Lending Club Case Study

Presented by,

- Prathima Kumari B V

- Seema Simoliya

Problem Statement

Lending Club (LC) which is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

When the company receives the loan application, it has to make a decision for loan approval based on applicant's profile.

Two types of risk are involved with the LC's decision:

1. If the applicant is likely to repay the loan, then not approving the loan might result in business loss.
2. If the applicant is not likely to repay the loan, i.e. he/she likely to default, then approving the loan might result in financial loss.

- 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.
- We will also find and provide our analysis on the driving factors behind loan default. The company can use this knowledge for risk assessment.

Business Understanding

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:

   a. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate).
   b. **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'.
   c. **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.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Data Description

- The dataset contains the complete loan data for all loans issued through the time period 2007 t0 2011 along with data dictionary for variable reference.

Objective of the Case Study¶

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.

1. Identification of risky loan applicant (defaulter)

2. Depending upon the amount of risk, which action could be taken
   a. denying the loan
   b. reducing the amount of loan
   c. lending (to risky applicants) at a higher interest rate

3. Understanding of driving factors(driver variables) behind loan default. i.e. Variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

4. Other conclusion which we can make by analyzing the data which will help the company to improve the business.

Data Cleaning

- Remove all the rows “Current ” loan_status rows which has
- Handling Null Values (Missing Values)
  - Drop all the columns which has more than 90% null values
- Remove all the columns which has single value
- Remove all the variables which does not add any importance to our analysis because these information the lender will get after the approval of the loan.
- Data Imputation with mode value for the null value columns (emp_length, revol_util, pub_rec_bankruptcies)

Numerical and Categorical Variables

- Numerical variables

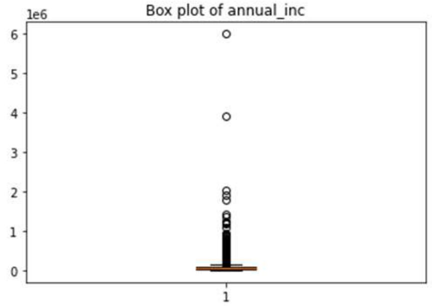
  'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'installment', 'annual_inc', 'dti', 'revol_bal', 'revol_util', 'int_rate'

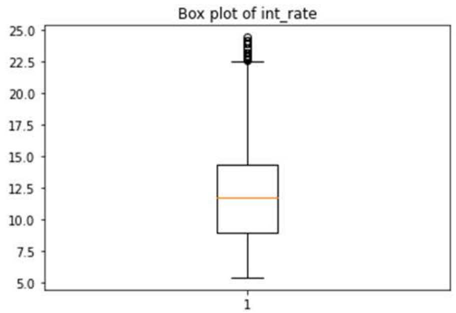
- Catagorical variabes

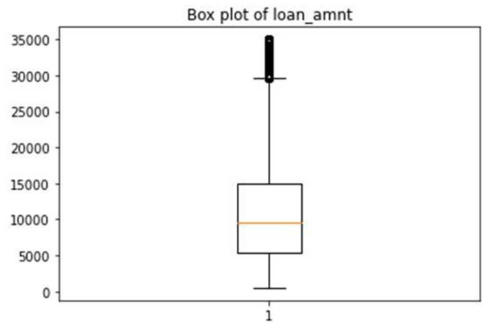
  'term', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'verification_status', 'issue_d', 'loan_status', 'purpose', 'delinq_2yrs', 'pub_rec', 'pub_rec_bankruptcies'

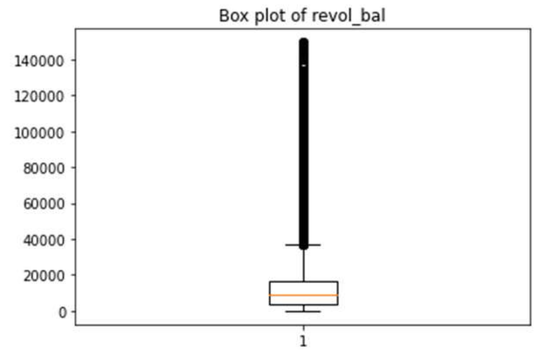
Exploratory Data Analysis (EDA)

Univariate Analysis of Numerical variables





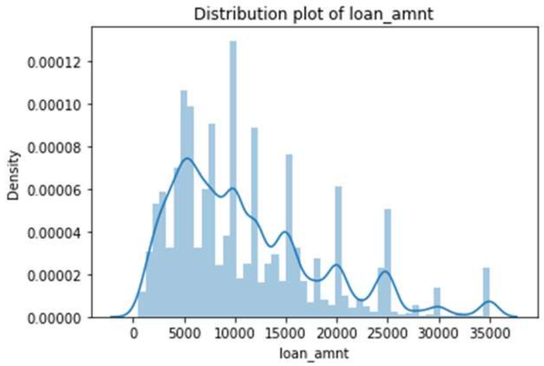


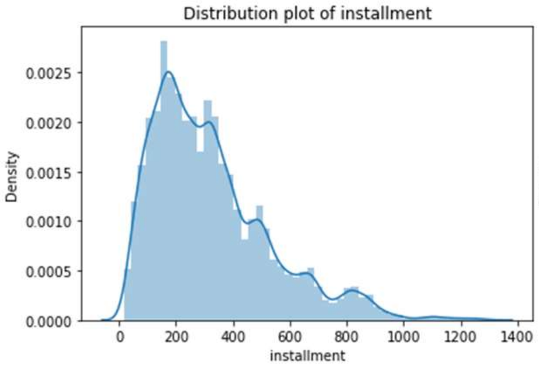


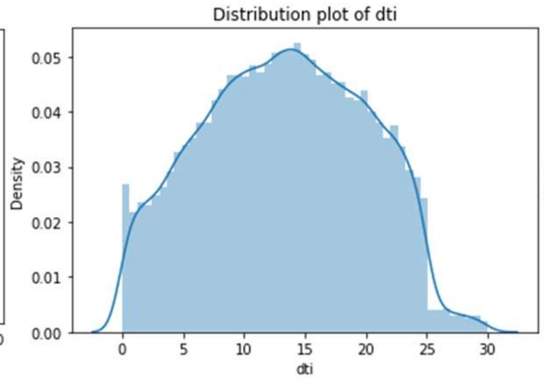
Observation

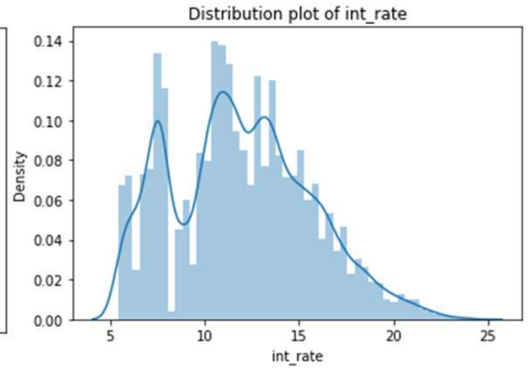
1. annual_inc has high value outlier points. Needs outlier treatment. (But we are not doing it because we want to check with true value for our analysis. We will convert it to categorical variable for our analysis)
2. The distribution of loan amount and interest rate is continuous and does not have much outlier points.
3. revol_bal boxplot shows more points outside the upper fence. But the distribution is continuous, so no need to remove outliers.

Univariate Analysis of Numerical variables





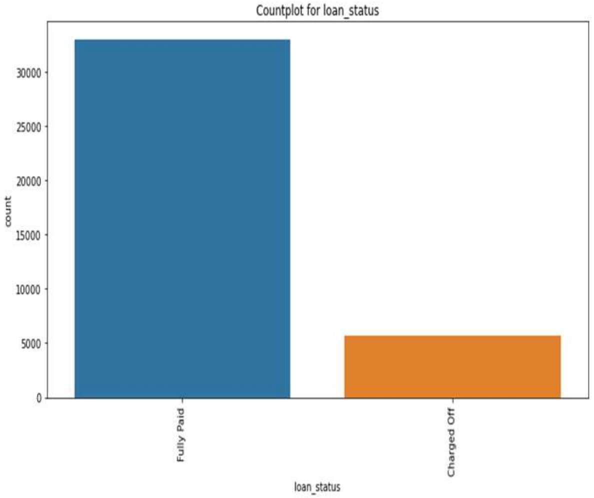


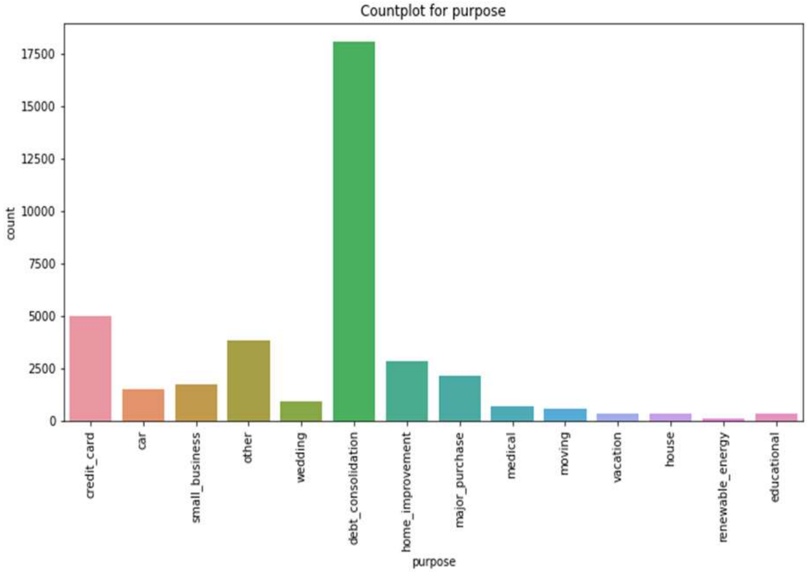


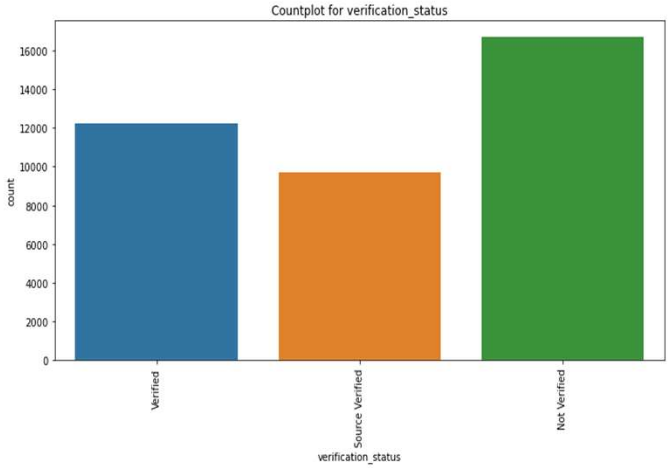
Observation:

- Large population is applying for small loan amount,very less people applying for high loan amount.
- LC also funding small loan amount more compare to high loan amount.
- More people paying low Installment amount
- Large population maintaining their dti ratio around 15
- Large population of applicants paying interest rate around 10% to 15% and 5% to 10%

Univariate Analysis for Categorical variables



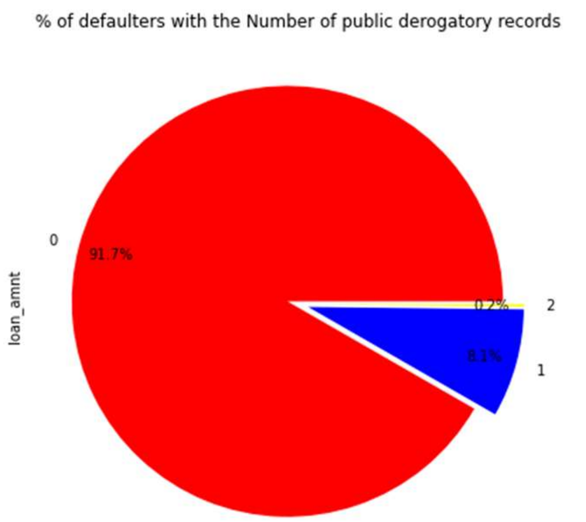


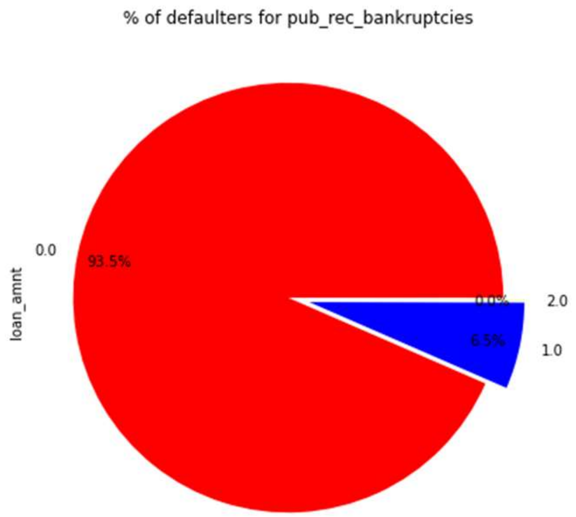


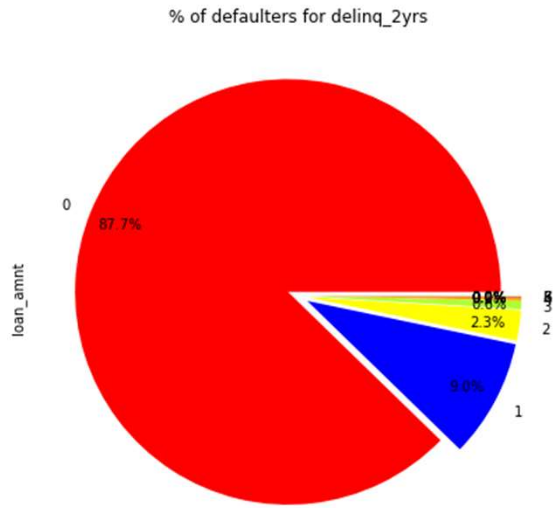
Observation

- Around 14% of applicants becomes defaulter according to the dataset
- Maximum people are lending loan for debt_consolidation (Its looks very risky)
- For more number of applicants, the income source was not verified.

The % of defaulters with Number of public derogatory records







Observation

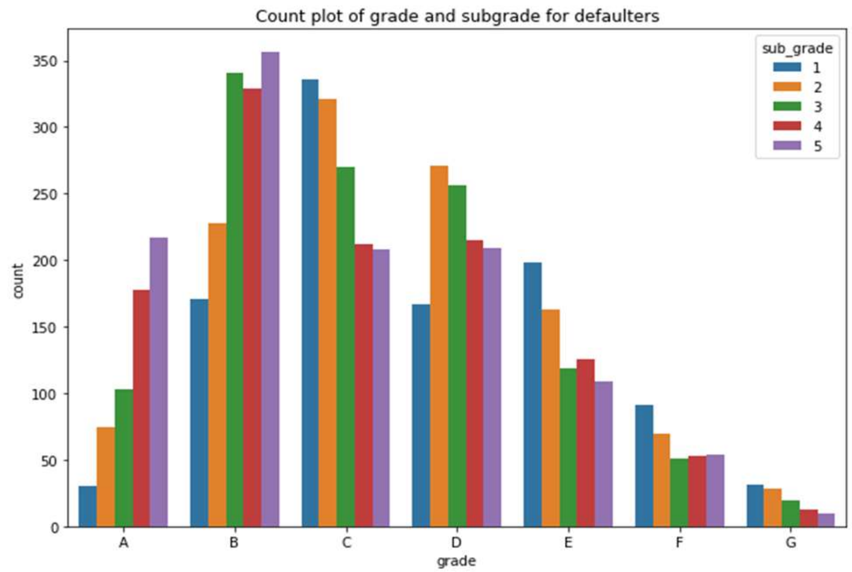
1.91.7 % of defaulters are having clean chit (do not have any derogatory records)

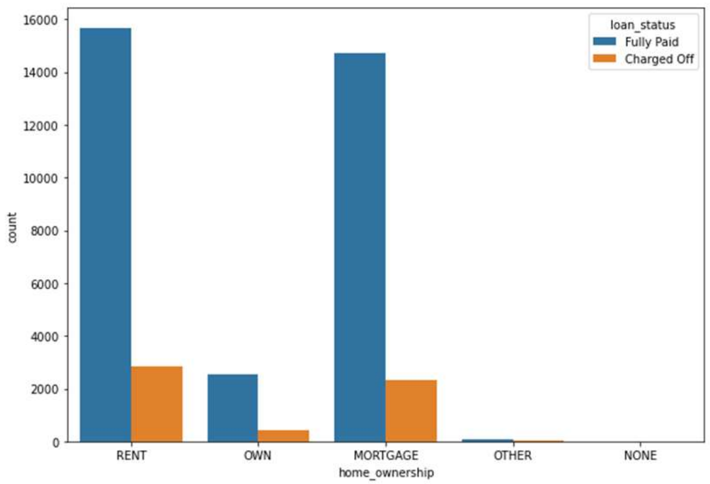
2.93.46 % of defaulters are having clean chit (do not have any public records of bankruptcies).

3.87.72% of defaulters are having clean chit (do not have any delinquency cases in 2 years)

4.New defaulters are more compare to the defaulters who already having history of bankruptcies, derogatory records.

We can give more importance to other parameters for our analysis





Observation

1. Maximum default cases comes under Grade B.
2. The level increases the default cases also increases in case of Grade A and B
3. In Grade C,D,E,F and G the almost the level increases the default cases are decreases.

Observation:

The people who's home_ownership is rented or mortgage are likely to be defaulters

Recommendation:

Lender should be little careful if the applicant is rented or mortgage

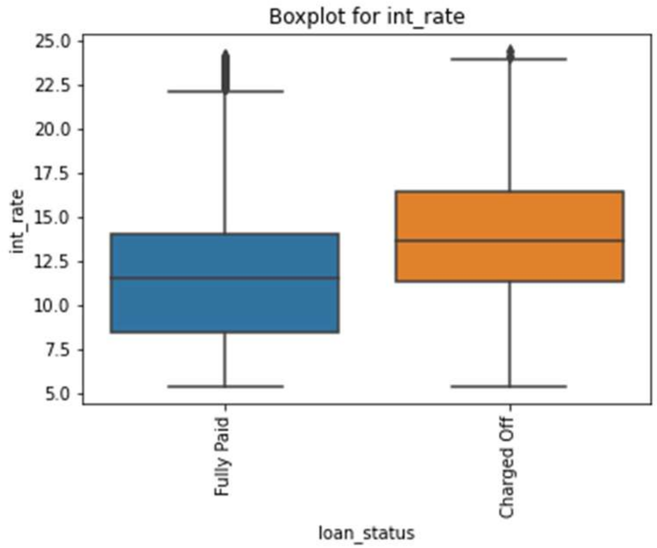
Observation

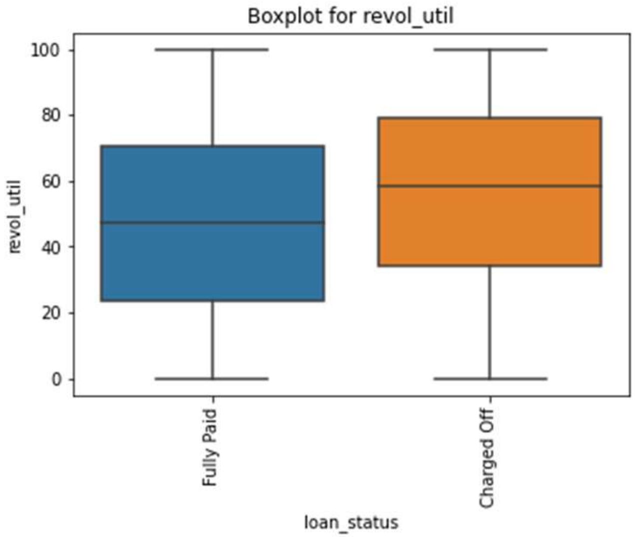
1. 50 % of defaulters staying in rented house. and 41.35% of defaulters are Mortgaged
2. 27% of defaulters having work experience greater than 10 years and 11.36 % of defaulters having work experience less than 1 year.
3. 49.17 % of defaulters are taking loan for debt_consolidation

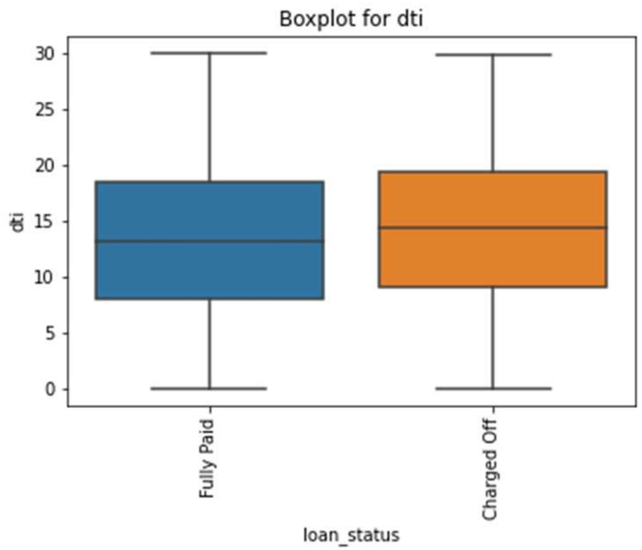
Insight

1. 10+ years of experience means the applicant is in middle age. He might be taking loan for home, car,home_improvement, debt _consolidation etc
2. less than 1 year experience means the applicant is a student or fresher. He might be taking loan for education,wedding

Bivariate Analysis



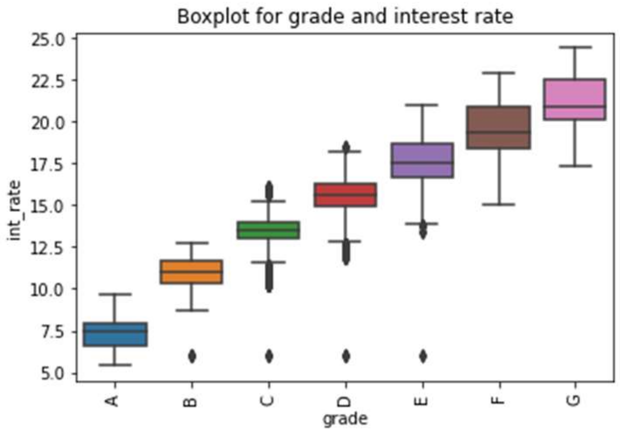


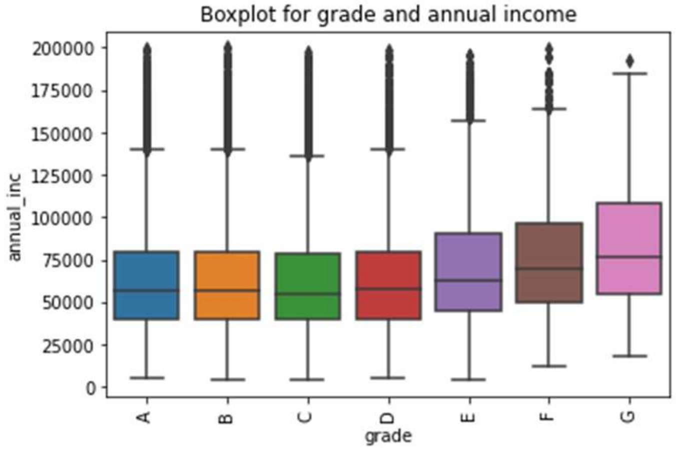


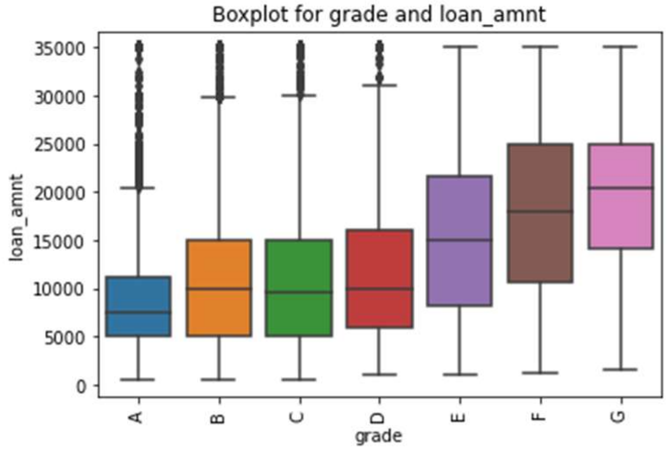
Observation

1. Interest rate is more for Charged Off applicants.
2. Defaulters have 10 units of higher Revolving line utilization rate(revol_util).
3. There is little difference in dti ratio of defaulters and non-defaulters.

Bivariate Analysis



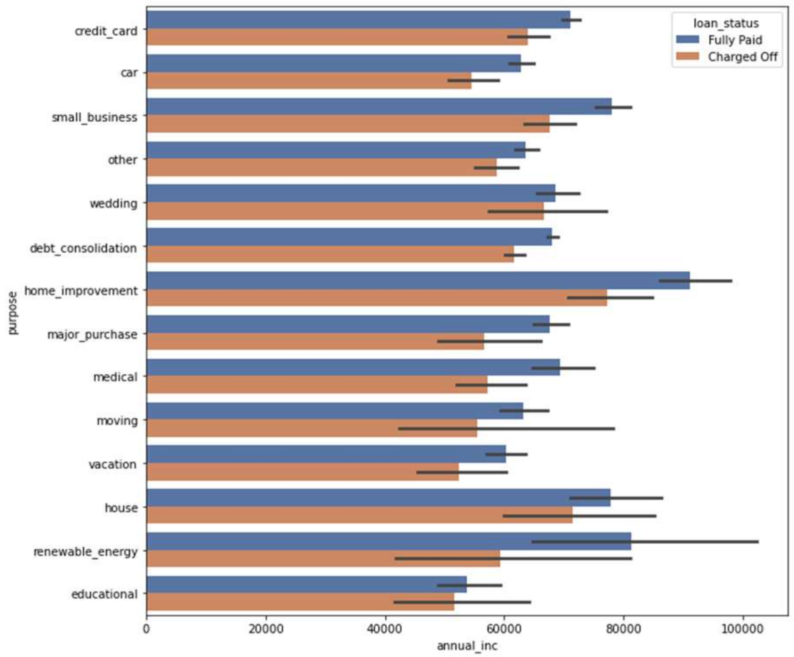


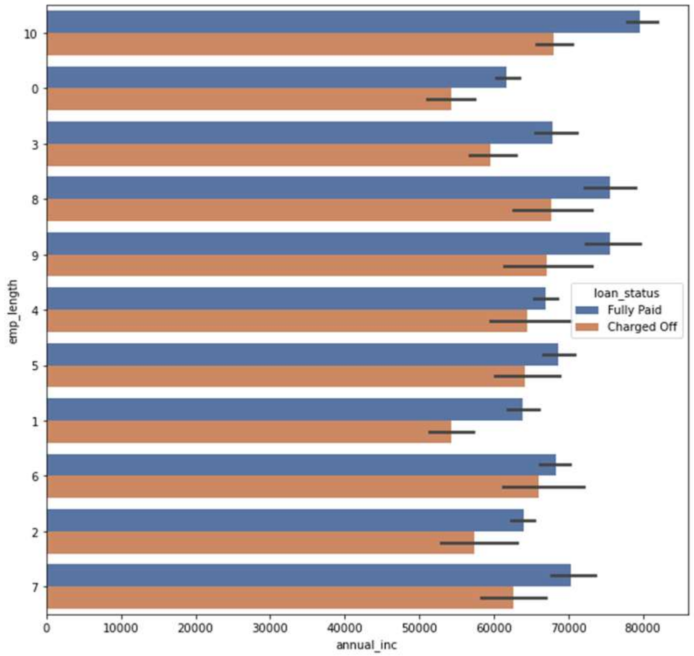


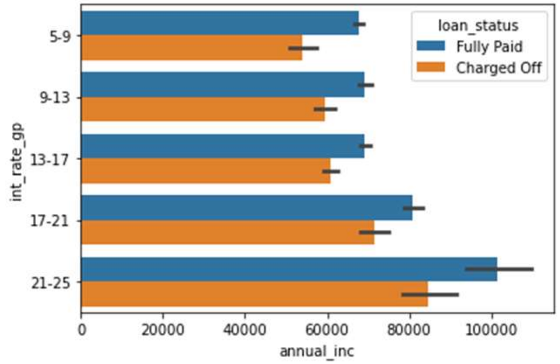
- **Observation**

1. The interest rate is increases from Grade A to G.
2. Grade A has less interest rate and Grade G has highest interest rate
3. Maximum number of people's salary ranges from 30000 to 80000
4. The loan_amount of grade G is high (range from 15K to 25K)
5. The loan_amount of grade A is low (range from 5K to 12K)
6. Grade G is having highest annual_income ,took high loan amount and paying high interest rate.

Bivariate Analysis



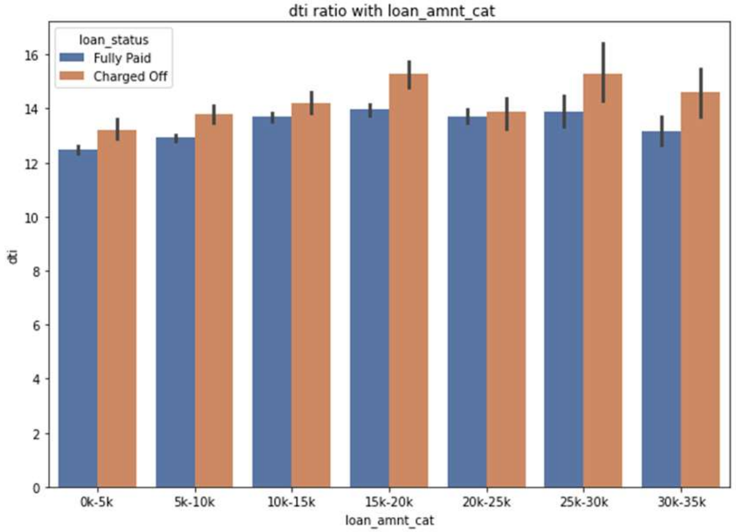


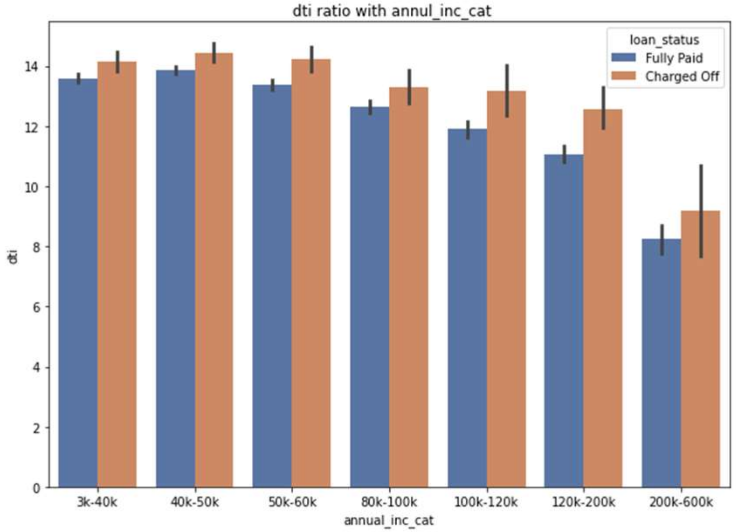


Observation:

1. From the above plot, we see that the people who are likely to default have took loan "home improvement", "house" and have salary less than 80K.
2. Also, the applicant who have loan fully paid has income more than 80K and have taken the loan for the same purpose.
3. People who's salary is less than 60K are more likely to be defaulter.
4. People with annual salary ranging from 60K to 70K with experience of more than 4 years are likely to default.
5. People with experience less than equal to 3 years is less likely to default.
6. People with low income has taken loan of higher amount with higher interest rate. Such people are more likely to default.

Bivariate Analysis

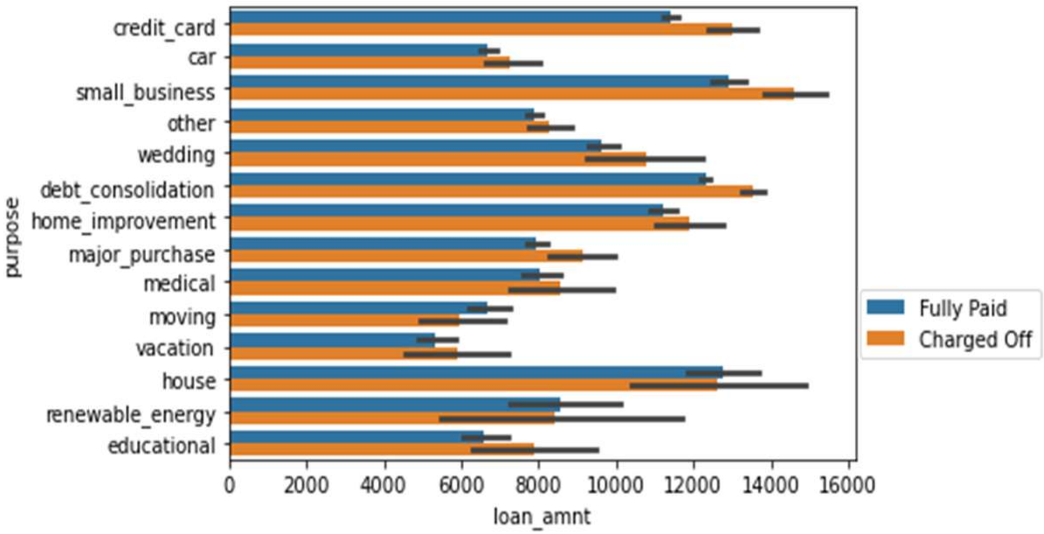


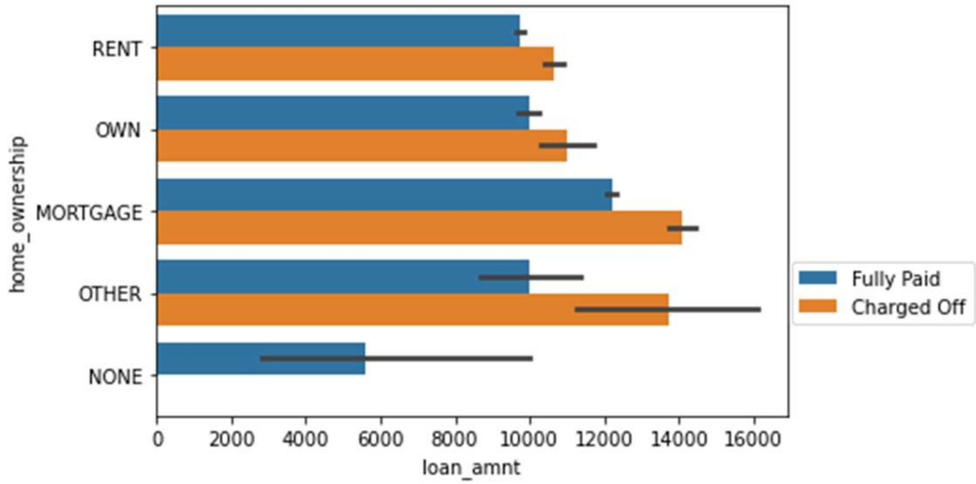


Observation

1. Who ever having high dti ratio(A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income) are likely to be defaulter.
2. The non defaulters who maintains their dti ratio around 12. People who maintains their dti ratio above 13 are likely to be defaulters
3. People who apply for big amount and having high dti ratio are likely to be defaulters
4. People who are having low salary are having high dti ratio likely to be defaulters.

Bivariate Analysis

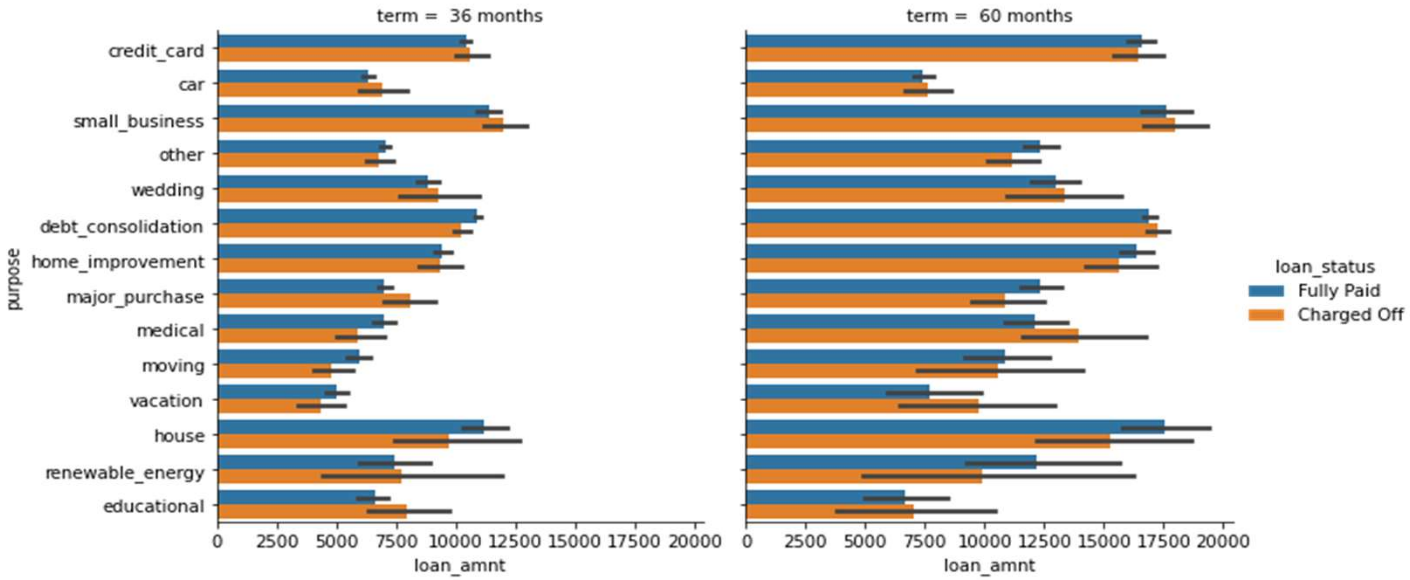




Observation:

1. Defaulters who applied for higher loan amount have taken loan for small business and have home_ownership as mortgaged.

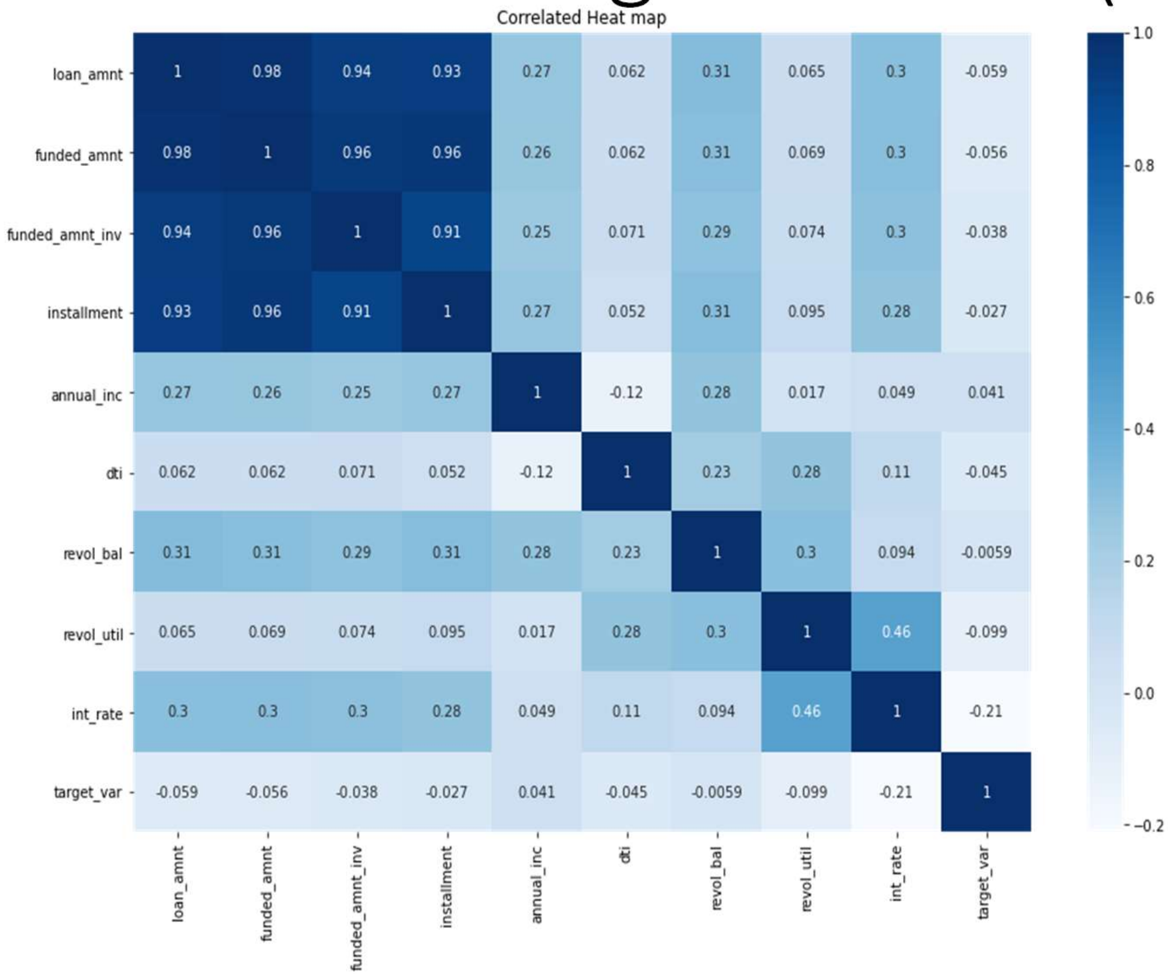
Multivariate Analysis



Observation

- The applicants applying high amount of loan for medical for 60 months term are likely to be defaulters.
- The applicants who took loan of 60 months term (High amount of loan) for small business are likely to be defaulters
- 36 months term is for low loan amount and 60 months term is for high loan amount.
- Applicants who are taking high amount for debt_consolidation are likely to be defaulters.

Correlation among all the numerical variables and with target variable(Loan_status)



Observation

1. People who are paying high interst rate are likely to be defaulter

2. dti,revol_bal and revol_util are also important but not for the first time loan applicants. In data set more new applicants are there, so the dti, revol_util not having much importance.

3. dti is negatively correlated with annual_inc.

4. The variables which are moderately correlated with default loan are the loan_amount and funded amount and dti.

Conclusion

1. The income source should be verified thoroughly before sanctioning the loan.
2. People who are having low annual salary and high dti ratio are more likely to be defaulter.
3. It is risky to give big loan amount to applicants with higher dti.
4. Small loan amounts are always non risky.
5. Large loan amount for small business should not have the home ownership as mortgaged.
6. People with higher dti meaning that their monthly income mostly goes to pay their debts. These debts are higher for the people who have taken loan to repay their credit card bills and Such people are likely to default a loan. On the other hand, the people with dti between 10-12 are less likely to default as they have taken loan for educational or house purpose.
7. The applicant who has more than 10 years of experience and having dti ratio high is likely to be defaulter.
8. The applicants having experience 8+ years applying for very high amount of loan having dti 0-6 or 24-30 are more likely to be defaulters.
9. Defaulting cases are high for big amount of loan.
10. The LC has to give more low amount of loan and reduce giving high loan amount.
11. Applicants who maintaining their dti ratio more than 13 are likely to be a defaulter. On the other hand who maintains dti between 10 to 12 are likely to be non-defaulter.

Business Driver Variables

1. loan_amnt : The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
2. purpose : A category provided by the borrower for the loan request.
3. home_ownership : The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
4. annual_inc : The self-reported annual income provided by the borrower during registration.
5. emp_length : Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
6. dti : A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.

Thank You