LENDING CLUB: EDA CASE STUDY

Abstract

- **Analysis background:**

We work for a Consumer finance company which specializes in lending various types of loans.

This company considers two categories of risk that contribute to its financial loss– 1) rejecting a loan for an applicant who is more likely to repay the loan/ not default 2) approve loan for an applicant who would most likely NOT repay the loan/ default.

The aim of the financial company is to minimize this risk by understanding the important factors which should determine whether a loan application should be accepted or rejected.

- **Objective:**

We need to come up with a list of variables which are strong indicators of Default by utilizing the loan.csv dataset which contains the profile of all those applicants whose loans were accepted.

These variables would then be considered by the company as the driving factors which decide whether a loan application should be rejected or not i.e. whether an applicant would default or not.

- **Analysis:**

We've identified the number of applicants that fall in the 'Charged Off' ,'Fully paid' and 'Current' categories .

We've then considered the 'Charged off' i.e. Applicants who have a 'Defaulted' loan status for further analysis.

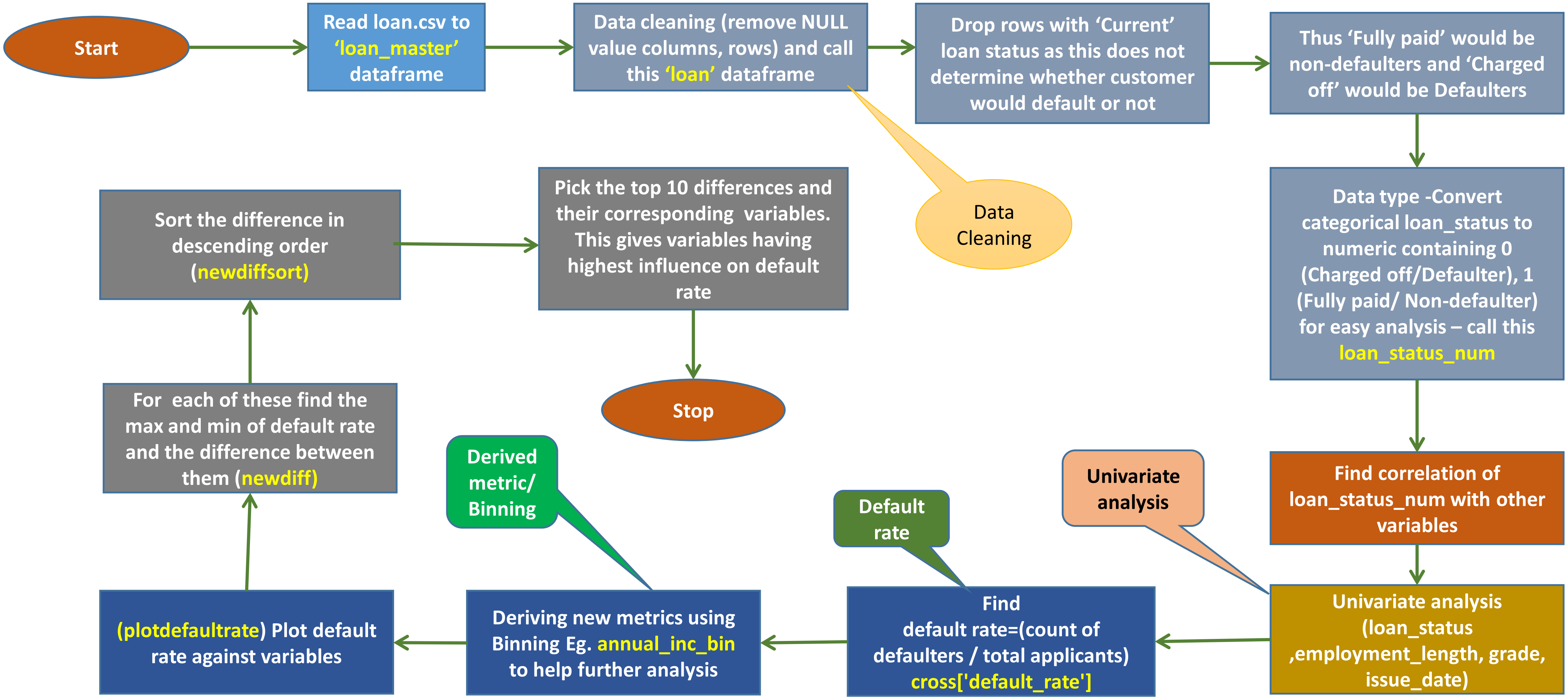
We have identified the relation this has, with other variables in the dataset. Using this we have come up with variables which have the highest influence on 'Defaulted/Charged Off' loan status.

We've used the important concepts of EDA to come up with the top factors influencing the Loan default, details of which you will find in the next few slides.

- **Analysis result:**

Our analysis indicates that home_ownership, purpose, sub_grade, int_rate , grade, annual_inc are some of the major factors contributing to Loan Default.

Problem solving methodology - Flow chart



Start

Read loan.csv to 'loan_master' dataframe

Data cleaning (remove NULL value columns, rows) and call this 'loan' dataframe

Data Cleaning

Drop rows with 'Current' loan status as this does not determine whether customer would default or not

Thus 'Fully paid' would be non-defaulters and 'Charged off' would be Defaulters

Data type -Convert categorical loan_status to numeric containing 0 (Charged off/Defaulter), 1 (Fully paid/ Non-defaulter) for easy analysis – call this loan_status_num

Find correlation of loan_status_num with other variables

Univariate analysis (loan_status ,employment_length, grade, issue_date)

Univariate analysis

Find default rate=(count of defaulters / total applicants) cross['default_rate']

Default rate

Deriving new metrics using Binning Eg. annual_inc_bin to help further analysis

Derived metric/ Binning

(plotdefaultrate) Plot default rate against variables

For each of these find the max and min of default rate and the difference between them (newdiff)

Sort the difference in descending order (newdiffsort)

Pick the top 10 differences and their corresponding variables. This gives variables having highest influence on default rate

Stop

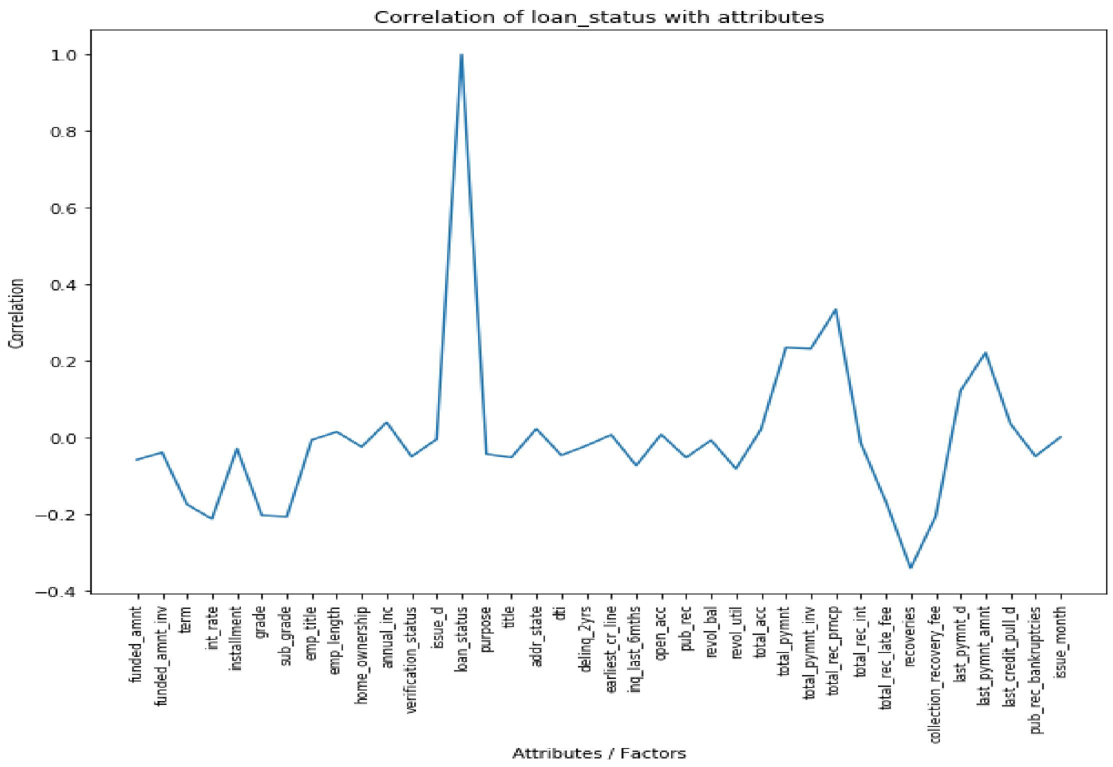
Analysis –Data cleaning

- Read the loan csv file as dataframe **‘loan_master’**
- Describe the data frame and find its shape. There are 39717 rows and 111 columns. Drop columns with all NULL values. Drop rows with all NULL values. Drop columns with more than 30 % NULL values.
- Drop columns which do not contribute to loan default at all i.e. id, member_id, zip_code (partial) , url.
- Clean the columns like int_rate (remove % and convert to numeric) and issue_date (extract month from date) and make them fit for analysis.
- We’ll be left with 49 columns and 39717 rows (number of rows remain the same as there are no rows with all NULL values).
- Drop the ‘Current’ value rows from the ‘loan_status’ column as records with ‘Current’ loan_status value do not determine whether or not customer would default.
- Convert loan_status from categorical to numeric column i.e. loan_status_num having values of 0 (Charged Off) and 1 (Fully paid).
- The cleansed ‘loan_master’ dataframe will be called **‘loan’.**
- Binning was performed for analysing continuous variables. E.g: annual_inc, funded_amnt_inv

Analysis - Correlation

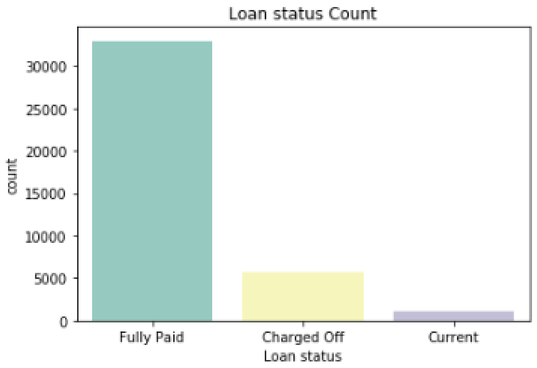
- Create a temporary dataframe of ‘loan’ called ‘loan_corr’ and convert all categorical columns to numeric for correlation.
- Find the correlation of all variables with other variables.
- Now find the correlation of loan_status with other variables in the dataframe. Drop the rows which are not correlated with loan_status (correlation value =0 or NaN)
- Sort the correlation series in ascending order. Find the columns with the highest and lowest values(highly positive and highly negative).
- However we are not concluding these as the top factors as few of them are customer behavioural attributes which are not available at the time of Loan application. Hence proceeding with further analysis using Default rate

```
funded_amnt              -0.056497
funded_amnt_inv          -0.037781
term                     -0.173487
int_rate                 -0.211390
installment              -0.027153
grade                    -0.201869
sub_grade                -0.205320
emp_title                -0.005041
emp_length                0.016068
home_ownership           -0.023099
annual_inc                0.040867
verification_status      -0.048262
issue_d                  -0.003457
loan_status               1.000000
purpose                  -0.041831
title                    -0.050776
addr_state                0.023912
dti                      -0.045078
delinq_2yrs              -0.020096
earliest_cr_line          0.008362
inq_last_6mths           -0.071878
open_acc                  0.009140
pub_rec                  -0.051001
revol_bal                -0.005854
revol_util               -0.080271
total_acc                 0.022608
total_pymnt               0.235898
total_pymnt_inv           0.232906
total_rec_prncp           0.334944
total_rec_int            -0.013008
total_rec_late_fee       -0.165115
recoveries               -0.339562
collection_recovery_fee  -0.204914
last_pymnt_d              0.123312
last_pymnt_amnt           0.222524
last_credit_pull_d        0.036695
pub_rec_bankruptcies     -0.047757
issue_month               0.002239
```

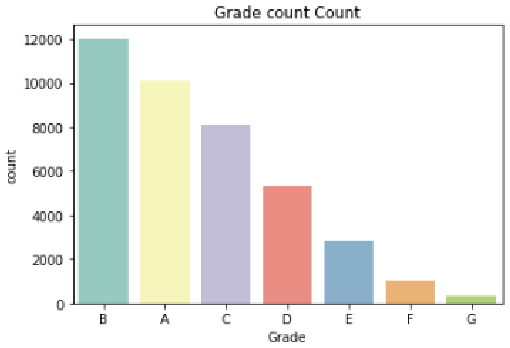


Analysis – Gaining insights

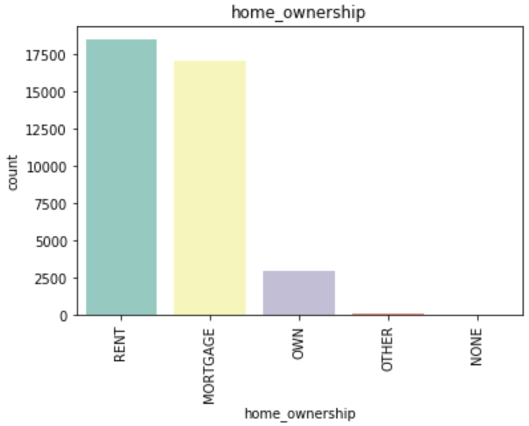
Now that we have the variables of interest for further analysis: Perform analysis (univariate) on categorical columns like term, purpose, loan_status, emp_length, verification_status, grade, subgrade, home_ownership, purpose. Find the frequency of these variables in the loan dataframe



Insight:
83% of applicants have Paid off their loans.



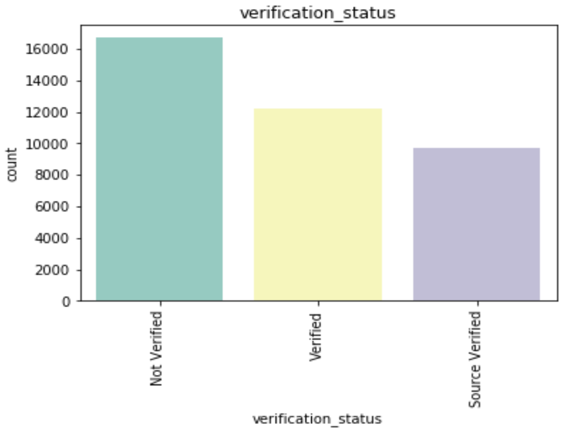
Insight:
Around 30% of applicants are in Grade B category, 25% in grade A



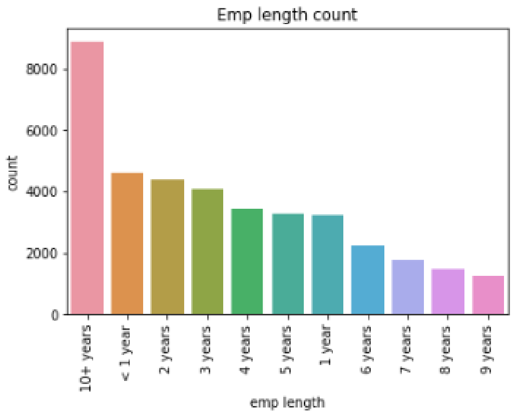
Insight:
We have 46% of applicants renting a home and 43% of applicants having a mortgage and only 7% who own a home.

Results:
Loan status (Loan attr) – 83% of the customers have fully paid off their loans
Grade(Loan attr – 30% i.e. highest number of applicants have Grade B
Home Ownership (Applicant attr) – 46% applicants rent a home, 43% have a mortgage, only 7% own a home

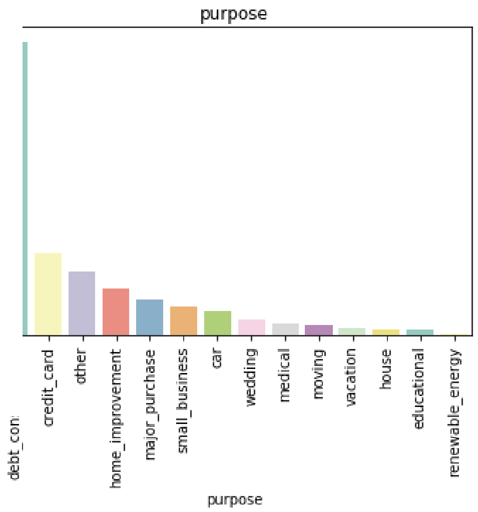
Analysis – Gaining insights



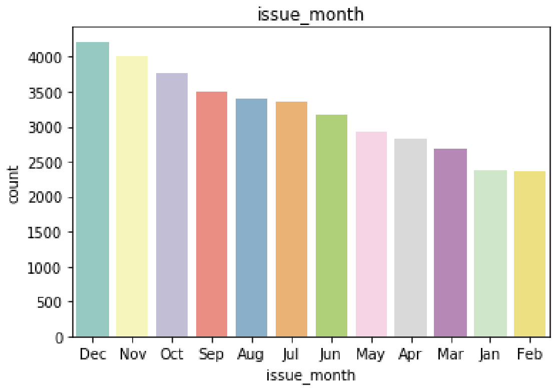
Insight:
Thus there are 42% (highest 16694) of customers whose income is Not Verified , 30.7% whose income is verified, 24% whose income source is verified



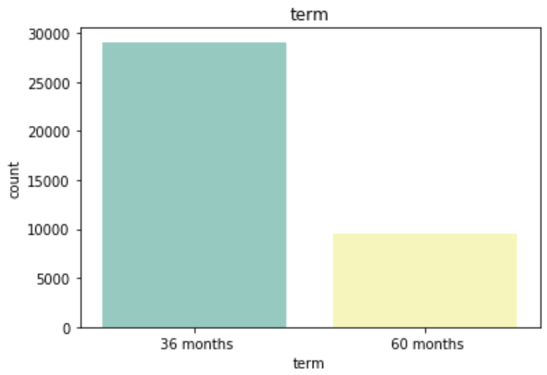
Insight:
We have 20% (8488 - highest) of applicants in the > 10 yrs experience category followed by less than 1 year category



Insight:
We have 45% (highest 18055) of applicants applying loan for the purpose of debt consolidation



Insight:
We have 10% (highest 4215) of applicants for whom the loan was funded in the month of December and 5% (least 2358) for whom the loan was funded in Feb.



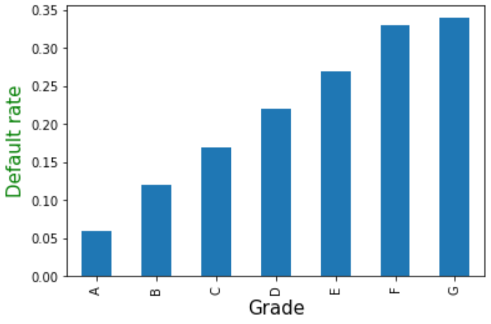
Insight:
We have 73% of applicants having 36 no. of payments on the loan and 27% applicants having 60 months

Results:

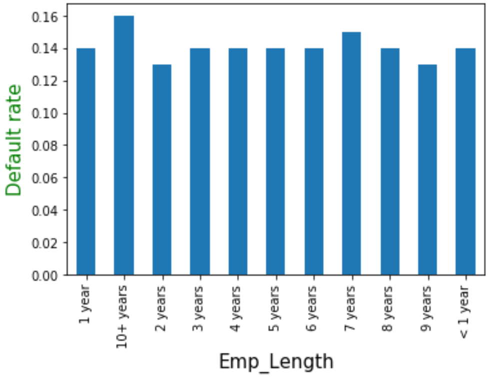
Emp_length (Appl icant attr) – Applicants with different no. of years of experience are there of which highest are from >10years category
Purpose (Loan char)– 45% applicants apply for purpose of debt consolidation
Issue_month (Loan char) - Highest number of loans are funded in the month of December
Term (Loan char)– – 73% applicants have 36 number of payments on loan
Verification status (Loan char)– – 42% applicants were not verified

Segmented Univariate Analysis

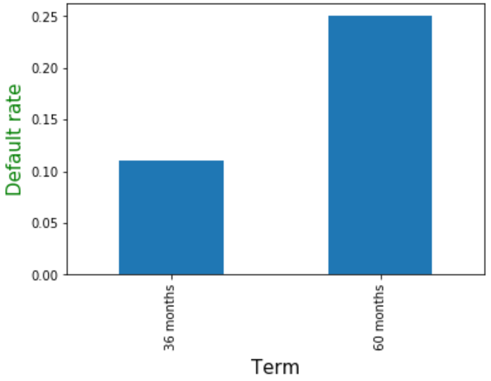
- **The following have been plotted to observe and infer how the default rate varies with each of these variables**



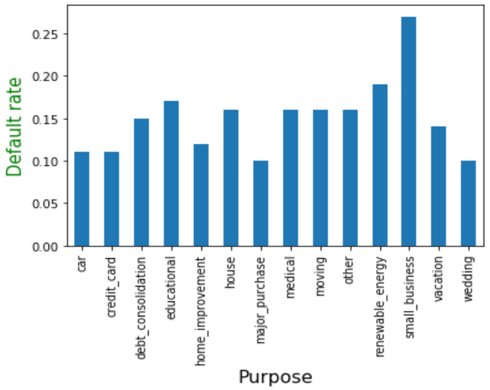
Insights: As grade value increase so is default rate. **Max – Min = 0.28**



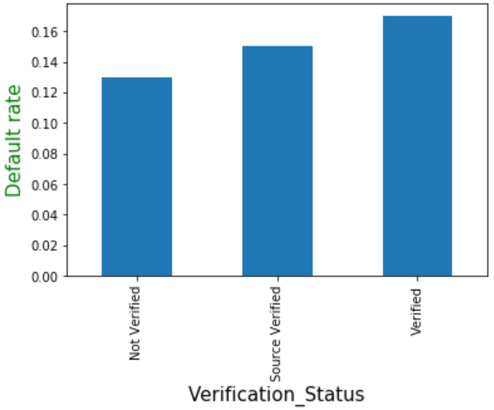
Insights: Applicants with 10+ years of employment have the highest default rate. **Max – Min = 0.03**



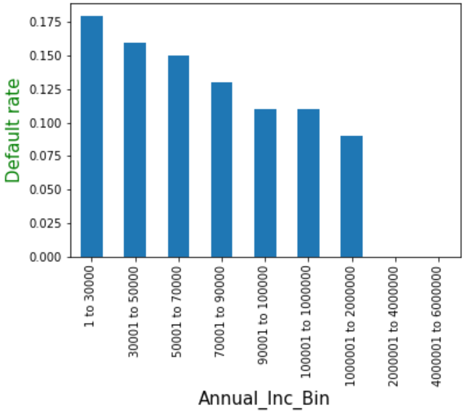
Insights: Default rate is comparatively higher in the applicants who have opted for 60 months Term. **Max – Min = 0.14**



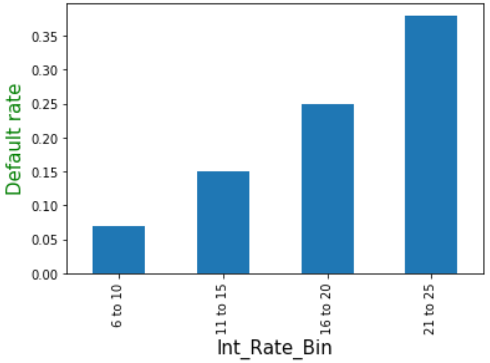
Insights: Small business applicants have higher chances of defaulting.
Max – Min = 0.17



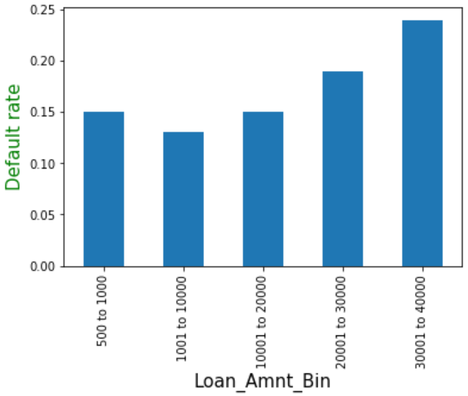
Insights: Surprisingly Default rate is higher in those applicants whose income was ‘Verified’.
Max – Min = 0.04



Insights: Apparently applicants whose annual income falls between1 and 30000 have highest chances of defaulting. Thus higher the annual income lower the Default rate . **Max – Min = 0.18**

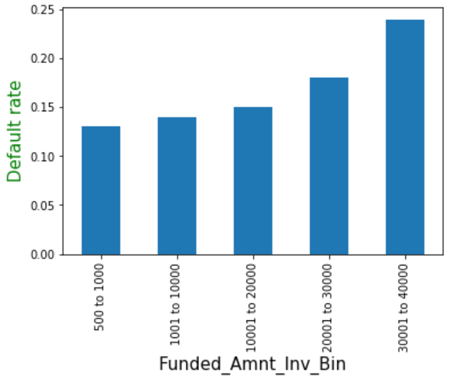


Insights: Loans lent with interest rate between 21 and 25 have the highest default rate. Lower the interest rate, lower is the default rate . **Max – Min = 0.31**

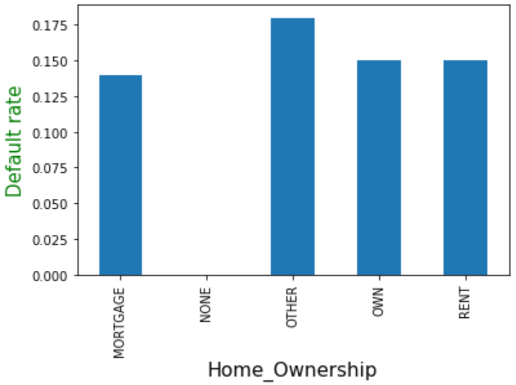


Insights: Loan amount between 30001 and 40000 have the highest default rate
Max – Min = 0.11

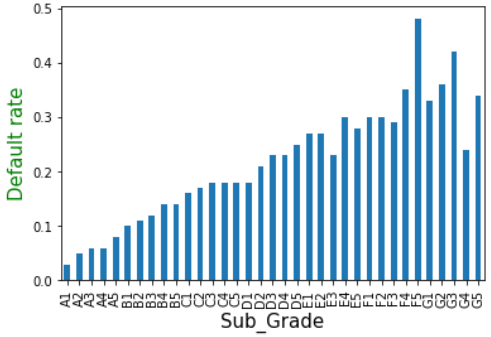
Segmented Univariate Analysis – continued..



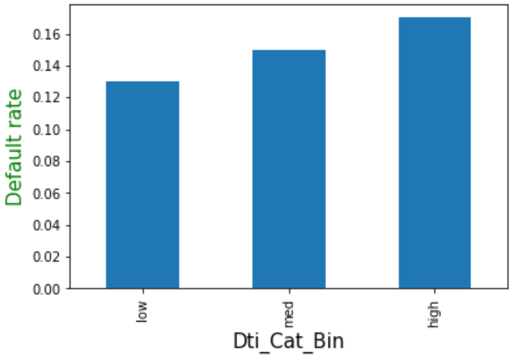
Insights: Fund amount invested between 30001 and 40000 have the highest default rate similar to Loan Amount
Max – Min = 0.11



Insights: Default rate is highest for 'Other 'category. **Max – Min = 0.18**

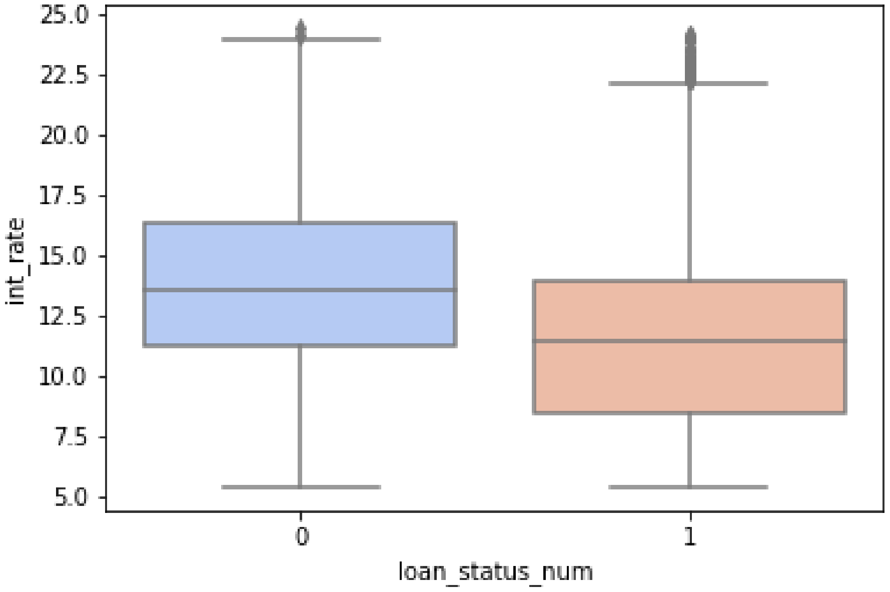


Insights: F5 sub_grade has the highest default rate followed by G3. **Max – Min = 0.45**



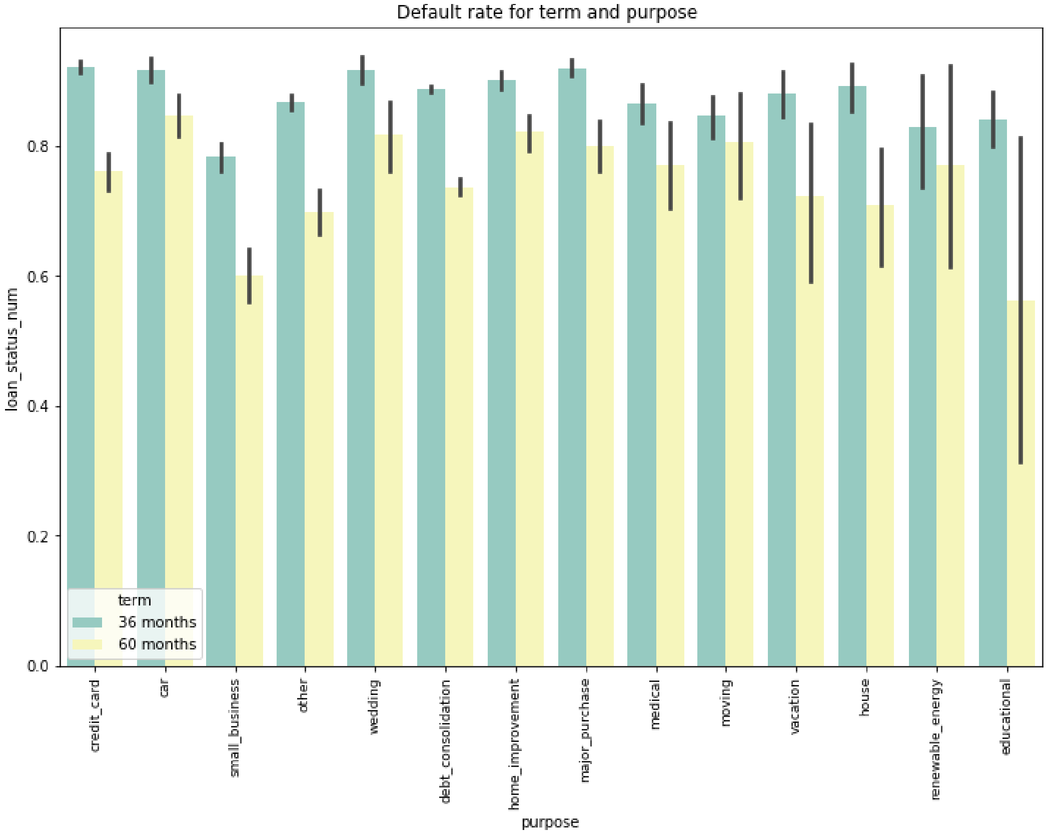
Insights: Higher the dti higher is the default rate.
Max – Min = 0.04

Bivariate / Multi Variate Analysis



Insights: Average interest rate of the applicants who default is great than those who paid fully

0 – Charged Off
1 – Fully Paid



Insights: Default rate is higher for the applicants who opted 36 months term for any purpose

Conclusion

- As part of the Lending Club Case Study we have analysed and would recommend that the following factors be considered by Consumer Finance Company to approve / reject loan of an applicant:
  - sub_grade
  - int_rate_bin
  - grade
  - annual_inc_bin
  - home_ownership
  - purpose
  - term
  - funded_amnt_inv_bin
  - loan_amnt_bin
  - dti_cat_bin
  - verification_status
  - emp_length
- Please refer to Slide – 8, 9 to see how these factors influence Default rate.
- The above recommendations are based on the following assumptions and considerations
  - Loan_status is the attribute from the dataset we have considered to determine whether an applicant is a defaulter or not
  - We have used correlation matrix to find those attributes that are highly correlated with loan_status to perform further analysis
  - Default rate = No of Charged Off customers / Total No of applicants
  - We plotted the default rate against each of these attributes to observe how default rate varies with each of these and have used the same to come up with a list of attributes that are strong indicators of loan default(factors influencing the default rate)
  - We have used the EDA techniques like Univariate(Categorical and Numeric), Bivariate analysis, Derived metrics, correlation for the above recommendations