Health Equity Explorer

Explore health equity by state, race, ethnicity, gender, and age. Based on the full `Lokahi_Innovation_in_Healthcare_Hackathon.zip` dataset.

Approach

The dataset consists of 907,464 members and their associated enrollment and services data. Processing the entire dataset using Pandas and Dask on my hardware, a Dell Inspiron laptop with 16 GB RAM and 4 cores, was not possible, causing memory allocation and out of disk space errors after minutes of processing.

As a result, I decided to split the member dataset into groups of 100,000, in order of parquet file name in the original dataset. I called each group a "cohort". This resulted in 10 cohorts, the first 9 having 100,000 rows, and the 10th having the remaining 7,464 rows.

I then analyzed each cohort separately and generated separate HTML reports for state, race, ethnicity, gender, and age comparisons, for a total of 10 reports per cohort, plus an additional report for providers by state. The analysis for each cohort takes several minutes but now could all run to completion on my laptop without crashing.

With all 10x10+1 = 101 reports generated from the analysis, I wrote a web UI to display the reports. The user can select a specific cohort by number (from 1 to 10) or can expand all 10 cohorts to see their corresponding reports, or the providers report.

Splitter script - split_rows.py

The `split_rows.py` script is located in directory `dataset/Claims_Member`. To run it first follow the `README.txt` in each `dataset` folder, which tells you to put the contents of the corresponding folder from the provided dataset.

Then run `python split_rows.py`. It creates subfolders `split_Claims_Member_1` to `split_Claims_Member_10`. Each subfolder contains a parquet file and a CSV file. The CSV is for reference; the analysis only uses the parquet file. These generated files are included in this repo.

Note this script only needs the original dataset files for `Claims_Member`. However, you should also provide the other requested original dataset files because they are needed by the analysis script, described next.

The .env file

A `.env_sample` file is provided in this repo. Copy it to a new `.env` file which you can edit if you choose to locate the dataset files in a custom folder.

Analysis script - analyze.py

To run `analyze.py`, first open the `.env` file and confirm the `MEMBERS_PATH_BASE` points to where the split subfolders are located (`split_Claims_Member_1` to `split_Claims_Member_10`). By default this is set to directory `dataset/Claims_Member`. Then run `python analyze.py` which will generate the 101 HTML reports in project output folder `html_files`. These generated files are included in this repo. Please be patient running this script and ensure you have maximum memory free.

Web app - ui.py

The web app is a FastAPI app that lets you select a specific cohort to see the HTML reports for it. There are 10 dropdowns for selecting cohorts to allow you to see the reports for all 10 cohorts on the same page if you wish. To make it convenient for you to see all reports for all cohorts, there is an `Expand all` button at the top you can click to expand all 10 cohorts at once. If you use this feature, please wait for all the reports to render, and note that the order of rendering is not necessarily from top to bottom or left to right.

Since not all reports can fit on one page, there are buttons at the top to open new browser tabs to display additional reports.

The web app is launched using `uvicorn` in the code itself, so to start the app simply run `python ui.py`. The app will start on port 8000.

Deployment

A Docker file is provided for deployment. I used it to build a Docker image which I successfully deployed to my cloud instance: https://health-equity-explorer.williamcheung.click (https://health-equity-explorer.williamcheung.click)