Project 01: Exploratory Data Analysis

By: Shubham Chaudhary, Kirun Haque, Michelle Baginski

Questions:

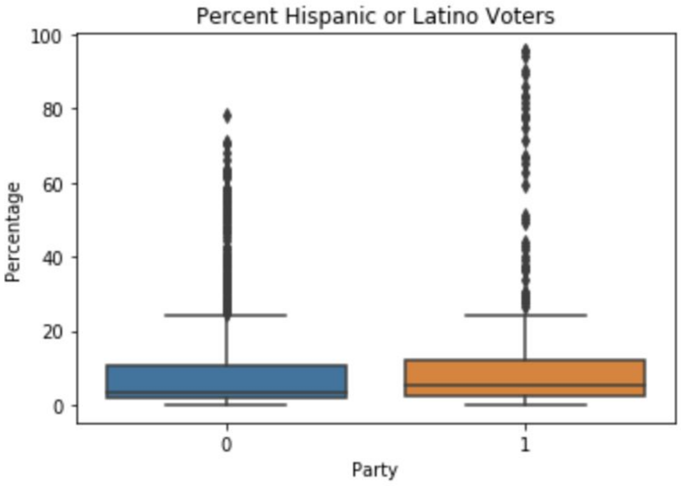
1. The merged dataset has 21 variables. The variable 'Year' is redundant because the entire dataset contains data solely for 2018. 'Office' is an irrelevant variable because all the cells for that variable contains the same value of "US Senator." The columns for 'Office' and 'Year' can be removed as they hold fixed values for the entire dataset and seem better fitted for a title or other information rather than a variable.

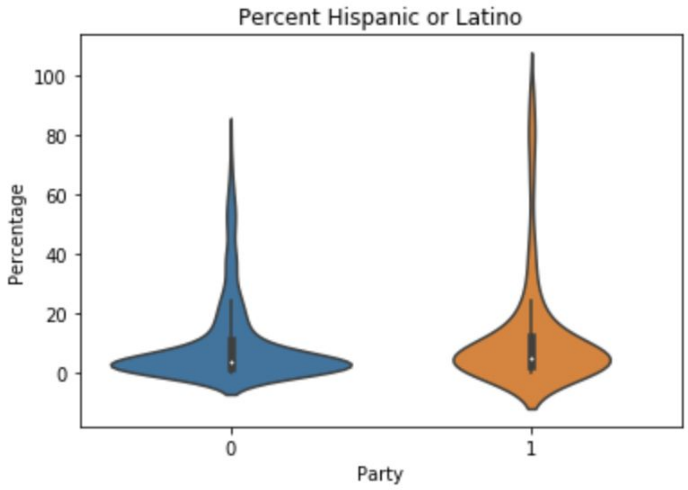
2. There are missing values for several variables. 'Citizen Voting-Age Population' has a significant amount of 680 missing values. Since roughly half of the values for this variable are missing, it is best to drop the column since estimating with only half of the data collected would result in inaccurate values. The variables 'Democratic' and 'Republican' each contain 5 missing values which both appear in the same observation. Since we cannot determine which party the votes should be placed in and the number missing is low, we can remove these rows.
'Percent Black, not Hispanic or 'Latino' has 45 missing values. These values may be kept in the dataset because it is a possibility that certain counties did not have any black voters. They should be estimated and replaced with the mean.
'Percent Unemployed' contains 3 missing values which can be removed.
'Percent Foreign Born' contains 3 missing values which can be removed.
'Percent Rural' contains 19 missing values which can be kept because some of the counties the data was collected in may be entirely urban and contain no rural voters.
'Percent Hispanic or Latino' contains 5 missing values which can be removed.

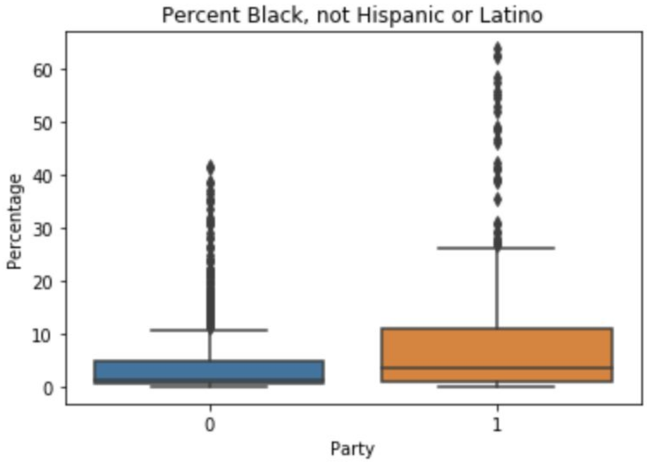
3. For task 6, our null hypothesis was that the mean population of the Democratic and Republican parties were equal to each other. The alternative hypothesis was that the two were not equal to one another. The mean population of the Democratic party was calculated as 30,998 and the mean population of Republican counties was calculated as 54,354. After calculating the test statistic, we found it was -7.988. After calculating the p-value, we found it to be $1.14 * 10^{-14}$ which is statistically significant, since it is less than the value of alpha at .05. Thus, we reject the null hypothesis.

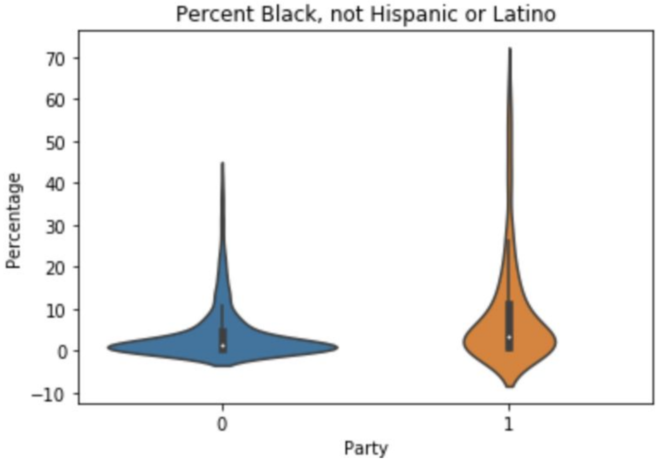
4. For task 7, our null hypothesis was that the mean for the Median Household Income of both the Democratic and Republican parties were equal to each other. The alternative hypothesis was that the two were not equal to each other. The mean for the Median Household income for the Democratic party was calculated as 53,798 and the Median Household Income of the Republican party is 48,724. After calculating the test statistic, we found it was 5.507, and the p-value to be 3.087^{-8} which is statistically significant, since it is less than the value of alpha at .05. Thus, we reject the null hypothesis.

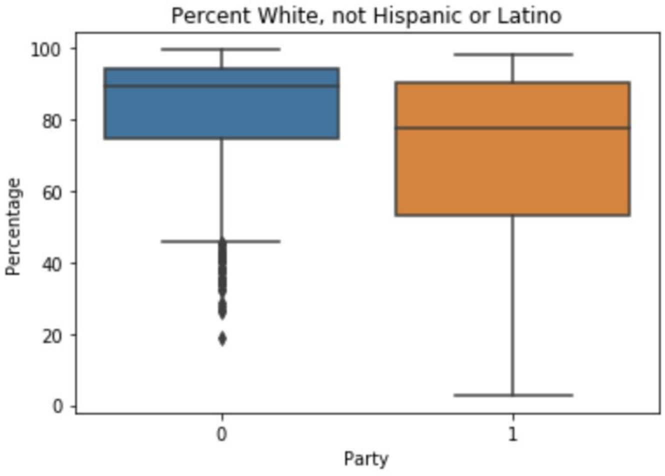
8. Plots and Analysis

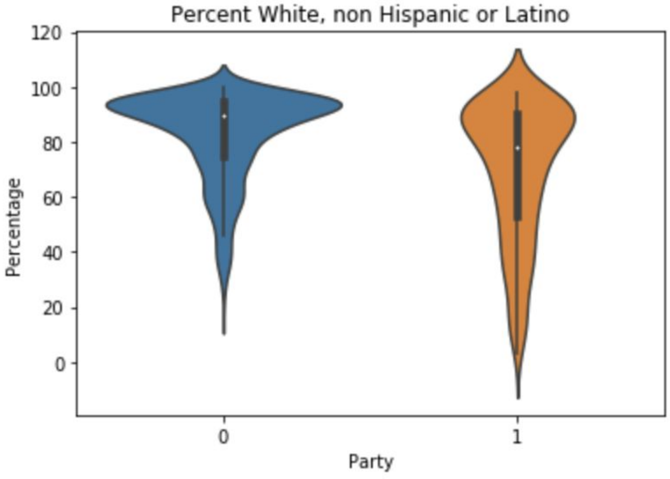


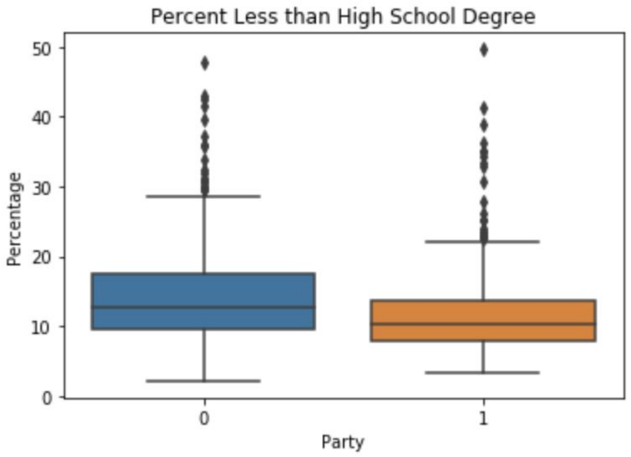


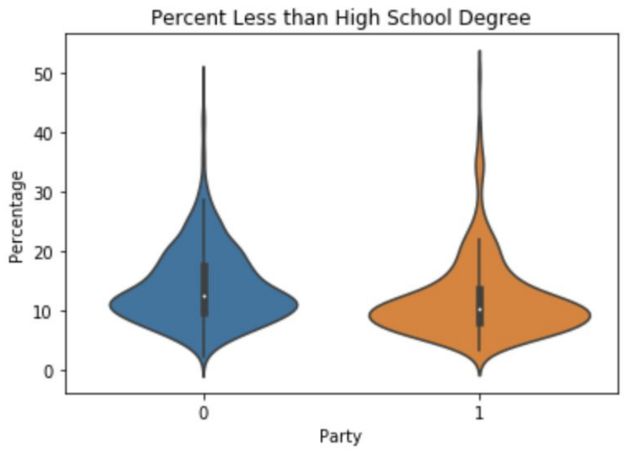


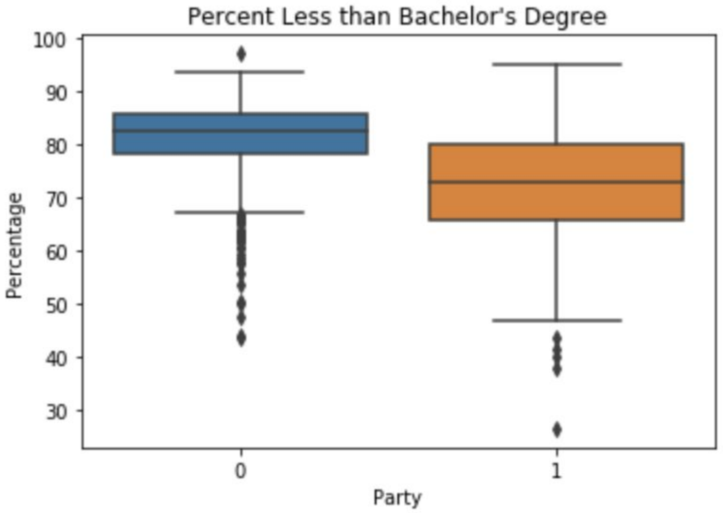


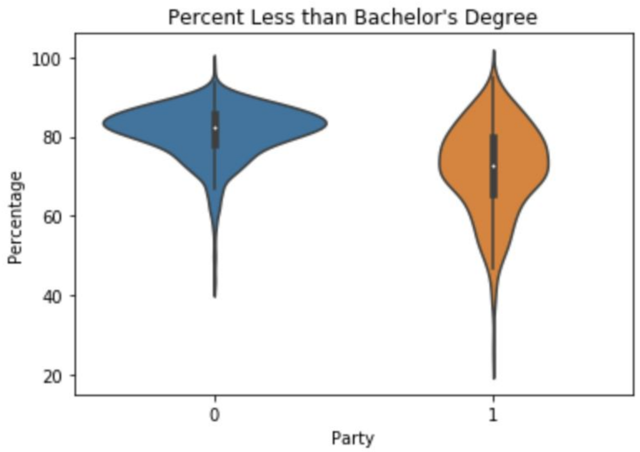


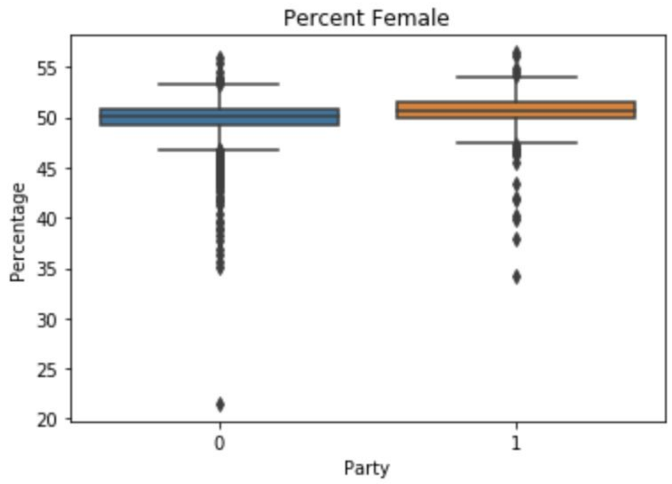


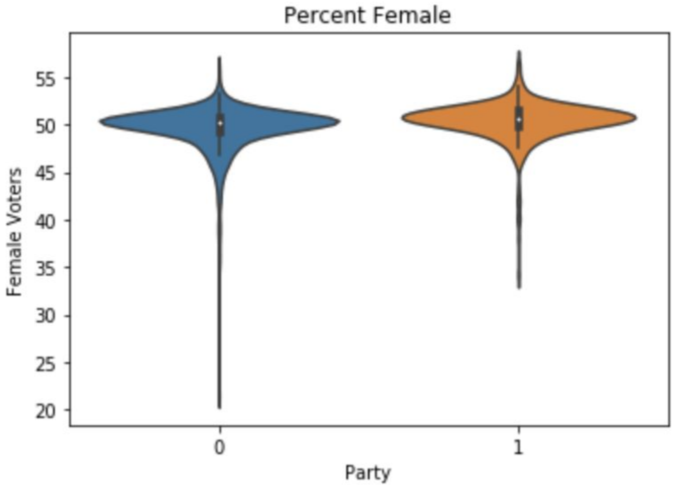

Percent Less than High School Degree

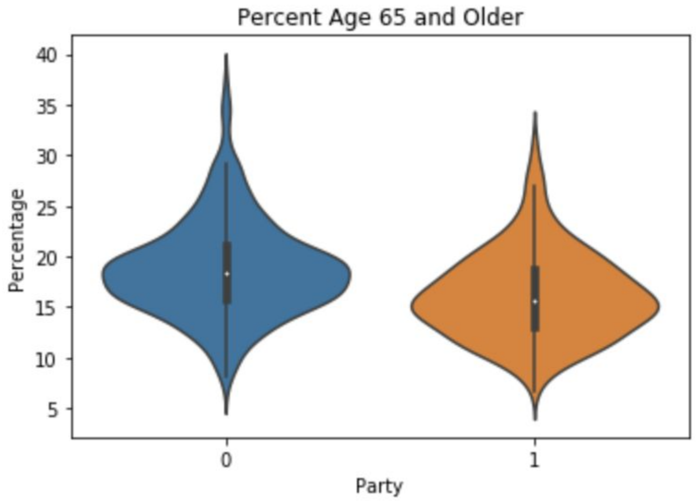

Percent Less than High School Degree

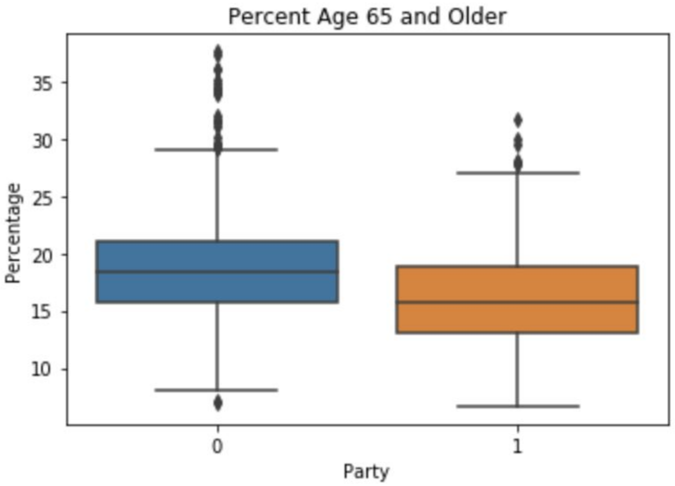

Percent Less than Bachelor's Degree

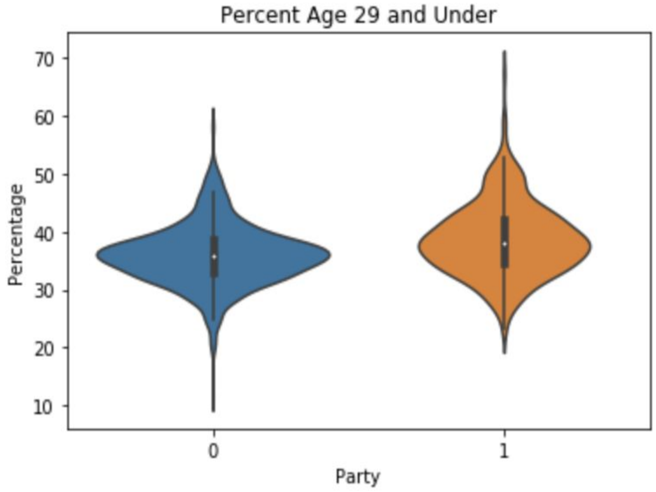

Percent Less than Bachelor's Degree

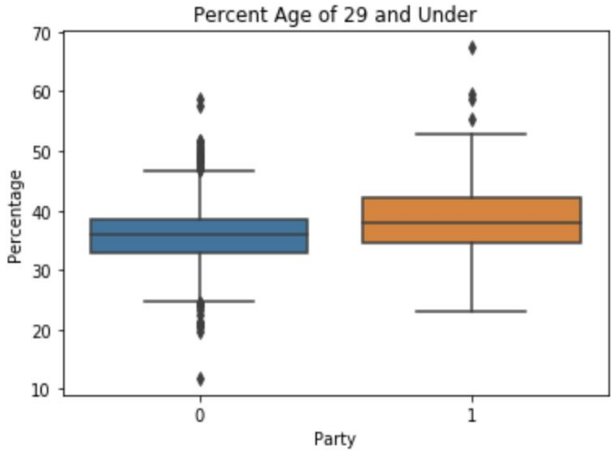

Percent Female


Percent Female









Race & Ethnicity

The median for White Republican voters is higher than that of White Democratic voters. Based on the density for Republican voters, the probability of white voters voting for the Republican party is higher than that of the democrats.

Gender

The gender for males and females in both Democratic and Republican parties observes a similar mean and median value. The gender appears to be evenly split between the two parties. The majority of the Republican and Democratic population lies in the range of 47 to 54. The Republicans contain more outliers

Education

In 50% of the Republican counties that voted, we observed the majority 82% of the people who voted had less than a bachelor’s degree. however 50% of the Democratic counties had the majority 70% of the votes from people who had less than a bachelor’s degree.

In 50% of the Republican counties that voted, we observed that the majority of 12% of the people that voted had less than a high school degree, however 50% of the Democratic counties had the majority 10% of the votes from people who had less than a high school degree

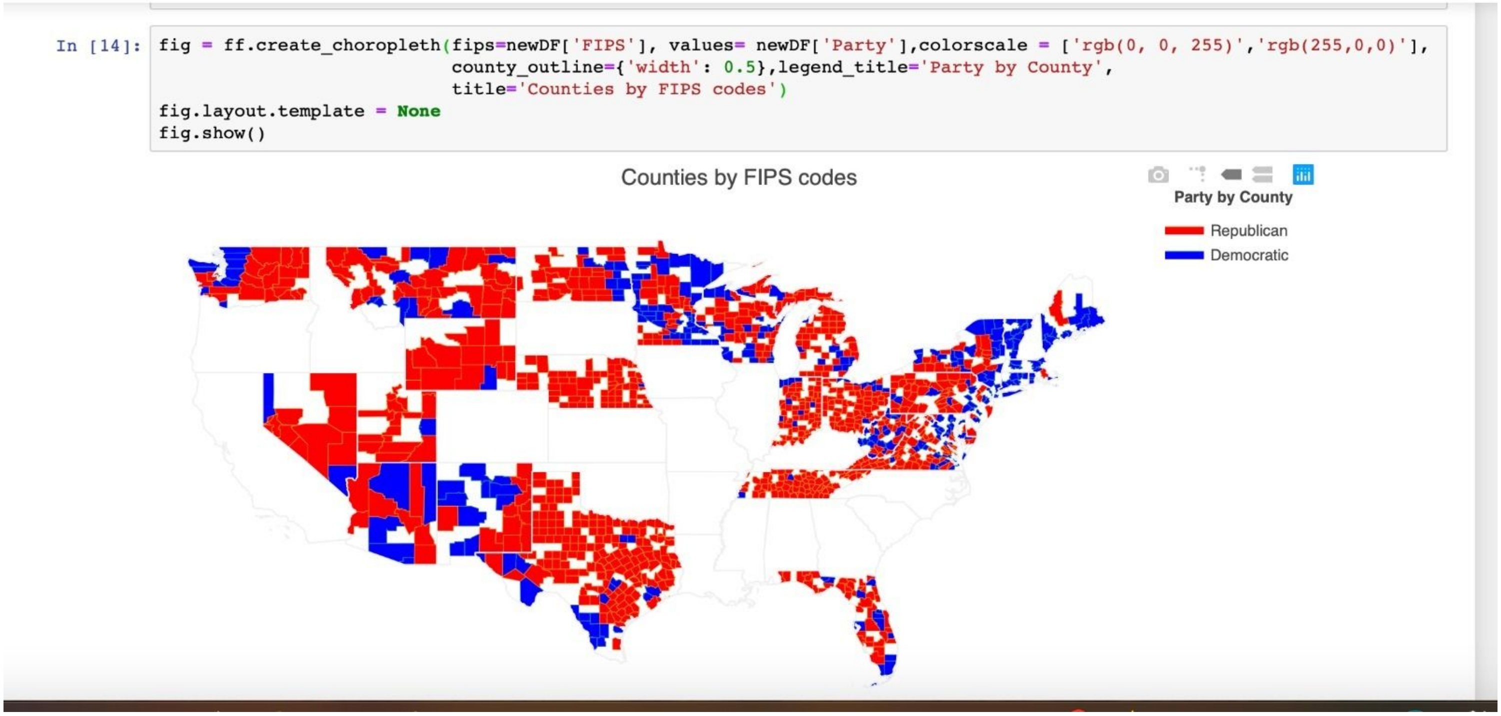
Age

Those of age 65 and older tend to vote as Republican because of the percentage distribution for Republicans is greater than that of the Democrats. Those who were 29 and under tend to vote for Democrats more than Republicans.

9.

Percent White and Percent less than Bachelor's degree are the most important variables because the percentage distribution of the observation is well defined, as well as the difference in the median. Also, in Percent White has fewer outliers in comparison to other variables.

10.



```
In [14]: fig = ff.create_choropleth(fips=newDF['FIPS'], values= newDF['Party'],colorscale = ['rgb(0, 0, 255)','rgb(255,0,0)'],
                                    county_outline={'width': 0.5},legend_title='Party by County',
                                    title='Counties by FIPS codes')
         fig.layout.template = None
         fig.show()
```