Midterm Exam

DATA 210

Week 5

1. Consider the following survey question:

   Who is to blame for the dysfunction and gridlock in Washington? (select one)

   - Donald Trump
   - Congress
   - Nancy Pelosi
   - Chuck Schumer
   - Obama
   - The media

   a. Discuss the problems with the way this survey question has been written.

   b. Rewrite the survey question and response options. Explain how your survey wording addresses the issues you identified.

2. One of the most important aspects of being a data scientist is knowing how to ask good research questions and find interesting results in your data. Imagine you are a data scientist working for Very Accurate Political Insights. You are given a dataset that contains survey responses from 1,100 US residents on the following topics:

   - 2016 presidential vote
   - Trump approval
   - Support for the Green New Deal
   - Support for Medicare for All
   - Age
   - Sex/gender
   - Race/ethnicity
   - Education

   a. Imagine you are asked to write a report about the survey. After cleaning the data, what would you do next? What research questions would be interesting to try to answer with the data? Come up with several examples of analyses you'd perform or research questions you may try to answer with the data.

   b. Imagine you had been involved in writing this survey. Write two questions that you think would be interesting to have included in this survey. One of these questions should be about political behavior or political attitudes (like the first four bullet points above). The other questions should be about the demographic or other non-political characteristics of the respondent (like the last four bullet points above). Write the questions exactly as they should appear in the survey and include answer choices.

3. Use the GenForward survey data (`genforward sept 2017.sav`) to answer the following questions. If you need to do any data cleaning before performing these calculations, please include that code in your R script.[1]

   a) What percent of the sample strongly approved or somewhat approved of the way that President Trump is handling his job as president (using question Q1)?

   b) What percentage of Republican men "strongly approve" or "somewhat approve" of the way Trump is handling his job as president? What is this percentage for Republican women? What percentage of Republican men and Republican women (separately) "somewhat disapprove" or "strongly disapprove" of Trump?[2]

   c) Which two issues did 2016 Trump voters indicate were the most important problems facing the country? What percentage of Trump voters listed each of these two issues as the top issue?

   d) What percentage of 2016 Clinton voters listed these two issues are the most important problem facing the country?

   e) What are the top three issues that women over 30 years old care about? Are these top issues the same for women aged 30 and under?

[1] You do not need to consider survey weights for any of the answers in this question. Also, be sure that you are careful about how you handle people who did not answer particular questions.

[2] For this question, you should use the `PartyID7` variable and consider anybody who is in the "Lean Republican", "Moderate Republican", or "Strong Republican" categories to be a Republican.

4. For this exercise, we'll be working with daily weather data from a weather station in New York City's Central Park. The station has been running continuously since January of 1869, so we'll be able to analyze 150 years of weather patterns. The data come from a dataset extraction tool provided by the National Oceanic and Atmospheric Administration.[3]

    a. Begin by loading the temperature data.[4] Use the `separate()` function to turn the `DATE` variable into three separate variables for year, month, and date. Which years are missing at least one day of temperature data, and how many days are missing?

    b. Create a variable that tells us the difference between the highest and lowest temperature for each day. Across the full dataset, what the average of this difference? Which day during this 150 year window had the biggest difference between the highest and lowest temperature? Averaging across years, which month tends to have the highest average difference in daily high and low temperatures?

    c. Load and merge in the precipitation data. What type of merge does it mark sense to perform? Which variable(s) will you merge on? Perform the merge, then use the results to figure out how many days in the past 150 years had a high temperature of at least 50 degrees and received at least 1 inch of snowfall.

    d. Aggregate the data by month to figure out what percentage of days have had preciptation since 1869. Your resulting dataset should have 12 rows (one per month). You should use the PRCP variable (and ignore the SNOW variable). Which month tends to have the most rainy days in New York City? What percentage of days does it usually rain in this month? And which month tends to be the dryest (i.e. fewest days with precipitation)? What percentage?

    e. Use aggregation to figure out how many days in each year since 1869 had a low temperature of 32 degree or below. Use the `plot()` function or ggplot2 to make a simple graph of the relationship between the year (on the x-axis) and the number of cold days in Central Park (on the y-axis). What pattern do you notice in this graph?

[3] You can download data for other cities or locations here: https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND.
[4] All temperatures are in terms of Fahreinheit.