The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. **The goal is to determine the early readmission of the patient within 30 days of discharge**. The problem is important for the following reasons. Despite high-quality evidence showing improved clinical outcomes for diabetic patients who receive various preventive and therapeutic interventions, many patients do not receive them. This can be partially attributed to arbitrary diabetes management in hospital environments, which fail to attend to glycemic control. Failure to provide proper diabetes care not only increases the managing costs for the hospitals (as the patients are readmitted) but also impacts the morbidity and mortality of the patients, who may face complications associated with diabetes.

ANALYSIS FROM THE DATA

The data contains 13 Numerical features and 37 categorical features.

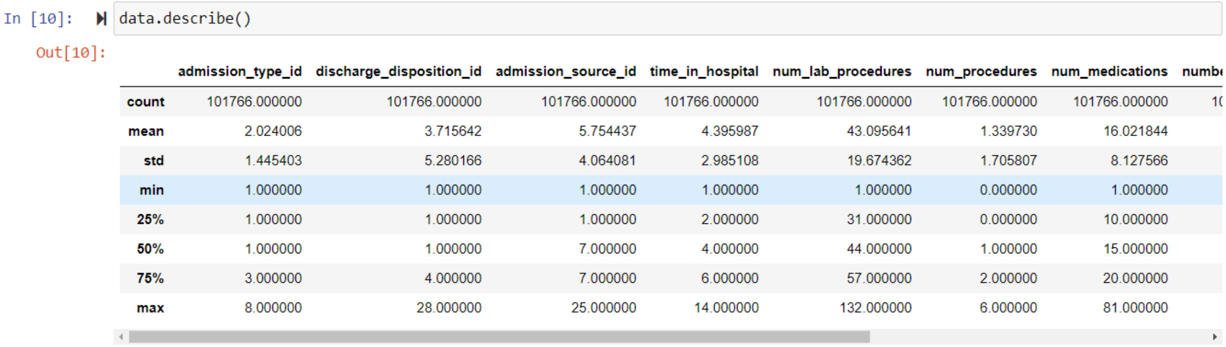
The data shape is (101766, 50) with readmitted the Target Column

Few columns are unique to each row like the 'id', 'encounter_id', 'patient_nbr','payer_code'

This Features are removed as they represent unique row and do not contribute to model building

Columns admission_type_id, discharge_disposition_id, admission_source_id are the categorical columns each entity mapping to a special category.

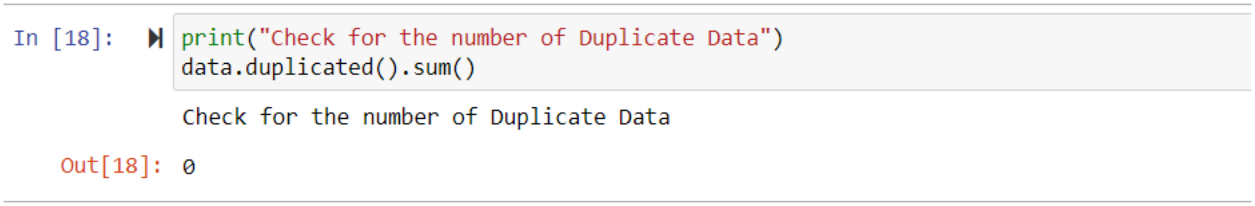
Description of the data



```
In [10]:    data.describe()
Out[10]:
```

	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	num_lab_procedures	num_procedures	num_medications	numbe
count	101766.000000	101766.000000	101766.000000	101766.000000	101766.000000	101766.000000	101766.000000	1(
mean	2.024006	3.715642	5.754437	4.395987	43.095641	1.339730	16.021844	
std	1.445403	5.280166	4.064081	2.985108	19.674362	1.705807	8.127566	
min	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	
25%	1.000000	1.000000	1.000000	2.000000	31.000000	0.000000	10.000000	
50%	1.000000	1.000000	7.000000	4.000000	44.000000	1.000000	15.000000	
75%	3.000000	4.000000	7.000000	6.000000	57.000000	2.000000	20.000000	
max	8.000000	28.000000	25.000000	14.000000	132.000000	6.000000	81.000000	

The data contained no duplicate value



```
In [18]:    print("Check for the number of Duplicate Data")
            data.duplicated().sum()

            Check for the number of Duplicate Data

Out[18]:    0
```

Missing values in this dataset are represented by '?'. Only very few of the columns have missing values.

race	2.233555
weight	96.858479
medical_specialty	49.082208
diag_1	0.020636
diag_2	0.351787
diag_3	1.398306

Six Features have missing values, a closer look at them to find a proper method to replacing such, either using mean, meadian or removing the features in entirety

Since weight has more than 70% missing it was dropped while diag_1, 2, 3 features where filled with the mode from each of the column

While Columns named 'examide','citoglipton' have only single value for all the entries In the data which does not add any value to the model training and hence removed

DATA TRANSFORMATION/ FEATURE ENGINEERING

There are majority of the columns that represent medications administered by the diabetic patients. This is important to observe the change in medications that might affect the readmission rates.

Discharge Disposition ID corresponding to [11 or 13 or 14 or 19 or 20 or 21] indicates patient has expired so there is no chance that it will readmit again so these records were removed.

Discharge Disposition ID has lots of distinct values using domain knowledge we will convert them into small number of categories.

As our focus is on hospital readmissions, we must disregard the entries of patients who might not return to the hospital. The discharge reason of the patient has been described in the discharge_disposition_id. We are removing the entries of patients whose discharge_disposition_id is Expired, Hospice / home, Hospice / medical facility, Expired at home. Medicaid only, hospice, Expired in a medical facility. Medicaid only, hospice, Expired, place unknown. Medicaid only, hospice.

The medical_specialty feature, which is crucial, has too many distinct values, so applying one hot encoding, it will unnecessarily create a lot of features, according to my research. I used a frequency-based method and domain knowledge, such as the idea that all types of surgeons should be included under the "surgeon" category, to divide them up into fewer categories.

While some was not grouped into any of the below medical specialist so, I grouped them into 'ungrouped' category.

- Endocrinology -- glands
- Gastroenterology --stomach
- Gynecology -- women reproduction system
- Hematology -- Blood
- Hematology/Oncology -- Blood

- Hospitalist -- one who takes care of admitted patients
- Oncology -- cancer
- Ophthalmology -- eye
- otolaryngology -- ears, nose, and throat
- Pulmonology -- respiratory
- Radiology -- diagnosing and treating injuries and diseases using medical imaging (radiology) procedures (exams/tests) such as X-rays

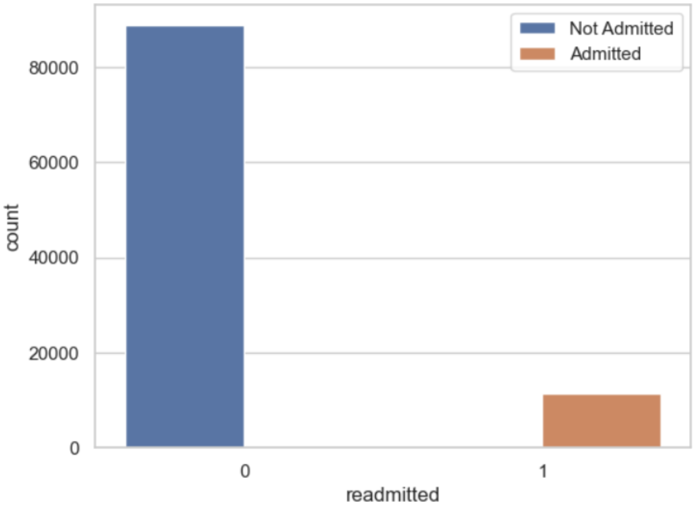
Domain Knowledge Type --> ICD Values --> Description

- Circulatory --> 390–459, 785 --> Diseases of the circulatory system
- Respiratory --> 460–519, 786 --> Diseases of the respiratory system
- Digestive --> 520–579, 787 --> Diseases of the digestive system
- Diabetes --> 250.xx --> Diabetes mellitus
- Injury --> 800–999 --> Injury and poisoning
- Musculoskeletal --> 710–739 --> Diseases of the musculoskeletal system and connective tissue
- Genitourinary --> 580–629, 788 --> Diseases of the genitourinary system
- Neoplasms --> 140–239 --> Neoplasms
- Pregnecy --> 630–679 --> Complications of pregnancy, childbirth, and the puerperium
- Other

Thsese features where mapped to diag_1, diag_2 and diag_3 for creating new features

The Target Column was an Imbalance Data

0	88757
1	11357



NEW FEATURE ENGINEERING IDEA

If the frequency of person's visit to the hospital is high then I can think of that person to be less healthier and less healthier patient tends to readmit quickly lets create health_index variable.Higher the health_index lesser the chance that person will readmit (indirectely propotional)

Health_index = (1 / (number_emergency + number_inpatient + number_outpatient))

Severity of disease is high if patient is spending lots of time in hospital and going through number of complicated test so, lets create severity of disease as feature. To get probablistic interpretation lets divide it by total values.

severity_of_disease = (time_in_hospital + num_procedures + num_medications + num_lab_procedures + number_of_diagnoses)

Research has found that the patient which keep going through changes(up/down) in proportion of medications is tend to readmit so we have engineered new variable called as 'number_of_changes'. This captures number of medications whose proportion have changed for each patient.

New Features were created which included Glucose Serum test, A1C test, total_medical_interactions, number_medications_per_diagnosis etc

STATISTICAL ANALYSIS

Spearsman correlation coefficient was used to check whether numerical features and readmitted column are dependant or independant if some features are found to be independant on readmitted there were simply remove.

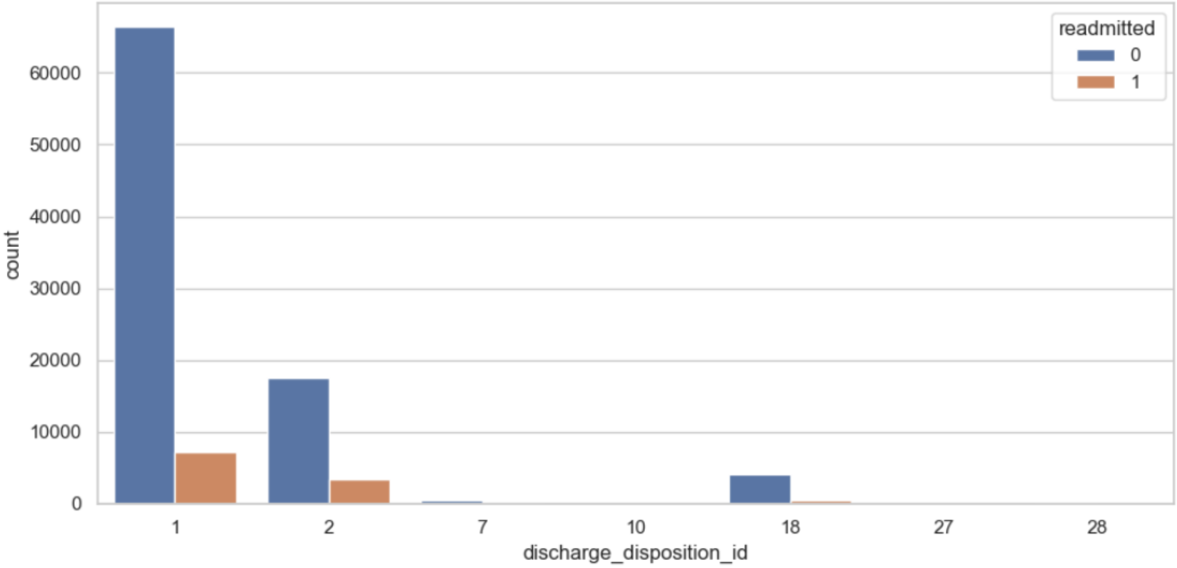
After analysis we found that correlation is always close to zero but, spearman doesn’t capture the non-linear relationships so, rather than using correlation coeff we used pvalue to get "rejected features list" Here hypothesis testing is done assuming null hypothesis to be "variables are independant" so assuming significance level = alpha = 0.35 if pvalue < alpha then reject null hypothesis that is we accept variables are dependant

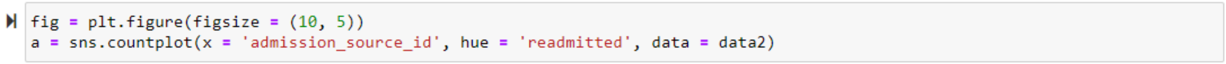
List of Features Rejected

['nateglinide', 'acetohexamide', 'glipizide', 'miglitol', 'troglitazone', 'tolazamide', 'examide', 'citoglipton', 'glyburide.metformin', 'glipizide.metformin', 'glimepiride.pioglitazone', 'metformin.rosiglitazone', 'metformin.pioglitazone', 'average_lab_procedure_cost']

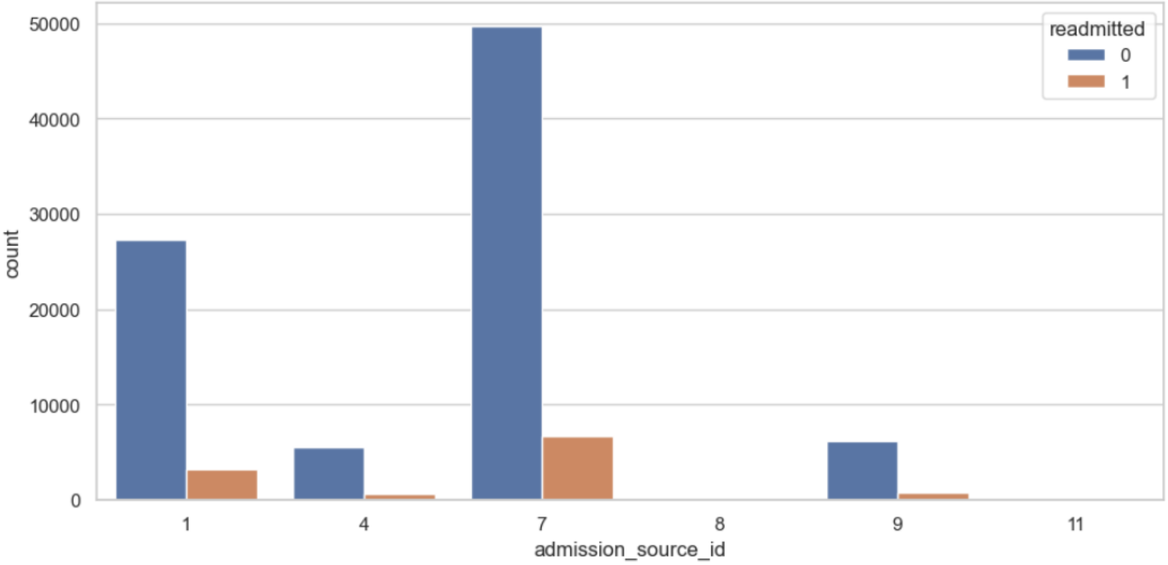
DATA VISUALIZATION

From the graph it clear that if discharge disposition id is 7 the patient wont readmit.



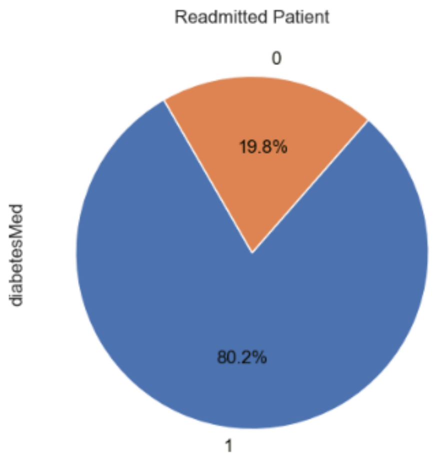


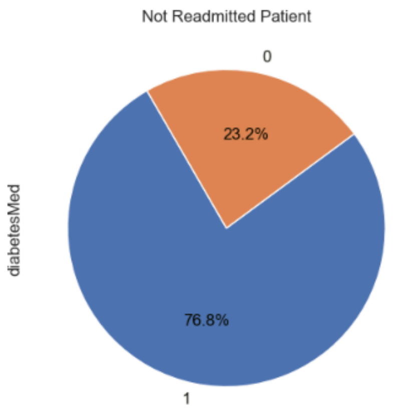
```
fig = plt.figure(figsize = (10, 5))
a = sns.countplot(x = 'admission_source_id', hue = 'readmitted', data = data2)
```



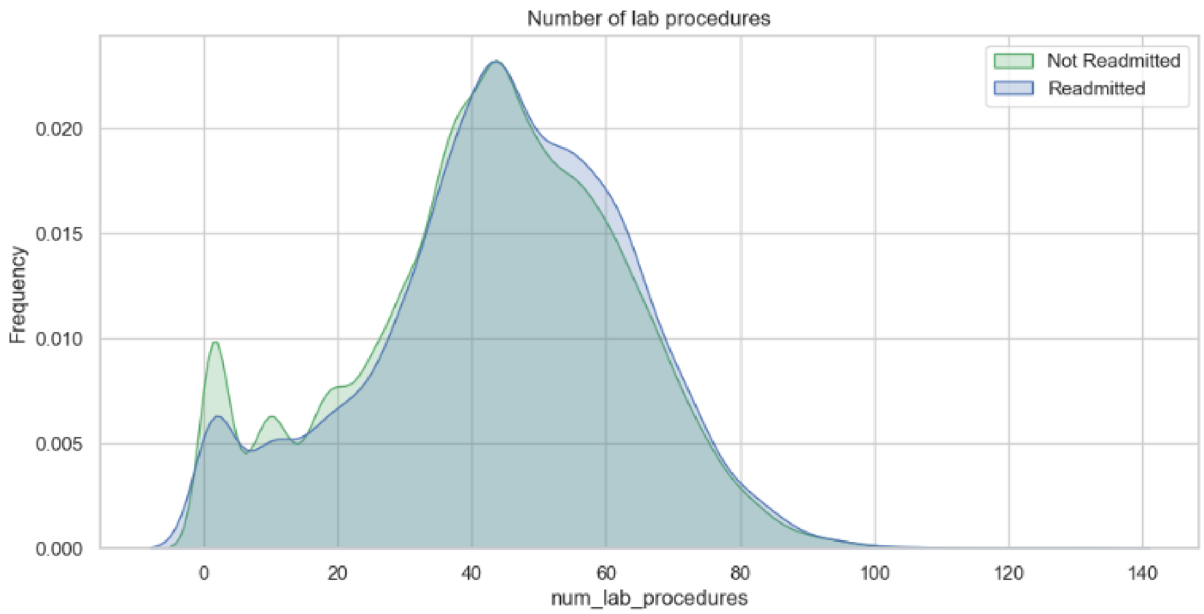
Most of the patient who readmitted have admission source as 1 and so if some patient has source id as 1 he is more likely be going to readmit.

This feature tells weather the patient has taken Diabetes Medication or not. In our dataset Number of patient taken Diabetes Medication and "readmitted" is almost same as number of patients taken Diabetes medication and "not readmitted".But the by iteracting with other feature Diabetes Med might reveal lot of information that is useful for given task.





Distribution Number of lab procedures for radmitted and not readmitted patient is exactely same. But it has high varience.High varience features are considered information rich features.



Summary:

From the EDA i found out that, The features like number of lab procedures, Diabetes Medication, admission are important in our task. By combining pairs like "change + DiabetesMed" , "age + time_in_hospital", "age + number_impatient" and "change + admission_source_id" can give us lots of information which is helpful in given task.

Found out that missing values with columns larger than 50% would add no value to Performance of the Data when trained with a model, and other categorical Features with missing values where filled with the Mode while Numerical were filled with the Mean.

In feature engineering domain knowledge was used for creating some features while other required feature interaction to engineer new features for columns like change, insulin, etc we have No , Up, Steady values but after experiments we found that it is good idea to convert them into numbers.

Statistical Analysis was used to reject Some Features using Null Hypothesis.