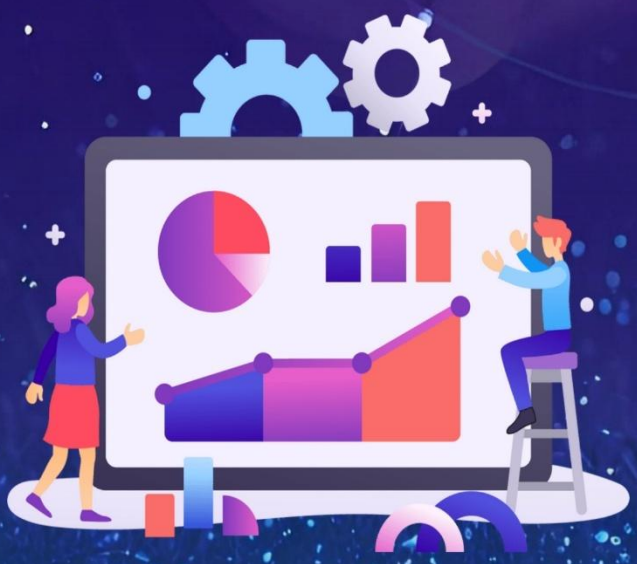
Prodigy InfoTech Internship



Data Science

Note:
It's not compulsory to use the provided dataset

Track: DS

Name: Suriya. B
Linkdin: www.linkedin.com/in/suriya0210
Email: bsuriya223@gmail.com

PRODIGY INFOTECH TASK-1

Create a bar chart or histogram to visualize the distribution of a categorical or continuous variable, such as the distribution of age, gender, rural vs urban, discharged vs Expired, Emergency vs OPD, Monthly Trends of Admissions, Histograms for Key Lab Values, Outcome vs Gender, Outcome vs Rural/Urban, Average Duration of Stay by Gender.

1. Abstract

This project explores a comprehensive hospital dataset spanning **two years** of cardiovascular patient admissions at a tertiary care facility in Ludhiana, India. Using **bar charts** for categorical data and **histograms** for continuous variables, the analysis uncovers demographic distributions, admission patterns, clinical outcomes, and laboratory value trends. These insights help in understanding patient characteristics, optimizing healthcare resources, and identifying priority areas for hospital management.

---

2. Introduction

Modern healthcare generates vast amounts of patient data daily. However, without effective analysis, such data remains underutilized. This project aims to **transform raw hospital admission records into meaningful visual insights**. By focusing on both categorical variables (e.g., gender, locality, admission type) and continuous variables (e.g., age, haemoglobin levels), we identify patterns that may influence patient care strategies and operational planning.

3. About the Dataset

License

This dataset is released under the **Creative Commons Attribution-Non-commercial-Share Alike 4.0 International (CC BY-NC-SA 4.0)** License.

🔗 License Details

Source & Context

The dataset originates from **Hero DMC Heart Institute**, a unit of Dayanand Medical College and Hospital, Ludhiana, Punjab, India. It covers admissions from **1 April 2017 to 31 March 2019**, capturing patient demographics, admission details, comorbidities, laboratory parameters, and outcomes.

Key Statistics:

- **Total Admissions:** 14,845
- **Unique Patients:** 12,238
- **Multiple Admissions:** 1,921 patients

Content

The dataset includes:

- **Demographics:** Age, Gender, Locality (Rural/Urban)
- **Admission Details:** Dates, Type (Emergency/OPD)
- **Medical History:** Smoking, Alcohol, Diabetes Mellitus, Hypertension, CAD, CMP, CKD
- **Lab Parameters:** haemoglobin, TLC, Platelets, Glucose, Urea, Creatinine, BNP, RCE, EF
- **Comorbidities:** Heart Failure, STEMI, Pulmonary Embolism, etc.

- **Shock Classification:** Shock, Cardiogenic Shock, Mixed Shock
- **Outcome:** Discharged or Expired

Reference

Bolle Palli, S.C., et al. *An Optimized Machine Learning Model Accurately Predicts In-Hospital Outcomes at Admission to a Cardiac Unit.* Diagnostics 2022, 12, 241.
🔗 DOI: 10.3390/diagnostics12020241

4. Data Preparation

Before visualization, the dataset was processed to ensure accuracy:

- **Loading & Inspection:** Using Pandas for initial exploration
- **Data Cleaning:** Handling missing values, correcting data types, removing duplicates
- **Date Formatting:** Converting admission/discharge dates for monthly trend analysis
- **Categorical Encoding:** Standardizing category names (e.g., "M" → "Male")
- **Filtering:** Removing irrelevant or extreme outlier entries where necessary

---

5. Visualizations & Findings

5.1 Gender Distribution *(Bar Chart)*

Shows the proportion of male and female patients. Often reflects gender-based health risks in cardiovascular diseases.

5.2 Residence: Rural vs Urban *(Bar Chart)*

Highlights the geographic distribution of patients, important for healthcare outreach planning.

5.3 Outcome: Discharged vs Expired *(Bar Chart)*

Provides a quick overview of recovery vs mortality rates.

5.4 Admission Type: Emergency vs OPD *(Bar Chart)*

Reveals whether patients mostly arrive for emergencies or scheduled visits.

5.5 Monthly Trends of Admissions *(Line/Bar Chart)*

Tracks seasonal or monthly patterns in admissions.

5.6 Age Distribution *(Histogram)*

Shows age groups most affected by cardiac conditions.

5.7 Histograms for Key Lab Values

Examples: haemoglobin, Glucose, Urea — useful for detecting common clinical abnormalities.

5.8 Outcome vs Gender *(Grouped Bar Chart)*

Compares survival and mortality rates between male and female patients.

5.9 Outcome vs Residence *(Grouped Bar Chart)*

Checks if rural or urban patients face different outcome rates.

5.10 Average Duration of Stay by Gender *(Bar Chart)*

Evaluates if hospitalization duration differs between genders.

---

6. Insights & Discussion

From the visual analysis:

- Males form most of admissions in this cardiac unit.
- Rural patients represent a significant proportion, possibly due to regional referral patterns.
- Emergency admissions dominate, indicating acute cardiac conditions.
- Mortality is higher in older age groups, aligning with known cardiac risk factors.
- Certain lab parameters (e.g., low haemoglobin, high creatinine) correlate with worse outcomes.

7. Conclusion

This analysis offers a **clear, data-driven snapshot** of patient demographics, clinical trends, and outcomes in a tertiary cardiac care setting. The insights can guide hospital administrators in **resource allocation, preventive care strategies, and targeted interventions**. Visual analytics has proven to be a powerful tool in making complex healthcare data more understandable and actionable.

---

8. References

- Dataset: Hero DMC Heart Institute – Ludhiana, Punjab, India
- DOI: 10.3390/diagnostics12020241

Importing the library function and csv file into python:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
pd.set_option('display.max_columns', None)

hos_ad = pd.read_csv("C:\\Users\\SURIYA\\Downloads\\HDHI Admission data.csv")
hos_ad
```

Nam	Type	Size	Value
hos_ad	DataFrame	[15757, 56]	Column names: SNO, MRD No., D.O.A, D.O.D, AGE, GENDER, RURAL, TYPE OF ...

Data cleaning:

```
hos_ad.info()
hos_ad.describe()
hos_ad.isna().sum()

num_cols = ['HB', 'TLC', 'PLATELETS', 'GLUCOSE', 'UREA', 'CREATININE', 'EF']
# Convert numeric columns to proper numeric type (float)
for col in num_cols:
    hos_ad[col] = pd.to_numeric(hos_ad[col], errors='coerce')

# Now fill missing values with median
for col in num_cols:
    hos_ad[col].fillna(hos_ad[col].median(), inplace=True)

# Fill BNP with placeholder (-1)
hos_ad['BNP'] = pd.to_numeric(hos_ad['BNP'], errors='coerce')  # ensure BNP is numeric
hos_ad['BNP'].fillna(-1, inplace=True)

# Verify no missing values remain
print(hos_ad.isna().sum())
```

```
In [6]: hos_ad.isna().sum()
Out[6]:
SNO                                  0
MRD No.                              0
D.O.A                                0
D.O.D                                0
AGE                                  0
GENDER                               0
RURAL                                0
TYPE OF ADMISSION-EMERGENCY/OPD      0
month year                           0
DURATION OF STAY                     0
duration of intensive unit stay      0
OUTCOME                              0
SMOKING                              0
ALCOHOL                              0
DM                                   0
HTN                                  0
CAD                                  0
```

```
PRIOR CMP                          0
CKD                                0
HB                                 0
TLC                                0
PLATELETS                          0
GLUCOSE                            0
UREA                               0
CREATININE                         0
BNP                                0
RAISED CARDIAC ENZYMES             0
EF                                 0
SEVERE ANAEMIA                     0
ANAEMIA                            0
STABLE ANGINA                      0
ACS                                0
STEMI                              0
ATYPICAL CHEST PAIN                0
HEART FAILURE                      0
HFREF                              0
```

```
CHB                                0
SSS                                0
AKI                                0
CVA INFRACT                        0
CVA BLEED                          0
AF                                 0
VT                                 0
PSVT                               0
CONGENITAL                         0
UTI                                0
NEURO CARDIOGENIC SYNCOPE          0
ORTHOSTATIC                        0
INFECTIVE ENDOCARDITIS             0
DVT                                0
CARDIOGENIC SHOCK                  0
SHOCK                              0
PULMONARY EMBOLISM                 0
CHEST INFECTION                    0
dtype: int64
```

Interpretation:

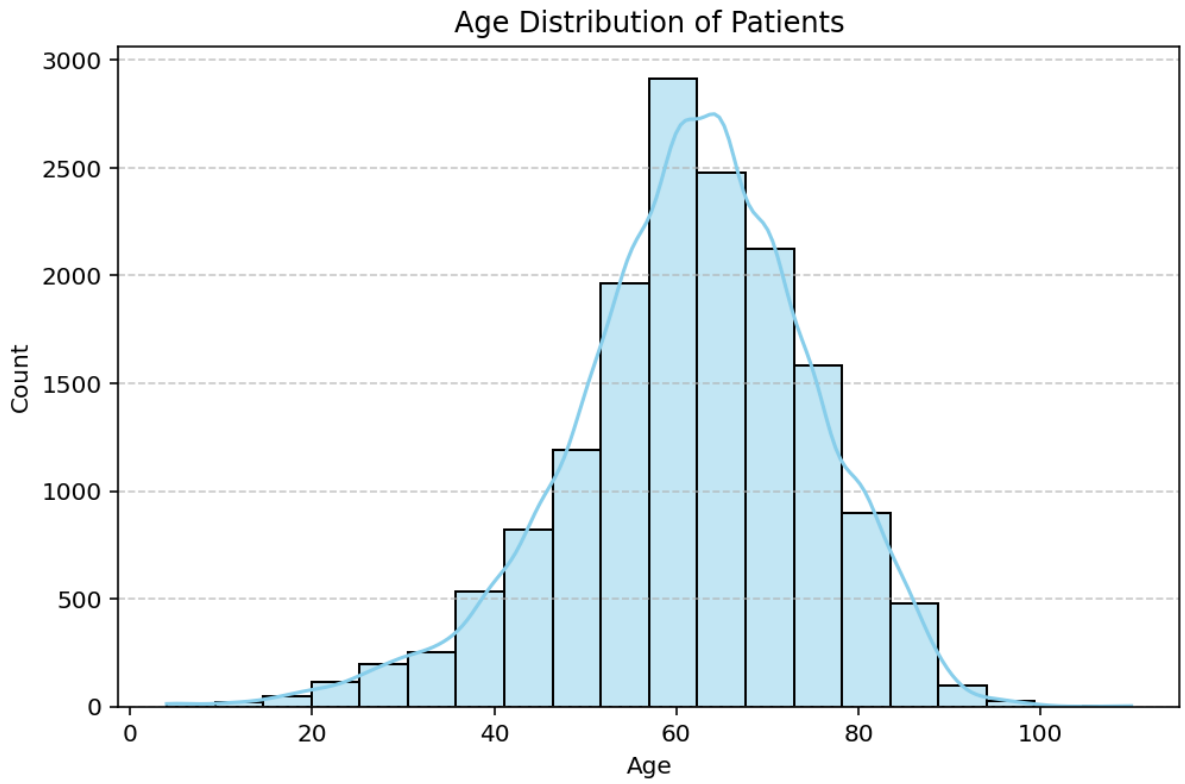
1. **Convert numeric columns to float:**
   For the lab test columns (HB, TLC, PLATELETS, GLUCOSE, UREA, CREATININE, EF), any non-numeric values are coerced into Nan to ensure consistent numeric data types.
2. **Fill missing values with median:**
   For each of these numeric columns, missing values are replaced with the **median** of that column. Median

imputation helps reduce the impact of outliers compared to mean filling.

3. **Special handling for BNP:**
   The BNP column is also converted to numeric, and missing values are replaced with a placeholder value **-1** to indicate "data not available" without confusing it with a real clinical value.

4. **Final check:**
   The code prints the count of missing values per column to confirm that all gaps have been addressed, and all columns are filled, and the missing values are computed with mean.

1.Histogram for Age distribution of patients



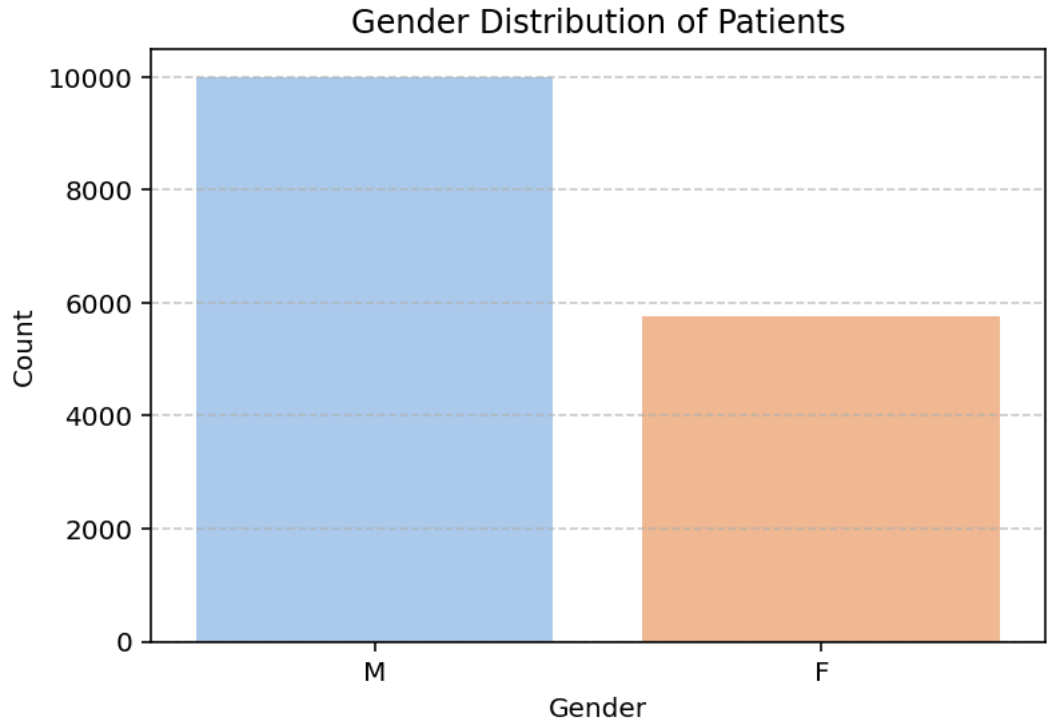
Interpretation:

1. Peak at 60-80 years: Most patients (3000) are aged 60-80, indicating higher healthcare needs in this group.

2. Fewer young patients: Ages 0-40 have under 1000 patients, likely due to lower medical demands.

3. Decline after 80: Patients over 80 are the smallest group (<500), possibly reflecting aging population limits.

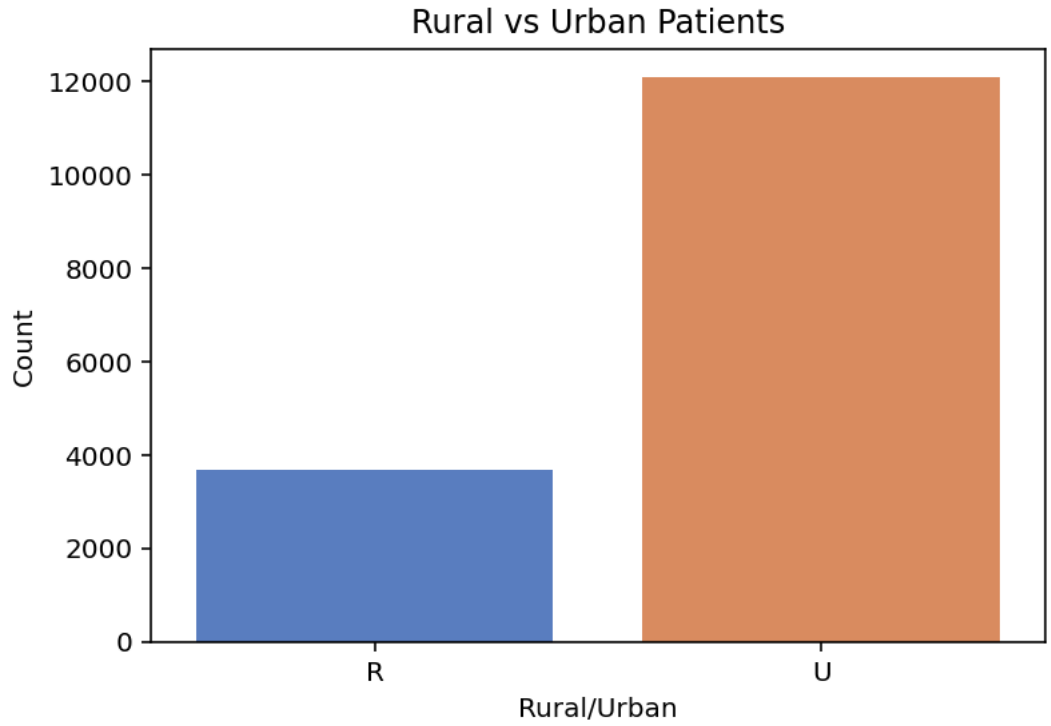
2.Bar plot for gender distribution of patients



Interpretation:

1. Based on the chart, there is a clear gender imbalance in the patient population.
2. Approximately 10,000 male patients and 5,800 female patients are represented.
3. This means that males make up roughly 63% of the patient total, while females constitute about 37%.
4. The total number of patients is around 15,800.
5. The difference in the number of patients between the two genders is approximately 4,200.
6. Therefore, the male patient population is nearly double that of the female patient population, indicating a significant disparity.

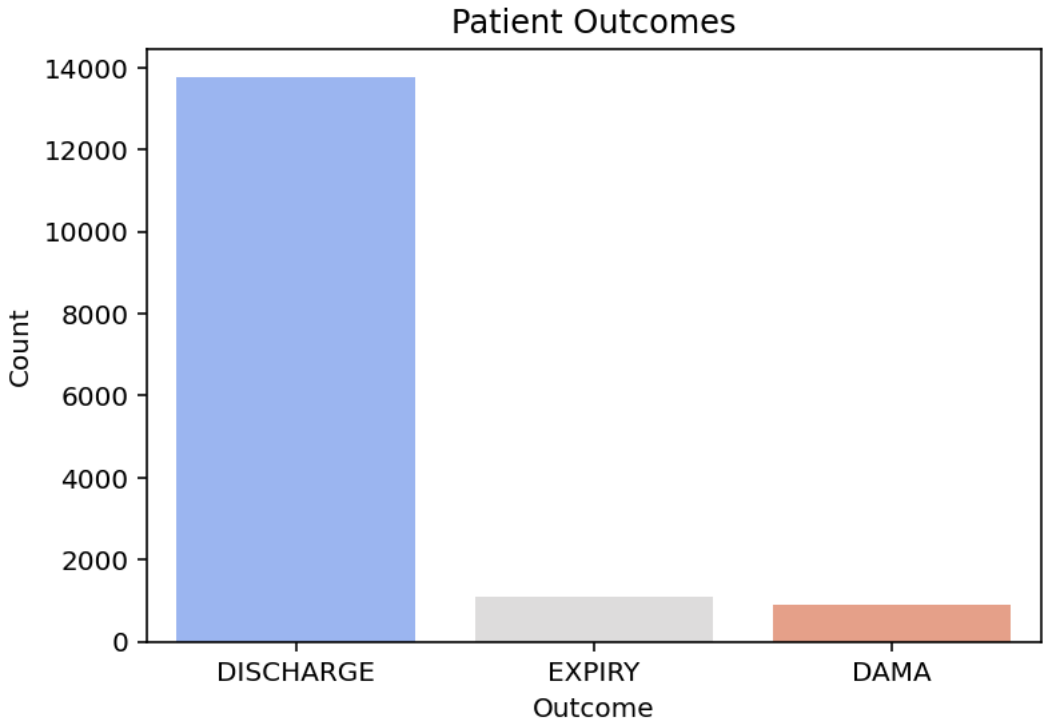
3.Bar chart for Rural vs Urban



Interpretation:

1. The chart compares the number of patients from rural and urban areas.
2. The blue bar labelled "R" represents rural patients, and its height indicates a count of approximately 3,800.
3. The orange bar labelled "U" represents urban patients, and its height indicates a count of 12,000.
4. The data shows a significant difference in the number of patients, with the urban patient population being much larger than the rural patient population. The number of urban patients is more than three times the number of rural patients.

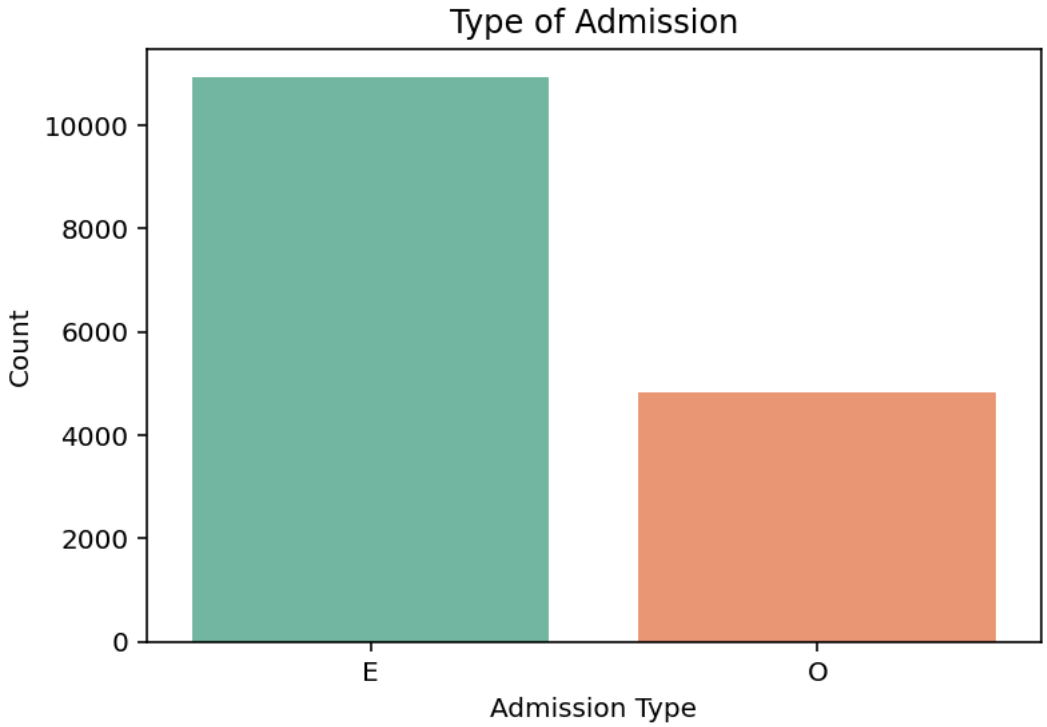
4.Bar chart for Outcome (e.g., Discharged vs Expired)



Interpretation:

1. The chart shows the final status of patients, categorized into three outcomes. The overwhelming majority of patients were discharged, with a count of approximately 13,800.
2. In contrast, a much smaller number of patients had the other two outcomes. There were roughly 1,200 expiry cases (patients who passed away) and about 900 cases of DAMA (discharged against medical advice). This clearly indicates that most patients successfully complete their treatment and are released from care.

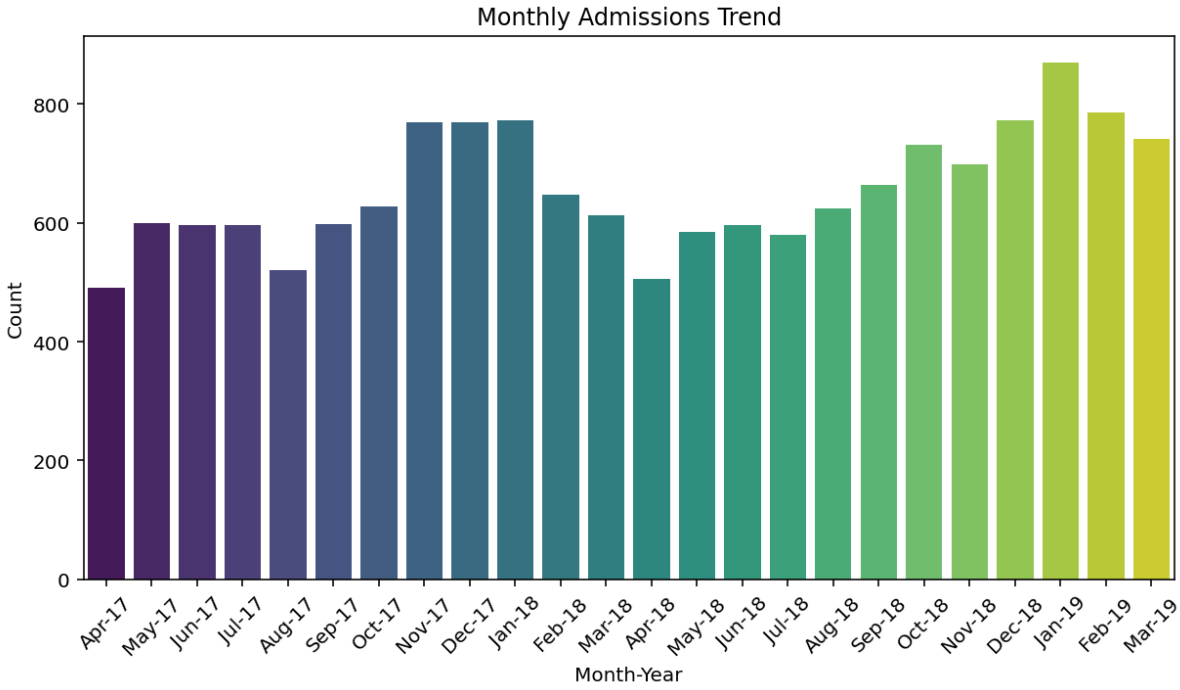
5.Bar chart for Type of Admission (Emergency vs OPD)



Interpretation:

The chart compares two types of patient admissions, labelled as "E" (Emergency) and "O" (Outpatient). The bar for E is significantly taller, showing a count of approximately 10,800. The bar for O is much shorter, with a count of around 4,800. This indicates that more than twice as many patients were admitted with an "E" type of admission compared to an "O" type of admission.

6. Monthly Trends of Admissions

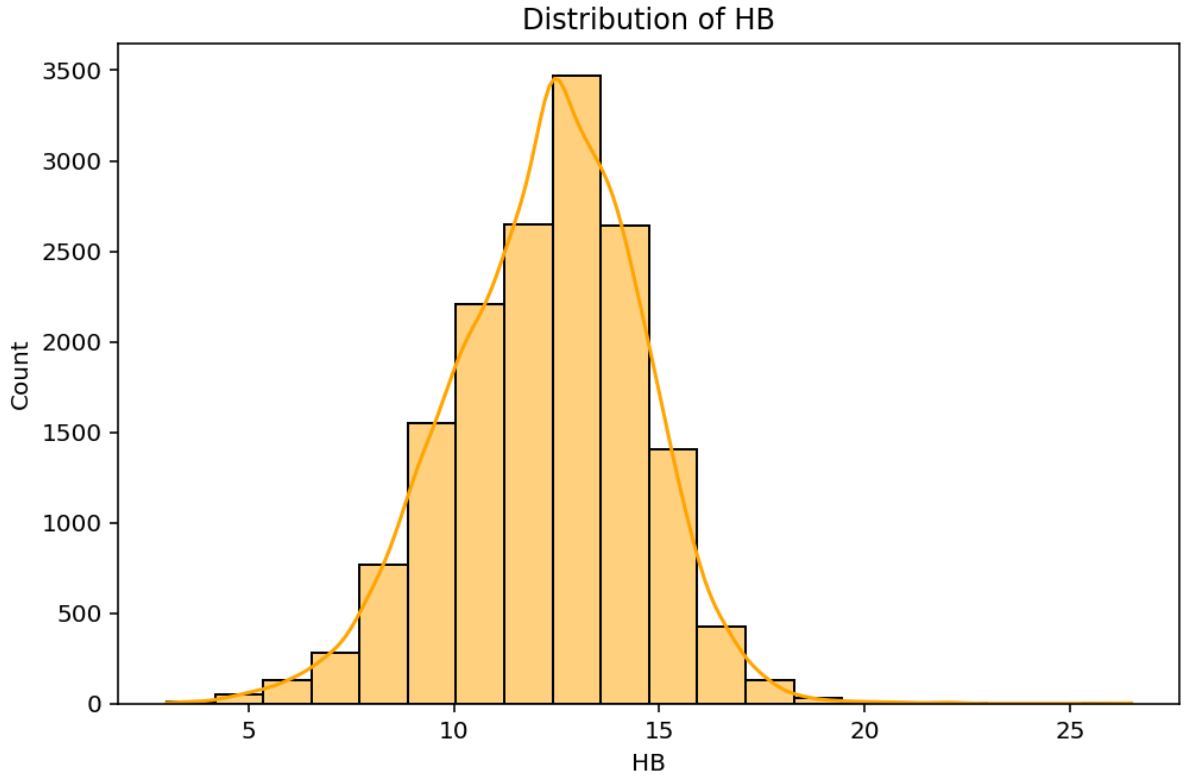


Interpretation:

The chart shows the number of admissions per month from April 2017 to March 2019. The overall trend appears to be cyclical and slightly increasing over time. The lowest admission counts are around April 2017 and April 2018, at approximately 500. The highest admission counts occur towards the end of the time, with a peak in January 2019 at over 850. There's a notable dip in admissions in April of both 2017 and 2018, followed by a general increase throughout the rest of each year. The admissions seem to peak during the winter months of late 2017 and late 2018/early 2019.

7. Histograms for Key Lab Values

Distribution of HB:



Interpretation:

Based on the histogram titled "Distribution of HB," the data appears to be approximately normally distributed. The graph shows a classic bell-shaped curve, with most of the data points clustered around the central value.

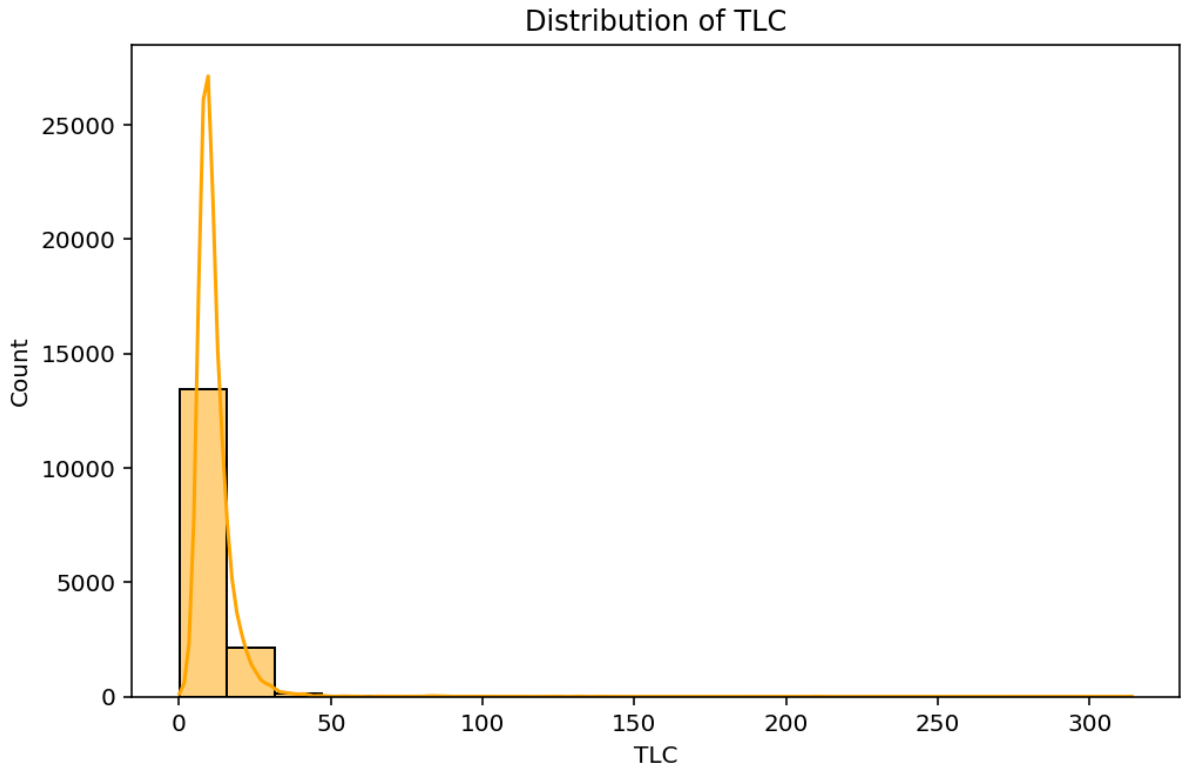
Key Observations

- Central Tendency: The highest frequency of "HB" values is concentrated between 12 and 14, which represents the mean or mode of the distribution.
- Symmetry: The histogram is roughly symmetrical. The frequency of values decreases as they move away from the central peak in both directions, forming a tail on the left and right.

- Range: The "HB" values range from less than 5 to a little over 20, with a few outliers.
- Peak: The peak of the distribution is at an "HB" value of approximately 13.

This shape is also reinforced by the superimposed orange curve, which is a kernel density estimate (KDE) and closely follows the outline of the histogram bars. This visual confirms that the data for "HB" (Haemoglobin) is distributed in a manner consistent with a normal distribution.

Distribution of TLC:



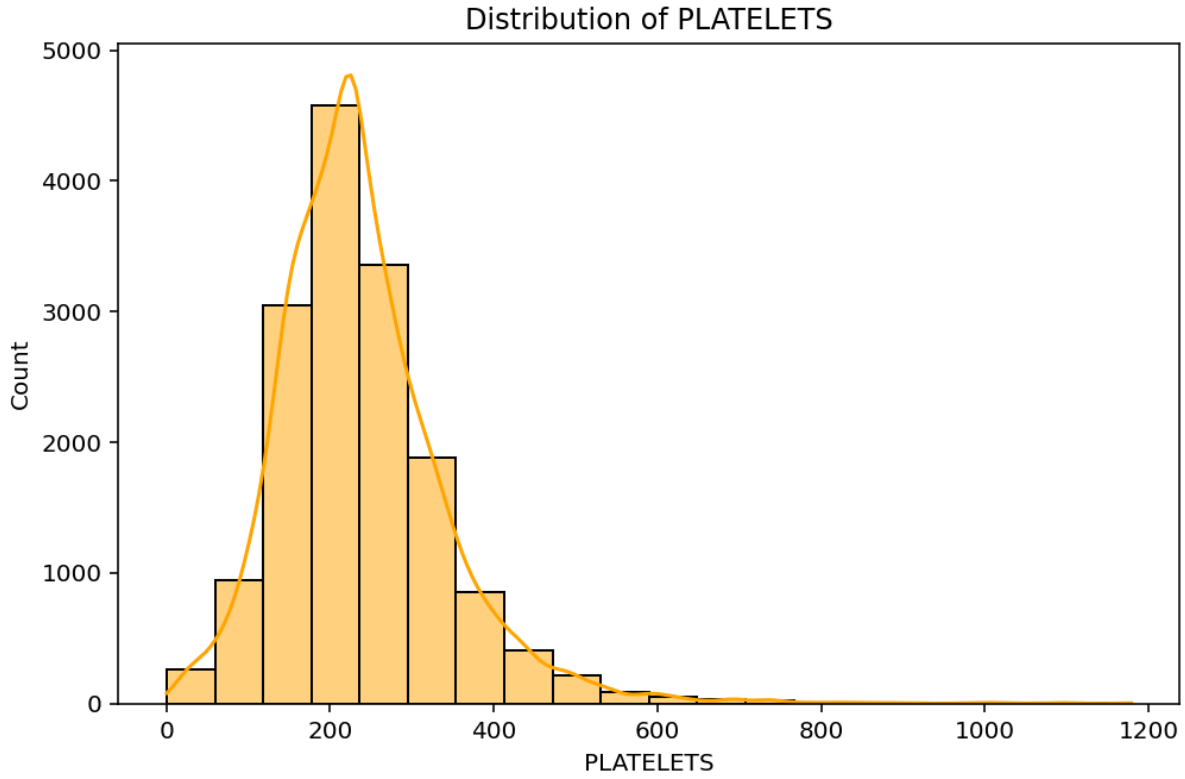
Interpretation:

Based on the histogram titled "Distribution of TLC," which stands for Total Leucocyte Count, the data is highly skewed to the right. This means most patients have a low Total Leucocyte Count, while a small number of patients have a significantly high count.

Key Observations

- Central Tendency: The highest frequency of patient counts for Total Leucocyte Count is heavily concentrated in the low-value range, specifically between 0 and 10.
- Skewness: The distribution is not symmetrical. It has a long tail extending to the right, indicating a few patients with very large values. This is characteristic of a positively skewed distribution.
- Outliers: The presence of a few data points with values extending past 50, and even up to 300, suggests some patients have extremely high Total Leucocyte Counts. The orange kernel density estimate curve visually reinforces this strong skew and the concentration of data at the lower end.

Distribution of Platelets:



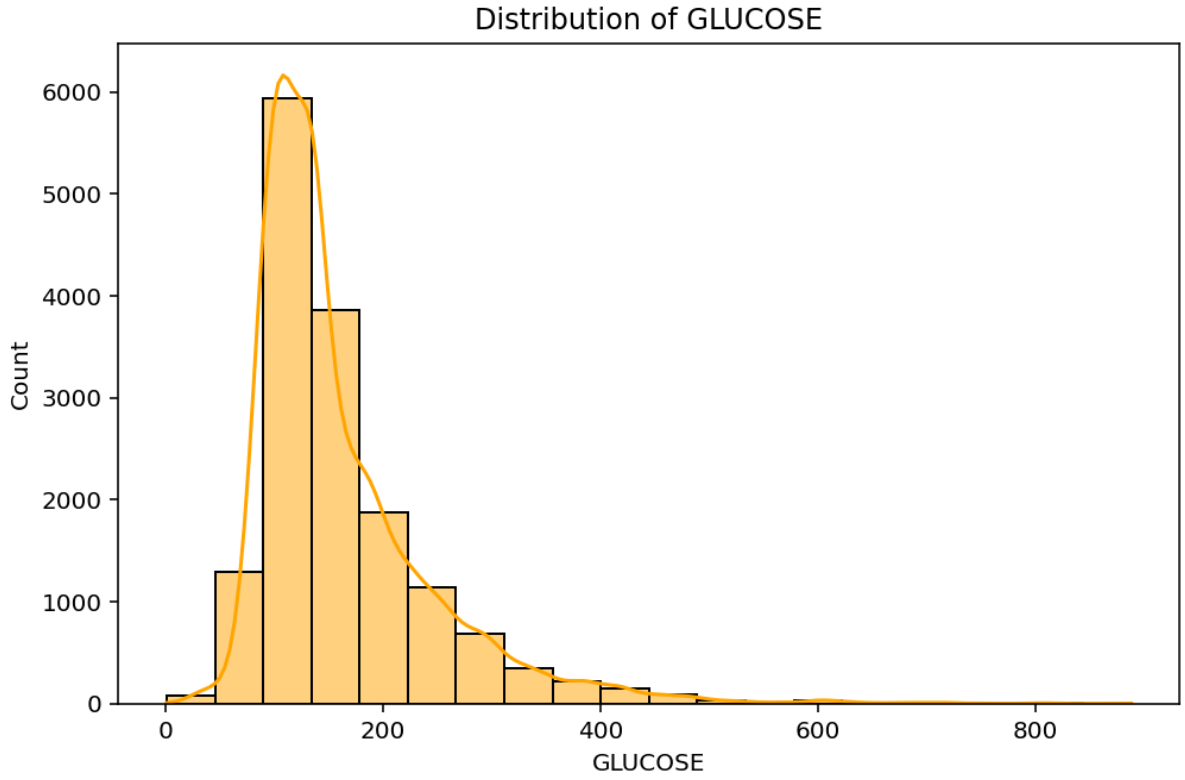
Interpretation:

Based on the histogram titled "Distribution of PLATELETS," the data is **skewed to the right**. The distribution has a clear peak, and a long tail extends towards the higher values.

Key Observations

- **Central Tendency:** Most patients have a platelet count clustered between 100 and 300, with the most frequent count occurring between 200 and 250.
- **Skewness:** The data is not symmetrical. The distribution has a long tail on the right side, which indicates that while most patients fall within a normal range, a smaller number have significantly higher platelet counts. This is characteristic of a **positively skewed** distribution.
- **Range:** While most counts are below 400, a few data points extend beyond 800 and up to 1200, representing a small number of patients with extremely high platelet levels.

Distribution of Glucose:



Interpretation:

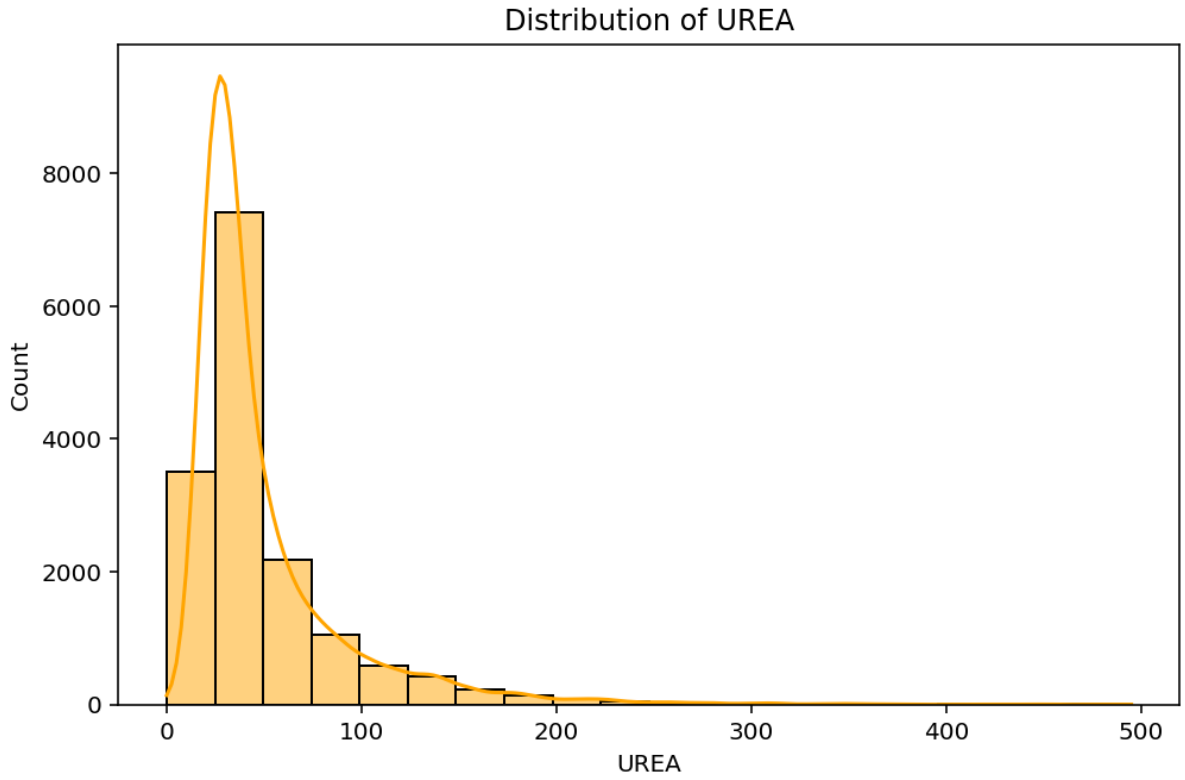
Based on the histogram titled "Distribution of GLUCOSE," the data is **highly skewed to the right**. Most of the values are concentrated at the lower end of the scale, with a long tail extending toward higher glucose values.

Key Observations

- **Central Tendency:** Most patients have glucose levels clustered between 100 and 150, with a prominent peak in the range of 100-125.
- **Skewness:** The distribution is not symmetrical. The long tail on the right side indicates that while most patients have glucose levels in a relatively narrow, low range, a smaller number of patients have significantly higher glucose levels. This is a classic example of a **positively skewed** distribution.

- **Range:** While the bulk of the data lies below 200, the distribution extends out to values over 800, representing a small portion of the patient population with very high glucose levels.

Distribution of Urea:



Interpretation:

Based on the histogram titled "Distribution of UREA," the data is highly skewed to the right. Most UREA values are concentrated at the low end of the scale, with a long tail of decreasing frequency extending toward much higher values.

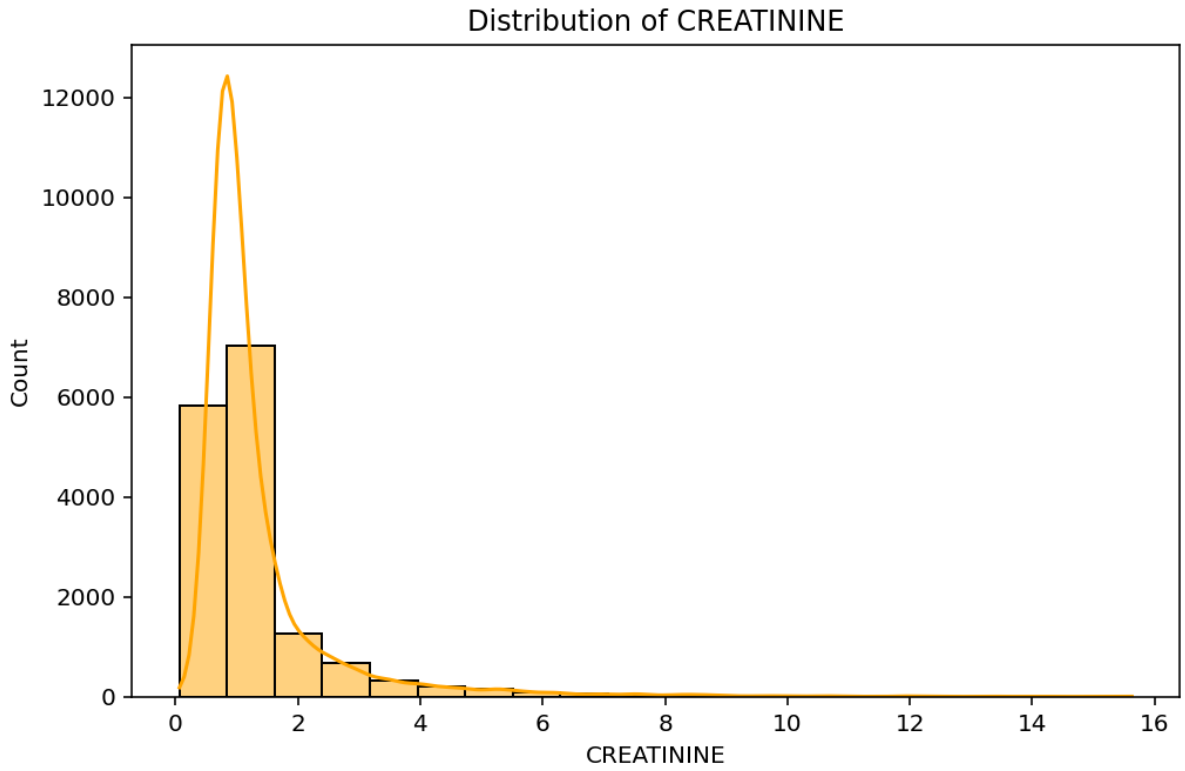
Key Observations

- Central Tendency: The most frequent UREA values are found in the first two bins, specifically between 0 and 50, where the count is highest.
- Skewness: The distribution is not symmetrical. It has a long tail on the right side, which is a key characteristic of a

positively skewed distribution. This indicates that while most patients have low UREA levels, a small number have significantly high levels.

- Range: The bulk of the data is below 100, but the distribution shows that some values extend beyond 400, representing a small portion of the patient population with extremely high UREA levels.

Distribution of Creatine:



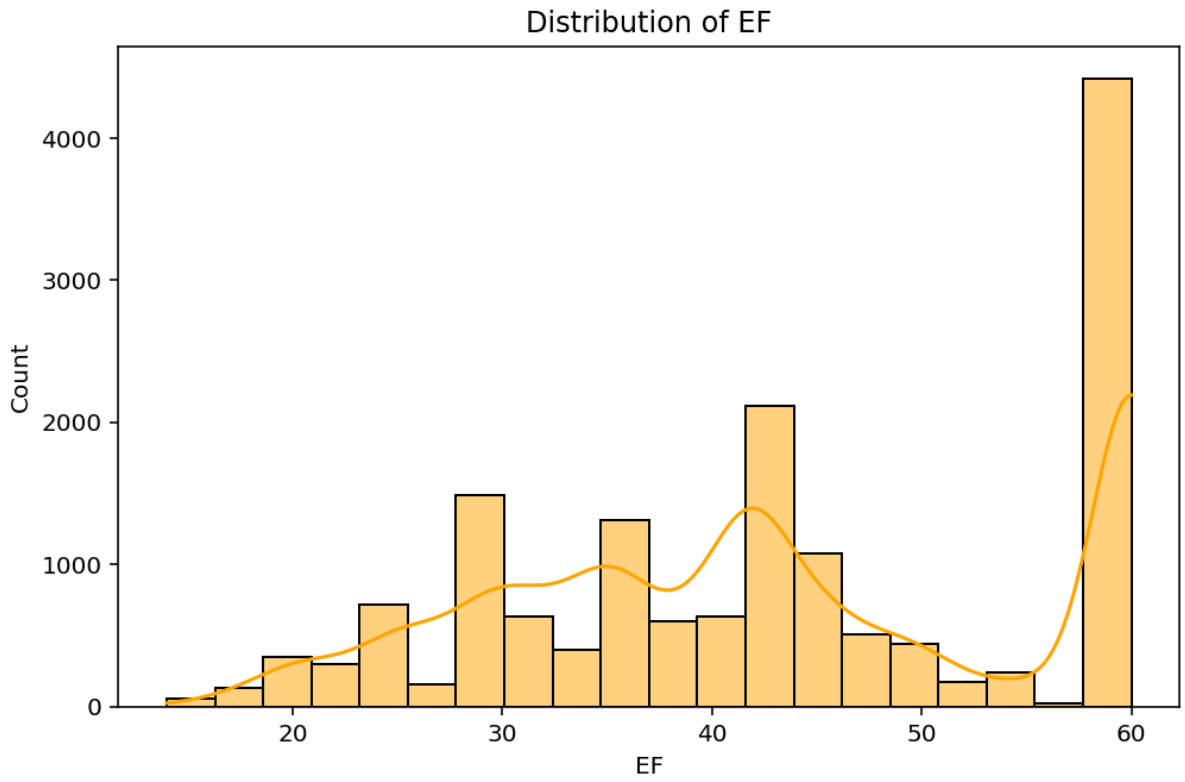
Interpretation:

Based on the histogram titled "Distribution of CREATININE," the data is highly skewed to the right. Most of the values are concentrated at the low end of the scale, with a long tail extending toward higher creatinine values.

Key Observations

- Central Tendency: The majority of patients have creatinine values clustered between 0 and 2, with the most frequent count occurring in the range of 0 to 1.
- Skewness: The distribution is not symmetrical. The long tail on the right side is a key characteristic of a positively skewed distribution. This indicates that while most patients have low creatinine levels, a small number have significantly high levels.
- Range: While the bulk of the data is below 4, the distribution shows that a few values extend beyond 10 and even up to 16, representing a small portion of the patient population with extremely high creatinine levels.

Distribution of EF:



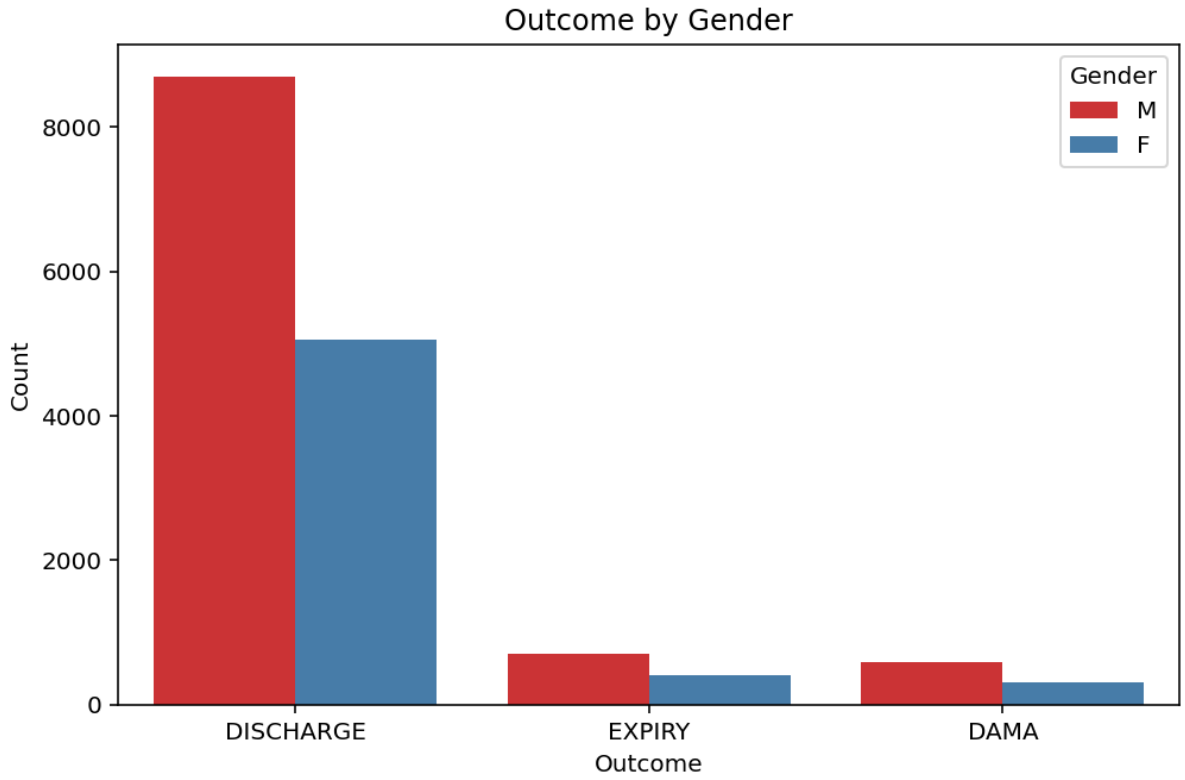
Interpretation:

Based on the histogram titled "Distribution of EF" (Ejection Factor), the data does not follow a single normal distribution and appears to be multimodal. The data is heavily skewed and clustered at the upper end of the scale.

Key Observations

- Multimodal Distribution: The histogram shows several peaks, suggesting that there are multiple distinct groups or populations within the data.
- Most Common Values: The highest concentration of EF values occurs at 60, with a count of over 4000. This is the dominant mode.
- Other Peaks: There are other smaller, but noticeable, peaks around 28-30, and again around 42-43, indicating secondary modes in the distribution.
- Skewness: The distribution is highly skewed, with a long, spread-out left tail and a sharp spike at the far-right end.
- Range: The values of EF range from below 15 to a distinct peak at 60.

8. Comparative Bar Plot: Outcome vs Gender



Interpretation:

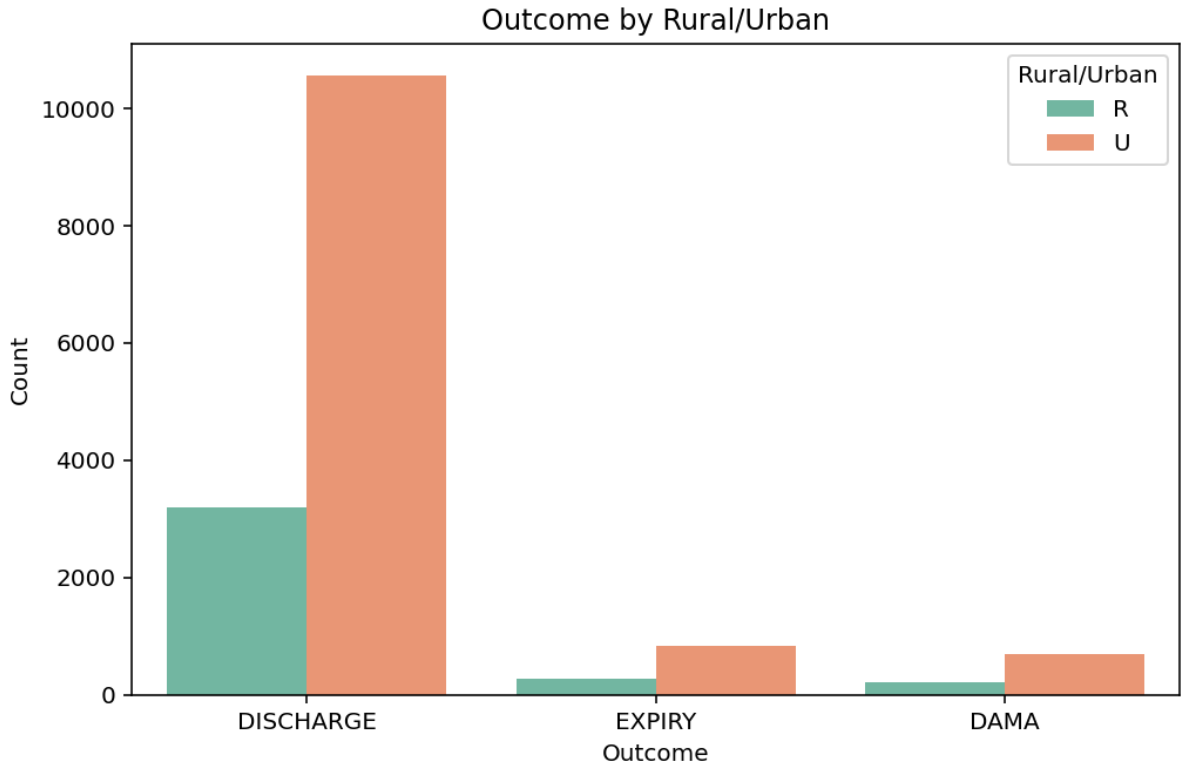
Based on the "Outcome by Gender" bar chart, the data shows that **males have a higher count for every outcome category** (Discharge, Expiry, and DAMA) compared to females.

Key Observations

- **Overall Disparity:** There's a significant difference in the total number of males versus female patients, with males outnumbering females. This is consistent across all outcomes.
- **Discharge:** Most patients for both genders were discharged. However, the number of males discharged (over 8,000) is substantially higher than the number of females discharged (around 5,000).
- **Expiry and DAMA:** While the counts for both expiry and DAMA (discharged against medical advice) are much lower

than for discharge, the pattern remains the same. A higher number of males passed away and a higher number of males left against medical advice compared to females.

9. Comparative Bar Plot: Outcome vs Rural/Urban



Interpretation:

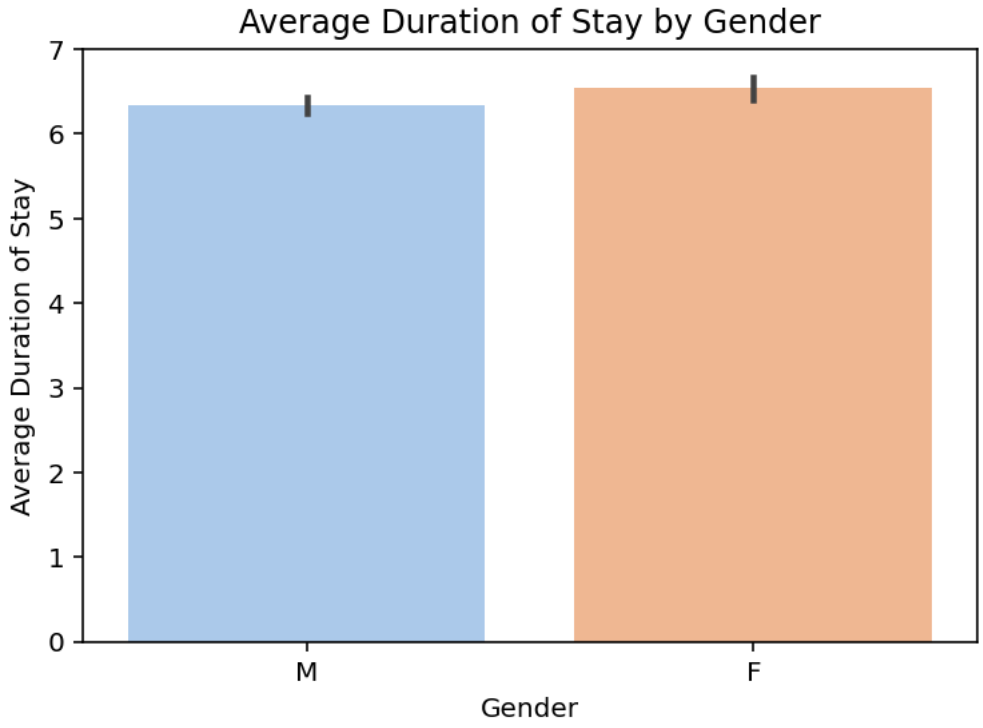
Based on the "Outcome by Rural/Urban" bar chart, the data shows that urban patients have a higher count for all outcome categories (Discharge, Expiry, and DAMA) compared to rural patients.

Key Observations

- Overall Disparity: There's a notable difference in the total number of urban versus rural patients, with urban patients outnumbering rural patients. This is consistent across all outcomes.

- Discharge: Most patients for both groups were discharged. However, the number of urban patients discharged (over 10,000) is substantially higher than the number of rural patients discharged (around 3,200).
- Expiry and DAMA: While the counts for both expiry and DAMA (discharged against medical advice) are much lower than for discharge, the pattern remains the same. A higher number of urban patients passed away and a higher number of urban patients left against medical advice compared to rural patients.

10. Average Duration of Stay by Gender (Bar Plot)



Interpretation:

Based on the bar chart titled "Average Duration of Stay by Gender," the average duration of stay for both male and female patients is very similar.

Key Observations

- The average duration of stay for male patients (M) is approximately 6.3 days.
- The average duration of stay for female patients (F) is approximately 6.6 days.
- The difference between the two genders is minimal, suggesting that a patient's gender does not have a significant impact on their length of hospital stay.
- The small black lines on top of each bar represent the confidence interval, which visually indicates the range of probable values for the average. The overlap of these intervals further reinforces that there is no statistically significant difference in the average stay between genders.

Conclusion:

In conclusion, the dataset provides a comprehensive view of patient demographics, admission patterns, clinical outcomes, and lab results, enabling both categorical and continuous data visualizations. These insights help identify trends, compare groups, and support data-driven healthcare decisions.