Final Report (One Page) — Hospital Length of Stay Dataset

1. Dataset and Objective

Dataset: Hospital Length of Stay Dataset
File: LengthOfStay.csv
Objective: To analyze the dataset and build a **Linear Regression** model that predicts **lengthofstay** (number of days a patient stays in the hospital).

2. Data Loading

The dataset was loaded from an external CSV file using Pandas. Basic exploration was conducted using `head()`, `shape`, and `info()` to understand the number of records, variables, and data types.

3. Descriptive Statistics

Descriptive statistics were produced using `describe()` for numerical variables (mean, standard deviation, min/max) and `describe(include="all")` to summarize categorical variables (unique values and frequency).
These statistics were selected to understand central tendency, variability, and the most common categories.

4. Data Visualization (EDA)

Histograms were created for numerical variables to observe distributions, identify skewness, and detect possible outliers.
The visualizations provided insights into how patient-related variables vary and how they may influence hospital stay duration.

5. Data Quality Checks

- **Missing Values:** The dataset was checked using `isnull().sum()`. (If no missing values were found, no imputation was needed.)
- **Duplicates:** Duplicate rows were checked using `duplicated().sum()` and removed where necessary.
  These steps ensured clean and consistent data for modeling.

6. Categorical Variables (Unique Values)

Unique values and frequencies of categorical variables were investigated to understand category levels and identify any inconsistent entries. The most frequent categories were noted for interpretation.

7. Predictive Modeling — Linear Regression

A Linear Regression model was built to predict **lengthofstay**.
Categorical variables were handled using one-hot encoding during preprocessing. The model was trained using a train/test split approach.

8. Model Evaluation and Interpretation

Model performance was evaluated using:

- **MAE:** Average prediction error in days
- **RMSE:** Calculated as √(MSE), penalizes larger errors more
- **R^2:** Measures how much variation in lengthofstay is explained by the model
  These metrics provide a baseline understanding of prediction quality.

9. Conclusion

The project successfully implemented a complete data analysis pipeline: loading, descriptive statistics, visualization, cleaning checks, categorical analysis, and Linear Regression modeling. This approach can support hospitals in planning resources such as beds and staff by estimating patient length of stay.