Lending Club – Case study

Submitted by

Naga Sai Gopi Krishna Lingamallu

Goutham Ganji

- ML C54 Course

Lending club – case study

Problem statement:

The company receives loan applications and the company like to identify patterns which indicate if a applicant is likely to default or not. Based on this decision denying the loan or reducing the amount lending the loan.

Understanding Data

Data understanding

- Broadly divide loan data into 3 categorical variables
  - Related to the applicant (id, member_id, zipcode, etc)
  - Loan characteristics (loan amount, purpose, loan status, interest rate etc)
  - Customer behaviour variables (Those which are generated after the loan is approved)
- The Loan_status becomes target variable for this case study
- The Loan status has 3 types of values, 'Fully Paid', 'Charged Off' and 'Current'. We will ignore 'Current' status as this is in progress and consider 'Fully Paid' and 'Charged Off' values. Here charged off means defaulters.
- Many columns are having null values which can be dropped
- There are some columns which may not be required for the analysis e.g. id, member id etc.

Data cleaning

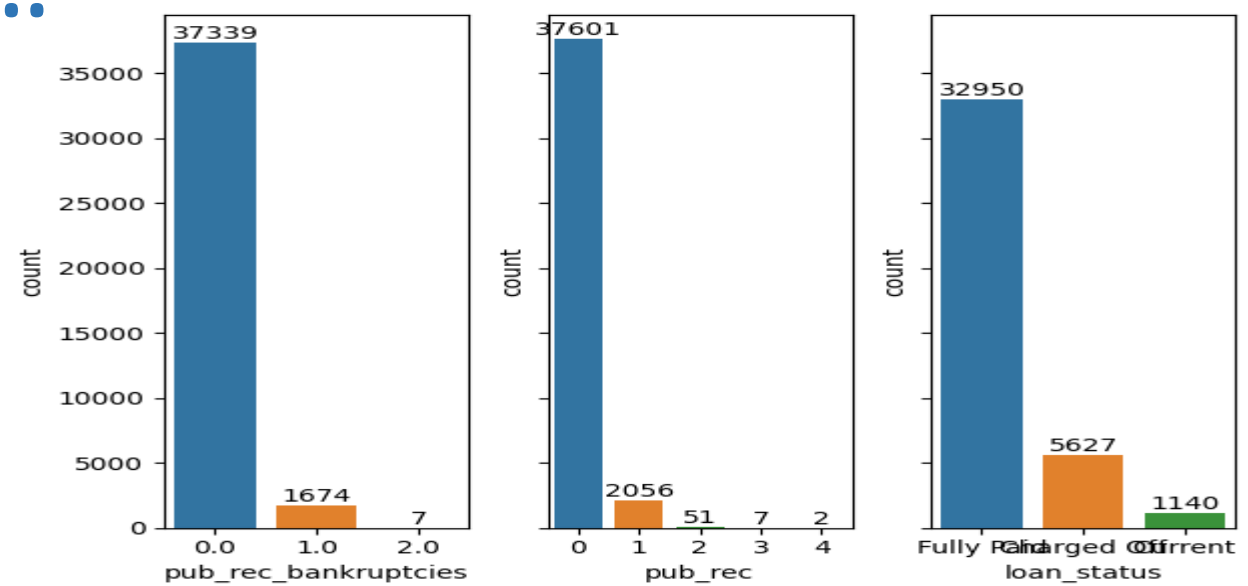
Data Cleaning

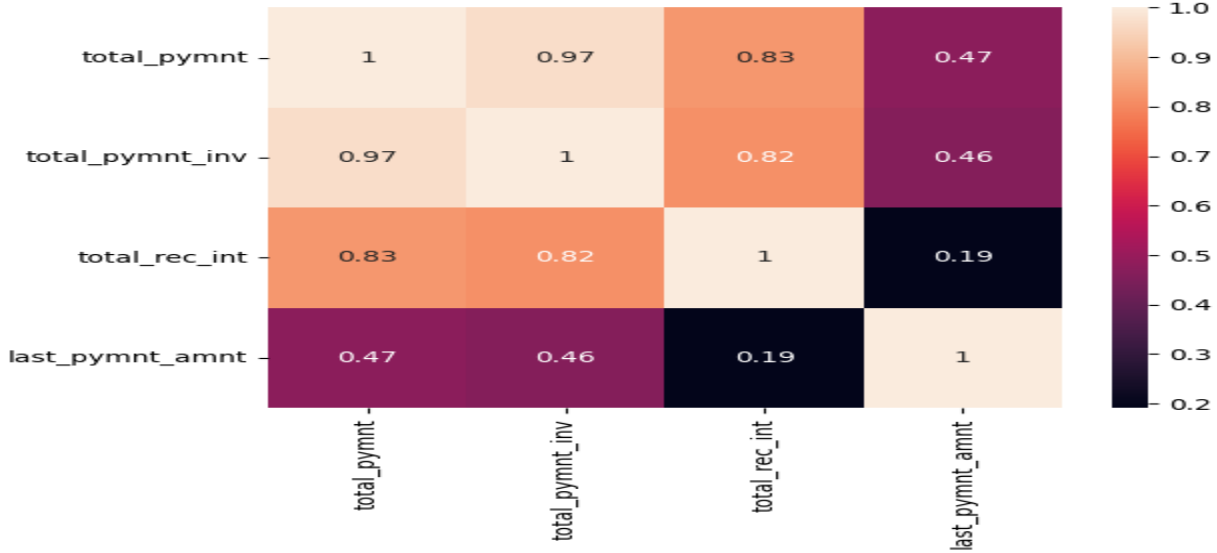
- Delete unnecessary columns
  - Many columns which have null values, hence dropping them
  - Some columns having missing values or nulls, hence dropping them
  - For the analysis purpose, taking ration number of 0.6 ie Having 60% or more null values dropped
- Dropping the following 57 columns which satisfy the above condition
- Dropping Columns which don't add any meaning to the analysis like id, member_id, desc, zip_code, url, addr_state etc
- Observed some of the columns high unique values and low unique values like total_pymnt, total_pymnt_inv, total_rec_int, last_pymnt_amnt, emp_title(high unique values) and term, pub_rec_bankruptcies, loan_status, verification_status, pub_rec (low unique values)
- Some values are unique values and discarding then as it cant be used for analysis
- Merge columns to create unique ids
- Split the columns for more data analysis
- Add missing column names
- Fixing the misaligned columns

Data Cleaning Contd…

Some of the observations and dropping the columns like

- public_rec_bankruptcies and pub_rec have very high skewed values which are not useful in drawing insights
- Although the variable/column loan_status is also included into the graph but since it is the target variable, we will not drop it.
- Columns total_pymnt & total_pymnt_inv have very high correlation and drop one of them.
- Although column emp_title has unique values and categorical in nature, the value of each category is not that significant and wouldn't help us much in our analysis, drop it as well.





Data Manipulation

Data Manipulation

Standardizing Values and Numbers

- Remove Outliers – This can be identified by box plot (eg, annual income)
- Standardize units, Converting lbs to kgs, miles to kms
- Scale values if required. Fit to percentage
- Standardize precision. Round to nearest decimal (eg to 2 decimal)

Fix Invalid Values

- Encode unicode properly
- Convert incorrect data types. String to Num, String to Date etc...(Eg, %' was removed from int_rate column and converted to int64 and issue_d - column which is a string type column has been converted to datetype)
- Correct values
- Correct wrong structure
- Correct values beyond range
- Validate internal rules

Data Manipulation Contd…

Fix Missing Values

- Set values as missing values i.e., NA or null
- Fill missing values with constant data, column functions or with external data
- Remove missing values
- Fill partially missing values
  - 1075 null values in "emp_length" column
  - 11 null values in "title" column
  - 50 null values in "revol_util" column
  - 71 null values in "last_pymnt_d" column
  - 2 null values in "last_credit_pull_d" column

  Count -- 39717

Data Analysis

- Univariate
- Bivariate

Univariate Data Analysis

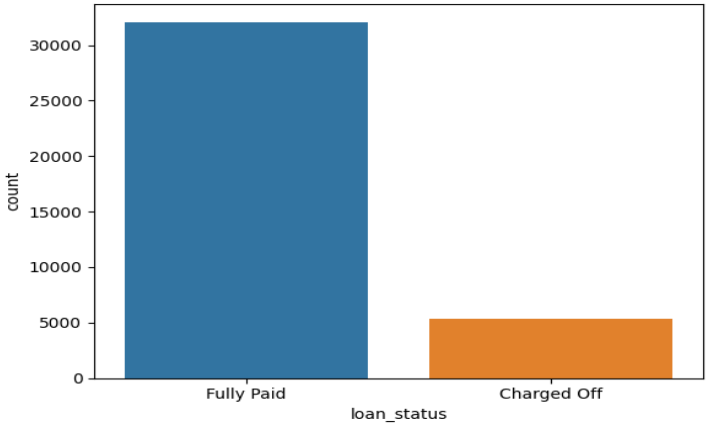
- Variables can be broadly divided as
  - Categorical (Again sub-divided into Ordered and Unordered)
  - Quantitative / Numerical

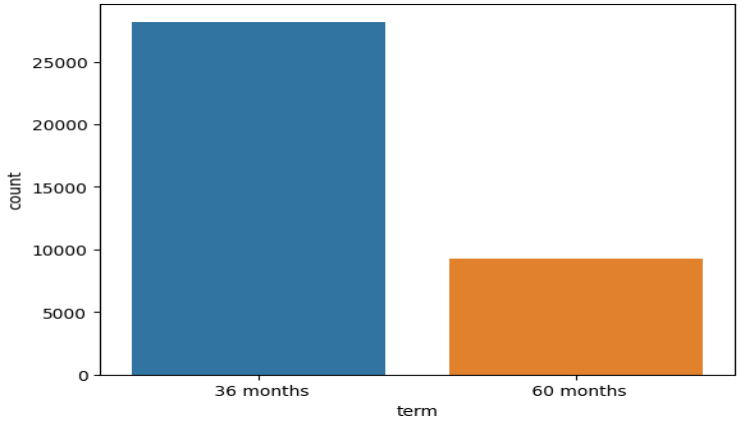
In the upcoming slides we show some graphs of a single variable

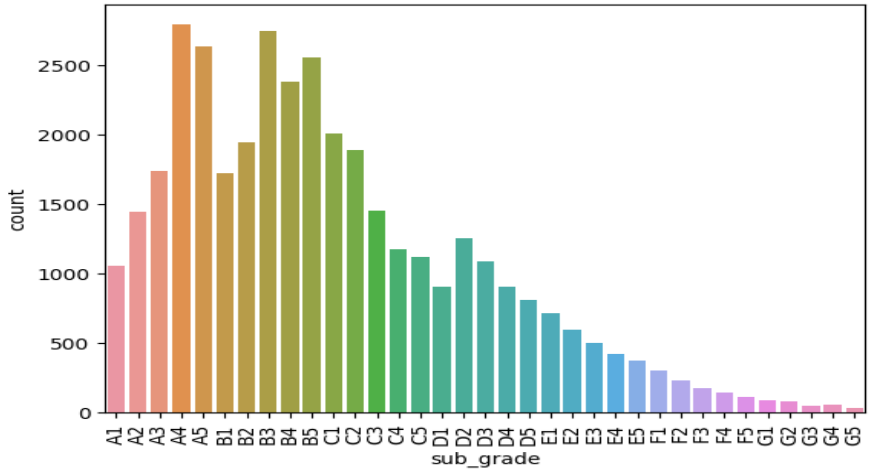
- For Categorical Variables – we use countplots
- For Numerical Variables – we use boxplots/histplots

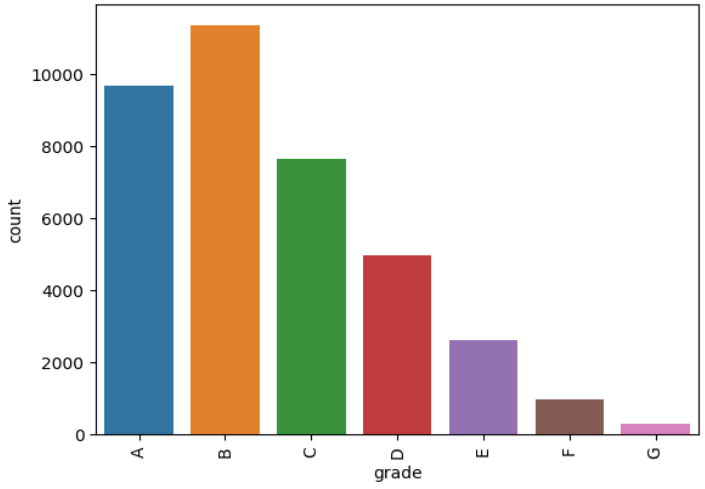
Univariate Data Analysis Contd…

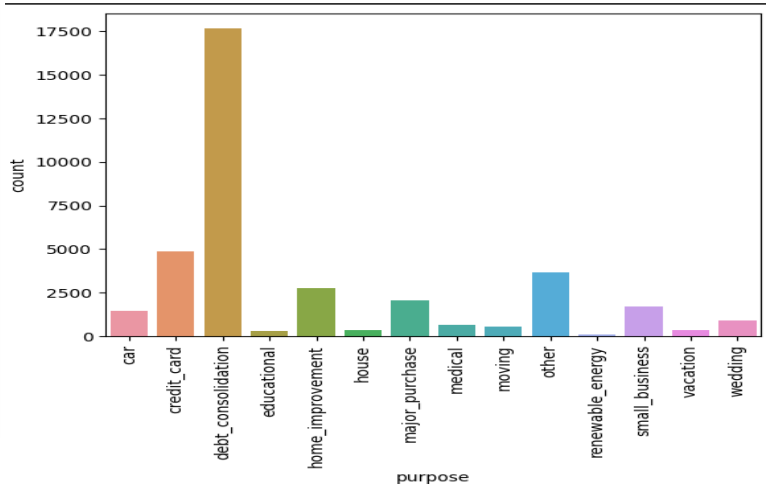
Some example countplot on categorical variables

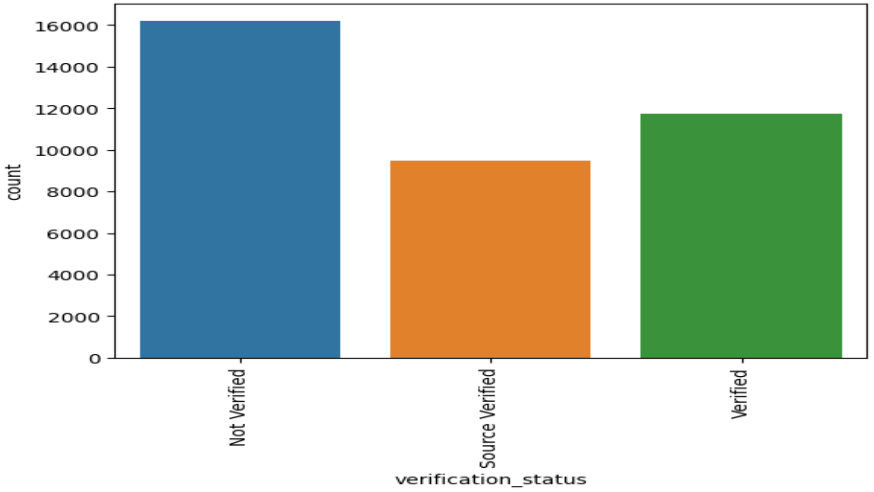






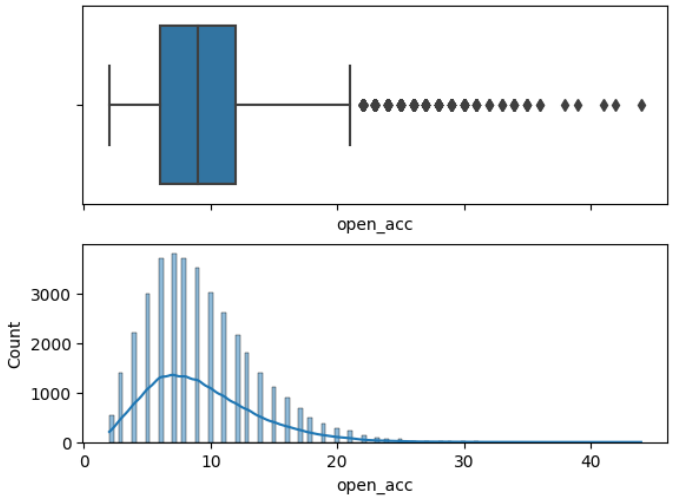


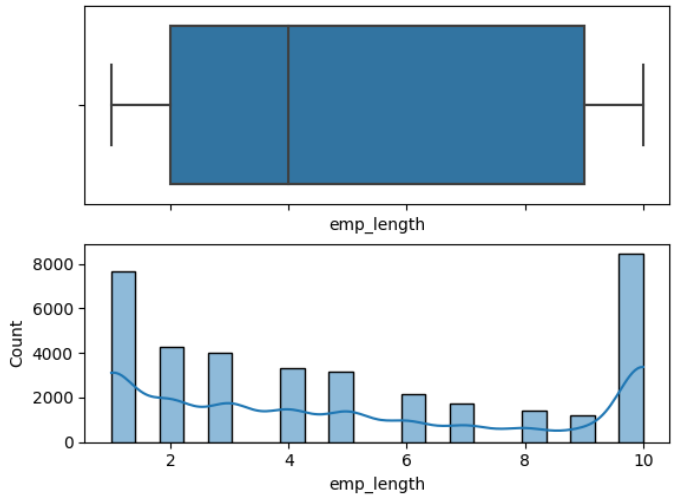


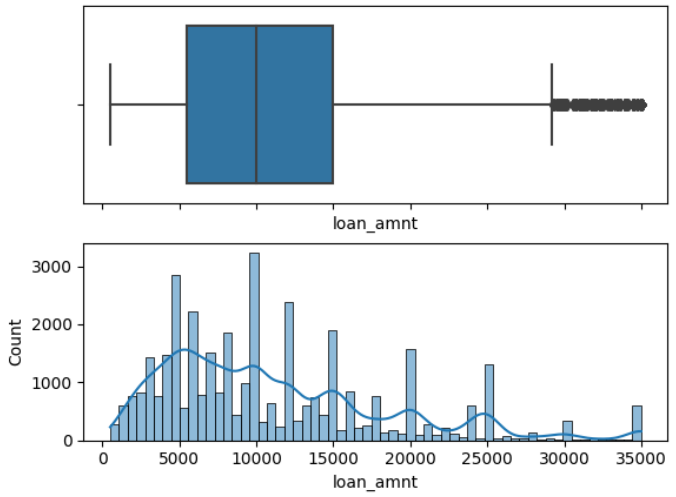


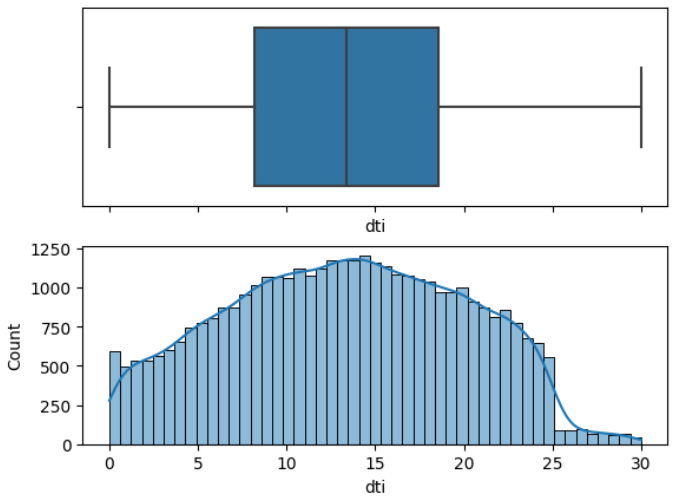
Univariate Data Analysis Contd…

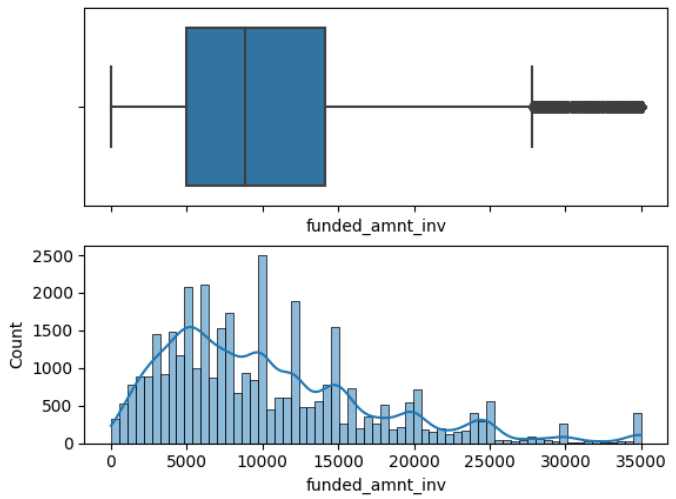
Some example boxplot & histplot on numerical variables

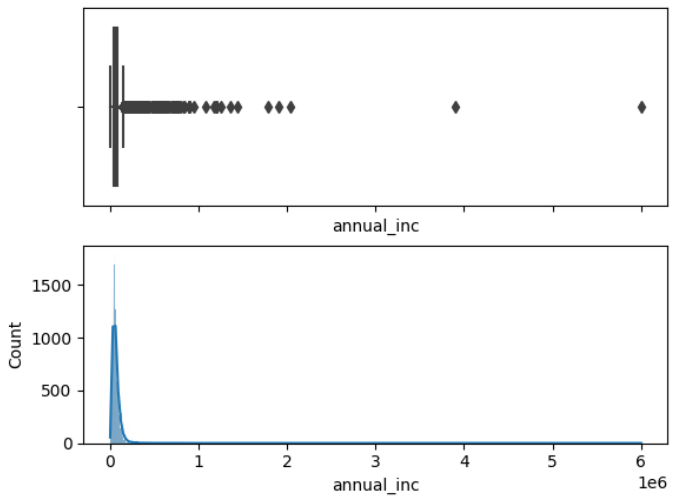










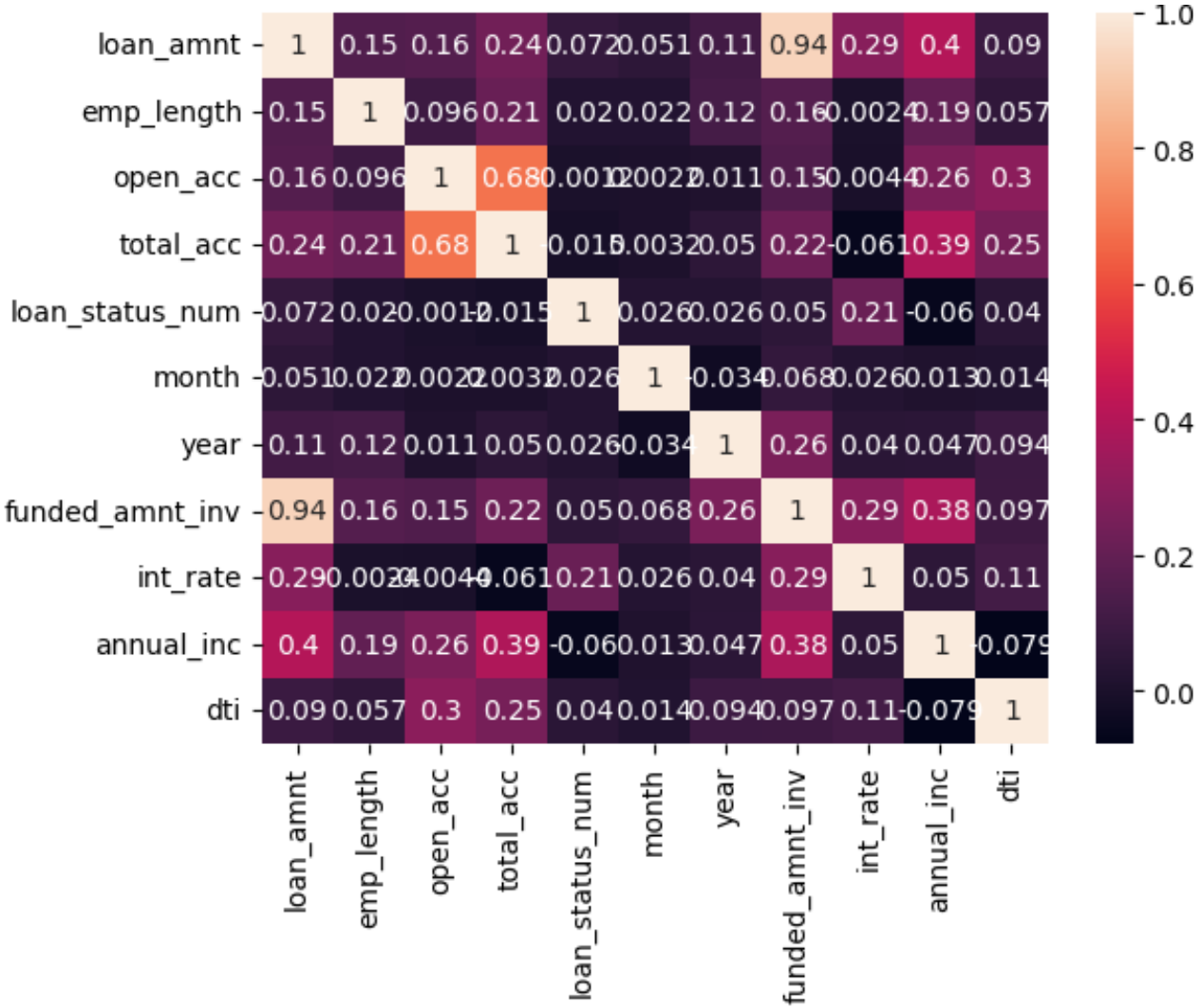


Data Analysis – Bivariate Analysis

Bivariate Analysis

- Plotted graphs for most variables to get a rough idea on how the data is distributed and if the data has high outliers we have also removed them by easily looking at the graphs.
- This is mostly the purpose of the univariate analysis for our case-study. For our case-study we would be able to derive more inferences once we start to compare these variables against the target variable i.e., by doing a Bi-variate Analysis.
- Three ways in which a Bivariate analysis can be done
  - Numerical - Numerical - scatter plots, correlation matrix, pairplots
  - Numerical - Categorical - Grouped box plots, Bar plots, violin plots
  - Categorical - Categorical - Pivot table, Crosstab
- For all numerical categories we can plot a pairplot which would give us the relationship between those variables, but wouldn't contribute much to our analysis in our case-study.
- Plot a heatmap which shows correlation between each of those numeric variables and we can exclude those with high correlation to make our initial analysis. We have already done this kind of analysis before in this sheet to remove the unnecessary rows.
- Observed significant impact on default rate for the variables term, grade, annual_inc, int_rate, purpose, home_ownership, funded_amnt_inv, annual_inc, dti - debt to income ratio

Data Analysis – Bivariate Analysis Contd…



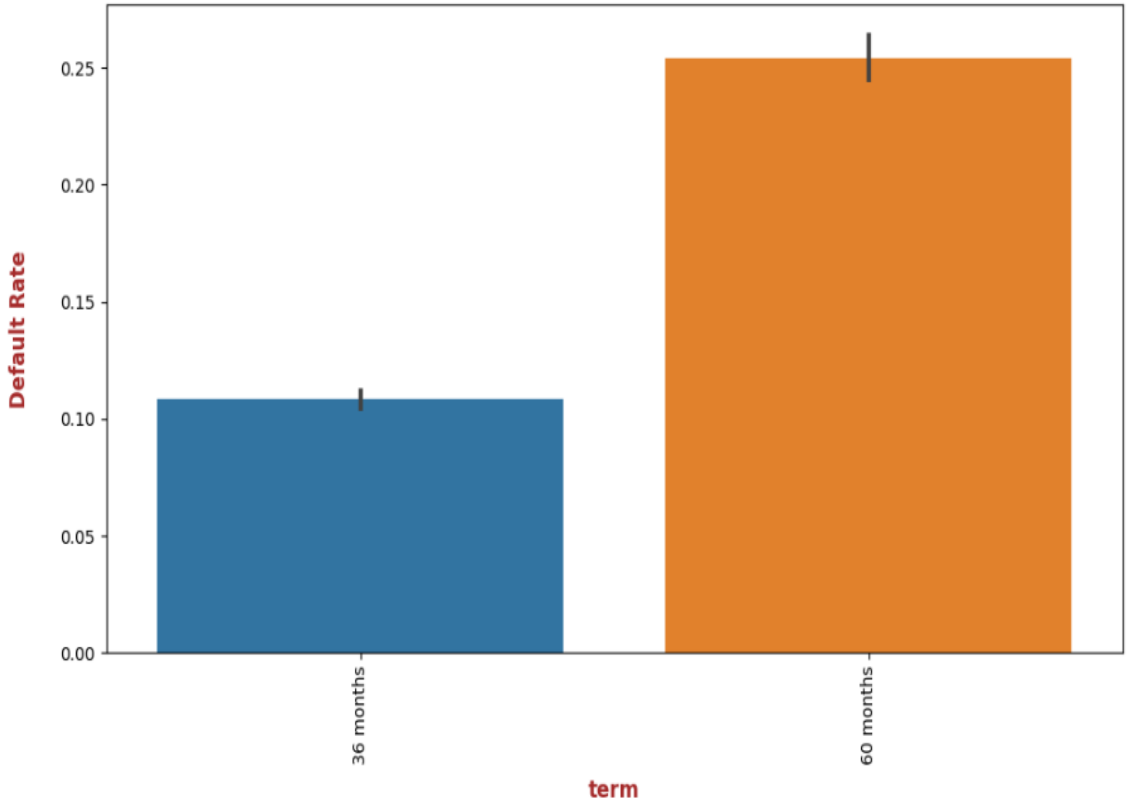
Correlation Heatmap of all the numeric variables

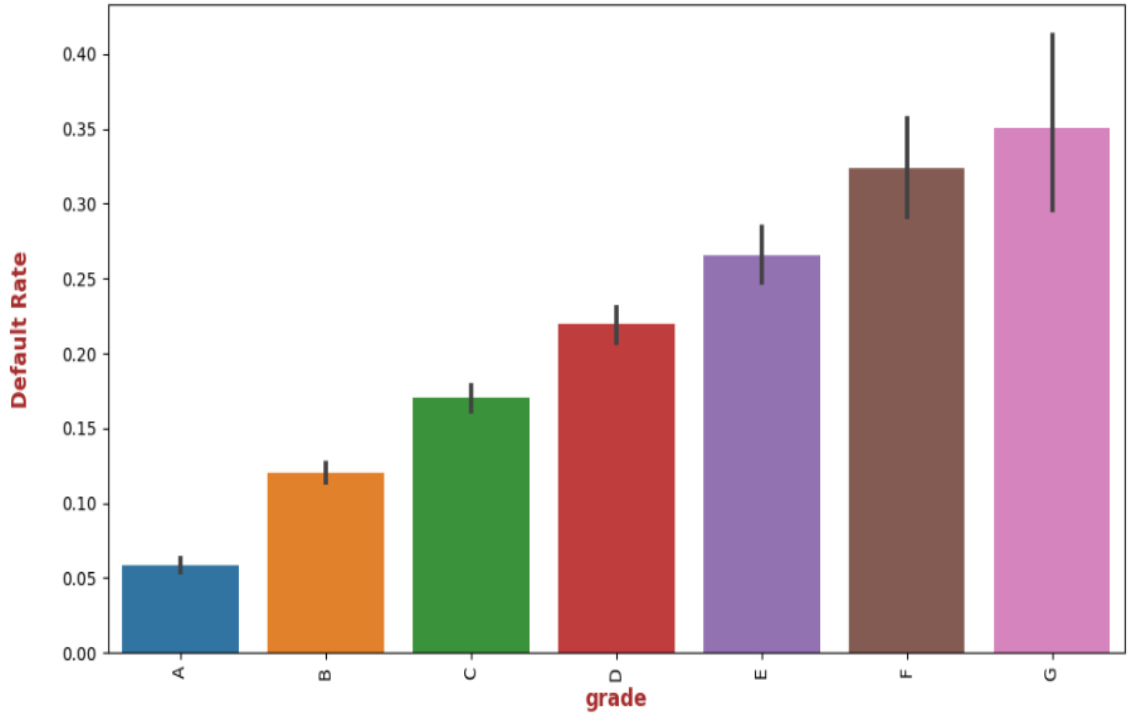
Correlation plots are quite useful to draw insights for the numerical-numerical type of bivariate analysis.

This was plotted to see if there are any highly correlated numeric variables.

Data Analysis – Bivariate Analysis Contd…

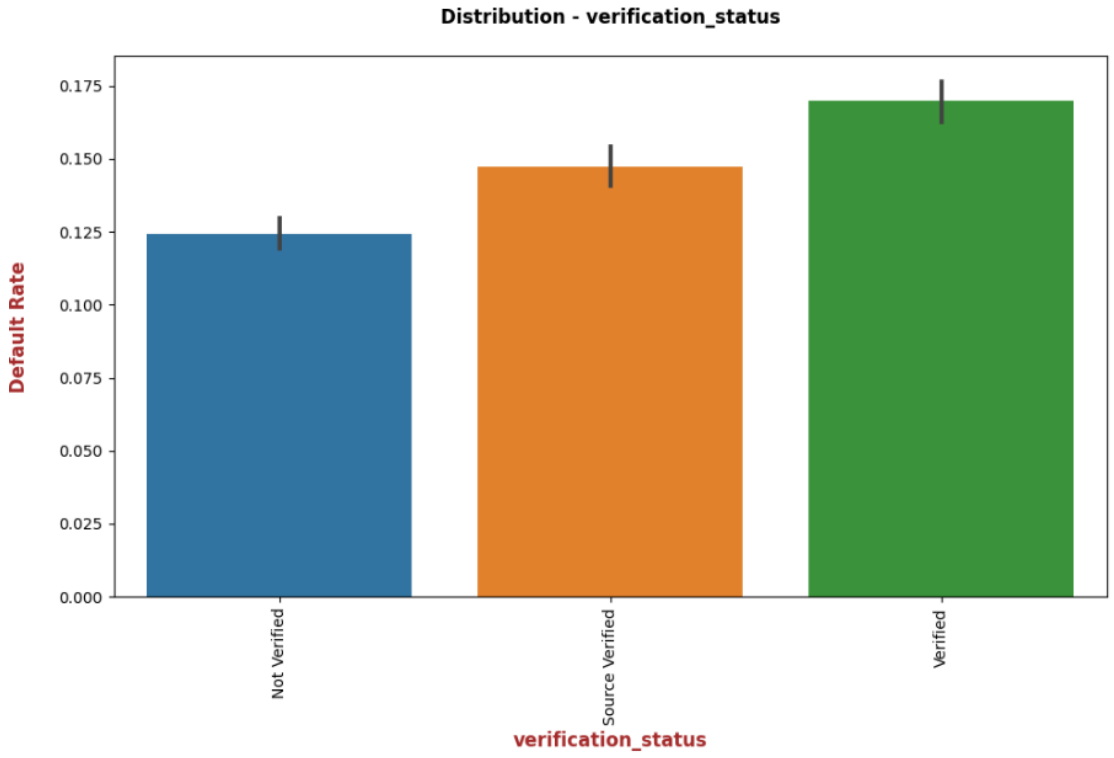
- Rate of defaulters is higher for 60 months than compared to 36 Months. That means we can say that the as the term increases the number of defaulters are also increasing.
- From the graph Grade vs Default Rate, it can be observed that the Grade **G** people are more likely to default than Grade **A** people, and the default rate is also increasing from Grades A to G

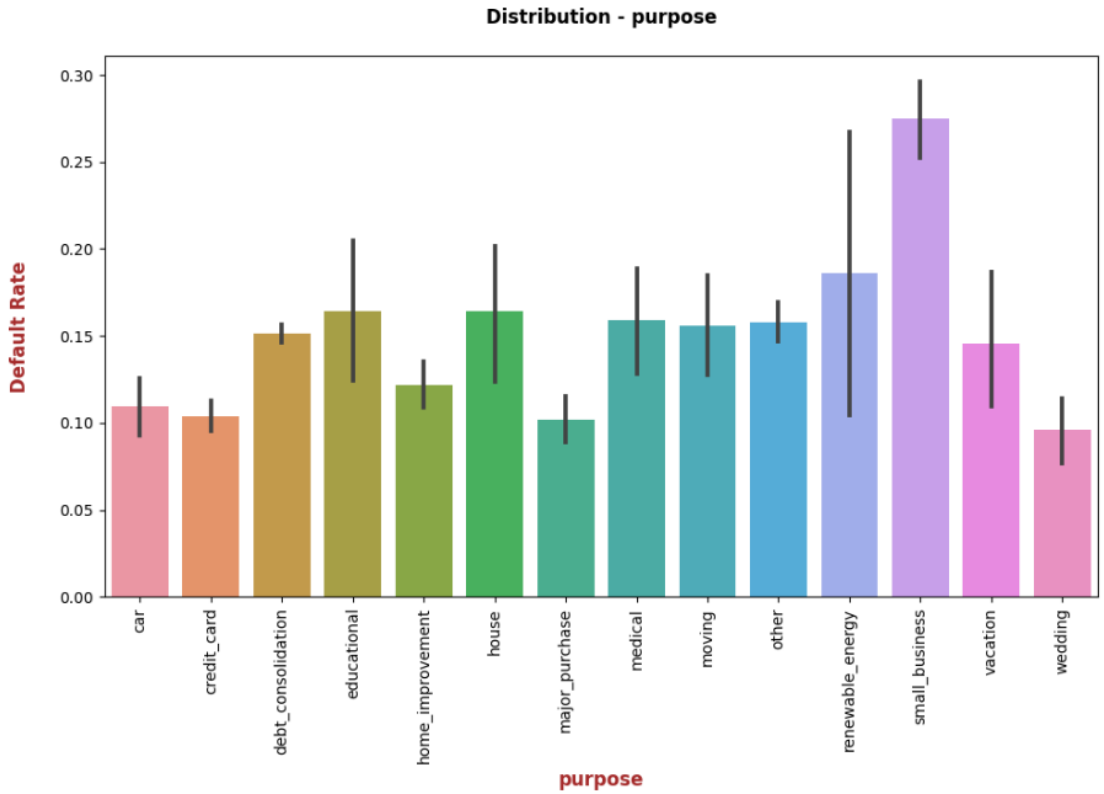




Data Analysis – Bivariate Analysis Contd…

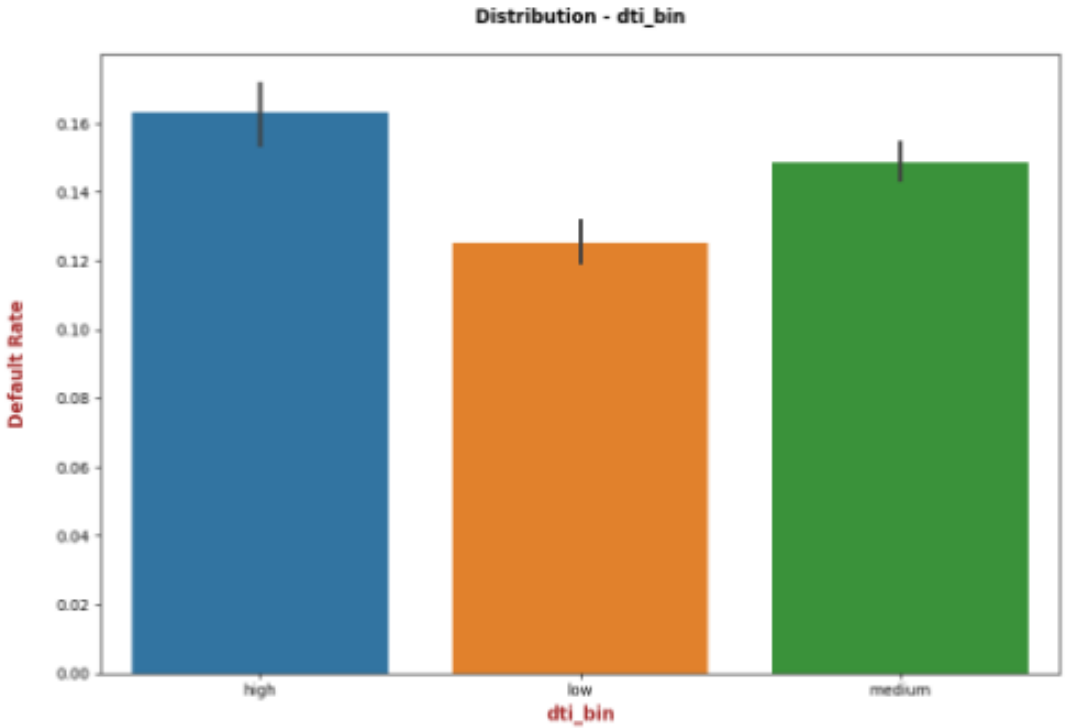
- Suprisingly, in the plot for 'verification_status' we observed that people whose status is **verified** and **source verified** are more prone to default than people with 'not verified' status
- And at last from the purpose distribution we observe that the people taking loan for the purpose of **small_business** are more likely to default than other purposes.

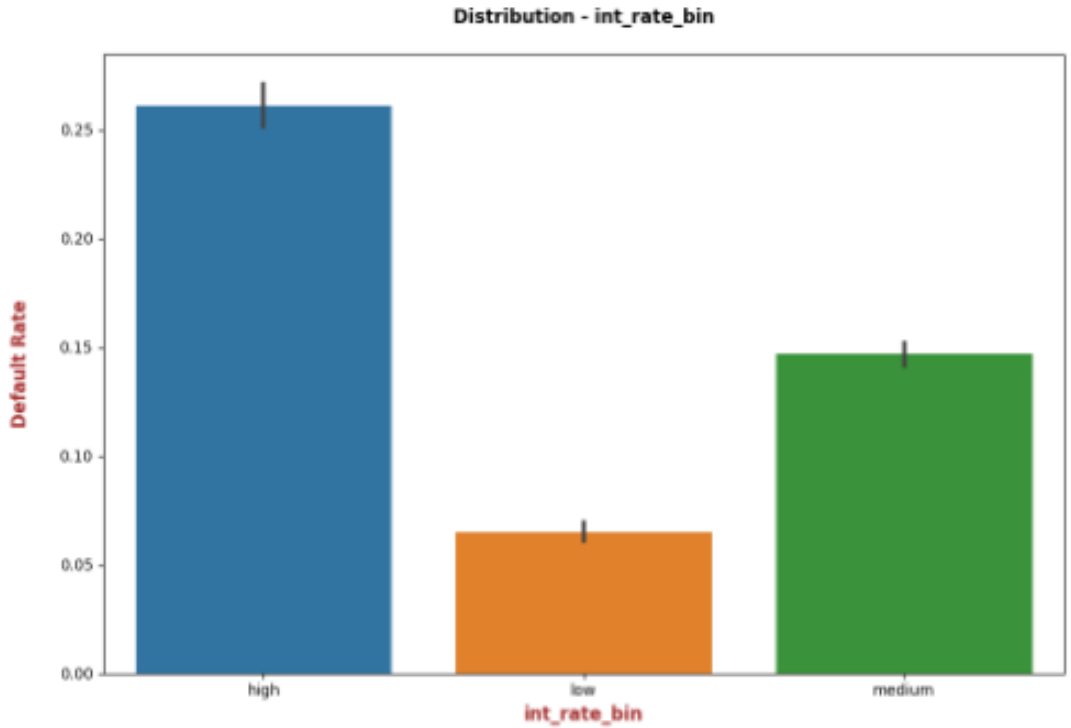




Data Analysis – Bivariate Analysis Contd…

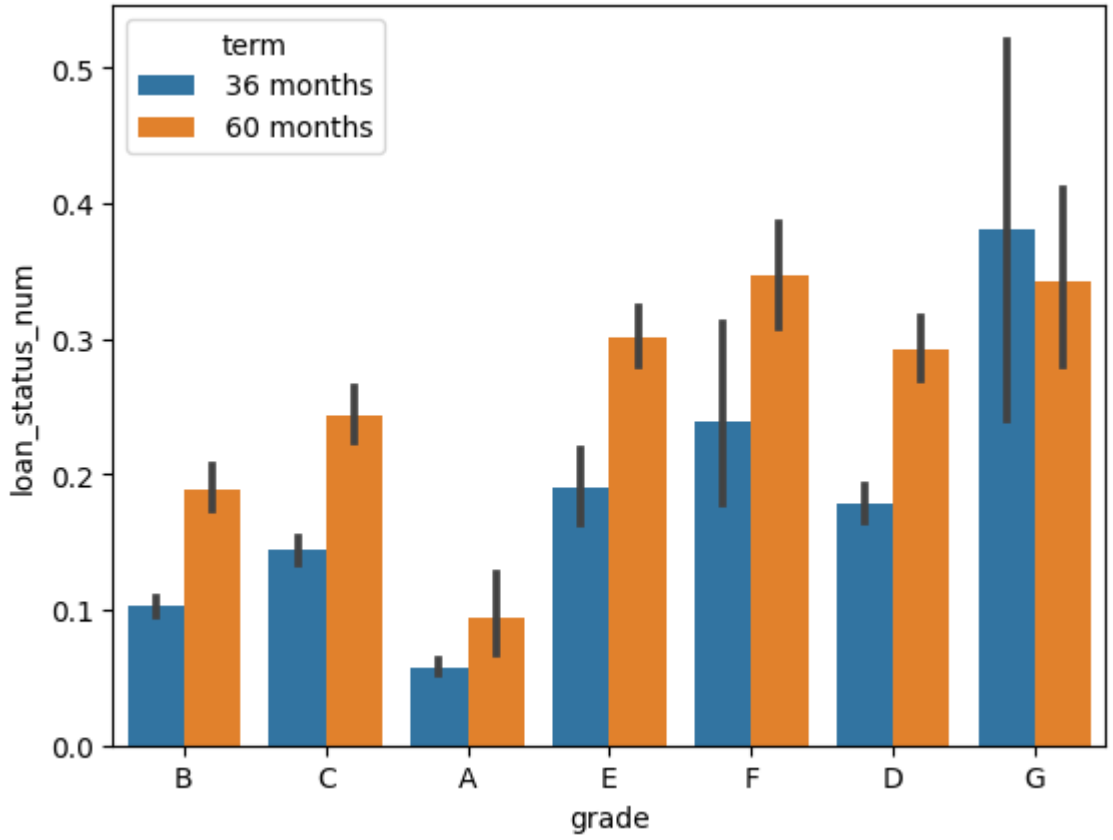
- Similar pattern as the Grade is observed in sub-grade category vs defaulters rate. The trend of defaulters is increasing from sub_grade **A1** to **G5** with minor exceptions.
- From above plot for ‘home_ownership’ we can infer that the defaulters' rate is more or less constant here. It is quite more for OTHERS, hence default rate will depend on home ownership if it is of type OTHER or not.





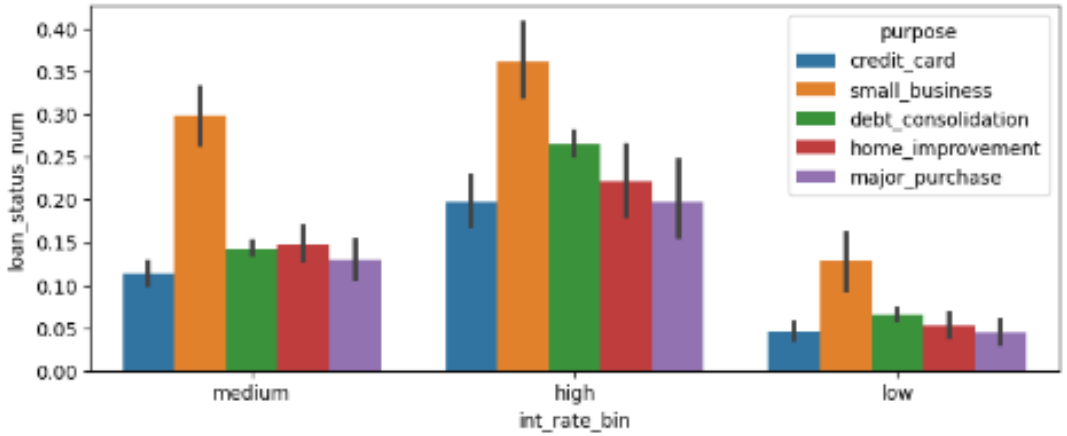
Data Analysis – Bivariate Analysis Contd…

- From the before analysis we found out that as the Grade increased, the default rate increased.
- From the just side plot of grouped barplot we could see that in all the grades except from G there is definite distinction between people with tenure is 36 & 60 months. At each grade, people with less loan tenure/period are less likely to default.



Data Analysis – Bivariate Analysis Contd…

- As observed before and also in this graph of int_rate vs loan_status grouped on purpose, the risk of defaulting increases with the interest rate increase and also interesting thing to observe is that small_business, renewable energy, educational and house loans also tend to default more as the int_rate increases.

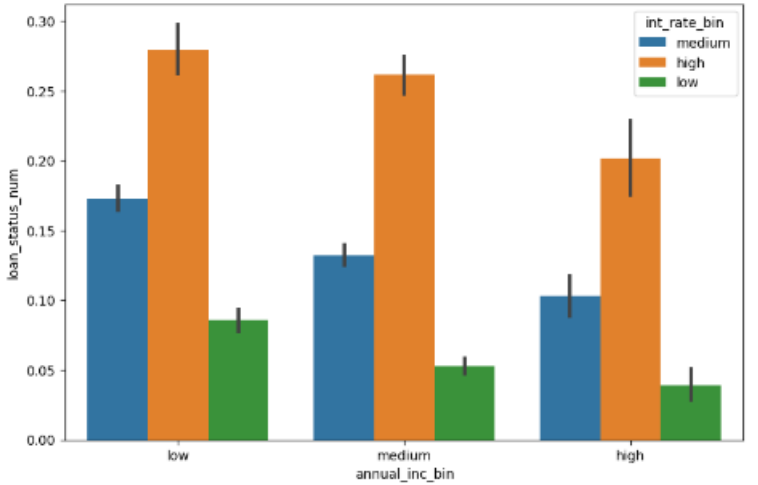


Comment on **purposes field**

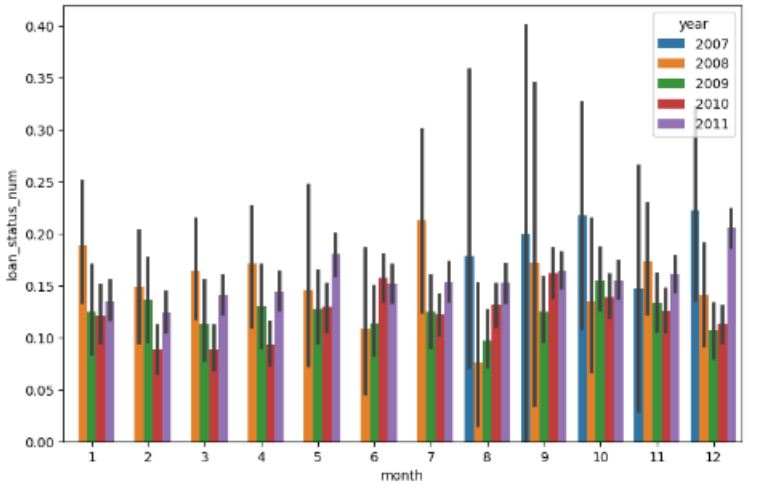
- Although the number of loans taken for small business purposes are comparitively lesser, but the probability of those loans to be defaulted is much higher than loans taken for any other purposes. In each and every plot irrespective of the other category, we observed this general trend.

Data Analysis – Bivariate Analysis Contd…

- From plot annual_inc_bin vs loan_status_num we can observe a linear relation between annual income , debt to income and default status. That means as the annual income of person given loan falls into the high range category and their dti (debt to income ratio) is low then that person is very unlikely to default and their defaulting risk increases as their dti increases or their annual income decreases.



- From the plot month vs loan_status, we previously observed that during the months of May, September and December the default risk is higher. But now from this graph its observed that during the years 2007 & 2011 the default risk is higher in september and december than other months which is the cause for that spike we observed.



Recommendations

Recommendations & Result

Result:

Top columns/variables which affect the default rate are in decreasing order of influence

1. sub_grade - LC assigned loan subgrade - **0.5**
2. grade - LC assigned loan grade - **0.29**
3. int_rate - Interest Rate on the loan - **0.2**
4. home_ownership - The home ownership status provided by the borrower during registration- **0.19**
5. purpose - A category provided by the borrower for the loan request - **0.18**
6. term - The number of payments on the loan. Values are in months and can be either 36 or 60 - **0.15**

Other influencing variables are loan_amount_bin, annual_inc_bin, funded_amnt_inv_bin, verification_status, dti_bin, emp_length

Recommendations & Result Contd…

Recommendation:

- As the sub_grade assigned to the loan type changes from A1 to G5 the default rate also increases
- As the grade assigned to the loan type changes from A to G the default rate also increases
- As the interest rate increases the default rate also increases
- While giving loans to a home_ownership type which comes under other category, LC has to be careful while giving loans because the highest default rate is observed for OTHER category
- If the purpose of loans is for small businesses, debt consolidation, credit card, home improvement, major_purchase, then higher the risk of defaulting
- If the term of the loan is higher, higher is the risk of defaulting

THANK YOU