EDA CASE STUDY

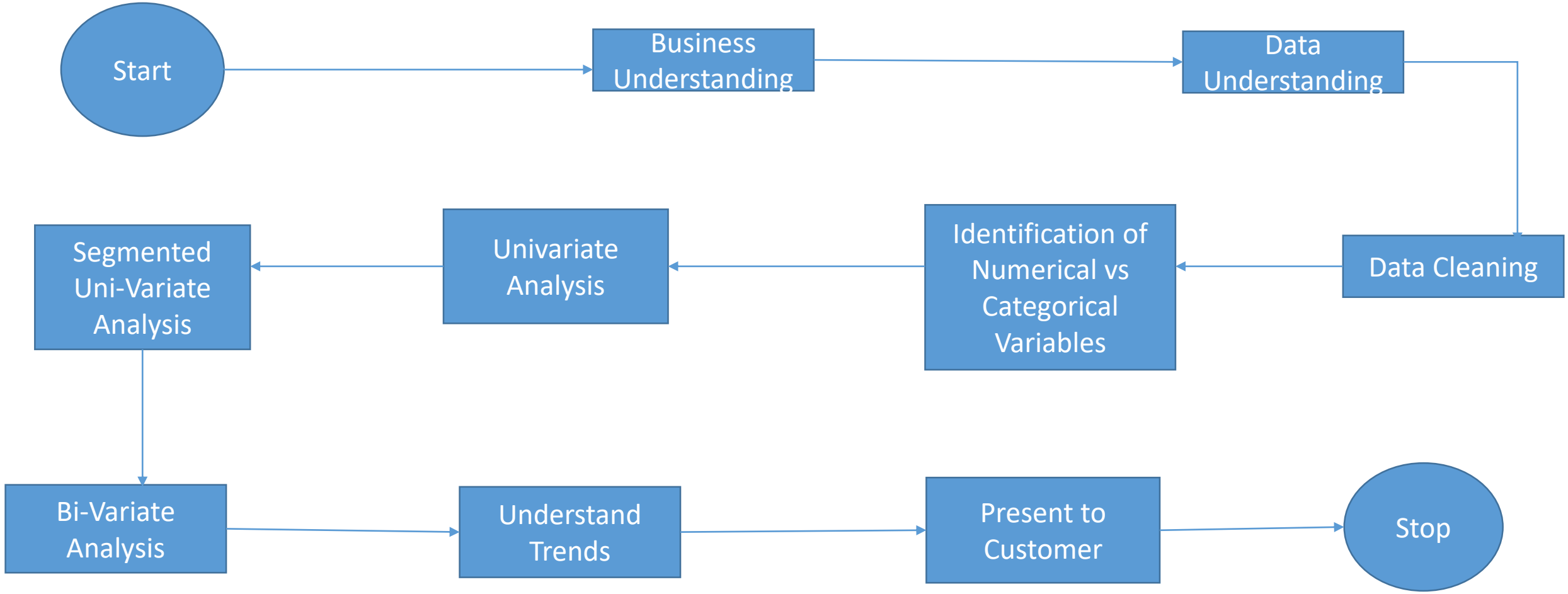
Group Members:

1. Gayathri Madhira
2. Apurva Deshpande
3. Karthik Ramakrishnan
4. Jayashri M

Problem Objectives :

- Finding the defaulters who will not be able to pay the loan back upon doing an EDA.
- Finding the factors responsible for people not paying the loan back (defaulting)
- These two will be used by banks to decide whether to grant loan to a particular person or not.

Methodology Flow Chart:



Business Understanding:

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.
- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

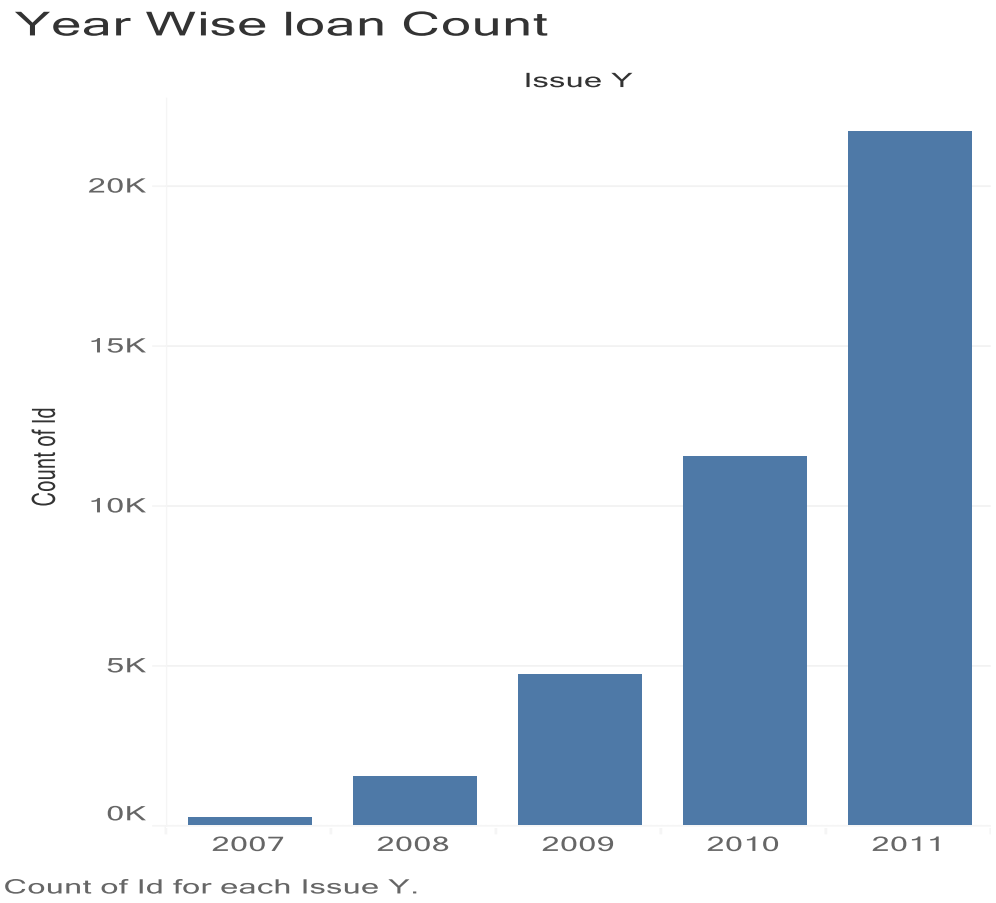
Data Understanding:

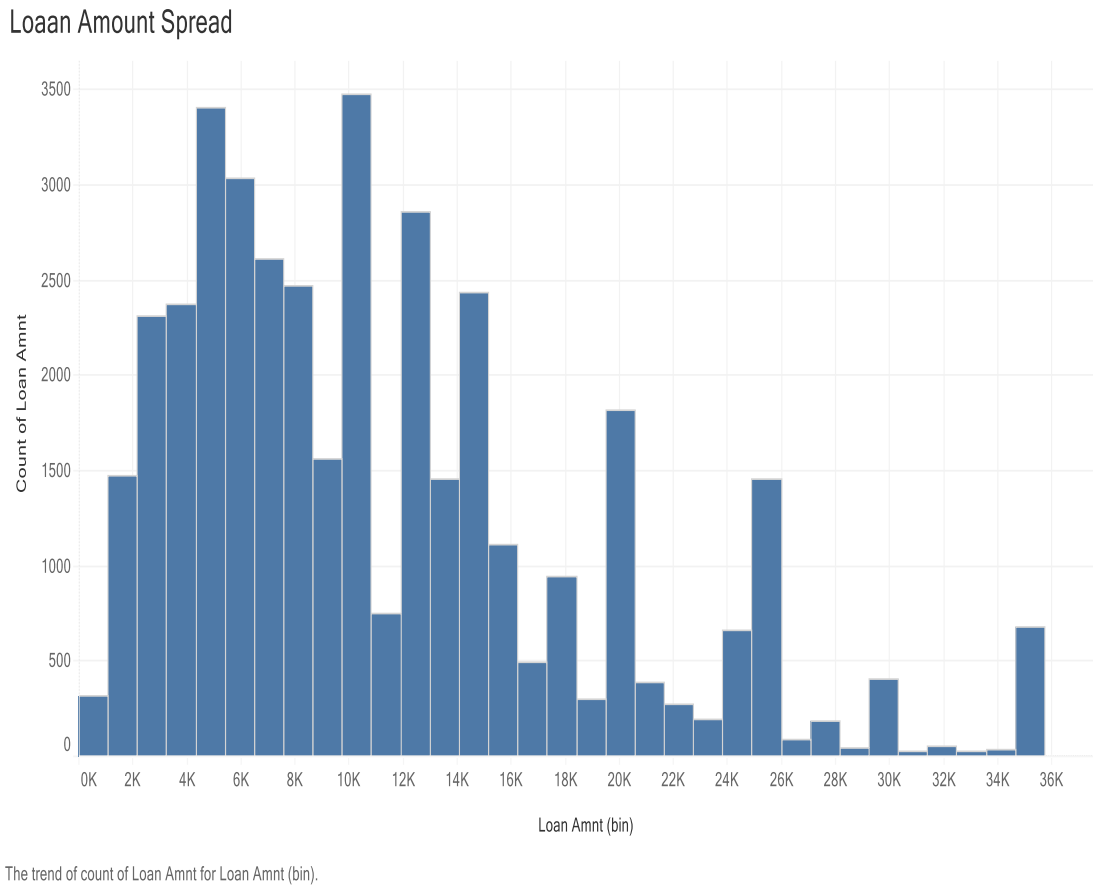
- The Data Set contains the complete loan data for all loans issued through the time period 2007 t0 2011.
- The Data Set has 111 attributes (coloumns) and 39717 data points (rows) .
- Each data points refer to a loan application and has details like member id, loan amount , annual income, current balance, home ownership, previous loan details, interest rate etc.,
- There’s a huge chunk of attributes with more null percentages.

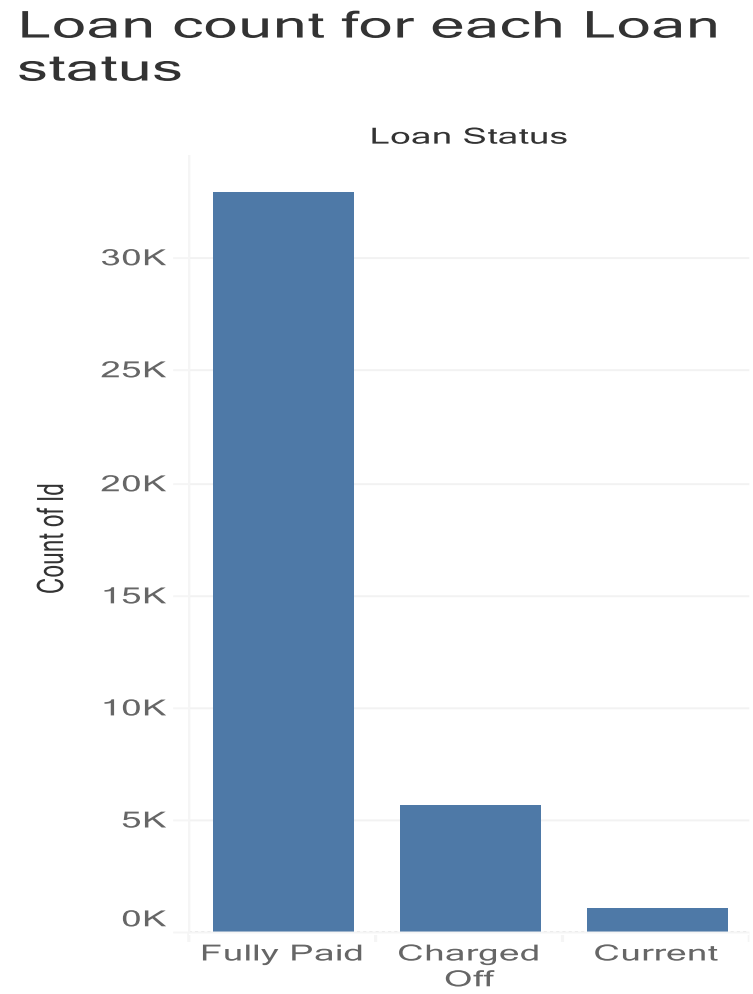
Data Cleaning:

1. We have removed all the columns that have constant values for all data points. Because , There is no variance.
2. We have also removed those columns which are irrelevant. E.g. url.
3. We have also removed those columns which needs NLP and advanced text processing as they are out of scope as of now
4. Redundant columns are removed.
5. Columns which does not have any other value other than NA or 0 are also removed.
6. Few column’s data points are trimmed to make numerical analysis easier. E.g, % has been trimmed.
7. We also have type casted the trimmed columns for numerical analysis.

Univariate Analysis







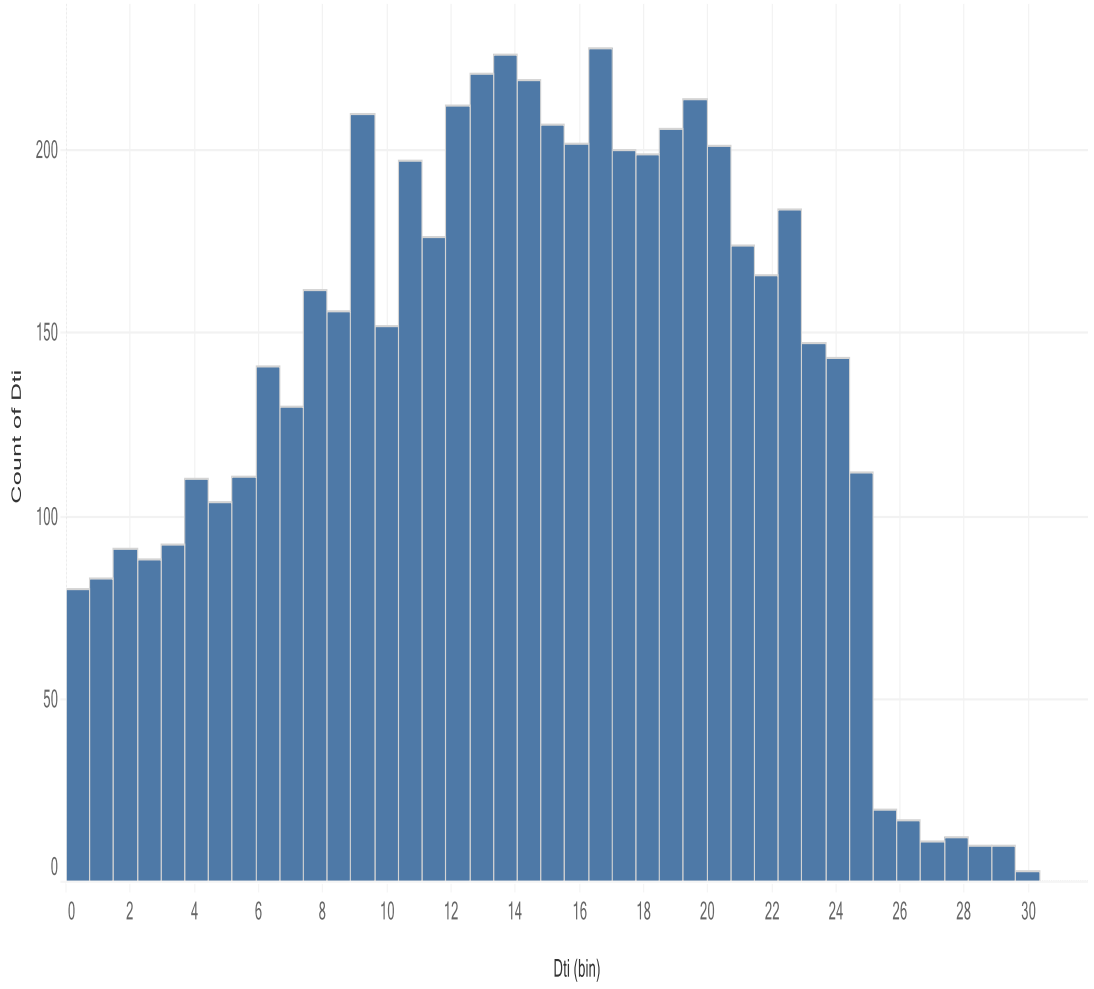
Count of Id for each Loan Status.

Analysis

- The maximum number of loans request were received in the 2011 of about 21656
- Maximum loan Amount Requested lies between (5k to 15k)
- We can also see that the maximum loan amount request is about 10 K (3477 Request)
- The No of Defaulters is 5627 which is 14% of the total request

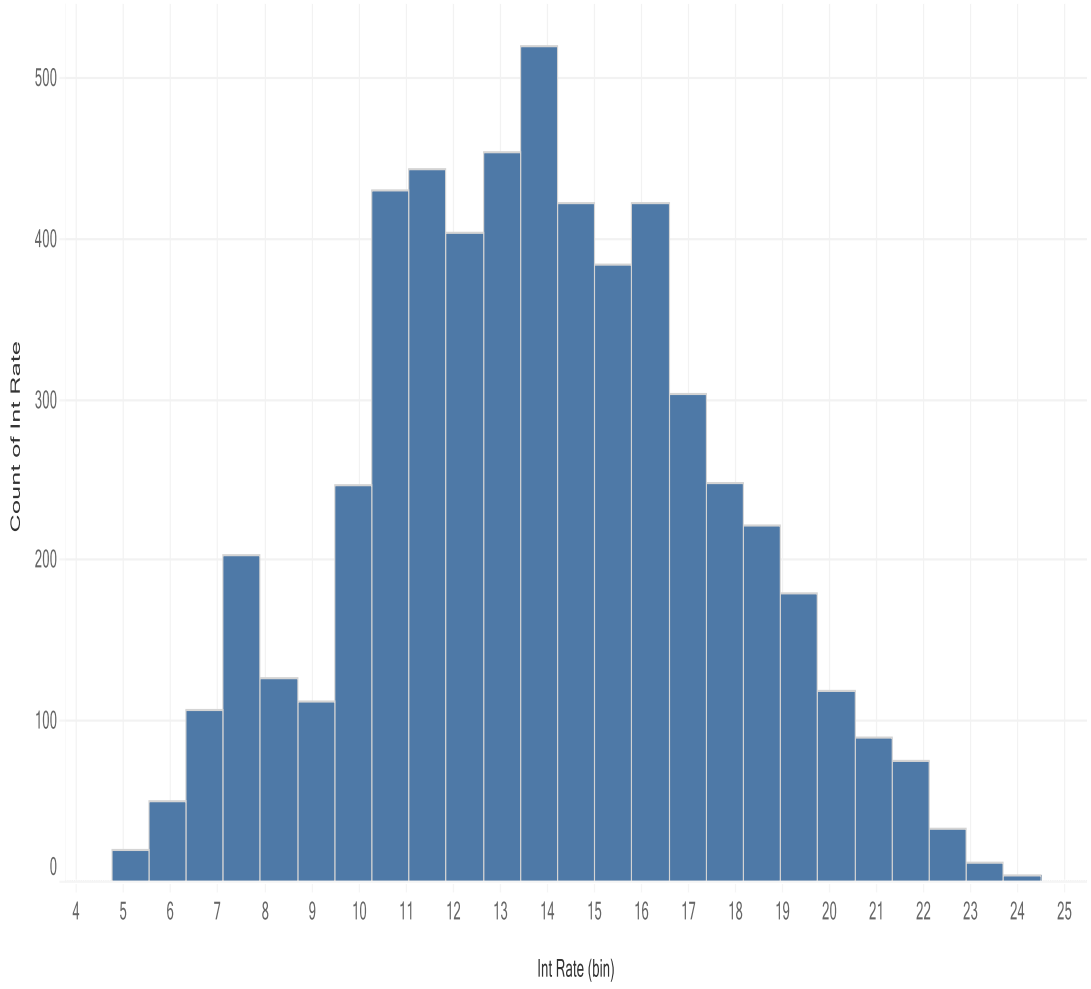
Univariate Analysis on Loan Defaulted Request

Debt to Income Ratio Frequency



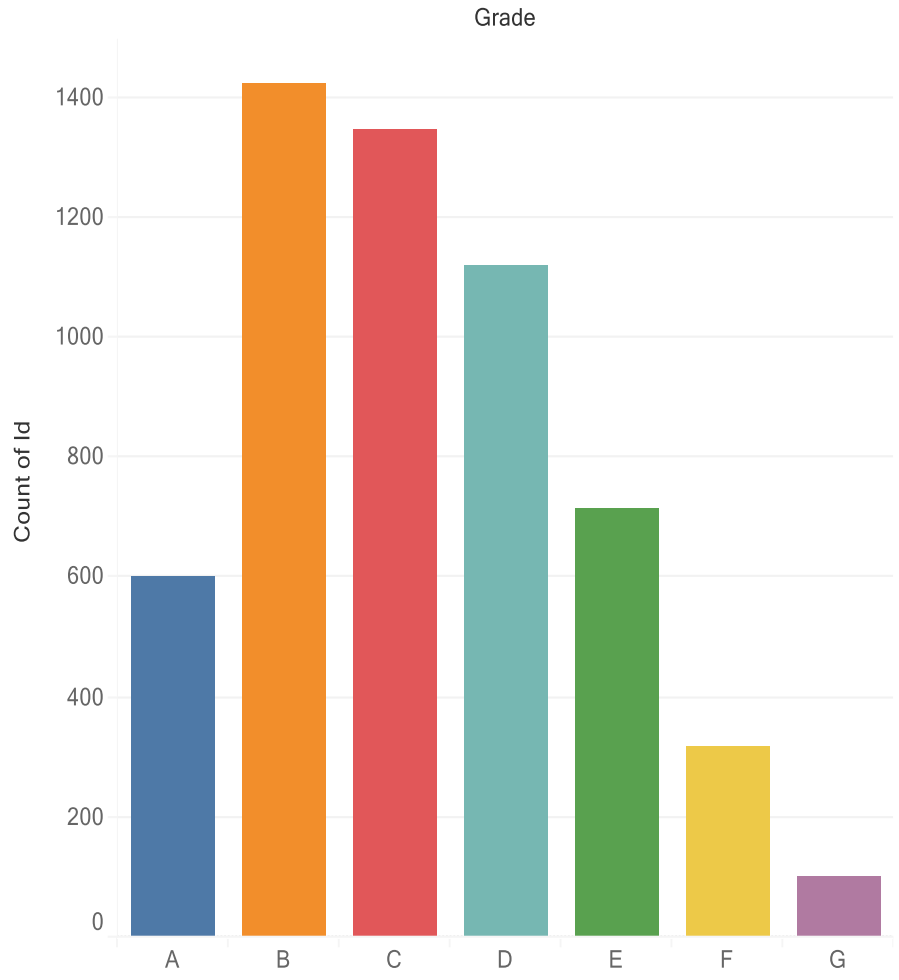
The trend of count of Dti for Dti (bin).

Number of Loans Via Interest Rate



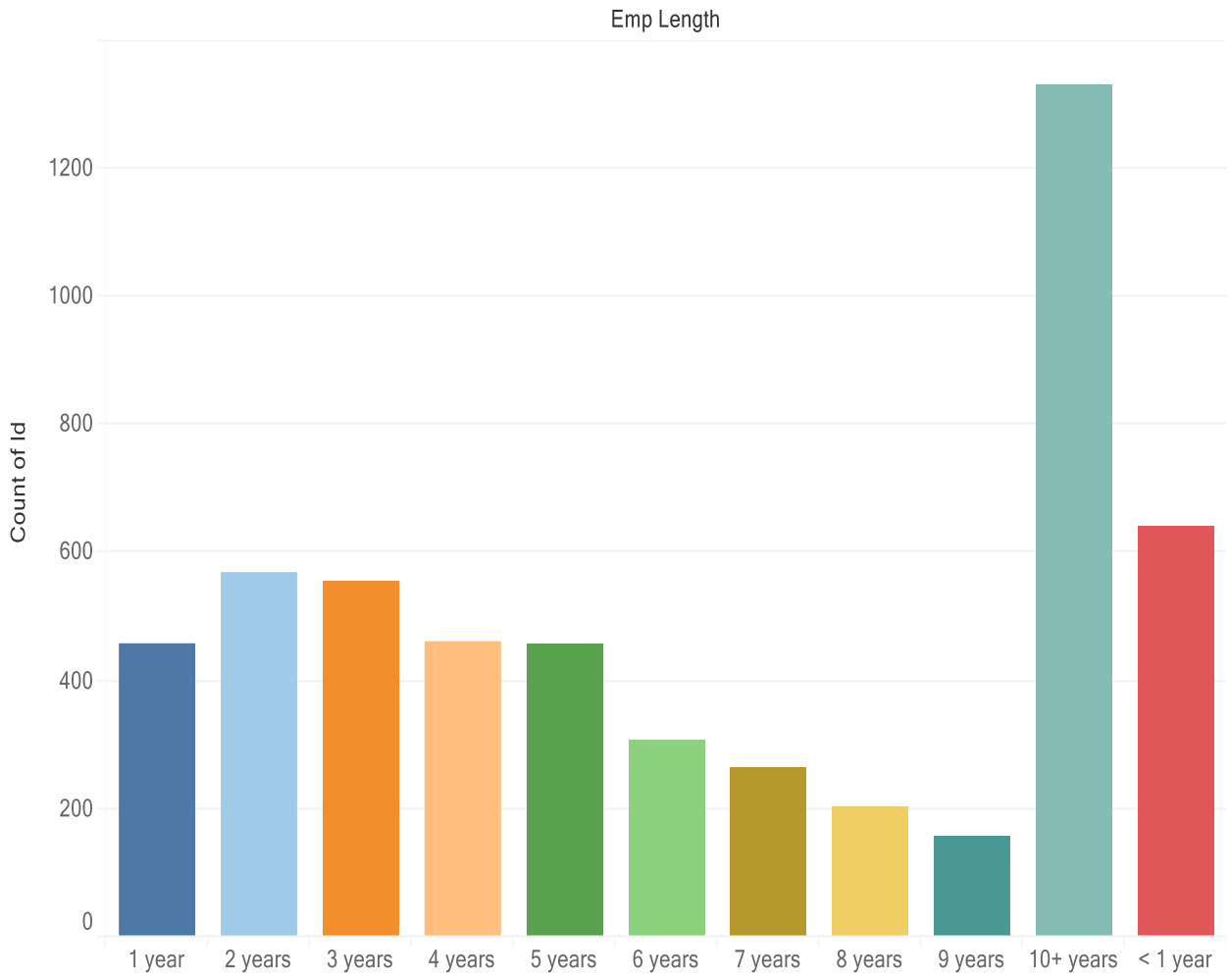
The trend of count of Int Rate for Int Rate (bin).

Number of Loans distributed via Grades



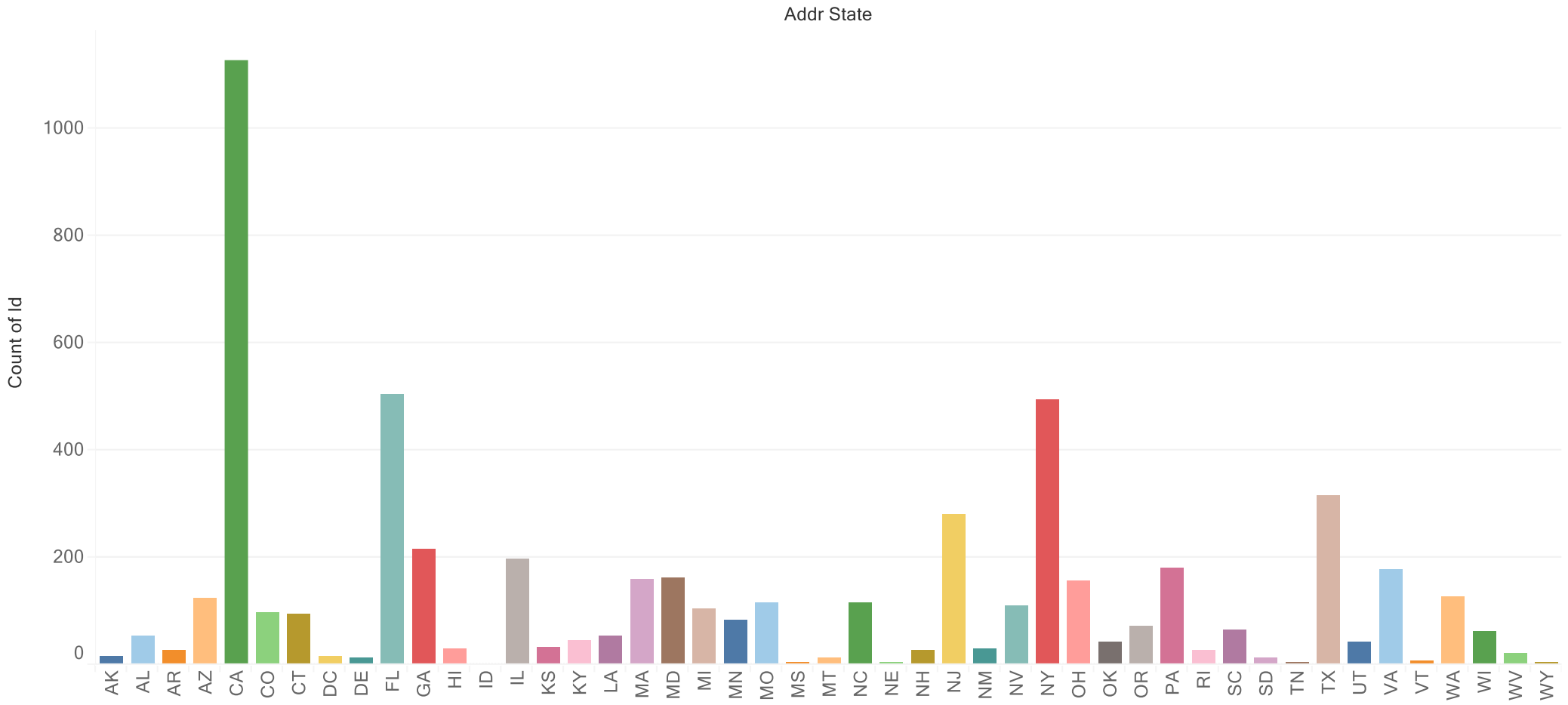
Count of Id for each Grade. Color shows details about Grade.

Count on the basis of experience



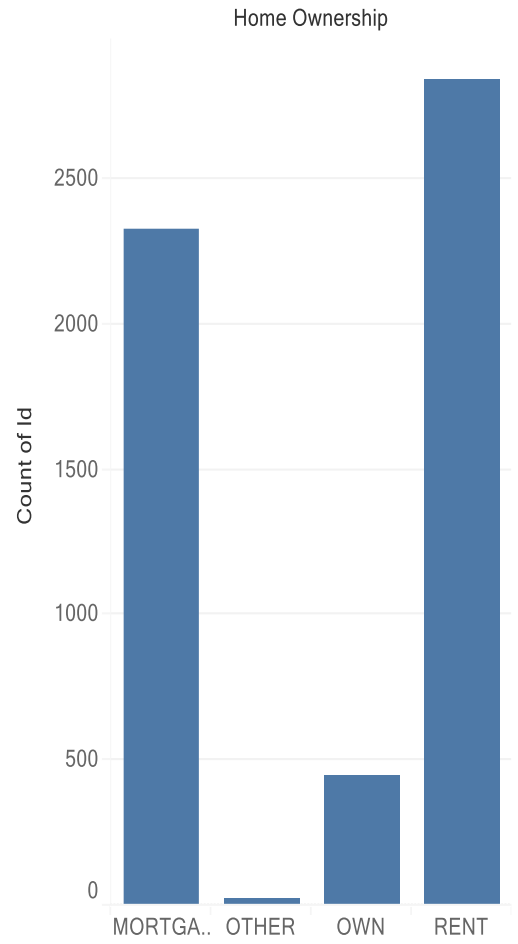
Count of Id for each Emp Length. Color shows details about Emp Length. The view is filtered on Emp Length, which excludes Null.

Region wise counts



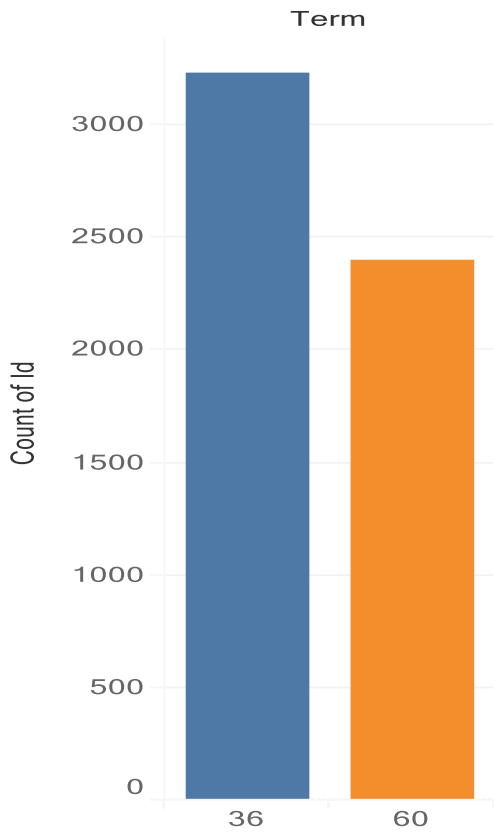
Count of Id for each Addr State. Color shows details about Addr State.

Home Ownership



Count of Id for each Home Ownership.

Number of applicants with term



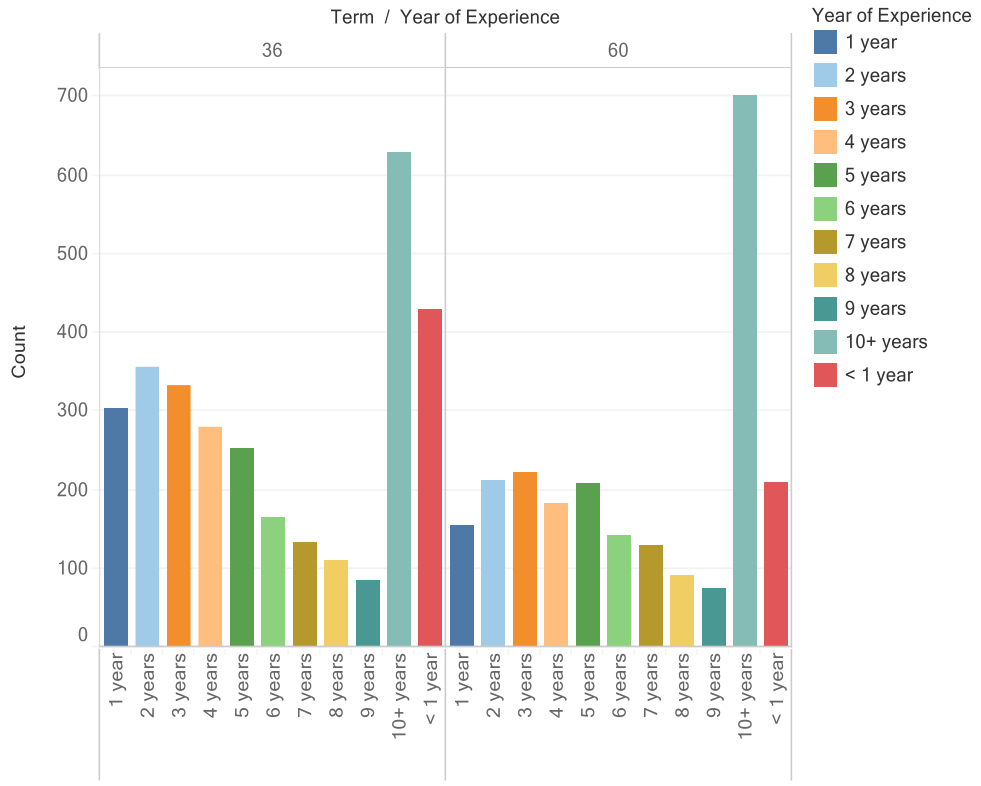
Count of Id for each Term. Color shows details about Term.

Analysis :

- We can see that DTI, Interest Rate, Home Ownership, Term, Grade, Years of Experience, Region are the factors which have huge impact on Loan Defaulters
- Detection of outliers in the sections of annual income and funded amount was found and the same has been carried with the plots too.
- We have seen significant difference between the annual income, funded amount etc. we can conclude that if we opine to give loan to only those income group which do not come under or fall under default_new, we can overcome from committing the type 1 error.

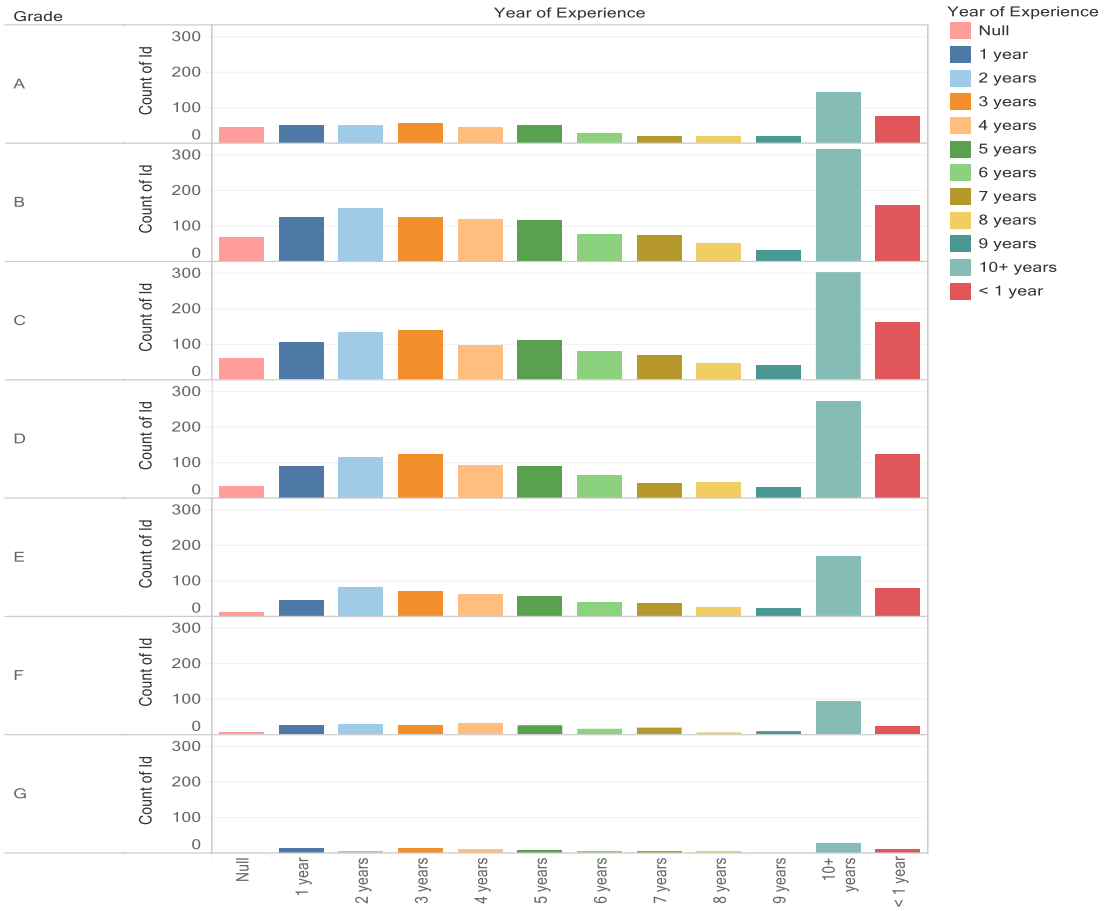
Bivariate Analysis

Count of applicatiants on basis of term and years of experience



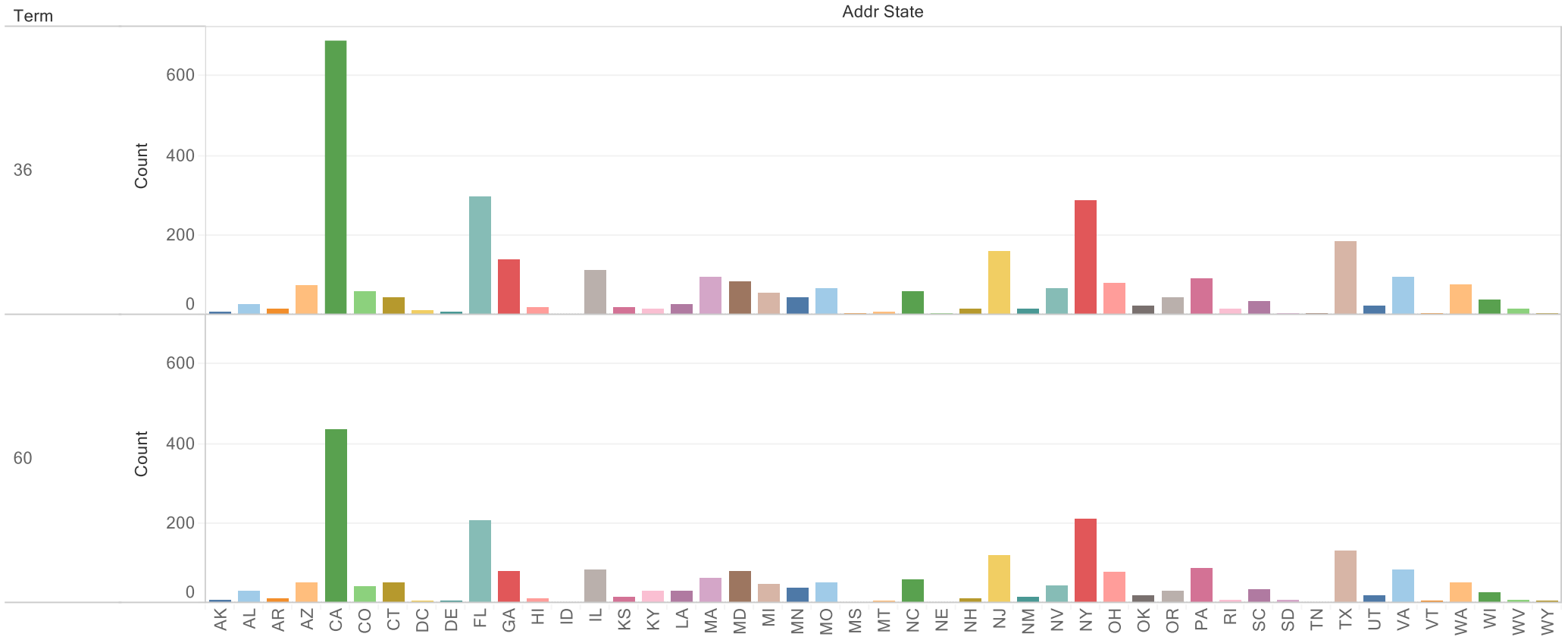
Count of Id for each Year of Experience broken down by Term. Color shows details about Year of Experience. The view is filtered on Year of Experience, which excludes Null.

Count of applicatiants on basis of grade and years of experience



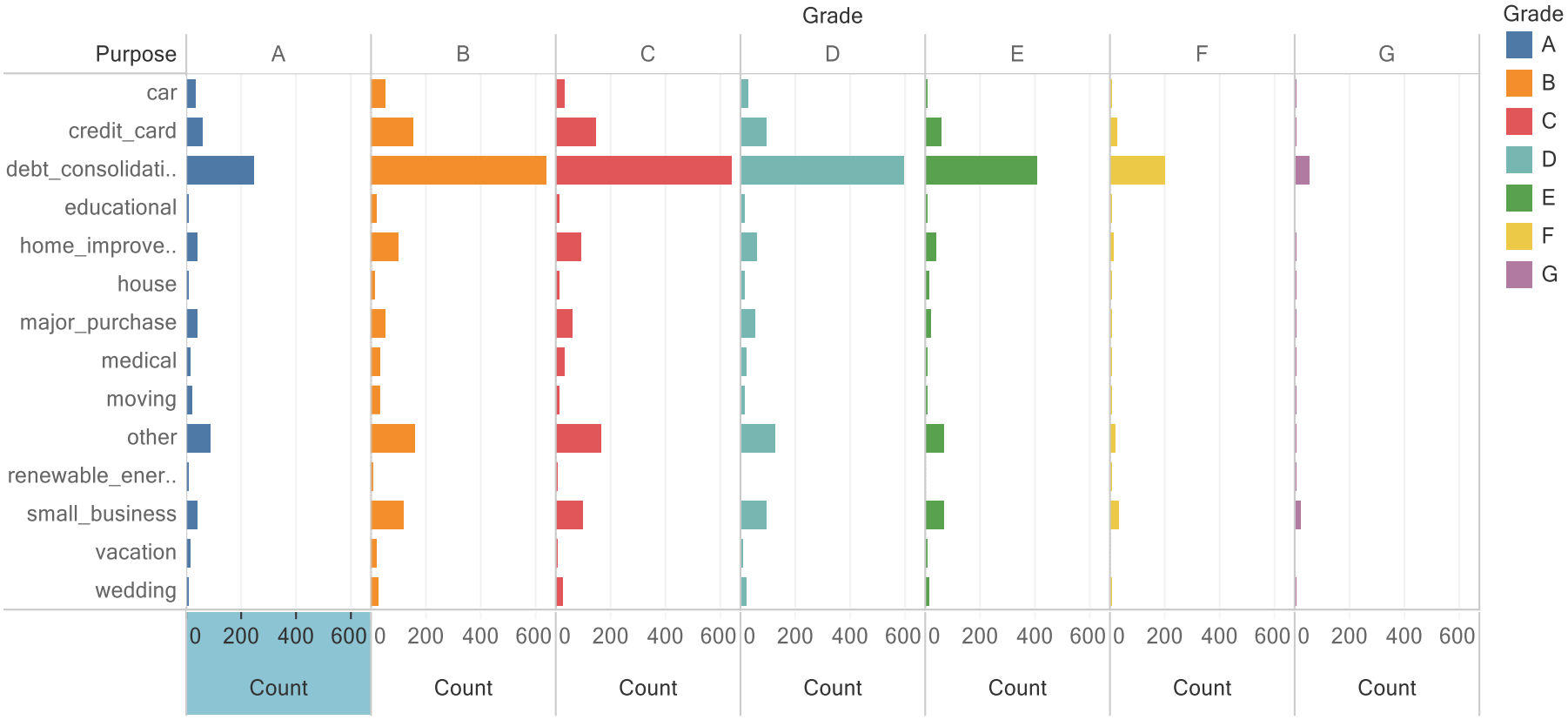
Count of Id for each Year of Experience broken down by Grade. Color shows details about Year of Experience.

Number of applicants with term for different Region



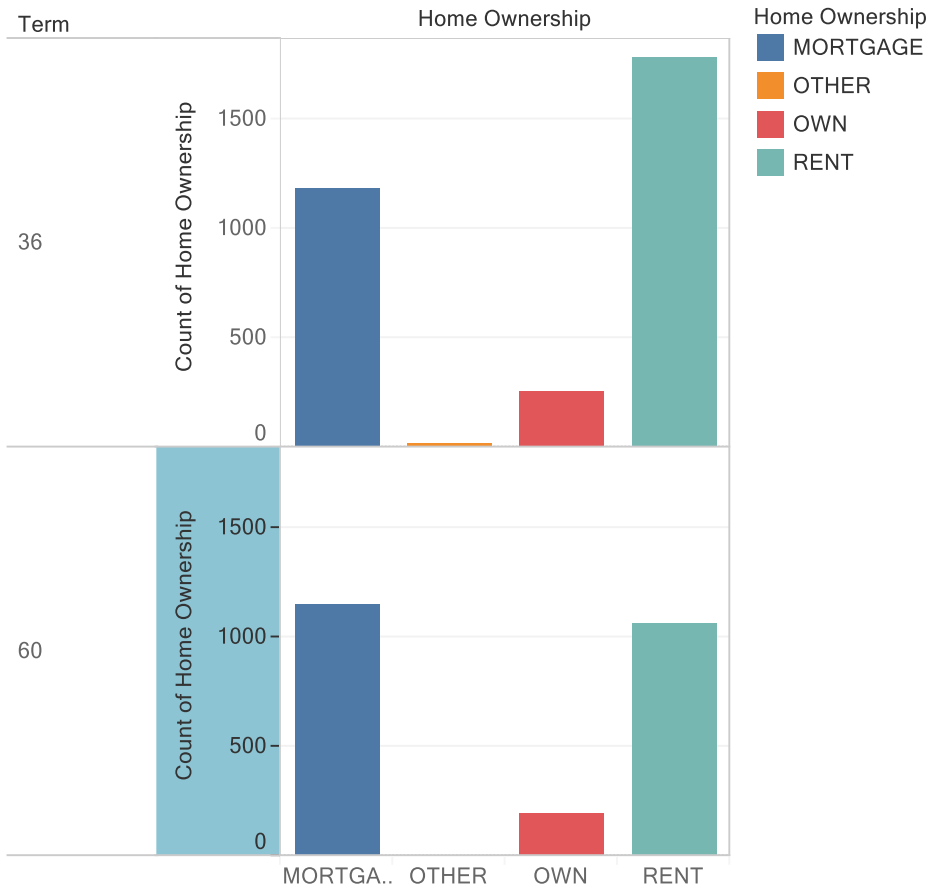
Count of Id for each Addr State broken down by Term. Color shows details about Addr State.

Count of applicants based on purpose and grade



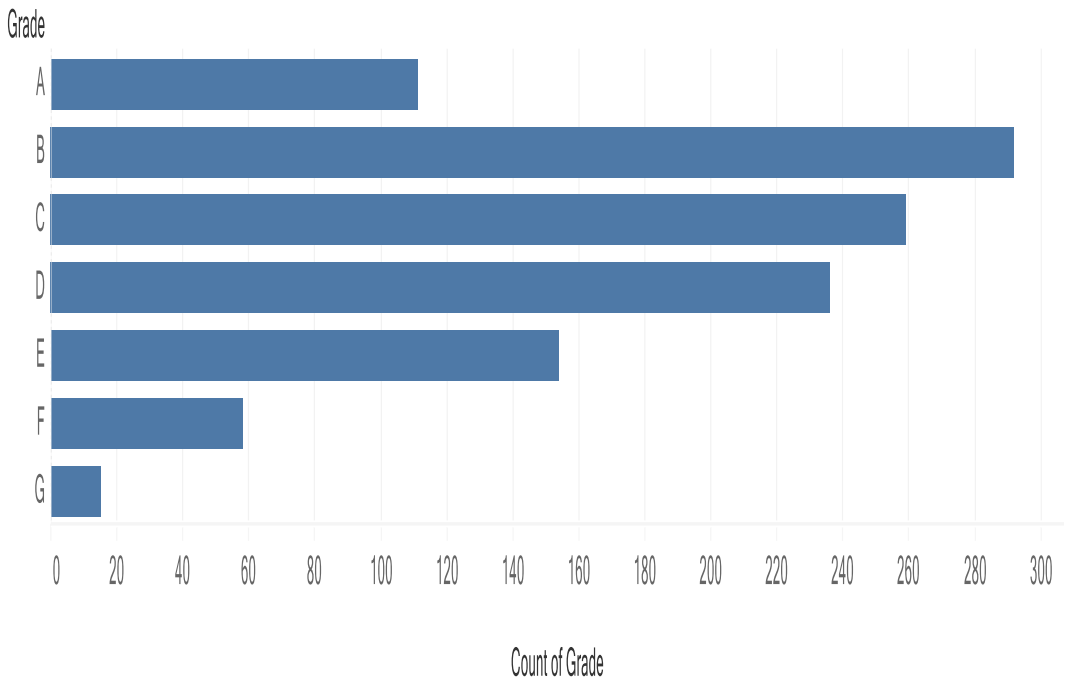
Count of Grade for each Purpose broken down by Grade. Color shows details about Grade.

Term vs House Ownershiop



Count of Home Ownership for each Home Ownership broken down by Term. Color shows details about Home Ownership.

Analysis for State CA



Count of Grade for each Grade. The data is filtered on Addr State, which keeps CA.

Analysis

- People with more than 10+ years of experience or less than 5 years experience are majorly defaulting in their loans with respect to term and grade as well
- We can also say that the regions CA and NY have more number of defaulters
- The loan purpose having maximum defaulters is ‘debt_consolidation’
- Home Ownership as ‘Rent’ has maximum defaulters
- DTI less than 30 and Interest Rate >= 6 are likely to default

Recommendations:

- Annual income, interest rate and verification sets are having strong bearing on the loan fully paid or charged off. So, these parameters or sections were thoroughly analyzed and it have to be monitored.
- Detailed analysis of annual income, funded amount and grades, our group has found that almost all the grades have similar distribution.
- Last minute loan approval or target based loan disbursal to be avoided.
- Income verification process to be rigorous for granting of a loan to control defaults
- DTI should be an important factor while disbursing the loan