GovEx

Data engineer
Interview Exercises

Instructions

Please take no longer than two days (48h) to complete the following exercises and return the answers, scripts and any other materials to mvaugha6@jhu.edu. In general, these exercises are not specifically designed to have a right or wrong answer, but more importantly to allow you to be creative. We are more interested in the process rather than the final answer.

Part 1: Calculating Youth Unemployment Rate

We have provided you with a file containing unemployment and population data disaggregated by age and sex for various cities from 2005 to 2019 from the American Community Survey (ACS) 1-year estimates. For each city, we would like you to convert this data to a single value for each year representing the percent of civilians aged 16-24 that are unemployed and output it into a particular format. Since ACS disaggregates data by sex and age we need to add some of these variables together to calculate a percentage. See calculation methods below:

To obtain the **total number of unemployed civilians aged 16-24**:
B23001_008E (16-19, males) + **B23001_015E** (20-21 males) + **B23001_022E** (22-24 males) + **B23001_094E** (16-19 females) + **B23001_101E** (20-21 females) + **B23001_108E** (22-24 females)

To obtain the **total number of civilians aged 16-24 in the labor force**:
B23001_006E + **B23001_013E** + **B23001_020E** + **B23001_092E**+ **B23001_099E**+ **B23001_106E**

The final output should be a single **csv** file with four fields:

city: city identifier
date: year which the data represents
value: percent of civilians aged 16-24 that are unemployed
delta: change from the previous year

See table below for example of what the output table should look like

city	date	value	delta
city 1	2005	###	
city 1	2006	###	###
…	…	…	…
city 1	2018	###	###
city 1	2019	###	###
city 2	2005	###	
city 2	2006	###	###
city 2	2007	###	###
…	…	…	…

Part 2: Scaling the process in a larger context

In the previous question, we asked you to create a script that ingests data, processes it, and produces a specific output. We're also interested in how you fit your work within a larger context/framework, and how you communicate your thoughts.

Using any form of media you prefer (document, presentation, commented code, video, etc), please describe your recommendations to scale it. You may pick 2-3 of the below scenarios, provide a few alternative scenarios based upon your own knowledge/experience/intuition, or a hybrid of the two.

- Ingesting a greater variety of data sources, processing them, and adding them to the same output.
- Ensuring that when data sources are updated, any previously processed data dependent upon those sources also gets updated.
- Managing multiple versions of the output data (for example, one for development, one for staging/testing, and one for production).
- Automation
- Performance optimizations for querying the output data
- Integrating with / leveraging metadata such as data source details, processing methodology and calculations, data, etc.