PROJECT 01 REPORT

Sai Bharath Kumar Bandi, Sachin Dattatraya Hedge, Martin Bautista, Mohmed Hira
CS 418 – Introduction to Data Science

INTRODUCTION:

The objective of this project is to prepare and explore a dataset containing results for the 2018 United States Senate elections and demographic information for United States counties. The given datasets (election_train.csv and demographics_train.csv) are reshaped, cleaned and explored for data analysis. The information gathered or the conclusions made on each task is given below along with code. We have used Python 3 for the given tasks and used Jupyter Notebook for implementation.

TASKS AND ANSWERS:

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

A. The election dataset is read using **pd.read_csv("election_train.csv")** and reshaped on the basis of Party and Votes columns. **pivot_table** function from pandas library is used to reshape the election dataset from long format to wide format.
pd.pivot_table(election,values="Votes",index=["Year","State","County","Office"],columns="Party",aggfunc=np.sum).reset_index()
After reshaping the data, the result contains **1205** observations and **6** variables.

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**.

A. The demographic data and the election data are merged based on 'State' and 'County' variables. For merging, the variable values must be same in both datasets.
The differences in 'County' values in both the datasets are for every value in demographics dataset there is an extra string ' County' for each value and few values are in upper case. The difference in the 'State' values in both the datasets is that the values in election dataset are in short forms.
So, the values in the 'County' variable is striped accordingly and converted into lower cased for similarity in two datasets and the values of 'State' variable in demographics dataset are mapped to their short forms.
After that, the two datasets are merged. The merged dataset contains **1200 observations** and **21 variables.**

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?**

A.

- The merged dataset consists of **21 variables**.
- The types of the variables are – {**object**, **int64**, **float64**}
- The irrelevant variables present in the datasets are **Year** and **Office**

- There are no redundant variables
- The irrelevant variables and redundant variables can be dropped or ignored

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

A. Yes, there are missing values in the dataset. The missing values are assigned with a value '0'. We have different ways to deal with the missing values such as dropping, replacing, estimating and ignoring. In this case we are dropping few observations and a variable.

1. Variable 'Citizen Voting-Age Population' is dropped.
    a. It has 681 missing values out of 1200 values. It has more than 50% missing values.
    b. So, we are dropping the variable because it will not provide a better information with missing values
2. Observations with both Democratic votes and Republican votes as '0' are dropped
    a. As we are analyzing the election results based on the Democratic and Republican votes, observations with Democratic votes as '0' and Republican votes as '0' will not provide any information for analysis. So, we can drop those observations.

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.

A. The variable 'Party' is created on the basis of Democratic and Republican votes by using groupby() function. The value of the 'Party' variable will be '1' if the number of votes for Democratic will be greater than Republican and it will be '0' if the number of votes for Republican will be greater than Democratic.

data_merged["Party"] = np.where(data_merged['Democratic'] >data_merged['Republican'], 1, 0)

6. (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?

A. The mean population for Democratic counties = 300998.3169230769
The mean population for Republican counties = 53864.6724137931

- The mean population for Democratic counties is higher than the mean population of Republican counties

Two-sampled T test is performed to determine whether this difference is statistically significant.
The null hypothesis : Mean_population(Democratic) = Mean_population(Republican)
The alternate hypothesis:
Mean_population(Democratic) != Mean_population(Republican)

Scipy.stats.ttest_ind() function is used to perform the t test. The result is:

t statistic = 8.004638577960957

pvalue = 2.0478717602973023e-14

Since α = 0.05 ➔ **pvalue < α**

So we reject the null hypothesis and conclude that the mean population of Democratic counties is different from the mean population of Republican counties.

7. **(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?**

A. The mean median household income for Democratic counties = 53798.732307692306

The mean median household income for Republican counties = 48746.81954022989

- The mean median household income for Democratic counties is higher than the mean median household income of Republican counties

Two-sampled T test is performed to determine whether this difference is statistically significant.

The null hypothesis :

Mean_median_household_income(Democratic) **=**

Mean_median_household_income(Republican)

The alternate hypothesis:

Mean_median_household_income(Democratic) **!=**

Mean_median_household_income(Republican)

Scipy.stats.ttest_ind() function is used to perform the t test. The result is:

t statistic = 5.479141589767388

pvalue = 7.149437363182572e-08

Since α = 0.05 ➔ **pvalue < α**

So we reject the null hypothesis and conclude that the mean median household income of Democratic counties is different from the mean median household income of Republican counties.

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?**

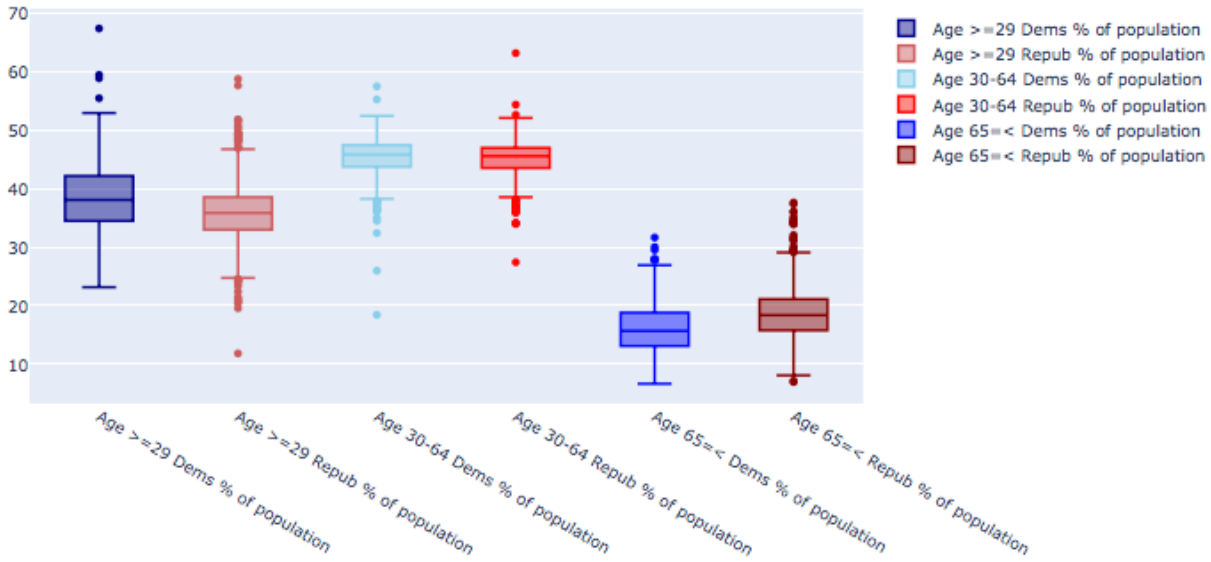
A. The descriptive statistics is calculated by applying describe() function on the respective column.

For age, the percentage of population between age 29 and age 65 is calculated from the variables ‘Percent Age 29 and Under’ and ‘Percent Age 65 and Older’.

For education, the percentage of the population with bachelor’s degree or higher is calculated from the variables ‘Percent Less than High School Degree’ and ‘Percent Less than Bachelor's Degree’.

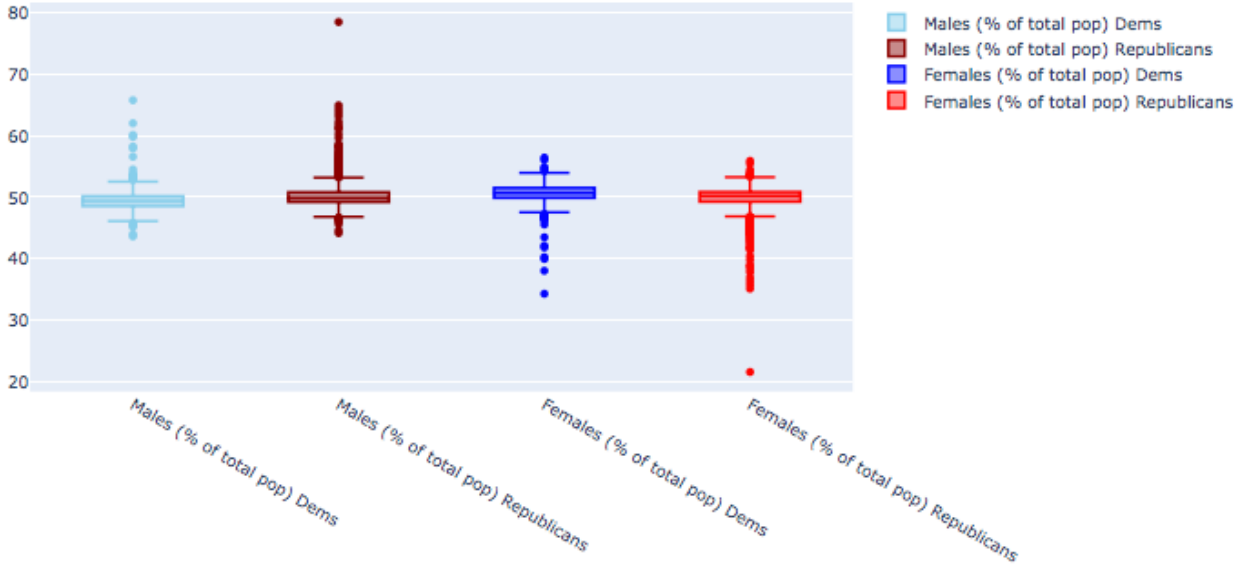
The results are available in the project01.pdf.

Age Comparison:



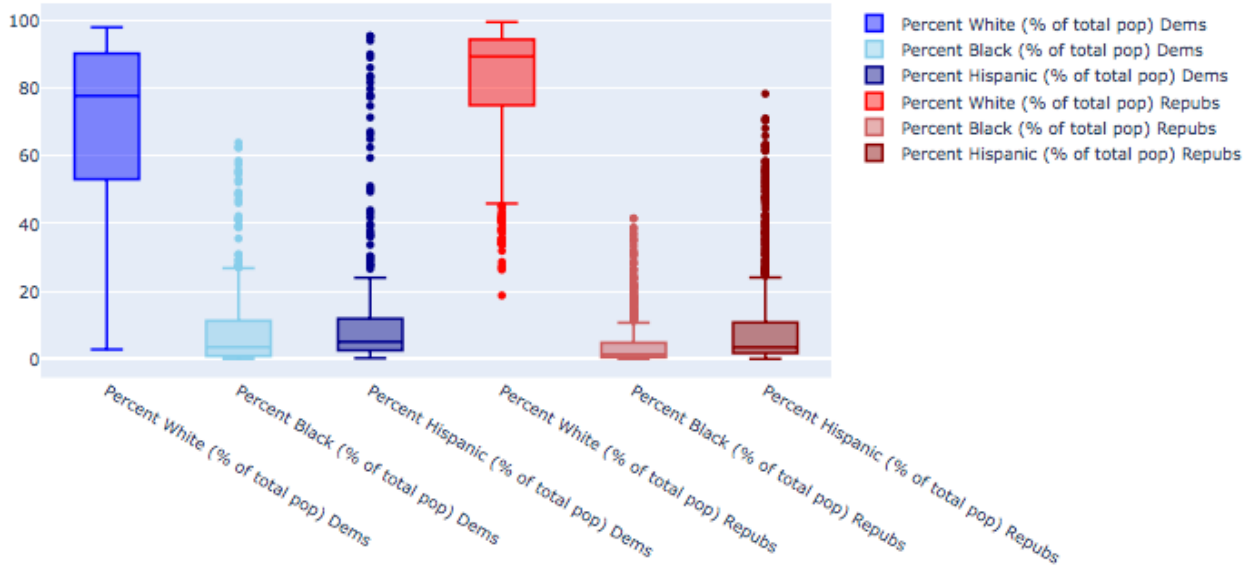
From the graph:
There is no much difference in the median of percentages of population in each age group when we compare between both the parties
Democratic Party: The population between age 29 and age 65 is more. Nearly 37% (median) below age 29, nearly 45%(median) between age 29 and age 65 and nearly16% (median) above age 65
Republican Party: The population between age 29 and age 65 is more. Nearly 35% (median) below age 29, nearly 45%(median) between age 29 and age 65 and nearly18% (median) above age 65

Gender Comparison:



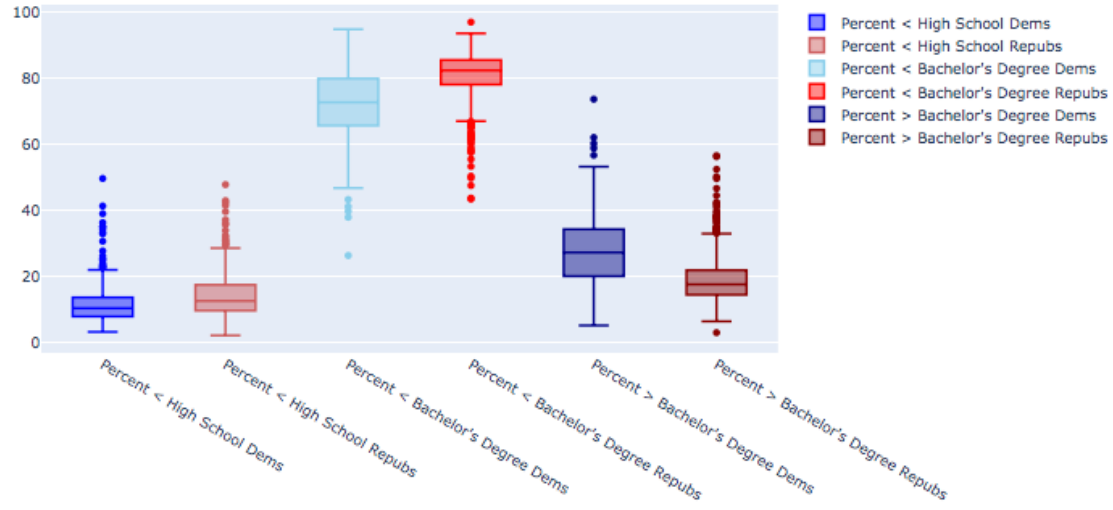
From the graph:
There is no much difference in the median of percentages in male and female when we compare between both the parties
Democratic Party: The median percentage for male is nearly 49% and for the female it is nearly 51%
Republican Party: The median percentage for male is nearly 49.5% and for the female it is nearly 50.5%

Ethnicity Comparison:



From the graph:
The percent white population is more in republican counties than in democratic counties. (near 90% in republican)
In Democratic counties, there many observations with percent white between 2.7% to 53.27%.

Education Comparison:



From the graph:
The median percentage of population less than bachelors degree is more in Republican counties than Democratics
The median percentage of population greater than bachelors degree is more in democratics than republican.

9. **(5 pts.) 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.**

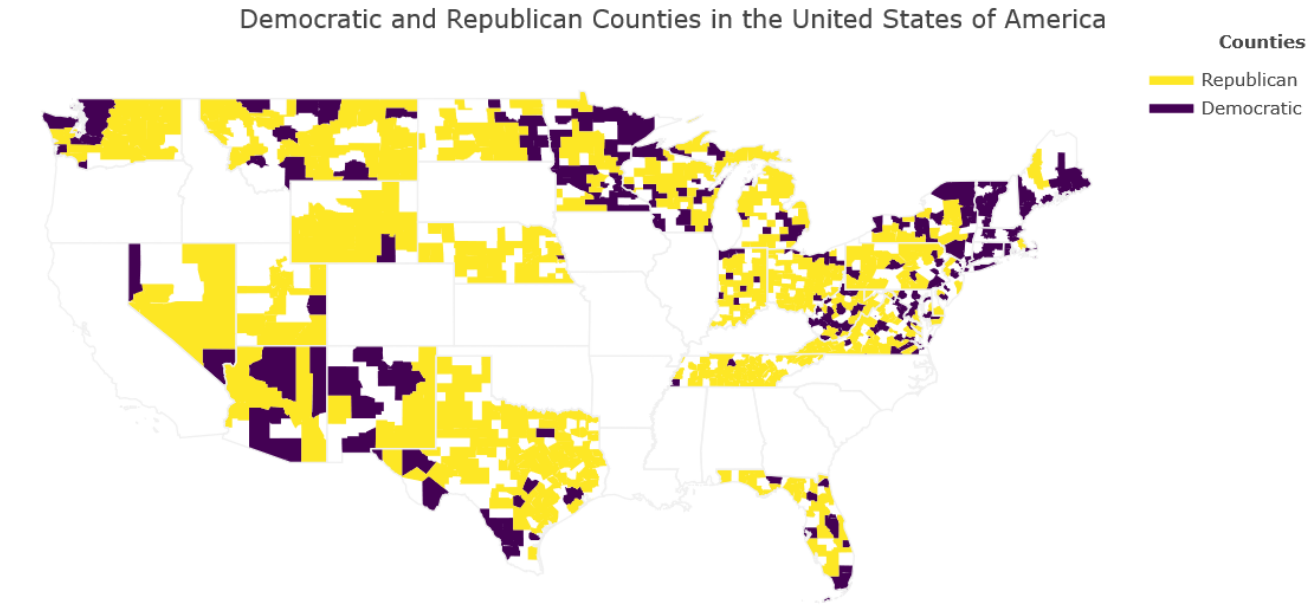
A. The important variables are the 'Percent White, not Hispanic or Latino', 'Percent Less than Bachelor's Degree'.
Percent White, not Hispanic or Latino: The percent white population is more in republican counties than in democratic counties. (nearly 90% in republican and 77% in democrartic)
Percent Less than Bachelor's Degree: The median percentage of population less than bachelors degree is more in Republican counties than Democratics (nearly 83% in republic and nearly 70% in democratic)
The median percentage of population greater than bachelors degree is more in democratics than republican(nearly 30% in democratic and nearly 18% in republican)

So, we consider these variables play a crucial role in deciding the win of the parties. So county with nearly 90% white can be considered as republican otherwise democratic.

On the basis of education, county with nearly 30% and more of population with > bachelors degree is considered as democratic otherwise republican

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.**

A. A map of the United States of America is plotted representing Democratic and Republican counties using the variable 'FIPS' and 'Party'. The plotly library is imported for ploting.


Democratic and Republican Counties in the United States of America
Counties
Republican
Democratic