Lending Club Case Study

Prepared by:
Lalithanjali
Dinesh

The Problem

Problem statement

You work for a consumer finance company which specialises in lending various types of loans to urban customers.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.

Dataset

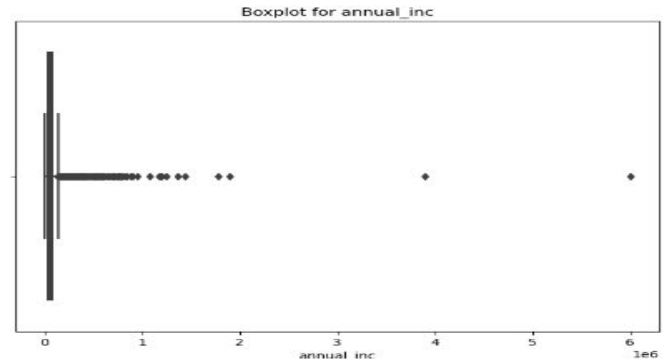
The dataset contains the complete loan data for all loans issued through the time period 2007 t0 2011.The data given contains information about past loan applicants and whether they ‘defaulted’ or not.

Data Understanding

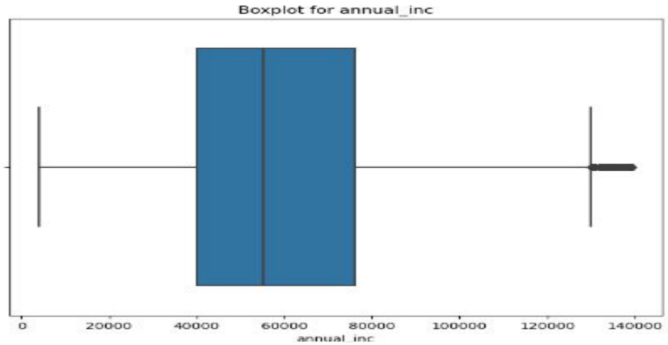
- The dataset contains information about past loan applicants and their loan statuses, categorized as 'Fully Paid', 'Charged Off', and 'Current'. The objective is to identify the factors influencing default tendencies. Since the status of 'Current' customers is uncertain, they will be excluded from analysis, focusing only on 'Fully Paid' and 'Charged Off' cases. 'Charged Off' denotes defaulters.
- Our analysis involves data preprocessing like handling missing values, handling outliers and visualizing the factors to get more insights for driving factors which results to loan default.
- In the dataset the target variable which we want to compare across the independent variables, is loan status.
- There are several columns contain null values. We need to drop the columns which contain missing values.

Data Cleaning

```
Before Outlier removal:
```



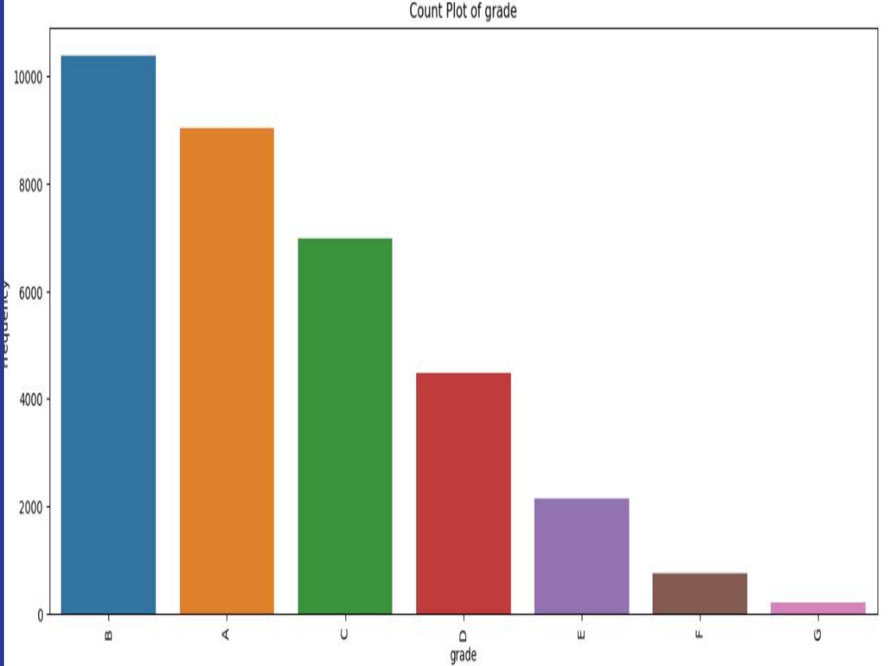
```
After Outlier removal:
```

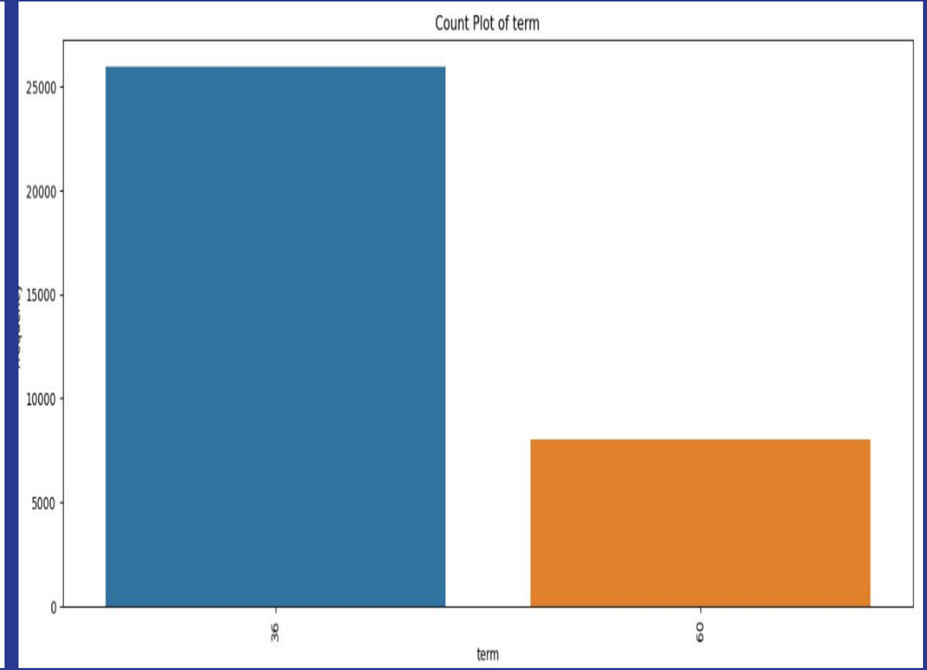


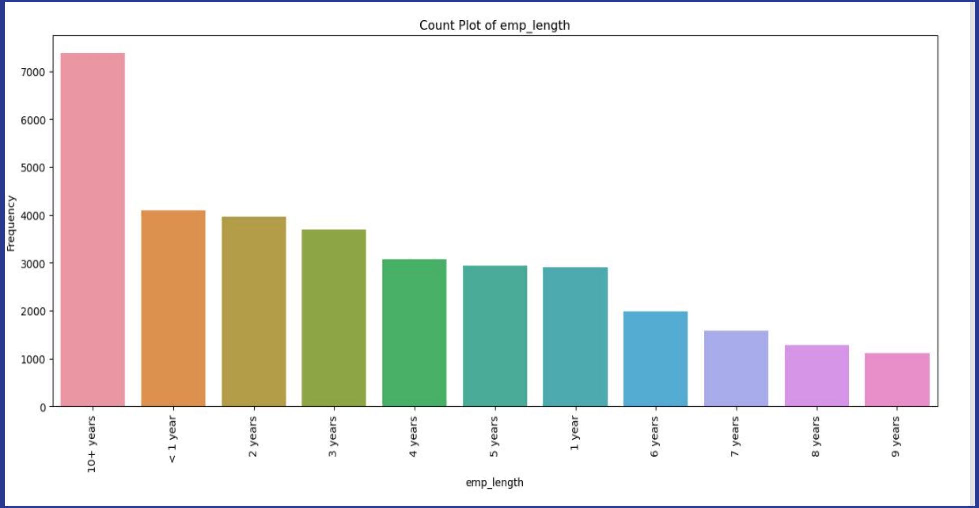
- Removed the columns which contain more missing values.
- Handled outliers with threshold 1.5 value.
- Extracted some new columns like issue_year,issue_month for better insights.
- Removed unnecessary columns which are not required for analysis.
- Handled datatype format types for int_rate,issue_d,annual_inc
- Filtering the dataset based on loan_status being either "Charged Off" which is default or "Fully Paid", as the status "Current" applies to customers whose loans are currently active and doesn't provide definitive information on whether they will be fully paid or charged off by the end of the loan term.

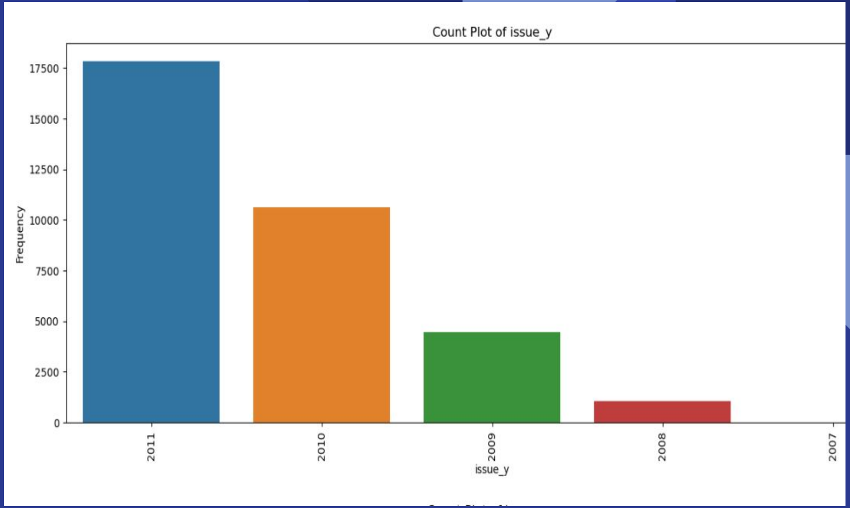
Univariate Analysis

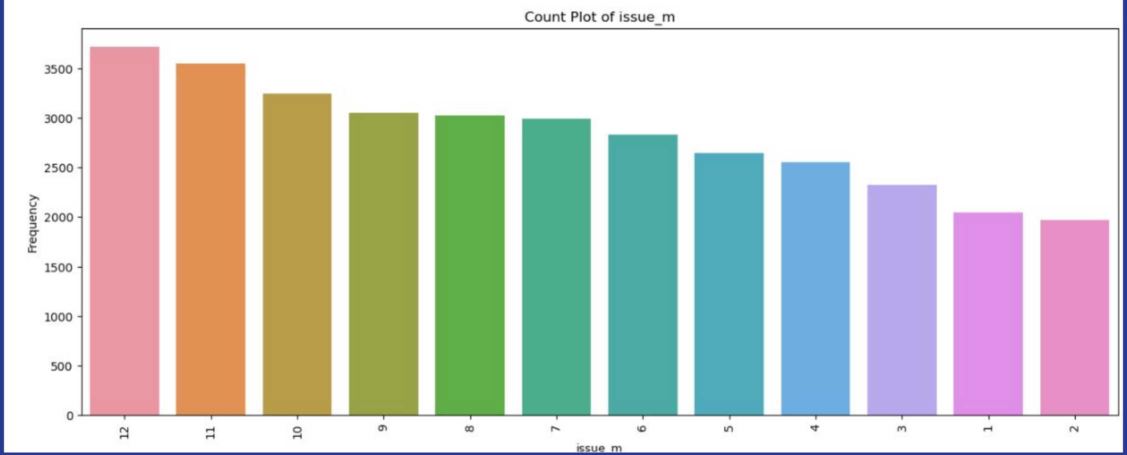
Conducted univariate analysis on individual columns for categorical(ordered and unordered) and numeric variables to get more insights about loan information.

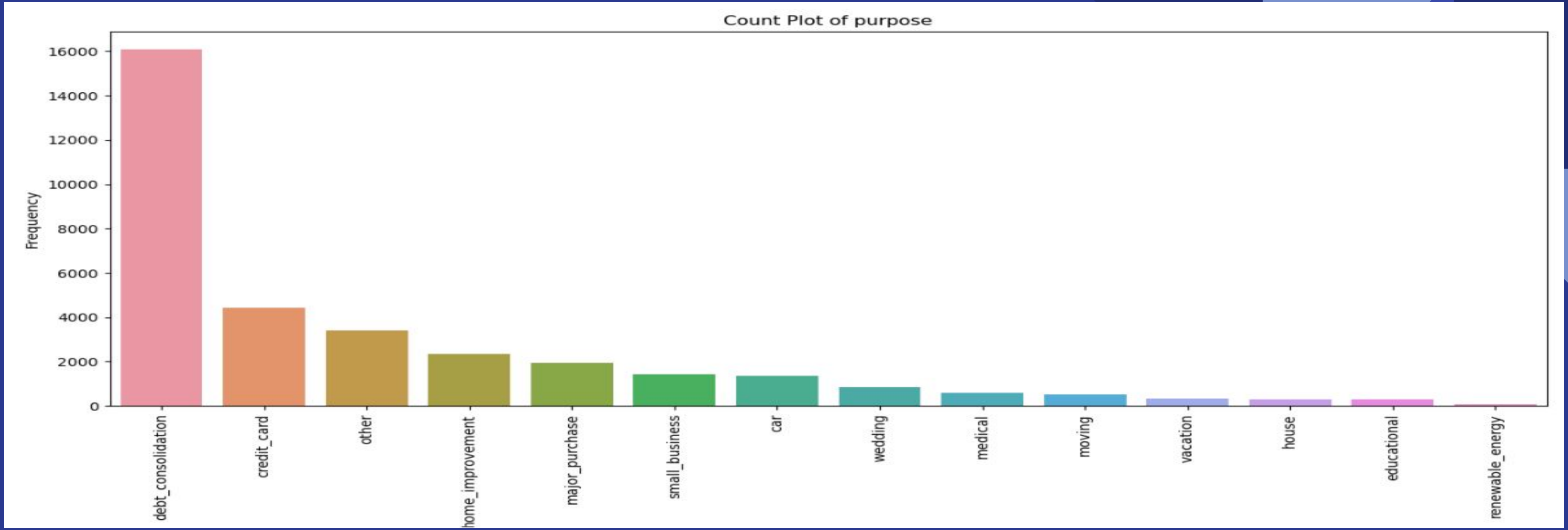


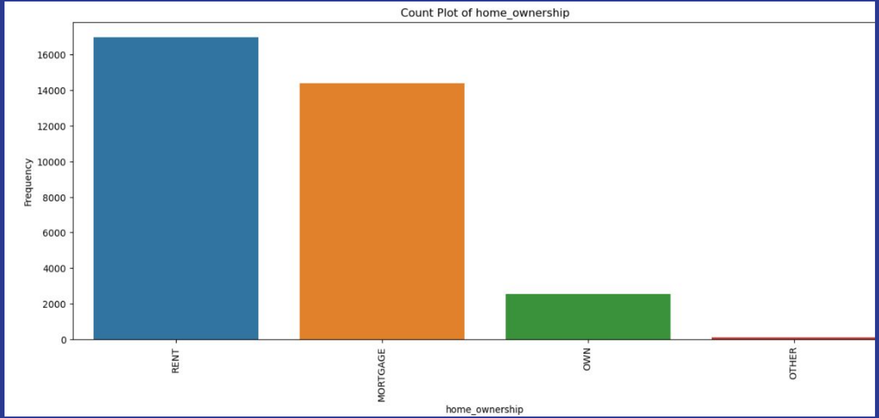


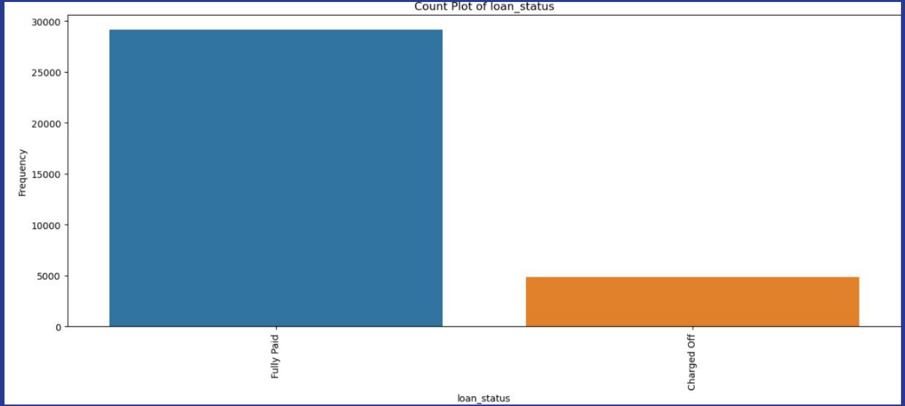

Count Plot of emp_length
Frequency
0
1000
2000
3000
4000
5000
6000
7000
10+ years
< 1 year
2 years
3 years
4 years
5 years
1 year
6 years
7 years
8 years
9 years
emp_length

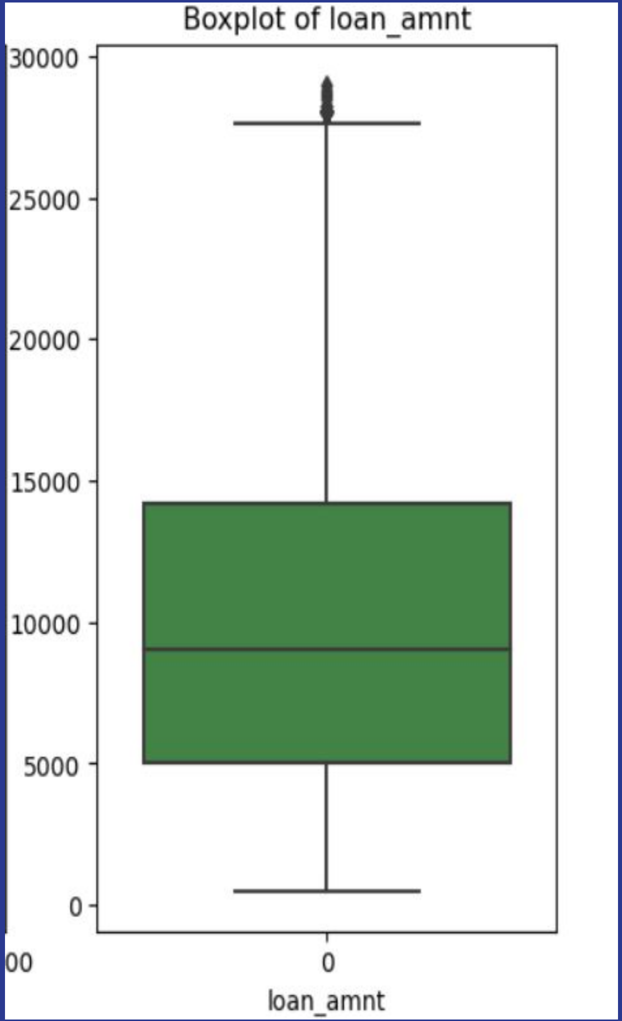

Count Plot of issue_y
Frequency
0
2500
5000
7500
10000
12500
15000
17500
2011
2010
2009
2008
2007
issue_y

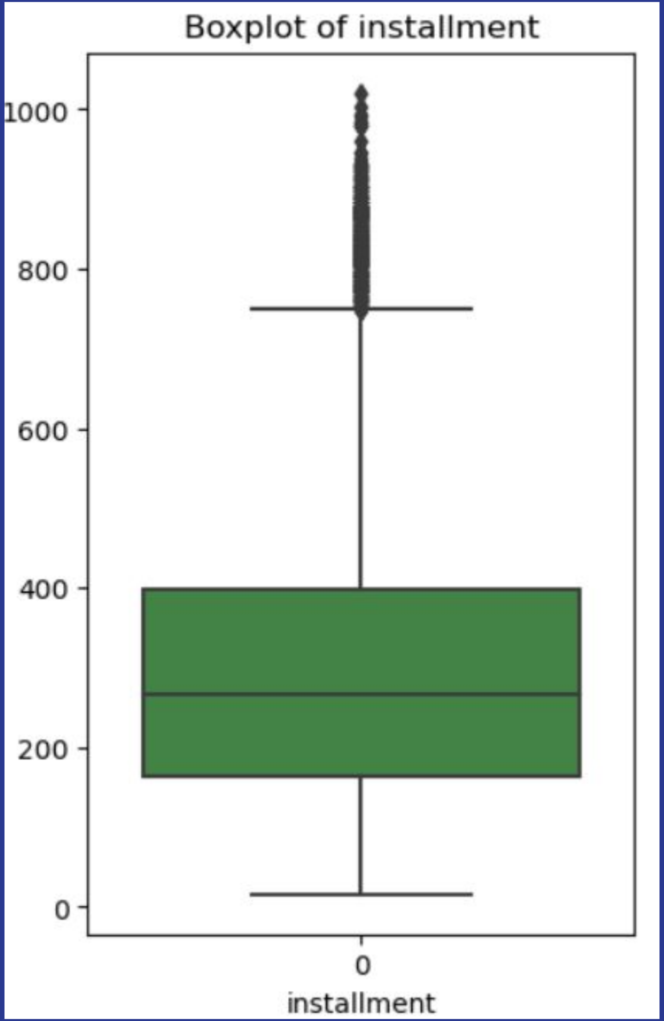

Count Plot of issue_m
Frequency
0
500
1000
1500
2000
2500
3000
3500
12
11
10
9
8
7
6
5
4
3
1
2
issue_m

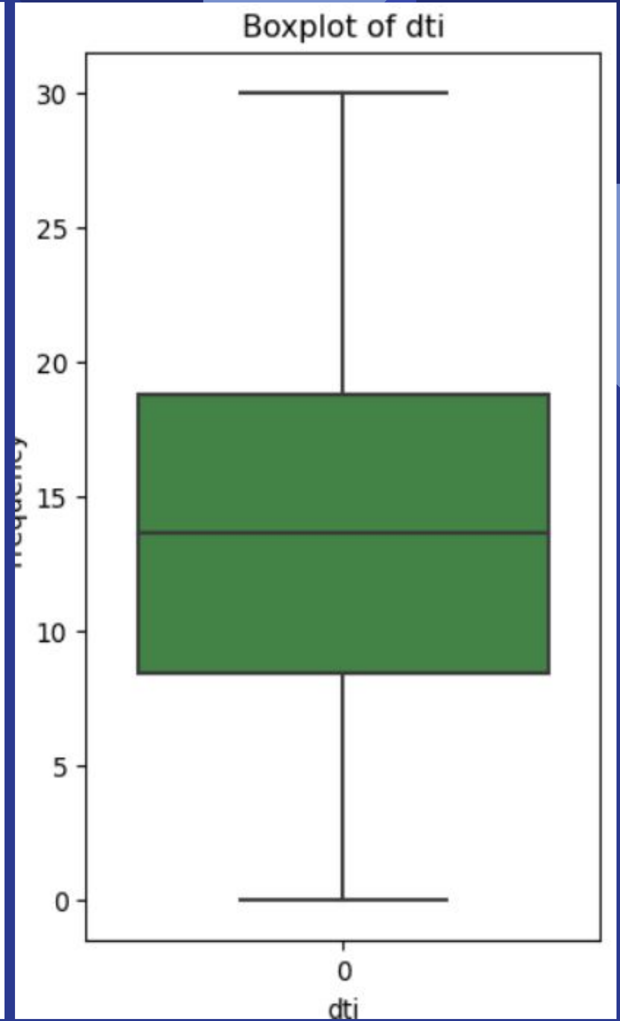

Count Plot of purpose
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
debt_consolidation
credit_card
other
home_improvement
major_purchase
small_business
car
wedding
medical
moving
vacation
house
educational
renewable_energy


Count Plot of home_ownership
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
RENT
MORTGAGE
OWN
OTHER
home_ownership


Count Plot of loan_status
Frequency
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
loan_status

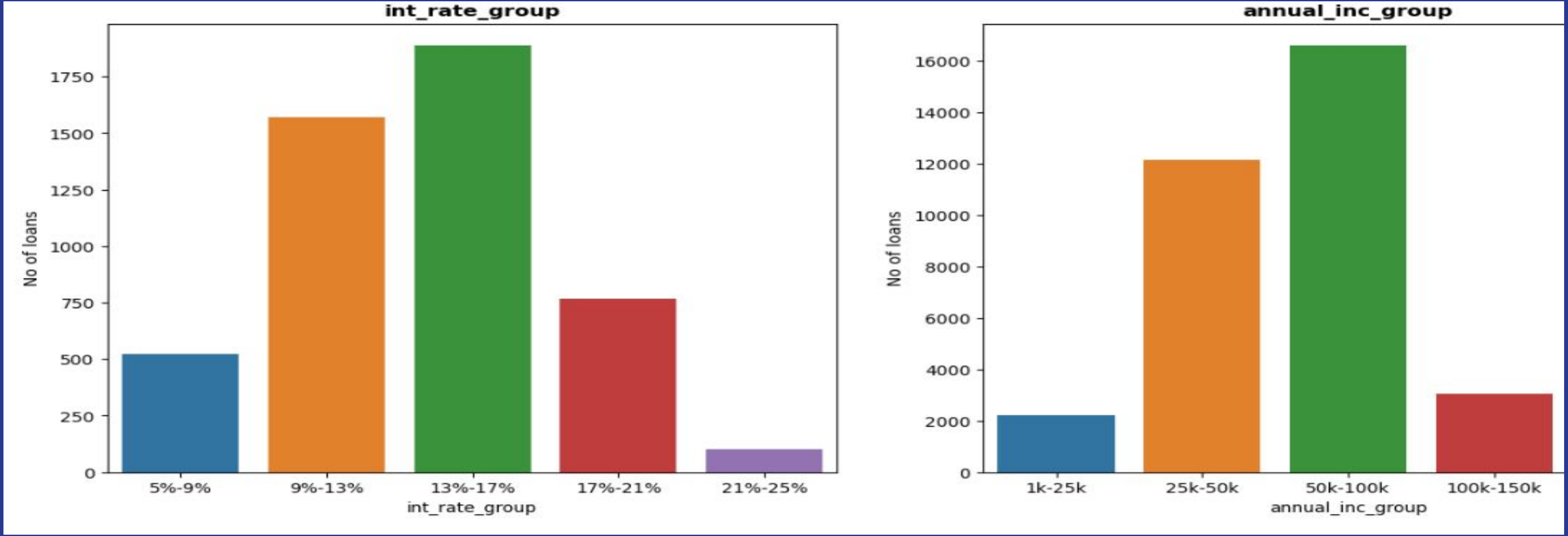

Boxplot of loan_amnt
30000
25000
20000
15000
10000
5000
0
00
0
loan_amnt


Boxplot of installment
1000
800
600
400
200
0
0
installment


Boxplot of dti
30
25
20
15
10
5
0
0
dti

Segmented Univariate Analysis

We have considered the Interest rates by grouping them to understand where we see the highest Charged off loans. It was observed that the loan applicants who are having 13% -17% interest rates are the ones who has highest charged off loan amounts. Same for annual_inc_group which has 50k-100k highest charged off.

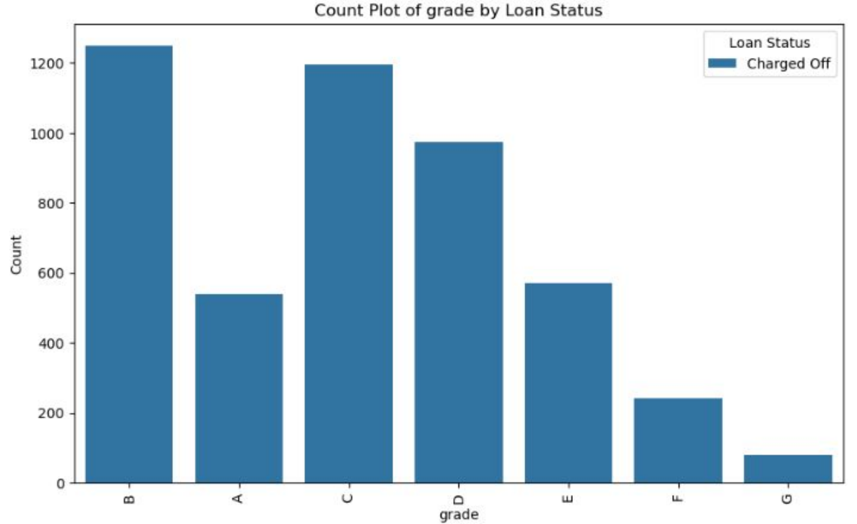


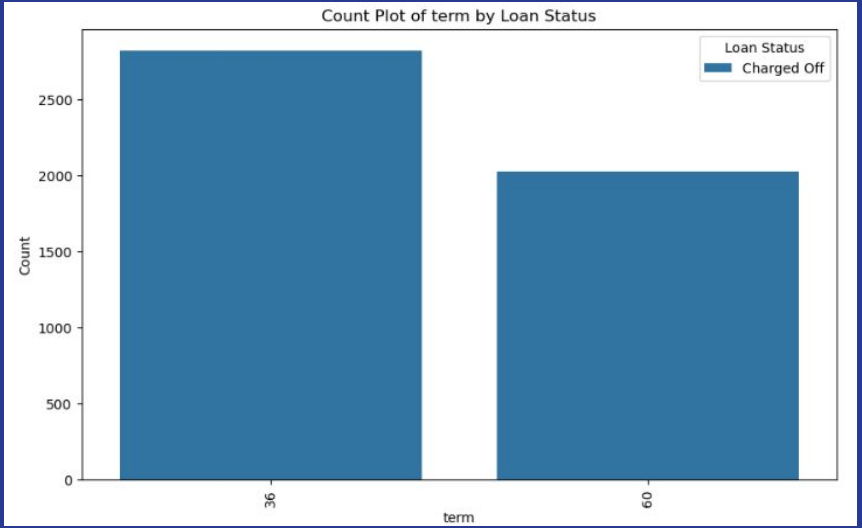
Observations from Univariate Analysis:

1. Grade A and B are given more loans compared to other grades
2. Grade A4, B3, A5, B5, B4 are given more loans compared to other grades
3. 36 months loans are issued more compared to 60 months loans.
4. Employees with 10 years and above are given loan compared with lesser experience.
5. Maximum loans were taken in the year 2011. The trend is increasing with the increase in the year,there is increasing trend in number of loans with increase in the months. Maximum loans were given in the month of Oct, Nov, Dec.
6. 14 % of the total loans are charged off
7. States CA, NY, FL and TX are the states for which maximum loans have been issued
8. Maximum loans are given for debt consolidation, paying off Credit card and 'other' reasons
9. Education and renewable energy is the least category where loans have been given
10. People who are in Rented house or Mortgage have availed maximum of the loans
11. Funded amount is ranging from 5000 to 15000 USD
12. Installment amount is ranging from 200 to 400 USD
13. Interest rate range 13 to 17% is the range where maximum default loans have been issued.
14. 21 - 25% is the range where minimum loans have been issued
15. The annual income group where 50k-100k have maximum loans issued and 1k-25k have minimum loans issued.

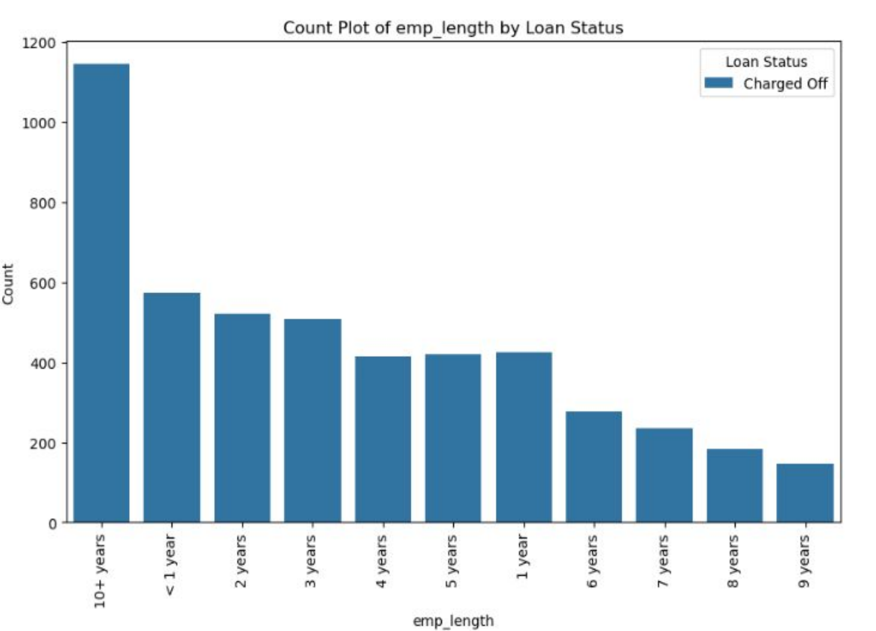
Bivariate analysis

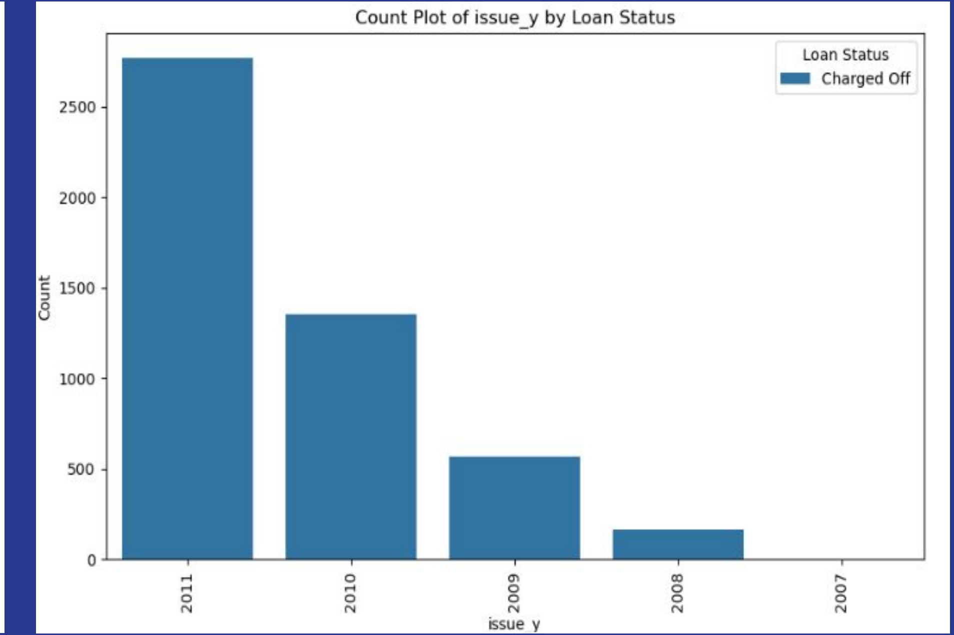
Variables: We examine two variables - grade and term- in relation to loan status (charged off)
Purpose: Identify patterns and insights that can help predict loan defaulters and minimize risk.
Analysis: We had performed bivariate analysis to understand how grade and term(36/60 months) influence loan status.
Outcome: We observe a significantly higher likelihood of default among individuals with grade ‘B’ and same for term ‘36 months’ loan.



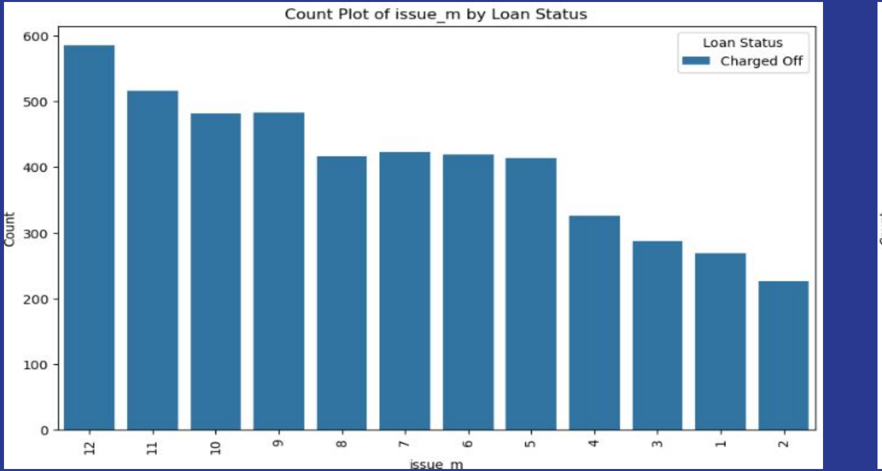


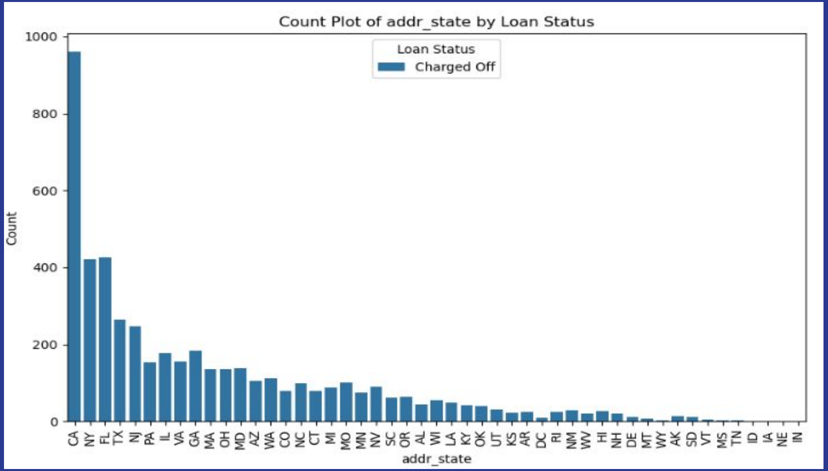
Variables: We examine two variables - emp_length and issue_yr - in relation to loan status (charged off)
Purpose: Identify patterns and insights that can help predict loan defaulters and minimize risk.
Analysis: We had performed bivariate analysis to understand how emp_length and issue_yr influence loan status.
Outcome: We observe a significantly higher likelihood of default among individuals with 10+ years and same for issue_yr 2011 year.





Variables: We examine two variables - issue_month and addr_state- in relation to loan status (charged off)
Purpose: Identify patterns and insights that can help predict loan defaulters and minimize risk.
Analysis: We had performed bivariate analysis to understand how issue_month and addr_state influence loan status.
Outcome: We observe a significantly higher likelihood of default among individuals who have taken loan in DEC,NOV..and same for addr_state who are from ‘CA’ state.



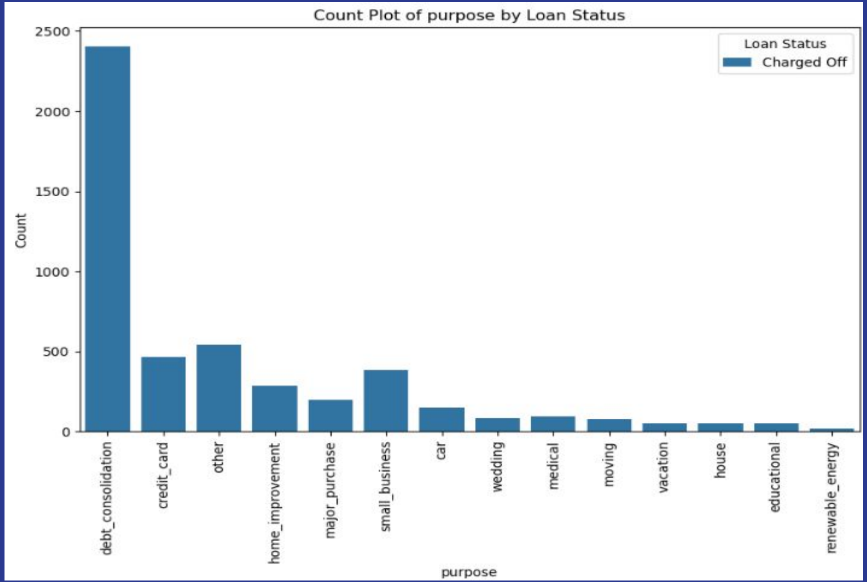


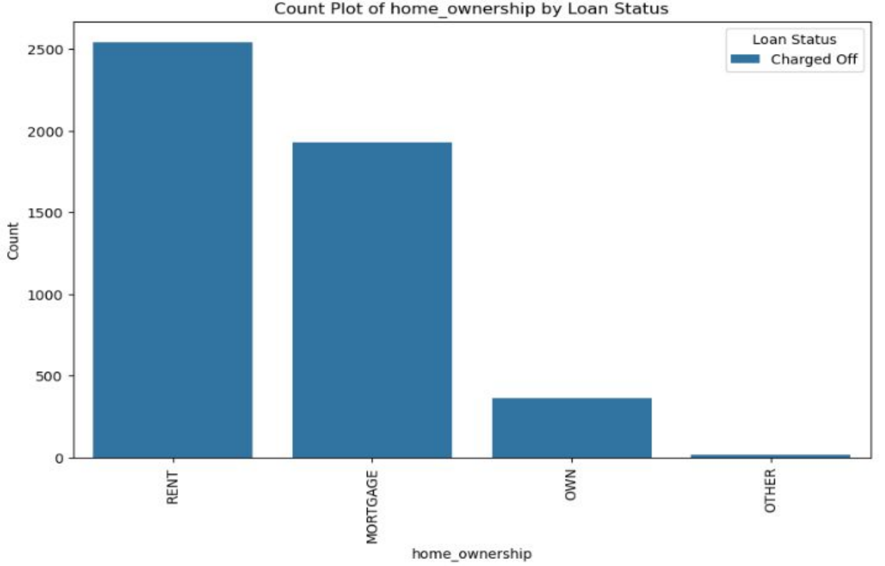
Variables: We examine two variables - purpose and home_ownership - in relation to loan status (charged off)

Purpose: Identify patterns and insights that can help predict loan defaulters and minimize risk.

Analysis: We had performed bivariate analysis to understand how purpose and home_ownership influence loan status.

Outcome: We observe a significantly higher likelihood of default among individuals whose purpose is debt_consolidation and same for home_ownership who are in 'RENT'/'MORTGAGE'.



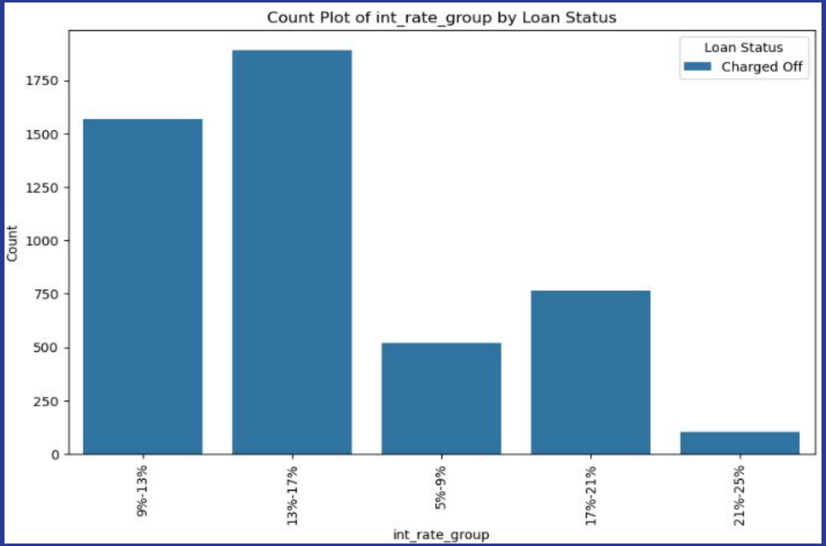


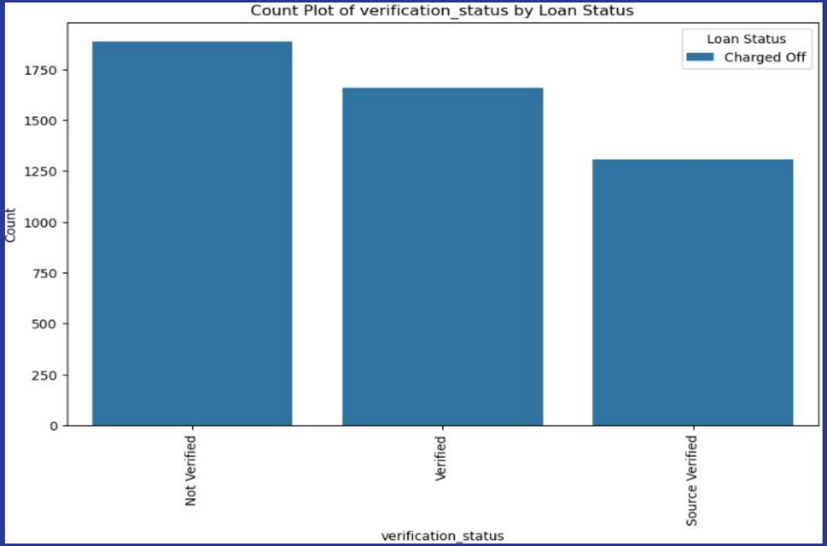
Variables: We examine two variables - int_rate and verification_status - in relation to loan status (charged off)

Purpose: Identify patterns and insights that can help predict loan defaulters and minimize risk.

Analysis: We had performed bivariate analysis to understand how int_rate and verification_status influence loan status.

Outcome: We observe a significantly higher likelihood of default among individuals who has taken interest around 13%-17% and same for verification_status whose status has Not verified.





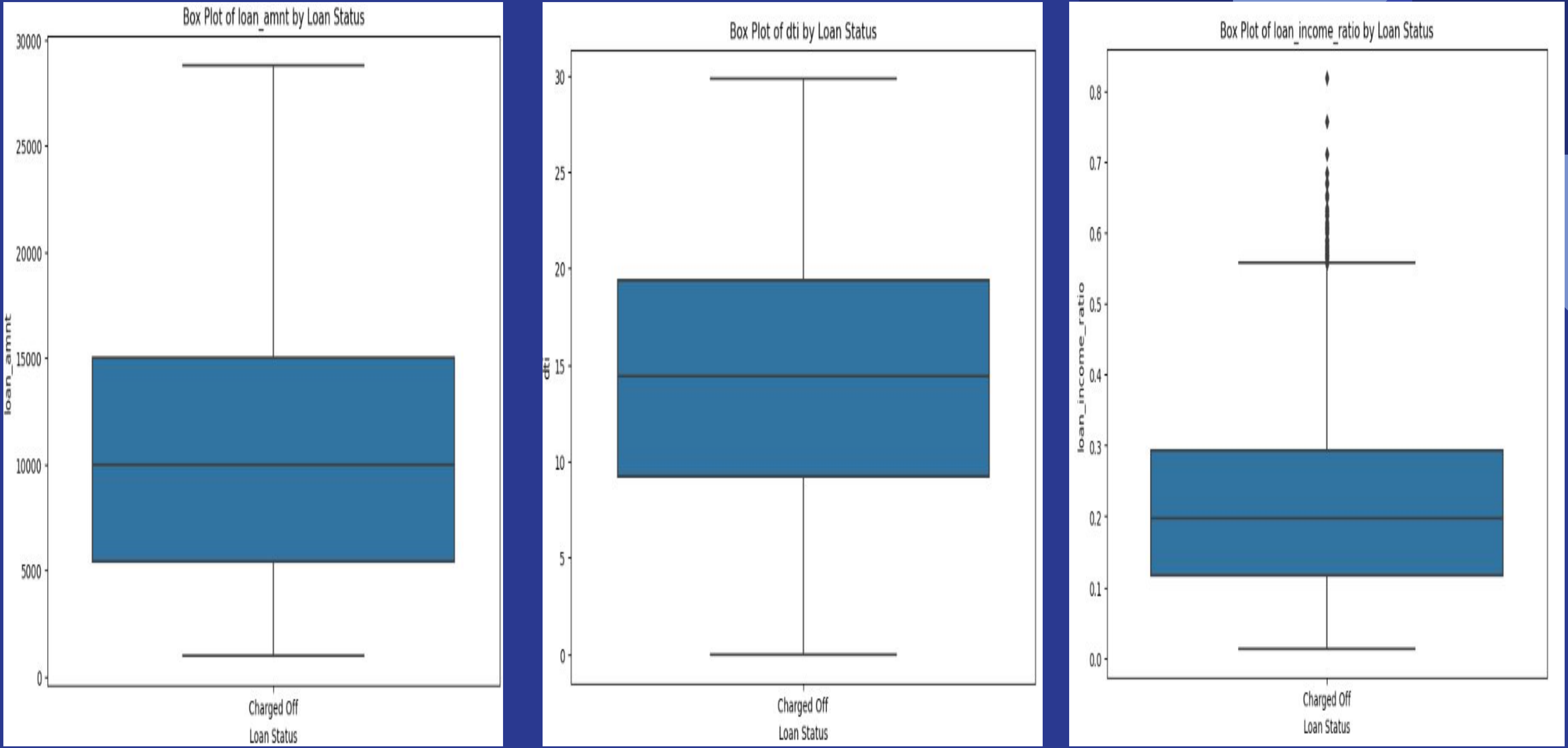
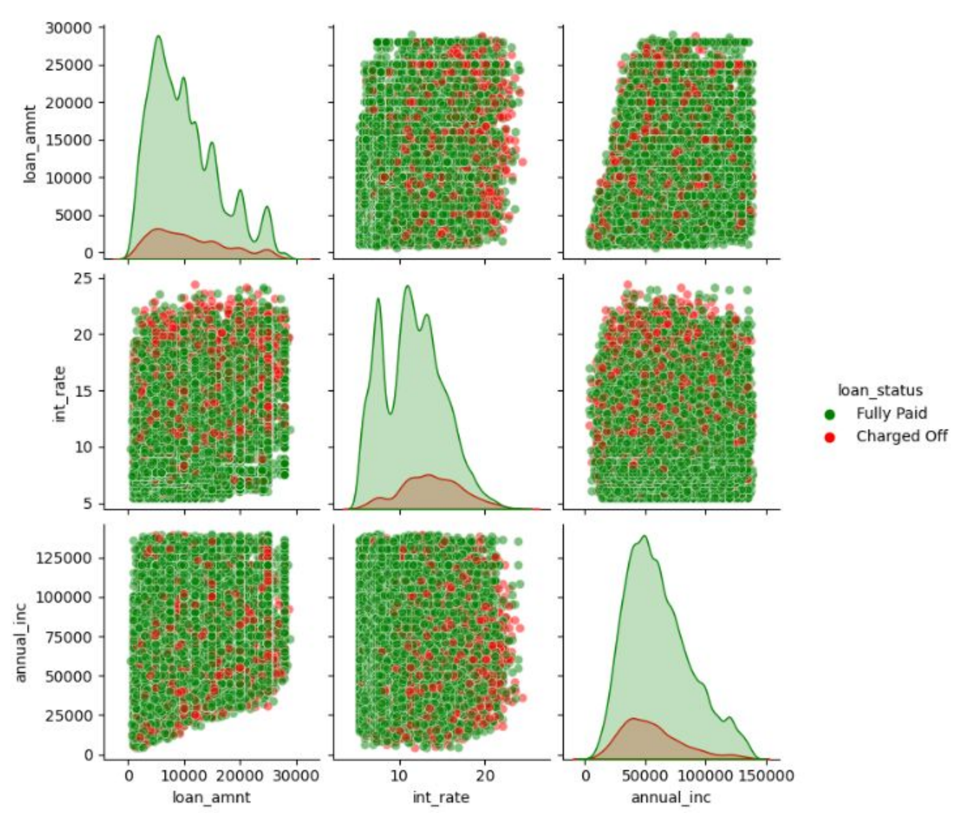


Fig: Bivariate analysis for quantitative variables loan_amnt,dti,loan_income_ratio

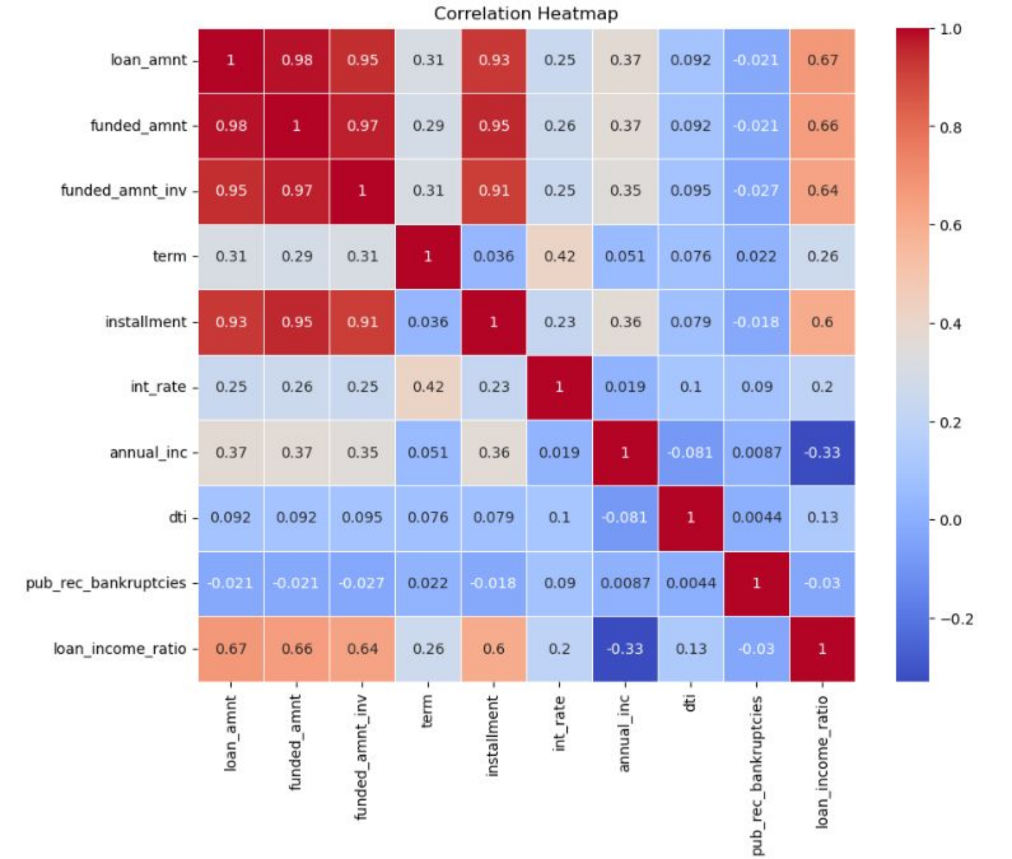
Multivariate Analysis



Observations:

- Less interest rate and less amount of loan amount are not likely to default.
- More interest for the loan are more likely to default.

Correlation Analysis:



	loan_amnt	funded_amnt	funded_amnt_inv	term	installment	int_rate	annual_inc	dti	pub_rec_bankruptcies	loan_income_ratio
loan_amnt	1	0.98	0.95	0.31	0.93	0.25	0.37	0.092	-0.021	0.67
funded_amnt	0.98	1	0.97	0.29	0.95	0.26	0.37	0.092	-0.021	0.66
funded_amnt_inv	0.95	0.97	1	0.31	0.91	0.25	0.35	0.095	-0.027	0.64
term	0.31	0.29	0.31	1	0.036	0.42	0.051	0.076	0.022	0.26
installment	0.93	0.95	0.91	0.036	1	0.23	0.36	0.079	-0.018	0.6
int_rate	0.25	0.26	0.25	0.42	0.23	1	0.019	0.1	0.09	0.2
annual_inc	0.37	0.37	0.35	0.051	0.36	0.019	1	-0.081	0.0087	-0.33
dti	0.092	0.092	0.095	0.076	0.079	0.1	-0.081	1	0.0044	0.13
pub_rec_bankruptcies	-0.021	-0.021	-0.027	0.022	-0.018	0.09	0.0087	0.0044	1	-0.03
loan_income_ratio	0.67	0.66	0.64	0.26	0.6	0.2	-0.33	0.13	-0.03	1

- There is a strong positive correlation between the loan amount, funded amount, and installment amount. This means that loans with higher loan amounts are also likely to have higher funded amounts and higher installment amounts.
- There is a negative correlation between loan_amnt and annual_inc.

Final Analysis

1. Grades B, C, and D have the highest occurrence of Charged Off loans, based on frequency counts.
2. Based on the counts, Grade B3,B4,B5 top sub grades in Charged Off.
3. Default rates are higher for 36-month loans compared to 60-month loans.
4. Borrowers with over 10 years of experience have the highest default rates, making them the primary defaulters.
5. The year 2011 saw the highest number of loans issued, with a significant portion of them being Charged Off during that year.
6. Plot of the loan issue month shows maximum loans were given in the month of Oct, Nov, Dec.Also high loans are being Charged Off for the loans issued in Sep,Oct,Nov,Dec.
7. Loans intended for debt consolidation, credit card payments, and home improvement, as well as small business ventures, are more likely to default compared to those for education or renewable energy.
8. Borrowers from California(CA) are more likely for defaulting on loans.
9. Individuals who rent or have a mortgage are more prone to loan default compared to homeowners.
10. The amount of Not verified loans which are Charged Off (default) is more compared to Verified,Source Verified.
11. Interest Rate with 13-17% have more loans default.
12. People who have annual income 50K-100k are more likely to loan default.