LENDING CLUB CASE STUDY

SUBMISSION

Name:

Sagar J

Deepak Sharma K

Abstract

- 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.

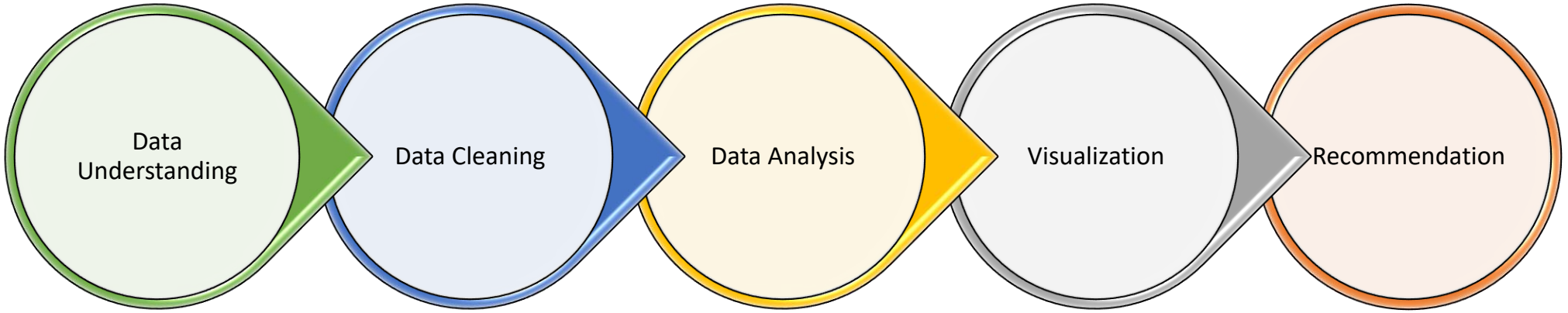
- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

- 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. In this case study, I have used EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Problem solving methodology

- The approach to flow of understanding and analyzing this case study is:
  - Data Understanding – Based on the various type of data and necessity of the columns in the provided Data Set.
  - Data Cleaning – Null Value Handling, Finding and Handling Outliers, Data Handling based on the requirement.
  - Data Analysis – Univariate, Bivariate and Segmented Analysis were performed to extract meaning from the data.
  - Conclusion – Recommendation to the Lending Club on driving variables as to how to take the decision on loan approvals.
- It takes the approach of the Waterfall Methodology to understand and analyze the data set.
- Visualization was used to make the best of the data set which was also a major step in this problem solving methodology.

Problem solving methodology



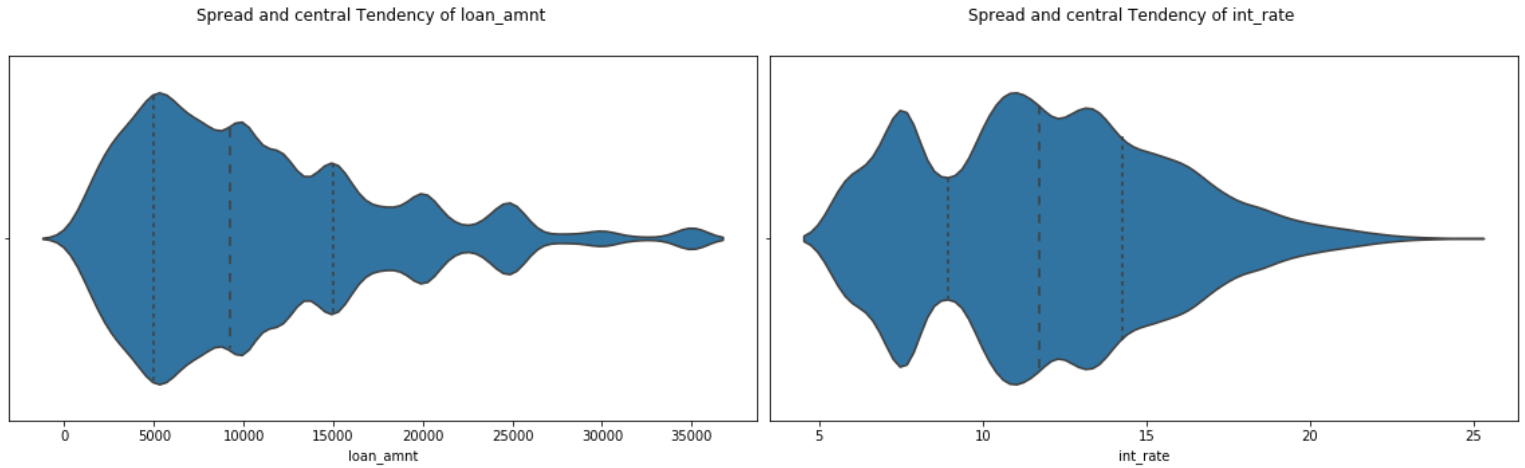
Assumptions and Treatments

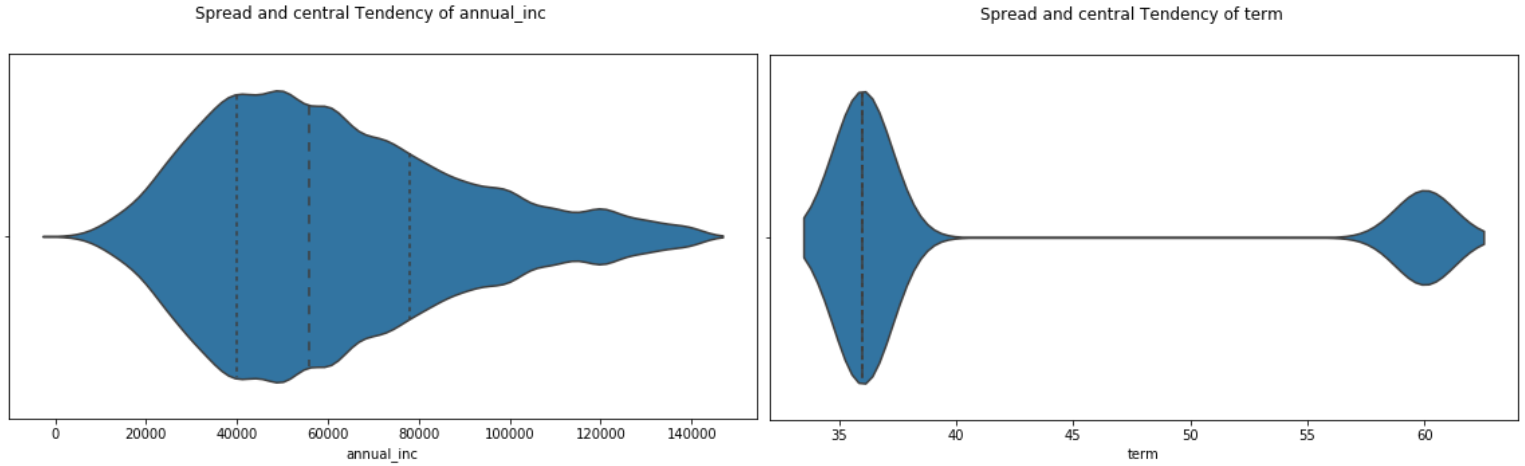
- In the provided Data Set,
  - 54 Columns had 100% NaN Values which was not necessary hence dropped those columns.
  - 11 Columns did not have any real predictive values hence dropped them from the dataset.
- In the remaining 46 Columns,
  - Non numeric data from the ‘term’ and ‘emp_length’ columns were stripped off along with the unnecessary whitespaces.
  - Null Values in ‘emp_title’ and ‘emp_length’ were assumed to be ‘Self Employed’ with ‘0’ months of experience.
- Issue date which was in ‘issue_d’ column was spilt into ‘issue_d_month’ and ‘issue_d_year’ for periodical understanding of data.
- Those loans which are currently under repayment process are not used in making any analysis as it doesn’t give a proper insight to the analysis.
- Outliers have been treated using the standard of 95% for annual income.
- Data in Loan Status column has been replaced as follows: Charged Off as ‘1’ and Fully Paid as ‘0’ for numerical assessment.

Analysis – Univariate (Complete Data Set)

Univariate Analysis is performed to understand the individual distribution of data in the dataset.

- Loan Amount distribution is higher between 5000 and 15000
- Interest Rate distribution is higher between 8 and 15
- Annual Income distribution is higher between 40000 and 80000
- Term of Loan is much higher as 36 months rather than 60 months.

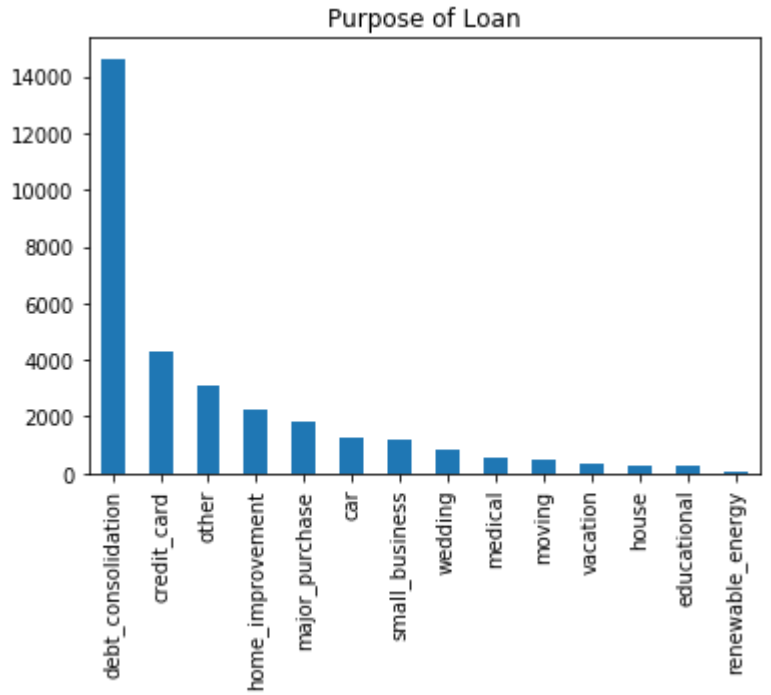


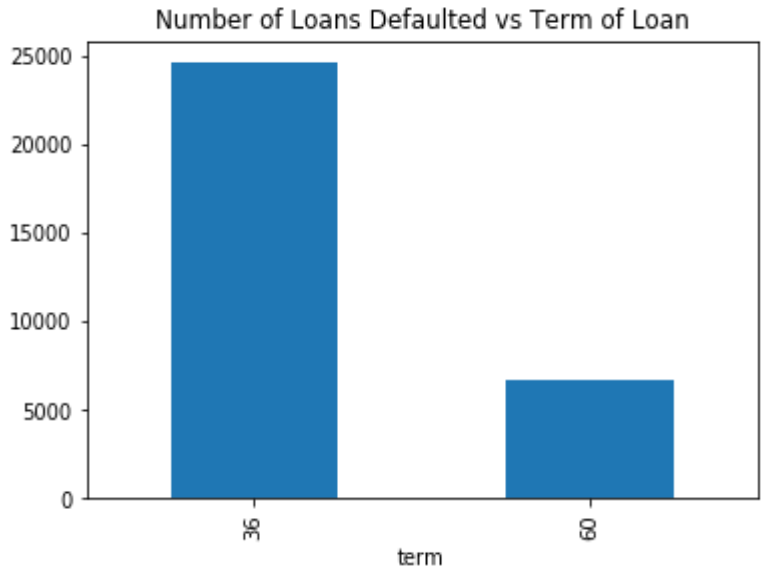


Analysis – Univariate (Charged Off Data Set)

Univariate Analysis is performed to understand the individual distribution of charged off data in the dataset.

- Purpose of Loan is debt_consolidation being the highest for the number of defaulters.
- Highest number of people defaulted from loan had a Term of Loan as 36 months.

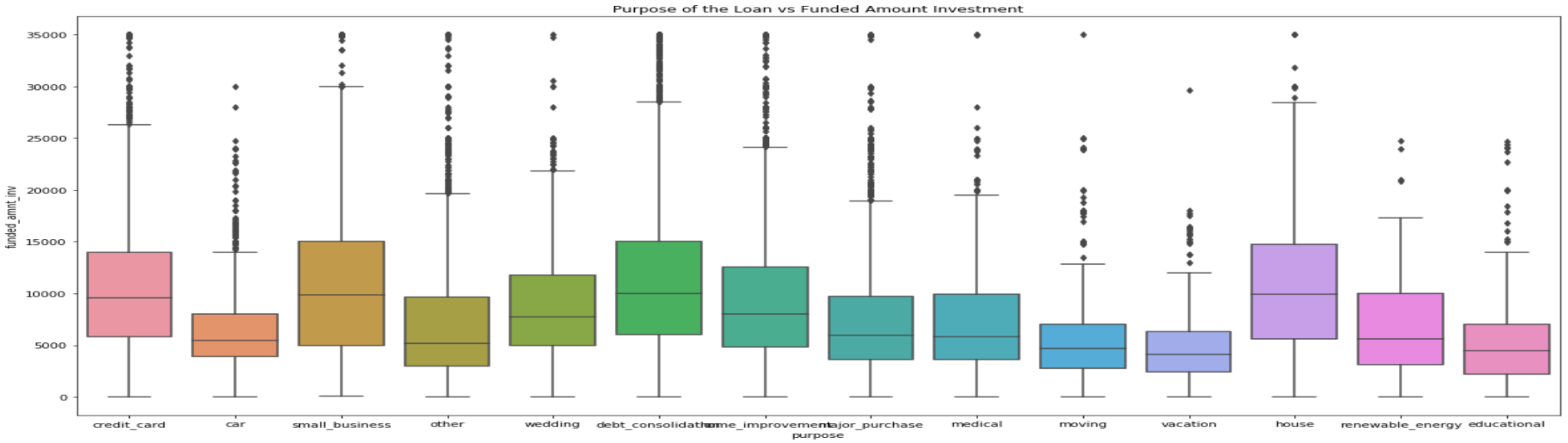




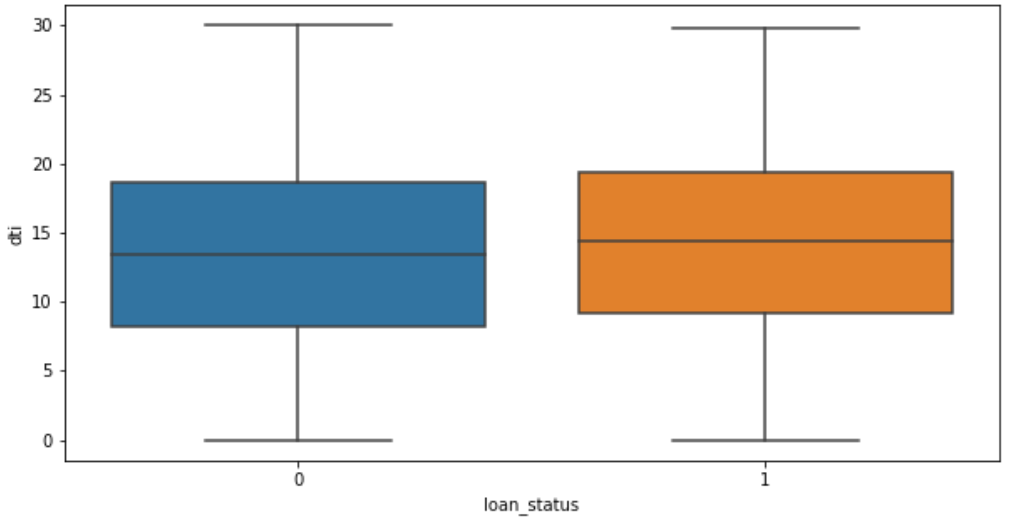
Analysis – Bivariate (Complete Data Set)

Bivariate Analysis is performed to understand the distribution of complete data in the dataset between Funded Amount Investor and Purpose.

- Purpose of Loan is debt_consolidation being the highest for the amount of defaulters.
- Small business has the widest spread of Loan.



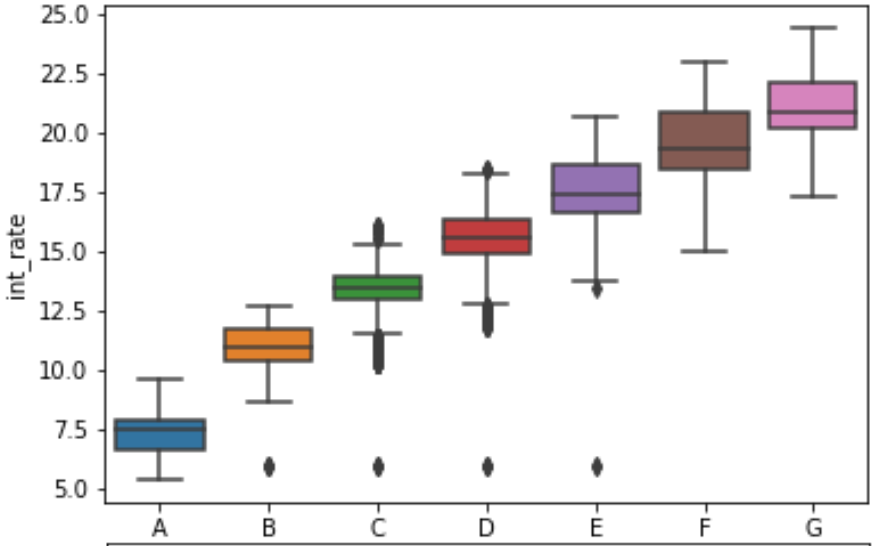
- Charged Off Loans have more debt to income ratio.

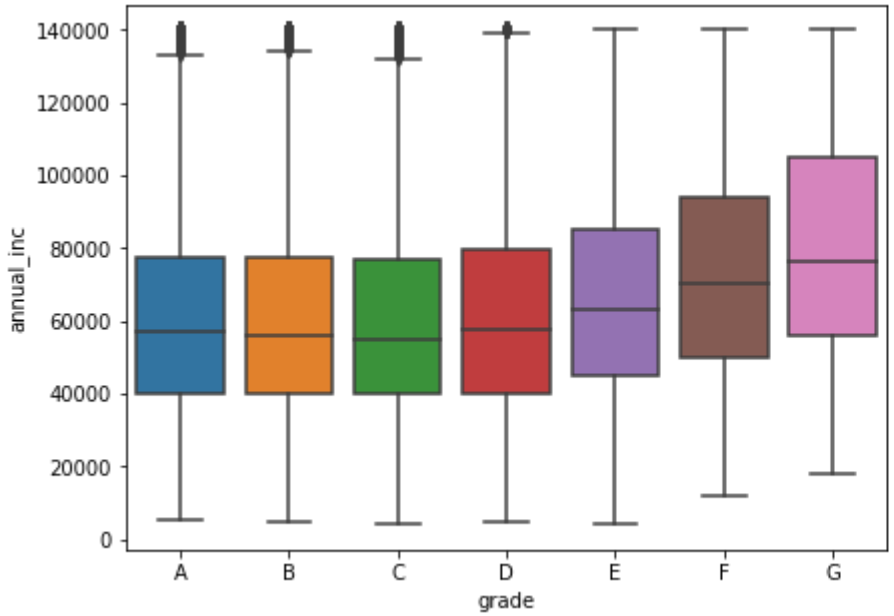


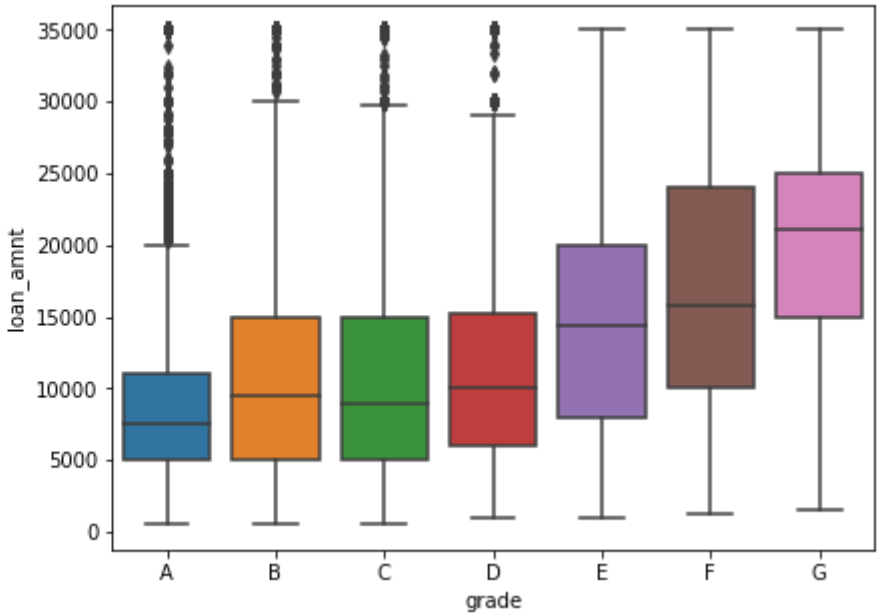
Analysis – Bivariate (Charged Off Data Set)

Bivariate Analysis is performed to understand the distribution of charged off data in the dataset between Grade and Rate of Interest and between Grade and Loan Status.

- Grade is directly proportional on the Rate of Interest. Grade is divided based on the Rate of Interest.
- Annual Income and Loan Amount are directly proportional. People with higher annual income were given higher loan amount as requested.
- People with Higher Annual Income were majorly those who were defaulting.



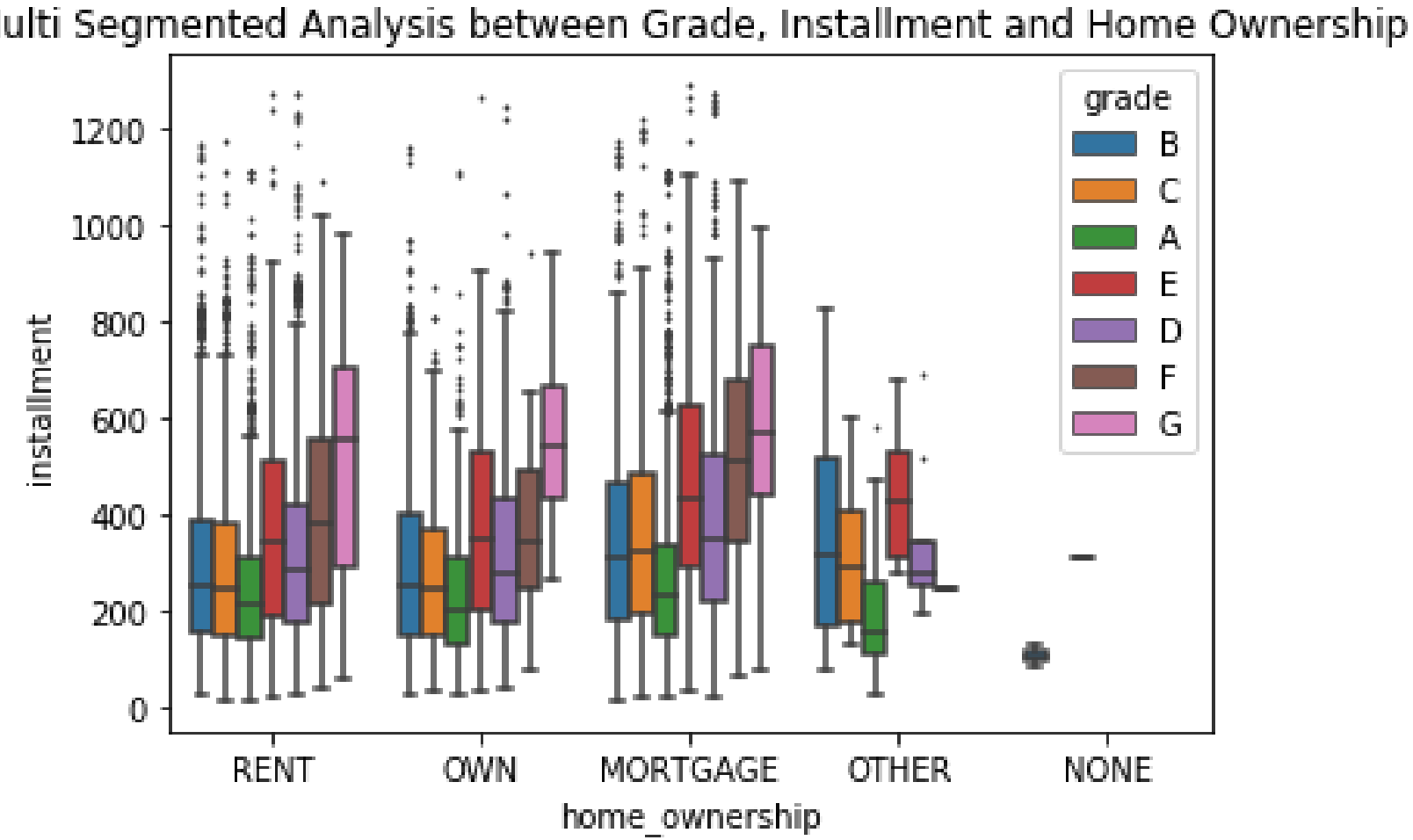




Multi – Segmented Analysis (Charged Off Data Set)

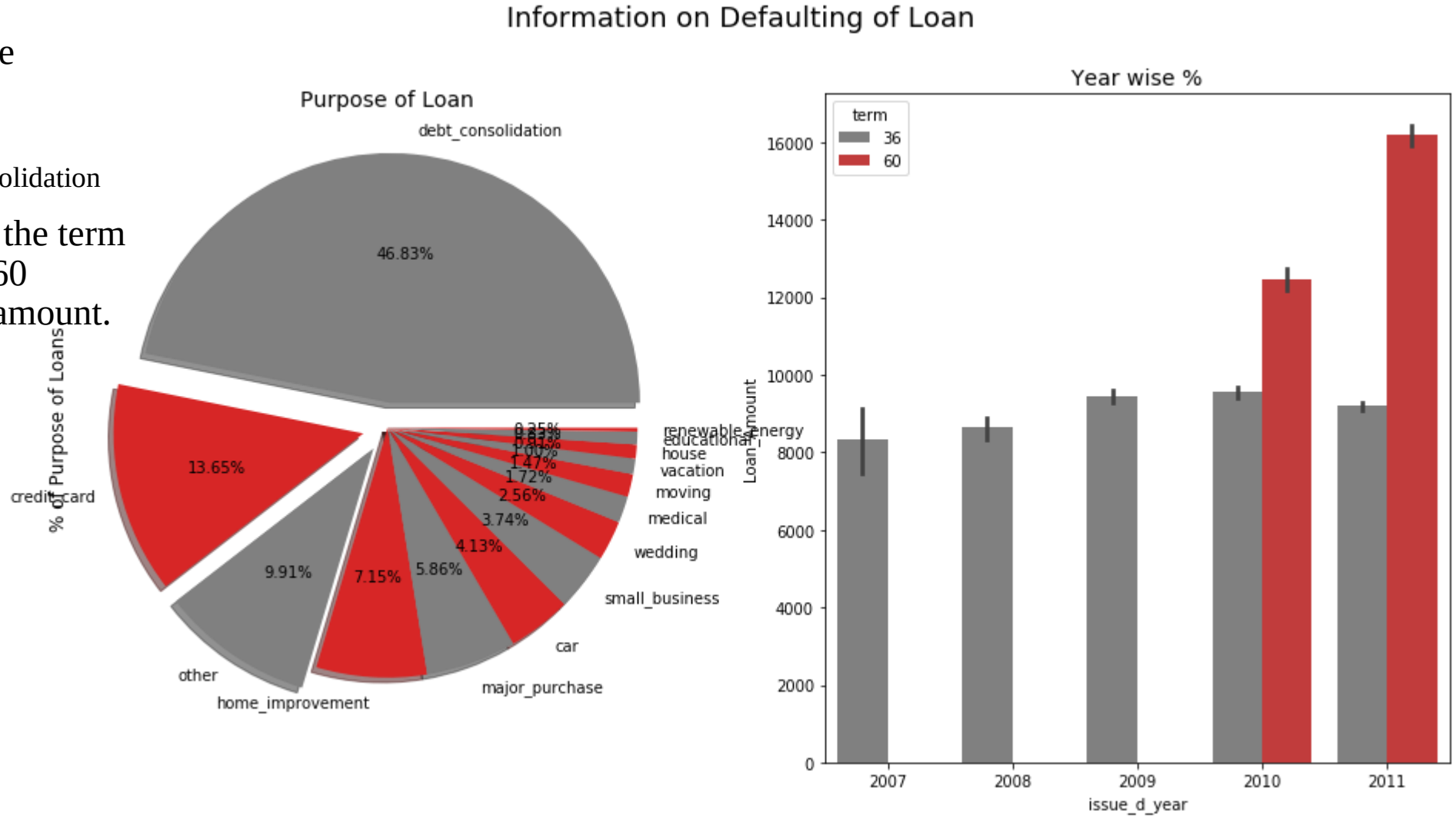
- People in charged off region are highest in installment amount from Rent, Mortage and Own Home Ownership and they fall in 'E','F','G' grades.

- The widest spread is found in rent home ownership.



Multi – Segmented Analysis (Charged Off Data Set)

- Charged Off loan has the highest purpose in Debt Consolidation.
  - i.e.: 46.83% - debt_consolidation
- In the recent past years, the term of loan is increasing to 60 months for higher loan amount.



Conclusions / Recommendation of Analysis

Based on the Analysis, Below mentioned are the driving factors for the consumer to Default.

- **Interest Rate:** Loans that have higher rate of interest.
- **Term of Loan:** Based on Current Trend, Loans that had longer term. i.e.: 64 Months
- **Home Ownership:** People who are staying in rented house are much prone to default.

Based on the Analysis, Decision can be taken using below attributes along with combination of the Major Factors as mentioned above are:

- **Purpose:** Debt consolidation is one of the major purpose of the loan with few certain combinations we can predict if they can be defaulter. i.e.: 40% - Debt Consolidation
- **Debt to Income Ratio:** Higher the dti ratio, Higher is the chances of defaulting.

The combination of the above mentioned driving factor can help the Lending Club take decision on the loan approval process.