variable “GUNS”, the swingers follow a right-skewed distribution similar to the Loyal Democrats and most of the swingers support the issue variable.

MARLEG

Again, in the above graph, it is evident that the swingers follow a right-skewed distribution similar to Loyal Democrats and also, most of the swingers support the “MARLEG” issue variable.

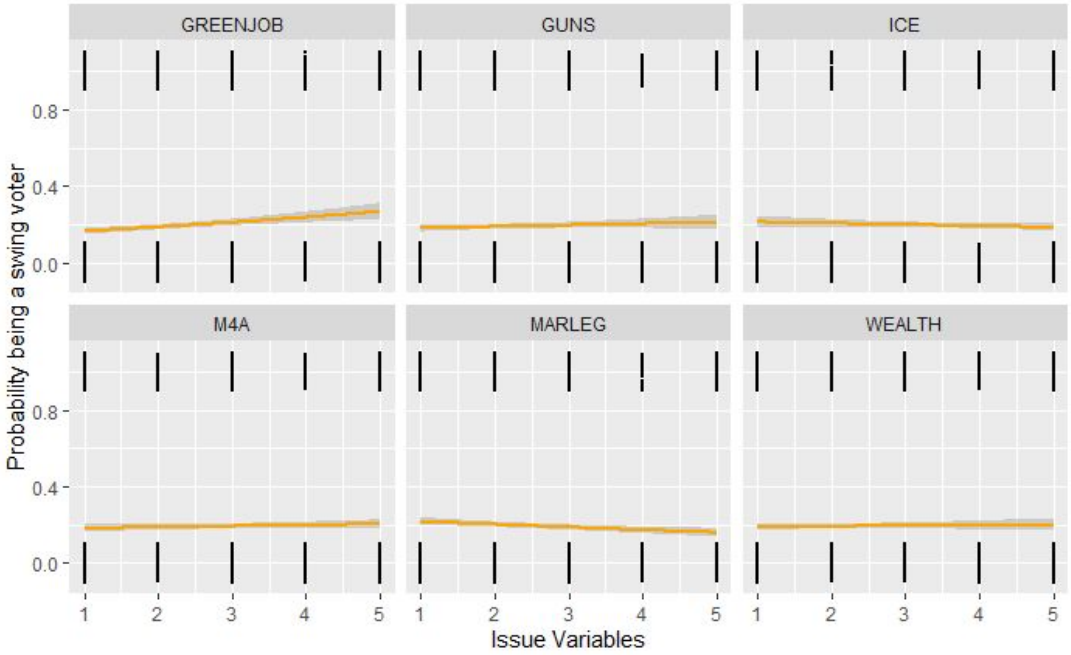
Hence, after visualizing all the plots, it can be said that the data is best described by the Hypothesis 3 - (Swing voters

think more like Democrats on some issues and more like Republicans on other issues) on all the issue variables except M4A.

RESEARCH QUESTION 3:

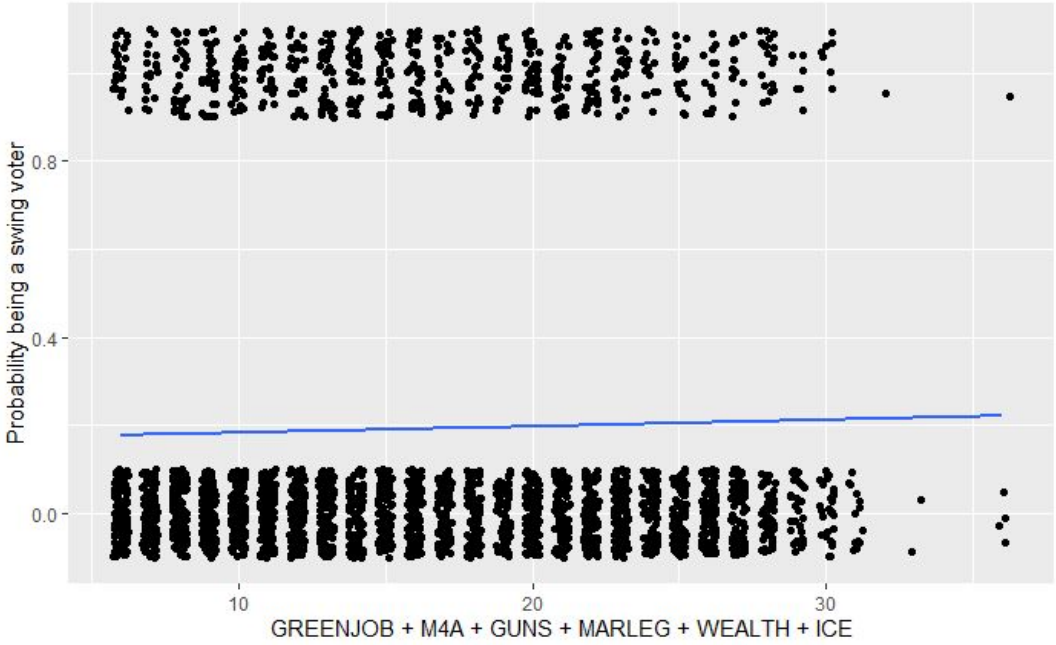
Model 1:Predict the probability of being a swing voter using only the "issue" variables as predictors

To build this model, we perform "Logistic Regression" and create a new column named "swing_voter" which stores the value 0 and 1 for "Not a swing voter" and "swing voter" respectively. This column acts as a target variable, so we checked its correlation with all the issue variables. The issue variable 'GREENJOB' has the maximum correlation of 0.086. Then we plotted the probability of being a swing voter faceted by each issue variable:



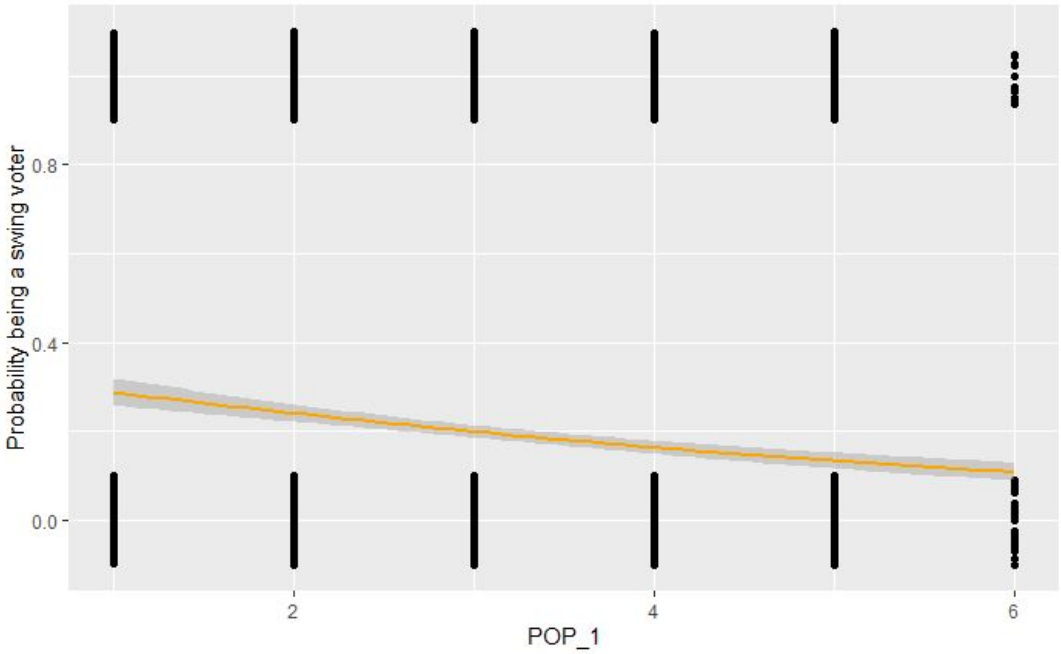
And now, looking at the slopes it is evident that 'GREENJOB' is the most important issue variable for predicting the 'Swing Voter'. For other issue variables, the probability seems to remain constant as the slope of the line is almost zero and hece, these variables do not act like good predictors.

Then, we combined all the issue variables to create a model and it can be visualized as follows:



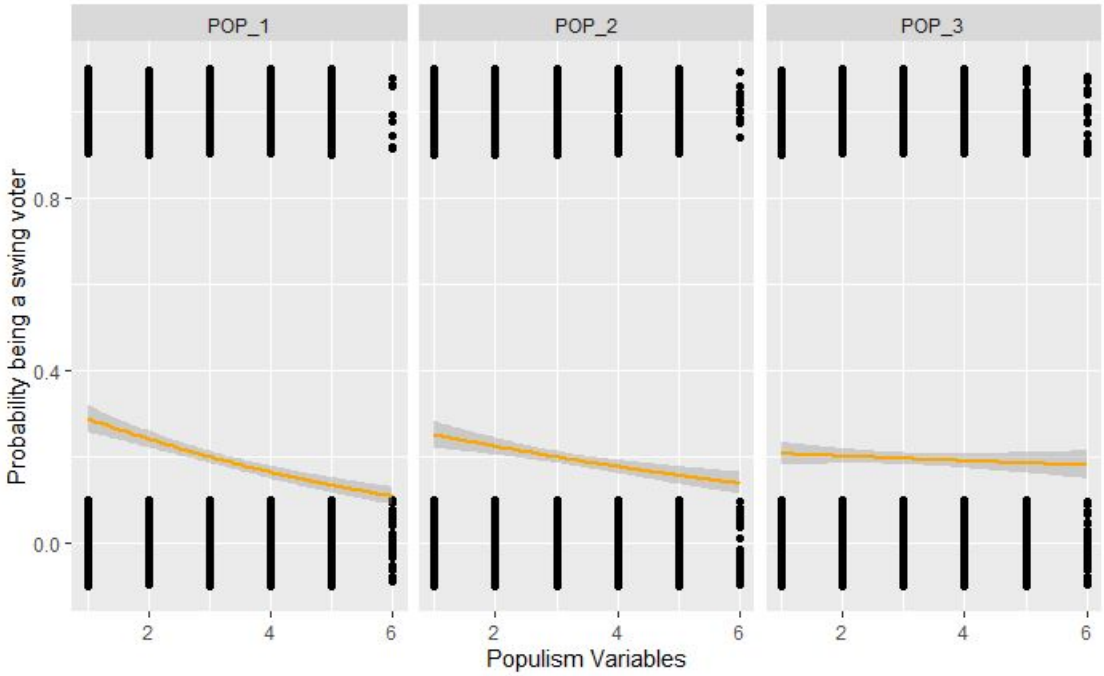
Model 2:Predict the probability of being a swing voter using only the "populism" variables as predictors

We follow the same approach to create logistic regression models for the data with the "populism" variables. When checking the correlation of all the populism variables with the target variable, it can be seen that the variable "POP_1" has the highest negative correlation of 0.13. So initially, we created the model with just the "POP_1" variable as the predictor and plotted it as follows:



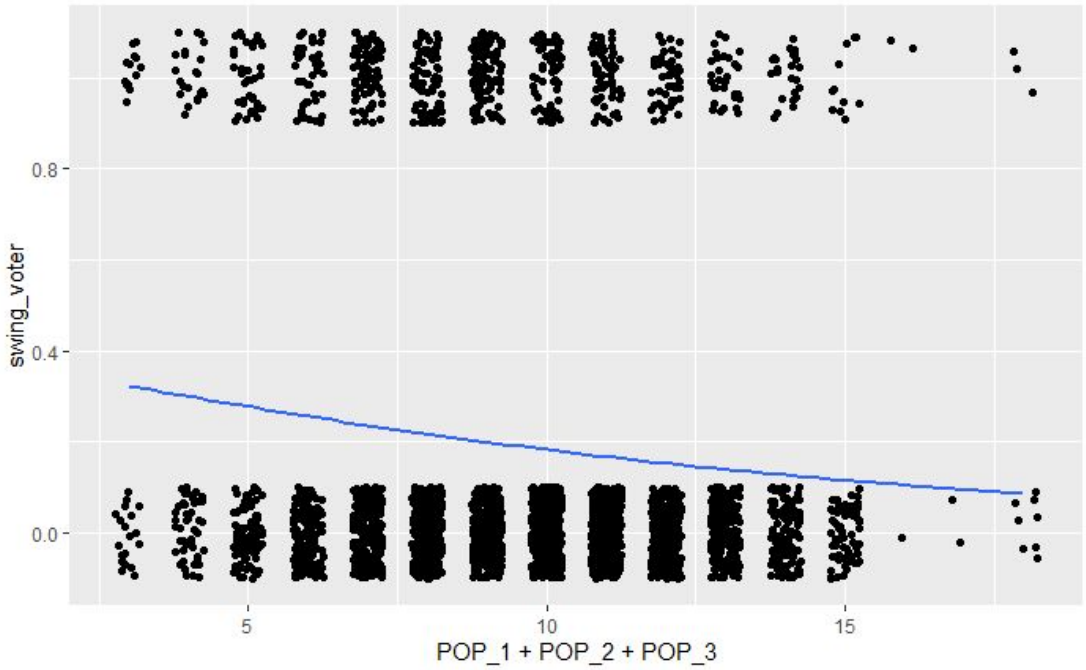
The plot shows the negative slope of the line.

Then, we checked the probability of being a swing voter with respect to the individual populism variables and plotted it as follows:



It can be seen that only the variables "POP_1" and "POP_2" have a negative slope and seem to be correlated to the target variable. The variable "POP_3" has zero slope and hence, no correlation with the target variable.

Then, we created a logistic regression model by combining all the “populism” variables and the trend for the predicted probability can be visualized as follows:



Comparing both the models:

The Model 1, using the issue variables, seems to be inefficient in predicting the probability of being a swing voter as the slope of the fitted line is close to zero.
Whereas, the Model 2, using the populism variables, seems to be more efficient in predicting the swing voter as the slope of the line is more than the one compared to Model 1

Model 2 seems to do a better job as compared to Model 1. This is due to the fact that 2 of the 3 variables in Model 2 are more correlated to the target variable. Whereas, in Model 1, there is only one variable which has a high correlation with the target variable.

There are multiple factors that could be relevant in determining what makes a voter a swing voter. Considering the facet plot of Model 1, it is observed that only the “GREENJOB” variable is doing a good job in predicting the swing voter. While considering the facet plot of Model 2, there are two variables - “POP_1” and “POP_2” that are helpful in predicting the probability of being a swing voter. Hence there are three factors namely - “GREENJOB”, “POP_1” and “POP_2” that are important for predicting whether the voter is a swing voter or not.

CONCLUSION

After the analysis of the swing voters and looking at the three research questions, it can be said that swing voters are mostly influenced by the “GREENJOB” issue variable. This is evident in the first and second research question as well, where the plots of “GREENJOB” usually follows a left-skewed distribution, which indicates that swing voters don’t have mixed opinions on this particular issue variable.