Heathcare Data Analysis Report

Muahmmad Ali

September 18, 2025

Introduction:

This dataset presents a comprehensive collection of hospital patient record metadata, capturing core details such as patient name, gender, blood type, insurance information, medication, and more. In addition, it also provides identifiers such as the doctor's name, hospital, billing amount, test results and other attributes that are relevant for a healthcare environment. Utilizing the metadata, this will enable analysis beyond the surface level revealing correlations between the identifiers and hospital operations when it comes to patients.

The central focus of this project is to identify and interpret patterns by using Python, SQL, and the visualization tool Power BI to answer questions such as:

- How do demographic factors such as gender and blood type relate to common treatments, lengths of stay, or medication regimens?
- Does a certain hospital have a different time length between admission and discharge date?
- Are there trends between blood types with certain medical conditions?

Through addressing these questions and future questions, this analysis seeks to deliver actionable insights into the functioning and optimization of patient care, administrative workflow, and whatever findings that are hoped to be unveiled through this analysis.

Ultimately, this report aims to elevate the understanding of hospital operations by transforming raw metadata into clear, actionable intelligence about patient needs, care quality, and institutional performance, fueling continuous improvement and innovative practice across the healthcare ecosystem.

Python:

Exploratory Data Analysis:

Initial Data Overview:

- There are 15 columns in this dataset.
- There are over 55,000 entries in this dataset with patient information.
- There are:
  - 2 int columns
  - 1 float column
  - 12 object columns
- There are no missing values or entries within this dataset.

Data Cleaning:

- As mentioned previously, there are no missing values.
- 534 duplicate entries are present in this dataset.
- Upon analyzing samples of the duplicate entries, they seem to contain the same information as the initial entry which will lead me to dropping the entries.
- Upon dropping the duplicate entries, the dataset now has 54966 entries in total.
- Additionally, the names of the `'Name'` column had to be set to all lower cases due to the casing of the metadata.
  For example: 'Samuel joYCe', would be unappealing to those reading the data, so forcing the entire column to change its casing to 'Samuel Joyce' or as it is called in the Python functions, `titles()`.
- In the `'Billing Amount'` column, the float values were represented as: '38142.109678'. Due to the column's nature of representing money, changing the decimal places was necessary and taken. All entries in this column have been rounded to 2 decimal places.
- In the `'Billing Amount'` column, it was also noted there were negative billings in the dataframe which is not accepted because it would indicate there was a misinput when given this data. The choice was made to drop those with negative billings as there were only 108 entries out of 54,966 in the dataframe.

Data Transforming:

- Appended a new column `'treatment_duration'` to show how long it took a patient to get treated from time of admission to discharge.
- There are identifiers for this dataframe such as, `'Name'` however, for simplicity sake and future SQL analysis, I have added new columns for this dataframe:
  - `'patient_id'`
  - `'doctor_id'`
- Changed all column names to all lowercases. Additionally added '_' to those column names that had spaces initially.
- Changed the column order to make it simpler for SQL use later.

SQL:

Using what was used in the Python segment of the project, new columns were added to the dataset to allow relations to be created between the columns and to make the schmea.