# Lending Club Case Study

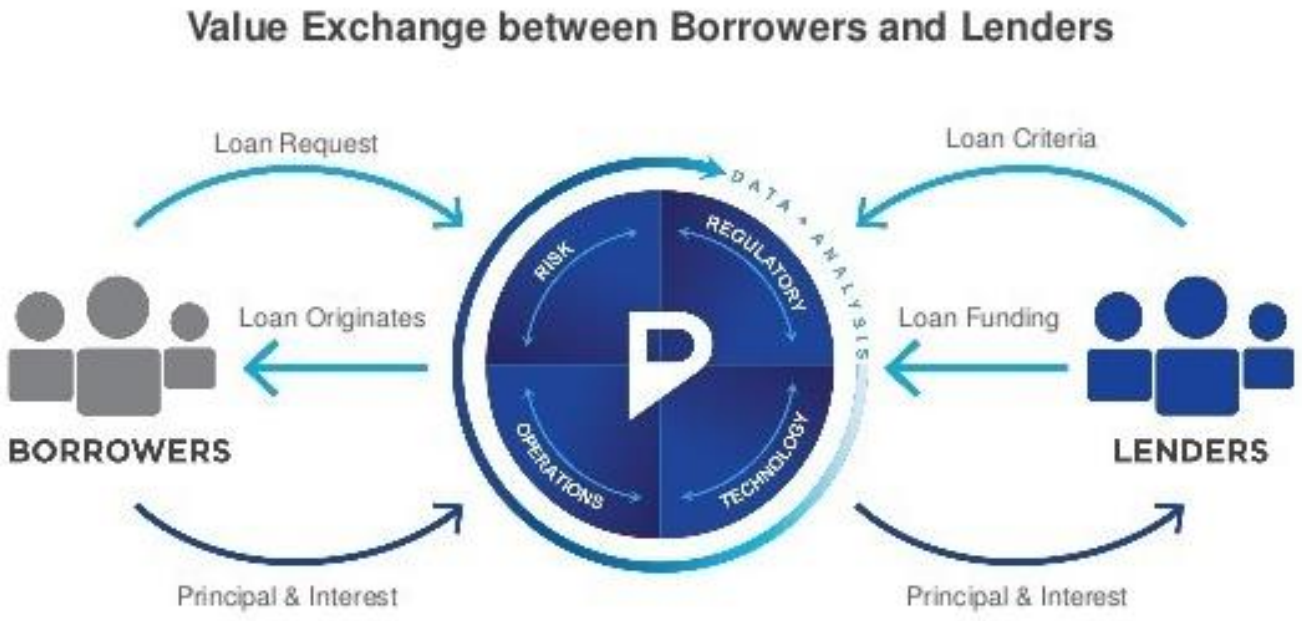
Group Members:

- ❑ Akash Agrawal
- ❑ Vishnu Vivek

## About Lending Club

Lending Club one of the world's largest peer-to-peer lending platform. It matches borrowers who are seeking for a loan with investors who are looking to lend money in the similar portfolio.

It Provides wide variety of loan like personal, business, medical etc



## Problem Statement

As the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile.

The company may face financial loss in either cases:

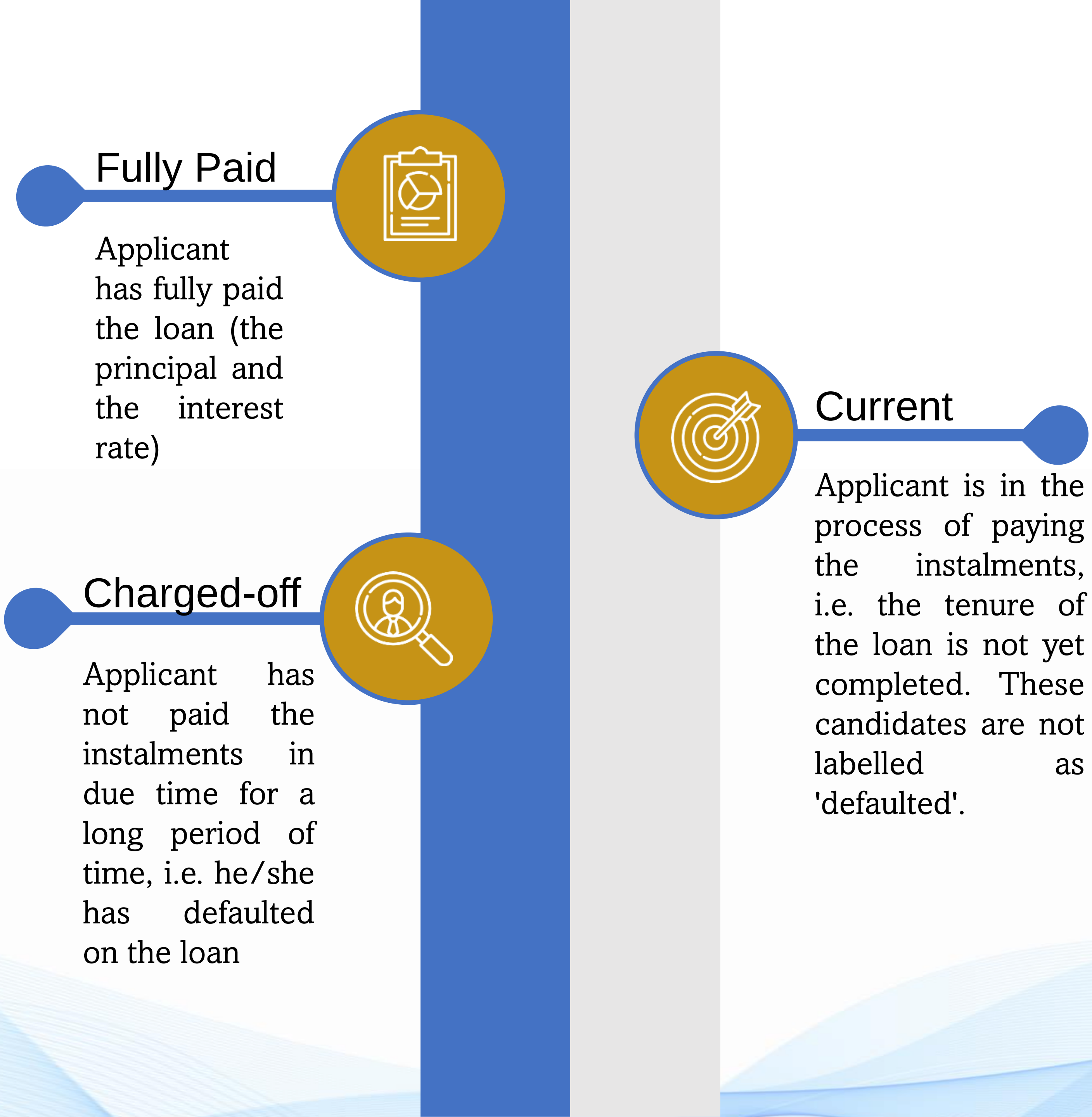
- If the applicant is likely to repay the loan, but the company is not approving the loan
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan

# Objective

The objective of the case study is to understand the driving factors/variables behind the loan default i.e., the variables which are strong indicators of default.

If the company approves the loan, there are 3 possible scenarios:

- Fully Paid
- Current
- Charged-off



## Fully Paid

Applicant has fully paid the loan (the principal and the interest rate)

## Charged-off

Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

## Current

Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

# ANALYSIS PLAN

01

**Data Understanding**
Knowing more about the columns with the help of Data Dictionary and getting knowledge of their domain specific uses

02

**Sanity check & Data preparation**
Irrelevant Variables , Missing value treatment, classifying the variables and treating the outliers

03

**Correlation**
If we keep highly correlated (assuming r > 0.8) variables in our data, then they will not add any extra information to the default status variable and will only add noise in the data

04

**Univariate Analysis**
We have used **Boxplot** for Quantitative variables and **Countplot** for Qualitative variables

05

**Bi-variate Analysis with target variable**
“Normalized stacked bar chart” was used to achieve the same

06

**Multi-variable Analysis**
Relationship of two independent variables with target variable (loan status) was plotted using heat map

07

**Recommendations**
Concluding observations and recommendations for minimizing the loss in business by finding factors contributing to default

# Sanity Check and Data Preparation

- **Sanity Check on imported data**
  - Matching  rows and columns count
  - Checking for shift of value in columns due to unwanted delimiters
  - None value correction

- **Excluding loans with loan_status 'Current':** Since our primary objective is to identify the triggers of loan default, loan in 'Current' category can either become default or fully paid in future.
  - Transform the target variable loan_status to *numerical categorical variable* and rename it to *default* And dropping the loan_status variable:
    - Fully Paid (Not Default) = 0
    - Charged Off (Default) = 1
  - This was done in order to create heatmap during multivariate analysis where depth of the heatmap is the mean default rate for a grid block (More explanation in respective section).

- **Missing values:**
  - Deleting Columns having all null/missing values
  - Deleting the Columns with higher percentage of missing values
  - Observing Missing values distribution using Heatmap

- **Irrelevant Columns and Rows:**
  - All unique and all constant value columns were dropped
  - Columns with  more than 70% unique values and nominal in nature are dropped
  - Post default/charged off variables were also dropped as default information already present in the target variable(default)
  - Remove identical rows

- **Outliers Treatment of Quantitative Variables:**
  - To identify the outliers, we have analyzed below p5 and above p95 quantiles values
  - Box plot created to analyze the data spread
  - The concept of flooring and ceiling used to replace the outliers

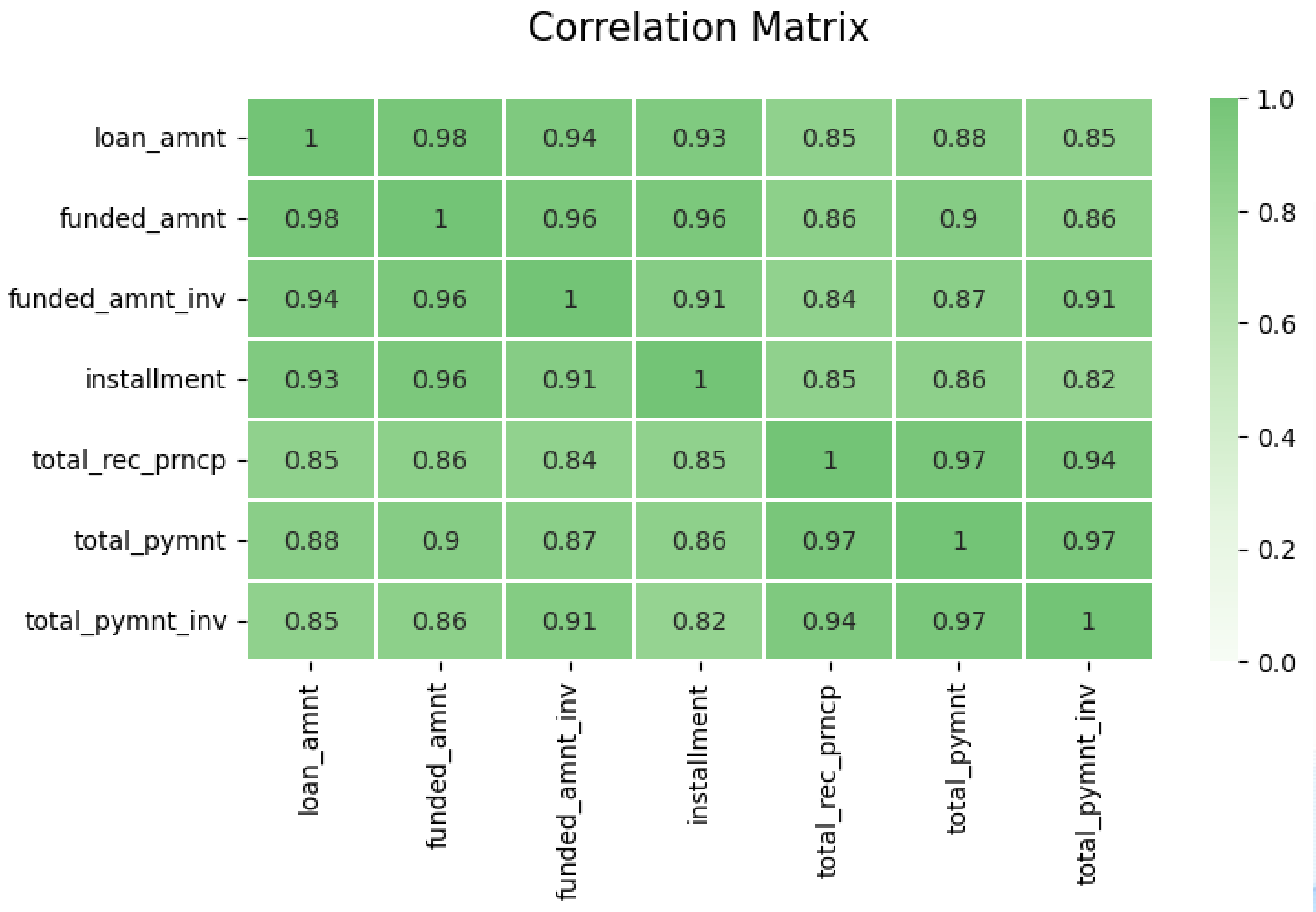
# Excluding variables with high Correlation

- Two highly correlated quantitative variables can have nearly the same ability to predict the outcome value. In our case, the outcome is loan default status. Correlation is scaled from 0 (no correlation) to 1 (max correlation).

- If we keep highly correlated (assuming r > 0.8) variables in our data then they will not add any extra information to the default status variable and will only add noise in the data.

- we have selected 'loan_amnt' variable from this block and all other variables are dropped from our further analysis.

- Similarly, co-relation matrix were drawn between grade and sub-grade, we found that sub-grade is made out of grade only.



Correlation Matrix

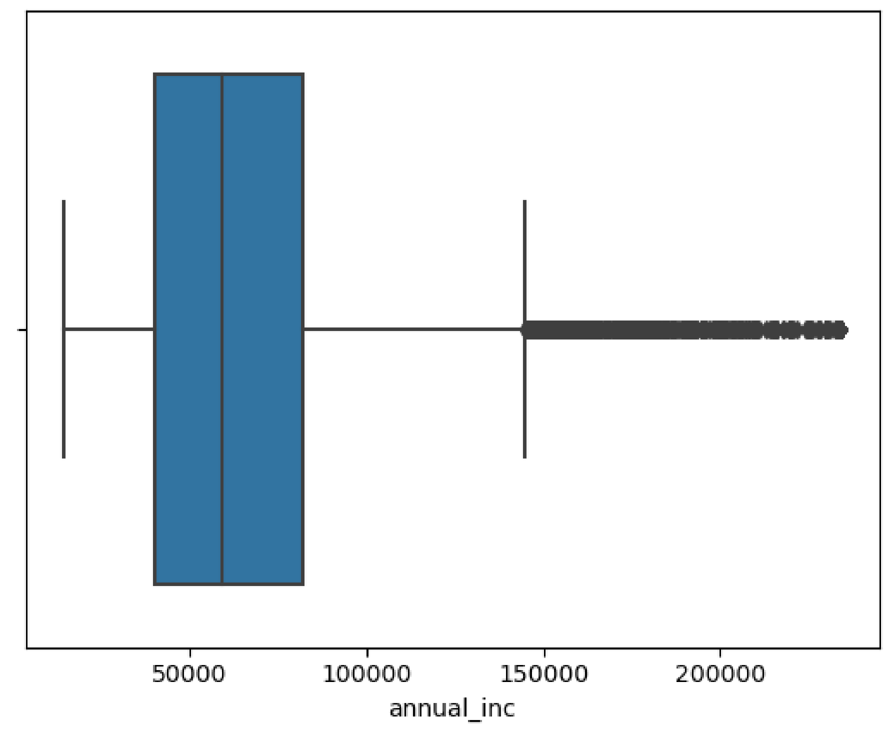
| | loan_amnt | funded_amnt | funded_amnt_inv | installment | total_rec_prncp | total_pymnt | total_pymnt_inv |
|---|---|---|---|---|---|---|---|
| loan_amnt | 1 | 0.98 | 0.94 | 0.93 | 0.85 | 0.88 | 0.85 |
| funded_amnt | 0.98 | 1 | 0.96 | 0.96 | 0.86 | 0.9 | 0.86 |
| funded_amnt_inv | 0.94 | 0.96 | 1 | 0.91 | 0.84 | 0.87 | 0.91 |
| installment | 0.93 | 0.96 | 0.91 | 1 | 0.85 | 0.86 | 0.82 |
| total_rec_prncp | 0.85 | 0.86 | 0.84 | 0.85 | 1 | 0.97 | 0.94 |
| total_pymnt | 0.88 | 0.9 | 0.87 | 0.86 | 0.97 | 1 | 0.97 |
| total_pymnt_inv | 0.85 | 0.86 | 0.91 | 0.82 | 0.94 | 0.97 | 1 |

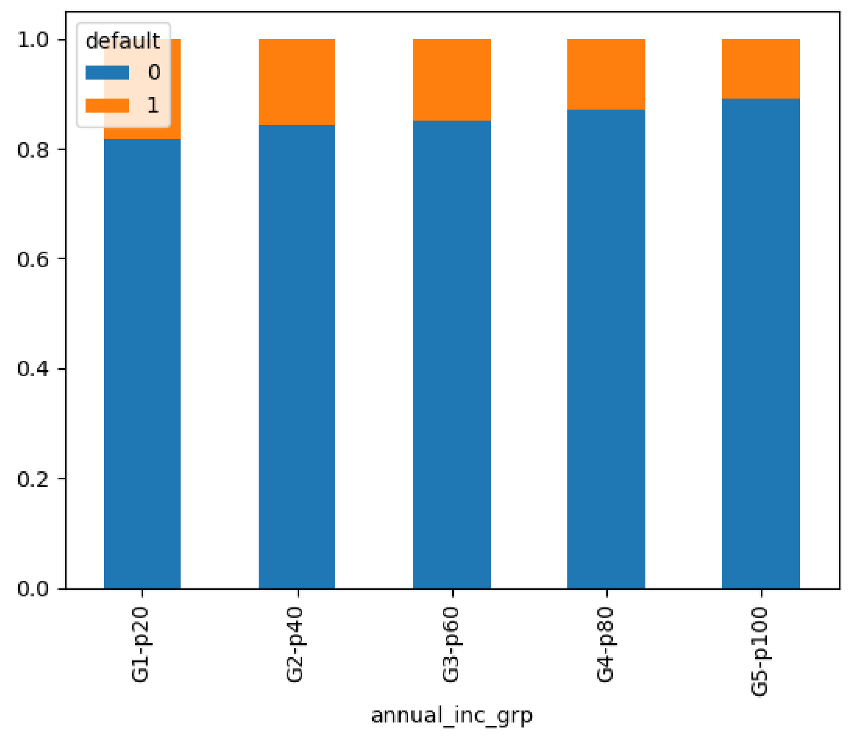
# Quantitative Variables Analysis

- Box-plot to observe skewness and spread of the data
- Normalized Stack Bar Chart is created to observe trend with target variable

## annual_inc & annual_inc_grp

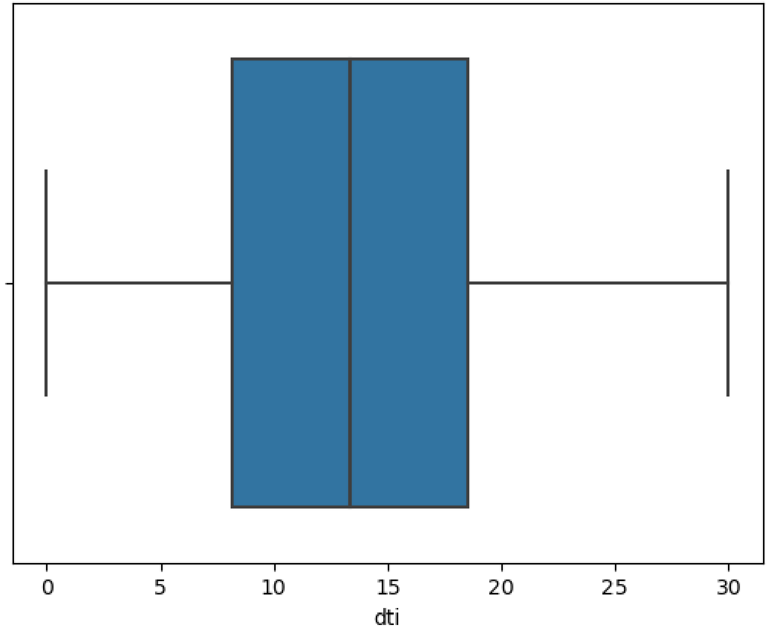
- ❑ Right Skewed Data
- ❑ ‘Strong trend’ observed with the target variable. As the *income in the group* increasing, default rate is decreasing. People with higher income are more likely to pay their loans.

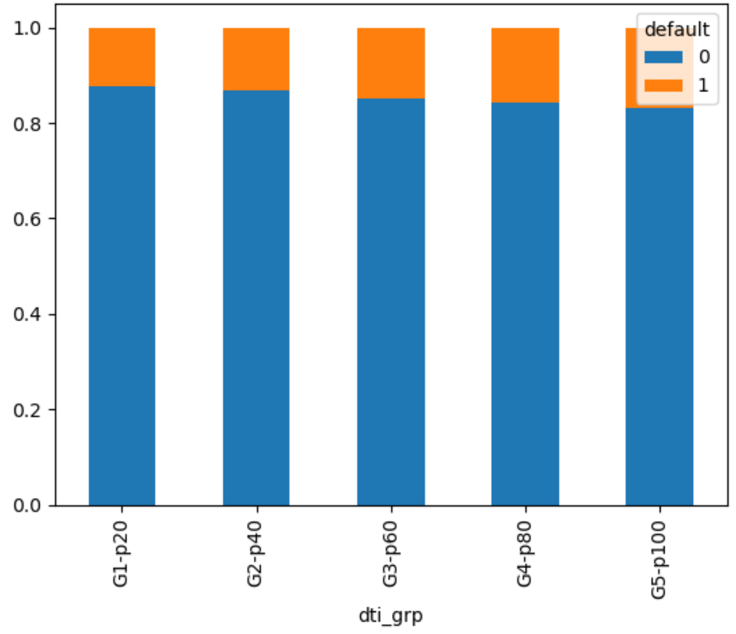




## dti & dti_grp

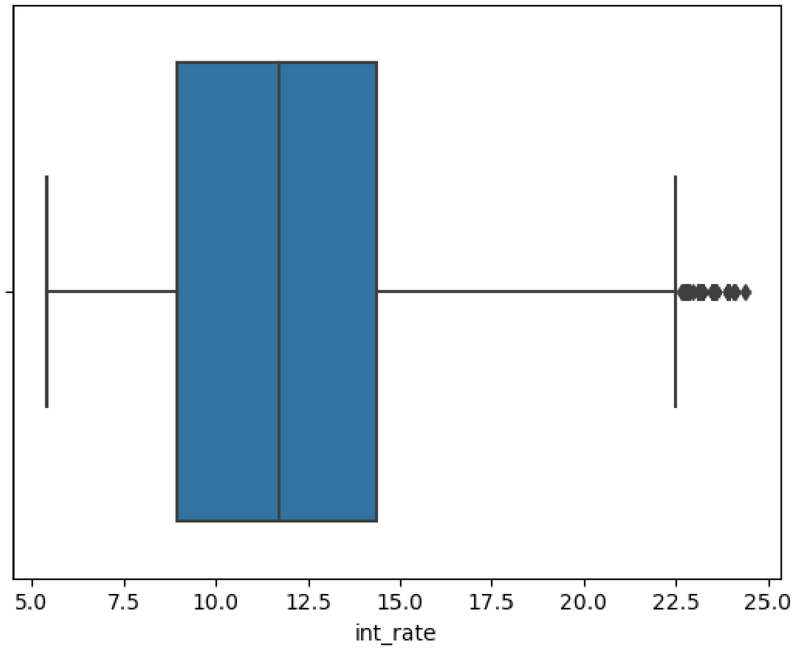
- ‘Strong trend’ observed. As the *debt-to-income* ratio increasing, default to count ratio is increasing.

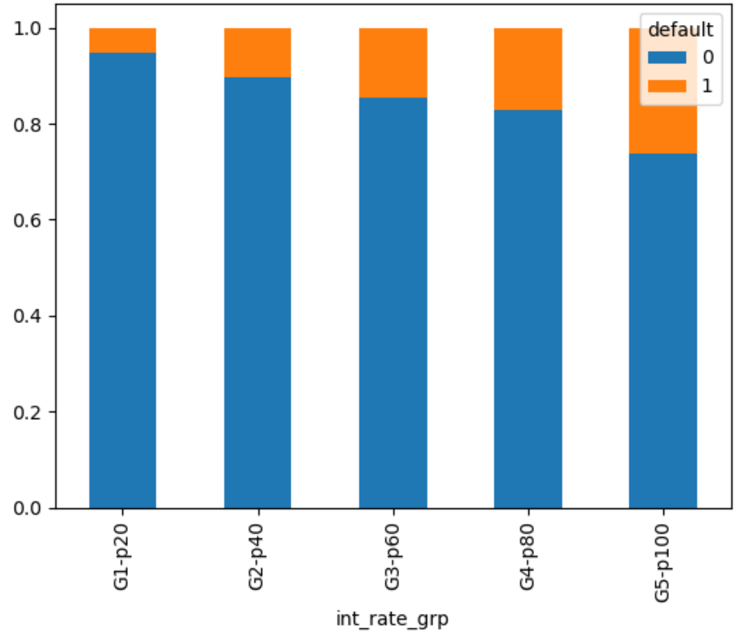




## int_rate & int_rate _grp

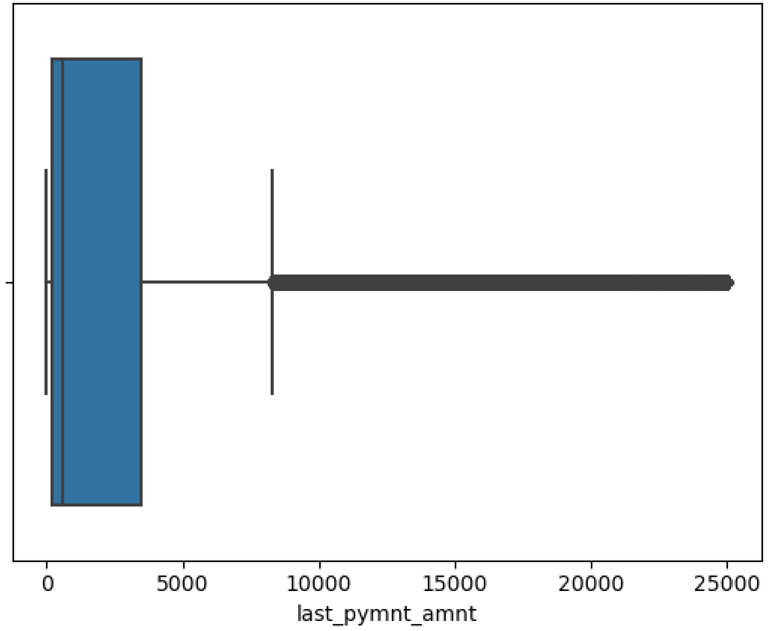
- ‘Strong trend’ observed. As the *loan interest rate* is increasing, default to count ratio is increasing. It means that higher risky loans have higher interest rate.

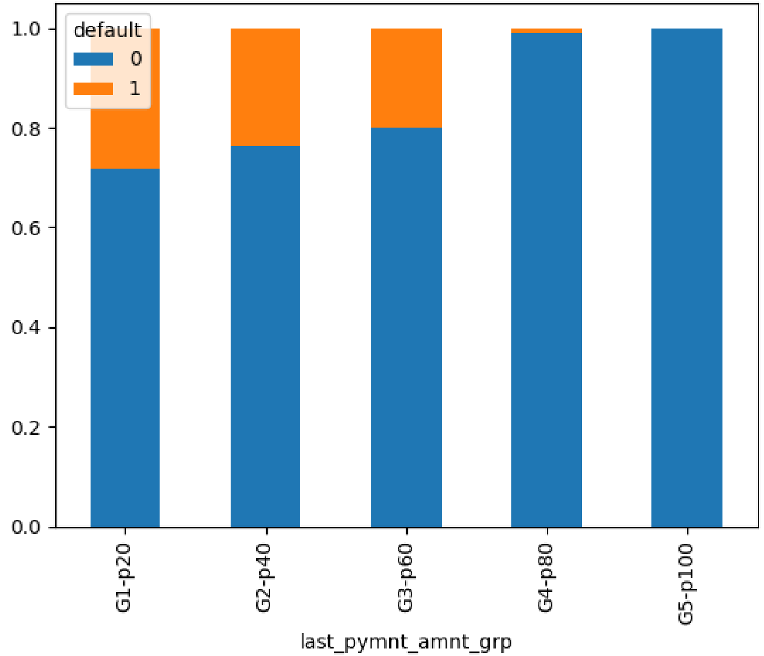




## last_pymnt_amnt & last_pymnt_amnt_grp

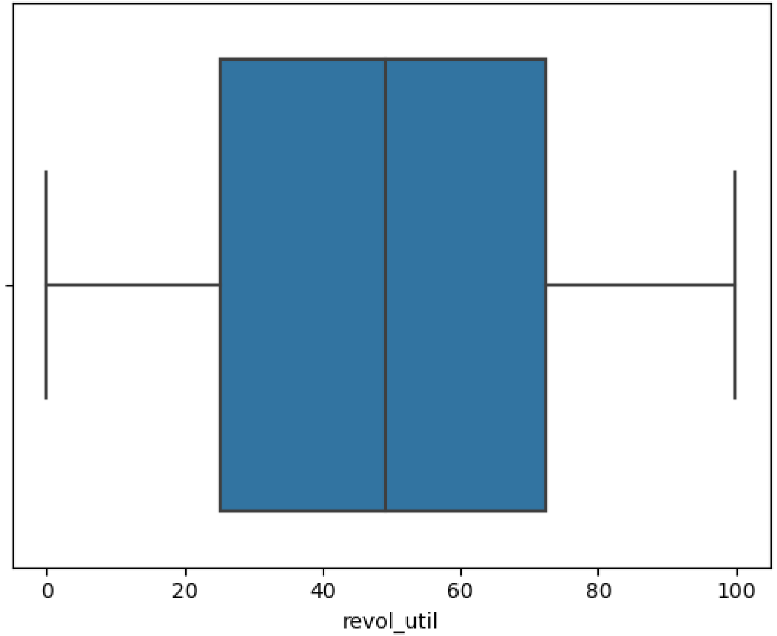
- ❑ Right Skewed Data
- ❑ **Strong trend** observed. People who last payment amount is higher their chance of getting default is lower. It means that these people are financially strong and therefore able to pay their obligations

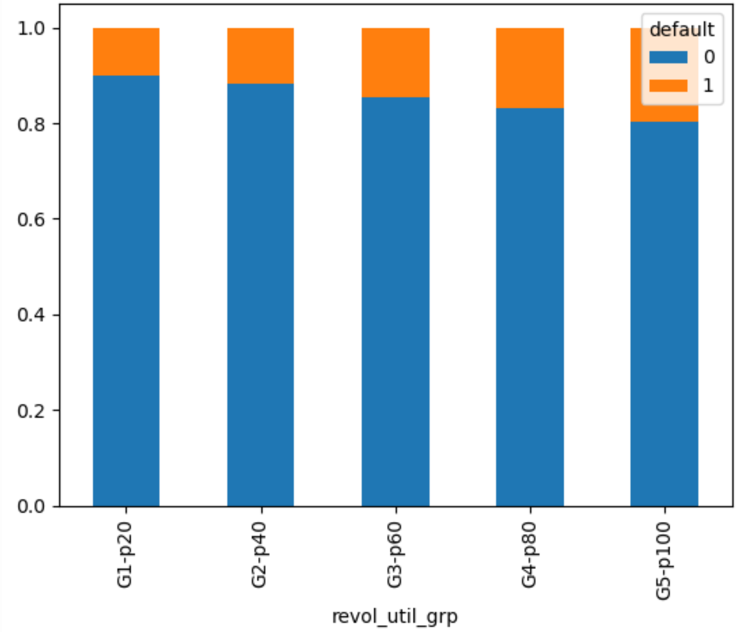




## revol_util & revol_util _grp

- ❑ **Strong trend** observed. As the utilization rate of revolving credit increasing, defaults occurrence is also increasing.

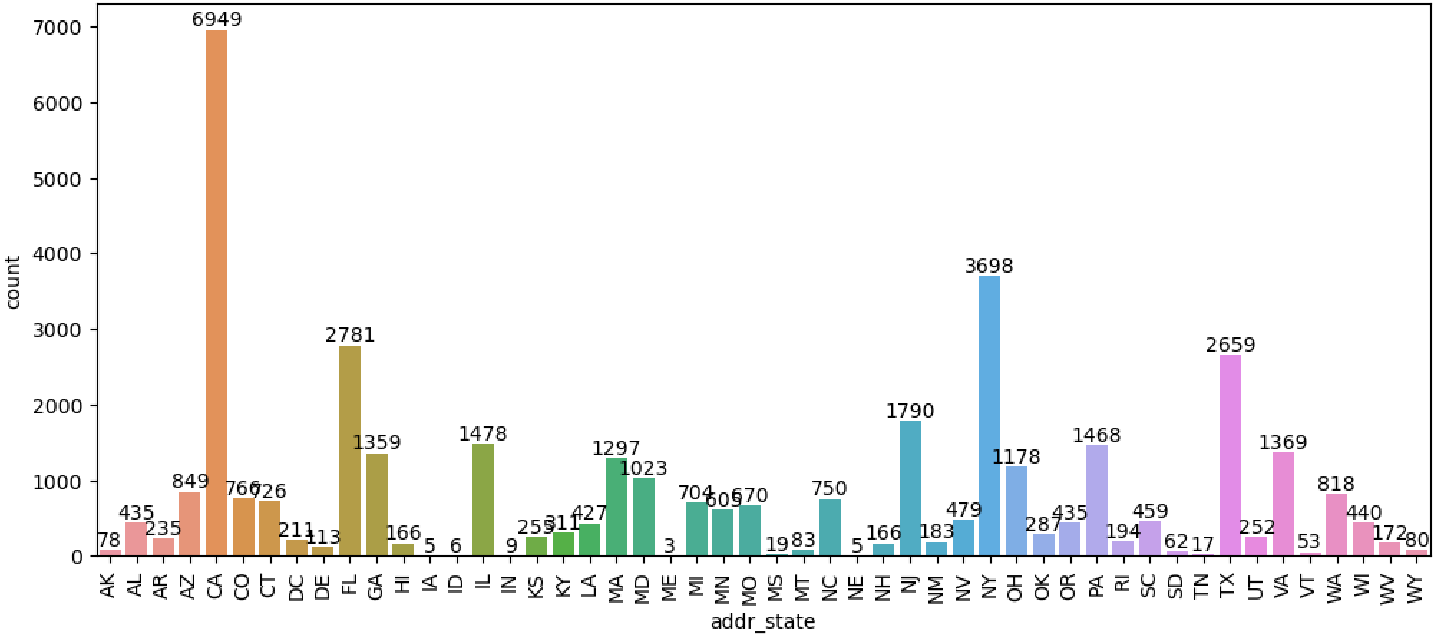


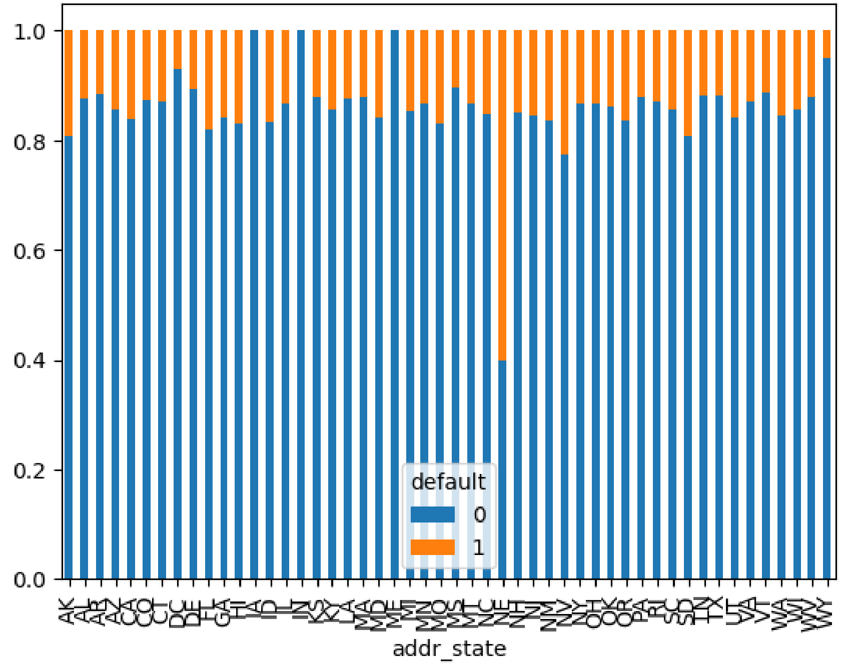


# Qualitative Variables Analysis

- Count-plot created to observe the frequency distribution of the data
- Normalized Stack Bar Chart created to observe relationship with target variable

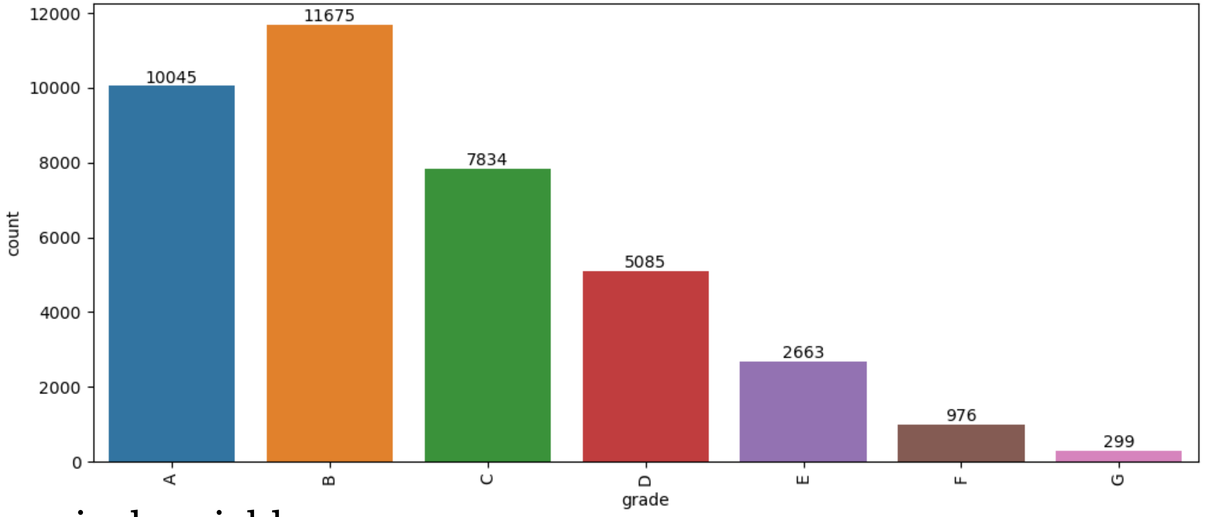
## Addr_state

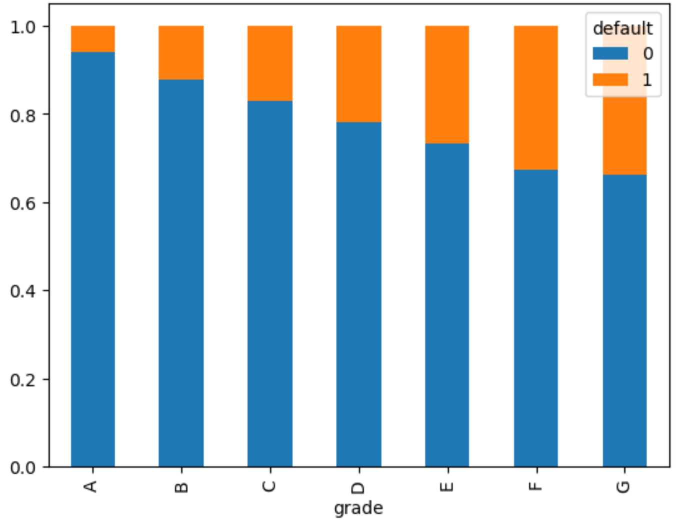




- Nominal categorical variable. These are the states of residence of loan applicants at the time of loan application.
- It can be observed that highest number of loans approved come from CA (California) state, followed by NY (New York), FL (Florida) and TX (Texas) states in descending order.
- The highest default rate is observed for NE (Nebraska) state, however for this state loan status available for only 5 loans. Therefore, this observation might not be true.
- There is **no trend** observed as loan default rate is similar across all the states.

## grade

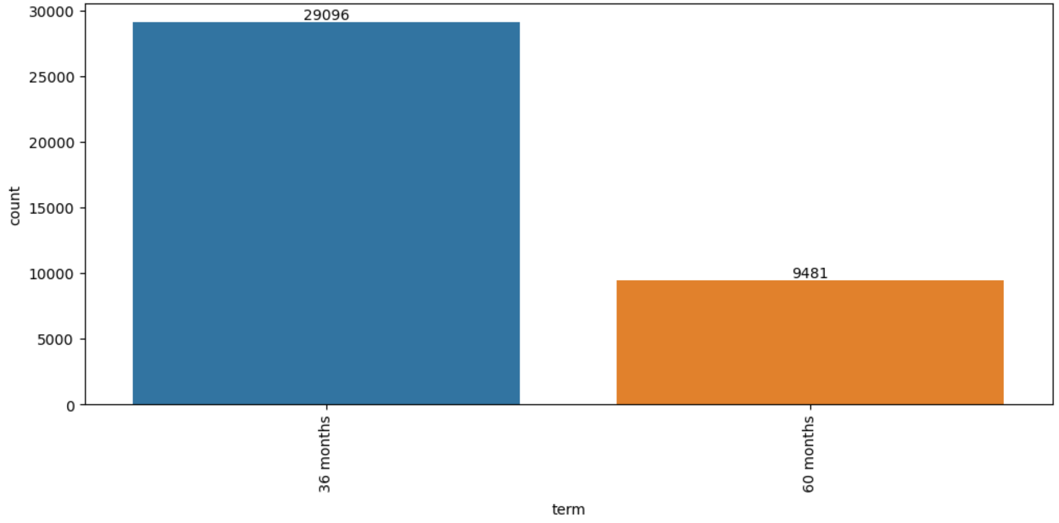


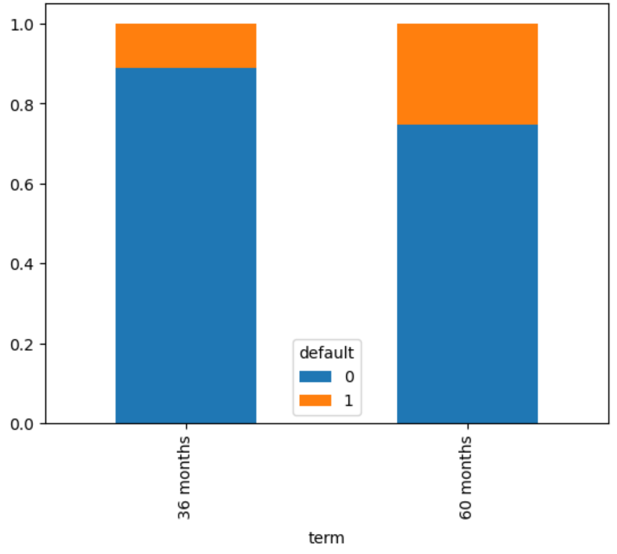


- Interval categorical variable
- As grade decreases, count of loans sanctioned also decreases. **Strong trend** observed. As the loan grade worsens, defaults to count ratio is increasing. Which means higher is grade, riskier is the loan.

## term

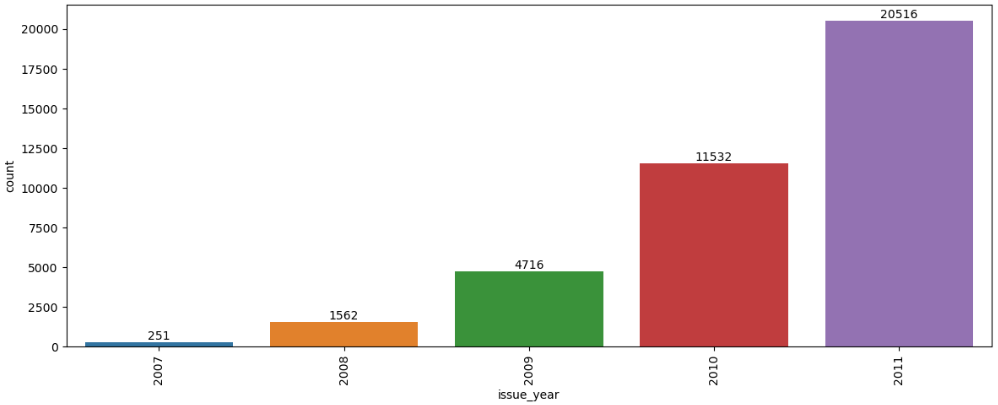
- Ratio categorical variable
- Count of loans is decreasing as loan tenure is increasing.
- **Strong trend** observed with the target variable. Loan with 60 months of tenure has higher chances of getting default.

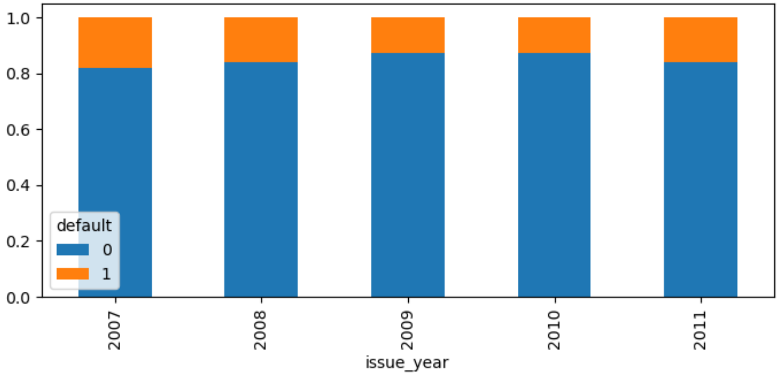




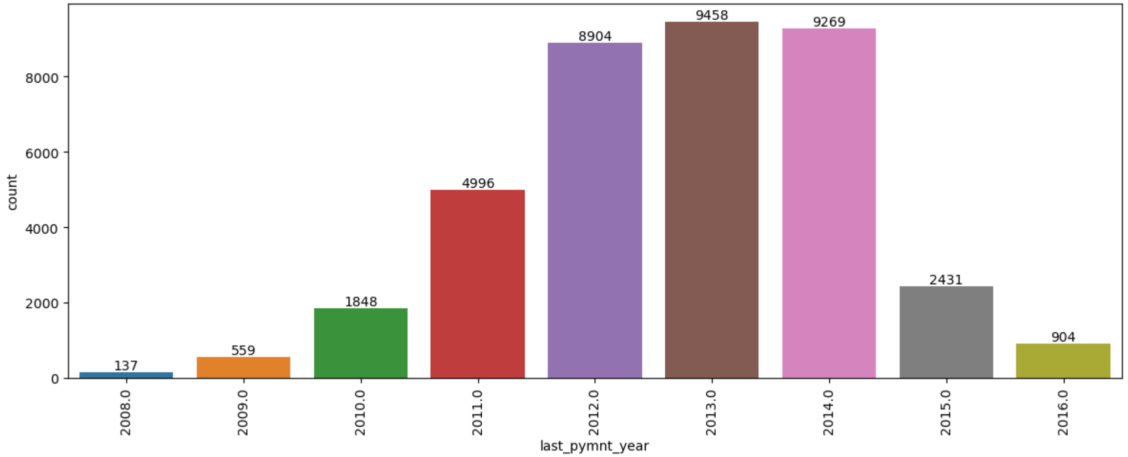
## issue_year

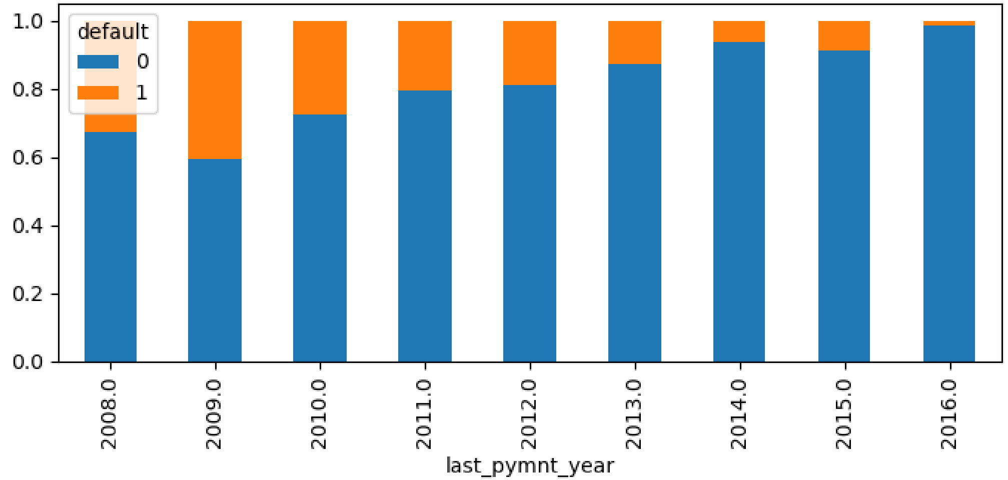
- We have observed an increasing trend of loan count for years 2007 to 2011.
- **No trend** observed with the target variable





## last_pymnt_year



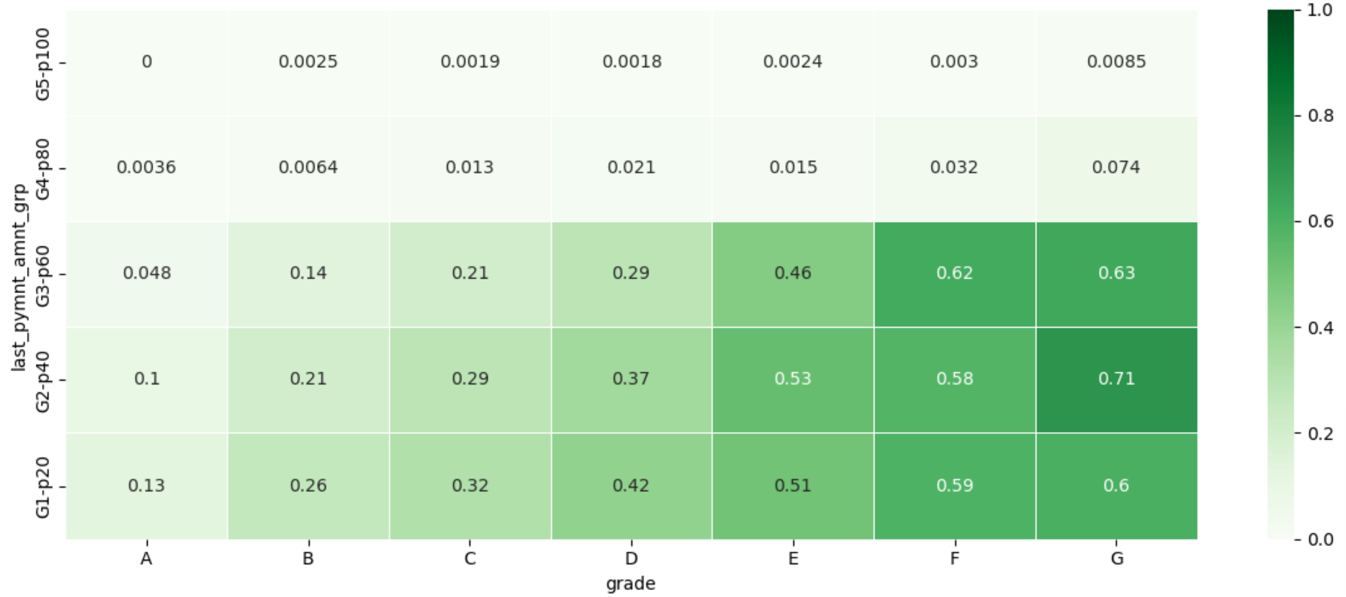


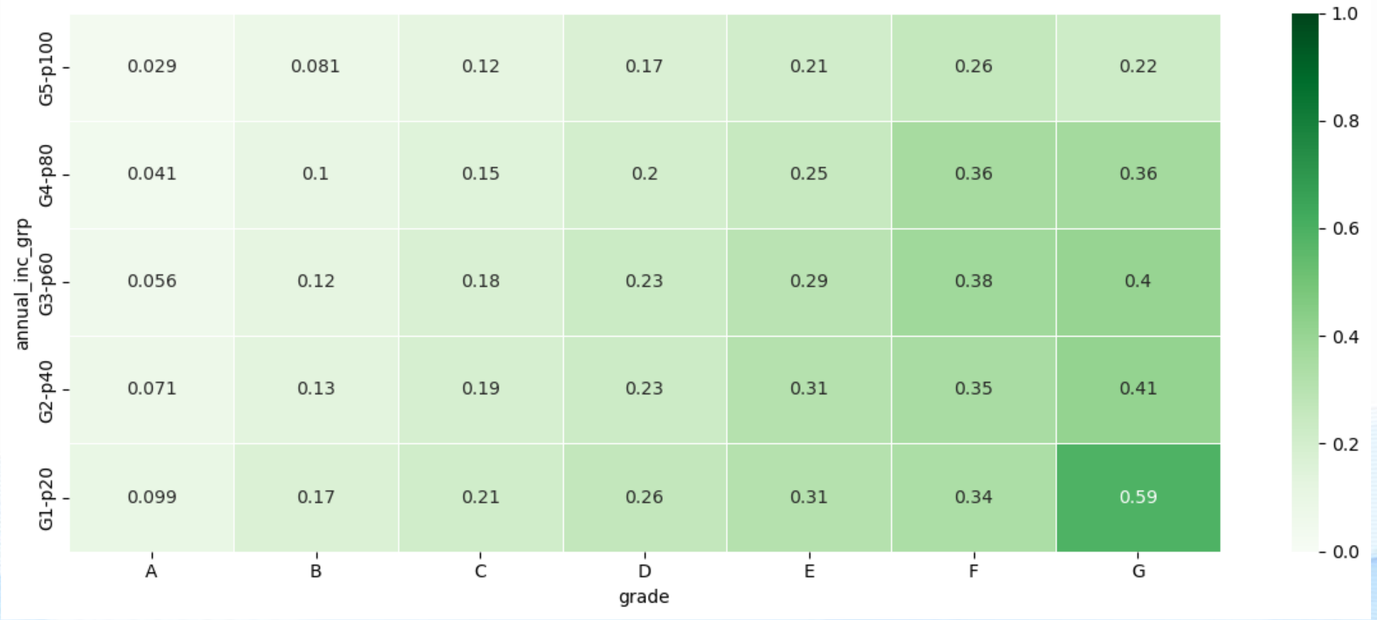
- It is clearly visible that from year 2008 to 2010, default to count ratio is decreasing despite having low count of payment done by borrower. This may be attributed to subprime crisis in USA in 2008 and people were unable to pay for their loans. However, this trend is not cyclic thus can not be predictor to loan default status.
- On the contrary, in 2012 to 2014, payment done by borrowers is maximum, which shows improvement in financial capacity of borrowers and lesser economic stress.

# Multi-variables analysis with target variable

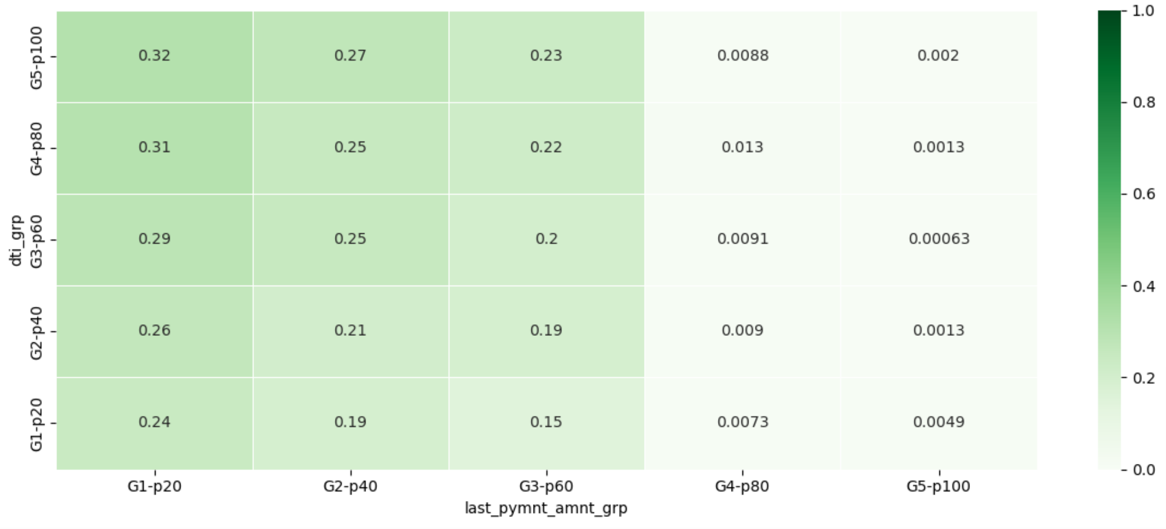
- To perform multivariable analysis with target variable, we have selected those variables who have shown strong trend with default variables.

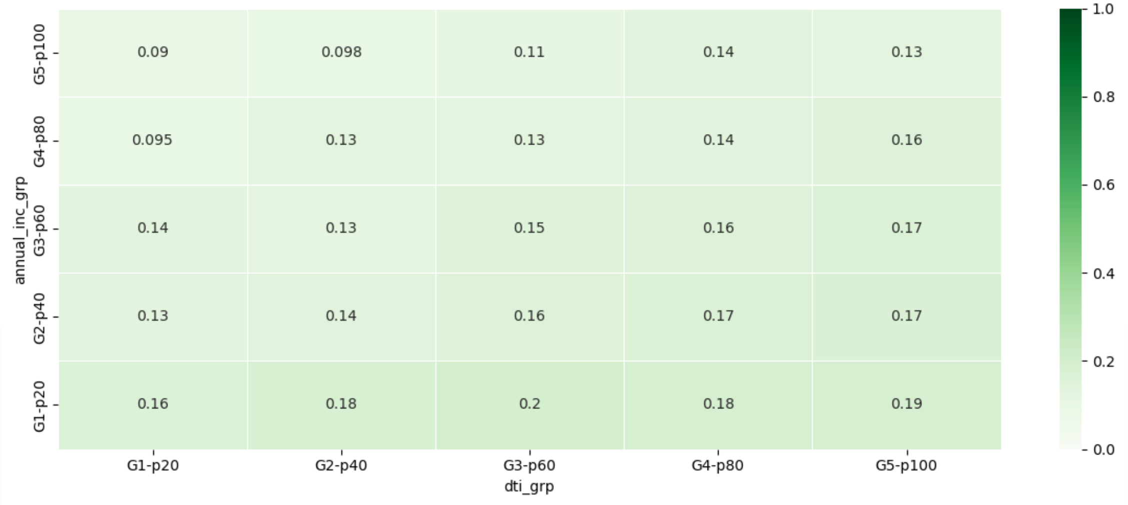
- We will be using heat map for conducting the Multi-variable analysis with target variable

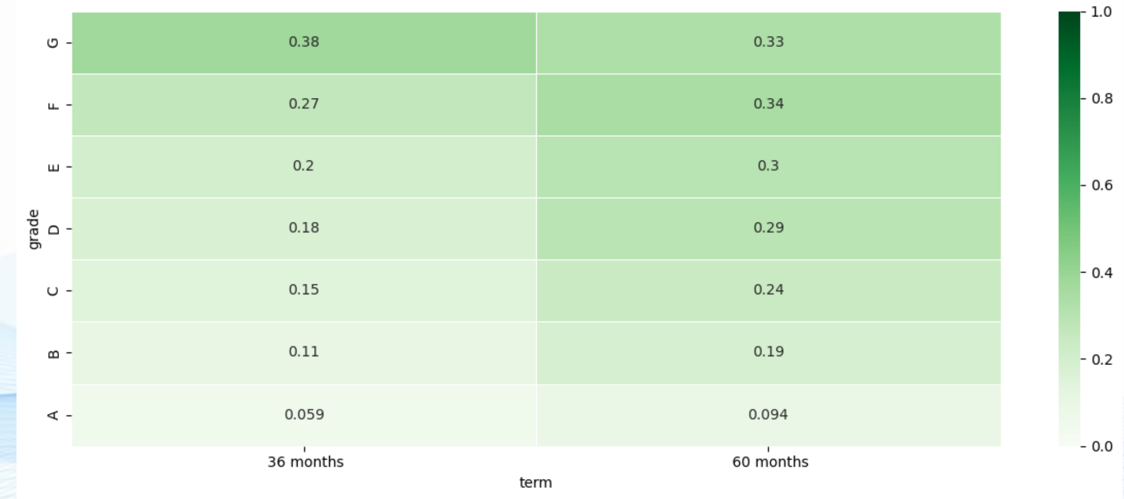


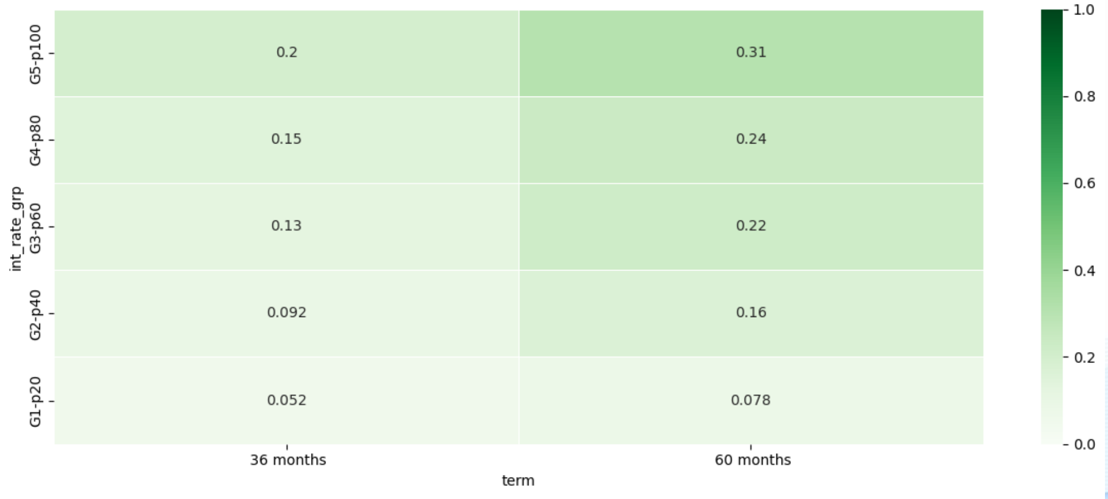


# Multi-variables analysis with target variable

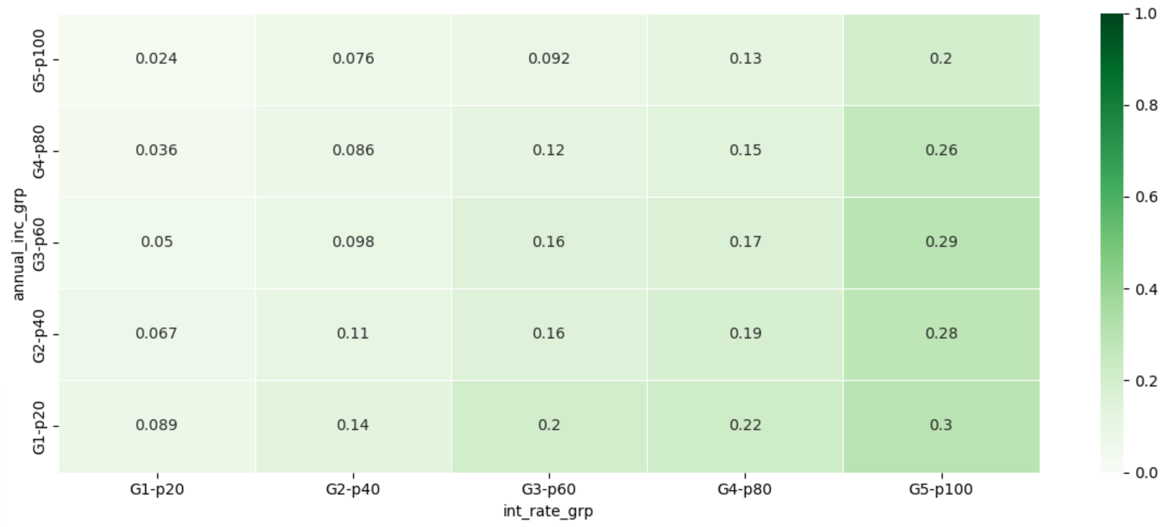




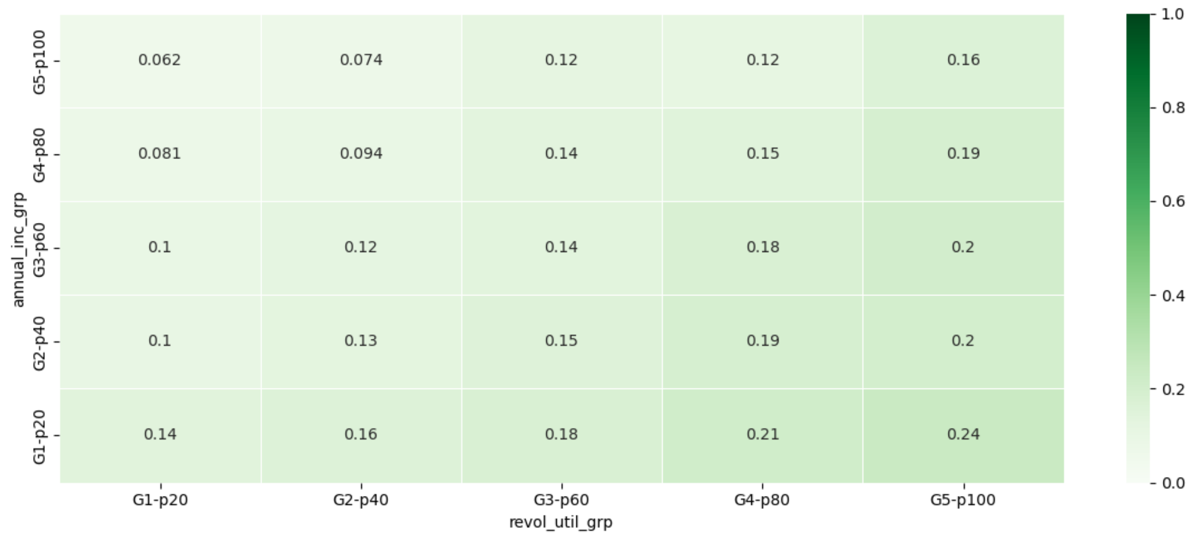




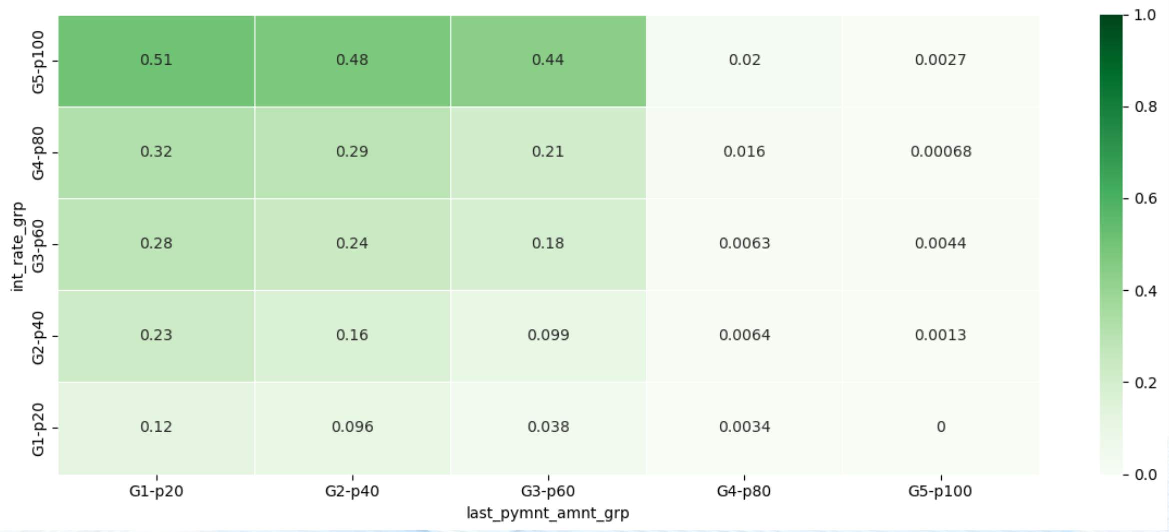
# Multi-variables analysis with target variable



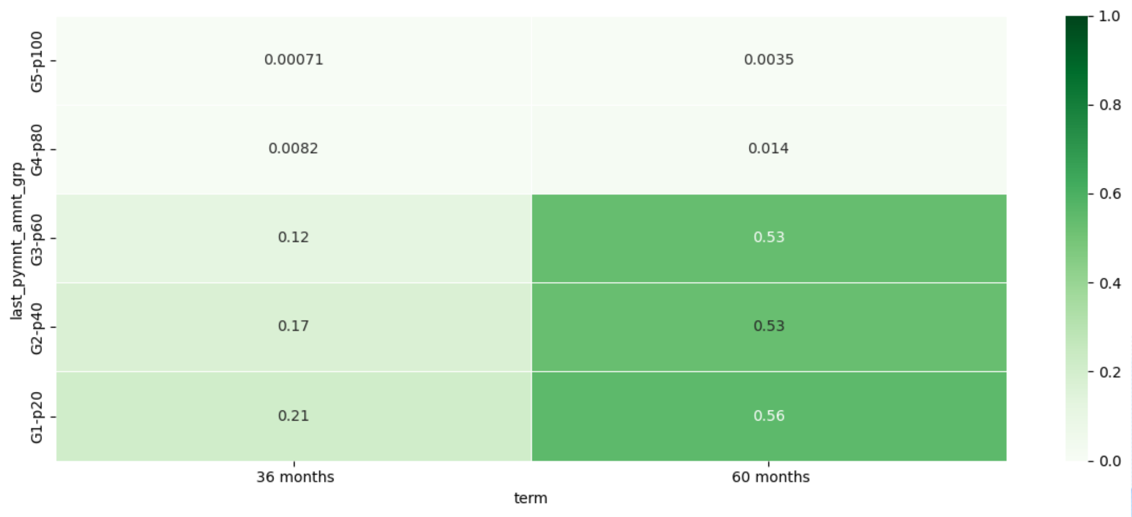
| annual_inc_grp \ int_rate_grp | G1-p20 | G2-p40 | G3-p60 | G4-p80 | G5-p100 |
|---|---|---|---|---|---|
| G5-p100 | 0.024 | 0.076 | 0.092 | 0.13 | 0.2 |
| G4-p80 | 0.036 | 0.086 | 0.12 | 0.15 | 0.26 |
| G3-p60 | 0.05 | 0.098 | 0.16 | 0.17 | 0.29 |
| G2-p40 | 0.067 | 0.11 | 0.16 | 0.19 | 0.28 |
| G1-p20 | 0.089 | 0.14 | 0.2 | 0.22 | 0.3 |



| annual_inc_grp \ revol_util_grp | G1-p20 | G2-p40 | G3-p60 | G4-p80 | G5-p100 |
|---|---|---|---|---|---|
| G5-p100 | 0.062 | 0.074 | 0.12 | 0.12 | 0.16 |
| G4-p80 | 0.081 | 0.094 | 0.14 | 0.15 | 0.19 |
| G3-p60 | 0.1 | 0.12 | 0.14 | 0.18 | 0.2 |
| G2-p40 | 0.1 | 0.13 | 0.15 | 0.19 | 0.2 |
| G1-p20 | 0.14 | 0.16 | 0.18 | 0.21 | 0.24 |

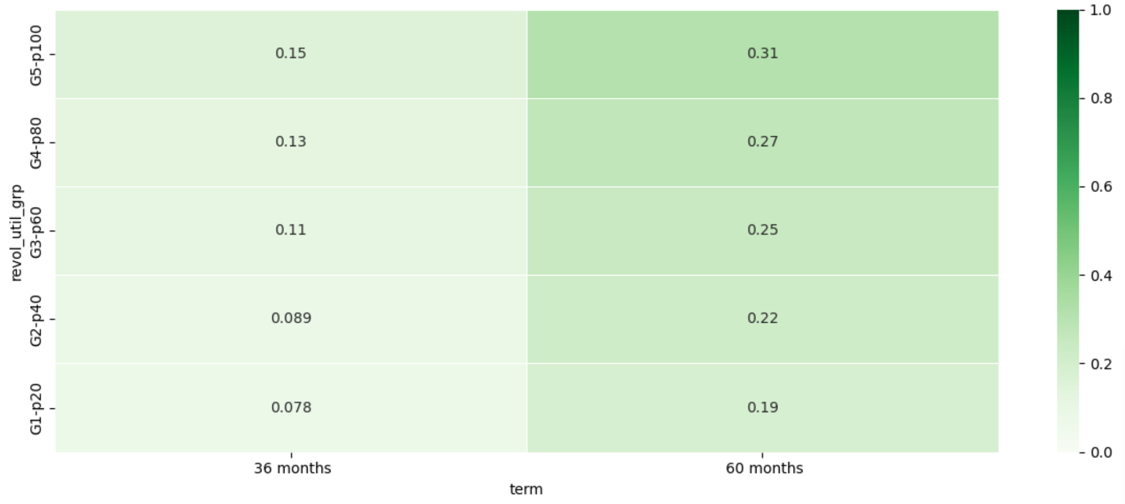


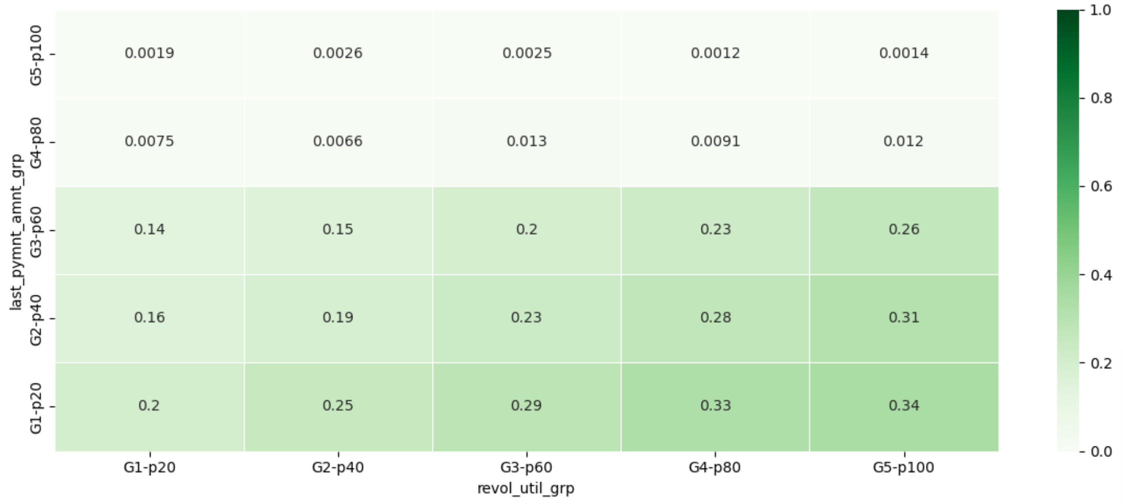
| int_rate_grp \ last_pymnt_amnt_grp | G1-p20 | G2-p40 | G3-p60 | G4-p80 | G5-p100 |
|---|---|---|---|---|---|
| G5-p100 | 0.51 | 0.48 | 0.44 | 0.02 | 0.0027 |
| G4-p80 | 0.32 | 0.29 | 0.21 | 0.016 | 0.00068 |
| G3-p60 | 0.28 | 0.24 | 0.18 | 0.0063 | 0.0044 |
| G2-p40 | 0.23 | 0.16 | 0.099 | 0.0064 | 0.0013 |
| G1-p20 | 0.12 | 0.096 | 0.038 | 0.0034 | 0 |

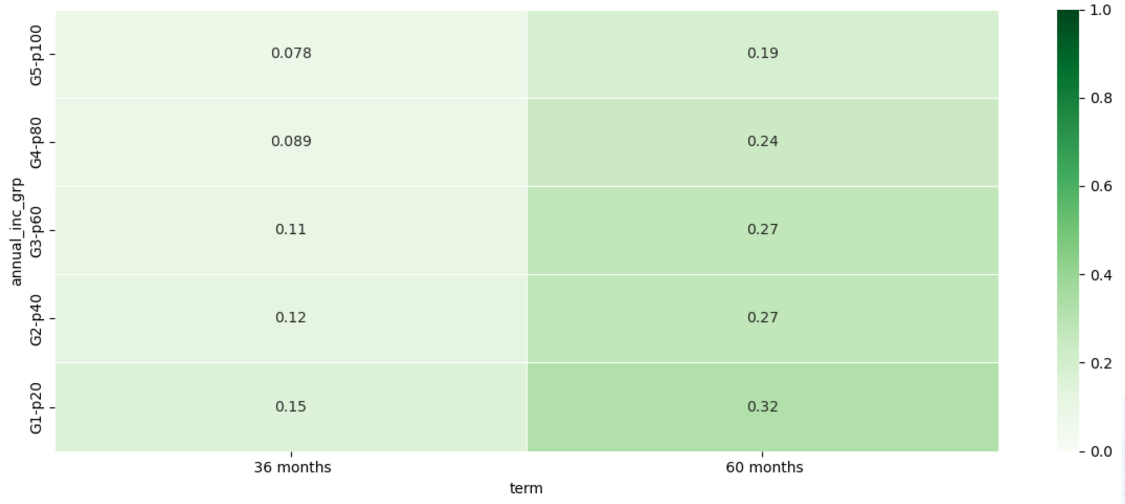


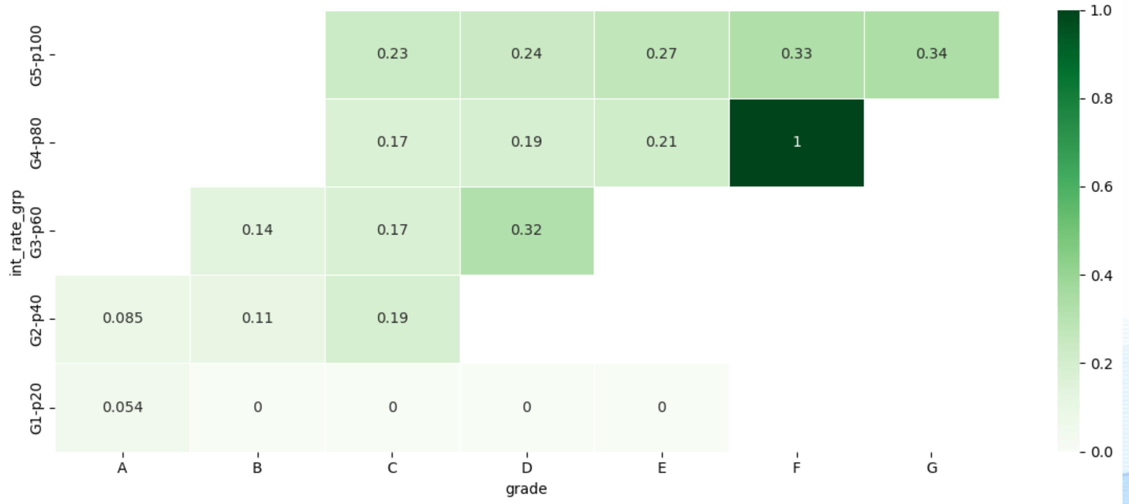
| last_pymnt_amnt_grp \ term | 36 months | 60 months |
|---|---|---|
| G5-p100 | 0.00071 | 0.0035 |
| G4-p80 | 0.0082 | 0.014 |
| G3-p60 | 0.12 | 0.53 |
| G2-p40 | 0.17 | 0.53 |
| G1-p20 | 0.21 | 0.56 |

# Multi-variables analysis with target variable

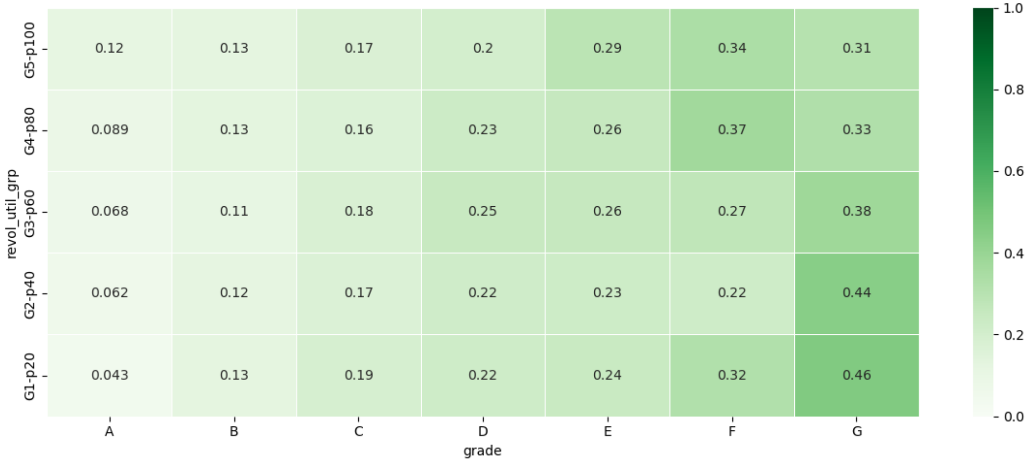


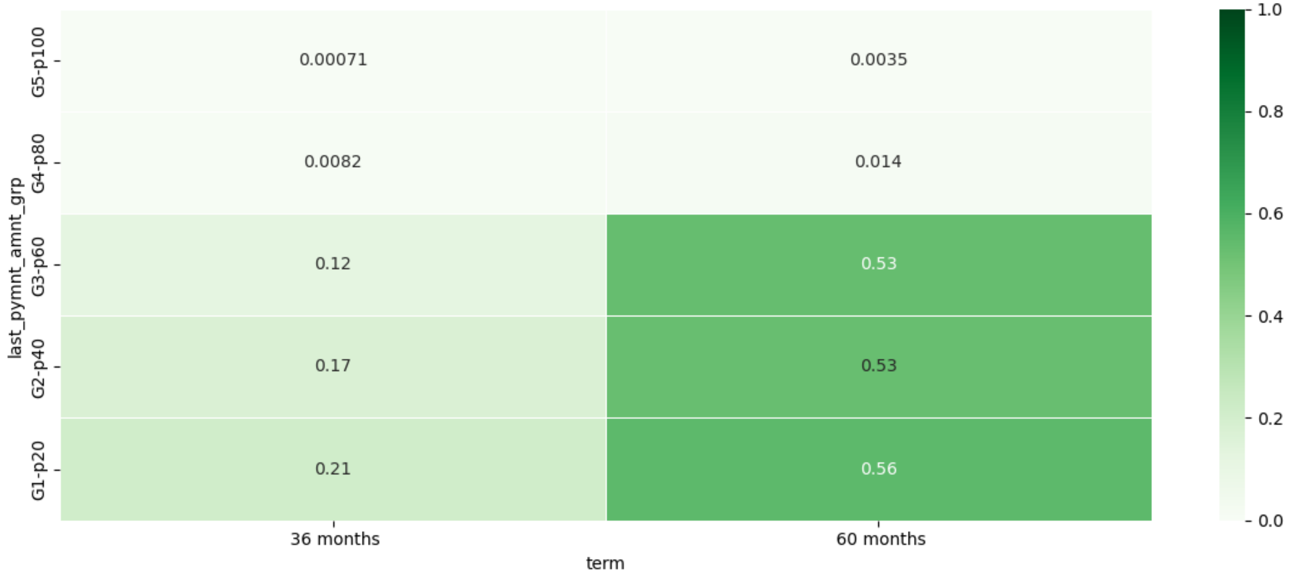


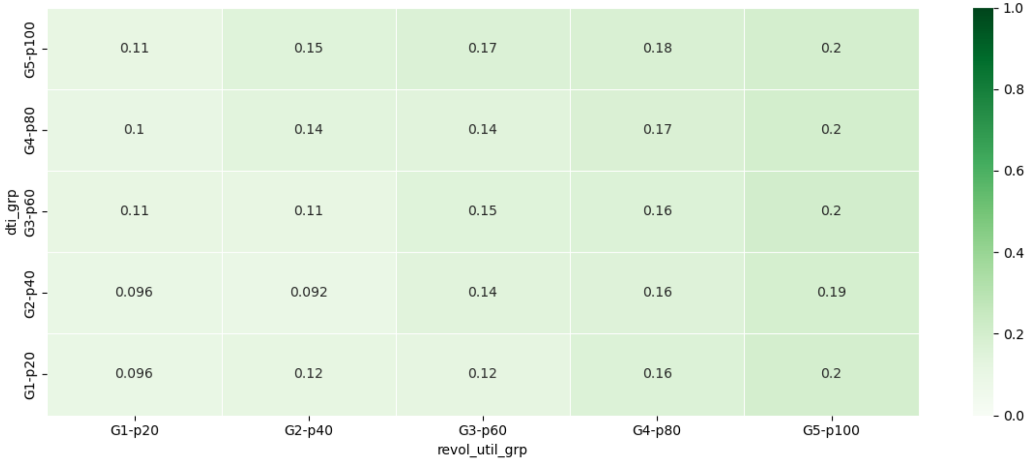




# Multi-variables analysis with target variable







# Conclusion and Recommendation

After analyzing all the **multivariate charts**, we can clearly observe the **potentially strong indicators/ predictors** of loan default. These are following:

- **annual_inc -** The self-reported annual income provided by the borrower during registration.
- **grade -** LC assigned loan grade
- **last_pymnt_amnt -** Last total payment amount received
- **term -** The number of monthly payments on the loan.
- **int_rate -** Interest Rate on the loan
- **dti -** A ratio calculated using the borrower's total monthly debt divided by the borrower's self-reported monthly income.
- **revol_util -** Amount of credit the borrower is using relative to all available revolving credit.

Out of these 7 indicators, **annual_inc, grade and last_pymnt_amnt** are most potential indicators for identifying loan defaults.

**Reason for their selection:**
There are few more variables which have shown strong trends with loan default status, however those variables are not as good indicators as above mentioned variables because these selected variables individually and in combination with each other, separate the loans with higher default rate and not nullifying their combining impact on loan default rate.
These variables, in together, have increased the degree of separability between loan default or not-default, thus are potential indicators.
***Therefore, achieving the objective of problem statement of Lending Club Case Study.***