Project Part 2 Report

## Logical Schema Design

For this part, I used methods in the textbook about converting ER(EER) diagrams to relational schemas, including how to convert relations to tables. So, the result relational schema includes table of entities and relations. And during the process, I also checked the normalization of my schema, and it is in the 3NF.

## Data Lake Design

It is hard to find the explicit dataset of insurance companies or medical expenditure, and I tried my best to find one dataset from MEPS, which contains medical information of individuals (such as condition types, age when conditioned, expenditure etc.) from 2018 to 2019, around 180000 records stored in xlsx files and a dataset from Kaggle which contains individuals’ information and their expenditure. I plan to use these information to train a model to predict the expenditure of an individual at a certain age with conditions and other info. And this information can be used by the insurance company when they provide the services as they can predict the amount of money the clients are likely to spend.

But this information doesn’t include any inter-relation with the relational schema I created. So, I plan to directly use these files as input to the model and train the model without access to the physical database system.

And from what I figured out and understood, the relational schema I created doesn’t inter-relate with any open-source dataset due to privacy and credentials. So the best way to do this is to set them apart and apply the trained model to the new data collected by the company.

## Cloud Application

To achieve what I mentioned above, I decided to use Microsoft Azure Cloud platform and store the data I collected in the storage and create a Azure MySQL server to build the physical database system of my relational schema.