CS 418: Introduction to Data Science
Project 01: Exploratory Data Analysis
Fall 2019

Instructions

This assignment is due Sunday, November 03, at 11:59PM (Central Time).

For this assignment, you must work in teams.

Deliverables for this assignment (see *Deliverables* section below) must be submitted on *Blackboard*. Only 1 submission per team is required.

Late submissions will be accepted within 0-12 hours after the deadline with a 5-point penalty and within 12-24 hours after the deadline with a 20-point penalty. No late submissions will be accepted more than 24 hours after the deadline.

Offering or receiving any kind of unauthorized or unacknowledged assistance is a violation of the University's academic integrity policies, will result in a grade of zero for the assignment, and will be subject to disciplinary action.

Project Description

Given the following datasets:

- ***election_train.csv* with results for the 2018 United States Senate elections, including the number of votes received by each party (Democratic or Republican).**
- ***demographics_train.csv* with demographic information for United States counties collected from 2012 to 2016 by the United States Census Bureau, including population, age, gender, race and ethnicity, education, income, and other statistics (www.census.gov/quickfacts/table/PST045215/00).**

Perform the following tasks:

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

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

Deliverables

Submit a compressed (zipped) folder on Blackboard containing the following files:

- **README text file with the name, NetID, and UIN of the members of the team, as well as the contribution of each member to the assignment. Also include all necessary instructions to run your code.**
- **Jupyter notebook (saved as both a PDF file and ipynb file) with your code and output for all the tasks in the project description.**
- **Report (3-5 pages, saved as a PDF file) with your answers to all the questions in the project description. Also include all corresponding results and plots. You cannot submit the PDF of your Jupyter notebook as your report.**