Project Report on

**UNDERSTANDING RISK ANALYSIS IN BANKING AND FINANCIAL SERVICES**



Submitted in partial fulfillment for the award of

**Post Graduate Diploma in High Performance Computing Application Programming from**

**CDAC-ACTS (Pune)**

**Under the Guidance of**

**Mr. Prakash Sinha**

**Submitted By:**

Aditya Nagose 220940141001

Anjali Sangode 220940141005

Pooja Jadhav 220940141017

Taksande Sandeep Ravindra 220940141028

**INDEX**

**Introduction ...............................................................................................................................................................03**

* About Lending Club .....................................................................................................................................03
* Purpose/ Objective of the Project .................................................................................................................03
* Dataset ..........................................................................................................................................................03

**Importing Essential Libraries ..................................................................................................................................04**

**Reading The Dataset .................................................................................................................................................05**

* Looking At The Columns .............................................................................................................................06

**Exploratory Data Analysis / Data Visualization .....................................................................................................08**

* Crosstabs .......................................................................................................................................................12
* Loan Status By Application Type .................................................................................................................14
* Loan Status By Home Ownership .................................................................................................................14
* **Mean Interest Rate by Grade and Sub-Grade .......................................................................................................15**
* **Grades For Each Employee Title ..........................................................................................................................16**
* Interest Rates According To Employee Lengths ..........................................................................................17
* DTI ...............................................................................................................................................................17

**Dealing With Missing Values ...................................................................................................................................18**

* Dropping Columns Containing Either 40% or More than 40% NaN Values ...............................................18

**Business Case 1 ..........................................................................................................................................................20**

* Encoding Categorical Variables ...................................................................................................................21
* Imputing Missing Values .............................................................................................................................22
* Correlation Matrix ........................................................................................................................................23
* Splitting Data into Test and Train ................................................................................................................24
* Standardizing Numerical Data .....................................................................................................................24
* Training The Data ........................................................................................................................................25

**Model Diagnostics ....................................................................................................................................................26**

* Logistic Regression .....................................................................................................................................26
  + Tuning Hyperparameters ...............................................................................................................27
  + Confusion Matrix ..........................................................................................................................28
  + ROC Curve ....................................................................................................................................28
* LightGBM Classification ............................................................................................................................29
  + Tuning Hyperparameters ..............................................................................................................29
  + Confusion Matrix .........................................................................................................................30
  + ROC Curve ...................................................................................................................................31

**Model Comparison ..................................................................................................................................................31**

**Business Case 2 .......................................................................................................................................................32**

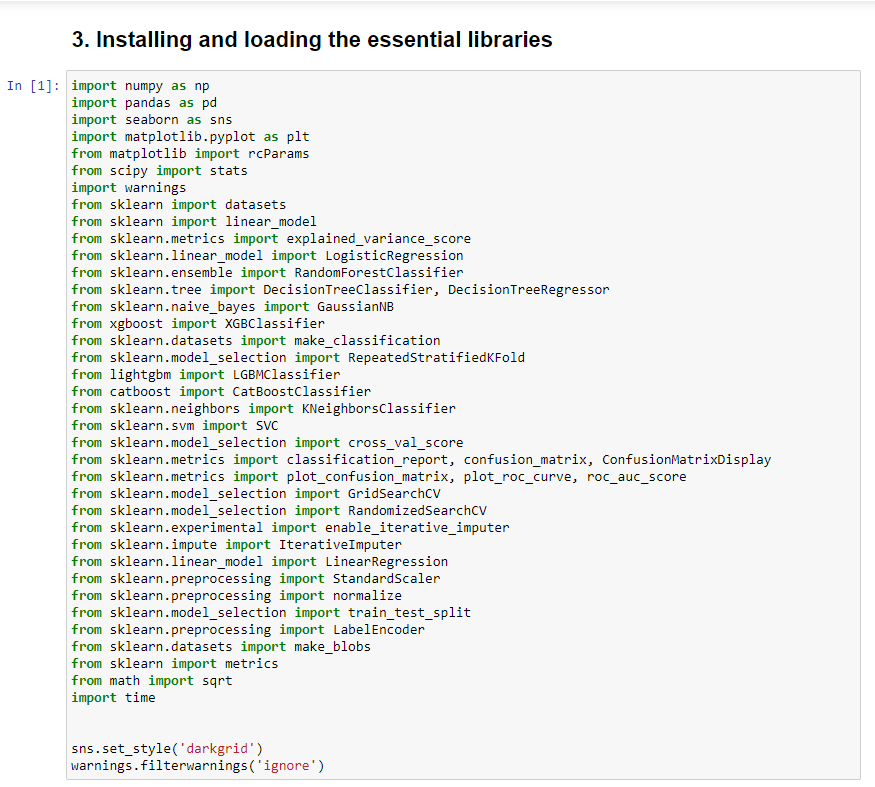
* Encoding Categorical Variables .................................................................................................................35
* Imputing Missing Values ...........................................................................................................................35
* Correlation Matrix .....................................................................................................................................36
* Splitting Data into Test and Train .............................................................................................................37
* Standardizing Numerical Data ..................................................................................................................37
* Linear Regression Model ..........................................................................................................................38
* Evaluating the Model ................................................................................................................................39
* Plotting Actual Vs Predicted ....................................................................................................................40
* Plotting Histogram of Residuals ...............................................................................................................40

**Limitations and Future Scope...............................................................................................................................41**

**INTRODUCTION**

* **About Lending Club**
  + LendingClub is a US peer-to-peer lending company, headquartered in San Francisco, California founded by Renaud Laplanche and Soul Htite in 2006. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. LendingClub is the world's largest peer-to-peer lending platform.
  + Borrowers were able to arrange unsecured personal loans between $1,000 and $40,000 via LendingClub. The typical loan term was three years.
  + According to the information provided about the borrower, loan amount, loan grade, and loan purpose, investors were able to search and browse the loan listings on the LendingClub website and choose loans that they wished to invest in.
  + The interest on these loans generated profits for the investors. By charging investors a service fee and borrowers an origination fee, LendingClub generated income.
  + As the largest technology IPO in the United States of 2014, the firm raised $1 billion in its initial public offering. Despite being one of the biggest fintech companies and regarded as a pioneer in the field, LendingClub ran into issues in early 2016. These issues included difficulty attracting investors, a scandal involving some of the company's loans, and board concerns over CEO Renaud Laplanche's disclosures, which caused a significant decline in the company's share price and Laplanche's resignation.
  + The peer-to-peer lending platform of LendingClub, which had recently purchased Radius Bank, would be shutting down in 2020. There are no new loans accessible for individual investment; however, existing account holders will still be able to receive interest on their existing notes until they are all repaid or default occurs. A secondary market, which was originally used to sell existing loans, is no longer available.
* **Purpose/ Objective of the Project**
  + The goal of risk analysis is to recognise, quantify, and reduce different risk exposures or hazards that may be encountered by a project, investment, or organization.
  + Solving this case study will provide an idea about how real business problems are solved using EDA and Machine Learning. In this case study, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
* **Dataset**
  + The dataset contains all entries of Lending Club from 2007 to 2018 having 2260701 entries and 151 features.

**IMPORTING ESSENTIAL LIBRARIES**



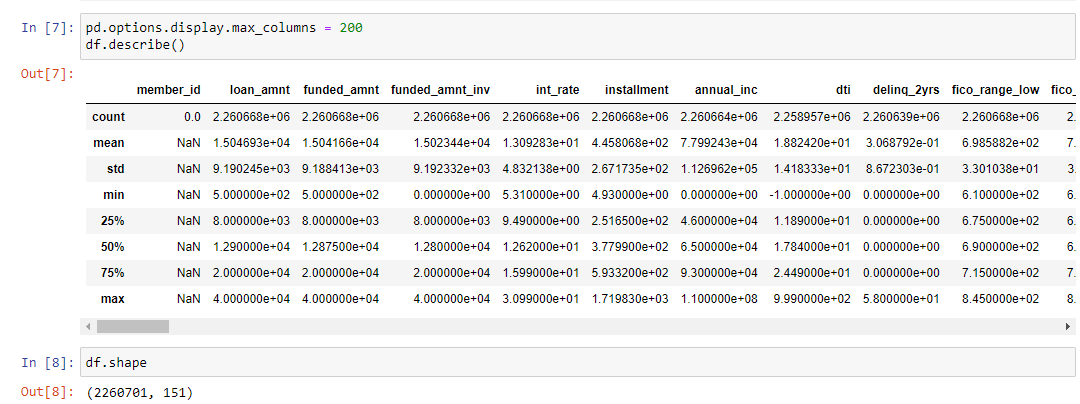
**READING THE DATASET**



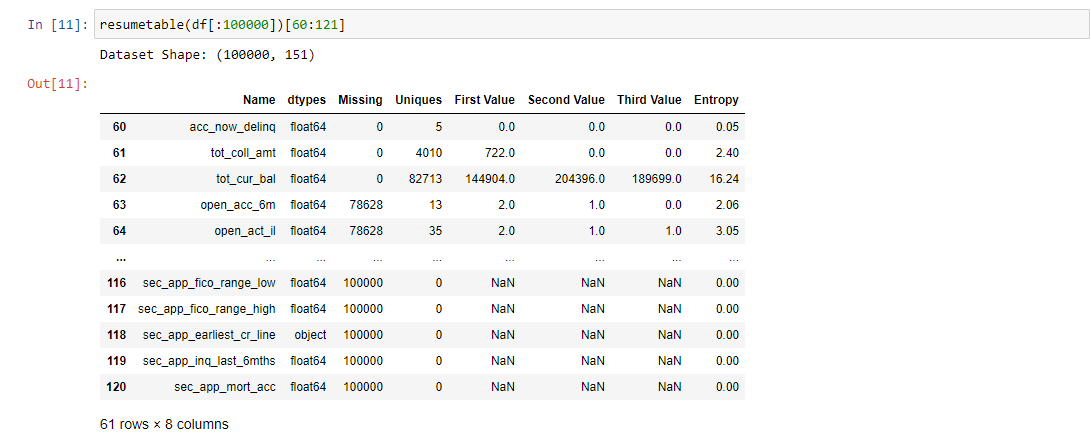
With Chunk size of 10000, the process takes 0.8149 Minutes (50.5176 seconds) to read the dataset. It is approximately 18 seconds LESS than executing without chunks.

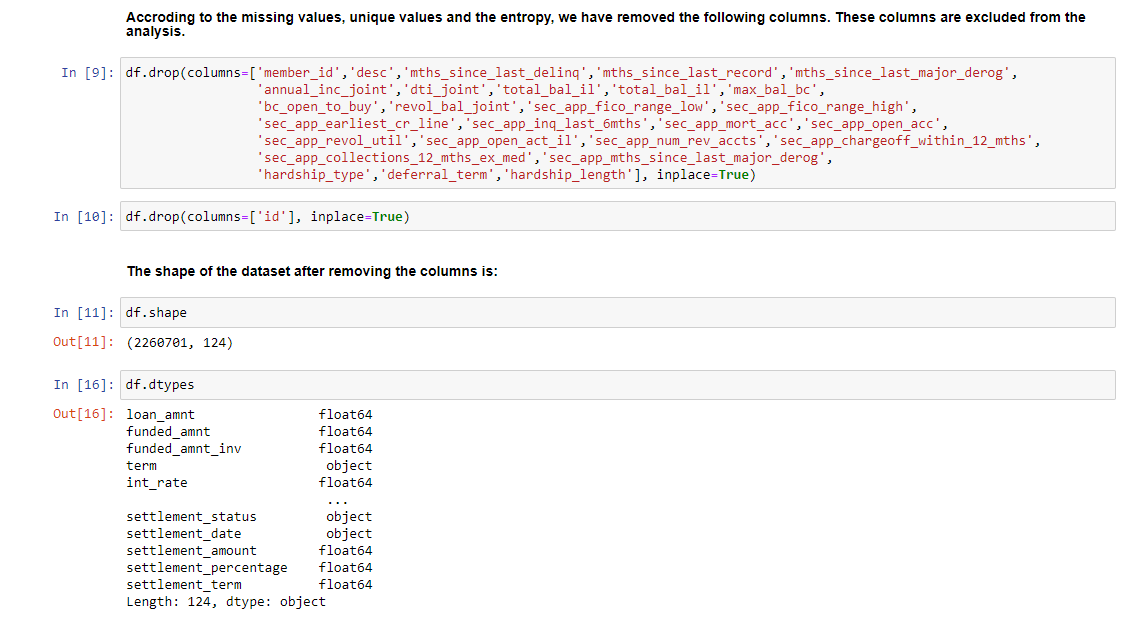


Without any chunk the process takes 1.13133 Minutes (67.8798 seconds) to read the dataset.





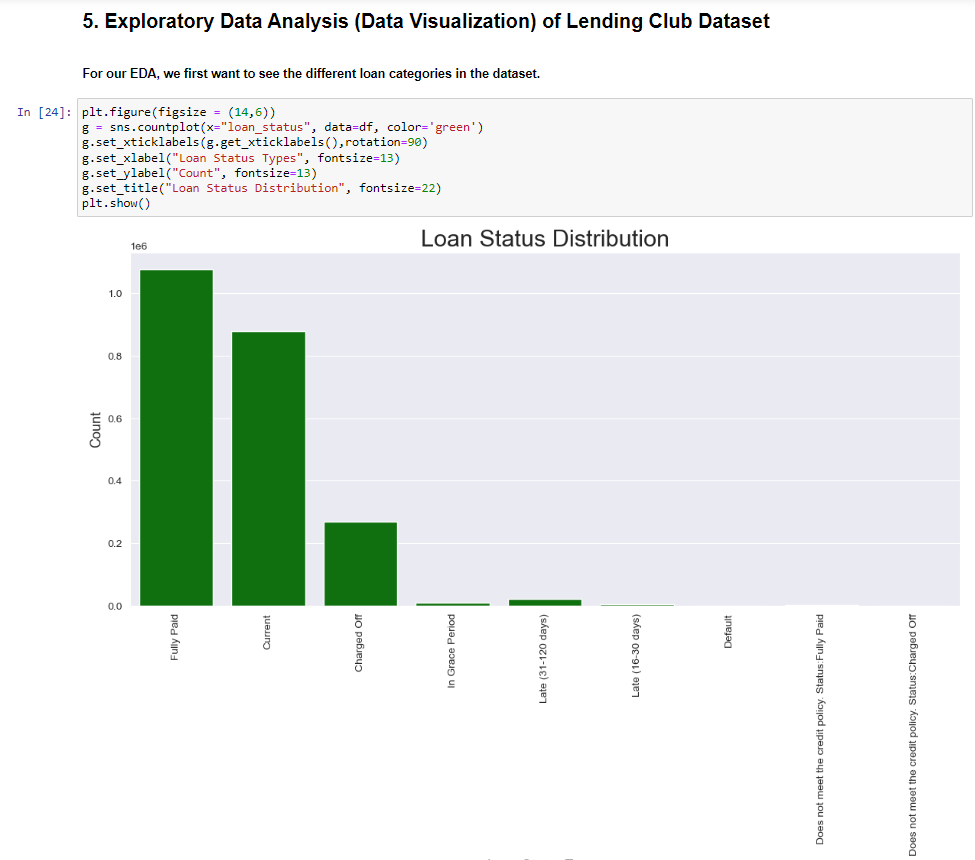


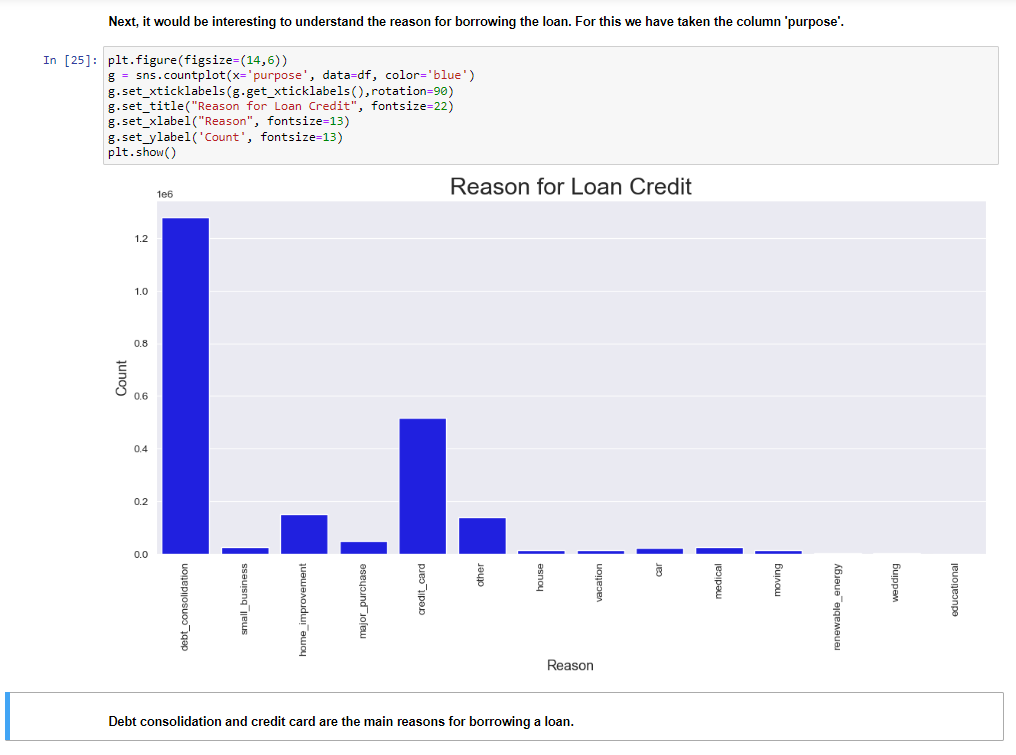


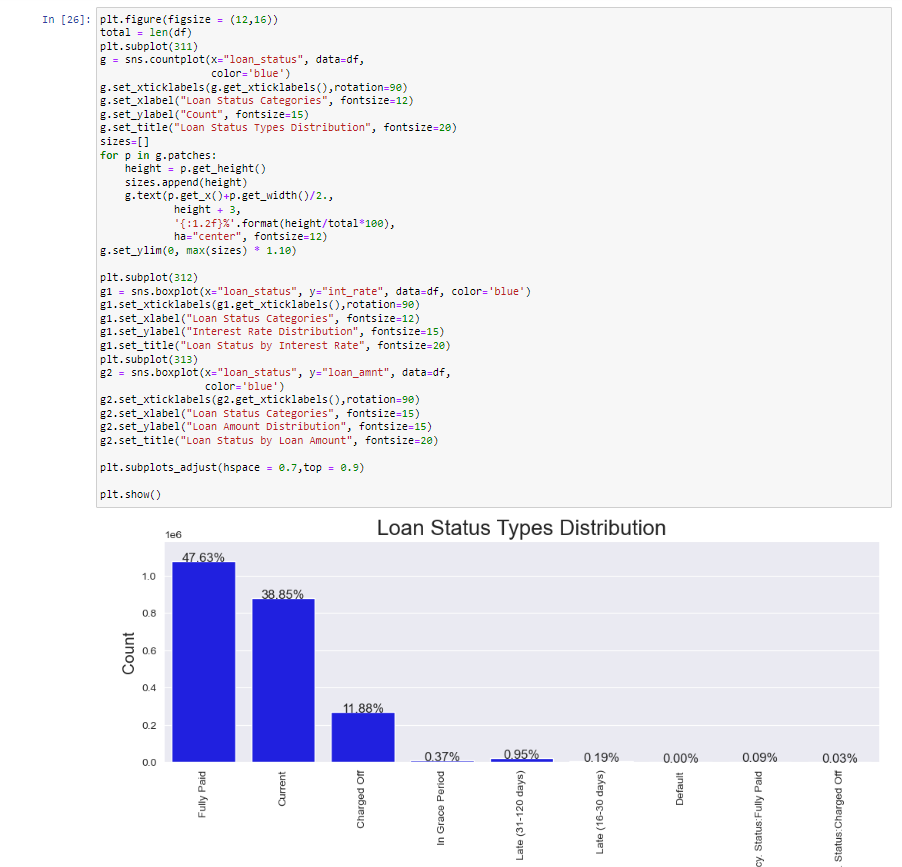
**The Program is divided into 3 stages:**

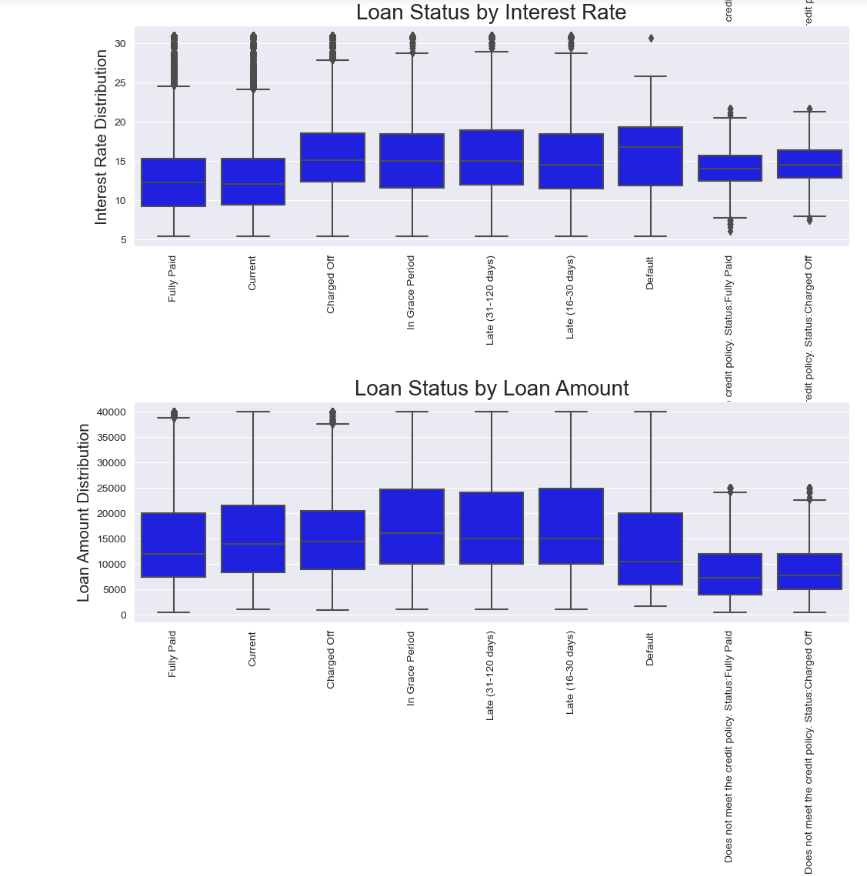
1. **Exploratory Data Analysis / Data Visualization**
2. **Model Designing by splitting dataset in to Train and Test**
3. **Model Diagnostics**

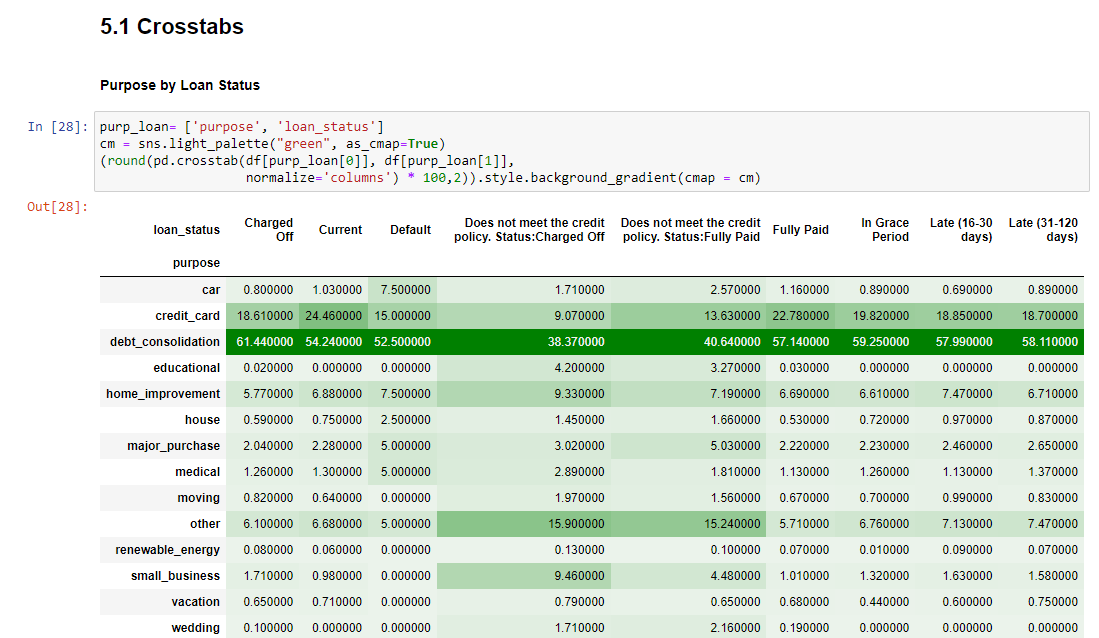
**EXPLORATORY DATA ANALYSIS/ DATA VISUALIZATION**

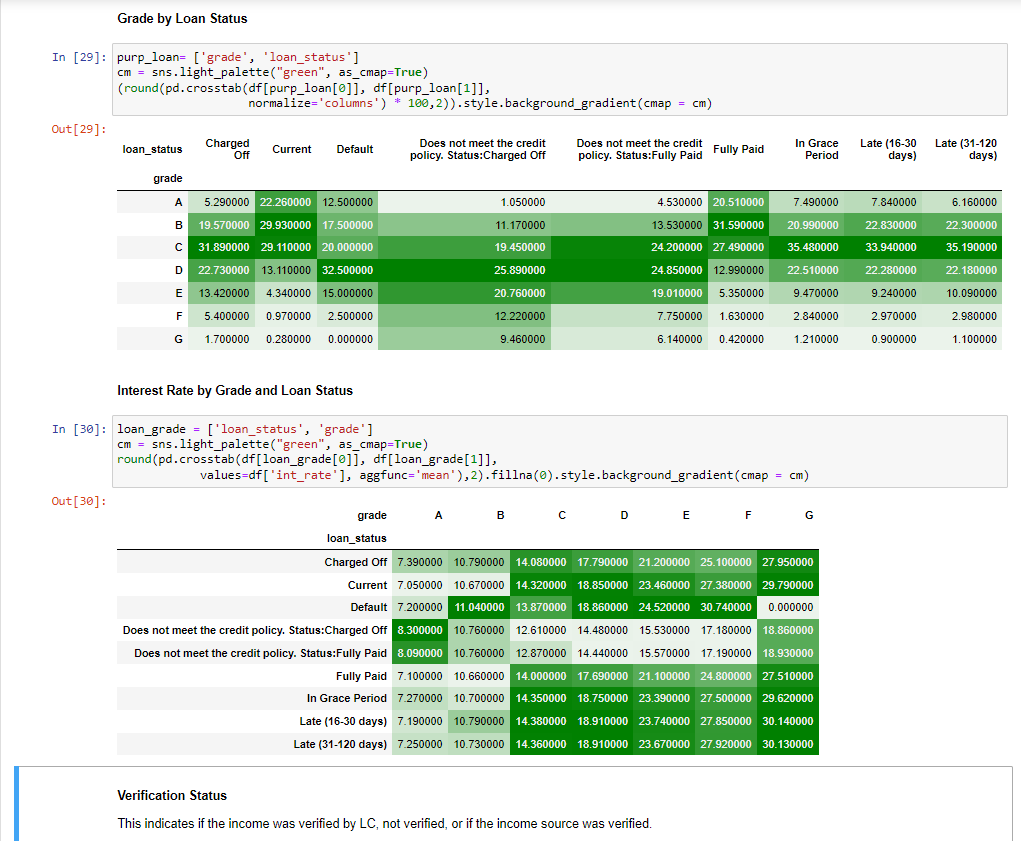
****

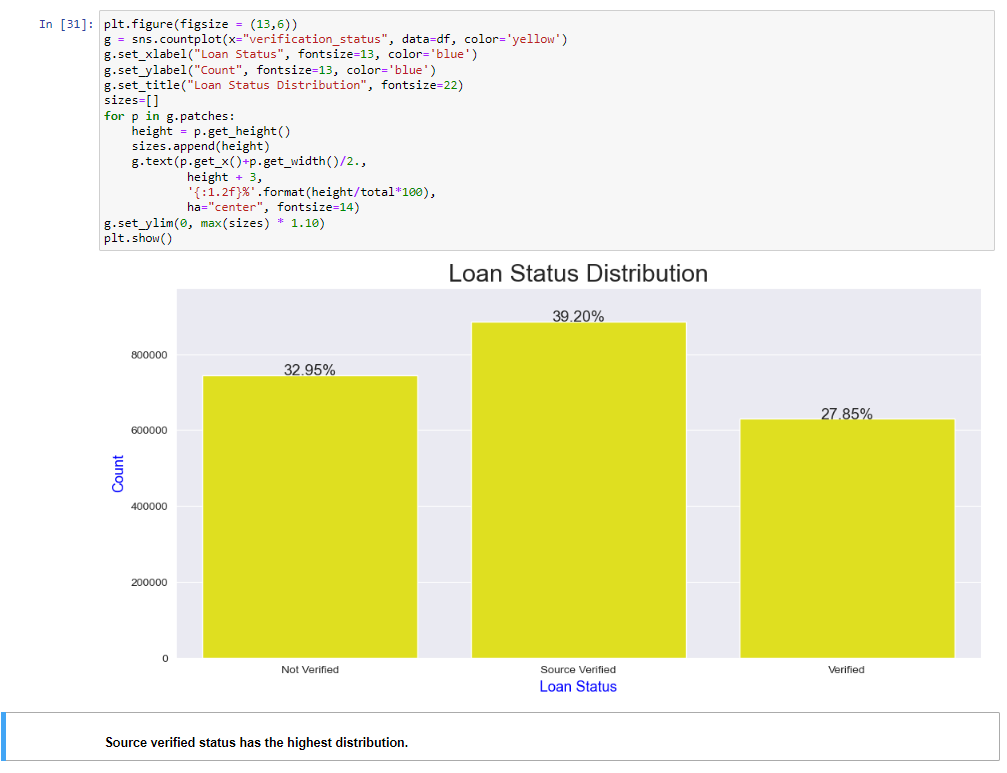
****

****

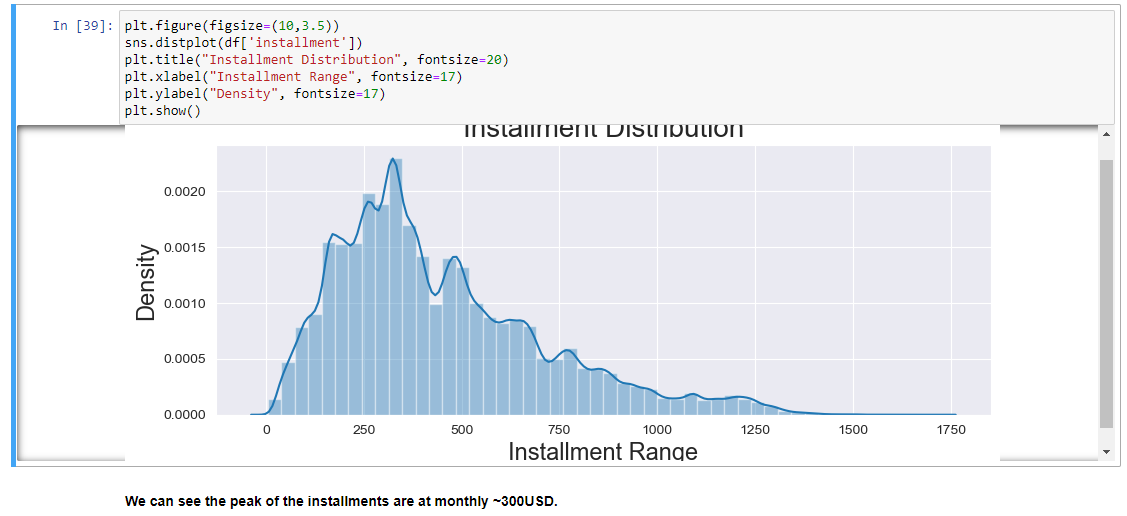
****

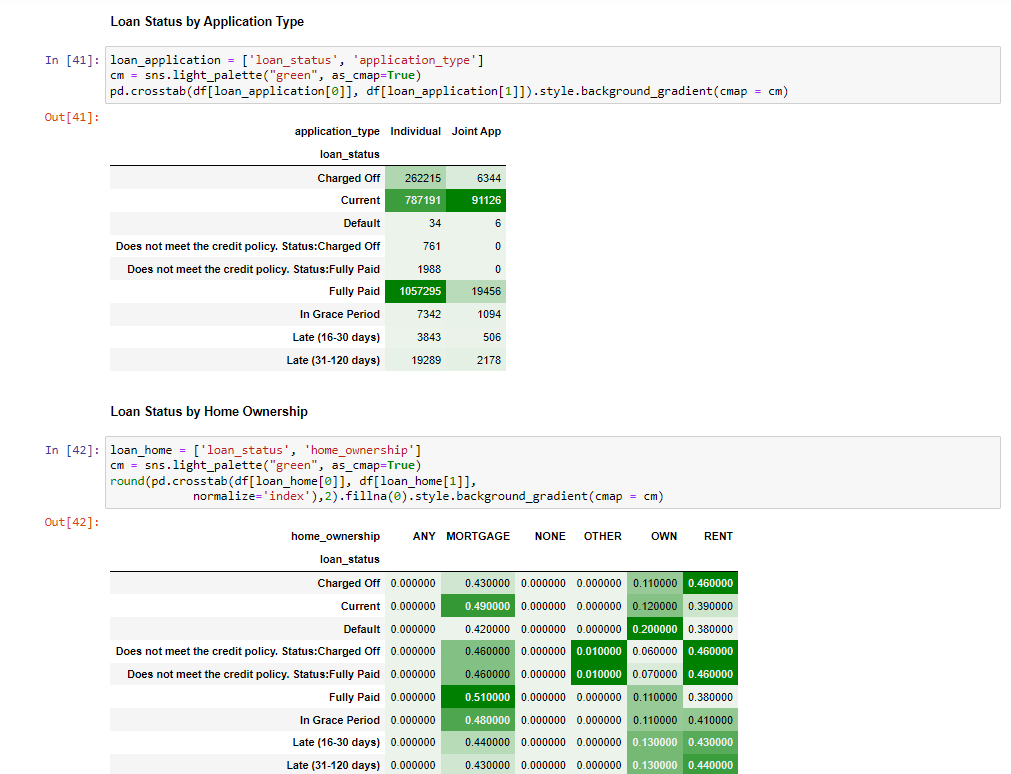
****

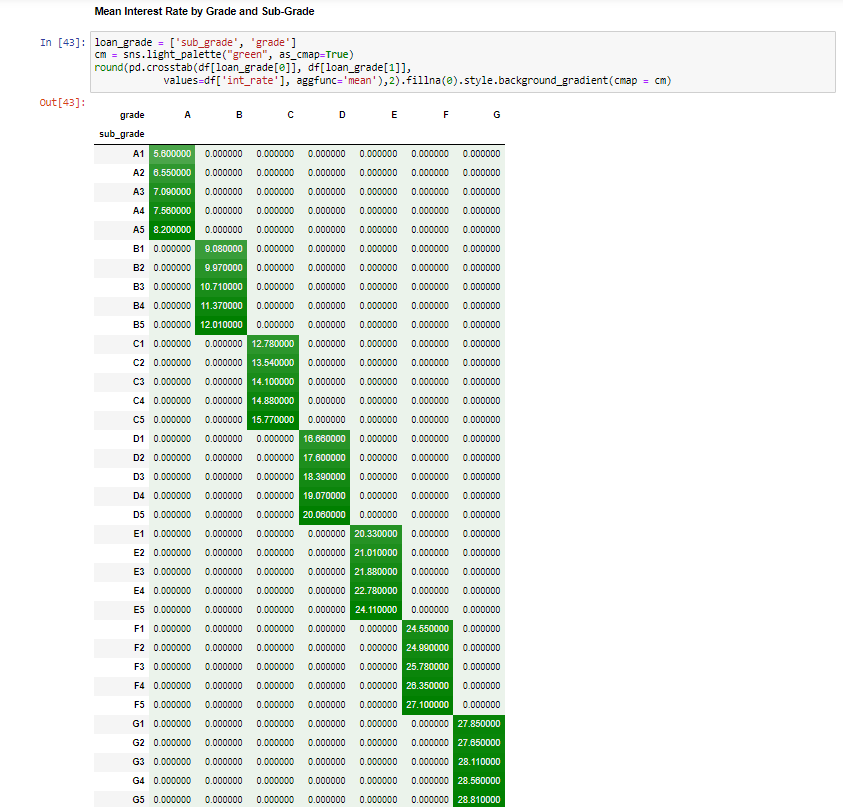
****

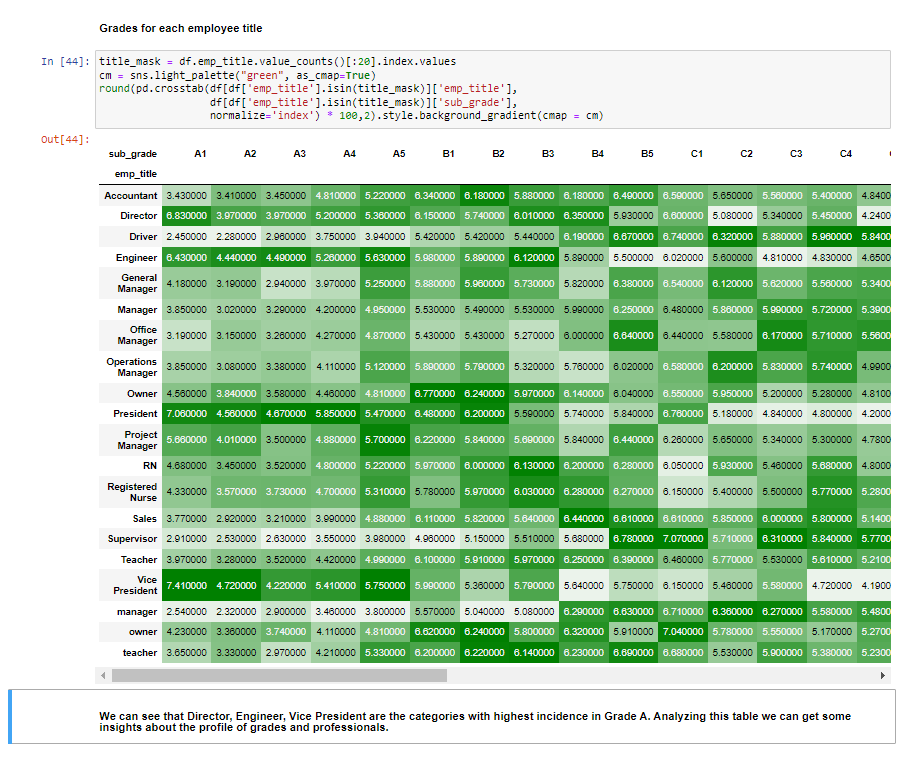
****

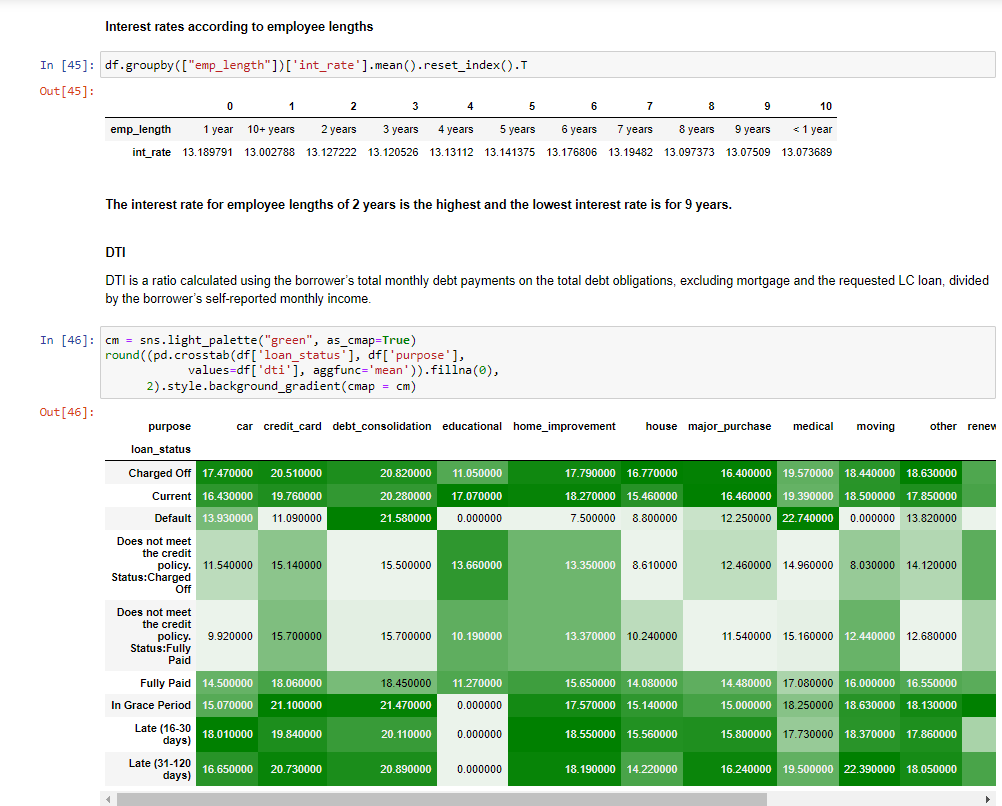
****

****

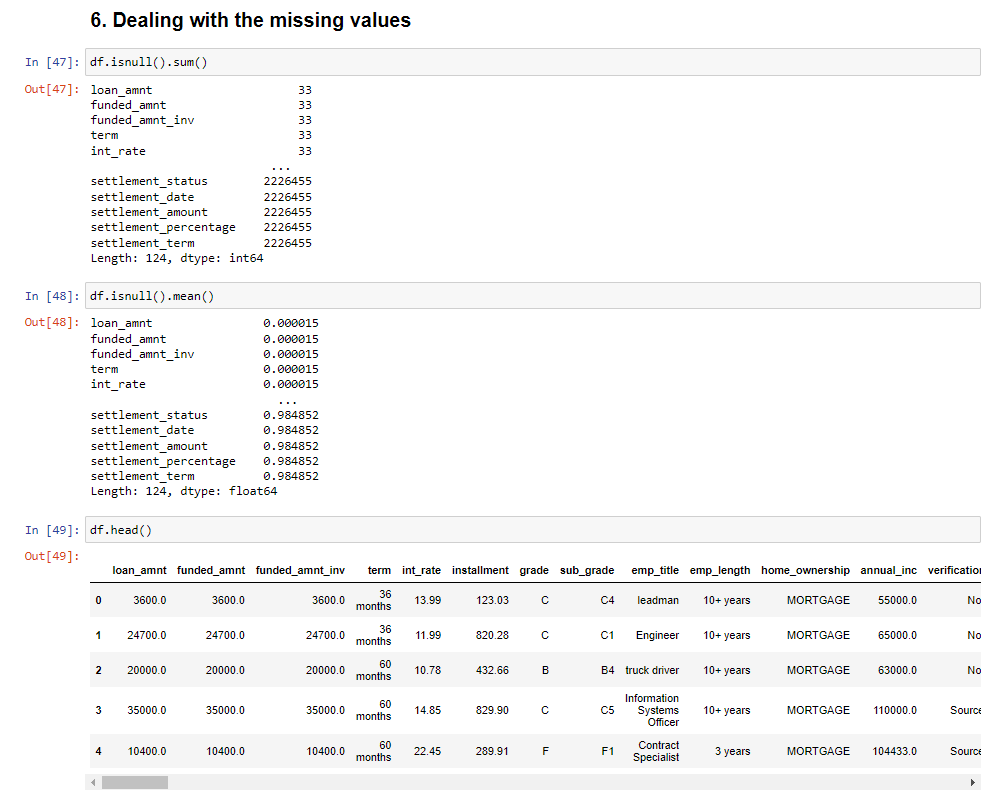
****

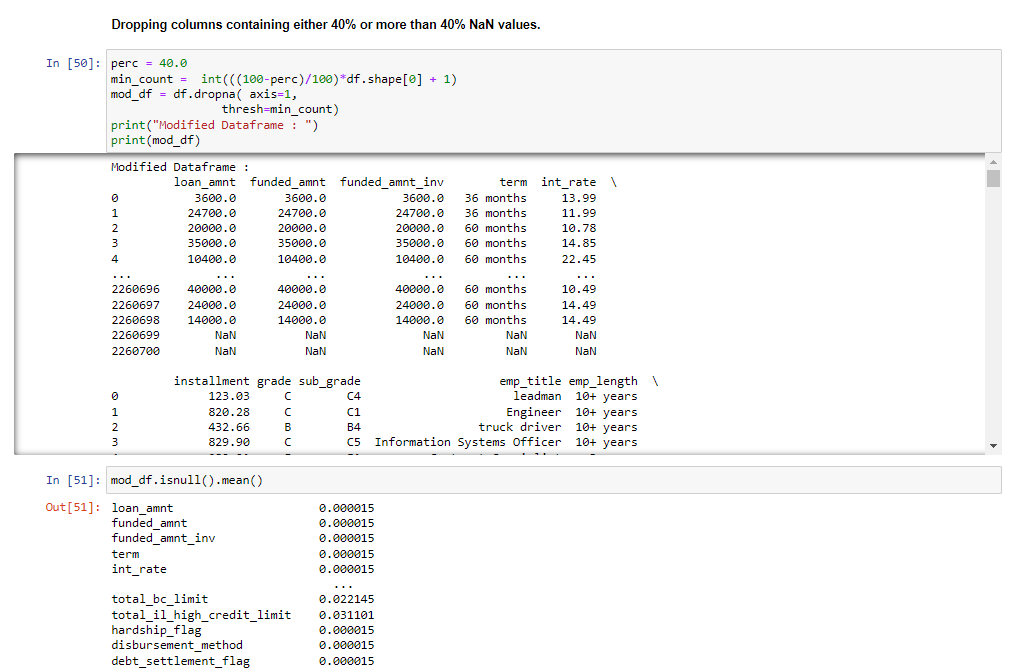
****

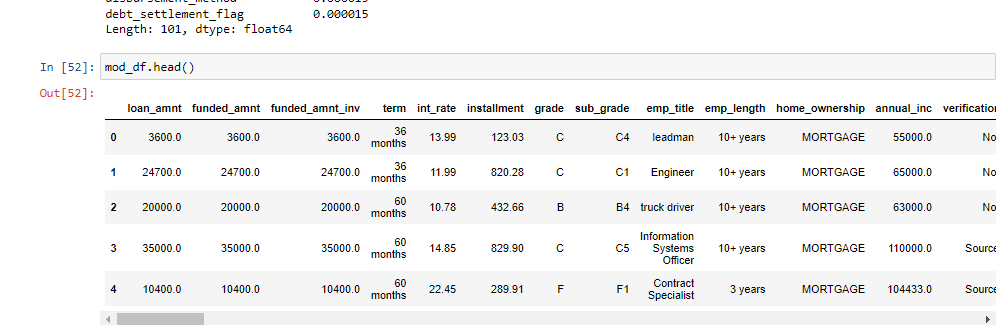
****

****

**DEALING WITH MISSING VALUES**







**BUSINESS CASE 1**

**Problem Statement:** Build a model to predict the probability of default for the applicants on the lending club app. Follow the ML pipeline of data cleaning, feature engineering, model training, K fold validation, Model diagnostics, and model selection.Try out the whole suite of classifier models that are available in scikit-learn.

**To predict if a borrower will clear his/ her loan, we chose the following variables:**

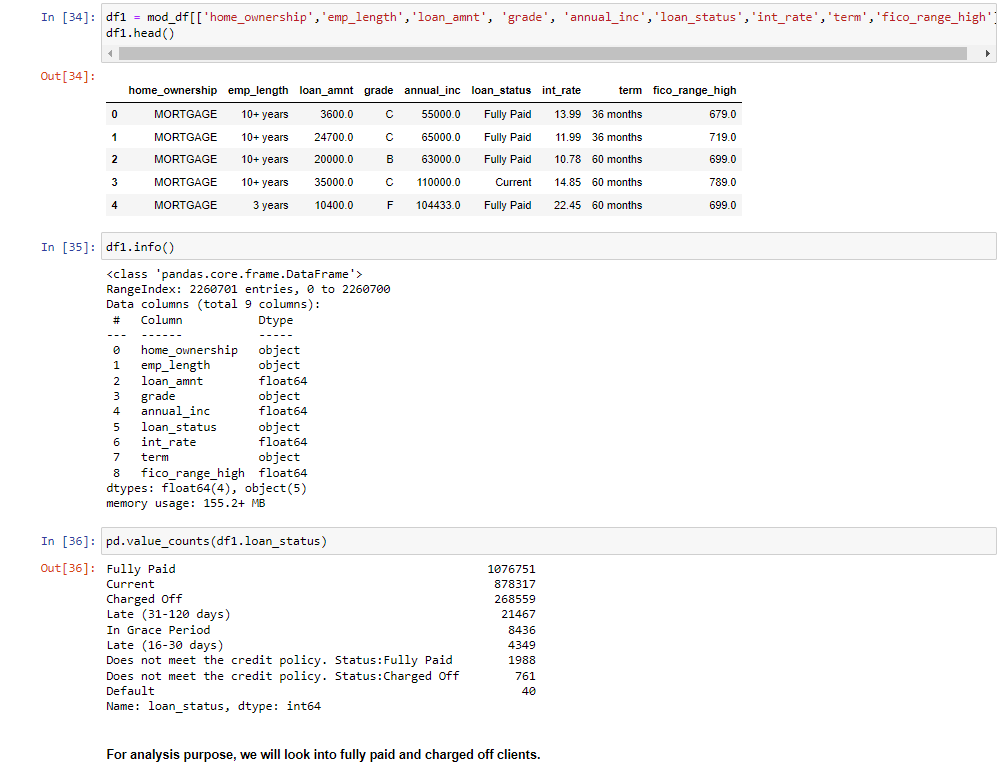
**Independent Variables:**

* home\_ownership : status provided by the borrower during registration or obtained from the credit report
* emp\_length:Employment length in Years
* loan\_amnt: the listed amount of loan applied for by the borrower,
* grade
* annual\_inc: the self reported annual income provided by the borrower during registration
* int\_rate:Interest Rate on the loan
* term: Number of payments on the loan (in months)
* fico\_range\_high: The upper boundry range the borowwer's FICO at loan originates belongs to

**Dependent Variable:**

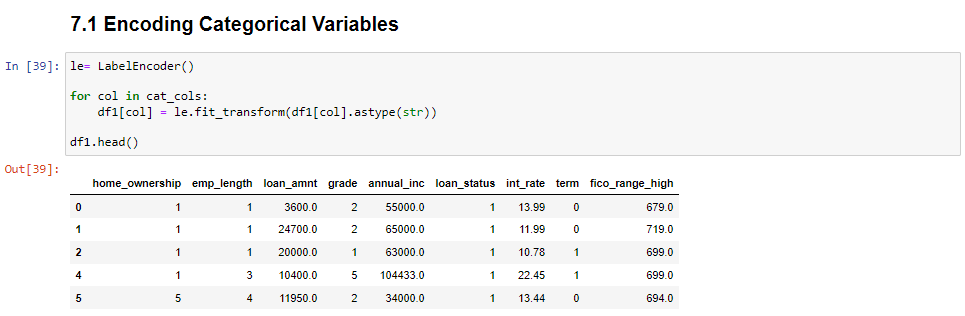
* loan\_status: Current state of the loan

**Summary:** In 2012, LendingClub reached a $1 billion loan milestone; the company facilitated these loans through three different vehicles: notes, certificates, and whole loan sales. Through its online standard program loan division, it offered these three or five-year unsecured personal loans to borrowers with a FICO range of at least 660 who met the other credit criteria.Applicants specified the amounts they wanted to borrow and provided information about their financial situation, employment, and intended use of the loans. Using the online platform, investors could browse notes and decide which ones to fund. The most common reasons for borrowing were debt consolidation, credit card debt refinance, and home renovation.LendingClub assigned grades ranging from A to G, which are further subdivided into sub-grades 1 to 5. For each entry, the data includes over 100 descriptive features such as application ID, credit score, loan amount, purpose, and so on.To predict new data with a reassuring level of confidence, the target variable is "loan status." Several models have been trained to classify a new loan application as default or repayment based on this target variable.

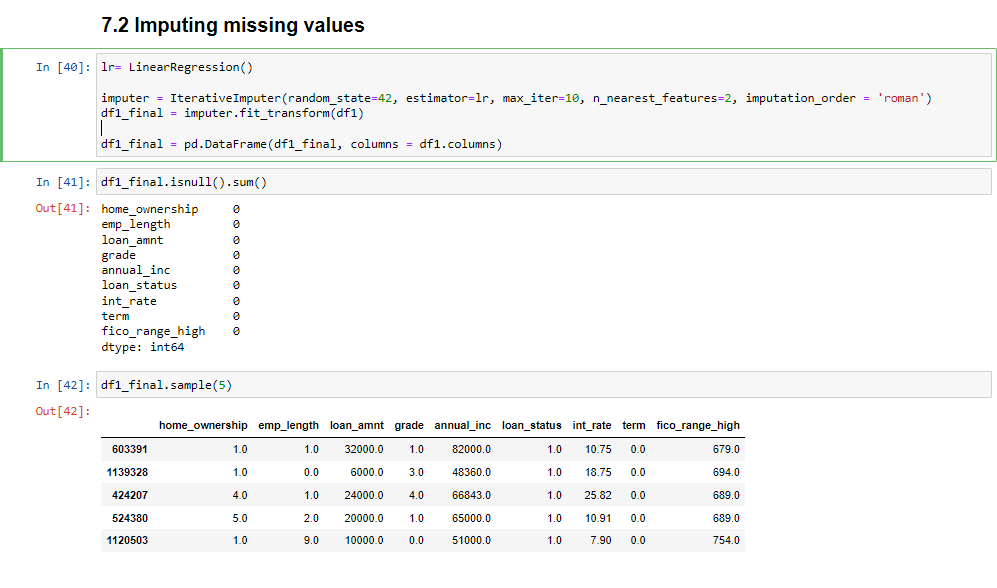




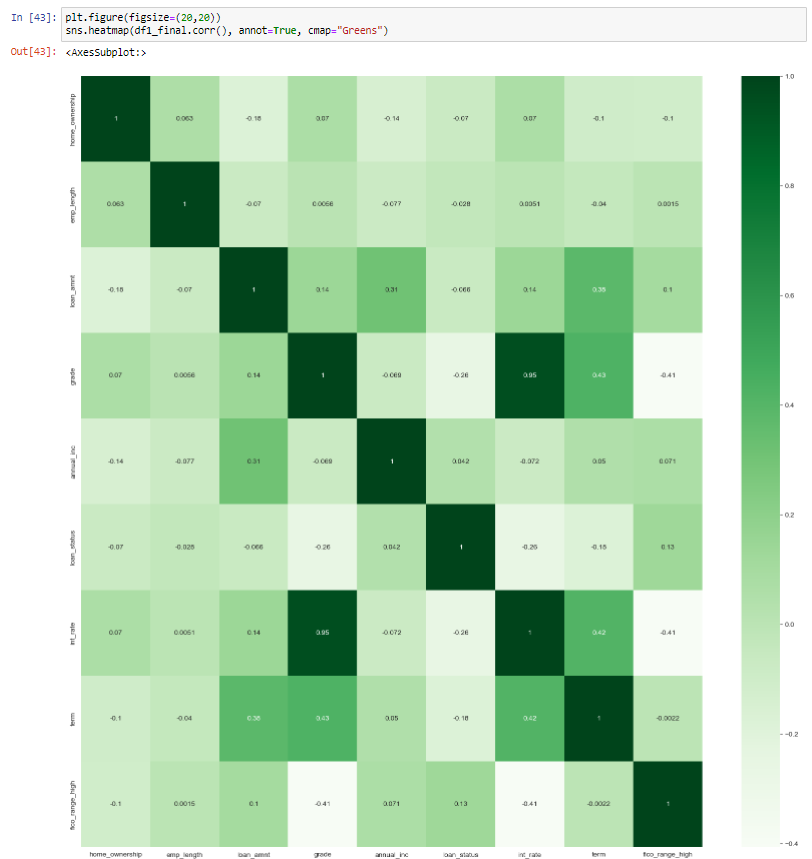
* + **Encoding categorical Values**
    - Working with various datasets, we discovered that some of the features are categorical; if we give the feature directly to our model, our model is unable to comprehend those feature variables. The incomprehensibility of categorical data to machines is well known. All independent and dependent variables, or input and output features, must be numerical for machines to function. To fit our data to the model, we must first encode any categorical variables in our data that are present.
    - Models can only be used with numbers. In order for the machine to learn from the data and provide the appropriate model, it is important to convert the categorical values of the feature values into numerical ones. This process of transforming category data into numerical data is called Encoding.



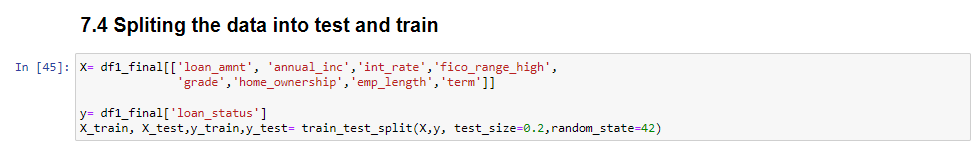
* + **Imputing Missing Values**
    - Imputation maintains every instance by substituting missing data with an estimated value based on other available information. Following the imputed values for all missing items, the data set can be analyzed using methods that are typical for complete data.
    - Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located.



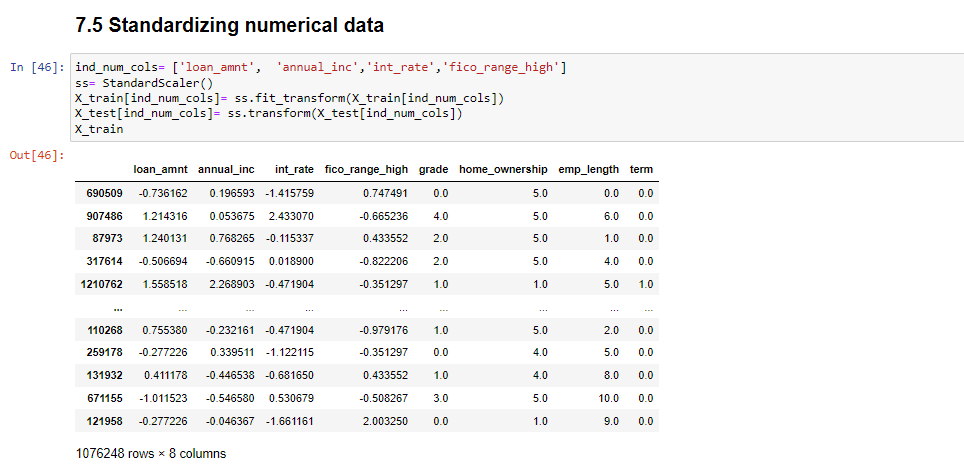
* + **Correlation Matrix**



* + **Splitting into Train & Test**
    - When machine learning algorithms are used to make predictions on data that was not used to train the model, their performance is estimated using the train-test split technique.
    - It is a quick and simple process to carry out, and the outcomes let you compare the effectiveness of machine learning algorithms for your particular predictive modeling issue. While being straightforward to use and understand, there are some circumstances in which the method shouldn't be applied, such as when the dataset is tiny and further configuration is needed, as when it is used for classification and the dataset is unbalanced.
    - The test size we have used is 0.2 which indicates 80% of the data will be trained and 20% will be used on testing models to get the accuracy.



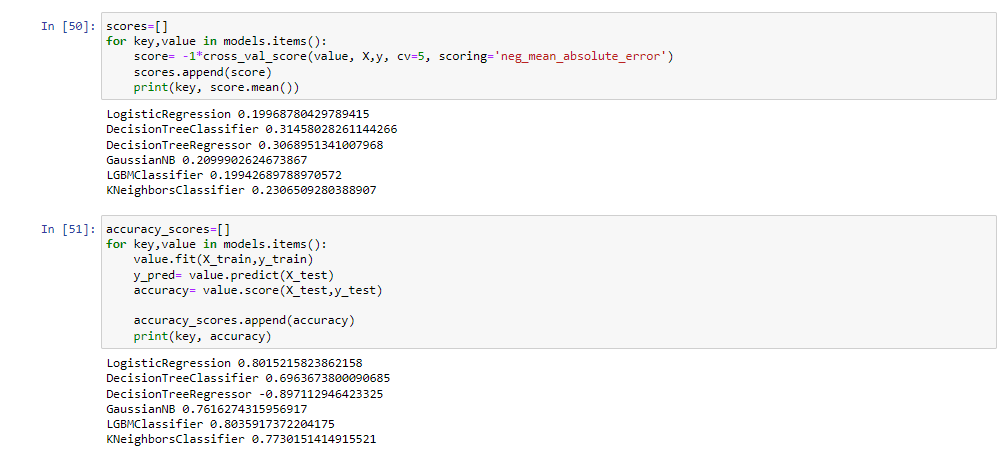
* + **Standardizing Numerical Data**
    - The process of transforming data into a standardized format so that users may process and evaluate it is known as data standardization.
    - For several reasons, data standardization is important. At the beginning, it aids in the establishment of distinct, clearly defined elements and attributes and offers a thorough catalog of your data. Understanding your data well is an essential first step, regardless of the insights or issues you're seeking to solve.



* + **Training Data (All models)**
    - We took following models to train the data:
      * LogisticRegression
      * DecisionTreeClassifier
      * DecisionTreeRegressor
      * GaussianNB
      * LGBMClassifier
      * KNeighborsClassifier
    - Also, we planned to use RandomForestClassifier, SVC, CatBoostClassifier & XGBClassifier for training but these models were taking a lot of time to process, some models were getting crashed too. Hence, we omitted these models.

****

* + **Generating Scores and Accuracy**

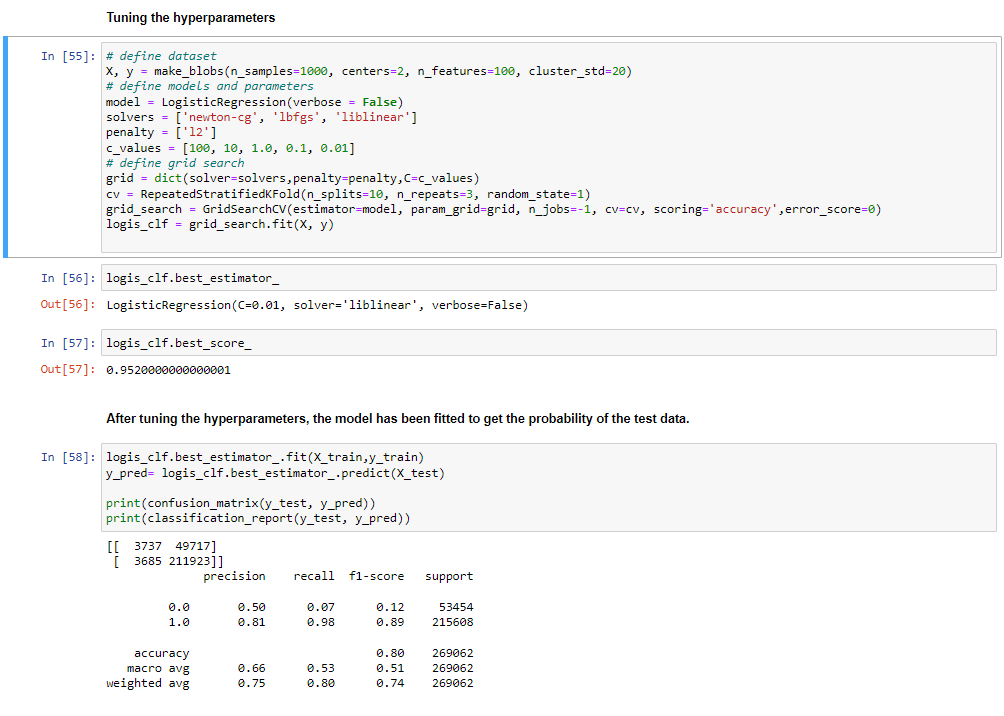
****

* **As it can be observed, LogisticRegression and LGBMClassifier gives accuracy of 0.8015215 and 0.8035915 respectively. Hence we used these two models for further processes.**

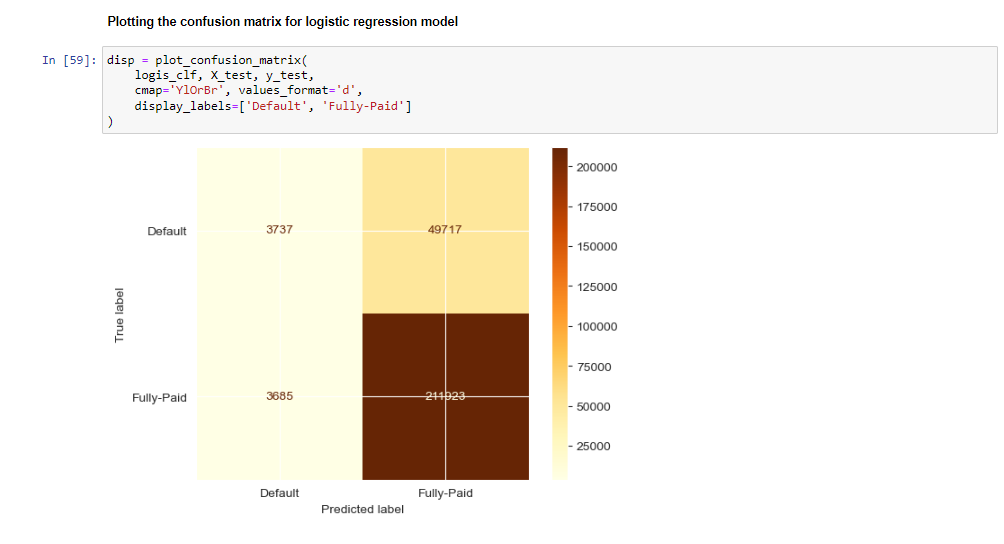
**MODEL DIAGNOSTICS**

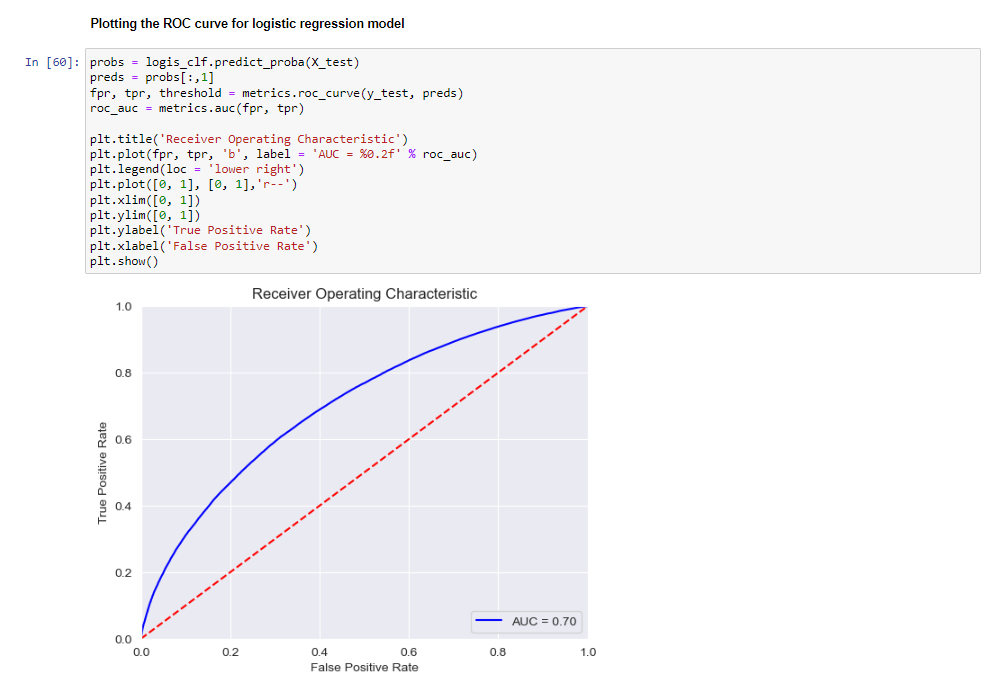
****

* + **Tuning Hyperparameter**
    - The control of a machine learning model's behavior requires hyperparameter adjustment. Our predicted model parameters will yield less-than-ideal outcomes if our hyperparameters aren't properly tuned to minimize the loss function. This indicates that our model has more flaws. The accuracy or confusion matrix will really be worse in practice.
    - As hyperparameters are unique to the algorithm, we are unable to determine their values from the data. To determine the model parameters, we employ hyperparameters. For a given data collection, various hyperparameter values result in various model parameter values.

****

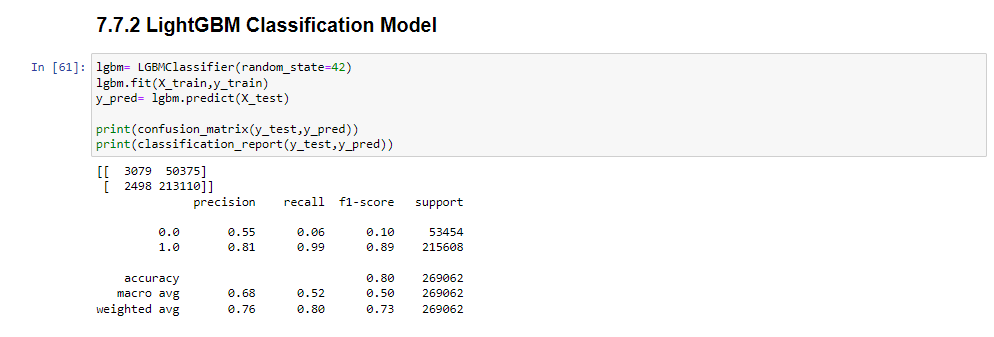
* + **Plotting Confusion Matrix**

****

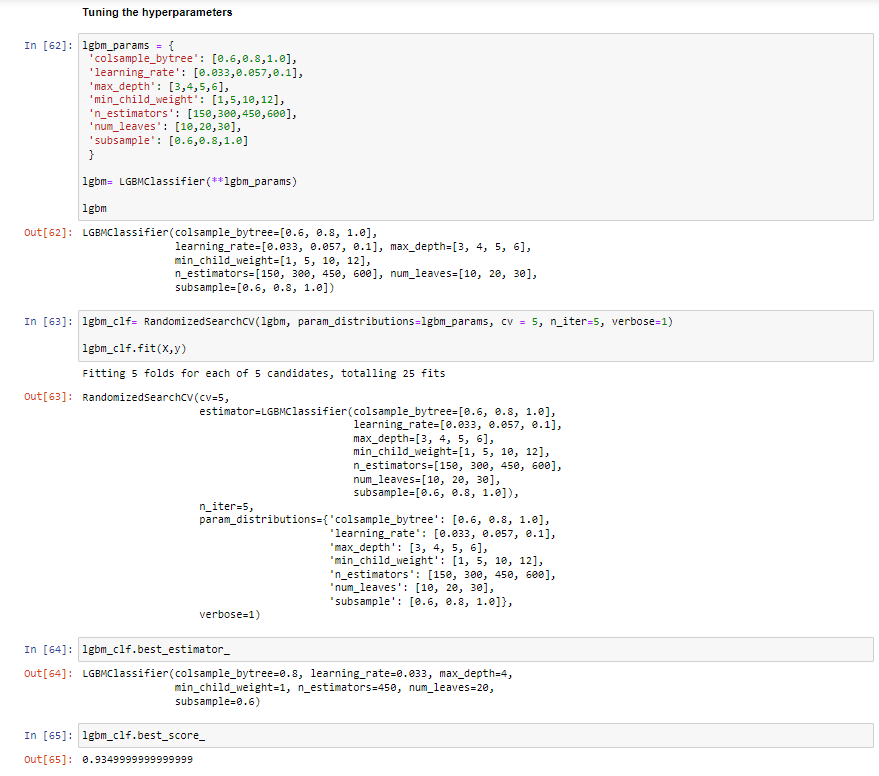
* + **Plotting ROC Curve**
    - An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
* True Positive Rate (TPR = TP / [TP + FN])
* False Positive Rate (FPR = FP / [FP + TN])

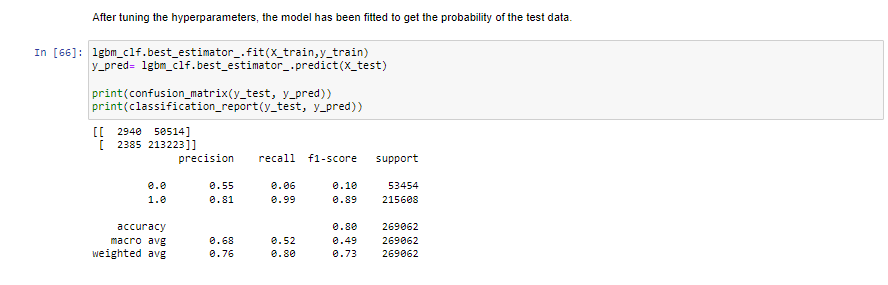
**NOTE : Similar Process is followed for LGBMClassifier**

* + **Model Diagnostics for LGBM Classifier**

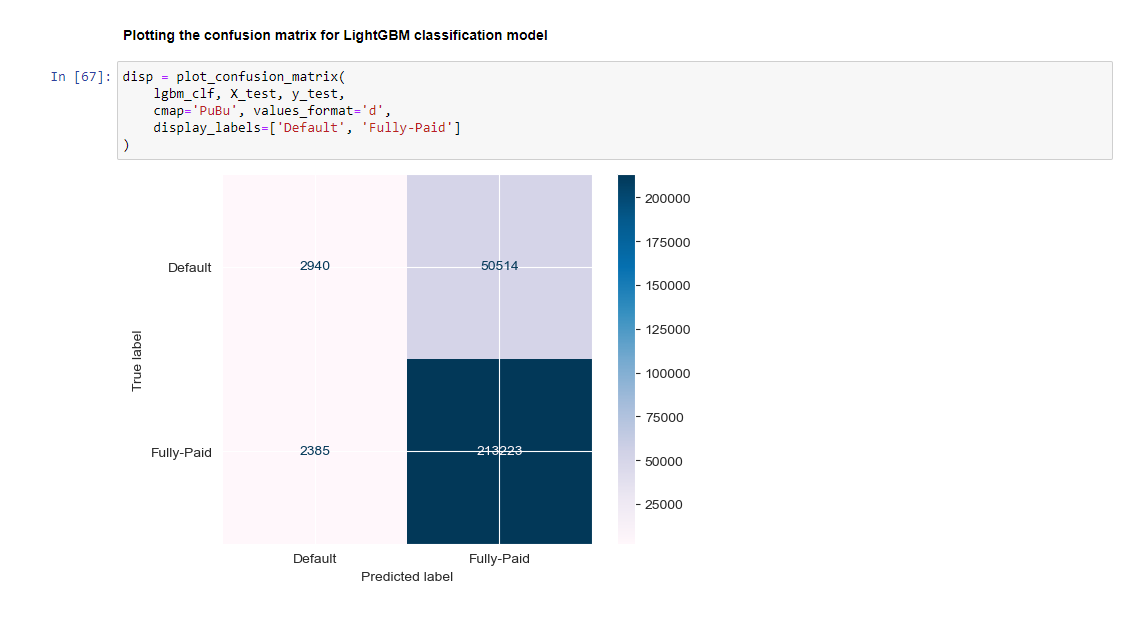
****

* + **Tuning Hyperparameter**

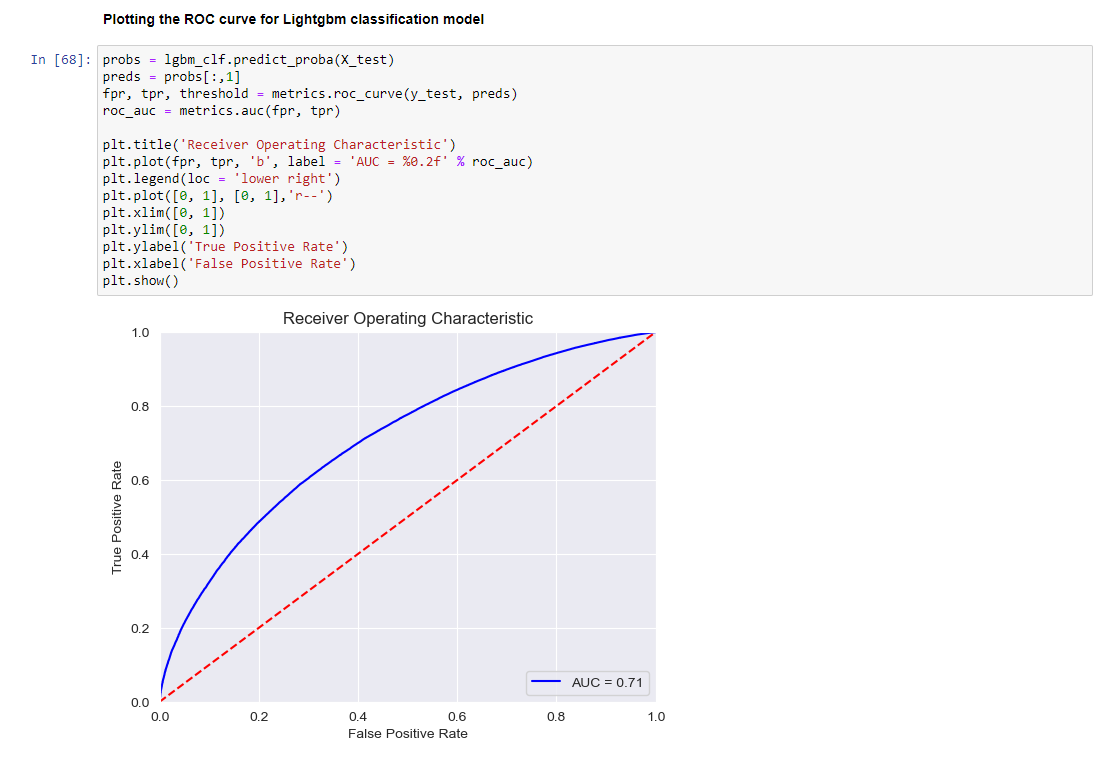
****

****

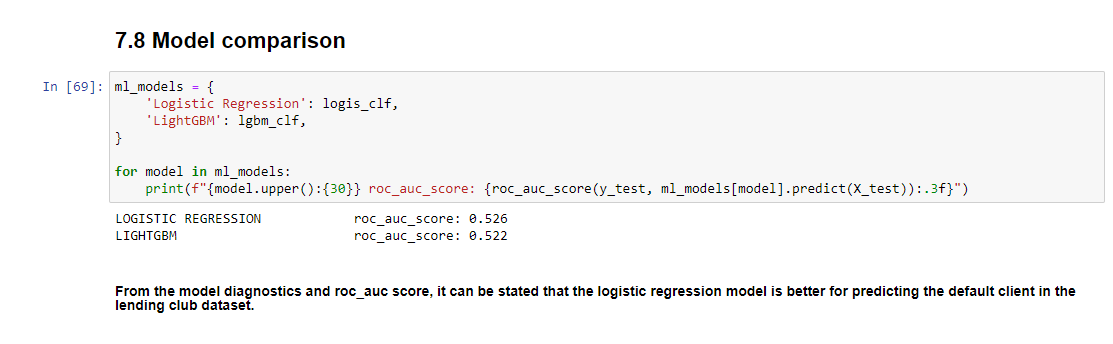
* + **Plotting Confusion Matrix**

****

* + **Plotting ROC Curve**

****

**MODELS COMPARISON**

****

**BUSINESS CASE 2**

**Problem Statement:** Defining a business case for building a linear regression model. Define Y and choose appropriate independent variables such that there should be a sound business reason to do this .Defining the use context for such a model clearly. Presenting the model and quantifying the benefits of the model.

**Business Model:** To predict the loan amount that can be borrowed.

To predict how much loan can be borrowed, we chose the following variables:

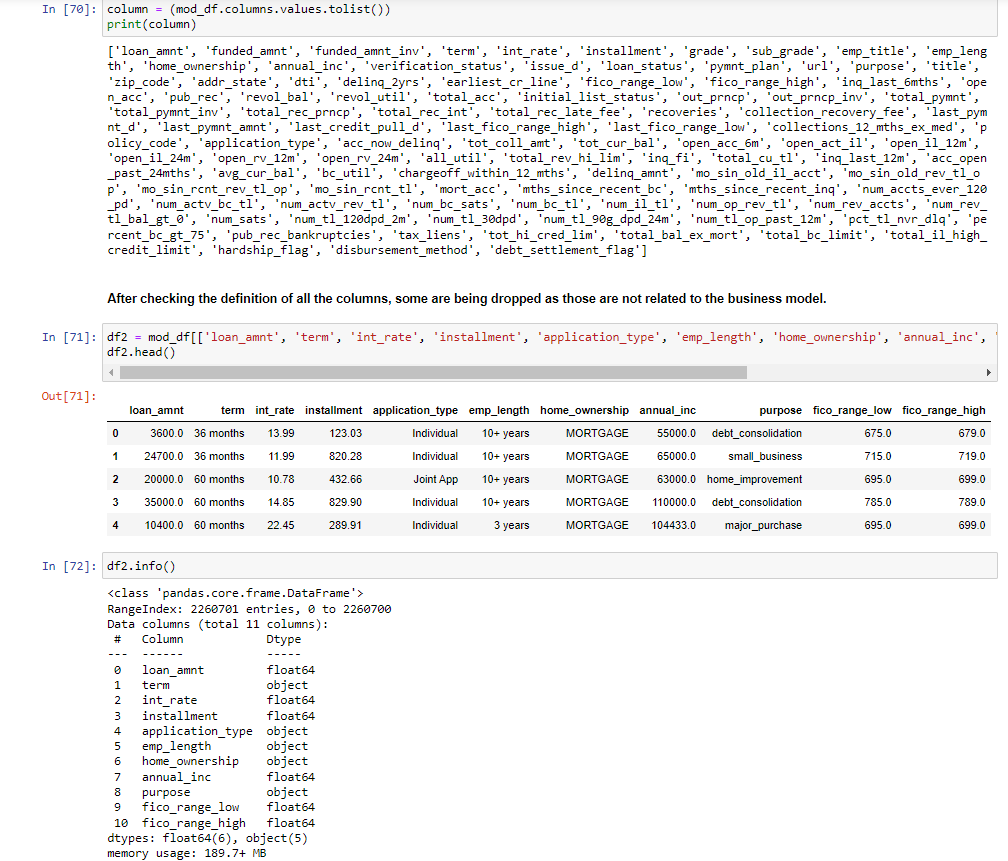
**Independent Variables:**

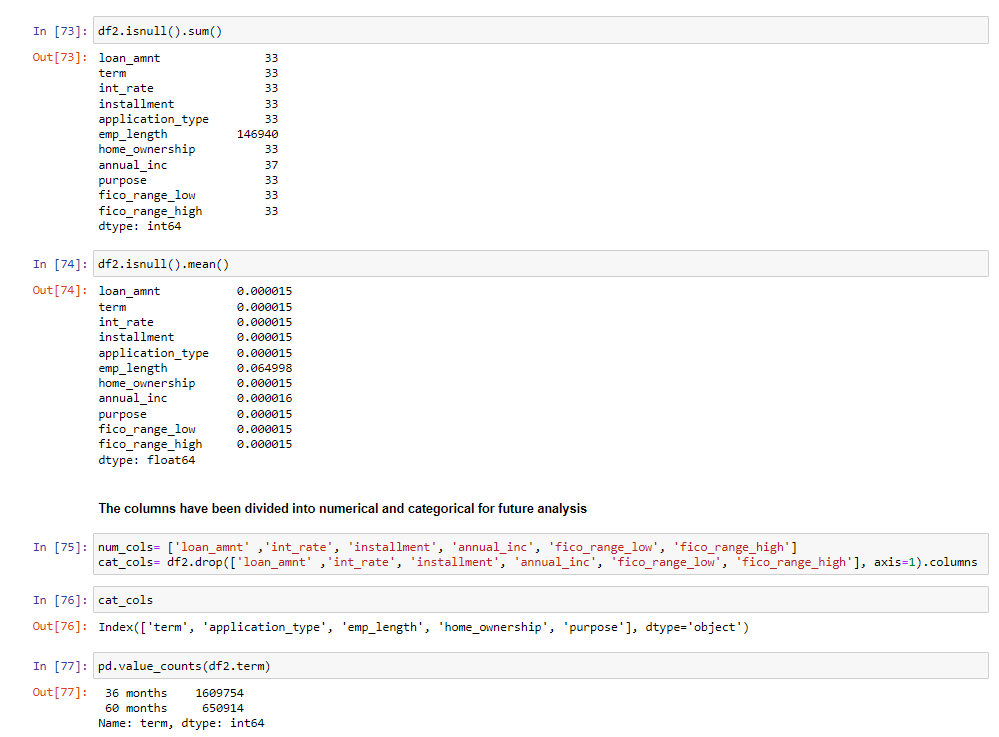
Term, interest rate, installment, employment length, home ownership, annual income, purpose, application type, and FICO range (low and high)

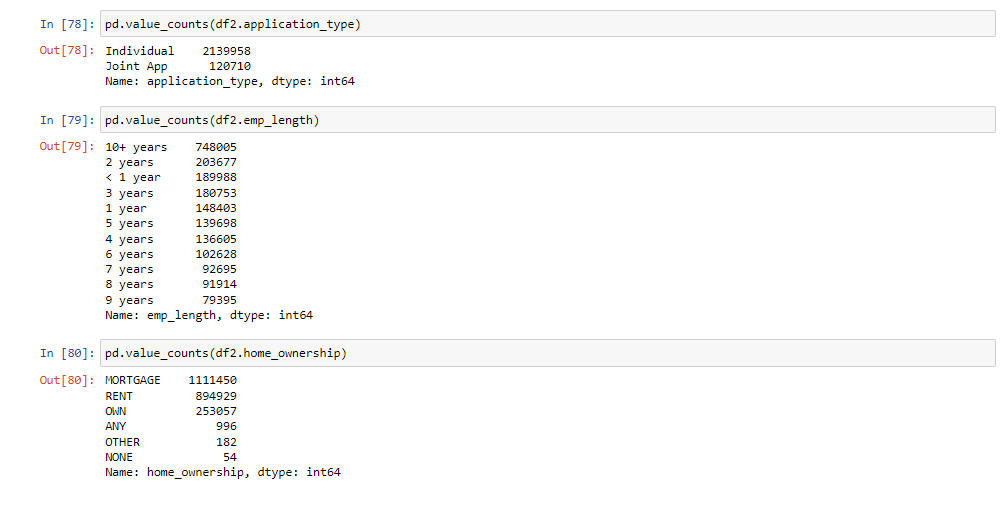
**Dependent Variable:**

loan\_amnt: the listed amount of loan applied for by the borrower

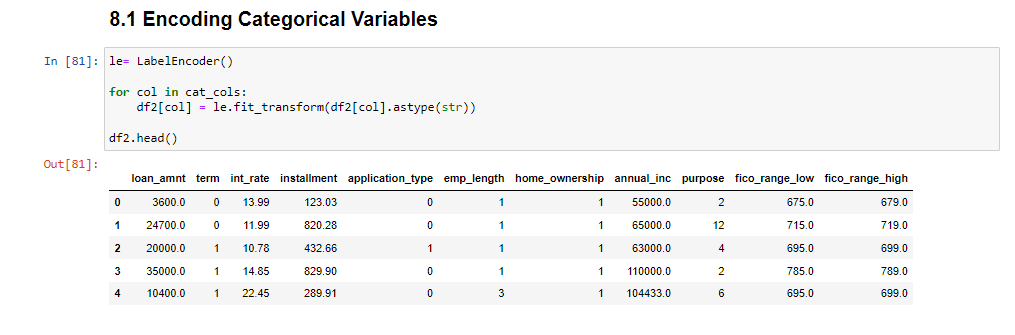
**Use Context:** A customer has approached Lending Club and requested a loan. Lending Club will collect the following information from her/him: Employee history, homeownership, annual income, the reason for the loan, loan type, high and low fico range, and so on.Investors can analyze the loan amount that it can provide, the interest rate, and the monthly installments based on the data shared by a customer and classify it as either default or repay.Based on the independent variables mentioned above, the model is designed to calculate an estimated loan that can be given to a customer.

****

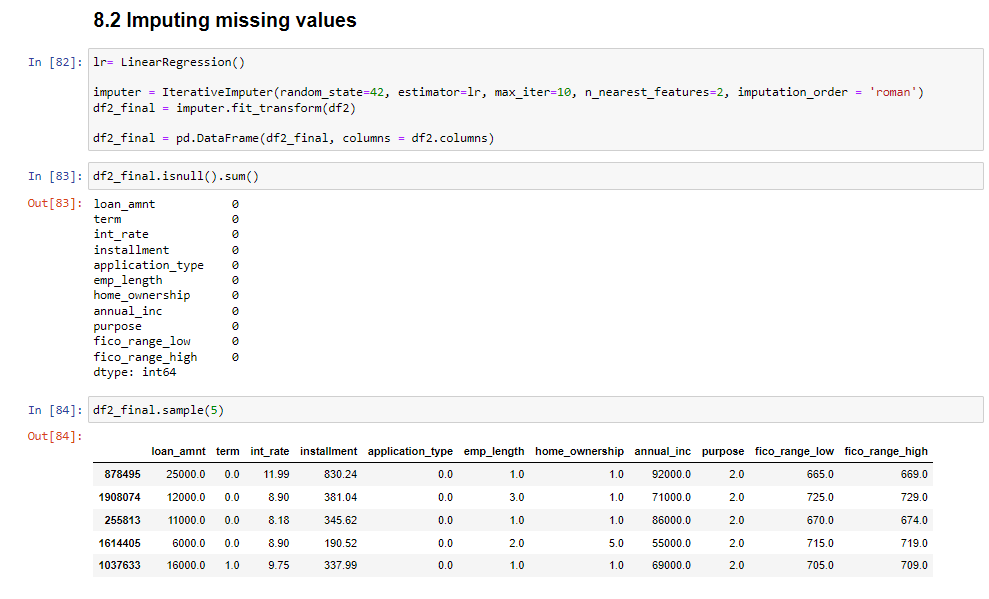
****

****

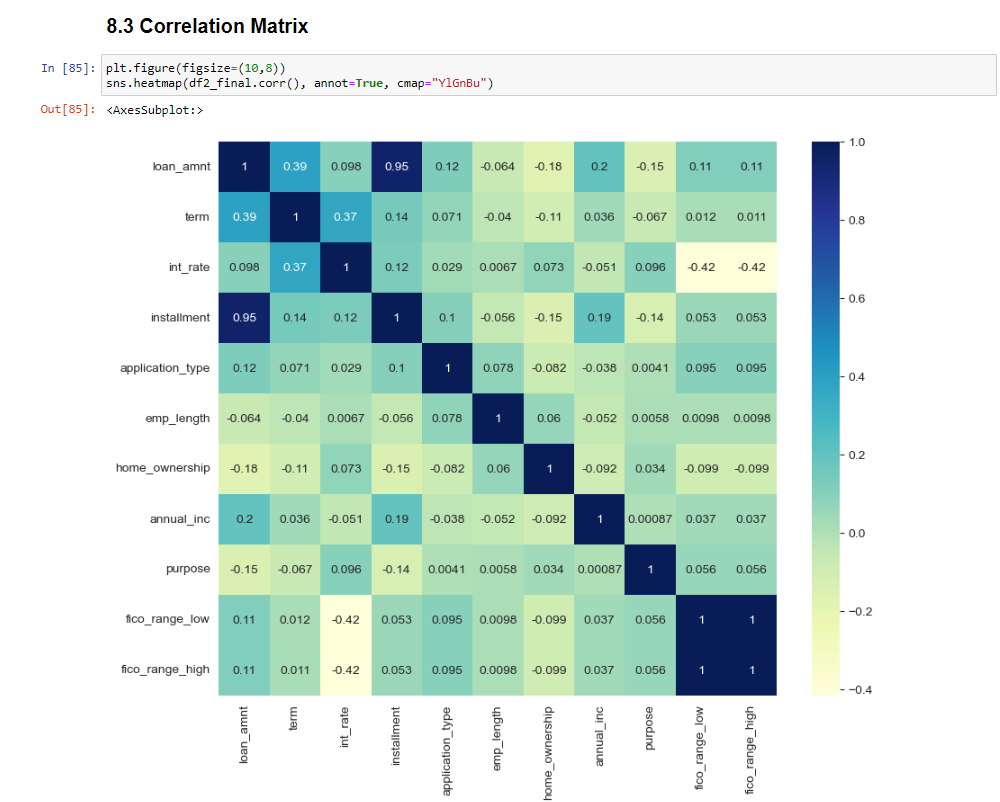
* + **Encoding categorical Values**

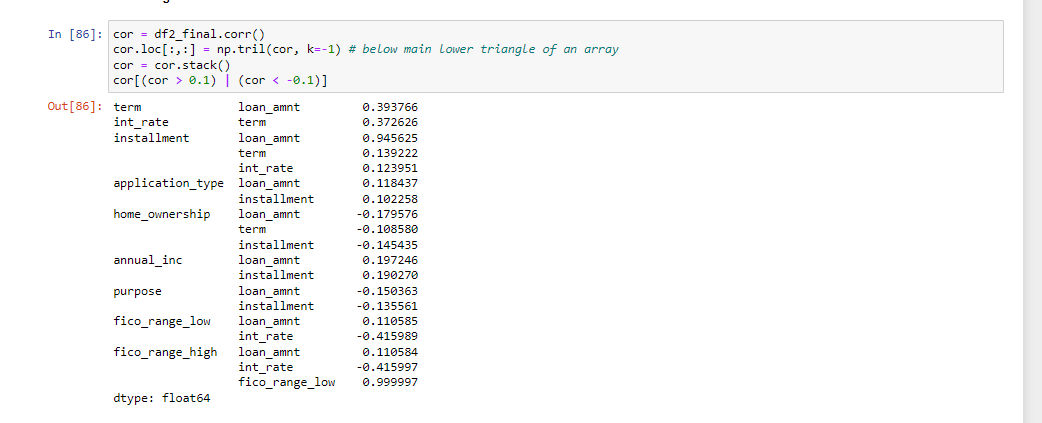
****

* + **Imputing Missing Values**

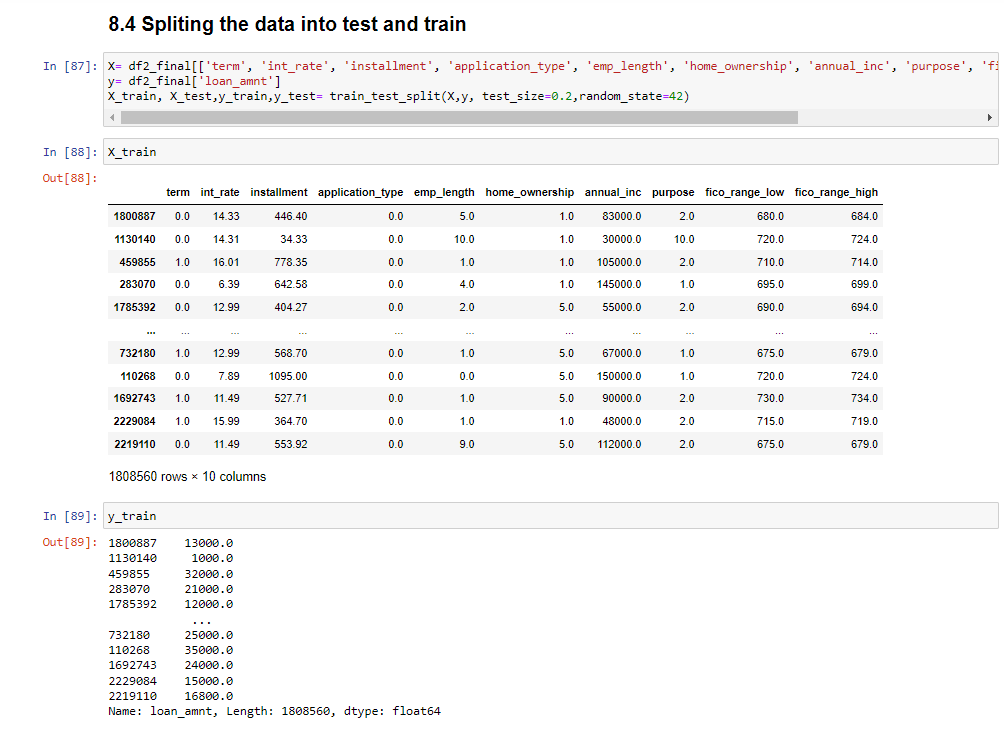
****

* + **Correlation Matrix**

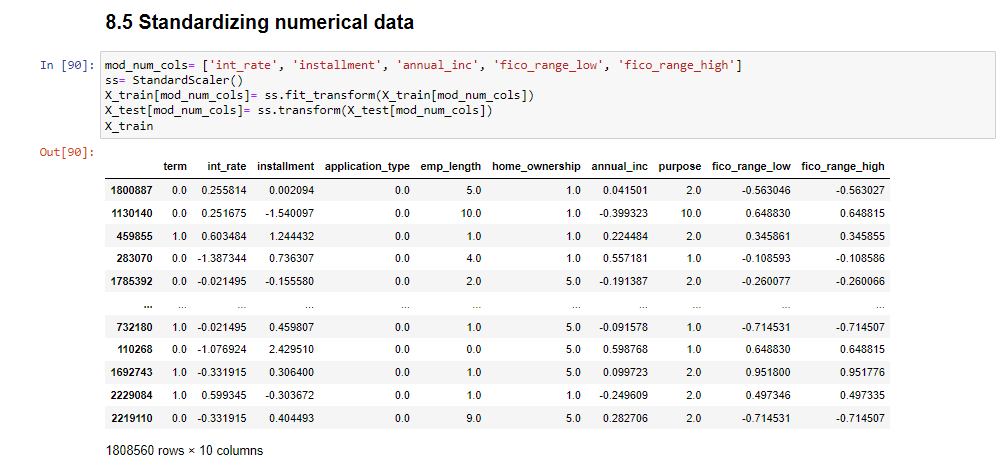
****

****

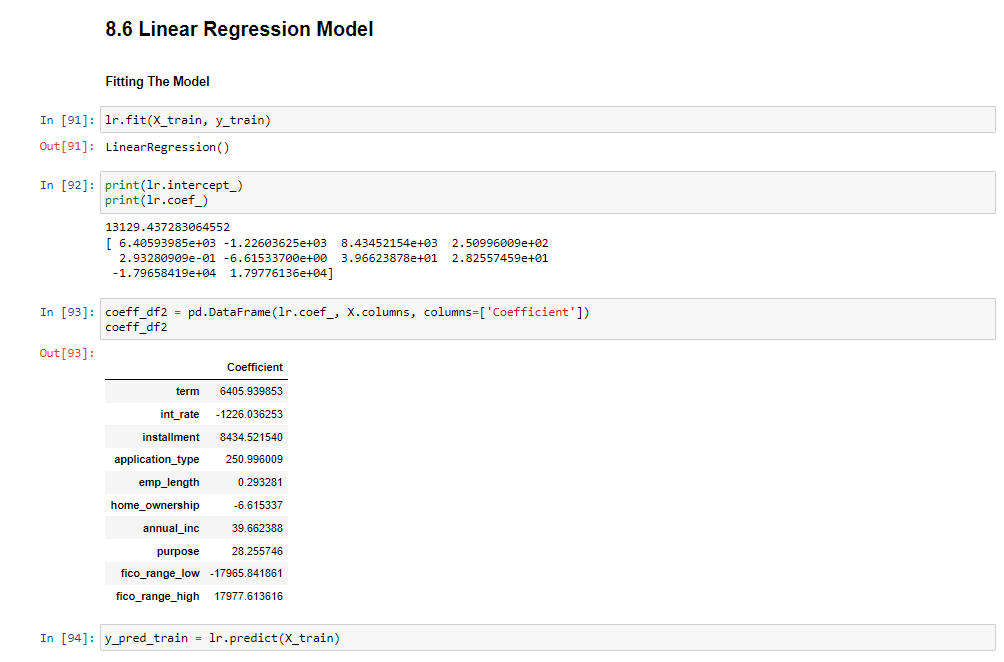
* + **Splitting into Train & Test**

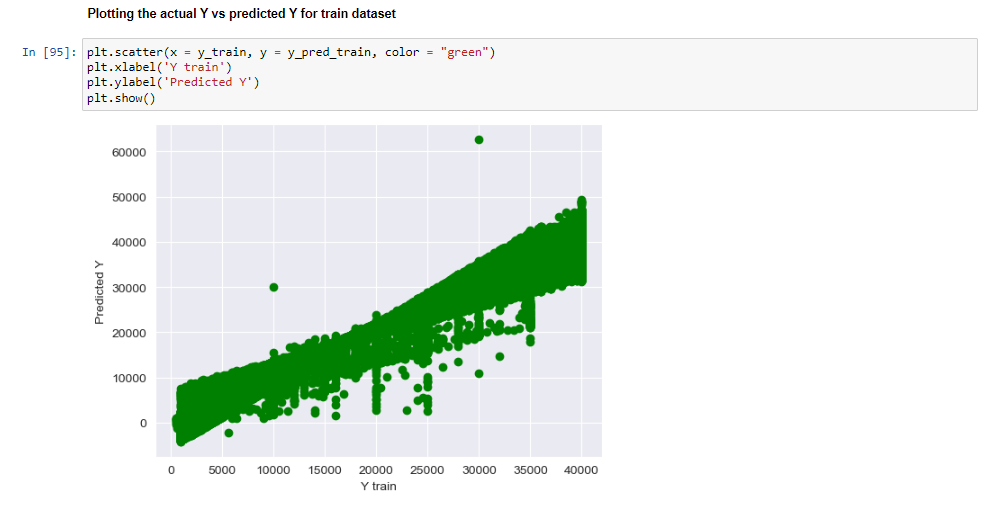
****

* + **Standardizing Numerical Data**

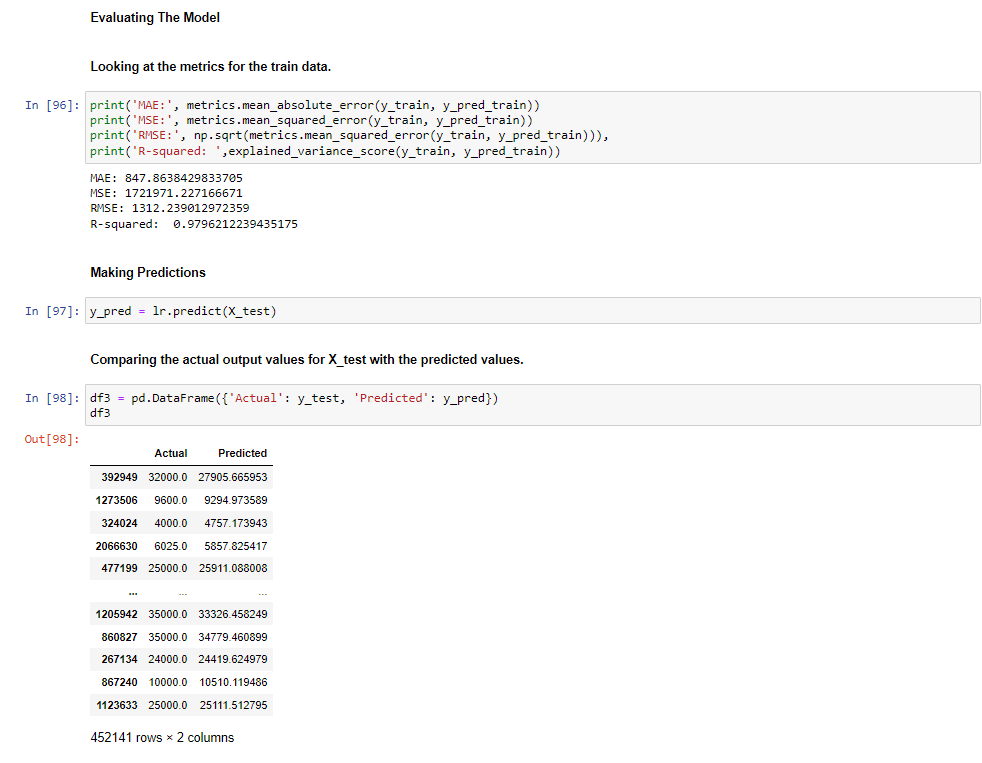
****

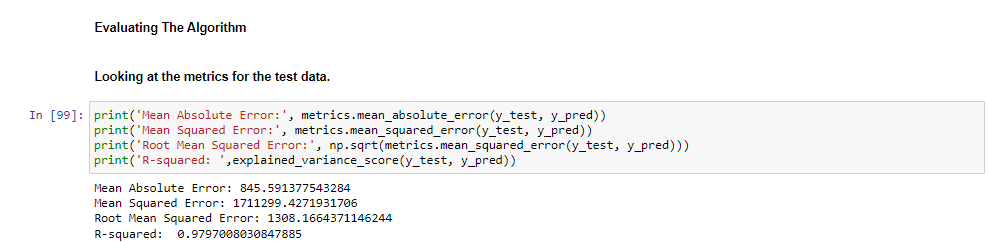
* + **Fitting Linear Regression Model**

****

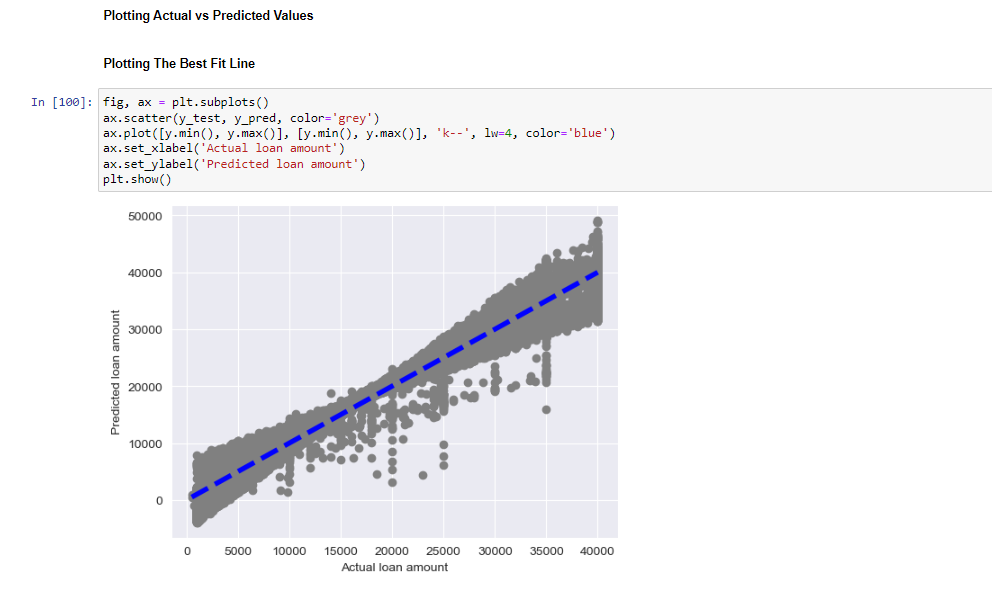
****

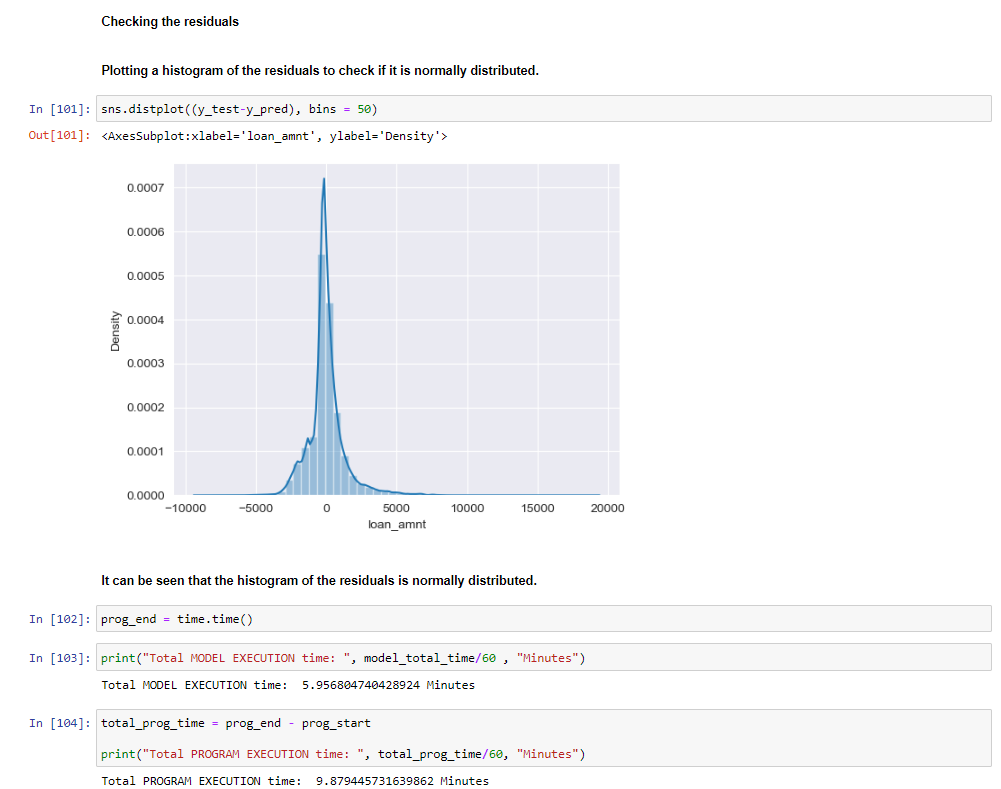
* + **Model Evaluation**

****

****

* **Plotting Actual vs Predicted**





* **Limitation**
  + Few models take more time to execute, sometimes the kernel gets crashed. Hence, Higher RAM is preferred.
* **Future Scope**

To increase the process time, the following processes can be executed on GPUs

as it takes most of the time running in series:

* + Reading the dataset
  + Calculating model scores
  + Calculating Accuracy