from mpl toolkits.mplot3d import Axes3D

In [96]:

```
from sklearn.preprocessing import StandardScaler
                                  import matplotlib.pyplot as plt
                                  import numpy as np
                                  import pandas as pd
                                  import seaborn as sns
                                  import pandas as pd
                                  import numpy as np
In [97]:
                                 df= pd.read csv("./diabetic data.csv")
                                 df.head(2)
Out[97]:
                                        encounter_id patient_nbr
                                                                                                                            race gender age weight admission_type_id discharge_disposition_i
                                                                                                                                                                      [0-
                              0
                                                    2278392
                                                                                     8222157 Caucasian
                                                                                                                                            Female
                                                                                                                                                                                                 ?
                                                                                                                                                                                                                                                   6
                                                                                                                                                                      10)
                                                                                                                                                                    [10-
                                                       149190
                                                                                   55629189 Caucasian Female
                                                                                                                                                                      20)
                            2 rows × 50 columns
In [98]:
                                 df.shape
Out[98]: (101766, 50)
In [99]:
                                 df.columns
Out[99]: Index(['encounter_id', 'patient_nbr', 'race', 'gender', 'age', 'weight', 'admission_type_id', 'discharge_disposition_id', 'admission_source_id',
                                                    'time_in_hospital', 'payer_code', 'medical_specialty',
'num_lab_procedures', 'num_procedures', 'num_medications',
'number_outpatient', 'number_emergency', 'number_inpatient', 'diag_1',
'diag_2', 'diag_3', 'number_diagnoses', 'max_glu_serum', 'A1Cresult',
                                                    'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide', 'glimepiride', 'acetohexamide', 'glipizide', 'glyburide', 'tolbutamide', 'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol', 'troglitazone', 'tolazamide', 'examide', 'citoglipton', 'insulin', 'tolbutamide', 'acarbose', 'miglitol', 'troglitazone', 'tolazamide', 'examide', 'citoglipton', 'insulin', 'tolbutamide', 'acarbose', 'miglitol', 'troglitazone', 'tolazamide', 'citoglipton', 'insulin', 'tolazamide', 'acarbose', 'miglitol', 'troglitazone', 'tolazamide', 'citoglipton', 'insulin', 'tolazamide', 'tolazamide', 'citoglipton', 'insulin', 'tolazamide', 'citoglipton', 'insulin', 'tolazamide', 'tolazamide', 'citoglipton', 'insulin', 'tolazamide', '
                                                     'glyburide-metformin', 'glipizide-metformin',
                                                     'glimepiride-pioglitazone', 'metformin-rosiglitazone',
                                                     'metformin-pioglitazone', 'change', 'diabetesMed', 'readmitted'],
                                                  dtype='object')
In [100...
                                 df=df.replace("?",np.nan)
In [101...
                                 #dropping columns with large number of missing values
                                  df = df.drop(['weight','payer_code','medical_specialty'], axis = 1)
```

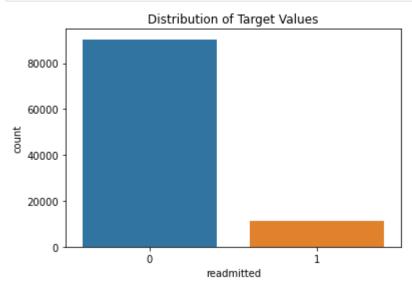
```
#variables (drugs named citoglipton and examide), all records have the same value
In [102...
           df = df.drop(['citoglipton', 'examide'], axis = 1)
In [103...
          #Analyzing race column
          df['race'].isnull().sum()
          df['race'].value_counts()
          df['race']=df['race'].fillna('UNK') #filling null with unk
In [104...
          df['readmitted'].value_counts()
         NO
                 54864
Out[104...
          >30
                 35545
                 11357
          <30
          Name: readmitted, dtype: int64
In [105...
          #Feature engineering
           #Generating output variable
          #we need to check whether a patient admitted within 30 days or not
          df['target']=(df['readmitted']=='<30').astype('int')</pre>
           #dropping readmitted column
           df.drop(['readmitted'],axis=1,inplace=True)
          df['readmitted']=df['target']
           df.drop(['target'],axis=1,inplace=True)
In [106...
          df['readmitted'].value_counts()
               90409
Out[106...
               11357
          Name: readmitted, dtype: int64
In [107...
           cleanup age = {"age":
                                    {"[0-10)": 5, "[10-20)": 15,"[20-30)": 25,"[30-40)": 35,"[40-
               "[60-70)": 65,"[70-80)": 75,"[80-90)": 85,"[90-100)": 95}}
          df.replace(cleanup_age, inplace=True)
In [108...
          #analyzing gender column
           df['gender'].value counts()
           #removing invalid/unknown entries for gender
           df=df[df['gender']!='Unknown/Invalid']
In [109...
           df.head()
Out[109...
             encounter_id patient_nbr
                                              race gender age admission_type_id discharge_disposition_id
          0
                 2278392
                            8222157
                                          Caucasian
                                                             5
                                                                              6
                                                                                                   25
                                                   Female
                  149190
                           55629189
          1
                                          Caucasian
                                                   Female
                                                            15
                                                                              1
                                                                                                    1
          2
                   64410
                           86047875 AfricanAmerican Female
                                                            25
                                                                              1
```

	encounter_id	patient_nbr	race	gender	age	admission_type_id	discharge_disposition_id	
3	500364	82442376	Caucasian	Male	35	1	1	
4	16680	42519267	Caucasian	Male	45	1	1	

5 rows × 45 columns

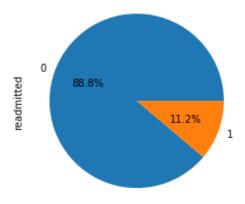
Exploratory Data Analysis

```
sns.countplot(x = "readmitted", data = df)
plt.title("Distribution of Target Values")
plt.show()
```



```
# Pie chart
df.readmitted.value_counts().plot.pie(autopct = "%.1f%%")
plt.title("Proportion of Target Value")
plt.show()
```

Proportion of Target Value



Our target variable is almost balanced distributed.

Race We have 5 different races value, these are;

Caucasian African American Hispanic Asian Other

Number of Race values Outher African American African American Other Asian Asian

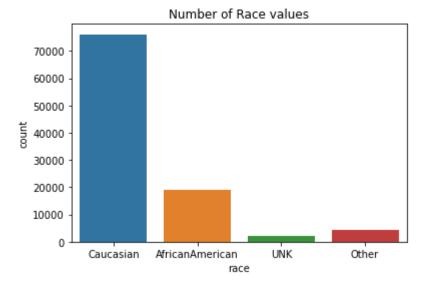
Proportion of Race
Caucasian 74.780618
AfricanAmerican 18.877195
UNK 2.231656
Hispanic 2.001710
Other 1.478927
Asian 0.629895
Name: race, dtype: float64

As we see, there is Caucasians in 76 percent of all our data. And other other 24 percent is divided into African Americans, Hispanics, Asians and Others. Here we decided to divide into 3 groups like Caucasian, African American and Other.

```
In [113...
    mapped_race = {"Asian":"Other","Hispanic":"Other"}
    df.race = df.race.replace(mapped_race)

sns.countplot(x="race", data = df)
    plt.title("Number of Race values")
    plt.show()

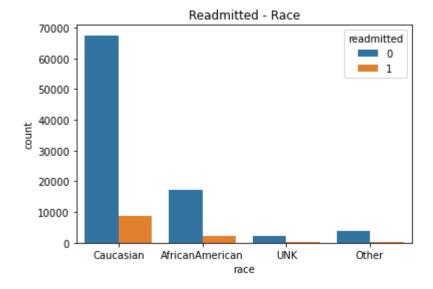
print("Proportion of Race After the Mapping")
    print(df.race.value_counts(normalize= True)*100)
```



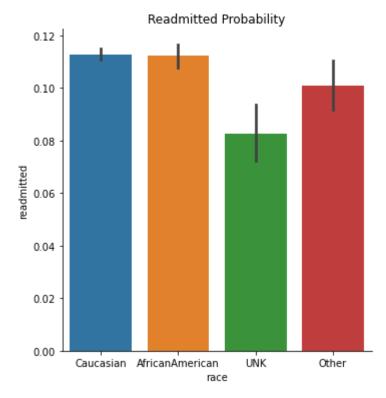
Proportion of Race After the Mapping

Caucasian 74.780618
AfricanAmerican 18.877195
Other 4.110531
UNK 2.231656
Name: race, dtype: float64

```
In [114...
sns.countplot(x="race", hue= "readmitted", data = df)
plt.title("Readmitted - Race")
plt.show()
```



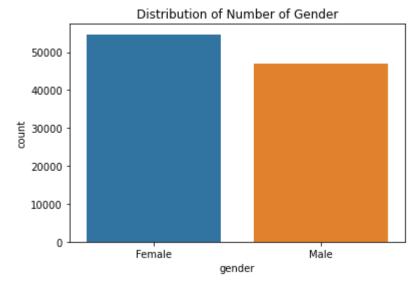
```
sns.catplot(x = "race", y = "readmitted",data = df, kind = "bar", height= 5)
plt.title("Readmitted Probability")
plt.show()
```



Most of the patients are Caucasian, followed by African Americans.

Although the Other values are few than Caucasian, we see that the Readmitted Probability almost close to Caucasian.

Gender



Proportions of Race Value Female 0.537602

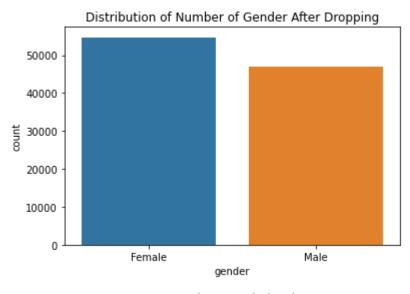
```
Male 0.462398
Name: gender, dtype: float64
```

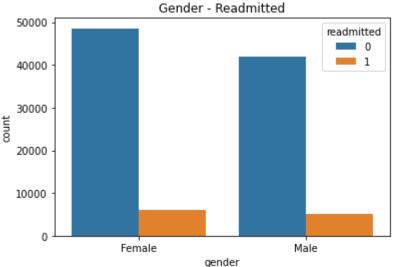
```
In [117...

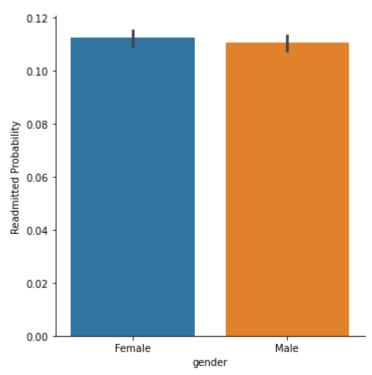
df = df.drop(df.loc[df["gender"]=="Unknown/Invalid"].index, axis=0)

sns.countplot(x = "gender", data = df)
plt.title("Distribution of Number of Gender After Dropping")
plt.show()

sns.countplot(x = "gender", hue = "readmitted", data = df)
plt.title("Gender - Readmitted")
plt.show()
```







We see a nearly equal distribution of Gender.

Also, we can state that Females are a little more prone than Males.

Age

In [120...

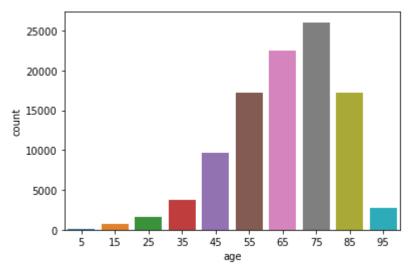
```
In [119...
            sns.countplot(x="age", data = df)
           plt.xticks(rotation = 90)
           plt.show()
             25000
             20000
          ig 15000
             10000
              5000
                 0
                                               55
                                25
                                     æ
                                                     8
                                                          75
                                                               88
                          12
                                            age
```

```
localhost:8888/nbconvert/html/Project/Untitled.ipynb?download=false
```

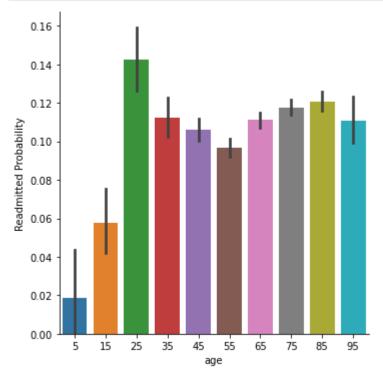
plt.show()

sns.countplot(x="age", data = df)

#plt.xticks(rotation = 90)



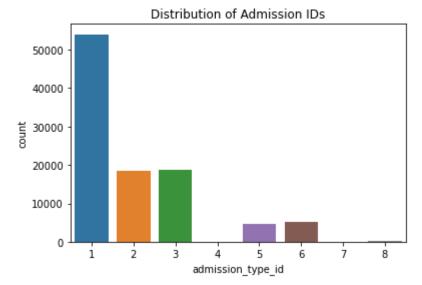
```
g = sns.catplot(x = "age", y = "readmitted", data = df,kind = "bar", height = 5)
g.set_ylabels("Readmitted Probability")
plt.show()
```



we can understand that we have an elderly population.

Weight

As with the age variable, we see the same situation for weight and decided to managed like below.



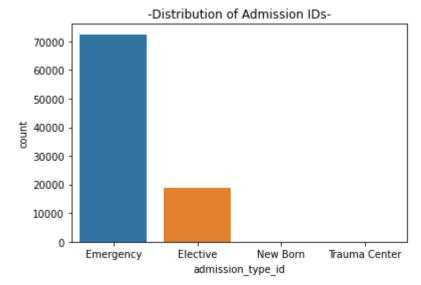
```
Distribution of ID's
1 53988
3 18868
2 18480
6 5291
5 4785
8 320
7 21
4 10
```

Name: admission_type_id, dtype: int64

In here we need to do mapping for:

NULL, Not Available and Not Mapped values.

In addition, we will map Urgent value as Emergency because they have same meaning

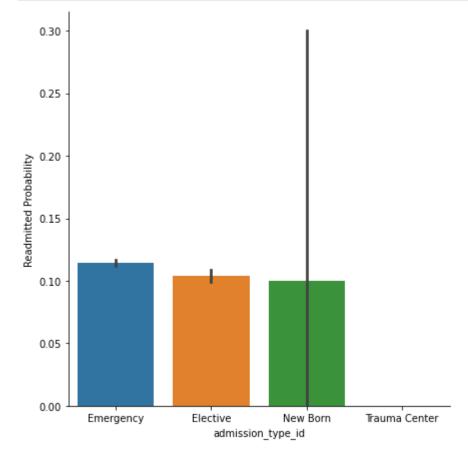


-Distribution of ID's-Emergency 72468 Elective 18868 Trauma Center 21 New Born 10

Name: admission_type_id, dtype: int64

```
In [124...
```

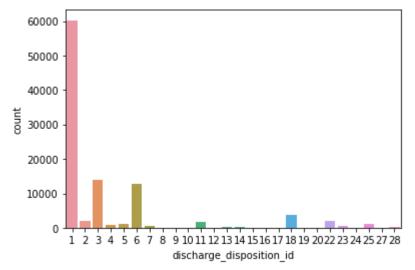
```
g = sns.catplot(x = "admission_type_id", y ="readmitted",data = df, height = 6, kind =
g.set_ylabels("Readmitted Probability")
plt.show()
```



Discharge Disposition ID

-Integer identifier corresponding to 29 distinct values. For example, discharged to home, expired, and not available

```
In [125...
sns.countplot(x ="discharge_disposition_id", data = df)
plt.show()
```



When we look the graph, we can see too much values. Getting rid of that situation, we applied this rules:

If any one includes "home" word I will grouping into one

If not, it will be as OTHER

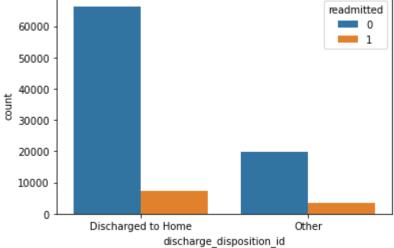
NAN = 18, 25, 26

Now it will be more clear and readable

```
sns.countplot(x ="discharge_disposition_id", data = df)
plt.show()
sns.countplot(x ="discharge_disposition_id", hue = "readmitted", data = df)
plt.show()
```

```
print("Proportions of ID's")
print(df.discharge_disposition_id.value_counts())
```

```
70000 - 60000 - 50000 - 40000 - 30000 - 20000 - 10000 - 0 Discharged to Home discharge_disposition_id
```

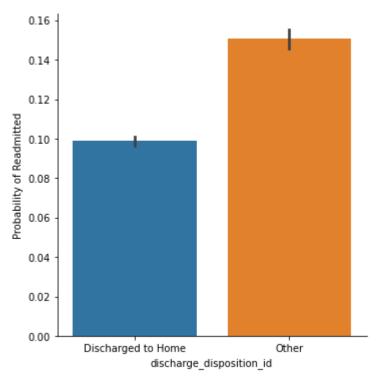


Proportions of ID's
Discharged to Home 73649
Other 23434

Name: discharge_disposition_id, dtype: int64

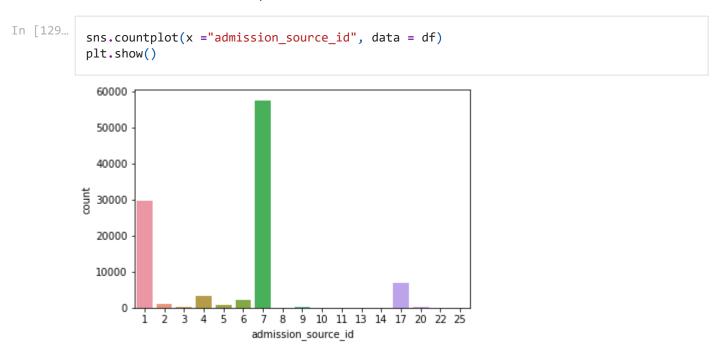
```
In [128...
```

g = sns.catplot(x = "discharge_disposition_id", y="readmitted",data = df, height = 5, k
g.set_ylabels("Probability of Readmitted")
plt.show()



Admission Source ID

Integer identifier corresponding to 21 distinct values. For example, physician referral, emergency room, and transfer from a hospital

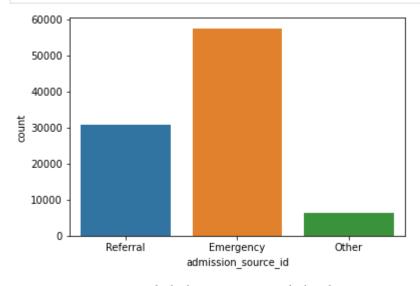


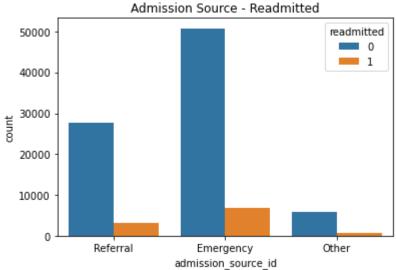
We can see that there is same problem here. Again we applied some map like:

we'll put the similar ones together like Referral or Transfer

we will replace Null, Not Mapped, Unknown values as NAN

```
15:np.nan,17:np.nan,20:np.nan,21:np.nan,
7:"Emergency"}
df.admission_source_id = df.admission_source_id.replace(mapped_adm)
sns.countplot(x = "admission_source_id", data = df)
plt.show()
sns.countplot(x = "admission_source_id", hue = "readmitted", data = df)
plt.title("Admission_Source - Readmitted")
plt.show()
print(df.admission_source_id.value_counts())
```

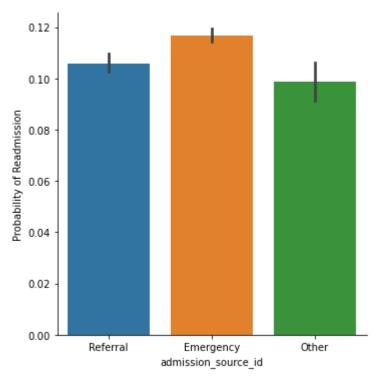




Emergency 57492 Referral 30855 Other 6474

Name: admission_source_id, dtype: int64

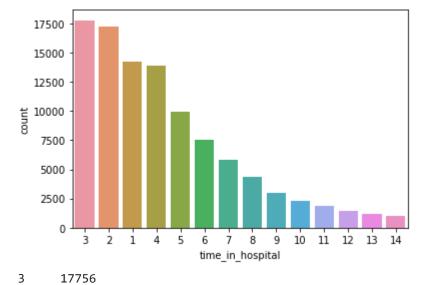
```
g = sns.catplot(x = "admission_source_id", y ="readmitted",data = df, kind = "bar", hei
g.set_ylabels("Probability of Readmission")
plt.show()
```



We see that Readmitted Probability of Referral is very close to Emergency, although Emergency is have more samples than other

Time in Hospital

Integer number of days between admission and discharge. Shortly it is "treatment time"



- 2 17224 1 14206 4 13924
- 5 99666 7539

ax.legend(loc="upper right")

ax.set_ylabel("Frequency")

plt.show()

ax.set_xlabel("Time in Hospital")

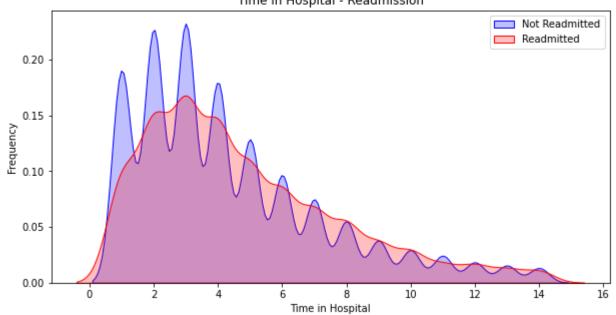
```
7
                 5859
          8
                 4390
          9
                 3002
          10
                 2342
          11
                 1855
          12
                 1448
          13
                 1210
          14
                 1042
          Name: time_in_hospital, dtype: int64
In [133...
          fig = plt.figure(figsize=(10,5))
          #readmitted = 0
          ax = sns.kdeplot(df.loc[(df.readmitted == 0), "time_in_hospital"],
                             color = "b", shade = True, label = "Not Readmitted")
```

ax.set title("Time in Hospital - Readmission")

ax = sns.kdeplot(df.loc[(df.readmitted == 1), "time in hospital"],

Time in Hospital - Readmission

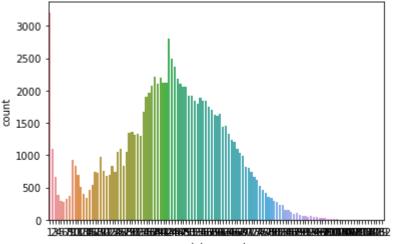
color = "r", shade = True, label = "Readmitted")



Medical Specialty Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon

We clearlt see that the Nephrology section has more probability than others

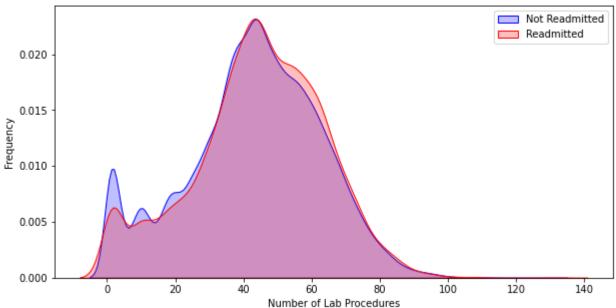
Number of Lab Procedures Number of lab tests performed during the encounter



num_lab_procedures

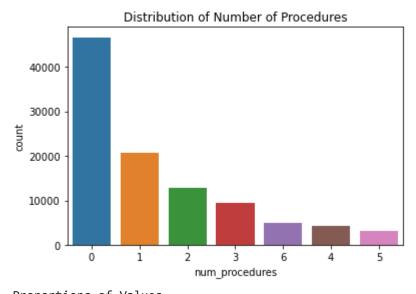
```
Proportions of Column
      3208
43
      2804
44
      2496
45
      2376
38
      2212
40
      2201
      2189
46
41
      2117
42
      2113
      2106
47
Name: num_lab_procedures, dtype: int64
```





Number of Procedures

Number of procedures (other than lab tests) performed during the encounter



```
Proportions of Values

0     45.843774

1     20.381671

2     12.495701

3     9.279404

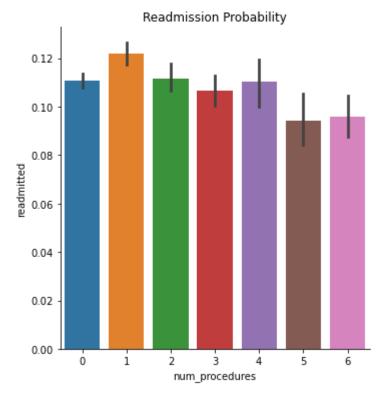
6     4.868174

4     4.107583

5     3.023692

Name: num_procedures, dtype: float64
```

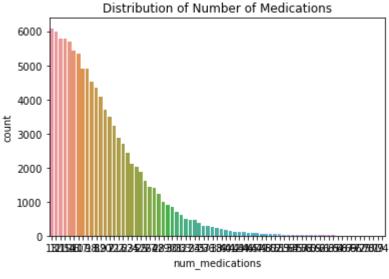
```
40000 - 35000 - 30000 - 25000 - 15000 - 10000 - 5000 - 0 0 1 2 3 6 4 5 num_procedures
```



Number of Medications

Number of distinct generic names administered during the encounter

```
plt.show()
print(df.num_medications.value_counts())
```



```
6086
13
12
      6004
      5795
11
15
      5792
14
      5707
75
70
          2
81
          1
79
74
Name: num medications, Length: 75, dtype: int64
```

Name: num_medications, Length: 75, dtype:

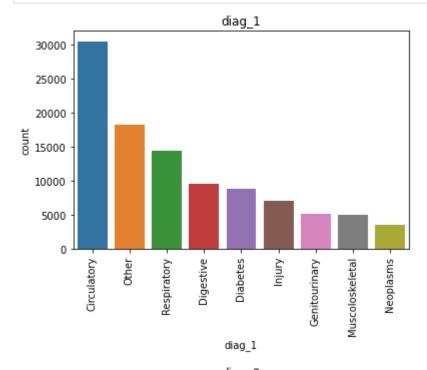
Diag1, Diag2 and Diag3

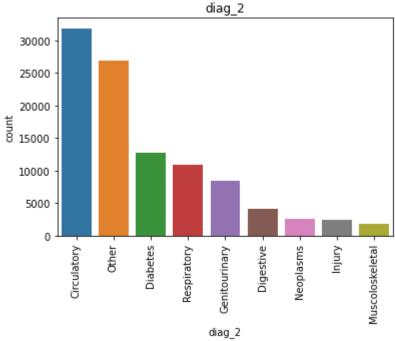
In diag section there are lots of ID that belong the specific name. So we'll map them

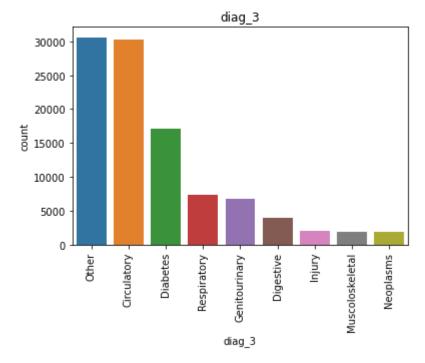
```
In [140...
            def map diagnosis(data, cols):
                for col in cols:
                     data.loc[(data[col].str.contains("V")) | (data[col].str.contains("E")), col] =
                     data[col] = data[col].astype(np.float16)
                for col in cols:
                     data["temp_diag"] = np.nan
                     data.loc[(data[col]>=390) & (data[col]<=459) | (data[col]==785), "temp_diag"] =</pre>
                     data.loc[(data[col]>=460) & (data[col]<=519) | (data[col]==786), "temp diag"] =</pre>
                     data.loc[(data[col]>=520) & (data[col]<=579) | (data[col]==787), "temp_diag"] =</pre>
                     data.loc[(data[col]>=250) & (data[col]<251), "temp_diag"] = "Diabetes"</pre>
                     data.loc[(data[col]>=800) & (data[col]<=999), "temp_diag"] = "Injury"</pre>
                     data.loc[(data[col]>=710) & (data[col]<=739), "temp_diag"] = "Muscoloskeletal"
data.loc[(data[col]>=580) & (data[col]<=629) | (data[col] == 788), "temp_diag"]</pre>
                     data.loc[(data[col]>=140) & (data[col]<=239), "temp diag"] = "Neoplasms"</pre>
                     data["temp_diag"] = data["temp_diag"].fillna("Other")
                     data[col] = data["temp_diag"]
                     data = data.drop("temp diag", axis=1)
                return data
```

for diag in diag_cols:
 plot_diags(diag,df)

```
In [141... df = map_diagnosis(df,["diag_1","diag_2","diag_3"])
```

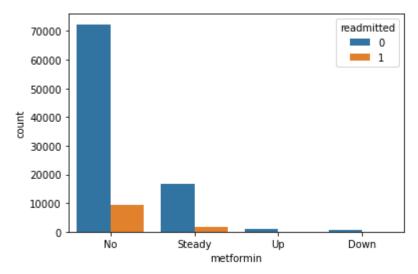






Diabetes medications

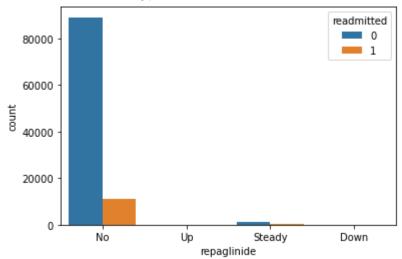
There was too many medications that belong the diabet. And some of them has just one or two value that does not any impact the model. So we decided to drop them. But firstly, lets look at the medications



METFORMIN

No 81776 Steady 18345 Up 1067 Down 575

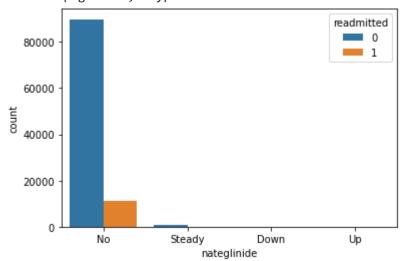
Name: metformin, dtype: int64



REPAGLINIDE

No 100224 Steady 1384 Up 110 Down 45

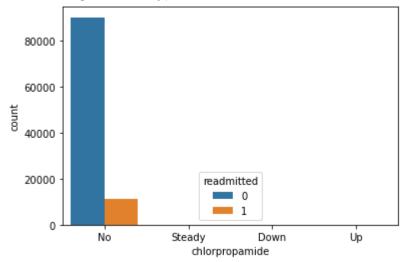
Name: repaglinide, dtype: int64



NATEGLINIDE

No 101060 Steady 668 Up 24 Down 11

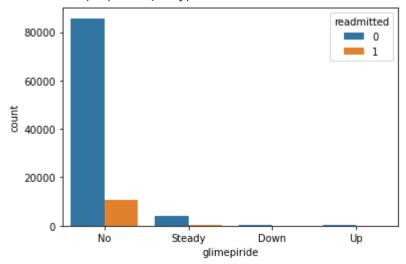
Name: nateglinide, dtype: int64



CHLORPROPAMIDE

No 101677 Steady 79 Up 6 Down 1

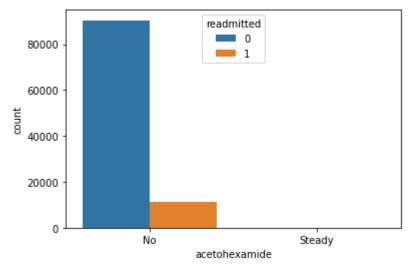
Name: chlorpropamide, dtype: int64



GLIMEPIRIDE

No 96572 Steady 4670 Up 327 Down 194

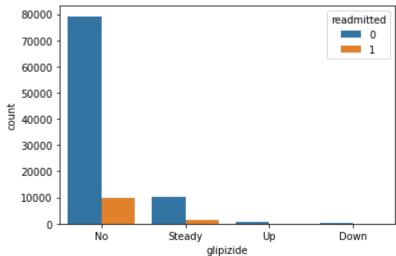
Name: glimepiride, dtype: int64



 ${\tt ACETOHEXAMIDE}$

No 101762 Steady 1

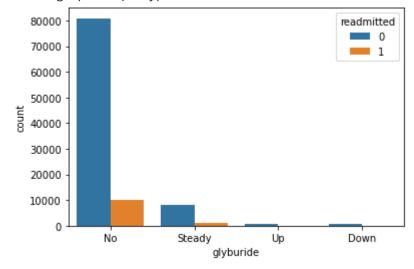
Name: acetohexamide, dtype: int64



GLIPIZIDE

No 89078 Steady 11355 Up 770 Down 560

Name: glipizide, dtype: int64

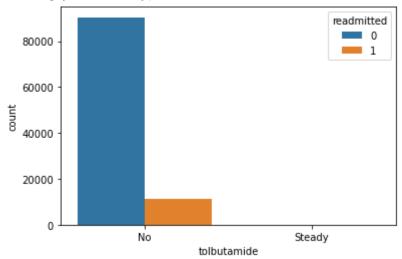


GLYBURIDE

No 91113

Steady 9274 Up 812 Down 564

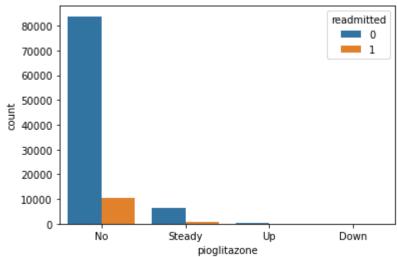
Name: glyburide, dtype: int64



TOLBUTAMIDE

No 101740 Steady 23

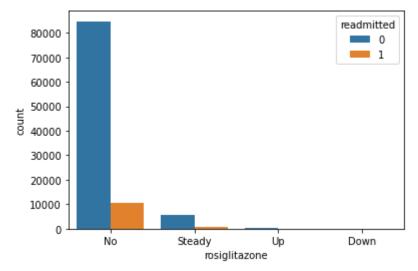
Name: tolbutamide, dtype: int64



PIOGLITAZONE

No 94436 Steady 6975 Up 234 Down 118

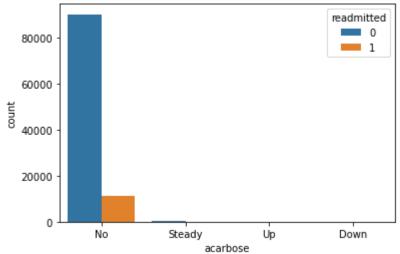
Name: pioglitazone, dtype: int64



ROSIGLITAZONE

No 95399 Steady 6099 Up 178 Down 87

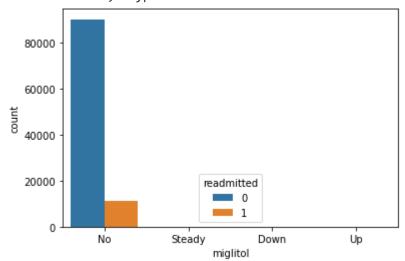
Name: rosiglitazone, dtype: int64



ACARBOSE

No 101455 Steady 295 Up 10

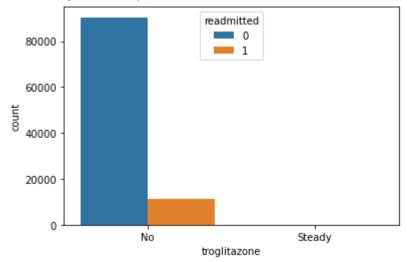
Name: acarbose, dtype: int64



MIGLITOL

No 101725 Steady 31 Down 5 Up 2

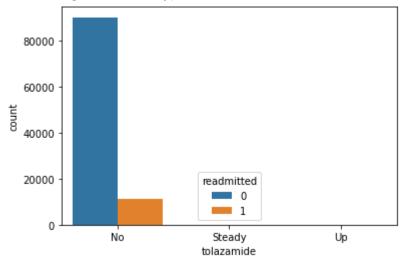
Name: miglitol, dtype: int64



TROGLITAZONE

No 101760 Steady 3

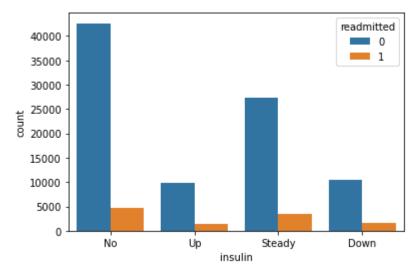
Name: troglitazone, dtype: int64



TOLAZAMIDE

No 101724 Steady 38 Up 1

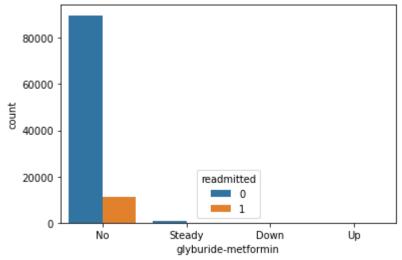
Name: tolazamide, dtype: int64



INSULIN

No 47380 Steady 30849 Down 12218 Up 11316

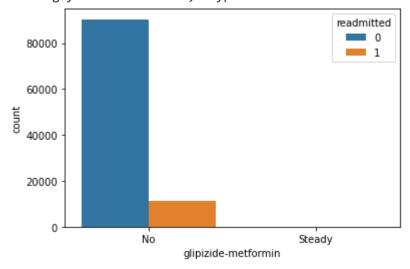
Name: insulin, dtype: int64



GLYBURIDE-METFORMIN

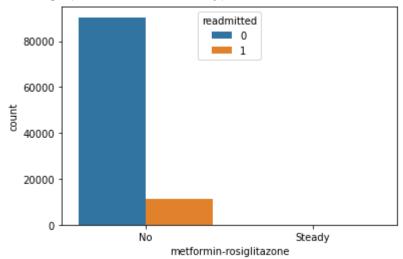
No 101057 Steady 692 Up 8 Down 6

Name: glyburide-metformin, dtype: int64



GLIPIZIDE-METFORMIN No 101750 Steady 13

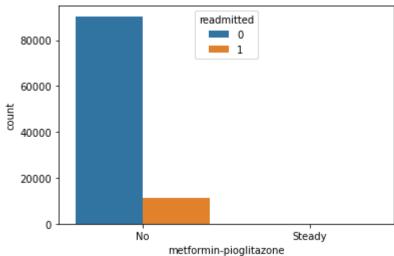
Name: glipizide-metformin, dtype: int64



METFORMIN-ROSIGLITAZONE

No 101761 Steady 2

Name: metformin-rosiglitazone, dtype: int64



METFORMIN-PIOGLITAZONE

No 101762 Steady 1

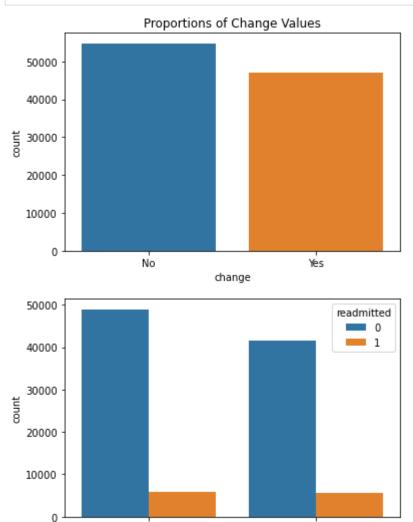
Name: metformin-pioglitazone, dtype: int64

Change

Indicates if there was a change in diabetic medications (either dosage or generic name). Values:

"change"

"no change"

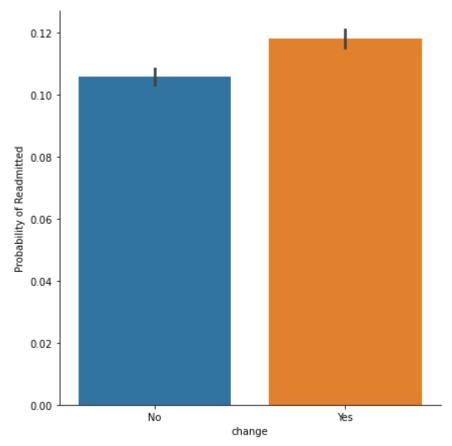


No

```
In [146...
g = sns.catplot(x = "change", y="readmitted", data = df, height = 6, kind ="bar")
g.set_ylabels("Probability of Readmitted")
plt.show()
```

change

Yes



Glucose Serum Test Result

Indicates the range of the result or if the test was not taken.

Values:

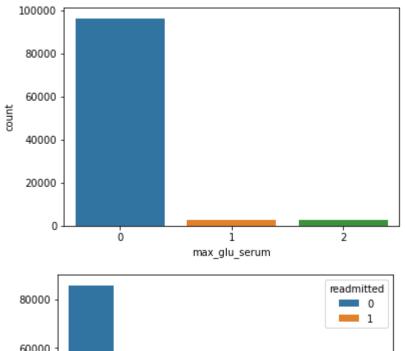
">200,"

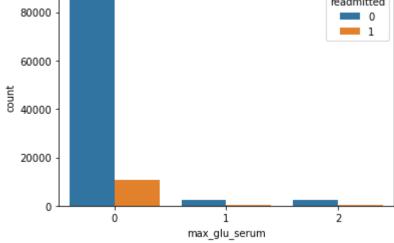
">300,"

"normal,"

"none" if not measured

We decided use the Glucose Serum Test Result like as follows:





9641727492597

Name: max_glu_serum, dtype: int64

A1c test result

Indicates the range of the result or if the test was not taken.

Values:

">8" if the result was greater than 8%,

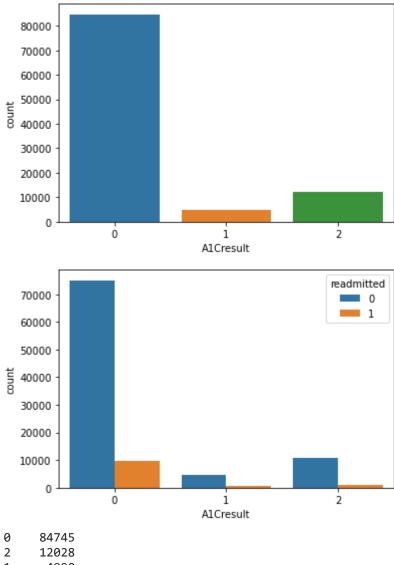
">7" if the result was greater than 7% but less than 8%, "normal"

if the result was less than 7%, and "none" if not measured.

We decided use the A1c test result like as follows:

Untitled 10/20/21, 5:13 AM

```
sns.countplot(x = "A1Cresult", hue = "readmitted", data = df)
plt.show()
print(df.A1Cresult.value_counts())
```



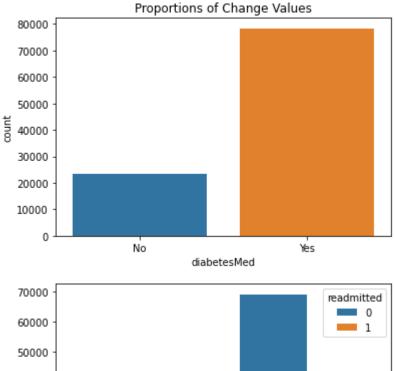
4990

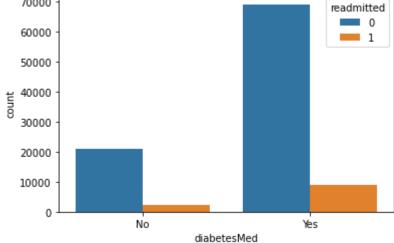
Name: A1Cresult, dtype: int64

Diabetes medications

Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"

```
In [149...
          sns.countplot(x = "diabetesMed", data = df )
          plt.title("Proportions of Change Values")
          plt.show()
          sns.countplot(x = "diabetesMed", hue = "readmitted", data = df)
          plt.show()
          print(df.diabetesMed.value_counts())
```



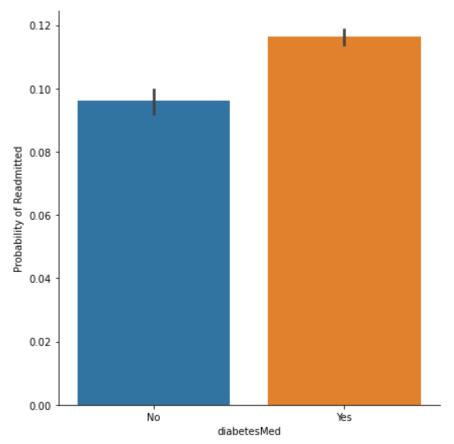


Yes 78361 No 23402

Name: diabetesMed, dtype: int64

```
In [150...
```

g = sns.catplot(x = "diabetesMed", y="readmitted",data = df, height = 6, kind ="bar")
g.set_ylabels("Probability of Readmitted")
plt.show()



[151	df.isnull().sum()	
	· · · · · ·	
ut[151	encounter_id	0
	patient_nbr	0
	race	0
	gender	0
	age	0
	admission_type_id	10396
	discharge_disposition_id	4680
	admission_source_id	6942
	time_in_hospital	0
	num_lab_procedures	0
	num_procedures	0
	num_medications	0
	number_outpatient	0
	number_emergency	0
	number_inpatient	0
	diag_1	0
	diag_2	0
	diag_3	0
	number_diagnoses	0
	max_glu_serum	0
	A1Cresult	0
	metformin	0
	repaglinide	0
	nateglinide	0
	chlorpropamide	0
	glimepiride	0
	acetohexamide	0
	glipizide	0
	glyburide	0
	tolbutamide	0
	pioglitazone	0

```
rosiglitazone
                                  0
acarbose
                                  0
miglitol
                                  0
troglitazone
                                  0
tolazamide
                                  0
insulin
                                  0
glyburide-metformin
                                  0
glipizide-metformin
                                  0
glimepiride-pioglitazone
                                  0
metformin-rosiglitazone
                                  0
metformin-pioglitazone
                                  0
change
                                  0
diabetesMed
                                  0
readmitted
                                  0
dtype: int64
```

```
In [152...

df['race'] = df['race'].fillna(df['race'].mode()[0])
```

```
In [153... df['admission_type_id'] = df['admission_type_id'].fillna(df['admission_type_id'].mode()
```

In [154... df['discharge_disposition_id'] = df['discharge_disposition_id'].fillna(df['discharge_di

In [155... df['admission_source_id'] = df['admission_source_id'].fillna(df['admission_source_id'].

In [156... df.head()

Out[156... encounter_id patient_nbr race g

	encounter_id	patient_nbr	race	gender	age	admission_type_id	discharge_disposition_id
0	2278392	8222157	Caucasian	Female	5	Emergency	Discharged to Home
1	149190	55629189	Caucasian	Female	15	Emergency	Discharged to Home
2	64410	86047875	AfricanAmerican	Female	25	Emergency	Discharged to Home
3	500364	82442376	Caucasian	Male	35	Emergency	Discharged to Home
4	16680	42519267	Caucasian	Male	45	Emergency	Discharged to Home

5 rows × 45 columns

CODE

```
In [157... df.shape
Out[157... (101763, 45)
In [158... df.head()
```

Out[158...

	encounter_id	patient_nbr	race	gender	age	admission_type_id	discharge_disposition_id
0	2278392	8222157	Caucasian	Female	5	Emergency	Discharged to Home
1	149190	55629189	Caucasian	Female	15	Emergency	Discharged to Home
2	64410	86047875	AfricanAmerican	Female	25	Emergency	Discharged to Home
3	500364	82442376	Caucasian	Male	35	Emergency	Discharged to Home
4	16680	42519267	Caucasian	Male	45	Emergency	Discharged to Home

5 rows × 45 columns

cat_data = df.select_dtypes('0')
num_data = df.select_dtypes(np.number)
cat_data

Ο.	-1-1		1
	IT.		h I
\sim \circ	4 6	-	O · ·

	race	gender	admission_type_id	discharge_disposition_id	admission_source_id	di
0	Caucasian	Female	Emergency	Discharged to Home	Referral	Dia
1	Caucasian	Female	Emergency	Discharged to Home	Emergency	(
2	AfricanAmerican	Female	Emergency	Discharged to Home	Emergency	(
3	Caucasian	Male	Emergency	Discharged to Home	Emergency	(
4	Caucasian	Male	Emergency	Discharged to Home	Emergency	Neopl
•••						
101761	AfricanAmerican	Male	Emergency	Other	Emergency	Dia
101762	AfricanAmerican	Female	Emergency	Other	Other	Dige
101763	Caucasian	Male	Emergency	Discharged to Home	Emergency	(
101764	Caucasian	Female	Emergency	Other	Emergency	1
101765	Caucasian	Male	Emergency	Discharged to Home	Emergency	Dig€

101763 rows × 31 columns

```
In [162...
          from sklearn.preprocessing import LabelEncoder
          le = LabelEncoder()
          for i in cat data:
               cat_data[i] = le.fit_transform(cat_data[i])
          <ipython-input-162-74041ae2a149>:6: SettingWithCopyWarning:
          A value is trying to be set on a copy of a slice from a DataFrame.
          Try using .loc[row_indexer,col_indexer] = value instead
         See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user
          guide/indexing.html#returning-a-view-versus-a-copy
            cat_data[i] = le.fit_transform(cat_data[i])
In [163...
          df = pd.concat([num_data,cat_data],axis=1)
          df.head()
Out[163...
             age time_in_hospital num_lab_procedures num_procedures num_medications number_outpatient n
          0
              5
                              1
                                                                0
                                                                                                  0
                                                41
                                                                                1
              15
                              3
                                                59
                                                                0
                                                                               18
                                                                                                  0
          1
          2
              25
                              2
                                                                5
                                                                               13
                                                                                                  2
                                                11
```

5 rows × 43 columns

Splitting the dependent and independent variable¶

```
from sklearn.model_selection import train_test_split
In [167...
          X train,X test,y train,y test = train test split(X,y,test size=0.3,random state=99)
In [168...
          SC = StandardScaler()
          X_train_scaled = pd.DataFrame(SC.fit_transform(X_train),columns=X_train.columns)
          X test scaled = pd.DataFrame(SC.transform(X test),columns=X test.columns)
In [169...
          X_train.shape,X_test.shape,y_train.shape,y_test.shape
Out[169... ((71234, 42), (30529, 42), (71234,), (30529,))
         PCA
In [170...
          from sklearn.decomposition import PCA
          pca = PCA()
          X_train_pca=pca.fit_transform(X_train_scaled)
In [171...
          total=sum(pca.explained_variance_)
          k=0
          current_variance=0
          while current variance/total < 0.99:</pre>
               current_variance += pca.explained_variance_[k]
          k
Out[171... 41
In [172...
          import matplotlib.pyplot as plt
          cum sum = pca.explained variance ratio .cumsum()
          cum sum = cum sum*100
          plt.bar(range(k+1), cum_sum)
          plt.title("Around 99% of variance is explained by the {} features".format(k));
              Around 99% of variance is explained by the 41 features
          100
           80
           60
           40
           20
```

In [173...

```
from sklearn.preprocessing import MinMaxScaler
           sc = MinMaxScaler(feature range = (0, 1))
           X_pca = sc.fit_transform(X_train_pca)
In [174...
           pca = PCA(2)
           projected = pca.fit_transform(X_train_pca)
           print(X train pca.shape)
           print(projected.shape)
          (71234, 42)
          (71234, 2)
In [175...
           plt.figure(figsize=(12, 6))
           plt.scatter(projected[:, 0],
           projected[:, 1],
           c=y_train,
           edgecolor='none',
           alpha=1,
           cmap=plt.cm.get_cmap('gnuplot', 5)
           plt.xlabel('Component 1')
           plt.ylabel('Component 2')
           plt.colorbar();
                                                                                                      1.0
             8
                                                                                                      - 0.8
             6
             4
                                                                                                      0.6
          Component 2
             2
                                                                                                      - 0.4
             0
            -2
                                                                                                      - 0.2
            -4
             -6
                                  -2
```

MODELS

LOGISTIC REGRESSION

```
from sklearn.linear_model import LogisticRegression
LR = LogisticRegression(fit_intercept=True, penalty='12')
```

Component 1

```
LR.fit(X_train_scaled,y_train)
          #LR.fit(X train, y train)
Out[176... LogisticRegression()
In [177...
          y pred LR = LR.predict(X test scaled)
           #y pred LR = LR.predict(X test)
In [178...
           from sklearn.metrics import confusion matrix
           confusion_matrix(y_test,y_pred_LR)
Out[178... array([[27089,
                            54],
                 [ 3340,
                            46]], dtype=int64)
In [179...
          from sklearn.metrics import classification report
          print("Classification report - \n", classification report(y test,y pred LR))
          Classification report -
                         precision
                                       recall f1-score
                                                           support
                             0.89
                                        1.00
                                                  0.94
                                                            27143
                     0
                             0.46
                                        0.01
                                                  0.03
                                                             3386
                     1
                                                  0.89
                                                            30529
              accuracy
                             0.68
                                        0.51
                                                  0.48
                                                            30529
             macro avg
                                                  0.84
                             0.84
                                        0.89
                                                            30529
         weighted avg
In [180...
          from sklearn.metrics import accuracy score
          accuracy_score(y_test,y_pred_LR)
Out[180... 0.8888270169347178
```

SMOTE LR

```
In [181... X.shape, y.shape

Out[181... ((101763, 42), (101763,))

In [182... print('Original dataset shape {}'.format(Counter(y)))
    smt = SMOTE(random_state=20)
    train_input_new, train_output_new = smt.fit_resample(X, y)
    print('New dataset shape {}'.format(Counter(train_output_new)))
    train_input_new = pd.DataFrame(train_input_new, columns = list(X.columns))
    X_train, X_test, y_train, y_test = train_test_split(train_input_new, train_output_new,
    Original dataset shape Counter({0: 90406, 1: 11357})
    New dataset shape Counter({0: 90406, 1: 90406})

In [184... from sklearn.linear_model import LogisticRegression
    LR = LogisticRegression(fit_intercept=True, penalty='12')
```

```
LR.fit(X_train,y_train)
          #LR.fit(X train, y train)
         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: Conver
         genceWarning: lbfgs failed to converge (status=1):
         STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
         Increase the number of iterations (max iter) or scale the data as shown in:
              https://scikit-learn.org/stable/modules/preprocessing.html
         Please also refer to the documentation for alternative solver options:
              https://scikit-learn.org/stable/modules/linear model.html#logistic-regression
           n iter i = check optimize result(
Out[184... LogisticRegression()
In [185...
          y pred LR = LR.predict(X test)
          #y pred LR = LR.predict(X test)
In [186...
          from sklearn.metrics import confusion matrix
          confusion matrix(y test,y pred LR)
Out[186... array([[12437, 5487],
                 [ 4977, 13262]], dtype=int64)
In [187...
          from sklearn.metrics import accuracy score
          accuracy_score(y_test,y_pred_LR)
Out[187... 0.7106434753753836
```

RANDOM FOREST

```
In [188...
          from sklearn.ensemble import RandomForestClassifier
          RF = RandomForestClassifier()
          #RF.fit(X_train_scaled,y_train)
          RF.fit(X_train,y_train)
Out[188... RandomForestClassifier()
In [189...
           #y pred RF = RF.predict(X test scaled)
          y pred RF = RF.predict(X test)
In [190...
          from sklearn.metrics import confusion matrix
           confusion_matrix(y_test,y_pred_RF)
Out[190... array([[16159, 1765],
                 [ 2415, 15824]], dtype=int64)
In [191...
          print("Classification report - \n", classification_report(y_test,y_pred_RF))
          Classification report -
```

```
precision
                            recall f1-score
                                                support
           0
                   0.87
                              0.90
                                        0.89
                                                  17924
           1
                   0.90
                              0.87
                                        0.88
                                                  18239
                                        0.88
                                                  36163
    accuracy
                              0.88
                                        0.88
                                                  36163
   macro avg
                   0.88
weighted avg
                   0.88
                              0.88
                                        0.88
                                                  36163
```

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,y_pred_RF)
```

Out[192... 0.8844122445593563

plt.figure(figsize=(12,24)) sns.barplot(sorted(RF.feature*importances*),X_train.columns);

z = pd.DataFrame([RF.featureimportances,X_train.columns]).T

z.columns = ['Feature', 'importance'] plt.figure(figsize=(13,12))

sns.barplot(y=z['Feature'],x=z['importance'])

important_features = pd.DataFrame({'Features': X_train.columns,'Importance':

RF.feature*importances*}) important_features = important_features.sort_values('Importance', ascending = False) plt.figure(figsize=(12,24)) sns.barplot(x = 'Importance', y = 'Features', data = important_features) plt.title('Feature Importance', fontsize = 15) plt.ylabel('Importance', fontsize = 15) plt.ylabel('Features', fontsize = 15) plt.show()

Decision Tree

```
from sklearn.tree import DecisionTreeClassifier#for checking testing results
from sklearn.metrics import classification_report, confusion_matrix#for visualizing tre
```

Defining the decision tree algorithm
dtree=DecisionTreeClassifier()
dtree.fit(X_train,y_train)
print('Decision Tree Classifier Created')

Decision Tree Classifier Created

```
# Predicting the values of test data
y_pred_DT = dtree.predict(X_test)
print("Classification report - \n", classification_report(y_test,y_pred_DT))
```

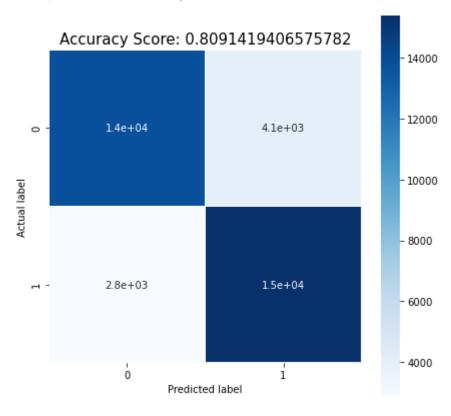
```
Classification report -
               precision
                            recall f1-score
                                                support
           0
                   0.83
                             0.77
                                        0.80
                                                 17924
           1
                   0.79
                                        0.82
                                                 18239
                             0.84
                                        0.81
    accuracy
                                                 36163
   macro avg
                                        0.81
                   0.81
                             0.81
                                                 36163
                   0.81
                                        0.81
                                                 36163
weighted avg
                              0.81
```

```
In [196... from sklearn.metrics import accuracy_score accuracy_score(y_test,y_pred_DT)
```

Out[196... 0.8091419406575782

```
cm = confusion_matrix(y_test, y_pred_DT)
plt.figure(figsize=(7,7))
sns.heatmap(data=cm,linewidths=.5, annot=True,square = True, cmap = 'Blues')
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
all_sample_title = 'Accuracy Score: {0}'.format(dtree.score(X_test, y_test))
plt.title(all_sample_title, size = 15)
```

Out[197... Text(0.5, 1.0, 'Accuracy Score: 0.8091419406575782')



SVM

from sklearn.svm import SVC classifier = SVC(kernel='rbf', random_state = 1)

classifier.fit(X_train,y_train)

Y_pred_SVM = classifier.predict(X_test)

from sklearn.metrics import confusion_matrix cm = confusion_matrix(y_test,Y_pred_SVM) cm

from sklearn.metrics import accuracy_score accuracy_score(y_test,Y_pred_SVM)

Neural Network

```
In [198... from keras.models import Sequential
```

```
from keras.layers import Dense
from keras.layers import Dropout
```

```
In [201...
```

```
Epoch 3/300
Epoch 4/300
Epoch 5/300
Epoch 6/300
Epoch 7/300
Epoch 8/300
Epoch 9/300
Epoch 10/300
Epoch 11/300
Epoch 12/300
Epoch 13/300
Epoch 14/300
Epoch 15/300
Epoch 16/300
Epoch 17/300
Epoch 18/300
Epoch 19/300
Epoch 20/300
Epoch 21/300
Epoch 22/300
Epoch 23/300
Epoch 24/300
```

```
Epoch 25/300
Epoch 26/300
Epoch 27/300
Epoch 28/300
Epoch 29/300
Epoch 30/300
Epoch 31/300
Epoch 32/300
Epoch 33/300
Epoch 34/300
Epoch 35/300
Epoch 36/300
Epoch 37/300
Epoch 38/300
Epoch 39/300
Epoch 40/300
Epoch 41/300
Epoch 42/300
Epoch 43/300
Epoch 44/300
Epoch 45/300
Epoch 46/300
Epoch 47/300
Epoch 48/300
Epoch 49/300
Epoch 50/300
Epoch 51/300
Epoch 52/300
Epoch 53/300
Epoch 54/300
Epoch 55/300
Epoch 56/300
```

```
Epoch 57/300
Epoch 58/300
Epoch 59/300
Epoch 60/300
Epoch 61/300
Epoch 62/300
Epoch 63/300
Epoch 64/300
Epoch 65/300
Epoch 66/300
Epoch 67/300
Epoch 68/300
Epoch 69/300
Epoch 70/300
Epoch 71/300
Epoch 72/300
Epoch 73/300
Epoch 74/300
Epoch 75/300
Epoch 76/300
Epoch 77/300
Epoch 78/300
Epoch 79/300
Epoch 80/300
Epoch 81/300
Epoch 82/300
Epoch 83/300
Epoch 84/300
Epoch 85/300
Epoch 86/300
Epoch 87/300
Epoch 88/300
Epoch 89/300
```

```
Epoch 90/300
Epoch 91/300
Epoch 92/300
Epoch 93/300
Epoch 94/300
Epoch 95/300
Epoch 96/300
Epoch 97/300
Epoch 98/300
Epoch 99/300
Epoch 100/300
Epoch 101/300
Epoch 102/300
Epoch 103/300
Epoch 104/300
Epoch 105/300
Epoch 106/300
Epoch 107/300
Epoch 108/300
Epoch 109/300
Epoch 110/300
Epoch 111/300
Epoch 112/300
Epoch 113/300
Epoch 114/300
Epoch 115/300
8: 0s - loss: 0.5095 - accura
Epoch 116/300
Epoch 117/300
Epoch 118/300
Epoch 119/300
Epoch 120/300
Epoch 121/300
```

```
Epoch 122/300
Epoch 123/300
Epoch 124/300
Epoch 125/300
Epoch 126/300
Epoch 127/300
Epoch 128/300
Epoch 129/300
Epoch 130/300
Epoch 131/300
Epoch 132/300
Epoch 133/300
Epoch 134/300
Epoch 135/300
Epoch 136/300
Epoch 137/300
Epoch 138/300
Epoch 139/300
Epoch 140/300
Epoch 141/300
Epoch 142/300
Epoch 143/300
Epoch 144/300
Epoch 145/300
Epoch 146/300
Epoch 147/300
Epoch 148/300
Epoch 149/300
Epoch 150/300
Epoch 151/300
Epoch 152/300
Epoch 153/300
```

```
Epoch 154/300
Epoch 155/300
Epoch 156/300
Epoch 157/300
Epoch 158/300
Epoch 159/300
Epoch 160/300
Epoch 161/300
Epoch 162/300
Epoch 163/300
Epoch 164/300
Epoch 165/300
Epoch 166/300
Epoch 167/300
Epoch 168/300
Epoch 169/300
Epoch 170/300
Epoch 171/300
Epoch 172/300
Epoch 173/300
Epoch 174/300
Epoch 175/300
Epoch 176/300
Epoch 177/300
Epoch 178/300
Epoch 179/300
Epoch 180/300
Epoch 181/300
Epoch 182/300
Epoch 183/300
Epoch 184/300
Epoch 185/300
Epoch 186/300
```

```
Epoch 187/300
Epoch 188/300
Epoch 189/300
Epoch 190/300
Epoch 191/300
Epoch 192/300
Epoch 193/300
Epoch 194/300
Epoch 195/300
Epoch 196/300
Epoch 197/300
Epoch 198/300
8: 0s - loss: 0.4970 - accu
Epoch 199/300
Epoch 200/300
Epoch 201/300
Epoch 202/300
Epoch 203/300
Epoch 204/300
Epoch 205/300
Epoch 206/300
Epoch 207/300
Epoch 208/300
Epoch 209/300
Epoch 210/300
Epoch 211/300
Epoch 212/300
Epoch 213/300
Epoch 214/300
Epoch 215/300
Epoch 216/300
Epoch 217/300
Epoch 218/300
```

```
Epoch 219/300
Epoch 220/300
Epoch 221/300
Epoch 222/300
Epoch 223/300
Epoch 224/300
Epoch 225/300
Epoch 226/300
Epoch 227/300
Epoch 228/300
Epoch 229/300
Epoch 230/300
Epoch 231/300
Epoch 232/300
Epoch 233/300
Epoch 234/300
Epoch 235/300
Epoch 236/300
Epoch 237/300
Epoch 238/300
Epoch 239/300
Epoch 240/300
Epoch 241/300
Epoch 242/300
Epoch 243/300
Epoch 244/300
Epoch 245/300
Epoch 246/300
Epoch 247/300
Epoch 248/300
Epoch 249/300
Epoch 250/300
```

```
Epoch 251/300
Epoch 252/300
Epoch 253/300
Epoch 254/300
Epoch 255/300
Epoch 256/300
Epoch 257/300
Epoch 258/300
Epoch 259/300
Epoch 260/300
Epoch 261/300
Epoch 262/300
Epoch 263/300
Epoch 264/300
Epoch 265/300
Epoch 266/300
Epoch 267/300
Epoch 268/300
Epoch 269/300
Epoch 270/300
Epoch 271/300
9: 0s - loss: 0.5027 - accuracy: 0.
Epoch 272/300
Epoch 273/300
Epoch 274/300
Epoch 275/300
Epoch 276/300
Epoch 277/300
Epoch 278/300
Epoch 279/300
Epoch 280/300
Epoch 281/300
Epoch 282/300
```

```
8: 0s - loss: 0.5011 - accuracy: 0.75
Epoch 283/300
Epoch 284/300
Epoch 285/300
Epoch 286/300
Epoch 287/300
Epoch 288/300
Epoch 289/300
Epoch 290/300
Epoch 291/300
Epoch 292/300
Epoch 293/300
Epoch 294/300
Epoch 295/300
Epoch 296/300
Epoch 297/300
Epoch 298/300
Epoch 299/300
Epoch 300/300
75970s - loss: 0.4
Accuracy: 75.97
```

In []: