# GRAMENER CASE STUDY

# SUBMISSION

Group Members:

1. Arpit Joshi
2. Harsh Tiwari
3. Harshit Tiwari
4. Vinay Dharwadkar

# Business Objective

The 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.

Thus, lending loans to ‘risky’ applicants is a Big source of Financial Loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

In other words, borrowers who default cause the largest amount of loss to the lenders.

Company's objective is to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

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 our case study.

**Business Constraints:**

To consider only those applicants , for analysis, whose application is approved.

To consider only “Chared Off” and “Fully Paid” borrowers in the analysis. ie, We do not consider borrowers with Loan Status as “Current”.

**Goals:**

To understand the Driving Factors behind Loan Default.

To Express the findings of  Univariate and Bi-Variate analysis using Neat Visualisations.

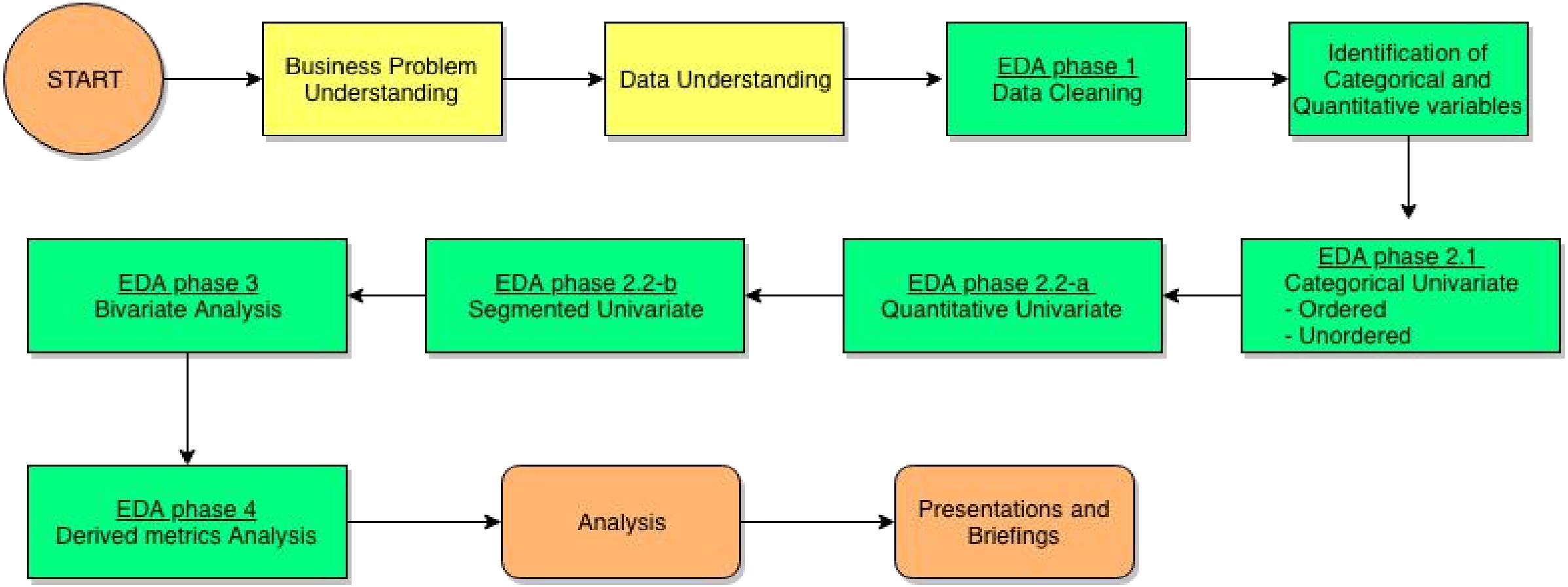
To make some recommendations, based on the findings, to reduce Loan defaults.

**Assumptions:**

Applicants with Loan Status as “Charged Off” are considered to have defaulted on Loan and assumed that they would not repay the Loan amount.

# Problem solving methodology

The various stages of the analysis done is depicted in the flow chart below:



# Data Cleaning and Preparation

We Drop All columns that have more than 80% NA values, since no significant analysis can be made from them.

We Drop All columns (such as paymnt_plan, tax_liens) that have only one unique value.

We also drop some columns such as (Zip Code, Title, Member Id, Id) since they were redundant and not useful for our Analysis.

We Fix certain Data Quality issues such as:

Remove "%" sign and text "months" from appropriate columns.

We process and clean the Employment Length column using Regular Expression, so that the values in the columns become Quantitative in nature, thereby open to analysis.

Regex is used to replace "year/years" with empty, "less than 1" with 0, and "10+" with 10.

The missing values in the Employment Length column is Imputed with the Median Value.

Similarly we impute the values for "pub_rec_bankruptcies" and "revol_util" columns.

We Fix and Change the Date formats from "Jan-15" to "2015/01", so that Analysis can be done on it.

For Outlier Treatment, we use the range of 5th percentile to 95th percentile of the values.

# Exploratory Data Analysis (EDA)

## Methodology:

1. Univariate Analysis on Categorical Variables (Un-Ordered and Ordered) followed by Univariate Analysis on Quantitative Variables.

2. Segmented Univariate Analysis.

3. Bi-Variate Ananlysis.

4. Further Analysis on some Derived Metrics.
   (Type Driven,Business Driven and Data Driven)

# Univariate Analysis - Categorical Variables

## (Unordered Categorical Variables)

**Univariate Analysis was done on Unordered Categorical Variables For the following variables:**

**Verification Status**

**Home Ownership**

**Purpose**

**Region**

**For each mentioned variable, the analysis was done on the Total Count of Defaulters as well as the Percentage of Defaulters. The Key observations are presented in the following slides.**

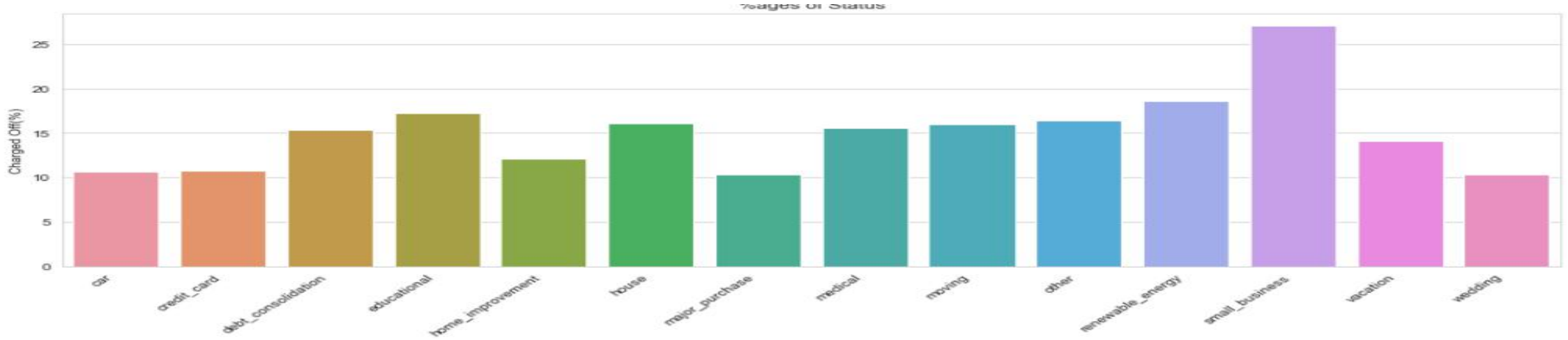
# Univariate Analysis - Categorical Variables
## (Unordered Categorical Variables)

Key Observations:

Loan taken for the Purpose of "Small Businesses" show very high percentage of Defaulters. (475 out of 1754)

ie. 27.08% of loans taken for the purpose of "Small Businesses" results in a default.

(As depicted in the below bar-graph)



# Univariate Analysis - Categorical Variables (Ordered Categorical Variables)

**Univariate Analysis was done on Ordered Categorical Variables For the following variables:**

**Grades and Sub-Grades**

**Employment Length**

**Term of Loan**

**Public Record Bankrupties**

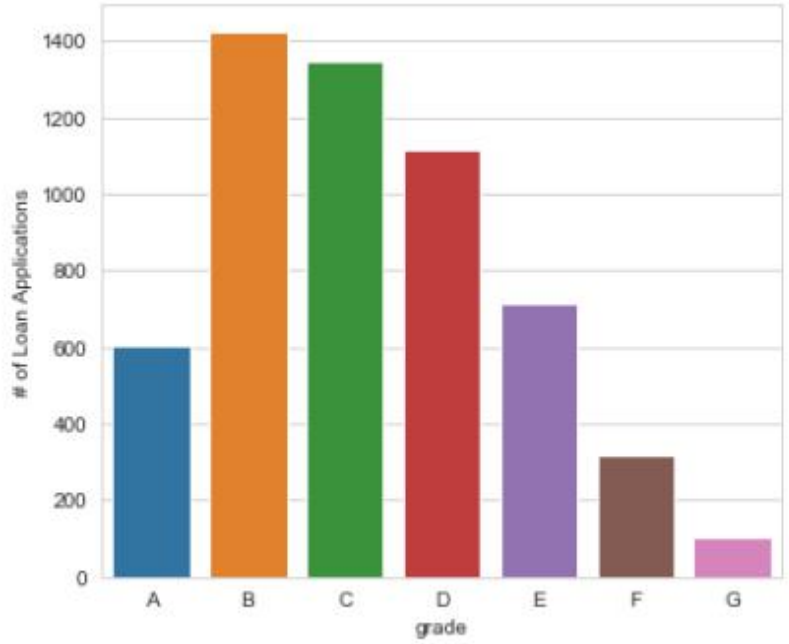
**For each mentioned variable, the analysis was done on the Total Count of Defaulters as well as the Percentage of Defaulters. The Key observations are presented in the following slides.**

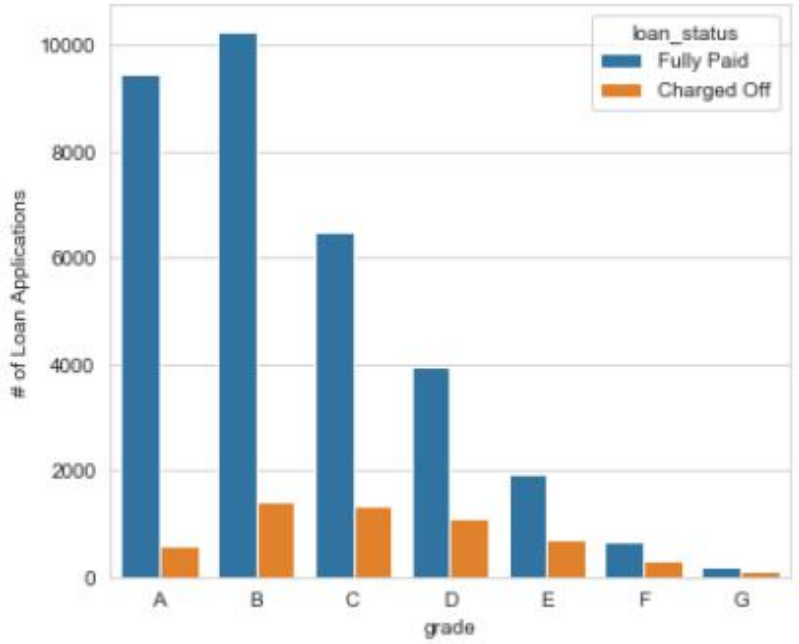
# Univariate Analysis - Categorical Variables (Ordered Categorical Variables)

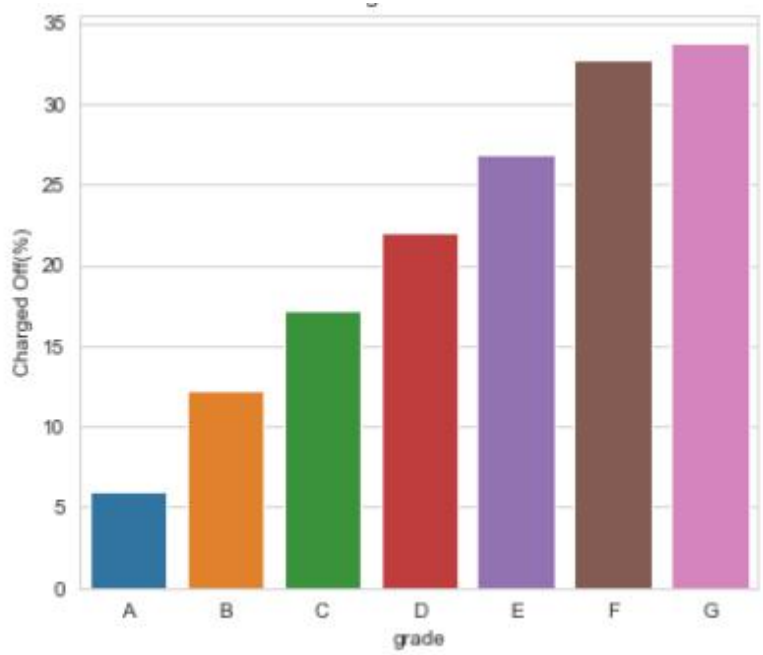
**Key Observations:**

**Grade G and Grade F Loans clearly show very high percentage of defaults.**

**33.78% for Grade G and 32.68% for Grade F result in loan default. (As depicted below)**



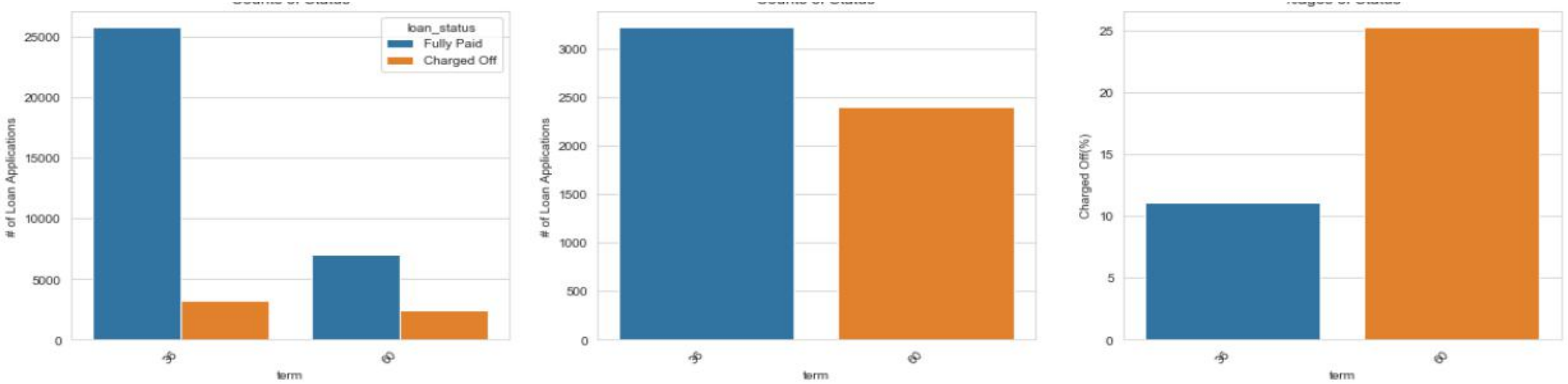



# Univariate Analysis - Categorical Variables (Ordered Categorical Variables)

**Key Observations:**

**The percentage of Loan defaults is Extremely high for a Term of 60 months, as compared to the term of 36 months. (As depicted below)**

**25.31% of borrowers default when given loan for a period of 60 months, as compared to 11.09% defaulters for a period of 36 months.**



# Univariate Analysis - Quantitative Variables

**Univariate Analysis was done on the following Quantitative Variables:**

**Loan Amount, Funded Amount,**
**Investor Funded Amount, Interest Rate,**
**Installment,  Annual Income,**
**Debt to Income Ratio, Revolving Line Utilisation Rate**

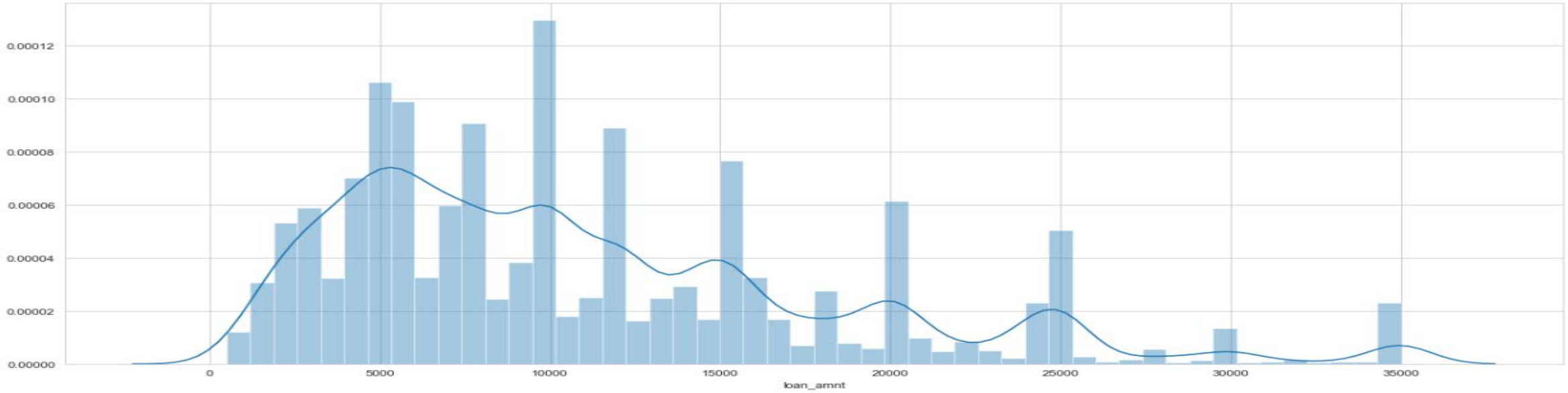
**For each mentioned variable, the analysis was done using a Box-Plot as well as using a ViolinPlot. The key observations are noted in the following slides.**

# Univariate Analysis - Quantitative Variables

**Key Observations:**

**Loans given for a smaller amount (4,000 to 10,000) are very likely to default as compared to Loans given for Higher amounts. (15,000 and more)**

**As clearly observed from the below graph, maximum default of Loans happen for a Loan Amount of Rs. 5,000.**

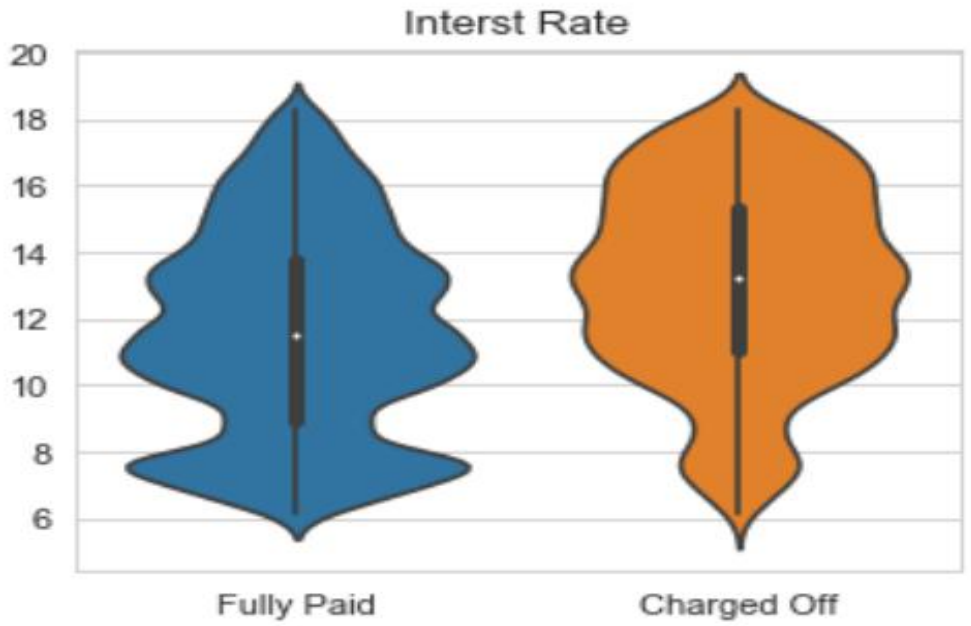


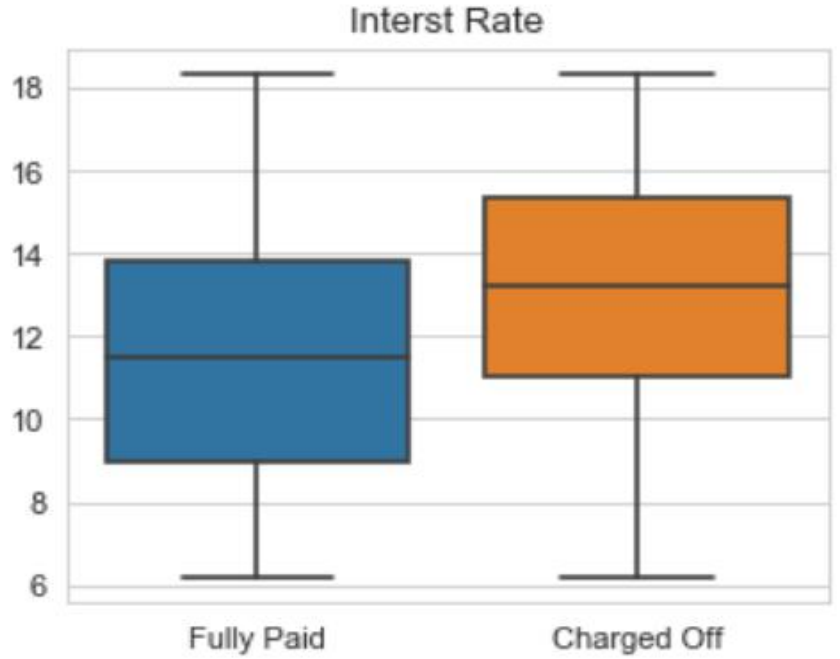
# Univariate Analysis - Quantitative Variables

**Key Observations:**

**Borrowers are highly likely to default on the Loan when the Rate of Interest is High. (12% and above).**

**The Higher the rate gets, more chances of a default. As can be clearly seen from the below Violin Plot and Box Plot.**

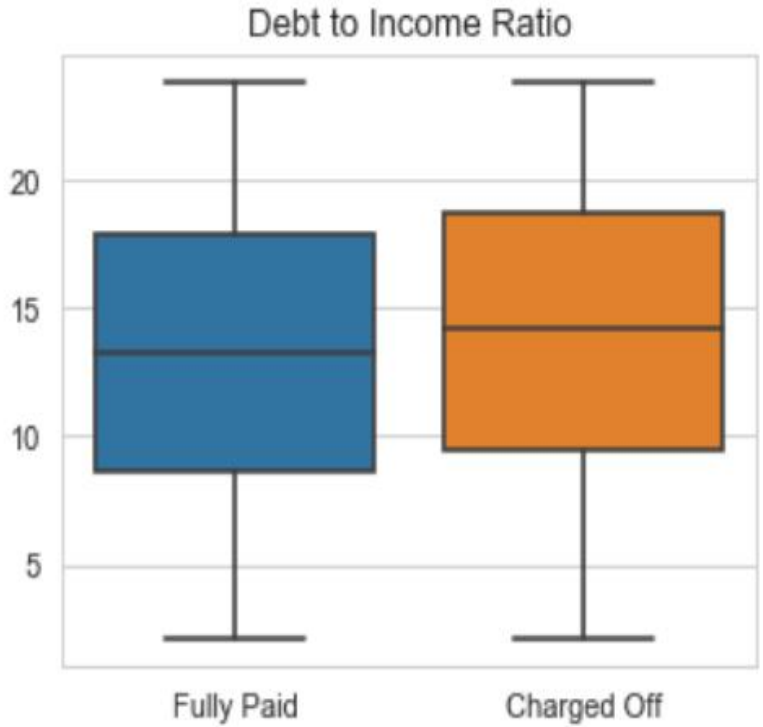


# Univariate Analysis - Quantitative Variables

**Key Observations:**

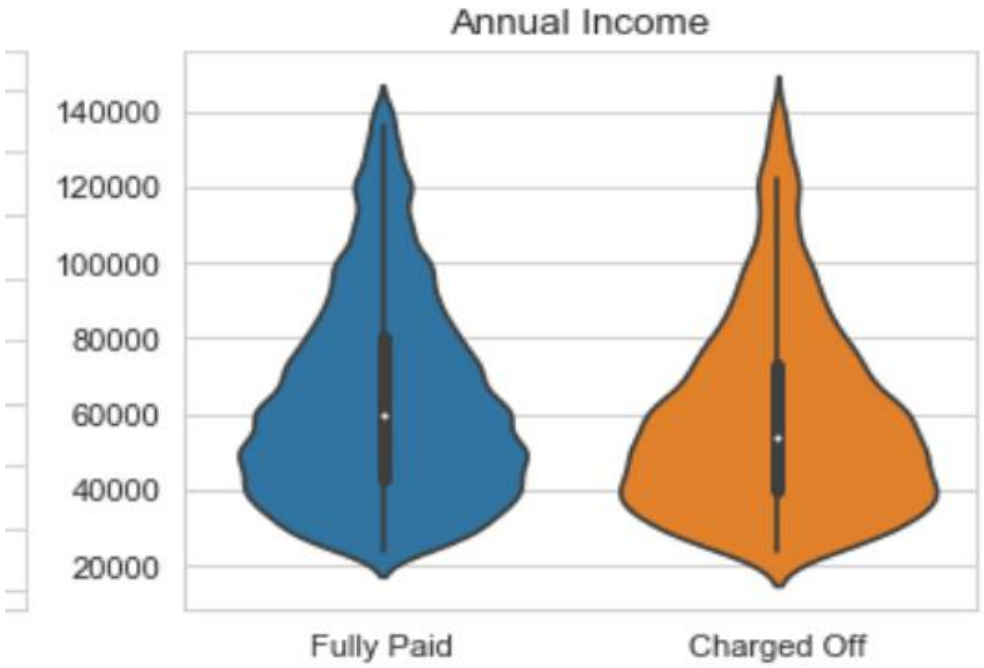
**As can be seen from the below Violin and Box plot, A borrower with High Debt to Income Ratio is more likely to default as compared to a borrower with low Debt to Income Ratio.**

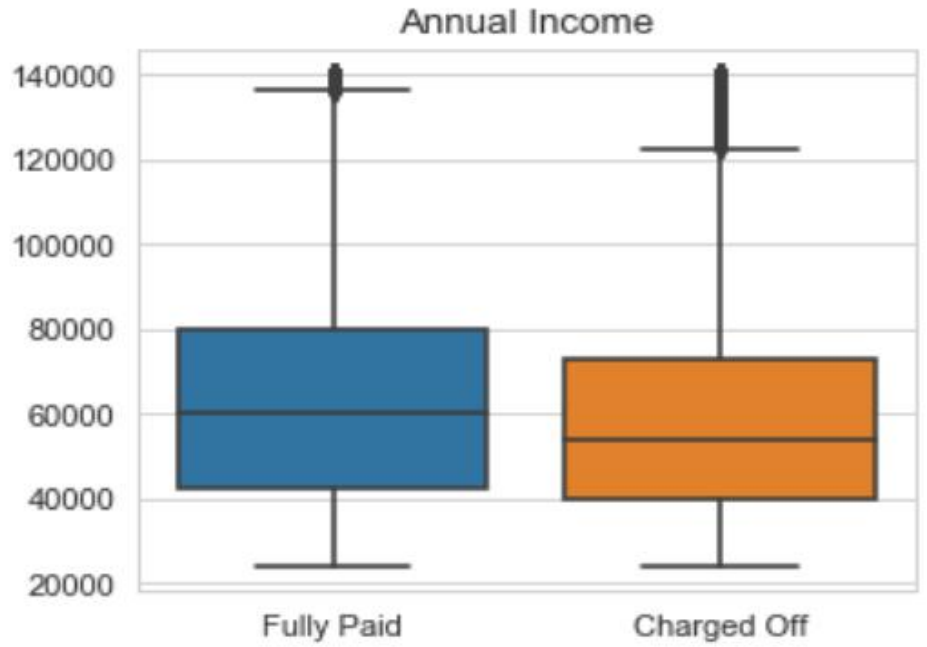



# Univariate Analysis - Quantitative Variables

**Key Observations:**

**As can be clearly seen from the below Violin and Box plot, A borrower with Low Annual Income is more likely to default as compared to a borrower with High Annual Income.**



# Segmented Univariate Analysis

**Segmented Univariate Analysis was done on the following Quantitative Variables:**

**Loan Amount,**

**Interest Rate,**

**Annual Income,**

**Debt to Income Ratio.**

**Each of the above mentioned variable was analysed with the following segments:**

**Purpose of Loan,**

**Home Ownership,**

**Grade of Loan,**

**Employment Length.**

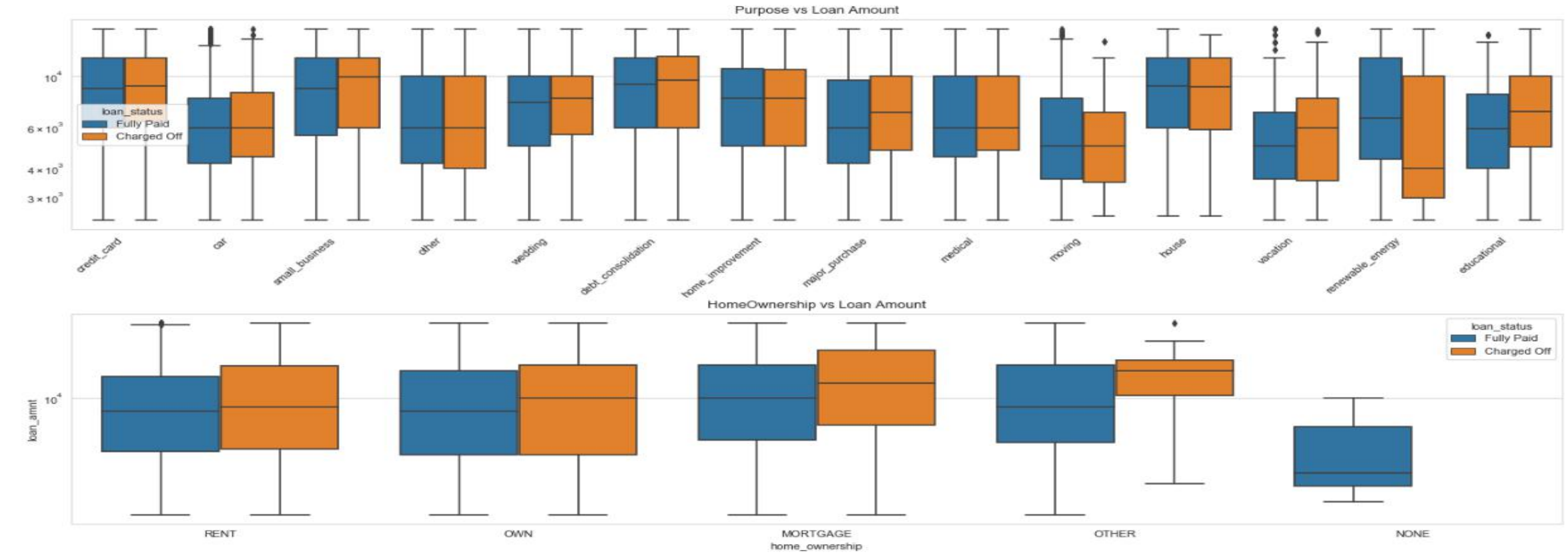
**The key obervations are mentioned in the following slides.**

# Segmented Univariate Analysis

## Key Observations (on Loan Amount):

**We observe that Small Businesses followed by Debt Consolidation as a purpose for Loan are the ones that are most likely to default. While Loans taken for the purpose of Renewable Energy are the Least Likely to default.**

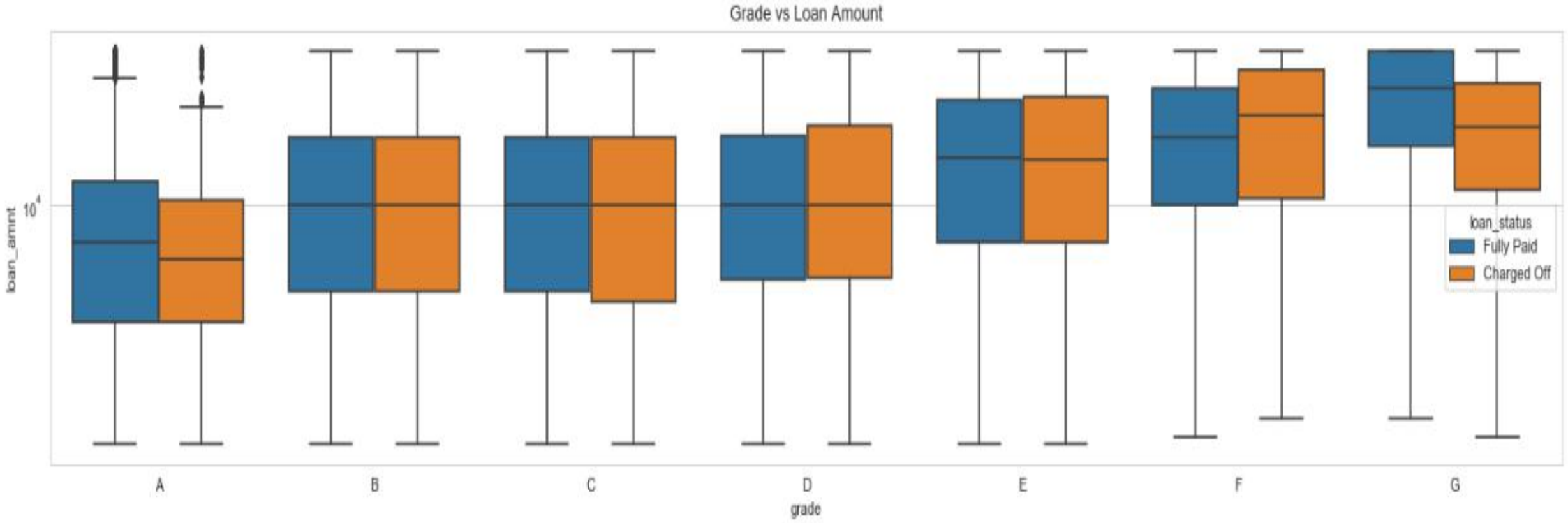
**We also observe that, people taking High Loan amount and having "Mortgage" as Home Ownership are very Likely to default.**



# Segmented Univariate Analysis

## Key Observations (on Loan Amount):

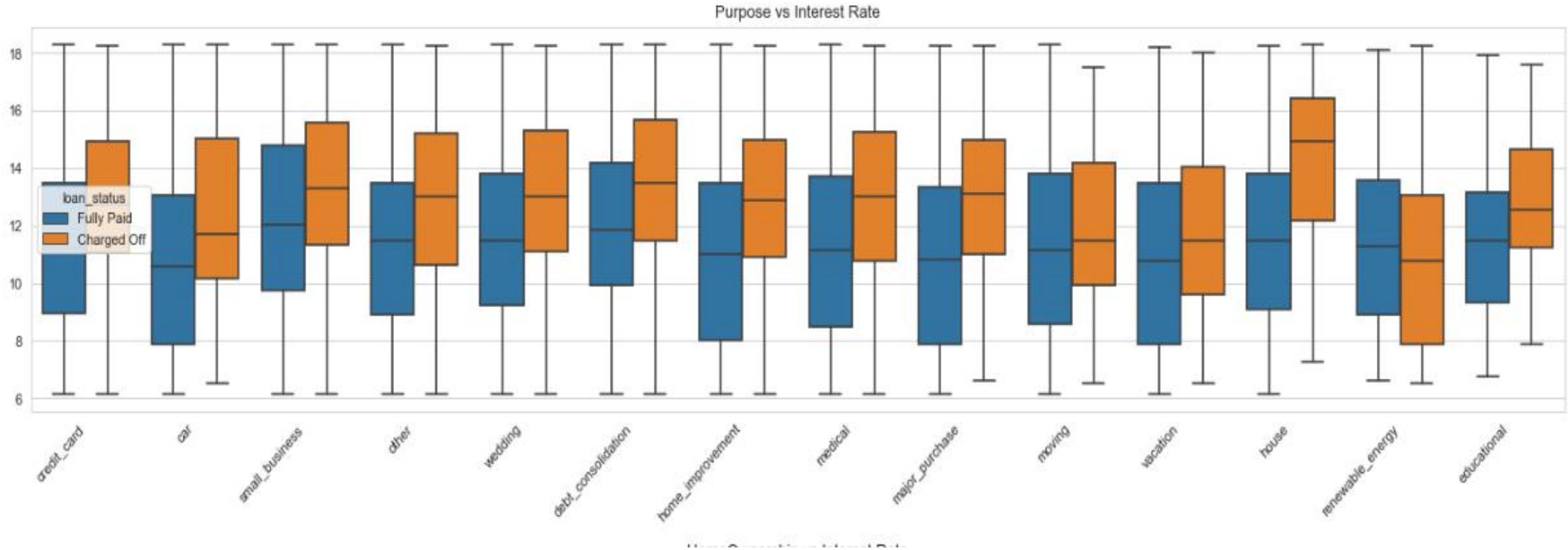
**We observe that Grade F and Grade G loans for relatively high amounts are very Highly likely to default.**



# Segmented Univariate Analysis

## Key Observations (on Interest Rates):

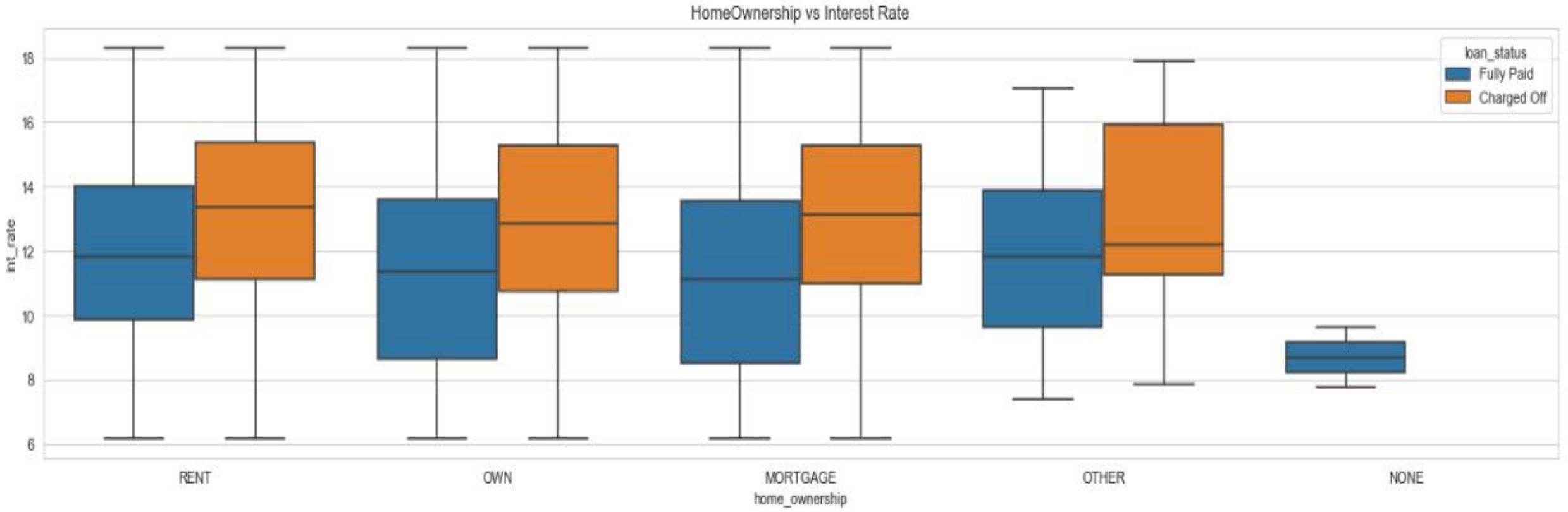
**We observe that Loan taken at a High Interest Rate for the Purpose of "House" is very Likely to Default.**



# Segmented Univariate Analysis

## Key Observations (on Interest Rates):

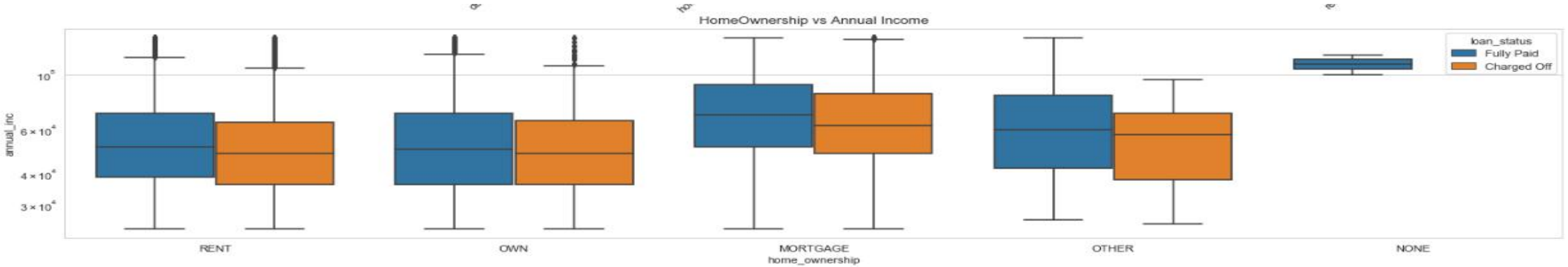
**We also observe that Loan taken at a High Interest Rate by People whose Home Ownership is "Others" are highly probable to Default.**

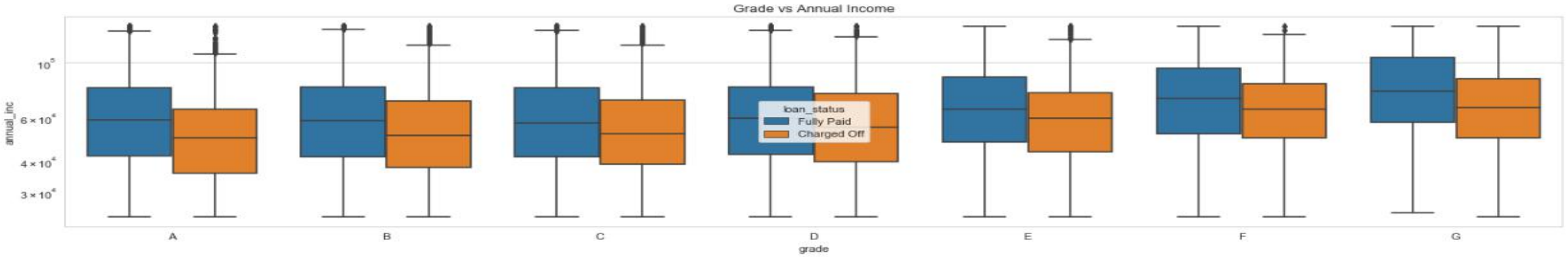


# Segmented Univariate Analysis

## Key Observations (on Annual Income):

**Loan given to High Annual income people who have "Mortgage" as home ownership or Grade G category loan are most likely to default.**

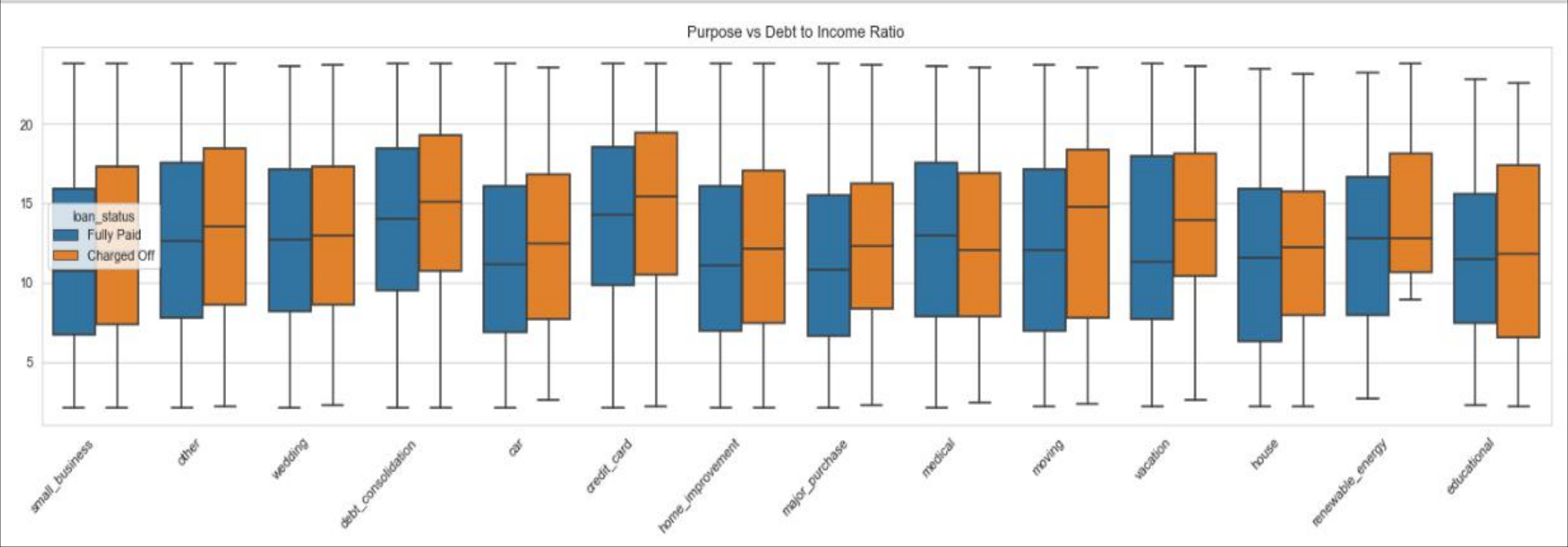


# Segmented Univariate Analysis

## Key Observations (on Debt to Income Ratio):

**We can observe that Loan given for the purpose of CrediCard and Debt Consolidation to people having High Debt to Income Ratio are very Likely to Default.**



# Univariate Analysis - Summary

## Driver Variables:

**Based on the Uni-Variate Analysis, the following Variables can be concluded as the Important Driver Variables (i.e. variables which are strong indicators of default):**

1. **Purpose of Loan (**27.08% of loans taken for the purpose of "Small Businesses" results in a default**)**

2. **Grade of Loan** (33.78% for Grade G and 32.68% for Grade F result in loan default)

3. **Term of Loan** (25.31% of borrowers default when given loan for a period of 60 months)

4. **Loan Amount** (Maximum default of Loans happen for a Loan Amount of Rs. 5,000.)

5. **Rate of Interest** (Borrowers are highly likely to default when the Rate of Interest is High. (12% up).

6. **Debt to Income Ratio** (A borrower with High Debt to Income Ratio is Highly likely to Default)

7. **Annual Income** (A borrower with Low Annual Income is Highly likely to Default)

# Univariate Analysis - Summary

## Recommendations:

**Based on the Uni-Variate Analysis, the following recommendations can be made:**

**Thorough scrutiny should be done on applicants citing "Small Business" as a purpose for the Loan.**
**Furthermore, High Loan amount should not be sanctioned for people giving "Small Business" or "Debt Consolidation" as a purpose for the Loan. Either a thorough background check on the applicant should be done or interest charged should be high enough to discourage such Loans.**

**Applicants with High Debt to Income ratio (DTI) should be treated with caution. Thorough scrutiny and background check must be done. Furthermore, Applicants with High DTI ratio citing "Credit Card" or "Debt Consolidation" as a purpose must be strictly dealt with. Security Mortgage must be taken, if the need be. Since they are Highly likely to Default. (as seen in slide 23)**

**Interest Rates should be reduced for people citing "House" as a purpose for Loan. Since High interest rate here implies high default rate. (as observed in Slide 20)**

**Caution should taken when Applicants having Home Ownership as "Others" and are willing to take a Loan at a High Interest Rate. Either a Security Deposit should be taken from such applicant or a thorough background check must be done. Since they are very likely to default. (as seen in slide 21)**

**Caution should be maintained when giving loans to applicants with High Annual Income BUT having "Mortgage" as Home Ownership. Since they are very Likely to default. (as seen in slide 22)**

**Verification checks should be made more stringent, since there seems some loopholes. As we observed verified people have high percentage of Loan Defaults.**

# Bi-Variate Analysis

**For Bi-Variate Analysis we created the following two derived columns:**

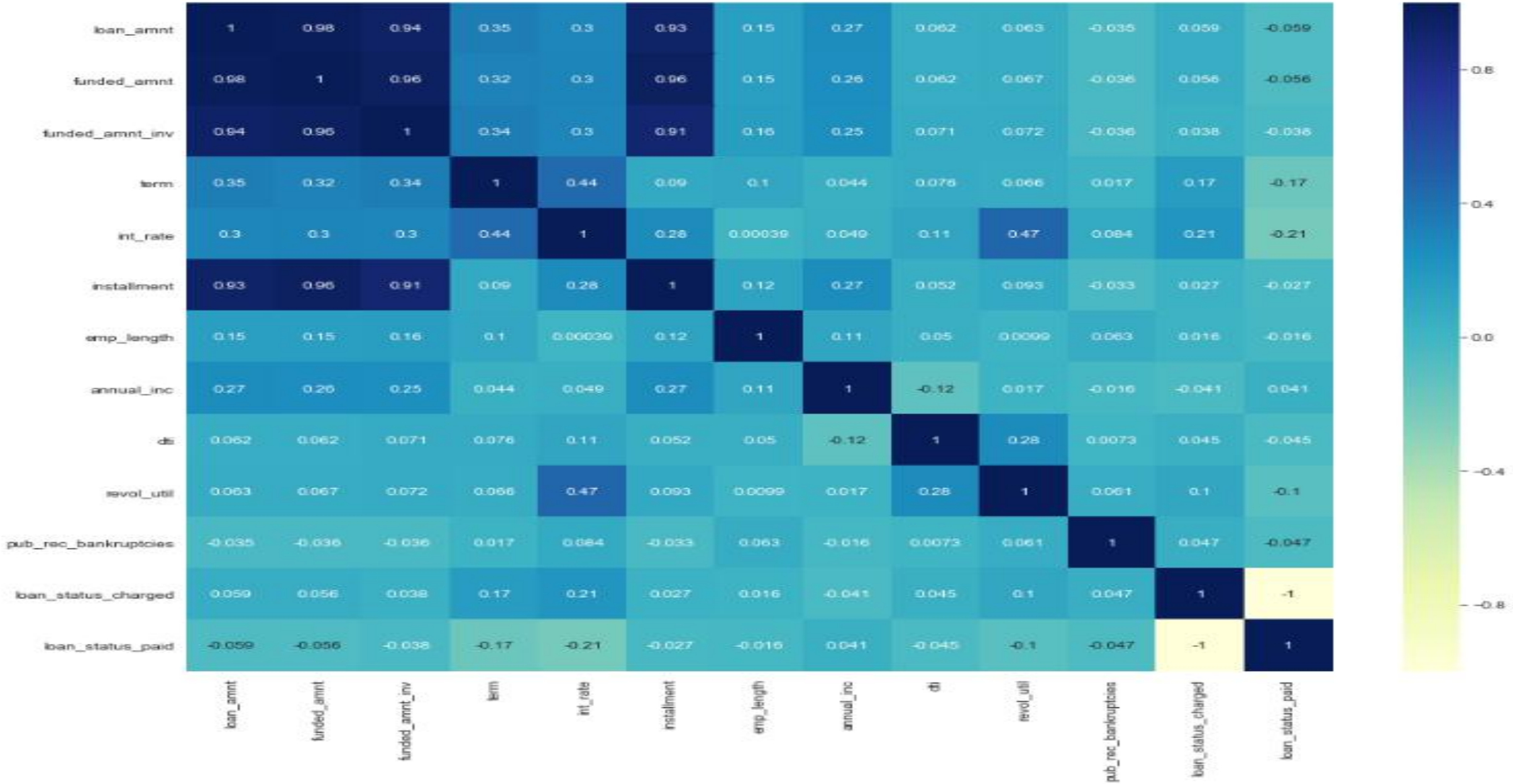
**loan_status_charged (1 if Charged Off)**
**loan_status_paid (1 if Fully Paid)**

**We Plot Heat Maps (shown in next slide) using the Pearson Correlation Method and observe the following:**

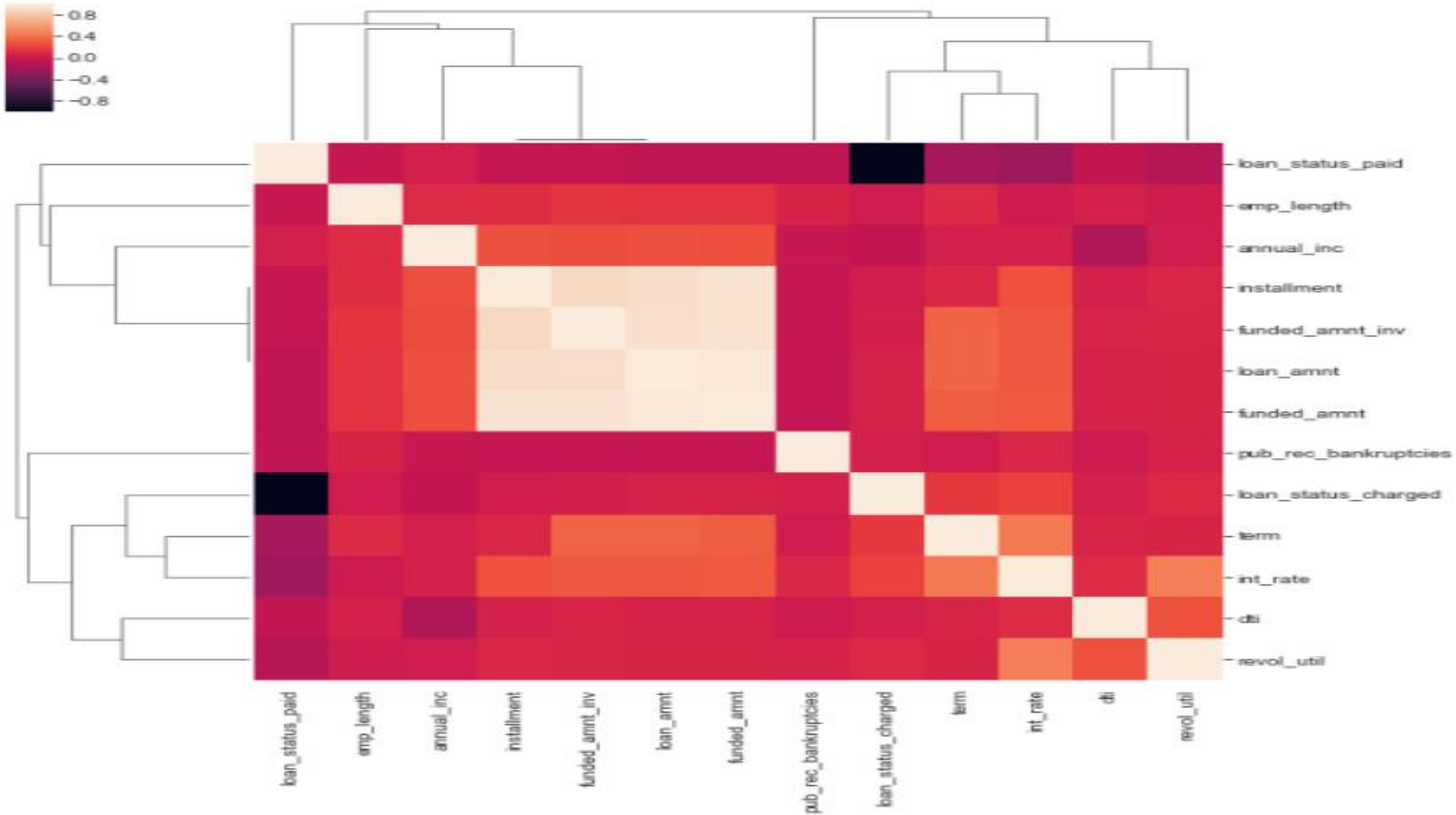
**1. charged off applications have close relation with term and interest rates**

**2. charged off applications have close relation with dti and revolving_util**

# Bi-Variate Analysis



# Bi-Variate Analysis



# Further Analysis - Type Driven Metrics

We do some further Analysis on Type Driven Metrics and observe the following: (Graph on Next Slide)

There is an Increase in the loan amount year on year.

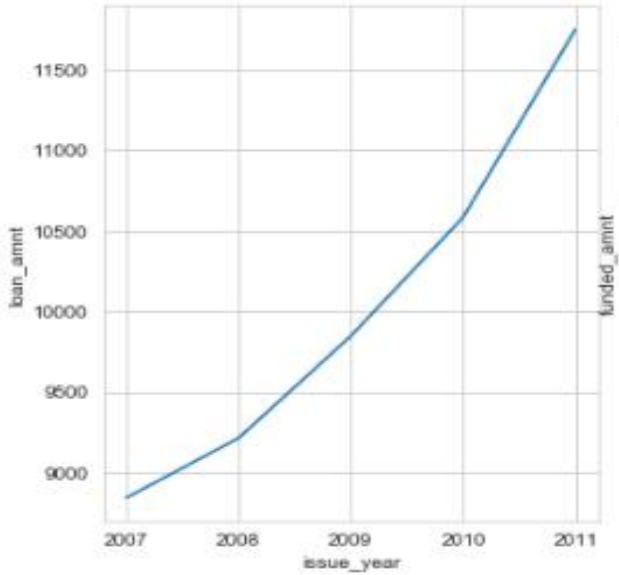
There is an Increase in the funded amount year on year.

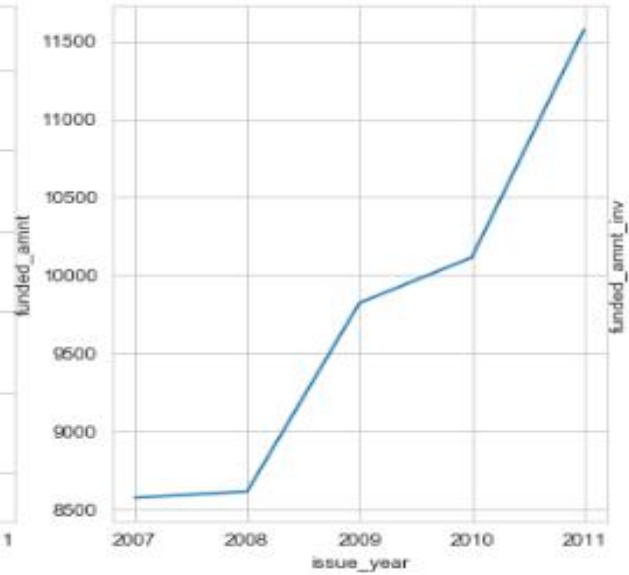
There is an Increase in the fundede amount investor year on year.

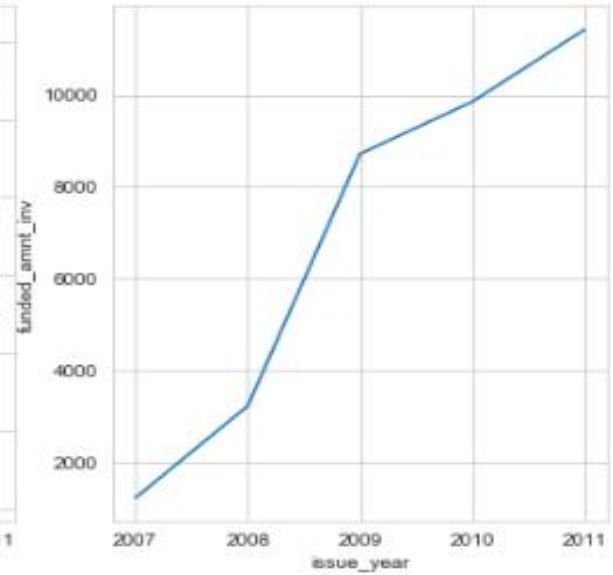
There is an Increase in revolving util year on year.

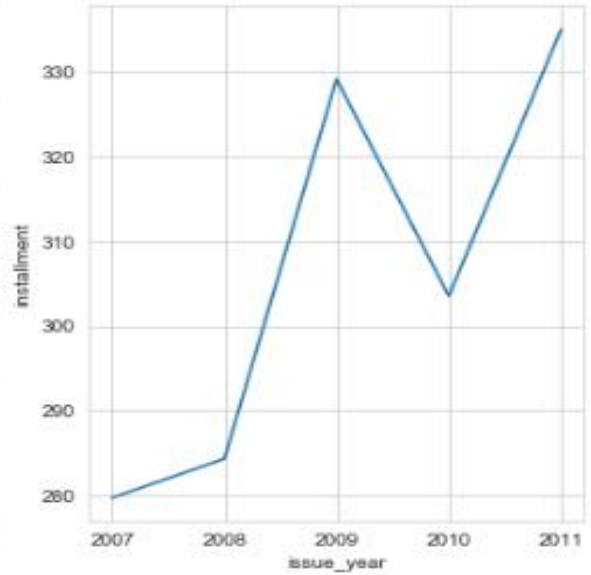
There is an Increase in debt to income ration year on year.

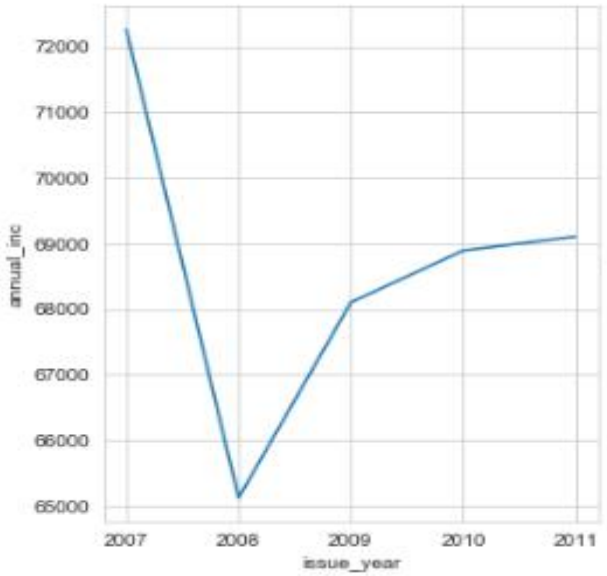
# Further Analysis - Type Driven Metrics

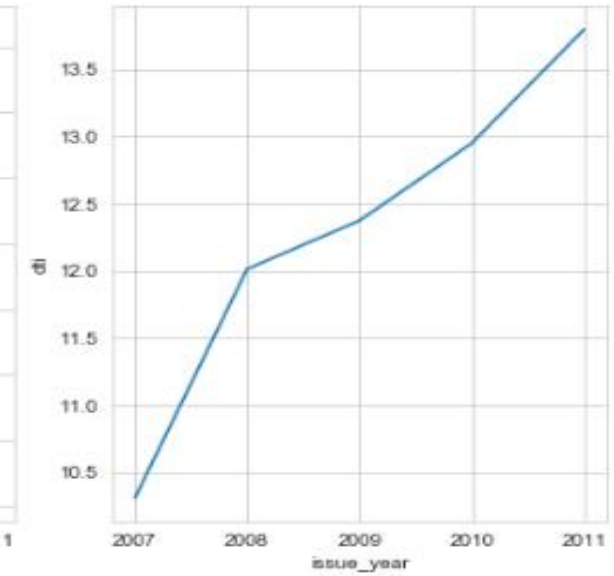


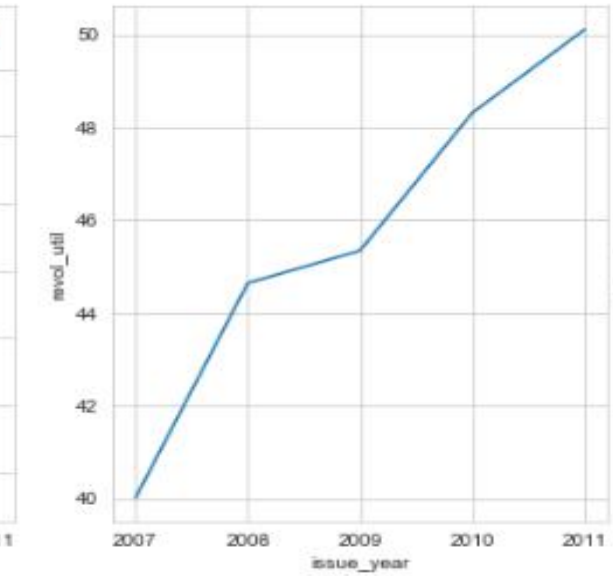



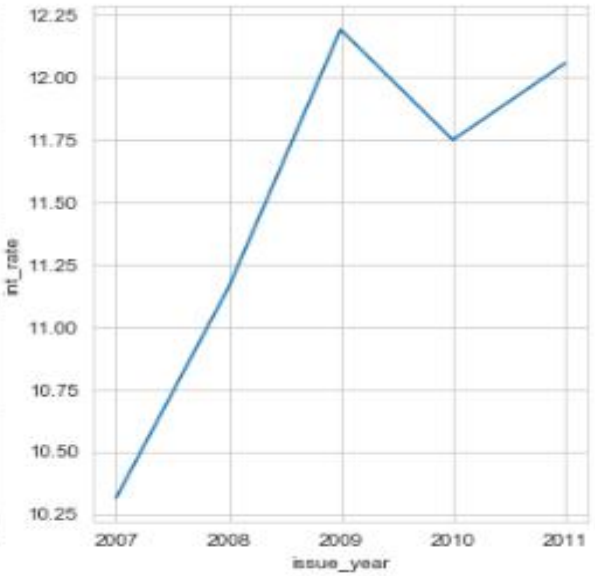



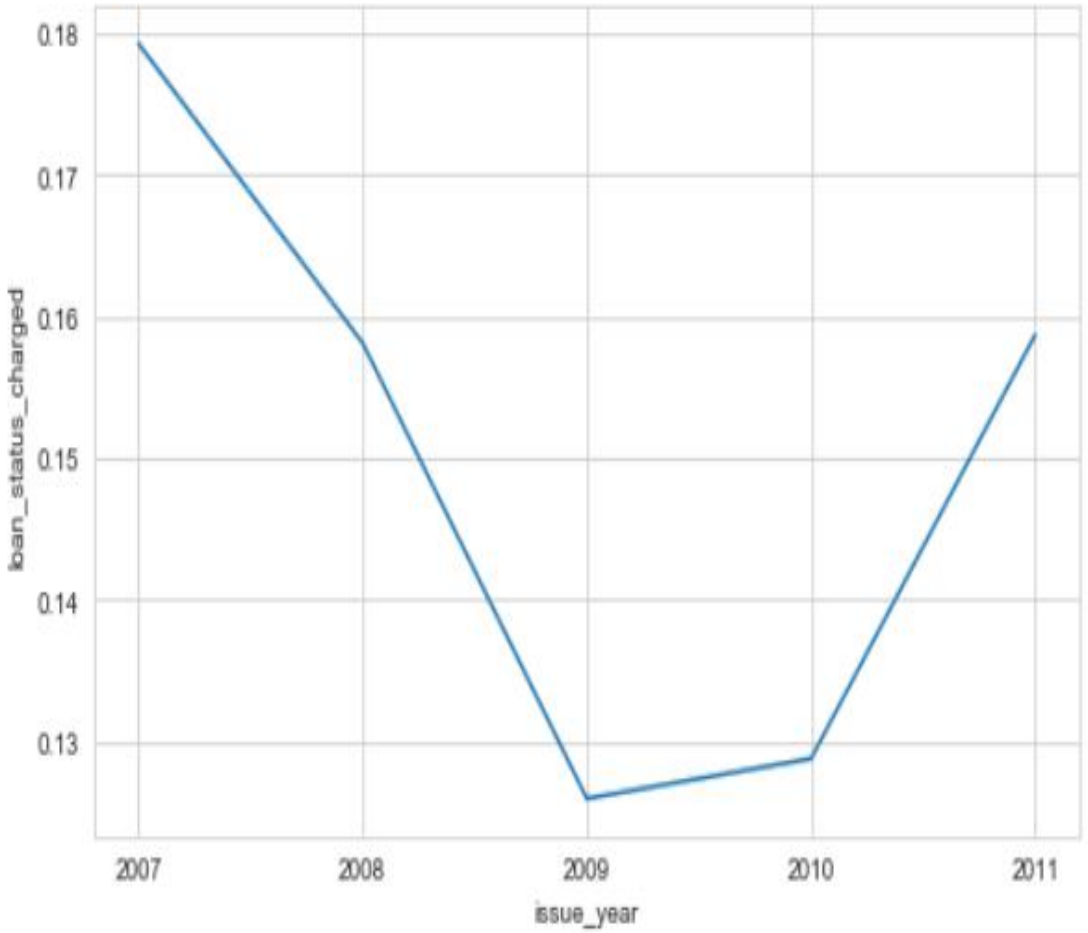



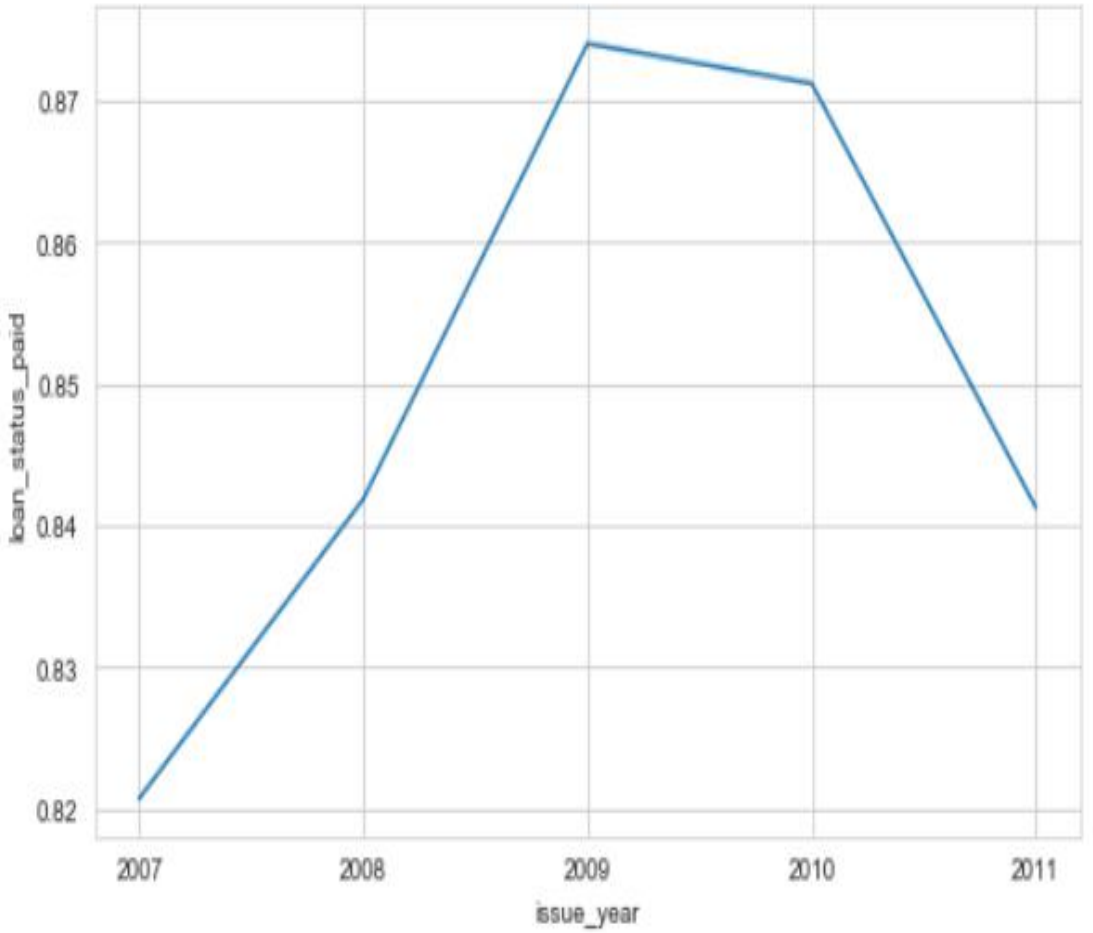



# Further Analysis - Type Driven Metrics

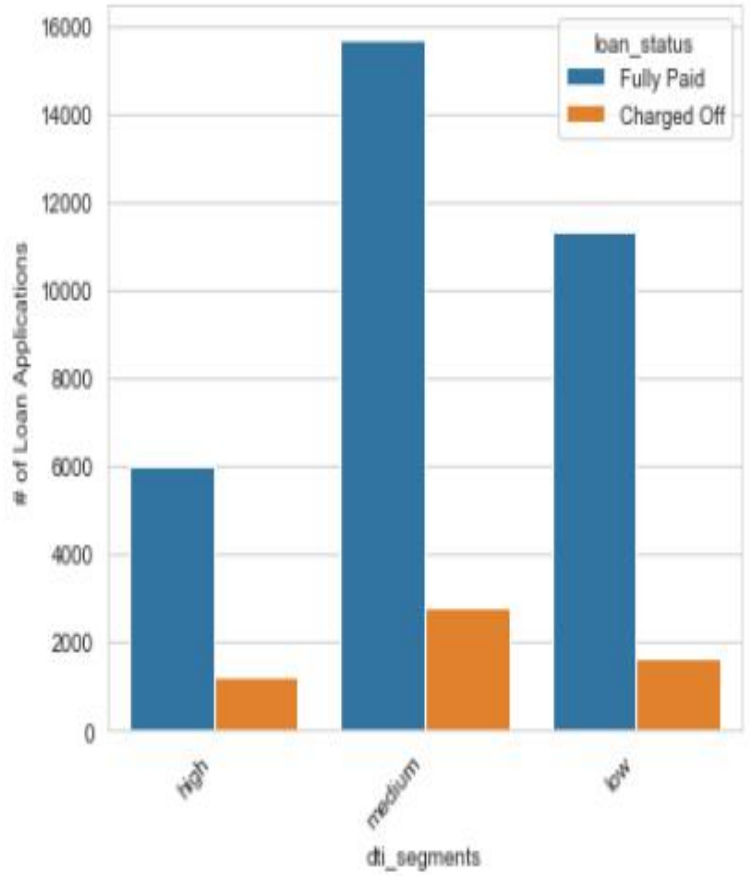
We can Observe a sudden increase in Loan Defaults from 2010 to 2011, as depicted in the below graphs

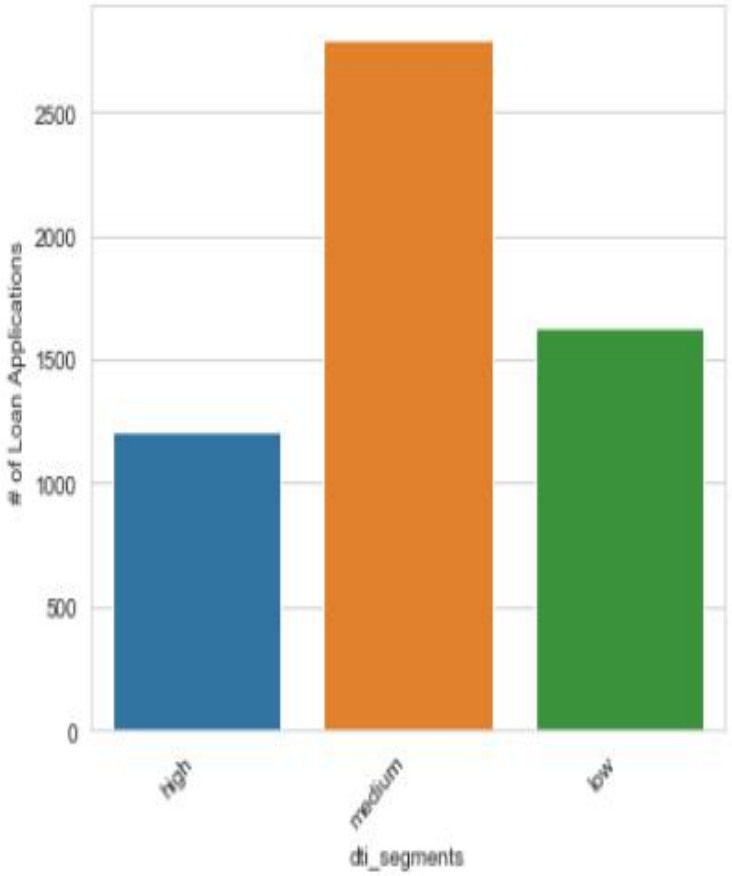


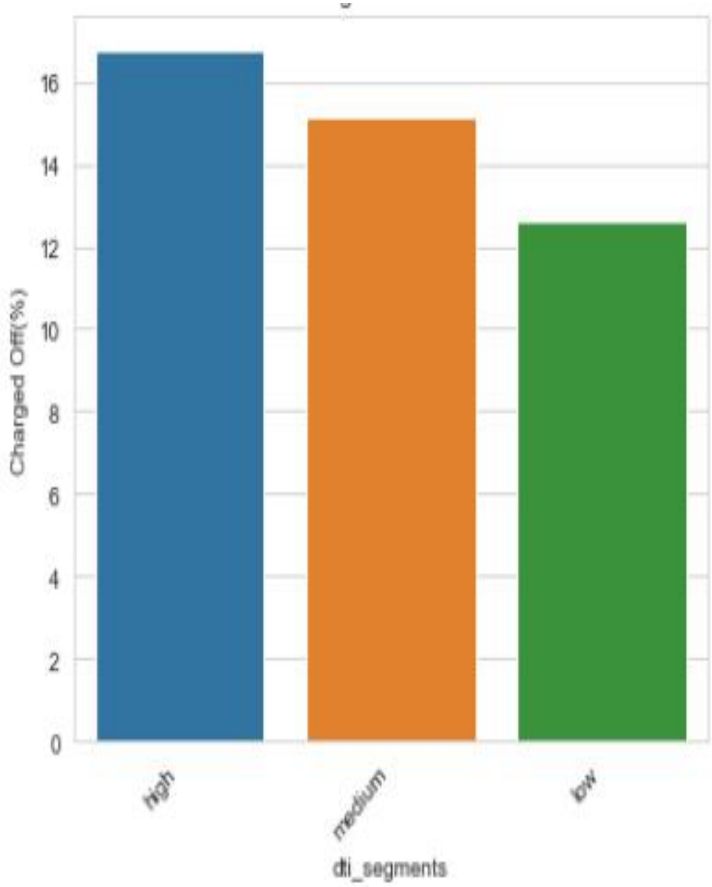



# Further Analysis - Business Driven Metrics

We can observe from the following graphs that a High Debt to Income Ratio results in more number of Loan Defaults.



# Further Analysis - Business Driven Metrics

We can observe from the following graphs that Both the ratios Income:Loan_Amount & Income:Installment show that lower income has high chances of defaults.