Patient Readmission Analysis Technical Report

**Table of Contents**

[Problem 2](#_66sz2oh9jwhy)

[Goal 2](#_722zj61cyp2t)

[Success Metric 2](#_rpzyurwkapxc)

[Audience 2](#_hqkok3a1ybzj)

[Data Source 2](#_5jbkq4s3yin2)

[Data Dictionary 3](#_6a2ksaokna9h)

[Risks & Assumptions 4](#_53clzcwnb055)

[Risks 4](#_f4oejbos6snk)

[Assumptions 4](#_z6dfotd2ck5t)

[Methodology 5](#_qcle6zkbhr7s)

[Exploratory Data Analysis 5](#_chl4fvrt3wpn)

[Numerical Features 6](#_wxs414o7eis2)

[Categorical Features 17](#_r7i0bcxayhrc)

[Summary of the EDA on features to be used in the Predictive Model 27](#_nqaq1vtohw0w)

[Data Cleaning 27](#_ef7iq12hcgbl)

[Feature Engineering 28](#_sce6whjpsash)

[Predictive Model 29](#_gcd11at58zbv)

[Model Evaluation 30](#_6mpil3rbzywp)

[Recommendations 32](#_gdi0osr0jn92)

[Next Steps 32](#_xgzmjxoyp89i)

[Appendix 32](#_e2yiovs6cskd)

# 

# Problem

A hospital has asked us to analyze a dataset of 25,000 patient records to understand its readmission rate. The doctor would like to know if primary diagnosis, time in hospital, and other variables plays a part in a patient's readmission.

* Which is the primary diagnosis with the highest readmission rate?
* Can we predict if a patient will be readmitted within 30-days after discharge?
* Which group of patients the hospital needs to focus its follow-up effort to better monitor patients with high probability of readmission?

# Goal

Using data analysis framework, we would want to achieve the following goals:

1. Analyze the primary diagnosis with the highest readmission rate and its pattern in the dataset.
2. Build a predictive model to predict if a patient will be readmitted within 30-days after discharge.
3. Identify the group of patients the hospital needs to focus its follow-up effort to better monitor patients with high probability of readmission.

# Success Metric

1. Identify the primary diagnosis with the highest readmission rate and its pattern in the dataset.
2. Built a predictive model with a strong Recall value which indicates the correctness of the model identifying patients who were actually readmitted.
3. Identify the group of patients the hospital needs to focus its follow-up effort to better monitor patients with high probability of readmission.

# Audience

The audience of our analysis will be the doctors, nurses, and hospital administrators. The main stakeholder would be the doctors who will decide the next actions from the analysis.

# Data Source

The data source is sourced from [DataCamp’s Competition](https://app.datacamp.com/learn/competitions/hospital-patient-readmissions).

***Acknowledgments****: Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.*

# 

# Data Dictionary

* The columns of the data set will be referred to as features in here onwards.
* There are 17 features (including the target feature to be classified by the prediction model).
  + 7 features are numerical features.
  + 10 features are categorical features.
* There are a total of 25,000 patient records in this data set.
* The missing value in diag\_1, diag\_2, diag\_3, and medical\_specialty has been labeled as ‘Missing’.

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Name** | **Description** | **Type** |
| 1 | age | age bracket of the patient.  **Note:** age groups are defined by a standard convention where the lower bound is **inclusive** and the upper bound is **exclusive**. For example, the 50-60 age group includes all patients who are 50 years old, up to but not including 60. | string |
| 2 | time\_in\_hospital | days (from 1 to 14) | number |
| 3 | n\_procedures | number of procedures performed during the hospital stay | number |
| 4 | n\_lab\_procedures | number of laboratory procedures performed during the hospital stay | number |
| 5 | n\_medications | number of medications administered during the hospital stay | number |
| 6 | n\_outpatient | number of outpatient visits in the year before a hospital stay | number |
| 7 | n\_inpatient | number of inpatient visits in the year before the hospital stay | number |
| 8 | n\_emergency | number of visits to the emergency room in the year before the hospital stay | number |
| 9 | medical\_specialty | the specialty of the admitting physician | string |
| 10 | diag\_1 | primary diagnosis (Circulatory, Respiratory, Digestive, etc.) | string |
| 11 | diag\_2 | secondary diagnosis | string |
| 12 | diag\_3 | additional secondary diagnosis | string |
| 13 | glucose\_test | whether the glucose serum came out as high (> 200), normal, or not performed | string |
| 14 | A1Ctest | whether the A1C level of the patient came out as high (> 7%), normal, or not performed | string |
| 15 | change | whether there was a change in the diabetes medication ('yes' or 'no') | string |
| 16 | diabetes\_med | whether a diabetes medication was prescribed ('yes' or 'no') | string |
| 17 | readmitted | if the patient was readmitted at the hospital ('yes' or 'no')  **Note**: this will be the target feature the prediction model will classify. | string |

# Risks & Assumptions

## Risks

* Changes in the healthcare landscape like treatment protocols and hospital policies over time could affect the readmission rate.
* Implementing the predictive model into the hospitals’ current workflow and ensuring its adoption can be challenging.

## Assumptions

* The dataset of 25,000 records is representative of the broader population of the hospital.
* The features in the dataset are relevant and sufficient for predicting patient readmission.
* The dataset is reasonably correct, without widespread systematic errors or critical missing values that would skew our analysis.
* The ‘Missing’ category in the medical\_specialty feature is a valid and useful signal for our Logistic Regression predictive model.
* The factors that influence patient readmission remain relatively stable over the time period covered in the dataset.

# Methodology

We have used a data analysis framework consisting of framing, preparing, analyzing, and communicating for this project.

In the framing stage, we define the problem of the analysis, goals, and success metrics.

For the preparation and analyzing stage, which is an iterative stage, we have done the following:

1. sourced the data set,
2. define the data dictionary of the data set,
3. perform exploratory data analysis, descriptive analysis, prediction model analysis,
4. perform the data cleaning & feature engineering,
5. build and evaluate the predictive model, and
6. define risks & assumptions.

For the communication stage, we document the findings as a technical report, iterate and storyboard the findings to create a non-technical presentation.

## Exploratory Data Analysis

We have used Tableau to have a better understanding of the features (columns) in the data set.

We have used:

1. Univariate analysis, which analyzes a single feature, to understand the data distribution of the feature and
2. Bivariate analysis, which analyzes two features to see their relationship, to understand if a change in a feature correlates with a change in another feature in this analysis.

Please refer to the [Patient Readmission EDA.twbx](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/03-Analyse/Patient%20Readmission%20EDA.twbx) file to see the charts in detail.

### Numerical Features

1. time\_in\_hospital

Univariate analysis

* In Figure 1.1, the distribution of the data is right-skewed, indicating that while the majority of patients have shorter stays centered around 2-4 days, a smaller number of patients have significantly longer hospitalizations.
* In Figure 1.2, the majority of the patients (11,777) have a short stay (1 - 3 days) in the hospital, representing 47.19% of the feature.

Figure 1.1

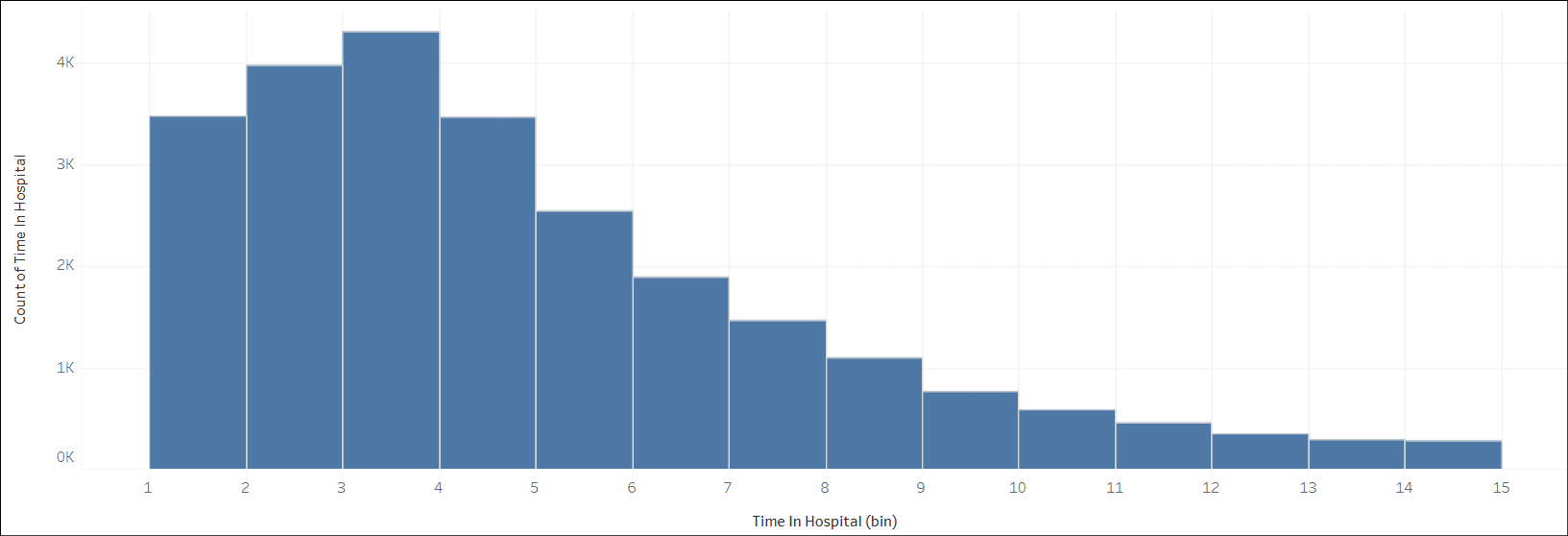
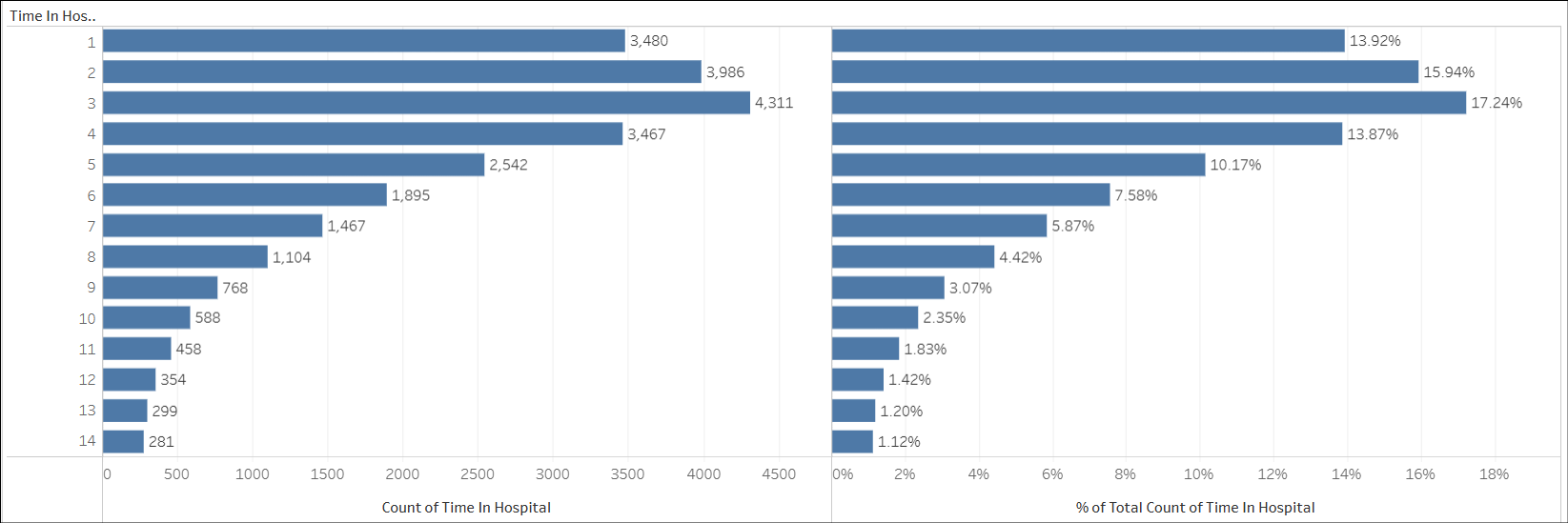


Figure 1.2



Bivariate analysis

* In Figure 1.3, the median for both readmitted and non-readmitted patients is 4 days.

The whiskers and the interquartile range are the same for both groups. Given no clear visual difference in the distribution of hospital stay duration between the two patient groups, this feature is a **weak indicator** for the predictive model.

* In Figure 1.4, the overall trend shows that the readmission rate increases slightly as the length of stay increases with a slight deviation for patients staying more than 5 days. This suggests a moderate positive correlation relationship between the length of stay and the patient readmission rate.

Figure 1.3

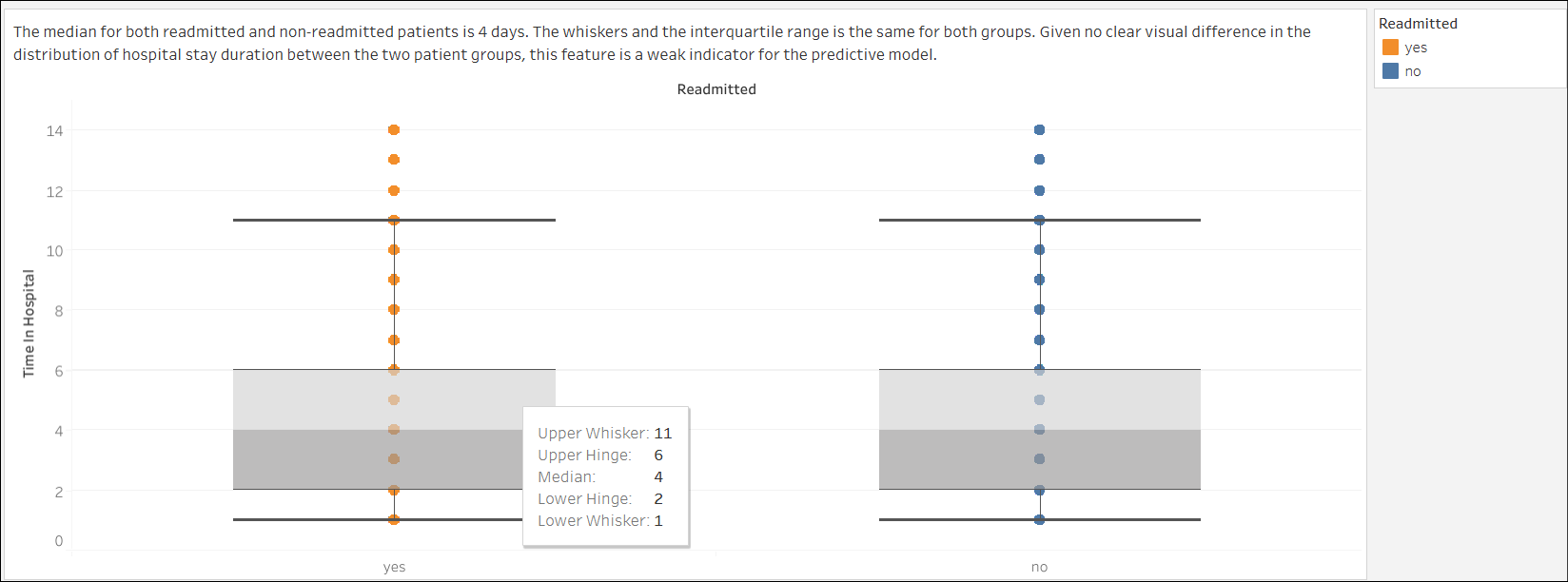
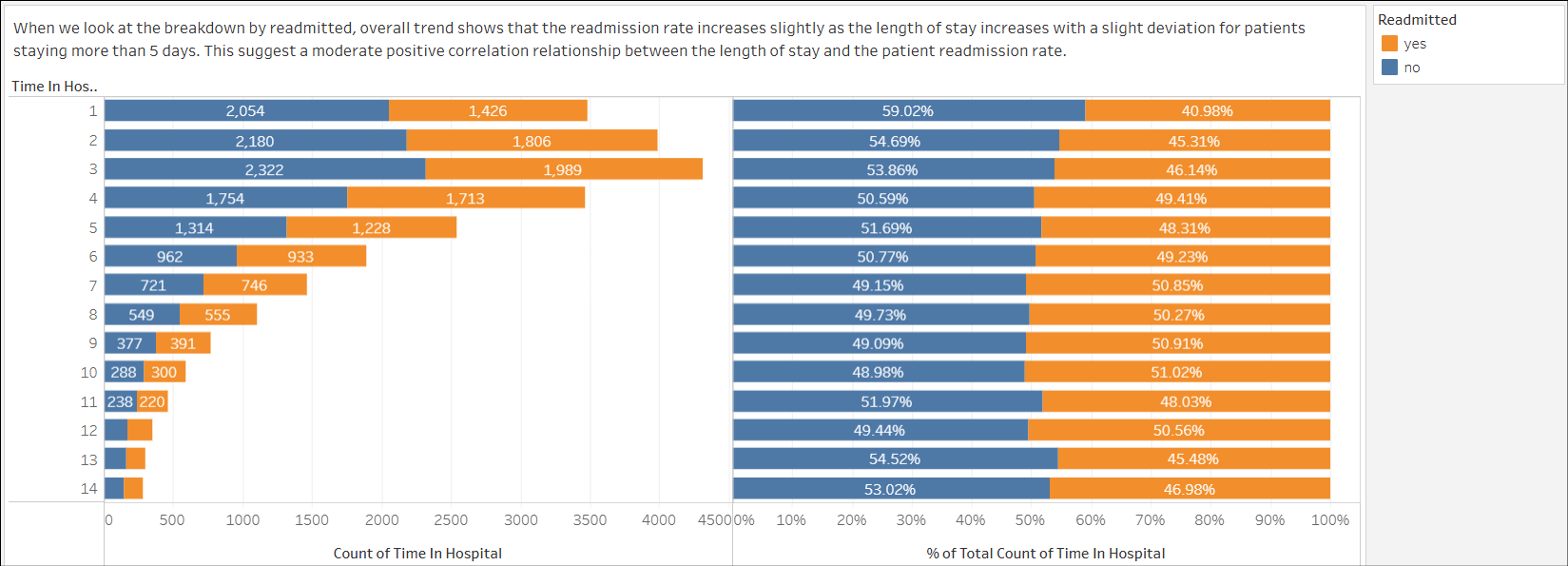


Figure 1.4



1. procedures

Univariate analysis

* In Figure 2.1, the distribution of the data is right-skewed, indicating that while the majority of patients have 0 to 1 procedures performed during the hospital stay, a smaller number of patients have significantly higher procedures performed.
* In Figure 2.2, the majority of patients (11,409) had no procedures performed during their hospital admission, representing 45.64% of the feature. The distribution is heavily skewed to the left.

Figure 2.1

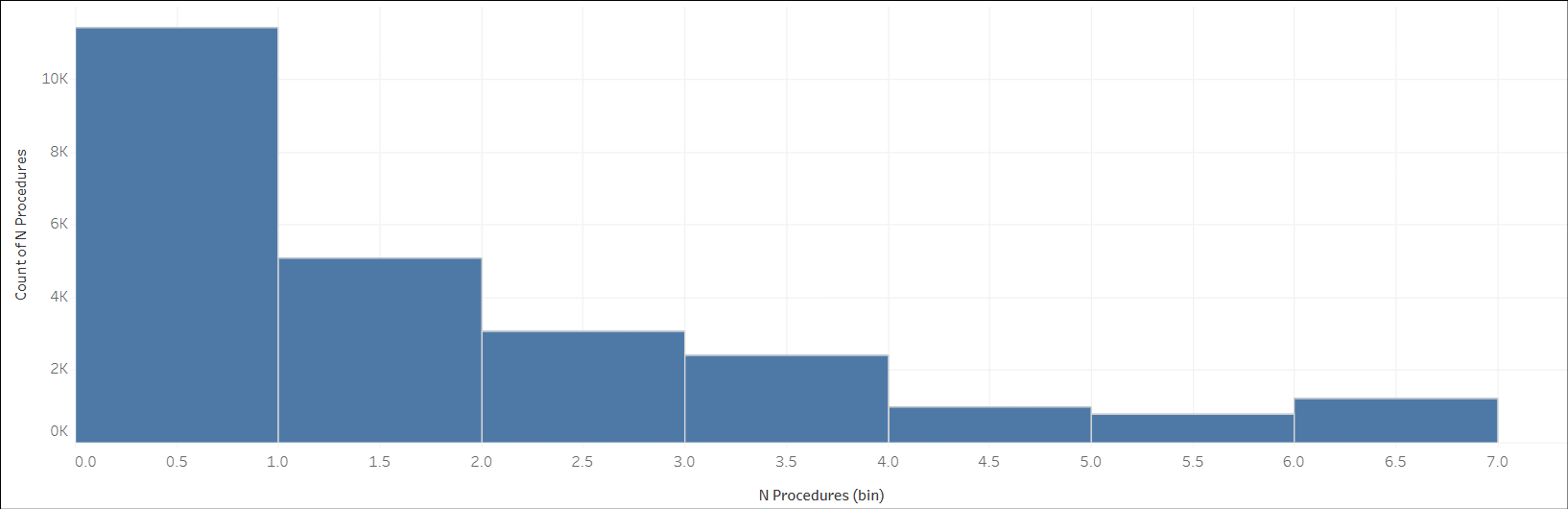
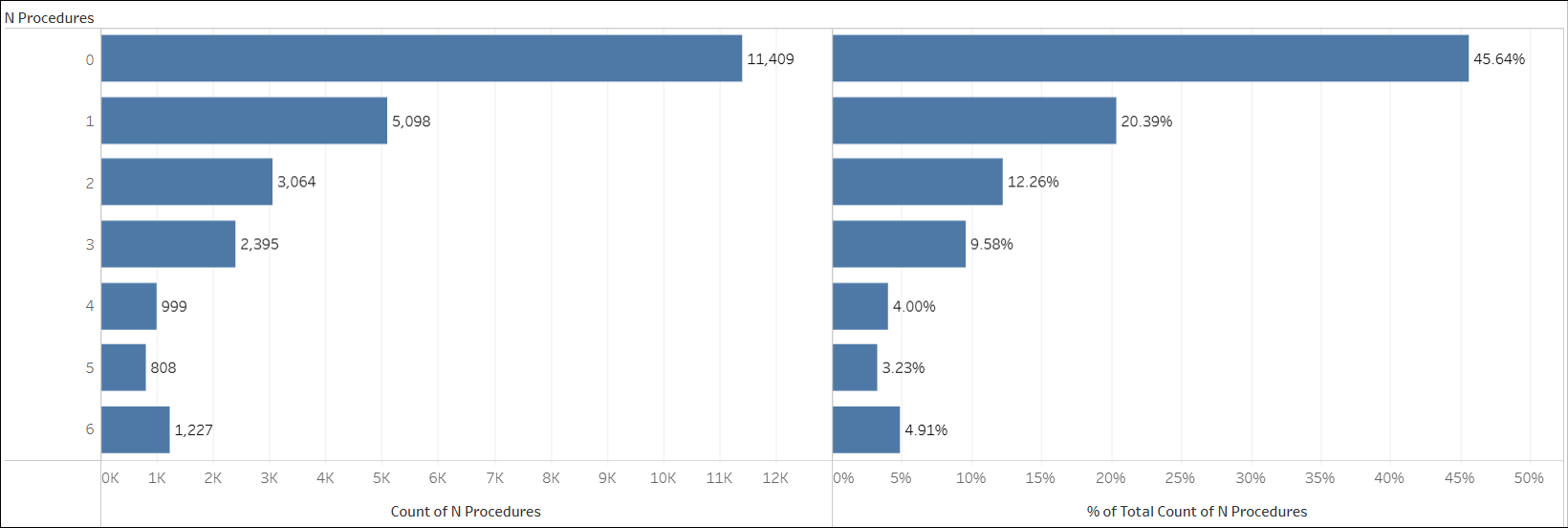


Figure 2.2



Bivariate analysis

* In Figure 2.3, the median for both readmitted and non-readmitted patients is 1 procedure. The whiskers and the interquartile range is the same for both groups. Given no clear visual difference in the distribution of procedures performed during the hospital stay between the two patient groups, this suggests the procedures feature is a **weak indicator** for the predictive model.
* In Figure 2.4, the readmission rate generally decreases as the number of procedures increases, with a slight deviation for patients with 4 procedures. This shows a moderate **negative correlation** between the number of procedures and the readmission rate.

Figure 2.3

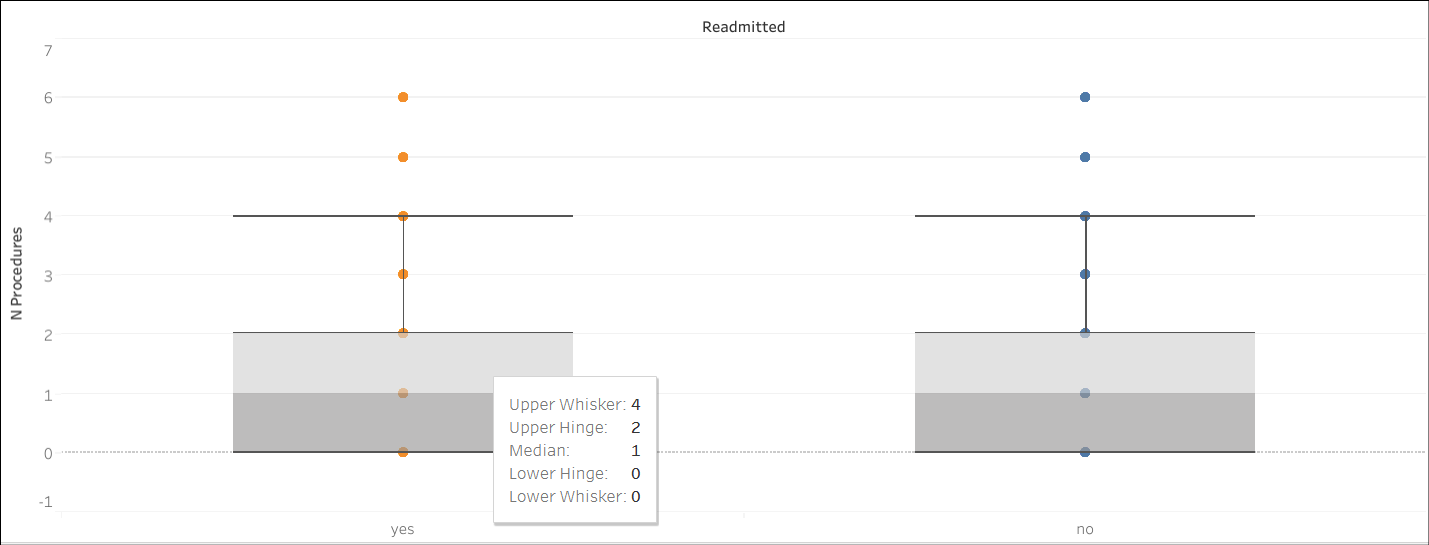
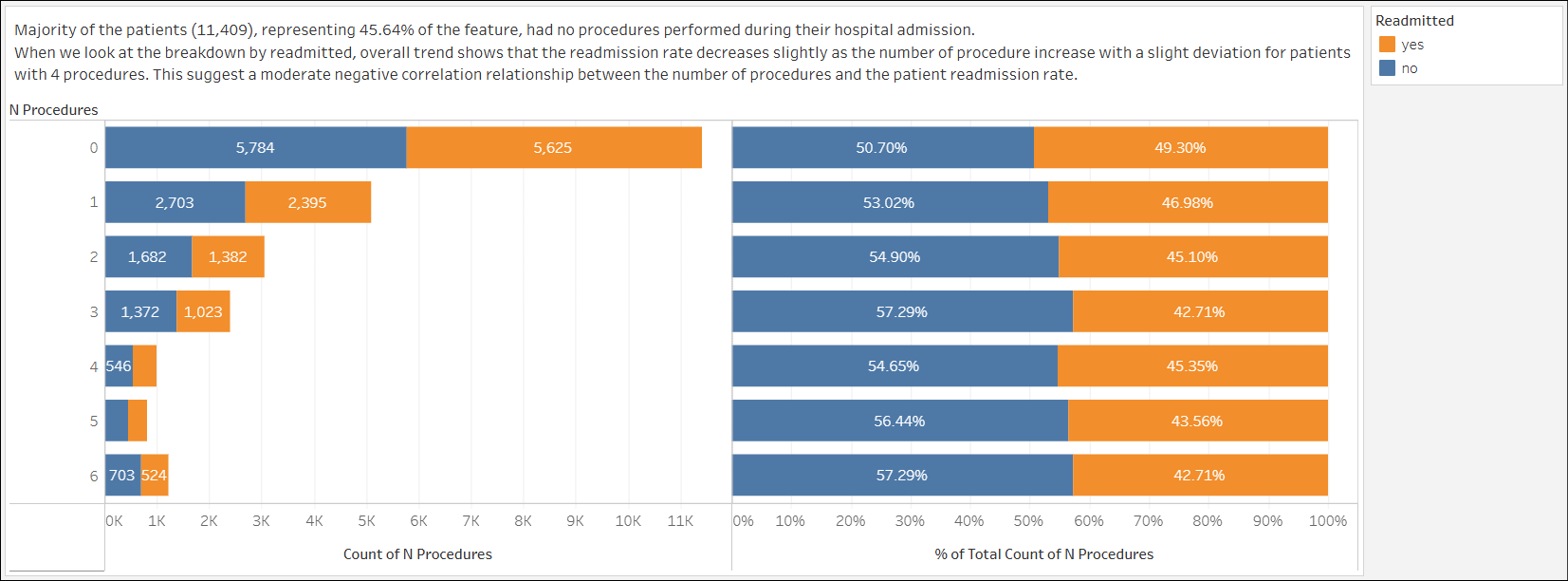


Figure 2.4

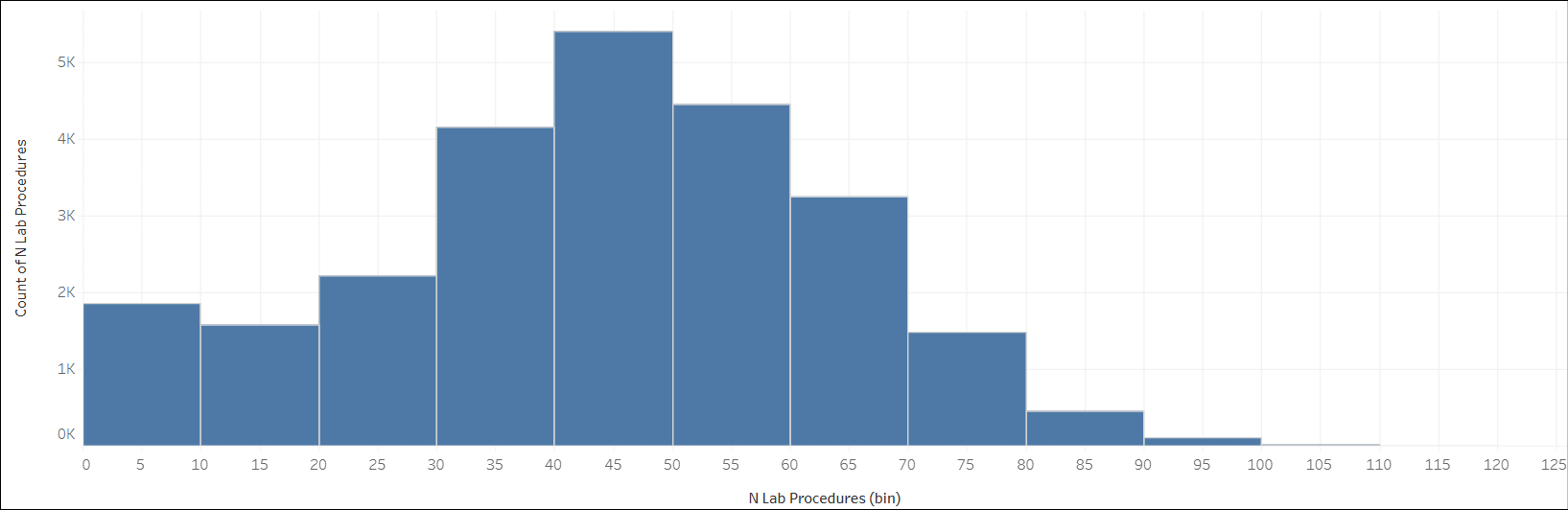


1. lab\_procedures

Univariate analysis

* In Figure 3.1, the number of lab procedures are centered around 40-50 lab procedures, which means this is the typical number of lab procedures conducted on a patient during a hospital stay. There are 3 bins where patients have more than 90 lab procedures conducted. The data is moderately right-skewed.

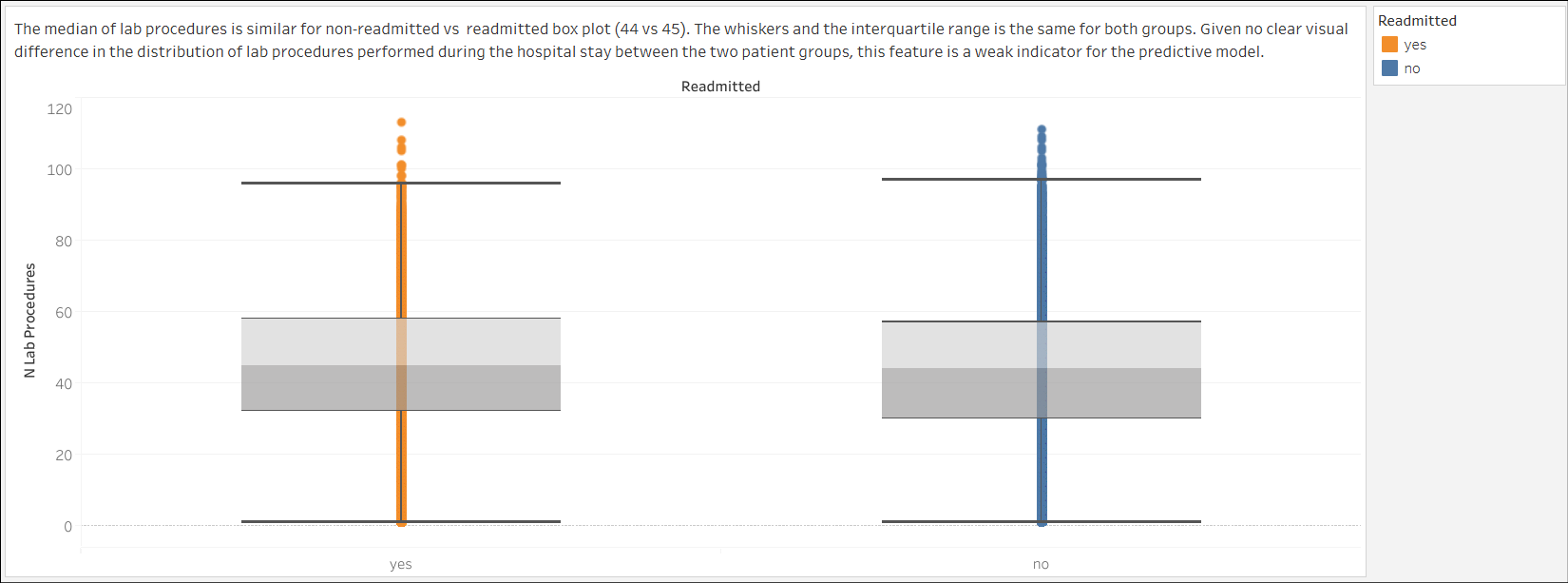
Figure 3.1



Bivariate analysis

* In Figure 3.2, the median of lab procedures is similar for non-readmitted vs readmitted box plot (44 vs 45). There are a number of individual dots in both box plots but it seems higher in the non-readmitted box plot. The whiskers and the interquartile range is the same for both groups. This suggests the lab\_procedures feature is a **weak indicator** for our prediction model.

Figure 3.2

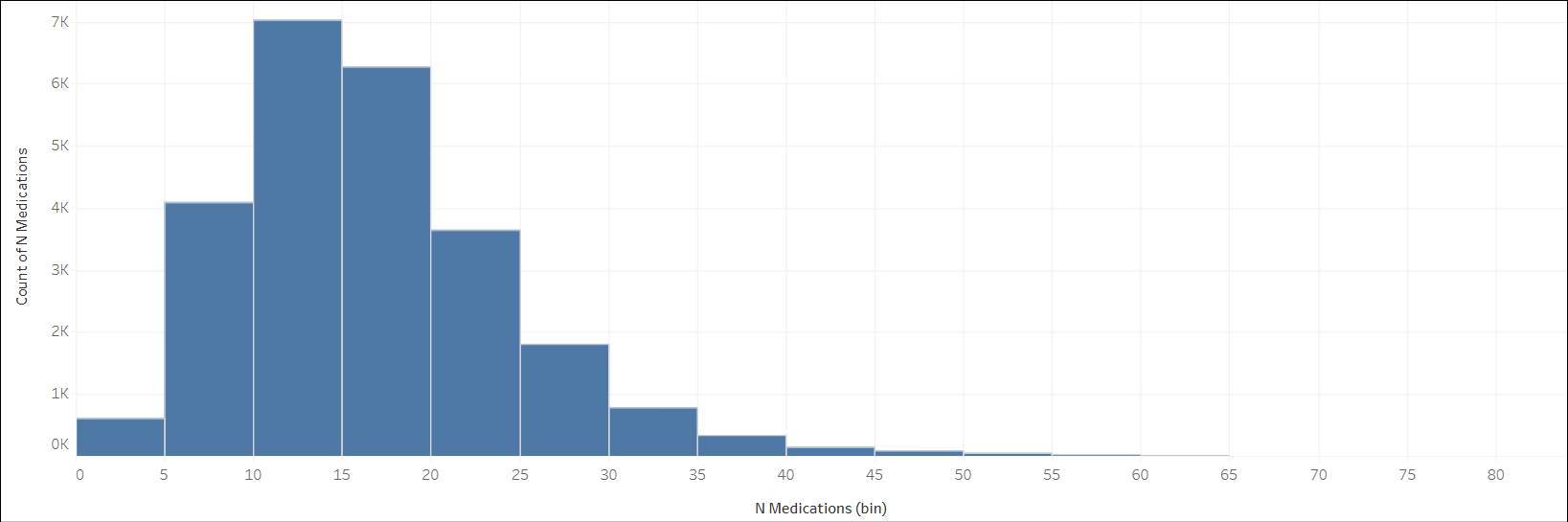


1. n\_medications

Univariate analysis

* In Figure 4.1, the number of medications administered to patients is centered around 10-15 medications. The distribution is right-skewed.

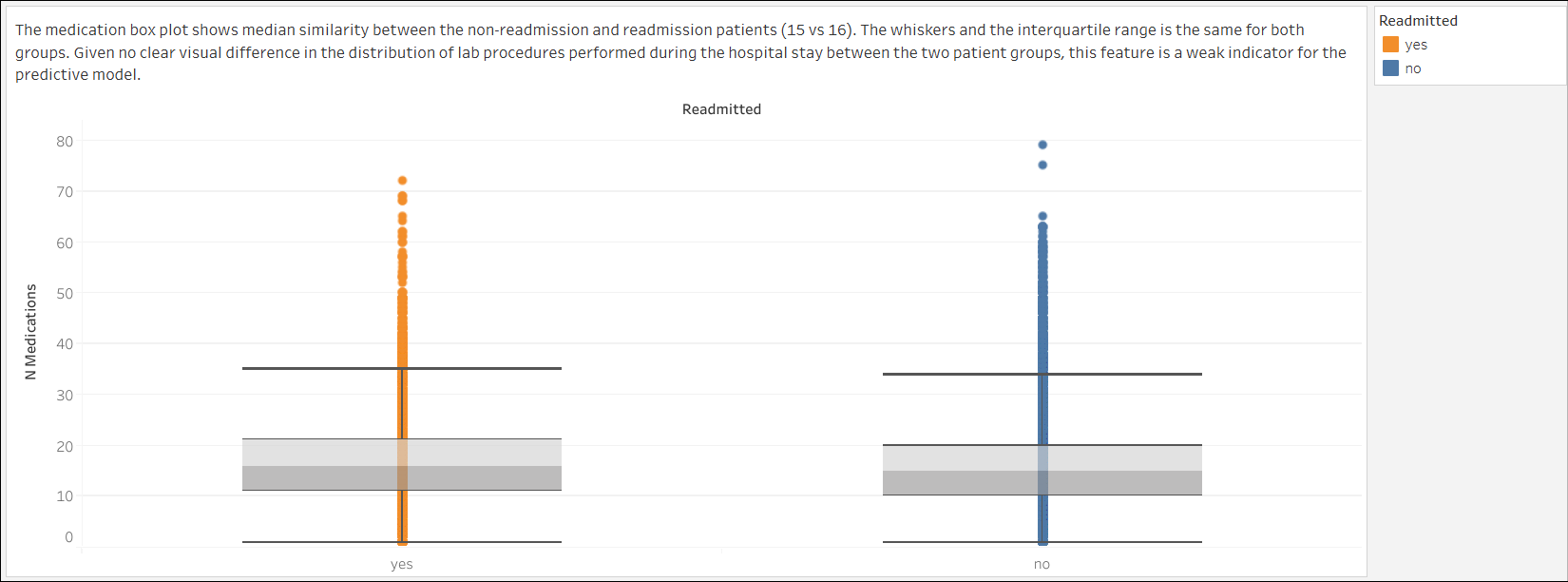
Figure 4.1



Bivariate analysis

* In Figure 4.2, the medication box plot shows median similarity between the non-readmission and readmission patients (15 vs 16). The whiskers and the interquartile range is the same for both groups. Given no clear visual difference in the distribution of lab procedures performed during the hospital stay between the two patient groups, this suggests the n\_medications feature is **a weak indicator** for the predictive model.

Figure 4.2



1. n\_outpatient

Univariate analysis

* In Figure 5.1, the distribution of the data is right-skewed, indicating that while the majority of patients have 0 to 4 outpatient visits before the hospital stay, a smaller number of patients have significantly higher outpatient visits.
* In Figure 5.2, the majority of patients have no outpatient visit, 20,859 total visits, in this feature. This represents 83.44% of this category.

Figure 5.1

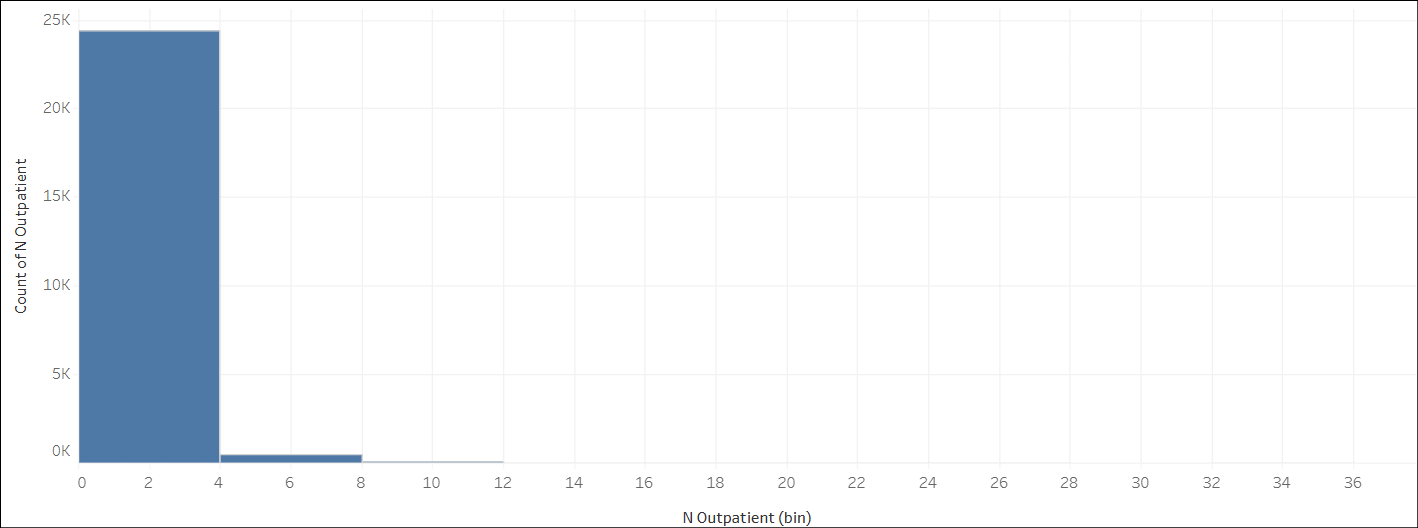
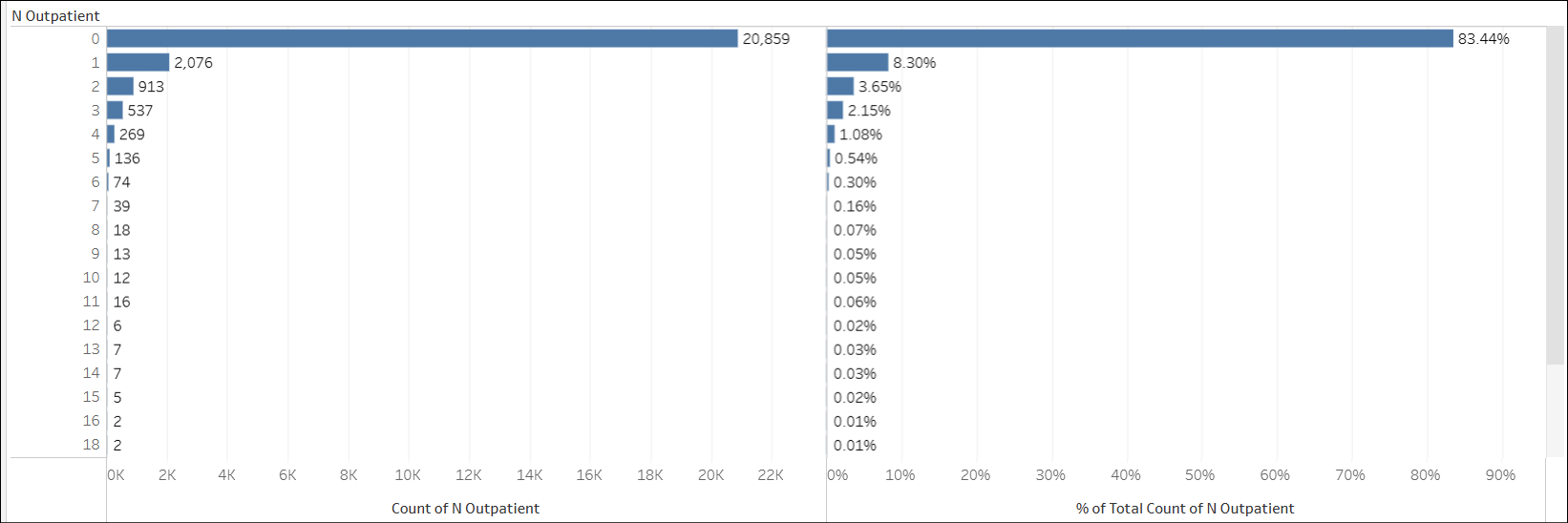


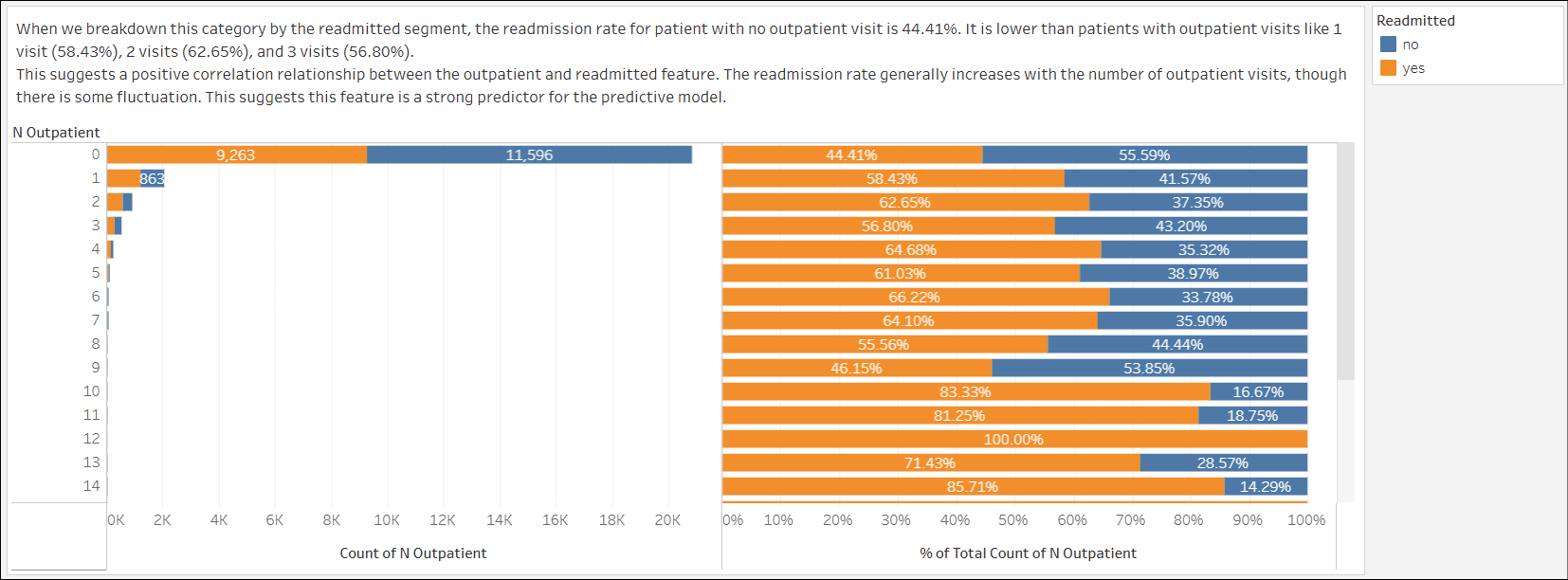
Figure 5.2



Bivariate analysis

* In Figure 5.3, the readmission rate for patients with no outpatient visit is 44.41%. It is lower than patients with outpatient visits like 1 visit (58.43%), 2 visits (62.65%), and 3 visits (56.80%).
* This suggests a positive correlation relationship between the outpatient and readmitted feature. The readmission rate generally increases with the number of outpatient visits, though there is some fluctuation. This suggests the n\_outpatient feature is a **strong predictor** for the predictive model.

Figure 5.3



1. n\_inpatient

Univariate analysis

* In Figure 6.1, the distribution of the data is right-skewed, indicating that while the majority of patients have 0 to 2 inpatient visits before the hospital stay, a smaller number of patients have significantly higher inpatient visits.
* In Figure 6.2, the majority patient has no inpatient visit prior to the hospital admission where it recorded 16,537 total visits. It represents 66.15% for this feature.

Figure 6.1

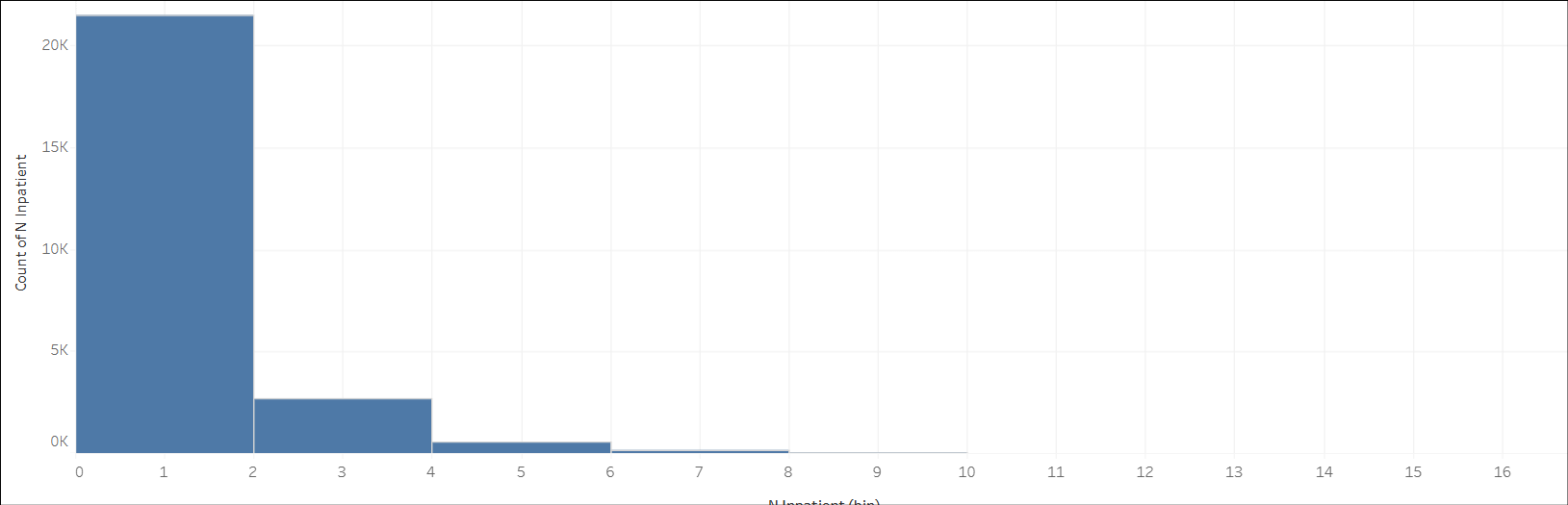
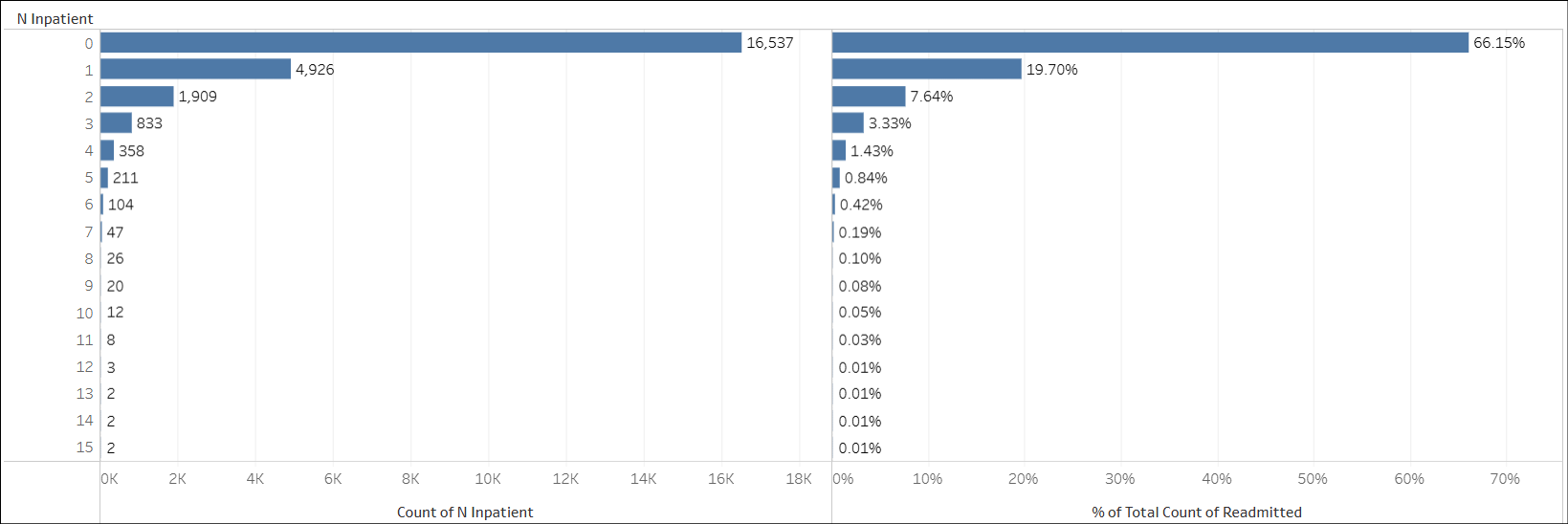


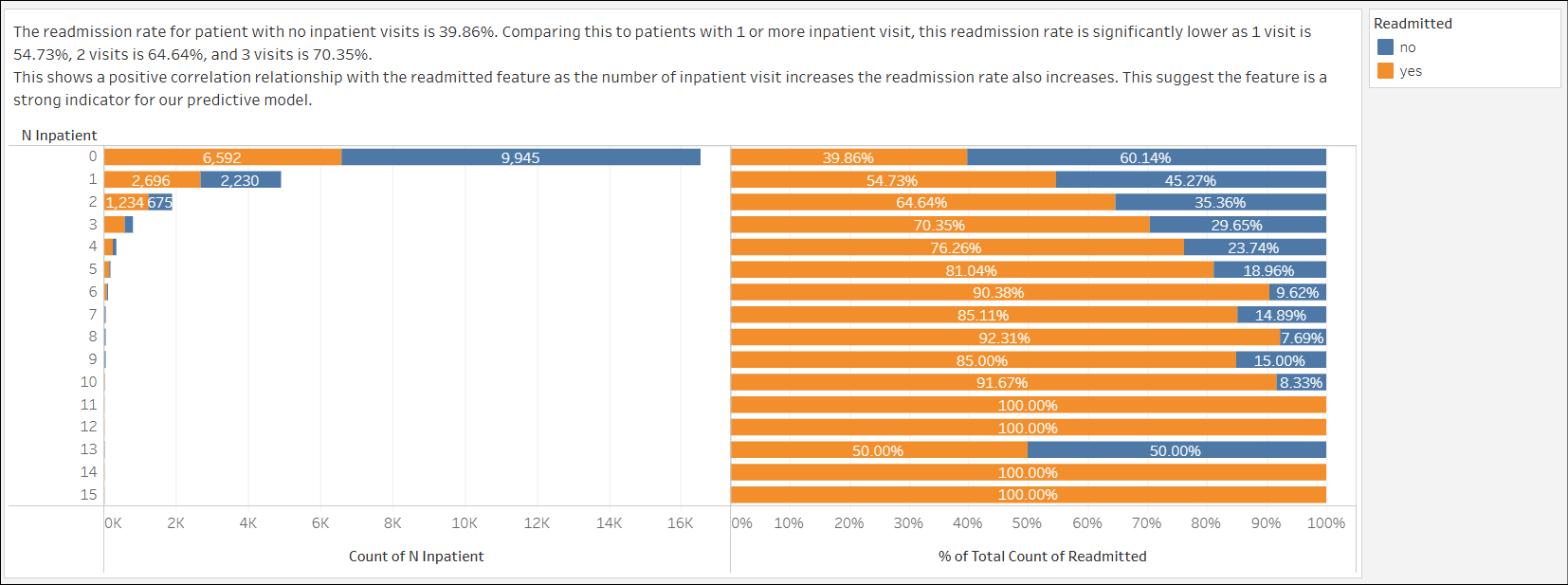
Figure 6.2



Bivariate analysis

* In Figure 6.3, the readmission rate for patients with no inpatient visits is 39.86%. Comparing this to patients with 1 or more inpatient visits, this readmission rate is significantly lower as 1 visit is 54.73%, 2 visits is 64.64%, and 3 visits is 70.35%.
* This shows a positive correlation relationship with the readmitted feature as the number of inpatient visits increases and the readmission rate also increases. This suggests the n\_inpatient feature is a **strong indicator** for our predictive model.

Figure 6.3



1. n\_emergency

Univariate analysis

* In Figure 7.1, the distribution of the data is right-skewed, indicating that while the majority of patients have 0 to 5 emergency visits before the hospital stay, a smaller number of patients have higher emergency visits.
* In Figure 7.2, the majority patient has no inpatient visit prior to the hospital admission where it recorded 16,537 total visits. It represents 66.15% for this feature.

Figure 7.1

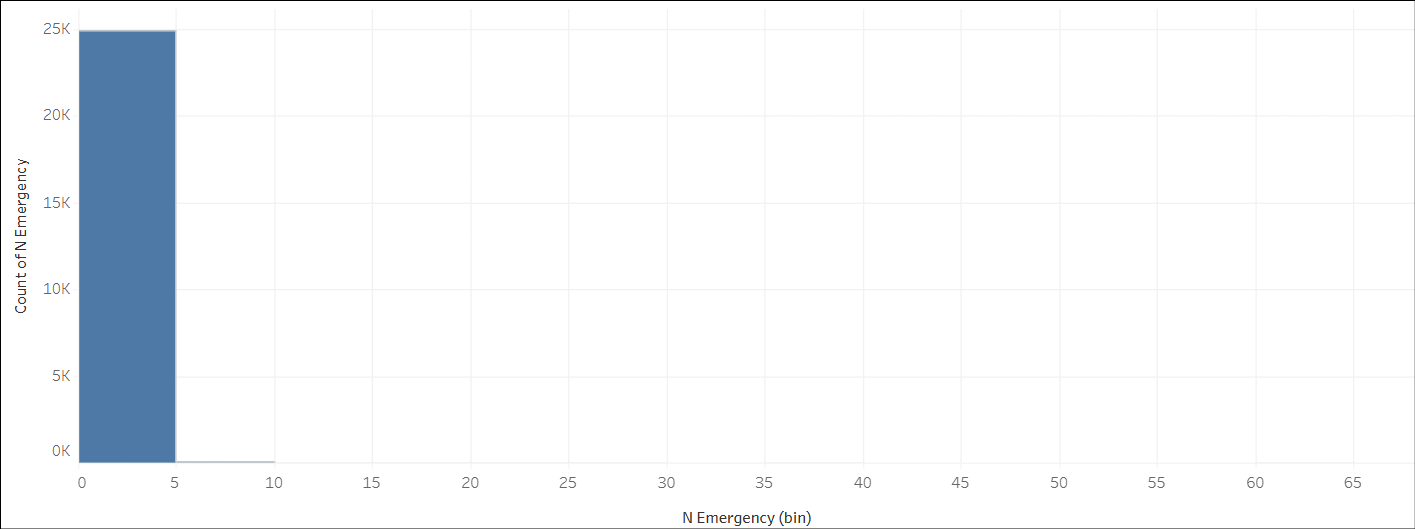
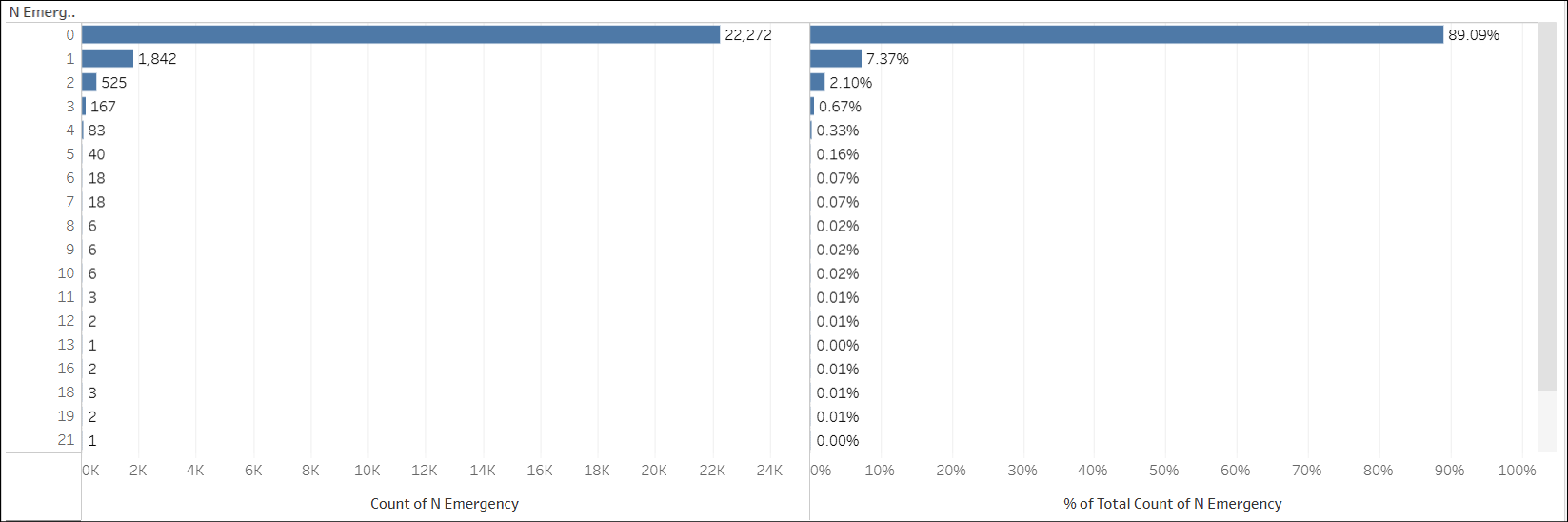


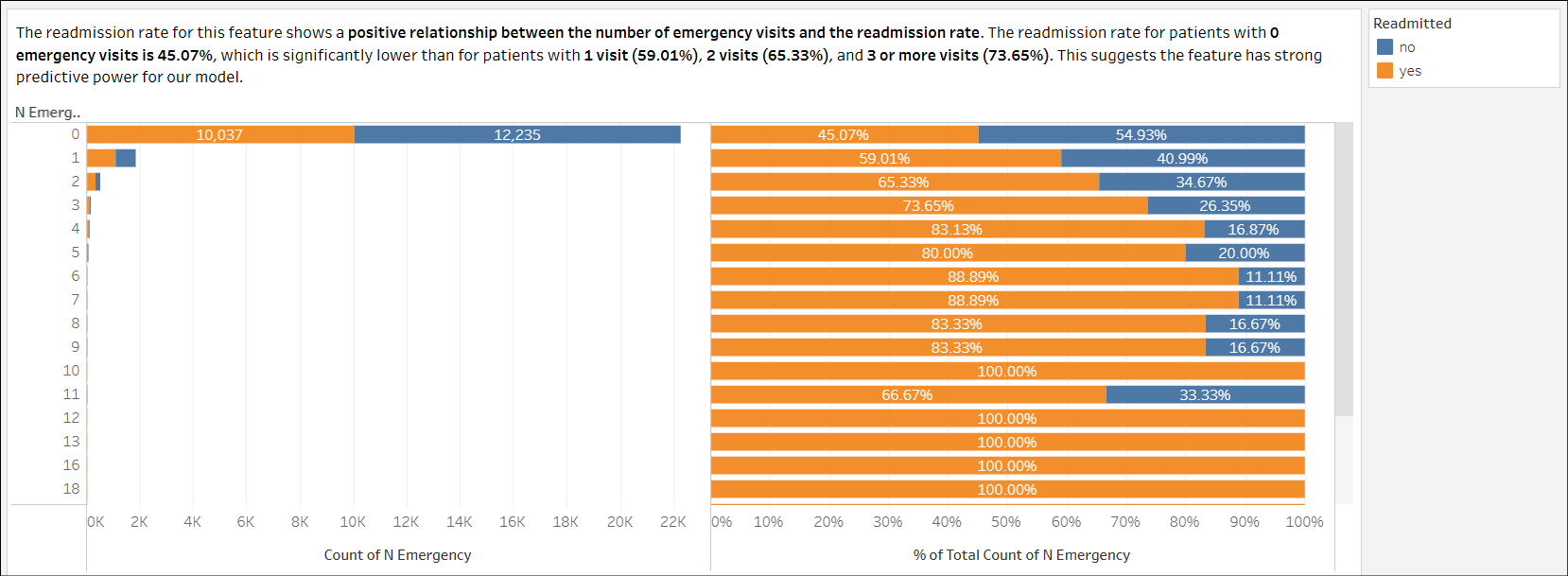
Figure 7.2



Bivariate analysis

* In Figure 7.3, the readmission rate for patients with no emergency visits is 45.07%, which is significantly lower than for patients with **1 visit (59.01%)**, **2 visits (65.33%)**, and **3 or more visits (73.65%)**.
* This shows a positive correlation relationship with the readmitted feature as the number of emergency visits increases and the readmission rate also increases. This suggests the n\_emergency feature is a **strong indicator** for our predictive model.

Figure 7.3



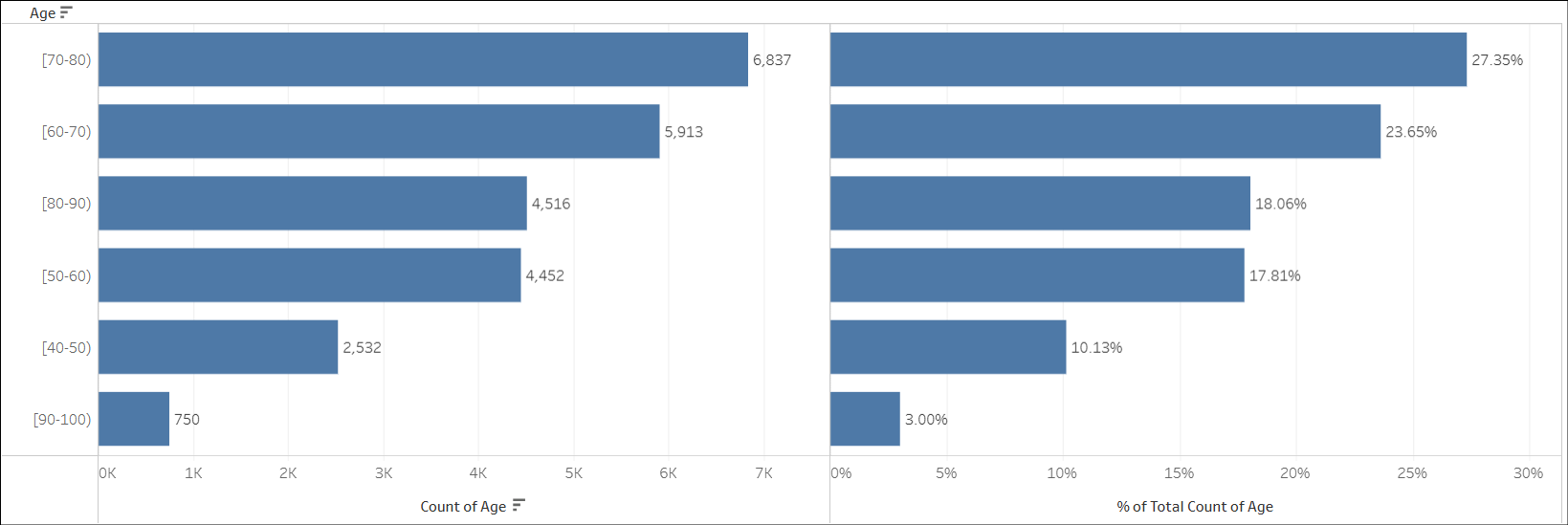
### Categorical Features

1. age

Univariate analysis

* In Figure 8.1, the 70-80 age group represents the largest group of patients (6,837), representing 27.35% of the feature.

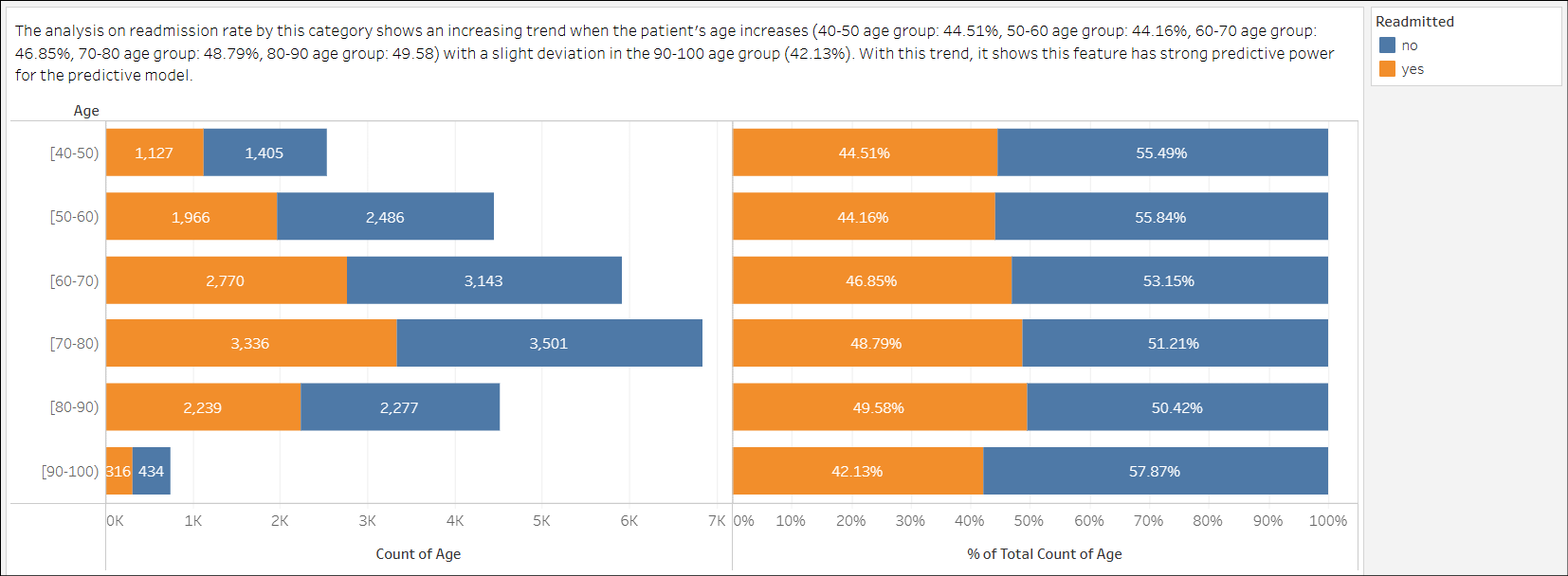
Figure 8.1



Bivariate analysis

* In Figure 8.2, the analysis on readmission rate by this category shows an increasing trend when the patient's age increases (40-50 age group: 44.51%, 50-60 age group: 44.16%, 60-70 age group: 46.85%, 70-80 age group: 48.79%, 80-90 age group: 49.58) with a slight deviation in the 90-100 age group (42.13%). With this trend, it shows the age feature has **strong predictive** **power** for the predictive model.

Figure 8.2

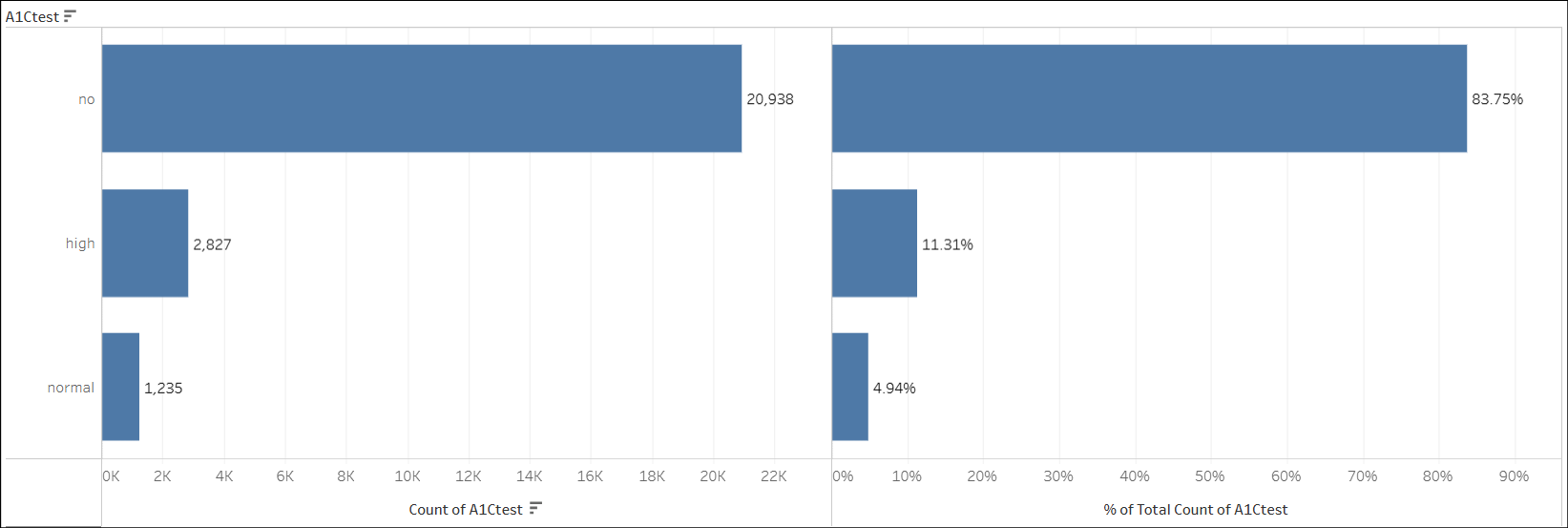


1. A1Ctest

Univariate analysis

* In Figure 9.1, most of the patients (20,938), representing 83.75% of the dataset, do not have the A1C test performed during the hospital admission.

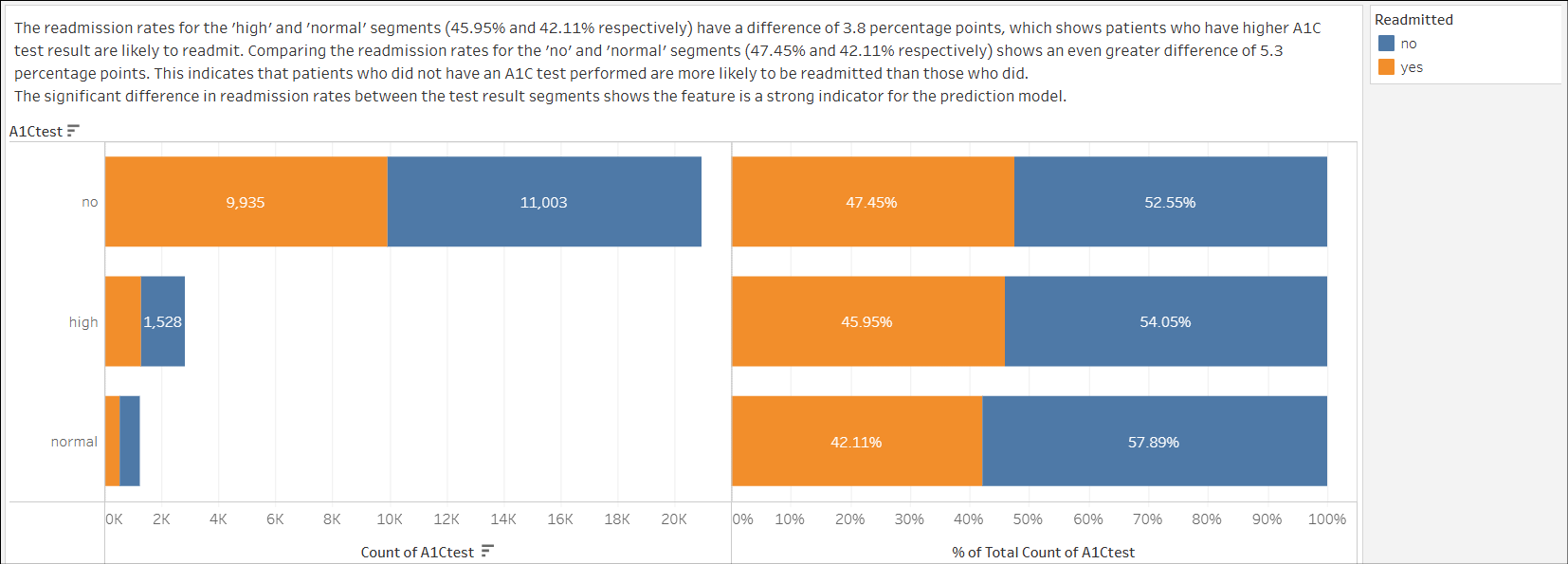
Figure 9.1



Bivariate analysis

* In Figure 9.2, the readmission rates for the 'high' and 'normal' segments (45.95% and 42.11% respectively) have a difference of 3.8 percentage points, which shows patients who have higher A1C test results are likely to readmit. Comparing the readmission rates for the 'no' and 'normal' segments (47.45% and 42.11% respectively) shows an even greater difference of 5.3 percentage points. This indicates that patients who did not have an A1C test performed are more likely to be readmitted than those who did.
* The significant difference in readmission rates between the test result segments shows the A1Ctest feature is a **strong indicator** for the prediction model.

Figure 9.2

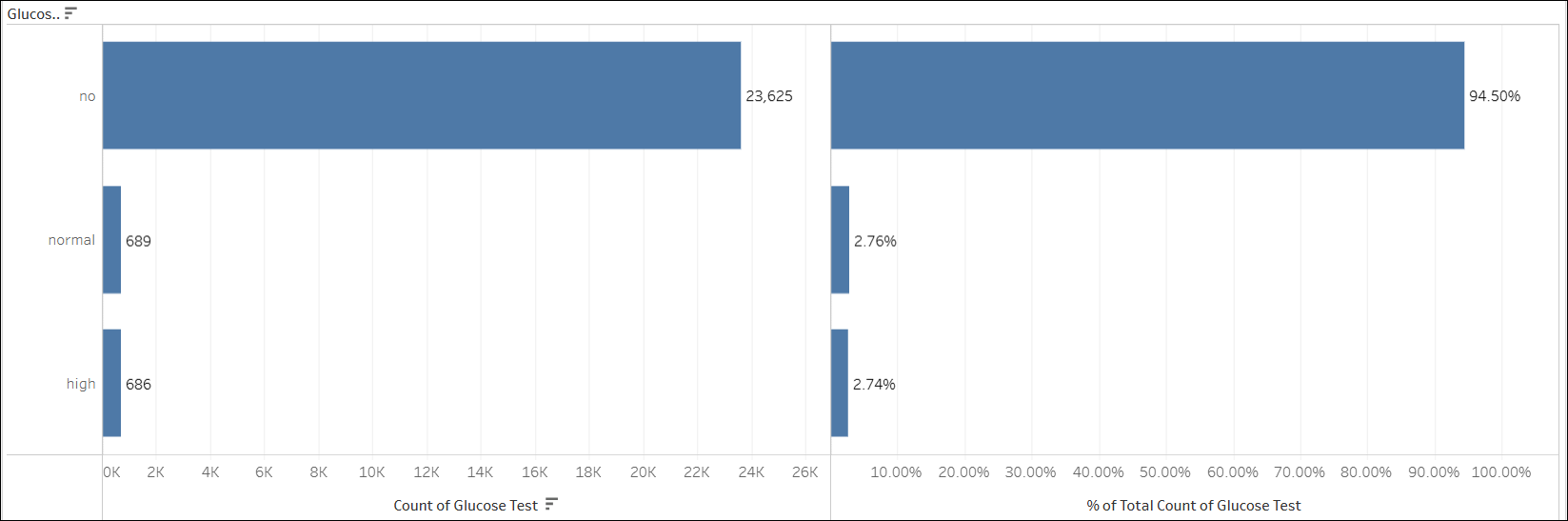


1. glucose\_test

Univariate analysis

* In Figure 10.1, there are 3 segments in the glucose\_test feature: 'high', 'normal', and 'no'. The majority of patients (23,625), representing 94.50% of the dataset, have no glucose test performed during the hospital admission.

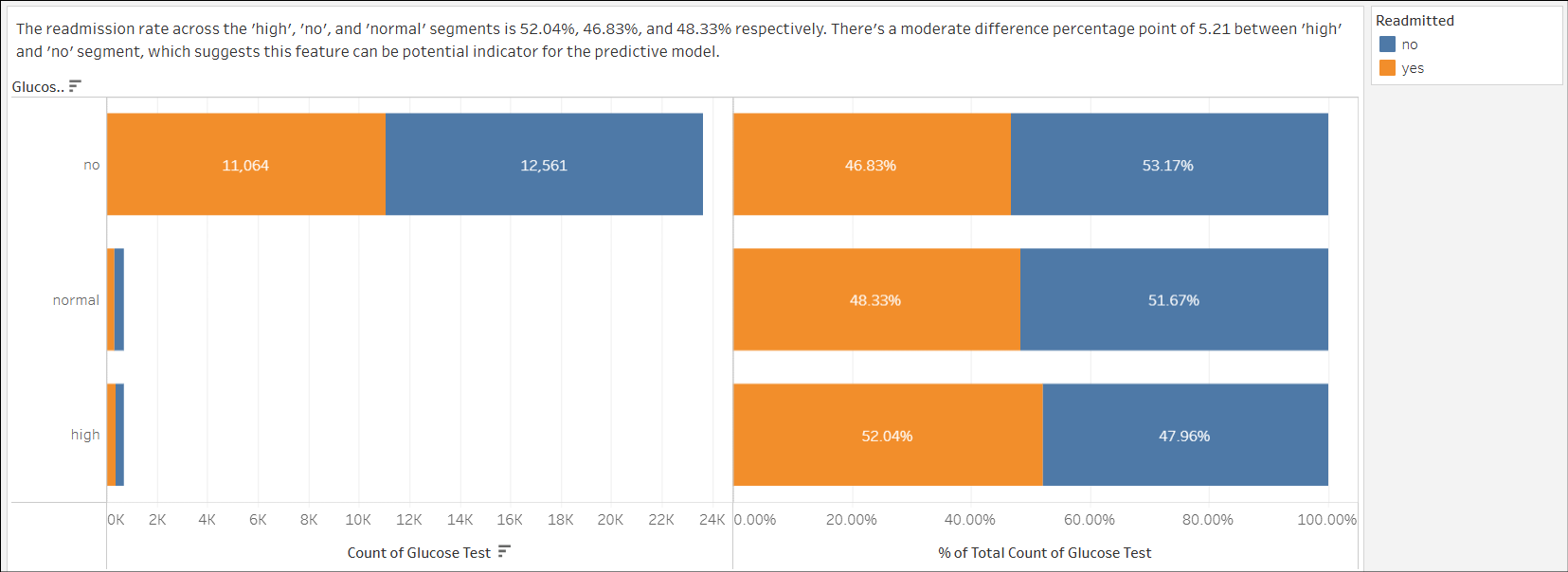
Figure 10.1



Bivariate analysis

* In Figure 10.2, the readmission rate across the 'high', 'no', and 'normal' segments is 52.04%, 46.83%, and 48.33% respectively. The difference percentage point of 5.21 between 'high' and 'no' segment, suggests the glucose\_test feature is a **strong indicator** for the predictive model.

Figure 10.2

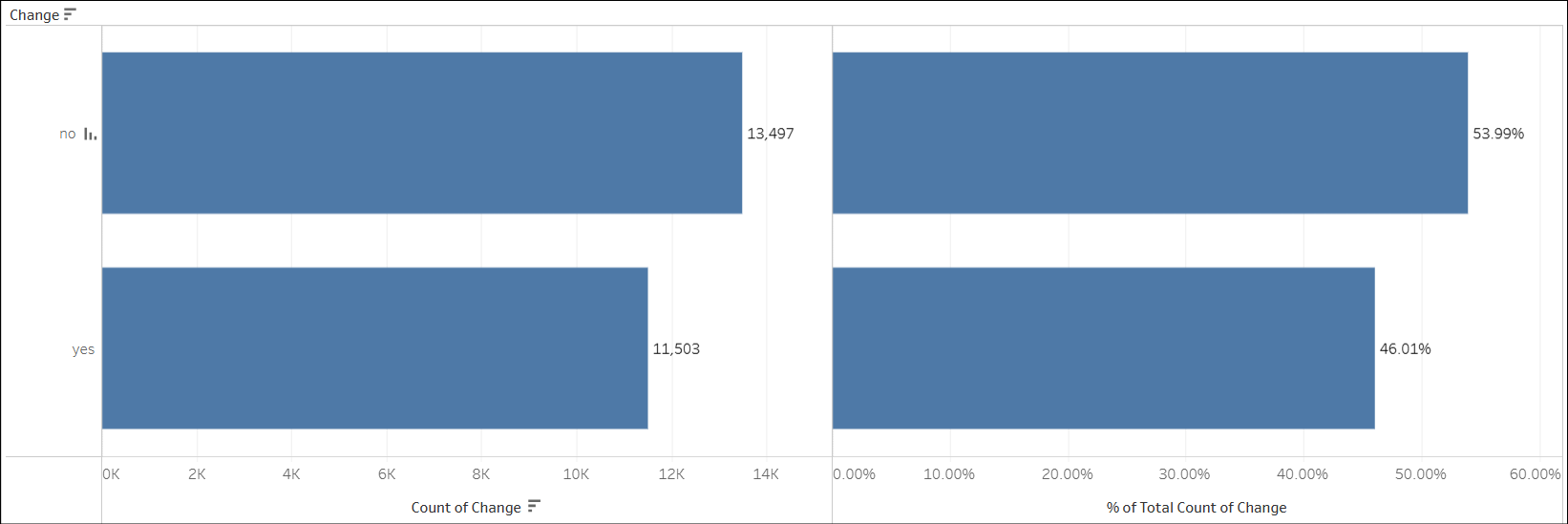


1. change

Univariate analysis

* In Figure 11.1, the majority of patients (13,497), representing 53.99% of the dataset, have no diabetes medication change.

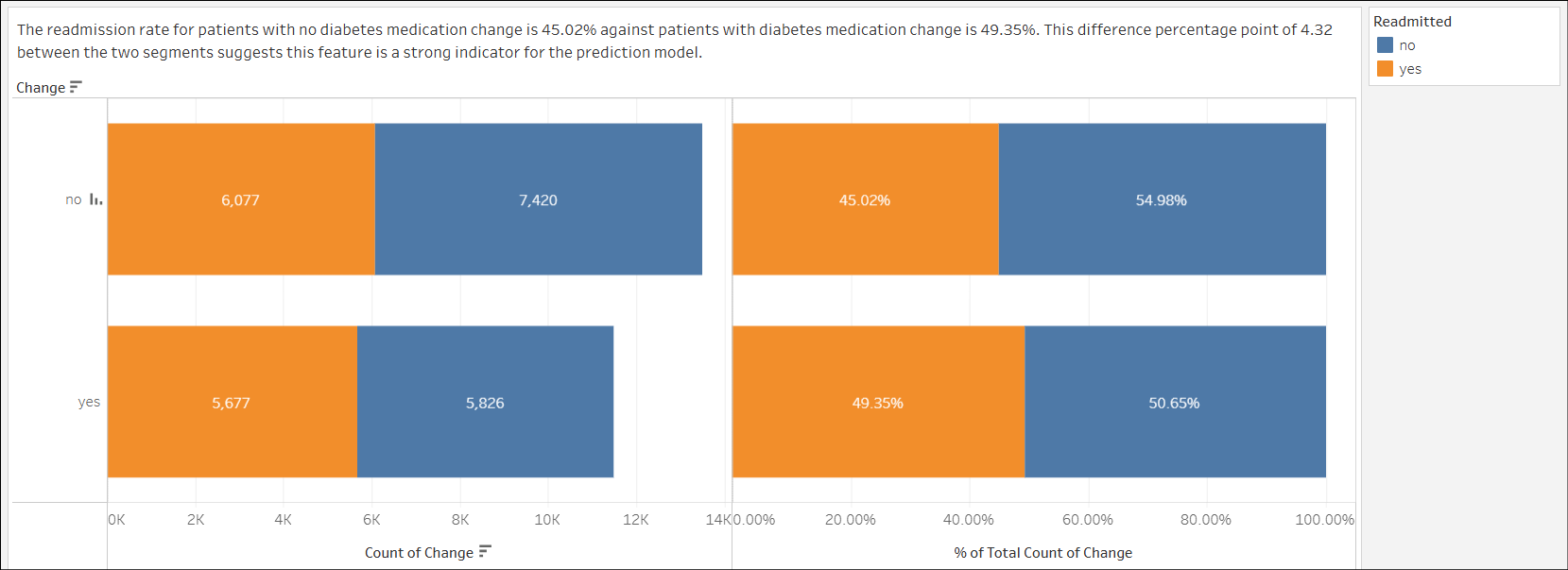
Figure 11.1



Bivariate analysis

* In Figure 11.2, the readmission rate for patients with no diabetes medication change is 45.02% against patients with diabetes medication change is 49.35%. This difference percentage point of 4.32 between the two segments suggests the change feature is a **strong indicator** for the prediction model.

Figure 11.2

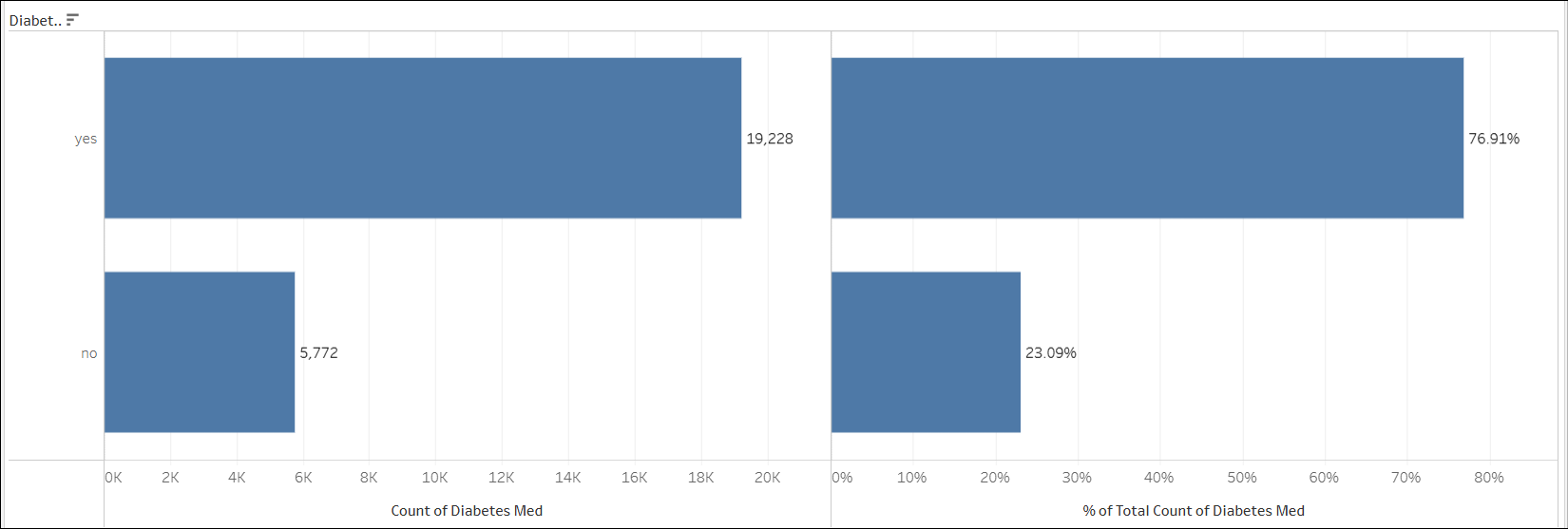


1. diabetes\_med

Univariate analysis

* In Figure 12.1, the majority of patients (19,228), representing 76.91% of the dataset, have been prescribed with diabetic medications.

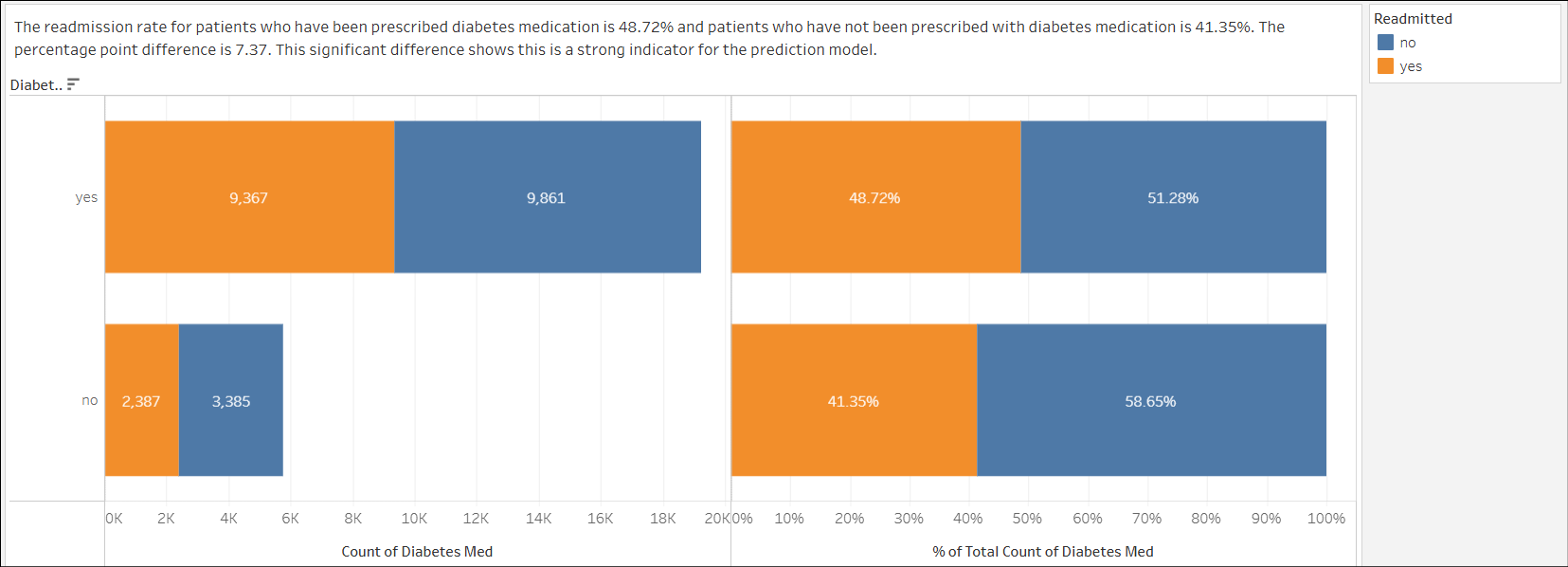
Figure 12.1



Bivariate analysis

* In Figure 12.2, the readmission rate for patients who have been prescribed diabetes medication is 48.72% and patients who have not been prescribed with diabetes medication is 41.35%. The percentage point difference is 7.37. This significant difference shows the diabetes\_med feature is a **strong indicator** for the prediction model.

Figure 12.2

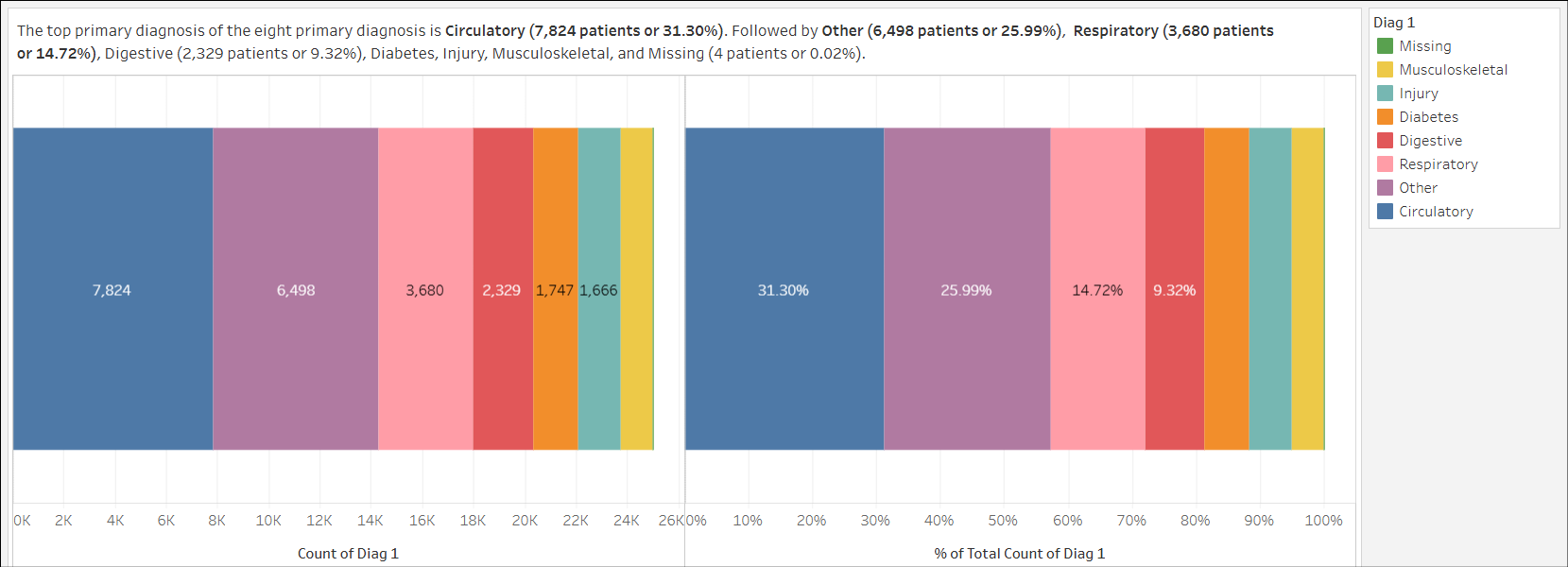


1. diag\_1

Univariate analysis

* In Figure 13.1, the top primary diagnosis of the eight primary diagnoses is **Circulatory (7,824 patients or 31.30%)**. Followed by **Other (6,498 patients or 25.99%)**, **Respiratory (3,680 patients or 14.72%)**, **Digestive (2,329 patients or 9.32%)**, Diabetes, Injury, Musculoskeletal, and Missing (4 patients or 0.02%).

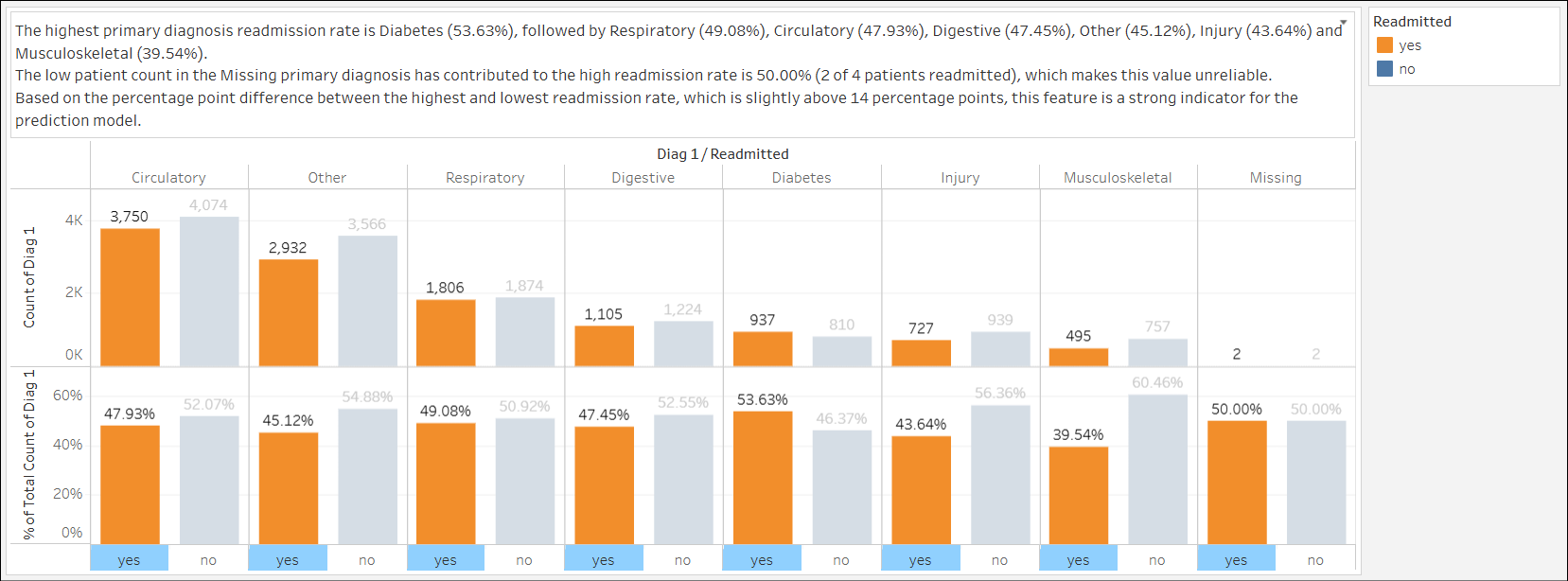
Figure 13.1



Bivariate analysis

* In Figure 13.2, the highest primary diagnosis readmission rate is Diabetes (53.63%), followed by Respiratory (49.08%), Circulatory (47.93%), Digestive (47.45%), Other (45.12%), Injury (43.64%) and Musculoskeletal (39.54%).
* The low patient count in the Missing primary diagnosis has contributed to the high readmission rate of 50.00% (2 of 4 patients readmitted), which makes this value unreliable.
* Based on the percentage point difference between the highest and lowest readmission rate, which is slightly above 14 percentage points, the diag\_1 feature is a **strong indicator** for the prediction model.

Figure 13.2

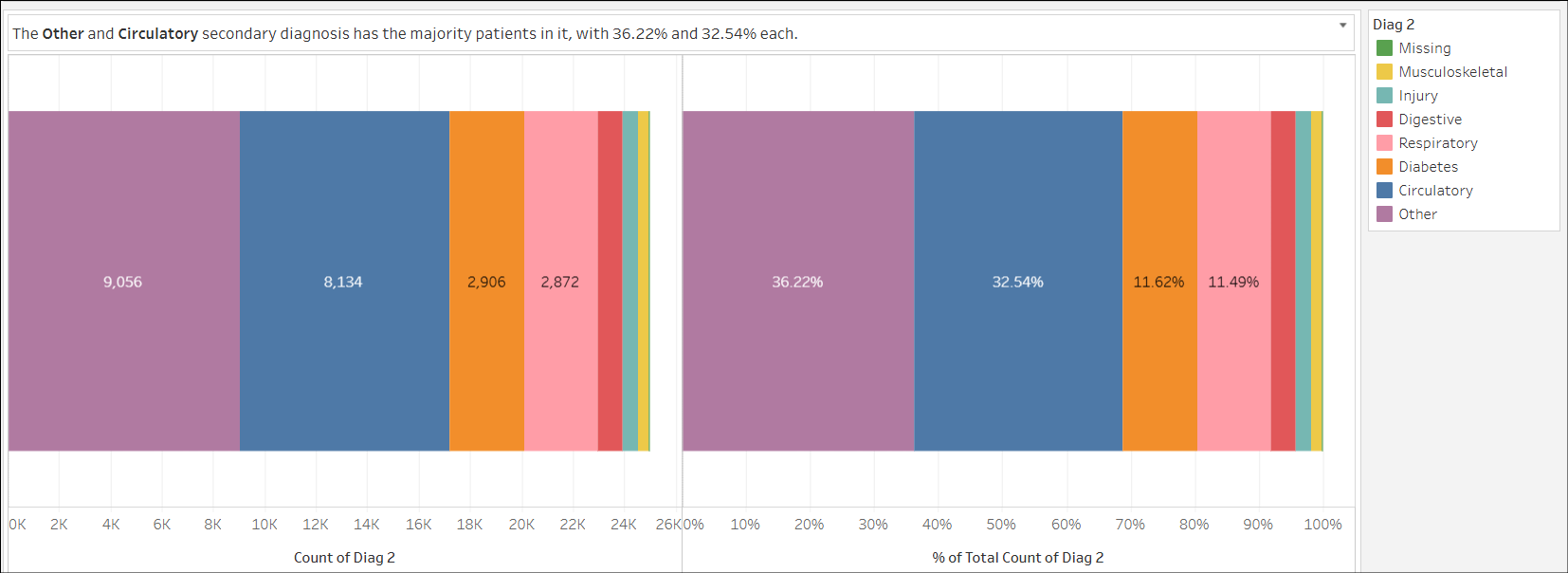


1. diag\_2

Univariate analysis

* In Figure 14.1, the **Other** and **Circulatory** secondary diagnosis has the majority of patients in it, with 36.22% and 32.54% each.
* The Missing label in the feature accounts for 42 patients or 0.17%.

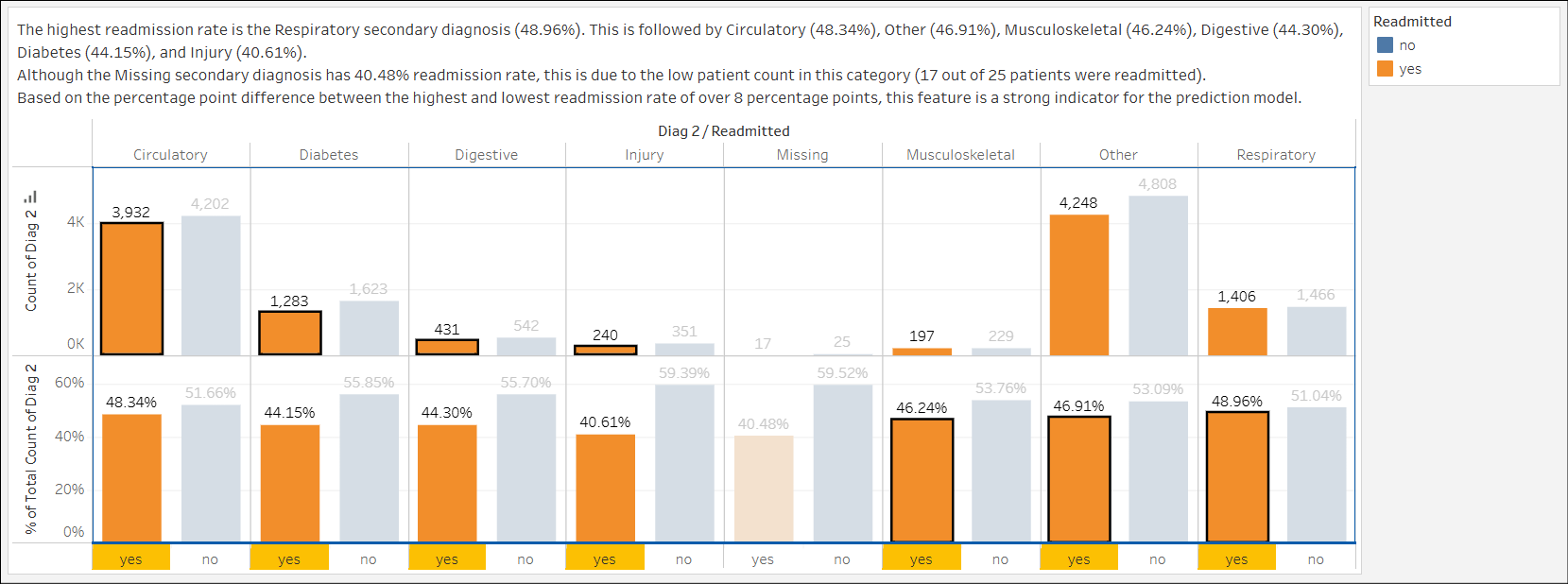
Figure 14.1



Bivariate analysis

* In Figure 14.2, the highest readmission rate is the Respiratory secondary diagnosis (48.96%). This is followed by Circulatory (48.34%), Other (46.91%), Musculoskeletal (46.24%), Digestive (44.30%), Diabetes (44.15%), and Injury (40.61%).
* Although the Missing secondary diagnosis has a 40.48% readmission rate, this is due to the low patient count in this category (17 out of 25 patients were readmitted).
* Based on the percentage point difference between the highest and lowest readmission rate of over 8 percentage points, the diag\_2 feature is a **strong indicator** for the prediction model.

Figure 14.2

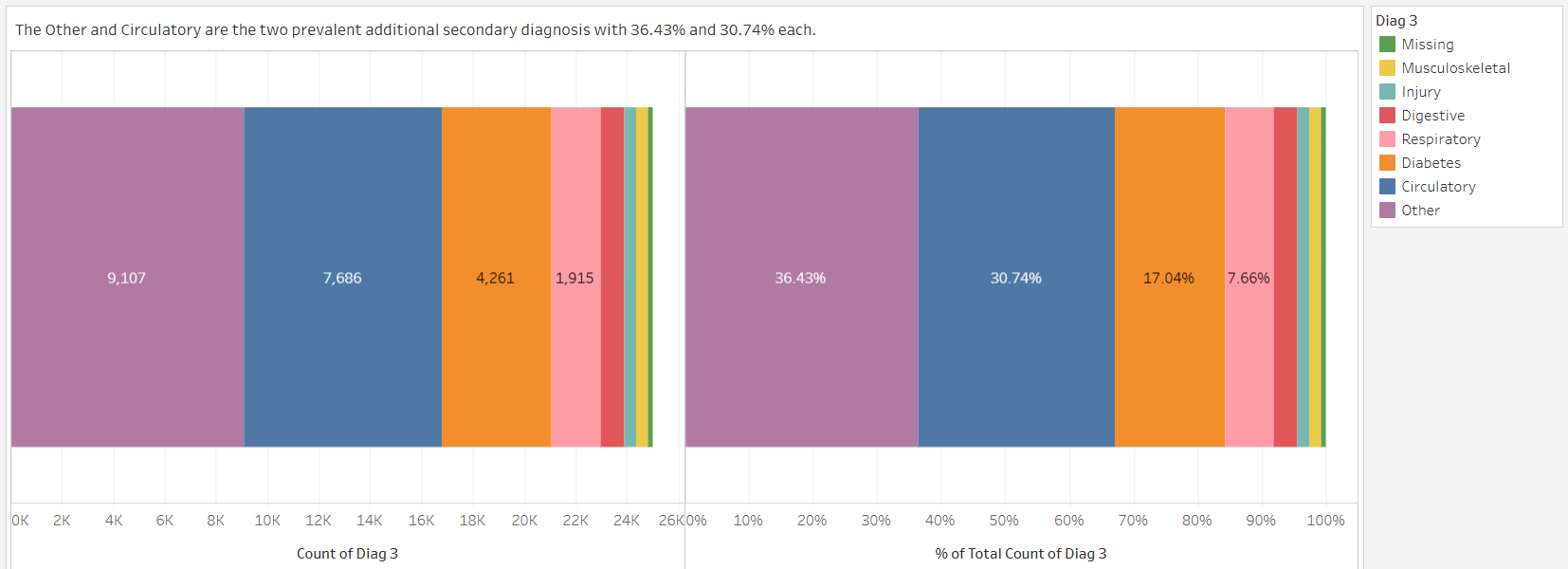


1. diag\_3

Univariate analysis

* In Figure 15.1, the **Other** and **Circulatory** are the two prevalent additional secondary diagnoses with 36.43% and 30.74% each.
* The Missing label in the feature accounts for 196 patients or 0.78%.

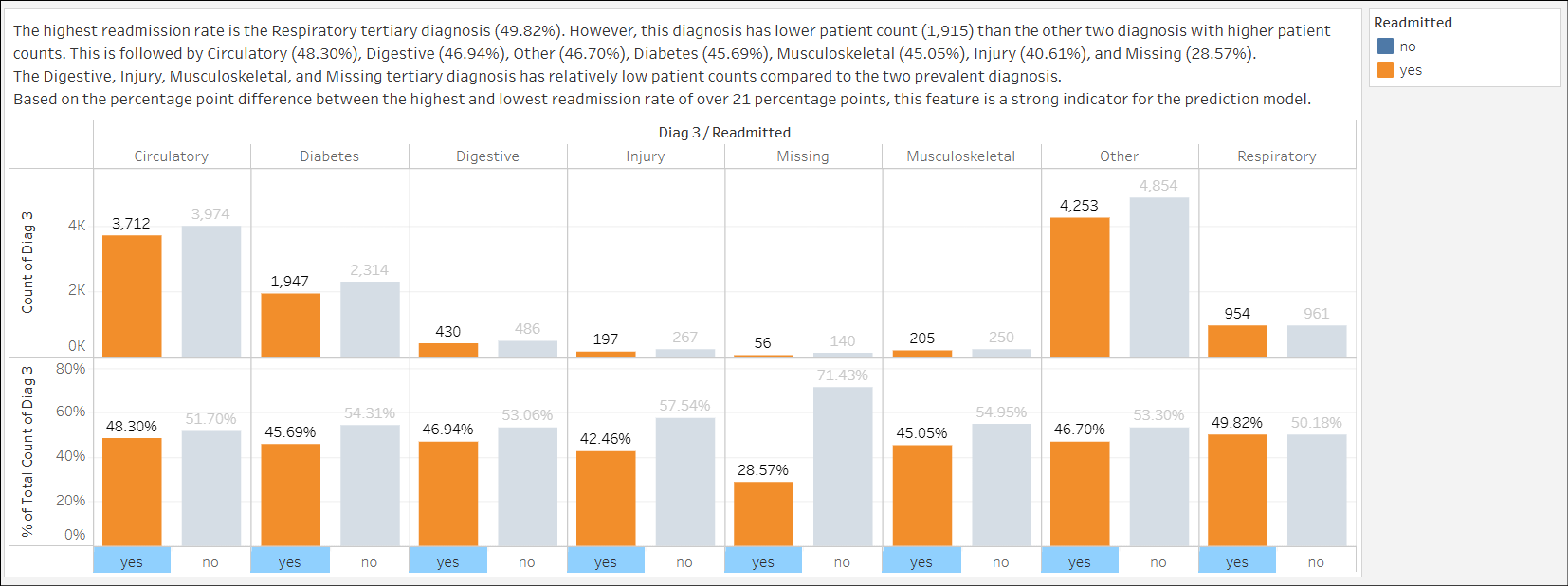
Figure 15.1



Bivariate analysis

* In Figure 15.2, the highest readmission rate is the Respiratory additional secondary diagnosis (49.82%). However, this diagnosis has lower patient count (1,915) than the other two diagnoses with higher patient counts. This is followed by Circulatory (48.30%), Digestive (46.94%), Other (46.70%), Diabetes (45.69%), Musculoskeletal (45.05%), Injury (40.61%), and Missing (28.57%).
* The Digestive, Injury, Musculoskeletal, and Missing tertiary diagnosis has relatively low patient counts compared to the two prevalent diagnoses.
* Based on the percentage point difference between the highest and lowest readmission rate of over 21 percentage points, the diag\_3 feature is a **strong indicator** for the prediction model.

Figure 15.2

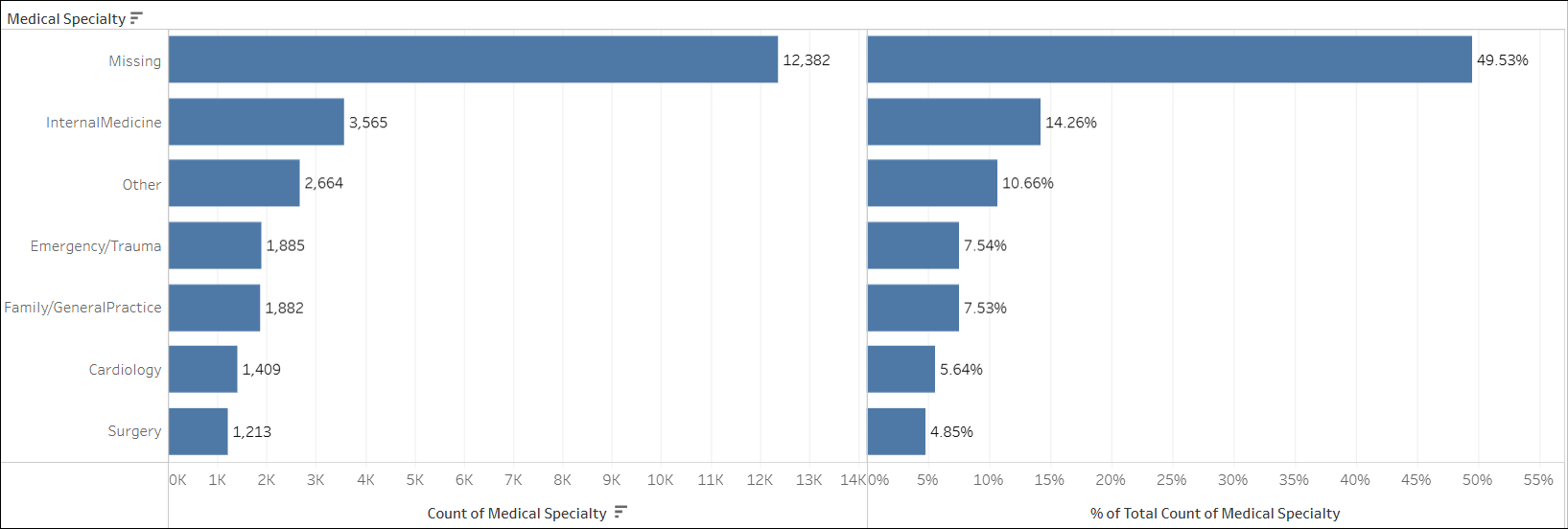


1. medical\_specialty

Univariate analysis

* In Figure 16.1, the highest medical specialty is the Missing specialty with 12,382 patients or 49.53%. This is followed by InternalMedicine (14.26%), Other (10.66%), Emergency/Trauma (7.54%).

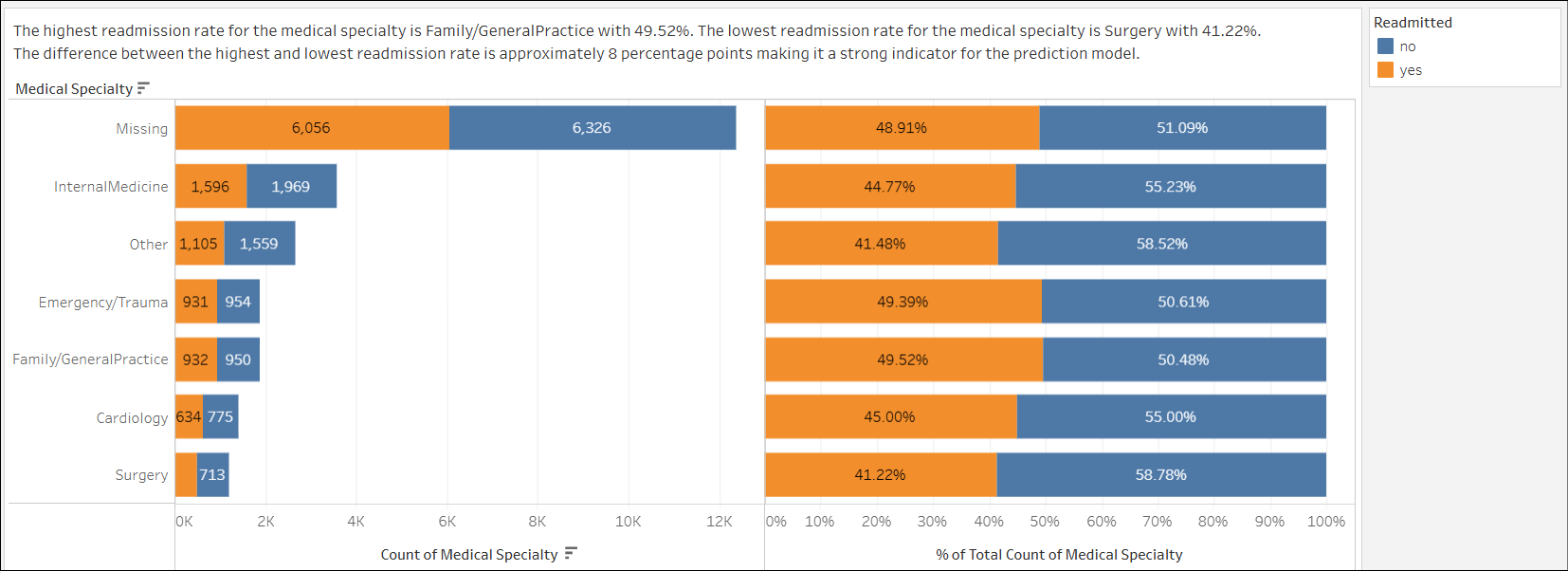
Figure 16.1



Bivariate analysis

* In Figure 16.2, the highest readmission rate for the medical specialty is Family/GeneralPractice with 49.52%. The lowest readmission rate for the medical specialty is Surgery with 41.22%.
* The difference between the highest and lowest readmission rate is approximately 8 percentage points making the medical\_specialty feature a **strong indicator** for the prediction model.

Figure 16.2

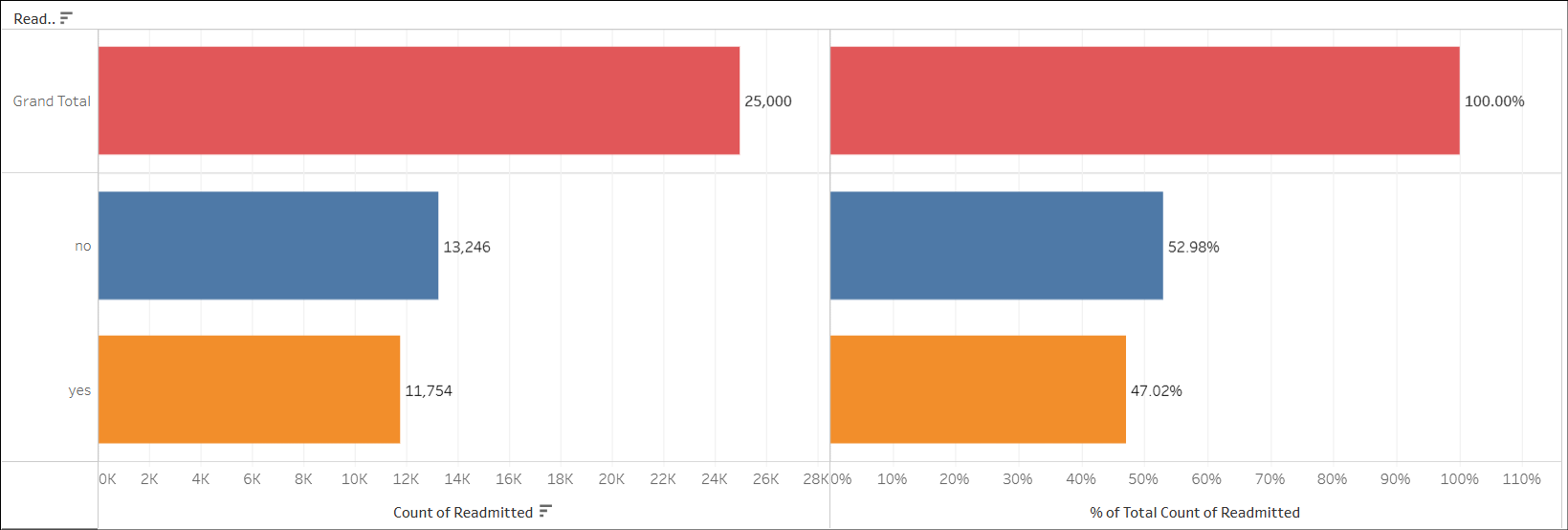


1. readmitted

Univariate analysis

* In Figure 17.1, the readmitted feature is relatively balanced in terms of the patients who are readmitted and not with 47.02% and 52.98% respectively.

Figure 17.1



### 

### Summary of the EDA on features to be used in the Predictive Model

|  |  |
| --- | --- |
| **Feature** | **Prediction Indicator** |
| time\_in\_hospital | Weak or provide additional prediction pattern |
| procedures | Weak or provide additional prediction pattern |
| lab\_procedures | Weak or provide additional prediction pattern |
| n\_medications | Weak or provide additional prediction pattern |
| n\_inpatient | Strong prediction indicator |
| n\_outpatient | Strong prediction indicator |
| n\_emergency | Strong prediction indicator |
| age | Strong prediction indicator |
| change | Strong prediction indicator |
| A1Ctest | Strong prediction indicator |
| glucose\_test | Strong prediction indicator |
| diag\_1 | Strong prediction indicator |
| diag\_2 | Strong prediction indicator |
| diag\_3 | Strong prediction indicator |
| medical\_specialty | Strong prediction indicator |
| diabetes\_med | Strong prediction indicator |

## Data Cleaning

Based on the EDA of the data set, we have found a large portion of the data (12,382 records or 49.53%) in the medical\_specialty feature is labeled as Missing. Since the Missing label is expected in the real-world dataset, we've decided to keep the data to train the final prediction model as it has its own pattern.

We'll be removing the Missing labels in the diag\_1, diag\_2, and diag\_3 feature since their proportion in the feature is insignificant (0.02%, 0.17%, 0.78% respectively) and have no benefit to the prediction model.

Please refer to the [Capstone - Patient Readmission Prediction Model.ipynb](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/03-Analyse/Capstone%20-%20Patient%20Readmission%20Prediction%20Model.ipynb) file for the code cell removing the Missing labels in the diag\_1, diag\_2, and diag\_3 feature.

## Feature Engineering

Based on the EDA, we have identified features like emergency, inpatient, outpatient, and procedures are highly imbalanced, with a large majority of patients having zero visits or procedures.

We will create binary bins for each of them to allow the prediction model to better learn the patterns from these features.

For the time\_in\_hospital, lab\_procedures and n\_medications feature, we will create category bins and transform it from a numeric variable into a categorical one. This will allow the model to learn from distinct patient groups rather than a continuous, weakly correlated number.

Please refer to the [Capstone - Patient Readmission Prediction Model.ipynb](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/03-Analyse/Capstone%20-%20Patient%20Readmission%20Prediction%20Model.ipynb) file for the code cells creating the 7 new features.

# Predictive Model

Based on the EDA, we will build the Logistic Regression model based on these features which is a strong indicator for the prediction model.

1. n\_inpatient
2. n\_outpatient
3. n\_emergency
4. age
5. change
6. A1Ctest
7. glucose\_test
8. diag\_1
9. diag\_2
10. diag\_3
11. medical\_specialty
12. diabetes\_med

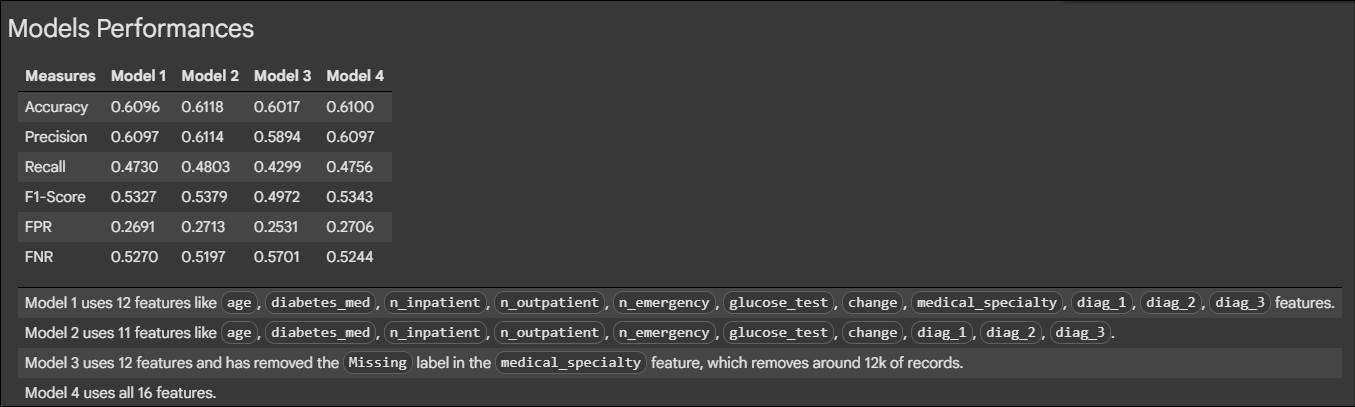
Based on the EDA, we have found that the medical\_specialty feature has a large proportion of Missing labels in it. We will build two models (one excluding the feature and one excluding the rows containing the Missing labels in the feature) to compare its performance with the first model.

Lastly, we will build the last model with all the features.

Please refer to the [Capstone - Patient Readmission Prediction Model.ipynb](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/03-Analyse/Capstone%20-%20Patient%20Readmission%20Prediction%20Model.ipynb) file for the code cells creating the predictive model.

### 

### Model Evaluation



The purpose of creating the predictive model is to identify patients who are at high-risk of readmission to allow doctors to focus on follow-up efforts.

We observed similar performance in terms of accuracy and precision across the models with Model 2 performing slightly better at 61.18% and 61.14% respectively .

Accuracy indicates how well the model predicts the outcome (readmitted or not) in actual dataset. This means the model is able to predict whether a patient is readmitted or not correctly around 6 out of 10 times.

Precision indicates how well the model predicts if a patient will be readmitted. This means among all the patients the model predicted would be readmitted, it was correct around 6 out of 10 times.

Given the priority of the hospital is to identify patients who have actual readmission, we look at the Recall metric for the models. Recall is how well the model can identify a patient who is actually readmitted. Based on the Model 2’s recall of 48.03%, the model 2 is able to identify close to 5 out 10 patients who were actually readmitted.

The Model 2 shows the highest False Positive Rate (27.13%) among the models but the difference among other models is relatively small. This value may come at the expense of small improvement in the Recall value for the model.

The Model 2 shows a lower False Negative Rate (51.97%) among the models, which means the model was better at identifying patients who were actually readmitted. A False Negative Rate is the proportion of actual readmissions that the model missed and a lower rate means it missed fewer actual readmissions.

Model 2’s feature importance provides us a guideline on which are the features that have an impact on the readmission probability. This helps the doctors to focus on which group of patients which are having higher risk of readmission. A positive coefficient increases the risk of readmission. The larger the absolute value, the more influential the feature. For example, a patient who had a prior inpatient visit is more than twice as likely to be readmitted compared to those who have not had a prior inpatient visit, all else being equal.

|  |  |
| --- | --- |
| **Feature** | **Coefficient** |
| has\_inpatient\_visit | 2.10 |
| has\_emergency\_visit | 1.54 |
| has\_outpatient\_visit | 1.53 |
| diabetes\_med\_yes | 1.27 |
| age\_[80-90) | 1.18 |
| diag\_1\_Diabetes | 1.18 |
| age\_[70-80) | 1.16 |
| diag\_2\_Musculoskeletal | 1.09 |
| age\_[60-70) | 1.07 |

# Recommendations

Based on the EDA and model evaluation, we suggest the hospital:

1. piloting the prediction model to help with patient readmission reduction effort. While the model currently identifies close to 5 out of 10 of patients who were actually readmitted, these represent the highest-risk individuals according to our analysis. Implementing targeted follow-up for these identified high-risk individuals could still contribute to reducing overall readmission rates.
2. to focus on the patient group who are strongly associated with the following characteristics as it has significant influence to a patient’s readmission probability:
   1. has previous hospital visit before the hospital stay (especially with inpatient, emergency , and outpatient visit),
   2. has been prescribed with diabetes medication during the hospital stay,
   3. has Diabetes diagnosed as the Primary diagnosis,
   4. has Musculoskeletal diagnosed as the Secondary diagnosis,
   5. are within the 70-90 age group.

Doctors can use the Prediction Dashboard to find out the readmission probability of the patient to plan for follow-up effort.

# Next Steps

For future improvement of the predictive model, we can:

1. explore strategies like adjusting the classification threshold to improve the Recall although it will increase the False Positive Rate.
2. explore feature engineering like categorising the intensity of procedures, lab procedures or medications.
3. train the predictive model using a larger dataset to allow the model to learn from a diverse dataset.

# Appendix

* Acknowledgments: [Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.](https://onlinelibrary.wiley.com/doi/10.1155/2014/781670)
* [Patient Readmission EDA Tableau Workbook](https://github.com/lxyong/dab-capstone/blob/3aff98bb3e531bed33c03c262055e46ba8141d22/03-Analyse/Patient%20Readmission%20EDA.twbx)
* [Patient Readmission Dashboard Tableau Workbook](https://github.com/lxyong/dab-capstone/blob/3aff98bb3e531bed33c03c262055e46ba8141d22/04-Communication/Patient%20Readmission%20Dashboard.twbx)
* [Guide - Integrating the Predictive Model with Tableau using TabPy](https://github.com/lxyong/dab-capstone/blob/825779b36e07d3d29e43df5c82dbae0db90b28a2/02-Data-Prep/Guide%20-%20Integrating%20the%20Predictive%20Model%20with%20Tableau%20using%20TabPy.docx)
* [Capstone - Patient Readmission Prediction Model Jupyter Notebook](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/03-Analyse/Capstone%20-%20Patient%20Readmission%20Prediction%20Model.ipynb)
* [Capstone - Patient Readmission TabPy Deploy Script Jupyter Notebook](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/03-Analyse/Capstone%20-%20Patient%20Readmission%20TabPy%20Deploy%20Script.ipynb)
* [Patient Readmission Prediction Dashboard Tableau Workbook](https://github.com/lxyong/dab-capstone/blob/b2cde31ddec713513eae7834a4dc6ff095ef306f/04-Communication/Patient%20Readmission%20Prediction%20Dashboard.twb)