Lending Club – Group Case Study

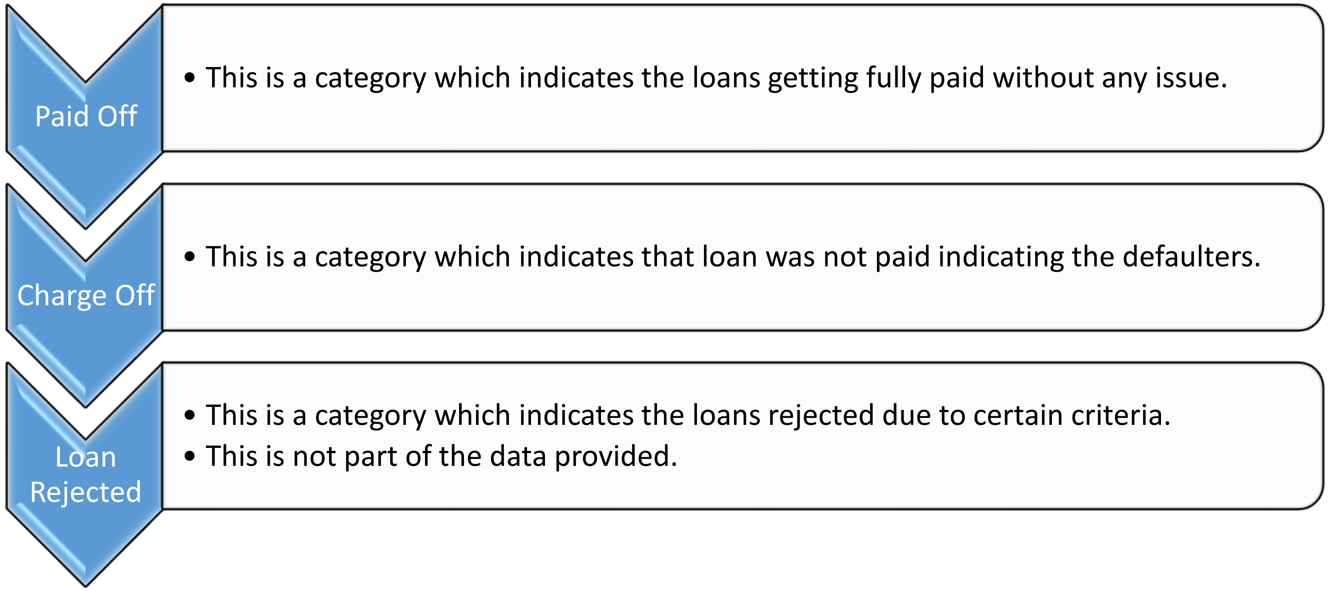
Name: Ashwin Rajagopalan & S Lakshmi Prasanna

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.



Paid Off

- This is a category which indicates the loans getting fully paid without any issue.

Charge Off

- This is a category which indicates that loan was not paid indicating the defaulters.

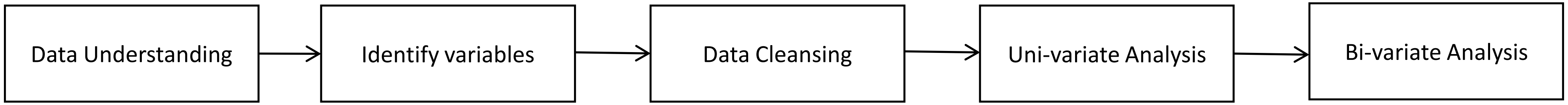
Loan Rejected

- This is a category which indicates the loans rejected due to certain criteria.
- This is not part of the data provided.

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:

1. All customer behaviour related variables are removed.
2. Based on the distinct values below are the different category of variables from the historical file provided:

   a. 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

   b. 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
3. Identified that all the columns with more than 60% empty values in a column are not useful for further analysis.
4. Target variable to be analysed is Loan Status.

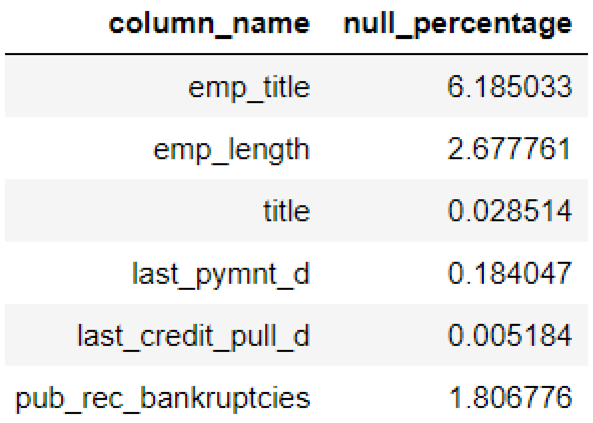
Data Cleansing

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. Also in the tabular column documented in the right hand side 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 of data



column_name	null_percentage
emp_title	6.185033
emp_length	2.677761
title	0.028514
last_pymnt_d	0.184047
last_credit_pull_d	0.005184
pub_rec_bankruptcies	1.806776

Loan.csv

Companies Dataset	Method to impute values
Emp_title	Mode; Most frequent value in the column
Emp_length	Mode; Most frequent value in the column
title	Mode; Most frequent value in the column
Last_paymnt_d	Mode; Most frequent value in the column
Last_credit_pull_d	Mode; Most frequent value in the column
Pub_rec_bankrupties	Mode; Most frequent value in the column

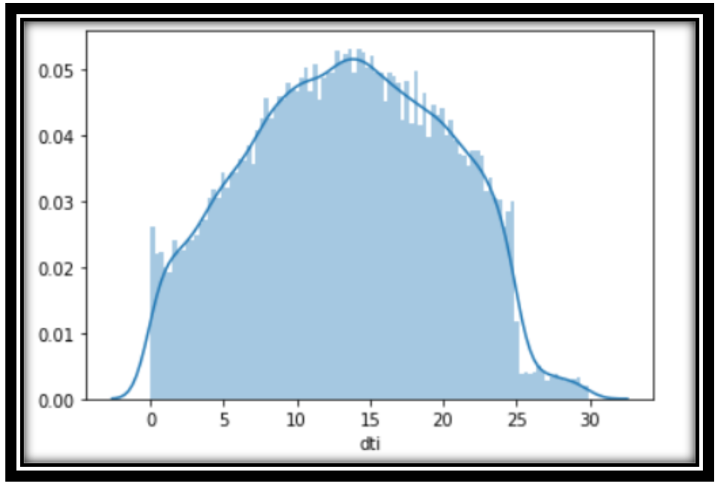
Uni-Variate Analysis

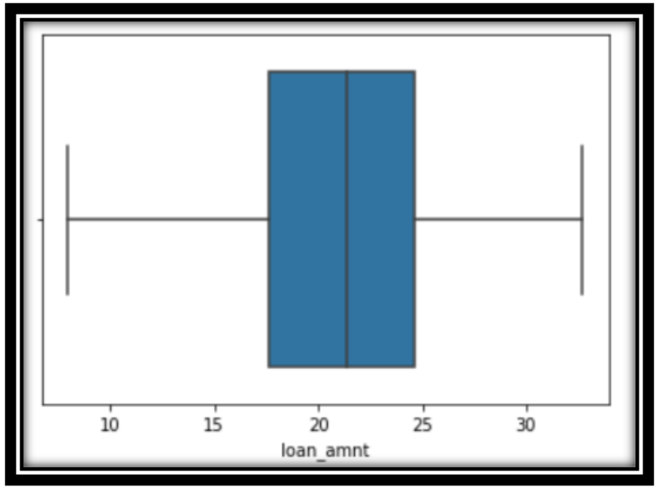
The analysis for uni variate is mostly done to identify outliers/ extreme values for all the columns identified. This is done to make sure that the analysis is not affected by those extreme value on the loan status.

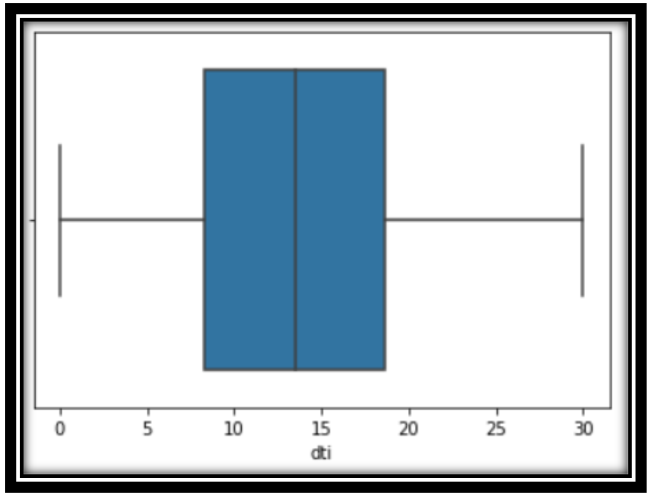
The outliers are treated with 2 techniques.

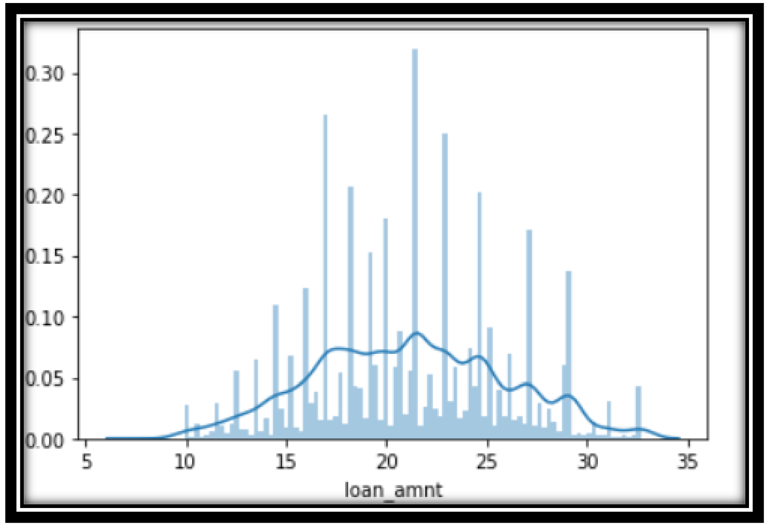
1. Scaling the value in the column.
2. Use Inter quartile technique to remove the extreme values.

Providing the below box plot for 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.







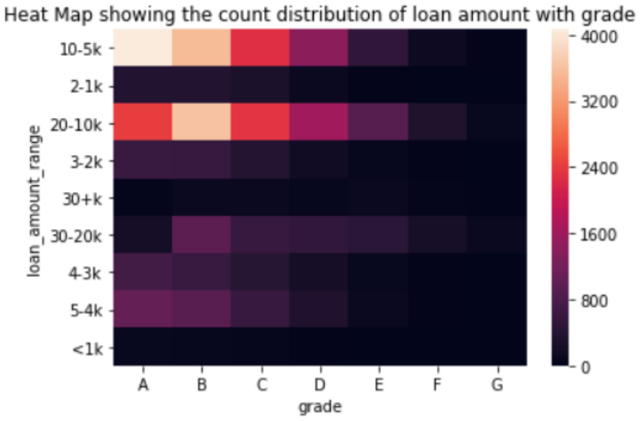


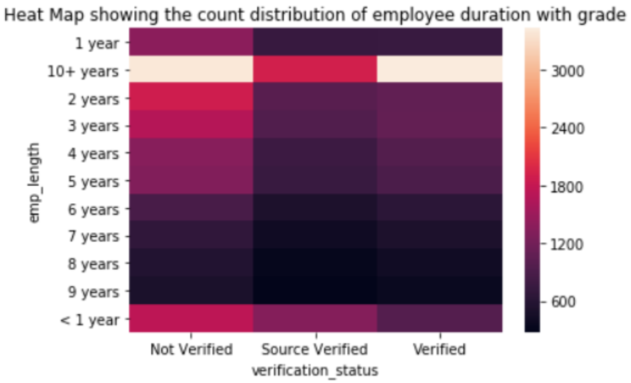
Bi-Variate Analysis – Between Categorical Variables

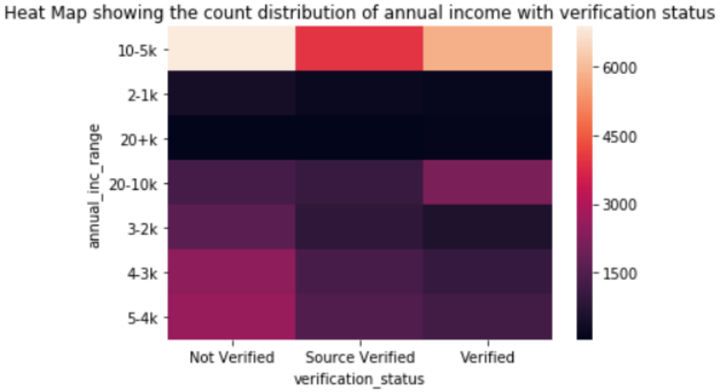
Analysis was performed on variables which have categorical values. Below are the major take away points.

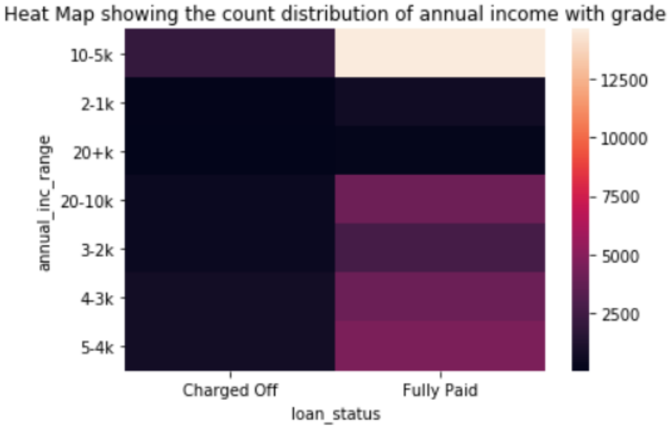
Also slide contains few heat maps to provide us the insight of data.

1. Majority are provided for loan amount between 10000 to 5000 with grade assigned as A.
2. Based on the count we see that loans are charged off for income range between 10000 to 5000.
3. Most loans provided with the borrows annual income between 10000 and 5000 are not verified.
4. Most of the borrowers having employment experience more than 10 plus years are verified.
5. Most of the borrowers loan requirements are ranging from 5000 to 20000.
6. 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.





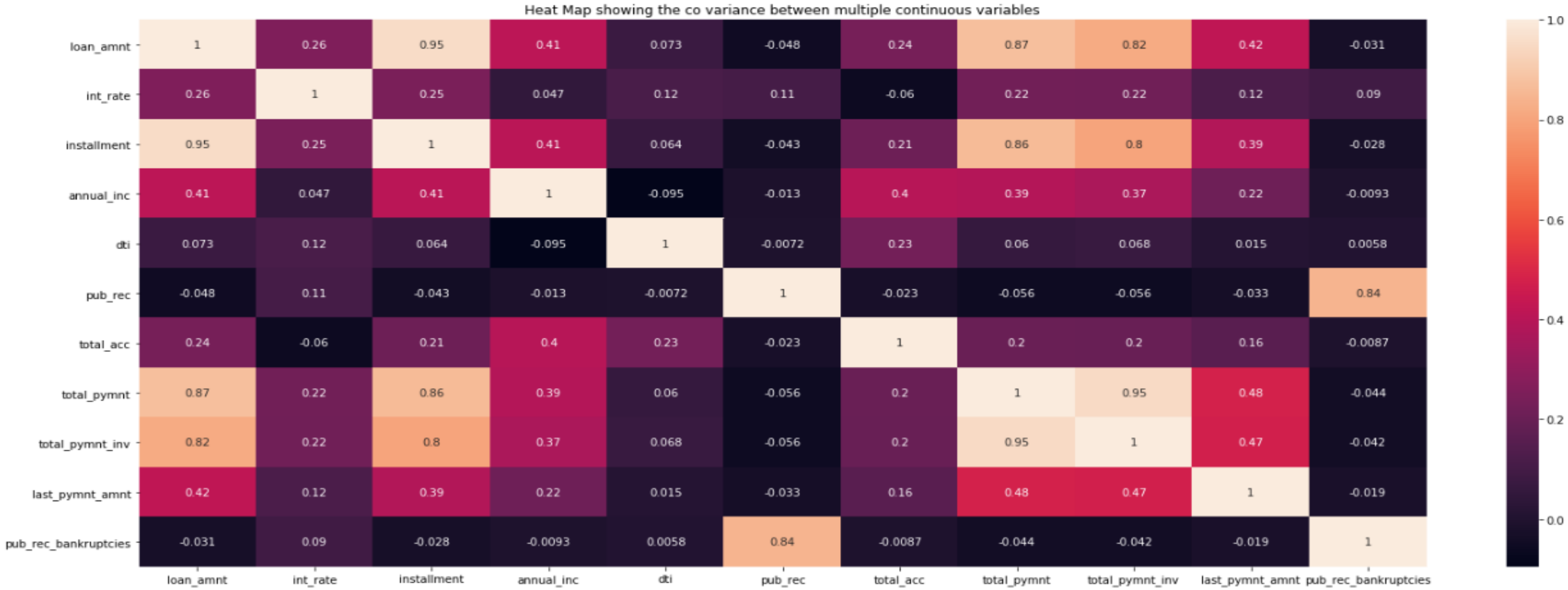




Bi-Variate Analysis – Between Continuous Variables

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

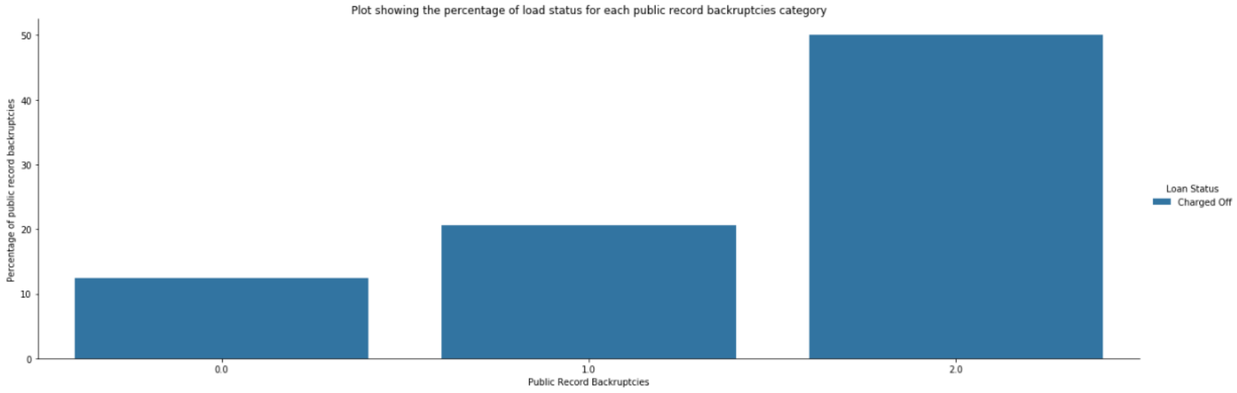
1. Total payment and total payment inv column has a clear positive correlation hence using any one of the variable is sufficient for analysis.
2. Loan amount and instalment column has a clear positive correlation hence using any one of the variable is sufficient for analysis.
3. Annual income and dti has almost no correlation, hence both can be used for analysis separately which can provide us more insights.
4. 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.

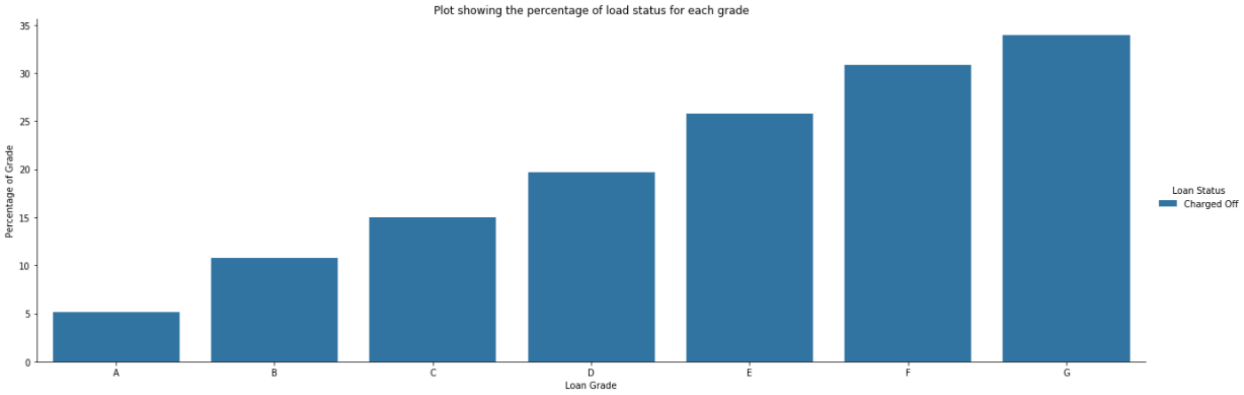

Heat Map showing the co variance between multiple continuous variables

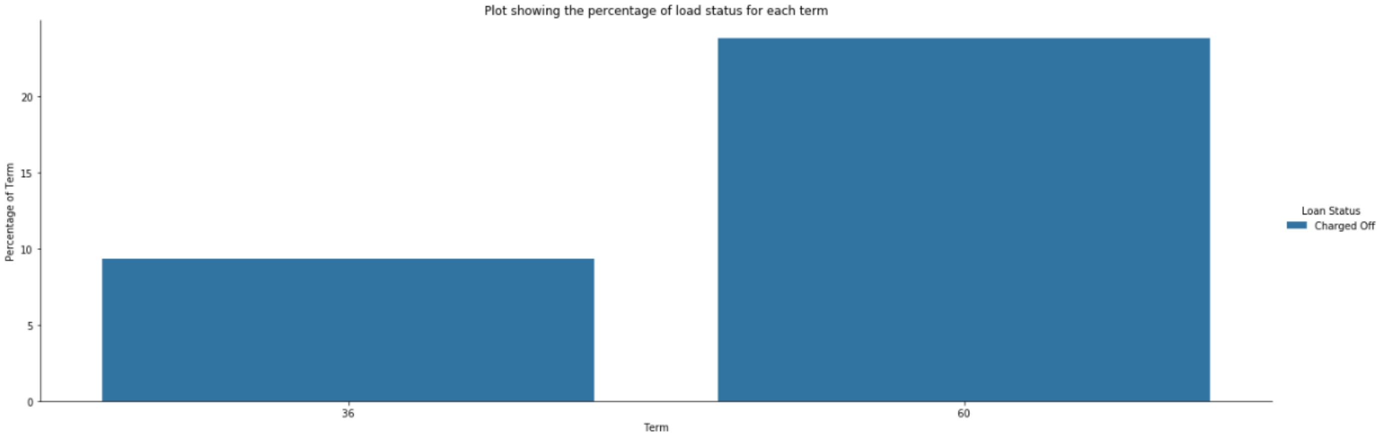
	loan_amnt	int_rate	installment	annual_inc	dti	pub_rec	total_acc	total_pymnt	total_pymnt_inv	last_pymnt_amnt	pub_rec_bankruptcies
loan_amnt	1	0.26	0.95	0.41	0.073	-0.048	0.24	0.87	0.82	0.42	-0.031
int_rate	0.26	1	0.25	0.047	0.12	0.11	-0.06	0.22	0.22	0.12	0.09
installment	0.95	0.25	1	0.41	0.064	-0.043	0.21	0.86	0.8	0.39	-0.028
annual_inc	0.41	0.047	0.41	1	-0.095	-0.013	0.4	0.39	0.37	0.22	-0.0093
dti	0.073	0.12	0.064	-0.095	1	-0.0072	0.23	0.06	0.068	0.015	0.0058
pub_rec	-0.048	0.11	-0.043	-0.013	-0.0072	1	-0.023	-0.056	-0.056	-0.033	0.84
total_acc	0.24	-0.06	0.21	0.4	0.23	-0.023	1	0.2	0.2	0.16	-0.0087
total_pymnt	0.87	0.22	0.86	0.39	0.06	-0.056	0.2	1	0.95	0.48	-0.044
total_pymnt_inv	0.82	0.22	0.8	0.37	0.068	-0.056	0.2	0.95	1	0.47	-0.042
last_pymnt_amnt	0.42	0.12	0.39	0.22	0.015	-0.033	0.16	0.48	0.47	1	-0.019
pub_rec_bankruptcies	-0.031	0.09	-0.028	-0.0093	0.0058	0.84	-0.0087	-0.044	-0.042	-0.019	1

Bi-Variate Analysis – Between Continuous Variables

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


Plot showing the percentage of load status for each public record backruptcies category


Plot showing the percentage of load status for each grade

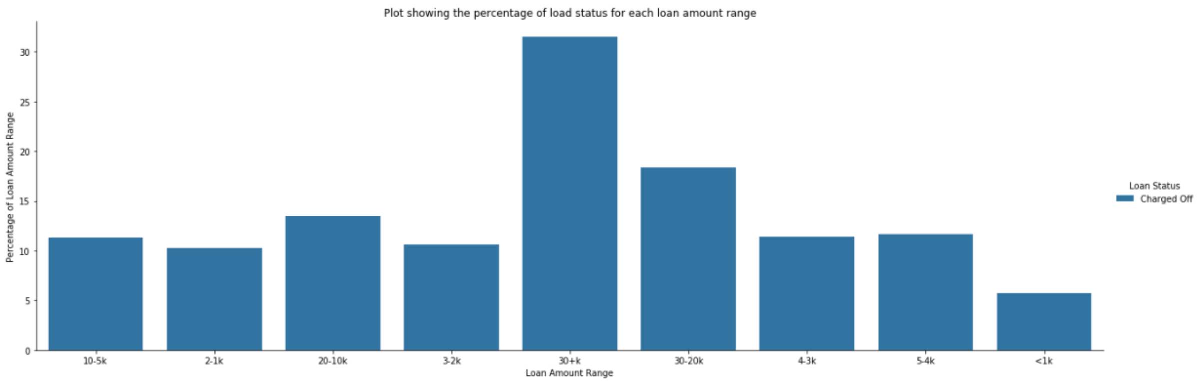

Plot showing the percentage of load status for each term

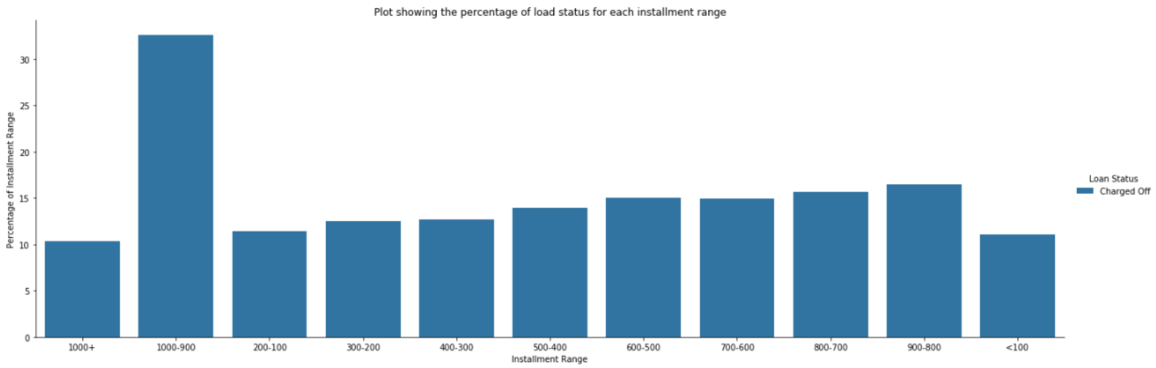
Below are the major take away from the plots provided in the slide:

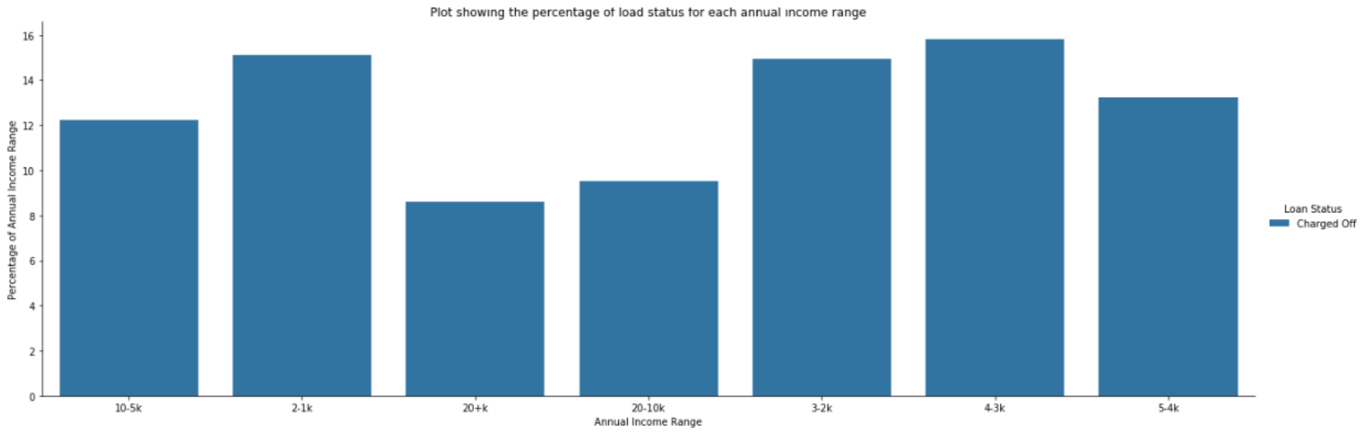
1. We see that higher bankruptcy record will lead to higher charge off
2. Longer the long duration higher is the possibility for charge off
3. Also we can see that higher the grade higher is the possibility of charge off. This in turn reveal that higher the interest rate higher is the possibility of charge off.

Bi-Variate Analysis – Between Continuous Variables

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


Plot showing the percentage of load status for each loan amount range


Plot showing the percentage of load status for each installment range

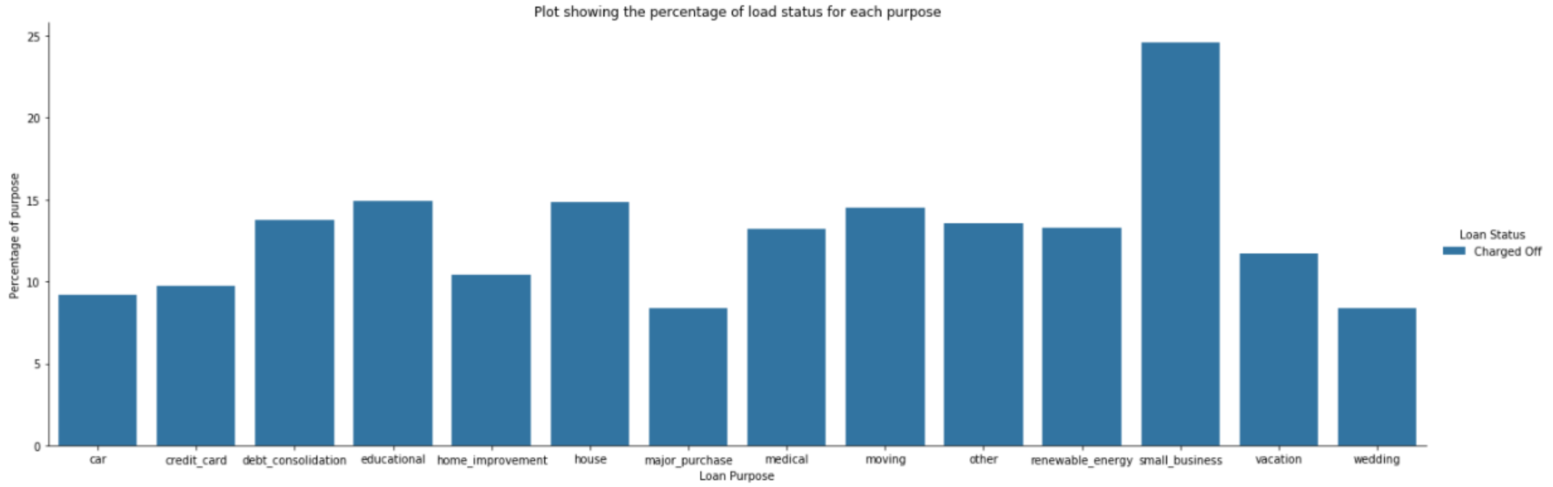

Plot showing the percentage of load status for each annual income range

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

1. We see that higher loan amount will lead to higher charge off
2. Higher the instalment then higher the charge off percentage
3. Lower income group have higher charge off hence falls into defaulter group.

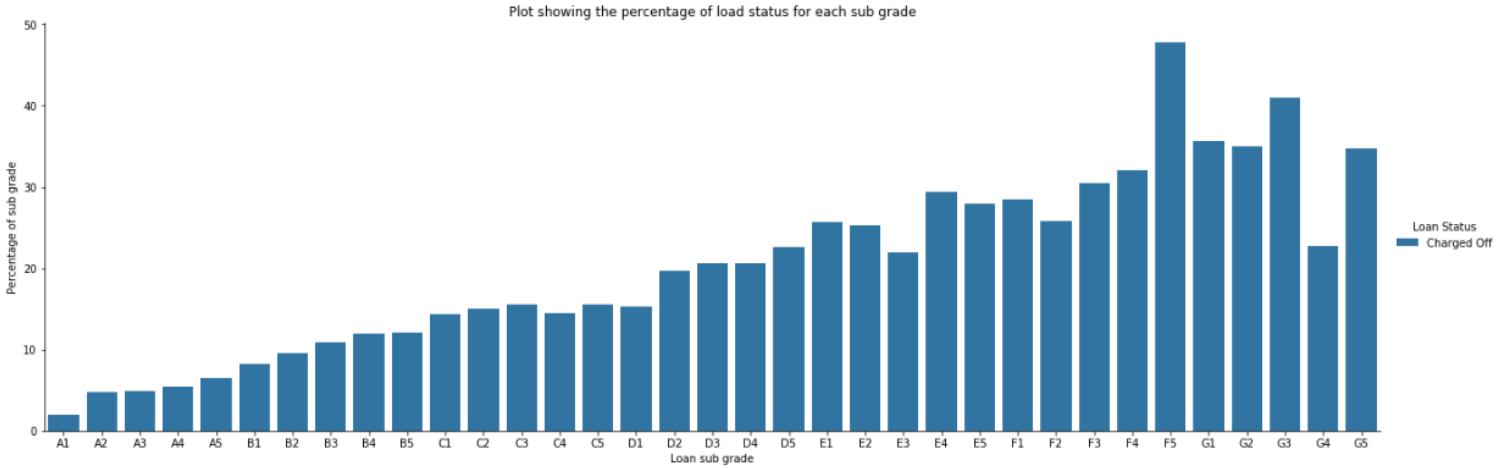
Bi-Variate Analysis – Between Continuous Variables

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



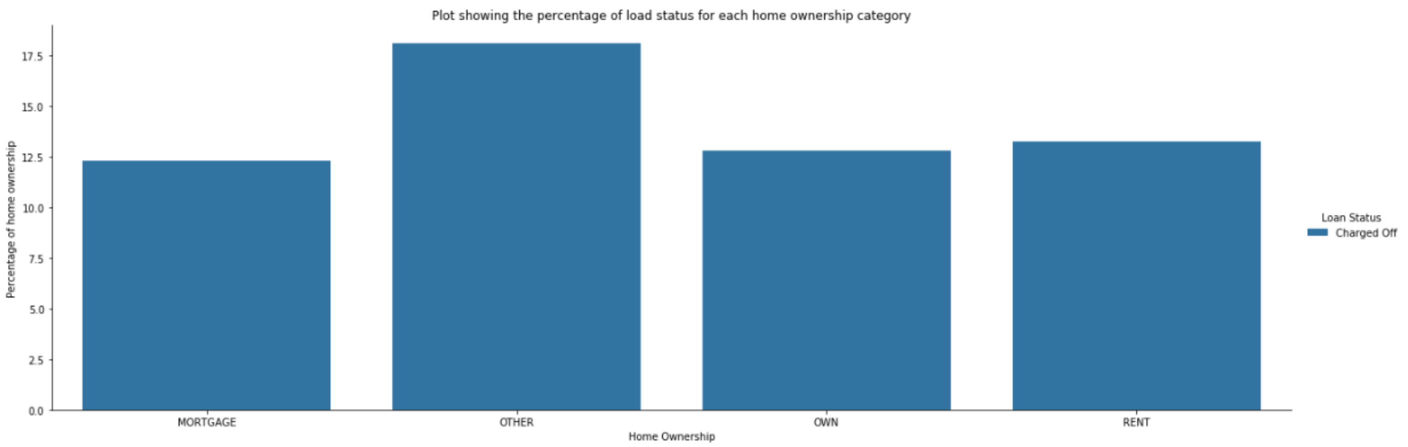
Below are the major take away from the plots provided in the slide:

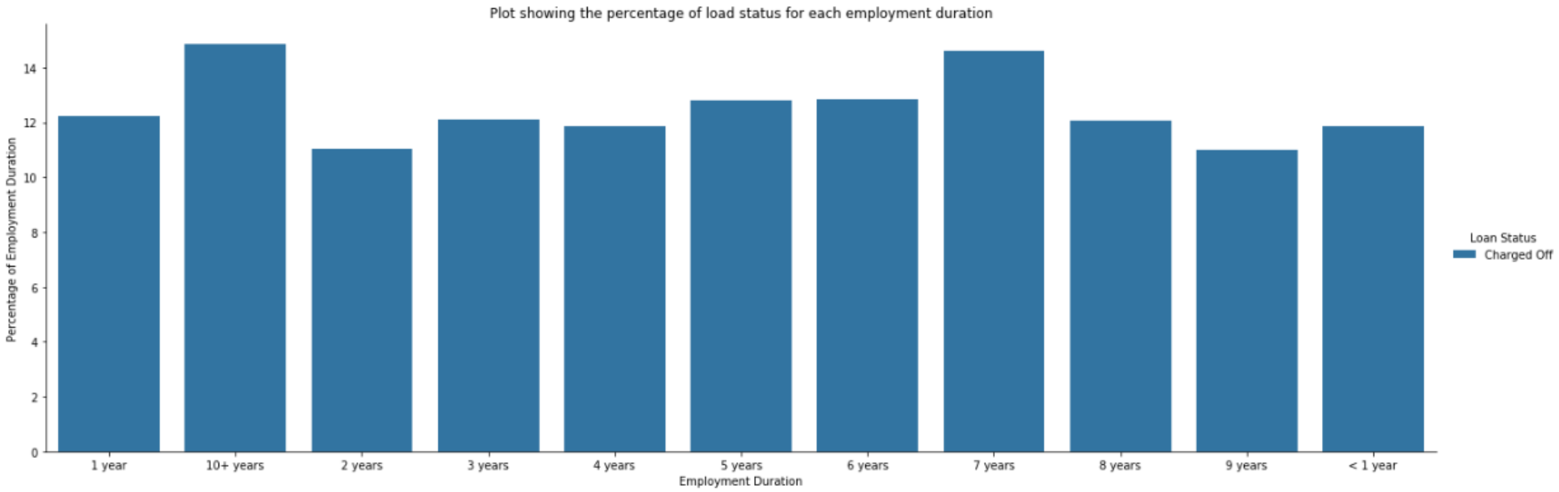
1. Small Business group mostly end up becoming defaulters.
2. As per the previous slide we found that higher the grade higher is the chance of charge off. Same is the case of sub grade as well.



Bi-Variate Analysis – Between Continuous Variables

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



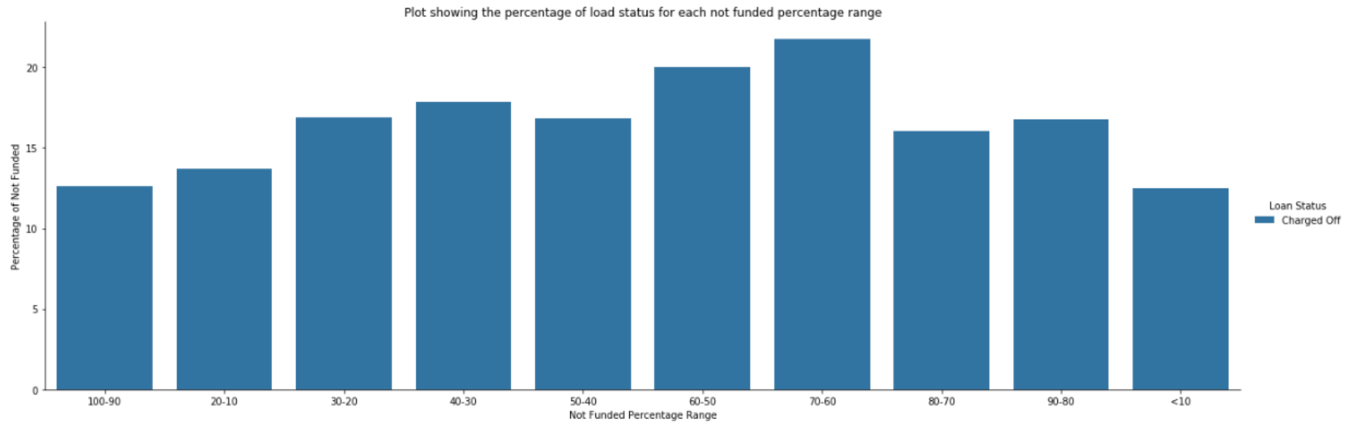


Below are the major take away from the plots provided in the slide:

1. Under home ownership other category has higher rate of charge off.
2. High employment experience borrower tend to default at high rate.

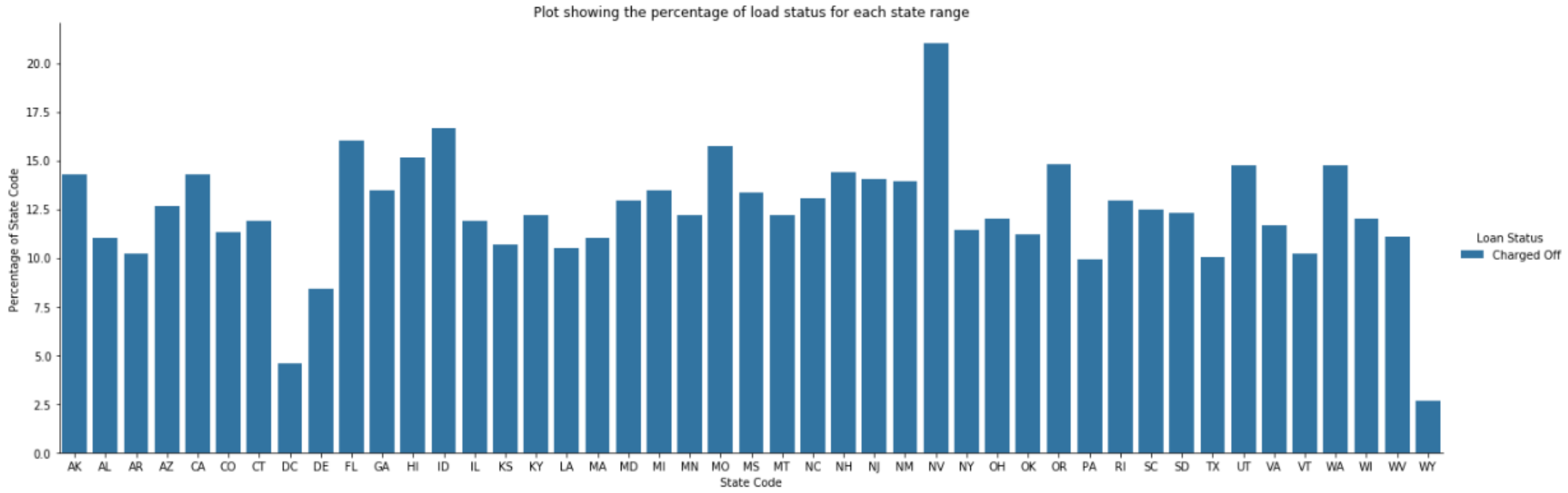
Bi-Variate Analysis – Between Continuous Variables

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



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

1. Borrowers who are partially funded are the once who default more.
2. NV state has higher charge off.



Recommendation – Lending club

Below is the conclusion:

1. The top variables which has the impact over the loan status are:
   - Public record Bankruptcies
   - Grade
   - Loan Amount
   - Loan Duration
2. The Lending club has to make sure that proper verification has to be done before sanctioning the loan.
3. Based on the previous records the loan amount has to be sanctioned instead of looking the annual income.

Below are the reason for the conclusion:

1. The reason to select these variables are that we found a large variation between different category.
2. Higher the variation indicates that there are many insights to the data.