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. Segmentation has been done on both continuous and categorical variables
2. We divided the continuous dataset into segments based on value range – High,low,etc
3. Categorical variables are segmented based on purpose of loan

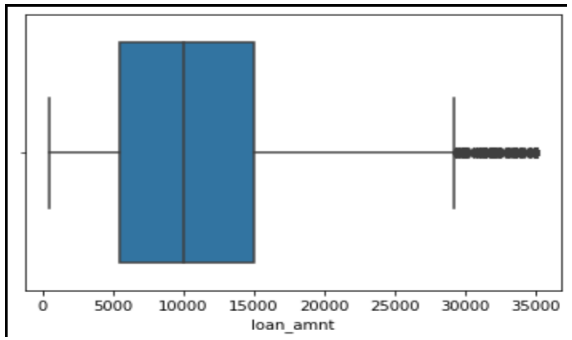
Bivariate Analysis

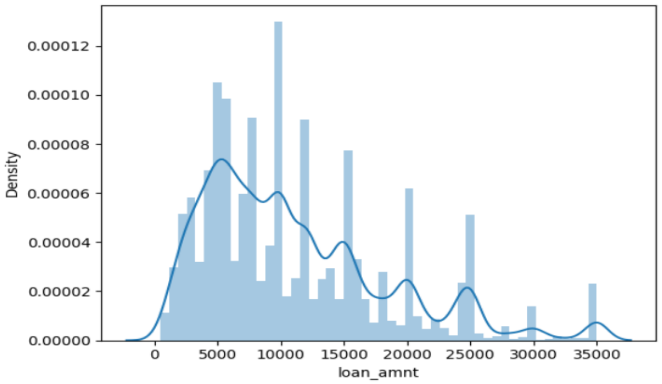
1. Here we did visualisations that show the relationship between two variables like barplots ,countplots,bar charts etc.
2. We found some results on correlation which measures strength of liner relationship
3. To observe trends and patterns to understand how changes in one variable correspond to changes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

Univariate Analysis Results – Continuous Variables

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

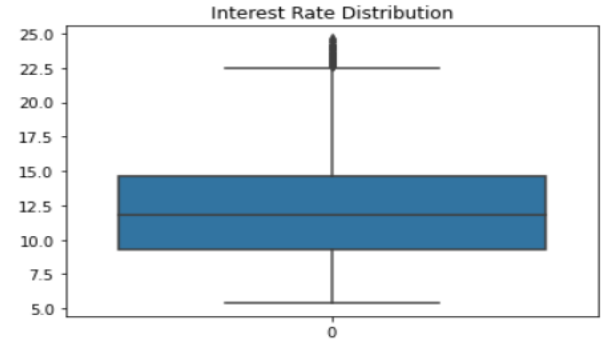
Loan amount

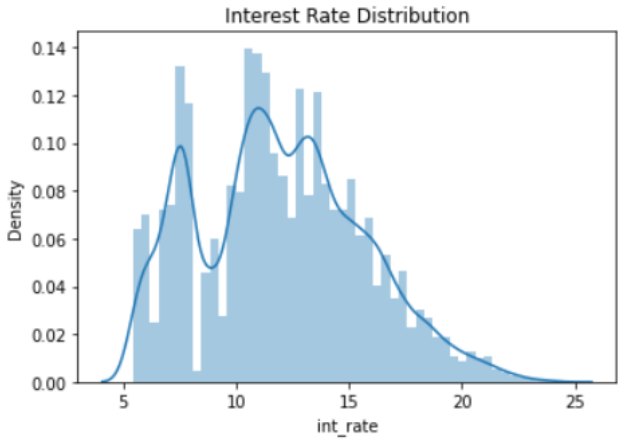




Observation : Majority of loans are between range 5000-15000. There is a pattern observed in amount which can be segmented for further analysis

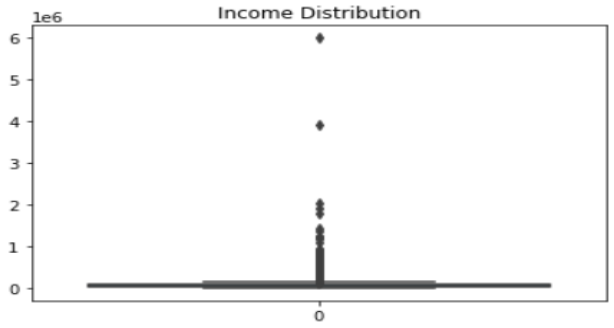
Interest rate

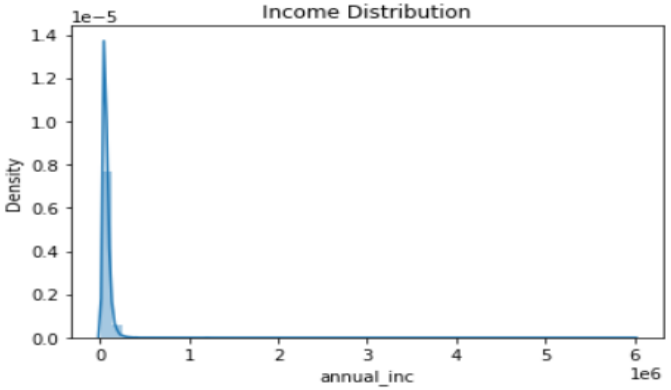




Observation : Maximum of people have loans with interest rate >8 and less than 15. Possibility of 2 segments of loans with interest rate

Annual Income



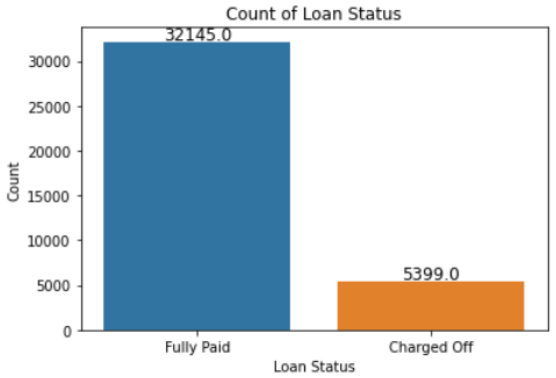


Observation : Annual Income is right skewed and it has outliers

Univariate Analysis Results – Categorical Variables

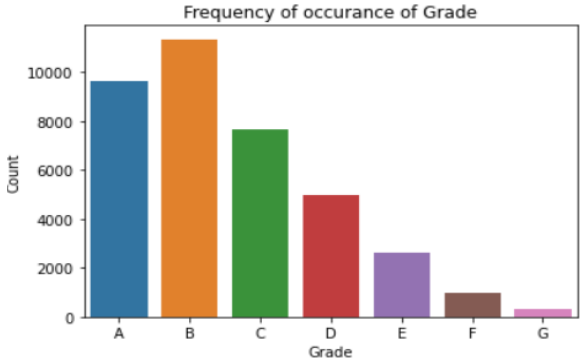
Plots and conclusion derived out of our analysis on variables are given below:

Loan Status



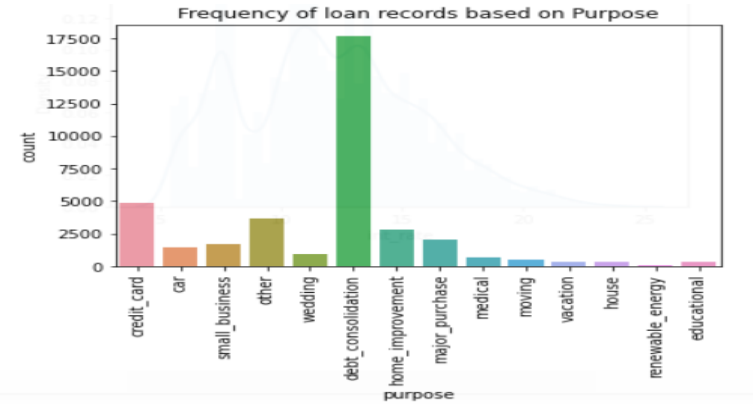
Observation : Source dataset has 5399 defaulter records.

Grade



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

Purpose

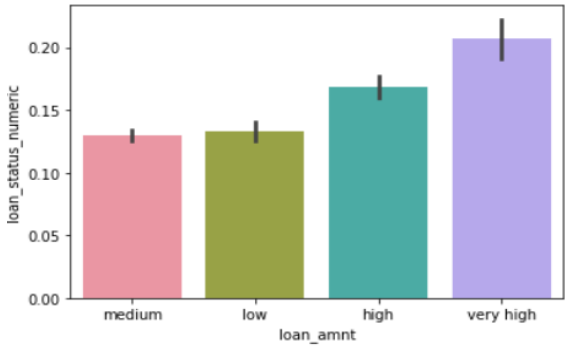


Observation : Top 4 types of loans are: consolidation, credit card, home improvement and major purchase.

Analysis -Segmented

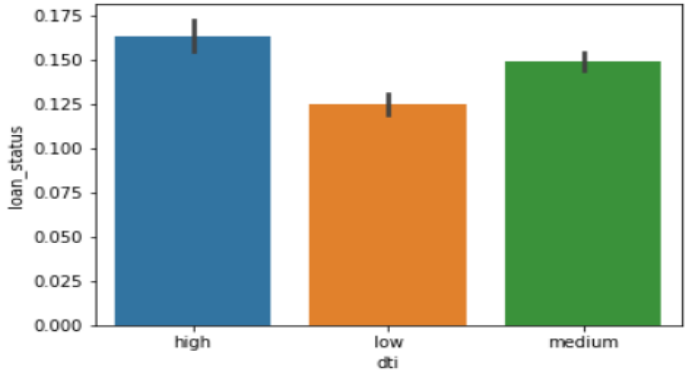
Plots and conclusion derived out of our analysis on variables are given below:

Loan amount



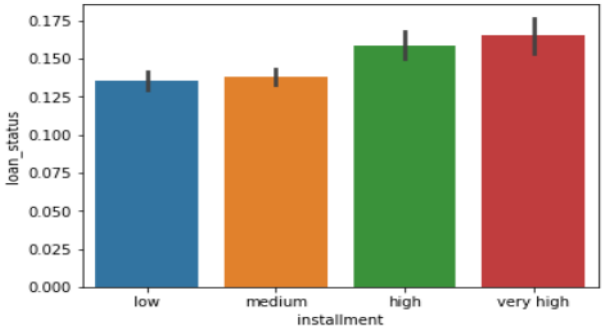
Observation : higher the loan amount, higher the default rate

DTI



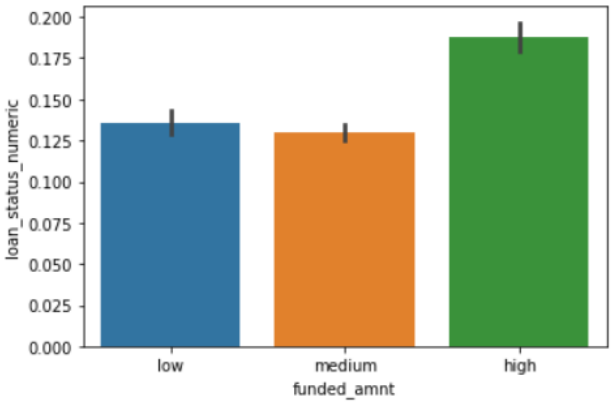
Observation : high dti translates into higher default rates

Installment



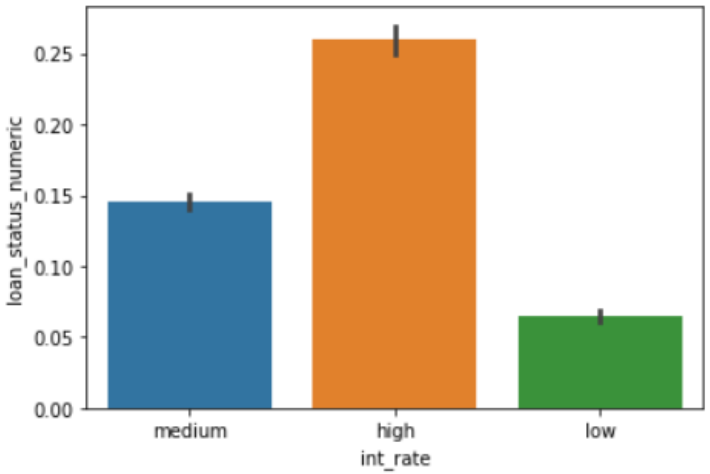
Observation : the higher the installment amount, the higher the default rate

Funded amount



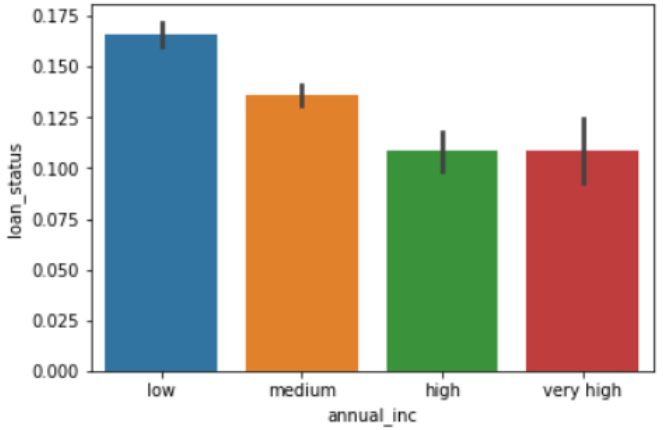
Observation : higher the funded amount, higher the default rate

Interest rate



Observation : high interest rates default more, as expected

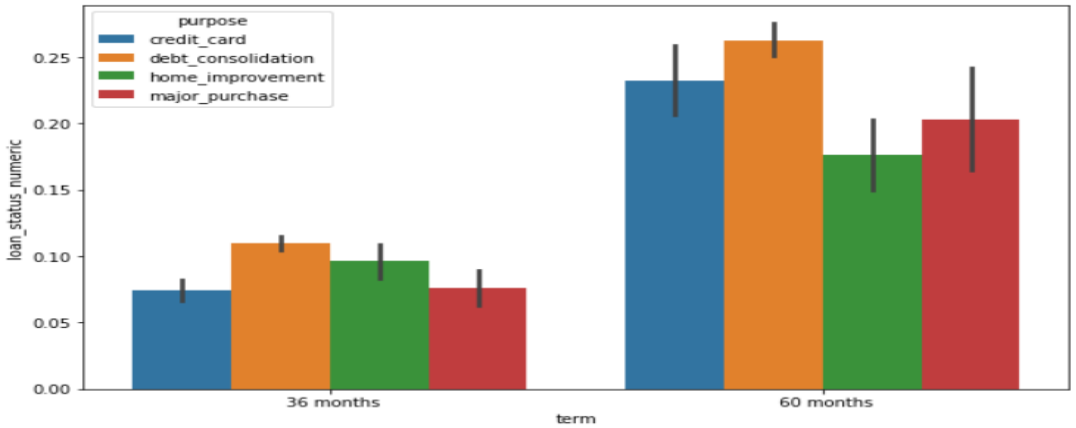
Annual Income



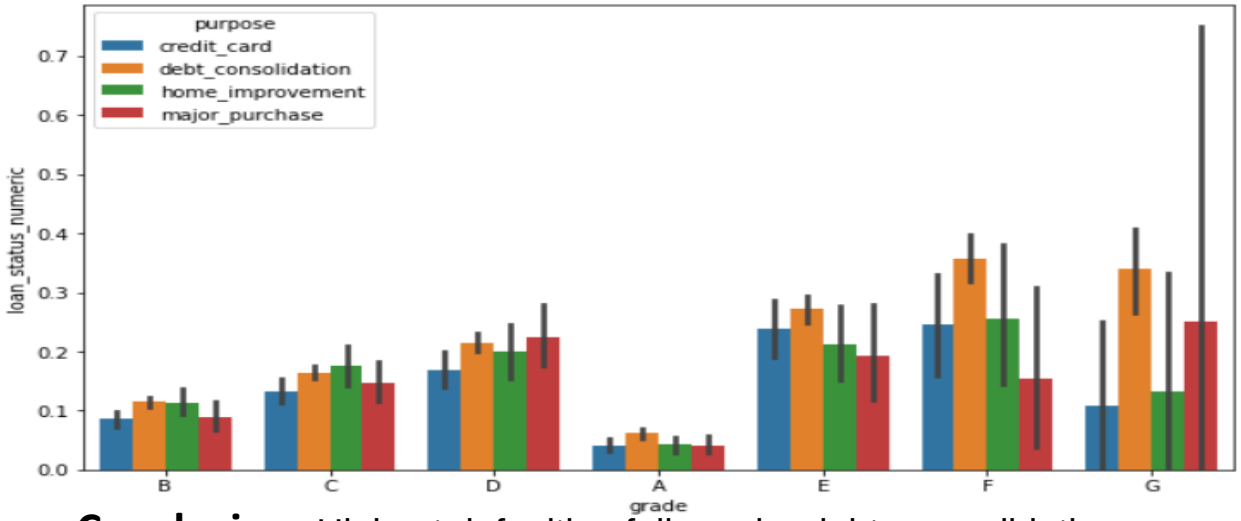
Observation :lower the annual income, higher the default rate

Analysis -Segmented

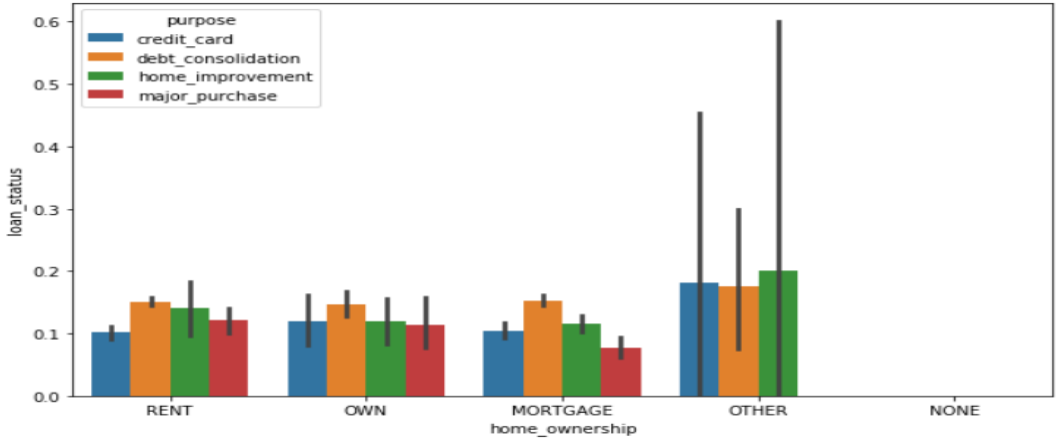
Plots and conclusion derived out of our analysis on variables are given below:



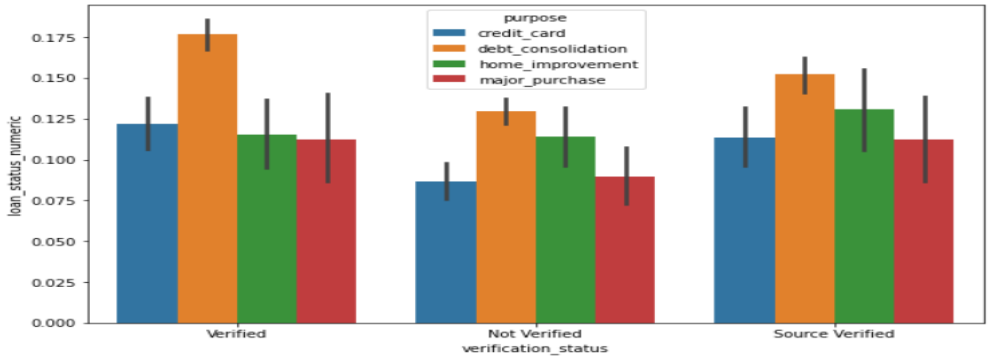
Conclusion: Highest defaulting falls under debt_consolidation category irrespective of term period



Conclusion: Highest defaulting falls under debt_consolidation category other than B and C



Conclusion: In general, debt consolidation loans have the highest default rates. We can ignore other category



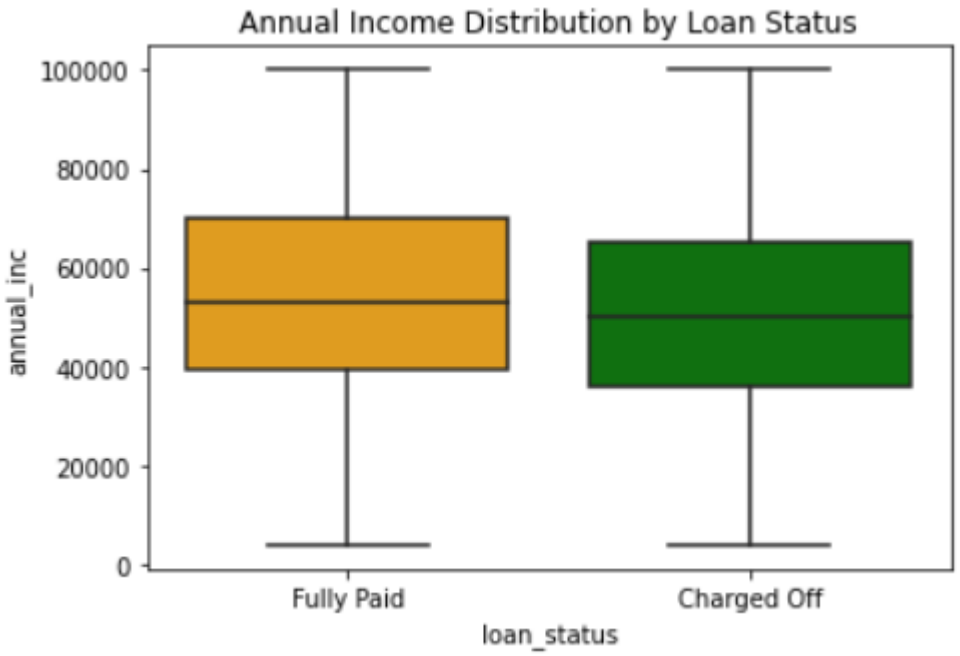
Conclusion: In general, debt consolidation loans have the highest default rates.

Bivariate Analysis Results of Continuous Variables with Target variable

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

Hypothesis: Borrower's annual income could influence their ability to repay the loan.

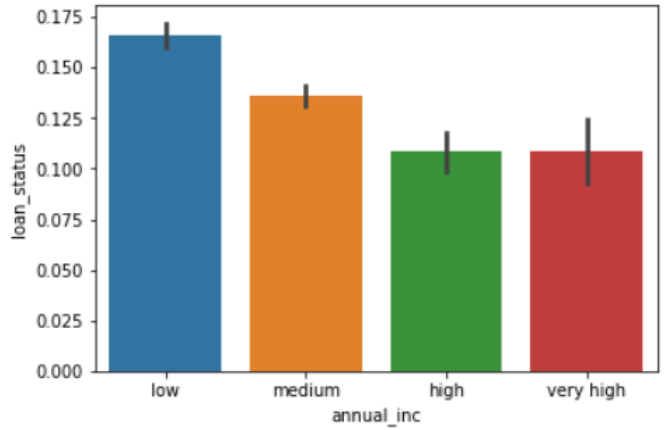
Loan status vs Annual Income



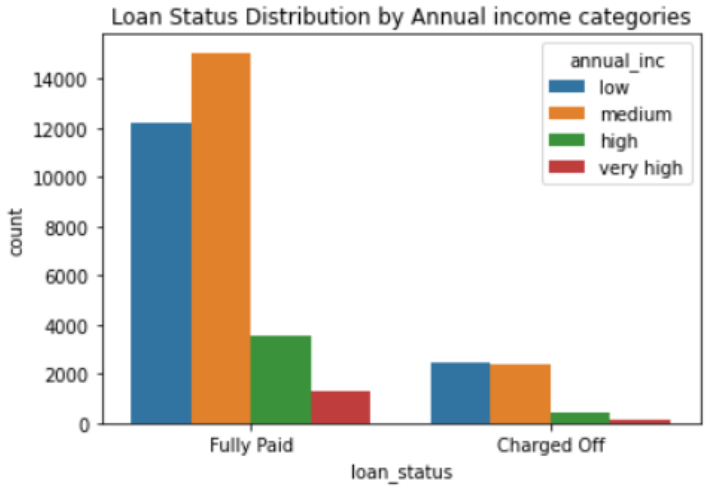
Conclusion: Higher annual income might slightly reduce the likelihood of default.

Analysis -Segmented

Categorized to bins: low, medium, high, very high



Observation :lower the annual income, higher the default rate



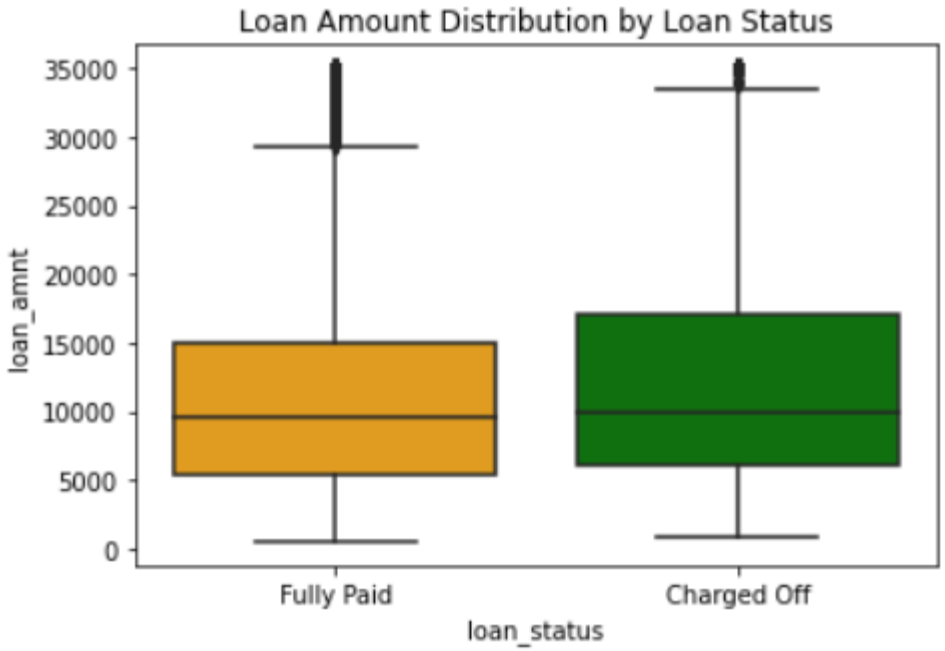
Observation: Highest frequency of defaulting has been identified for category ‘low '(Range: <= 50000) and 'Medium'(Range: 50000 -100000)

Bivariate Analysis Results of Continuous Variables with Target variable

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

Hypothesis: The amount of the loan could be a strong indicator.

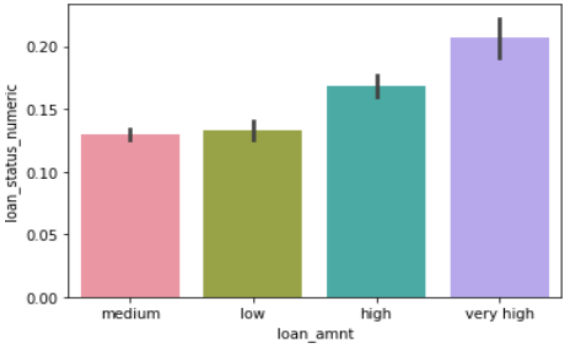
Loan status vs loan amount



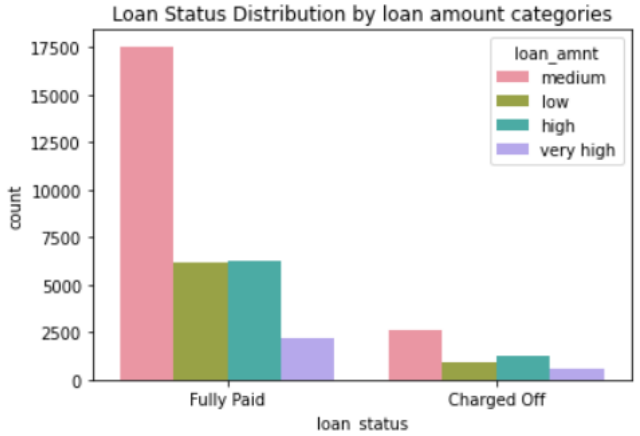
Conclusion: Higher loan amounts might slightly increase the chance of default, but the effect is minimal.
Max people who defaulted loan has relatively higher loan amount range. Both IQR box sizes and upper range is different, it indicates a difference in the variability (spread) of loan amount.

Analysis -Segmented

Categorized to bins: low, medium, high, very high



Observation : higher the loan amount, higher the default rate

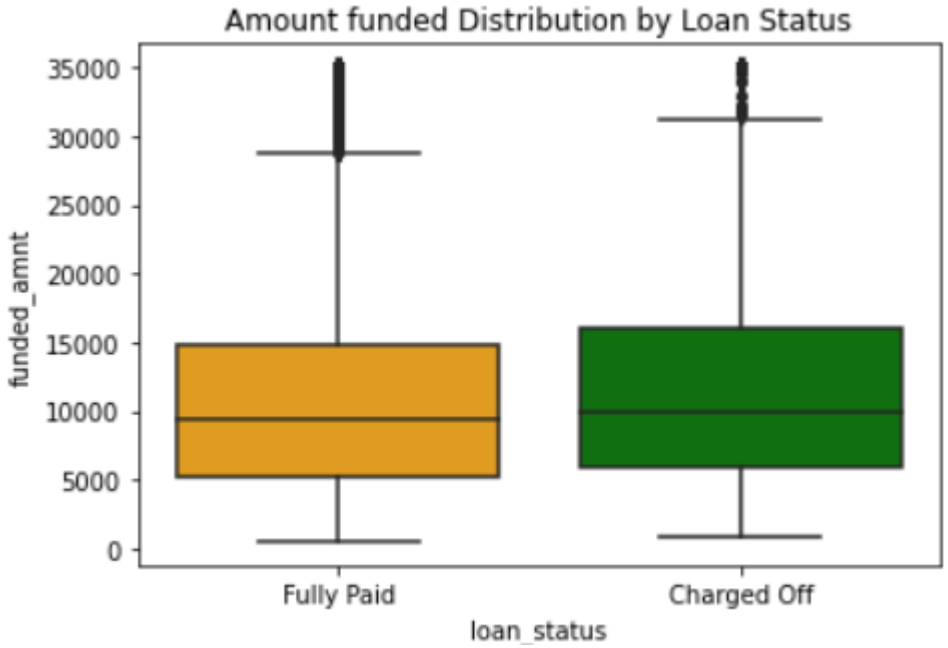


Conclusion: Most frequency of defaulting has been identified for loan amount category 'Medium'.

Bivariate Analysis Results of Continuous Variables with Target variable

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

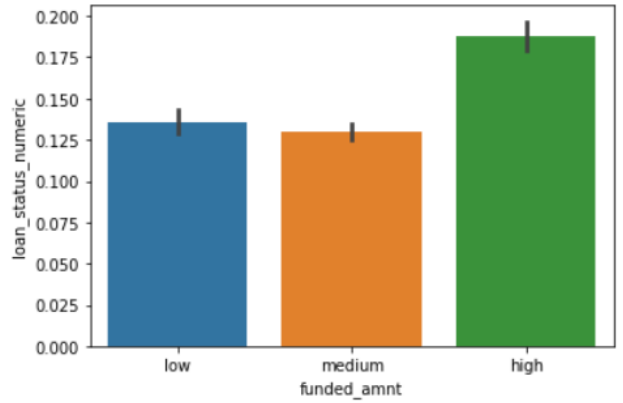
Loan status vs funded_amnt



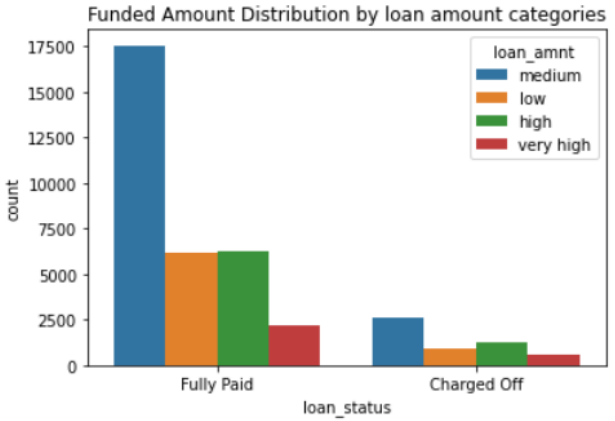
Conclusion: Like with loan_amnt, higher funded amounts might have a slight impact on default rates.
Defaulter's upper amount range increased compared to fully paid customers.

Analysis -Segmented

Categorized to bins: low, medium, high, very high



Observation : higher the funded amount, higher the default rate



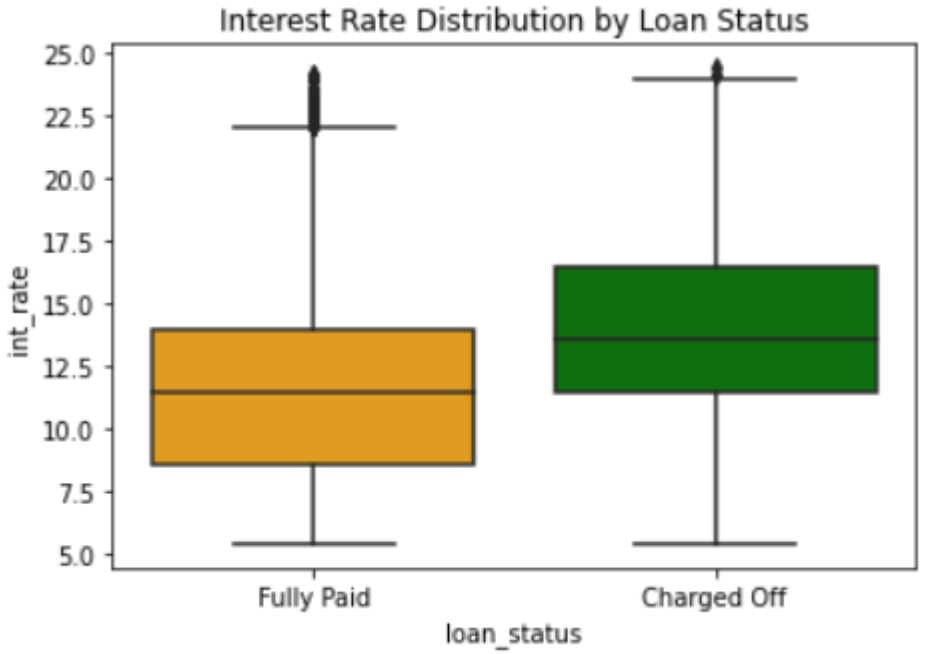
Conclusion: Highest frequency of defaulting has been identified for category 'Medium'(Range: 5000 - 15000)

Bivariate Analysis Results of Continuous Variables with Target variable

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

Hypothesis: Higher interest rates might lead to a higher likelihood of default.

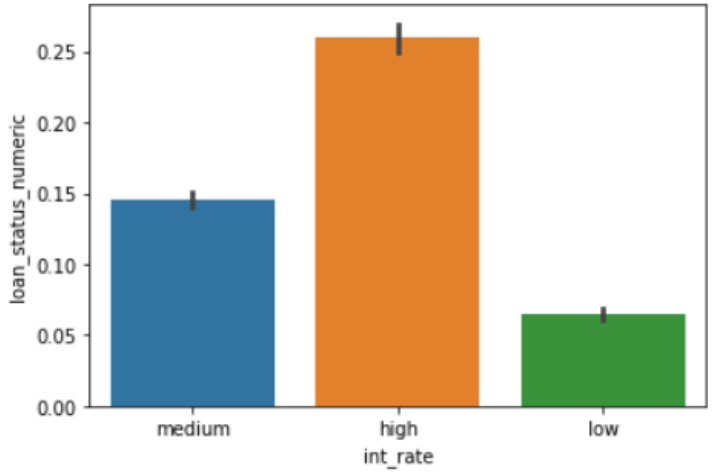
Loan status vs Interest rate



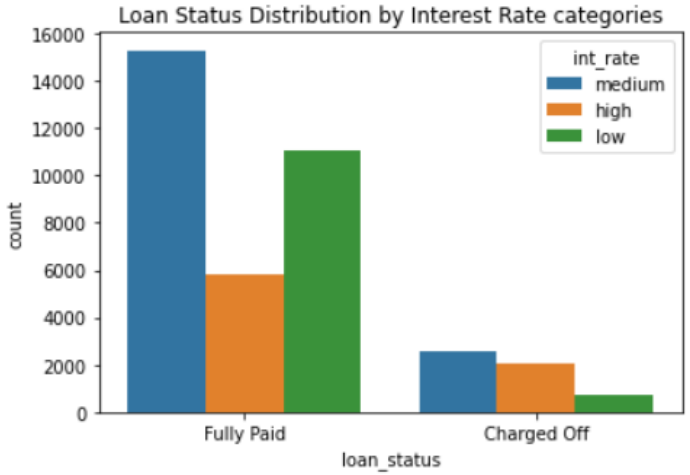
Conclusion: Higher interest rates might lead to a higher likelihood of default.
Majority of the Defaulter's has higher interest rate compared to fully paid customers. Upper 25% of defaulters has higher interest rate than fully paid customer's interest rate

Analysis -Segmented

Categorized to bins: low, medium, high,



Observation : high interest rates default more, as expected



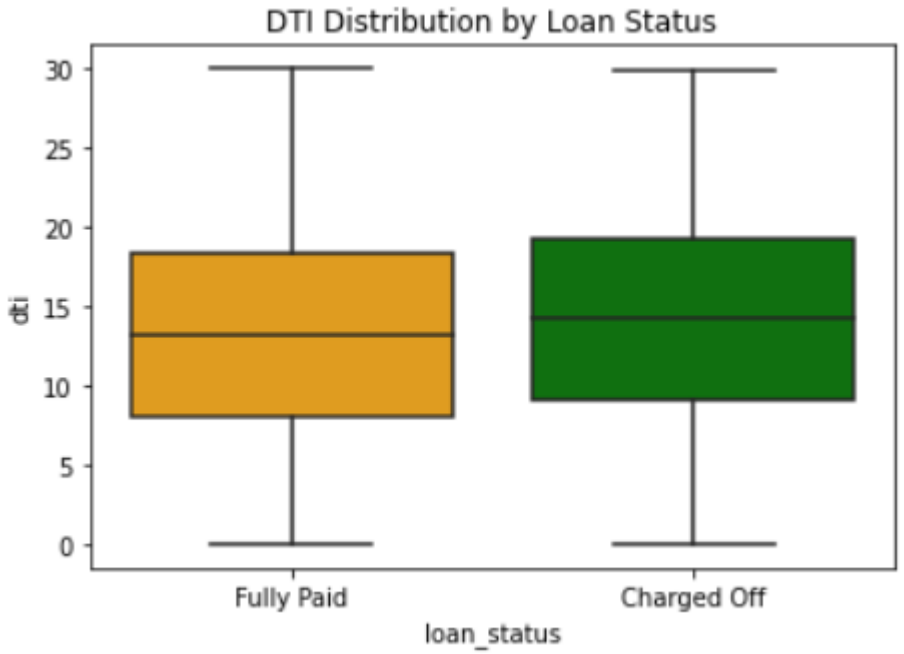
Observation: Highest frequency of defaulting has been identified for category 'Medium'(Range: 10-15)

Bivariate Analysis Results of Continuous Variables with Target variable

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

Hypothesis: Higher DTI ratios could be associated with higher default rates.

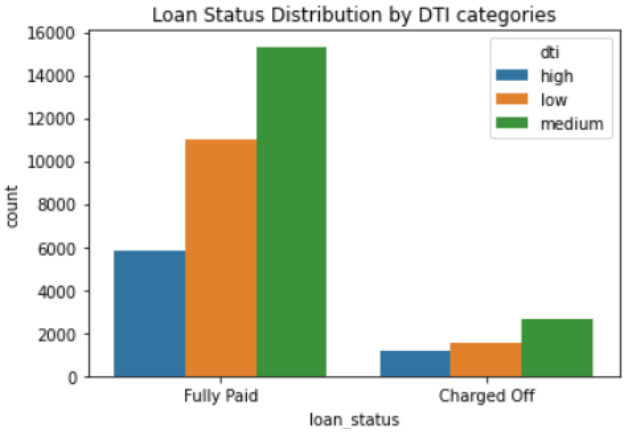
Loan status vs DTI

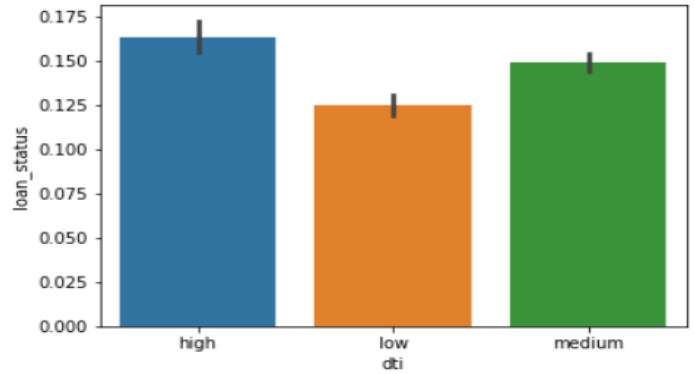


Conclusion: Higher DTI ratios might have a small impact on default rates.

Analysis -Segmented

Categorized to bins: low, medium, high,





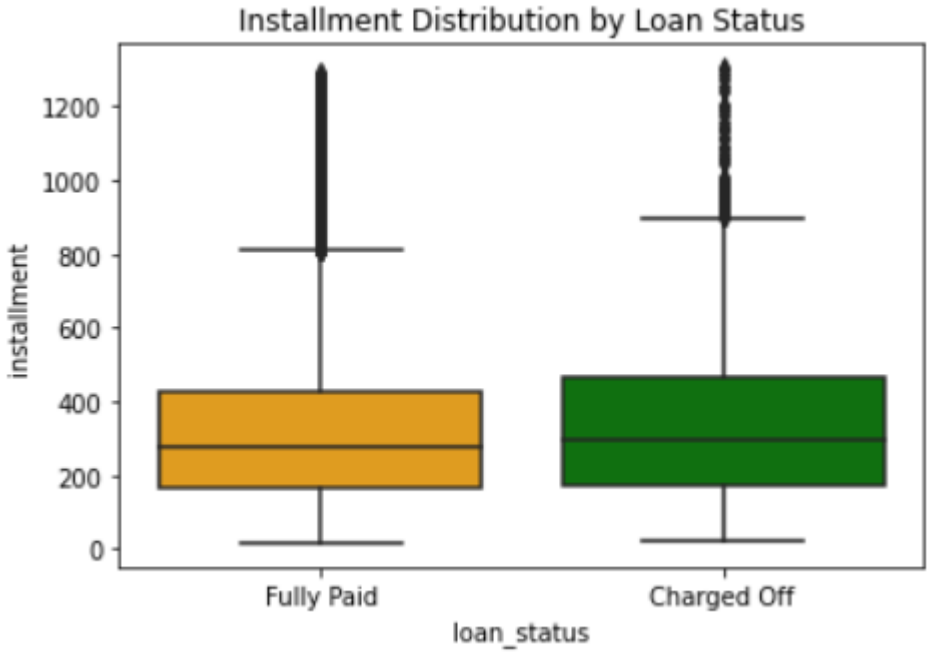
Observation : high dti translates into higher default rates

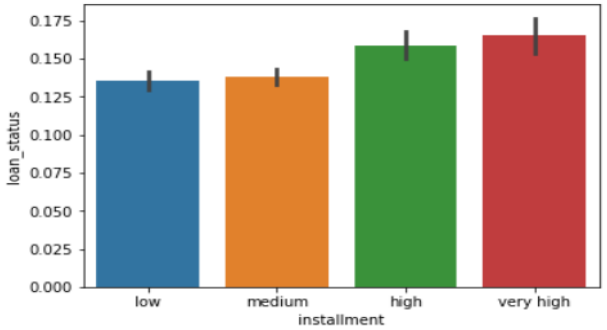
Observation: Highest frequency of defaulting has been identified for category 'Medium'(Range: 10-20)

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

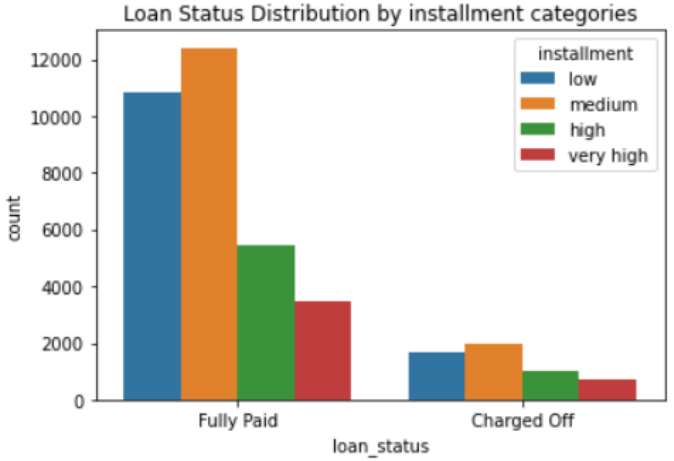




Analysis -Segmented

Categorized to bins: low, medium, high, very high

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



Observation: Highest frequency of defaulting has been identified for category 'Medium'(Range: 200-400)

Conclusion:

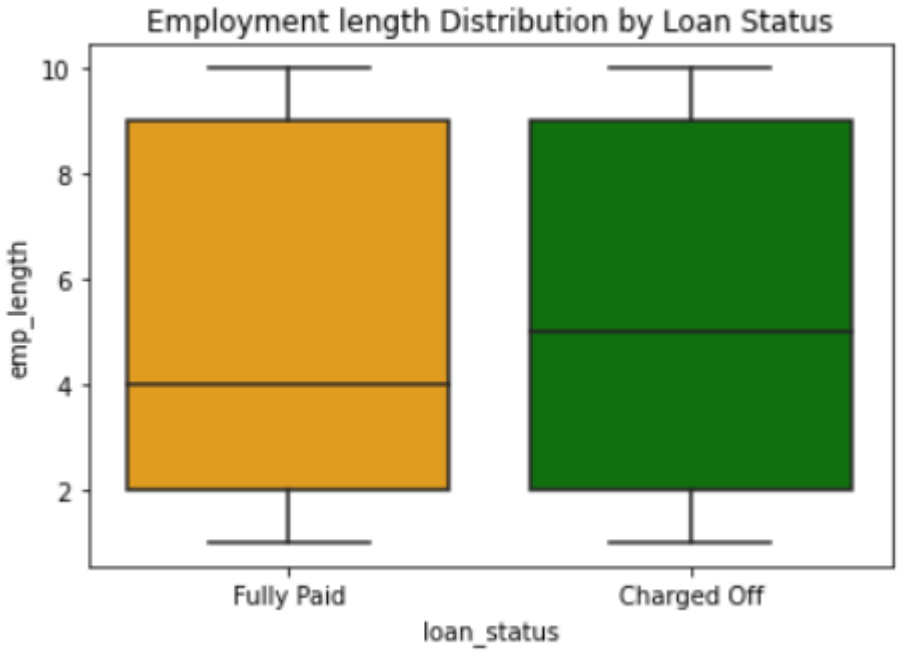
The impact of installment on default rates seems to be minimal.
1) Both IQR box size difference is relatively small, it indicates less difference in the variability (spread) of 'installment' amounts between the two loan statuses.
2) Defaulter's upper installment range increased compared to fully paid customers.

Bivariate Analysis Results of Continuous Variables with Target variable

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

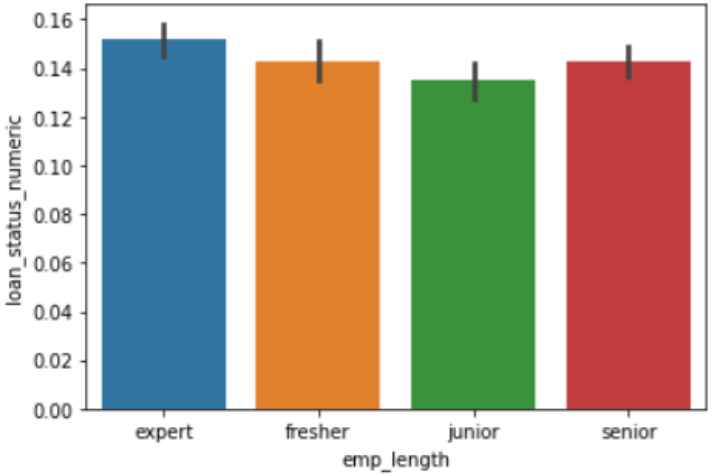
Hypothesis: The length of employment could reflect stability and the ability to repay loans. Shorter employment lengths might be associated with higher default rates.

Loan status vs Employment length

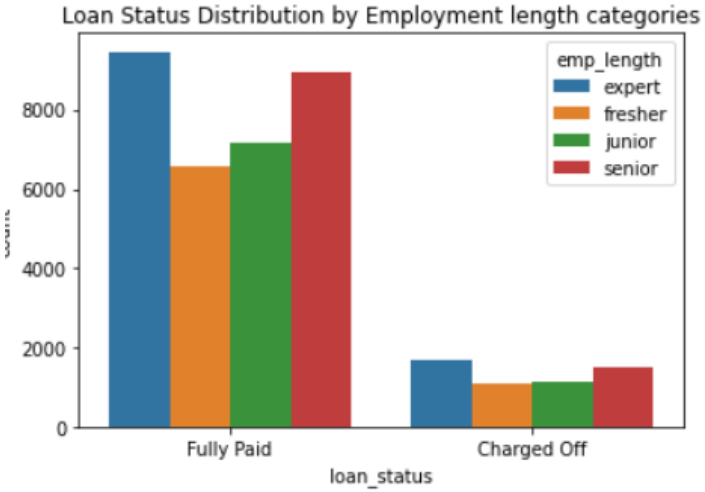


Analysis -Segmented

Categorized to bins: expert, fresher, junior, senior



Observation : Employment length is not much of a predictor of default



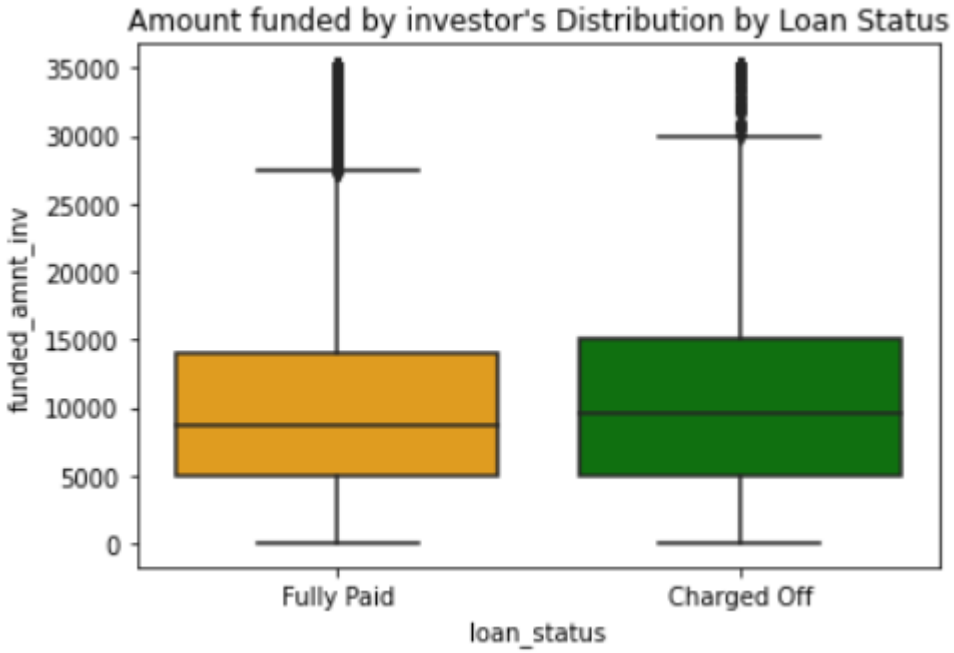
Conclusion: There seems to be a minimal impact of employment length on default rates.
Median of Defaulter's employment length is higher than fully paid customers.
For fully paid customers we have more values between 4 and 9 , where as defauters's employment length are kind of equally distributed.

Observation: Highest frequency of defaulting has been identified for expert category(>7 years of expr) followed by seniors(3-7 years of expr)

Bivariate Analysis Results of Continuous Variables with Target variable

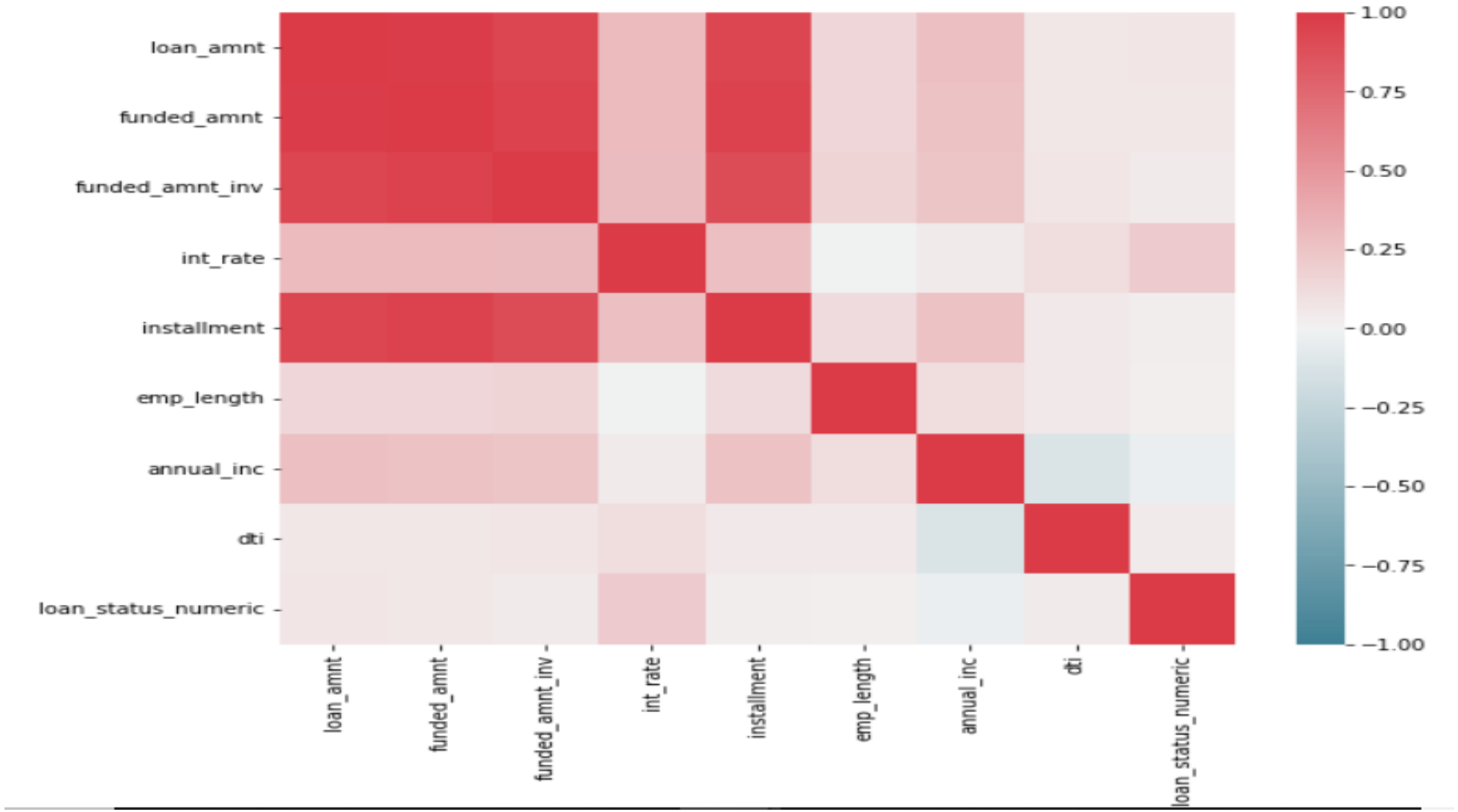
We have done analysis along with segmentation and details are given below:

Loan status vs funded_amnt_inv



Conclusion: The impact of funded_amnt_inv on default rates appears to be minimal.Defaulter's upper amount range increased compared to fully paid customers

Correlation Metrics – Continuous Variables



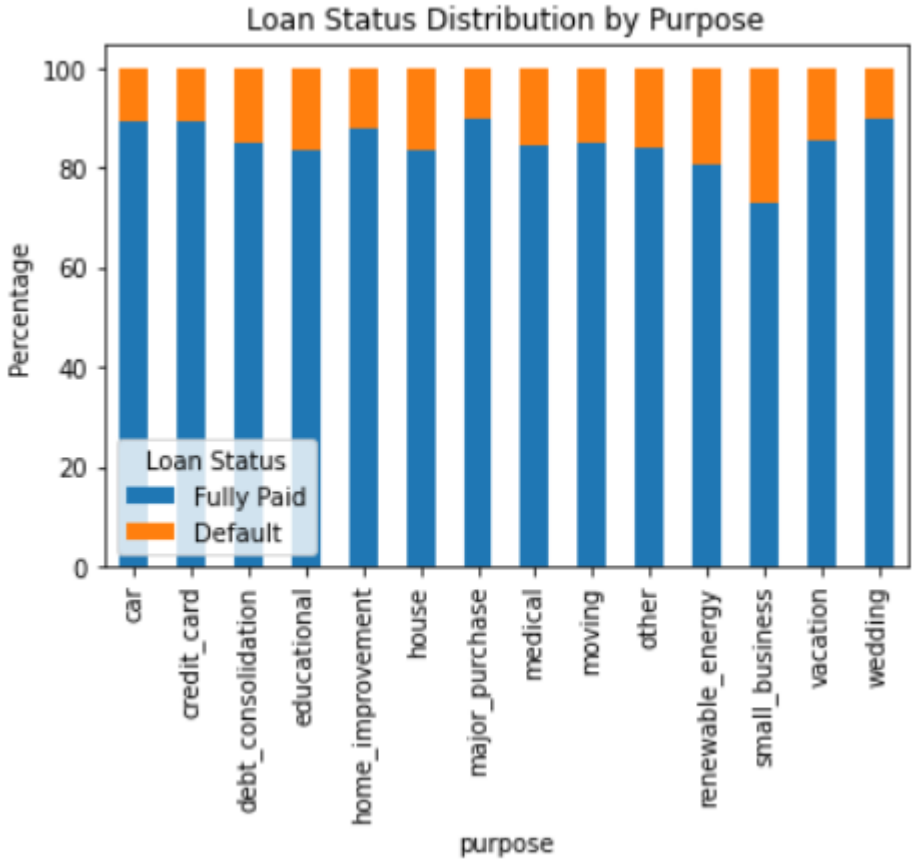
Correlation Metrics – Interpretation

1.loan_amnt: A correlation of 0.06268 suggests a very weak positive linear relationship between the loan amount and the likelihood of loan default. This could imply that higher loan amounts might slightly increase the chance of default, but the effect is minimal.
2.funded_amnt: A correlation of 0.059535 indicates a similarly very weak positive linear relationship between the funded loan amount and loan default. Like with loan_amnt, higher funded amounts might have a slight impact on default rates.
3.funded_amnt_inv: A correlation of 0.040097 suggests a very weak positive linear relationship between the funded loan amount by investors and loan default. The impact of funded_amnt_inv on default rates appears to be minimal.
4.term: A correlation of 0.1752 indicates a moderate positive linear relationship between the loan term and loan default. Longer loan terms might be associated with a higher likelihood of default.
5.int_rate: A correlation of 0.213497 suggests a moderate positive linear relationship between the interest rate and loan default. Higher interest rates might be associated with a higher chance of default.
6.installment: A correlation of 0.029868 indicates a very weak positive linear relationship between the installment amount and loan default. The impact of installment on default rates seems to be minimal.
7.emp_length: A correlation of 0.016656 suggests a very weak positive linear relationship between employment length and loan default. There seems to be a minimal impact of employment length on default rates.
8.annual_inc: A correlation of -0.038501 suggests a very weak negative linear relationship between annual income and loan default. However, this negative correlation is also weak, implying that higher annual income might slightly reduce the likelihood of default.
9.dti: A correlation of 0.042803 indicates a very weak positive linear relationship between debt-to-income ratio (DTI) and loan default. Higher DTI ratios might have a small impact on default rates.

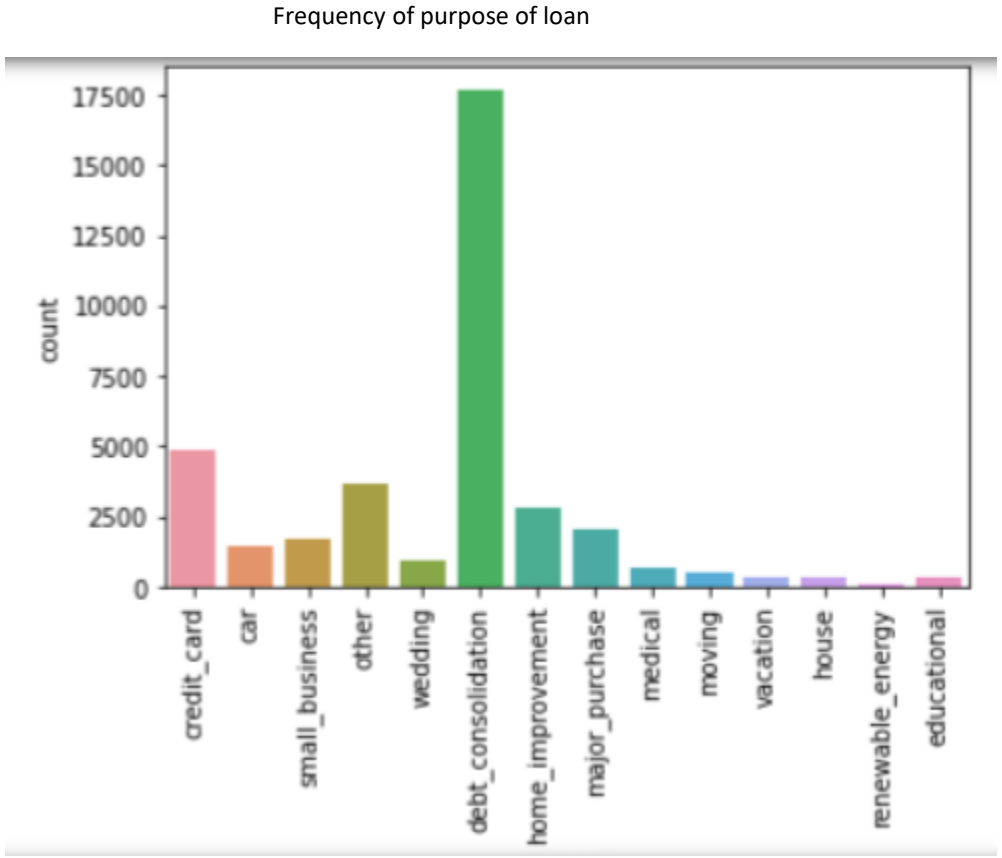
Bivariate Analysis Results of Categorical Variables with Target variable

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



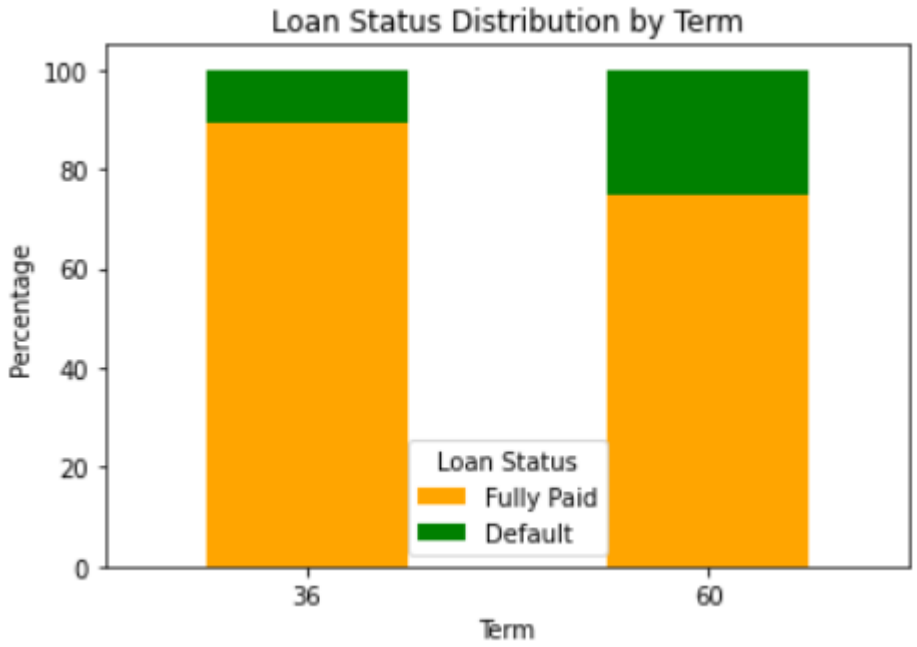
Conclusion: Top 4 types of loan based on frequency is "credit_card","debt_consolidation","home_improvement","major_purchase". Since purpose of the loan is an important predictor, this can be segmented for more insights

Bivariate Analysis Results of Categorical Variables with Target variable

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

Hypothesis: Longer terms might have higher default rates.

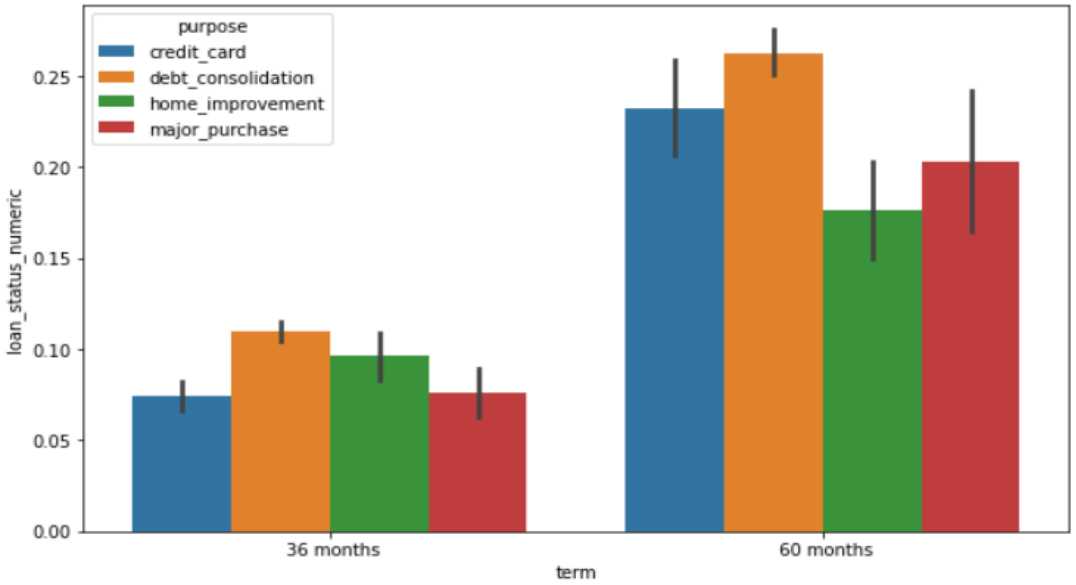
Loan status vs Term



Conclusion: Higher the Term higher the defaulting rate. 60 months loans default more than 36 months loans

Analysis -Segmented

Categorized based on top 4 types of loan

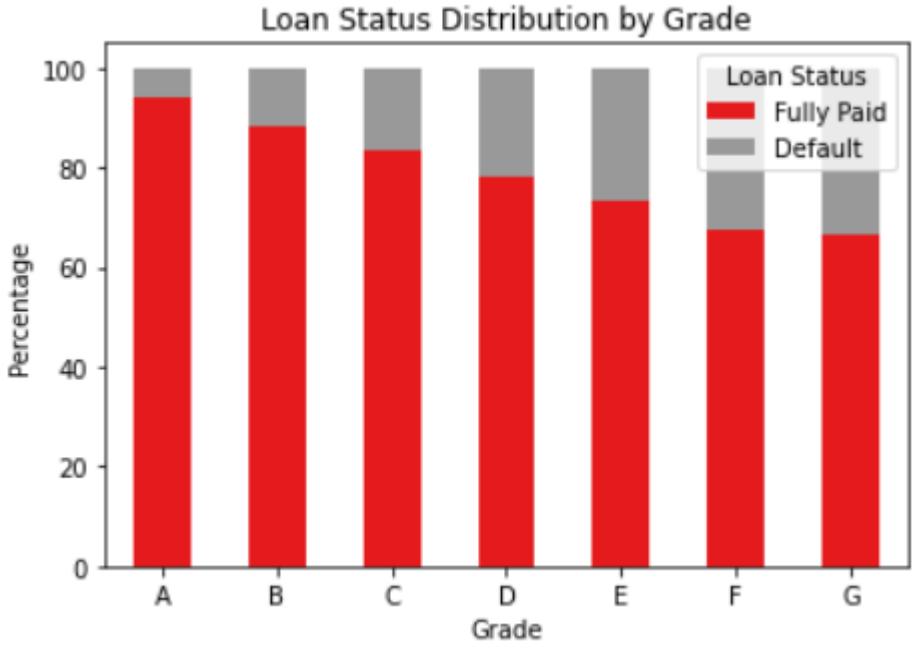


Conclusion: Highest defaulting falls under debt_consolidation category irrespective of term period

Bivariate Analysis Results of Categorical Variables with Target variable

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

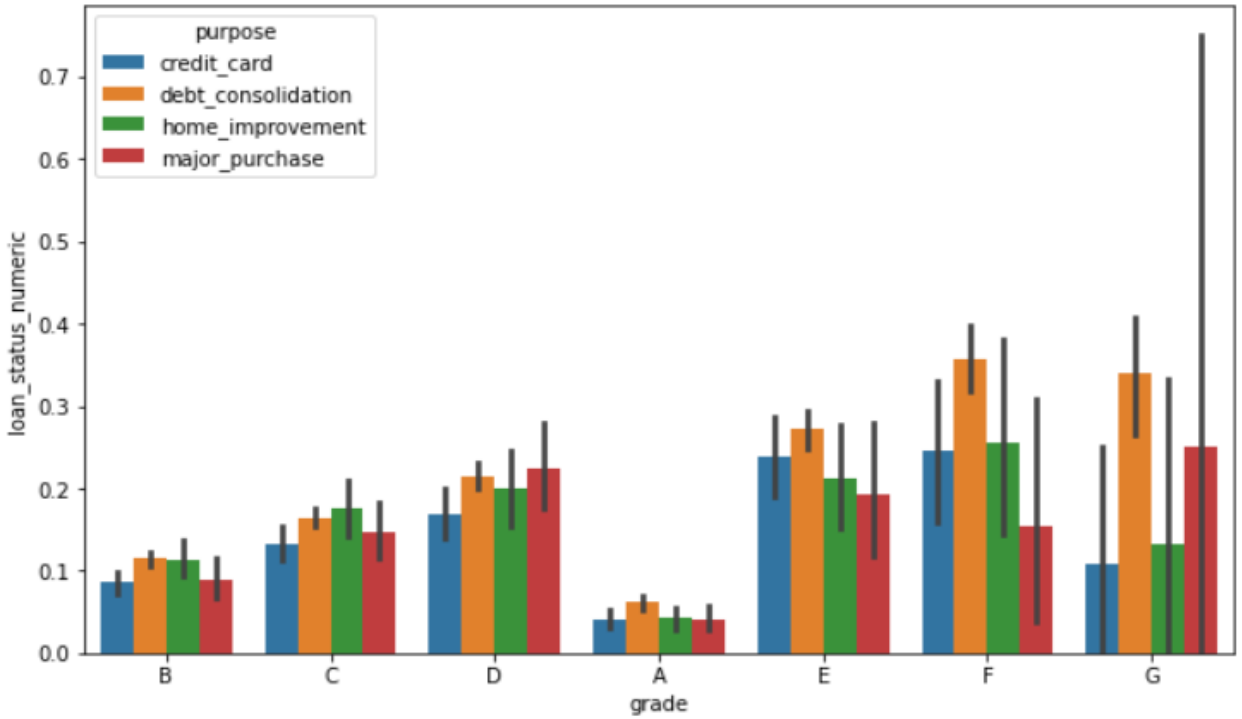
Loan status vs Grade



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.

Analysis -Segmented

Categorized based on top 4 types of loan

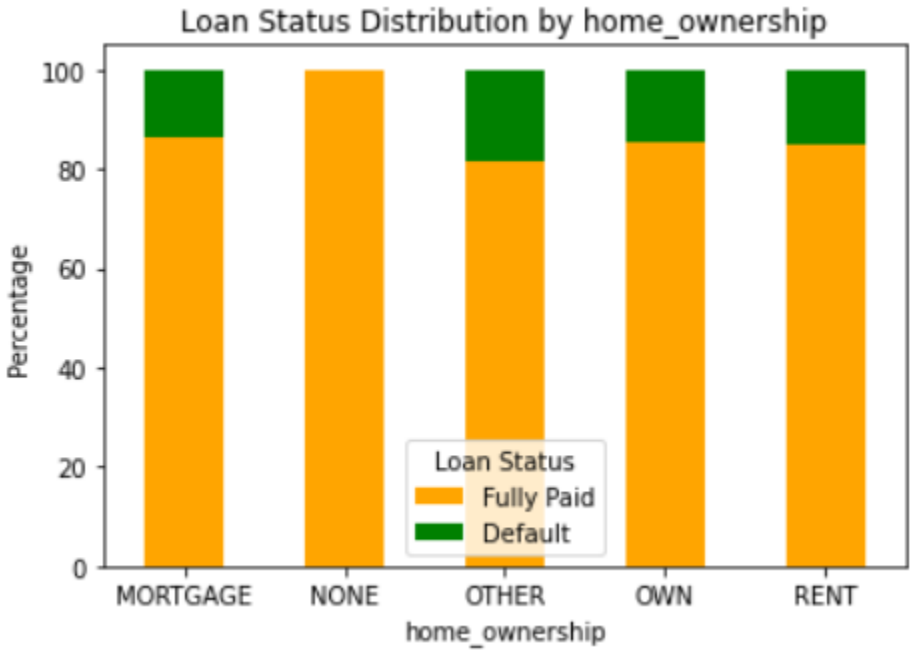


Conclusion: Highest defaulting falls under debt_consolidation category other than B and C

Bivariate Analysis Results of Categorical Variables with Target variable

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

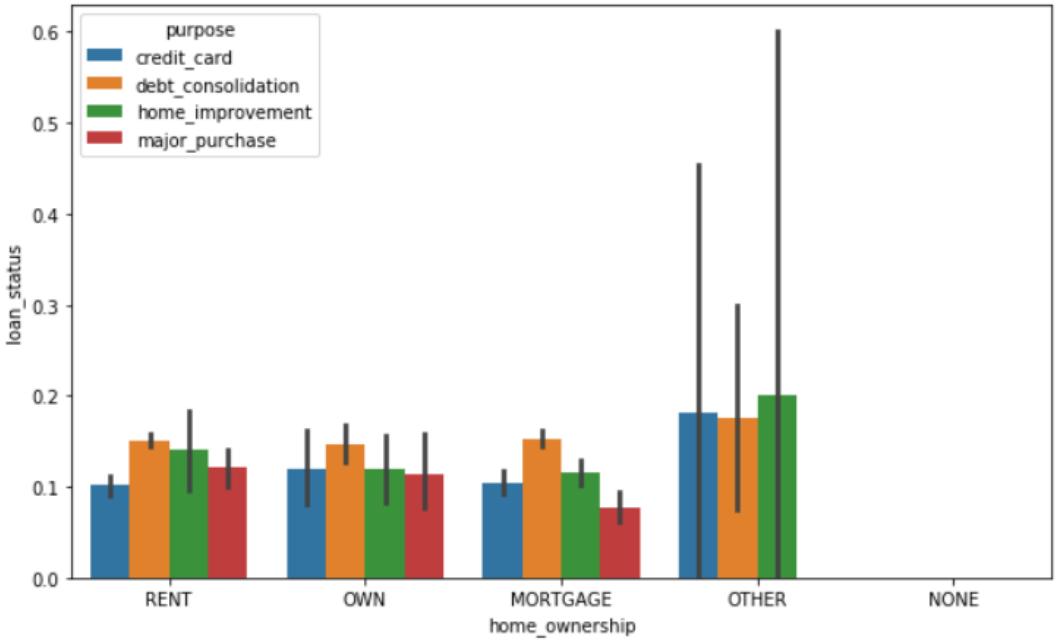
Loan status vs Home ownership



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

Analysis -Segmented

Categorized based on top 4 types of loan

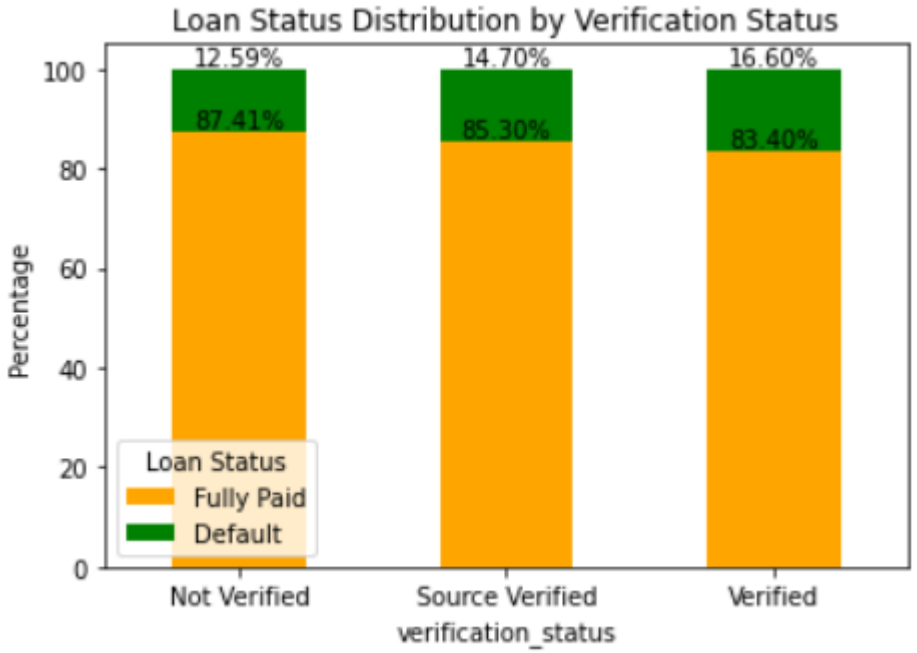


Conclusion: In general, debt consolidation loans have the highest default rates. We can ignore other category

Bivariate Analysis Results of Categorical Variables with Target variable

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

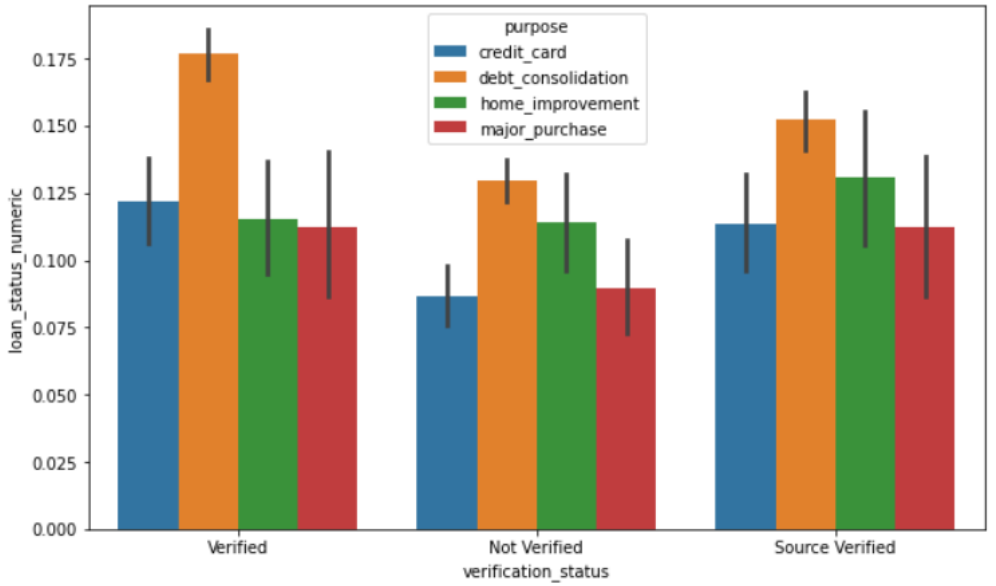
Loan status vs Verification status



Conclusion: There is no much significant difference in defaulter rate irrespective of verified or not

Analysis -Segmented

Categorized based on top 4 types of loan

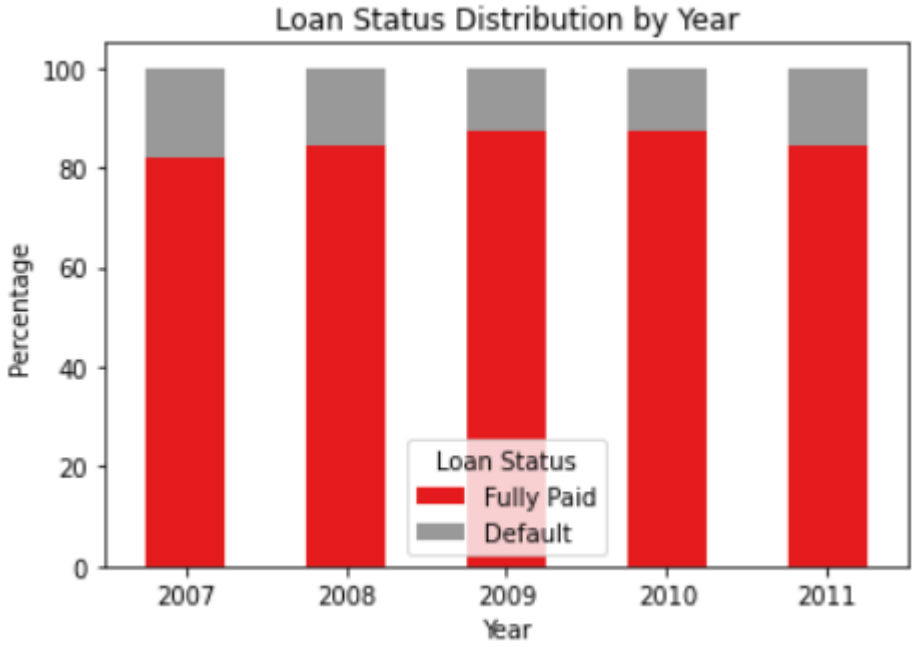


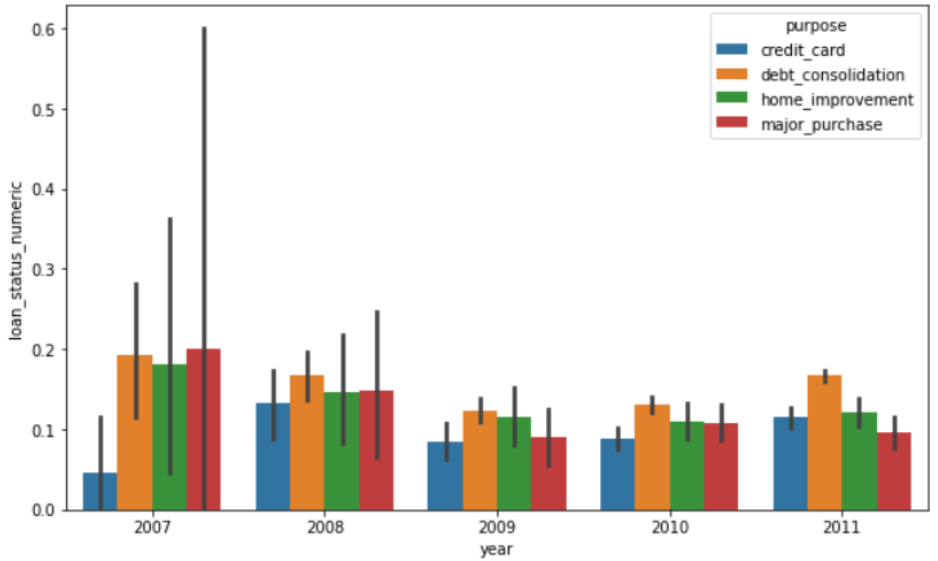
Conclusion: In general, debt consolidation loans have the highest default rates.

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





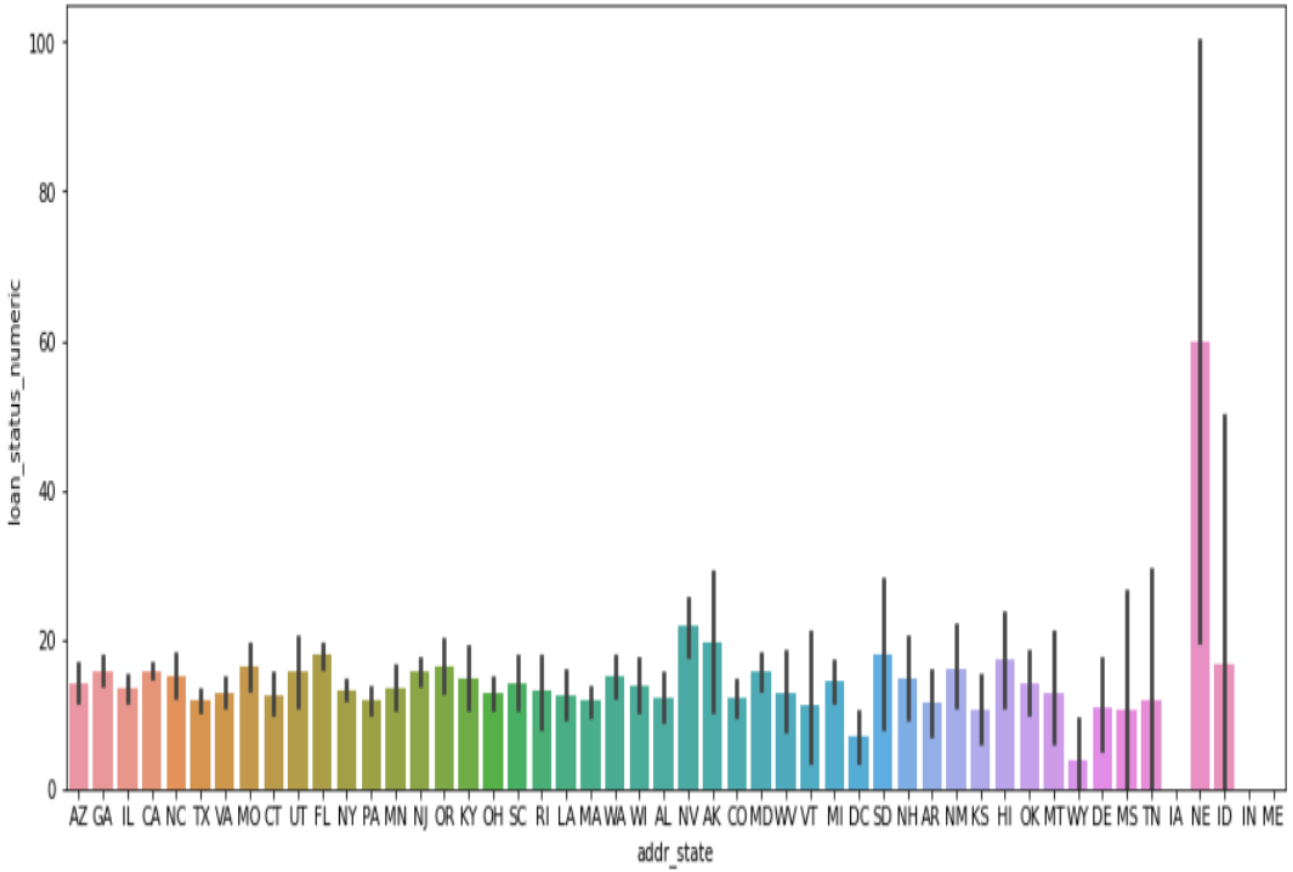
Conclusion: The defaulter rate had suddenly dipped in 2011 inspite of hike in 2009 & 2010

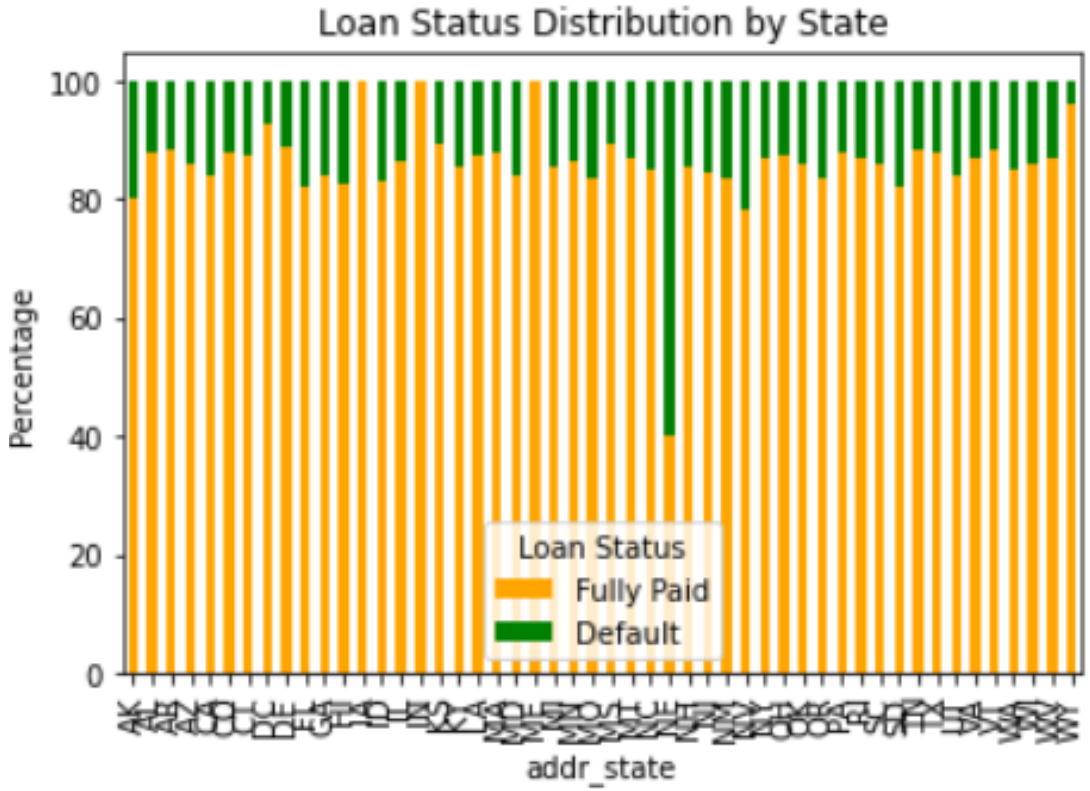
Conclusion: As we move from 2007 to 2011, debt consolidation loans have the highest default rates.

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



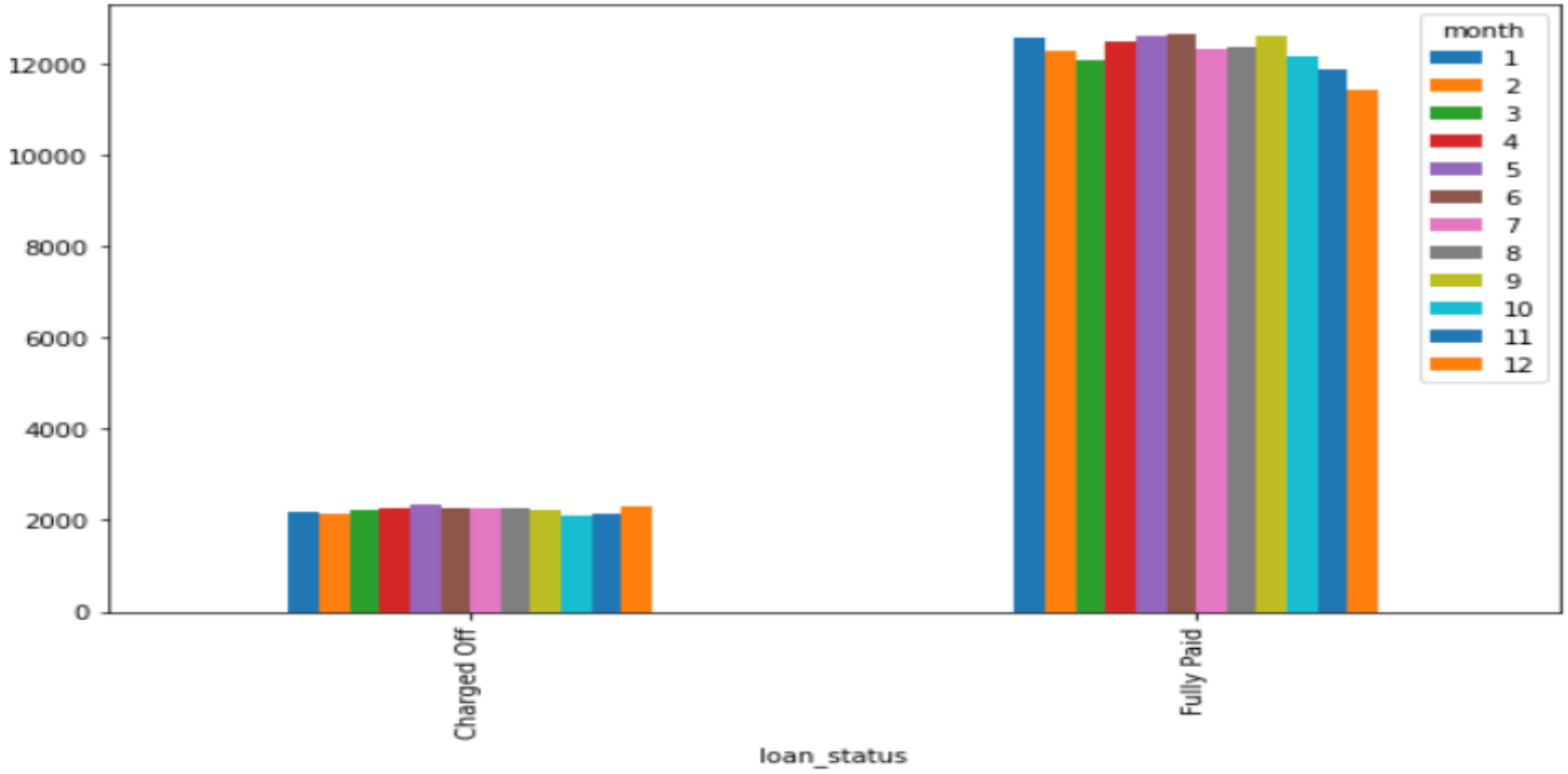


Conclusion: NE(Nebraska) state has higher defaulter rate

Bivariate Analysis Results of Continuous Variables with Target variable

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

Loan status vs Month(Issue Date)



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