CS109A Project Milestone 3: EDA, Adjusted project plan and Set Goals

Project Team: Karim Gowani, Ryan McGillicuddy, Yaseen Mohmand, Steven Worthington
Project Title: Predicting Hospital Readmission Rates for Diabetes.

Data Description & Preprocessing
The data consist of inpatient hospitalizations of diabetic patients with the response being whether they were readmitted early (within 30 days) after being discharged from their current hospitalization. The dataset was sourced from the UC Irvine ML repository.

We preprocessed the data to be at the patient-level, rather than at the 'encounter' level, since most patients only had a single encounter. We selected the most recent encounter and aggregated any encounter-level statistics (such as total number of encounters) in each patient record. Additional preprocessing steps included:

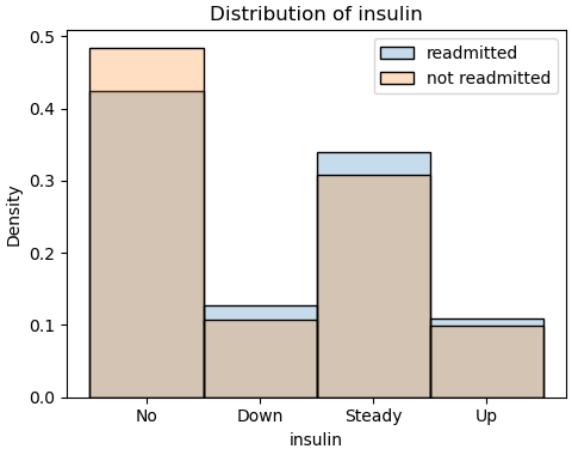
- Filtering out patients who expired or were transferred to hospice
- Filling in missing values especially for categorical variables with 'Unknown'
- Dropping columns such as weight that had 90%+ missing values
- Collapsing values of high cardinal columns into more meaningful groups

Additional details of preprocessing are in the accompanying notebook and our MS2 writeup.

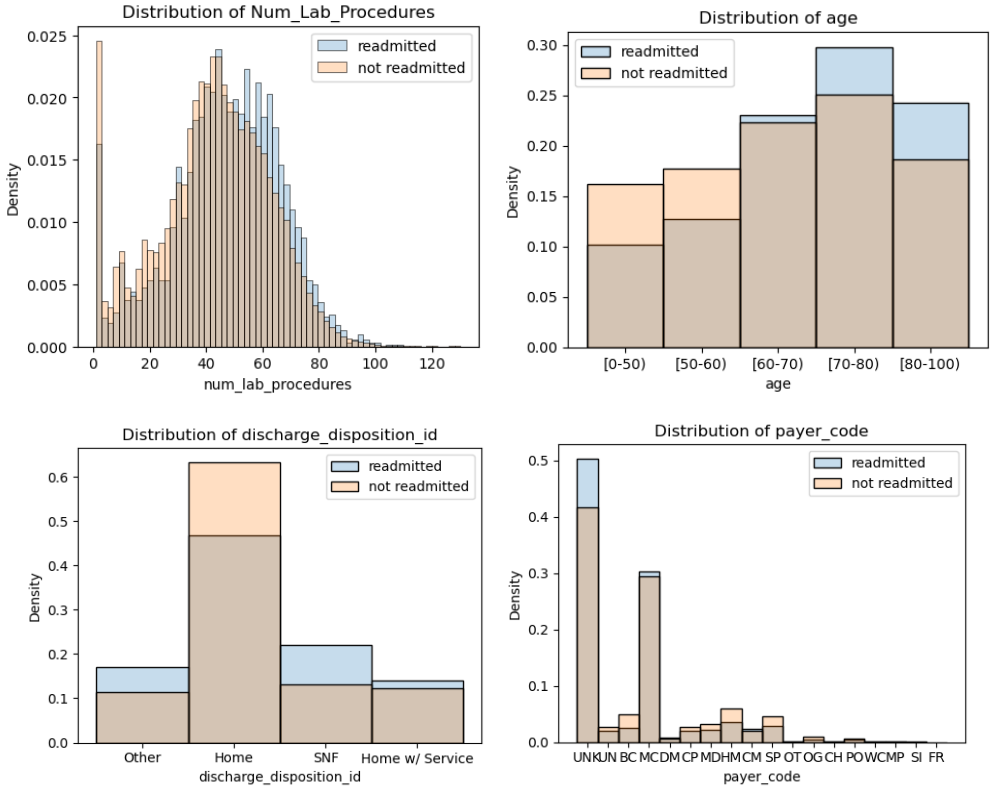
Exploratory Data Analysis
Following preprocessing, the data have 85 features, 69,990 observations, and no missing values, with the following counts of data types across the features.

Feature Data Type	Count
object	36
int64	39
float64	10

The 85 features can be grouped into three types. One group of features describe 23 drugs and their dosage changes, or lack of prescribing, resulting from a visit, with the possible categories being 'No', 'Steady', 'Up', and 'Down'. Exploratory data analysis of these drug features does not show a strong dependence of readmission on the drug categories for the drugs (see Appendix), except for insulin, which shows a different distribution between the readmitted and not readmitted categories.

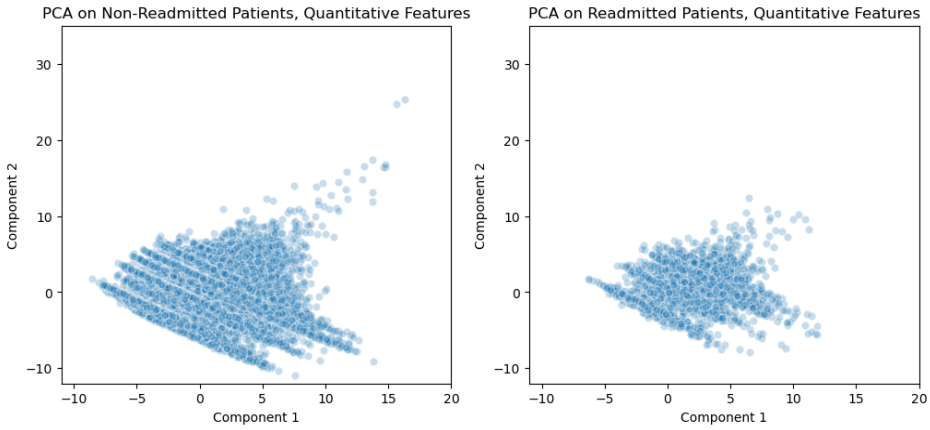


The remainder of the original features (seen in black in the Appendix) cover qualitative and quantitative features, with some taking on categories and others taking on quantitative values. Several features here show contrast in the distribution between readmitted and not readmitted, including the age of the patient, the number of lab procedures performed during the visit, the location to which a patient is discharged, and the payer code.



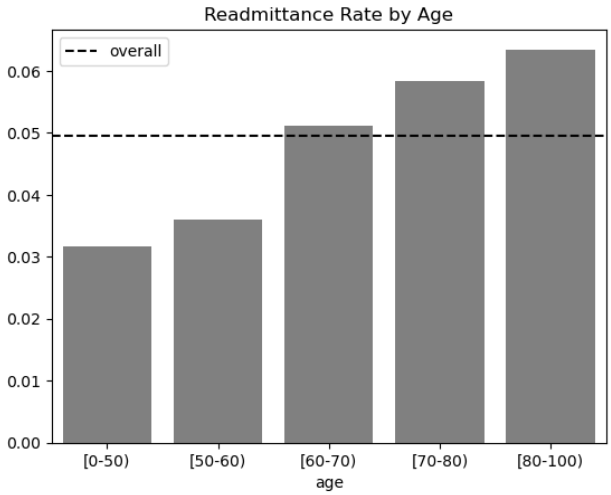
The remaining features are derived features (shown in blue in the Appendix) and are generated from the historical records of patients who appear multiple times in the dataset (the generation of these features is described in the Preprocessing section). For brevity, EDA on the derived features will not be shown here.

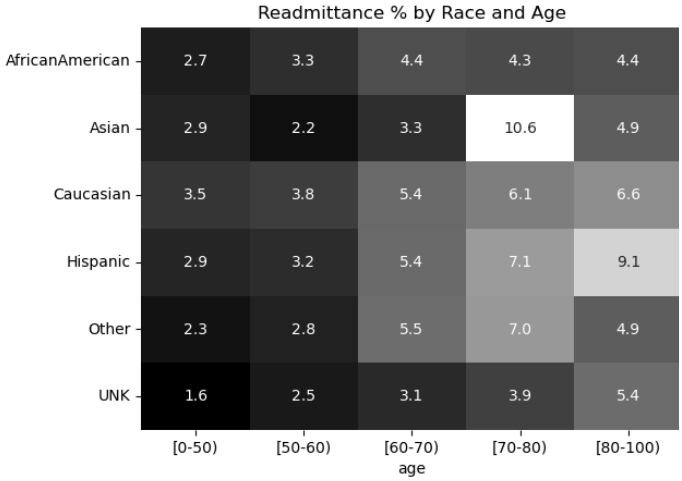
Given the large dimensionality of the dataset, PCA was attempted to see whether the readmitted classes would separate in a dimensionality reduced PC space. PCA was done on only the quantitative variables (all features that had non-object data types). We found that the readmitted classes were not well separated in the first vs. second principal component space, which is as expected given this noisy and complex dataset. The PCA, however, does reveal some outliers in this new representation of the feature space. As the full design matrix is a mix of categorical and quantitative data, PCA, which is not well suited for one hot encoded categorical data, was not attempted on the full dataset.



Meaningful Insights

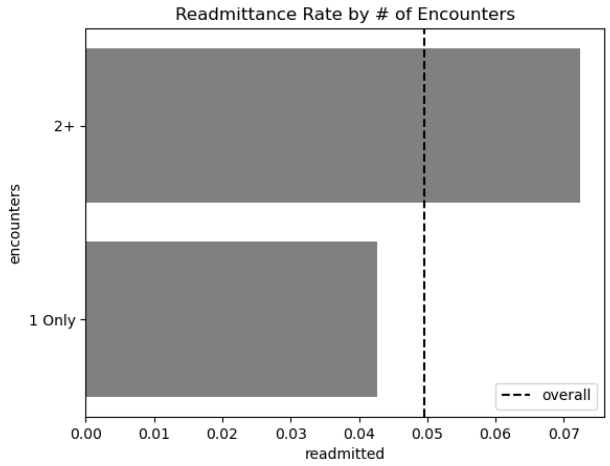
We analyzed early readmittance rates in terms of demographics and also medical condition and admission sources. In terms of demographics, readmittance rate, which is ~5% on average, increases with age (as one would expect). Also, patient race by itself is not an important driver but when crossed with age, it reveals combinations where the interaction of the two make a difference: Asians and Hispanics dominate the 80-90 and 90+ age groups respectively in terms of readmittance rate. Interestingly, African Americans show little variation across various age groups.

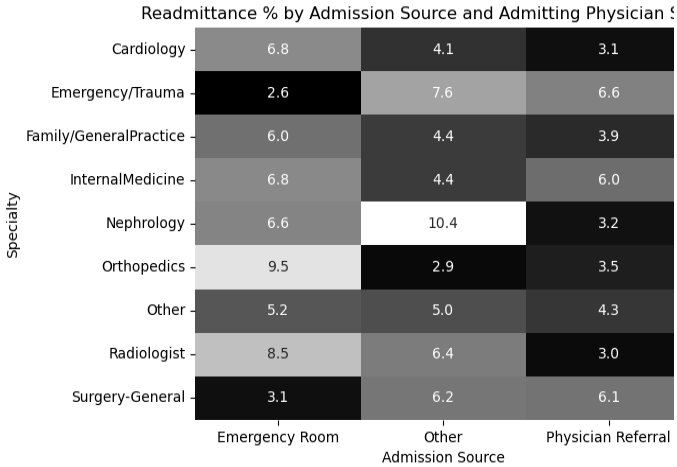




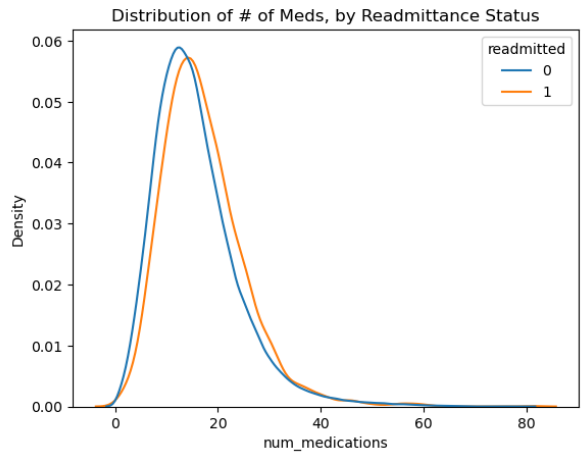
In terms of patients' previous medical history, those who had two or more previous hospitalizations (not surprisingly) were at a higher risk of being readmitted than otherwise: 7% vs. ~4%.

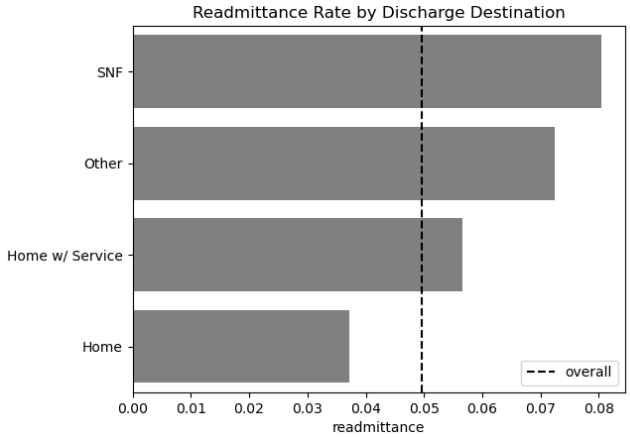
When patients did get admitted, there seems to be a link between how they got admitted and the admitting physician specialty and their chances of being readmitted: those that nephrologists admitted neither through a referral nor through the emergency room had a much higher readmittance rate of 10% (2x the average rate); perhaps these were cases of diabetic nephropathy where kidney function deteriorates with diabetes. Of those that were admitted through ER, orthopedists and radiologists top the list. Radiology is not surprising, as radiologists help make determinations after reviewing imaging, however, the fact that orthopedics shows high readmittance was at first confusing. Upon further research, we learned that diabetes increases the risk of falls through neuropathy, bone mineral density, etc.





Finally, patients that were on more medications were at more risk of being admitted. What is also interesting is that those patients that were discharged to a Skilled Nursing Facility (SNF) after their current hospitalization were also at higher risk: 8% vs. overall rate of 5%, as shown below:

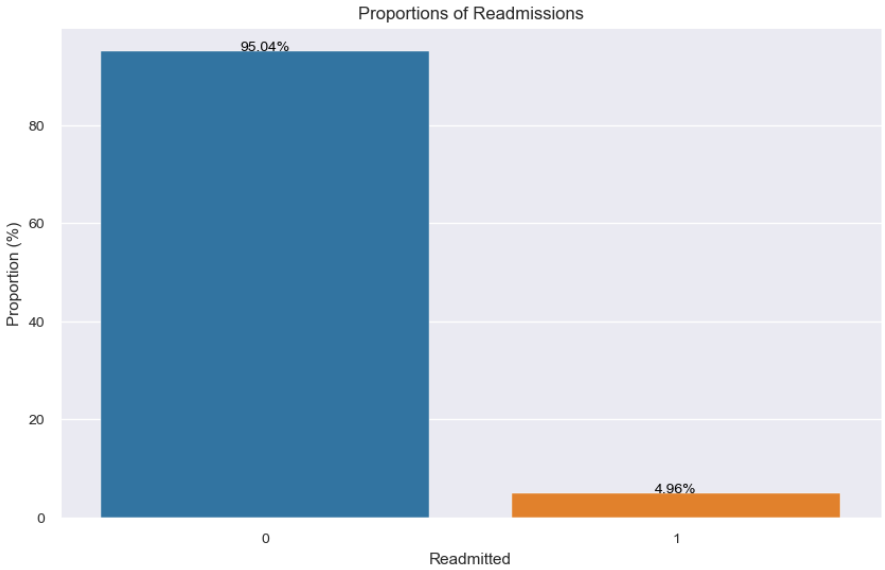




We expect all of these variables to play an important part in the modeling.

Class imbalance

We have 66,521 patients (95%) who were not readmitted within 30 days of their visit and 3,469 patients (5%) who were readmitted within the same period. The class imbalance in our response variable actually reflects the expected outcome in a hospital setting, where the majority of patients are not typically readmitted within 30 days of their visit. Therefore, our class distribution is representative of this reality.



Project Question

Based on the insights from the EDA, there are several features such as age, race, source of last admission, etc. that can help us build risk profiles of patients - beyond predicting likelihood of early readmission. Such risk profiles–for example, types of diabetes such as diabetic nephropathy, falls indirectly caused by progression of diabetes, etc.- can help practitioners design effective and targeted intervention and outreach programs for high risk patients. Identification of these important drivers will also help with acceptance of the model by the medical community as it will not be considered a black box.

Baseline Model / Implementation Plan

Given that we plan to predict a categorical response (early readmittance rate), this is a classification problem. For a baseline model, we will implement a simple logistic regression without any polynomial or interaction terms, and without regularization. Given the imbalance in the dataset and cross-sectional nature of the data, we will partition the data using stratified train/test splits. We will use cross validation for hyperparameter tuning.

Our primary performance metric will be AUC, for a couple of reasons. First, our data have class imbalance and AUC is invariant to this imbalance. Second, AUC is also a good metric vs. F1 score given that the cost of false positives is different than that of false negatives in this problem. It is better in this case to predict someone as positive even if we are incorrect: predicting someone as not needing follow-ups when they actually are in need of immediate follow-ups can be disastrous for their health outcomes vs. predicting them positive (when they are actually not) at the much lower costs of outreach.

Appendix/Supplementary Figures

Features with int64 data types (derived features):

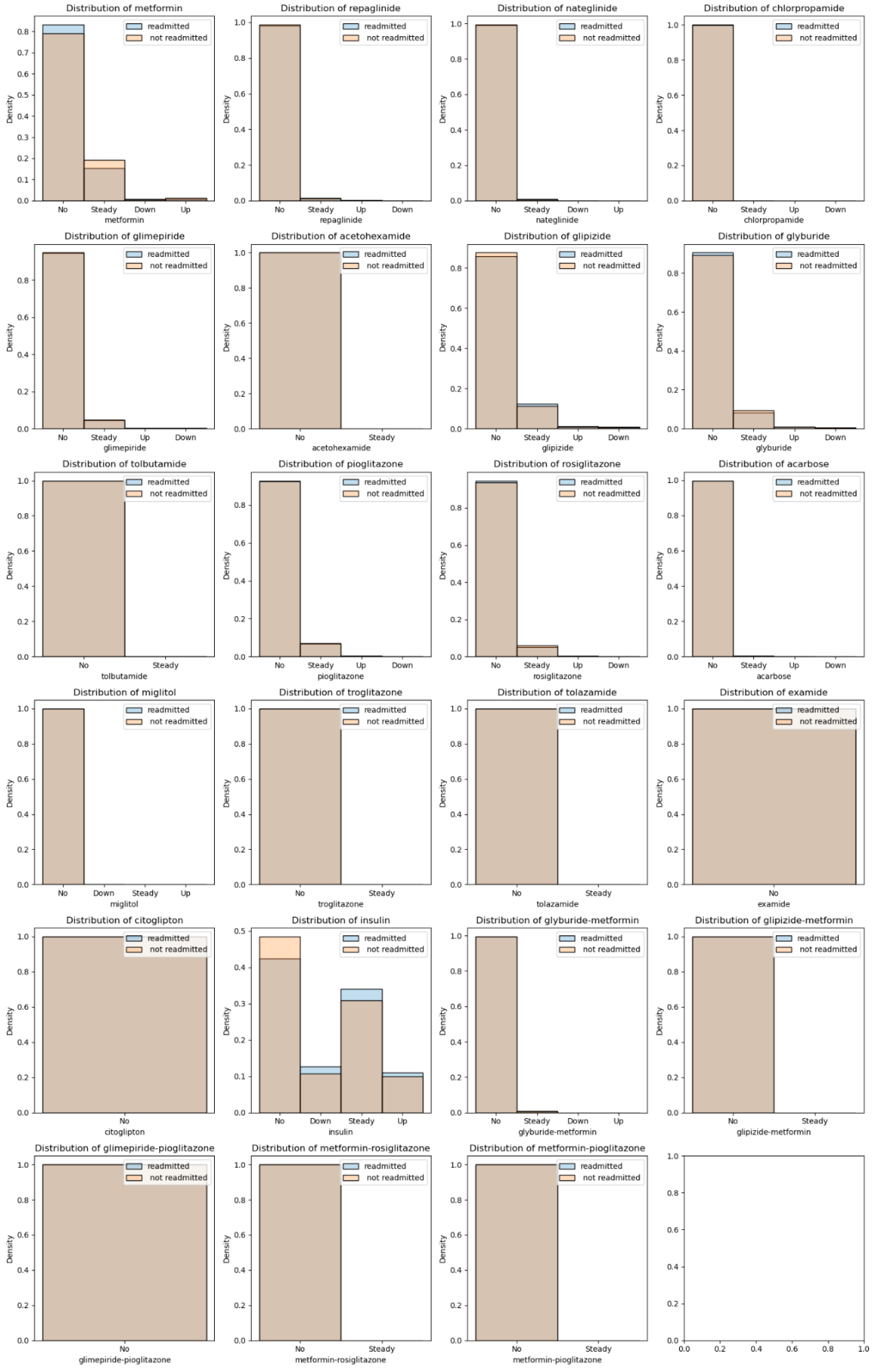
```
'patient_nbr', 'time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications',
'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses', 'change',
'diabetesMed', 'num_encounters', 'min_time_in_hospital', 'max_time_in_hospital',
'min_num_lab_procedures', 'max_num_lab_procedures', 'min_num_procedures', 'max_num_procedures',
'min_num_medications', 'max_num_medications', 'min_diagnoses', 'max_diagnoses',
'unique_glu_measurements', 'num_times_glu_high', 'glu_always_high', 'glu_ever_high',
'unique_a1c_results', 'num_times_a1c_high', 'a1c_always_high', 'a1c_ever_high',
'num_times_med_changed', 'med_always_changed', 'med_ever_changed',
'num_times_diabetic_med_prescribed', 'diabetic_med_always_prescribed',
'diabetic_med_ever_prescribed', 'num_times_readmitted', 'always_readmitted', 'ever_readmitted'
```

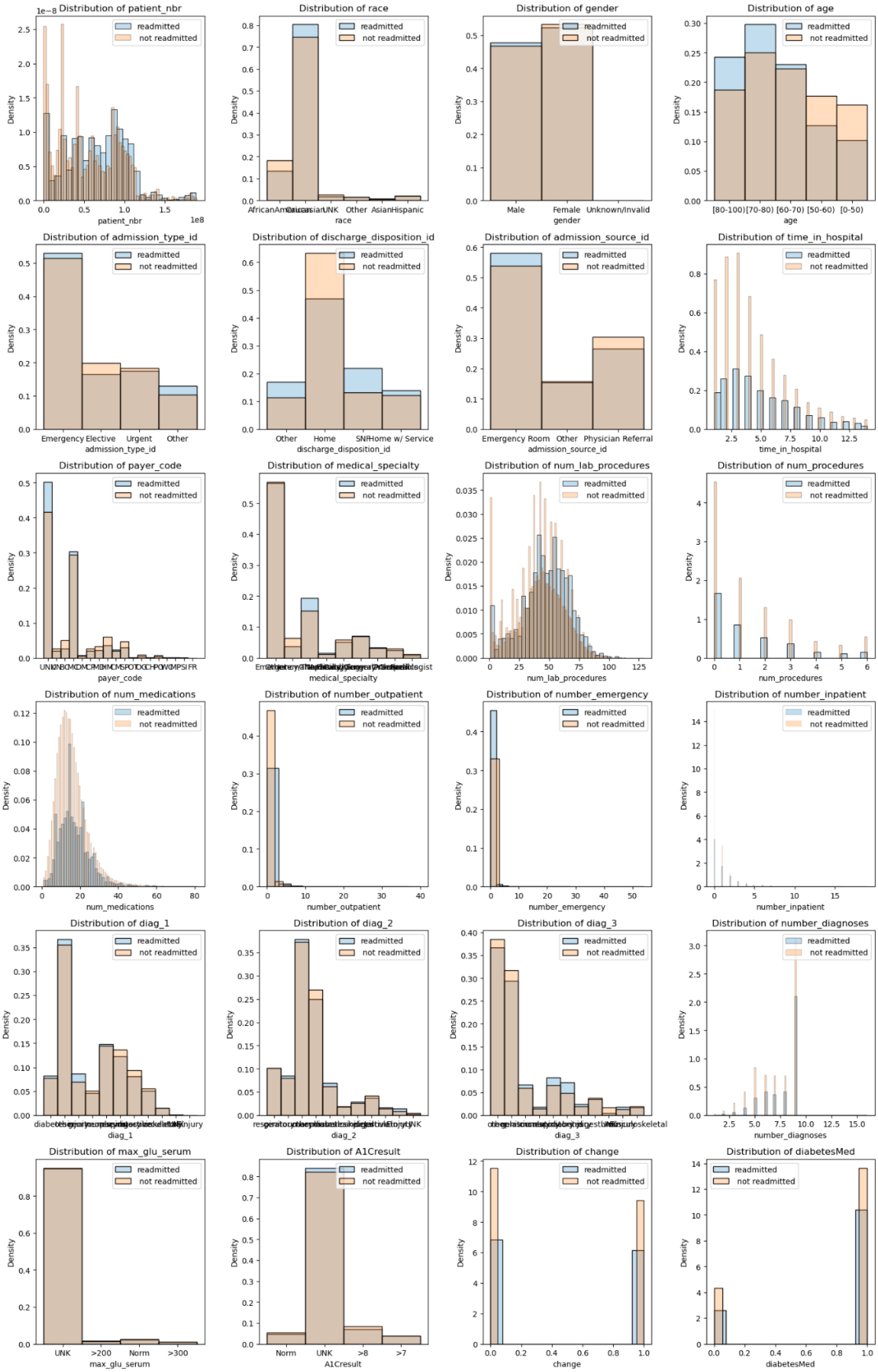
Features with float64 data types:

```
'avg_time_in_hospital', 'avg_num_lab_procedures', 'avg_num_procedures', 'avg_num_medications',
'mean_diagnoses', 'avg_times_glu_high', 'avg_times_a1c_high', 'avg_times_med_changed',
'avg_times_diabetic_med_prescribed', 'avg_times_readmitted'
```

Features with object data types (drug features):

```
'race', 'gender', 'age', 'admission_type_id', 'discharge_disposition_id',
'admission_source_id', 'payer_code', 'medical_specialty', 'diag_1', 'diag_2', 'diag_3',
'max_glu_serum', 'A1Cresult', 'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide',
'glimepiride', 'acetohexamide', 'glipizide', 'glyburide', 'tolbutamide', 'pioglitazone',
'rosiglitazone', 'acarbose', 'miglitol', 'troglitazone', 'tolazamide', 'examide', 'citoglipton',
'insulin', 'glyburide-metformin', 'glipizide-metformin', 'glimepiride-pioglitazone',
'metformin-rosiglitazone', 'metformin-pioglitazone'
```


Distribution of metformin
readmitted
not readmitted
Density
No
Steady
Down
Up
metformin
Distribution of repaglinide
repaglinide
Distribution of nateglinide
nateglinide
Distribution of chlorpropamide
chlorpropamide
Distribution of glimepiride
glimepiride
Distribution of acetohexamide
acetohexamide
Distribution of glipizide
glipizide
Distribution of glyburide
glyburide
Distribution of tolbutamide
tolbutamide
Distribution of pioglitazone
pioglitazone
Distribution of rosiglitazone
rosiglitazone
Distribution of acarbose
acarbose
Distribution of miglitol
miglitol
Distribution of troglitazone
troglitazone
Distribution of tolazamide
tolazamide
Distribution of examide
examide
Distribution of citoglipton
citoglipton
Distribution of insulin
insulin
Distribution of glyburide-metformin
glyburide-metformin
Distribution of glipizide-metformin
glipizide-metformin
Distribution of glimepiride-pioglitazone
glimepiride-pioglitazone
Distribution of metformin-rosiglitazone
metformin-rosiglitazone
Distribution of metformin-pioglitazone
metformin-pioglitazone


Distribution of patient_nbr
Distribution of race
Distribution of gender
Distribution of age
Distribution of admission_type_id
Distribution of discharge_disposition_id
Distribution of admission_source_id
Distribution of time_in_hospital
Distribution of payer_code
Distribution of medical_specialty
Distribution of num_lab_procedures
Distribution of num_procedures
Distribution of num_medications
Distribution of number_outpatient
Distribution of number_emergency
Distribution of number_inpatient
Distribution of diag_1
Distribution of diag_2
Distribution of diag_3
Distribution of number_diagnoses
Distribution of max_glu_serum
Distribution of A1Cresult
Distribution of change
Distribution of diabetesMed
readmitted
not readmitted
Density