Data Engineering - Assignment 09

For this assignment, you will implement and then compare and contrast imperative and declarative programming styles within the context of a dbt project. There is one data file included in the GitHub repository and a set of requirements below. You will create a single data pipeline to accomplish those requirements but must use a combination of imperative and declarative steps. In addition to writing the code for each step you devise; you must also include comments that describe why you chose that programming paradigm and what turned you off using the other. Be descriptive. You can’t simply say “I chose imperative because I thought it would be clearer. Declarative would have too complex.”

# Input Description

The included data file, complications.csv, describes the rate of specific complications (measure\_id, measure\_name) for specific hospitals (provider\_id, hospital\_name) for procedures performed in a specific time period (measure\_start\_date, measure\_end\_date).

The reported measurement value (score) is an estimated value for the population (denominator) observed in that period. You can assume that measures described by their name as “rates” have a value reported in percent.

# Output Requirements

The output should be a Snowflake database table with one row for each measure. Each year should have two columns, city with the best scoring hospital and city with the worst scoring hospital for that measure in that time period. An example is provided below. Not that these are not real based on the data—only shown for structural purposes.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| measure\_id | 2015-worst | 2015-best | 2016-worst | 2016-best | 2017-worst | 2017-best |
| COMP\_HIP\_KNEE | Fayette | Gadsden | Fayette | Hamilton | Fayette | Dothan |
| MORT\_30\_AMI | Gadsen | Dothan | Fayette | Dothan | Hamilton | Dothan |
| MORT\_30\_CABG | Wilcox | Dothan | Wilcox | Hamilton | Dothan | Wilcox |

1. You should load the raw data into Snowflake table and use that as the input to your dbt project
2. Your dbt project output should be another table you create through tow or more layers of coding in dbt
3. Only consider hospitals in Alabama
4. Only include measurement periods that begin in July of one year and run at least 12 months
5. If there are overlapping periods (for example periods that start on different dates, but end on the same date, you can treat those as distinct time periods)
6. The year number should be the year in which the sample period begins
7. Ignore those records where the score or the denominator are “Not Available”
8. If the denominator the best and the denominator of the worst are more than 5x difference for any of the time periods for that measure, add an asterisk to the end of the measure name
9. Format the city name in Title Case / Proper Case
10. While the required output is described above, consider ways to manage your data so that other outputs can also be easily created if needed. For example:
    1. Show hospital name or provider id instead of city name
    2. Show worst, best, and median
    3. Show values instead of city name
    4. Show zip code instead of city name
    5. Show measure description instead of name
    6. Show if a given city has moved up or down from the previous period
    7. *You don’t have to actually do these things, just think about how to best structure your work so it wouldn’t be hard to do these kinds of things in the future.*