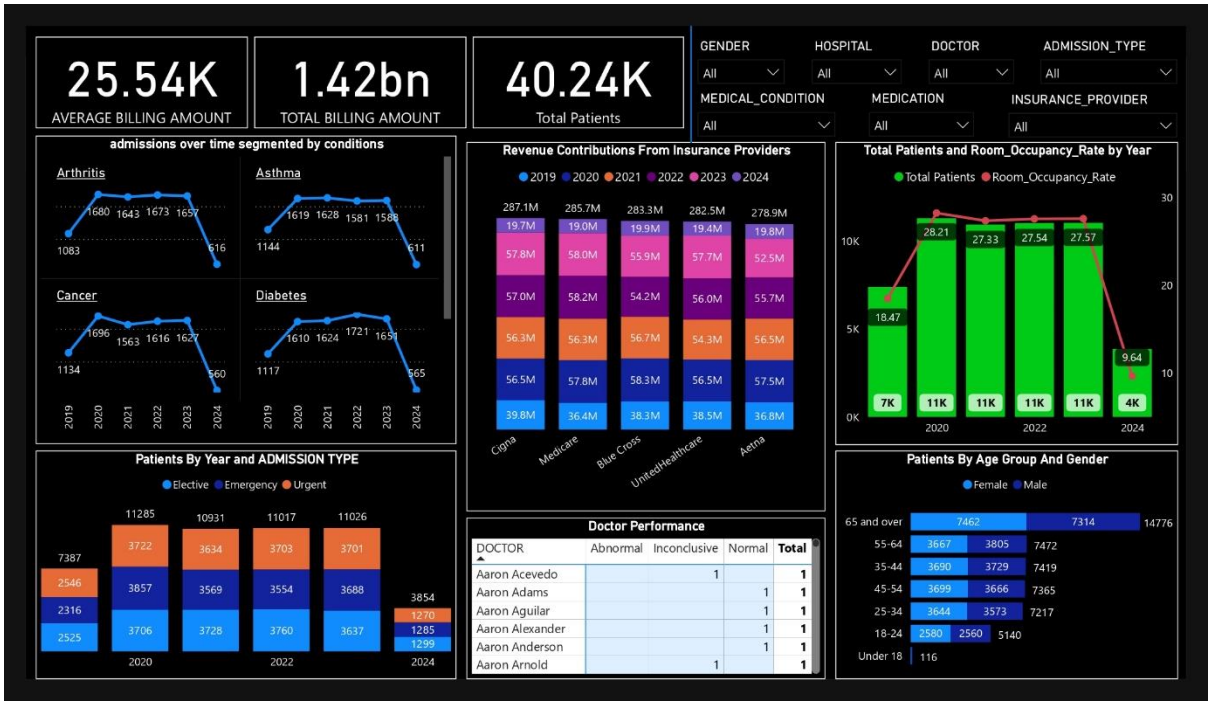
Healthcare Data Insights Dashboard



PROJECT OVERVIEW

The Healthcare Data Insights Dashboard project aims to leverage public healthcare datasets extracted from Kaggle to derive meaningful insights and facilitate data-driven decision-making in the healthcare sector. This project encompasses several key phases, including data extraction, cleaning, storage, analysis, and visualization.

Project Phases

1. Data Extraction:
    - Utilize Python to extract healthcare datasets from Kaggle. The datasets may include various health metrics, patient demographics, and treatment outcomes.
2. Data Cleaning:
    - Employ the Pandas library to clean the extracted data. This involves handling missing values, correcting data types, and ensuring consistency across the dataset.
3. Data Storage:
    - Save the cleaned dataset as a CSV file for easy access and further processing.
    - Load this cleaned data into Snowflake, a cloud-based data warehousing service, to facilitate scalable storage and querying.
4. Data Analysis:
    - Use SQL within Snowflake to perform complex queries on the dataset. This analysis will help derive important insights such as trends in patient demographics, treatment efficacy, and healthcare costs.
    - Generate reports that summarize these insights for stakeholders.
5. Data Visualization:
    - Connect Snowflake with Power BI to create an interactive and informative dashboard. This dashboard will visualize key performance indicators (KPIs), trends, and other significant findings from the analysis.
    - The dashboard will allow users to filter data based on various parameters such as time periods, patient groups, or treatment types.

Expected Outcomes

- A comprehensive understanding of healthcare trends derived from public datasets.
- An interactive dashboard that provides stakeholders with real-time insights into healthcare metrics.
- Enhanced ability for healthcare professionals to make informed decisions based on data analysis.

This project not only showcases technical skills in Python, SQL, and data visualization but also emphasizes the importance of data in improving healthcare outcomes.

DATA OVERVIEW

Patient Demographics

- Name: The full name of the patient.
- Age: The patient's age, which is crucial for understanding health risks and treatment responses.
- Gender: Gender identity, important for certain medical conditions and treatment considerations.
- Blood_Type: Essential for transfusions and certain medical treatments.

Medical History

- Medical_Condition: The primary diagnosis or health issues that led to admission.
- Date_of_Admission: The date when the patient was admitted to the hospital.
- Doctor: The attending physician responsible for the patient's care.
- Hospital: The facility where the patient is receiving treatment.

Treatment and Care

- Insurance_Provider: The health insurance company covering the patient's medical expenses.
- Billing_Amount: Total charges incurred during the hospital stay, essential for financial records and insurance claims.
- Room_Number: Indicates where the patient is accommodated during their stay.
- Admission_Type: Specifies whether the admission was planned (elective) or emergency-based.
- Discharge_Date: The date when the patient was released from the hospital.

Medications and Tests

- Medication: List of medications prescribed during the hospital stay.
- Test_Results: Outcomes from diagnostic tests conducted during admission, critical for ongoing treatment decisions.

DATA EXTRACTION AND CLEANING(PYTHON(PANDAS)

[1]:
```
pip install kaggle
```

```
Requirement already satisfied: kaggle in c:\users\saura\appdata\roaming\python\python312\site-packages (1.6.14)
Requirement already satisfied: six>=1.10 in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (1.16.0)
Requirement already satisfied: certifi>=2023.7.22 in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (2024.2.2)
Requirement already satisfied: python-dateutil in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (2.9.0.post0)
Requirement already satisfied: requests in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (2.32.2)
Requirement already satisfied: tqdm in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (4.66.4)
Requirement already satisfied: python-slugify in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (8.0.4)
Requirement already satisfied: urllib3 in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (2.2.1)
Requirement already satisfied: bleach in c:\users\saura\appdata\roaming\python\python312\site-packages (from kaggle) (6.1.0)
Requirement already satisfied: webencodings in c:\users\saura\appdata\roaming\python\python312\site-packages (from bleach->kaggle) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in c:\users\saura\appdata\roaming\python\python312\site-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\saura\appdata\roaming\python\python312\site-packages (from requests->kaggle) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in c:\users\saura\appdata\roaming\python\python312\site-packages (from requests->kaggle) (3.7)
Requirement already satisfied: colorama in c:\users\saura\appdata\roaming\python\python312\site-packages (from tqdm->kaggle) (0.4.6)
Note: you may need to restart the kernel to use updated packages.

[notice] A new release of pip is available: 24.1.2 -> 24.3.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

[6]:
```
import kaggle
import pandas as pd
import os
from kaggle.api.kaggle_api_extended import KaggleApi
```

[7]:
```
api = KaggleApi()
api.authenticate()
dataset_name = 'prasad22/healthcare-dataset'
download_path = './healthcare-dataset'
api.dataset_download_files(dataset_name, path=download_path, unzip=True) # Download dataset
```

```
Dataset URL: https://www.kaggle.com/datasets/prasad22/healthcare-dataset
```

[20]:
```
df = pd.read_csv(os.path.join(download_path, 'healthcare_dataset.csv'))
df.info()
print(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Name                55500 non-null  object
 1   Age                 55500 non-null  int64
 2   Gender              55500 non-null  object
 3   Blood Type          55500 non-null  object
 4   Medical Condition   55500 non-null  object
 5   Date of Admission   55500 non-null  object
 6   Doctor              55500 non-null  object
 7   Hospital            55500 non-null  object
 8   Insurance Provider  55500 non-null  object
 9   Billing Amount      55500 non-null  float64
 10  Room Number         55500 non-null  int64
 11  Admission Type      55500 non-null  object
 12  Discharge Date      55500 non-null  object
 13  Medication          55500 non-null  object
 14  Test Results        55500 non-null  object
dtypes: float64(1), int64(2), object(12)
memory usage: 6.4+ MB
             Name  Age  Gender Blood Type Medical Condition Date of Admission  \
0  Bobby JacksOn   30    Male         B-            Cancer        2024-01-31
1   LesLie TErRy   62    Male         A+           Obesity        2019-08-20
2    DaNnY sMiTH   76  Female         A-           Obesity        2022-09-22
3   andrEw waTtS   28  Female         O+          Diabetes        2020-11-18
4  adrIENNE bEll   43  Female        AB+            Cancer        2022-09-19

            Doctor                   Hospital Insurance Provider  \
0     Matthew Smith             Sons and Miller         Blue Cross
1   Samantha Davies                    Kim Inc           Medicare
2  Tiffany Mitchell                   Cook PLC              Aetna
3      Kevin Wells  Hernandez Rogers and Vang,           Medicare
4    Kathleen Hanna                White-White              Aetna

   Billing Amount  Room Number Admission Type Discharge Date   Medication  \
0   18856.281306          328         Urgent     2024-02-02  Paracetamol
1   33643.327287          265      Emergency     2019-08-26    Ibuprofen
2   27955.096079          205      Emergency     2022-10-07      Aspirin
3   37909.782410          450       Elective     2020-12-18    Ibuprofen
4   14238.317814          458         Urgent     2022-10-09   Penicillin

   Test Results
0        Normal
1  Inconclusive
2        Normal
3      Abnormal
4      Abnormal
```

[9]:
```
missing_values = df.isnull().sum()
print(missing_values)
```

```
Name                  0
Age                   0
Gender                0
Blood Type            0
Medical Condition     0
Date of Admission     0
Doctor                0
Hospital              0
Insurance Provider    0
Billing Amount        0
Room Number           0
Admission Type        0
Discharge Date        0
Medication            0
Test Results          0
dtype: int64
```

*[25]:
```
df.columns = df.columns.str.replace(' ', '_')
print(df)
print(df.columns)
```

```
                    Name  Age  Gender Blood_Type Medical_Condition  \
0          Bobby JacksOn   30    Male         B-            Cancer
1           LesLie TErRy   62    Male         A+           Obesity
2            DaNnY sMitH   76  Female         A-           Obesity
3           andrEw waTtS   28  Female         O+          Diabetes
4          adrIENNE bEll   43  Female        AB+            Cancer
...                  ...  ...     ...        ...               ...
55495  eLIZABeTH jaCkSOn   42  Female         O+            Asthma
55496         KYle pEREz   61  Female        AB-           Obesity
55497       HEATher WaNG   38  Female         B+      Hypertension
55498     JENniFER JOneS   43    Male         O-         Arthritis
55499       jAMES GARCiA   53  Female         O+         Arthritis

      Date_of_Admission            Doctor                     Hospital  \
0            2024-01-31     Matthew Smith               Sons and Miller
1            2019-08-20   Samantha Davies                      Kim Inc
2            2022-09-22  Tiffany Mitchell                     Cook PLC
3            2020-11-18       Kevin Wells    Hernandez Rogers and Vang,
4            2022-09-19    Kathleen Hanna                  White-White
...                 ...               ...                          ...
55495        2020-08-16     Joshua Jarvis               Jones-Thompson
55496        2020-01-23   Taylor Sullivan                 Tucker-Moyer
55497        2020-07-13    Joe Jacobs DVM  and Mahoney Johnson Vasquez,
55498        2019-05-25    Kimberly Curry      Jackson Todd and Castro,
55499        2024-04-02     Dennis Warren                Henry Sons and

      Insurance_Provider  Billing_Amount  Room_Number Admission_Type  \
0            Blue Cross    18856.281306          328         Urgent
1              Medicare    33643.327287          265      Emergency
2                 Aetna    27955.096079          205      Emergency
3              Medicare    37909.782410          450       Elective
4                 Aetna    14238.317814          458         Urgent
...                 ...             ...          ...            ...
55495        Blue Cross     2650.714952          417       Elective
55496             Cigna    31457.797307          316       Elective
55497  UnitedHealthcare    27620.764717          347         Urgent
55498          Medicare    32451.092358          321       Elective
55499             Aetna     4010.134172          448         Urgent

      Discharge_Date   Medication  Test_Results
0         2024-02-02  Paracetamol        Normal
1         2019-08-26    Ibuprofen  Inconclusive
2         2022-10-07      Aspirin        Normal
3         2020-12-18    Ibuprofen      Abnormal
4         2022-10-09   Penicillin      Abnormal
...              ...          ...           ...
55495     2020-09-15   Penicillin      Abnormal
55496     2020-02-01      Aspirin        Normal
55497     2020-08-10    Ibuprofen      Abnormal
55498     2019-05-31    Ibuprofen      Abnormal
55499     2024-04-29    Ibuprofen      Abnormal

[55500 rows x 15 columns]
Index(['Name', 'Age', 'Gender', 'Blood_Type', 'Medical_Condition',
```

[26]:
```
df.to_csv(r'C:\Users\saura\Downloads\tokyo-olympic-azure-data-engineering-project-main\healthcare_dataset.csv', index=False)
```

SQL QUERIES

Patient Demographics and Trends

1. **Age Distribution**:

- Analyze the age distribution of patients to identify the most common age groups admitted.

```
SELECT Age, COUNT(*) AS PatientCount
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY Age
ORDER BY Age;
```

	AGE	PATIENTCOUNT
1	13	14
2	14	18
3	15	28
4	16	24
5	17	32
6	18	772
7	19	805
8	20	750
9	21	822
10	22	817
11	23	790
12	24	837
13	25	836
14	26	815
15	27	786
16	28	760
17	29	787

2. **Gender Analysis**:

- Compare the number of male and female patients.

```
SELECT Gender, COUNT(*) AS PatientCount
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY Gender;
```

	GENDER	PATIENTCOUNT
1	Male	27774
2	Female	27726

3. **Blood Type Distribution**:

- Assess the distribution of blood types among patients.

```
SELECT Blood_Type, COUNT(*) AS Count
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY Blood_Type;
```

	BLOOD_TYPE	COUNT
1	B-	6944
2	A+	6956
3	A-	6969
4	O+	6917
5	AB-	6945
6	O-	6877
7	AB+	6947
8	B+	6945

4. **Prevalent Medical Conditions**:

- Identify the most common medical conditions among patients.

```
SELECT Medical_Condition, COUNT(*) AS ConditionCount
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY Medical_Condition
ORDER BY ConditionCount DESC;
```

	MEDICAL_CONDITION	CONDITIONCOUNT
1	Arthritis	9308
2	Diabetes	9304
3	Hypertension	9245
4	Obesity	9231
5	Cancer	9227
6	Asthma	9185

4. **Medication Prescriptions**:

- Analyze which medications are prescribed for specific conditions.

```
SELECT Medical_Condition, Medication, COUNT(*) AS PrescriptionCount
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY Medical_Condition, Medication;
```

	MEDICAL_CONDITION	MEDICATION	PRESCRIPTIONCOUNT
1	Cancer	Paracetamol	1853
2	Obesity	Ibuprofen	1851
3	Obesity	Aspirin	1865
4	Diabetes	Ibuprofen	1861
5	Asthma	Ibuprofen	1827
6	Asthma	Aspirin	1802
7	Hypertension	Lipitor	1848
8	Diabetes	Penicillin	1881
9	Arthritis	Paracetamol	1877
10	Obesity	Paracetamol	1793
11	Hypertension	Aspirin	1865

Financial Insights

6. **Billing Analysis**:

- Nxnnx

SELECT Medical_Condition, AVG(Billing_Amount) AS AverageBilling

FROM

"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"

GROUP BY Medical_Condition;

	MEDICAL_CONDITION	AVERAGEBILLING
1	Cancer	25161.79268776
2	Obesity	25805.97121222
3	Diabetes	25638.40557502
4	Asthma	25635.24930866
5	Hypertension	25497.09573824
6	Arthritis	25497.32706167

7. **Insurance Provider Analysis**:

- Determine which insurance providers are most frequently associated with hospital admissions.

SELECT Insurance_Provider, COUNT(*) AS AdmissionCount

FROM

"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"

GROUP BY Insurance_Provider;

	INSURANCE_PROVIDER	ADMISSIONCOUNT
1	Blue Cross	11059
2	Medicare	11154
3	Aetna	10913
4	UnitedHealthcare	11125
5	Cigna	11249

8. **Admission Trends Over Time**:

- Analyze patient admissions over specific periods (e.g., monthly or yearly).

```
SELECT YEAR(Date_of_Admission) AS AdmissionYear,
MONTH(Date_of_Admission) AS AdmissionMonth,
COUNT(*) AS TotalAdmissions
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY AdmissionYear, AdmissionMonth
ORDER BY AdmissionYear, AdmissionMonth;
```

	ADMISSIONYEAR	ADMISSIONMONTH	TOTALADMISSIONS
1	2019	5	686
2	2019	6	907
3	2019	7	957
4	2019	8	1001
5	2019	9	936
6	2019	10	1013
7	2019	11	959
8	2019	12	928
9	2020	1	950
10	2020	2	881
11	2020	3	937

9. **Length of Stay Analysis**:

- Calculate the average length of stay for patients based on admission type.

```
SELECT Admission_Type,
AVG(DATEDIFF(DAY, Date_of_Admission, Discharge_Date)) AS
AverageLengthOfStay
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
GROUP BY Admission_Type;
```

	ADMISSION_TYPE	AVERAGELENGTHOFSTAY
1	Urgent	15.408000
2	Elective	15.525328
3	Emergency	15.595052

Risk Assessment and Predictive Analytics

10. **Patient Risk Categorization**:

- Create a risk category based on medical conditions and test results.

```
SELECT Name,
 CASE
  WHEN Medical_Condition IN ('Critical Condition') THEN 'High Risk'
  WHEN Medical_Condition IN ('Chronic Condition') THEN 'Medium Risk'
  ELSE 'Low Risk'
  END AS RiskCategory
  FROM
 "HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET";
```

	NAME	RISKCATEGORY
1	Bobby Jackson	Low Risk
2	Leslie Terry	Low Risk
3	Danny Smith	Low Risk
4	Andrew Watts	Low Risk
5	Adrienne Bell	Low Risk
6	Emily Johnson	Low Risk
7	Edward Edwards	Low Risk
8	Christina Martinez	Low Risk
9	Jasmine Aguilar	Low Risk
10	Christopher Berg	Low Risk
11	Michelle Daniels	Low Risk

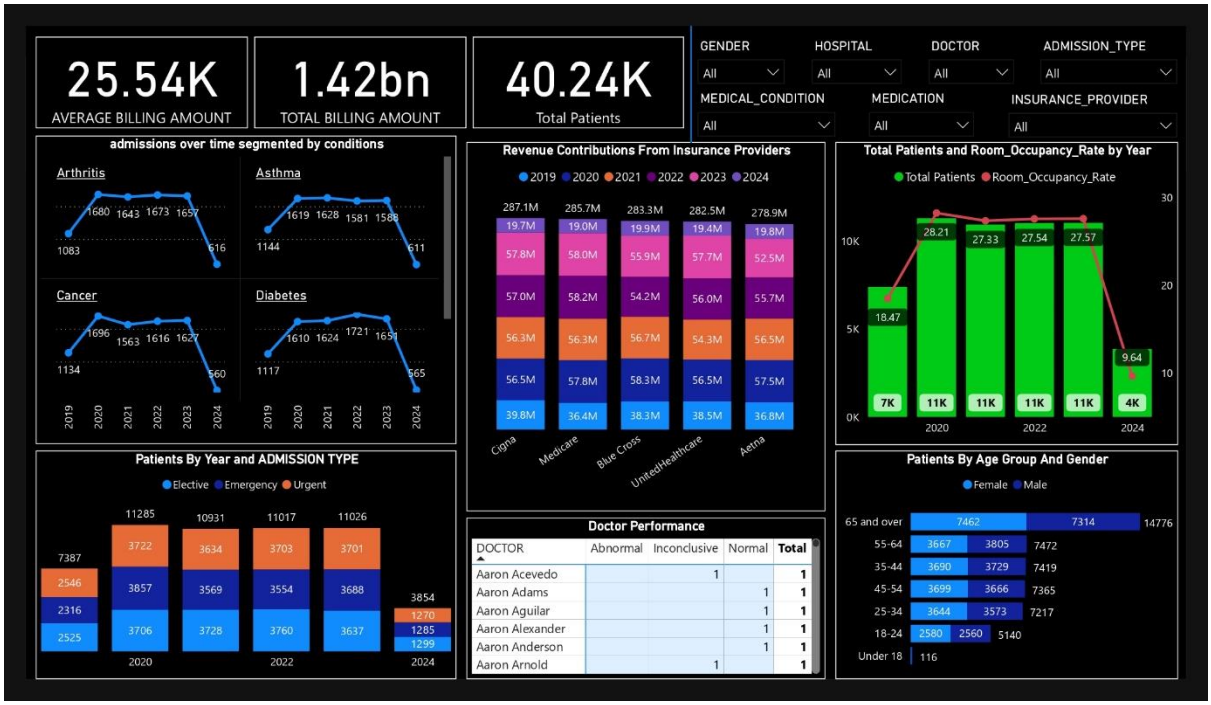
11. **Predictive Analytics for Readmissions**:

- Analyze factors that predict readmissions based on previous admissions data.

```
SELECT Doctor,
COUNT(*) AS ReadmissionCount,
AVG(Billing_Amount) AS AverageBilling,
AVG(DATEDIFF(DAY, Date_of_Admission, Discharge_Date)) As AverageStayDuration
FROM
"HEALTHCARE"."PUBLIC"."HEALTHCARE_DATASET"
WHERE Discharge_Date IS NOT NULL
GROUP BY Doctor
HAVING ReadmissionCount > 1;
```

	DOCTOR	READMISSIONCOUNT	AVERAGEBILLING	AVERAGESTAYDURATION
1	Matthew Smith	17	24790.30390584479700	13.764706
2	Kathleen Hanna	2	14238.31781393762300	20.000000
3	Kenneth Fletcher	2	29391.70105873364100	19.000000
4	Justin Kim	2	32433.63496332101750	16.000000
5	James Ellis	2	26852.12122299229250	19.500000
6	Emily Taylor	4	22229.08570820728210	14.250000
7	Matthew Thomas	3	28371.78207264323833	12.666667
8	John Smith	22	27732.25473534015377	14.363636
9	Scott Grant	2	3908.94656794631370	23.000000
10	Jeremiah Wolf	2	25425.72786260709000	8.000000
11	Brenda Lopez	3	34595.79149733048000	16.000000

POWERBI DASHBOARD



*The Room Occupancy Rate (ROR) is a crucial performance indicator in healthcare that measures the percentage of available hospital beds occupied by patients at a given time.

Patient Demographics

- **Age Groups**: The data categorizes patients into age groups, including:
  - **65 and over**: 14,776 patients
  - **55-64**: 7,472 patients
  - **45-54**: 7,365 patients
  - **35-44**: 7,419 patients
  - **25-34**: 7,217 patients
  - **18-24**: 5,140 patients
  - **Under 18**: 116 patients
- **Gender Breakdown**:
  - Female admissions: 7,462
  - Male admissions: 3,667

Admissions by Medical Condition

The report details hospital admissions segmented by specific medical conditions:

- **Arthritis**: 1,083 admissions
- **Asthma**: 616 admissions
- **Cancer**: 1,680 admissions
- **Diabetes**: 1,673 admissions

Financial Overview

Revenue Contributions from Insurance Providers (in millions)

- Cigna: $287.1M (2024)
- Medicare: $285.7M (2024)
- Blue Cross: $283.3M (2024)
- UnitedHealthcare: $282.5M (2024)
- Aetna: $278.9M (2024)

Total Billing Amount

The total billing amount reported is **$1.42 billion**, with an average billing amount of **$25.54K**.

Doctor Performance

The document also includes a performance summary for doctors based on patient outcomes:

- Each doctor is listed with the number of patients categorized as abnormal, inconclusive, or normal.

Admission Types by Year

- A breakdown of patient admissions by type (Elective, Emergency, Urgent) is provided for the years analyzed.

Patient Volume and Room Occupancy Rate

The report tracks total patient numbers and room occupancy rates over the years:

- Total Patients by Year shows fluctuations from **4K to over 11K**, indicating changes in patient volume.
- Room Occupancy Rates are also documented, highlighting trends in hospital capacity utilization.
- The Room Occupancy Rate is calculated using the formula:
  Room Occupancy Rate=(Total Available BedsTotal Occupied Beds)×100