HEALTHCARE PATIENT DATA ANALYSIS

BY HITESH ANANTH.UC

DATA PREPROCESSING

- Data preprocessing is a crucial step in healthcare data analysis that involves cleaning, transforming, and organizing raw patient data into a structured and usable format.
- In real-world healthcare datasets, raw data often contains:
- Duplicate patient records
- Inconsistent data formats
- Invalid or missing date values
- Raw numerical values that are difficult to interpret
- Unstructured categorical data
- To ensure accurate analysis and meaningful insights, these issues must be addressed before visualization.
- For this project, **Python** was used as the primary tool for data preprocessing due to its efficiency, flexibility, and powerful data-handling libraries such as **Pandas**.

```python
import pandas as pd

# Load dataset
df = pd.read_csv("healthcare_dataset.csv")

# Standardize column names
df.columns = (
    df.columns
    .str.lower()
    .str.strip()
    .str.replace(" ", "_")
)
```

- The healthcare dataset is loaded using the **Pandas** library, which allows efficient handling of large tabular data.
- The dataset is stored in a DataFrame to enable further cleaning and analysis.
- Column names are standardized to improve consistency and avoid errors during coding.
- All column names are converted to **lowercase** for uniformity.
- Extra spaces are removed to prevent formatting issues.
- Spaces in column names are replaced with **underscores**, making them Python-friendly.
- This step ensures clean, readable, and consistent column naming for smooth data processing.

```
# Convert date columns to datetime
df['date_of_admission'] = pd.to_datetime(df['date_of_admission'])
df['discharge_date'] = pd.to_datetime(df['discharge_date'])

# Clean text columns
text_columns = [
    'gender', 'blood_type', 'medical_condition',
    'admission_type', 'insurance_provider', 'test_results'
]
```

- Date columns such as **Date of Admission** and **Discharge Date** are converted into **datetime format** using Pandas.
- Converting dates enables accurate **time-based analysis**, such as calculating length of stay and monthly trends.
- Text-based columns like **gender, blood type, medical condition, admission type, insurance provider, and test results** are identified for cleaning.
- Grouping text columns allows consistent formatting and easier categorical analysis.
- This step improves data accuracy and ensures reliable results during visualization and dashboard creation.

•Text-based categorical columns are cleaned by:

•Removing unnecessary spaces

•Converting text into a consistent title case format

•This ensures uniform values across categories such as gender, medical condition, and admission type.

•Invalid numeric records are removed by filtering:

•Age values greater than zero

•Billing amounts greater than zero

•Duplicate patient records are identified and removed to avoid repeated analysis.

•A new feature, **Length of Stay**, is created by calculating the difference between discharge date and admission date.

•This feature helps measure hospital efficiency and patient stay duration.

```
for col in text_columns:
    df[col] = df[col].str.strip().str.title()

# Remove invalid numeric values
df = df[(df['age'] > 0) & (df['billing_amount'] > 0)]

# Remove duplicate rows
df.drop_duplicates(inplace=True)

# Feature engineering: Length of hospital stay
df['length_of_stay'] = (
    df['discharge_date'] - df['date_of_admission']
).dt.days
```

- Patient age values are grouped into meaningful **age categories** such as:
  - Child
  - Young Adult
  - Adult
  - Middle Age
  - Senior
- Age grouping helps simplify demographic analysis and improves dashboard readability.
- The cleaned and transformed dataset is exported to a new CSV file for further use in **Excel visualization**.
- Exporting the cleaned data ensures a clear separation between raw data and processed data.
- A final validation step is performed to confirm that data cleaning was completed successfully and to verify the dataset size.

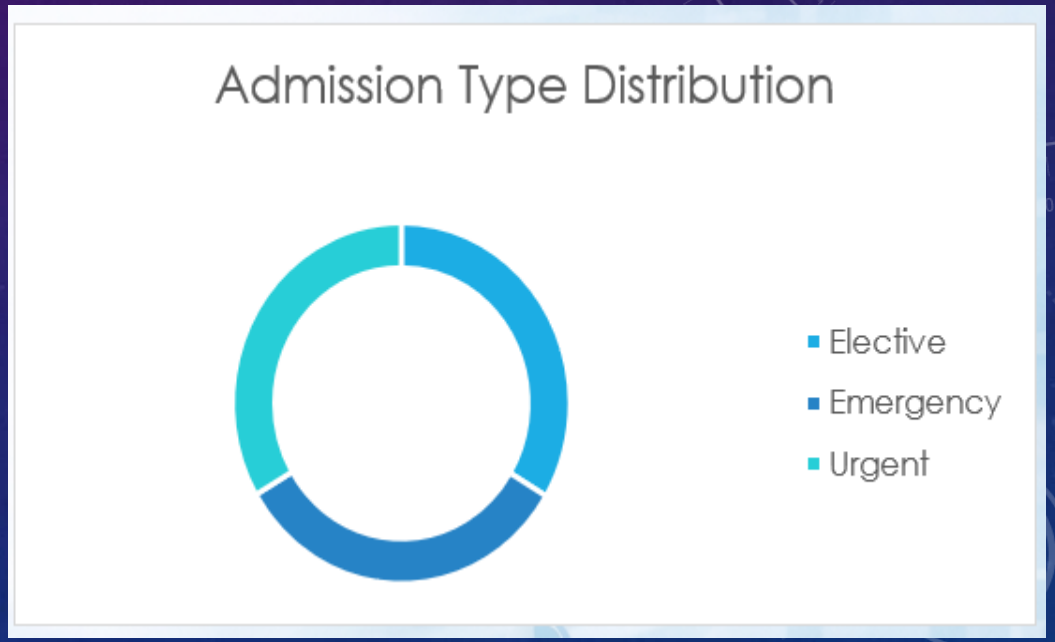
```
# Feature engineering: Age groups
bins = [0, 18, 35, 50, 65, 100]
labels = ['Child', 'Young Adult', 'Adult', 'Middle Age', 'Senior']
df['age_group'] = pd.cut(df['age'], bins=bins, labels=labels)

# Export cleaned dataset for Excel visualization
df.to_csv("cleaned_healthcare_data.csv", index=False)

# Final check
print("Cleaning completed successfully!")
print("Final dataset shape:", df.shape)
```

CLEANED DATASET

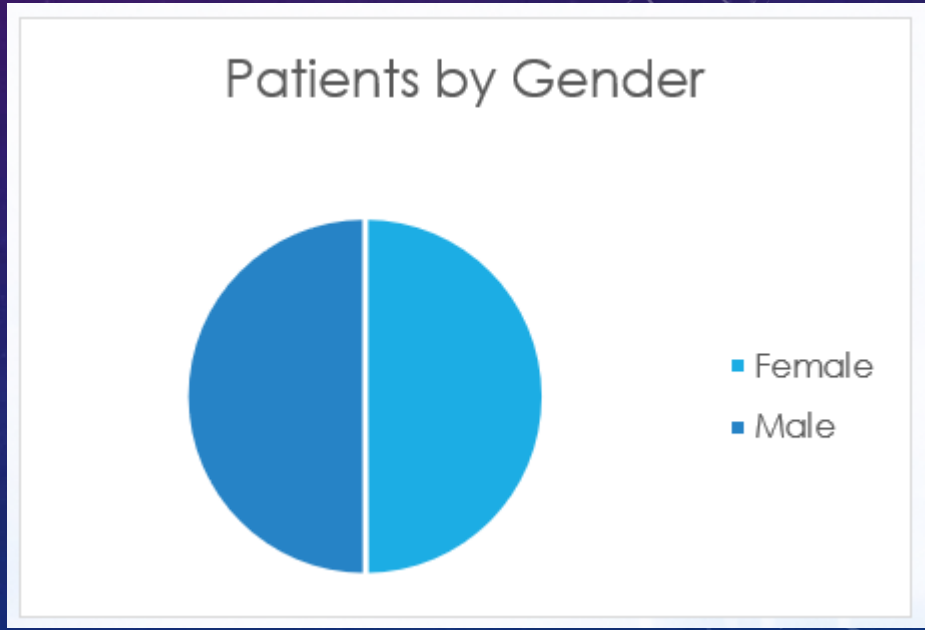
	A	B	C	D	E	F	G	H	I	J	K	L	
1	name	age	gender	blood_type	medical_condition	date_of_admission	doctor	hospital	insurance_provider	billing_amount	room_number	admission_type	discha
2	Bobby JacksOn	30	Male	B-	Cancer	31-01-2024	Matthew Smith	Sons and Miller	Blue Cross	18856.28131	328	Urgent	0
3	LesLie TErRy	62	Male	A+	Obesity	20-08-2019	Samantha Davies	Kim Inc	Medicare	33643.32729	265	Emergency	2
4	DaNnY sMitH	76	Female	A-	Obesity	22-09-2022	Tiffany Mitchell	Cook PLC	Aetna	27955.09608	205	Emergency	0
5	andrEw waTtS	28	Female	O+	Diabetes	18-11-2020	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.78241	450	Elective	1
6	adrIENNE bEll	43	Female	Ab+	Cancer	19-09-2022	Kathleen Hanna	White-White	Aetna	14238.31781	458	Urgent	0
7	EMILY JOHNSOn	36	Male	A+	Asthma	20-12-2023	Taylor Newton	Nunez-Humphrey	Unitedhealthcare	48145.11095	389	Urgent	2
8	edwArD EDWaRDs	21	Female	Ab-	Diabetes	03-11-2020	Kelly Olson	Group Middleton	Medicare	19580.87234	389	Emergency	1
9	CHrisTInA MARtinez	20	Female	A+	Cancer	28-12-2021	Suzanne Thomas	Powell Robinson and Valdez,	Cigna	45820.46272	277	Emergency	0
10	JASmINe aGuIlaR	82	Male	Ab+	Asthma	01-07-2020	Daniel Ferguson	Sons Rich and	Cigna	50119.22279	316	Elective	1
11	ChRISTopher BerG	58	Female	Ab-	Cancer	23-05-2021	Heather Day	Padilla-Walker	Unitedhealthcare	19784.63106	249	Elective	2
12	mIchElLe daniELs	72	Male	O+	Cancer	19-04-2020	John Duncan	Schaefer-Porter	Medicare	12576.79561	394	Urgent	2
13	aaRon MARtiNeZ	38	Female	A-	Hypertension	13-08-2023	Douglas Mayo	Lyons-Blair	Medicare	7999.58688	288	Urgent	0
14	connOR HANsEn	75	Female	A+	Diabetes	12-12-2019	Kenneth Fletcher	Powers Miller, and Flores	Cigna	43282.28336	134	Emergency	2
15	rObeRt bAuer	68	Female	Ab+	Asthma	22-05-2020	Theresa Freeman	Rivera-Gutierrez	Unitedhealthcare	33207.70663	309	Urgent	1
16	bROOkE brady	44	Female	Ab+	Cancer	08-10-2021	Roberta Stewart	Morris-Arellano	Unitedhealthcare	40701.59923	182	Urgent	1
17	MS. nAtalIE gAMble	46	Female	Ab-	Obesity	01-01-2023	Maria Dougherty	Cline-Williams	Blue Cross	12263.35743	465	Elective	1
18	haley perkins	63	Female	A+	Arthritis	23-06-2020	Erica Spencer	Cervantes-Wells	Unitedhealthcare	24499.8479	114	Elective	1
19	mRS. jamiE cAMPBELl	38	Male	Ab-	Obesity	08-03-2020	Justin Kim	Torres, and Harrison Jones	Cigna	17440.46544	449	Urgent	0
20	LuKE BuRgEss	34	Female	A-	Hypertension	04-03-2021	Justin Moore Jr.	Houston PLC	Blue Cross	18843.02302	260	Elective	1
21	dANIEL schmIdt	63	Male	B+	Asthma	15-11-2022	Denise Galloway	Hammond Ltd	Cigna	23762.20358	465	Elective	2
22	tIMOTHY burNs	67	Female	A-	Asthma	28-06-2023	Krista Smith	Jones LLC	Blue Cross	42.51458855	115	Elective	0
23	ChRISToPHEr BRiGhT	48	Male	B+	Asthma	21-01-2020	Gregory Smith	Williams-Davis	Aetna	17695.91162	295	Urgent	0
24	KatHRYn StewArt	58	Female	O+	Arthritis	12-05-2022	Vanessa Newton	Clark-Mayo	Aetna	5998.102908	327	Urgent	1
25	dR. EilEEn thomPsoN	59	Male	A+	Asthma	02-08-2021	Donna Martinez MD	and Sons Smith	Aetna	25250.05243	119	Urgent	1
26	PAUl hEndERsOn	72	Female	Ab+	Hypertension	15-05-2020	Stephanie Kramer	Wilson Group	Medicare	33211.29542	109	Emergency	0
27	PeTER fiTzgeRaLd	73	Male	Ab+	Obesity	15-05-2020	Angela Contreras	Garner-Bowman	Medicare	19746.83201	162	Urgent	2
28	cathy sMaLl	51	Female	O-	Asthma	23-12-2023	Wendy Glenn	Brown, and Jones Weaver	Blue Cross	26786.52956	401	Elective	1

- **Total Patients (54,860):**
Represents the total number of patient records analyzed in the dataset, indicating the overall scale of healthcare operations.
- **Average Billing Amount (₹25,595):**
Shows the average cost incurred per patient, helping evaluate hospital billing patterns and financial performance.
- **Average Length of Stay (15.49 days):**
Indicates the average number of days patients spend in the hospital, reflecting hospital efficiency and resource utilization.
- **Average Age (51.53 years):**
Represents the mean age of patients, providing insights into the dominant age group seeking healthcare services.
- These KPIs provide a **high-level summary** of patient demographics, operational efficiency, and financial trends.
- All KPIs are **dynamic** and update automatically based on slicer selections in the dashboard.

- This chart represents the distribution of patient admissions across different **admission types**.
- The three main admission categories analyzed are:
  - **Emergency**
  - **Urgent**
  - **Elective**
- Emergency admissions form a significant portion of total hospital visits, indicating a high demand for immediate medical care.
- Urgent admissions represent cases requiring prompt attention but not immediate emergencies.
- Elective admissions account for planned procedures and scheduled treatments.
- This analysis helps hospitals understand **patient inflow patterns** and plan resources accordingly.



- This chart illustrates the distribution of patients based on **gender**.
- It compares the number of **male** and **female** patients in the dataset.
- The visualization shows a relatively balanced distribution between male and female patients.
- Understanding gender-wise distribution helps healthcare providers:
  - Plan gender-specific healthcare services
  - Identify demographic patterns in hospital visits
- This insight supports better patient care planning and resource allocation.



•This chart shows the distribution of patients across different **age groups**.

•Patient ages are grouped into meaningful categories such as:

•0–19

•20–30

•31–40

•41–50

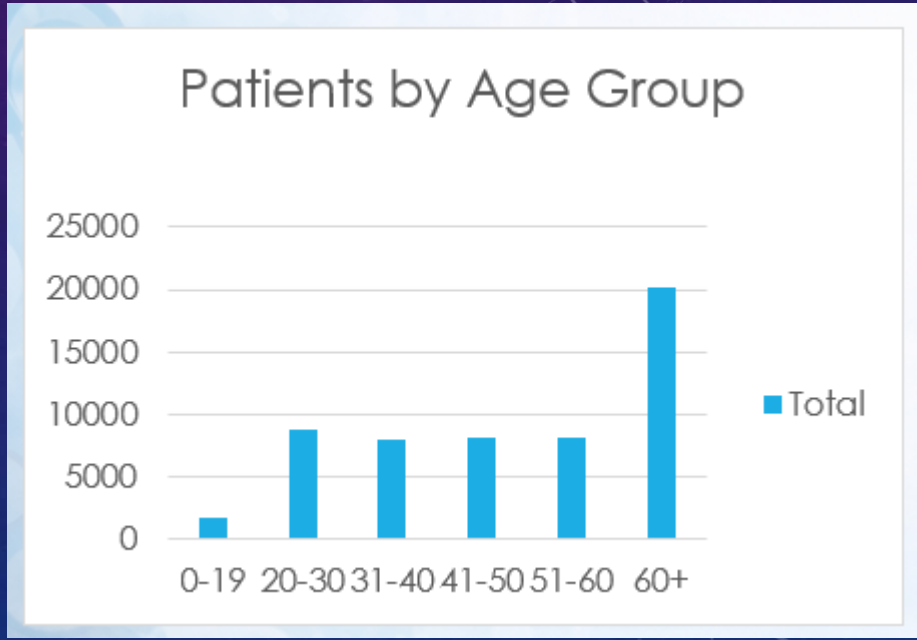
•51–60

•60+

•The **60+ age group** represents the highest number of patients, indicating a higher healthcare demand among senior citizens.

•Adult and middle-aged groups also show significant patient volumes.

•This analysis helps healthcare providers understand which age groups require more medical attention and resources.



•This chart displays the number of patients categorized by different **medical conditions**.

•Common medical conditions analyzed include:

•Diabetes

•Arthritis

•Hypertension

•Cancer

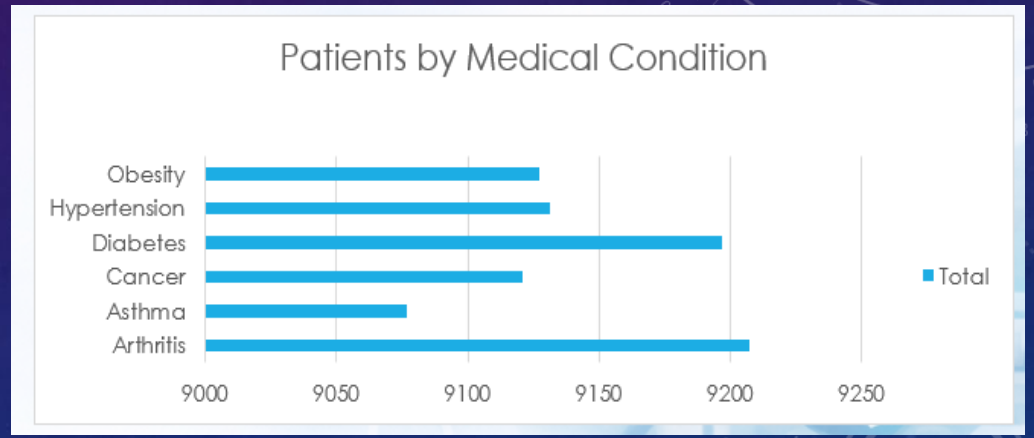
•Obesity

•Asthma

•**Diabetes and Arthritis** show the highest patient counts, indicating a high prevalence of chronic conditions.

•Conditions such as asthma and obesity have comparatively lower patient counts.

•This analysis helps healthcare providers identify **high-impact medical conditions** that require focused treatment plans and long-term care strategies.



•This chart shows the **average number of days patients stay in the hospital** for different medical conditions.

•It helps analyze hospital efficiency and resource utilization across conditions.

•**Asthma and Arthritis** patients have the longest average hospital stays, indicating more intensive or prolonged treatment.

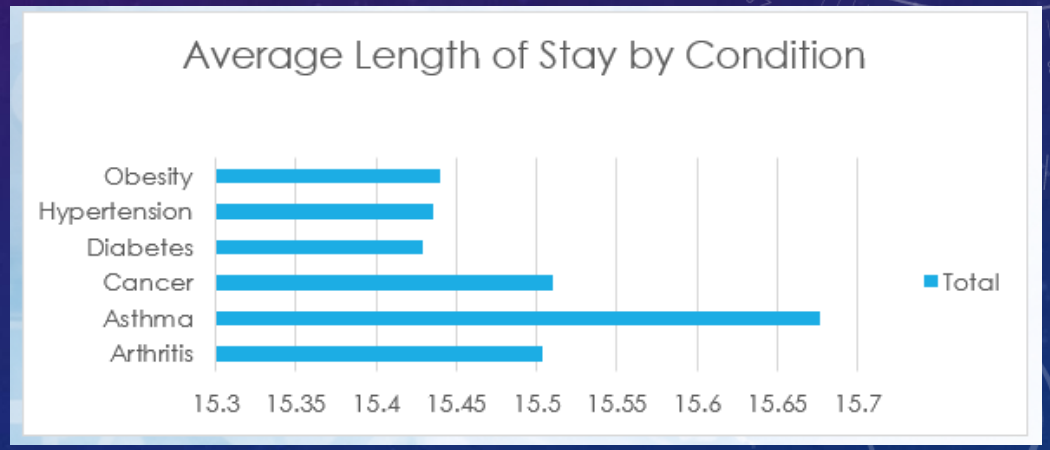
•Conditions such as **Obesity, Hypertension, and Diabetes** have comparatively shorter hospital stays.

•Understanding length of stay by condition helps hospitals:

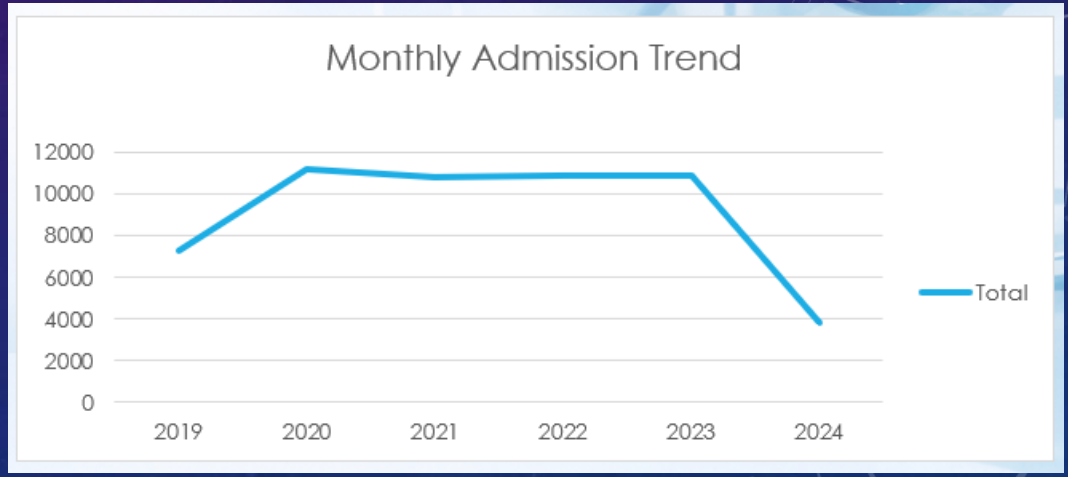
•Optimize bed utilization

•Improve treatment planning

•Allocate medical resources effectively



- This chart shows the **trend of patient admissions over time**.
- Admissions increased significantly from 2019 to 2020, indicating a rise in hospital visits.
- From 2020 to 2023, patient admissions remained relatively stable.
- A noticeable decline in admissions is observed in 2024.
- Time-based trend analysis helps hospitals:
  - Identify seasonal or yearly patterns
  - Plan staffing and resource allocation
  - Anticipate future admission demands



•This chart compares the **average billing amount per patient** across different insurance providers.

•Insurance providers analyzed include:

•Aetna

•Blue Cross

•Cigna

•Medicare

•UnitedHealthcare

•**Medicare** shows the highest average billing amount, indicating higher treatment costs or longer care duration for covered patients.

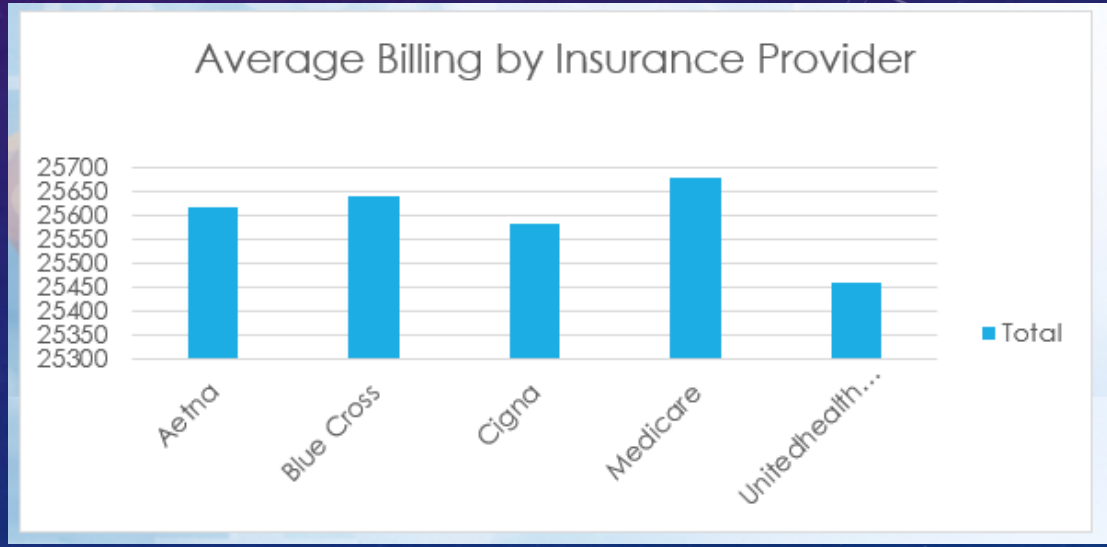
•**UnitedHealthcare** has the lowest average billing amount among the providers.

•This analysis helps hospitals:

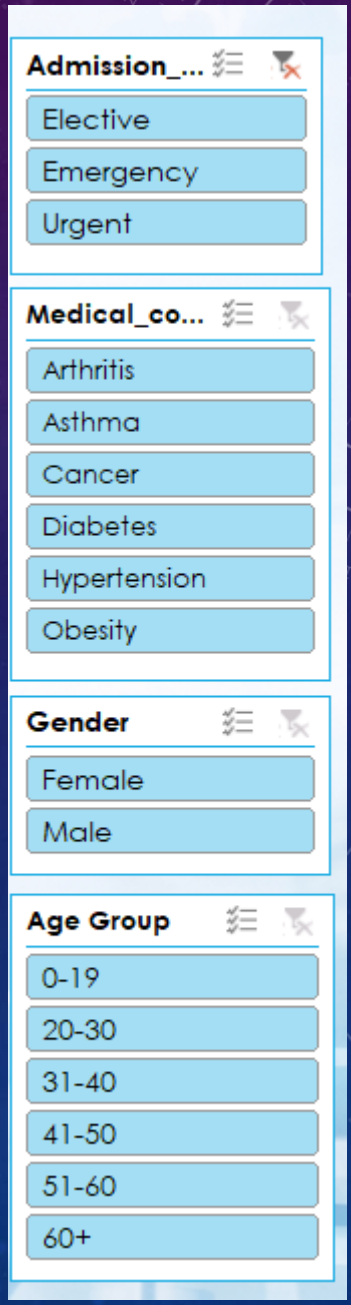
•Understand insurance-wise revenue patterns

•Identify high-cost insurance plans

•Support financial planning and negotiations with insurers



- Slicers are used to make the dashboard **interactive and user-friendly**.
- Users can filter the dashboard based on:
- **Admission Type** (Elective, Emergency, Urgent)
- **Medical Condition** (Arthritis, Asthma, Cancer, Diabetes, Hypertension, Obesity)
- **Gender** (Male, Female)
- **Age Group** (0–19, 20–30, 31–40, 41–50, 51–60, 60+)
- All charts and KPIs update dynamically based on slicer selection.
- This allows users to:
- Focus on specific patient segments
- Compare trends across demographics and conditions
- Perform quick, scenario-based analysis without changing the data



CONCLUSION

- In this project, I analyzed a healthcare patient dataset to understand **patient demographics, admission patterns, billing behavior, and hospital efficiency**.
- After cleaning and preprocessing the raw data using **Python**, I identified key factors such as **age groups, admission types, medical conditions, and length of hospital stay** that influence healthcare operations.
- I observed that **senior and adult patients** form the majority of hospital visits, **emergency admissions** contribute significantly to hospital workload, and certain **medical conditions require longer hospital stays**.
- Based on these findings, I designed an **interactive Excel dashboard** with KPIs, charts, and slicers to present the insights in a clear and actionable way.
- This dashboard helps healthcare stakeholders quickly identify trends, compare patient segments, and support **data-driven decision-making**.