Phase 1: Data Loading and Initial Exploration

Masheia Dzimba and Peter Mangoro

2025-11-26

Contents

1 Introduction

This document presents Phase 1 of the project: Data Loading and Initial Exploration. In this phase, we load the hospital readmissions dataset, examine its structure, identify missing values, and create initial visualizations.

2 Data Loading

```
Dataset Dimensions:

Rows (observations): 25000

Columns (variables): 17
```

3 Dataset Structure

```
’data.frame’:   25000 obs. of  17 variables:
 $ age              : chr  "[70-80)" "[70-80)" "[50-60)" "[70-80)" ...
 $ time_in_hospital : int  8 3 5 2 1 2 4 1 4 8 ...
 $ n_lab_procedures : int  72 34 45 36 42 51 44 19 67 37 ...
 $ n_procedures     : int  1 2 0 0 0 0 2 6 3 1 ...
 $ n_medications    : int  18 13 18 12 7 10 21 16 13 18 ...
 $ n_outpatient     : int  2 0 0 1 0 0 0 0 0 0 ...
 $ n_inpatient      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ n_emergency      : int  0 0 0 0 0 0 0 1 0 0 ...
 $ medical_specialty: chr  "Missing" "Other" "Missing" "Missing" ...
 $ diag_1           : chr  "Circulatory" "Other" "Circulatory" "Circulatory" ...
 $ diag_2           : chr  "Respiratory" "Other" "Circulatory" "Other" ...
 $ diag_3           : chr  "Other" "Other" "Circulatory" "Diabetes" ...
 $ glucose_test     : chr  "no" "no" "no" "no" ...
 $ A1Ctest          : chr  "no" "no" "no" "no" ...
 $ change           : chr  "no" "no" "yes" "yes" ...
 $ diabetes_med     : chr  "yes" "yes" "yes" "yes" ...
 $ readmitted       : chr  "no" "no" "yes" "yes" ...
```

```
 Column Names

 [1] "age"               "time_in_hospital"  "n_lab_procedures"
 [4] "n_procedures"      "n_medications"     "n_outpatient"
 [7] "n_inpatient"       "n_emergency"       "medical_specialty"
[10] "diag_1"            "diag_2"            "diag_3"
[13] "glucose_test"      "A1Ctest"           "change"
[16] "diabetes_med"      "readmitted"
```

```
 Data Types Summary

data_types
character   integer
       10         7
```

4 Missing Values Analysis

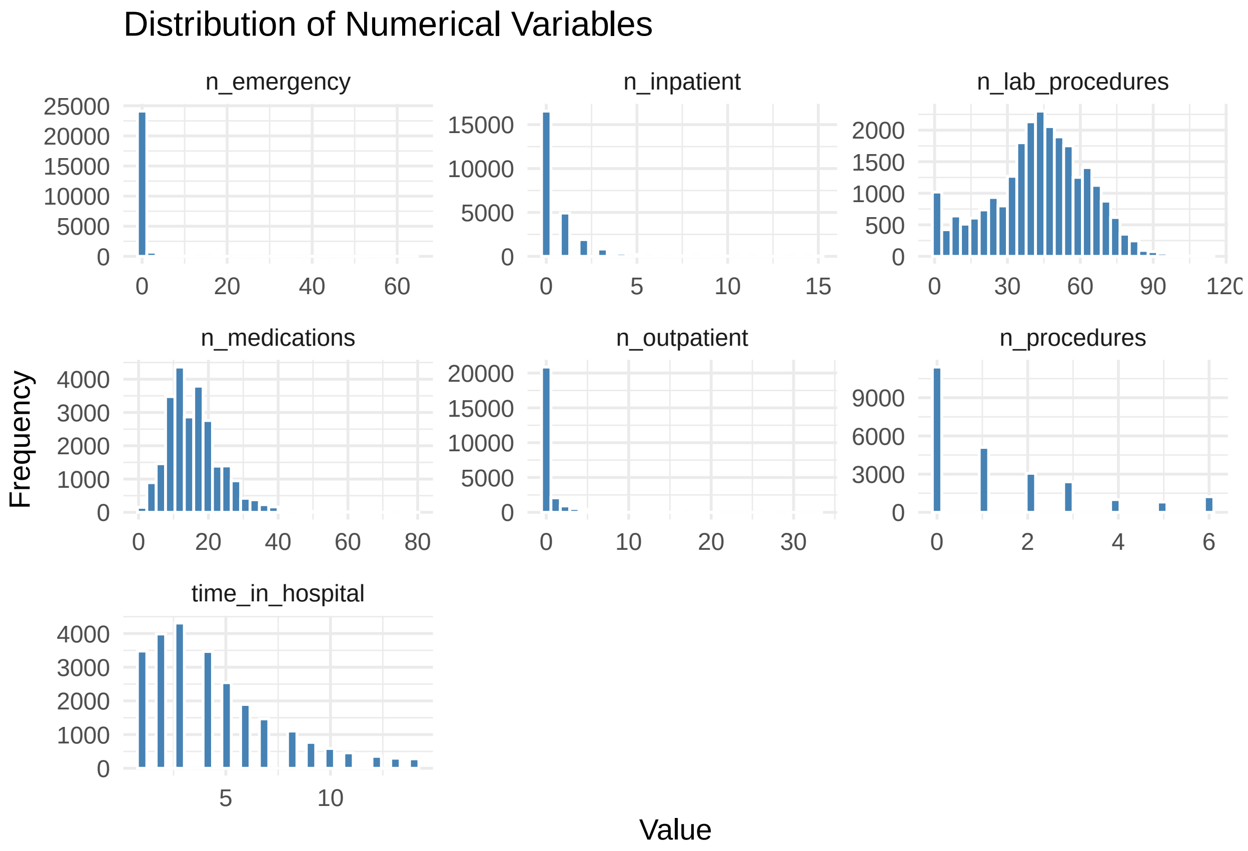
```
Columns with missing values:
```

Table 1: Missing Values Summary

	Variable	Missing_Count	Missing_Percentage
medical_specialty	medical_specialty	12382	49.53
diag_1	diag_1	4	0.02
diag_2	diag_2	42	0.17
diag_3	diag_3	196	0.78

5 Summary Statistics

```
 Numerical Variables Summary
 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     
 Min.   : 0.0000   Min.   : 0.000   Min.   : 0.0000  
 1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.0000  
 Median : 0.0000   Median : 0.000   Median : 0.0000  
 Mean   : 0.3664   Mean   : 0.616   Mean   : 0.1866  
 3rd Qu.: 0.0000   3rd Qu.: 1.000   3rd Qu.: 0.0000  
 Max.   :33.0000   Max.   :15.000   Max.   :64.0000  
```



# Response Variable Distribution

```
Unique values in 'readmitted' column:
```

```
[1] "no"  "yes"
```

```

   no   yes 
13246 11754 
```

```
Proportions:
```

```
   no   yes
52.98 47.02
```

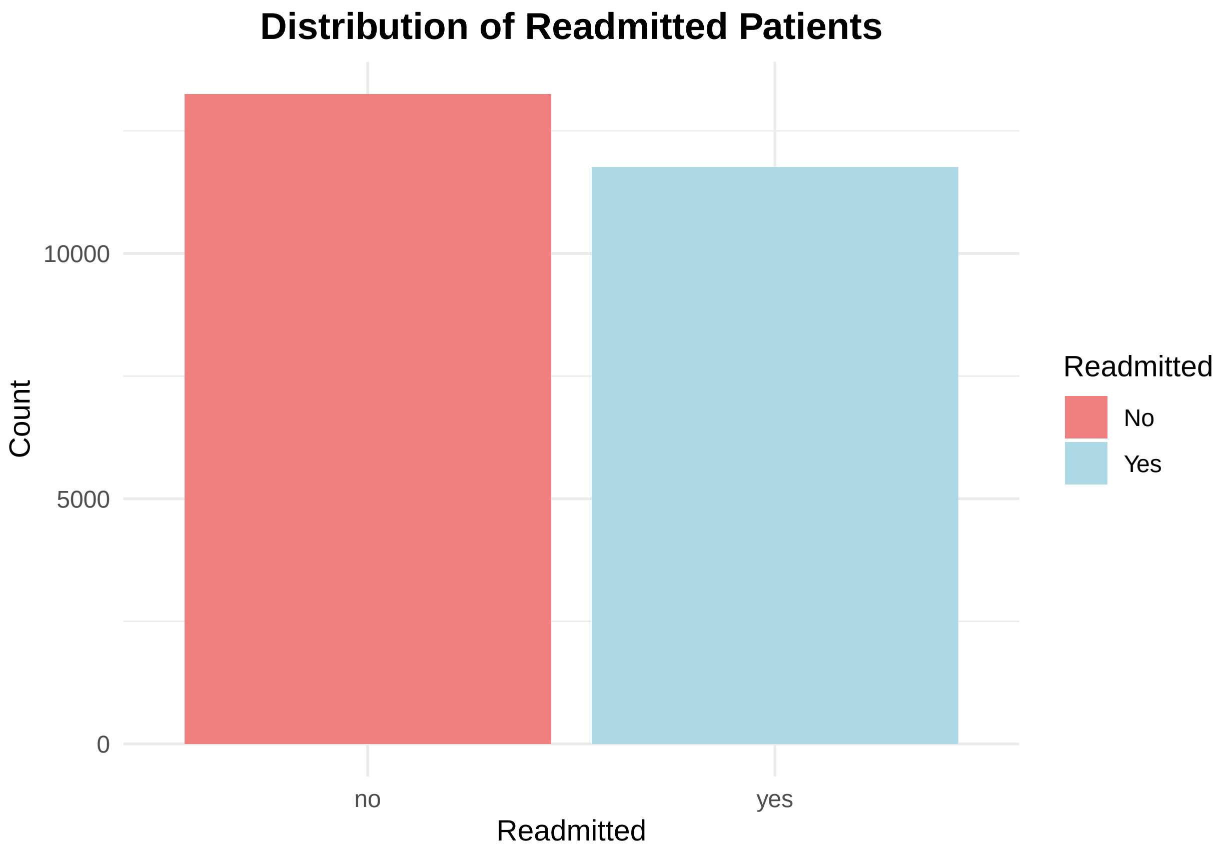


Figure 1: Distribution of Readmitted Patients

6 Summary

This phase successfully loaded and explored the dataset. Key findings:

- **Dataset size**: 25,000 observations, 17 variables
- **Response variable**: Binary (yes/no) with 47.02% readmission rate
- **Missing values**: Found in medical_specialty (49.53%), diag_1 (0.02%), diag_2 (0.17%), diag_3 (0.78%)
- **Data types**: 10 categorical, 7 numerical variables