# CS 418: Project-1 Report

Team 18: Ashwin Deshpande, Apurva Raghunath, Nandana Shimoga Prasad

**Task 1 (5 pts.)** Reshape dataset election_train from long format to wide format. Hint: the reshaped dataset should contain 1205 rows and 6 columns

**Solution**:

- We use pivot table so that every two rows where 'State' and 'County' are the same are converted to a single row.
- Each value for 'Votes' in the previous two rows are converted to two variables in the new merged row.

**Task 2 (20 pts.)** 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.

**Solution**:

- *County* in election_train has the word 'County' whereas the demographics_train has only the name of county.
- We will therefore remove word county in election_train to match the *County* names in demographics_train
- *State* in election_train is abbreviated whereas demographics_train has full state name.
- We will therefore replace the state abbreviation in election_train to its full state name to match the *State* in demographics_train.

We can now merge election_train_tidy & demographics_train such that 'State' and 'County' together are unique in the new table.

**Task 3 (5 pts.)** 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?

**Solution**:

- Number of variables: 21
- Types of variables: object, int64, float64
- Redundant Variables: *Year* and *Office* are redundant across all observation and thus do not give any information about a county. So, we can drop these variables.

**Task 4 (10 pts.)** Search the merged dataset for missing values. Are there any missing values? If so, how will you deal with these values?

**Solution**:

- In the merged dataset there are 5 observations with missing values for *Democratic* or *Republican* votes: we will be ignoring these observations
- *Citizen Voting-Age Population* has implicit missing values for 675 out of 1200 observations. We will be dropping this variable.
- We have zero values in *Percent Black, not Hispanic or Latino* and *Percent Rural* but these contain useful information and are not treated as missing values.

**Task 5 (5 pts.)** 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.

**Solution**: We can see that Party contains 1 when Democratic > Republican and 0 otherwise.

| State | County | Democratic | Republican | . . . . | Party |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Arizona | apache | 16298 | 7810 | .... | 1 |
| Arizona | cochise | 17383 | 26929 | .... | 0 |
| Arizona | coconino | 34240 | 19249 | .... | 1 |
| Arizona | gila | 7643 | 12180 | .... | 0 |
| Arizona | graham | 3368 | 6870 | .... | 0 |

Top 5 rows of the dataset after creating the new variable named 'Party'

**Task 6 (10 pts.)** 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 $\alpha=0.05$ significance level. What is the result of the test? What conclusion do you make from this result?

**Solution**:

Mean median houshold income:

$$\mu_{\text{Democratic Counties}} = 53798.732307692306$$
$$\mu_{\text{Republican Counties}} = 48746.81954022989$$

Therefore, mean median household income for democratic counties is *greater* than mean population for republican counties.

$$H_0 : \mu_{\text{Democratic Counties}} = \mu_{\text{Republican Counties}}$$
$$H_a : \mu_{\text{Democratic Counties}} \neq \mu_{\text{Republican Counties}}$$
$$\text{t-test} = 5.479141589767387$$
$$\text{p-value} = 7.149437363182598e\text{-}08$$

Since p-value< α, we reject $H_0$(Null hypothesis).

There is sufficient evidence to conclude that the mean median household income of democratic counties is different from republican counties.

**Task 7 (10 pts.)** 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 $\alpha=0.05$ significance level. What is the result of the test? What conclusion do you make from this result?

**Solution**:

Mean population:

$$\mu_{\text{Democratic Counties}} = 300998.3169230769$$
$$\mu_{\text{Republican Counties}} = 53864.6724137931$$

Therefore, mean median household income for democratic counties is *greater* than mean population for republican counties.

$$H_0 : \mu_{\text{Democratic Counties}} = \mu_{\text{Republican Counties}}$$
$$H_a : \mu_{\text{Democratic Counties}} \neq \mu_{\text{Republican Counties}}$$
$$\text{t-test} = 8.004638577960957$$
$$\text{p-value} = 2.0478717602973023e\text{-}14$$

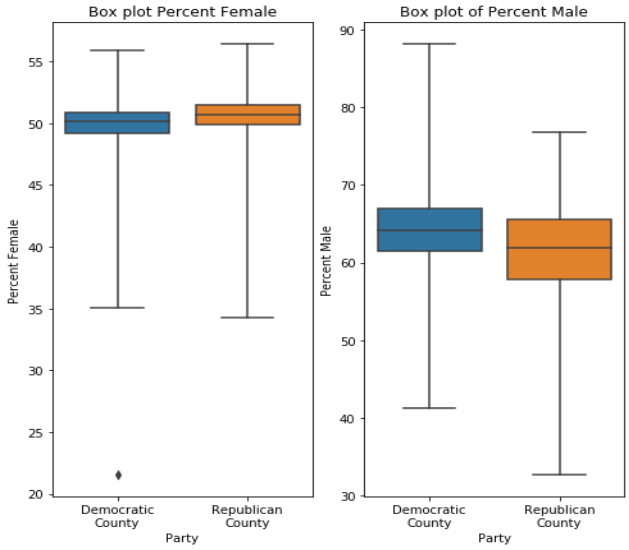
Since p-value< α, we reject $H_0$(Null hypothesis). There is sufficient evidence to conclude that the mean population of democratic counties is different from republican counties.

**Task 8 (20 pts.)** 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?

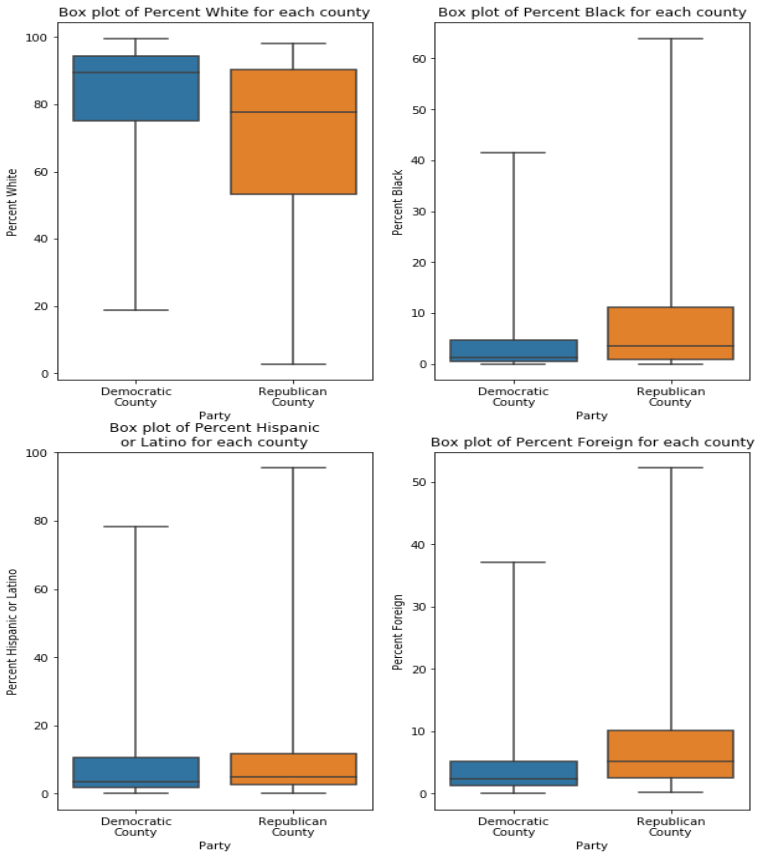
**Solution**:

- We group observations by Party=1 and Party=0.
- We describe count, mean, standard deviation, minimum, 1st Quartile, Median, 3rd Quartile and maximum for each age, gender, race and ethnicity and education variables for the two groups.

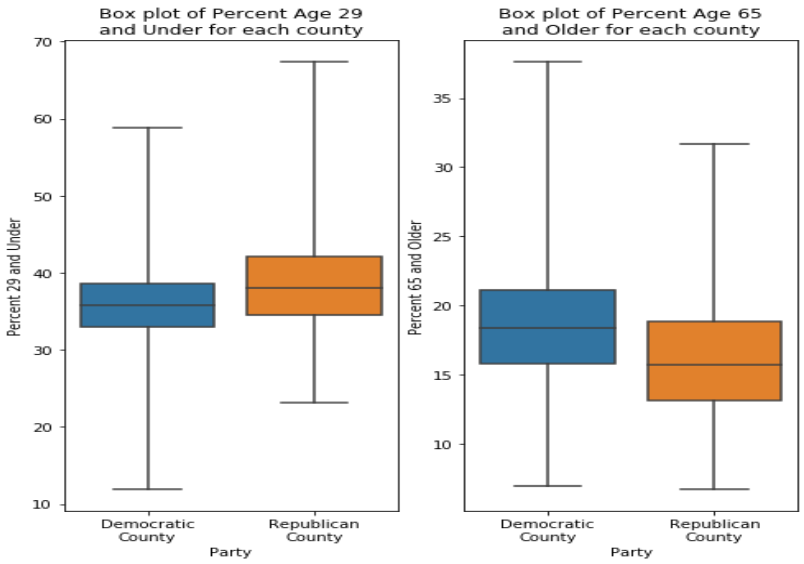
1. Gender:



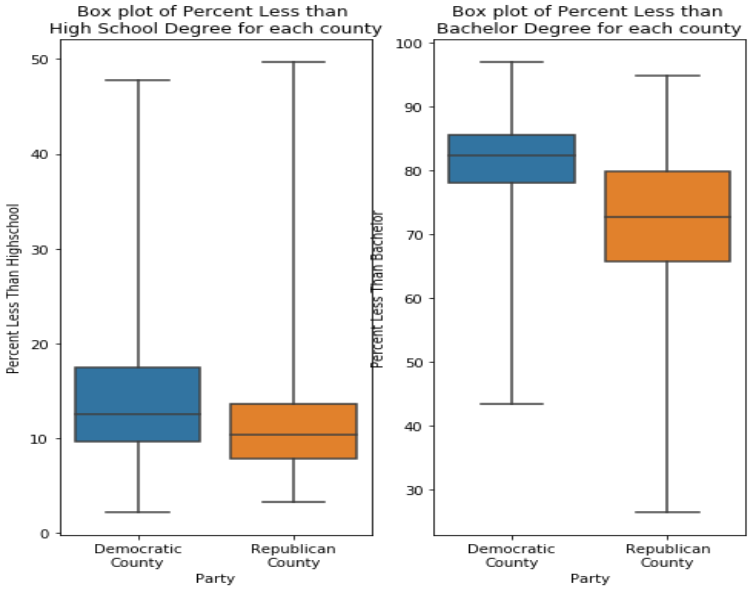

2. Ethnicity & Race




3. Age




4. Education




Conclusions:

- Gender - The number of males and females are equal for both parties.
- Race and Ethnicity - White is the dominant race for both parties.
- Age - Younger people tend to be more politically active than older people in both parties.
- Education - People with more education tend to be involved with both parties more than people with less education.

**Task 9 (5 pts.)** Based on your results for tasks 6-8, 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.

**Solution**:

- The variables *Race and Ethnicity*, *Education and Age* are important to determine whether a county is marked Democratic or Republican.
- The Percent White make up about
  - 83% of Democrats
  - 70% of Republicans
- In Education
  - 72 % of Democrats have a degree less than a bachelor's degree
  - 81 % of Republicans have a degree less than a bachelor's degree.
- In age
  - 39 % of Republicans are 29 and younger
  - 36 % of Democrats are age 29 and younger.

**Task 10 (10 pts.)** 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.