EDA CASE STUDY

Understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

By

1. Rajani M
2. Sandesh
3. Vinaykumar V
4. Srikar

About Problem:

1. With this case study, Understanding types of risks associated while accepting the loan applications from customers.
2. The aim is to identify patterns which indicate if a person is likely to default, 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. under following problem conditions:

Two **types of risks** are associated with the bank's decision:

1. If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
2. If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
   - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
   - **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
   - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.).

PROBLEM STATEMENT:

A consumer finance company which specializes in lending various types of loans to urban customers wants to identify loan default patterns . When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

Two types of **risks** are associated with the bank's decision:

- If the **applicant is likely to repay the loan**, then not approving the loan results in a loss of business to the company.
- If the **applicant is not likely to repay the loan**, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions like –

- **denying the loan**
- **reducing the amount of loan**
- **lending (to risky applicants) at a higher interest rate**

PROBLEM SOLVING METHODOLOGY:

1. EDA is used to understand how consumer attributes and loan attributes influence the tendency of default.
2. The dataset used for analysis contains the information about past loan applicants and whether they 'defaulted'
3. Univariate Analysis is used on major consumer and loan attributes.
4. On the basis of Univariate Analysis, we have identified driver variables.
5. Bivariate Analysis is used to find correlation between driver variables.

Assumptions:-

1. There are 54 columns in dataset having all observations as NA. They were excluded from EDA.

2. There are 06 columns having redundant values for all observations . They too were excluded .

3. For this Analysis, we have left outliers and missing values as it was in raw file, and immuted the values in two columns.

ANALYSIS ON VARIABLES:

- Driver Variables:-

Business Insight - Tells about return on funds invested by investors.

- Borrower Attributes:-
  1. Grade & Sub-Grade
  2. Loan Purpose
  3. Annual income
- Loan Attribute :-
  1. Interest Rate
  2. Loan Amount
  3. Loan Payment Term
  4. Public Record & Bankruptcies
  5. Delinquent Status

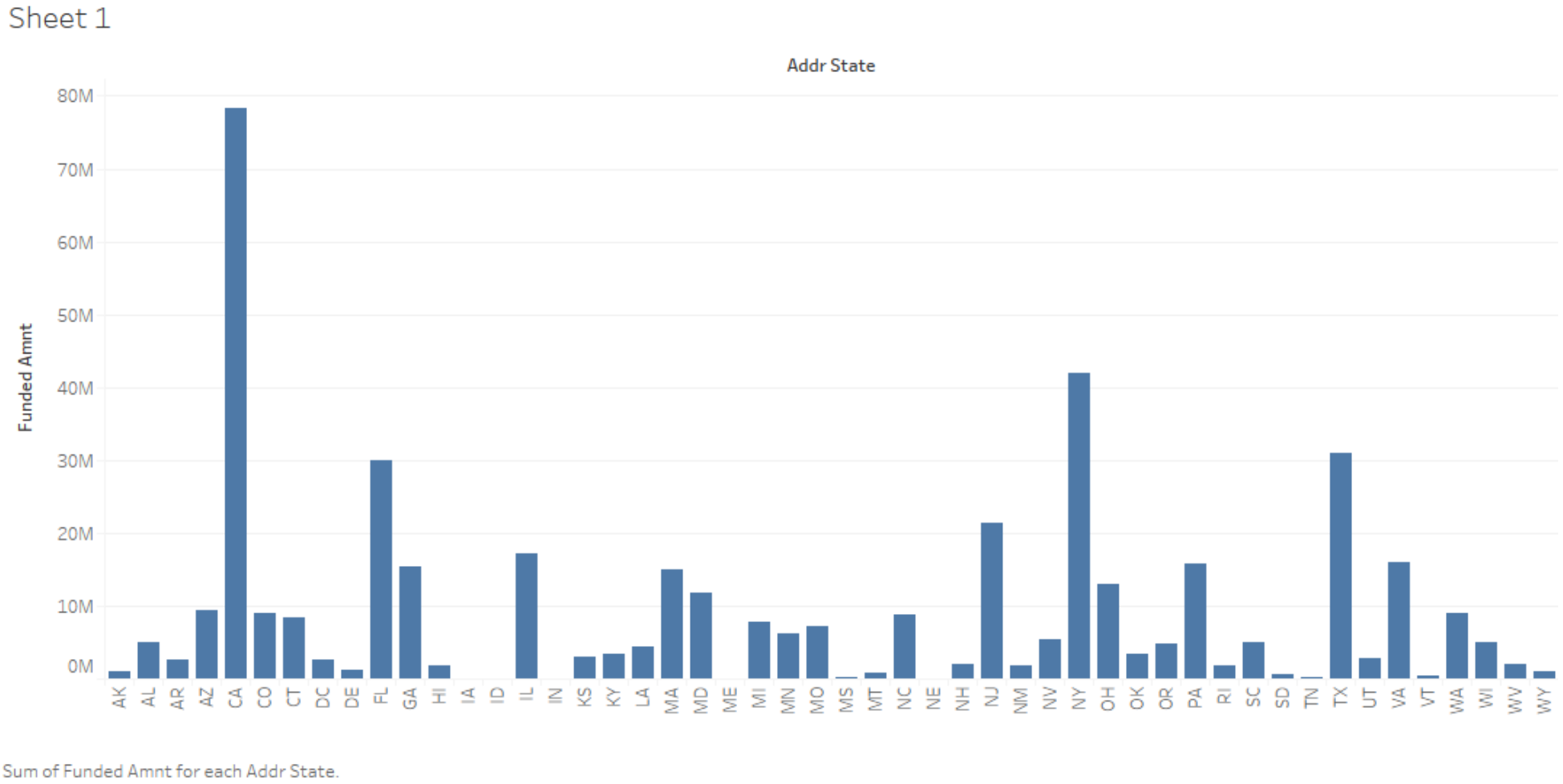
Derived variable:-

1. Return on Investment (ROI)

Note :- DTI is one of the major driver variable for loan approval. The dataset used for EDA have data for approved loans only, So we are not considering it as driver variable for this analysis.

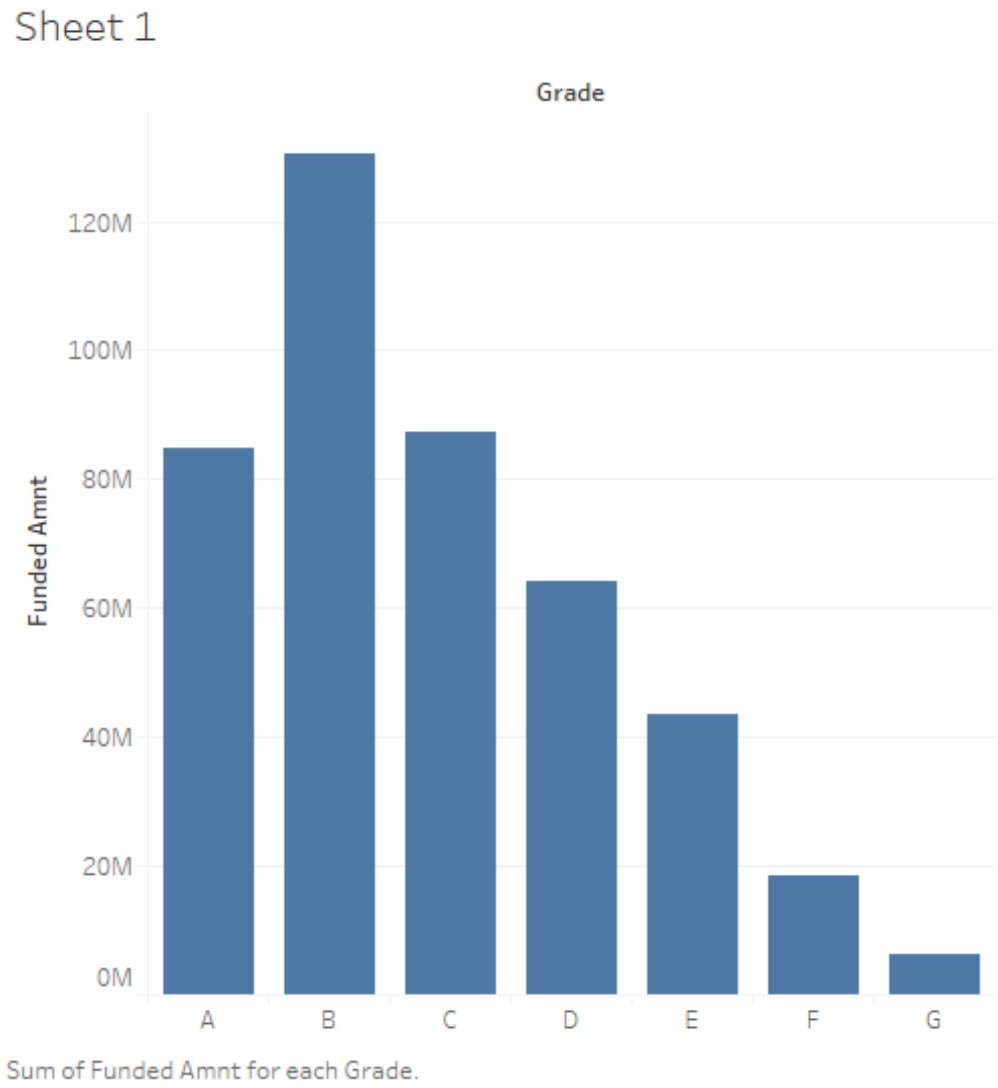
Univariate Analysis

Funded amount Vs State's:



This chart represents the funded amount(in Millions) to the different States. On analyzing we can say that, maximum funding happened in CA(California) and less funding in few states(like IA,ID,MS,TN).

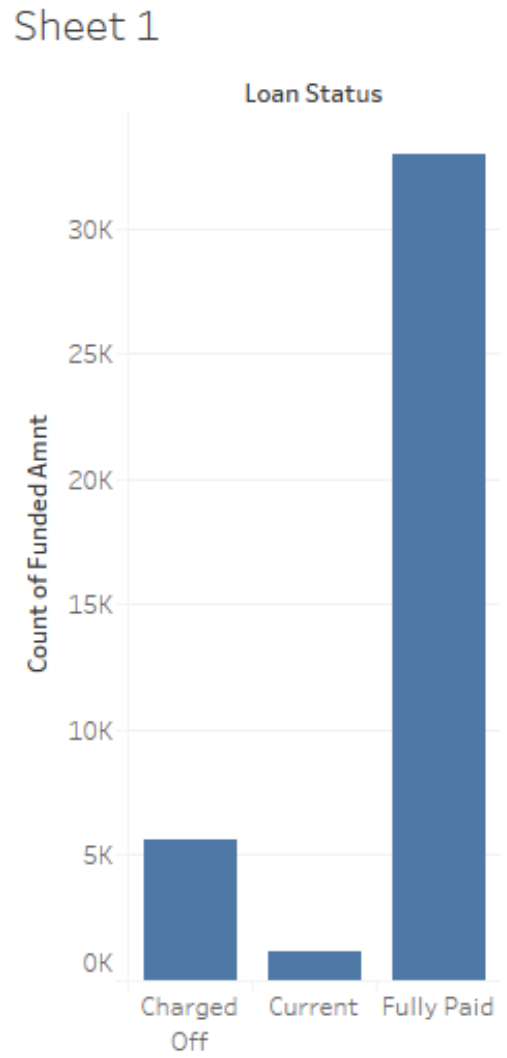
Funded amount Vs Grades:



Each state is classified into different grades. Here we have classified into 7 grades from A to G.

This plotting represents the funded amount across different grades. Based on the graph, it is observed that grade B got maximum funded followed by grade C and A, and the minimum funding happened to G.

LOAN STATUS VS FUNDED AMOUNT :



Count of Funded Amnt for each Loan Status.

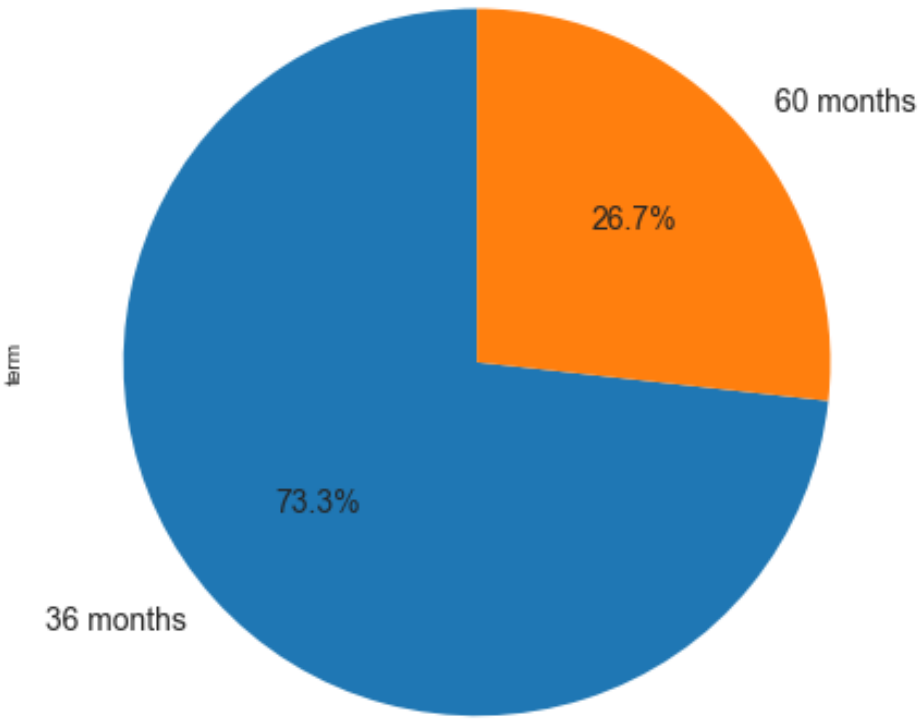
Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. Fully paid: Maximum Applicants has fully paid the loan (the principal and the interest rate)

2. Current: Minimum Applicants is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

3. Charged-off: Average Applicants has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

ABOUT TERMS OF INSTALLMENT:

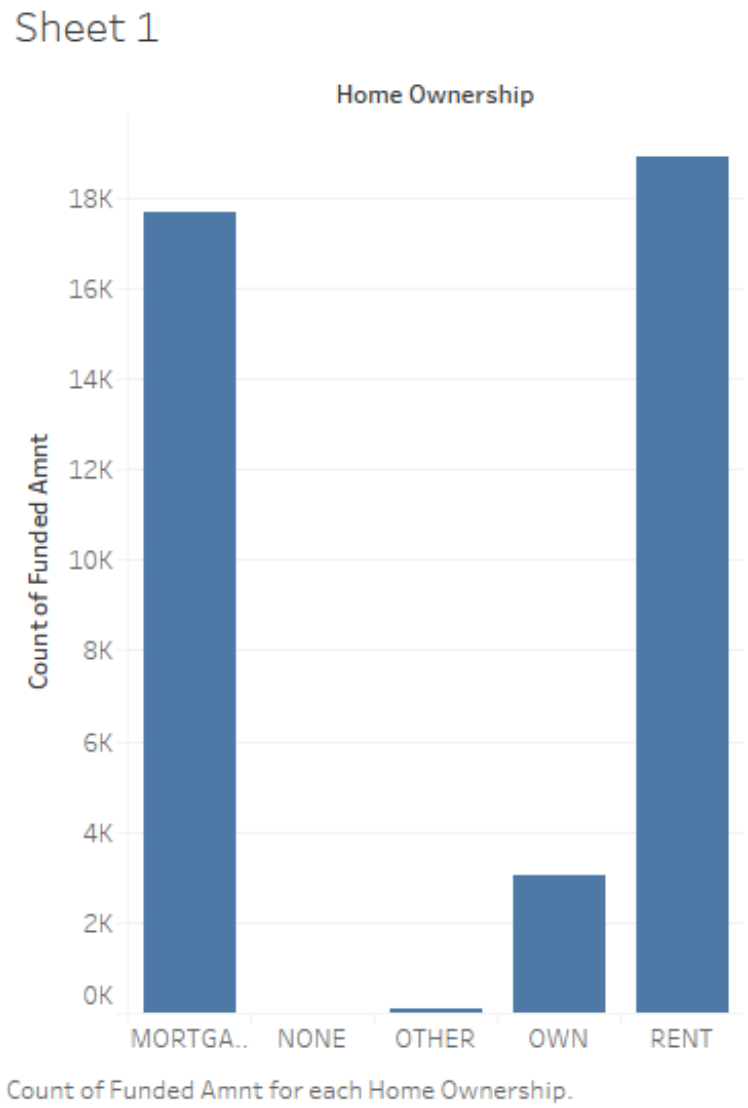


Here, terms of the installments is classified into 2 categories,

1. 36 months
2. 60 months

From the entire applications, 73.3% of applicates opted 36months for installments of the loan amount. And 26.7% applicates chosen 60months term to pay the installments of the loan amount.

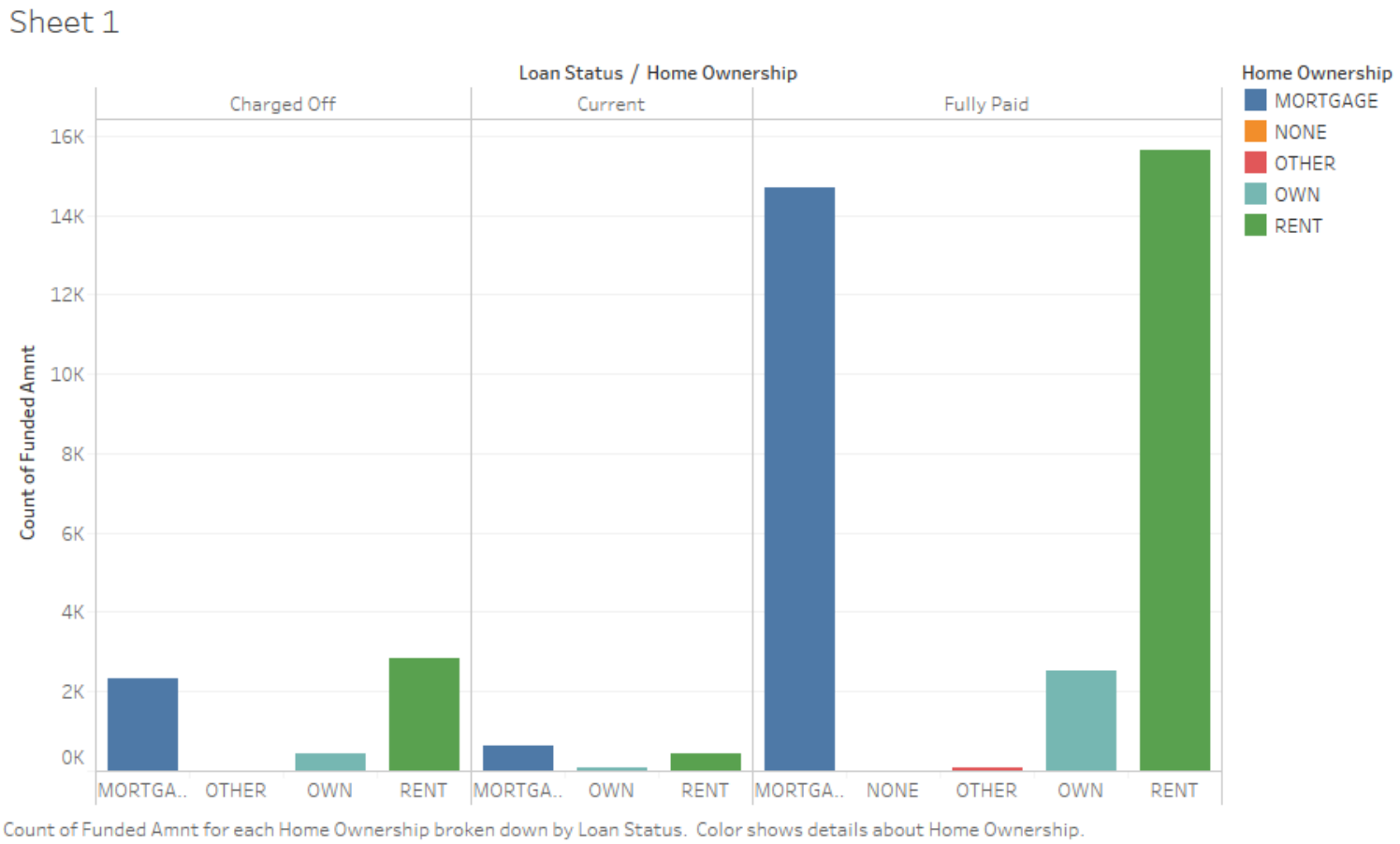
HOME OWNERSHIP VS FUNDED AMOUNT:



Loan amount funded based on the Home ownership, in this chart it showing that for rent/ mortgage applicants got the maximum funding, and those home ownership is none and other applicants got very less funding of loan, and also we can say that own home ownership applicates showed less interest to take loans when compared with rented applicants.

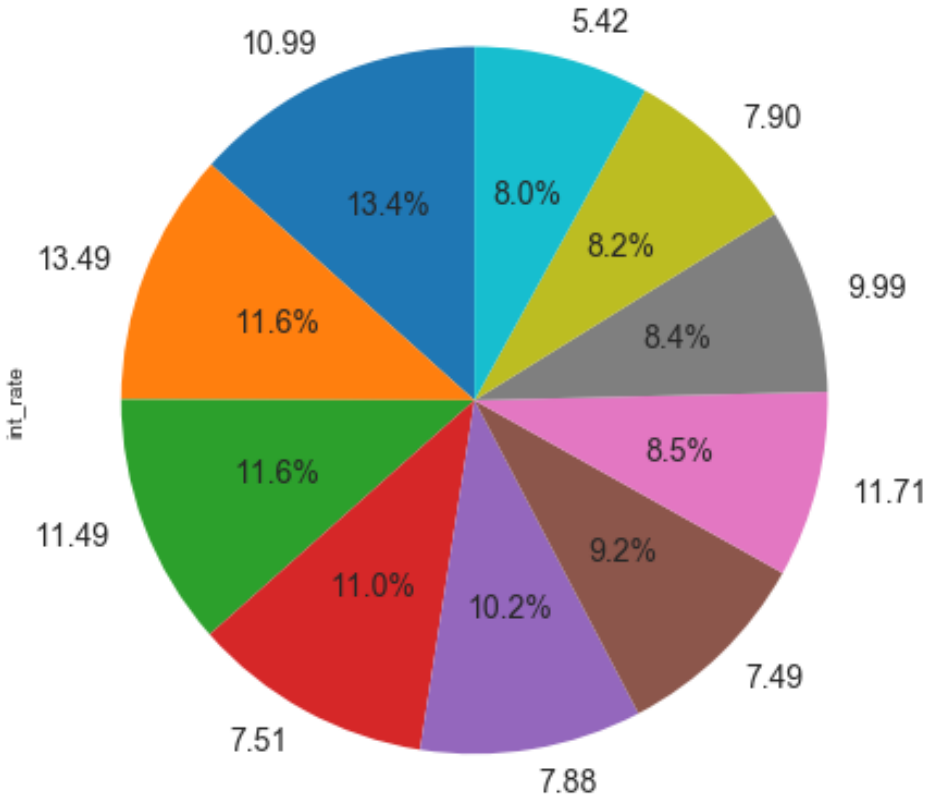
Bivariate Analysis

HOME OWNERSHIP VS LOAN STATUS:



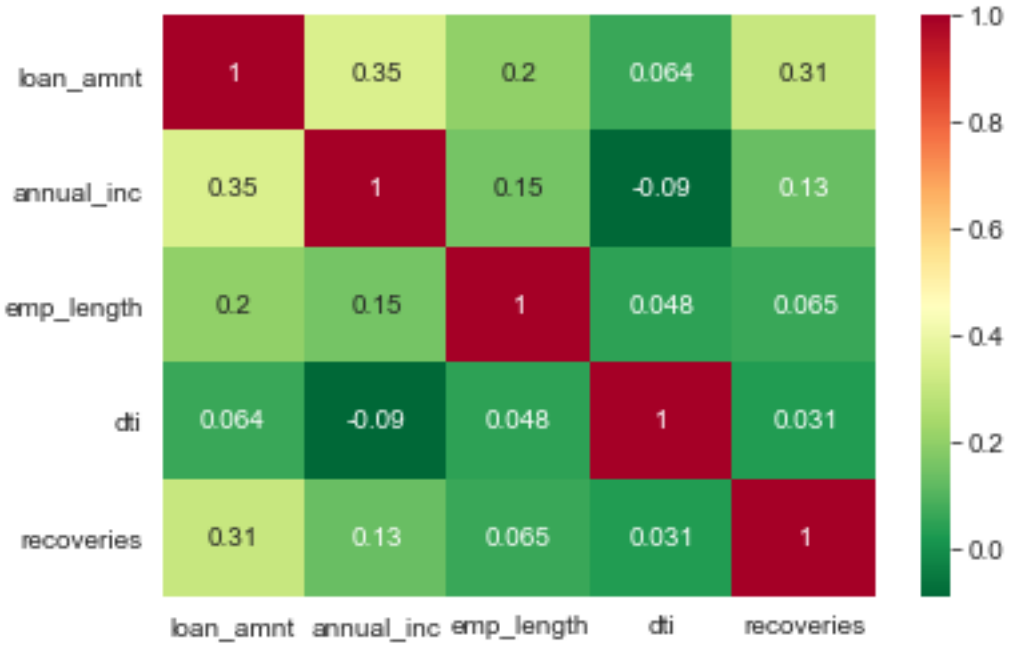
This chart is representation of the home ownership and loan status.
Home ownership is classified into 5 categories and loan status classified into 3 types which is charged off, current, fully paid.
By the observation from charged off loan status, count of applicants who is rent/mortgage almost same and also we can say that there is no funding is happened to rent home applicants.

INTEREST RATE :



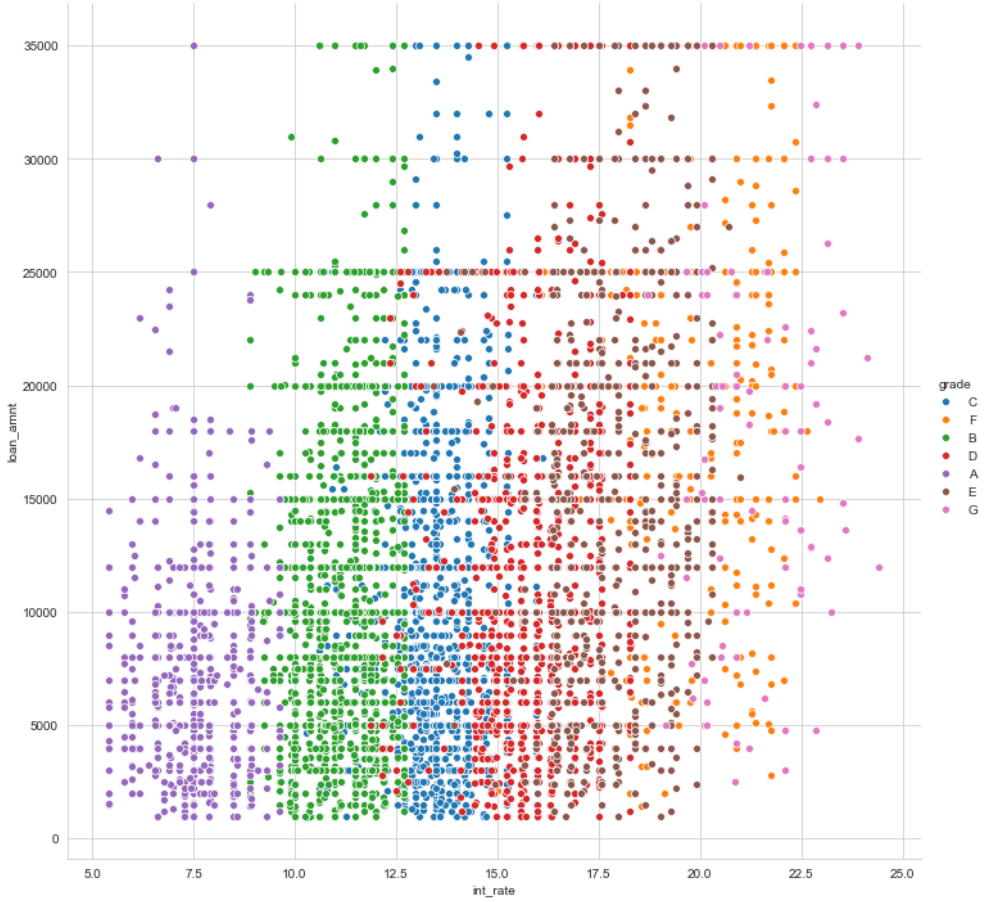
There are different interest rates on the different purpose of loans, by this chart we can say that 11.6% applicants taken loan on 13.49 maximum rate of interest, similarly 8.0% of applicants taken loan on 5.42 rate of interest.

CORRELATION BETWEEN DIFFERENT CATEGORIES :



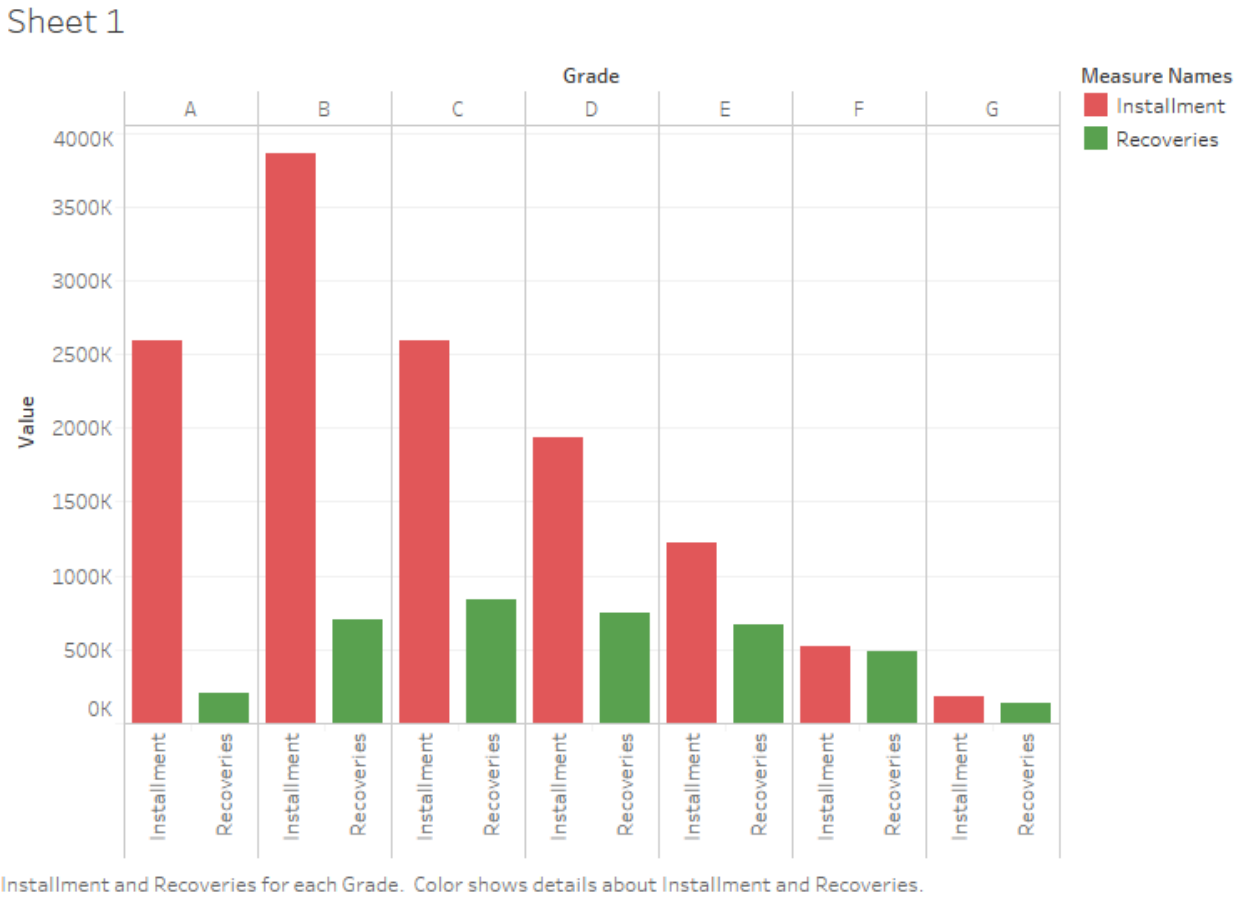
This plot representation for the relationship between the different categories like loan amount, annual income, employee experience length, etc. the relation between two variables called correlation, the correlation between two variables measured as correlation coefficient, the correlation coefficient between two variables always between -1 and 1, if its towards positive 1 then it said to be positive relation, and if its towards negative 1 then there is negative relation is there. By this plot we can say that 35% positive relation between loan amount and annual income.

LOAN AMOUNT VS GRADE:



This scatter plot showing the density of the Loan amount funded to different grades, in this plot each grade is represented with different colours, and each dotted point on this plotting is representation of the loan amount taken by that respective grade.

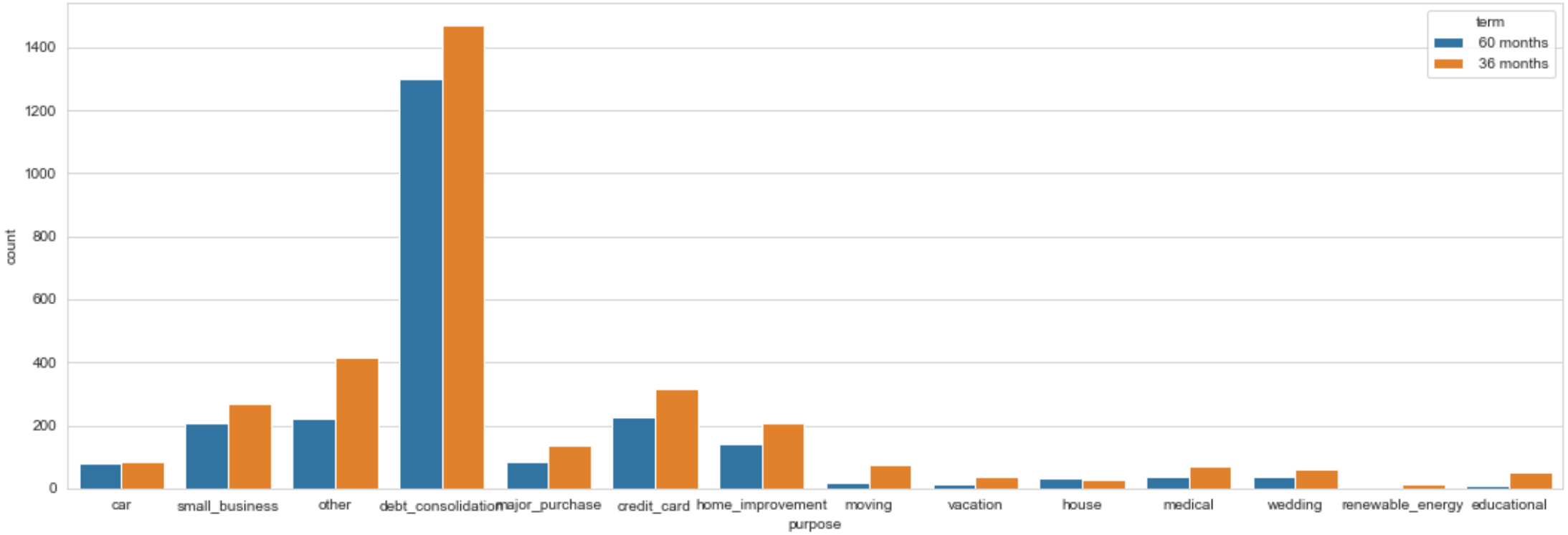
INSTALLMENT VS RECOVERIES:



In each grade red bar represents the installment and green bar represents the recoveries. It is observed that maximum installment is from grade B. In case of grade F, installments and recoveries are almost same, resulting in no loss to the organization. In cases where installment is more and recoveries is less then it leads to the loss to the organization which is applicable to grade A and B.

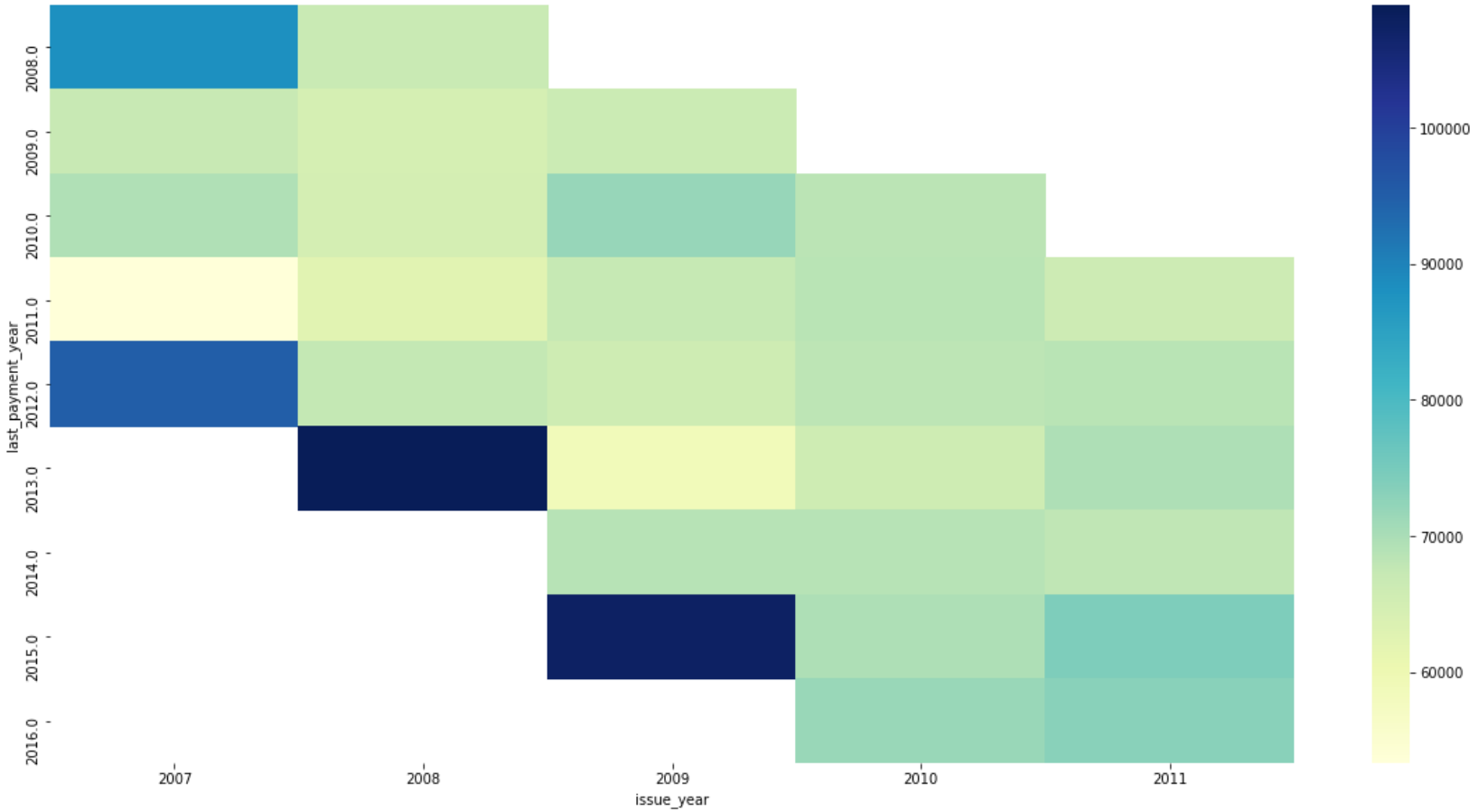
Segmented & Derived matrix - Analysis

PURPOSE OF LOAN VS INSTALLMENT TERMS:



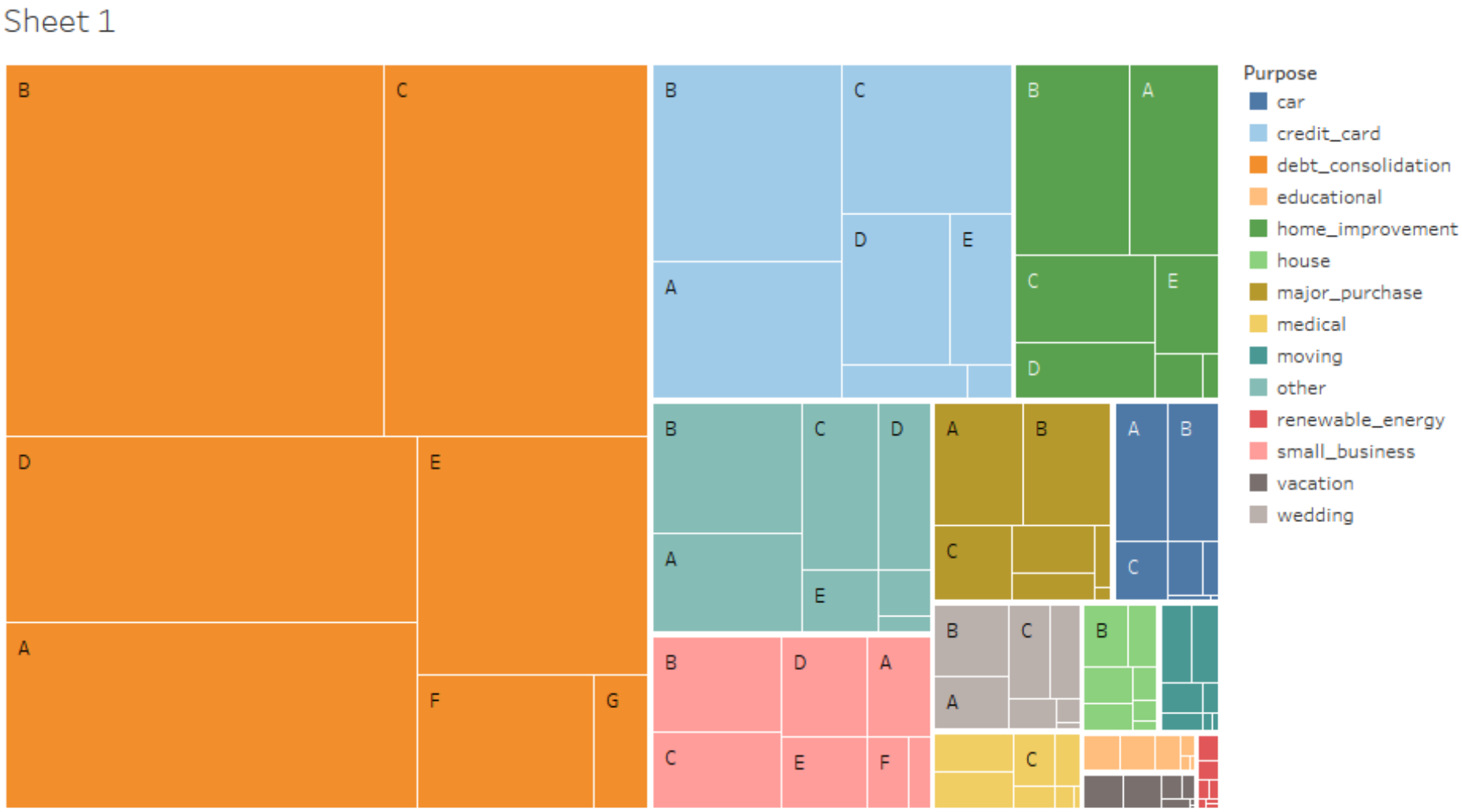
The loan amount is taken by the applicants with different purposes, here x-axis is placed with purpose of the loan, and y-axis representation of the count of the loan amount taken, in this graph blue bar representation of terms which is 60 months, and Orange bar representation of 36 months.

ISSUE YEAR VS LAST PAYMENT YEAR:



This plot showing that, the relationship between issue year of the loan amount and last payment year of the loan amount. This plot showing that light colour to dark colour, where dark colour is placed means that it has maximum funding happened to that respective year. The light shaded box representation for minimum funding is happened.

FUNDED AMOUNT VS PURPOSE/GRADE:



By observation on this plotting we can find that maximum funded amount of loan application for the purpose of debt consolidation from grade B. In this chart each colour represents the purpose of the loan in each grade respectively.

CONCLUSION:

- Lending loans to ‘risky’ applicants is the largest source of financial loss (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.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. We have understand how consumer attributes and loan attributes influence the tendency of default.

- Driver variables (Grade, Subgrade, loan purpose, annual income, loan amount, interest rate, public records, payment terms) are the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

- Accordingly, they can decide on denying the loan, reducing the amount of loan or lending (to risky applicants) at a higher interest rate.

THANK YOU.