ESILV LÉONARD DE VINCI | ÉCOLE D'INGÉNIEURS PARIS-LA DÉFENSE

Final Project
Python Data Analysis

Diabetes 130-US hospitals for years 1999-2008

Alexandre Grosse - 5 January 2022

Introduction

- Hospitalized patients with diabetes are at higher risk of readmission than those without diabetes. Therefore, reducing readmission rates for diabetic patients has a great potential to reduce medical cost significantly.

- The objective of this project is to **predict the likelihood of a diabetic patient being readmitted**.

- The dataset was obtained from the Center for Machine Learning and Intelligent Systems at University of California, Irvine, and contains over **100,000 attributes and 50 features**.
- It represents **10 years** (from 1999 to 2008) of clinical care at 130 US hospitals and integrated delivery networks.

The attributes

The data contains 101,766 instances and 50 attributes, such as:

- Patient number
- Race
- Gender and age range
- Admission type
- Time spent in the hospital
- Diabetic medications
- HbA1c test result
- And many others: diagnosis, number of medication, number of outpatient, inpatient, emergency visits in the year before the hospitalization, etc.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 101766 entries, 0 to 101765
Data columns (total 50 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   encounter_id              101766 non-null  int64
 1   patient_nbr               101766 non-null  int64
 2   race                      101766 non-null  object
 3   gender                    101766 non-null  object
 4   age                       101766 non-null  object
 5   weight                    101766 non-null  object
 6   admission_type_id         101766 non-null  int64
 7   discharge_disposition_id  101766 non-null  int64
 8   admission_source_id       101766 non-null  int64
 9   time_in_hospital          101766 non-null  int64
 10  payer_code                101766 non-null  object
 11  medical_specialty         101766 non-null  object
 12  num_lab_procedures        101766 non-null  int64
 13  num_procedures            101766 non-null  int64
 14  num_medications           101766 non-null  int64
 15  number_outpatient         101766 non-null  int64
 16  number_emergency          101766 non-null  int64
 17  number_inpatient          101766 non-null  int64
 18  diag_1                    101766 non-null  object
 19  diag_2                    101766 non-null  object
 20  diag_3                    101766 non-null  object
 21  number_diagnoses          101766 non-null  int64
 22  max_glu_serum             101766 non-null  object
 23  A1Cresult                 101766 non-null  object
 24  metformin                 101766 non-null  object
 25  repaglinide               101766 non-null  object
 26  nateglinide               101766 non-null  object
 27  chlorpropamide            101766 non-null  object
 28  glimepiride               101766 non-null  object
 29  acetohexamide             101766 non-null  object
 30  glipizide                 101766 non-null  object
```

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), …, [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%

Feature name	Type	Description and values	% missing
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
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. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.	0%

The target:

- It is important to know if a patient will be readmitted in some hospital in order to know if it's a better choice to change the treatment, to avoid a possible readmission.
- In this database, we can have 3 different outputs:
  1. **No readmission** (seems to be a good situation);
  2. **A readmission in less than 30 days** (this situation is not good, because maybe the treatment was not appropriate);
  3. **A readmission in more than 30 days** (this one is not so good as well the last one, however, the reason can be the state of the patient.

readmitted
>30
NO
NO
NO
>30
...

We can merge the readmission after 30 days and no readmission into a single outcome

Cleaning data

- **Missing data**: we drop the columns which have a lot of missing data and seem to be irrelevant in order to predict our target

- We drop **weight** (missing in over 96% values), **payer_code** and **medical_specialty** (40% and 49% missing)

- Now we can remove all the rows with missing data: we lost only 3,713 rows, 3.65% of the dataset.

Feature engineering

- We remove the patients who died during the admission, because they obviously can’t be readmitted
- We also remove the **encounter_id** column because some patients in the dataset may had more than one encounter, and we remove the duplicates with the **patient_nbr** column
- The binary qualitative columns are transformed into 0 and 1.
- We regrouped the admission types, the discharge dispositions and the admission source into larger categories, which make more sense.

Categories for admission_type, discharge_disposition and admission_source

```
Emergency          46529
Elective           13318
Other               7700
Trauma Center         17
Newborn                9
Name: admission_type, dtype: int64
```

```
Discharged to home    42230
Other                 25343
Name: discharge_disposition, dtype: int64
```

```
Emergency Room               36429
Physician Referral           21869
Other                         4918
Transfer from a hospital      4347
Court/Law Enforcement           10
Name: admission_source, dtype: int64
```

Categories for diagnoses

- We did the same for the 3 diagnoses columns: example of repartition for the diag_1 column

```
Diseases Of The Circulatory System                                                  20800
Diseases Of The Respiratory System                                                   6310
Diseases Of The Digestive System                                                     6121
Symptoms, Signs, And Ill-Defined Conditions                                          5338
Diabetes mellitus                                                                    5149
Injury And Poisoning                                                                 4496
Diseases Of The Musculoskeletal System And Connective Tissue                         3879
Diseases Of The Genitourinary System                                                 3355
Neoplasms                                                                            2555
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders without diabetes  1777
Diseases Of The Skin And Subcutaneous Tissue                                         1711
Infectious And Parasitic Diseases                                                    1650
Mental Disorders                                                                     1456
Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services  895
Diseases Of The Nervous System And Sense Organs                                       838
Diseases Of The Blood And Blood-Forming Organs                                        643
Complications Of Pregnancy, Childbirth, And The Puerperium                            559
Congential Anomalies                                                                   40
Supplementary Classification Of External Causes Of Injury And Poisoning                 1
Name: diag_1, dtype: int64
```

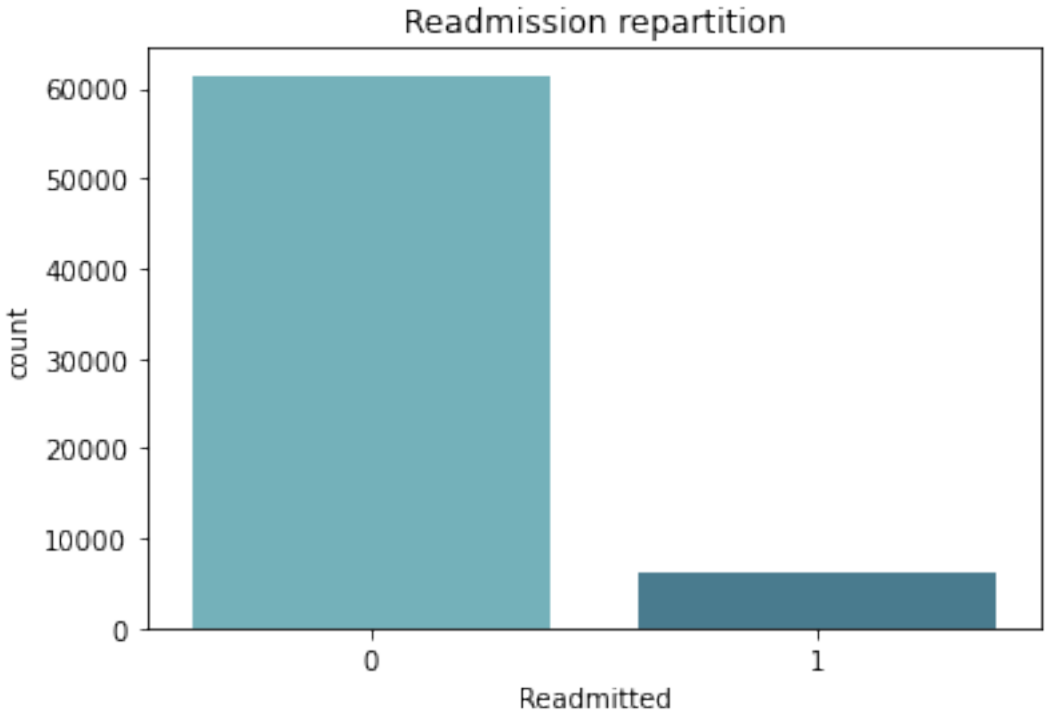
Medications features

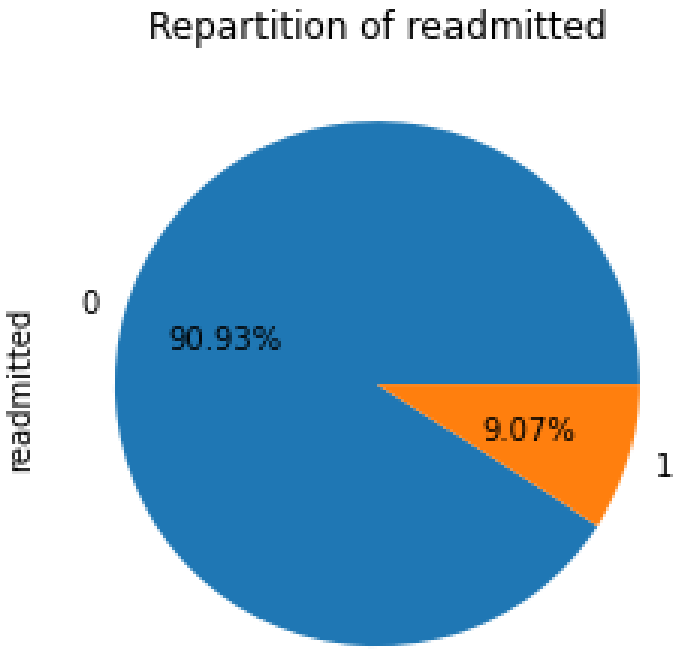
- Some of the medications columns don't give important informations: we can see some of them have an insignficant number of other values compared to the number of "No" values. We can remove those columns.

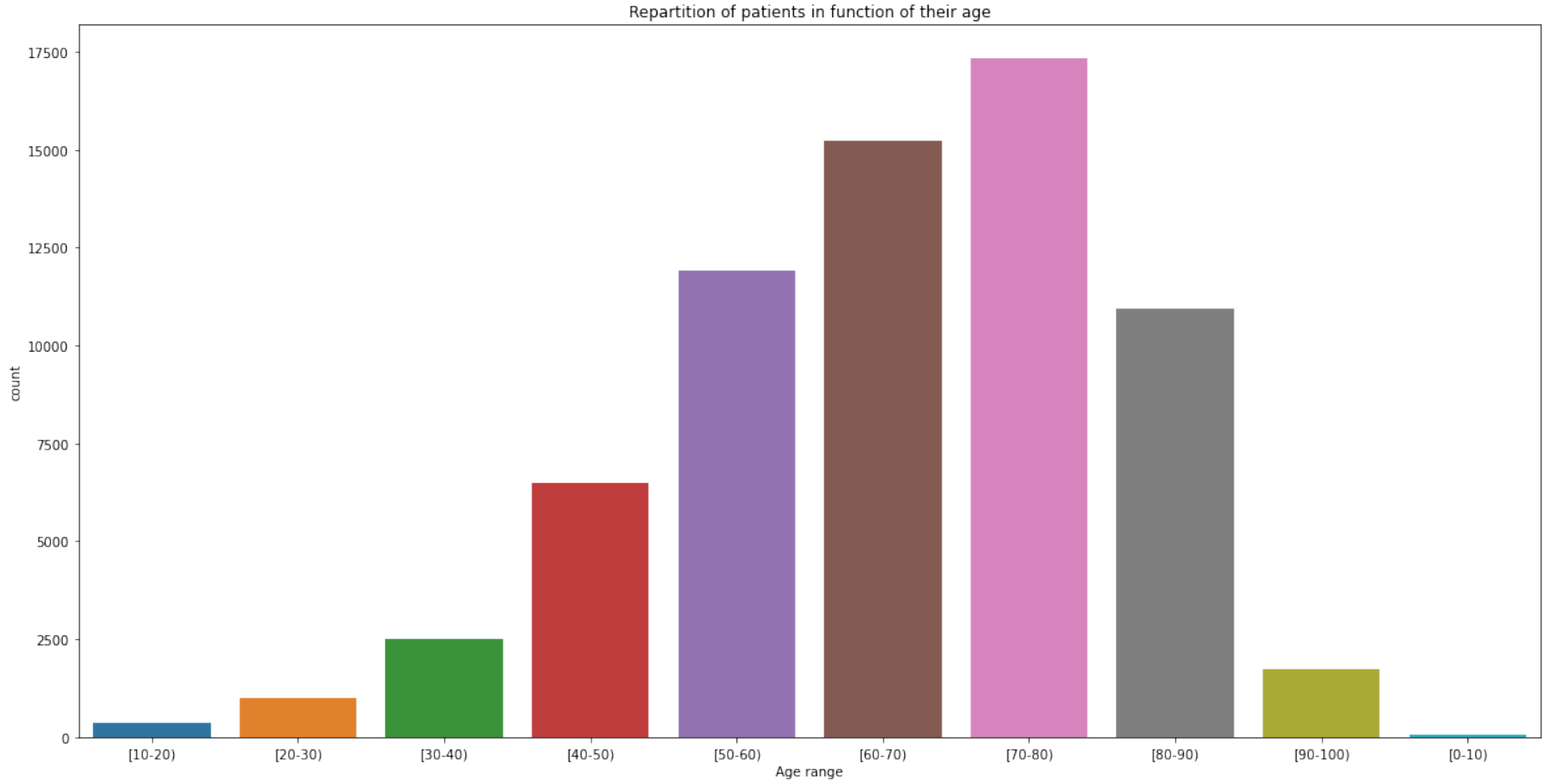
- After recoding all those features, we are down to 67,573 rows and 37 features

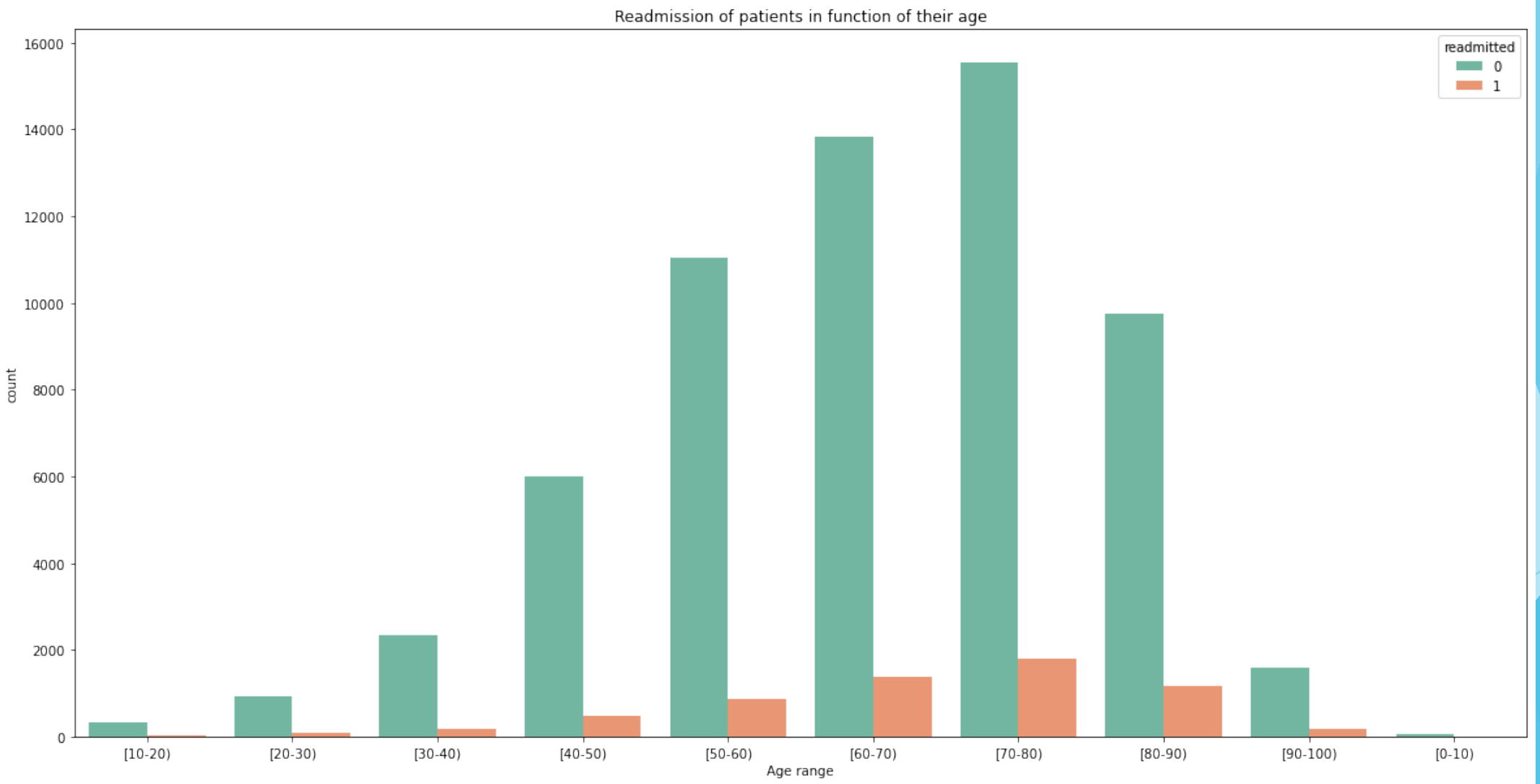
```
Entrée [50]: dataset.shape

 Out[50]: (67573, 37)
```

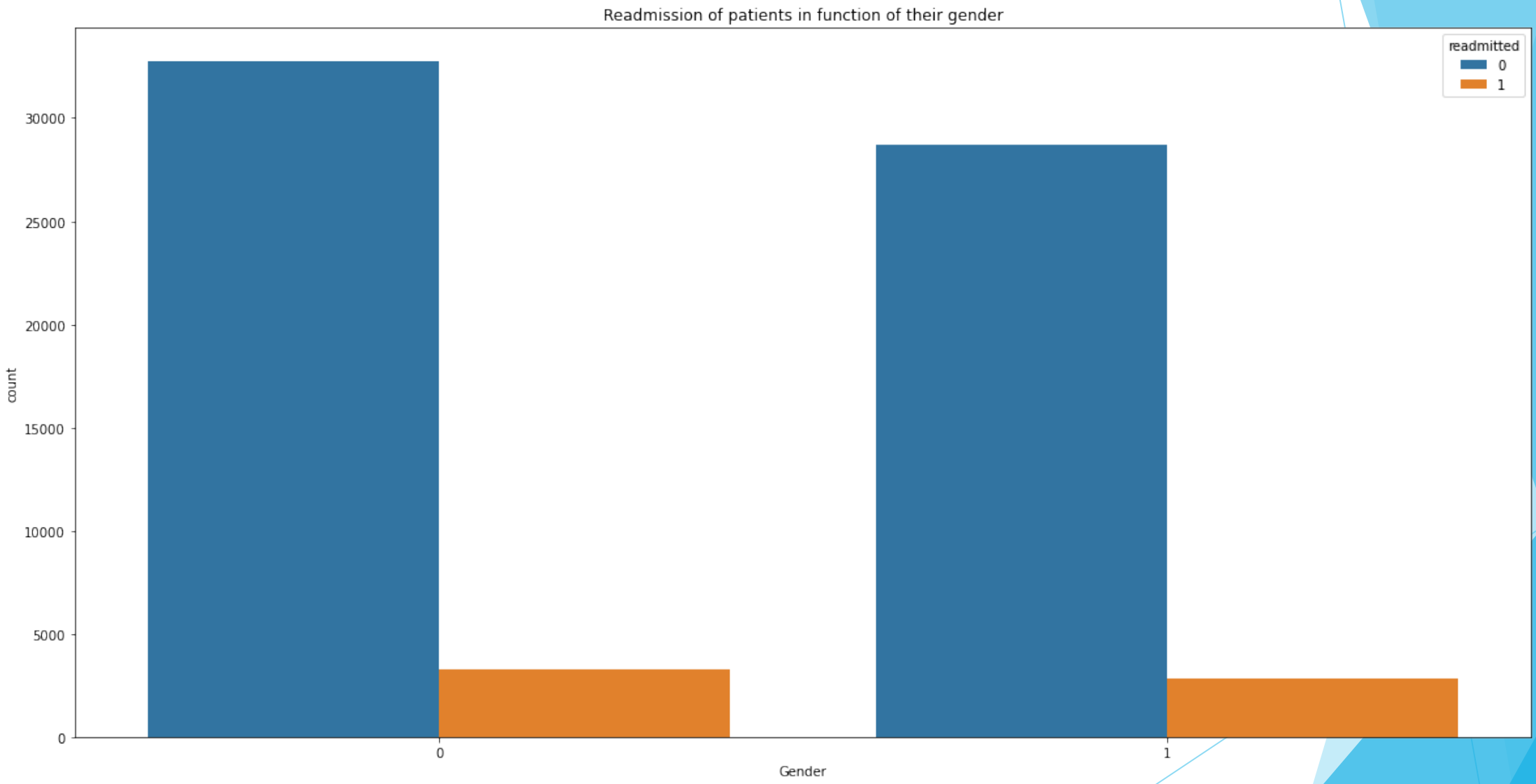
Data visualization

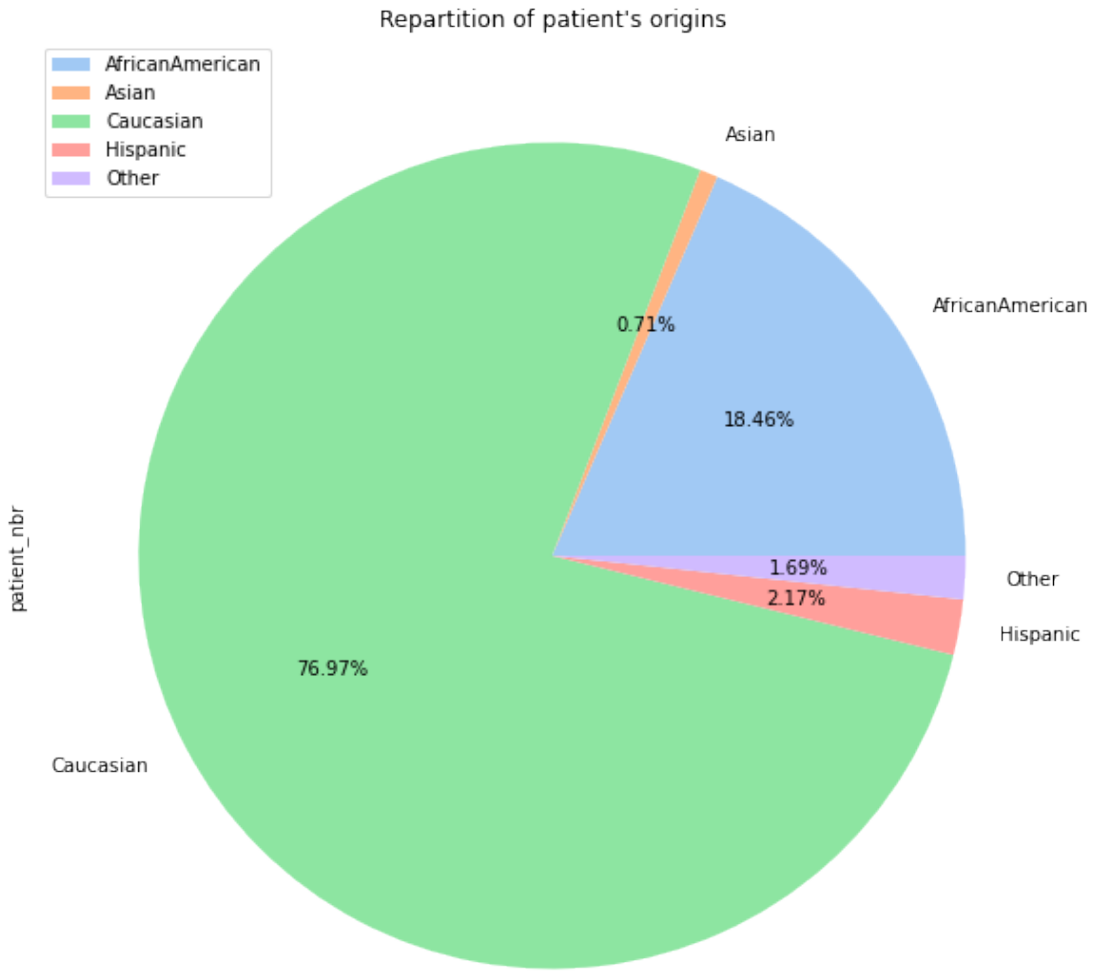


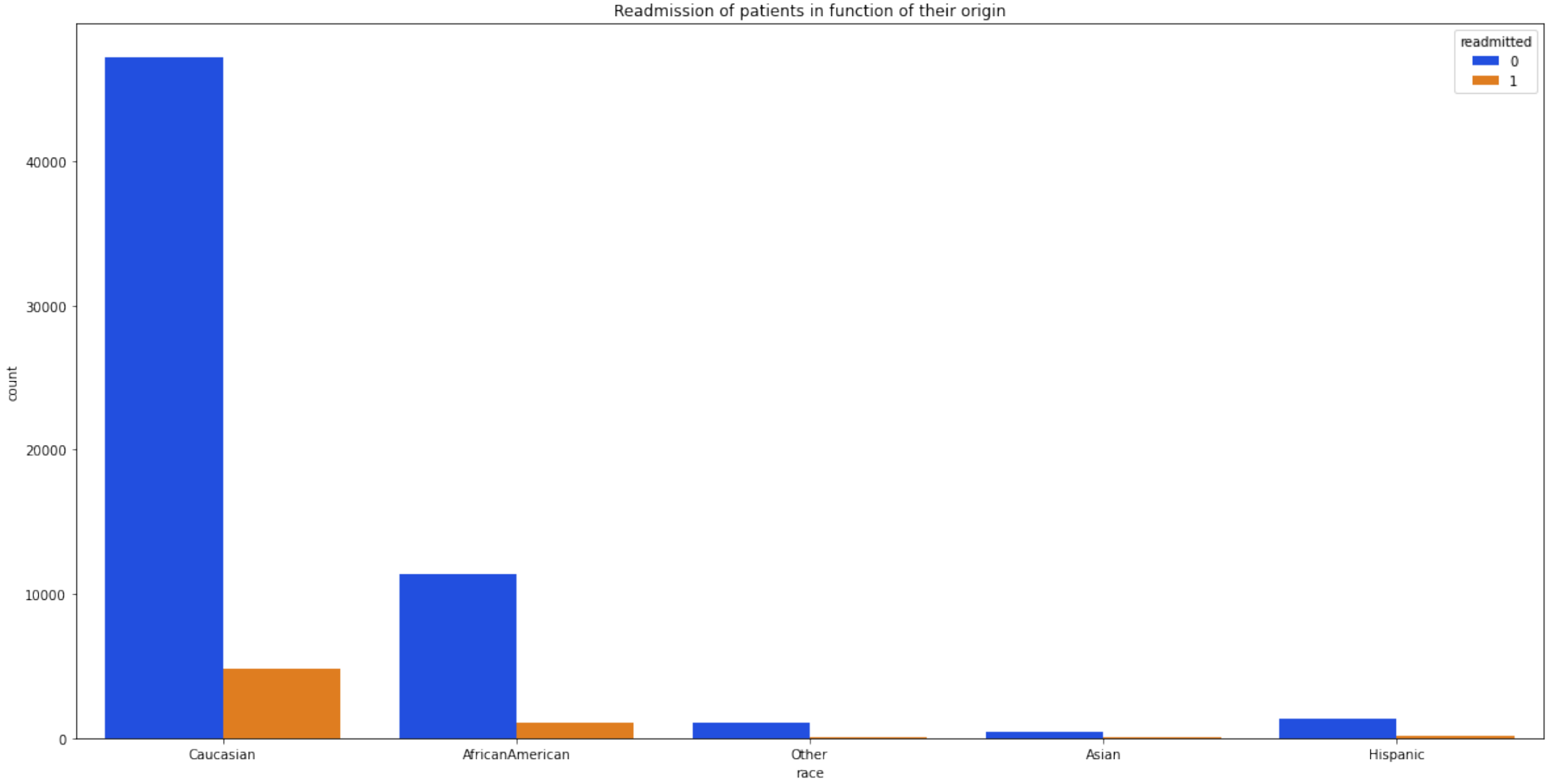


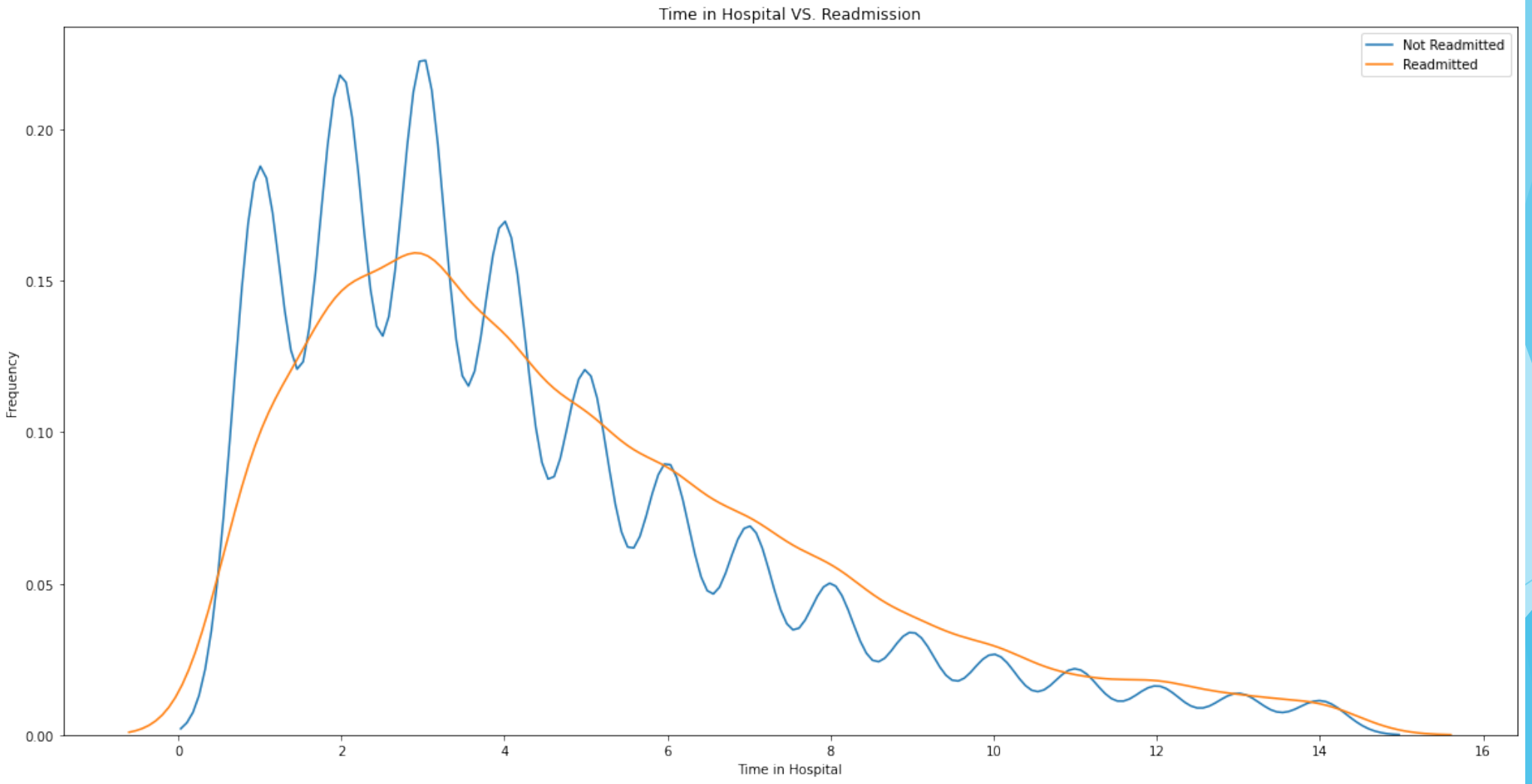

Repartition of patients in function of their age
count
17500
15000
12500
10000
7500
5000
2500
0
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
[0-10)
Age range

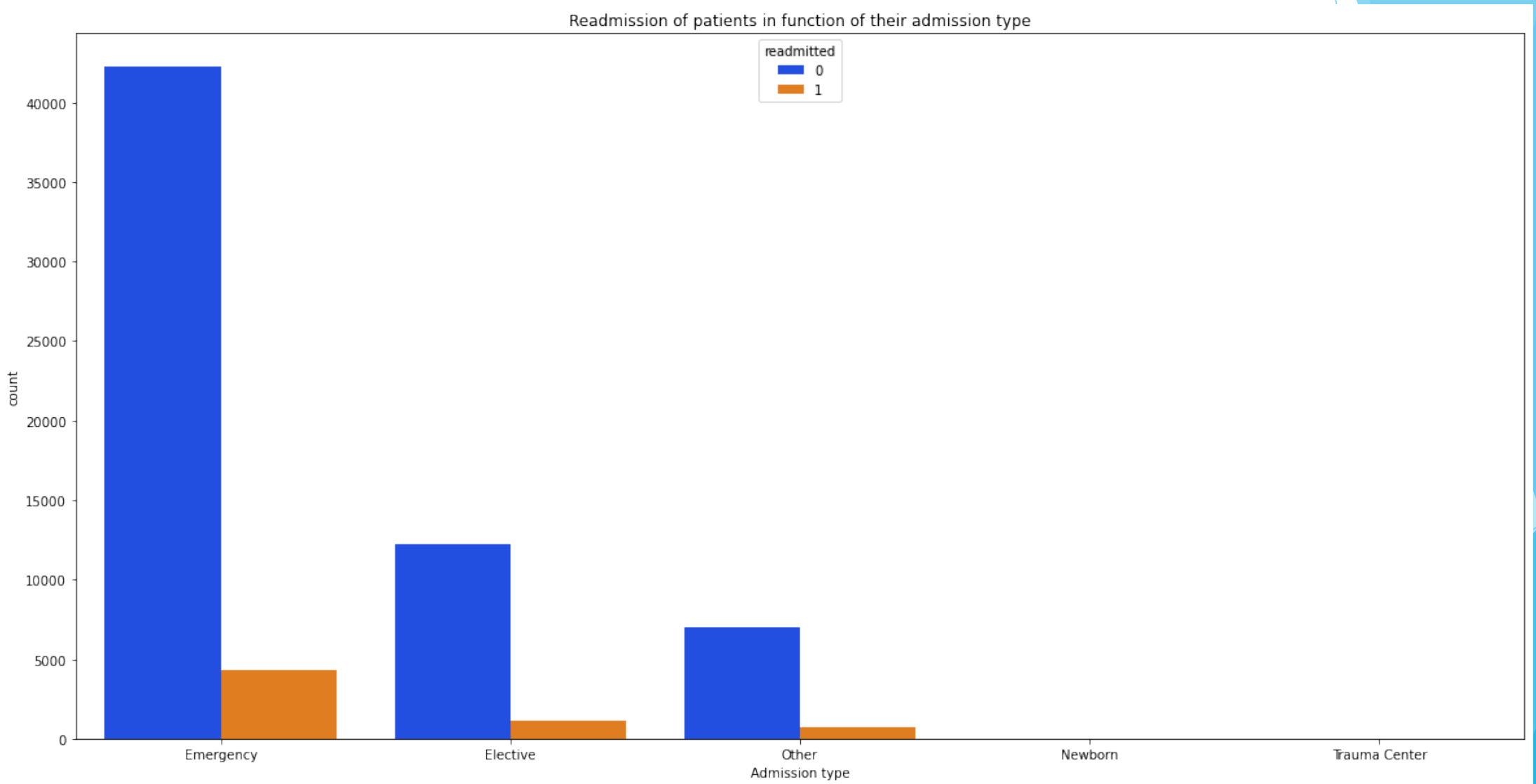

Readmission of patients in function of their age
readmitted
0
1
count
16000
14000
12000
10000
8000
6000
4000
2000
0
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
[0-10)
Age range

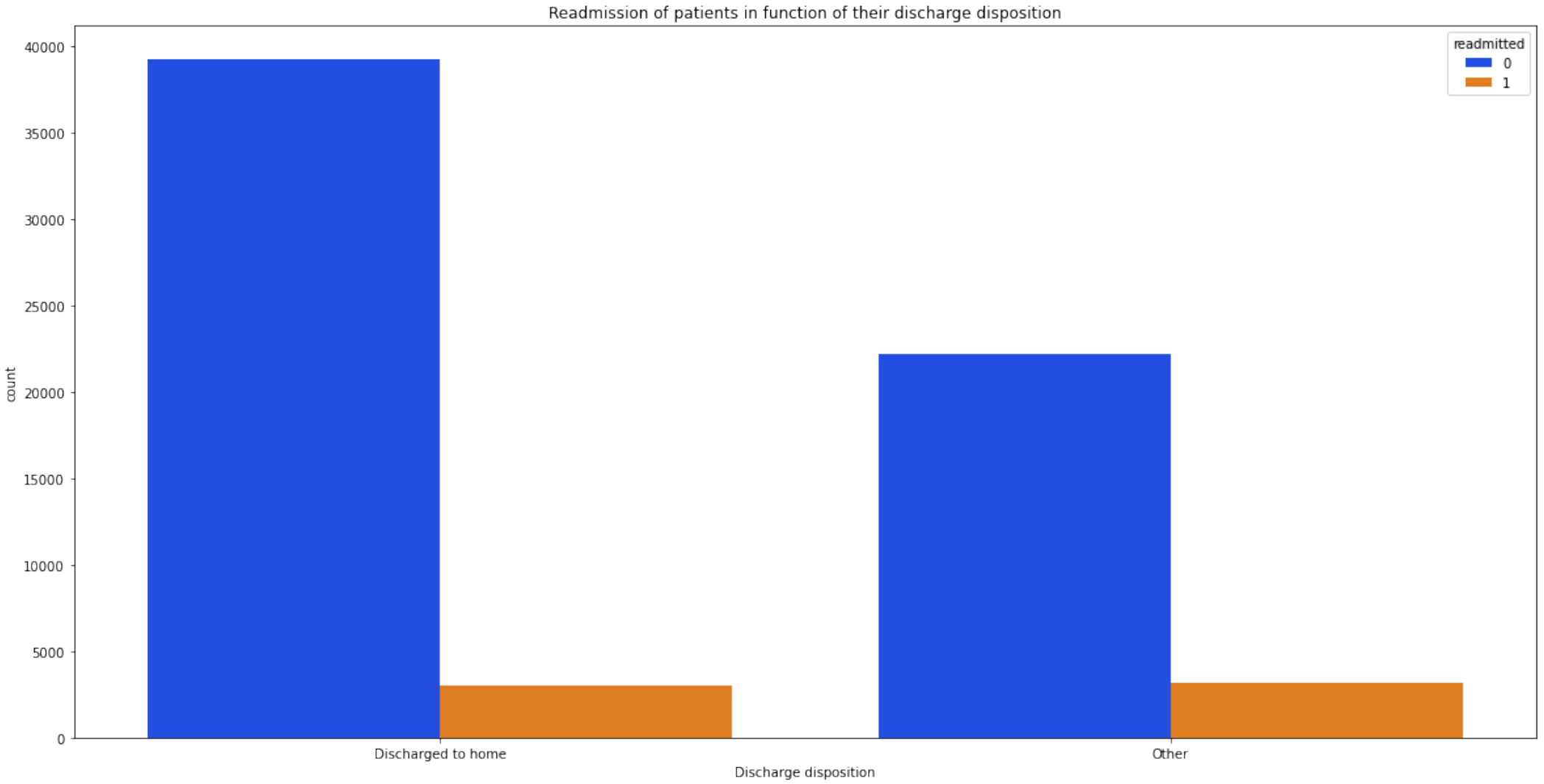

Readmission of patients in function of their gender
readmitted
0
1
30000
25000
20000
15000
10000
5000
0
count
0
1
Gender

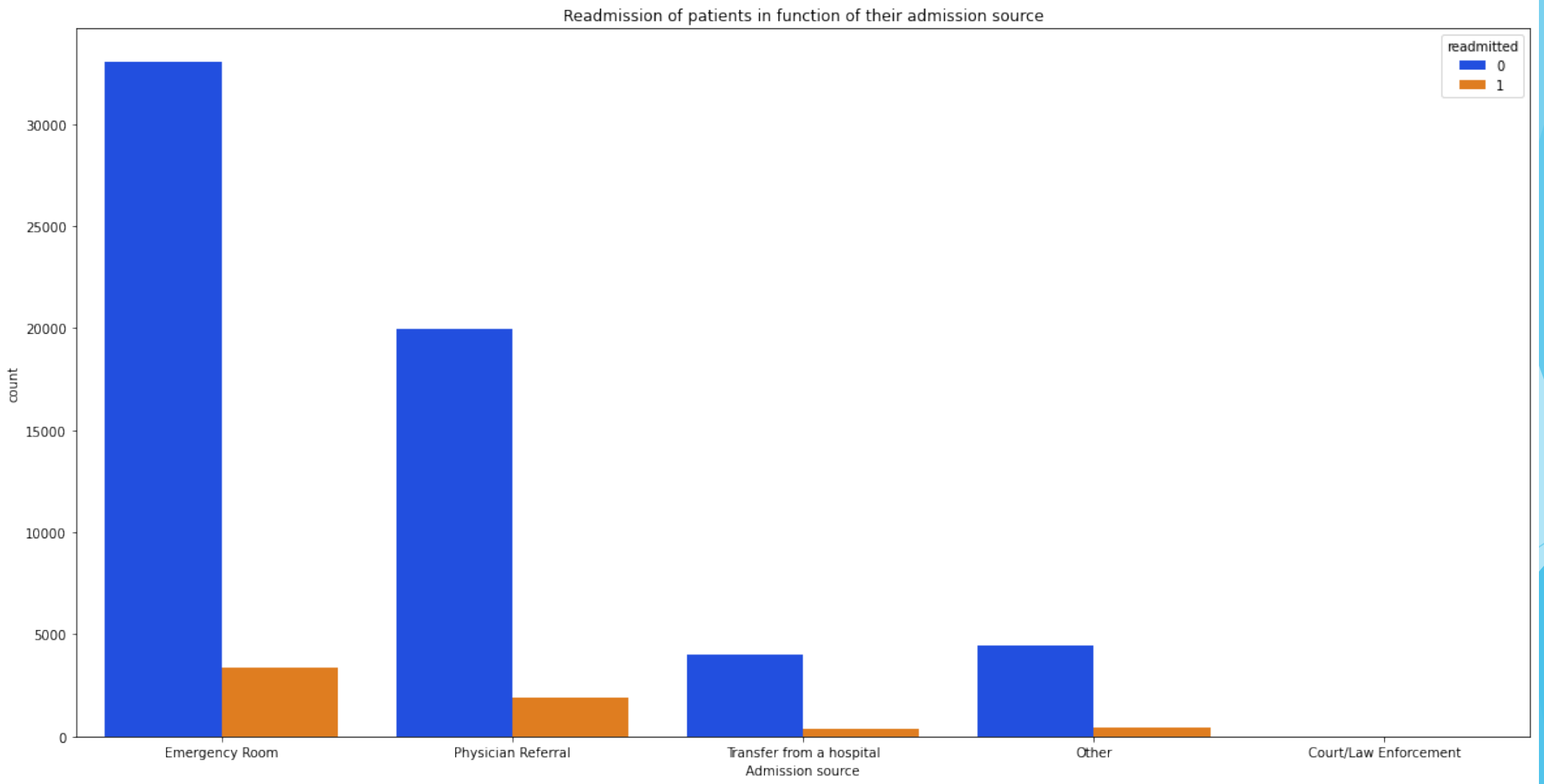

Repartition of patient's origins
AfricanAmerican
Asian
Caucasian
Hispanic
Other
patient_nbr
0.71%
18.46%
1.69%
2.17%
76.97%
Asian
AfricanAmerican
Other
Hispanic
Caucasian

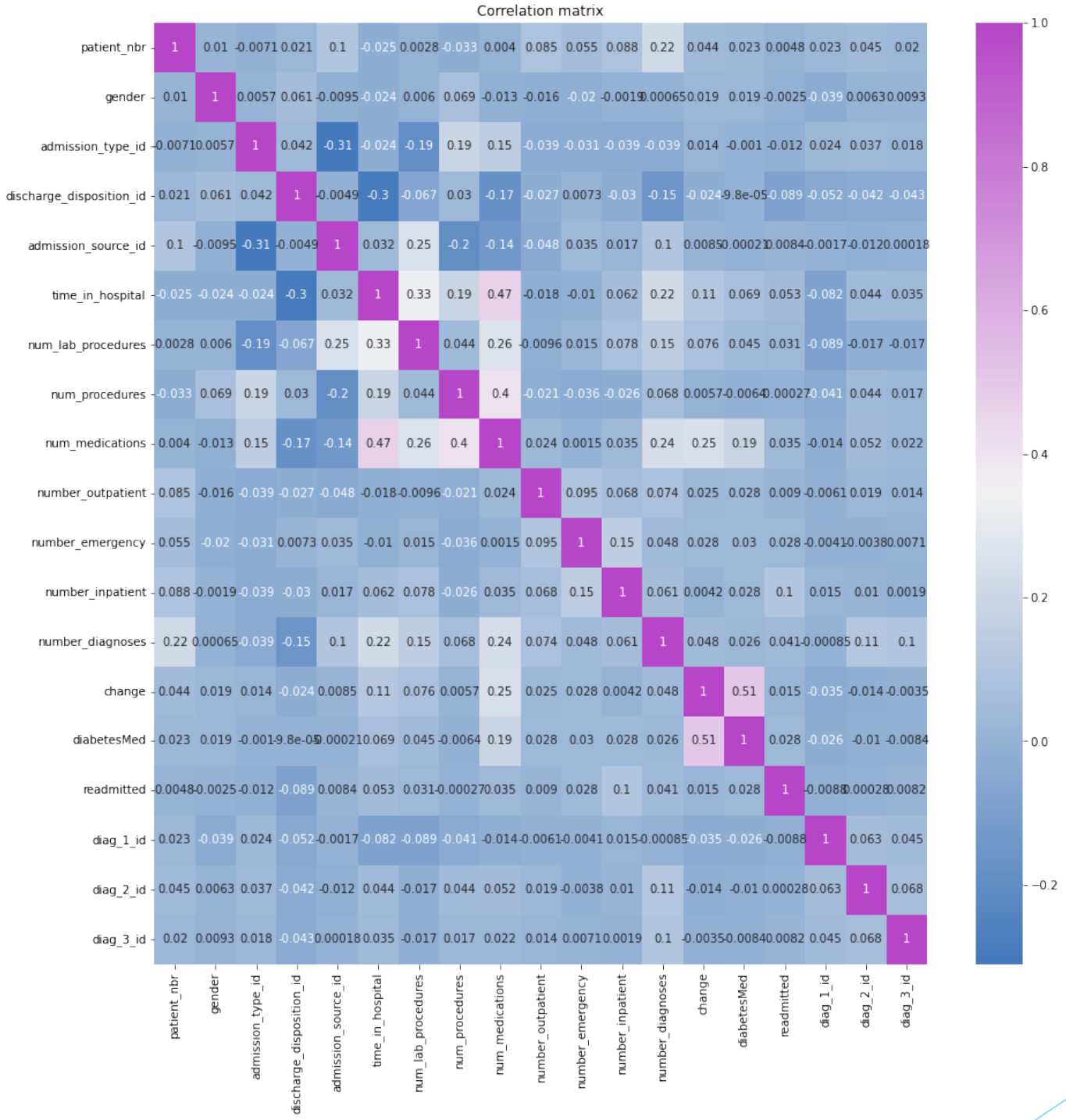

Readmission of patients in function of their origin
readmitted
0
1
count
40000
30000
20000
10000
0
Caucasian
AfricanAmerican
Other
Asian
Hispanic
race


Time in Hospital VS. Readmission
Not Readmitted
Readmitted
Frequency
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
14
16
Time in Hospital


Readmission of patients in function of their admission type
readmitted
0
1
count
40000
35000
30000
25000
20000
15000
10000
5000
0
Emergency
Elective
Other
Newborn
Trauma Center
Admission type


Readmission of patients in function of their discharge disposition
readmitted
0
1
count
40000
35000
30000
25000
20000
15000
10000
5000
0
Discharged to home
Other
Discharge disposition


Readmission of patients in function of their admission source
readmitted
0
1
count
30000
25000
20000
15000
10000
5000
0
Emergency Room
Physician Referral
Transfer from a hospital
Other
Court/Law Enforcement
Admission source


Correlation matrix
patient_nbr 1 0.01 -0.0071 0.021 0.1 -0.025 0.0028 -0.033 0.004 0.085 0.055 0.088 0.22 0.044 0.023 0.0048 0.023 0.045 0.02
gender 0.01 1 0.0057 0.061 -0.0095 -0.024 0.006 0.069 -0.013 -0.016 -0.02 -0.0019 0.00065 0.019 0.019 -0.0025 -0.039 0.0063 0.0093
admission_type_id -0.0071 0.0057 1 0.042 -0.31 -0.024 -0.19 0.19 0.15 -0.039 -0.031 -0.039 -0.039 0.014 -0.001 -0.012 0.024 0.037 0.018
discharge_disposition_id 0.021 0.061 0.042 1 -0.0049 -0.3 -0.067 0.03 -0.17 -0.027 0.0073 -0.03 -0.15 -0.024 9.8e-05 -0.089 -0.052 -0.042 -0.043
admission_source_id 0.1 -0.0095 -0.31 -0.0049 1 0.032 0.25 -0.2 -0.14 -0.048 0.035 0.017 0.1 0.0085 0.00021 0.0084 -0.0017 -0.012 0.00018
time_in_hospital -0.025 -0.024 -0.024 -0.3 0.032 1 0.33 0.19 0.47 -0.018 -0.01 0.062 0.22 0.11 0.069 0.053 -0.082 0.044 0.035
num_lab_procedures 0.0028 0.006 -0.19 -0.067 0.25 0.33 1 0.044 0.26 -0.0096 0.015 0.078 0.15 0.076 0.045 0.031 -0.089 -0.017 -0.017
num_procedures -0.033 0.069 0.19 0.03 -0.2 0.19 0.044 1 0.4 -0.021 -0.036 -0.026 0.068 0.0057 -0.0064 -0.00027 0.041 0.044 0.017
num_medications 0.004 -0.013 0.15 -0.17 -0.14 0.47 0.26 0.4 1 0.024 0.0015 0.035 0.24 0.25 0.19 0.035 -0.014 0.052 0.022
number_outpatient 0.085 -0.016 -0.039 -0.027 -0.048 -0.018 -0.0096 -0.021 0.024 1 0.095 0.068 0.074 0.025 0.028 0.009 -0.0061 0.019 0.014
number_emergency 0.055 -0.02 -0.031 0.0073 0.035 -0.01 0.015 -0.036 0.0015 0.095 1 0.15 0.048 0.028 0.03 0.028 -0.0041 -0.0038 0.0071
number_inpatient 0.088 -0.0019 -0.039 -0.03 0.017 0.062 0.078 -0.026 0.035 0.068 0.15 1 0.061 0.0042 0.028 0.1 0.015 0.01 0.0019
number_diagnoses 0.22 0.00065 -0.039 -0.15 0.1 0.22 0.15 0.068 0.24 0.074 0.048 0.061 1 0.048 0.026 0.041 -0.00085 0.11 0.1
change 0.044 0.019 0.014 -0.024 0.0085 0.11 0.076 0.0057 0.25 0.025 0.028 0.0042 0.048 1 0.51 0.015 -0.035 -0.014 -0.0035
diabetesMed 0.023 0.019 -0.001 9.8e-05 0.00021 0.069 0.045 -0.0064 0.19 0.028 0.03 0.028 0.026 0.51 1 0.028 -0.026 -0.01 -0.0084
readmitted 0.0048 -0.0025 -0.012 -0.089 0.0084 0.053 0.031 -0.00027 0.035 0.009 0.028 0.1 0.041 0.015 0.028 1 -0.0088 0.00028 0.0082
diag_1_id 0.023 -0.039 0.024 -0.052 -0.0017 -0.082 -0.089 -0.041 -0.014 -0.0061 -0.0041 0.015 -0.00085 -0.035 -0.026 -0.0088 1 0.063 0.045
diag_2_id 0.045 0.0063 0.037 -0.042 -0.012 0.044 -0.017 0.044 0.052 0.019 -0.0038 0.01 0.11 -0.014 -0.01 0.00028 0.063 1 0.068
diag_3_id 0.02 0.0093 0.018 -0.043 0.00018 0.035 -0.017 0.017 0.022 0.014 0.0071 0.0019 0.1 -0.0035 -0.0084 0.0082 0.045 0.068 1
patient_nbr
gender
admission_type_id
discharge_disposition_id
admission_source_id
time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_outpatient
number_emergency
number_inpatient
number_diagnoses
change
diabetesMed
readmitted
diag_1_id
diag_2_id
diag_3_id
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Modeling

- We tested the logistic regression model, the decision tree and the random forest model

- We used different techniques to avoid the problems caused by imbalanced data, with reasmpling techniques

- The best model is the random forest, with

  - Accuracy : 0.96
  - Precision: 0.99
  - Recall: 0.92
  - F1-score: 0.96

Bibliography

- The table with all the features and their descriptions: https://www.hindawi.com/journals/bmri/2014/781670/tab1/

- The 2008 ICD-9-CM Diagnosis Codes: http://www.icd9data.com/2008/Volume1/default.htm https://www.cdc.gov/nchs/icd/icd9cm.htm

- Dealing with Imbalanced Data: https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18