Lending Club Case Study

Group Members :
Akankshya Abhilipsa
Labeeb Ali Koleth

Problem Statement

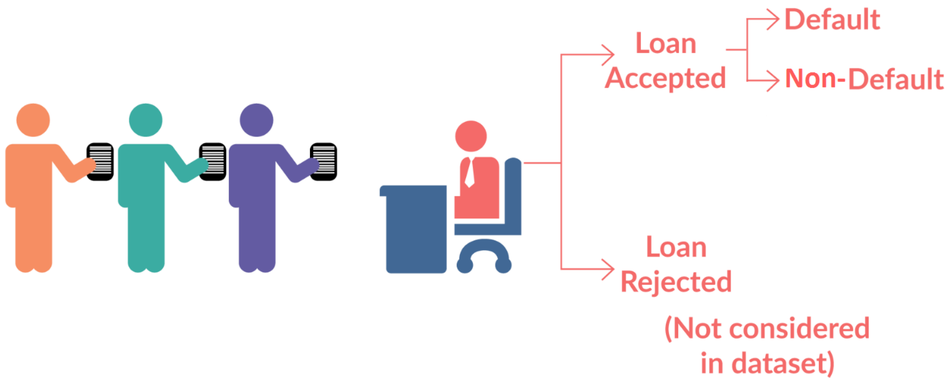
A consumer finance company specialises in lending various types of loans to urban customers. When the company receives a loan application, it has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- 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



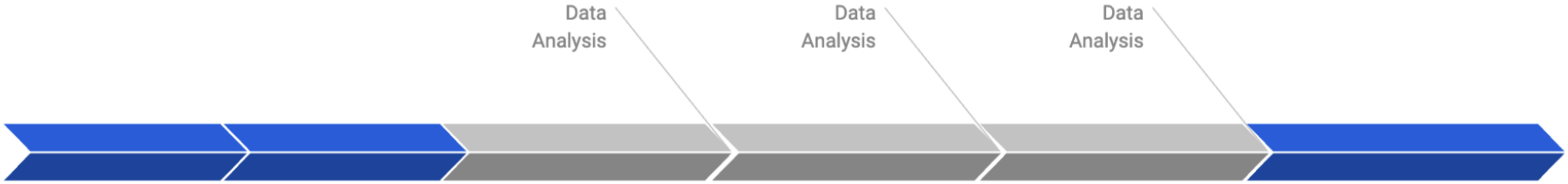
When a person applies for a loan, there are two types of decisions that could be taken by the company:

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

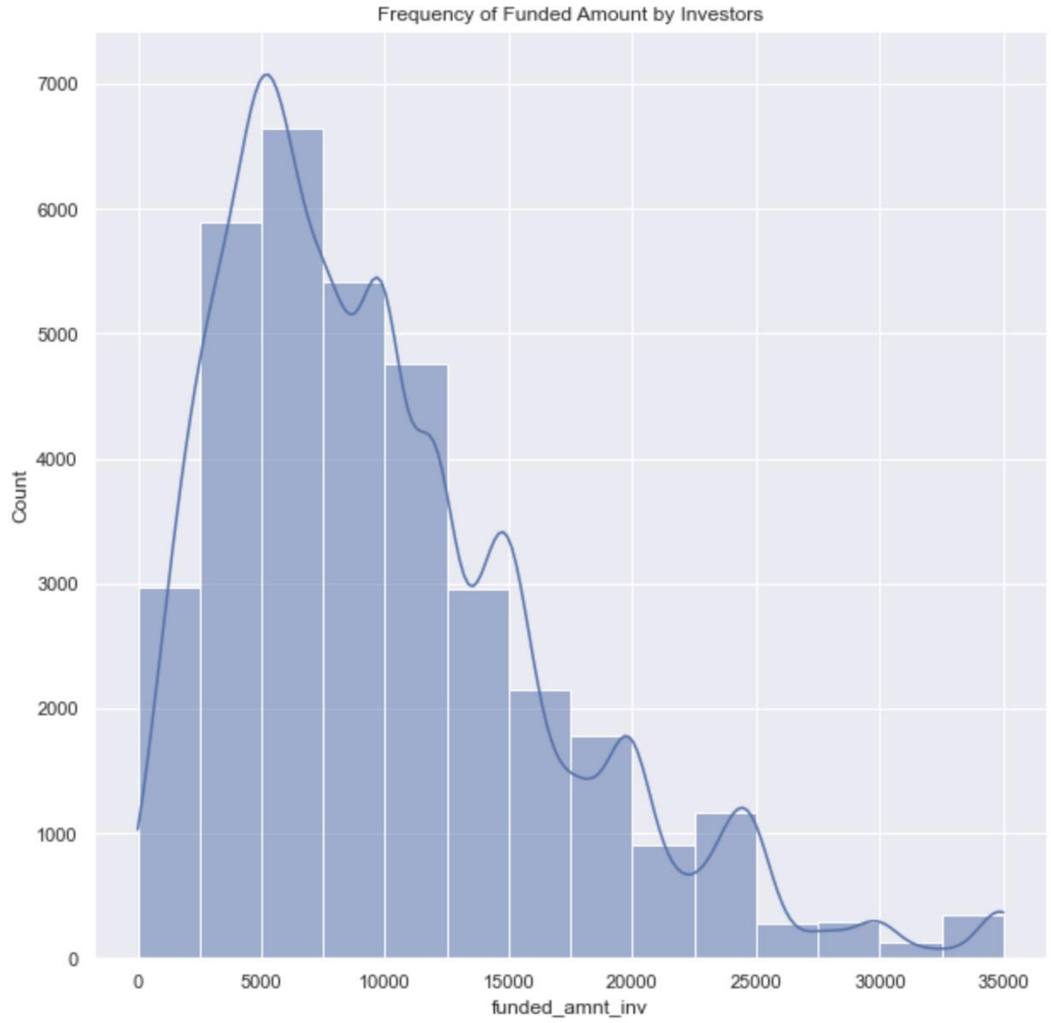
Analysing the two variable behaviour like term and loan status with respect to loan amount.

Recommendations

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

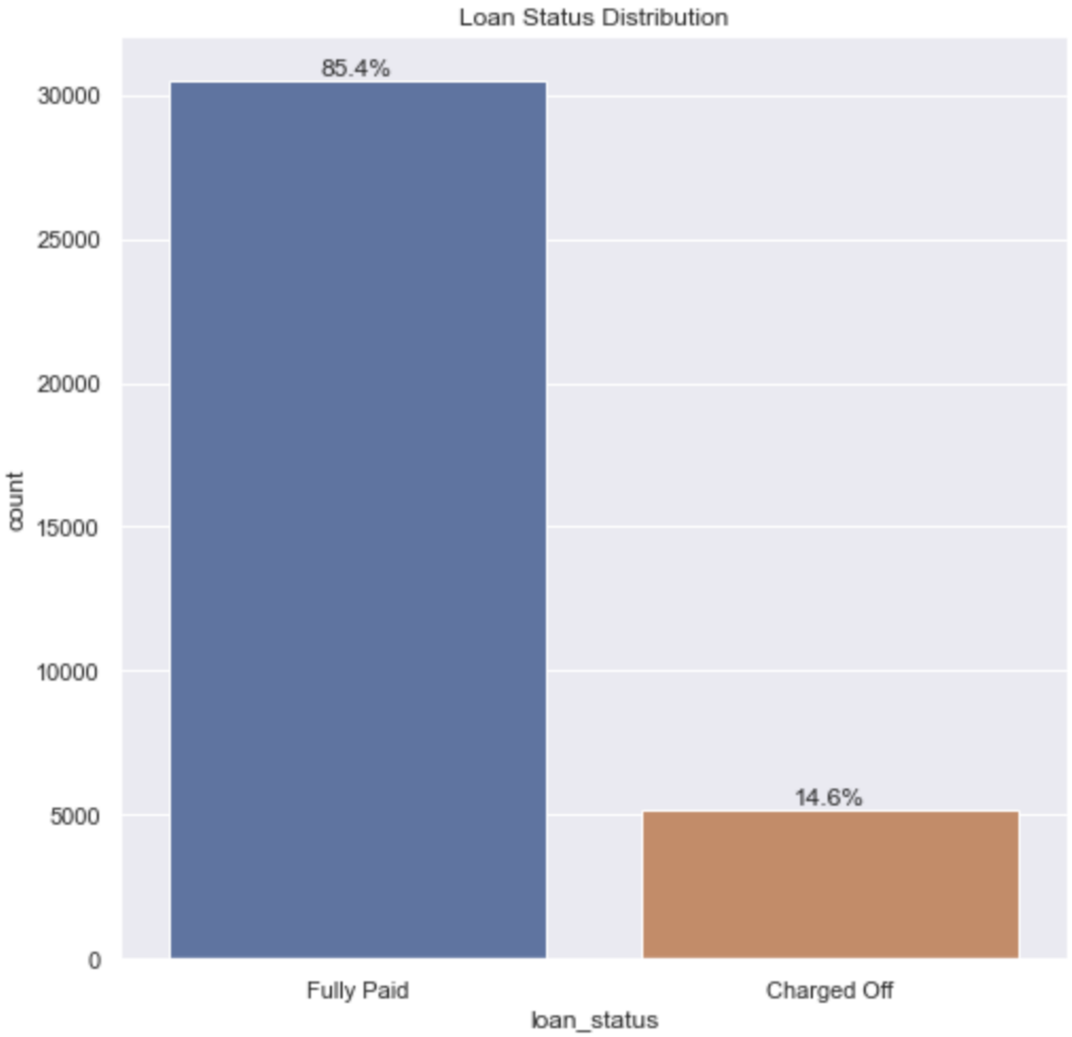
Univariate Analysis

1. Frequency of Funded Amount by Investors



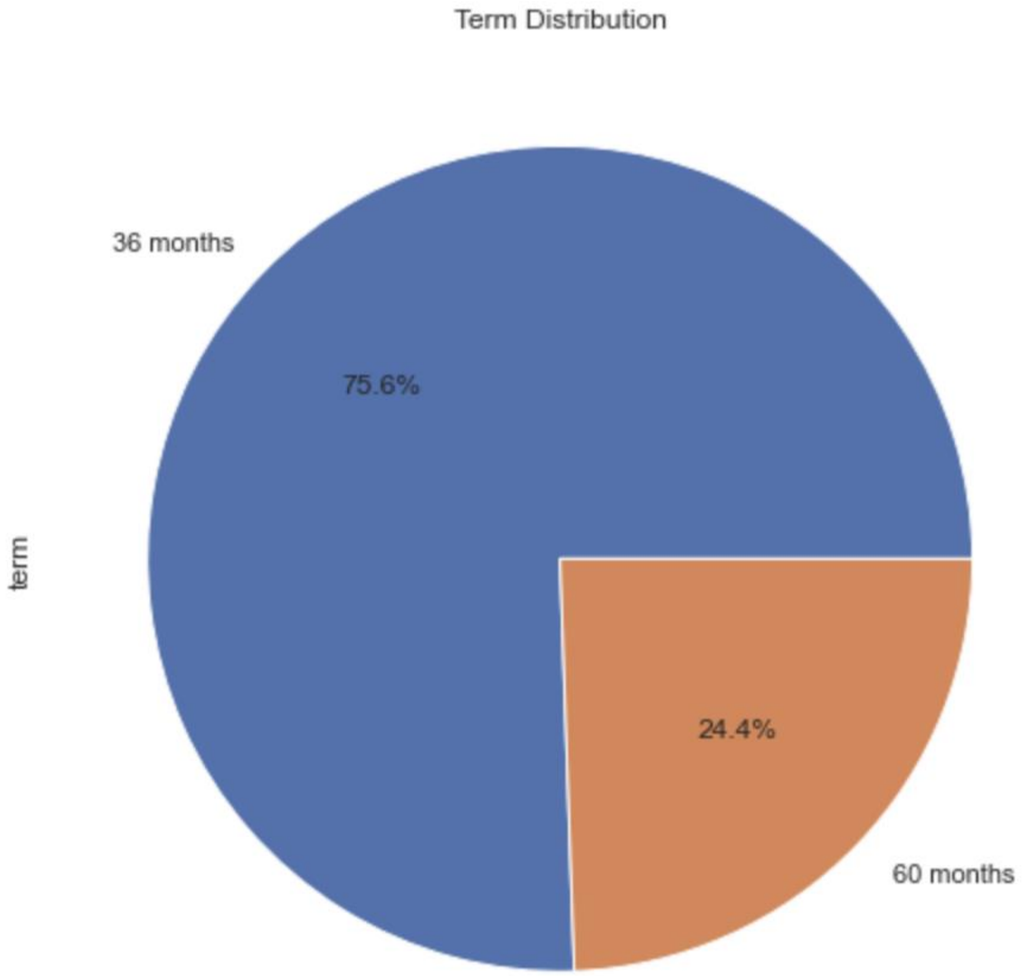
- The histogram indicates that most of the funding was between **2500 - 10000** range.

2. Loan Status Distribution



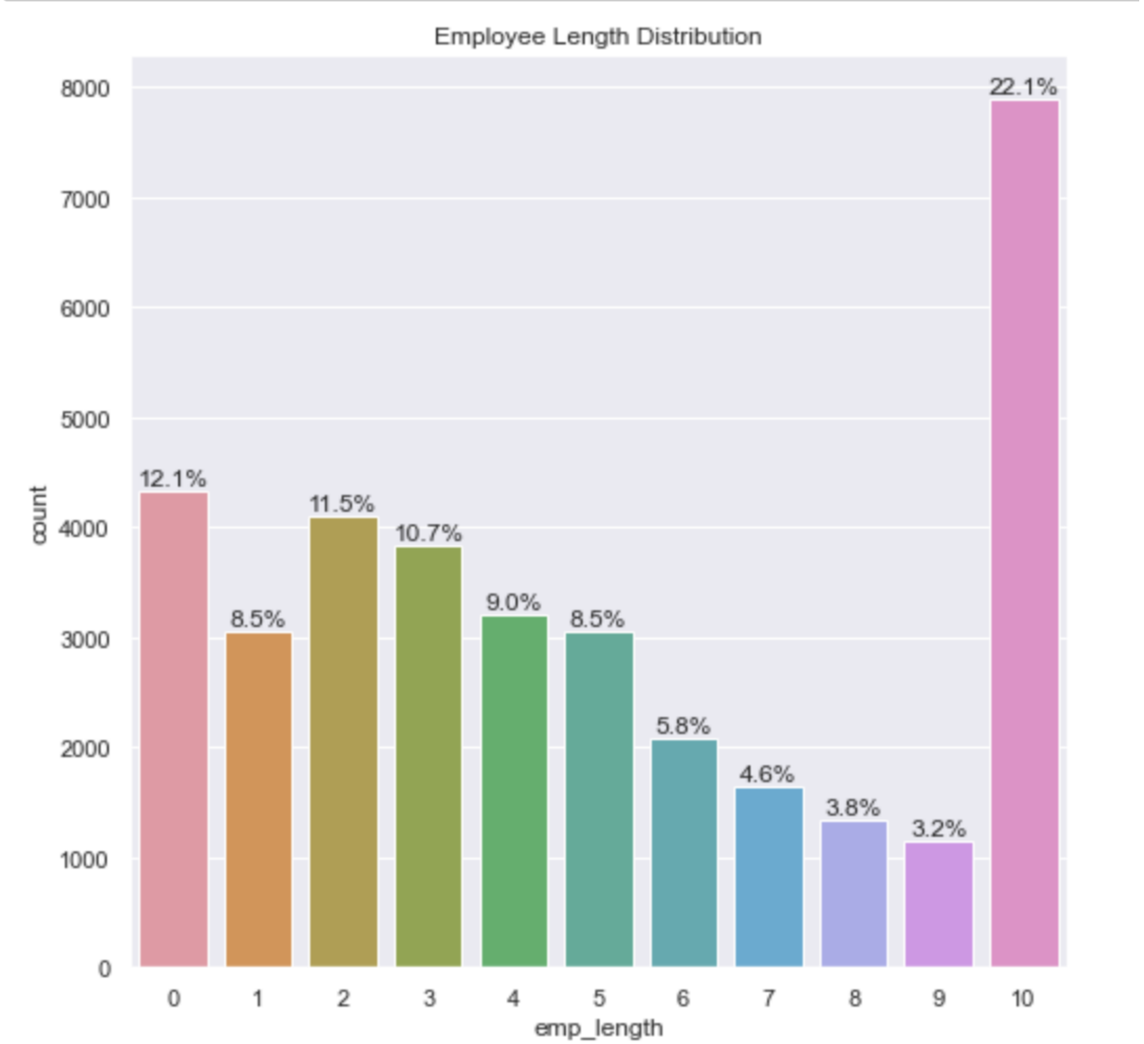
- The count plot indicates that **85.4%** of the loans approved was **Fully Paid** and a **14.6%** of loans was **defaulted**

3. Term Distribution



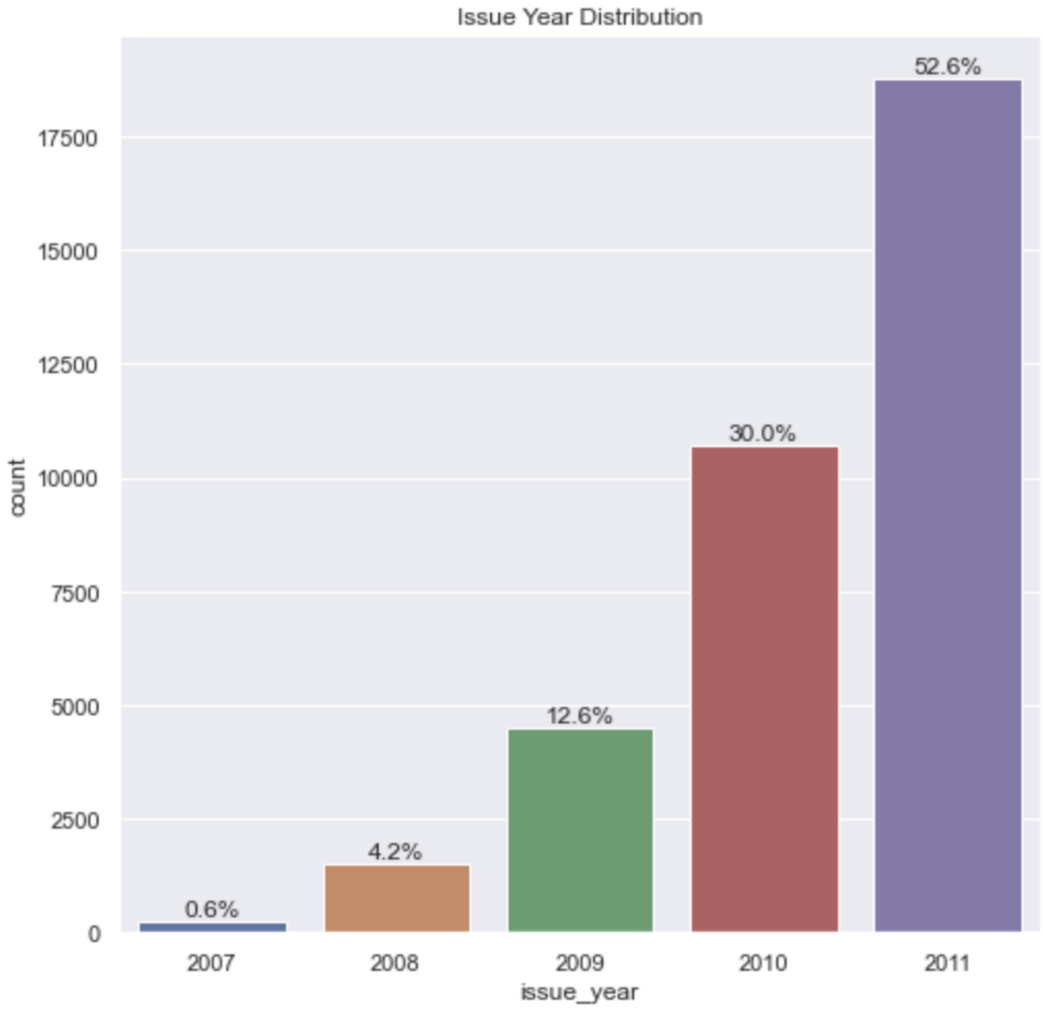
- The pie chart shows that around **75.6%** of loans was taken under **36 months** term and **24.4%** under **60 months** term

4. Employee Length Distribution



- 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.

5. Issue Year Distribution



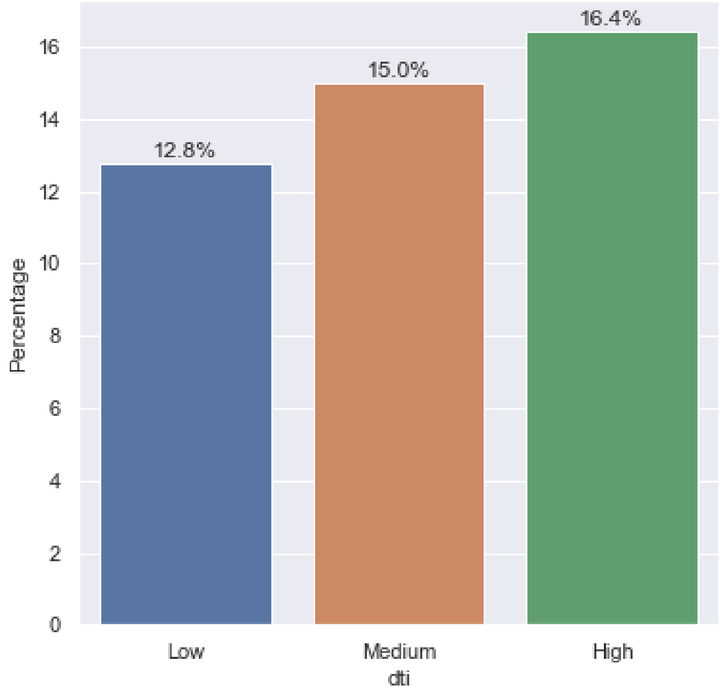
- The loan issued increased drastically year by year, 2011 has over 50% of all issued loans. This can be due to several reasons.
    - Life Getting Tougher Over Years
    - Recession in 2011
    - LC became popular over the years

Segmented Univariate Analysis

1. Relation btw segmented dti vs defaulted percentage in the respective category

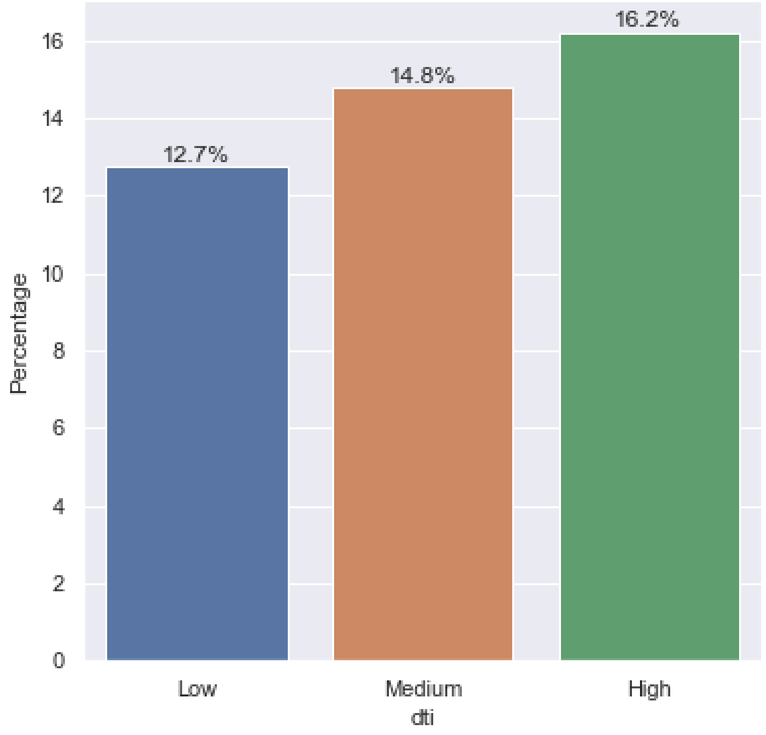
```
Bins considered in Equal Width Binning method are :
0 - 10 (Low)
10 - 20 (Medium)
20 - 30 (High)
```

Equal Width Binned Bar Plot showing relation btw dti and percenatge of loans defaulted in those segment




```
Bins considered in Equal Frequency Binning (Quantile Binning) are :
0.0 - 10.22 (Low)
10.22 - 16.89 (Medium)
16.89 - 29.99 (High)
```

Quantile Binned Bar Plot showing relation btw dti and percenatge of loans defaulted in those segment

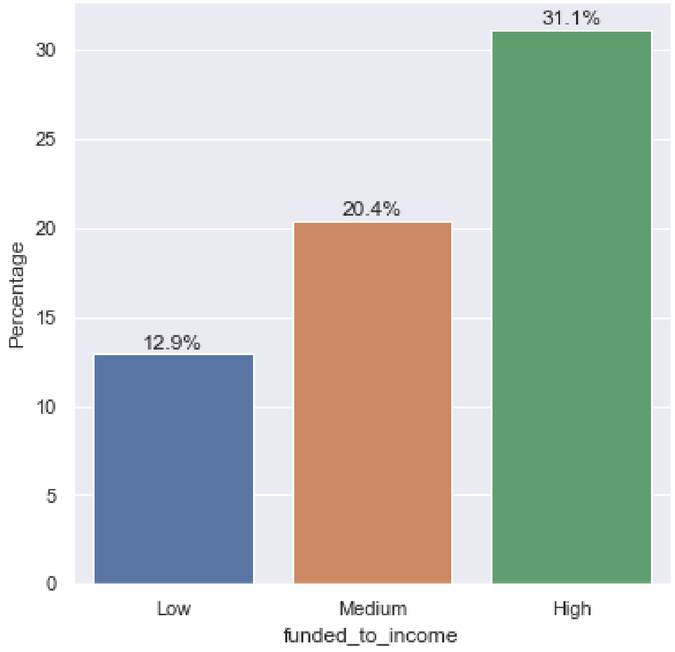


- The above 2 plots indicate a relationship btw **dti and the Percentage of defaulters in each segment**, binned in 2 different methods **Equal Width Binning** and **Quantile Binning** show almost similar patterns.
- From the analysis of 2 binning methods one can conclude that as the **dti increases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

2. Relation btw segmented funded_to_income vs defaulted percentage in the respective category

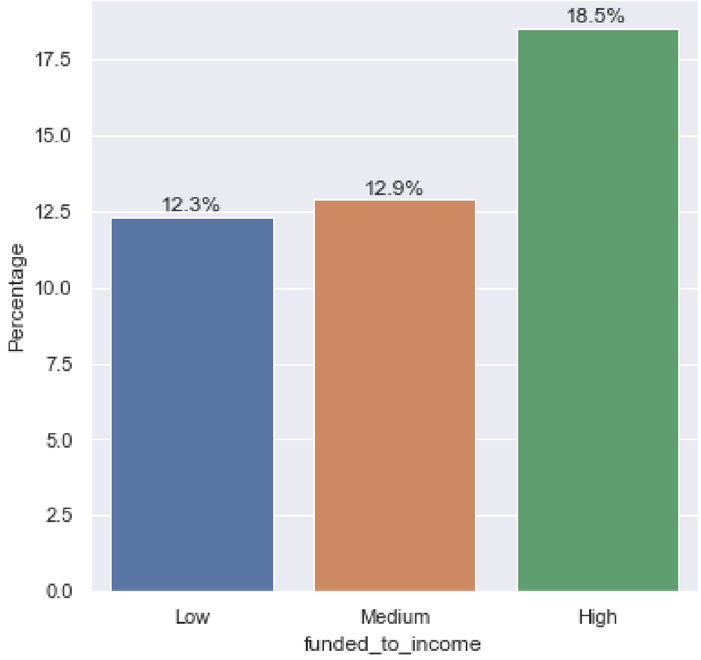
```
Bins considered in Equal Width Binning method are :
0.00 - 0.26 (Low)
0.26 - 0.52 (Medium)
0.52 - 0.78 (High)
```

Equal Width Binned Bar Plot showing relation btw funded_to_income and percenatge of loans defaulted in those segment




```
Bins considered in Equal Frequency Binning (Quantile Binning) are :
0.00 - 0.11 (Low)
0.11 - 0.21 (Medium)
0.21 - 0.78 (High)
```

Quantile Binned Bar Plot showing relation btw funded_to_income and percenatge of loans defaulted in those segment



- The above 2 plots indicating the relationship btw **funded_to_income and the Percentage of defaulters in each segment**, binned in 2 different methods **Equal Width Binning** and **Quantile Binning** show almost similar patterns.
- From the analysis of 2 binning methods one can conclude that as the **funded_to_income increases chances of defaulting also increases**
- From the plot generated using Equal Width Binning, it’s clear that almost **31.1% of loans got defaulted whose funded_to_income ratio was above 0.52**
- So lending out loans to **higher funded_to_income ratio applications can be reduced**

Inferences from Univariate Analysis

- Majority of the Loan Amount Requested/Sanctioned falls between **2500 - 10000** range.
- **85.4%** of the loans approved were **Fully Paid** and **14.6%** of loans were **defaulted**.
- Around **22%** of the total loans are taken by people who had **10+ Years of employee length**, indicating that people tend to take loans more at a later stage of life.
- Term Distribution Pie chart shows that around **75.6%** of loans were taken under **36 months term** and **24.4%** under **60 months term**
- The loan issued increased drastically year by year, **2011 had over 50%** of 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 the years

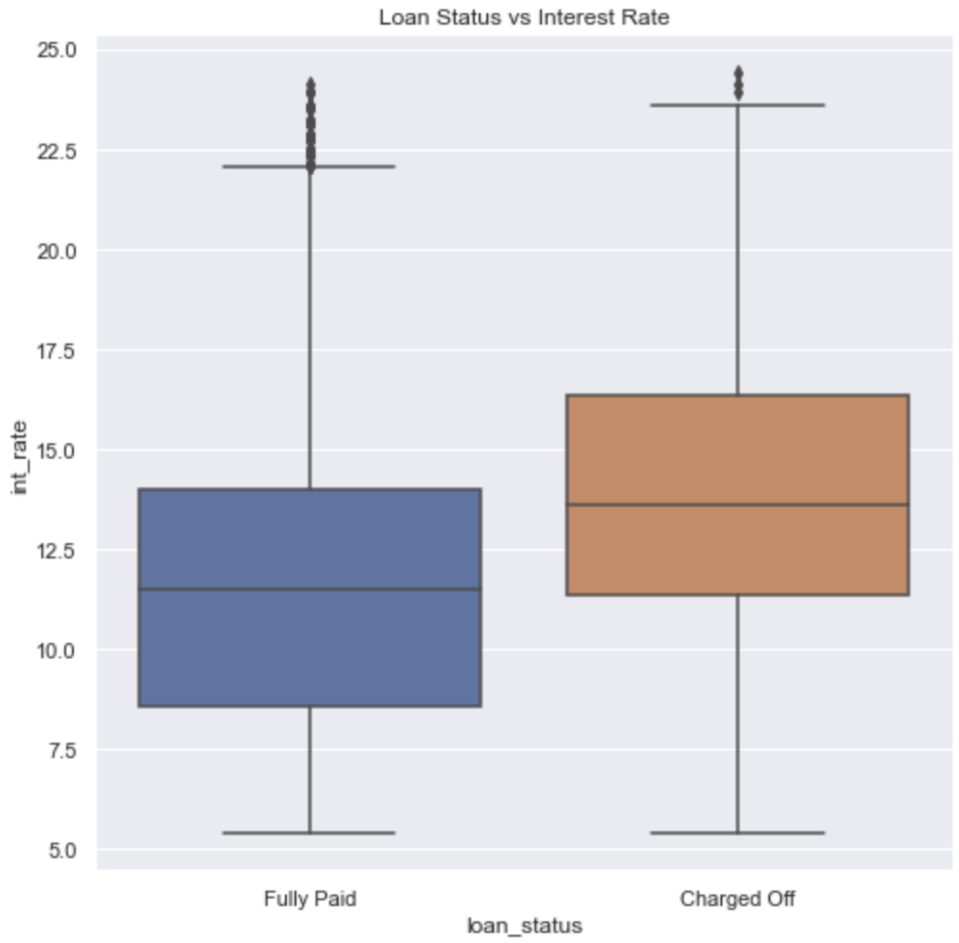
Continuation :

1) dti vs loan_status

- From dti vs Loan Status analysis, binned in 2 different methods Equal Width Binning and Quantile Binning show almost similar patterns, that as the **dti increases chances of defaulting also increases**
- 2) funded_to_income vs loan_status
- Similar analysis was made from the relationship btw funded_to_income vs loan_status, binned in 2 different methods Equal Width Binning and Quantile Binning.
- From the analysis of 2 binning methods one can conclude that as the **funded_to_income increases chances of defaulting also increases**
- From the plot generated using Equal Width Binning for funded_to_income vs loan_status, its clear that almost **31.1% of loans defaulted whose funded_to_income ratio was above 0.52**
- Loan Status vs Funded To Income Ratio, the Box plot gives a slight indication that most of the defaulters fall on a high f_to_i ratio value, whereas the majority of the Fully Paid is on the lower ratio end

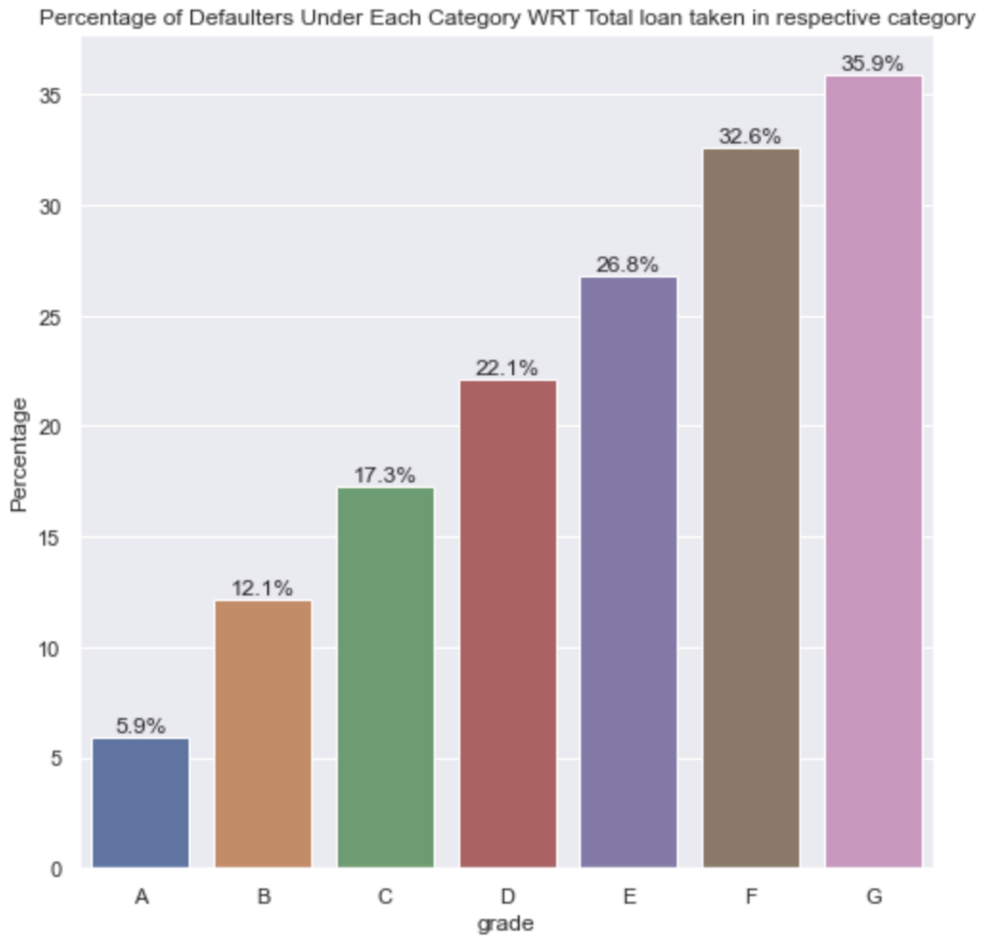
Bivariate Analysis

1. Loan Status vs Interest Rate



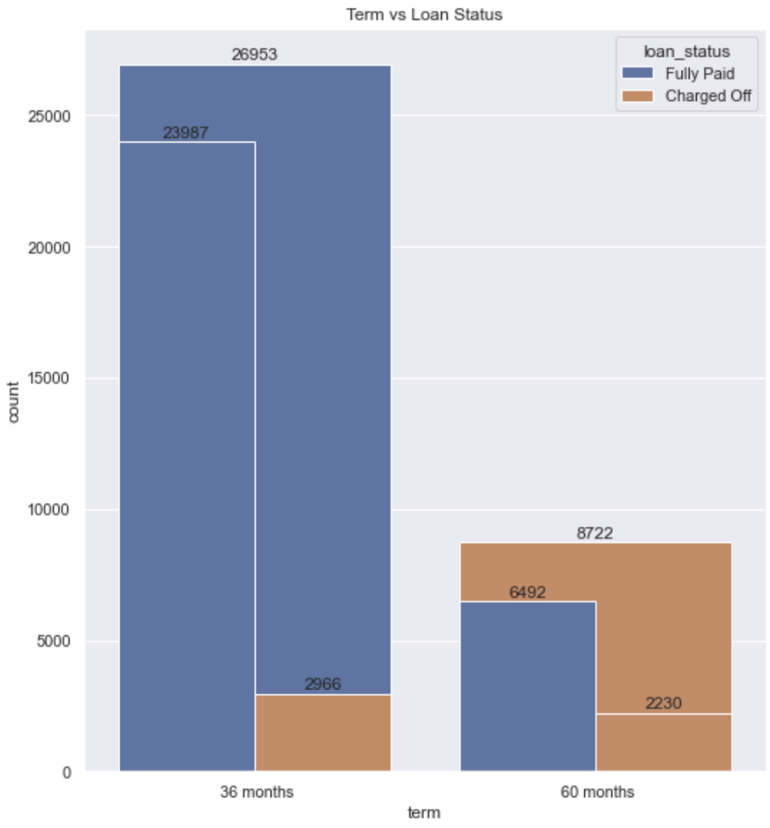
- Loan Status vs Interest Rate Box plot gives a strong indication that most of the **defaulters tend to fall on higher interest rates when compared to non-defaulters.**

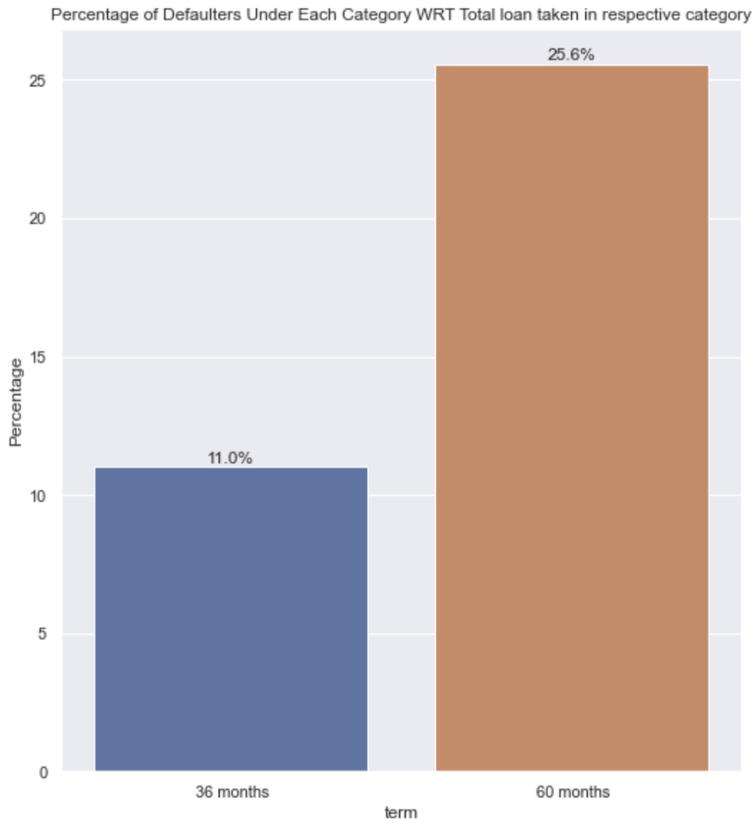
Percentage of Defaulters Under Each Grade wrt Total loan taken in respective grades



- The bar plot gives a clear conclusion/insight that the **higher the grade at which the loans are taken, the more the chance of defaulting**.
- Around **36**% of the loan takers under the **G** category have defaulted
- The above **box plot of Loan Status vs Interest Rate** also indicates the same, that the **higher the interest rates higher the chance of defaulting**
- Grades and Interest Rate are closely linked, as the interest rate increases, grades increase and vice versa, indicating that **Grades are a bucketed version of the interest rate**
- Hence from these 2 plots, we can conclude that the loans taken **under high interest rates or grades tend to default more than the others**.

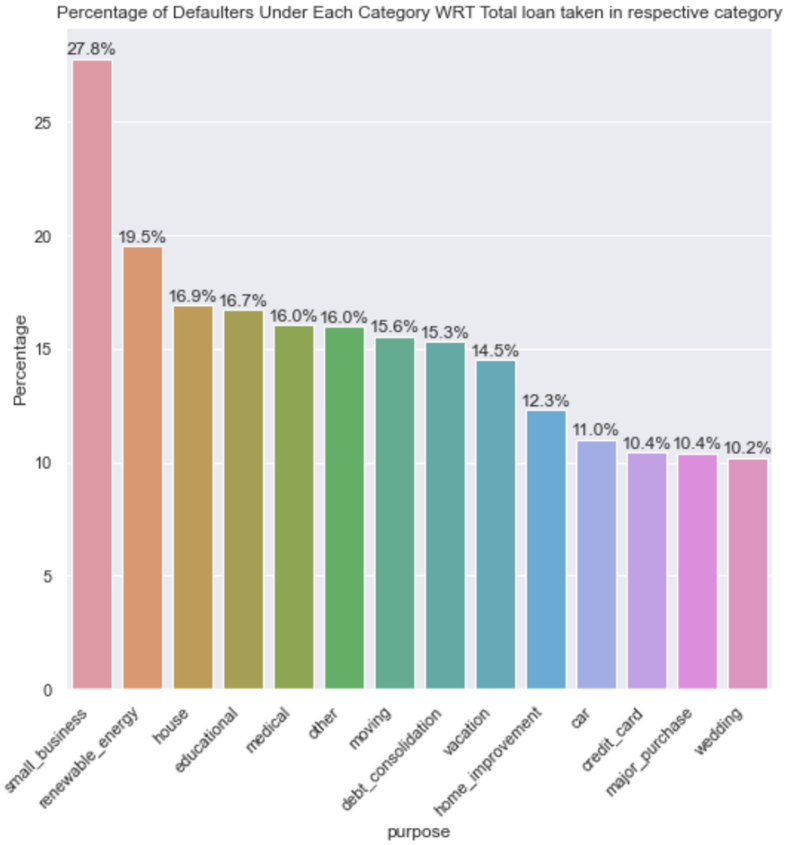
2. Term vs Loan Status (Count/Percentage)

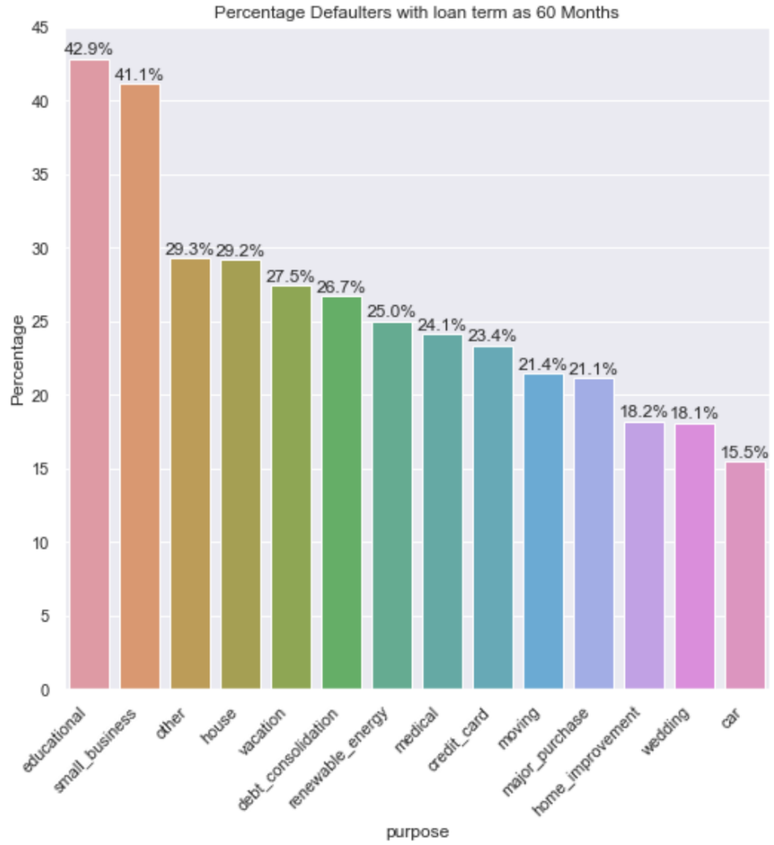




- From the count plot for **Term vs Loan Status**, it is clear that out of **8722** who opted for **60 Months** as term **2230** has defaulted, which means around **25.6 %**, whereas for those who opted **36 Months** only **2966** out of **26953** defaulted, that's just **11%**
- The **Percentage of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

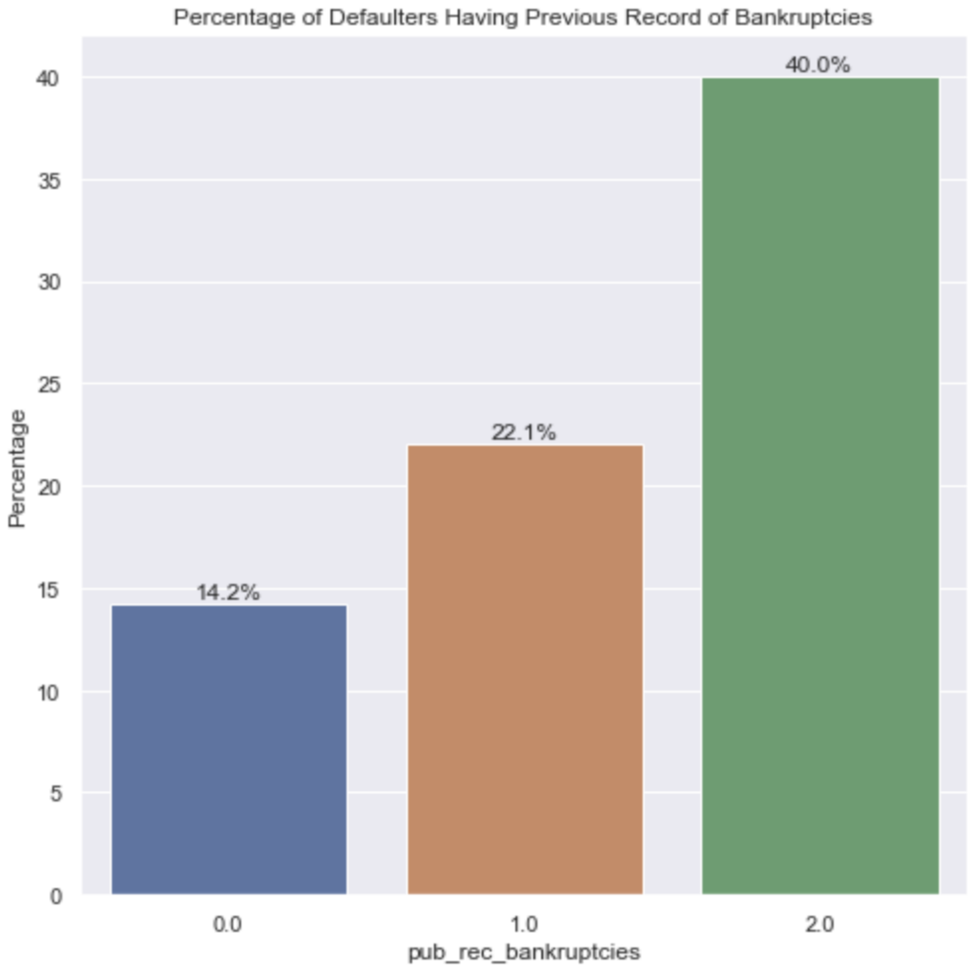
3. Purpose vs Loans Defaulted Under that purpose





- The analysis, with **purpose vs loan_status** gives interesting insights.
- **27.8%** of loans taken for **small_business** end up as defaulters. This might be because of the failure of the business
- Another insight is that loans taken under **60 months as term** and purpose as **educational** and **small_business** show very high default rates of about **42%**
- So lending loans for purposes such as **education** and **small_business** for longer terms of **60 months** have a very huge chance of defaulting
- There is another inference that, the 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.

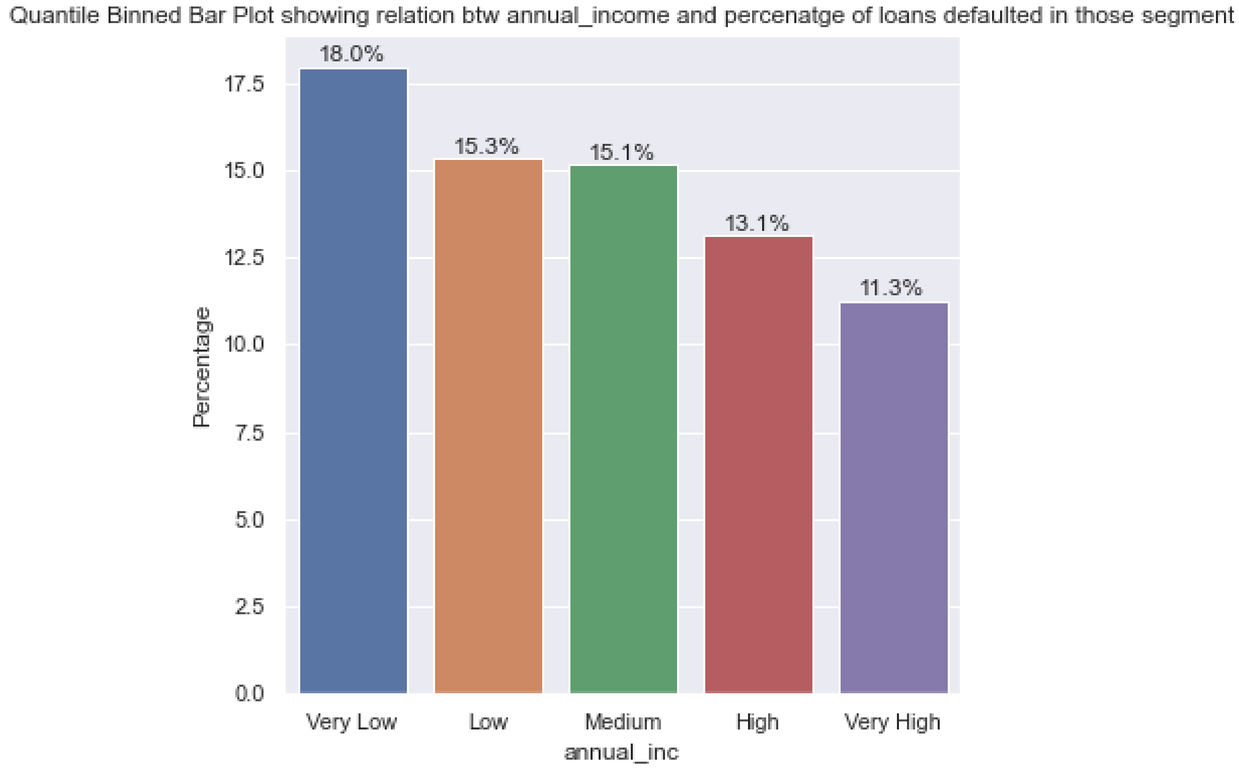
4. Record of Bankruptcies vs Loans Defaulted In Percentage



- The bar plot of **pub_rec_bankruptcies vs percentage of defaulters** shows an indication that people having previous records of bankruptcies tend to repeat that in the future.
- **40%** of those who take loans with a history of bankruptcies of **2** tend to default.
- So it’s better not to provide loans for those having previous records of bankruptcies.

5. Annual Income vs Loans Defaulted In Percentage

```
Bins considered in Equal Frequency Binning (Quantile Binning) are :
4000 - 37000 (Very Low)
37000 - 50000 (Low)
50000 - 64661 (Medium)
64661 - 84996 (High)
84996 - 142000 (Very High)
```

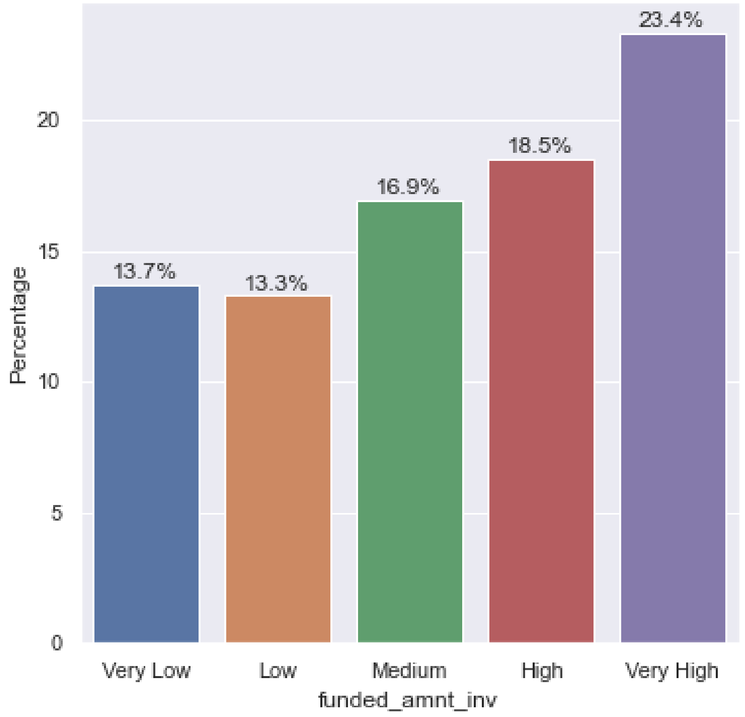


- The bar plot indicating the relationship btw **annual_income and the Percentage of defaulters in each segment**, binned using the Quantile Binning technique shows some interesting analysis.
- Most of the **defaulters lie in the lowest income range**
- There is a trend that as the **annual_inc decreases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

6. Funded Amount By Investors vs Loans Defaulted In that segment in Percentage

```
Bins considered in Equal Width Binning method are :
0 - 7000 (Very Low)
7000 - 14000 (Low)
14000 - 21000 (Medium)
21000 - 28000 (High)
28000 - 35000 (Very High)
```

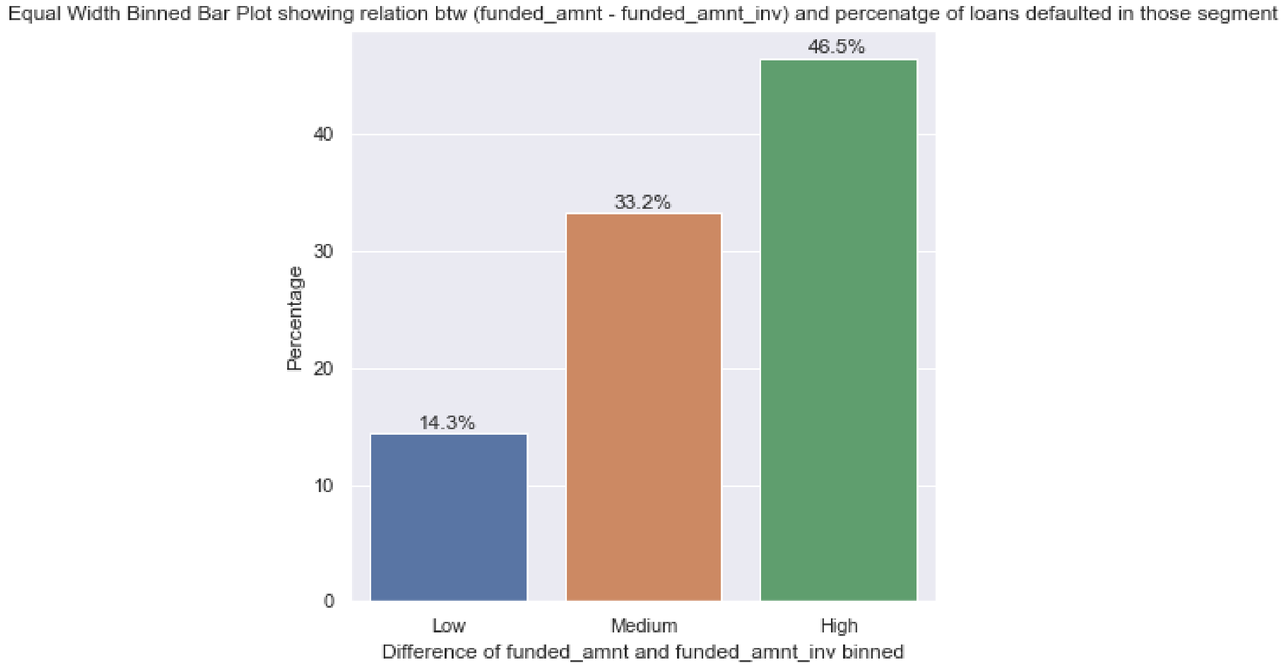
Equal Width Binned Bar Plot showing relation btw funded_amnt_inv and percenatge of loans defaulted in those segment



- The plot indicating the relationship btw **funded_amnt_inv and the Percentage of defaulters in each segment**, binned using the Equal Width Binning technique shows some interesting analysis.
- Most of the **High Funded loans are the ones most tend to default**
- There is a trend that as the **funded amount by investors increases chances of defaulting increases**

7. Difference of Funded Amount and Funded amount by investors vs Loans Defaulted In that segment in Percentage

```
Bins considered in Equal Width Binning method are :
0 - 10825 (Low)
10825 - 21650 (Medium)
21650 - 32475 (High)
```



- The plot indicates relationship btw **difference of funded_amnt and funded_amnt_inv, and the Percentage of defaulters in each segment**, binned using the Equal Width Binning technique shows a very interesting analysis.
- So if the difference btw the **approved amount from LC and the amount funded by investors** increases, means the tendency for that loan to default 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

Inferences from Bivariate Analysis

1) **Loan Status vs Interest Rate**

- 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** bar plot gives a clear conclusion/insight that the higher the grade at which the loans are taken, the more the chance of defaulting.
- Around **36%** of the loan takers under **G** category has defaulted
- Grade/Sub Grade is linked to Interest rate, The Higher the grade higher the interest rate

- 2) **Term vs loan_status**
- People who opted for longer duration installments i.e. **60 months** are going to default more, than people who opted for shorter duration i.e. **36 months**
- From Term vs Loan Status Analysis, it's clear that out of **8722** who opted for **60 Months** as term **2230** has defaulted, which means around **25.6 %**, whereas, for those who opted **36 Months** only **2966** out of **26953** defaulted, that's just **11%**

- 3) **Purpose vs loan_status**
- From Purpose vs Loan_status analysis, it's clear that **27.8%** of loans taken for **small_business** end up as defaulters. This might be because of the failure of the business.
- Another insight is that loans taken under **60 months as term** and purpose as **educational** and **small_business** show very high default rates of about **42%**

Continuation :

4) **Annual Income vs Loans Defaulted In that segment in Percentage**

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

5) **Funded Amount By Investors vs Loans Defaulted In that segment in Percentage**

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

6) **Record of Bankruptcies vs Loans Defaulted In that segment in Percentage**

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

7) **Difference of Funded Amount and Funded amount by investors vs Loans Defaulted In that segment in Percentage**

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

upGrad

Thank You !