Lending Club Case Study

Group Members :
Diwansh Sharma
Shankar Naik

Problem Statement

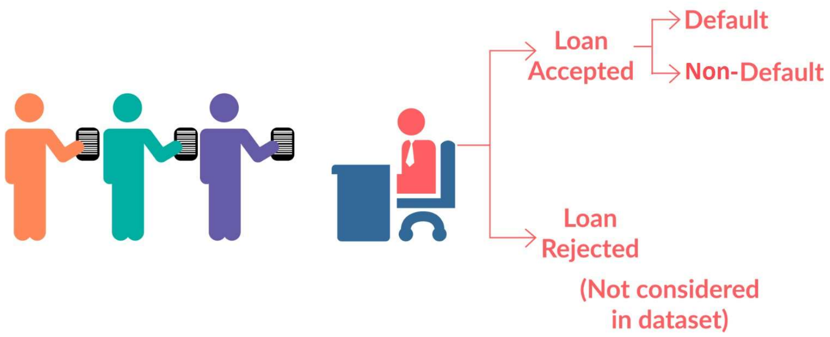
A consumer finance company specialises in lending various types of loans to urban customers. When the company receives a loan application, it has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- 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 such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

LOAN DATASET



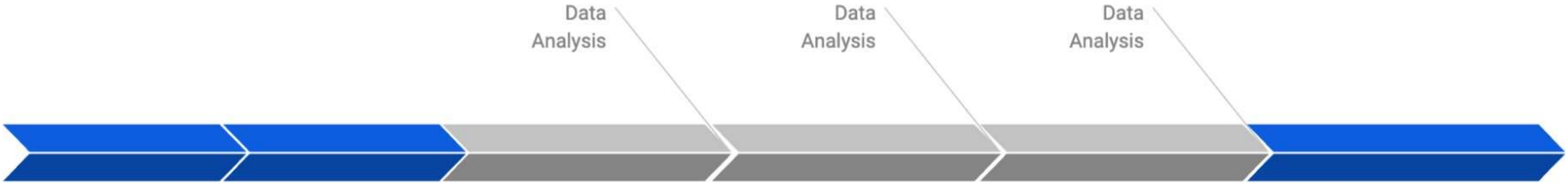
When a person applies for a loan, there are two types of decisions that could be taken by the company:

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

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)

Problem solving methodology



Data Cleaning

Removing the null valued columns, unnecessary variables and checking the null value percentage and removing the respective rows.

Data Understanding

Working with the Data Dictionary and getting knowledge of all the columns and their domain specific uses

Univariate Analysis

Analysing each column, plotting the distributions of each column.

Segmented Univariate Analysis

Analysing the continuous data columns with respect to the categorical column

Bivariate Analysis

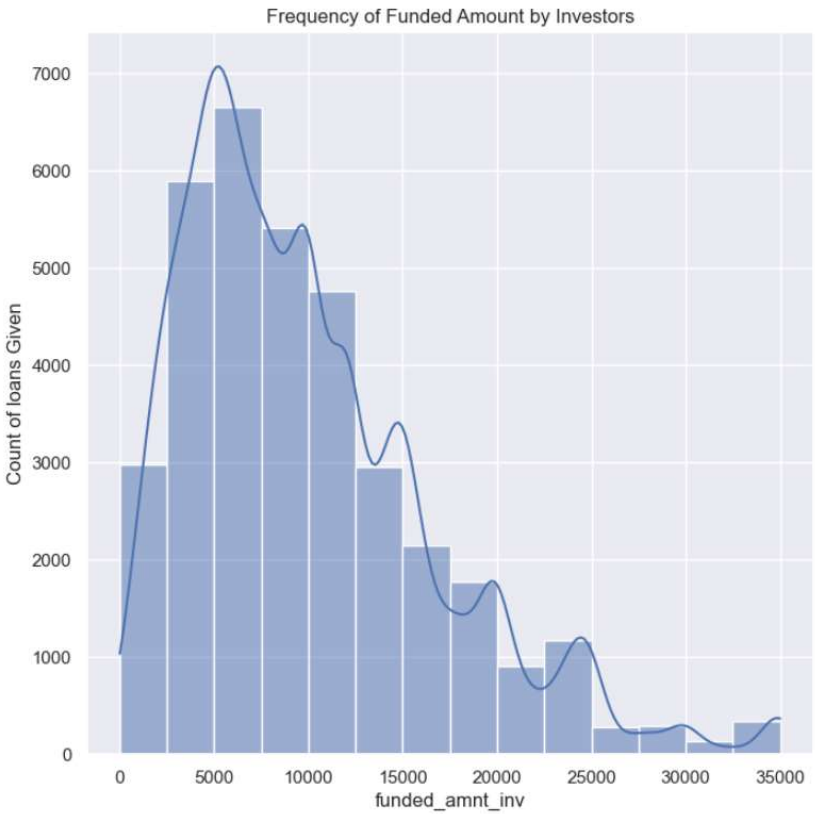
Analysing the two variable behaviour like term and loan status with respect to loan amount.

Recommendations

Analysing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

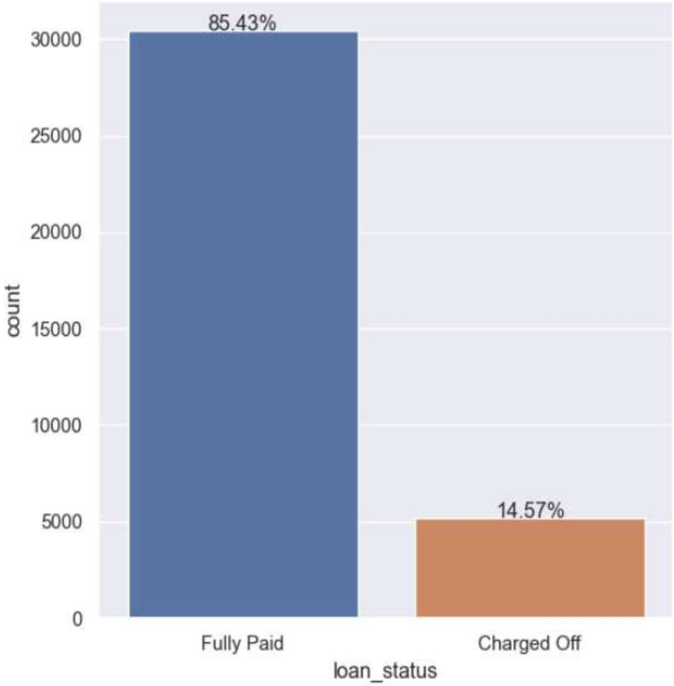
1. Univariate Analysis

Frequency of Funded Amount by Investors



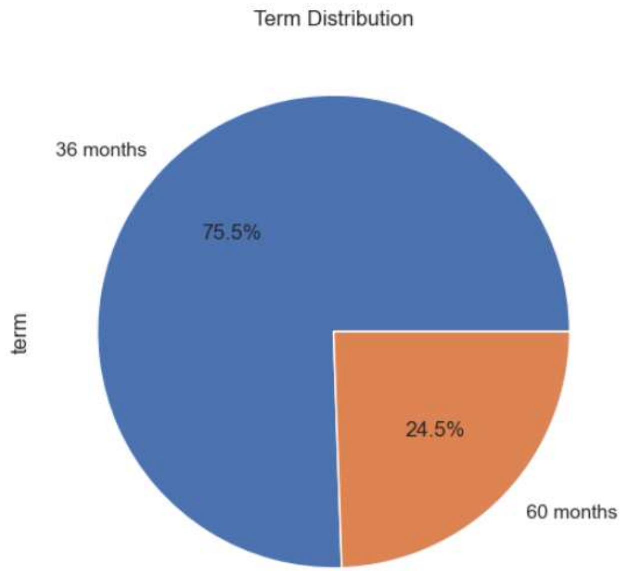
- The above histogram indicates that most of the funding was between 2500 - 10000 range.

Loan Status Distribution



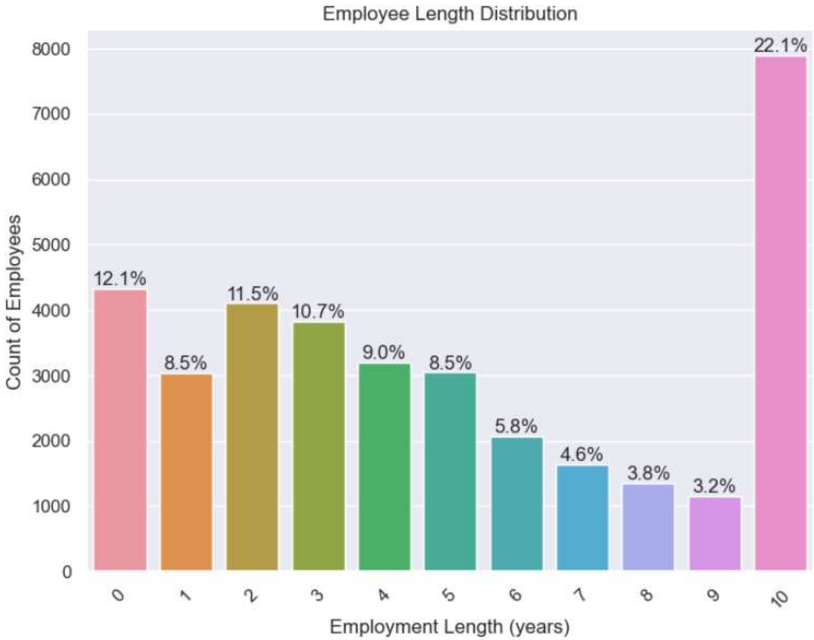
- The above count plot indicates that **85.43%** of the loans approved were **Fully Paid** and a **14.57%** of loans were **defaulted**

Term Distribution



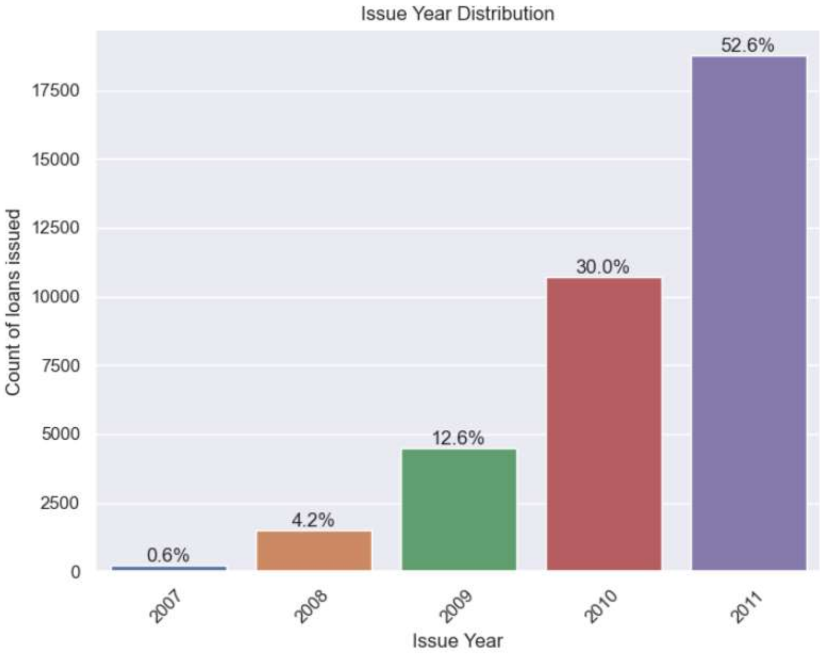
- The above pie chart shows that around **75.5%** of loans was taken under **36 months** term and **24.5%** under **60 months** term

Employee Length Distribution



- The above chart shows that around **22%** have employee length period of 10 years and only **3.2%** have employee length of 9 years.

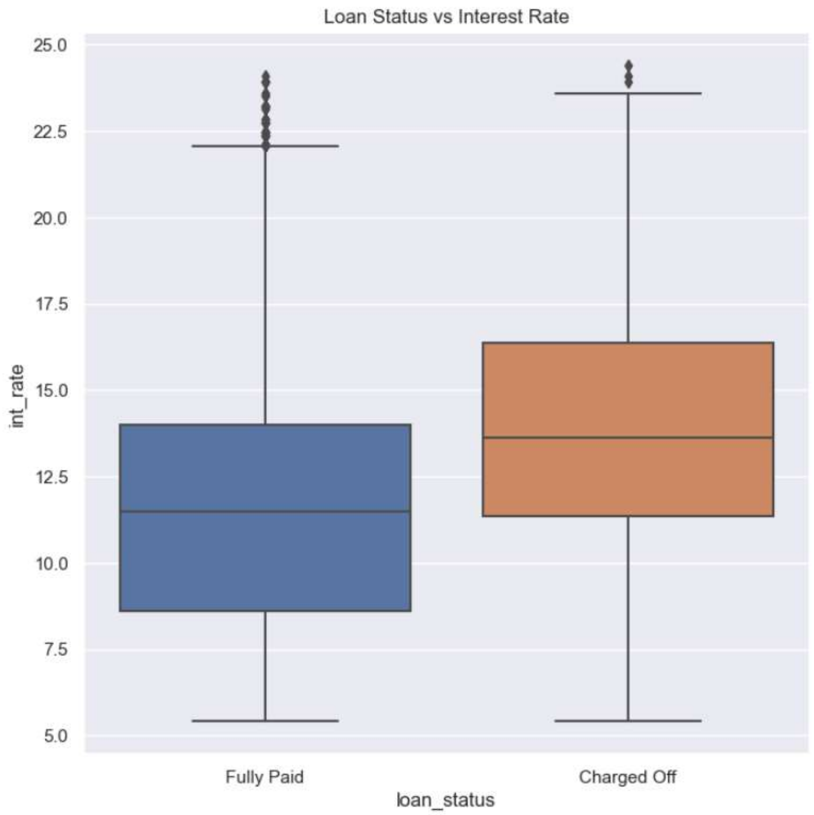
Issue Year Distribution



- The above chart shows that around year 2011 had highest issue count at **52.6%** where as 2007 had the least at **0.6%**.

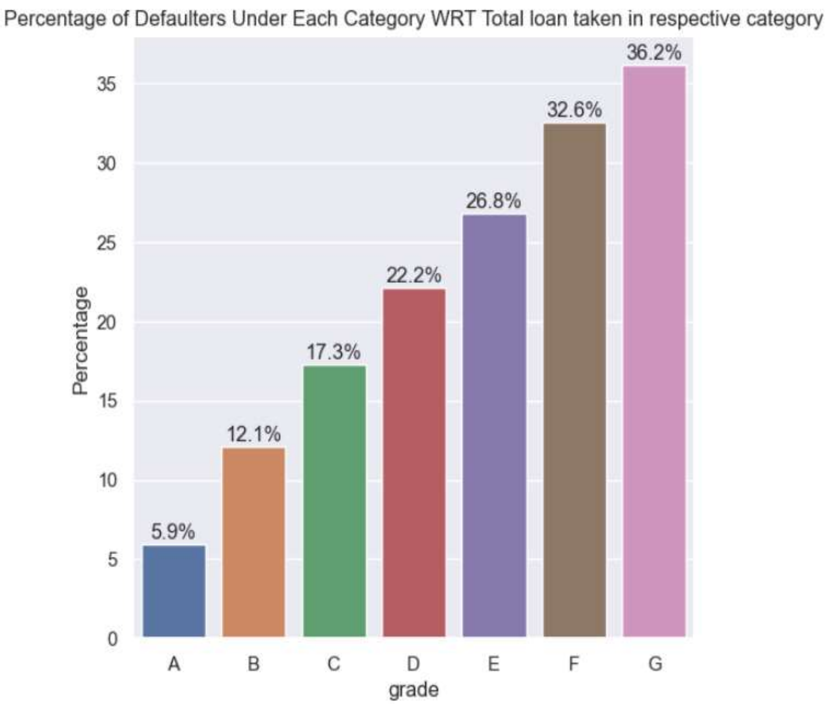
2. Segmented Univariate Analysis

Loan Status vs Interest Rate



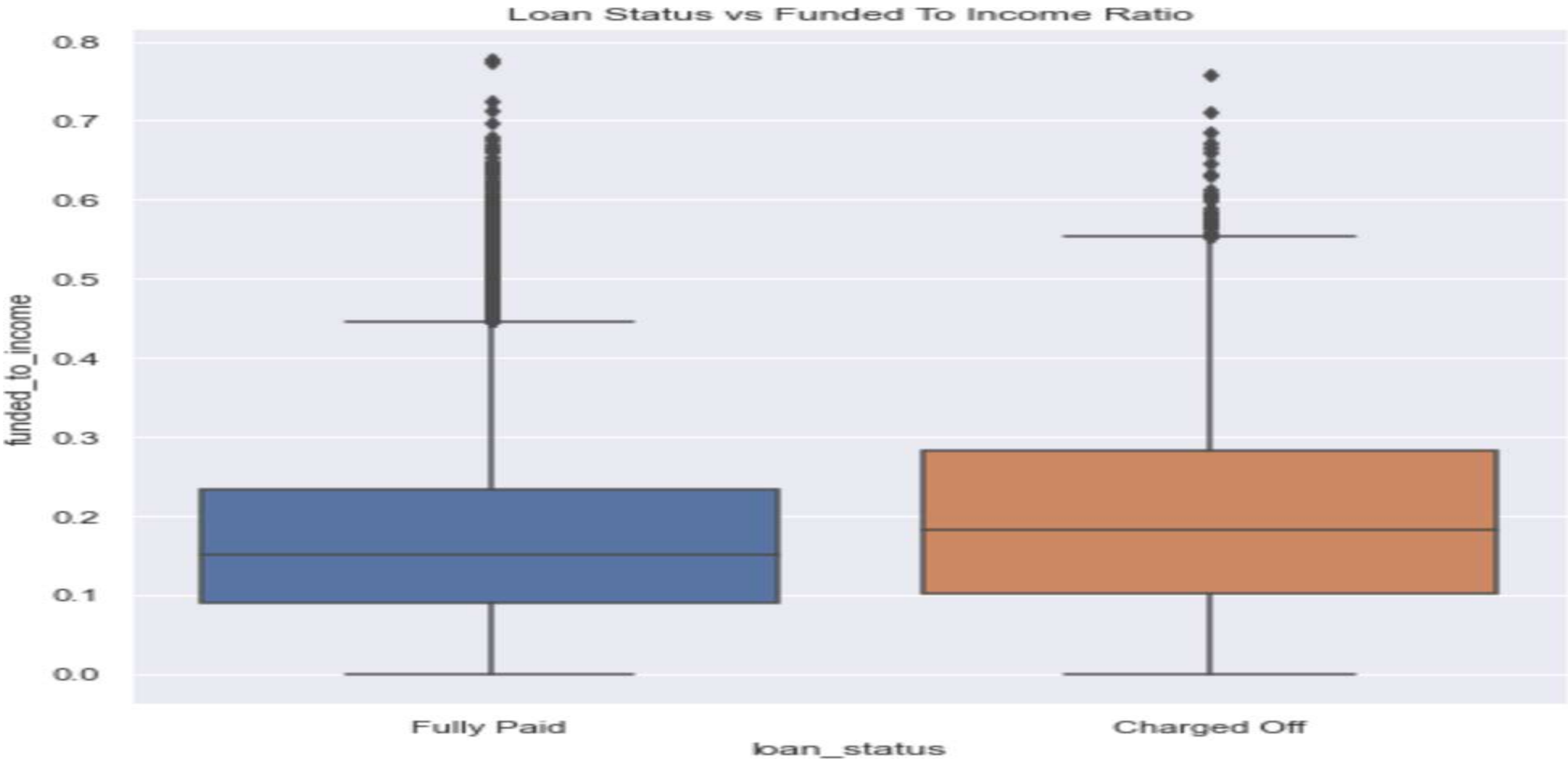
- Plotting a bar graph with the percentage of people defaulted across each grade category by deriving the percentage values of defaulters

Percentage of Defaulters Under Each Category wrt Total loan taken in respective category



- The above chart shows that around **22%** and above barplot gives a clear conclusion/insights that higher the grade at which the loans are taken, more the chance of defaulting.
- Around **36.2%** of the loan takers under **G** category has defulated
- The above box plot of Loan Status vs Interest Rate also indicate the same, that higher the interest rates higher the chance of defaulting
- Grades and Interest Rate is closely linked, as the interest rate increases, grades increase and vice versa, indicating that Grades is a bucketed version of interest rate
- Hence from these 2 plots, we can conclude that the loan taken under high interest rate or grades are tend to default more than the others.

Loan Status vs Funded to Income Ratio

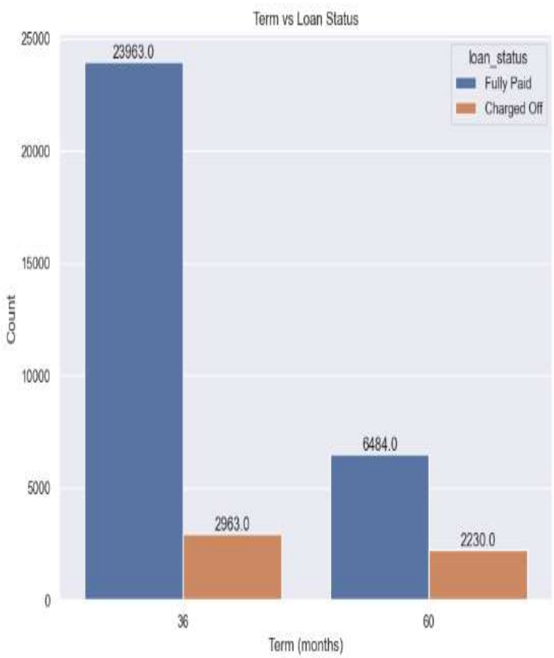


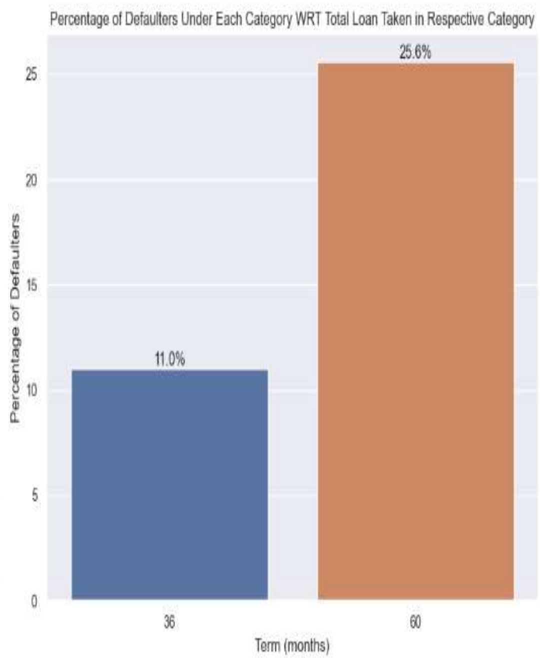
Inferences from Univariate Analysis

- Majority of the Loan Amount Requested/Sanctioned falls between **2500 - 10000** range.
- **85.4%** of the loans approved was Fully Paid and a 14.6% of loans was defaulted.
- Around **22%** of the total loan are taken by people who had 10+ Years of employee length, indicating that people tend to take loans more on a later stage of life.
- The loan issued increases drastically year by year, **2011** has over **50%** of the all issued loans. This can be due to several reasons.
    1. Life Getting Tougher Over Years
    2. Recession in 2011
    3. LC became popular over years
- **Loan Status vs Interest Rate** Box plot gives a strong indication that most of the defaulters tend to fall on higher interest rates when compared to non defaulters
- The **Percentage of Defaulters Under Each Category WRT Grade** barplot gives a clear conclusion/insights that higher the grade at which the loans are taken, more the chance of defaulting.
- Around **36%** of the loan takers under **G** category has defulated
- **Loan Status vs Funded To Income Ratio** Box plot gives a slight indication that most of the defaulters fall on high f_to_i ratio value, whereas majority of the Fully Paid are on the lower ratio end
- Grade/Sub Grade is linked to Interest rate, Higher the grade higher the interest rate
- Term Distribution Pie chart shows that around **75.6%** of loans was taken under **36 months term** and **24.4%** under **60 months** term.

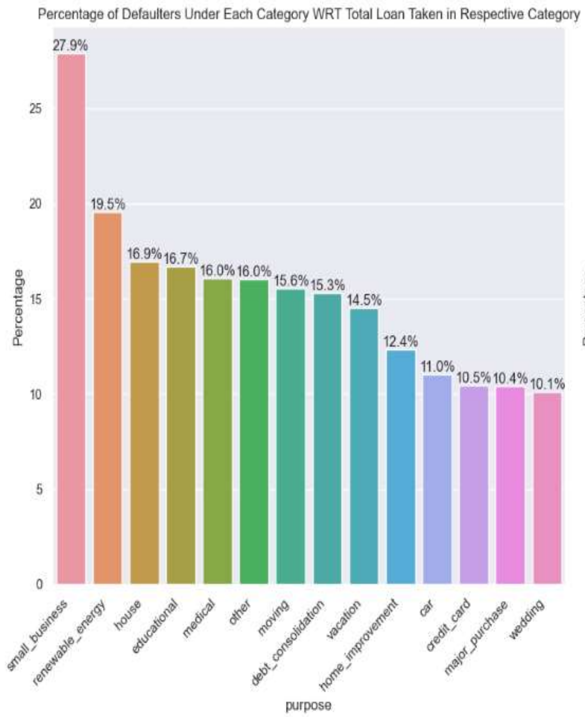
3. Bivariate Analysis

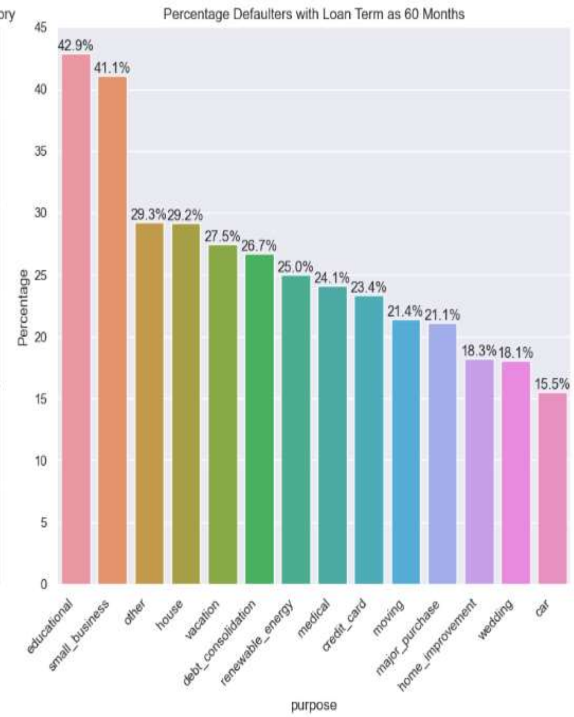
Percentage of Defaulters Under Each Category WRT Total loan taken in respective category



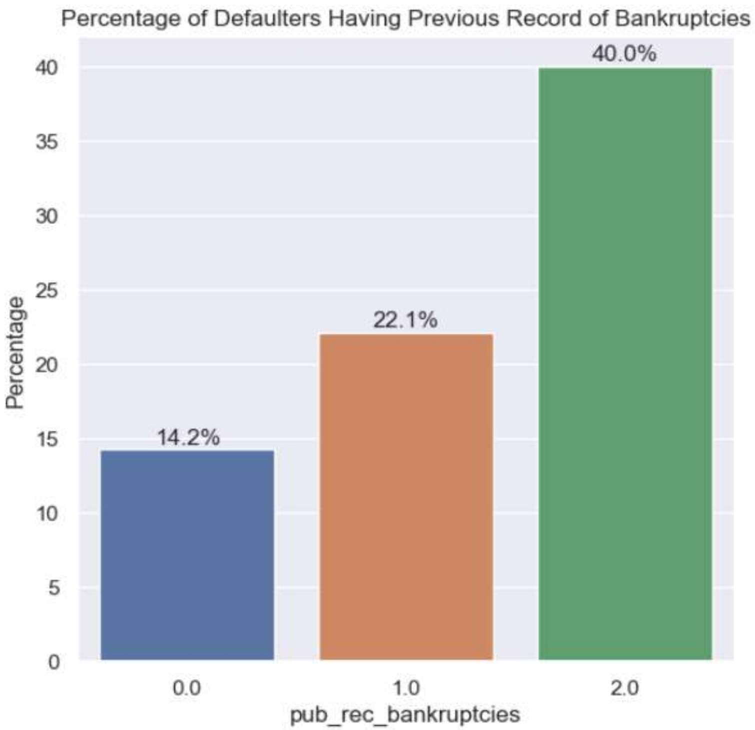


- From count plot for Term vs Loan Status, its clear that out of 8722 who opted for 60 Months as term 2230 has defaulted, means around 25.6 %, where as for those opted 36
- Months only 2966 out of 26953 defaulted , that's just 11%
- The Percentage of Defaulters Under Each Category WRT Total loan taken in respective category shows the same information with respect to percentage values in a barplot.
- This gives a clear indication that people opted for longer duration installments are going to default more, than people opted shorter duration
- So always insist on lending money for shorter
- duration.

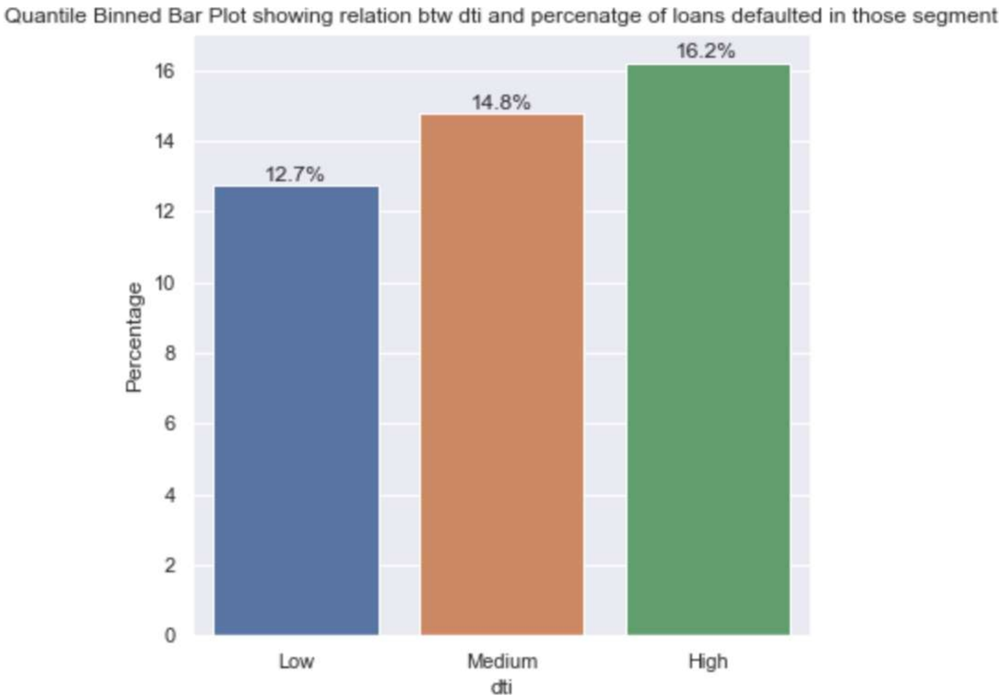




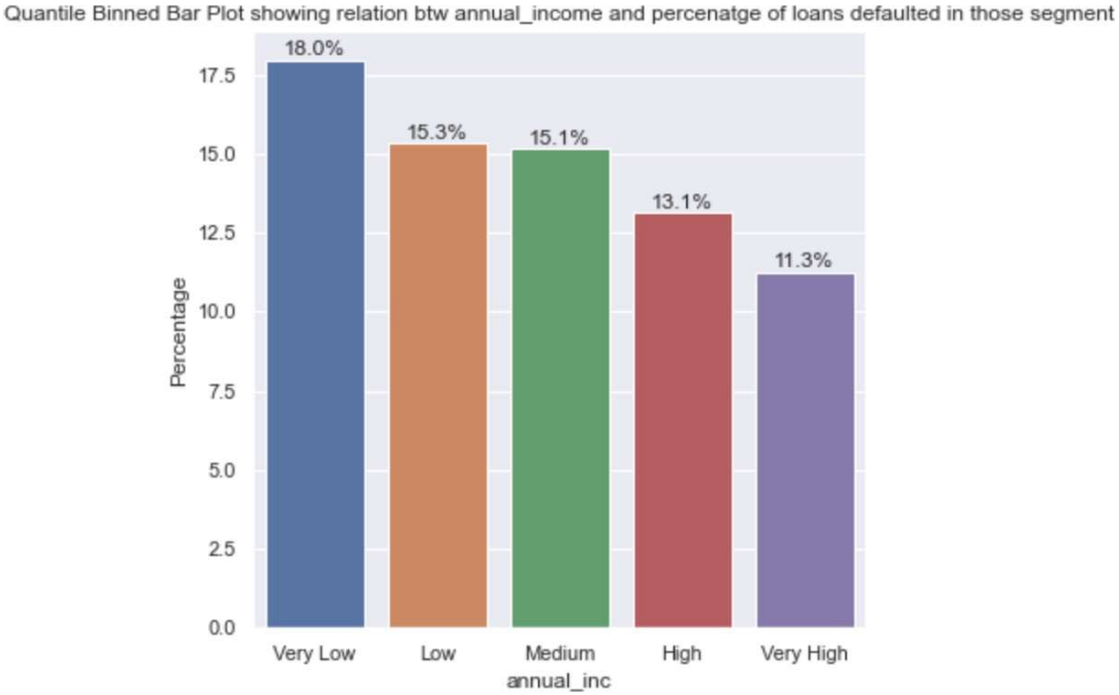
- The above analysis with **purpose vs loan_status** gives interesting insights.
- **27.9%** of loans taken for the purpose of **small_business** end up as defaulters. This might be because of the failure of the business
- Another insight is that for loans taken under **60 months as term** and purpose as **educational** and **small_business** shows very high default rates of about **42.9%**
- So lending loans for purposes such as **educational** and **small_business** for longer terms of **60 months** have a very huge chance of defaulting
- There is another inference that, majority of loans taken for small_business, are taken under **high interest**, **G Grade**, and longer term **60 Months**, resulting in high chances of defaulting.



- The above barplot of **pub_rec_bankruptcies vs percentage of defaulters** shows a indication that, people having previous record of bankruptcies tend to repeat that again in future.
- **40%** of those who take loans with a history of bankruptcies of **2** are tend to default.
- So its better not to provide loans for those having previous records of bankcruptcies.



- The above 2 plots indicating relationship between **dti and Percentage of defaulters in each segment**, binned in 2 different methods **Equal Width Binning** and **Quantile Binning** shows almost similar patterns.
- From analysis of 2 binning method one can come to a conclusion that as the **dti increases chances of defaulting also increases.**
- So lending out loans to **higher dti applications can be reduced.**



- The above plot indicating realtionship btw annual_income and Percentage of defaulters in each segment, binned using Quantile Binning technique shows some interseting analysis.
- Most of the **defaulters lie on the lowest income range**
- There is a trend that as the **annual_inc decreases chances of defaulting increases**

Inferences from Bivariate Analysis

1.term vs loan_status

- People opted for longer duration installments i.e. **60 months** are going to default more, than people opted shorter duration i.e. **36 months**
- From **Term vs Loan Status** Analysis, its clear that out of **8716** who opted for **60 Months** as term **2230** has defaulted, means around **25.6 %**, where as for those opted **36 Months** only **2963** out of **26953** deafulted, thats just **11%**

2.purpose vs loan_status

- From **Purpose vs Loan_status** analysis, its clear that **27.9%** of loans taken for the purpose of **small_business** end up as defaulters. This might be because of the failure of the business.
- Another insight is that for loans taken under **60 months as term** and purpose as **educational** and **small_business** shows very high default rates of about **42.9%**

3.dti vs loan_status

- From **dti vs Loan Status** analysis, binned in 2 different methods **Equal Width Binning** and **Quantile Binning** shows almost similar patterns, that as the **dti increases chances of defaulting also increases**

4.funded_to_income vs loan_status

- Similar analysis was made from realtionship btw **funded_to_income vs loan_status**, binned in 2 different methods **Equal Width Binning** and **Quantile Binning**.
- From analysis of 2 binning method one can come to a conclusion that as the **funded_to_income increases chances of defaulting also increases**
- From plot generated using **Equal Width Binning** for **funded_to_income vs loan_status**, its clear that almost **31.1%** of loans got defaulted whose funded_to_income ratio was above **0.52**

5.annual_inc vs loan_status

- From **Annual Income vs Loan Status** analysis, it was found that as **annual_inc decreases chances of defaulting increases**

6.funded_amnt_inv vs loan_status

- From **Funded_amnt_inv vs loan_status** analysis, it was found that as **funded amount by investors increases chances of defaulting increases**

7.pub_rec_bankruptcies vs percentage of defaulters

- **pub_rec_bankruptcies vs percentage of defaulters** shows a indication that, people having previous record of bankruptcies tend to repeat that again in future.
- Around **40%** of those who take loans with a history of bankruptcies of **2** are tend to default.

8.difference of funded_amnt and funded_amnt_inv vs loan_status

- Analysising realtionship btw the **difference of funded_amnt and funded_amnt_inv**, and Percentage of defaulters in each segment, binned using Equal Width Binning technique shows very interseting analysis.
- So if the difference btw **approved amount from LC and amount funded by investors** increases, means the tendency for that loan to deafult is very high
- For loans which had a difference in **approved amount from LC and amount funded by investors** greater than **21.6K**, around **46.5%** of such loans was defaulted

Thank You !