Using Machine Learning for Prediction of Early Readmission of Diabetic Patients

Literature Review

CIND 820- Big Data Analytics Project

Supervised by: Ceni Babaoglu

Presented by: Nehal Gamal Mohamed (501278190)

Ryerson University

Table of Contents

Revised Abstract

Diabetes management poses a significant challenge in healthcare, with patient readmission within 30 days of discharge serving as a critical metric for assessing care quality. Despite advances in preventive interventions, many diabetic patients experience readmissions due to suboptimal glycemic control and inadequate care.

This study will try to predict the likelihood of early readmission (within 30 days) for patients diagnosed with diabetes using clinical data collected over a ten-year period (1999-2008) from 130 US hospitals and integrated delivery networks.

The study objective is to answer the following questions:

- By using machine learning models can we accurately predict early readmission of diabetic patients?
- What are the patient and hospital factors which strongly influence early readmission?
- By using predictive models how can we improve diabetes management and reduce readmission rates?

The dataset has patient records from diabetic encounters, including demographic information, medical history, admission details, laboratory results, medications given, and hospital outcomes. It has 101,766 instances and 50 features, the data is multivariate, consisting of categorical and integer variables.

The study will commence with preparing the dataset, then employing classification techniques to predict early readmission outcomes based on patient and hospital features. Machine learning algorithms including logistic regression, decision trees, and random forests will be utilized.

After splitting the dataset and training the model, the performance will be evaluated using standard metrics such as accuracy, precision, recall and F1 score. Moreover, feature selection and dimensionality reduction techniques will be applied to identify the most informative factors and enhance model performance. Additionally, clustering algorithms will be explored to check if there are hidden patterns that could reveal any patient subgroups with different readmission risks.

Python programming will be used for implementation, to take advantage of various python libraries and tools.

Introduction

Hospital readmission, especially among diabetic patients, creates a challenge due to its implications for patient outcomes and healthcare resources. According to the Centers for Medicare & Medicaid Services (CMS), nearly one in five Medicare patients discharged from a hospital is readmitted within 30 days, resulting in approximately $26 billion in annual costs. Diabetes, affecting over 34 million Americans, frequently leads to complications that necessitate hospital readmissions.

The COVID-19 pandemic has deepened this issue, as diabetic patients are at a higher risk for severe complications from COVID-19. Studies have shown that the pandemic has led to an increase in hospital admissions and readmissions among diabetic patients due to the virus's direct and indirect effects. For instance, research indicates that diabetic patients with COVID-19 are more likely to experience severe outcomes, including hospital readmission, due to worsened glycemic control and increased inflammation. A study published in Diabetes Care found that COVID-19 patients with diabetes had a 30-day readmission rate of 10.4%, significantly higher than the general population.

Readmissions not only indicate potential issues in the quality of care but also place a significant financial burden on the healthcare system. Studies have shown that diabetic patients are twice as likely to be readmitted compared to non-diabetic patients. Common factors contributing to readmissions include inadequate post-discharge care, medication non-adherence, and the presence of comorbid conditions. The pandemic has further highlighted these issues, as the disruptions in routine care, delayed medical attention, and increased psychological stress have contributed to poorer health outcomes for diabetic patients.

Addressing these issues through effective management and predictive modeling could potentially reduce the high rates of readmission, thereby improving patient outcomes and reducing healthcare costs. By utilizing machine learning techniques to predict readmissions, healthcare providers can identify high-risk patients early and implement targeted interventions to improve care quality, particularly during challenging times like the COVID-19 pandemic.

Methodology and Contribution of Current Study

This study builds on the existing body of research by integrating advanced Machine Learning (ML) techniques and comprehensive data preprocessing methods to enhance predictive accuracy. By using a robust, large dataset and employing comprehensive feature selection, along with

exploring clustering algorithms to check for hidden patterns that could reveal patient subgroups with different readmission risks, the study aims to provide actionable insights for healthcare providers.

Moreover, previous studies have addressed the same problem, with some utilizing the same dataset. These studies employed a variety of tools such as R statistical software, Hadoop and Spark. However, this study utilizes Python programming and its libraries, providing a unique approach to the analysis. The study is valuable in the context of previous studies because it addresses existing gaps and leverages Python programming for enhanced predictive accuracy and healthcare impact.

Comprehensive Findings from Reviewed Research Papers

As mentioned before diabetic patients have higher readmission rates compared to the general population, often due to poor disease control and complications. Reducing readmissions can significantly lower healthcare costs and improve patient outcomes. Various ML models, such as Random Forest (RF), Naive Bayes (NB), and decision tree, have been used to predict 30-day readmissions. RF models have consistently shown superior performance in predicting readmissions. Important features include patient demographics (age, sex, race), clinical factors (number of diagnoses, length of stay, medication use), and healthcare utilization patterns (number of inpatient admissions, emergency visits).

Although RF models outperformed other algorithms in predicting readmissions, there was a potential for overfitting which needs careful management. RF models were robust in handling various types of data, and their ability to provide feature importance measures made them a preferred choice. Thorough data preprocessing is crucial for building accurate predictive models, including handling missing values, normalizing data, and selecting relevant features. Studies emphasized removing attributes with high missing values or irrelevant features to improve model performance. Techniques like down-sampling and over-sampling Synthetic Minority Oversampling Technique (SMOTE) are employed to address class imbalances in the datasets, enhancing model accuracy.

Overview of Previous Research Papers

Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records

This study highlighted the importance of HbA1c measurement in predicting readmissions among diabetic patients. The analysis revealed that better diabetes management, reflected in HbA1c levels, could potentially reduce readmission rates. ML models were used to identify high-risk patients, focusing on the influence of HbA1c levels. However, the study focused on a single predictor (HbA1c) and lacked the need for incorporating multiple predictors to enhance model accuracy.

Implementation of Big Data Analytics on Diabetes 130-US Hospitals for Year 1999-2008 for Predicting Patient Readmission & The 30-days Hospital Readmission Risk in Diabetic Patients: Predictive Modeling with Machine Learning Classifiers

Both studies utilized the same Health Facts Database to develop predictive models for 30-day readmissions. Various ML algorithms, including RF, NB, and decision trees, were employed, with RF showing the best performance. Key predictors identified include race, sex, age, admission type, admission location, length of stay, and drug use. Both studies concluded that RF is more suitable for making readmission predictions and emphasized the importance of identifying high-risk patients to reduce the probability of readmission within 30 days. These studies had a comprehensive approach and used a large dataset, but the potential for overfitting with RF models needs to be considered carefully. The focus on multiple predictors enhances the model's accuracy, making it more applicable in real-world settings. However, findings need to be validated in clinical settings.

Hospital Readmission and Length-of-Stay Prediction Using an Optimized Hybrid Deep Model

The study introduced the Genetic Algorithm-Optimized Convolutional Neural Network GAOCNN model, a hybrid deep learning model, for predicting readmissions and length of stay. GAOCNN proved robust to missing values and performed well on imbalanced data. Future improvements suggested including optimizing feature extraction and classifier training time. The complexity and computational intensity of the model may limit its practical application in clinical settings, and further research is needed to streamline the model for real-time predictions.

Effective Hospital Readmission Prediction Models Using Machine-Learned Features

The study showed that combining machine-learned features with manual features improved prediction accuracy over traditional models. The ML model was effective in identifying high-risk patients for targeted interventions. The integration of manual and machine-learned features

highlights the importance of comprehensive feature selection. Further validation with larger datasets is necessary to generalize its findings.

Forecasting Hospital Readmissions with Machine Learning

The study employed various ML techniques, including support vector machines and random forests, to predict readmissions using data from a Greek hospital. Balanced RF models achieved the best performance in terms of sensitivity and generalization. However, the study's focus on a single hospital limits the generalizability of the results.

Application of Machine Learning in Predicting Hospital Readmissions: A Scoping Review of the Literature

The study found that tree-based methods, neural networks, and regularized logistic regression are commonly used for predicting hospital readmissions. The performance of these algorithms varies due to different factors, emphasizing the need for external validation. The review highlights the variability in model performance and the need for standardized evaluation metrics. Future efforts should focus on optimizing ML algorithms for clinical integration to improve care quality and reduce costs.

Descriptive Statistics of the Selected Dataset

Total Records: 101,767

Key Attributes: Age, sex, race, number of diagnoses, length of stay, medication use

Missing Values: During data preprocessing, missing values were identified in several attributes, including race, weight, payer_code, medical_specialty, max_glu_serum, A1Cresult, diag_1, diag_2, and diag_3. To manage these missing values, they were categorized into a separate "Missing" subgroup. However, attributes with a high percentage of missing values, such as Weight and Payer_code, will be excluded from the analysis to preserve data quality and ensure robust results.

Total Number of Attributes: 53

Numeric Attributes: 16

Categorical Attributes: 33

Text Attributes: 1

Boolean Attributes: 3

Duplicated Records: 0

Dataset Overview & variables distribution

YData Profiling Report | Overview | Variables | Interactions | Correlations | Missing values | Sample

Overview

Overview | Alerts 44 | Reproduction

Dataset statistics

Number of variables	53
Number of observations	101766
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	248.3 MiB
Average record size in memory	2.5 KiB

Variable types

Numeric	13
Categorical	33
Text	4
Boolean	3

Variables

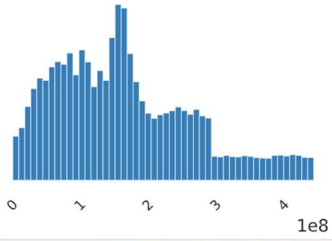
encounter_id

encounter_id

Real number (ℝ)

HIGH CORRELATION UNIQUE

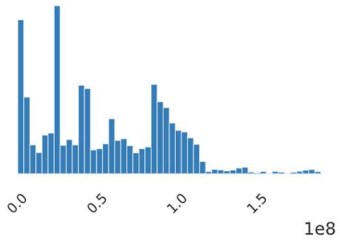
Distinct	101766	Minimum	12522
Distinct (%)	100.0%	Maximum	4.4386722×10^8
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1.6520165×10^8	Memory size	795.2 KiB



patient_nbr
Real number (ℝ)

HIGH_CORRELATION

Distinct	71518	**Minimum**	135
Distinct (%)	70.3%	**Maximum**	1.8950262×10^8
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	54330401	**Memory size**	795.2 KiB



More details

race
Categorical

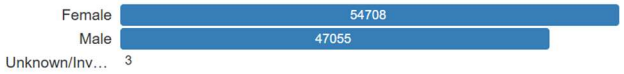
IMBALANCE

Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.5 MiB

More details

gender
Categorical

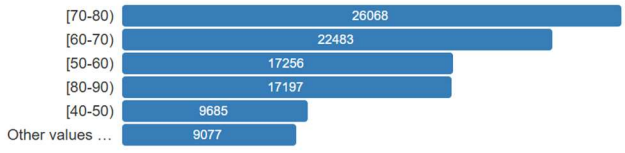
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.0 MiB



More details

age
Categorical

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.2 MiB



More details

weight

Categorical

IMBALANCE

Distinct	10
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.2 MiB

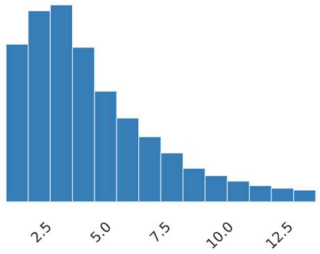
More details

time_in_hospital

Real number (ℝ)

Distinct	14
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	4.3959869

Minimum	1
Maximum	14
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	795.2 KiB

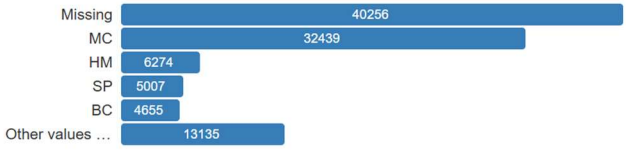


More details

payer_code

Categorical

Distinct	18
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.9 MiB



More details

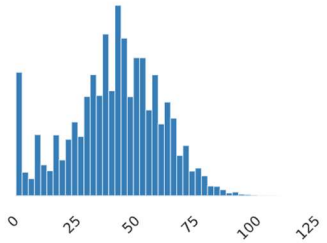
medical_specialty

Text

Distinct	73
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.7 MiB

num_lab_procedures
Real number (ℝ)

Distinct	118	**Minimum**	1
Distinct (%)	0.1%	**Maximum**	132
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	43.095641	**Memory size**	795.2 KiB

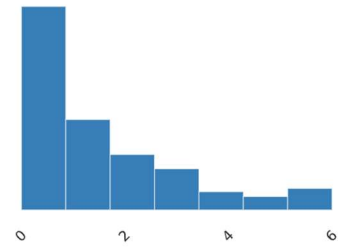


More details

num_procedures
Real number (ℝ)

ZEROS

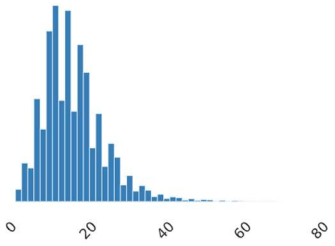
Distinct	7	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	6
Missing	0	**Zeros**	46652
Missing (%)	0.0%	**Zeros (%)**	45.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.3397304	**Memory size**	795.2 KiB



More details

num_medications
Real number (ℝ)

Distinct	75	**Minimum**	1
Distinct (%)	0.1%	**Maximum**	81
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	16.021844	**Memory size**	795.2 KiB



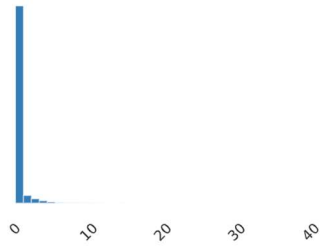
More details

number_outpatient

Real number (ℝ)

ZEROS

Distinct	39	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	42
Missing	0	**Zeros**	85027
Missing (%)	0.0%	**Zeros (%)**	83.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.36935715	**Memory size**	795.2 KiB



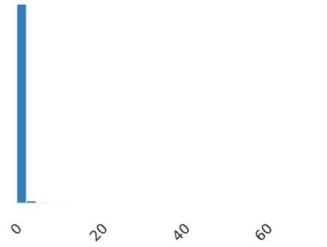
More details

number_emergency

Real number (ℝ)

SKEWED ZEROS

Distinct	33	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	76
Missing	0	**Zeros**	90383
Missing (%)	0.0%	**Zeros (%)**	88.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.19783621	**Memory size**	795.2 KiB



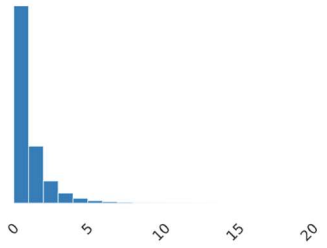
More details

number_inpatient

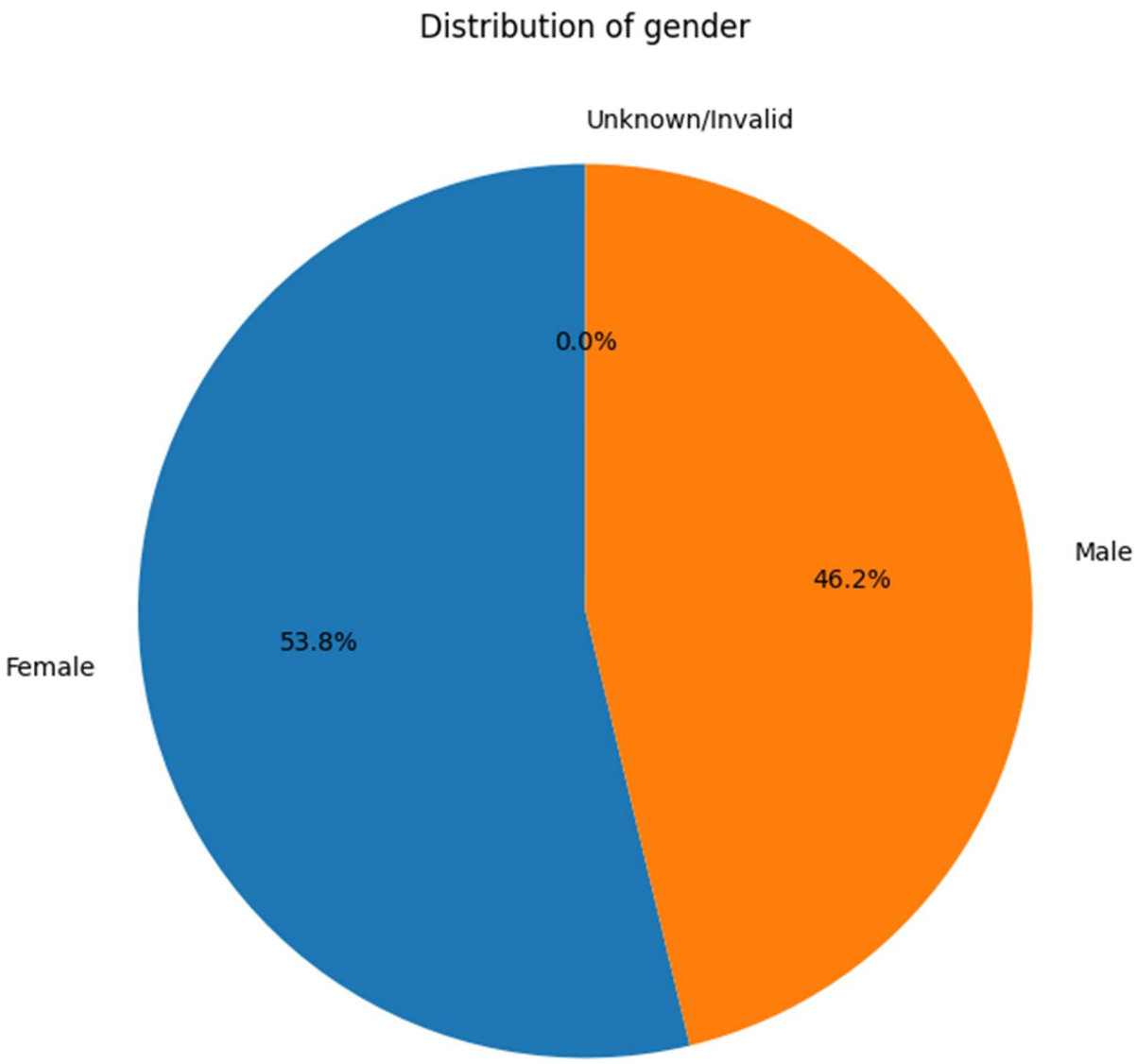
Real number (ℝ)

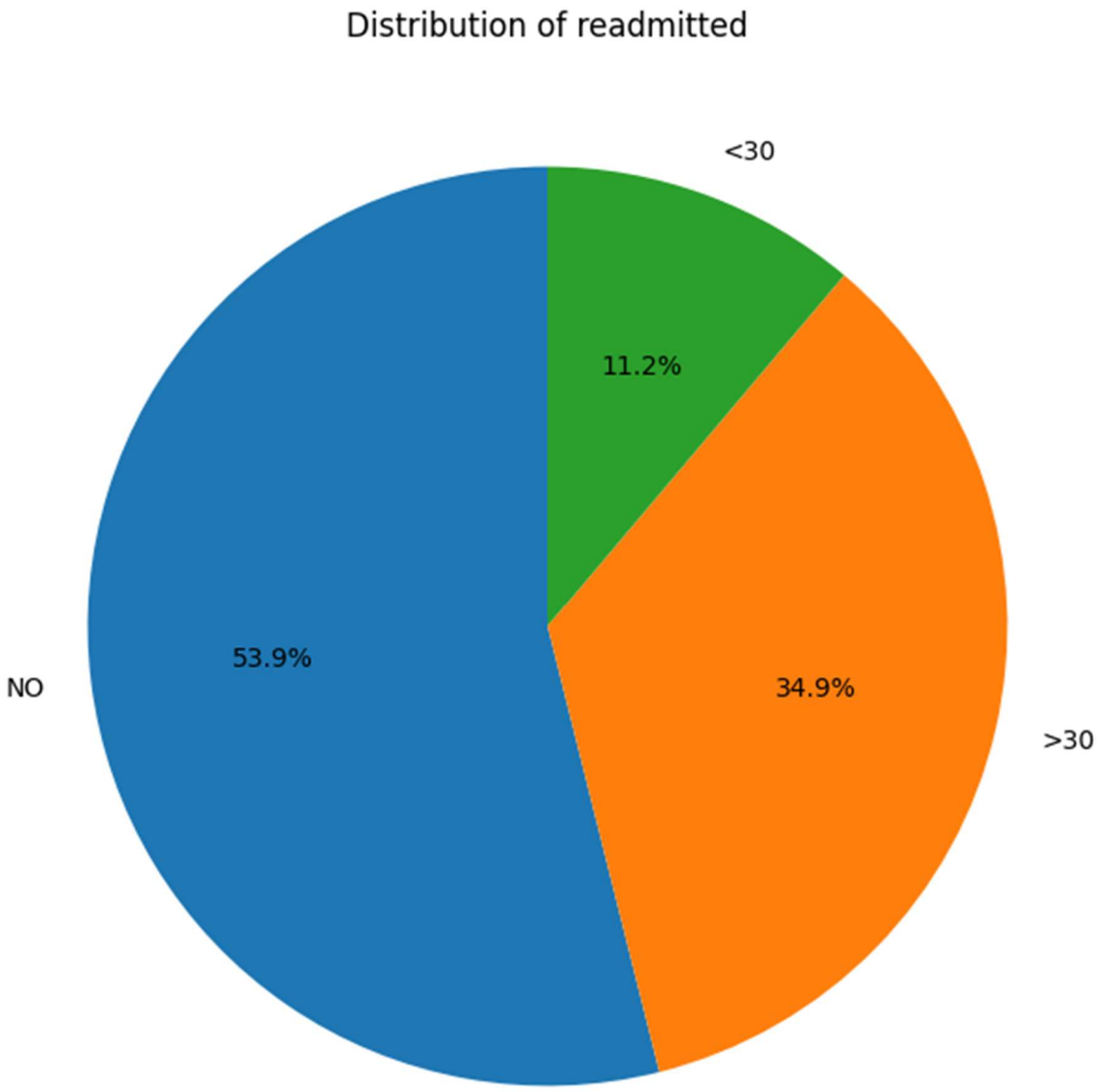
ZEROS

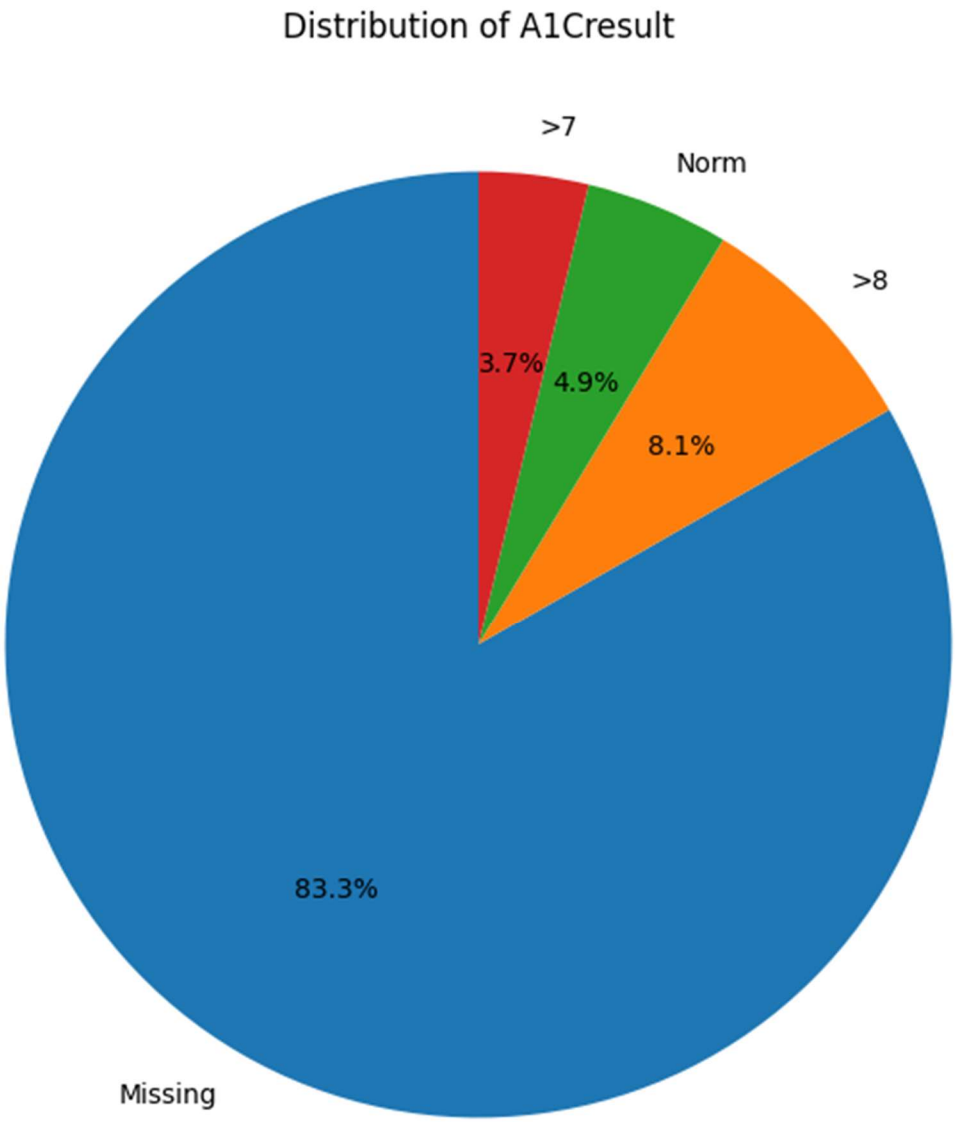
Distinct	21	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	21
Missing	0	**Zeros**	67630
Missing (%)	0.0%	**Zeros (%)**	66.5%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.63556591	**Memory size**	795.2 KiB

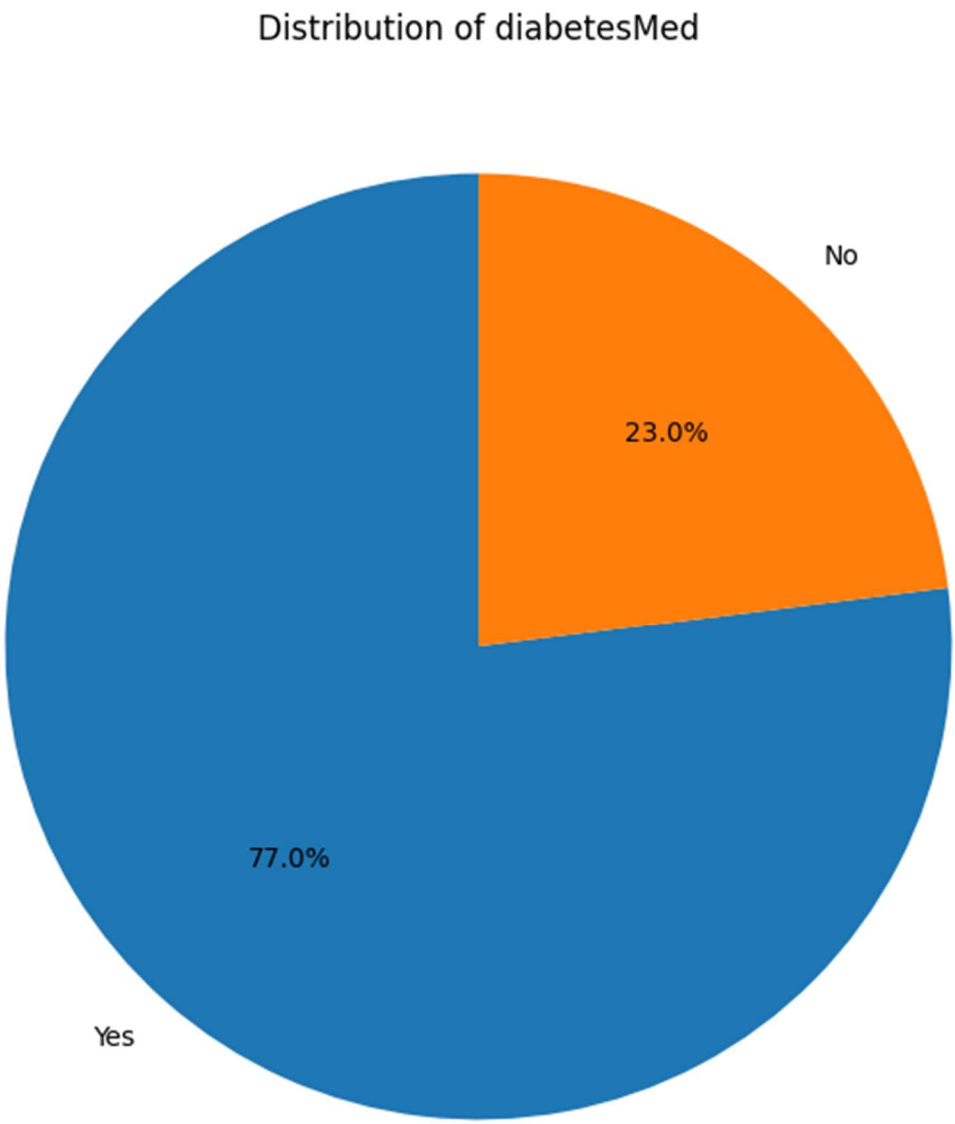


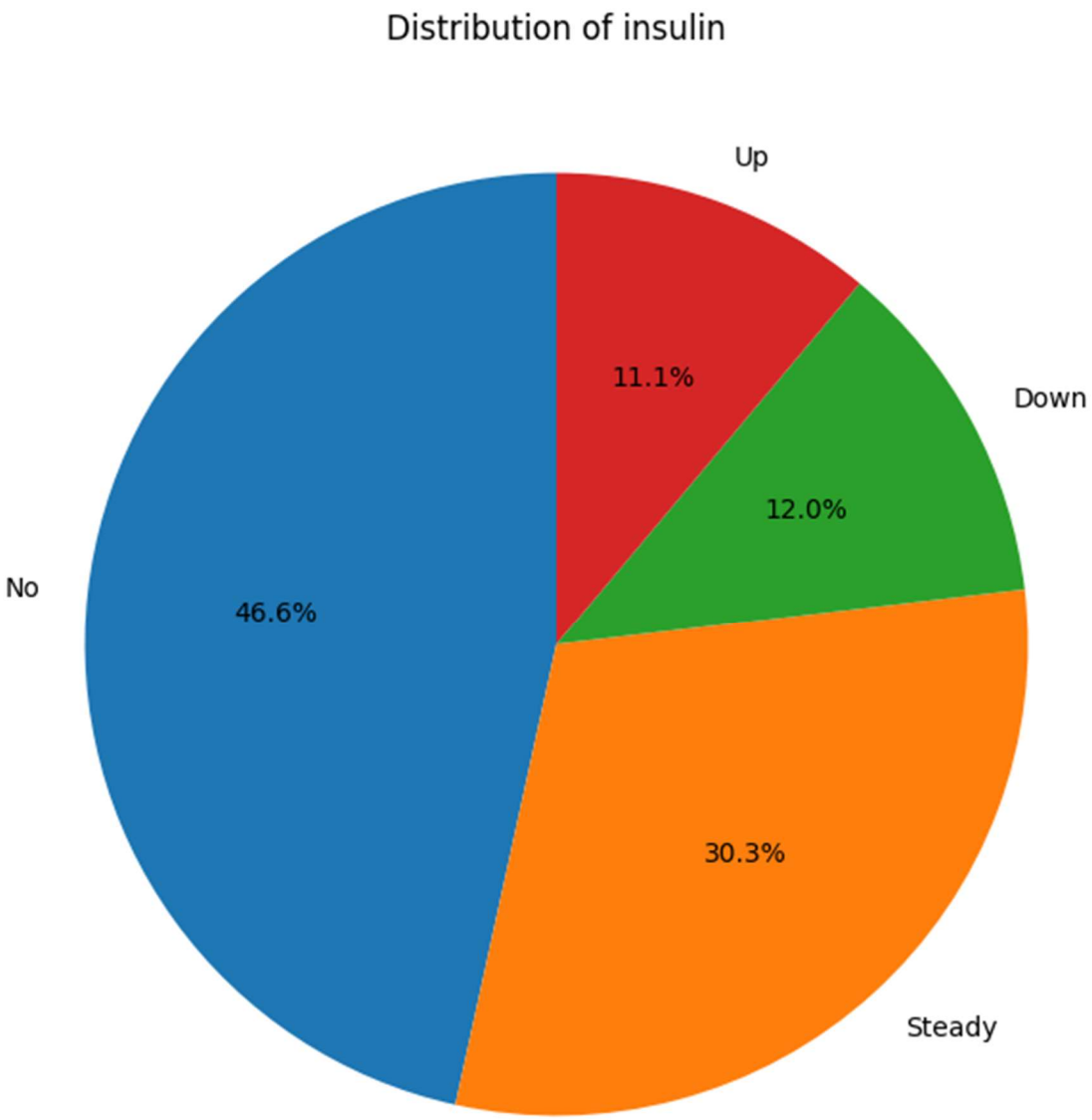
More details


Distribution of gender
Unknown/Invalid
0.0%
Male
46.2%
Female
53.8%


Distribution of readmitted
<30
11.2%
NO
53.9%
34.9%
>30


Distribution of A1Cresult
>7
Norm
>8
3.7%
4.9%
8.1%
83.3%
Missing


Distribution of diabetesMed
No
23.0%
77.0%
Yes


Distribution of insulin
Up
11.1%
Down
12.0%
No
46.6%
30.3%
Steady

diag_1
Text

Distinct	717
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

diag_2
Text

Distinct	749
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

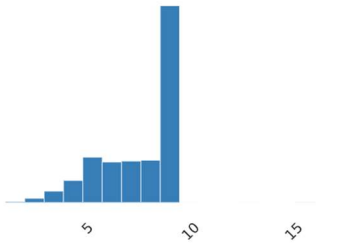
diag_3
Text

Distinct	790
Distinct (%)	0.8%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

number_diagnoses

Real number (ℝ)

Distinct	16	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	16
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	7.4226068	**Memory size**	795.2 KiB



More details

max_glu_serum

Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.2 MiB

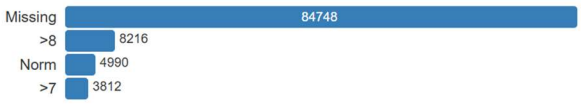
More details

A1Cresult

Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.1 MiB



More details

metformin

Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

More details

metformin
Categorical

IMBALANCE

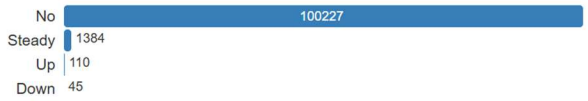
Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

More details

repaglinide
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB



More details

nateglinide
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

chlorpropamide
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

glimepiride
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

acetohexamide
Categorical

IMBALANCE

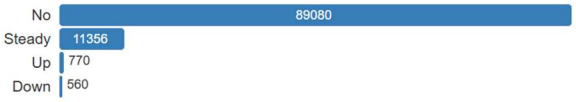
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

glipizide
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB



More details

glyburide
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

More details

tolbutamide
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

pioglitazone
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB

More details

rosiglitazone
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

acarbose
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

miglitol
Categorical

IMBALANCE

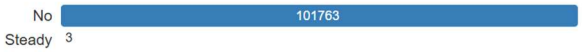
Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

troglitazone
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB



More details

tolazamide
Categorical

IMBALANCE

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

examide
Boolean

CONSTANT

Distinct	1
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	99.5 KiB

More details

citoglipton
Boolean

CONSTANT

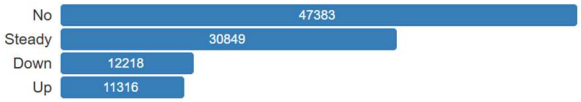
Distinct	1
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	99.5 KiB

More details

insulin
Categorical

HIGH CORRELATION

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.9 MiB



More details

glyburide-metformin
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

glipizide-metformin
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

glimepiride-pioglitazone
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

metformin-rosiglitazone
Categorical

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

metformin-pioglitazone
Categorical

IMBALANCE

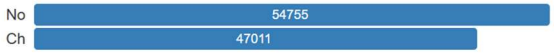
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB

More details

change
Categorical

HIGH CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.7 MiB



More details

diabetesMed
Boolean

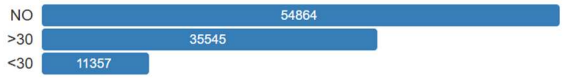
HIGH CORRELATION

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	99.5 KiB

More details

readmitted
Categorical

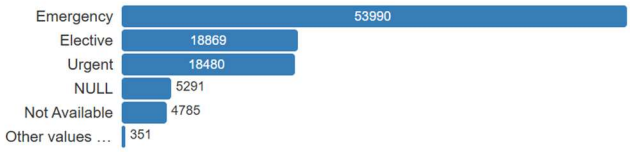
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	5.8 MiB



More details

admission_type
Categorical

Distinct	8
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	6.3 MiB



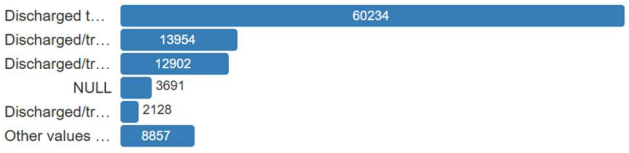
More details

discharge_disposition

Categorical

HIGH CORRELATION IMBALANCE

Distinct	26
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.1 MiB



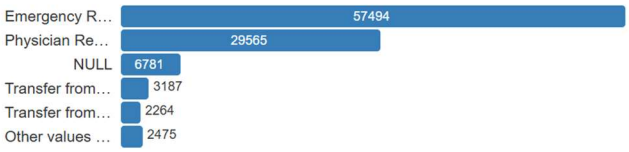
More details

admission_source

Categorical

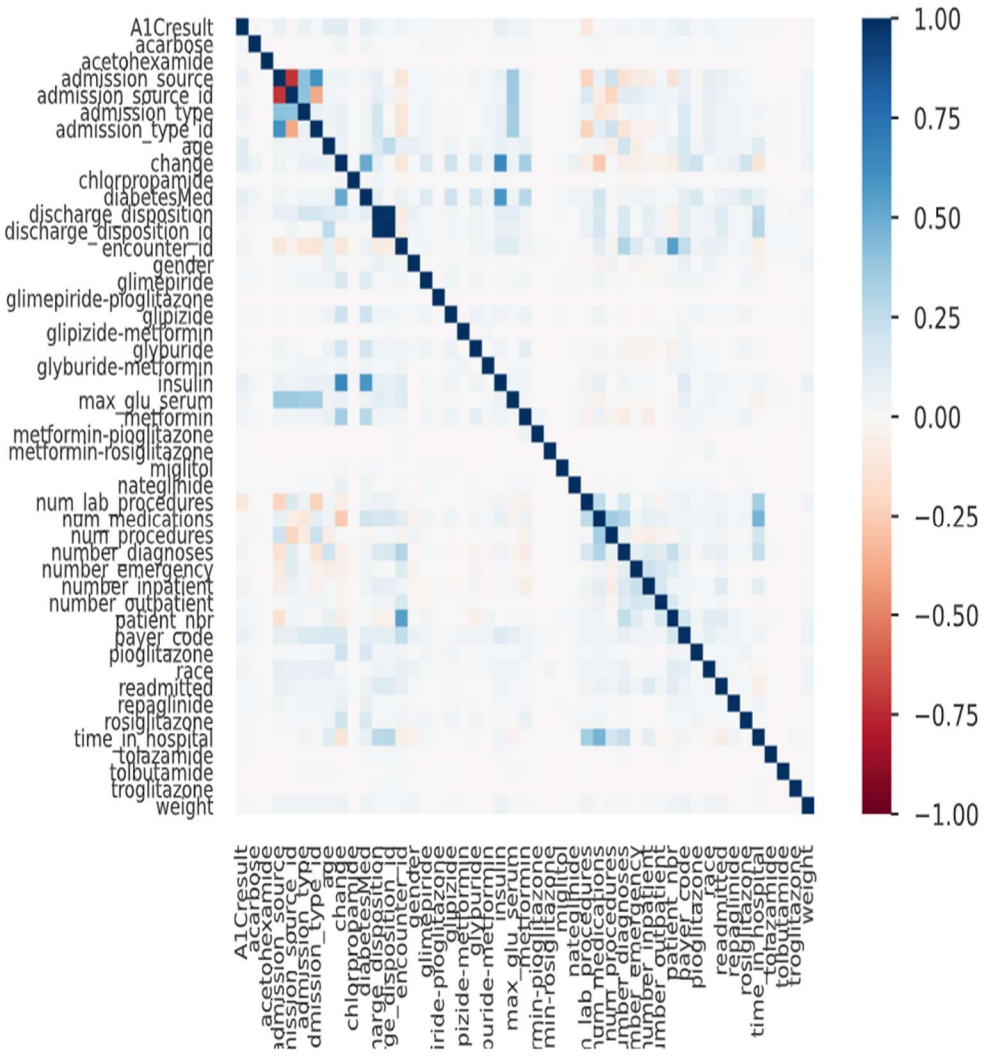
HIGH CORRELATION IMBALANCE

Distinct	17
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	7.1 MiB

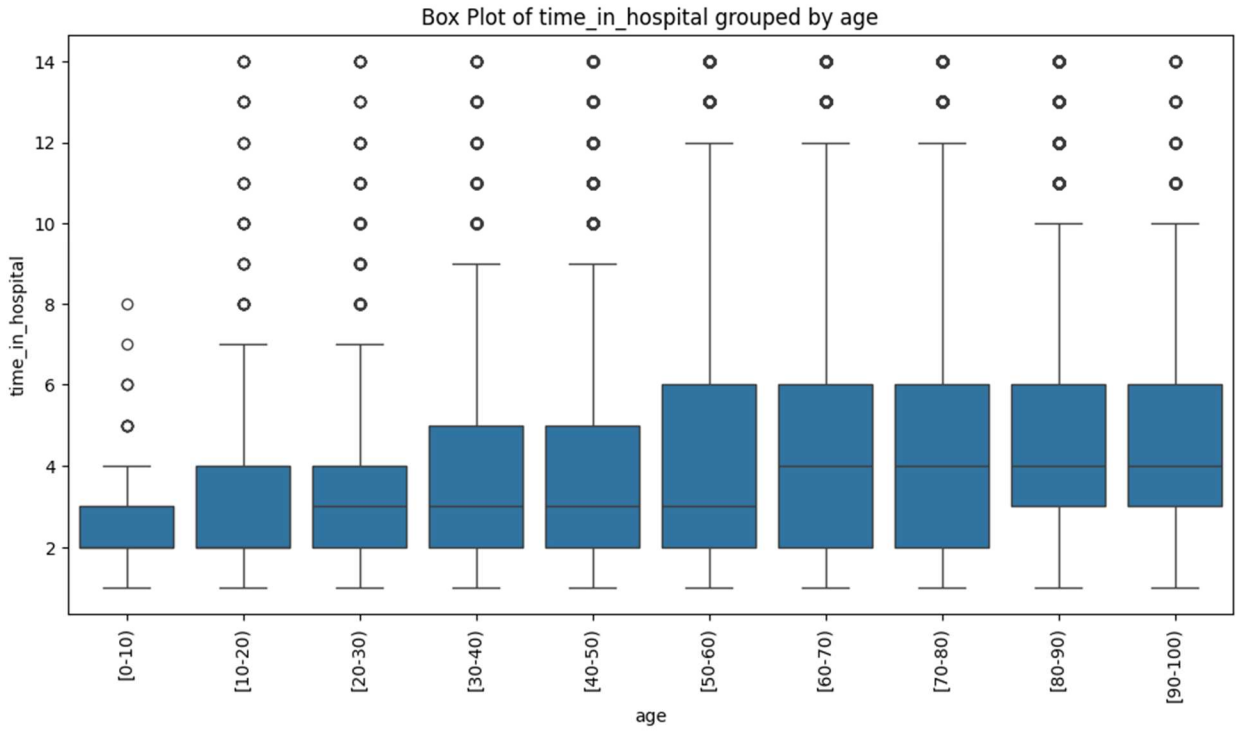


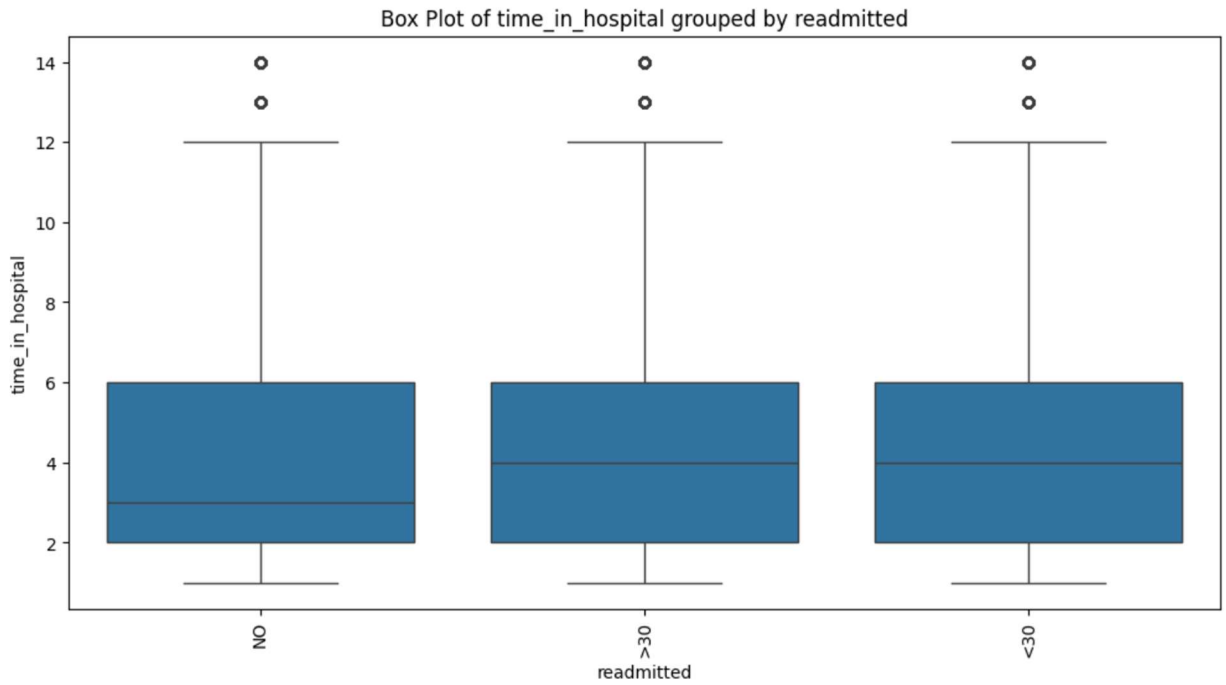
More details

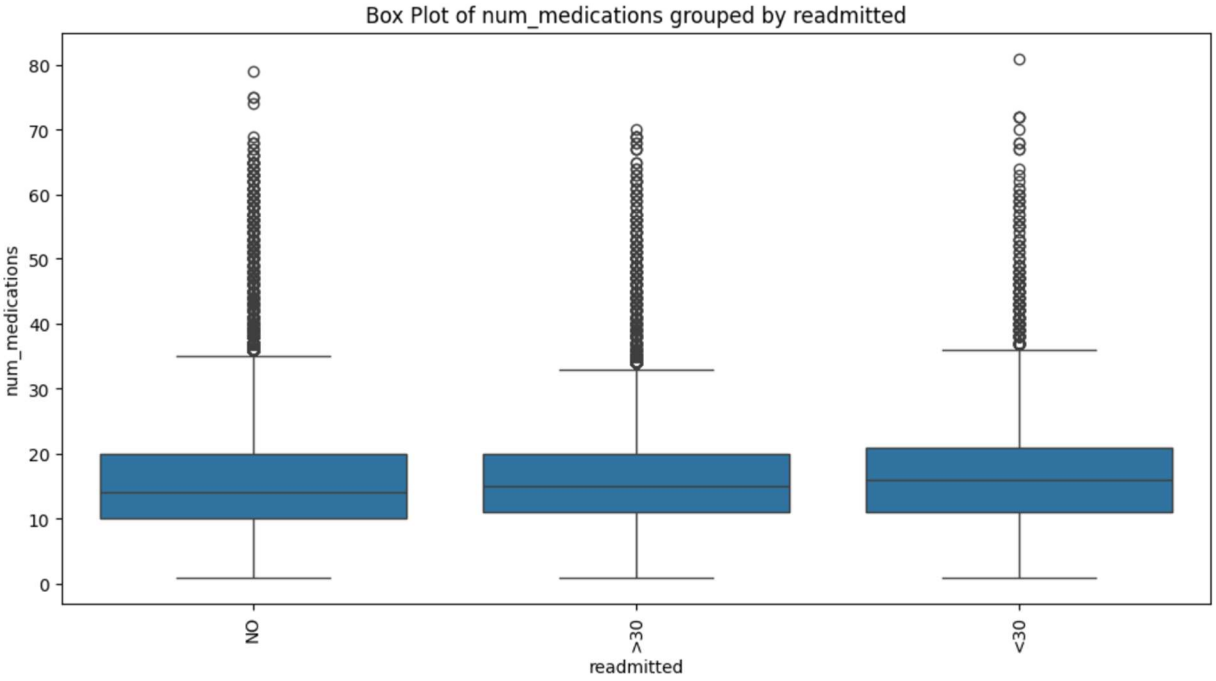
Correlation matrix

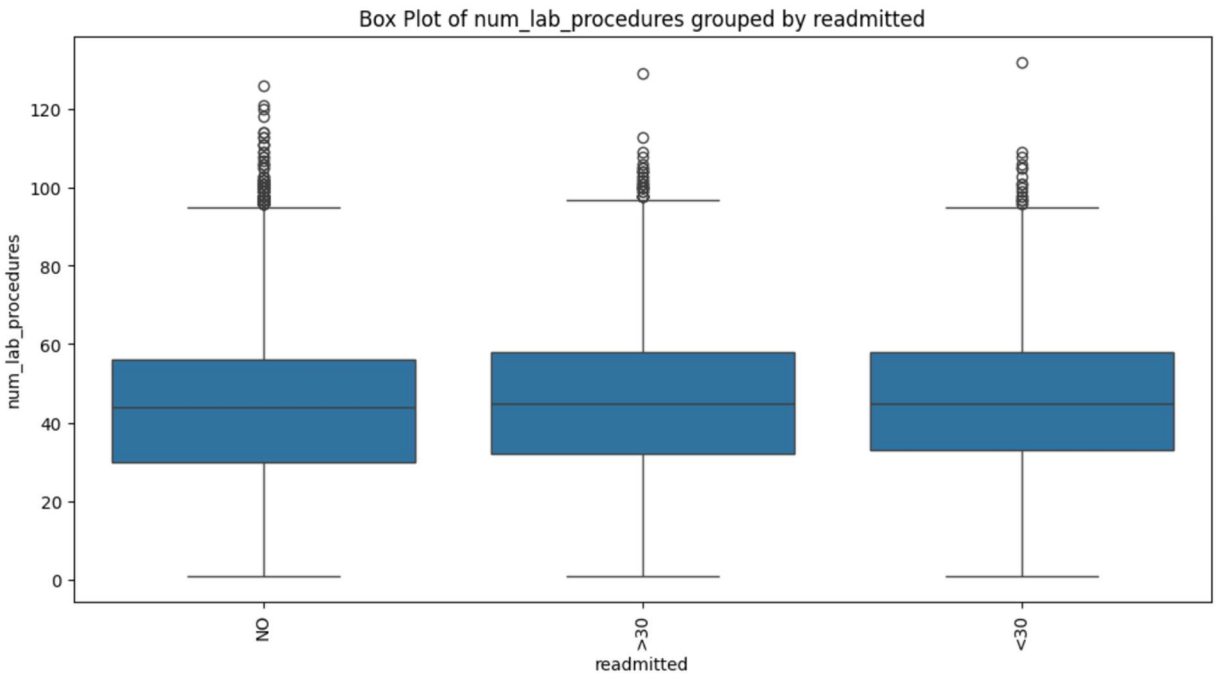


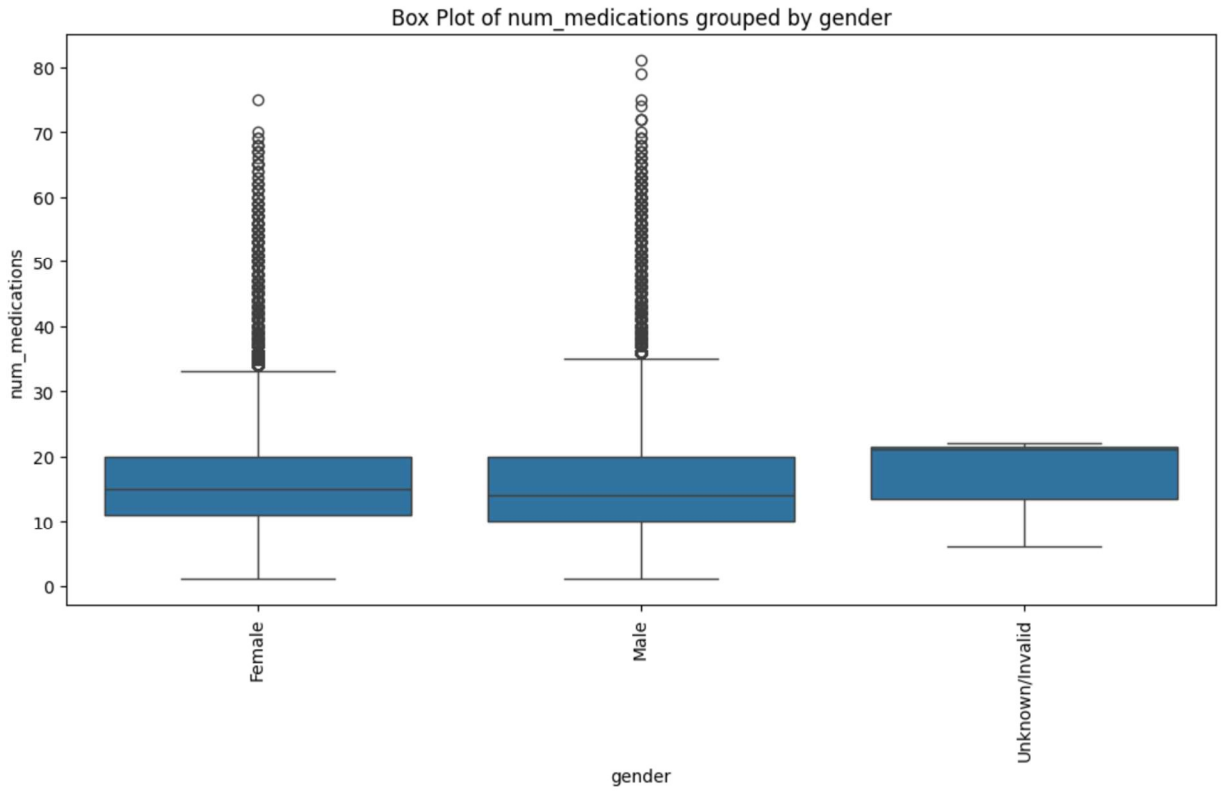
Boxplots & Outliers

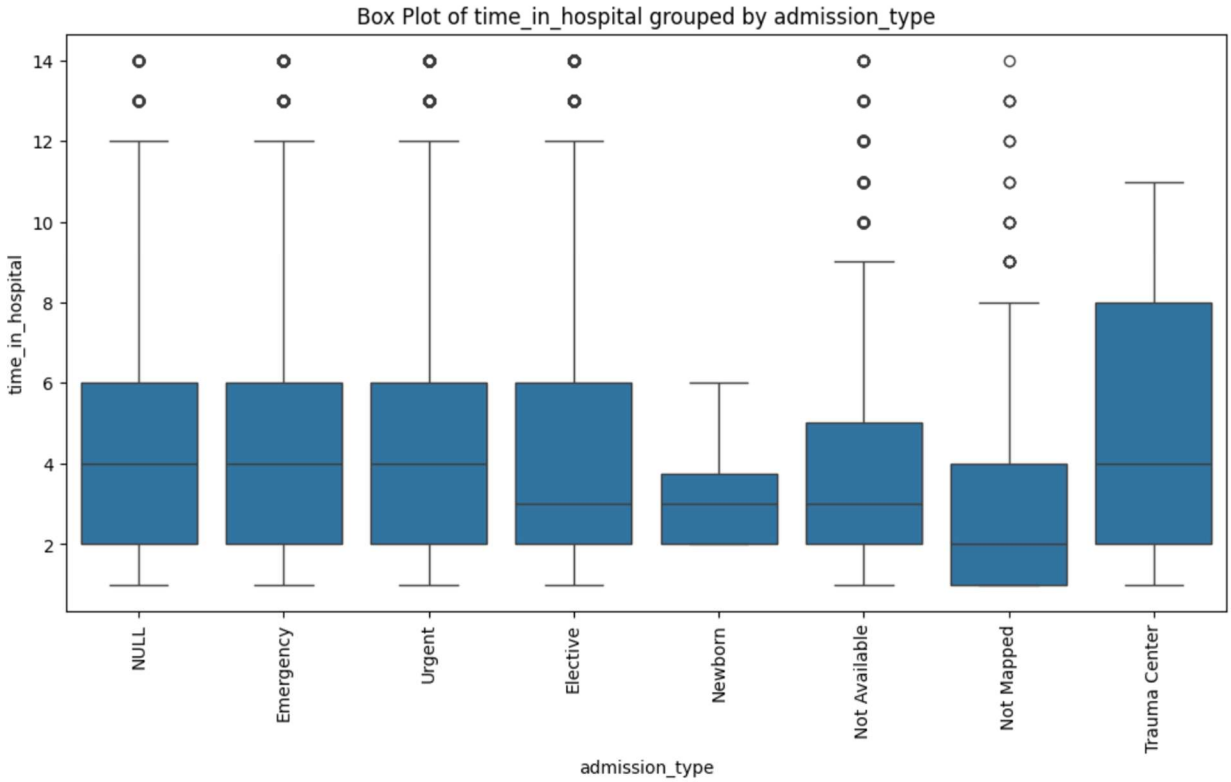


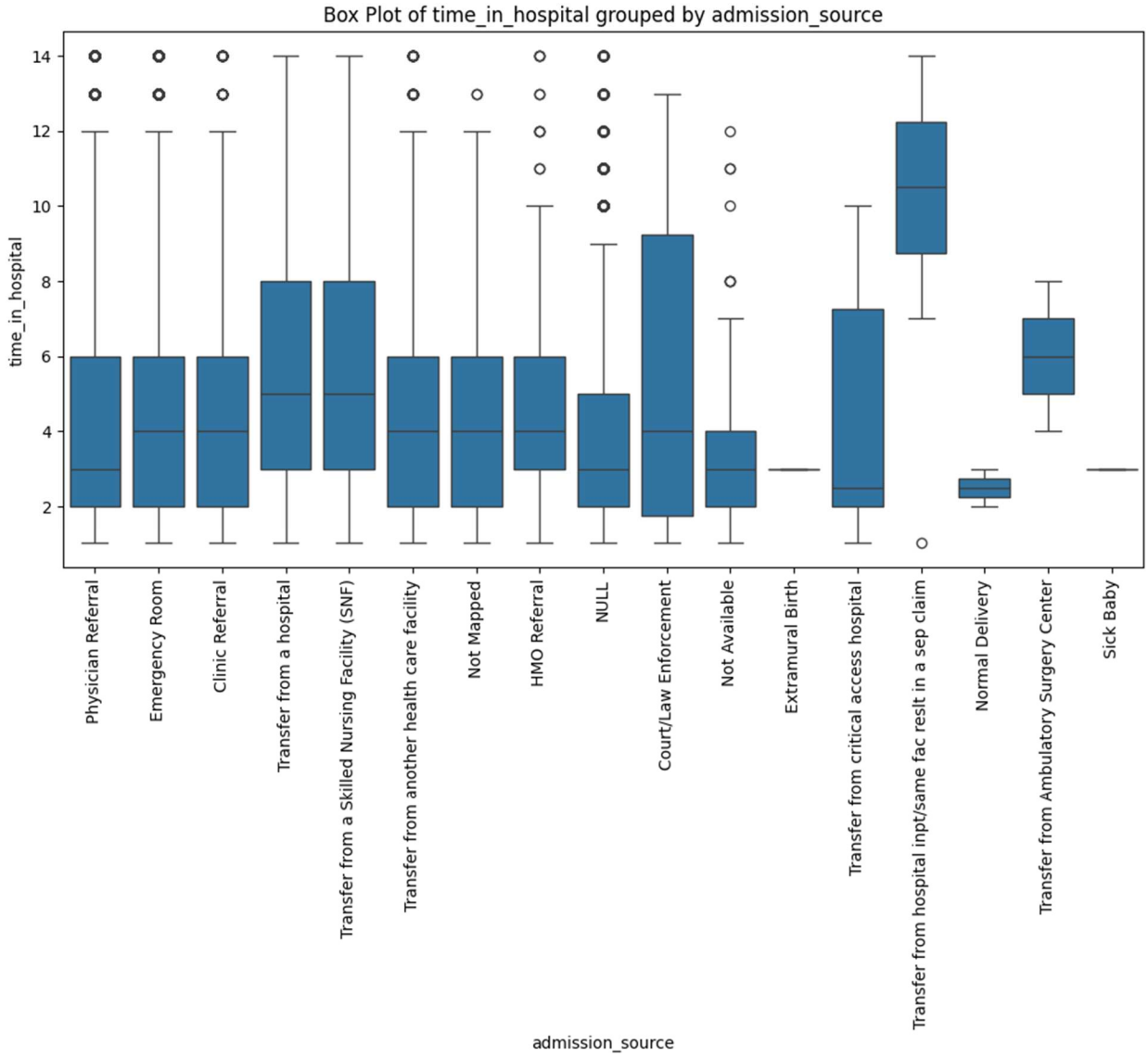

Box Plot of time_in_hospital grouped by readmitted
time_in_hospital
14
12
10
8
6
4
2
NO
>30
<30
readmitted

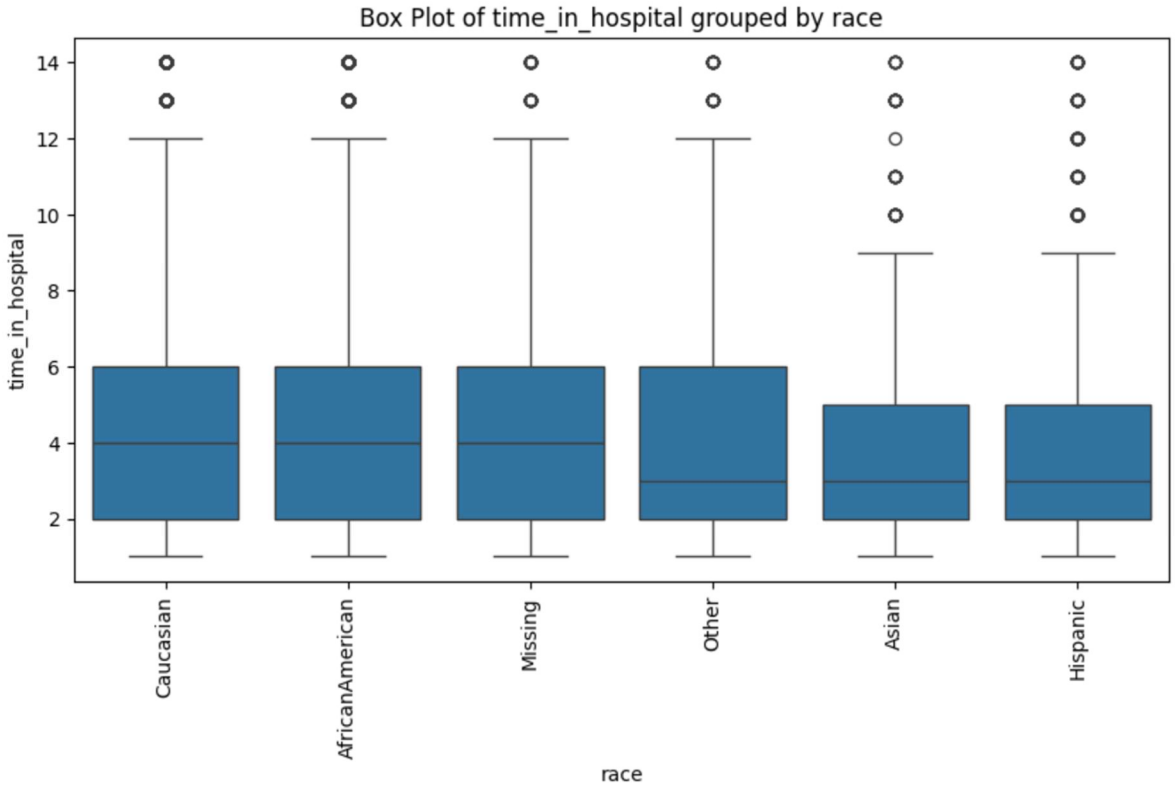

Box Plot of num_medications grouped by readmitted
num_medications
80
70
60
50
40
30
20
10
0
NO
>30
<30
readmitted

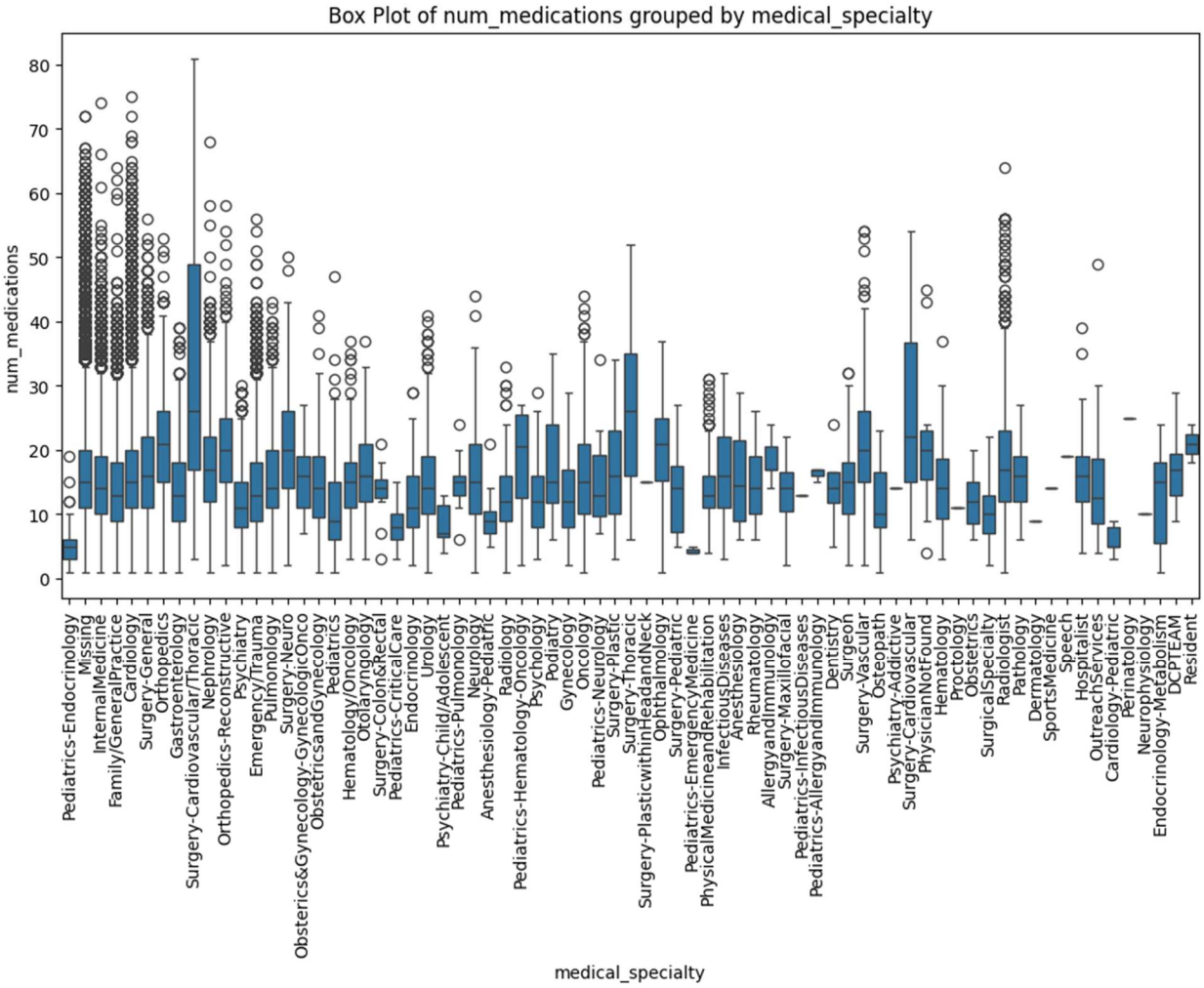

Box Plot of num_lab_procedures grouped by readmitted
num_lab_procedures
0
20
40
60
80
100
120
NO
>30
<30
readmitted

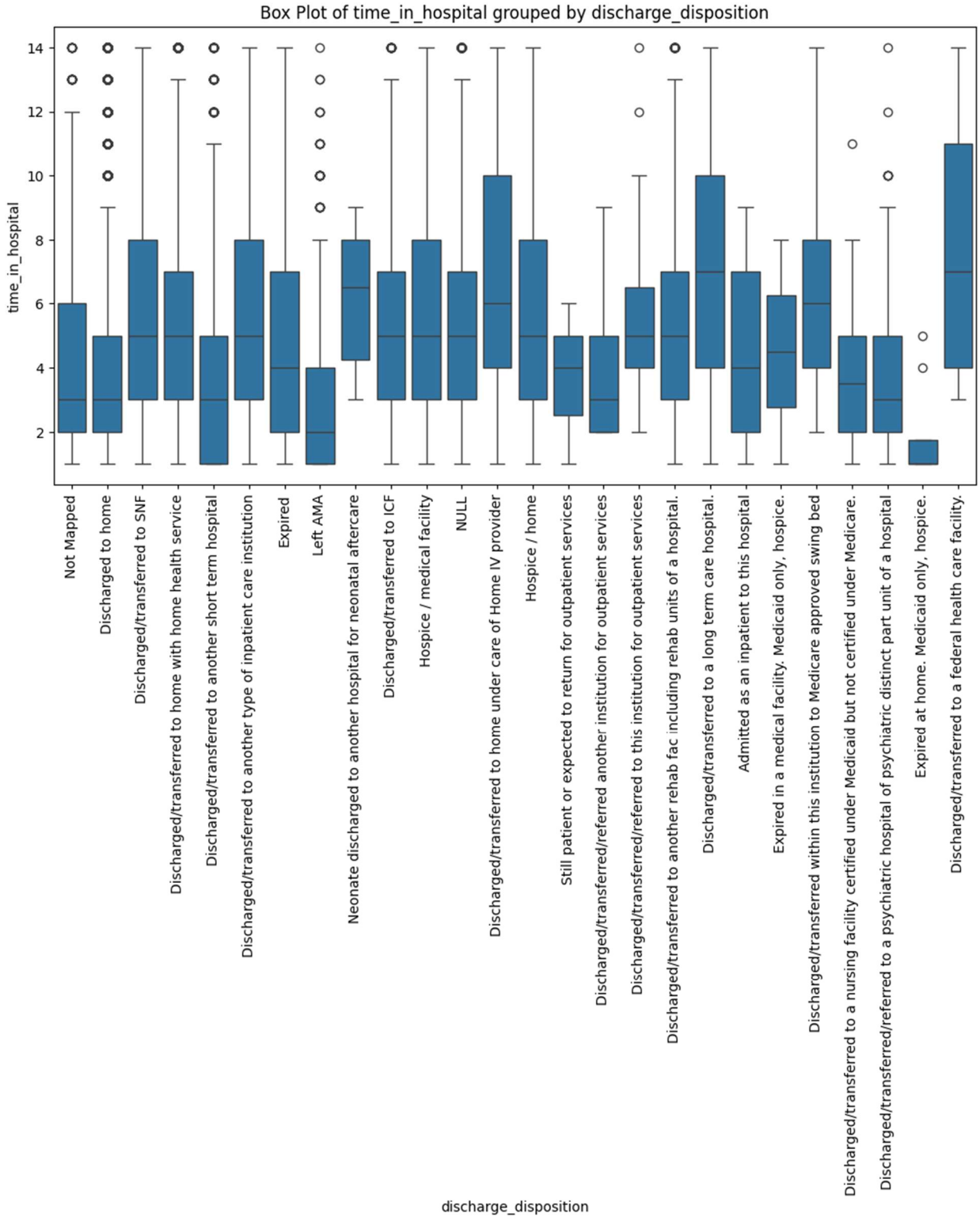

Box Plot of num_medications grouped by gender
num_medications
0
10
20
30
40
50
60
70
80
Female
Male
Unknown/Invalid
gender

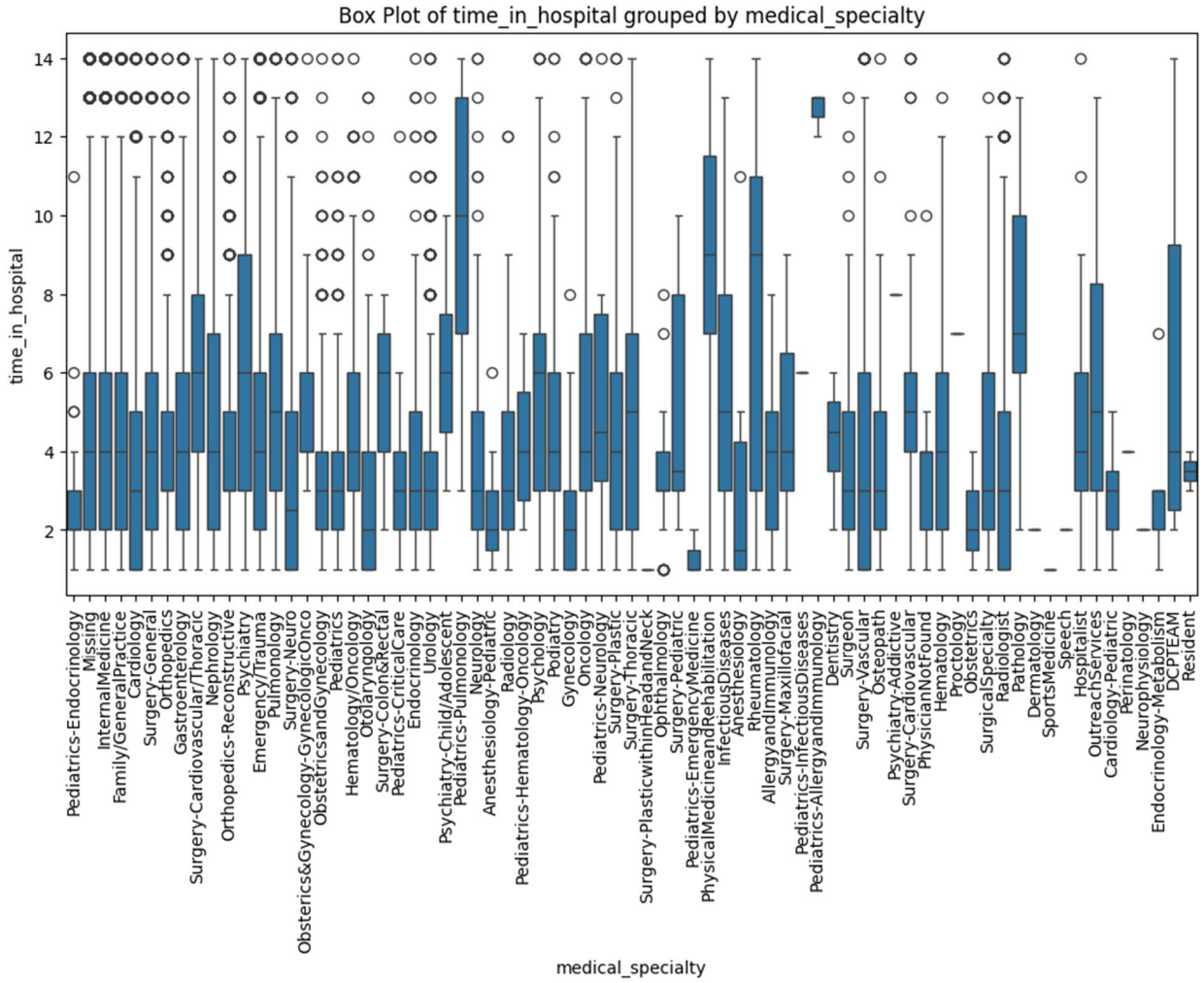

Box Plot of time_in_hospital grouped by admission_type
time_in_hospital
2
4
6
8
10
12
14
NULL
Emergency
Urgent
Elective
Newborn
Not Available
Not Mapped
Trauma Center
admission_type

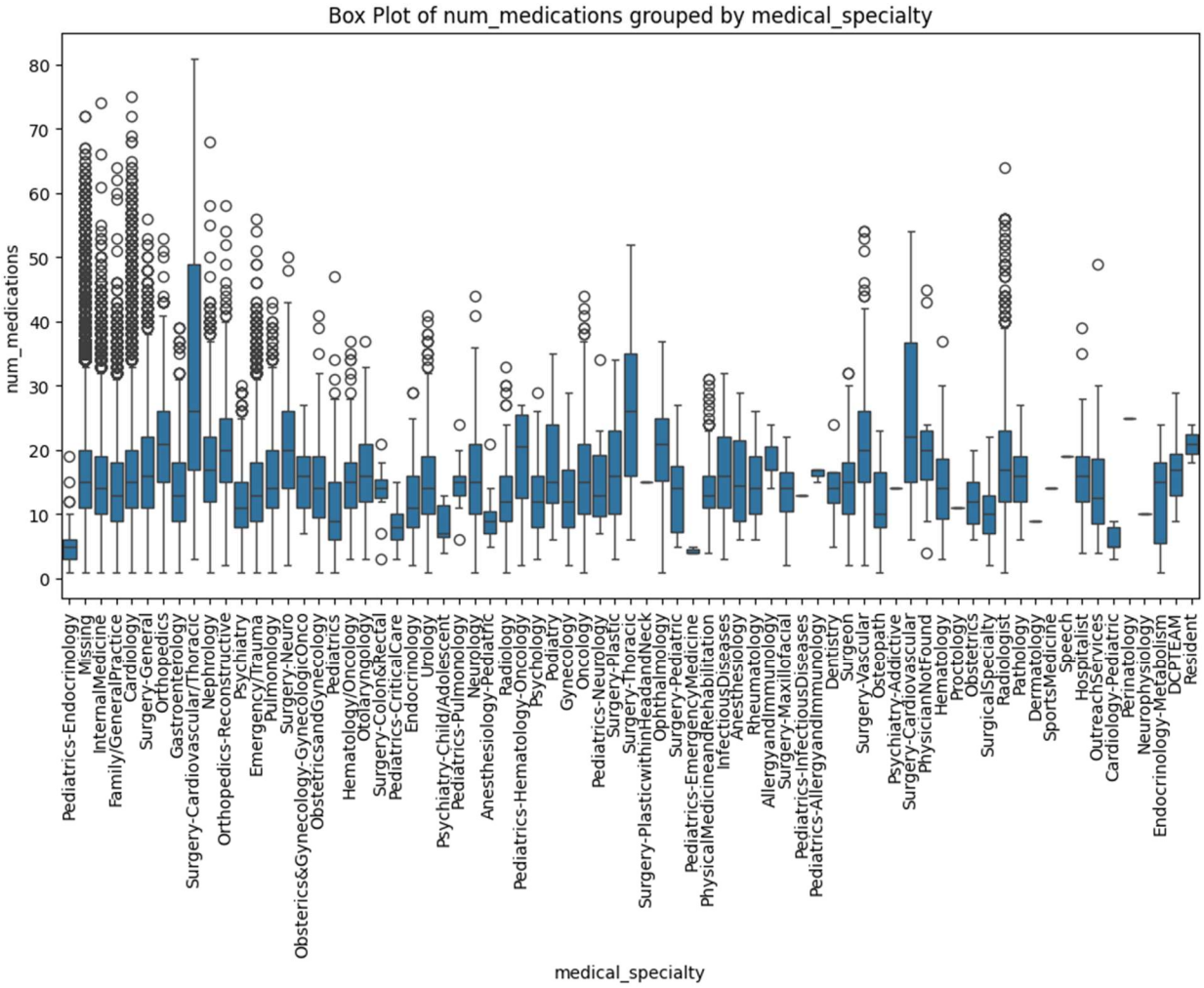

Box Plot of time_in_hospital grouped by admission_source
time_in_hospital
14
12
10
8
6
4
2
Physician Referral
Emergency Room
Clinic Referral
Transfer from a hospital
Transfer from a Skilled Nursing Facility (SNF)
Transfer from another health care facility
Not Mapped
HMO Referral
NULL
Court/Law Enforcement
Not Available
Extramural Birth
Transfer from critical access hospital
Transfer from hospital inpt/same fac reslt in a sep claim
Normal Delivery
Transfer from Ambulatory Surgery Center
Sick Baby
admission_source

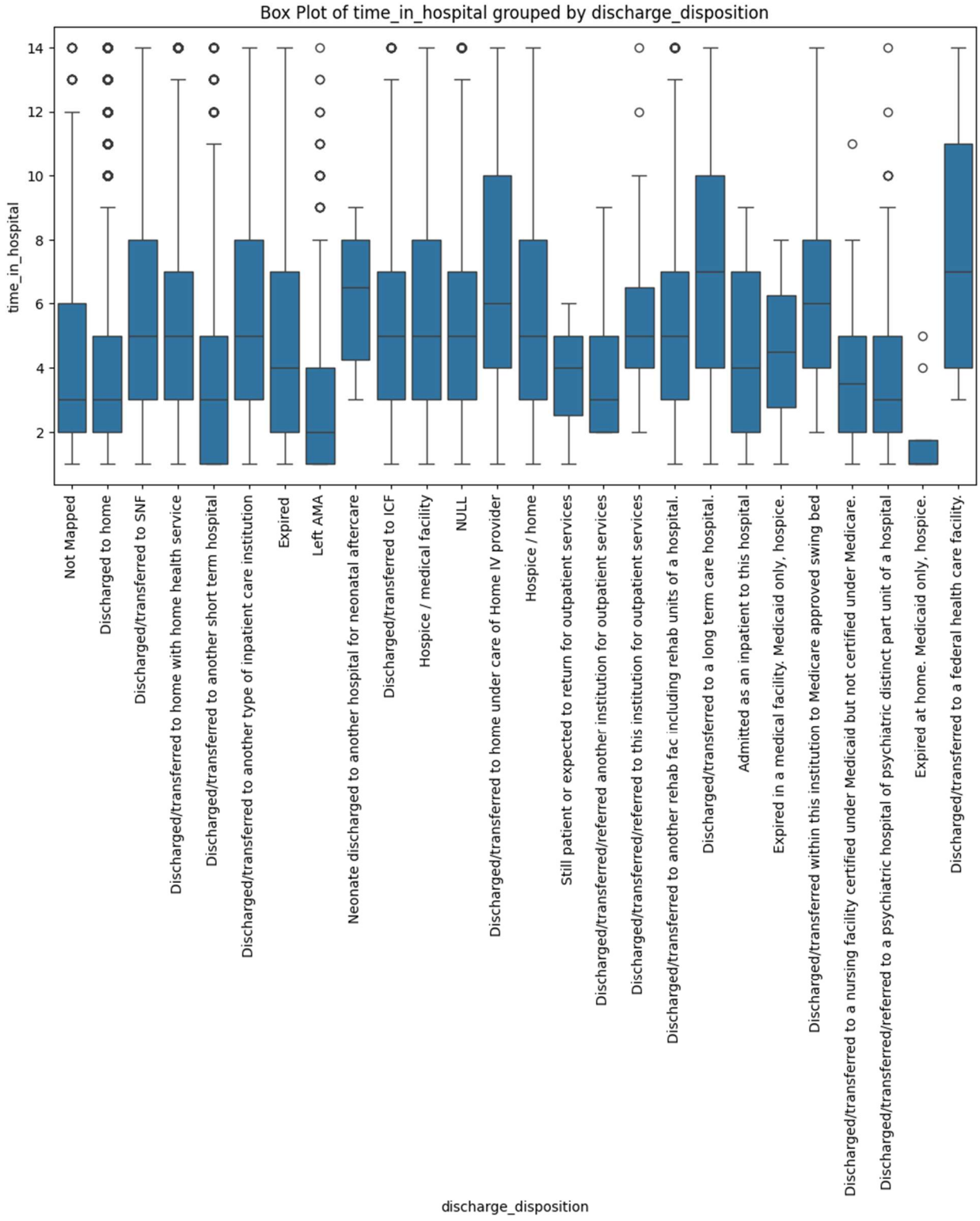

Box Plot of time_in_hospital grouped by race
time_in_hospital
14
12
10
8
6
4
2
Caucasian
AfricanAmerican
Missing
Other
Asian
Hispanic
race

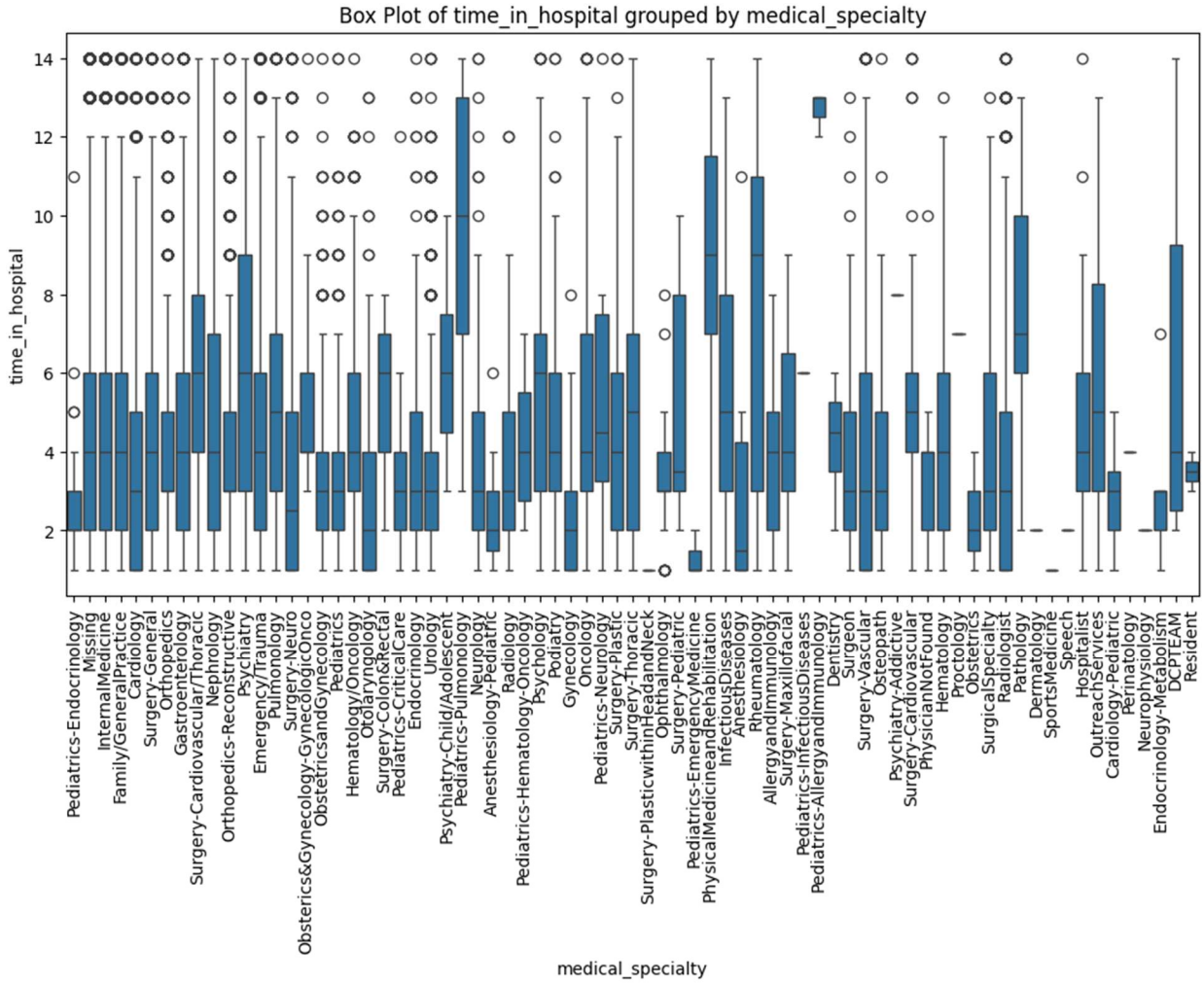

Box Plot of num_medications grouped by medical_specialty
num_medications
medical_specialty
Pediatrics-Endocrinology
Missing
InternalMedicine
Family/GeneralPractice
Cardiology
Surgery-General
Orthopedics
Gastroenterology
Surgery-Cardiovascular/Thoracic
Nephrology
Orthopedics-Reconstructive
Psychiatry
Emergency/Trauma
Pulmonology
Surgery-Neuro
Obstetrics&Gynecology-GynecologicOnco
ObstetricsandGynecology
Pediatrics
Hematology/Oncology
Otolaryngology
Surgery-Colon&Rectal
Pediatrics-CriticalCare
Endocrinology
Urology
Psychiatry-Child/Adolescent
Pediatrics-Pulmonology
Neurology
Anesthesiology-Pediatric
Radiology
Pediatrics-Hematology-Oncology
Psychology
Podiatry
Gynecology
Oncology
Pediatrics-Neurology
Surgery-Plastic
Surgery-Thoracic
Surgery-PlasticwithinHeadandNeck
Ophthalmology
Surgery-Pediatric
Pediatrics-EmergencyMedicine
PhysicalMedicineandRehabilitation
InfectiousDiseases
Anesthesiology
Rheumatology
AllergyandImmunology
Surgery-Maxillofacial
Pediatrics-InfectiousDiseases
Pediatrics-AllergyandImmunology
Dentistry
Surgeon
Surgery-Vascular
Osteopath
Psychiatry-Addictive
Surgery-Cardiovascular
PhysicianNotFound
Hematology
Proctology
Obstetrics
SurgicalSpecialty
Radiologist
Pathology
Dermatology
SportsMedicine
Speech
Hospitalist
OutreachServices
Cardiology-Pediatric
Perinatology
Neurophysiology
Endocrinology-Metabolism
DCPTEAM
Resident


Box Plot of time_in_hospital grouped by discharge_disposition
time_in_hospital
14
12
10
8
6
4
2
Not Mapped
Discharged to home
Discharged/transferred to SNF
Discharged/transferred to home with home health service
Discharged/transferred to another short term hospital
Discharged/transferred to another type of inpatient care institution
Expired
Left AMA
Neonate discharged to another hospital for neonatal aftercare
Discharged/transferred to ICF
Hospice / medical facility
NULL
Discharged/transferred to home under care of Home IV provider
Hospice / home
Still patient or expected to return for outpatient services
Discharged/transferred/referred another institution for outpatient services
Discharged/transferred/referred to this institution for outpatient services
Discharged/transferred to another rehab fac including rehab units of a hospital.
Discharged/transferred to a long term care hospital.
Admitted as an inpatient to this hospital
Expired in a medical facility. Medicaid only, hospice.
Discharged/transferred within this institution to Medicare approved swing bed
Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
Expired at home. Medicaid only, hospice.
Discharged/transferred to a federal health care facility.
discharge_disposition


Box Plot of time_in_hospital grouped by medical_specialty
time_in_hospital
medical_specialty
Pediatrics-Endocrinology
Missing
InternalMedicine
Family/GeneralPractice
Cardiology
Surgery-General
Orthopedics
Gastroenterology
Surgery-Cardiovascular/Thoracic
Nephrology
Orthopedics-Reconstructive
Psychiatry
Emergency/Trauma
Pulmonology
Surgery-Neuro
Obstetrics&Gynecology-GynecologicOnco
ObstetricsandGynecology
Pediatrics
Hematology/Oncology
Otolaryngology
Surgery-Colon&Rectal
Pediatrics-CriticalCare
Endocrinology
Urology
Psychiatry-Child/Adolescent
Pediatrics-Pulmonology
Neurology
Anesthesiology-Pediatric
Radiology
Pediatrics-Hematology-Oncology
Psychology
Podiatry
Gynecology
Oncology
Pediatrics-Neurology
Surgery-Plastic
Surgery-Thoracic
Surgery-PlasticwithinHeadandNeck
Ophthalmology
Surgery-Pediatric
Pediatrics-EmergencyMedicine
PhysicalMedicineandRehabilitation
InfectiousDiseases
Anesthesiology
Rheumatology
AllergyandImmunology
Surgery-Maxillofacial
Pediatrics-InfectiousDiseases
Pediatrics-AllergyandImmunology
Dentistry
Surgeon
Surgery-Vascular
Osteopath
Psychiatry-Addictive
Surgery-Cardiovascular
PhysicianNotFound
Hematology
Proctology
Obstetrics
SurgicalSpecialty
Radiologist
Pathology
Dermatology
SportsMedicine
Speech
Hospitalist
OutreachServices
Cardiology-Pediatric
Perinatology
Neurophysiology
Endocrinology-Metabolism
DCPTEAM
Resident


Box Plot of num_medications grouped by medical_specialty
num_medications
medical_specialty
Pediatrics-Endocrinology
Missing
InternalMedicine
Family/GeneralPractice
Cardiology
Surgery-General
Orthopedics
Gastroenterology
Surgery-Cardiovascular/Thoracic
Nephrology
Orthopedics-Reconstructive
Psychiatry
Emergency/Trauma
Pulmonology
Surgery-Neuro
Obstetrics&Gynecology-GynecologicOnco
ObstetricsandGynecology
Pediatrics
Hematology/Oncology
Otolaryngology
Surgery-Colon&Rectal
Pediatrics-CriticalCare
Endocrinology
Urology
Psychiatry-Child/Adolescent
Pediatrics-Pulmonology
Neurology
Anesthesiology-Pediatric
Radiology
Pediatrics-Hematology-Oncology
Psychology
Podiatry
Gynecology
Oncology
Pediatrics-Neurology
Surgery-Plastic
Surgery-Thoracic
Surgery-PlasticwithinHeadandNeck
Ophthalmology
Surgery-Pediatric
Pediatrics-EmergencyMedicine
PhysicalMedicineandRehabilitation
InfectiousDiseases
Anesthesiology
Rheumatology
AllergyandImmunology
Surgery-Maxillofacial
Pediatrics-InfectiousDiseases
Pediatrics-AllergyandImmunology
Dentistry
Surgeon
Surgery-Vascular
Osteopath
Psychiatry-Addictive
Surgery-Cardiovascular
PhysicianNotFound
Hematology
Proctology
Obstetrics
SurgicalSpecialty
Radiologist
Pathology
Dermatology
SportsMedicine
Speech
Hospitalist
OutreachServices
Cardiology-Pediatric
Perinatology
Neurophysiology
Endocrinology-Metabolism
DCPTEAM
Resident


Box Plot of time_in_hospital grouped by discharge_disposition
time_in_hospital
14
12
10
8
6
4
2
Not Mapped
Discharged to home
Discharged/transferred to SNF
Discharged/transferred to home with home health service
Discharged/transferred to another short term hospital
Discharged/transferred to another type of inpatient care institution
Expired
Left AMA
Neonate discharged to another hospital for neonatal aftercare
Discharged/transferred to ICF
Hospice / medical facility
NULL
Discharged/transferred to home under care of Home IV provider
Hospice / home
Still patient or expected to return for outpatient services
Discharged/transferred/referred another institution for outpatient services
Discharged/transferred/referred to this institution for outpatient services
Discharged/transferred to another rehab fac including rehab units of a hospital.
Discharged/transferred to a long term care hospital.
Admitted as an inpatient to this hospital
Expired in a medical facility. Medicaid only, hospice.
Discharged/transferred within this institution to Medicare approved swing bed
Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
Expired at home. Medicaid only, hospice.
Discharged/transferred to a federal health care facility.
discharge_disposition


Box Plot of time_in_hospital grouped by medical_specialty
time_in_hospital
medical_specialty
Pediatrics-Endocrinology
Missing
InternalMedicine
Family/GeneralPractice
Cardiology
Surgery-General
Orthopedics
Gastroenterology
Surgery-Cardiovascular/Thoracic
Nephrology
Orthopedics-Reconstructive
Psychiatry
Emergency/Trauma
Pulmonology
Surgery-Neuro
Obstetrics&Gynecology-GynecologicOnco
ObstetricsandGynecology
Pediatrics
Hematology/Oncology
Otolaryngology
Surgery-Colon&Rectal
Pediatrics-CriticalCare
Endocrinology
Urology
Psychiatry-Child/Adolescent
Pediatrics-Pulmonology
Neurology
Anesthesiology-Pediatric
Radiology
Pediatrics-Hematology-Oncology
Psychology
Podiatry
Gynecology
Oncology
Pediatrics-Neurology
Surgery-Plastic
Surgery-Thoracic
Surgery-PlasticwithinHeadandNeck
Ophthalmology
Surgery-Pediatric
Pediatrics-EmergencyMedicine
PhysicalMedicineandRehabilitation
InfectiousDiseases
Anesthesiology
Rheumatology
AllergyandImmunology
Surgery-Maxillofacial
Pediatrics-InfectiousDiseases
Pediatrics-AllergyandImmunology
Dentistry
Surgeon
Surgery-Vascular
Osteopath
Psychiatry-Addictive
Surgery-Cardiovascular
PhysicianNotFound
Hematology
Proctology
Obstetrics
SurgicalSpecialty
Radiologist
Pathology
Dermatology
SportsMedicine
Speech
Hospitalist
OutreachServices
Cardiology-Pediatric
Perinatology
Neurophysiology
Endocrinology-Metabolism
DCPTEAM
Resident

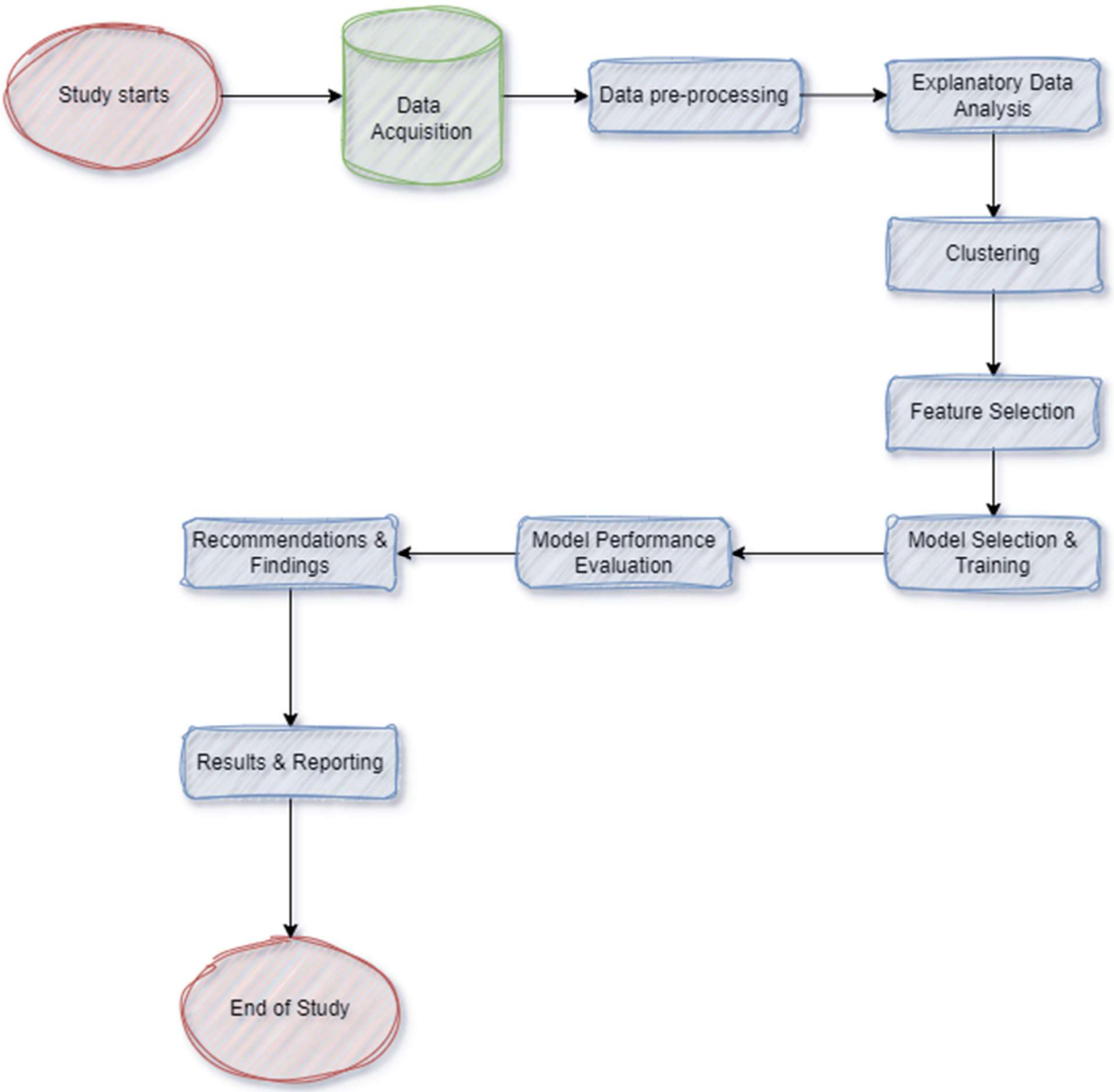
```python
#Identifying outliers in numeric columns
#Function to calculate and print outliers
def outliers(df, numeric_columns):
    outliers_count = {}

    for col in numeric_columns:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        outliers = df[(df[col] < lower_bound) | (df[col] > upper_bound)]
        outliers_count[col] = outliers[col]

        print(f"{col}:")
        print(f"  Q1: {Q1}")
        print(f"  Q3: {Q3}")
        print(f"  IQR: {IQR}")
        print(f"  Lower Bound: {lower_bound}")
        print(f"  Upper Bound: {upper_bound}")
        print(f"  Outliers: {len(outliers)}")
        print()

    return outliers_count

#print outliers
outliers_count = outliers(df, numeric_columns)
```

```
encounter_id:
  Q1: 84961194.0
  Q3: 230270887.5
  IQR: 145309693.5
  Lower Bound: -133003346.25
  Upper Bound: 448235427.75
  Outliers: 0

patient_nbr:
  Q1: 23413221.0
  Q3: 87545949.75
  IQR: 64132728.75
  Lower Bound: -72785872.125
  Upper Bound: 183745042.875
  Outliers: 247

admission_type_id:
  Q1: 1.0
  Q3: 3.0
  IQR: 2.0
  Lower Bound: -2.0
  Upper Bound: 6.0
  Outliers: 341

discharge_disposition_id:
  Q1: 1.0
  Q3: 4.0
  IQR: 3.0
  Lower Bound: -3.5
  Upper Bound: 8.5
  Outliers: 9818

admission_source_id:
  Q1: 1.0
  Q3: 7.0
  IQR: 6.0
  Lower Bound: -8.0
  Upper Bound: 16.0
  Outliers: 6956

time_in_hospital:
  Q1: 2.0
  Q3: 6.0
  IQR: 4.0
  Lower Bound: -4.0
  Upper Bound: 12.0
  Outliers: 2252

num_lab_procedures:
  Q1: 31.0
  Q3: 57.0
  IQR: 26.0
  Lower Bound: -8.0
  Upper Bound: 96.0
  Outliers: 143

num_procedures:
  Q1: 0.0
  Q3: 2.0
  IQR: 2.0
  Lower Bound: -3.0
  Upper Bound: 5.0
  Outliers: 4954

num_medications:
  Q1: 10.0
  Q3: 20.0
  IQR: 10.0
  Lower Bound: -5.0
  Upper Bound: 35.0
  Outliers: 2557

number_outpatient:
  Q1: 0.0
  Q3: 0.0
  IQR: 0.0
  Lower Bound: 0.0
  Upper Bound: 0.0
  Outliers: 16739

number_emergency:
  Q1: 0.0
  Q3: 0.0
  IQR: 0.0
  Lower Bound: 0.0
  Upper Bound: 0.0
  Outliers: 11383

number_inpatient:
  Q1: 0.0
  Q3: 1.0
  IQR: 1.0
  Lower Bound: -1.5
  Upper Bound: 2.5
  Outliers: 7049

number_diagnoses:
  Q1: 6.0
  Q3: 9.0
  IQR: 3.0
  Lower Bound: 1.5
  Upper Bound: 13.5
  Outliers: 281
```

Link to GitHub Repository

https://github.com/NehalTMU/TMU-Capstone-Project

Tentative Methodology Flowchart



References

Dataset Source

https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International, 2014*, 781670–781611. https://doi.org/10.1155/2014/781670

URL: https://www.hindawi.com/journals/bmri/2014/781670/

Kumar Sah, D., & Khanal, M. (2023, November). Implementation of big data analytics on diabetes 130-US hospitals for the year 1999-2008 for predicting patient readmission. Preprint. https://doi.org/10.13140/RG.2.2.18564.30081

URL: https://www.researchgate.net/publication/375690075_Implementation_of_Big_Data_Analytics_on_Diabetes_130-US_Hospitals_for_year_1999-2008_for_predicting_patient_readmission

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers. *BMC Medical Informatics and Decision Making, 21*(Suppl 2), 57. https://doi.org/10.1186/s12911-021-01423-y

URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01423-y

Tavakolian, A., Rezaee, A., Hajati, F., & Uddin, S. (2023). Hospital readmission and length-of-stay prediction using an optimized hybrid deep model. *Future Internet, 15*(9), Article 304. https://doi.org/10.3390/fi15090304

URL: https://www.mdpi.com/1999-5903/15/9/304

Davis, S., Zhang, J., Lee, I., Rezaei, M., Greiner, R., McAlister, F. A., & Padwal, R. (2022). Effective hospital readmission prediction models using machine-learned features. *BMC Health Services Research*. https://doi.org/10.1186/s12913-022-08748-y

URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-022-08748-y

Michailidis, P., Dimitriadou, A., Papadimitriou, T., & Gogas, P. (2022). Forecasting hospital readmissions with machine learning. *Healthcare, 10*, 981. https://doi.org/10.3390/healthcare10060981

URL: https://www.mdpi.com/2227-9032/10/9/981

Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021). Application of machine learning in predicting hospital readmissions: A scoping review of the literature. *BMC Medical Research Methodology*. https://doi.org/10.1186/s12874-021-01284-z

URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12874-021-01284-z

Centers for Medicare & Medicaid Services. (n.d.). CMS hospital readmissions reduction program (HRRP). Retrieved from https://www.cms.gov

American Diabetes Association. (2022). Statistics about diabetes. Retrieved from https://www.diabetes.org

Zhu, Z., Hinton, N., Zhao, Y., & He, Y. (2021). 30-Day readmission risk for diabetic patients with COVID-19. *Diabetes Care, 44*(7), 1530-1533. https://doi.org/10.2337/dc21-0104

URL: https://care.diabetesjournals.org/content/44/7/1530

Hirsch, J. S., Ng, J. H., Ross, D. W., Sharma, P., Shah, H. H., Barnett, R. L., ... & Northwell COVID-19 Research Consortium. (2020). Acute kidney injury in patients hospitalized with COVID-19. *Kidney International, 98*(1), 209-218. https://doi.org/10.1016/j.kint.2020.05.006

URL: https://www.kidney-international.org/article/S0085-2538(20)30612-9/fulltext

Dandachi, D., Geiger, G., Montgomery, M. W., Kharfen, M., & Rodriguez-Barradas, M. C. (2021). Characteristics, outcomes, and mortality among persons with diabetes hospitalized with COVID-19: Experience from a large New York City health system. *Journal of Diabetes and Its Complications, 35*(10), 107966. https://doi.org/10.1016/j.jdiacomp.2021.107966

URL: https://www.journalofdiabetesanditscomplications.com/article/S1056-8727(21)00140-2/fulltext