Assignment: Data Analytics Process and Interpretation

Student : K.A.D.K.D. Dharmathilake | 21020191 | 2021/IS/019
Repo Link:
https://github.com/Kavithma-Dharmathilake/BIS_Assignment_Data_Analytics_Process_and_Interpretation/tree/main
Dataset: Diabetes 130-US Hospitals (1999–2008), UCI Machine Learning Repository
https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

Business Domain: Healthcare Analytics / Hospital Management

1. Introduction

The dataset covers 1999–2008 hospital records from 130 US hospitals for patients with diabetes, including labs, medications, and stays up to 14 days. The goal is to analyze and predict early readmission within 30 days. Inconsistent diabetes care often leads to readmissions, higher hospital costs, and increased patient morbidity and mortality.

Business Context: As a data analyst within a hospital network, the focus is on supporting management to reduce 30-day readmission rates, which is critical because:

- Readmissions significantly increase hospital costs.
- They negatively impact hospital quality ratings.
- They affect insurance reimbursements and financial incentives.

Business Question: Which patient, clinical, and hospital factors influence 30-day readmission among diabetic patients, and can we identify patients at high risk of early readmission?

2. Data & Analytical Process

1. **Understanding the Dataset and Dimensions:** Size: 100,000+ encounters, Features: Demographics, hospital stay details, clinical measures, medications, and readmission outcome. It has a total of 50 columns with 47 features.
2. **Data Preprocessing:**
    a. Handled missing values.(Dropped columns with more than 90% missing values, Rename the missing values with new label, Compute the missing value with mode)
    b. Encoded categorical variables numerically or grouped into meaningful categories(This dataset includes lot of categorical data)
    c. Mapped ICD-9 diagnosis codes to broad disease groups (By referring this link special encoding was applied for diagnostic data)
    d. Checked for duplicates and removed redundant rows (Duplicated rows and Columns with very less variations(99% similar values in columns) were dropped as such features were not helpful)
3. **Descriptive Analysis:**
    1. Univariate Analysis: Examined individual feature under four categories (Patient Identifiers & Demographics, Hospital Stay Details, Clinical Measures & Diagnostic Data,Medication Features)

2. **Bivariate / Multivariate Analysis:** Explored relationships between features to identify potential patterns and correlations:
    a. Numeric vs Numeric
    b. Categorical vs Categorical
    c. Numeric vs Categorical

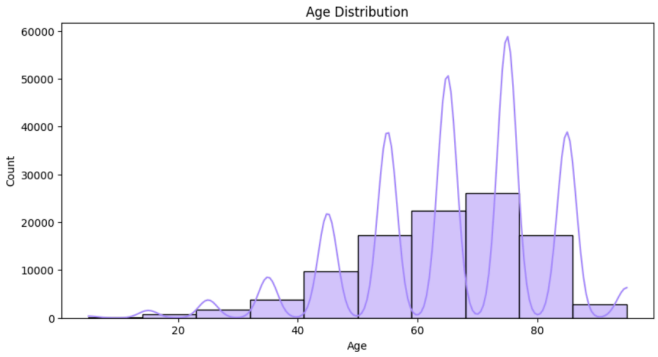
3. **Target-Focused Analysis (Readmission):** Examined how all features relate to the outcome variable (readmitted), highlighting trends, high-risk patient segments, and key predictors.

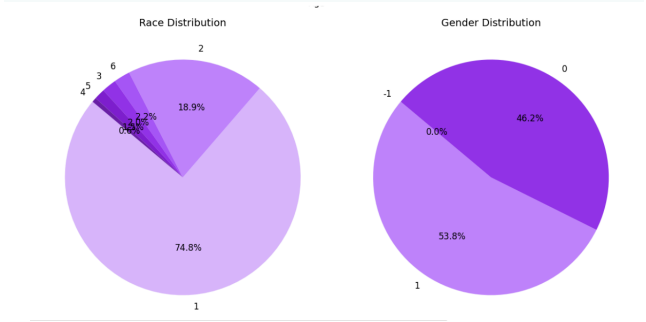
4. **Visualization & Interpretation**

More through explanation with application can be found in the Colab Notebook file in the repository

3.Key Visualization

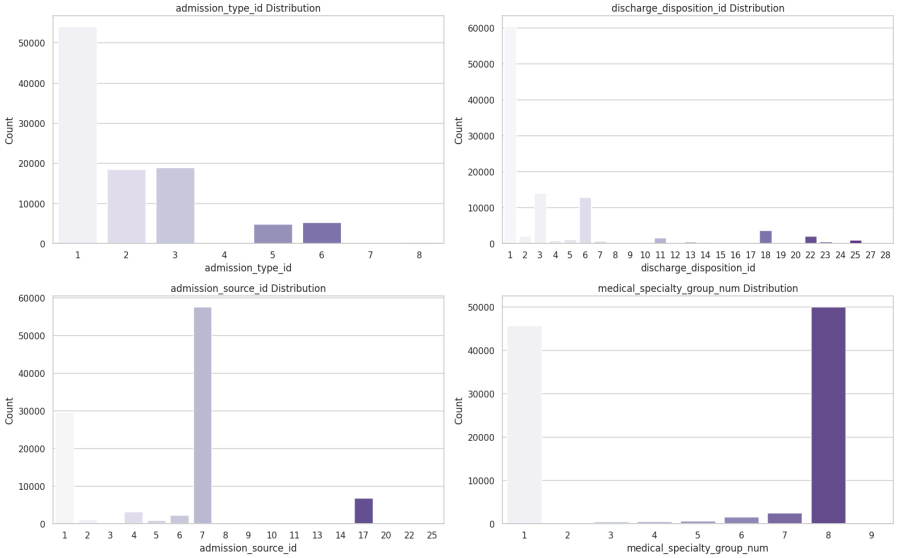
4.1 Demographics - Univariate Analysis





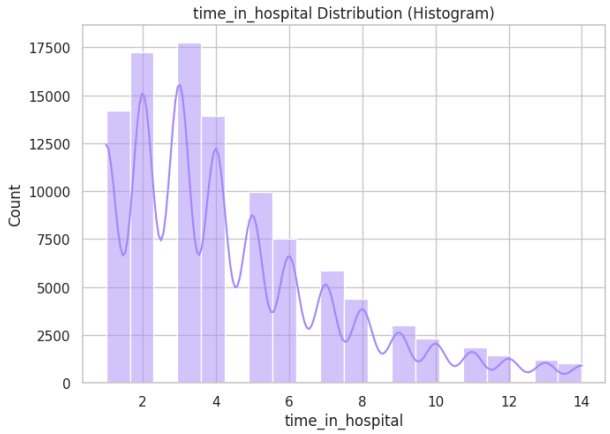
The highest number of patient encounters are between 60-80 in age. Most encounters with Caucasian while least from Asians. The number of encounters in females is higher than male.

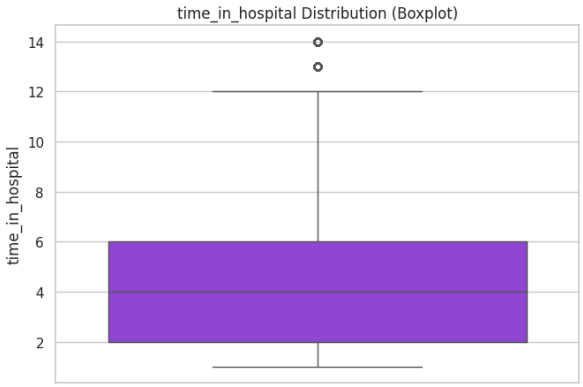
4.2 Hospital Stay & Admission - Univariate Analysis



- The dominant admission type was 1 (Emergency Admission). This hospital likely handles a lot of unplanned or acute encounters.
- Looking at admission_source_id, ID 7 ("Emergency Room") and ID 1 ("Physician Referral") are the high values. This reinforces that most patients are coming in through the Emergency Room.

- In the discharge_disposition_id chart, ID 1 ("Discharged to home") is the majority.
- The medical_specialty_group_num shows a bipolar distribution. Most patients fall into Group 1(Infectious and parasitic diseases) or Group 8(Respiratory system). This often indicates a data split where most patients are either under any of the categories mentioned.

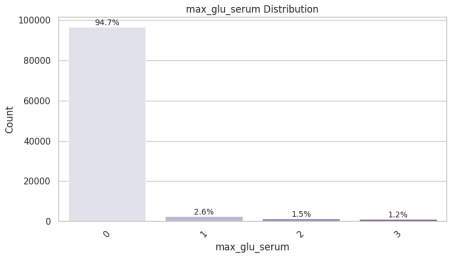


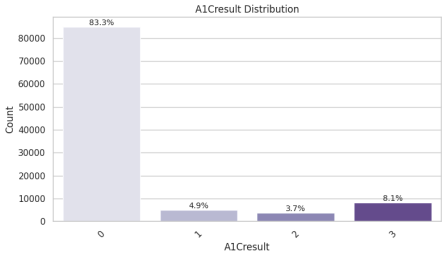


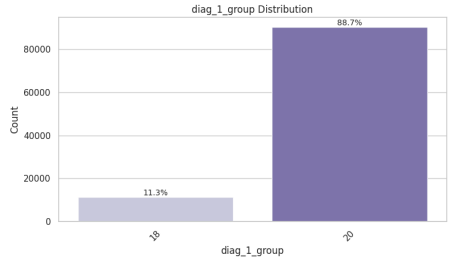
The histogram shows that the majority of patients stay in the hospital for a short duration, specifically between 2 and 4 days. The highest peak (mode) occurs at 3 days.According to the box plot half of all patients are discharged within 4 days. The tail of the histogram stretches far to the right, meaning as the number of days increases, the number of patients significantly decreases. This is typical for medical data where most cases are routine, but a few are highly complex.

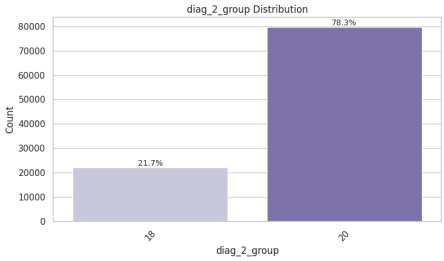
More numerical features were analysed and can be accessed through the Notebook file in GitHub Repository.

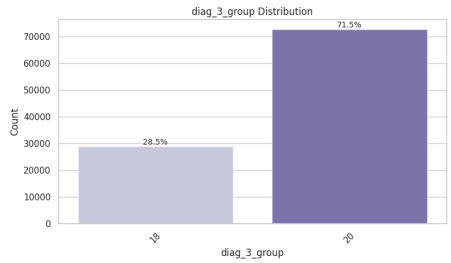
4.3 Clinical Measures & Diagnoses - Univariate Analysis









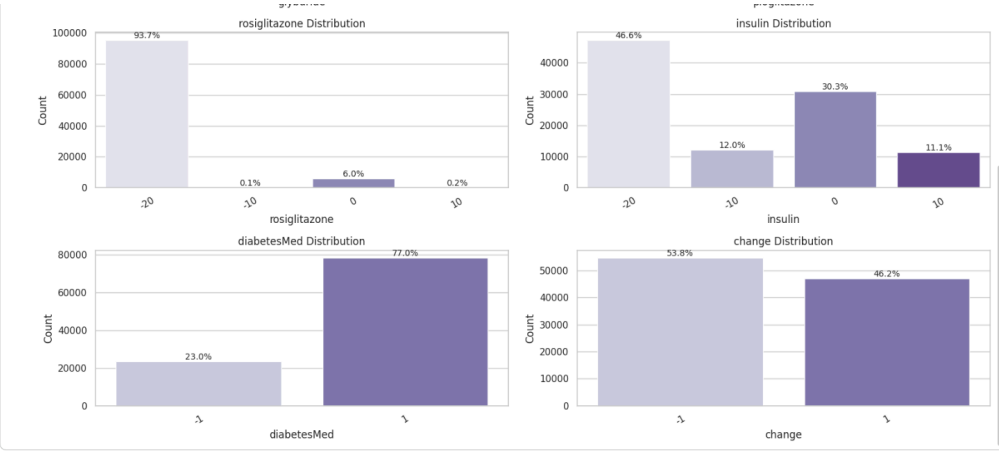


- Most patients stay between 2 and 6 days, with the median stay being 4 days.

- A massive majority of patients did not have a Max Glucose Serum test (94.7%) or an A1C result (83.3%) recorded in this dataset.A1C Results: For the small percentage of patients who were tested, 8.1% returned a result categorized as "3" (typically indicating high risk or >8%).

- Across all three diagnostic categories (diag_1, diag_2, and diag_3), Group 20(Other) is consistently the most frequent.

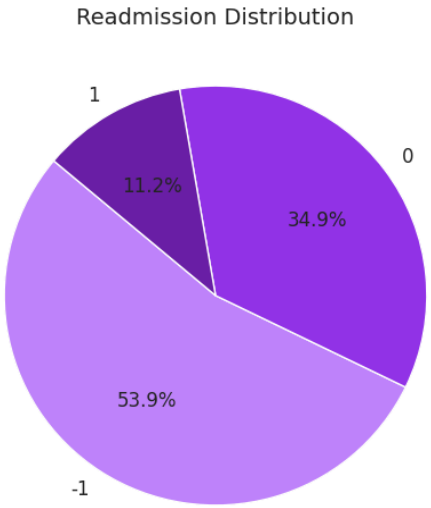
4.4 Medications - Univariate Analysis



- Insulin is the most utilized specific medication, with 53.4% of patients having it as part of their treatment plan (combining steady, increased, or decreased dosages).
- Most oral diabetes medications (like Metformin, Glipizide, and Glyburide) are not used by 80%–98% of the population in this dataset.

- Metformin is the most common oral option, used by 19.6% of patients.
- There is a near-even split in treatment stability: 46.2% of patients experienced a change in their diabetic medication dosage or type during their stay, while 53.8% remained on a stable regimen.

4.5 Target Variable



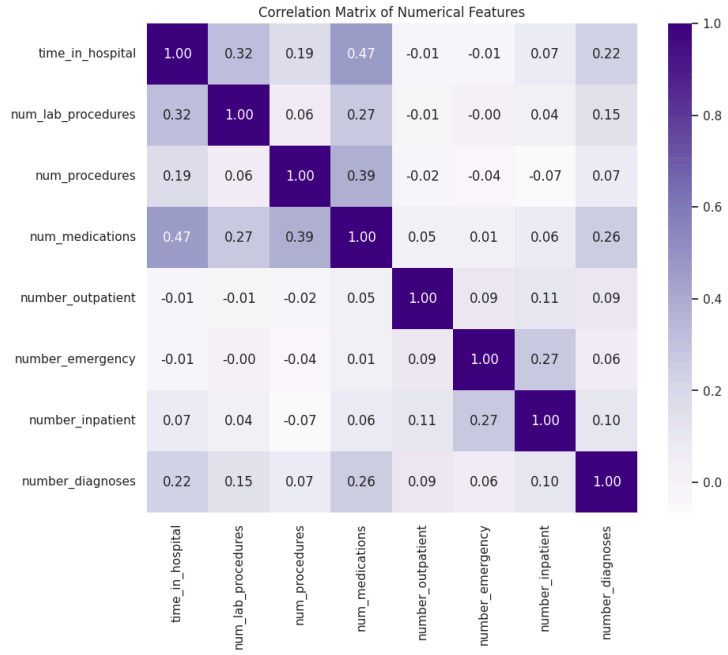
The Readmission Distribution pie chart reveals the core challenge in the dataset.

1. Less than <30 days': 1,
2. Greater than >30 days: 0,
3. Never Admitted again: -1

- Only 11.2% of patients fall into Class 1 (Readmitted in <30 days). This is the most critical class to predict in healthcare but is significantly underrepresented.
- Over a third of patients (34.9%) fall into Class 0 (Readmitted after> 30 days).
- No Readmission: The majority of patients (53.9%) are in Class -1, meaning they were never readmitted during the study period.

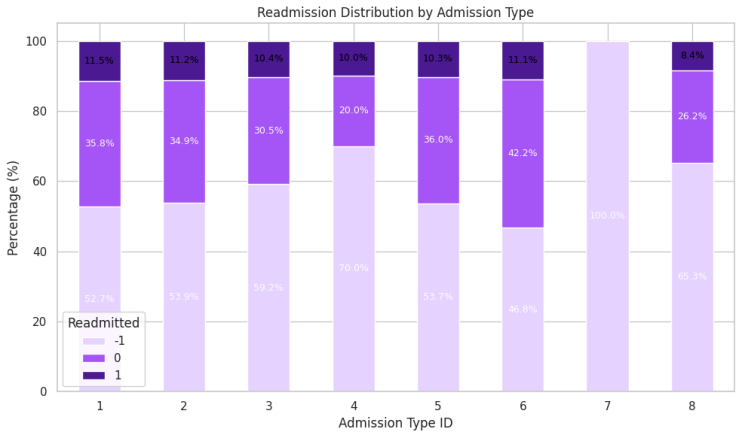
The target variable is significantly imbalanced, meaning the categories are not distributed equally.

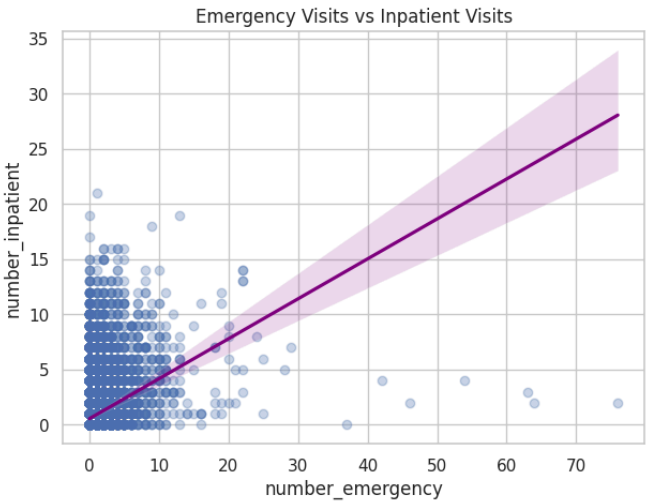
4.6 Numerical vs Numerical Features - Multivariate Analysis



- There is a strong relationship (0.47) between time_in_hospital and num_medications. This confirms that the longer a patient remains in care, the more complex their pharmacological regimen becomes.
- A correlation of 0.39 between num_procedures and num_medications indicates that clinical interventions are frequently paired with medical treatments. A correlation of 0.32 identified between time_in_hospital and num_lab_procedures, along

- Admission Types 1 (Emergency), 2 (Urgent), and 5 (Trauma Center) show remarkably similar readmission profiles.
- In these common admission types, the high-risk group (Class 1, readmitted <30 days) consistently hovers around 11.2% to 11.5%.
- These groups are dominated by patients who are never readmitted (Class -1), making up approximately 53% to 54% of their respective totals.
- Admission Type 7 is a total outlier as 100.0% of recorded cases resulted in no readmission.

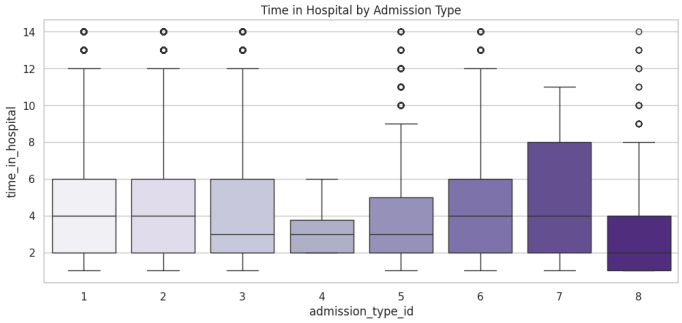




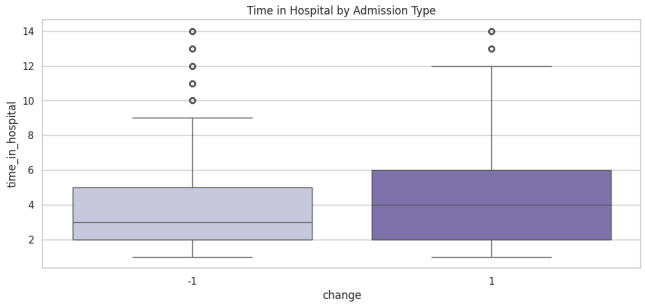
- The upward-sloping purple regression line indicates a positive relationship between the two variables. The regression line suggests that as the number of emergency visits increases, the number of inpatient visits typically follows suit. This makes sense clinically. Patients who frequently require emergency intervention are more likely to have chronic or acute conditions necessitating full hospital admission.

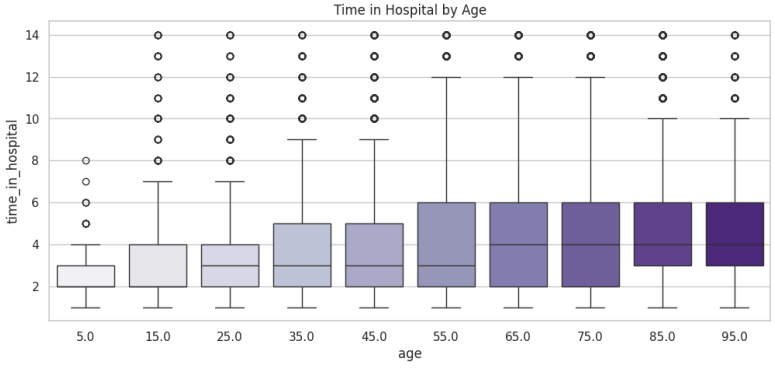
- The dense cluster near the origin (0,0) shows that most patients have very few of either visit type, but as you move along the x-axis, the spread on the y-axis increases, showing the predictive difficulty at higher visit counts.

4.7. Numerical vs Categorical Features - Multivariate Analysis



- Admission Type 7 shows the highest median stay (~4 days) with a much larger IQR, reaching up to 8 days. Types 1, 2, and 6 also maintain a consistent median of 4 days.
- Admission types 3, 5, and 8 generally have shorter stays, with medians typically around 2 to 3 days. Type 8 shows the most compact distribution, with the bulk of patients leaving very quickly.
- Type 4 has a very narrow range, indicating highly predictable stay durations.



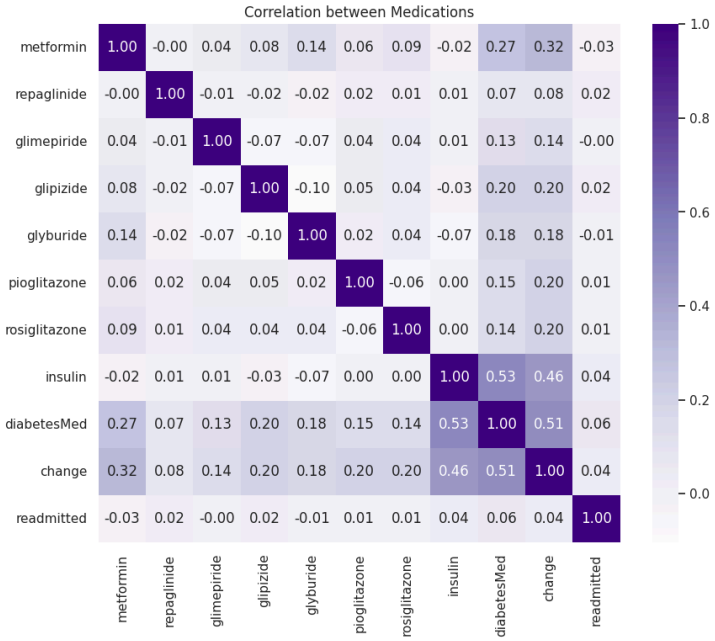


Age and Hospital Duration: There is a clear, incremental relationship between age and the length of stay. Patients under 50 typically have a median stay of **3 days**, whereas patients aged 65 and older see that median rise and stabilize at **4 days**.Older age groups (75–95) show much larger Interquartile Ranges (IQR) and higher "whiskers," indicating that elderly patients are far more likely to experience prolonged, complex hospitalizations.

Medication Changes vs. Stay Length: Patients who underwent a change in their diabetic medication (represented by **1**) stay longer than those whose medication remained stable (**-1**).

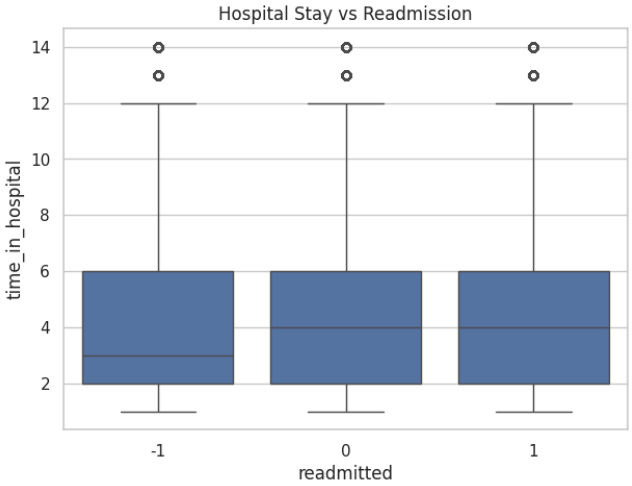
4.7. Numerical vs Categorical Features - Multivariate Analysis

- DiabetesMed & Insulin (0.53): This is the strongest correlation in the chart. It indicates that patients prescribed diabetes medication are highly likely to specifically be on an insulin regimen.
- DiabetesMed & Change (0.51): There is a strong link between being on diabetes medication and experiencing a "change" in treatment (dosage or drug type) during the hospital stay.



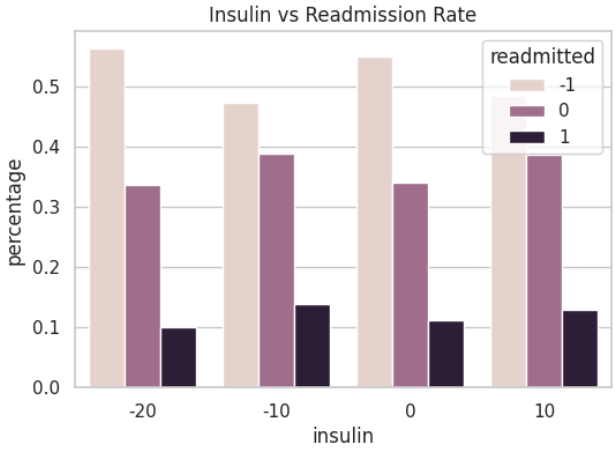
- Insulin & Change (0.46): Insulin use is a major driver of treatment adjustments, suggesting that insulin levels are frequently titrated or modified during inpatient care.
- Metformin's Influence: Metformin shows the strongest relationship with treatment change (0.32) and the presence of diabetes medication (0.27) among the oral drug group.
- No single medication has a strong direct correlation with readmission. This suggests that "which drug" a patient takes is not a strong independent predictor of readmission; instead, readmission is likely driven by a complex combination of features rather than a single medication type.

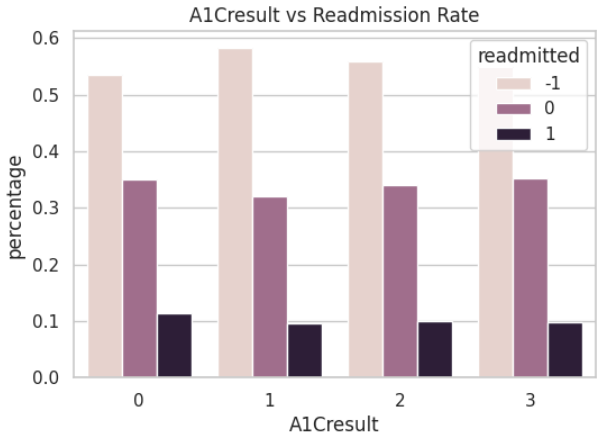
4.7. Target Variable - Multivariate Analysis

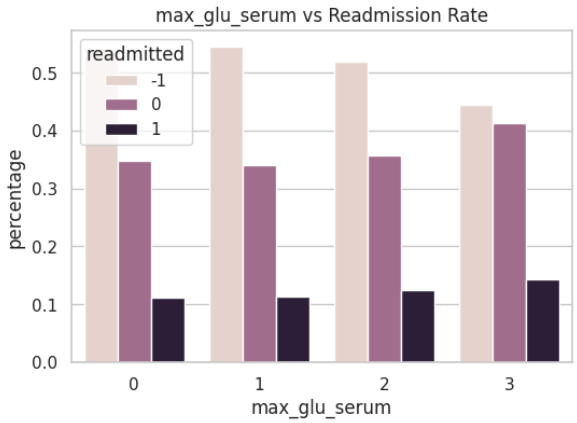


- For all three categories (Class 1: <30 days, Class 0: >30 days, and Class -1: NO), the **median stay is 4 days** and the interquartile range (IQR) remains centered between **2 and 6 days**.
- The nearly identical distributions suggest that **length of stay alone is not a strong discriminator** for predicting which specific readmission class a patient will fall into.

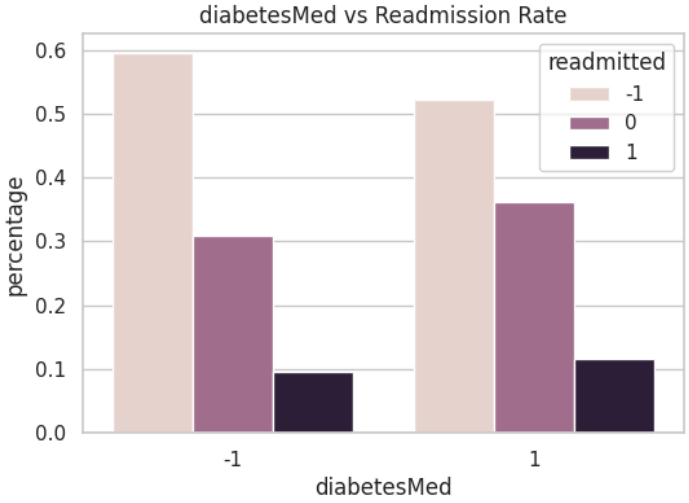
- Patients who had their insulin dosage **decreased (mapped as -10)** or **increased (mapped as 10)** show a higher percentage of early readmissions compared to those whose insulin was not prescribed (-20) or remained steady (0).
- Patients not on insulin or on a steady dose have the highest rates of **no readmission** , while any titration (up or down) noticeably increases the likelihood of a return visit within 30 days.

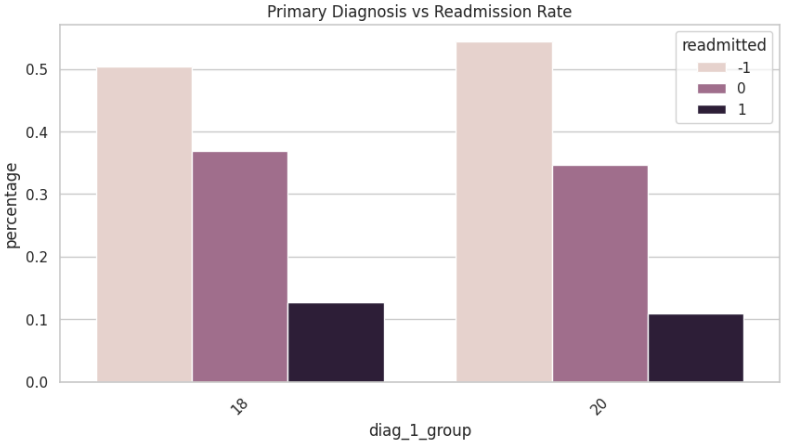






- Patients with a result of **1 (Norm)** show the lowest percentage of early readmissions (Class 1). Those with a result of **0 (None)** or higher levels like **2 (>200)** and **3 (>300)** exhibit slightly higher, comparable risks of returning within 30 days.
- There is a clearer upward trend in readmission risk as glucose levels increase. Patients with a result of **3 (>300)** have the highest percentage of early readmissions (Class 1) compared to those with normal or no recorded results.
- For both tests, patients with no result recorded (**0**) or a normal result (**1**) consistently have the highest rates of **no readmission (Class -1)**.





Diabetes Medication and Readmission: Patients prescribed diabetes medication (marked as **1**) show a higher percentage of early readmissions (Class 1) compared to those not on medication (**-1**). Conversely, patients not on diabetes medication have a higher rate of **no readmission (Class -1)**, exceeding **0.58** compared to approximately **0.52** for medicated patients.

Primary Diagnosis and Readmission: Patients admitted with a primary diagnosis in **Group 18** show a higher likelihood of early readmission (Class 1) than those in **Group 20**. While both groups have a majority of patients who are not readmitted, **Group 20** has a higher percentage of stable, non-returning patients (Class -1) at over **0.54**.

5. Key Findings

5.1. Major Findings

1. Early readmission is notably higher among patients who:
    a. Have a primary diagnosis in Group 18 compared to Group 20.
    b. Undergo insulin titration (dosage increases or decreases) rather than a stable or non-insulin regimen.
    c. Are prescribed diabetes medication generally, showing higher risk than unmedicated patients.
    d. Present with high Max Glucose Serum results (>300).
2. There is a clear positive relationship between the number of emergency visits and subsequent inpatient admissions, suggesting that frequent ER users are a core high-risk demographic.
3. Hospital stays increase in duration and complexity with age, specifically for patients aged 65 to 95, who have a median stay of 4 days and higher variability in care.
4. **Target Imbalance:** The high-risk "Class 1" (<30 day readmission) is a minority at 11.2%, while 53.9% of patients are never readmitted. This makes identifying the specific "Class 1" characteristics critical for financial incentives.

5.2. Operational Findings

The data reveals how the hospital currently functions and where information gaps exist:

- 94.7% of patients lacked a Max Glucose Serum test and 83.3% had no A1C result.
- Most encounters are for Caucasian females aged 60–80, admitted via the Emergency Room (Admission Type 1, Source 7), staying for a median of 4 days, and discharged to home.
- There is a strong relationship (0.47) between time in the hospital and the number of medications, confirming that longer stays involve higher clinical complexity and resource use.
- Interestingly, the median stay (4 days) is identical across all readmission categories, meaning time in hospital alone cannot predict if a patient will return early.

5.3. Implications for Business Decisions

To reduce readmissions and protect insurance reimbursements, management should consider the following strategic shifts:

- Since insulin adjustments (up or down) are linked to higher readmission, the hospital should implement enhanced discharge counseling or 48-hour follow-up calls specifically for patients whose insulin dosage was changed during their stay.
- The 83%+ gap in A1C and Glucose testing represents a significant missed opportunity for risk assessment. Hospital should mandate these tests for all diabetic admissions to provide the objective data needed for better predictive modeling.
- Patients with a primary diagnosis in Group 18 (External Causes / Injury) should be flagged for a multidisciplinary review before discharge, as they show a higher return rate than Group 20 (Other).
- Because prior Emergency Room visits predict inpatient admissions, the hospital could save costs by investing in outpatient diabetic clinics to manage "heavy users" before they require emergency care.

6. Conclusion

In this study, a systematic descriptive analysis using statistical methods ranging from univariate to multivariate exploration was carried out on the UCI US Diabetes database to identify the factors influencing 30-day readmission. By examining the relationships between patient demographics, hospital stay details, and clinical measures, key insights were extracted to define high-risk patient segments.

The analysis ultimately reveals that to effectively reduce readmission rates, the hospital must move beyond "standard care" for the 60–80 age bracket. Instead, management should implement targeted monitoring and enhanced discharge protocols specifically for patients with fluctuating insulin levels and high-frequency Emergency Room histories, as these factors serve as the most significant predictors of early return.

7. References / Appendix

- Dataset: UCI Diabetes Dataset
- Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
- ICD-9 Classification: Wikipedia