**Voting Behavior Analysis of the US citizens in 2018**

**--- Project Team 3**

1. **Business Understanding**

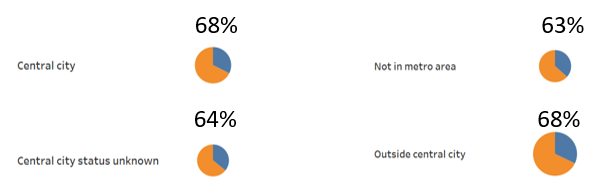
Our client on this project would be the US Government. What the Government wants is to increase the percentage of voters in the coming election of 2020. The country’s average vote rate on the last election (2018) was 66% with a range of 25 points going from 54% in West Virginia to 79% in Colorado. The standard deviation of the sample was 6%. These sample facts let us know that not only the ratio of people who voted is low but also there is a big dispersion between states.

Into the database provided we count with personal information about a sample of citizen’s liker age, gender, race and level of education as well as the reason why they do not vote or not registered in the last election. Additionally, we have demographic information like the region of the voters and if the voters live in a central city or outside.

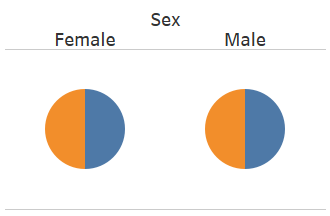
The objective of these analysis is to understand if there are a group with particular characteristics that are the cause of the lower ratio and the reason why they did not vote or register. Also, we need to understand if there are some attributes that have more influence than others over the voters’ behaviors. As long as we could identify and understand that, we are able come up with conclusions to the problem and recommendations to our client.

Using tableau, we first analyze each attribute isolated related to the vote/not vote behavior in the last election. By doing this, we can have a broad sense about which are the most influencing ones and which are better to be discard. (showed below)

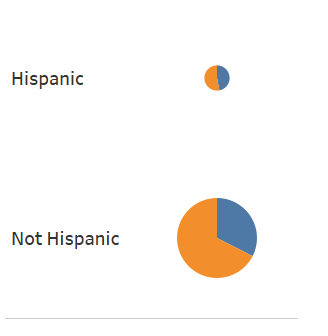
**Metropolitan Central City Status:**

**(yellow – voted, blue – not voted):**

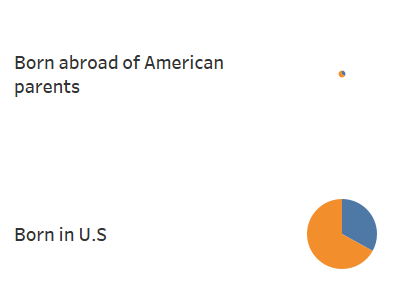
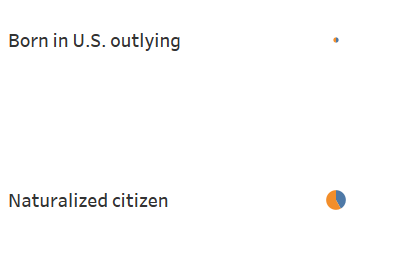
It seems that the Metropolitan Central City Status has not influence in the decision of going or not going to vote.

**Gender:**

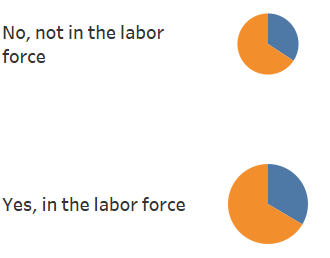
It seems that the Gender does not affect the decision of going or not going to vote.

**Hispanic or not:**

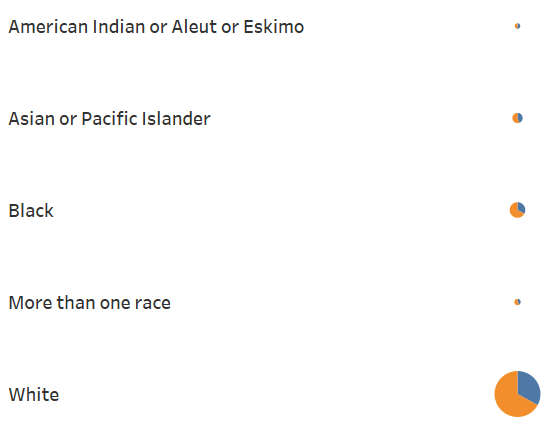
Even though there is a difference in the percentage of voted or not between the Hispanic and not Hispanic, we prefer to discarded it because the sample size is very unbalance.

**Citizenship:**

Same reason explained before, although there is a difference in the percentage of we prefer to discarded it because the sample size is very unbalance.

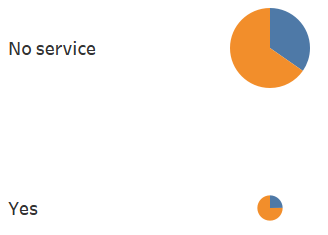
**Labor force**:

It seems that the labor force is not an attribute does affect the decision of going or not going to vote.

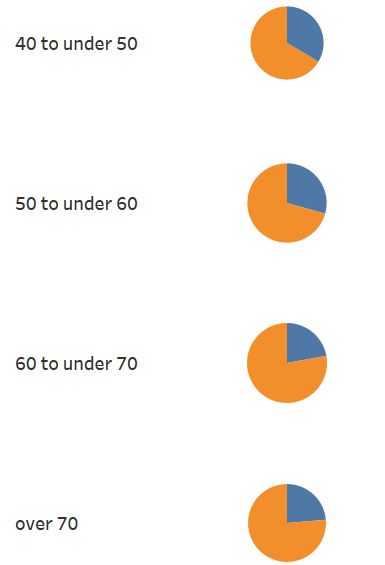
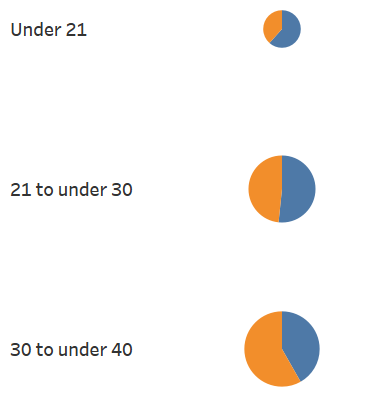
**Race Group**

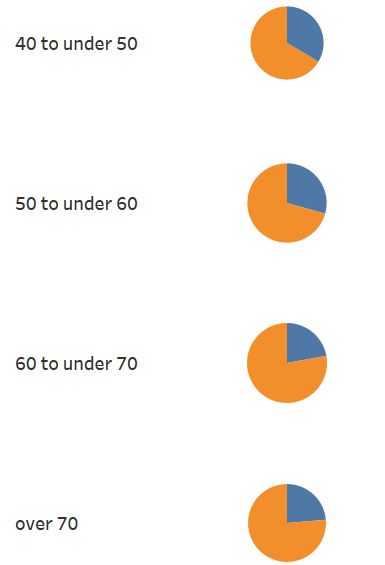
Same reason explained before, although there is a difference in the percentage of we prefer to discarded it because the sample size is very unbalance being the number of white people almost 10 times bigger than the second group that are black people.

**Military Service**:



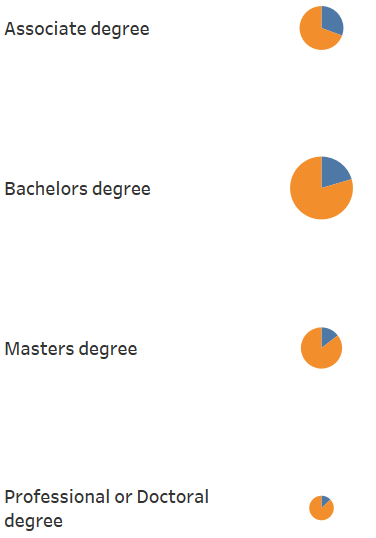
Same reason explained before, although there is a difference in the percentage of we prefer to discarded it because the sample size is very unbalance.

**Age Group**:

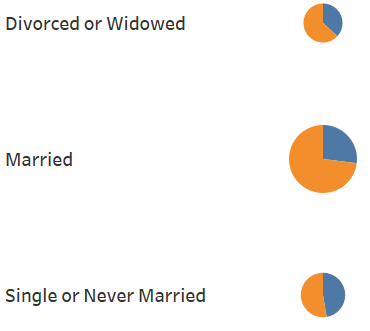


This pie charts show it clear that while the age increase, the percentage of voters over the population also increase. It seems that the age could be a factor to be analyze in more detail.

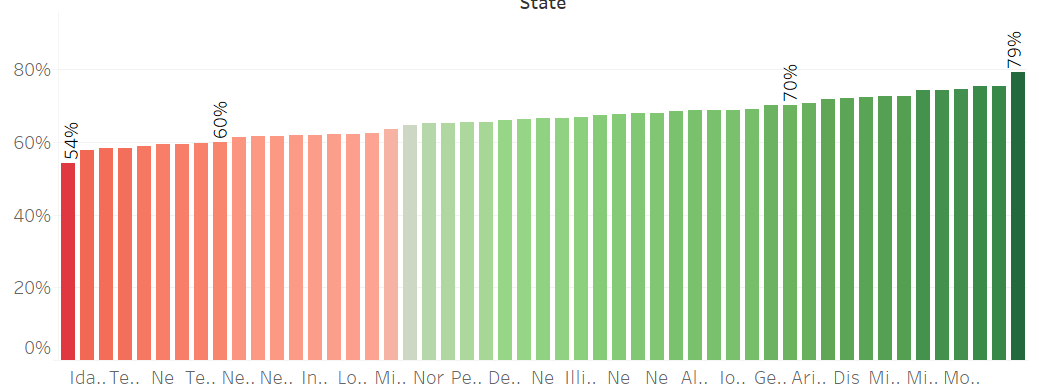
**Education Group**:



This pay charts show it clear that while the level of education goes up, the ratio of voters also goes up. It seems that level of education could be a factor to be analyze in more detail.

**Marital Status**:

In these graphs we can identify that the single or never married people has a biggest percentage of non-voters.

**State:**

In this bar chart we could see that there is a difference between the state performance in this aspect too.

Because of this analysis with Tableau, we decided to make focus on the following two main topics:

**Topic1: The relationship between people’s voting behaviors and their age, educational groups and marital status and the reason why they didn’t vote/registered for vote.**

In solving this problem, we compare vote rate for the most recent election of the 7 different education groups and also of different age groups and marital status. The reason why we put those three attributes together is that they are all closely related to each other. A person’s age may restrict his/her education status as well as marital status.

Also, by applying some classification techniques, we can find interesting rules that would decide whether a person would vote or not. With that, we can better help our client (the government) to decide which age group, marital status group or education group to focus more on propaganda and encourage them to “get out the vote”.

For example, the vote rate is the lowest for the education groups of “No school”, “Some School but no diploma”, and “Highschool graduate”. We may analyze the reason why they didn’t register or vote. For example, the biggest reason for why they didn’t register of the three groups is simply “Not interested”, then we can give recommendations to the government that they could make some direct dissemination about the election to the groups and may even provide them some benefits if they could get out the vote.

The client will benefit a lot from our analysis of this topic. They could have a clearer direction on which groups to focus and save the cost of marketing by making it more efficient.

**Topic 2: The relationship between people’s voting behaviors with their region. It includes the analysis of performance of the government of each state and the difference in people’s voting behavior of the west part and the east part of the US.**

In solving this problem, we compare the registration rate, vote rate and registered but not vote rate for the most recent election of all the 50 states in the dataset. By making visualizations, we can clearly see which states are underperforming and we can also try to find out the reason why those states have obviously lower vote rate than the other ones.

This would help our client (the government) to better focus their propaganda on certain regions to improve efficiency. Also, even though “State” is not a critical attribute in this dataset, it could give us additional insights about the performance of the government of each state and also the possible reason for the difference in vote rate of the west and the east. By doing the analysis, we can figure out possible recommendations for the government to boost future vote rate in the particular regions.

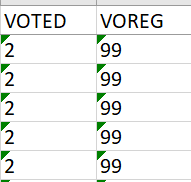
**2.** **Data Understanding and Preparation**

This dataset contains comprehensive information of potential voters, including their demographic information, voting information and also some details about other information. We only keep the data of 2018 here because human behavior has changed a lot in the last few years so we did not want to mix old information with the new ones.

The dataset contains complete information that we need, however, there are also some data quality issues that needs efforts in data preprocessing.

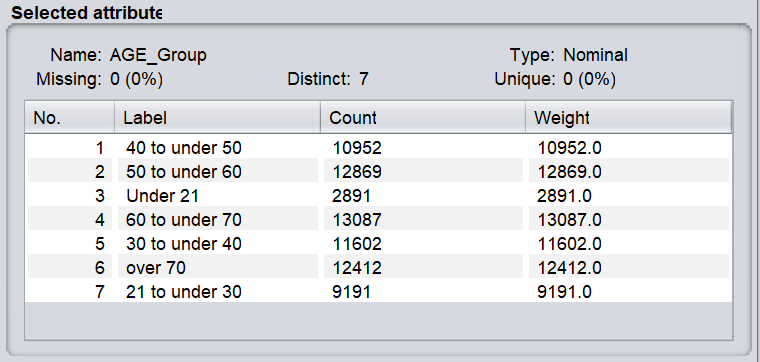
Firstly, the attributes are all coded in numbers and it’s hard for us to understand in a straight forward way, therefore, we decided to align the data with its dictionaries. We changed all the “numbers” in this dataset into its real meanings in excel.

Also, there are values that have no actual meanings in the dataset like 99 (not in universe), we should also replace those attributes into their real meanings of for example “Registered” in this case.



Besides, there are some symbols in this dataset also, includes dots, bars and so on. In order to make the data understandable for WEKA, we also removed the symbols in Excel.

After we aligned the dictionaries with the dataset solved the primary quality issues, we should then consider proper data preprocessing before apply data mining techniques in WEKA. After the alignment with dictionary, the only attribute that needs to be discretized in this dataset is the “Age”. It ranges from 18-85, and we choose to discretize it into 7 groups of “Under 21”, “21 to under 30”, “30 to under 40”, “40 to under 50”, “50 to under 60”, “60 to under 70”, and “Over 70”. This discretization is more in line with people’s education status, marital status etc., which we think is reasonable.



Also, before we apply classification to predict whether a person would vote or not, we should eliminate the attributes that are the premises or derivations of the vote condition. For example, the attributes like “Reason why not vote” or “Voting on or before election day” etc. should all be removed, since they already denote whether a person voted or not. We don’t want those attributes to mislead the models when applying classification.

Also, we ignore the attribute of “YEAR”, since we only about the data in 2018.

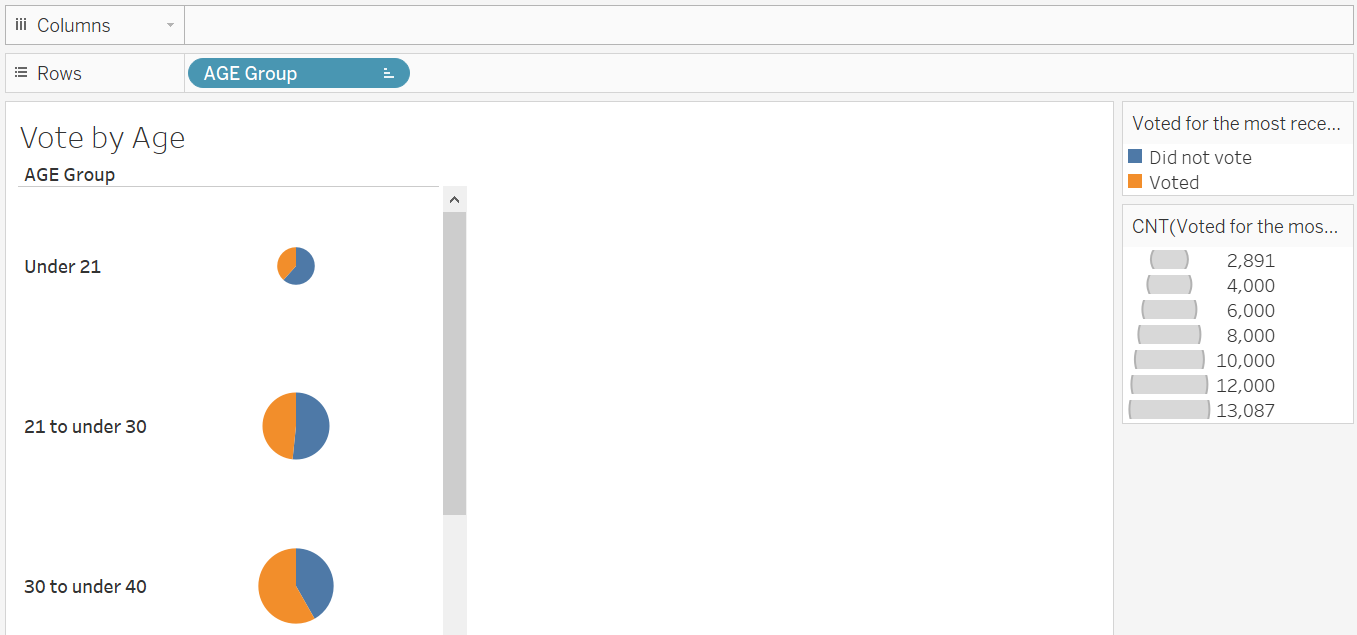
For potential data analysis or data mining techniques that are suitable for this data, we think the primary techniques we would use are visualization and classification. With visualization, we can have a clear understanding of the characteristics of the topics, and with classification, we can either verify our findings or add details to it.

**3.** **Data Mining and Evaluation**

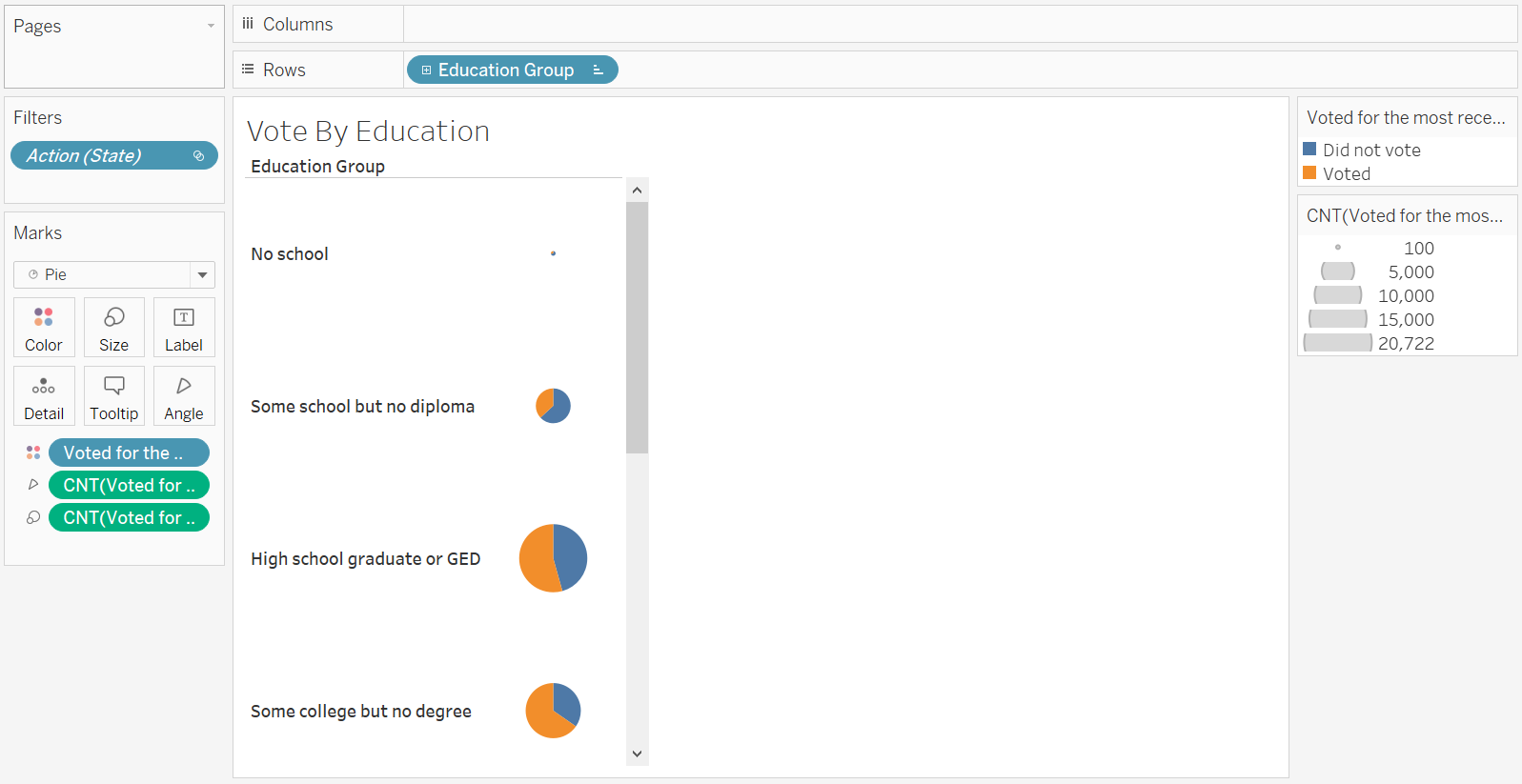
In this part, we organize our contents based on the two different topics and evaluate them separately.

**[Topic 1]** **Analysis of relationship between people’s voting behaviors and their age, marital status and educational groups**

First of all, we made separate visualizations based on the three aspects independently, to see their influences on the voting behaviors generally.

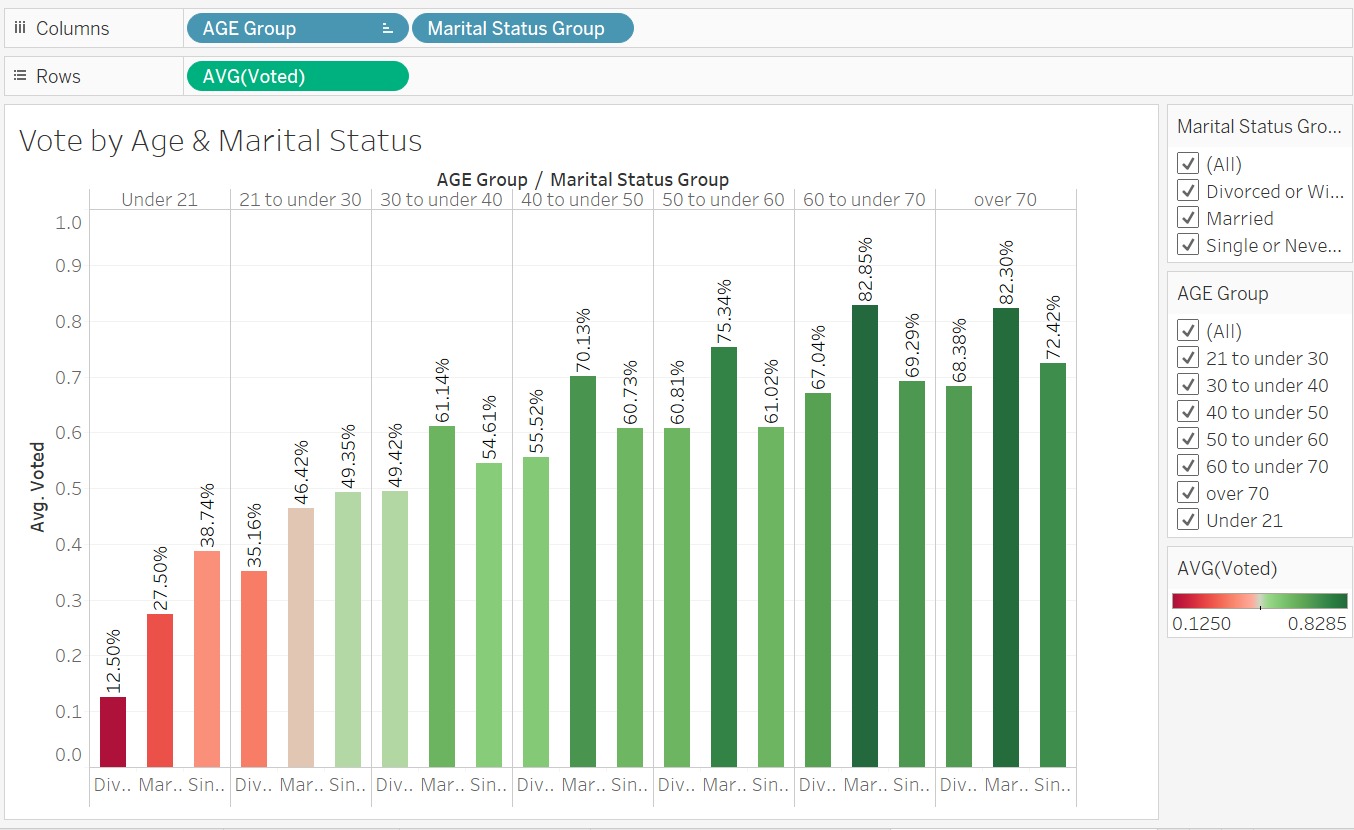




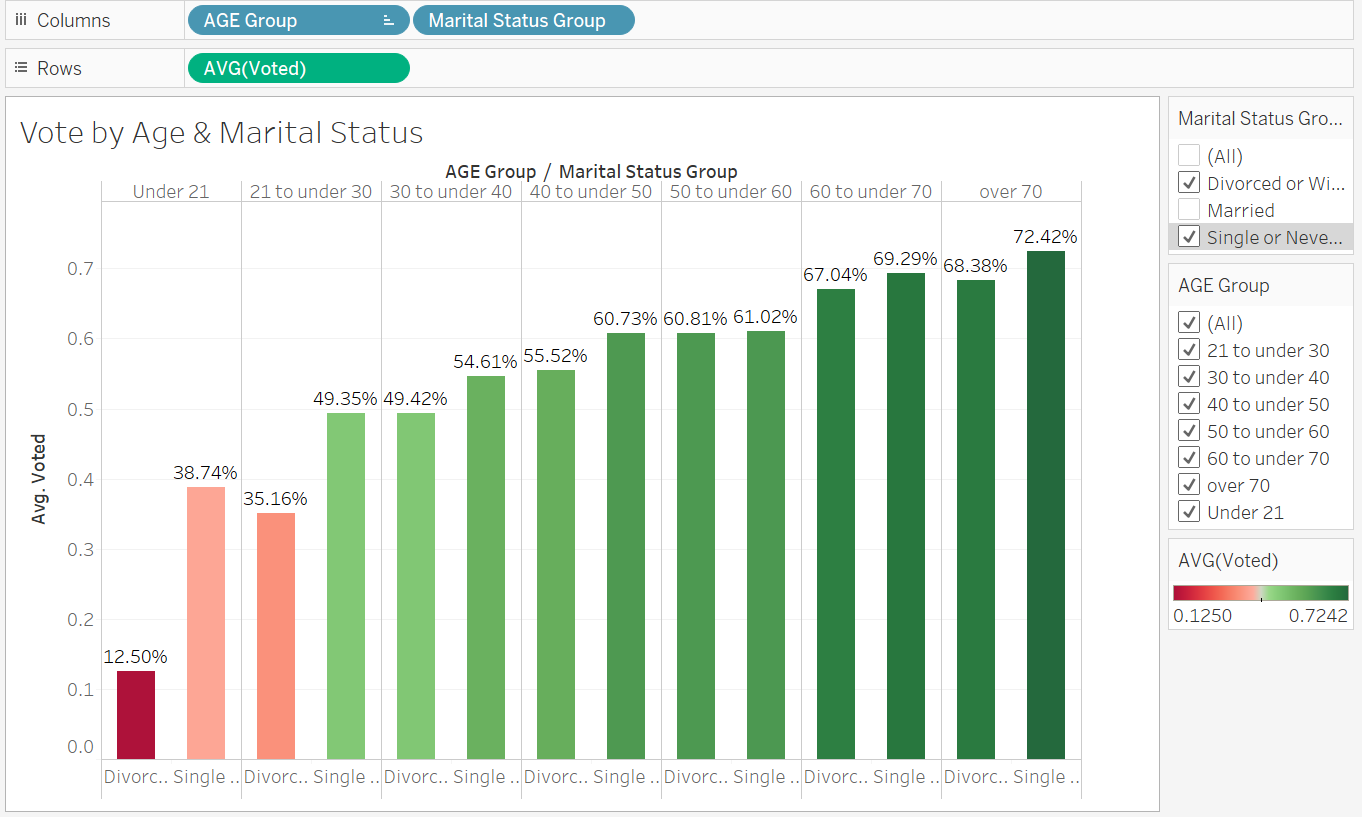


From the above visualizations, we can have a brief idea that people who are young, single and have lower degrees tend to give out less vote. This is in line with our common sense that people who are young are more probable to be single and have lower degrees, and they may care less about the political election.

However, the visualizations we made above are quite general and it may disguise some interesting insights. Therefore, we then dig deeper into this problem by making a visualization on vote rate of different age groups and marital status groups.

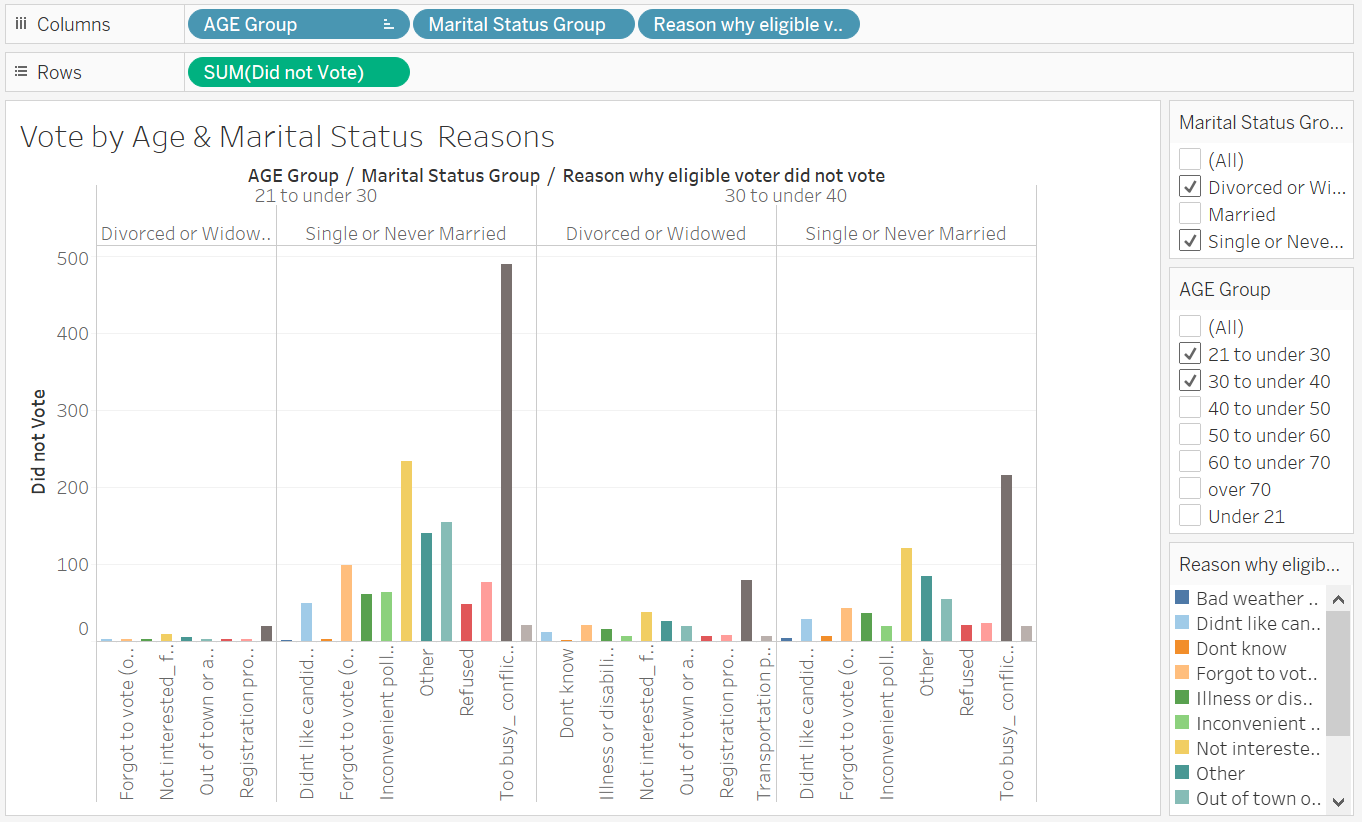


We can see from the above visualization that still, people who are under 21 and 21 to under 30 have the least vote rate. However, we can see an interesting thing here that for each age group, the marital status group that have the least vote rate is always the group of “Divorced or Widowed” rather than “Single”.



In the pie chart, people who are single appear to have the lowest vote rate, so why in this case people who are divorced or widowed seem to take place of that? We then filtered the data to “Divorced and Widowed” and “Single”, and verified our finding from the above visualization that people who are divorced or widowed actually have lower vote rate than people who are single in each age group. From our point of view, this is caused by the Simpson's paradox. Even though for the entire group, people who are single have the lowest vote rate, when we divide them into different age groups, the people who are divorced or widowed take charge.

After determining what group of people have the lowest vote rate (young, single or divorced), we then want to analyze their reason for why they didn’t get out the vote.



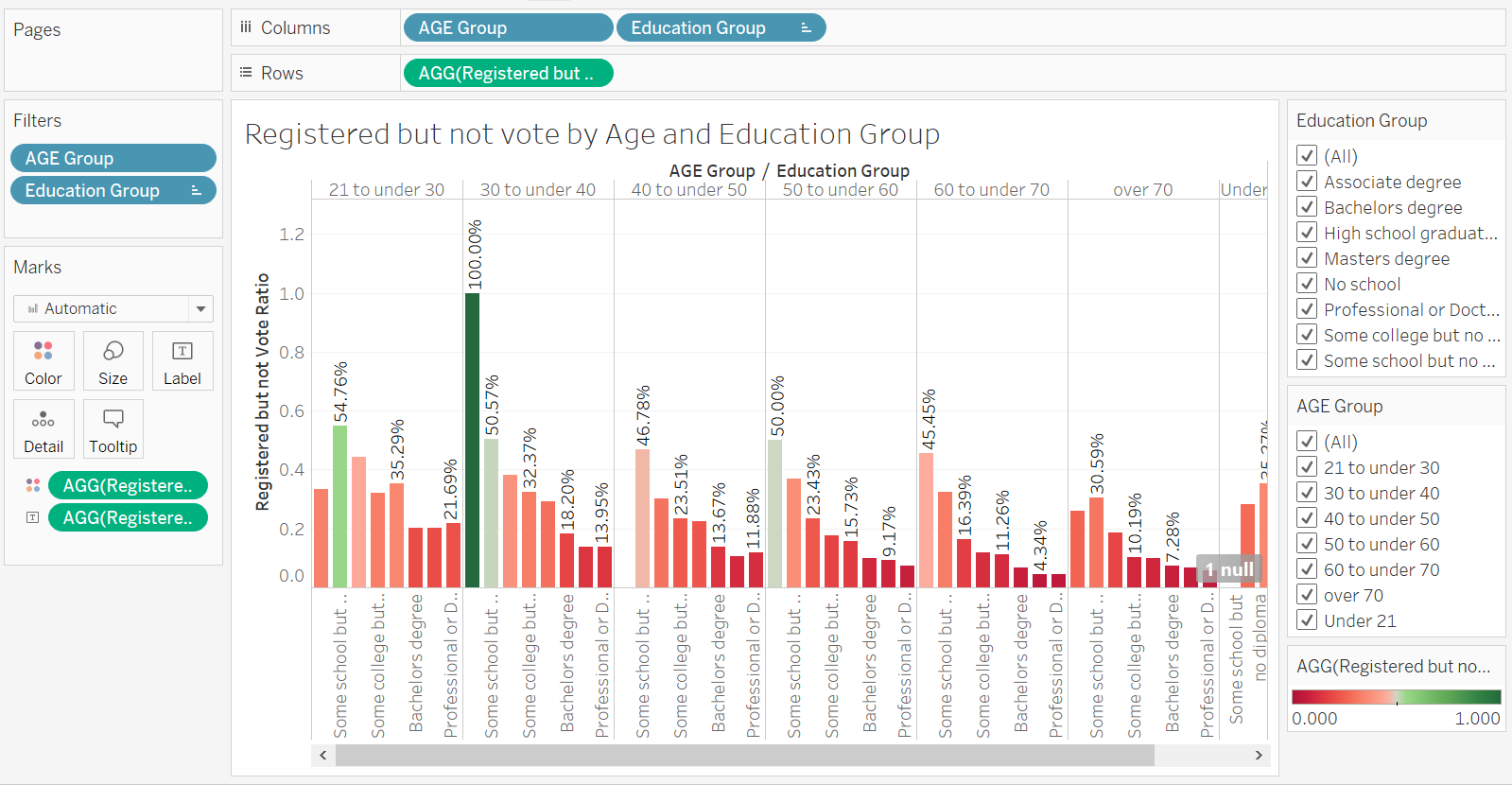
From the above visualization, we can see that the top 2 reasons for the people who didn’t vote in these groups are “Too busy, conflicting work or school schedule” and “Not interested, felt my vote wouldn’t make a difference”.

Then, another aspect we are interested about is the relationship between people’s voting behavior and their education groups, so we made a visualization based on the vote rate of different age groups and education groups as below.



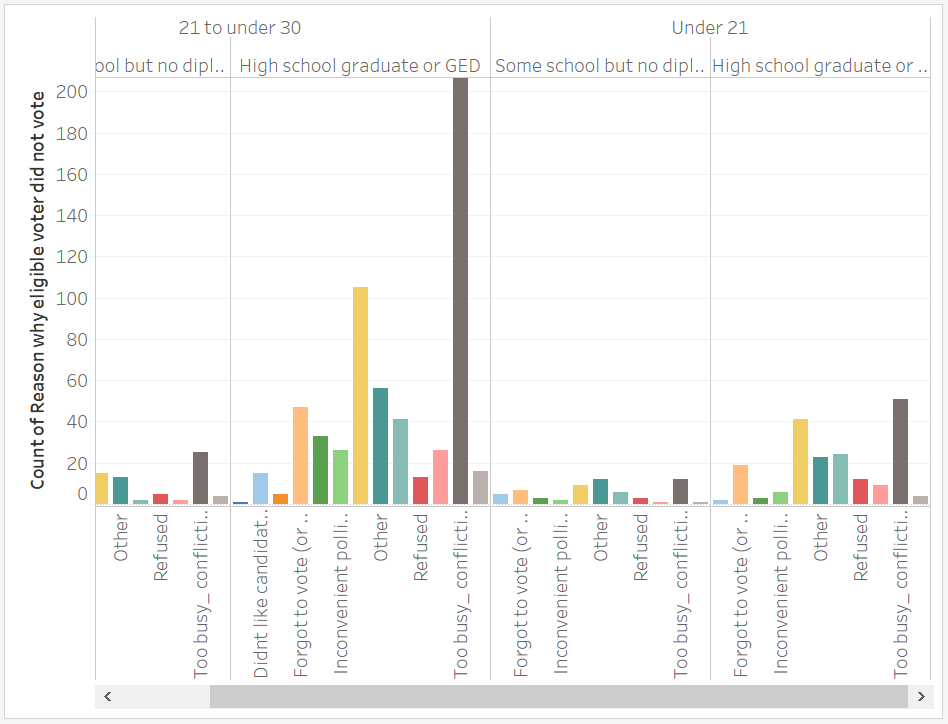
We can see it clearly from the above visualization that in each age group, people who have lower degrees have lower vote rate. The vote rates are especially low in the education groups of “No school”, “Some school but no diploma” and “High school graduate or GED”.

Also, when we turn the measurement from vote rate to the rate of registered but not vote, the education group of “No school”, “Some school but no diploma” and “High school graduate or GED” also stands out among all in each age group.



Therefore, we can conclude that people who are young have lower degrees tend to give out less vote, and they are also more likely to give up their vote even though they have registered.

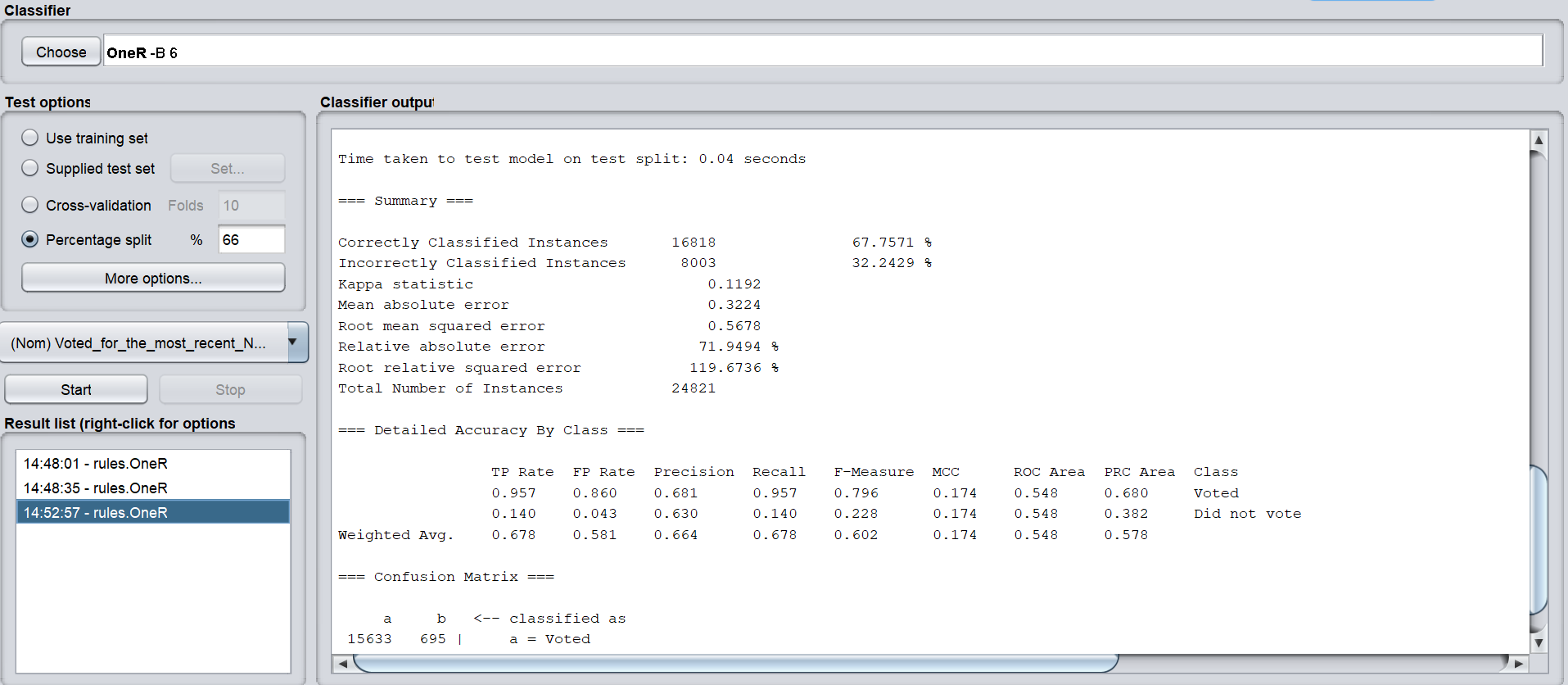
To analyze their reasons for them of not voting, we also made a visualization like below.



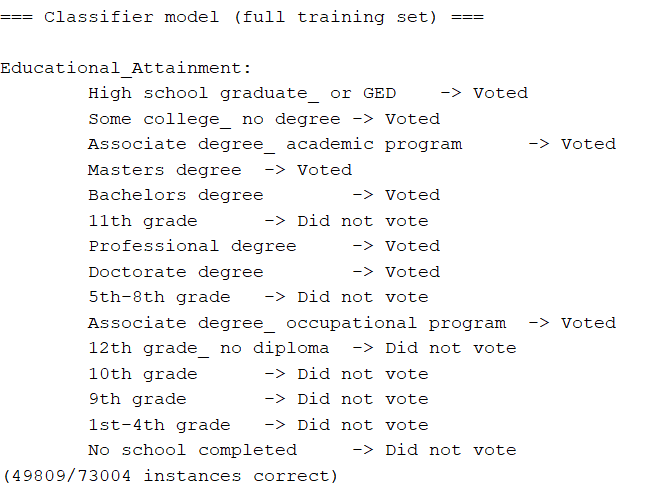
We can then see from the visualization that the same reasons stand out, that are “Too busy, conflicting work or school schedule” and “Not interested, felt my vote wouldn’t make a difference”.

After we have done about visualizations and found some interesting insights about the relationship between people’s voting behavior and their age, marital status and education groups, we also want to add more details to our findings using classification.

Firstly, we would like to use OneR as a beginning trial to predict whether a person would vote or not, and analyze the classifier model.

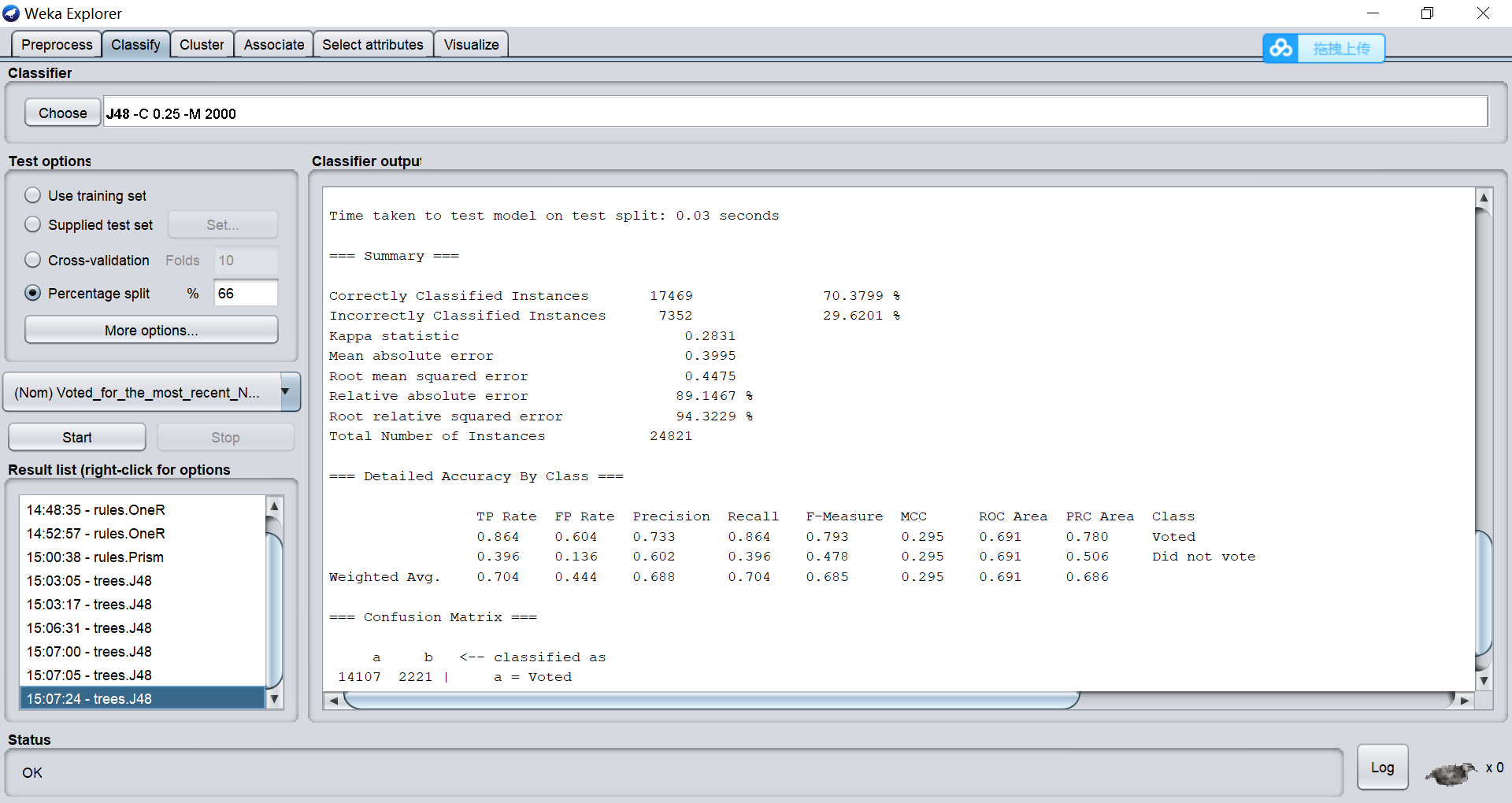


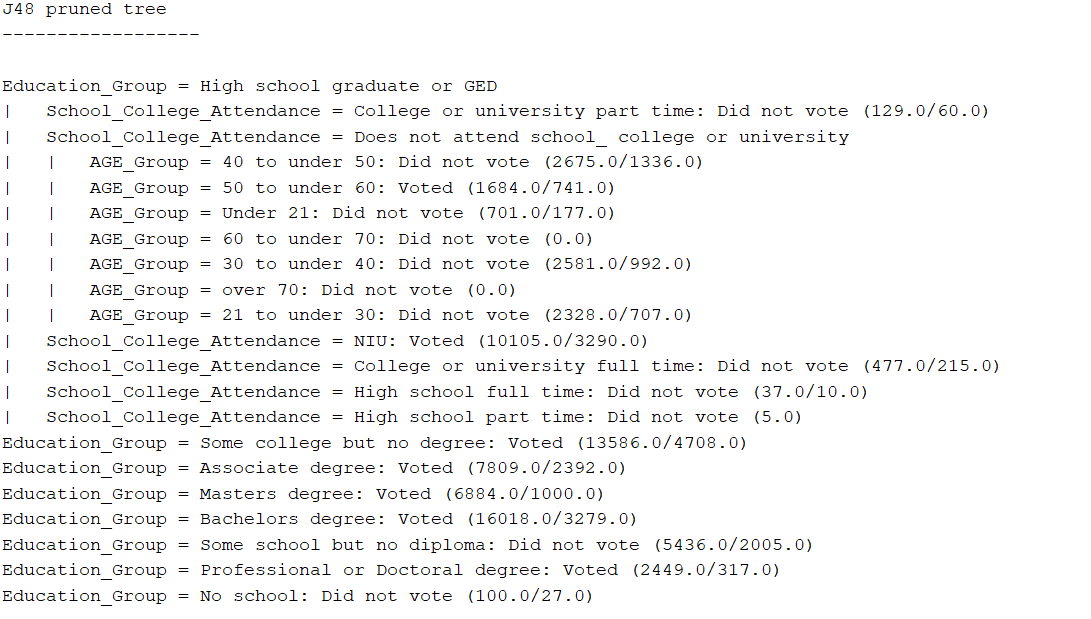
We used percentage split as our test option, and the model accuracy is round 67.76%.



The classifier model of OneR is as above. The model is based on the attribute of “Education Attainment”, and we can see a clear trend that as people get more years of education, their classified class start to change from “Did not vote” to “Voted”, which is also in line with our findings using visualizations.

Then, we changed the scheme to J48 and set the minNumObj value to 2000, since there are more than 70,000 instances in this dataset with only two classes, we want the decision tree we get to be smaller and simpler for the convenience of visualization. And the result is as below.



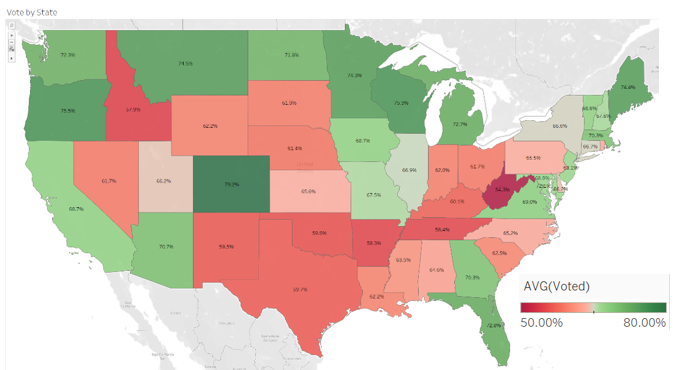


The model accuracy is improved to around 70.38%. And we can see from the structure of the tree that “Education Group” is the determinant attribute in this dataset.

In this dataset, applying J48 could give us better results than OneR, and the visualization in J48 would also give us more insights. For example, the higher the degree the higher the possibility of the instance to be classified as “Voted”.

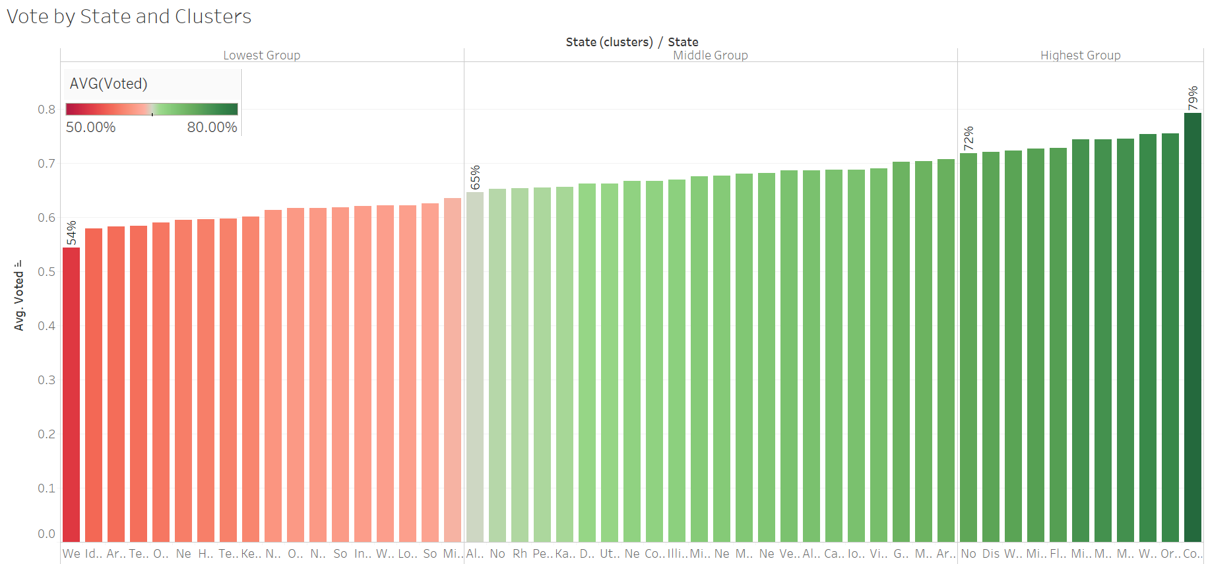
To conclude, we analyzed the relationship between people’s voting behavior with their age group, marital status and education groups using both visualizations and classifications in this topic. We’ve found various interesting insights in this process, and we would further summarize and extract business meanings from them in following parts.

**[Topic 2] The relationship between people’s voting behaviors with their region (geographic analysis)**

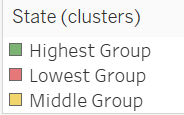
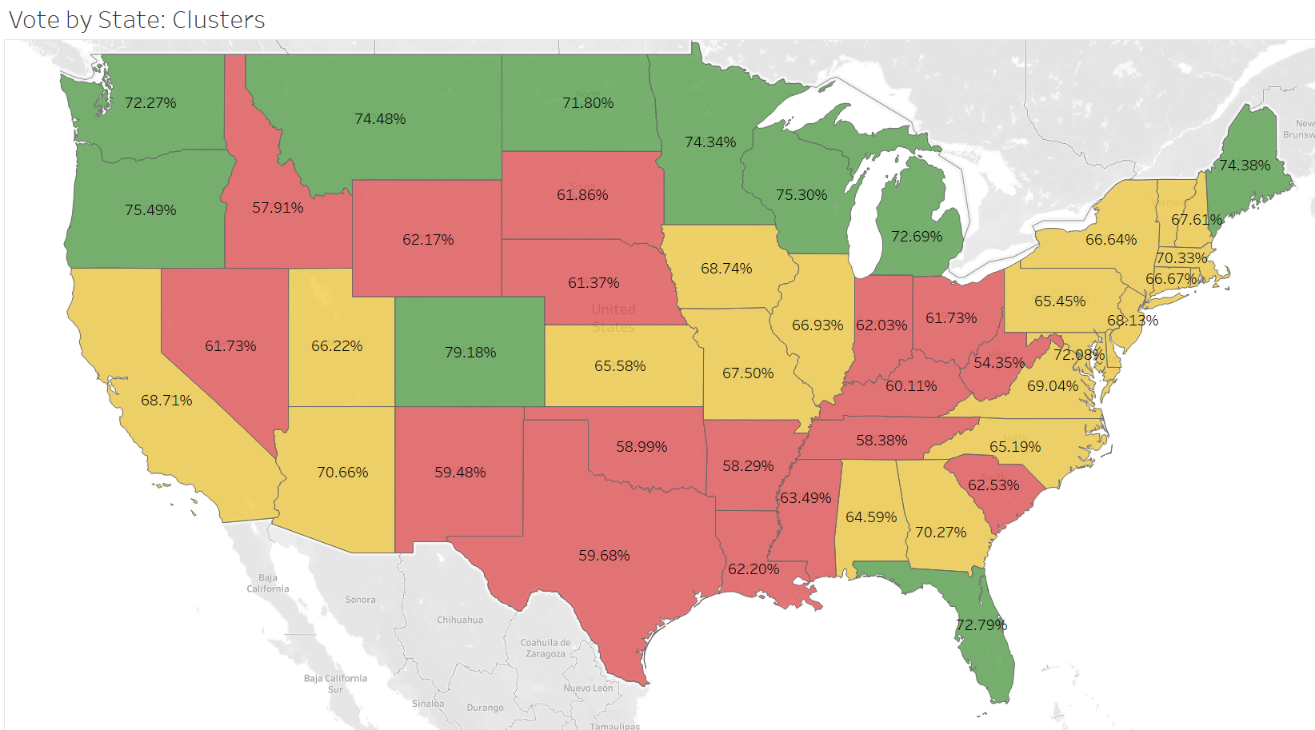
We want to analyze the voting rate among the different states.

There is a big dispersion on the voting rate among the country states going from West Virginia with 54% of people that voted in the las election to Colorado with 79%.

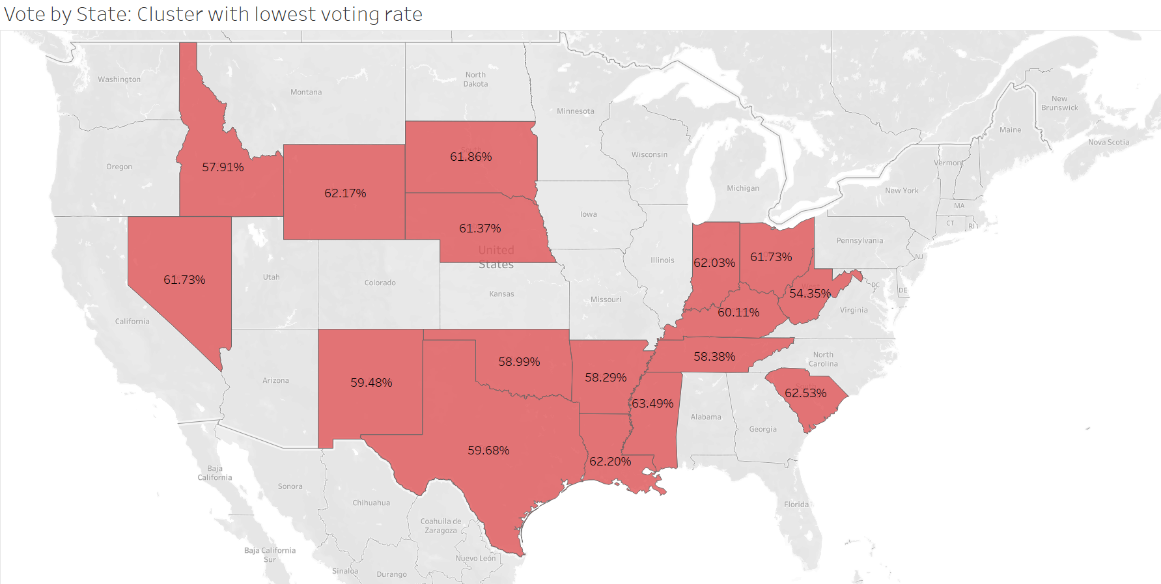
To start this analysis, we first identify three main clusters of states based on their percentage of voting rate.



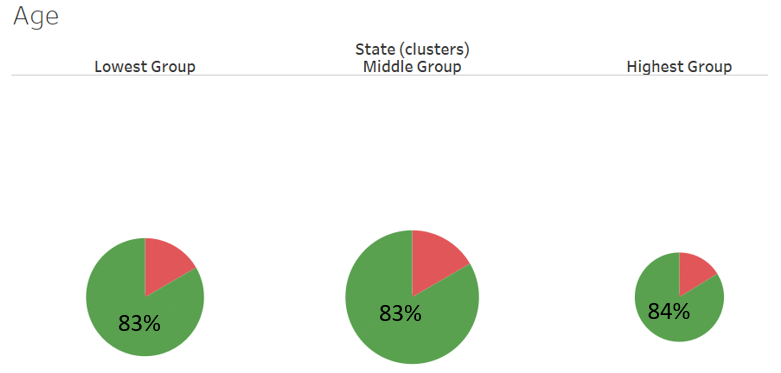
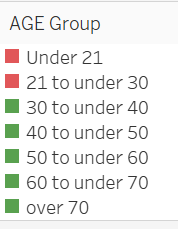
Then we visualize these cluster into a map. The green one is the group of states with the highest voting rate, the yellow one is for the states in the middle and the red one is for the states with the lowest voting rate.



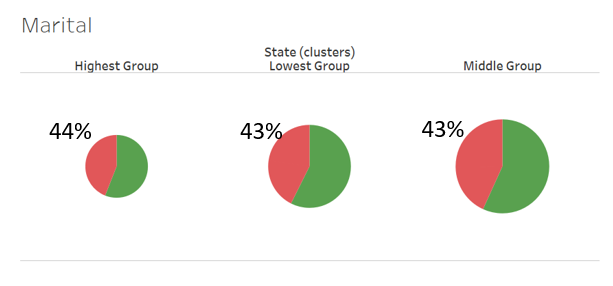
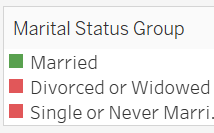
By focus in the red group, the ones with the lowest voting rate, we can see that is not a matter of a geography. The states with the lowest voting rate are all across the country.



Using these three clusters, we analyze separately the main attributes of influence identify before (age group, marital status, labor force and education) to find some common patter in the red ones.

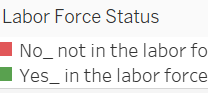
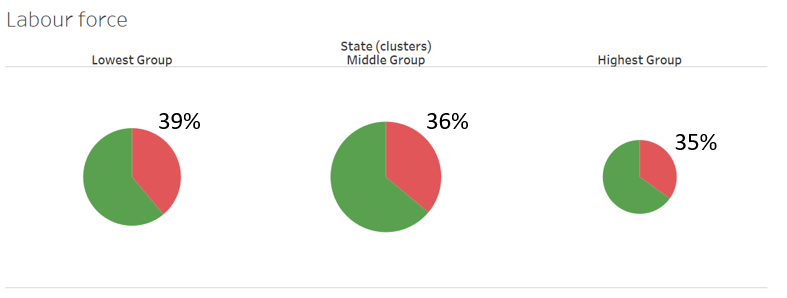
As we know from the analysis done above, people under 30 years old have least vote rate. Thus, if these cluster of states have a bigger percentage of people under 30, could be expected a lower voting rate.

As the above makes clear, there is no difference between the percentage of people under 30 in any of the three clusters. Thus, age could not be the reason for a lower voting rate.

Then, we analyze the weight of “single and divorced or widowed” vs married people.

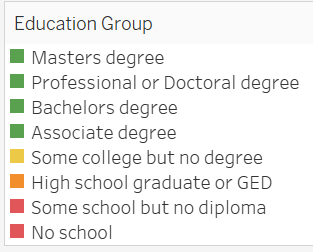
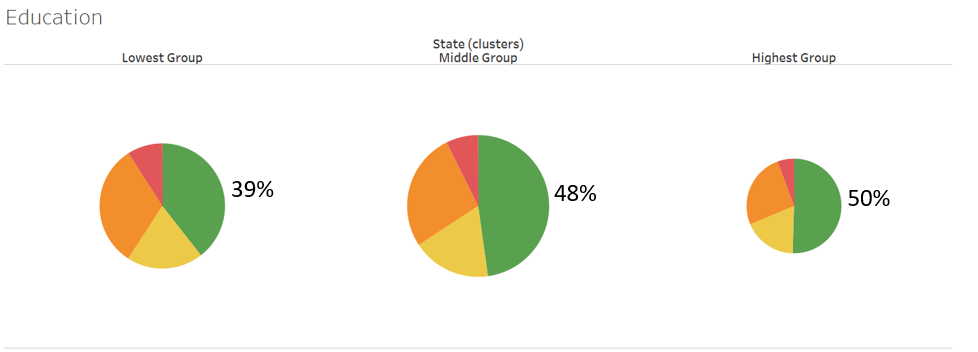
The graph shows clear that the proportion between clusters of married vs not married is the same. Then, the marital is not reason for a lower voting rate.

In third place, we analyze the proportion of people in labor force vs the people that are not in the labor force between cluster. We were trying to identify if people in the cluster with lowest voting rate were busier than the other ones because one of the main reason for not voting is being too busy.

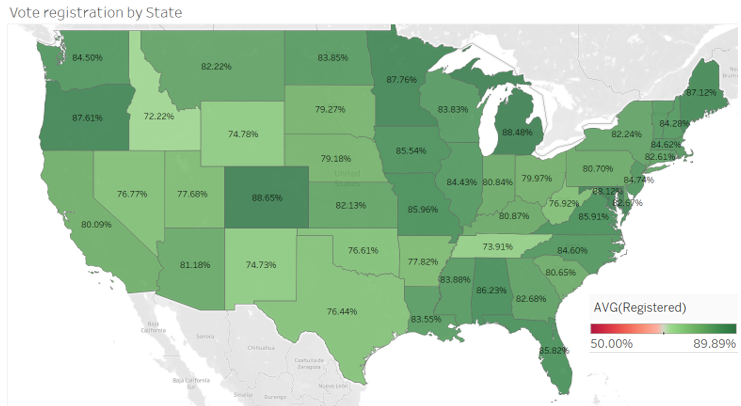


Despite of there is a little difference between clusters, it is not significant. Thus, labor force is not a factor that influence in the voting behavior.

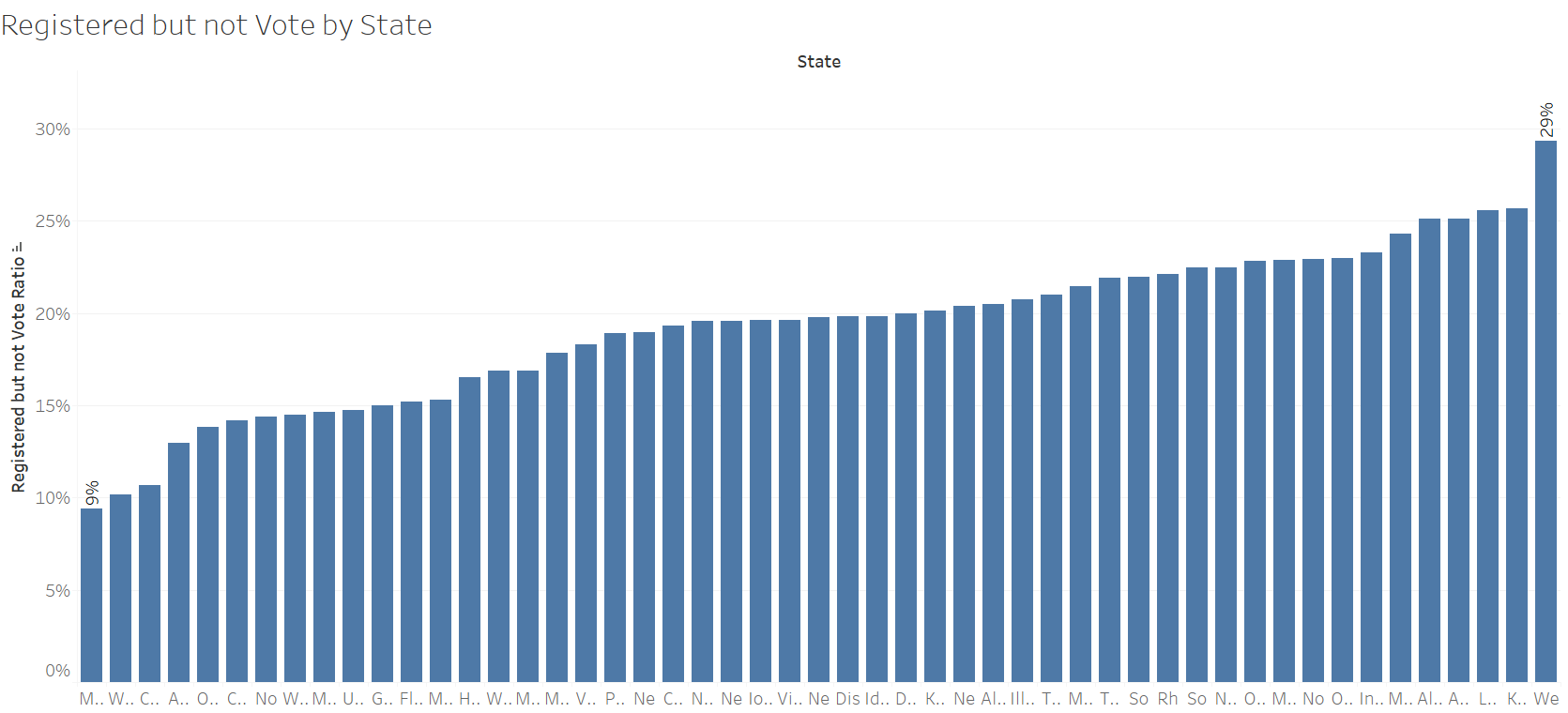
Fourthly, we analyze the percentage of less educated people vs more educated people. To simplify the visualization, we compare the people with associate degree, bachelor degree, professional or doctoral degree and master degree vs all the others.



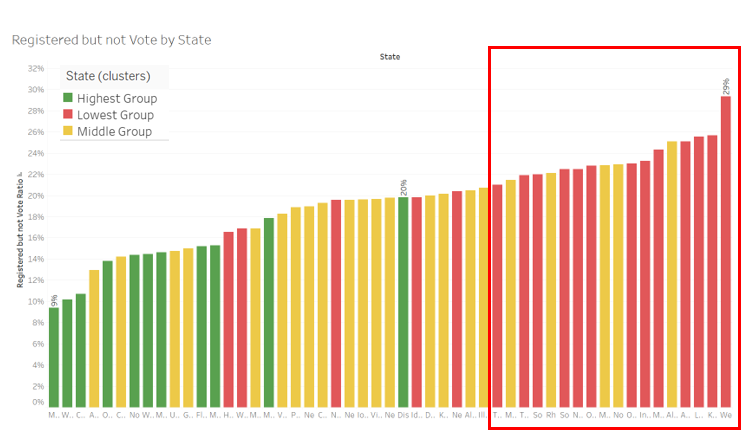
As we can see in the graph above, there is a difference of more than 10 points in the percentage of people with high education between the cluster of states with higher voting rate vs the cluster of states with the lowest voting rate. This graph makes it clear, again, that education is one of the key factors that influence in the voting rate.

Another aspect that we are interested about is to analyze if the difference on the voting rate between clusters is because of the people on those states does not registered to vote or because they registered but then they do not vote.

As the graph make it clear, the registration rate by state is much higher than the voting rate.

The graph below shows the percentage of people that were registered to vote but did not vote in the las elections.

There is a big GAP between Montana were only 9 % of the registered people to vote did not do it and West Virginia were 29% of them did not do it.

We can understand where is the main problem by identify to which cluster each of the states belongs to.

As we can see, the states with the higher “registered but not voted” rate are all from the cluster of states where the voting rate is the lowest. Thus, the problem is not people registration for voting but the ones that are registered and do not vote.

***Now the idea is continuing to analyze if there is something that makes registered people not vote.***

**4.** **Analysis of Models**

**[Topic 1]** **Analysis of relationship between people’s voting behaviors and their age, educational groups and marital status**

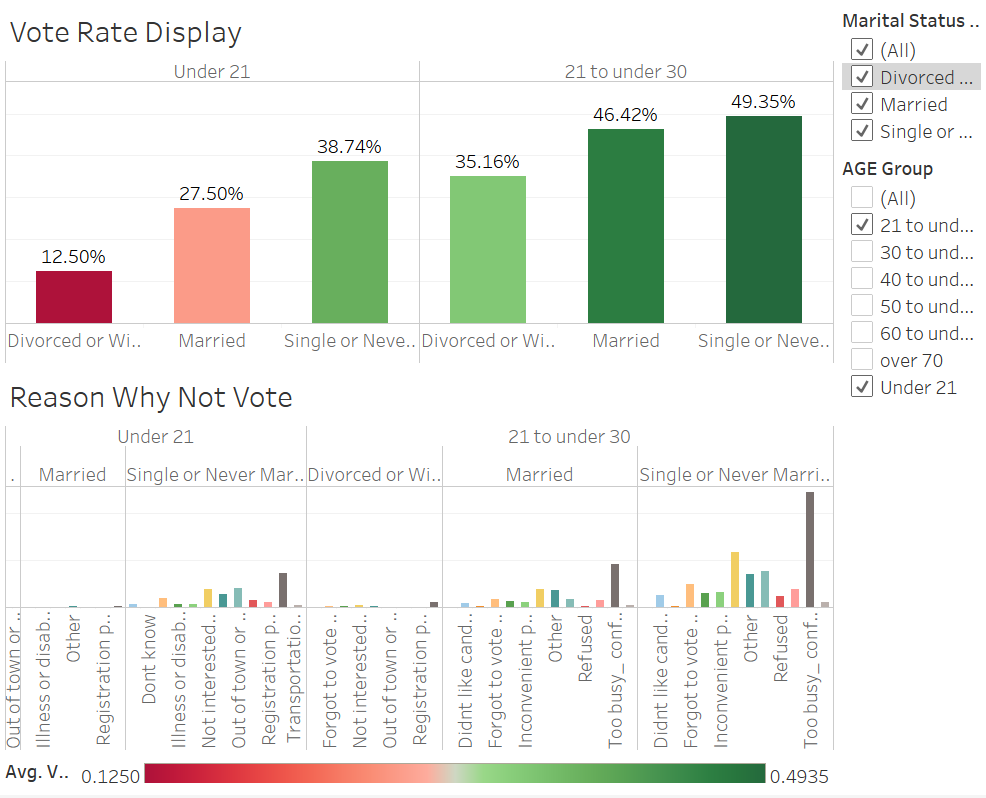
In topic 1, we first analyze the relationship between people’s voting behavior with their age and marital status. And we created a dashboard displaying what we’ve found.



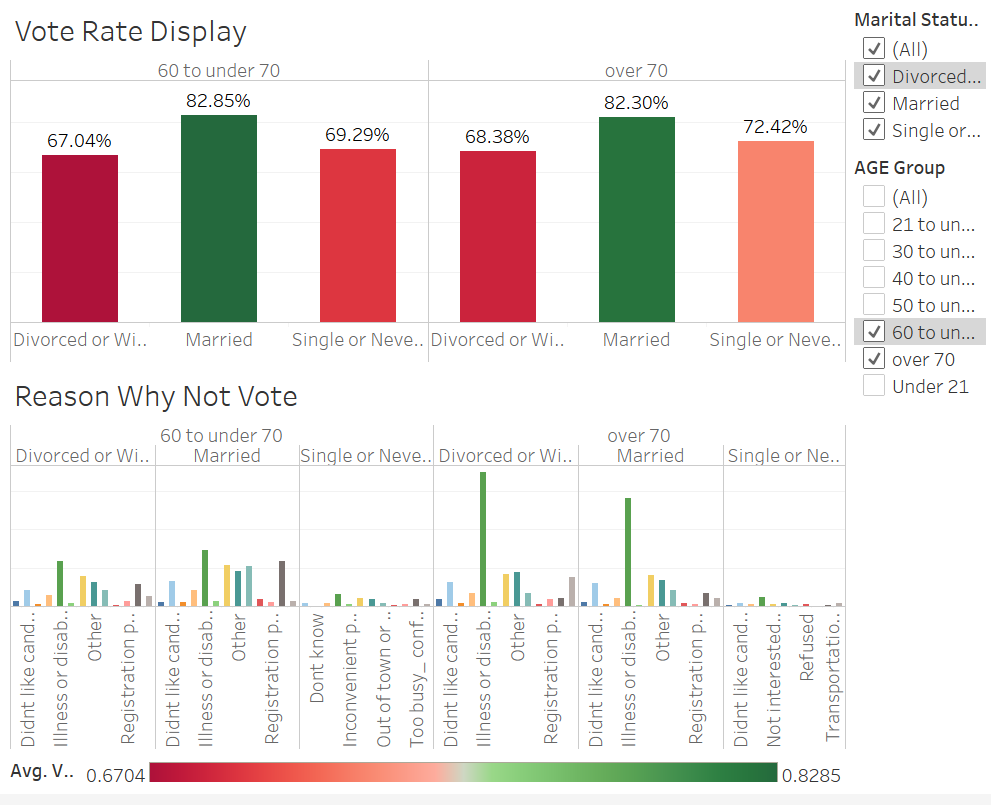
From the above dashboard we can see that people who are younger turns out to have lower vote rate. And when we come to focus on the marital status, we can see that in each age group, people who are divorced or widowed generates the lowest vote rate.

Therefore, we found the group of people who dragged the vote rate down among all, that is the group of young people and divorced or widowed people. And we then focus on the reasons why they didn’t get out the vote. From the lower part of the dashboard, we can see roughly that the reason for people who didn’t vote are generally “Too busy” or “Not interested”, but how would the reasons differentiate among different age groups, we can still find something interesting.

We first filtered the data into young people as below.



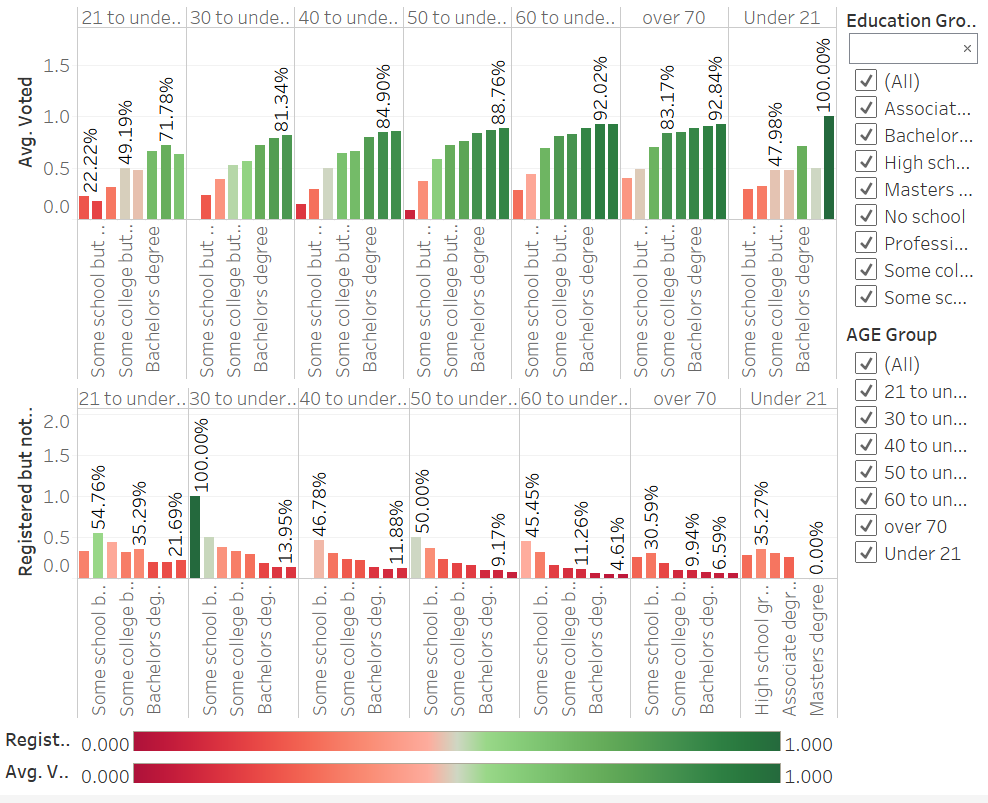
We can see that for young people, no matter what their marital status is, the main reasons for no voting are also “Too busy” and “Not interested”. And the same reasons apply to the middle aged people of 30-60. However, when we then focus on the elder people, the reason they give starts to be different.



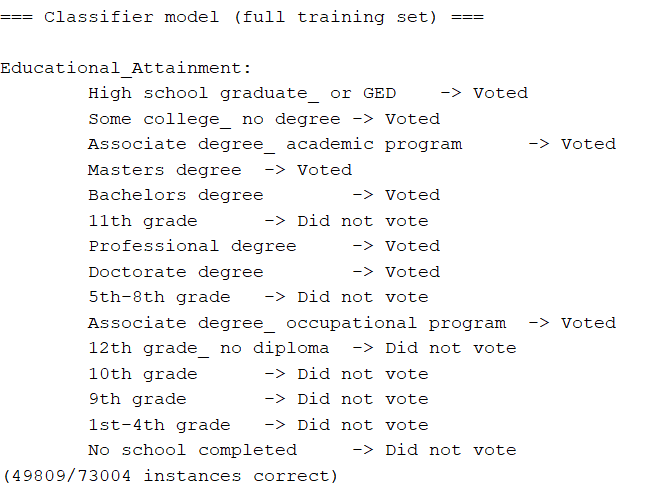
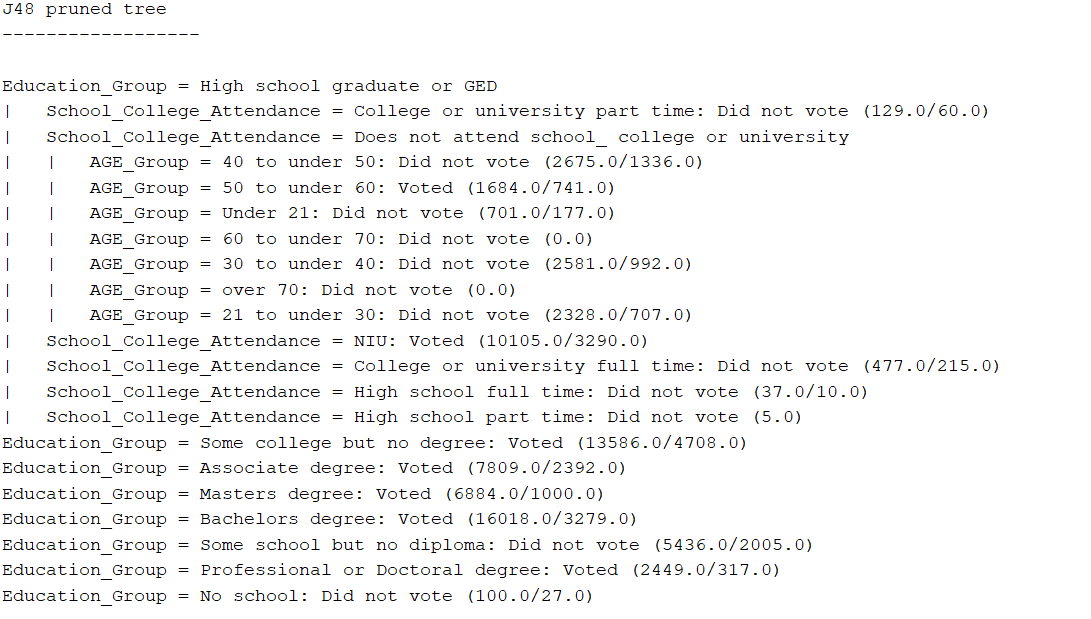
When we filtered the table into elder people (60 or above), we can see that the main reason for them to not vote changed to “Illness or disability”.

For this part of display, we want to show our client that when we want to boost the vote rate, we should specifically focus our attention on the group of young people and divorced or widowed people. Also, in general, for the young people and middle-aged people, we should focus on how to solve the problem of them being “Too busy, conflicting work or school schedule” and “Not interested, felt my vote wouldn’t make a difference”. For the elder people, we should find ways to help them vote when they have illness or disabilities

In topic 1, we also analyzed the relationship between people’s voting behavior and their education groups.

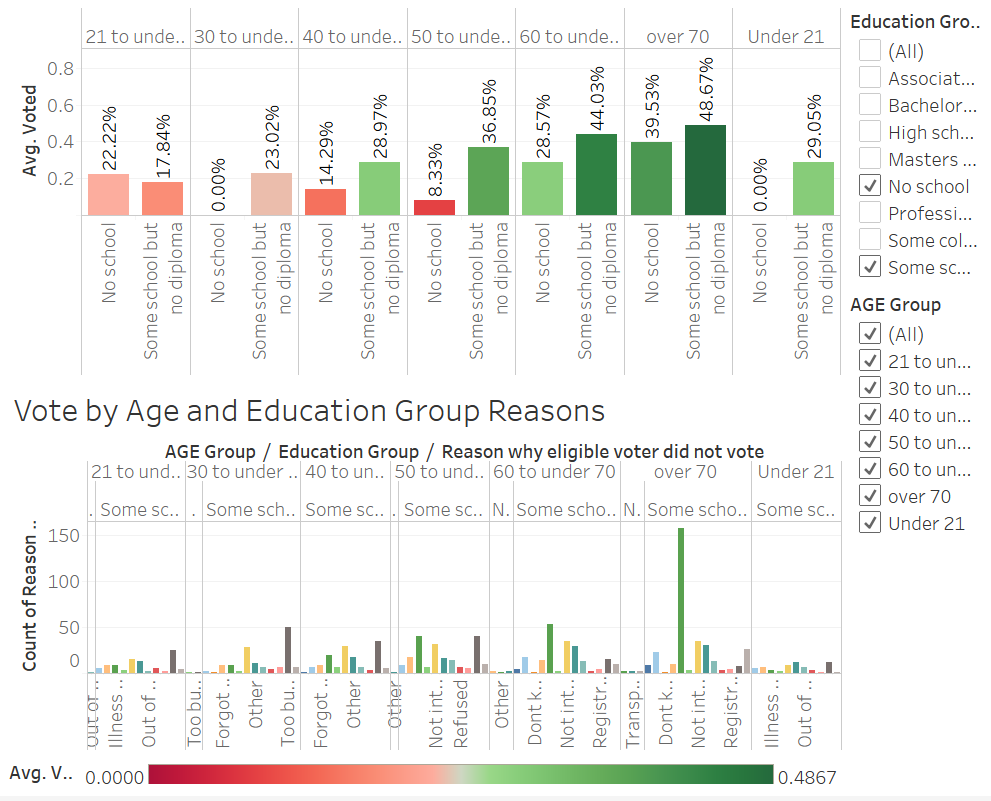


From the above dashboard we can see that people with lower degrees (No school, Some school but no diploma, High school graduate or GED) tend to have lower vote rates as well as higher registered but not vote rates. And the trend is reversed for people who have higher degrees like Doctoral degree or Master’s degree.

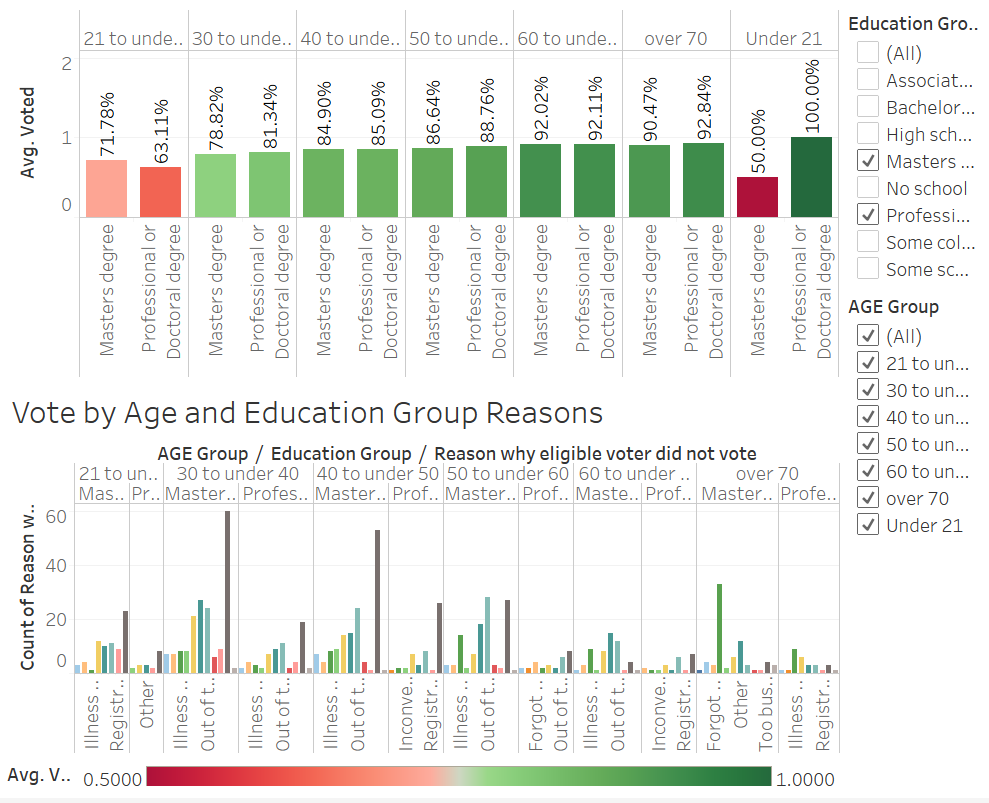
Additionally, we also applied classification schemes to the data, and from the above rules we get from classification scheme of OneR and J48, we can also see that the higher the degrees, the higher the possibility of the voter to be classified as “Voted” and vice versa.

Then, we further analyzed the reason why people didn’t vote when people’s education group differentiates. We first filtered the data into people who have lower degrees.



We can see that for people who have no school or have some school but no diploma, the main reasons for no voting are also “Too busy” and “Not interested” for young people, and “Illness and disability” for elder people.

However, when we changed the filter to people who have doctoral degrees and master’s degrees, the reasons for no voting also changed.



We can see from the above dashboard that for people with higher degrees, the reason of “Not interested” has lower proportion now, what take majority are the reasons of “Too busy” and “Out of town or away from home” for young people and “Illness and disability” for elder people. People with higher degrees would be more interested in voting and what prevent them from getting out the vote are more due to irresistible factors.

For this part of display, we want to show our client that apart from focusing on voters who are young or divorced/widowed, we should also focus on voters who have lower degrees to encourage them to get out the vote. In general, for people with lower degrees, we should try to avoid the problem for them being “Too busy” on the election day and try to build up their interests. While for people with higher degrees, we should also avoid the problem for them being “Too busy” on the election day and try to fix the problem of “Out of town or away from home” for them.

All in all, for topic 1, what we would like our client to know is that when we want to boost up vote rates, we should focus our attention on the people who are young,

divorced/widowed or have low degrees. Also, for people who are both young and have low degrees, we should solve the problem of “Too busy” and “Not interested” for them, and for people who are both young and have higher degrees, we should solve the problem of “Too busy” and “Out of town or away from home” for them. While for people who are elder, we should try to solve the problem of “Illness and disability” for them.

**5.** **Recommendations**

**6.** **Conclusions**