DA5020 - Week 10 SQLite and comparing dplyr to SQL

2019-09-21

This week you are responsible for chapters 10, 11, 12 in the "Data Collection, Integration and Analysis" textbook. Review each chapter separately and work through all examples in the text BEFORE starting the assignment. You will use the schema you developed in homework 6 to store data in SQLite.

This week's assignment you use the relational schema you designed in week 6 and store data into the SQLite relational database system. Load the Unemployment and Educational data files into R studio. One file contains yearly unemployment rates from 1970 to 2015, for counties in the US. The other file contains aggregated data percentages on the highest level of education achieved for each census member. The levels of education are: "less than a high school diploma", "high school diploma awarded", "attended some college", "college graduate and beyond". The census tracks the information at the county level and uses a fips number to represent a specific county within a U.S. state. The fips number is a 5 digit number where the first two digits of the fips number represents a U.S. state, while the last three digits represent a specific county within that state.

Questions

1. Revisit the census schema you created for homework 6. After installing SQLite, implement the tables for your database design in SQLite and load the data into the correct tables using either SQL INSERT statements or CSV loads. Make sure the database design is normalized (at least 3NF) and has minimal redundancy. Make sure your SQLite tables have primary keys as well as foreign keys for relationships. (20 points)

2. Write SQL expressions to answer the following queries: (40 points)

- 2.0 In the year 1970, what is the population percent that did not earn a high school diploma for the Nantucket county in Massachusetts? What about the year 2015?
- 2.1 What is the average population percentage that did not earn a high school diploma for the counties in Alabama for the year 2015?
- 2.2 What is the average percentage of college graduates for the counties in the state of Massachusetts for the year 2015?
- 2.3 Determine the average percentage of the population that did not earn a high school diploma for the counties in Alabama for each year within the dataset. The result should return the calendar year and the average percentage drop out rate for that year.
- 2.4 What is the most common rural_urban code for the U.S. counties?
- 2.5 Which counties have not been coded with a rural urban code? Return a result that contains two fields: County, State for the counties that has not been assigned a rural urban code. Do not return duplicate values in the result. Order the result alphabetically by state.

- 2.6 What is the minimal percentage of college graduates for the counties in the state of Mississippi for the year 2010?
- 2.7 Which state contains the most number of counties that have not been provided a rural urban code?
- 2.8 In the year 2015, which fip counties, U.S. states contain a higher percentage of unemployed citizens than the percentage of college graduates? List the county name and the state name. Order the result alphabetically by state.
- 2.9 Return the county, U.S. state and year that contains the highest percentage of college graduates in this dataset?

3. Compare your SQL SELECT statements to your dplyr statements written to answer the same questions. Do you have a preference between the two methods? State your reasons for your preference. (10 points)

4. Write a R function named get_state_county_education_data_dplyr(edf, state), it accepts a data frame containing education data and a state's abbreviation for arguments and produces a chart that shows the change in education across time for each county in that state. Use dplyr to extract the data. Write a few R statements that call the function with different state values. (5 points)

5. Write a R function named get_state_county_education_data_sql(edSQL, state), it accepts a SQL database connection containing education data and a state's abbreviation for arguments and produces a chart that shows the change in education across time for each county in that state. Use SQL SELECT to extract the data from the database. Write a few R statements that call the function with different state values. (10 points)

6. Write a R function named get_state_county_unemployment_data_dplyr(udf, state), it accepts a data frame containing unemployment data and state's abbreviation and produces a chart that shows the change in unemployment across time for each county in that state. Use dplyr to extract the data. Write a few R statements that call the function with different state values. (5 points)

7. Write a R function named get_state_county_unemployment_data_sql(udfSQL, state), it accepts a SQL database oject containing unemployment data and state's abbreviation and produces a chart that shows the change in education across time for each county in that state. Use SQL SELECT to extract the data. Write a few R statements that call the function with different state values. (10 points)

Submission

You need to submit an .Rmd extension file as well as the generated pdf file. Be sure to state all the assumptions and give explanations as comments in the .Rmd file wherever needed to help us assess your submission. Please name the submission file LAST_FirstInitial_1.Rmd for example for John Smith's 1st assignment, the file should be named Smith_J_1.Rmd.