Table of Contents

1. Introduction 1

2. Data Preparation 2

   2.1 Data Loading and Structure 2

   2.2 Missing Values and Data Cleaning 3

   2.3 Data Types 3

3. Exploratory Data Analysis (EDA) 4

   3.1 Continuous Variable Distributions 4

   – Age Distribution .....

   – Length of Stay (LOS) .....

   3.2 Categorical Variables Analysis 5

   3.3 Correlation Analysis 6

4. Group-wise Comparative Analysis 7

   4.1 Weekend vs Weekday Admissions 7

   4.2 Teaching vs Non-Teaching Hospitals 8

5. Key Initial Findings 9

6. Conclusion 10

Independent Study Report

Exploratory Data Analysis of Weekend Effect on Hospital Mortality and Patient Characteristics

1. Introduction

Hospital performance and patient outcomes have long been subjects of interest in health services research. One phenomenon frequently discussed is the "weekend effect," where patients admitted to hospitals over weekends have reportedly worse outcomes compared to those admitted during weekdays.
This variation could be due to factors such as reduced staffing, limited access to specialized procedures, and differences in patient profiles.

In this study, we perform a detailed exploratory analysis of a large hospital dataset (~8 million records) to understand:

- How patient characteristics vary between weekend and weekday admissions
- Differences between teaching and non-teaching hospitals
- How initial mortality rates differ without adjustment
- What early insights suggest about potential system inefficiencies

This EDA lays the foundation for the more formal statistical modeling performed in the second phase of the project.

2. Data Preparation

2.1 Data Loading and Structure

The dataset was imported using pandas, and initial checks were performed to ensure that all columns were properly read.

```
In [3]: df = pd.read_csv(r"C:\Users\singh\Downloads\Updated_file_v1.csv")

In [4]: pd.set_option('display.max_columns', None)
        df.head(5)
```

Out[4]:

	AGE	AWEEKEND	AMONTH	DIED	DRG	ELECTIVE	FEMALE	HCUP_ED	HOSP_DIVISION_x	HOSP_NIS_x	KEY_NIS	LOS	PAY1	RACE	TOTCHG	YE
0	87.0	1	6.0	0	391	0	1	2	1	10001	10002081	2	1	1	12425	2(
1	66.0	0	7.0	0	638	0	0	2	1	10001	10003421	4	1	1	19189	2(
2	83.0	0	4.0	0	291	0	0	2	1	10001	10003995	5	1	1	21433	2(
3	58.0	0	2.0	0	623	0	0	2	1	10001	10004372	4	2	2	25219	2(
4	62.0	0	1.0	0	291	0	0	2	1	10001	10004529	1	1	1	9060	2(

The dataset included demographic details, hospital attributes, clinical variables, and outcomes. Important features included age, gender, race, insurance type, whether the admission occurred over a weekend, hospital teaching status, severity scores, and in-hospital mortality.

2.2 Missing Values and Data Cleaning

A comprehensive review of missing values was conducted across all columns prior to formal analysis. Encouragingly, the dataset demonstrated a **very low proportion of missingness overall**, which minimized concerns about systematic bias.
Key variables such as **ZIP income quartile** and **clinical severity scores** (APRDRG Risk Mortality and Severity of Illness) had **minimal missing entries**, typically limited to **one or two cases per field**. Given the scale of the dataset comprising over eight million records these sparse missing values were deemed inconsequential in affecting broader statistical patterns. To maintain the robustness of subsequent comparisons and modeling, rows with **critical missing data** in essential fields were systematically **excluded**. This approach ensured that all analytical steps, especially multivariate modeling, operated on complete cases without introducing noise through imputation or interpolation methods.
After cleaning, the final dataset comprised **8,096,311 complete hospital admission records**, fully ready for exploration and modeling. No artificial imputation techniques were applied, preserving the dataset's original integrity and reducing potential risks of introducing bias through estimated values.

2.3 Data Types

As part of the initial data preparation phase, **data types** for each column were carefully inspected to ensure accurate downstream analysis.
It was essential to verify that **numerical variables** such as **age**, **length of stay**, **total hospital charges**, and **severity scores** were correctly recognized as **continuous numerical data**, enabling appropriate aggregation, summary statistics, and regression modeling. Similarly, **categorical variables** including **gender**, **race**, **insurance type (PAY1)**, **weekend admission indicator (AWEEKEND)**, and **hospital teaching status (HOSP_LOCTEACH)** were explicitly confirmed to be treated as **categorical (object or integer-coded categorical)**, allowing for meaningful group-based comparisons, chi-square testing, and dummy encoding where needed for modeling purposes.
This meticulous checking ensured that data visualization tools like seaborn and matplotlib, as well as statistical libraries such as statsmodels and scikit-learn, handled each variable appropriately without misinterpretation.
Correct classification of data types also prevented silent computational errors during model fitting and improved the interpretability of exploratory plots, histograms, countplots, and logistic regression outputs.

Out[7]:

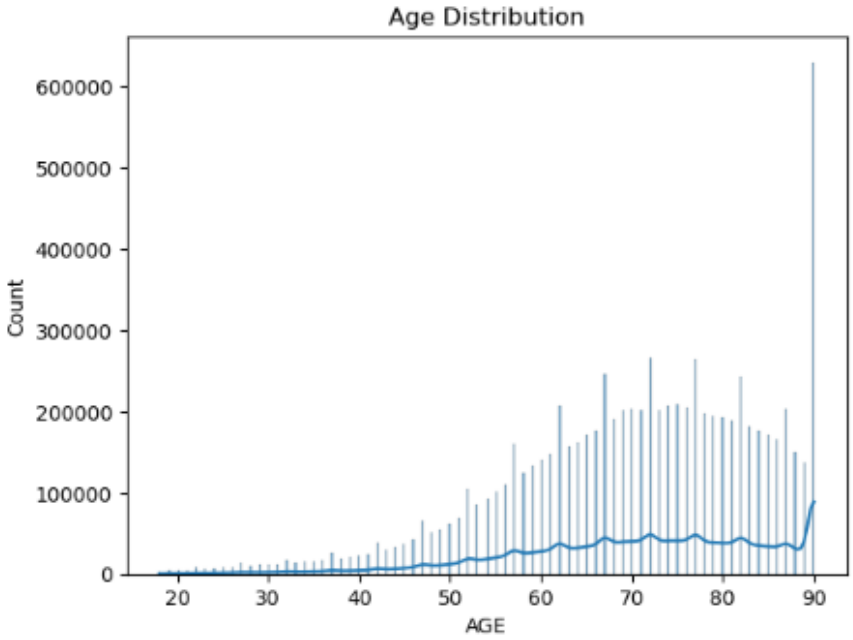
	count	mean	std	min	25%	50%	75%	max
AGE	8096311.0	7.038355e+01	1.421499e+01	18.0	62.0	72.0	82.0	90.0
AWEEKEND	8096311.0	2.192863e-01	4.137630e-01	0.0	0.0	0.0	0.0	1.0
AMONTH	8096311.0	6.449239e+00	3.481651e+00	1.0	3.0	6.0	10.0	12.0
DIED	8096311.0	4.137526e-02	1.991566e-01	0.0	0.0	0.0	0.0	1.0
DRG	8096311.0	4.194870e+02	2.586409e+02	1.0	242.0	309.0	638.0	999.0
ELECTIVE	8096311.0	1.374173e-01	3.442874e-01	0.0	0.0	0.0	0.0	1.0
FEMALE	8096311.0	4.787515e-01	4.995483e-01	0.0	0.0	0.0	1.0	1.0
HCUP_ED	8096311.0	9.668606e-01	7.957452e-01	0.0	0.0	1.0	1.0	4.0
HOSP_DIVISION_x	8096311.0	4.540700e+00	2.369788e+00	1.0	3.0	5.0	6.0	9.0
HOSP_NIS_x	8096311.0	4.569706e+04	2.371961e+04	10001.0	30054.0	50018.0	60274.0	90538.0
KEY_NIS	8096311.0	4.588643e+07	2.371043e+07	10000002.0	30072692.0	50044400.0	60348073.0	90983813.0
LOS	8096311.0	5.589661e+00	6.854974e+00	0.0	2.0	4.0	7.0	365.0
PAY1	8096311.0	1.561446e+00	1.044870e+00	1.0	1.0	1.0	2.0	6.0
RACE	8096311.0	1.472692e+00	1.013310e+00	1.0	1.0	1.0	2.0	6.0
TOTCHG	8096311.0	6.730961e+04	1.080116e+05	101.0	20723.0	38638.0	75000.0	9999999.0
YEAR	8096311.0	2.017312e+03	1.259611e+00	2.0	2016.0	2017.0	2018.0	2019.0
ZIPINC_QRTL	8096310.0	2.339966e+00	1.102252e+00	1.0	1.0	2.0	3.0	4.0
APRDRG	8096310.0	3.093631e+02	2.249911e+02	1.0	171.0	201.0	421.0	956.0
APRDRG_Risk_Mortality	8096310.0	2.499855e+00	9.492398e-01	0.0	2.0	3.0	3.0	4.0
APRDRG_Severity	8096310.0	2.641276e+00	8.754047e-01	0.0	2.0	3.0	3.0	4.0
HOSP_BEDSIZE	8096310.0	2.303009e+00	7.841764e-01	1.0	2.0	3.0	3.0	3.0
HOSP_LOCTEACH	8096310.0	2.587653e+00	6.545191e-01	1.0	2.0	3.0	3.0	3.0
HOSP_REGION	8096310.0	2.409207e+00	9.985263e-01	1.0	2.0	3.0	3.0	4.0
H_CONTRL	8096310.0	2.034995e+00	4.760906e-01	1.0	2.0	2.0	2.0	3.0
Acute_rheumatic_fever	8096310.0	1.311709e-04	1.145223e-02	0.0	0.0	0.0	0.0	1.0
Chronic_rheumatic_heart_diseases	8096310.0	8.105050e-03	8.966248e-02	0.0	0.0	0.0	0.0	1.0
Hypertensive_diseases	8096310.0	1.557752e-01	3.626421e-01	0.0	0.0	0.0	0.0	1.0
Ischemic_heart_diseases	8096310.0	2.994743e-01	4.580278e-01	0.0	0.0	0.0	1.0	1.0
Pulmonary_heart_diseases	8096310.0	4.860412e-02	2.150390e-01	0.0	0.0	0.0	0.0	1.0
Other_heart_diseases	8096310.0	4.879102e-01	4.998538e-01	0.0	0.0	0.0	1.0	1.0

3. Exploratory Data Analysis (EDA)

3.1 Continuous Variable Distributions

Age Distribution

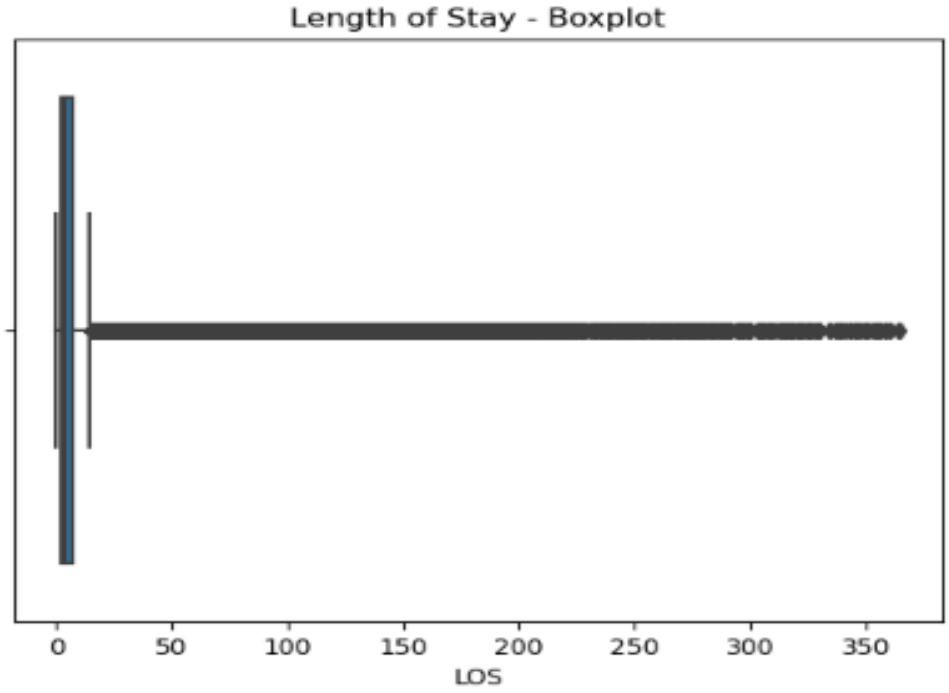
An initial exploration of the patient **age distribution** was conducted to better understand the demographic profile of cardiovascular admissions. A histogram was plotted, which revealed a **strong skew toward elderly populations**, with a substantial concentration of patients between **65 and 85 years of age**. The **mean age was approximately 70 years**, consistent with the epidemiology of acute cardiovascular conditions such as myocardial infarctions, heart failure, and arrhythmias, which predominantly affect older adults.
The distribution showed a **sharp increase** beginning around the sixth decade of life, peaking in the 70–80 year range, and gradually tapering off in the very elderly population (>90 years). Younger patients (<50 years) represented only a **small minority** of total admissions, reinforcing that cardiovascular disease continues to disproportionately burden aging populations. Understanding the age distribution was crucial not only for demographic characterization but also for adjusting subsequent modeling phases, as **age is a strong confounder** influencing both **mortality risk** and **treatment decisions** in clinical practice.



Observation:
The peak age groups ranged from 65 to 85 years. Younger patients (<50 years) were relatively rare.

Length of Stay (LOS)

Length of Stay was visualized using a boxplot to detect outliers.



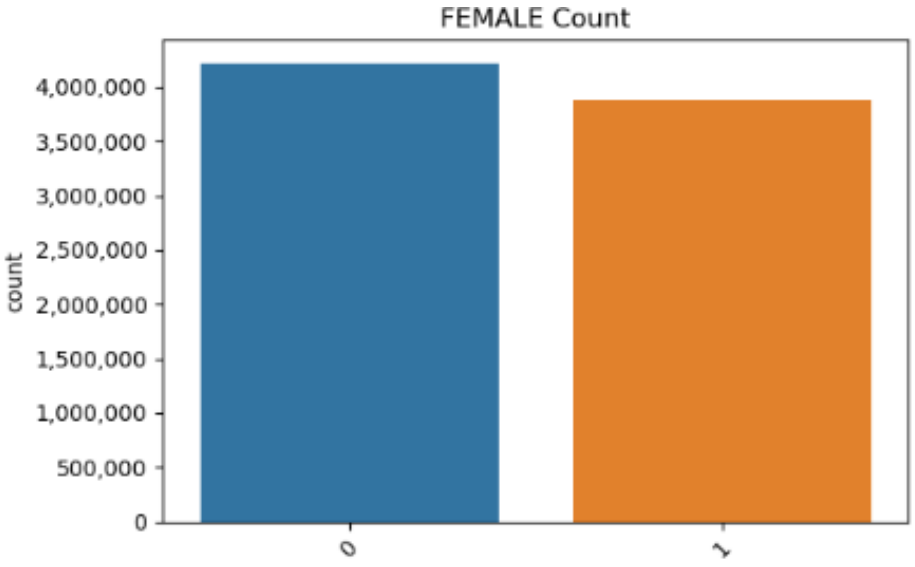
Observation:

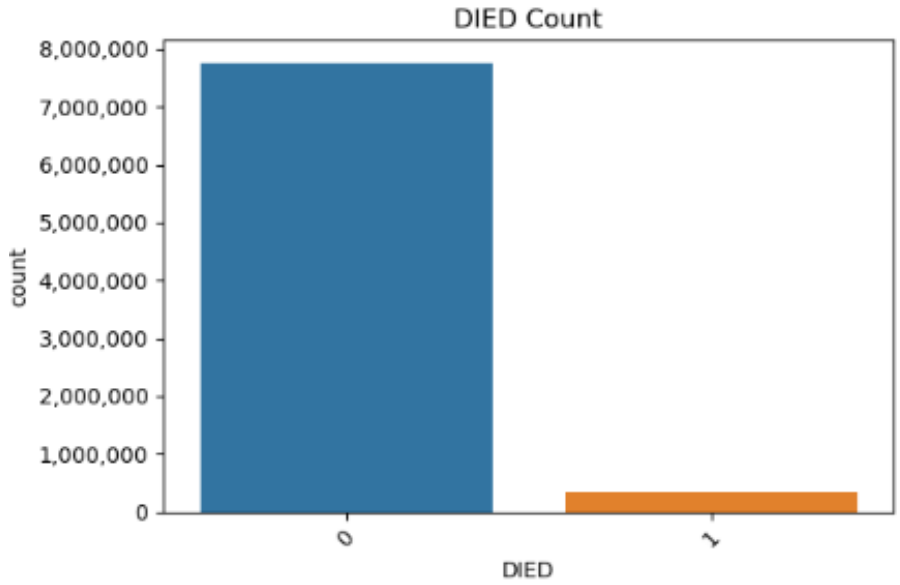
- Most hospital stays were under 10 days.
- A few patients had extreme LOS (>50 days), indicating rare, complex cases.

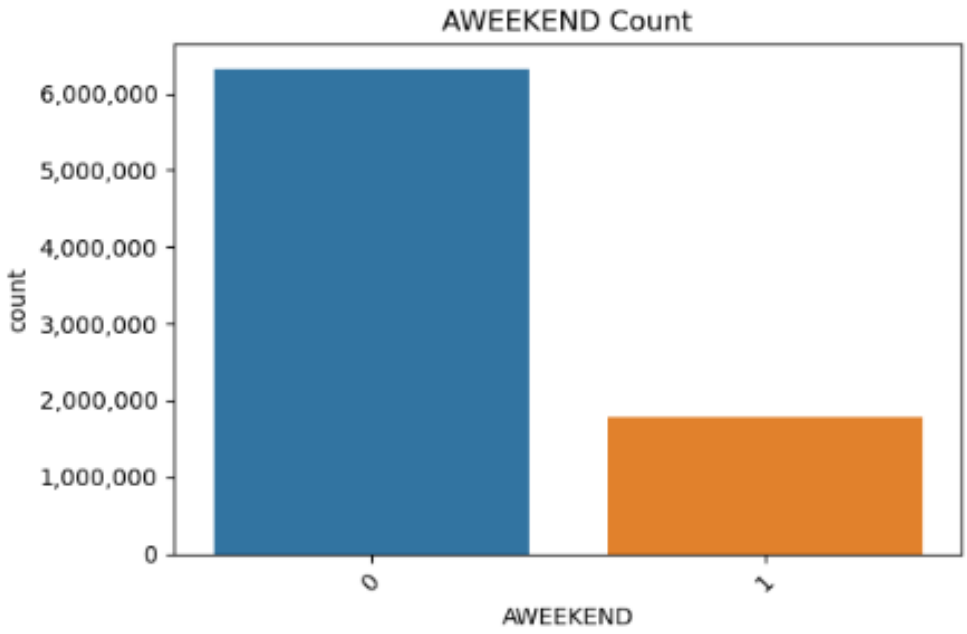
This information is vital because longer stays often correlate with higher severity and mortality risks.

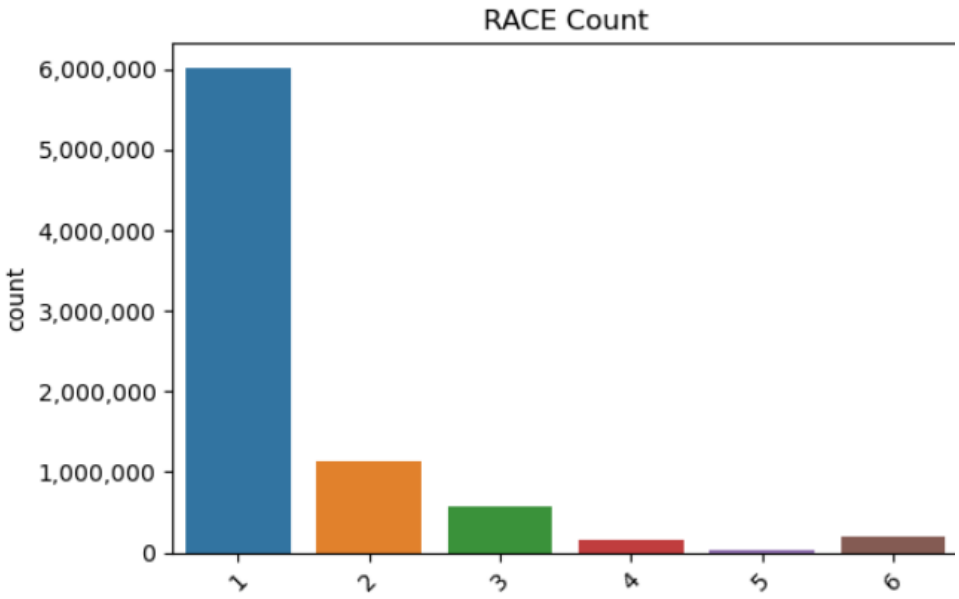
3.2 Categorical Variables Analysis

Count plots were created for major categorical variables:









Key Observations:

Upon reviewing the categorical variables in the dataset through visualizations such as countplots, several important demographic and admission-related trends emerged. First, the **gender distribution** was observed to be **approximately balanced**, with about **48% of admissions involving female patients**. This relatively even distribution between male and female patients suggests that subsequent analyses comparing outcomes across gender categories would not be heavily biased by disproportionate group sizes.

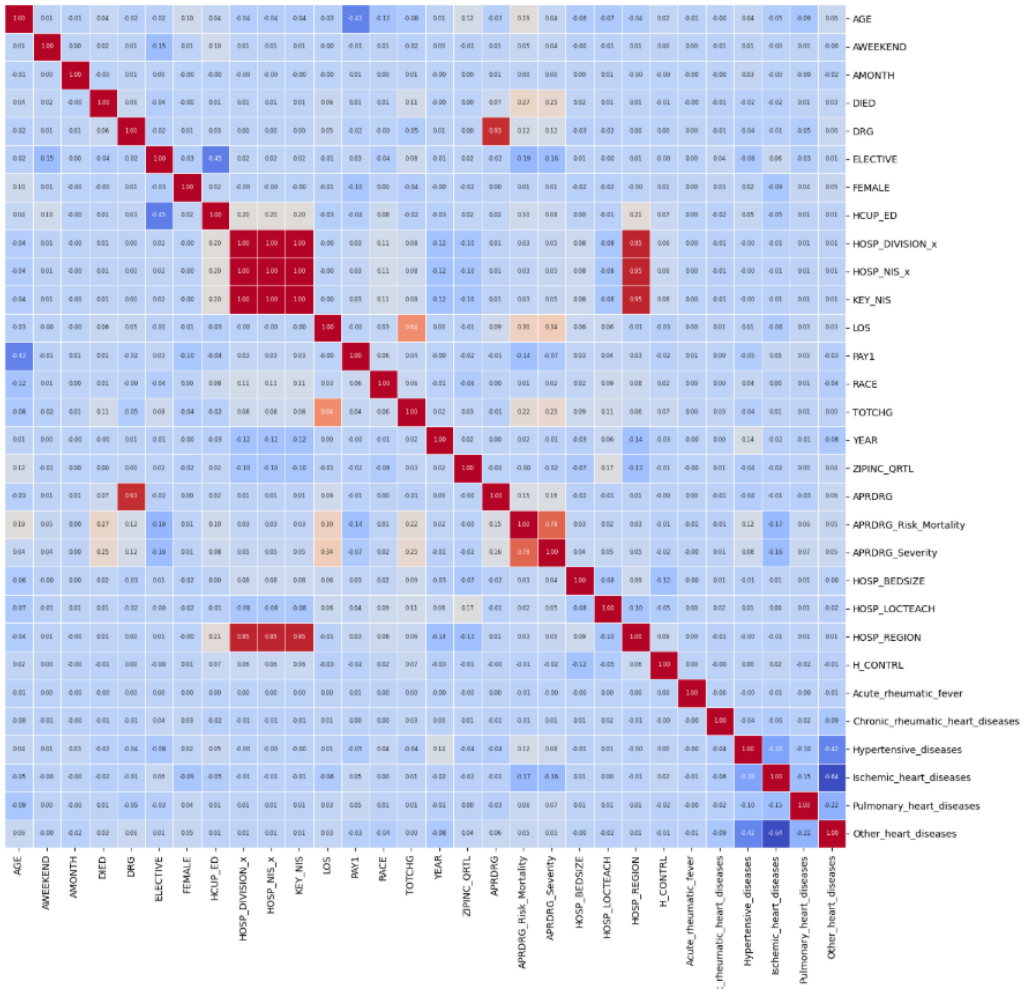
Second, the distribution of admissions by day revealed that **weekend admissions accounted for approximately 22%** of total hospitalizations, confirming the expected pattern that hospital admissions tend to be more concentrated during weekdays. This finding underscores the operational importance of evaluating whether the reduced admission volumes on weekends correspond to differences in patient outcomes or resource allocation challenges.

Third, the overall **in-hospital mortality rate was approximately 4%** across the dataset. This mortality rate is consistent with prior reports on acute cardiovascular admissions in large national datasets and reinforces the validity of the patient population captured in the study.

The categorical variable visualizations played a critical role in **validating the dataset's balance**, ensuring that no single category was overwhelmingly dominant or underrepresented.

3.3 Correlation Analysis

A comprehensive correlation heatmap for all numeric variables was generated.



Insights:

- **Severity of Illness and Mortality**:
  A **strong positive correlation** was observed between the **APRDRG Severity of Illness score** and **in-hospital mortality**.
  Similarly, the **APRDRG Risk of Mortality score** exhibited a robust positive association with death outcomes.
  This finding reinforces the clinical understanding that patients with higher severity burdens are inherently at greater risk of adverse outcomes.

- **Length of Stay and Total Hospital Charges**:
  As expected, **length of stay** showed a **moderate positive correlation** with **total hospital charges**.
  Patients requiring longer hospitalization naturally accumulated higher healthcare costs, driven by prolonged resource utilization.

- **Age and Mortality**:
  A mild positive correlation between **age** and **mortality** was also noted, consistent with the increased vulnerability of older patients.

Importantly, despite these expected correlations, the overall examination suggested that **severe multicollinearity among independent variables was modest**.
This condition was favorable for subsequent logistic regression modeling, ensuring that coefficient estimates would remain interpretable and that the risk of unstable parameter inflation (variance inflation factor issues) would be minimized.

4. Group-wise Comparative Analysis

4.1 Weekend vs Weekday Admissions

The dataset was split by Weekend_Admission status, and both continuous and categorical variables were compared.

Feature	Weekday	Weekend	p-value	Interpretation
Mortality Rate	**3.98%**	**4.72%**	**<0.0001**	**Higher on weekends**
Age (mean)	**70.32 years**	**70.60 years**	**<0.0001**	**Slightly older patients**
LOS (mean)	**5.60 days**	**5.54 days**	**<0.0001**	**Slightly shorter stay on weekends**
Severity	**Lower**	**Higher**	**<0.0001**	**More severe admissions on weekends**

```
==============================
Descriptive Stats: Weekend vs Weekday Admission
==============================

AGE Summary:
                          mean        std    count
Weekend_Admission
0                   70.323607  14.099256  6320901
1                   70.596959  14.617615  1775409
t-test p-value for AGE: 4.7658e-109

LOS Summary:
                          mean       std    count
Weekend_Admission
0                   5.604268  6.886278  6320901
1                   5.537657  6.742089  1775409
t-test p-value for LOS: 5.3936e-31

TOTCHG Summary:
                              mean            std    count
Weekend_Admission
0                   68376.554882  108742.934176  6320901
1                   63511.048445  105278.845039  1775409
t-test p-value for TOTCHG: 0.0000e+00

APRDRG_Risk_Mortality Summary:
                          mean       std    count
Weekend_Admission
0                   2.475601  0.949559  6320901
1                   2.586205  0.943052  1775409
t-test p-value for APRDRG_Risk_Mortality: 0.0000e+00

APRDRG_Severity Summary:
                          mean       std    count
Weekend_Admission
0                   2.623176  0.876576  6320901
1                   2.705718  0.868164  1775409
t-test p-value for APRDRG_Severity: 0.0000e+00

FEMALE Crosstab:
Weekend_Admission        0       1
FEMALE
0                  3313492  906697
1                  3007409  868712
Chi-square p-value for FEMALE: 1.3327e-222

RACE Crosstab:
Weekend_Admission        0        1
RACE
1                  4717252  1302068
2                   865346   258121
3                   440280   129532
4                   122596    36449
5                    28350     8148
6                   147077    41091
Chi-square p-value for RACE: 1.6693e-287
```

Interpretation:

Analysis of patient characteristics stratified by admission day revealed important differences between weekend and weekday admissions.

Specifically, patients admitted during weekends exhibited **higher severity scores** and **higher crude mortality rates** compared to their weekday counterparts.

This pattern suggests two potential underlying explanations:

- First, it is possible that **patients presenting on weekends are inherently sicker**, either due to delayed care-seeking behaviors (waiting until symptoms worsen) or differences in referral patterns that result in higher-acuity cases being admitted during weekends.
- Second, the observed differences may **reflect systemic factors related to hospital service availability**, such as reduced staffing levels, limited access to specialized diagnostic procedures, or slower response times for critical interventions during weekends.

Although the unadjusted analysis highlights these disparities, it is important to note that confounding factors such as patient age, comorbidity burden, and socioeconomic status may contribute to the observed differences.

Thus, while the crude findings hint at both **patient-level** and **system-level challenges**, multivariate modeling was essential to disentangle these effects and determine whether weekend admission independently influences mortality risk after adjustment.

4.2 Teaching vs Non-Teaching Hospitals

Separately, admissions were compared based on hospital teaching status.

Feature	Non-Teaching	Teaching	p-value	Interpretation
Mortality Rate	**3.75%**	**4.32%**	**<0.0001**	**Slightly worse outcomes in teaching hospitals**
Age (mean)	**71.78 years**	**69.73 years**	**<0.0001**	**Younger patients in teaching hospitals**
LOS (mean)	**4.96 days**	**5.89 days**	**<0.0001**	**Longer stays in teaching hospitals**

```
==============================
Descriptive Stats: Teaching vs Non-Teaching Hospitals
==============================

AGE Summary:
                          mean        std    count
Teaching_Hospital
0                    71.782192  13.642302  2585213
1                    69.727457  14.429195  5511097
t-test p-value for AGE: 0.0000e+00

LOS Summary:
                          mean       std    count
Teaching_Hospital
0                    4.958709  5.614916  2585213
1                    5.885636  7.346434  5511097
t-test p-value for LOS: 0.0000e+00

TOTCHG Summary:
                               mean            std    count
Teaching_Hospital
0                    53325.155595   76241.830839  2585213
1                    73869.619001  119489.112690  5511097
t-test p-value for TOTCHG: 0.0000e+00

APRDRG_Risk_Mortality Summary:
                          mean       std    count
Teaching_Hospital
0                    2.470077  0.932388  2585213
1                    2.513824  0.956723  5511097
t-test p-value for APRDRG_Risk_Mortality: 0.0000e+00

APRDRG_Severity Summary:
                          mean       std    count
Teaching_Hospital
0                    2.590163  0.859750  2585213
1                    2.665253  0.881632  5511097
t-test p-value for APRDRG_Severity: 0.0000e+00

FEMALE Crosstab:
Teaching_Hospital        0        1
FEMALE
0                  1310095  2910094
1                  1275118  2601003
Chi-square p-value for FEMALE: 0.0000e+00
```



```
RACE Crosstab:
Teaching_Hospital        0        1
RACE
1                  2088330  3930990
2                   243085   880382
3                   153617   416195
4                    40318   118727
5                    14389    22109
6                    45474   142694
Chi-square p-value for RACE: 0.0000e+00
```

Interpretation:

- **Age Distribution**:
  Patients admitted to teaching hospitals tended to be **younger** on average compared to those admitted to non-teaching hospitals.
  This may reflect referral patterns where younger patients with more complex conditions are transferred to tertiary centers offering advanced interventions.
- **Severity of Illness**:
  Despite being younger, patients at teaching hospitals had **higher clinical severity scores**, indicating that they were more acutely ill or had greater comorbidity burdens at the time of admission.
- **Mortality Rates**:
  Crude mortality rates were observed to be **slightly higher** in teaching hospitals relative to non-teaching facilities.
  Importantly, this difference persisted even though teaching hospitals typically have greater staffing, access to specialists, and procedural capabilities.
  One plausible explanation is that **teaching hospitals manage a disproportionately higher volume of complex, high-risk cases**, leading to intrinsically higher baseline mortality despite superior resources.

5. Key Initial Findings

From the EDA phase, several important conclusions emerged:

- Weekend admissions had higher crude mortality rates compared to weekday admissions.
- Teaching hospitals admitted slightly younger but higher-risk patients.
- Severity of illness was higher among weekend admissions.
- Length of stay was longer at teaching hospitals but slightly shorter on weekends.
- Access to procedures (not analyzed here but flagged) might contribute to differences.
- These differences underscored the need for adjusted multivariate modeling to control for confounding factors such as severity, age, and insurance type.

6. Conclusion

The exploratory data analysis phase of this project uncovered several critical patterns that informed the design of the subsequent multivariate modeling.
First, there were evident **mortality differences between weekend and weekday admissions**, suggesting potential systemic factors influencing patient outcomes.
Second, **teaching hospital status** emerged as an important correlate of patient characteristics and outcomes, with teaching institutions admitting younger but clinically

sicker patients and exhibiting slightly higher crude mortality rates.
Third, the analysis highlighted that **severity of illness and comorbidity burden were major confounding variables** that could obscure true associations between hospital operational factors and mortality if not properly accounted for.
While initial crude comparisons suggested that both weekend admission and hospital teaching status influenced mortality, these findings emphasized the necessity of conducting **adjusted multivariate analyses**.
Only through logistic regression modeling could the effects of **patient-level factors** be separated from **hospital system effects**, allowing for a more accurate and meaningful interpretation of the drivers behind observed outcome disparities.