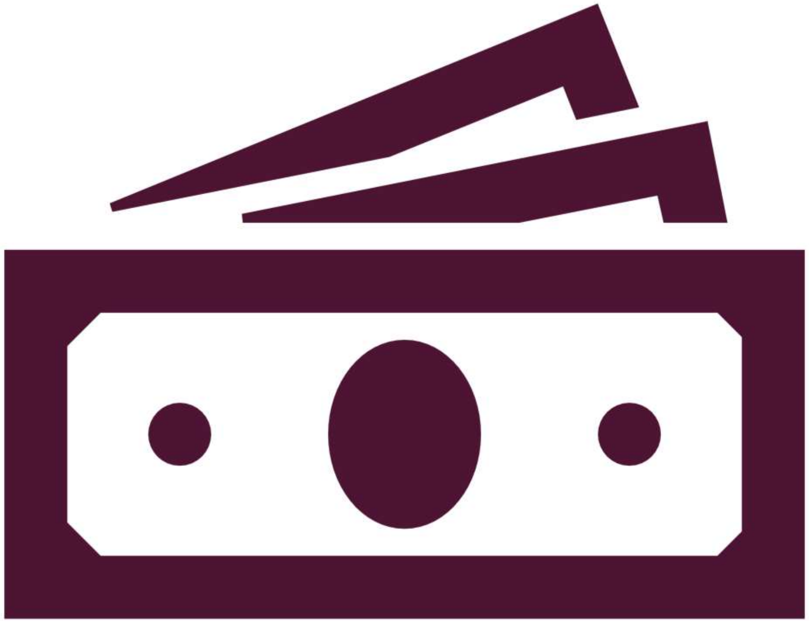


LENDING CLUB CASE STUDY

SATYAJIT MAHAPATRA

V RAJAN KUMAR RAJU

COURSE 2: STATISTICS ESSENTIALS

PROBLEM STATEMENT

- We have been provided a dataset of historical loan offerings by a financial institutions. For each loan id, the dataset identifies whether the loan was paid off or defaulted.

- The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

APPROACH

METHODOLOGY AND PACKAGES USED

The following steps were followed to analyze the pattern in the data.

A. Understanding the Data
B. Data Exploration and Cleaning
C. Univariate Analysis
D. Segmented Univariate Analysis to find key insights
E. Bivariate Analysis to conclude insights generated from above step and finding new insights
F. Business Derived Metric

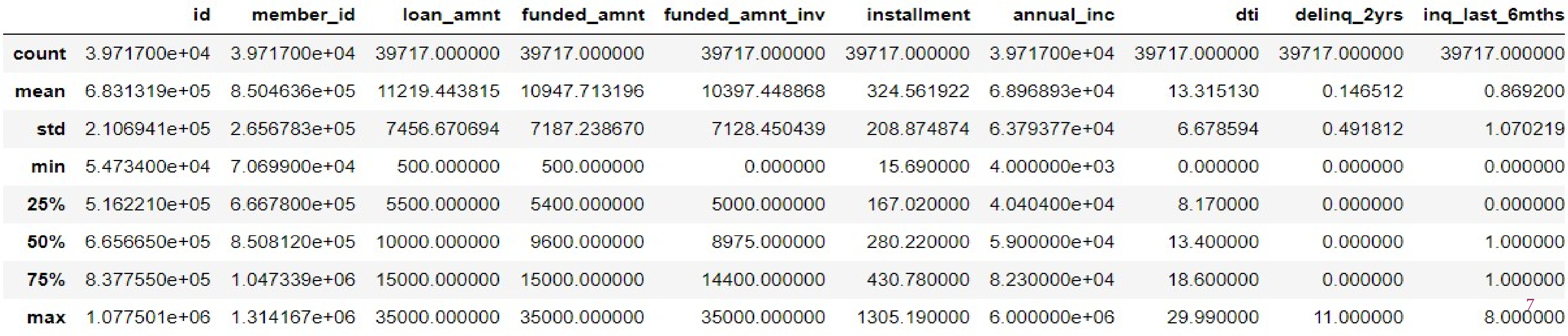
Libraries Used:

1. Pandas
2. Numpy
3. Matplotlib
4. Datetime

DATA EXPLORATION AND CLEANING

Metadata Exploration

1. The loans.csv file contains 111 columns and 39,717 rows. Each row corresponds to one loan id and the corresponding 111 columns contain multiple data points describing the features of the loan as well as the borrower.
2. The column ‘loan_status’ is our target variable which contains the status of the loan. (Current: Ongoing loan, Charged off: Default, Fully Paid: Cleared loan).
3. The data has 87 numerical columns (float64(74), int64(13)) and 24 categorical columns (object(24)).



	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000	39717.000000	3.971700e+04	39717.000000	39717.000000	39717.000000
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868	324.561922	6.896893e+04	13.315130	0.146512	0.869200
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439	208.874874	6.379377e+04	6.678594	0.491812	1.070219
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	0.000000
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000	167.020000	4.040400e+04	8.170000	0.000000	0.000000
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000	280.220000	5.900000e+04	13.400000	0.000000	1.000000
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000	430.780000	8.230000e+04	18.600000	0.000000	1.000000
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	11.000000	8.000000

Table 1: The describe function gives an idea of the data spread of the caontinous columns.

Blank, empty and single value columns were identified in the data, data was cleaned to change the data type of the columns, and loan status column was transformed to a numerical value column.

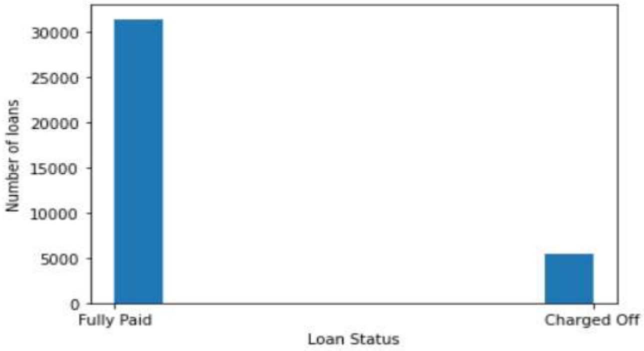
1. The data is unique at the column ‘id’ level. There were no duplicates at an overall level.
2. Based on data understanding and its business application, removed purely descriptive columns ‘url’ and ‘desc’ which do not contribute to the analysis.
3. Removed columns which were completely null or had 0. This removed 63 columns.
4. Removed columns which have only 1 unique value. This removed 9 columns.
5. Removed identifier columns: id and loan id.
6. Changed data type to Datetime from string for 2 columns: ‘issue_d’ and ‘earliest_cr_line’.
7. Upon completing the above cleaning steps, we were left with 44 columns.
8. The data was filtered out for ‘loan_status’ as current to limit the scope of analysis to just ‘Fully Paid’ and ‘Charged off’ loans.
9. Outlier Treatment was performed for Annual Income
10. Interest rate column was also treated by string "%" and making it a float

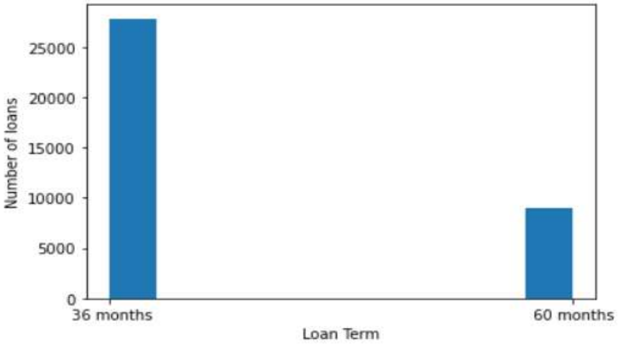
Data Transformation:

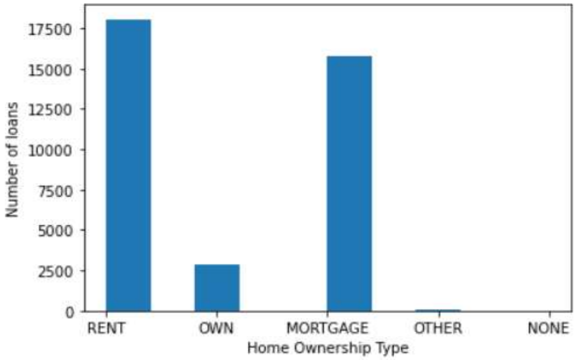
1. A charged off loan has been marked as 1 and a successfully paid loan has been marked as 0. Any summarized value closer to 1 suggests that given the factors, the loan has a higher tendency of defaulting.

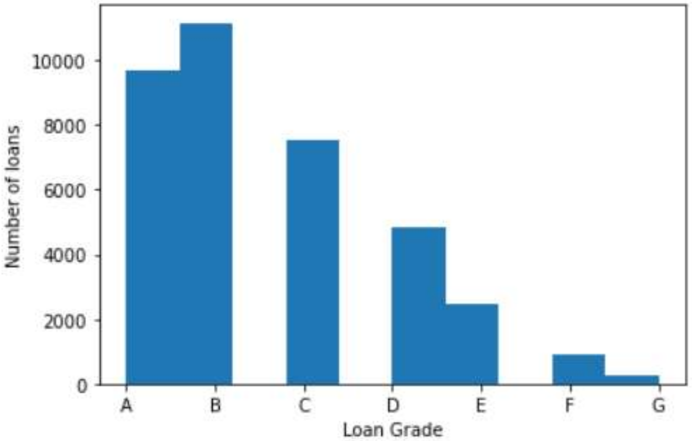
UNIVARIATE ANALYSIS

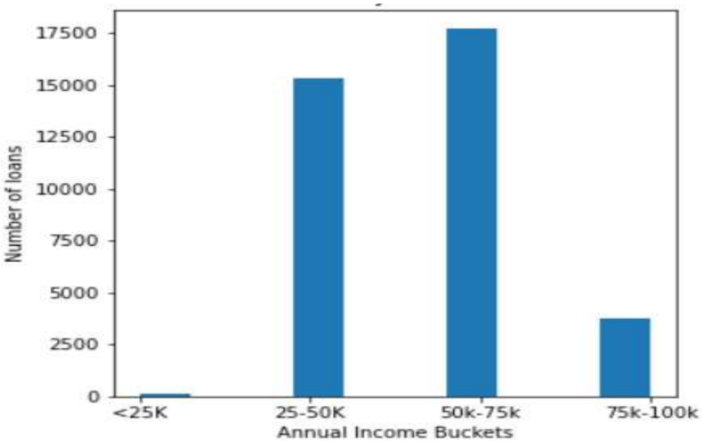
The loan status was evaluated across multiple categorical and continuous variables to understand the spread and impact of each variable on the loan status.

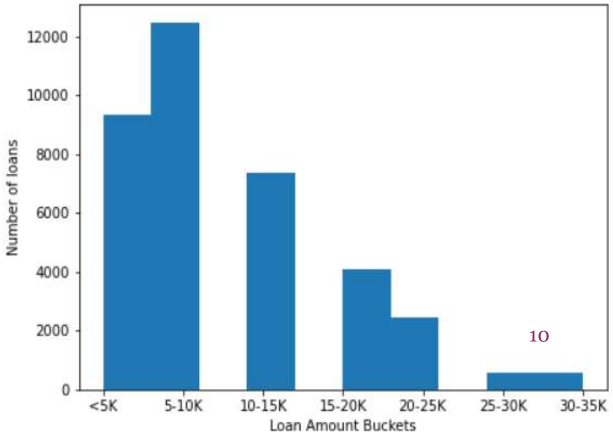






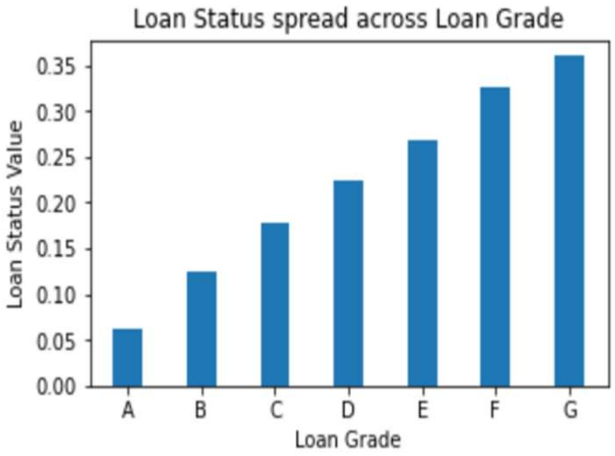






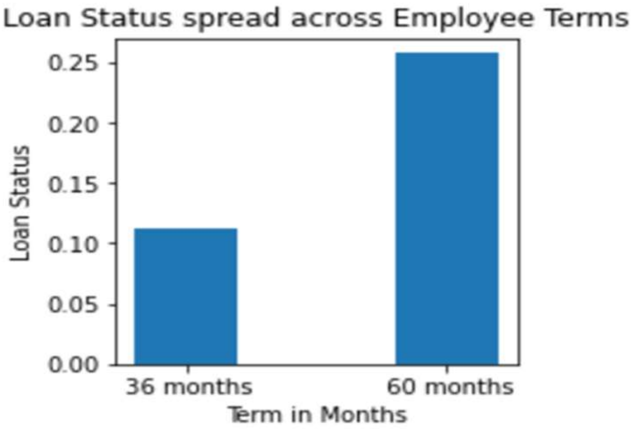
SEGMENTED UNIVARIATE ANALYSIS

Across Categorical Variables

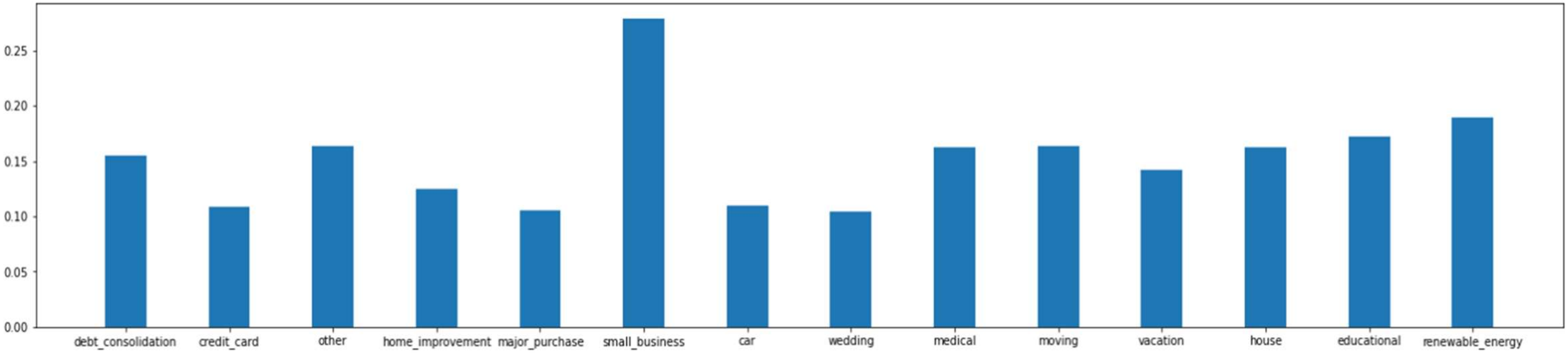


Defaulting chances of a loan increases as we move from Grade A to Grade G.

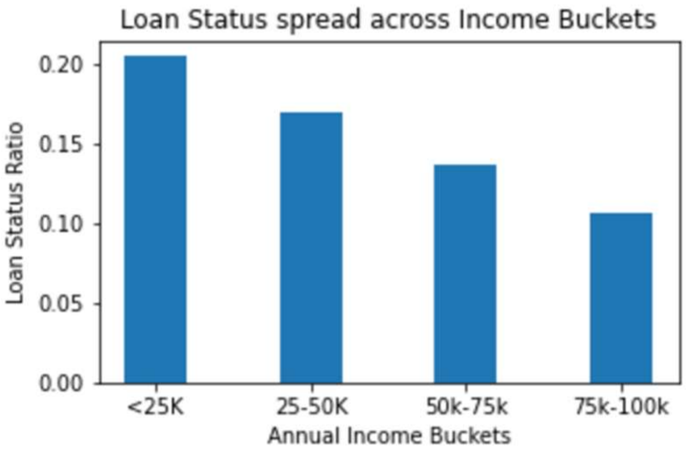
Small business purposed loans have at least 50% higher chances of defaulting as compared to any other loan purpose .



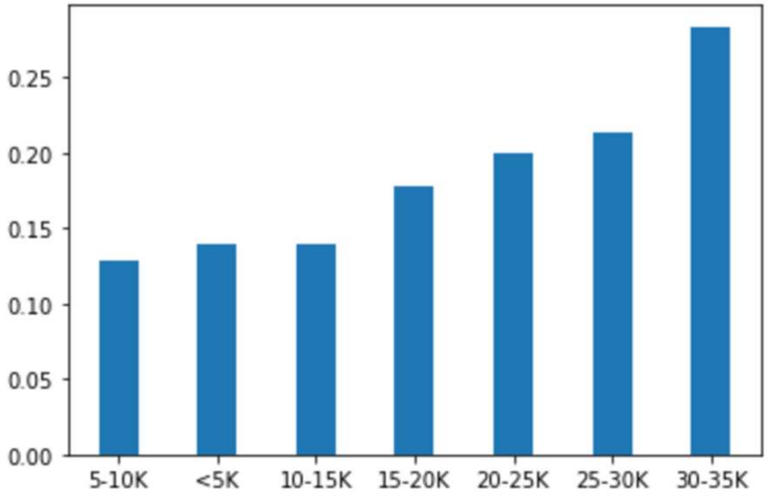
Longer duration loans are more likely to be charged off / defaulted.



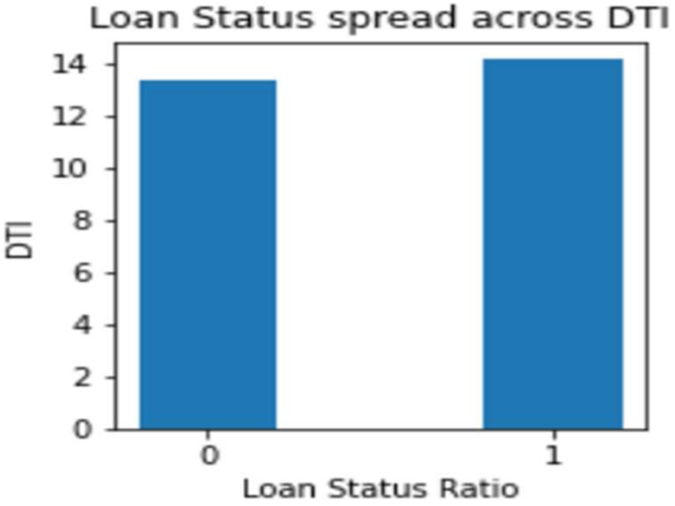
Across Continuous Variables



Lower the annual income higher becomes the chances of defaulting.



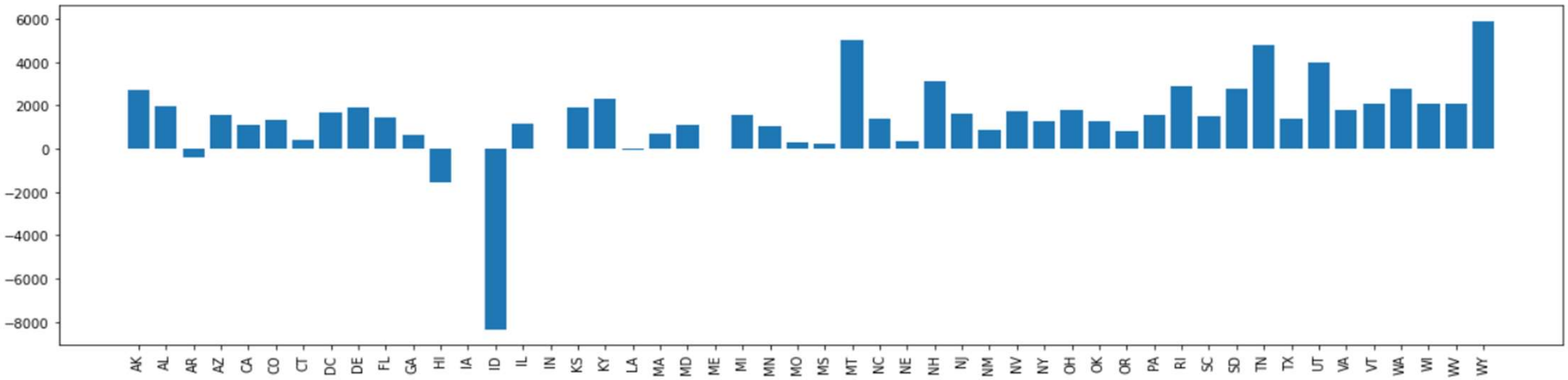
Higher loan amounts have a higher tendency of defaulting



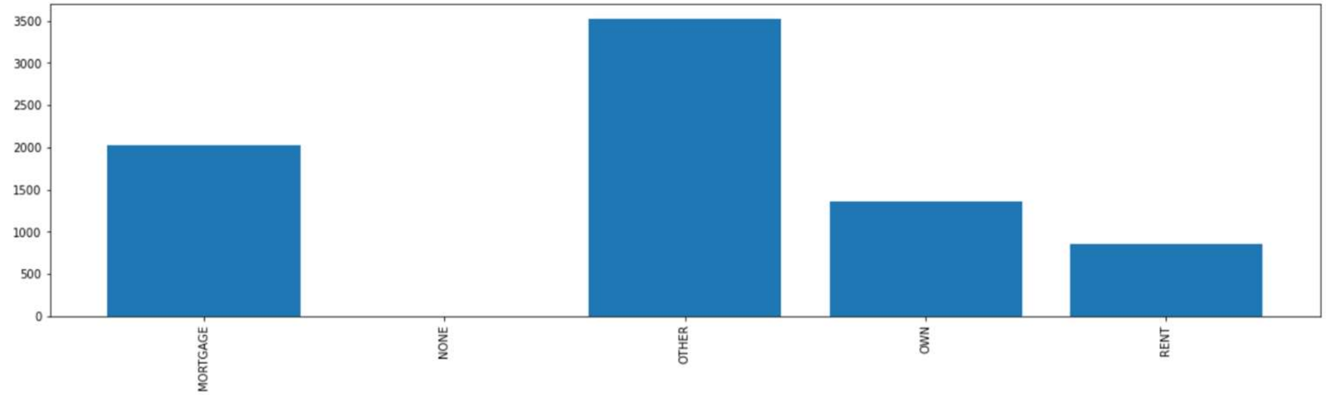
Loans that are not repaid have a higher dti, meaning the ratio of the installments or EMI amount per month is higher for unpaid loans.

BIVARIATE ANALYSIS

Performing Bivariate analysis to further validate the insight "Higher Loan amount leads to Defaulting"

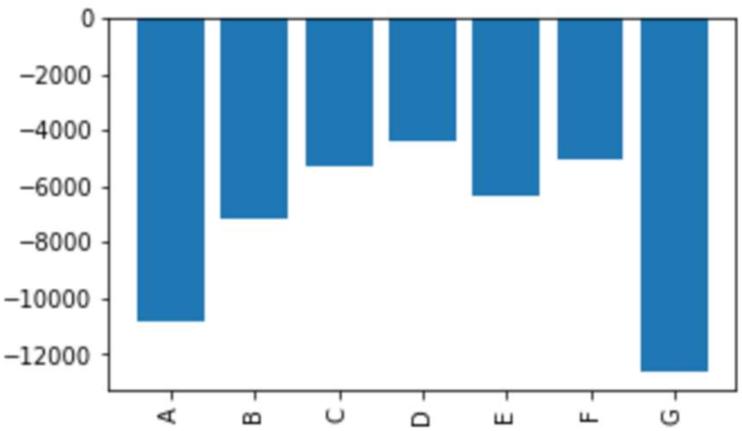


By seeing a Positive difference between the mean of loan amounts between "Charged Off" and "Fully Paid" loans across majority of the states , we further validate that higher loan amounts leads to defaulting of loans

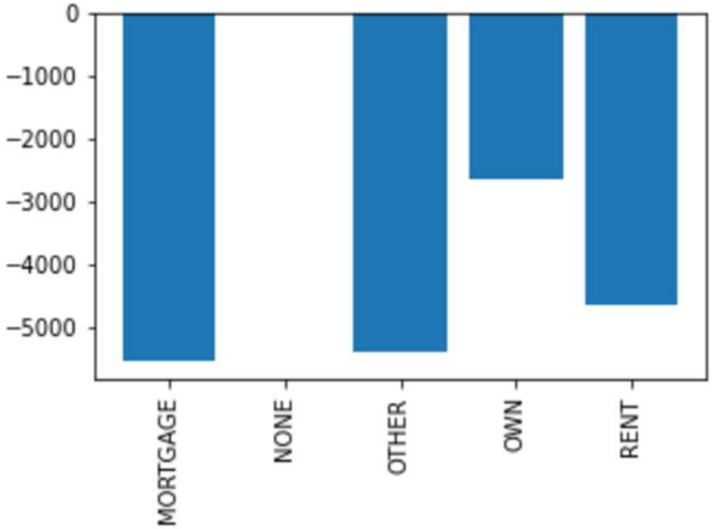


Even in case of Home Ownership there is positive difference hence further validating our insight.

Performing Bivariate analysis to further validate the insight "Lower Annual Income leads to Defaulting"

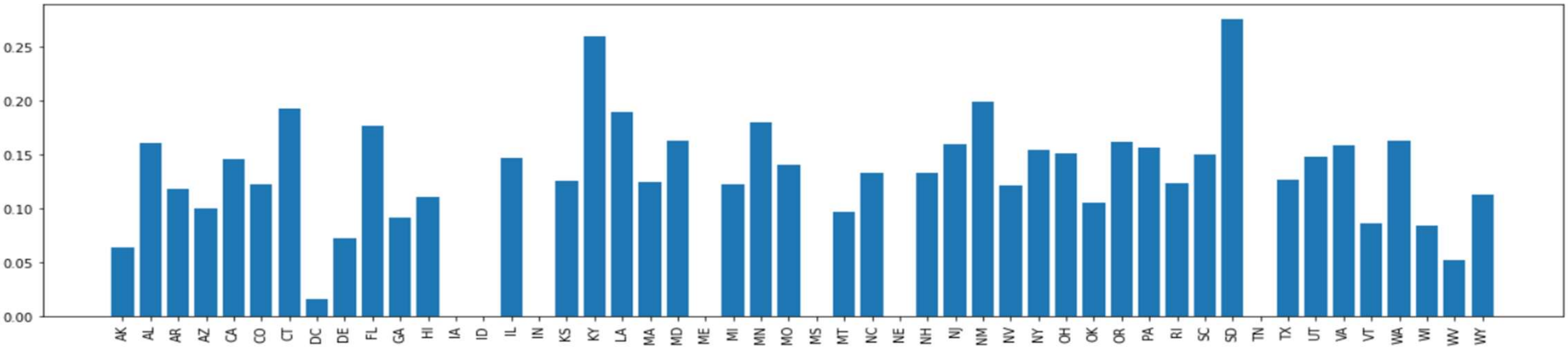


By seeing a Negative difference between the mean of Annual Incomes between "Charged Off" and "Fully Paid" loans across all the Grades , we further validate that lower Annual income leads to defaulting of loans.

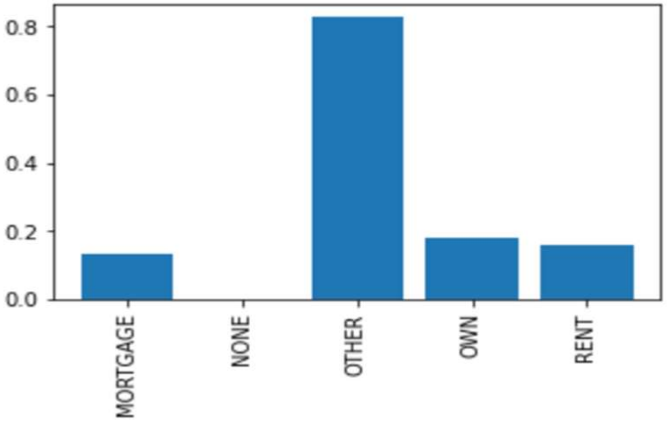


A negative difference across Home ownership helps us further validate our insight.

Performing Bivariate analysis to further validate the insight "Higher loan term leads to Defaulting"

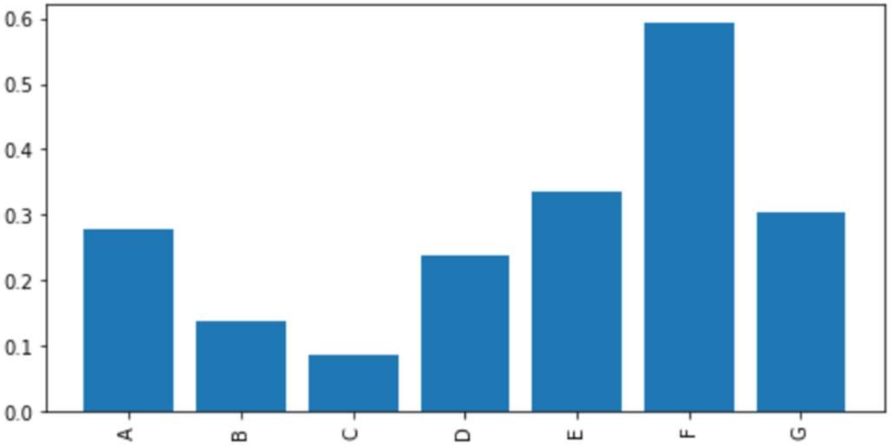


By seeing a Positive difference in terms of loan status for 60 and 36 months for all the states , we can be sure that higher loan term leads to defaulting.

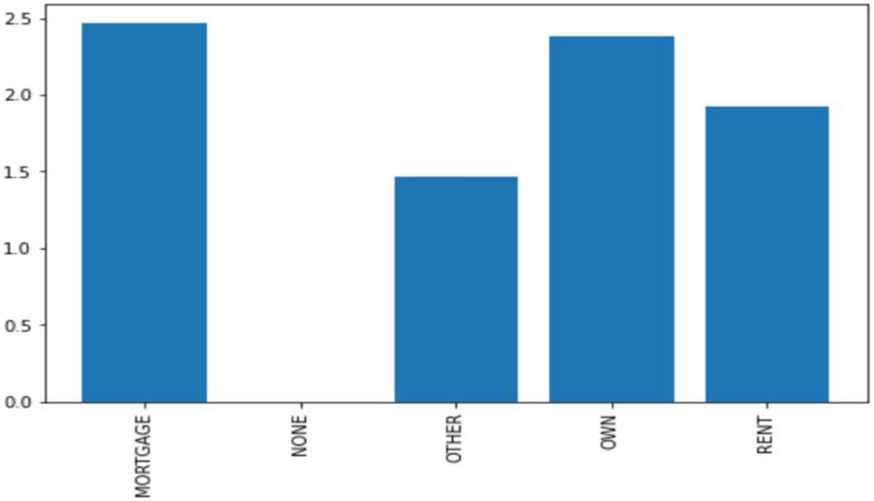


A positive difference across home ownership validates our above insight.

Performing Bivariate analysis to conclude new insight "Higher interest rate leads to Defaulting"



By seeing a positive difference between the interest rates for Charged Off and Fully paid loans across grades , we can conclude that higher interest rates create higher defaulting tendency.



A positive difference between the interest rates between Charged Off and Fully Paid loans across home ownerships conclude our above insight.

BUSINESS DERIVED METRIC

Creating new Derived Business Metrics

- The ratio between the monthly installment and monthly income also creates a simple deciding metric to prevent defaulting.

- As per industry standards our research suggested the above ratio should lie between 30-35 %

RESULTS

Insights Generation

Deciding Factors while providing loans:-

- Higher is the loan amount, higher are the chances of loan default.
- Lower is the annual income, higher are the chances of loan default.
- Higher is the interest rate, higher are the chances of loan default.
- Longer is the loan term, higher are the chances of loan default.
- Higher is the dti, higher are the chances of loan default.
- Higher is the alphabet value of the loan, higher is the probability of default.

Observation:-

Small business loans as a purpose has shown much higher defaulting rate as compared to any other loan purpose

Business Metric Suggestion:-

Monthly installment by Monthly income ratio should lie between 30-35%