Lending Club – Group Case Study

Group:

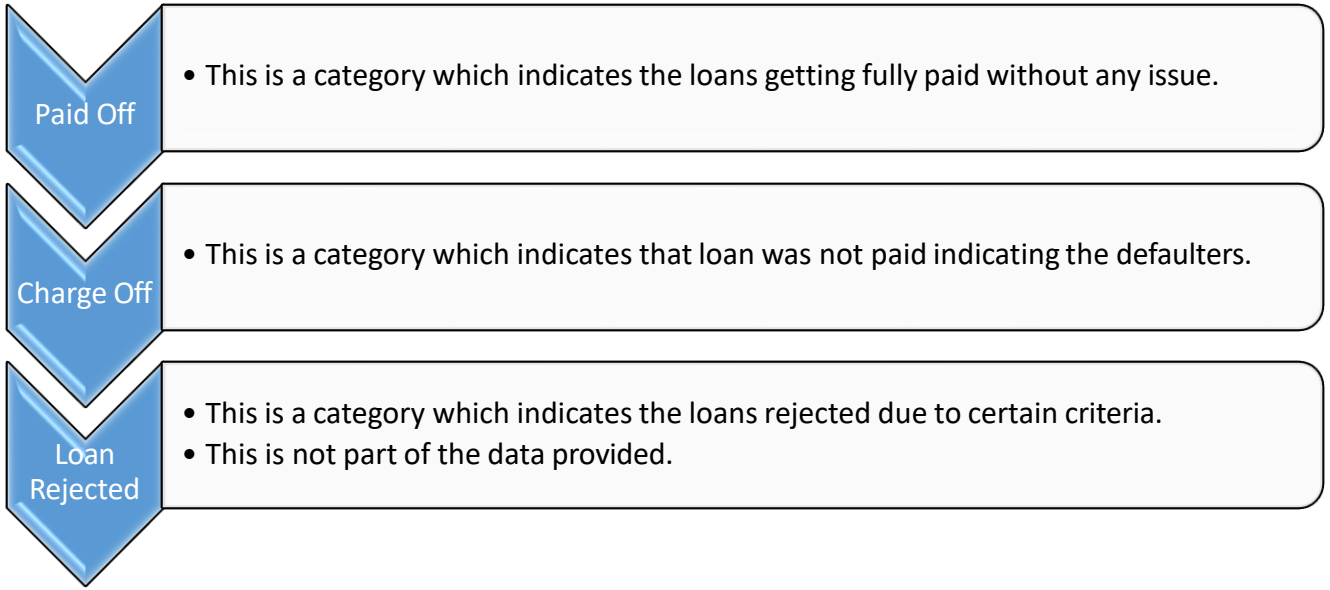
Sunil Kumar Veldurthi
Alexander Krishna

Lending Club – Problem Statement

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

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.

Analysis has to happen in a systematic way to get the best fit understanding on the data and finally provide recommendation.



Business objective: The objective is to identify these risky loan applicants, then such loans can be reduced there by cutting down the amount of credit loss using EDA and understand how consumer attributes and loan attributes influence the tendency of default.

Lending Club – Problem Solving Methodology

Flow Diagram

Data Understanding

Based on initial look of the data below were the conclusion made:

- All customer behaviour related variables are removed.
- Based on the distinct values below are the different category of variables from the historical file provided:
  - Categorical variables: home_ownership,loan_status,verification_status,pub_rec,annual_inc_range,emp_title,grade,term,sub_grade,title,purpose,addr_state, pub_rec_bankruptcies,emp_length
  - Continuous variables: loan_amnt,int_rate,installment,annual_inc,dti,delinq_2yrs,inq_last_6mths,open_acc,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_amnt
- Identified that all the columns with more 80% empty values in a column are not useful for further analysis.
- Target variable to be analysed is Loan Status.

Data Cleansing

The very first step toward solving any analytics problem is to have clean data to understand the insights. Hence to begin with we have found out the missing values in every column of the dataset.

Below are missing value percentage in the provided data below. **Also** in the tabular column documented in the below has the column to impute values/ drop.

Below are the method adopted to impute the values:
Mode Method
Mean Method
Update the data based on understanding

Identify Single values columns and drop columns from Data to explore.

column_name	null_percentage
mths_since_last_record	92.985372
next_pymnt_d	97.129693
mths_since_last_major_derog	100.000000
annual_inc_joint	100.000000
dti_joint	100.000000
verification_status_joint	100.000000
tot_coll_amt	100.000000
tot_cur_bal	100.000000
open_acc_6m	100.000000
open_il_6m	100.000000

Single Valued Columns

['pymnt_plan',
"initial_list_status",'collections_12_mths_ex_med',
'policy_code','acc_now_delinq', 'application_type',
'pub_rec_bankruptcies', 'tax_liens', 'delinq_amnt'
]

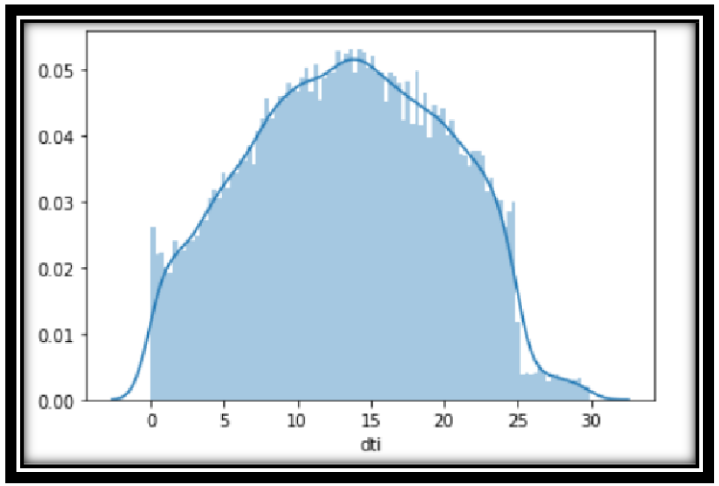
Uni-Variate Analysis

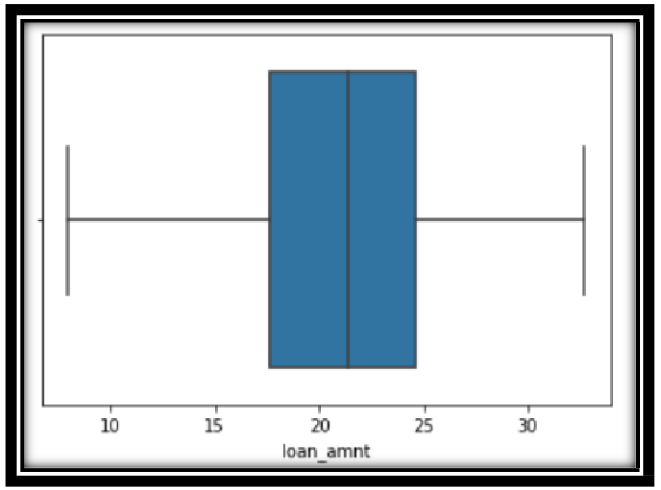
The analysis for uni variate is mostly done to identify outliers/ extreme values for all the columns identified. This is done to make sure that the analysis is not affected by those extreme value on the loan status.

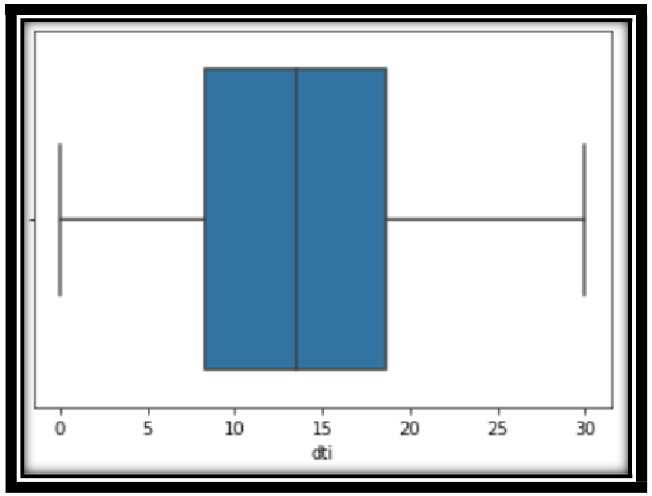
The outliers are treated with 2 techniques.

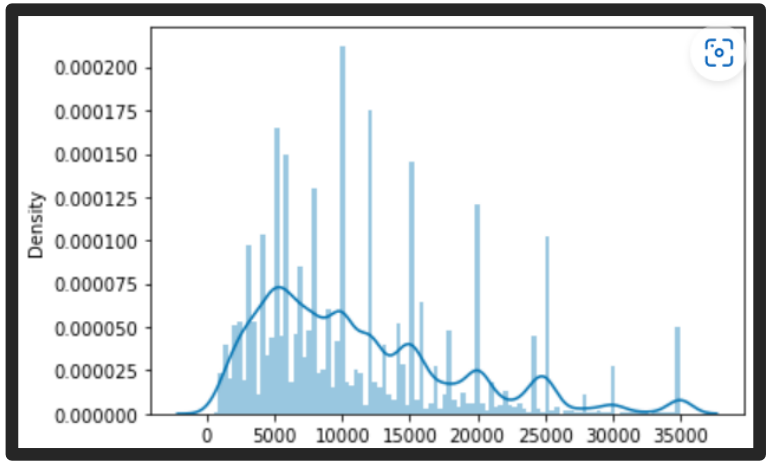
1. Scaling the value in the column.
2. Use Inter quartile technique to remove the extreme values.

Providing the below box plot for annual_inc, loan amount and dti column. These plots are post outlier treatment. To confirm its distribution we even plot the histogram plot.
Similar to this all others columns are plotted to check if the outliers are removed.







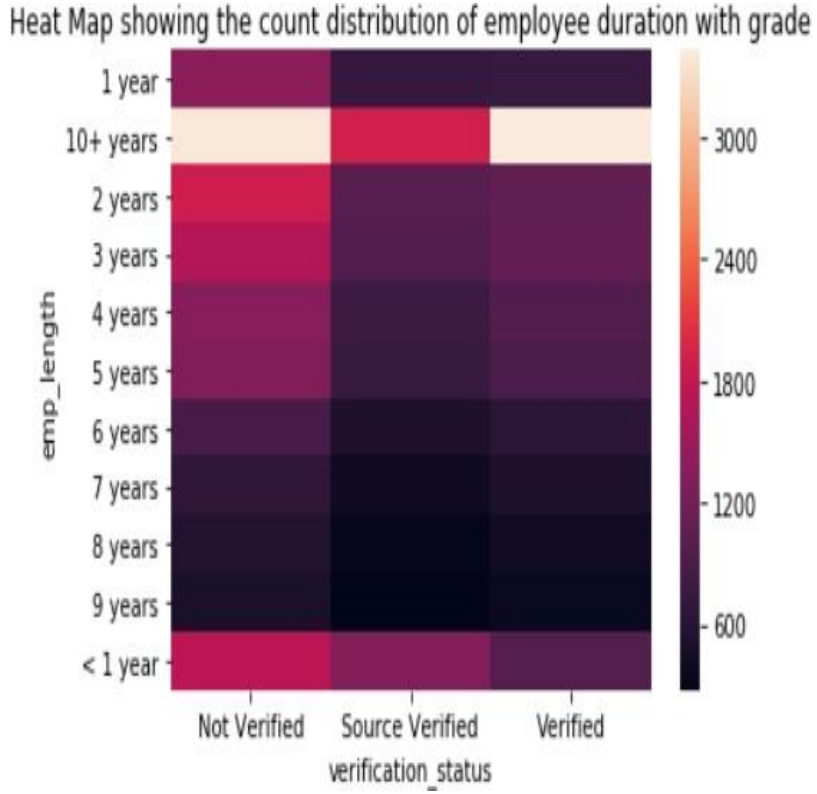


Bi-Variate Analysis – Between Categorical Variables

- Analysis was performed on variables which have categorical values. Below are the major take away points.

- Also slide contains few heat maps to provide us the insight of data.

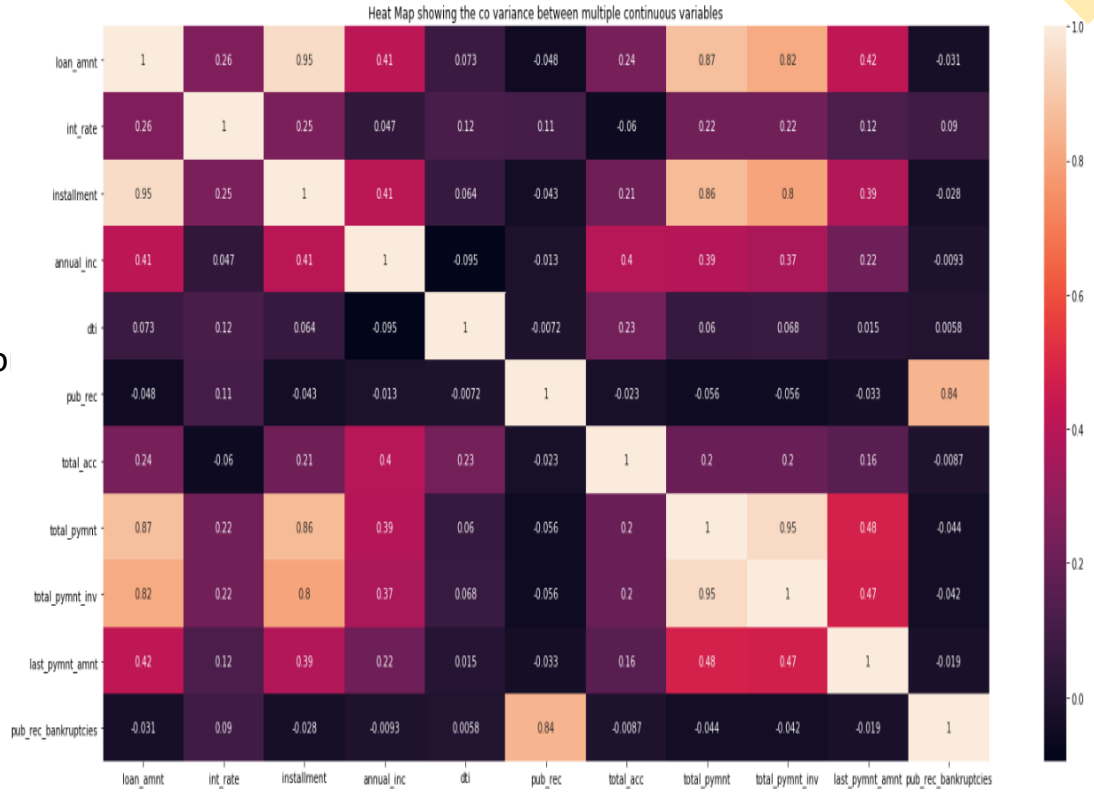
- Majority are provided for loan amount between 10000 to 5000 with grade assigned as A.
- Based on the count we see that loans are charged off for income range between 10000 to 5000.
- Most loans provided with the borrows annual income between 10000 and 5000 are not verified.
- Most of the borrowers having employment experience more than 10 plus years are verified.
- Most of the borrowers loan requirements are ranging from 5000 to 20000.
- Based on the reference grade is related to the interest rate, hence most of the interest rate are very less.
- That is grade A has very low interest rate and grade G has high interest rate for the loan.



Bi-Variate Analysis – Between Continuous Variables

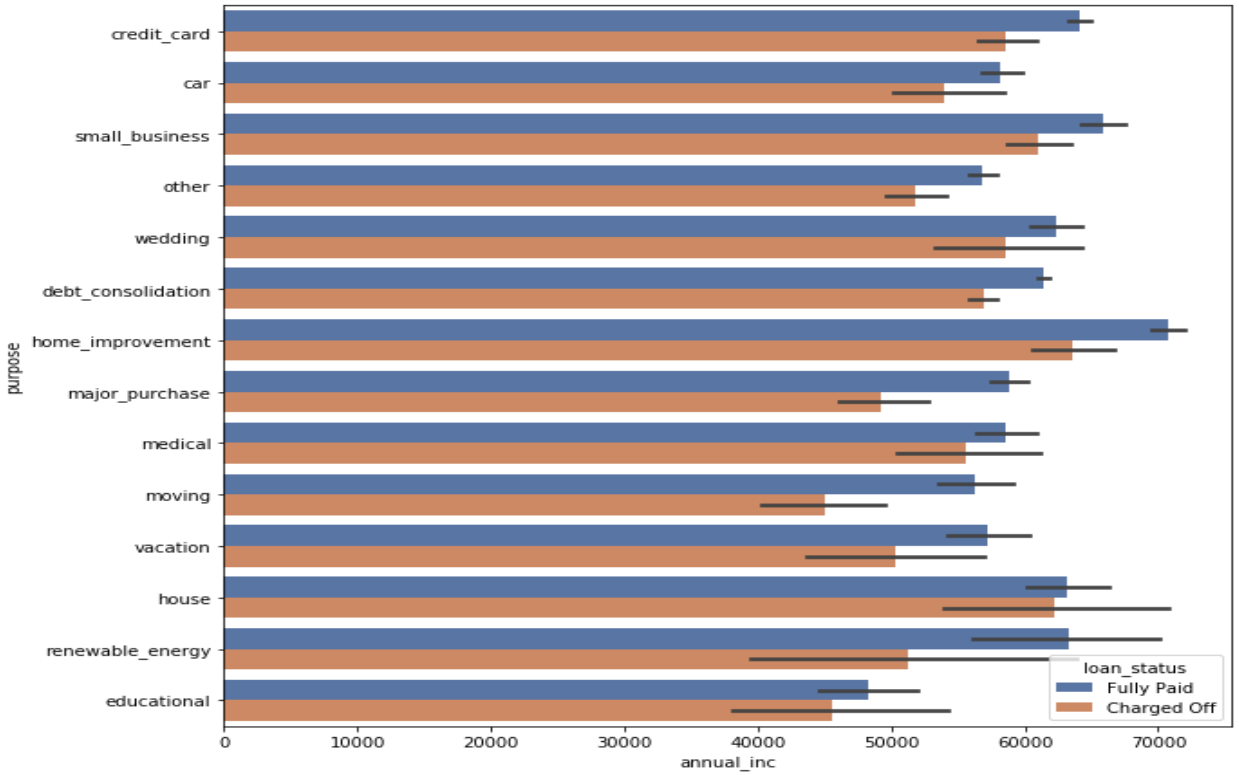
Below is the heat map plotted between continuous variables, also listing down the major take away points:

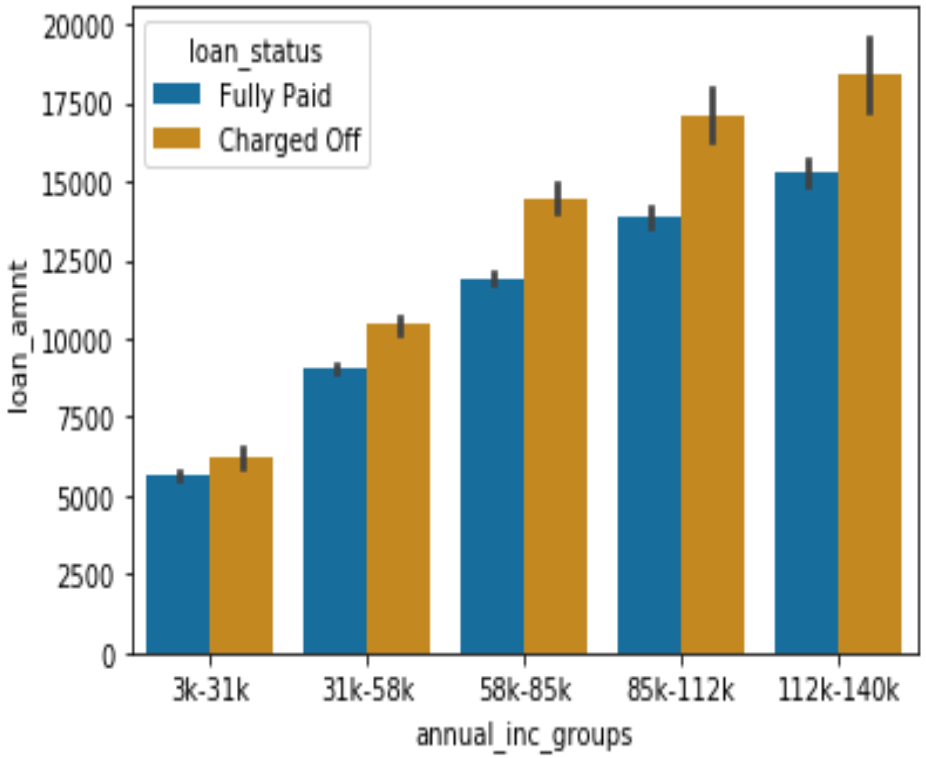
- Total payment and total payment inv column has a clear positive correlation hence using any one of the variable is sufficient for analysis.
- Loan amount and instalment column has a clear positive correlation hence using any one of the variable is sufficient for analysis.
- Annual income and dti has almost no correlation, hence both can be used fo analysis separately which can provide us more insights.
- From loan amount ,dti, total payment, annual income and pub rec bankruptcies can be used for analysis with loan status as they are not correlated. This can give us more patterns and useful information to make a proper recommendation.



Bi-Variate Analysis – Between Continuous Variables

Below are the major plots and the conclusion we can make:



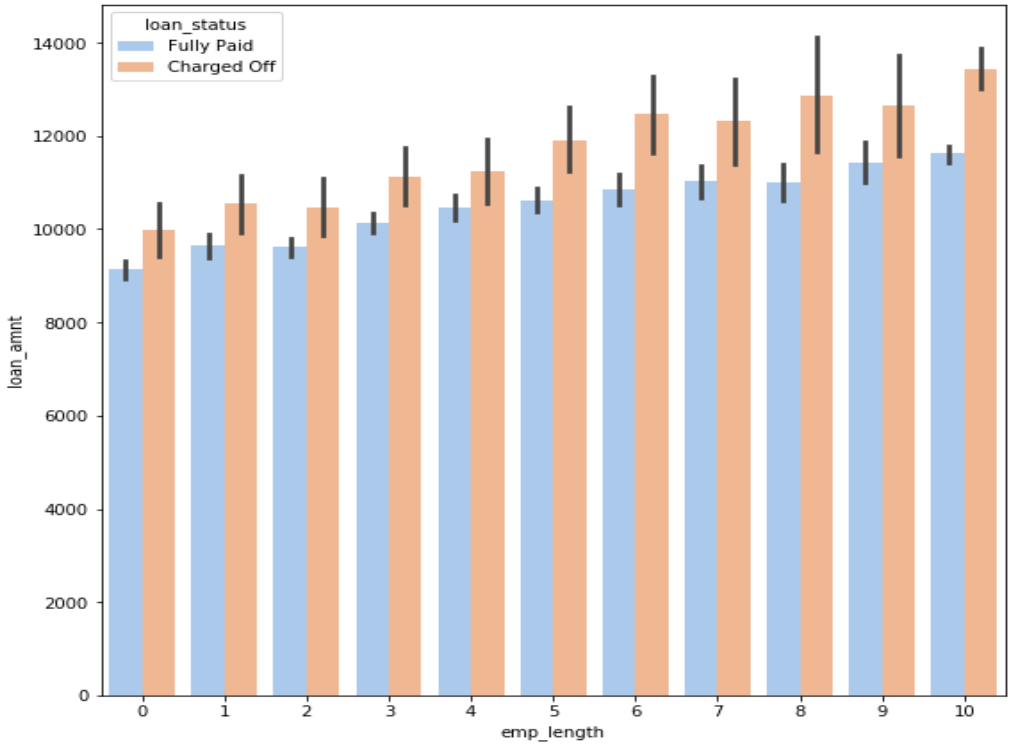


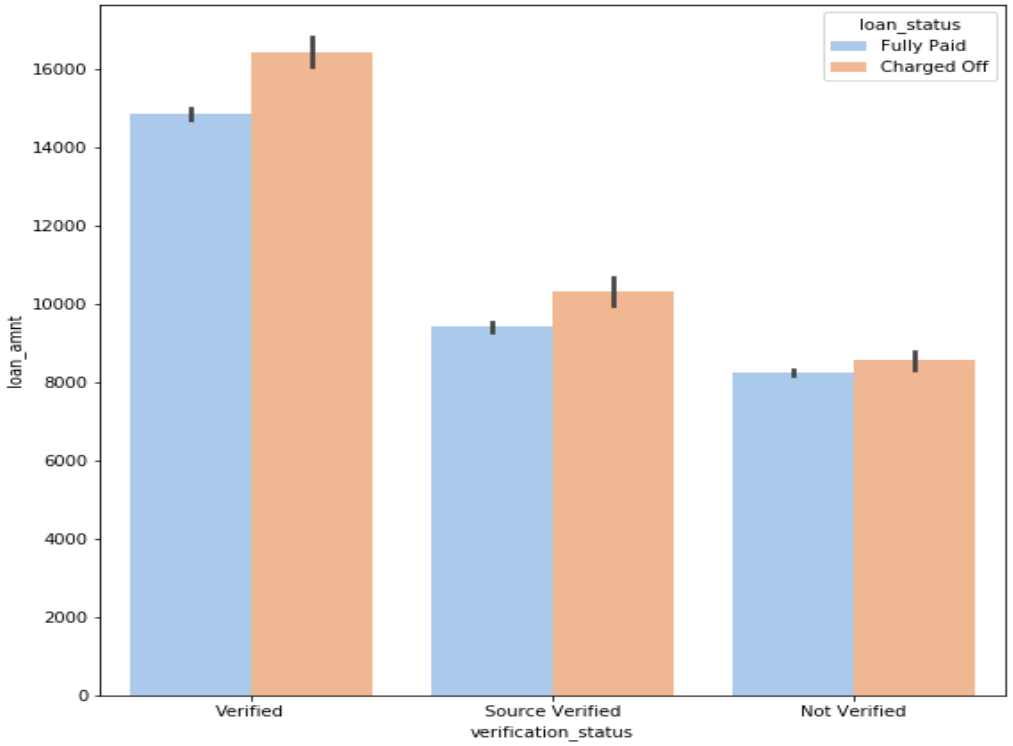
Below are the major take away from the plots provided in the slide:

- Applicants with higher salary mostly applied loans for "home_improvment", "house", "renewable_energy" and "small_businesses“
- Across all the income groups, the loan_amount is higher for people who defaulted

Bi-Variate Analysis – Between Continuous Variables

Below are the major plots and the conclusion we can make:



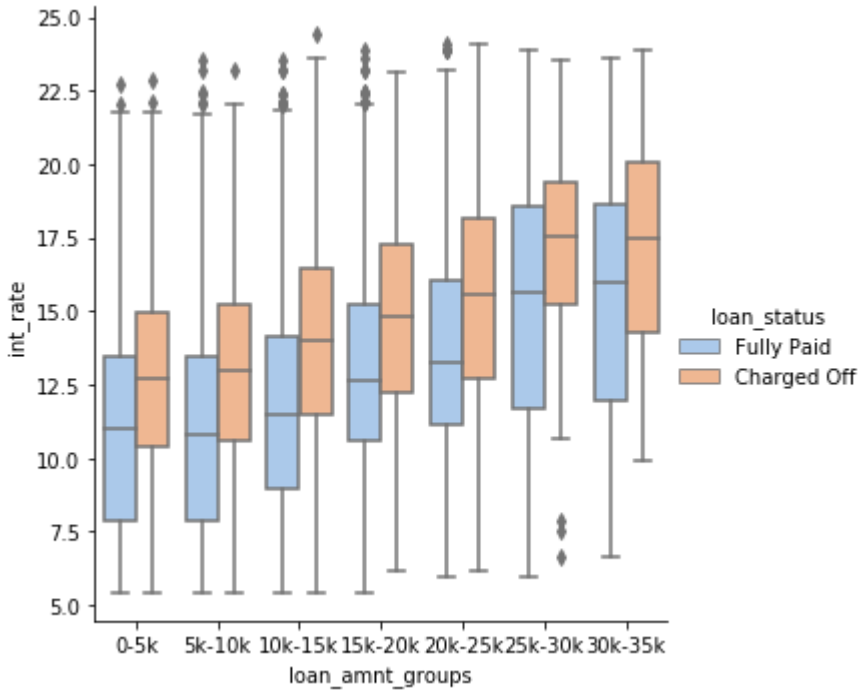


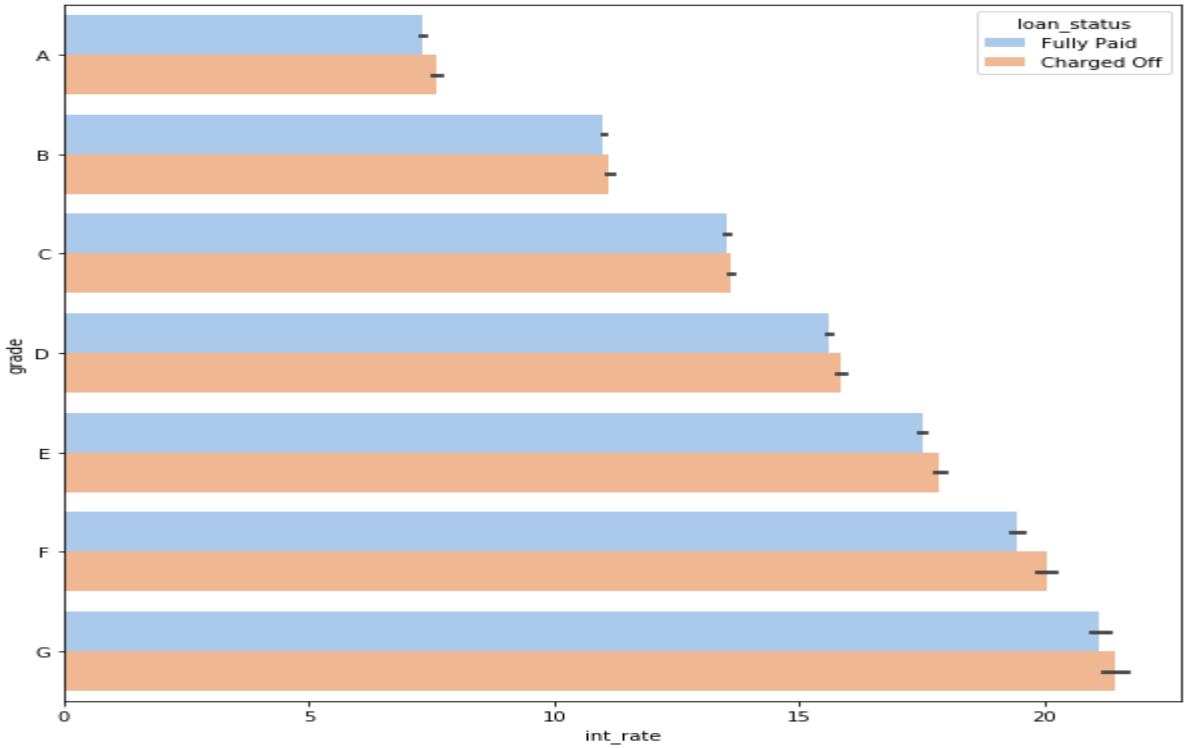
Below are the major take away from the plots provided in the slide:

- - Employees with longer working history got the loan approved for a higher amount.
- - Looking at the verification status data, verified loan applications tend to have higher loan amount. Which might indicate that the firms are first verifying the loans with higher values.

Bi-Variate Analysis – Between Continuous Variables

Below are the major plots and the conclusion we can make:



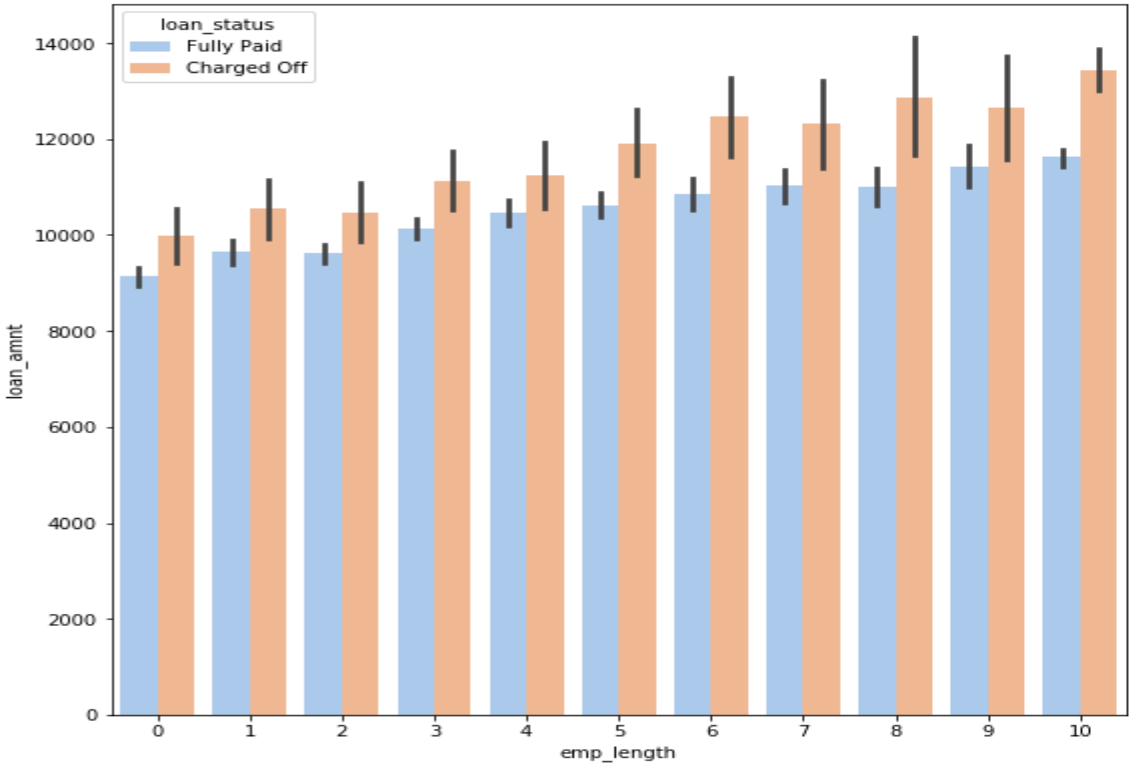


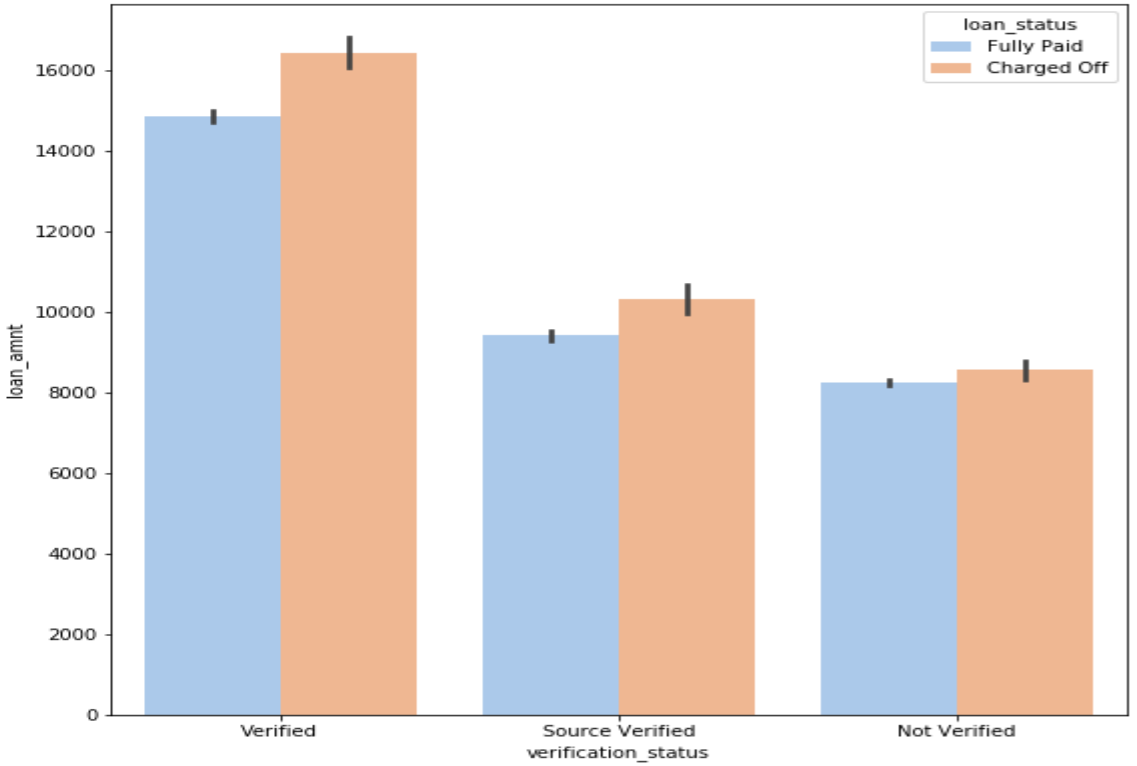
Below are the major take away from the plots provided in the slide:

- The interest rate for charged off loans is pretty high than that of fully paid loans in all the loan_amount groups.
- This can be a pretty strong driving factor for loan defaulting.

Bi-Variate Analysis – Between Continuous Variables

Below are the major plots and the conclusion we can make:





Below are the major take away from the plots provided in the slide:

- Employees with longer working history got the loan approved for a higher amount.
- Looking at the verification status data, verified loan applications tend to have higher loan amount. Which might indicate that the firms are first verifying the loans with higher values.

Recommendation- Lending club

- Below is the conclusion:
  - The top variables which has the impact over the loan status are:
    - ❑ Interest rate
    - ❑ Grade
    - ❑ Loan Amount
    - ❑ Loan Duration
    - ❑ Employee length

  - The Lending club has to make sure that proper verification has to be done before sanctioning the loan.
  - Based on the previous records the loan amount has to be sanctioned instead of looking the annual income.
  - Loan of Higher Interest rate have more defaulters, check the background of applicant if interest rate is vey high
  - Low grade loans have high tendency to default.