IRisk Lab Data Discovery and Consolidation

Week 10 Report

Rohit Valmeekam

What Was Accomplished

- ❖ Successfully located datasets related to insurance premiums, specifically focusing on pre-existing conditions and health insurance marketplace information. The identified datasets are as follows:
  - ➢ https://data.world/carlvlewis/pre-exisiting-conditions-by-state-congressional-district
  - ➢ https://data.world/data-ny-gov/472d-zats
  - ➢ https://data.world/data-ny-gov/qkrk-6v78
  - ➢ https://data.world/data-ny-gov/u925-8y2g
  - ➢ https://data.world/data-society/health-insurance-marketplace
- ❖ Started the process of evaluating the identified datasets to ensure relevance, reliability, and completeness for our project. This involves examining the metadata, data quality, and understanding the variables within each dataset.
  - ➢ Started looking at the data categories and feeding it as input data in the data pipeline I created earlier
  - ➢ Began documenting the key features and attributes of each dataset, including data sources, update frequency, and any notable limitations. This documentation will serve as a reference for the team and future stages of the project.
    - Although the focus is now on Lessor's Risk, we can expand on our project next semester by feeding these datasets into my data pipeline and further into our database schema as well
    - This is easy
  - ➢ Explored additional external resources and research papers related to insurance premiums, pre-existing conditions, and health insurance market trends. This broader exploration aims to provide a comprehensive understanding of the subject matter and identify any gaps in the currently identified datasets.

Next Steps

- ❖ Data Consolidation: Facilitate process of consolidating relevant information from the identified datasets. This involves merging and cleaning data to create a unified dataset that can be used for analysis. Collaborate with Group 1 to help merge datasets

❖ Data Validation: Perform thorough validation of the consolidated dataset to ensure data accuracy and consistency. This step is critical for producing reliable results in subsequent analyses that we will do in collaboration with group 1.