Lending club case study:

PROBLEM STATEMENT : To identify variables which are strong indicators of default and potentially use the insights in approval /rejection loan status ,here approval is fully paid and rejected are charged off customers.

Journey of our Analysis

Understand Data/Domain

1.Data Sourcing by loading data
2.Understand Data dictionary
3.Identify Behaviour variables which will not be available at the time of loan application
4. Identify columns which are not needed
5.Identify rows which are not needed

Data cleaning/readiness

1. Removed columns having missing values for 75% of records
2.No rows found with more than 5 missing values
3.Removed columns with constant value
4. Fixed Datatype and cleaned data for int_rate, employment length, term,etc
5. Removed behaviour columns as it is not needed
6. Created column with numeric value for loan status

Univariate Analysis

1. Performed analysis on target variable – loan status
2. Performed analysis on categorical variable. Eg purpose
3. Performed analysis on continuous variable. Eg loan amount

Segmented Univariate

1. In this step we choose one or more variables which could influence the relationship between target and other variables
2. We divided the dataset into segments based on chosen segmentation variables
3. Compared the result of univariate analysis across different segments

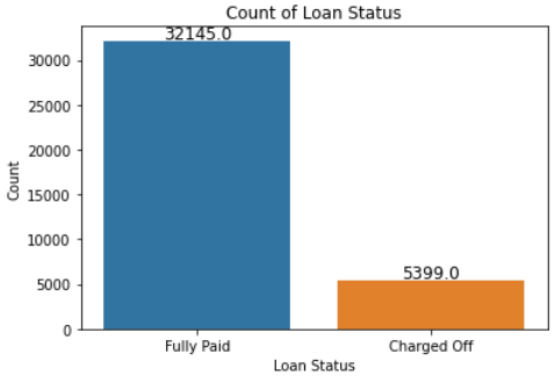
Bivariate Analysis

1. Here we did visualisations that show the relationship between two variables like barplots ,countplots,bar charts etc.
2. We found some results on correlation which measures strength of liner relationship
3. To observe trends and patterns to understand how changes in one variable correspond to changes in the other.

Univariate Analysis Results

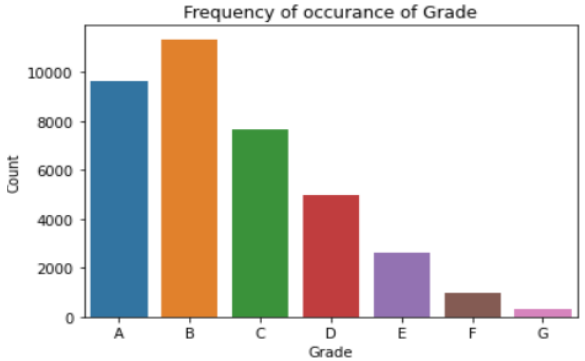
Plots and conclusion derived out of our analysis on variables are given below:

Loan Status



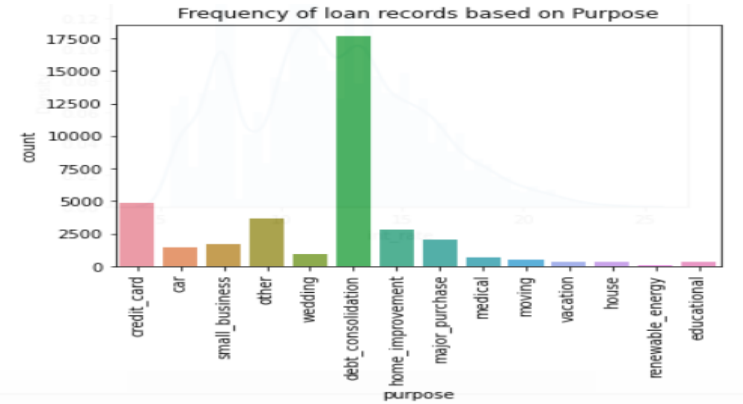
Observation : Source dataset has 5399 defaulter records.

Grade



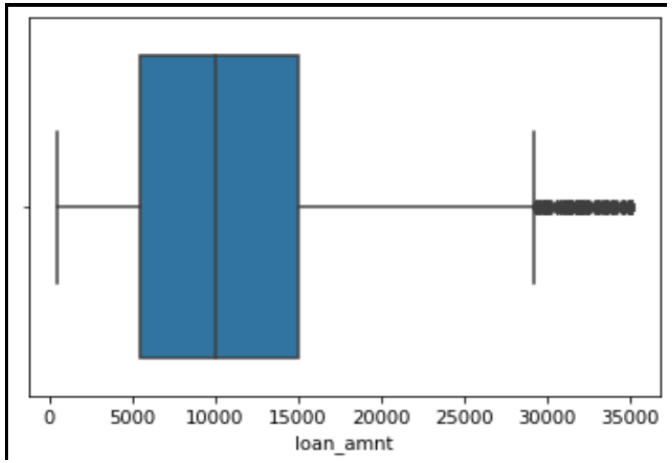
Observation : Out of fully paid/Defaulted customers, more loans has been provided for Grade B

Purpose

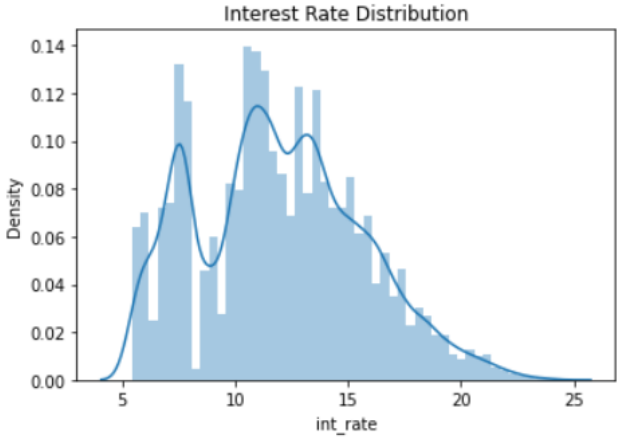


Observation : Top 4 types of loans are: consolidation, credit card, home improvement and major purchase.

Loan amount



Interest rate

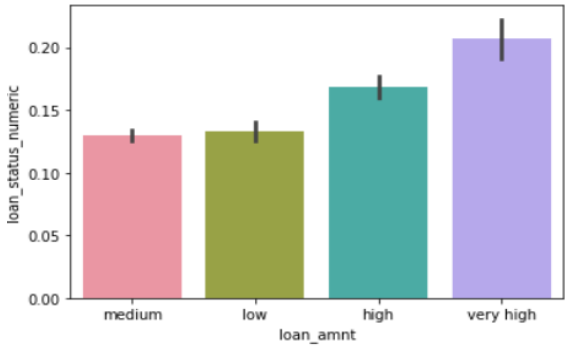


Observation : Possibility of 2 segments of loans with interest rate

Univariate Analysis -Segmented

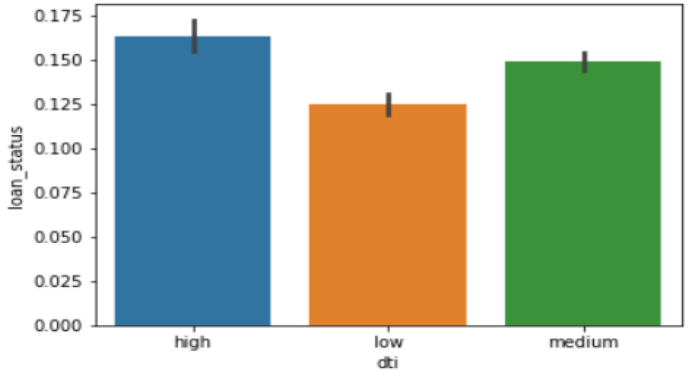
Plots and conclusion derived out of our analysis on variables are given below:

Loan amount



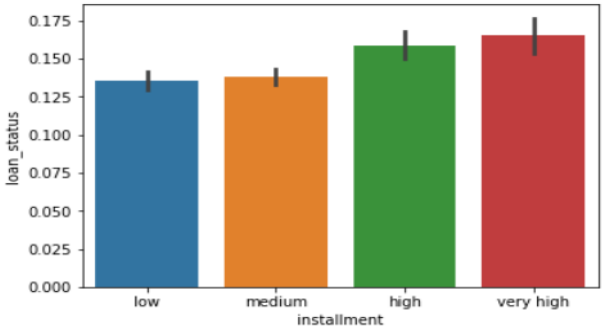
Observation : higher the loan amount, higher the default rate

DTI



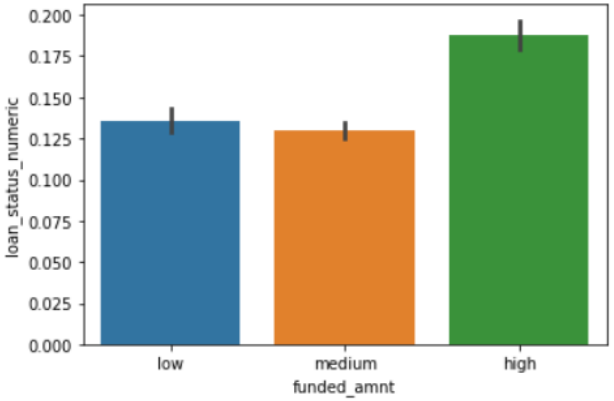
Observation : high dti translates into higher default rates

Installment



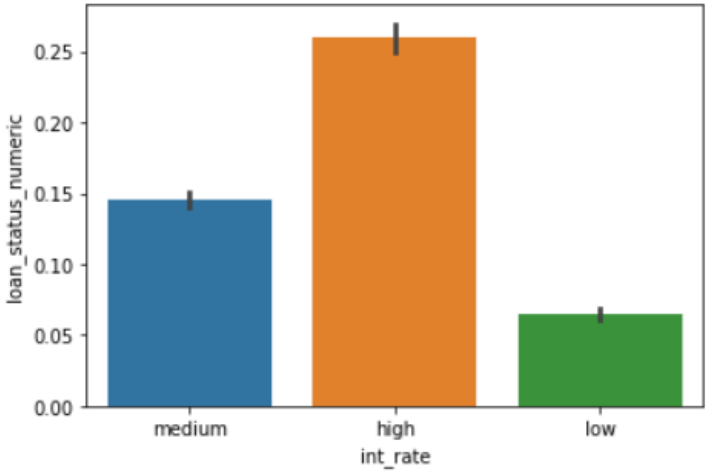
Observation : the higher the installment amount, the higher the default rate

Funded amount



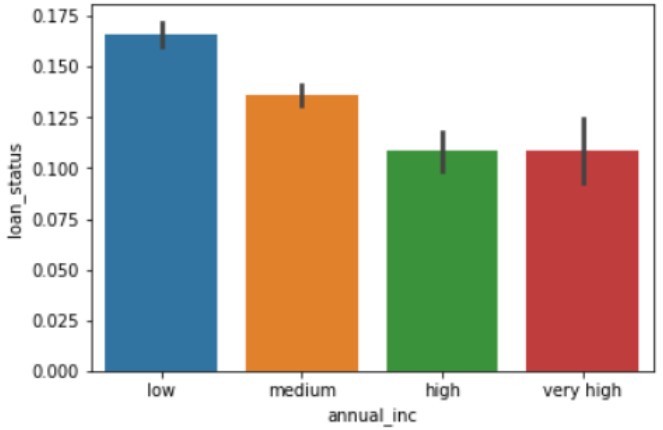
Observation : higher the funded amount, higher the default rate

Interest rate



Observation : high interest rates default more, as expected

Annual Income

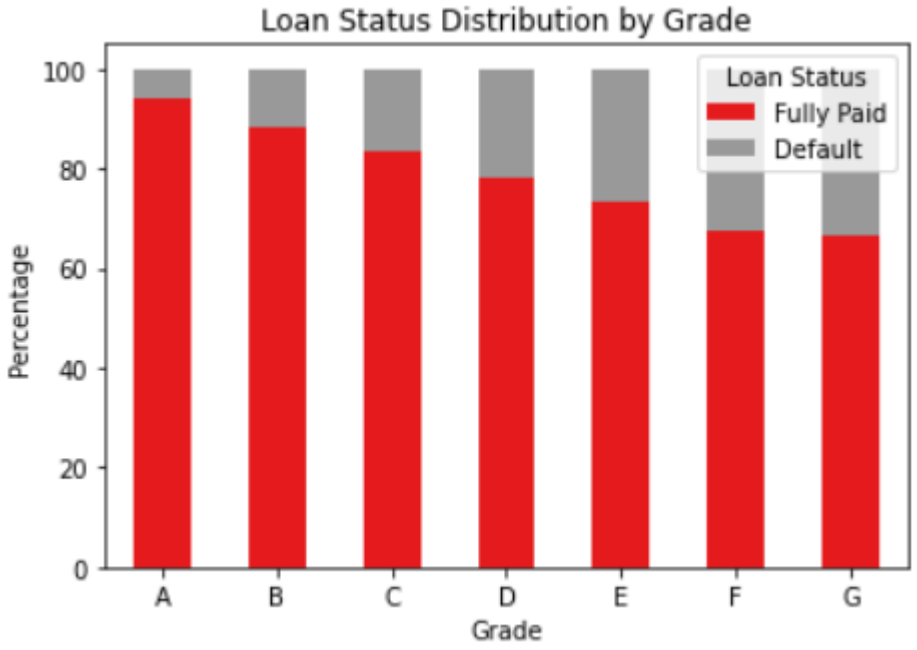


Observation :lower the annual income, higher the default rate

Bivariate Analysis Results of Categorical Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Grade

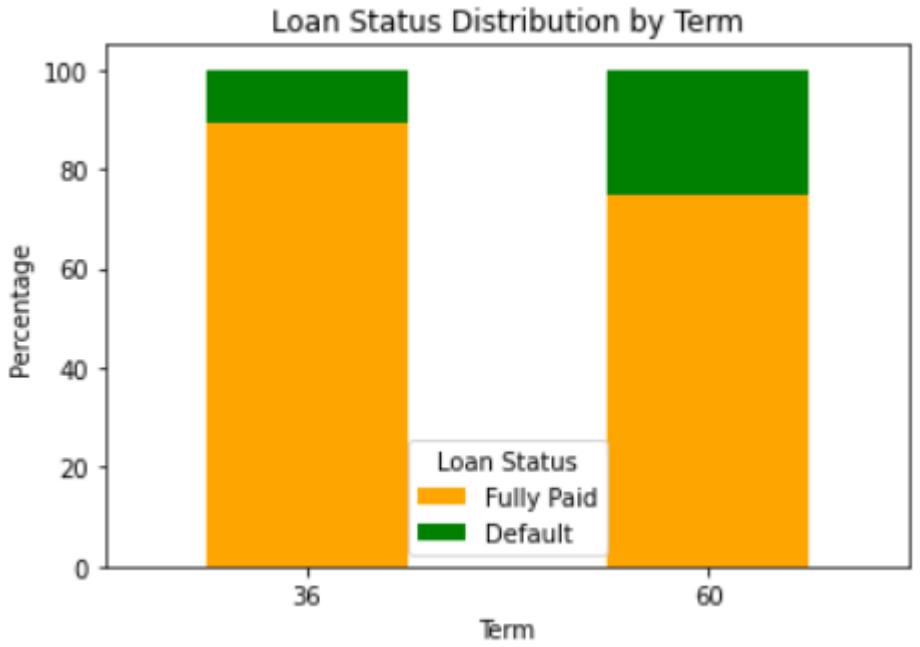


Conclusion: As the grade of loan goes from A to G, the default rate increases. This is expected because the grade is decided by Lending Club based on the riskiness of the loan.

Bivariate Analysis Results of Categorical Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Term

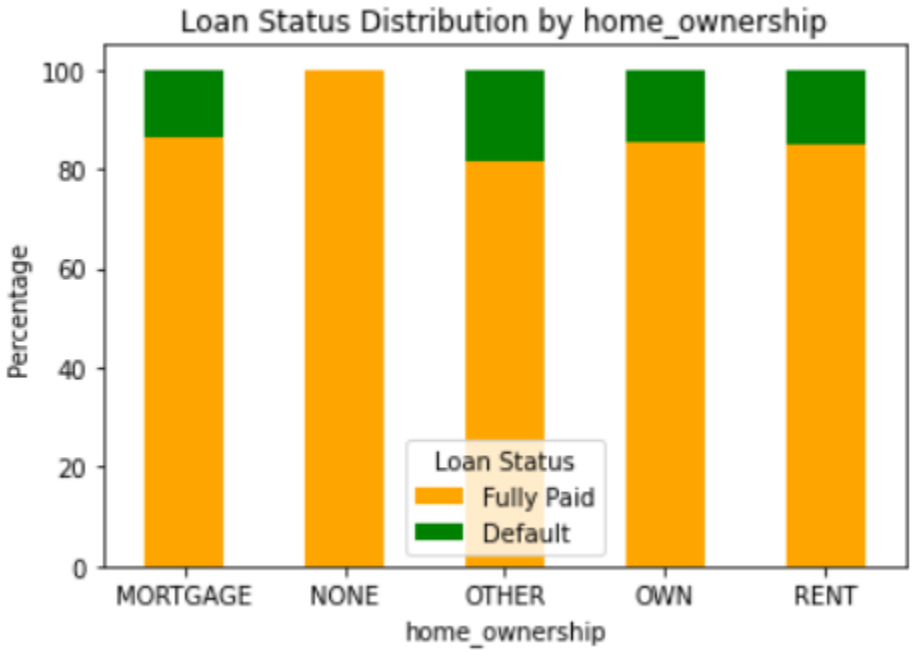


Conclusion: 60 months loans default more than 36 months loans

Bivariate Analysis Results of Categorical Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Home ownership

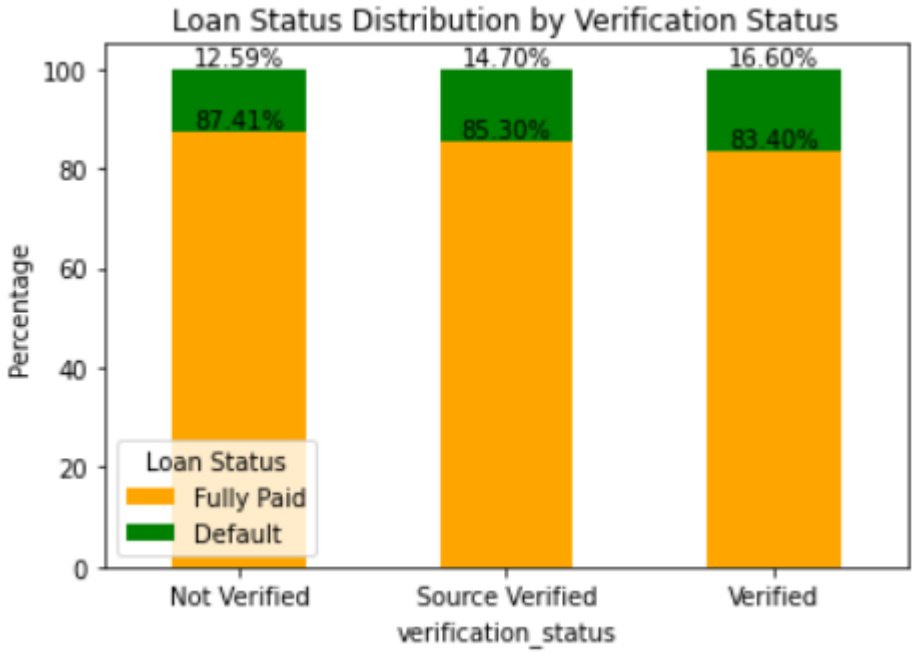


Conclusion: Irrespective of whether someone have owned home or not, there is no variation on defaulter rate

Bivariate Analysis Results of Categorical Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Verification status

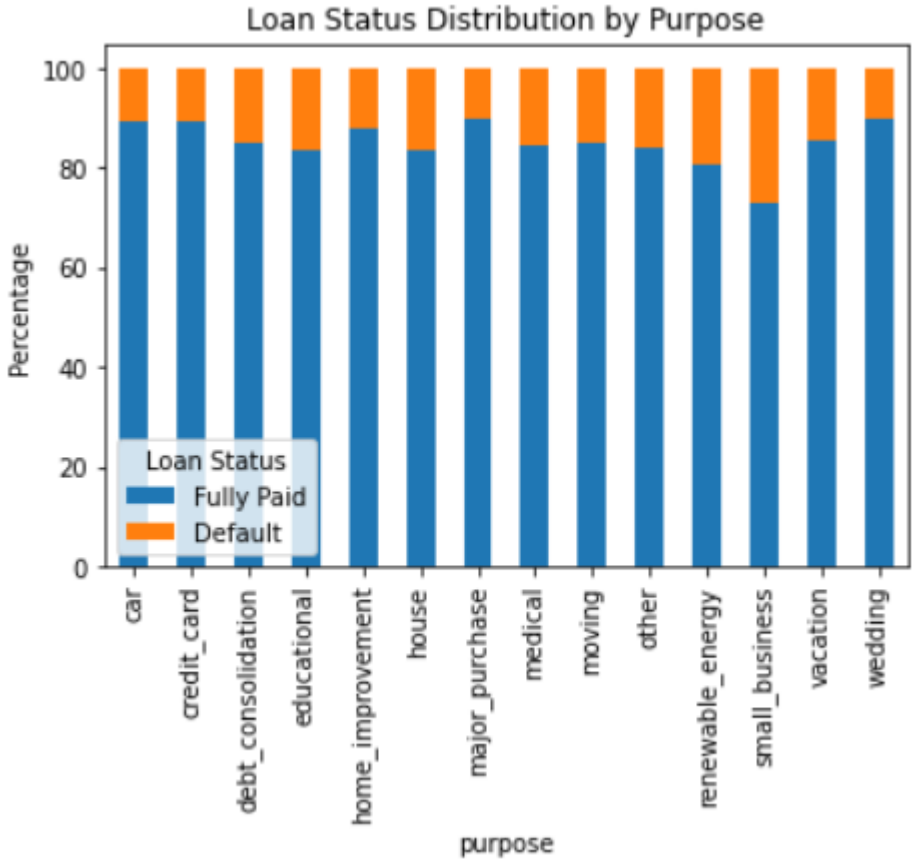


Conclusion: Verified sources has relatively higher default rate

Bivariate Analysis Results of Categorical Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Purpose

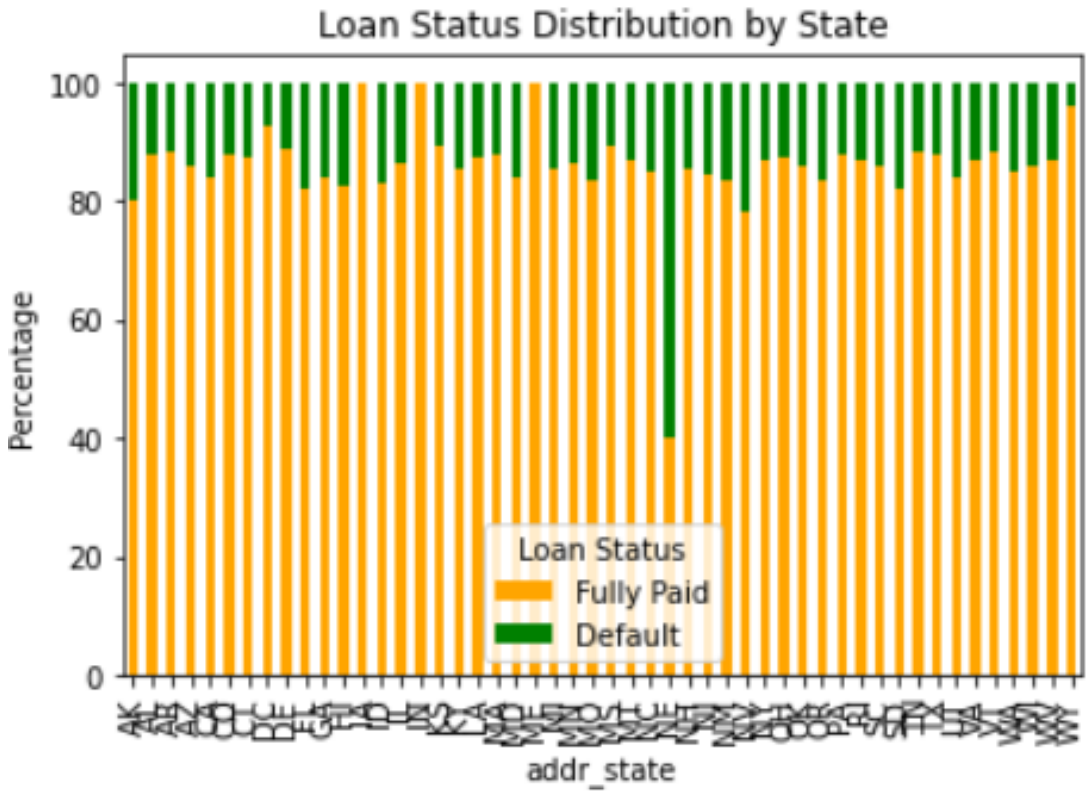


Conclusion: Small business has high defaulter rate followed by renewable energy and education

Bivariate Analysis Results of Categorical Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Address State

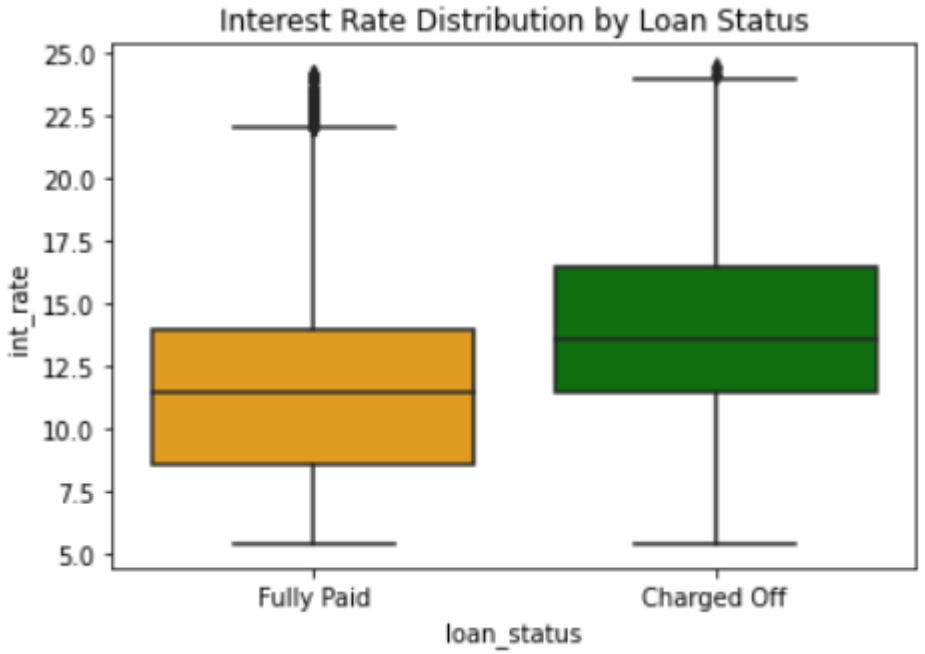


Conclusion: NE(Nebraska) state has higher defaulter rate

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Interest rate

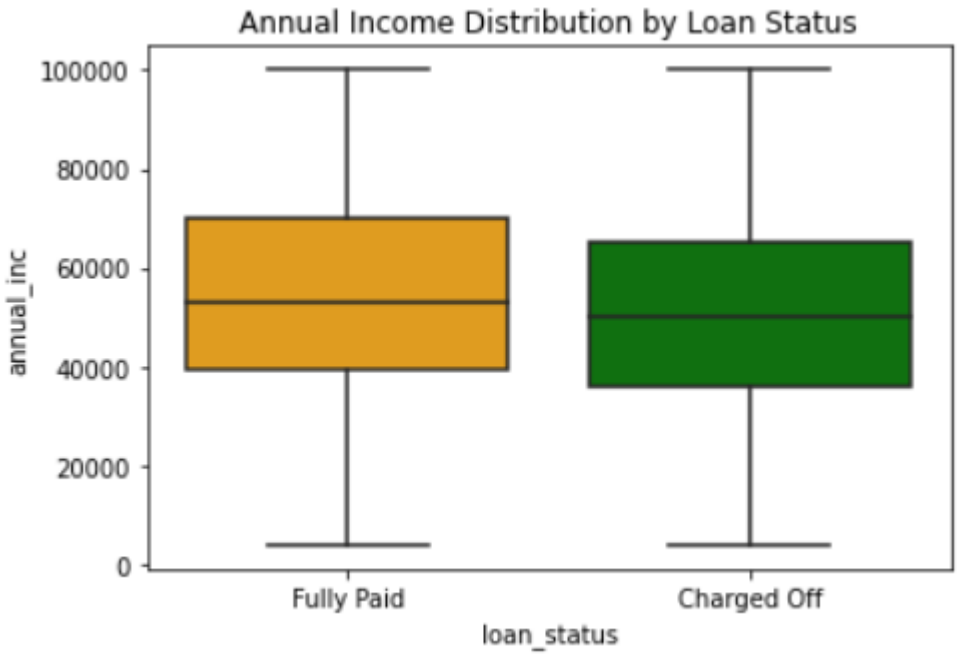


Conclusion: Majority of the Defaulter's has higher interest rate compared to fully paid customers. Upper 25% of defaulters has higher interest rate than fully paid customer's interest rate

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Annual Income

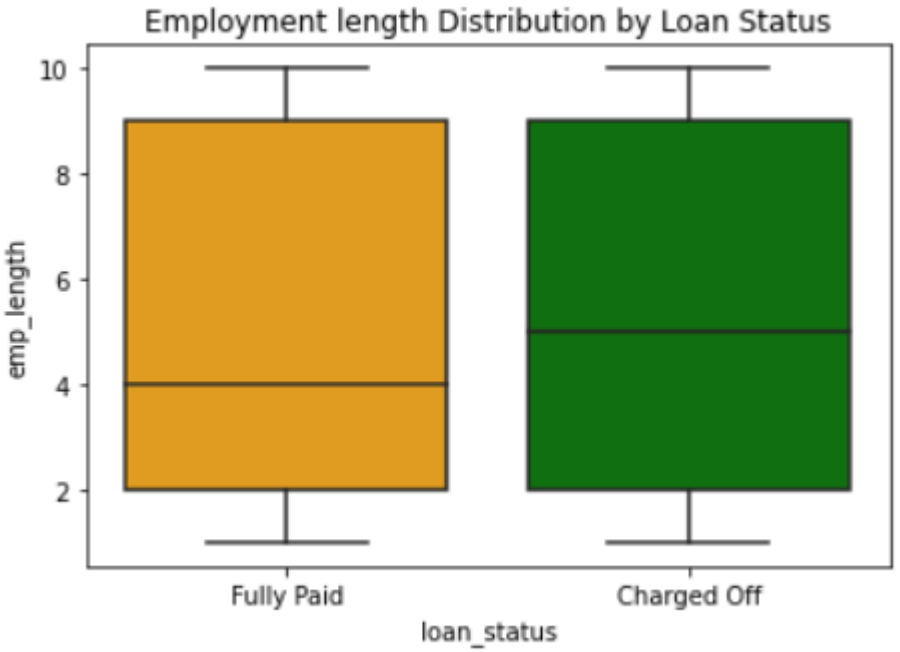


Conclusion: Defaulter’s annual income range is comparatively lesser than most of the fully paid customers.

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Employment length

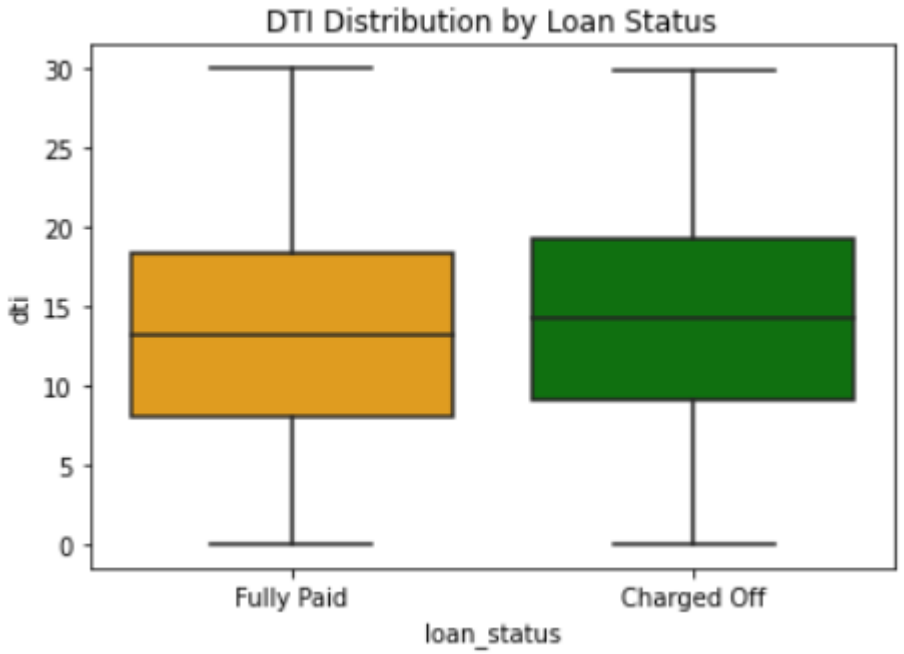


Conclusion: Median of Defaulter's employment length is higher than fully paid customers.
For fully paid customers we have more values between 4 and 9 , where as defauters's employment length are kind of equally distributed.
emp_length is not much influencing loan_status.

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs DTI

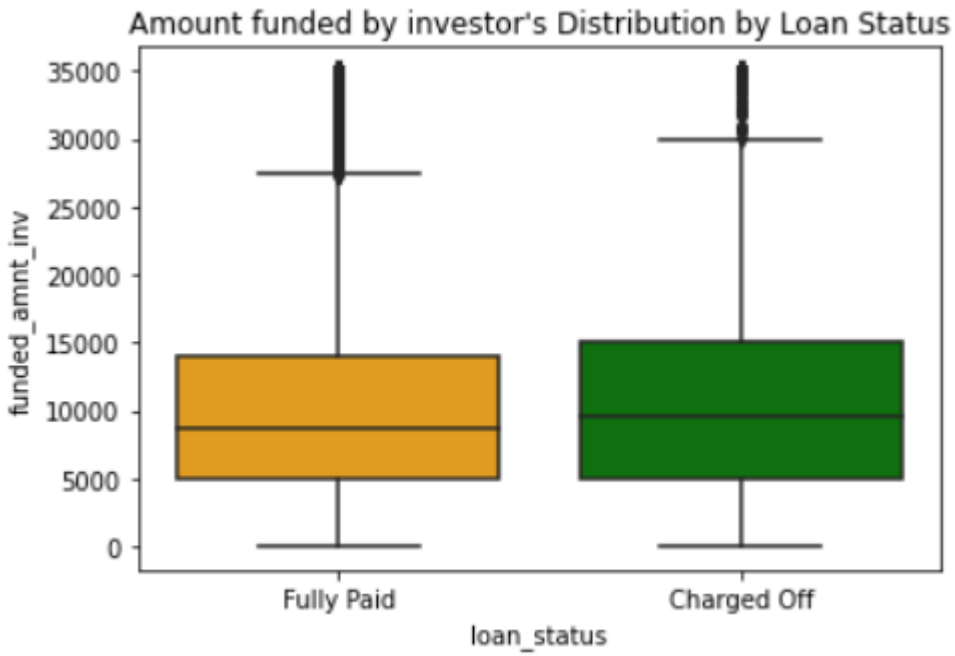


Conclusion: Most of the defaulters has higher dti lower range compared to fully paid customers

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs funded_amnt_inv

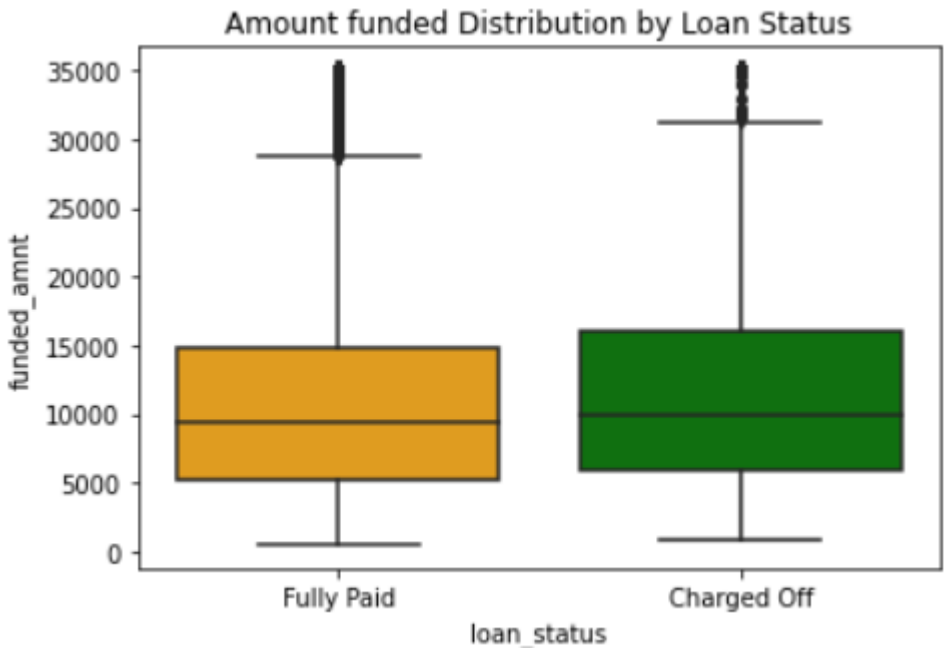


Conclusion: More the funded amount, higher the possibility of defaulters. Defaulter's upper amount range increased compared to fully paid customers

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs funded_amnt

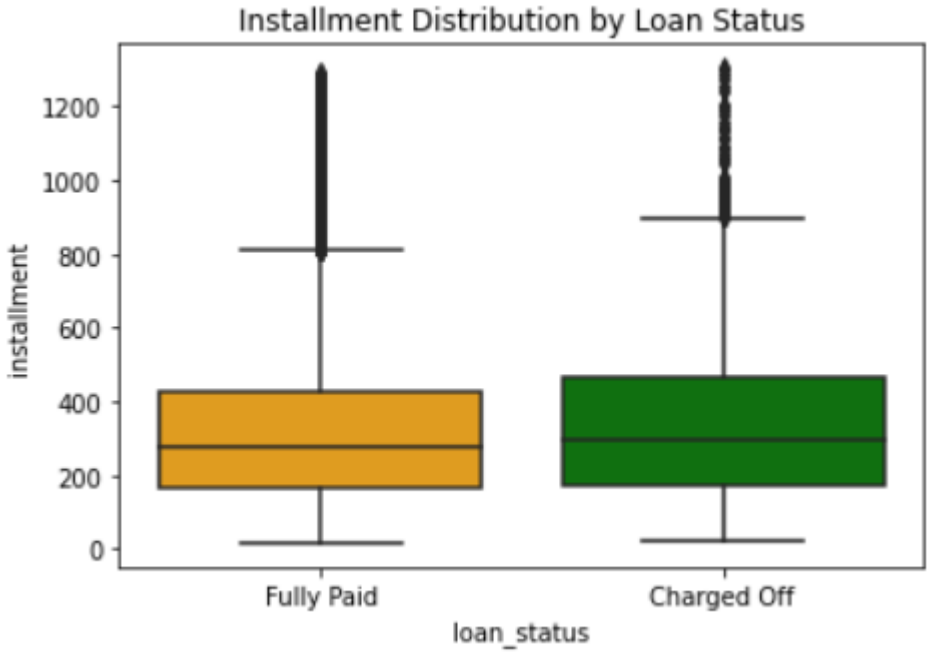


Conclusion: When the funded amount increase, higher is the chances to get defaulted. Defaulter's upper amount range increased compared to fully paid customers.

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Installment



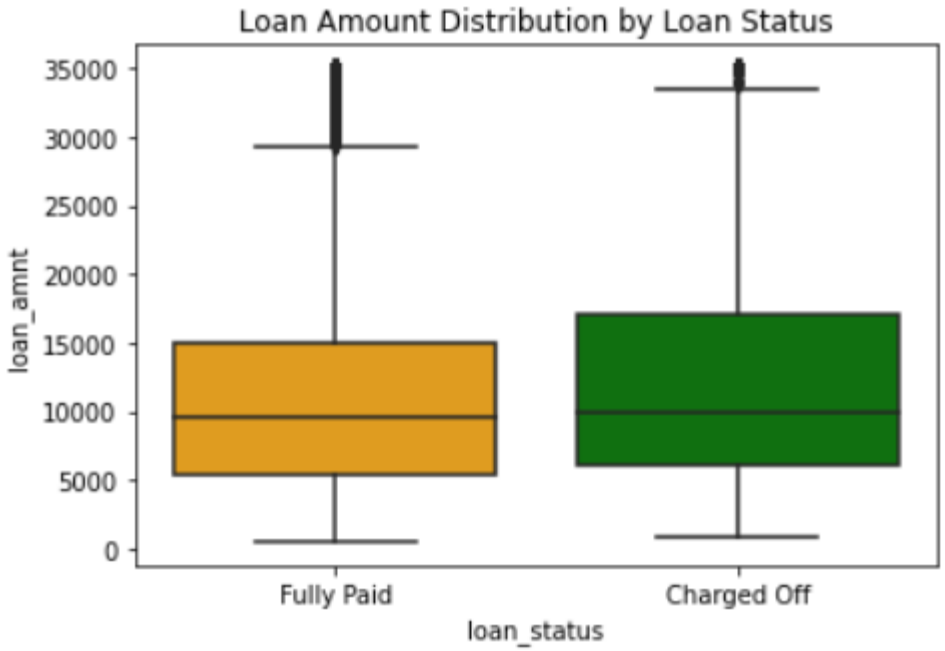
Conclusion: 1)Max people who fully paid loan has lesser installment compared to defaulters. Both IQR box sizes are different, it indicates a difference in the variability (spread) of 'installment' amounts between the two loan statuses.

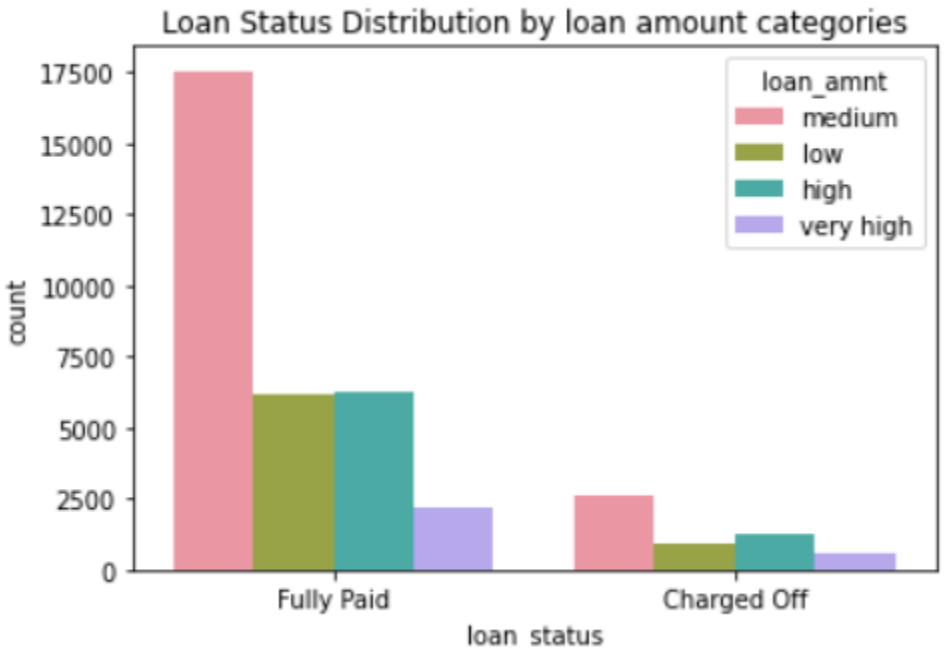
2)When the Installment increase, higher is the chances to get defaulted. Defaulter's upper installment range increased compared to fully paid customers.

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs loan amount





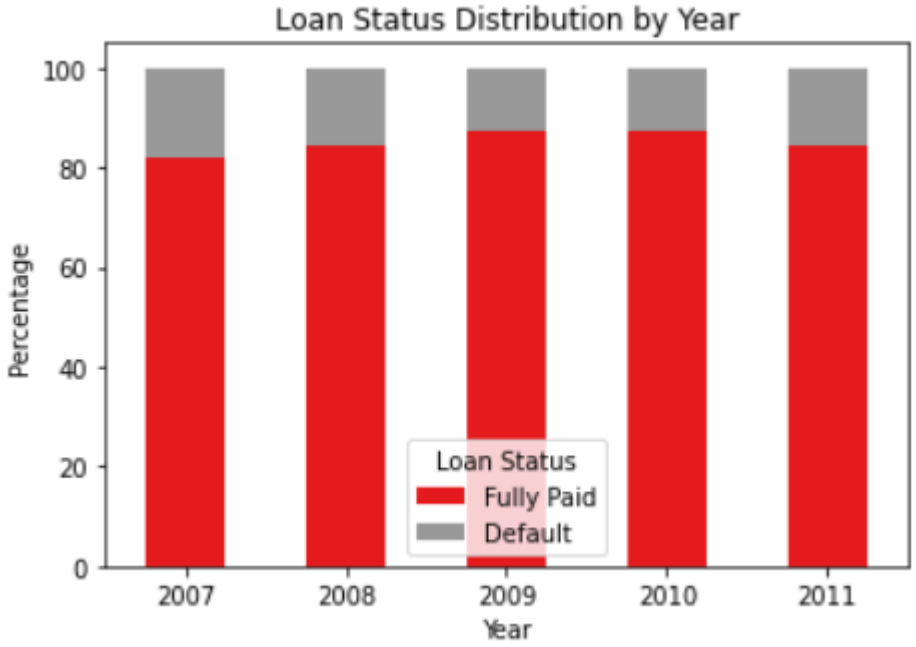
Conclusion: 1) When the loan amount increase, higher is the chances to get defaulted.
2)Max people who fully paid loan has lesser loan amount range. Both IQR box sizes and upper range is different, it indicates a difference in the variability (spread) of loan amount.

Conclusion: Most of the people are paying fully or defaulting for loan amount category 'Medium'.

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Year

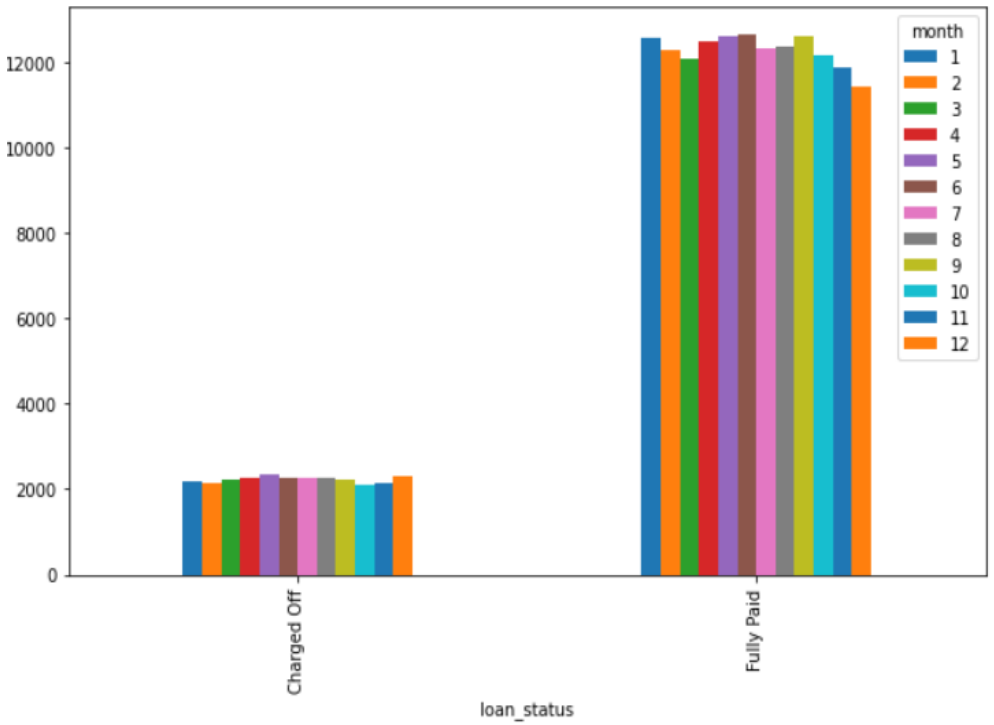


Conclusion: The defaulter rate had suddenly dipped in 2011 inspite of hike in 2009 & 2010

Bivariate Analysis Results of Continuous Variables with Target variable

We have done analysis along with segmentation and details are given below:

Loan status vs Month(Issue Date)



Conclusion: Defaulter rate is not much varying across months of the year