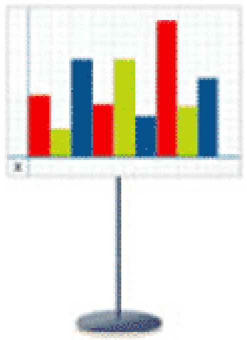
LENDING CLUB LOAN ASSIGNMENT

Name: Dharvi Kumra & Supriya Raman

LENDING CLUB HAVE A PROBLEM STATEMENT !!!

Borrowers apply for loans.
Investors open an account.

Borrowers get funded.
Investors build a portfolio.



Borrowers repay automatically.
Investors earn & reinvest.

Lending Club is a consumer finance company which specializes in lending various types of loans to urban customers. 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

BUSINESS OBJECTIVE

Lending club aims to reduce the credit loss by reviewing and identifying loan applicants data. 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'.

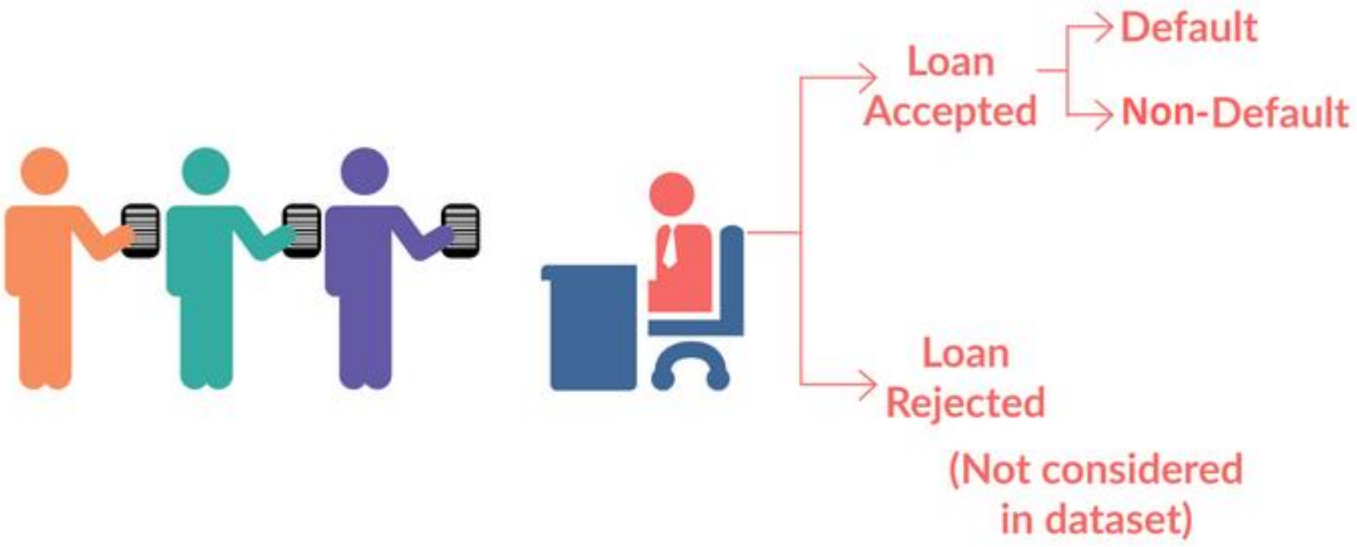
The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

The company want to utilize this knowledge for its portfolio and risk assessment.

DATA FILE

Loan file stores all data observations and all attributes of all loan applicants.

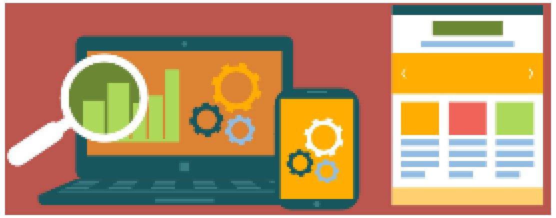
LOAN DATASET



- Comprise all information on all the loans issued by Lending Club

- Information available for each loan comprises of all the details of the loans at the time of the issuance along with more info related to the revised status of loan for example how much principal has been paid until now, the interest, whether the loan was fully paid, or it's defaulted, or if the debtor is late on payments etc.

DATA EXPLORATION AND ANALYSIS



Objective: To identify all driver variables (attributes) which should be considered to identify the loan default.

The analysis is divided in following business steps –

1. **Data Understanding**
2. **Data Cleaning**
3. **Data Analysis**
4. **Univariate Analysis**
5. **Bivariate Analysis**
6. **Conclusion and Analysis Summary**

DATA UNDERSTANDING AND CLEANING

Following steps were taken to understand and clean the data for further analysis -

1. Input data has 39717 observations.
2. Input data has 111 attributes.
3. Data has 54 attributes which has missing data in all observations. These attributes can be dropped.
4. Two attributes with more than 90% missing data were dropped and certain other attributes not fit for analysis were discarded.
5. Remaining 17 attributes were identified for further data analysis out of which id is not considered and loan_status is found to be used as target variable.
6. The data file has no duplicate observations. No row was deleted.

DATA ANALYSIS

Analyze data observations and remaining attributes to find out suitable attributes which can be predicators of default -

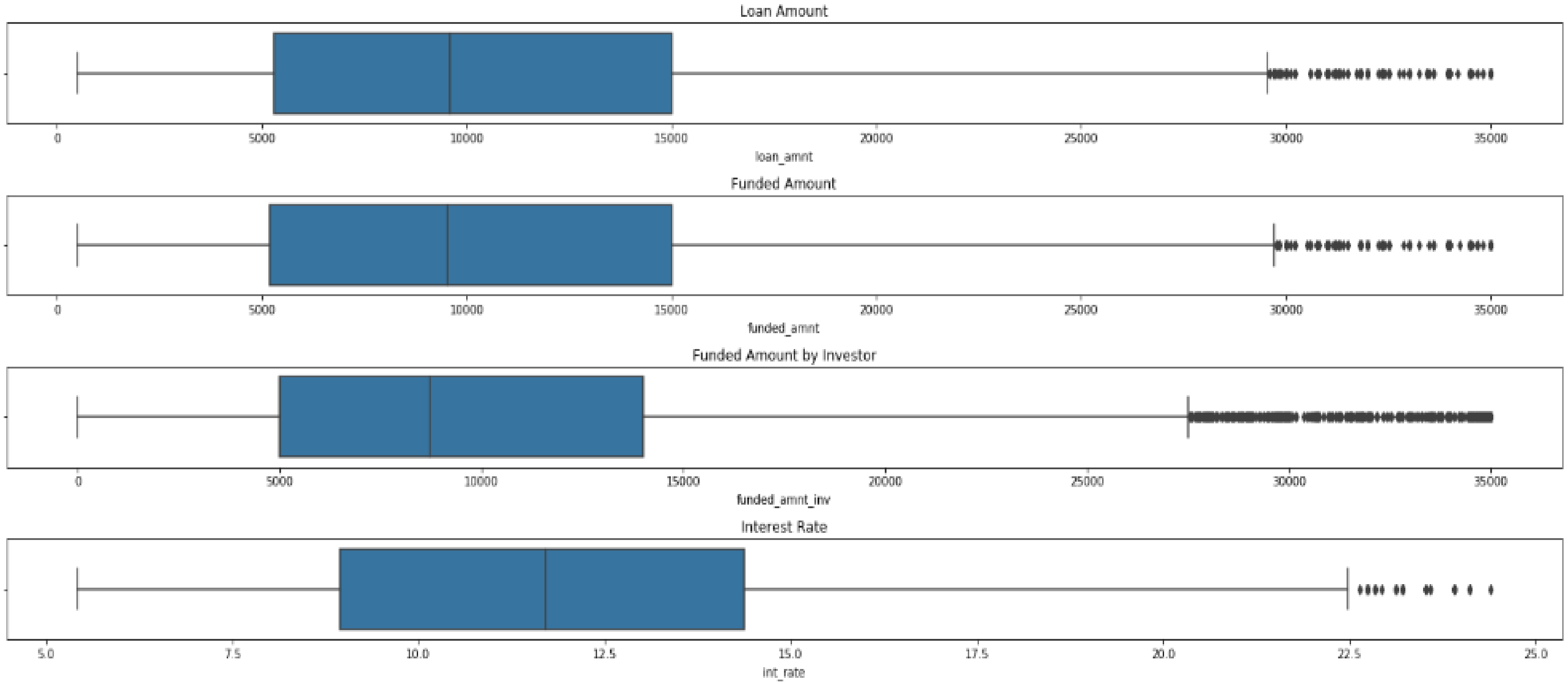
1. Nine columns were found to have only one unique value in all observations. These attributes were dropped from further analysis.
2. By referring the metadata of the input file in Data Dictionary, it was found to have three types of attributes
   - Applicant related attributes like age, occupation etc.
   - Loan related attributes
   - Customer Behavior related attributes which was observed post loan approval.
3. Only loan related attribute was found to be suitable for further analysis and hence all other attributes were dropped. This left us with only 18 attributes for further analysis.
4. Loan status field is found to have three types of loan - Fully Paid, Charged Off and Current. "Current" loans were not found applicable for further analysis of Loan Defaults and hence those observations are dropped. "Fully Paid" and "Charged Off" loans are hot encoded with 0 and 1 respectively for the ease of further analysis.

UNIVARIATE ANALYSIS

In univariate analysis we are trying to find out if the chosen attributes have any outliers and what is the count of people in each of those attribute's category.

<u>These are the results for various attributes</u> : -

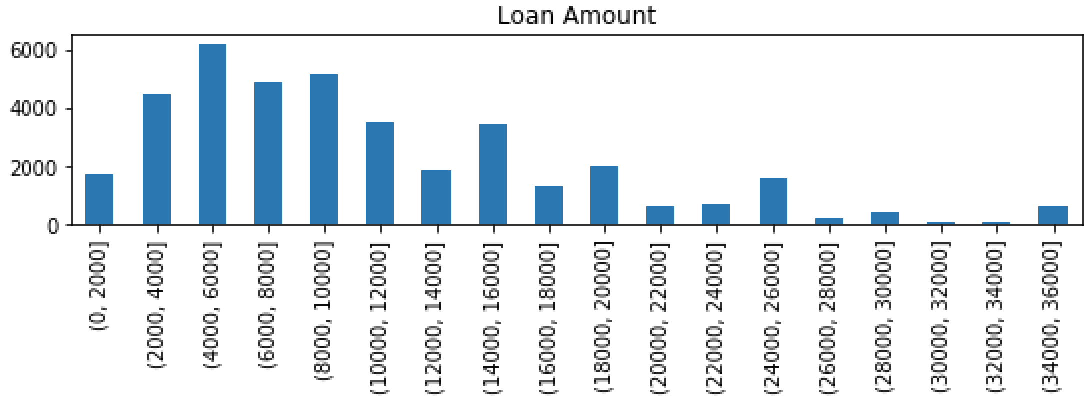
"loan_amnt", "funded_amnt", "funded_amnt_inv" and "int_rate" : -

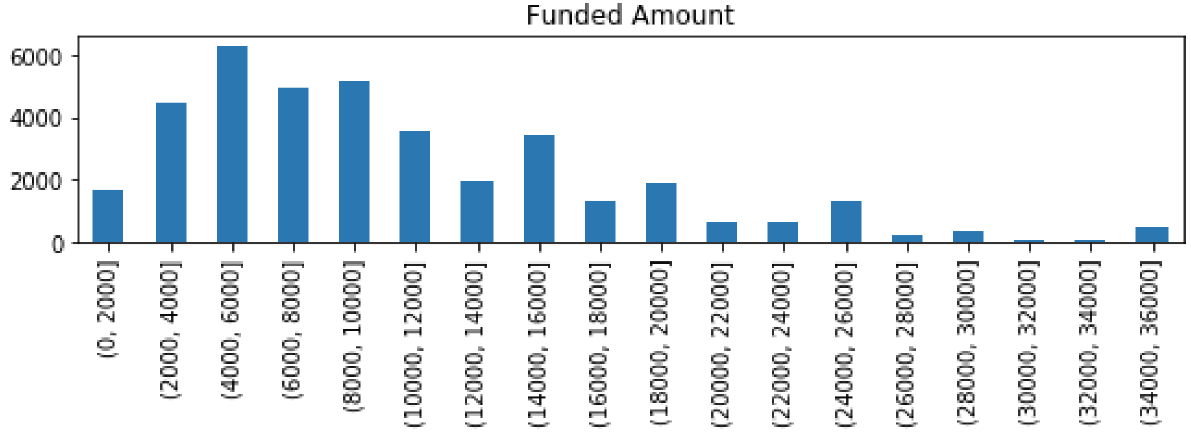


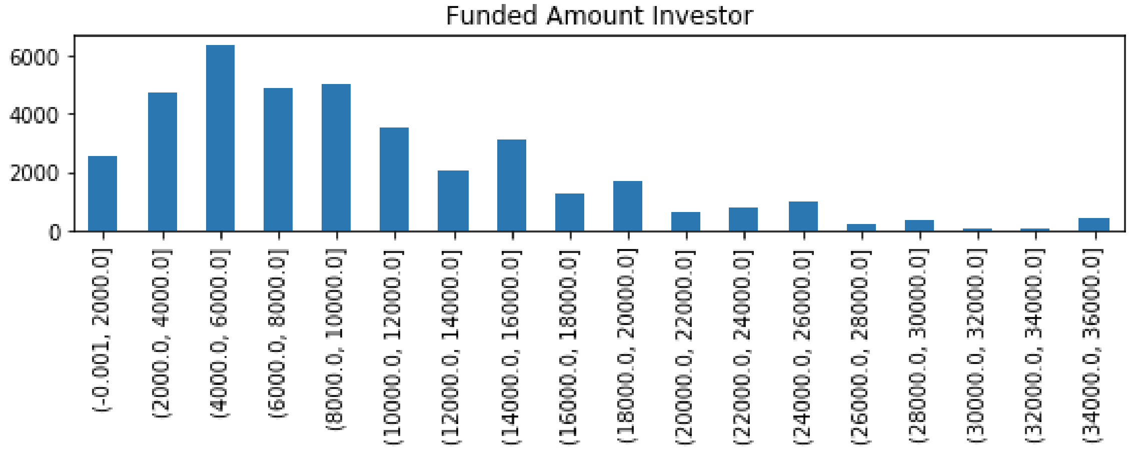
These fields are found to have outliers hence which are huge in no. Hence can't be discarded because this will lead to data loss .

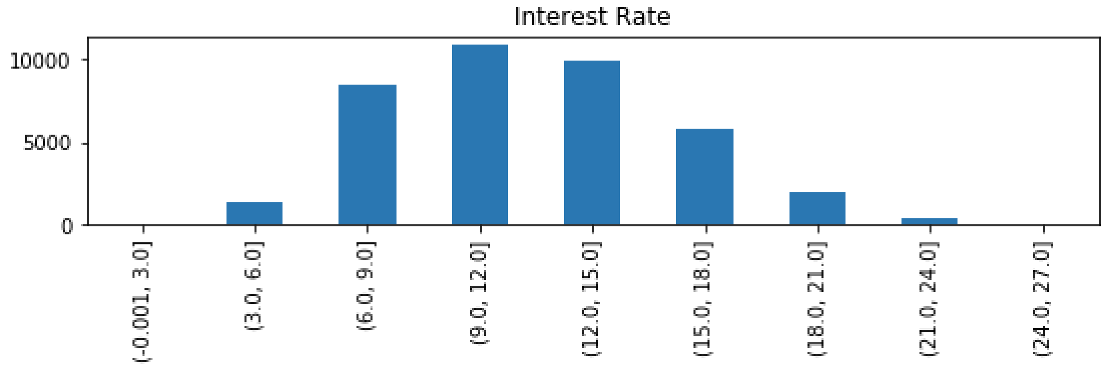
And for the NaN values we have imputed them with the medians of the respective fields.

"loan_amnt", "funded_amnt", "funded_amnt_inv" and "int_rate" : -


Loan Amount

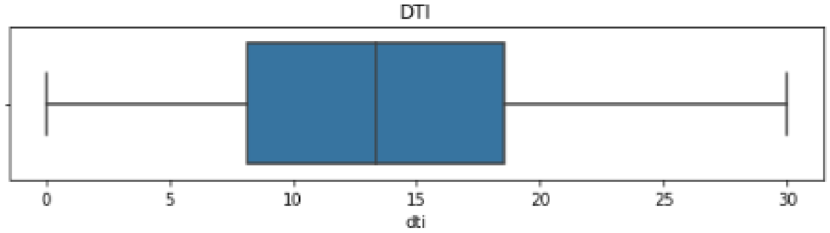

Funded Amount

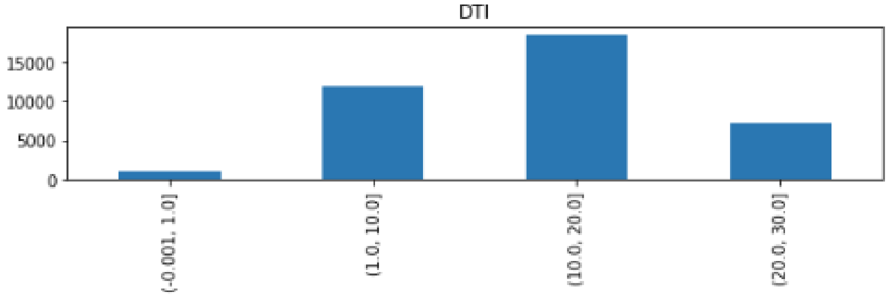

Funded Amount Investor


Interest Rate

Loan amnt, Funded amnt and Funded amnt inv have same values which specifies that usually the people whose loans are passed get exact amount as they demand. In the interest rate graph we can see that most of the loans are passed on interest rate (9 to 12]

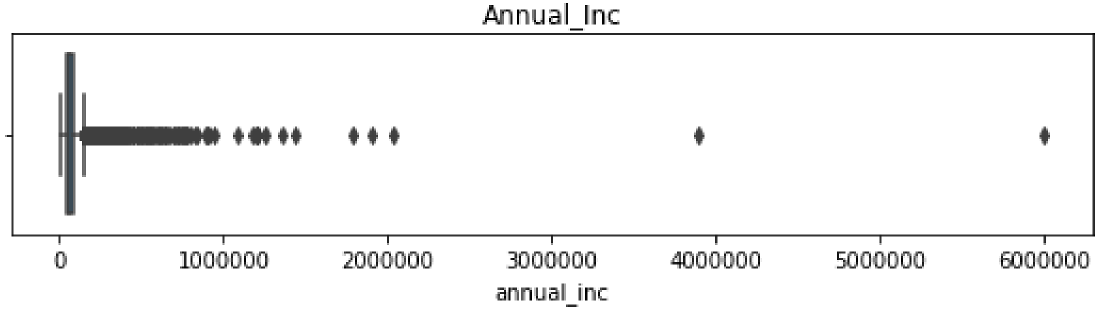
DtI(debt to income ratio)

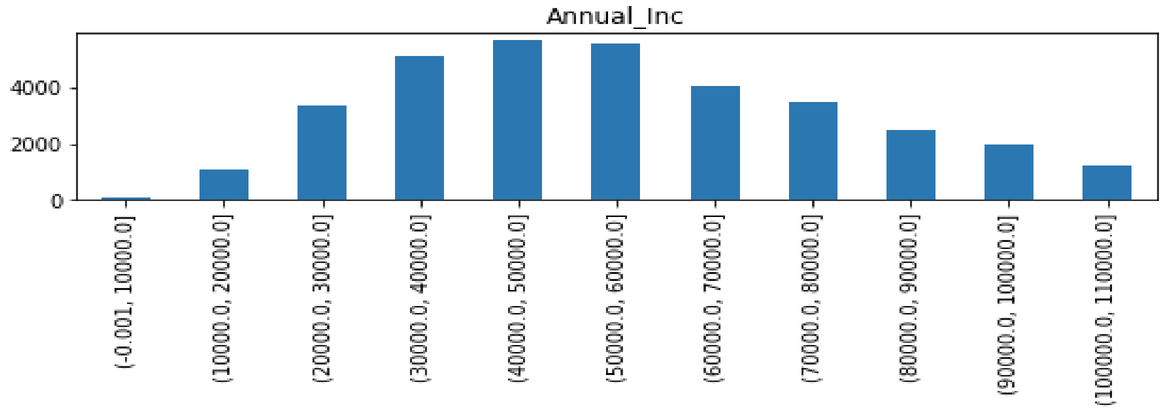




DTI is found to be uniformely distributed and there are huge no. of people who have dti in range (10.0 to 20.]

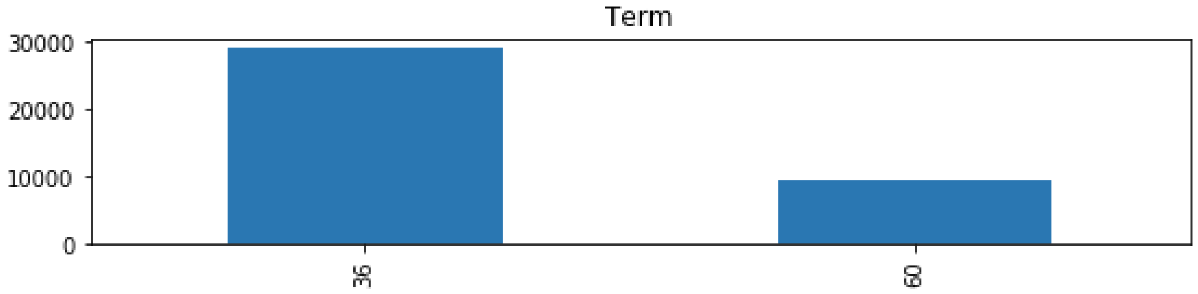
Annual_inc





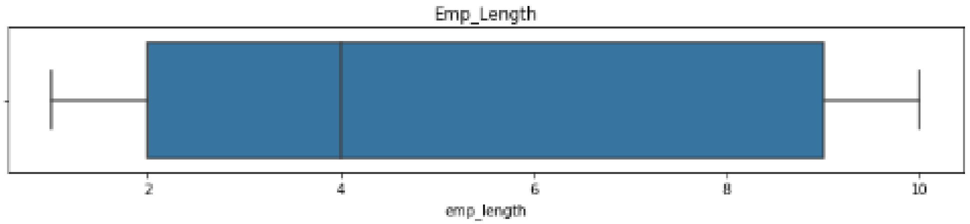
Annual Income is found to have outliers hence NaN values are imputed using median and there are huge no. of people who have annual_income in range (40000 usd to 50000 usd)

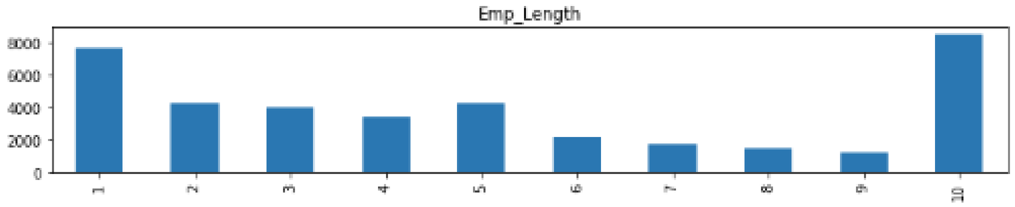
Term



usually of the loans are for 36 months term

Emp_length

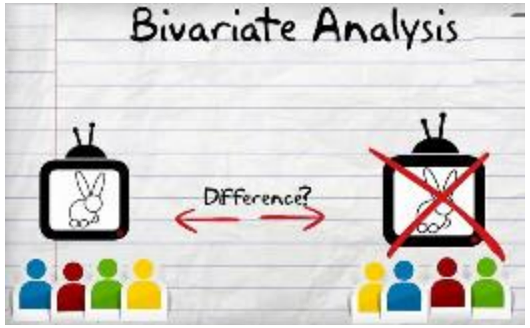




The people who usually apply for loan are either having emp_length 1 year or approx. 10 years

Other variables being considered are Grade and Subgrade, state of the person, home_ownership , purpose, installment, pub_rec_bankruptsy

BIVARIATE ANALYSIS

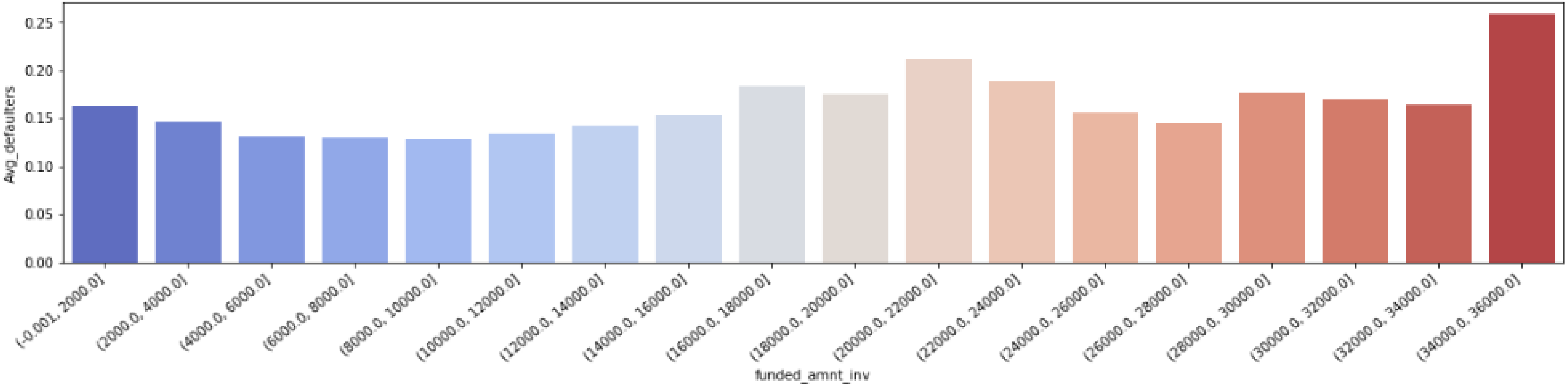


In bivariate analysis we are trying to find out how the chosen attributes are found to be related to target variable which in our case is loan_status because we want to find out what are the factors that are leading to default.

These are the results for various attributes : -

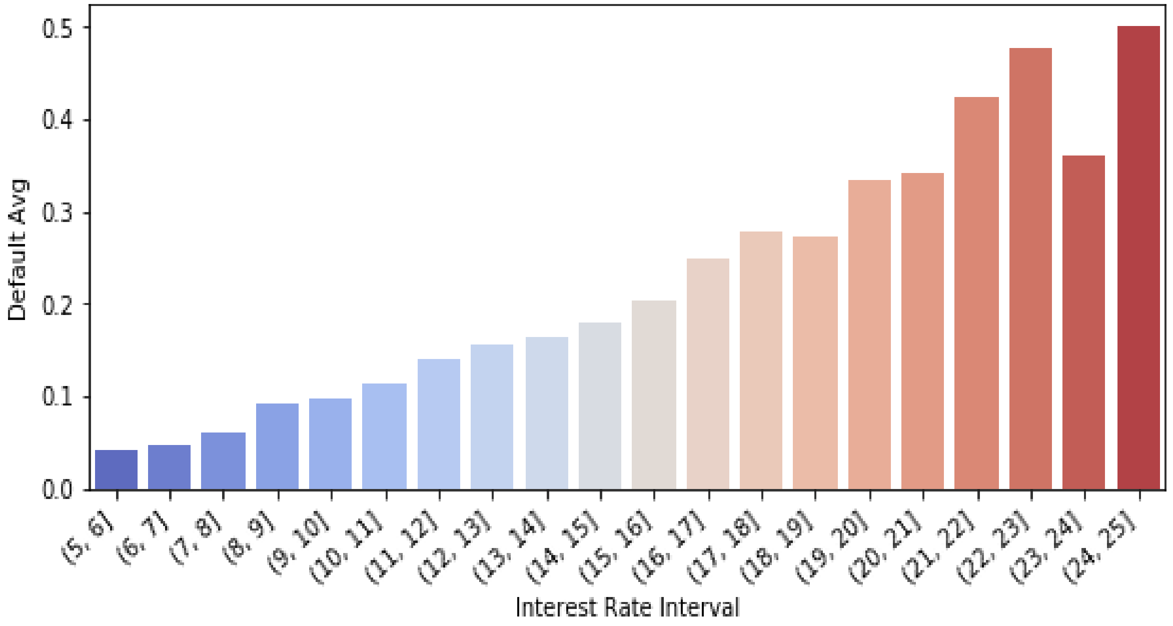
"loan_amnt", "funded_amnt", "funded_amnt_inv" :-

The graphs for these three variables are approx. Same and lead to the inference that people who demand high loan amounts are probably causing more defaults than others



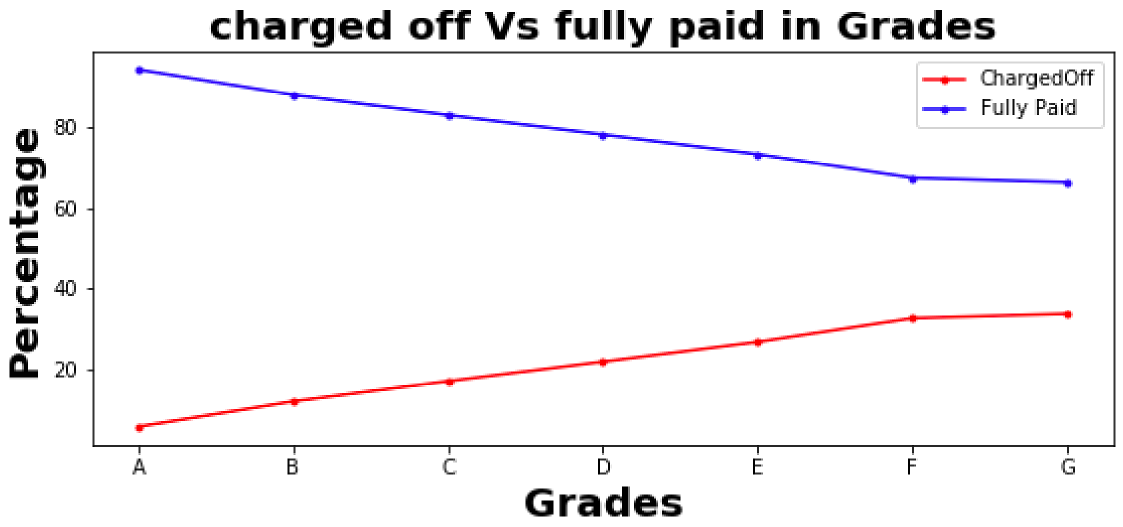
"int_rate" :-

The graph shows that as the interest rate increases people start causing defaults



"grades" :-

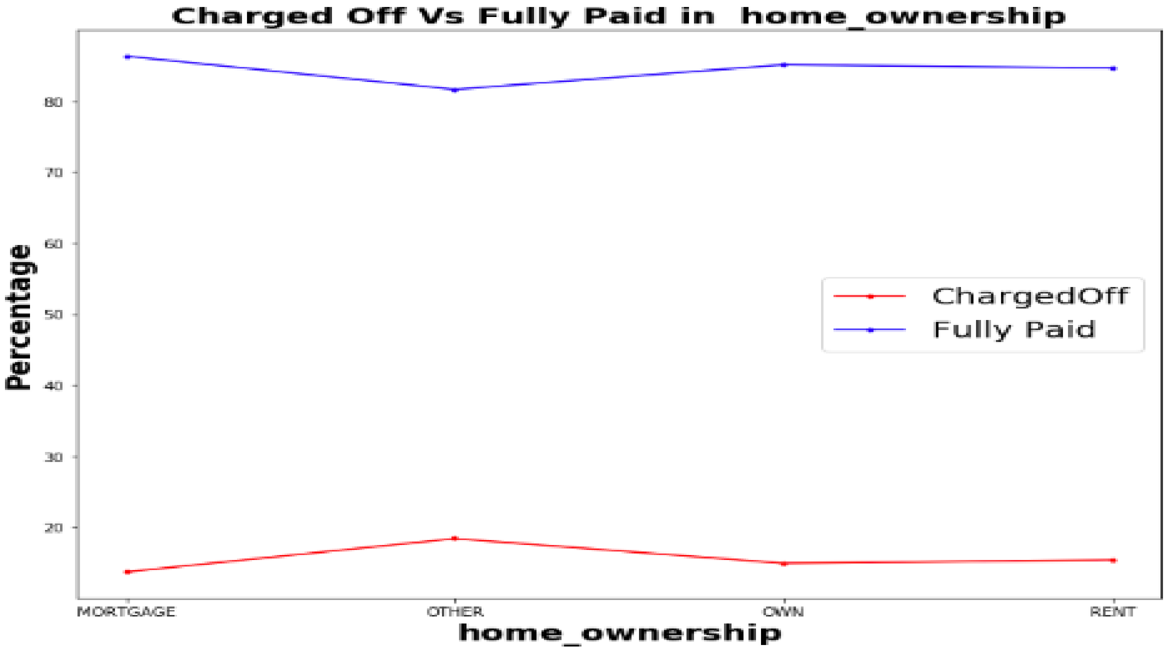
The graph shows that lower the grade higher the probability of default by that person



This implies that higher interest rates and lower grades are a driving factor of default.

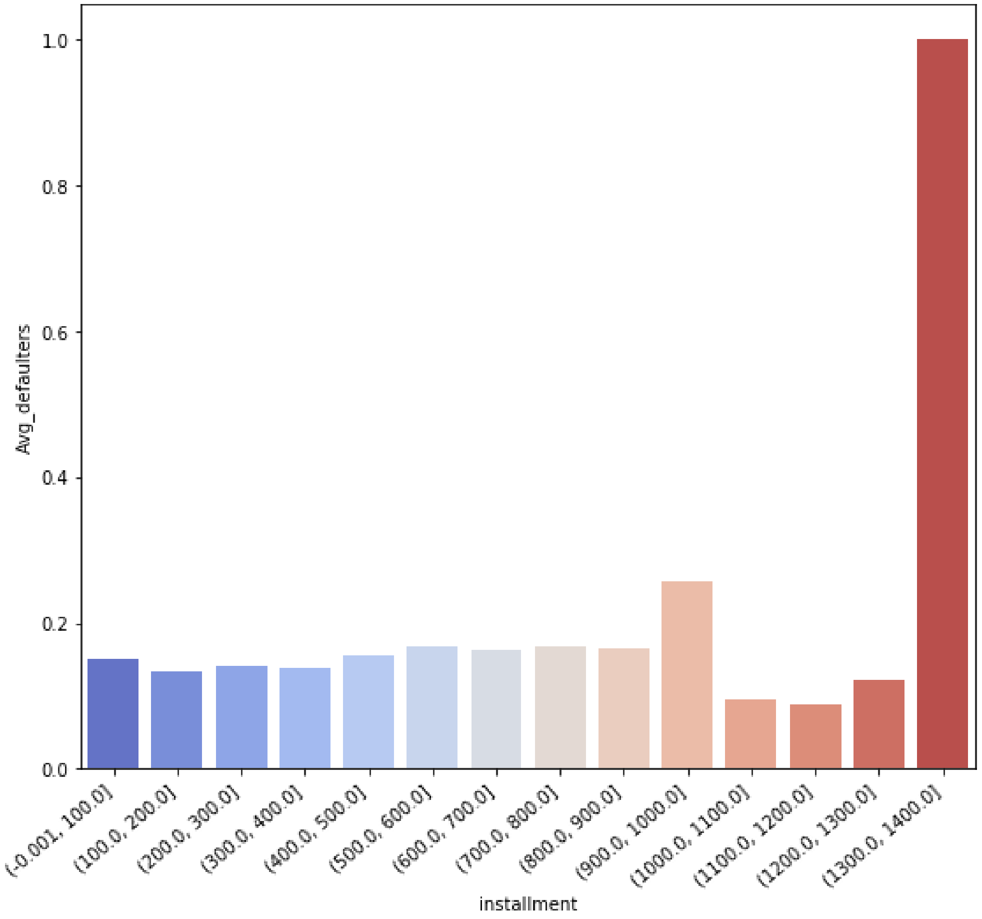
"home_ownership" :-

The graph shows that people who have "other" as accomodation , chances are high they will cause default .



"installment" :-

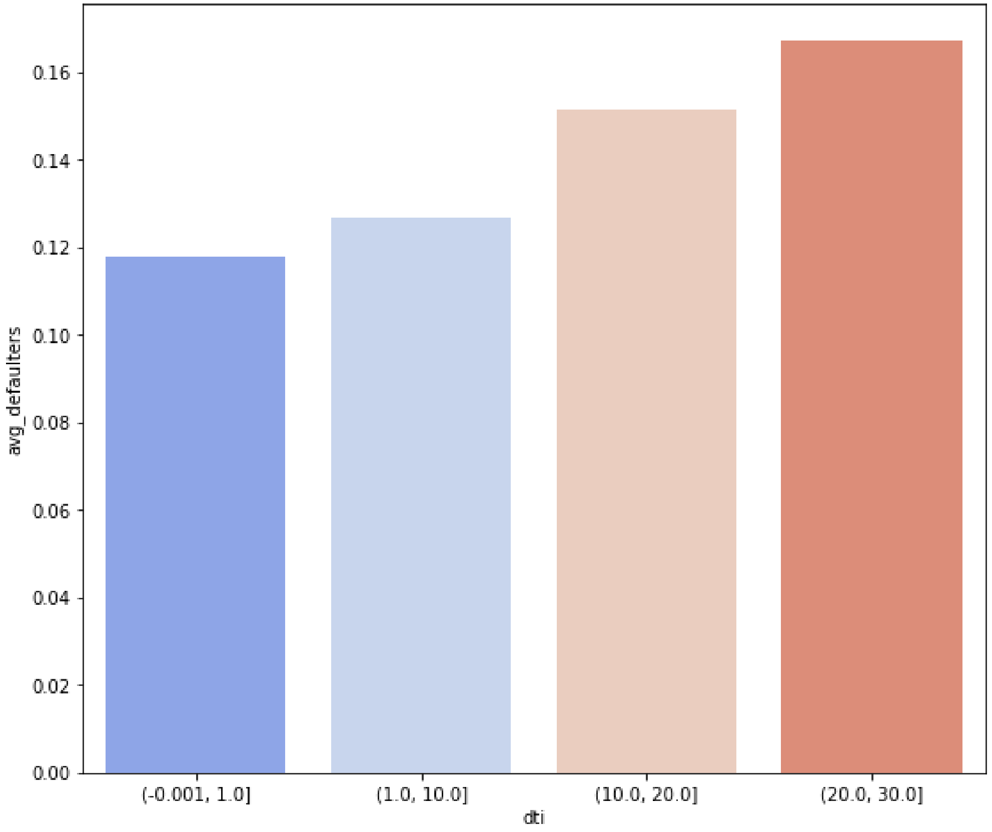
The graph shows that higher the installment amount higher are the chances people will default



Hence high installment amounts and "OTHER" accomodation are driving factor of default

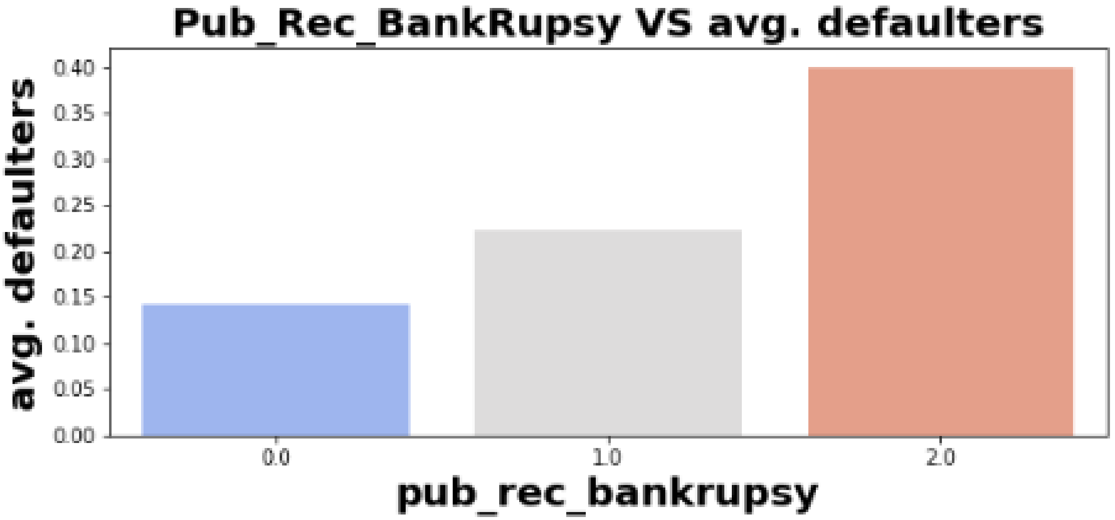
"DTI" :-

The graph shows that higher the debt to income ratio chances are high that person will be charged off



"Public Record of bankruptcies" :-

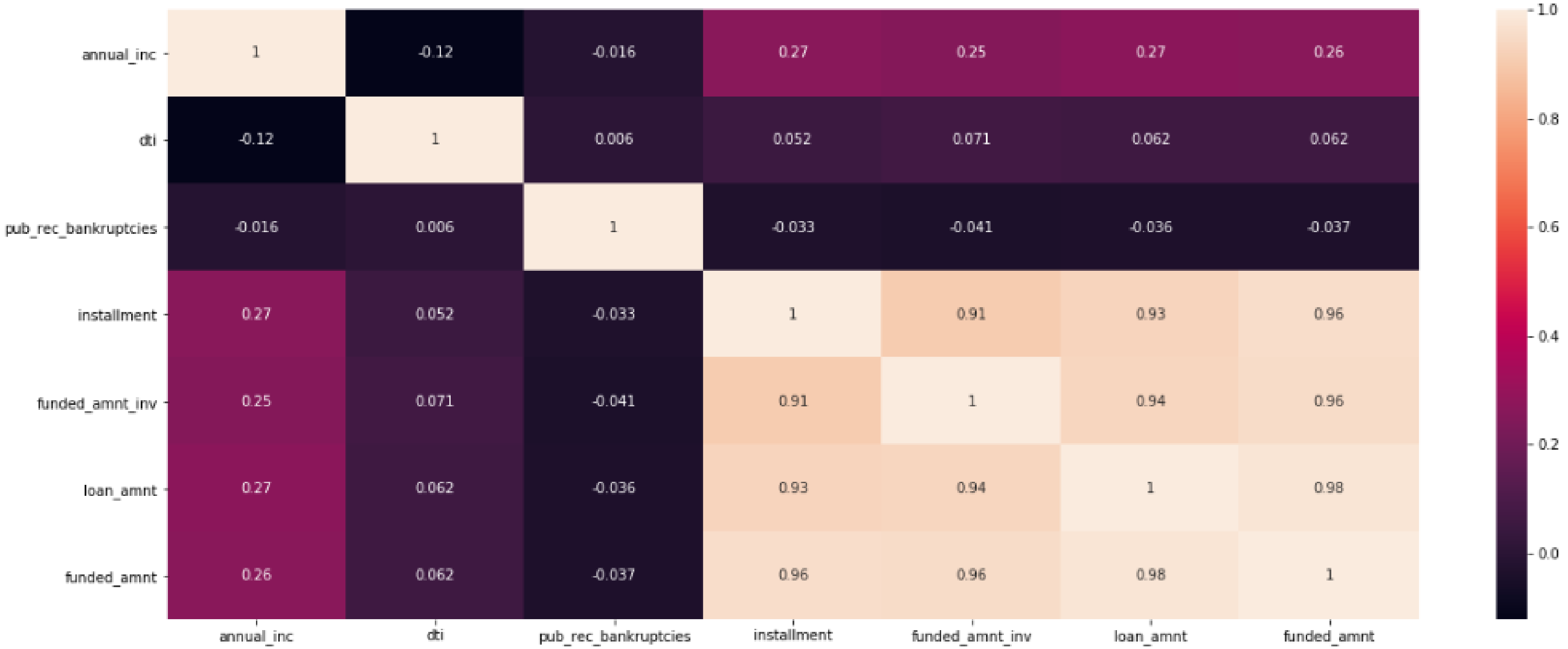
The graph shows that higher the record point (implies large no. Of bankruptcies) higher is the chance of that person being charged off



Hence high dti and high bankruptcy point are driving factor of default

Other variables being considered in the analysis are address state of the person in Nebaraska people are found to have high chances of being charged off as defaulter and sub grade lower the subgrade higher the chances of default

CORRELATION BETWEEN DIFFERENT VARIABLES



Inference: Correlation matrix was generated for all existing numerical attributes in the dataset to find how strongly the attributes are related. So, obtained correlation co-efficients are plotted in the heatmap and pairplot below. The heatmap and pairplot shows the following -

1. Loan Amount (loan_amnt) and the Funded Amount (funded_amnt) are highly co-related attributes where Correlation Co-efficient is 0.98
2. Other strongly co-related attributes are funded amount provided by investors (funded_amnt_inv)and instalment amount(The monthly payment owed by the borrower if the loan originates)
3. dti (debt to income ratio) is found strongly correlated to all the above attributes.

ANALYSIS SUMMARY

From the analysis, following loan attributes are found to be drivers or play a key role in identification of strong indicators of default.

- **GRADE - lower the grade higher the chances of default**
- **INTEREST RATES – Higher the rate of interest higher the chances of default**
- **LOAN PURPOSE - People who apply for small businesses are the ones who usually cause defaults**
- **ANNUAL INCOME – Lower the annual income chances are high that the person will cause default**
- **DEBT TO INCOME RATIO - People with high debt to income ratio usually cause more defaults than others**

- **HOME OWNERSHIP – People who put their accomodation as "OTHER" are tend to cause higher defaults than others**
- **INSTALLMENTS – Higher installments are also a driving factor causing a person to be charged off**
- **FUNDING AMOUNT INVESTOR , LOAN AMOUNT predict that people who apply for higher amounts of loan tend to default more.**
- **PUBLIC RECORD OF BANKRUPTCY – People who have high point of previous bankruptcies are more likely to cause defaults**

SUGGESTION: Lending Club should use a good model to assign grades to people because it was found that people with a good grade that is A or B are also found to have defaulted.