Diabetes Readmittance: Final Report

Problem Statement

Hospitals and physicians want to know the likelihood of a diabetic patient being readmitted after receiving treatment at the hospital. Understanding what the odds are of a readmission helps doctors provide better treatments for patients, given the medical background and previous visit information. All hospitals aim to have low readmission rates since it reflects on the quality of medical care received at the facility and for patients, frequent readmissions indicate a possible instability in their medications or disease management, which is dangerous for a condition like diabetes. Therefore, the purpose of this project is to analyze anonymous patient hospital visits and provide a model that predicts the likelihood of a patient being readmitted to the hospital given their current situation.

Data Description and Wrangling

The data comes from the UCI Machine Learning Repository and represents 10 years of hospital visits by anonymous diabetic patients. Vital information about the patient's demographics and hospital visit is included in the data, including any diagnoses received and medication intake. For a full list of all variables, please see the table at the end of this section.

Data Wrangling

The purpose of cleaning the data is to present a viable dataset that can provide reliable information about diabetic patients and their hospital admission patterns. Looking at the dataset as a whole, there were no duplicates, meaning that each visit was unique and the encounter ID variable could be removed. The weight and payer code variables were missing for around half of all patients with no real method of finding their true values, and they were dropped. Taking a closer look at the data, it became clear that some patients visited the hospital multiple times and were overrepresented in this dataset. Multiple visits were limited to one per patient by retaining their first visit.

NaN Values

Five variables contained missing values: race, medical specialty, diagnosis 1, diagnosis 2, and diagnosis 3. Patients without a listed race were dropped from the dataset. Since the size of the set is 71,518 rows and 50 columns and those missing a race numbered 1,948, they could be dropped without contaminating results. Those patients could not be added to the Other group either because we do not know the true value. Rows without a medical specialty had them filled with an Unknown string since a significant portion of patients, over 34,000, did not have a listed value.

The three diagnosis columns represent only the first three conditions that a patient is diagnosed with during their visit with the number of diagnoses column stating the total number. Each of the three diagnoses columns contain an ICD-9 code that correlates to a specific condition. If the patient is only diagnosed with one condition, then the second and third diagnosis columns will be empty and the number of diagnoses column will have a 1 listed. To approach this problem, I looked at the primary diagnosis column and isolated the ones with NaN. They had codes listed for secondary and tertiary diagnoses and the one row missing codes for all three columns had five total diagnoses, so those rows were dropped from the dataset on account of missing vital data. I repeated this process for secondary and tertiary diagnosis columns, checking the number of diagnoses columns to determine if a row should be kept or dropped. The remaining empty rows received a None string since some patients only had one or two diagnoses total.

Grouping Diagnosis Codes

Each string in the three diagnosis columns were replaced with a category that represented the code they had. The categories are suggestions from the providers of the dataset.

Treating Admission Type ID, Discharge Disposition, and Admission Source

Each of the three variables contained an ID that correlated with a categorical value. To help understand the patient's visit, for the admission type column, each ID was replaced with the corresponding text condition. Discharge disposition and admission source contained descriptions instead of categories so they retained their IDs. In each of the three columns, there are unknowns described in different ways, i.e. NULL, not mapped, not available. Since they all mean that there is no information on the patient for that variable, I grouped them (the unknowns) together by replacing the IDs with a 0 to show that the information is missing. One thing to note is that discharge disposition tells us what happened to the patient after their hospital visit. The patients who expired and were transferred to hospice were removed from the dataset since will not (unlikely) be readmitted due to death (expired) and being terminally ill (hospice).

Medications

Four medications, 'metformin-rosiglitazone', 'glimepiride-pioglitazone', 'sitagliptin', 'examide', were removed from the dataset since they contained a single value, No, which means that they are not being used by patients.

Final Remarks

The cleaned dataset was saved as a new file for data visualization and analysis.

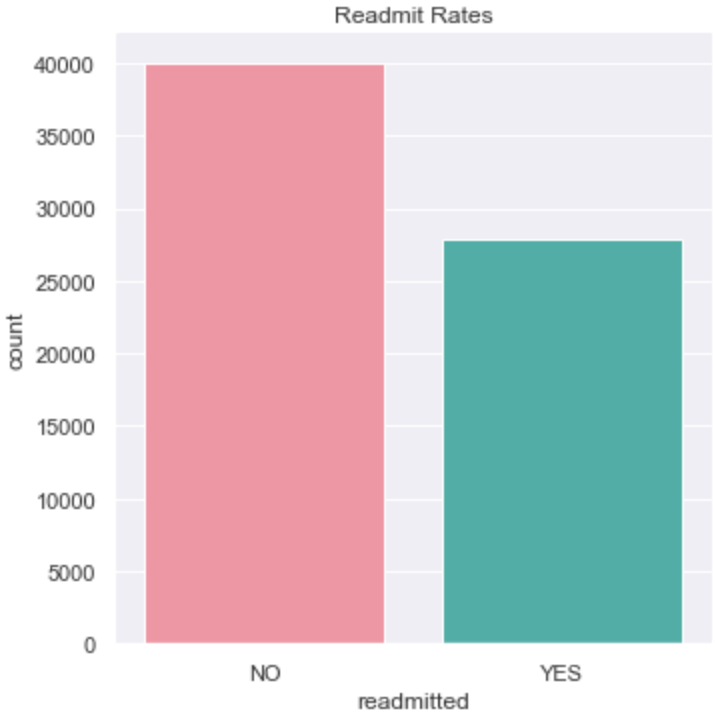
Feature Name	**Type**	**Description**
~~Encounter ID~~	~~Numeric~~	~~Unique identifier of an encounter~~
~~Patient number~~	~~Numeric~~	~~Unique identifier of a patient~~
Race	Nominal	Caucasian, African American, Hispanic, Asian, Other

Gender	Nominal	Male, Female
Age	Nominal	Grouped in 10-year intervals
~~Weight~~	~~Numeric~~	~~Weight in pounds~~
Admission type	Nominal	Integer identifier corresponding to 9 distinct values
~~Discharge disposition~~	~~Nominal~~	~~Integer identifier corresponding to 29 distinct values~~
Admission source	Nominal	Integer identifier corresponding to 21 distinct values
Time in hospital	Numeric	Integer number of days between admission and discharge
~~Payer code~~	~~Nominal~~	~~Integer identifier corresponding to 23 distinct values~~
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values
Number of lab procedures	Numeric	Number of lab tests performed during the encounter
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter
Number of medications	Numeric	Number of distinct generic names administered during the encounter
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9)
~~Diagnosis 2~~	~~Nominal~~	~~Secondary diagnosis (coded as first three digits of ICD9)~~
~~Diagnosis 3~~	~~Nominal~~	~~Additional secondary diagnosis (coded as first three digits of ICD9)~~
Number of diagnoses	Numeric	Number of diagnoses entered to the system
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken
A1c test result	Nominal	Indicates the range of the result or if the test was not taken
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name)
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed

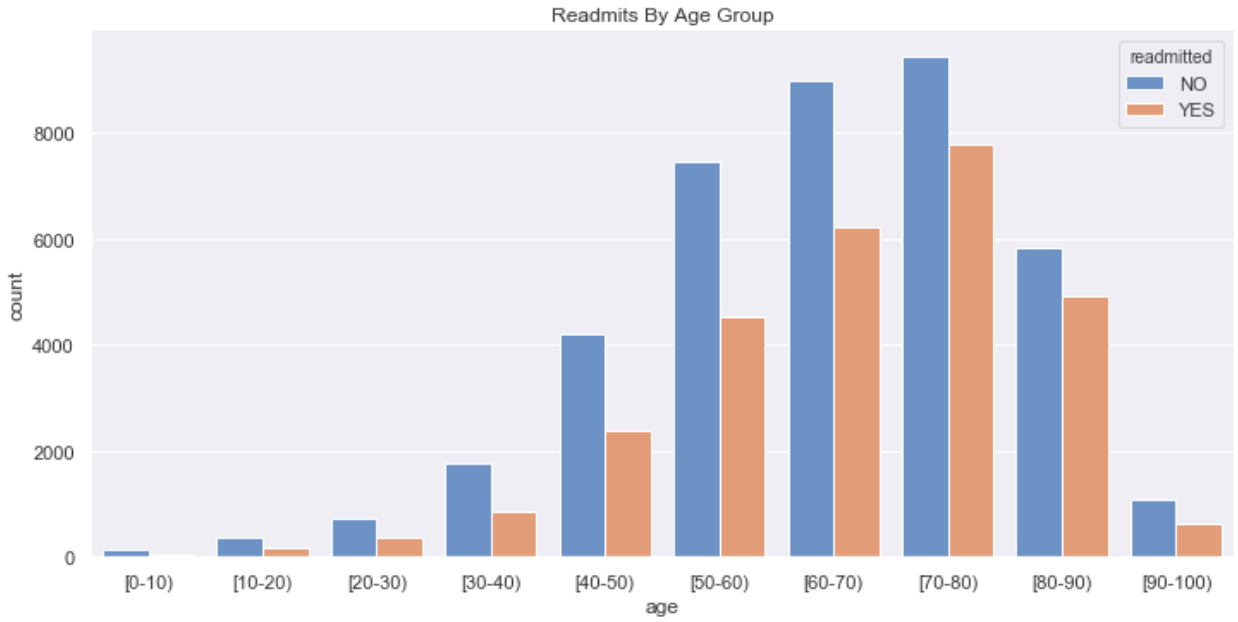
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage
Readmitted	Nominal	Days to inpatient readmission

Exploratory Data Analysis

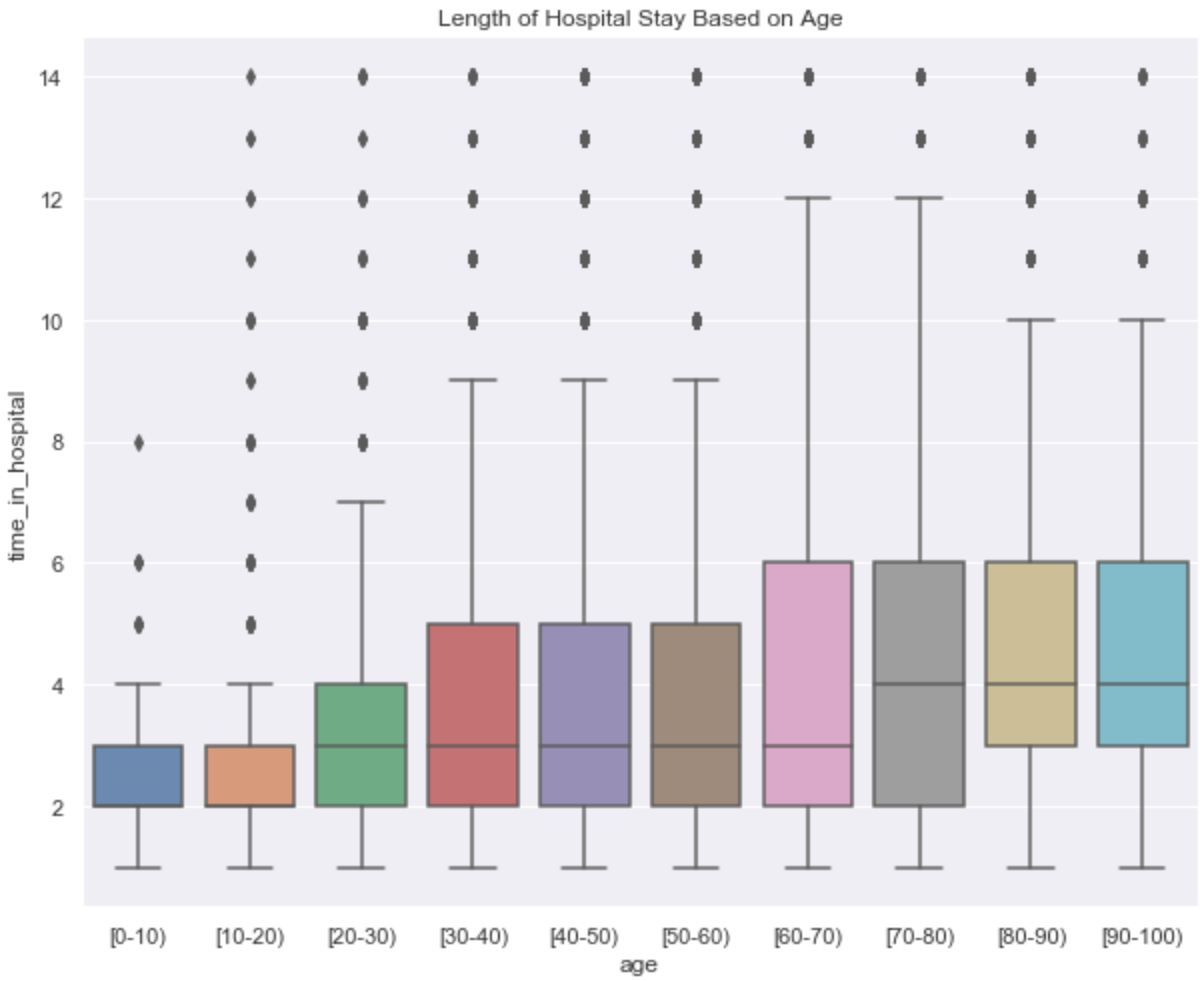
The goal here is to visualize and understand the impact of the target variable, readmission probability, and each of the independent variables. About 41% of patients are readmitted.



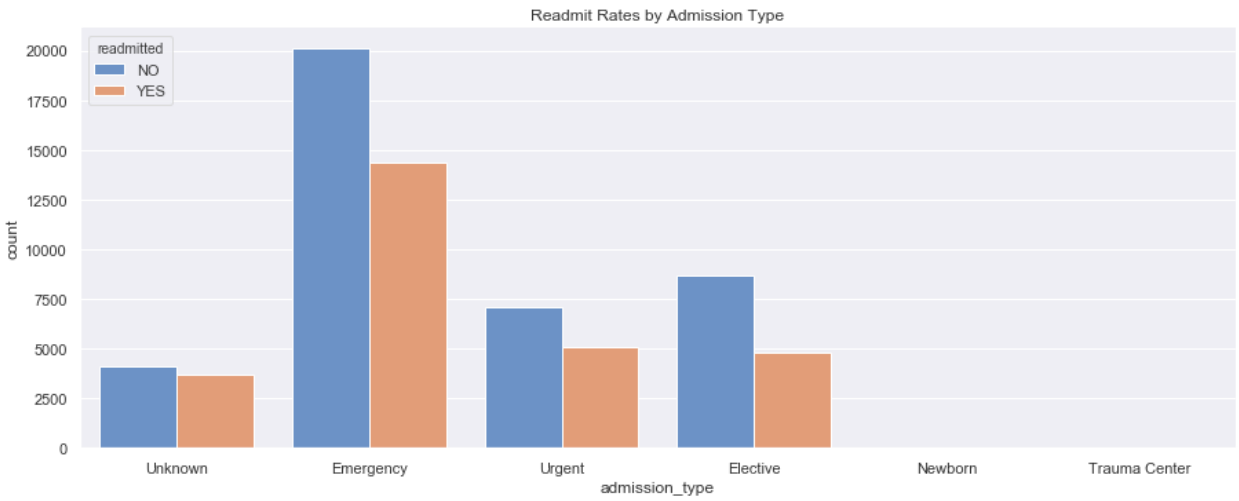
Demographics

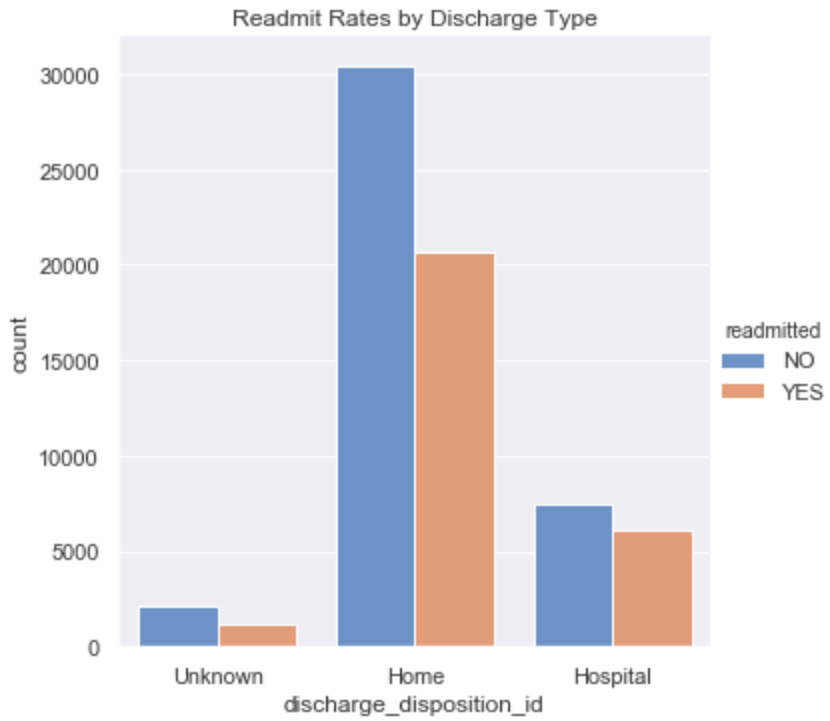


Patients are mostly caucasian followed by African Americans. Over half of all patients are over 50 years of age, though there aren't many patients over 90 years old. Older groups do have higher chances of returning to the hospital and they stay longer too (see below), but overall, most receive adequate care and are not readmitted.

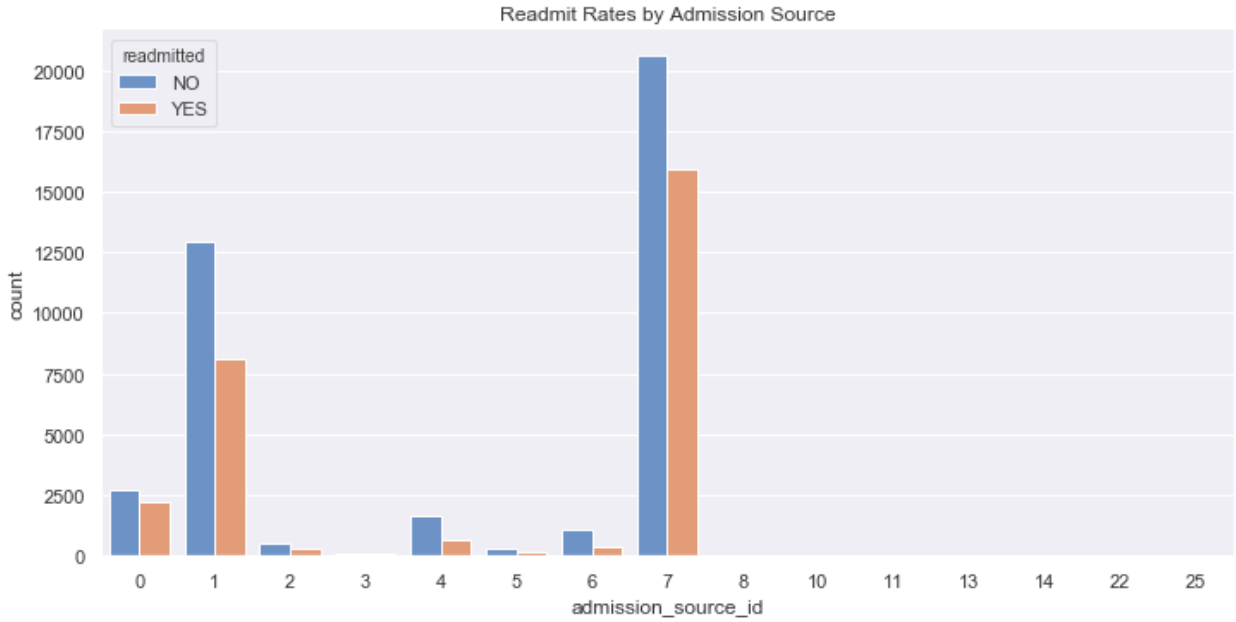


Admission and Discharge



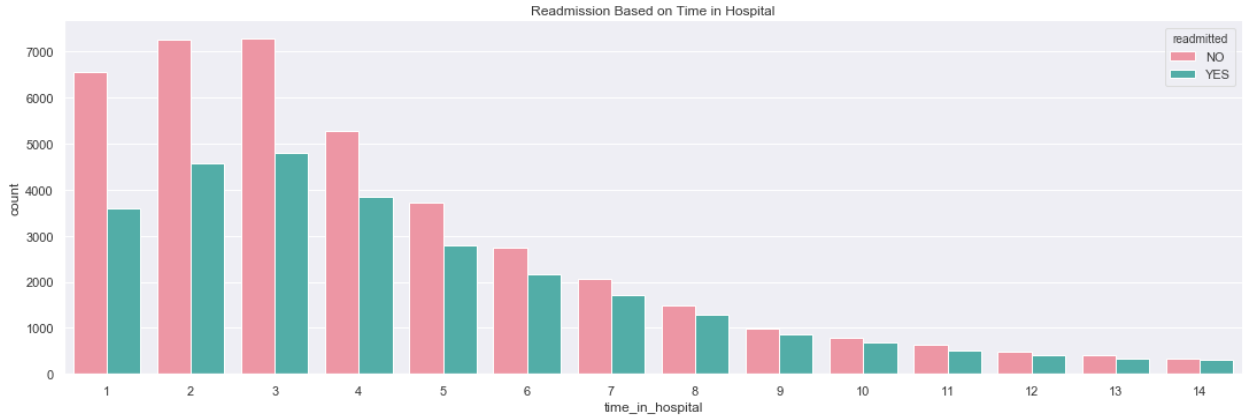


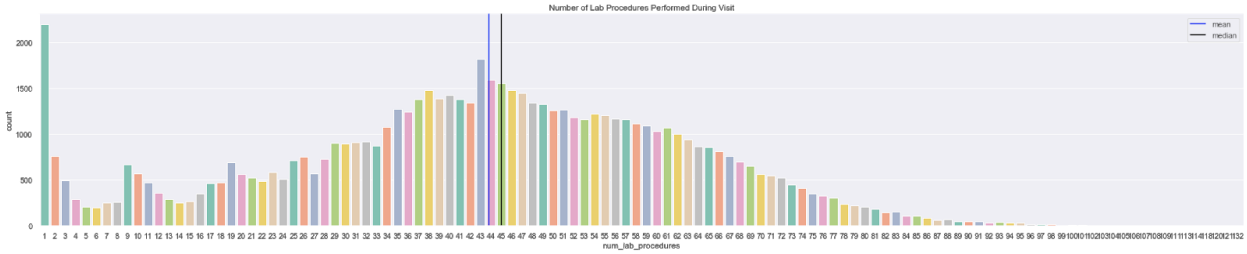
There are multiple outcomes for discharged, so I grouped them into three categories since patients either need more care, go home, or there is no data available for the outcome.

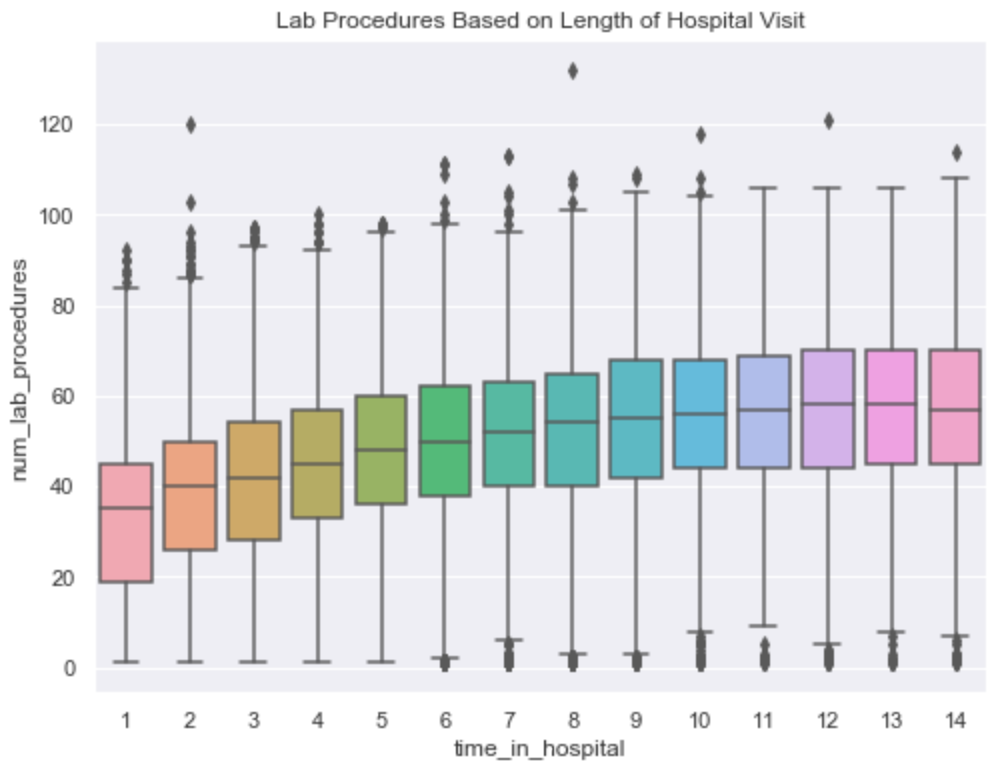


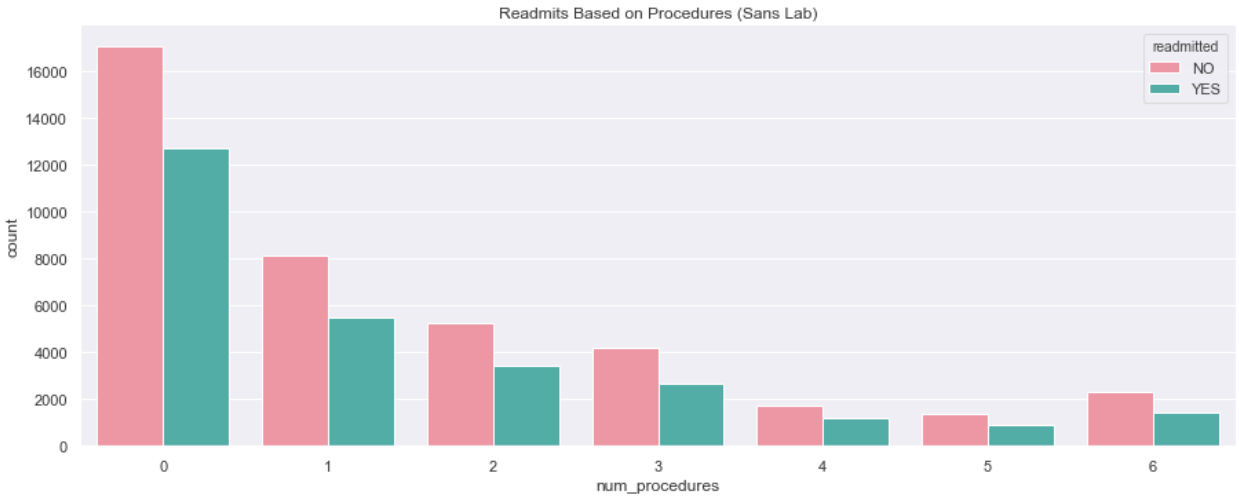
The most popular sources of admission are either from a physician referral (ID 1) or a visit to the emergency room (ID 7).

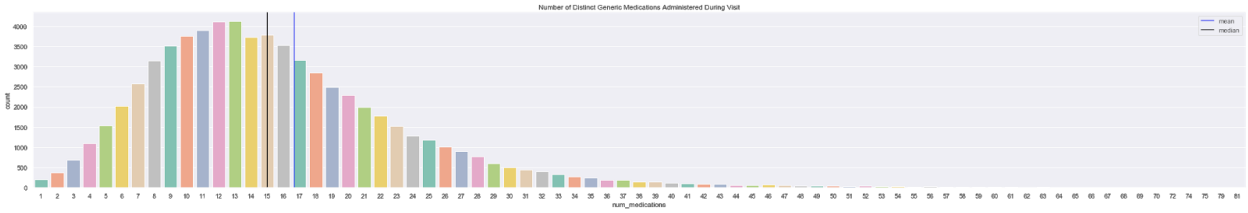
Hospital Visit Events



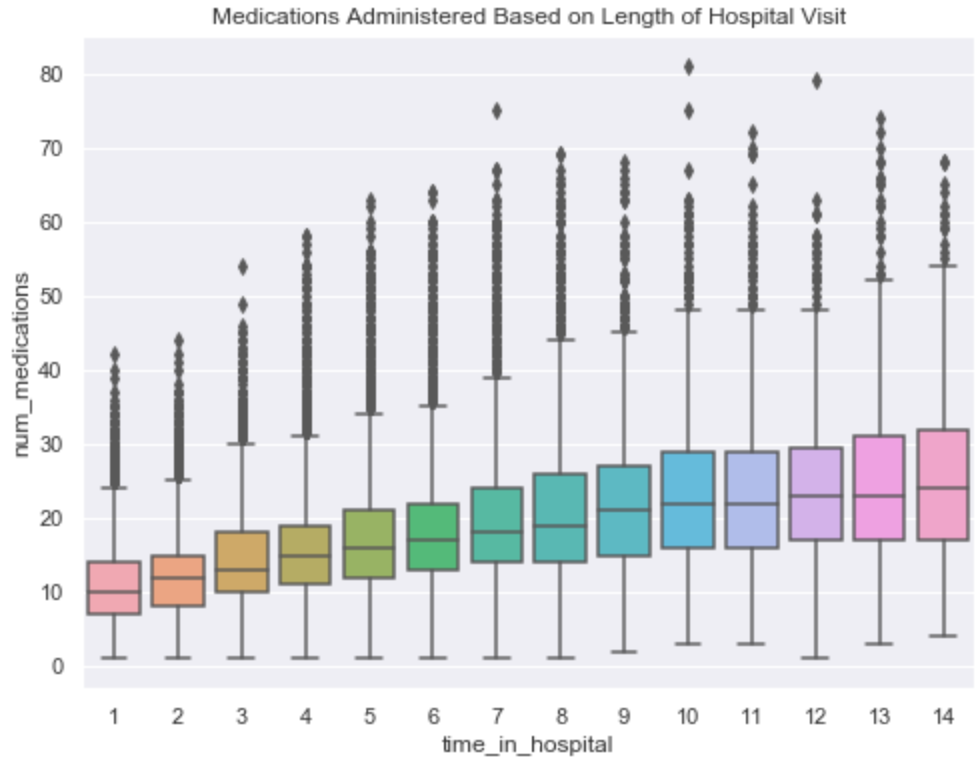




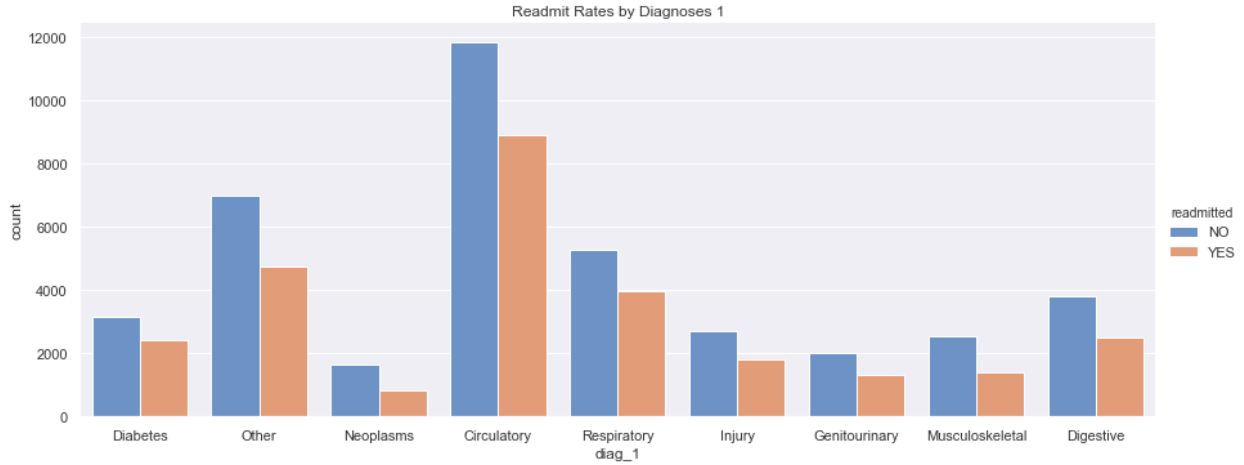


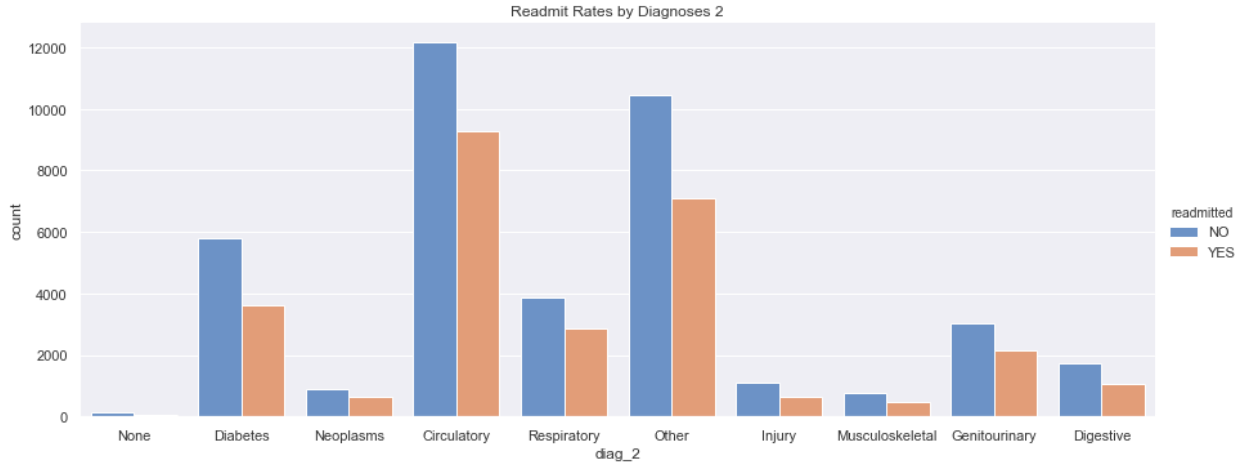


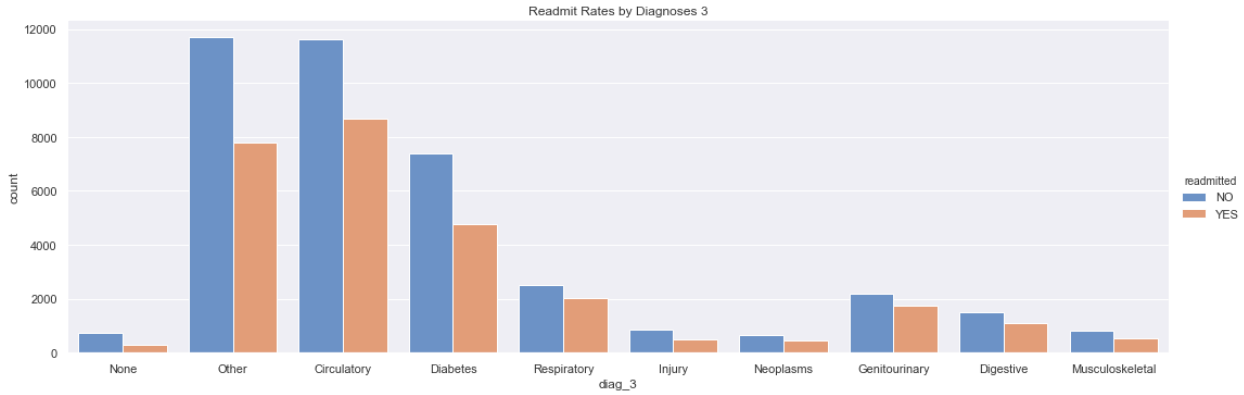
When a patient visits the hospital, doctors will order tests to be run on the patient in order to determine the causes for their ailments.



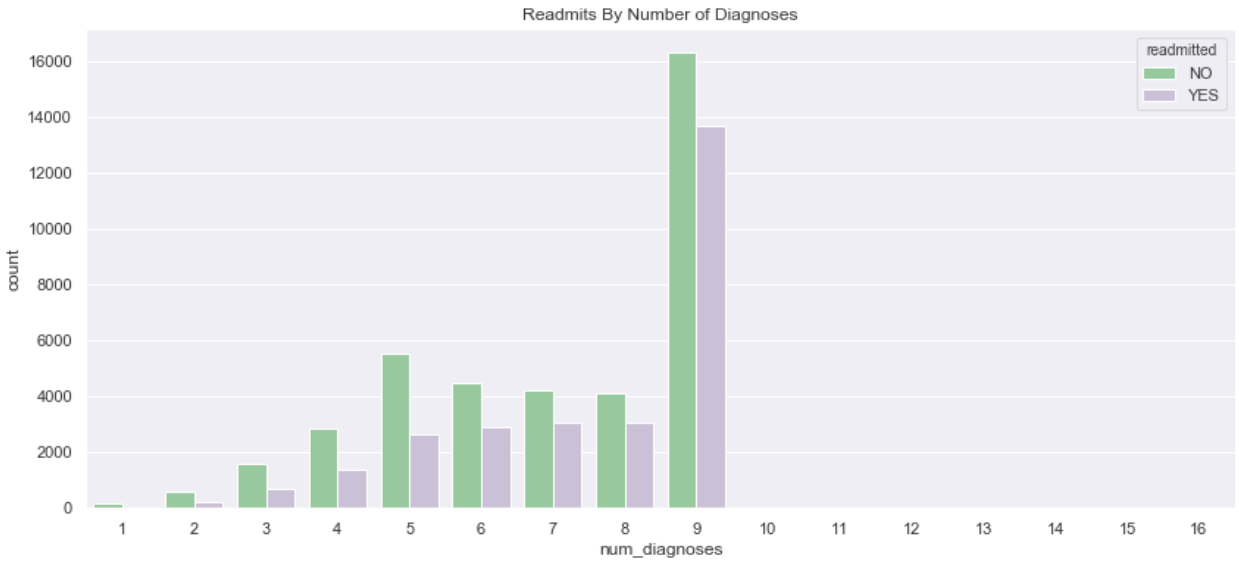
Diagnosed Conditions





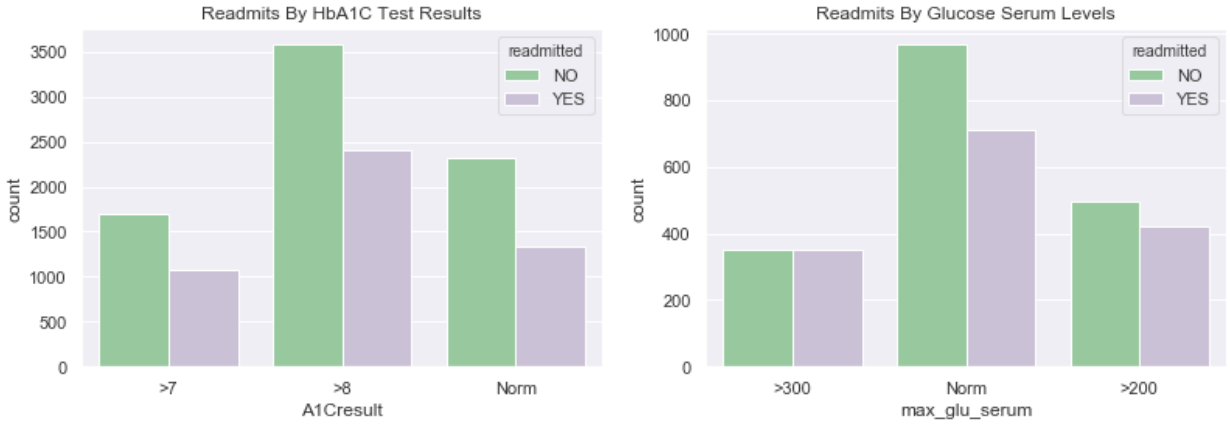


For each of the three listed diagnoses, the most popular options are circulatory, other, respiratory, and diabetes.



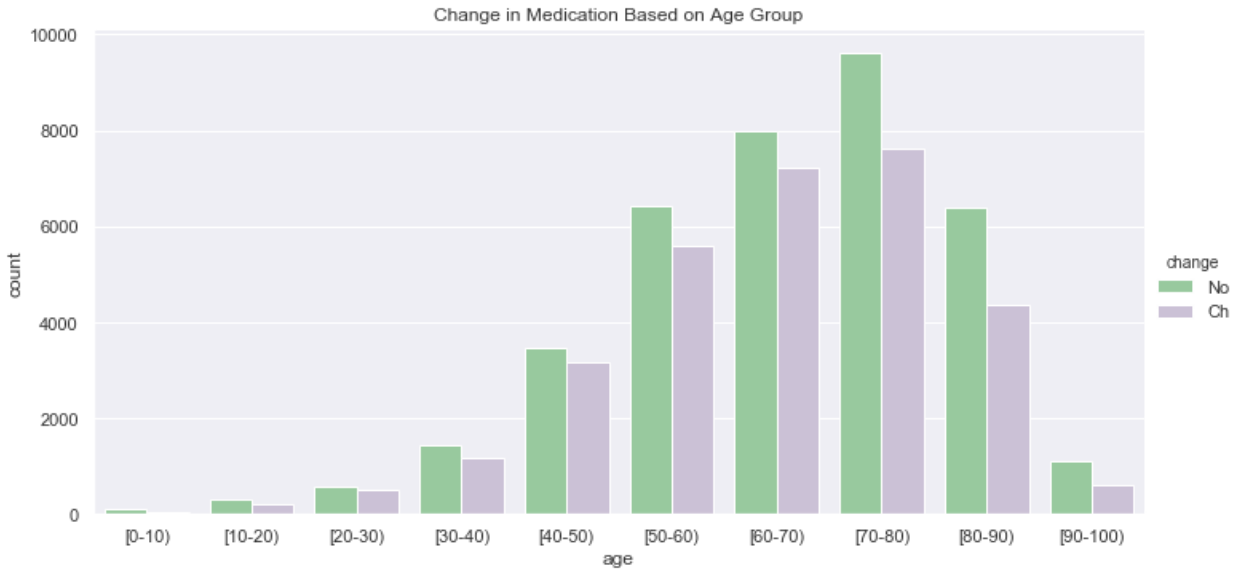
A large number of patients are diagnosed with 9 conditions during their visit!

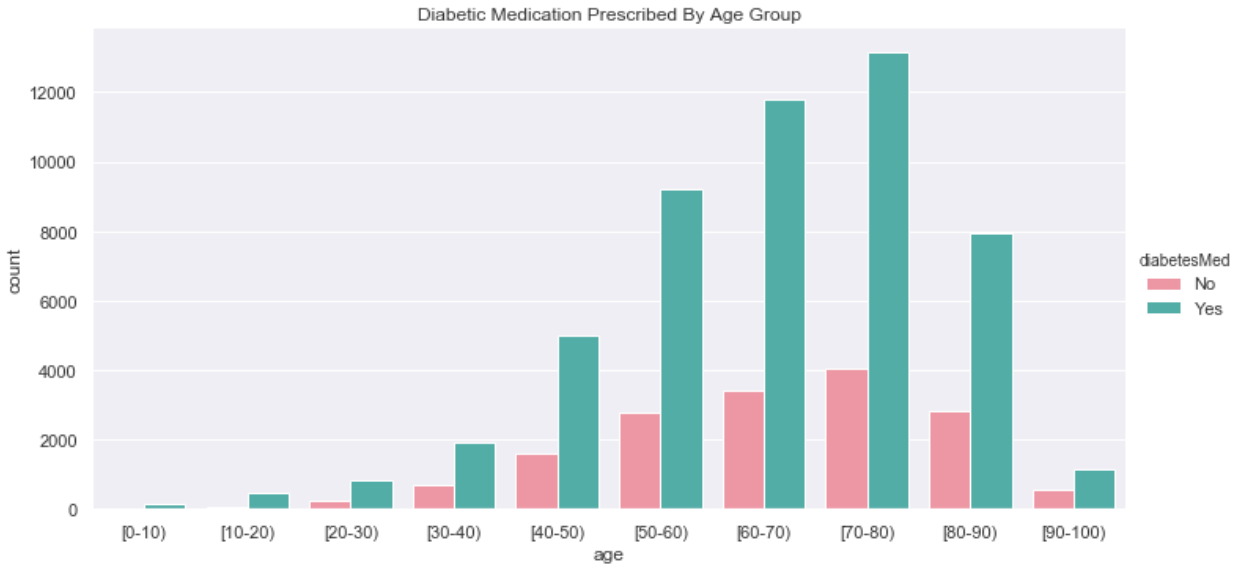
Glucose Serum and HbA1C Test Results



Since the majority of patients do not have a reading for the HbA1c and glucose serum tests, the plots above are only referring to patients who took either test. The HbA1c test, or glycated hemoglobin test, shows an individual's average blood sugar levels for the past 2-3 months. Higher percentages (>8%) indicate that a patient has consistently had high blood sugar levels. The glucose serum test measures the amount of glucose in a patient's body and it is noticeable that higher levels of found glucose indicates a stronger likelihood of returning to the hospital.

Medications





Medications often keep a diabetic's blood sugar levels under control. In every age group, patients are not receiving new medications but the same prescriptions for their current ones, which is a good sign because that means the medication is stabilizing the patient.

Outliers

After visualizing the data, I couldn't help but notice the outliers found in most numerical variables. They often exceeded the variable average by at least three standard deviations. To keep the data representative of sample majorities, I removed any samples that were more than three standard deviations from the average.

Statistical Analysis

The goal of statistical analysis is to analyze the significance of the relationships between the independent variables and the target variable, whether a patient will be readmitted to the hospital. Our data is a mix of numerical and categorical (nominal) variables, and we approach establishing statistical significance differently based on type. Based on our target variable, this is a binary classification problem. Patients are either readmitted or not readmitted. Previously, we combined the two readmitted columns together so the patients that are readmitted within 30 days and after 30 days are now in one class. The final step before we start building our models is to encode all categorical variables using dummy variables.

Categorical Variables

We use the chi-square test for association (or chi square test for independence) to compute the chi-square statistics and p-value between two categorical variables, the independent and target variables. For each comparison, we establish a null hypothesis that there is no relationship between the two variables with a p-value threshold of 0.005. If the calculated p-value is less than our threshold value, then we may safely reject the null hypothesis and claim that there is a relationship between the variables.
The medications: nateglinide, chlorpropamide, glimepiride, acetohexamide, glyburide, tolbutamide, acarbose, miglitol, troglitazone, tolazamide, glyburide-metformin, glipizide-metformin, and metformin-pioglitazone all failed to pass the test since they have p-values greater than 0.005. We removed these columns from the dataset.

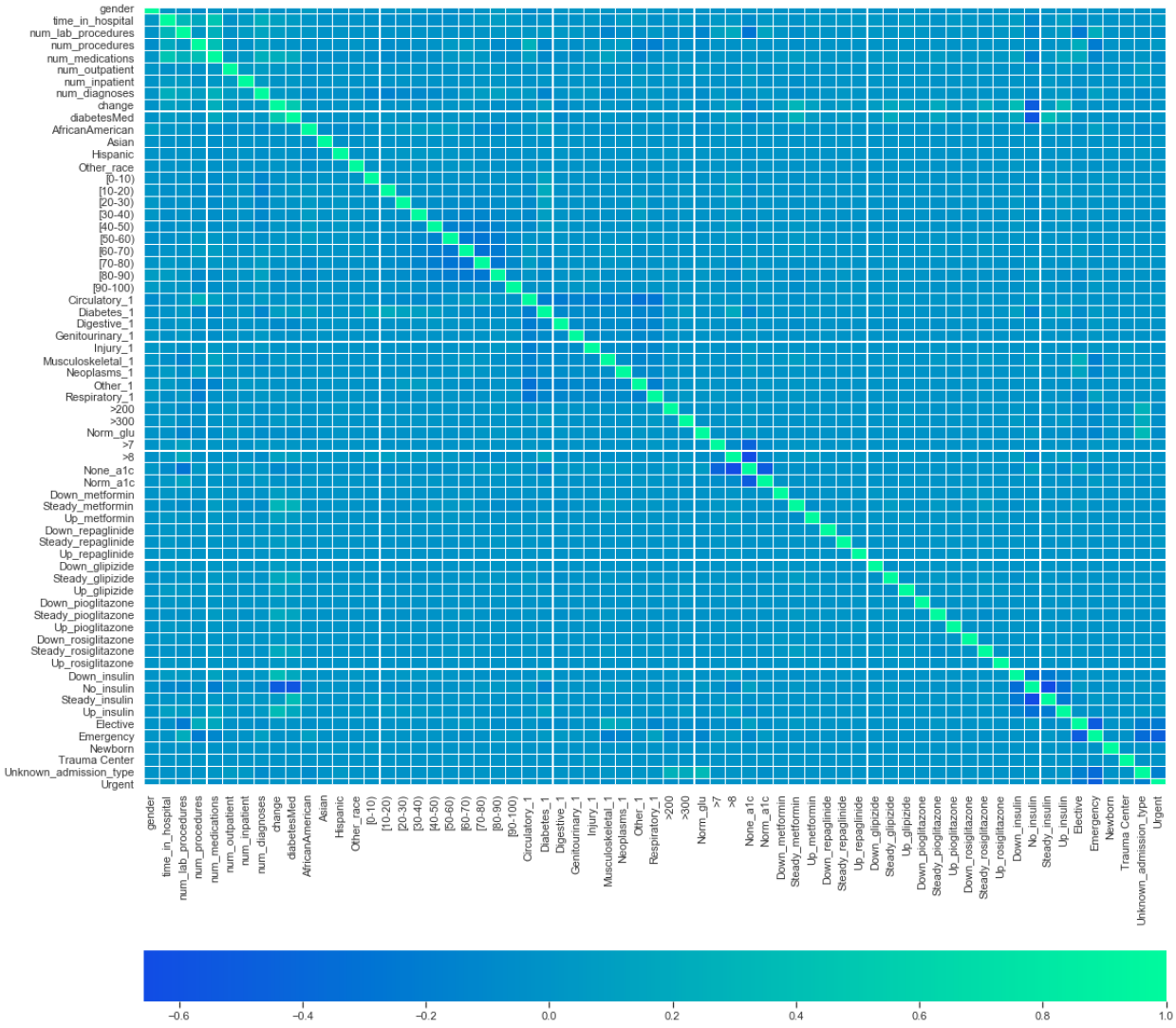
Numerical Variables

For the numerical variables, we use the analysis of variance test (an extension of the t-test) to compare the means of the two classes and understand the significance of their difference. Once again, our p-value threshold is 0.005 and we establish the null hypothesis as the mean for admitted patients is the same as the mean for not admitted patients. All of the numerical variables (admission_source_id, time_in_hospital, num_lab_procedures, num_procedures, num_medications, num_outpatient, num_emergency, num_inpatient, diag_1, diag_2, diag_3, and num_diagnoses) had p-values < 0.005 and thus, we safely rejected the null hypothesis and accepted the alternate hypothesis that the means for each class are not the same.

Label Encoding

Machine learning models are better adapted to working with numerical columns as opposed to categorical ones. Before building the models, we have to convert categorical variables to numeric columns using dummy variables and label encoding. Binary columns are replaced with 0 for No and 1 for Yes. Complex categorical variables, such as max_glu_serum, are split into columns, one for each option, and replaced with 0 and 1s depending on which option is stated in the original column. Once we finished encoding our variables, we combined the numerical

and newly encoded variables together. Finally, we created a correlation table and addressed the correlation coefficients. If a pair of columns have a coefficient greater than 0.7 or less than -0.7, one of them must be removed since they are too closely related and may influence the model outcome. See the correlation table below and note that due to the numerous variables, we omitted the coefficients.



Machine Learning

The diabetes dataset presents a binary classification problem. For that, we are using (binary) logistic regression, decision tree classifier, random forest classifier, Gaussian Naive Bayes, SVC (support vector classifier), and Extra Trees Classifier. Since the percentage of readmits and not readmitted patients is 60% to 40%, the classes are not heavily imbalanced. To aid in improving the precision-recall-f1 scores, we tuned the parameters of each model.

Parameter Tuning

For each of the machine learning models, I laid out a set of parameters that, put together in the right combination, optimizes how well the model performs. Not all hyperparameter optimizations led to an increase in precision, recall, and f1 scores since the default settings already created the best model possible. While testing these models, I noticed that Decision Tree Classifier, Random Forest Classifier, and Extra Trees Classifier had high accuracy scores for training sets but not test sets, which is known as overfitting. In an effort to counteract this behavior, I used the Bagging Classifier, which fits base classifiers on random subsets of the original dataset and aggregates their individual predictions to make a final prediction. I also used the Gradient Boosting Classifier, which optimizes for arbitrary differentiable loss functions. These helped increase the precision and recall scores for each of the algorithms.

Model Selection and Next Steps

Using only the training and test sets (without any modifications), the precision scores hovered around 50-60 while recall scores fluctuated between 18-90. Once under and over sampling methods shaped the training and test sets, scores improved. Randomly under sampled data, which randomly removes samples from the majority class, combined with the Extra Trees Classifier model produced the most reliable results (see below).

	Precision	**Recall**	**f1-score**
NO	0.70	0.73	0.71
YES	0.72	0.69	0.70

Random over sampler and random forest produced another reliable model.

	Precision	**Recall**	**f1-score**
NO	0.71	0.71	0.71
YES	0.71	0.72	0.71

Further investigation is needed to create an even more optimal model for representing this data, including using new methods as a way of organizing the samples into groups.