Capstone Project 1- Predicting Loan Default

Data Wrangling

Contents

[The Lending Club Data Set 2](#_Toc3559187)

[Libraries Used 2](#_Toc3559188)

[Data Wrangling 2](#_Toc3559189)

# The Lending Club Data Set

The Dataset used in this project is The Lending Club dataset available from kaggel website. (https://www.kaggle.com/wordsforthewise/lending-club/home). It is a real world data set which contains 2004126 rows of loan listing and 150 columns (attributes) of the each loan listing from year 2007 to 2018 Q2. Out of 150 columns, some columns have missing data and some columns are not needed for the analysis. After applying several data wrangling method, data set contains 2004095 rows and 83 columns. 4 columns are of data type *category*, 4 columns are of data type *datetime64*, 66 are of data type *float64*, and 9 are of data type *object*.

# Libraries Used

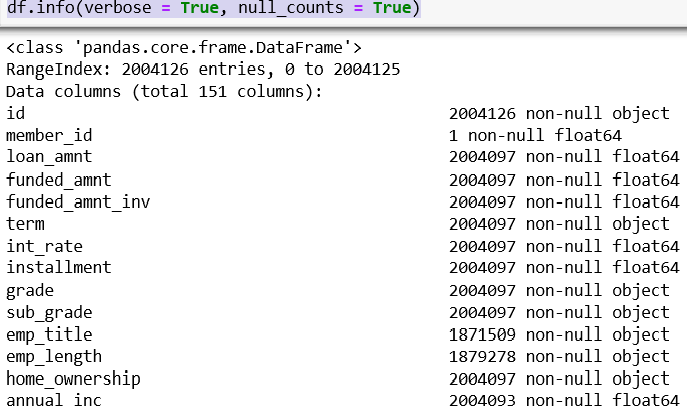
Besides Python standard libraries, Pandas, Matplotlib.pyplot, Numpy and Seaborn libraries are used.

# Data Wrangling

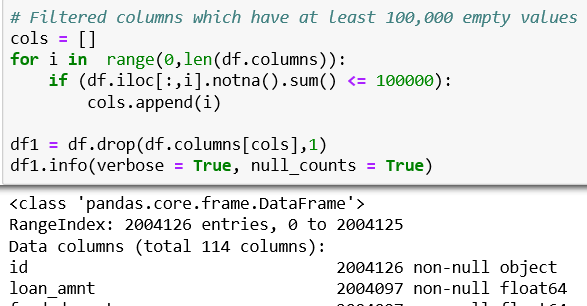
* The dataset is available in csv format which is loaded in memory as Dataframe using pandas library function read\_csv().

df = pd.read\_csv('accepted\_2007\_to\_2018Q2.csv', low\_memory = False)

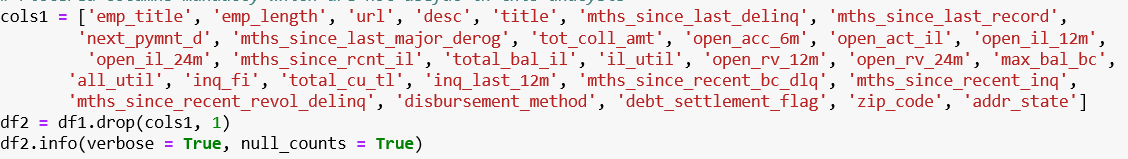
* df.info() shows that the data set has relatively large and has missing values and columns that are not useful for the analysis. Both columns and rows are dropped to reduce the size of table.



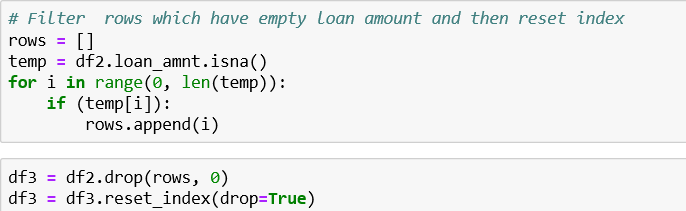
* Drop Columns:
  1. Dropped columns which have more than 100000 missing values.



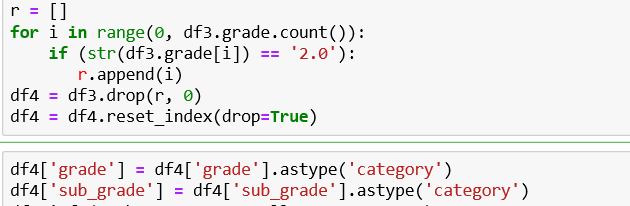
* 1. Using manual inspection, identified columns which are not useful for this analysis such as ‘emp\_title’, ‘url’, ‘desc’ etc. Dropped these columns.



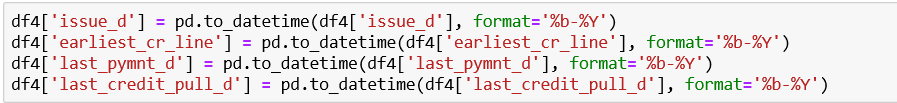
* Filter Rows
  1. Df.info() shows that there are rows in which ‘loan\_amnt’ column has missing values. Loan\_amnt is a critical column for the analysis. So filtered those rows.



* Convert data type from type Object to type Categorical.
  1. Columns ‘grade’, sub\_grade’, ‘loan\_status’, ‘applicaion\_type’ have category type data. So converted those into categorical data type.
  2. In this process found some row which have wrong values. Filtered those rows.



* Convert data type from Object to type datetime64.
  1. Columns ‘issue\_d’, ‘earliest\_cr\_line’, ‘last\_pymnt\_d, and ‘last\_credit\_pull\_d’ show dates. Converted them into datetime64 data type.



* Fill missing values of ‘annual\_inc’ column.

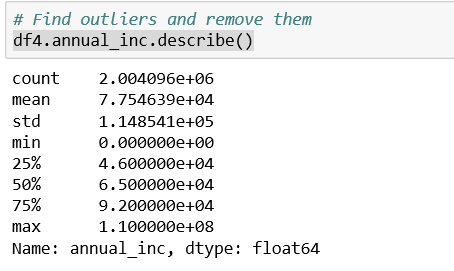
*. 1.* df4.isnull().sum() to display if there are missing values in ‘annual\_inc’ column.

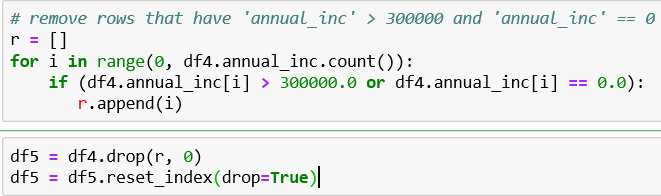


1. Filled the missing values with the median value of the ‘annual\_inc’ column.

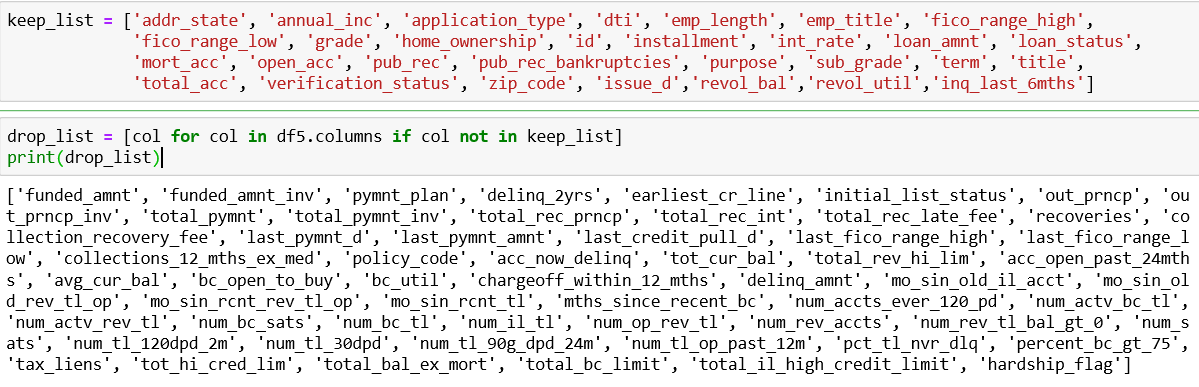


* Find outliers and remove them.

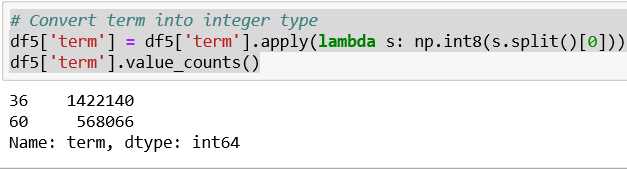




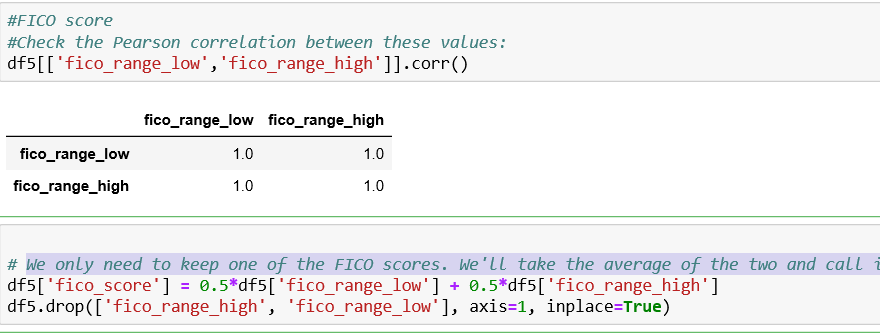
* For each of the features below, I checked the description in the Data Dictionary and only keep the features that would have been available to investors considering an investment in the loan. These include features in the loan application, and any features added by The Lending Club when the loan listing was accepted, such as the loan grade and interest rate.



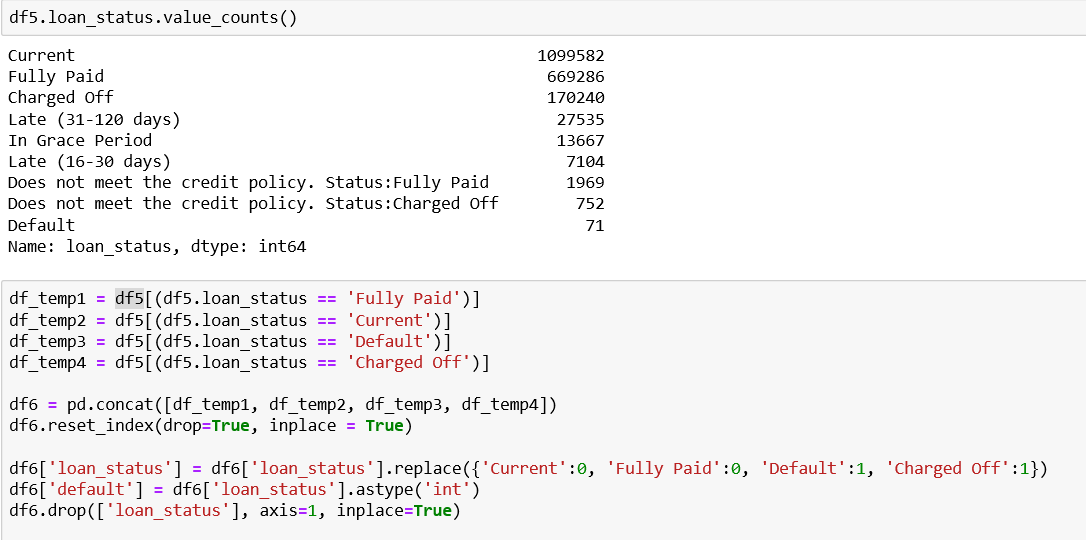
* ‘term’ is the loan period which has 36 months and 60 months possible values. Converted this from object data type to integer type.



* There are two columns for 'fico\_range\_low' and 'fico\_range\_high'. After checking the Pearson correlation coefficient, decided that both values are correlated. So we only need to keep one of the FICO scores. We took the average of the two and called it fico\_score.



* Change the response variable loan\_status to a 0/1 variable, where 0 indicates fully paid and 1 indicates default. Since actual number of defaults are really small, used charged off status as an indicator of default. Called this column Default because it makes more sense for our analysis.



* Dataset is ready for further Exploratory Data Analysis and Data Modelling. Saved the dataset as clean\_loan.csv.

