Heathcare Data Analysis Report

Muhammad Ali

October 4, 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:

Data Overview:

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.

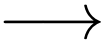
Data Transforming:

- 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.

Listed below is the initial dataframe information of columns and the dtypes.

Column	Dtype
Name	object
Age	int64
Gender	object
Blood Type	object
Medical Condition	object
Date of Admission	object
Doctor	object
Hospital	object
Insurance Provider	object
Billing Amount	float64
Room Number	int64
Admission Type	object
Discharge Date	object
Medication	object
Test Results	object

Table 1: Initial Dataset



Column	Dtype
patient_id	int64
doctor_id	int64
patient_name	string
age	int64
gender	category
blood_type	category
medical_condition	string
admission_date	datetime64[ns]
discharge_date	datetime64[ns]
treatment_days	int64
admission_type	category
room_number	int64
doctor_name	string
hospital_name	string
insurance_provider	string
medication	string
test_results	category
billing_amount	float64

Table 2: Transformed Dataset

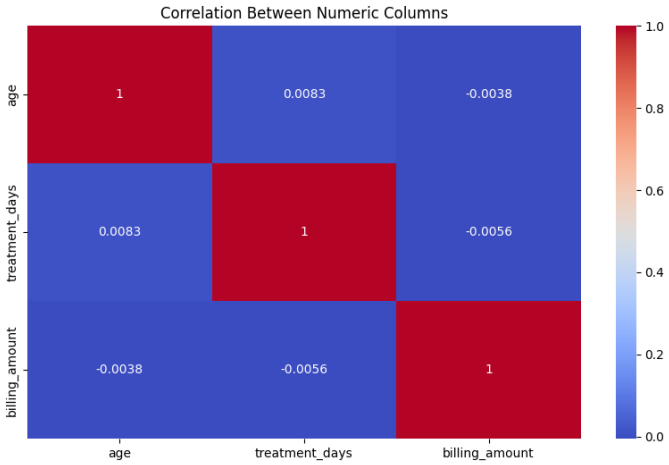
- Looking at the differences in these two tables, the changes are clear as the data has been transformed to satisfy the needs of this data analysis and for further development in SQL.
- As mentioned, new columns were added and the data types were changed to help make sense of the columns and what they represent than just and 'object' type.

Exploratory Data Analysis:

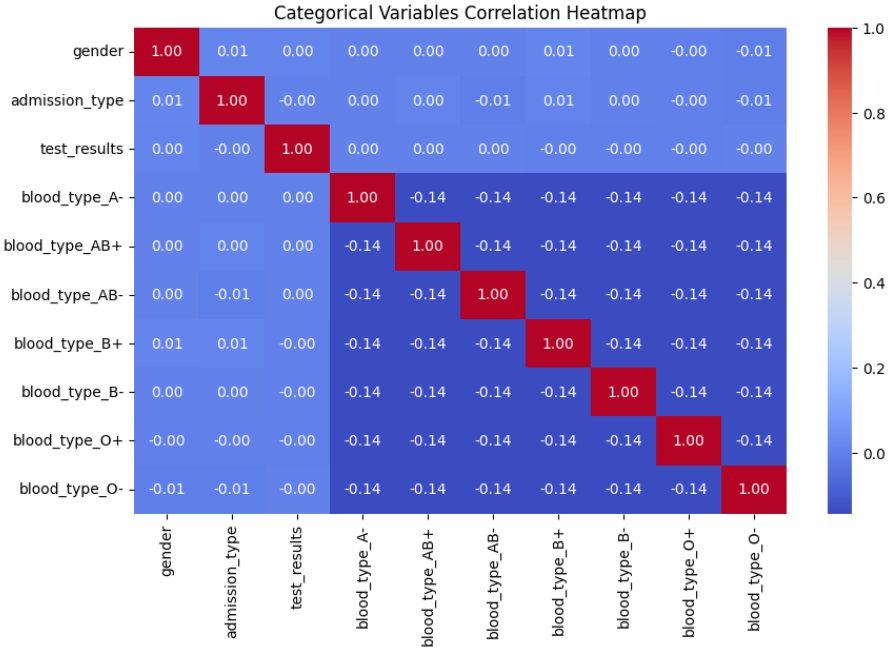
Correlation Heatmaps:

In this portion of the analysis, it needs to be determined if there are correlations between the columns i.e. if one affects the other more than the other, etc.

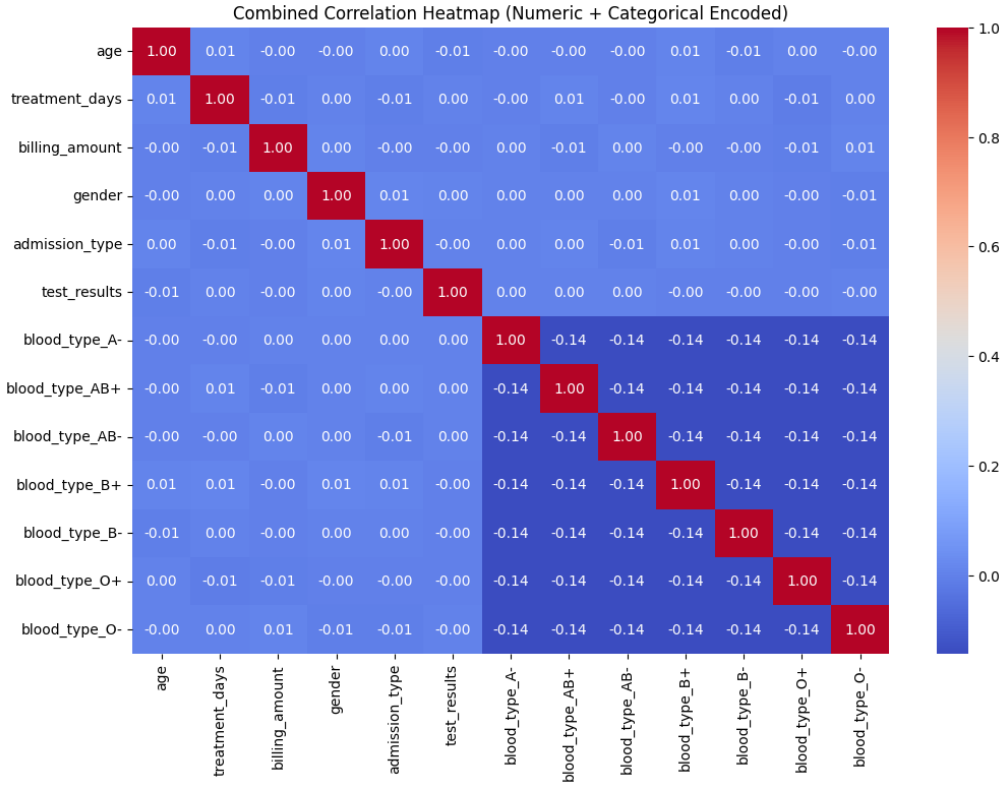
1. **Numeric Correlation:**



2. **Categorical Correlation:**



3. **Numeric & Categorical Correlation:**

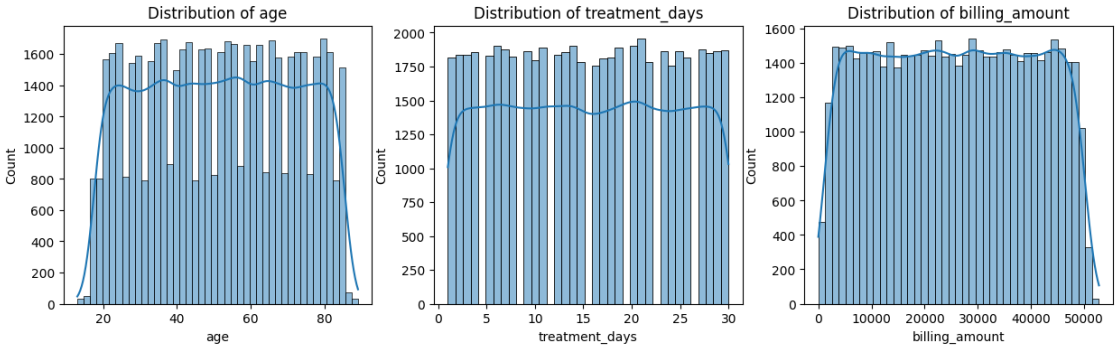


What can be gathered from the correlation heatmaps:

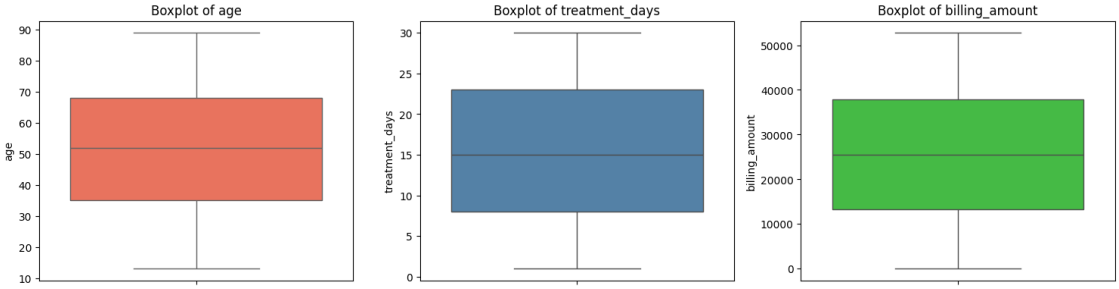
- Overall in all 3 correlation heatmaps, there are no signs of strong positive or negative correlations.
- The only thing present is either no correlation, or a weak negative or positive correlation.
- Interestingly enough, there seems to be a small weak positive correlation with *age and treatment days*, which will have statistical testing to see how significant it is.
- Billing amount and treatment days have a weak negative correlation, which statistical testing will be conducted.
- As the analysis goes on, statistical testing will be conducted.
- As far as blood types, they all seem to be consistent with each other, showing 0.14 with each other.
- Nothing too eye catching from the combined correlative heatmap.

Distributions & Counts:

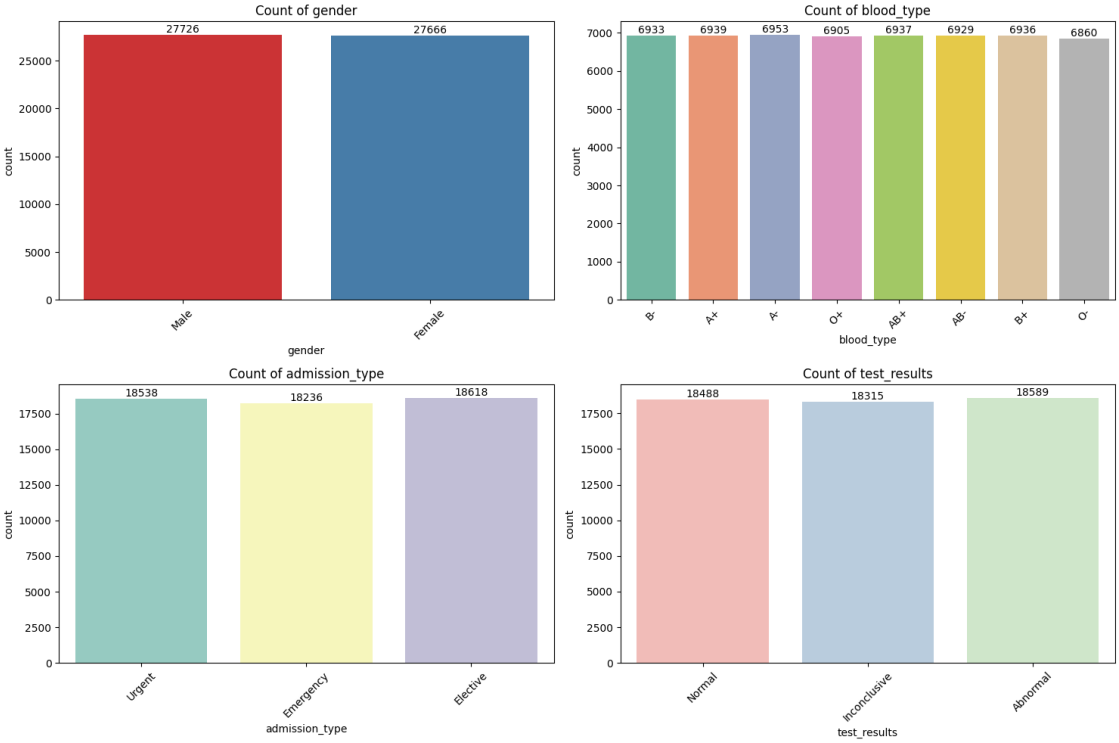
1. **Numeric Columns:**



2. **Numeric Box Plots:**



3. **Categorical Counts:**



What can be taken from these images?

- In the numeric columns distribution graph, it shows a uniform distributions among the 3 separate graphs. This is showing this has a wide spread of age, treatment days, and the billing amount.

- In the box plot graphs, it was tested to see if there were any outliers in the data for the numeric columns, and there seems to be none.
- Finally the categorical count illustrates the differences between the category columns in 4 separate graphs and the specifics within them, and they seem to be even between all the categories within the column.

Statistical Testing:

Relationship between Age and Treatment Days

Test Used: Pearson correlation coefficient (Pearson's r)

Null Hypothesis (H_0):
There is no correlation between `age` and `treatment_days`.

$$\rho = 0$$

Alternative Hypothesis (H_1):
There is a significant correlation between `age` and `treatment_days`.

$$\rho \neq 0$$

Test Results:
Pearson correlation coefficient: $\rho = 0.0083$
p-value: 0.0495

From this statistical test, the 0.0083 value from the correlation heatmap and this statistic value does indeed correspond to a **weak** positive correlation. Additionally with the p-value being less than 0.05, it also concludes it is statistically significant which leads to rejecting the null hypothesis, as there is some correlation but not too much.

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 schema.

Entity Relational Diagram: