**Basic EDA Steps:**

In our dataset we have 101766 rows and 50 columns.

In our dataset we have 13 numerical columns and 37 categorical columns.

In our dataset we have null values in 7 columns:

weight 96.858479

medical\_specialty 49.082208

payer\_code 39.557416

race 2.233555

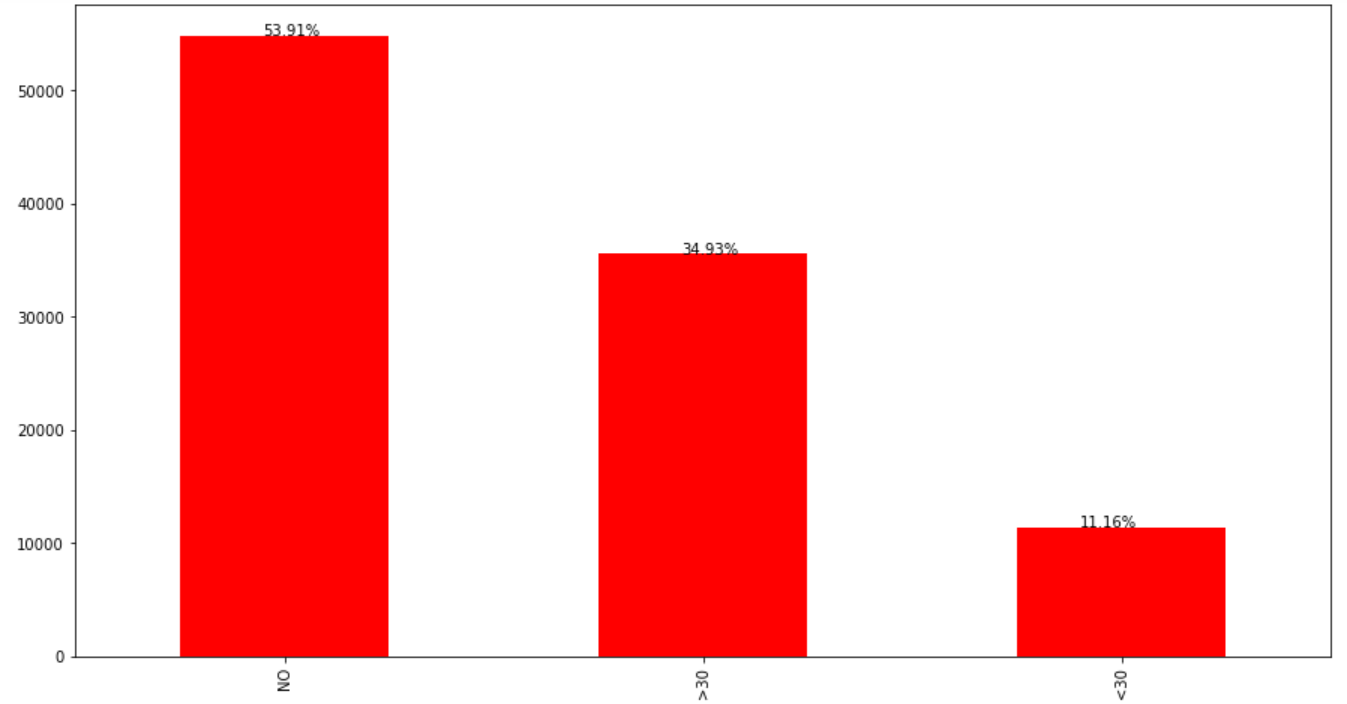
diag\_3 1.398306

diag\_2 0.351787

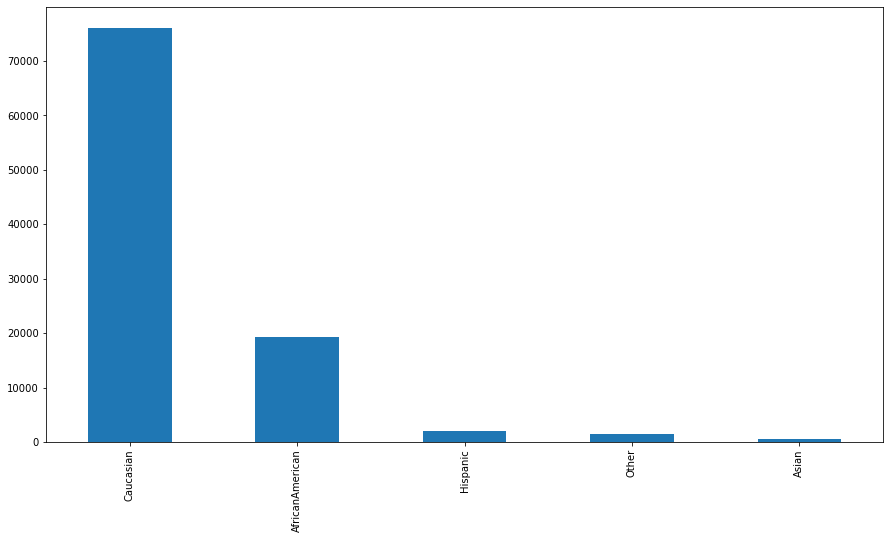
diag\_1 0.020636

**Univariate analysis for categorical variables:**

1. **Readmitted:**

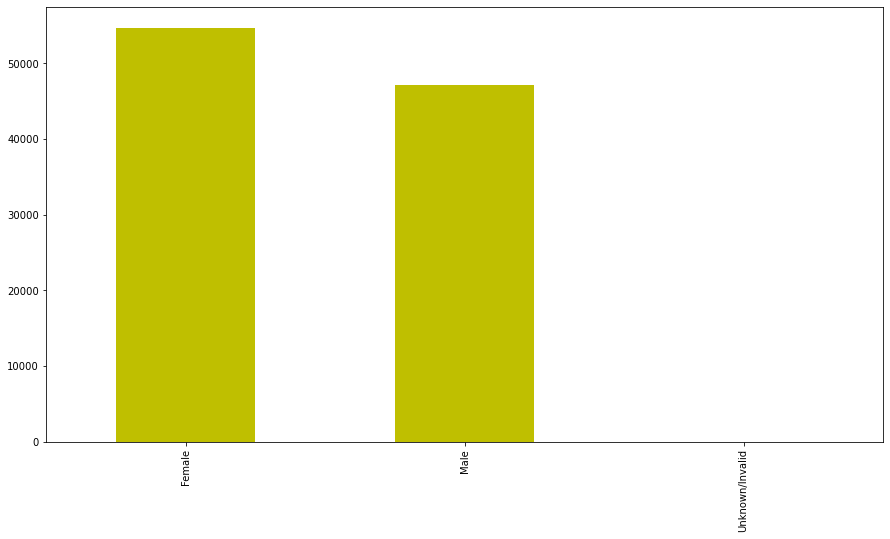
****

Days to inpatient readmission. “No” for no record of readmission has the most counted observation.

2.Gender: 

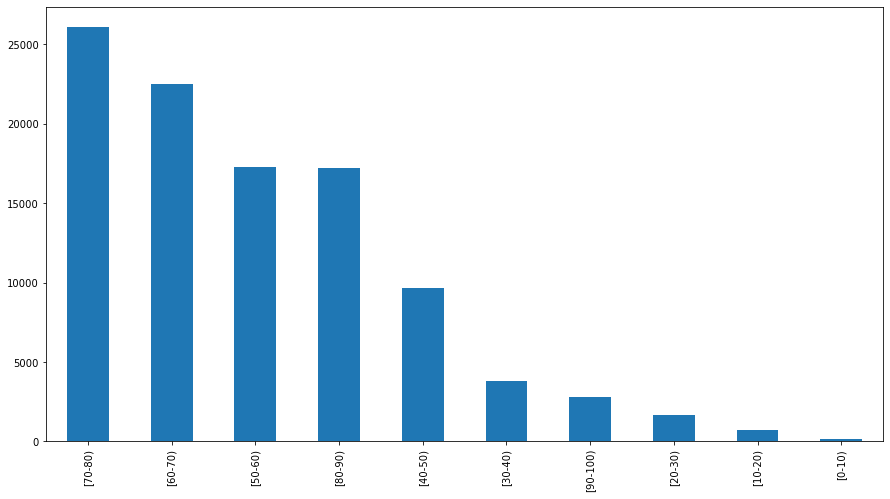
'Caucasian' is more in race category than 'AfricanAmerican' and

Others.

3.Gender: 

Female has high observation than male and unknown/invalid is in negligible.

4.Age:



70-80 age group has maximum and 0-10 are the minimum inpatients.

# Bi-variate analysis:

Since there is no high co-relation between any variable. So, no action can be taken.

# Bivariate analysis with categorical variables:

# Race and Readmitted:

# 

# As the Race does not show any significant variations on the readmission obseravtions. There is no impact inferenced.

# Gender and Readmitted:

# 

# As the distribution of Gender on the target variable is same so the Gender don’t have any impact on target.

# Age and Readmitted:

# 

# As the Age does not show any significant variation over the different target variations. It don’t have any significance.

# Payer Code and Readmitted:

# 

# As the Payer\_code does not show any significant variation over the different target variations. It doesn’t have any significance.

# Feature Engineering for categorical variables:

# Data contained 22 columns on medication and had categorical values, for the sake of simplicity we calculated the sum of the medication category for each record and converted the 22 columns into 4 columns.

# For Readmitted columns we have convert 3 categories (>30, <30, No) into 2 categories(Yes, No).