Diabetes Hospital Readmission

Presentation 12

Kusum Sai Chowdary Sannapaneni (G25941197)

Srivallabh Siddharth Nadadhur (G32021287)

Visualization of Complex Data

Professor. Sarah Gates

May 3rd, 2024

Table of contents

Introduction

Diabetes is a prevalent chronic condition globally, affecting an estimated 422 million adults and leading to significant morbidity and mortality. The management of diabetes is fraught with challenges, particularly the prevention of acute complications that can result in hospital admissions and readmissions. Hospital readmissions are not only distressing for patients but also impose substantial financial burdens on healthcare systems. Given the growing prevalence of diabetes and the associated healthcare costs, there is an urgent need for strategies that effectively reduce the incidence of hospital readmissions. This project focuses on leveraging advanced predictive analytics to identify diabetic patients at risk of readmission, aiming to provide timely interventions that can enhance patient outcomes and optimize healthcare resources.

Motivation

The motivation for this research arises from the critical need to improve healthcare outcomes for diabetic patients who are particularly prone to hospital readmissions. These readmissions often lead to worsened health outcomes, increased mortality rates, and higher healthcare costs. With the shift towards value-based care models, reducing unnecessary hospital readmissions has become a priority for healthcare providers and payers alike. Employing predictive modeling techniques to identify high-risk patients early in their care journey enables healthcare practitioners to implement preventive measures and personalized care plans. This proactive approach not only improves the quality of life for patients but also aligns with economic incentives to enhance the efficiency of healthcare delivery systems. By addressing the root causes and predictors of readmissions, the project seeks to contribute to the broader goals of sustainable healthcare practices and improved patient care for the diabetic population.

Business Questions

The project is structured around the following robust business questions, which guide the investigation and the application of data analytics:

1. **How do demographic variables such as age and gender influence the likelihood of readmission among diabetic patients?**
- This question aims to uncover patterns between demographic factors and readmission rates, providing insights that could lead to more personalized patient care strategies.

2. **Is there a relationship between the length of hospital stay and readmission rates among diabetic patients?**
- Understanding this relationship can help in optimizing the duration of hospital stays to minimize risks without compromising patient recovery, thus efficiently utilizing healthcare resources.

3. **Does the number of in-hospital procedures correlate with the readmission rates of diabetic patients?**
- Investigating this correlation will reveal whether the complexity or intensity of care (as reflected by the number of procedures) influences the likelihood of a patient's readmission, which in turn could guide clinical decision-making processes.

These questions are designed to dissect various facets of the hospital readmission issue among diabetic patients, leveraging data-driven insights to propose actionable strategies for improvement.

Description of Dataset

The dataset employed in this study is a comprehensive collection of data from 130 U.S. hospitals, focusing on the readmission of diabetic patients over a ten-year period from 1999 to 2008. Sourced from the UC Irvine Machine Learning Repository, this dataset includes 101,766 records and 50 distinct columns, encompassing a wide array of variables. These variables provide detailed insights into various facets of patient care and management, including demographic information (age, gender), clinical parameters (lab test results, number of procedures), admission and discharge details, diagnosis codes, and the primary outcome of interest—readmission status.

Source : UCI Machine Learning Repository

Data Collection and Access

Access to this dataset is facilitated through the UC Irvine Machine Learning Repository, a well-regarded source for machine learning datasets used by researchers worldwide. The repository provides this dataset under specific terms of use, which generally include requirements for proper citation and adherence to ethical guidelines concerning patient confidentiality and data usage. This ensures that the data is not only accessible for research and educational purposes but also handled with the utmost integrity and respect for patient privacy. Each entry in the dataset represents a unique patient admission event, with detailed records of their treatment during the stay. The rich granularity of the dataset allows for an in-depth analysis of patterns and trends in hospital readmissions, specifically for the diabetic patient population.

Analysis of Data Quality

The data quality is critical for ensuring the reliability and validity of the research findings. A data dictionary is provided to explain each variable, enhancing understanding and transparency regarding the data.

Data Dictionary

Feature	Description
Encounter ID	Unique identifier of an encounter
Readmitted	**Days to inpatient readmission. Values: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.**
Race	Values: Caucasian, Asian, African American, Hispanic, and other
Gender	Values: male, female, and unknown/invalid
Age	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)
Weight	Weight in pounds.
Admission type	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Discharge disposition	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Admission source	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
Time in hospital	Integer number of days between admission and discharge
Payer code	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
Medical specialty	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
Number of lab procedures	Number of lab tests performed during the encounter
Number of procedures	Number of procedures (other than lab tests) performed during the encounter
Number of medications	Number of distinct generic names administered during the encounter
Number of outpatient visits	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Number of emergency visits of the patient in the year preceding the encounter

Number of inpatient visits	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Diagnosis 2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
Diagnosis 3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
Number of diagnoses	Number of diagnoses entered to the system
Glucose serum test result	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
A1c test result	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured
Change of medications	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
Diabetes medications	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
24 features for medications	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed

Table.1 contains the description of each variable in the dataset

50 columns 101,766 records

This data dictionary serves as a crucial tool for understanding the variables within the dataset, their significance, and the completeness of the data, ensuring a robust analysis framework for the research study.

Variables of Interest: Description and Analysis

In our research focusing on the predictors of hospital readmission among diabetic patients, we have selected a subset of variables from the comprehensive dataset. These variables are critical for understanding the factors influencing readmission rates. Below is a detailed explanation of both the qualitative and quantitative aspects of these variables.

1. Age:

- Type: Nominal (Categorical)
- Description: Age of the patients grouped into 10-year intervals (0-10, 10-20, ..., 90-100).
- Qualitative Information: The age groups are distinct categories with no intrinsic ranking apart from their natural numerical order. Each category represents a decade of life.
- Quantitative Information: While the data is categorical, the midpoint of each range might be used for certain types of analysis to approximate the age numerically.

2. Gender:

- Type: Nominal (Categorical)
- Description: Gender of the patient recorded as male, female, or unknown/invalid.
- Qualitative Information: There are three categories with no ranking. This variable is essential for analyzing demographic differences in readmission rates.

3. Readmission:

- Type: Nominal (Categorical)
- Description: Indicates if the patient was readmitted, and if so, whether it was within 30 days ("<30"), after 30 days (">30"), or not at all ("No").
- Qualitative Information: The categories here are distinct and critical for the outcome of the study. They are used to classify the readmission status, which is the primary variable of interest.

4. Time in Hospital:

- Type: Numeric (Continuous)

- Description: The total number of days the patient stayed in the hospital during the encounter.
- Quantitative Information: This variable is measured as an integer count of days. Summary statistics such as mean, median, range, and standard deviation are relevant and provide insights into the typical and extreme durations of hospital stays, which could influence readmission.

5. Number of Procedures

- Type: Numeric (Continuous)
- Description: The total number of medical procedures (other than lab tests) performed during the encounter.
- Quantitative Information: Another continuous integer count, this variable offers a measure of the healthcare services provided during a hospital stay. The summary statistics for this variable reveal the commonality and distribution of medical interventions, which might correlate with readmission rates.

Summary Statistics

Age:

- Most common category: 70-80

Gender:

- Distribution: 54,708 females, 47055 males

Readmission:

- Distribution: 54,864 not readmitted, 35,545 readmitted after 30 days, 11357 readmitted within 30 days

Time in Hospital:

- Mean: 4.3 days
- Min: 1 days
- Max: 14 days
- Standard Deviation: 2.9 days
- Range: 1-14 days

Number of Procedures:

- Mean: 1.3

- Min: 0
- Max: 6
- Standard Deviation: 1.7
- Range: 0-6

These variables form the core of our analytical framework, each contributing essential insights into the dynamics of hospital readmission among diabetic patients. The qualitative and quantitative analyses will help in drawing conclusions about the relationships between these variables and readmission outcomes.

Main Analysis: Python

Data Cleaning and Preparation:

The first step involved reading the dataset and cleaning it to ensure consistency and completeness. We found that some categorical variables, such as gender and readmission status, had unexpected values, which needed to be handled. For example, the gender variable contained an 'Unknown/Invalid' category, which was removed from further analysis. Similarly, the readmission variable was transformed into a binary variable, where 'NO' represented no readmission, and '1' represented any form of readmission, either within or after 30 days.

Descriptive Statistics:

The cleaned dataset consisted of 101,763 records with variables of interest including age, gender, readmission status, length of stay, and number of in-hospital procedures. We generated descriptive statistics for each variable to understand their distribution and central tendencies.

- Age: Patients had an average age of 66 years, with a standard deviation of 16 years. Ages ranged from 5 to 95, with a median of 65, reflecting an older patient population.
- Gender: The dataset had more female (54,708) than male (47,055) patients.
- Readmission: 54,864 patients were not readmitted, while 46,902 were readmitted either within or after 30 days.

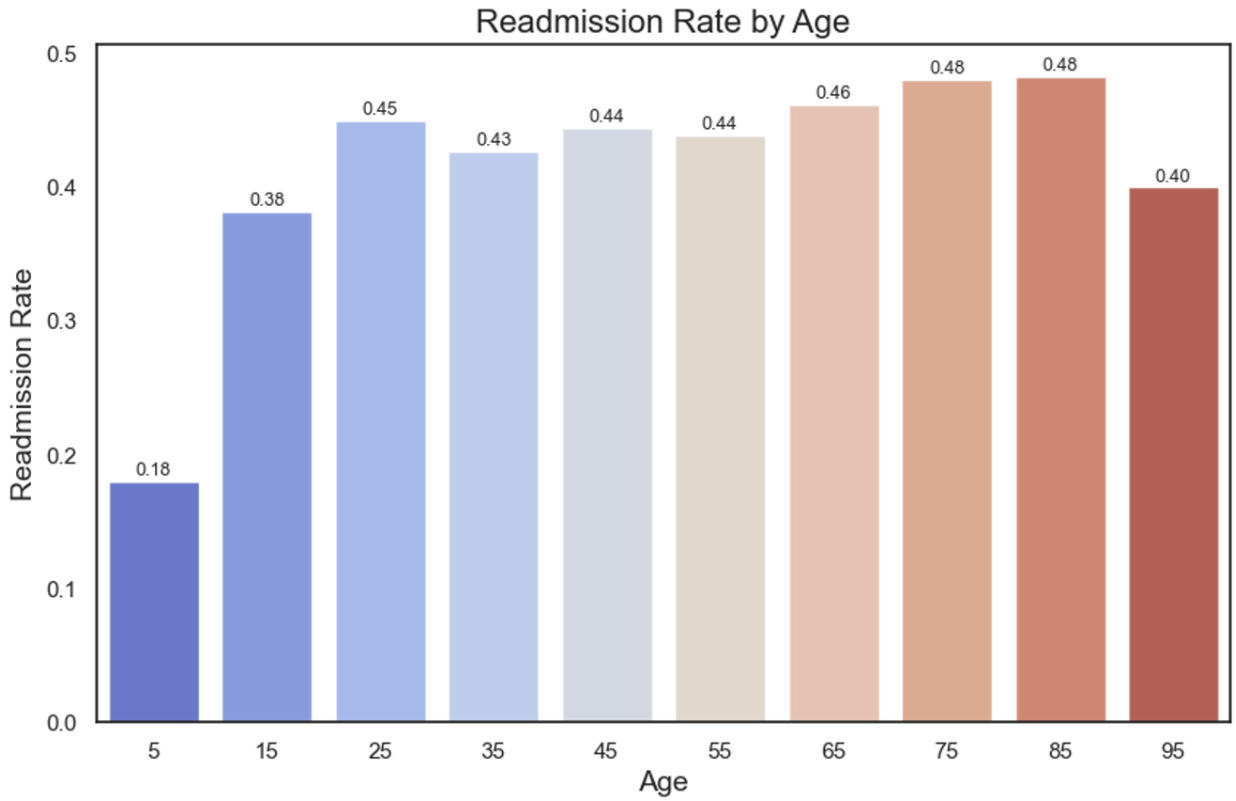
- Time in Hospital: The average length of stay was 4.4 days, with a range from 1 to 14 days, and a median of 4 days.
- Number of Procedures: Patients underwent an average of 1.34 procedures, with half having at least one and a quarter having two or more.

Visualization and Analysis:

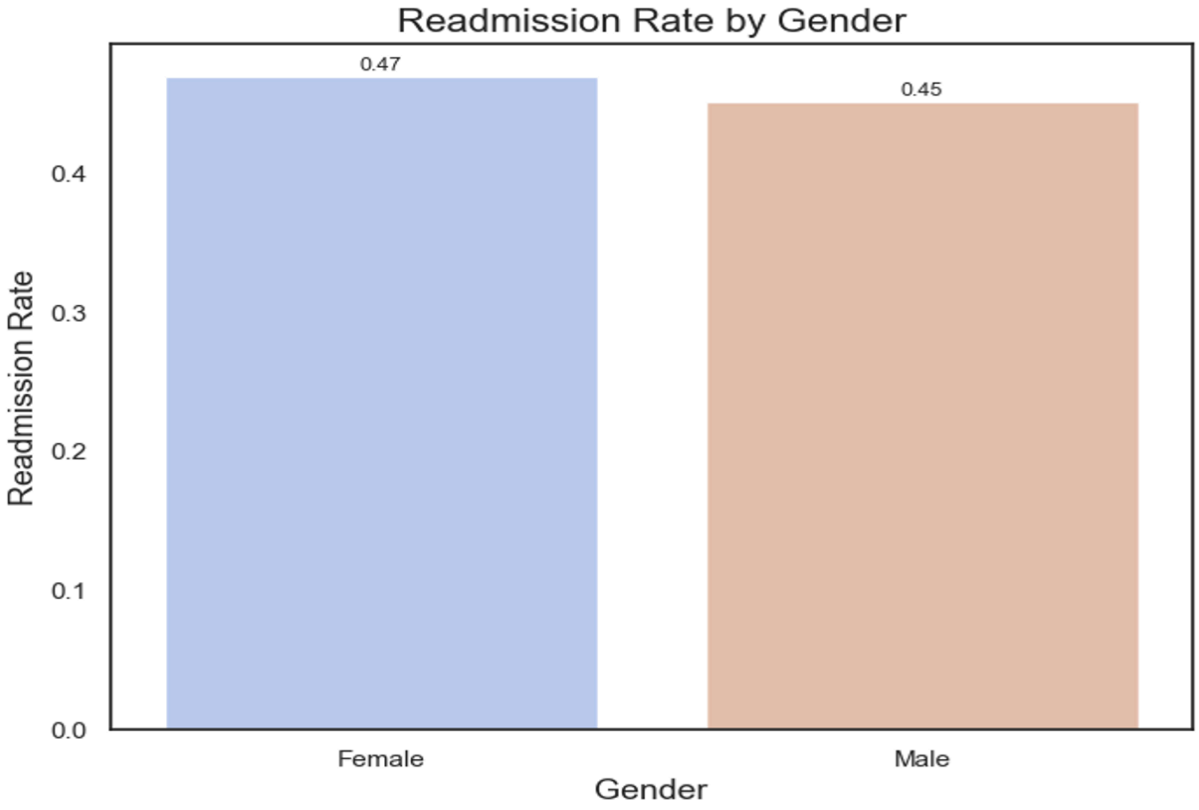
1. Demographic Factors and Readmission Rates (Q1):

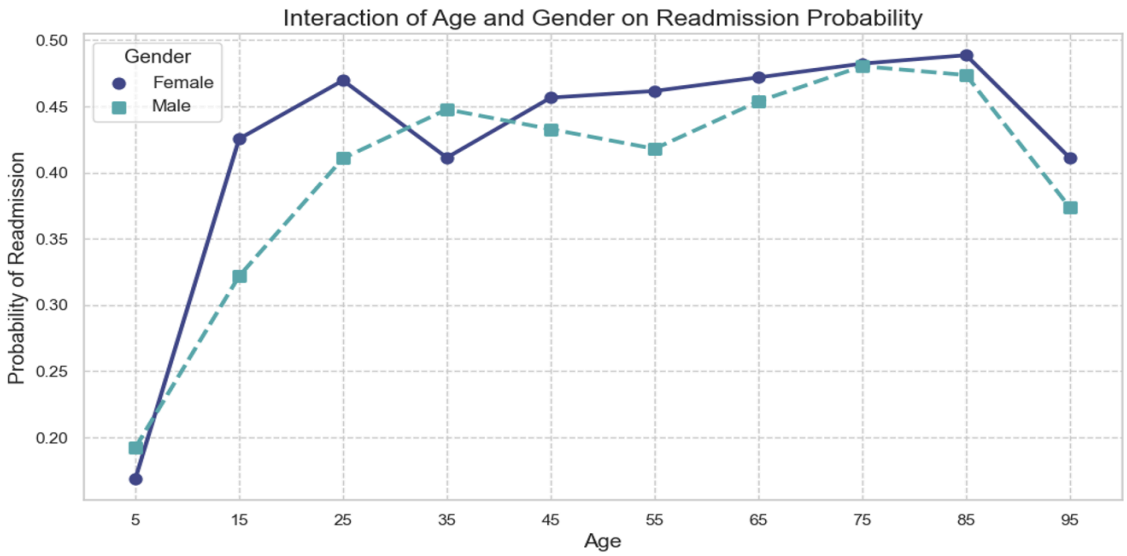
Bar plots and point plots were created to examine the relationship between age, gender, and readmission rates:

- Age: The bar plot showed an increase in readmission rates with age, suggesting older patients have higher readmission risks. The logistic regression analysis confirmed a significant positive relationship between age and readmission probability.



- Gender: The gender plot showed a marginally higher readmission rate for females, but logistic regression indicated males had a slightly lower likelihood of readmission, which contradicted the bar plot's initial observation. A point plot revealed that this difference persisted across all age groups.





Interpretation of the visualizations and how they answer Q1:

Age Influence: The first bar graph demonstrates a clear trend where readmission rates increase progressively with age, indicating that older individuals are at a higher risk of hospital readmission.

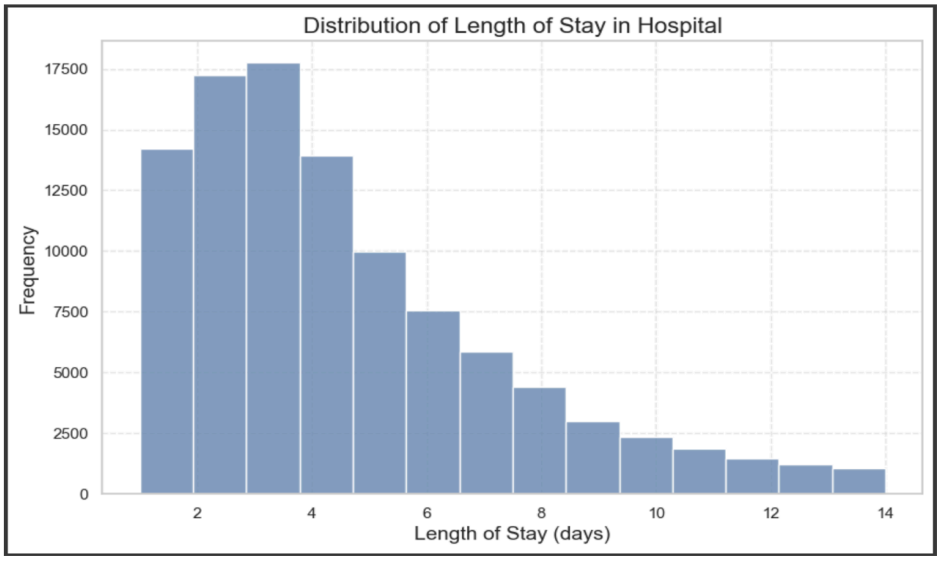
Gender Influence: The second graph shows a slightly higher readmission rate for females compared to males, suggesting gender differences in readmission likelihood.

Combined Influence of Age and Gender: The third graph highlights the interaction between age and gender, showing that both factors together influence readmission rates. While the probability of readmission increases with age for both genders, females consistently exhibit a higher probability of readmission across most age groups.

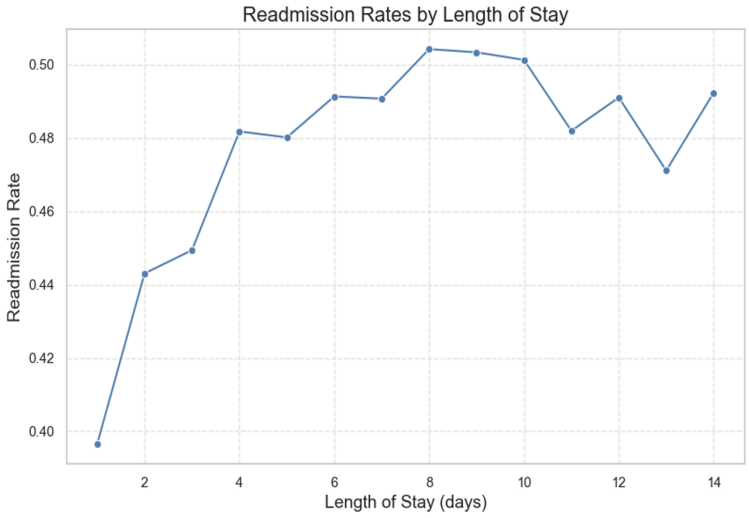
2. Length of Stay and Readmission Rates (Q2):

Several visualizations, including histograms, line plots, hexbin plots, and box and violin plots, were created:

- Histogram: This showed most stays were concentrated between 1 and 6 days, with fewer longer stays.

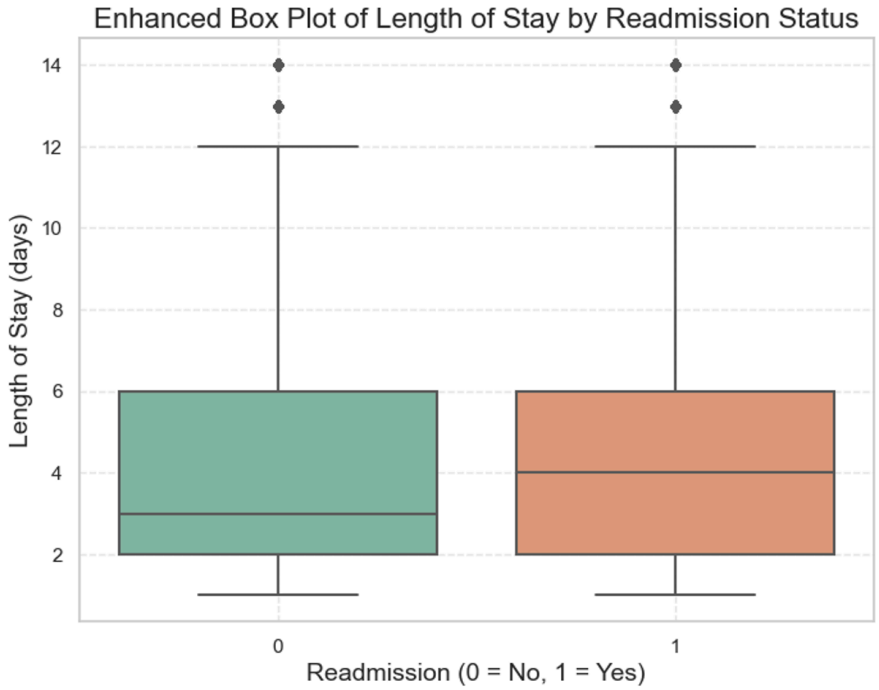


- Line Plot: A trend emerged where readmission rates increased with longer stays, indicating that more extended hospitalizations may signal complex health issues leading to readmission.



- Hexbin Plot: This confirmed the concentration of higher readmission rates around stays of 2 to 4 days, aligning with earlier observations.

- Box Plot: The median length of stay is slightly higher for readmitted patients, indicating a potential relationship between longer stays and readmission rates.



- Violin Plot: The density distribution shows the length of stay with wider sections indicating a higher density of longer stays between 1 to 6 days.

Interpretation of the visualizations and how they answer Q2:

Histogram: Shows the distribution of lengths of hospital stays, with most staying between 1 to 6 days. This sets the context by indicating the typical durations of hospital stays.

Line Plot: Demonstrates an initial increase in readmission rates as the length of stay increases, peaking at 4 days, then stabilizing, and showing some fluctuations. This suggests that longer stays, potentially indicating more severe or complex health issues, are associated with higher readmission rates.

Hexbin Plot: Confirms the line plot insights, showing a moderate concentration of higher readmission rates for stays of 2 to 4 days. It uses color density to show where most data points lie, providing a visualization of the readmission probability across different lengths of stay.

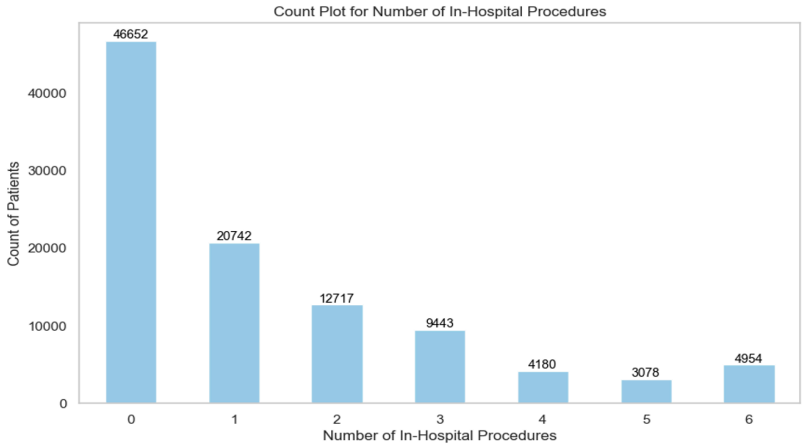
Box Plot and Violin Plot: These plots compare the length of stay by readmission status, with both showing that readmitted patients tend to have longer stays. The box plot highlights the median and spread of stays for readmitted and non-readmitted patients, while the violin plot gives a more detailed view of the density distribution of stays, with wider sections indicating a higher density of longer stays for readmitted patients.

Together, these graphs effectively show that there is indeed a relationship between the length of hospital stay and readmission rates, with longer stays generally associated with higher readmission risks.

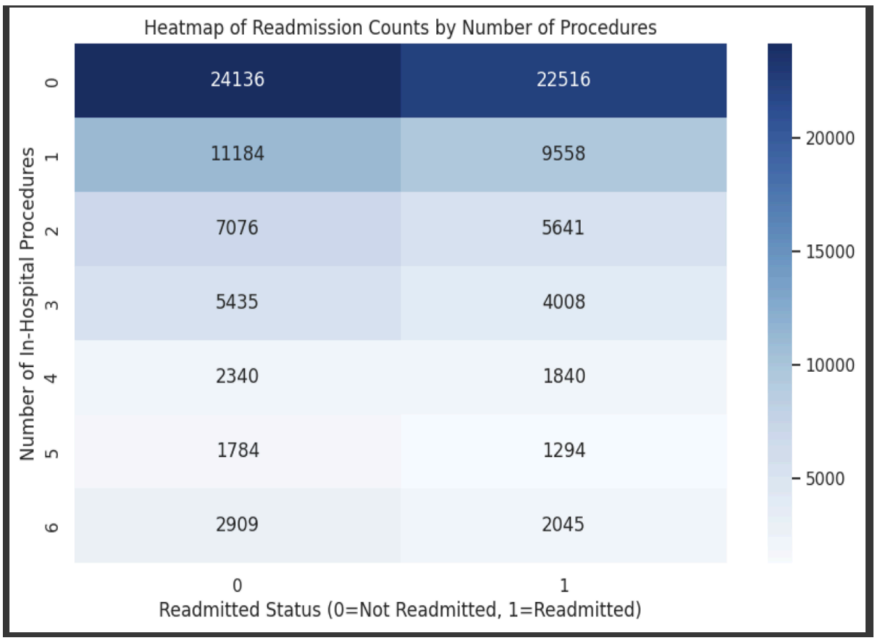
3. Number of Procedures and Readmission Rates (Q3):

We created bar plots, heatmaps, and point plots to explore this relationship:

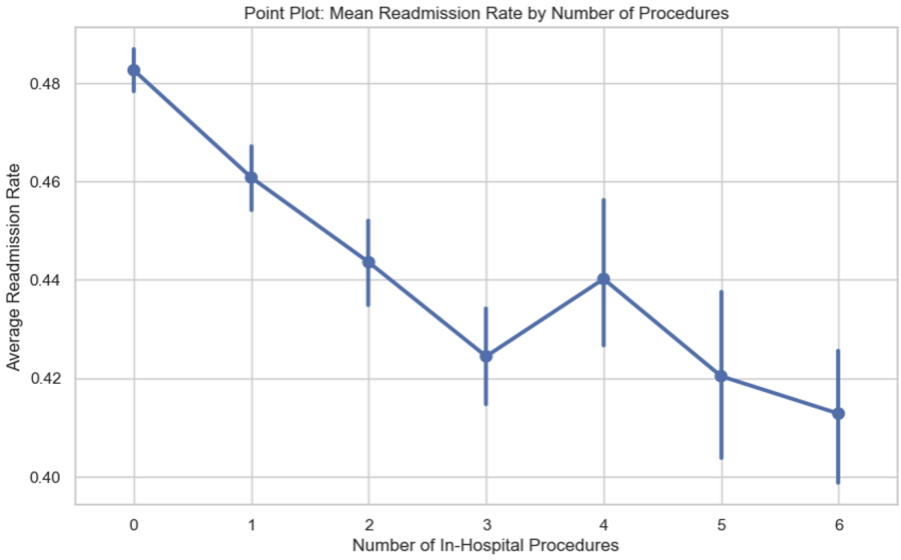
- Bar Plot: This showed a high frequency of zero procedures, with a consistent decrease in patients as procedures increased, suggesting multiple procedures were reserved for severe cases.



- Heatmap: This showed higher readmission rates among patients with at least one procedure, though the correlation analysis indicated a weak negative correlation (-0.045).



- Point Plot: Each point shows the average readmission rate per category, with trend lines illustrating patterns and fluctuations. The lack of a clear trend indicates a complex relationship influenced by multiple variables.



Interpretation of the visualizations and how they answer Q3:

Count Plot: Shows that the majority of patients undergo few or no in-hospital procedures, with a steep decline in frequency as the number of procedures increases. This provides a basis for understanding the typical range of procedures patients undergo.

Heatmap: Illustrates readmission counts by the number of procedures, showing that patients who undergo at least one procedure have higher readmission rates. This suggests a correlation where more procedures might be associated with more complex or severe cases, leading to higher readmission rates.

Point Plot: Displays the average readmission rate by the number of in-hospital procedures. Initially, readmission rates decrease as the number of procedures increases from one to three, but then fluctuate without a clear trend. This indicates a complex relationship where the initial decrease could suggest effective treatment or intervention, but the fluctuations imply other influencing factors.

Together, these graphs suggest that there is a relationship between the number of in-hospital procedures and readmission rates, although the relationship is not strictly linear and may be influenced by the nature and severity of the patient's condition.

Additional Analysis: Tableau

Tableau is a powerful data visualization tool that facilitates the transformation of complex datasets into clear, intuitive visual insights. Its drag-and-drop interface allows users to create diverse charts, graphs, and dashboards that effectively convey data patterns and relationships. Tableau's interactivity enables viewers to explore data details dynamically, making it ideal for presenting comprehensive analyses to both technical and non-technical audiences.

Dashboard 1:

1. Data Preparation: Loaded the dataset into Tableau, providing a foundation for quick visualization creation.

2. Sheets:

- **Age vs. Readmission:** A bar chart and line chart combination visualizing average readmission rates by age groups, with the line chart further segmented by gender.
- **Gender vs. Readmission:** A bar chart comparing readmission rates between male and female patients.

3. Calculated Field: Created a "Readmission Indicator" field using the formula:

IF [Readmitted] = 'NO' THEN 0 ELSE 1 END

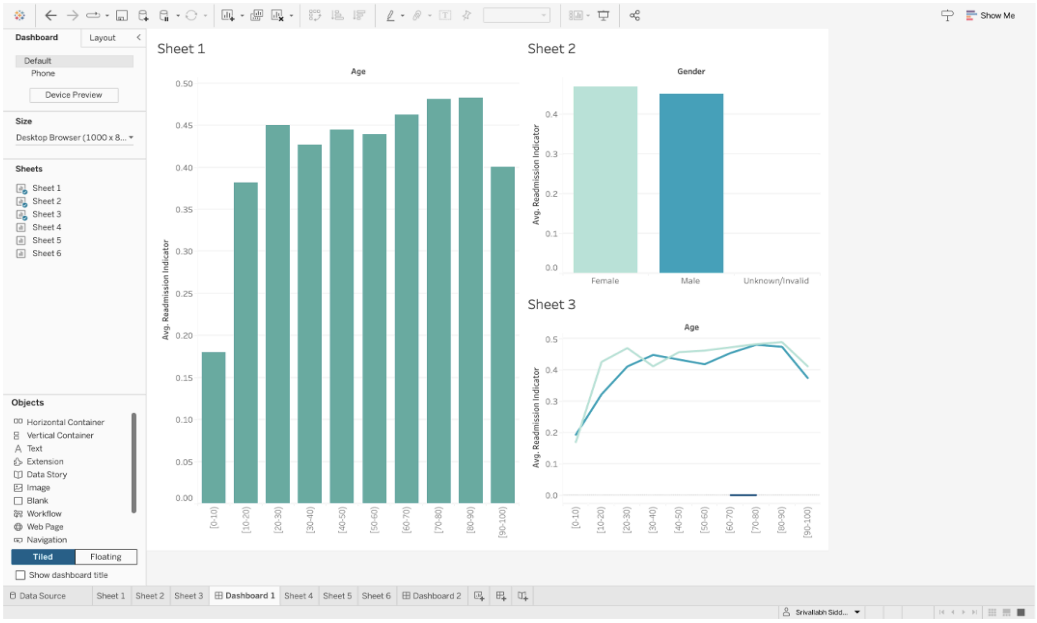
The field was aggregated using "**Average**," enabling readmission rates to be visualized as averages.

4. Interactivity: Enabled "Use as Filter" for each chart, allowing them to interact dynamically, updating based on selections made.

5. Dashboard Assembly: Integrated the sheets into a dashboard, arranging them for an intuitive view of the factors influencing readmission.

Observations from Dashboard 1:

- **Gender Difference:** When clicking on the age 10-20 bracket, a significant difference in readmission rates between male and female patients was observed. Females showed a higher average readmission rate than males in age group.
- **Gender Distribution:** The second chart shows there are more female than male patients in this age bracket, which might contribute to the observed differences in readmission rates.



Dashboard 2:

1. Data Preparation: Dataset loaded similarly to Dashboard 1.

2. Sheets:

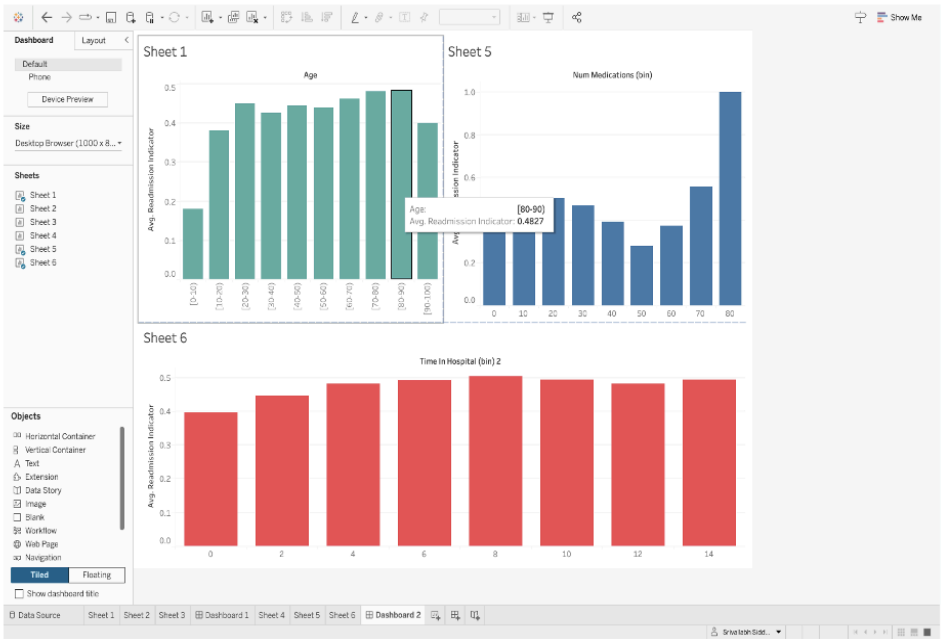
- **Age vs. Readmission:** A bar chart visualizing average readmission rates by age groups.
- **Time in Hospital vs. Readmission:** A bar chart visualizing average readmission rates based on the duration of hospital stays.
- **Num Medications vs. Readmission:** A bar chart visualizing average readmission rates based on the number of medications taken by patients.

3. Interactivity: Enabled "Use as Filter" for each chart, making them interactive and updating dynamically based on selections.

4. Dashboard Assembly: Arranged sheets in a coherent manner, offering an intuitive overview of additional factors influencing readmission.

Observations from Dashboard 2:

- **Medications vs. Age:** When selecting 70 as the number of medications in the Num of Medications vs. Readmission chart, the age distribution for this bracket became visible, ranging from 50 to 80 years. This indicates that higher medication counts are predominantly associated with elderly patients.
- **Time in Hospital:** The Time in Hospital chart revealed that for 70 medications, the hospital stay ranged from 10 days, with no higher values visible before 10 days, suggesting a longer duration of stay correlates with higher med counts.



By interacting with different elements across various graphs within these dashboards, we uncover multiple interesting insights. These insights provide a comprehensive overview of the relationship between patient demographics, medications, and hospital stays, offering valuable information for healthcare analysis and decision-making.

Key Findings:

1. Demographics: Age significantly impacts readmission rates, with older patients having higher risks. Gender also plays a role, though inconsistently, with logistic regression indicating a lower risk for males.

2. Length of Stay: The trend of increasing readmission rates with longer stays indicates that longer hospitalizations reflect more complex health issues, leading to higher readmission probabilities.

3. **Number of Procedures:** The weak negative correlation between procedures and readmission suggests that while patients undergoing multiple procedures are readmitted more frequently, the overall effect size is small, indicating additional factors may play a larger role.

Limitations:

- **Data Preprocessing:** Tableau's limited capabilities posed challenges in handling raw datasets effectively.
- **Class Imbalance:** This created potential biases in the analysis, particularly in correlating readmission with different variables.

Future Directions:

- **Class Balancing:** Addressing class imbalance using techniques like SMOTE could provide a more balanced dataset.
- Predictive Modeling: Further modeling and predictions could enhance our understanding of readmission probabilities.

Lessons Learned:

- **Visualization Techniques:** Choosing appropriate charts helps convey insights effectively, and color and annotations enhance clarity.
- **Complex Relationships:** Understanding the multifaceted relationships between variables requires thorough statistical and visual exploration to avoid oversimplifying findings.

Equal Contributions Throughout the project