# EDA – CASE STUDY - LOAN

Anu Raje

Vinay Venkatesh

# AGENDA

Introduction

Data Cleaning and Outlier Management

Analysis of Univariate and Bi/Multi variate

Observations

Driver Variables

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

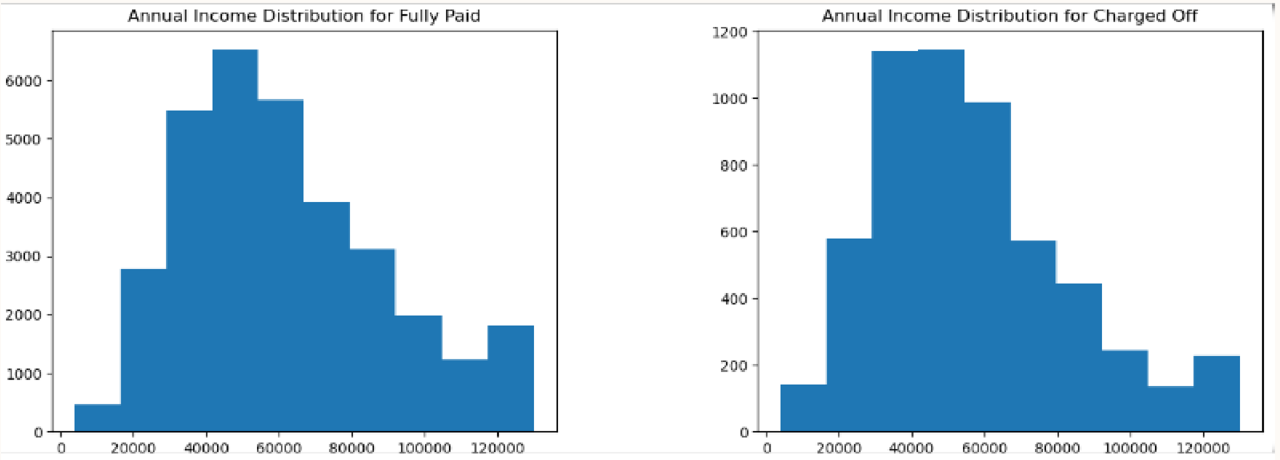
**Data cleaning is important because of the following reasons:**

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 outliers. So outlier have been treated and replaced with percentile based employment length
2. Int_rate - Removed % from the text, replaced blanks with 0 and changed data type to float.
3. Emp_length -
    1. Replace '<1 year' and NA by 0
    2. '10 + Years' = 10
    3. Remaining (Example: 5 Years) - Remove years (Example: 5).
    4. Convert to int64
4. Term - Remove "months" convert to int64-
5. For 70 columns with more than 95% missing data or all rows containing same data are dropped

# UNIVARIATE ANALYSIS

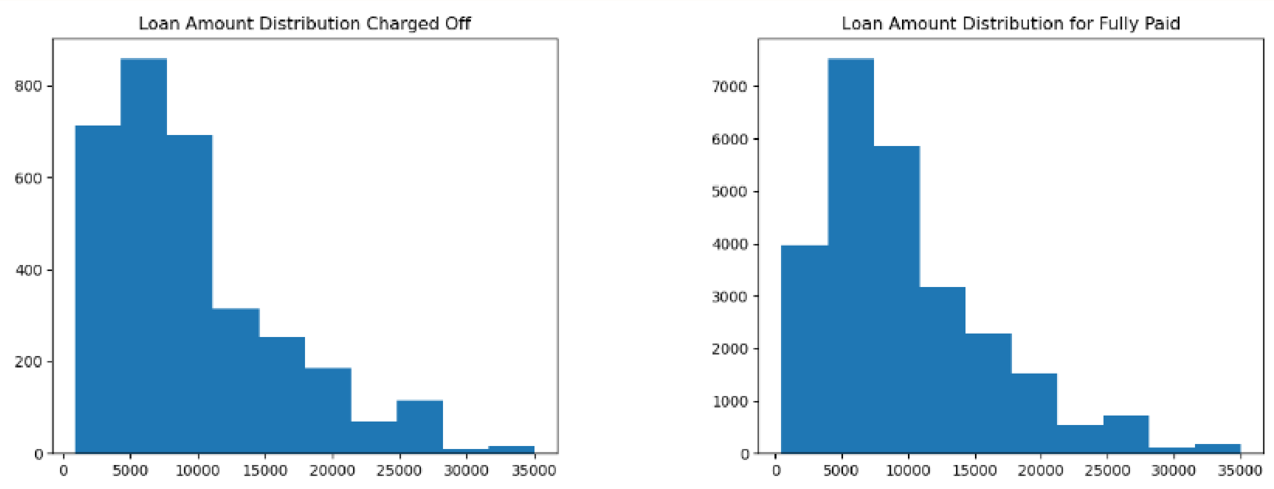


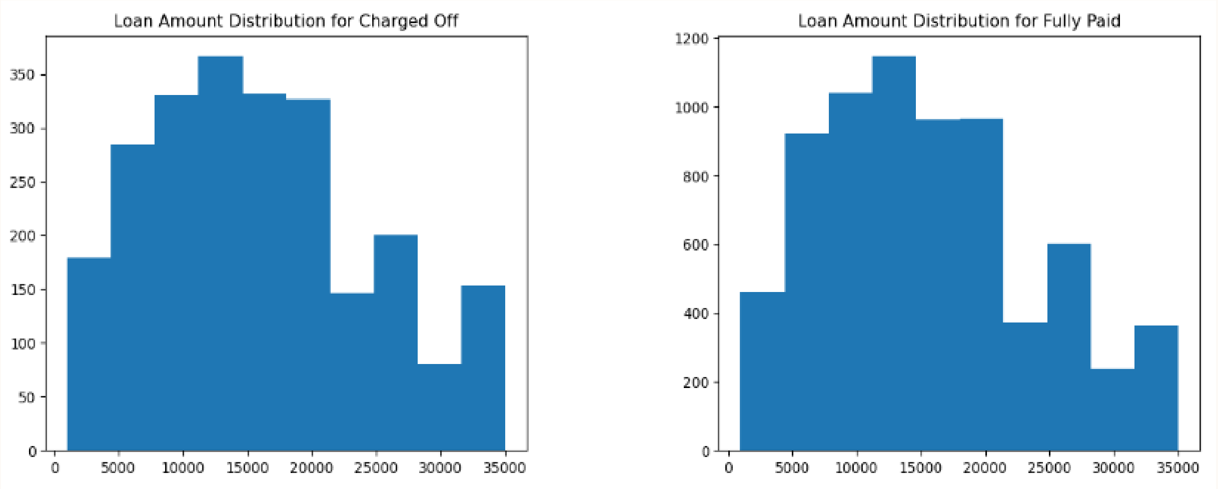
**Annual Income distribution**

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

## Short Term Loan Tenure

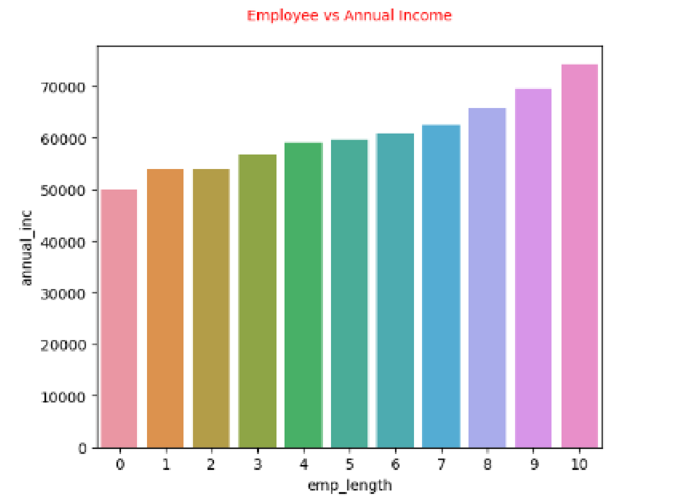
## Long Term Loan Tenure

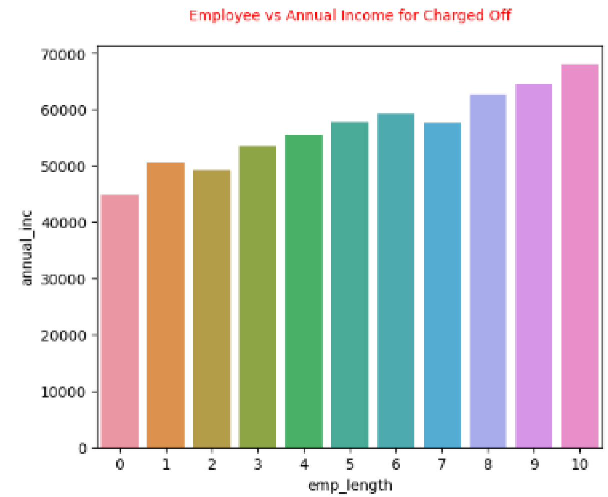




- For Short Term tenure, among charged off members, higher number of charged off is when the loan amount is between **500-8000**
- For Long Term tenure, among charged off members, higher number of charged off is when the loan amount is between **500-5000**

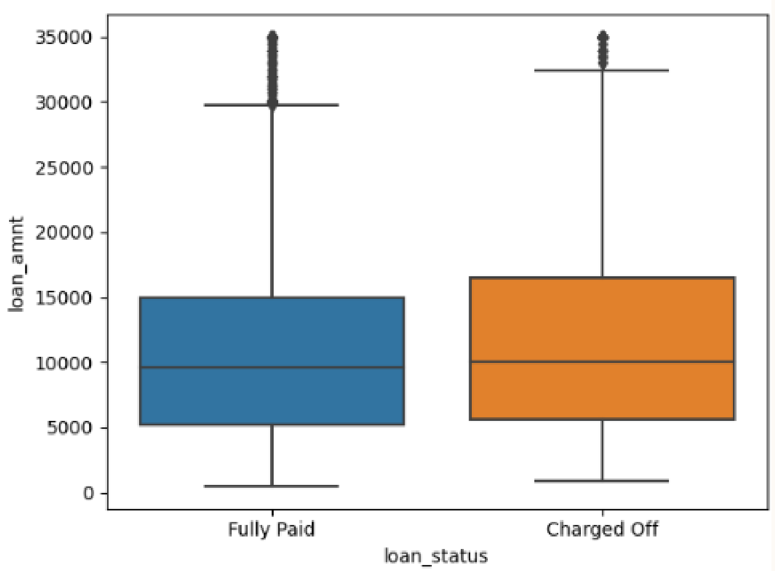
# BI/MULTI VARIATE ANALYSIS





## Employee Length vs Average Annual Income.

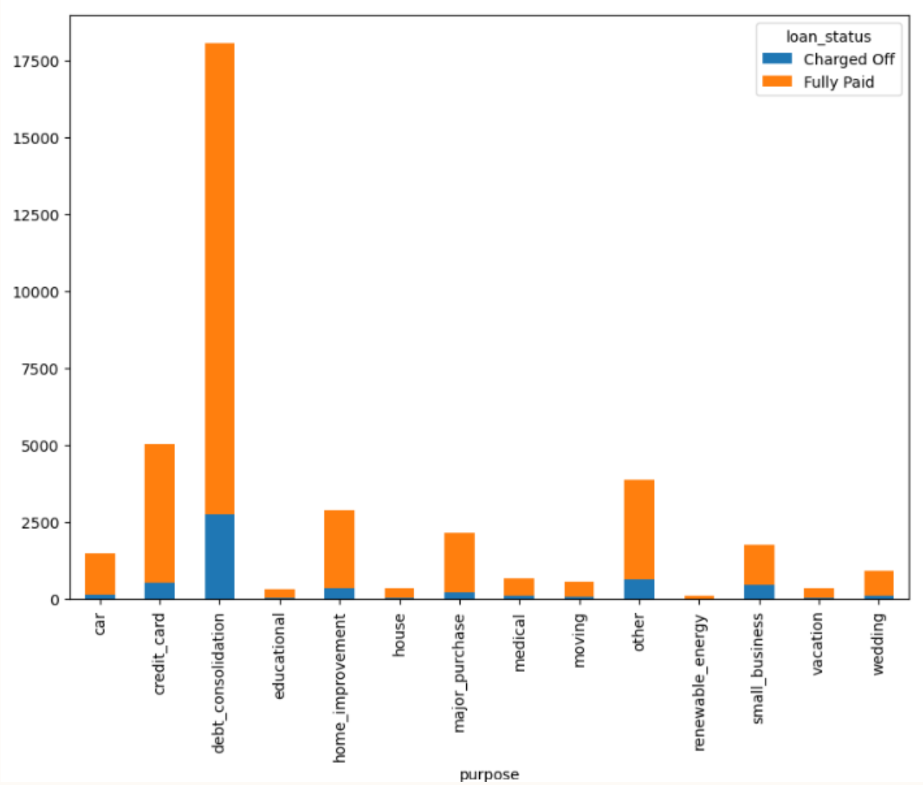
- For all 3 loan status (Current, Charged off and Fully paid), the average annual income is increasing with the Employee Length.
- For Loan Status - Charged Off, the average annual income is lower compared to all other Loan status considered.

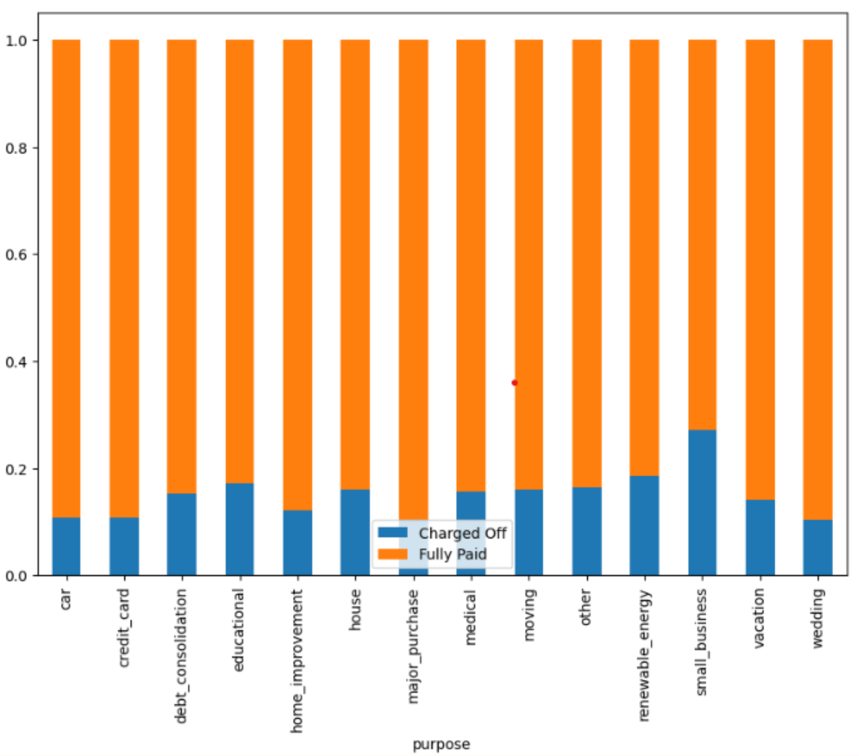


## Loan Amount Box plot for Fully Paid Vs Charged off

- 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

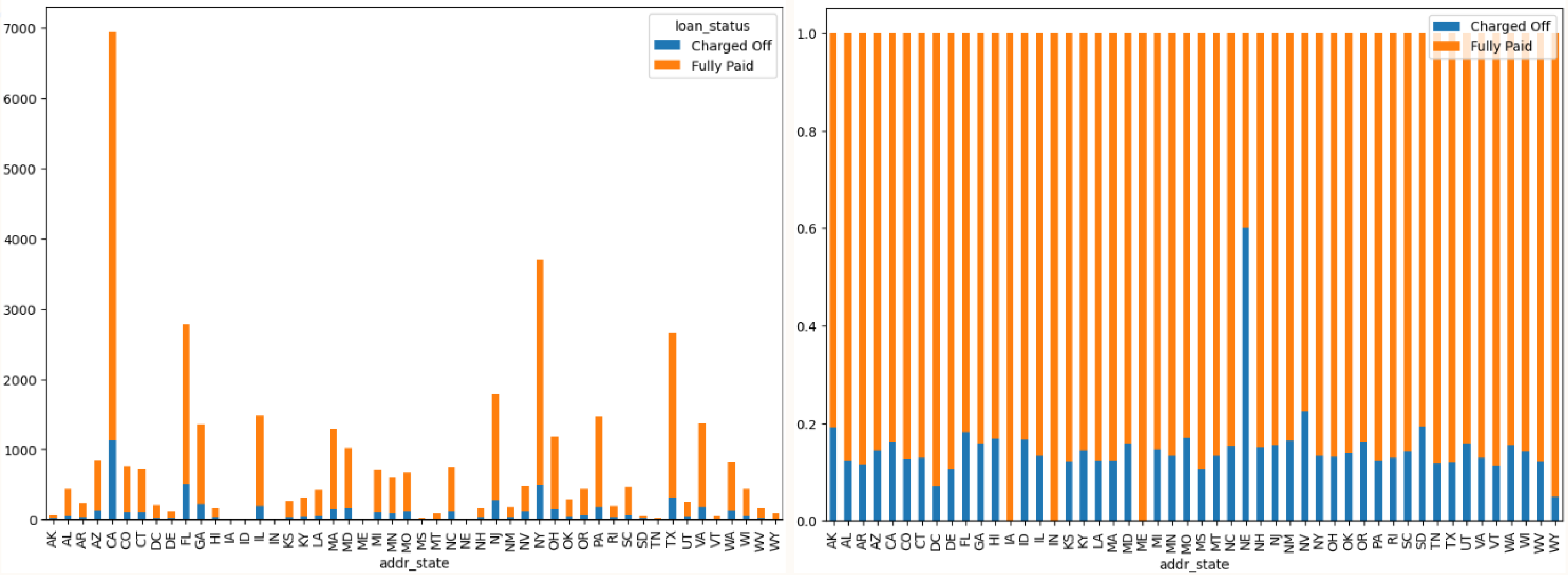




**Count fully paid and charged off members across various purpose for which they avail loans**

- 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

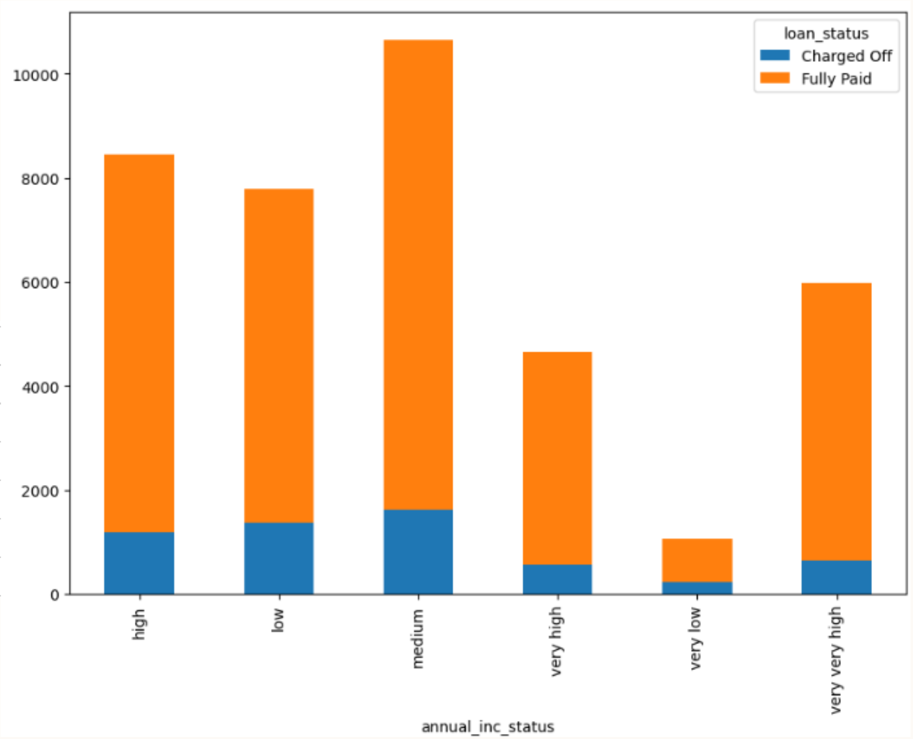


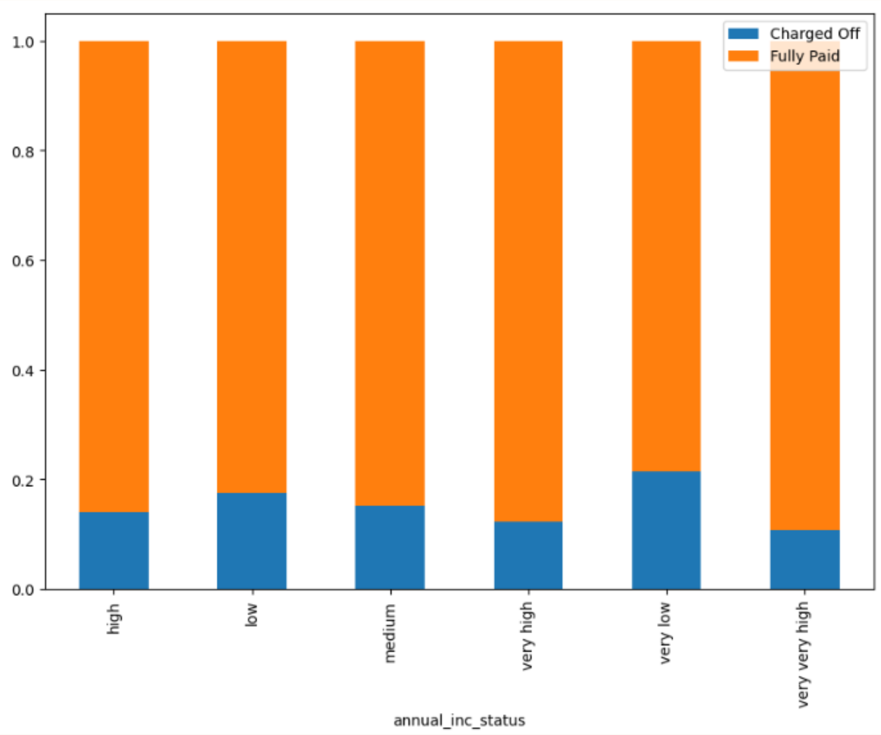
**Region wise split of Fully Paid Vs Charged Off – Count and Percentages**

- 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.

# BI/MULTI VARIATE  ANALYSIS

| Annual Income Range | Classification |
|---|---|
| < 20000 | Very Low |
| Between 20000 - 40000 | Low |
| Between 40000 - 60000 | Medium |
| Between 60000 - 80000 | High |
| Between 80000 - 100000 | Very High |
| > 100000 | Verry Very High |





- Percentage of Charged off decreases with higher salary ranges.
- Typical risk of charged off is there with members with low and very low income range

# OBSERVATIONS

**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. Percentage of Charged off decreases with higher salary ranges. Typical risk of charged off is there with members with low and very low income range
3. For Loan Status - Charged Off, the average annual income is lower compared to all other Loan status considered.
4. 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.
5. There is a clear outlier that members with purpose as "Credit Card" and "Debt consolidation" **are at higher risk of loan defaulting**
6. It can be observed from percentages pf defaulters that members with purpose as "Small Business" **are at higher risk of loan defaulting**
7. Members from state of CA, FL, NY have higher count of defaulters. M**embers availing loan from the mentioned regions are at risk of defaulting.**
8. 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.**

# DRIVER VARIABLES

**Following are driver variables observed from analysis:**

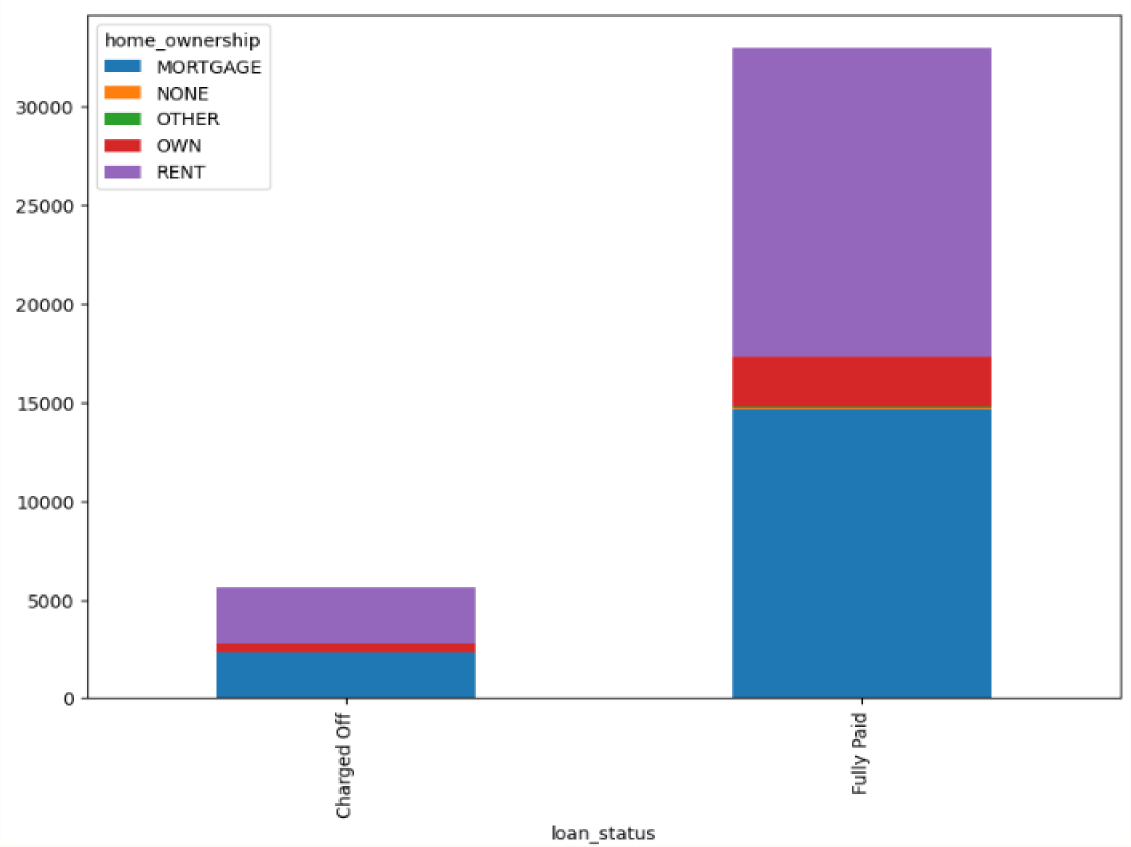
1. Annual Income – Charged off rate is increasing with lower salary ranges
2. Region - This variable is used to determine Charged off rates for members across regions and their respective charged off percentages.
3. Term – Charged off rate for loan amount ranging from 500 to 2000 is higher for 36 months tenure
4. Loan Amount - Charged off rate for loan amount ranging from 500 to 2000 is higher for 36 months tenure
5. Employee Length - Annual Income outliers are treated using emp_length

# THANK YOU

# ANNEXURE

NOTE: GRAPHS WHERE WE COULD NOT DRAW CONCLUSION

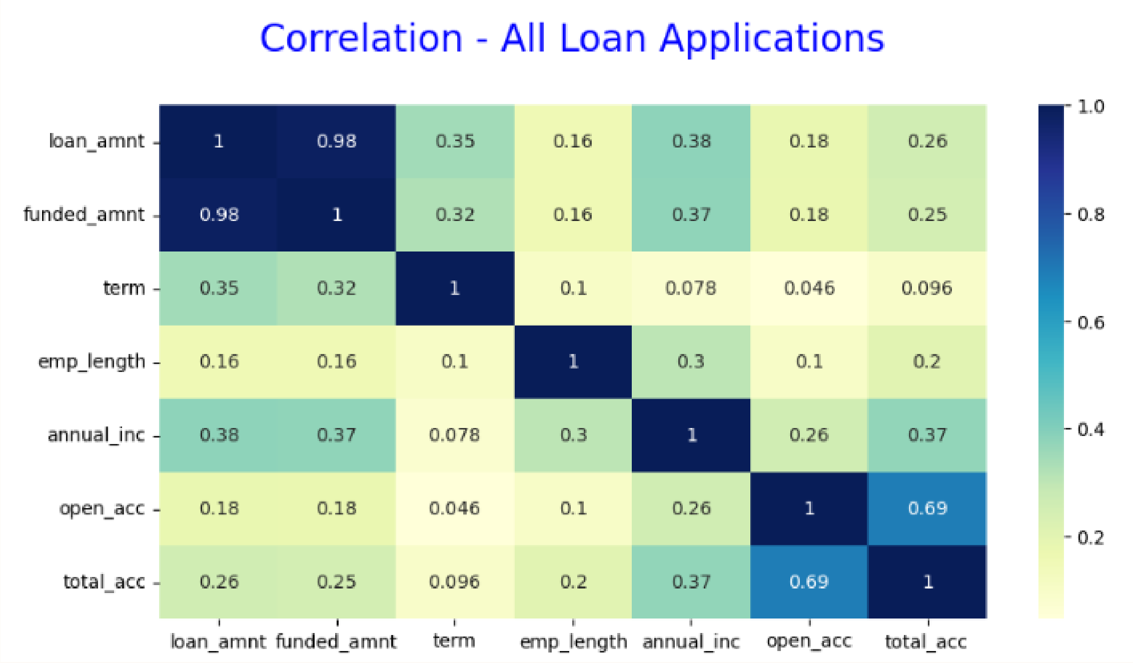
# BIVARIATE/MULTIVARIATE ANALYSIS



**Home Ownership analysis**

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

# BIVARIATE/MULTIVARIATE ANALYSIS



Correlation - All Loan Applications

| | loan_amnt | funded_amnt | term | emp_length | annual_inc | open_acc | total_acc |
|---|---|---|---|---|---|---|---|
| loan_amnt | 1 | 0.98 | 0.35 | 0.16 | 0.38 | 0.18 | 0.26 |
| funded_amnt | 0.98 | 1 | 0.32 | 0.16 | 0.37 | 0.18 | 0.25 |
| term | 0.35 | 0.32 | 1 | 0.1 | 0.078 | 0.046 | 0.096 |
| emp_length | 0.16 | 0.16 | 0.1 | 1 | 0.3 | 0.1 | 0.2 |
| annual_inc | 0.38 | 0.37 | 0.078 | 0.3 | 1 | 0.26 | 0.37 |
| open_acc | 0.18 | 0.18 | 0.046 | 0.1 | 0.26 | 1 | 0.69 |
| total_acc | 0.26 | 0.25 | 0.096 | 0.2 | 0.37 | 0.69 | 1 |

**Correlation**

- Unable to derive correlation across numerical data.