Project 1: Exploratory Data Analysis

(Conor Harrigan, Deblina Das, Paul Dougherty)

1) Reshape dataset election_train from long format to wide format. Hint: the reshaped dataset should contain 1205 rows and 6 columns.

First, we checked if there are missing values in any columns. The result obtained were 10 rows with NaN for the column Votes. Replaced the missing values with 0.

We have used pivot_table for converting long to wide format on the column Party and the Values is Votes. The number of rows returned are 1205 and columns are 6.

Converted the 0's back to NaN.

2) Merge reshaped dataset election_train with dataset demographics_train. Make sure that you address all inconsistencies in the names of the states and the counties before merging. Hint: the merged dataset should contain 1200 rows.

Created a dictionary of US States and abbreviations. Used merge function to merge the two datasets. Some preprocessing is done on County column which includes removal of the term ' County' and converting it to lower case. The State column in demographics_train dataset was mapped to US state abbreviation and the values in County variable is converted to lowercase.

We are obtaining 1200 rows and 25 columns.

Note: The number of columns of the merged dataset might vary depending on the preprocessing of the columns. Other team members got different number of columns. But, the number of rows remains the same.

3) Explore the merged dataset. How many variables does the dataset have? What is the type of these variables? Are there any irrelevant or redundant variables? If so, how will you deal with these variables?

There are 25 variables. The datatypes are float64, int64 and object. Yes, there are redundant variables. Variables like State, key_1, State_x, County_x, State_y and County_y is redundant. Only one variable for State and County will be kept and the rest will be dropped. Also, columns Year and Office can be dropped since the values are constant in all the observations.

4) Search the merged dataset for missing values. Are there any missing values? If so, how will you deal with these values?

There are 5 observations with missing values in the columns Democratic and Republican. Since, the number of such observations is less hence we can drop these observations. Also, another reason to drop these rows could be that the column values on which we are doing analysis are missing hence, there is no sense to keep them.

Citizen Voting-Age Population variable is 0 for 680 observations. We cannot drop these observations as that will result into dropping maximum data. Hence, we will drop the variable itself.

After dropping redundant variable and removing observations we are left with 1195 rows and 18 columns.

5) Create a new variable named "Party" that labels each county as Democratic or Republican. This new variable should be equal to 1 if there were more votes cast for the Democratic party than the Republican party in that county and it should be equal to 0 otherwise.

Created a new column Party by comparing the values of 2 columns: Democratic and Republican. The value of new variable value is 1 if votes cast for Democratic > votes Cast for Republican else 0.

6) Compute the mean population for Democratic counties and Republican counties. Which one is higher? Perform a hypothesis test to determine whether this difference is statistically significant at the $\boldsymbol{\alpha} = \mathbf{0.05}$ significance level. What is the result of the test? What conclusion do you make from this result?

Mean population for Democratic counties = 300998.316923

Mean population for Republican counties = 53864.672414

Mean population of Democratic counties is greater than Republican counties.

p-value is **2.0478717602973023e-14**.

Here, Ho : μ1 = μ2

Hα : μ1 ≠ μ2

where μ1 is mean total population of Democratic counties

μ2 is mean total population of Republican counties

Since p-value < significance level hence, we reject the null hypothesis and hence, μ1 ≠ μ2

7) Compute the mean median household income for Democratic counties and Republican counties. Which one is higher? Perform a hypothesis test to determine whether this difference is statistically significant at the $\boldsymbol{\alpha} = \mathbf{0.05}$ significance level. What is the result of the test? What conclusion do you make from this result?

Mean median household income for Democratic counties = 53798.732308

Mean median household income for Republican counties = 48746.819540

Mean median household income of Democratic counties is greater than Republican counties.

p-value is **7.149437363182598e-08**.

Here, Ho : μ1 = μ2

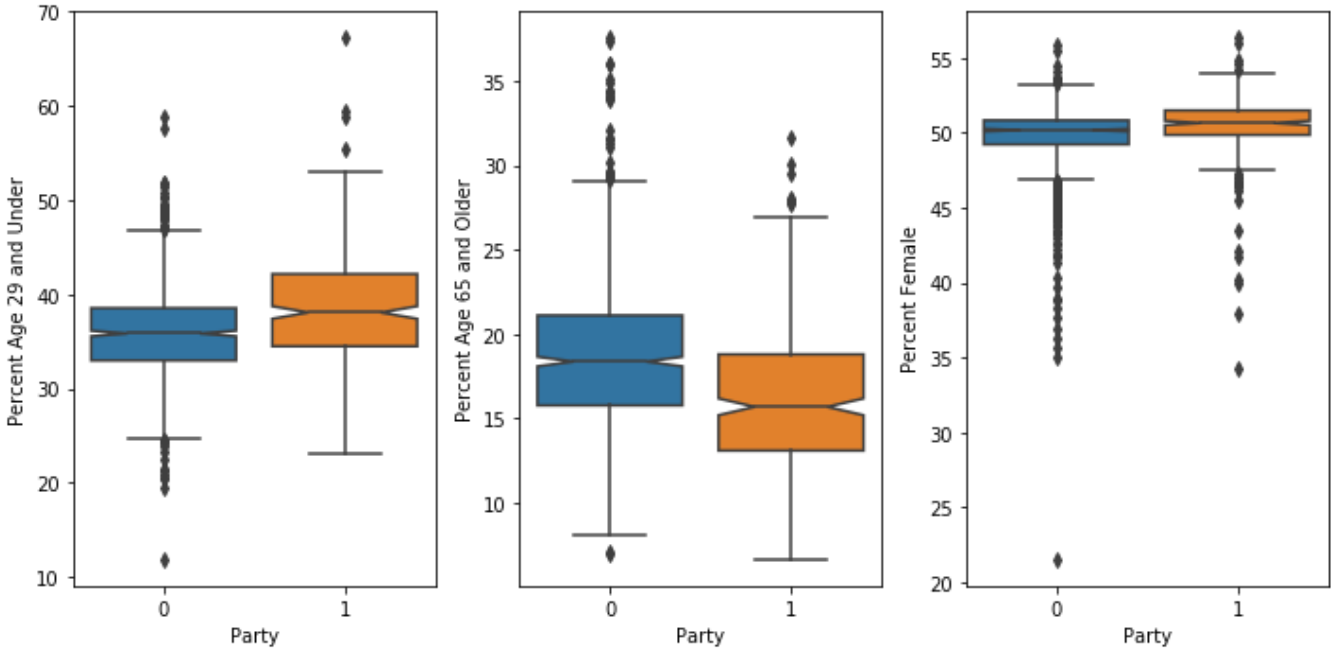
Hα : μ1 ≠ μ2

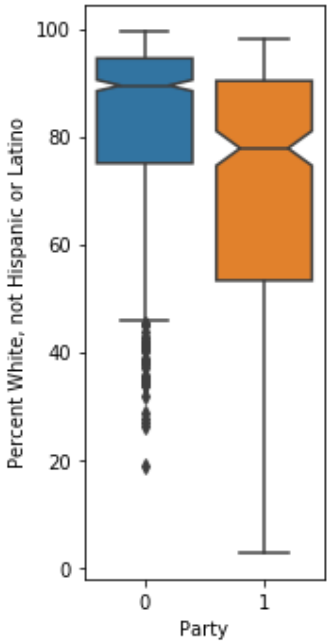
where μ1 is mean median household income of Democratic counties

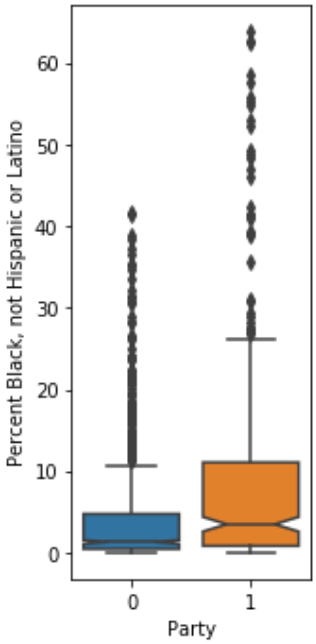
μ2 is mean median household income of Republican counties

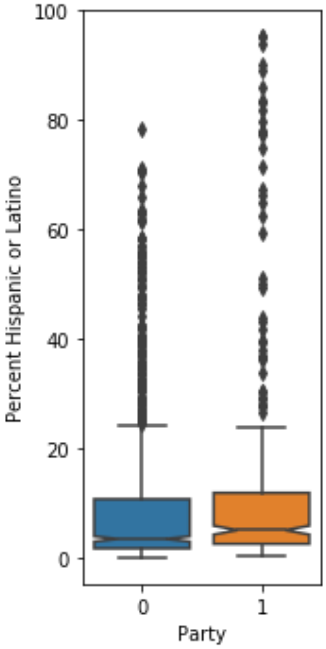
Since p-value < significance level hence, we reject the null hypothesis and hence, μ1 ≠ μ2

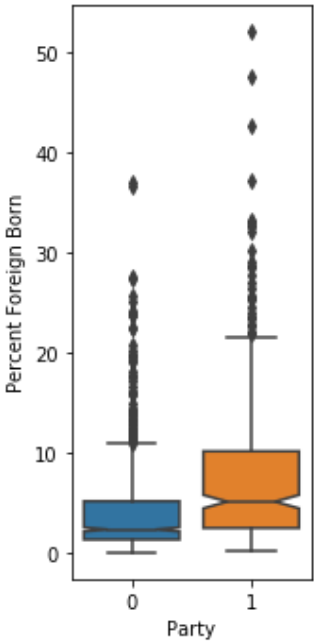
8) Compare Democratic counties and Republican counties in terms of age, gender, race and ethnicity, and education by computing descriptive statistics and creating plots to visualize the results. What conclusions do you make for each variable from the descriptive statistics and the plots?

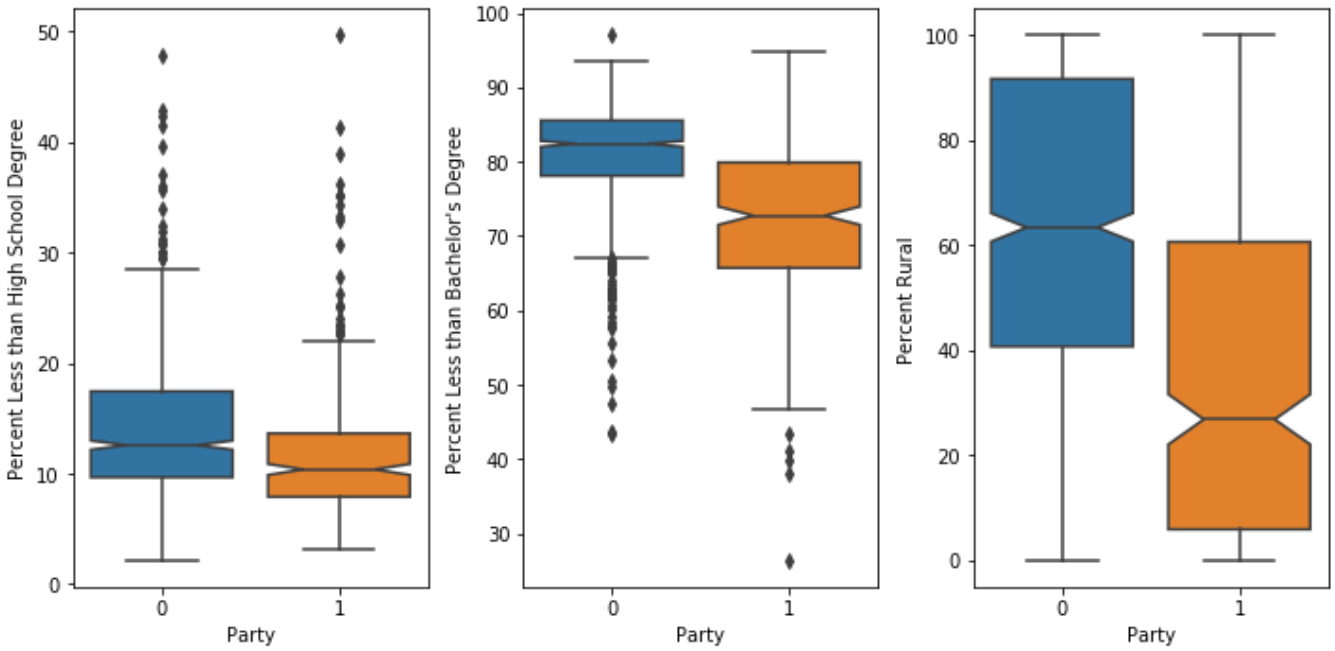












	Republican (0)	Democratic (1)	Remarks
Percent Age 29 and Under	larger outliers	greater variability	medians do not overlap. since the notches in the boxplots do not overlap, we can conclude that with 95% confidence, the true medians differ
Percent Age 65 and Older	larger outliers	Median less than republican	Same as above
Percent Female	larger outliers	larger outliers	Medians slightly overlap
Percent White, not Hispanic or Latino	larger outliers	Greater variability	Medians don't overlap
Percent Black, not Hispanic or Latino	larger outliers	larger outliers	Prefer democrats
Percent Hispanic or Latino	larger outliers	larger outliers	Medians slightly overlap
Percent Foreign Born	larger outliers	larger outliers	Prefer democrats
Percent Less than High School Degree	larger outliers	larger outliers	Prefer republicans
Percent Less than Bachelor's Degree	larger outliers	larger outliers	Prefer republicans
Percent Rural	Greater variability	Greater variability	Prefer democrats

Conclusion:

It appears that democrats tend to be younger, marginally more female, less white and more blacks, more educated with bachelor's degree and high school degree, more foreign born, and less rural than republicans.

Republicans tend to be more whites (Percent White, not Hispanic or Latino) and aged people who are 65 and above.

Percent Hispanic does not appear to vary between the two, and Percent Female, as well as Aged 29 and Under vary quite slightly. Although, if we take smaller details into consideration, then "Percent Black, not Hispanic or Latino", "Percent Hispanic or Latino" prefer democratic party. All these conclusions have been inferred by considering mean, median and values in quartile 3 and 4 for both the counties.

9) Based on your previous analysis, which variables in the dataset do you think are more important to determine whether a county is labeled as Democratic or Republican? Justify your answer.

Size of county and median household income have already shown themselves to be significant. Percent Rural also appears to vary greatly between the two (although it is related to size of county and median income). We can also observe that the number of counties where Republicans have more votes (870) is higher than counties where Democrats have more votes (325).

The other important variables are:

- Percent Less than Bachelor's Degree
- Percent White, not Hispanic or Latino
- Percent Less than High School Degree

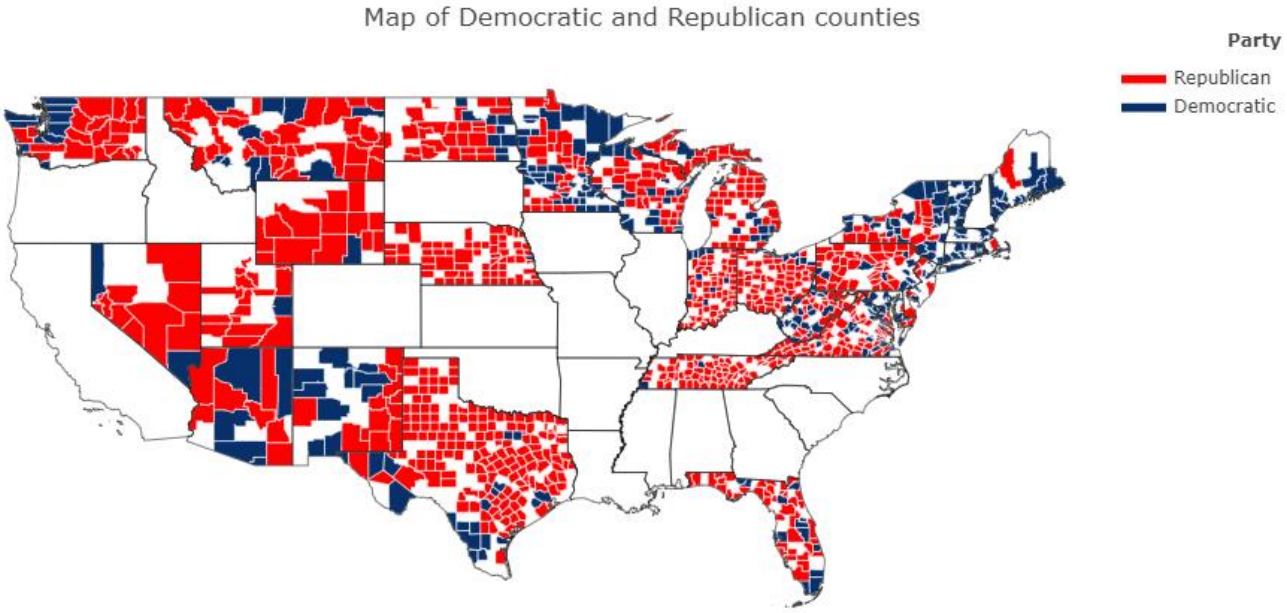
The reasons to coin these variables as important are:

- the median and descriptive statistics range of population under this category is higher than rest of the other categories
- the median of the box-plots for both parties are not overlapping and hence its clearly visible which category people are favoring which party.
- Significant variation is observed in the descriptive stats

"Percent Age 29 and Under" and "Percent Female" are the variables which do have effect given the quartile values. However, the medians are almost overlapping and hence it's difficult conclude which parties they favored.

Overall, Republican Counties have higher values (Percentage values) when compared to Democratic Counties which can be also observed in Task 10.

10) Create a map of Democratic counties and Republican counties using the counties' FIPS codes and Python's Plotly library (plot.ly/python/county-choropleth/). Note that this dataset does not include all United States counties.



From the map we can see majority of counties are Republicans which was also inferred in the previous tasks.