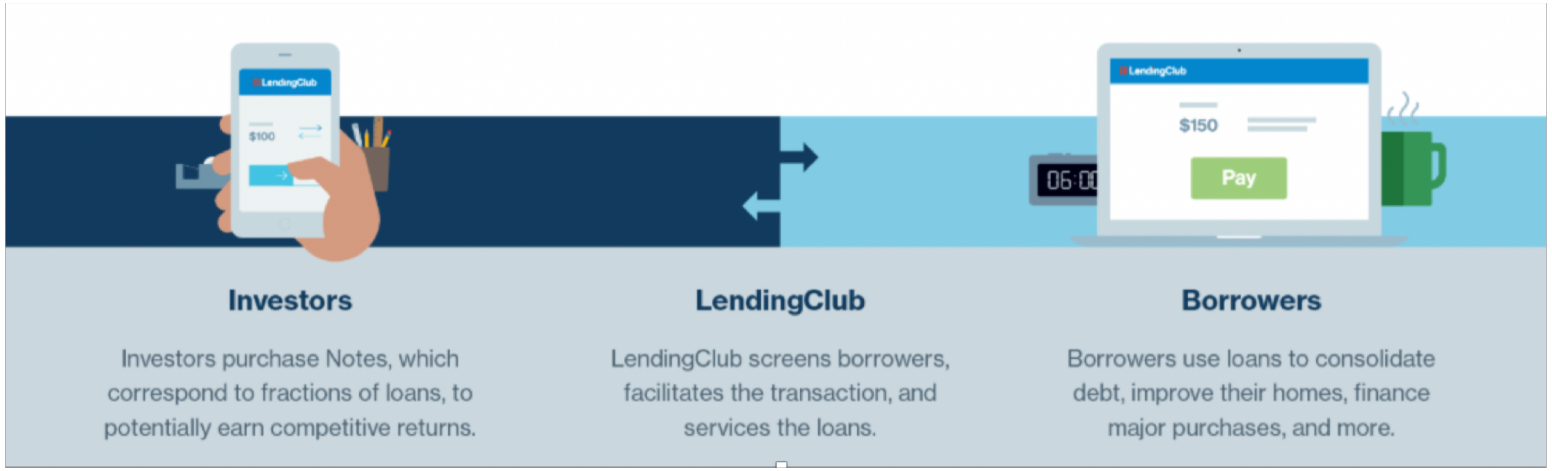


LENDING CLUB CASE GROUP STUDY

SUBMISSION

Name:

- ADYASHA RATH(adyasha016@gmail.com)
- BINNY GARG(binny.garg9@gmail.com)

Abstract

Business Objective:

Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. The objective is to identify risky applicants who can cause financial loss (called credit loss) to the company. To identify the risky applicants it is needed to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Goals of Data Analysis:

1. Univariate Analysis
2. Bivariate Analysis

Problem solving methodology

- A step-by-step approach has been followed as shown in the below diagram to find the variables that are strong indicators of Loan Default borrowers.
- The motive is to identify these risky loan applicants, so that such loans can be reduced thereby cutting down the amount of credit loss to Lending Club.

Note: All the constraints of Lending Club case study has been kept in mind throughout the data analysis,

- Only **"Fully Paid" & "Charged Off"** loan status is preferred in our analysis**.**
- Loan data for analysis is taken from **loan.csv** as provided**.**
- Also, we have tag **"Fully Paid" & "Charged Off"** as numeric values 0 and 1 correspondingly to make our analysis simple and clean.

Data Sourcing & Data Understanding

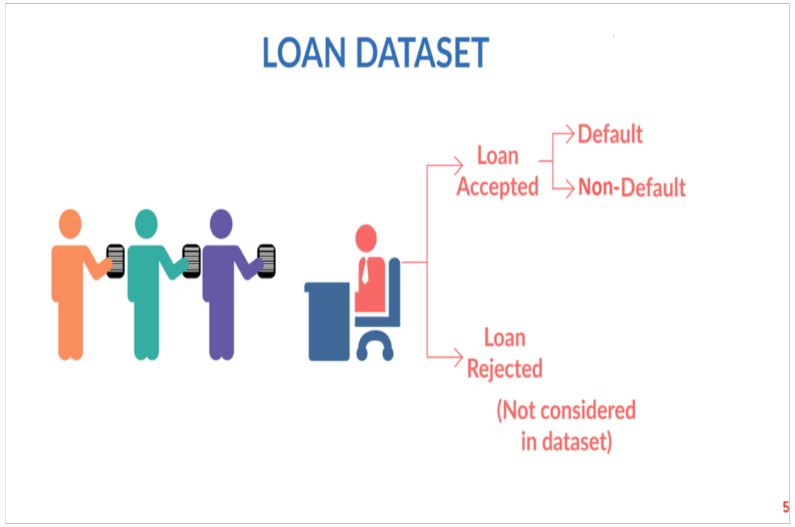
- The data is taken is imported from a csv file named “loan.csv”. The raw data consists of 39717 rows & 111 columns.
- We checked the datatype of various columns.
- There are different types of variables in the data as mentioned below:
  - Customer (applicant) demographic
  - Loan related information & characteristics
  - Customer behaviour(if the loan is granted)

Customer's Demographics
Employement Length
Employement title
Annual Income
Zip Code
Description

Loan Information & Characteristics
Loan Amount
Funded Amount
Funded Amount Investment
Interest Rate
Loan Status
Laon Grade

Customer Behaviour variables
Delinquency year -2
earliest credit line
Revolving balance
Recoveries
Application type
Loan purpose

- We also checked columns where we may need to require to extract numerical data for further data insights.
- We identified the target column. In our case it is ‘loan_status’.



Data Cleaning

We have followed the below methods to clean our data:

- Deleted columns:
  - Deleted unnecessary columns such as ‘zip_code’, ‘desc’,’url’, etc since these are irrelevant for our analysis.
- Removed outliers:
  - Removed extreme high and low values that would disproportionately affect the results of our analysis.
- Removed Missing values:
  - Dropped columns where there is missing values of more than 90%.
- Removed Duplicate data:
  - Removed identical duplicate data.
- Filtered rows:
  - Filtered the data further by loan_status to drop loan_status as ”Current”, as we only need “Fully Paid” & “Charged Off” data for our analysis.
- Imputed the NA values for all the variables.

After performing the above Data Cleaning steps, we proceed for EDA analysis over the cleaned data.

Data Analysis

- As part of our exploratory data analysis we tried to find the interesting fact and findings about the loans from the historical loan data that has been provided in the csv file.
- Since the customer behaviour variables are not available at the time of loan application, hence they cannot be used as predictors for credit approval.
- The ones whose loan_status are marked as 'current' are neither fully paid not defaulted, so we get rid of the current loans.
- We performed 2 types of EDA Analysis:
    1. Univariate Analysis
    2. Bivariate Analysis
- Univariate Analysis : In this type of analysis the default rate across various categorical and continuous features is checked. In case of continuous features, we performed binning to proceed with univariate analysis.
- Bivariate Analysis : Two or more features are used to understand the default variable(in our case ‘loan_status’).
- During our analysis we have created several plots to notice the relationship of variables as mentioned below:
    - Histograms and Bar charts to check out the distribution of all the driver variables
    - Box plots to detect the Outliers
    - Performed the Multivariate analysis to understand how different variables interact with each other.

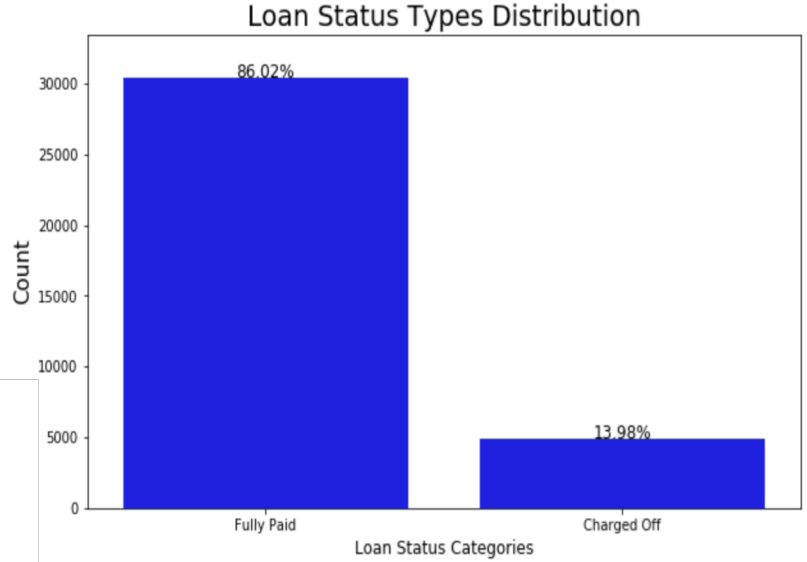
Analysis of LOAN STATUS with different variables

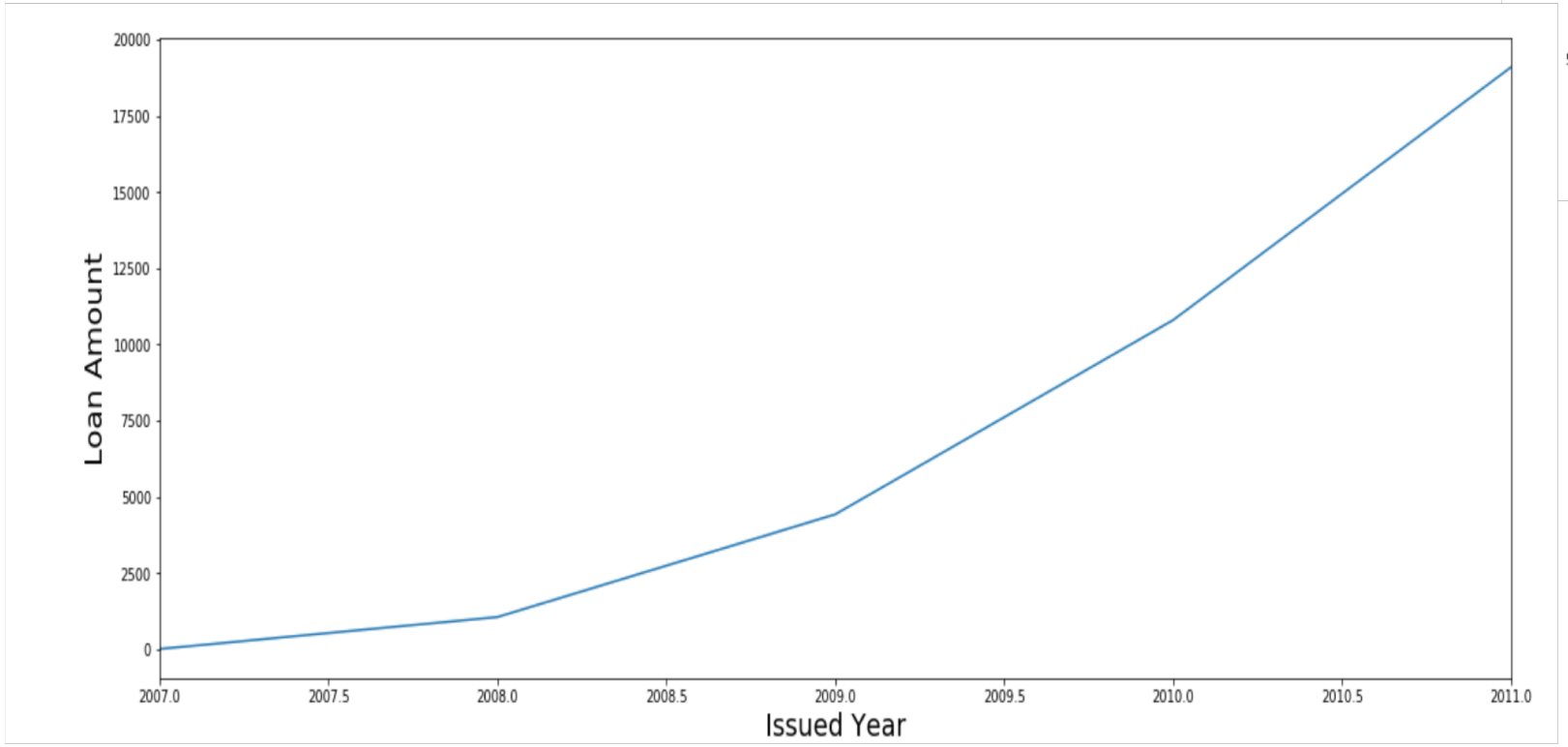
As we can see there are two categories of Loan Borrowers:

1- Fully Paid

2- Charged Off/Defaulter

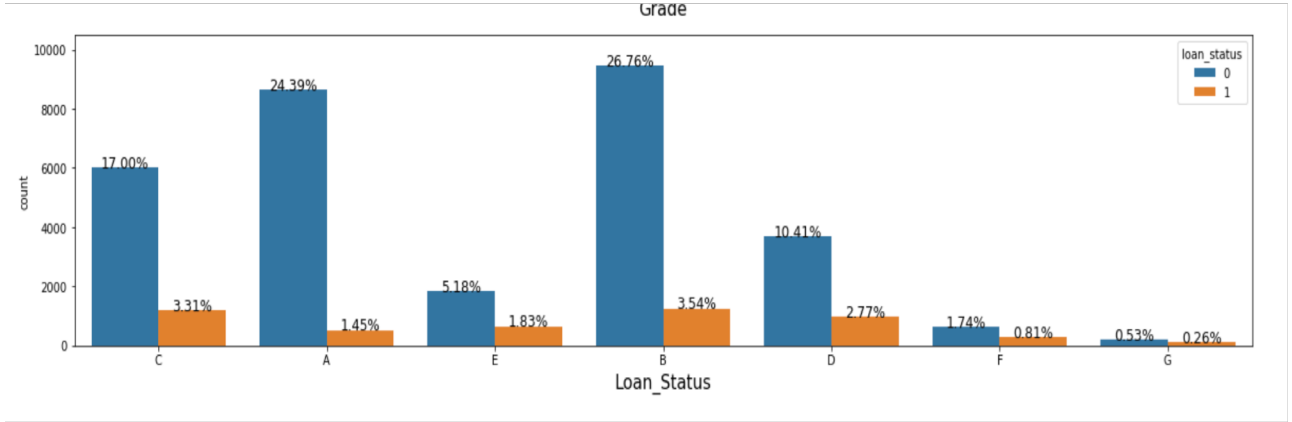
1. Most of the loans are Fully Paid as we see from the plot the frequency is -86.02%.
2. About 13.98% of loan are having status as “Charged Off”.
3. The loan has been increasing exponentially over the years as show in the below figure.

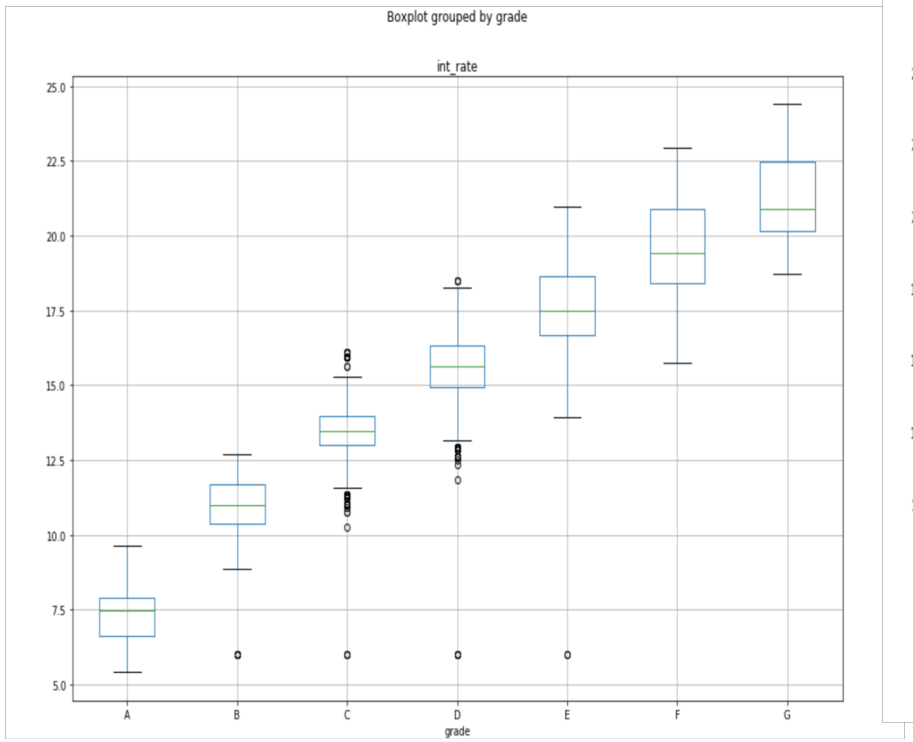


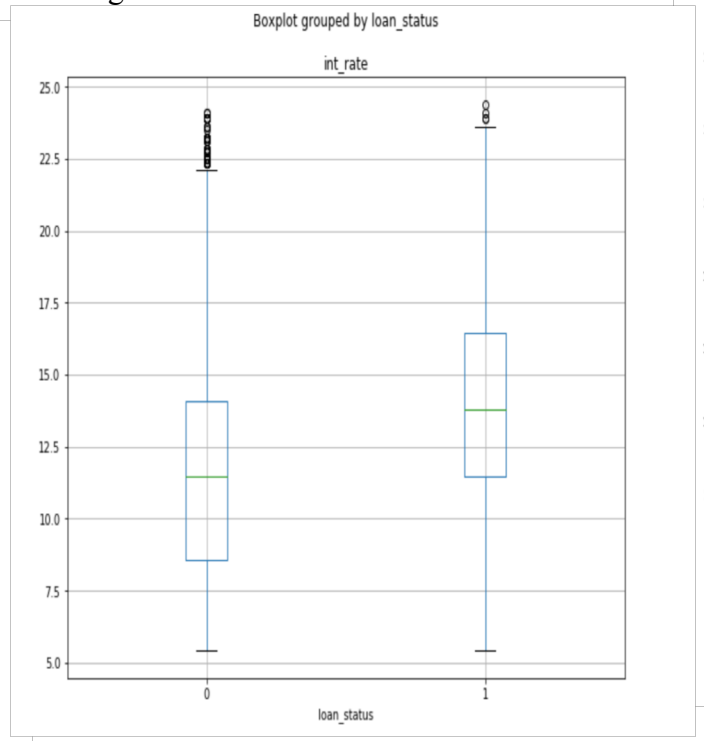


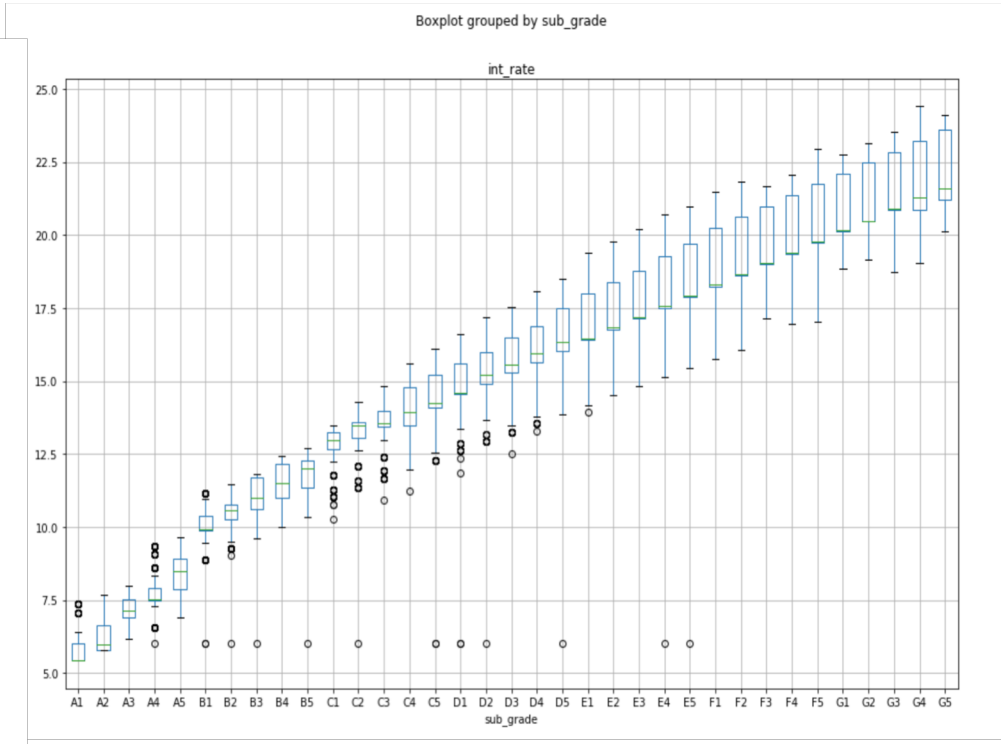
Analysis of Loan with Grade, Sub-Grade & Interest Rate

1. Most of the loans have grade of A and B as can be seen from the graph beside. This indicates most of the loans are high graded loans.
2. Grade A and B loans are considered as safe. The percentages in full dataset are much higher than percentages in Charged Off loans. Grade D, E, F, G loans are less safe.
3. The plot below shows that high graded loans have low interest rate.
4. While comparing with interest rate we notice that, the higher the interest rate , higher is the tendency to default the loan.
5. The percentage of interest rates increases as we go down the lower sub-grades.

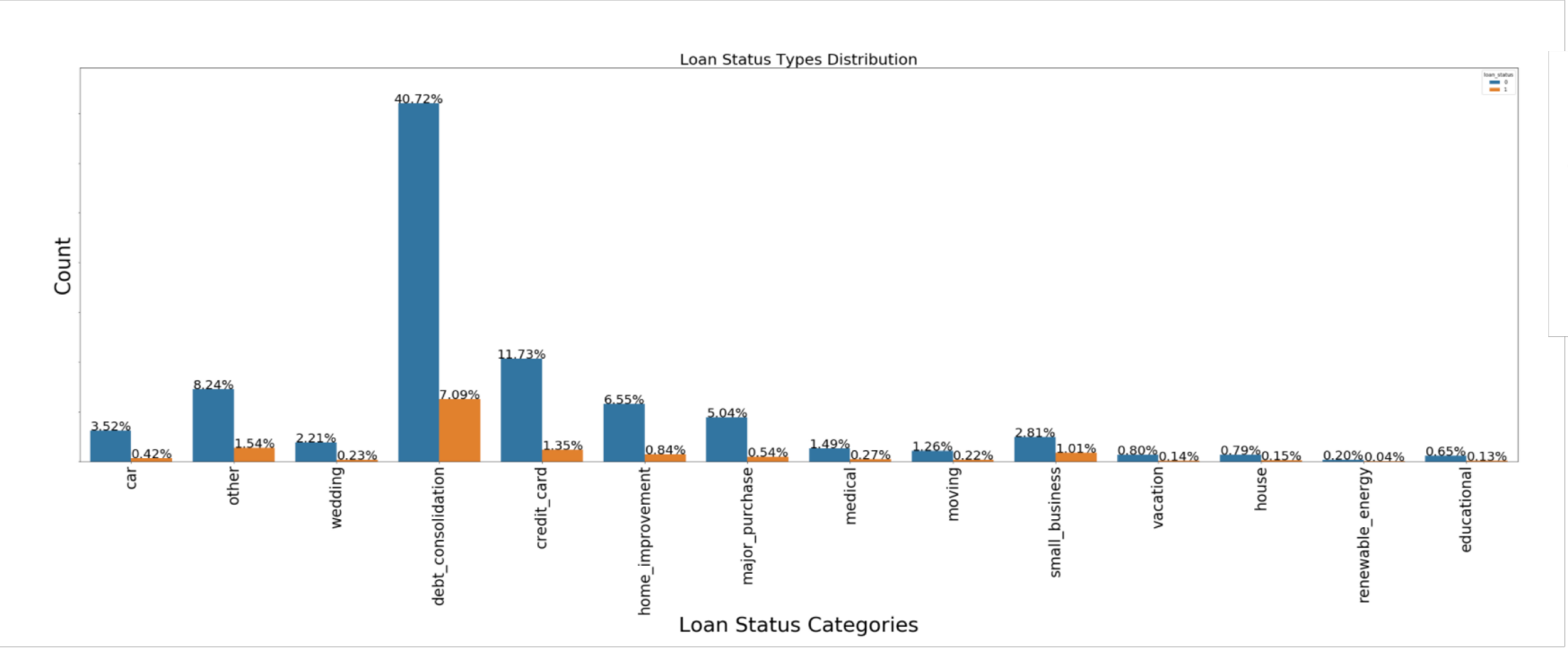


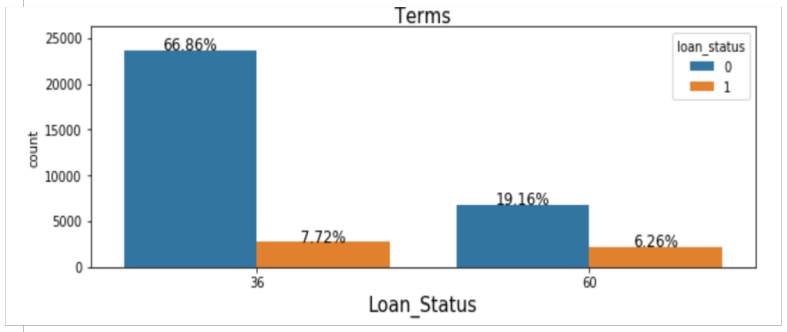




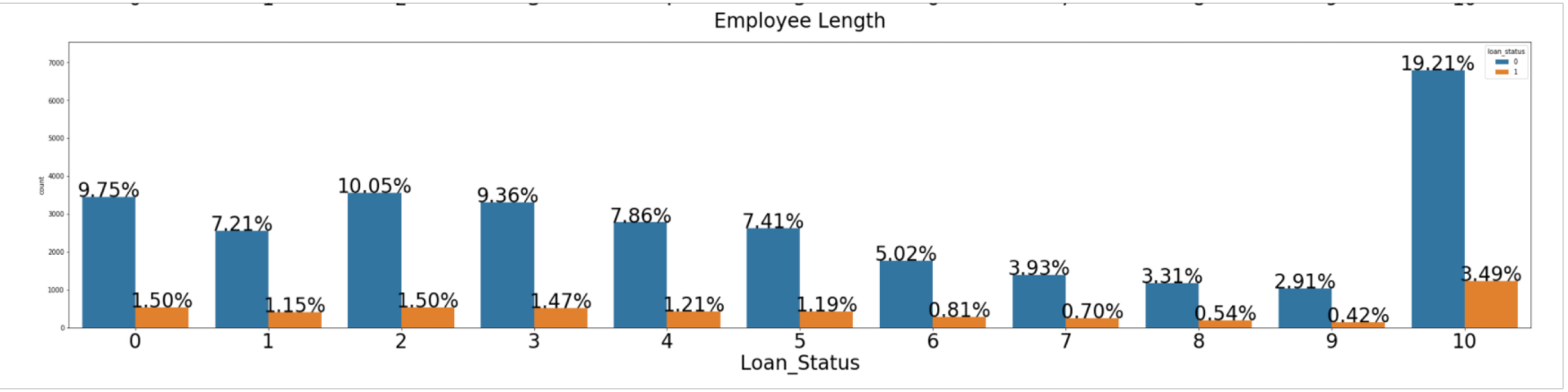


Analysis of Loan with Purpose, Term & Employee Length



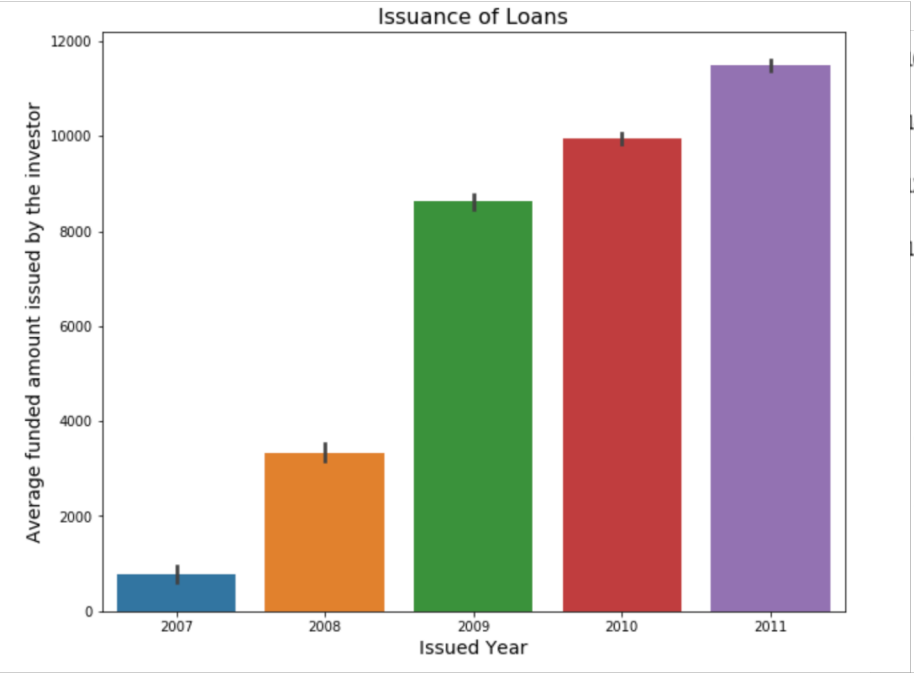


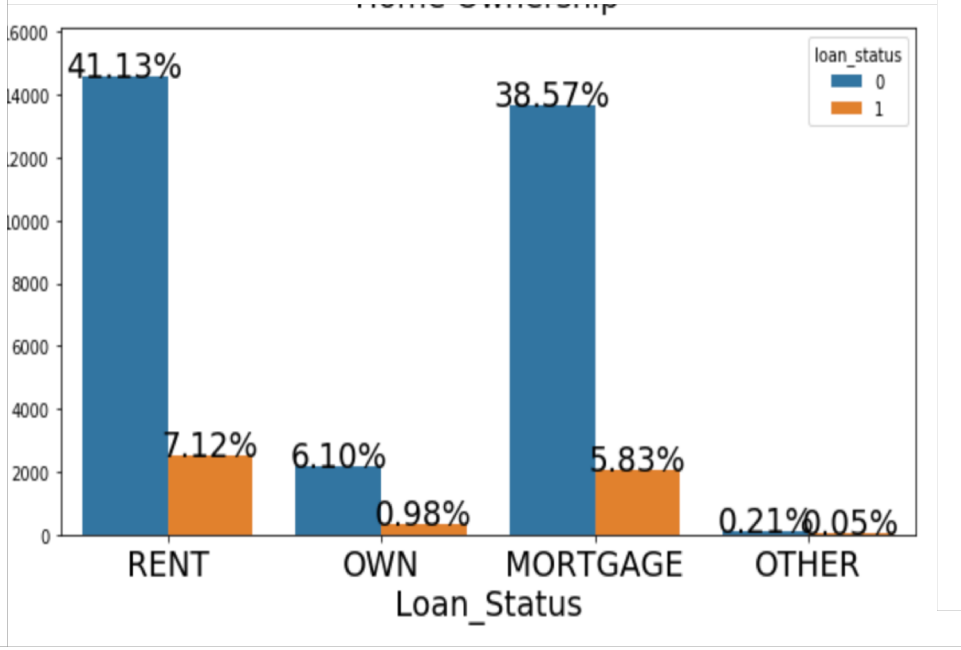
1. Most of the loans are granted were for '**debt consolidation**' purpose.
2. Most of the loans granted were of **36 months** term.
3. The loans that are granted for 36 months have slightly high tendency to default.
4. Majority of employees applying for the loan have more than 10 years of experience.
5. Loan Default rate is high with 10 years of employee length.

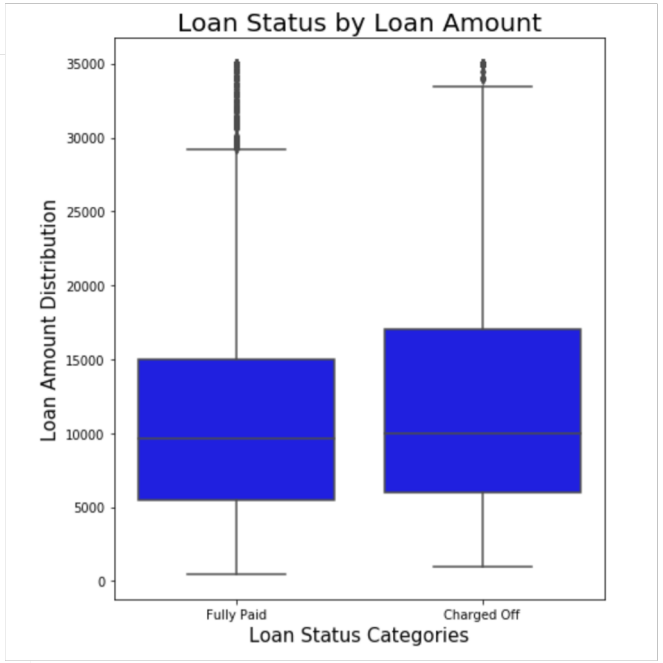


Analysis of Loan Default with Loan Amount, Home Ownership & Funded Amount

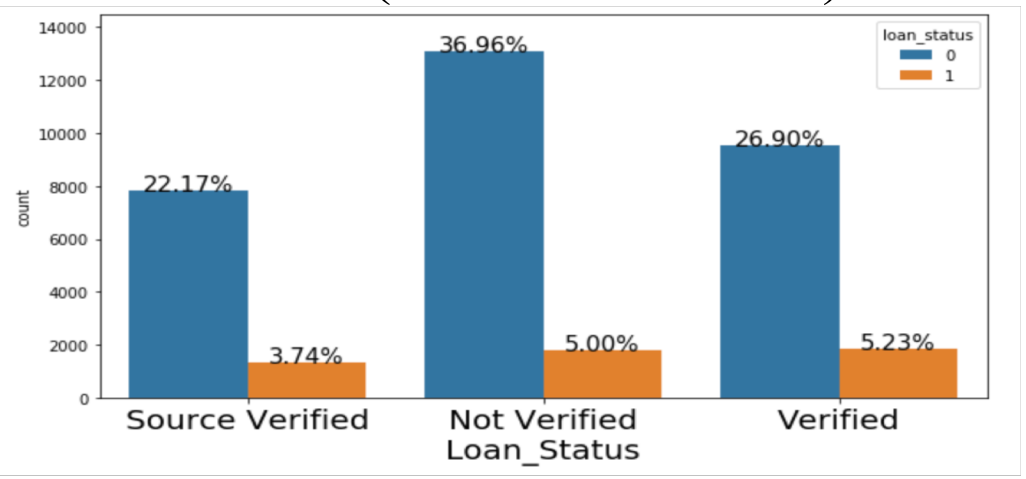
1. The loan applicants living in rented home or on mortgage have almost equal tendency to default the loan.
2. The range of loan amount taken as loan is higher for “Charged Off”(or Defaulters) as compared to “Fully Paid” borrowers.
3. The investors have invested the highest funded amount is in the year 2011 as we can see from the plot of Issued Year vs Average Funded Amount Invested.

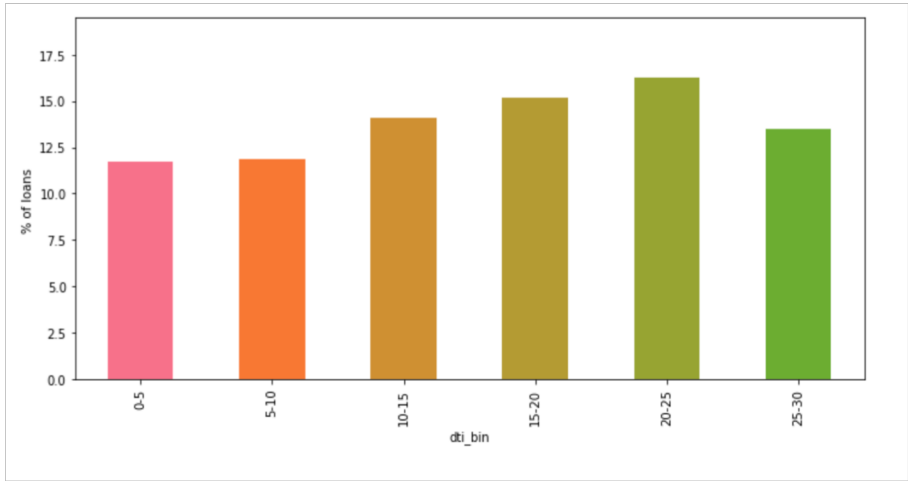


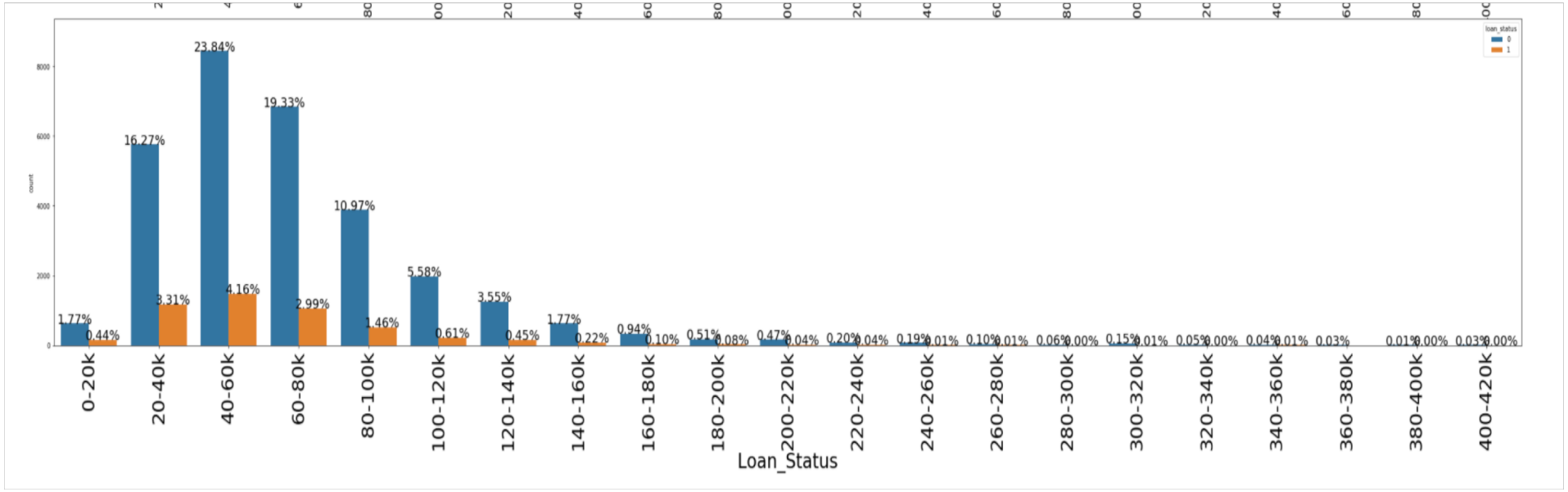




Analysis of Loan Default with Annual Income, Verification Status, DTI(Debt To Income)







- Majority of the loan applicants are not verified as we can see in the plot.
- The Loan Default borrowers are more or less same across the ‘Verified’ and ‘Not Verified’ category. We notice “Verification Status” does not much impact on the loan default borrowers behaviour.
- From the plot of dti_bin vs %age of loan defaults, we notice that when the dti (debt to income ratio) or debt payment is higher then higher is the chance of loan being Charged Off. Hence when applicants whose DTI ratio is higher can be considered as ‘risky’ applicants.
- The highest annual income group is $40k-$60k.
- Most of the defaulters are also present in this group. The default percent is 4.16%.

Conclusions

- The percentage of loan defaults is found to be nearly 14%(~ 13.98%_) .
- The high grade loans ‘A’ & ’B’ are the most common loans and can be considered safe to invest. The low grade loans(D,E,F,G) have a high tendency to default.
- The loan amount taken borrowed by “Charged Off”(or Defaulters) is more as compared to “Fully Paid” borrowers.
- The loan with high Interest Rate have more defaulters.
- The lower Sub-grades are seen to have high interest rates.
- ‘Debt consolidation’ is the most common purpose. Also it has high tendency to default.
- Majority of loan applicants applying for the loan have more than 10 years of service employment. They also contribute to maximum number of loan defaults as well.
- Loans taken for a period of 36 months has default rate higher than those taken for 60 months.
- Most of the loans were issued in the year 2011.
- Most of the loan applicants lie in the annual income group of $40k-$60k. Also the number of loan defaults is seen to be high in this annual income group.
- The loan applicants living in rented home or having mortgage property have almost equal tendency to default the loan.
- Also the verified and non-verified borrowers both have almost equal tendency of loan defaults. Hence the verification status should be checked thoroughly.
- Hence we can conclude that the Top-5 Major purchase loan variables:
    1. Grade
    2. Bin Interest Rate
    3. Term
    4. Home Ownership
    5. Year

Recommendations

- The higher the loan amount, the higher the likelihood of default. Investors should invest in loans that are approximation $10000 or less.
- Loans with term of 36 months have a high default rate as compared to loans with term of 60 months. So it is recommended the investor should invest in long term loans.
- Loan taken for the purpose of ‘debt_consolidation’ is most common . Hence it is recommended for a thorough background investigation before approving the loans for debt consolidation purpose.
- We noticed certain sub-grades are most likely to default compared to other sub-grades. It is recommended to select loans of subgrade B5 and higher which will result in a 90% chance of repayment.
- We also noticed that most of the loan defaults have a home ownership of “Rent” or “Mortgage”. It is better to investigate on background details more of the loan applicant.
- Loan Applicants whose DTI(Debt To Income) ratio is high are ‘risky’ applicants.
- Loan applicants having employment length of 10 or more years have taken the most loans. Also we see the default rate is high for such borrowers. Thorough investigation has to be done before giving loans to applicants of this category in order to avoid loan default.