Analysis of Lending Club Defaulter Data

By,

Samarpan Das

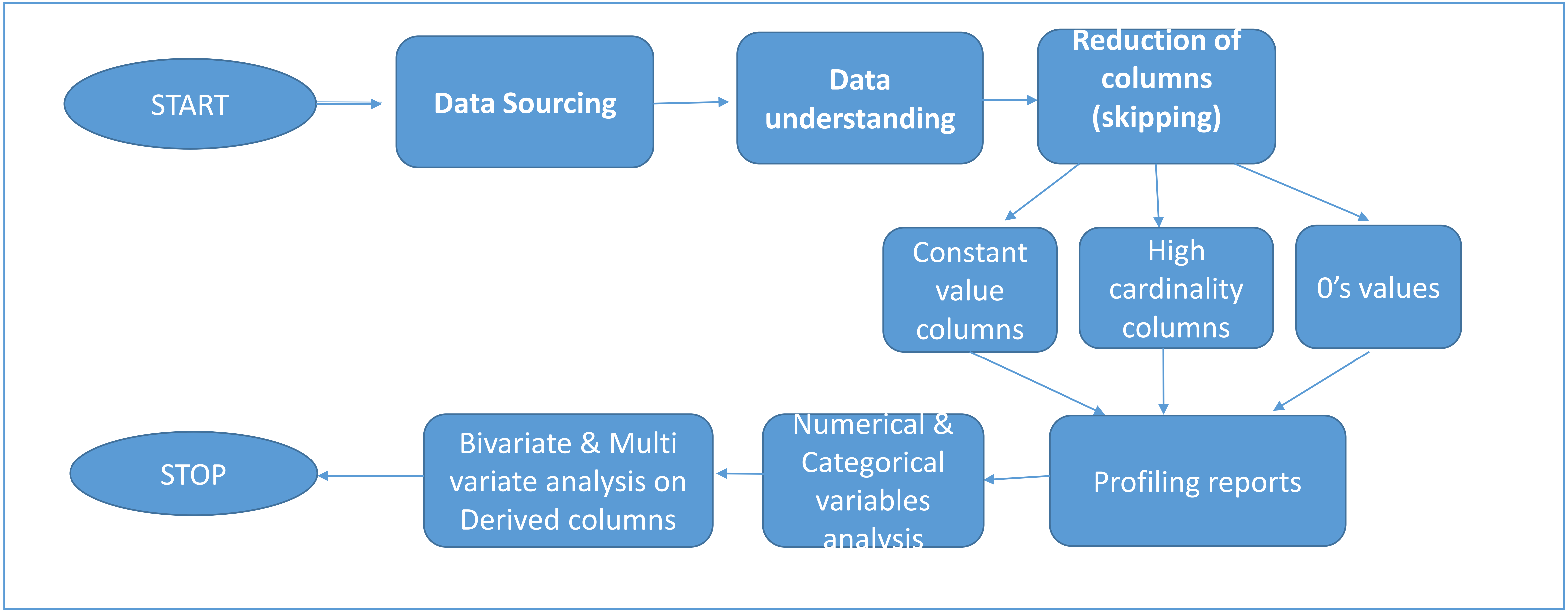
Venkatesh P

Objectives

Business objective: The objective is to identify the **defaulters** .

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Problem solving methodology


START
Data Sourcing
Data understanding
Reduction of columns (skipping)
Constant value columns
High cardinality columns
0's values
Profiling reports
Numerical & Categorical variables analysis
Bivariate & Multi variate analysis on Derived columns
STOP

Data Sourcing

- The data was provided by the use case.
- The same data was kept in the local file in the csv format.
- The data was read into a pandas data frame.
- The shape command was ran to make sure the count of rows in the dataframe matched with the count of records provided in the csv file.

Data Understanding

- The file which was provided had a very high no. of columns.
- The first action was taken to reduce the number of data columns based on the below categories:
    - Columns having constant values
    - Columns having high cardinality values
    - Columns having a high percentage of missing values or 0’s (which has no business meaning)
    - Customer behavioral columns which are not known at the time of issuing the loan
    - Columns with very high correlation with loan amount which is not required for analysis

Reduction of Columns

- Constant Value Columns:

  64 Columns which are dropped for having constant values

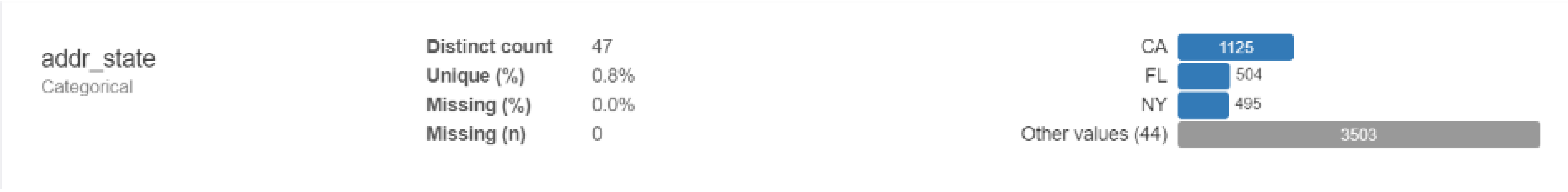
- High Cardinality Columns :

  Most of the high cardinality non date timestamp columns like desc, emp_title and zip code was removed.

- 0's Values
  - Most of the columns which are numeric and containing high percentage of 0's were removed ex: delinq_2yrs
- Behavioural Columns
  - All the columns which are not known at the time of analysis was dropped as that does not help with the problem statement

Profiling Report

- The best way to understand the data is to create a profiling report where the below details are explained:
  - Unique Values
  - Missing Values and Percentages
  - Distribution in case of numeric value
  - Frequent Value Analysis
  - Mean/Mode/Median analysis for the numeric values
- Here is an example of a categorical value result. This clearly shows that the state CA, FL and NY has most of the customers and their distributions with the total population.

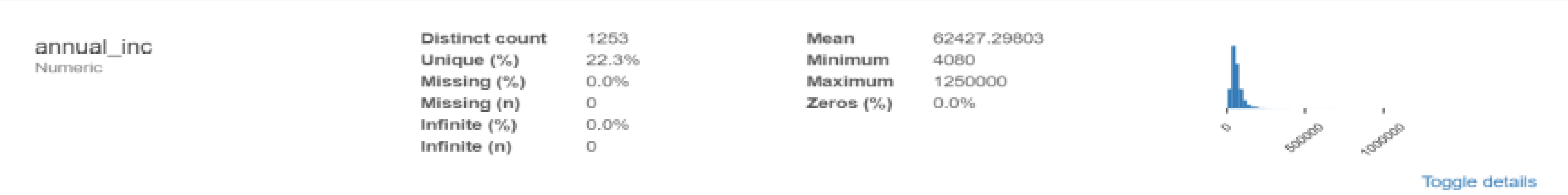


addr_state
Categorical

Distinct count	47
Unique (%)	0.8%
Missing (%)	0.0%
Missing (n)	0

Profiling Report Numerical Variable

For a numerical column, the profiling report shows below details:



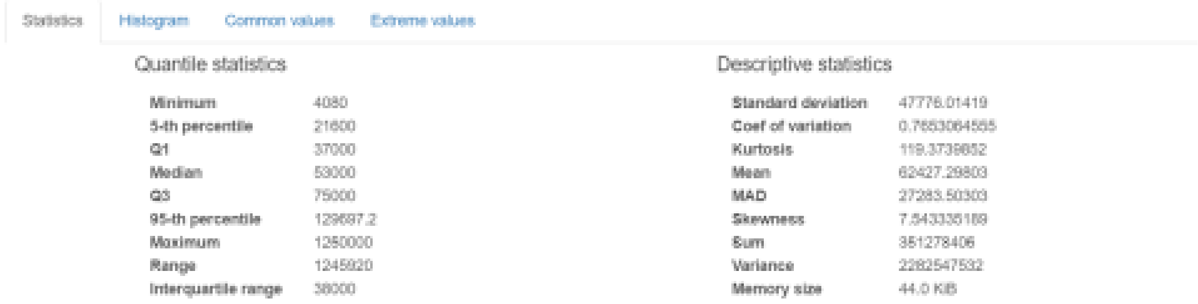
annual_inc
Numeric

Distinct count	1253
Unique (%)	22.3%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	62427.29803
Minimum	4080
Maximum	1250000
Zeros (%)	0.0%

Toggle details

This also shows the summary statistics, outlier analysis.

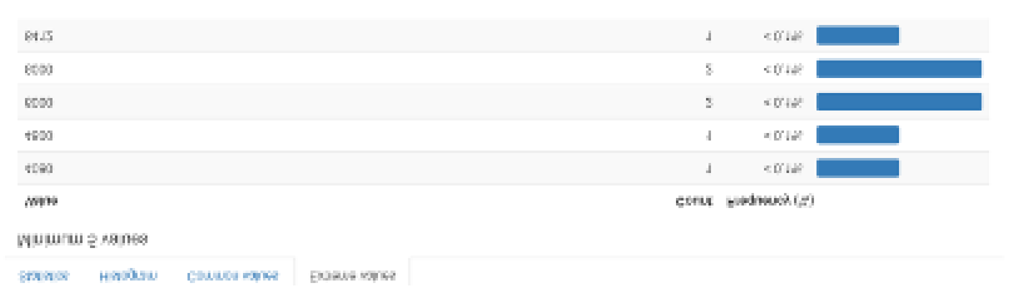


Quantile statistics

Minimum	4080
5-th percentile	21600
Q1	37000
Median	53000
Q3	75000
95-th percentile	129687.2
Maximum	1250000
Range	1245920
Interquartile range	38000

Descriptive statistics

Standard deviation	47776.01419
Coef of variation	0.7653064555
Kurtosis	119.3739852
Mean	62427.29803
MAD	27283.50303
Skewness	7.543335189
Sum	351278406
Variance	2282547532
Memory size	44.0 KiB



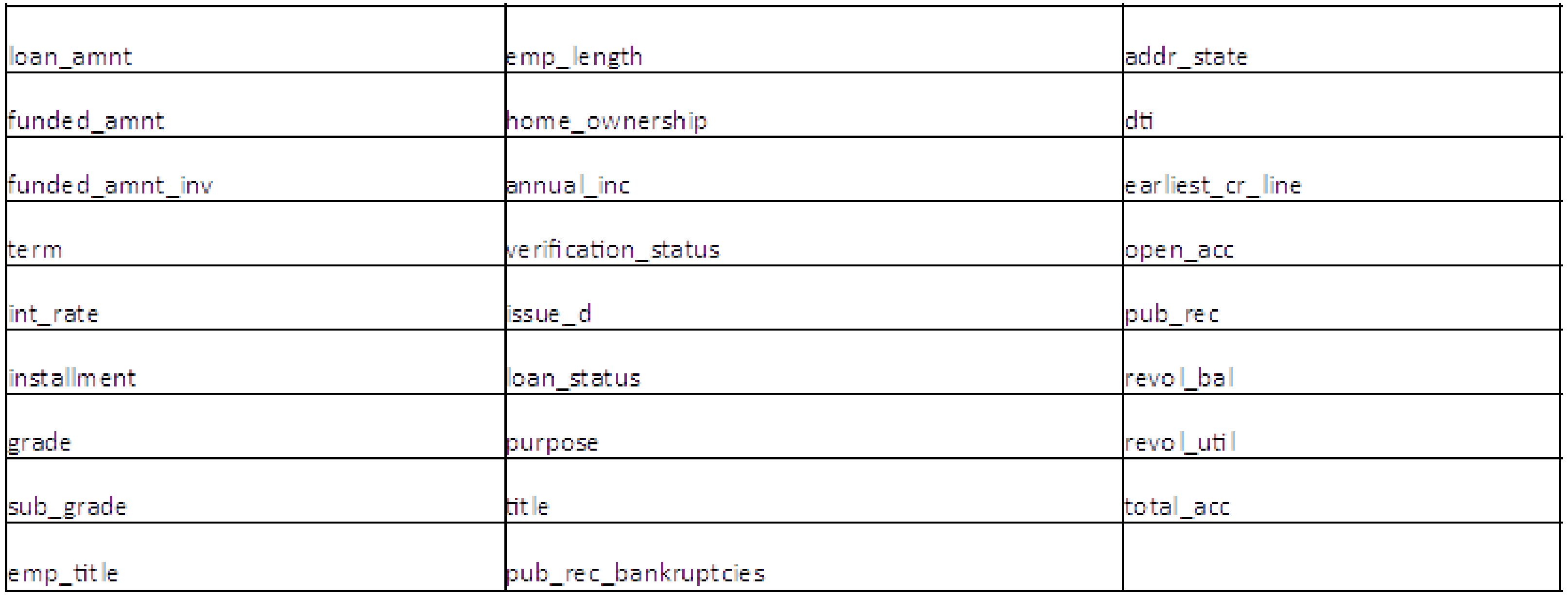
Minimum 5 values

Value	Count	Frequency (%)
4080	1	< 0.1%
4800	1	< 0.1%
5000	2	< 0.1%
6000	2	< 0.1%
[illegible]	1	< 0.1%

A complete profiling report for the entire dataset is attached along with the submission.

Final List of Variables Considered

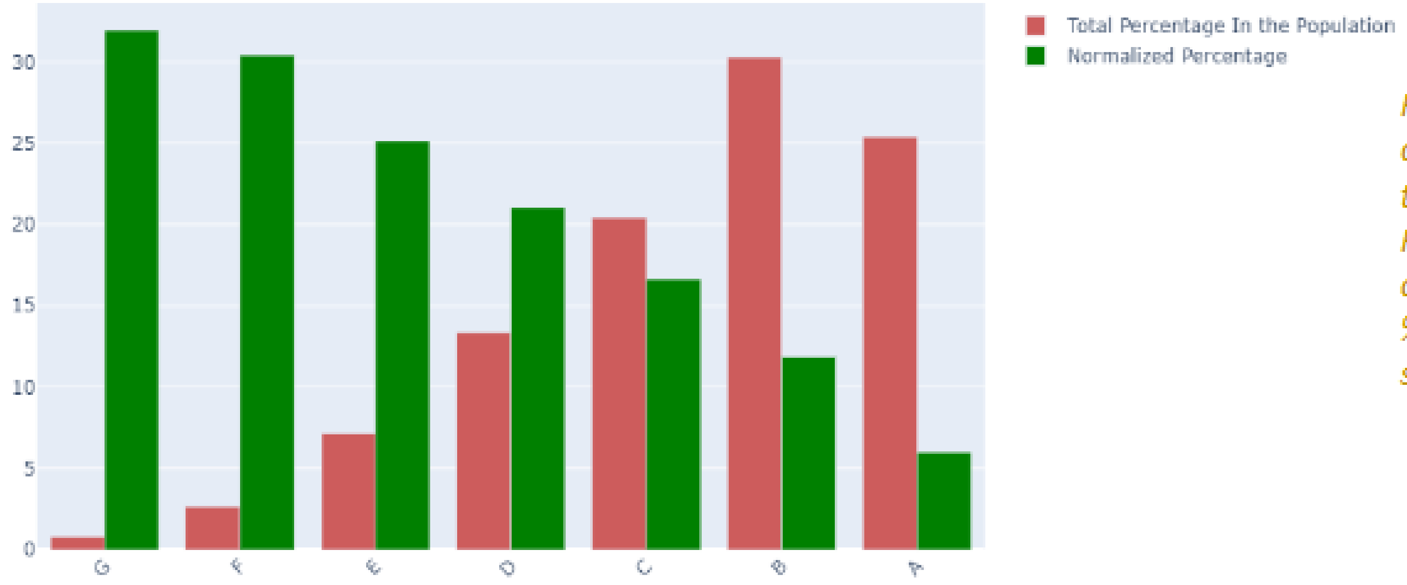
Below are the columns considered for the final analysis. Few of the highly correlated columns are retained for few business driven calculations to see if they have any impact.



loan_amnt	emp_length	addr_state
funded_amnt	home_ownership	dti
funded_amnt_inv	annual_inc	earliest_cr_line
term	verification_status	open_acc
int_rate	issue_d	pub_rec
installment	loan_status	revol_bal
grade	purpose	revol_util
sub_grade	title	total_acc
emp_title	pub_rec_bankruptcies	

Strategy for Categorical Variables Analysis

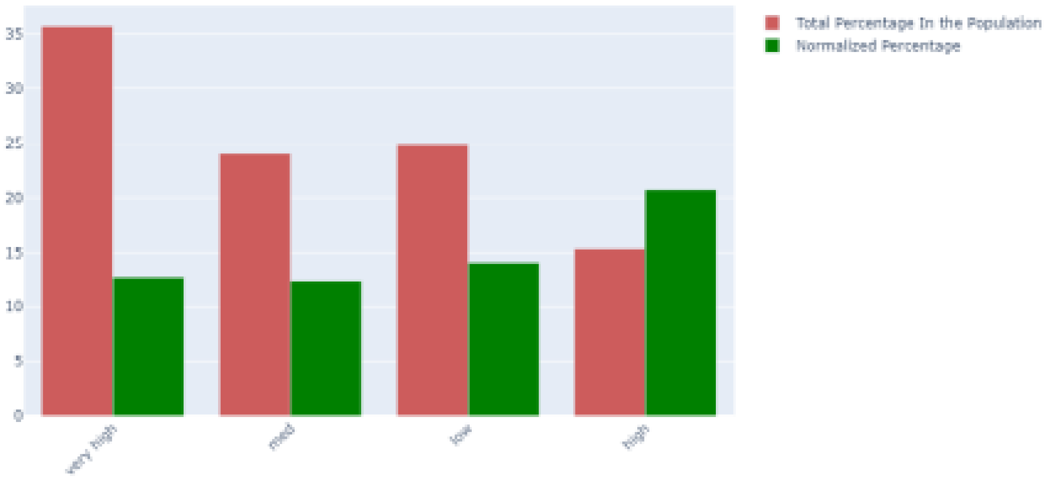
- Just by looking at the number of defaulters we should not be arriving at any conclusions. For example, if you look at the graph below, maximum number of defaulters have grade is B, but the overall percentage with the total population if you consider the number of customers with grade B is also highest. So we need to normalize the data by categories converting to percentage and then we should conclude.
- In the graph below, the company needs to be careful about giving loans to people with grade = G or F. If going by sheer number of defaulters (whose Grade = B) then LendingClub would lose business.

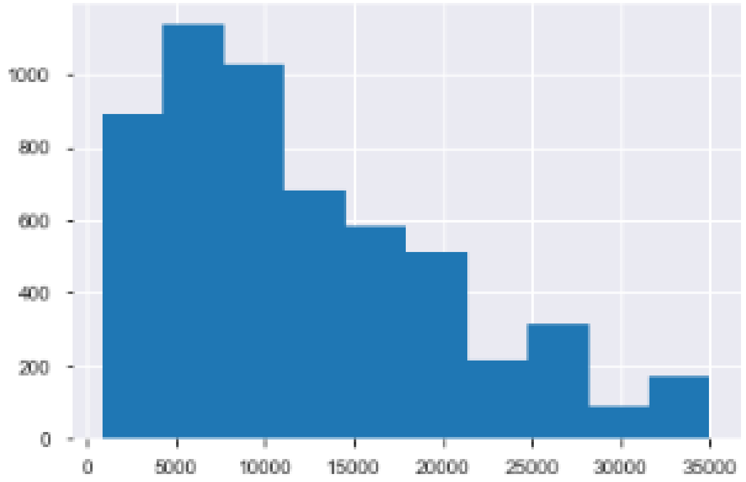


For any category with less than 50 defaulters are rejected from analysis as they are too small groups to consider. For ex. For the state NE there are 5 loans out of which 3 are charged off where the %ge defaulters is 60% but this is a very small category to take it seriously.

Strategy for Numerical Variables

- We have already figured out the required numerical variables from the profile report.
- The first step is to create a histogram for the defaulter population to find out the spread of the defaulters for that variable.
- The next step is to divide the numerical variable into low, medium, high, very high categories and plotting the graph after normalizing the data with the population percentage. By doing this we would take care of the outliers as bins are created and outliers (as they are not in significant number do not affect the conclusion much)
- Here are such analysis on Loan Amount column



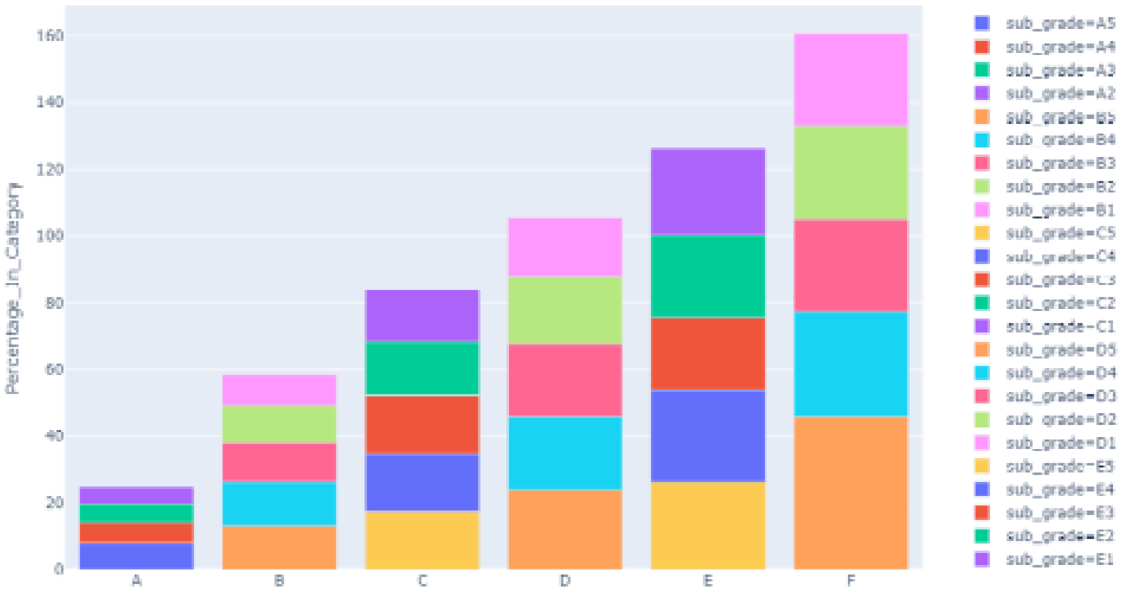


Analysis of Derived Column

- The example of the derived column analysis is performed by below cases
    - By breaking the Loan Issue date into month and year separate columns to see how the number of defaulters are spanned across time
    - By deriving a flag called Fully_Funded_Loans which is set for all those columns where the requested amount is not fully funded and see if this feature has an affect on the loan defaulters.

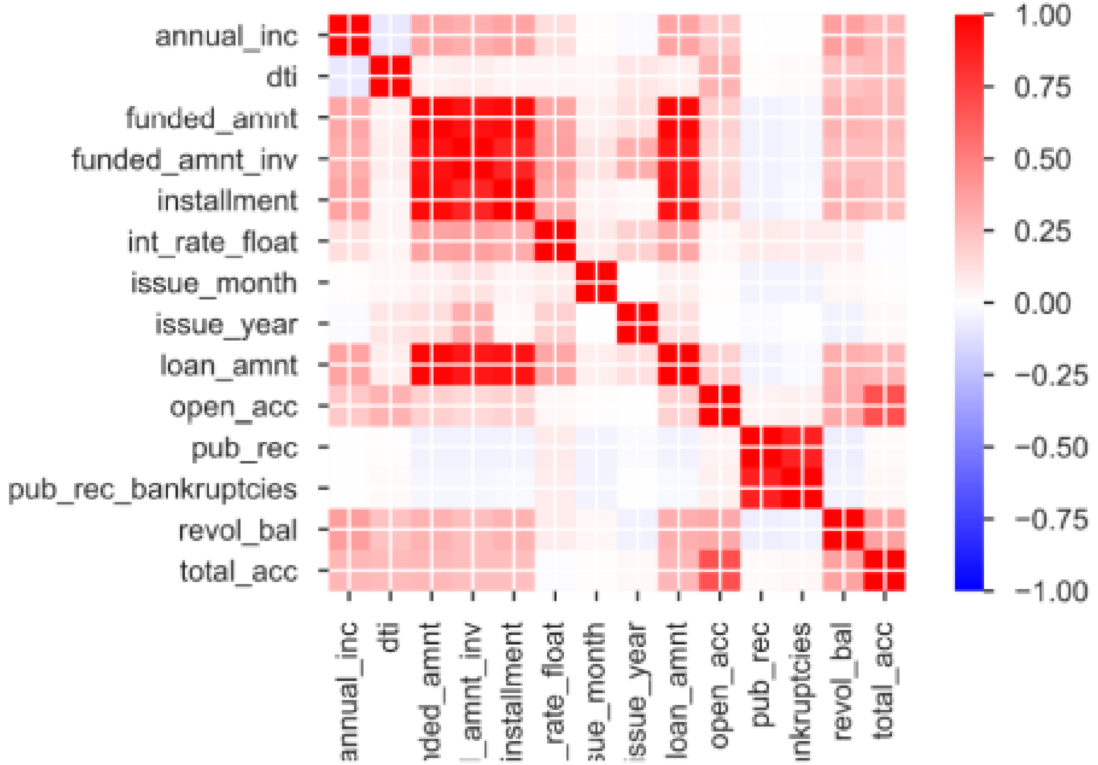
Bivariate Analysis

- Since the number of categorical variables are too high so picking variables for bivariate analysis is tricky as the number of combinations are high.
- So, to keep it simple we considered only the variables which are logically related like Grade and Sub-Grade. The output of this analysis looks like:



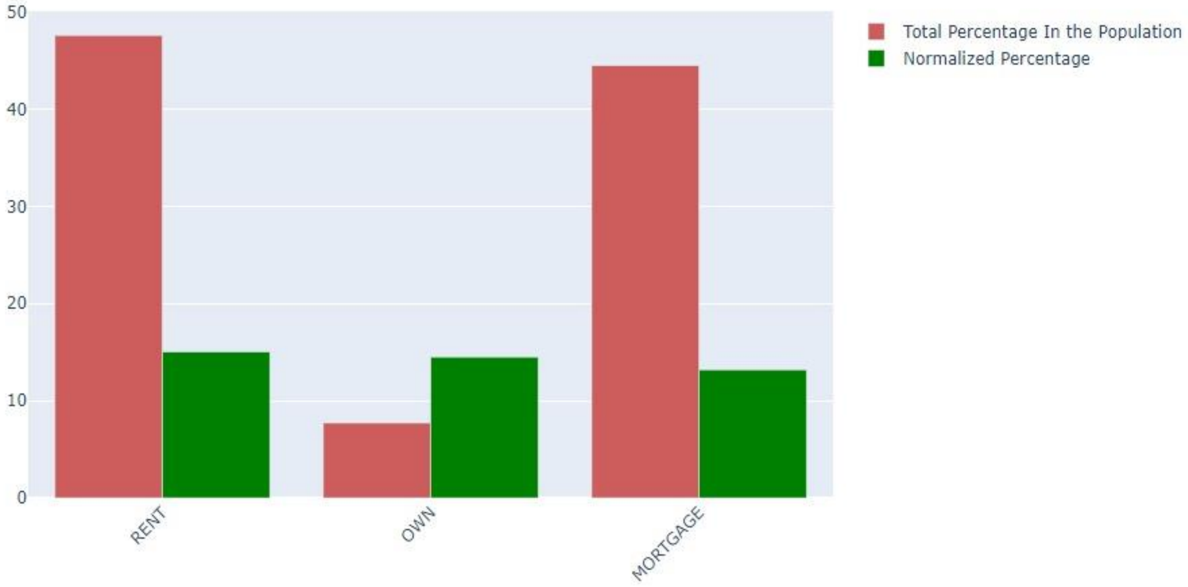
Numerical Multi Variate Analysis

- For multi variate correlation analysis we plotted the correlation graph between the numerical categories and here are the results

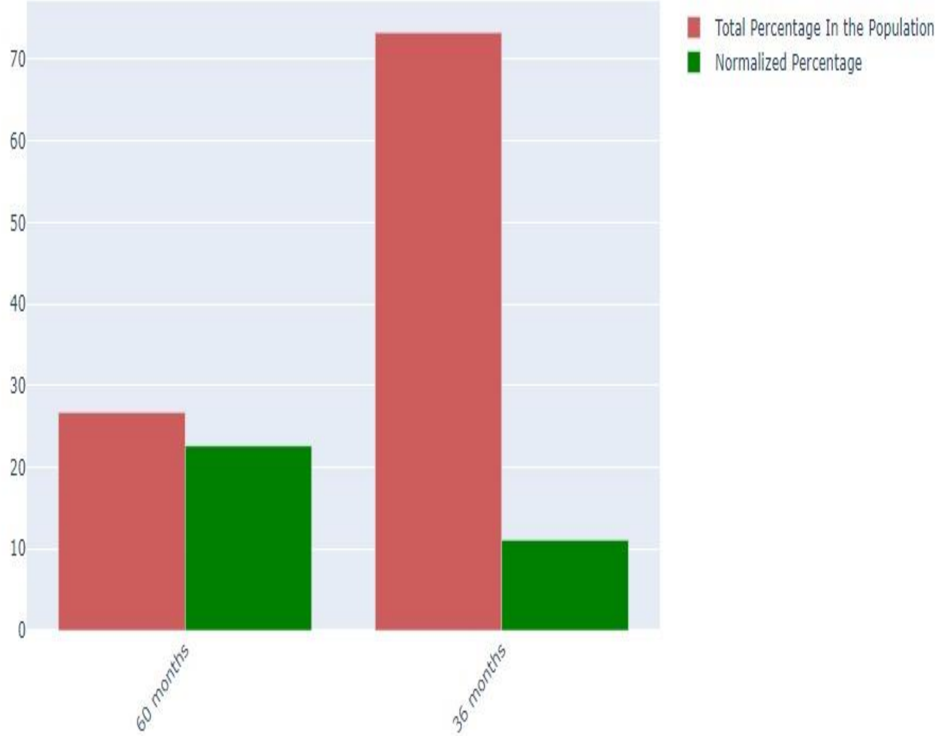


Results supporting graphs

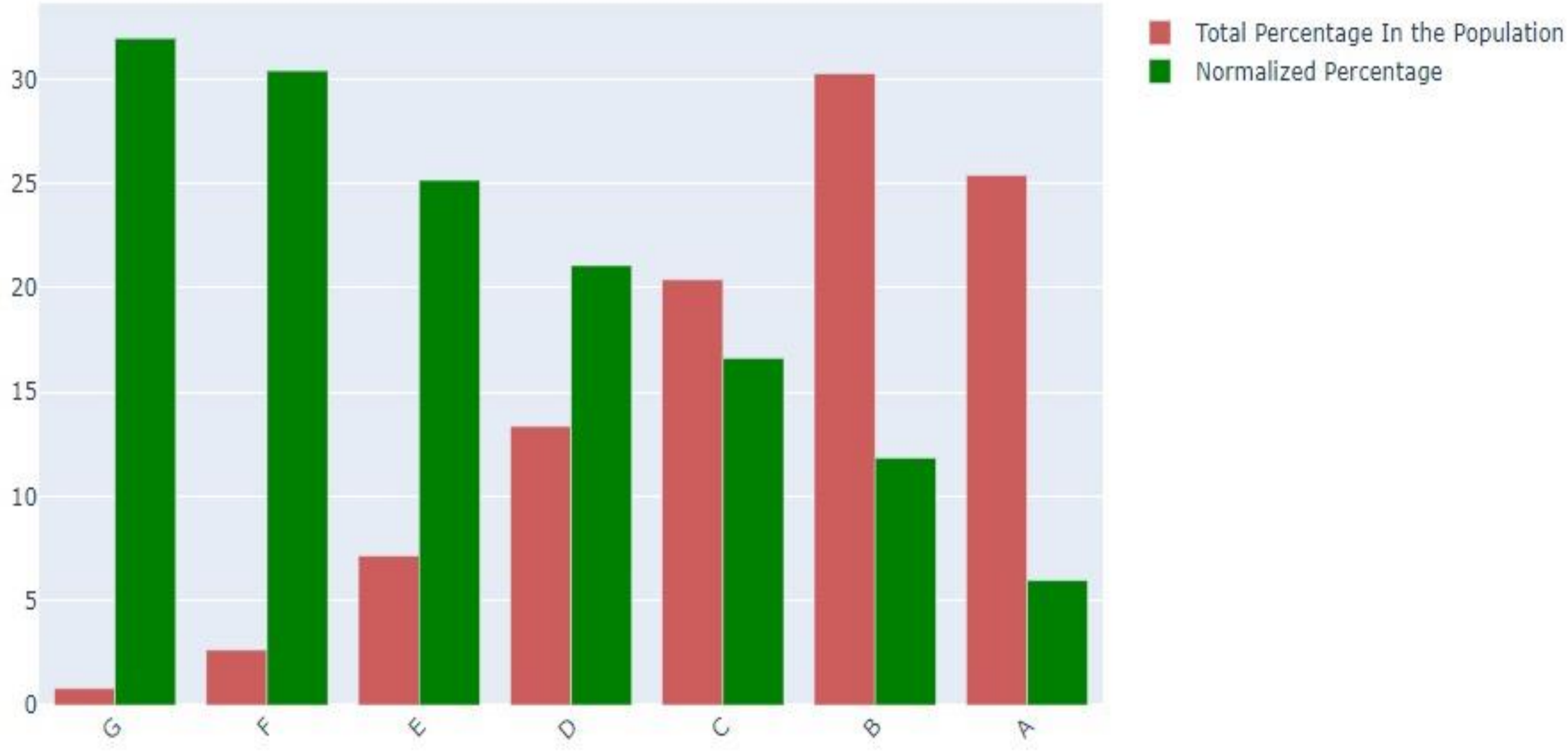
Graph on home ownership



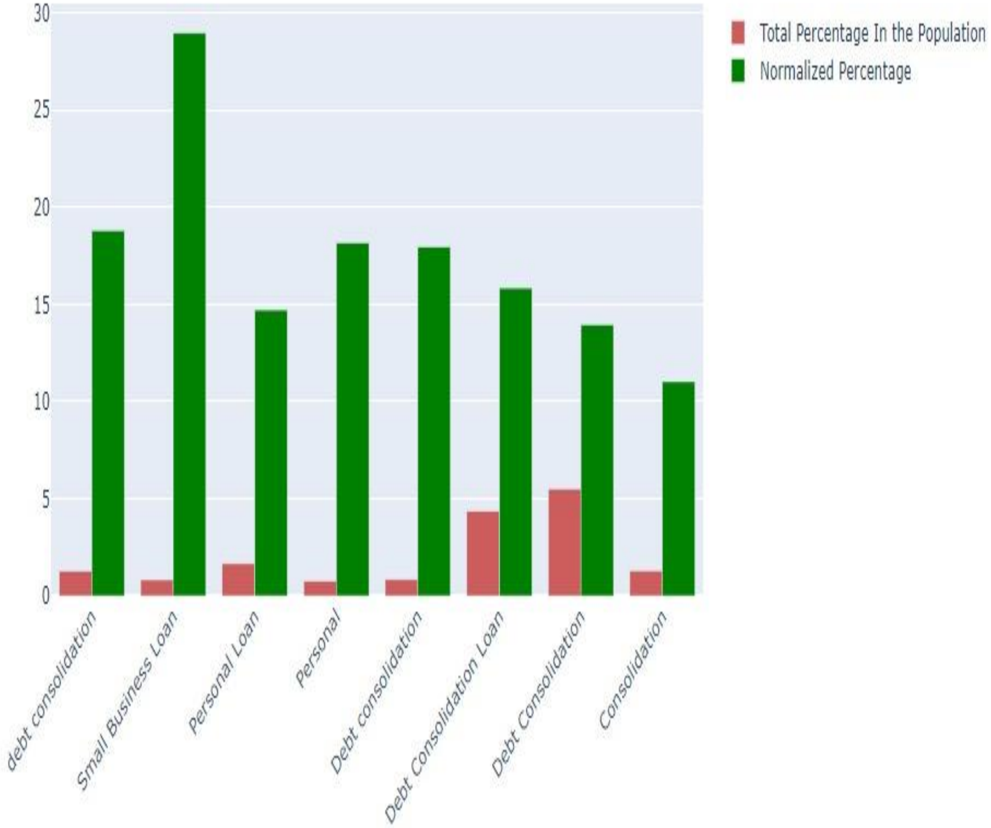
Graph on term

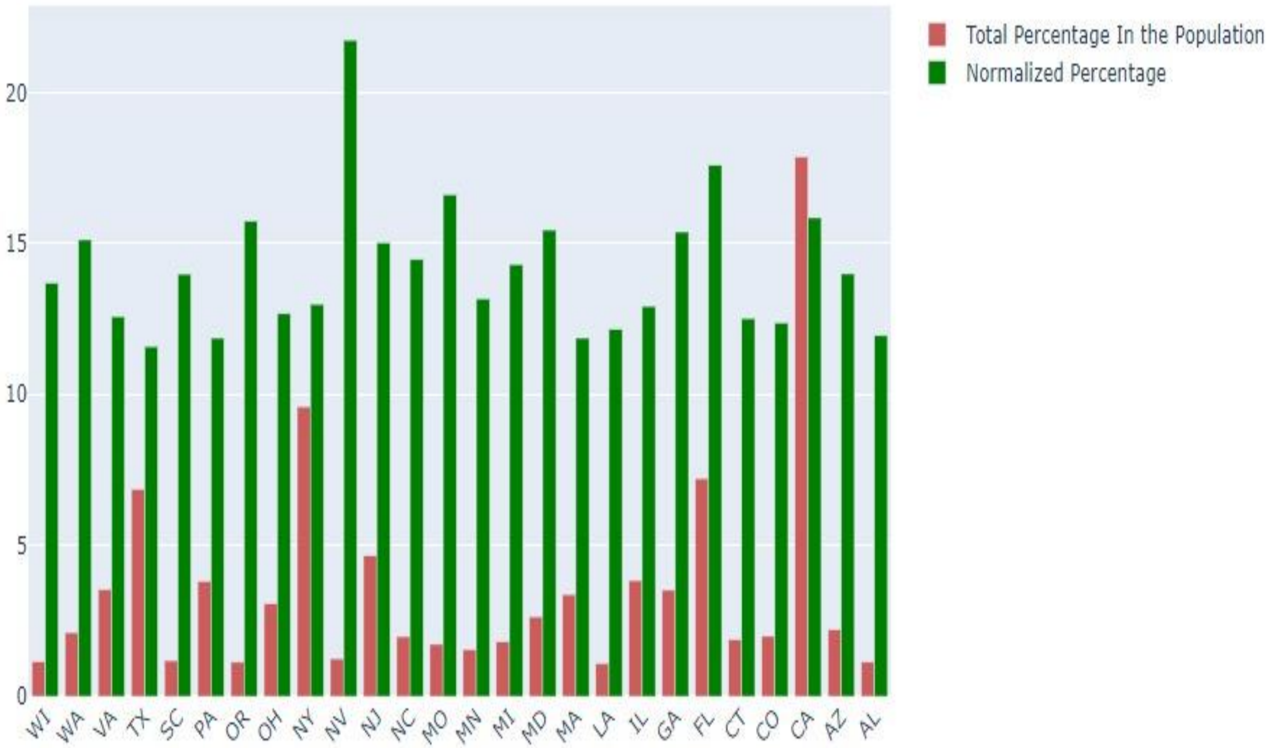


Graph on grades

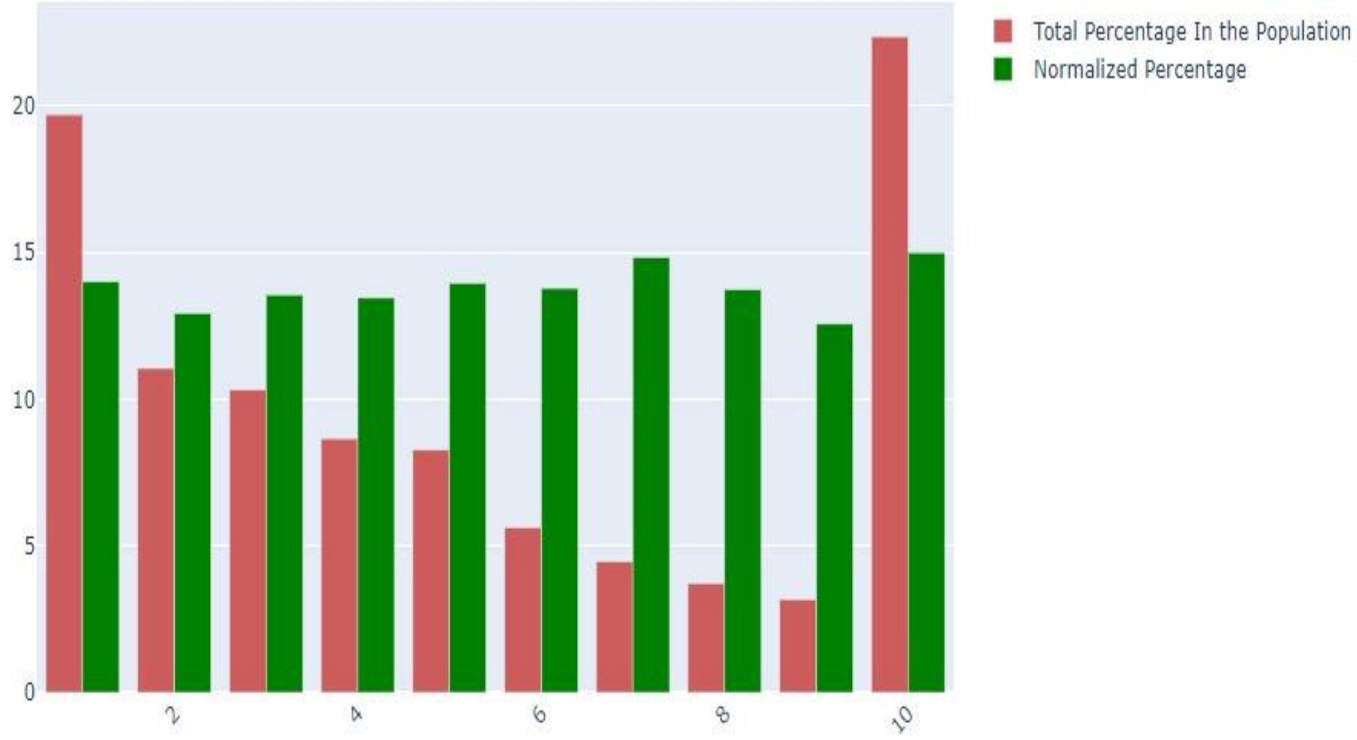


Graph on charged off

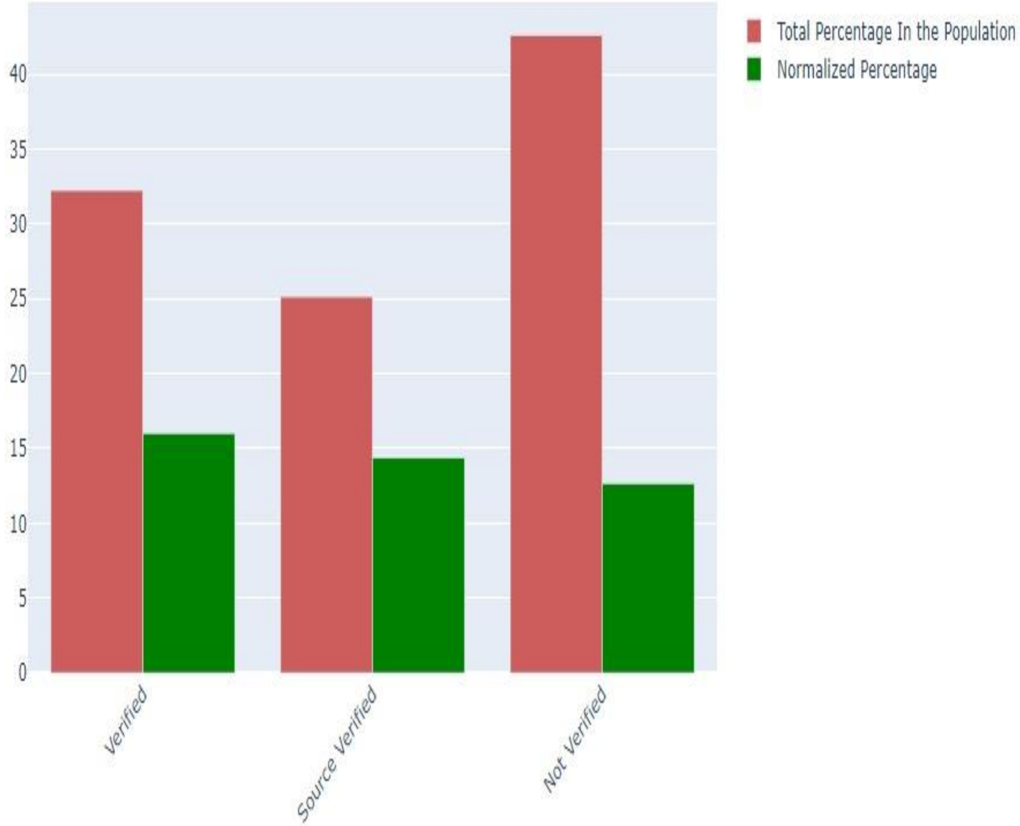


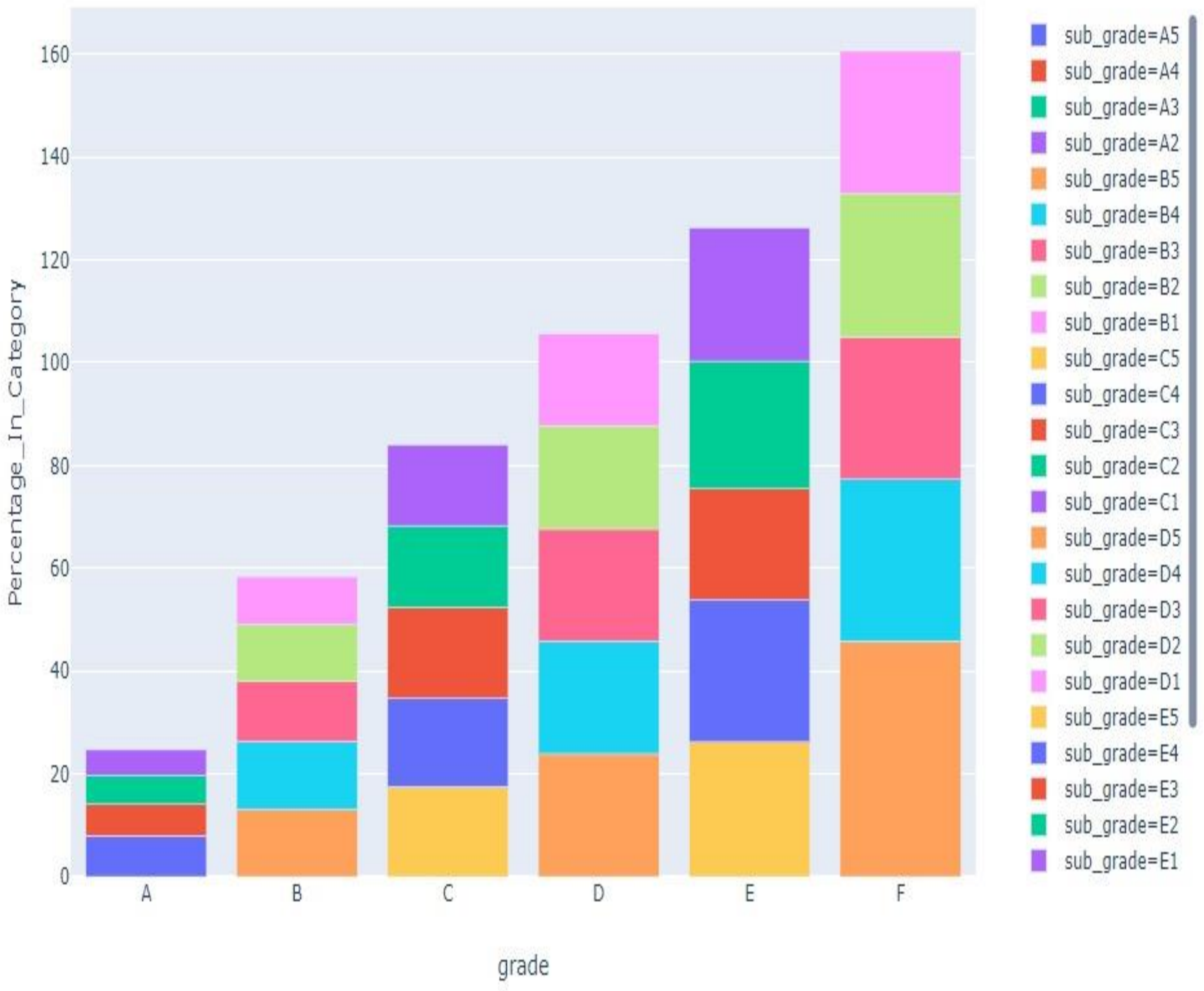


Graph on employee length

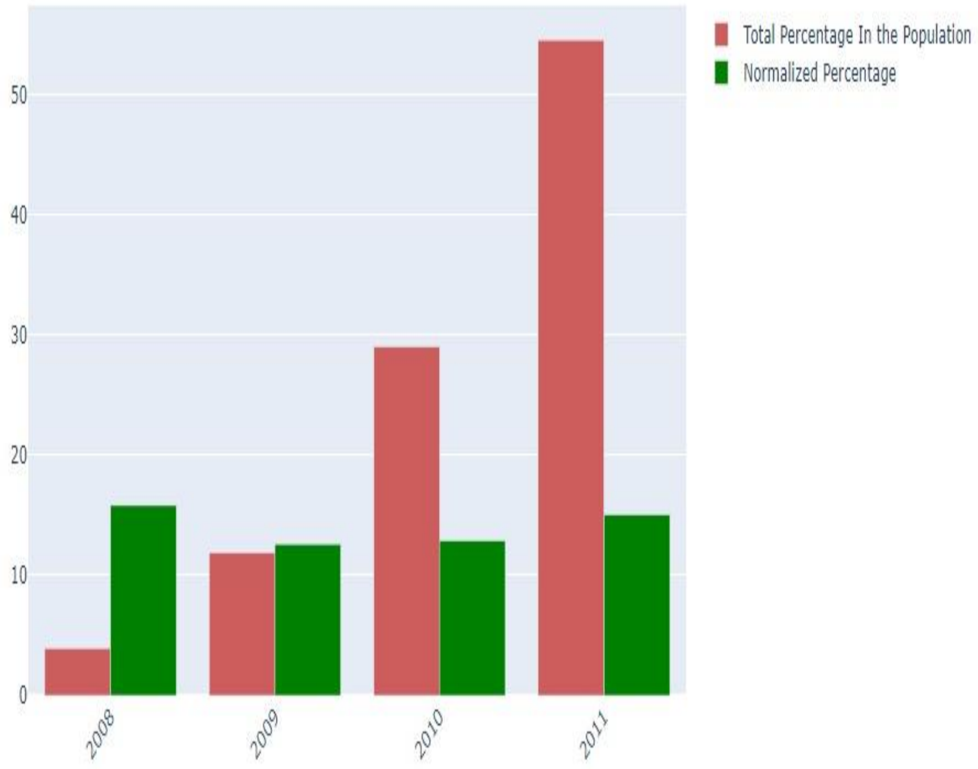


Graph on verification status



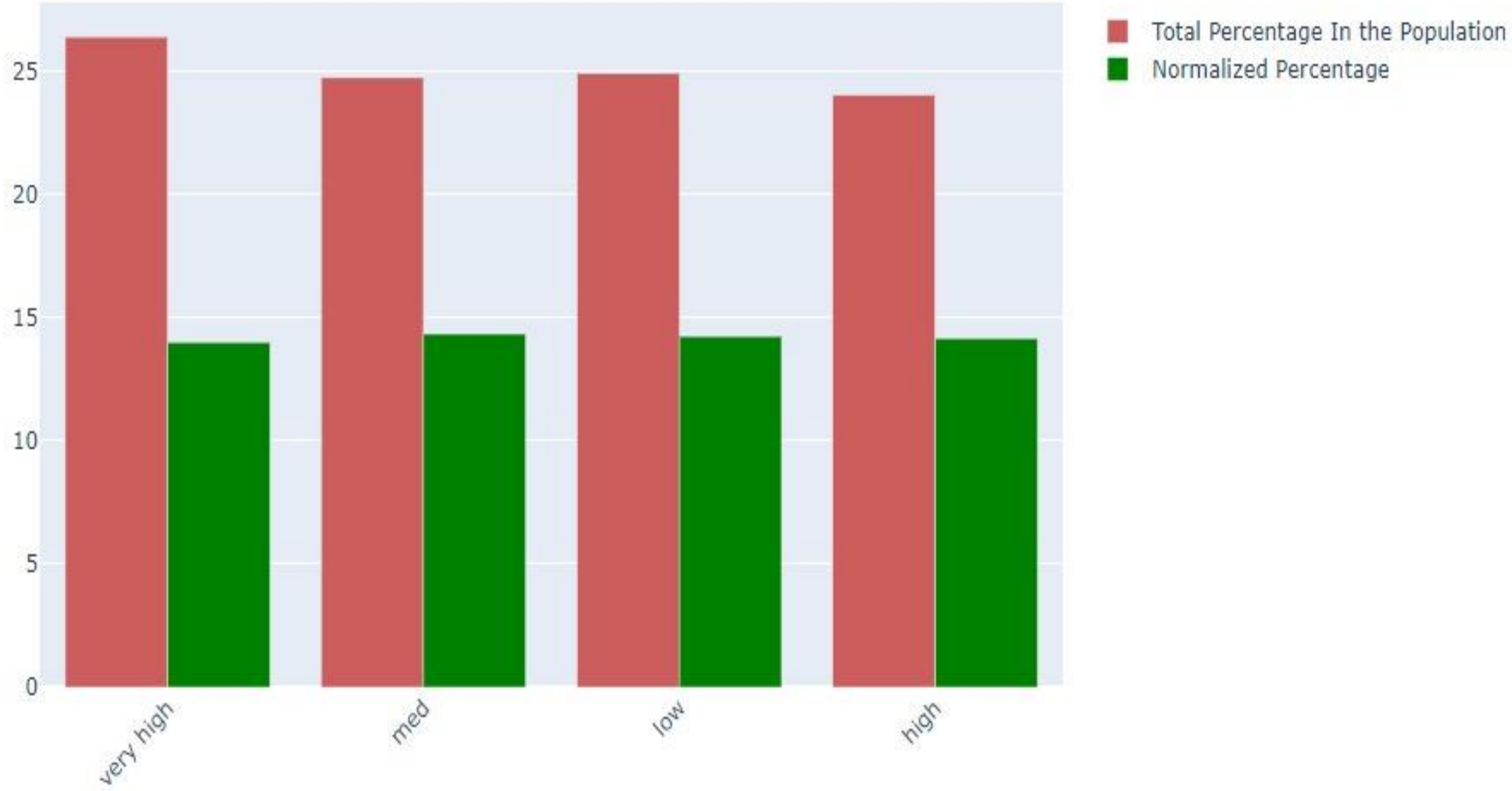


Graph on Sub-Grades

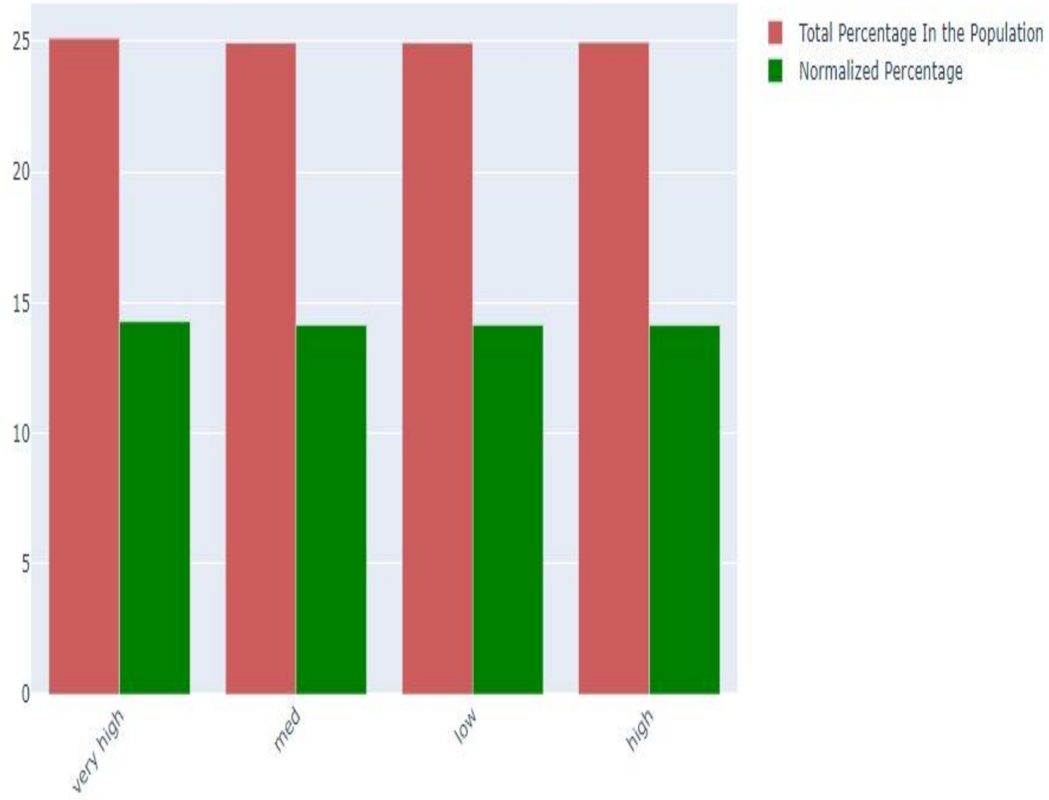


Graph on year

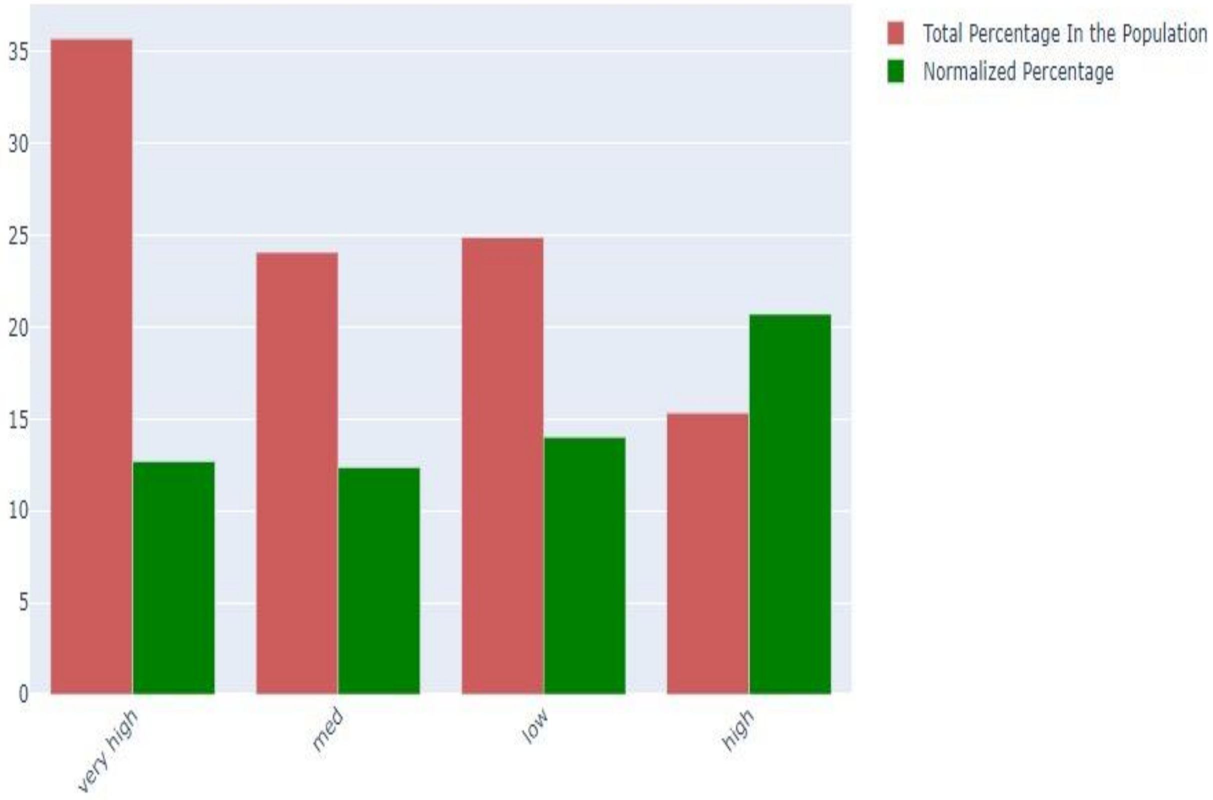
Graph on interest rate



Graph on DTI category



Graph on loan amount



Conclusion

- Approximately ~47% of the loans are issued for the purpose of debt consolidation.
- 26% of the loans for small business are charged Off.
- 18% of the loans for renewable_enrgy are charged Off.
- 17% of the loans for educational are charged Off.
- Certain sub-grades were almost certain to default compared to other sub-grades. Company need to be carefull about giving loans to people with grade = G & F.
- Loans with term of 60 months tended to be defaulted a lot more than loans with term of 36 months. Choose loans with 36 month terms.
- The home ownership does not seem to have an impact on the defaulters.
- The customers with bankruptcy records are more prone to be defaulters.
- States of Nevada (1 out of 5 people are defaulters), Florida (1 out of 6 people are defaulters) and Missouri (1 out of 6 people are defaulters) are the top defaulters.
- No. of years in experience has little or no impact on the no of defaulters.

Conclusion – Contd.

- Verification status has little impact on the no of defaulters.
- Defaulter percentage improved little bit from 2008 to 2010 but again they increased in 2011.
- Interest rate and DTI has no significant impact on the no of defaulters.
- Annual income which are in the 3rd quartile are more prone to being defaulters.