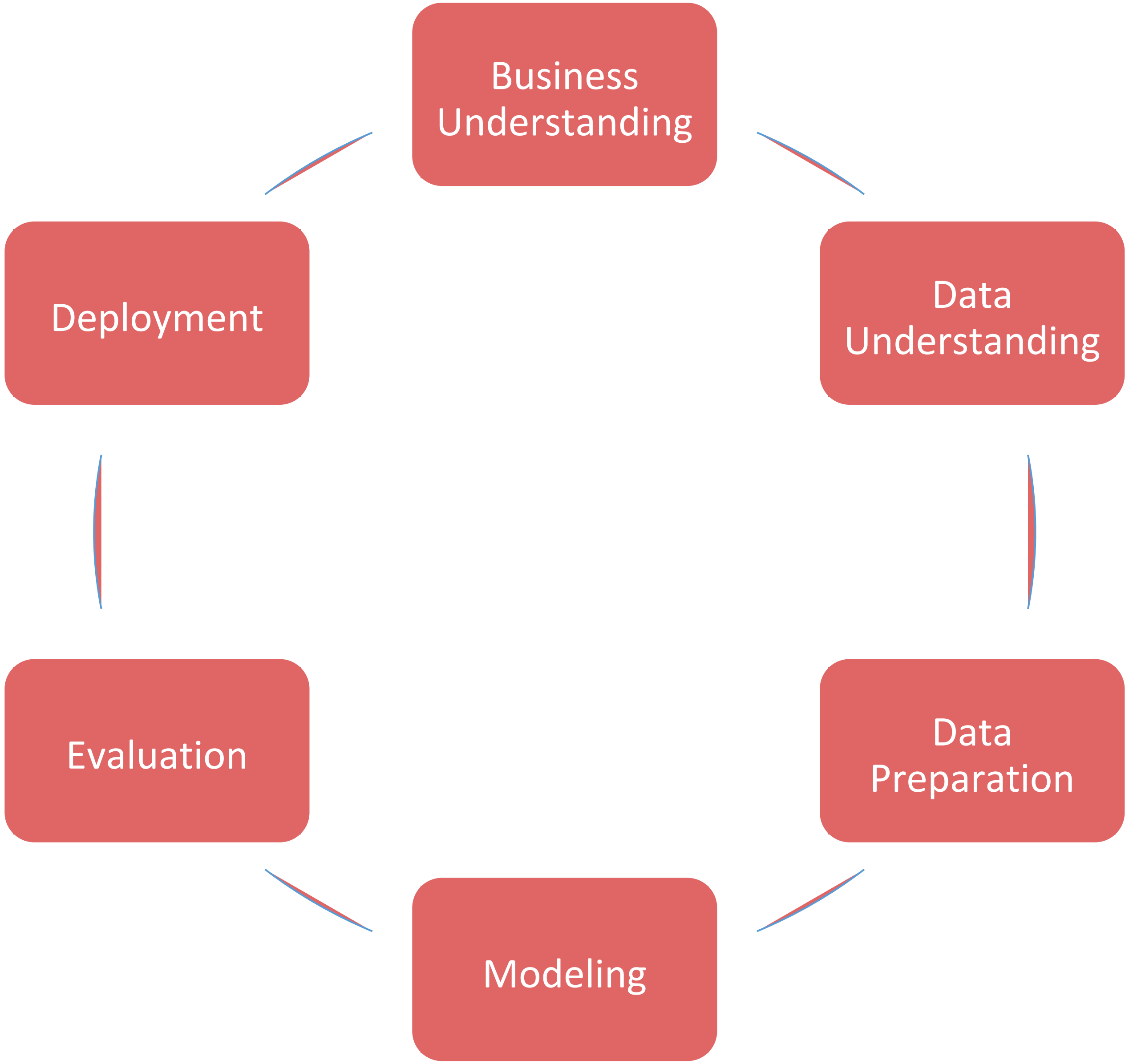
Gramener Case Study

Group Name: DataCatalysts

1. Sudarshan Bhattacharjee
2. Sagnik Banerjee
3. Huzaifa Hareeri
4. Yashjit Gangopadhyay

CRISP Data-mining process flow



- **Business Objective:**

The finance company is looking for indicators in an applicant’s profile in order to help them decide whether to approve or decline a loan application.

- **Goal of Analysis:**

To find out the relation among the various indicators and their impact on loan default. And suggest which attributes contributes a significant difference in Loan Default.

Data Understanding

The following indicators have some important attributes that will help us understand the behavior of the approved loan customers w.r.t. loan default.

annual_inc - Annual Income of applicant

loan_amnt - The listed amount of the loan applied for by the borrower

funded_amnt - The total amount committed to that loan at that point in time

int_rate - Interest Rate on the loan

Grade - LC assigned loan grade

Dti - Debt to income ratio

emp_length - Employment length in years

Purpose - A category provided by the borrower for the loan request.

home_ownership - The home ownership status provided by the borrower during registration

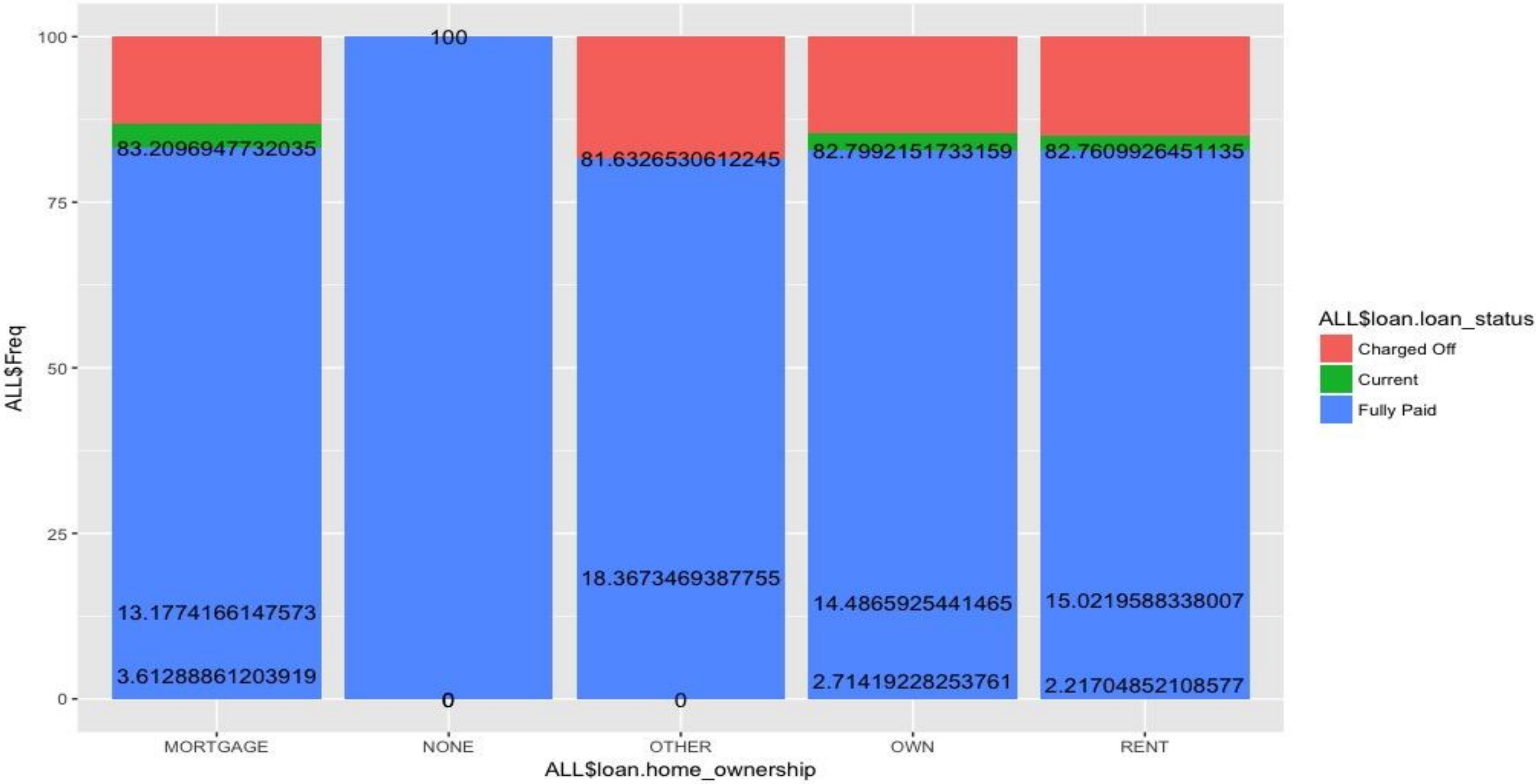
loan_status - Current status of the loan

Data Cleaning and Preparation

- The file (loan.csv) was loaded in R and it was noted that there are 39717 observations of 111 variables.
- ID is the unique variable in the dataset. We check for blanks and remove all the blanks rows from all the driving variables.
- We found that there are NA variables are present in the dataset thus they too were removed.
- We have removed the loan$url from our data sets as it was not required for our analysis. We also converted all the character variable to factor variable, and labelled them as per the instruction provided and created new variables with new labels on which we performed our analysis.

Status of loan per various categories

- We see that about 18% of the applicants tends to default when they apply for “OTHER” loan.

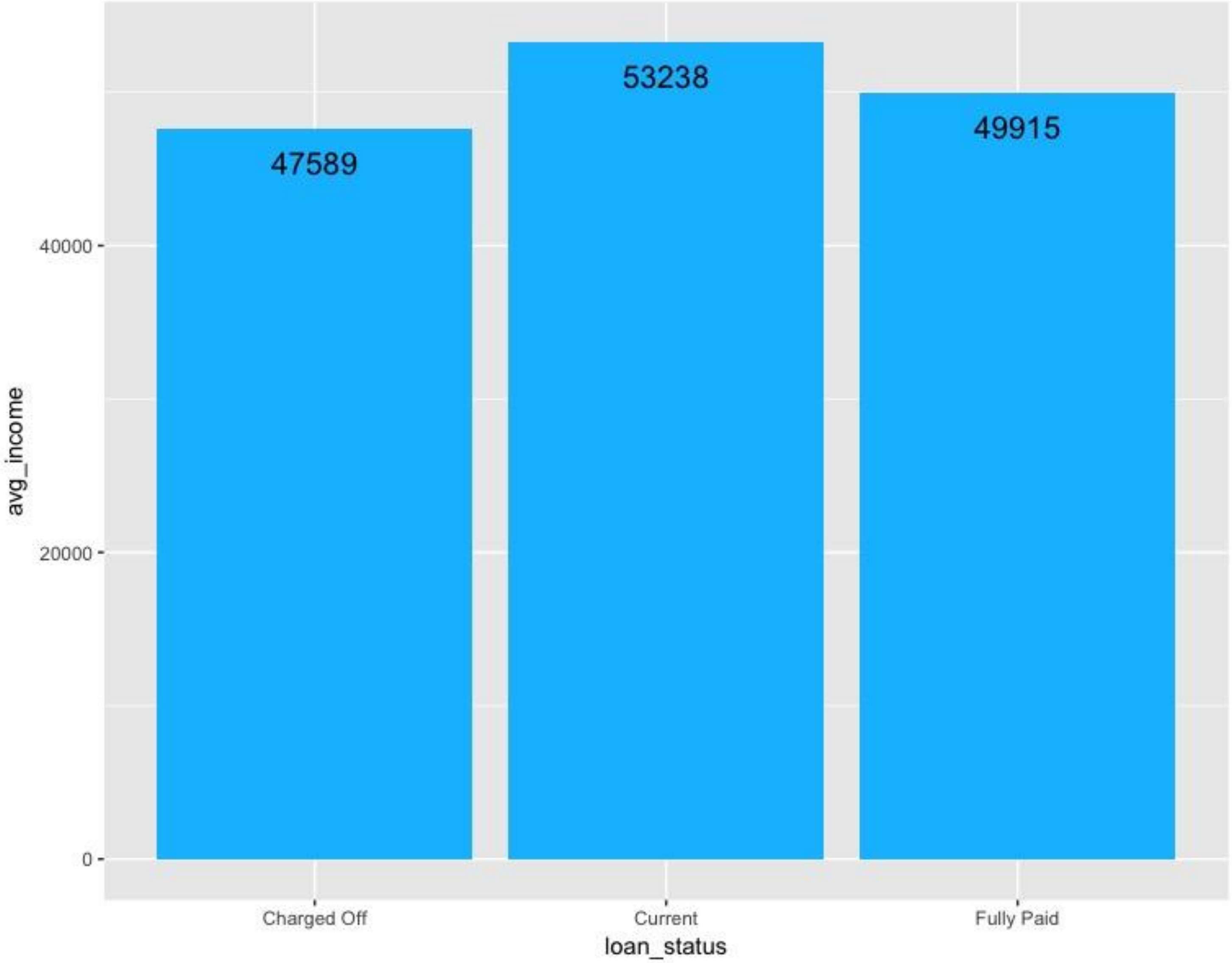


Finding relationship with the annual income average

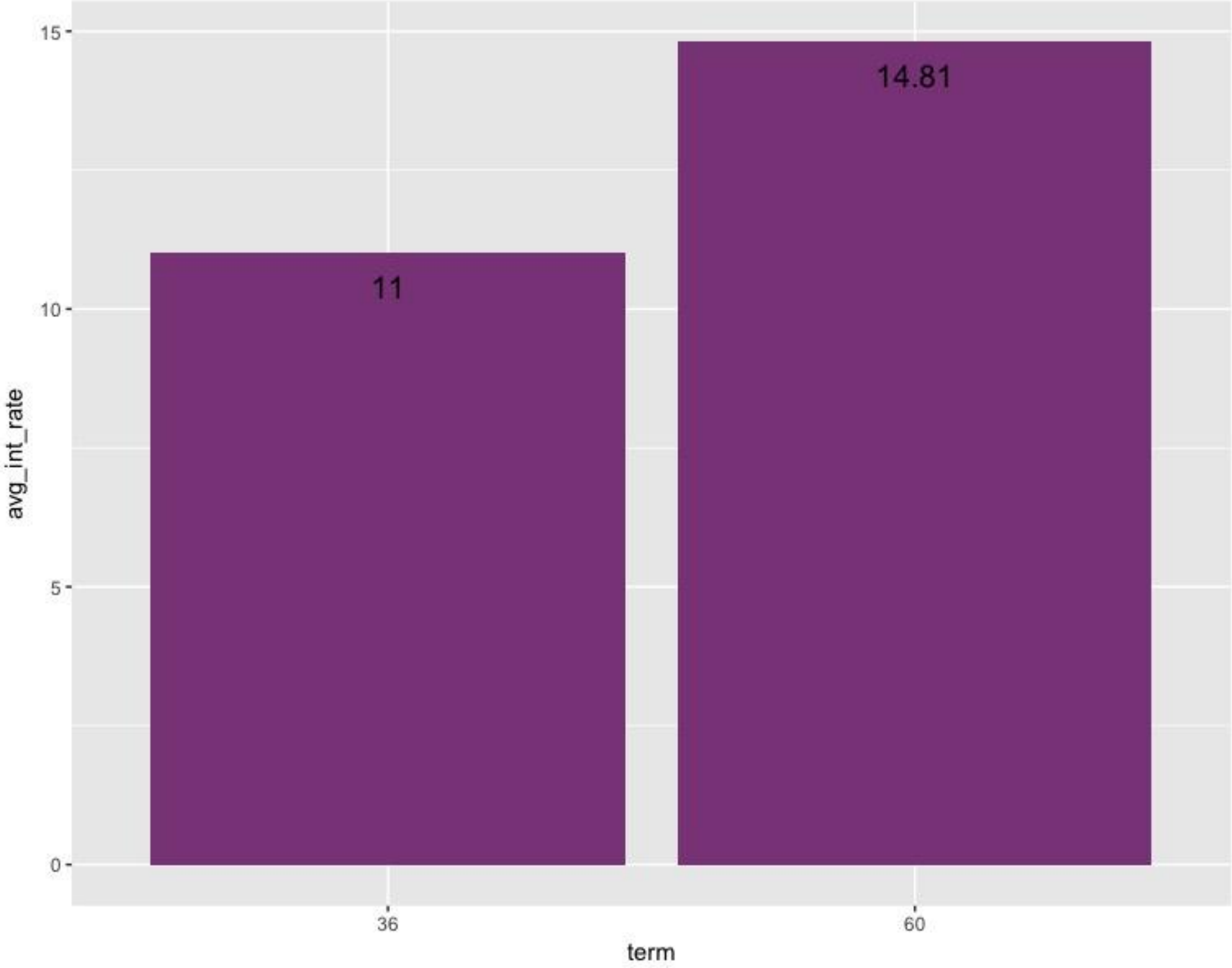
- It can be seen that loan takers of average annual income of less than 47589 generally have to get their loans charged off.

- So it can be assumed that loan requests within this range of income has historically been a reason for financial loss for the firm.

- As a remedial measure, we can suggest giving loans at higher interest rates than usual for this sector of the applicants.



Relationship between term and interest rate

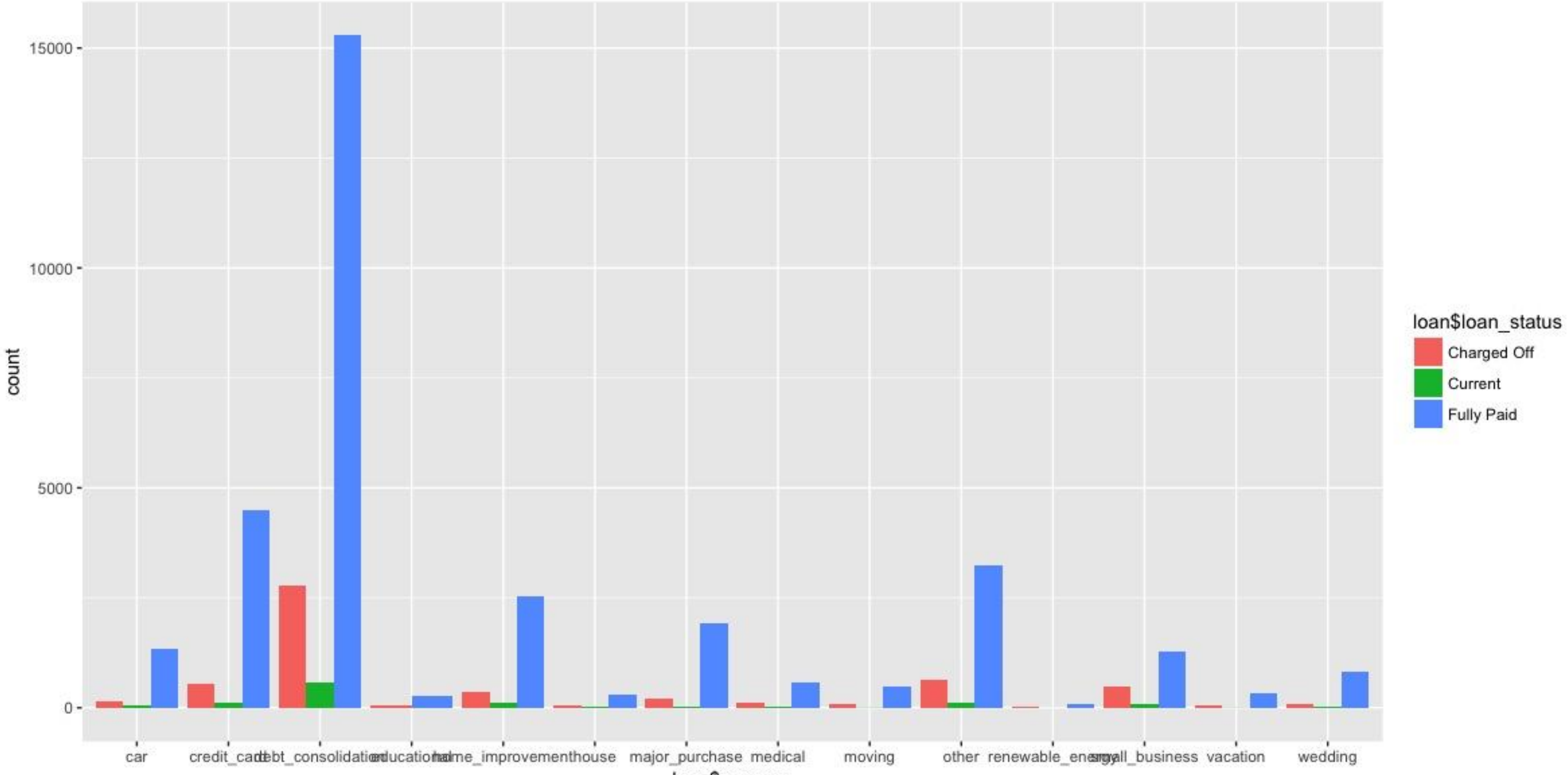


It has been observed that a loan with a payment term of 60 months is more likely to be defaulted as compared to a loan with a payment term of 36 months.

(NOTE: The current loan portfolio doesn’t have any loan with a payment term of 36 months)

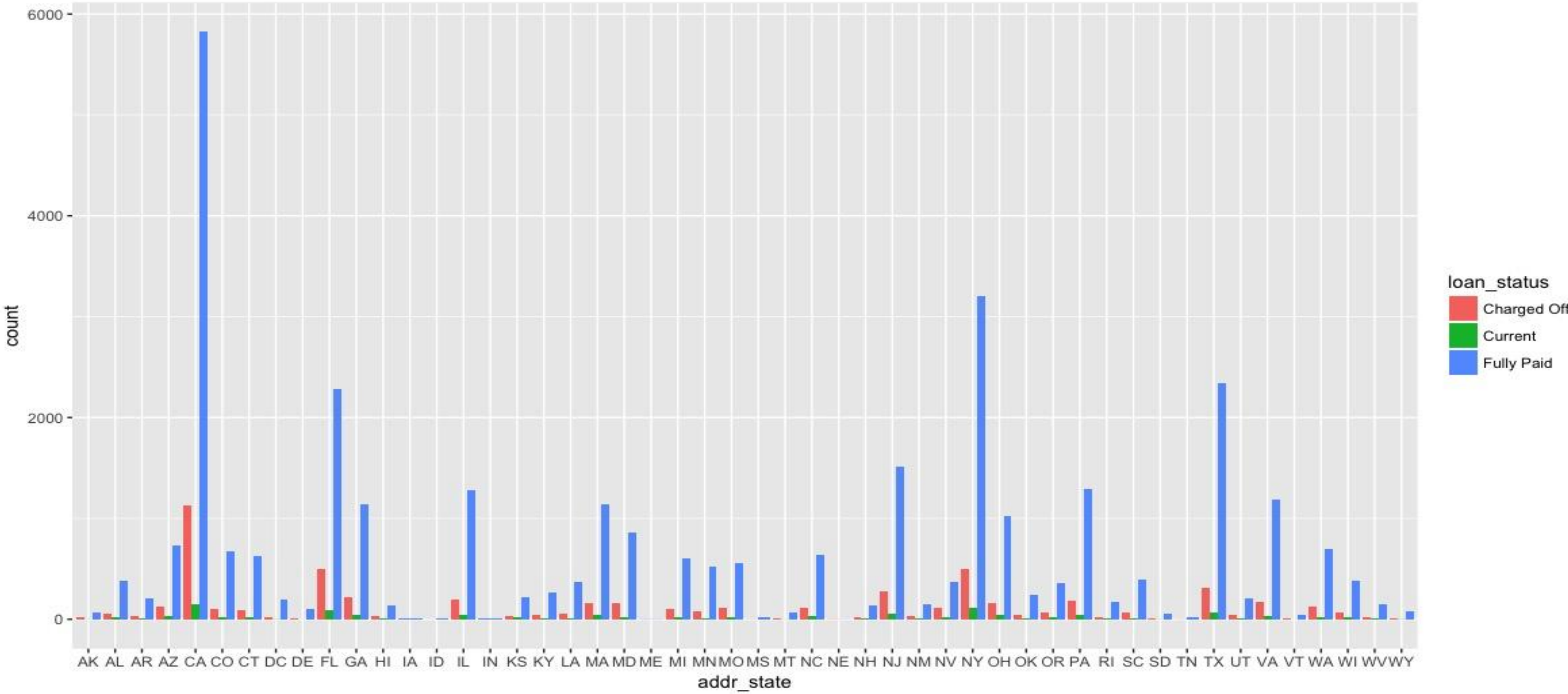
Status of loan categorised by purpose

- debt_consolidation shows the highest amount of loans that have been defaulted, followed by credit_card loans and other_renewable_energy loans.



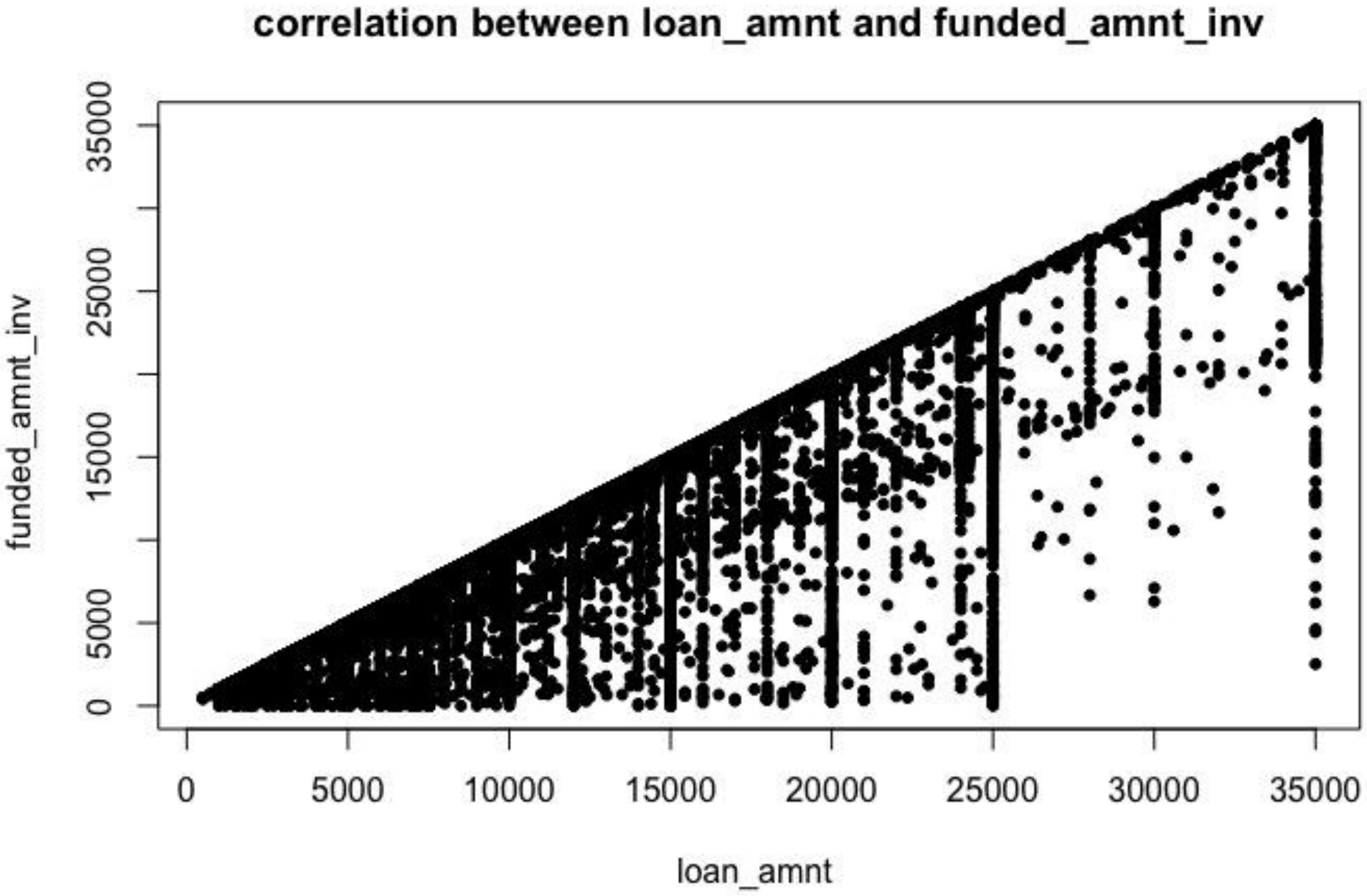
Status of loan categorised by state (USA)

California (CA) and Florida (FL) have the most amount of loan defaulters. New York (NY), Illinois (IL) and Texas (TX) are the states where the decisions for approving loans are taken most correctly.

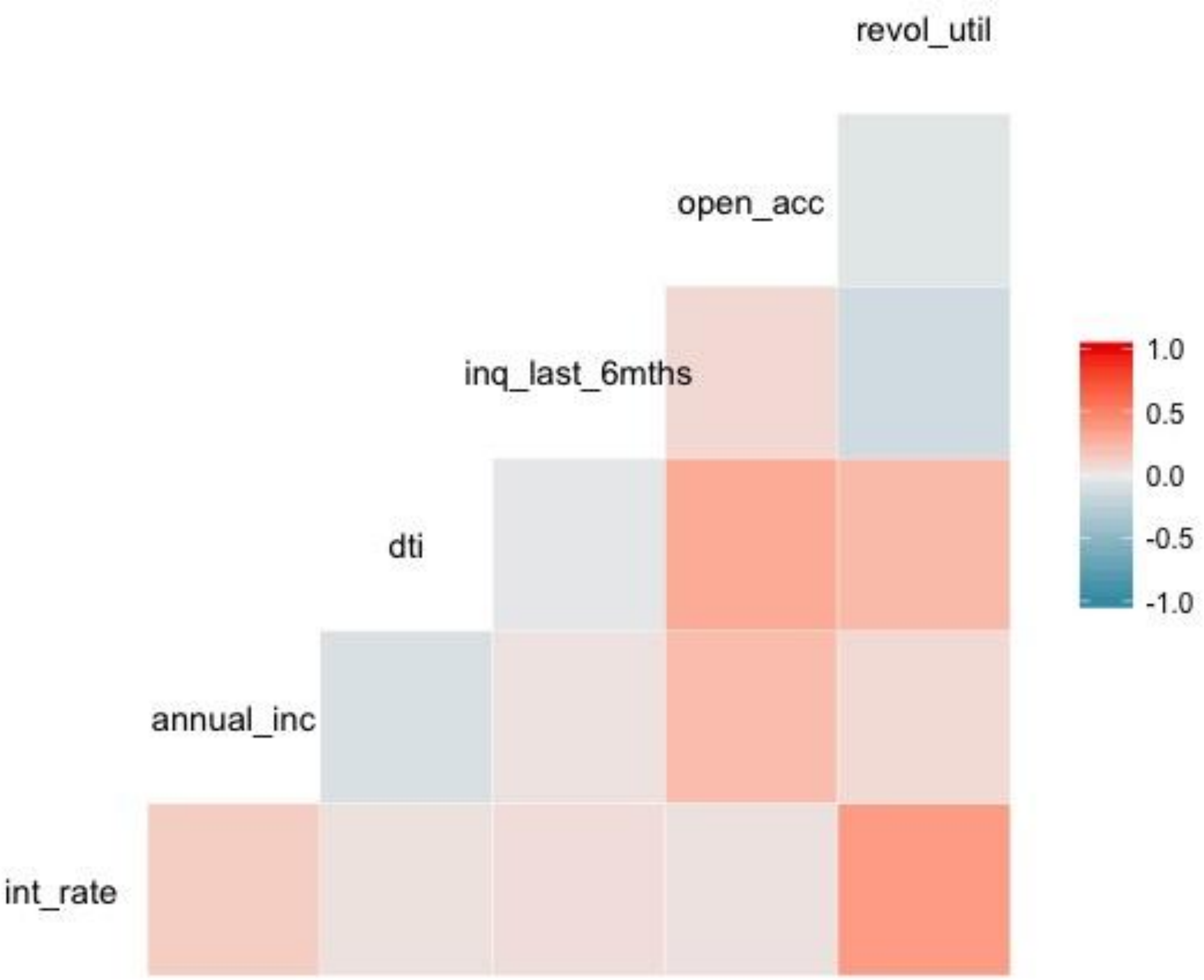


Exploratory Data Analysis: Correlations

There is correlation between funded_amt & Loan_amt, which means %increase in funded_amount tends to %increase in loan_amount.



Correlation Matrix: Continuous variables



- We have used correlation Matrix to understand which continuous variables have positive, negative & zero correlation with each other.
- Revol_util & int_rate are closely related to each other. Interest rates on loans are higher for people with a higher revolving credit line utilisation rate. This makes sense as people who are utilizing a significant amount of available credit all the time, have the risk of running out of credit. So, the firm must recover these loans fast.
- Annual_inc is most negatively correlated to DTI. A higher DTI is an indicator that it is highly probable the loan which is current now may need to be charged_off. This correlation & the analysis in slide #7 reiterates that loan applicants with income less than 47589 should be given loan on higher rates / not given at all.

Conclusion:

1. 18% of the loan applicants default when they apply for “OTHER” loans, 15% applicants default when they apply for “RENT” loans and 14.5% applicants default when they apply for “OWN” loans.

2. Annual_inc is most negatively correlated to DTI. loan takers of average annual income of less than $47,589 generally have to get their loans charged off. These applicants must be charged high interest rates to recover the loan amount fast. The firm in some case might also decide not to give a loan to an applicant in this income group.

3. Loans with a payment period of 60 months are most likely to be defaulted as compared to the loans with payment period of 36 months.

4. CA region have highest number of charged off as well as fully paid loans. FL region is where the loan approval decisions are taken the most incorrectly.