TEAM -10 : Analysis of Health Insurance Claim Data

Introduction

Healthcare is a booming industry in today's market. As the value of healthcare is increasing the insurances are getting more premium. According to the government, the total Medicare spending increased exponentially due to increase in presence of chronic diseases. People affected by high blood pressure, heart diseases, depression are the highest insurance claimers.

As the cost of diagnosis is quite ambiguous, insurance companies have raised their premiums as more and more patients are claiming their insurance due to bad health conditions.

Some of the supporting facts are:

- People having Medical Insurance coverage in USA – 299.5 million
- Average number of claims in the country that saves lives – 68,000
- People having Medical Insurance coverage in Maryland - 5.9 million
- Most common disease or accident on which insurance was claimed – 1+ million

Problem Statement

Our problem is to analyze the trends of the claims from data and helping insurance companies in their decision making.

Role of Data Management

Data plays a crucial part to understand the pattern of the fraud claims. It is important to process and manage the data efficiently. Relational Databases are the simplest one which helps when is data is small. It removes the redundancy and help in organizing the data in a better way. As cloud data bases are emerging and are easy to configure, we have leveraged AWS to store and query the data.

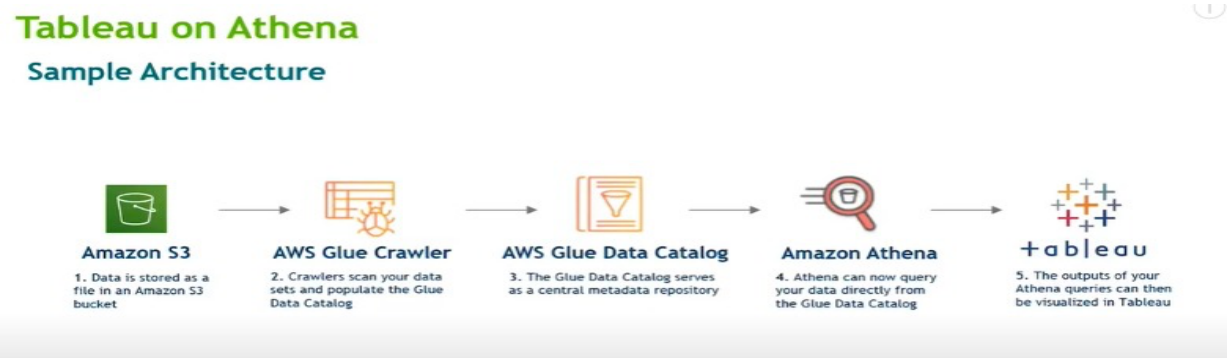


Fig.1 : Flow of DataPipeline

TEAM -10 : Analysis of Health Insurance Claim Data

We have used the AWS standard architecture to integrate data pipeline with tableau for data visualization. The architecture has helped to seamless data flow from S3 to tableau without any hassle. The above picture has a brief explanation of the components.

Data Collection

Source: KAGGLE

Link: https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

In the dataset we have three common separated files which are as follows:

1. Inpatient Data: This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit diagnosis code.

   Volume of data: 50K records and 27 attributes

2. Outpatient Data: This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.

   Volume of data: 643k records and 30 attributes

3. Beneficiary Details Data: This data contains beneficiary KYC details like health conditions, region they belong to etc.

   Volume of data: 202k records and 27 attributes

Steps involved in Data Pipeline

- RDS Database created in AWS named 'database-aws'
- Assigned respective IAM roles for the team members
- We have stored the **3** CSV files in **S3 bucket** of AWS to store the data.
- After that, we have crawled the data from AWS Glue Crawler and Connected Athena service to our cloud data.
- Integrated Tableau with AWS RDS database to plot charts.

TEAM -10 : Analysis of Health Insurance Claim Data

- Ran data queries on AWS Athena for Data Analysis.

ER DIAGRAM

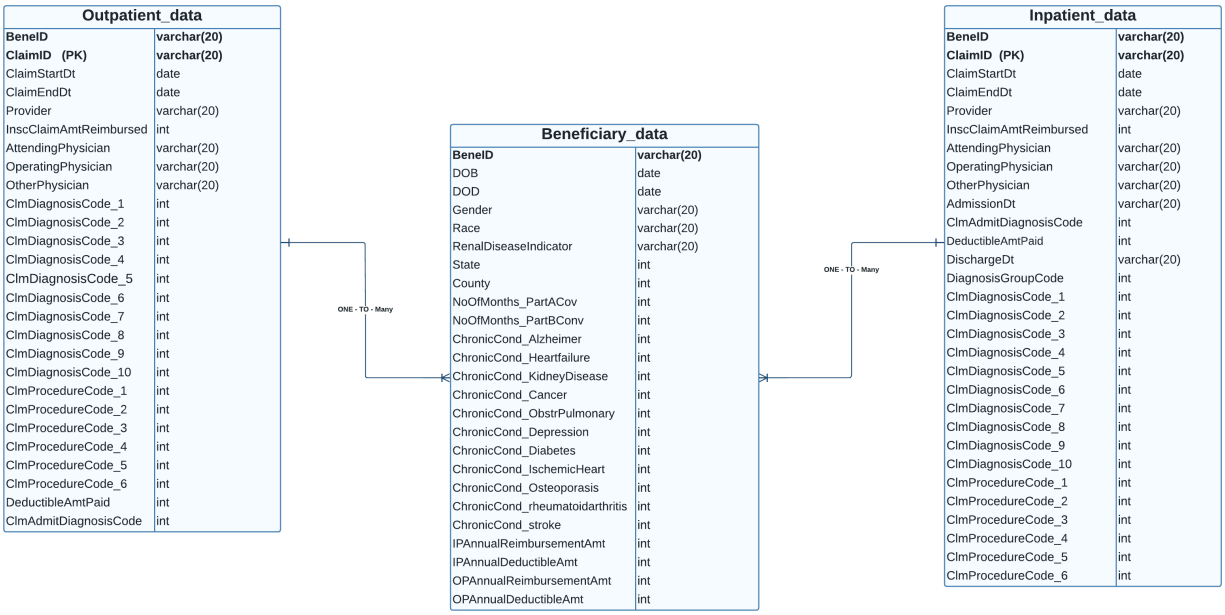


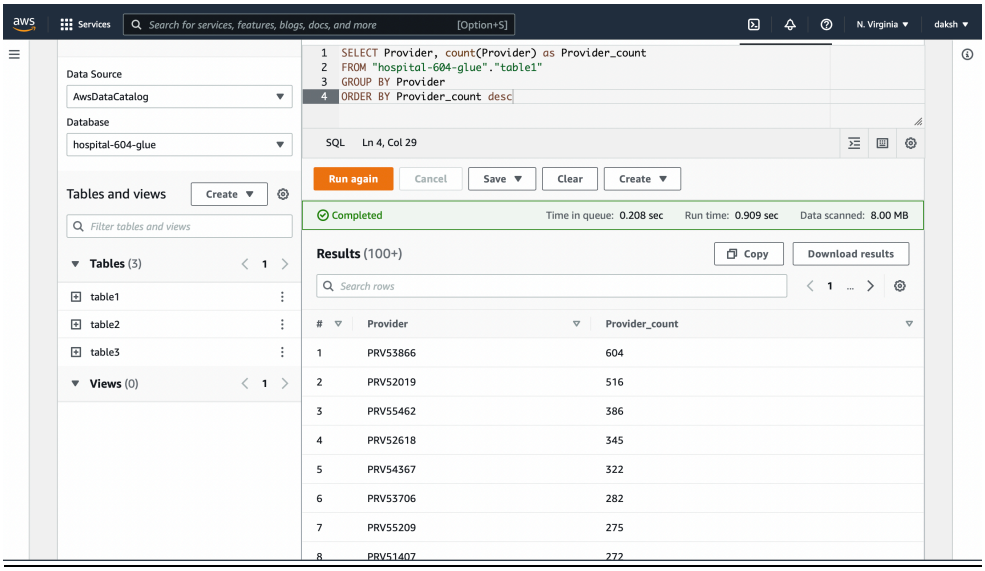
Fig. 2: Entity Relational Diagram

The ERD diagram shows the relational between the 3 tables. This helped us to understand the relation between three entities.

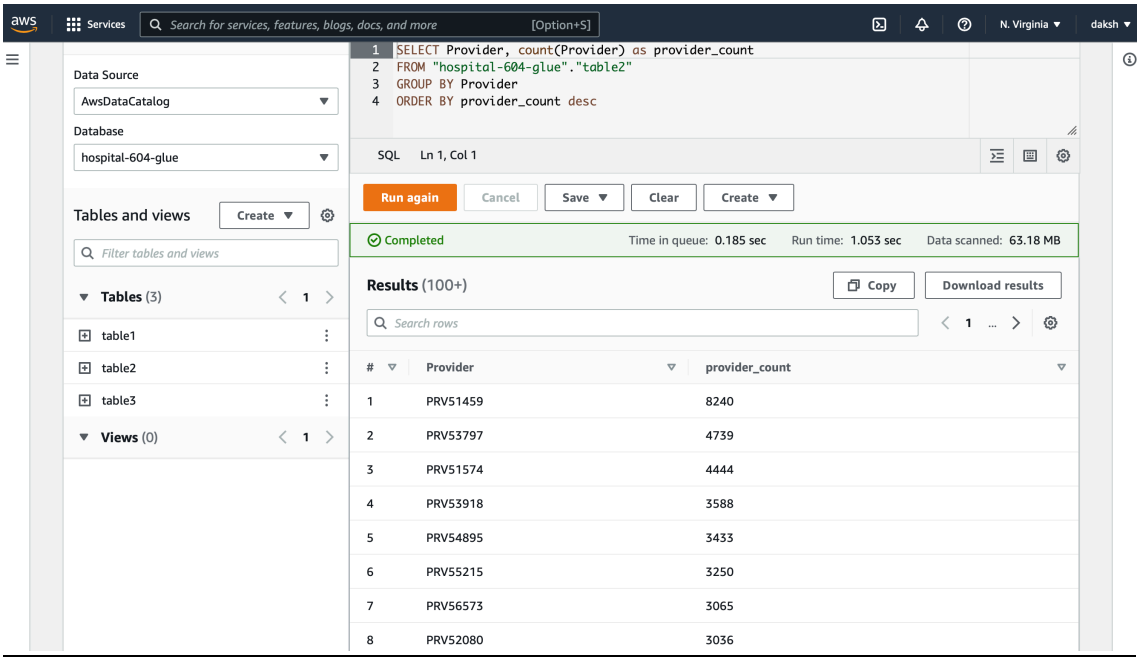
TEAM -10 : Analysis of Health Insurance Claim Data

Data Analysis using SQL

1. Count of unique providers in Inpatient Data

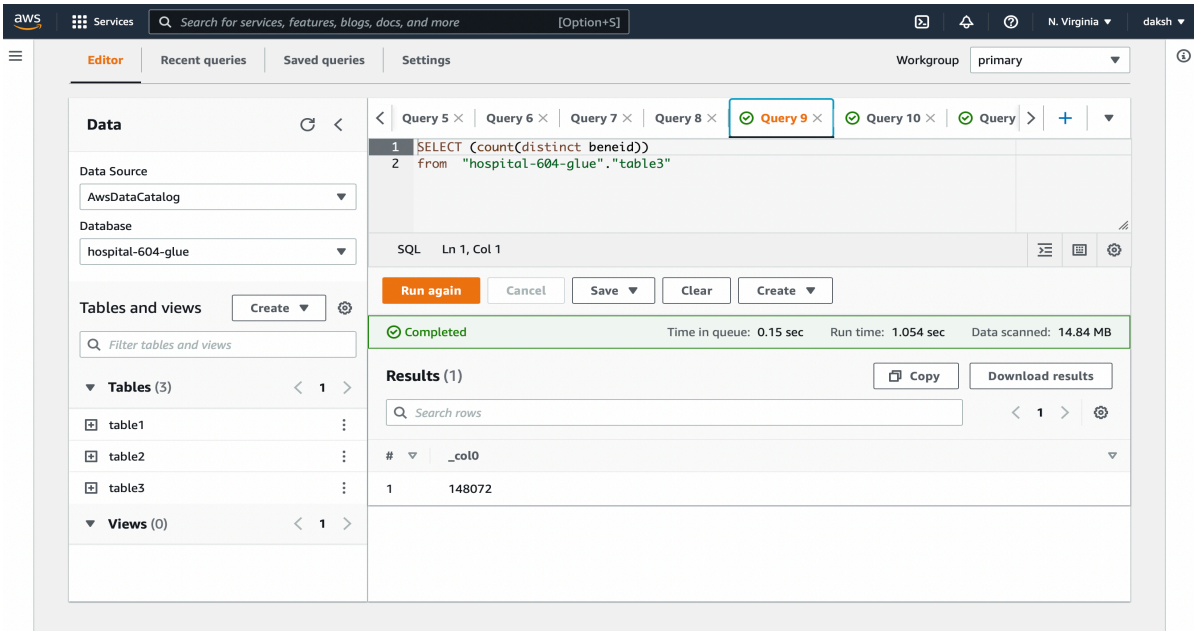


2. Count of unique providers in Outpatient Data



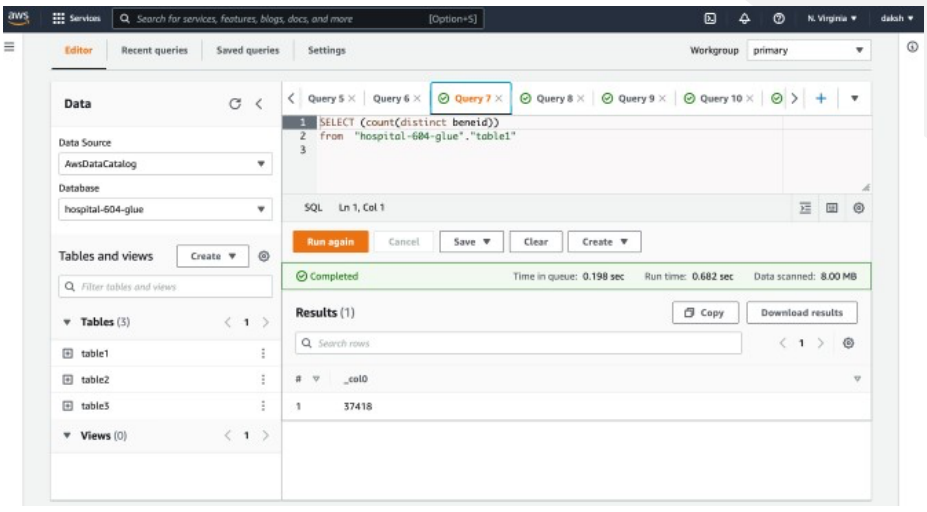
TEAM -10 : Analysis of Health Insurance Claim Data

3. Count of unique Beneficiary ID in Inpatient Data



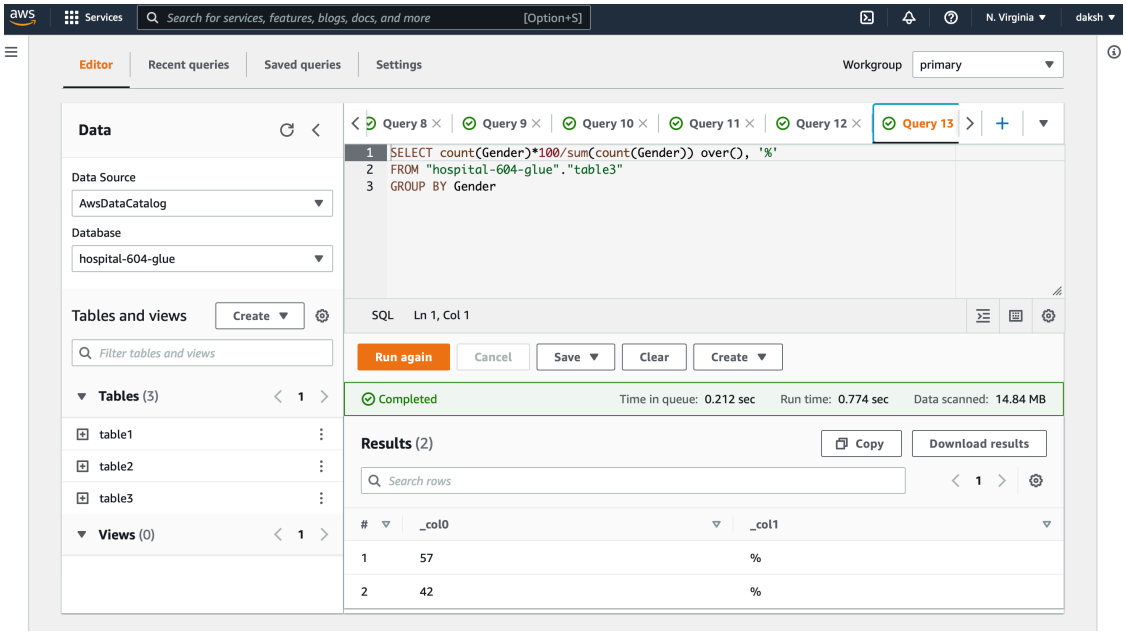
4. Count of unique Beneficiary ID in Outpatient Data

AWS

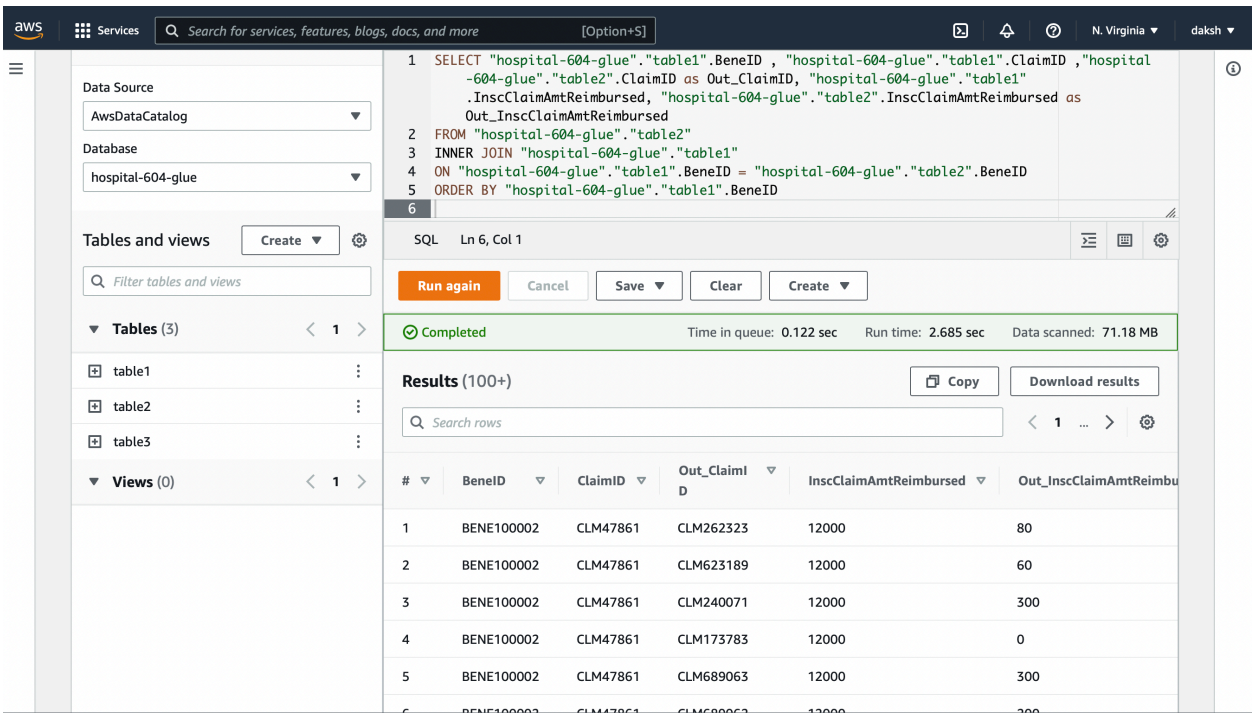


TEAM -10 : Analysis of Health Insurance Claim Data

5. %age of the Gender

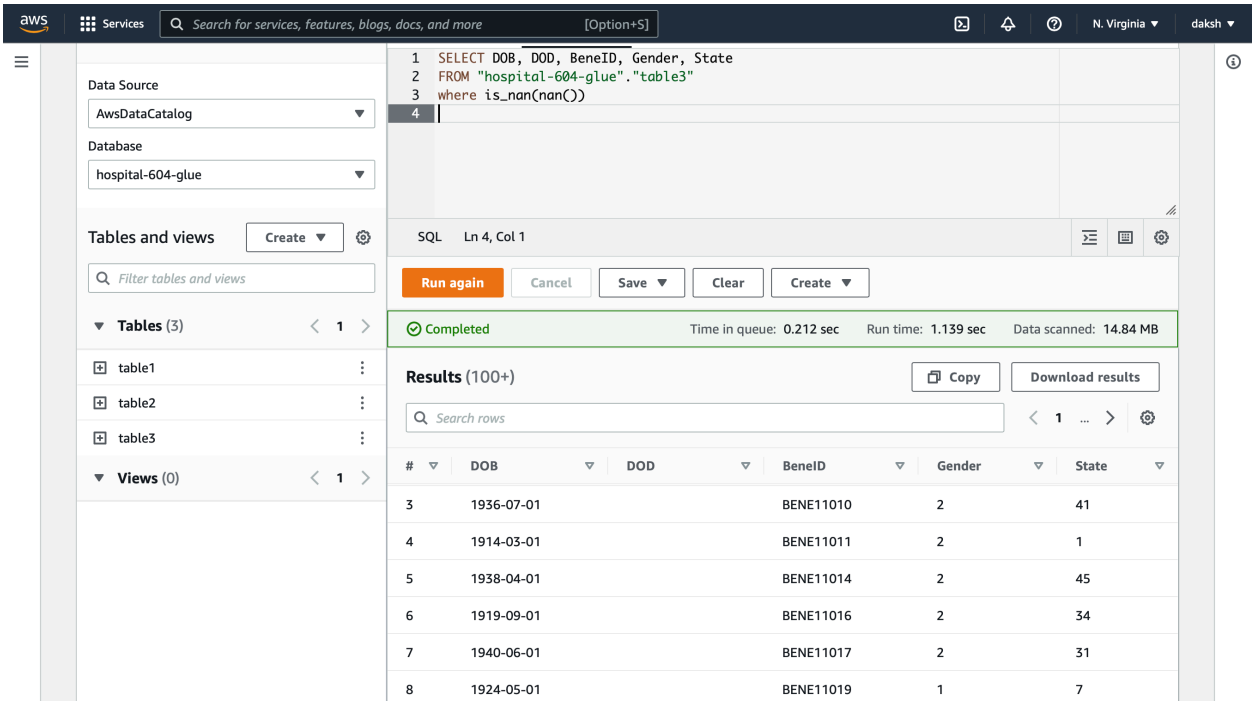


6. Inner Join Query



TEAM -10 : Analysis of Health Insurance Claim Data

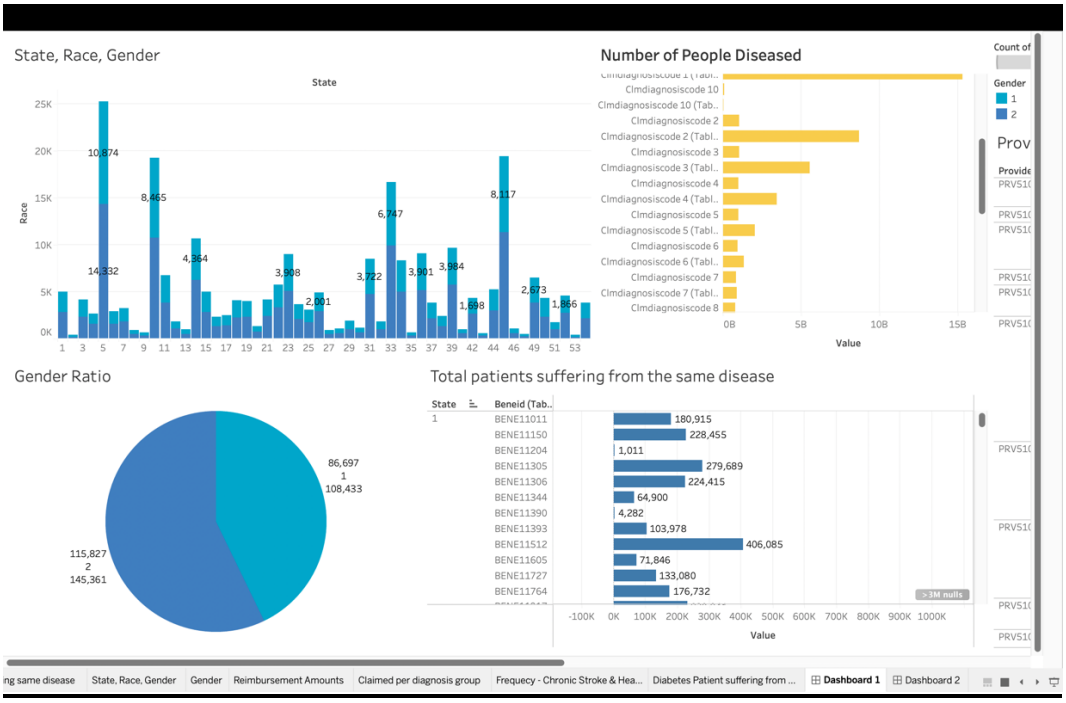
7. Finding Null Values



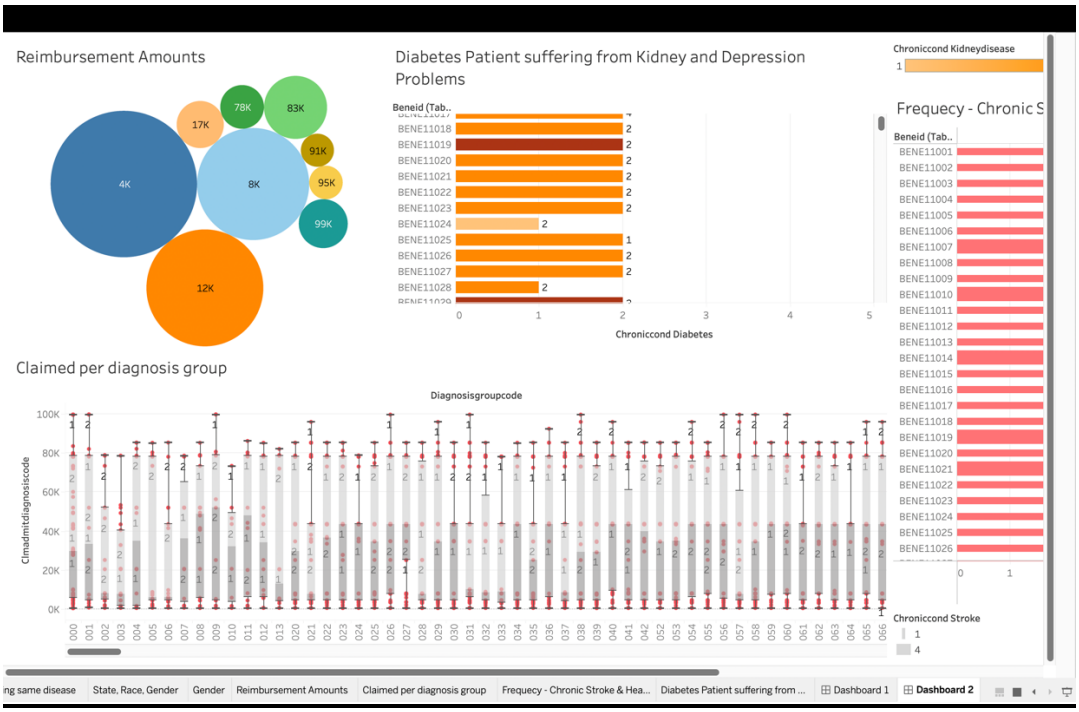
#	DOB	DOD	BeneID	Gender	State
3	1936-07-01		BENE11010	2	41
4	1914-03-01		BENE11011	2	1
5	1938-04-01		BENE11014	2	45
6	1919-09-01		BENE11016	2	34
7	1940-06-01		BENE11017	2	31
8	1924-05-01		BENE11019	1	7

Data Visualization

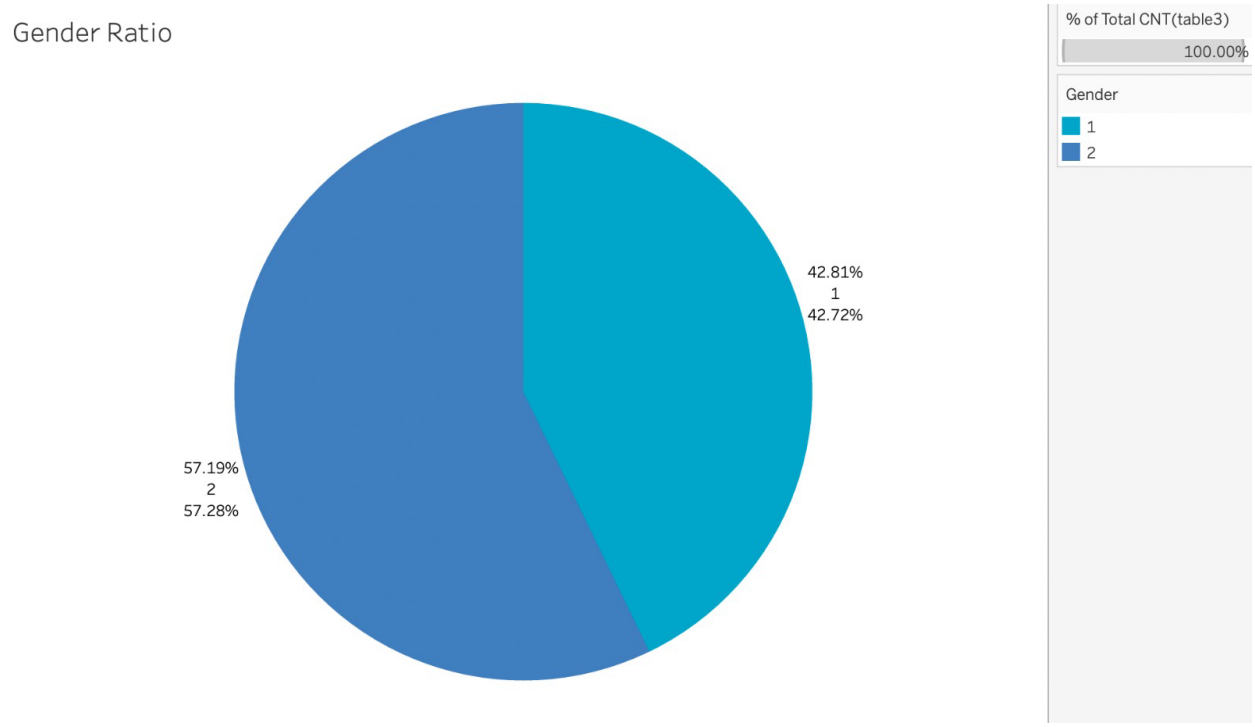
Dashboard:



TEAM -10 : Analysis of Health Insurance Claim Data

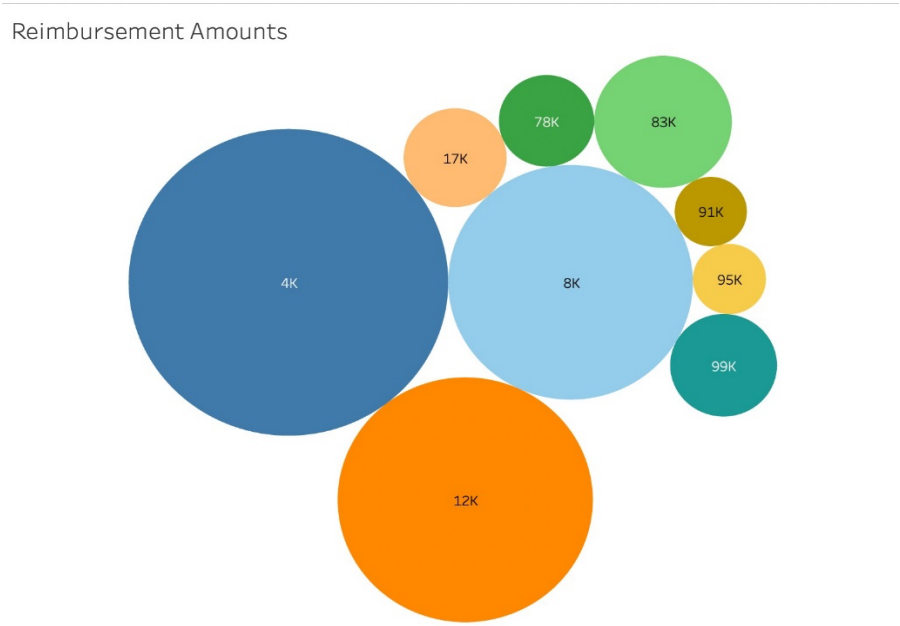


1. Gender Ratio

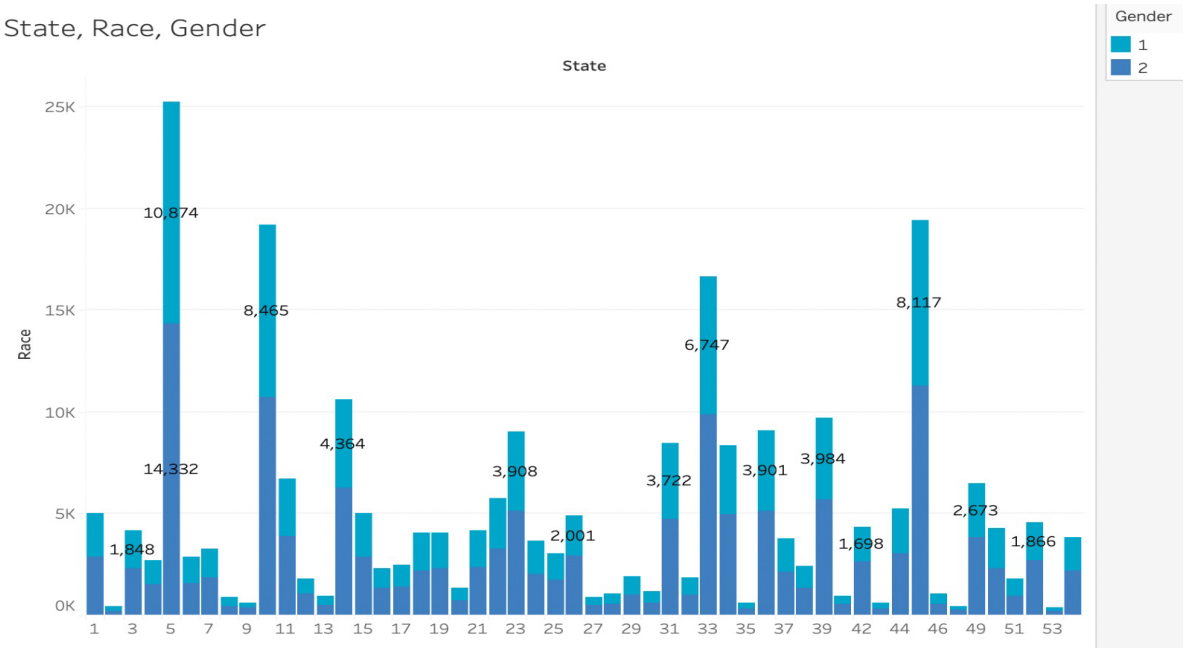


TEAM -10 : Analysis of Health Insurance Claim Data

2. Reimbursement Amounts

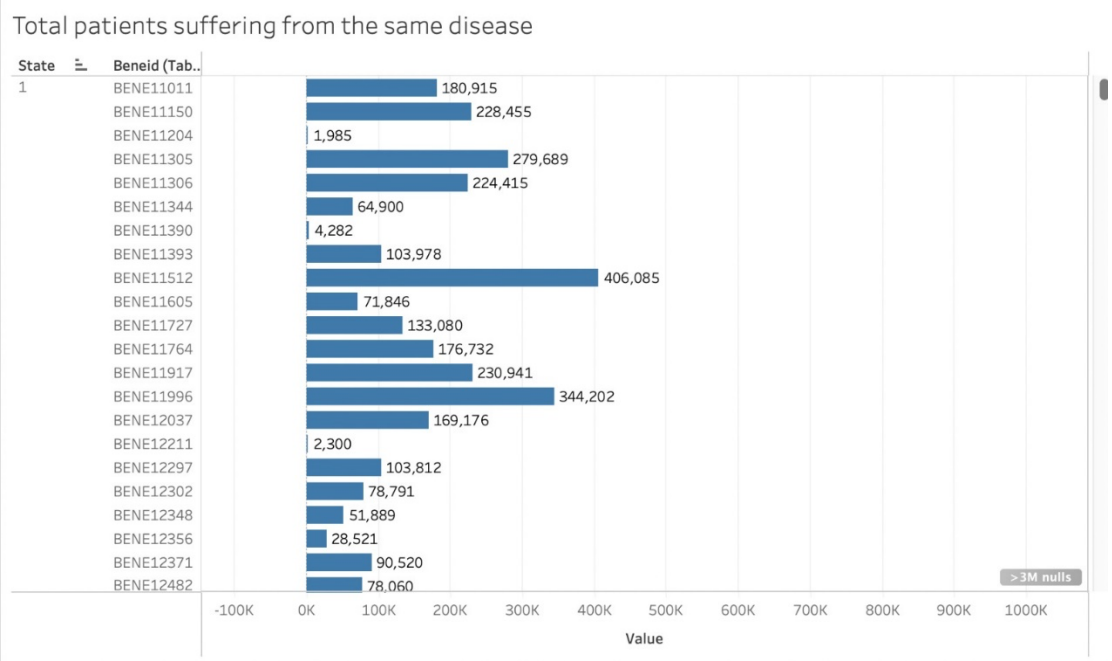


3. State VS Race VS Gender

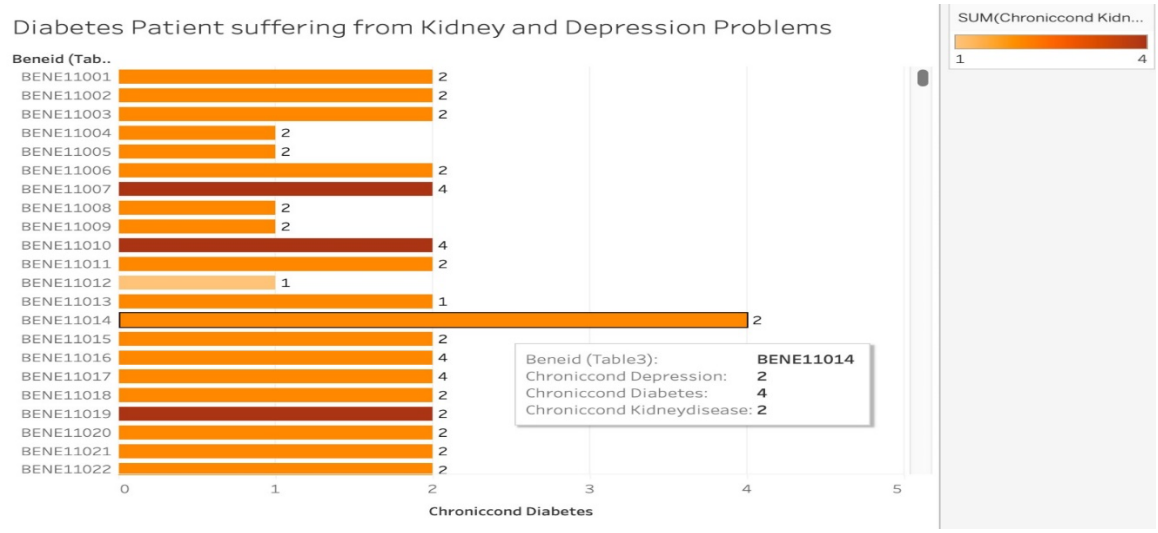


4. Patients suffering from disease

TEAM -10 : Analysis of Health Insurance Claim Data

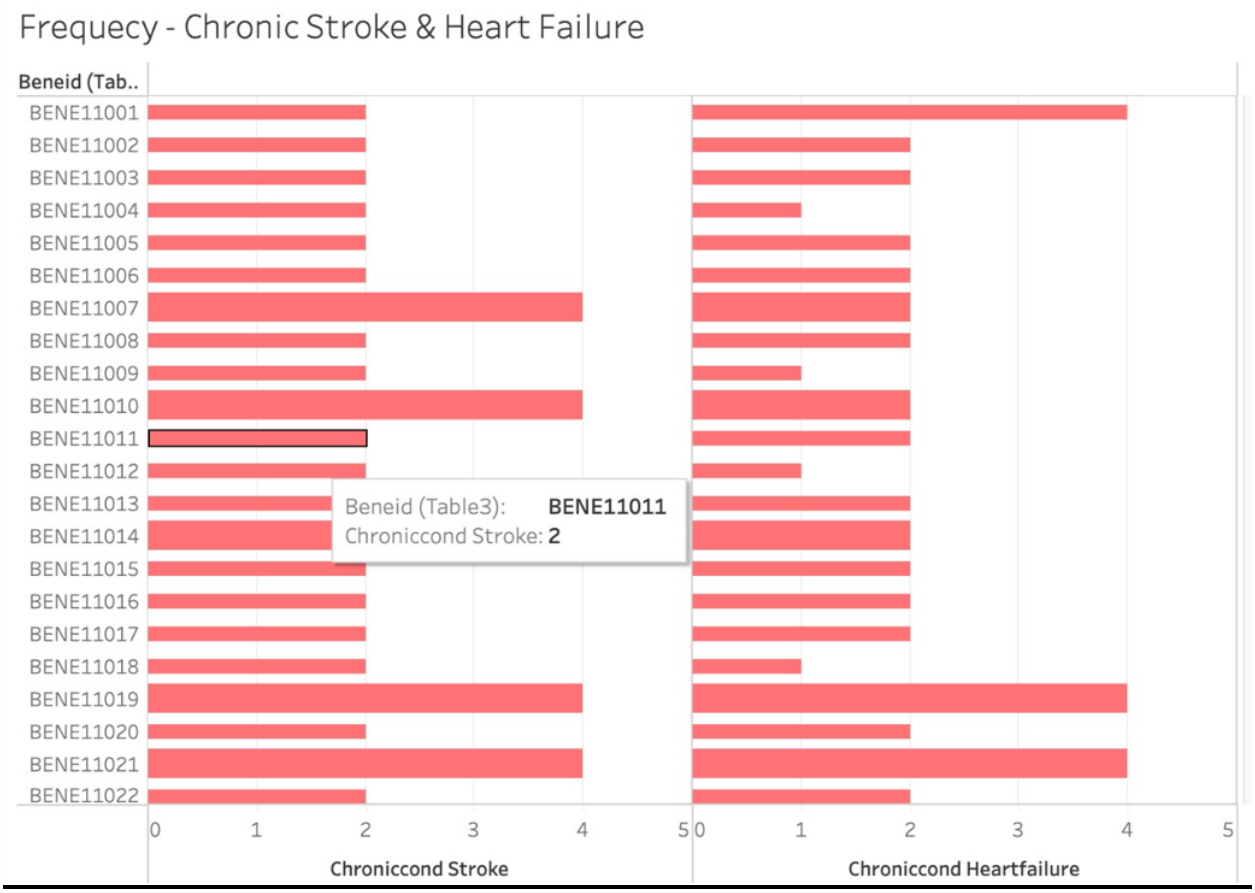


5. Diabetes Patients suffering other diseases



6. Number of Chronic Stroke & Heart Failures

TEAM -10 : Analysis of Health Insurance Claim Data



TEAM -10 : Analysis of Health Insurance Claim Data

References

1.How Many Americans Don't Have Health Insurance In 2022?

https://www.simplyinsurance.com/how-many-americans-dont-have-health-insurance/

2. Galvani, A. P., Parpia, A. S., Foster, E. M., Singer, B. H., & Fitzpatrick, M. C. (2020). Improving the Prognosis of Healthcare in the United States. Lancet (London, England), 395(10223), 524.

https://doi.org/10.1016/S0140-6736(19)33019-3

3. Health insurance status Maryland population 2020 | Statista.

https://www.statista.com/statistics/238765/health-insurance-status-of-the-total-population-of-maryland/

4. *Top 5 critical illness claims | BenefitsPRO*.

https://www.benefitspro.com/2012/06/22/top-5-critical-illness-claims/?slreturn=20220416173943