Gramener Case Study

SUBMISSION

Candidate Name:

1. Venkata Nihar Tatavarti
2. Prasanth Rasam
3. Deepak Malhotra

Business Objective

The aim of analysis is to identify the root cause of the problem

- Driving factors (or driver variables) behind loan default
- Inferential analysis on loans which are “Charged off” and “Fully paid”
- Predictive analysis on “Current Loans”

Problem solving methodology

- Data Understanding
- Data clean-up and correction
- Identify Risky loan applicants {Inferential}
    - Univariate
    - Bi-variate
    - Muti-variate
- How to reduce Credit Loss {Predictive}
    - Apply model on Current
- Data Analysis and Inferences
- Presentation

Problem solving methodology

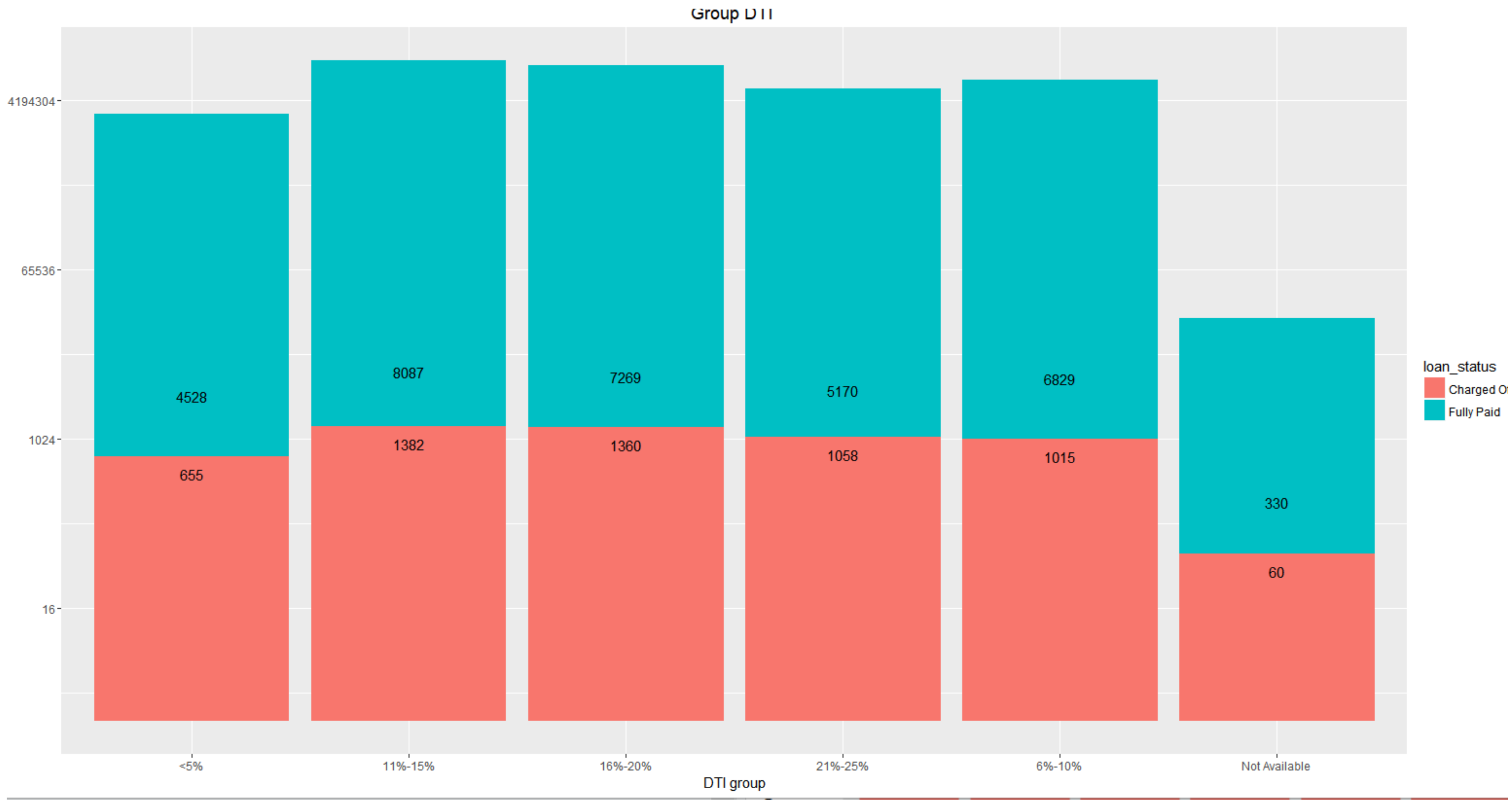
- Data clean-up and correction
  - 54 columns have NA and hence removed
  - 11 columns have values but are not useful for analysis
  - Two key fields identified
    - Debt-to-income ratio (dti) - Lower the better
      - Loans are approved where dti is less than 30%
    - Revolving balance also known as debt-to-limit ratio (revol_util), lower the better
      - Revolving balance (debt-to-limit ratio) is not considered for Loan approval
    - Grouping – Categorization applied on
      - Revolving balance
      - Interest Rate
      - Annual Income
      - Total Accounts (Active credit lines)
    - Data anomalies observed between Title & Purpose. Many loans which are Home Improvement, Buying House, Medical and Student loans are mapped to different Purpose. Manual correction is not possible hence excluded these two columns from Analysis
- Data Analysis
  - Inferential Statistics to be applied on loan status Fully Paid & Charged Off - What went wrong
  - Predictive Analytics to be applied on loan status Charged Off - How to avoid

Inferential Analysis

Data Analysis

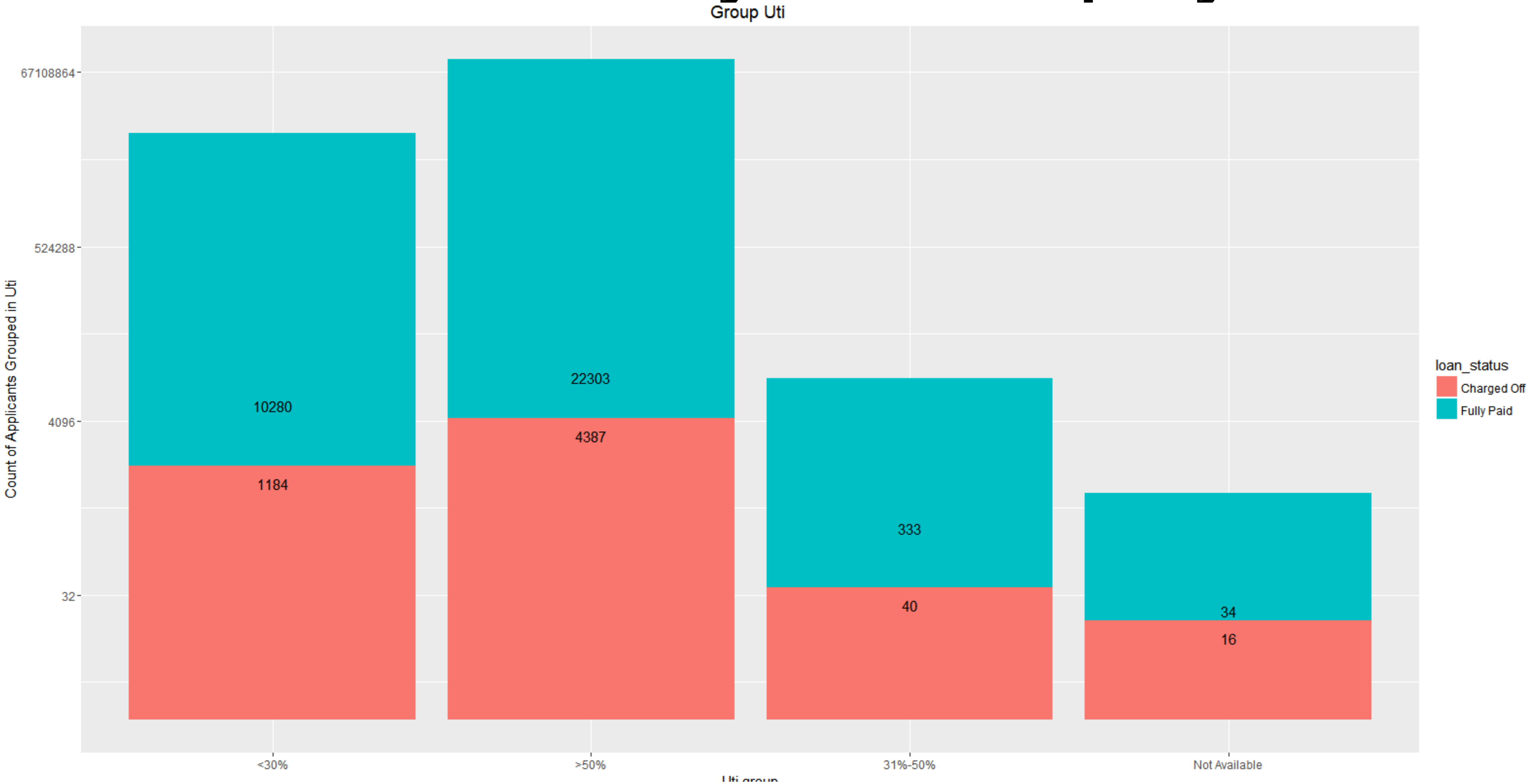
Analysis Type	Field(s)	Inferences
Univariate	Group_util	>50% - 78% of Charged Loans are in this category
Univariate	Group_DTI	Across all categories we same mostly same % of Bad Loans
Univariate	Group_int_rate	11% - 20% Interest Rate 80% but inferences can't be made
Univariate	Group_annual_inc	Income groups less than 100K stands as outlier
Univariate	Group_creditlines	Between 10-50 (total_acc - Lower the credit accounts higher the bad loans. (Customers with 7-25 Accounts have more badloans))
Univariate	emp_length	Unable to able to arrive at inference due to low standard deviation
Univariate	verification_status	Where source is verified is 4% of badloans else 11%
Univariate	addr_state	Only for Demographics, CA is highest 3%
Univariate	delinq_2yrs	Bad loans are found with people less than 3 defaults (0-2)
Bi-Variate	delinq_2yrs & Group_DTI	When DTI Grouping and deling_2years are combined, customers with 0 defaults in last 2 years and in category between 6%-25% have most number of charged loans
Bi-Variate	Group_DTI & home_ownership	Where Home Ownership is RENT or MORTGAGE, DTI Grouping 11%-20% is found as high contributors for charged loans
Bi-Variate	delinq_2yrs & verification_status	In both Not Verified and Verified, DTI Grouping 11%-20% is found as high contributors for charged loans
Bi-Variate	delinq_2yrs & Income	Where Annual Income is less than 100K and DTI is between 11%-20%, high contributors for Charged loans
Multi-Variate	verification_status, Group_util, Group_creditlines, Group_DTI	Charged Loans Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 16-20% -> 395 Not Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 16-20% -> 375 Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 21-25% -> 371 Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 11-15% -> 354 Not Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 11-15% -> 340 Calculated the Probability of Default for Each of the combinations - verification_status, Group_util, Group_creditlines, Group_DTI

Univariate Analysis- Group_DTI



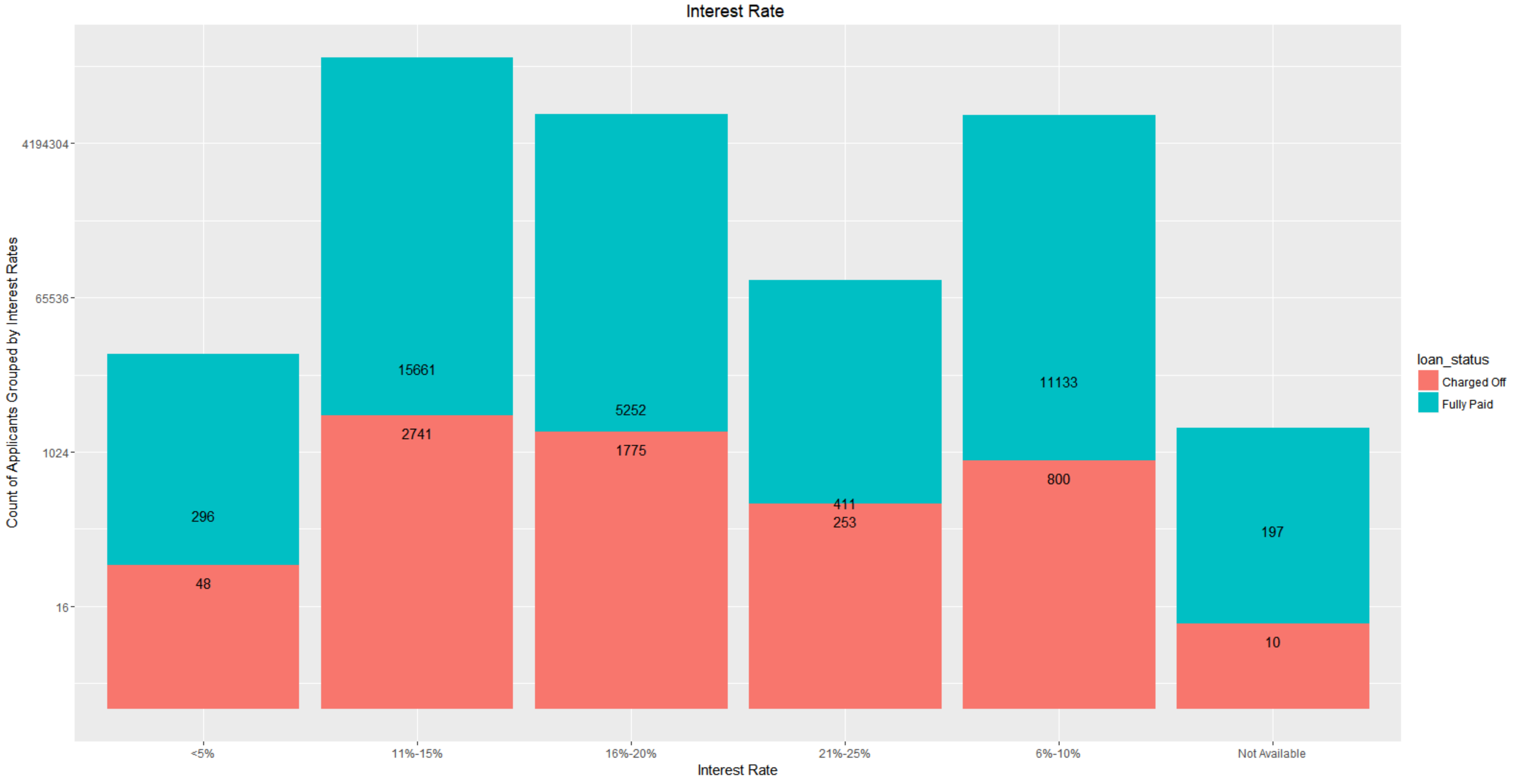
- Debt-to-income ratio (dti) - Lower the better
- Looks like banks approve loans where dti is <30% as maximum value is 29.9%
- Charged Loans are higher in 11% to 20% group

Univariate Analysis- Group by UTI



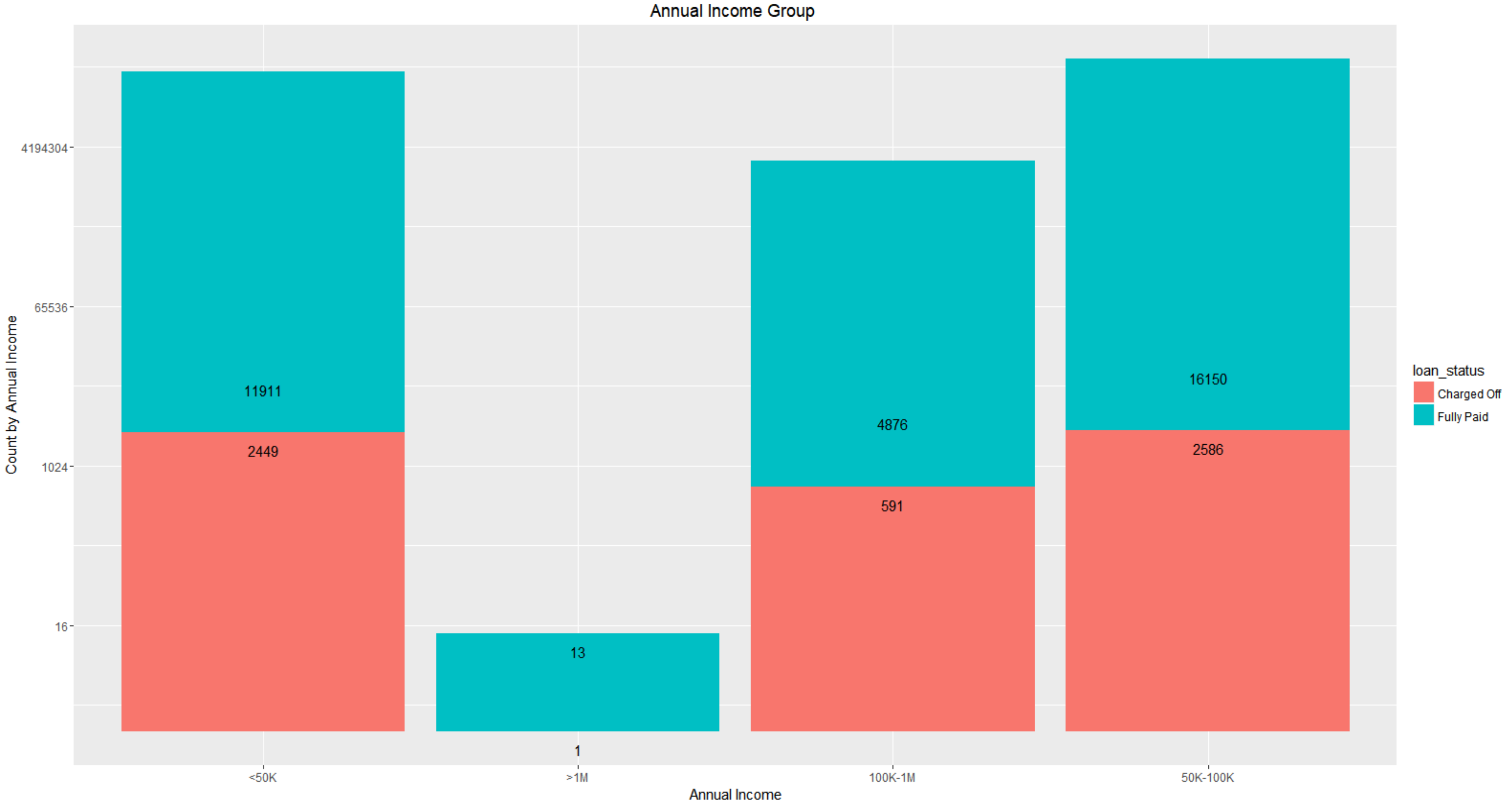
- 75 % of charged of loans are in the category ,where, UTI greater than 50%
- UTIL -**revolving balance** is the portion of **credit** card spending that goes unpaid at the end of a billing cycle. The **debt**-to-**limit ratio** is the **ratio** of a consumer's total credit card balances versus total credit card limits, expressed as a percentage

Univariate Analysis- Group by interest rate



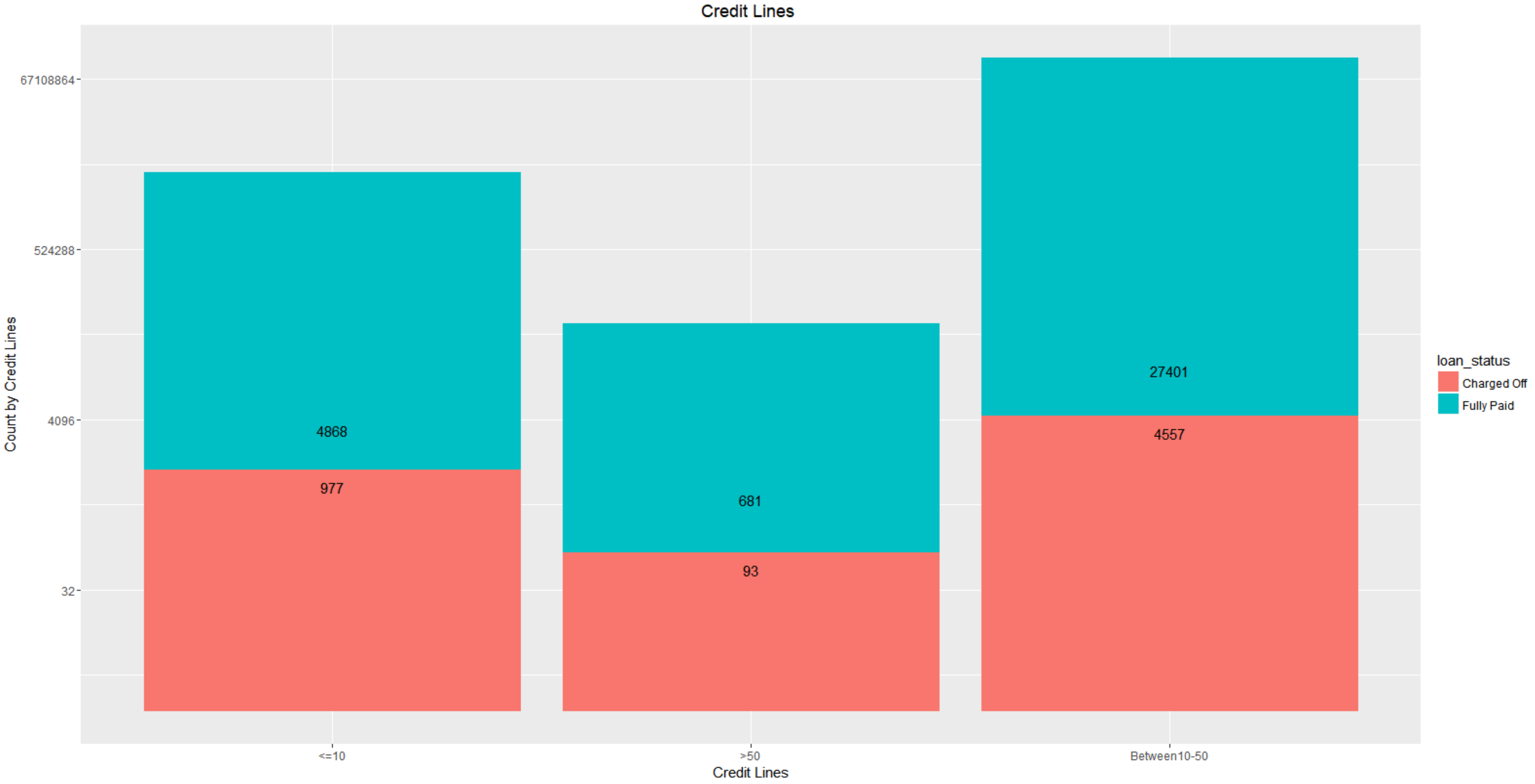
- 11% - 20% Interest Rate have 80% of Charged Off Loans but inferences can't be made

Univariate Analysis- Group by Annual income



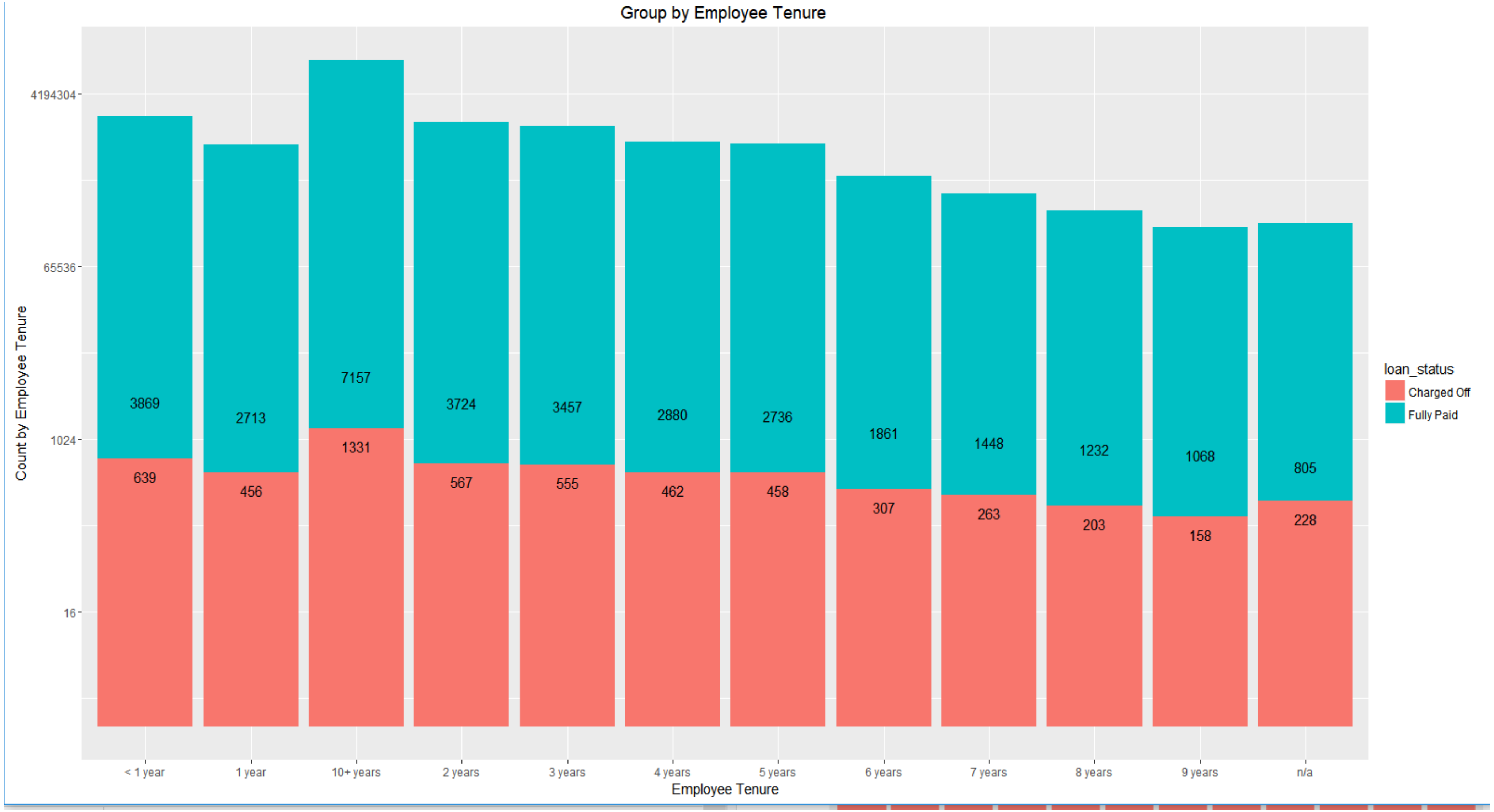
- Income groups less than 100K stands as outlier

Univariate Analysis- credit lines



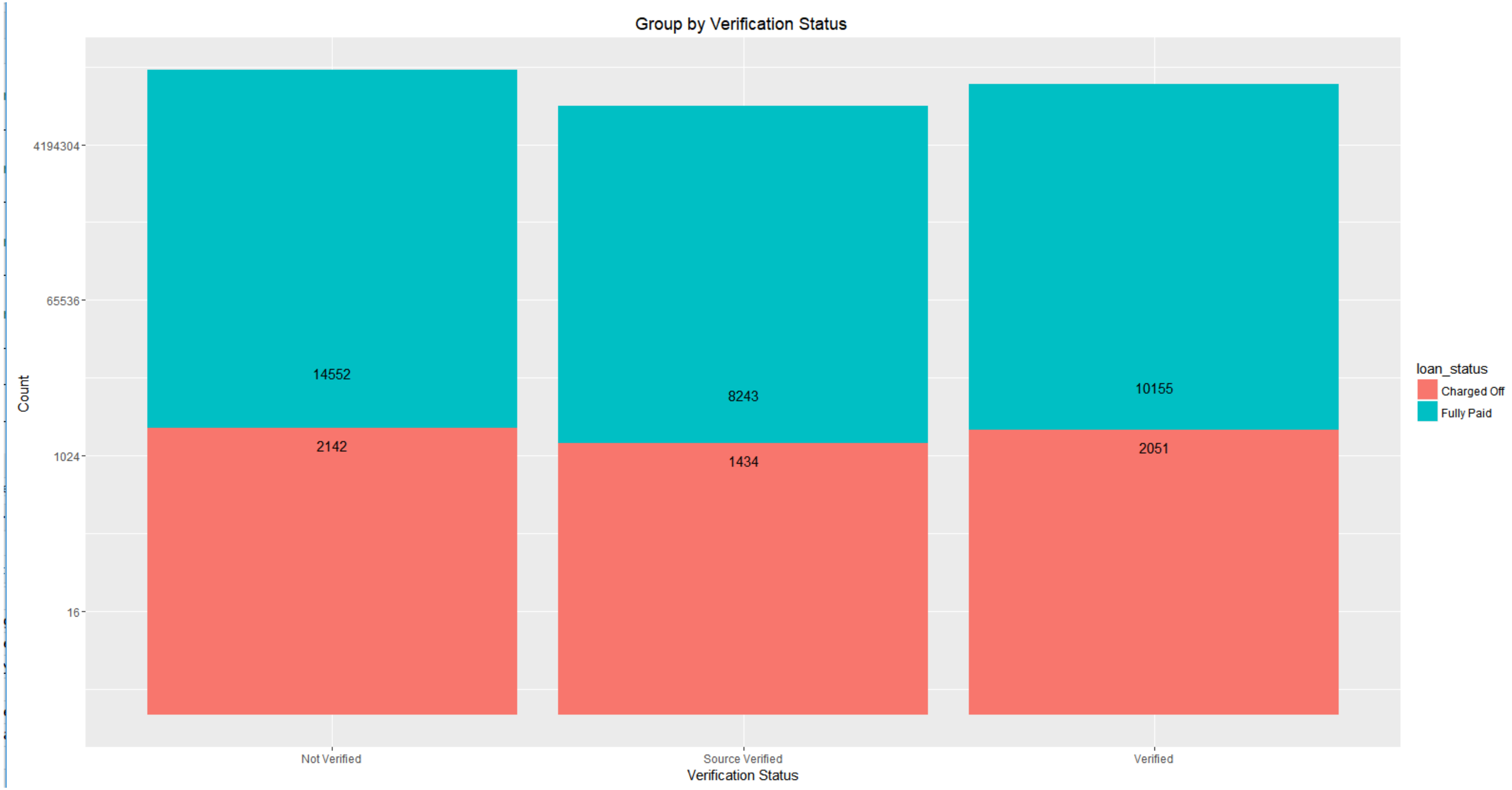
- Between 10-50 (total_acc - Lower the credit accounts higher the bad loans. (Customers with 7-25 Accounts have more badloans))

Univariate Analysis- Employee Length



Not much deviation in the data-No inferences can be drawn

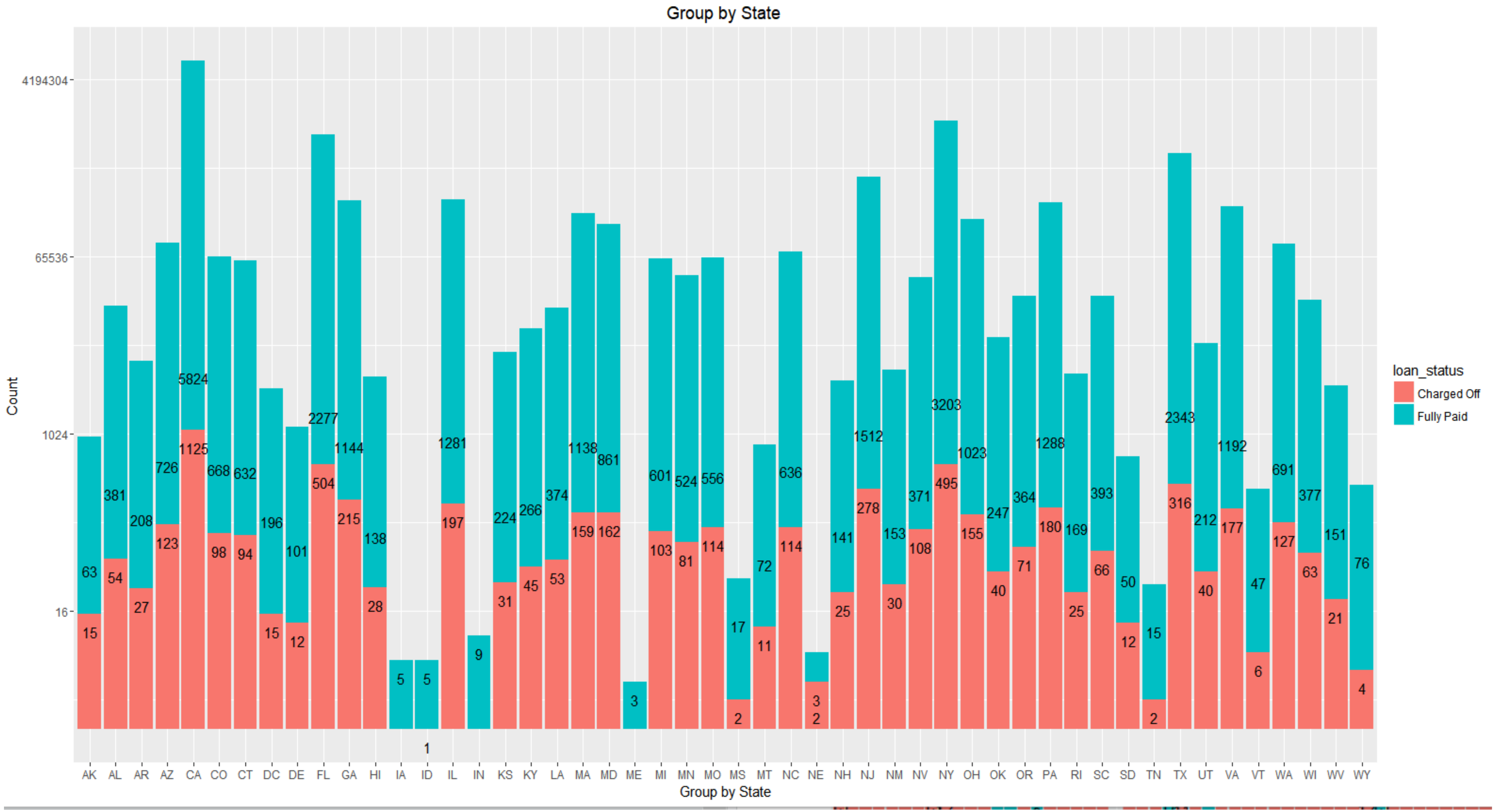
Univariate Analysis- Verification Status



11% of the charged off loans are either Not Verified or Verified.

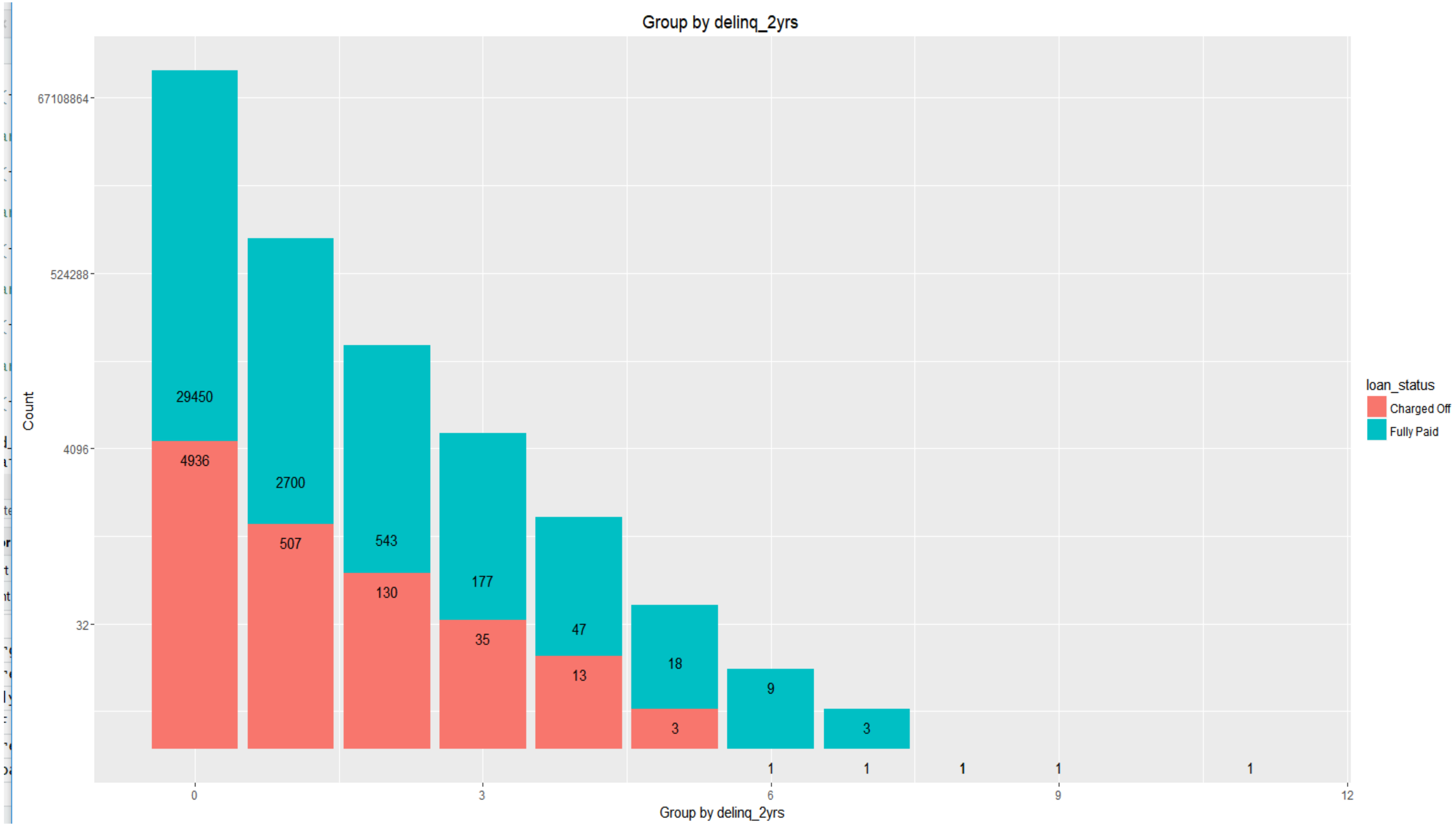
Source Verified status has 4% of loan indicating that 'Source Verification' is needed to avoid Charged off loans

Univariate Analysis- Applicant State



Demographic distribution of loan , California has the most number of loans indicating more Credit based behaviour

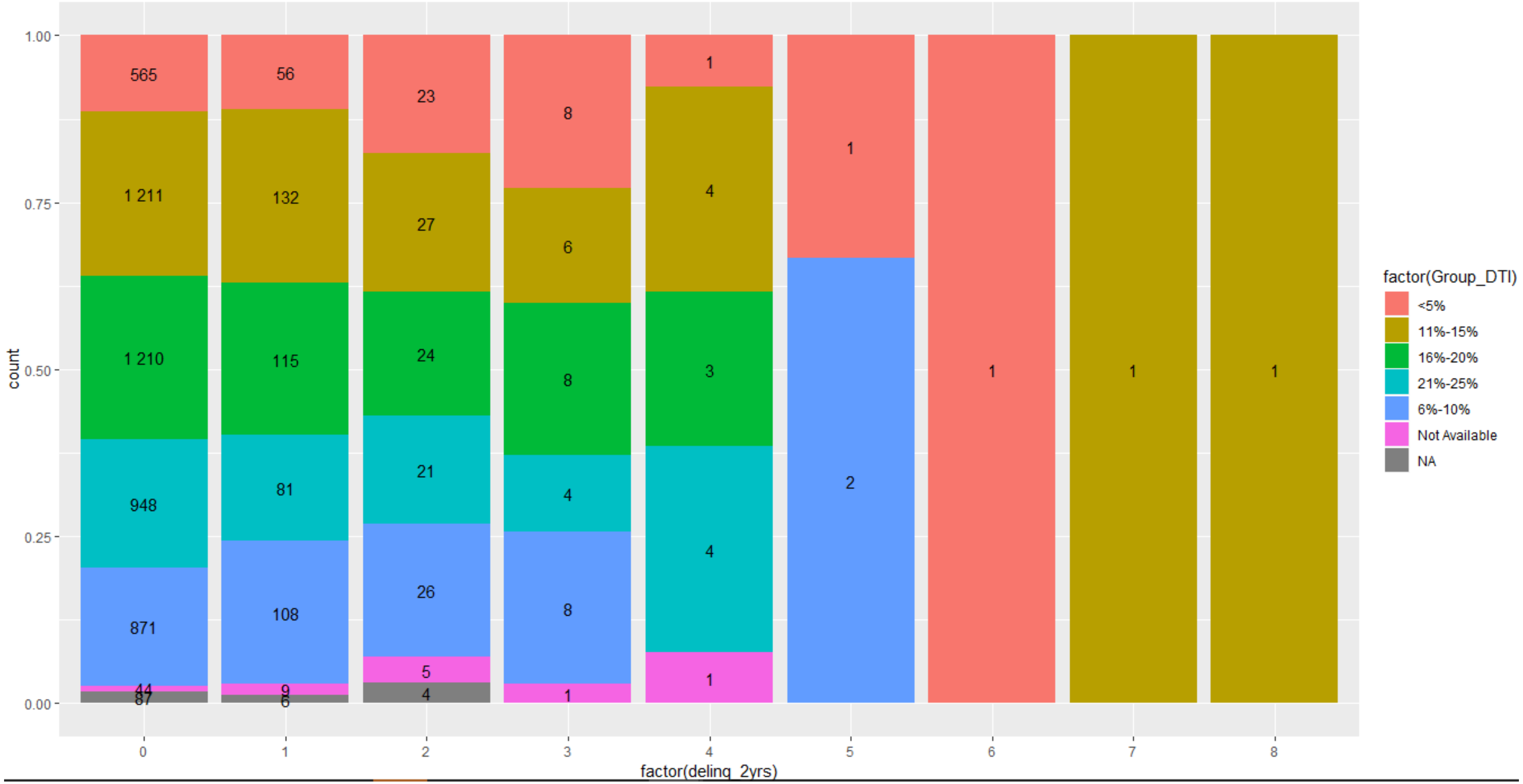
Univariate Analysis- Deling 2 years



Charged Off loans are found with Customers less than 3 defaults (0-2)

Bivariate Analysis- Charged off Loans

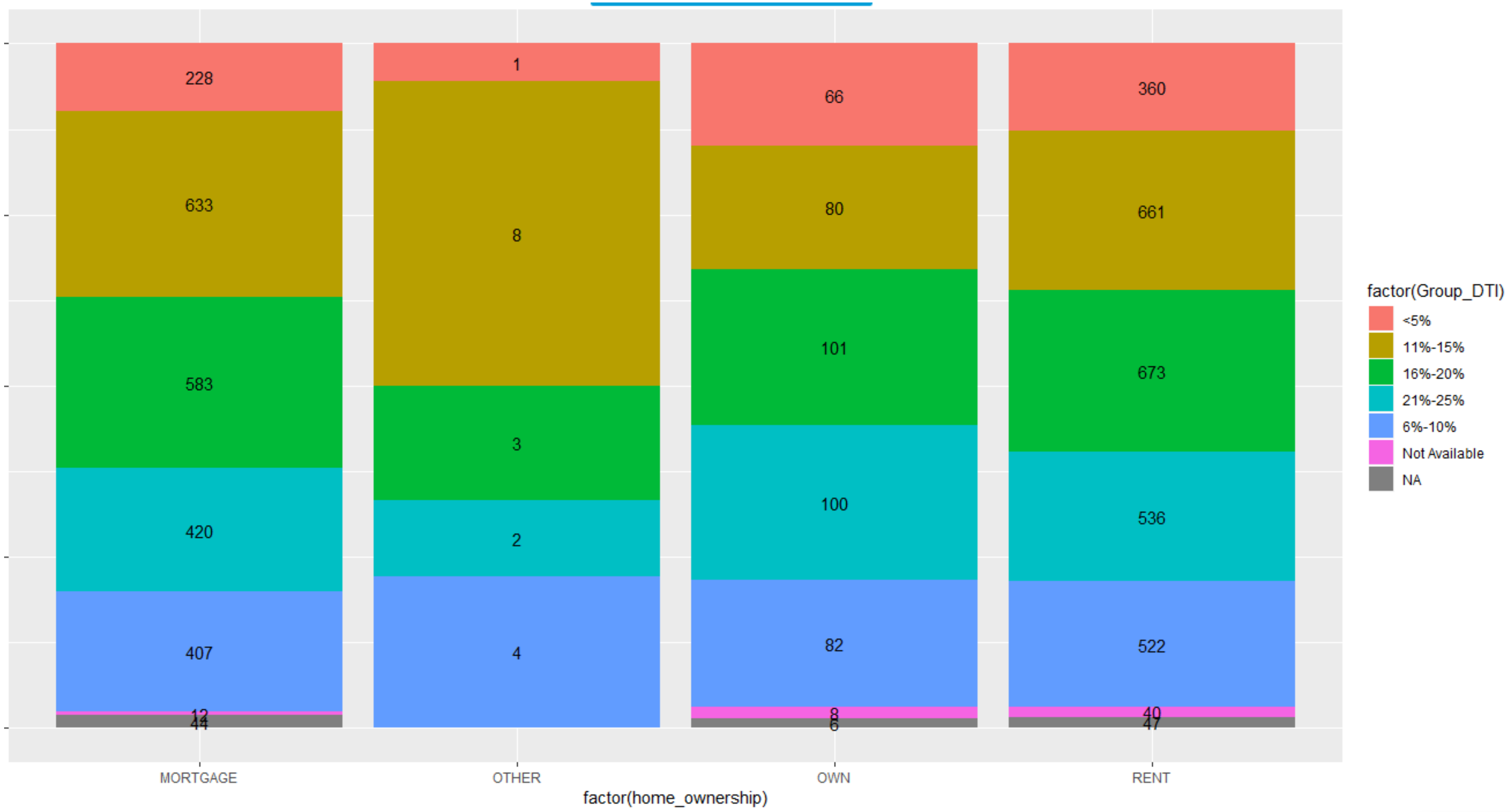
DTI and delinquency



- Customers with 0 and 1 past-due incidences of delinquency in the borrower's credit file for the past 2 years tend to default more indicating a behavior pattern

Bivariate Analysis- Charged off Loans

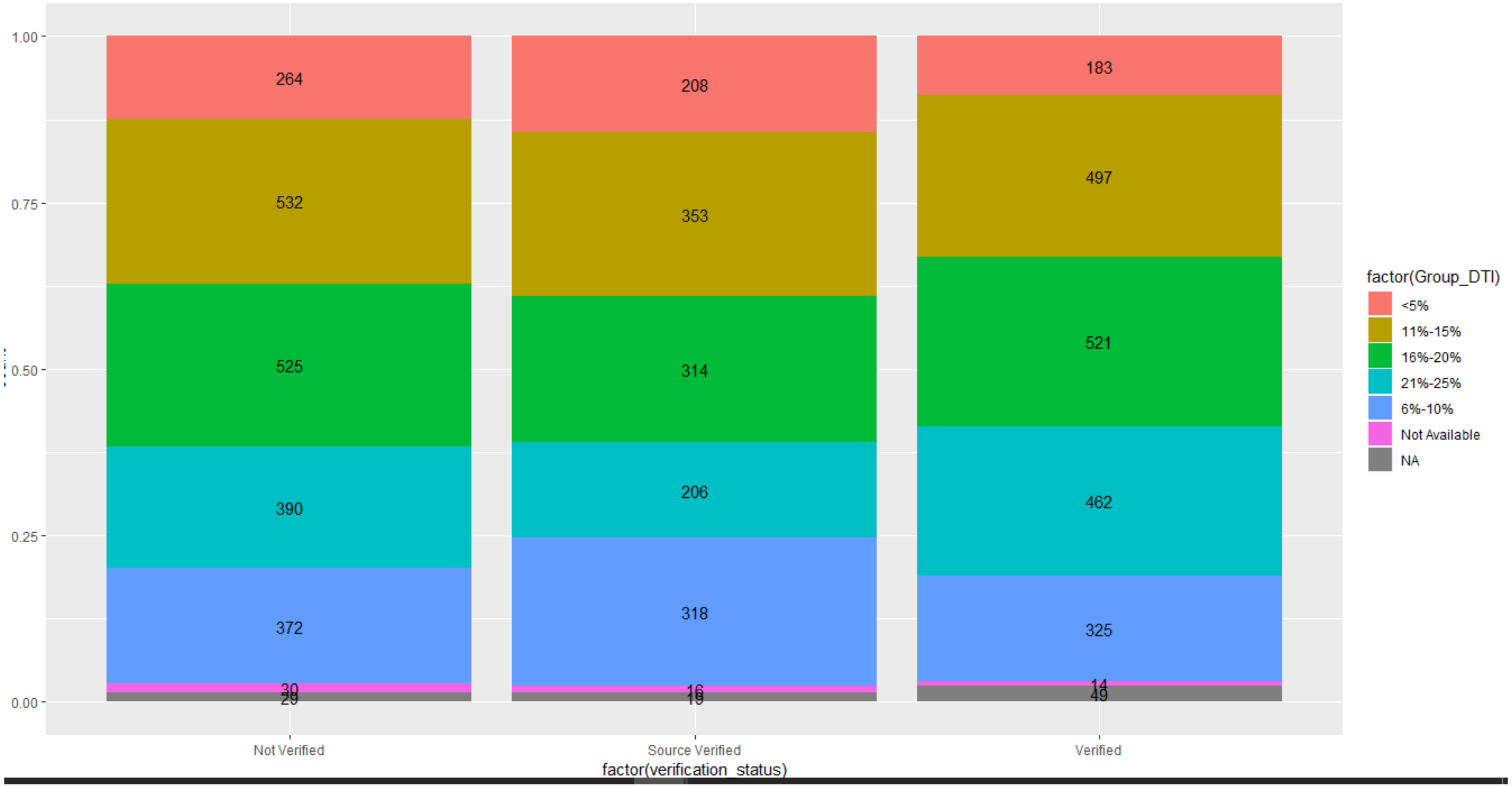
DTI and Home Ownership



- Group DTI of 11%-15% and 16%-20% identified as key segments in Rent and Mortgage Home Ownership indicating that Customers with monthly spend on Rent or Mortage tend default

Bivariate Analysis- Charged off Loans

DTI and Verification status



- Group DTI of 11%-15% and 16%-20% identified as key segments. Source Verified cases are 50% of Not Verified and Verified cases put together.
- It is required to verify the 'Source' before approving loans to avoid Charged Off Loans

Multivariate Analysis- Charged off Loans

- Univariate Outliers
  - Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit, Total Credit Accounts and Loan Verification Status are major contributors to Charged off loans
  - Fields - Group_util, Group_creditlines and verification_status

- Bi Variate Outliers
  - A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income, plays a major role when combined with Ownership of Home, delinquency in the borrower's credit file and Verification status
  - Field - Group_DTI
  - Delinquency is already captured in 'Group_creditlines' with respect to number of credit accounts open and Ownership of House classifies under category of Income linking to 'revol_util'. Hence using the field 'Group_DTI' in conjunction with the three fields Group_util, Group_creditlines and verification_status will provide us with a good prediction model to assess Current Loans

Multivariate Analysis- Charged off Loans

	verification_status	Group_util	Group_creditlines	Group_DTI	LoanNum
145	Verified	>50%	Between10-50	16%-20%	395
144	Not Verified	>50%	Between10-50	16%-20%	375
143	Verified	>50%	Between10-50	21%-25%	371
142	Verified	>50%	Between10-50	11%-15%	354
141	Not Verified	>50%	Between10-50	11%-15%	340
140	Not Verified	>50%	Between10-50	21%-25%	314
139	Source Verified	>50%	Between10-50	11%-15%	234
138	Source Verified	>50%	Between10-50	16%-20%	227
137	Verified	>50%	Between10-50	6%-10%	193
136	Not Verified	>50%	Between10-50	6%-10%	190

Charged Loans – Topliers

- Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 16-20% -> 395
- Not Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 16-20% -> 375
- Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 21-25% -> 371
- Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 11-15% -> 354
- Not Verified + UTIL > 50% + Creditlines between 10-50 + DTI between 11-15% -> 340

Inferences & Corrective actions

- Source Verification is needed to avoid Charged Loans
- Revolving line utilization rate is a key KPI which needs to be incorporated in screening while dispensing loans
- Customers with more Credit Accounts tend to miss loan repayments
- Though Bank is giving loans where DTI is <30%, they need to see DTI in conjunction with Utilization rate and Credit Lines.

Predictive Analysis

Probability of Default on Current Loans

- Calculate the Probability of Default on Multivariate Analysis fields using Fully Paid loans and Charged off loans
  - Verification Status, Revolving line utilization rate , Number of Total Credit Accounts and Debt-to-income ratio

- Apply the Probability on Current loans to predict loans that would turn into Charged Off Loans
  - mulVariatePredictiveDF data frame contains the probability for further action
  - Prob_ChargedIn column indicates the Probability of Current loans translating into Charged off loans

verification_status	Group_util	Group_creditlines	Group_DTI	Current Loan	Fully_Paid	Charged_Off	ProbabilityOfPayment	Prob_ChargedIn
Verified	>50%	Between10-50	16%-20%	127	1679	395	80.95468	24
Verified	>50%	Between10-50	21%-25%	120	1459	371	79.72678	24
Verified	>50%	Between10-50	11%-15%	80	1628	354	82.13925	14
Source Verified	>50%	Between10-50	11%-15%	63	1291	234	84.65574	10
Verified	>50%	Between10-50	6%-10%	61	1047	193	84.43548	9
Source Verified	>50%	Between10-50	16%-20%	42	1158	227	83.61011	7
Source Verified	>50%	Between10-50	6%-10%	41	845	185	82.03883	7
Not Verified	>50%	Between10-50	11%-15%	42	1978	340	85.33218	6
Not Verified	>50%	Between10-50	16%-20%	43	2110	375	84.90946	6
Not Verified	>50%	Between10-50	21%-25%	34	1500	314	82.69019	6
Source Verified	>50%	Between10-50	21%-25%	31	791	155	83.61522	5
Verified	<30%	Between10-50	16%-20%	25	406	68	85.65401	4
Verified	>50%	Between10-50	NA	26	199	33	85.77586	4

Summary

Inferential Analytics

- Improve Verification process
  - Source Verification is needed to avoid Charged Loans
  - Reduce No Verification cases

- Revolving line utilization rate is a key KPI which needs to be incorporated in screening while dispensing loans. The **debt**-to-**limit ratio** is the **ratio** of a consumer's total credit card balances versus total credit card limits, expressed as a percentage, Lower the better.

- Customers with more Credit Accounts tend to miss loan repayments. Factor in while dispensing loans

- Though Bank is giving loans where DTI is <30%, they need to see DTI in conjunction with Utilization rate and Credit Lines.

Predictive Analytics

- 178 of 1140 Current loans identified to convert into Charged off loans based on Probability of default.

- Key factors are Revolving line utilization rate > 50% & Credit Accounts between 10 - 50

- Banks will have to closely monitor these categories to avoid Exposure at Default.

THANK YOU