EDA Case Study on Risk analysis of Bank loan

By Naveen Hosamani
Upgrad Campus

Files and Resources

- Dataset Name : loan.csv
- Dataset Format : .csv
- Analysis done in : Jupyter Notebook
- Notebook Name : loan_risk_analysis.ipynb
- Reference file : Data_dictionary.xlsx

Risk Analytics In Banking And Financial Services:

Case Study on Bank's lending process by analyzing the dataset containing the past loan applicants while identifying the patterns which indicates risk associated applicants which may default on loan.

Problem Statement:

A Consumer Finance Company that specializes in providing various types of loans to urban customers. When the company receives a loan application, it has to decide whether to approve or reject it based on the applicant’s profile.

Company experience a heavy loss if a borrower defaults on his loan. In another case if applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

Company wants to understand the driving factors (or driver variables) behind loan default, thereby cutting down the amount of credit loss.

Columns Considered for analyzing dataset:

	0	1	2	3	4
funded_amnt	5000	2500	2400	10000	3000
int_rate	10.65	15.27	15.96	13.49	12.69
installment	162.87	59.83	84.33	339.31	67.79
emp_length	10+ years	< 1 year	10+ years	10+ years	1 year
home_ownership	RENT	RENT	RENT	RENT	RENT
annual_inc	24000.0	30000.0	12252.0	49200.0	80000.0
verification_status	Verified	Source Verified	Not Verified	Source Verified	Source Verified
issue_d	Dec 2011	Dec 2011	Dec 2011	Dec 2011	Dec 2011
loan_status	Fully Paid	Charged Off	Fully Paid	Fully Paid	Current
purpose	credit_card	car	small_business	other	other
addr_state	AZ	GA	IL	CA	OR
dti	27.65	1.0	8.72	20.0	17.94
delinq_2yrs	0	0	0	0	0
open_acc	3	3	2	10	15
total_acc	9	4	10	37	38

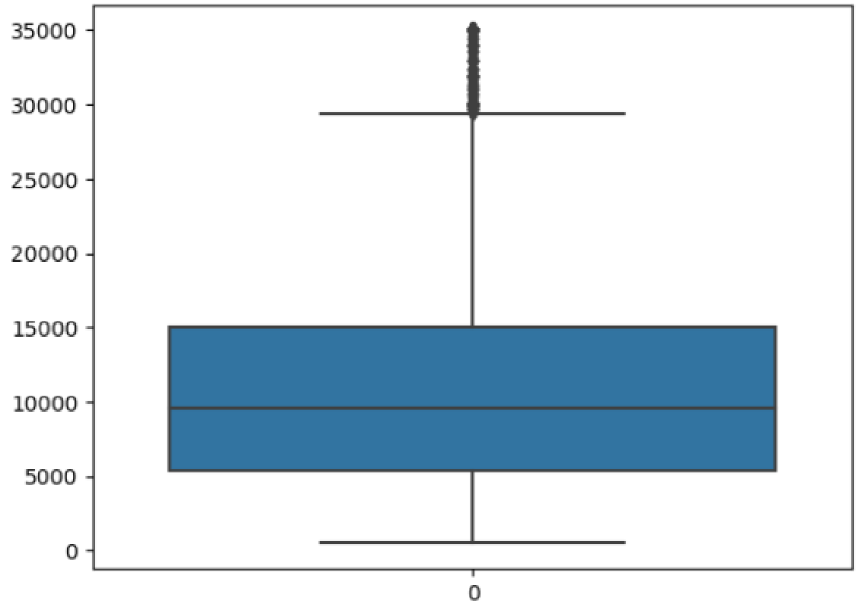
Columns Considered for analyzing dataset:

```
In [7]: df_loan.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   funded_amnt          39717 non-null  int64  
 1   term                 39717 non-null  object 
 2   int_rate             39717 non-null  object 
 3   installment          39717 non-null  float64
 4   emp_length           38642 non-null  object 
 5   home_ownership       39717 non-null  object 
 6   annual_inc           39717 non-null  float64
 7   verification_status  39717 non-null  object 
 8   issue_d              39717 non-null  object 
 9   loan_status          39717 non-null  object 
 10  purpose              39717 non-null  object 
 11  addr_state           39717 non-null  object 
 12  dti                  39717 non-null  float64
 13  delinq_2yrs          39717 non-null  int64  
 14  open_acc             39717 non-null  int64  
 15  total_acc            39717 non-null  int64  
dtypes: float64(3), int64(4), object(9)
memory usage: 4.8+ MB
```

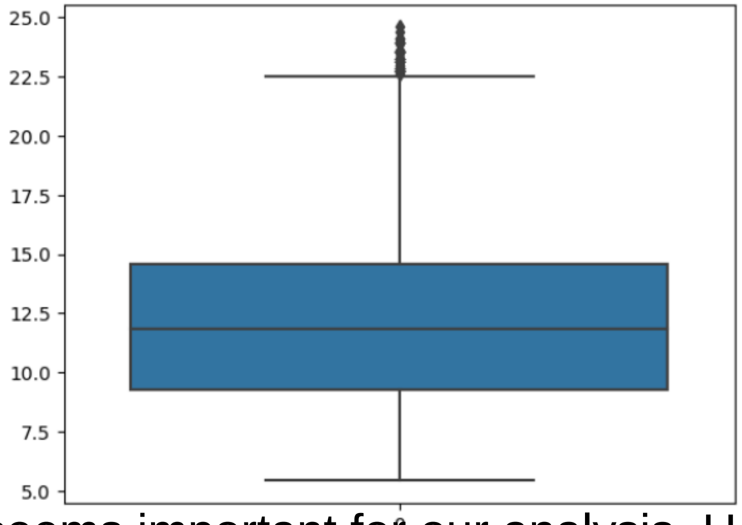
Data Cleaning:

- Converting 'term' column to 'int' datatype from 'object' datatype and renaming ‘term’ to ‘term in months’
- Convert 'int_rate' column to 'float' datatype from 'object' datatype by removing ‘%’ symbol from all rows
- Replacing null values of ‘emp_length’ column to ‘Not_Disclosed’
- Changing date format
- Analyzing outliers in bank funded amount. There are 686 borrowers who borrowed high fund(more than 30000) from bank. These outliers are important for the analysis which is useful in future hence retaining these rows is necessary.



Data Cleaning:

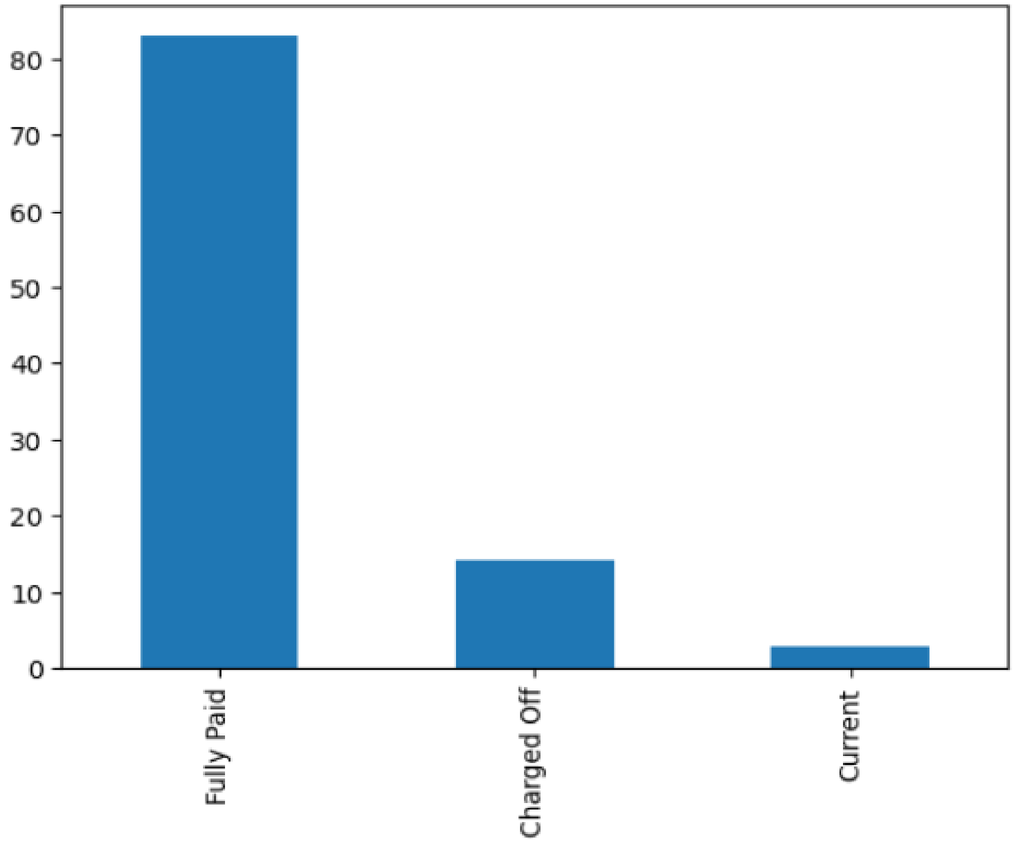
- Analyzing outliers in interest rate. There are 78 borrowers who borrowed fund with high interest(more than 22.5%) from bank. These outliers are important for the analysis which is useful in future hence retaining these rows is necessary.



- 'installment' column does not seems important for our analysis. Hence dropping the column 'installment'

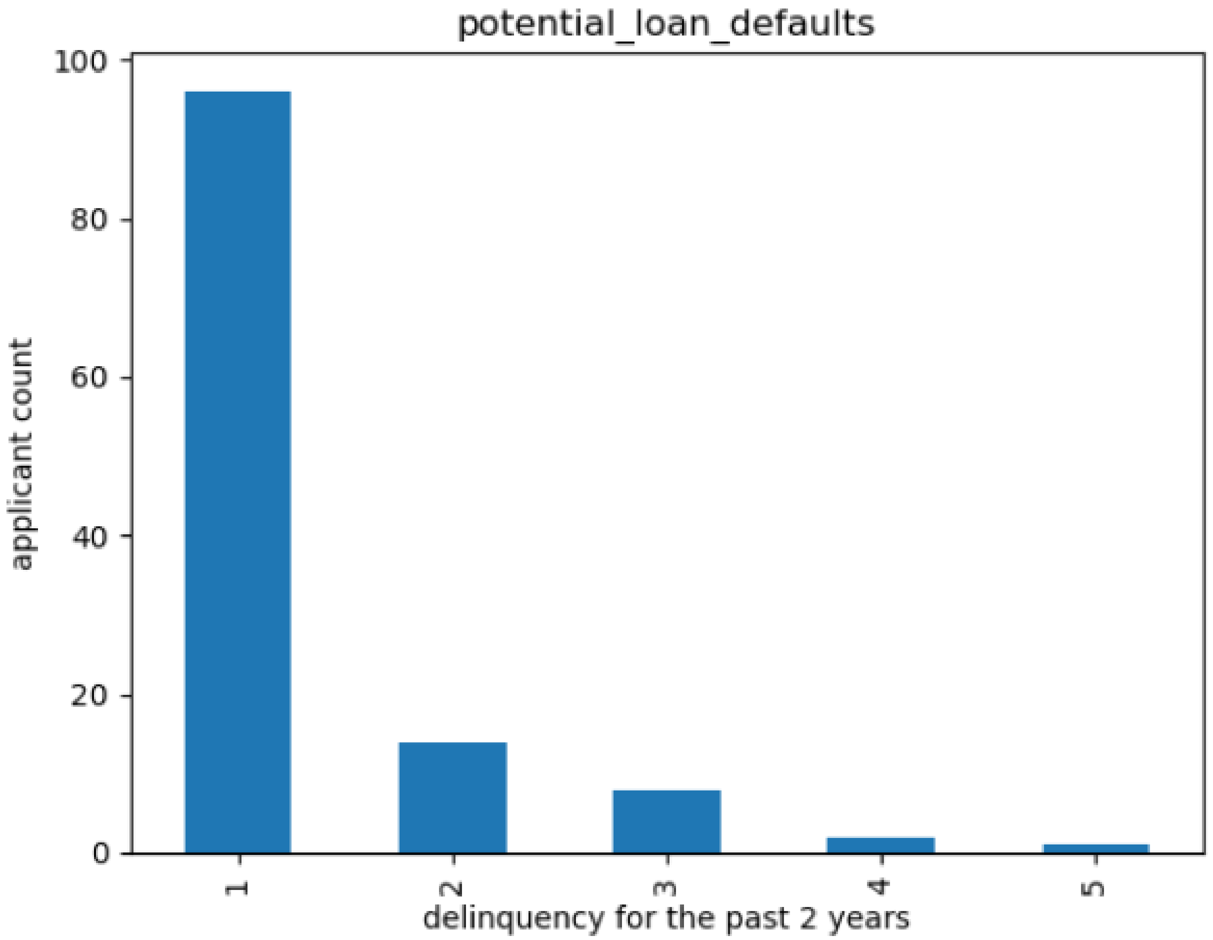
Univariate Analysis:

1) Loan Category



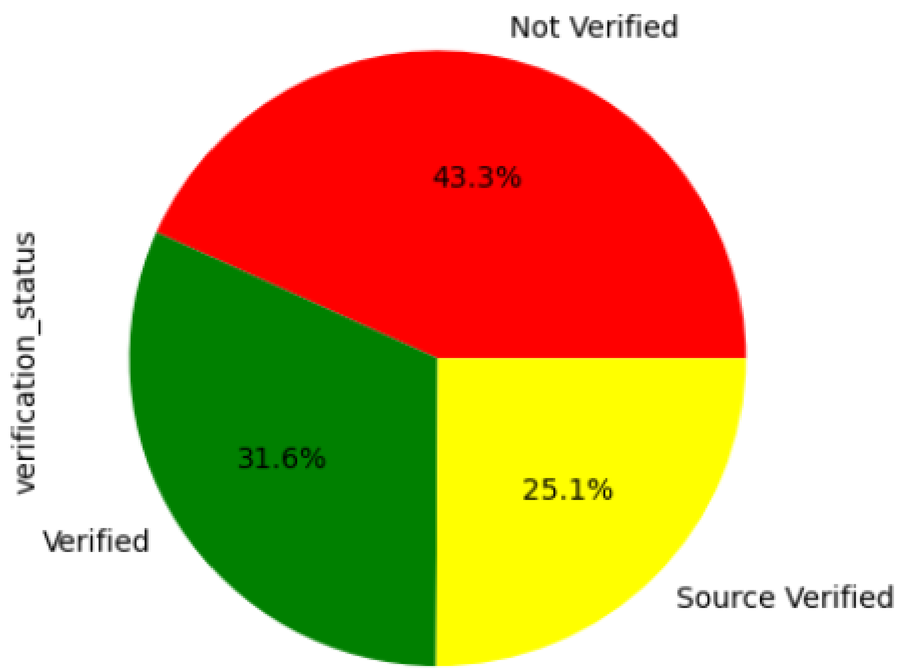
- Since we need loan default entries the current borrowers who presently crediting installment are not needed for analysis, hence dropping the Current category is necessary for analysis.

2) Potential loan defaults from currently active borrowers



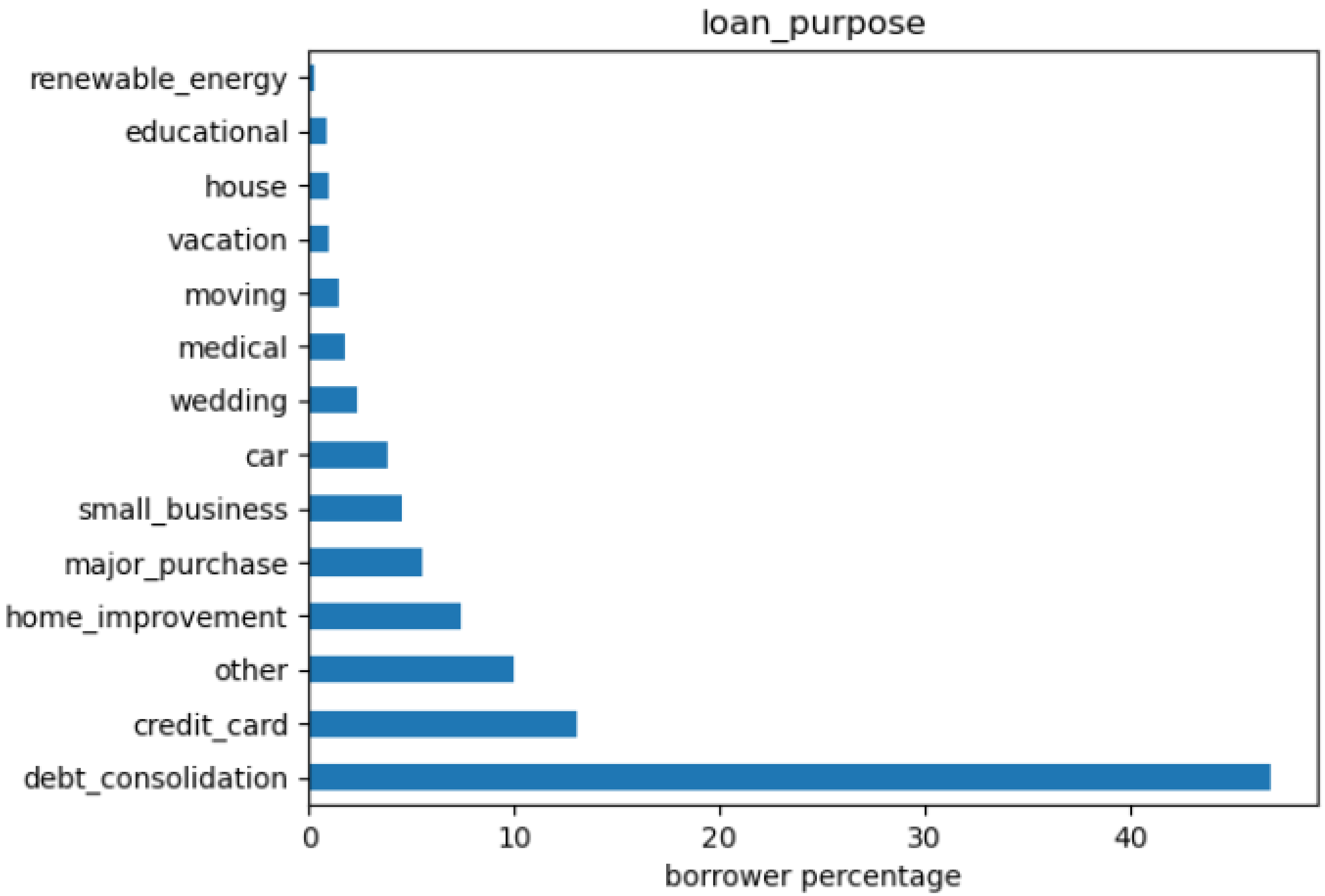
- There are 121 borrowers who are having 30+ days past-due incidences of delinquency.
- There are 25 borrowers who are having 30+ days past-due incidences of delinquency more than once, may be a potential_loan_defaults.

3) Analyzing bank's efficiency in applicant's income source verification



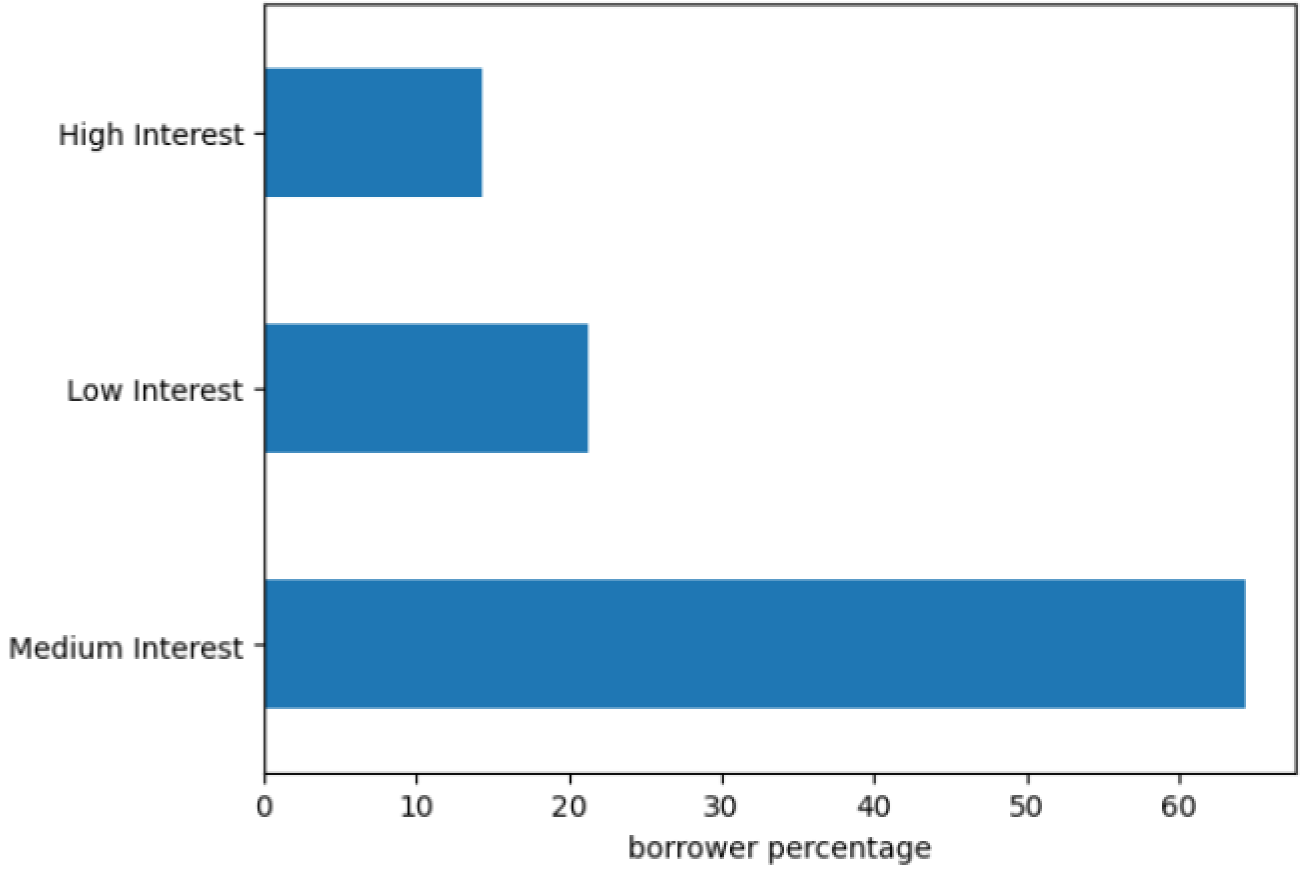
- Above graph shows over 43% of applicants received a loan without their income verification by bank before lending loan.
- This may increase the loan default or delay in loan repayment or delinquent.

4) Analyzing lending diversification



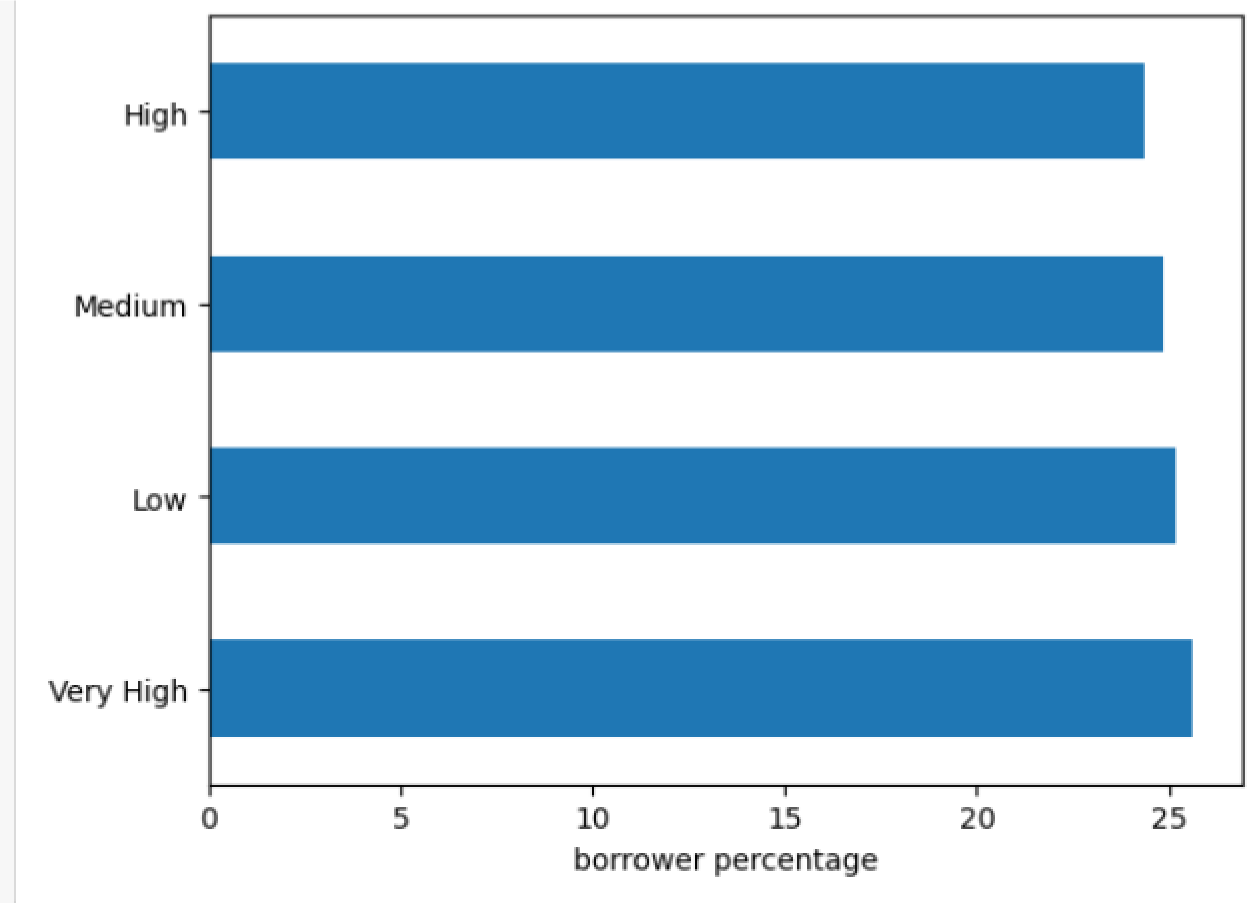
- Above graph shows over 46% of loan is issued for debt_consolidation
- Loan taken for Debt Consolidation is loan taken by person to pay off all existing loans. This shows a person's inability to pay loans
- It is better to diversify lending money to other purpose or sectors to expect better debt repayments and enhance company return

5) Categorization of Interest rate column and analysis



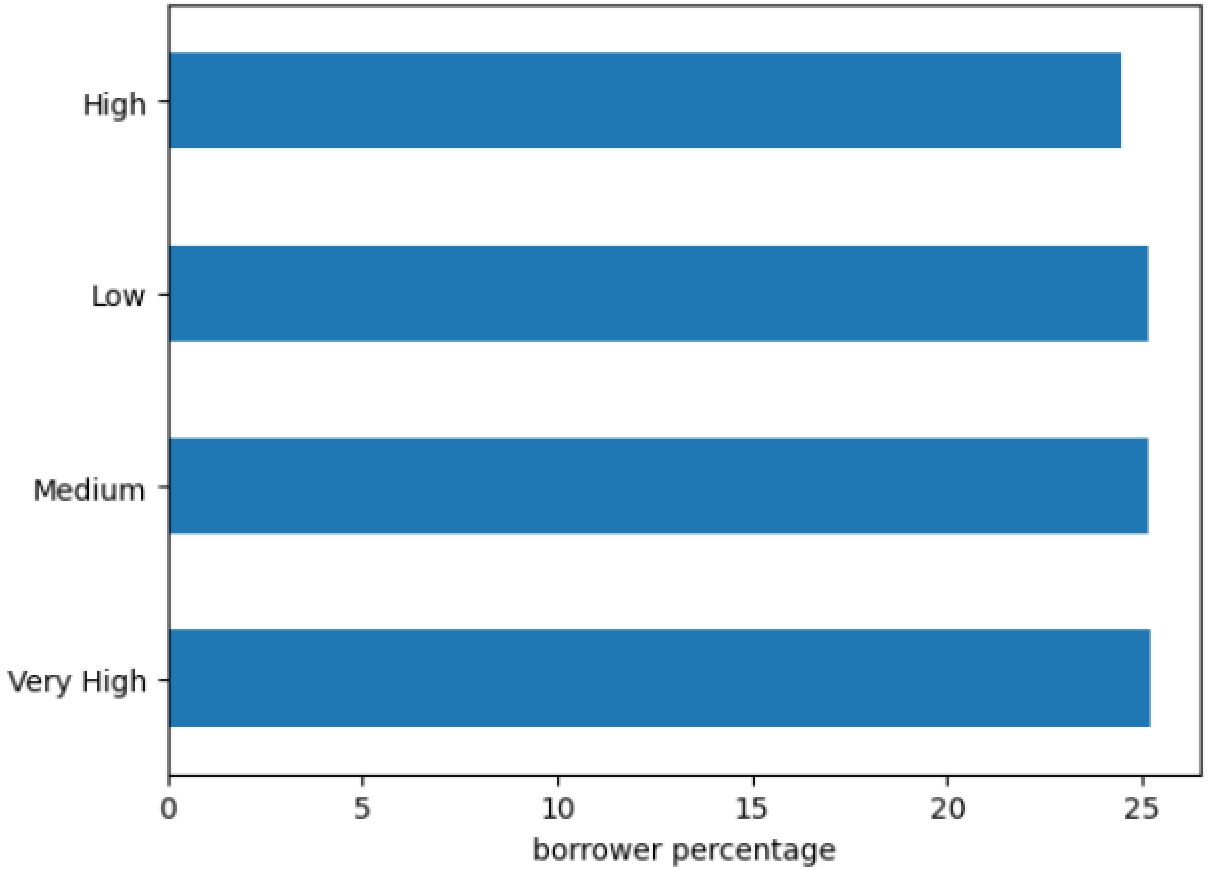
- Above graph shows over 60% of loan is issued with medium interest and less people got loan for high interest relatively.

6) Categorization of loan amount column



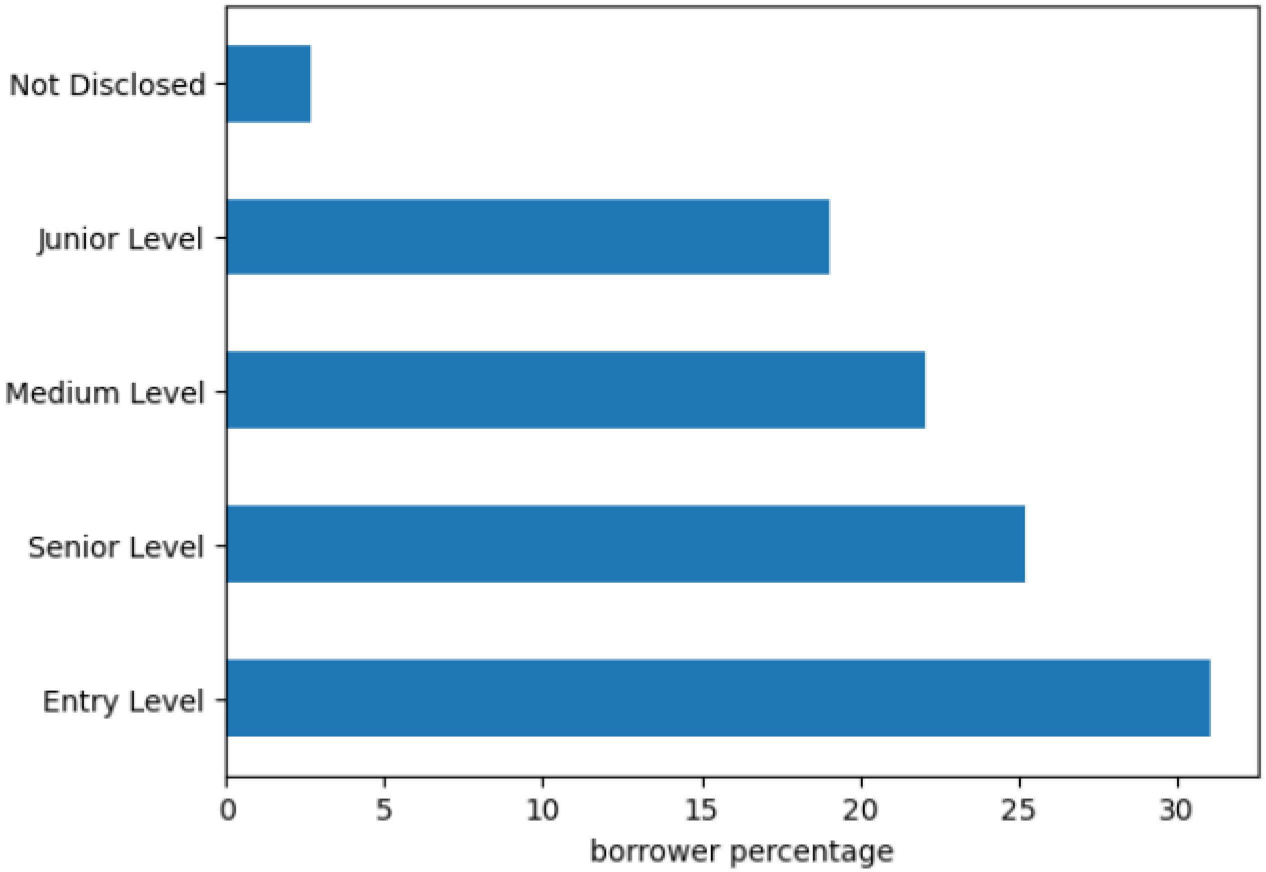
- Above graph shows almost all loan amount catagory has equal number of borrowers. This shows bank lending is not specific to any one amount category.

7) Categorization of Applicant's annual income column



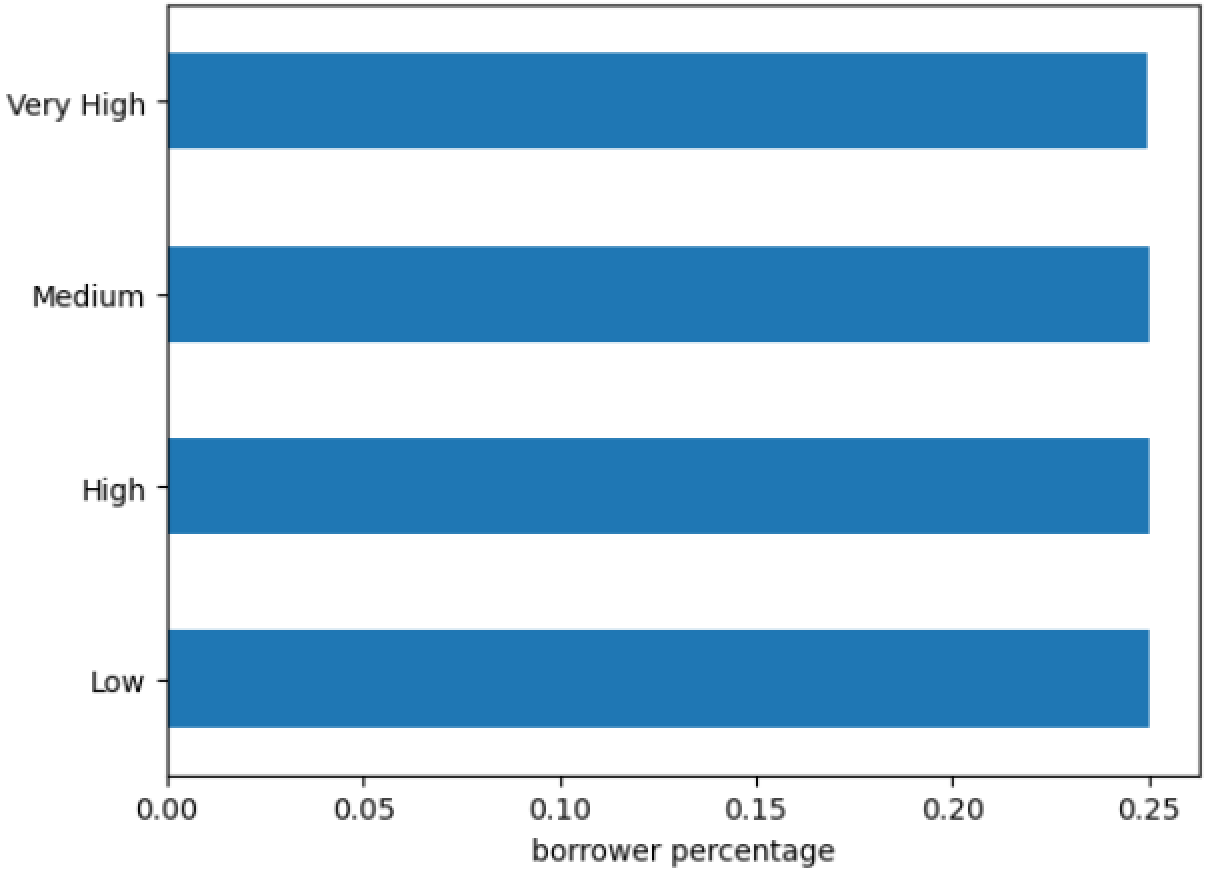
- Above graph shows almost all loan amount category has equal number of borrowers. This shows bank lending is not specific to any one amount category

8) Categorization of Applicant's working experience column



- Above graph shows significant portion of loan has been issued to Entry level and Junior level, even though their income will be less compared to middle and senior level borrowers

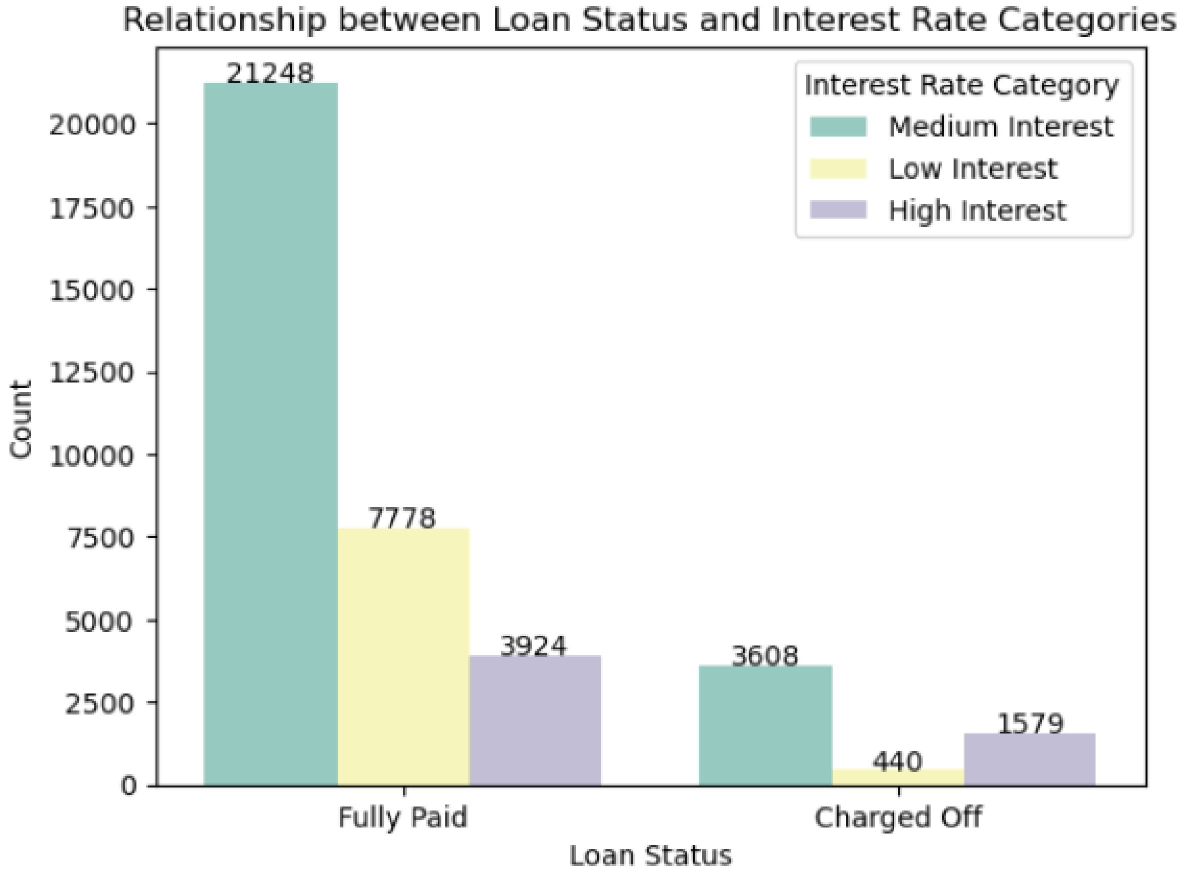
9) Categorisation of Applicant's DTI(Debt-To-Income) column



- Number of people borrowed for each category of DTI ratio are equal.

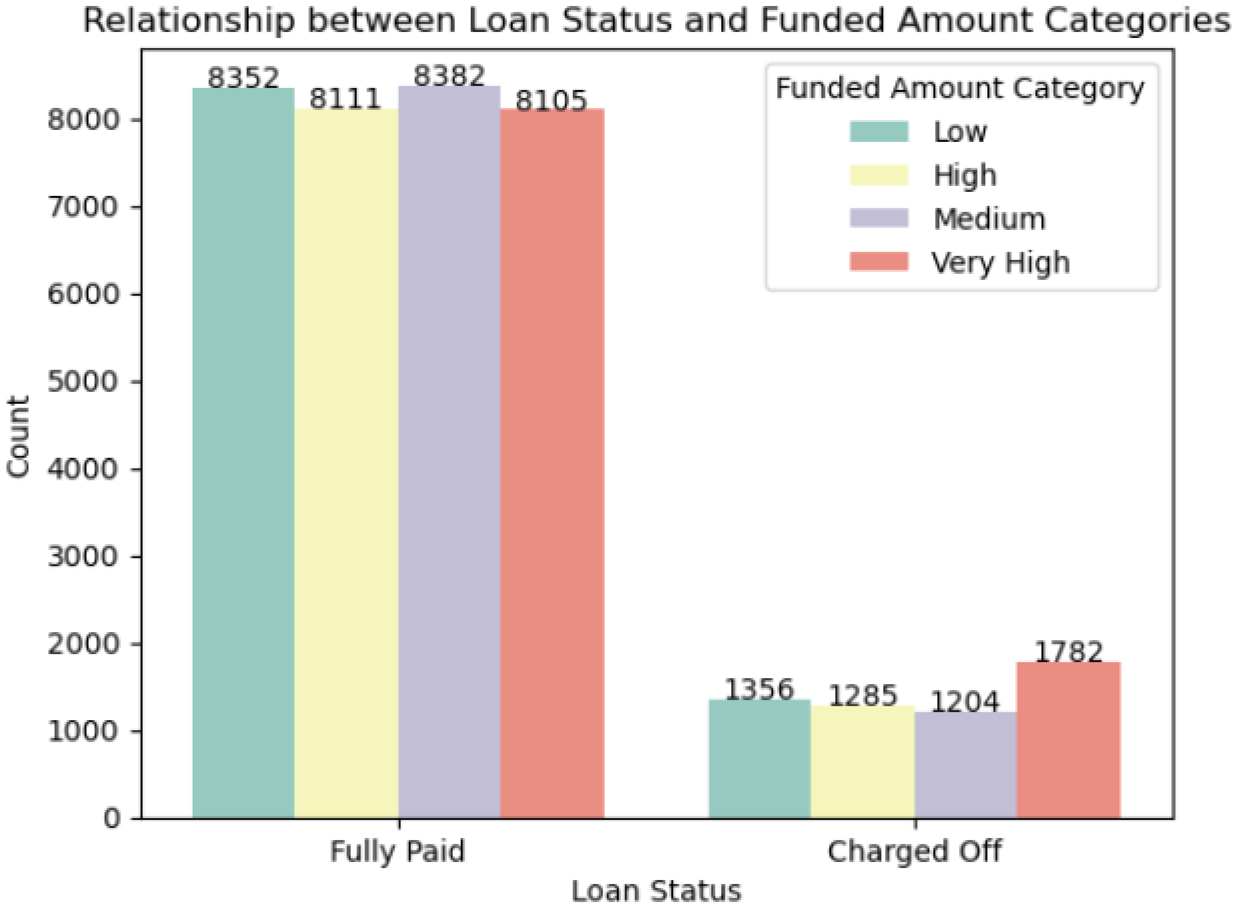
Bivariate Analysis:

1) Relationship between Loan Status and Interest Rate Categories



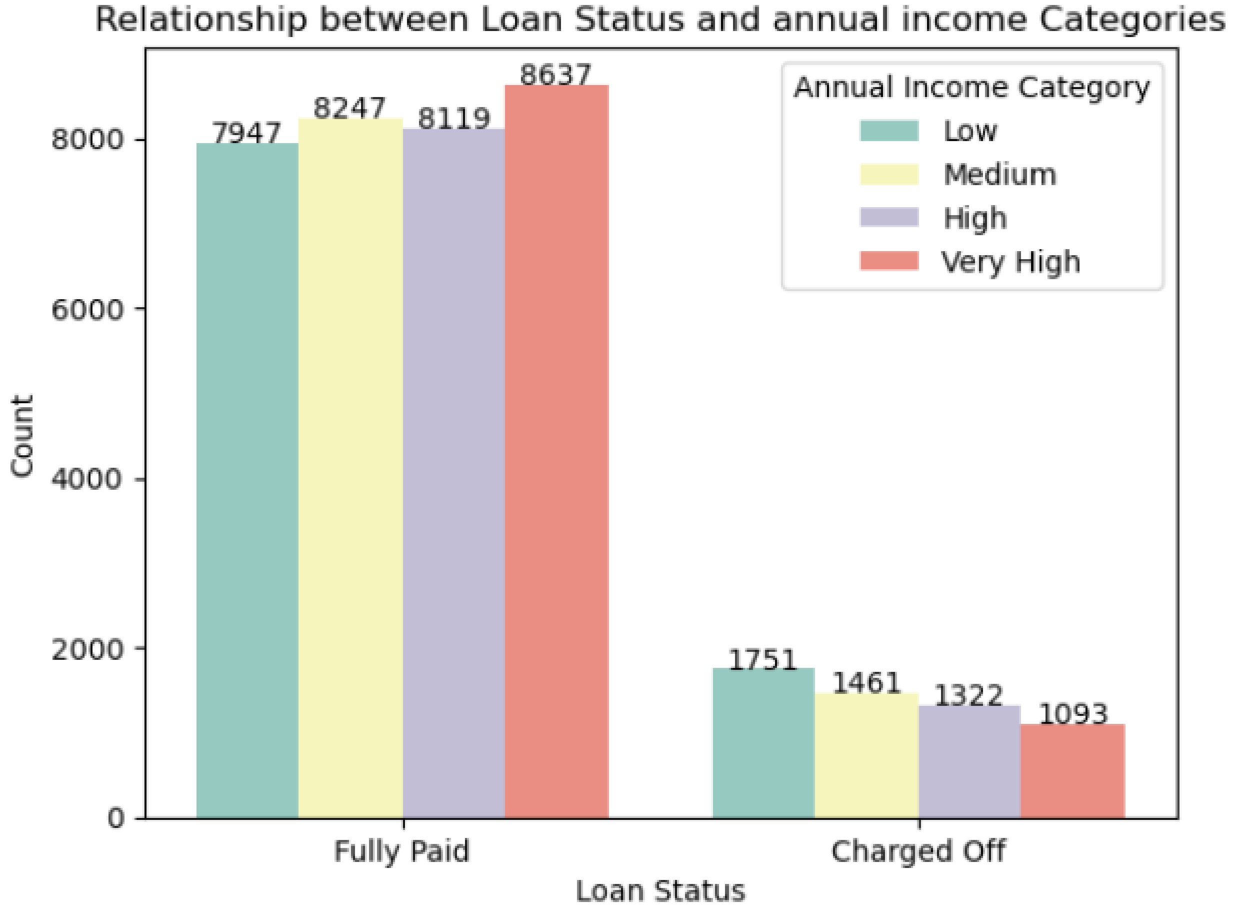
- Loan default probability for Medium Interest Rate is 1 person in 7 borrowers.
- Loan default probability for Low Interest Rate is 1 person in 19 borrowers.
- Loan default probability for High Interest Rate is 1 person in 3 borrowers.
- It is evident that if the interest rate is high then the loan default chances are considerably high

2) Relationship between Loan Status and Funded Amount Categories



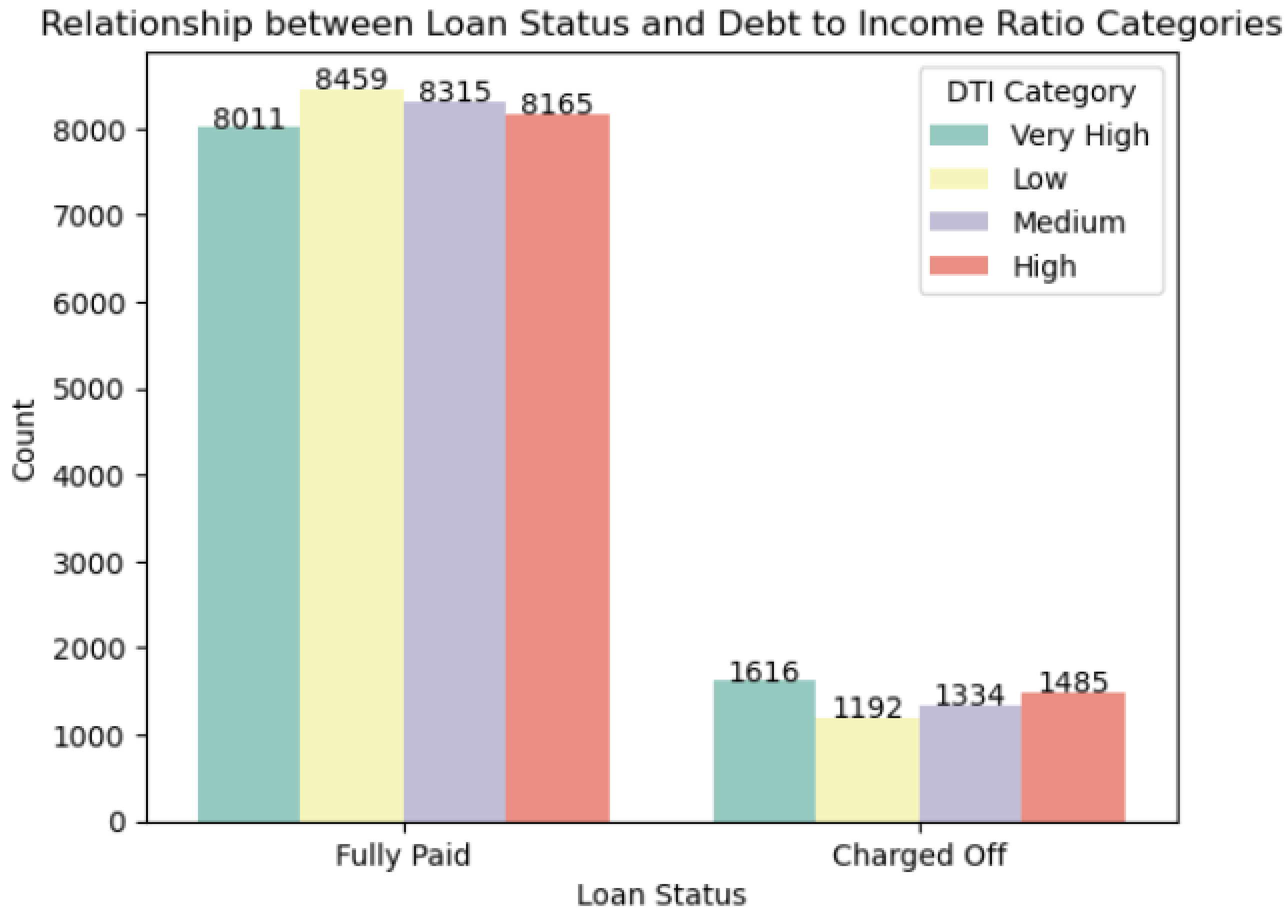
- There is minimal affect on loan default with respect to size of loan borrowed as graph shows similar Charged Off ratio for all categories

3) Relationship between Loan Status and Borrower's annual income Categories



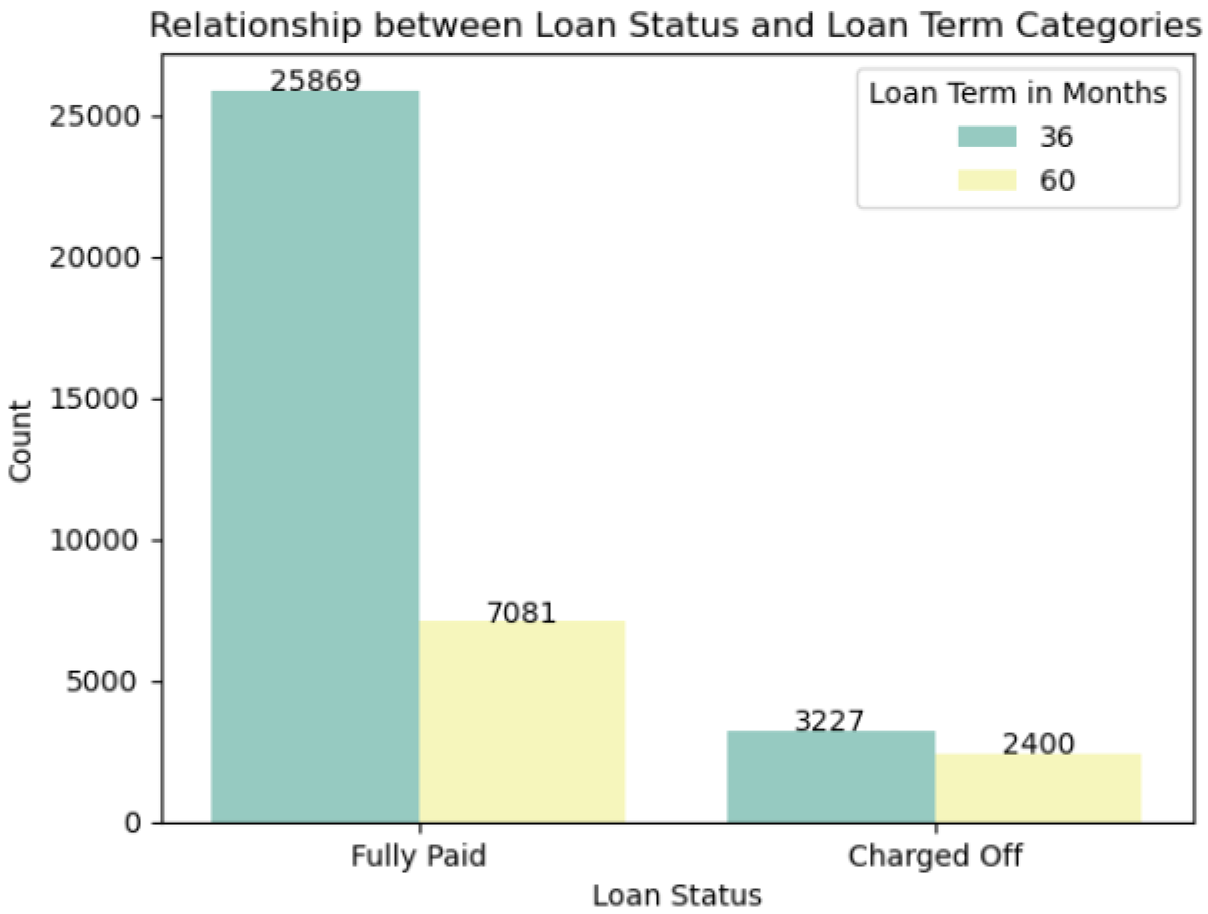
- There is increasing loan default risk trend with reducing annual income of a borrower. The difference between relative income category is less but low to very high income difference when considered Charged Off is significant.

4) Relationship between Loan Status and Debt to Income Ratio Categories



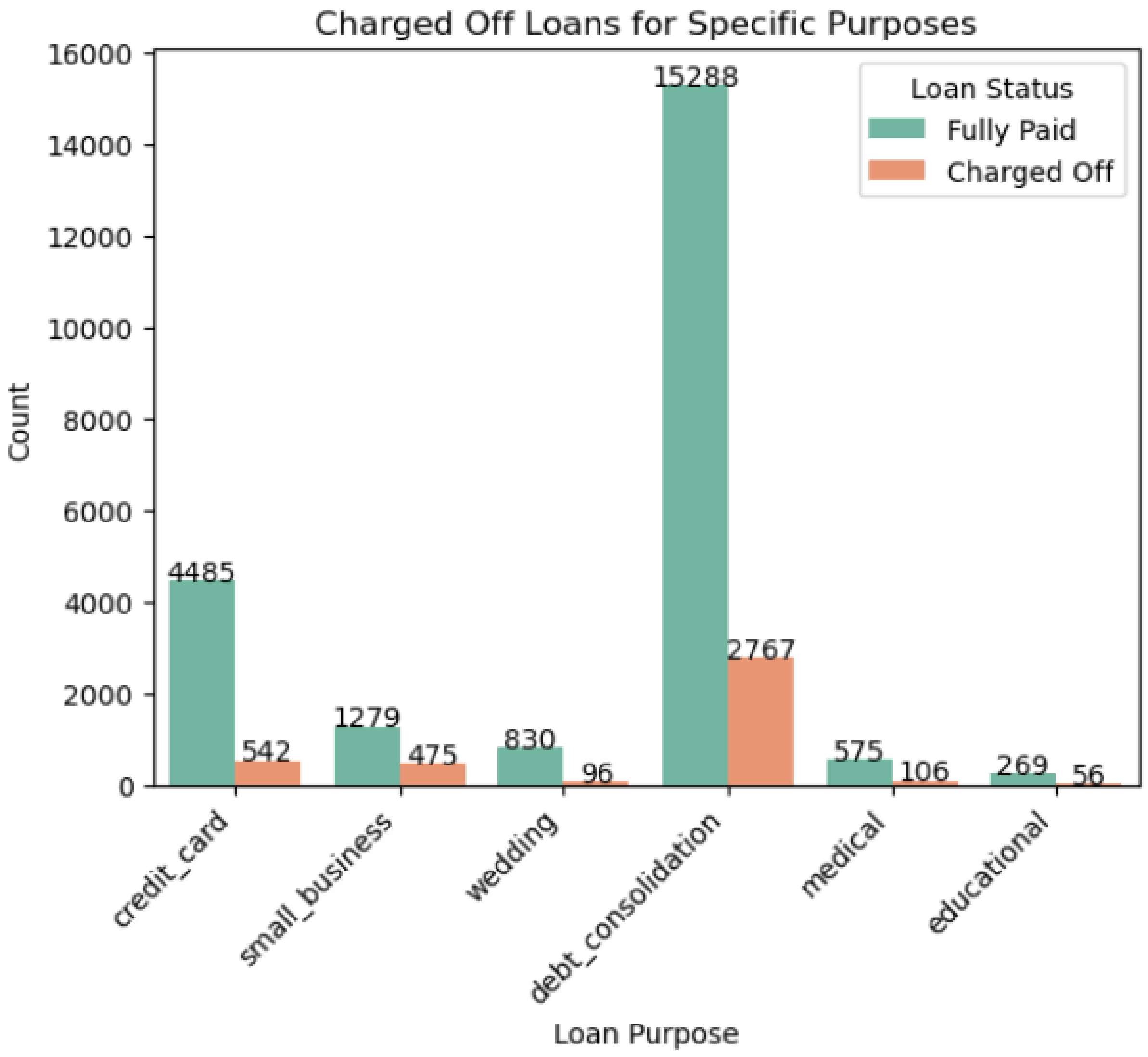
- DTI is directly proportionate with Charged Off. when DTI is low there are less charged off cases, when DTI is high the chances of loan default is high.

5) Relationship between Loan Status and Loan Term



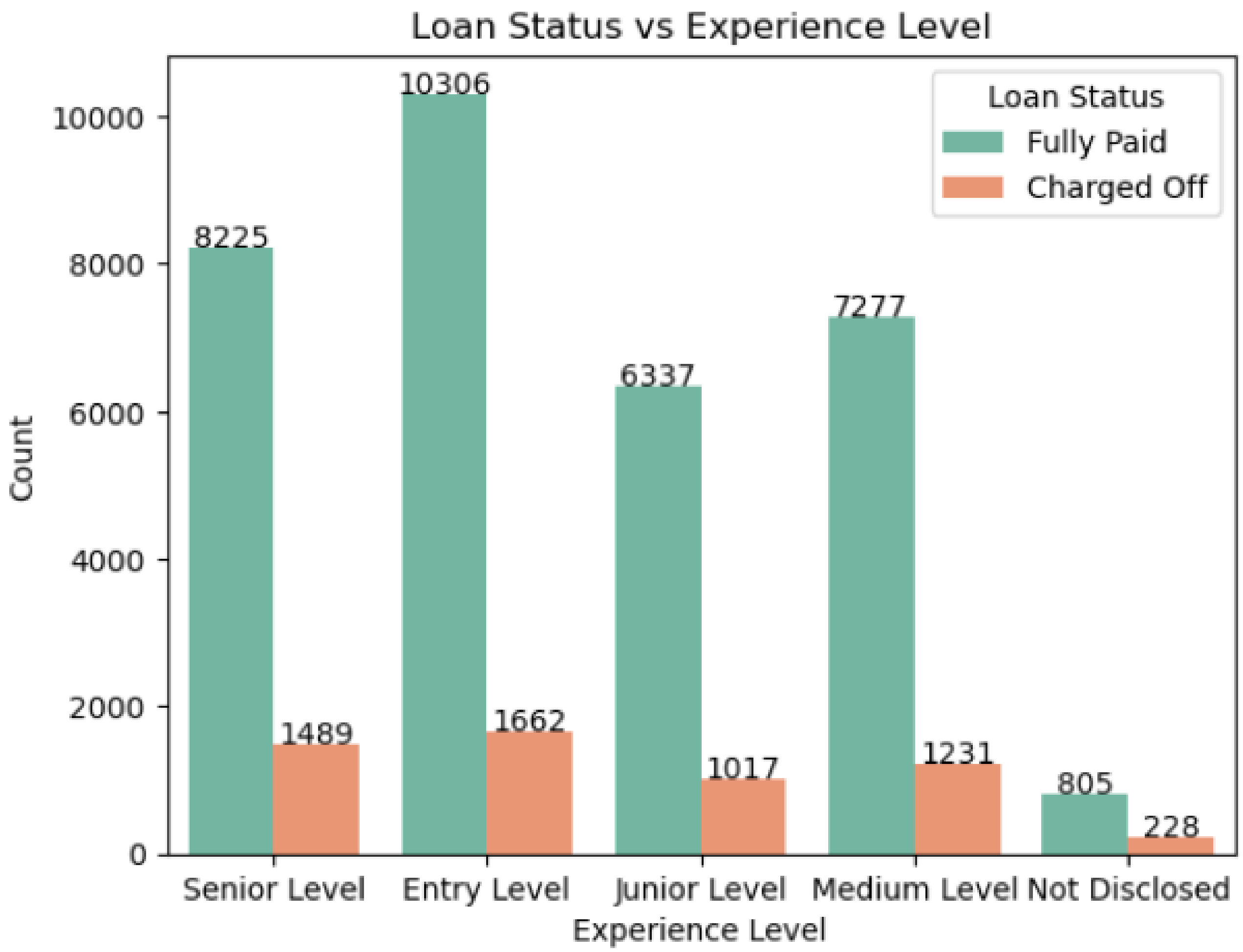
- There is 2.5 Times more risk of default when the loan term is 60 Months compared to 36 months.

6) Relationship between Loan Status and Specific Purposes



- credit card loan default ratio 1: 9
- small business loan default ratio 1: 4
- wedding loan default ratio 1: 10
- debt consolidation loan default ratio 1: 7
- medical loan default ratio 1: 6
- educational loan default ratio 1: 6
- Loan taken for small business, medical, educational are of high chance to be charged off than other purposes.

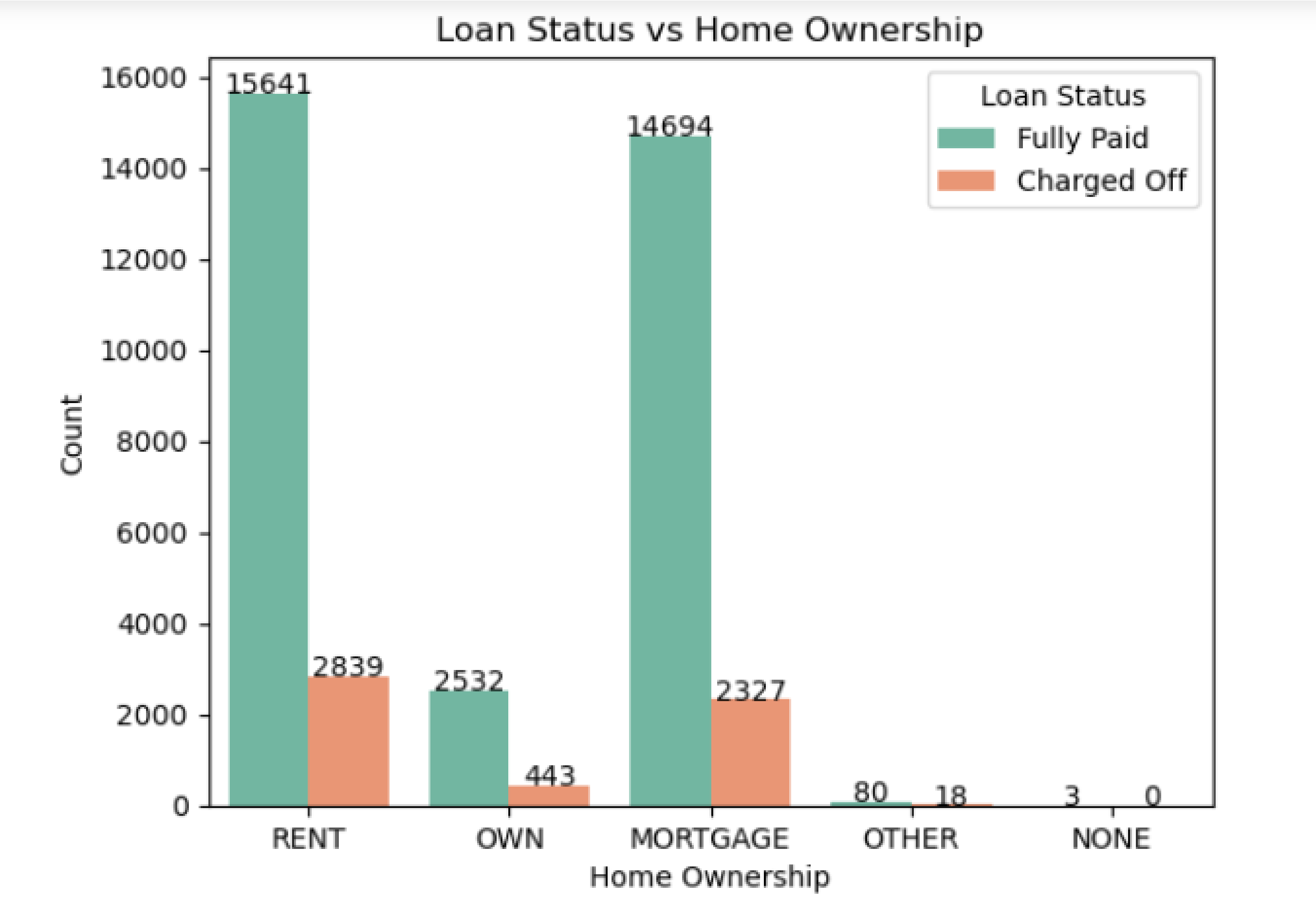
7) Relationship between Loan Status, Experience level



- Senior level loan default ratio 1: 7
- Entry level loan default ratio 1: 7
- Junior level loan default ratio 1: 7
- Medium level loan default ratio 1: 7

- Loan taken by all experience level borrower's loan default ratio is same 1 person for 7 borrowers.

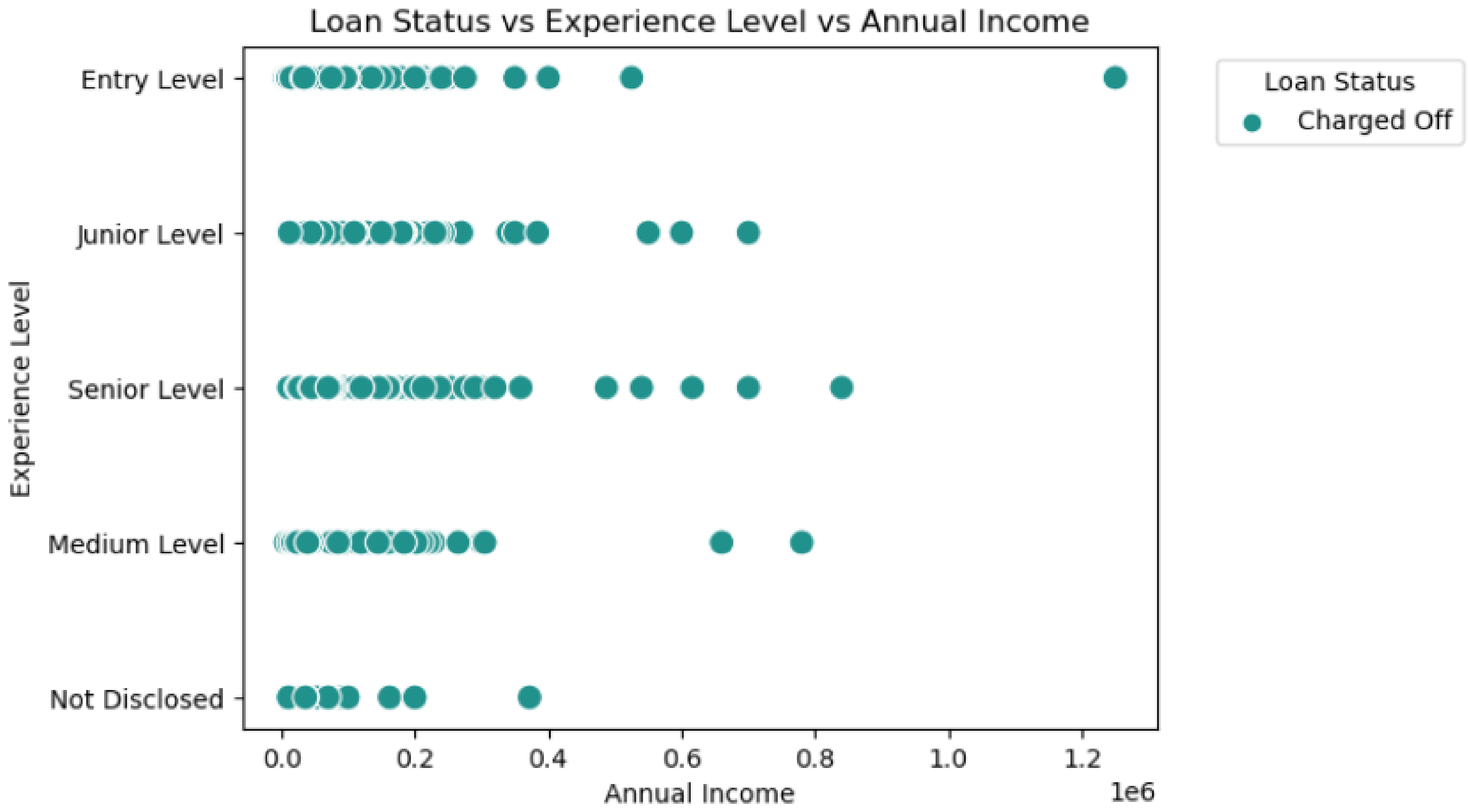
8) Relationship between Loan Status and Home ownership



- Borrowers who have rent and mortgage category are more likely to take loans.
- And the default is considerably high in numbers of both rent and mortgage category

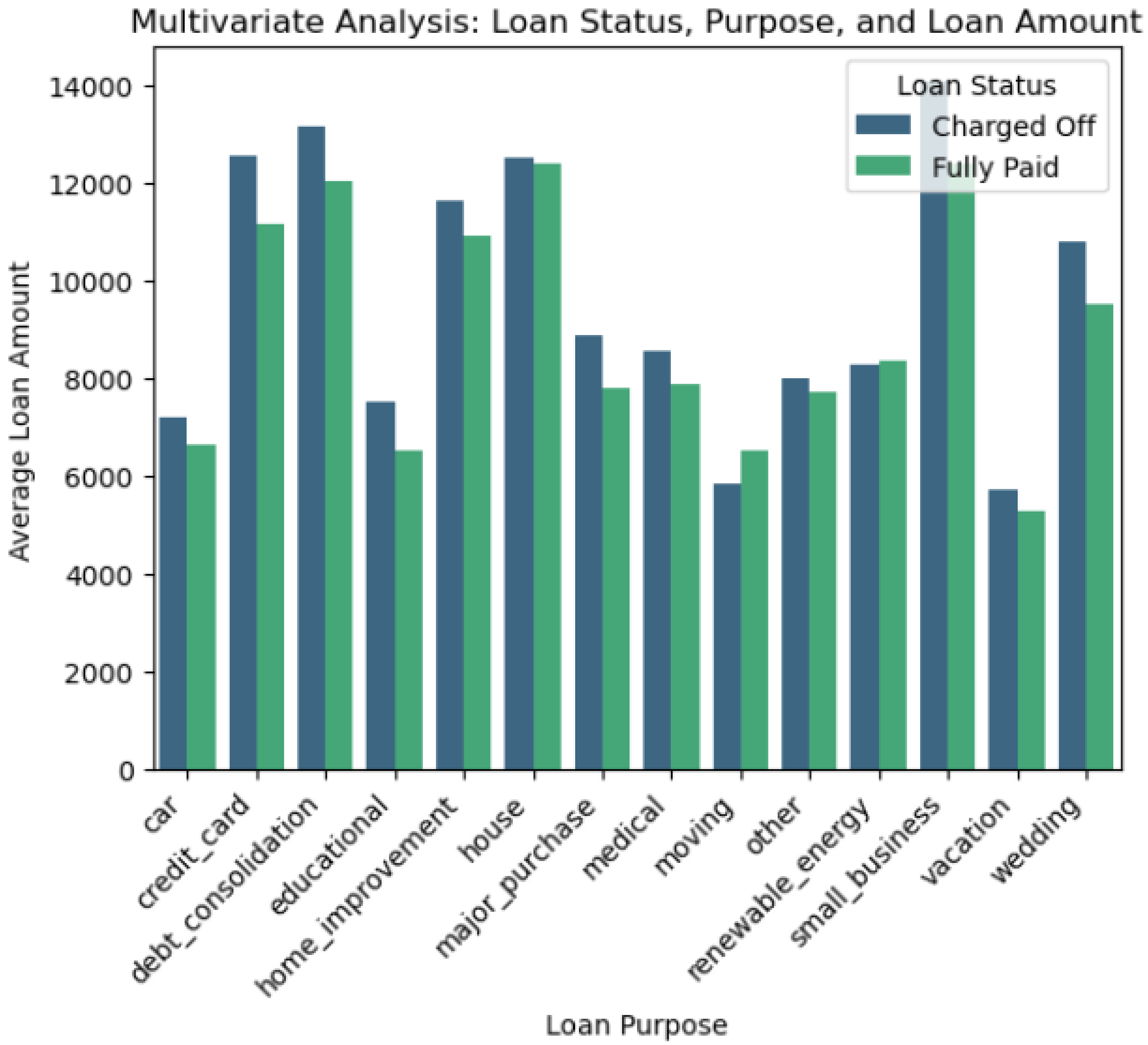
Multivariate Analysis

1) Relationship between Annual income, Experience and loan status == Charged Off



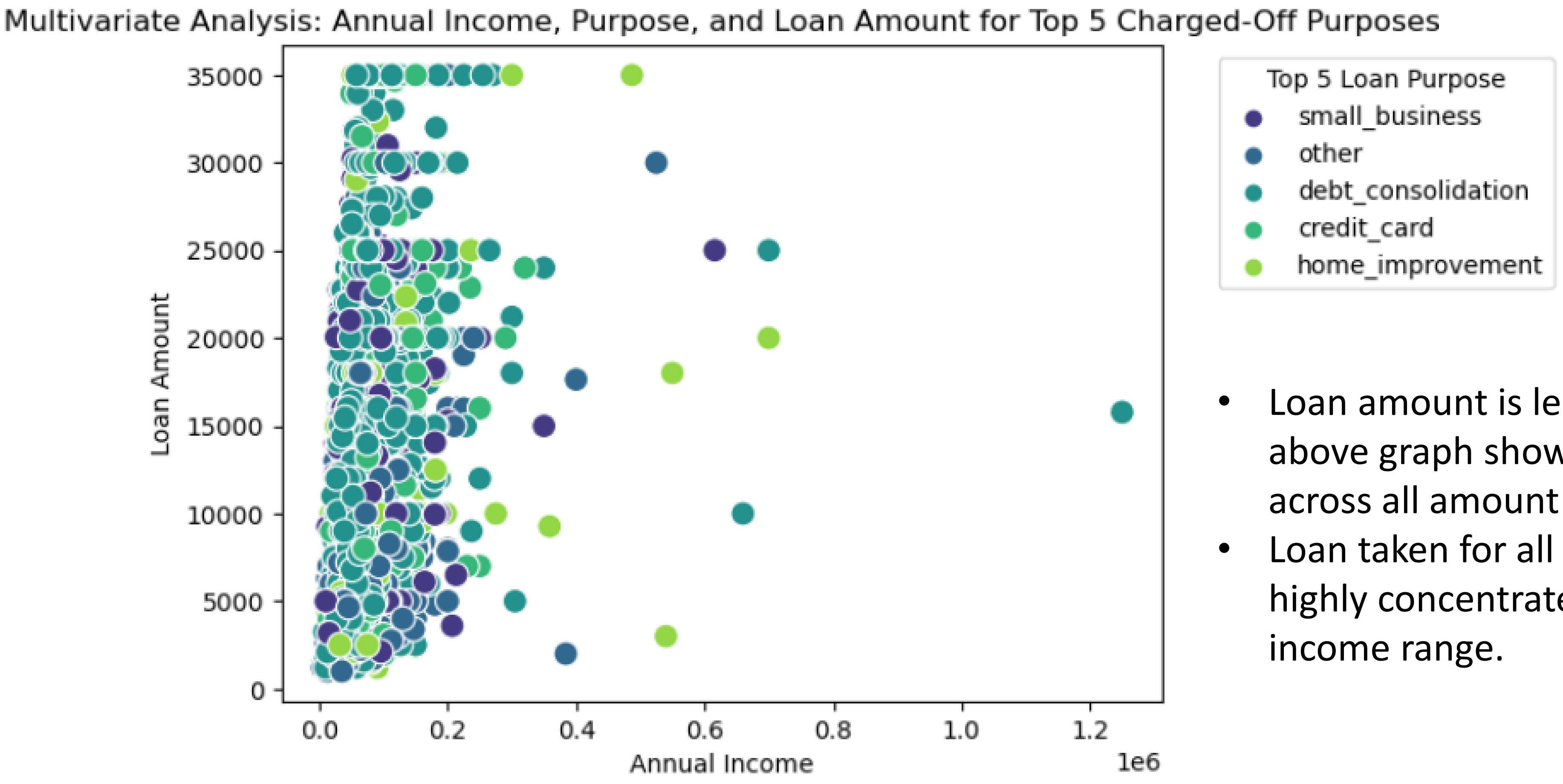
- The loan default is concentrated at the low income borrowers of each experience level, higher the income lesser the chances of default for all experience level category

2) Loan status v/s Purpose v/s Loan amount



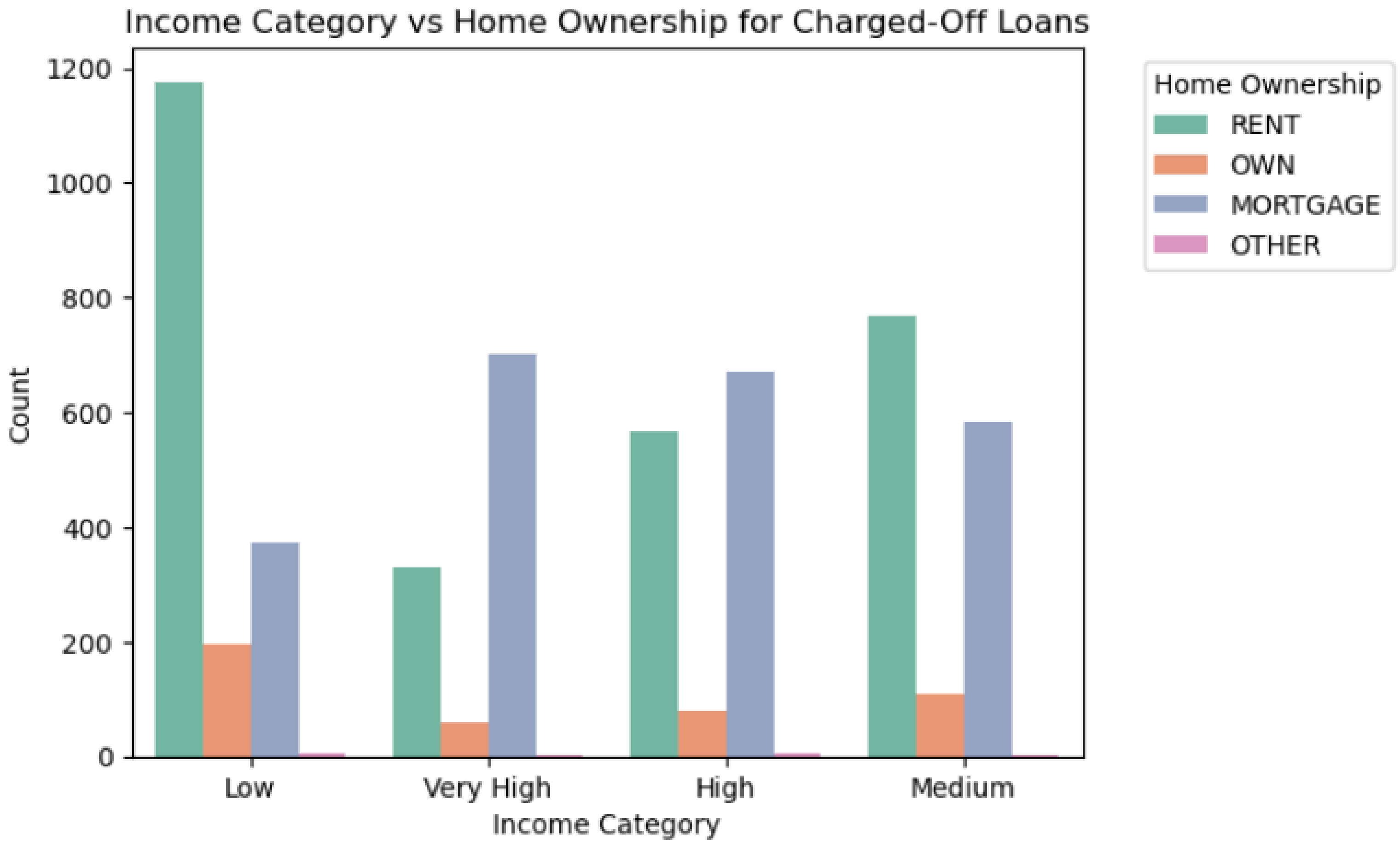
- The mean charged off amount is relatively higher than paid off mean amount in almost all the loan purpose categories.

3) Annual Income, Purpose, and Loan Amount for Top 5 Charged-Off Purposes



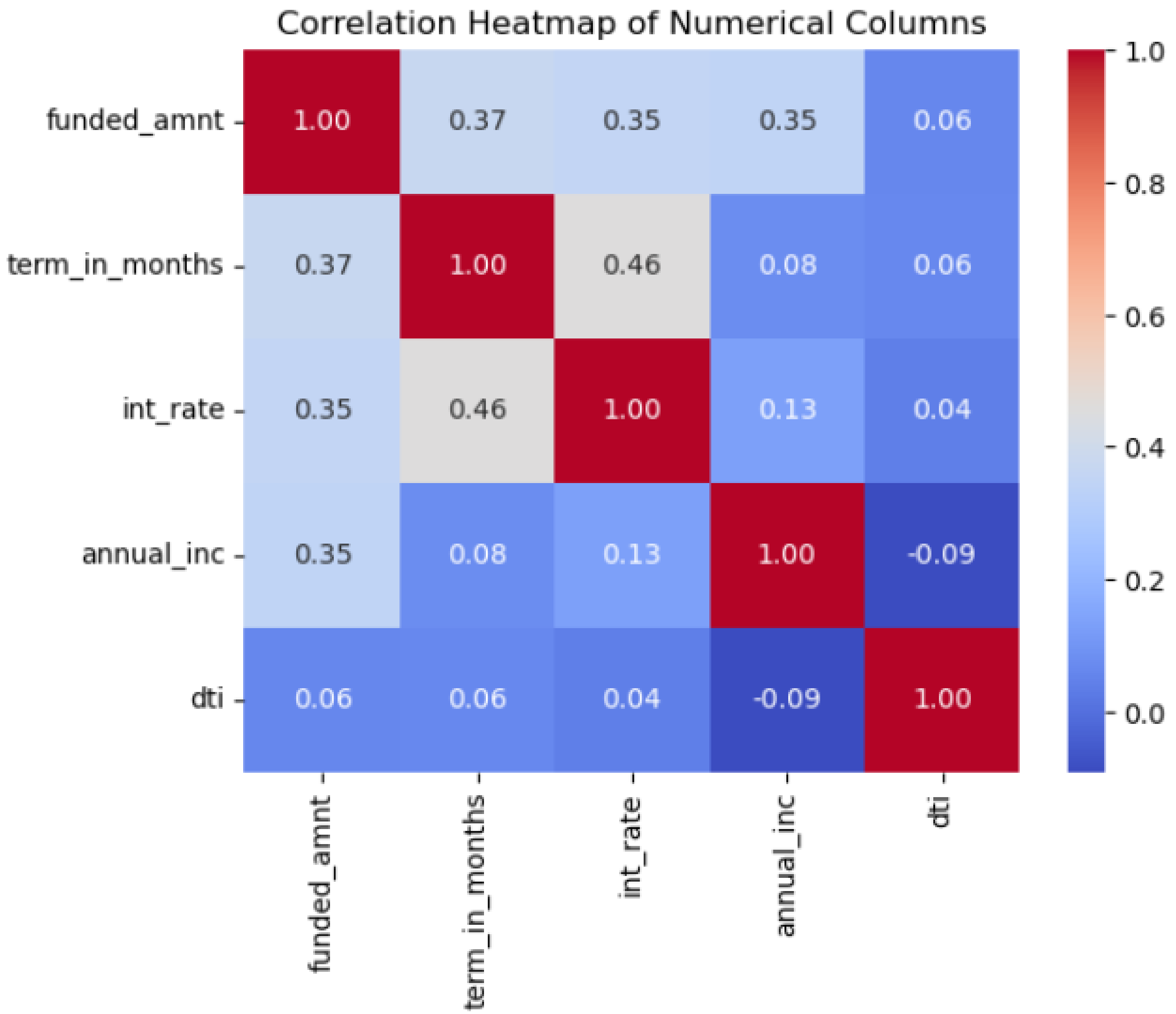
- Loan amount is least concerned as the above graph shows distribution of loan across all amount range
- Loan taken for all the purposes is highly concentrated at low annual income range.

4) Annual Income category , Home ownership, and Loan Default Relationship



- Loan Default is considerably high in low and medium Income category of Rented home applicants
- Loan default is higher in high and very high Income category of those who has home mortgage

5) Correlation Heatmap of Numerical Columns



Conclusion:

The following conclusions/suggestions can be drawn from the analysis

- Loan Default is considerably high in low and medium Income category of Rented home applicants and high and very high Income category of those who has home mortgage.
- Over 43% of applicants received a loan without their income verification
- Loan taken for all the purposes is highly concentrated at low annual income range. Hence concentrating on low and medium income level applicants with proper Income verification is necessary.
- Also loan default is high at low level income and entry or less job experience applicants. Hence careful assessment is needed before lending loan to this category
- Borrowers who have rent and mortgage category are more likely to take loans & default rate is also high with these category
- Small business loan default ratio 1: 4(1 in 4 applicants) Hence risk to issue loan to Small Business
- There is 2.5 Times more risk of default when the loan term is 60 Months compared to 36 months.(long term loan attracts more risk of defaulting)
- DTI is directly proportionate with Charged Off. Hence low DTI is better for lending
- Loan default probability for High Interest Rate is 1 person in 3 borrowers. Hence high interest may increase the risk of loan default