BI Project

Asna Farooqui-26918, Saad Thaplawala-27172,
Uzair Nadeem-24928, Zainab Hasan- 27127

Data Set: Healthcare

Dataset Description:

Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here's a brief explanation of each column in the dataset:

- **Name:** This column represents the name of the patient associated with the healthcare record.
- **Age:** The age of the patient at the time of admission, expressed in years.
- **Gender:** Indicates the gender of the patient, either "Male" or "Female."
- **Blood Type:** The patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.).
- **Medical Condition:** This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.
- **Date of Admission:** The date on which the patient was admitted to the healthcare facility.
- **Doctor:** The name of the doctor responsible for the patient's care during their admission.
- **Hospital:** Identifies the healthcare facility or hospital where the patient was admitted.
- **Insurance Provider:** This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare."
- **Billing Amount:** The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.
- **Room Number:** The room number where the patient was accommodated during their admission.
- **Admission Type:** Specifies the type of admission, which can be "Emergency," "Elective," or "Urgent," reflecting the circumstances of the admission.
- **Discharge Date:** The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range.

- **Medication:** Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor."
- **Test Results:** Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal," "Abnormal," or "Inconclusive," indicating the outcome of the test.

Industry Background:

Healthcare is an increasingly data-driven industry as hospitals generate huge amounts of patient data with admissions, diagnostics, treatments, and billing. However, due to the departmentalization of hospitals, this data is very fragmented which leads to inefficiency in operations as well as lack of decision-making analysis.

As healthcare systems grow more complex, there is a rising demand for Business Intelligence (BI) tools that can consolidate data, uncover trends, and support strategic decision-making. Hospitals now rely on BI to monitor treatment costs, predict seasonal admission spikes, and identify high-risk patient demographics for targeted intervention.

Data Cleaning and EDA:

A thorough data cleaning and enrichment process was conducted in order to remove inconsistencies and problems from the dataset. EDA was then done to achieve a better understanding of the data and the distributions of different numerical and categorical columns.

The following steps outline how the cleaning and enrichment process was done:

1. Incorrect capitalization in the Name (patient name) and Doctor (doctor name) columns was corrected, and the unique values in each categorical column were verified to have valid entries.
2. The Admission Date and Discharge Date columns were converted to a datetime data type and a new column was calculated, representing the Length of Stay of a patient in days, by subtracting the Admission Date from the Discharge Date.
3. A significant number of rows had a negative value in the Length of Stay column, and since there was no logical way to deal with these values, the rows had to be dropped.
4. Rows with a negative billing amount were also dropped.
5. A new column, Age Group, was created. Patients between the ages of 13-19 were categorized as teens, patients between 20-30 as young adults, patients between 31-50 as middle aged, patients between 51-69 as middle aged, and patients older than 69 as seniors.

The resulting clean and improved dataset was analyzed to observe how the data is distributed among different categories and numerical ranges. The key findings of this process are given below:

- While billing amounts and ages had a relatively symmetric distribution, length of stay was skewed to the right, indicating a higher number of stays of shorter length.
- All categorical columns had an approximately similar, equal distribution.
- There was no correlation found between age, billing amount, or length of stay.
- There was a statistically significant difference found between the mean billing amounts of blood groups A+ and O+ and blood groups O+ and O-.
- There was a statistically significant dependency found between the Age Group and Medical Condition columns.

The dataset was then exported to a csv file which was to be loaded into Power BI in order to build a solution to our problems.

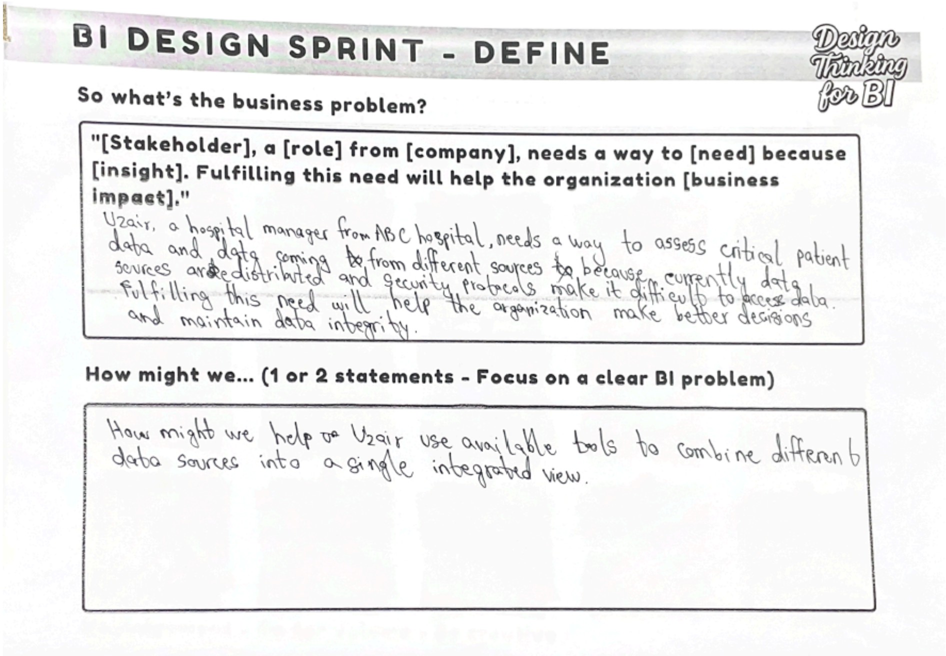
Problem Statements:

The final set of problems that our BI solution tries to solve is:

1. Are there any specific age groups, genders, or blood groups that are more frequently associated with high-cost conditions?
2. For the identified demographics associated with high-cost conditions, what are the common patterns and historical trends related to their admissions (stay durations, common medical conditions, billing trends, etc.)?

Implementing the Design Thinking Framework

Design Sprint:



BI DESIGN SPRINT - DEFINE

Design Thinking for BI

So what's the business problem?

"[Stakeholder], a [role] from [company], needs a way to [need] because [insight]. Fulfilling this need will help the organization [business impact]."

Uzair, a hospital manager from ABC hospital, needs a way to assess critical patient data and data coming ~~to~~ from different sources ~~to~~ because currently data sources are distributed and security protocols make it difficult to access data. Fulfilling this need will help the organization make better decisions and maintain data integrity.

How might we... (1 or 2 statements - Focus on a clear BI problem)

How might we help ~~to~~ Uzair use available tools to combine differenb data sources into a single integrated view.

Our project began with defining a clear and impactful business problem:

> “Uzair, a hospital manager from ABC Hospital, needs a way to assess critical patient data and data coming from different sources because currently data sources are distributed and security protocols make it difficult to access data. Fulfilling this need will help the organization make better decisions and maintain data integrity.”

This problem emphasizes challenges around **data silos**, **inconsistent access**, and the lack of **integrated, actionable insights** for hospital management.

We conducted AI based empathy interview and user research to understand key frustrations and goals of stakeholders like hospital managers and analysts:

Uzair, the hospital manager, works in a **hectic and often tense environment**, where **decision-making is challenged** by inconsistent and scattered data. He **hears concerns** from department heads about **untrained staff and data problems**, and regularly **sees fragmented systems** across departments.

In his day-to-day, he **attends meetings, reviews reports, checks KPIs**, and **coordinates with vendors and partners**. However, the **lack of integrated, real-time data** slows down insights and creates frustration.

This revealed a clear need for a **centralized BI dashboard** that consolidates data sources, improves KPI tracking, and supports timely, informed decisions.

From user pain points, we crafted **“How Might We”** questions to guide ideation:

- How might we integrate data from multiple departments into one dashboard?
- How might we use analytics to identify high-risk patients based on demographics?
- How might we support preventive care through targeted insights?

Each team member contributed **idea sketches and Crazy 4s**, brainstorming possible solutions such as:

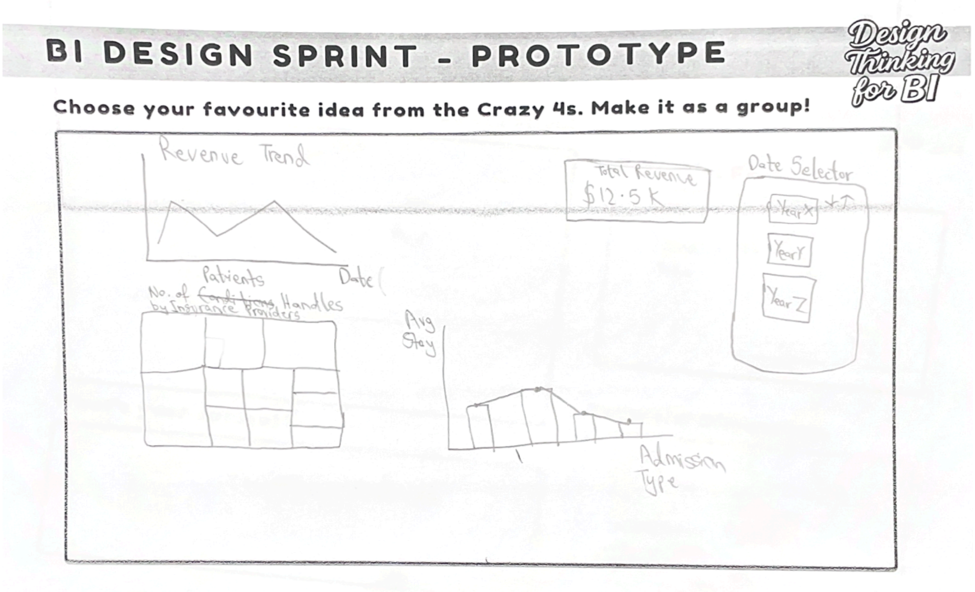
- A dashboard comparing billing amounts across doctors and hospitals
- A predictive panel for identifying high-cost demographics
- A time-series view of seasonal admission trends
- A summary of medication effectiveness by test result outcomes

The team converged on a prototype design that included:

- A centralized **BI dashboard** with drilldowns for patient demographics, cost analysis, admission types, and test outcomes
- Filters for age group, gender, blood type, and condition
- Visualizations to identify **billing anomalies**, **seasonal patterns**, and **preventive care opportunities**

This design sprint helped narrow the focus to two primary BI goals:

1. **Cost transparency by demographics and providers** to support data-driven decision-making
2. **Predictive and seasonal insights** for planning resources, improving patient care, and reducing preventable high-cost treatments


BI DESIGN SPRINT - PROTOTYPE

Choose your favourite idea from the Crazy 4s. Make it as a group!

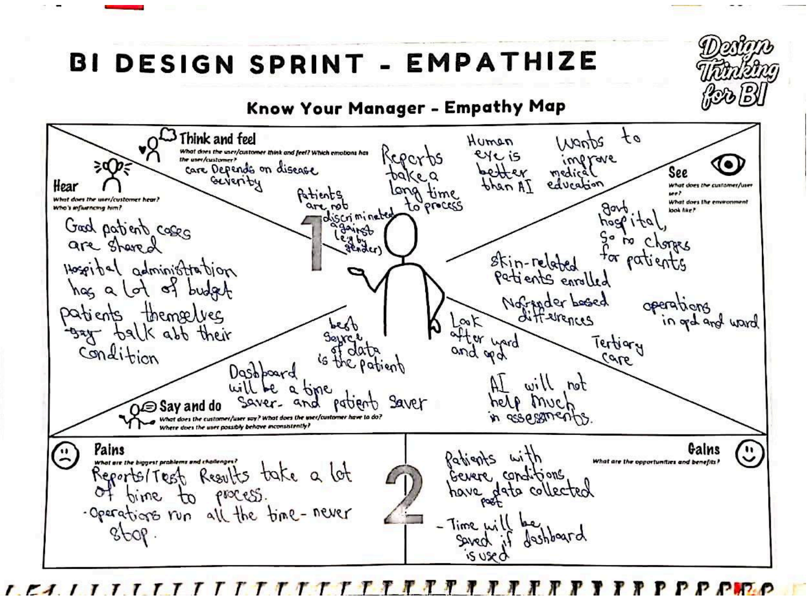
Interviewing a Domain Expert:

In order to obtain a clear picture and understanding of how patient billing and care is handled at hospitals and the challenges involved with managing information related to admitted patients, Dr. Humaira Talat, Head of the Dermatology Department at Civil Hospital, Karachi, was requested to provide insights into how patients are managed at the hospital. Her answers were also translated onto an empathy map to get a deeper look into the informational needs of stakeholders in related positions.

According to Dr. Humaira, patient care is majorly dependent on the severity of their condition, and not on other factors such as gender. The higher the severity, the more care a patient gets. Since the hospital admits a large number of patients that need extreme care, operations at the hospital almost never stop. This means workers are always trying to track patient information, and the faster they can access accurate information, the faster they can act on it and provide appropriate treatment to patients. Doctors currently rely on no real time source of patient information, causing delays in diagnosis and treatment. An example of this is doctors waiting to receive a patient's test results to determine their condition. A real time solution that provides them with accurate data would not only save time, but also save patients.

Dr. Humaira's LinkedIn

Empathy Map:



BI DESIGN SPRINT - EMPATHIZE

Design Thinking for BI

Know Your Manager - Empathy Map

Think and feel
What does the user/customer think and feel? Which emotions has the user/customer?
care Depends on disease severity
Patients are not discriminated against (e.g by gender)
Reports take a long time to process
Human eye is better than AI
Wants to improve medical education

Hear
What does the user/customer hear? Who's influencing him?
Good patient cases are shared
Hospital administration has a lot of budget
patients themselves ~~say~~ talk abt their condition

See
What does the customer/user see? What does the environment look like?
govt hospital, So no charges for patients
skin-related patients enrolled
No gender based differences
operations in opd and ward
Tertiary care

Say and do
What does the customer/user say? What does the user/customer have to do? Where does the user possibly behave inconsistently?
best source of data is the patient
Dashboard will be a time saver- and patient saver
Look after ward and opd
AI will not help much in assesments.

1

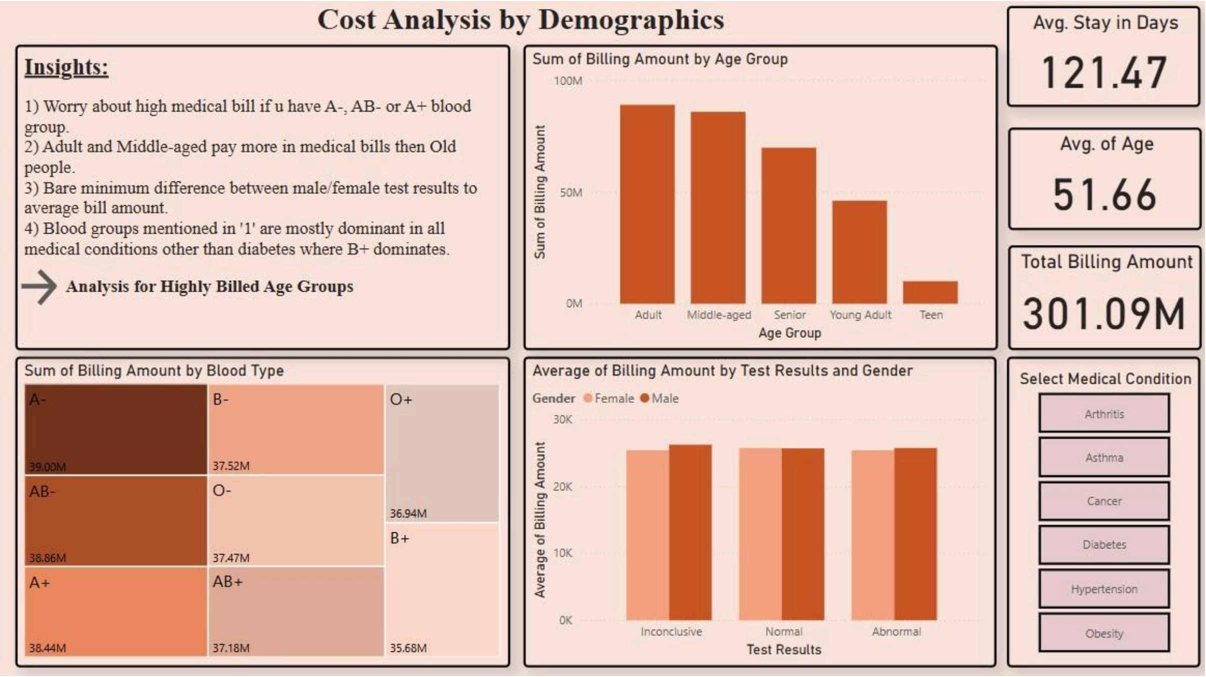
Pains
What are the biggest problems and challenges?
Reports/Test Results take a lot of time to process.
·Operations run all the time- never stop.

2

Gains
What are the opportunities and benefits?
Patients with severe conditions have data collected post
- Time will be saved if dashboard is used

This empathy map has also been submitted along with other required deliverables on google drive.

Dashboard and Chart Explanation:



Cost Analysis by Demographics

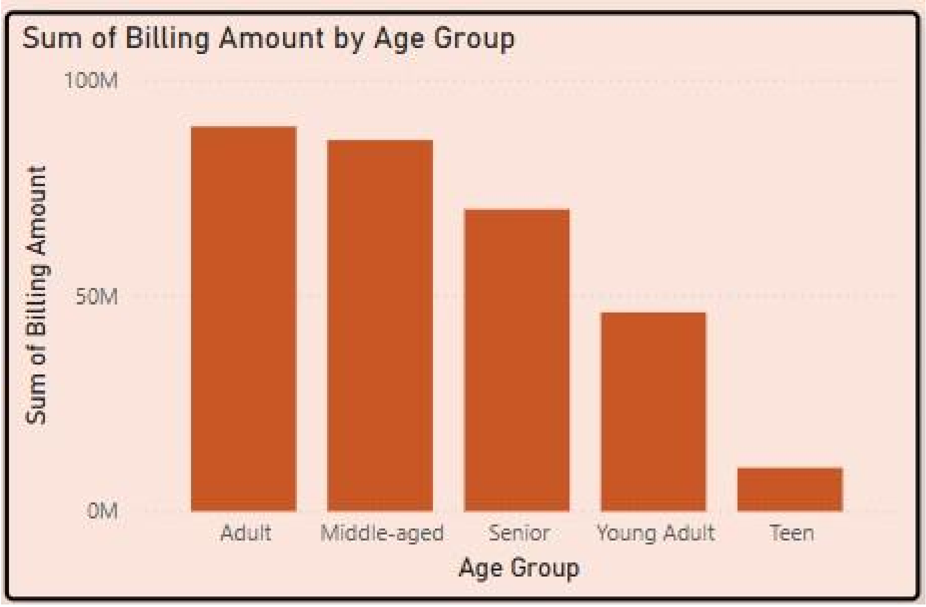
Insights:

1) Worry about high medical bill if u have A-, AB- or A+ blood group.
2) Adult and Middle-aged pay more in medical bills then Old people.
3) Bare minimum difference between male/female test results to average bill amount.
4) Blood groups mentioned in '1' are mostly dominant in all medical conditions other than diabetes where B+ dominates.

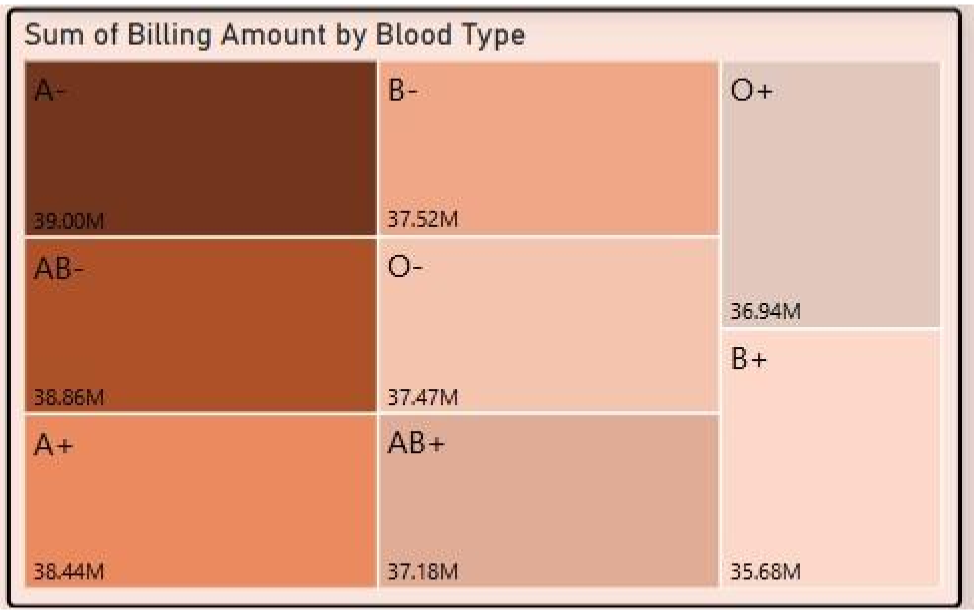
→ **Analysis for Highly Billed Age Groups**

Avg. Stay in Days	Avg. of Age	Total Billing Amount
121.47	51.66	301.09M

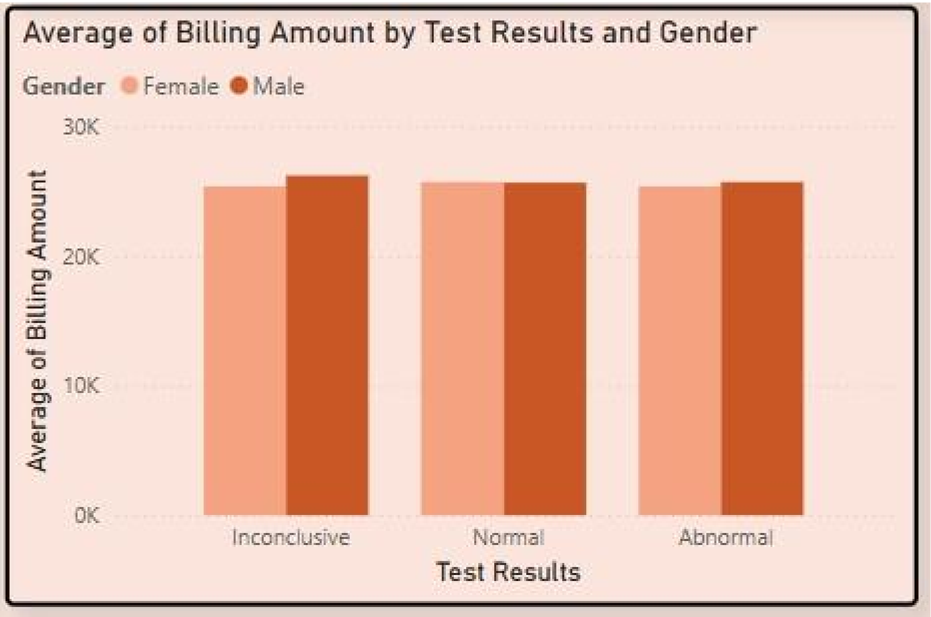
Page1: Overall view of dataset, analysis by sum of billing amount and medical condition.



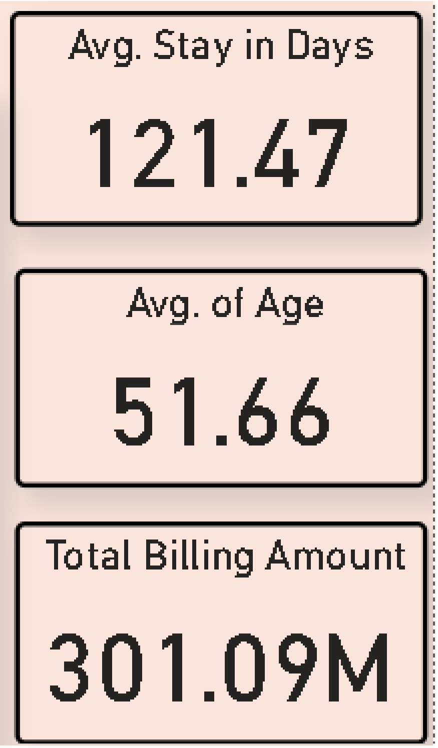
Total bill for each age group visual view for further analysis into differences by age group.



Treemap to show proportional view of total bill by each blood group to identify and breakdown top groups.



Seeing if there is any relation to test results to average amount spent in hospital bills and difference between gender in this case.



Avg. Stay in Days

121.47

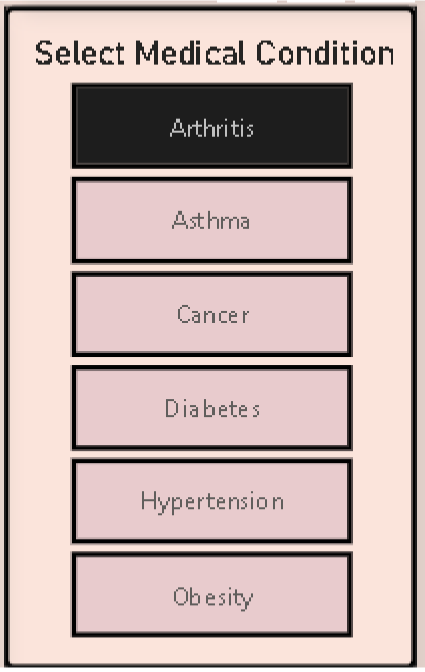
Avg. of Age

51.66

Total Billing Amount

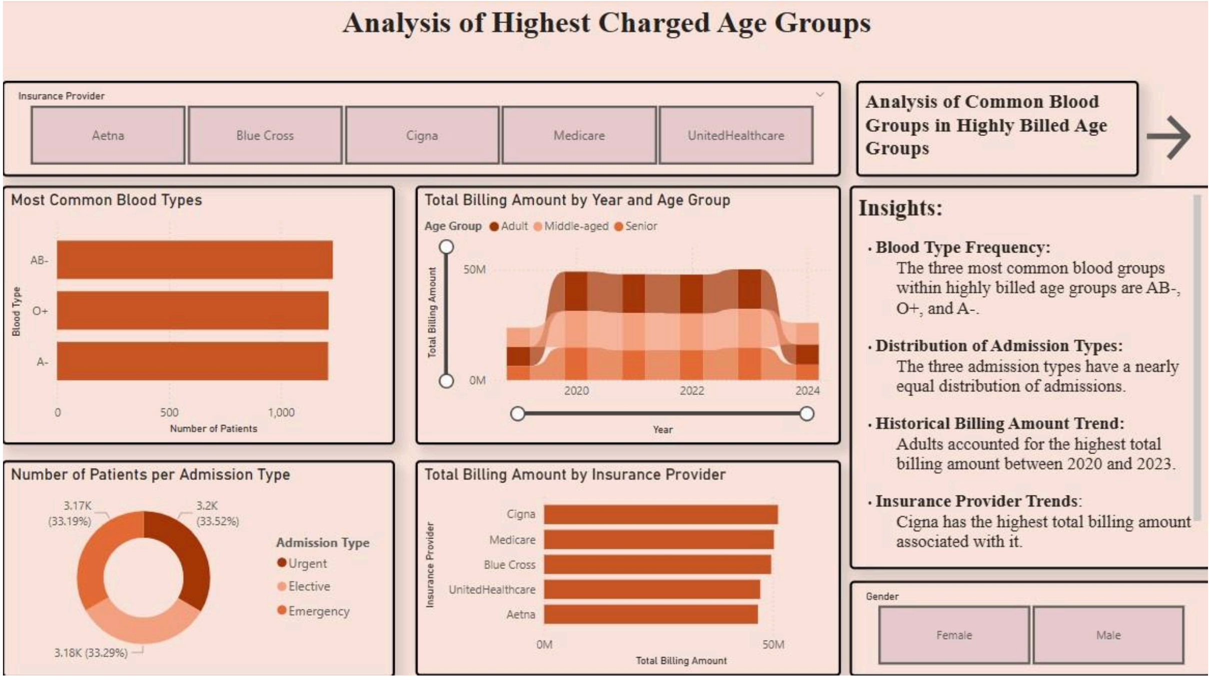
301.09M

KPI’s to show important numbers directly.



Select Medical Condition

Select medical conditions to drill down or see changes by different medical conditions or see by combination of both.



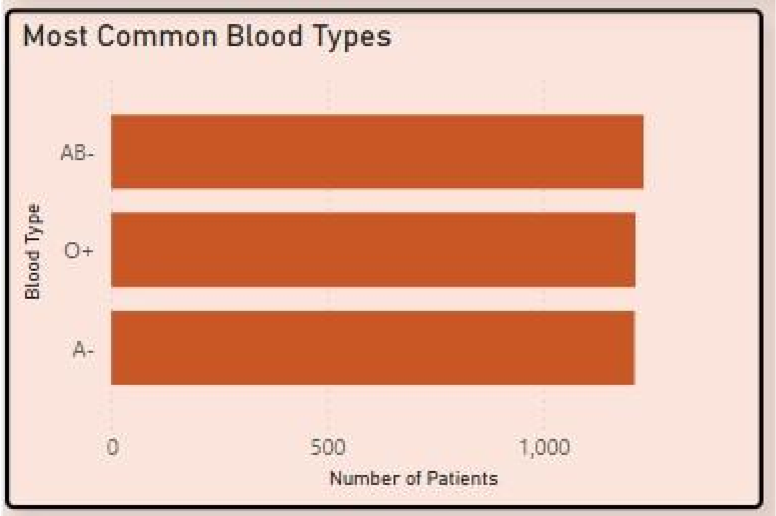
Analysis of Highest Charged Age Groups

Analysis of Common Blood Groups in Highly Billed Age Groups

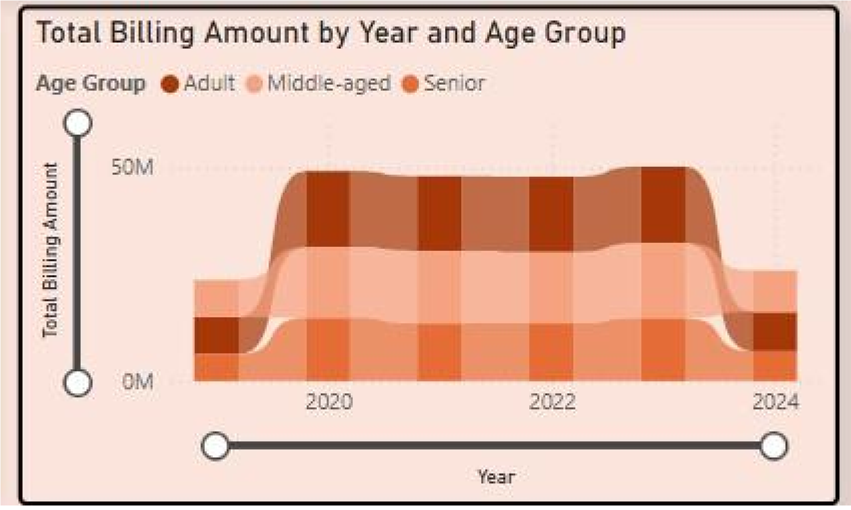
Insights:

- **Blood Type Frequency:** The three most common blood groups within highly billed age groups are AB-, O+, and A-.
- **Distribution of Admission Types:** The three admission types have a nearly equal distribution of admissions.
- **Historical Billing Amount Trend:** Adults accounted for the highest total billing amount between 2020 and 2023.
- **Insurance Provider Trends**: Cigna has the highest total billing amount associated with it.

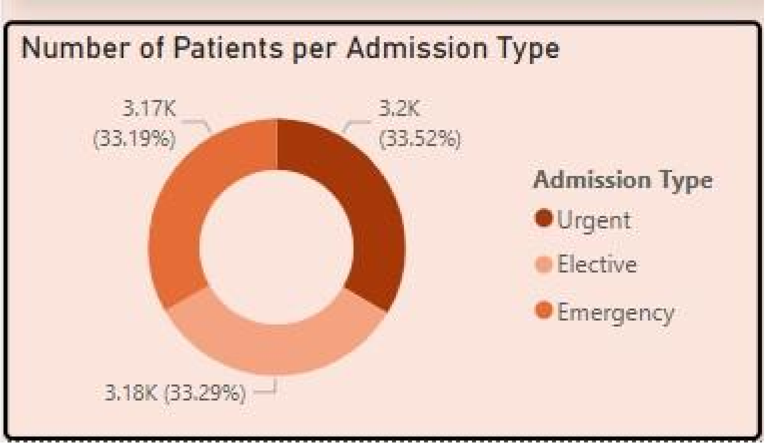
Page2: After identifying the highest billed age groups to be adults, middle-aged individuals, and seniors from the previous page, analyse potential factors driving high costs.



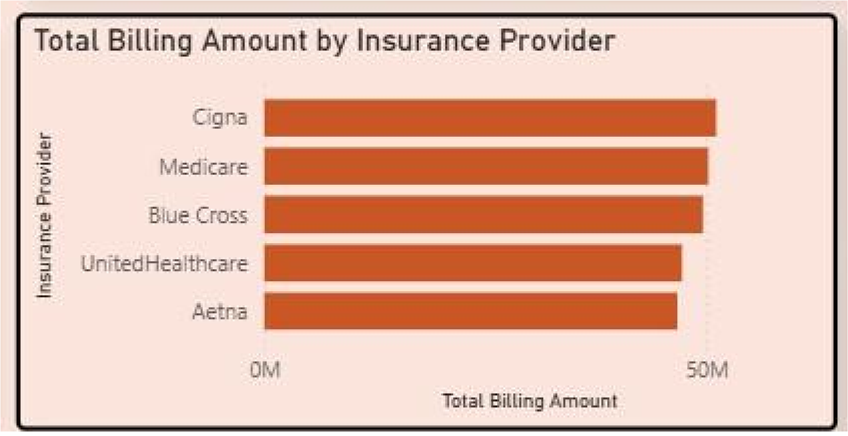
The three most common blood types in adults, middle-aged individuals, and seniors are AB-, O+, and A-. These can be further drilled into for analysis.



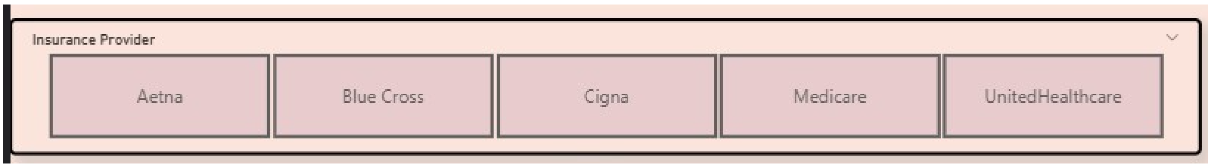
Between 2020-2023, adults have been associated with the highest total billing amount among the three age groups.



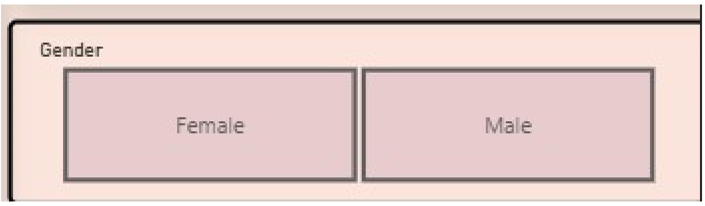
For the identified age groups, the distribution of admission types is relatively even. No single admission type is more common.



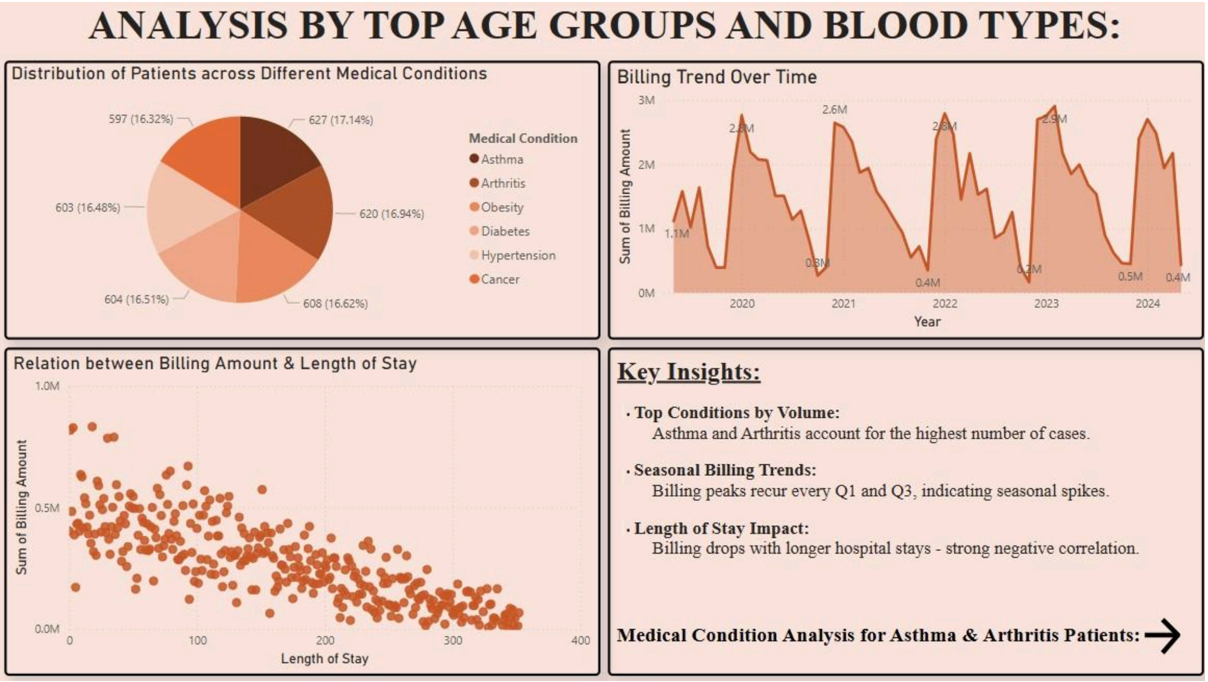
Among the three identified age groups, Cigna is the insurance provider with the highest total billing amount associated with it.



A slicer to allow analysis of the page for a particular insurance provider.



A slicer to allow filtering by gender. This may be used with the insurance provider slicer to analyse gender specific and provider specific trends.



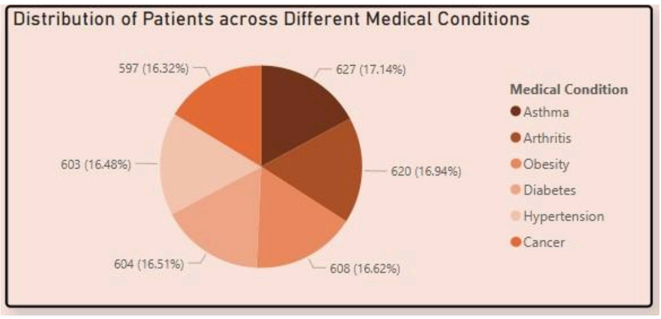
ANALYSIS BY TOP AGE GROUPS AND BLOOD TYPES:

Key Insights:

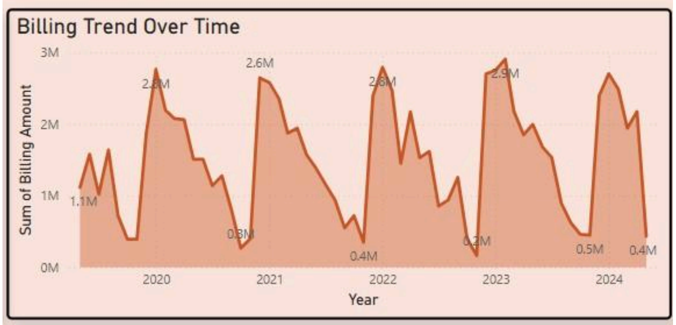
- **Top Conditions by Volume:**
  Asthma and Arthritis account for the highest number of cases.
- **Seasonal Billing Trends:**
  Billing peaks recur every Q1 and Q3, indicating seasonal spikes.
- **Length of Stay Impact:**
  Billing drops with longer hospital stays - strong negative correlation.

Medical Condition Analysis for Asthma & Arthritis Patients: →

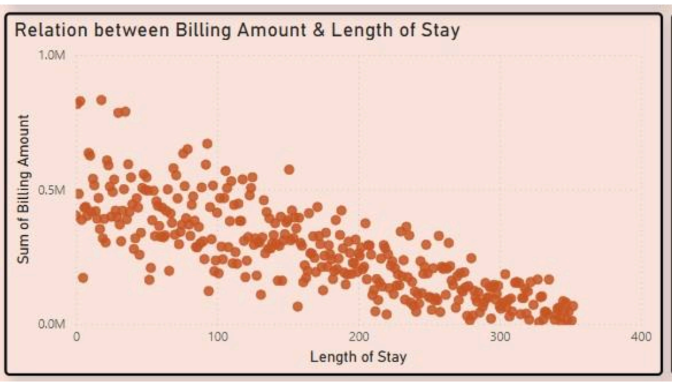
Page3: Shows overall analysis of patients and billing amount for top 3 age groups and blood types.



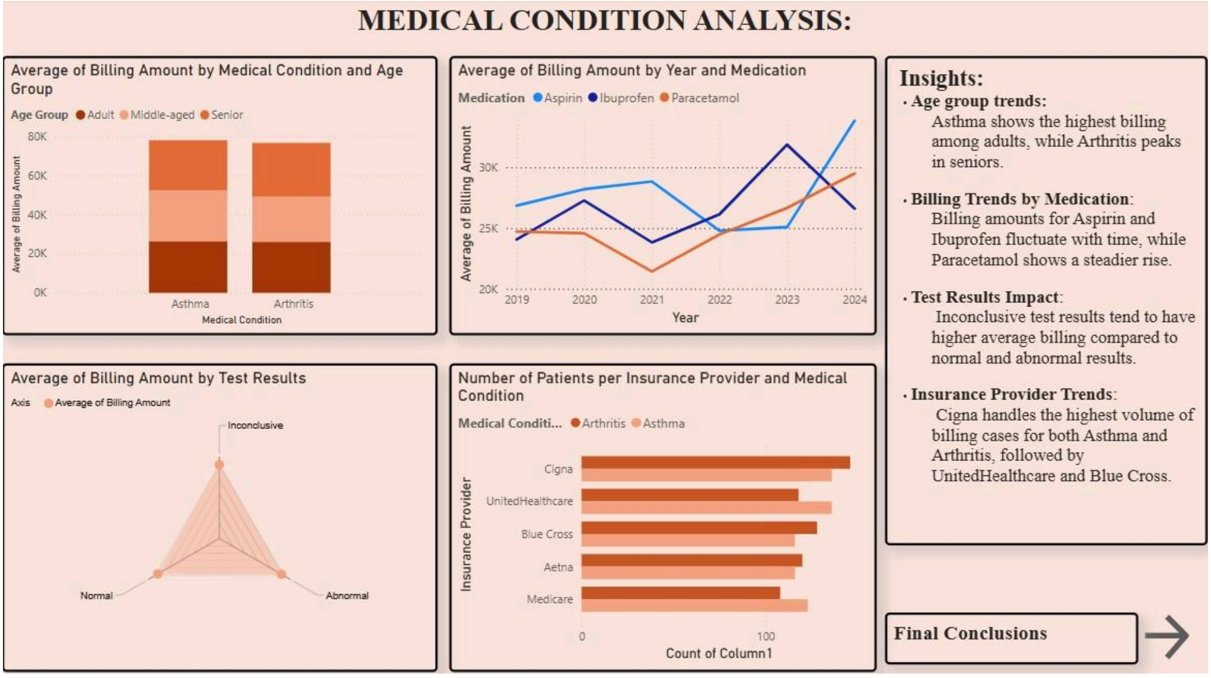
Pie Chart to show the number of patients with each medical condition for Age Groups Adult, Middle Age, and Senior as well as the most common blood groups (AB-, O+, A-).



Area Chart which shows the trend of billing amount over time. Can be drilled down to date.



The scatter chart shows us the correlation between length of stay and billing amount of patients.



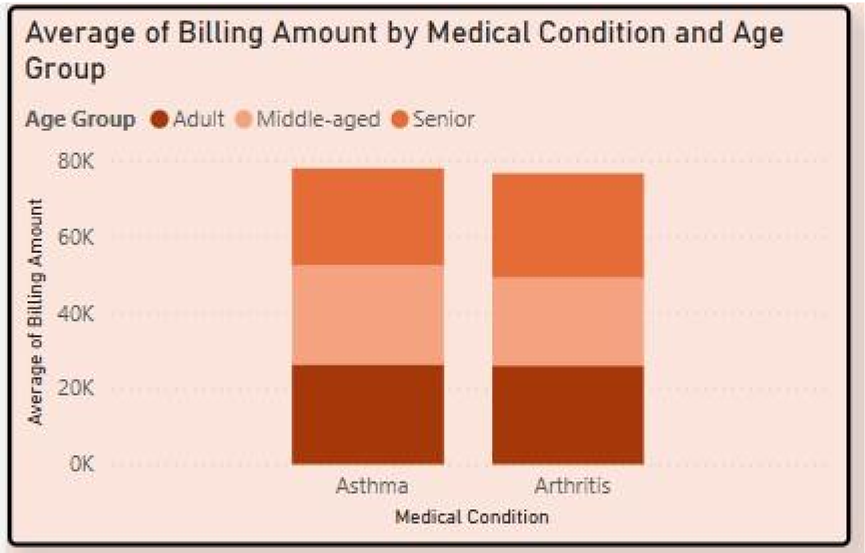
MEDICAL CONDITION ANALYSIS:

Average of Billing Amount by Medical Condition and Age Group
Age Group ● Adult ● Middle-aged ● Senior
80K
60K
40K
20K
0K
Average of Billing Amount
Asthma
Arthritis
Medical Condition
Average of Billing Amount by Year and Medication
Medication ● Aspirin ● Ibuprofen ● Paracetamol
30K
25K
20K
Average of Billing Amount
2019
2020
2021
2022
2023
2024
Year
Average of Billing Amount by Test Results
Axis ● Average of Billing Amount
Inconclusive
Normal
Abnormal
Number of Patients per Insurance Provider and Medical Condition
Medical Conditi... ● Arthritis ● Asthma
Cigna
UnitedHealthcare
Blue Cross
Aetna
Medicare
Insurance Provider
0
100
Count of Column1
Final Conclusions

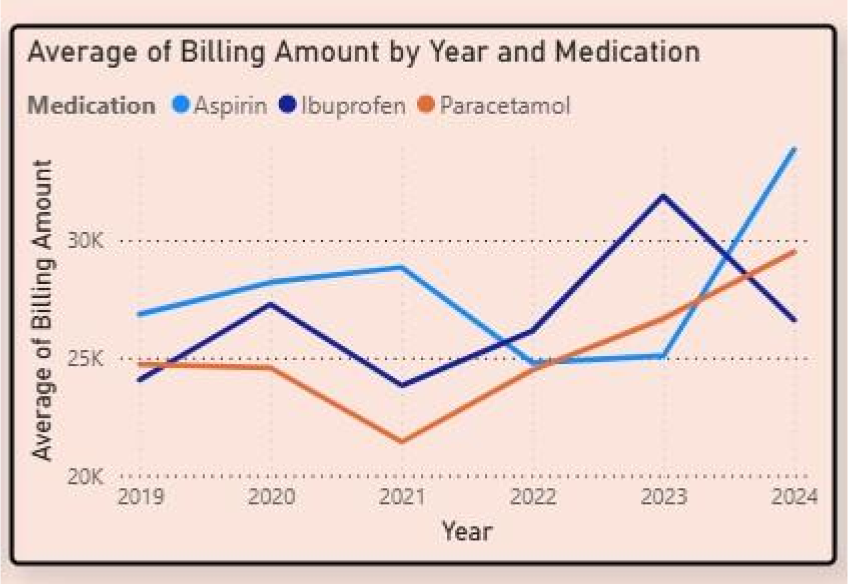
Insights:

- **Age group trends:** Asthma shows the highest billing among adults, while Arthritis peaks in seniors.
- **Billing Trends by Medication**: Billing amounts for Aspirin and Ibuprofen fluctuate with time, while Paracetamol shows a steadier rise.
- **Test Results Impact**: Inconclusive test results tend to have higher average billing compared to normal and abnormal results.
- **Insurance Provider Trends**: Cigna handles the highest volume of billing cases for both Asthma and Arthritis, followed by UnitedHealthcare and Blue Cross.

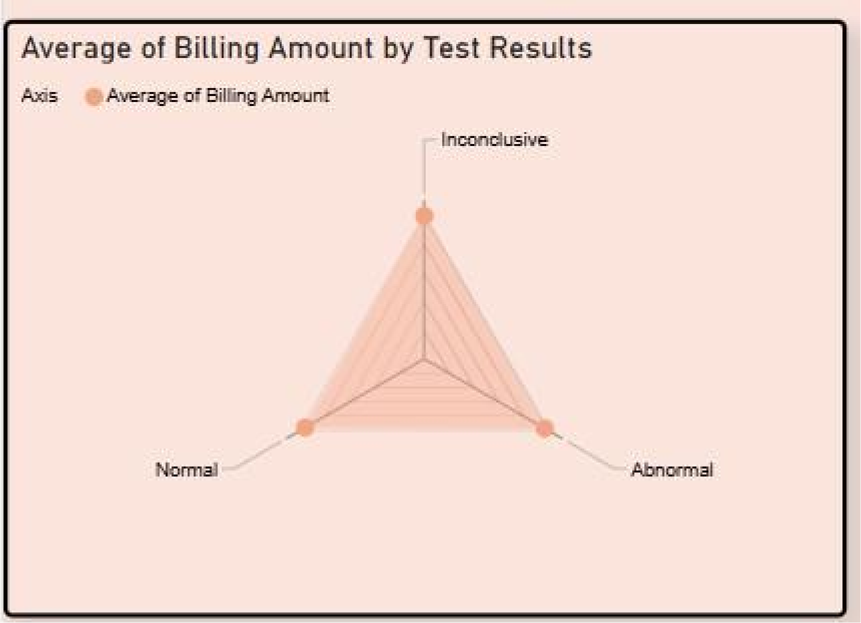
Page4: Shows medical Condition Analysis by examining how factors like medications, test results, age groups, and insurance providers influence billing patterns.



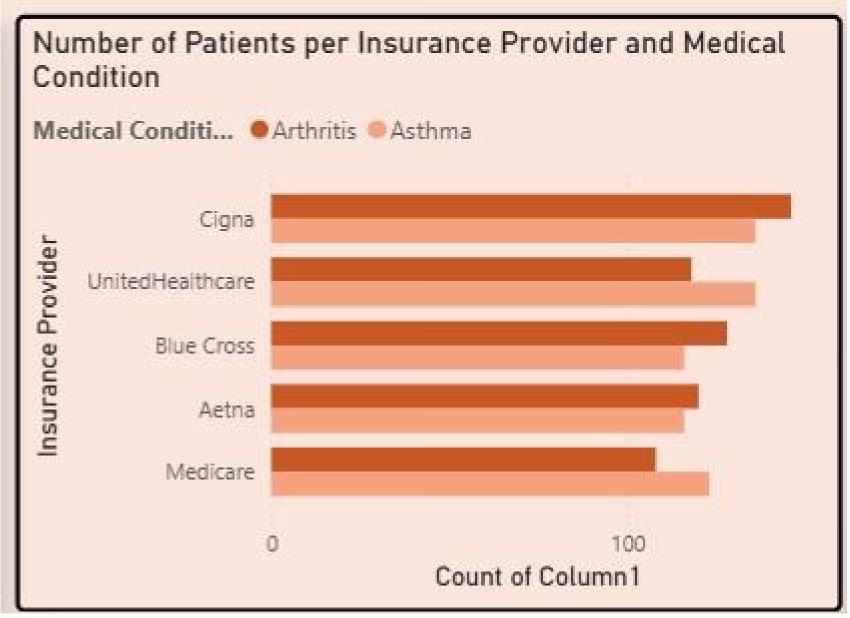
In the stacked column chart we can see that Asthma shows the highest billing among adults, while Arthritis peaks in seniors.



In this line chart we can see that Paracetamol shows a steady increase in billing over the years, while Aspirin and Ibuprofen fluctuate.

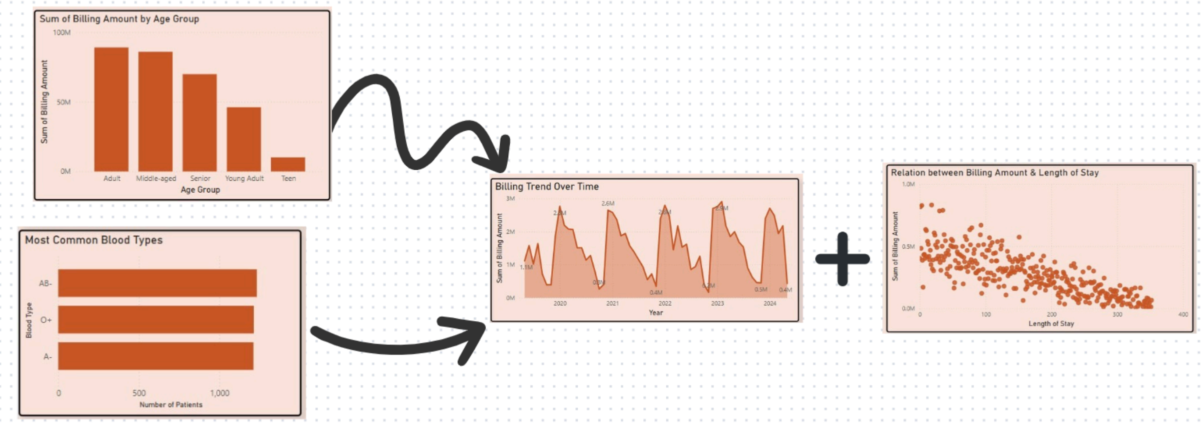


This Radar chart shows that inconclusive test results are associated with the highest average billing amounts.

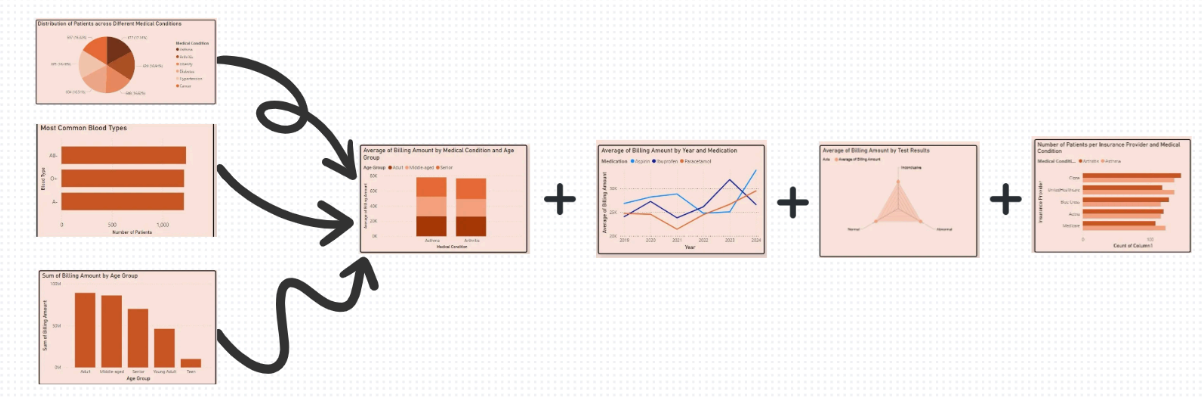


This clustered chart shows that Cigna leads in billing volume for both Asthma and Arthritis, followed by UnitedHealthcare and Blue Cross.

Datastory:



Highest total billing amount is contributed the most by the adult and middle-aged groups, while by minimal differences the blood types that contribute the most within these are AB- ,O+ and A-. These filters on further charts then show that billing is higher in the months of December, January and February, while total bill has a negative correlation with length of stays, mainly due to the reason for expensive surgeries, operations, and other procedures that are done in a few days, or extended to only a week if any complications arise.



Adults, middle-aged individuals, and seniors with the blood types AB-, O+, and A- are most commonly diagnosed with asthma and arthritis. Among such patients (people with the identified characteristics and diagnosed with asthma or arthritis) the billing amount distribution is similarly split between different age groups (for both asthma and arthritis). Aspirin has been the primary medication associated with the highest bills for such patients, though ibuprofen and paracetamol are also on the higher end. Test results for these patients are likely not a contributing factor, since each type of result is linked to a similar billing amount on average. The insurance providers UnitedHealthcare and Medicare deal more with asthma cases compared to arthritis cases. On the other hand, Cigna, Blue Cross, and Aetna deal with more cases of arthritis.

Recommendations:

Based on the insights found, some action steps to optimize operations and improve billing for patients are:

- Create programs tailored to adults and middle aged people, specially for arthritis or asthma patients.
- High cost procedures such as surgeries and operations should be investigated and steps should be taken to reduce costs (Note: Optimizing cost is extremely important for hospitals entirely managed by the government. This will help redirect funds to other important areas).
- Collaborate with insurance providers (UnitedHealthcare, Medicare for asthma, and Cigna, Aetna for arthritis) to streamline claims for high-bill demographics.

Work Contributions:

Uzair Nadeem:

Data Cleaning and EDA
Formatting and Designing of one page of the Power BI report
Conducting the Domain Expert Interview and Filling the Empathy Map

Zainab Hasan:

Design Sprint
Formatting and designing of one page in power bi (medical condition analysis)
Insights of charts in this page
Conclusion page
Video demo

Saad Thaplawala:

My part in the design sprint, created the title page and created the first page theme.
First page of report charts suggestions and format and overall power BI report theme suggestion.
Finalized the word document report and also added the datastory.

Asna Farooqui:

Formatting and designing of one report page.
Design Sprint
Business Data Knowledge Research for document
Report: Dataset Description, Data Sprint, Industry Background