Final Exam

DATA 210

Due: Monday December 20th, at 11:59pm

Part A

1. In this question, you will use a series of datasets to investigate population density in the United States.

a) Load in population data for Alabama ("sub-est2016_1.csv") and Alaska ("sub-est2016_2.csv"), then append the two datasets together so that all of the information is within one dataframe.

b) Read in the csv file that already contains population information for each state. Check to see which unique states are included in this dataset.

c) There's a lot of interesting data in this population dataset, but for our purposes in this problem set, we are only interested in a few columns. Use the subset() function to subset the "NAME", "STNAME", and "POPESTIMATE2012" columns into a new dataset. (Using a different function to complete the same task will result in partial credit.)

d) This new subsetted dataset definitely makes our lives easier, but it still includes the population stats for each city and town. You'll notice, however, that the first observation for each new state is the population total for the entire state where the states name appears in both the NAME and STNAME columns. Use the subset() function to choose only these rows. Make sure that your new data set doesn't have any repeating/redundant observations or columns (The resulting dataframe should be 51 X 2)

e) We're going to try to find the population density of each state. Our first step in doing this is to read in some online data about the square mileage of each state from this link:(https://raw.githubusercontent.com/jakevdp/PythonDataScienceHandbook/master/notebooks/data/state-areas.csv) Once the data is read in, merge that data set with our 2012 state populations dataset from the last question. Which observations can be matched? Make sure to not merge observation(s) that have no match.

f) Next, we are going to create a new variable in this merged dataset that tells us each state's population density in 2012. Do this by dividing the population variable by the state size variable.

g) Finally we've finished preparing our dataset, now we're going to get into some more interesting investigative work. Let's first load in the "ECN_2012_US_52A1.csv" dataset which includes economic data for each sector within each state. Get rid of the first row, as this merely gives us descriptions of each variable.

h) Find the total revenue per sector by state.

i) Now merge this dataset with our population density dataset.

j) Plot the relationship between state population density and the state's total revenue to see if there's a relationship. Comment on your findings.

Part B

For this question, you will use the data file ‘nes.rda’ (which will require you to use load(“nes.rda”) to read in the data.) The codebook, called ‘nes2012_codebook.pdf’ is also available to you on Canvas. This data comes from the ANES 2012 Time Series Study, which looks at attitudes toward political ideologies and groups, among many other things.

1. According to this survey, of those who claimed to have voted in the 2008 election, what percentage of survey respondents voted for Barack Obama in 2008? (Hint: you will need to search the codebook to find the variables ‘interest_voted2008’ and ‘interest_whovote2008’ in order to clean them correctly.)

2. A ‘Feeling Thermometer’ is a type of survey question that asks respondents to rate how warmly or cool they feel toward an individual or group. A feeling thermometer score of 100 indicates a respondent feels the most positive toward that entity. A feeling score of 0 indicates the respondent feels most negative about that entity. A score of 50 indicates indifference. Using the variable that records the feeling thermometer score towards the ‘Federal Government in Washington,’ clean the variable to only include scores between 0 and 100. (Use the codebook to locate the ‘ftgr_fedgov’ variable to clean it properly.)

3. Using the cleaned variable, what is the average feeling thermometer for the Federal Government in Washington, according to this survey?

4. Using the ‘prevote_regpty’ variable, create a new variable that indicates whether a respondent is a Democrat or a Republican. All other political affiliations or unknowns should be set to ‘NA.’ (Use the codebook to clean this variable correctly.)

5. Find the difference in means between the average feeling thermometer score for Democrats vs. Republicans. What do you conclude?