Lending Club Case Study

EDA Presentation

By

Group Facilitator: Mohammad Shahael Shah

Team Member: Saurabh Kumar

Table of Content

Problem Statement

We work for a Consumer Finance Company which specialises in lending various type of loans to urban customers. We have been provided data for past loan applicants with various attributes and data points for them along with information if they defaulted or successfully paid the loan.

Through Exploratory Data Analysis we need to identify how various factors impact the loan repayment of a customer, also how various customer and loan attributes influence tendency of load default. In this case study we will apply our knowledge of **Univariate analysis**, **bivariate analysis**, **derived fields/ deriving factors** to find insights in data set.

To start with different data analysis we would need to pre-process, clean and make data ready for analysis.

Technology Used:

We have done our analysis in Jupyter Notebook using Python 3.

Packages to be imported:

Before we start data cleaning we need to import all libraries that we will need in our EDA.

```
import pandas as pd
pd.set_option('display.max_rows', 130, 'display.max_columns', 130)
pd.options.display.float_format = '{:,.2f}'.format

import matplotlib.pyplot as plt
import seaborn as sns

import numpy as np
```

Data Cleaning & Pre – processing:

1. First thing would be to import the data in a data frame
2. Then we checked for data rows and columns , there are 111 columns with 39717 columns
3. Then we identified columns having all nulls or NaN (missing values)
4. We identified about 54 columns that were having missing values, we removed. Dropped then from dataset.
5. We further dropped columns which have too many missing in them.
6. After that we further identified columns which were having same values at most of places we dropped them as well.
7. After removing all columns we are left with 40 from 111 columns.

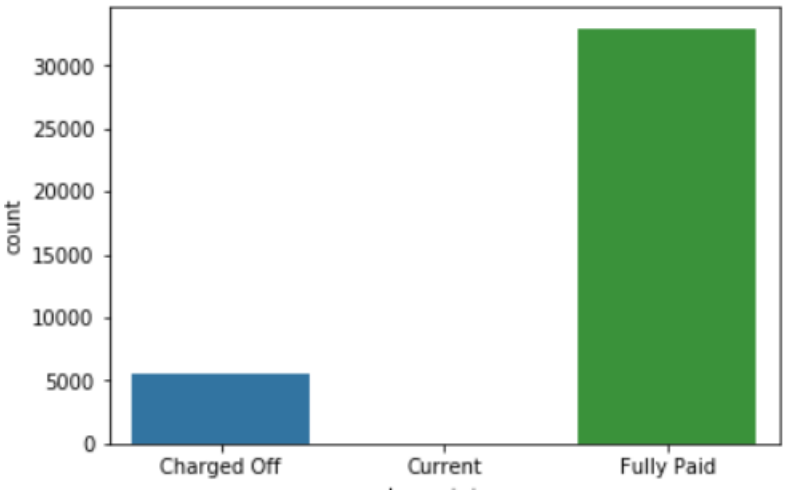
Derived Columns :

1. Let's create derived columns now : term_month to term , remove % from int_rate and create new column : int_rate_percentage
2. Derive emp_length_months from emp_length , remove years and + from values

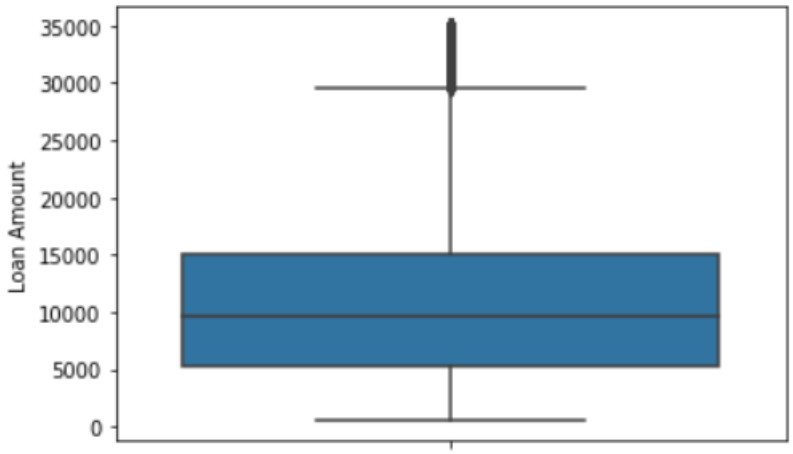
3. Remove % from revol_util and create a column revol_util_percentage
4. Similarly we need to create few more and drop old ones
5. Create another derived column PnL and drop old ones.
6. Now Update columns to float
7. Now identify the types of these columns. There are 25 numeric al , 12 categorical and 14 string features in the data set.

Univariate Analysis

1. For Loan Status, there are 32950 Fully paid records and 5627 ch arged off.

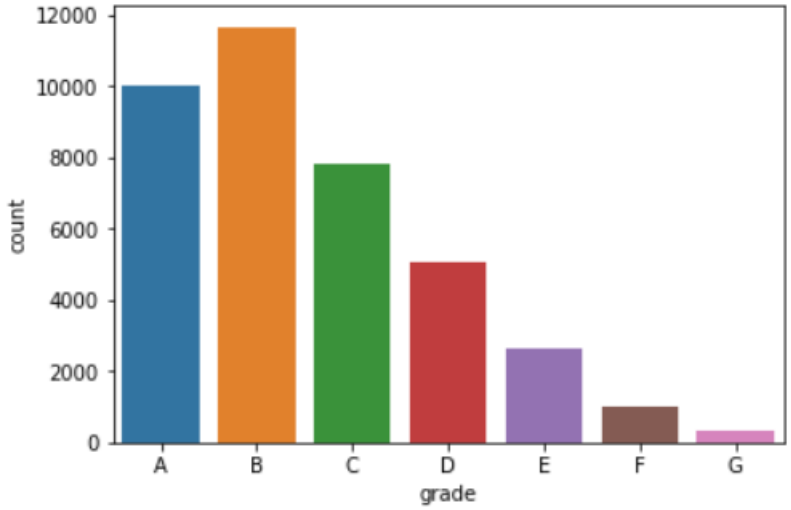


2. The loan amount varies from 0 to 35,000 having mean of 10,00 0.

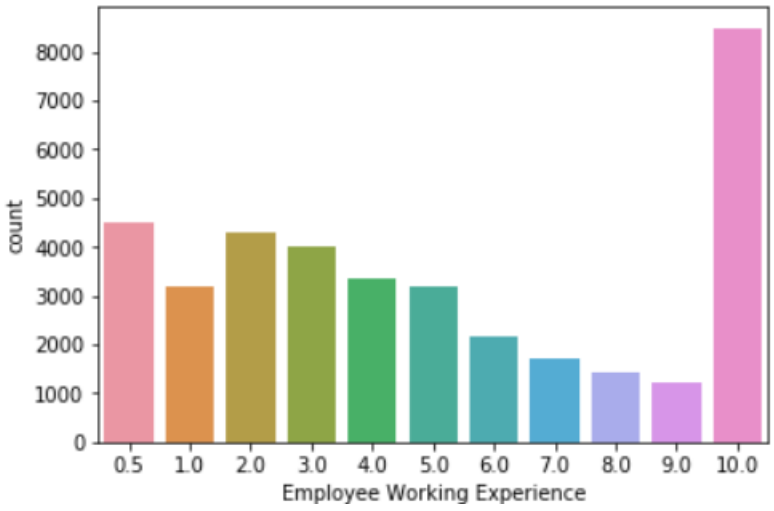


3. Most of the loans are Fully Paid.

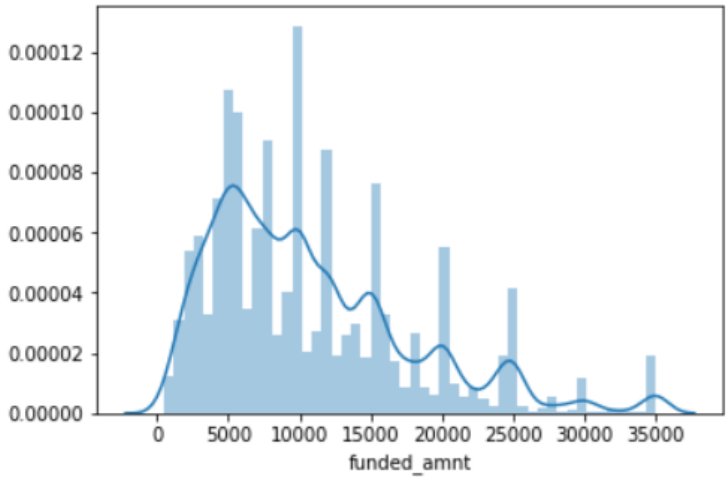
4. About 14% of loan are having status as defaulters.
5. We have a class imbalance here.
6. Most of the loans have grade of A and B. Therefore stating mos t of the loans are high graded loans



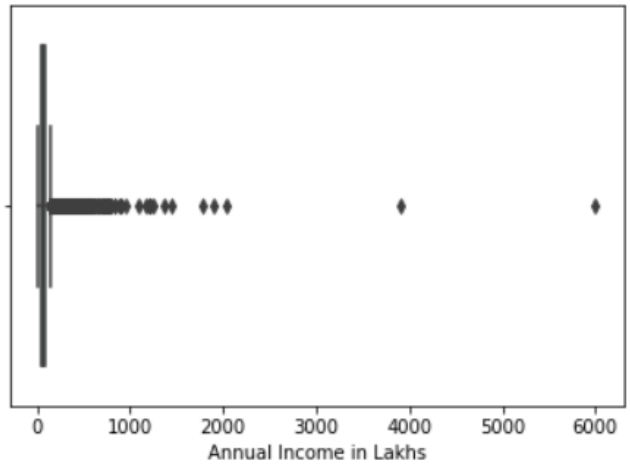
7. Majority of employees applying for the loan have more than 10 years of experience



8. Funded amount is left skewed. Most of the loan amount given i s 5 lakhs

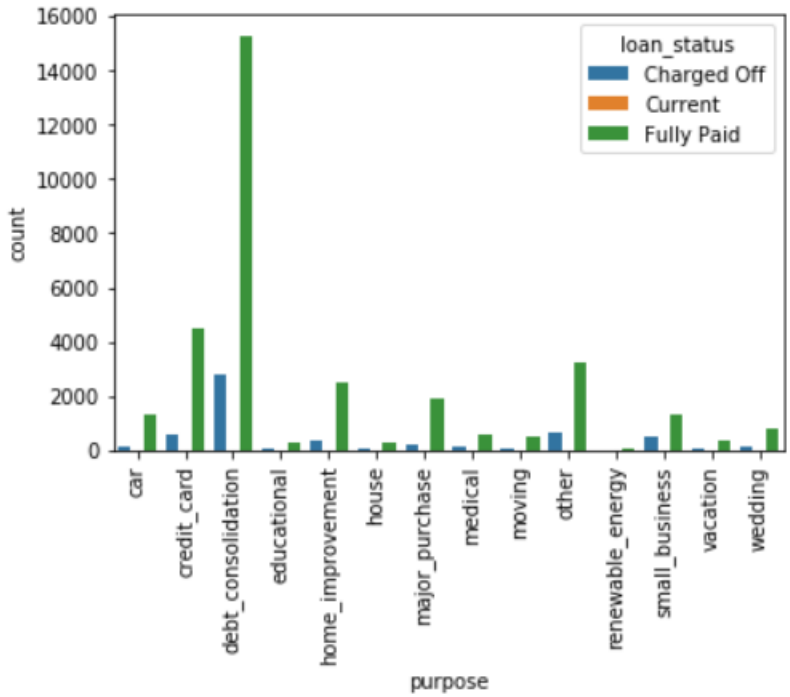


9. There are only two applicants having annual income of more th an 30 lakhs

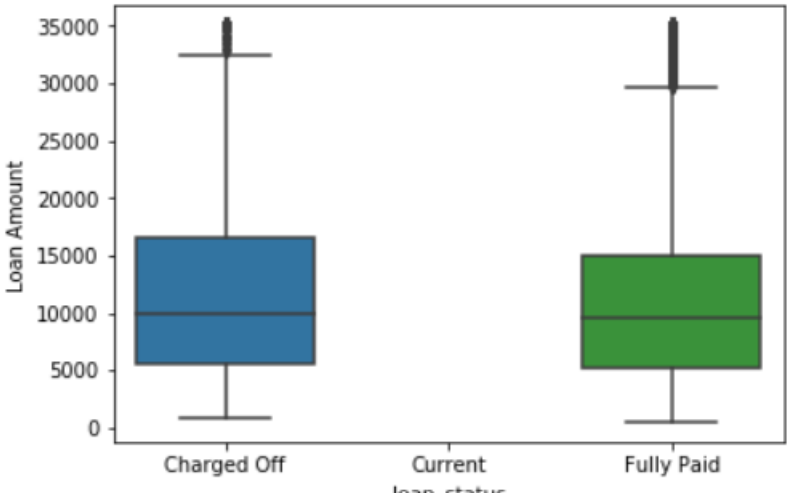


Segmented Univariate Analysis

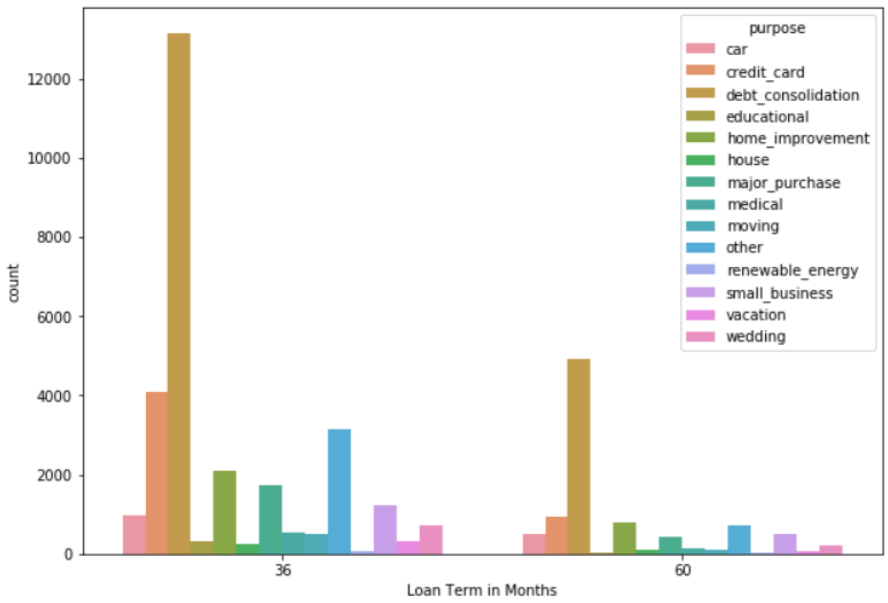
1. Majority of loan has been given for the debt consolidation p urpose and has been fully paid.



2. Mean, 25% and 75% Loan amount of Fully paid and charged off is exactly same

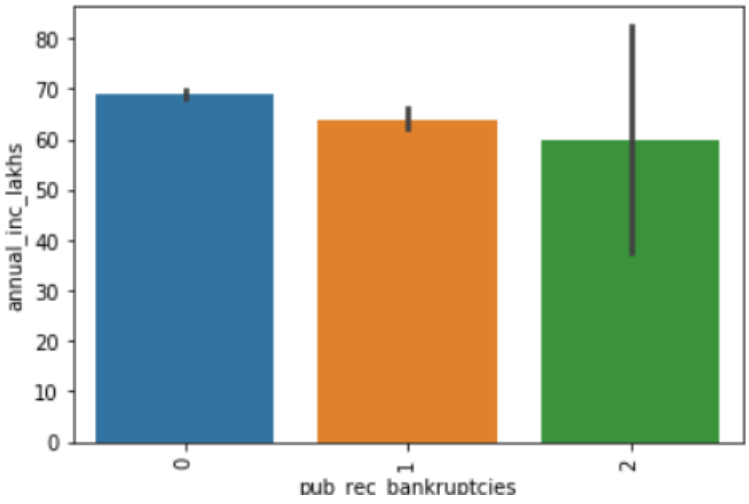


3. Tenure of 36 months have high chances to be defaulters

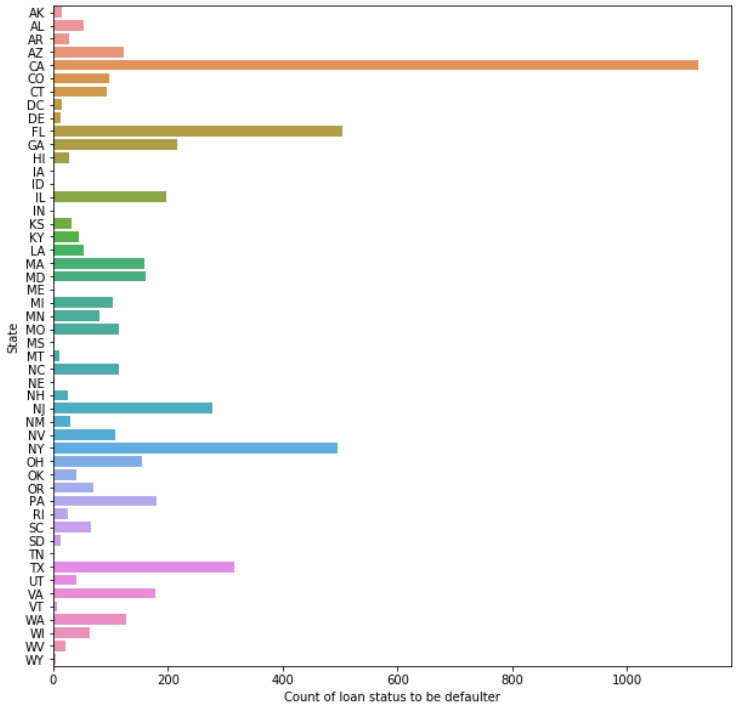


Bivariate Plots

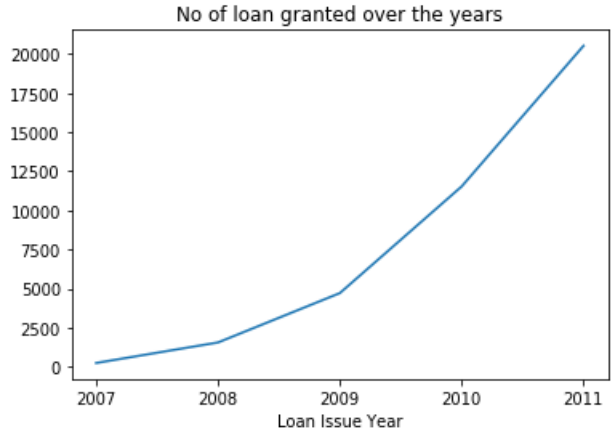
1. Off All Publicly recorded Bankruptcies , Customers with slightly higher income has no bankruptcies .



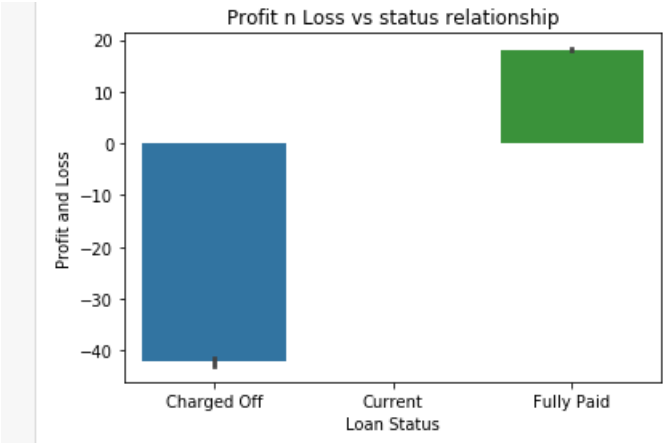
2. Applicants from the state CA are having high probability to be d efault .



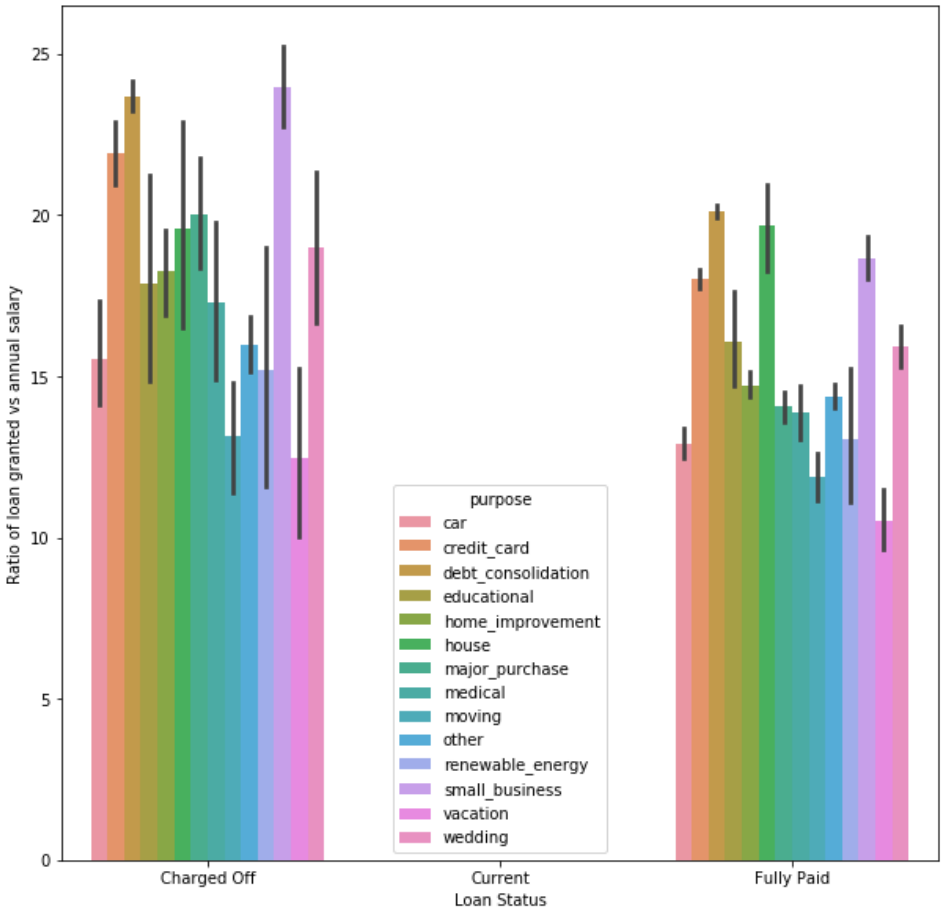
3. There is an increment in loan applicants from 2007 which incre ases with much higher rate from 2010 to 2011



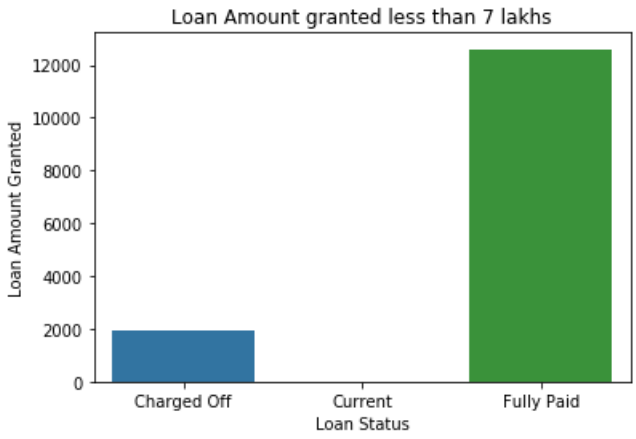
4. Most customers who have defaulted suffers losses.

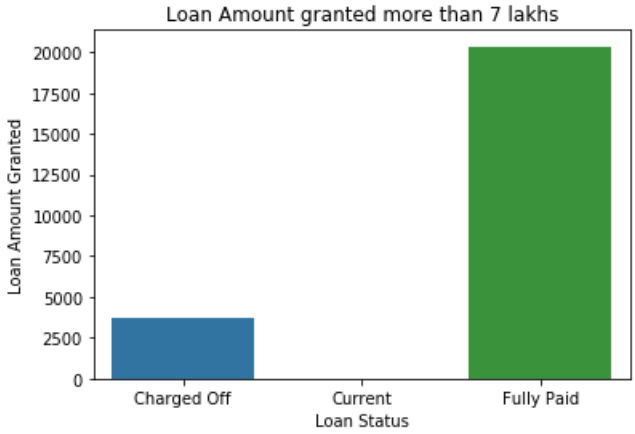


5. If we see from Ratio of loan granted to annual income , custom ers with higher ratio have defaulted more .

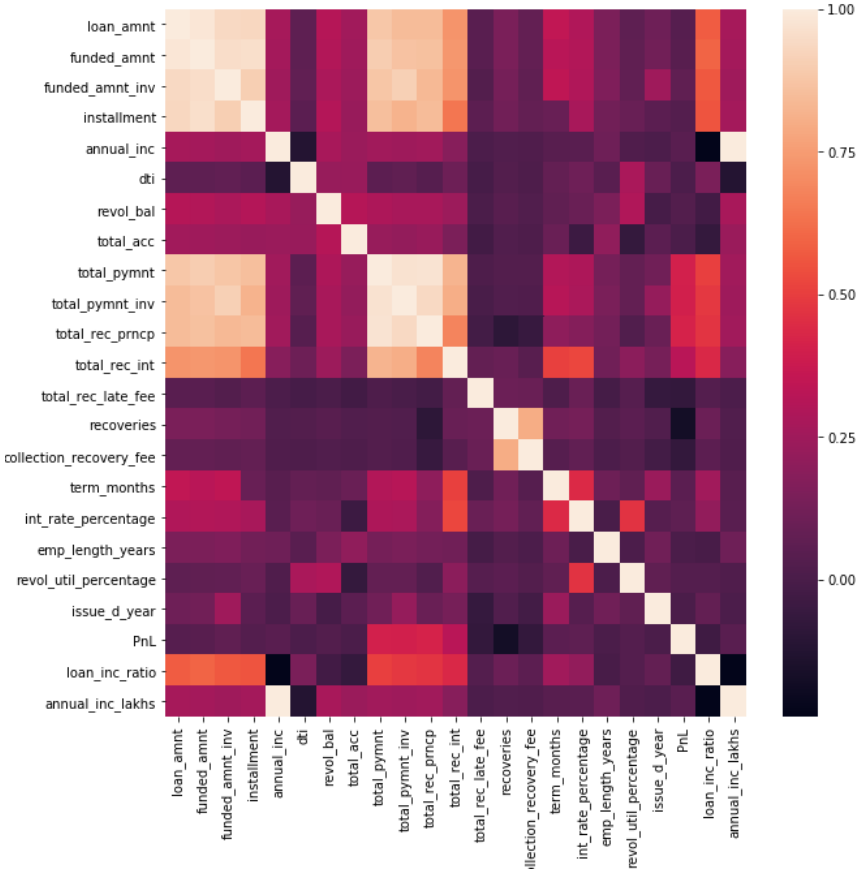


6. Another insight is customers with higher annual income have fu lly paid loan amount.





7. From Correlation Metric we can identify that Loan Amount, Fun ded Amount and Instalment are highly correlated positively , Tota l Payment columns too are significantly related in a positive way t o all mentioned in prior statement .Loan Amount is not too much related to Emp_Length years , interest rate and term months.



Conclusion

As part of case study we first processed the data which takes most of the time in analysis, included fixing rows , columns , deriving new columns , managing missing values etc.

Then we identified which are numerical variables and which are categorical, further identifying Integer, float and strings.

Then we performed Univariate analysis numeric variables to see how various variable are distributed to find out ranges and quartiles by graph.

Similarly we did Univariate analysis of segmented variables to find how they impact loan status in data set.

Finally we did bivariate analysis to identify how one variable is related / correlated with other and how change in values for one impact other.

Acknowledgement/References:

https://www.kaggle.com/code/lonewolf95/eda-101-univariate-analysis-for-beginners

https://www.geeksforgeeks.org/get-unique-values-from-a-column-in-pandas-dataframe/

https://www.kaggle.com/code/kashnitsky/topic-1-exploratory-data-analysis-with-pandas

https://www.tutorialspoint.com/how-to-select-all-columns-except-one-in-a-pandas-dataframe#:~:text=To%20select%20all%20columns%20except%20one%20column%20in%20Pandas%20DataFrame,%5D

.