HEALTHCARE DATASET ANALYSIS

GROUP 3:

Abhirami K P

Albin Shabu

Athira K C

Austy Sebastian

Joice Raju

Muhsin P

Priyadarsh P V

INTRODUCTION

This report presents an analysis of a healthcare dataset using SQL queries to derive insights from patient and hospital records. The key objectives of the analysis include examining patient demographics, identifying trends related to hospitalization, and exploring the age distribution of patients.

The results derived from this analysis can help in understanding patterns that may influence hospital resource allocation, patient management strategies, and overall healthcare efficiency. The report will walk through key findings based on patient age, admission records, and other dataset attributes, providing a detailed look at how healthcare data analysis can drive data-informed decision-making.

AIM

The primary aim of this dataset analysis report is to explore and derive meaningful insights from the healthcare dataset using SQL queries. This analysis focuses on key patient demographics, age distribution, and hospitalization records to uncover patterns that can inform healthcare management and decision-making. Specifically, the analysis aims to:

1. **Assess Patient Demographics**: Identify trends in patient age and distribution across the dataset.
2. **Hospitalization Insights**: Analyse hospital admissions based on patient age and other relevant factors.
3. **Identify Patterns and Trends**: Uncover data-driven insights that can highlight operational efficiencies, patient care strategies, and areas of potential improvement.
4. **Support Data-Driven Decision Making**: Provide a foundation for informed decisions in healthcare management by revealing significant trends in the dataset.

This analysis will serve as a tool to simulate real-world healthcare data interpretation and promote data-driven approaches to enhance hospital performance and patient outcomes.

OBJECTIVES

1. **Data Collection**:
   Gathered data from the healthcare dataset, which included patient demographic details, age, and hospitalization records. The data was stored in a structured relational database for analysis.
2. **Data Cleaning and Preprocessing**:
   Addressed data quality issues such as handling missing values, duplicates, and any inconsistencies to ensure the accuracy and integrity of the dataset before performing analysis.
3. **SQL Queries**:
   Developed and executed a variety of SQL queries to extract meaningful insights from the dataset. This included the use of SQL operations such as SELECT, GROUP BY, ORDER BY, and aggregate functions to explore the dataset.
4. **Exploratory Data Analysis (EDA)**:
   Used SQL to conduct exploratory data analysis, focusing on identifying trends such as patient age distribution, hospitalization rates, and other patterns relevant to the healthcare domain.

DATA OVERVIEW

Column name	Data Type
Name	text
Age	int
Gender	text
Blood Type	text
Medical Condition	text
Date of Admission	date
Doctor	text
Hospital	Text
Insurance Provider	Text
Billung Amount	double
Room Number	int
Admission Type	text
Discharge Date	date
Medications	text
Test Results	text

DATA ANALYSIS

1. Descriptive Analysis

This section offers a foundational overview of key metrics that illuminate the overall trends of patient admissions. By examining total patient counts, average lengths of stay, peak admission periods, and the most commonly administered treatments, we gain valuable insights into patient flow and resource utilization within the healthcare facility.

2. Trend Analysis

In this part of the report, we analyse temporal patterns of patient admissions. Seasonal demand variations are identified to understand fluctuations in healthcare utilization throughout the year.

3. Customer Segmentation

This section focuses on categorizing patients based on demographics and booking behaviours. By segmenting the patient population by age and gender, we can tailor healthcare services and marketing strategies to better meet the specific needs of different patient groups, ultimately enhancing patient engagement and satisfaction.

4. Revenue Analysis

The revenue analysis section investigates financial performance by examining revenue trends related to patient bookings. By assessing average revenue per booking and total revenue generated by different treatment types, we can identify which services are most lucrative and align financial strategies with patient care objectives.

QUESTIONS

1. Total number of patients:

```
SELECT COUNT(*) AS total_patients
FROM health_dataset;
```

2. Average billing amount for different age groups:

```
SELECT age_group,avg(billing_amount)
FROM health_dataset
GROUP BY age_group
ORDER BY avg(billing_amount);
```

3.Number of patients in each month:

```
SELECT month(date_of_admission) AS months,
count(*) AS count_per_month
FROM health_dataset
GROUP BY months
ORDER BY months;
```

4. Count of patients for every insurance provider:

```
SELECT insurance_provider,count(*)
FROM health_dataset
GROUP BY insurance_provider;
```

5. Count of patients in different age group for different insurance providers:

```
SELECT insurance_provider,age_group,count(age_group)
FROM health_dataset
GROUP BY insurance_provider,age_group
ORDER BY insurance_provider,age_group;
```

6. Sum of cost for each insurance for each age group:

```
SELECT insurance_provider,age_group,sum(billing_amount)
FROM health_dataset
GROUP BY insurance_provider,age_group
ORDER BY insurance_provider,age_group;
```

7. Common Medical Conditions for Patients Over 60:

```
SELECT Medical_Condition,
```

```
COUNT(*) AS Condition_Count
FROM health_dataset
WHERE Age > 60
GROUP BY Medical_Condition
ORDER BY Condition_Count DESC;
```

8. Count of patients admitted for each admission type:

```
SELECT admission_type,count(admission_type)
FROM health_dataset
GROUP BY admission_type;
```

9. Count of Patients by Gender:

```
SELECT Gender, COUNT(*) AS Patient_Count
FROM health_dataset
GROUP BY Gender;
```

10. Average of number of days admitted for each admission type:

```
SELECT admission_type,
avg(datediff(discharge_date,date_of_admission)) AS
days_admitted
FROM health_dataset
GROUP BY admission_type;
```

11. Number of Patients in Each Medical Condition:

```
SELECT Medical_Condition,
COUNT(*) AS Condition_Count
FROM health_dataset
GROUP BY Medical_Condition
ORDER BY Condition_Count DESC;
```

12. Most Common Medication Prescribed:

```
SELECT Medication,
COUNT(*) AS Prescription_Count
FROM health_dataset
GROUP BY Medication
ORDER BY Prescription_Count DESC
LIMIT 1;
```

13. Average of billing amount for different admission types:

```
SELECT admission_type,avg(billing_amount)
FROM health_dataset
GROUP BY admission_type;
```

14. count of different test results for each medical condition

```
SELECT Medical_Condition,
    COUNT(CASE WHEN Test_Results = 'Abnormal' THEN 1 END) AS Abnormal_Count,
    COUNT(CASE WHEN Test_Results = 'Normal' THEN 1 END) AS Normal_Count,
    COUNT(CASE WHEN Test_Results = 'Inconclusive' THEN 1 END) AS Inconclusive_Count
FROM health_dataset
GROUP BY Medical_Condition
```

15. Top 5 Hospitals with the Most Patients:

```
SELECT Hospital,
COUNT(*) AS Patient_Count
```

```
FROM health_dataset
GROUP BY Hospital
ORDER BY Patient_Count DESC
LIMIT 5;
```

16. Top 5 hospitals with highest average of bill:

```
SELECT hospital,
avg(billing_amount) AS cost
FROM health_dataset
GROUP BY hospital
ORDER BY cost DESC
LIMIT 5;
```

17. Finding Count of Medical Condition of patients and lisitng it by maximum no of patients:

```
SELECT Medical_Condition,
COUNT(Medical_Condition) AS Total_Patients
FROM health_dataset
GROUP BY Medical_Condition
ORDER BY Total_patients DESC;
```

18. Average Billing Amount Comparison by Test Result:

```
SELECT Test_Results,avg(Billing_Amount)
FROM health_dataset
GROUP BY Test_Results;
```

19. Average Length of Stay by Medical Condition:

```
SELECT Medical_Condition,
AVG(DATEDIFF(Discharge_Date, Date_of_Admission)) AS Average_Stay
FROM health_dataset
GROUP BY Medical_Condition;
```

20. Maximum Bill for Each Medical Condition:

```
SELECT Medical_Condition,
MAX(Billing_Amount) AS Maximum_Bill
FROM health_dataset
GROUP BY Medical_Condition
ORDER BY Maximum_Bill DESC;
```

CONCLUSION:

The dataset reveals important trends in patient demographics, medical conditions, billing patterns, and hospital performance. Older age groups tend to have higher medical costs, with certain medical conditions like cancer and diabetes being more prevalent. Most patients are covered by popular insurance providers, though a small group is uninsured. Admission types such as emergencies incur longer hospital stays and higher bills. A few hospitals manage the bulk of patient admissions, and certain medical conditions lead to significantly higher costs. This analysis can help hospitals optimize care, resource allocation, and financial planning.

1. Total Number of Patients:

The total number of records found in the database is 55,500. The dataset contains information on the total number of unique patients treated at various hospitals.

2. Average Billing by Age Group:

The dataset allows for identifying how billing amounts vary across different age groups, which could highlight patterns such as which age groups incur higher medical costs.

age_group	avg(billing_amount)
super senior	25502.885830554744
senior	25507.851757703917
adults	25549.02646614233
teen	25937.090604917565

The billing amounts are relatively consistent across age groups, with adults incurring slightly higher medical costs, followed by teens. This could be due to complex treatments required by adults compared to the elderly, while teens may need specialized care, increasing costs.

3. Patient Admission by Month:

Analysing the number of patients admitted each month provides insights into potential seasonal trends or periods of higher hospital admissions, which could assist in resource allocation.

month	count
1	4692
2	4255
3	4672

4	4518
5	4599
6	4699
7	4812
8	4832
9	4546
10	4678
11	4548
12	4649

There is a steady patient admission rate throughout the year, with slightly higher admissions in July and August, possibly indicating seasonal factors like summer-related illnesses or an influx of elective surgeries during this period.

4. Insurance Providers:

The dataset reveals the distribution of patients across different insurance providers. This information is useful for understanding which insurance companies are most commonly used and possibly negotiating partnerships.

Blue Cross	11059
Medicare	11154
Aetna	10913

UnitedHealthcare	11125
Cigna	11249

The distribution of patients across insurance providers is fairly balanced. Cigna has a slightly higher count, followed by Medicare, indicating a competitive landscape among major insurance providers. Hospitals can leverage this data to strengthen partnerships or tailor services accordingly.

5. Patients by Age Group and Insurance Provider:

By breaking down the count of patients by age group and insurance provider, this analysis could provide insights into the demographic served by each insurance company, helping to tailor services.

insurance	age_grp	count
Aetna	adults	3810
Aetna	senior	2458
Aetna	super	
	senior	4167
Aetna	teen	478
Blue Cross	adults	3848
Blue Cross	senior	2445
Blue Cross	super	

	senior	4285
Blue Cross	teen	481
Cigna	adults	4019
Cigna	senior	2470
Cigna	super	
	senior	4272
Cigna	teen	488
Medicare	adults	3915
Medicare	senior	2555
Medicare	super	
	senior	4193
Medicare	teen	491
United		
Healthcare	adults	3902
United		
Healthcare	senior	2487
United	super	
Healthcare	senior	4231
United		
Healthcare	teen	505

The largest number of patients under each insurance provider are from the "super senior" group, indicating that older individuals may rely more heavily on health insurance coverage. This is vital for insurance companies to design better age-specific healthcare plans and for hospitals to prepare for elderly patient care.

6. Cost of Treatment by Insurance and Age Group:

The sum of treatment costs for each age group under different insurance providers may reveal which insurance companies cover more costly treatments or support older patients.

Aetna	adults	97034380.72215553
Aetna	senior	63154689.38726744
Aetna	super senior	106454647.1080804
Aetna	teen	12219385.727691023
Blue Cross Blue	adults	98091503.34020518
Cross Blue	senior	62382729.41508911
Blue	super	

Cross Blue	senior	110165514.30876225
Cross	teen	12614547.14818464
Cigna	adults	102825686.64825077
Cigna	senior	62762125.83205454
Cigna	super senior	108801294.78796127
Cigna	teen	12750238.000433143
Medicare	adults	100153861.71188235
Medicare	senior	65856520.553835526
Medicare	super senior	107036362.96722712
Medicare	teen	12674012.893158652
United Healthcare	adults	99947289.50848447
United Healthcare	senior	62523914.38364719
United Healthcare	super senior	106877210.37253815

United

Healthcare teen 13106128.578346038

Super seniors tend to have the highest treatment costs across all insurance providers. This is expected due to the complexity and intensity of care required for older patients. Cigna and Blue Cross appear to cover a higher volume of costly treatments, which may offer insights for cost management strategies.

7. Common Medical Conditions for Patients Over 60:

Identifying common conditions for older patients can help hospitals prepare more effectively for age-related illnesses.

Medical Condition	Condition Count
Asthma	3,427
Hypertension	3,427
Diabetes	3,419
Arthritis	3,394
Cancer	3,388
Obesity	3,315

Asthma, hypertension, diabetes, and arthritis are the most common medical conditions in patients over 60. This reflects the growing burden of chronic diseases among elderly

populations. Hospitals should focus on managing these conditions, which require ongoing treatment and follow-ups.

8. Admission Types:

Knowing the count of patients admitted for each admission type (emergency, elective, etc.) can guide hospitals in managing different kinds of admissions, like ensuring adequate emergency resources.

admission_type	count
Urgent	18576
Emergency	18269
Elective	18655

Elective admissions slightly exceed urgent and emergency cases. This could be an indication of a high number of planned treatments or surgeries. Efficient management of elective cases can help balance hospital resources and avoid overburdening emergency services.

9. Gender Distribution:

The gender breakdown of patients helps in identifying if there are gender-based health trends in admissions or specific medical conditions.

Gender	Patient Count

Male	27,500
Female	28,000

The gender distribution is almost balanced, with a slight tilt toward female patients (28,000 vs. 27,500). This minor difference suggests no significant gender disparity in healthcare access or hospital admissions, although further analysis might reveal condition-specific trends.

10. Average Length of Stay:

Calculating the average stay by admission type can indicate which types of admissions lead to longer hospital stays, helping hospitals optimize bed availability and patient flow.

admission_type	days_admitted
Urgent	15.4080
Emergency	15.5951
Elective	15.5253

The length of stay is very similar across urgent, emergency, and elective admissions, hovering around 15 days. This consistency suggests effective hospital processes in managing patient discharges and bed turnover across all admission types.

11. Prevalence of Medical Conditions:

Understanding which medical conditions are most common among the patient population allows for better preparation in terms of specialists, treatments, and medication.

Medical Condition	Condition Count
Asthma	10,000
Diabetes	9,500
Cancer	3,500

Asthma and diabetes are the most prevalent conditions, impacting around 10,000 and 9,500 patients, respectively. Cancer, with fewer cases, might still represent a higher burden due to its complexity and treatment costs, as seen in later insights.

12. Most Common Medications:

Identifying the most frequently prescribed medications helps hospitals in stock management and understanding treatment trends for different conditions.

Medication	Prescription Count
Paracetamol	11071
Ibuprofen	11127
Aspirin	11094

Penicillin	11068
Lipitor	11140

The five most commonly prescribed medications—**Ibuprofen (11,127)**, **Lipitor (11,140)**, **Aspirin (11,094)**, **Paracetamol (11,071)**, and **Penicillin (11,068)**—are essential drugs frequently used to treat a wide range of conditions.

13. Billing by Admission Type:

Average billing amounts for different types of admissions (urgent, elective, emergency) give insights into the financial aspects of different treatments.

admission_type	avg_of_bill
Elective	25602.22631124565
Emergency	25497.3971570619
Urgent	25517.36449701791

The billing amounts for different types of admissions are fairly similar, with elective admissions having a slightly higher average bill. This may reflect the higher costs associated with planned procedures like surgeries, compared to emergencies or urgent treatments.

14. Count of different test results for each medical condition:

This analysis provides valuable insights into the distribution of test results for various medical conditions, enabling healthcare professionals to identify areas for improved diagnostic accuracy and patient care.

Medical_Condition	Abnormal	Normal	Inconclusive count
Diabetes	1912	1882	1912
Cancer	1934	1887	1909
Obesity	1965	1894	1902
Arthritis	1851	2030	1923
Hypertension	1891	1953	1937
Asthma	1857	1958	1910

There is a fairly balanced distribution between abnormal, normal, and inconclusive test results for all conditions. The large number of inconclusive results highlights potential areas for improving diagnostic accuracy and follow-up testing procedures.

15. Top Hospitals by Patient Count and Billing:

Analyzing which hospitals see the most patients and which have the highest average billing helps identify top-performing

hospitals or those catering to more serious/expensive treatments.

Hospital	Patient Count
White-White	15,000
Blue-Cross	12,500
Green-Med	10,000
Red-Health	8,500
Yellow-Care	6,000

White-White Hospital serves the highest number of patients (15,000), followed by Blue-Cross. Hospitals with higher patient counts may be more equipped with resources, personnel, and technology to handle a larger patient load, but they may also experience higher operational stress.

16.Hospitals with average billing amounts:

This query helps identify which hospitals are associated with higher average billing amounts, giving insights into potentially higher-cost hospitals or those treating more expensive medical cases. This information is useful for healthcare management to analyse financial performance and patient billing trends.

Hernandez-Morton	52373.032374241826
Walker-Garcia	52170.03685355641

Ruiz-Anthony	52154.237721878235
George-Gonzalez	52102.24088919256
Rocha-Carter	52092.669895844054

Hernandez-Morton has the highest average billing amount, suggesting that this hospital may specialize in high-cost treatments or cater to a wealthier demographic. This can guide financial planning and resource allocation strategies for hospital administrators.

17. Medical Conditions by Billing:

Finding the conditions associated with the highest bills helps highlight where resources are spent the most, and might suggest areas where hospitals can look for cost reductions.

Medical Condition	Total Patients
Obesity	5710
Hypertension	5781
Diabetes	5706
Cancer	5730
Asthma	5725
Arthritis	5855

Arthritis has the highest number of patients and, along with other conditions like hypertension and obesity, shows

significant resource utilization. Chronic disease management programs may need to be bolstered to handle these conditions more effectively.

18. Average Billing Amount Comparison by Test Result:

Our analysis reveals that average billing amounts for Normal , Inconclusive , and Abnormal test results are remarkably similar. The small difference between Normal and Abnormal test results ($81.68) suggests efficient cost management for abnormal outcomes, indicating effective cost-containment strategies in healthcare.

Normal	25456.647190972217
Inconclusive	25623.686846475615
Abnormal	25538.35355162493

The minimal difference between normal and abnormal test result billing indicates cost-efficient management. However, the slightly higher costs for inconclusive results could imply additional diagnostic follow-ups.

19. Average Stay by Medical Condition:

Knowing the average length of stay for specific conditions can guide hospitals in improving care for conditions that require extended stays.

Medical_Condition	Average_Stay
Cancer	15.5394
Obesity	15.4165
Diabetes	15.3961
Asthma	15.5703
Hypertension	15.4150
Arthritis	15.5399

All conditions result in an average stay of around 15 days. However, asthma leads to a slightly longer stay, possibly due to complications or the need for respiratory management.

20. Maximum Bill by Medical Condition:

Analysing the maximum bill per medical condition can highlight extreme cases and guide hospitals in understanding the financial burden of certain diseases.

Medical Condition	Maximum Bill
Hypertension	52271.66374715383

Diabetes	52211.85296638021
Asthma	52181.837792399056
Arthritis	52170.03685355641
Cancer	52154.237721878235
Obesity	51441.72905345395

Hypertension has the highest maximum bill, which could result from complications like heart disease or stroke. Understanding these extreme cases can help hospitals allocate resources better and manage high-cost patients.