Lending Club Case Study

Gurumoorthi Ramanathan

Primary Analysis

- Data Understanding & Exploration
- We have 111 columns with 39717 entries
- There are some un-necessary columns which may not help us to analysis the defaults, Let's filter out only the useful columns for the analysis such as,

- loan_status
- loan_amnt
- funded_amnt
- funded_amnt_inv
- Grade
- sub_grade
- emp_length
- annual_inc
- Purpose
- total_acc
- open_acc
- home_ownership
- term
- int_rate
- addr_state
- verification_status
- total_pymnt
- issue_d
- pub_rec_bankruptcies

- Goal is to predict the defaulters, So the "Current" status loan entries will not help us in such case, Let's **eliminate** the entries having "Current" status in loan_status.
- Now we have narrow down the variables to 17.

Null Values & Normalization

- We have **1033** null values null values in emp_length
- The values have been filled with most common values in the column that is "**10+ years**"
- Since the **emp_length** represent the total no of experience the employee (borrower) holds. Considering this as numeric let's normalize this **variable to numeric**.
- Same normalization has been processed for **term** variable

```
loan_status              0
loan_amnt                0
funded_amnt              0
funded_amnt_inv          0
int_rate                 0
grade                    0
sub_grade                0
term                     0
emp_length            1033
annual_inc               0
purpose                  0
total_acc                0
open_acc                 0
home_ownership           0
addr_state               0
verification_status      0
total_pymnt              0
dtype: int64
```

	term	emp_length
36	months	10+ years
60	months	< 1 year
36	months	10+ years
36	months	10+ years
60	months	1 year
36	months	3 years
60	months	8 years
36	months	9 years
60	months	4 years
60	months	< 1 year

term	emp_length
36	10
60	0
36	10
36	10
36	3
60	8
36	9
60	4
60	0
60	5

Conti..

- Considering the **sub_grade** represents the grade number along the grade, Let's make this values to numeric by slicing
- int_rate represent the rate of interest for the loan. Let's normalize this as interval group (0-8%, 8-12% and so on)
- Similarly, "**home_ownership**" variables has 3 **NONE** values, considering the numbers – let's remove the NONE valued entries from the data frame.
- Issue date has month-year, let's create a new columns to separate the month and year

int_rate	grade	sub_grade
10.65%	B	B2
15.27%	C	C4
15.96%	C	C5
13.49%	C	C1
7.90%	A	A4
15.96%	C	C5
18.64%	E	E1
21.28%	F	F2
12.69%	B	B5
14.65%	C	C3

Year	Month
Dec	11
Dec	11
Dec	11
Dec	11
Dec	11
Dec	11
Dec	11
Dec	11
Dec	11
Dec	11

grade	sub_grade
B	2
C	4
C	5
C	1
A	4
C	5
E	1
F	2
B	5
C	3

int_rate_interval
8-12%
12-16%
12-16%
12-16%
0-8%
12-16%
16-20%
20-24%
8-12%
12-16%

Conti..

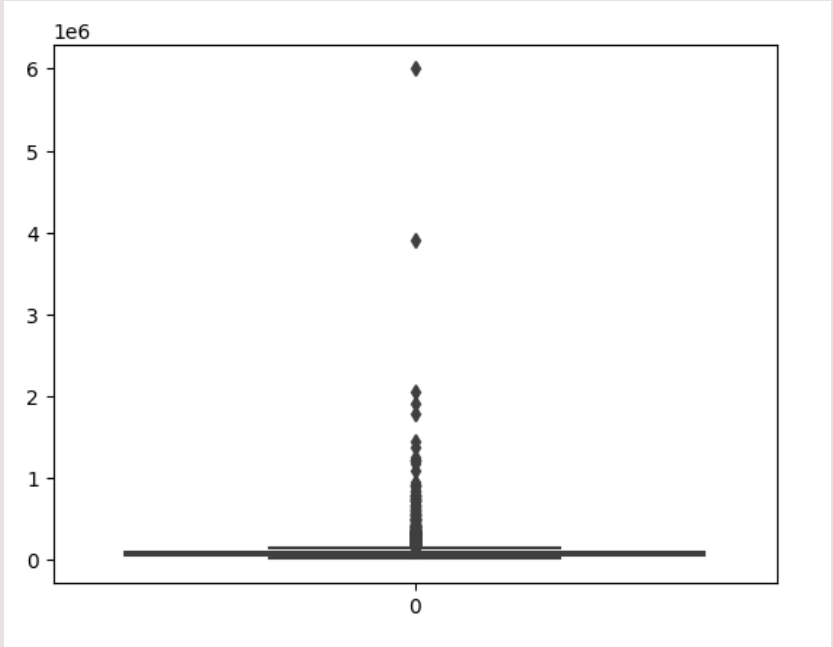
- annual_inc has multiple values, let's narrow it down with interval values
- Same for Loan amount

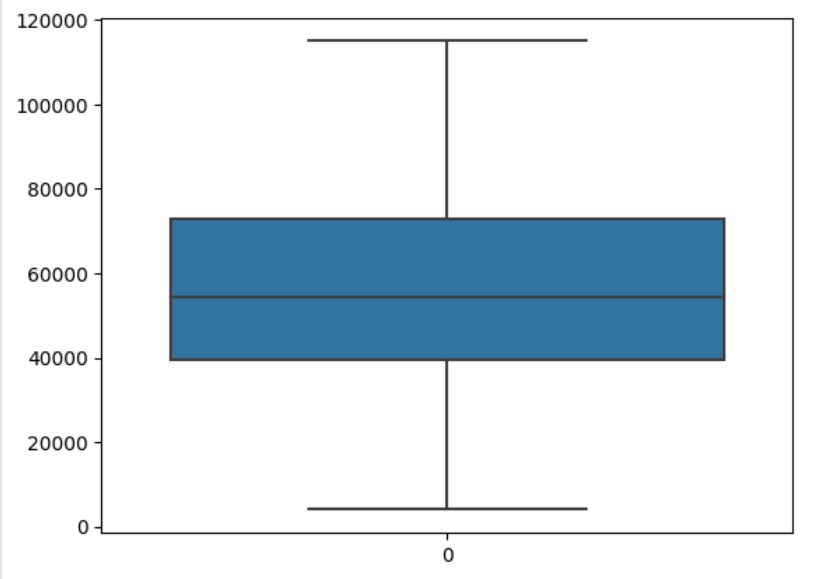
annual_inc	annual_inc_int	loan_status	loan_amnt_int
24000.0	15000-30000	Fully Paid	0-5000
30000.0	15000-30000	Charged Off	0-5000
12252.0	0-15000	Fully Paid	0-5000
49200.0	45000-60000	Fully Paid	5000-10000
36000.0	30000-45000	Fully Paid	0-5000
47004.0	45000-60000	Fully Paid	5000-10000
48000.0	45000-60000	Fully Paid	0-5000
40000.0	30000-45000	Charged Off	5000-10000
15000.0	0-15000	Charged Off	5000-10000
72000.0	60000- 75000	Fully Paid	5000-10000

Outliers

- By visualizing the “annual_inc”, annual income of the borrow into boxplot, it’s clearly shows us that the column has some outliers.
- Also, the quantiles against the variable shows that the values after 90% seems to be dropped.
- Let’s take the values below 90% to remove the outliers.

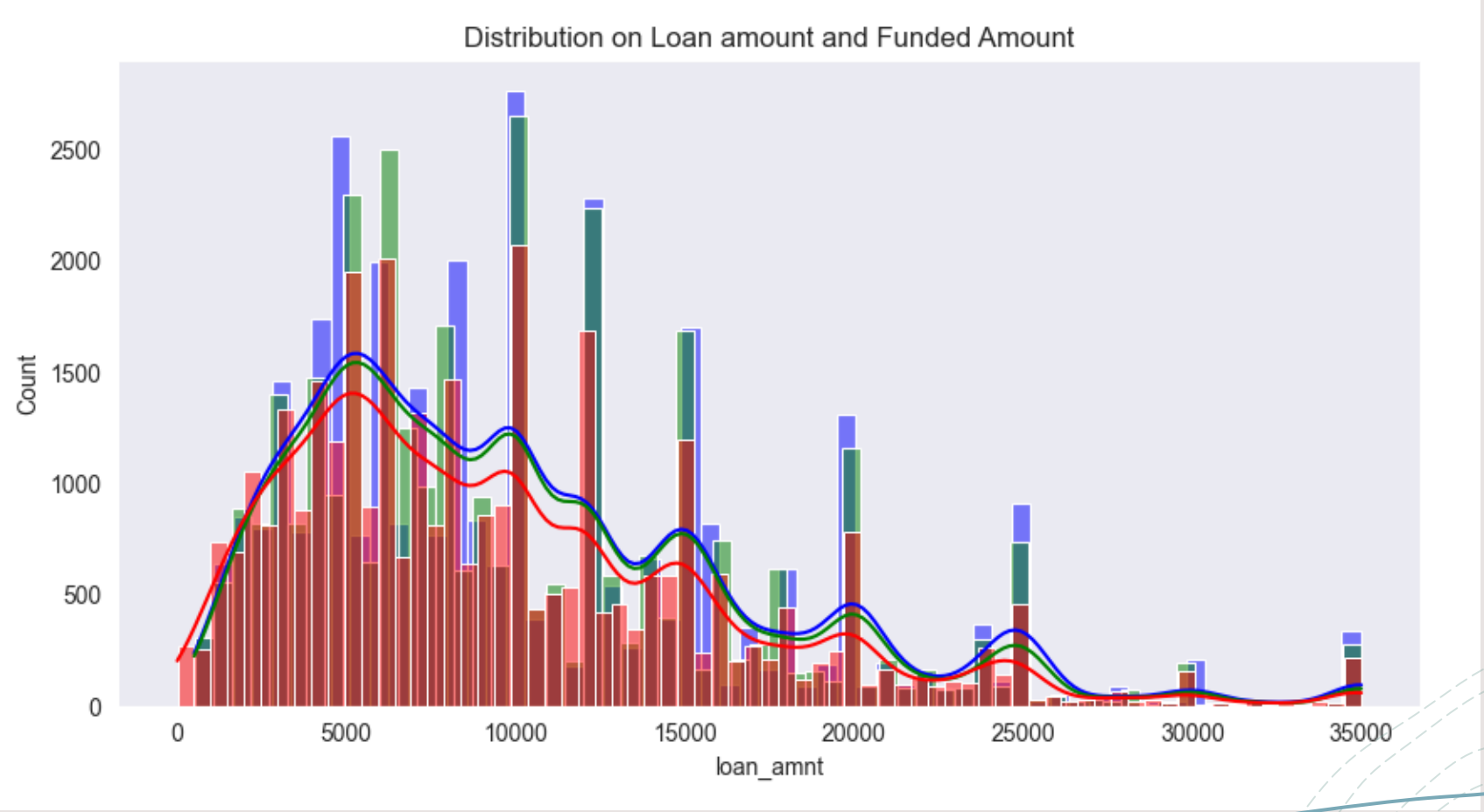
```
0.50     58860.28
0.60     65004.00
0.70     75000.00
0.75     82000.00
0.80     90000.00
0.85    100000.00
0.90    115000.00
0.95    140004.00
Name: annual_inc, dtype: float64
```





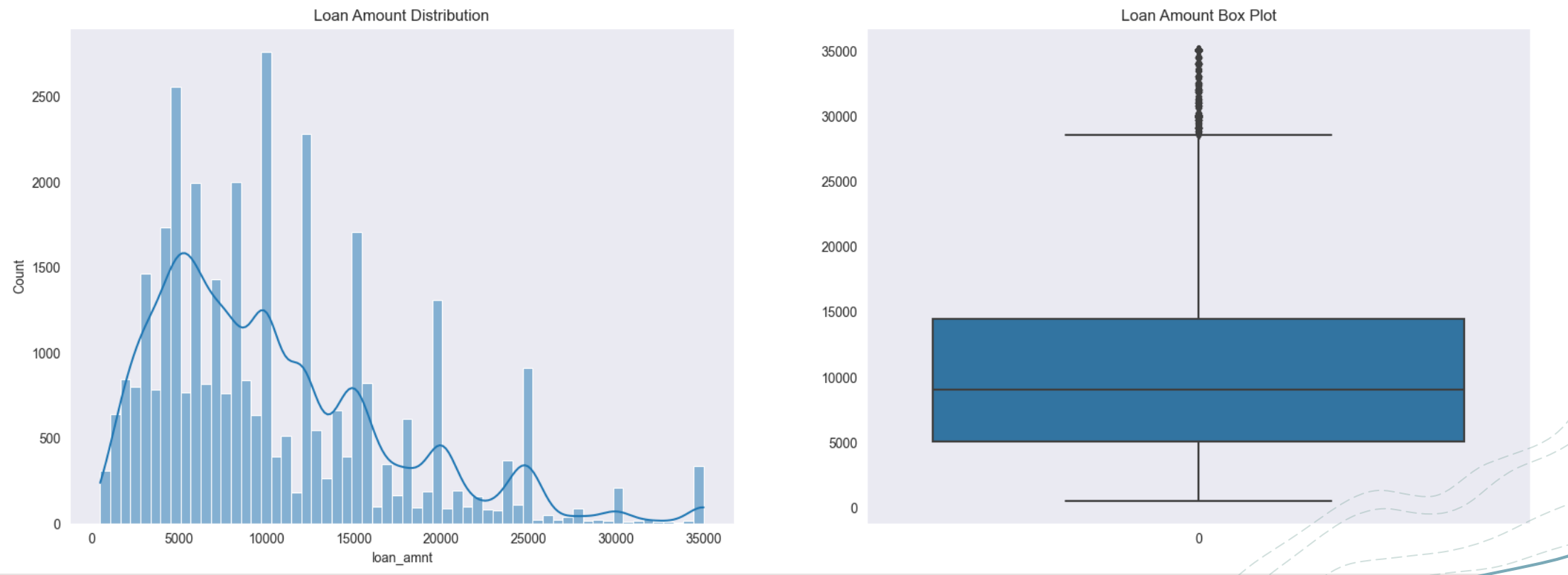
Visualizations

All the variables have almost similar values



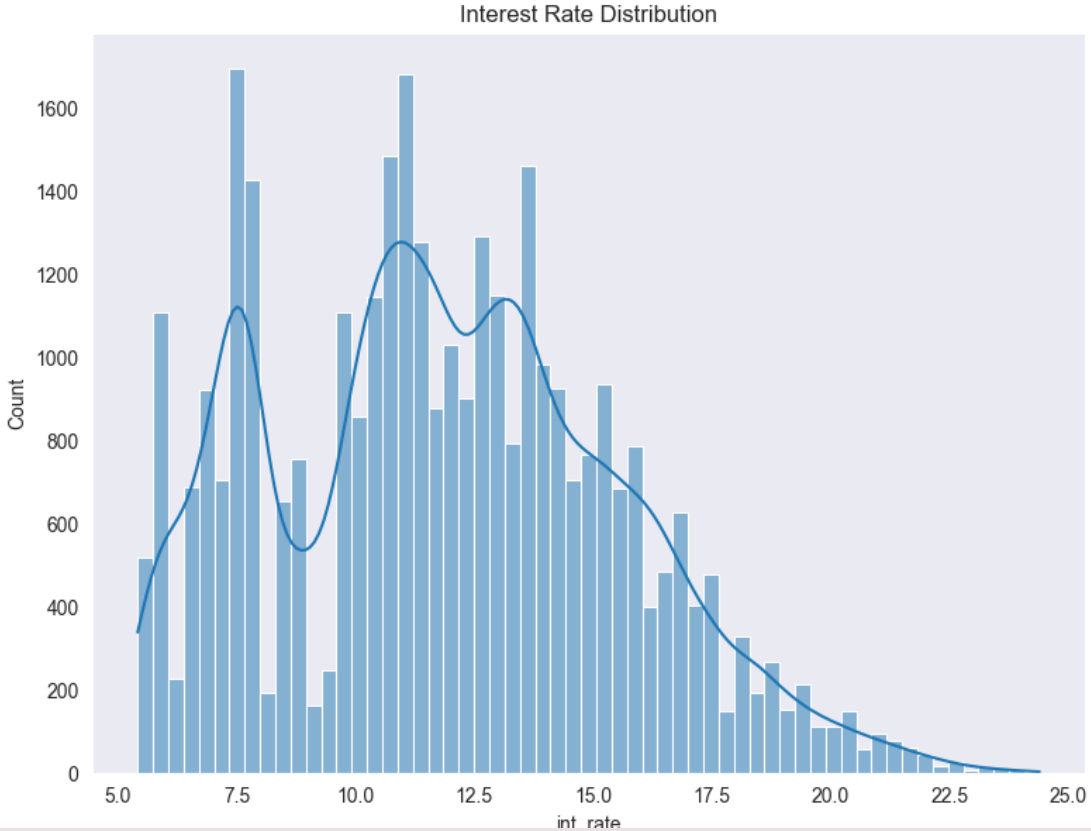
Univariate Analysis (Loan Amount)

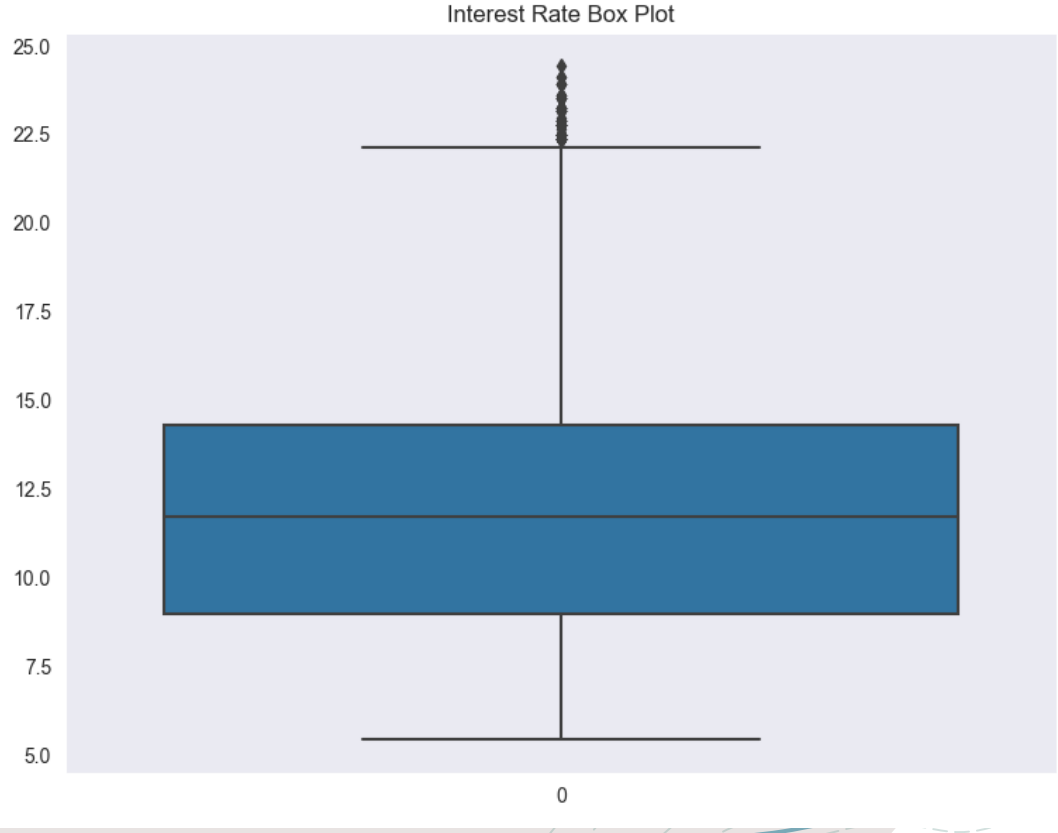
Most of the loans were given between 5000 to 15000



Conti.. Univariate Analysis (Interest Rate)

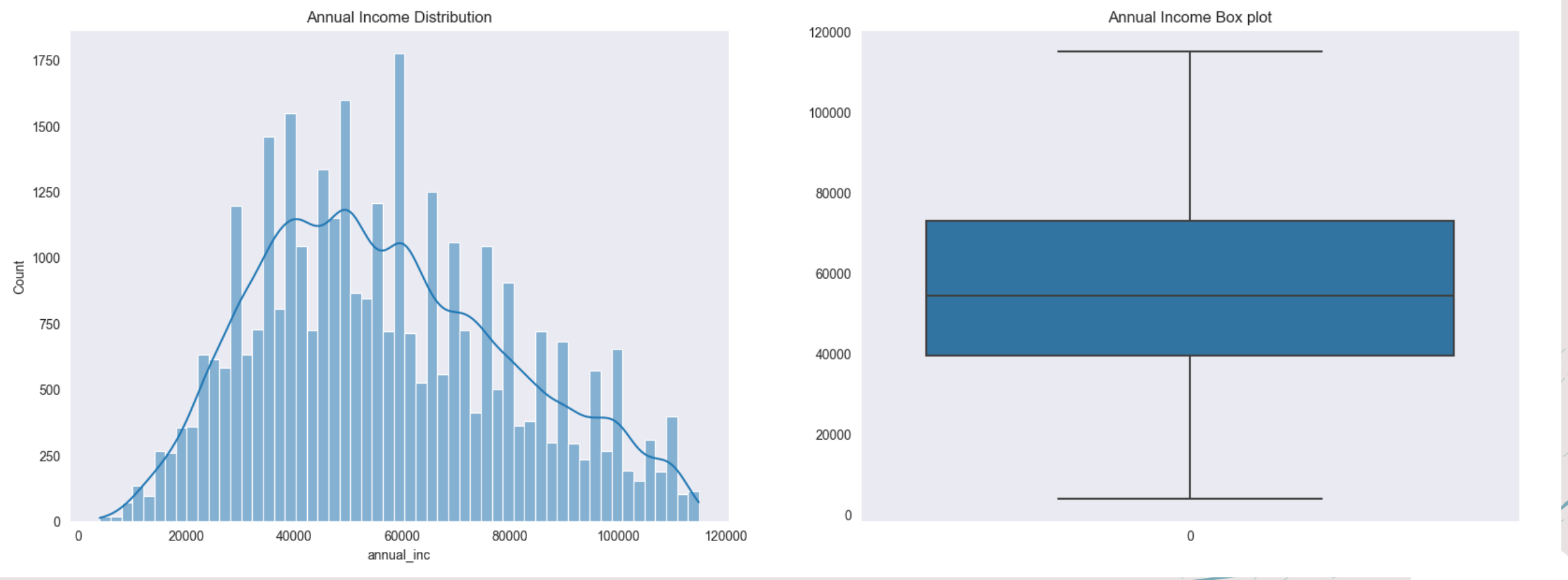
Interest rate ratio is high between 10.0% to 15.0%





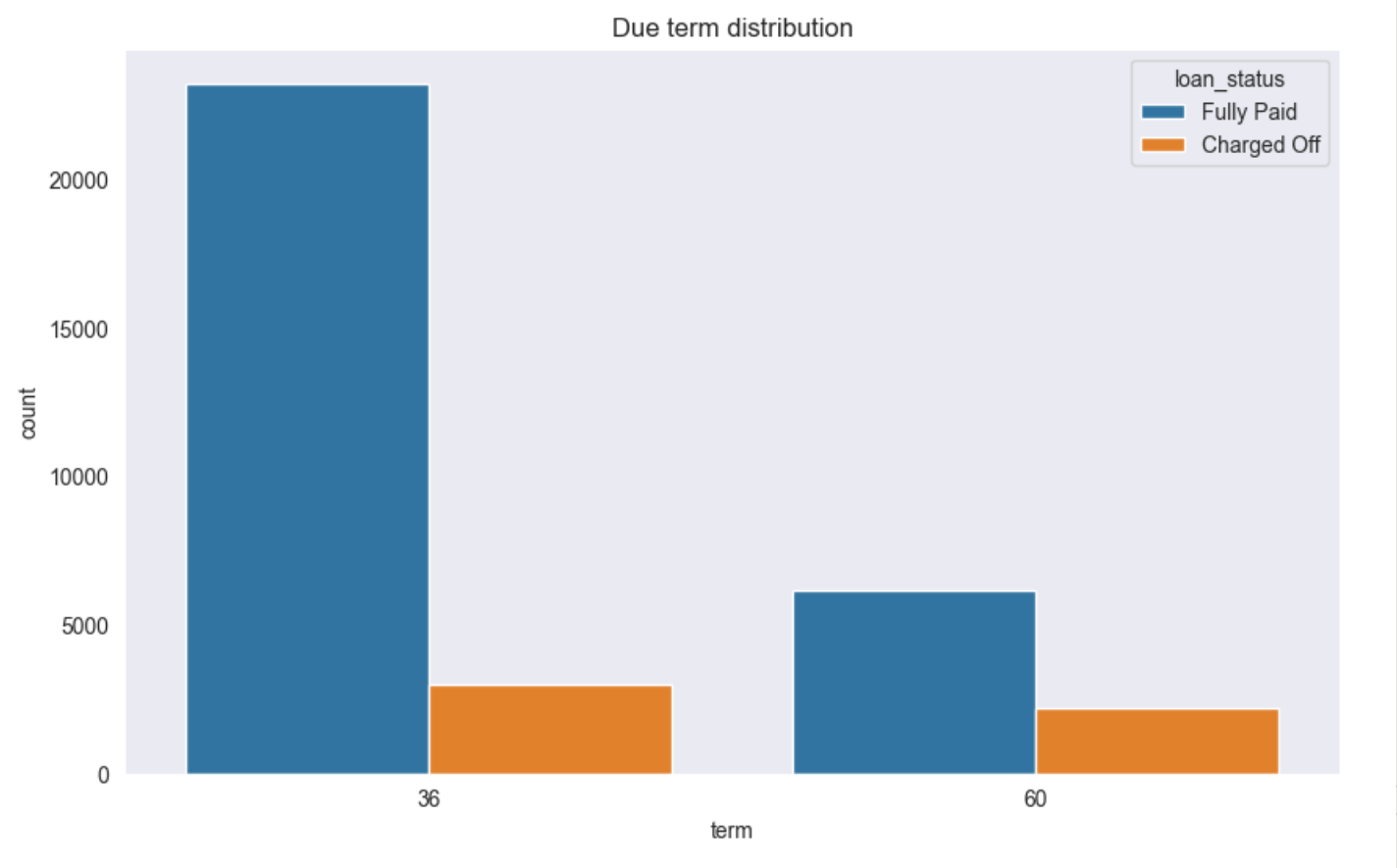
Conti.. Univariate Analysis (annual Income)

Interest rate ratio is high between 10.0% to 15.0%

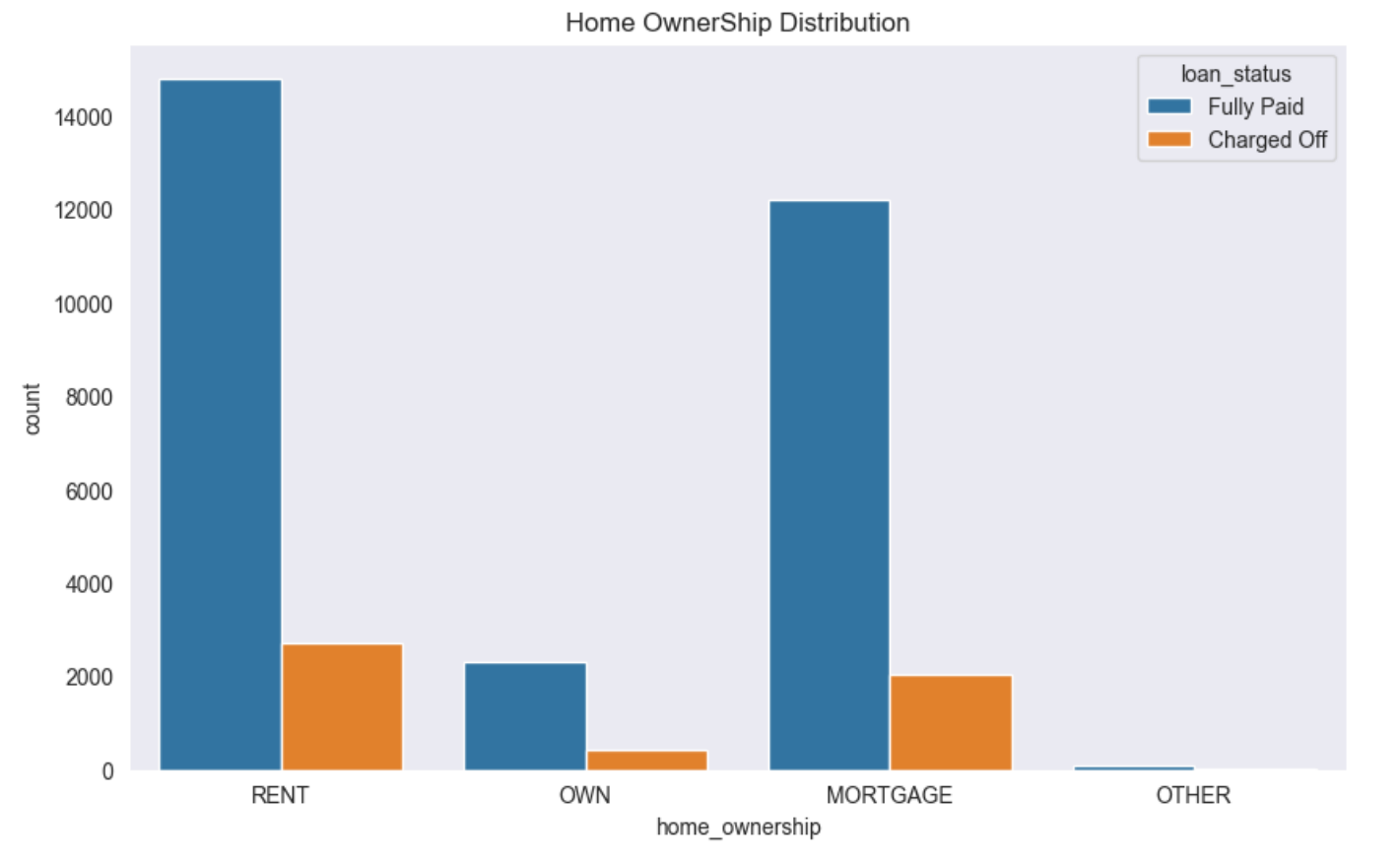


Conti.. Univariate Analysis (Due term)

Due term is high on 36 months, whereas charged off values were similar for each



Conti.. Univariate Analysis (Home Ownership)

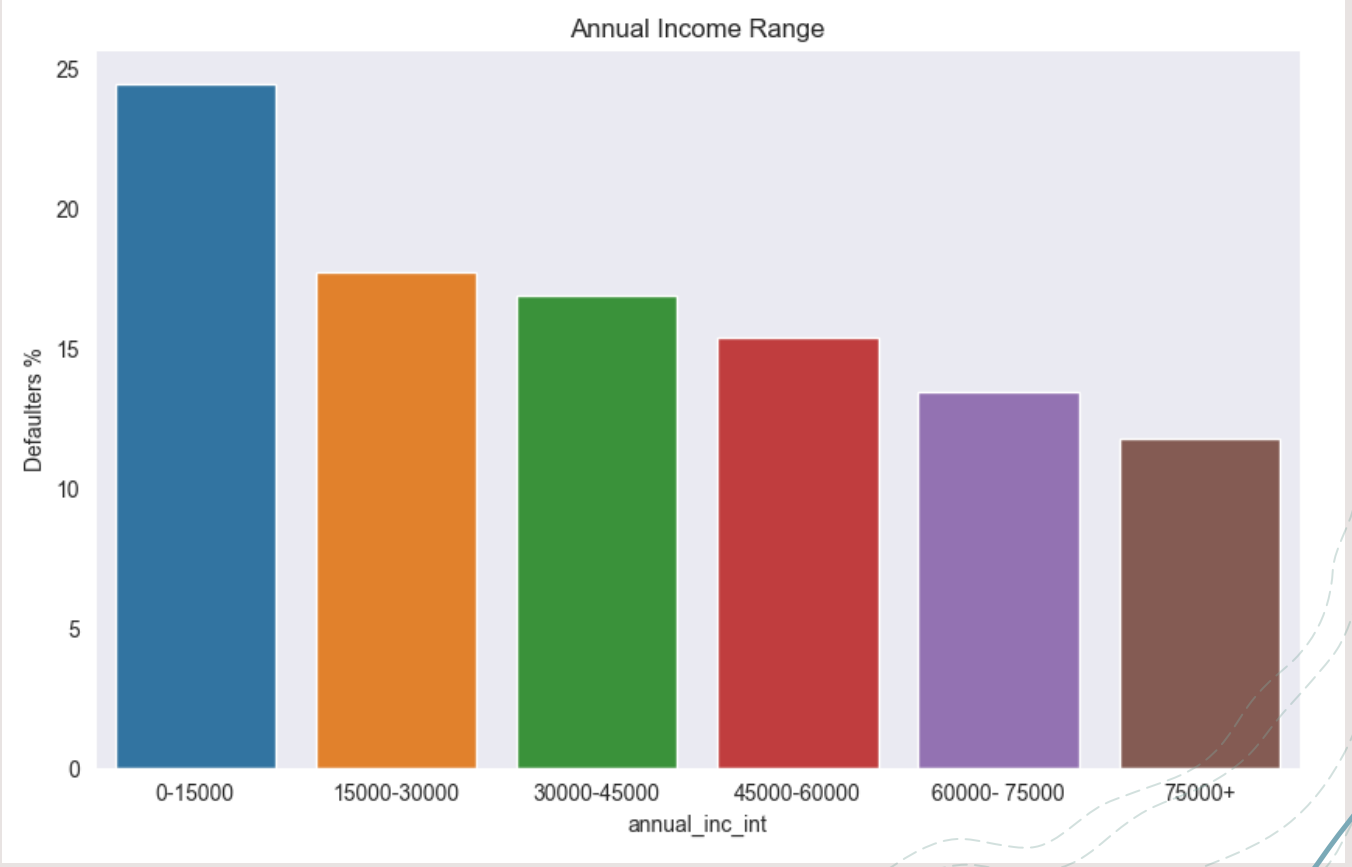


+ Applicants mostly from Rented and Mortgage

+ Notable thing is that those two variables having high no of Charged Off borrowers than others

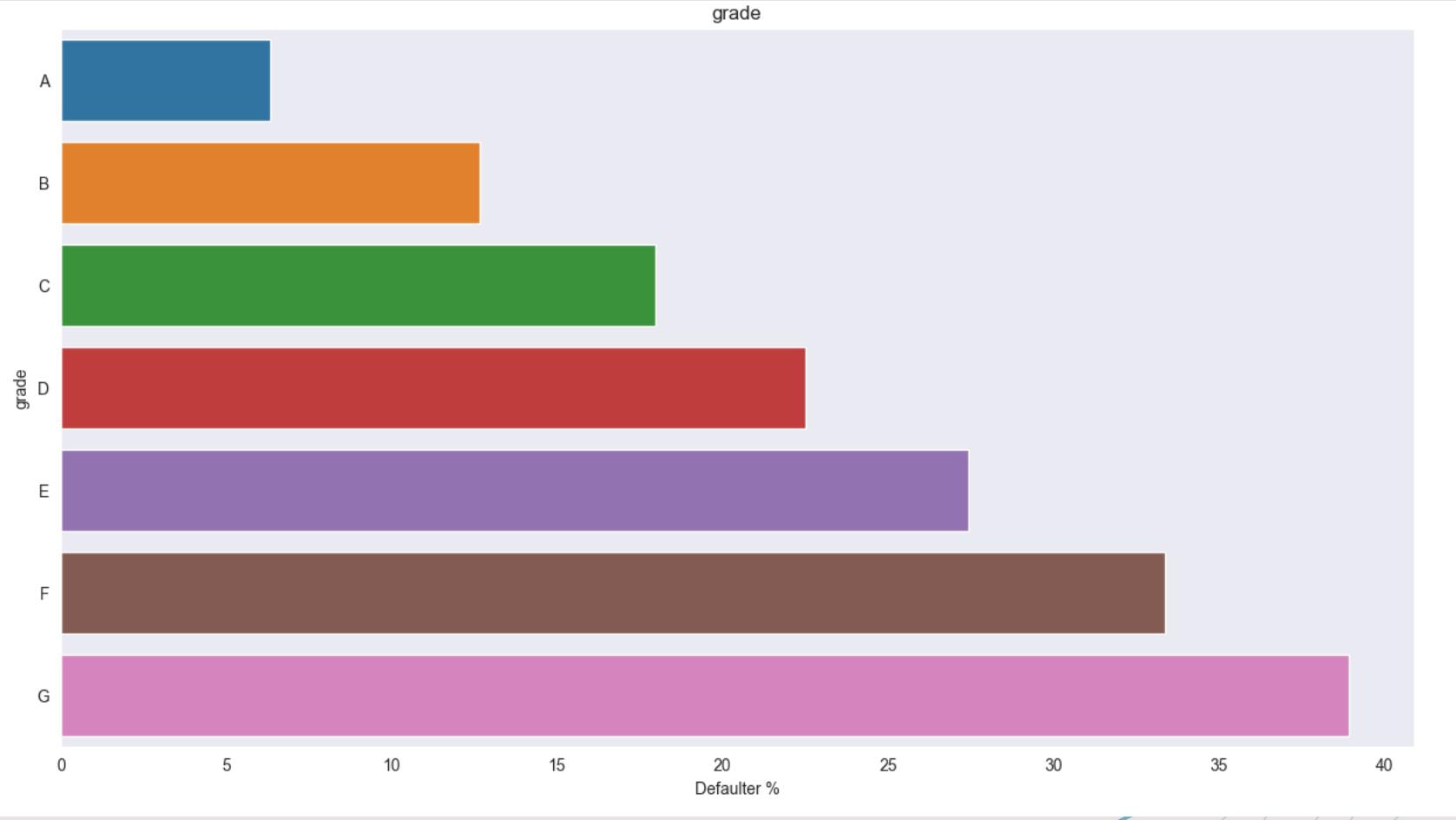
Bivariate Analysis (Defaulter vs Annual Inc)

- Borrowers earning 0 to 15000 annually has a highest rate of Defaulters
- Borrowers earning more than 75K+ are likely to be not defaulters



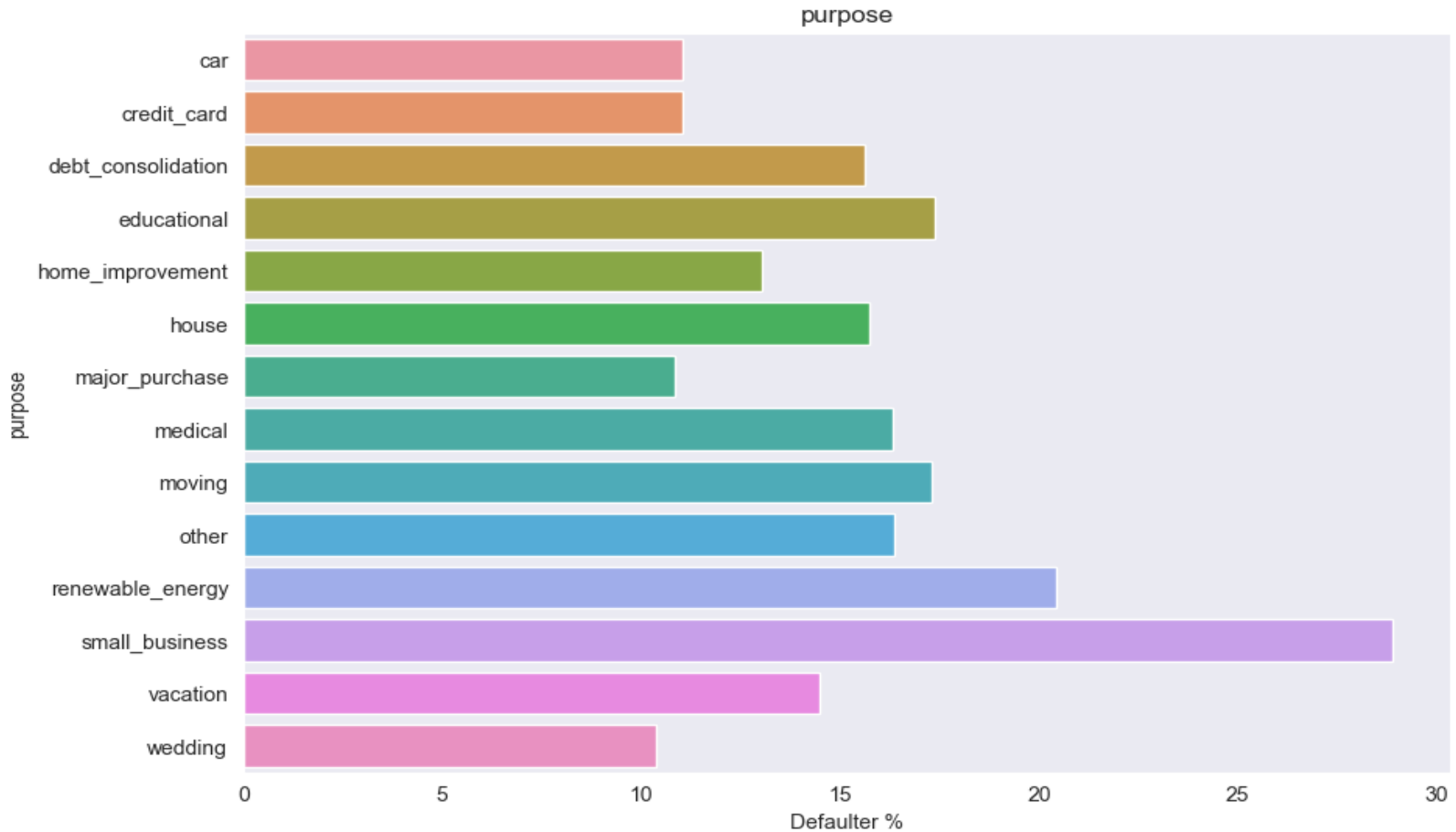
Conti.. Bivariate Analysis (Defaulter vs Grade)

Borrowers from G and F grade were found to be highest defaulters



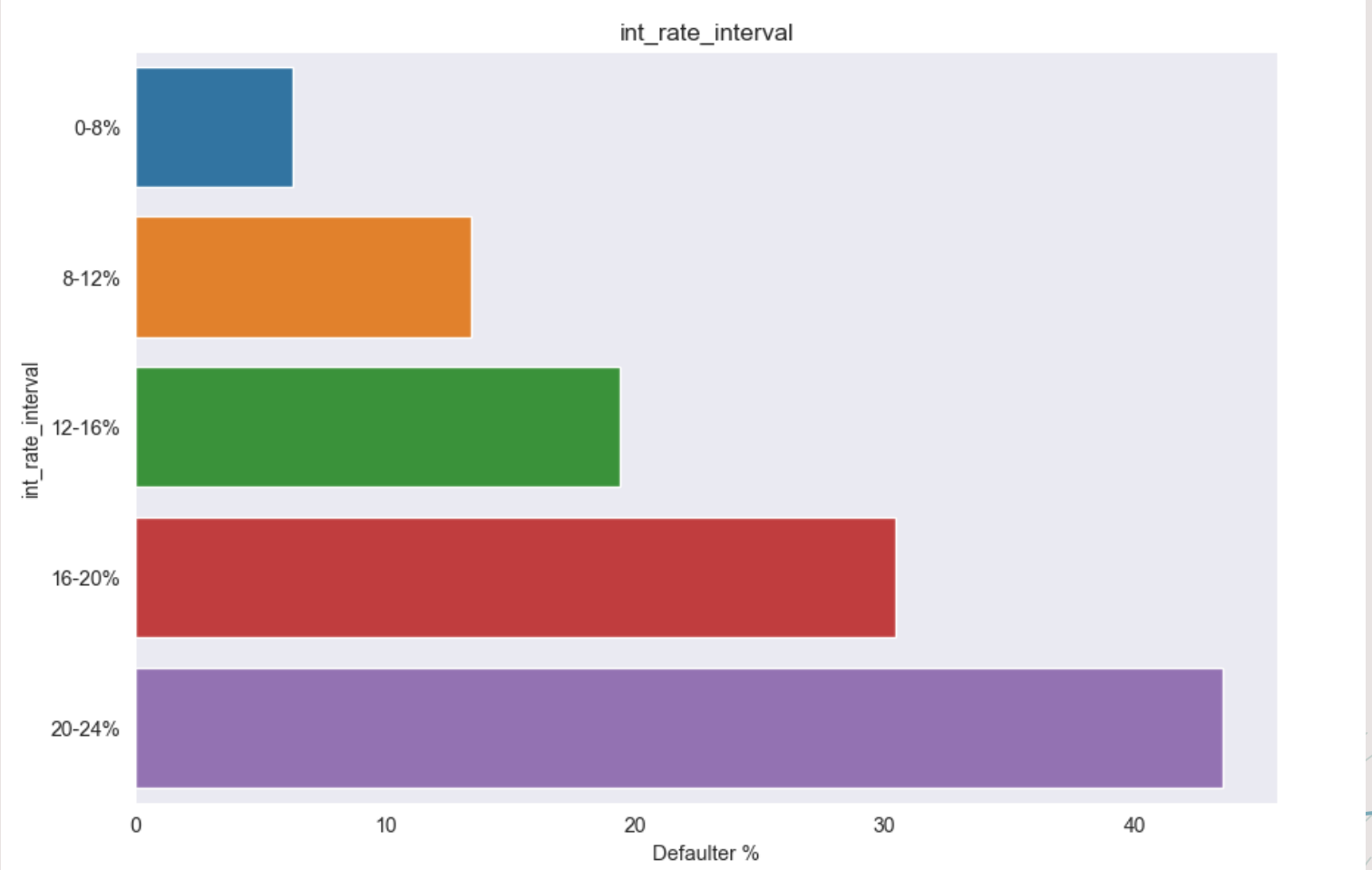
Conti.. Bivariate Analysis (Defaulter vs Purpose)

The purpose for "Small business" has high rate of defaulters compare to others



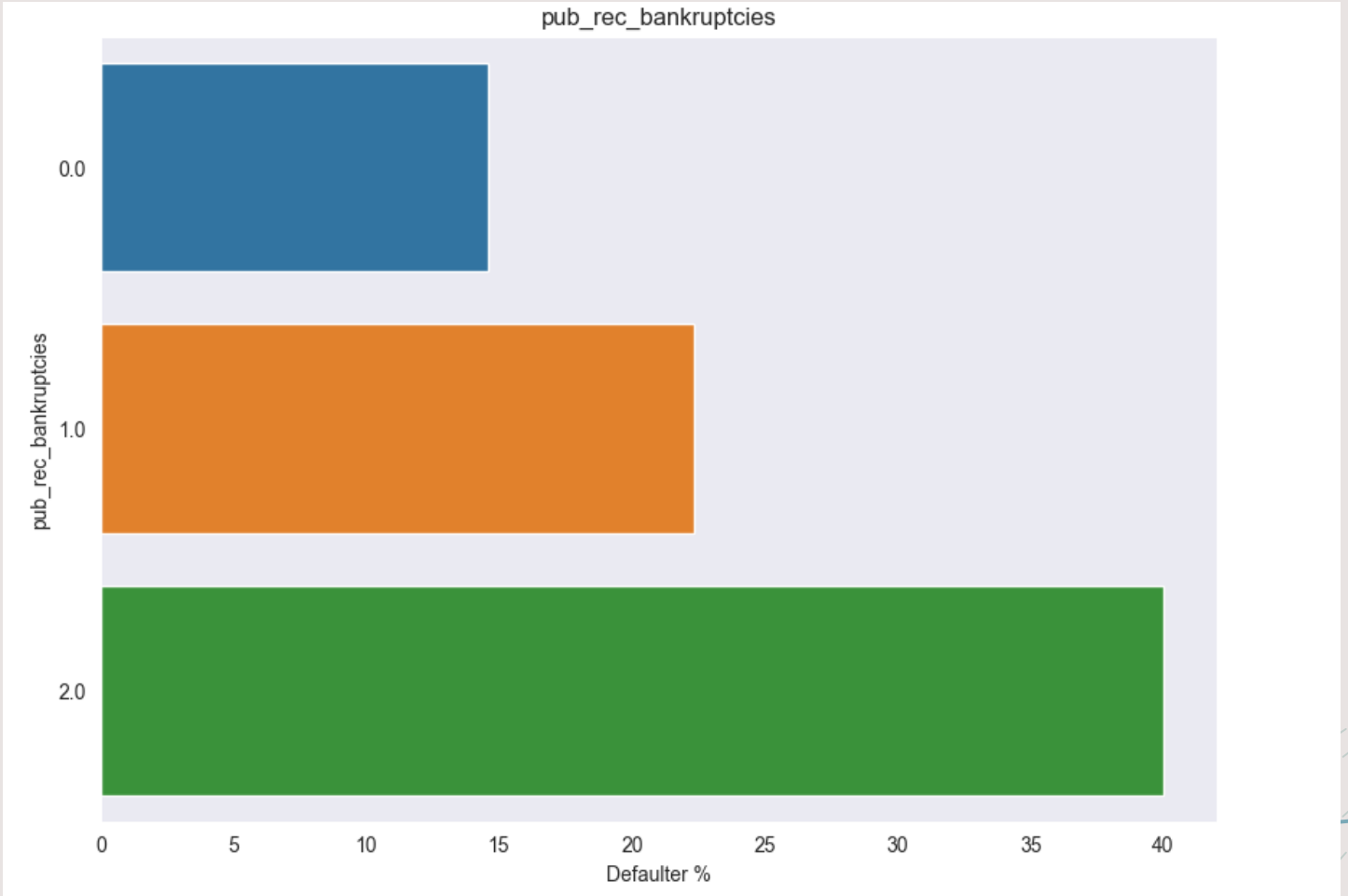
Conti.. Bivariate Analysis (Defaulter vs Interest Rate Range)

Interest rate more than 20% has the highest defaulters compare to others



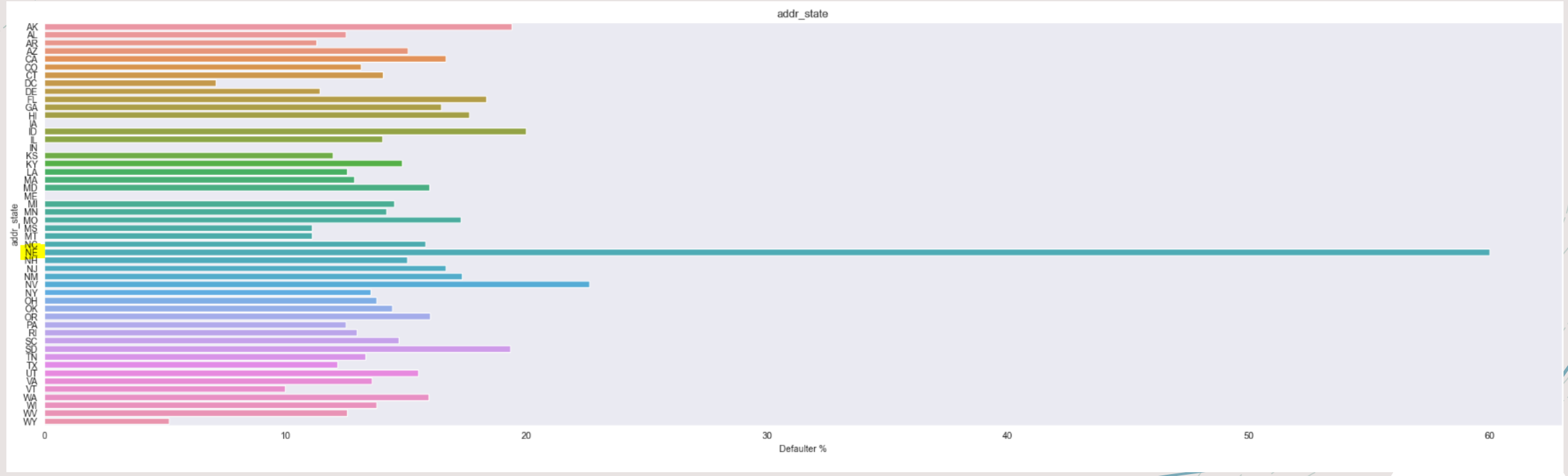
Conti.. Bivariate Analysis (Defaulter vs Bank rupts)

borrowers are defaulters who has been defaulted before and that has ratio of more than 40%



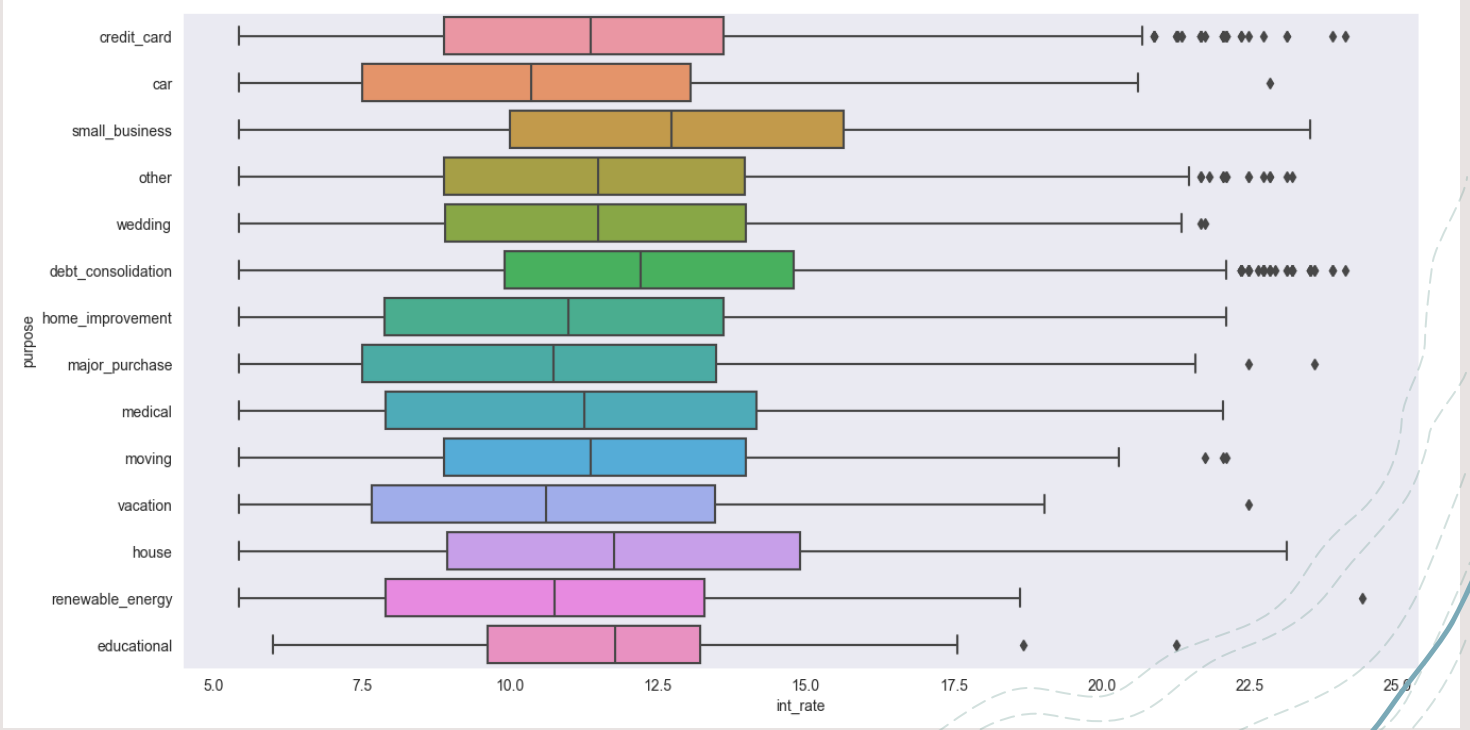
Conti.. Bivariate Analysis (Defaulter vs Address State)

More than 60% of borrowers were charged off in NE state



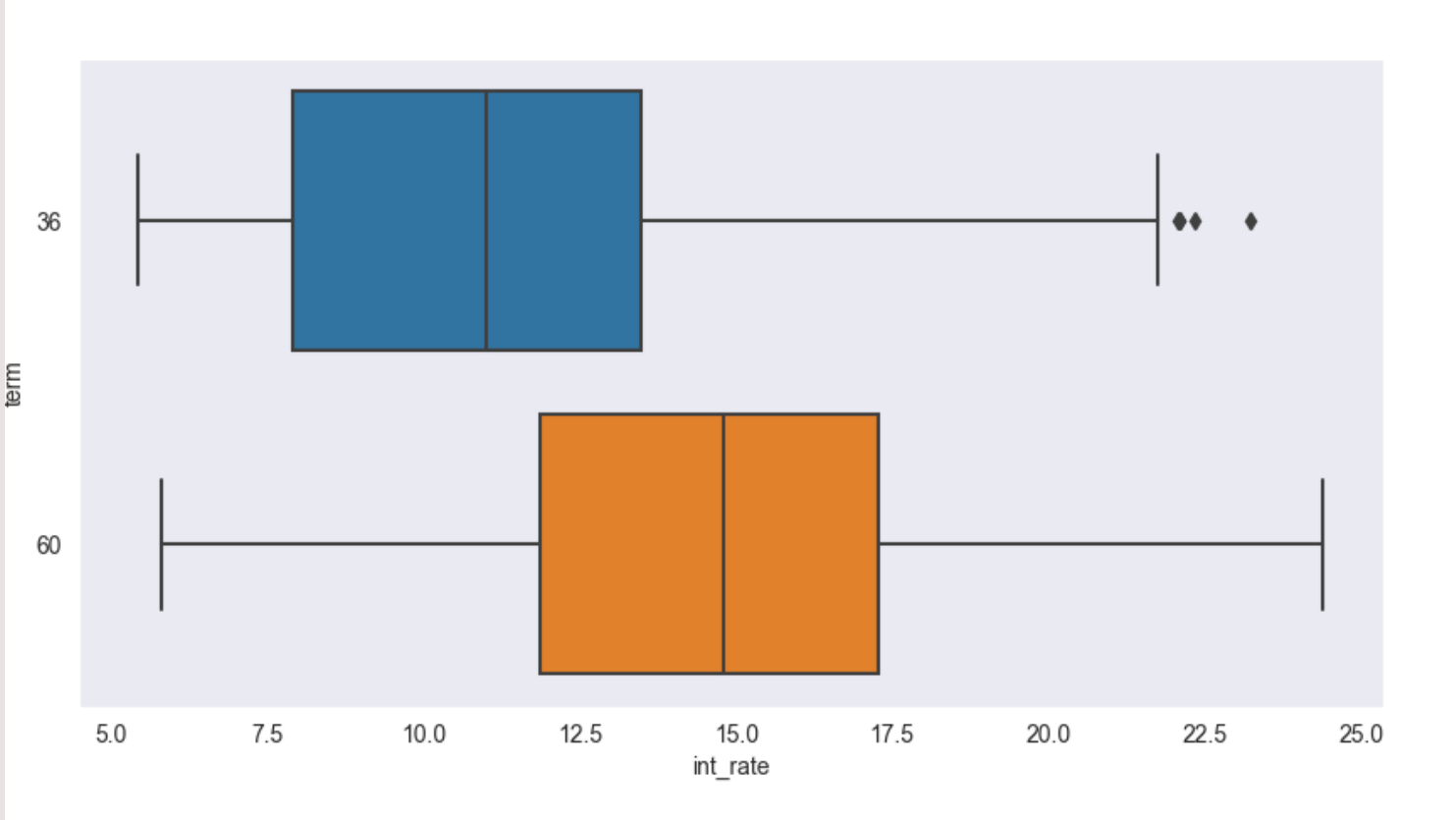
Bivariate Analysis – Box Plot (Purpose vs Interest Rate)

- The loan amount obtained for small business purpose has the highest median, 95th percentile, and 75th percentile values of any reason.
- whereas house seems to be 2nd and Credit card seems to be as 3rd line



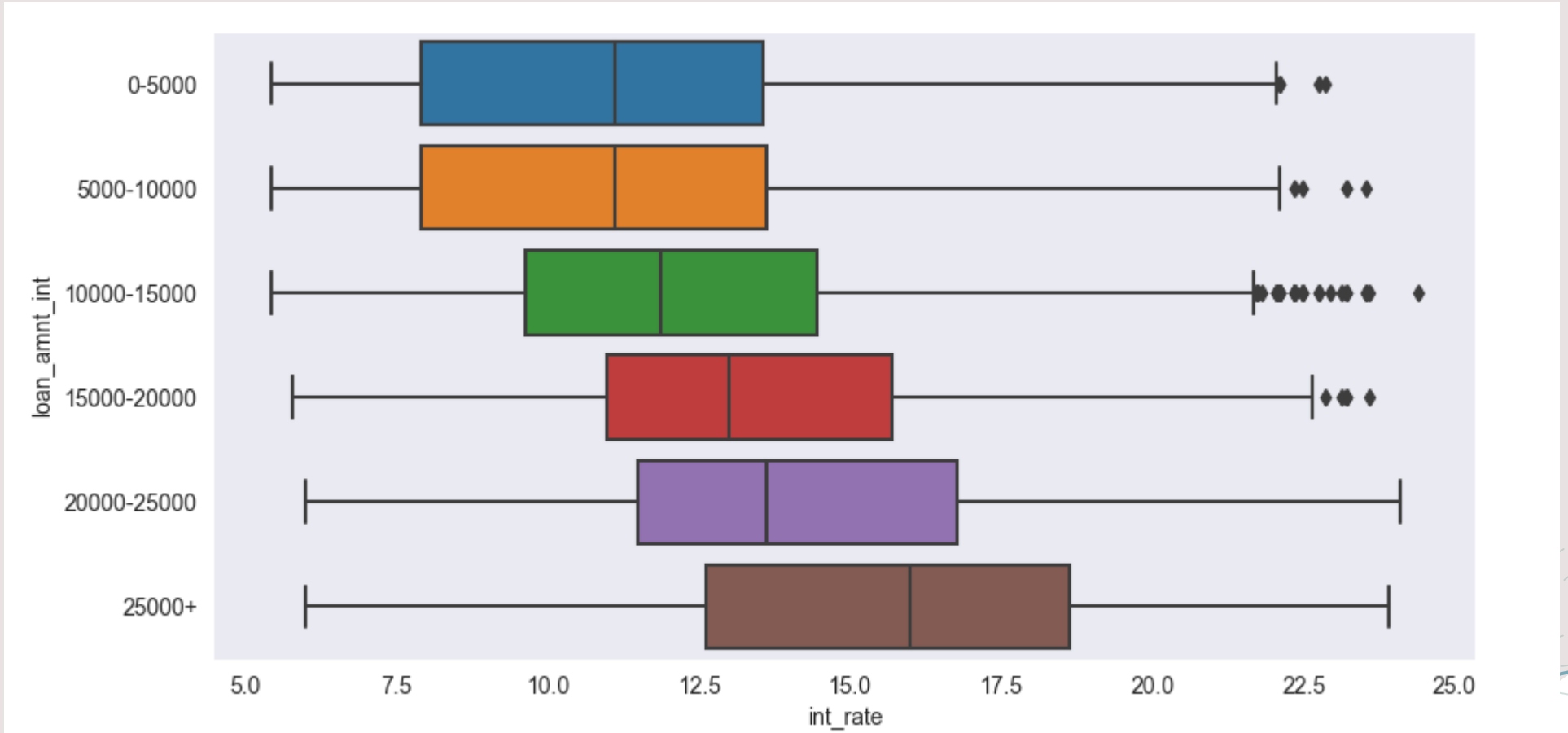
Conti..Bivariate Analysis – Box Plot (Due Term vs Interest Rate)

Higher the term duration higher the interest rate



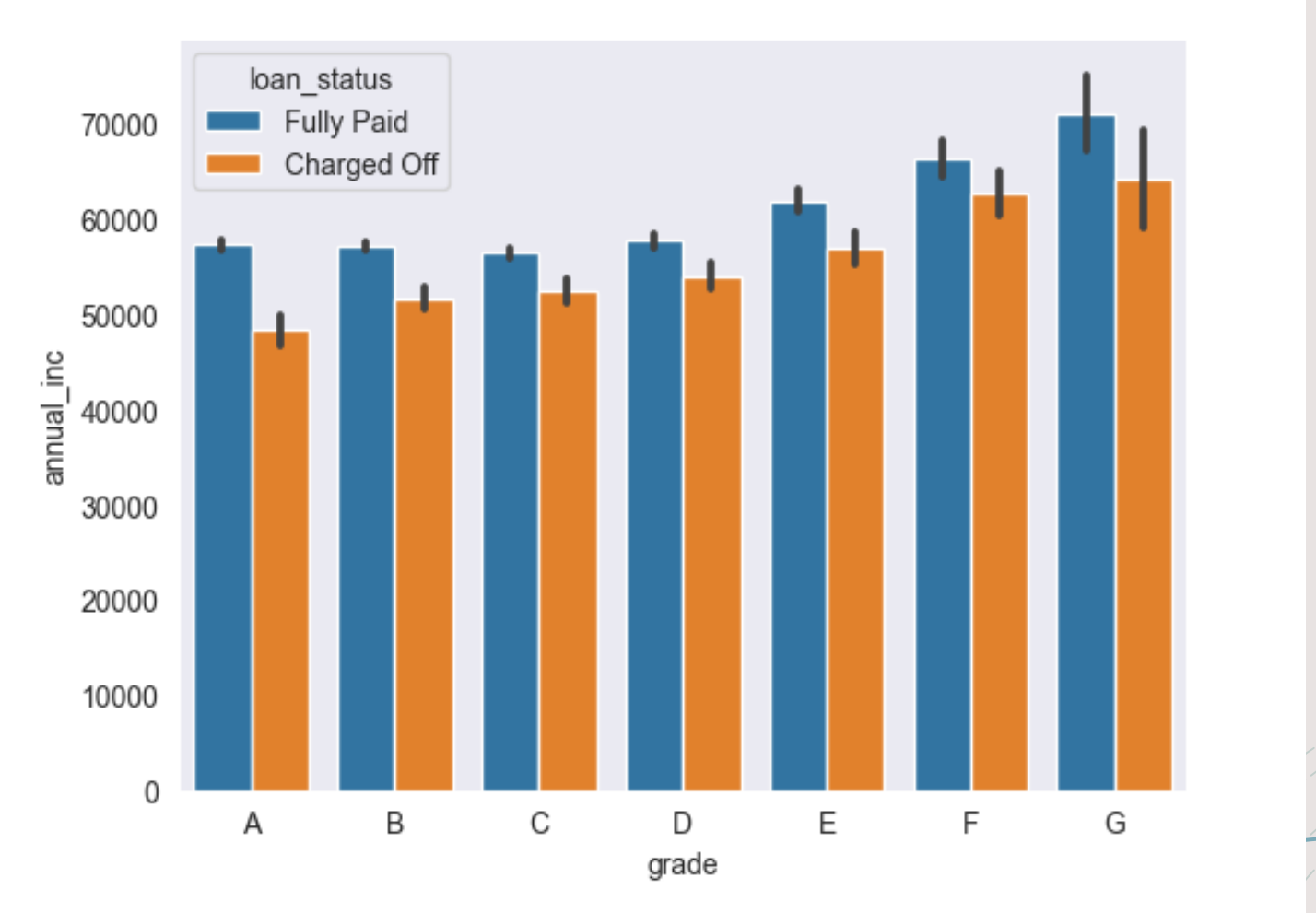
Conti.. Bivariate Analysis – Box Plot (Due Term vs Interest Rate)

As loan amount increases. Interest rate increase



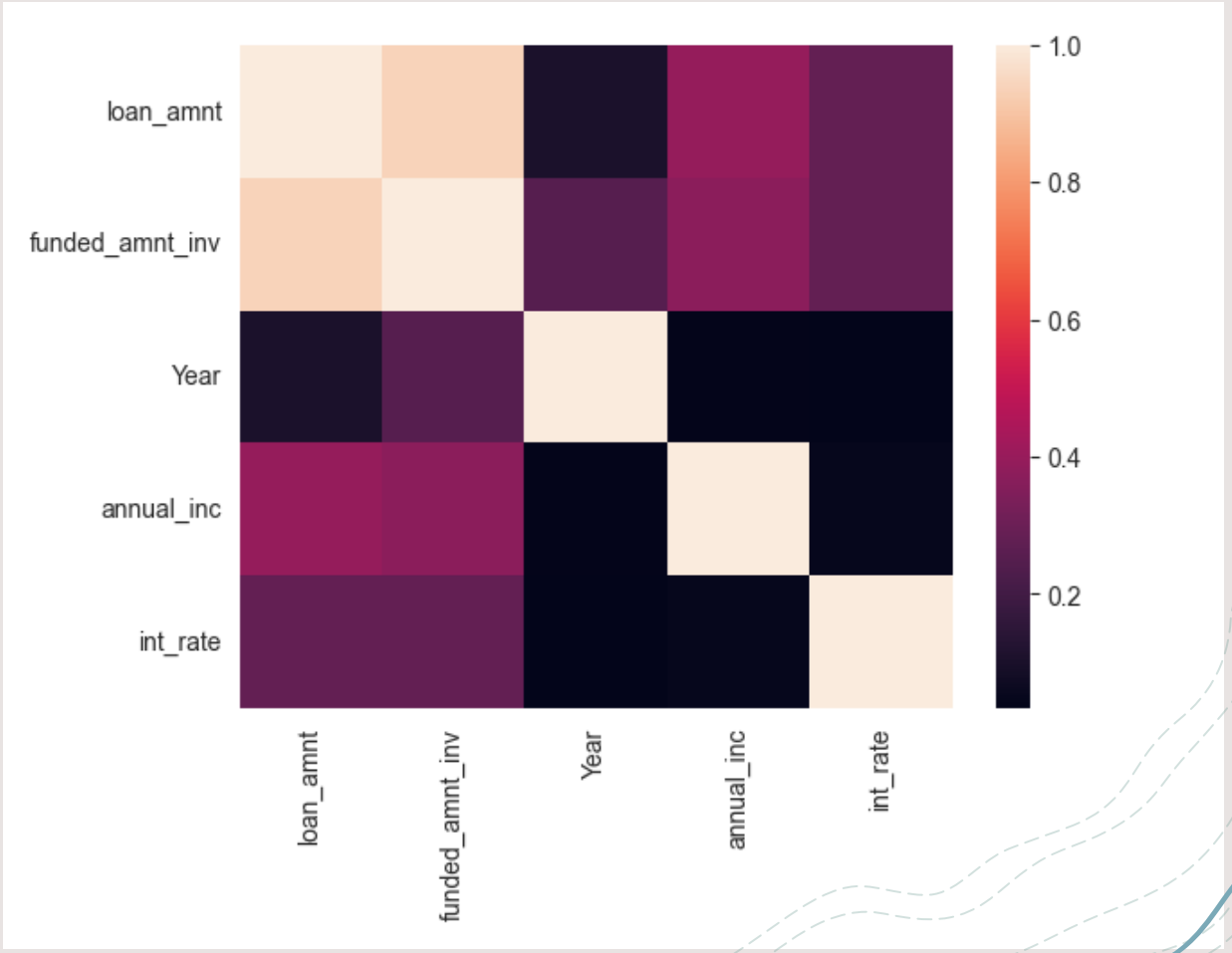
Conti.. Bivariate Analysis – Box Plot (Annual Income vs Grade)

Charged off borrowers earning less annual income than the other one each



Correlation between Continuous variables

The more annual income, the more chance of loan amount sanctioned



Business terms against loan status