Exploration of Election data from 2018

- **Data exploration and cleaning**

We are given two datasets which contains data pertaining to 2018 elections in American counties.The first dataset, *electrion_train* provides an overview on the dataset, whereas the second dataset, *demographics_data* provides a detailed look at the collected data. Since the two datasets needed to be merged to carry out further analysis, reshaping was necessary. This was achieved by reshaping the *election_data* to the wide format, using **pivot_table** function of pandas. On reshaping the *election_data*, we got a dataset of *1205 rows X 6 columns*.

- **Merging the two datasets**

Before merging the datasets however, we have to make sure the data is consistent. The first inconsistency is present in the column State. We created a dictionary with the key as the acronym for the state and the values are the full form of the states and stored it in a file named *state_map*. Since *election_data* and *demographics_data* had State names written in the 2-letter code and actual state name respectively, we made the two data frames consistent by replacing the 2-letter codes with the actual name. For example, AZ turned into Arizona and WY turned into Wyoming. Now, the State column was consistent among the two data frames.

- **Handling missing values**

After merging the datasets on *State* and *County* columns, we observed that there are 21 variables and the type of these variables are object, int64 and float64. There are also irrelevant or redundant variables in the dataset. Year has a value of only 2018. Office has a value of 'US Senator' only. More than 50% of Citizen Voting-Age Population has missing values filled with 0. Hence, these are irrelevant/redundant variables. We deleted the Year, Office and Citizen Voting-Age Population column and inserted the year 2018 and US Senator in the table header.
There are missing values in Democratic and Republican columns. We removed the 5 entries of Democratic and Republican since changes in small number of observations won't impact the data analysis.

- **Analysis**

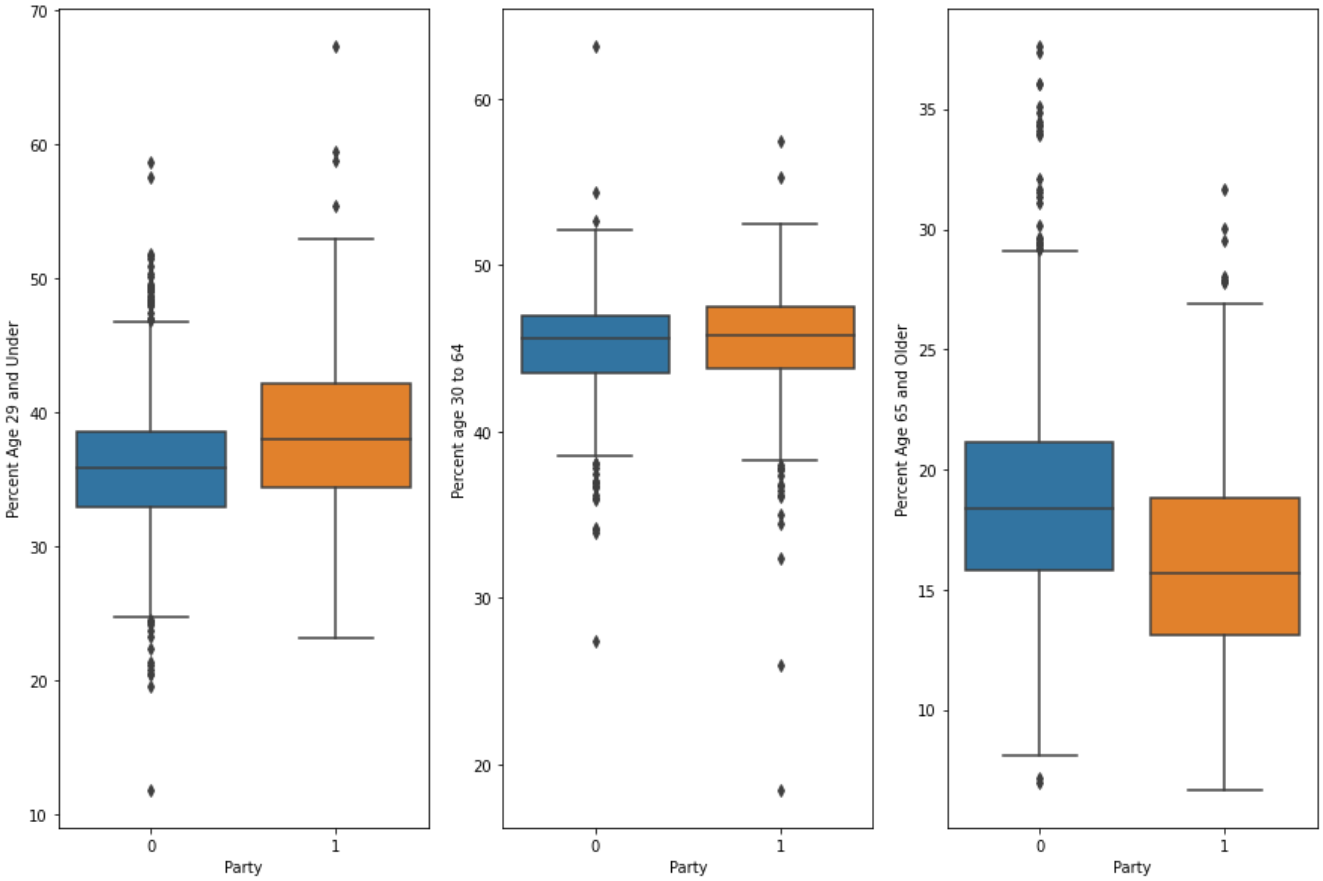
We created a variable named *Party* and assigned it a value of *1 if the county received more Democratic* votes than Republican and a *value of 0 if the county received more Republican votes* than Democratic, we calculated the median household income for Democratic and Republican values based on their Party variable value being either 1 or 0. The mean median household income for Democratic counties came out to be higher.
However, since p-value was less than the significance value 0.05, we had sufficient evidence to reject the null hypothesis. We did the same analysis to calculate the mean population of the Democratic and Republican counties. The mean population came out to be higher for

Republican Counties. However, since p-value is less than the significance value 0.05 we had sufficient evidence to reject the null hypothesis.

- **Plotting the results**

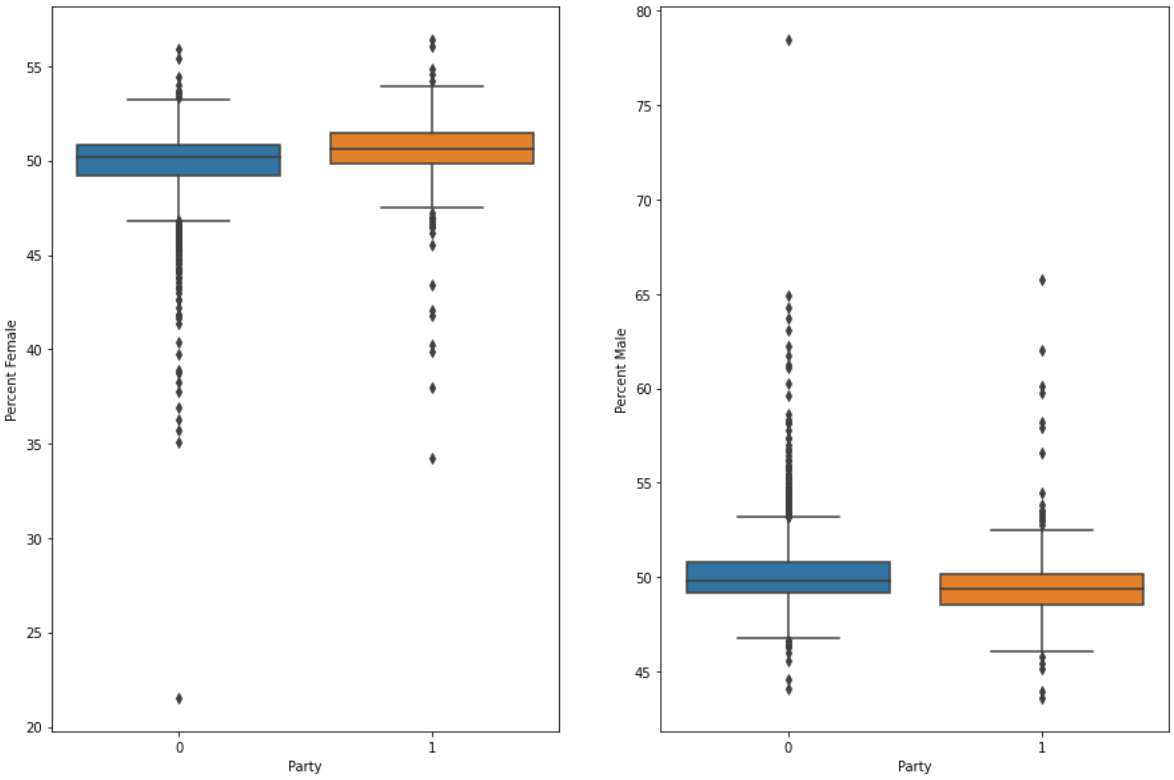
We compared the Democratic and Republican counties in terms of age, gender, race and ethnicity, and education by computing the descriptive analysis and visualizing the results by creating plots.

- ***Comparison of Republican and Democratic Counties in terms of Age***:



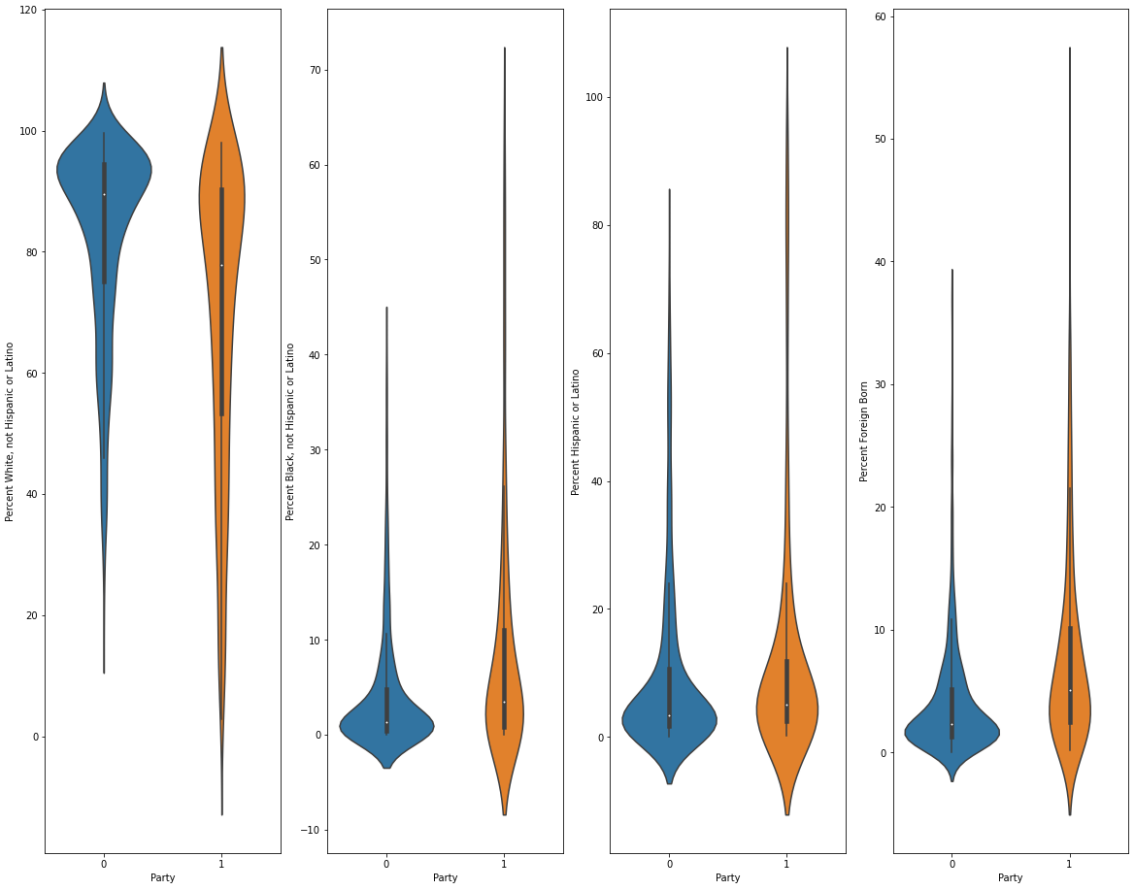
Although there is not much of a difference. By observing the descriptive statistics & plots we conclude that counties with 'Percent Age 29 and Under' are democratic and counties with 'Percent Age 65 and Older' are Republicans counties. We also see that counties under 'Percent age 30 to 64' are almost equally divided having the same mean.

- ***Comparison of Republican and Democratic Counties in terms of Gender:***



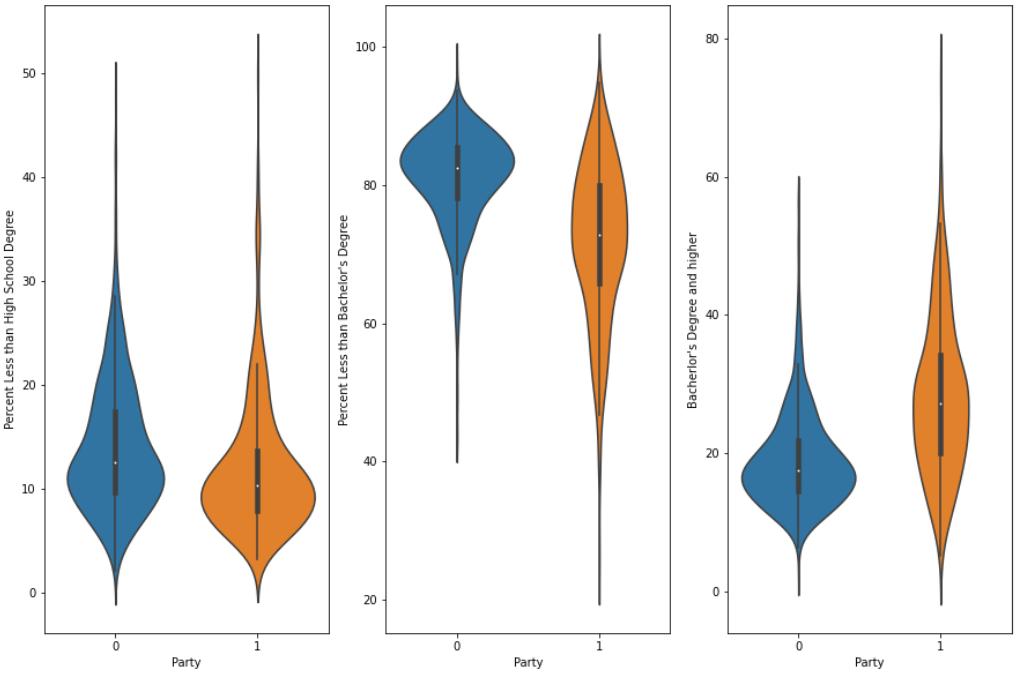
We can see that the mean of 'Percent female' & 'Percent male' voting are very close, although its more in the case of Democrats in 'Percent female' and republicans in case of 'Percent male'. We cannot conclude any county to be republican or democratic just on the basis of gender.

- ***Comparison of Republican and Democratic Counties in terms of ethnicity:***



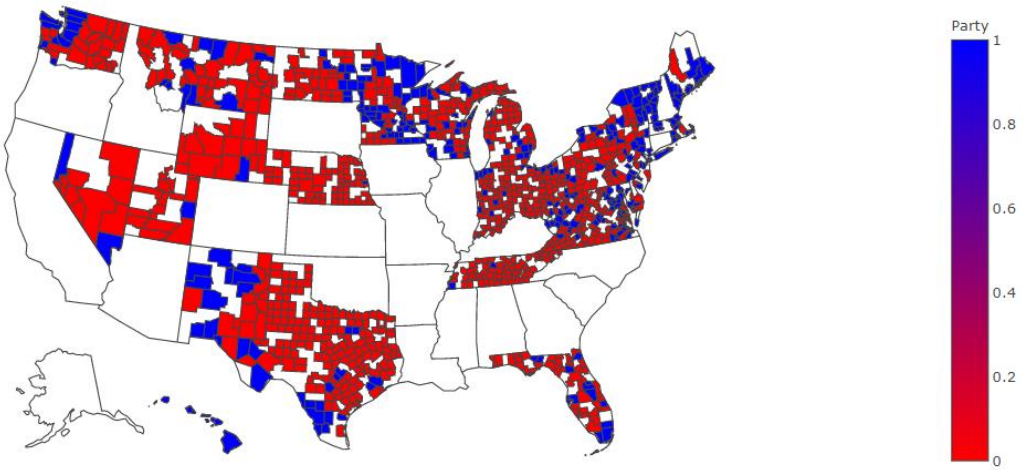
In case of Percent white, not Hispanic or Latino we can see that the counties with a greater number of them are republican counties whereas in the case of all other three categories counties with higher number of other ethnicities are more inclined towards Democrats.

- ***Comparison of Republican and Democratic Counties in terms of education:***



Counties with more percent of 'Percent Less than High School Degree' and 'Percent Less than Bachelor's Degree' people are inclined towards Republicans. Counties with more percent of people having 'Bachelor's Degree and higher' are inclined towards Democrats.

- ***Map of Democratic and Republican counties***



The variable that we think was more important than others to determine whether a county is labelled as a Democratic or a Republican was the *Total Population*. This is because the mean population of democratic counties is a lot higher than the republican counties which means the higher total population counties are inclined towards Democrats. *Education level (Percent Less than Bachelor's Degree)* and *Ethnicity* are also important variables because according to the plots the values of democrats and republicans in these variables vary a lot.