Lending Club Case Study

Group Members:

Sneha Kumari

Govind Vishnoi

Content

- About Lending Club
- Problem Statement
- Fix missing Values
- Standardise Values
- Removing Outliers
- Univariate Analysis on Continuous Variables
- Univariate Analysis on Categorical Variables
- Segmented Univariate on Categorical Variables
- Segmented Univariate on Continuous Variables
- Bivariate on Continuous Variables
- Bivariate on Categorical Variables
- Derived Metric
- Recommendations based on Univariate Analysis
- Recommendations based on Bivariate Analysis

About Lending Club

- **LendingClub** is a peer-to-peer lending company headquartered in San Francisco, California.
- It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market.
- The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015.

Problem Statement

- The data given contains the information about past loan applicants and whether they ‘defaulted’ or not.

- The aim is to identify patterns which indicate if a person is likely to default by understanding how consumer attributes and loan attributes influence the tendency of default.

- 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:
  1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
  2. **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'.
  3. **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

Data Cleaning

Fix Missing Values

- Loading the loan dataset containing 39717 rows and 111 columns

	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
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	

5 rows × 111 columns

- Dropping columns containing all null values, duplicate values, huge percentage of null values and which are irrelevant for analysis. Shape of dataset after dropping: 39719 rows and 14 columns

Standardize Values

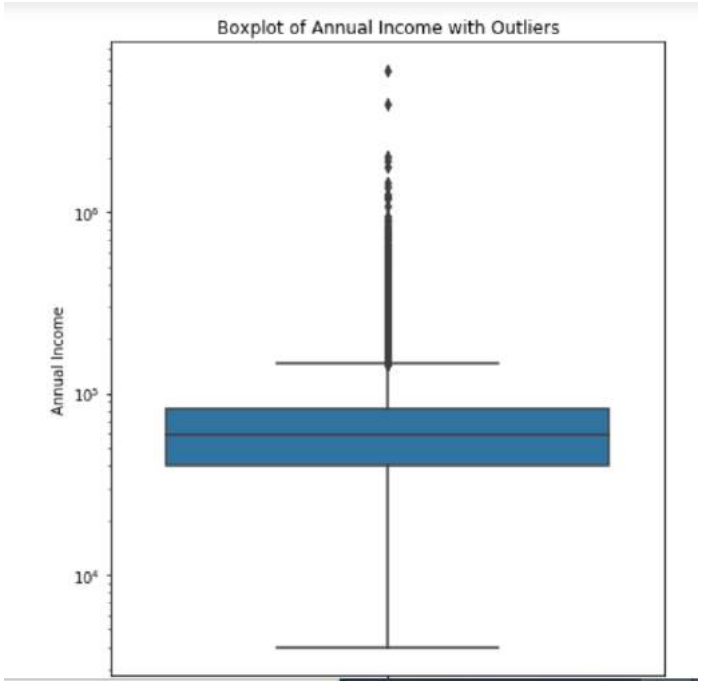
- Removing % from int_rate column

	id	member_id	loan_amnt	funded_amnt_inv	term	int_rate	grade	sub_grade	emp_length	home_ownership	annual_inc	verification_status	issue
0	1077501	1296599	5000	4975.0	36 months	10.65	B	B2	10+ years	RENT	24000.0	Verified	01-2
1	1077430	1314167	2500	2500.0	60 months	15.27	C	C4	< 1 year	RENT	30000.0	Source Verified	01-2
2	1077175	1313524	2400	2400.0	36 months	15.96	C	C5	10+ years	RENT	12252.0	Not Verified	01-2
3	1076863	1277178	10000	10000.0	36 months	13.49	C	C1	10+ years	RENT	49200.0	Source Verified	01-2
4	1075358	1311748	3000	3000.0	60 months	12.69	B	B5	1 year	RENT	80000.0	Source Verified	01-2

- Changing datatype of int_rate column to float
- Dropping data points where the loan_status is current as it not useful for study. New shape of data frame: 38577 rows and 14 columns
- Converting issue_d column into datetime format

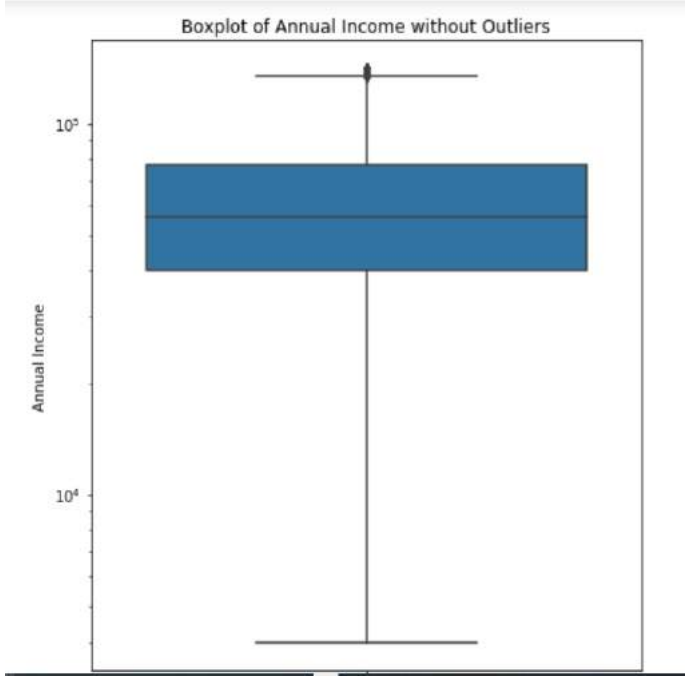
Removing Outliers

- Annual income contains a large outliers as shown in the 1st plot

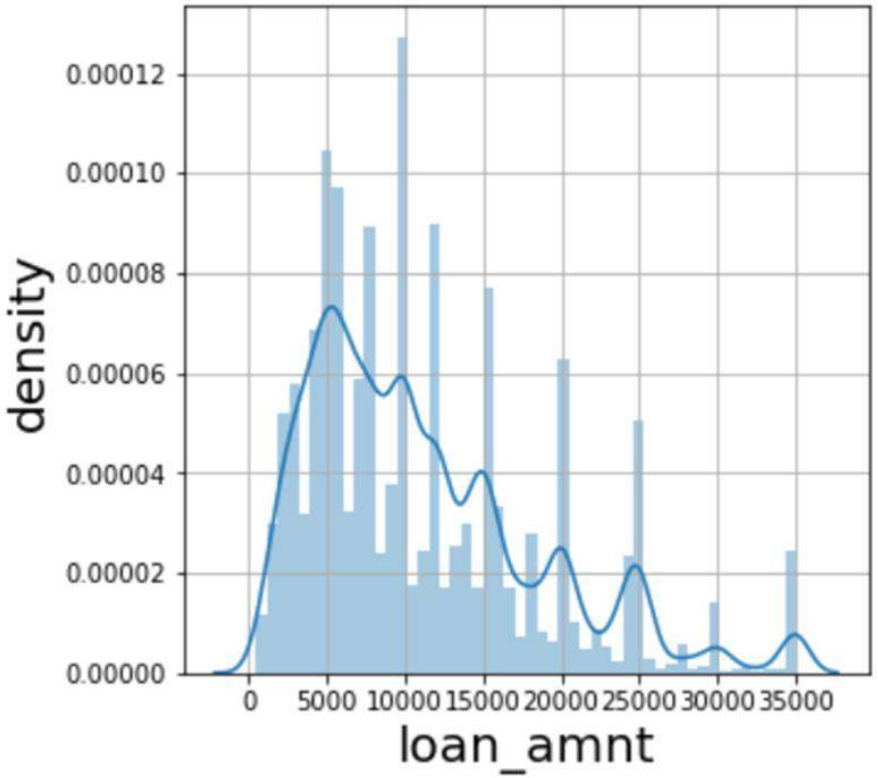


- Outlier are those whose values are falling after .95 percentile

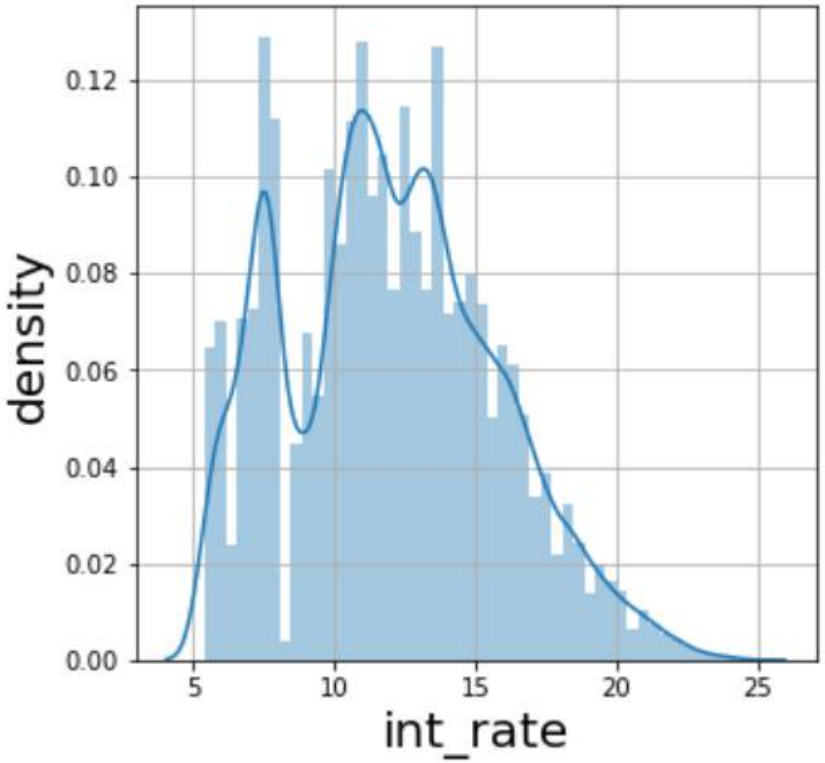
- Data shape after dropping outliers of annual income: 36642 rows and 14 columns



Univariate Analysis on Continuous Variables

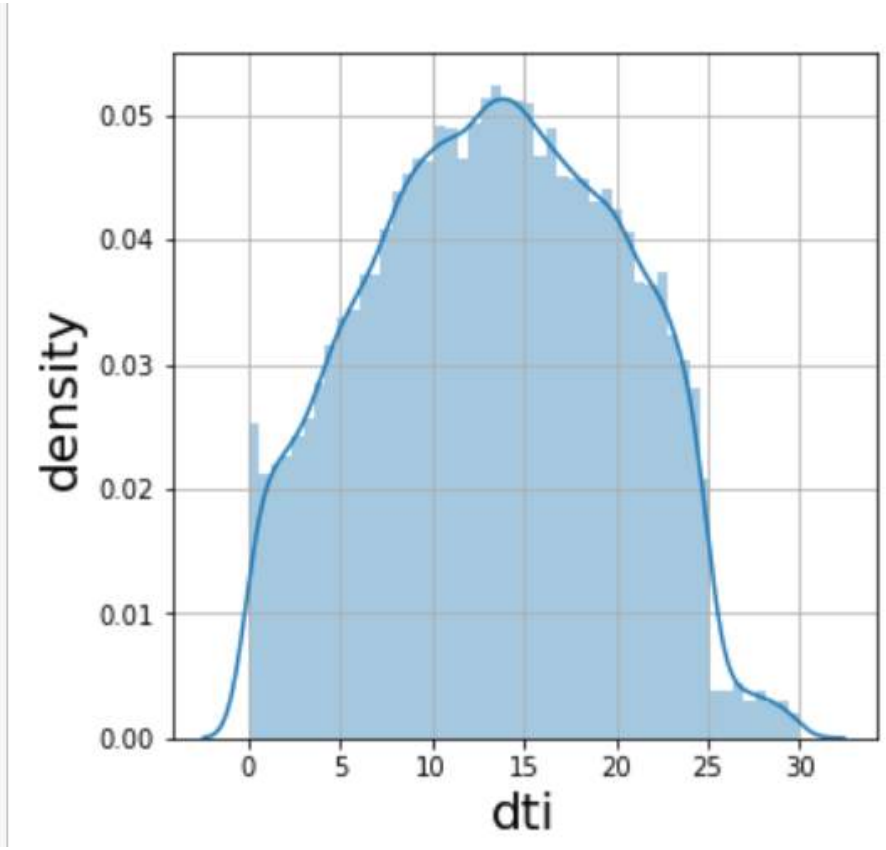


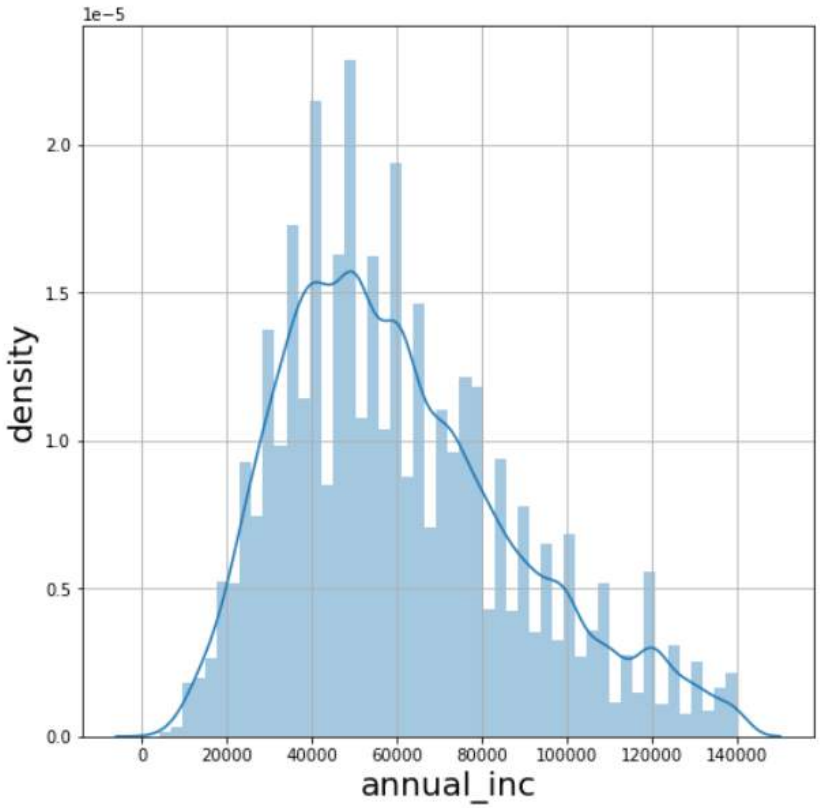
- Majority of the loan amount lies between 4000 to 12000



- Majority of the loan interest rate lies in the range 10% to 15%

Univariate Analysis on Continuous Variables

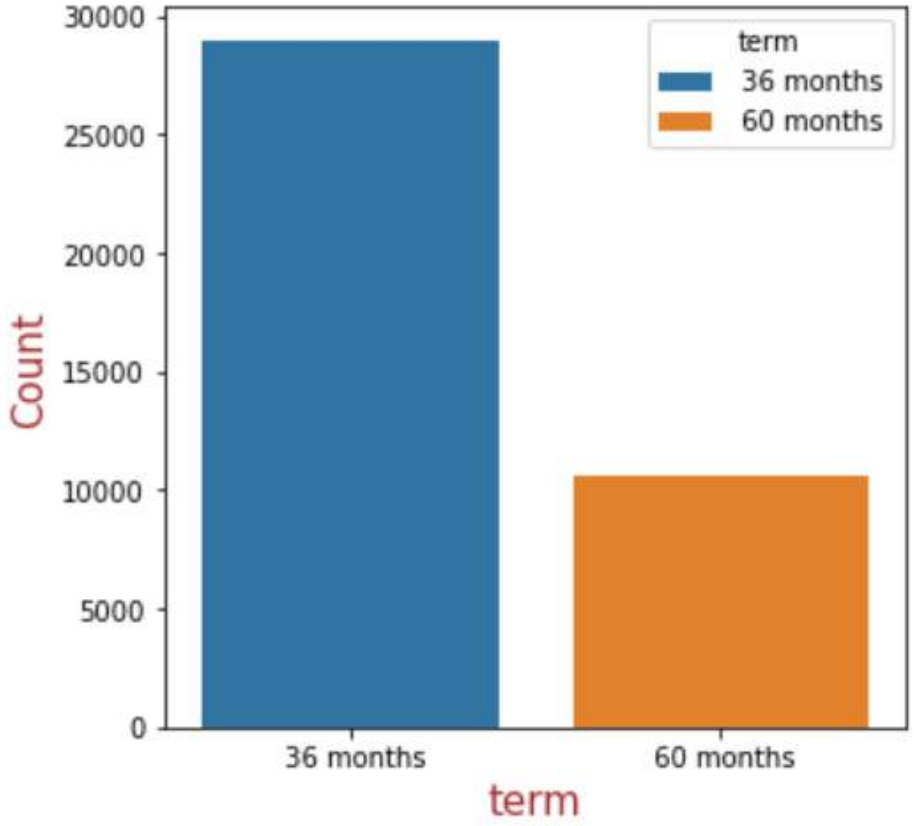


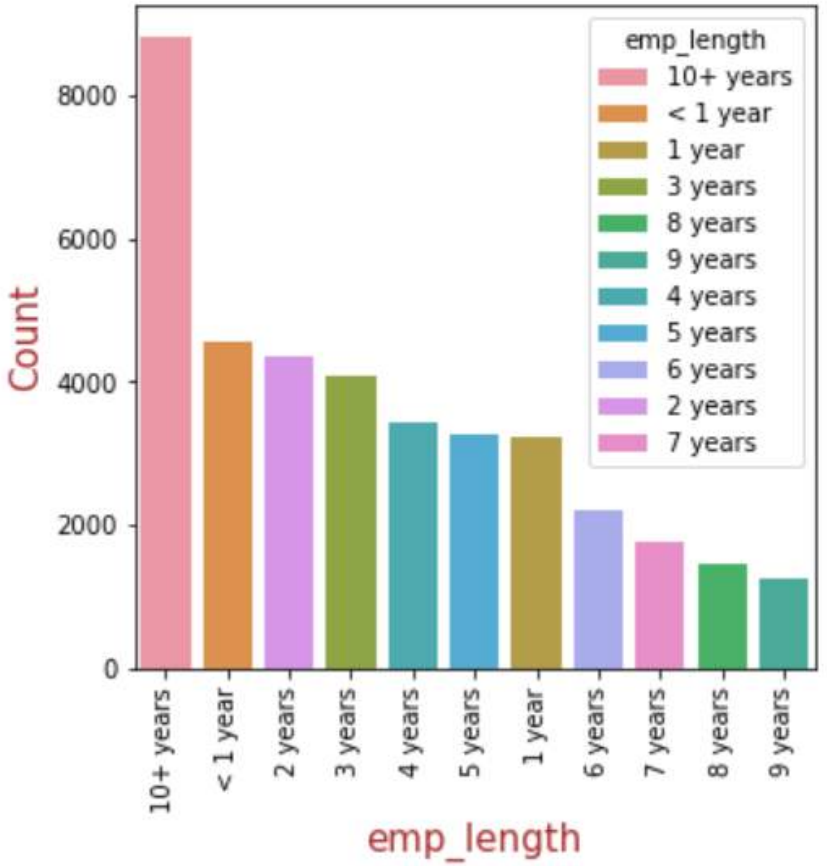


- Majority of the debt to income ratio lies between 4 to 24

- The annual income of majority of the borrowers are in the range 30000 to 80000

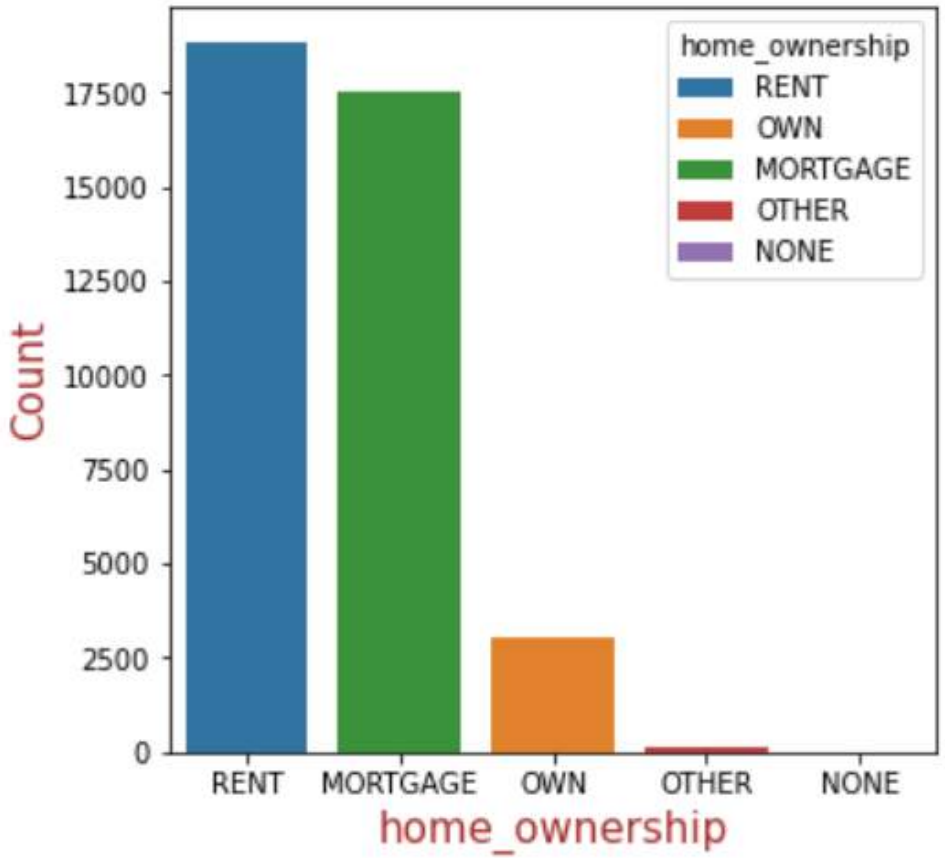
Univariate Analysis on Categorical Variables

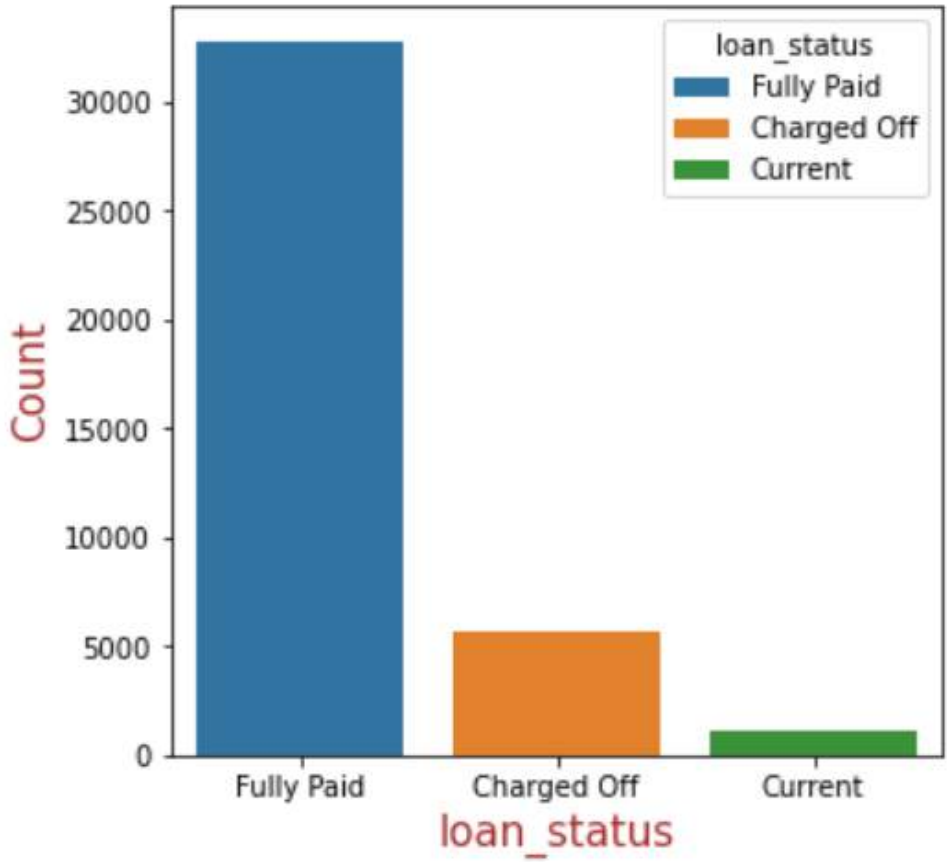




- Majority of the borrowers prefer to choose term of 36 months (~29000) whereas less borrowers opt for 60 months (~10000)
- Number of years of Employment for majority of the borrowers are 10+ years (~8900)

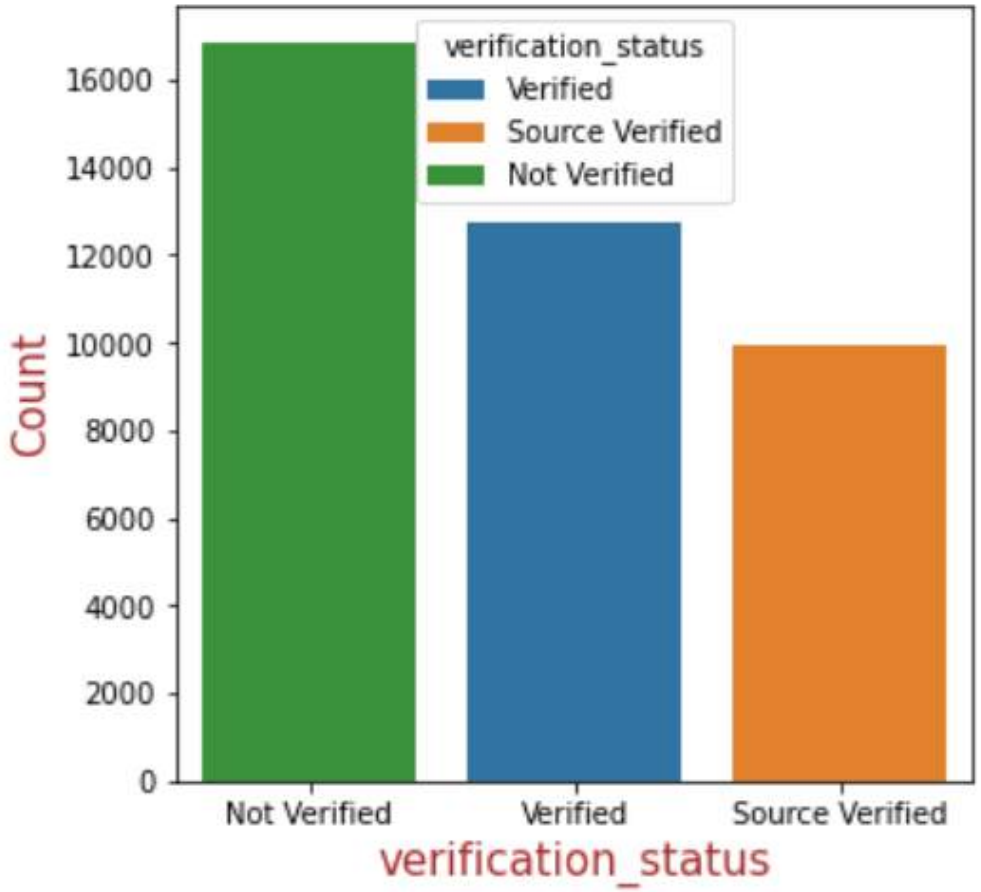
Univariate Analysis on Categorical Variables



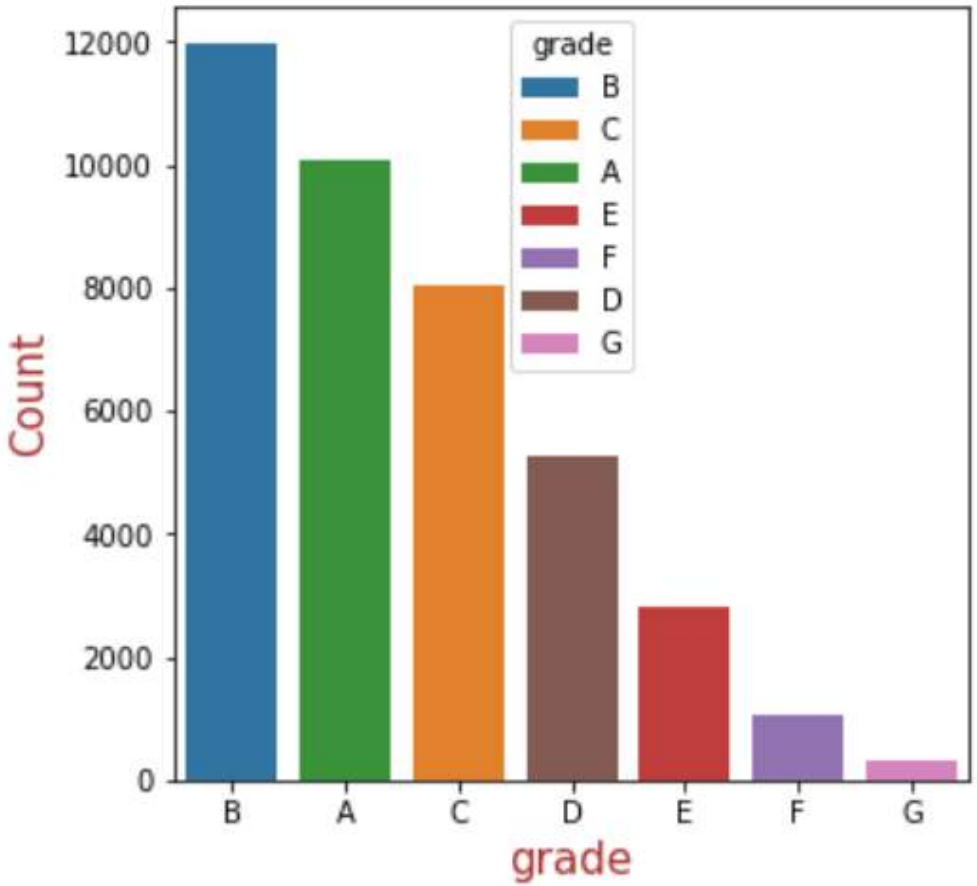


- Majority of the borrowers live in a Rented home 18900 whereas 17600 borrowers have a Mortgage on their property
- Majority of the loan status is Fully paid (33000) whereas very less borrowers have a status as Charged off (5600)

Univariate Analysis on Categorical Variables

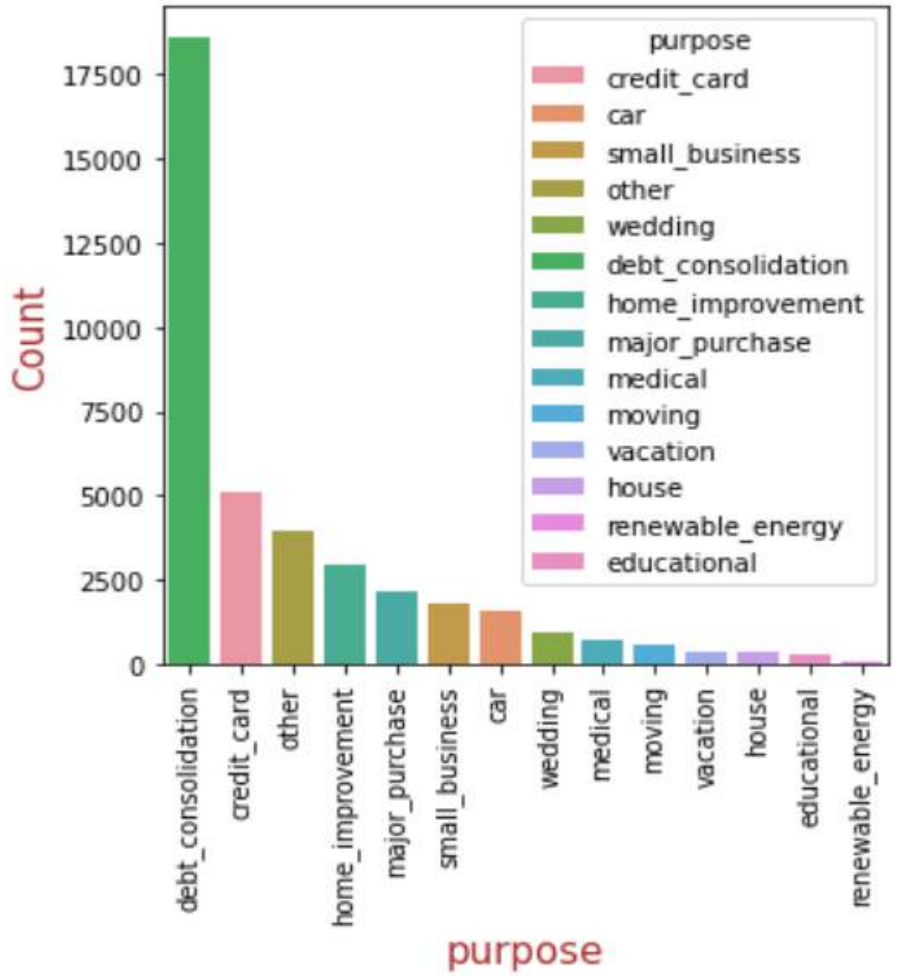


- Majority of the borrower's income is Not Verified (~16930)

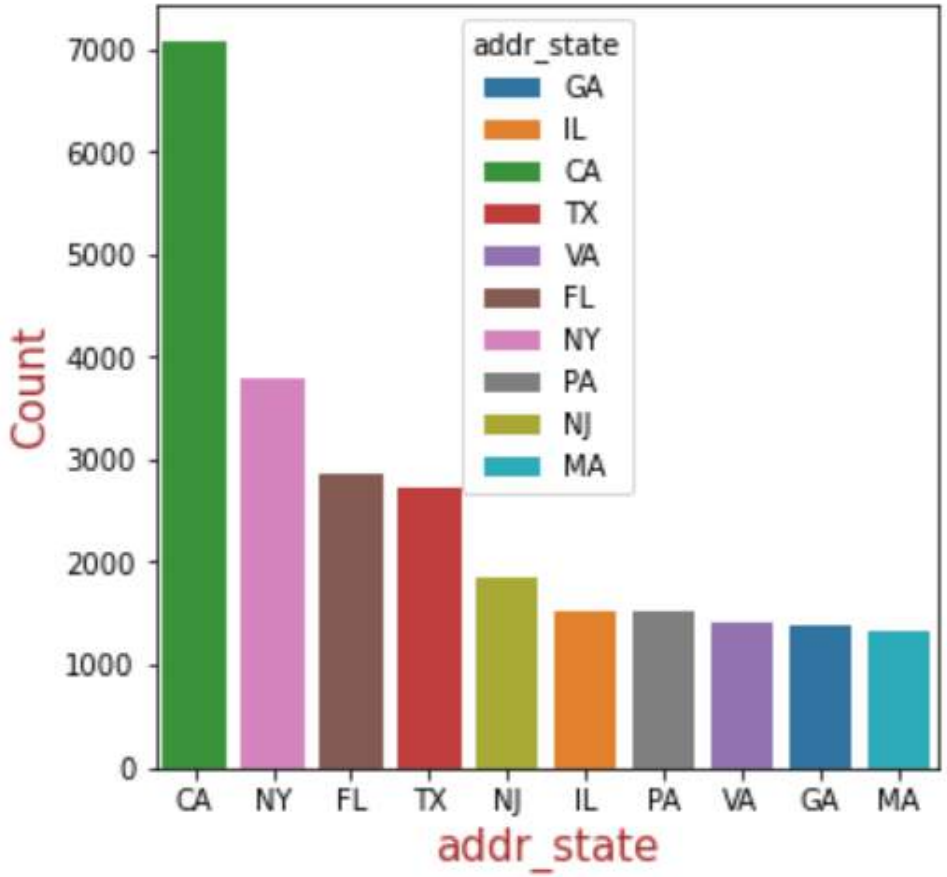


- Most of the borrowers are categorised under Grade B (12020) whereas least are categorised as Grade G (316)

Univariate Analysis on Categorical Variables

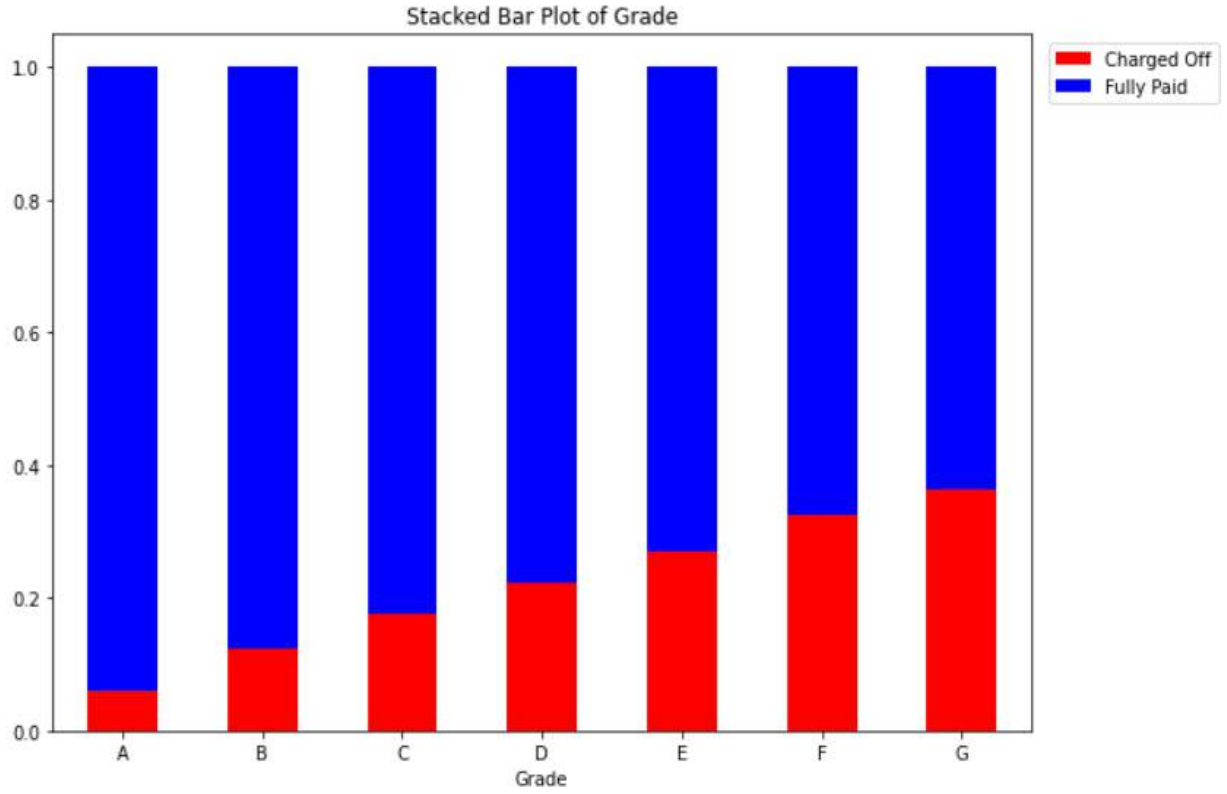


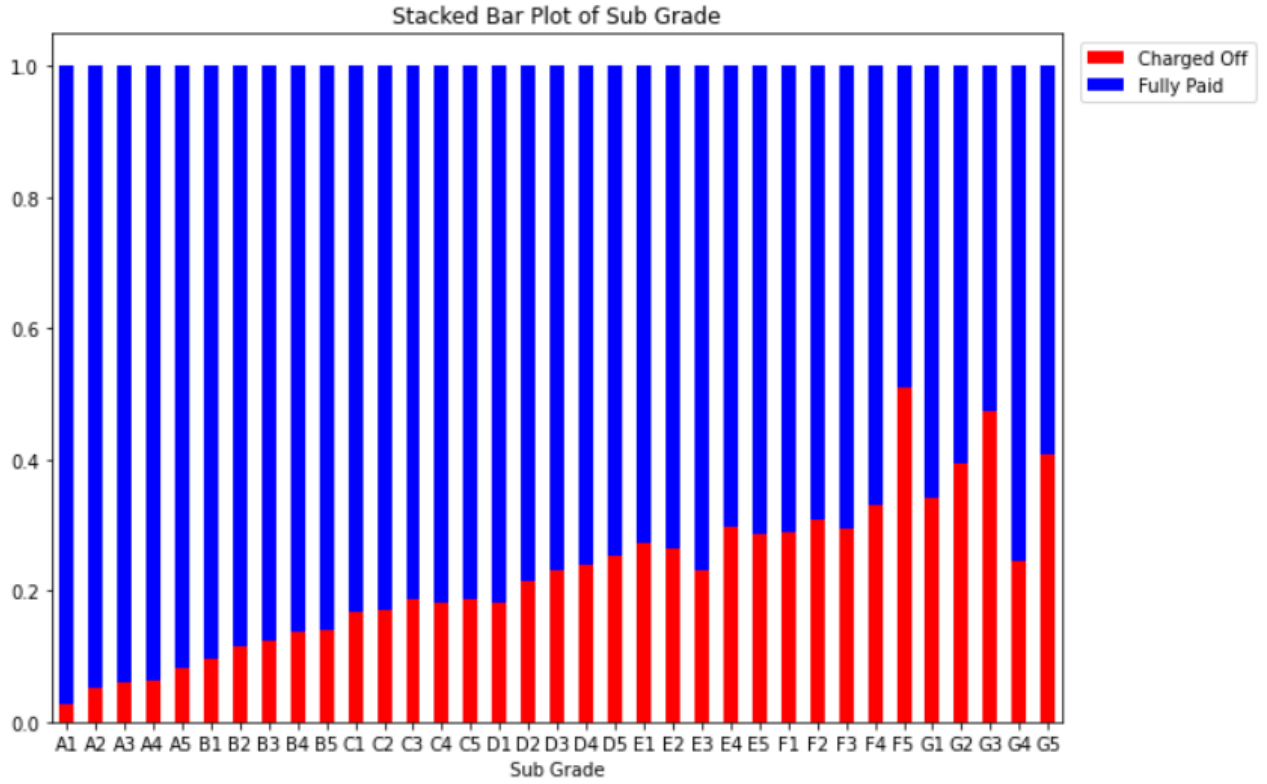
- The purpose of applying loan for majority of the borrowers is dept consolidation (18641) followed by credit card (5130)



- Majority of the borrowers applying for loan are from California state (7099)

Segmented Univariate Analysis on Categorical Variables

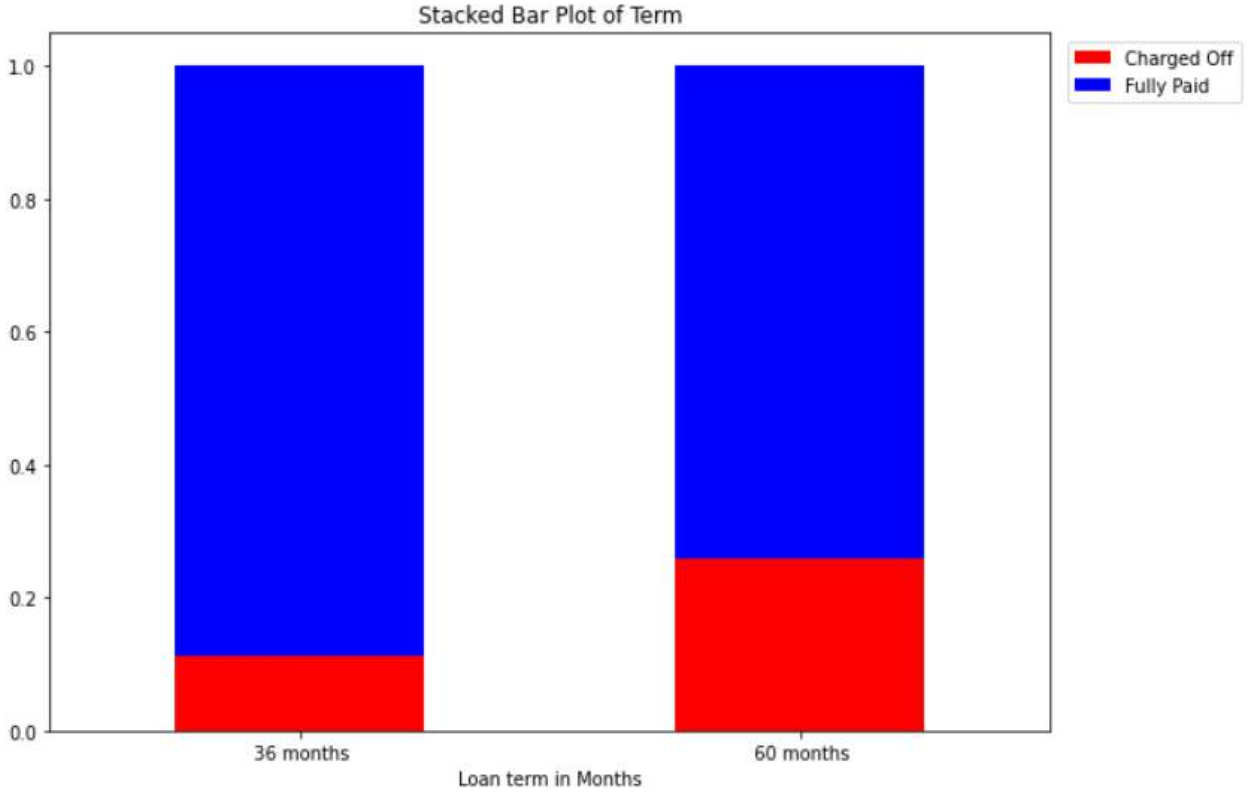


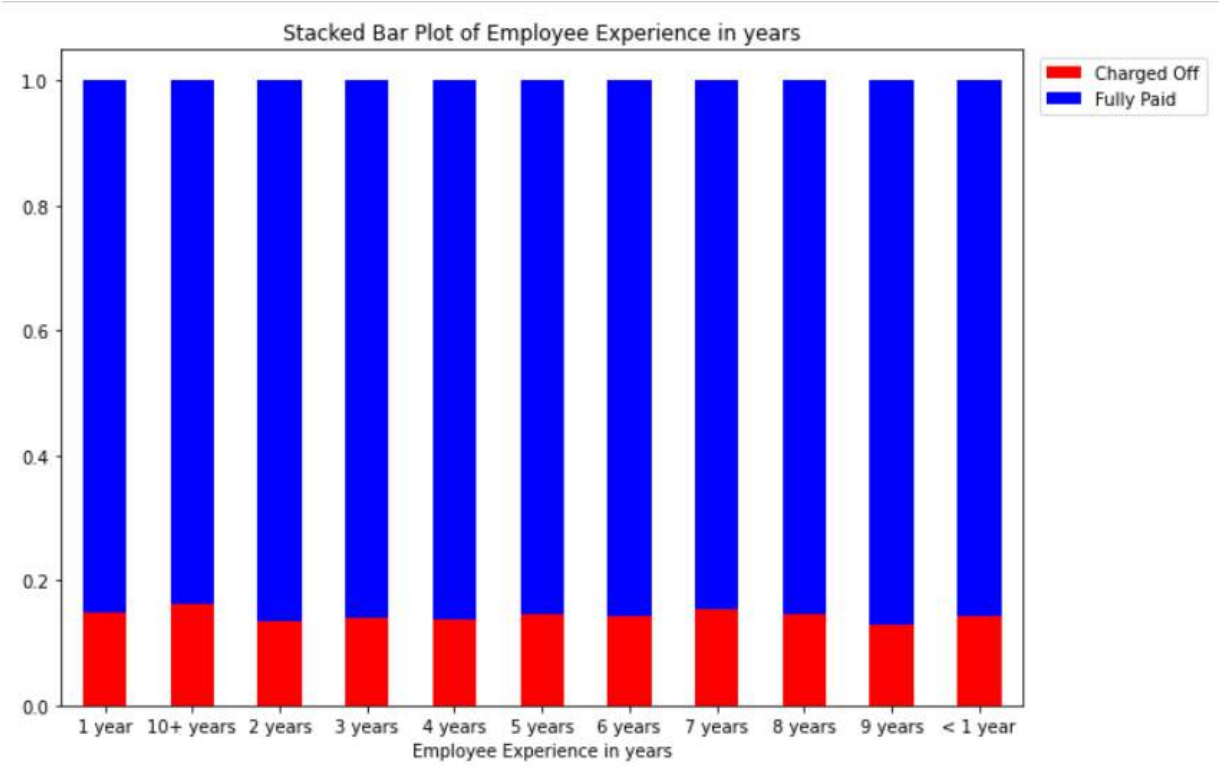


- We can see from the above graph that the Grade A has less prone to Defaulters.
- Going forward from A to G the rate of defaulters are increasing.

- As we can see G3, G5, F5 is more prone to defaulters

Segmented Univariate Analysis on Categorical Variables

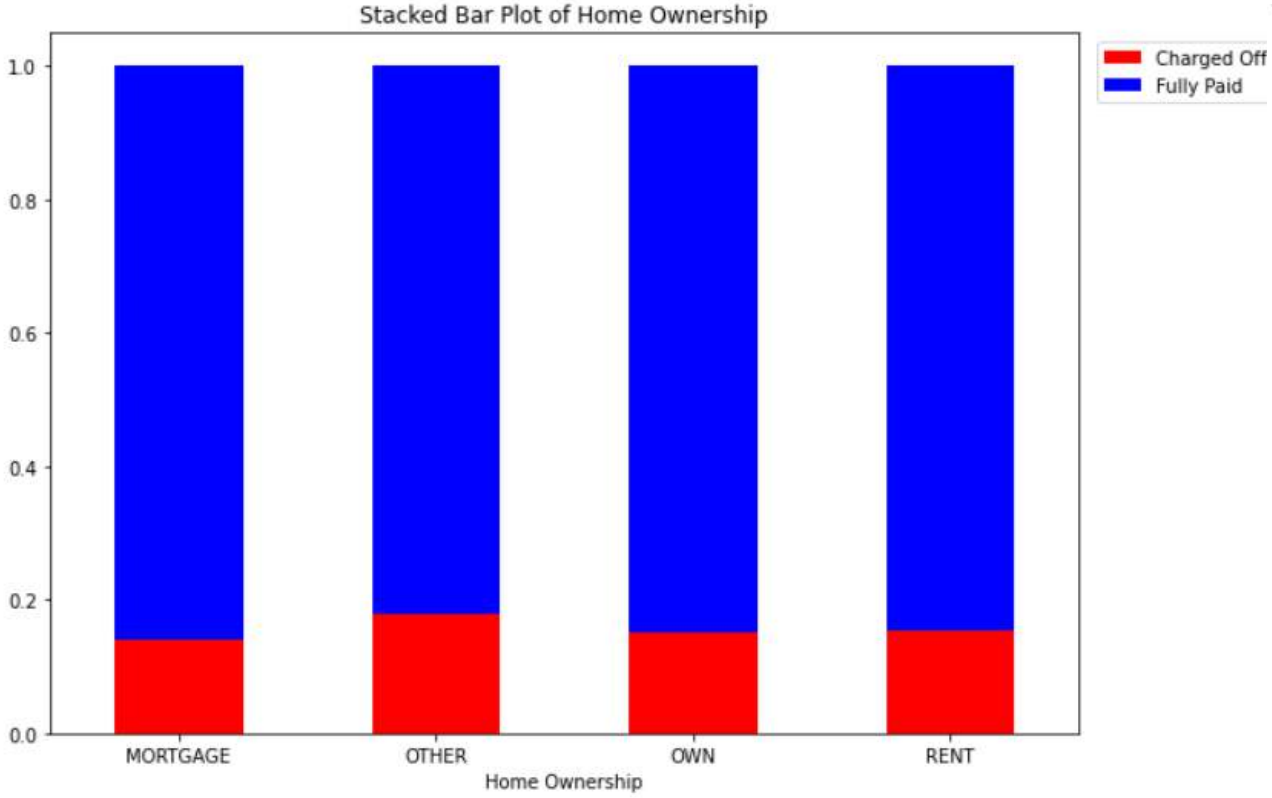




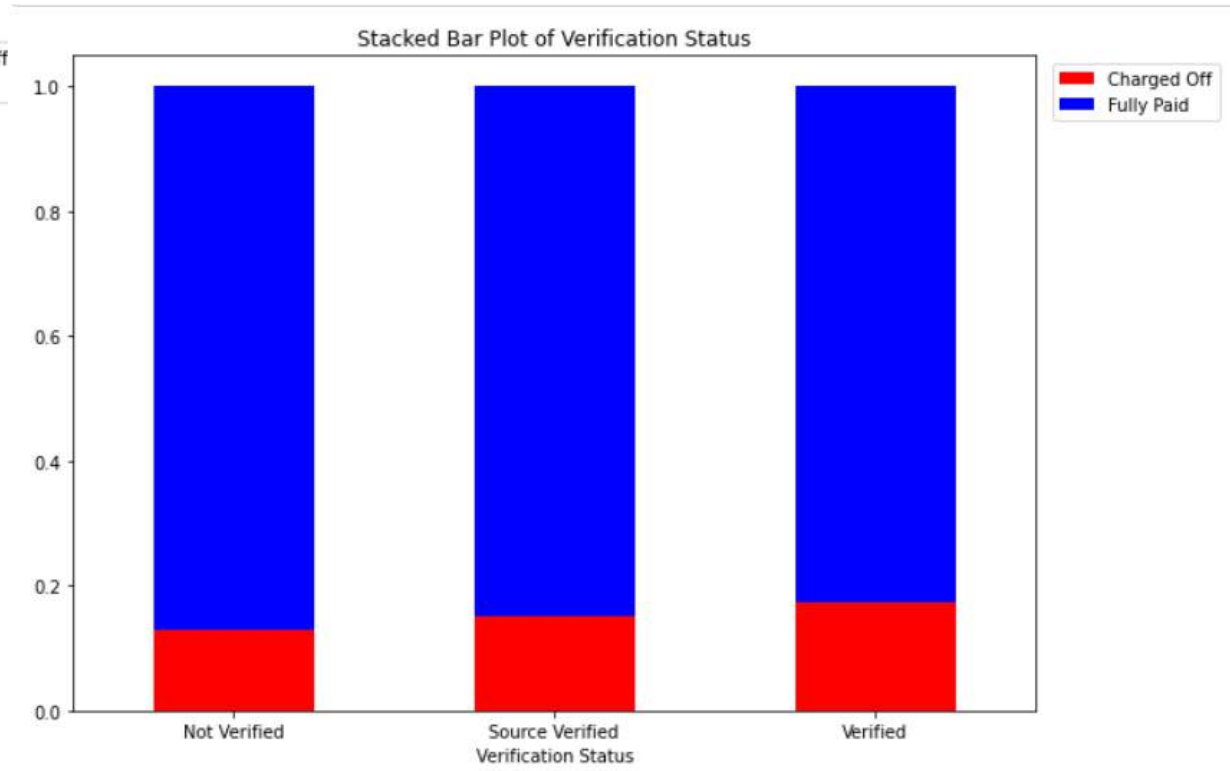
- We can see from the above plot that the 60 months loan term is more prone to charge off

- We can see from the above plot that there is not significant affect of no. of years of experience on loan status

Segmented Univariate Analysis on Categorical Variables

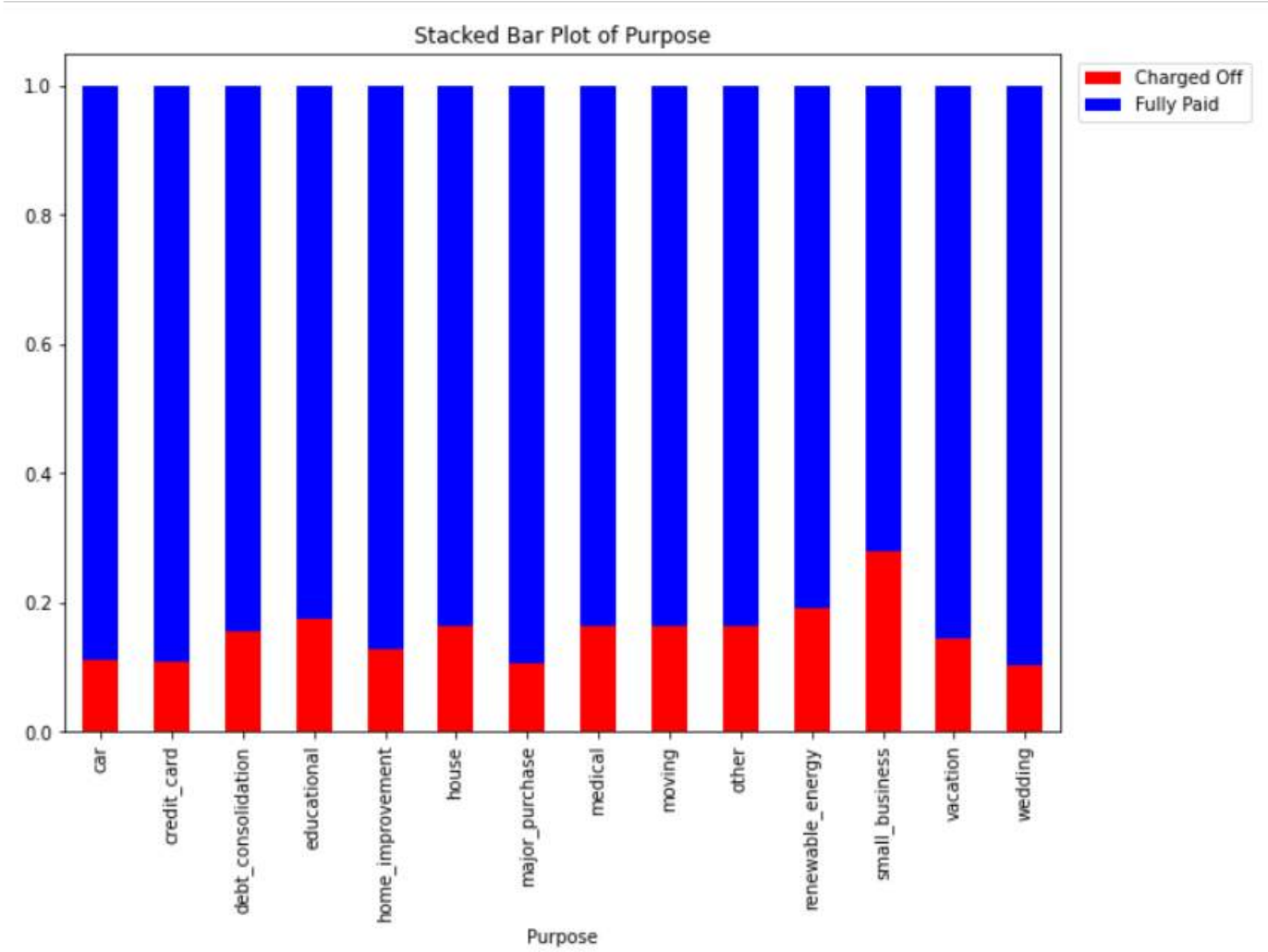


- We can see from the above plot that 'OTHER' category is more prone to charge off but we cannot comment on that as the 'OTHER' category is not defined properly.



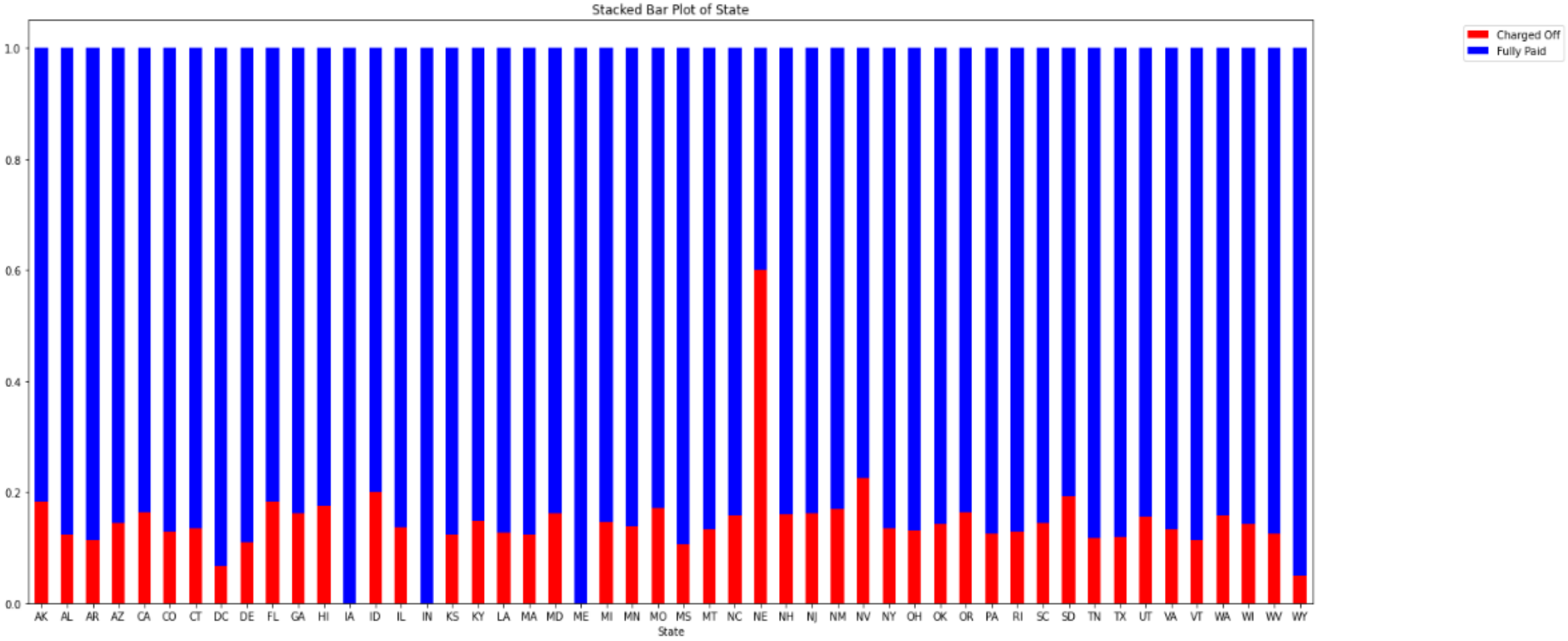
- We can see from above plot that the verified loans are more prone to defaulters.

Segmented Univariate Analysis on Categorical Variables



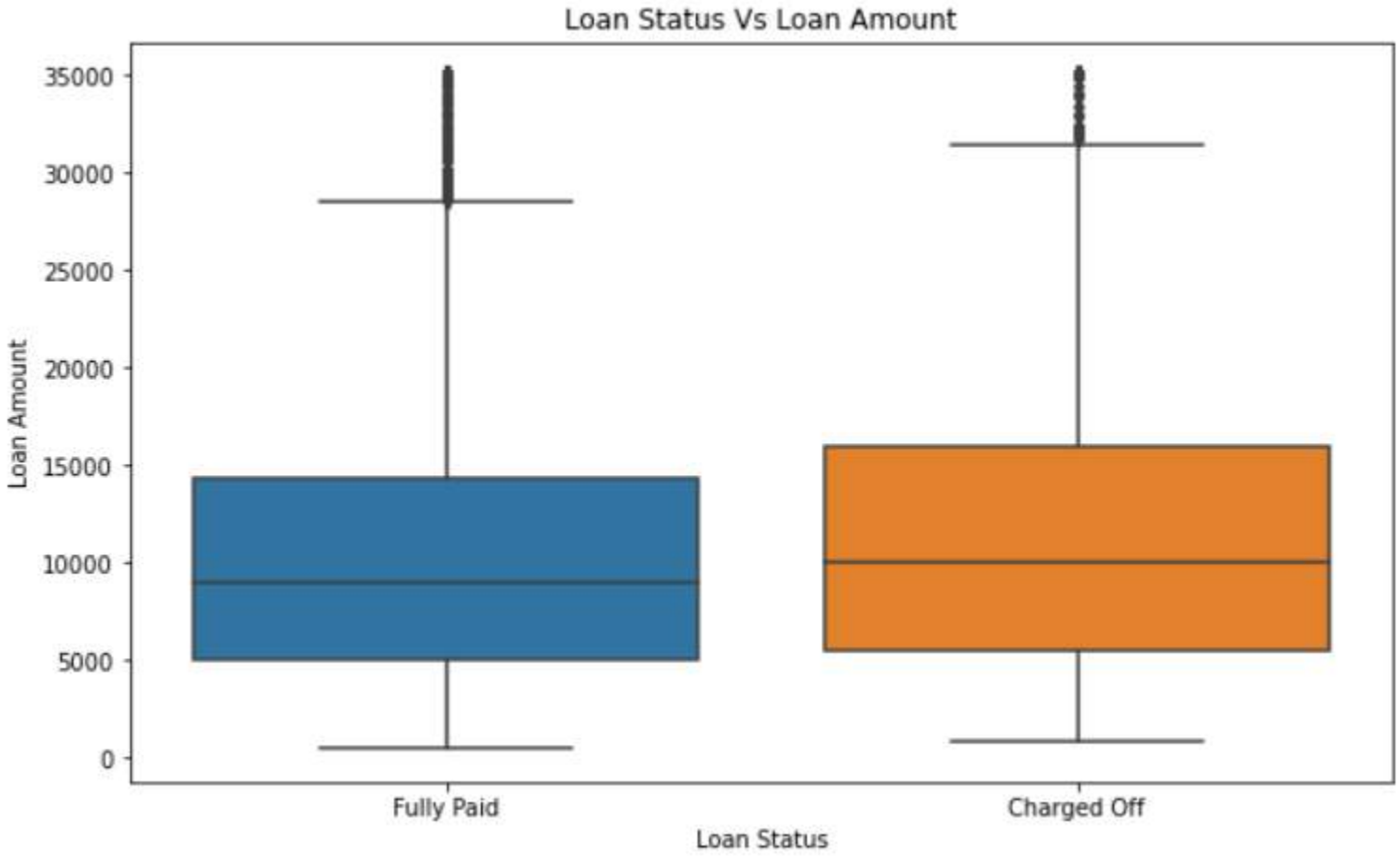
- We can see from above plot that the loan taken for the purpose of small business has more prone to defaulters.

Segmented Univariate Analysis on Categorical Variables



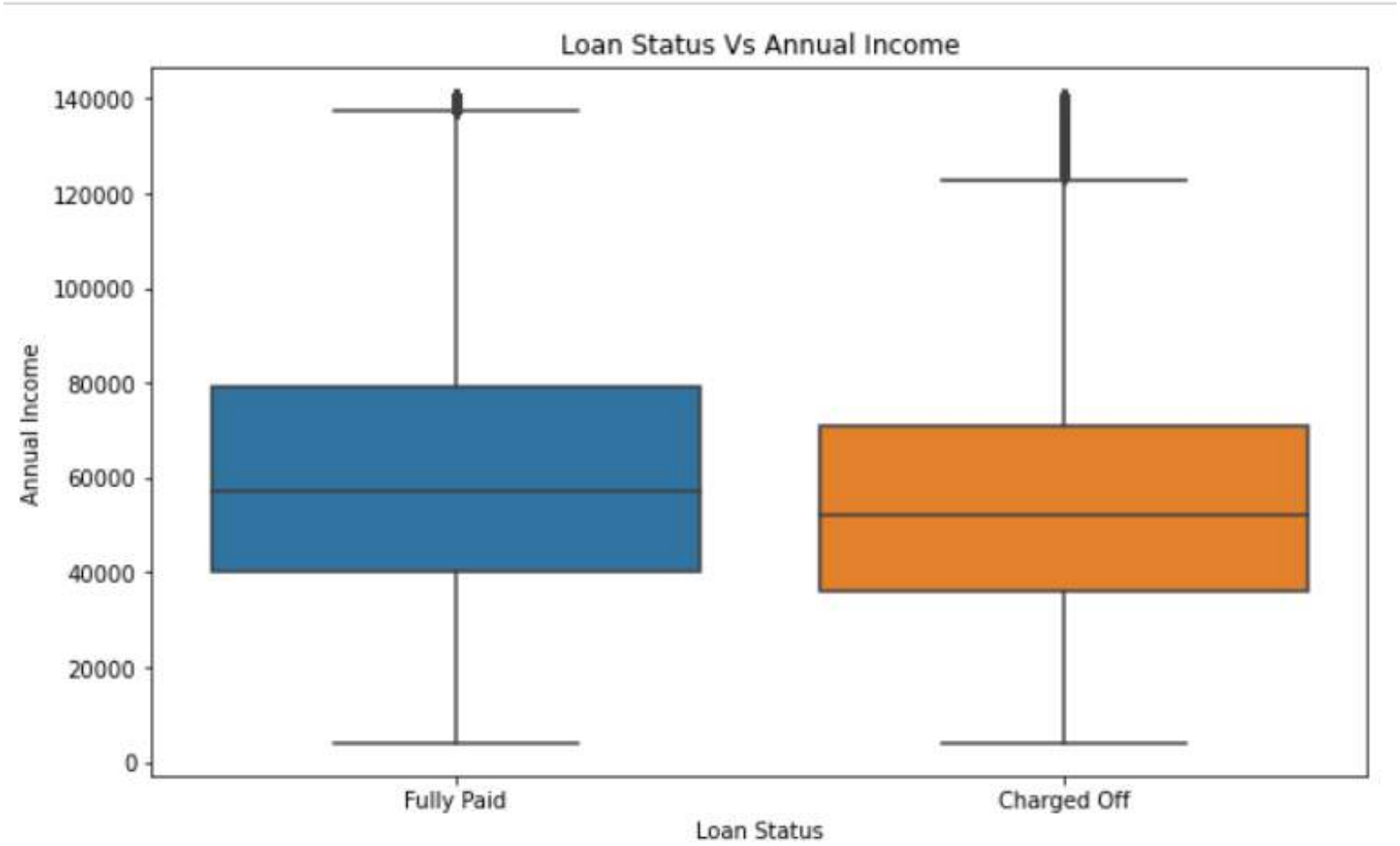
- We can see from above plot that Nebraska (NE) state borrowers are more prone to defaulters

Segmented Univariate Analysis on Continuous Variables



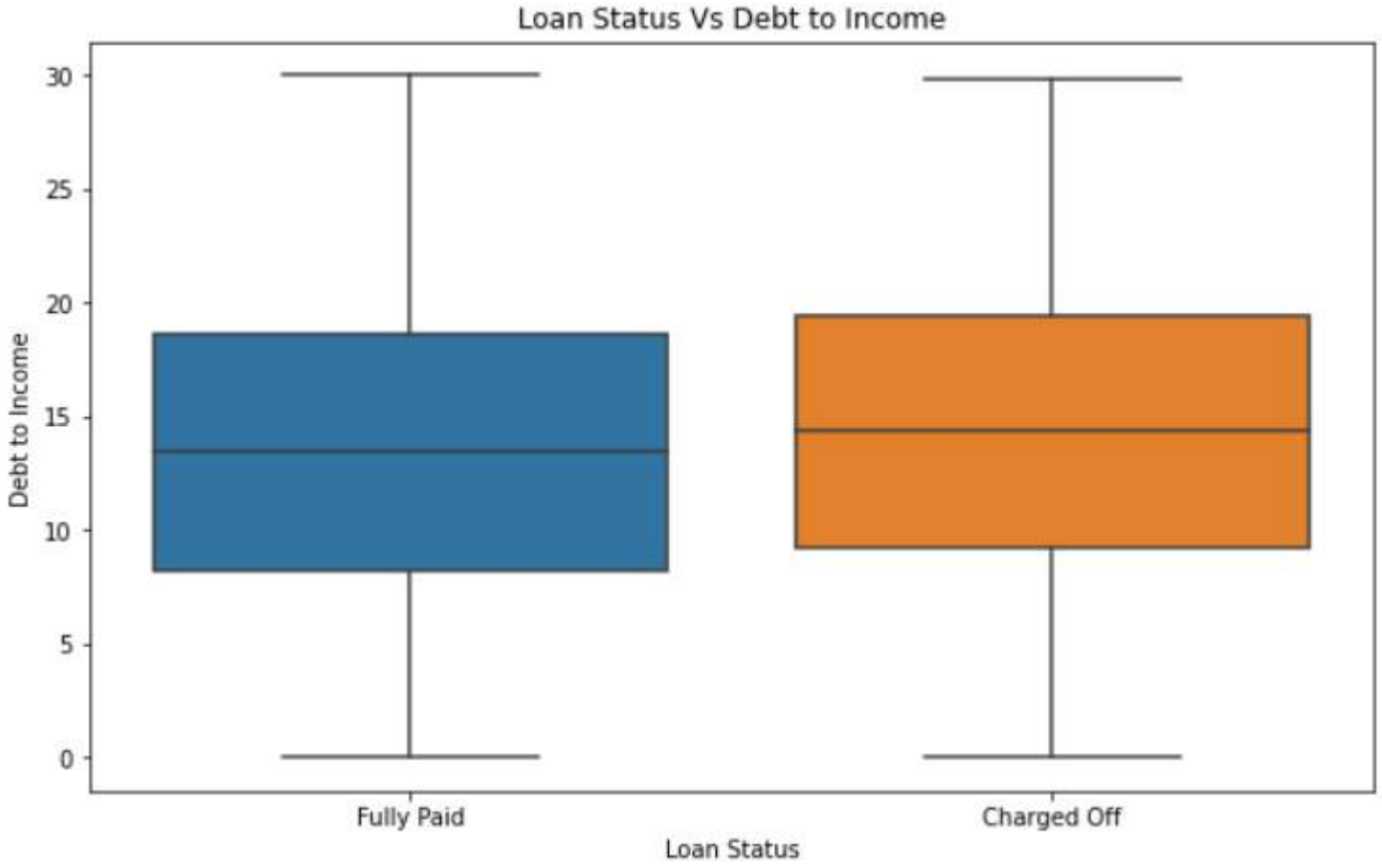
- We can see from the graph that higher the loan amount higher the chance of default.

Segmented Univariate Analysis on Continuous Variables



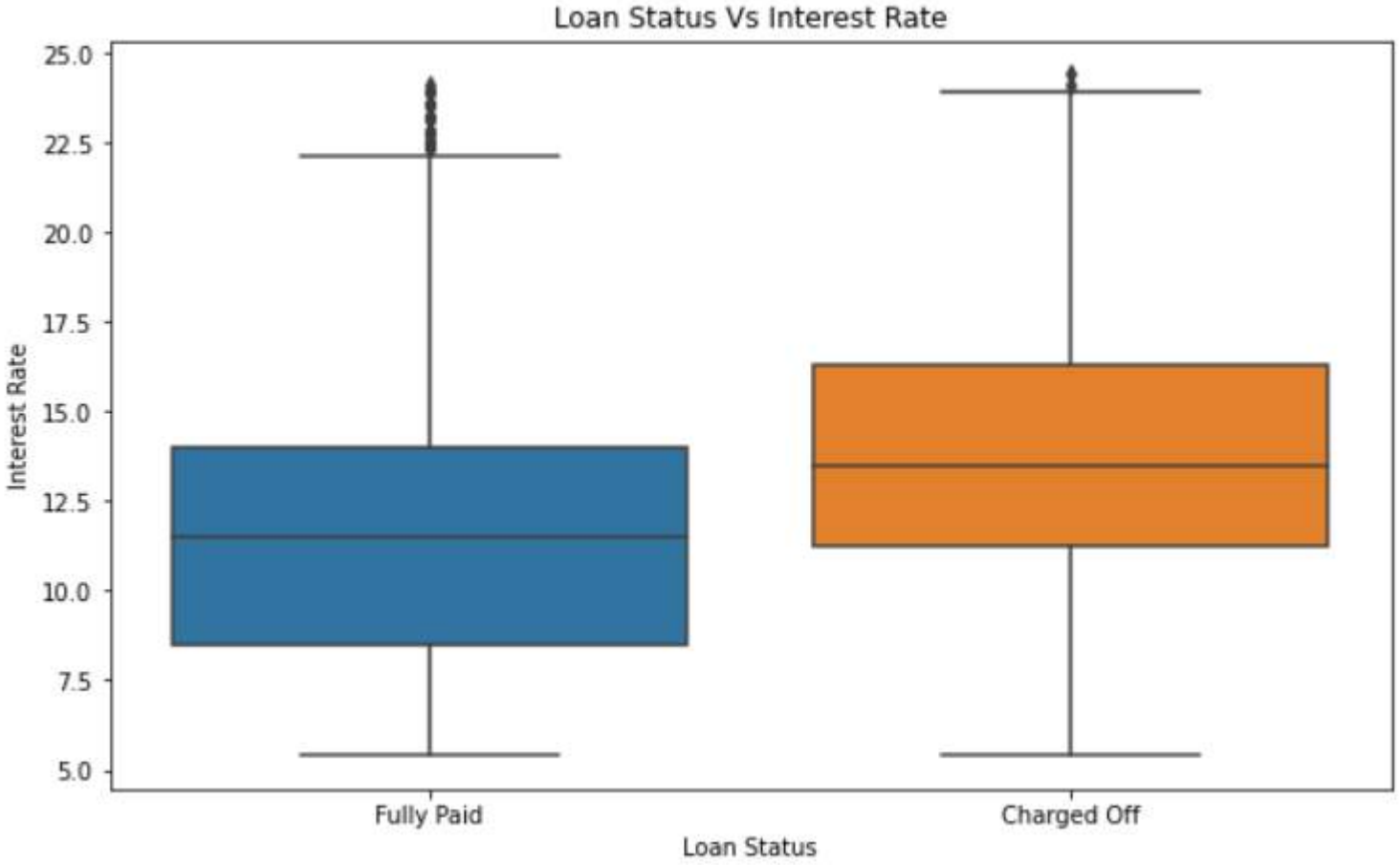
- We can see from the plot that lower the income higher the chance of charge off.

Segmented Univariate Analysis on Continuous Variables



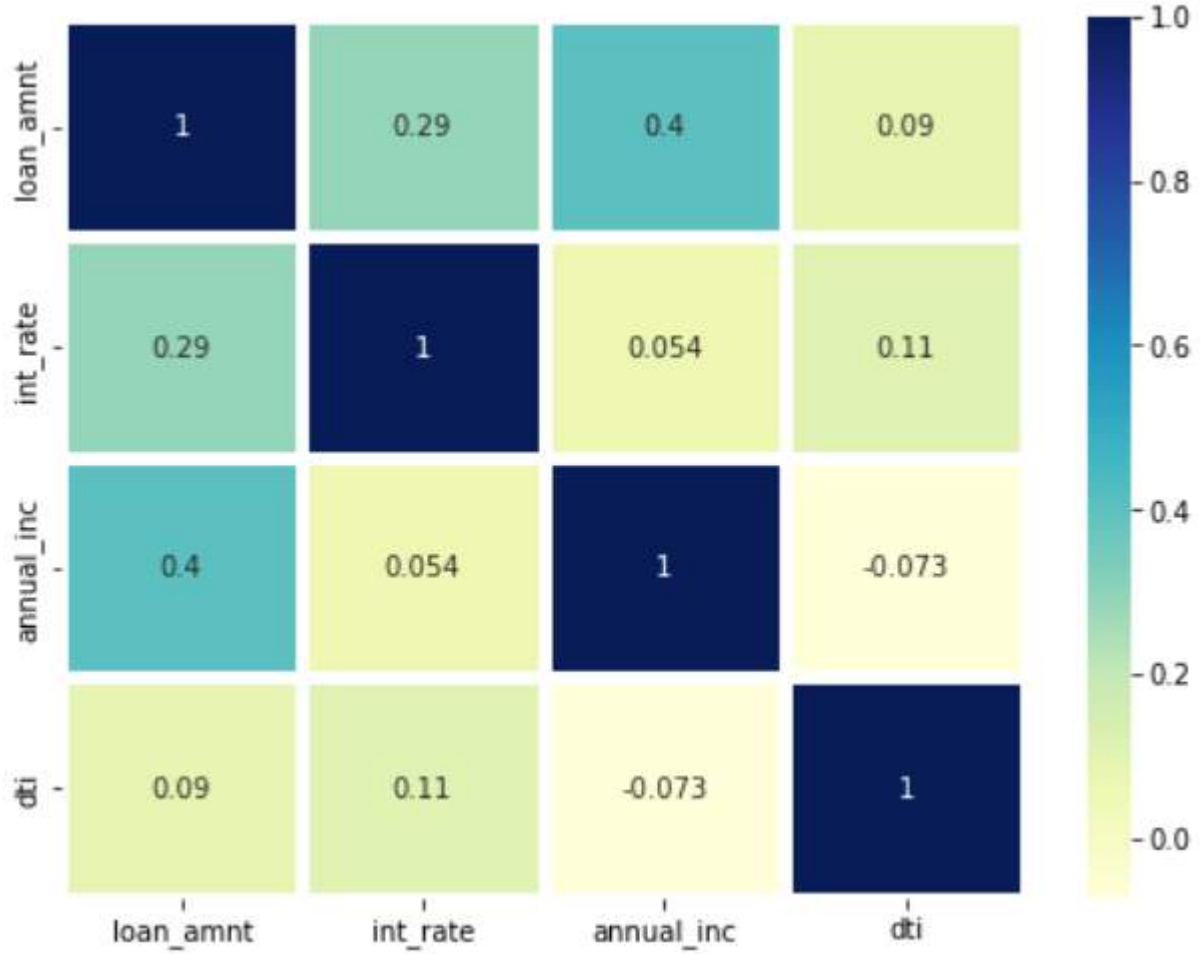
- We can see from the above boxplot that the low debt to income is slight less prone to charge off.

Segmented Univariate Analysis on Continuous Variables



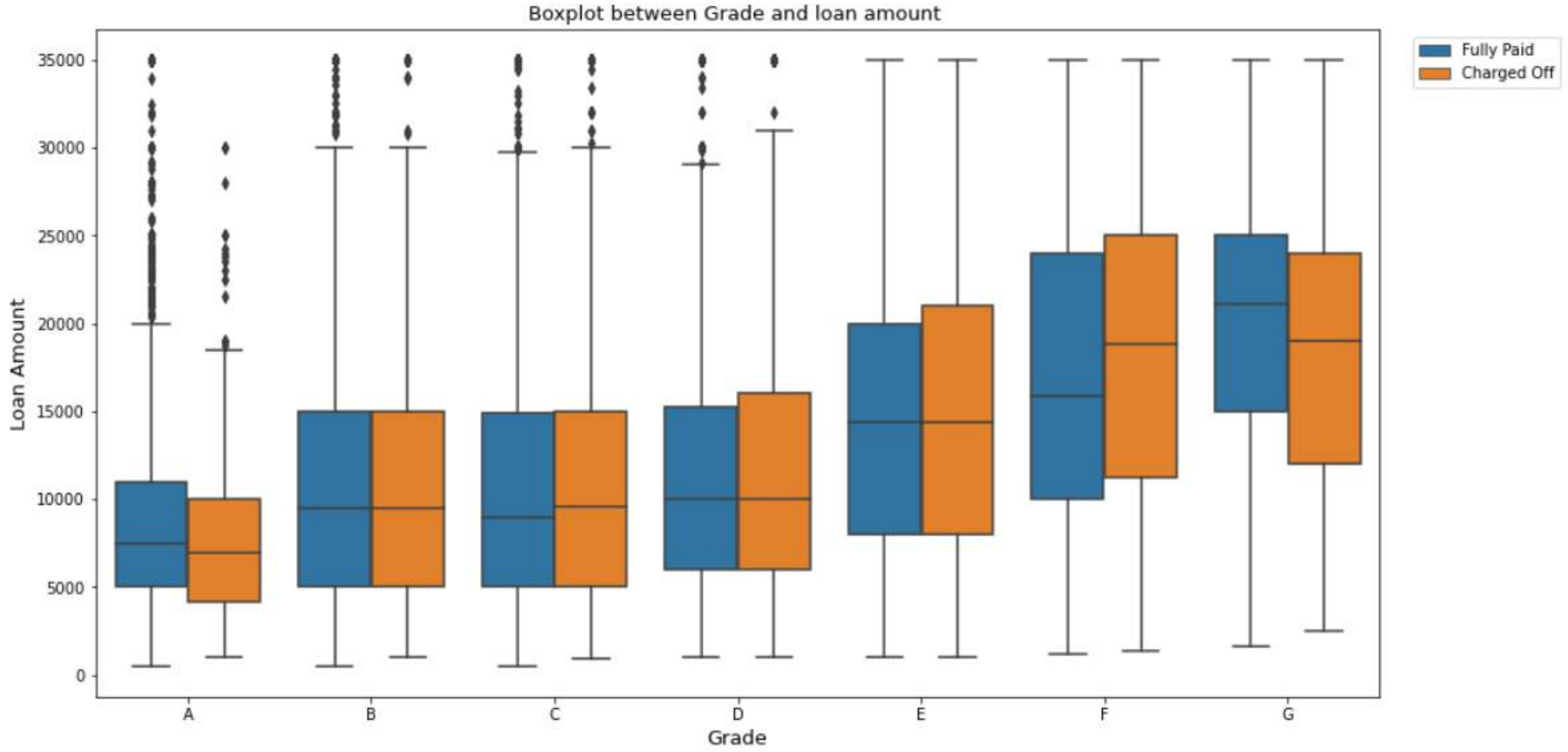
- We can see from the above boxplot that the higher interest rates are more prone to charge off loan

Bivariate Analysis on Continuous Variables



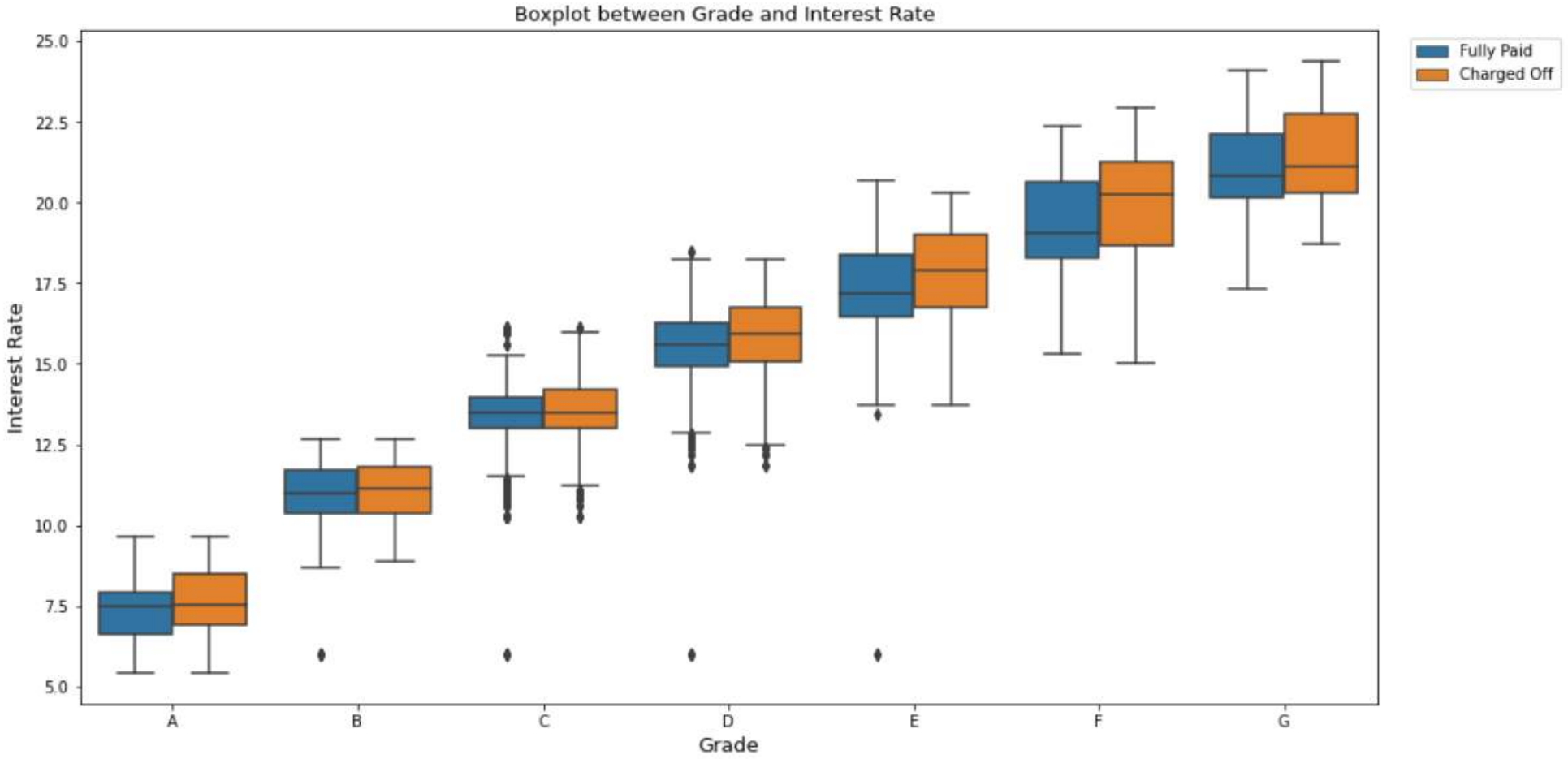
- Loan amount and Annual Income is positively correlated with the correlation value of .4
  Other correlation values are very small

Bivariate Analysis on Categorical Variables



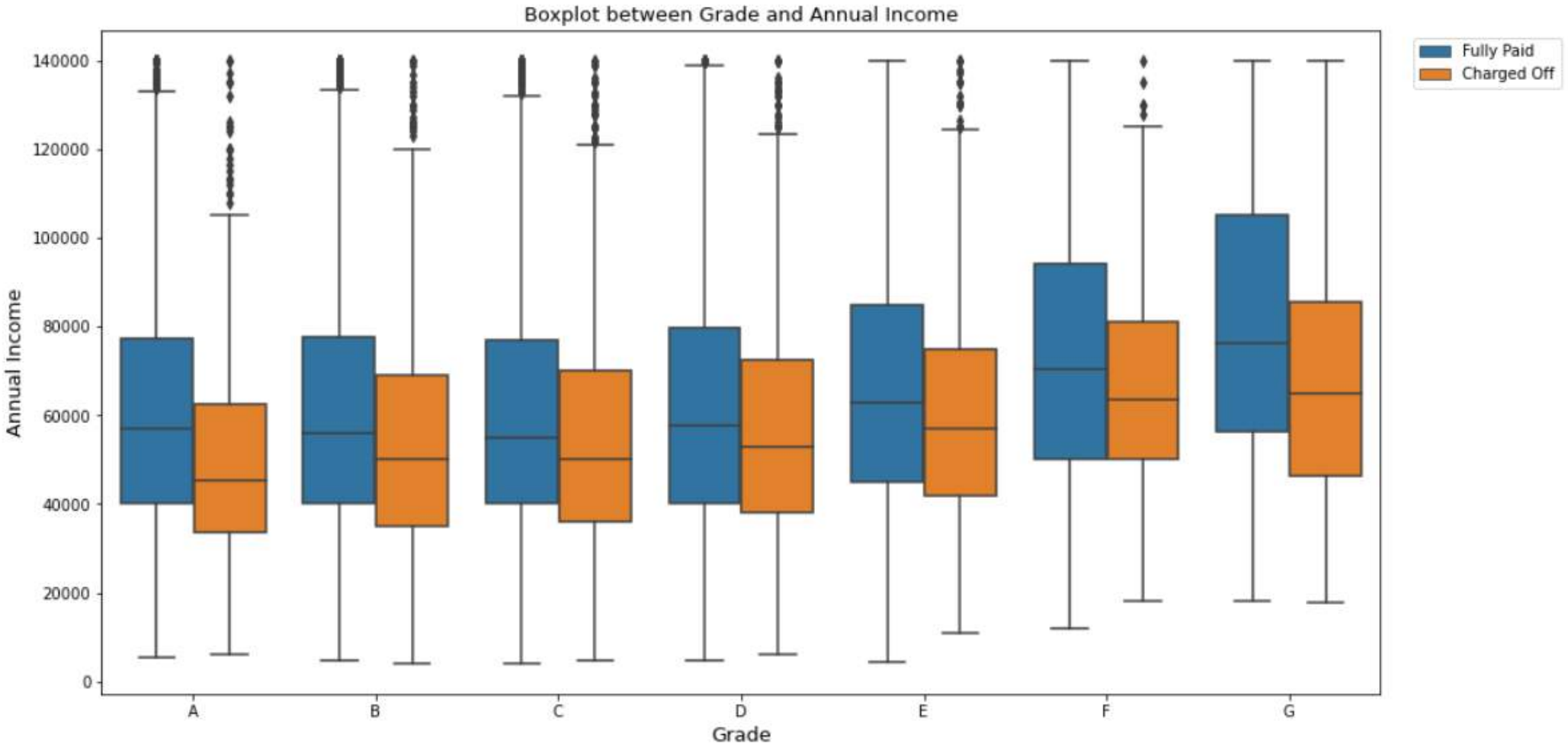
- We can se from the plot that for Grade F, higher the loan amount higher are the chances of default

Bivariate Analysis on Categorical Variables



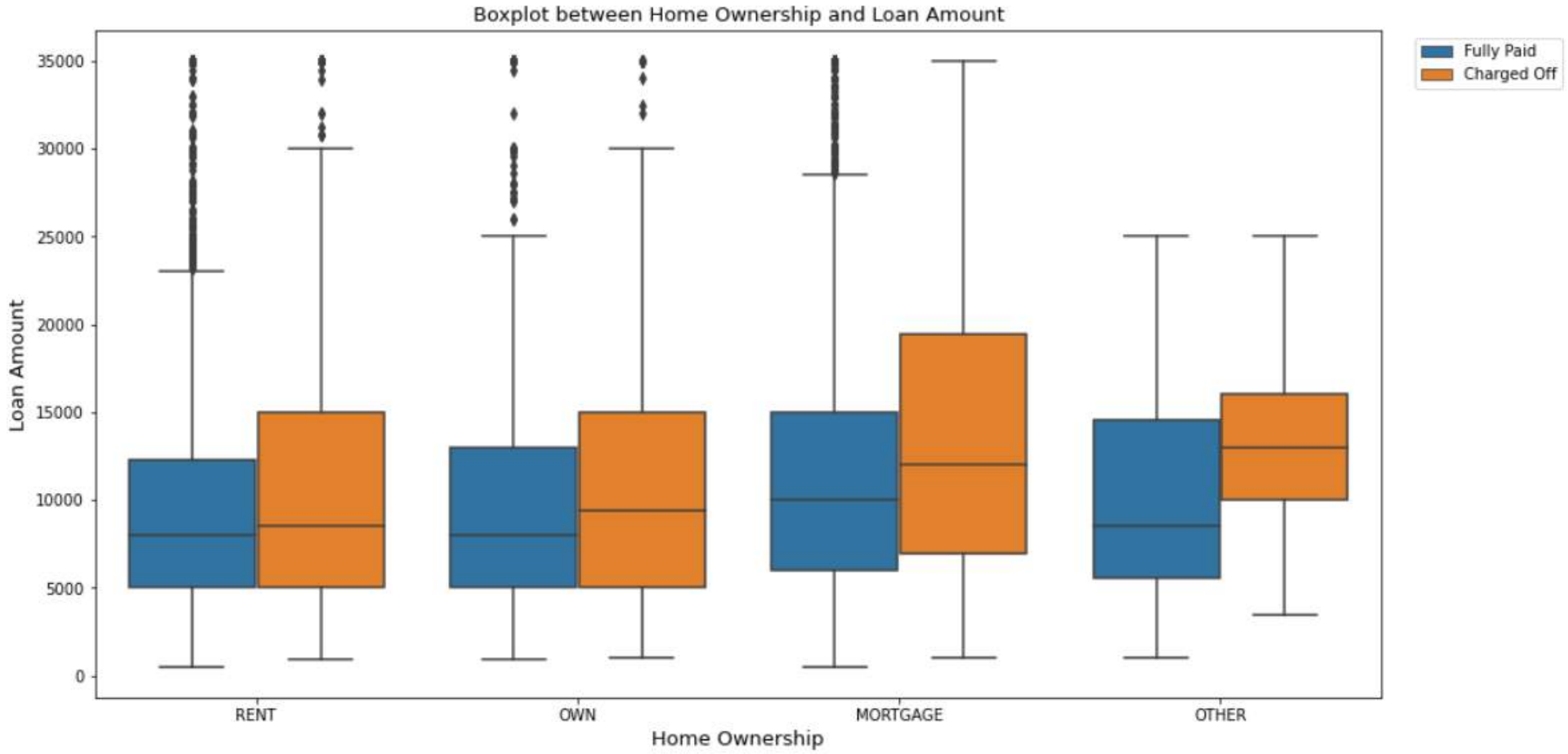
- We can see as the grade is moving from A to G, the interest rate is also increasing
- We can see from the plot for Grade F, G, higher the interest amount, higher the chances of default.

Bivariate Analysis on Categorical Variables



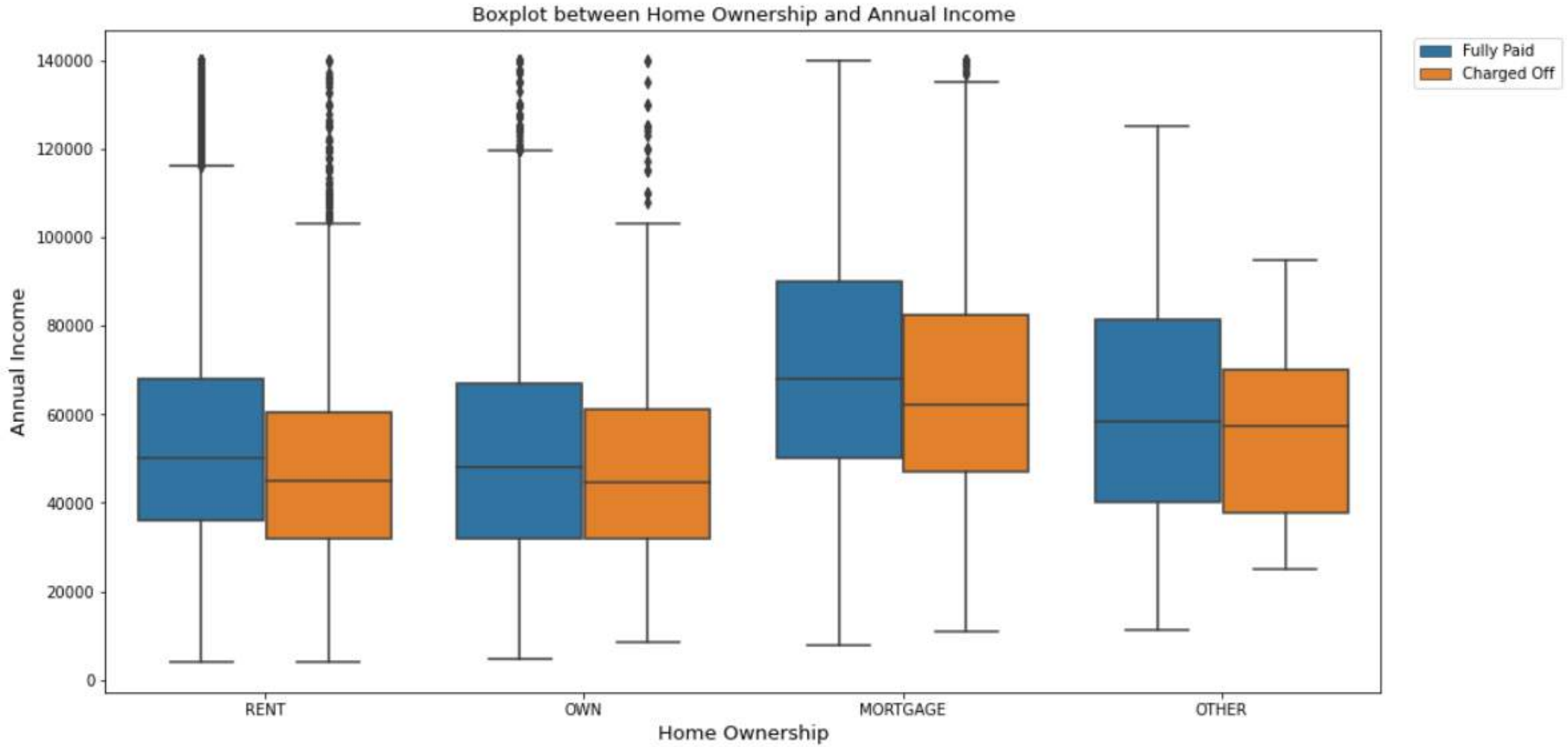
- We can see from the graph that the low annual income leads to charge off in each grade category.

Bivariate Analysis on Categorical Variables



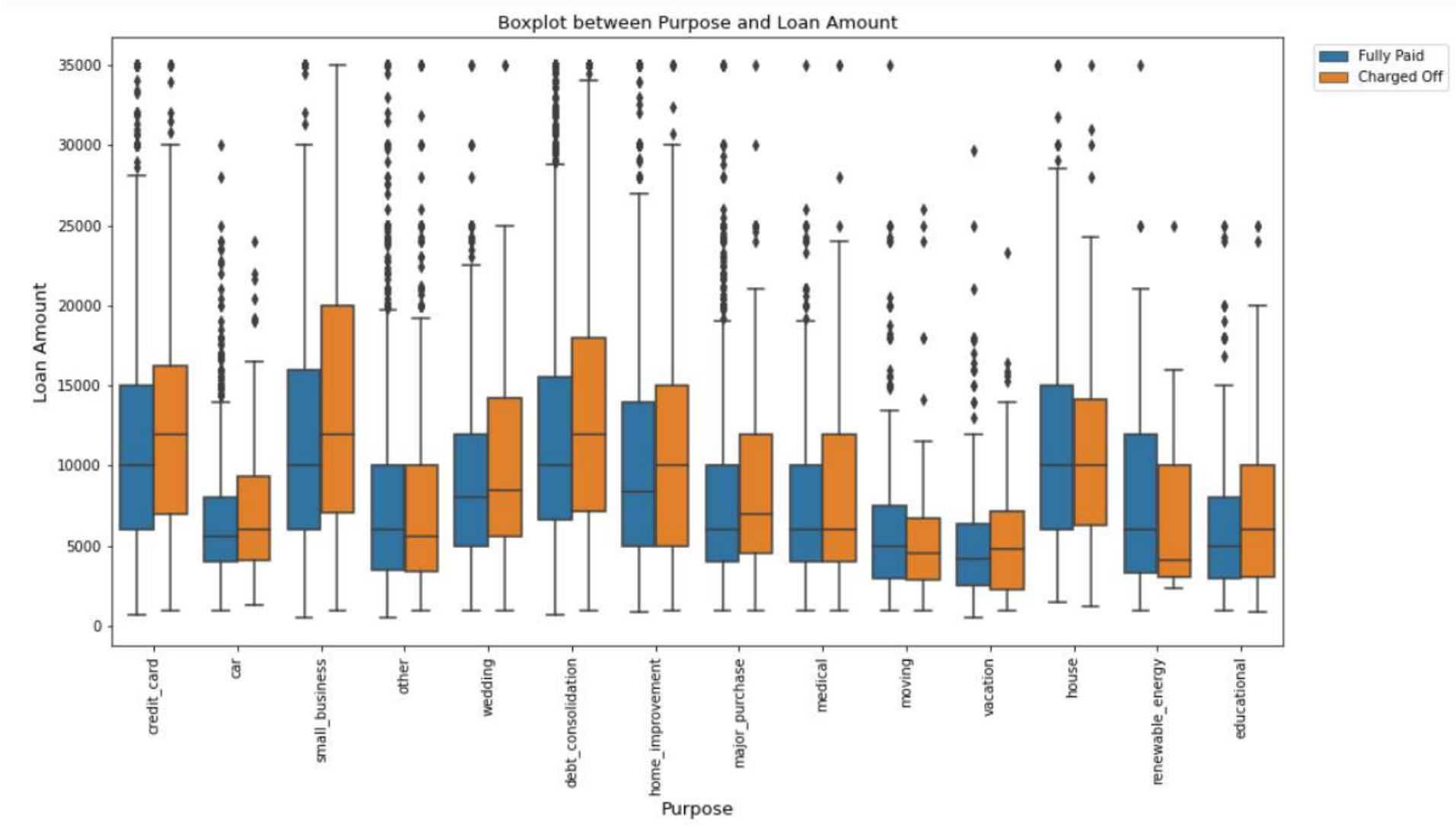
- For Mortgage homes, higher the loan amount, higher the chances of default.

Bivariate Analysis on Categorical Variables



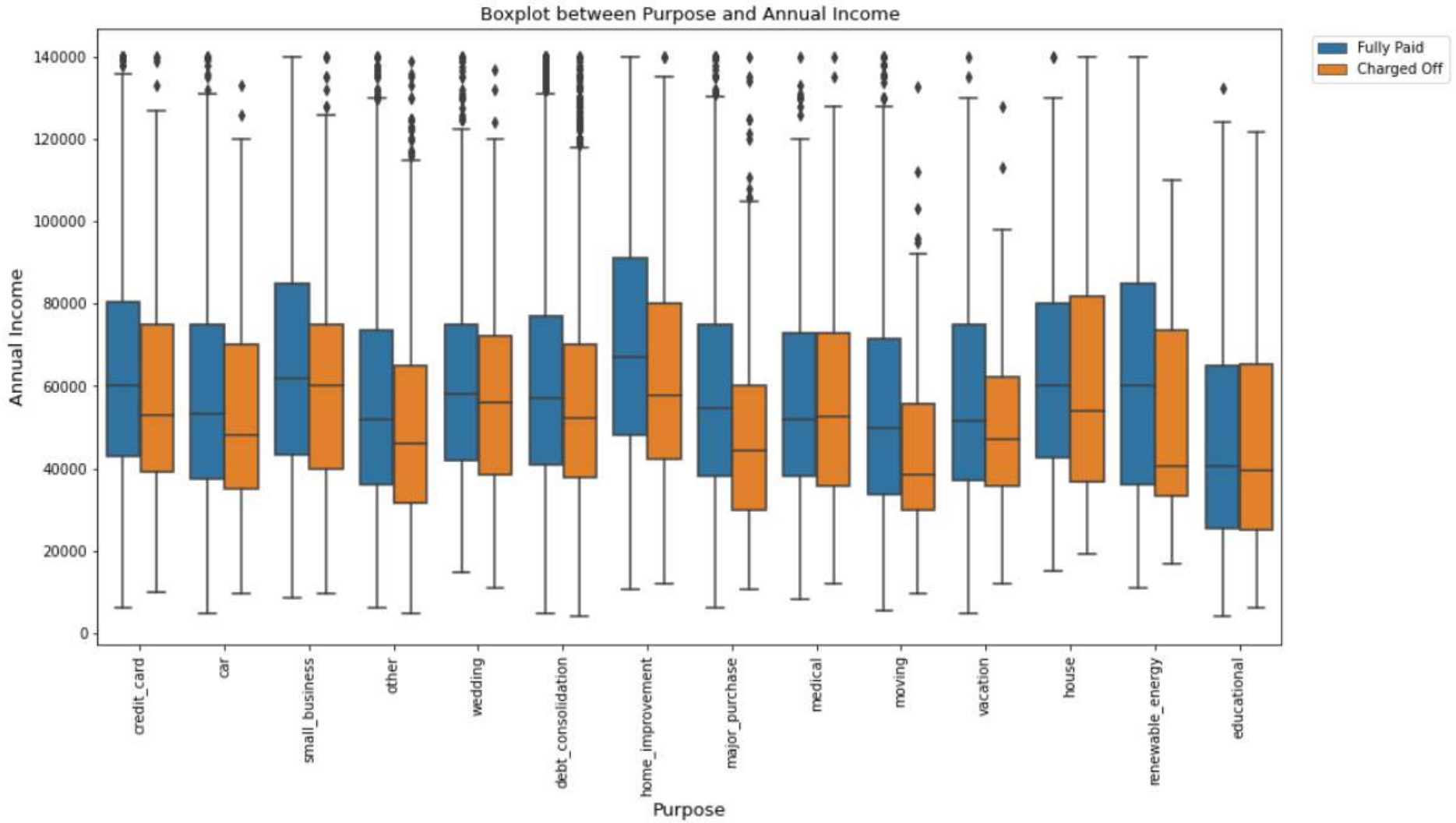
- We cannot comment here anything as it shows same for all types ownership, lesser the income, higher the chance of default

Bivariate Analysis on Categorical Variables



- We can see from the plot that for Small business, Debt consolidation and education loan if the loan amount is high. higher the chance of default.

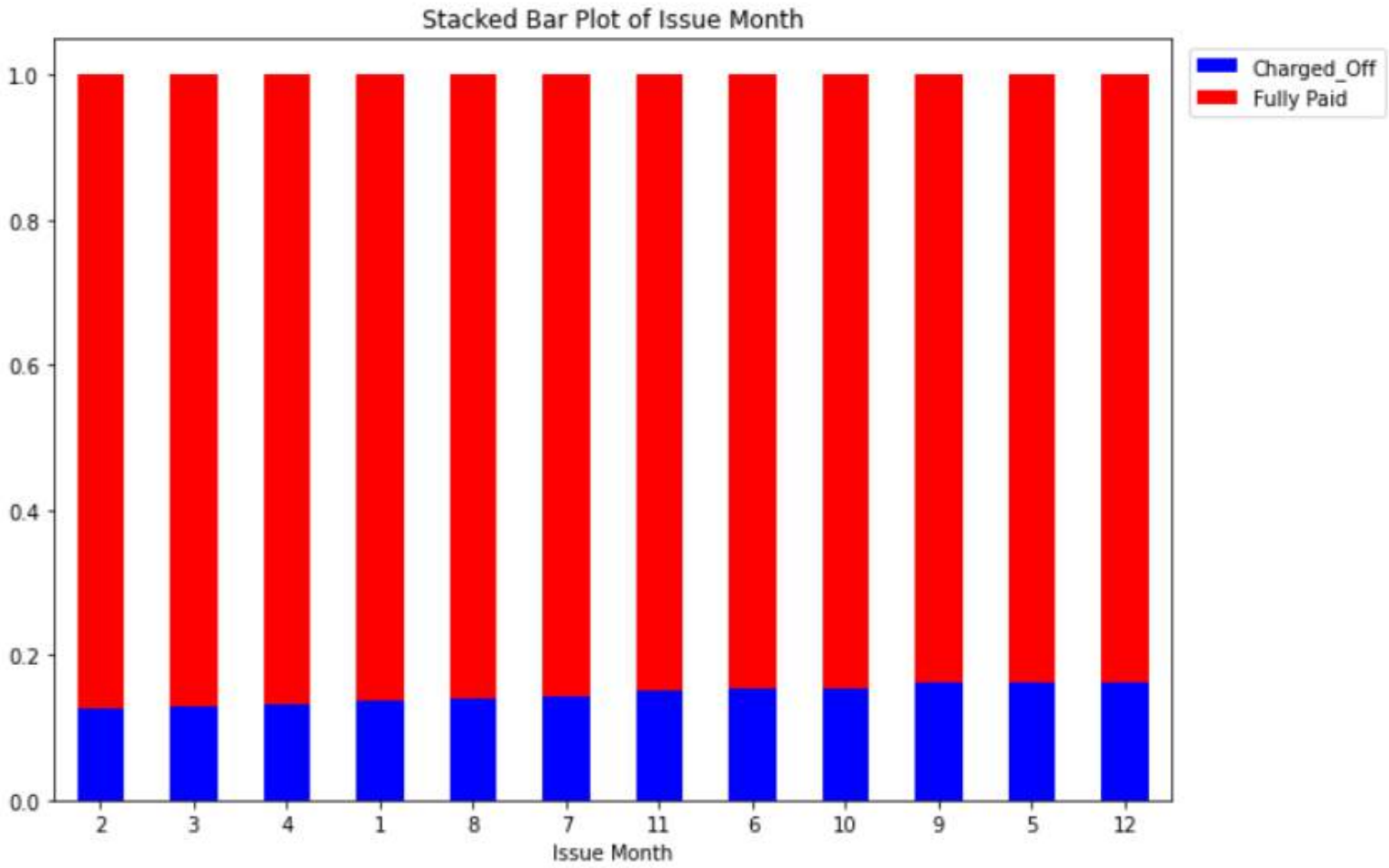
Bivariate Analysis on Categorical Variables



- We can see from the graph for the renewable energy if the annual income is low, higher the chance of default.

Derived Matric

- Derived issue month and issue year from issue_d column



- The above plot shows 'December' and 'May' has the high possibility of defaulting

Recommendations based on Univariate Analysis

- Borrowers under Grade G has highest chances of defaulters : 36%
- Borrowers under sub grade F5 is more prone to defaulters : 51%
- Most of the defaulters are the one who has term of 60 months : 25%
- Verified customers are more prone to defaulters : 17%
- Loan taken for the purpose of small business has more prone to defaulters : 28%
- Nebraska (NE) state borrowers are more likely to be charged off
- Higher the loan amount, higher the chances of default
- Lower the income, higher the chances of charge off
- Higher interest rates are more prone to charge off loan

Recommendations based on Bivariate Analysis

- Borrowers under Grade F and G has higher loan amount with high interest resulting in higher chances of defaulters
- Lower annual income leads to charge off in each grade category
- For Mortgage homes, higher the loan amount, higher the chances of default.
- Loan taken for the purpose of Small business, Debt consolidation and education loan, if the loan amount is high there are higher chance of charged off
- Higher the Loan amount, higher the Annual Income. They are positively correlated with the correlation value of .4
- In the month of 'December' and 'May’ there is high possibility of defaulting