Healthcare Dataset: Exploratory Data Analysis and Feature Engineering Report

Executive Summary

This report presents a comprehensive analysis of a healthcare dataset encompassing information on patients admitted to a hospital.

Dataset Overview

The dataset captures demographic details, medical conditions, treatments, and billing data for patients. Key features include:

- Demographics: Age, Gender, Blood Type
- Admission Details: Date of Admission, Admission Type (Emergency, Elective, Transfer)
- Medical Information: Medical Condition, Doctor, Hospital
- Treatment Details: Medication, Test Results
- Billing Information: Billing Amount, Room Number

Data Cleaning and Preparation

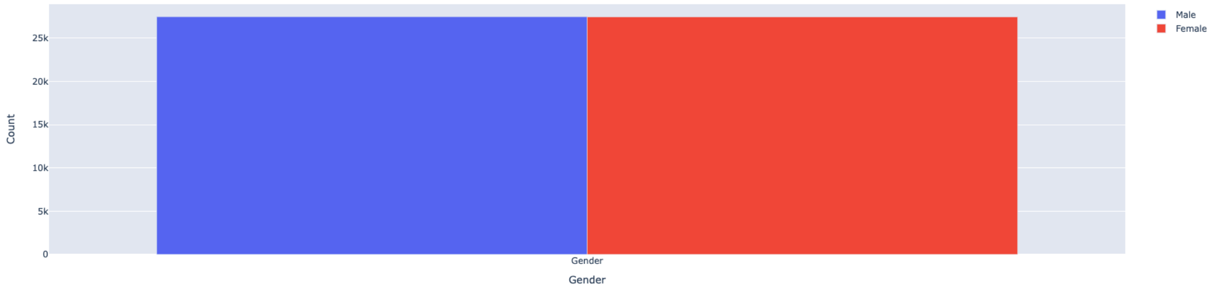
- **Duplicate Removal:** Redundant records were eliminated, resulting in a clean dataset with 54,966 unique patient records.
- **Data Formatting:** Patient names were standardized to lowercase. Date formats were transformed for consistency.

Key Observations

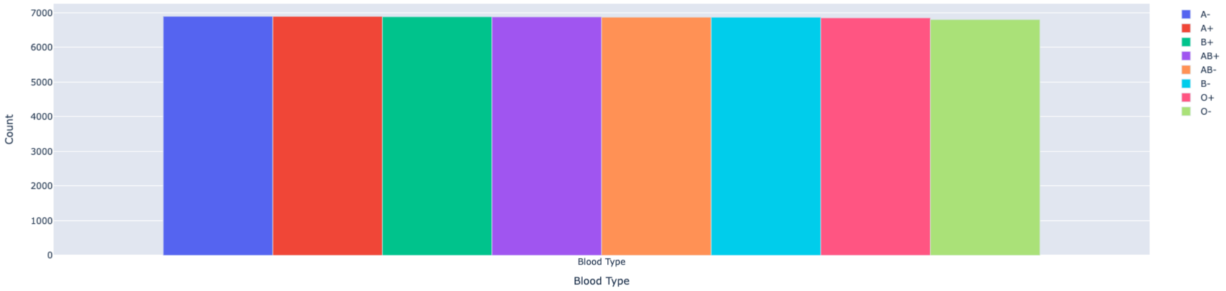
- **Patient Age:** Age ranges from 13 to 89 years, with an average of 52.
- **Hospital Capacity:** Room numbers suggest varied accommodation capacities across facilities (101-500).
- **Temporal Coverage:** The dataset spans a five-year period, encompassing admissions from May 8th, 2019 to May 7th, 2024.
- **Admission Types:** Three primary admission types exist: emergency, elective, and transfer.
- **Blood Type Distribution:** A- is the most prevalent blood type.
- **Hospital and Doctor Distribution:** Patients were admitted across 44 hospitals, with LLC Smith having the highest frequency. Michael Smith treated the most patients among 27 doctors.

Visual Insights

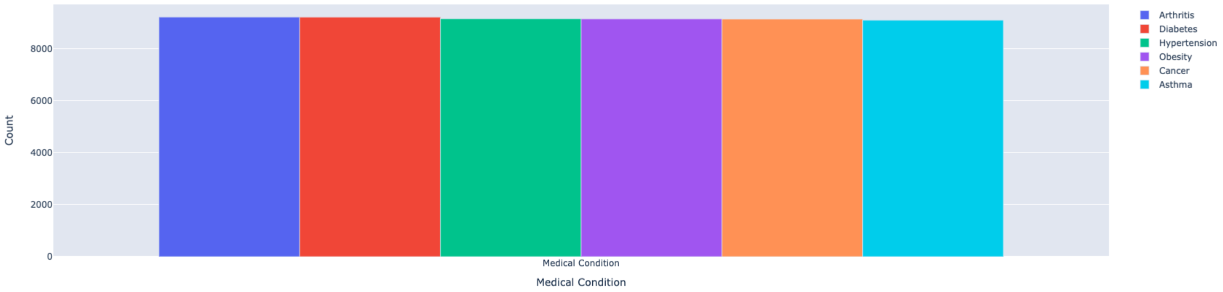
Distribution of Gender

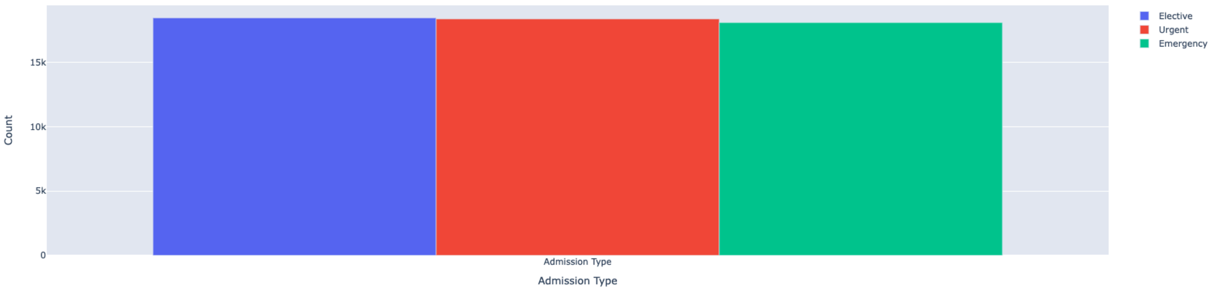


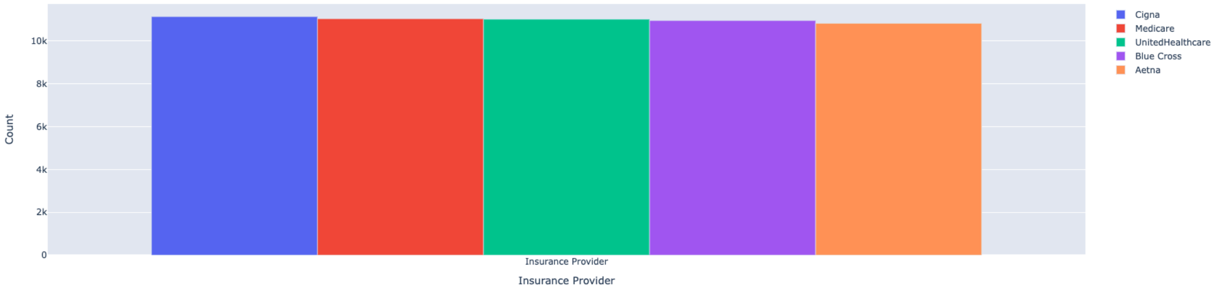
Distribution of Blood Type

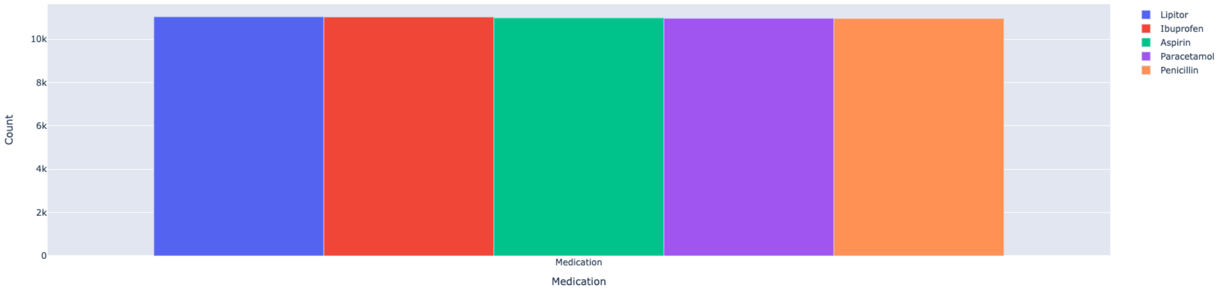


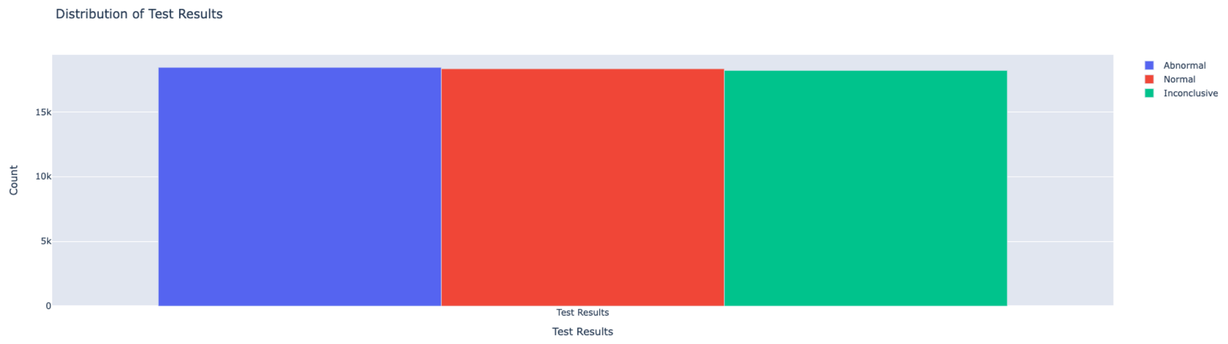
Distribution of Medical Condition

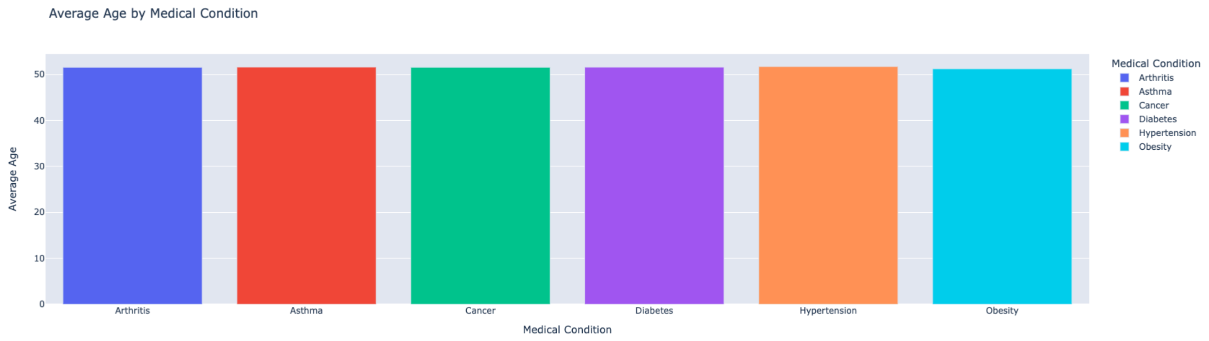


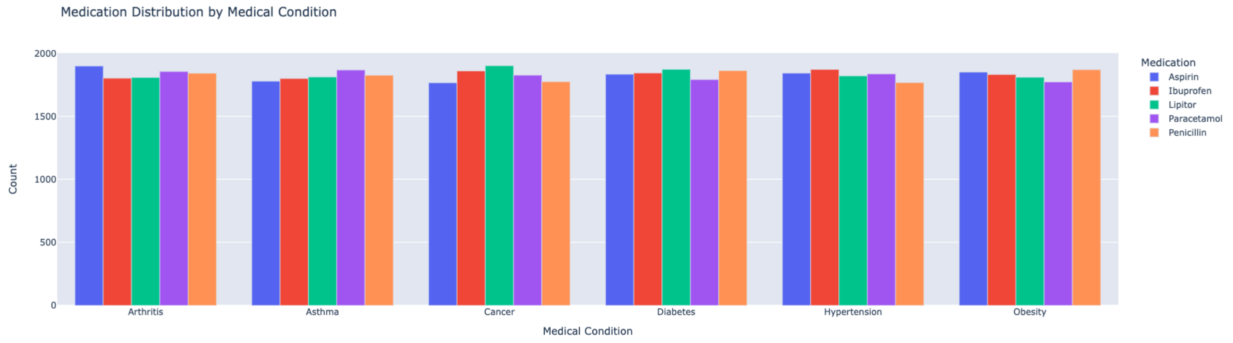

Distribution of Admission Type

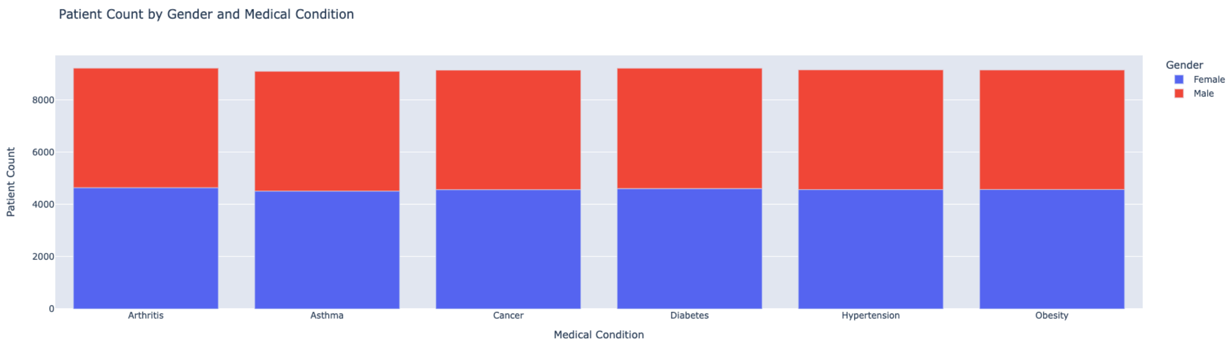

Distribution of Insurance Provider

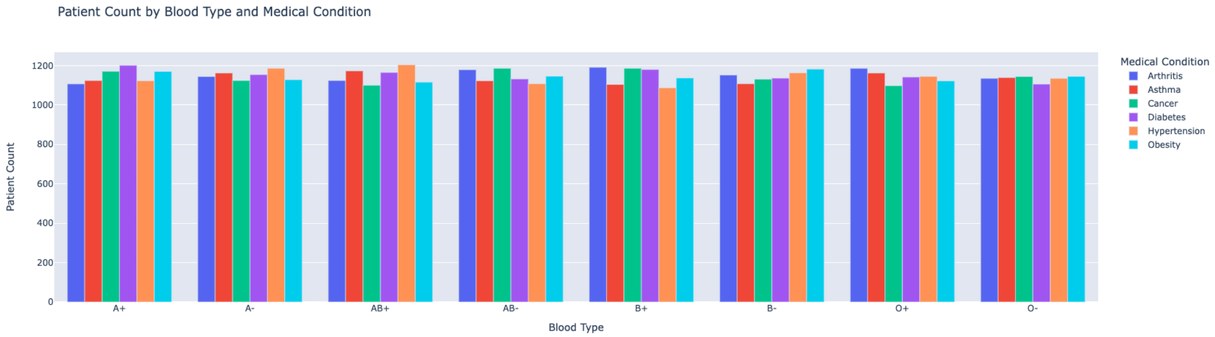

Distribution of Medication


Distribution of Test Results


Average Age by Medical Condition


Medication Distribution by Medical Condition





Explorations through visualizations revealed insights on:

- **Gender Distribution:** Balanced representation of male and female patients across various categories (admission types, medical conditions).
- **Medical Condition and Age:** Average age varied by medical condition, potentially indicating age-related health issues.
- **Blood Type and Gender:** Gender distribution was fairly balanced across blood types.
- **Admission Type and Medical Condition:** Patterns emerged in hospital admissions based on admission type and associated medical conditions.
- **Test Results and Admission Type:** Certain test results were more prevalent based on the type of admission.
- **Medication Distribution:** Gender-based medication distribution provided insights into potential prescription patterns.

Additional Insights

- The most frequent blood type is A-.
- There are 44 unique hospitals represented in the dataset.
- The oldest patient is 89 years old.
- Michael Smith is the doctor treating the highest number of patients.
- Lipitor is the most frequently prescribed medication.

- Peak bed occupancy occurs in August.
- The average billing amount is $25,544.31.
- Gender distribution is nearly equal with 27,496 male and 27,470 female patients.
- Arthritis, Diabetes, and Hypertension are the top three most common medical conditions.
- Insurance providers include Cigna, Medicare, UnitedHealthcare, Blue Cross, and Aetna.
- Admission types are categorized as Elective, Urgent, and Emergency.
- Patient types are evenly distributed among Abnormal, Normal, and Inconclusive categories.

Feature Engineering

- **Date Features:** Created datetime features from date of admission and discharge data.
- **Binning:** Bucketized the continuous billing amount variable into groups with various ranges.
- **Data Split:** Separated 20% of the data for validation and retained the remaining 80% for training.
- **High Cardinality Feature Removal:** Dropped columns like Name, Doctor, and Hospital due to their high number of unique values.
- **Billing Amount Bucketing:** Grouped billing amounts into ranges to capture high-cost treatments potentially associated with high-risk patients.
- **Categorical Feature Encoding:** Employed count encoding, ordered integer encoding, and mean encoding for categorical columns.

Feature Selection

- Removed constant and quasi-constant features.
- Eliminated duplicate features.
- Removed correlated features to optimize the feature space.
- Used row index as a unique identifier.

Hyperparameter Tuning and Model Selection

Several machine learning models were trained, including KNN Classifier, CatBoost Classifier, LightGBM Classifier, and XGBoost Classifier. Initial accuracy scores were obtained before hyperparameter tuning:

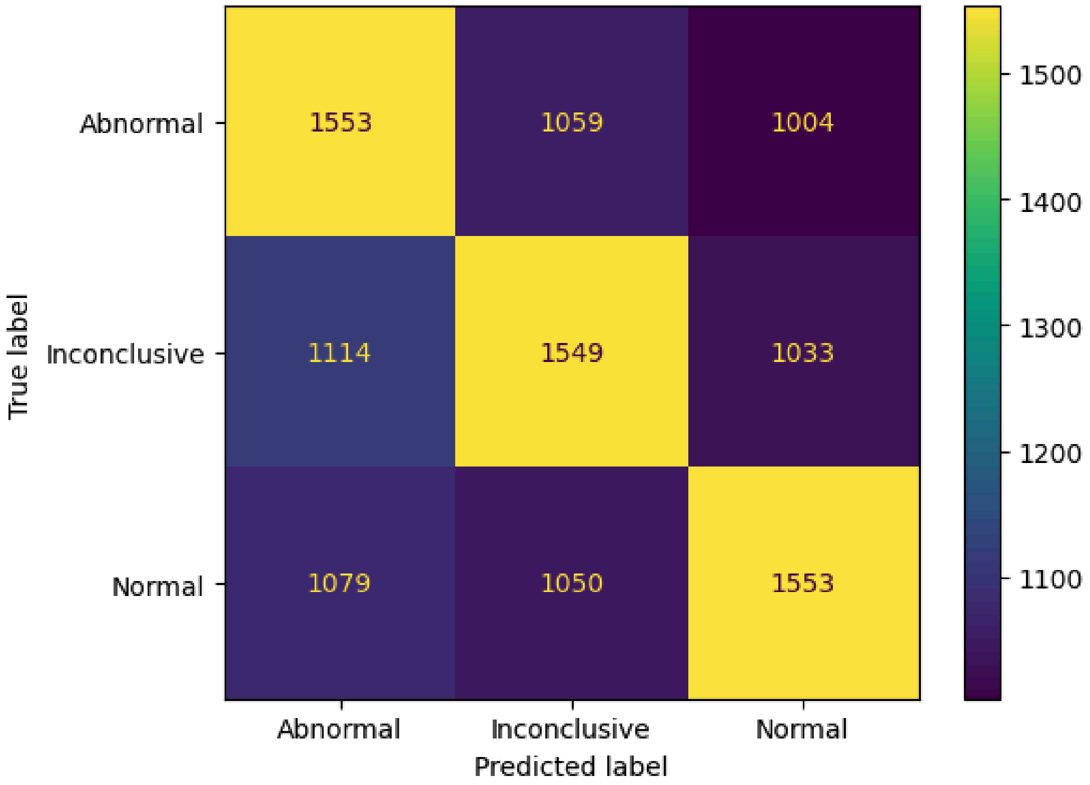
- XGBoost Classifier: 0.3832
- LGBM Classifier: 0.3666
- CatBoost Classifier: 0.3782

Hyperparameter tuning was then applied to the XGBoost Classifier, resulting in a significant improvement to an accuracy of 42.34%. Grid Search CV was utilized for this hyperparameter optimization process.

Model Performance

To comprehensively evaluate model performance, we have included a confusion matrix (attached). This matrix allows us to calculate precision, recall, and F1-score for each class. While the report does not include these metrics explicitly, the confusion matrix provides the foundation for their calculation.

It's important to note that the model exhibits similar performance across different classes. This suggests a relatively balanced classification capability. Further analysis can be conducted to delve deeper into specific class performance if needed.


True label
Abnormal
Inconclusive
Normal
Abnormal
Inconclusive
Normal
Predicted label