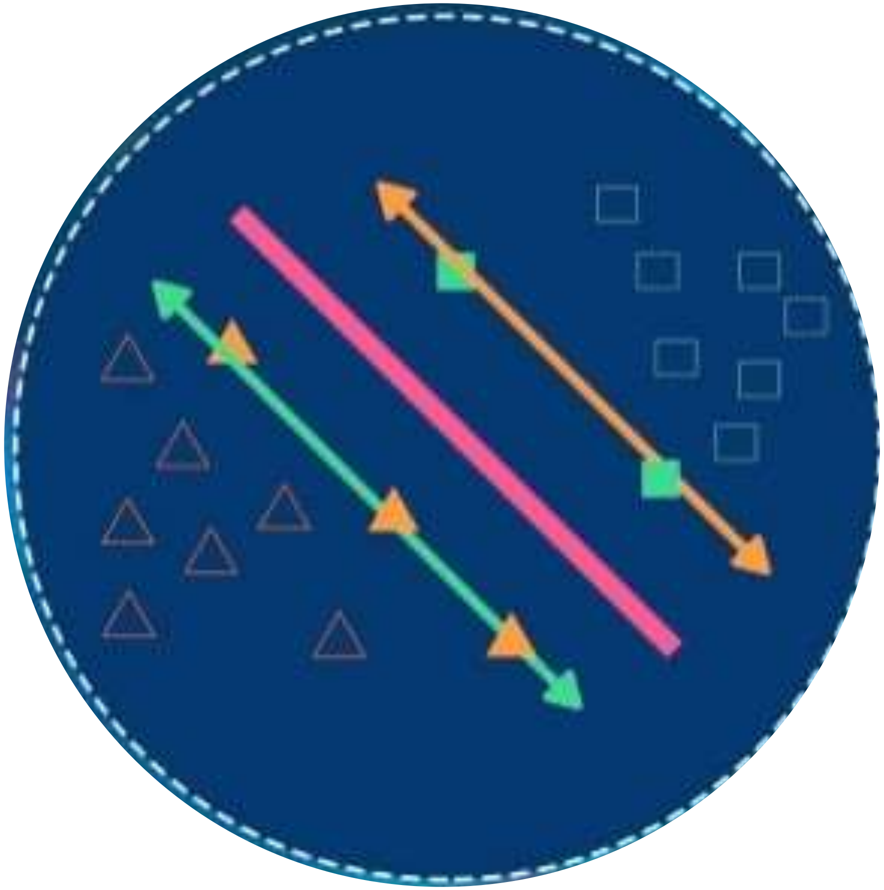


Lending Club Case Study

10th August, 2022

Graison Thomas

Table of Contents

Problem Statement

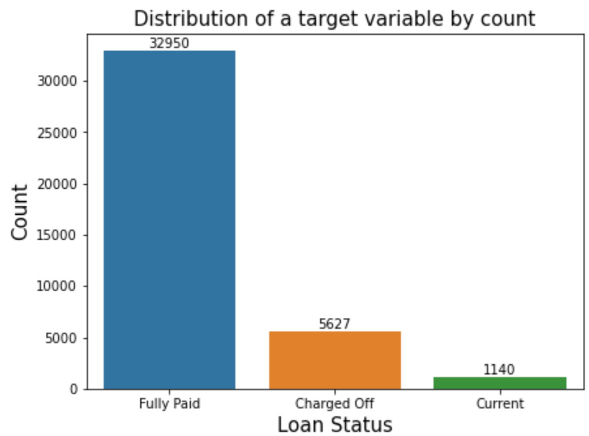
- Lending Club is a marketplace for personal loans that matches individual borrowers with investors looking to lend money.
- To be profitable the company should be able to correctly price credit risk. For this it needs to correctly quantify the credit risk and charge appropriate interest rate to account for the expected defaults.
- The business objective of this assignment is to look at the historical loan data set provided by the company and try to identify key factors which determine credit risk (likelihood of default in the future).

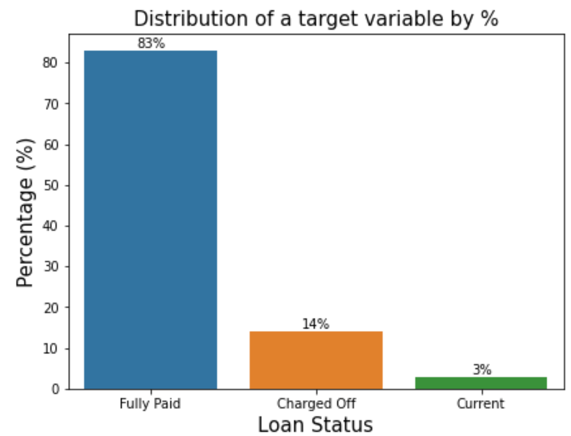
Analysis Approach

- We perform an Exploratory Data Analysis on the historical data asset provided to identify key drivers behind default
- As part of this we perform the following
  - Target Variable Identification (Loan Status)
  - Data Cleaning
    - Correct any data quality issues
    - Identify missing values
    - Convert data to more usable format where required
    - Date formats are correctly captured
    - Drop non-relevant variables
    - Perform outlier treatment where required
  - Derived Metrics
    - New variables are created from existing variables for better information
  - Univariate Analysis
    - Univariate analysis is performed to better understand each variable
  - Bivariate Analysis
    - Bivariate Analysis is performed to assess the relationship between individual variables and target variable

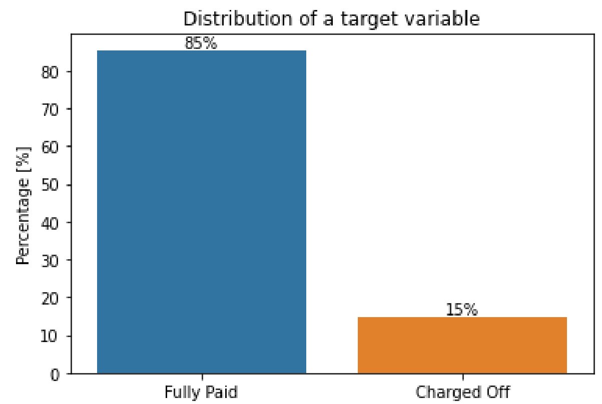
Target Variable Identification

Provided below is the distribution of the target variable Loan_Status countwise and % wise





We drop the records with Loan Status "Current" as the performance history is not complete and it comprises only 3% of data. Provided below is distribution of the target variable % wise after dropping Current accounts.



Target Variable Transformation

For the purpose of our analysis it is better to convert the target variable to numeric with 1 denoting default/chargeoff and 0 denoting fully paid.

We create a new variable based on this logic

The summary statistics for the new target variable is shown below

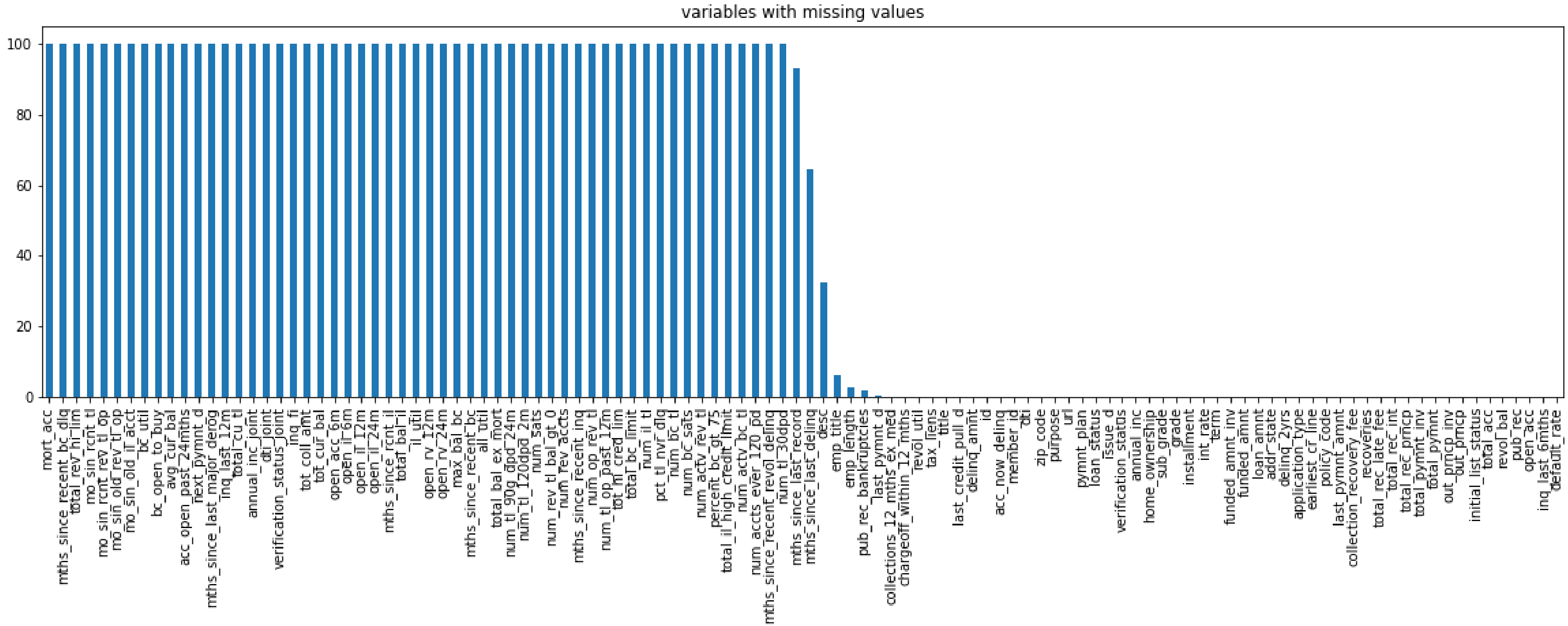
count	38577
mean	0.146
std	0.353
min	0
25%	0
50%	0
75%	0
max	1

It can be seen that the mean for this variable is same as the default rate for the overall data set after removing Current loans.

We will be using this new target variable for further analysis and drop Loan_Status variable.

Data Quality Issues – Columns with missing values

Below you can see a visualization for the percentage of missing values for various columns



We drop columns which have 100% missing values.

Also dropped is the col 'mths_since_last_record' as it has 90%+ missing values

Data Quality Issues – Unique Values

- Given 'id' and 'member_id' have all unique values no aggregation is required and each record is for a single account
- For the purpose of our analysis we do not need both so will drop "member_id"
- Additionally url column also does not have any useful information so will be dropped
- We also drop the 11 variables with only 1 unique value since they don't add any value to the analysis

Variable	Unique Values
collections_12_mths_ex_med	1
initial_list_status	1
out_prncp	1
out_prncp_inv	1
pymnt_plan	1
policy_code	1
tax_liens	1
acc_now_delinq	1
chargeoff_within_12_mths	1
delinq_amnt	1
application_type	1
term	2
loan_status	2
default_rate	2
pub_rec_bankruptcies	3
verification_status	3
home_ownership	5
pub_rec	5
grade	7
inq_last_6mths	9
emp_length	11
delinq_2yrs	11
purpose	14
sub_grade	35
open_acc	40
addr_state	50
issue_d	55
total_acc	82
mths_since_last_delinq	95
last_pymnt_d	101
last_credit_pull_d	106
int_rate	370
earliest_cr_line	524
zip_code	822
loan_amnt	870
funded_amnt	1019
revol_util	1088
total_rec_late_fee	1320
collection_recovery_fee	2616
dti	2853
recoveries	4040
annual_inc	5215
total_rec_prncp	6841
funded_amnt_inv	8050
installment	15022
title	19297
revol_bal	21275
desc	25803
emp_title	28027
total_rec_int	34025
last_pymnt_amnt	34418
total_pymnt_inv	36387
total_pymnt	36714
member_id	38577
url	38577
id	38577

Data Quality Issues – Correctly import data

To ensure data is correctly imported we look at the below list of variables of object type to see which ones need to be converted to a suitable format for further analysis

#	Column	Non-Null	Count	Dtype
4	term	38577	non-null	object
5	int_rate	38577	non-null	object
7	grade	38577	non-null	object
8	sub_grade	38577	non-null	object
9	emp_title	36191	non-null	object
10	emp_length	37544	non-null	object
11	home_ownership	38577	non-null	object
13	verification_status	38577	non-null	object
14	issue_d	38577	non-null	object
15	loan_status	38577	non-null	object
16	desc	26050	non-null	object
17	purpose	38577	non-null	object
18	title	38566	non-null	object
19	zip_code	38577	non-null	object
20	addr_state	38577	non-null	object
23	earliest_cr_line	38577	non-null	object
29	revol_util	38527	non-null	object
38	last_pymnt_d	38506	non-null	object
40	last_credit_pull_d	38575	non-null	object

Based on this analysis few fields like interest rate, revolver utilization, date columns etc. are corrected

Data Quality Issues – Drop non relevant variables

- Free form text fields like Employer name, title are dropped due to lack of usable data.
- Variables "recoveries", and, "collection_recovery_fee" are only applicable to defaults and based on business logic these variables are post default indicators and hence are not relevant to the problem statement so removed.
- zip code is categorical value with 822 levels and is difficult to analyze with the available data so dropped.
- Cols like 'total_pymnt', 'total_pymnt_inv','total_rec_prncp', 'total_rec_int', 'total_rec_late_fee’, 'last_pymnt_d', 'last_pymnt_amnt', "last_credit_pull_d“are only available once the loan is closed out either due to default or repayment amd as a result they do not have much predictive power so are dropped
- Not much difference can be observed between funded amount and funded amount by investor as a result dropping "funded_amnt_inv"

Derived Metrics

We create derived metrics from existing variables for better information as follows

- **Length of Credit History**

Based on business logic length of Credit History is a good indicator of credit quality. Longer the duration lower is the expected default. This variable is created based on the difference of earliest credit line and issue date

- **Transform/standardize level variables to relative values for better usability**

Level variables like loan_amnt, funded_amnt, installment, and revol_bal are standardized by dividing them by relevant variables

funded_amnt is divided by loan_amnt while rest are divided by annual_income

- **Removal of Categorical levels with insufficient observations**

Categorical variables like pub_rec_bankruptcies, and pub_rec (derogatory comments) have very few observations in levels other than 0 and 1 which don't add much value given low observations. As a result any observation with value > 1 is imputed to 1 thus reducing the levels to just 0 and 1.

Before imputation:

```
df2.pub_rec_bankruptcies.value_counts()

0.0    36163
1.0     1629
2.0        5
Name: pub_rec_bankruptcies, dtype: int64
```

```
df2.pub_rec.value_counts()

0    36431
1     2005
2       47
3        7
4        2
Name: pub_rec, dtype: int64
```

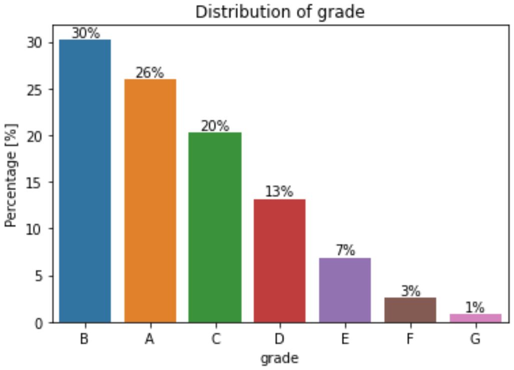
After imputation:

```
0.0    36163
1.0     1634
Name: pub_rec_bankruptcies, dtype: int64
```

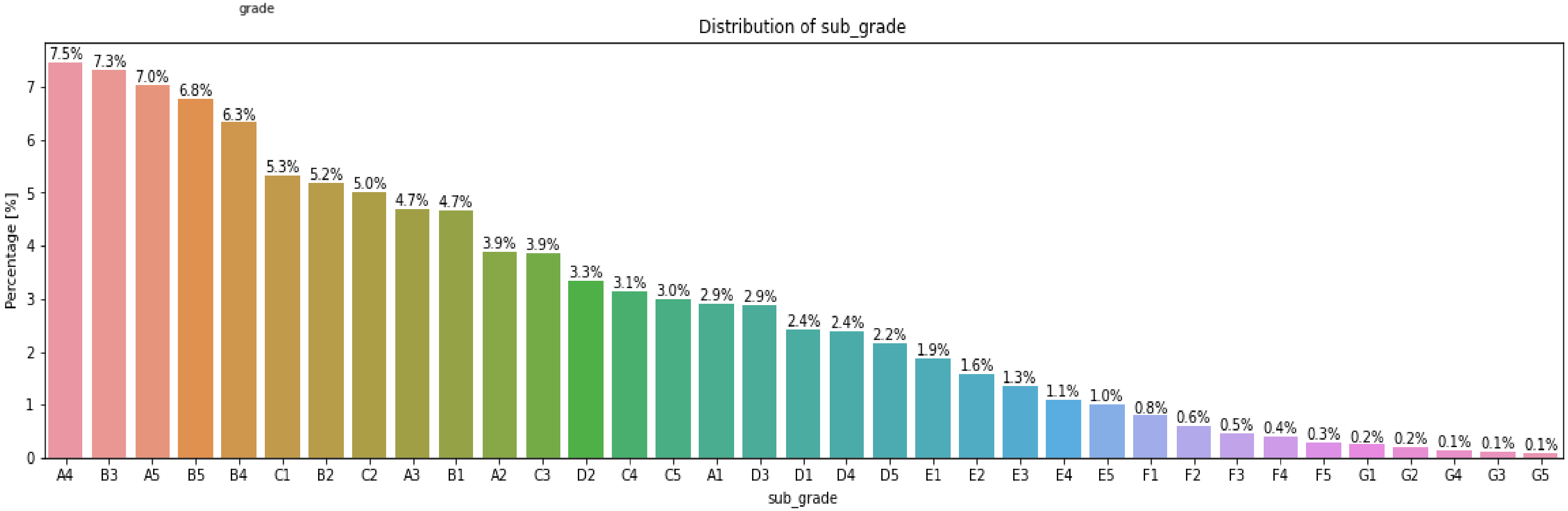
```
0    36431
1     2061
Name: pub_rec, dtype: int64
```

Univariate Analysis – Categorical Variables

As part of Univariate analysis for categorical variable we look at the population distribution across different categorical levels below



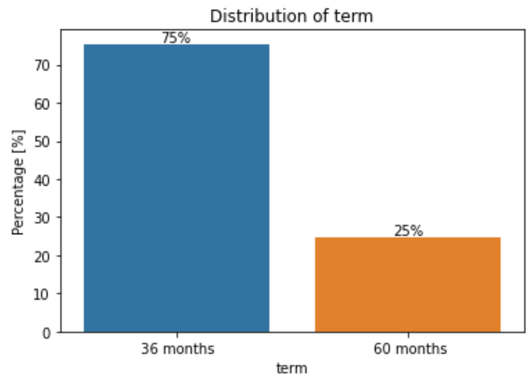
- It can be seen that Grade A & B together contribute more than 50% of the portfolio. As we will see in later Section Grade A denotes best credit Quality and Grade G worst credit quality.
- Subgrades map directly to grades for example A Grade corresponds to A1 to A5 sub grades.

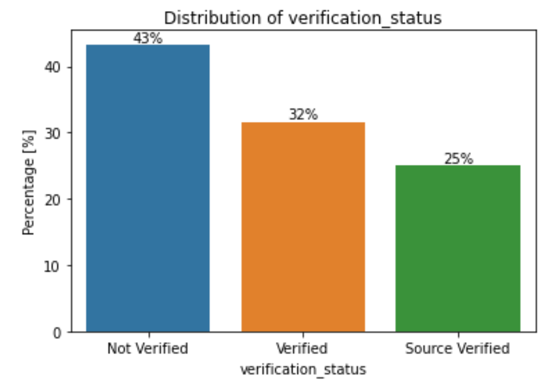


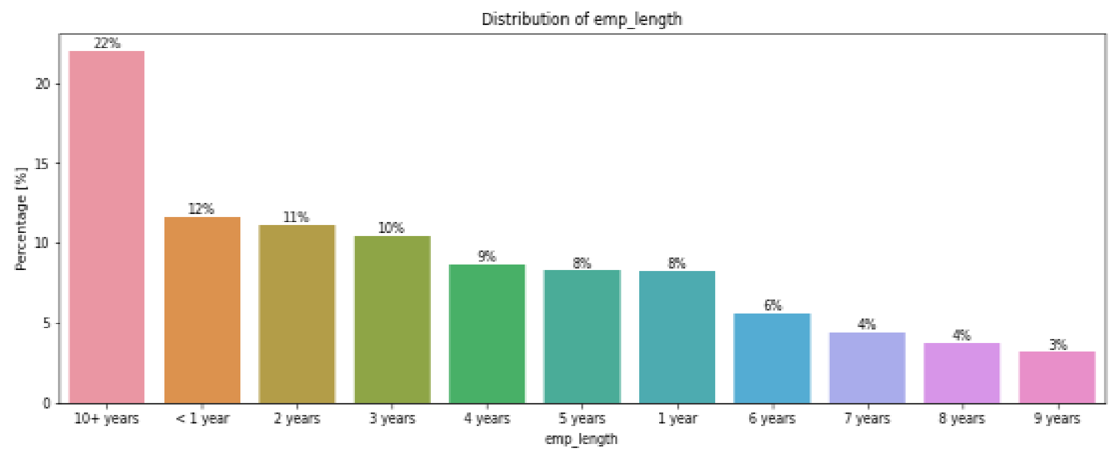
Univariate Analysis – Categorical Variables

As part of Univariate analysis for categorical variable we look at the population distribution across different categorical levels below

- It can be seen that 75% of the loans are of 3 year duration while rest 25% are of 5 years.
- From the verification status variable distribution it can be seen that around 43% of the loans are unverified.
- Similarly other categorical variables are also further assessed.

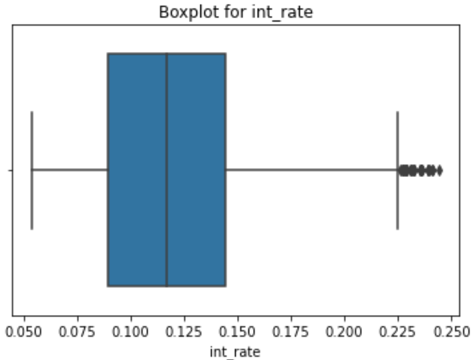


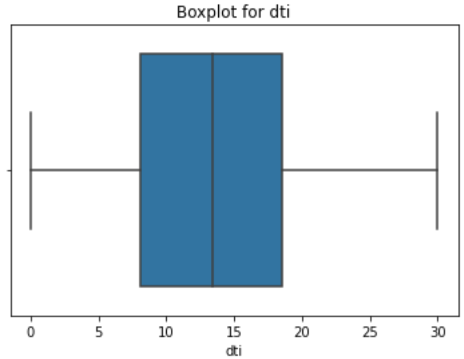


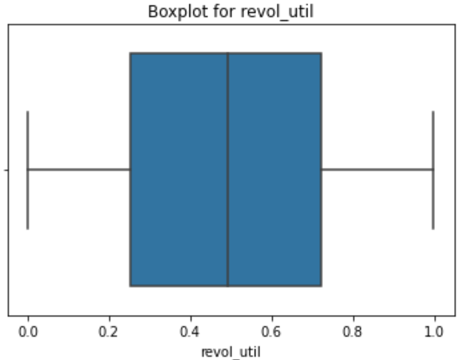


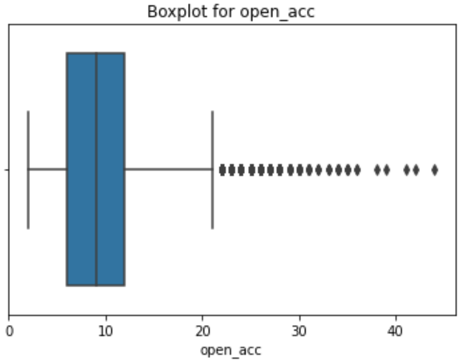
Univariate Analysis – Numeric Variables

As part of Univariate analysis for numeric variable we look at the box plot to visualize the summary statistics like mean, median, quartiles, outliers etc. as shown below









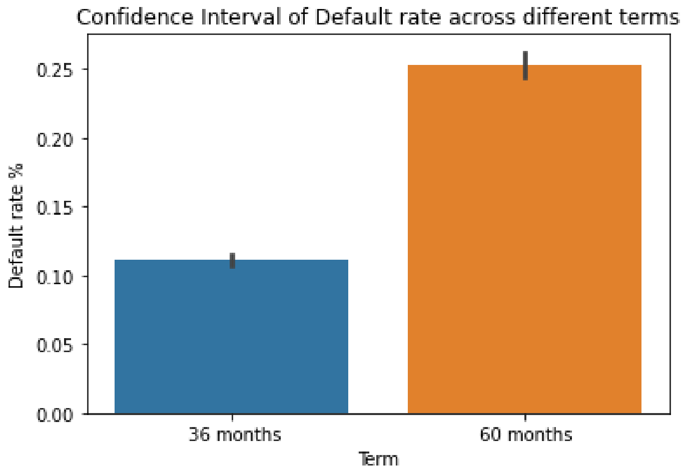
- DTI and revolver utilization can be seen to be well bounded with no outliers as per the box plots
- For interest rates it can be seen that half the loans are priced in the range of 8% - 15%, while there are some loans which have interest rates in upwards of 22%
- For number of open credit lines in Credit Report (open_acc) it can be seen that most of the loans have less than 20 open credit lines

Segmented Univariate Analysis – Categorical Variables

For purpose of segmented univariate analysis we visualize mean default rates using bar plots. The 95% Confidence Interval is also displayed by a line on top of the bar plot.

Term

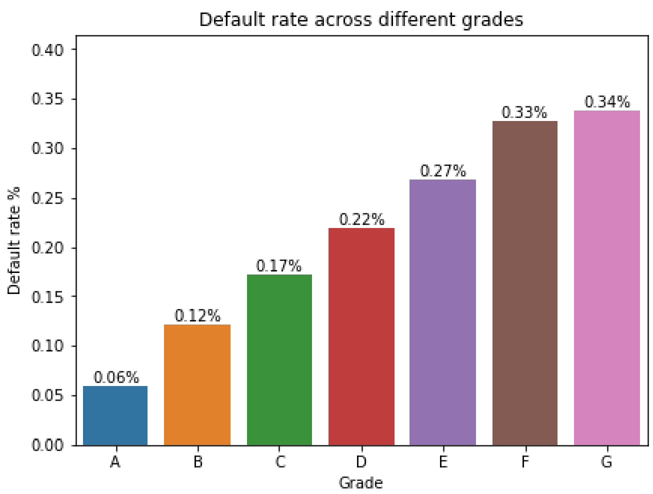
- Among all the variables Term shows the starkest difference between default rates.
- 5 year (60 months) loans have average default rate of 25% while 3 year (36 months) loans have default rate around 11%.
- The 5 year term loans have almost 2.5 times the default rate of the shorter term loans for 3 years.
- Also the 95% CI is widely separated showing clear difference in default rates across the 2 terms.

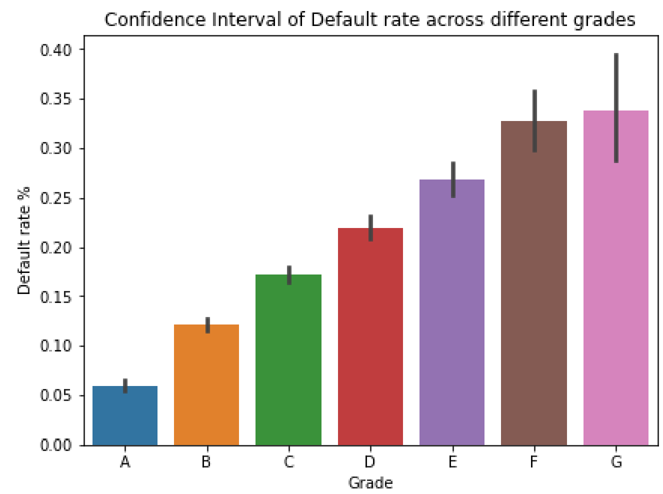


Segmented Univariate Analysis – Categorical Variables

Grade

- Clear discrimination in default rates can be seen for Grades A to F.
- Grades F and G seem to have overlapping default rates based on the spread of the 95% Confidence Interval implying that in terms of default rate the difference between F and G is not statistically significant at 95% confidence interval.
- For the remaining grades there is clear distinction between default rates across grades with A signifying the best credit quality



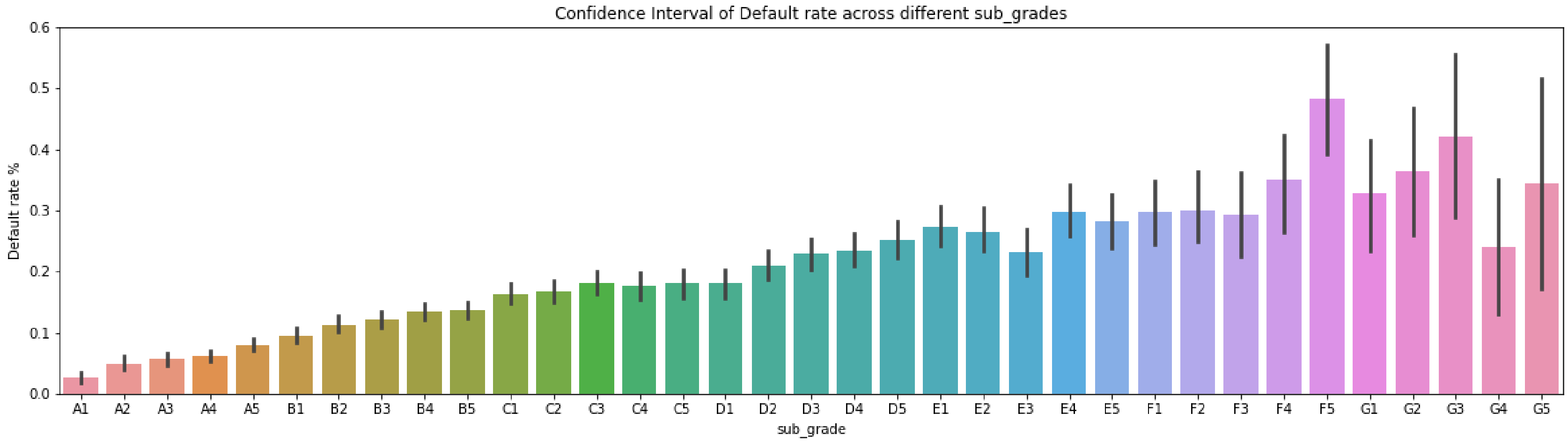


Segmented Univariate Analysis – Categorical Variables

Sub Grade

Sub Grade also shows similar characteristics.

However the granular nature of Sub Grade only helps in default rate discrimination in the initial buckets with later buckets seen to have overlapping default rate confidence interval.



Segmented Univariate Analysis – Categorical Variables

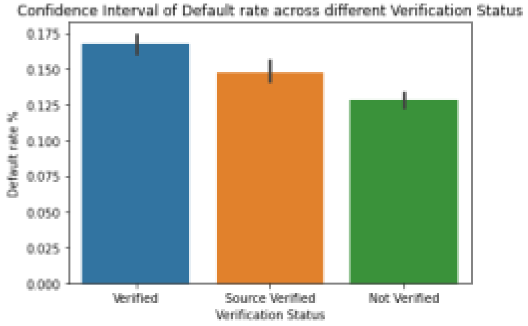
Verification Status

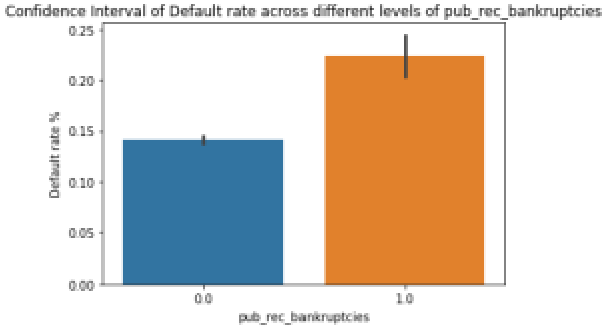
Number of derogatory public records (pub_rec)

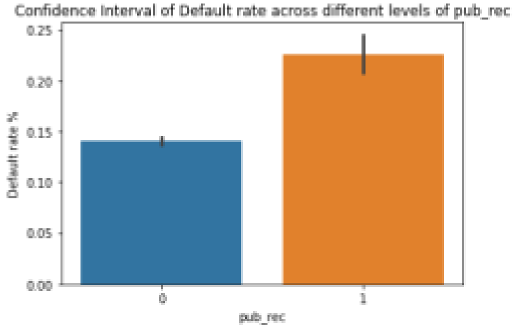
Number of public record bankruptcies (pub_rec_bankruptcies)

Clear discrimination in default rates is seen across different levels for all these 3 categorical variables.

Under data cleaning step explained earlier for both pub_rec and pub_rec_bankruptcies values greater than 1 were imputed with 1 given the low observations for such records







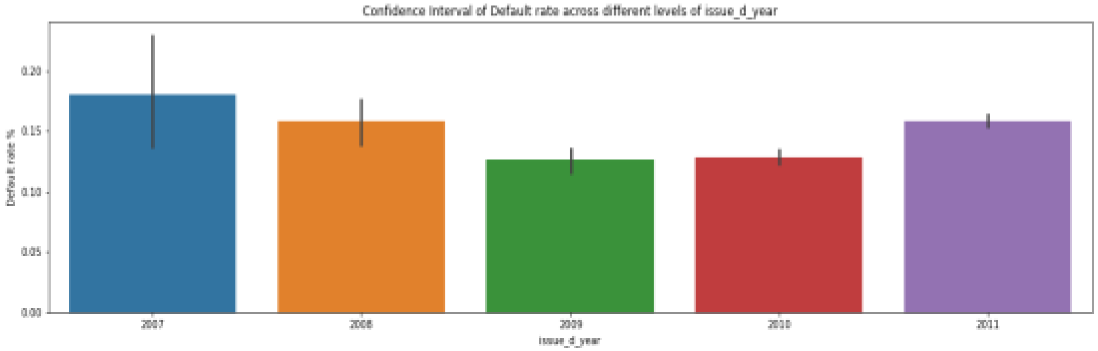
Verification Status

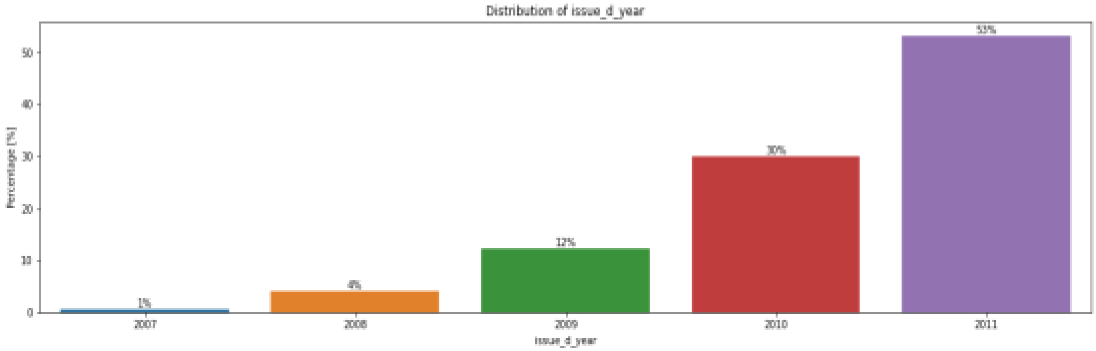
Oddly Not Verified loans have lower default rates that is counter intuitive. We will further assess this in bivariate analysis

Segmented Univariate Analysis – Categorical Variables

Issue Year

While not helpful in future loan approval decisions it is interesting to see the differences in default rate across different issue years. This seems to be both a function of underwriting standards as well as the number of loans issued as can be seen in the 2nd figure where it can be seen that the number of loans have kept on increasing year after year.

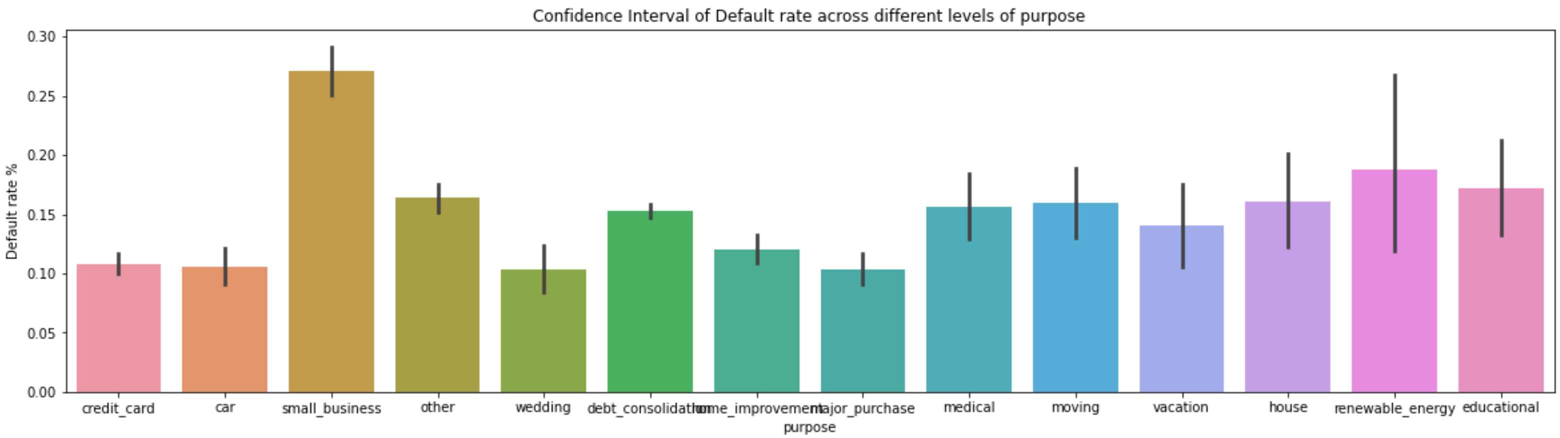




Segmented Univariate Analysis – Categorical Variables

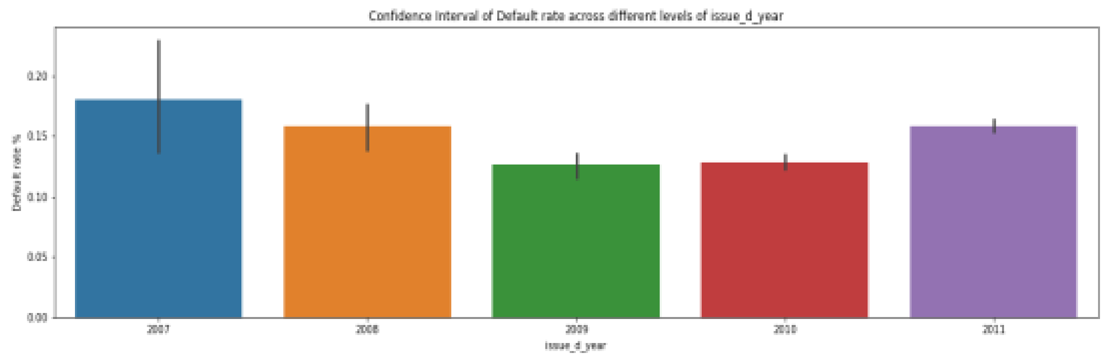
Purpose

Some interesting observations can be made regarding Purpose. Like for example Loans for Wedding generally have the lowest default rates, while small business loans have very high default rates



Issue Year

While not helpful in future loan approval decisions it is interesting to see the differences in default rate across different issue years. This is a direct result of underwriting standards with the post recession year 2009 having strict underwriting standards



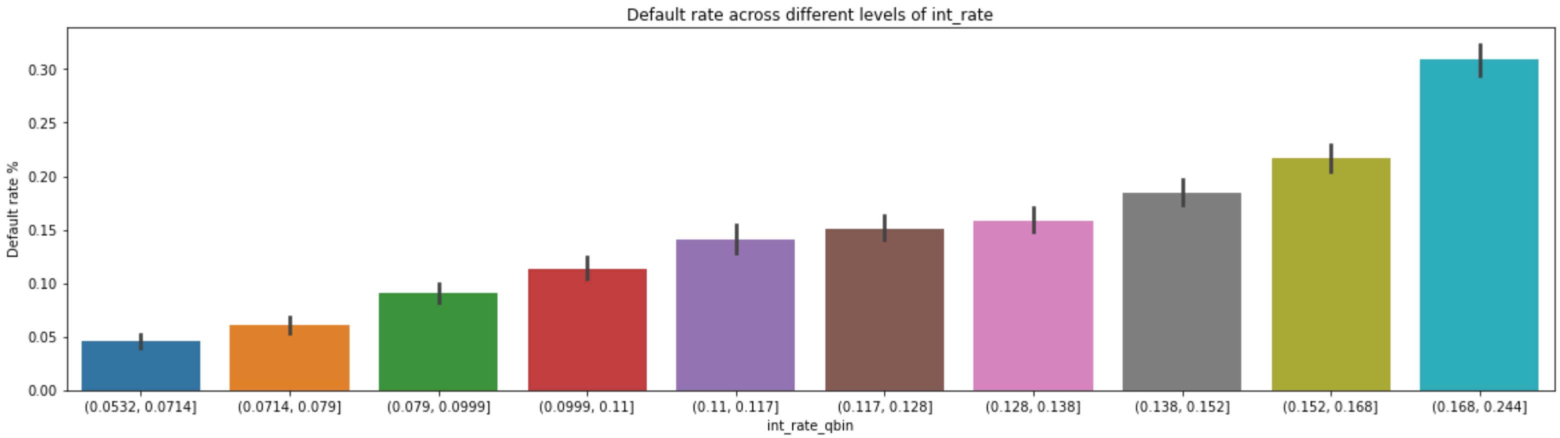
Segmented Univariate Analysis – Continuous Variables

- For purpose of segmented univariate analysis on Continuous variables we bin some of the numeric variables based on quantiles into 10 buckets for further analysis. This leads to each bin having roughly 10% of the population.
- Binning based on quantiles ensures each bin has sufficient records to make a meaningful and robust conclusion
- An alternative approach for binning would be linearly based on variable range. However, in this case sometimes some bins might not have sufficient records which can lead to spurious results

Segmented Univariate Analysis – Continuous Variables

Interest Rate

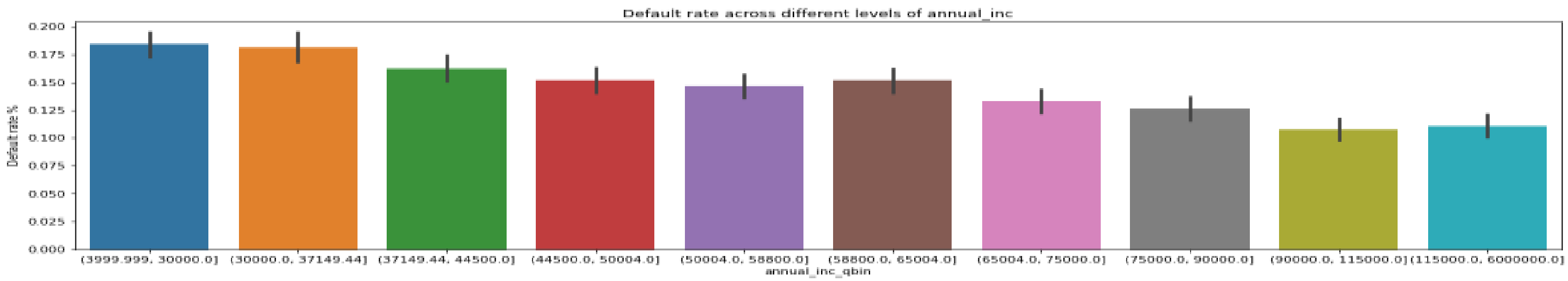
Among all continuous variables Interest Rate is seen to have the strongest relationship with default rate which is also expected as interest rate is how Credit Risk is priced. It just shows the companies underwriting department is doing a decent job. However, for the purpose of this analysis interest rate is not relevant as that is something set by the bank on the basis of the estimated credit risk of a borrower.



Segmented Univariate Analysis – Continuous Variables

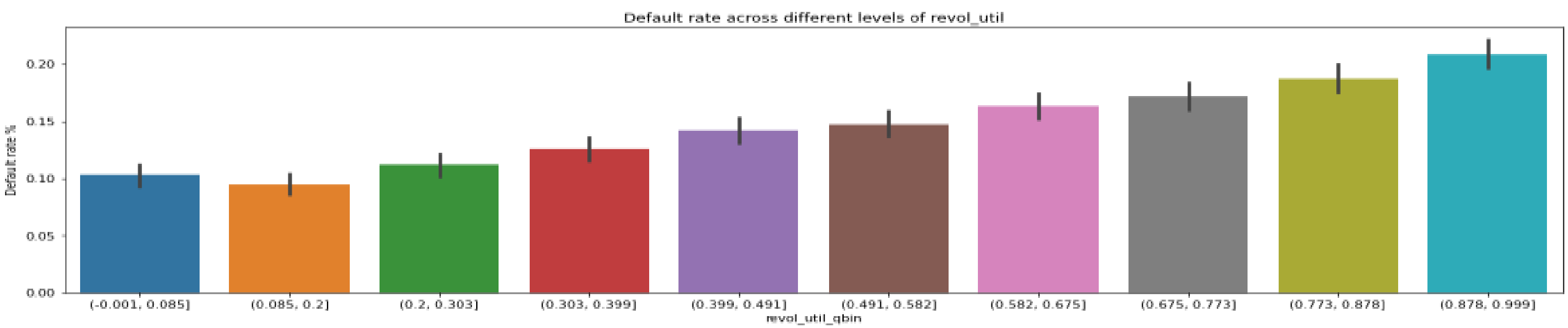
Annual Income

Default rates are generally seen to go down with increase in annual income



Revolver Utilization

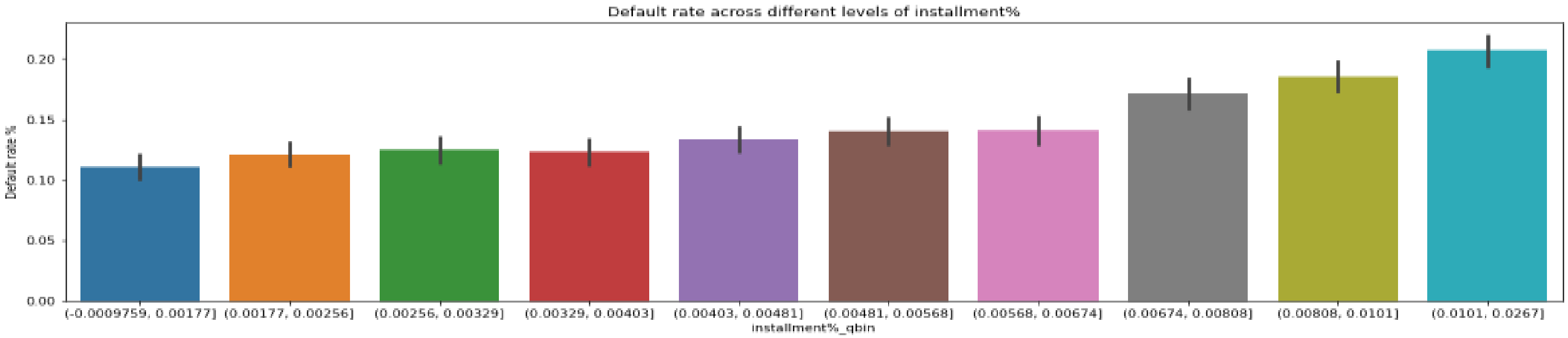
Revolver Utilization seems to be a good indicator of default rate with lower revolver utilization signifying lower credit risk and vice versa. Default rate for high utilization loans is almost double that of lower utilization loans.

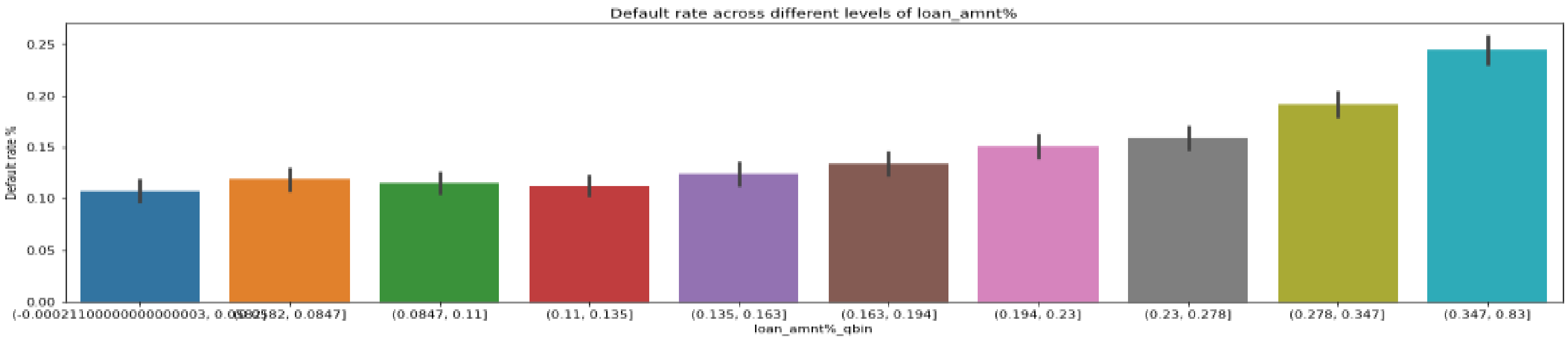


Segmented Univariate Analysis – Continuous Variables

Loan Amount and Installment

For our analysis we had standardized these 2 variables by dividing them by Annual Income for better results. From the plots below it can be seen that default rates are low when Loan Amount and Installment are a smaller fraction of Annual Income and vice versa.

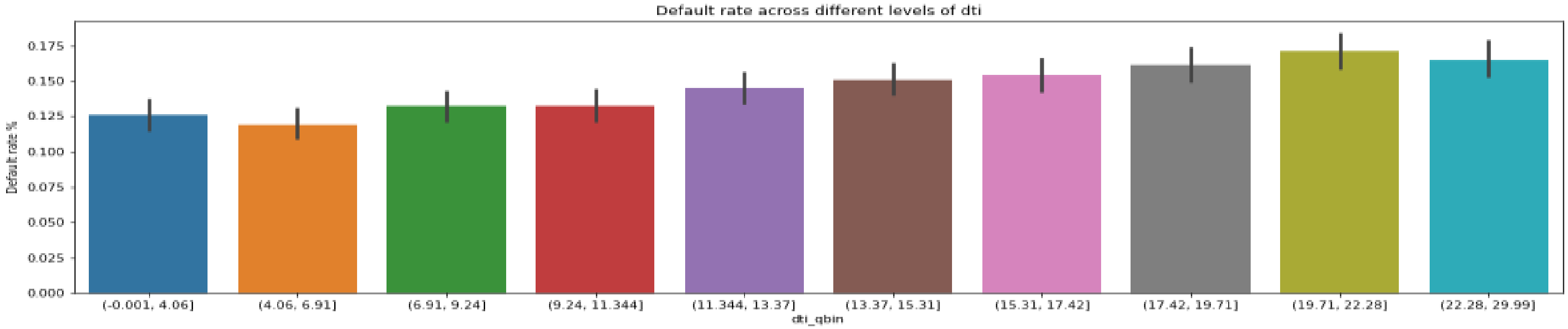




Segmented Univariate Analysis – Continuous Variables

Debt to Income Ratio

DTI ideally should have shown much higher discrimination then what is seen in the below bar plot. This might be due to the fact that the income used in the calculation is self declared and hence less reliable



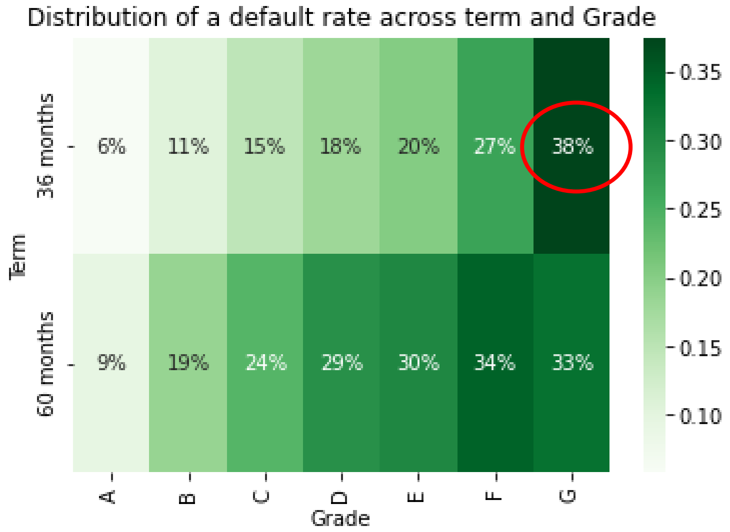
Bivariate Analysis

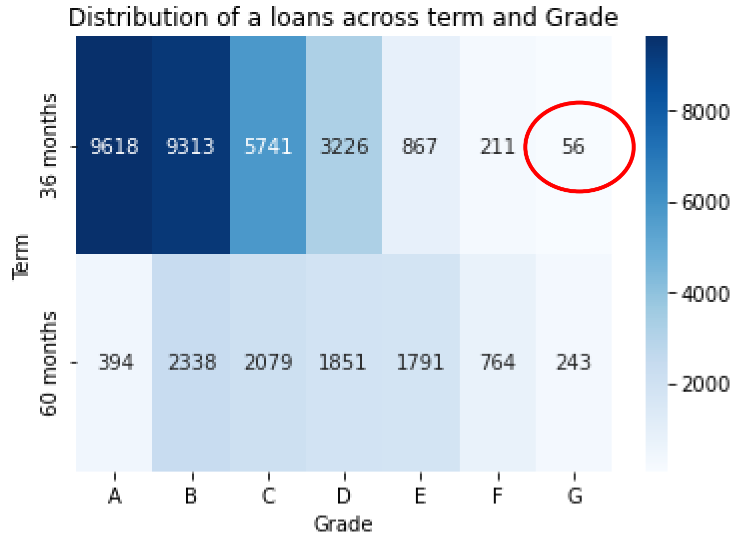
We do a bivariate analysis between 2 variables (Term and Grade) to see the distribution of default rates and loan counts across these 2 variables.

Left plot shows distribution of default rate while right plot shows distribution of loans

We need to look at the distribution of loans to ensure that the average default rates for the cross section is reliable.

For example there are only 56 loans for term=36 months and grade=G, hence the average default rate of 38% seen might not be very reliable as it is based on a very small sample size

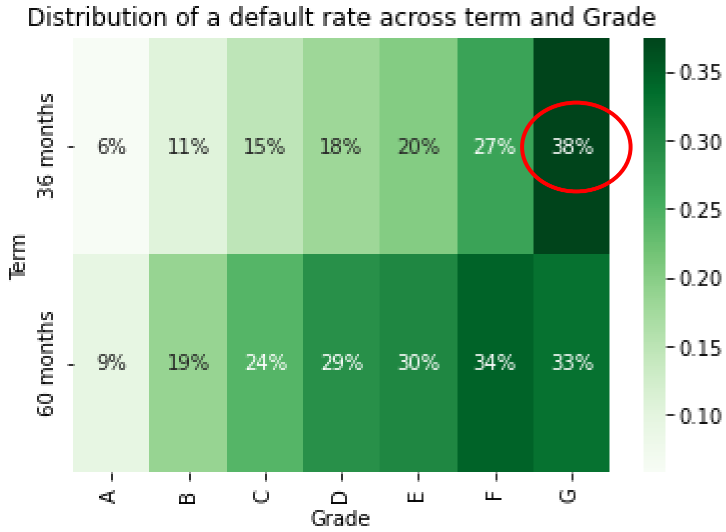


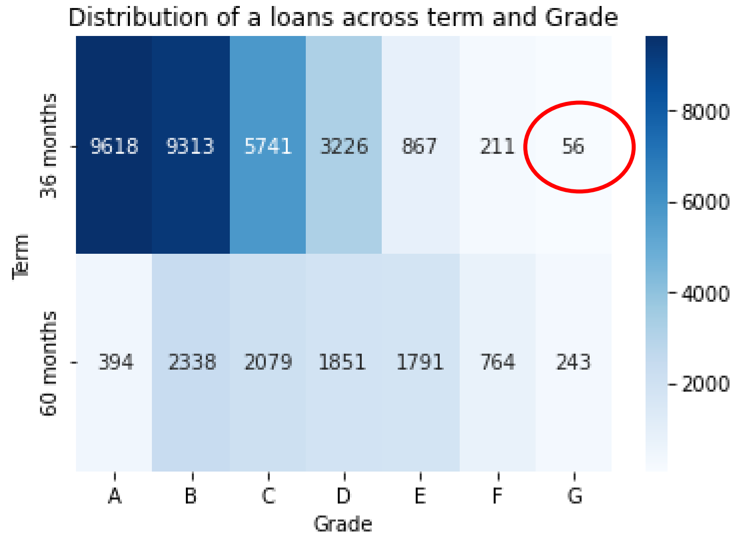


Bivariate Analysis – Term and Grade

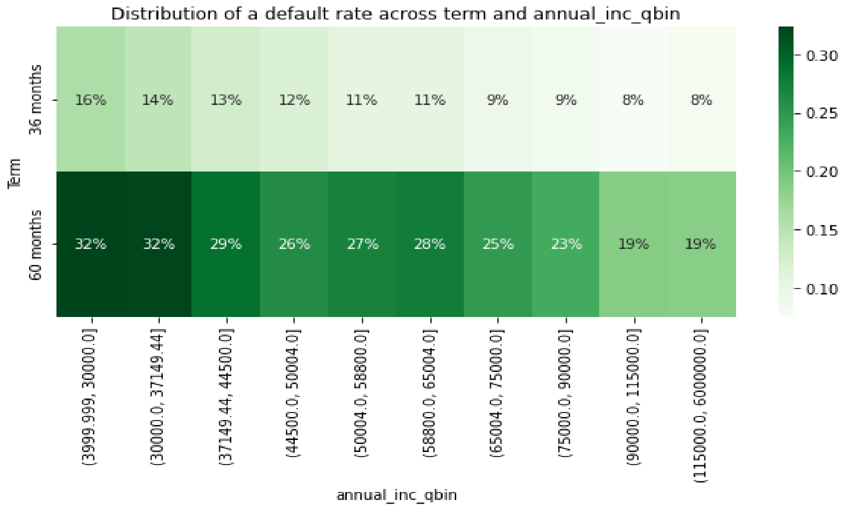
It can be seen that the trend from univariate analysis continues to hold with 3 year term seeing lower default rate then 5 year term across different grades. Similarly trend in default rate is seen across grades.

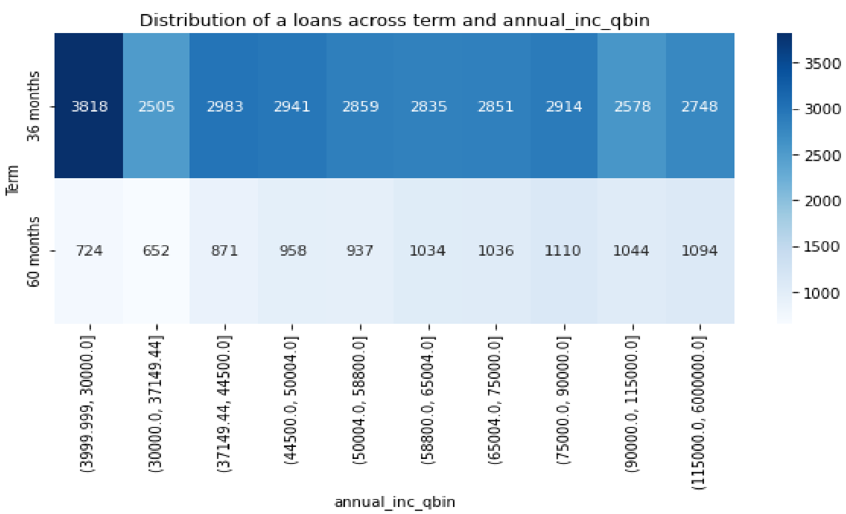
As discussed there might be some unintuitive result like 38% default for 3 year term vs 33% default for 5 year term for Grade G. This is due to the low number of loans matching this criteria.





Bivariate Analysis – Term and Annual Income

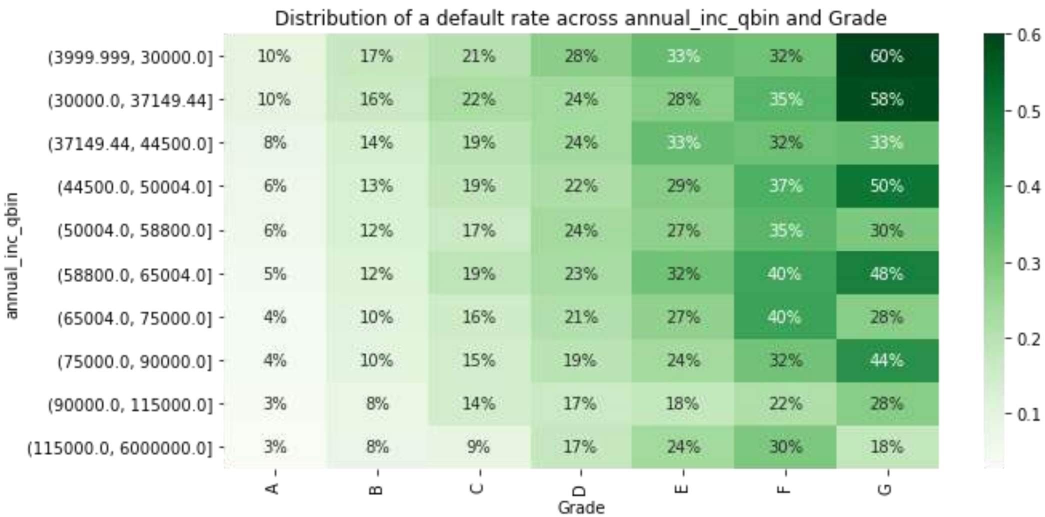


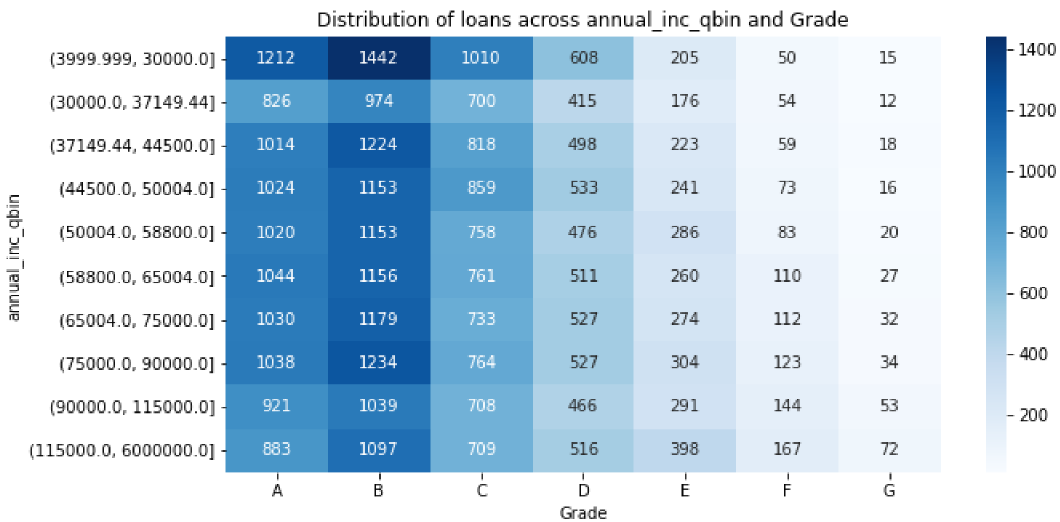


It can be seen that the trend from univariate analysis continues to hold consistently for both Term and Annual Income.

An interesting thing to note is that 5 year term loans are not as risky for high income borrowers

Bivariate Analysis – Annual Income and Grade





It can be seen that the trend from univariate analysis continues to hold consistently for both Grades and Annual Income.

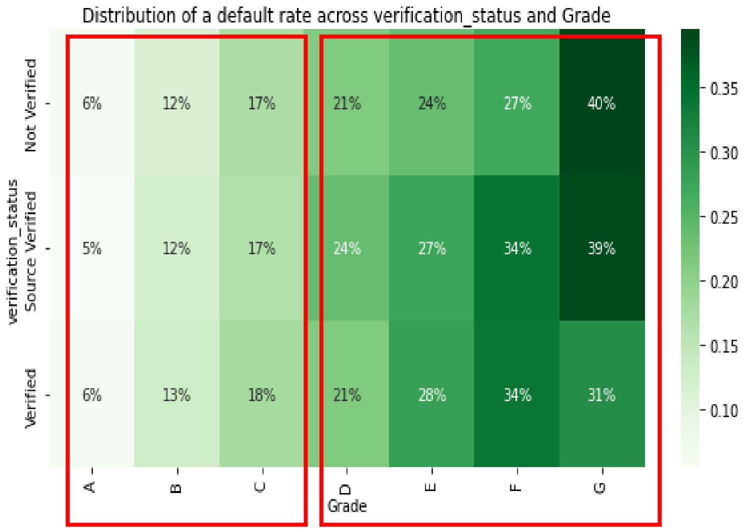
This is generally true even in instances where there are sufficient loans.

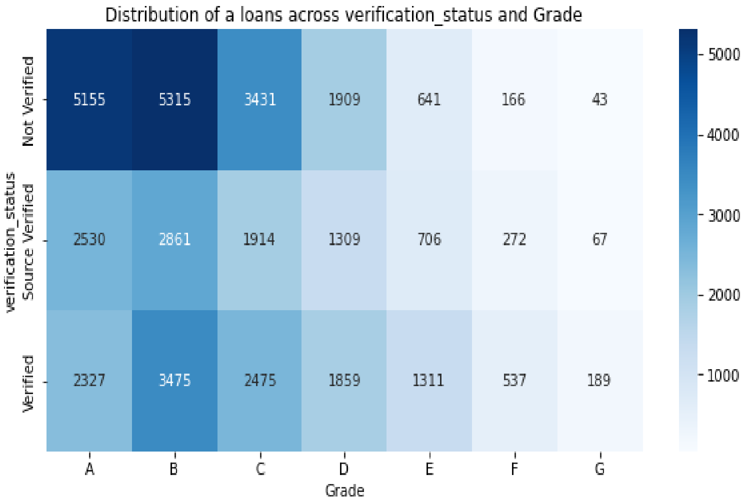
This indicates these 2 variables complement each other vey well

Bivariate Analysis – Verification Status and Grade

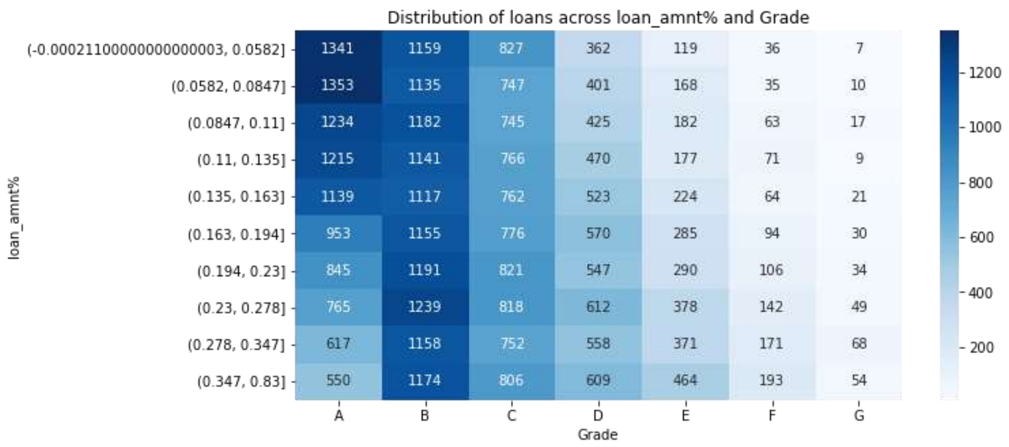
It can be seen that the trend from univariate analysis continues to hold for Grades

It is continued to be seen that verified status is not showing clear results even when combined with Grades





Bivariate Analysis – Loan amnt% and Grade



It can be seen that the trend from univariate analysis continues to hold for both Grades and Loan amnt% in instances where there are sufficient loans.

Key Drivers

Based on the analysis done so far the most important 5 drivers are

1. **Term**
2. **Grade**
3. **Annual Income**
4. **Purpose**
5. **Revolver Utilization**
6. **Pub_rec, pub_rec_bankruptcies** (default is higher for non zero records)
7. **Loan Amount and Installment as a proportion of Annual Income**

Business Recommendation

While lending the Bank should pay special attention to the following variables

- **Term –** Lower Terms have lower default rate then Higher Term loans
- **Grade –** Defaults are lowest for Grade A and increase for each Grade upto G. Sub Grades provide some further granularity for better grades but loose discriminatory power for worse grades
- **Annual Income –** Defaults are generally lower for people with higher income
- **Revolver Utilization –** this is a good indicator of default rate with lower revolver utilization signifying lower credit risk and vice versa. Default rate for high utilization loans is almost double that of lower utilization loans.
- **Purpose –** Loans for Wedding purpose are generally better than loans for Small Business which have a much higher default rate.
- **Loan Amount and Installment as a proportion of Annual Income -** default rates are low when Loan Amount and Installment are a smaller fraction of Annual Income and vice versa.
- An interesting thing to note is that 5 year term loans are not as risky for high income borrowers.
- **Verification Status and DTI** show unintuitive results, the company can try and explore the underlying reasons for this