Heathcare Data Analysis Report

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 analysis 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.

For full report reference, click the link to direct you to the full findings and analysis for each programming language:

- Python
- SQL
- Patient Information Dashboard
- Doctor Information Dashboard
- Financial Information Dashboard

Python:

In the data cleaning portion, only the columns and duplicate entries had to be changed as it would give the data more clarity and less skewness when conducting analysis on Python and SQL. Additionally, data types had to be changed in order to make the columns viable for analysis, name changes are mentioned on the referenced Python analysis.

Following transforming and cleaning the data, correlations were conducted among the columns. Seeing what column affects another is crucial to understanding what can help in this specific healthcare environment, if not specific a general basis.

While doing a correlation visualization, I came with up with the idea to do 3 correlations, numeric, categorical and numeric + categorical respectively. The insights that were drawn out before further analysis were the following:

- No strong correlations, meaning there were no columns that had a stronger impact on the other.
- It was either no correlation or a small impact.
- There was a slight weak positive correlation with age and the amount of days a patient had to stay.
- The longer a patient stays, it would affect their bill.

After the correlation visualization, another series of visualizations was applied. This was in order to see if there were any anomalies, outliers in the dataset. After conducting these graphs, it was concluded that the data contained a normal distribution with no outliers, and everything was balanced meaning nothing was higher or lower than the other; everything was evenly split.

Following this, statistical testing was conducted on the eye-catching correlations, and then there were other statistical testing conducted to make sure there was nothing hidden from the visualizations.

There were 4 distinct statistical testings conducted with different types of tests between columns to determine validity of the correlations and appropriate testing. The following tests were conducted:

- Pearson correlation coefficient (Pearson's r)
- Welch's T-Test (Unequal variance t-test)
- One-way ANOVA
- Chi-Squared Test of Independence

Using these statistical testing methods, the following insights were uncovered:

- Against treatment days and age, there is no correlation between the two categories.
- Treatment days vs Billing amount, there is no correlation between the two.
- There is no difference in billing between the two genders.
- The average billing amount is the same across the admission types.
- There is no correlation between gender and medical condition.
- Between test results and insurance provider, there is no correlation between each other.
- There is no relationship between blood type and test results.

From these tests, it is shown that majority of the relationships are independent of each other and/or have a weak correlation with each other. Following this, one column does not affect another so everything can be focused as one entity instead of worrying about another category being affected.

The last segment of this analysis was clustering columns to see what useful insights would be gathered. In this process, 4 groups of clusters were conducted as there were relevant columns needed separately. The following cluster groups are as follows:

Patient Groups
Age
Treatment Days
Billing Amount

Clinical Severity
Age
Treatment Days
Various Medical Conditions
Test Results
Medications

Utilization & Cost
Treatment Days
Billing Amount
Room Number
Admission Types

Insurance & Payment
Billing Amount
Treatment Days
Various Insurance Providers

With these groups in mind, the following interpretations were taken for each of the clustered groups:

- Patient Groups:
  - Younger patients with longer treatment durations and lower-than-average billing amounts.
  - Very young patients with slightly shorter treatment durations and higher-than-average billing amounts.
  - Older patients with longer stays and higher-than-average billing amounts.
  - Older patients with short treatment durations and lower-than-average billing amounts.
- Clinical Severity:
  - The analysis revealed four distinct patient clusters, which were created with similarities in age and length of hospital stay.
  - Younger patients tended to have shorter hospital stays, whereas older patients stayed longer, but other factors like medical conditions or medications did not differ much between the groups.
  - Age and treatment duration were the main factor of patient grouping, while clinical and medication patterns remained largely similar across all groups.
- Utilization & Cost:
  - Patients were divided into four distinct groups based on treatment duration, billing data, and admission details to uncover shared patterns.

  - Two clusters consisted of individuals with extended hospital stays, while the other two reflected shorter treatment periods. Despite these differences in length of stay, billing amounts, room assignments, and admission types remained largely consistent across all groups.
  - Treatment duration emerged as the key factor distinguishing patient categories, with the other variables playing only a minor role in this cluster distribution.

- Insurance & Payment:
  - Patients were grouped into four categories based on treatment duration, billing amounts, and insurance coverage.
  - The clusters portrayed distinct patterns: one group had extended stays with lower costs, another had short, low-cost treatments, a third showed brief but high-cost care, and the last combined long stays with high expenses.
  - Overall, length of stay and billing amount were the main factors distinguishing the groups, while insurance provider distribution such as Blue Cross, UnitedHealthcare, and Medicare showed only minor variation across clusters.

From the results of my Python analysis, it was revealed the true relationships between columns, if there were any abnormalities within the data in which there wasn't as this dataset's entries was accurate enough to draw valuable results.

SQL:

Starting the SQL analysis required relational tables to draw out insights on based on the tables created. In this instance, 6 tables were created which consisted of:

1. Hospital
2. Doctor
3. Patient
4. Admission
5. Medical Record
6. Billing

Each of these tables contained specific information that only pertains to that category with some relationships that connects to another table. An example of this would be `Patient` and `Admission`, the connections between these two would be a patient's identification with their admission identification which is connected to the patient's identification in the `Admission` table.

Following creating relationships between tables, thus began the analysis from basic investigation to advanced investigations. To start with this, an investigation over counts and average distributions over certain questions. This provides an adequate start to what the data represents, how much is given to work with, and if deeper analysis is needed.

With the analysis starting off with patient demographics, it was found that there were:

- 55,972 Patients
- Almost an equal split amongst male and female patients
- Age ranging from the youngest being 13 and the oldest being 99
- The top condition being arthritis followed by: obesity, asthma, hypertension, cancer, and diabetes
- Blood type distribution is uniform, meaning they are all around 6,800 patients
- The average age of both genders are 51

From this analysis, it shows a balanced scale of genders and blood type distributions.

Following patient demographics, the next analysis conducted was most common condition based off age group, the following was found:

- Ages 10-20: Arthritis & Cancer
- Ages 21-30: Obesity
- Ages 31-60: Balanced spread of all conditions
- Ages 61+: An increase in hypertension & arthritis

Based off the data, younger patients face obesity and arthritis whereas older patients face hypertension and arthritis.

After discovering conditions per age group, the next thing that was necessary to investigate were the hospital & doctor analytics:

- There are 33,602 distinct hospitals with a range of 1 to 37 patients administered.
- After investigating how many patients each doctor works with, there were instances where a spread of doctors had to deal with more than 20 patients, the highest was 27.
- In each hospital, it was only shown 1-2 conditions for most of them, however as the results show, its distributions expand.

To sum up, administered patients spread is high across all hospitals with certain doctors working with 1 or more.

Working off hospital and doctor analytics, the next relevant piece of information needed was treatment and billing information.

- The average treatment days across all conditions and admission types was around 15 days

- By insurance the accumulated billing were:
    - Cigna ($287M)
    - Medicare ($285M)
    - Blue Cross ($283M)
    - UnitedHealthcare ($282M)
    - Aetna ($278M)
- For patients that stay longer than 15 days (long-stay), their billings would be around $25,000 per patient
- Long stays generate higher bills across admission types

Cigna and Medicare have the highest accumulated bill as an insurance provider but Aetna has the lowest. In addition, longer stays generate higher billings.

With treatment information, it builds a segue to the next analysis, how long a doctor's average treatment is and monthly billings:

- Doctors with many patients have an average of 13 to 18 days
- The month with the lowest billing is February ($107.9M) whereas the highest is July ($122.7M)
- This creates a question of seasonal months causing peaks.

Billings are based off seasonally, meaning heat stroke, the flu or seasonal illnesses can be a factor of high monthly billing. Doctor workload and treatment days vary.

To sum up:

- Balanced split of patient demographics
- There are age related trends with conditions per age group
- Certain doctors are burdened with more patients along with many hospitals with only 1 patient per doctor
- Treatment duration affects billing
- Seasonal billing peaks can be linked to natural occurences in the month; the flu, heat stroke, or public health events

Power BI:

Three dashboards were created, each presenting key information in distinct areas:

1. Patient Information
2. Doctor Information
3. Financial Information

Patient Information:

This dashboard provides an overview of patient-related data. Filters for year and month allow comparisons across different time periods. Key metrics are displayed using cards, including:

- Total patients
- Average age
- Average treatment days
- Median treatment days
- Number of long-stay patients
- Total billing accumulated

Visualizations include:

- Blood type distribution
- Frequency of stays in 5-day intervals
- Pie chart of admission types
- Line graph showing monthly admission trends
- Combo chart: admission types (bar) with average stay trends (line)
- Combo chart: admission volumes by condition (bar) with average stay trends (line)

All visualizations can be filtered by year or month for customized comparisons.

Doctor Information:

This dashboard highlights data relevant to doctors and management, ensuring workloads are balanced. Summary cards include:

- Patients per doctor
- Total doctor count
- Treatment days per doctor
- Average billing per doctor

Filters are available for year, month, and condition. Visualizations include:

- Bar chart showing doctor effectiveness (patients and admissions)
- Bar chart illustrating doctor workload
- Scatter plot of caseload vs. efficiency, with an average reference line
- Treemap of medical conditions by patient volume

- Treemap of blood type distributions
- Stacked horizontal bar chart of top 10 doctors by patient count, split by condition
- T-chart of doctors and their average treatment days

All charts can be filtered by year or month for comparative analysis.

Financial Information:

This dashboard focuses on insurance and billing data. Filters include:

- Condition
- Insurance provider
- Year
- Month
- Admission type

Key metrics displayed are:

- Average billing
- Patient counts
- Average treatment days
- Total billing
- Lowest-paying insurance provider
- Highest-paying insurance provider

Visualizations include:

- Ribbon chart of billing across insurance providers by month
- Combo chart: billing sum per provider (bar) with monthly average billing (line)
- Ribbon chart of total billing by condition and insurance provider
- Scatter plot of billing vs. average treatment days, illustrating provider billing relative to patient counts
- Stacked column chart of total billing by provider with admission types
- Heatmap showing insurance overpayments with total sums

All visualizations can be filtered by year or month to support detailed financial comparisons.

Final Remarks:

This analysis demonstrates not only how important data analysts' work is, but also how much synergy they may have with a worker, in this instance, doctors. Being able to go through data to find a story inside this dataset revealed patient patterns, doctor workload, correlations, and how everything can be linked together to form this one story. Using tools and features like as statistical testing, modeling, interactive dashboards, and so on, it revealed that there are hidden truths behind the surface of the data, which can help generate efficiency and suit the demands of patients or doctors.

In the future, there is opportunity for expansion of this analysis with ideas such as predictive analysis, real-time monitoring with dashboards to anticipate demands, balance workloads in real time, and manage monetary resources. Such advancements in analysis not only increase efficiency, but also improve quality of care and anticipation.

Lastly, this analysis highlights how raw hospital metadata can be transform into actionable intelligence. It incorporates clinical practice, administration, and finance, reminding us that healthcare analytics is more than just data; it's about making better decisions, enhancing patient experiences, and driving innovation throughout the healthcare ecosystem. This applies not only to this analysis, but to any other as well.