Base Solution:

This solution provides answers to the requirements in Part 2.

The solution uses Python as the primary language to preprocess the required data and ingest it into the SQLite library database in memory. SQL is then passed to the database, and query data needed and later post-process into Pandas data frame that is then finally saved as a CSV file for result viewing. The database can be stored on disk, which has been demonstrated at the bottom of the code file, but for faster processing speed, it has been put in memory. Solutions can also be stored in a new table, but for direct viewing purpose, they have been exported as CSV.

Assumptions:

- Dataset1 data includes both number of sample and the sampled values, depending on whether the name of the dimension includes the dollar sign
- There is enough memory to contain data in this exercise; speed is the priority
- CSV for direct result reading is preferred over creating a new table and store the values as results or append repeated values for the same regions as new columns
- The income dimension with name “Total – Income Statistics” is not a dollar value. All features under this dimension represent number sampled within the feature
- Part C requires the proportion of male and female within the identities itself, not over the total population as a whole. (if it is, proportion can be obtained by applying results from part A)

Please see comments in the code to get more information and CSV files for results.