Lending Club Case Study - Credit Risk Analysis

Name: Umamaheswara Rao & Anand

Lending Club Case Study – Credit Risk Analysis

Business Problem - Business Understanding:

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

How Lending Club works?

1.Customers interested in a loan complete a simple application at LendingClub.com.

2. Lending Club evaluates each borrower's credit score using past historical data (and their data science process!) and assigns an interest rate to the borrower.

3. Qualified applicants receive loan offers in just minutes and can evaluate their options with no impact to their credit score.

4. Investors select the loans they want to invest in based on their own risk tolerance, investment portfolio goals, and time horizon

While doing application process by Bank's ,There are 2 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. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

To overcome these risks by Consumer Finance Company, Banker/ Credit risk analyst should analyze customer loan/lending historical data and generate the insights of customer behavior of loan repayment(whether customer paying ever month promptly without pending EMIs ,his financial commitments and sources of income etc., those insights details will help to banker’s to make correct decision of loan approval /reject.

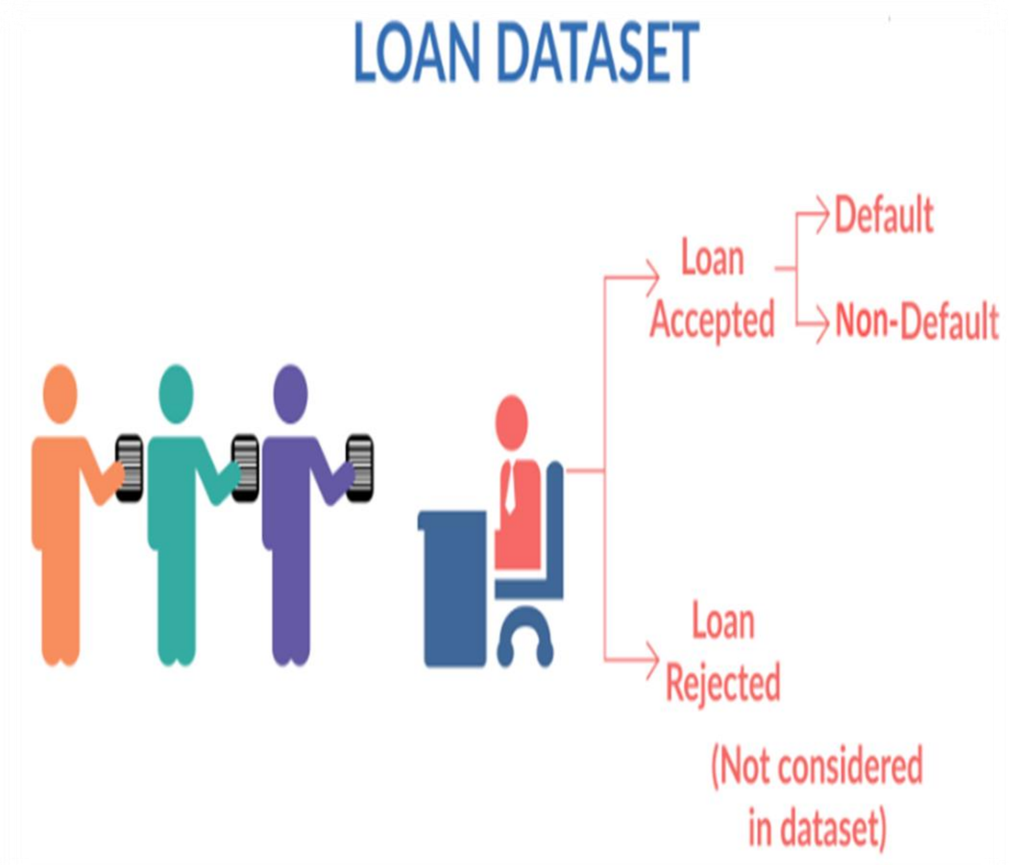
.

Lending Club Case Study – Credit Risk Analysis

Business Understanding

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:
  - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
  - Current: Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
  - Charged-off: Applicant has not paid the installments 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 etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company.



We should understand the key business attributes:

There are broadly three types of variables

1. those which are related to the applicant (demographic variables such as age, occupation, employment details etc.),
2. Loan characteristics (amount of loan, interest rate, purpose of loan etc.) and
3. Customer behavior variables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.).

Lending Club Case Study – Credit Risk Analysis

Business Objectives

- Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

- Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). 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. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labeled as '**charged-off'** are the **'defaulters'**.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using Exploratory Data Analysis.

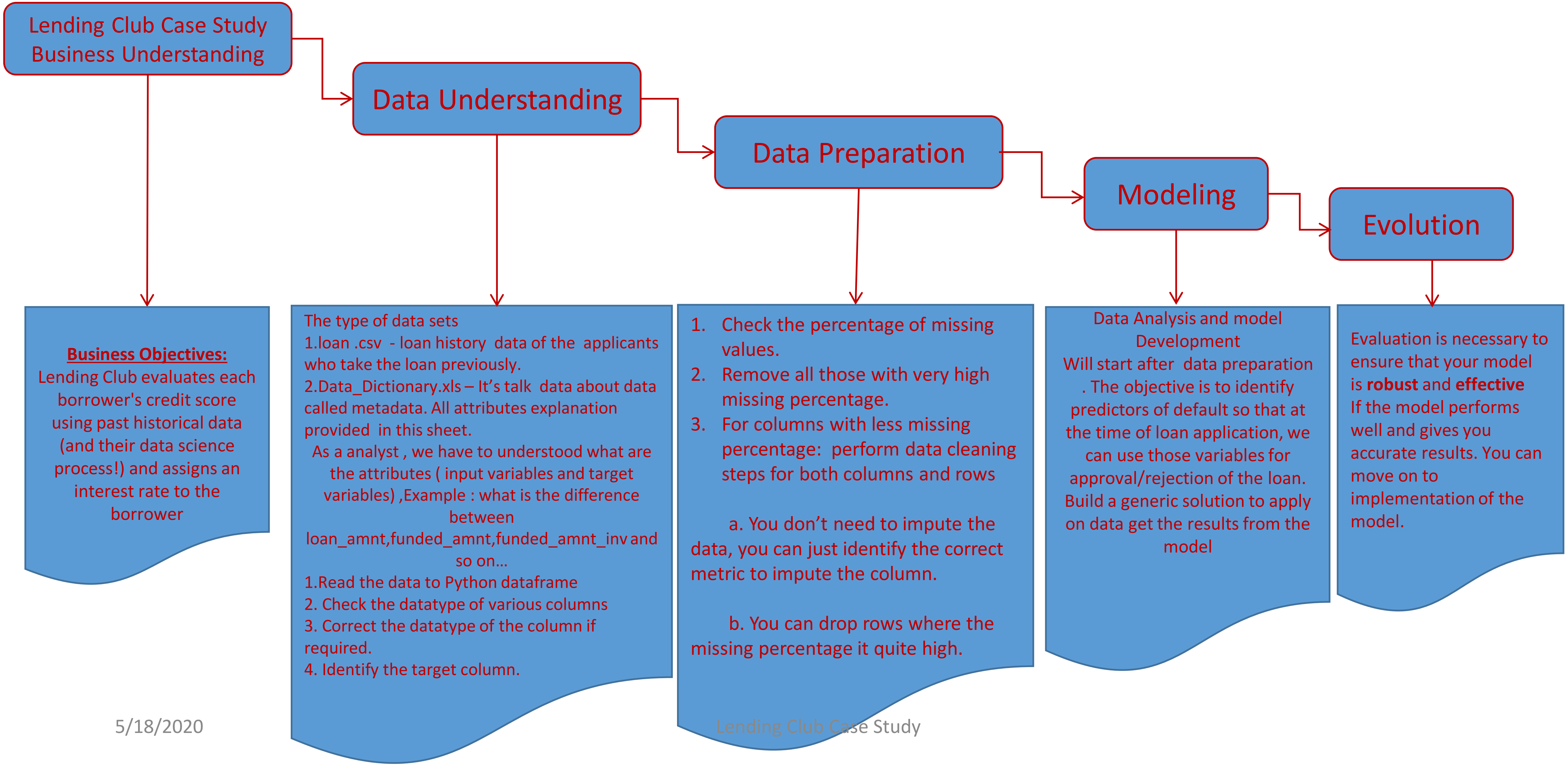
- In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

- To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

Lending Club Case Study – Credit Risk Analysis

Cross Industry Standard Process for Data Mining (CRISP–DM) framework.

Use flow chart



Lending Club Case Study Business Understanding

Data Understanding

Data Preparation

Modeling

Evolution

Business Objectives:
Lending Club evaluates each borrower's credit score using past historical data (and their data science process!) and assigns an interest rate to the borrower

The type of data sets
1.loan .csv - loan history data of the applicants who take the loan previously.
2.Data_Dictionary.xls – It's talk data about data called metadata. All attributes explanation provided in this sheet.
As a analyst , we have to understood what are the attributes (input variables and target variables) ,Example : what is the difference between loan_amnt,funded_amnt,funded_amnt_inv and so on…
1.Read the data to Python dataframe
2. Check the datatype of various columns
3. Correct the datatype of the column if required.
4. Identify the target column.

1. Check the percentage of missing values.
2. Remove all those with very high missing percentage.
3. For columns with less missing percentage: perform data cleaning steps for both columns and rows

a. You don't need to impute the data, you can just identify the correct metric to impute the column.

b. You can drop rows where the missing percentage it quite high.

Data Analysis and model Development
Will start after data preparation . The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan. Build a generic solution to apply on data get the results from the model

Evaluation is necessary to ensure that your model is **robust** and **effective**
If the model performs well and gives you accurate results. You can move on to implementation of the model.

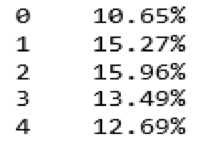
Lending Club Case Study – Data Understanding

Understand the Data Set and Data Cleancing

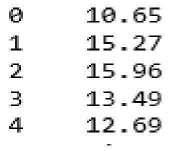
1.We have identified 54 columns out of 111 columns with 100 % null values and dropped those columns.

2. Corrected the Data types where ever applicable , example “int_rate” data format is alphanumeric like 10.65% , so we removed % sign and converted into float. And emp_lenght has values like “10+ years”, <1 year and so on .. We extracted only numeric value from that data and modified data type of the column and so on …

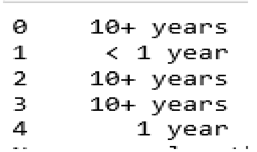
Before Cleaning **After Cleaning**



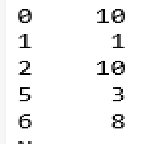
```
0      10.65%
1      15.27%
2      15.96%
3      13.49%
4      12.69%
```



```
0      10.65
1      15.27
2      15.96
3      13.49
4      12.69
```



```
0     10+ years
1      < 1 year
2     10+ years
3     10+ years
4        1 year
```



```
0    10
1     1
2    10
5     3
6     8
```

3. **These columns also we are not using for our analysis to find the defaulters**

#"mths_since_last_delinq (The number of months since the borrower's last delinquency.)-------64.66%

#" desc" : (Loan description provided by the borrower) ----32.5% and columns_not_used : ["title","id","member_id","zip_code","addr_state","url"] ['delinq_2yrs', 'earliest_cr_line', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d', 'last_pymnt_amnt', 'last_credit_pull_d']

So, we dropped those columns.

4. Below are the input variables for loan prediction :

Loan Status (target variable) and below are the input variables:

1.purpose 2.emp_length 3.Grade 4.int_rate 5.term 6.ChargeOff 7.loan_amount 8.funded_amunt 9.funded_amout_inv 10.int_rate 11.installment 12.anual_inc 13.dti 14.loan_income_ratio 15.issue_d 16.annual_inc The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan

Univariate Analysis:

We can see that fully paid comprises most of the loans (28952). The ones marked 'current' are not defaulted, let's tag the other two values as 0 for Fully Paid and 1 for Charged Off. we can exclude "Currnet" status records For univariate analysis, you have to check the default rate across various categorical features. For continuous features, you have to perform binning and then you may perform univariate analysis.

#Exclude the records where loan_status as 'Current' since these records are not get help our analysis since these records represents loan EMI's paying as usual (correctly)

Lending Club Case Study – Credit Risk Analysis

Data Analysis:

- The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan.

There are broadly three types of variables

1. Those which are related to the applicant (demographic variables such as age, occupation, employment details etc.),

2. Loan characteristics (amount of loan, interest rate, purpose of loan etc.) and

3. Customer behaviourvariables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.).

- Now, the customer behaviour variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.
- The ones marked 'current' are neither fully paid not defaulted, so get rid of the current loans. Also, tag the other two values as 0 or 1 to make your analysis simple and clean.

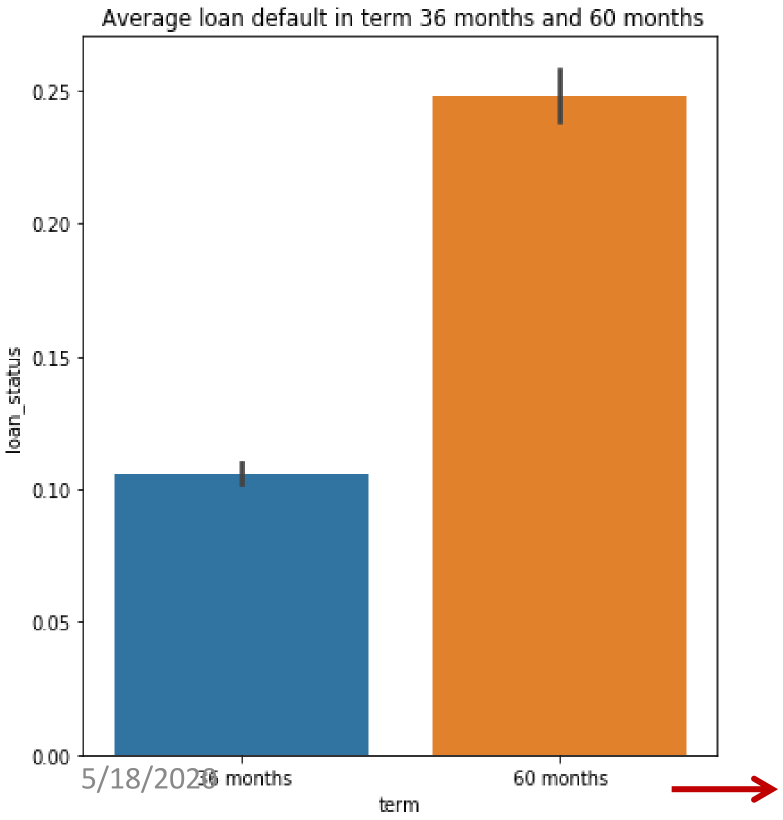
- Explain the results of univariate, bivariate analysis etc. in business terms

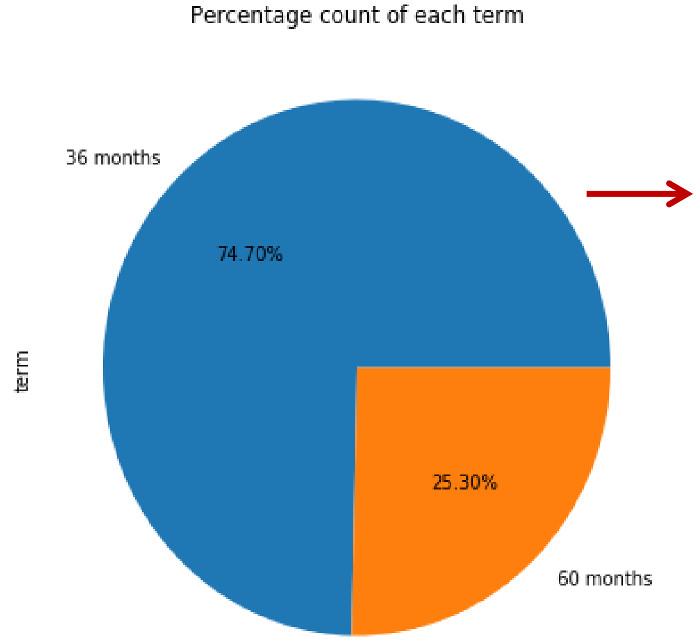
Lending Club Case Study – Detailed Analysis

Classify the variables into low frequency variables (categorical) and High frequency (Continues data variables).

Below are the categorical variables which we consider for our analysis
1.term,2.grade,3.purpose,4.sub_grade,5.loan_status,6.home_ownership,7.verification_status
We can see that fully paid comprises most of the loans (28952). The ones marked 'current' are not defaulted, let's tag the other two values as 0 for Fully Paid and 1 for Charged Off. we can exclude "Currnet" status records For univariate analysis, you have to check the default rate across various categorical features. For continuous features, you have to perform binning and then you may perform univariate analysis.

Categorical – Loan Term



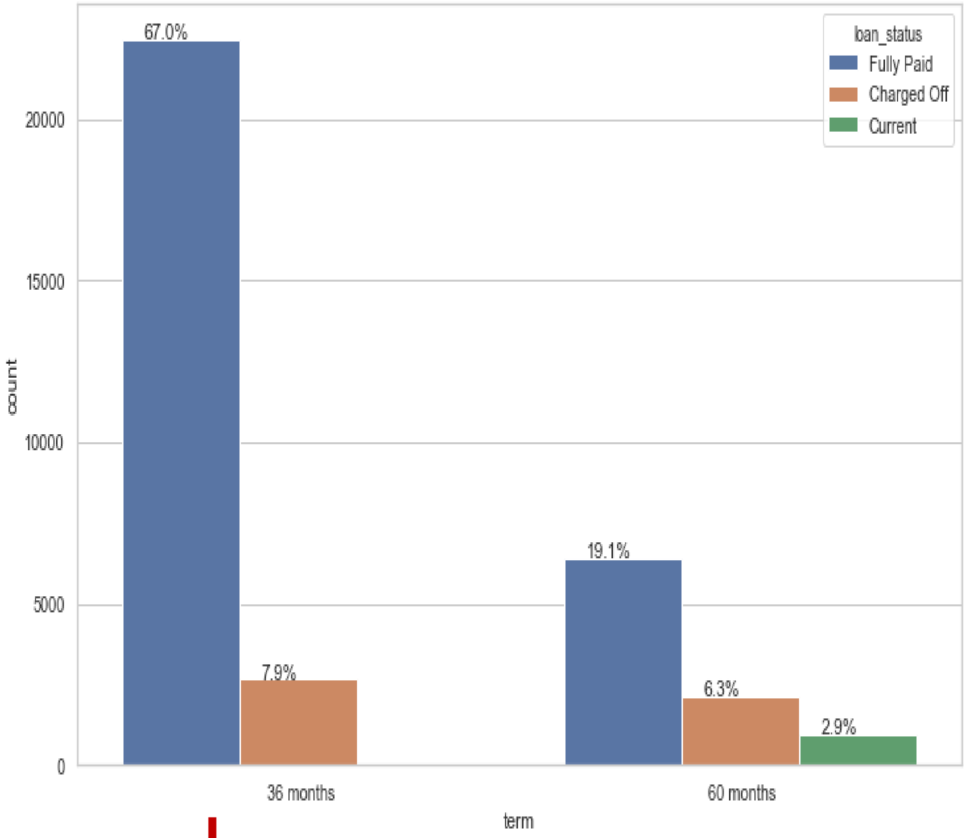


Insights:
5 Years (60 months) term loans distribution is 25.30 %
3 Years (36 months) term loans distribution is 74.70%
we should try to attract those customers towards 3 years with low interest rates to reduce defaulters

Insights: Loans with 5 years term (60 months) contribute more defaults as compared to 3 years (36-months) as per the bar chart .

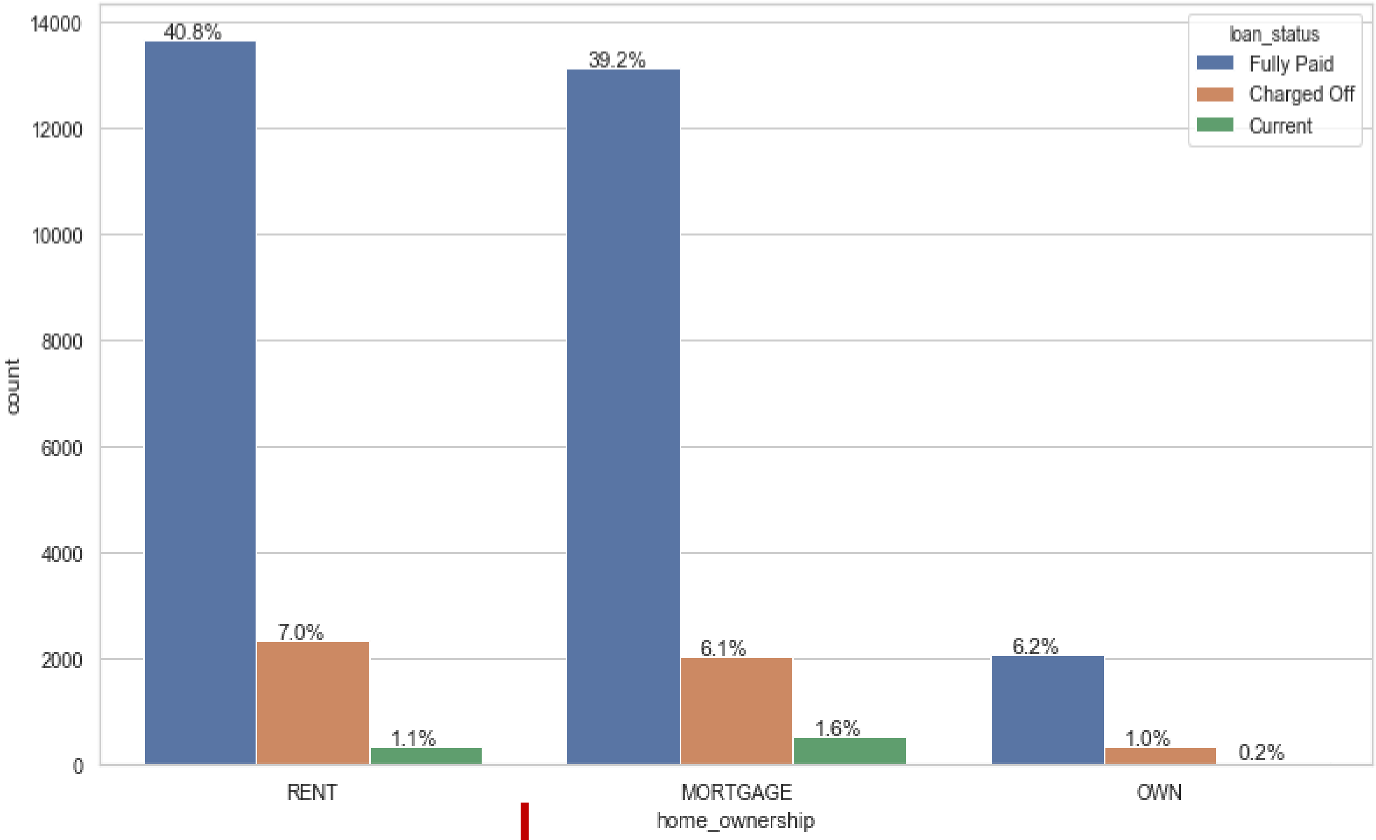
Lending Club Case Study – Detailed Analysis

Categorical -loan Term



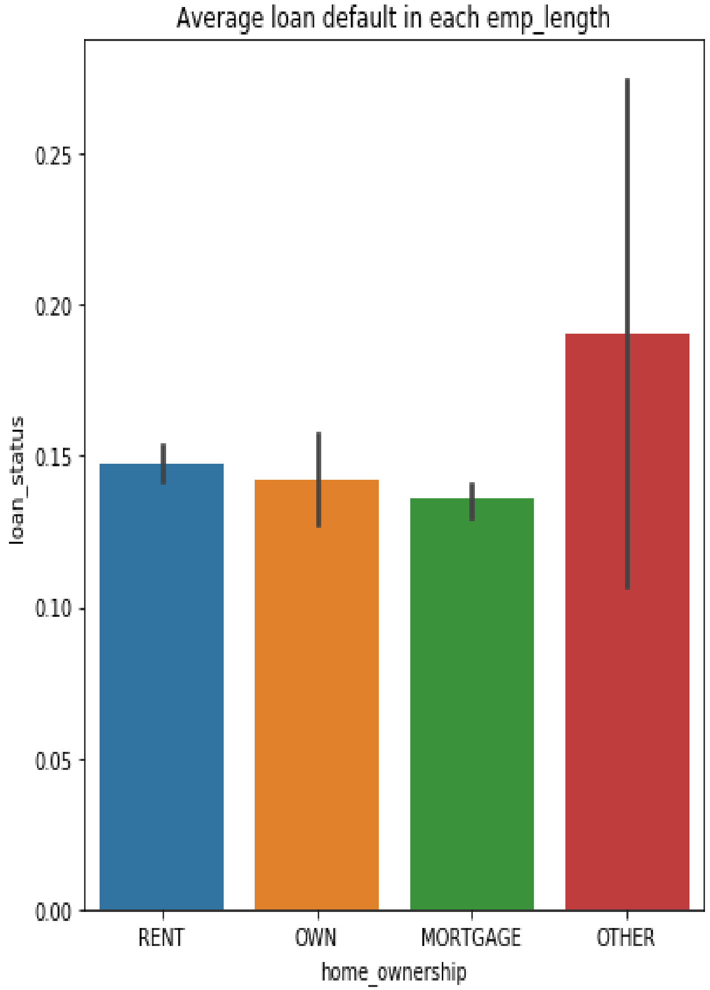
Insights: 75.3% of applications applied loan for 36 months term period

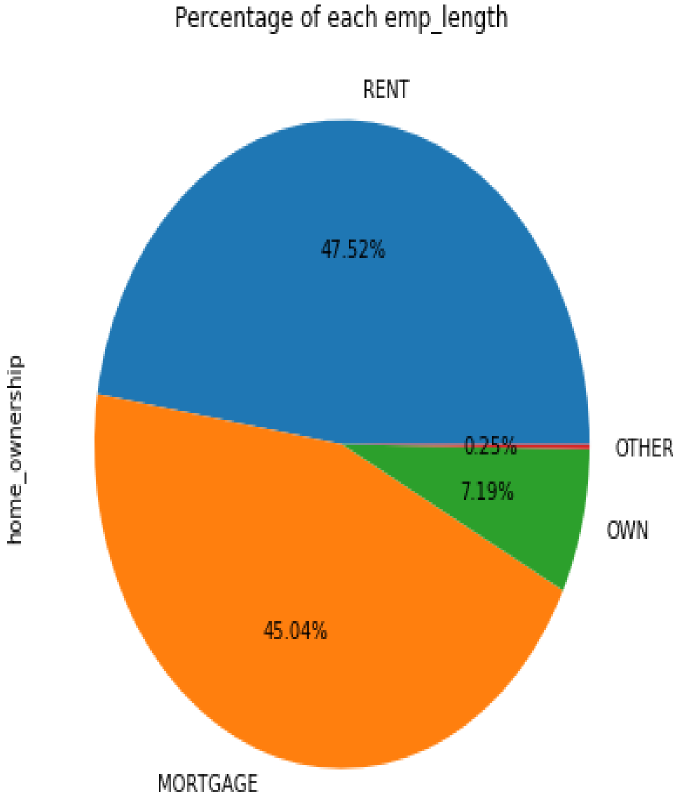
Categorical –Homeownership



Insight: 49.2% applicants are living in rented homes whereas 47% applicants were mortgaged their home

Lending Club Case Study – Detailed Analysis

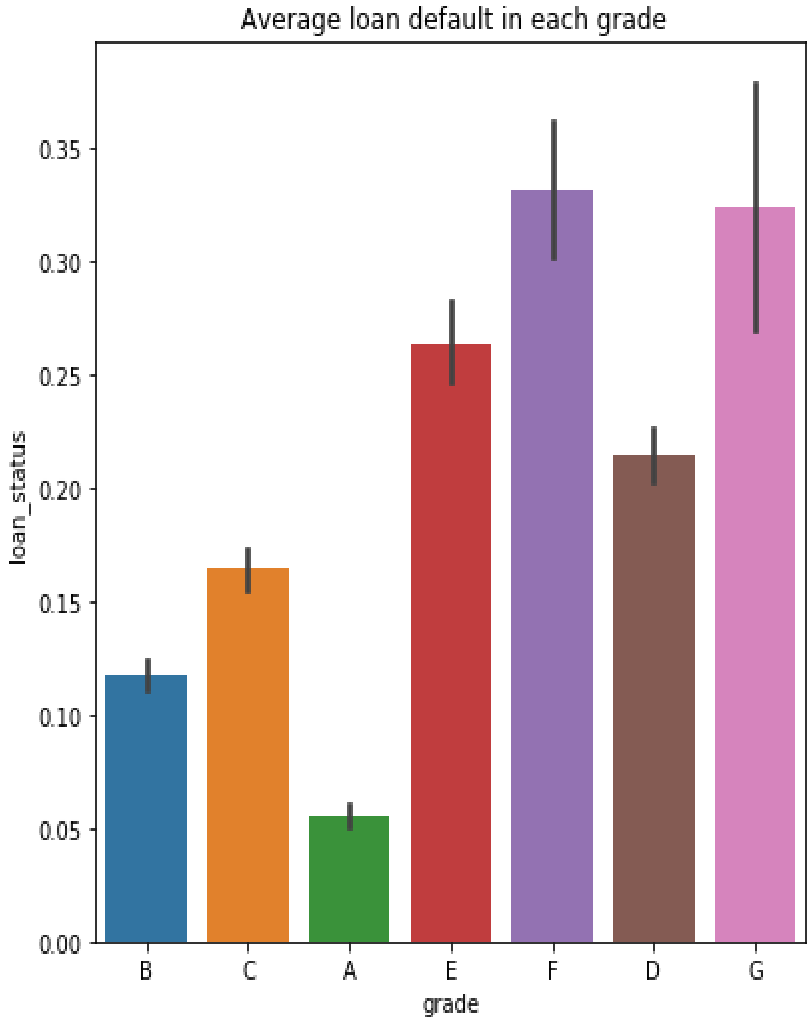


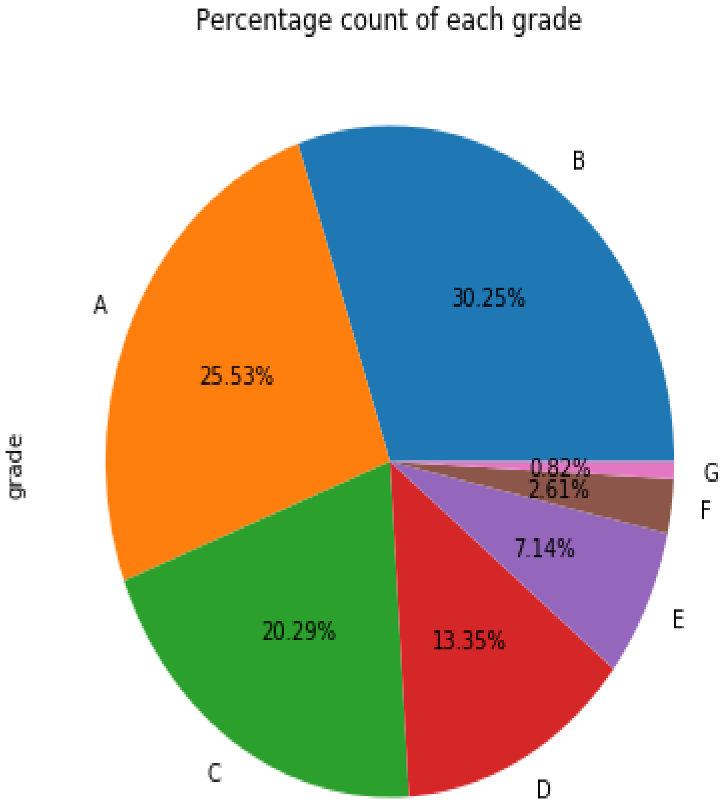


Insights : It's interesting to see that group of people who has not declared their exact homeownership status (others) , are the one who defaulted more on loans.

It should be restricted to not approve loan for this category (others) as it also constitutes just a fraction of percent of total loans. (however % of distribution of loans to other category is very low (0.25%)

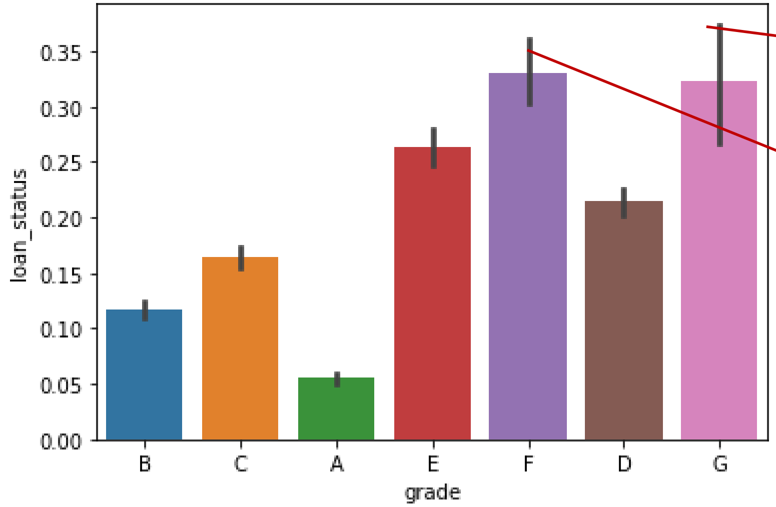
Lending Club Case Study – Detailed Analysis



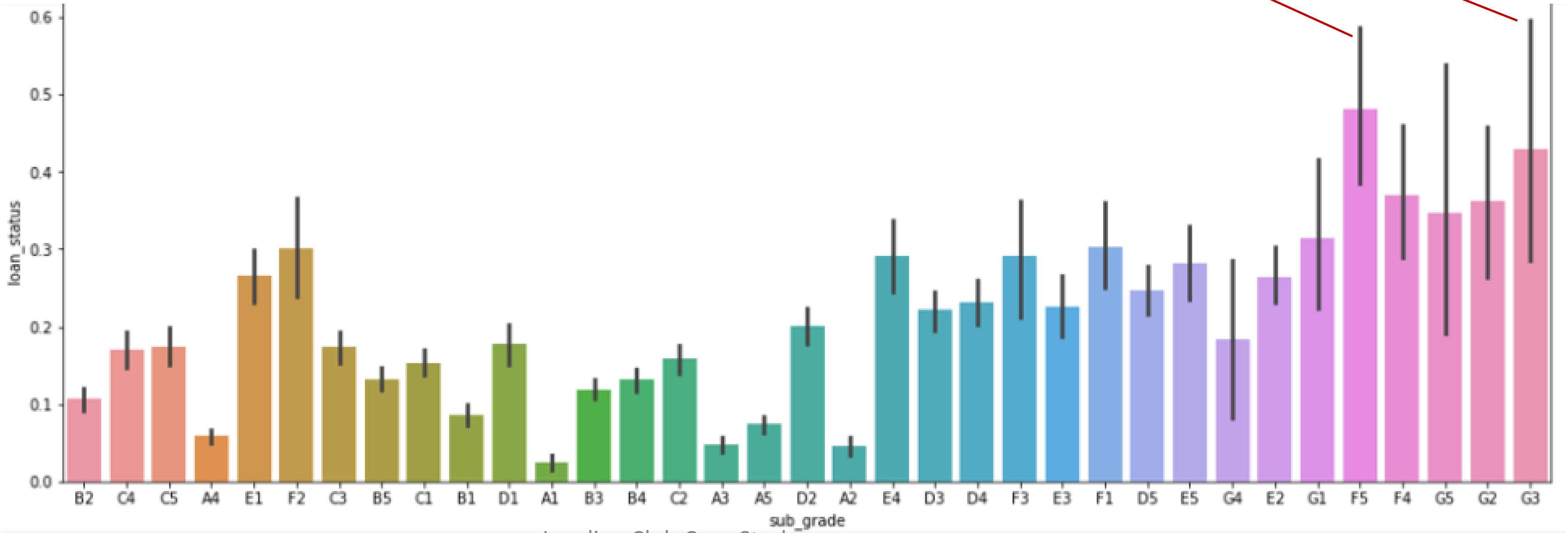


Insights: As per the graph charts , pie chart explains you each grade how much % of applicants falling and below 10% of the applicants are more defaulters in Grade E,F,G ,We should have more scrutiny during loan process to these kind of loans

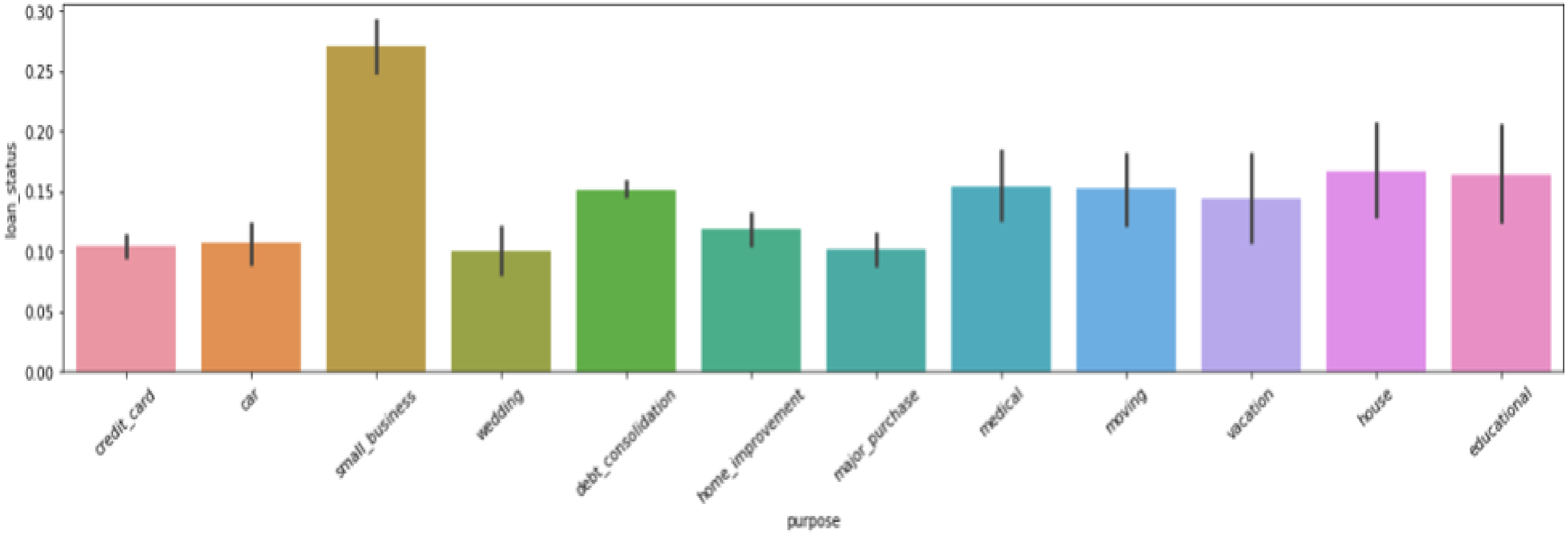
Lending Club Case Study – Detailed Analysis



Insights: As per above charts between Grade and sub grade level .. In **Grade F**, subgrade **F5** is more defaulters after that F4 an In **Grade G** ,subgrade **G3** has more defaulters and we have to take care while considering loan approval especially **F5** and **G3** sub grades.

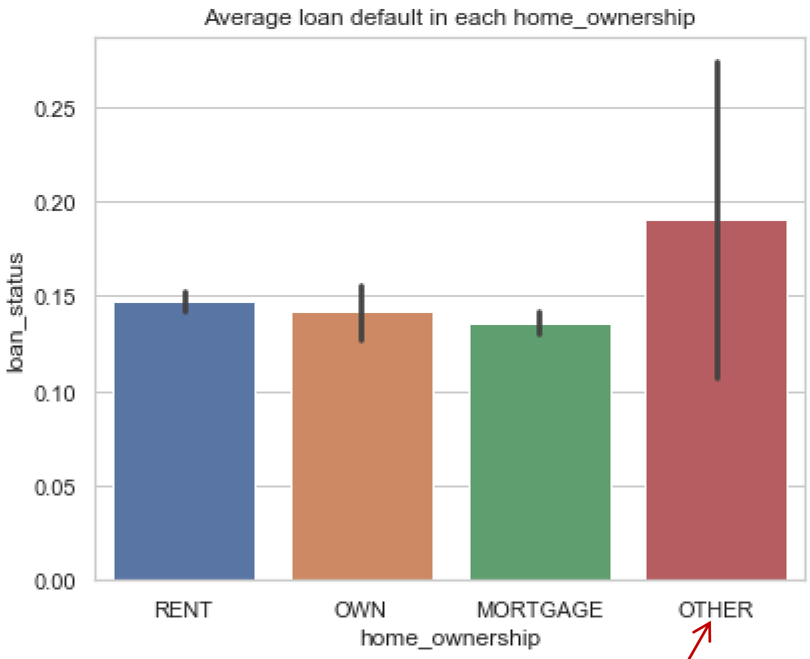


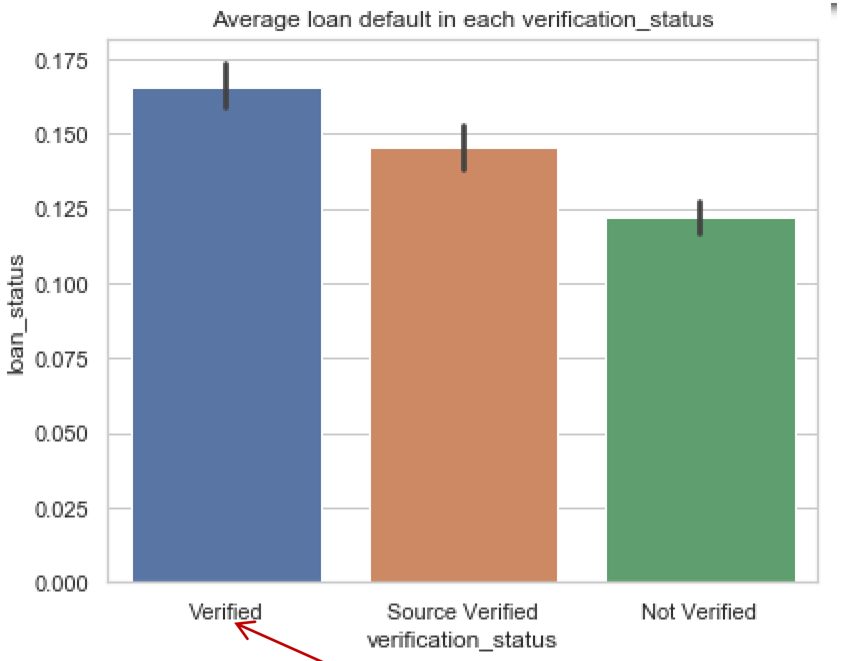
Lending Club Case Study – Detailed Analysis



Insights: As per the above graph “small_business” applicants are more defaulters

Lending Club Case Study – Detailed Analysis

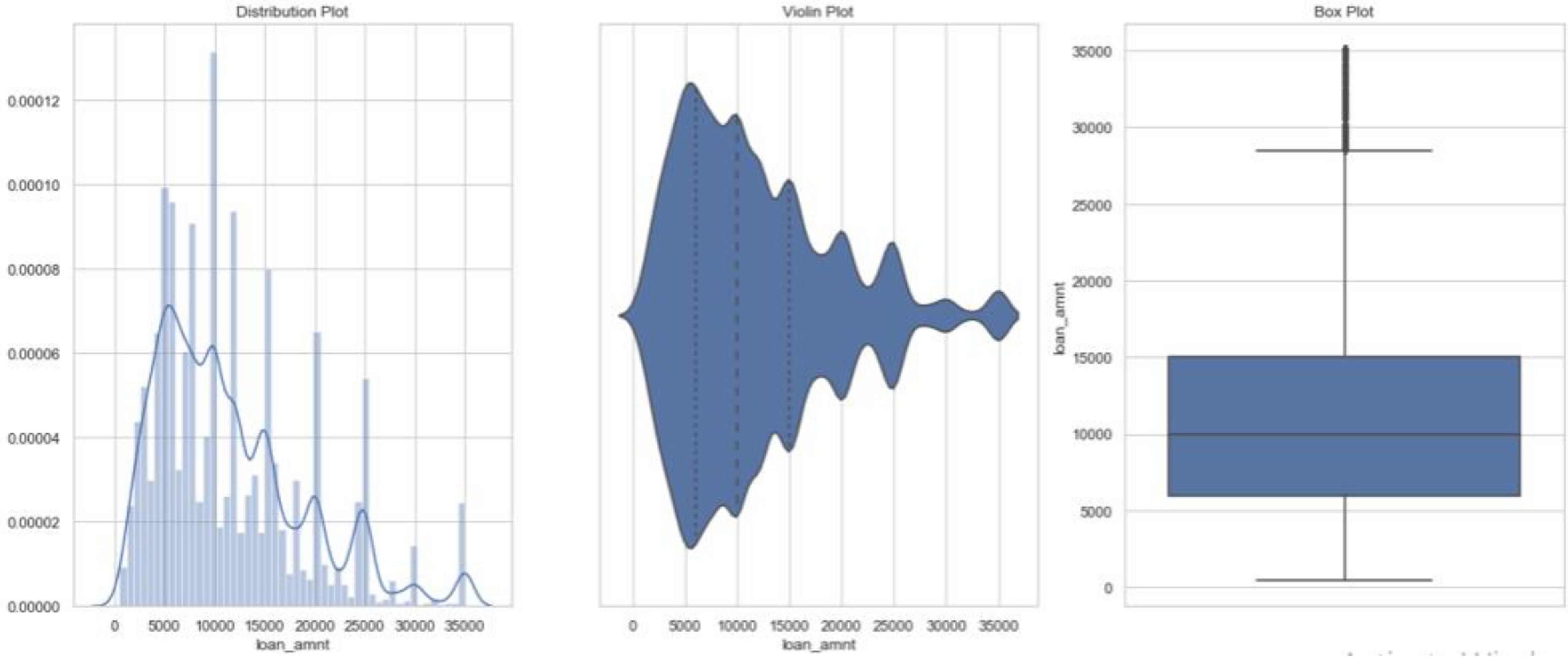

Average loan default in each home_ownership


Average loan default in each verification_status

Insights: As per above graphs “homeownership” **Others** are more defaulters and “Verification_Status” - **verified** more defaulters than secure verified and not verified.

Lending Club Case Study – Detailed Analysis

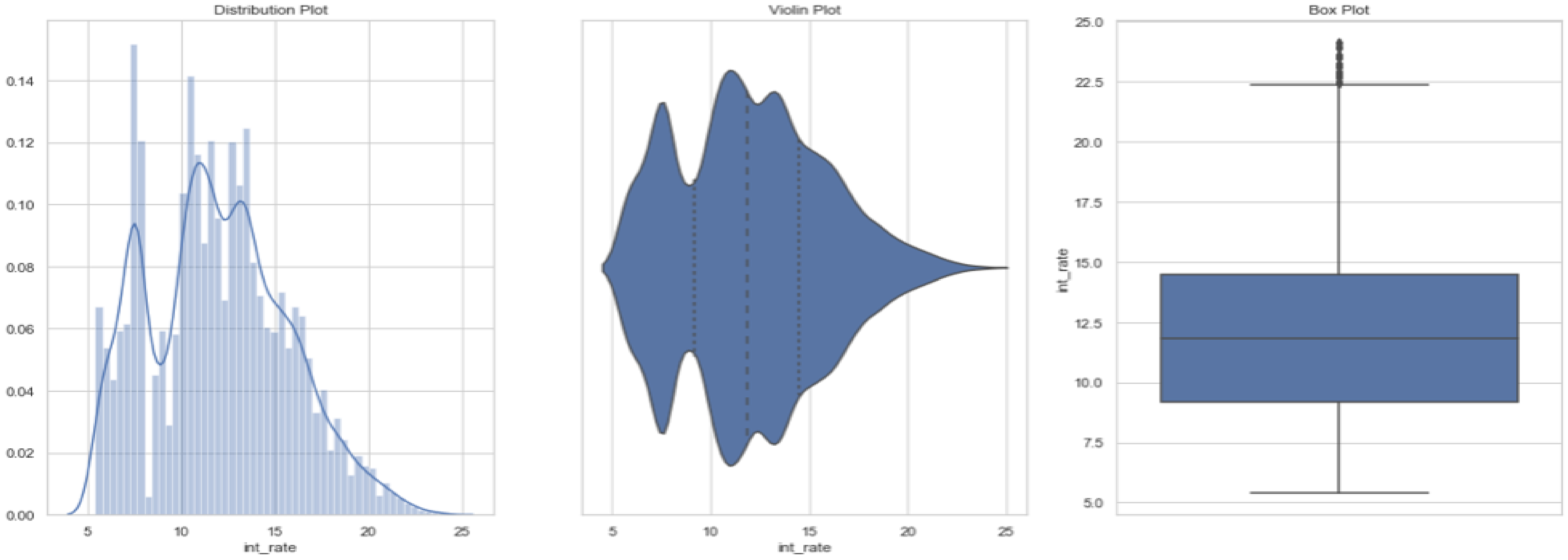
For continuous Variables - Loan Amount



Insights of this plots represents loan amounts are distributed between 8k and 20k

Lending Club Case Study – Detailed Analysis

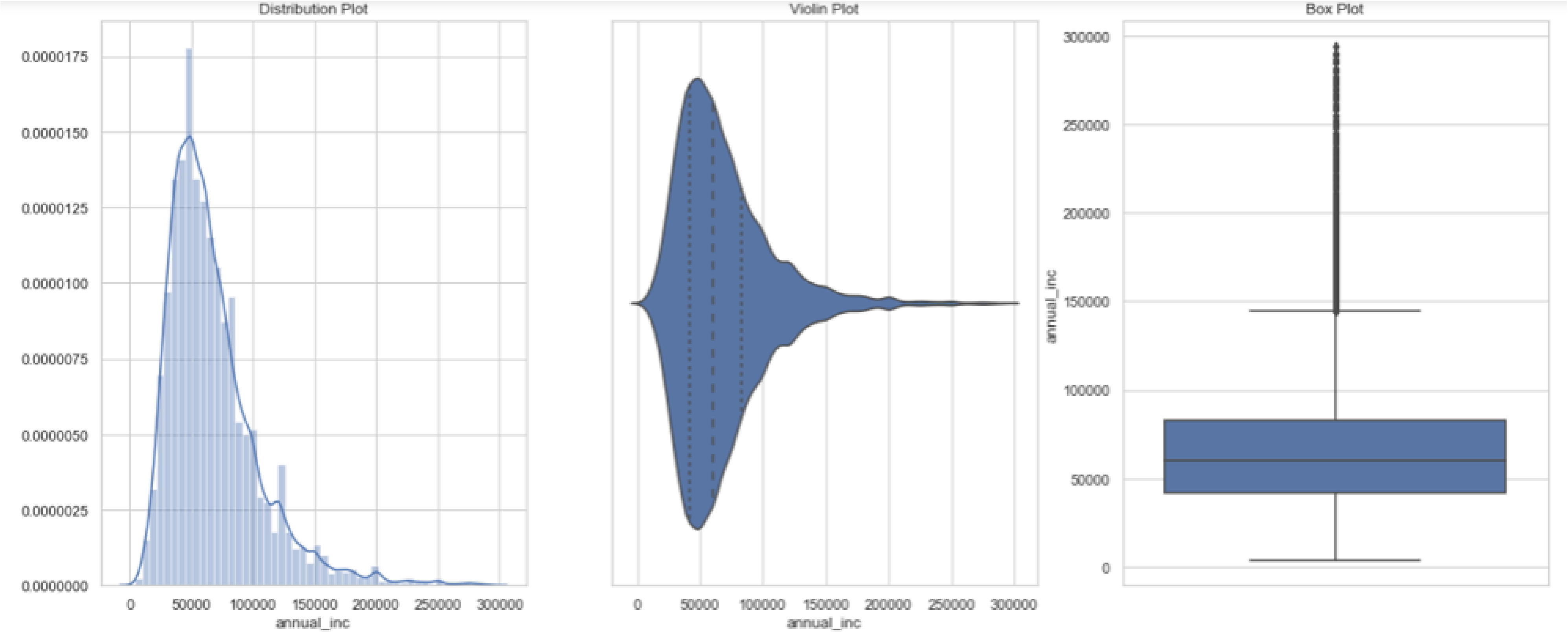
For continuous Variables – Interest Rate



Useful insight from this plots are that interest rates are distributed between 10% to 16%

Lending Club Case Study – Detailed Analysis

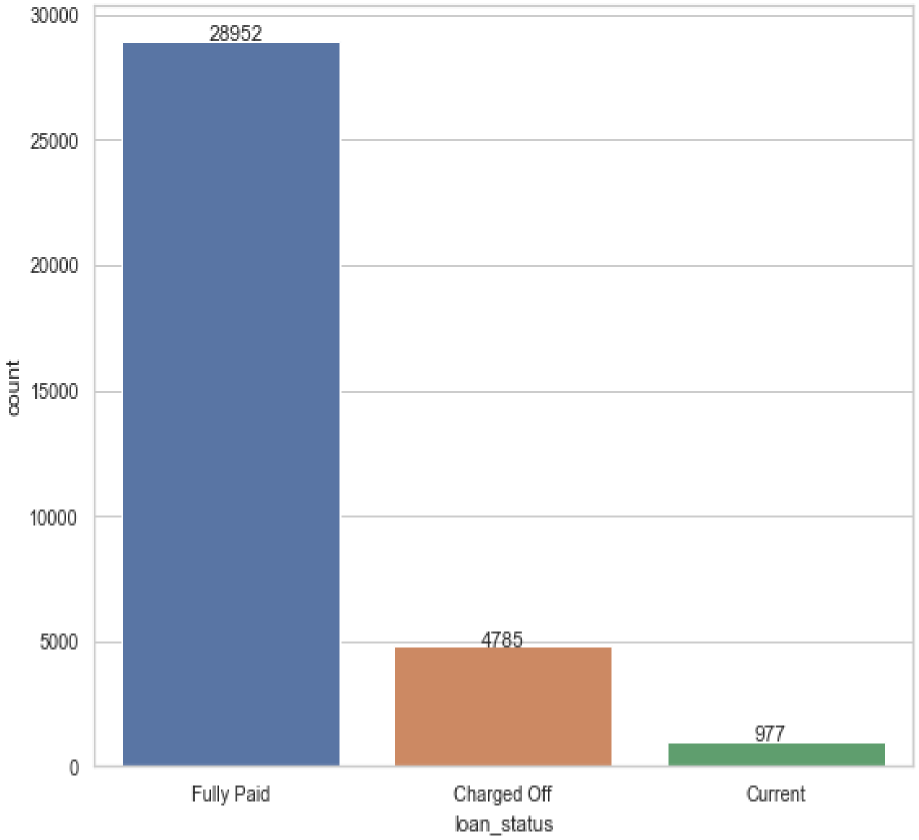
For continuous Variables - Annual Income



Insights: most of the applicants earns between 40000 to 90000 USD annually

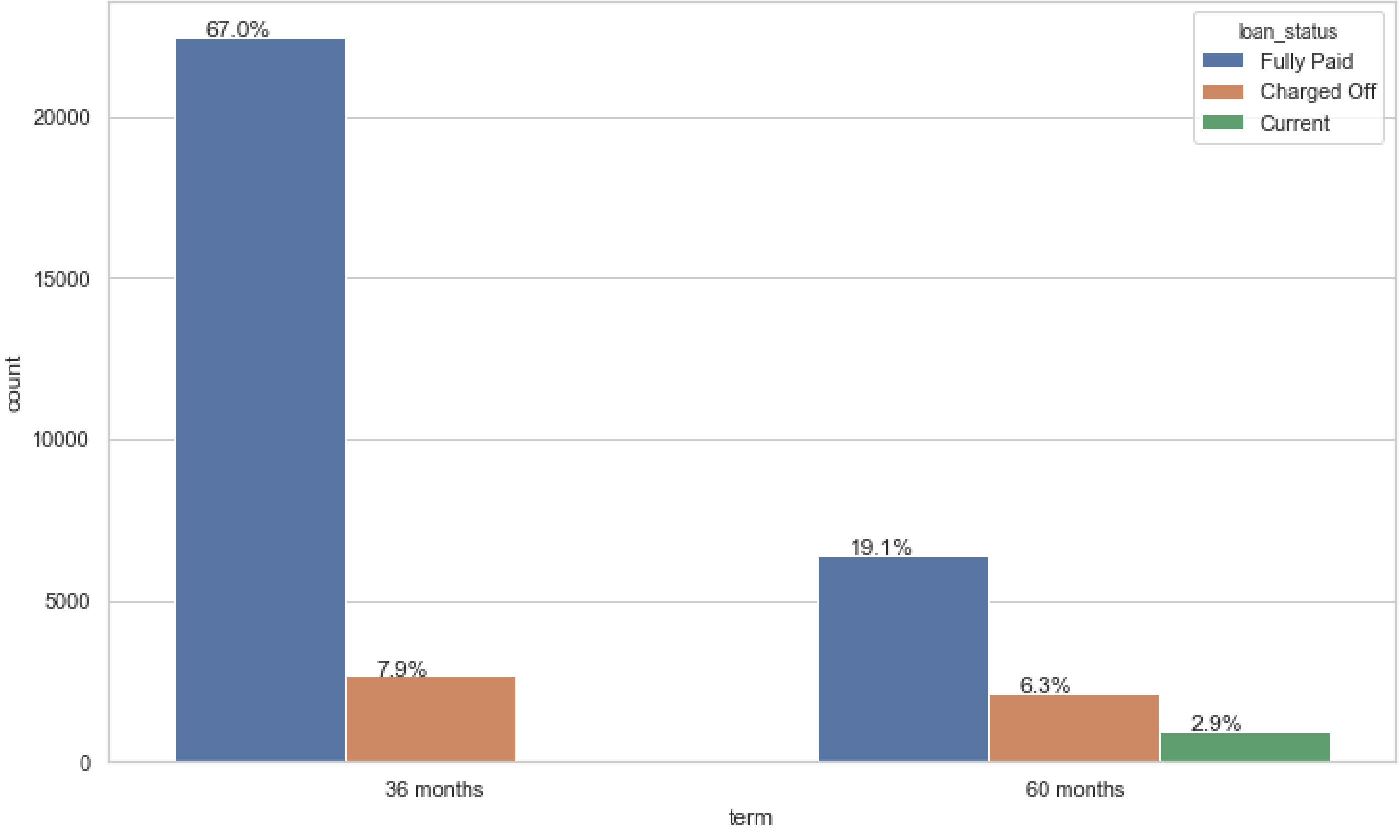
Lending Club Case Study – Detailed Analysis

Categorical variables - Loan Status



Insights is 14% of the applicants Charged off.

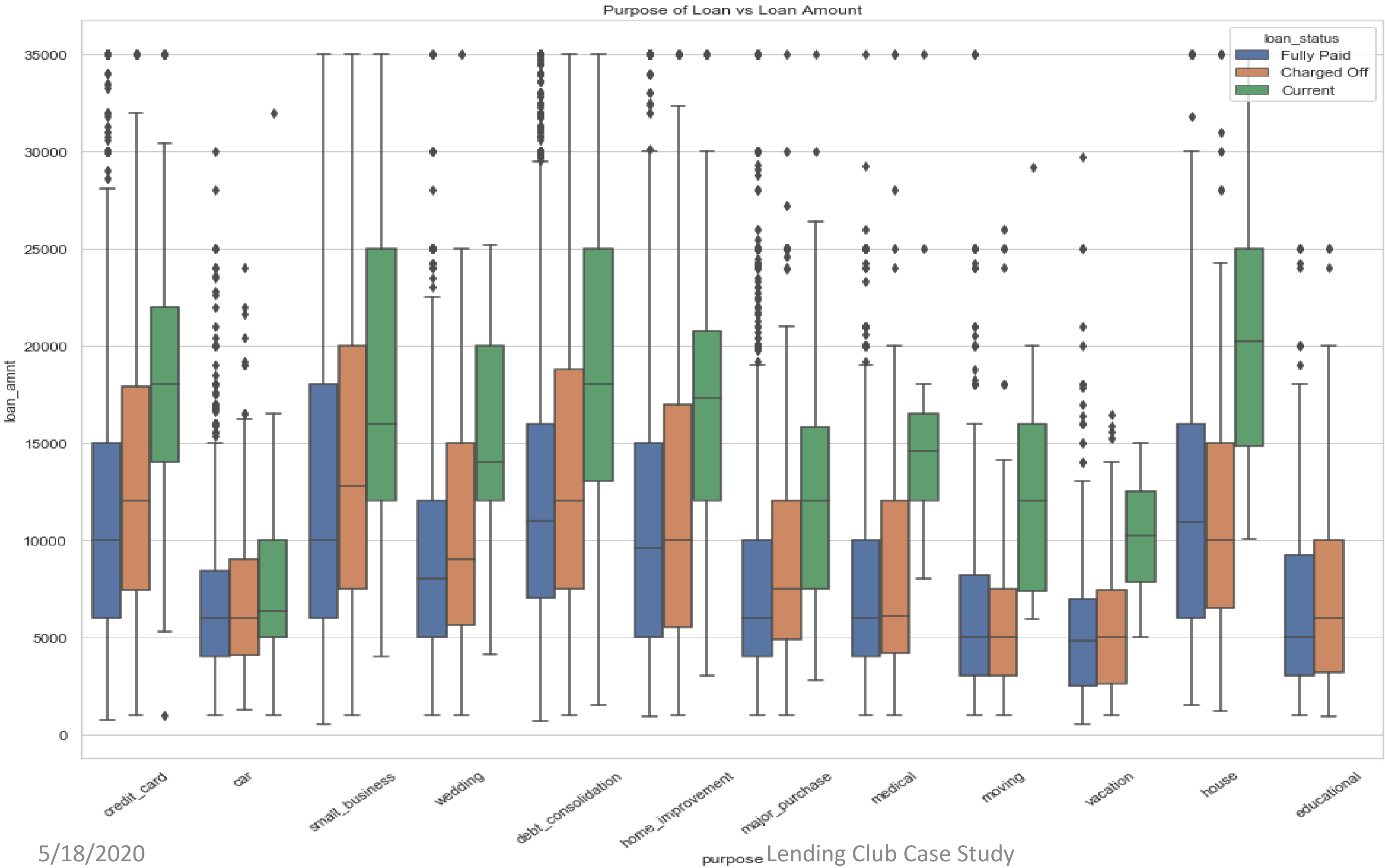
Categorical variables – Loan Term and no of applicants



Insights: 75.3% of applications applied loan for 36 months term period

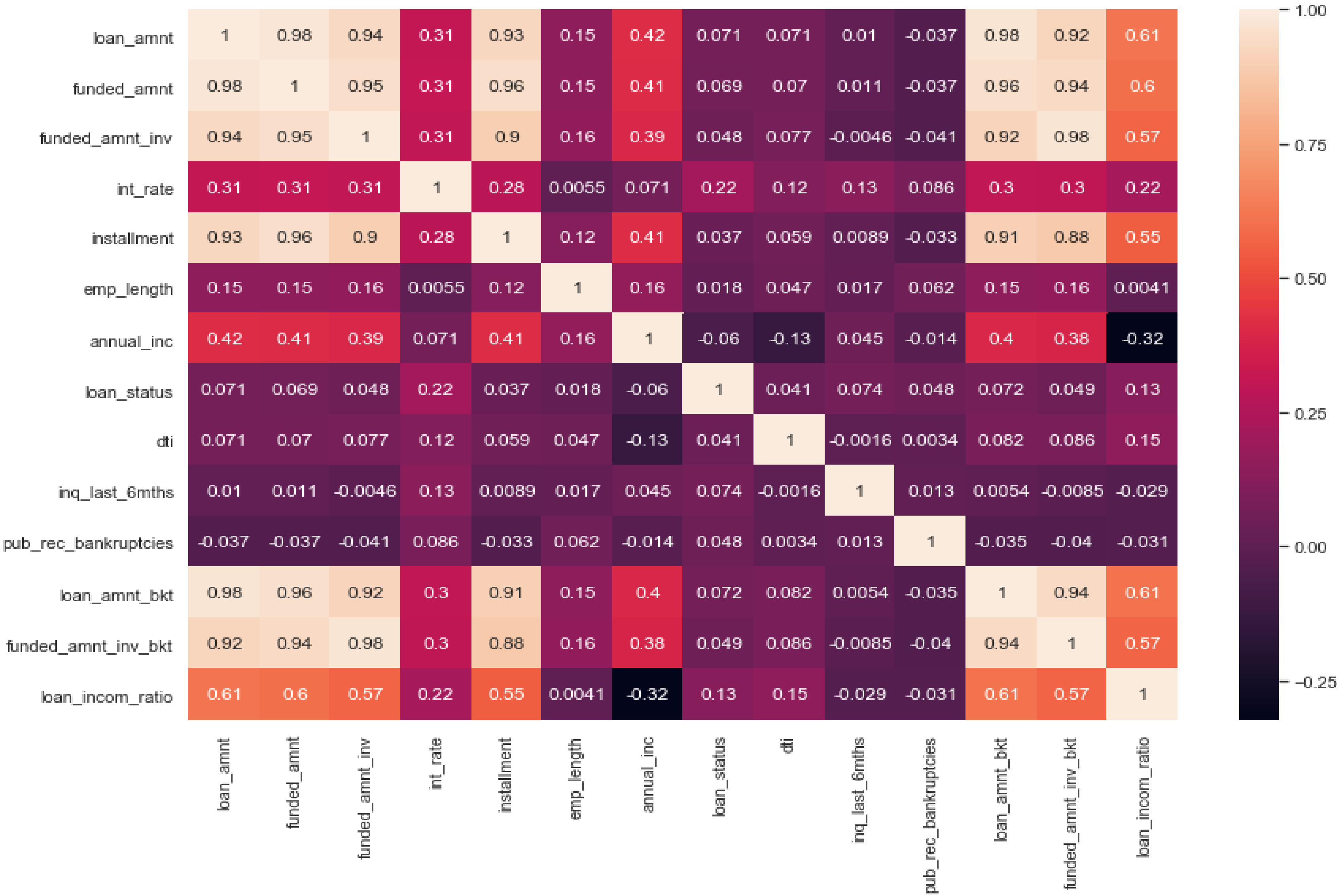
Lending Club Case Study – Detailed Analysis

Bivariate Analysis : purpose of Loan vs. Loan Amount for each Loan Status



Insights: purpose of Loan vs Loan Amount between applicants loan status (Fully Paid, Charged Off and Current).if you see over all charged off applicants took more loan amount than loan paid applicants

Lending Club Case Study – Detailed Analysis



	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	emp_length	annual_inc	loan_status	dti	inq_last_6mths	pub_rec_bankruptcies	loan_amnt_bkt	funded_amnt_inv_bkt	loan_incom_ratio
loan_amnt	1	0.98	0.94	0.31	0.93	0.15	0.42	0.071	0.071	0.01	-0.037	0.98	0.92	0.61
funded_amnt	0.98	1	0.95	0.31	0.96	0.15	0.41	0.069	0.07	0.011	-0.037	0.96	0.94	0.6
funded_amnt_inv	0.94	0.95	1	0.31	0.9	0.16	0.39	0.048	0.077	-0.0046	-0.041	0.92	0.98	0.57
int_rate	0.31	0.31	0.31	1	0.28	0.0055	0.071	0.22	0.12	0.13	0.086	0.3	0.3	0.22
installment	0.93	0.96	0.9	0.28	1	0.12	0.41	0.037	0.059	0.0089	-0.033	0.91	0.88	0.55
emp_length	0.15	0.15	0.16	0.0055	0.12	1	0.16	0.018	0.047	0.017	0.062	0.15	0.16	0.0041
annual_inc	0.42	0.41	0.39	0.071	0.41	0.16	1	-0.06	-0.13	0.045	-0.014	0.4	0.38	-0.32
loan_status	0.071	0.069	0.048	0.22	0.037	0.018	-0.06	1	0.041	0.074	0.048	0.072	0.049	0.13
dti	0.071	0.07	0.077	0.12	0.059	0.047	-0.13	0.041	1	-0.0016	0.0034	0.082	0.086	0.15
inq_last_6mths	0.01	0.011	-0.0046	0.13	0.0089	0.017	0.045	0.074	-0.0016	1	0.013	0.0054	-0.0085	-0.029
pub_rec_bankruptcies	-0.037	-0.037	-0.041	0.086	-0.033	0.062	-0.014	0.048	0.0034	0.013	1	-0.035	-0.04	-0.031
loan_amnt_bkt	0.98	0.96	0.92	0.3	0.91	0.15	0.4	0.072	0.082	0.0054	-0.035	1	0.94	0.61
funded_amnt_inv_bkt	0.92	0.94	0.98	0.3	0.88	0.16	0.38	0.049	0.086	-0.0085	-0.04	0.94	1	0.57
loan_incom_ratio	0.61	0.6	0.57	0.22	0.55	0.0041	-0.32	0.13	0.15	-0.029	-0.031	0.61	0.57	1

All continues variables

Insights: From the above Heatmap , we can see 'loan_amnt','funded_amnt','funded_amnt_inv' are closely interrelated.

Recommendation to Lending Club

Recommendations :

- Always prefer **low term (36)** loan that **high term (60)** loans because high term loan have found more defaulters as per the analysis between term and loan_status. Almost **75.3% of applicants applied for term-36 months**. So we can convince applicants to take short term loans with low interest rate to reduce defaulters based on his income.

- We observed the defaulters , 49.2% applicants are living in rented homes whereas 47% applicants were mortgaged their home , Our recommendation is that give loans against some security property to avoid loan defaulters in case **rented and mortgage loan applicants** else advised to reduce these kind of applicants or give loans high interest who stayed at rented and mortgage loan applicants. One more strong reason is that the applicants whose take loans are **52% of the applicants purpose is Debt Consolidation**, So , loan against **security (property)** is the advise.

- As per the insights, It's interesting to see that group of people who has not declared their exact homeownership status, are the one who defaulted more on loans. It should be restricted to not approve loan for this category as it also constitutes just a fraction of percent of total loans.

- As per the insights , Verification status - **verified** slightly more defaulters than secure verified and not verified , So it's better re do re verification strongly and Is verified properly or not (suspecting verification process).

- There seems to be not much difference based on employment length. However it's interesting to note that there are huge increase in loan applications after 10 years (25%) . It may be because of kids education support etc.,

- Interesting Insights are "small_business" applicants are more defaulters, keep an eye and enquire about their business profits /loss.
- Purpose of loans vs Loan Amount , more defaulters in educational purpose, please hold necessary documents against education loans with low interest rates.

- Purpose of loan like "Credit Card" and "debt consolidations" has found more defaulters as well as loan amount also more than 5000 USD , please keep an eye to reduce loans for these 2 purposes.

- Insights as per the heat map , we can see 'loan_amnt','funded_amnt','funded_amnt_inv' are closely interrelated. Almost all these 3 variables loan requested amount and funded amount is almost same ,We can recommending that funded amount (approved or disbursed loan amount would be max 75 % to 80% percentage of then loan amount to avoided loses .
- Insights: ,Grade and sub grade level .. In **Grade F**, subgrade **F5** is more defaulters after that F4 an In **Grade G** ,subgrade **G3** has more defaulters and we have to take care while considering loan approval especially **F5** and **G3** sub grades. Keep an eye and try to reduce to approval these sub categories.