LENDING CLUB CASE STUDY

BY

MOHANA KRISHNA

MONALISA PATRA

AGENDA

- Problem Statement & Business Objective
- Data Understanding
- Data Cleaning
- Data Conversion
- Univariate Analysis
- Bivariate Analysis
- Correlation Analysis
- Proposals
- References

PROBLEM STATEMENT & OBJECTIVES

- Lending Club is a company specializes in lending various types of loans to urban customers.
- When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile.
- Two types of risks are associated with the bank’s decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company
- The data given in the CSV contains information about past loan applicants and whether they ‘defaulted’ or not.
- 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.
- When a person applies for a loan, there are two types of decisions that could be taken by the company:
    - Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
        - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
        - Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
        - Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
    - Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

DATA UNDERSTANDING

Dataset Attributes:

- Primary Attribute : loan_status : This is the column that we can solely depend on for our analysis. This column has three values
  - Fully_Paid : Customers who have fully paid off the loan
  - Charged_Off : Defaulters who did not pay the installments
  - Current : Active applicants who are paying the installments on regular basis
- We are excluding “Current” values from our analysis as this category of customers will not help in our analysis
- All these attributes provided by the customer while filling the loan application form are useful for our analysis
- These attributes can be further classified into two categories
  - Customer related : DTI, annual_income, house_ownership, emp_length etc
  - Loan business related: loan_amnt, term, issue_d, purpose, int_rate etc

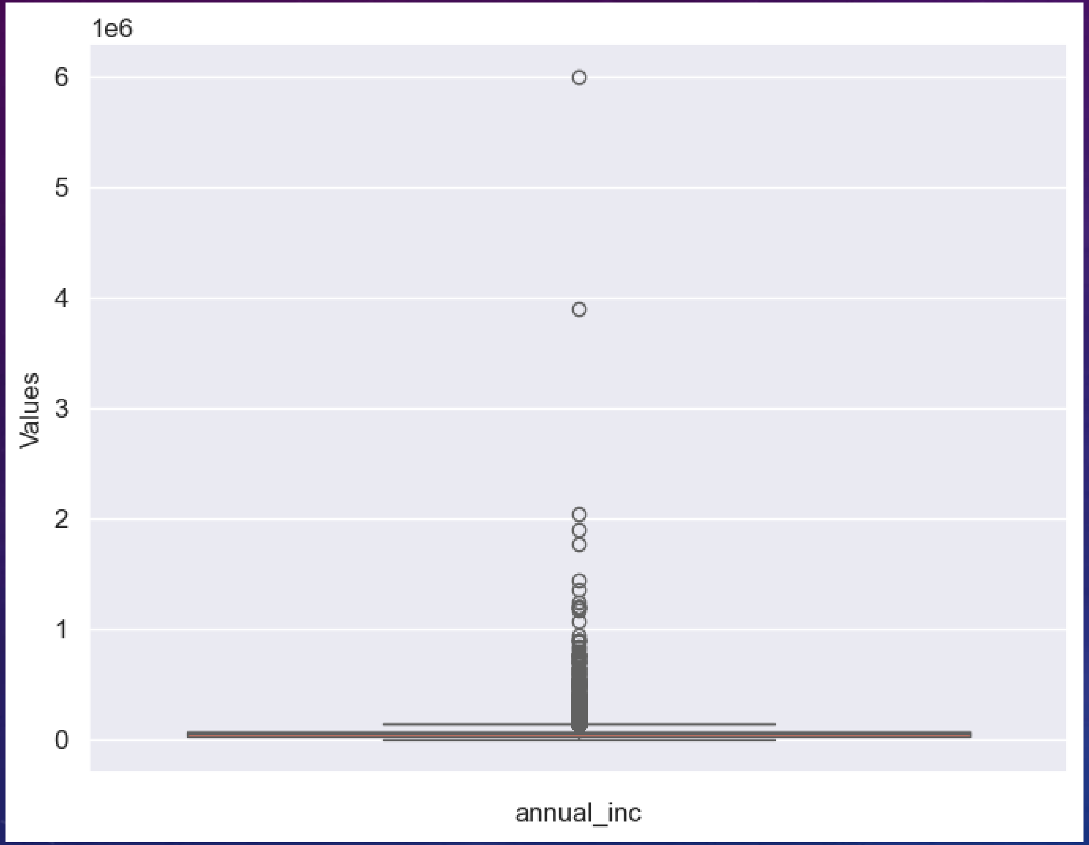
DATA CLEANING

- We have 111 columns in the provided data.
- On careful analysis, we have performed the below operations to clean up the data
  - **Dropped Customer behaviour columns :** 'collection_recovery_fee', 'delinq_2yrs', 'desc', 'earliest_cr_line', 'emp_title', 'id', 'inq_last_6mths', 'last_credit_pull_d', 'last_pymnt_amnt', 'last_pymnt_d', 'member_id', 'open_acc', 'out_prncp', 'out_prncp_inv', 'pub_rec', 'recoveries', 'revol_bal', 'revol_util', 'title', 'total_acc', 'total_pymnt', 'total_pymnt_inv', 'total_rec_int', 'total_rec_late_fee', 'total_rec_prncp', 'url'
  - 54 columns were having empty values. Directly dropped those columns
  - There are 9 columns with unique values which were dropped as they wont make any sense for analysis
  - No duplicate rows were found
  - Dropped all the rows whose loan_status = 'current'
  - Dropped 3 columns are the % of missing values is >=65%

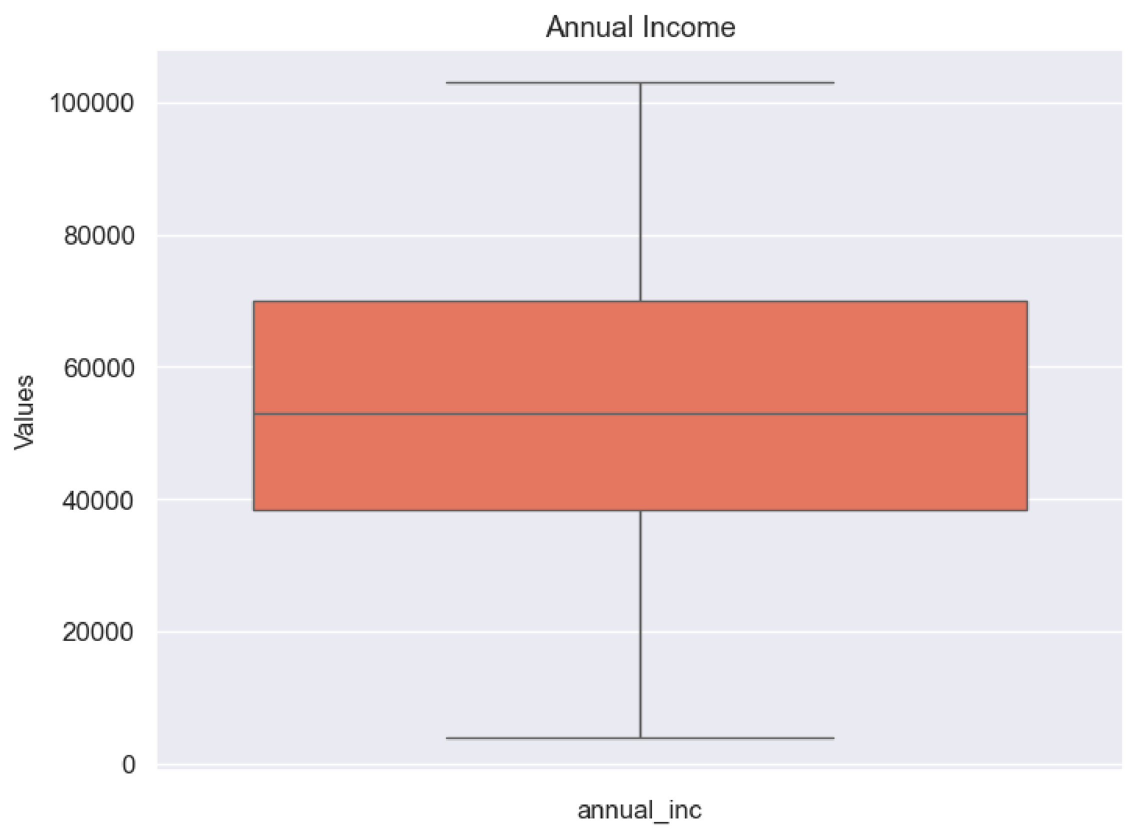
DATA CONVERSION

- Converted all the numeric columns to float to maintain the uniformity
- Int_rate column was of object type and contained % symbol in the values. Removed % symbol and converted the values into int
- Removed ‘months’ string from term column and converted into int
- Rounded off the useful numeric column values to 2 digits
- Converted issue_d column to YYYY-MM-DD format

HANDLING OUTLIERS

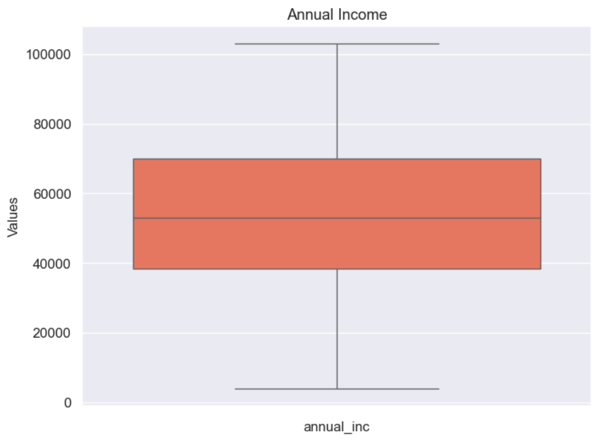


Before Outlier Removal

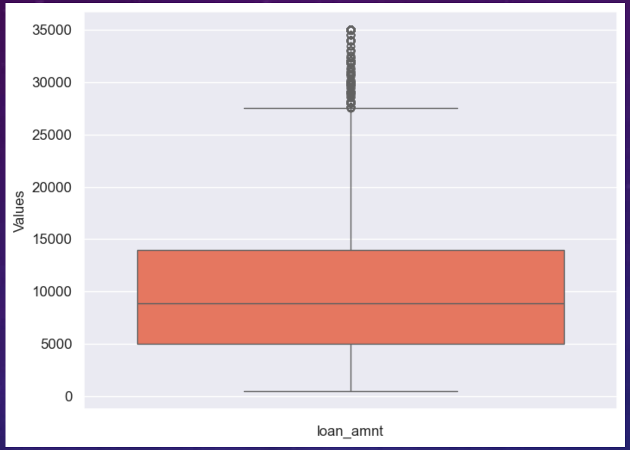


After Outlier Removal

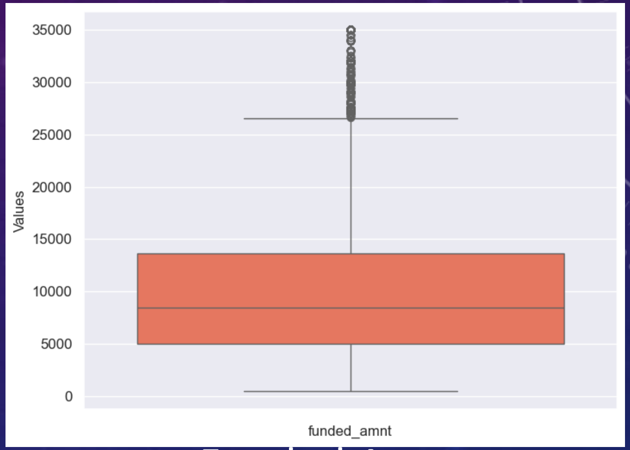
HANDLING OUTLIERS



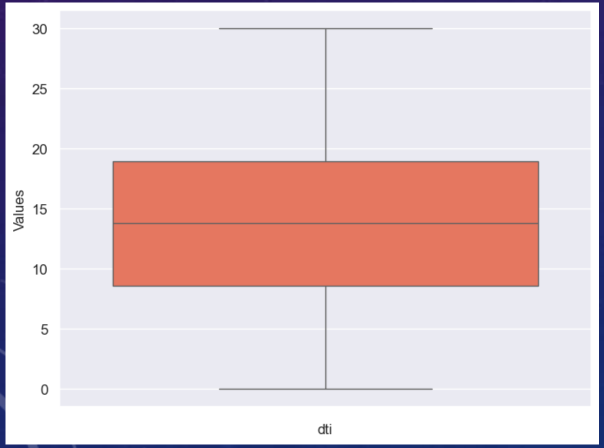
Annual Income



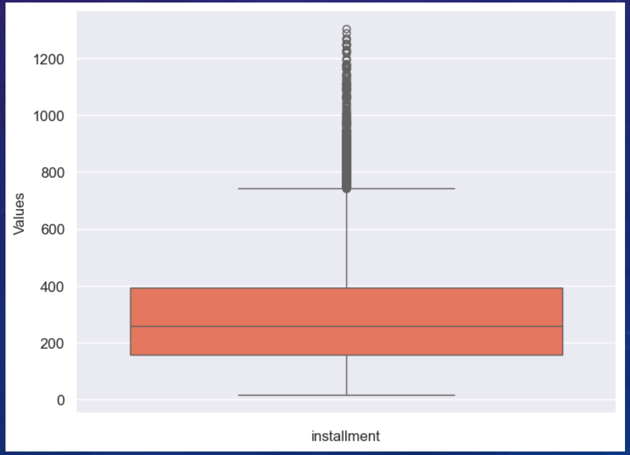
Loan Amount



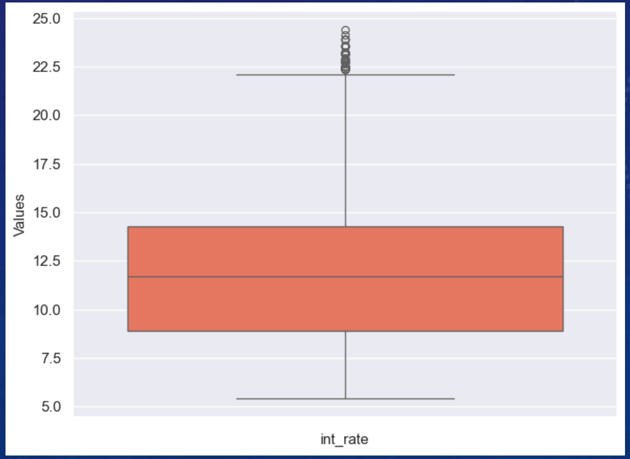
Funded Amount



Debt to Income



Instalment



Interest Rate

HANDLING OUTLIERS

OBSERVATIONS ON MOST OF THE APPLICANTS

- Annual Income – 40K – 70K USD
- Loan Amount - 5K – 15K USD
- Funded Amount 5K – 14K USD
- Interest Rate – 9% - 14%
- Installment – 160 – 440 USD
- DTI – 8 - 18

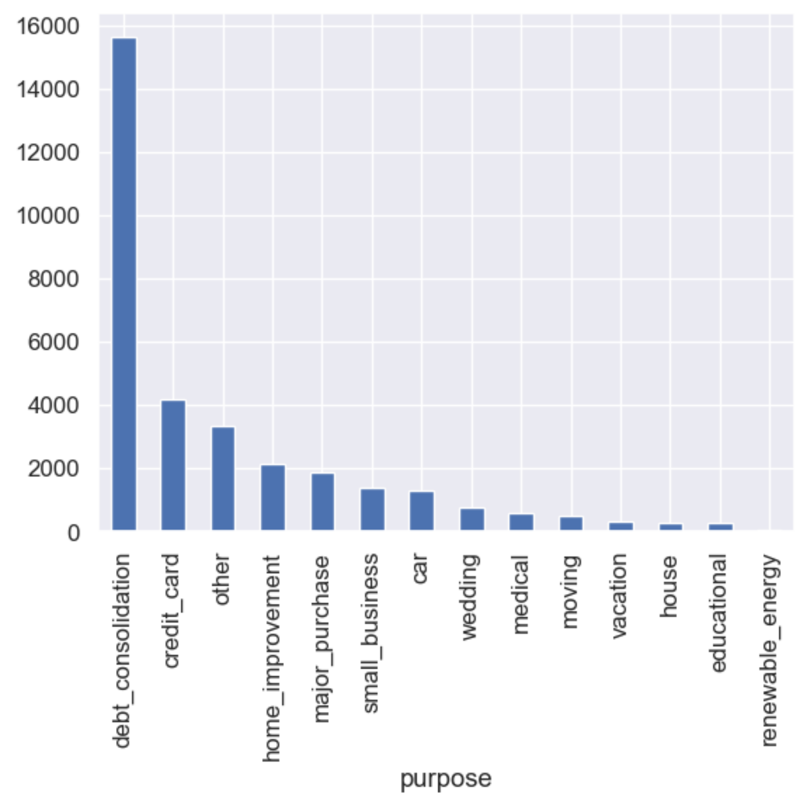
IMPUTING VALUES

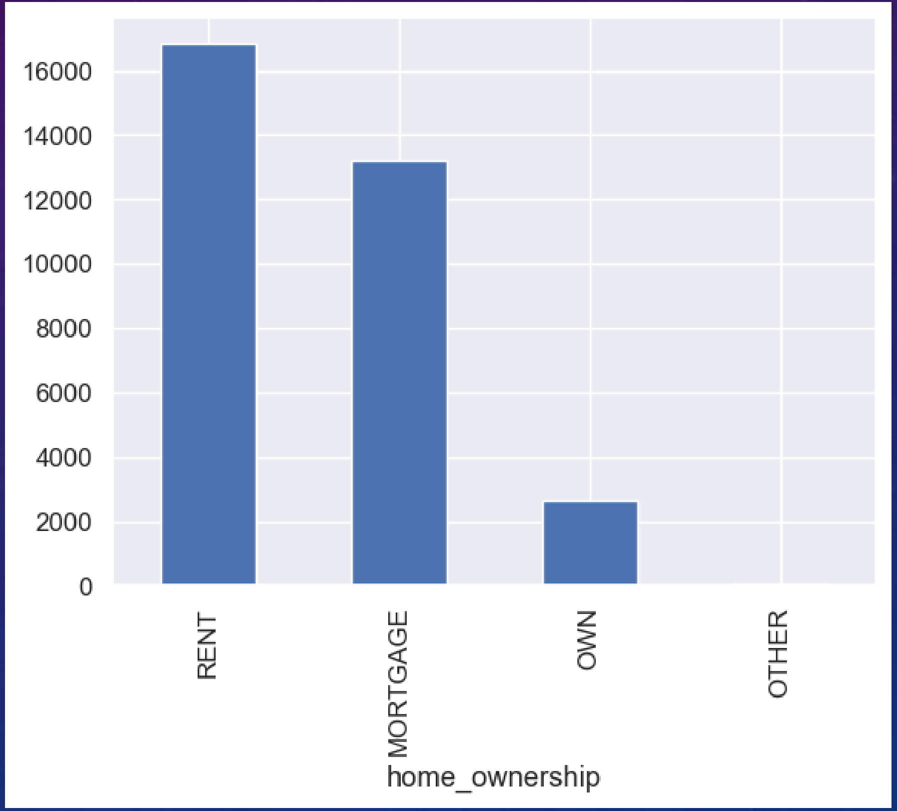
- 986 rows had None value for **emp_length** column
- Found the average salary of those applicants whose **emp_length** is null & Also found the most frequent **annual_inc** where **emp_length** is missing
  - **lending_df[lending_df.emp_length.isna()]['annual_inc'].mode()**
- Though the **emp_length** is missing, the **annual_inc** of these records are 36000, which is a decent income.Hence, we can infer that, these applicants are self employed.
- We can fill all the null values with the mode() of **emp_length** which is 10+ years
- **home_ownership** column has 2 records with NONE value. Imputed these two values with OTHER value
- **verification_status** column has two values ‘Verified’ & ‘Source Verified’ with the same meaning. So, combined them
- **emp_length** column values are converted to numeric : 0,1,2,3,4,5,6,7,8,9,10
- 602 rows has None values for **pub_rec_bankruptcies** column. Dropped all these rows

DERIVED COLUMNS

- Derived **issues_month** & **issues_year** from **issue_d**
- Created buckets for **loan_amnt, int_rate, annual_inc , dti** columns

UNIVARIATE ANALYSIS
UNORDERED CATEGORICAL VARIABLE ANALYSIS

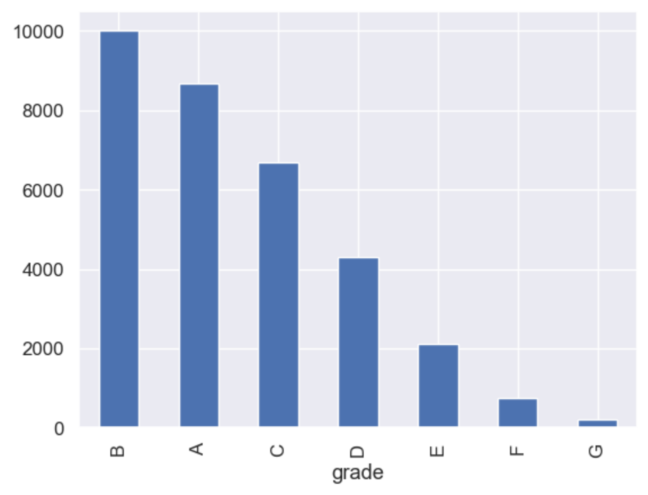


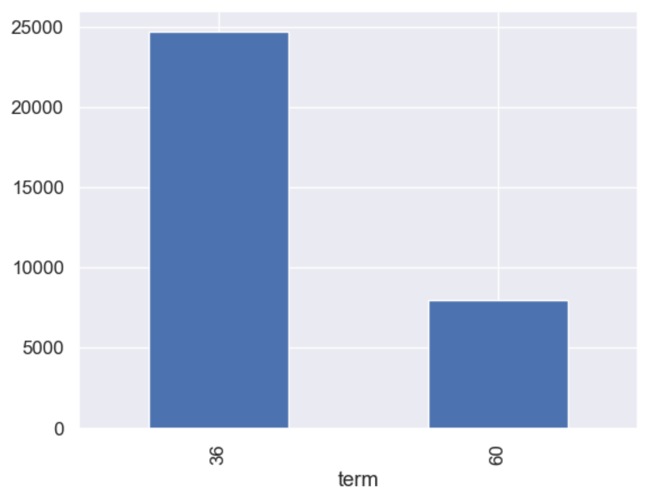


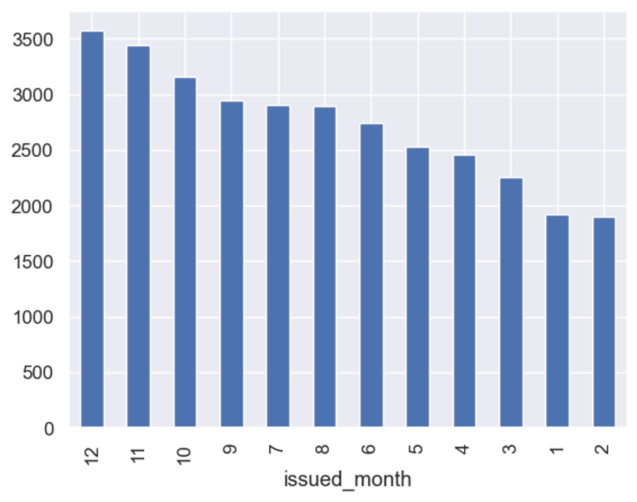
OBSERVATIONS

- Most of the loan applicants are staying in Rented house
- Most common purpose for the loan is ‘debt_consolidation’

UNIVARIATE ANALYSIS
ORDERED CATEGORICAL VARIABLE ANALYSIS



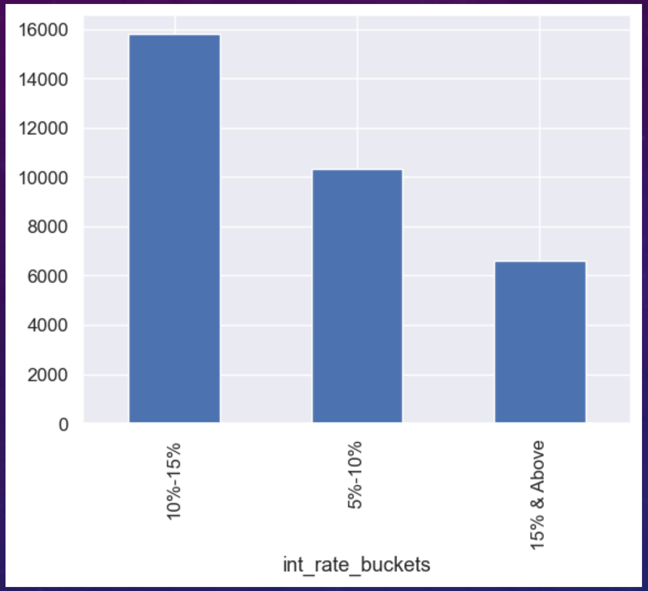


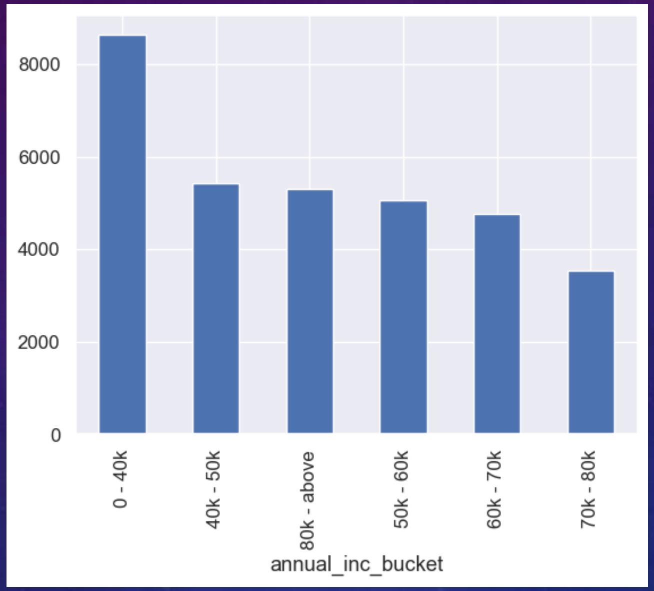


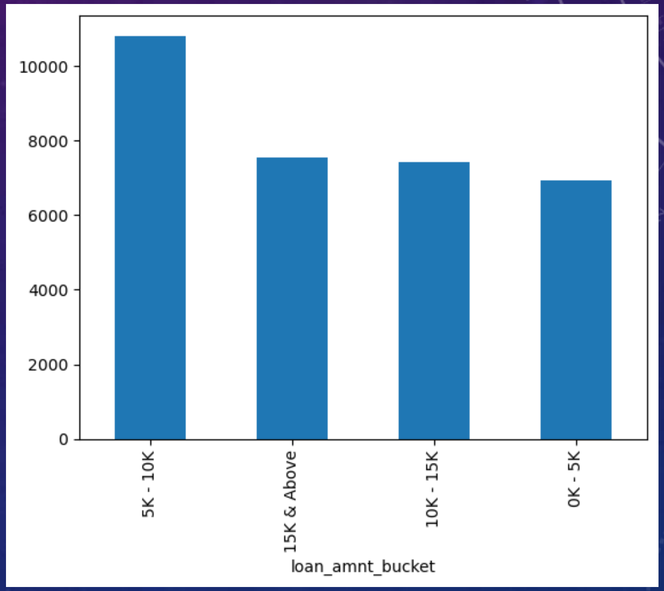
OBSERVATIONS

- Most of the applicants are from Grade B
- Most of the loan applications are of 36 months
- Maximum number of loans were availed in the month of December

UNIVARIATE ANALYSIS
QUANTITATIVE VARIABLE ANALYSIS







OBSERVATIONS

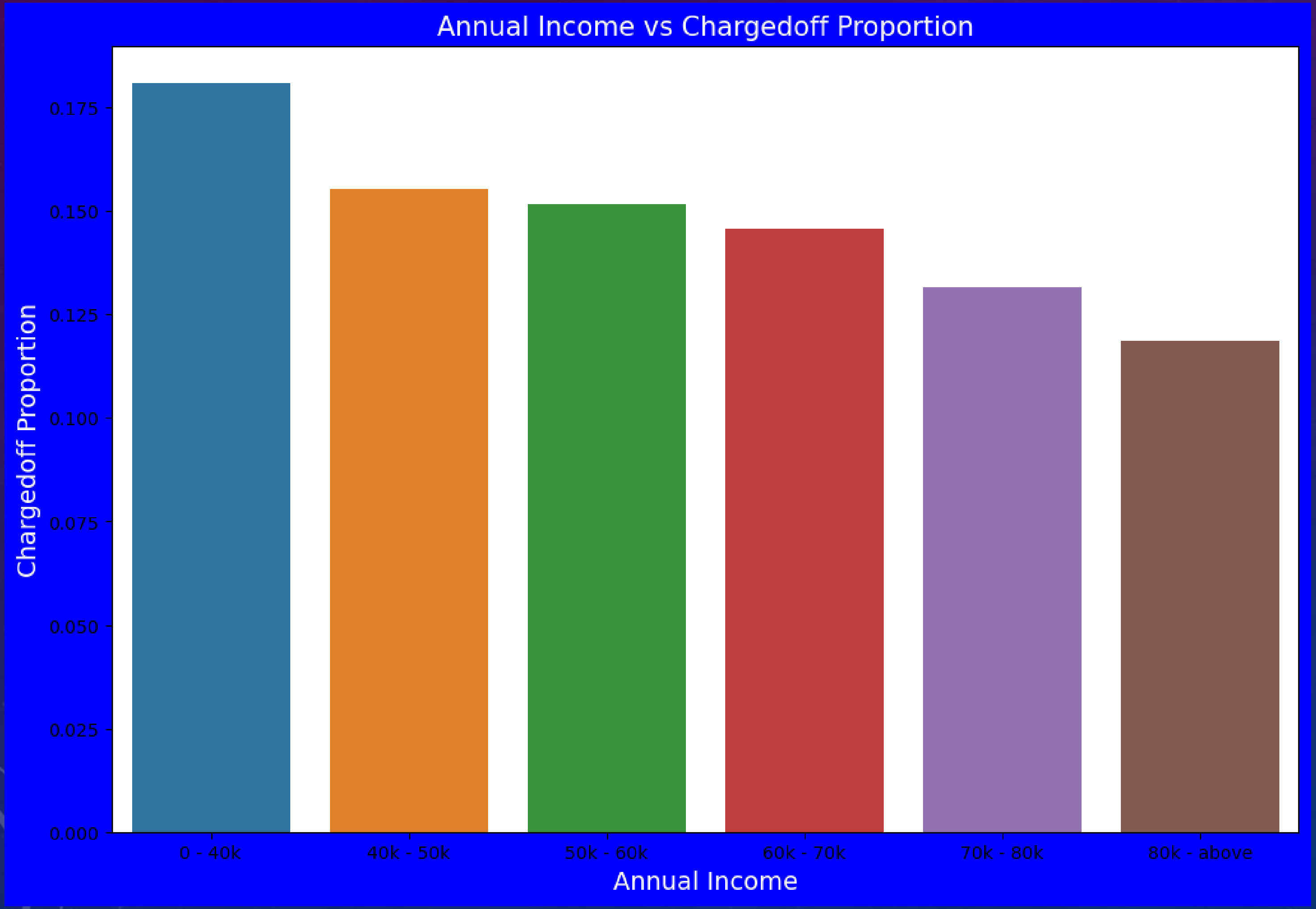
- Most of the applicants' interest rate is between 10-15%
- Most of the applicants are having annual income below 40K USD
- Maximum number of applicants applied for loan amount between 5-10K USD

BIVARIATE ANALYSIS

- Primary Attribute for Bivariate analysis is “Chargedoff_Proportion”. This value directly indicates the defaulter count (proportion)

$$\textbf{Chargedoff\_Proportion} = \frac{\text{No.of.Rows with loan\_status = ‘Charged Off’}}{\text{(No.of.Rows with loan\_status = ‘Charged Off’+ No.of.Rows with loan\_status = ‘Fully Paid’)}}$$

BIVARIATE ANALYSIS
ANNUAL INCOME VS CHARGEDOFF_PROPORTION



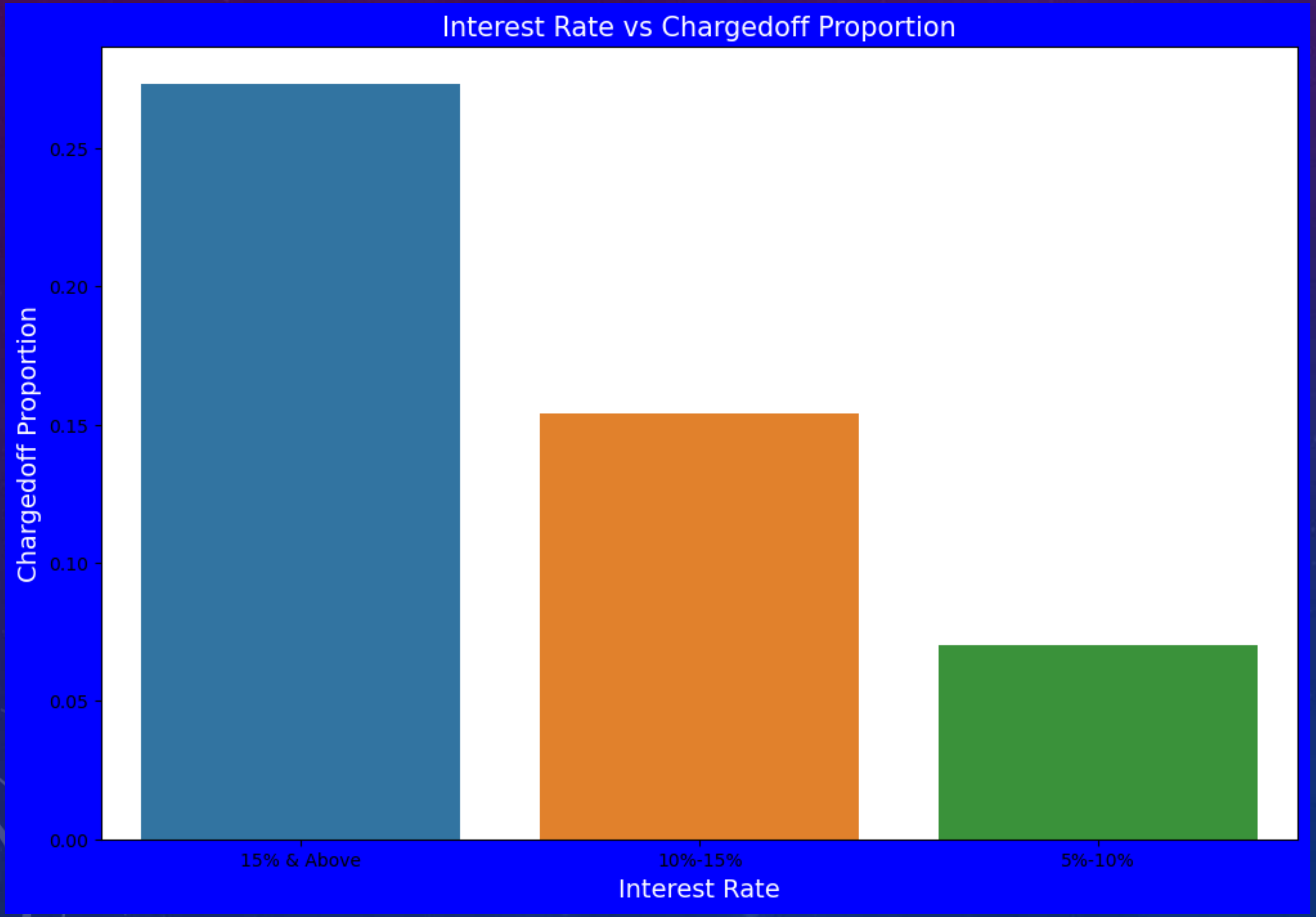
OBSERVATIONS

- Applicants whose income range up to 40K are more likely to default the loan repayment

INFERENCE

- Lending Club company should make sure to consider other attributes like house ownership grade etc while approving loans for those applicants whose income is below 40K USD

BIVARIATE ANALYSIS
INTEREST RATE VS CHARGEDOFF_PROPORTION



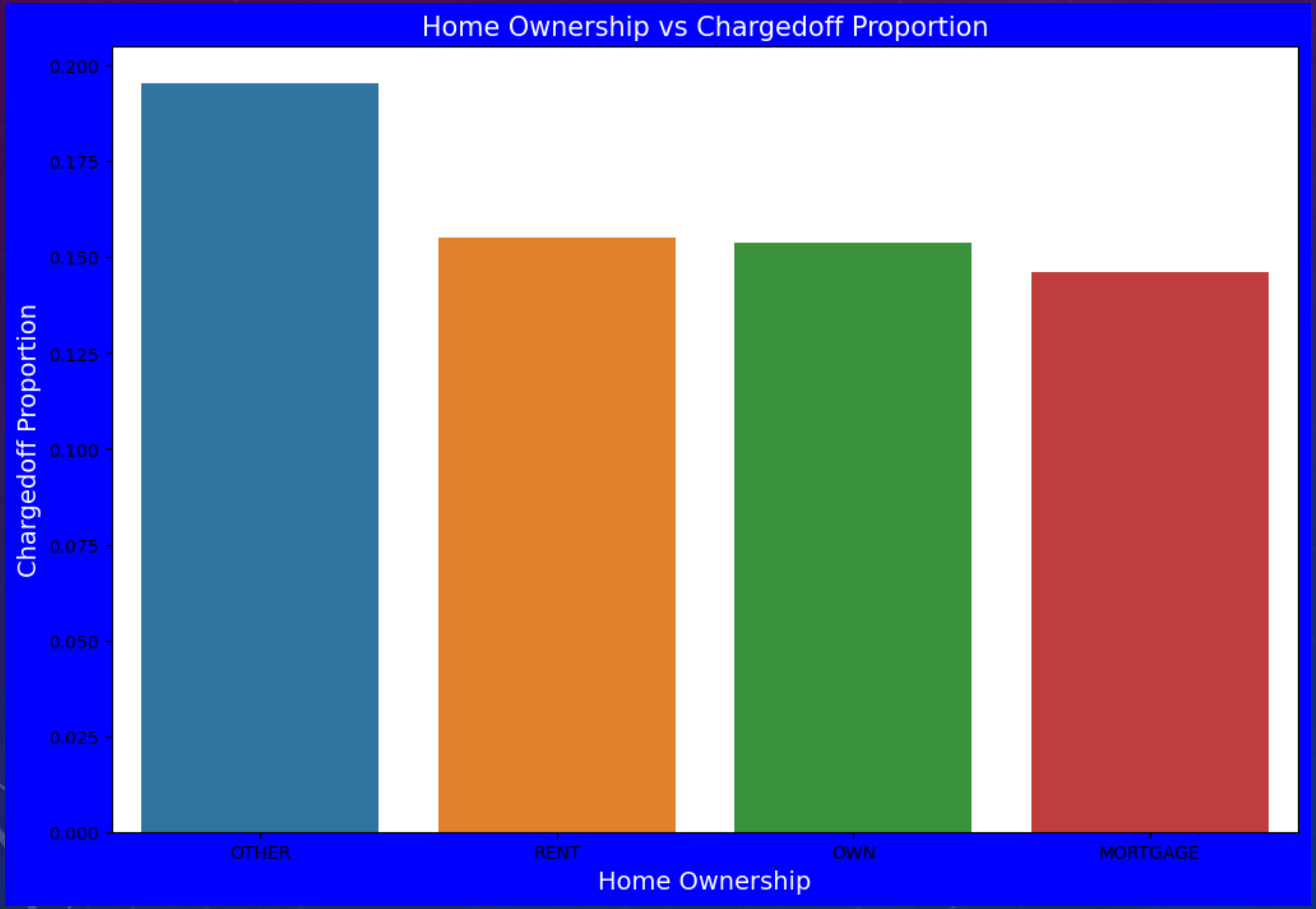
OBSERVATIONS

- More number of defaulters are observed with interest rate more than 15%

INFERENCE

- Company should reconsider the interest rates and adjust the interest rates based on the DTI score

BIVARIATE ANALYSIS
HOUSE OWNERSHIP VS CHARGEDOFF_PROPORTION



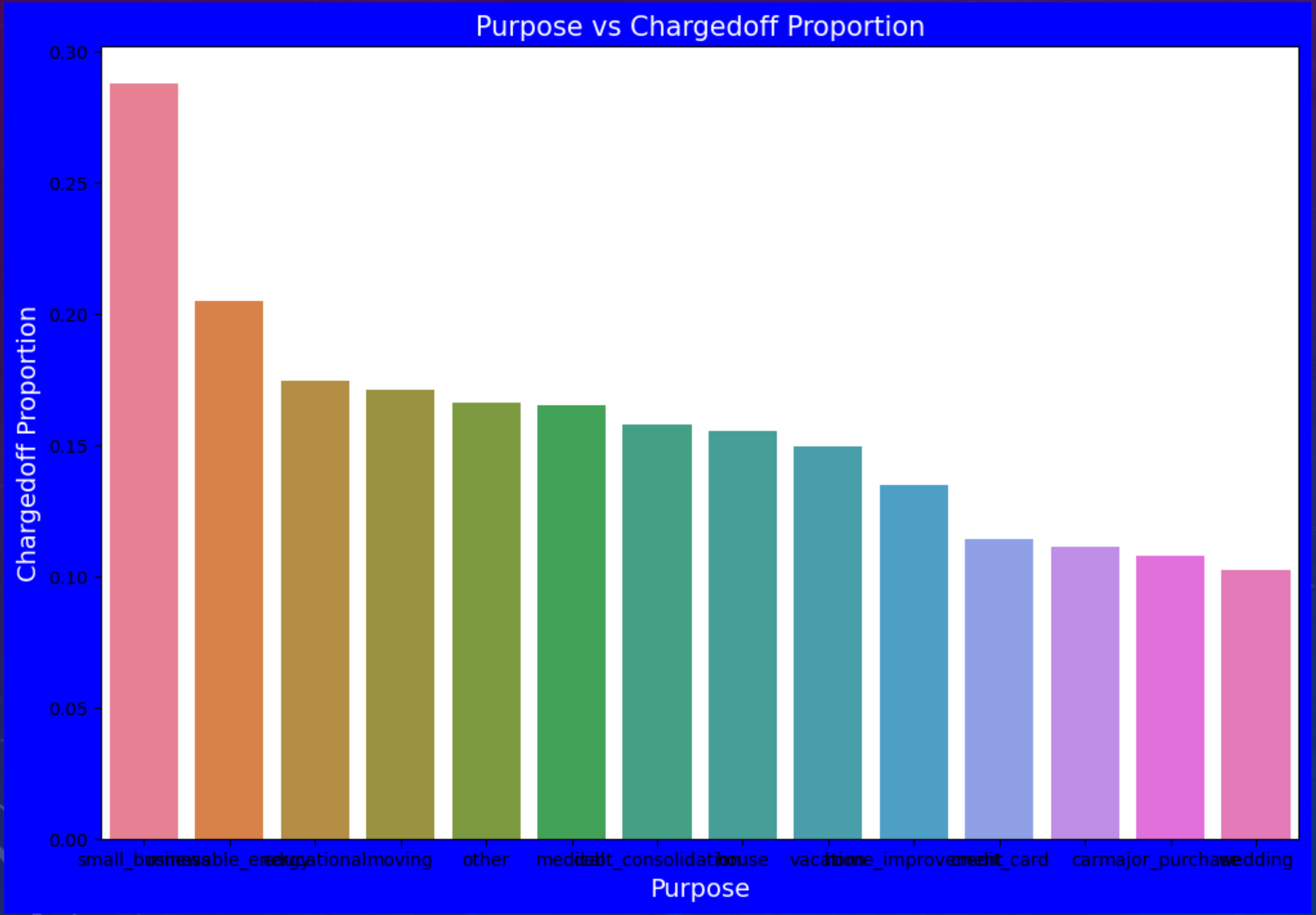
OBSERVATIONS

- Applicants who live in own house are very disciplined in repayment of loan

INFERENCE

- Company should be extra cautious and consider the house ownership as one of the key aspect while sanctioning the loans

BIVARIATE ANALYSIS
LOAN PURPOSE VS CHARGEDOFF_PROPORTION



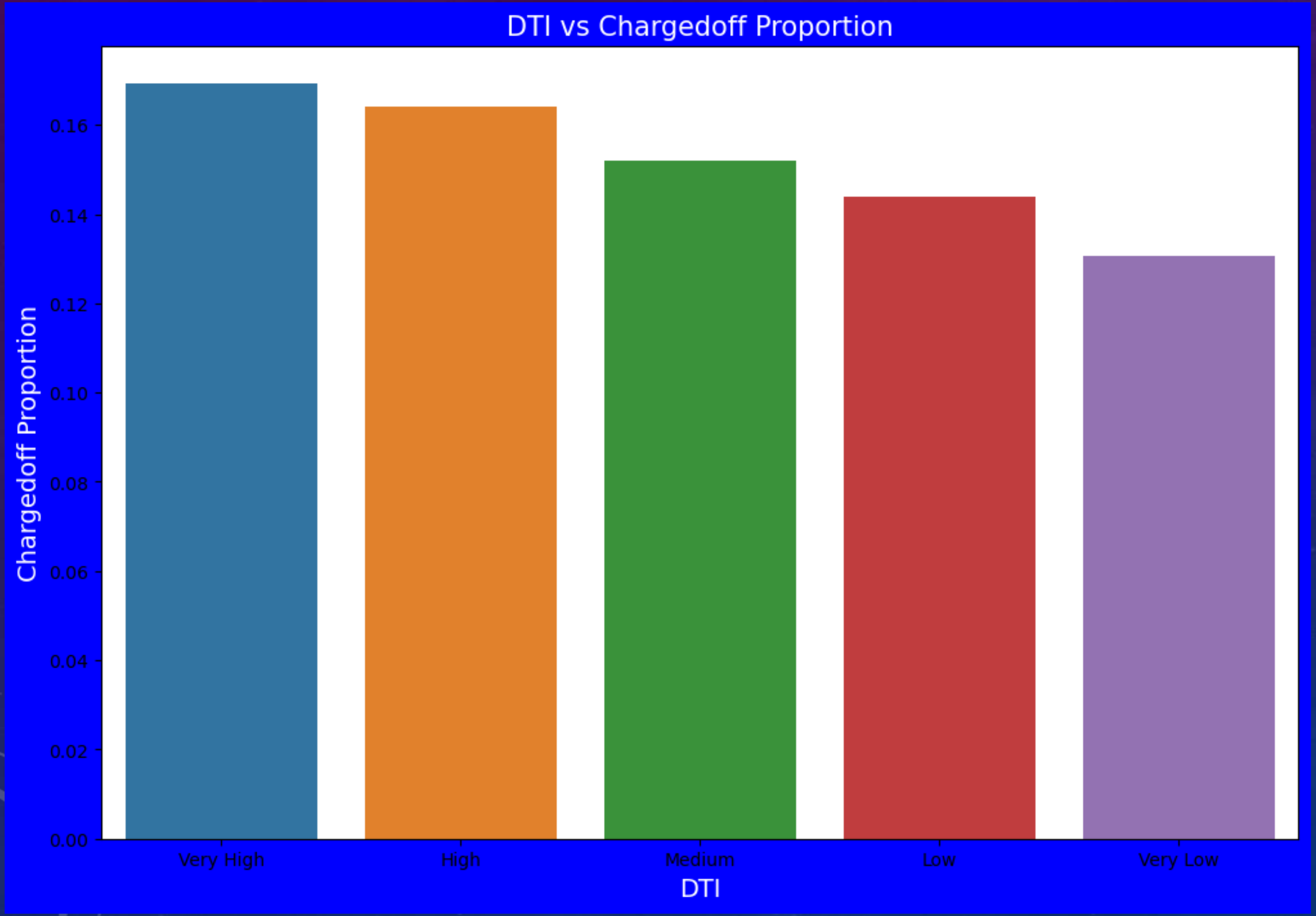
OBSERVATIONS

- Applicants who availed loan for small business purpose are more likely to be the defaulters

INFERENCE

- Company should consider less interest rates or other promotional discounts for those who are availing loans for small business purpose

BIVARIATE ANALYSIS
DTI VS CHARGEDOFF_PROPORTION



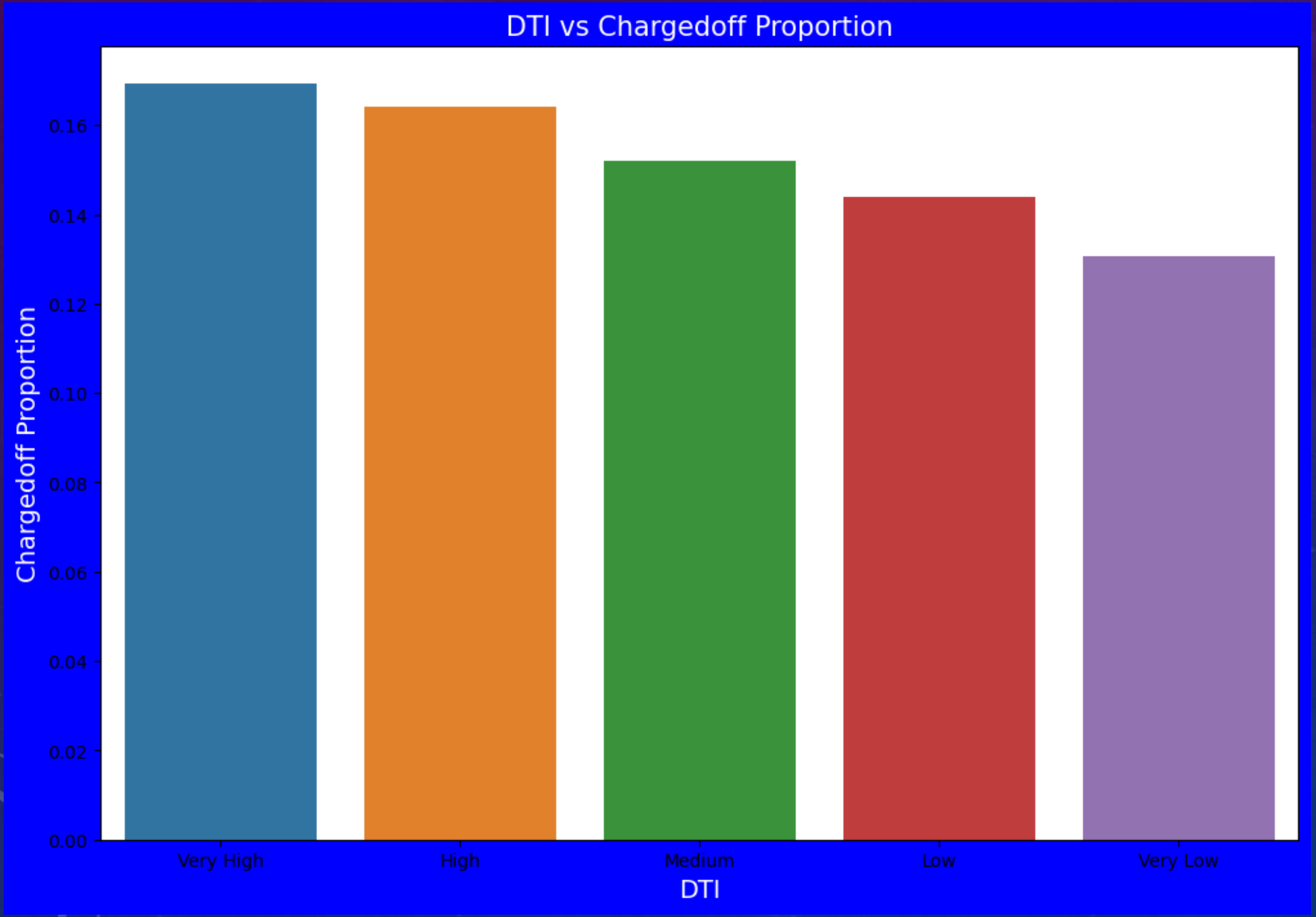
OBSERVATIONS

- Applicants who are already paying instalments for existing loans are more likely to be the defaulters

INFERENCE

- DTI indicates that the current financial situation of the applicant. So, the Company should minimize the loan amount while approving the loans for those customers whose DTI is high

BIVARIATE ANALYSIS
DTI VS CHARGEDOFF_PROPORTION



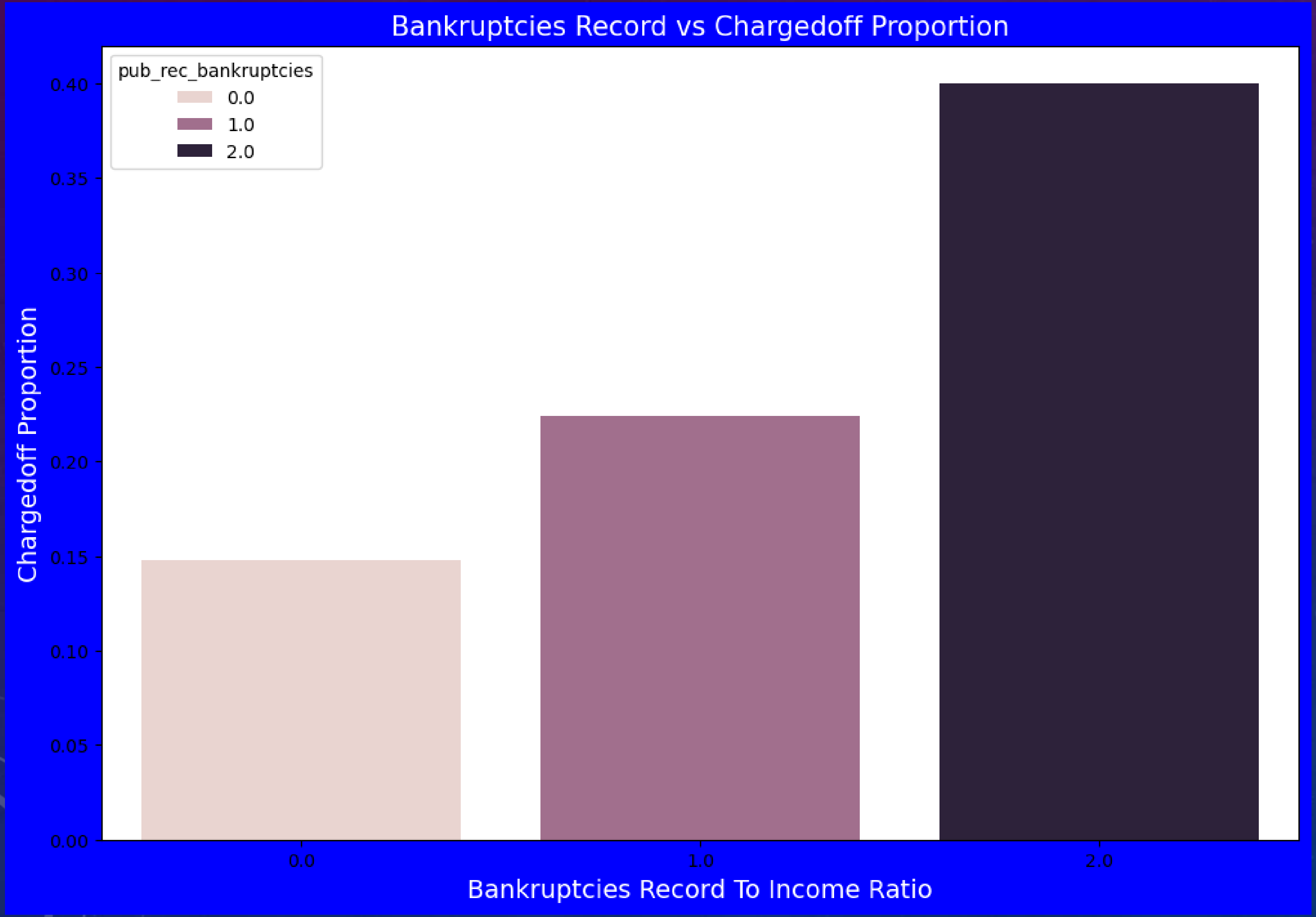
OBSERVATIONS

- Applicants who are already paying instalments for existing loans are more likely to be the defaulters
- Lower the DTI having low chances loan defaults.

INFERENCE

- DTI indicates that the current financial situation of the applicant. So, the Company should minimize the loan amount while approving the loans for those customers whose DTI is high

BIVARIATE ANALYSIS
BANKRUPTCIES RECORD VS CHARGEDOFF_PROPORTION



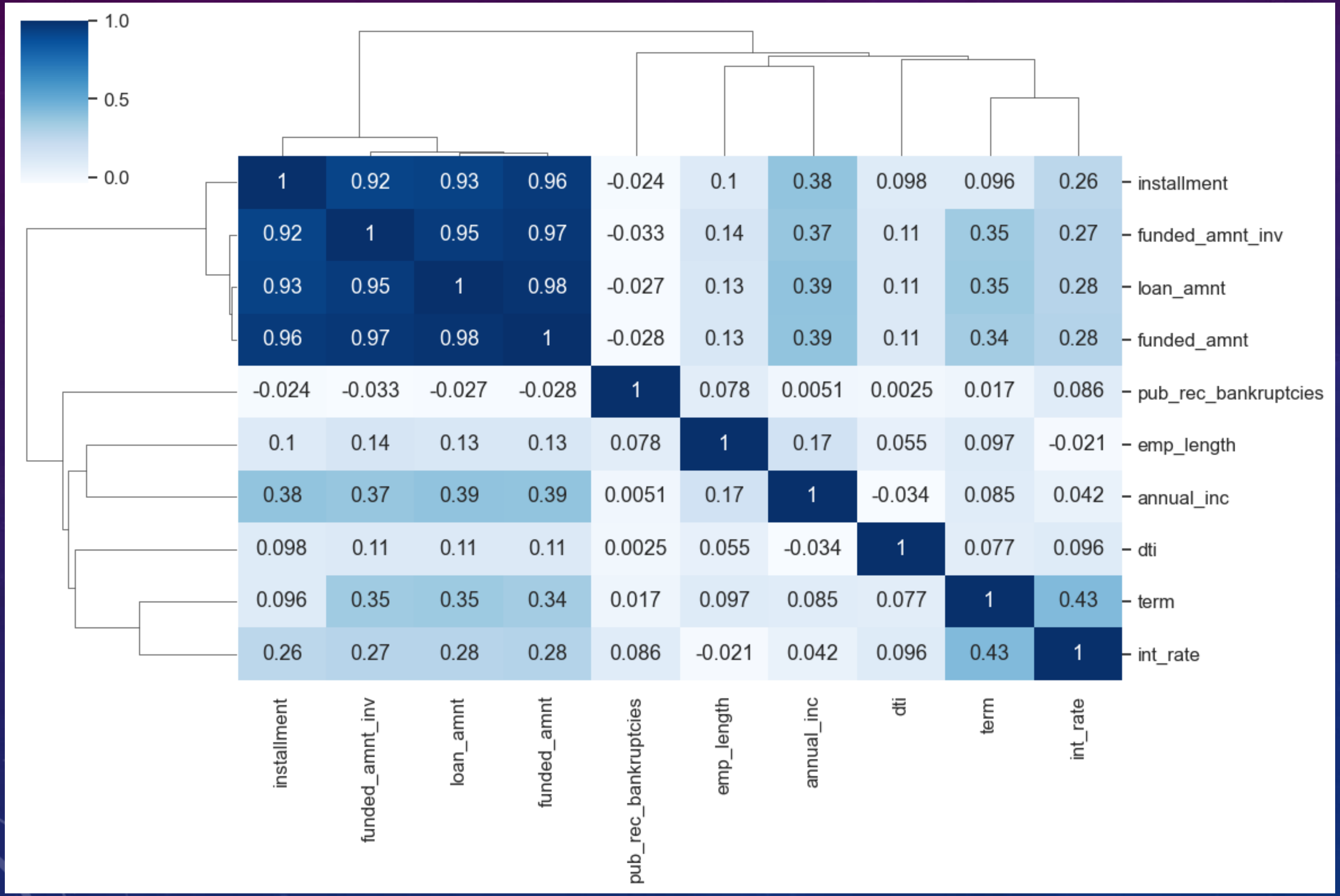
OBSERVATIONS

- Bankruptcies Record with 2 is having high impact on loan defaults
- Lower the bankruptcies, lower is the risk

INFERENCE

- Company should make sure to verify the bankruptcies records before sanctioning the loan

CORRELATION ANALYSIS



	installment	funded_amnt_inv	loan_amnt	funded_amnt	pub_rec_bankruptcies	emp_length	annual_inc	dti	term	int_rate
installment	1	0.92	0.93	0.96	-0.024	0.1	0.38	0.098	0.096	0.26
funded_amnt_inv	0.92	1	0.95	0.97	-0.033	0.14	0.37	0.11	0.35	0.27
loan_amnt	0.93	0.95	1	0.98	-0.027	0.13	0.39	0.11	0.35	0.28
funded_amnt	0.96	0.97	0.98	1	-0.028	0.13	0.39	0.11	0.34	0.28
pub_rec_bankruptcies	-0.024	-0.033	-0.027	-0.028	1	0.078	0.0051	0.0025	0.017	0.086
emp_length	0.1	0.14	0.13	0.13	0.078	1	0.17	0.055	0.097	-0.021
annual_inc	0.38	0.37	0.39	0.39	0.0051	0.17	1	-0.034	0.085	0.042
dti	0.098	0.11	0.11	0.11	0.0025	0.055	-0.034	1	0.077	0.096
term	0.096	0.35	0.35	0.34	0.017	0.097	0.085	0.077	1	0.43
int_rate	0.26	0.27	0.28	0.28	0.086	-0.021	0.042	0.096	0.43	1

STRONG CORRELATION

- installment has a strong correlation with funded_amnt, loan_amnt, and funded_amnt_inv
- Term has strong correlation with interest_rate
- Annual_inc has strong correlation with loan_amount

WEAK CORRELATION

- DTI & emp_length has weak correlation with most of the other fields

NEGATIVE CORRELATION

- pub_rec_bankrupticies has negative correlation with almost all the other fields
- DTI has is negatively correlated with annual_inc

PROPOSALS

- Thorough assessment for High loan amounts (> 15K USD)
- Careful Consideration for those who are taking loans for Debt Consolidation purpose
- Dynamic interest rates : Consider having flexible interest rates based on the DTI
- Home Ownership : Minimize the loan approval process and make it easy to get loans for those applicants who owns a house.
- Annual Income consideration : introduce maximum affordable capping of loan amount for those applicants whose annual income < 40K USD

REFERENCES

Python packages & Technology	Reference Links
Python	https://www.python.org/
Mathplotlib	https://matplotlib.org/
Numpy	https://numpy.org/
Pandas	https://pandas.pydata.org/
Seaborn	https://seaborn.pydata.org/
Jupyter notebook	https://jupyter.org/