GRAMENER CASE STUDY

SUBMISSION

Group Members:

1. *Anoop K*
2. *Arunima Menon*
3. *Ketan Koul*
4. *Manish Muralidharan*

Objective

To determine risk factors that can have an impact on repayment status of loan customers.

The study will contribute by providing useful current state insights into the factors that cause high default rates in loans. It will assist the management in decision making towards loan sanctioning for applicants which display these behaviours by either denying the loan, reducing the amount of loan , lending at a higher interest rate , etc.

Methodology: Exploratory Data Analysis

1. Source Data : Loan data of all loans issued between 2007 to 2011.
2. Data Cleaning : Performed data cleaning to make the data suitable for analysis.
3. Univariate Analysis : Summarize the data and find pattern.
4. Bivariate Analysis : Explore relationship between two variables ; strength of association or significance of difference.
5. Derived Metrics Analysis : Derive useful metric which aid in analysis.
6. Conclusion : Derive conclusions based on the analysis done.

Overview: Data Set

- The loan data set comprises of unique 39717 records of loan issued between 2007 to 2011.
- The data can be broadly categorized as :
  - Customer Characteristics like emp_title , Emp_length , Annual_inc , Home Ownership , Addr_state ,etc.
  - Loan Characteristics like loan_amount, interest_rate, loan_status, loan_grade, loan_sub_grade,dti,issue_date , etc.
  - Customer Credit Behaviour like delinq_2yrs, open_acc, total_acc, recoveries, out_prncp, etc.
- Of the dataset provided 14% of the loan records are 'Charged Off' in nature.
- Outliers w.r.t annual income have been identified and removed for analysis.
- Records with loan status as 'Current' have been filtered as they don't provide any meaningful insight.
- All columns which have only NAs have been removed from analysis.
- Columns having more than 15% of redundant data has been removed from analysis.
- Derived Metrics added: Annual Income Category , FOIR (Monthly Obligations to Income Ratio) and Leverage (Ratio of Loan Amount to Annual Income) have been added to analysis

Analysis of the fields of the dataset provided

Field Name	To Be Considered for Analysis	Remarks
id	N	As it's the sequence number of the entry in the file , hence can be ignored for analysis.
member_id	N	As it denotes the ID of the member and is unique in nature , hence can be ignored for analysis.
loan_amnt	Y	It is the amount for which the loan was applied.
funded_amnt	N	Doesn't provide an insight as to if the applicant will default or not.
funded_amnt_inv	N	Doesn't provide an insight as to if the applicant will default or not.
term	N	Doesn't have a direct impact on the loan status.
int_rate	N	Is a variable related to Loan status but doesn't have an impact on Loan Status, hence can be ignored from analysis.
installment	Y	Is a variable related to Loan status
grade	N	As the grades are assigned to each loan seeking member by the loan agency and is not directly related to the member , it can be ignored from our analysis point of view.
sub_grade	N	As the grades are assigned to each loan seeking member by the loan agency and is not directly related to the member , it can be ignored from our analysis point of view.
emp_title	N	Is the name of the company in which the applicant is working so can be ignored from analysis.
emp_length	Y	Is a variable related to Loan status
home_ownership	Y	Is a variable related to Loan status
annual_inc	Y	Is a variable related to Loan status
verification_status	Y	As the verification status of the member has no direct impact on his capability to repay the loan. This attribute can be ignored from analysis.
issue_d	N	Date of Issue has no impact on the loan status hence can be ignored for analysis.
loan_status	Y	Is a variable related to Loan status but for our analysis will be used to filter relevant data.
pymnt_plan	N	As this attribute holds only 1 value of N , hence can be ignored for our analysis.
url	N	As this field stores the link to the data listing on the site and doesn't contain any loan related attribute , we can ignore this column from analysis.
desc	N	As this field stores the description of the purpose and we have a separate column with purpose ,we can ignore this column from analysis.
purpose	Y	Is a variable related to Loan status
title	N	Can be ignored
zip_code	N	Can be ignored as it has masked data with XXs.
addr_state	Y	Can be considered to analyse any pattern
dti	Y	Is a variable related to Loan status
delinq_2yrs	Y	Is a variable related to Loan status
earliest_cr_line	Y	Is a variable related to Loan status
inq_last_6mths	Y	Is a variable related to Loan status

Analysis Contd..

Field Name	To Be Considered for Analysis	Remarks
mths_since_last_delinq	N	This is derived once the loan account is delinquent. Hence can be ignored for analysis.
mths_since_last_record	N	Doesn’t have an impact on Loan Status , hence can be ignored from analysis.
open_acc	Y	Is a variable related to Loan status
pub_rec		
revol_bal	Y	Is a variable related to Loan Status
revol_util	Y	Is a variable related to Loan Status
total_acc	Y	Is a variable related to Loan Status
initial_list_status	N	Doesn’t have an impact on Loan Status , hence can be ignored from analysis.
out_prncp	Y	Is a variable related to Loan Status
out_prncp_inv	Y	Is a variable related to Loan Status
total_pymnt	Y	Is a variable related to Loan Status
total_pymnt_inv	Y	Is a variable related to Loan Status
total_rec_prncp	N	Is a variable related to Loan but has no impact on the status of Loan. Hence can be ignored for analysis.
total_rec_int	N	Is a variable related to Loan but has no impact on the status of Loan. Hence can be ignored for analysis.
total_rec_late_fee	N	Is a variable related to Loan but has no impact on the status of Loan. Hence can be ignored for analysis.
recoveries	N	Is a variable related to Loan but has no impact on the status of Loan. Hence can be ignored for analysis.
collection_recovery_fee	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
last_pymnt_d	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
last_pymnt_amnt	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
next_pymnt_d	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
last_credit_pull_d	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
collections_12_mths_ex_med	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
mths_since_last_major_derog	N	contains only NA as values , hence can be ignored from analysis.
policy_code	N	Doesn't have an impact on the loan status , hence can be ignored from analysis.
application_type	N	Contains only INDIVIDUAL as value , hence can be ignored from analysis.
annual_inc_joint	N	contains only NA as values , hence can be ignored from analysis.
dti_joint	N	contains only NA as values , hence can be ignored from analysis.
verification_status_joint	N	contains only NA as values , hence can be ignored from analysis.

Analysis Contd..

Field Name	To Be Considered for Analysis	Remarks
acc_now_delinq	N	contains only NA as values , hence can be ignored from analysis.
tot_coll_amt	N	contains only NA as values , hence can be ignored from analysis.
tot_cur_bal	N	contains only NA as values , hence can be ignored from analysis.
open_acc_6m	N	contains only NA as values , hence can be ignored from analysis.
open_il_6m	N	contains only NA as values , hence can be ignored from analysis.
open_il_12m	N	contains only NA as values , hence can be ignored from analysis.
open_il_24m	N	contains only NA as values , hence can be ignored from analysis.
mths_since_rcnt_il	N	contains only NA as values , hence can be ignored from analysis.
total_bal_il	N	contains only NA as values , hence can be ignored from analysis.
il_util	N	contains only NA as values , hence can be ignored from analysis.
open_rv_12m	N	contains only NA as values , hence can be ignored from analysis.
open_rv_24m	N	contains only NA as values , hence can be ignored from analysis.
max_bal_bc	N	contains only NA as values , hence can be ignored from analysis.
all_util	N	contains only NA as values , hence can be ignored from analysis.
total_rev_hi_lim	N	contains only NA as values , hence can be ignored from analysis.
inq_fi	N	contains only NA as values , hence can be ignored from analysis.
total_cu_tl	N	contains only NA as values , hence can be ignored from analysis.
inq_last_12m	N	contains only NA as values , hence can be ignored from analysis.
acc_open_past_24mths	N	contains only NA as values , hence can be ignored from analysis.
avg_cur_bal	N	contains only NA as values , hence can be ignored from analysis.
bc_open_to_buy	N	contains only NA as values , hence can be ignored from analysis.
bc_util	N	contains only NA as values , hence can be ignored from analysis.
chargeoff_within_12_mths	N	contains only NA as values , hence can be ignored from analysis.
delinq_amnt	N	contains only NA as values , hence can be ignored from analysis.
mo_sin_old_il_acct	N	contains only NA as values , hence can be ignored from analysis.
mo_sin_old_rev_tl_op	N	contains only NA as values , hence can be ignored from analysis.
mo_sin_rcnt_rev_tl_op	N	contains only NA as values , hence can be ignored from analysis.
mo_sin_rcnt_tl	N	contains only NA as values , hence can be ignored from analysis.
mort_acc	N	contains only NA as values , hence can be ignored from analysis.
mths_since_recent_bc	N	contains only NA as values , hence can be ignored from analysis.
mths_since_recent_bc_dlq	N	contains only NA as values , hence can be ignored from analysis.
mths_since_recent_inq	N	contains only NA as values , hence can be ignored from analysis.
mths_since_recent_revol_delinq	N	contains only NA as values , hence can be ignored from analysis.

Analysis Contd..

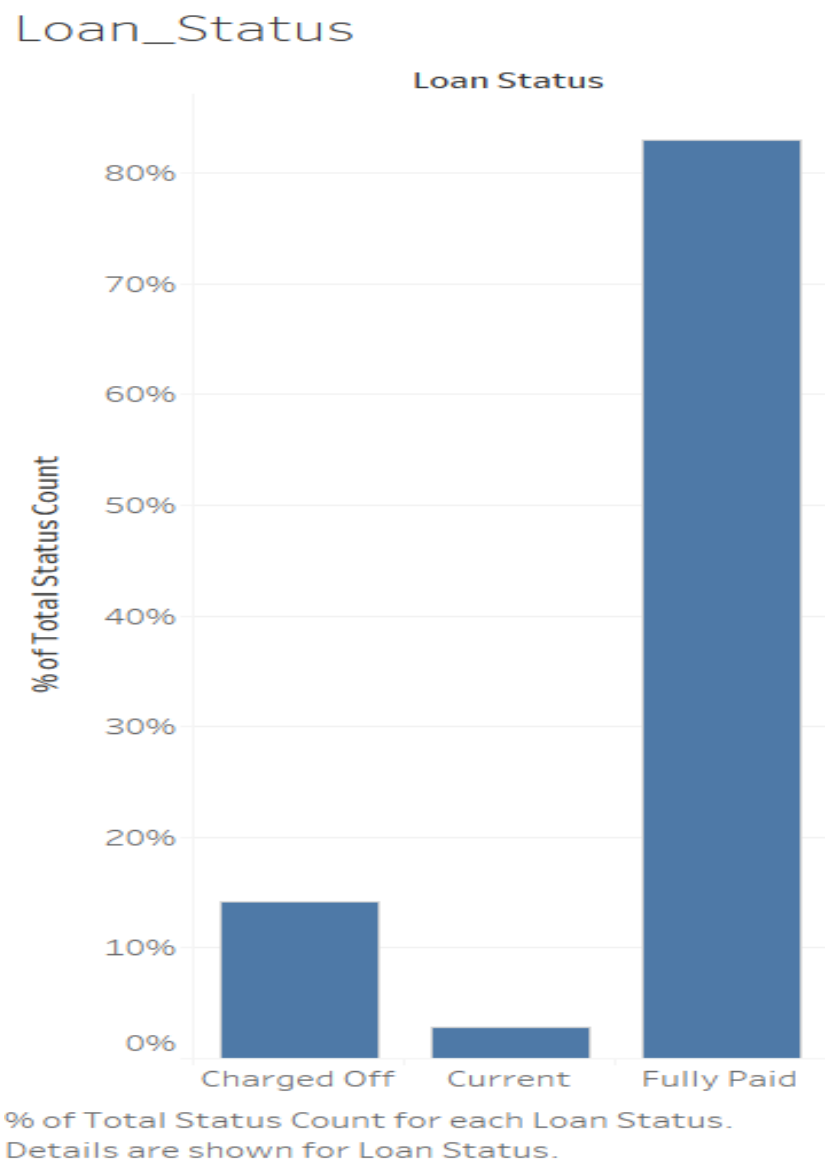
Field Name	To Be Considered for Analysis	Remarks
num_accts_ever_120_pd	N	contains only NA as values , hence can be ignored from analysis.
num_actv_bc_tl	N	contains only NA as values , hence can be ignored from analysis.
num_actv_rev_tl	N	contains only NA as values , hence can be ignored from analysis.
num_bc_sats	N	contains only NA as values , hence can be ignored from analysis.
num_bc_tl	N	contains only NA as values , hence can be ignored from analysis.
num_il_tl	N	contains only NA as values , hence can be ignored from analysis.
num_op_rev_tl	N	contains only NA as values , hence can be ignored from analysis.
num_rev_accts	N	contains only NA as values , hence can be ignored from analysis.
num_rev_tl_bal_gt_0	N	contains only NA as values , hence can be ignored from analysis.
num_sats	N	contains only NA as values , hence can be ignored from analysis.
num_tl_120dpd_2m	N	contains only NA as values , hence can be ignored from analysis.
num_tl_30dpd	N	contains only NA as values , hence can be ignored from analysis.
num_tl_90g_dpd_24m	N	contains only NA as values , hence can be ignored from analysis.
num_tl_op_past_12m	N	contains only NA as values , hence can be ignored from analysis.
pct_tl_nvr_dlq	N	contains only NA as values , hence can be ignored from analysis.
percent_bc_gt_75	N	contains only NA as values , hence can be ignored from analysis.
pub_rec_bankruptcies	N	contains only NA as values , hence can be ignored from analysis.
tax_liens	N	contains only NA as values , hence can be ignored from analysis.
tot_hi_cred_lim	N	contains only NA as values , hence can be ignored from analysis.
total_bal_ex_mort	N	contains only NA as values , hence can be ignored from analysis.
total_bc_limit	N	contains only NA as values , hence can be ignored from analysis.
total_il_high_credit_limit	N	contains only NA as values , hence can be ignored from analysis.

The fields marked as Y for ‘To be considered for Analysis’ have been analyzed to find patterns and relationships.

Univariate Analysis

1. Loan Status

Analysis on Loan Status to find the default rate in the dataset shows 14.6% of applicants to be charged off.

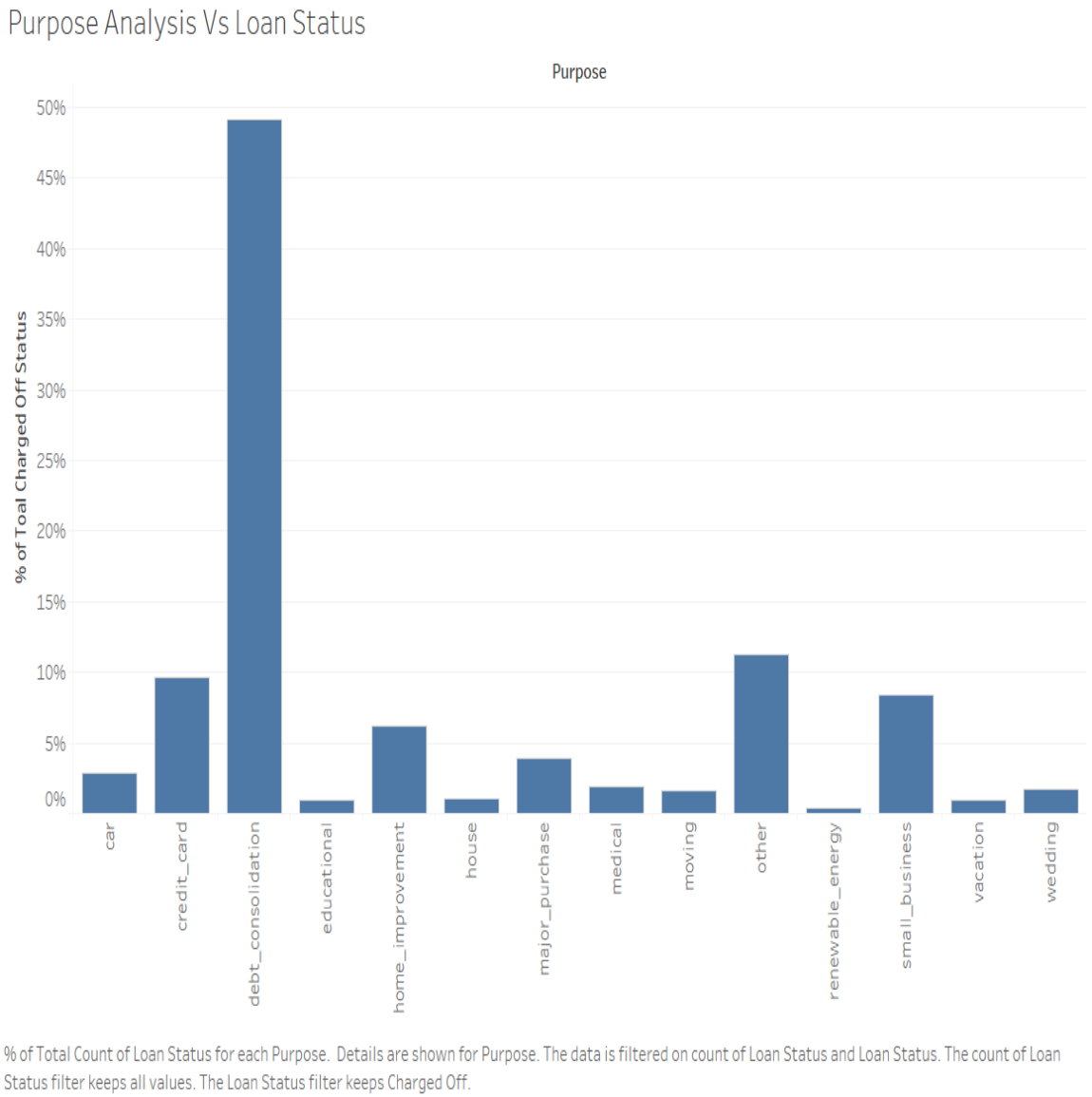


Univariate Analysis Contd..

2. Loan Purpose

A segmented univariate Analysis of the field “purpose” for only “charged off” loan status shows 46.8% loan was applied for debt_consolidation.

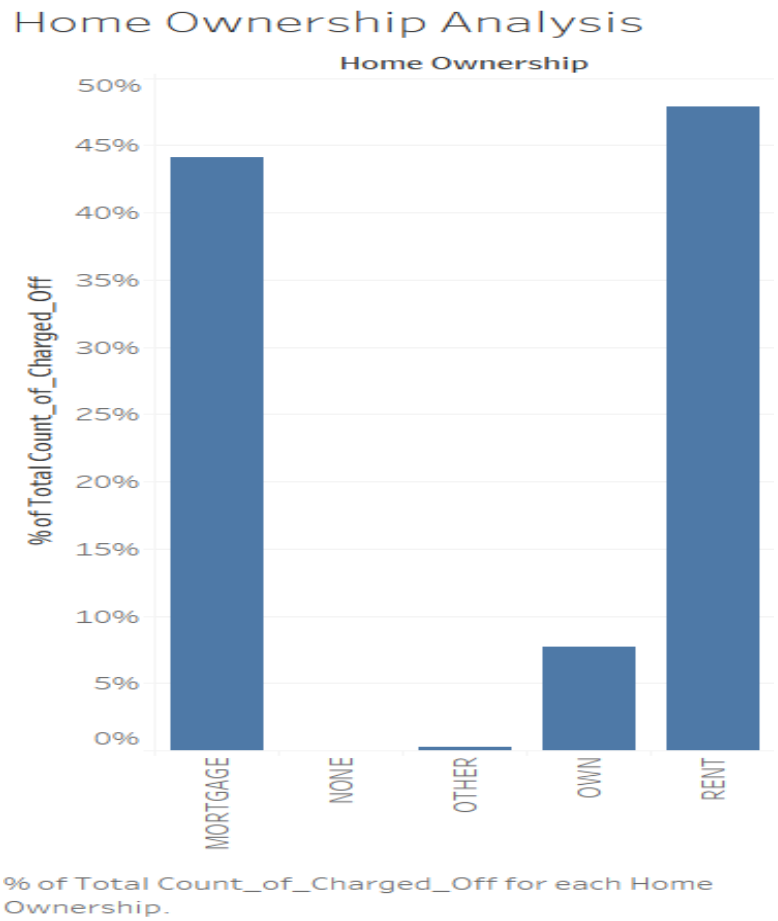
The top 5 purposes resulting in defaults are:
1. Debt Consolidation
2. Others
3. Credit Card
4. Small Business
5. Home Improvement.



Univariate Analysis Contd..

3. Home Ownership
A segmented univariate Analysis of the field “home ownership” for only “charged off” loan status reveals applicants with rented or mortagaged homes show a trend of defaulting.

Wherein rented ownership shows 47% default rate closely followed by mortgaged ownership at 44%.

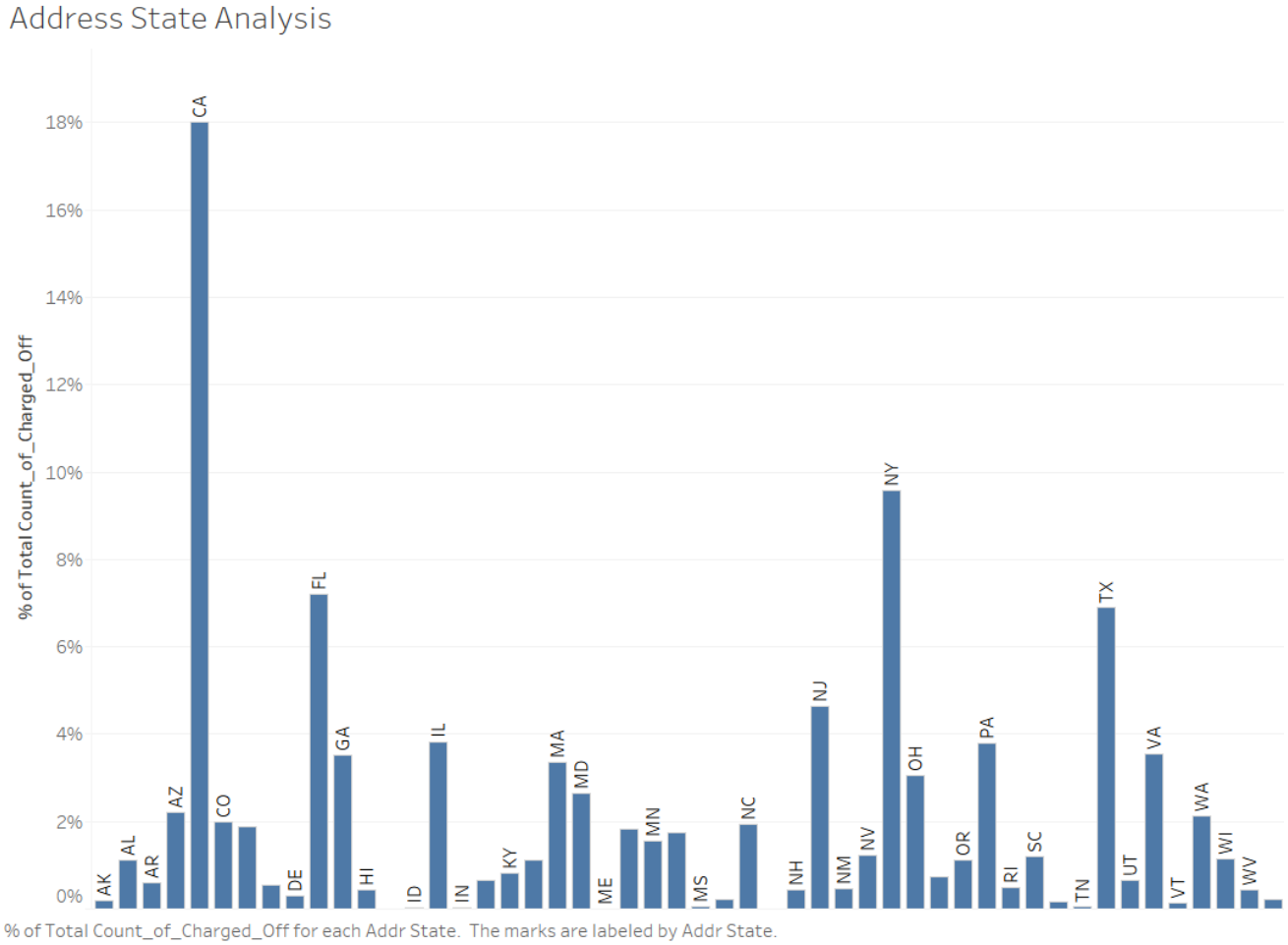


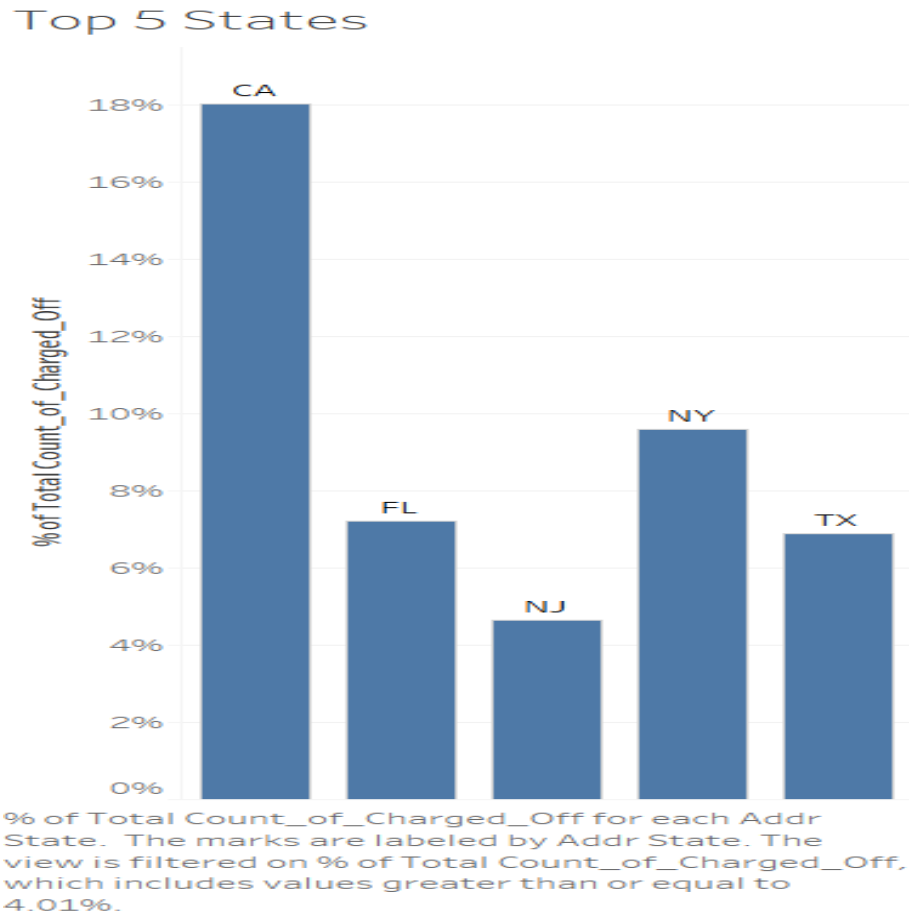
Univariate Analysis Contd..

4. State of residence

This plot tries to study if the state of residence has any pattern of default.

We see that people hailing from California have shown maximum cases of Charge Offs followed by people of



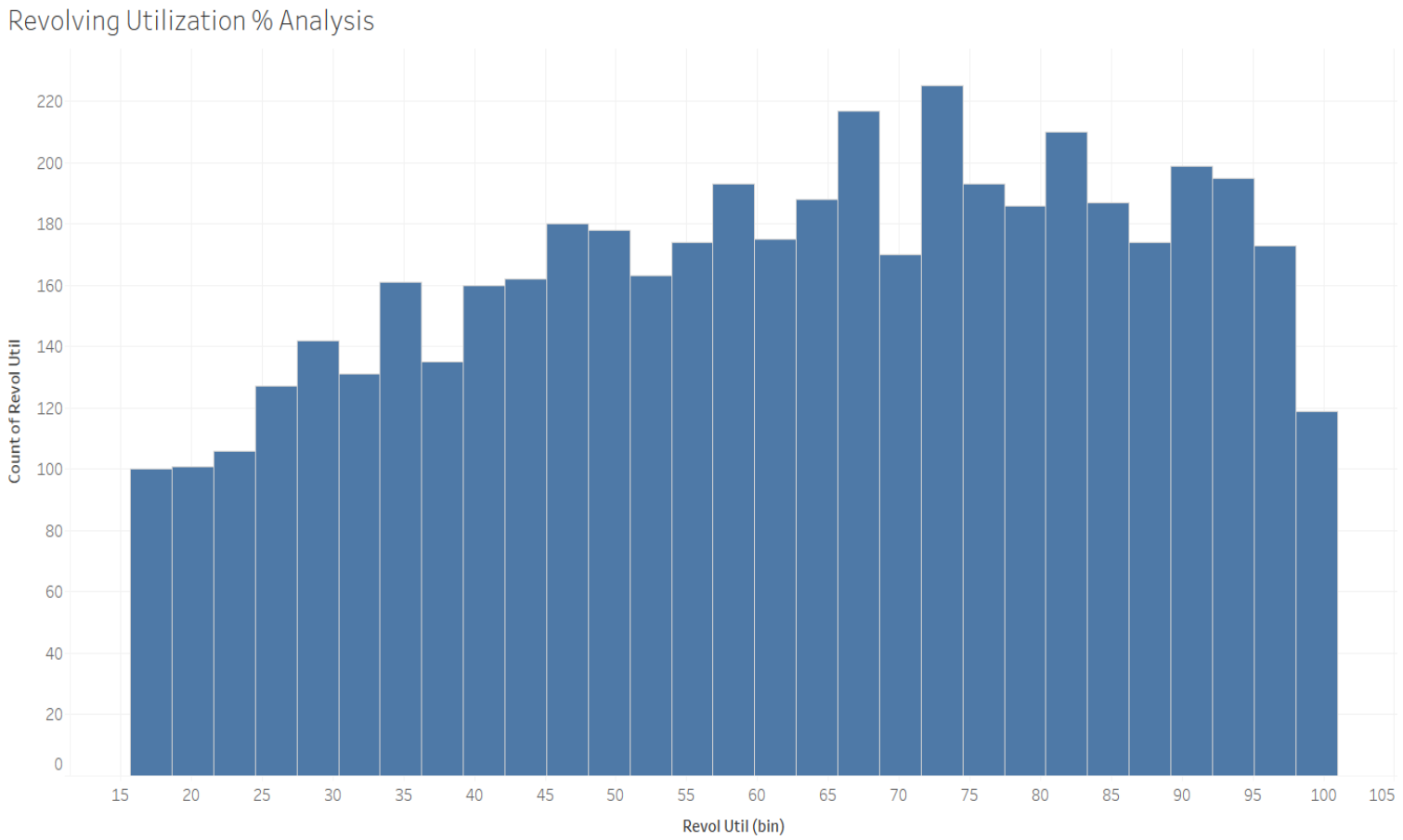


Univariate Analysis Contd..

5. Revolving Utilization Percentage

This plot tries to study if revolving utilization percentage has a pattern during charge offs.

We see that higher the revolving utilization percentage higher the charge offs.

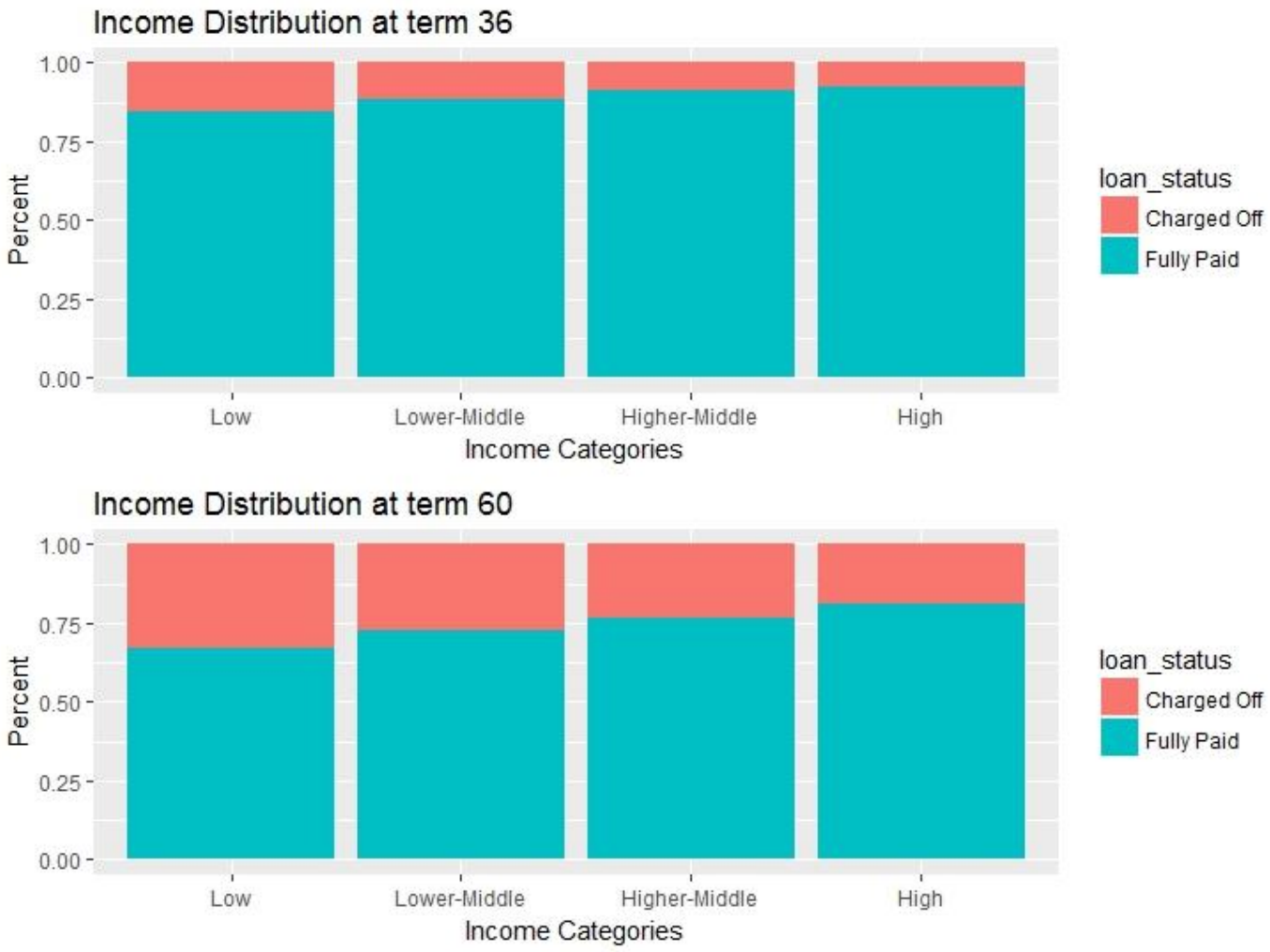


The trend of count of Revol Util for Revol Util (bin). The data is filtered on Loan Status, which keeps Charged Off. The view is filtered on Revol Util (bin), which ranges from 15 to 98.

Bivariate Analysis Contd..

1. Income Category with Term
The plot shows that the Income is inversely proportional to the risk of default, i.e., the higher the income bracket, the lower the chances of default.

Also, for the same income brackets, loans of longer terms tend to be defaulted more than the loans with shorter terms.

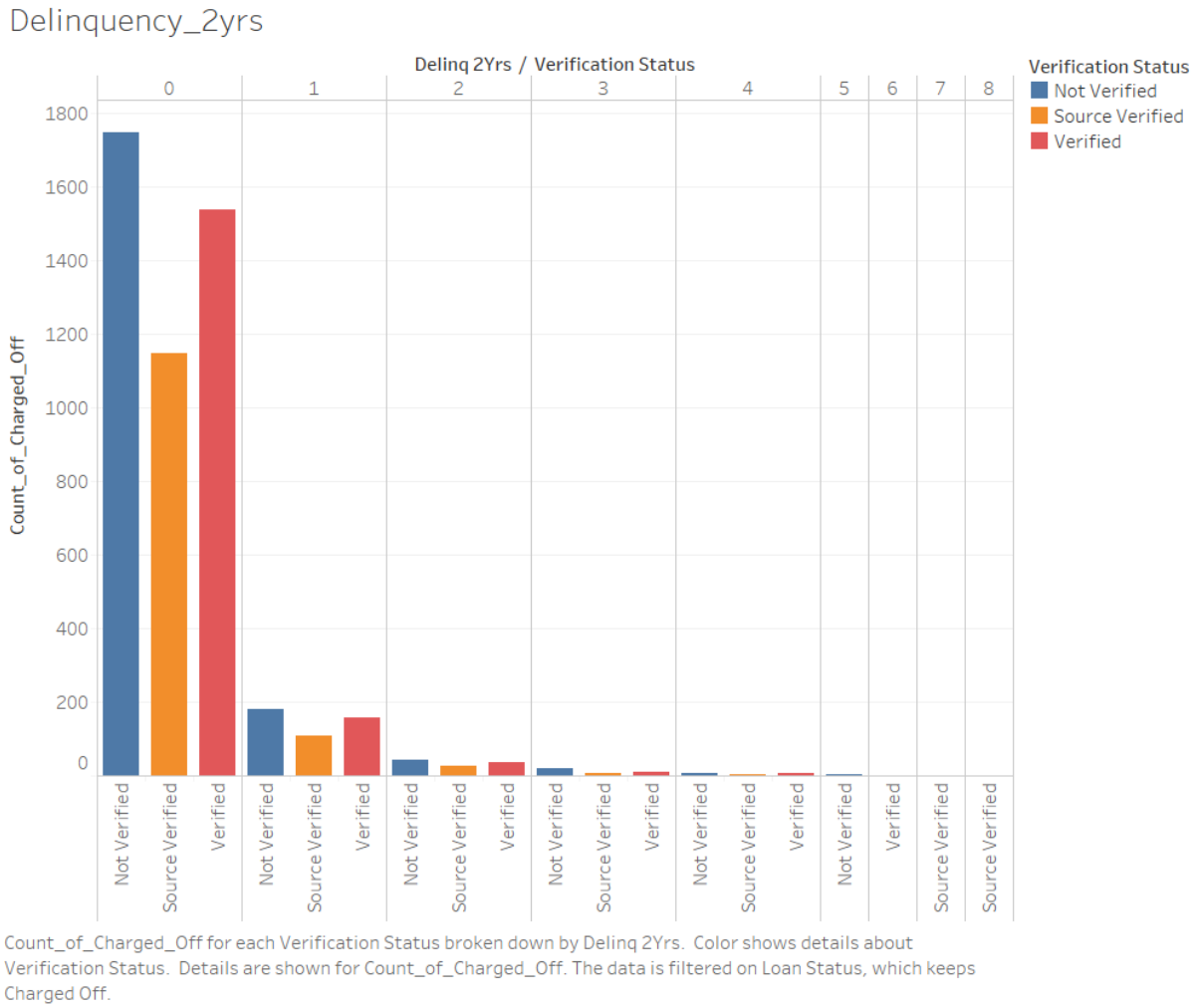


Bivariate Analysis Contd..

2. Delinquency_2yrs with Verification Status

It is seen that the loans where Income Source has not been verified tend to be riskier across all Delinquency Bins.

Historically, loans where income source is unverified tend to default on loans more compared to loans where the source is verified by some means.

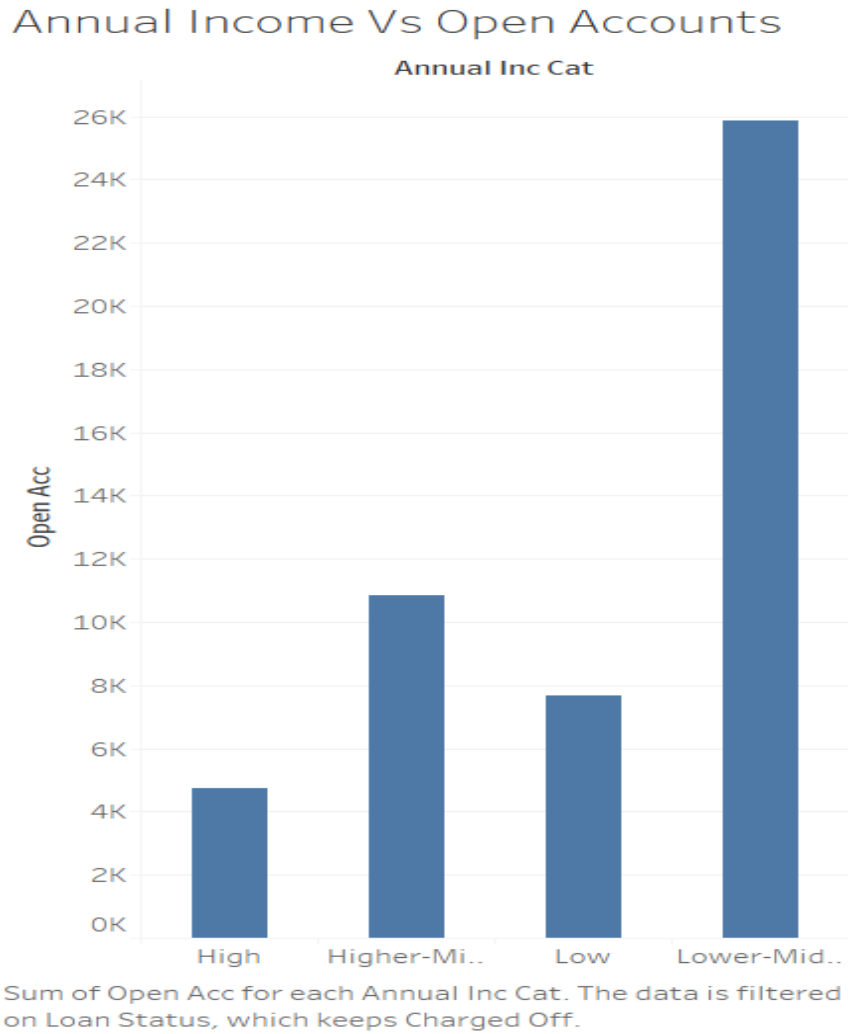


Derived Metrics

1. Annual Income Category.

Annual income has been broadly categorized as Low (Income upto 35K), Lower Medium(35K-70K), Higher Medium (70K - 100K), High (>100K).

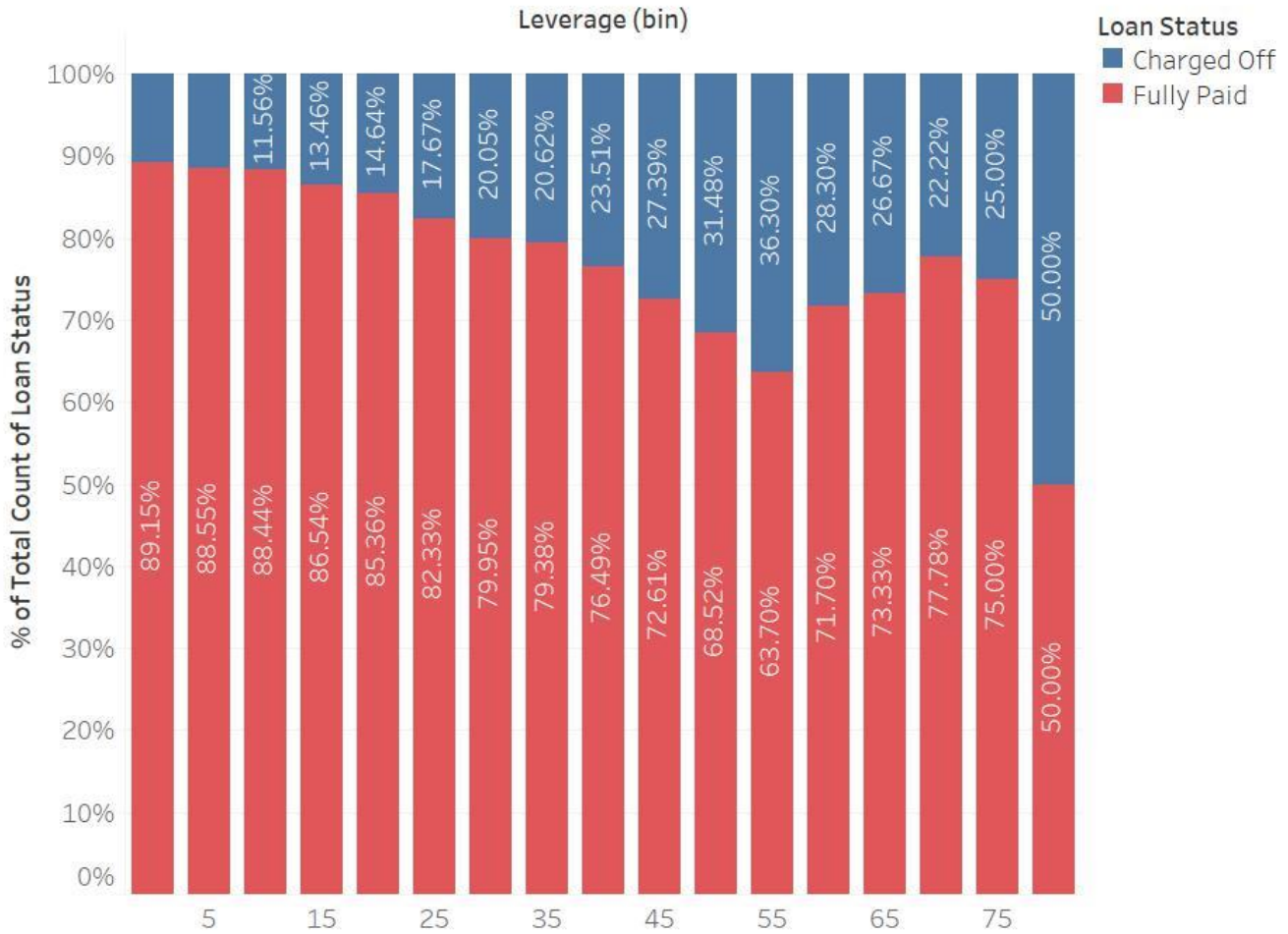
On plotting the same against “Open_acc” which indicates number of open credit lines for the applicant. It shows applicants falling under lower-middle income range have the maximum open accounts and show tendency of default.



Derived Metrics Contd..

3. Analysis of leverage on defaults.

The plot shows a general trend of higher Leverage leading to more defaults.



% of Total Count of Loan Status for each Leverage (bin). Color shows details about Loan Status. The view is filtered on Loan Status, which keeps Charged Off and Fully Paid.

Conclusions

The following conclusions were derived from the Data Analysis of the dataset:

- Income is inversely proportional to the default risk. Higher the income, lower the chance of default
- The term of the loan is directly proportional to the risk of default. The longer the term, the higher the chances of default on loan payments.
- Borrowers living in rented houses or with mortgages on houses tend to be riskier.
- Loans given without income source verification are riskier. Hence the company should strictly ensure verification of income source.
- Wherever the purpose of the loan is debt consolidation, the loan is the riskiest. Hence such loans should be charged higher interest rate to mitigate the risk.
- Loans which are rated lower sub grade riskier hence to be charged higher.

Conclusions :: Driving factors

Based on the analysis, the following are identified as the major drivers for determining the risk of a customer:

1. Income Bracket
2. Purpose of Loan
3. Type of Residence
4. Verification of Source of Income
5. Tenure of the Proposed Loan
6. Leverage on Income
7. Grading of the Customer