FEATURE NAME	Data Type	TRANSFORMATION DESCRIPTION
encounter_id	*Number*	*Dropped as it doesn't affect output*
patient_nbr	*Number*	*Dropped as it doesn't affect output*
payer_code	*Number*	*Dropped as it doesn't affect output*
race	*Category*	*1) Changed '?' to 'Other'* *2) ('Caucasian','AfricanAmerican','Hispanic') were kept same and rest are changed to 'Other'* *3) Later on Hot Encoding is applied. Discussed in next section*
gender	*Category*	*1) Dropped all rows which contained 'Unknown/Invalid' value* *2) Later on Hot Encoding is applied. Discussed in next section*
age	*Number*	*Converted range values to mid average value.* *For example -> "10-20" is changed to 15 and so on.*
A1Cresult	*Category*	*1) Changed 'None' to 'Undefined'* *2) Changed '>8' to '8+'* *3) Changed '>7' to '7+'* *4) Later on Hot Encoding is applied. Discussed in next section*
max_glu_serum	*Category*	*Dropped as there were 95% missing values.*

weight	*Category*	*Dropped as there were 95% missing values.*
examide	*Category*	*As there is only one value and it doesn't affect output*
citoglipton	*Category*	*As there is only one value and it doesn't affect output*
diabetesMed	*number*	*'Yes' is converted to 1 and rest is changed to*
change	*Number*	*'Ch' is converted to 1 and rest is converted to 0.*
medical_specialty	*Category*	1) *('InternalMedicine','Emergency/Trauma','FamilyGeneralPractice','Cardiology','Surgery-General') are kept same and rest are changed to 'Other'.* 2) *Later on Hot Encoding is applied. Discussed in next section*
admission_type_id	*Number*	*No changes*
discharge_disposition_id	*Number*	*No changes*
admission_source_id	*Number*	*No changes*
time_in_hospital	*Number*	*No changes*
num_lab_procedures	*Number*	*No changes*
num_procedures	*Number*	*No changes*

num_medications	*Number*	*No changes*
number_outpatient	*Number*	*No changes*
number_emergency	*Number*	*No changes*
number_inpatient	*Number*	*No changes*
number_diagnoses	*Number*	*No changes*
readmitted	*Number*	*1) "NO" is replaced by '>30'* *2) '<30' is converted to 1 and rest is converted to 0.*
diag_1	*Category*	*1) Extracted the number part from string.* *2) Applied diag_1_apply function to convert to "circulatory","diabetes","digestive","genitourinary","injury","musculoskeletal","neoplasms","respiratory" or "other"* *3) Later on Hot Encoding is applied. Discussed in next section*
diag_2	*Category*	*1) ('250','401','276','305') are kept same and rest is converted to "other".* *2) Later on Hot Encoding is applied. Discussed in next section*

diag_3	*Category*	*1) ('250','401','276','305') are kept same and rest is converted to "other".* *2) Later on Hot Encoding is applied. Discussed in next section*
metformin	*Number*	*Converted "No" to 0 else converted to 1.*
repaglinide	*Number*	*Converted "No" to 0 else converted to 1.*
nateglinide	*Number*	*Converted "No" to 0 else converted to 1.*
chlorpropamide	*Number*	*Converted "No" to 0 else converted to 1.*
glimepiride	*Number*	*Converted "No" to 0 else converted to 1.*
acetohexamide	*Number*	*Converted "No" to 0 else converted to 1.*
glipizide	*Number*	*Converted "No" to 0 else converted to 1.*
glyburide	*Number*	*Converted "No" to 0 else converted to 1.*
tolbutamide	*Number*	*Converted "No" to 0 else converted to 1.*
pioglitazone	*Number*	*Converted "No" to 0 else converted to 1.*
rosiglitazone	*Number*	*Converted "No" to 0 else converted to 1.*
acarbose	*Number*	*Converted "No" to 0 else converted to 1.*
miglitol	*Number*	*Converted "No" to 0 else converted to 1.*
troglitazone	*Number*	*Converted "No" to 0 else converted to 1.*
tolazamide	*Number*	*Converted "No" to 0 else converted to 1.*

insulin	*Number*	*Converted "No" to 0 else converted to 1.*
glyburide-metformin	*Number*	*Converted "No" to 0 else converted to 1.*
glipizide-metformin	*Number*	*Converted "No" to 0 else converted to 1.*
glimepiride-pioglitazone	*Number*	*Converted "No" to 0 else converted to 1.*
metformin-rosiglitazone	*Number*	*Converted "No" to 0 else converted to 1.*
metformin-pioglitazone	*Number*	*Converted "No" to 0 else converted to 1.*

ONE HOT ENCODING TRANSFORMATION

Applied To – Categorical Data Columns.

Transformed Columns – diag_1, diag_2, diag_3, medical_specialty, A1Cresult, race, gender

Description

- One hot encoded columns are created from Categorical Columns.
  For example -> Column race contains values ('Caucasian','AfricanAmerican','Hispanic' and 'other') will create 4 different columns with column names as race _Caucasian, race _AfricanAmerican, race _digestive and race_Hispanic.
- One hot encoded columns are concatenated with dataset.
- Categorical Columns were dropped from dataset
- Column Names were replaced "<" with 'less' , ">" with 'greater' and '/' with '_'.

Changes Done After One – Hot Encoding Transformations

ADDED COLUMNS	SOURCE COLUMN	DESCRIPTION
race_AfricanAmerican race_Caucasian race_Hispanic race_Other	race	Total 4 new columns are created from race column in one hot encoding format.
gender_Female gender_Male	gender	Total 2 new columns are created from gender column in one hot encoding format.
medical_specialty_Cardiology medical_specialty_Emergency_Trauma medical_specialty_Family_GeneralPractice medical_specialty_InternalMedicine medical_specialty_Other medical_specialty_Surgery-General	medical_specialty	Total 6 new columns are created from medical_specialty column in one hot encoding format.
diag_1_Other diag_1_circulatory diag_1_diabetes diag_1_digestive diag_1_genitourinary diag_1_injury diag_1_musculoskeletal diag_1_neoplasms diag_1_other diag_1_respiratory	diag_1	Total 10 new columns are created from diag_1 column in one hot encoding format.

diag_2_250 diag_2_276 diag_2_305 diag_2_401 diag_2_other	diag_2	Total 5 new columns are created from diag_2 column in one hot encoding format.
diag_3_250 diag_3_276 diag_3_305 diag_3_401 diag_3_other	diag_3	Total 5 new columns are created from diag_3 column in one hot encoding format.
A1Cresult_7+ A1Cresult_8+ A1Cresult_Norm A1Cresult_Undefined	A1Cresult	Total 4 new columns are created from A1Cresult column in one hot encoding format.

Output – All Categorical Columns are now converted to One Hot Encoding format with either 0 or 1 values. Total 36 new columns are added inplace of 7 columns [diag_1, diag_2, diag_3, medical_specialty, A1Cresult, race, gender].