

Phase 3: Exploratory Data Analysis

Masheia Dzimba and Peter Mangoro

2025-11-30

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 1 |
| 2 | Load Cleaned Data | 1 |
| 3 | Univariate Analysis(summary statistics for numerical columns) | 2 |
| 4 | Bivariate Analysis | 2 |
| 4.1 | Summary Statistics by Readmission Status | 2 |
| 4.2 | Boxplots: Numerical Variables by Readmission Status | 4 |
| 4.3 | Categorical Variables by Readmission Status | 4 |
| 4.4 | Readmission Rates by Category | 4 |
| 5 | Statistical Tests | 5 |
| 5.1 | T-tests for Numerical Variables | 5 |
| 5.2 | Chi-square Tests for Categorical Variables | 6 |
| 6 | Key Findings | 7 |
| 6.1 | Summary of Key Findings | 7 |
| 7 | Summary | 7 |

1 Introduction

This document presents Phase 3: Exploratory Data Analysis (EDA). We examine relationships between predictors and readmission status, create comprehensive visualizations, and perform statistical tests.

2 Load Cleaned Data

Dataset: 24996 observations, 21 variables

3 Univariate Analysis(summary statistics for numerical columns)

```

time_in_hospital n_lab_procedures n_procedures n_medications
Min. : 1.000 Min. : 1.00 Min. :0.000 Min. : 1.00
1st Qu.: 2.000 1st Qu.: 31.00 1st Qu.:0.000 1st Qu.:11.00
Median : 4.000 Median : 44.00 Median :1.000 Median :15.00
Mean : 4.453 Mean : 43.24 Mean :1.352 Mean :16.25
3rd Qu.: 6.000 3rd Qu.: 57.00 3rd Qu.:2.000 3rd Qu.:20.00
Max. :14.000 Max. :113.00 Max. :6.000 Max. :79.00

n_outpatient n_inpatient n_emergency n_diagnoses
Min. : 0.0000 Min. : 0.0000 Min. : 0.0000 Min. :1.00
1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.:3.00
Median : 0.0000 Median : 0.0000 Median : 0.0000 Median :3.00
Mean : 0.3664 Mean : 0.6161 Mean : 0.1865 Mean :2.99
3rd Qu.: 0.0000 3rd Qu.: 1.0000 3rd Qu.: 0.0000 3rd Qu.:3.00
Max. :33.0000 Max. :15.0000 Max. :64.0000 Max. :3.00

medications_per_day total_previous_visits
Min. : 0.1429 Min. : 0.000
1st Qu.: 2.6250 1st Qu.: 0.000
Median : 4.0000 Median : 0.000
Mean : 5.1289 Mean : 1.169
3rd Qu.: 6.3333 3rd Qu.: 2.000
Max. :40.0000 Max. :68.000

```

4 Bivariate Analysis

4.1 Summary Statistics by Readmission Status

Table 1: Summary Statistics: Original Proposal Variables

| | time_in_hospital | n_lab_procedures | n_procedures | n_medications |
|----------------|------------------|------------------|--------------|---------------|
| readmitted | | | | |
| Not Readmitted | 4.33 (3) | 42.62 (20.09) | 1.42 (1.75) | 15.97 (8.45) |
| Readmitted | 4.59 (3) | 43.94 (19.49) | 1.27 (1.68) | 16.57 (7.58) |

Table 2: Summary Statistics: Previous Visits

| | n_outpatient | n_inpatient | n_emergency |
|----------------|--------------|-------------|-------------|
| readmitted | | | |
| Not Readmitted | 0.26 (0.93) | 0.38 (0.82) | 0.11 (0.55) |
| Readmitted | 0.49 (1.43) | 0.88 (1.43) | 0.27 (1.15) |

Table 3: Summary Statistics: Feature-Engineered Variables

| | n_diagnoses | medications_per_day | total_previous_visits |
|----------------|-------------|---------------------|-----------------------|
| readmitted | | | |
| Not Readmitted | 2.99 (0.12) | 5.22 (4.04) | 0.75 (1.51) |
| Readmitted | 2.99 (0.08) | 5.03 (3.72) | 1.64 (2.62) |

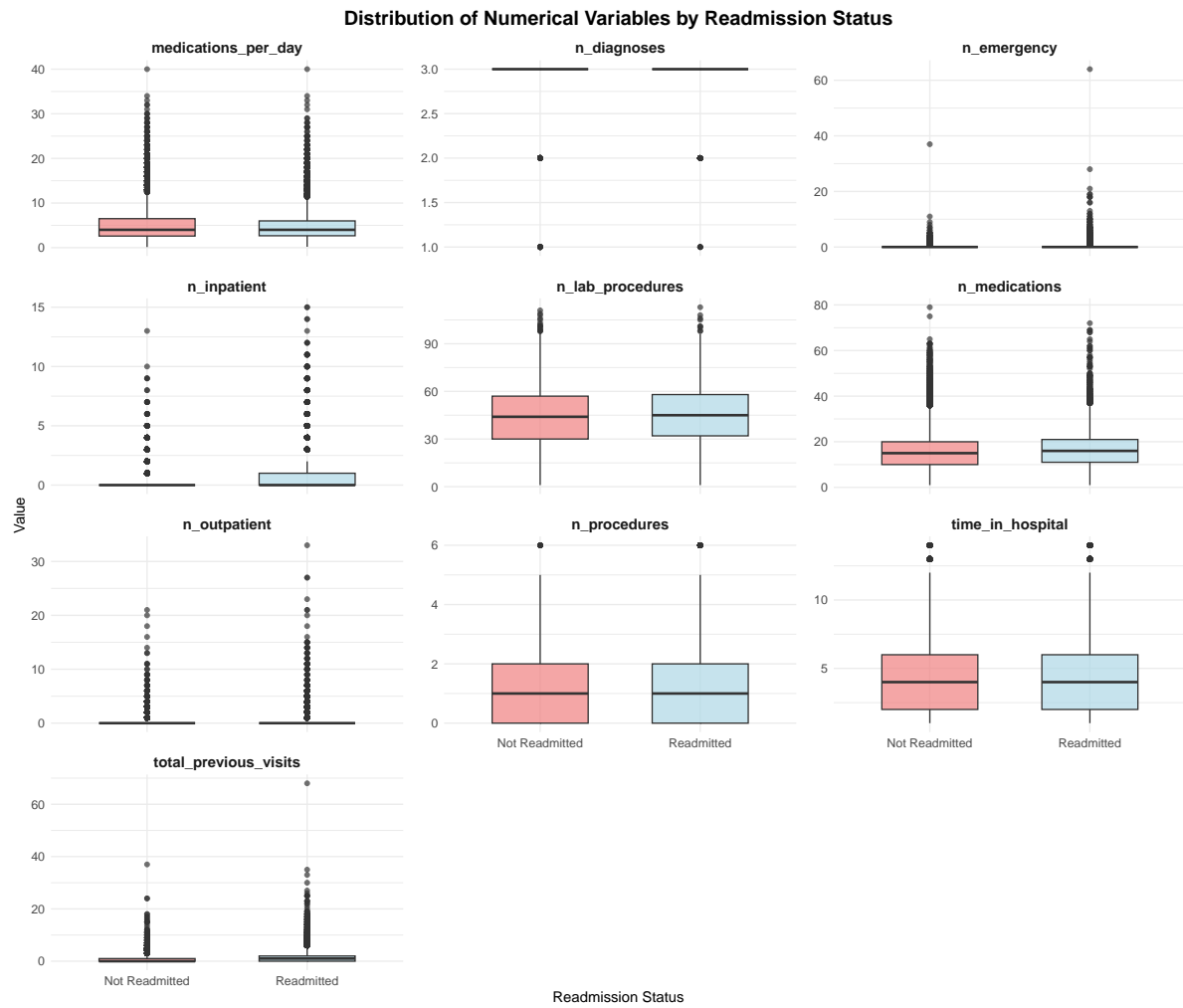


Figure 1: Distribution of Numerical Variables by Readmission Status

4.2 Boxplots: Numerical Variables by Readmission Status

4.3 Categorical Variables by Readmission Status

Table 4: Age by Readmission Status

| | Not Readmitted | Readmitted |
|----------|----------------|------------|
| [40-50) | 1405 | 1127 |
| [50-60) | 2486 | 1966 |
| [60-70) | 3142 | 2770 |
| [70-80) | 3501 | 3335 |
| [80-90) | 2276 | 2238 |
| [90-100) | 434 | 316 |

4.4 Readmission Rates by Category

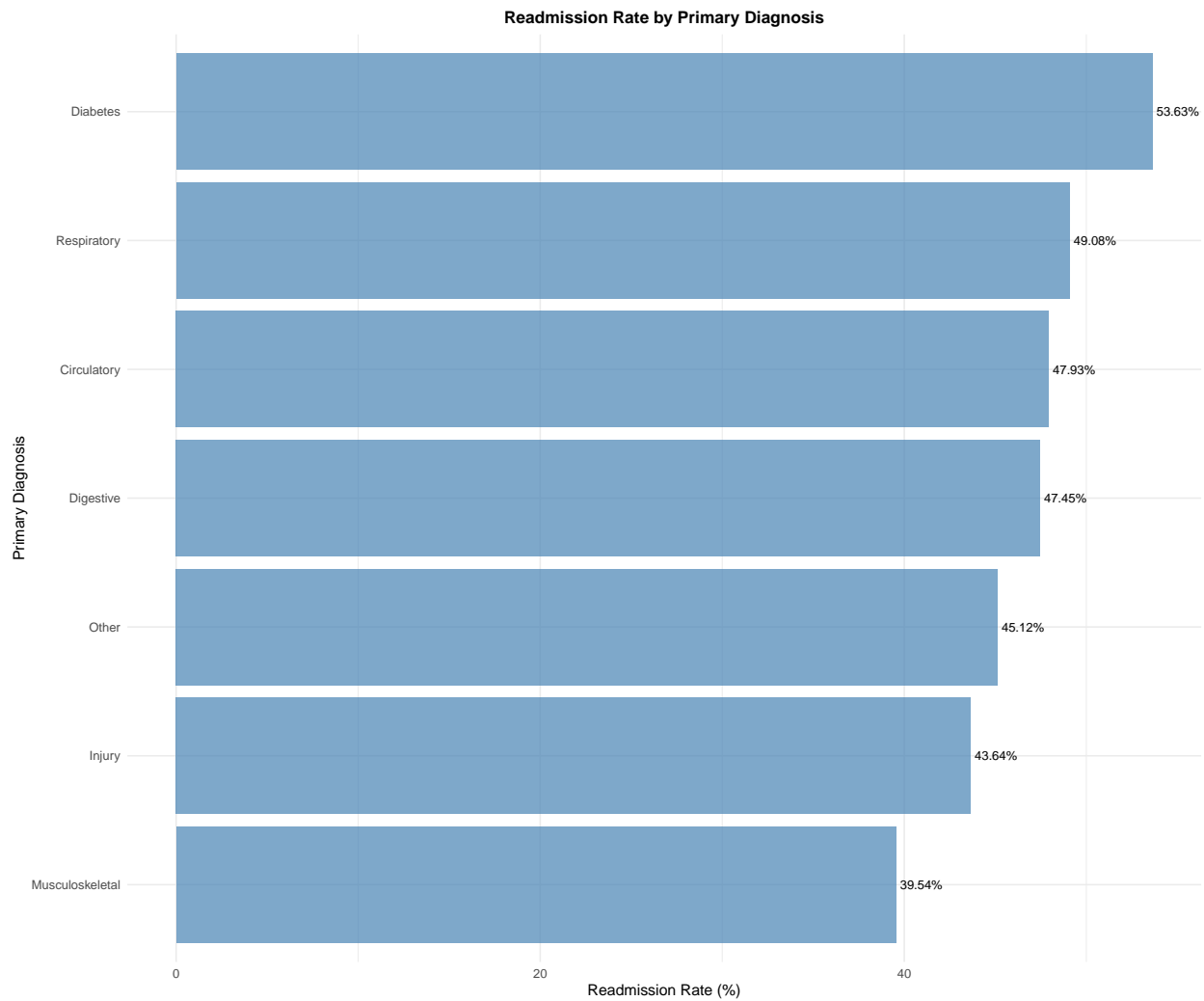


Figure 2: Readmission Rate by Primary Diagnosis

5 Statistical Tests

5.1 T-tests for Numerical Variables

time_in_hospital:
Mean (Readmitted): 4.59
Mean (Not Readmitted): 4.33
P-value: 0.0000
Significant difference ($p < 0.05$)

n_lab_procedures:
Mean (Readmitted): 43.94
Mean (Not Readmitted): 42.62
P-value: 0.0000
Significant difference ($p < 0.05$)

n_procedures:
Mean (Readmitted): 1.27
Mean (Not Readmitted): 1.42
P-value: 0.0000
Significant difference ($p < 0.05$)

n_medications:
Mean (Readmitted): 16.57
Mean (Not Readmitted): 15.97
P-value: 0.0000
Significant difference ($p < 0.05$)

n_outpatient:
Mean (Readmitted): 0.49
Mean (Not Readmitted): 0.26
P-value: 0.0000
Significant difference ($p < 0.05$)

n_inpatient:
Mean (Readmitted): 0.88
Mean (Not Readmitted): 0.38
P-value: 0.0000
Significant difference ($p < 0.05$)

n_emergency:
Mean (Readmitted): 0.27
Mean (Not Readmitted): 0.11
P-value: 0.0000
Significant difference ($p < 0.05$)

n_diagnoses:
Mean (Readmitted): 2.99
Mean (Not Readmitted): 2.99
P-value: 0.0000
Significant difference ($p < 0.05$)

medications_per_day:

Mean (Readmitted): 5.03
Mean (Not Readmitted): 5.22
P-value: 0.0002
Significant difference ($p < 0.05$)

total_previous_visits:
Mean (Readmitted): 1.64
Mean (Not Readmitted): 0.75
P-value: 0.0000
Significant difference ($p < 0.05$)

5.2 Chi-square Tests for Categorical Variables

Chi-square tests for categorical variables:

age:
Chi-square statistic: 48.700
P-value: 0.0000
Significant association ($p < 0.05$)

medical_specialty:
Chi-square statistic: 85.557
P-value: 0.0000
Significant association ($p < 0.05$)

diag_1:
Chi-square statistic: 84.895
P-value: 0.0000
Significant association ($p < 0.05$)

change:
Chi-square statistic: 46.867
P-value: 0.0000
Significant association ($p < 0.05$)

diabetes_med:
Chi-square statistic: 96.244
P-value: 0.0000
Significant association ($p < 0.05$)

glucose_test:
Chi-square statistic: 7.755
P-value: 0.0207
Significant association ($p < 0.05$)

A1Ctest:
Chi-square statistic: 14.822
P-value: 0.0006
Significant association ($p < 0.05$)

6 Key Findings

6.1 Summary of Key Findings

1. Overall Readmission Rate: 47.02 %

2. Numerical Variables - Mean Differences:

time_in_hospital: Readmitted (4.59) vs. Not Readmitted (4.33), Difference: 0.26
n_lab_procedures: Readmitted (43.94) vs. Not Readmitted (42.62), Difference: 1.32
n_procedures: Readmitted (1.27) vs. Not Readmitted (1.42), Difference: -0.15
n_medications: Readmitted (16.57) vs. Not Readmitted (15.97), Difference: 0.60
n_outpatient: Readmitted (0.49) vs. Not Readmitted (0.26), Difference: 0.23
n_inpatient: Readmitted (0.88) vs. Not Readmitted (0.38), Difference: 0.50
n_emergency: Readmitted (0.27) vs. Not Readmitted (0.11), Difference: 0.17
n_diagnoses: Readmitted (2.99) vs. Not Readmitted (2.99), Difference: 0.01
medications_per_day: Readmitted (5.03) vs. Not Readmitted (5.22), Difference: -0.18
total_previous_visits: Readmitted (1.64) vs. Not Readmitted (0.75), Difference: 0.90

7 Summary

This phase revealed important patterns:

- **All numerical variables** show significant differences between readmitted and not readmitted groups
- **All categorical variables** show significant associations with readmission
- **Previous visits** appear to be a key predictor
- **Age and diagnosis** are important factors