Lending Club Case Study

EPGML C55 July 2023

Group Members:

Syed Abdul Rahim
Leena Orpe

Introduction

Lending Club is a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- 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

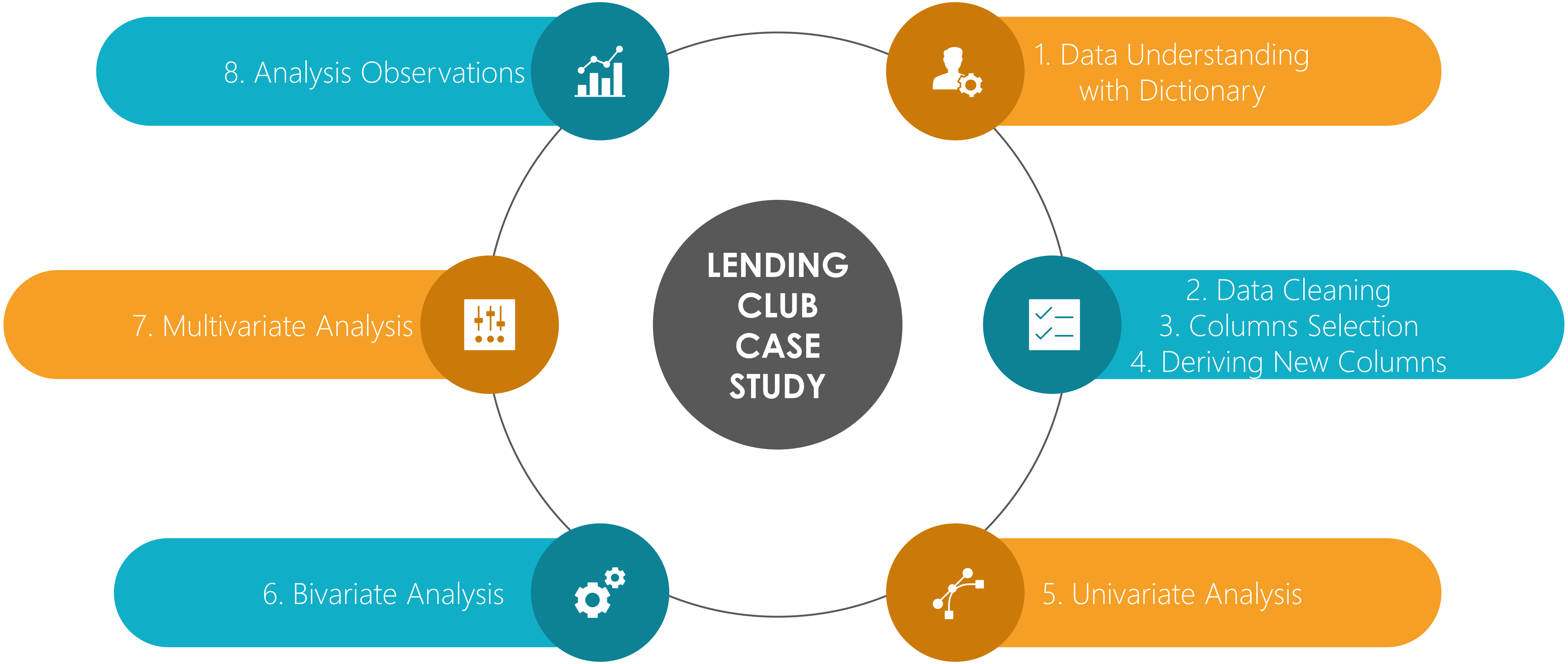
The data we are analyzing here contains information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Objective:

To uncover critical insights into loan default risk factors, we will use Exploratory Data Analysis (EDA) to understand how **consumer attributes** and **loan attributes** influence the tendency to default.

Along side exploring the key drivers behind loan defaults, evaluating historical borrower behaviors, and identifying actionable strategies for risk assessment and portfolio management, through data-driven visualizations and in-depth exploration. And to empower financial professionals with the knowledge needed to make informed decisions and reduce credit losses, for both Rejectable and Approvable Loans.

Analysis Methodology



LENDING CLUB CASE STUDY

1. Data Understanding with Dictionary

2. Data Cleaning
3. Columns Selection
4. Deriving New Columns

5. Univariate Analysis

6. Bivariate Analysis

7. Multivariate Analysis

8. Analysis Observations

Analysis Steps

DATA UNDERSTANDING

Going through the Data Dictionary and understand the columns and expected values in them, gives us a broader view of which columns to choose for this analysis

DATA CLEANING

Dropping columns with unique values

Dropping Rows with 30% threshold for NA or NULL values

Filling columns with mode or mean or 0 for NA or NULL

DERIVING COLUMNS

Deriving Year / Month from Date Columns

Binning certain columns for Graded Analysis

EDA - VARIATE ANALYSIS

Univariate Analysis on Continuous & Categorical Columns

Bivariate Analysis on combination

Multivariate Analysis with cross combination

ANALYSIS OBSERVATIONS

Result of all insights pooled into an observation for the client as an answer to this business problem

Data Understanding

Understanding the structure and meaning of each column is paramount in conducting meaningful analysis. To aid us in this endeavor, we have been provided with metadata, which includes column names, data types, and brief descriptions, to better understand the columns and the expected data in it. In particular we can deduce the business aspect of the data and its importance for the EDA which need to perform. Which is the starting point for any analysis, as too much data or unnecessary scope can ruin the analysis.

1. Identifying columns which are required primarily for the analysis. Stressing more on data cleaning activity for these columns.
2. Identifying columns which can be straightaway dropped, like Id, Member Id, etc, which have no impact on analysis.
3. Identifying columns which needs to be derived before using, like Year and Month from Issue Date, stripping % from Interest Rate, etc.
4. Identifying null valued columns with certain threshold to be dropped
5. Identifying null valued rows with certain threshold to be dropped
6. Identifying columns which have similar values, which needs a correlation, to see if a primary column identified from above can replace the presence of other columns, like Funded Amount and Funded Amount by Investor can be dropped and Loan Amount can be used in that place.

Data Cleaning

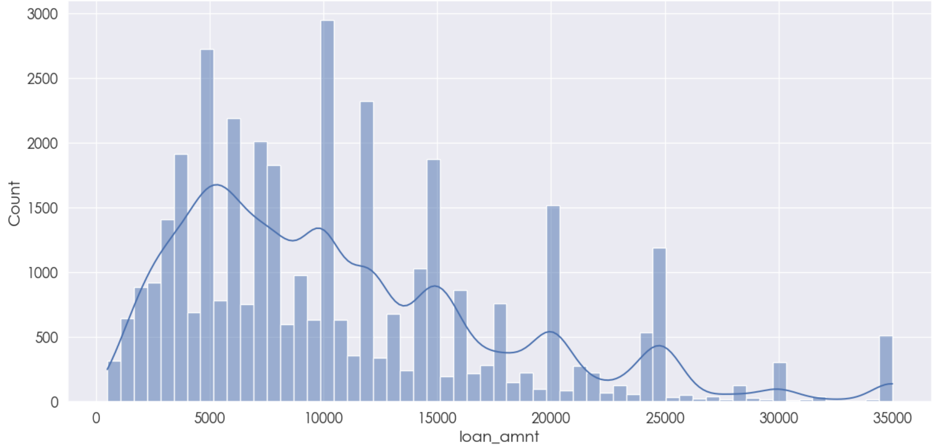
1. Find columns with null values more than 30% and drop
2. Find rows with null values more than 30% and drop
3. Find Columns with Unique values and drop
4. Replace Employment Length missing values with mode value
5. With box plot identify outliers and drop rows where the values is greater than (75th Quantile + (1.5 x 75th Quantile - 25th Quantile), Example: Annual Income
6. Mark columns as required basing on Data Dictionary pertaining to this case and drop all else columns

Deriving Columns

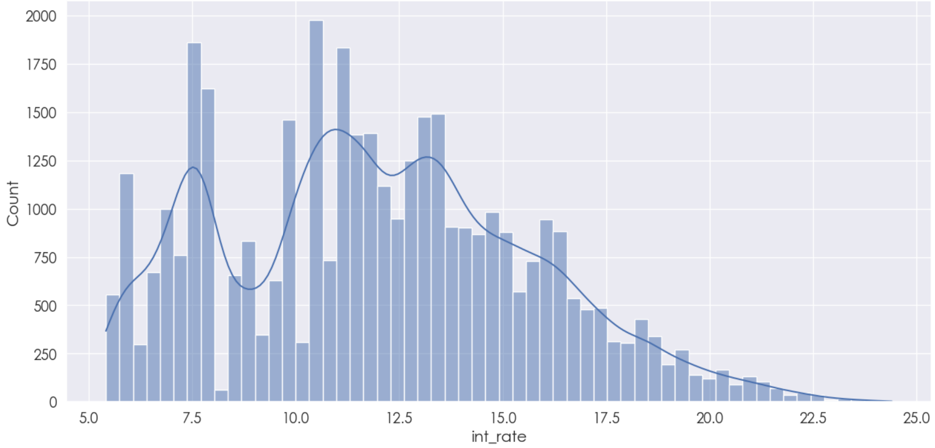
Derive columns to make them numeric, appropriate for analysis:

1. Interest rate column from existing one by stripping % from it
2. Revolving line utilization rate column from existing one by stripping % from it
3. Issued Year from Issue Date column
4. Issued Month from Issue Date column
5. Finally dropping the column which are remanent because of derivation
6. Convert following column to numeric
   'loan_amnt','int_rate','installment','annual_inc','dti','delinq_2yrs','open_acc', 'pub_rec', 'revol_util', 'total_acc', 'issue_d_year'
7. Binning columns for proper segmented analysis, example: interest rate

Univariate Analysis – 1/2

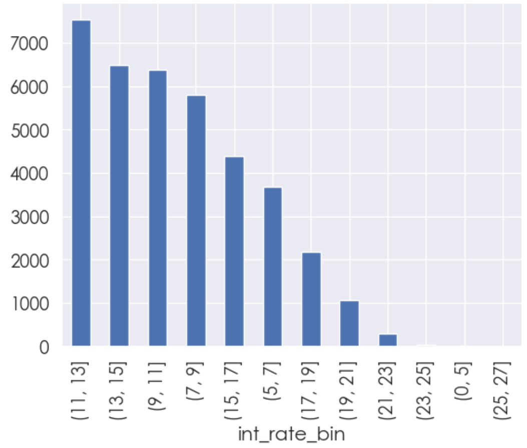


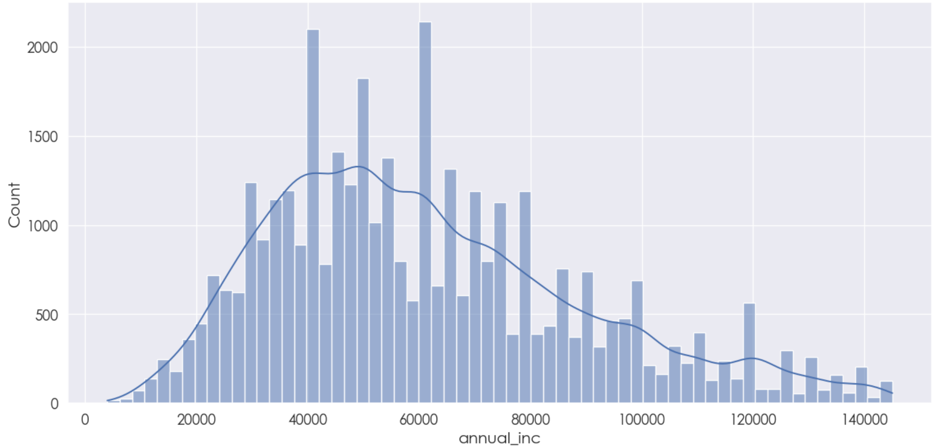
loan_amnt vs count: Most of the loans range from 4000 - 20000



int_rate or int_rate_bin vs count:

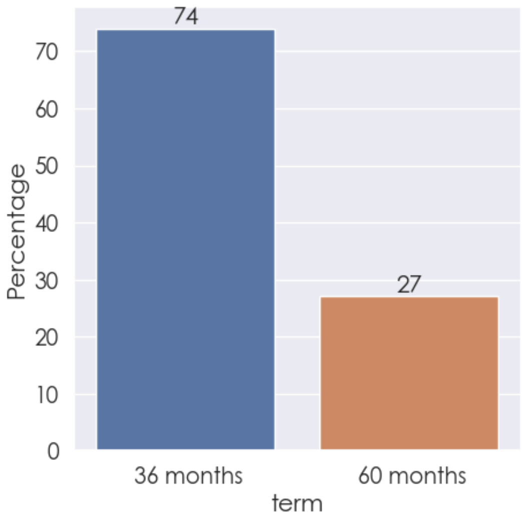
Most of the interest rates are distributed between 5% to 17%





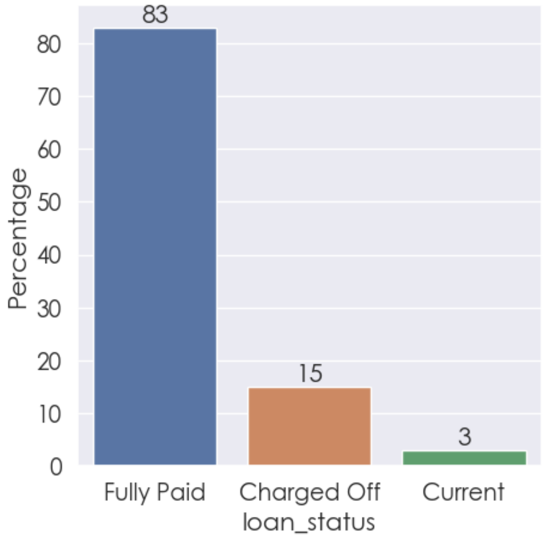
annual_inc vs count: Most of the applications income lies between 30000 to 80000

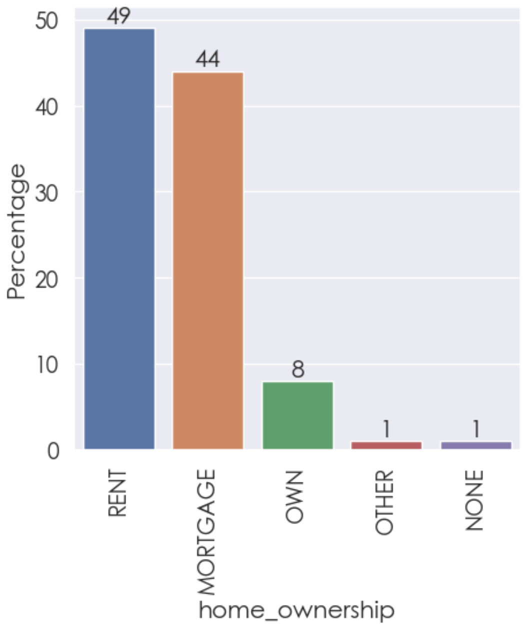
Univariate Analysis – 2/2



term by percentage:
Most of the loans are issued for 36 months (74%)

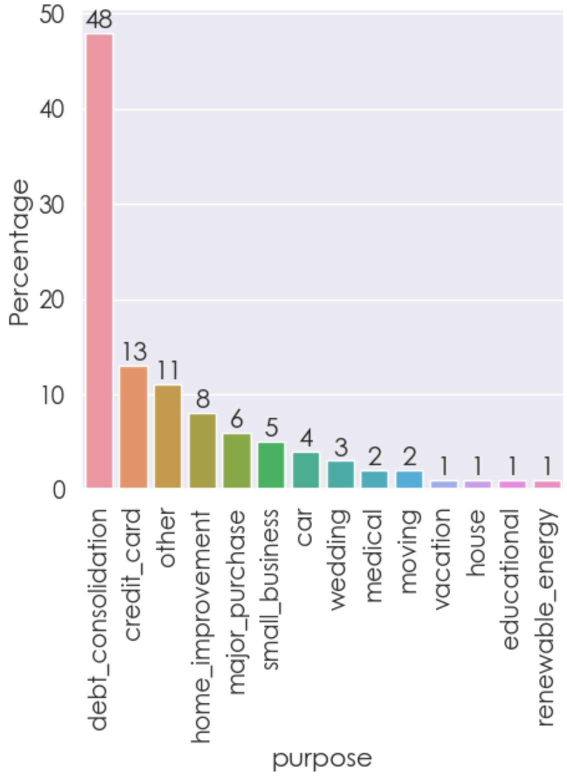
loan_status by percentage:
Fully Paid: 83%, Charged Off: 15%, Current: 3%



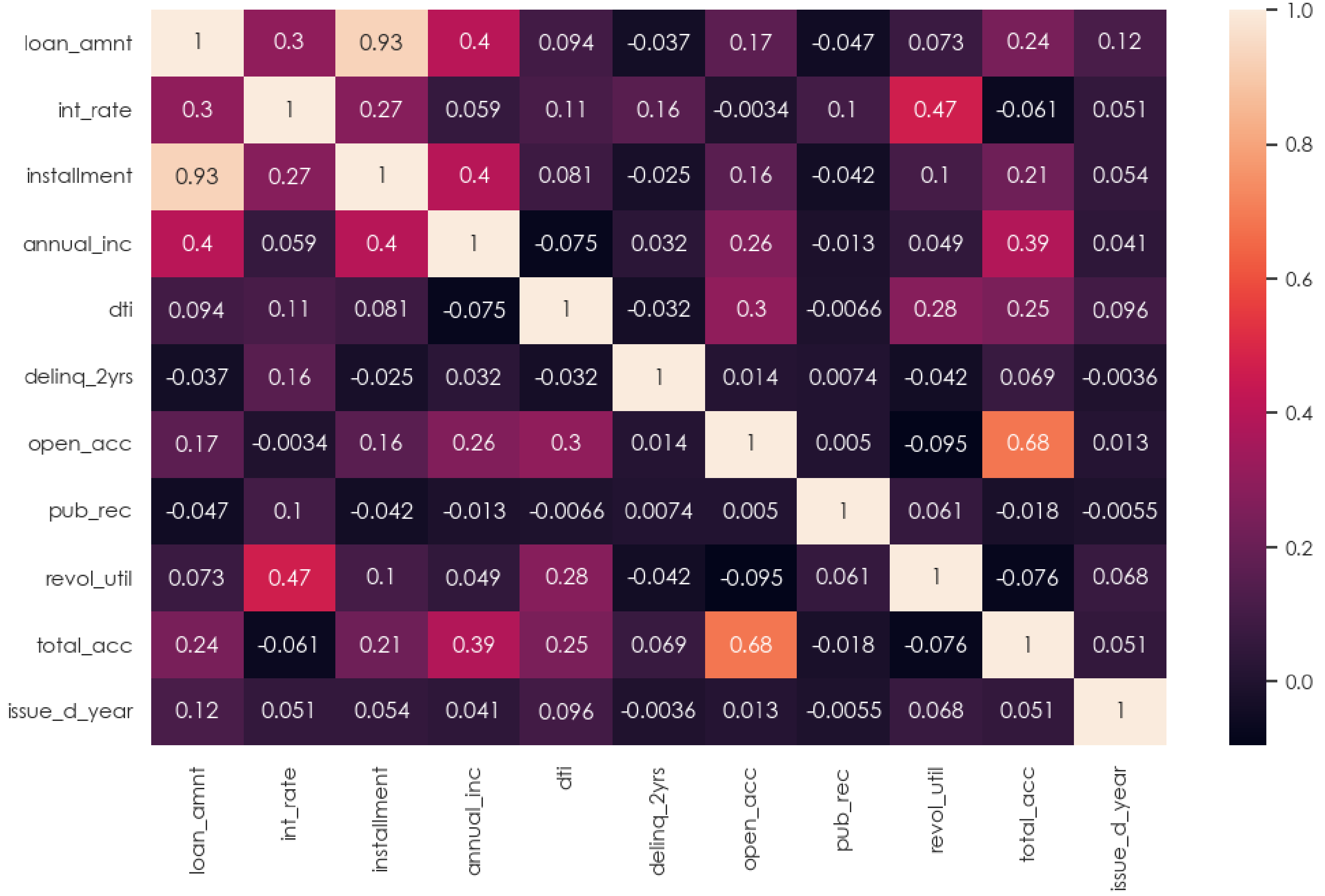


home_ownership by percentage:
Most of the owners are either on Rent (49%) or Mortgage (44%)

purpose by percentage:
61% applied for paying other loans, i.e., debit consolidation (48%) / credit card (13%)



Bivariate Analysis – 1/4

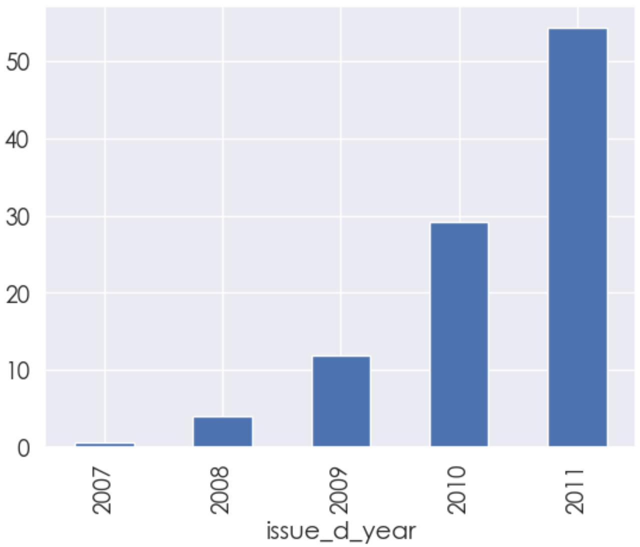


	loan_amnt	int_rate	installment	annual_inc	dti	delinq_2yrs	open_acc	pub_rec	revol_util	total_acc	issue_d_year
loan_amnt	1	0.3	0.93	0.4	0.094	-0.037	0.17	-0.047	0.073	0.24	0.12
int_rate	0.3	1	0.27	0.059	0.11	0.16	-0.0034	0.1	0.47	-0.061	0.051
installment	0.93	0.27	1	0.4	0.081	-0.025	0.16	-0.042	0.1	0.21	0.054
annual_inc	0.4	0.059	0.4	1	-0.075	0.032	0.26	-0.013	0.049	0.39	0.041
dti	0.094	0.11	0.081	-0.075	1	-0.032	0.3	-0.0066	0.28	0.25	0.096
delinq_2yrs	-0.037	0.16	-0.025	0.032	-0.032	1	0.014	0.0074	-0.042	0.069	-0.0036
open_acc	0.17	-0.0034	0.16	0.26	0.3	0.014	1	0.005	-0.095	0.68	0.013
pub_rec	-0.047	0.1	-0.042	-0.013	-0.0066	0.0074	0.005	1	0.061	-0.018	-0.0055
revol_util	0.073	0.47	0.1	0.049	0.28	-0.042	-0.095	0.061	1	-0.076	0.068
total_acc	0.24	-0.061	0.21	0.39	0.25	0.069	0.68	-0.018	-0.076	1	0.051
issue_d_year	0.12	0.051	0.054	0.041	0.096	-0.0036	0.013	-0.0055	0.068	0.051	1

Continuous Columns Correlation:

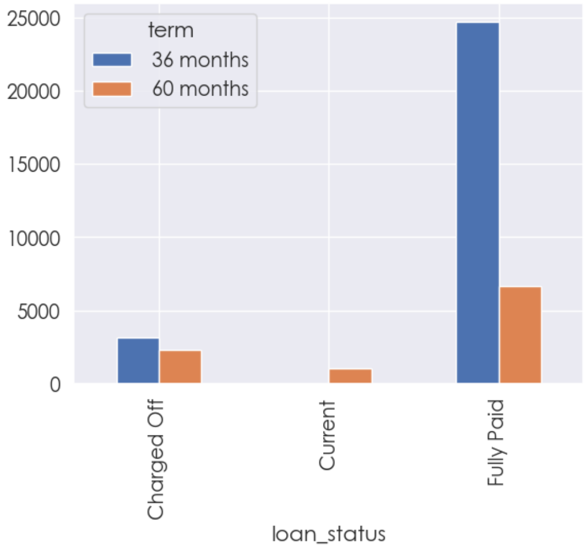
- Interest rate has positive correlation with loan amount
- Loan amount is highly correlated with installment
- Annual income has negative correlation with DTI
- DTI has small positive correlation with loan amount and installment
- Annual income has positive correlation with installment, loan amount
- Annual income has negative correlation with DTI

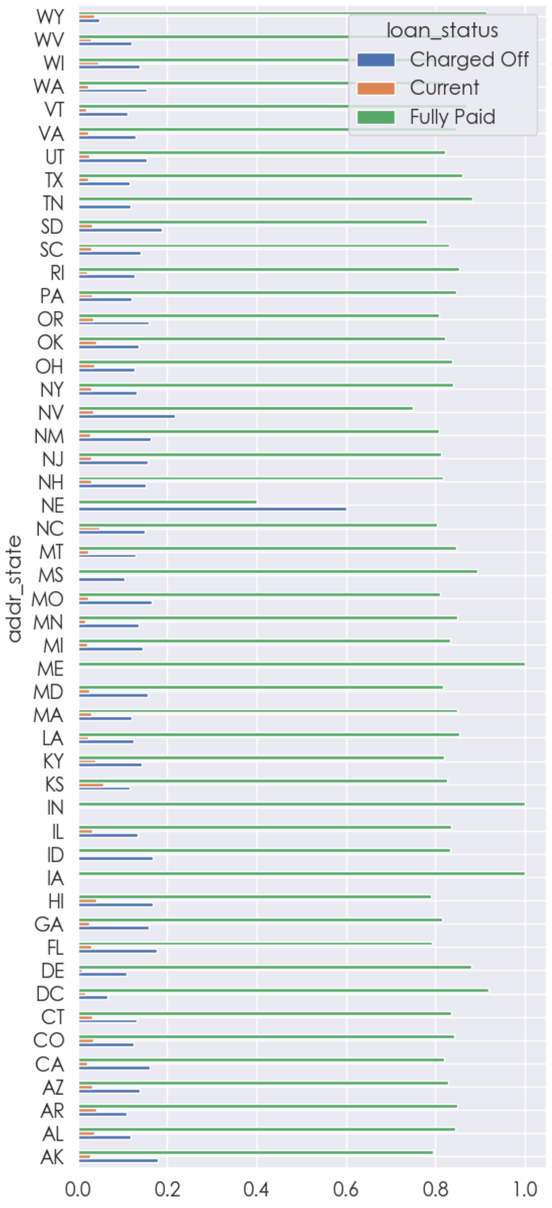
Bivariate Analysis – 2/4



issue_d_year by year: Loan applicants increases year after year, reaching over 55% in 2011

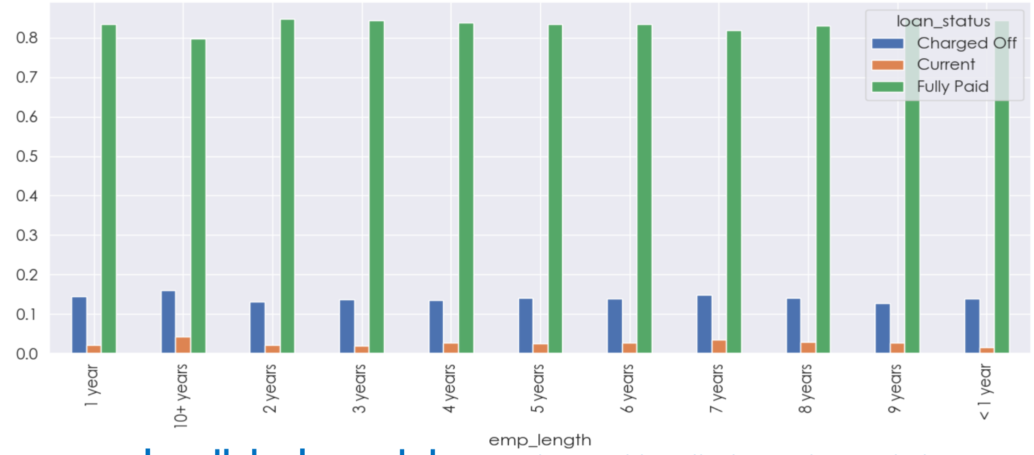
loan_status by term: Charged Off is slightly more for 36 months vs 60 months tenure
But the counter argument is significantly high too, as Fully Paid is much higher for 36 month term



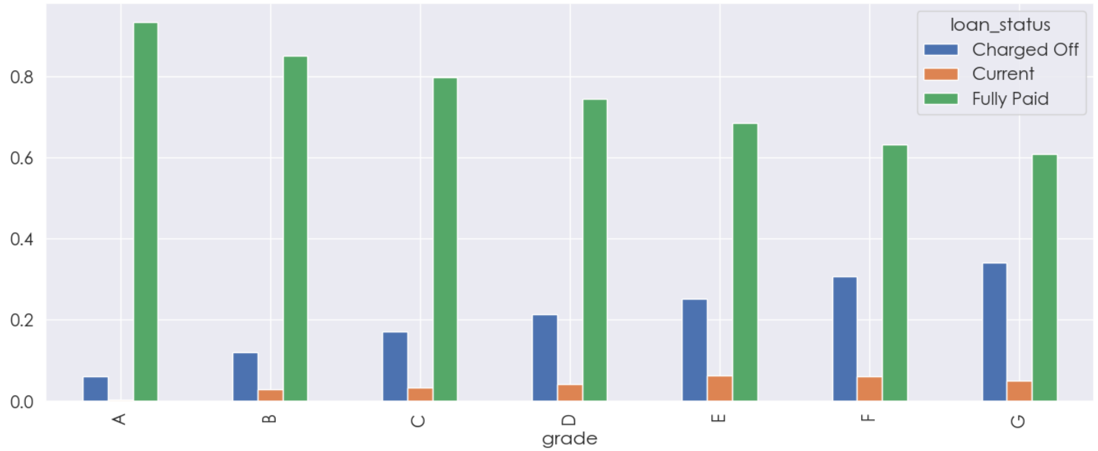


addr_state by loan_status: State NE (Nebraska) stands out as the highest

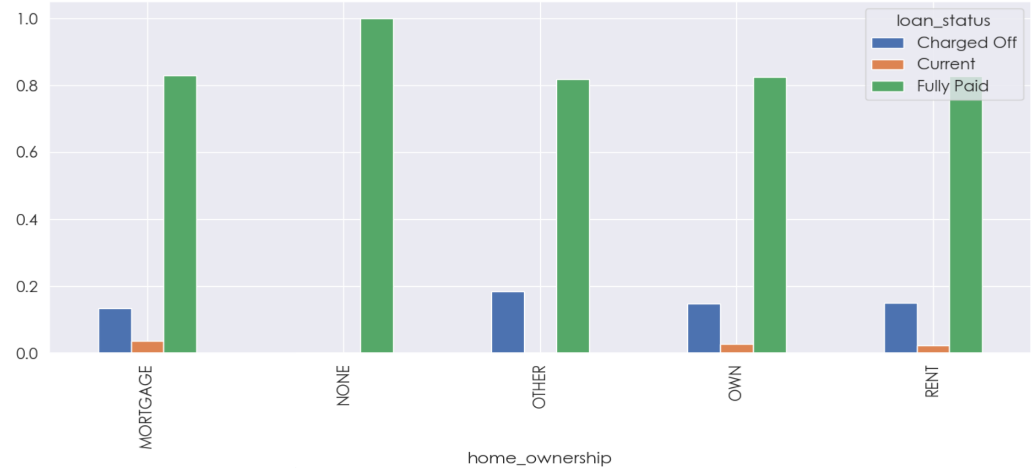
Bivariate Analysis – 3/4



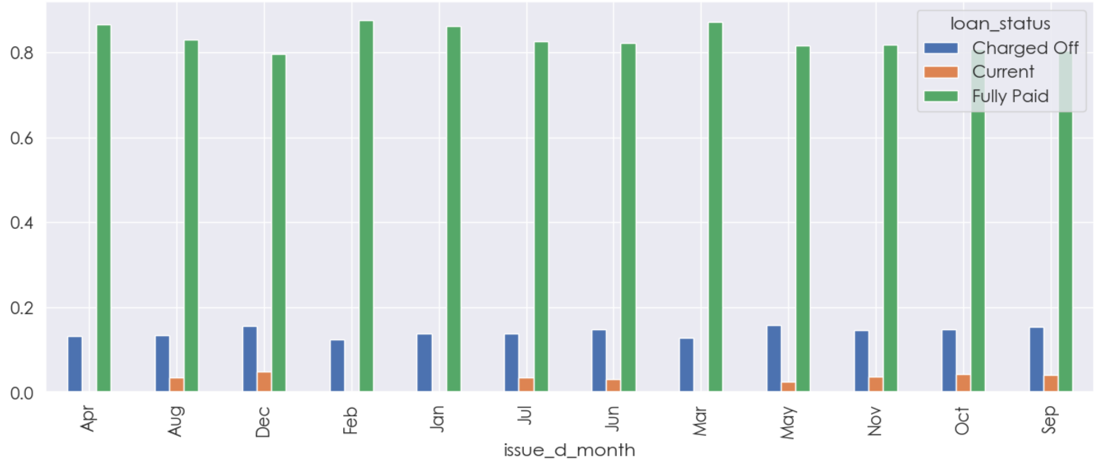
emp_length by loan_status: Employment length doesnot seem to have an influence



grade by loan_status: 'Charged Off' increases as grades go from A to G

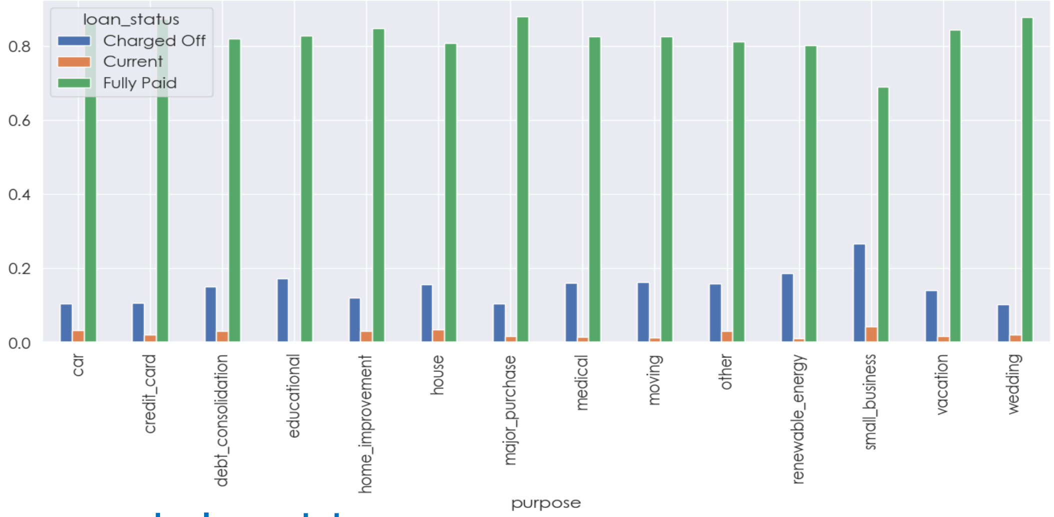


home_ownership by loan_status: Except for none, all other categories have similar rate

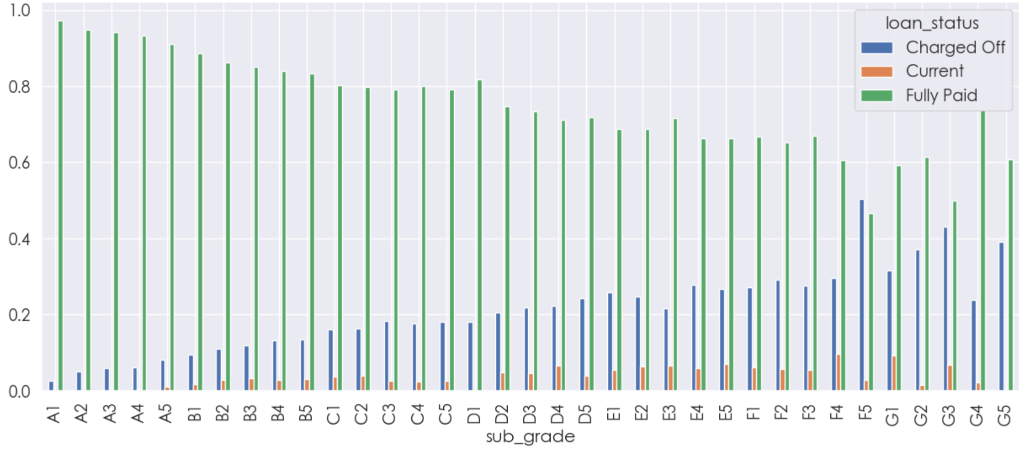


issue_d_month by loan_status: Occasionally a bit higher for some months, but not significantly different

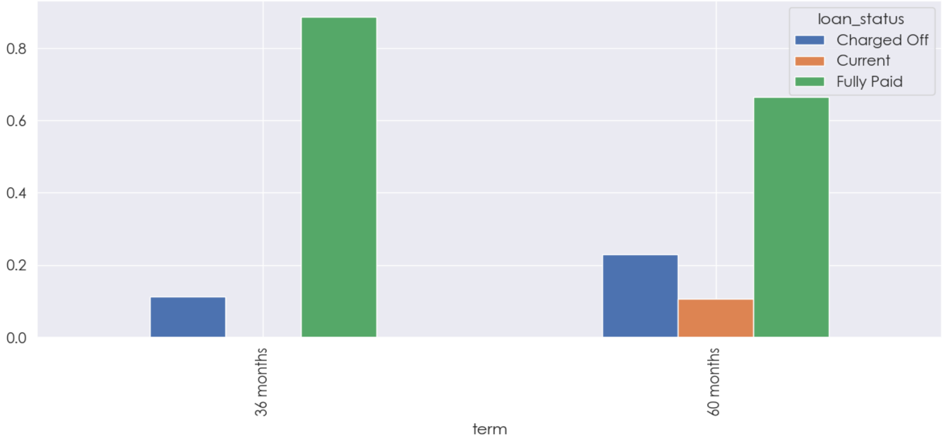
Bivariate Analysis – 4/4



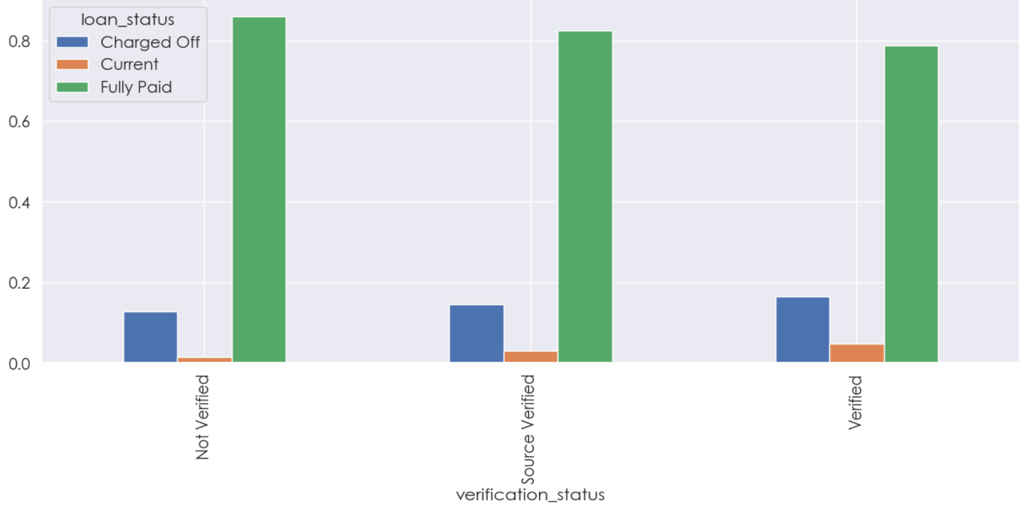
purpose by loan_status: Charged Off has significant increase for small_business



sub_grade by loan_status: Incremental trend moving from A to G, but significant for F5

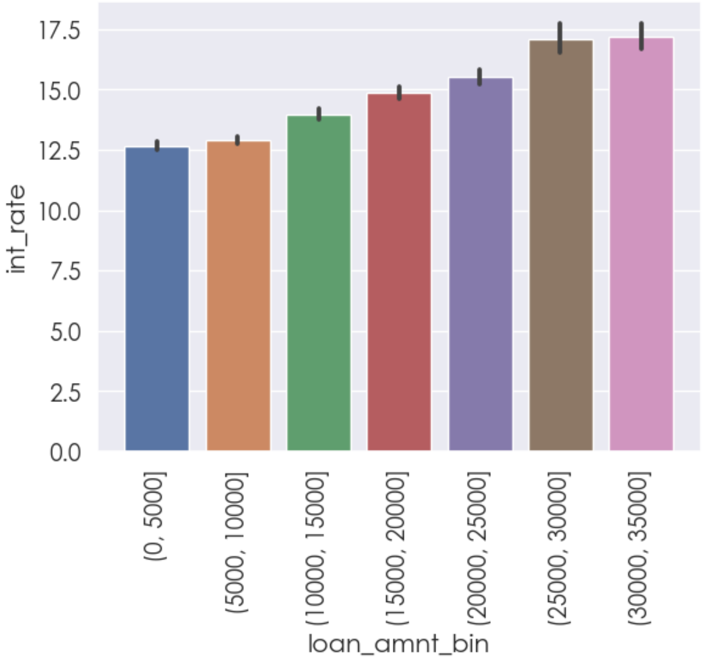


term by loan_status: Charged Off is higher for 60 months compared with 30 months



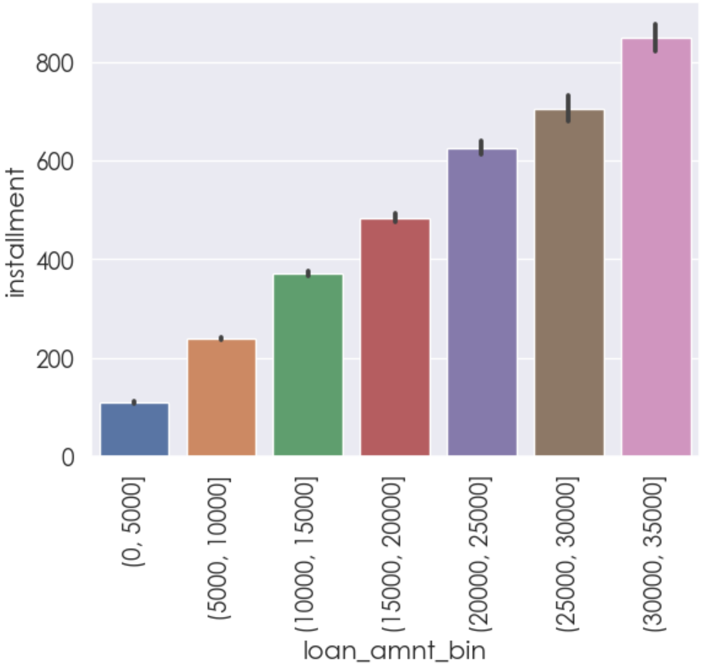
verification_status by loan_status: Ironically Verified has higher than Not Verified for Charged Off

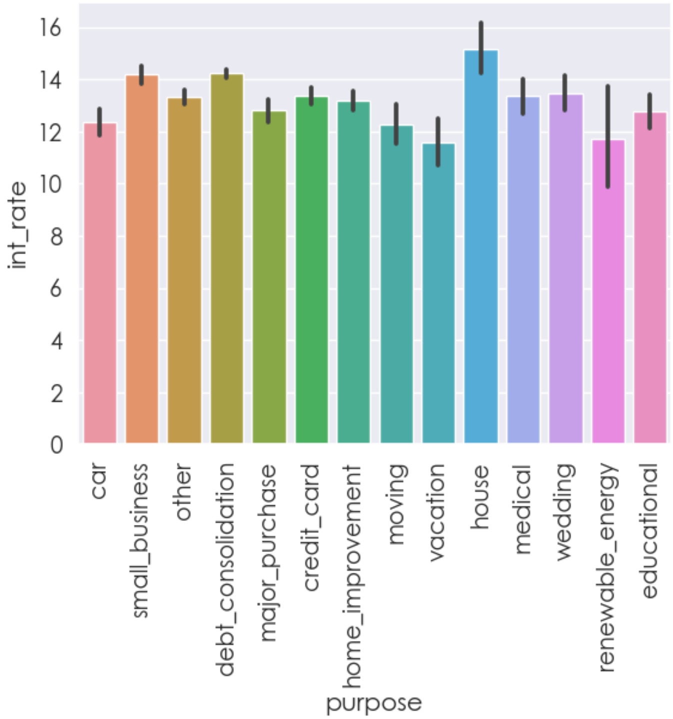
Multivariate Analysis – 1/2



loan_amnt_bin vs int_rate:
[data filtered on Charged Off]
Loans above 25000 and interest rate more than 15% have higher chance for Charged Off

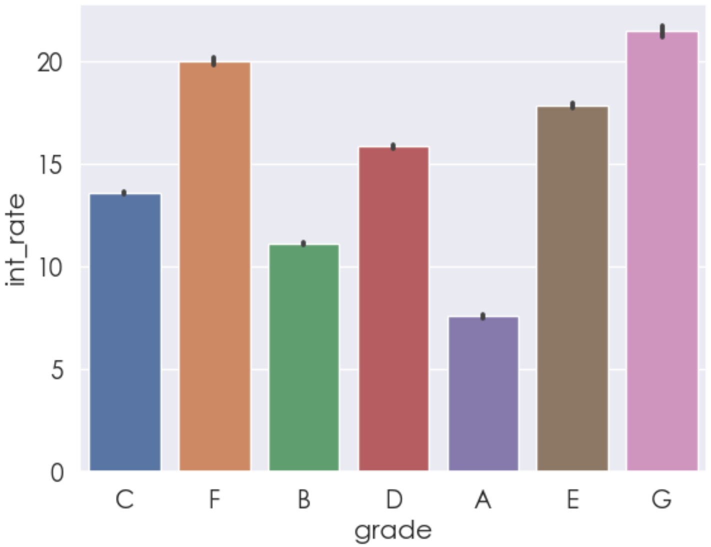
loan_amnt_bin vs installment:
[data filtered on Charged Off]
Loans above 30000 and installment above 800 have higher chance for Charged Off





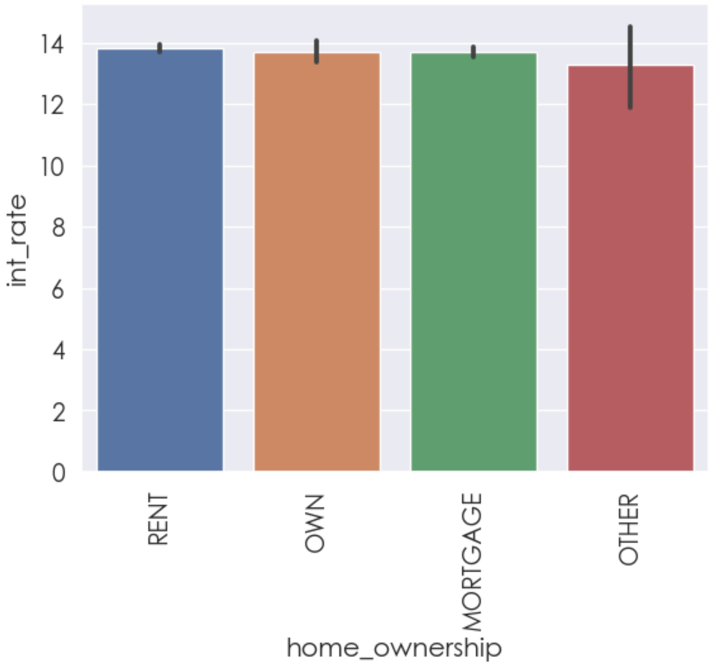
purpose vs int_rate:
[data filtered on Charged Off]
Loans applied for 'house' with Interest rate above 15% have higher chance for Charged Off

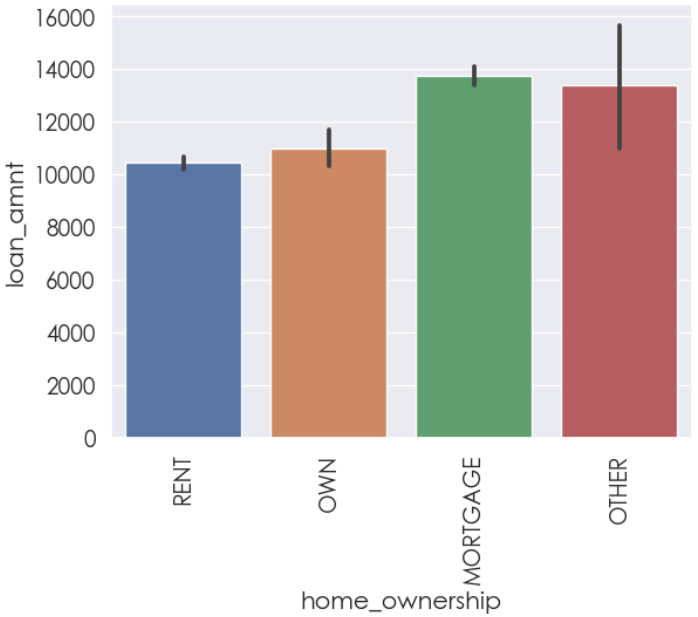
Multivariate Analysis – 2/2



grade vs int_rate:
[data filtered on Charged Off]
Loans under grade 'G' with Interest rate above 20% have higher chance for Charged Off

home_ownership vs int_rate:
[data filtered on Charged Off]
No significant impact basing on home ownership for Charged Off





home_ownership vs loan_amnt:
[data filtered on Charged Off]
Applicants on 'Mortgage' home and loan above 13000 have higher chance for Charged Off

Analysis Observations

Rejectable Loans:

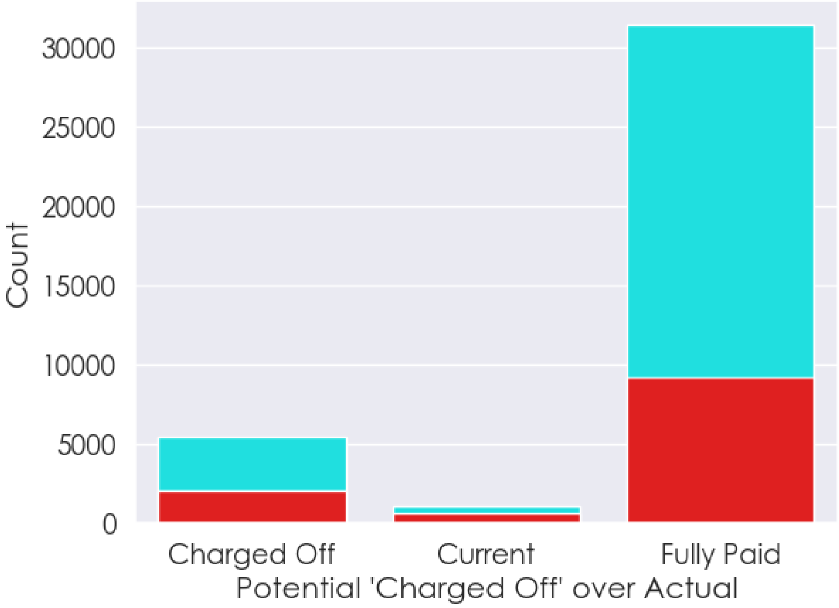
1. Loans above 25000 and interest rate more than 15% have higher chance for Charged Off
2. Loans above 30000 and installment above 800 have higher chance for Charged Off
3. Loans applied for 'house' with Interest rate above 15% have higher chance for Charged Off
4. Loans under grade 'G' with Interest rate above 20% have higher chance for Charged Off
5. Applicants on 'Mortgage' home and loan above 13000 have higher chance for Charged Off

Approvable Loans:

1. Loans below 25000, interest rate below 15% and not for 'house'
2. Loans below grade 'G' and interest rate below 20%
3. Loans below 13000 and not under 'Mortgage'

loan_status vs count:
overlayed potential 'Charged Off' on top of the actual values

Additional Observation: 'Current' has the potential of 'Charged Off' cases basing on the findings



Thank You