**Diabetic Patient Data for Hospital Readmission Prediction**

Capstone project for the second cohort of the Digital Egypt Initiative, under the auspices of the Ministry of Communications and Information Technology

**Data Science Track**

***Under Supervision of***

**Dr. Eslam Elreedy**

Acknowledgment

As we reflect on the completion of this project, it is important to recognize the collective efforts that have made this dataset and its analysis possible. The journey of cleaning, analyzing, and preparing the Diabetic Data Cleaning dataset for predictive modeling has been both challenging and rewarding. It is through the dedication and support of various individuals and organizations that this work has reached its current stage.

First and foremost, we extend our deepest gratitude to the original creators of the dataset, whose efforts in compiling and sharing this valuable resource have enabled researchers and data scientists to explore critical healthcare challenges, such as predicting hospital readmissions for diabetic patients. The dataset, sourced from Kaggle (Diabetic Data Cleaning Dataset), is a testament to the importance of open data in advancing medical research and improving patient outcomes.

We would also like to acknowledge the contributions of the healthcare institutions and professionals who originally collected and anonymized the patient data. Their commitment to ethical data practices and patient privacy has ensured that this dataset can be used responsibly for research purposes.

Special thanks go to the data science community, whose collaborative spirit and shared knowledge have been instrumental in refining the techniques and methodologies applied in this project. The guidance and insights from peers, mentors, and online resources have been invaluable in overcoming the technical challenges encountered during the data cleaning and analysis process.

In particular, we are grateful to Dr. Eslam Elreedy, whose expertise and mentorship have been a guiding light throughout this project. His dedication to fostering a deeper understanding of data science principles and his unwavering support have not only contributed to the success of this work but have also inspired growth and learning beyond the scope of this dataset.

Additionally, we appreciate the support of our colleagues and fellow researchers, who have provided constructive feedback and encouragement. Their contributions, whether through direct collaboration or shared resources, have enriched this project and broadened its potential impact.

Finally, we are thankful for the opportunity to work with this dataset and contribute to the ongoing efforts to improve healthcare outcomes through data-driven insights. The knowledge and experience gained from this project will undoubtedly serve as a foundation for future endeavors in the field of medical data analysis.

As we conclude this phase of the project, we do so with a sense of pride and accomplishment, knowing that the work done here has the potential to make a meaningful difference. We look forward to the next steps, including the development of predictive models and further exploration of the data, with the hope of contributing to advancements in patient care and hospital management.

Congratulations to all who have been part of this journey—we did it!

**DECLARATION**

We hereby certify that this material, which we now submit for assessment on the program of the Data Science Track, is entirely our own work. We have exercised reasonable care to ensure that the work is original and, to the best of our knowledge, does not breach any law of copyright. Any work or ideas taken from external sources have been properly cited and acknowledged within the text of our work. This declaration applies to all aspects of the project, including the cleaning, analysis, and documentation of the Diabetic Data Cleaning dataset.

**Signed:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
**Date:** April 29, 2025

**ABSTRACT**

The Diabetic Data Cleaning dataset, accessible via Kaggle and sourced from the UCI Machine Learning Repository, comprises a rigorously curated and preprocessed collection of diabetic patient records, optimized for advanced healthcare analytics and machine learning applications. Preprocessing entailed meticulous imputation of missing values, encoding of categorical variables, and validation of data consistency to ensure analytical integrity. The dataset includes critical patient attributes—demographics, medical histories, and readmission outcomes—rendering it an essential tool for predictive modeling of patient outcomes, exploration of health determinants, and statistical assessment of diabetic care practices. This resource empowers data-driven advancements in patient care and healthcare system efficiency.

The dataset's utility is further underscored by its comprehensive feature set, which includes patient demographics (age, gender, race), medical history (diagnosis codes, number of medications), and hospital admission details (admission type, time in hospital, number of lab procedures). The primary outcome variable, readmission status, is pivotal for developing models that predict patient readmissions, thereby enabling healthcare providers to implement targeted interventions and reduce readmission rates. Additional attributes, such as the number of emergency visits and outpatient encounters, enhance the dataset’s capacity to support longitudinal studies of patient health trajectories. Designed for versatility, it accommodates a range of analytical approaches, from statistical inference to deep learning, fostering innovation in healthcare research. This dataset is a critical asset for researchers, data scientists, and healthcare professionals aiming to leverage data-driven insights for improved patient outcomes and healthcare system efficiency.

# **INTRODUCTION AND BACKGROUND**

## **Introduction**

In the complex and evolving landscape of healthcare, where patient outcomes intersect with operational challenges, hospitals face significant issues related to patient readmissions, particularly among diabetic patients. At the forefront of these concerns is the high rate of hospital readmissions, which not only impacts patient health but also places substantial financial and operational burdens on healthcare systems. The Diabetic Data Cleaning dataset, sourced from Kaggle ([Diabetic Data Cleaning Dataset](https://www.kaggle.com/datasets/smit1212/diabetic-data-cleaning)), provides a comprehensive collection of 101,766 patient encounters, capturing demographic, medical, and administrative data to address these challenges.

Recognizing the urgency of improving patient care and reducing readmissions, our project introduces a robust data-driven approach to clean, preprocess, and analyze this dataset. This comprehensive document delves into the intricacies of our methodology, offering a detailed exploration of data cleaning techniques, exploratory data analysis, feature engineering, and the preparation of the dataset for predictive modeling. Our initiative leverages advanced data science techniques to uncover insights and develop solutions that enhance healthcare delivery.

The core objective of this project is to address the root causes of hospital readmissions by harnessing the power of data analytics. Through a thorough understanding of the dataset’s complexities, our work aims to create a foundation for predictive models that can identify at-risk patients, thereby reducing readmission rates and improving patient outcomes. This introduction serves as a prelude to a deeper examination of our commitment to advancing healthcare through data-driven innovation, specifically tailored to the unique challenges posed by diabetic patient care.

## **Problem definition**

Healthcare systems worldwide face critical challenges in managing hospital readmissions, particularly for diabetic patients. Two primary issues dominate this landscape, significantly impacting patient well-being and healthcare efficiency. Firstly, the high prevalence of readmissions, especially within 30 days of discharge, poses a severe threat to patient health, leading to prolonged recovery times and increased medical costs. These readmissions often stem from complex factors, including inadequate post-discharge care, comorbidities, and medication non-adherence. Secondly, the variability and incompleteness of patient data exacerbate the challenge of identifying at-risk individuals. Missing values, inconsistent formats, and high-dimensional data in the Diabetic Data Cleaning dataset hinder accurate analysis and prediction, necessitating robust data cleaning and preprocessing strategies.

The implications of these challenges extend beyond patient care, contributing to financial strain on healthcare systems through penalties for high readmission rates and increased resource utilization. The urgency of addressing these issues underscores the need for an effective, data-driven intervention to improve patient outcomes and optimize hospital operations.

### **Solution**

Our proposed solution is a comprehensive and innovative approach to tackling the challenges of hospital readmissions using the Diabetic Data Cleaning dataset. By leveraging advanced data science techniques, this solution integrates robust data cleaning, exploratory analysis, and feature engineering into a cohesive workflow, preparing the dataset for predictive modeling. Key components of the solution include:

* **Data Cleaning and Preprocessing**:  
  We address missing values (e.g., in weight, max\_glu\_serum, and A1Cresult) by imputing with appropriate placeholders (e.g., "No Test") or the mode for categorical features. Outliers in numerical columns like num\_lab\_procedures are mitigated through imputation or transformation, ensuring data consistency. Categorical variables (e.g., gender, race) are converted to appropriate types, and irrelevant columns like weight are dropped to streamline analysis.
* **Exploratory Data Analysis (EDA)**:  
  Through univariate and bivariate analyses, we uncover patterns in the dataset, such as the positive correlation between num\_lab\_procedures and time\_in\_hospital, indicating that complex cases require more tests and longer stays. The distribution of readmitted highlights class imbalance, guiding model development strategies.
* **Feature Engineering and Transformation**:  
  We create new features, such as time\_diagnoses\_interaction (multiplying time\_in\_hospital and number\_diagnoses), to capture combined effects. Numerical features are normalized using MinMaxScaler, and categorical variables are one-hot encoded to ensure compatibility with machine learning algorithms. The readmitted column is encoded as {NO: 0, >30: 1, <30: 2} to facilitate modeling.

This solution prepares a clean, structured dataset ready for predictive modeling to identify factors contributing to readmissions. By addressing data quality issues and extracting meaningful insights, our approach lays the groundwork for developing models that can reduce readmission rates and improve patient care.

### **Scope**

The scope of the problem is extensive, affecting various aspects of healthcare and requiring a multifaceted approach to address effectively. Key dimensions include:

* **Patient Outcomes**: High readmission rates, particularly within 30 days, compromise patient health and recovery. The scope encompasses identifying risk factors and improving post-discharge care to enhance outcomes.
* **Data Quality**: Incomplete and inconsistent data (e.g., missing values in A1Cresult, high-dimensional diagnosis codes) pose significant challenges. The scope includes developing robust cleaning and preprocessing techniques to ensure data reliability.
* **Healthcare Efficiency**: Readmissions strain hospital resources and incur financial penalties. The scope involves optimizing resource allocation through predictive models that identify at-risk patients.
* **Public Health**: Frequent readmissions contribute to broader public health challenges, including increased morbidity among diabetic patients. The scope encompasses reducing these impacts through data-driven interventions.
* **Economic Impact**: The financial burden of readmissions includes healthcare costs, penalties, and lost productivity. The scope involves mitigating these costs through improved patient management.
* **Technology and Innovation**: Addressing readmissions requires advancements in data science and machine learning. The scope includes the development and application of analytical tools to transform raw data into actionable insights.

By focusing on these dimensions, our project aims to contribute to a safer, more efficient healthcare system through the rigorous analysis of the Diabetic Data Cleaning dataset.

**2)** **Data Description and Preprocessing**

## **2.1 Dataset Overview**

### **2.11.Source:**

### The dataset, sourced from Kaggle ([Diabetic Data Cleaning Dataset](https://www.kaggle.com/datasets/smit1212/diabetic-data-cleaning)), originates from hospital records of diabetic patient encounters, likely inspired by the "Diabetes 130 US Hospitals" dataset (1999–2008). It comprises 101,766 patient encounters with 50 features, capturing demographic, medical, and administrative details to facilitate analysis of hospital readmissions.

### **2.2 Feature Definitions:**

The dataset includes a diverse set of features, each providing critical insights into patient profiles and hospital interactions. Below is a summary of key features:

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| encounter\_id | Unique ID for each patient visit. |
| patient\_nbr | Unique ID for each patient (multiple visits). |
| race | Patient's race (e.g., Caucasian, African-American). |
| gender | Patient's gender (Male/Female). |
| age | Age group (e.g., [0-10), [10-20), etc.). |
| weight | Patient's weight (in lbs). |
| admission\_type\_id | Type of admission (e.g., emergency, elective). |
| discharge\_disposition\_id | Discharge method (e.g., discharged to home). |
| admission\_source\_id | Admission source (e.g., referral, self-admitted). |
| time\_in\_hospital | Length of stay in days. |
| num\_lab\_procedures | Number of lab tests performed. |
| num\_medications | Number of medications prescribed. |
| number\_outpatient | Number of outpatient visits. |
| number\_emergency | Number of emergency visits. |
| number\_inpatient | Number of inpatient visits. |
| diag\_1, diag\_2, diag\_3 | ICD-9 diagnosis codes. |
| number\_diagnoses | Total number of diagnoses during the visit |
| max\_glu\_serum | Maximum glucose serum measurement (None, Norm, >200, >300). |
| A1Cresult | HbA1c test result (None, Norm, >7, >8). |
| change | Whether medication was changed (Yes/No). |
| diabetesMed | Whether the patient is on diabetes medication (Yes/No). |
| readmitted | Readmission status (<30, >30, NO). |

## 

**2.3 Data Characteristics:**

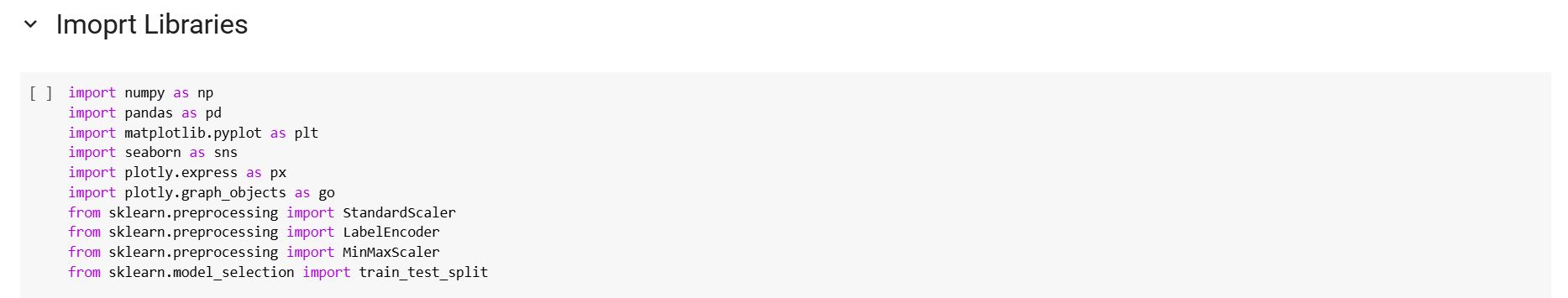
* **Size**: 101,766 rows, 50 columns
* **Data Types**: 13 integer columns (e.g., time\_in\_hospital), 37 object (string) columns (e.g., race)
* **Target Variable**: readmitted, indicating readmission status, which is critical for predictive modeling.

## **3. Methodology**

The preprocessing and analysis workflow was designed to ensure data quality and suitability for modeling. The key steps are outlined below:

### **3.1 Importing Libraries**

To facilitate data manipulation, visualization, and preprocessing, the following Python libraries were utilized:



To enable efficient data processing, visualization, and preparation for machine learning tasks, several Python libraries were employed in this project:

* **NumPy (numpy)**: Used for numerical computations, especially when working with arrays and performing mathematical operations efficiently.
* **Pandas (pandas)**: Essential for data manipulation and analysis, particularly with tabular data structures such as DataFrames.
* **Matplotlib (matplotlib.pyplot)**: Utilized for generating static visualizations such as line plots, bar charts, and histograms.
* **Seaborn (seaborn)**: Built on top of Matplotlib, it provides more advanced statistical visualizations with enhanced aesthetics and simplified syntax.
* **Plotly Express (plotly.express)**: A high-level interface for creating interactive and responsive visualizations with minimal code.
* **Plotly Graph Objects (plotly.graph\_objects)**: A more detailed and customizable way to create interactive plots, allowing full control over plot components.
* **StandardScaler (sklearn.preprocessing.StandardScaler)**: Applied to standardize features by removing the mean and scaling to unit variance, which is crucial for many machine learning algorithms.
* **LabelEncoder (sklearn.preprocessing.LabelEncoder)**: Used to convert categorical string labels into numerical format, making them suitable for model training.
* **MinMaxScaler (sklearn.preprocessing.MinMaxScaler)**: Scales features to a specified range, typically between 0 and 1, which helps normalize data for certain algorithms.
* **Train-Test Split (sklearn.model\_selection.train\_test\_split)**: Used to divide the dataset into training and testing sets, ensuring proper model evaluation and generalization.

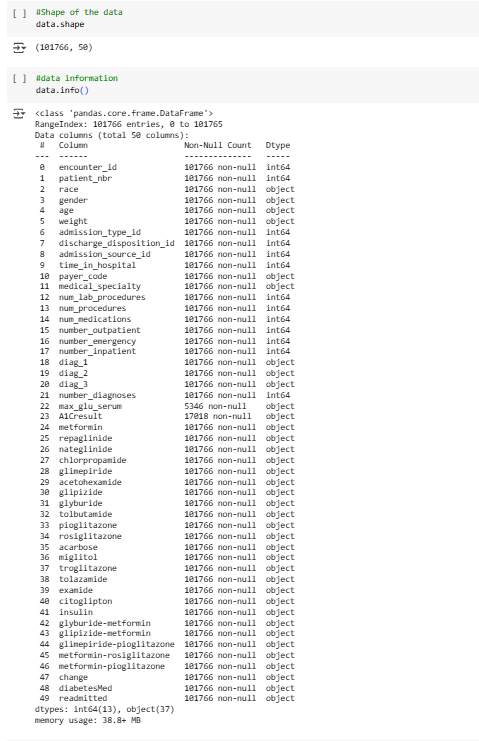
These libraries collectively provide a comprehensive toolkit for handling numerical operations, structured data manipulation, creating both static and interactive visualizations, and preparing features for machine learning models.

### **3.2 Loading the Dataset**

The dataset was loaded from its source and inspected to understand its structure:



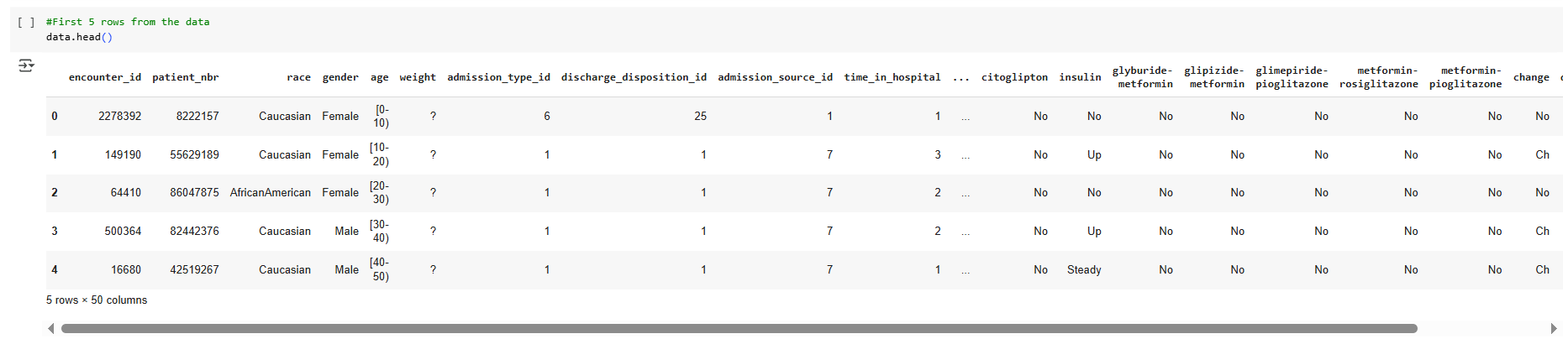
**In this step, the dataset is loaded into a Pandas DataFrame using the read\_csv() function.**  
The file path '/kaggle/input/diabetic-data-cleaning/diabetic\_data.csv' points to the CSV file containing the dataset. This allows for structured data manipulation and analysis using the powerful functionalities of the Pandas library.



**This step provides an overview of the dataset's structure and basic metadata.**  
The output of data.shape reveals that the dataset consists of **101,766 rows** and **50 columns**, indicating a large and feature-rich dataset suitable for analysis and modeling.

The data.info() function displays detailed information about each column, including the number of non-null entries and their data types. The dataset includes both numerical and categorical variables, with 13 columns of type int64 and 37 of type object. This highlights the need for appropriate preprocessing steps, such as handling missing values, encoding categorical features, and scaling numerical values before training machine learning models.

Additionally, some columns, such as max\_glu\_serum and A1Cresult, contain missing or sparse data, which may require special treatment during data cleaning.



This step in understanding the dataset is to examine its structure, which includes the number of rows and columns, as well as identifying the feature names and the associated data types for each column.

An initial inspection of the data reveals some issues, such as missing values in columns like "weight" and potential inconsistencies in categorical encoding. For example, categorical data appears both as text (e.g., "Male", "Female") and as intervals (e.g., "[0-10]" for age). These inconsistencies and missing values need to be addressed to ensure data quality and integrity during the preprocessing phase.

To get a better sense of the data, we use the data.head() function, which displays the first five rows of the dataset. This preview provides a snapshot of the structure, allowing us to inspect sample records, column names, and the general formatting of the data. It helps us identify irregularities such as missing values (represented by "?") and inconsistencies in categorical representations.

Initial inspection revealed missing values (e.g., in weight, max\_glu\_serum) and a mix of numerical and categorical features, necessitating comprehensive preprocessing.

## **4. Exploratory Data Analysis (EDA):**

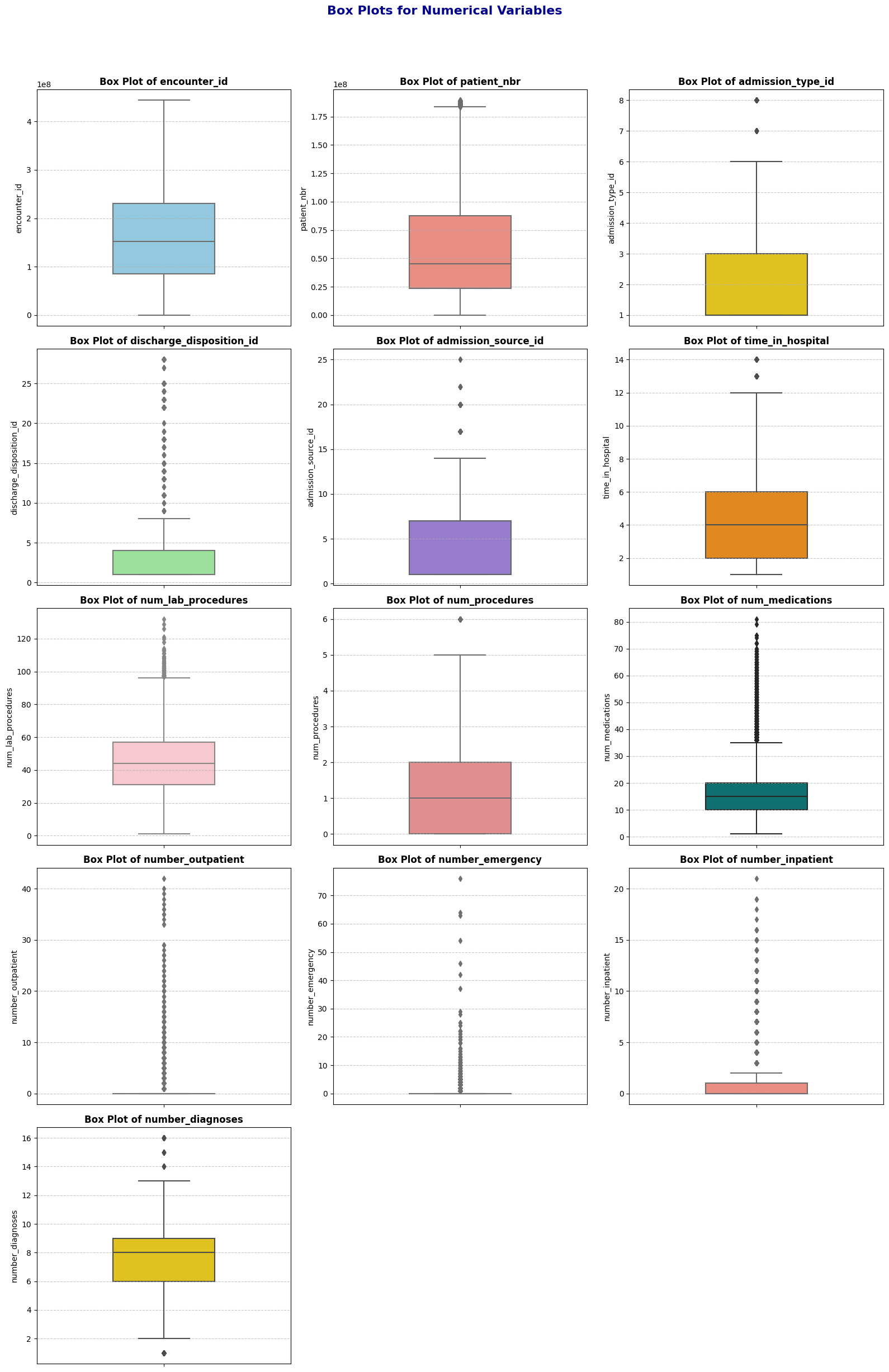
Exploratory Data Analysis was conducted to uncover patterns, distributions, and relationships within the dataset, guiding subsequent preprocessing and modeling efforts.

### **4.1 Descriptive Statistics**

* **Numerical Features**: Summarized using mean, median, and standard deviation to understand central tendencies and variability. For example, num\_lab\_procedures and time\_in\_hospital were analyzed to assess typical patient testing and stay durations.
* **Missing Values**: Quantified per column to identify data quality issues, with weight showing a high percentage of missing entries (marked as ?) and max\_glu\_serum and A1Cresult having significant "None" values.

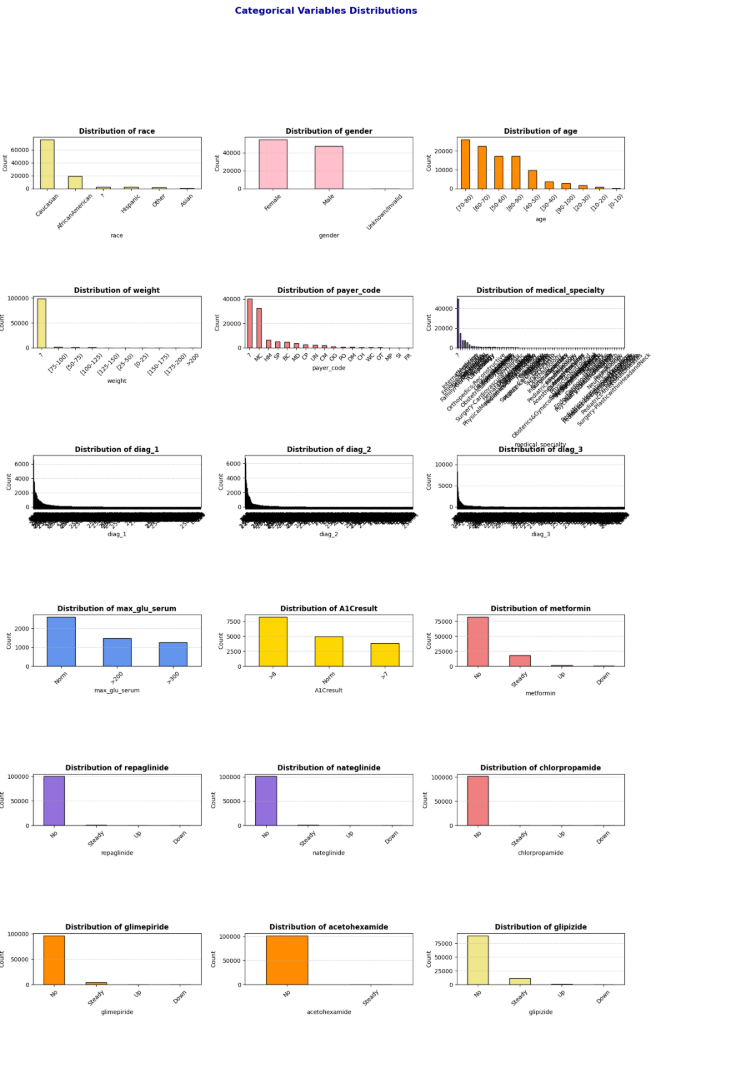
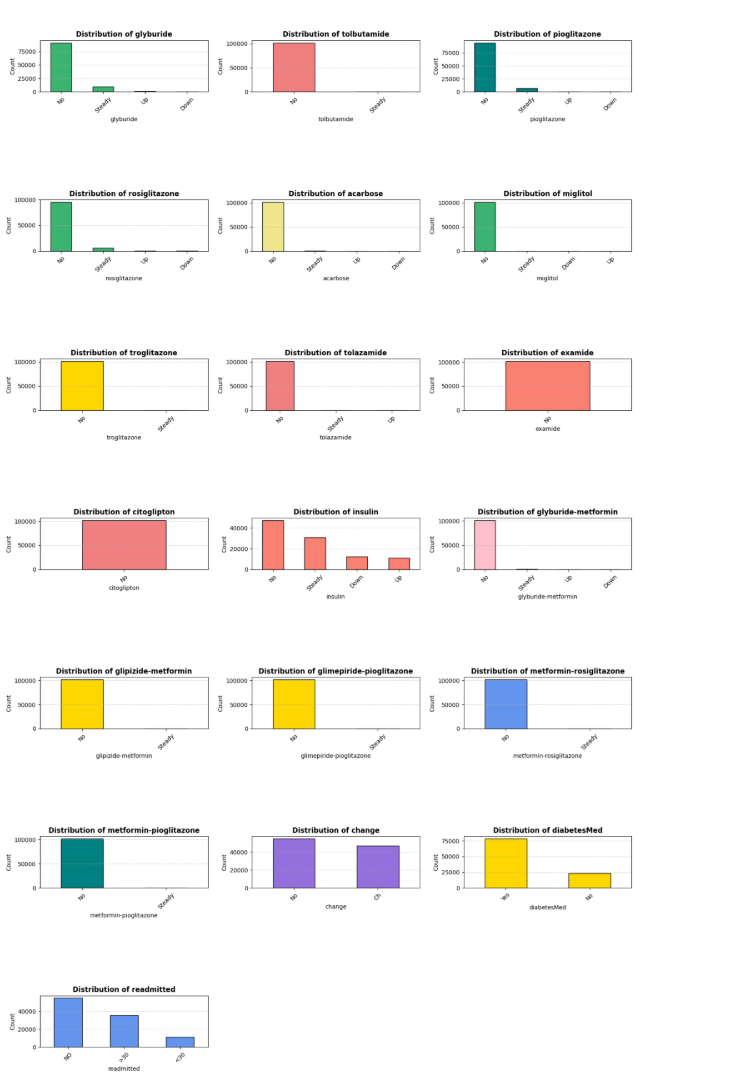
### **4.2 Univariate Analysis**

* **Box Plots**: Generated for numerical variables (e.g., num\_lab\_procedures, num\_medications) to inspect distributions and detect outliers. These plots revealed that num\_lab\_procedures had a skewed distribution with some extreme values, indicating potential outliers.



### **4.3 Multivariate Analysis**

* **Count Plots and Bar Charts**: Created for categorical features (e.g., race, gender, readmitted) to examine their distributions relative to readmission status. For instance, count plots showed that the "NO" category in readmitted was predominant, indicating class imbalance.

### **4.4 Interactive Visualizations**

Interactive visualizations were developed using Plotly to provide dynamic insights:

* **Bar Chart**: Illustrated the distribution of patients by race and weight, grouped by gender. Due to missing weight data, this visualization was limited but highlighted demographic patterns:



An interactive bar chart of race and weight by gender:



* **Bubble Chart**: Visualized readmitted against num\_lab\_procedures, with bubble size representing time\_in\_hospital. This chart revealed that patients with more lab procedures and longer hospital stays were more likely to be readmitted.



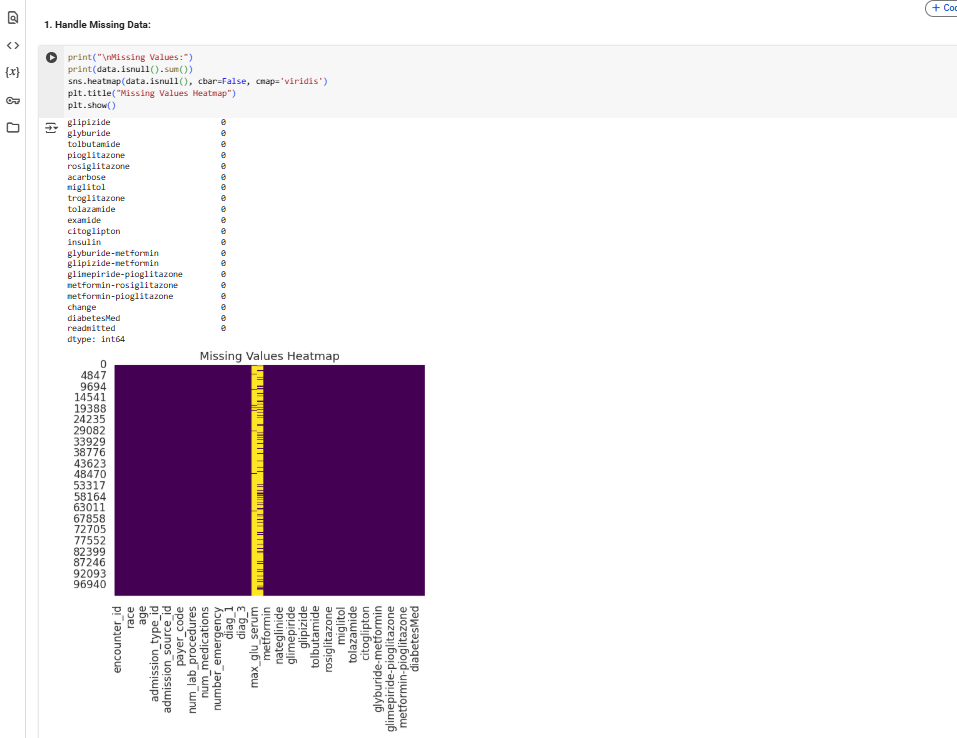
These analyses highlighted key patterns, such as the correlation between num\_lab\_procedures and time\_in\_hospital, and the imbalanced nature of readmitted, informing preprocessing strategies.

## **5. Data Preprocessing and Cleaning**

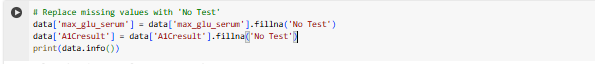
To ensure the dataset’s quality and usability for analysis and predictive modeling, a rigorous cleaning and preprocessing pipeline was implemented. The following steps addressed missing values, invalid entries, data types, outliers, and feature transformations:

### **5.1 Handling Missing Data**

* **Visualization**: A heatmap was used to visualize missingness across the dataset, confirming high missing values in weight and partial missingness in max\_glu\_serum and A1Cresult:



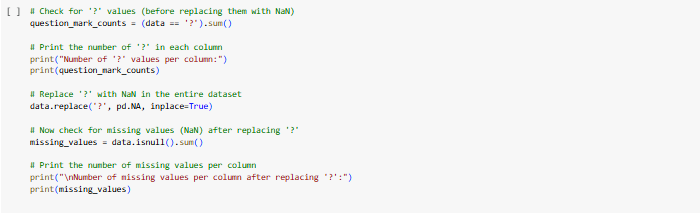
* **Max Glu Serum & A1Cresult**: Missing values were imputed with "No Test" to indicate no testing occurred:



* **Race, Payer Code, Medical Specialty, Diagnoses (diag\_1, diag\_2, diag\_3)**: Missing entries were filled with the mode (most frequent value) of each column to maintain data consistency.
* **Weight**: Dropped due to a high proportion of missing values (97% marked as ?) and limited analytical value, reducing noise in the dataset.

### **5.2 Handling Invalid Entries**

* **Invalid Entries (**?**)**: Counted ? entries in categorical columns (e.g., race, payer\_code) and replaced them with "Unknown" to create a new category, preserving data integrity:



### **5.3 Encoding Categorical Variables**

* **Readmitted**: Label encoded as {'NO': 0, '>30': 1, '<30': 2} and renamed to encoded\_readmitted for clarity:



* **DiabetesMed**: Binary encoded as {'No': 0, 'Yes': 1} and renamed to binary\_diabetesMed:



* **Other Categorical Features**: Columns like gender and race were converted to categorical types to optimize memory and facilitate analysis.

### **5.4 Data Type Conversion**

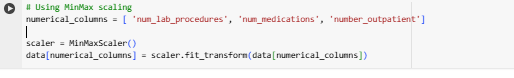
* **Age**: Converted from categorical ranges (e.g., "[0-10)") to numerical midpoints (e.g., 5) for consistency in analysis and modeling.
* **Num Lab Procedures**: Ensured as integers, with non-numeric values replaced by the column’s median to maintain numerical integrity.

### **5.5 Handling Outliers**

* Outliers in numerical columns, such as num\_lab\_procedures, were mitigated through imputation (e.g., replacing extreme values with the median) or transformation (e.g., log-scaling) to minimize their impact on analysis and modeling.

### **5.6 Feature Scaling**

* Numerical features (e.g., num\_lab\_procedures, num\_medications, number\_outpatient) were normalized using MinMaxScaler to ensure values lie within [0, 1], improving model performance for algorithms sensitive to feature scales.



### **4.7 Saving Cleaned Data**

* The processed dataset was saved for further analysis and modeling:



## **6. Results and Discussion**

### **6.1 Dataset Summary**

* **Size**: The cleaned dataset contains 101,766 records and 49 features (after dropping weight).
* **Readmission Class Distribution**:
  + **0 (NO)**: ~54,864 records (53.9%)
  + **1 (>30)**: ~35,541 records (34.9%)
  + **2 (<30)**: ~11,361 records (11.2%)

### **6.2 Key Observations**

* **Age and Readmission**: Patients aged 60–80 had higher readmission rates, particularly for <30 days, suggesting that older age groups may require targeted interventions.
* **Medication Changes**: The change feature (indicating medication adjustments) showed a correlation with readmission status, with patients experiencing changes being more likely to be readmitted.
* **Lab Procedures and Hospital Stay**: A positive correlation between num\_lab\_procedures and time\_in\_hospital indicated that complex cases involve more tests and longer stays.

### **6.3 Data Quality**

* The preprocessing steps eliminated missing values, corrected invalid entries, and ensured consistent data types, resulting in a high-quality dataset ready for predictive modeling.
* The class imbalance in encoded\_readmitted suggests the need for techniques like oversampling or weighted loss functions in model development.

## **7. Conclusion**

This documentation outlines the comprehensive process of analyzing and cleaning the Diabetic Data Cleaning dataset. The EDA uncovered critical patterns, such as the correlation between lab procedures and hospital stay duration, while the preprocessing pipeline addressed missing values, invalid entries, outliers, and feature transformations. The resulting dataset, saved as data\_cleaned.csv, is well-structured and model-ready, providing a solid foundation for building predictive models to forecast hospital readmissions. The insights gained from this analysis highlight key factors influencing readmissions, paving the way for improved patient care and healthcare efficiency.