Gramener Case Study

(Loans Dataset)

Problem statement

This case study will give an idea of how real-life business problems like how consumer attributes and loan attributes influence the tendency of defaulting are solved using EDA. In this case study, we will apply the techniques of EDA. Also will develop a basic understanding of risk analytics in banking and financial services and understand how data minimises the risk of losing money while lending it to customers.

Consumer finance company that specialises in providing various types of loans to urban customers. When the company receives a loan application, it has to decide whether to approve or reject it based on the applicant's profile. Two types of risks are associated with the bank's decision -

1) If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

2) If the applicant is not likely to repay the loan, i.e., they are likely to default, then approving the loan may lead to financial loss for the company.

Problem Summary

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 three possible scenarios, as described below:

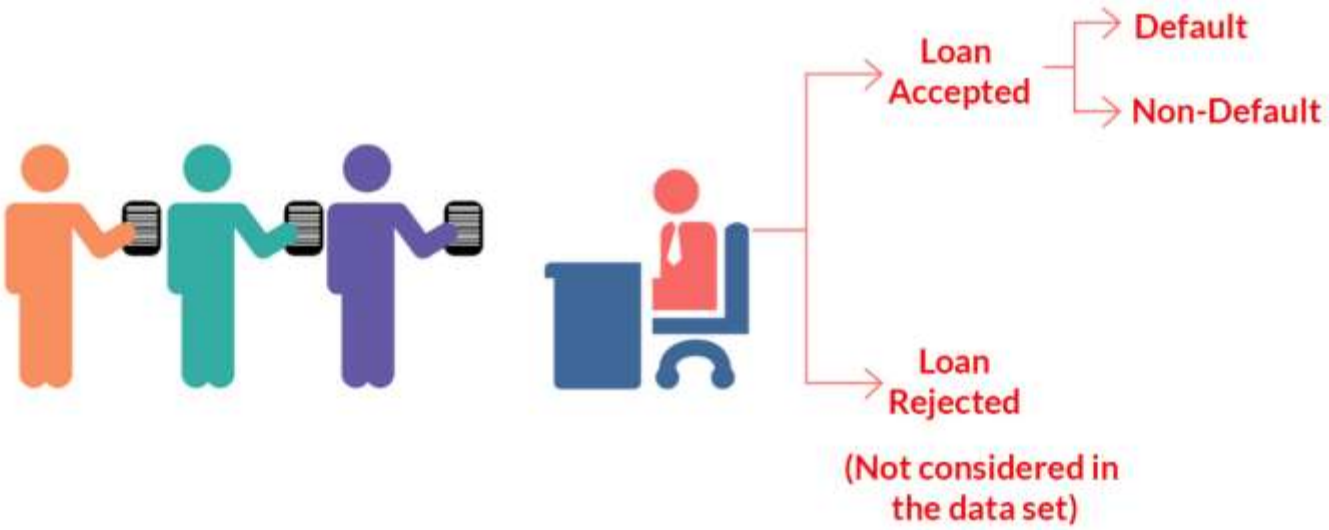
- **Fully paid**: The applicant has fully paid the loan (the principal and the interest amount).
- **Current**: The applicant is in the process of paying the instalments, i.e., the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- **Charged-off**: The applicant has not paid the instalments in due time for a long period, i.e., they have **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; so, this data is not available with the company (and thus, in this data set).

Process Flowchart

We will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of defaulting.

The following image depicts the decisions that could be undertaken by the firm.



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 utilise this knowledge for its **portfolio and risk assessment.**

Data Understanding

- **Importing packages**

```
[1]: import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt

     import seaborn as sns
```

- **Loading CSV file and checking few rows of data.**

```
[2]: #Loading csv file
     data = pd.read_csv(r"C:\Users\HP\Downloads\loan.csv" , low_memory=False)
```

```
[3]: data.head()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_past_12m	pc
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	NaN	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	NaN	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	NaN	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	NaN	
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	NaN	

5 rows × 111 columns

- **Print the structure of the data, There are 39717 observations and 111** characteristics .

```
[4]: print (data.info() , "\n")
     print(data.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Columns: 111 entries, id to total_il_high_credit_limit
dtypes: float64(74), int64(13), object(24)
memory usage: 33.6+ MB
None

(39717, 111)
```

There are 39717 observations and 111 characteristics .

Data Understanding

❖ Checking Datatypes

```
[6]: # check data tyes
     data.dtypes
```

```
[6]: id                                int64
     member_id                         int64
     loan_amnt                         int64
     funded_amnt                       int64
     funded_amnt_inv                 float64
                                      ...
     tax_liens                       float64
     tot_hi_cred_lim                 float64
     total_bal_ex_mort               float64
     total_bc_limit                  float64
     total_il_high_credit_limit      float64
     Length: 111, dtype: object
```

Overall there are int, float ,object datatypes

❖ Checking null values in percentage

```
#Checking null values in percentage
round (100*( data.isnull().sum() / len(data.index) ) ,2 )
```

```
id                              0.0
member_id                       0.0
loan_amnt                       0.0
funded_amnt                     0.0
funded_amnt_inv                 0.0
                               ...
tax_liens                       0.1
tot_hi_cred_lim               100.0
total_bal_ex_mort             100.0
total_bc_limit                100.0
total_il_high_credit_limit    100.0
Length: 111, dtype: float64
```

There are many columns having null values , we will fill them ahead

Also there are no duplicates value in columns

Data Cleaning

In this part , we had change datatype and fixed some values of columns .Also dropped unnecesssary columns and filled missing values .We had some derived columns as well

We had **change format of columns** - issue_d , last_pymnt_d , last_credit_pull_d to '%b-%y'
Also , We had remove '%' sign from int_rate,revol_util column by using rstrip().

Dropped unnecessary columns like – 'id , 'member_id','pymnt_plan','emp_title','title','url' ,'zip_code','last_credit_pull_d','out_prncp','out_prncp_inv','collections_12_mths_ex_med' ,'earliest_cr_line','policy_code','initial_list_status','acc_now_delinq','chargeoff_within_12_ mths','delinq_amnt','tax_liens'

To fill missing values we have use mean for numerical columns, and mode for categorical columns

Derived columns by doing operation on existing columns - num_of_mnts_paid (total_pymnt/installment), issued_month and issued_year from issue_d column, return of Investment(ROI) column by (total_pymnt-collection_recovery_fee/funded_amnt)-1

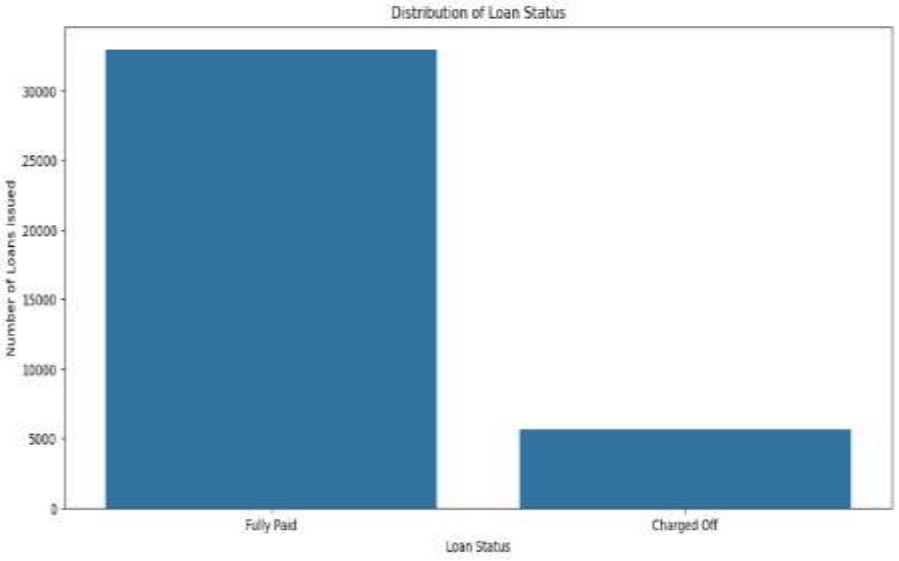
Univariate Analysis

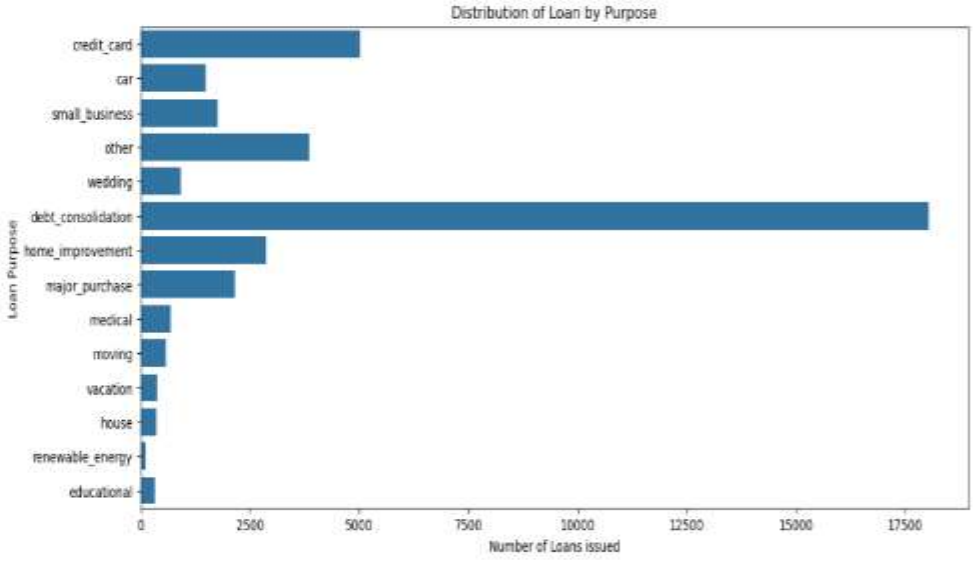
- Now we will see univariate analysis of variables with distribution of loans

There are 2 types -

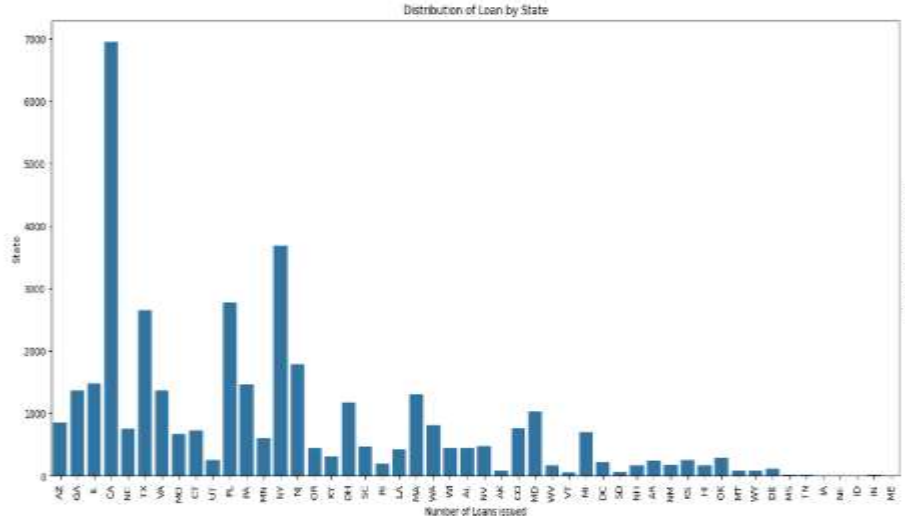
1) Unordered categorical variables
2) Ordered categorical variables

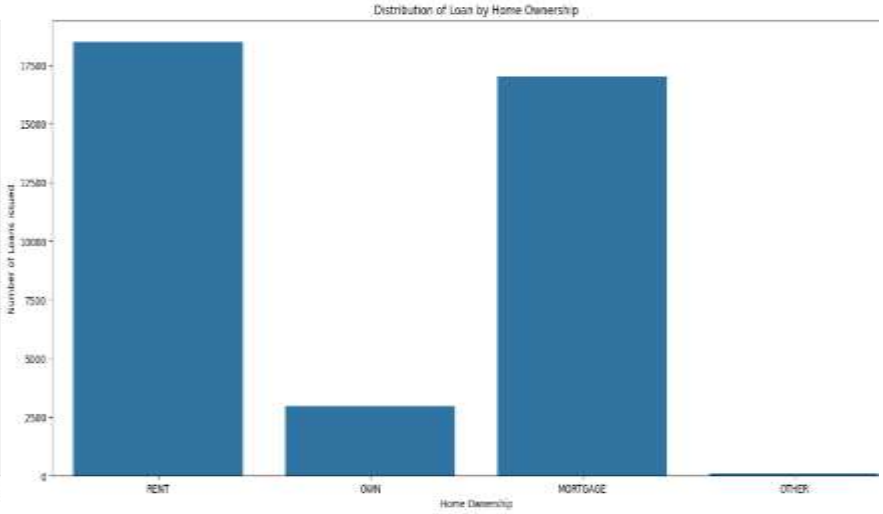
Univariate Analysis of unordered categorical variables

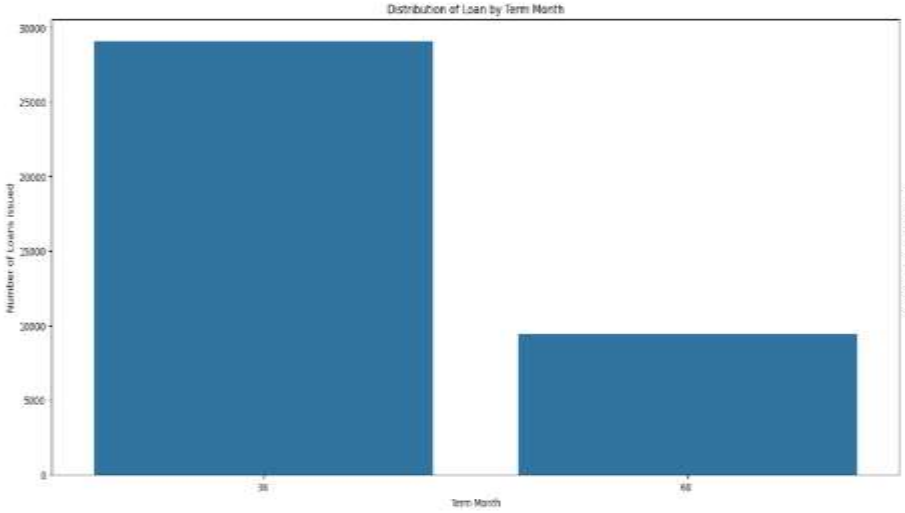


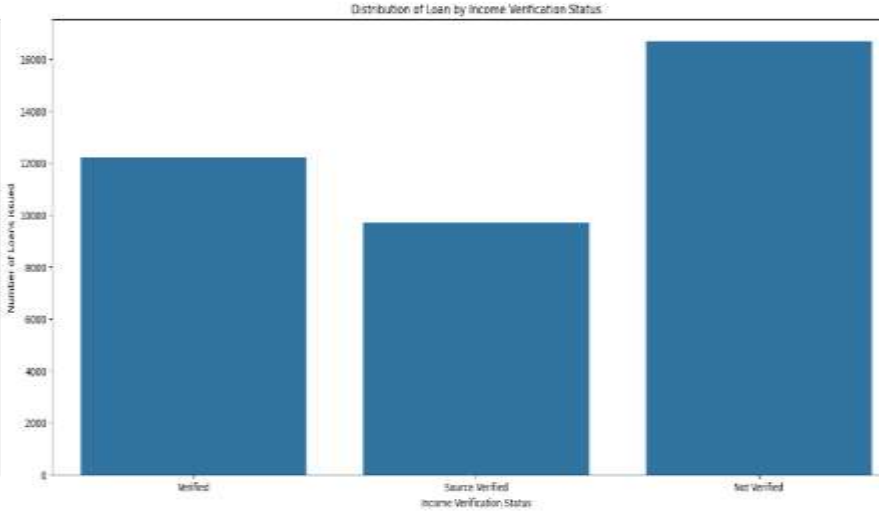


Univariate Analysis of unordered categorical variables









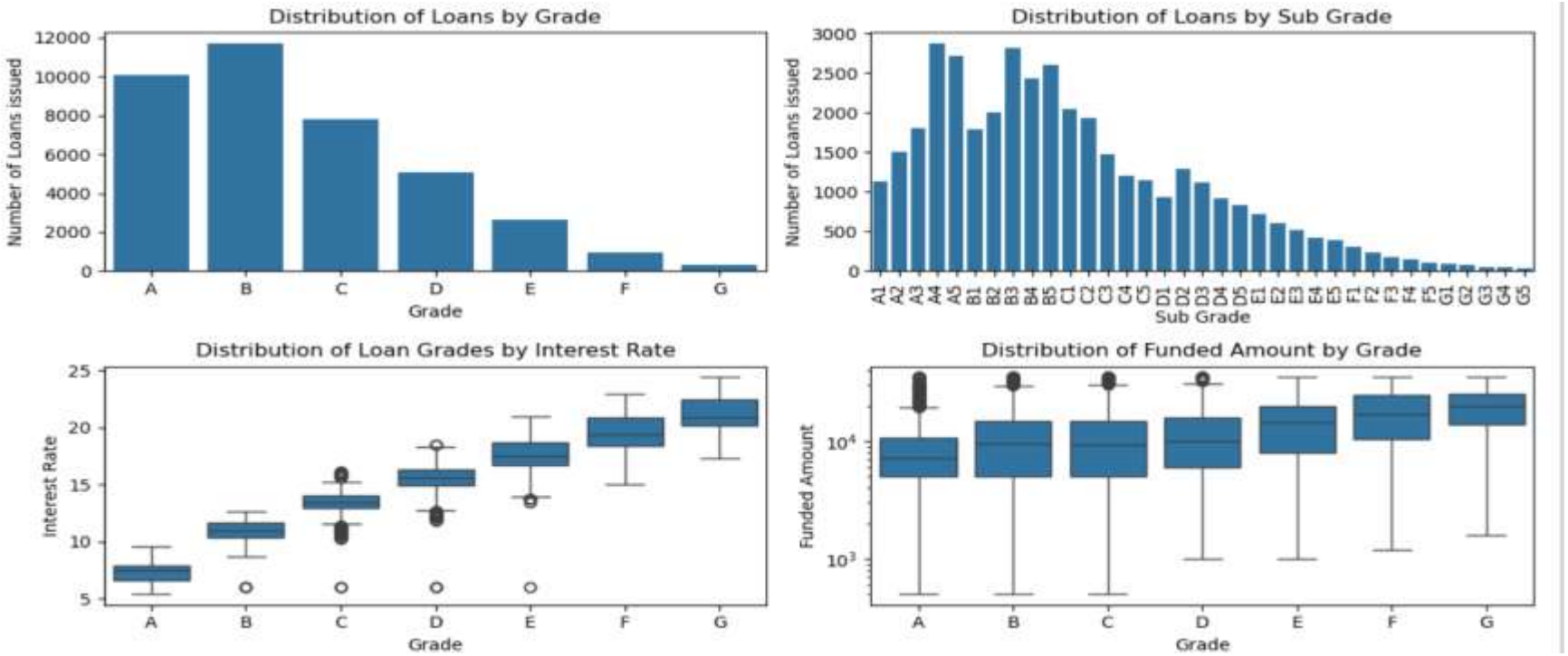
Univariate Analysis of unordered categorical variables

Observations are as below –

1) From this dataset, we have more observations(85%) from "Fully Paid" status.
2) There are more loan applicant's with purpose of debt consolidation.There are more loan applicant's from California state.
3) Most of the loan applicant's are rented and mortgage.
4) There the more number of loan with 36 month term.
5) There the more number of loan which were not verified the annual income.
6) Borrowers with own house and want to consolidate debt are not at much risk, but borrower with rent,mortgage and want to consolidate debt are at high risk applicants.

Univariate Analysis on Ordered Categorical Variables

- Now we will see Distribution of loans by ordered categorical variable



Univariate Analysis on Ordered Categorical Variables

Observations of Distribution of loans by order categorical varible :

1) From above plots, it shows that more number of loans were from B,A and C grade's and least from G grade.

2) From Sub grades A4, B3 have more number of loans.

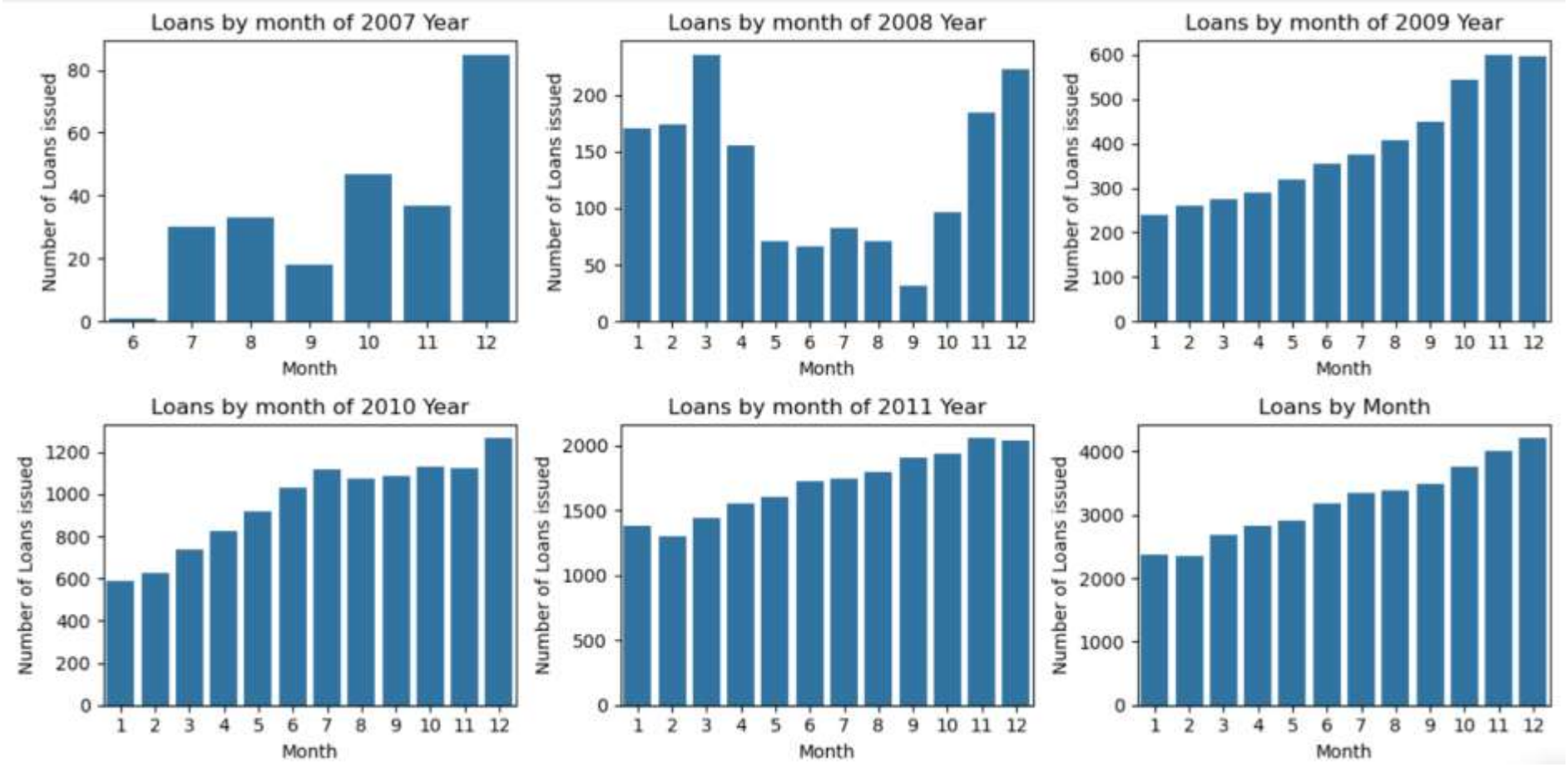
3) From 3rd plot, it shows that A,B,C grade loans have less interest rate and E,F,G have high interest rate. From 1st, 2nd plots there are more number of loans from A,B,C grade(granularity check from sub-grades). It might be the reason that the loan applicant's from A,B,C grades have better credit score and lower risk.

4) From 4th plot, it shows that there are high funded amount in A,B,C and D grades as the applicant's from these grades have better credit score and lower risk.

5) From 5^{th} plot , it shows that the majority of borrowers have been employed for at least 10 years.

Univariate Analysis on Ordered Categorical Variables

- Now we will see Distribution of loans by ordered categorical variable



Univariate Analysis on Ordered Categorical Variables

Observations of Loans Vs month of years relationship :

1) From above plots, it shows that more number of loans were issued in 11,12(November, December) months, the reasons could be Festivals(Thanks giving Day, Christmas and New Year).

2) In 2008 Year, there was huge spike in 3rd Month.

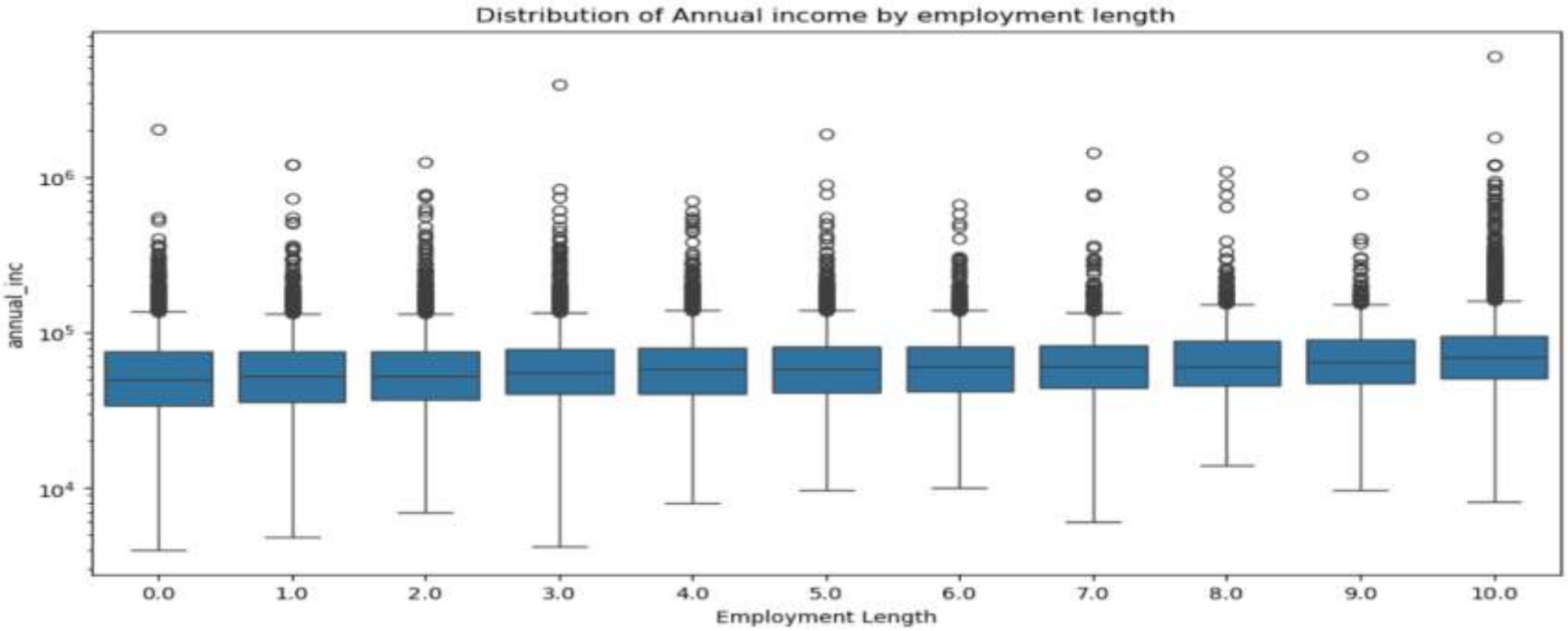
3) Number of loans issued increased steadily with slight descrease in 2008 year.

Univariate Analysis on Quantitative Variables

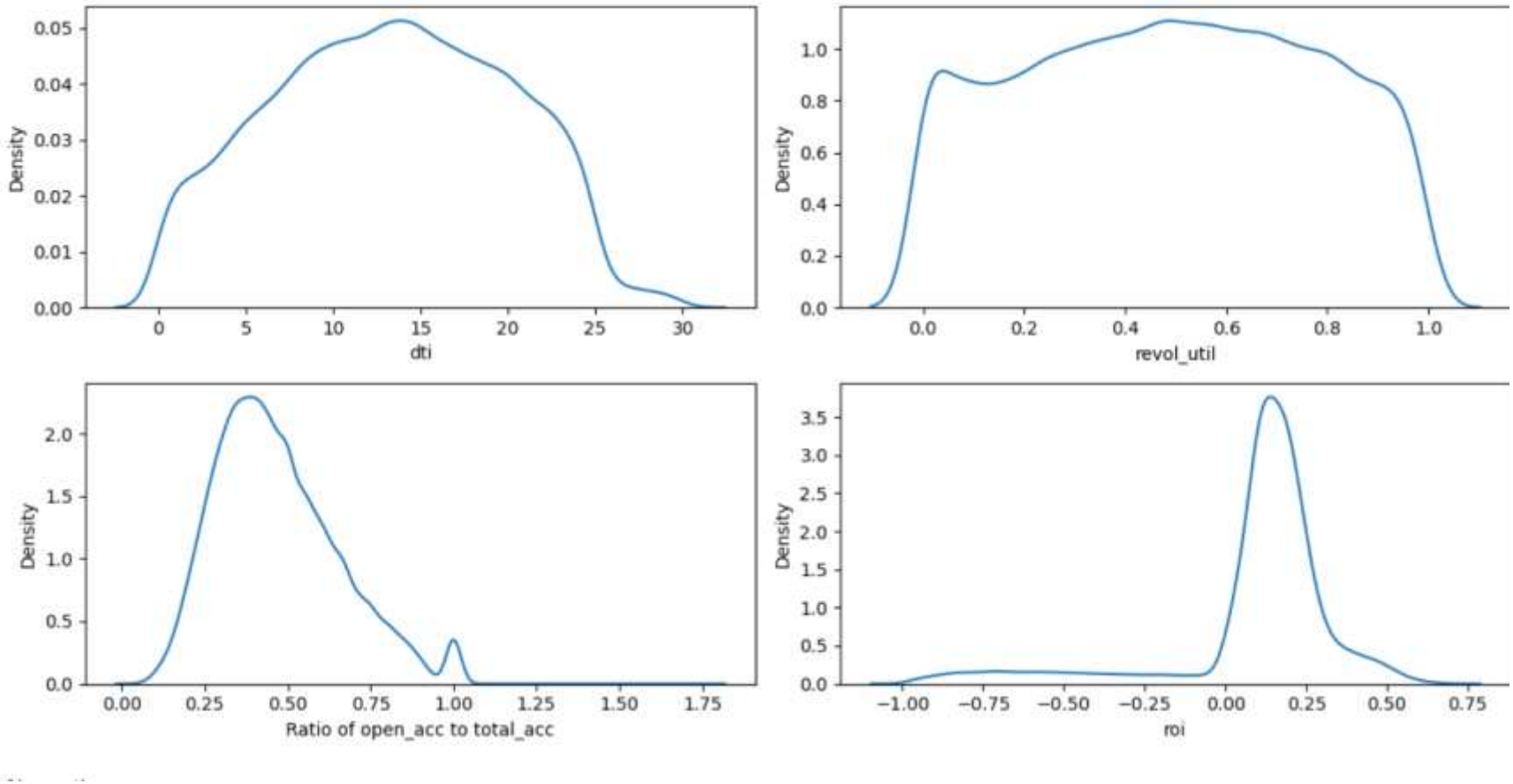
- Here we will see distribution of Annual income by employment length

1) Almost all 'employment length' category has same level of 25th, 50th(Median) and 75th precentile's.

2) It clearly shows that there are high annual income values in almost all 'employment length' category. We can treat these as outliers.



Frequency Distribution of Quantitative Variables



Frequency Distribution of Quantitative Variables

- **Observations:**

1) The average debt-to-income ratio is 13%. There doesn't seem to be much skew, considering the median is so close to the mean.

2) The mean of revolving credit utilization is 49%, which means the average borrower is using most of their revolving credit at a time when they are seeking the loan. Also, there is large spread of data but not much skewed.

3) The percentage of open accounts to total accounts seems left skewed.

4) Negative ROI indicates defaulted loans while almost all of the loans with positive ROI were fully paid.

5)Summary of Univariate Observations: There are 40 variables (so far) with 39717 rows.

6)Number of loans issued increased steadily by every year with a slight decrease in 2008.

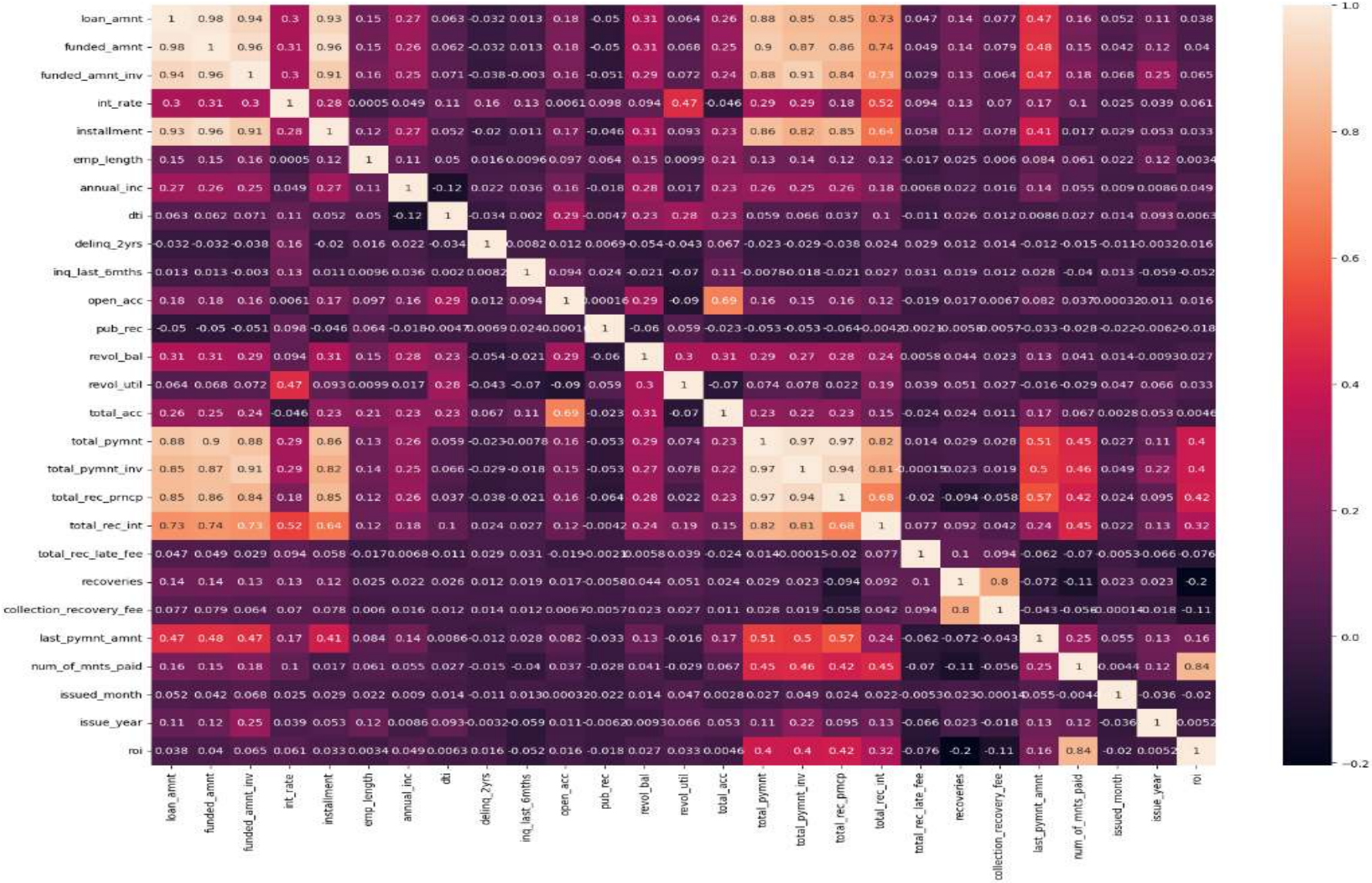
7) Of settled loans, 83% were Fully Paid and 14% were Charged Off.

8)Borrowers with own house and want to consolidate debt are not at much risk, but borrower with rent,mortgage and want to consolidate debt are at high risk applicants.

9) Majority of loans were from A, B, and C grade.

10) There is an inverse relationship between interest rate and loan grade - lower grades have higher interest rate.

Bivariate Analysis on Continuous Variables



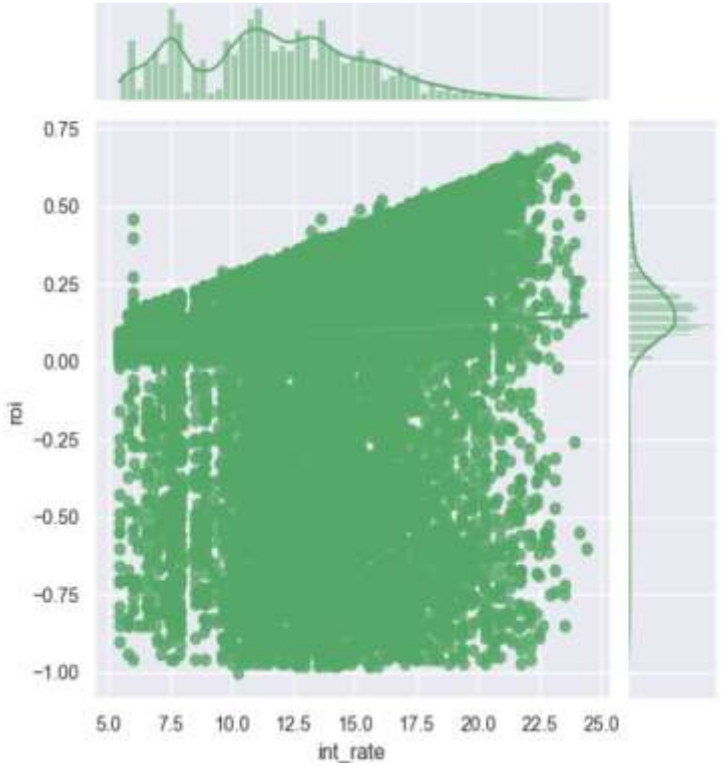
Bivariate Analysis on Continuous Variables

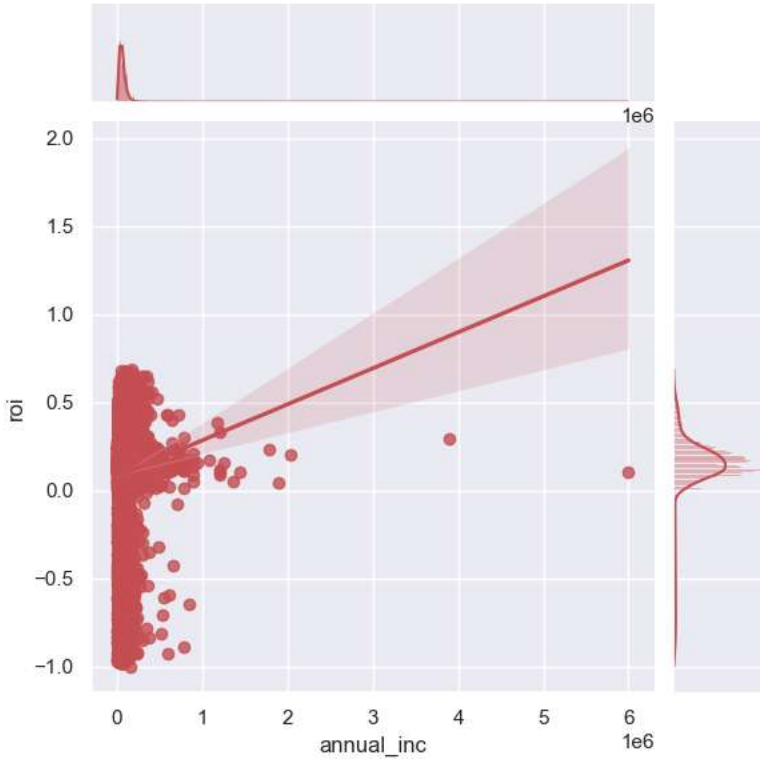
- From above correlation plot it show : -

1) There are no highly negative correlation between columns.
2) There are some highly correlated columns at top and middle left block.
3) There is 0.45 correlation between term_months and interest rate(int_rate).
4) There is 0.47 correlation between revol_util and interest rate(int_rate).
5) There is no much correlation between Annual Income, DTI with other columns.
6) Let's check the bivariate distribution for below columns:
7) Interest rates vs ROI Annual income vs ROI

Bivariate Analysis on Continuous Variables

ROI Vs Int_rate & ROI Vs Annual_income





There was a positive correlation between interest rates and ROI for borrowers who paid their loans, but for those who didn't pay the ROI was negative and it shows no correlation with interest rates.

Borrowers with higher income has positive ROI while borrowers with lower income has negative ROI.

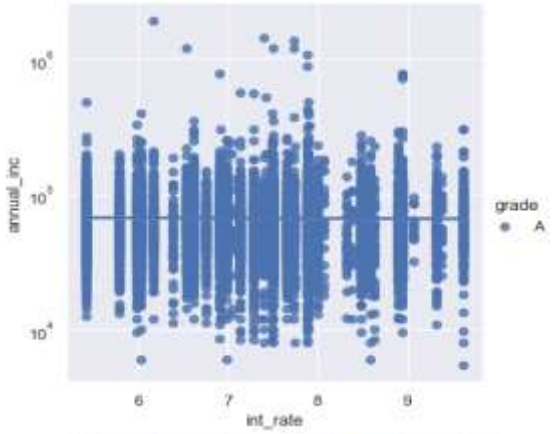
Bivariate Analysis on Continuous Variables

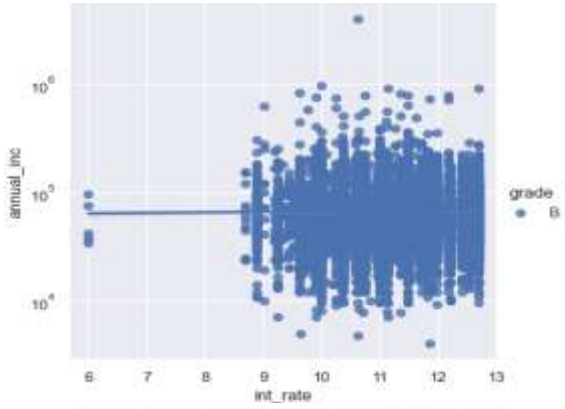
Annual Income Vs Interest Rate distribution by each grade

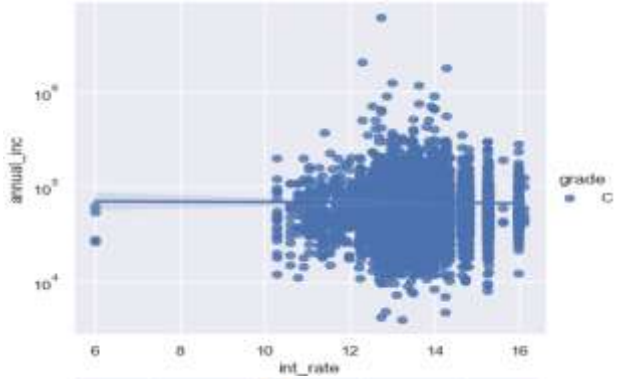
- Grade A has borrowers with interest rate between 5-10%.
- Grade B has borrowers with interest rate between 8-13%, but few are there with near 6%.
- Grade C has borrowers with interest rate between 10-16%, but few are there with near 6%.
- Grade D has borrowers with interest rate between 12-18%, but few are there with near 6%.
- Grade E has borrowers with interest rate between 14-20%, but few are there with near 6%.
- Grade F has borrowers with interest rate between 15-23%.
- Grade G has borrowers with interest rate between 17-24%.
- It is evident that the interest rate range changes from one grade to other, and also there are few borrowers with near 6% falling in B,C,D and E grade's.

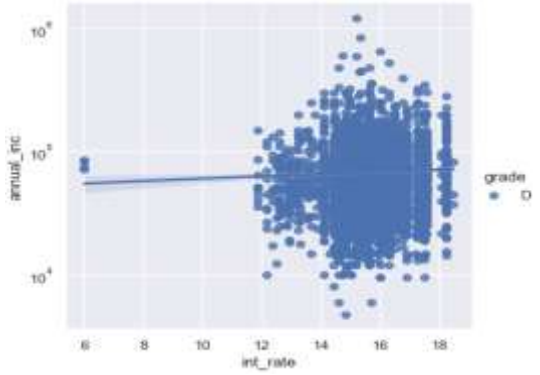
Bivariate Analysis on Continuous Variables

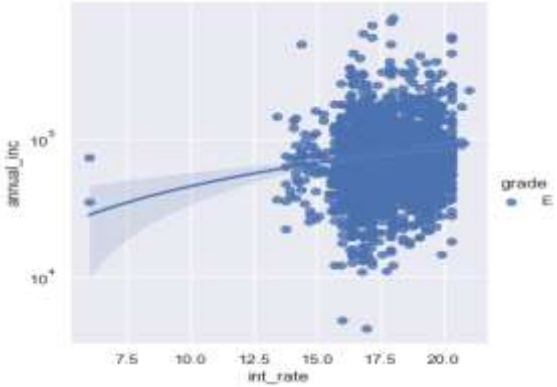
Annual Income Vs Interest Rate distribution by each grade

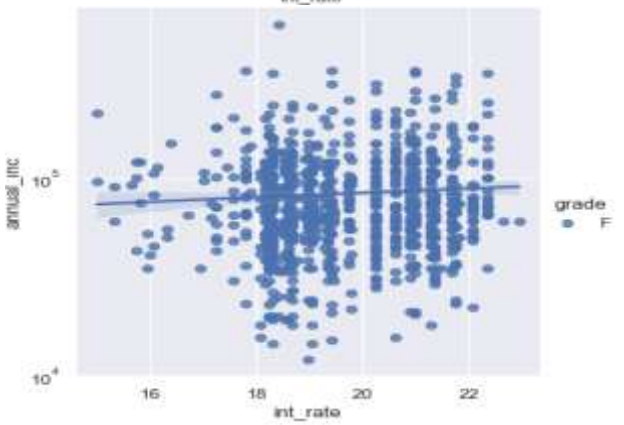


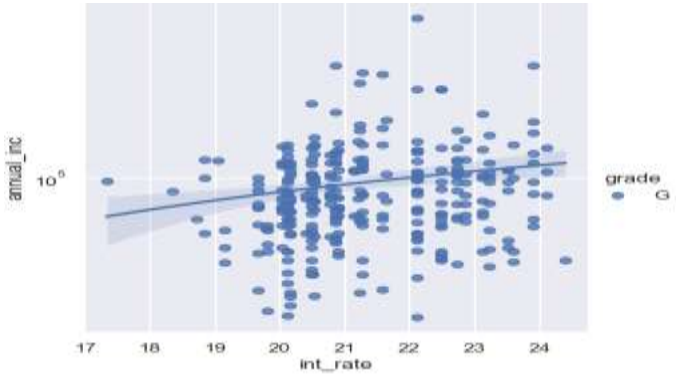












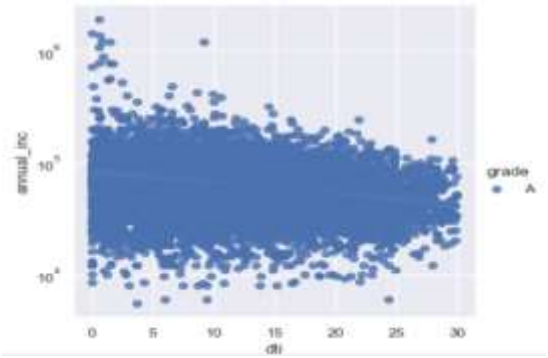
Bivariate Analysis on Continuous Variables

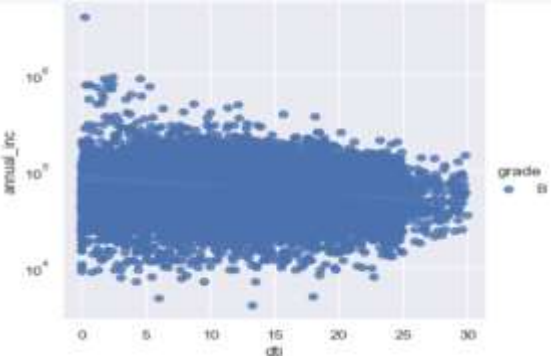
Annual Income Vs Interest Rate distribution by each grade

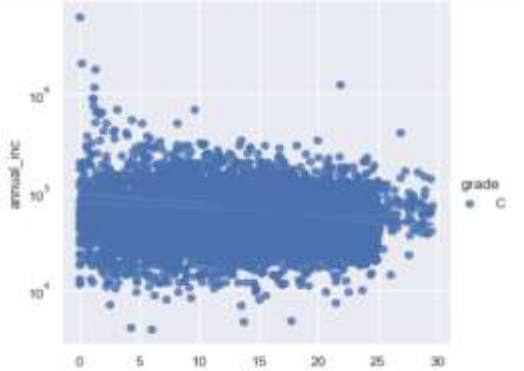
- Grade A has borrowers with interest rate between 5-10%.
- Grade B has borrowers with interest rate between 8-13%, but few are there with near 6%.
- Grade C has borrowers with interest rate between 10-16%, but few are there with near 6%.
- Grade D has borrowers with interest rate between 12-18%, but few are there with near 6%.
- Grade E has borrowers with interest rate between 14-20%, but few are there with near 6%.
- Grade F has borrowers with interest rate between 15-23%.
- Grade G has borrowers with interest rate between 17-24%.
- It is evident that the interest rate range changes from one grade to other, and also there are few borrowers with near 6% falling in B,C,D and E grade's.

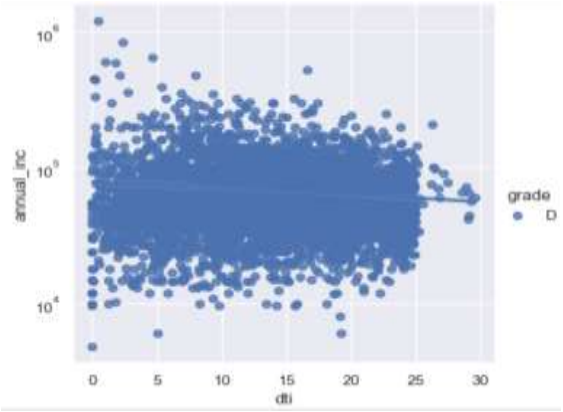
Bivariate Analysis on Continuous Variables

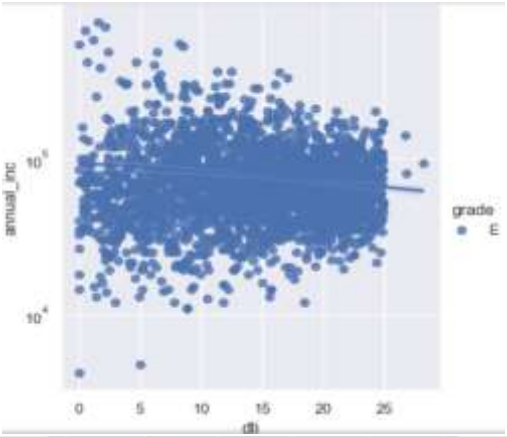
DTI VS annual income for each grade

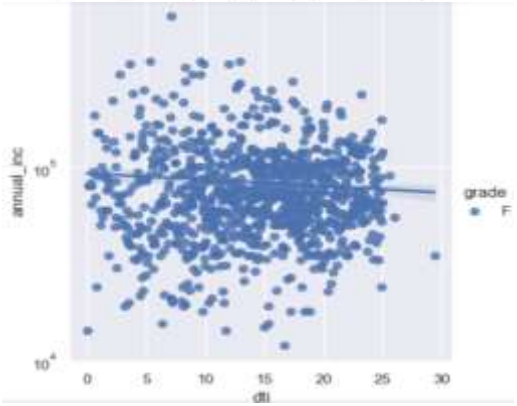


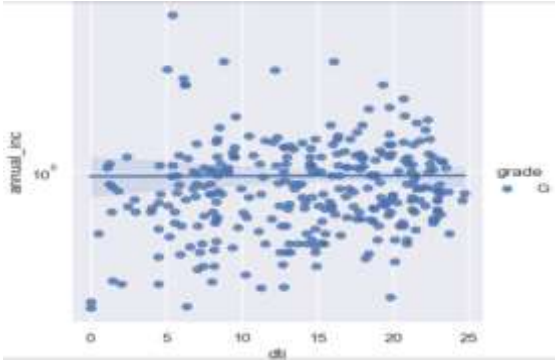












Bivariate Analysis on Continuous Variables

- Observations:

1) It seems in A,B,C grades the data is spreaded and have DTI from 0-30.

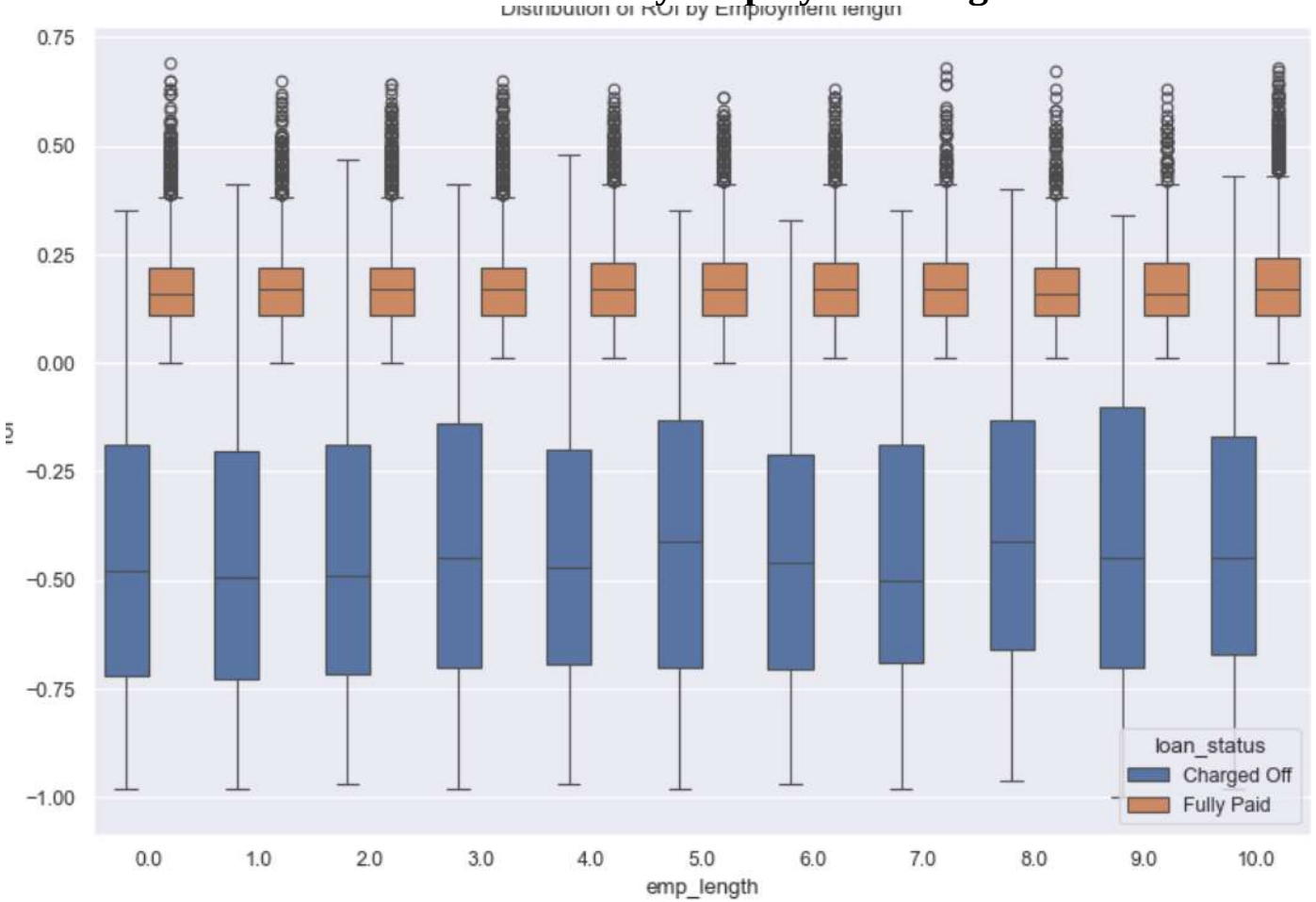
2) Lower number of borrowers from 25-30 DTI in D,E grades.

3) Grade G has maximum 25 DTI.

4) It seems there is no DTI range constraint from one grade to other similar to interest rate which we have seen above.

Bivariate Analysis on Categorical Variables

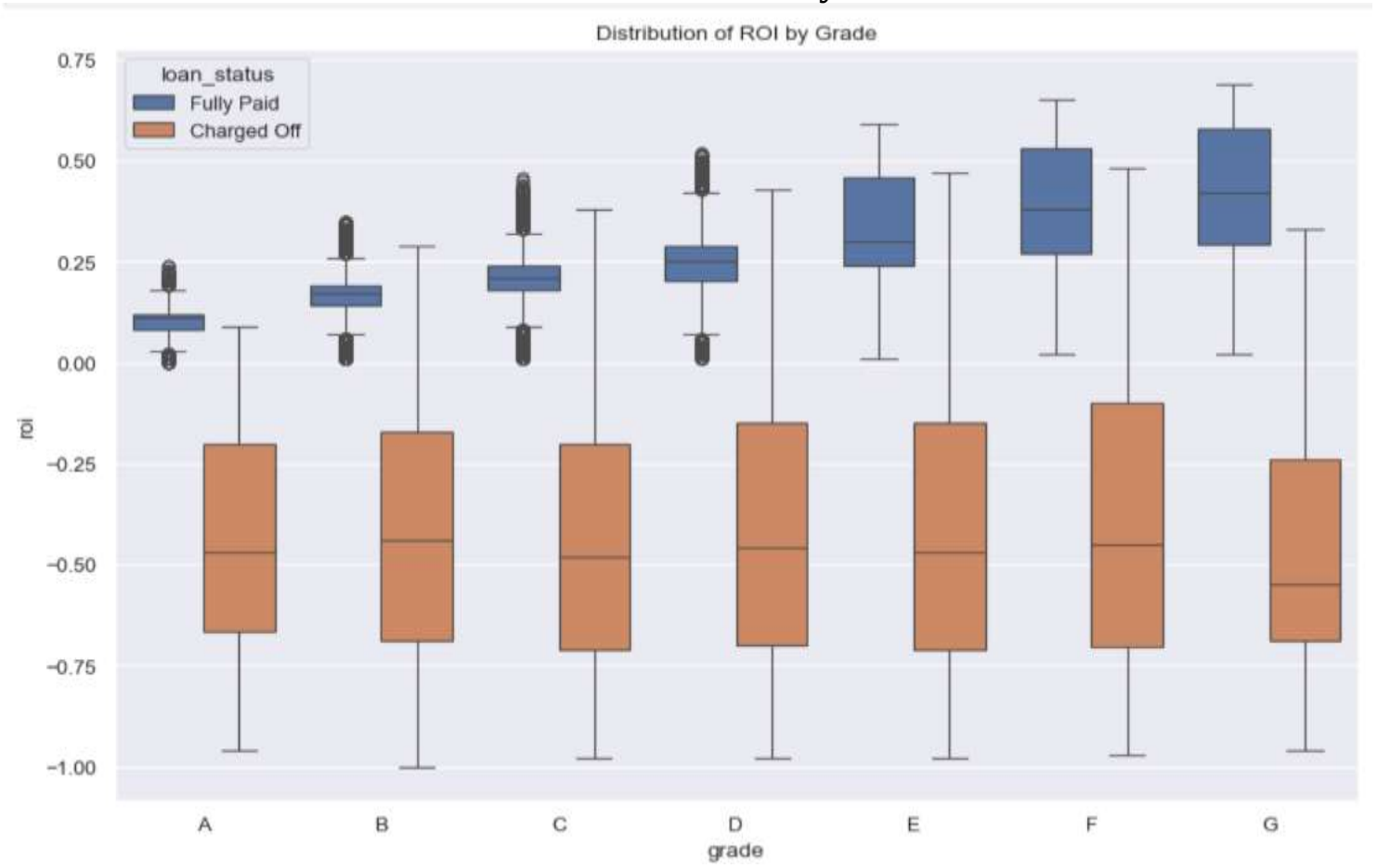
Distribution of ROI by Employment length



1) From fully_paid status loan, almost all employment length have same mean ROI.
2) From Charged_off status loan 1 years, 2 years and 7 years have low mean ROI.

Bivariate Analysis on Categorical Variables

Distribution of ROI by Grade

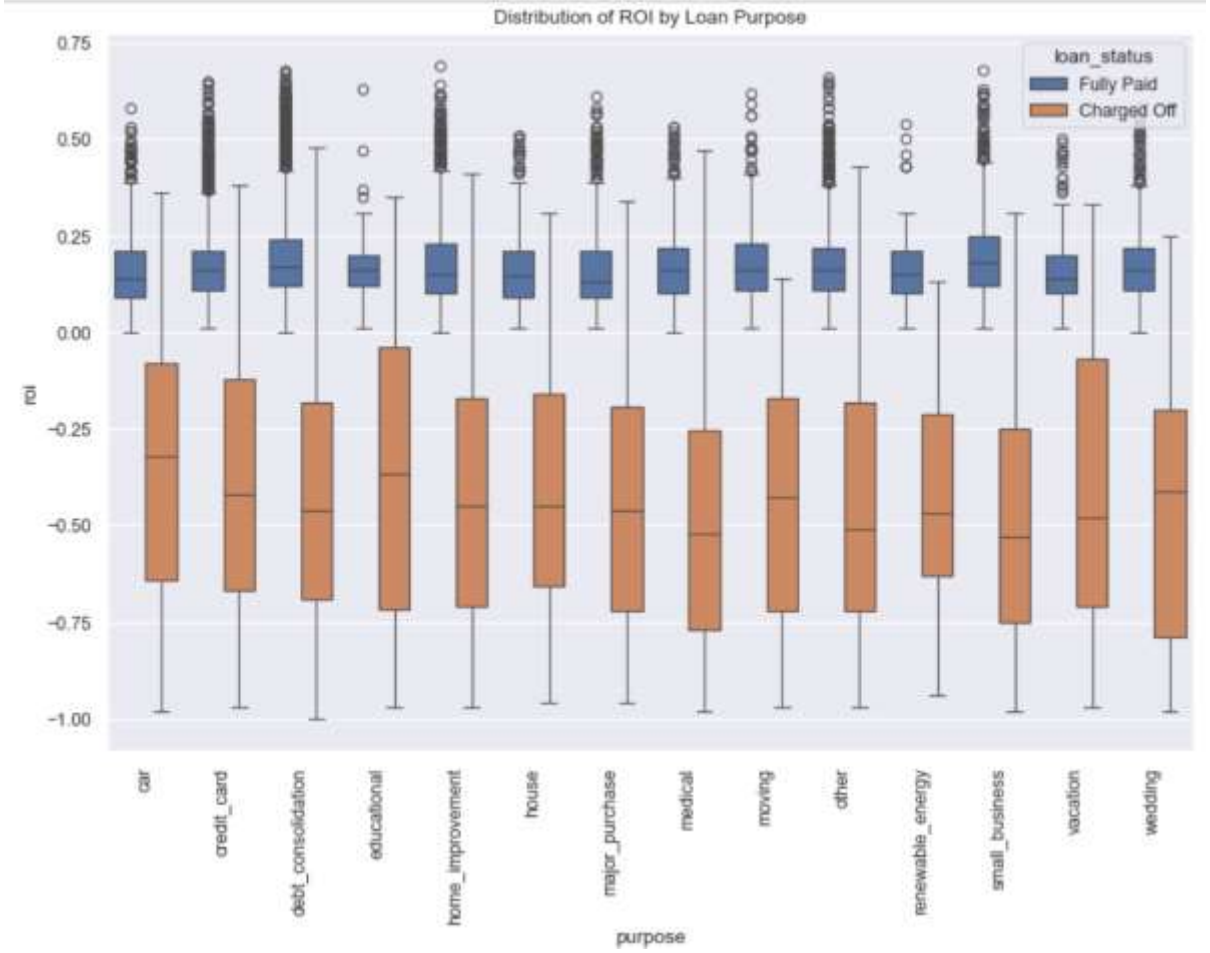


1) From fully_paid status loan, grades E,F,G have high mean ROI, this might be high interest rates.

2) From Charged_off status loan, grade G has low mean ROI.

Bivariate Analysis on Categorical Variables

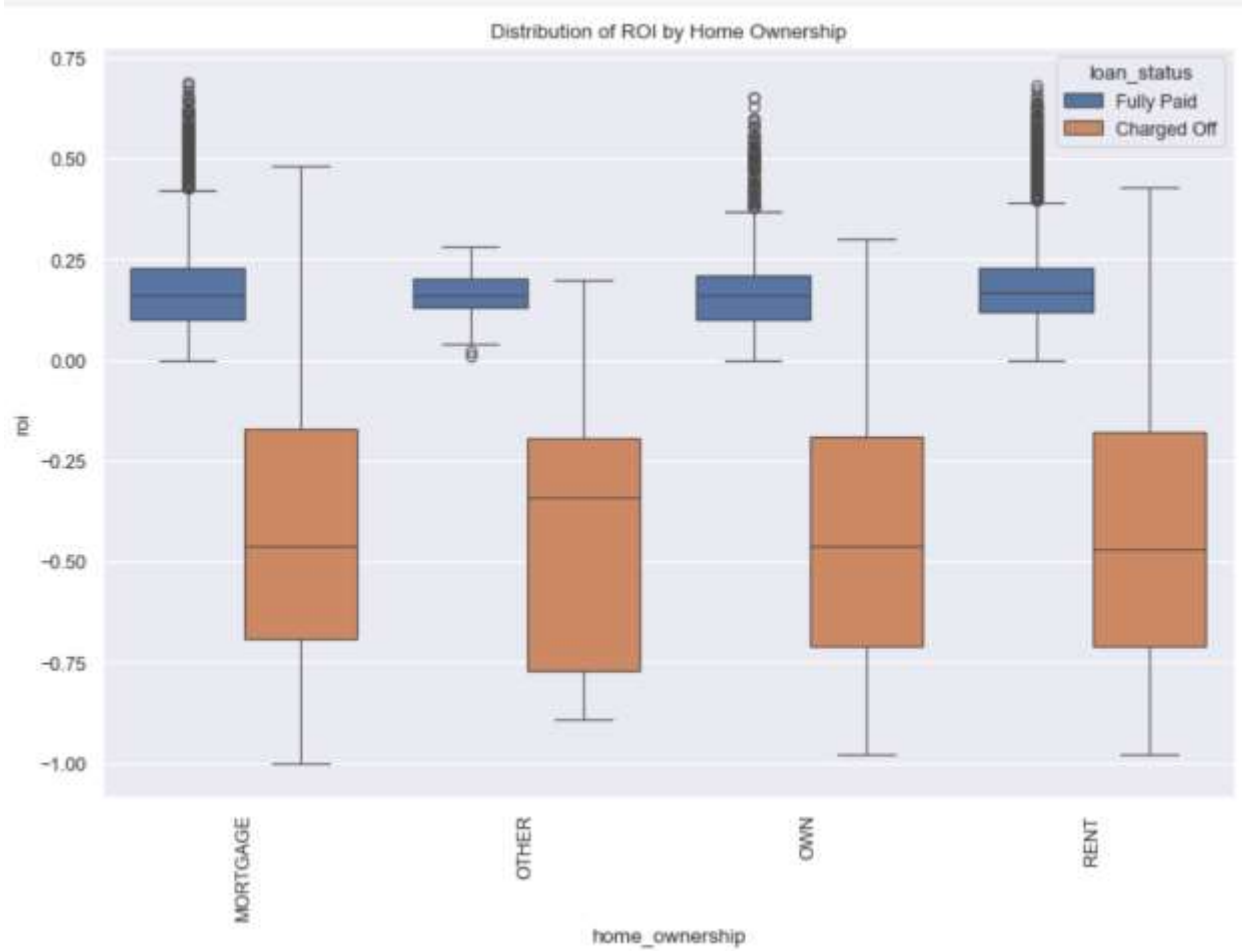
Distribution of ROI by Loan Purpose



1) From fully_paid status loan, almost all categories have nearby mean ROI.
2) From Charged_off status loan, small_business, other and medical have low mean ROI.

Bivariate Analysis on Categorical Variables

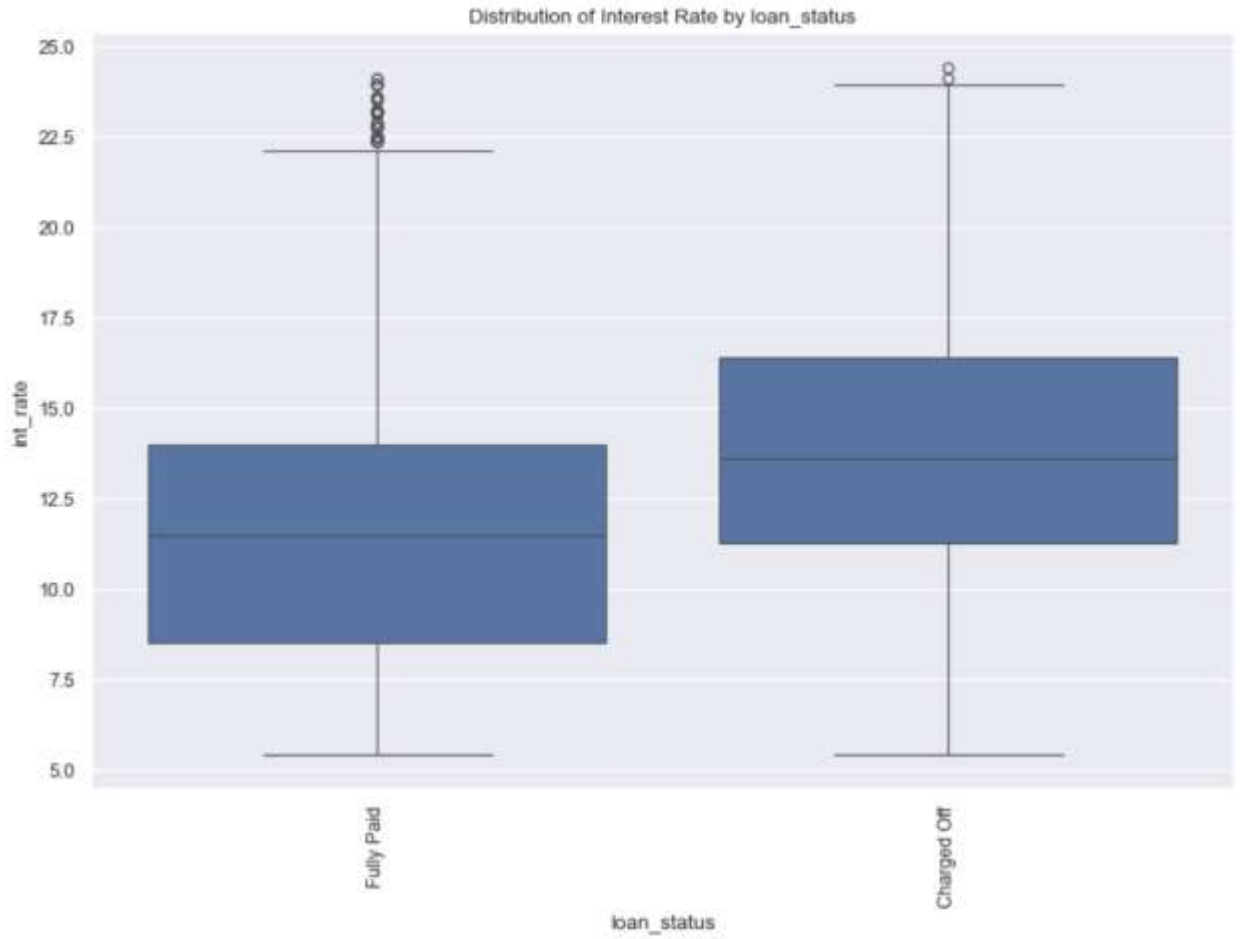
Distribution of ROI by Home Ownership



From Charged_off status loan, Rent and Mortgage have low mean ROI.

Interest Rate Bivariate Distribution Plots

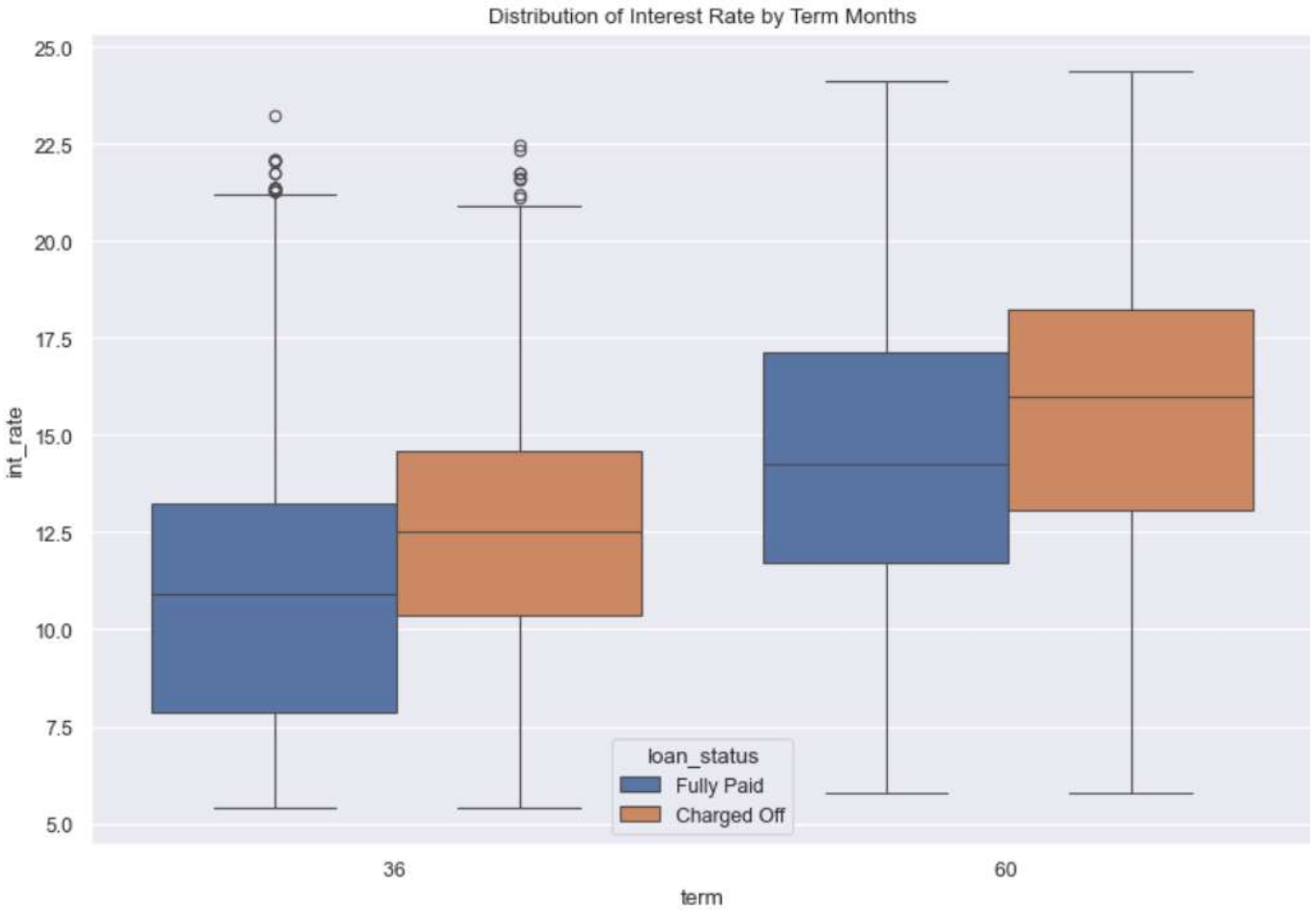
Distribution of Interest Rate by loan_status



Observations: It shows that the interest rate is more from defaulters.

Interest Rate Bivariate Distribution Plots

Distribution of Interest Rate by Term Months



1) It is evident that 60 months term loans have more interest rate.
2) There are more defaulters in both 36, 60 month terms because of high interest rates.

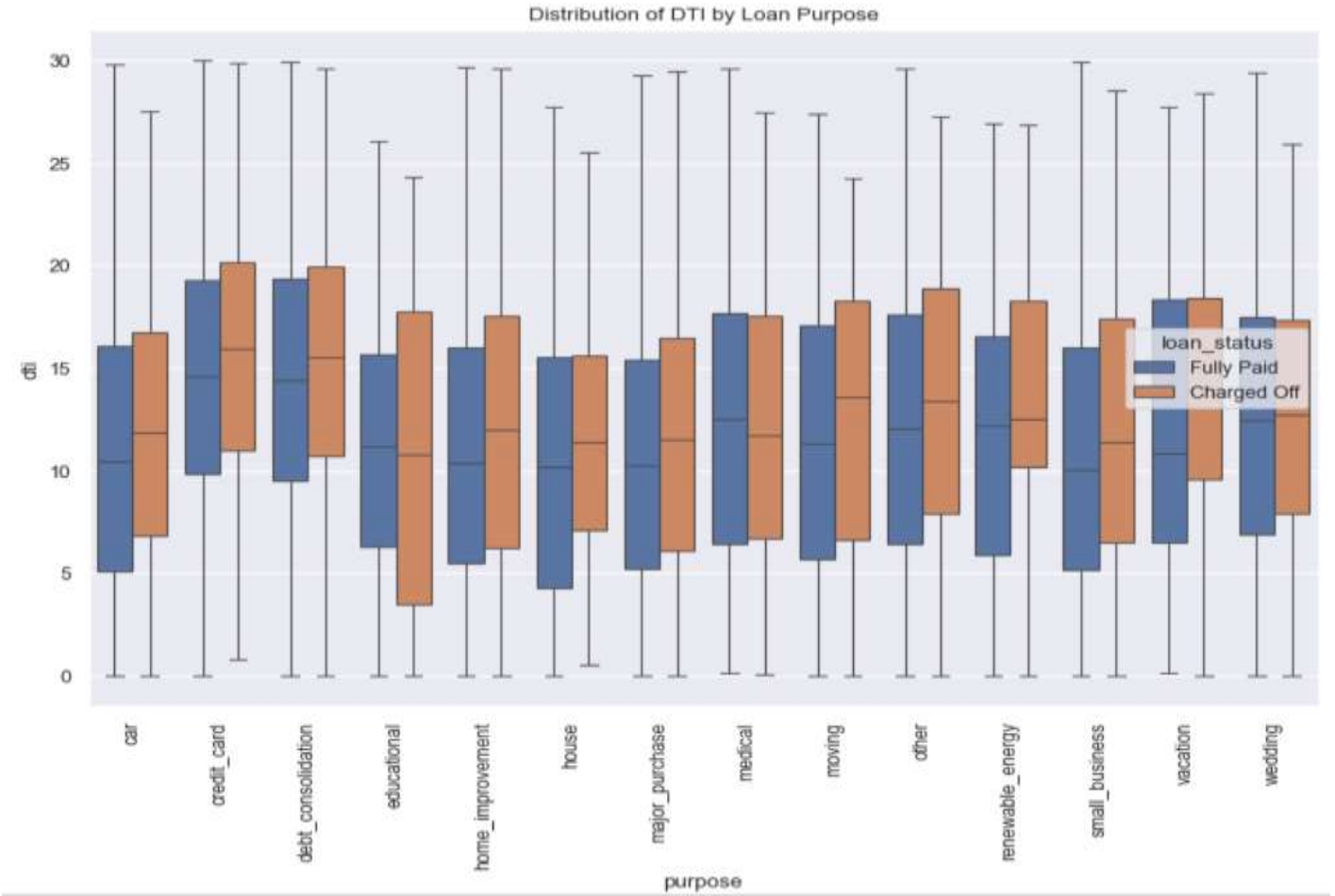
DTI Bivariate Distribution Plots

- DTI is less in lower segment, as the annual income is less and they might not have any loans and hence less DTI. If we have age details then we can predict more with borrower age, income and employment length.
- DTI is high in lower middle and middle segments.
- DTI is less in upper middle and upper segments, as the income is high they might have financial freedom. Note that there are some borrower in upper segment who have DTI more than 25.
- Almost in all categories of purpose, defaulter's DTI is high than fully paid borrowers.

DTI Bivariate Distribution Plots

Distribution of DTI by Loan Purpose

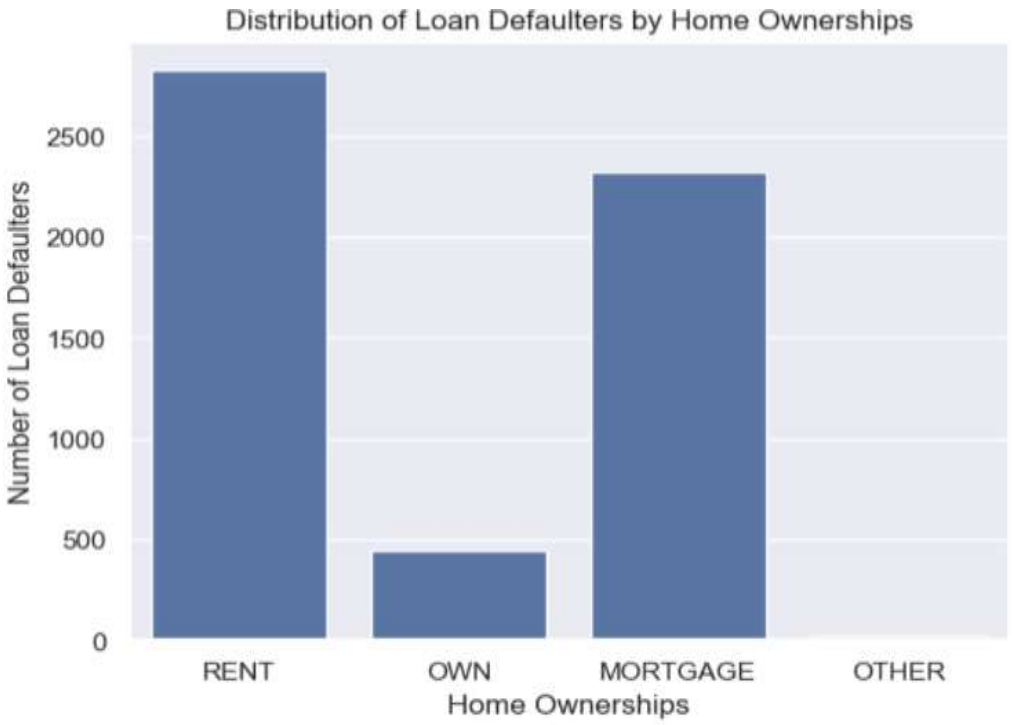
- Almost in all categories of purpose, defaulter's DTI is high than fully paid borrowers.



Loan Defaulters Analysis

Distribution of Loan Defaulters by Home Ownerships

- It shows there are more defaulters in RENT and MORTGAGE. let's check it in granular level.

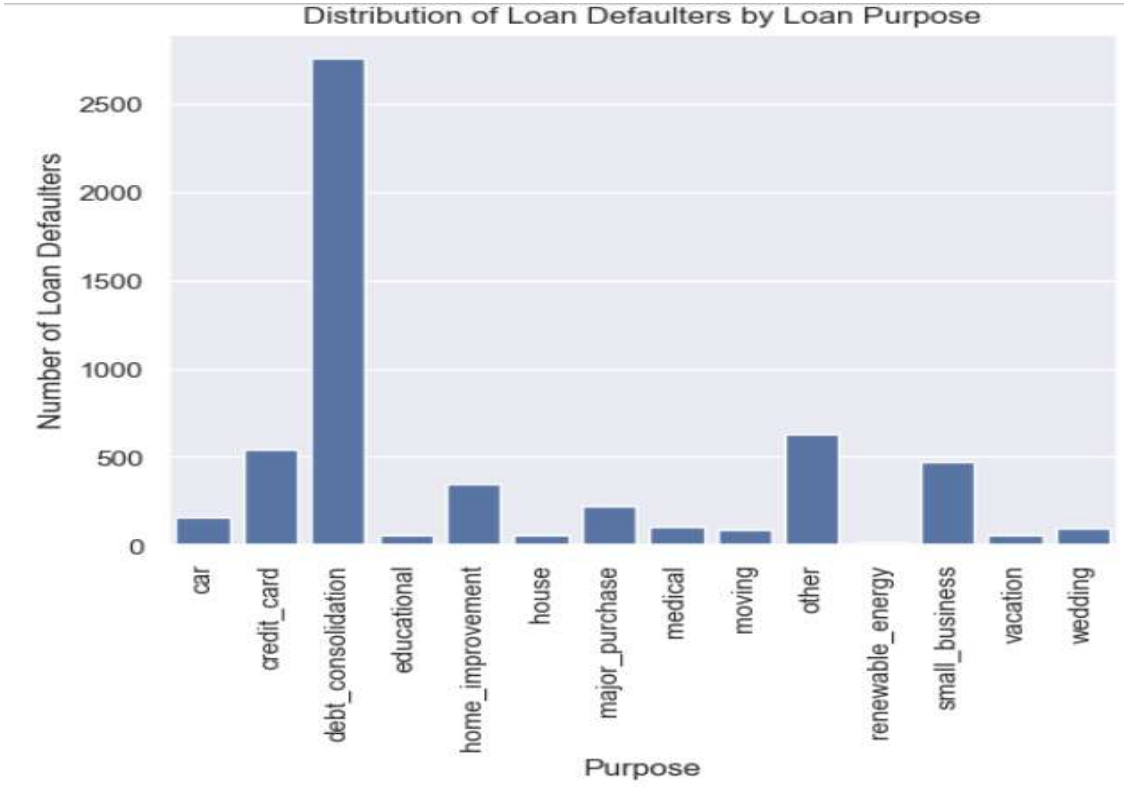



```
home_ownership
MORTGAGE    2325
OTHER         18
OWN          441
RENT        2827
Name: home_ownership, dtype: int64
```

Loan Defaulters Analysis

Distribution of Loan Defaulters by Loan Purpose

- There are more defaulters from 'debt_consolidation','other', 'credit_card' and 'small_business'. Let's check this at granular level by combining with Home Ownership and by each Grade.

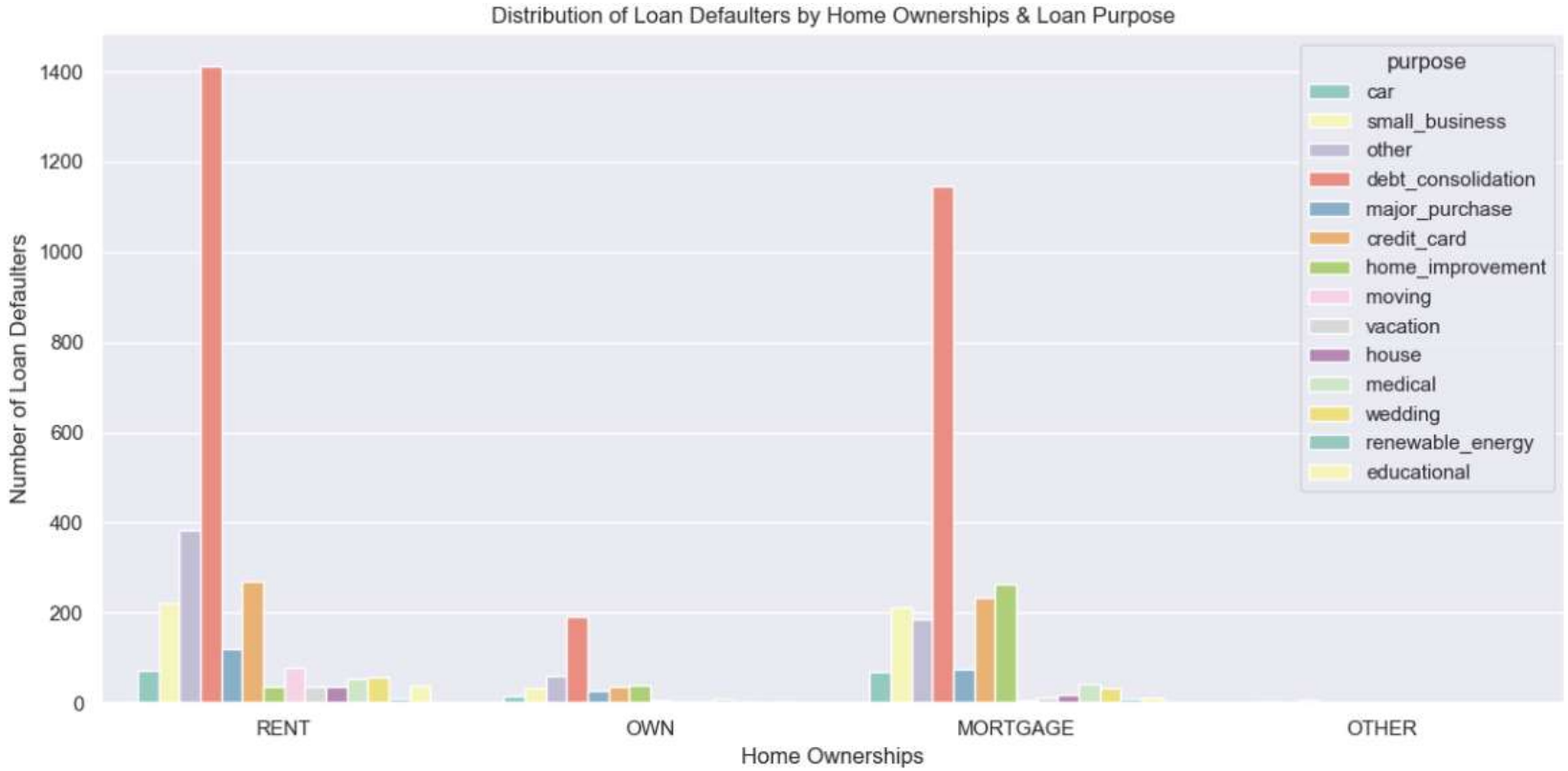



```
purpose
car                    160
credit_card            542
debt_consolidation    2757
educational             56
home_improvement       345
house                   59
major_purchase         222
medical                106
moving                  92
other                  632
renewable_energy        19
small_business         473
vacation                53
wedding                 95
Name: purpose, dtype: int64
```

Loan Defaulters Analysis

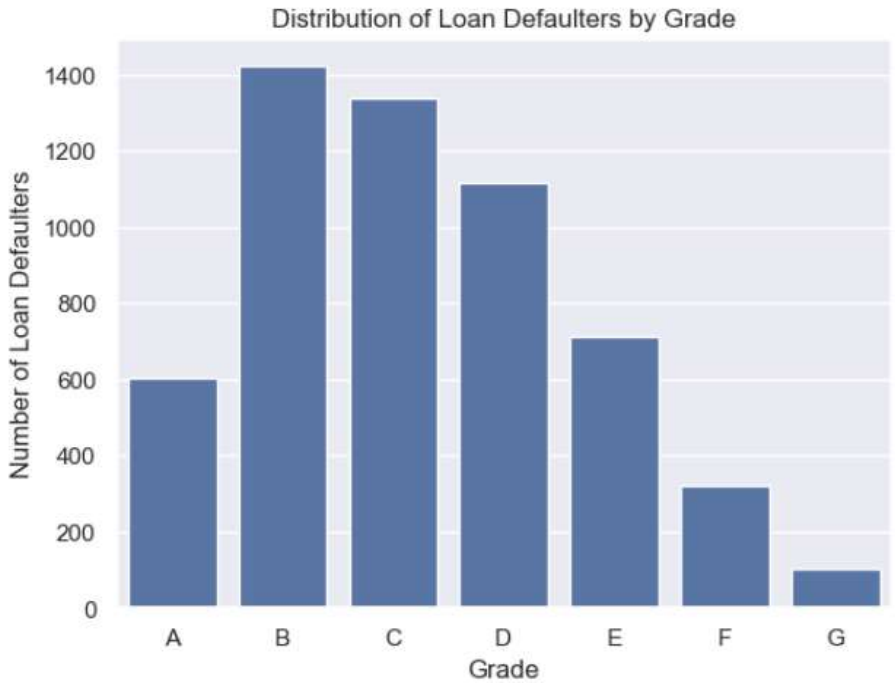
Distribution of Loan Defaulters by Home Ownerships & Loan Purpose

- From RENT category, there are more defaulters from 'debt_consolidation','other', 'credit_card' and 'small_business'.
- From MORTGAGE category, there are more defaulters from 'debt_consolidation','home_improvement', 'credit_card' and 'small_business'.
- Overall, one should be carefull with 'debt_consolidation', 'credit_card' and 'small_business' loans when the borrowers dont have own house.



Loan Defaulters Analysis
Distribution of Loan Defaulters by Grade

- It shows there are more defaulters in B,C and D grades.
- Grades F,G(more interest rate grades) are having less defaulters which is a good indicator.
- Let's check it in granular level.

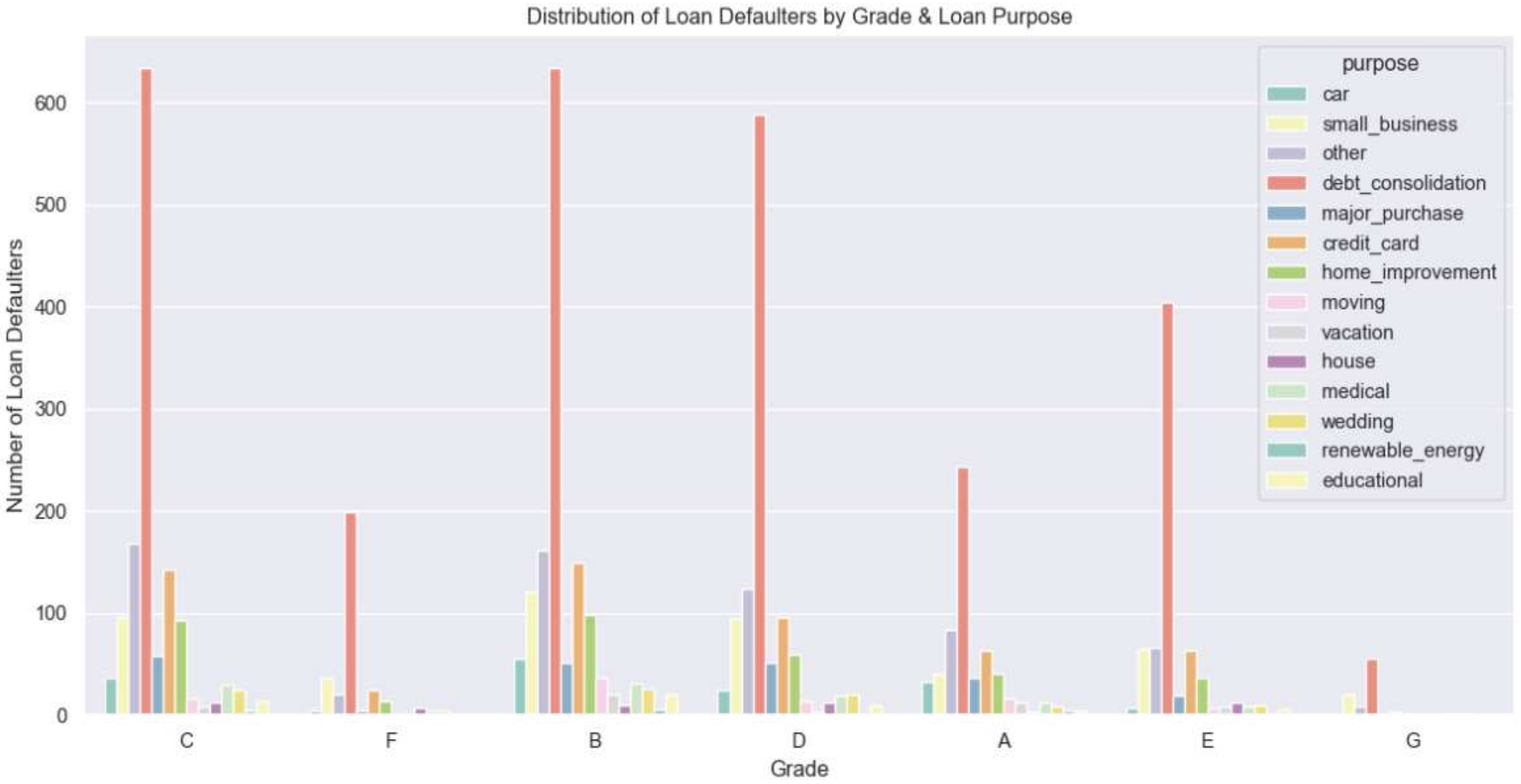



```
grade
A     602
B    1422
C    1339
D    1116
E     712
F     319
G     101
Name: grade, dtype: int64
```

Loan Defaulters Analysis

Distribution of Loan Defaulters by Grade & Loan Purpose

From all grades, there are more defaulters from 'debt_consolidation', 'others', 'credit_card' and 'small_business' purpose loans.



Summary of Bivariate Observations:

- **Bivariate Analysis on Continuous Variables:**

Correlation Plot:

a) There were no highly negative correlation between numerical columns.
b) 0.45 correlation between term_months and interest rate(int_rate).
c) 0.47 correlation between revol_util and interest rate(int_rate).
d) No much correlation between Annual Income, DTI with other numerical columns.

ROI vs Interest rates and Annual income:

a) There was a positive correlation between interest rates and ROI for borrowers who paid their loans, but for those who didn't pay the ROI was negative and it shows no correlation with interest rates.
b) Borrowers with higher income has positive ROI while borrowers with lower income has negative ROI.

Annual Income Vs Interest Rate by each grade:

By charts, it is evident that the interest rate range changes from one grade to other, and also there are few borrowers with near 6% falling in B,C,D and E grade's.

Annual Income Vs DTI distribution by each grade:

By charts, it shows there is no DTI range constraint from one grade to other similar to interest rate which we have seen above.

Bivariate Analysis on Categorical Variables:

- **ROI Bivariate Distribution plots:**

1) ROI vs Grade by loan_status: From fully_paid loan, grades E,F,G have high mean ROI, this is because of high interest rates.

2) ROI vs Loan Purpose by loan_status: a) From fully_paid loans, almost all categories have nearby mean ROI. b) From Charged_off loan, small_business, other and medical have low mean ROI.

3) ROI vs Home Ownership by loan_status: From Charged_off loans, Rent and Mortgage have low mean ROI.

4) ROI vs Income Segment by loan_status: a) From fully_paid loans, ROI is bit high for upper middle and upper Income_segment. b) From Charged_off loan, ROI is very low for lower Income_segment.

- **Interest Rate Bivariate Distribution Plots:**

1) Interest Rate vs Term Months by loan_status: a) It is evident that 60 months term loans have more interest rate. b) There are more defaulters in both 36, 60 month terms because of high interest rates.

2) Interest Rate vs Income Segment by loan_status: a) By charts, it shows that, at every income segment, the defaulters are due to high interest rates.

- **DTI Bivariate Distribution Plots:**

- DTI vs Income Segment by loan_status:

a) DTI is less in lower segment, as the annual income is less and they might not have any loans and hence less DTI. If we have age details then we can predict more with borrower's age, income and employment length.

b) DTI is high in lower middle and middle segments.

c) DTI is less in upper middle and upper segments, as the income is high they might have financial freedom. Note that there are some borrower's in upper segment who have DTI more than 25.

- **Loan Defaulters Analysis:**

- Distribution of Loan Defaulters by Loan Purpose:

a) There are more defaulters from 'debt_consolidation','other', 'credit_card' and 'small_business'. Let's check this at granular level by combining with Home Ownership and by each Grade.

- Distribution of Loan Defaulters by Home Ownerships:

a) By charts, it shows there are more defaulters in RENT and MORTGAGE.

- Distribution of Loan Defaulters by Home Ownerships & Loan Purpose:

a) There are more defaulters with 'debt_consolidation', 'credit_card' and 'small_business' purpose from Rent, Mortgage category and hence, should be carefull when borrowers don't have own house.

- Distribution of Loan Defaulters by Grade:

a) By charts, it shows there are more defaulters in B,C and D grades.

b) Grades F,G(more intereste rate grades) are having less defaulters which is a good indicator.

- Distribution of Loan Defaulters by Grade & Loan Purpose:

- By charts, from all grades, there are more defaulters from 'debt_consolidation', 'others', 'credit_card' and 'small_business' purpose loans.

- **Conclusion:**

1) Number of loans issued increased steadily by every year with a slight decrease in 2008.

2) Of settled loans, 83% were Fully Paid and 14% were Charged Off.

3) Borrowers with own house and the purpose of loan with consolidate debt, 'credit_card' and 'small_business' are not at much risk, but borrower with rent,mortgage are high risk applicants.

4) Majority of loans were from A, B, and C grade.

5) There is an inverse relationship between interest rate and loan grade - lower grades(E,F,G) have higher interest rate.

6) From ROI analysis, it shows that borrowers with the best credit profiles or the lowest loan amounts will not end up being the most profitable and the borrowers who seemed to have worst credit indicators ended up being more profitable from E,F,G grades.

7) Overall, there are more defaulters from 'debt_consolidation', 'others', 'credit_card' and 'small_business' purpose loans from all grades.