Lending Club Case Study

Abhinav Anand

Data Understanding

What is Lending Club ?

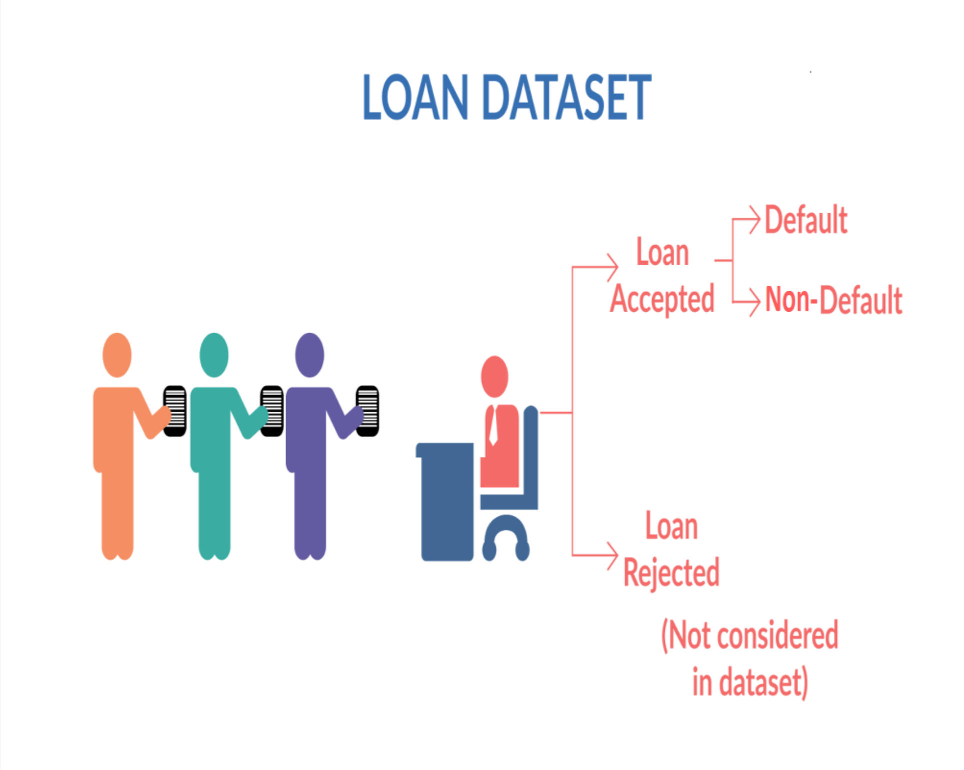
- It is a marketplace for lending various type of loans like home loans, personal loans, business loans etc that matches borrower who are seeking a loan with investors looking to lend money and make a return.
- Basically when an applicant apply for loans, the company has to make decision for loan approval based on applicant's Profile:
  - If Person likely to pay the loan then not approving will result to business loss to the company.
  - If Person not likely to pay the loan i.e. loan likely to get default then approving the loan will result to business loss to the company.

Business Problem

- Lending loans to any risky customer can leads to financial losses.
- 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 since borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters’.
- The Company need to find out what are driving factors for loan to getting defaulted.
- Help company to analyse the pattern in data and provide solution on how can they reduce loan defaulting hence cutting down credit loss. Identification of such applicants using EDA is the aim of this case study.

LOAN DATASET

- There are 39417 rows and 111 columns in the dataset.
- Loan status is the target column.
- It has following Categories:
- Fully paid: Applicant has fully paid the loan (the principal and the interest rate).
- 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'.
- 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.



Data Cleaning

Data Cleaning

- We have multiple column having missing value.
- Some of column having more than 30 percentage data missing so will be dropping them.
- For column having few missing value we use imputation.

id	0.000000
member_id	0.000000
loan_amnt	0.000000
funded_amnt	0.000000
funded_amnt_inv	0.000000
term	0.000000
int_rate	0.000000
installment	0.000000
grade	0.000000
sub_grade	0.000000
emp_title	6.191303
emp_length	2.706650
home_ownership	0.000000
annual_inc	0.000000
verification_status	0.000000
issue_d	0.000000
loan_status	0.000000
pymnt_plan	0.000000
url	0.000000

desc	32.580507
purpose	0.000000
title	0.027696
zip_code	0.000000
addr_state	0.000000
dti	0.000000
delinq_2yrs	0.000000
earliest_cr_line	0.000000
inq_last_6mths	0.000000
mths_since_last_delinq	64.662487
mths_since_last_record	92.985372
open_acc	0.000000
pub_rec	0.000000
revol_bal	0.000000
revol_util	0.125891
total_acc	0.000000
initial_list_status	0.000000
out_prncp	0.000000
out_prncp_inv	0.000000
total_pymnt	0.000000

total_pymnt_inv	0.000000
total_rec_prncp	0.000000
total_rec_int	0.000000
total_rec_late_fee	0.000000
recoveries	0.000000
collection_recovery_fee	0.000000
last_pymnt_d	0.178765
last_pymnt_amnt	0.000000
next_pymnt_d	97.129693
last_credit_pull_d	0.005036
collections_12_mths_ex_med	0.140998
mths_since_last_major_derog	100.000000
policy_code	0.000000
application_type	0.000000
annual_inc_joint	100.000000
dti_joint	100.000000
verification_status_joint	100.000000
acc_now_delinq	0.000000
tot_coll_amt	100.000000

tot_cur_bal	100.000000
open_acc_6m	100.000000
open_il_6m	100.000000
open_il_12m	100.000000
open_il_24m	100.000000
mths_since_rcnt_il	100.000000
total_bal_il	100.000000
il_util	100.000000
open_rv_12m	100.000000
open_rv_24m	100.000000
max_bal_bc	100.000000
all_util	100.000000
total_rev_hi_lim	100.000000
inq_fi	100.000000
total_cu_tl	100.000000
inq_last_12m	100.000000
acc_open_past_24mths	100.000000
avg_cur_bal	100.000000
bc_open_to_buy	100.000000

bc_util	100.000000
chargeoff_within_12_mths	0.140998
delinq_amnt	0.000000
mo_sin_old_il_acct	100.000000
mo_sin_old_rev_tl_op	100.000000
mo_sin_rcnt_rev_tl_op	100.000000
mo_sin_rcnt_tl	100.000000
mort_acc	100.000000
mths_since_recent_bc	100.000000
mths_since_recent_bc_dlq	100.000000
mths_since_recent_inq	100.000000
mths_since_recent_revol_delinq	100.000000
num_accts_ever_120_pd	100.000000
num_actv_bc_tl	100.000000
num_actv_rev_tl	100.000000
num_bc_sats	100.000000
num_bc_tl	100.000000
num_il_tl	100.000000
num_op_rev_tl	100.000000

num_rev_accts	100.000000
num_rev_tl_bal_gt_0	100.000000
num_sats	100.000000
num_tl_120dpd_2m	100.000000
num_tl_30dpd	100.000000
num_tl_90g_dpd_24m	100.000000
num_tl_op_past_12m	100.000000
pct_tl_nvr_dlq	100.000000
percent_bc_gt_75	100.000000
pub_rec_bankruptcies	1.754916
tax_liens	0.098195
tot_hi_cred_lim	100.000000
total_bal_ex_mort	100.000000
total_bc_limit	100.000000
total_il_high_credit_limit	100.000000
dtype: float64	

Data Cleaning - Dropping Column with null values

- After Dropping all the columns having more than 30 percent null value we left with 53 columns.
- Below the list of columns that present in the dataset.

```
class 'pandas.core.frame.DataFrame'>
angeIndex: 39717 entries, 0 to 39716
ata columns (total 53 columns):
#    Column                          Non-Null Count   Dtype
--   ------                          --------------   -----
0    id                              39717 non-null   int64
1    member_id                       39717 non-null   int64
2    loan_amnt                       39717 non-null   int64
3    funded_amnt                     39717 non-null   int64
4    funded_amnt_inv                 39717 non-null   float64
5    term                            39717 non-null   object
6    int_rate                        39717 non-null   object
7    installment                     39717 non-null   float64
8    grade                           39717 non-null   object
9    sub_grade                       39717 non-null   object
10   emp_title                       37258 non-null   object
11   emp_length                      38642 non-null   object
12   home_ownership                  39717 non-null   object
13   annual_inc                      39717 non-null   float64
14   verification_status             39717 non-null   object
15   issue_d                         39717 non-null   object
16   loan_status                     39717 non-null   object
17   pymnt_plan                      39717 non-null   object
18   url                             39717 non-null   object
19   purpose                         39717 non-null   object
20   title                           39706 non-null   object
21   zip_code                        39717 non-null   object
22   addr_state                      39717 non-null   object
23   dti                             39717 non-null   float64
24   delinq_2yrs                     39717 non-null   int64
25   earliest_cr_line                39717 non-null   object
```

```
26   inq_last_6mths                  39717 non-null   int64
27   open_acc                        39717 non-null   int64
28   pub_rec                         39717 non-null   int64
29   revol_bal                       39717 non-null   int64
30   revol_util                      39667 non-null   object
31   total_acc                       39717 non-null   int64
32   initial_list_status             39717 non-null   object
33   out_prncp                       39717 non-null   float64
34   out_prncp_inv                   39717 non-null   float64
35   total_pymnt                     39717 non-null   float64
36   total_pymnt_inv                 39717 non-null   float64
37   total_rec_prncp                 39717 non-null   float64
38   total_rec_int                   39717 non-null   float64
39   total_rec_late_fee              39717 non-null   float64
40   recoveries                      39717 non-null   float64
41   collection_recovery_fee         39717 non-null   float64
42   last_pymnt_d                    39646 non-null   object
43   last_pymnt_amnt                 39717 non-null   float64
44   last_credit_pull_d              39715 non-null   object
45   collections_12_mths_ex_med      39661 non-null   float64
46   policy_code                     39717 non-null   int64
47   application_type                39717 non-null   object
48   acc_now_delinq                  39717 non-null   int64
49   chargeoff_within_12_mths        39661 non-null   float64
50   delinq_amnt                     39717 non-null   int64
51   pub_rec_bankruptcies            39020 non-null   float64
52   tax_liens                       39678 non-null   float64
```

Data Cleaning – Drop unnecessary columns

- There was some columns having only one value, so these are column which is not significant and does not impact the target variable. We have dropped those column.

```
for columns unique values are in  term column
2
****************************************************************************
for columns unique values are in  int_rate column
371
****************************************************************************
for columns unique values are in  grade column
7
****************************************************************************
for columns unique values are in  sub_grade column
35
****************************************************************************
for columns unique values are in  emp_title column
28820
****************************************************************************
for columns unique values are in  emp_length column
11
****************************************************************************
```

```
****************************************************************************
for columns unique values are in  home_ownership column
5
****************************************************************************
for columns unique values are in  verification_status column
3
****************************************************************************
for columns unique values are in  issue_d column
55
****************************************************************************
for columns unique values are in  loan_status column
3
****************************************************************************
for columns unique values are in  url column
39717
****************************************************************************
for columns unique values are in  purpose column
14
****************************************************************************
```

```
for columns unique values are in  purpose column
14
****************************************************************************
for columns unique values are in  title column
19615
****************************************************************************
for columns unique values are in  zip_code column
823
****************************************************************************
for columns unique values are in  addr_state column
50
****************************************************************************
for columns unique values are in  earliest_cr_line column
526
****************************************************************************
for columns unique values are in  revol_util column
1089
****************************************************************************
```

```
****************************************************************************
for columns unique values are in  last_pymnt_d column
101
****************************************************************************
for columns unique values are in  last_credit_pull_d column
106
****************************************************************************
```

Data Cleaning – Drop unnecessary columns

- There are some Customer behaviour variables which are not available at time of application of loan so, there are not important column which is deciding the loan status.
- We have dropped those columns and after dropping them our dataset is having 21 columns

```
loan_amnt                   int64
funded_amnt_inv           float64
installment               float64
annual_inc                float64
dti                       float64
inq_last_6mths              int64
open_acc                    int64
pub_rec                     int64
total_acc                   int64
total_pymnt               float64
term                       object
int_rate                  float64
grade                      object
sub_grade                  object
emp_length                 object
home_ownership             object
verification_status        object
issue_d                    object
loan_status                object
```

```
]:   1  df_loan_new.shape

]:  (39717, 21)
```

```
purpose                    object
revol_util                float64
dtype: object
```

Data Cleaning – Missing Value Treatment

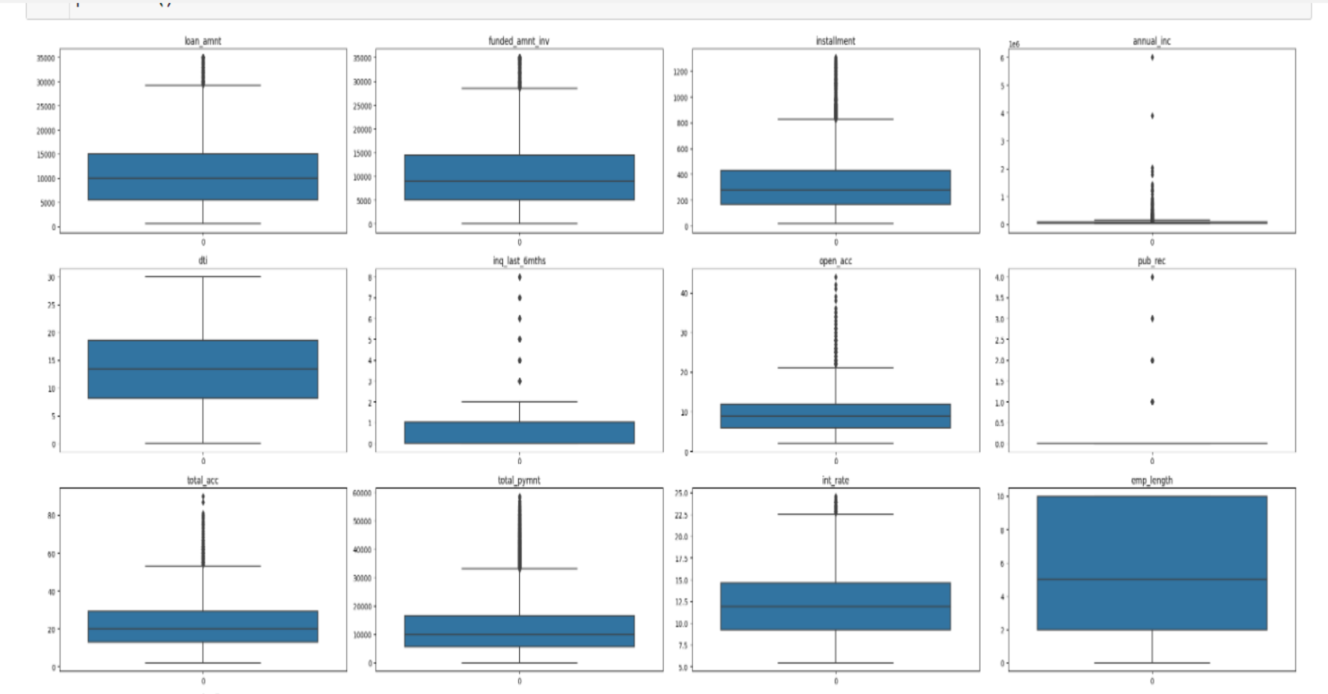
- There were few missing value present in dataset for two columns, so we imputed values for revol_util by mean since it was numerical column and there was no outliers and emp_length by mode since it was categorical column.

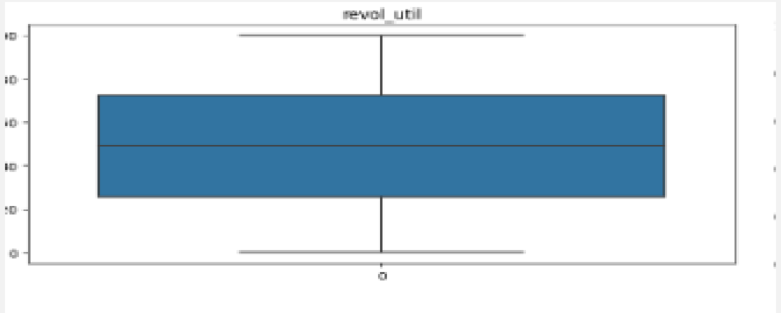
```
In [219]:  1  df_loan_new.isna().sum()

Out[219]: loan_amnt                0
          funded_amnt_inv          0
          installment              0
          annual_inc               0
          dti                      0
          inq_last_6mths           0
          open_acc                 0
          pub_rec                  0
          total_acc                0
          total_pymnt              0
          term                     0
          int_rate                 0
          grade                    0
          sub_grade                0
          emp_length            1075
          home_ownership           0
          verification_status      0
          issue_d                  0
          loan_status              0
          purpose                  0
          revol_util              50
          dtype: int64
```

```
:  1  df_loan_new.isna().sum()

:  loan_amnt              0
   funded_amnt_inv        0
   installment            0
   annual_inc             0
   dti                    0
   inq_last_6mths         0
   open_acc               0
   pub_rec                0
   total_acc              0
   total_pymnt            0
   term                   0
   int_rate               0
   grade                  0
   sub_grade              0
   emp_length             0
   home_ownership         0
   verification_status    0
   issue_d                0
   loan_status            0
   purpose                0
   revol_util             0
   dtype: int64
```

Data Cleaning – Outliers Detections and Treatments

- There were some columns where outliers were present, but we can see that much are value are close together.
- Instead of using IQR technique, we removed outliers based on percentile so that only the most extreme values removed from data set.
- After Removing outliers, we have 36758 rows in our final data set.






```
1 df_loan_new.shape
```

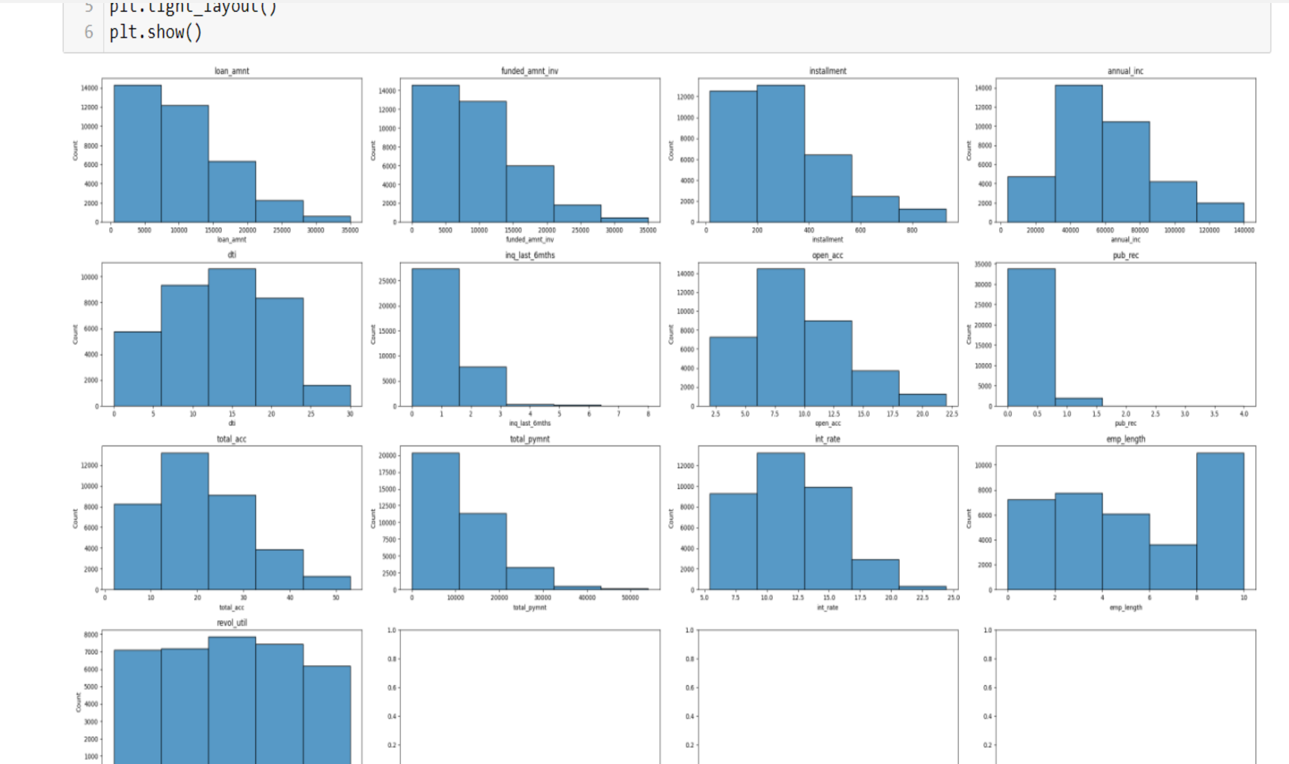
```
(36758, 21)
```

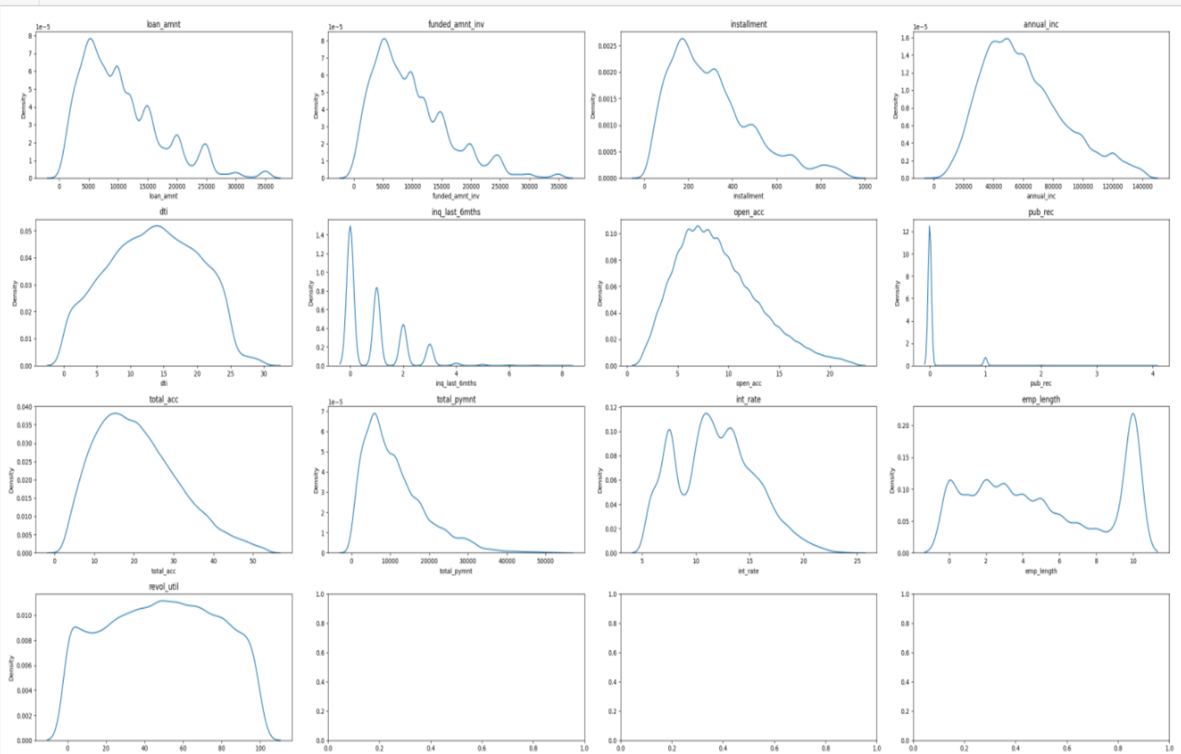
Data Analysis

Data Analysis – Univariate Analysis

For Numerical columns:

- We have plot histogram and distribution plot to check the spread of data.
- We Observed that most the data is right skewed because of presence of extreme values towards higher side.
- For loan amount and funded_amnt_inv most the loan is distributed around 5000.

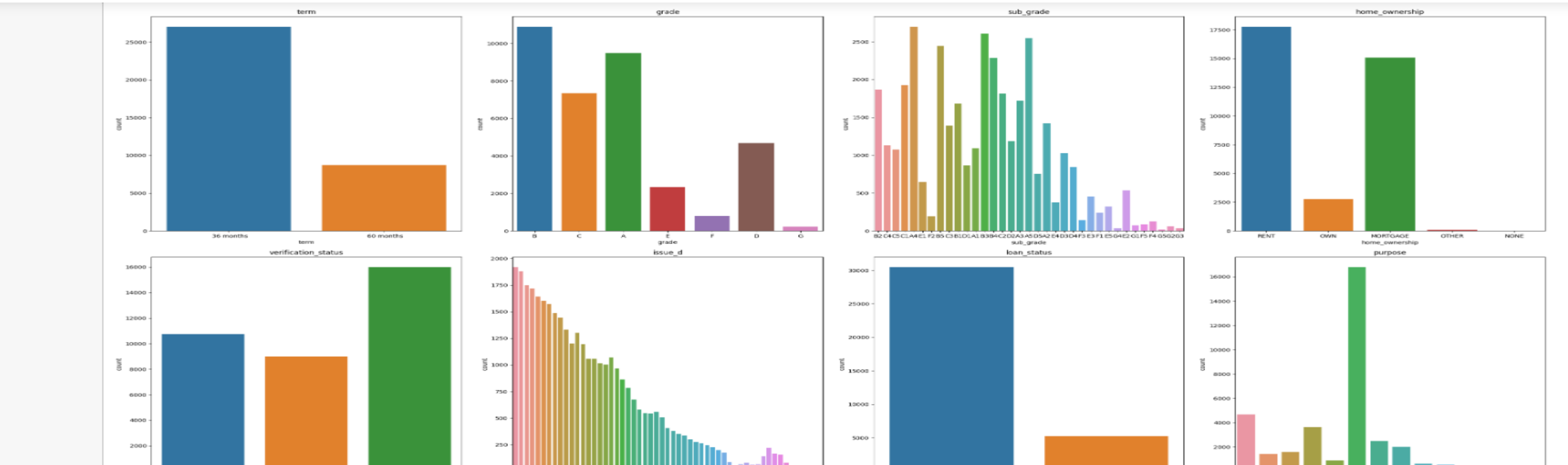




Data Analysis – Univariate Analysis

For Categorical Columns:

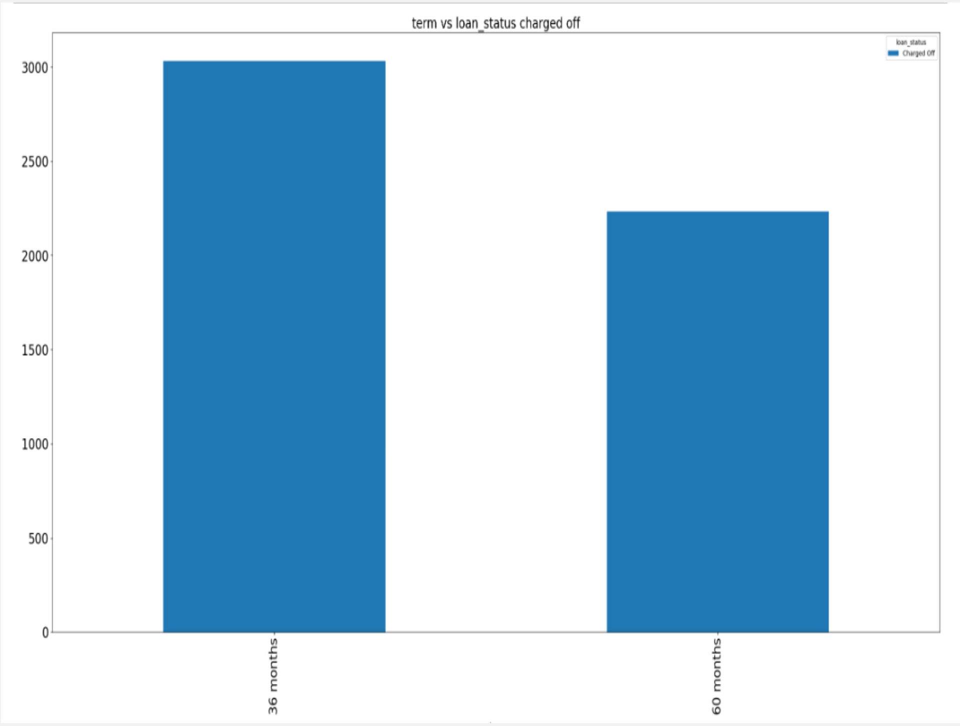
- We observed that there is huge data imbalance for loan status and term columns.
- Most of loans are fully paid and most of them are for 36 months period.
- Most of the loans has taken for debt_consolidation purpose.
- In home_ownership column we can see that None has only 3 values, so we decided to merge value with other.

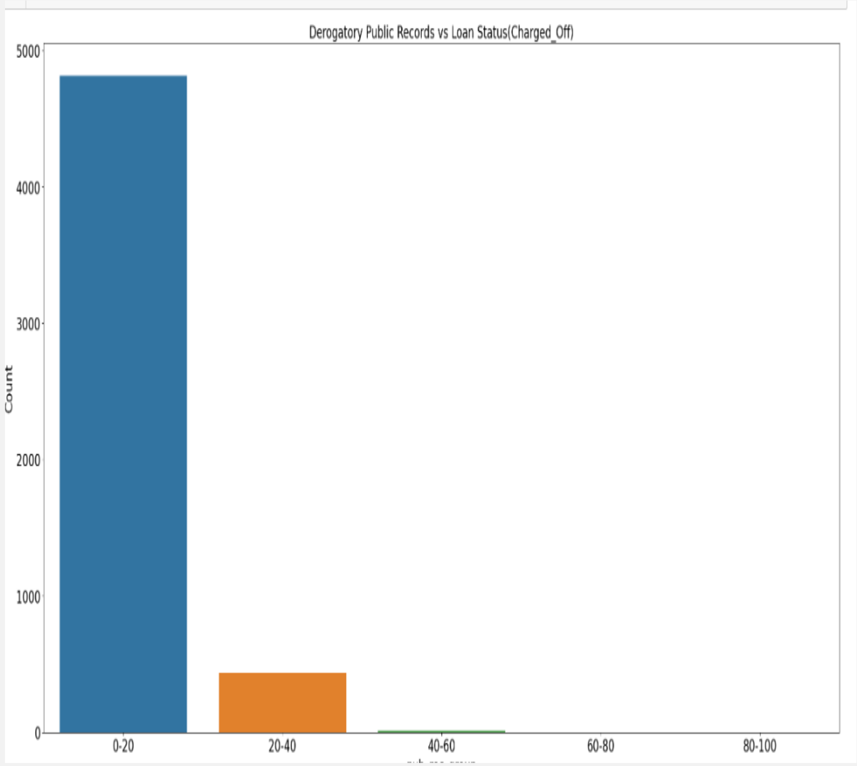


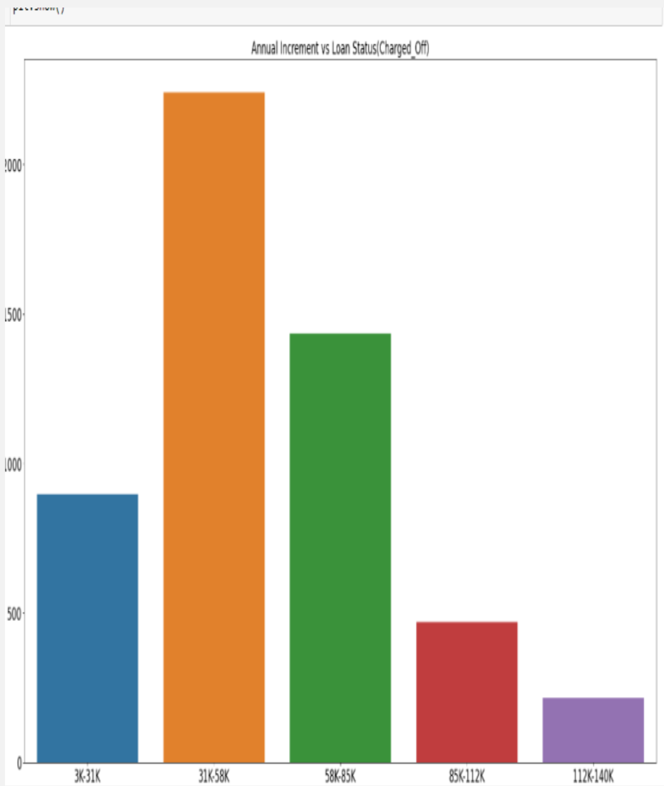
Data Analysis – Bivariate Analysis

We observed the following observation when loan are likely to get charge off:

- Loans amount which lend for 36 months.
- Derogatory Public Records between 0-20.
- Annual_inc_group between 31K-58K.



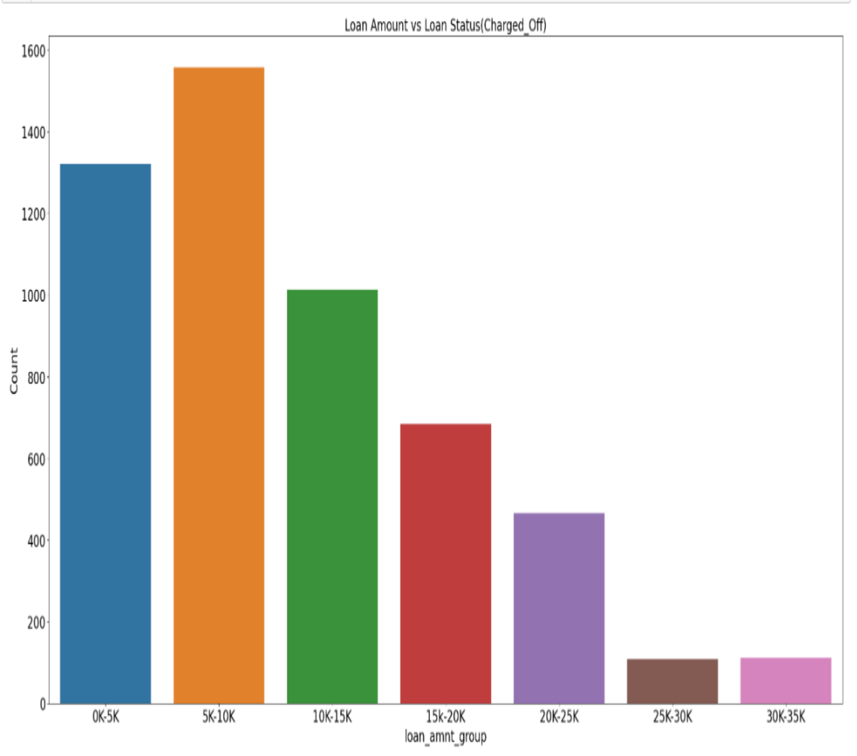


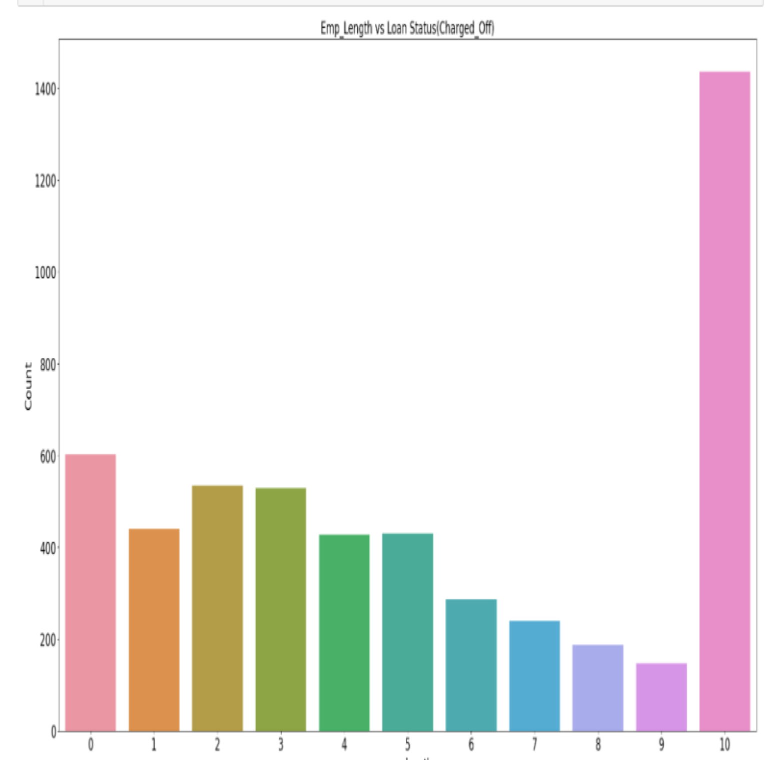


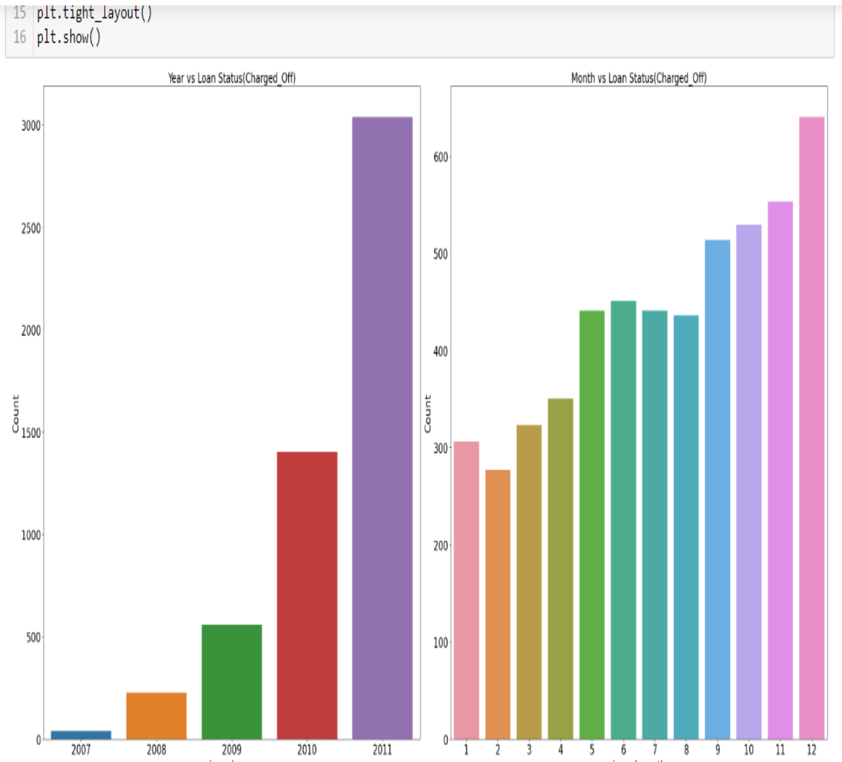
Data Analysis – Bivariate Analysis

We observed the following observation when loan are likely to get charge off:

- When loan amount is less than 10K.
- When Emp_length is 10.
- In 2011 maximum loan amount get defaulted may be because of financial crisis in USA.
- Loan issued in the month of December.



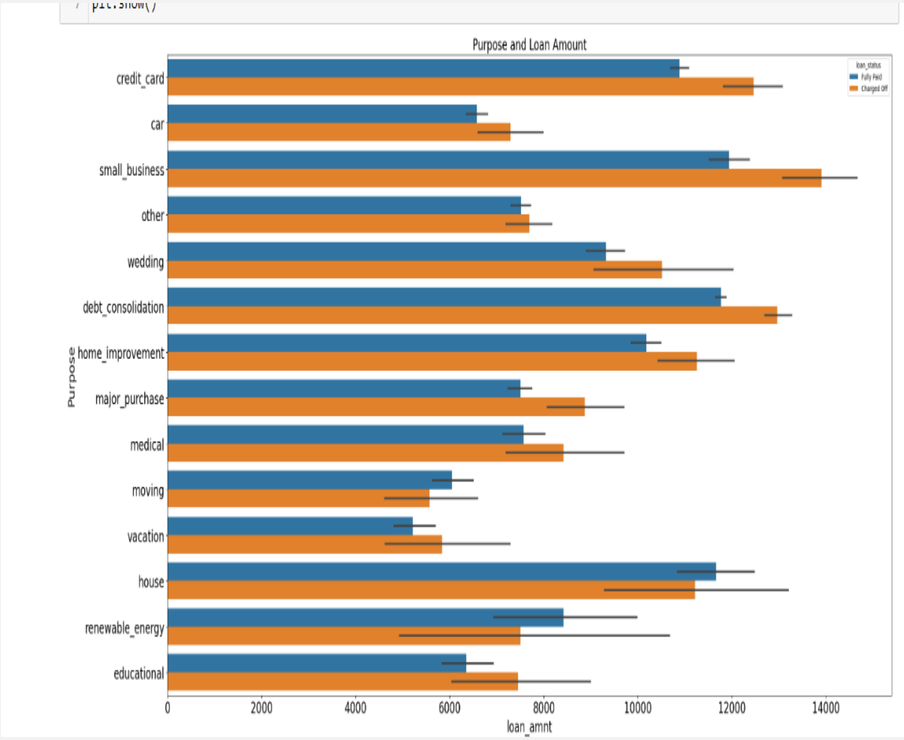


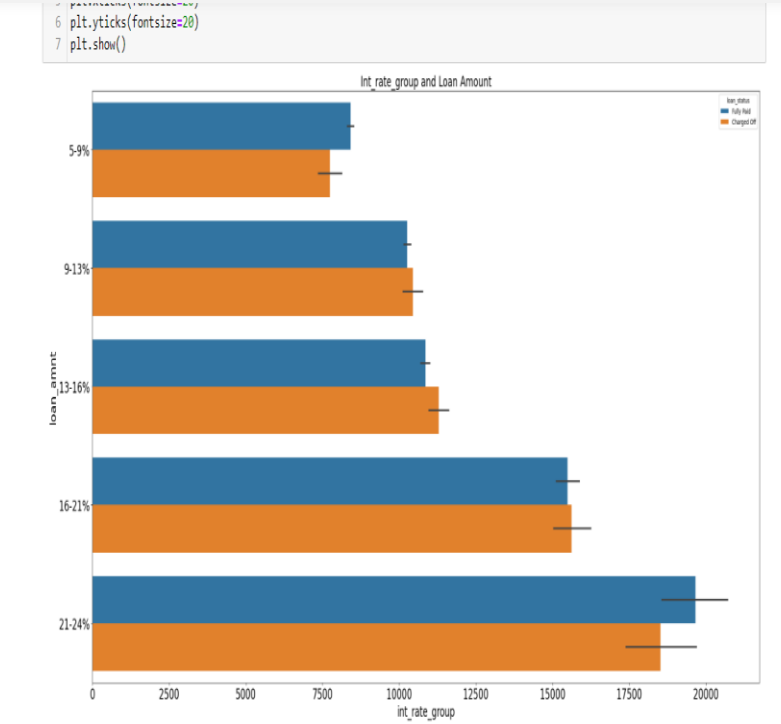


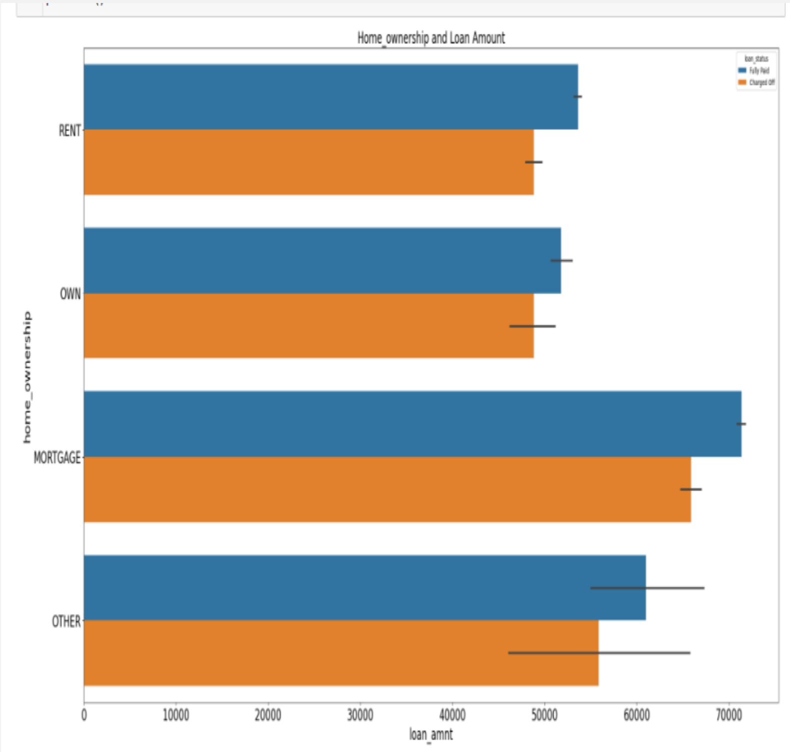
Data Analysis – Multivariate Analysis

We observed the following observation when loan are likely to get charge off:

- When loan amount more than 11K when loan lend for small business.
- There is no much difference for int_rate between 9-21% on loan_status but there is less chance for loan to get defaulted for 5-9% and 21-24% .
- There is less chance getting defaulted when loan amount is greater than 60000 for all home_ownership.

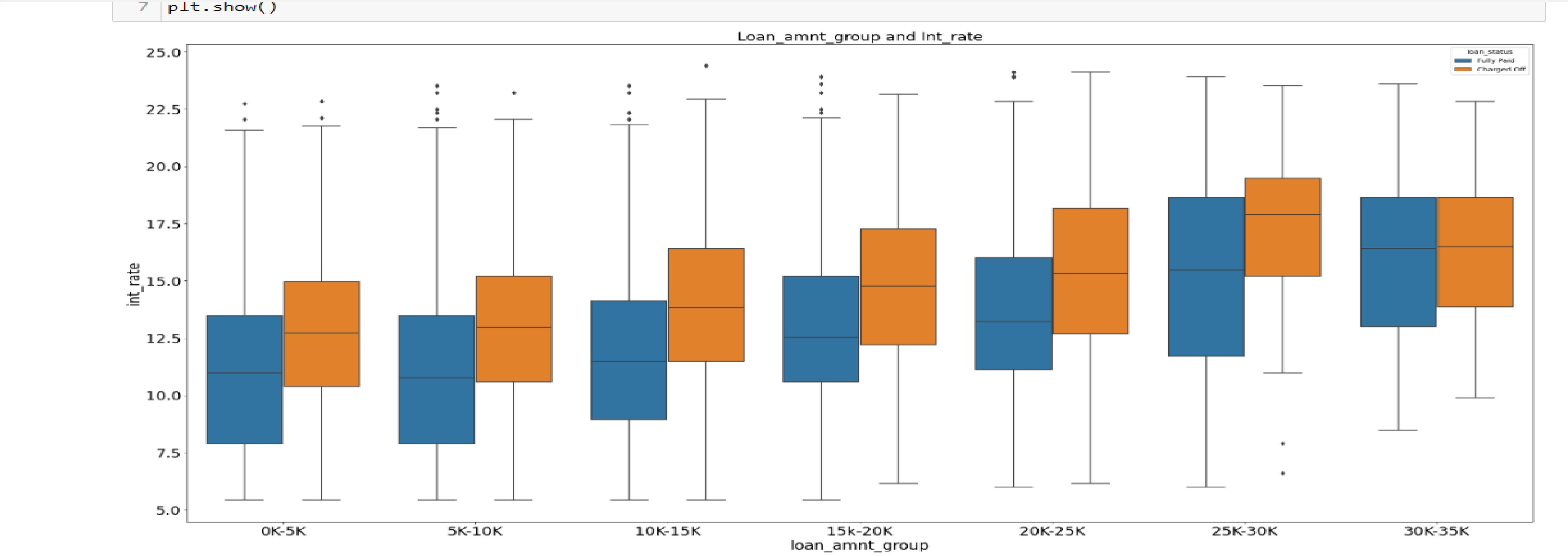






Data Analysis – Multivariate Analysis

- We observed that average of interest rate for the loan status having charged off is higher than loan status having fully paid.
- There some extreme value present for interest_rate for loan amount less than 30K.



Recommendations

Recommendations

Below is the scenario of loan likely to get defaulted :

- Loan taken for small business purpose.
- Loan issued for term that has 60 months slight chance but there is very less difference.
- Loan issued for verification_status inrespect the verification_status when amount is greater than 16K.
- For annual_inc_group inrespect the annual_inc_group.
- For int_rate_group when interest rate between 9%-21%.
- For annual_inc_group inrespect the revol_util_group.
- For annual_inc_group inrespect the installment_group.
- INT Rate Group having interest having between 13-16% likely to get charged off (defaulted)
- Open_acc_group having 6-10.
- Pub_rec_group having between 0-20.
- Annual_inc_group having between 31-58K.
- Revol_util_group having 60-80.
- Installment_Group having 107-199.
- Funded_Amnt_Inv_Group having 5K-10K but close for 0K-5K.
- Dti_group having 12-18.
- Inq_last_6mths having 0.
- Loan_amnt_group having 0-5K.
- Emp_length having 10.

THANK YOU