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

**--- Project Team 3**

1. **Business Understanding**

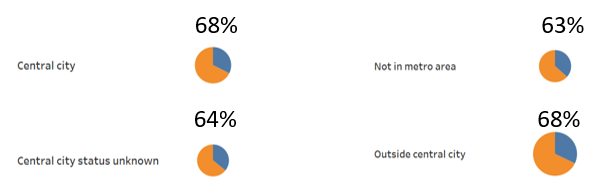
Our client on this project would be the USA Government. What the Government wants is to increase the percentage of voters in the coming election of 2020. The country’s average vote rate on the last election (2018) was 66% with a range of 25 points going from 54% in West Virginia to 79% in Colorado. The 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.

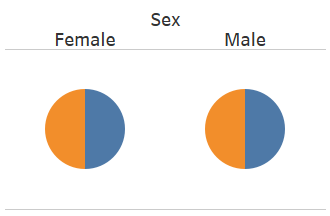
The objective of these analysis is to understand if there are a group with particular characteristics that are the cause of the lower ratio and the reason why they did not vote or register. Also, we need to understand if there are some attributes that have more influence than others over the voters’ 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)

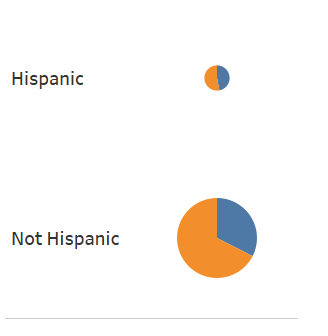
**Metropolitan Central City Status:**

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

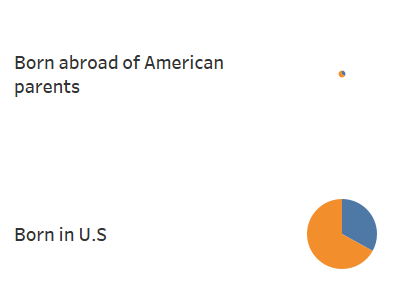
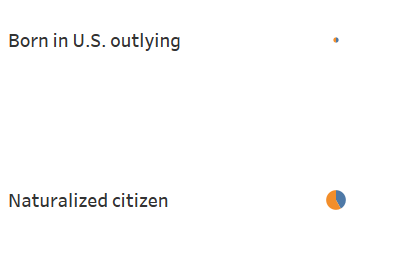
It seems that the Metropolitan Central City Status has not influence in the decision of going or not going to vote.

**Gender:**

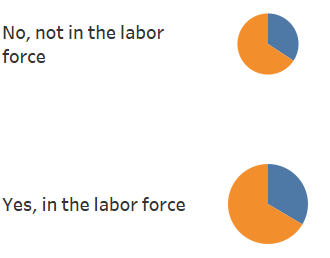
It seems that the Gender does not affect the decision of going or not going to vote.

**Hispanic or not:**

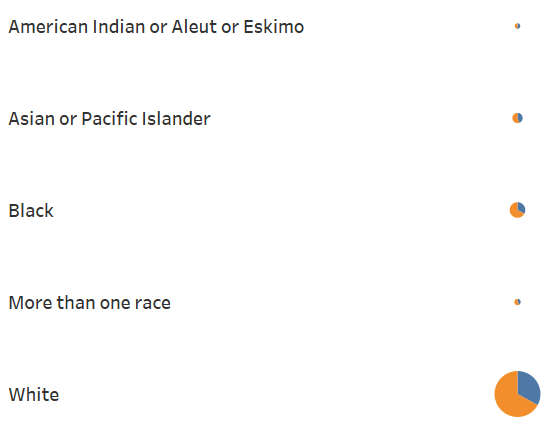
Even though there is a difference in the percentage of voted or not between the Hispanic and not Hispanic, we prefer to discarded it because the sample size is very unbalance.

**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.

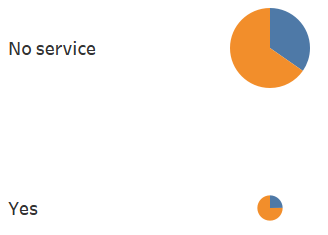
**Labor force**:

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

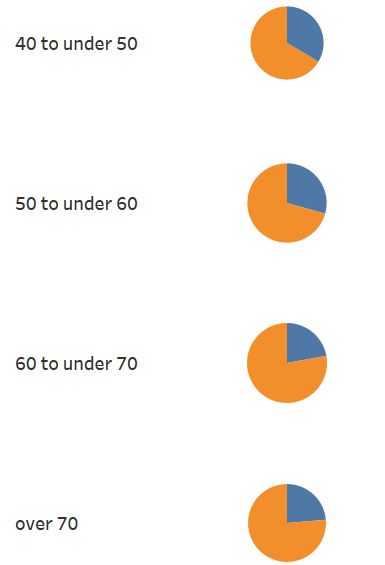
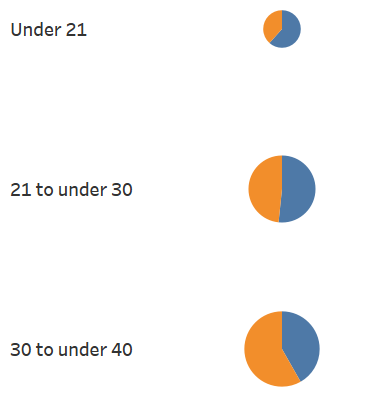
**Race Group**

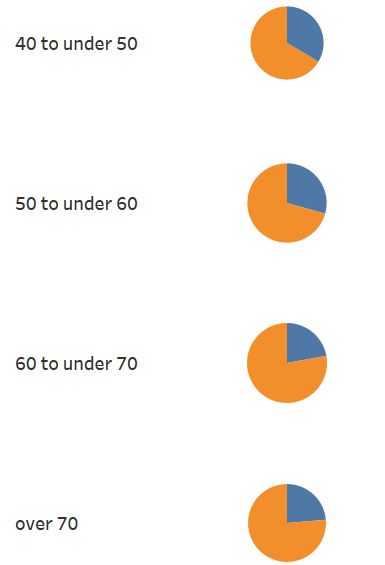
Same reason explained before, although there is a difference in the percentage of we prefer to discarded it because the sample size is very unbalance being the number of white people almost 10 times bigger than the second group that are black people.

**Military Service**:



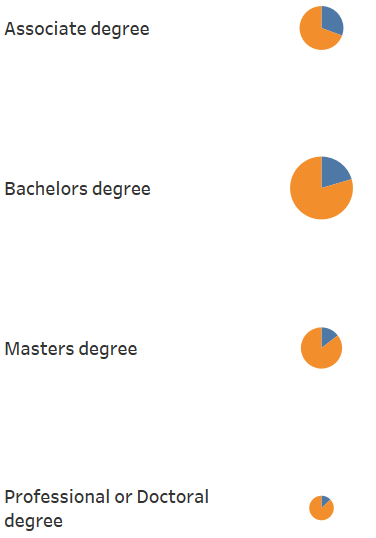
Same reason explained before, although there is a difference in the percentage of we prefer to discarded it because the sample size is very unbalance.

**Age Group**:

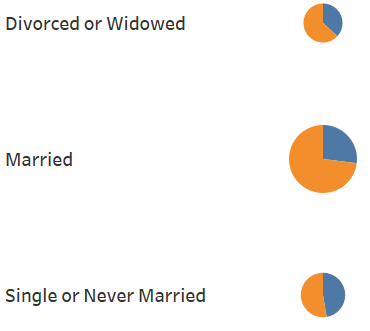


This pay charts show it clear that while the age increase, the percentage of voters over the population also increase. It seems that the age could be a factor to be analyze in more detail.

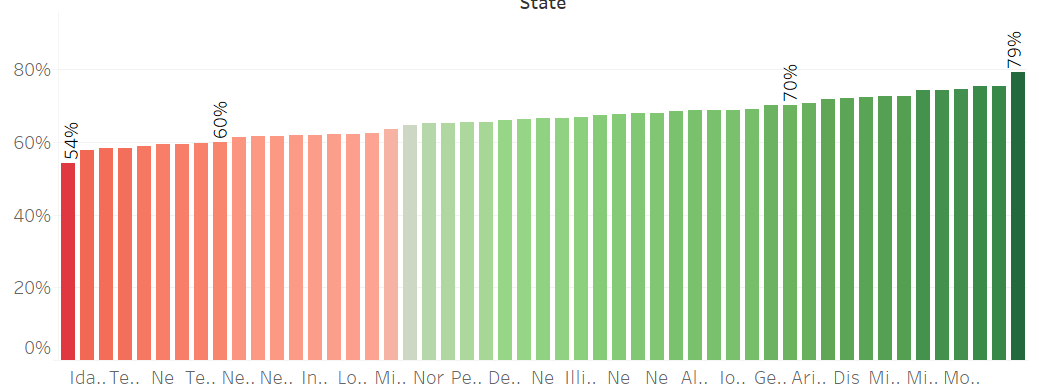
**Education Group**:



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 the registration rate, vote rate and registered but not 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 also closely related to each other. A person’s age may restrict his/her education status as well as marital status.

Also, by applying classification techniques, we can find some interesting rules that would decide whether a person would vote or not. With that, we can better help our customer (the government) to decide which 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.

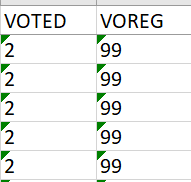
**2.** **Data Understanding and Preparation**

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.

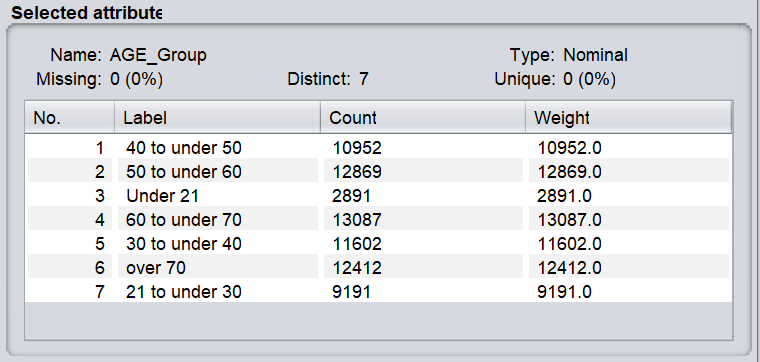
Firstly, the attributes are all coded in numbers and its hard for us to understand in a straight forward way, therefore, we decided to align the data with its dictionaries. We changed all the “numbers” in this dataset into its real meanings in excel. **(Something needs to be added here)**

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**

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

**5.** **Recommendations**

**6.** **Conclusions**