CS 418: Introduction to Data Science
Project 01: Exploratory Data Analysis
Fall 2018

Instructions

This assignment is due Monday, November 19, at 11:59PM (Central Time).

For this assignment, you must work in groups of 2-3 students.

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

- ***demographics_train.csv*, containing demographic information for counties in the United States collected from 2007 to 2014 by the United States Census Bureau. This information includes population, age and gender, race and ethnicity, education levels, income, and other miscellaneous statistics (www.census.gov/quickfacts/table/PST045215/00).**
- ***election_train.csv*, containing election results for the 2016 United States Presidential primaries. This information includes the number of votes received by each candidate in each county, as well as the candidate's political party.**

Perform the following tasks:

1. **Merge the two datasets using an inner join. Make sure that you address any inconsistencies in the names of the counties and the states before merging. *Hint*: the merged dataset should contain 2115 rows.**
2. **Explore the merged dataset. *How many variables does the dataset have? What is the type of these variables? Are there any missing values? If so, how will you deal with these missing values?***
3. **Create a new variable named "Democratic" that contains the number of votes cast for candidates from the Democratic party in each county.**

4. **Create a new variable named "Republican" that contains the number of votes cast for candidates from the Republican party in each county.**
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 candidates from the Democratic party than the Republican party in that county and it should be equal to 0 otherwise.**
6. **Compute the mean population in 2014 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. **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. **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 from the descriptive statistics and the plots?***
9. ***Out of all the variables in the dataset, which ones do you think are more important to determine whether a county is labeled as Democratic or Republican?* Justify your answer.**
10. **(*BONUS*) Create a map of Democratic counties and Republican counties using Python's Plotly library (plot.ly/python/county-choropleth/).**

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 group, as well as the contribution of each member to the assignment. Also include all necessary instructions to run your code.**
- **Jupyter notebook (saved as a ipynb file) with your code 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 a PDF of your Jupyter notebook as your report.**