

EDA – CASE STUDY - LOAN

Anu Raje
Vinay Venkatesh

AGENDA

Introduction

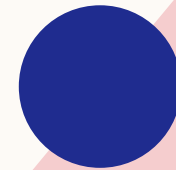
Data Cleaning and Outlier Management

Analysis of Univariate and Bi/Multi variate

Observation

Q&A

Annexure



INTRODUCTION

In this case study, we will apply the EDA techniques towards developing a basic understanding of risk analytics in financial services

Understand how data is used to minimize the risk of defaulting customers when financial institutes lend money.

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

DATA CLEANING AND OUTLIER MANAGEMENT

4

Data cleaning is important because of the following reasons:

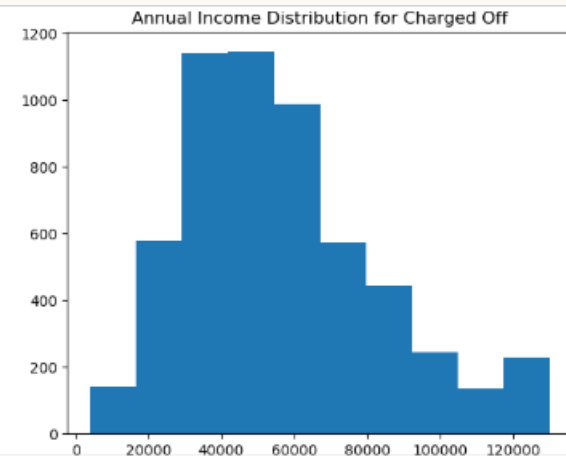
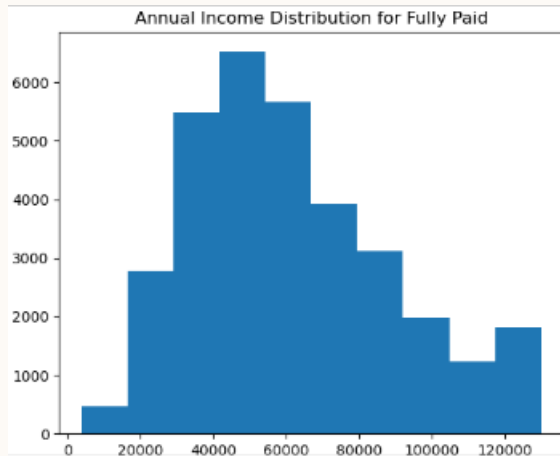
1. Improve data accuracy
2. Improve data quality
3. Helps reduce bias
4. Save time and resources

Brief description of data cleaning activities done in this case study is given below:

1. Annual_Inc - Column had multiple outlier. So outlier have been treated
2. Int_rate - Removed % from the text, replaced blanks with 0 and changed data type to float
3. Issue_d, last_credit_pull, earliest_cr_line - Split the field into Year and Month
4. Emp_length -
 - If <1 and NA yes, consider 0
 - 10 + Years = 10
 - Remaining (Example: 5 Years) - Remove years (Example: 5).
 - Convert to INT
5. Term - Remove "months" convert to int64-
6. For 70 columns with more than 50% missing data or all rows containing same data are dropped

UNIVARIATE ANALYSIS

5

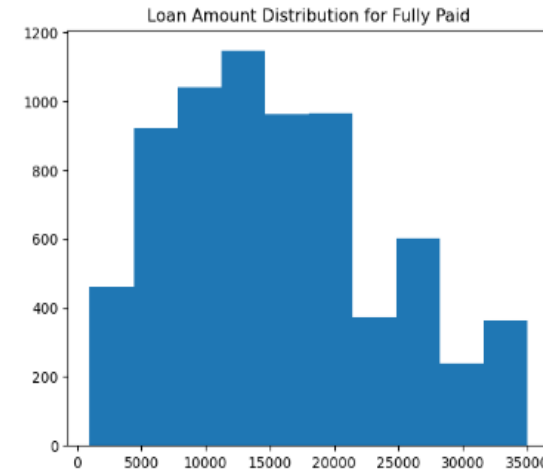
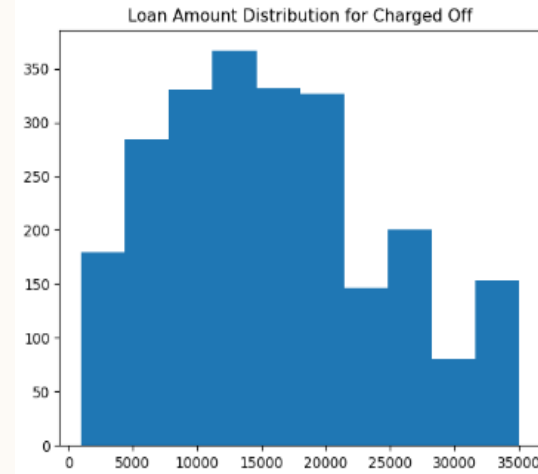
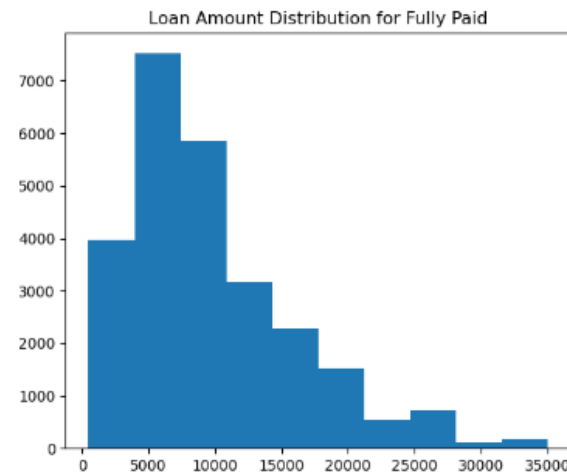
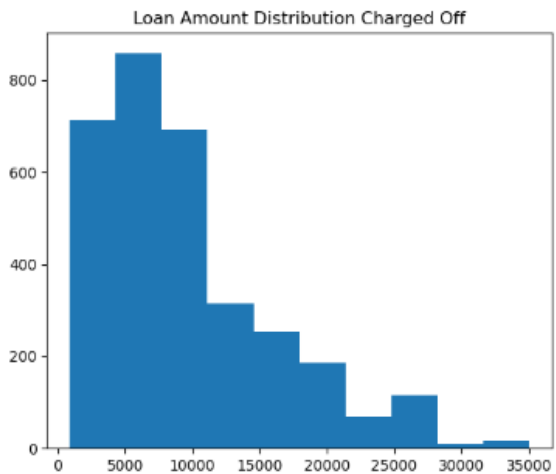


Annual Income distribution

- 30k-50k is the peak income of charged off members
- 50k-60k is the peak income of fully paid members.

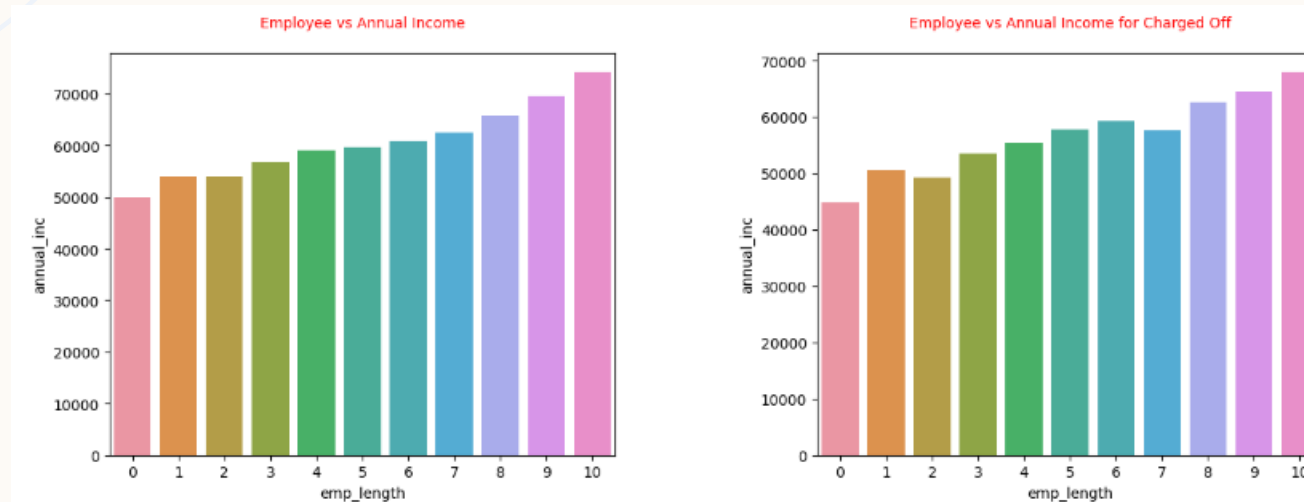
Short Term Loan Tenure

Long Term Loan Tenure



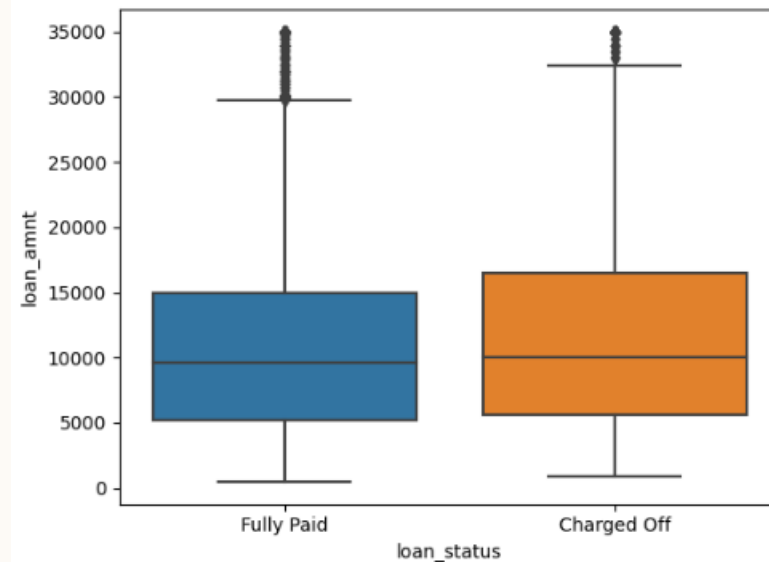
- For Short Term tenure, among charged off members, higher number of charged off is when the loan amount is between 500-8000
- For Long Term tenure, among charged off members, higher number of charged off is when the loan amount is between 500-5000

BI/MULTI VARIATE ANALYSIS



Employee Length vs Average Annual Income.

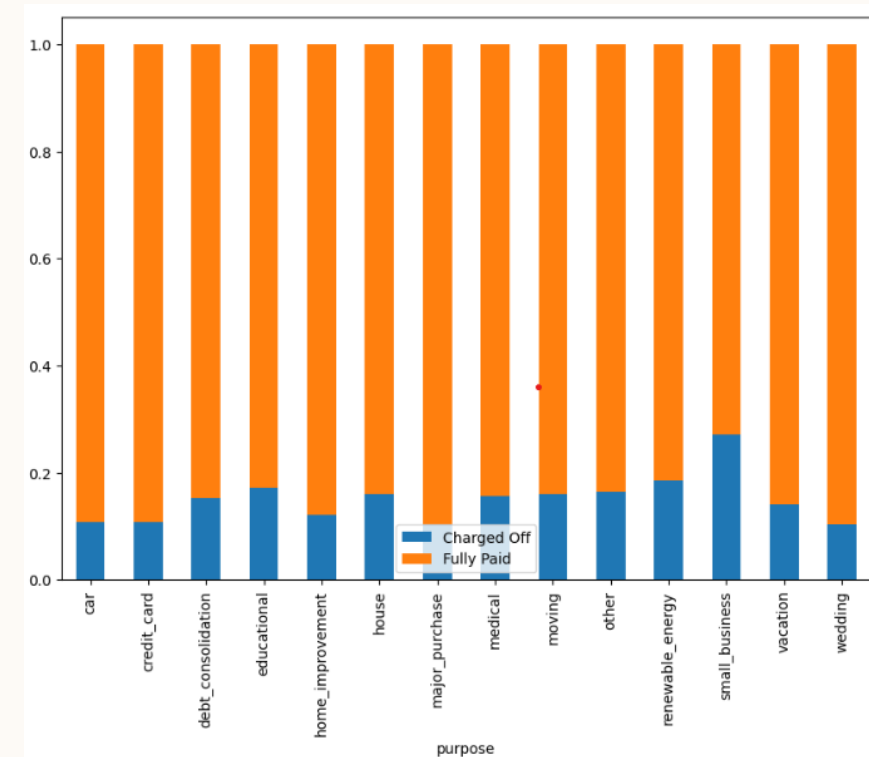
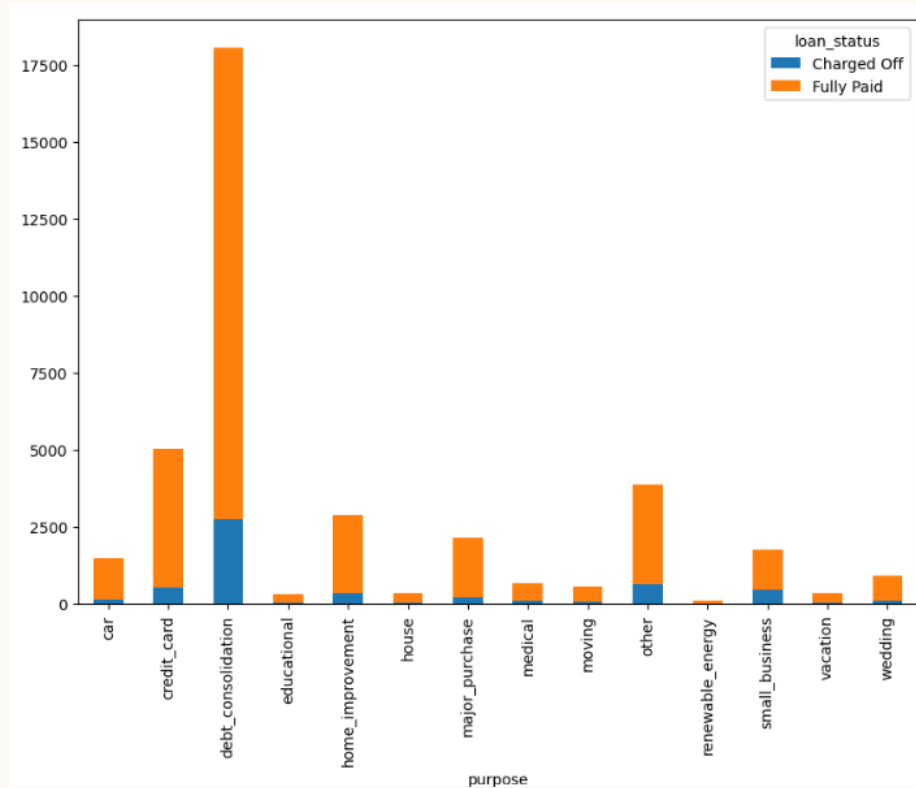
- For all 3 loan status (Current, Charged off and Fully paid), the average annual income is increasing with the Employee Length.
- For Loan Status - Charged Off, the average annual income is lower compared to all other Loan status considered.



Loan Amount Box plot for Fully Paid Vs Charged off

- The loan amount taken by Charged off members between 25-75 percentile range is slightly higher than the loan amount taken by Fully paid members.

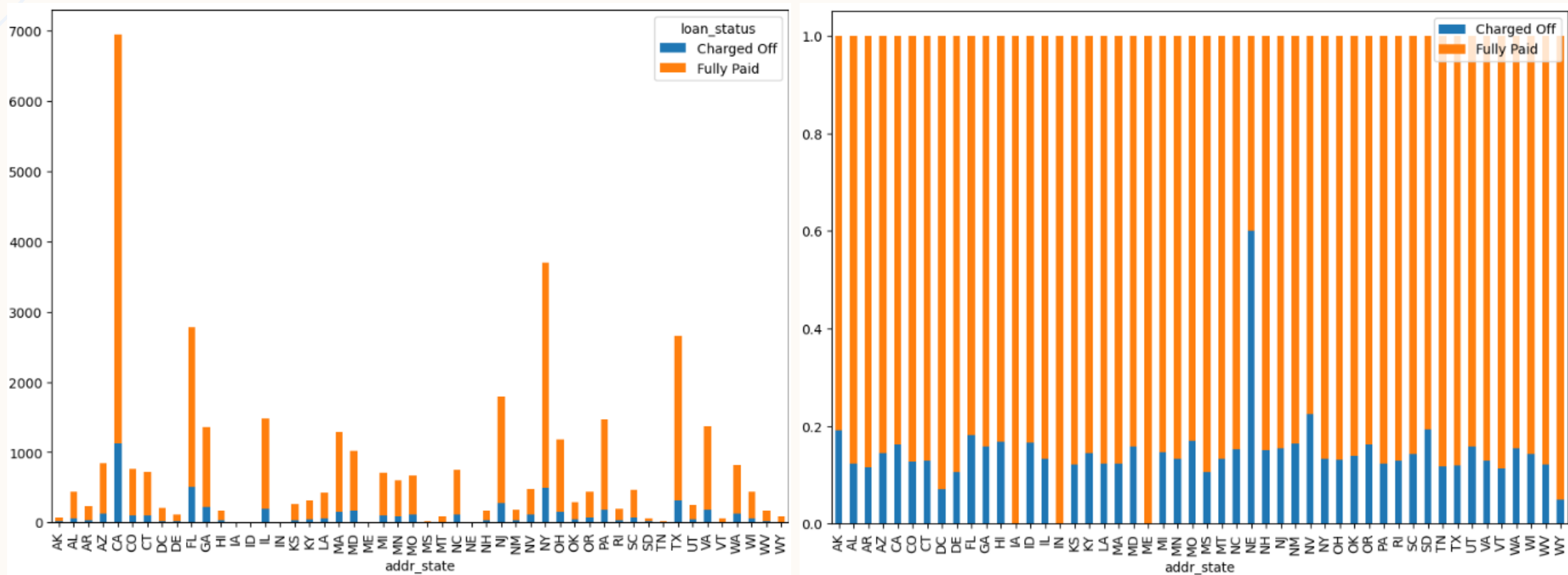
BI/MULTI VARIATE ANALYSIS



Count fully paid and charged off members across various purpose for which they avail loans

- There is an outlier that members with purpose as "Credit Card" and "Debt consolidation" are at higher risk of loan defaulting.
- There is a clear outlier from percentages that members with purpose as "Small Business" are at higher risk of loan defaulting

BI/MULTI VARIATE ANALYSIS



Region wise split of Fully Paid Vs Charged Off – Count and Percentages

- Members from state of CA, FL, NY have higher count of defaulters.
- From a percentages perspective the states (NE (count is low), NV, SD and FL) have higher defaulters and thus a risk states to grant loans to members.

OBSERVATIONS

9

Following driver variables were used for analysis:

1. Annual Income, Region, Term, Loan Amount and Employee Length

Univariate:

1. For Short Term tenure, among charged off members, higher number of charged off is when the loan amount is between 500-8000.
2. 30k-50k is the peak income of charged off members while 50k-60k is the peak income of fully paid members.

Bi/Multi Variate Observations:

1. For all 3 loan status (Current, Charged off and Fully paid), the average annual income is increasing with the Employee Length.
2. For Loan Status - Charged Off, the average annual income is lower compared to all other Loan status considered.
3. The loan amount taken by Charged off members between 25-75 percentile range is slightly higher than the loan amount taken by Fully paid members.
4. There is a clear outlier that members with purpose as "Credit Card" and "Debt consolidation" **are at higher risk of loan defaulting**
5. It can be observed from percentages pf defaulters that members with purpose as "Small Business" **are at higher risk of loan defaulting**
6. Members from state of CA, FL, NY have higher count of defaulters. **Members availing loan from the mentioned regions are at risk of defaulting.**
7. From a percentages perspective the states (NE (although count is low), NV, SD and FL) **have higher defaulters and thus risk states to grant loans to members.**

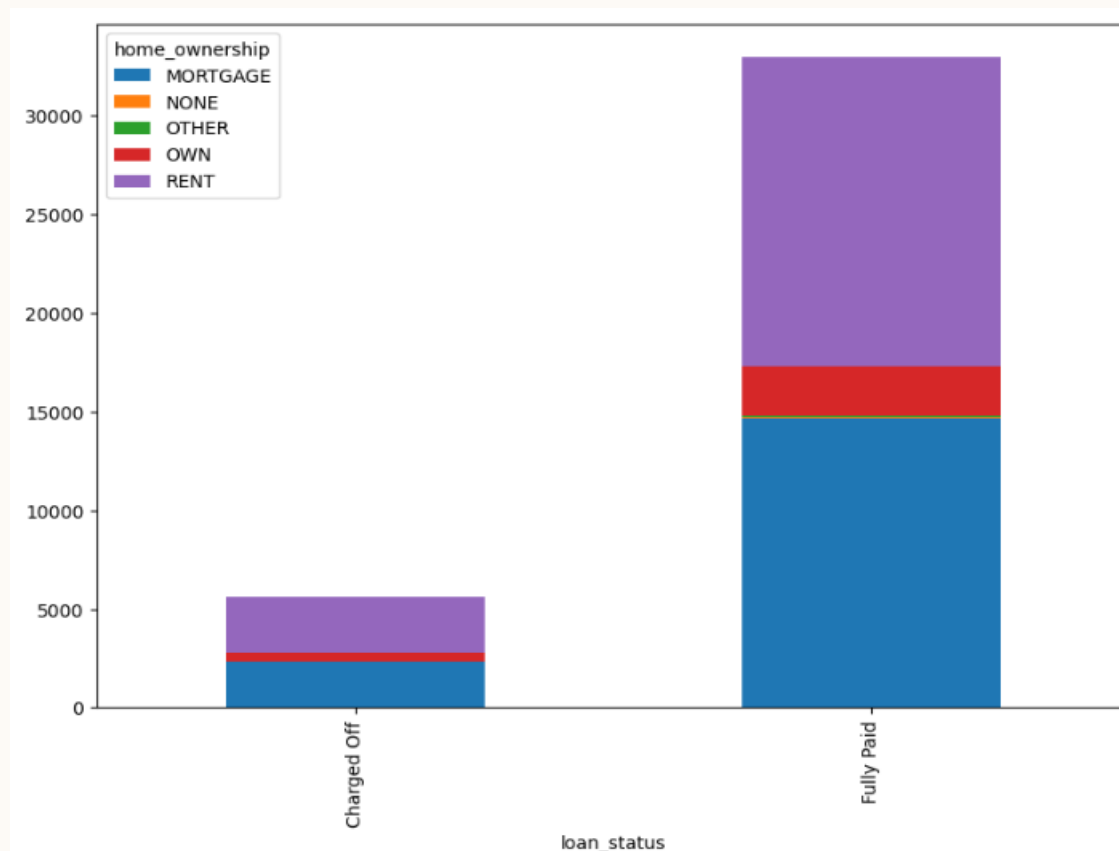
THANK YOU



ANNEXURE

NOTE: GRAPHS WHERE WE COULD NOT DRAW CONCLUSION

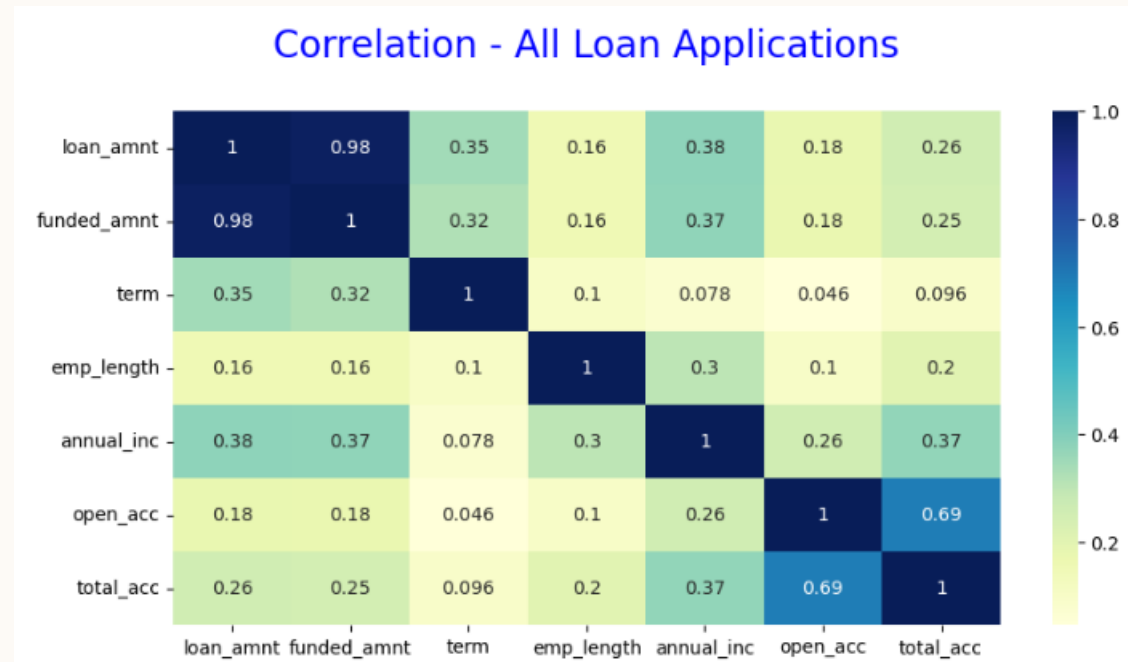
BIVARIATE/MULTIVARIATE ANALYSIS



Home Ownership analysis

- There is no significant difference in home ownership of members who avail loan.

BIVARIATE/MULTIVARIATE ANALYSIS³



Correlation

- Unable to derive correlation across numerical data.