Graduation Project: Healthcare data analysis

BY:

Nour Alaa

Nesma Emad

Wesam Abo Bakr

Mahmoud Ramadan

Mohammed Khaled

Ahmed Rabea

DEPI round 3

Data Analysis Track

Abstract

This project analyzes a structured dataset of patient admission records, containing demographic, clinical, administrative, and financial information. The dataset includes columns such as ID, Name, Age, Gender, Blood Type, Medical Condition, Date of Admission, Doctor, Hospital, Insurance Provider, Billing Amount, Room Number, Admission Type, Discharge Date, Medication, and Test Results.

The project documents each step of the data analysis process, beginning with data cleaning and preparation, followed by exploratory analysis, visualization, and interpretation of results. Key objectives include identifying patient demographics and health trends, evaluating hospital resource utilization, assessing financial and insurance coverage, and examining treatment outcomes.

By systematically transforming raw healthcare data into meaningful insights, this project demonstrates how structured records can support decision-making in healthcare management, public health studies, and operational planning.

Introduction

Healthcare systems worldwide face growing challenges due to increasing patient volumes, the prevalence of chronic diseases, and rising treatment costs. To address these pressures, hospitals and public health organizations are adopting data-driven strategies to improve efficiency, allocate resources effectively, and enhance patient outcomes. Patient admission records—containing demographic, clinical, administrative, and financial details—offer a rich source of information for understanding health trends and operational performance.

This project leverages a structured dataset of patient admissions to uncover actionable insights that support evidence-based decision-making in healthcare management and public health planning. The dataset includes key attributes such as patient demographics, medical conditions, admission and discharge details, billing amounts, insurance coverage, and treatment outcomes. By applying systematic data analysis techniques, including data cleaning, exploratory analysis, statistical evaluation, and visualization, this study demonstrates how raw healthcare data can be transformed into meaningful intelligence for improving hospital workflows, resource utilization, and patient care.

This project is built around a structured dataset that represents patient admission records in a healthcare setting. The dataset is organized into the following columns:

- ID – unique identifier for each patient
- Name – patient's full name
- Age – patient's age at admission
- Gender – male or female classification
- Blood Type – recorded blood group
- Medical Condition – primary diagnosis or health issue
- Date of Admission – when the patient was admitted
- Doctor – attending physician responsible for care

- Hospital – healthcare facility of admission
- Insurance Provider – coverage organization for billing
- Billing Amount – financial charges associated with treatment
- Room Number – assigned inpatient room
- Admission Type – emergency, elective, or other classification
- Discharge Date – when the patient left the hospital
- Medication – prescribed drugs during admission
- Test Results – outcomes of diagnostic investigations

Together, these columns provide a comprehensive view of patient care, combining demographic, medical, administrative, and financial information.

Project objective

The goal of this project is not only to describe the dataset but also to document every step of the data analysis process. This includes:

- Cleaning and preparing the data for analysis
- Exploring patterns and relationships across patient demographics, medical conditions, and hospital operations
- Performing statistical and visual analyses to uncover insights
- Evaluating financial and insurance trends
- Summarizing treatment outcomes through medication and test results

Why This Matters

By walking through each stage of the project, readers will gain a clear understanding of how raw healthcare data can be transformed into meaningful insights. This introduction sets the foundation for the detailed steps that follow, ensuring that anyone new to the project can easily grasp what the dataset contains and how it will be used.

Tools and Technologies Used

To ensure a complete and reliable analysis of the patient admission dataset, several tools were employed. Each tool served a distinct purpose in the workflow:

I. **Python**

1. Check the nulls and duplicates
2. Capitalize each word
3. Remove spaces
4. Replace values
5. Filling nulls

II. SQL (Structured Query Language)

1. what is the most common disease?
2. What is the impact of disease across different age groups?
3. How are different diseases distributed between male and female patients?
4. How do test results vary by disease or medication
5. Which hospitals have the highest patients intake
6. What is the average billing amount for patients at each hospital?
7. How does a doctor's performance vary based on patients volume or treatment outcomes?
8. Which medical condition does each doctor treat most frequently?
9. What is the common admission type per hospital?
10. What is the average treatment codt per disease?
11. How do insurance providers compare in terms of the number of patients they cover and total treatment cost?

12. What are total hospital billing amounts per month and year
13. Year over year change in case volume
14. Identify seasonal admission patterns

III. Power BI

1. Utilized for interactive dashboards and visualizations.
2. Allowed stakeholders to explore patient demographics, hospital resource utilization, and financial trends through dynamic charts and reports.

IV. Documentation Tools (Microsoft Word)

1. Employed to record each step of the project, ensuring transparency and reproducibility.
2. Provided a clear narrative of the workflow, methods, and findings for readers and stakeholders.

Methodology

1. Python

Using the Pandas library the following was done

1.1. Data exploration

Same procedures for all the files

1.1.1. Importing the file using

```
dh=pd.read_excel('/content/Doctor_Healthecare.xlsx')
```

1.1.2. Checking first five rows

```
dh.head()
```

1.1.3. Checking last 5 rows

dh.tail()

1.1.4. # to check columns info and data type

dh.info()

```
# to check columns info and data type
ph.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54944 entries, 0 to 54943
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   P_ID        54944 non-null  int64
 1   Name        54944 non-null  object
 2   Age         54940 non-null  float64
 3   Gender      54944 non-null  object
 4   Blood Type  54944 non-null  object
dtypes: float64(1), int64(1), object(3)
memory usage: 2.1+ MB
```

1.1.5. # to check the nulls

dh.isnull().sum()

1.1.6. # to check the duplicates

dh.duplicated().sum()

```
# to check the duplicates
ph.duplicated().sum()
```

```
np.int64(0)
```

1.2. Cleaning

1.2.1. # to capitalize first letter of each word

dh.Doctor=dh.Doctor.str.title()

1.2.2. # to remove extra space from beginning and end

dh['Doctor']=dh['Doctor'].str.strip()

1.2.3. # to remove extra spaces from the beginning and end of hospital names

hh['Hospital']=hh['Hospital'].str.strip(',').str.strip()

1.2.4. # to replace '-' with ',' in the hospital column

hh['Hospital']=hh['Hospital'].str.replace('-',',')

1.2.5. # to fill missing ages with the avg age of each gender

ph['Age']=ph['Age'].fillna(ph.groupby('Gender')['Age'].transform('mean'))

1.2.6. # to change age column data type

ph.Age=ph.Age.astype(int)

2. SQL

2.1. Analysis and Questions answers

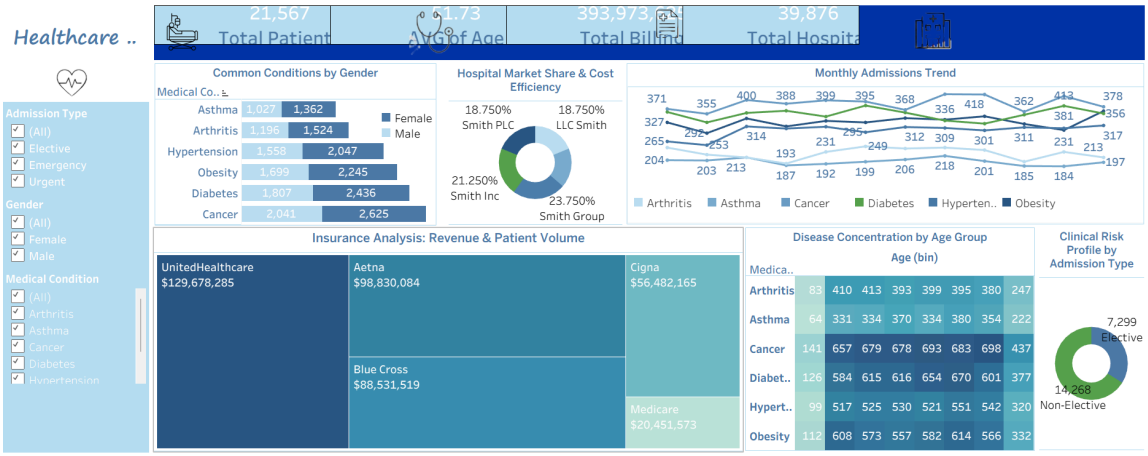
SQL queries were used to determine and answer all questions raised, which was 14 questions already put to gain insights from the data.

```sql
-----------------------------------------------
-- Question 1: WHAT IS THE MOST COMMON DISEASES?
-----------------------------------------------

 -- PROBLEM STATEMENT:
    --  Identify the most common diseases among patients to prioritize
    --  healthcare interventions and understand disease burden.

 -- GOAL:
    -- To determine which diseases affect the most patients,
    -- helping to guide decision-making for healthcare planning, prevention strategies, and resource

-- Frequency count of each disease
SELECT
    Medical_Condition,
    COUNT(*) AS Total_Common_Diseases
FROM PatientsData_healthecare_clean
GROUP BY Medical_Condition
ORDER BY Total_Common_Diseases DESC;
```

```
-------------------------------------------------------------------------------
-- Question 3: HOW ARE DIFFERENT DISEASES DISTRIBUTED BETWEEN MALE & FEMALE PATIENTS?
-------------------------------------------------------------------------------

 -- PROBLEM STATEMENT:
    -- Assess whether certain diseases disproportionately affect males or females.

 -- GOAL:
    -- Disease patterns by gender to identify whether certain conditions affect males or females more


SELECT
    PD.Medical_Condition,
    P.Gender,
    COUNT(*) AS Patient_Count
FROM Patients_healthecare_clean P
JOIN PatientsData_healthecare_clean PD
    ON P.P_ID = PD.P_ID
GROUP BY PD.Medical_Condition, P.Gender
ORDER BY PD.Medical_Condition, Patient_Count DESC;


-------------------------------------------------------------------------
-- Question 17: WHAT ARE THE SEASONAL ADMISSION TRENDS IN THE HOSPITAL?
-------------------------------------------------------------------------

  -- Problem Statement:
     -- Determine how patient admissions vary across different seasons
     -- (Winter, Spring, Summer, Fall) to identify seasonal patterns in hospital usage.

  -- Goal:
     -- To analyze seasonal admission trends, which can help in resource
     -- allocation, staff planning, and preparing for periods of high or low patient volume.

ALTER TABLE PatientsData_healthecare_clean
ADD Season VARCHAR(10);

UPDATE PatientsData_healthecare_clean
SET Season = CASE
                WHEN MONTH(Date_of_Admission) IN (12, 1, 2) THEN 'Winter'
                WHEN MONTH(Date_of_Admission) IN (3, 4, 5) THEN 'Spring'
                WHEN MONTH(Date_of_Admission) IN (6, 7, 8) THEN 'Summer'
                WHEN MONTH(Date_of_Admission) IN (9, 10, 11) THEN 'Fall'
             END;
--------------------
```

3. Tableau

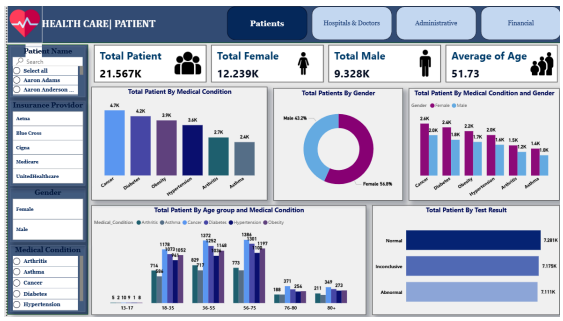
Tableau was used to show and visualize the overview

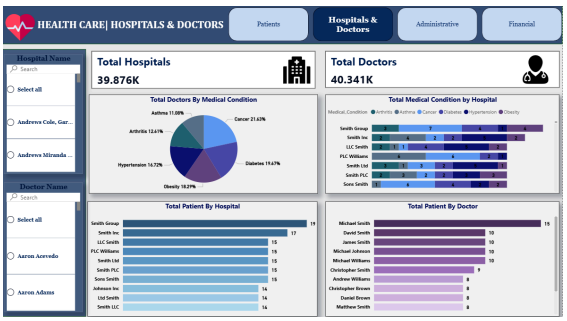


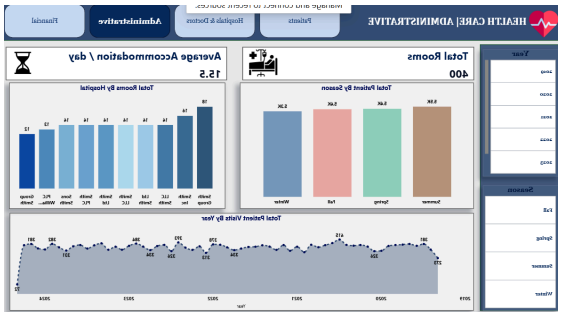
4. Power BI

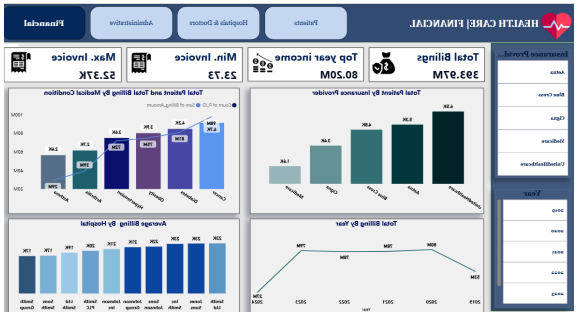
Power bi was our main visualization tool,

Used to visualize the data through 4 detailed dashboards with filters for more digging into the data.









A column named ‘Length of Stay’ has been added to determine the length of stay for each patient at hospital.

using this DAX measure: Length of stay = DATEDIFF(('Patient Data'[Date_of_Admission]),('Patient Data'[Discharge_Date]),DAY)

Insights

Behind every row of data lies a patient's journey through the healthcare system. Our analysis uncovers the stories hidden in the numbers — from the most common medical conditions to the financial burdens patients face.

1. Most common diseases

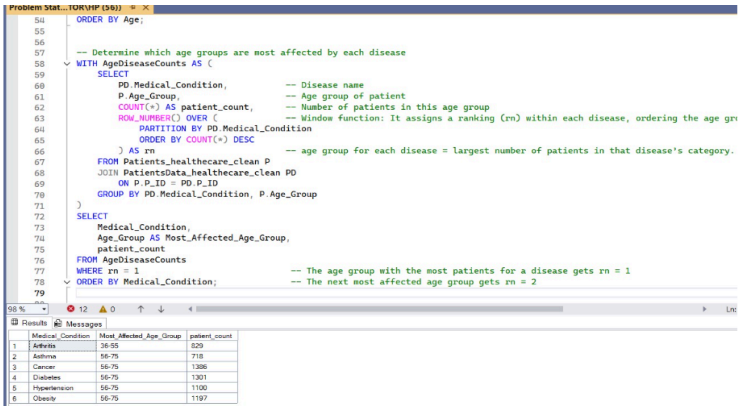
   Cancer accounted for the major health problem with 21.63% of all cases

   While diabetes occupied second place with the percentage of 19.67

   Followed by obesity, hypertension, arthritis and asthma with the percentage of 18.29%, 16.72%, 12.61% and 11.08% respectively.

2. Age groups most affected

   Both age groups (36-55) and (56-75) were found to be the most affected age groups affected by each disease.

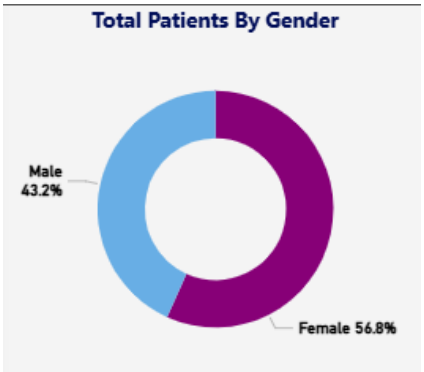


```
ORDER BY Age;

-- Determine which age groups are most affected by each disease
WITH AgeDiseaseCounts AS (
    SELECT
        PD.Medical_Condition,           -- Disease name
        P.Age_Group,                    -- Age group of patient
        COUNT(*) AS patient_count,      -- Number of patients in this age group
        ROW_NUMBER() OVER (             -- Window function: It assigns a ranking (rn) within each disease, ordering the age gr
            PARTITION BY PD.Medical_Condition
            ORDER BY COUNT(*) DESC
        ) AS rn                         -- age group for each disease = largest number of patients in that disease's category.
    FROM Patients_healthecare_clean P
    JOIN PatientsData_healthecare_clean PD
      ON P.P_ID = PD.P_ID
    GROUP BY PD.Medical_Condition, P.Age_Group
)
SELECT
    Medical_Condition,
    Age_Group AS Most_Affected_Age_Group,
    patient_count
FROM AgeDiseaseCounts
WHERE rn = 1                            -- The age group with the most patients for a disease gets rn = 1
ORDER BY Medical_Condition;             -- The next most affected age group gets rn = 2
```

	Medical_Condition	Most_Affected_Age_Group	patient_count
1	Arthritis	36-55	829
2	Asthma	56-75	718
3	Cancer	56-75	1386
4	Diabetes	56-75	1301
5	Hypertension	56-75	1100
6	Obesity	56-75	1197

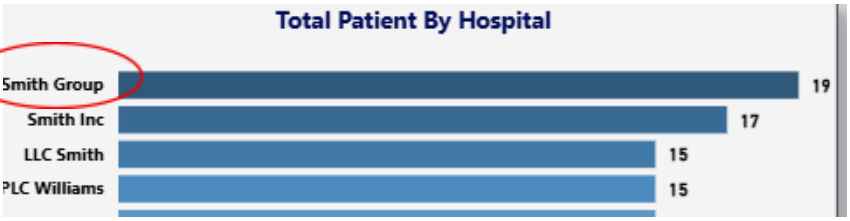
3. Gender distribution per disease

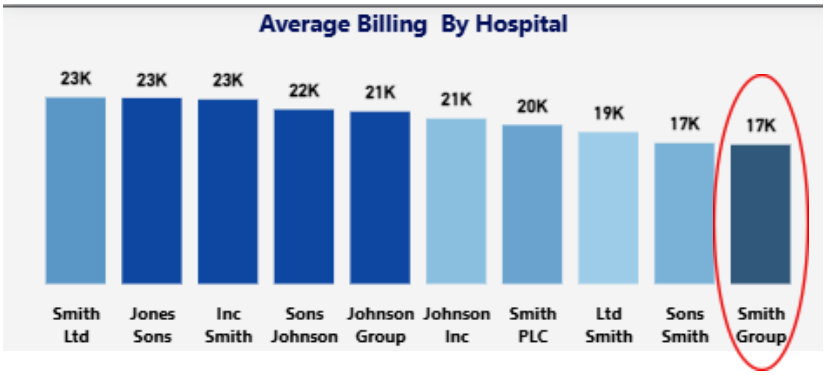
Female gender accounted for 56.75% of all cases while the male gender accounted for 43.25%



4. Hospital with highest patient intake

Smith group hospital was found to have the highest patient intake which was interpreted through the financial dashboard to be due to the implemented pricing policy

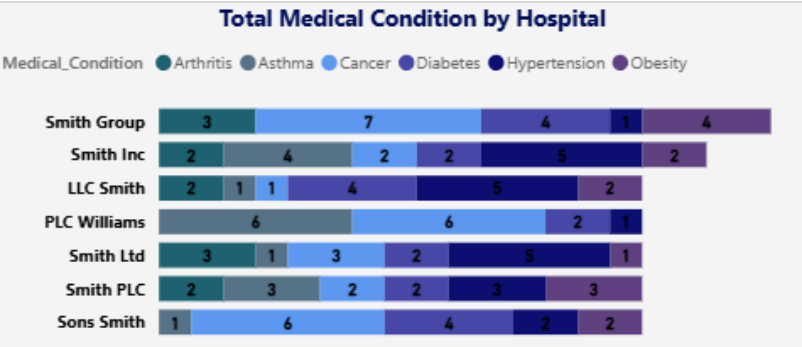




5. Doctor performance based on patient volume

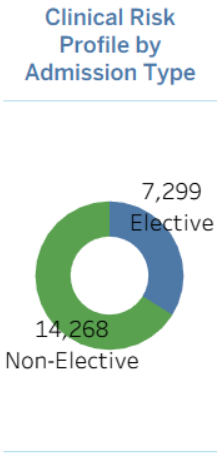
Oncology Doctors were the highest in number, while dr Michael smith was found to be the one with highest patient number.

6. Disease specialization by hospital or doctor



7. Most common admission types

   it was found that most admissions were emergency rather than elective

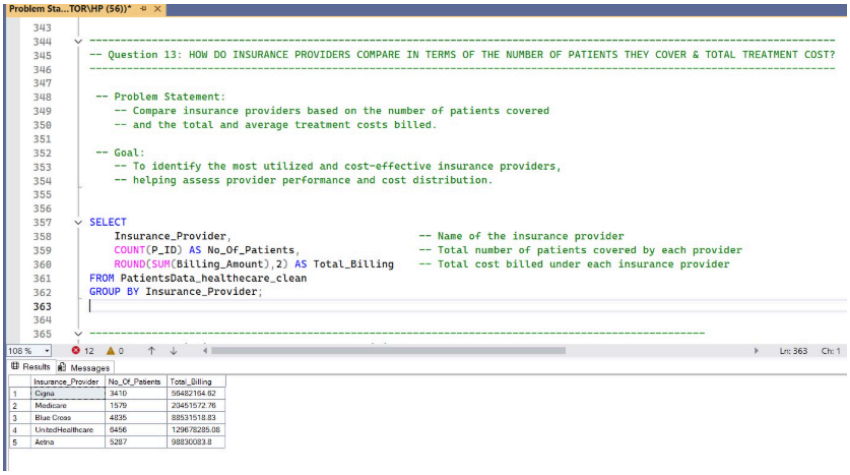


8. Treatment cost per disease

   Cancer was found to have the highest cost.

9. Comparison of insurance providers by client count and cost

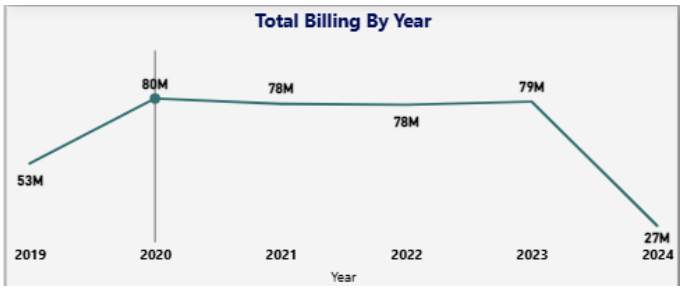
   United health care insurance has the highest subscribers.



```
-- Question 13: HOW DO INSURANCE PROVIDERS COMPARE IN TERMS OF THE NUMBER OF PATIENTS THEY COVER & TOTAL TREATMENT COST?

 -- Problem Statement:
    -- Compare insurance providers based on the number of patients covered
    -- and the total and average treatment costs billed.

 -- Goal:
    -- To identify the most utilized and cost-effective insurance providers,
    -- helping assess provider performance and cost distribution.


SELECT
    Insurance_Provider,                          -- Name of the insurance provider
    COUNT(P_ID) AS No_Of_Patients,               -- Total number of patients covered by each provider
    ROUND(SUM(Billing_Amount),2) AS Total_Billing    -- Total cost billed under each insurance provider
FROM PatientsData_healthecare_clean
GROUP BY Insurance_Provider;
```

	Insurance_Provider	No_Of_Patients	Total_Billing
1	Cigna	3410	96482164.62
2	Medicare	1579	20451572.76
3	Blue Cross	4835	88531518.83
4	UnitedHealthcare	6456	129678285.08
5	Aetna	5287	98830083.8

10. Monthly and yearly billing totals track spending trends.



11. Year-over-year change in case volume

```sql
-- Question 19: year-over-year change in case volume
-------------------------------------------------------------------------
  -- Problem Statement:
     -- Analyze how the number of patient admissions changes from year to year
     -- Determine whether admissions are increasing, decreasing, or stable over time.

  -- Goal:
     -- To calculate year-over-year (YoY) changes in admissions both in absolute
     -- numbers and percentages, providing insights for hospital planning, and resource allocation.

-- Yearly admissions with YoY changes, replacing NULL with 0 for the first year
-- Calculate yearly admissions with Year-over-Year (YoY) changes
WITH YearlyAdmissions AS (
    -- Step 1: Count total admissions per year
    SELECT
        YEAR(Date_of_Admission) AS Admission_Year,  -- Extract year from admission date
        COUNT(*) AS Total_Admissions                -- Total admissions in that year
    FROM PatientsData_healthecare_clean
    GROUP BY YEAR(Date_of_Admission)
)
SELECT
    Admission_Year,
    Total_Admissions,

    -- Step 2: Get the previous year's admissions using LAG()
    -- LAG(column) returns the value of 'column' from the previous row based on ORDER BY
    -- ISNULL(..., 0) replaces NULL (for the first year) with 0
    ISNULL(LAG(Total_Admissions) OVER (ORDER BY Admission_Year), 0) AS Prev_Year_Admissions,
    -- Step 3: Calculate absolute YoY change
    Total_Admissions - ISNULL(LAG(Total_Admissions) OVER (ORDER BY Admission_Year), 0) AS YoY_Change,
    -- Step 4: Calculate YoY percent change
    -- Percent change, cast as DECIMAL(5,2)
  CAST(
        CASE
            WHEN ISNULL(LAG(Total_Admissions) OVER (ORDER BY Admission_Year), 0) = 0 THEN 0
            ELSE (Total_Admissions - LAG(Total_Admissions) OVER (ORDER BY Admission_Year))
                 * 100.0 / CAST(LAG(Total_Admissions) OVER (ORDER BY Admission_Year) AS DECIMAL(10,2))
        END AS DECIMAL(5,2)
    ) AS YoY_Percent_Change
FROM YearlyAdmissions
ORDER BY Admission_Year;
```

67 % 12 0

Results Messages

	Admission_Year	Total_Admissions	Prev_Year_Admissions	YoY_Change	YoY_Percent_Change
1	2019	2855	0	2855	0.00
2	2020	4355	2855	1500	52.54
3	2021	4258	4355	-97	-2.23
4	2022	4280	4258	22	0.52
5	2023	4308	4280	28	0.65
6	2024	1511	4308	-2797	-64.93

Recommendations

. Disease Burden & Prevention

- **Cancer & Diabetes (Top Diseases)**

  • Hospitals should prioritize screening programs and early detection campaigns for cancer and diabetes.

• Public health authorities can invest in community awareness initiatives targeting lifestyle risk factors (diet, smoking, physical activity).

• Encourage partnerships with NGOs and insurance providers to subsidize preventive care.

- . **Age Group Focus**

  • Most affected: 36–55 and 56–75 years

  • Allocate more hospital resources (beds, specialized staff) to departments serving middle-aged and elderly patients.

  • Develop age-specific treatment protocols and chronic disease management programs.

  • Introduce telemedicine services for elderly patients to reduce hospital admissions.

- **Gender Distribution**

  • Higher female case percentage (56.75%)

  • Investigate gender-specific health needs (e.g., reproductive health, osteoporosis, autoimmune conditions).

  • Tailor awareness campaigns to women's health issues while ensuring equitable access for men.

Hospital Resource Utilization

- **Smith Group Hospital highest intake**
  - Benchmark Smith Group's **pricing policy and operational efficiency** for replication in other hospitals.
  - Balance patient loads by redistributing admissions across facilities to avoid overcrowding.

- **Financial & Insurance Trends**

  • Cancer highest treatment cost; United Healthcare most subscribers

  • Negotiate with insurance providers to expand coverage for high-cost diseases like cancer.

  • Introduce cost transparency dashboards for patients to understand billing.

• Explore alternative financing models (government subsidies, charity funds) for patients without insurance.

- Admission type

  More health awareness campaigns should be raised to inform people with how much early detection and routine checkups are important.

Appendix

Codes used

1. Python

```
import pandas as pd
dh=pd.read_excel('/content/Doctor_Healthecare.xlsx')
dh.head()
dh.tail()
dh.info()
dh.isnull().sum()
# to check the duplicates
dh.duplicated().sum()
# to capitalize first letter of each word
dh.Doctor=dh.Doctor.str.title()
# to remove extra space from beginning and end
dh['Doctor']=dh['Doctor'].str.strip()
hh=pd.read_excel('/content/Hospital_Healthcare.xlsx')
hh.head()
hh.tail()
# to check columns info and data type
hh.info()
# to check the nulls
hh.isnull().sum()
# to check the duplicates
hh.duplicated().sum()
# to check number of rows and columns
```

```
hh.shape
# to check number of unique values
hh.nunique()
# to replace 'and' with ',' in the hospital column
hh['Hospital']=hh['Hospital'].str.replace('and',',').str.replace(
'And',',')
# to remove extra spaces from the beginning and end of hospital
names
hh['Hospital']=hh['Hospital'].str.strip(',').str.strip()
# to replace '-' with ',' in the hospital column
hh['Hospital']=hh['Hospital'].str.replace('-',',')
ph=pd.read_excel('/content/Patients_Healthcare.xlsx')
# to check columns info and data type
ph.info()
# to check number of unique values
ph.nunique()
# to check the nulls
ph.isnull().sum()
# to check the duplicates
ph.duplicated().sum()
#to capitalize first letter of each word and remove extra space
from beginning and end
ph.Name=ph.Name.str.strip().str.title()
#to capitalize first letter, and remove extra space from
beginning and end
ph.Gender=ph.Gender.str.strip().str.capitalize()
# to find rows that are duplicated in all columns except age
```

```
dupes=ph[ph.duplicated(subset=['P_ID','Name','Gender','Blood
Type'],keep=False)]
print(len(dupes))
# to fill missing ages with the avg age of each gender
ph['Age']=ph['Age'].fillna(ph.groupby('Gender')['Age'].transform(
'mean'))
# to check the nulls
ph.isnull().sum()
# to change age column data type
ph.Age=ph.Age.astype(int)
# to check data types
ph.dtypes
pdh=pd.read_excel('/content/PatientsData_Healthcare.xlsx')
# to check columns info and data type
pdh.info()
# to check number of unique values
pdh.nunique()
# to check the nulls
pdh.isnull().sum()
# to check the duplicates
pdh.duplicated().sum()
pdh.drop_duplicates(inplace=True)
pdh.duplicated().sum()
# to drop rows with missing ID in P_ID column
pdh.dropna(subset=['P_ID'],inplace=True)
pdh.isnull().sum()
# to change p_ID column data type
```

```
pdh.P_ID=pdh.P_ID.astype(int)
pdh.dtypes
pdh['Test Results'].fillna('Unknown',inplace=True)
pdh.isnull().sum()
```

2. SQL

```
-------  SECTION 1: PATIENTS & CONDITION ANALYSIS

-----------------------------------------------
-- Question 1: WHAT IS THE MOST COMMON DISEASES?
-----------------------------------------------

 -- PROBLEM STATEMENT:
   --  Identify the most common diseases among patients to
prioritize
   --  healthcare interventions and understand disease burden.

 -- GOAL:
    -- To determine which diseases affect the most patients,
    -- helping to guide decision-making for healthcare planning,
prevention strategies, and resource distribution.

-- Frequency count of each disease
SELECT
    Medical_Condition,
    COUNT(*) AS Total_Common_Diseases
FROM PatientsData_healthecare_clean
```

```
GROUP BY Medical_Condition
ORDER BY Total_Common_Diseases DESC;


-------------------------------------------------------------------
--------
-- Question 2: WHAT IS THE IMPACT OF DISEASE ACROSS DIFFERENT AGE
GROUPS?
-------------------------------------------------------------------
--------

 -- PROBLEM STATEMENT:
    -- Determine which age groups are most affected by each
medical condition
    -- to improve disease prevention and age-targeted treatment
strategies.

 -- GOAL:
    -- Relationship between age and medical conditions in order
to identify the most affected age groups
    -- and support the development of targeted prevention and
treatment strategies.

-- Add a new column for age grouping
ALTER TABLE Patients_healthecare_clean
ADD Age_Group VARCHAR(10);
```

```
-- Assign age group categories based on patient age
UPDATE Patients_healthecare_clean
SET Age_Group = CASE
        WHEN Age BETWEEN 13 AND 17 THEN '13-17'
        WHEN Age BETWEEN 18 AND 35 THEN '18-35'
        WHEN Age BETWEEN 36 AND 55 THEN '36-55'
        WHEN Age BETWEEN 56 AND 75 THEN '56-75'
        WHEN Age BETWEEN 76 AND 80 THEN '76-80'
        ELSE '80+'
    END;

-- Verify age group assignments
SELECT DISTINCT Age, Age_Group
FROM Patients_healthecare_clean
ORDER BY Age;


-- Determine which age groups are most affected by each disease
WITH AgeDiseaseCounts AS (
    SELECT
        PD.Medical_Condition,           -- Disease name
        P.Age_Group,                    -- Age group of patient
        COUNT(*) AS patient_count,      -- Number of patients in
this age group
        ROW_NUMBER() OVER (             -- Window function: It
assigns a ranking (rn) within each disease, ordering the age
groups by number of patients in descending order.
```

```
            PARTITION BY PD.Medical_Condition
            ORDER BY COUNT(*) DESC
        ) AS rn                          -- age group for each
disease = largest number of patients in that disease's category.
    FROM Patients_healthecare_clean P
    JOIN PatientsData_healthecare_clean PD
        ON P.P_ID = PD.P_ID
    GROUP BY PD.Medical_Condition, P.Age_Group
)
SELECT
    Medical_Condition,
    Age_Group AS Most_Affected_Age_Group,
    patient_count
FROM AgeDiseaseCounts
WHERE rn = 1                             -- The age group with
the most patients for a disease gets rn = 1
ORDER BY Medical_Condition;              -- The next most
affected age group gets rn = 2



-------------------------------------------------------------------
--------------------
-- Question 3: HOW ARE DIFFERENT DISEASES DISTRIBUTED BETWEEN
MALE & FEMALE PATIENTS?
-------------------------------------------------------------------
--------------------
```

```
 -- PROBLEM STATEMENT:
    -- Assess whether certain diseases disproportionately affect
males or females.

 -- GOAL:
    -- Disease patterns by gender to identify whether certain
conditions affect males or females more



SELECT
    PD.Medical_Condition,
    P.Gender,
    COUNT(*) AS Patient_Count
FROM Patients_healthecare_clean P
JOIN PatientsData_healthecare_clean PD
    ON P.P_ID = PD.P_ID
GROUP BY PD.Medical_Condition, P.Gender
ORDER BY PD.Medical_Condition, Patient_Count DESC;



-----------------------------------------------------------------
-- Question 4: HOW DO TEST RESULTS VARY BY DISEASE OR MEDICATION?
-----------------------------------------------------------------

    -- PROBLEM STATEMENT:
       -- Determine how patients lab test results (Normal,
Abnormal, Inconclusive)
```

```
       -- vary by medical condition and medication to identify
patterns.

    -- GOAL:
       -- Monitor treatments and test outcomes to improve patient
care,
       -- medication effectiveness, and early treatment for
abnormal results.

-- Aggregate test results per disease       -- VARY BY DISEASE
SELECT
    Medical_Condition,
    Test_Results,
    COUNT(*) AS Result_Count
FROM PatientsData_healthecare_clean
GROUP BY Medical_Condition, Test_Results
ORDER BY Medical_Condition, Result_Count DESC;

-- Aggregate test results per medication    -- VARY BY MEDICATION
SELECT
    Medication,
    Test_Results,
    COUNT(*) AS Result_Count
FROM PatientsData_healthecare_clean
GROUP BY Medication, Test_Results
ORDER BY Medication, Result_Count DESC;
```

```
------- SECTION 2: Hospital & Doctor inisghts


-----------------------------------------------------------------
-- Question 5: WHICH HOSPITALS HAVE THE HIGHEST PATIENTS INTAKE?
-----------------------------------------------------------------


  -- PROBLEM STATEMENT:
       -- Identify which hospitals have the highest number of
patients.
       -- Understanding patient distribution helps hospitals
manage resources efficiently.


   -- GOAL:
      -- Allocate hospital resources, optimize staffing, and plan
healthcare services efficiently at high-volume hospitals.
      -- To determine which hospitals receive the most patients,
enabling better resource management and efficient hospital
planning.


SELECT P.H_ID, H.Hospital, COUNT(P.P_ID) AS PatientCount
FROM PatientsData_healthecare_clean P
JOIN Hospital_healthecare_clean H
ON H.H_id = P.H_ID
GROUP BY P.H_ID, H.Hospital
```

```
HAVING COUNT(P_ID) > 1
ORDER BY PatientCount DESC;



-------------------------------------------------------------------
---------------
-- Question 6: WHAT IS THE AVERAGE BILLING AMOUNT FOR PATIENTS AT
EACH HOSPITAL?
-------------------------------------------------------------------
---------------

 -- Problem Statement:
   -- Determine the average billing amount for patients at each
hospital to understand cost patterns
   -- and financial workload across healthcare facilities.

 -- Goal:
    -- Find the average patient charges, helping hospitals manage
finances,
    -- plan budgets, and identify areas for cost optimization.

SELECT P.H_ID,H.Hospital,ROUND(AVG(P.Billing_Amount),2) AS
AVG_Billing_Amount
FROM PatientsData_healthecare_clean P
JOIN Hospital_healthecare_clean H
ON P.H_ID = H.H_id
GROUP BY P.H_ID,H.Hospital
```

```
-----------------------------------------------------------------
---------------------------------
-- Question 7: HOW DOES A DOCTOR'S PERFORMANCE VARY BASED ON
PATIENT VOLUME OR TREATMENT OUTCOMES?
-----------------------------------------------------------------
---------------------------------
  -- Problem Statement:
     -- The objective is to assess each doctor's performance by
analyzing their patient volume
     -- and corresponding treatment outcomes based on test
results.

  -- Goal:
     -- To calculate a consistent, numeric success rate for every
doctor by quantifying
     -- their treatment outcomes and patient volume.
WITH DoctorPerformance AS (
    SELECT
        D_ID,
        COUNT(DISTINCT P_ID) AS Total_patients,          -- Count
how many unique patients each doctor treated
        SUM(CASE WHEN Test_Results = 'Normal' THEN 1 ELSE 0 END)
AS Normal_results,
        SUM(CASE WHEN Test_Results = 'Abnormal' THEN 1 ELSE 0
END) AS Abnormal_results,
```

```
        SUM(CASE WHEN Test_Results = 'Inconclusive' THEN 1 ELSE 0
END) AS Inconclusive_results,

        AVG(DATEDIFF(day, [Date_of_Admission], [Discharge_Date]))
AS AVG_Treatment_Days      -- Calculate the average treatment
duration
    FROM PatientsData_healthecare_clean
    GROUP BY D_ID    -- Grouping by doctor ensures we calculate
per doctor
)
-- CASE: Count how many of the doctor's patients had each type of
test result
-- CASE: Each CASE checks the Test_Results column and adds 1 if
the condition matches, else adds 0.
SELECT
    D_ID,
    Total_patients,         -- Number of patients treated
    Normal_results,         -- Patients with normal (successful)
outcomes
    Abnormal_results,       -- Patients with abnormal test
results
    Inconclusive_results,   -- Patients whose results were
inconclusive

    CAST(100.0 * Normal_results / NULLIF(Total_patients, 0) AS
DECIMAL(5,2)) AS Normal_result_rate_percentage,
```

```
                                        -- Calculate the percentage
of patients who had normal (successful) test results.
                                        -- NULLIF prevents division
by zero in case a doctor has zero patients.
    AVG_Treatment_Days                  -- Average number of
treatment days per doctor (from the CTE)
FROM DoctorPerformance
ORDER BY Normal_result_rate_percentage DESC, Total_patients DESC;
-- Sort doctors so the best performers (highest success rate)
appear first.
-- If two doctors have the same success rate, the one with more
patients comes first.




-------------------------------------------------------------------
---------------
-- Question 8: WHICH MEDICAL CONDITIONS DOES EACH DOCTOR TREAT
MOST FREQUENTLY?
-------------------------------------------------------------------
---------------


 -- Problem Statement:
    -- Determine which medical conditions are most commonly
treated by each doctor
    -- to understand their areas of expertise and patient care
focus.
```

```
 -- Goal:
    -- To identify the most frequently treated medical conditions
for each doctor,
    -- helping hospitals recognize doctor specializations, and
improve healthcare delivery.

SELECT D.Doctor, Medical_Condition, COUNT(P_ID) AS PatientCount
FROM PatientsData_healthecare_clean P
JOIN Doctor_healthecare_clean D
ON P.D_ID = D.D_ID
GROUP BY D.Doctor, Medical_Condition
HAVING COUNT(P_ID) > 1
ORDER BY PatientCount DESC;



-------------------------------------------------------------------
---------------
-- Question 9: WHAT IS THE COMMON ADMISSION TYPE PER HOSPITAL?
-------------------------------------------------------------------
---------------
  -- Problem Statement:
     -- Determine the most frequent admission type for each
hospital to understand
     -- patient flow patterns and hospital service utilization.
     -- This can help in identifying which admission types are
most common per hospital
```

```
     -- and guide resource allocation, staffing, and operational
planning.

  -- Goal:
     -- To identify the top admission type per hospital based on
historical data,
     -- allowing hospital administrators to prioritize resources,
     -- optimize scheduling, and plan for seasonal or recurring
trends in admissions.

SELECT H_ID, Admission_Type, admission_count
FROM (
    SELECT
        H_ID,
        Admission_Type,
        COUNT(*) AS admission_count,    -- counts how many
admissions of that type occurred at the hospital
        ROW_NUMBER() OVER (
            PARTITION BY H_ID           -- Start a new ranking
for each hospital
            ORDER BY COUNT(*) DESC      -- Rank by number of
admissions, descending
        ) AS rn
    FROM PatientsData_healthecare_clean
    GROUP BY H_ID, Admission_Type     -- Aggregate by hospital
and admission type
) AS ranked
```

```
WHERE rn = 1;                          -- Only pick the top
admission type per hospital



------- SECTION 3: Financial & Insurance Analysis

-----------------------------------------------------------------
-- Question 10: WHAT IS THE AVERAGE TREATMENT COST PER DISEASE?
-----------------------------------------------------------------

  -- Problem Statement:
     -- Determine the average treatment cost associated with each
disease
     -- to analyze cost variations and identify which medical
conditions
     -- require higher financial resources for patient care.

  -- Goal:
     -- To calculate and compare the average treatment cost per
disease,
     -- helping hospitals optimize budgeting, evaluate
cost-effectiveness,
     -- and improve financial planning for different medical
conditions.
```

```
SELECT Medical_Condition, ROUND(AVG(Billing_Amount),2) AS
AVG_Billing_Amount
FROM PatientsData_healthecare_clean
GROUP BY Medical_Condition



-------------------------------------------------------------------
-----------------------------------------------------
-- Question 11: HOW DO INSURANCE PROVIDERS COMPARE IN TERMS OF
THE NUMBER OF PATIENTS THEY COVER & TOTAL TREATMENT COST?
-------------------------------------------------------------------
-----------------------------------------------------

 -- Problem Statement:
    -- Compare insurance providers based on the number of
patients covered
    -- and the total and average treatment costs billed.

 -- Goal:
    -- To identify the most utilized and cost-effective insurance
providers,
    -- helping assess provider performance and cost distribution.



SELECT
    Insurance_Provider,                                  -- Name of
the insurance provider
```

```
    COUNT(P_ID) AS No_Of_Patients,                    -- Total
number of patients covered by each provider
    ROUND(SUM(Billing_Amount),2) AS Total_Billing    -- Total
cost billed under each insurance provider
FROM PatientsData_healthecare_clean
GROUP BY Insurance_Provider;



-------------------------------------------------------------------
-----------------------------------
-- Update the Billing_Amount column to be positive
SELECT Billing_Amount
FROM PatientsData_healthecare_clean
WHERE Billing_Amount < 0;

BEGIN TRANSACTION

UPDATE PatientsData_healthecare_clean
SET Billing_Amount = ABS(Billing_Amount);

COMMIT TRANSACTION
-------------------------------------------------------------------
-----------------------------------



-------------------------------------------------------------------
-------
```

```
-- Question 12: WHAT ARE TOTAL HOSPITAL BILLING AMOUNS PER MONTH
& YEAR?
-------------------------------------------------------------------
-------

  -- Problem Statement:
     -- Determine the total hospital billing amounts per month
and per year
     -- to understand how spending changes over time and identify
trends in hospital revenue.

  -- Goal:
     -- To track monthly and yearly billing totals, monitor
spending patterns,
     -- and provide insights for financial planning, budgeting,
and resource allocation.


-- Total hospital billing amounts per Month
SELECT
    MONTH(Date_of_Admission) AS Month,                -- Month of
admission
    ROUND(SUM(Billing_Amount),2) AS Total_Billing,    -- Total
billing for that month
    COUNT(*) AS Patient_Count                         -- Number
of admissions in that month
FROM PatientsData_healthecare_clean
```

```
GROUP BY MONTH(Date_of_Admission)
ORDER BY Total_Billing DESC;


-- Total hospital billing amounts per year
SELECT
    YEAR(Date_of_Admission) AS Year,                    -- Year of
admission
    ROUND(SUM(Billing_Amount),2) AS Total_Billing,             --
Total billing for that year
    COUNT(*) AS Patient_Count                           -- Number
of admissions in that year
FROM PatientsData_healthecare_clean
GROUP BY YEAR(Date_of_Admission)
ORDER BY Total_Billing DESC;



------- SECTION 4: Administrative Patterns

-------------------------------------------------------------------
------
-- Question 13: year-over-year change in case volume
-------------------------------------------------------------------
------

 -- Problem Statement:
```

```
     -- Analyze how the number of patient admissions changes from
year to year
     -- Determine whether admissions are increasing, decreasing,
or stable over time.

  -- Goal:
     -- To calculate year-over-year (YoY) changes in admissions
both in absolute
     -- numbers and percentages, providing insights for hospital
planning, and resource allocation.

-- Yearly admissions with YoY changes, replacing NULL with 0 for
the first year
-- Calculate yearly admissions with Year-over-Year (YoY) changes

WITH YearlyAdmissions AS (
    -- Step 1: Count total admissions per year
    SELECT
        YEAR(Date_of_Admission) AS Admission_Year,  -- Extract
year from admission date
        COUNT(*) AS Total_Admissions               -- Total
admissions in that year
    FROM PatientsData_healthecare_clean
    GROUP BY YEAR(Date_of_Admission)
)

SELECT
```

```
    Admission_Year,
    Total_Admissions,

    -- Step 2: Get the previous year's admissions using LAG()
    -- LAG(column) returns the value of 'column' from the
previous row based on ORDER BY
    -- ISNULL(..., 0) replaces NULL (for the first year) with 0
    ISNULL(LAG(Total_Admissions) OVER (ORDER BY Admission_Year),
0) AS Prev_Year_Admissions,

   -- Step 3: Calculate absolute YoY change
    Total_Admissions - ISNULL(LAG(Total_Admissions) OVER (ORDER
BY Admission_Year), 0) AS YoY_Change,

    -- Step 4: Calculate YoY percent change
    -- Percent change, cast as DECIMAL(5,2)
  CAST(
        CASE
            WHEN ISNULL(LAG(Total_Admissions) OVER (ORDER BY
Admission_Year), 0) = 0 THEN 0
            ELSE (Total_Admissions - LAG(Total_Admissions) OVER
(ORDER BY Admission_Year))
                * 100.0 / CAST(LAG(Total_Admissions) OVER (ORDER
BY Admission_Year) AS DECIMAL(10,2))
        END AS DECIMAL(5,2)
    ) AS YoY_Percent_Change
FROM YearlyAdmissions
```

```
ORDER BY Admission_Year;
```