S 670 - EXPLORATORY DATA ANALYSIS

MINI PROJECT 2

INTRODUCTION

The mini project deals with survey data collected by *thinktank* data for progress. The data represents the population of people registered to vote in the 2018 midterm elections. We create various subsets of voters from the original data to study how each subset of voters supports different programs and to study the similarities and the differences between the voters in terms of support to the various programs.

DATA

The dataset used for the analysis is *DFP_WTHH_release.csv*. This contains more than 200 columns out of which we need three sets of variables for the study. They are the three basic variables, six issue variables and three populism variables.

The three basic variables are *presvote16post*, *WEIGHT_DFP* and *house3*. The types of issue variables and populism variables will be discussed as the study proceeds.

The 5 subsets of voters are:
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 candidate in 2018.
Swing voters: All other people who voted in 2018.
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.

STUDY

- How do Switch to D and Switch to R voters differ on the issue variables?

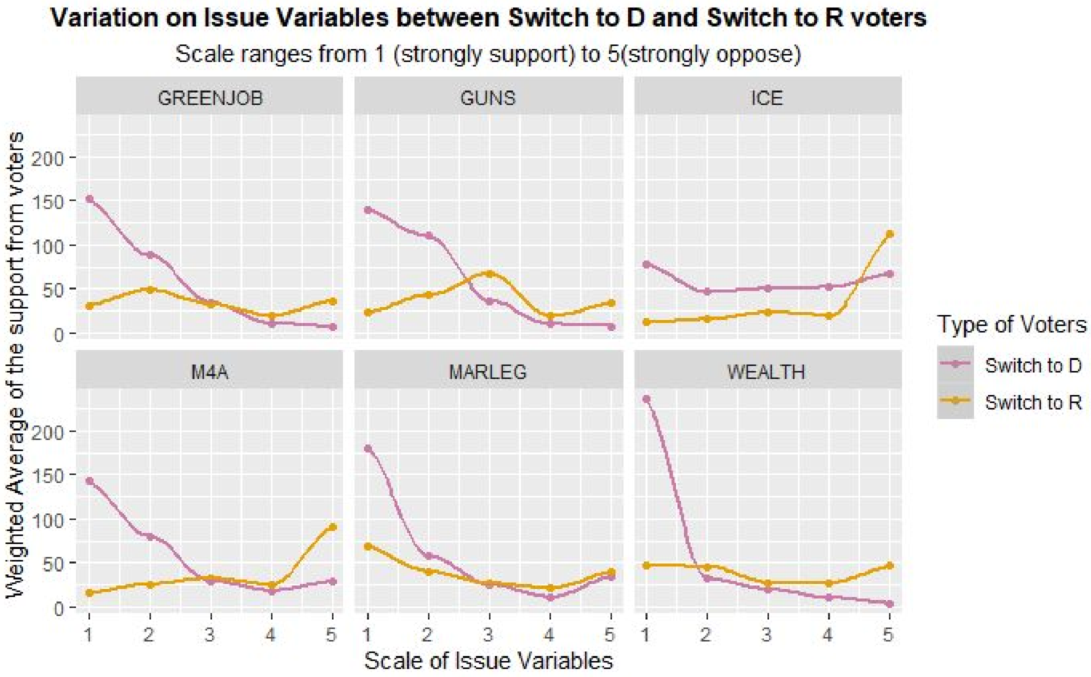

Fig 1

The above faceted plot shows how the Switch to D and Switch to R voters differ in the way they support the six programs.

GREENJOB: A Green Jobs program

Majority of the voters who switched to Democratic House extended strong support for the Green Jobs program thereby supporting the idea that every unemployed American who wants a job building energy-efficient infrastructure should be given.

The voters who switched to Republican House do not seem to show a clearer support or opposition for the Green Job program, as there is no peak throughout the scale from 1 to 5. It looks like Switch to R voters do not have a clear idea on the program and therefore the average number of voters supporting and opposing the program are almost the same.

GUNS: Gun control

The average number of voters supporting the Gun program is almost like the Green Job program. Among the voters who switched to Democratic House, the majority of them support the Gun Program. By supporting, they mean that the Gun program should make the purchase of some to all types of guns difficult. Therefore, switch to D voters support the regulation to avoid the purchase of guns and want stricter regulation on the purchase of guns.

The switch to R voters do not show any peak in the trend and some of the voters feel the existing regulation on Gun control is right while some of the voters are supporting and the rest of them are opposing the current Gun regulation. We cannot clearly conclude if the switch to R voters support or oppose this Gun program.

ICE: Defunding Immigration and Customs Enforcement

The voters who switch to Democratic House do not have a clearer opinion on the ICE program. They do not exhibit strong support or opposition to the program as the curve is almost flat throughout. The switch to D voters are neutral about defunding the Immigration and Customs Enforcement program although a very few of them show support to the program. But that does not conclude that the switch to D voters support the program

Switch to R voters seem to oppose the defunding of Immigration and Customs Enforcement as the trend slowly increases and suddenly reaches a peak for the scale of 5. Although the magnitude of the supporters is less than switch to D till scale of 4, the sudden increase in the average number of voters in opposition to the program makes us conclude that switching to R voters do not support the program.

M4A: Medicare for All

Majority of the switch to D voters are in favor of expanding Medicare such that it becomes the main health insurance provider for all Americans. Although some of the switch to D voters are against this program, the average number of voters in support of this program outweigh those who are against the program. Thus, we can safely conclude the switch to D voters support M4A.

The average number of switch to R voters who support the M4A program are way lesser than the average number of voters who are opposing this program. Thus, we can say that switching to R voters have an opposing opinion on the expansion of Medicare.

MARLEG: Legalizing marijuana

The average support from the switch to D voters for legalizing Marijuana in the nation is far greater than those who oppose the program. Thus, we can conclude that switching to D voters are in favor of legalizing marijuana.

A similar kind of trend is observed in switch to R voters as well although the magnitude of support and opposition is lesser compared to switch to D for every scale from 1 to 5.

WEALTH: A tax on wealth over $100 million

Voters who switched to Democratic House strongly support imposing wealth tax on wealth over $100 million whereas the voters who switched to Republican House do not show any clear opinion on the Wealth program as the curve lay flat throughout the scale.

To conclude, we can say that voters who switched to Democratic House and voters who switched to Republican House share the same opinion on Legalizing Marijuana in the nation. They do not share similar opinions in other programs. While looking at only those who switched to Democratic House, we can see that they have a clear idea on all programs if they support to oppose except for the ICE program where they have a neutral opinion about it. While looking at only those who switched to Republican House, we can see that they show neutral interest in Green Job, Guns and Wealth programs, while they strongly oppose Medicare and ICE programs.

□ How do swing voters differ from loyal Democrats and loyal Republicans on the issue variables?

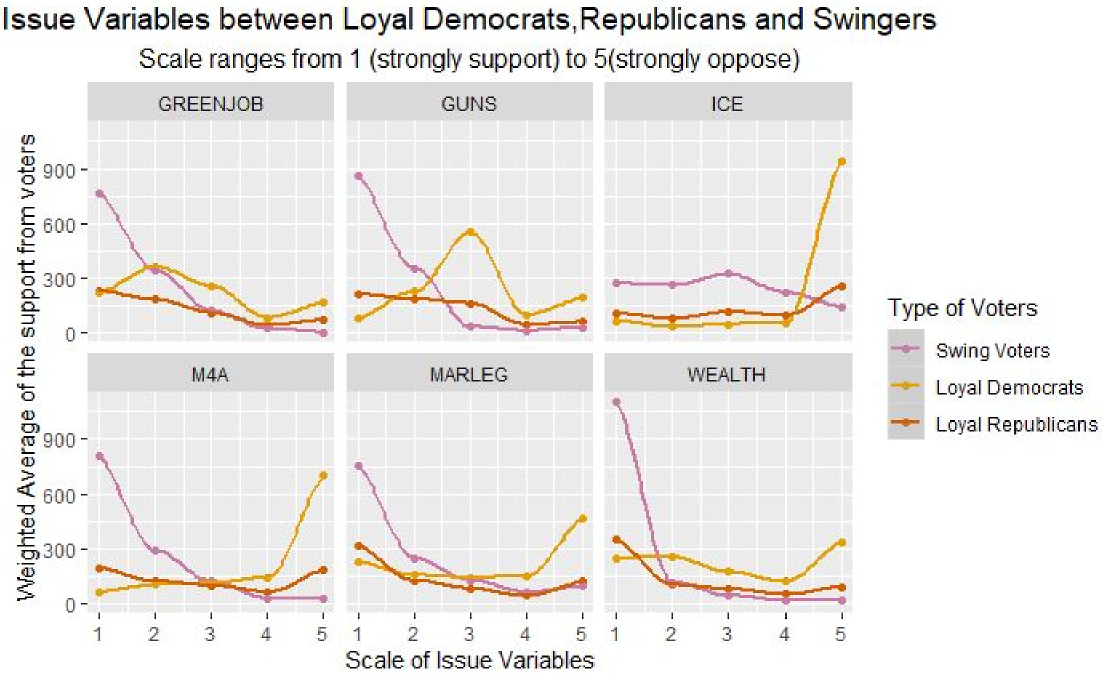

Fig 2

We study the Swing Voters and how they differ in ideologies from the Loyal Democrats and Loyal Republicans from the above plot. From the faceted plot, we see that the Swing Voters are in strong favor of all the programs. This rules out the last hypothesis which says swing voters are ideologically incoherent and don't have consistent patterns in their issue positions. The second hypothesis which says on most issues, swing voters are split, with some of them acting more like Democrats and others acting more like Republicans seem to suit the swing voters better.

GREENJOB: A Green Jobs program

All three of the subsets of the voters show strong support for the Green Jobs program. The average number of supporters from Swing Voters for this program is way too high compared to the average number of supporters from Loyal Democrats and Loyal Republicans.

GUNS: Gun control

The Swing voters and Loyal Republicans strongly support the Guns Program which says that these voters think the regulation should make purchase of guns should be made difficult. A Larger number of Loyal Democrats have neutral opinions on this program.

ICE: Defunding Immigration and Customs Enforcement

Loyal Republicans and Loyal Democrats show strong opposition to the defunding of the Immigration and Customs Enforcement program, whereas the average support from Swing voters for this program is slightly higher than the average opposition. Thus, we can say Swing voters support the program unlike the other two subsets of voters.

M4A: Medicare for All

Swing Voters exhibit strong support in favor of the Medicare expansion in the country, whereas Loyal Republicans have no clearer opinion on this program. On the other hand, Loyal Democrats seem to be in opposition to the program thus saying they do not want Medicare expansion in the country which is unusually odd for Democrats.

MARLEG: Legalizing marijuana

The opinion for legalizing Marijuana is the same as the opinion for Medicare expansion for Swing and Loyal Democrats, i.e the Swing voters support the program, whereas the Loyal Democrats oppose the program. Loyal Republicans seem to support the legalization of Marijuana in the nation.

WEALTH: A tax on wealth over $100 million

Swing Voters and Loyal Republicans strongly support the tax on wealth over $100 million, although the magnitude of support for swing voters are swooping high when compared to Loyal Republicans. On the other hand, Loyal Democrats seem to oppose the program as the average number of supporters is slightly lesser on the scale when compared to the average number of opposition.

● What predicts being a swing voter?

We create two models to probabilistically predict whether a registered voter is a swing voter or not. For this, the first model we create is using,

Model using issue variables as predictors:

There are six predictors here, namely,

1. GREENJOB
2. GUNS

3. ICE
4. M4A
5. MARLEG
6. WEALTH

We check the predicted probability for each predictor when we add them one by one. This is also done under the assumption that the order in which the predictors are added do not matter. Hence, 6 models and their predicted probability are created instead of taking into consideration their arrangement as well. The variables are added in the decreasing order of their correlation. Next we create a model with an ‘n’ number of predictors. Then we predict and train the dataset for that model and print the accuracy of each.

- Model with 1 predictor (GREENJOB)

```
## Accuracy for model with 1 predictor:  81.14 %
```

- Model with 2 predictors (GREENJOB, GUNS)

```
## Accuracy for model with 2 predictors: 81.65 %
```

Here, we can observe that there is an increase in accuracy by 0.5%.

- Model with 3 predictors (GREENJOB, GUNS, MARLEG)

```
## Accuracy for model with 3 predictors: 81.69 %
```

Here, again we observe an increase in accuracy by 0.04% when compared to the previous model.

- Model with 4 predictors (GREENJOB, GUNS, MARLEG, M4A)

```
## Accuracy for model with 4 predictors: 81.88 %
```

The increase in accuracy when compared to the previous mode is 0.2%

- Model with 5 predictors (GREENJOB, GUNS, MARLEG, M4A, ICE)

```
## Accuracy for model with 5 predictors: 81.82 %
```

However, here there is a decrease in accuracy when compared to the previous model by 0.06%

- Model with 6 predictors (GREENJOB, GUNS, MARLEG, M4A, ICE, WEALTH)

```
## Accuracy for model with 6 predictors: 81.95 %
```

The increase in accuracy when compared to the previous mode is 0.13%

We created another model, however this time without the variable ICE.

- Model with 5 predictors (GREENJOB, GUNS, MARLEG, M4A, WEALTH)

```
## Accuracy for model with 5 predictors: 81.95 %
```

There is no change in accuracy when the variable ICE is removed. We can conclude from this that ICE is not a very important predictor when trying to predict a swing voter.

Among all of the first six models, the model with all six issue variables performs the best with the highest accuracy.

Plot for Issue Variables:

When we plot the probability using the issue variables, the other variable values are set to the median value of the sequence.

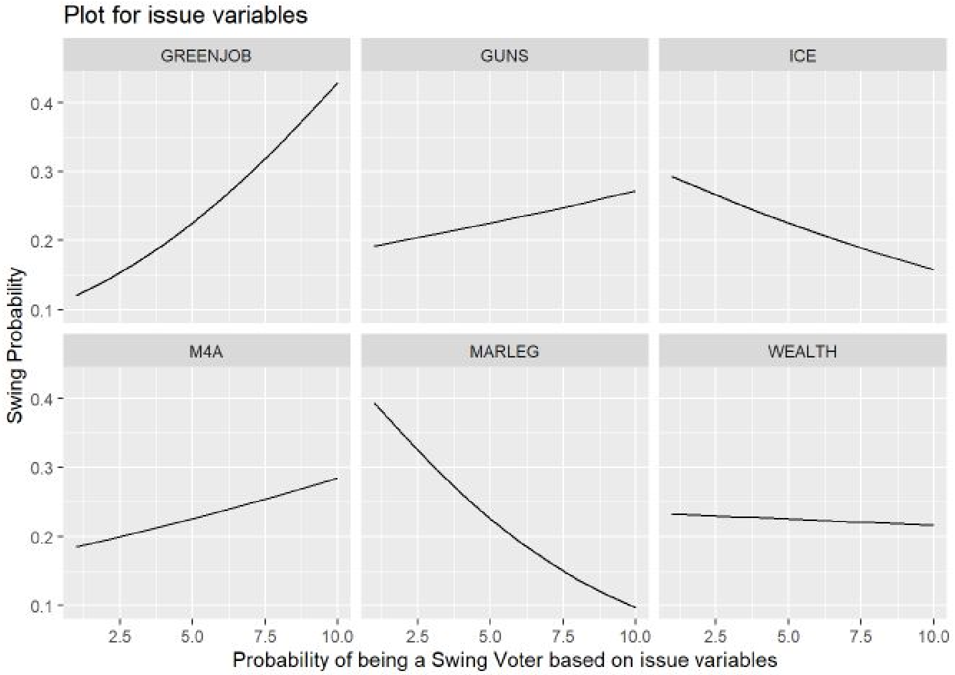

Fig 3

From the above graphs, we can see how each variable affects the probability of being a swing voter. It is evident that the variable GREENJOB has the highest affect whereas WEALTH is the least affecting factor. Supporting GREENJOB, GUNS and M4A make a voter less likely to be a swing voter, whereas supporting ICE, MARLEG and WEALTH make them more likely to be a swing voter.

Model using populism variables as predictors:

Here, respondents indicate their agreement with the given statements on a 1–5 scale, where 1 means strongly agree and 5 means strongly disagree. (6 means "Not sure.")

The three variables along with their respective statements are as follows,

POP_1: "It doesn't really matter who you vote for because the rich control both political parties."
POP_2: "The system is stacked against people like me."
POP_3: "I'd rather put my trust in the wisdom of ordinary people than in the opinions of experts and intellectuals."
We check the predicted probability for each predictor when we add them one by one. This is also done under the assumption that the order in which the variables are added do not matter. Hence, 3 models and their predicted probability are created instead of taking into consideration their arrangement as well.

- Model with 1 predictor (POP_1)

```
## Accuracy for model with 1 predictor: 80.67 %
```

- Model with 2 predictors (POP_1, POP_2)

```
## Accuracy for model with 2 predictors: 80.83 %
```

The increase in accuracy when compared to the previous mode is 0.16%

- Model with 3 predictors (POP_1, POP_2, POP_3)

```
## Accuracy for model with 3 predictors: 80.98 %
```

The increase in accuracy when compared to the previous mode is 0.15%

Among these three models, the model with all the three populism variables is the best as it has the highest accuracy for the training dataset.

When the two models, Model with issue variables and Model with populism variables are considered, the former performs better with a higher train accuracy.

Plot for Populism Variables:

When we plot the probability using the populism variables, the other variable values are set to the median value of the sequence.

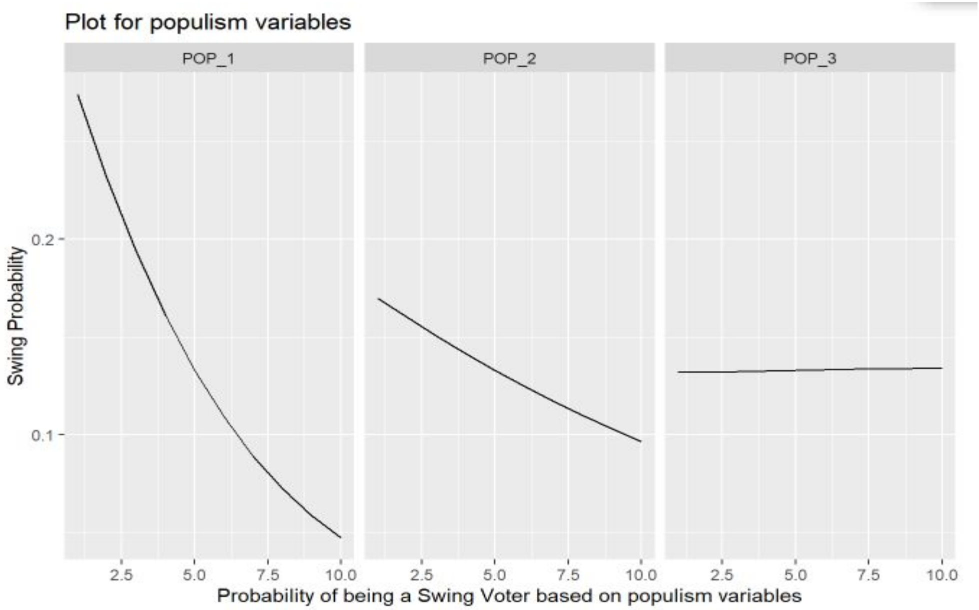

Fig 4

From the above graphs, we can see how each variable affects the probability of being a swing voter. In this case, it is evident that the variable POP_1 has the highest affect whereas POP_2 is the least affecting factor. In this case, support for POP_1 and POP_2 implies that they are more likely to be swing voters, whereas, the vice-versa is true for the POP_3 variable.

To best predict the swing voter, the model with all six issue variables and the model with all three populism variables should be used since they have the highest accuracy.

References:

1. https://rpubs.com/dvdunne/ggplot_two_bars
2. https://gerardnico.com/viz/ggplot/weight
3. https://statisticsbyjim.com/regression/model-specification-variable-selection/
4. https://stats.stackexchange.com/questions/238501/how-do-i-determine-the-bestpredictor-
5. https://ggplot2.tidyverse.org/reference/discrete_scale.html
6. https://ggplot2.tidyverse.org/reference/discrete_scale.html