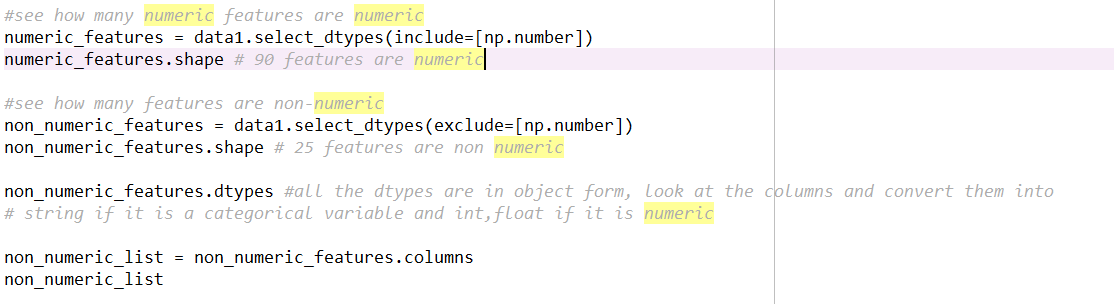
**Data Cleaning**:

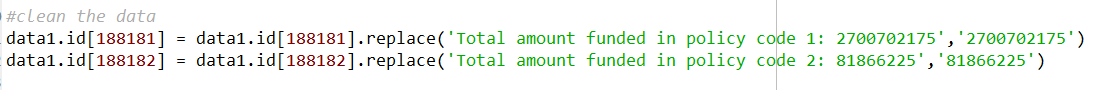
#read the dataset using pandas read\_csv method

data1 = pd.read\_csv('LendingClub2012to2013.csv',header=1)

data1.describe() is used for describing the dataset, to see the count, standard deviation, mean, minimum, maximum of the variables



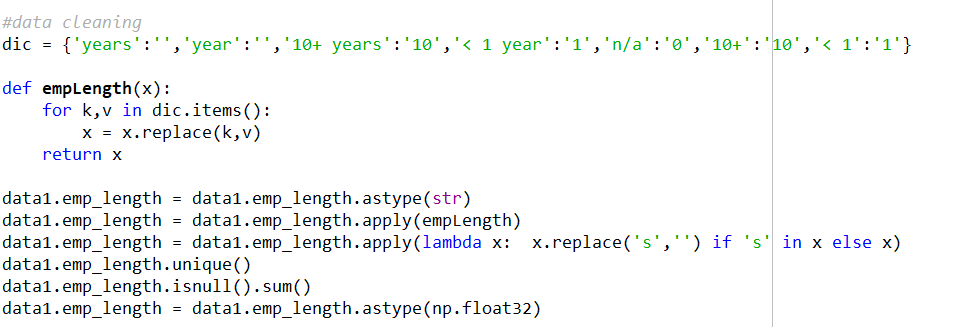
The above code lets us see how many numeric features and non-numeric features are present in the dataset



To clean the string values, replace method has been used



Pandas.isnull().sum() method has been used to see the count of null values and Null values are filled with ‘other’ string for emp\_title variable



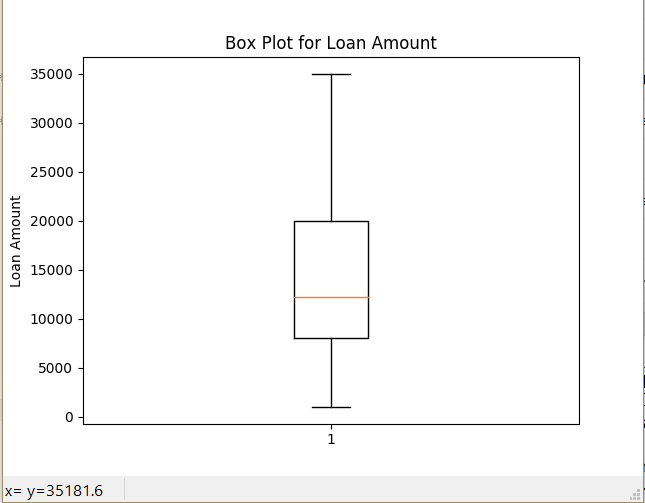
Dataframe.Apply function is used to apply a function on each cell of the dataframe

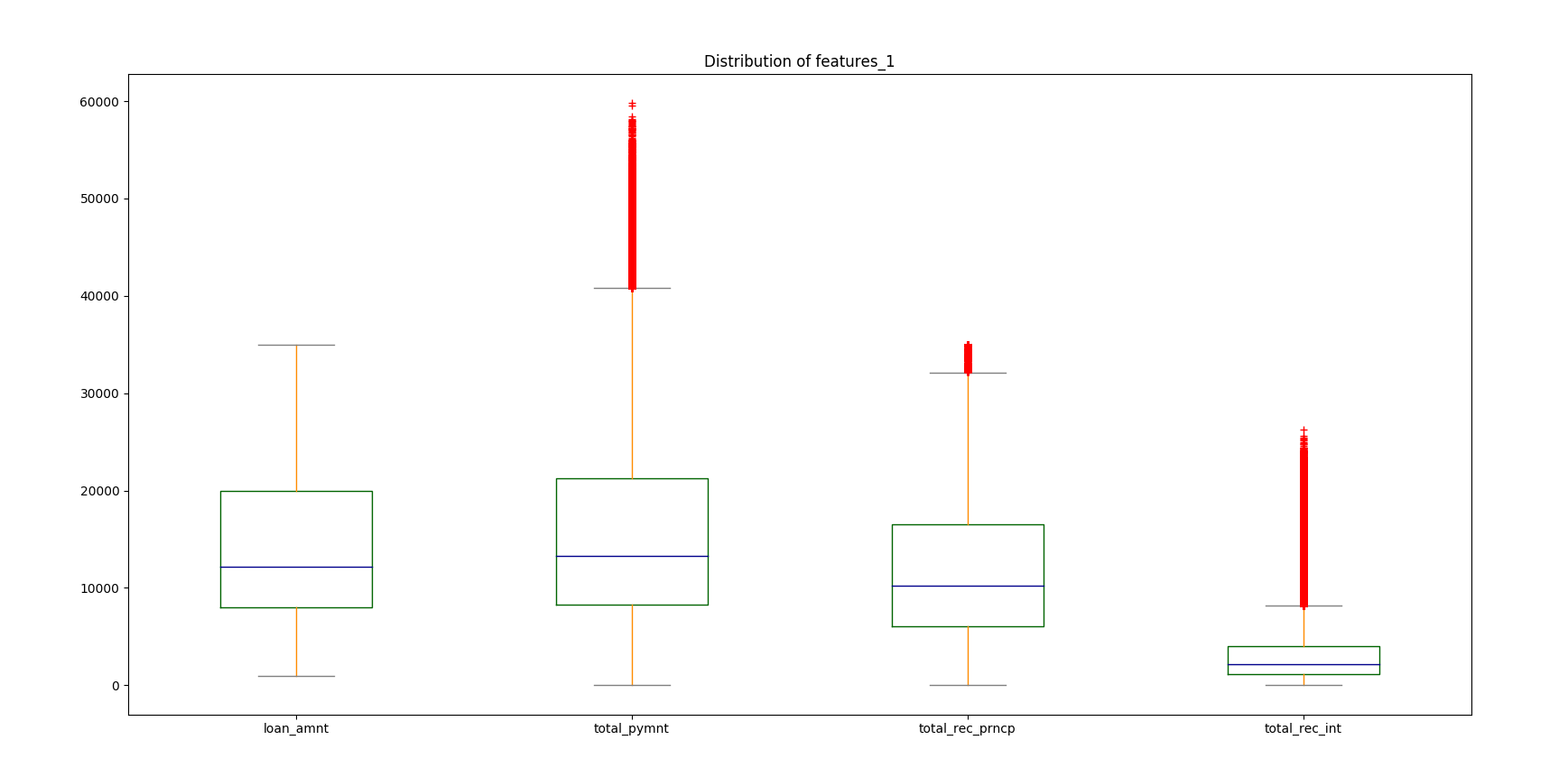
Further cleaning is done by forward filling date columns for the missing values, replacing % sign with empty string

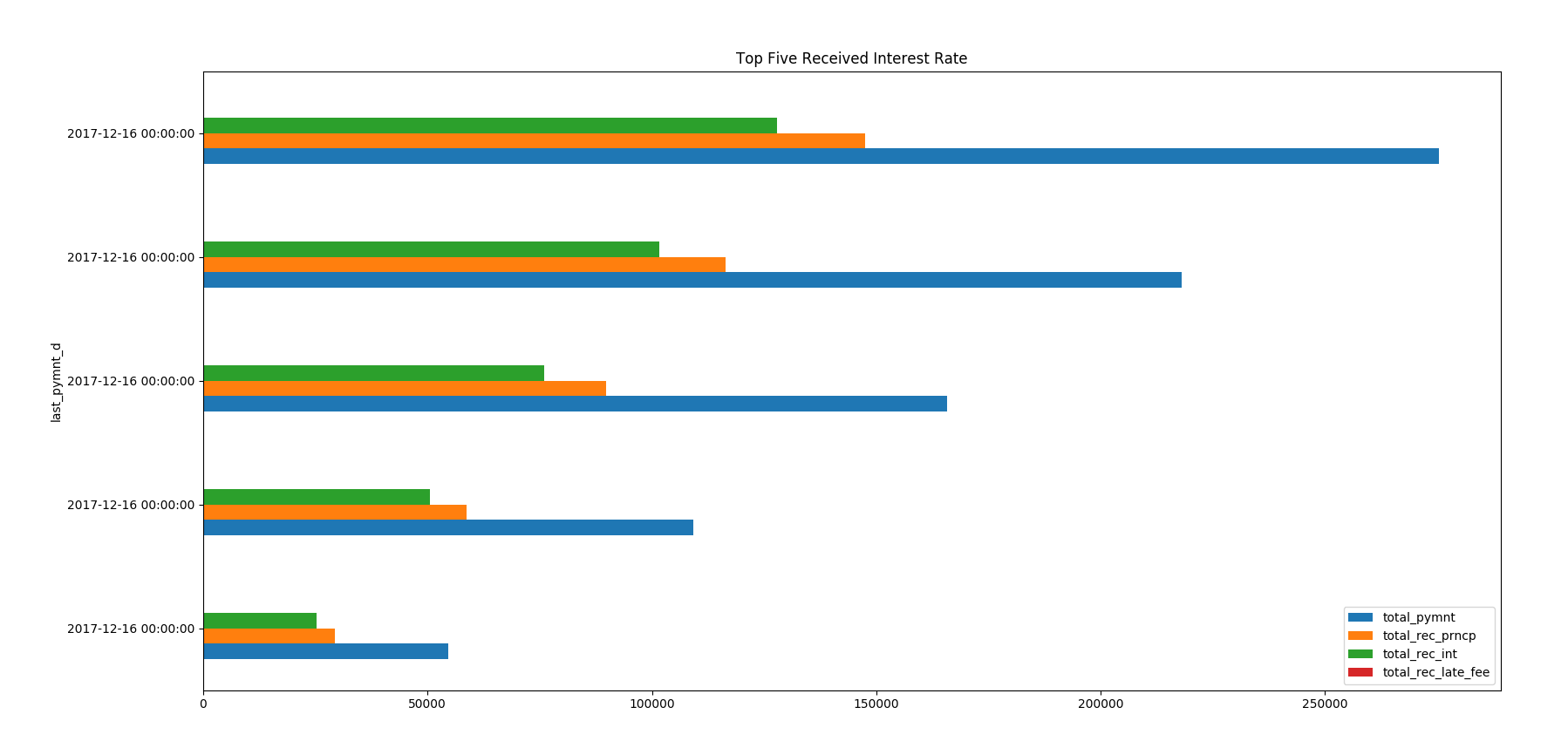
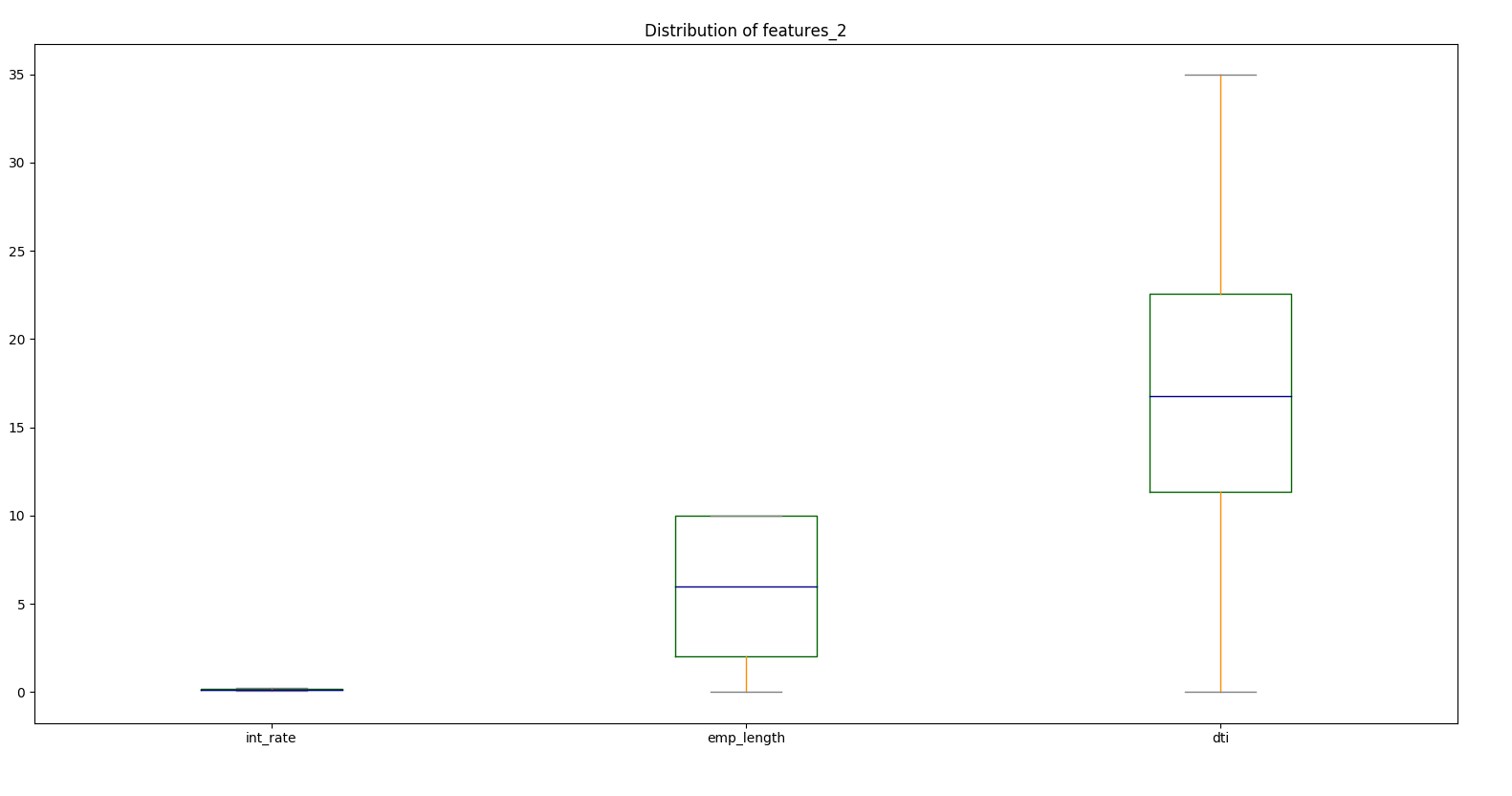
X\_train input values are normalized by using standardscalar class from sklearn.preprocessing

Confusion Matrix is used for evaluating the model by calculating precision, recall, f1 score, accuracy

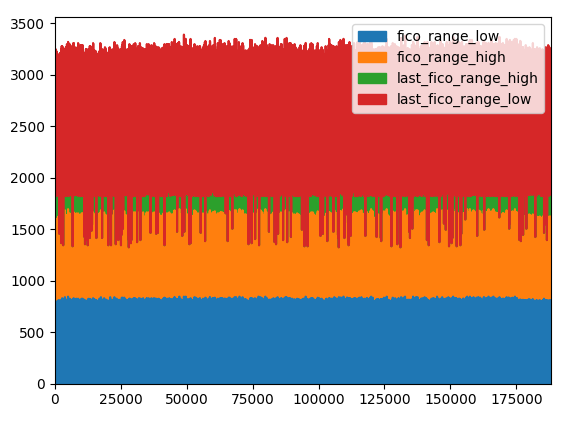
**Visualization**:

From the below figure we can say that median is around 12500, minimum amount is approximately 1000 and maximum is 35000

The below figure shows the distribution of independent variables loan amount,total received payment,total received principal,total received interest.

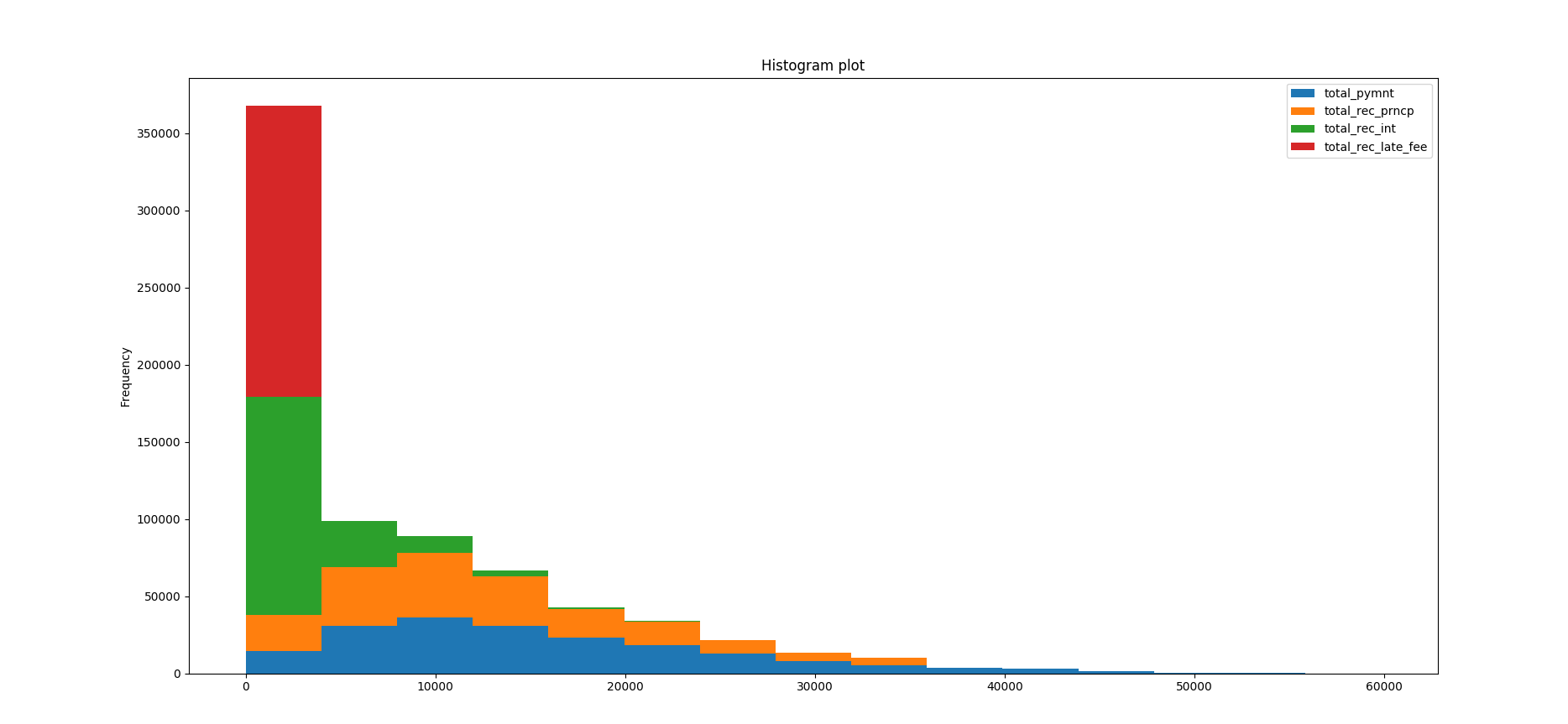
The below figure shows the distribution of independent variables interest rate,employee length, and dti.

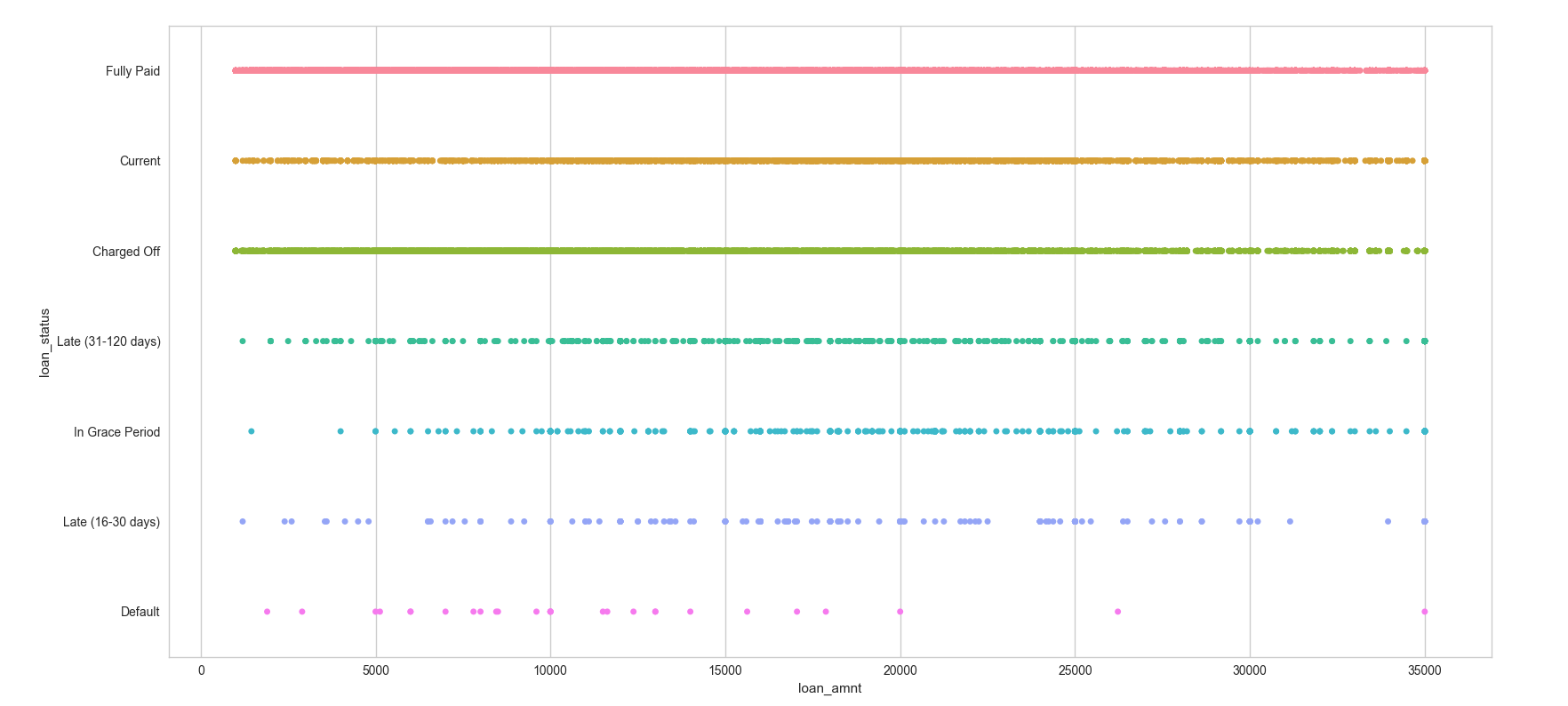
The above plot shows top five received interest rate



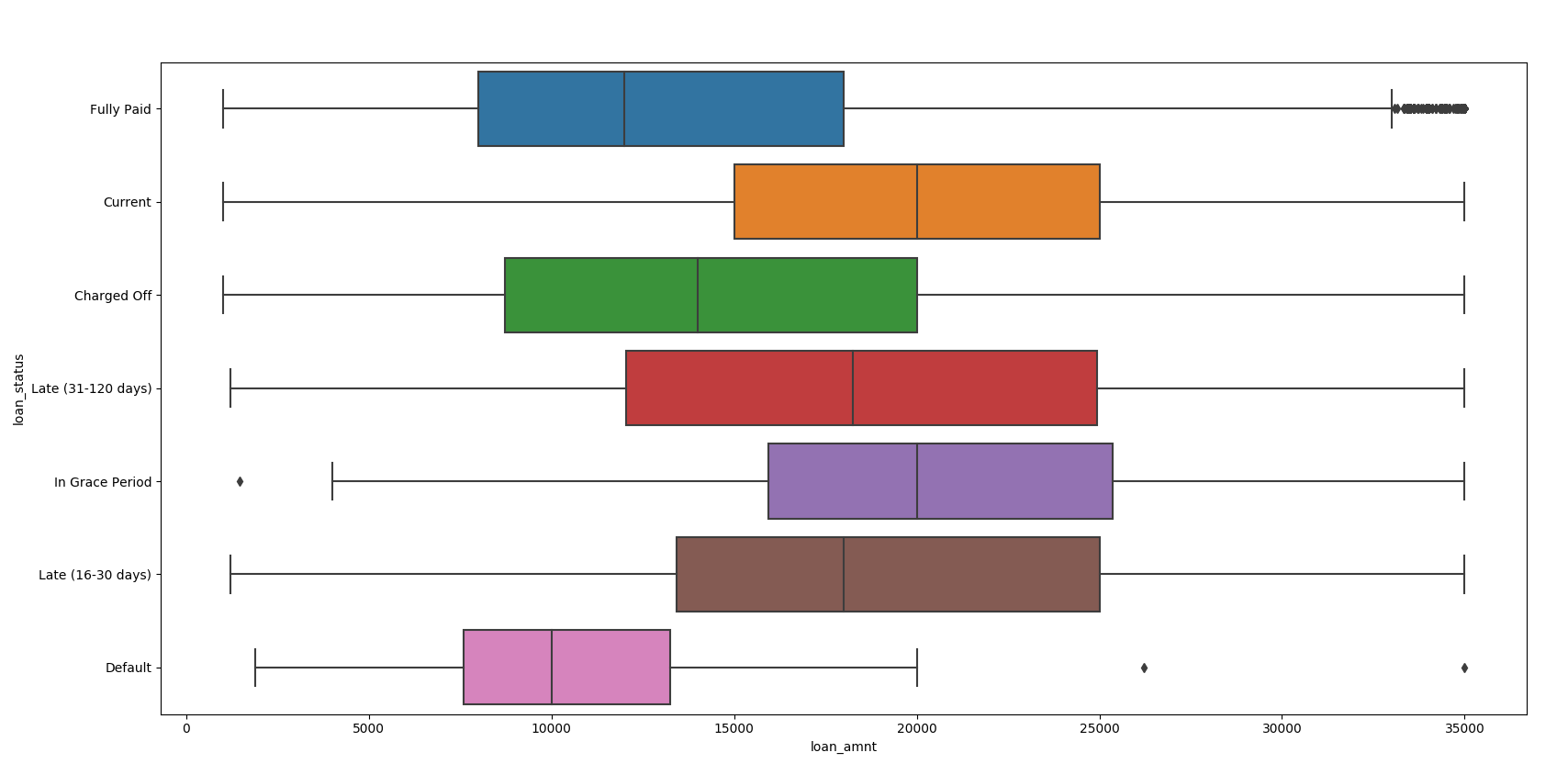
The above plot shows the area distribution of variable 'fico\_range\_low','fico\_range\_high','last\_fico\_range\_high','last\_fico\_range\_low'

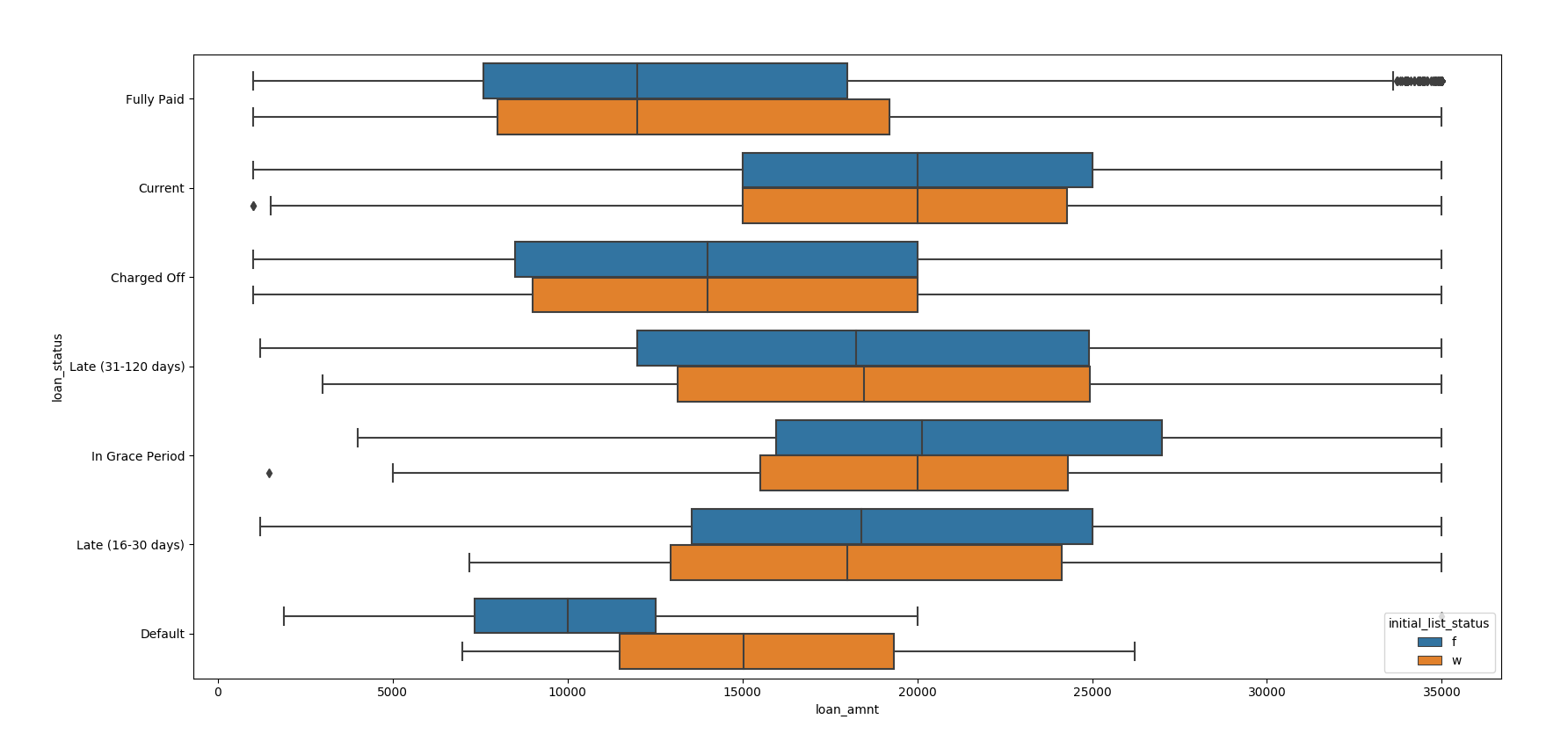
The below plot shows the histogram plot for total payment received, Late fees received to date, Interest received to date, Principal received to date

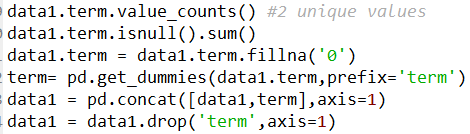




A simple way to show the the values of some quantitative variable across the levels of a categorical variable uses stripplot(), which generalizes a scatterplot to the case where one of the variables is loan status categorical variable and other quantitative variable is loan\_amount







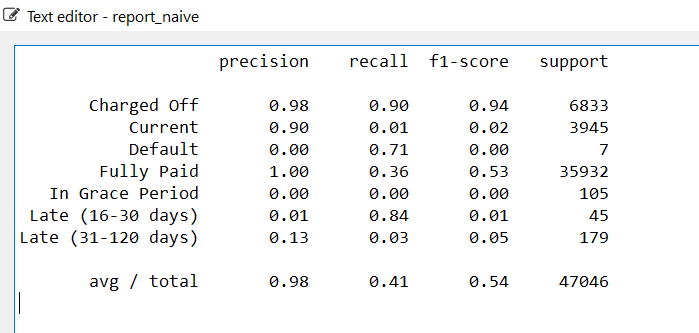
**Adding New Variables:**

New columns have been added to the dataset. Few categorical variables such as term are converted into binary values and added as columns to the dataset. These categorical variables are important foe classification because they have high correlation with the dependant variable loan status, grade or sub grade.

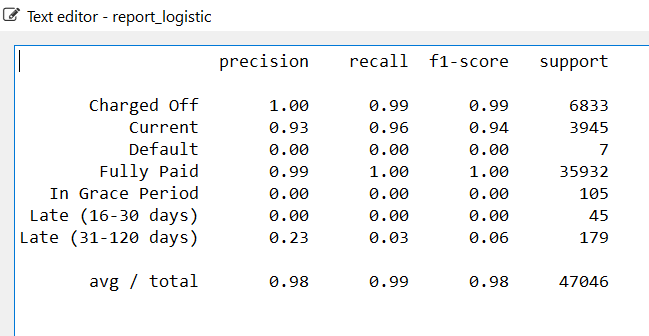
**Forecasting Features:**

The grade column, subgrade column and the loan status column are the best dependent variables for forecasting. As these features have multiple classes and they can be classified using Decision tree, logistic regression, random forest, KNN, and Naïve bayes.

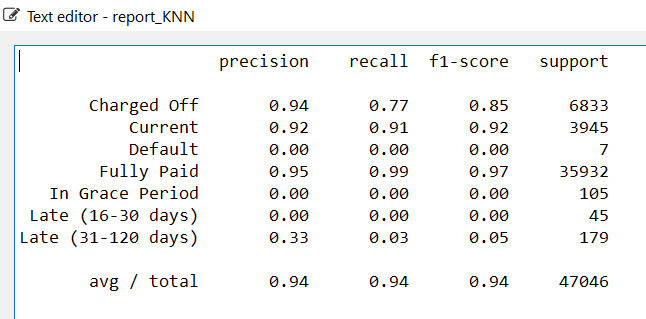
**Evaluating the Model Performance:Naïve bayes**



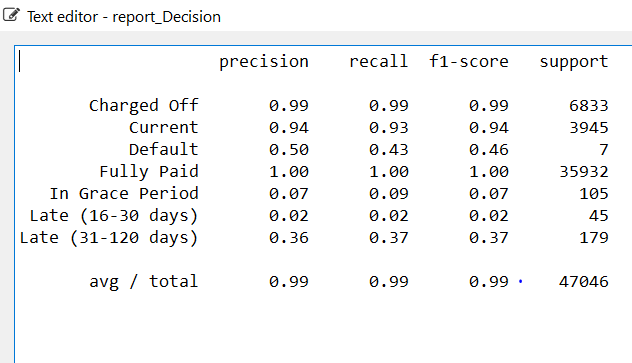
**Logistic regression**

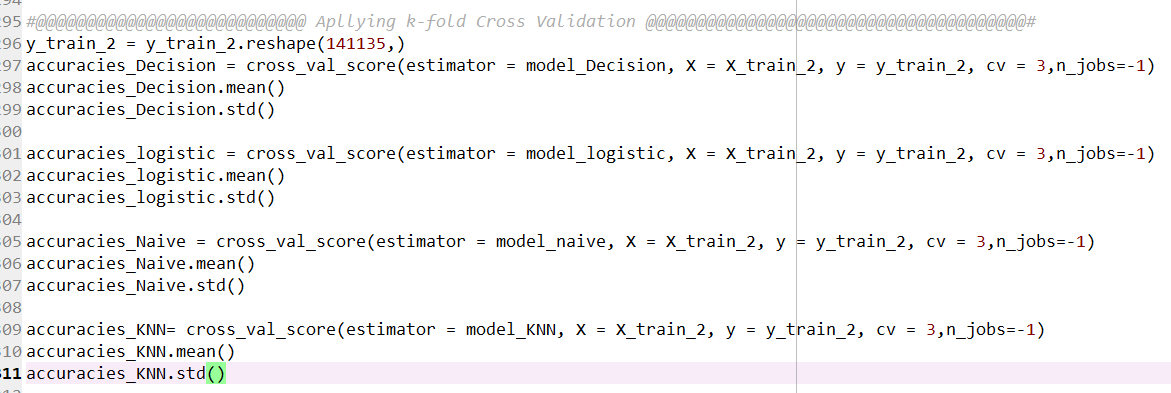


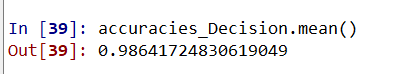
**KNN**



**Decision Tree**













**Conclusion:**

Decision Tree:From the above screenshots we can conclude that decision tree is the best model for predicting the loan status as the acuracy obtained is around 99%