INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

DATA ANALYSIS ON HEALTHCARE DATASET

Presented By: K.Bhargavi

INNOMATICS
RESEARCH LABS

About me

I am KUNDRAPU BHARGAVI , a recent graduate with a Bachelor of Science, specializing in Computer Science, from Chaitanya Degree College for Women.

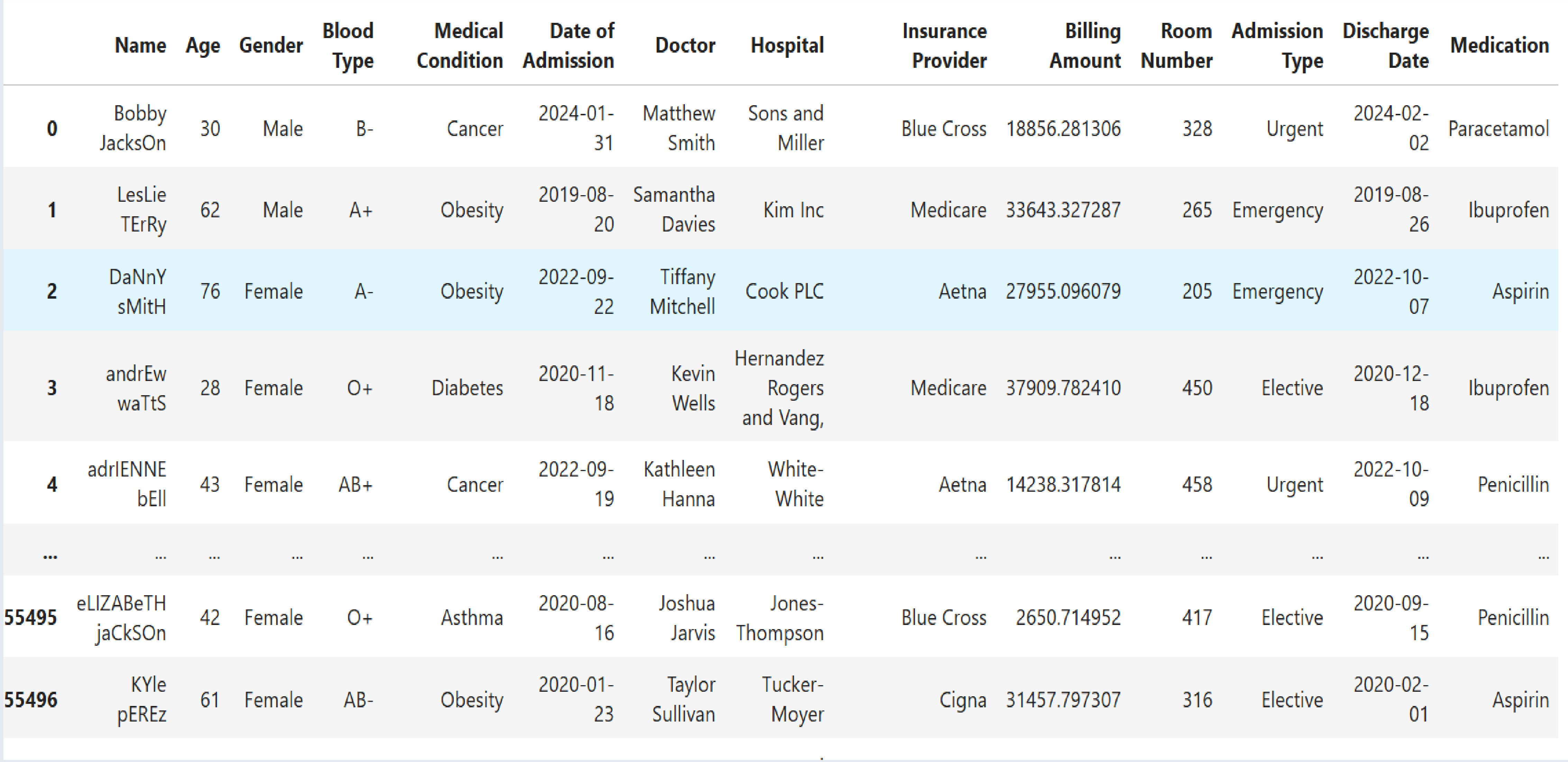
Why Data Analysis?

I choose Data Analysis because it blends problem-solving, technology, and decision-making—three things I truly enjoy. I love working with data to find patterns and use them to solve real-world problems. It’s a field full of learning, innovation, and impact across industries.

Analysis of Healthcare Dataset

- To analyze how **Hospital Patient Records Dataset** provides insights into admissions, treatments, billing, and overall healthcare operations across multiple hospitals.
- The dataset helps understand **patient demographics, medical conditions, healthcare provider involvement, and financial aspects** of treatment.
- The dataset consists of **55,500 rows** and **15 columns**.
- It includes **3 numerical columns** such as Age, Room Number, Billing Amount and **12 categorical columns** such as Name, Gender, Blood Type, Medical Condition, Date of Admission, Discharge Date, Doctor, Insurance Provider, Admission Type, Medication, Test Results.
- **Key Features Include:** Patient Demographics, Clinical Information, Hospitalization Details, Healthcare Provides, Financial Aspect.
- No columns does not contain any kind of missing values.

Understanding the Columns



	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen
2	DaNnY sMitH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin
3	andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen
4	adrIENNE bEll	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
55495	eLIZABeTH jaCkSOn	42	Female	O+	Asthma	2020-08-16	Joshua Jarvis	Jones-Thompson	Blue Cross	2650.714952	417	Elective	2020-09-15	Penicillin
55496	KYle pEREz	61	Female	AB-	Obesity	2020-01-23	Taylor Sullivan	Tucker-Moyer	Cigna	31457.797307	316	Elective	2020-02-01	Aspirin

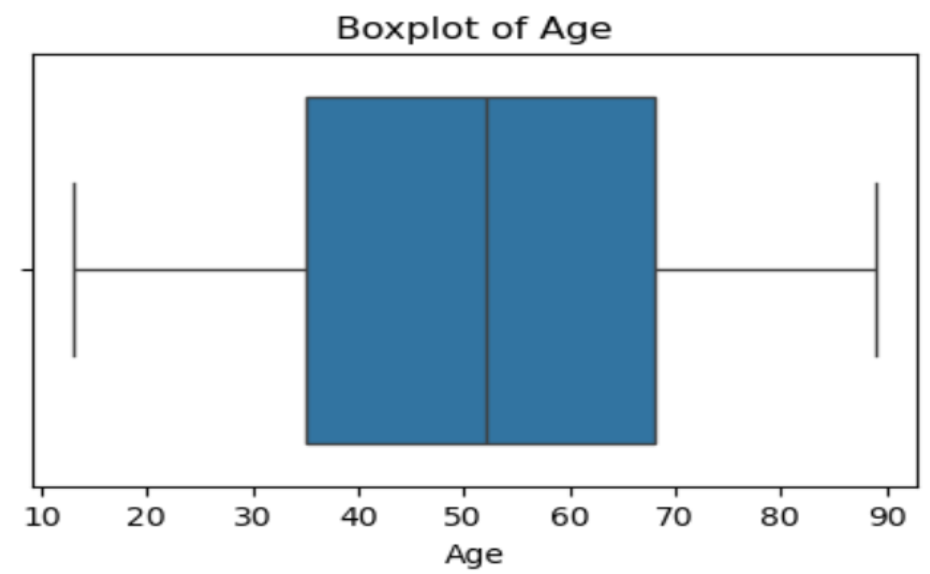
Handling Outliers

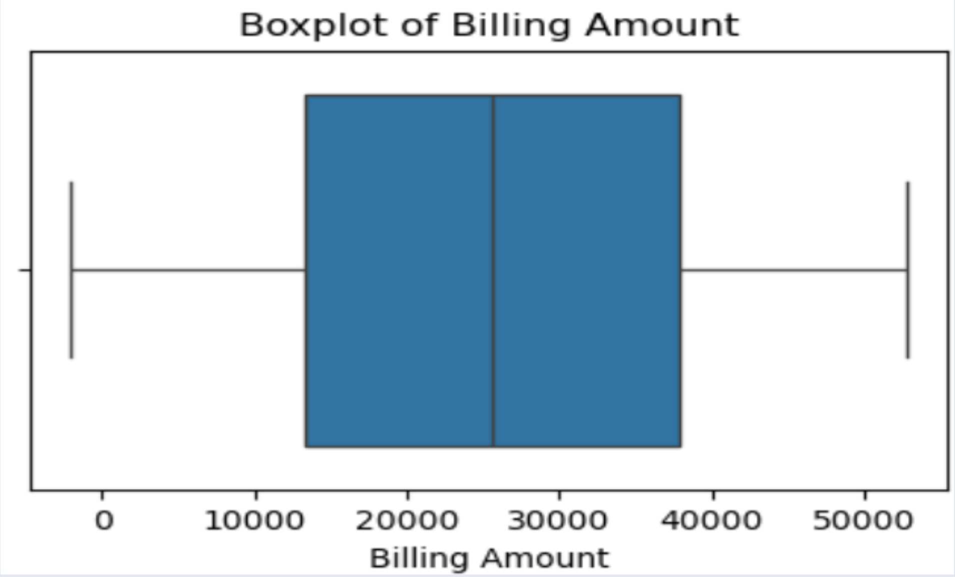
- Checked Outliers in numerical features such as Age, Billing Amount, and Length of Stay
- Techniques Used:

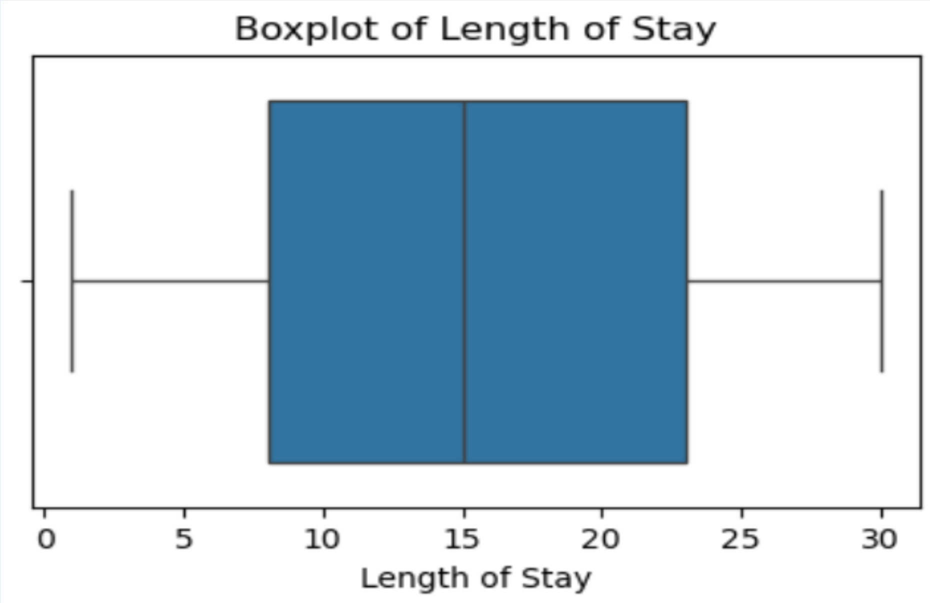
  Boxplot and Z-score to detect unusual values

  IQR method to validate statistical thresholds
- Findings: No significant outliers were detected → dataset already clean
- Why It Matters: Outliers can distort averages & reduce model accuracy
- Outcome: Clean dataset ensures reliable insights and better predictions

Handling Outliers- Box Plot







Handling Missing Data

- Check Performed: Verified all columns (Billing Amount, Age, Length of Stay, etc.) for missing values
- Techniques Considered: Mean/Median imputation for numerical & mode for categorical (not required as no gaps found)
- Findings: No missing values detected in the dataset
- Why It Matters: Completeness of data avoids bias, ensures higher accuracy
- Outcome: Dataset is consistent and now ready for analysis & modeling

Handling Duplicates

- Issue Found: Duplicate records in Name, Age, Admission Type, and Billing Amount
- Techniques Used:

  Checked duplicates with {.duplicated()} function

  Removed exact duplicate rows to avoid data redundancy
- Why It Matters: Duplicates can skew analysis and misrepresent patient statistics
- Outcome: Final dataset contains only unique patient records, ensuring reliability in healthcare insights.

Fixing Inconsistencies in Categorical Data:

- Standardized values : (e.g., “male”, “Male”, “MALE” → “Male”)
- Removed formatting issues: Extra spaces, Special characters cleaned
- Unified categories :(e.g., “normal”, “Normal”, “NORMAL” → “Normal”)
- Ensured consistency: Gender, Blood Type, Admission Type, and Test Results standardized for reliable analysis.

Data Type Conversion:

- Converted numeric fields (*Age*, *Billing Amount*, *Length of Stay*, *Year*, *Room Number*) into proper int/float types
- Converted date fields (*Date of Admission*, *Discharge Date*) into datetime format for time-based analysis
- Categorical fields (*Gender*, *Blood Type*, *Admission Type*, *Medical Condition*, *Insurance Provider*, *Medication*, *Test Results*, *Age Group*) changed to category for efficiency
- Identifiers (*Name*, *Hospital*, *Doctor*) retained as string/object to preserve uniqueness
- Ensured correct data types for accurate analysis, reduced memory usage, and faster processing.

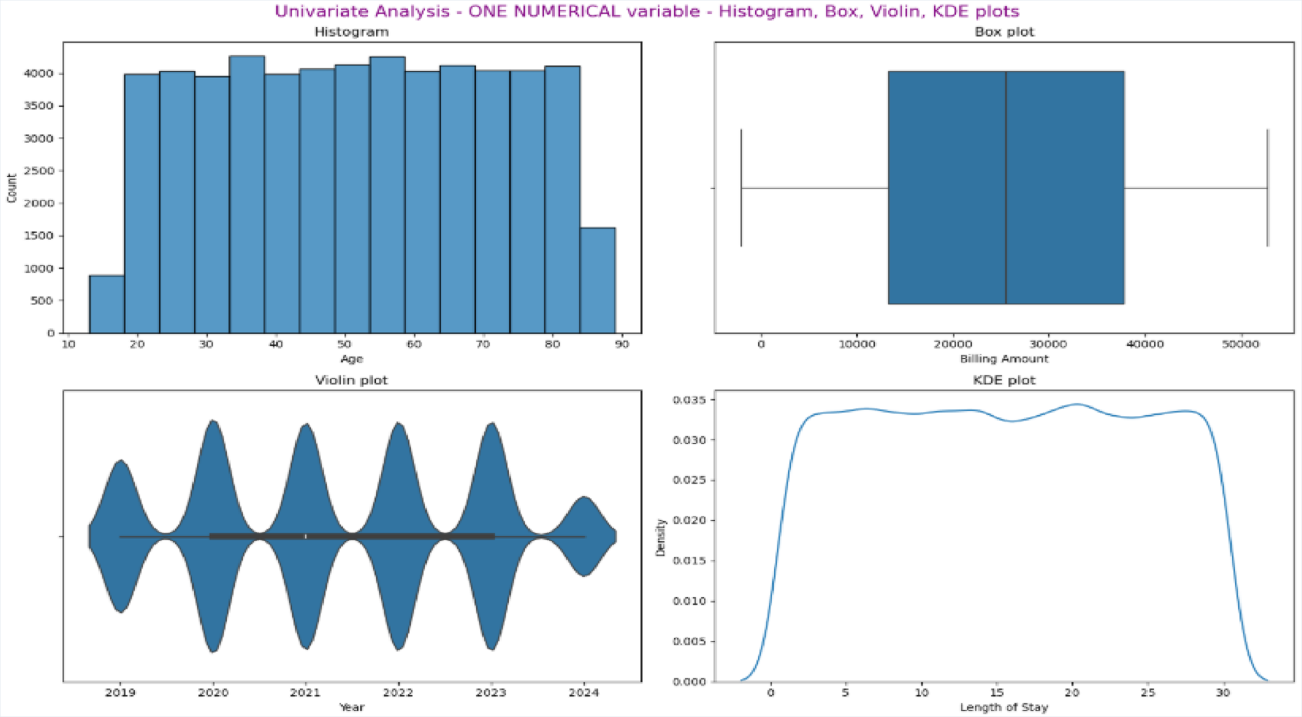
INNOMATICS RESEARCH LABS

Univariate Analysis

Focused on individual features to understand their distribution and patterns.

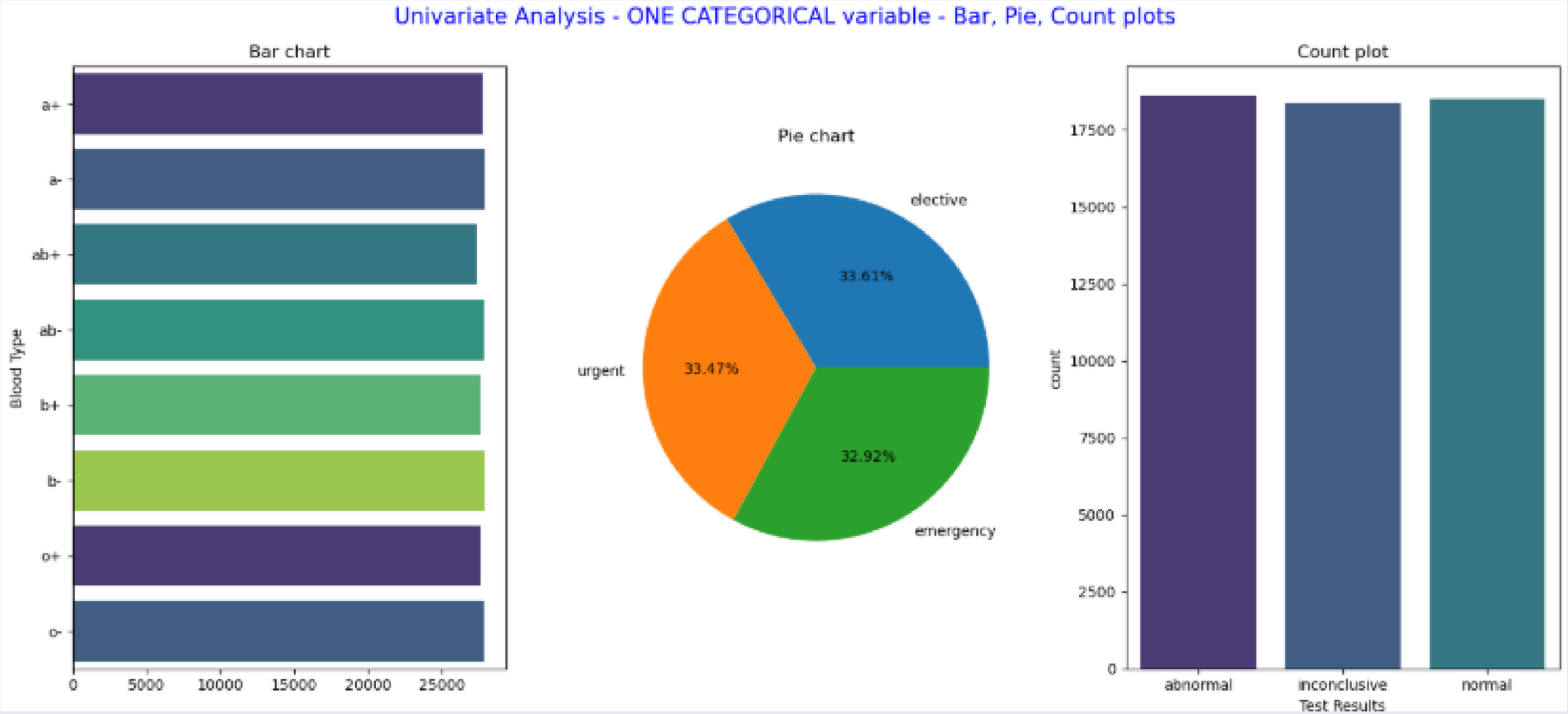
One numerical column

- Numerical Features: *Age*, *Billing Amount*, *Year*, *Length of Stay* showed slightly skewed distributions.
- Detected unrealistic values: Used histograms, boxplots, KDE plots to visualize spread and outliers



One Categorical column

- Blood Type, Admission Type, Test Results analyzed with bar plots, pie charts and count plots

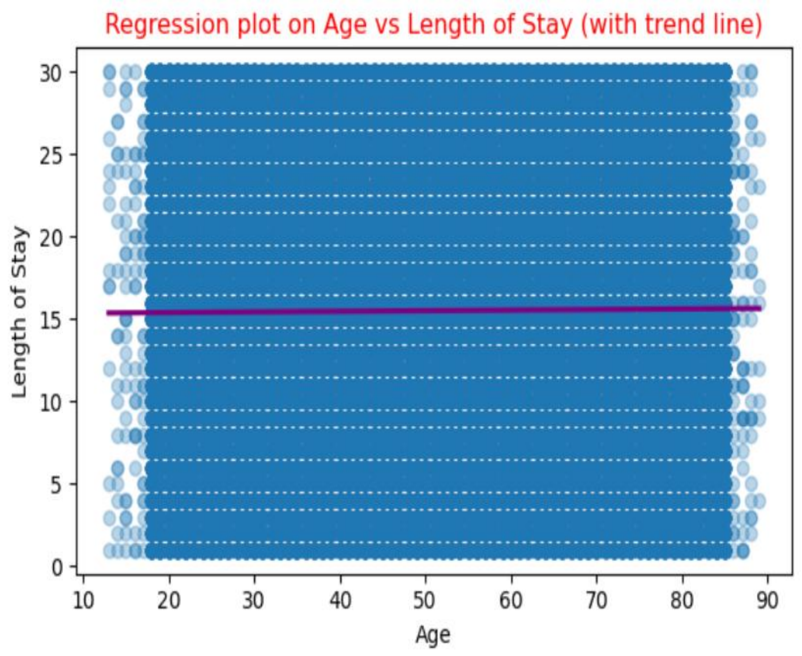


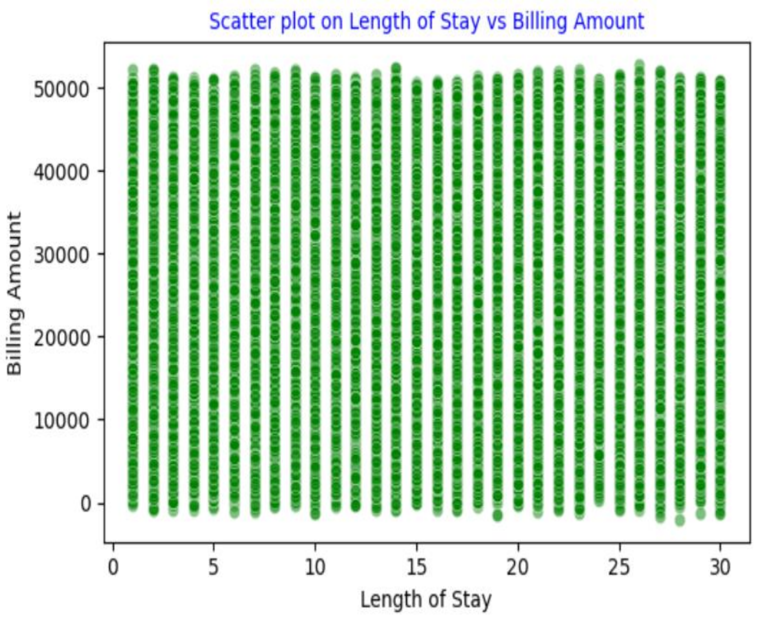
Bivariate Analysis

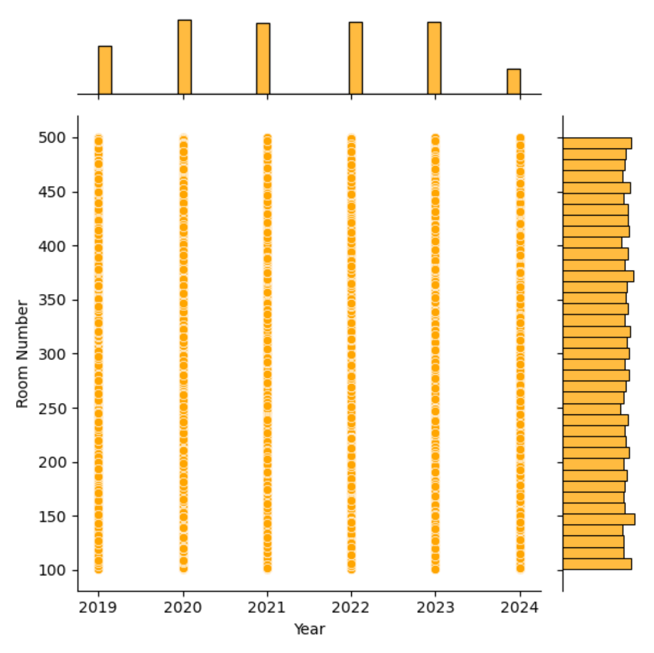
Focused on the relationship between two variables to detect patterns and dependencies..

Two numerical column

Bivariate analysis was performed using scatter, joint, and regression plots to study relationships Between numerical features. Explored relationships between key variables such as Age vs Length of stay , Billing amount vs length of stay and room number.

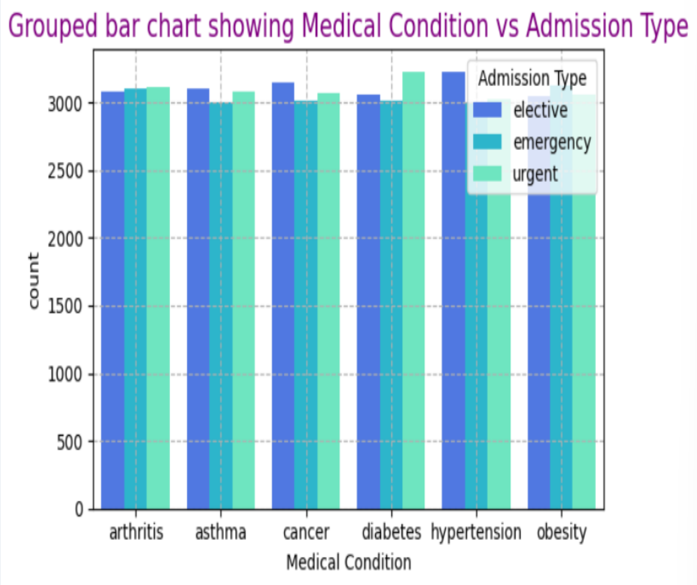


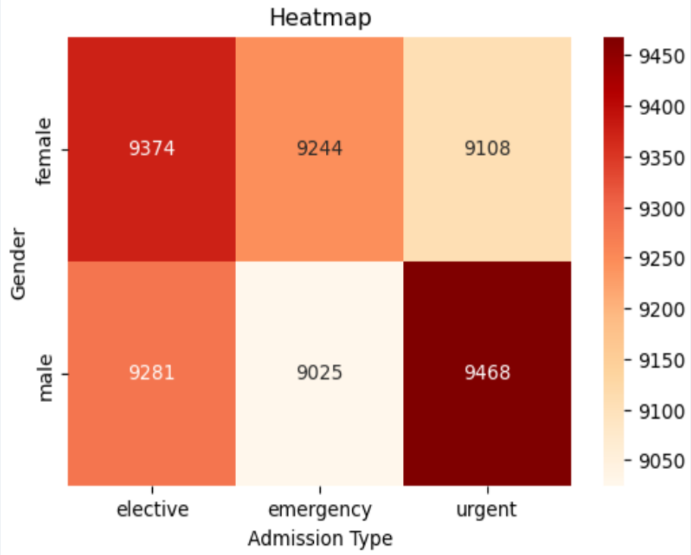


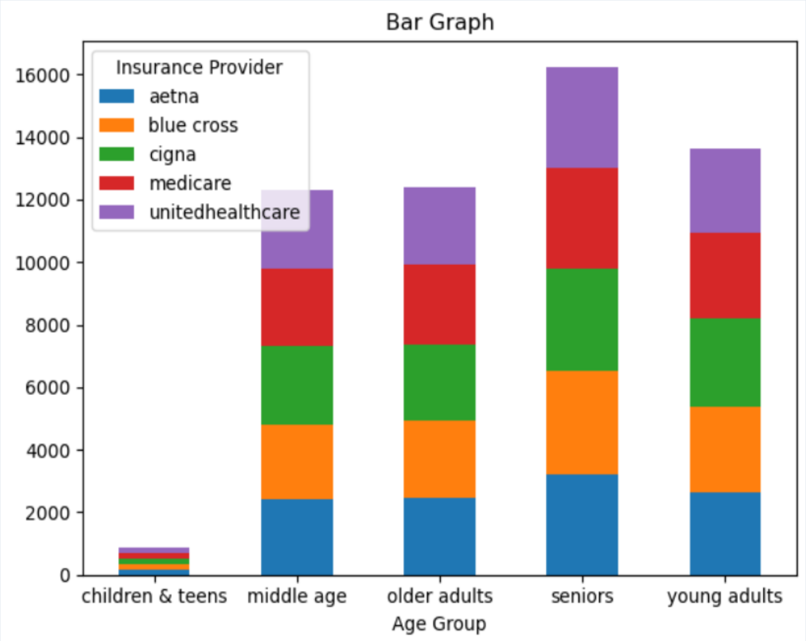


Two categorical column

Categorical analysis was carried out using grouped bar charts, heatmaps, and bar graphs To compare distributions across patient categories. These visualizations showed how Factors like gender , medical condition, Test results , Insurance providers , admission behaviour and health conditions.



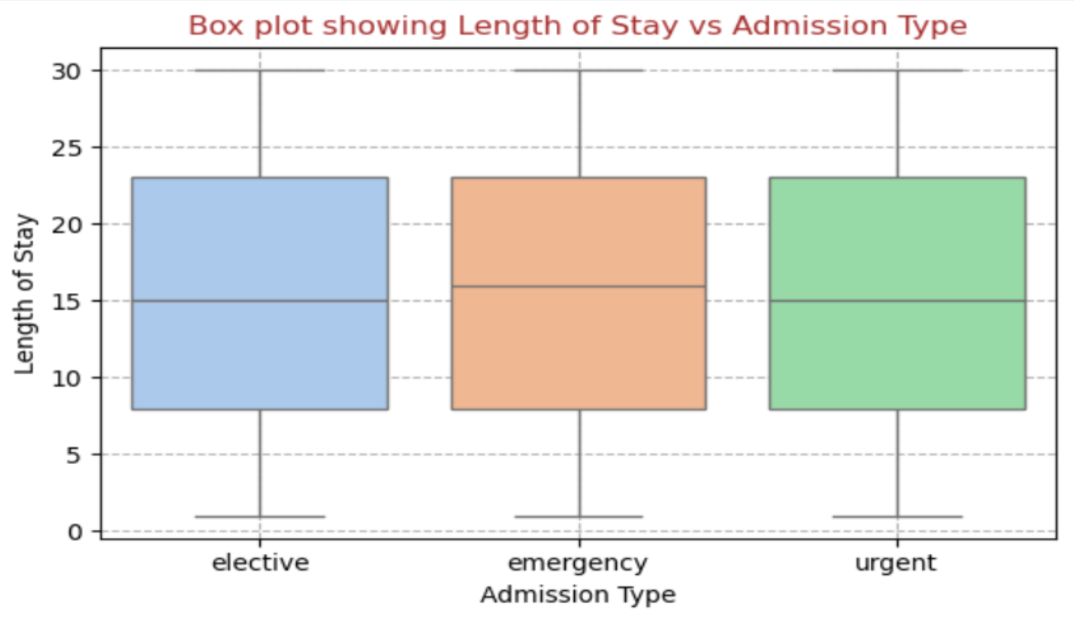


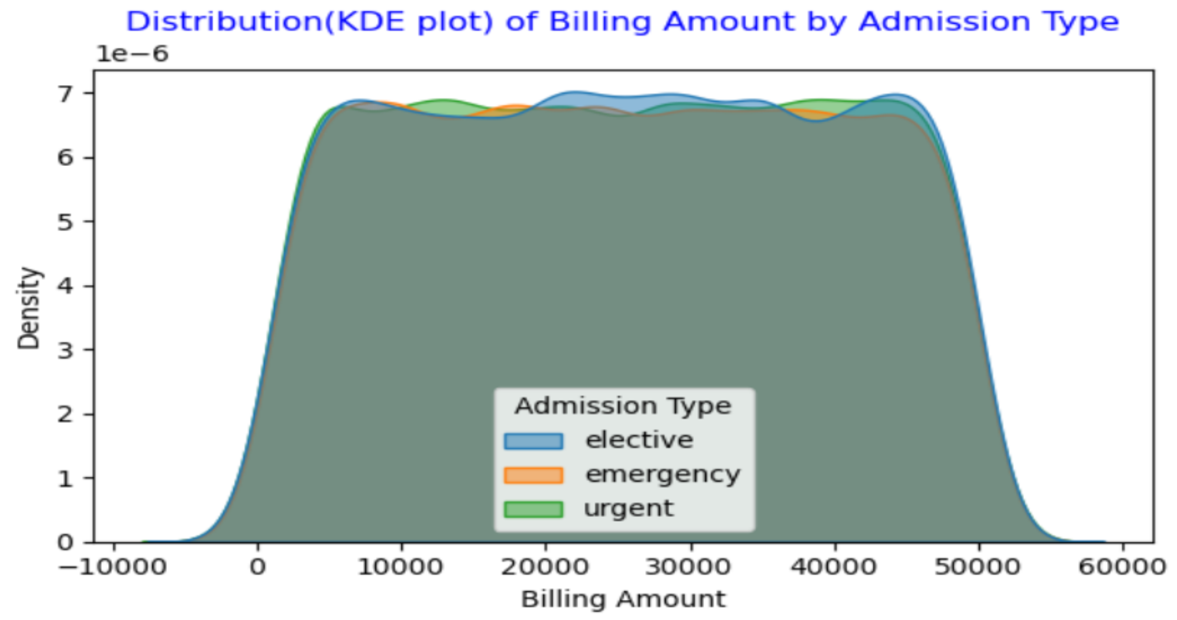


INNOMATICS
RESEARCH LABS

One categorical and one numerical column

Box and KDE plots were used to analyze numerical features across different admission types. These plots showed clear variations in billing amount and length of stay among Emergency, Elective, and Urgent patients. Older patients generally had higher billing, while emergency cases showed longer hospital stays.

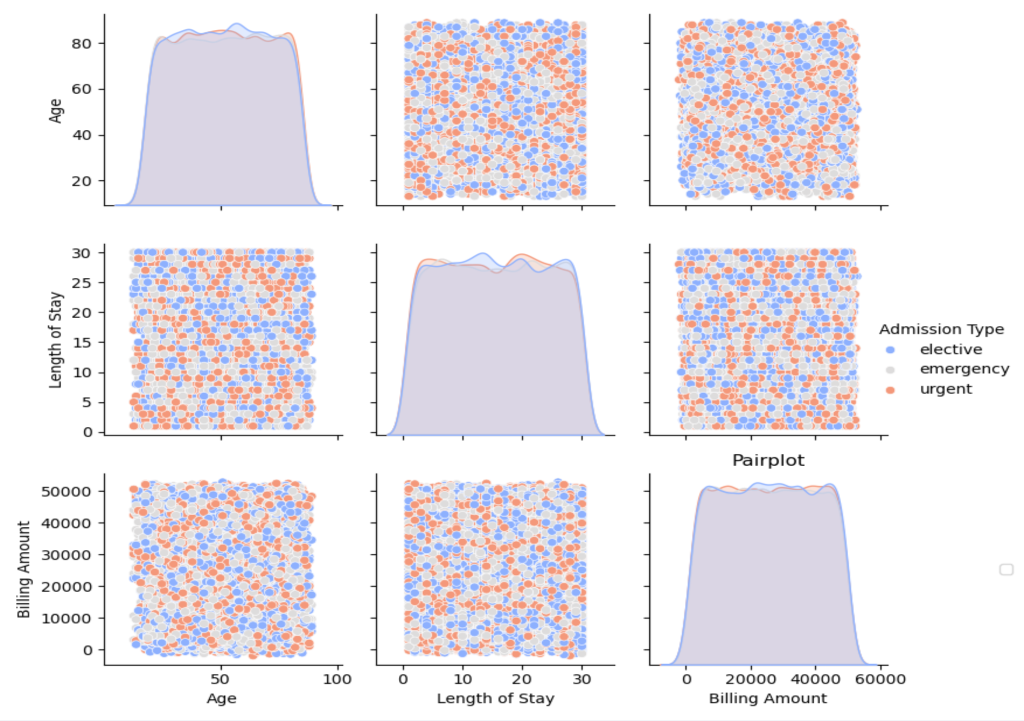


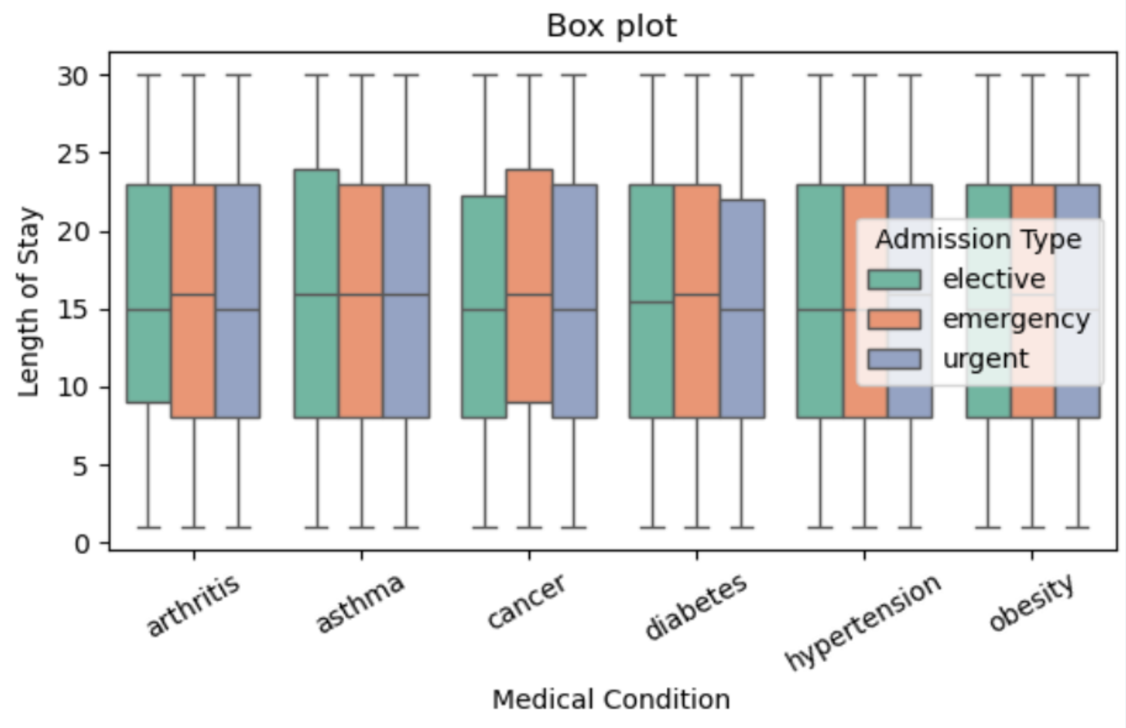


Multivariate Analysis

Focused on the relationship between three or more variables to understand combined effects and complex dependencies..

- Box plots showed that emergency admissions had longer hospital stays across most medical conditions, while routine cases were shorter with less variability.
- Pair plots analysis revealed that longer hospital stays are generally linked with higher billing amounts, while older patients tend to incur slightly more costs.





Insights:

- Cardiology, Orthopaedics , and Neurology dominate patient admissions.
- Emergency admissions have the longest hospital stays.
- Critical conditions (Cardiology, Oncology) show the highest billing amounts.
- Few insurance providers cover most patients, while smaller ones handle costly cases.
- Older patients (65+) stay longer but don't always generate the highest bills.
- Patient distribution varies widely across hospitals
- Admissions fluctuate(changing) across years, with seasonal spikes in certain months
- Some insurance providers cover fewer patients but show higher average bills
- Critical specialties (Cardiology, Orthopaedics) handle the highest patient volumes

Conclusion:

- The dataset provides a comprehensive overview of healthcare records including patients, admissions, and discharges.
- Admission and discharge distribution highlights hospital activity patterns
- Length of stay analysis shows both short-term treatments and extended care cases, reflecting efficiency.
- Age, gender, medical condition, and test results reveal patient characteristics and treatment trends.
- The dataset enables understanding of healthcare trends, hospital operations, and patient care patterns

Future Scope:

- **Predictive Analytics** – Identify disease outbreaks, patient risk factors, and support early diagnosis
- **Smart Healthcare Operations** – Optimize hospital resources (beds, staff, medicines) and improve efficiency
- **Policy & Research Support** – Provide data-driven insights for public health and medical innovation
- **Next-Gen Integration** – Expand with AI, IoT, and wearables for personalized care and real-time monitoring

THANK YOU

INNOMATICS
RESEARCH LABS