Optum Technical Interview Task

**Task Overview:**

In the Out of Network affordability team, we are focused on making the US healthcare system work better for everyone by reducing the cost of healthcare and finding affordability opportunities for our clients.

The task at hand is to perform an exploratory data analysis (EDA) and transform this into a story (data viz) for business users. The data to be used for this ask is publicly available claim data for Medicare members in the US. Medicare is the government healthcare program for US citizens aged 65+.

The aim is to understand how the cost of treating certain chronic conditions varies across different providers.

**Guidelines**

* Please complete this task before the interview.
* The candidate can spend as much time as desired on the tasks however, it is recommended that no more than 2 to 3 hours be spent on this task.
* Be prepared to talk through what you did during the task to a technical audience.
* Be prepared to tell a story to a non-technical business audience.
* If there are parts of this EDA that you would do differently, or with more time, would try something else, please discuss these during the interview.
* Please share code & relevant materials (slides, workbooks etc..) at least 24 hours prior to the interview

**Data**

The data can be found on the CMS (Medicare) website.

Member Benefit Data

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/DE1_0_2009_Beneficiary_Summary_File_Sample_20.zip>

Outpatient Claim Data

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/DE1_0_2008_to_2010_Outpatient_Claims_Sample_20.zip>

User Documentation

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/Downloads/SynPUF_DUG.pdf>

Additional Info

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs>

**Data Cleaning**

The beneficiary summary file has several chronic illness columns for each member. These are Boolean fields.

* Convert these columns into a single categorical variable, concatenating multiple true diagnoses.
* If a member has 3 or more chronic conditions, categorise these as “Multiple”
* Join claims & benefit data

**Story Telling (Data Viz). What stands out in the data? What do you want your business partners to take away from this data?**

**Basic Stats**

* What is the distribution of races?
* What is the most common chronic illness combination?
* Which chronic illness combination has the total highest cost?
* Which chronic illness combination has the highest cost per member?
* Any other significant metrics you think the business should be aware of?

**Benchmarking**

The aim here is to understand the distribution of cost across providers treating members with these chronic illnesses. Benchmarking providers across types of care is often a helpful starting point to begin solutioning for areas of high cost.

* For each provider (use AT\_PHYSN\_NPI) & chronic illness, calculate the cost per member.
* For each chronic illness combination, represent the distribution of costs per provider.
* How does this change if we filter out cases where a given Chronic Illness & Provider NPI combination only has 1 member?
* Which providers are consistently expensive across chronic illnesses they treat?
* Any other significant metrics you think the business should be aware of?

During the interview, please be prepared to discuss coding techniques, libraries used etc. to answer the questions.