**REPORT**

**INFO 7390 Assignment 2 – Fall 2017 (Lending Club dataset)**

**Professor - Sri Krishnamurthy**

**By - TEAM 03**

**Sonali Chaudhari**

**Madhumitha Prakash**

**Problem Statement**

You are working at a bank and you are considering investing in Lending club. Since there are no standard models, you are expected to build prediction models that will help you predict the interest rates based on various parameters users would input.

**Part 1: Data wrangling and exploratory data analysis 1(a)Data Download and pre-processing:**

**Data Download:**

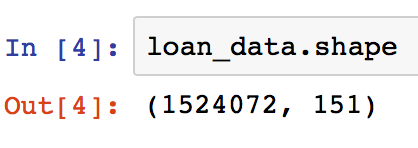
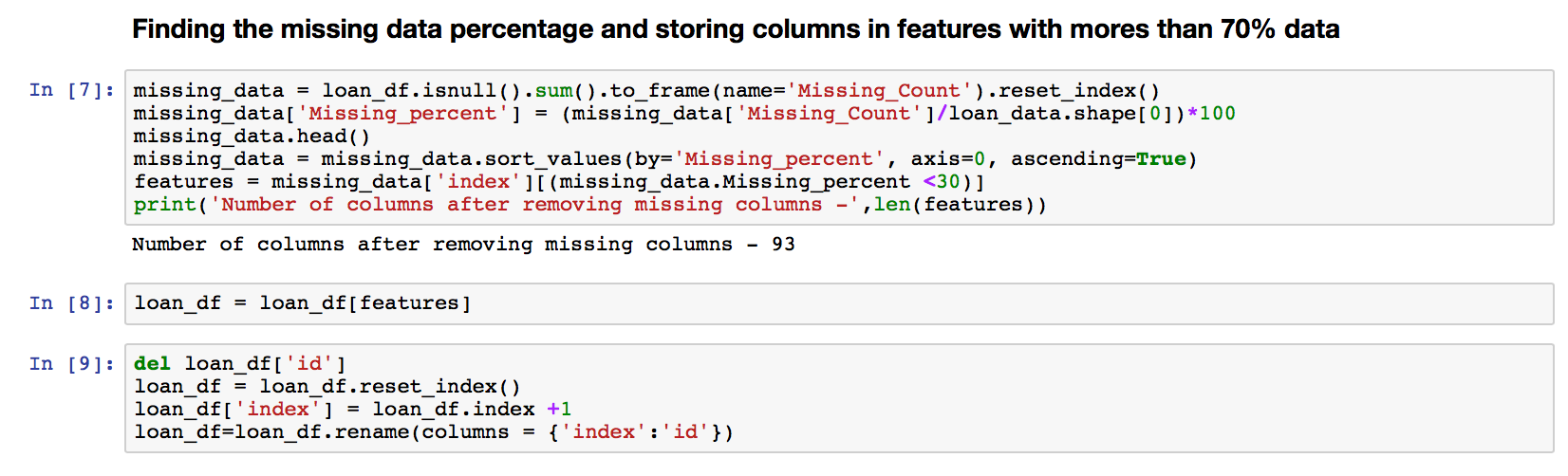
**Steps:**

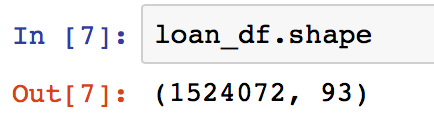
**madhu**

**Data Preprocessing:**

**Loan Data**

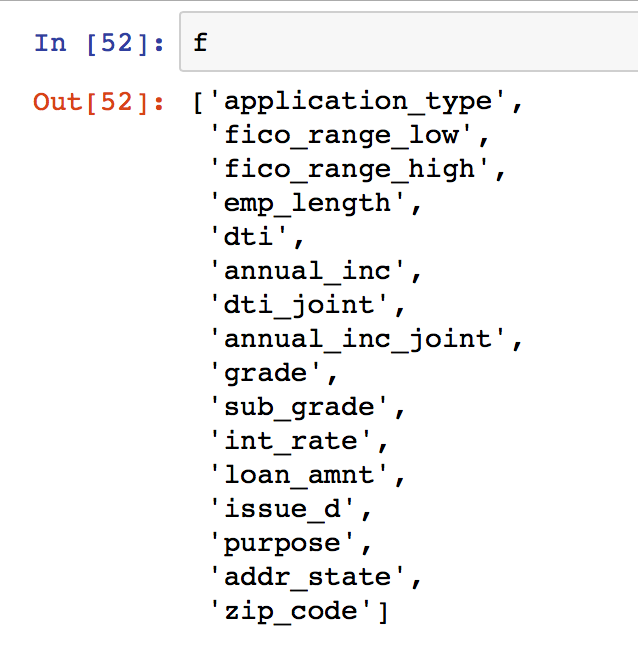
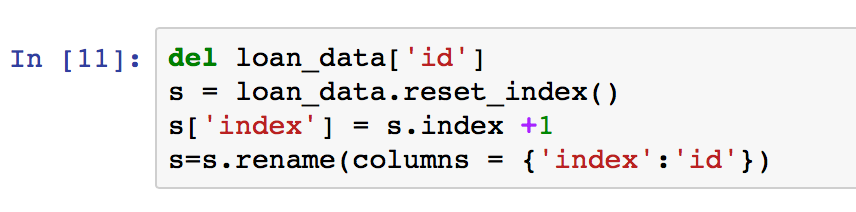
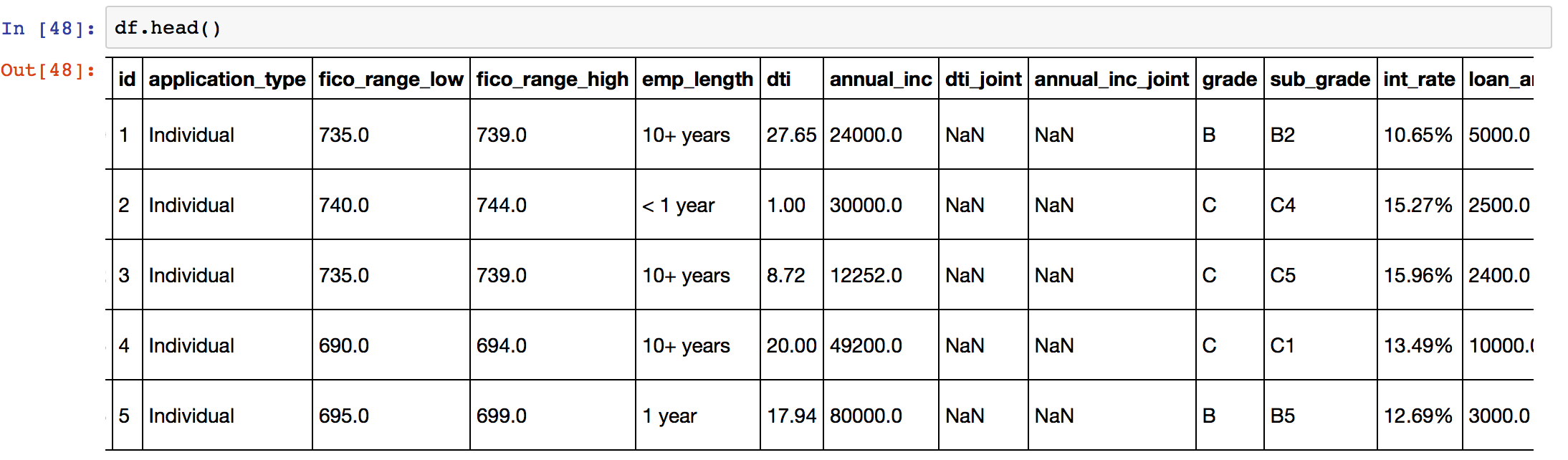
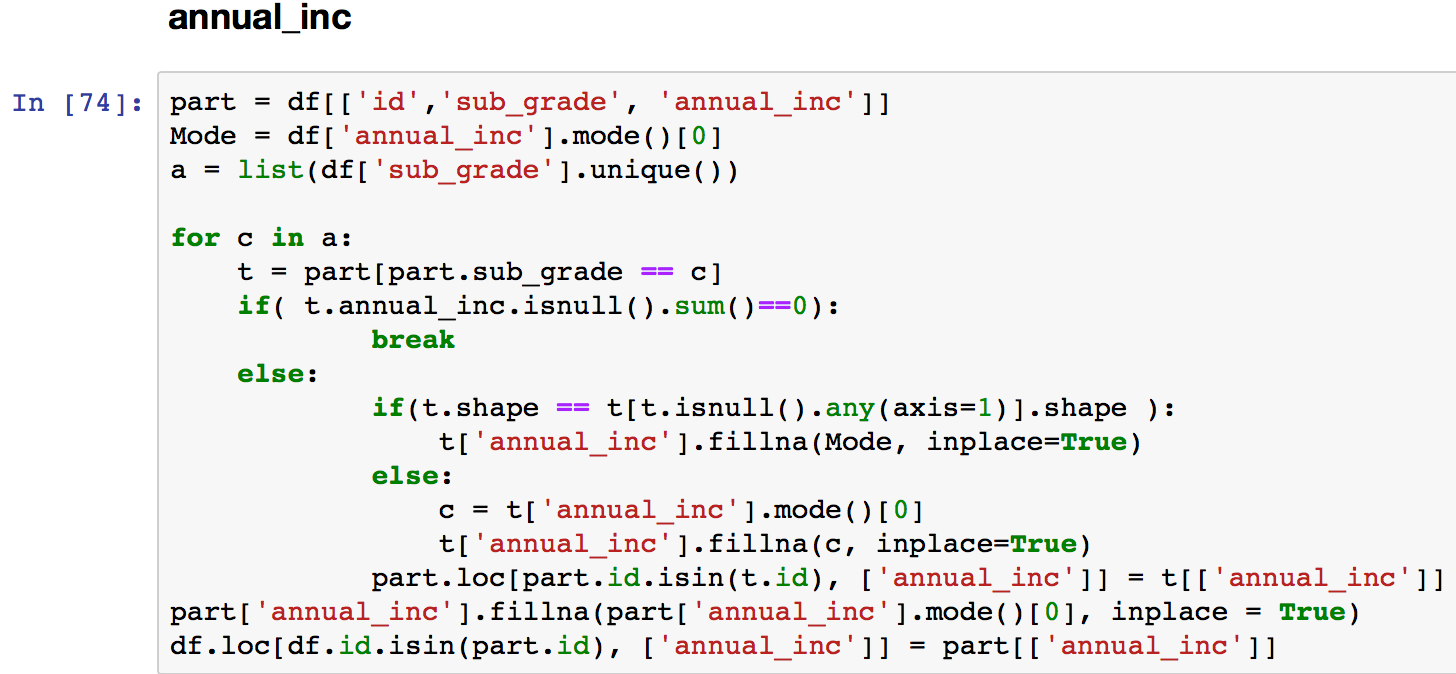
Steps:

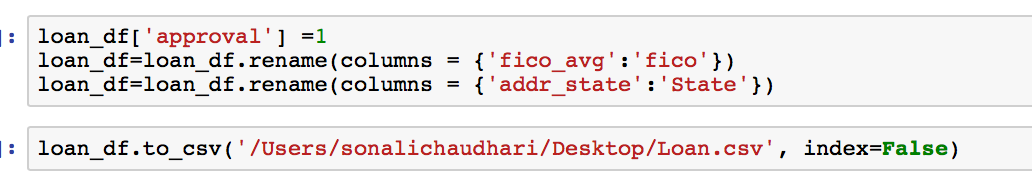
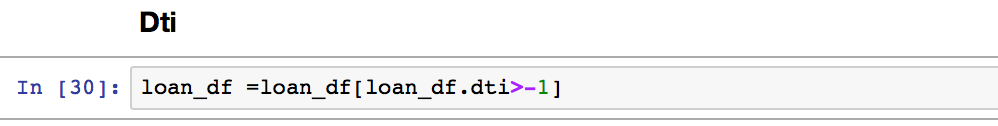
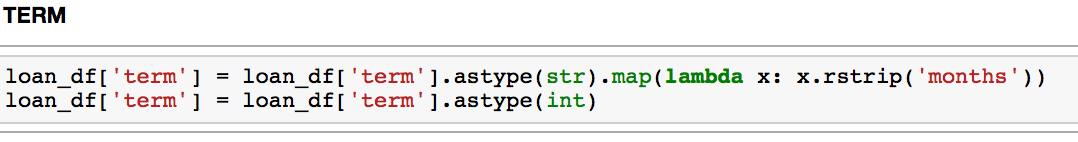
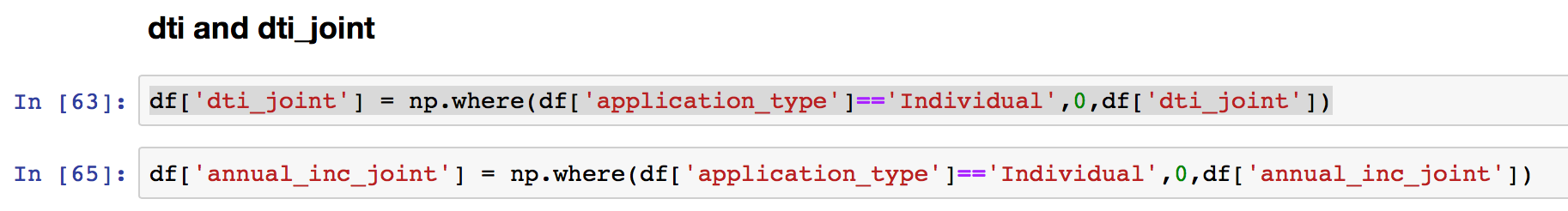
1. Loaded each programmatically downloaded CSVs in separate dataframe.
2. Concatenated each dataframe in a single dataframe ‘loan\_data’.
3. Going through each column and understanding the meaning of each column through the dictionary provided on the LC website ‘LCDataDictionary.xlsx’
4. Getting rid of the columns have more than 70% missing data. During this process we get rid of the column ‘id’ which has all rows null.



1. Out of the 151 columns most of the columns are generated after the loans application procedure has taken place. So these columns leak data from the future. We need columns data that potential customer provided to LC during their application.

**Feature Engineering**

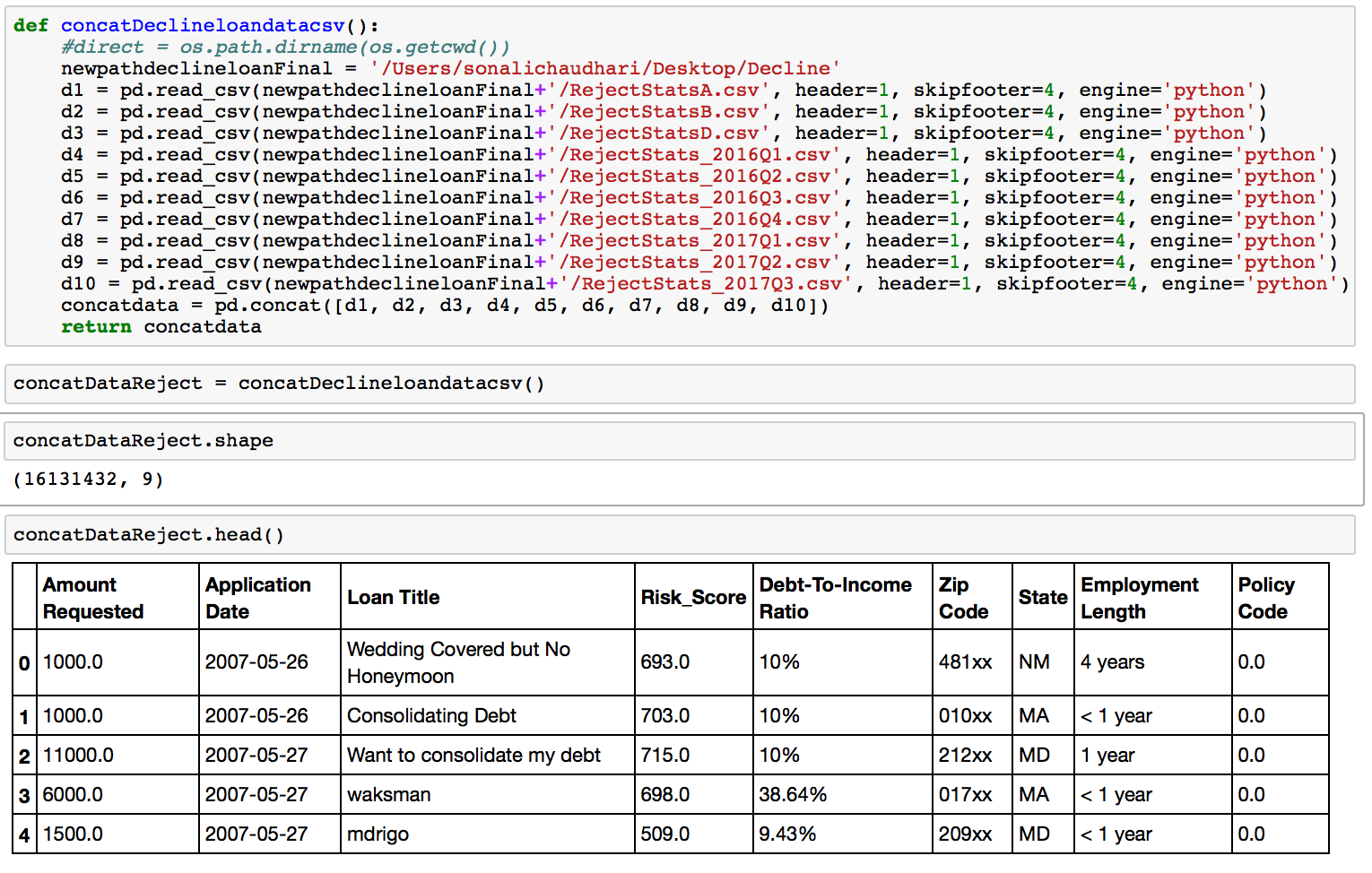
1. Created a list ‘f’ a features that are important for and received during initial borrower’s application.
2. Created a dataframe ‘df’ consisting of all the features in the list ‘f’ and generated a column ‘id’ for uniquely identifying each entry.
3. Retrieved the count of nulls in each column in ‘df’ to deal with missing data in every column.
4. Getting rid of the row having most of the columns null.
5. Getting the dataframe to have consistent formats for each column entry and logically replacing the NaN values for some categorial columns
   1. The columns like emp\_length, int\_rates, zipcode, dti… are of type object and has no consistent format
   2. Replaced dti\_joint, annual\_inc\_joint null values to ‘0’ for the application\_type ‘Individual’ since those application has no co-borrower and hence dti\_joint, annual\_inc\_joint becomes 0.
   3. Filled the remaining null values for annual\_inc with the mode annual\_inc value in a particular sub\_grade column that the row belongs to.
   4. Similarly for dti\_joint column
   5. There are columns ‘fico\_range\_high’ and ‘fico\_range\_low’; so generated a new column ‘fico\_avg’ having the average of both the column values
   6. Change the datatype of necessary columns and adding new column approval having value 1 for the loan data.

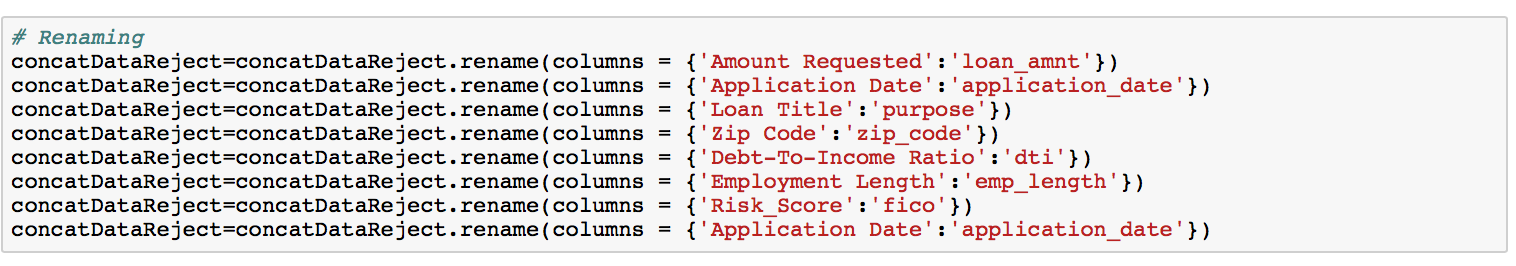


**Decline Data**

Steps:

1. Loaded the csvs in each data frames and then concatenating it one data frames.
2. Clean the data
3. Changing the column names and datatype
4. Adding new column ‘approval’ having values ‘0’ for the decline data
5. Exporting to Csv

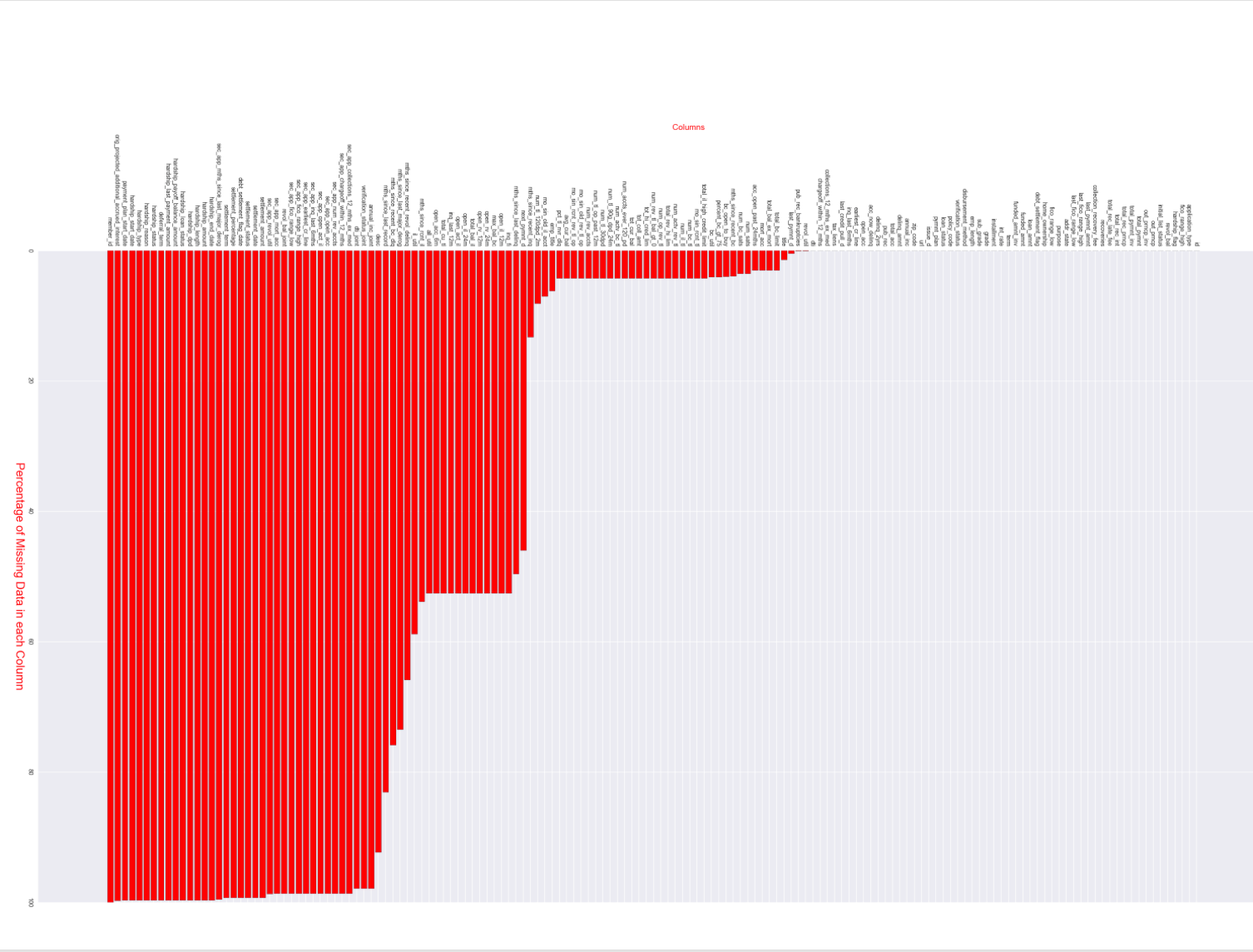
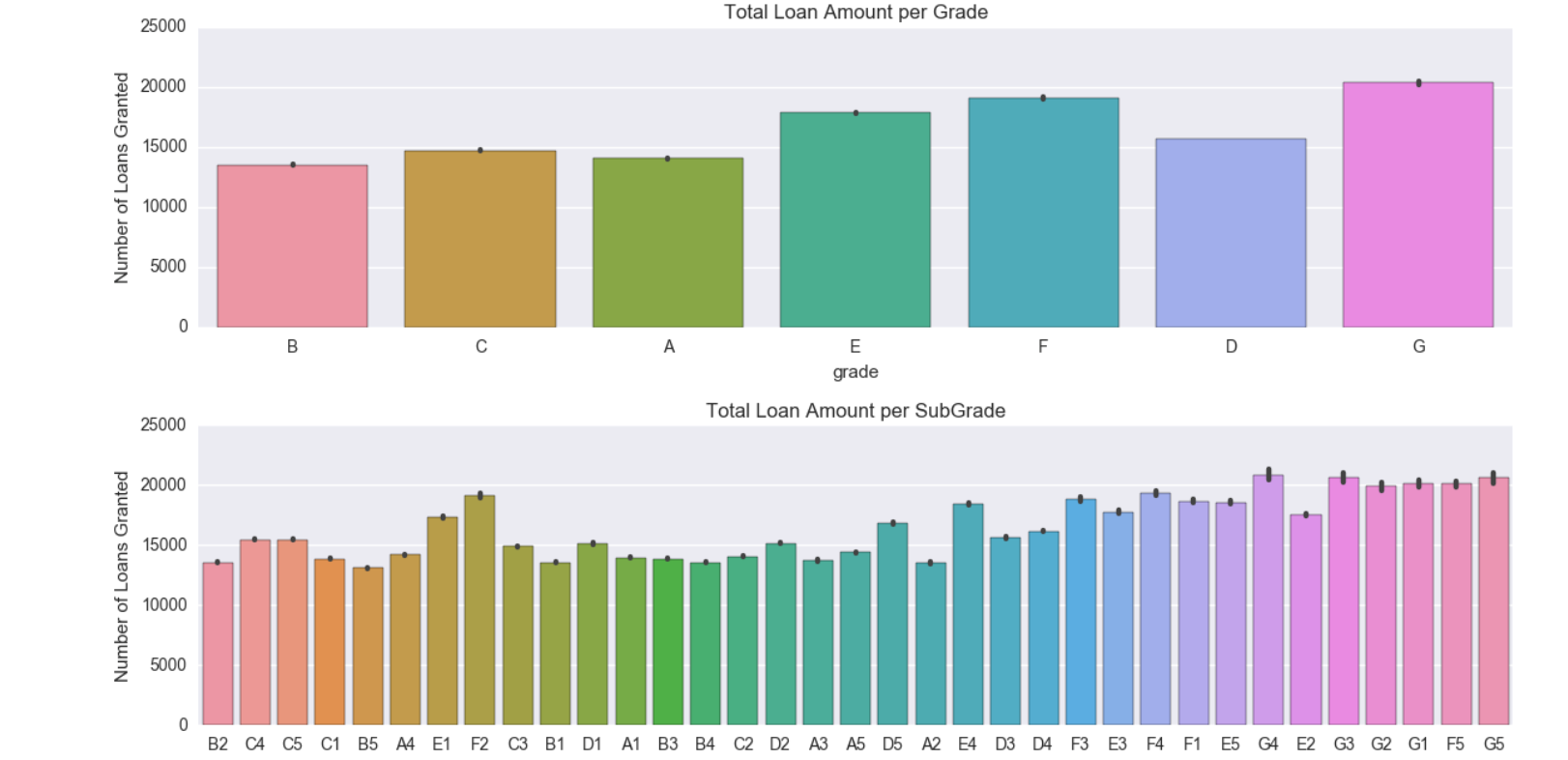
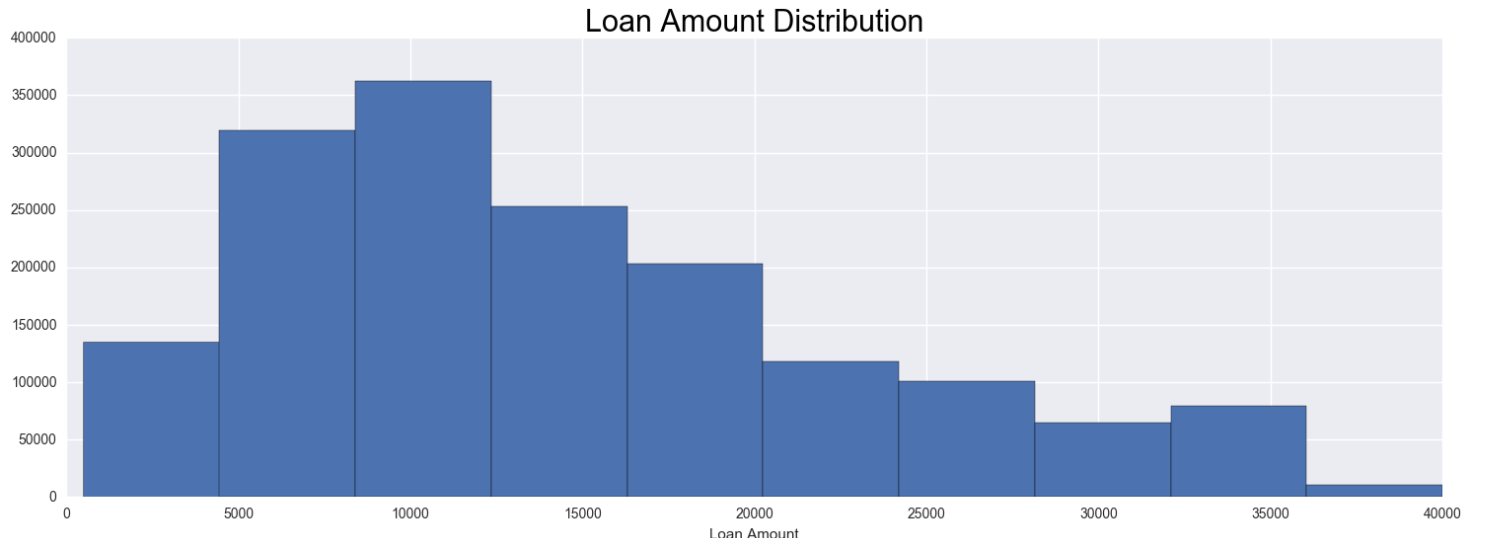
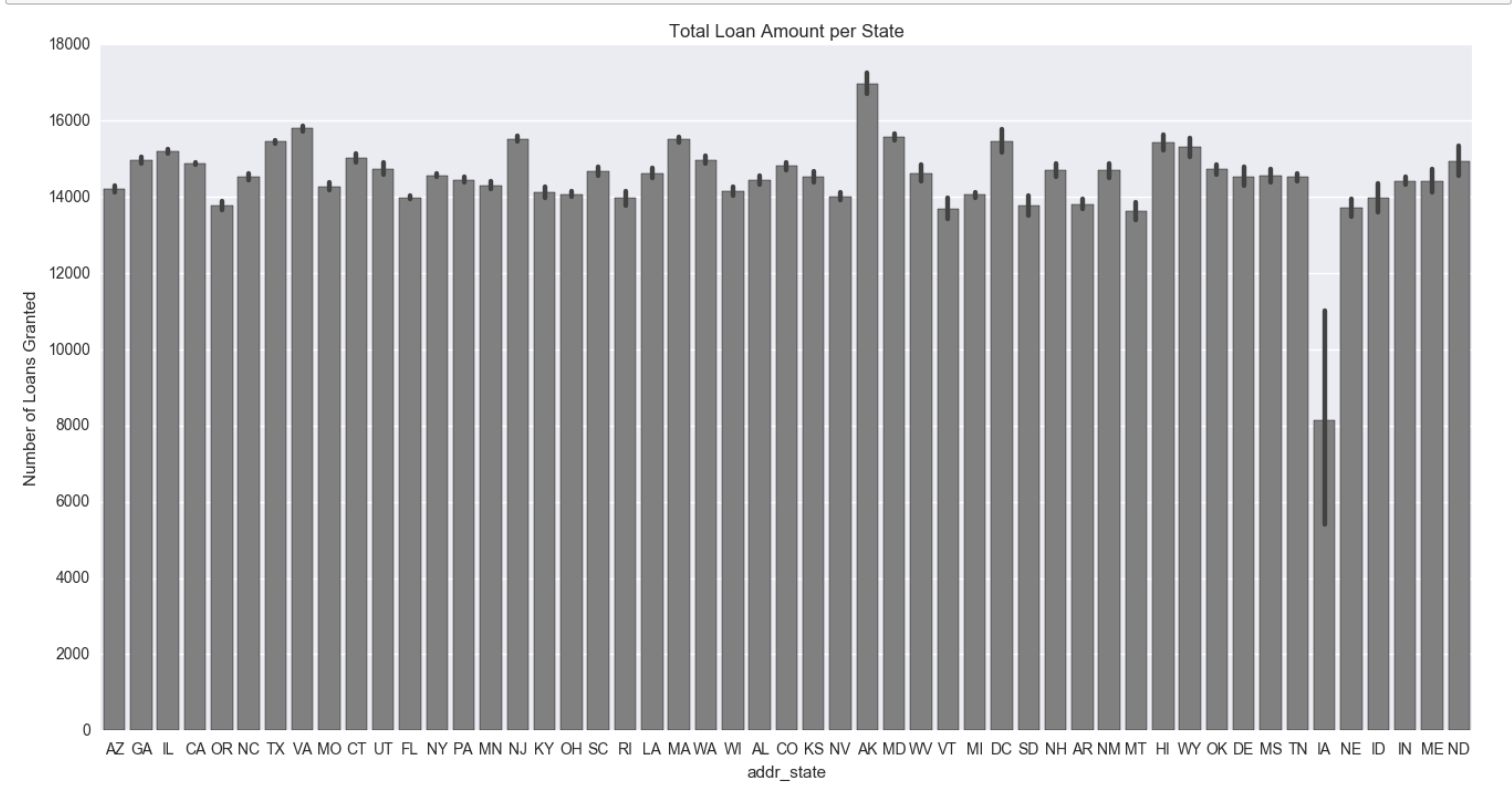
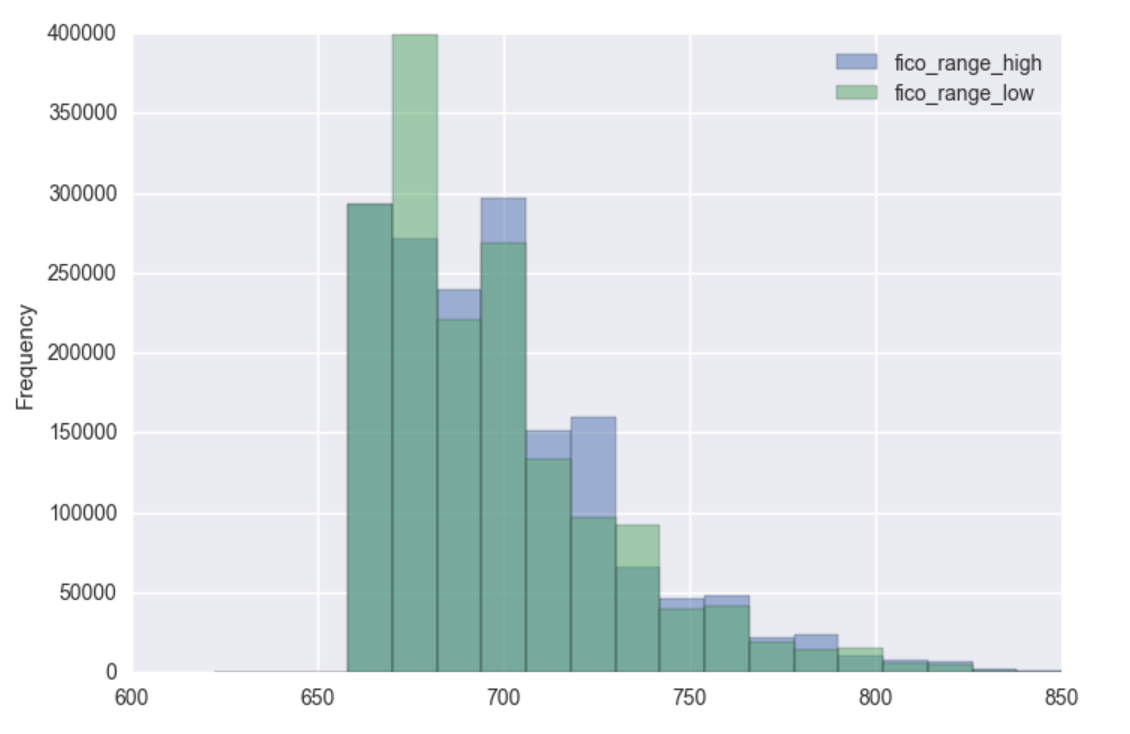
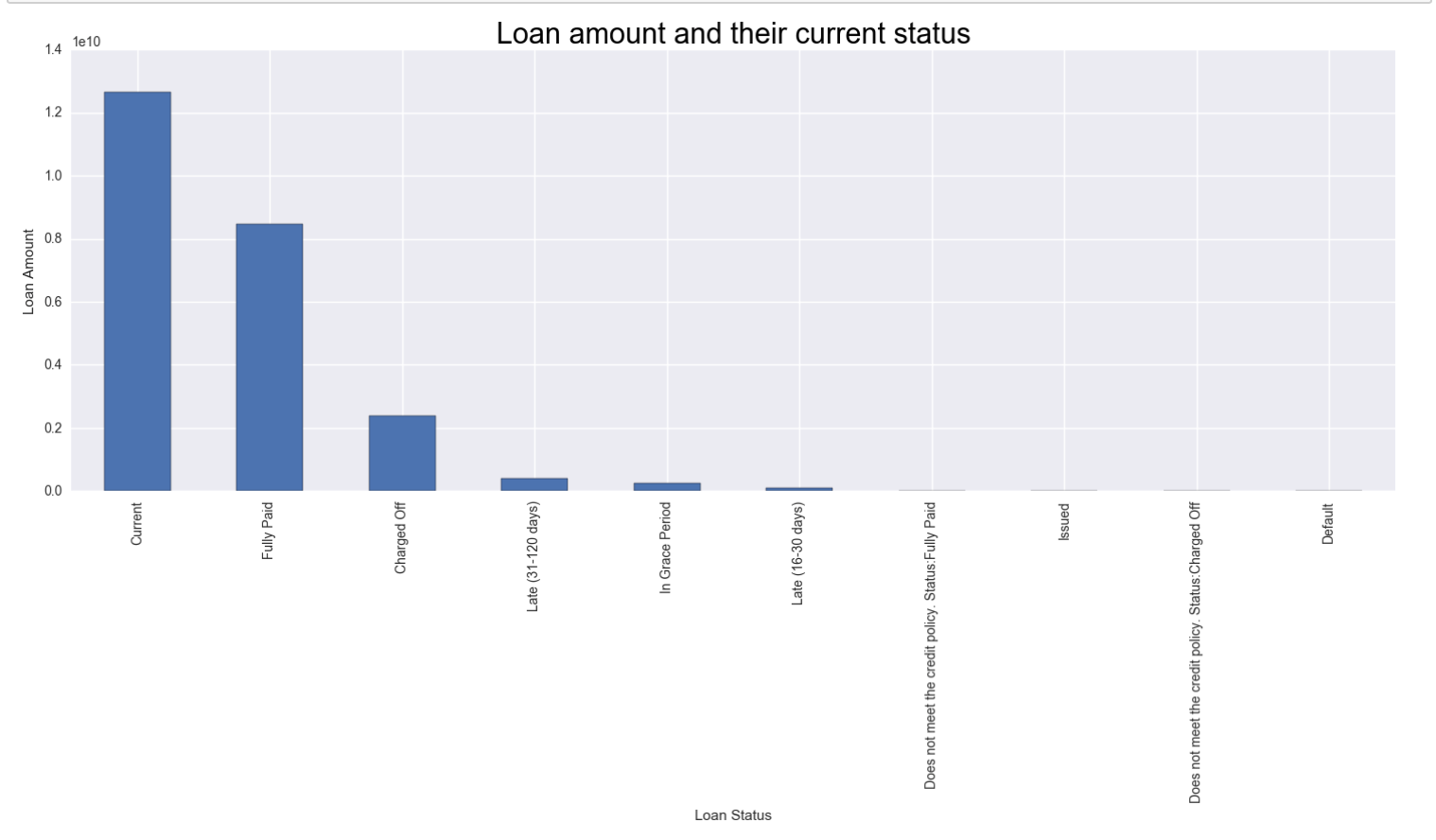
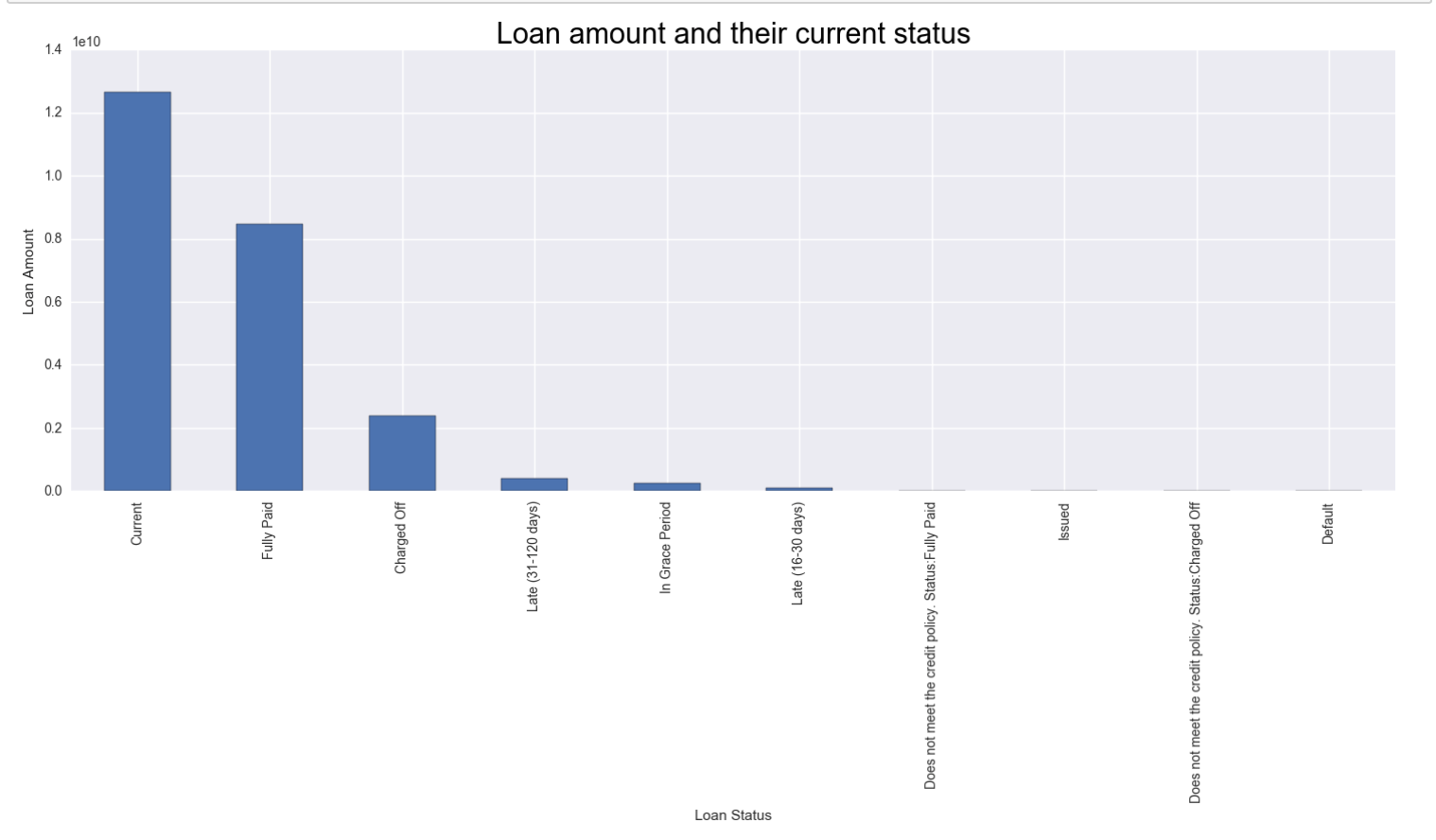
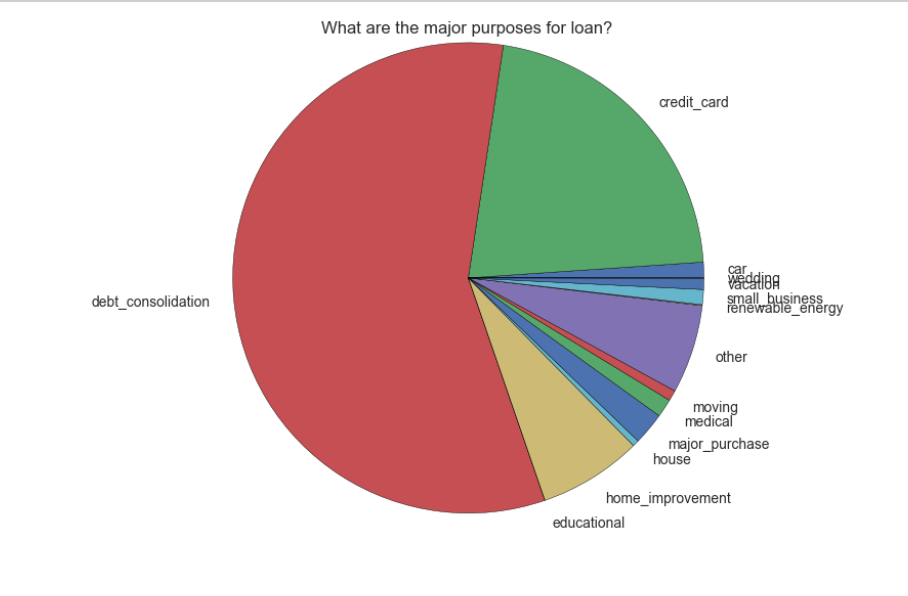
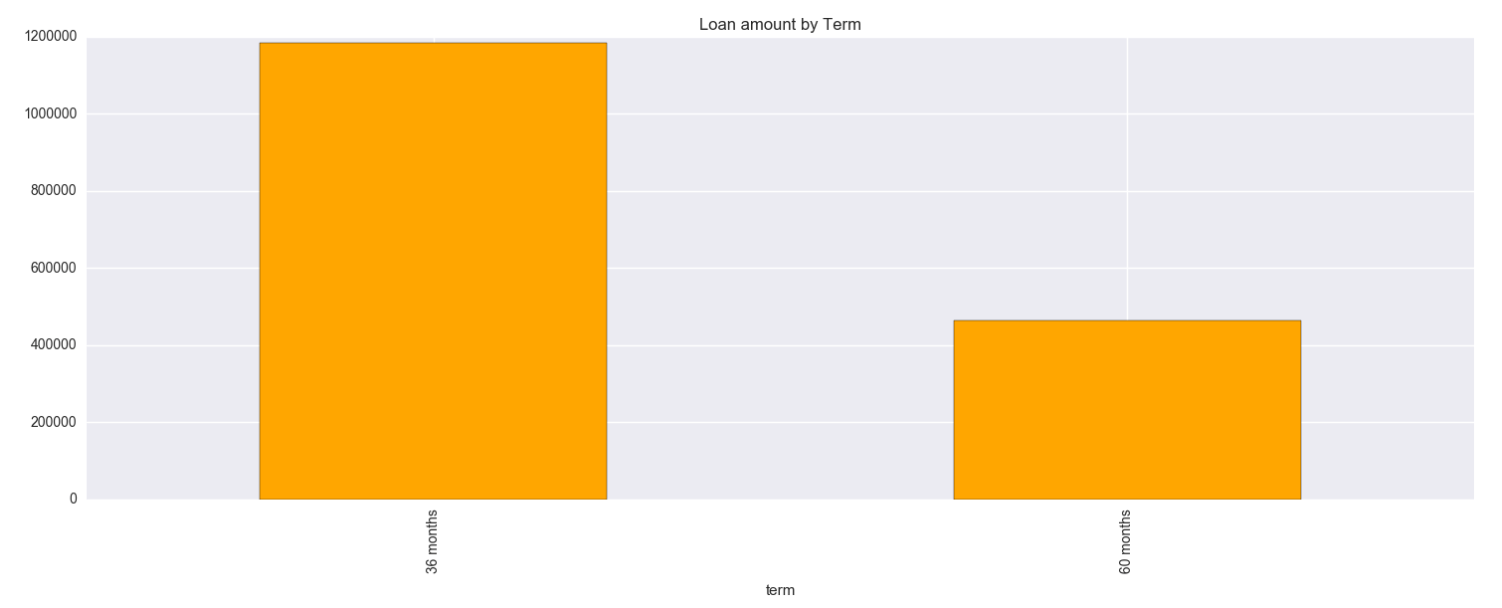
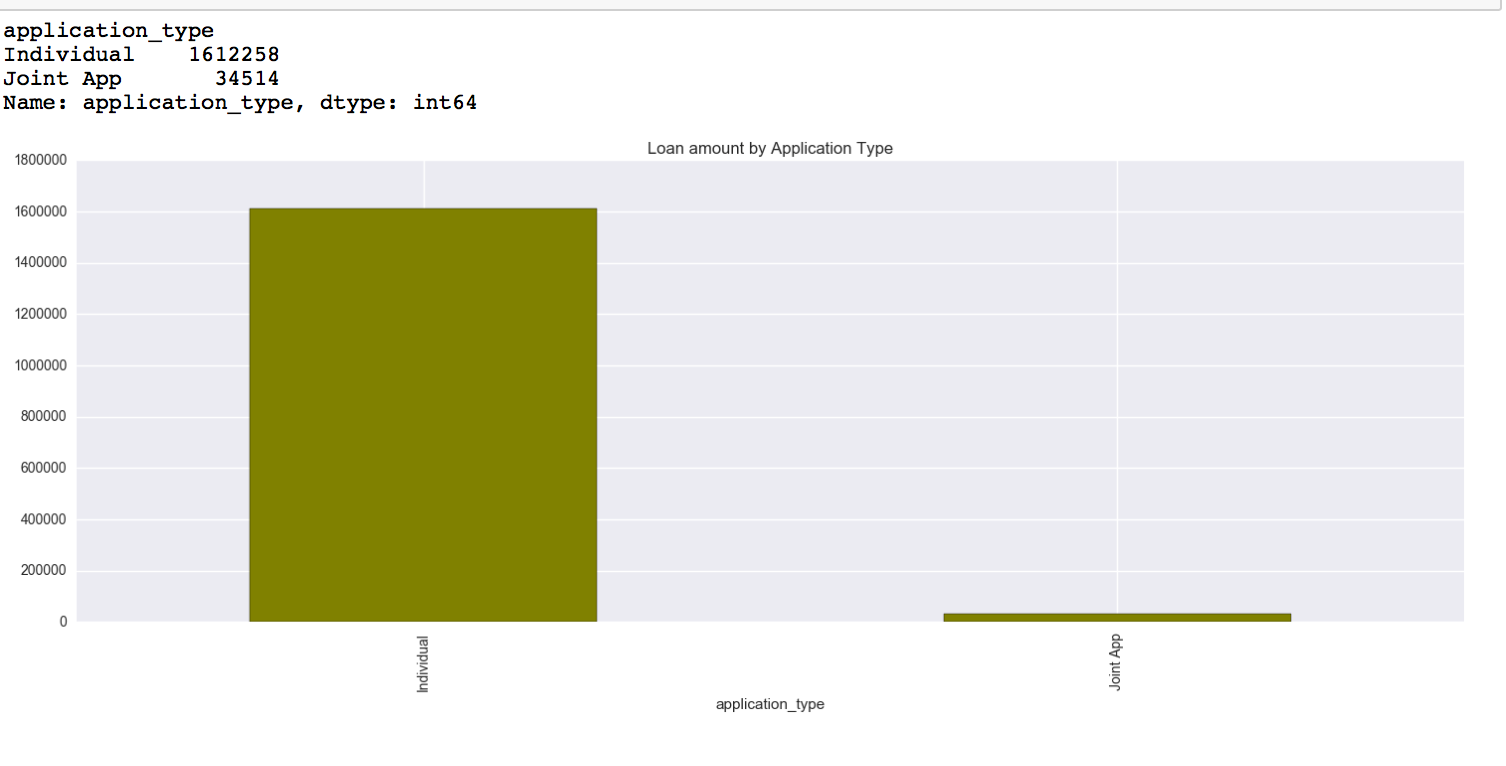
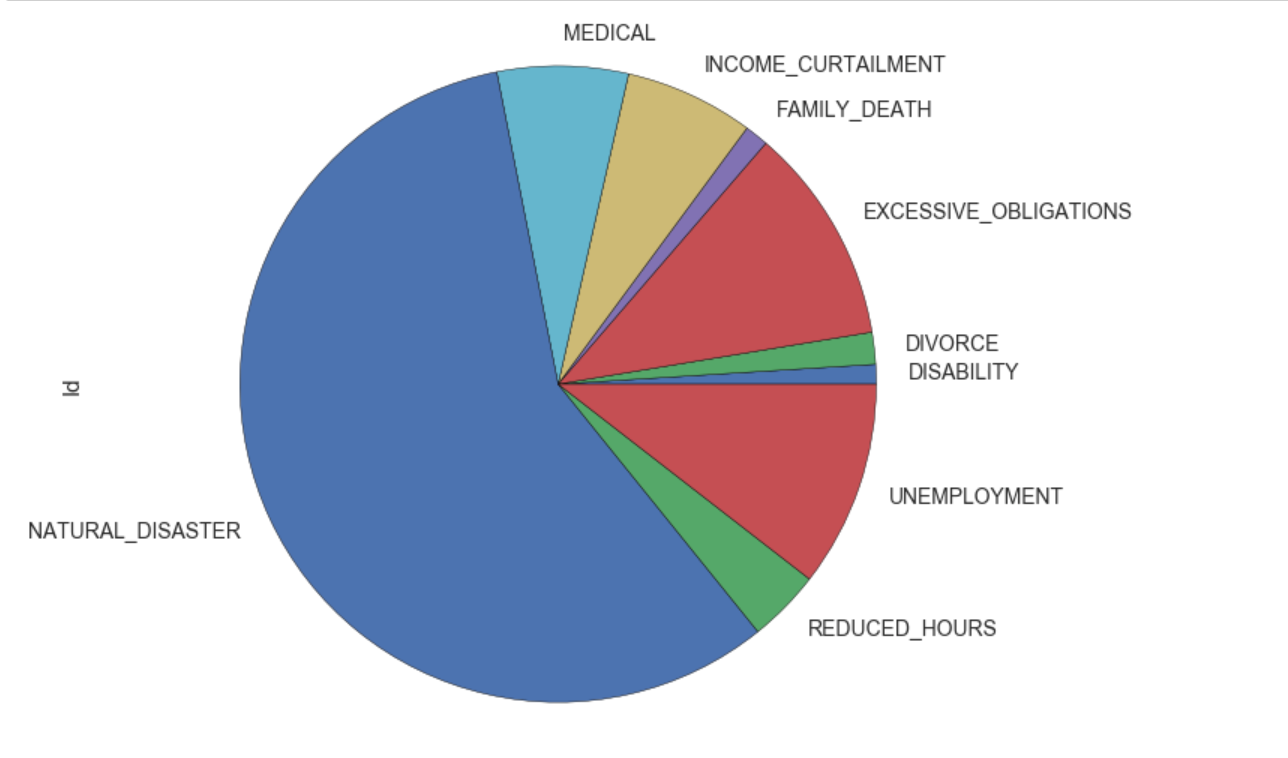
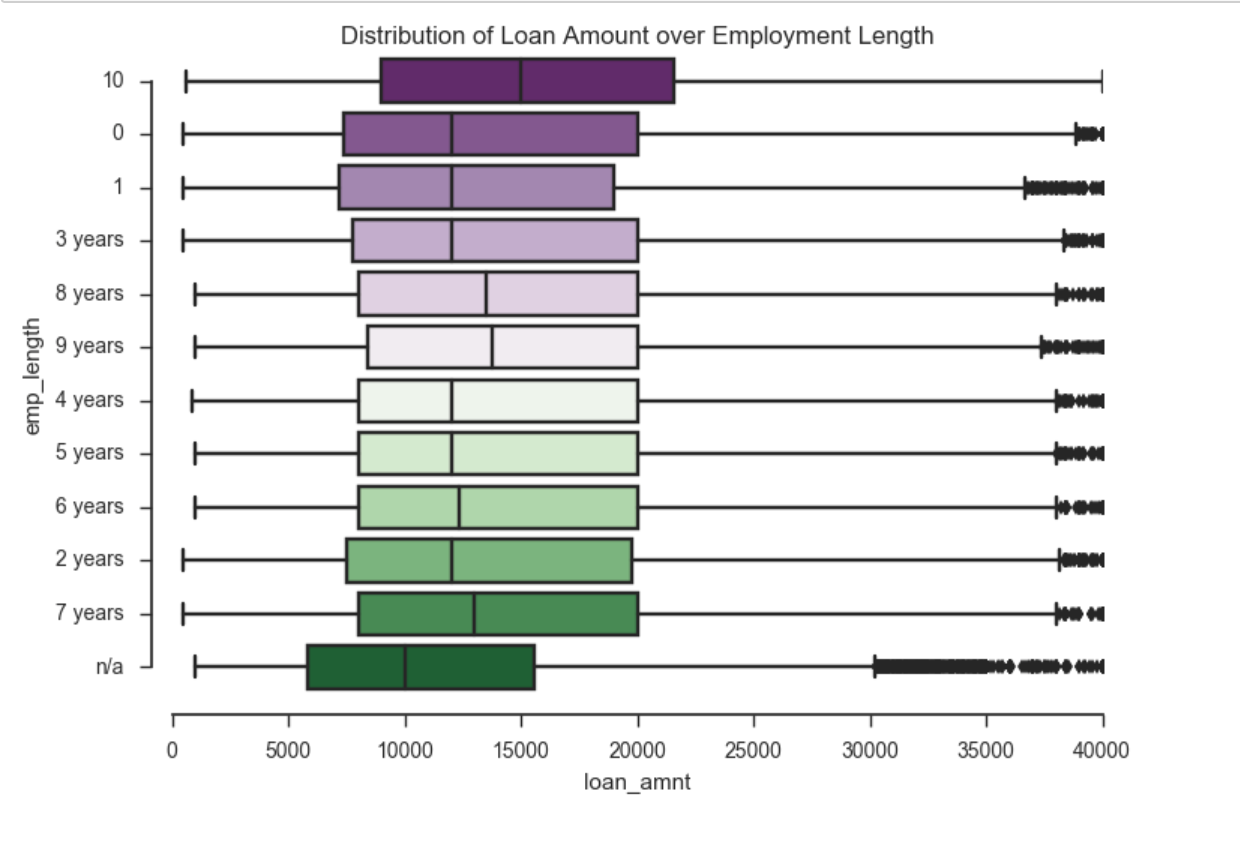
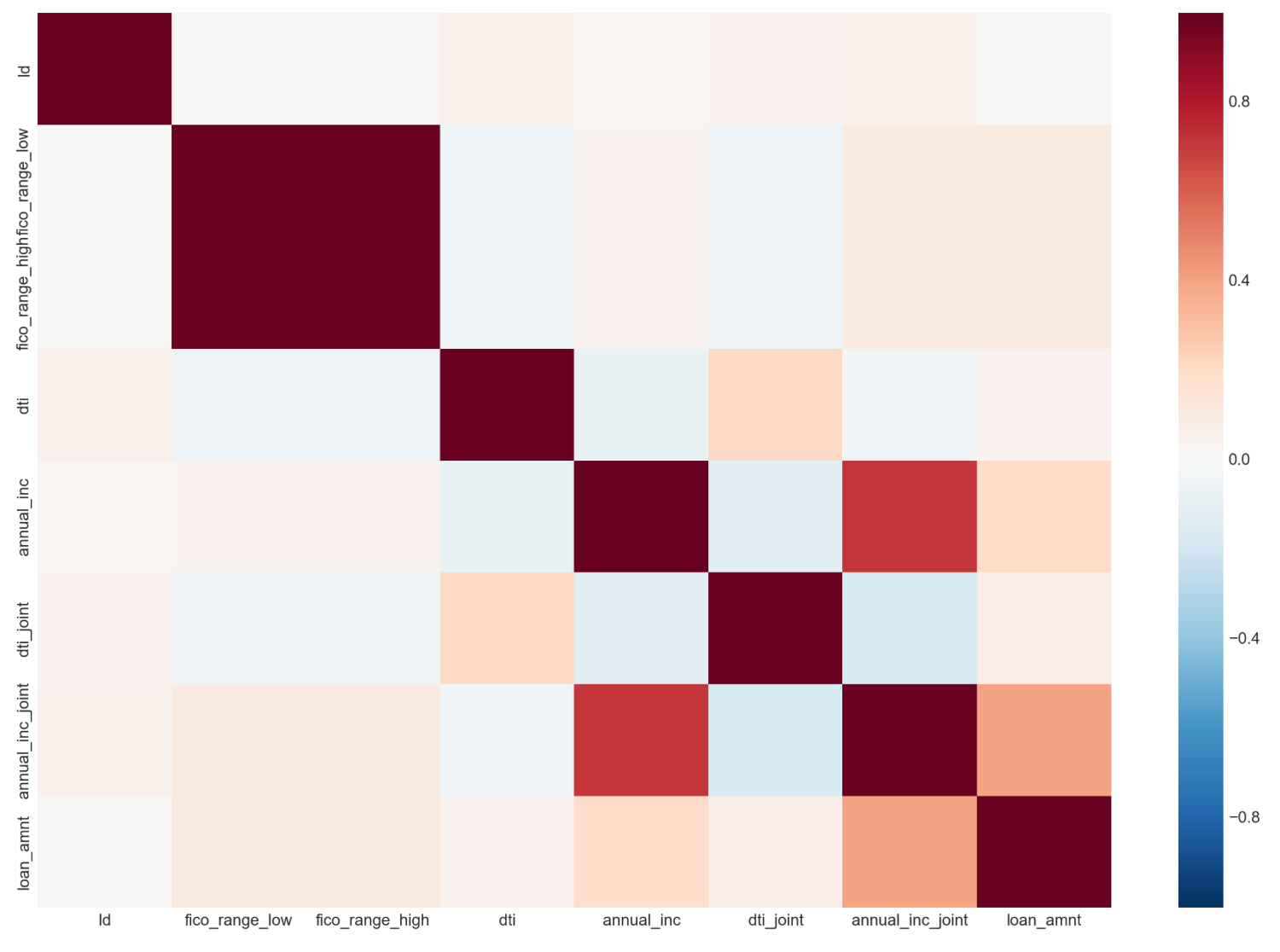


****

**1(b)Exploratory Data analysis:**

We performed exploratory data analysis the loan data and decline data separately and few EDAs for combine data.

**Loan data**

1. Missing data Analysis
2. How many loans have been approved per grade and sub grade?
3. Distribution of Interest Rate
4. Distribution of Loan Amount
5. Loans Granted per State
6. Zico high and fido low analysis
7. Loan Amount by loan status
8. Distribution of Loan amount over purpose
9. Distribution of Loan amount over term
10. 
11. Distribution of Loan amount over application type
12. Loan Amount count per Hardship Reason
13. Distribution of Loan Amount over Employment Length
14. Heat Map of columns details provided during application for Loan

**MISSING DATA**

**madhu**

**Part 2: Building and evaluating models**

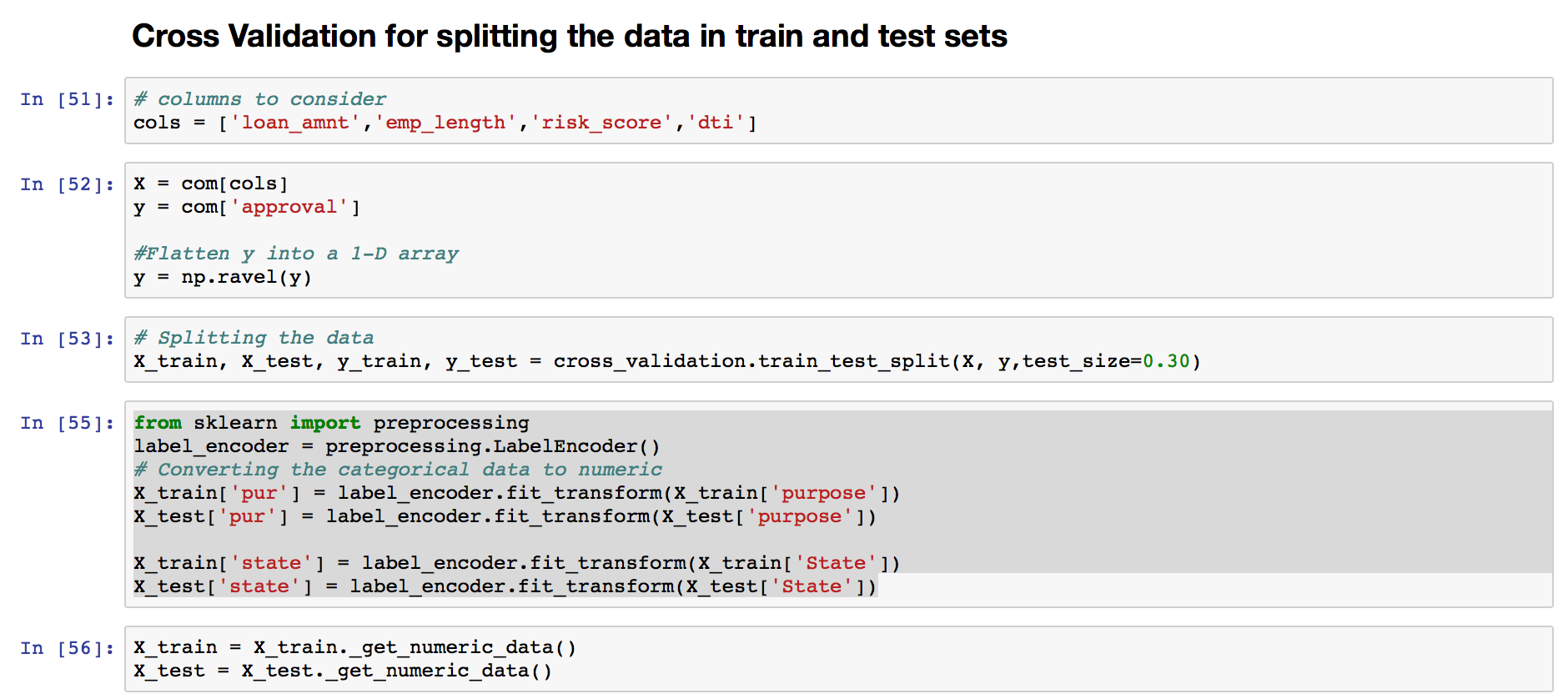
**2(a): Classification**

* For classification, we combined the loan and decline data keeping only the columns in both the datasets. We have generated a flag column named ‘approval’ during the cleaning process.

approval ->1 Loan data  
 approval ->0 Decline data

* We create a dataframe having equal number of rows(1640753) from loan and decline data.
* Features we are dealing with for classification:

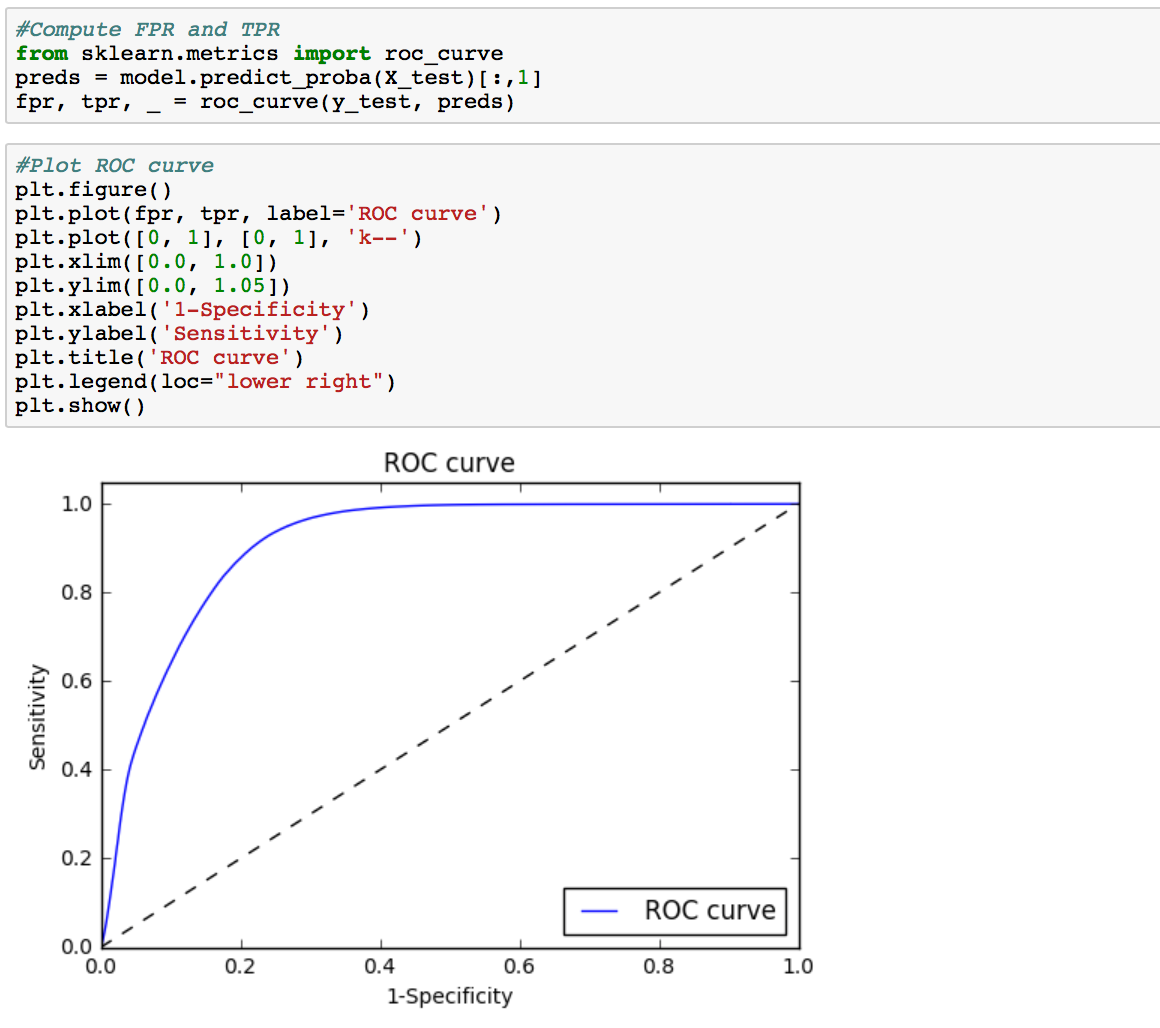
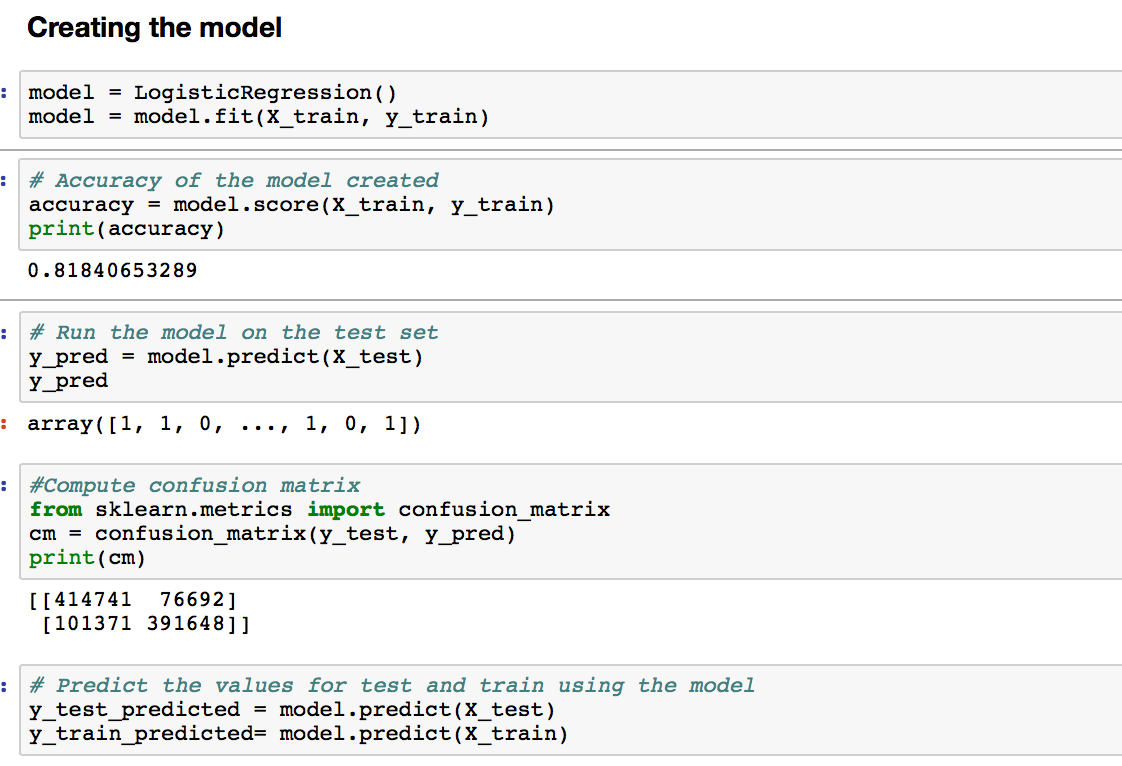
['id','loan\_amnt', 'purpose', 'fico', 'dti','zip\_code', 'State', 'emp\_length', 'Year', 'Month', ‘approval']]

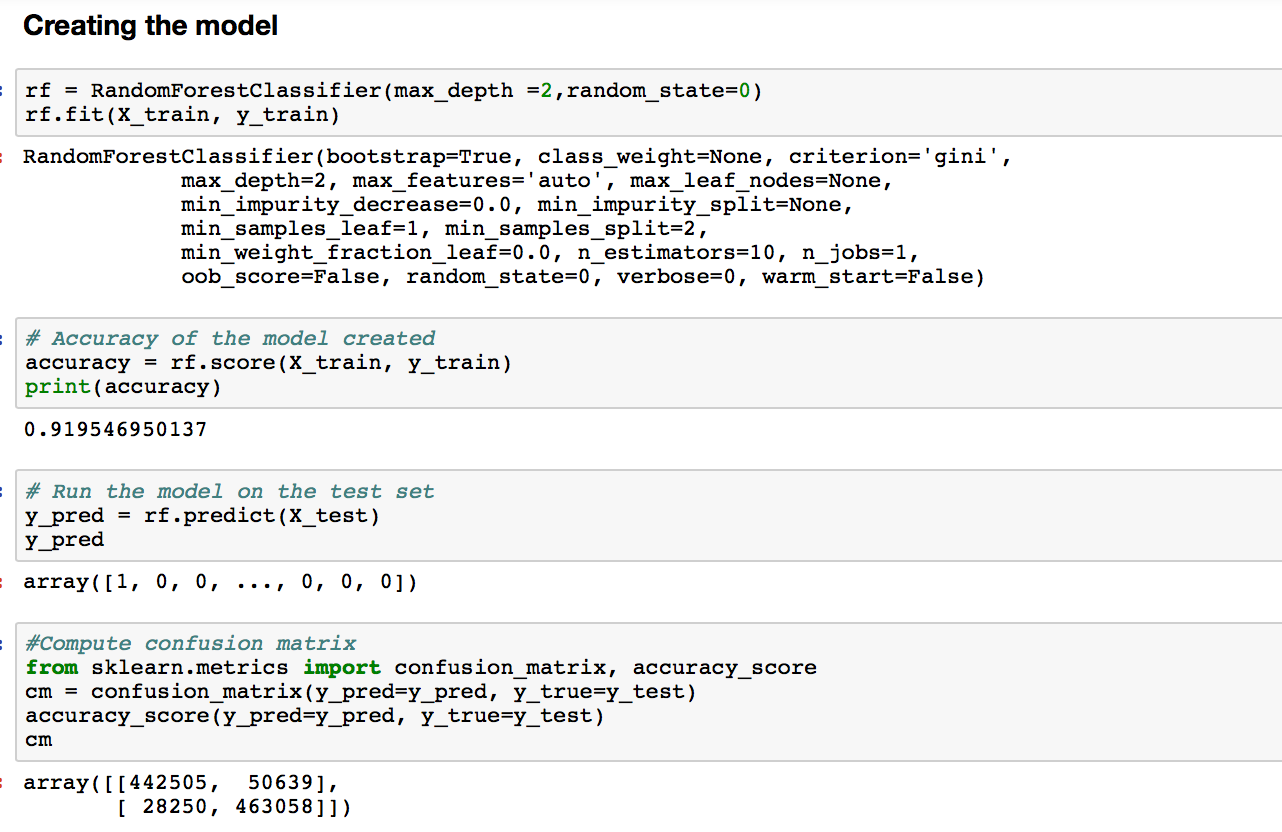
* Then we spilt the data into train and test having 70/30 split.

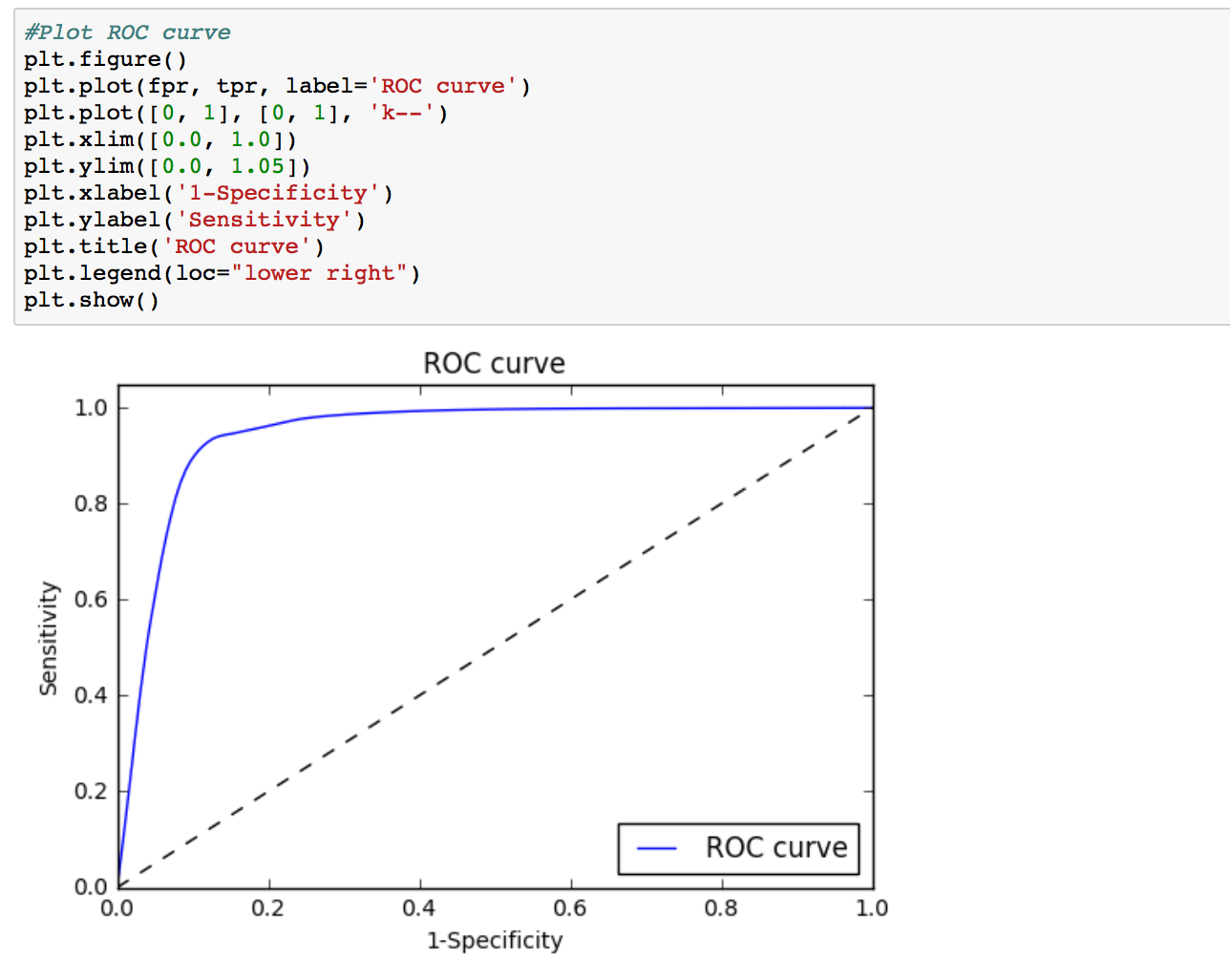
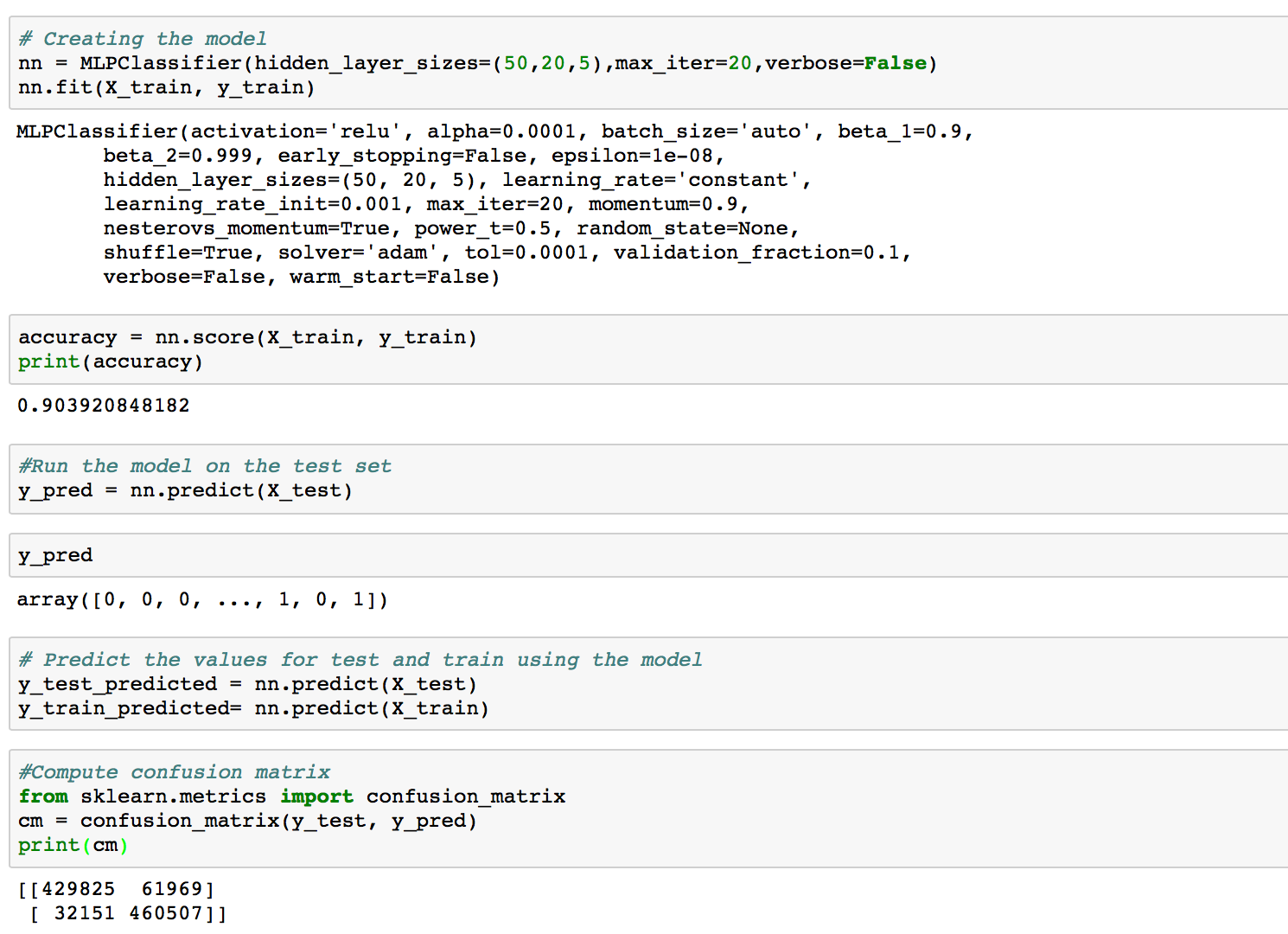
**Models Implemented**

We implement **Logistic regression, Random Forest, Neural Network** models algorithms.

**Random Forest** worked best for the classification.

**Logistic Regression:**

**Random Forest:**

**Neural Networks:**

**2(b): Clustering**

We have segmented our loan data broadly in ways:

* 1. No cluster(Zero Cluster)
  2. Manually
  3. Clustering Algorithm(K-Means)

**NO Cluster:**

We loaded the loan data as is.

**MANUALLY:**

We cluster data based on the fico score in to four clusters. And generated a new column ‘cluster’ for loan data frame. Based on the fido score in each row of loan data we fill in the cluster column with the corresponding cluster name. Then exporting each cluster to a separate CSV.