UPGRAD PROJECT
III-T BANGLORE

EDA

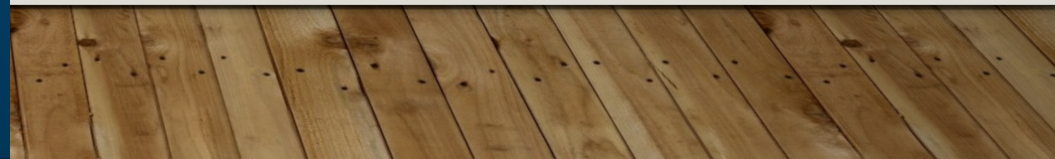
OF

LENDING CLUB DATA

(LENDING CLUB CASE STUDY)

Jojo Jacob

Cohort: EPGP ML C 37



SCOPE

Introduction

Business Objective

Data Cleaning

- Missing Values
- Data Types
- Redundant Values
- Outlier Detection

EDA

- Univariate Analysis
- Bivariate Analysis

Conclusion

- Summary.
- Recommendations

INTODUCTION

- Lending Club is a financial company that issues loans and based in USA
- It also deals with peer to peer lending enabling investors gain profit by lending money to applicants.
- Lending club wants to analyse the data of its customers to identify potential defaulters and thereby avoiding the risks of giving loans to them.
- They also want to maximise their profits.

Risk

- Risk of Loosing business and profit by not giving loans to potential customers
- Risk of giving bad loans and loosing capital

PROJECT APPROACH

- **The project shall be carried out in the steps enumerated below.**
  - **Import Libraries**
  - **Create Custom Function**
  - **Loading Data**
  - **Cleaning Data**
  - **Univariate Analysis**
  - **Bivariate Analysis**

Data Cleaning

- Missing Values
- Analyse Data Types
- Remove Outliers
- Remove Redundant Values

Univariate Analysis

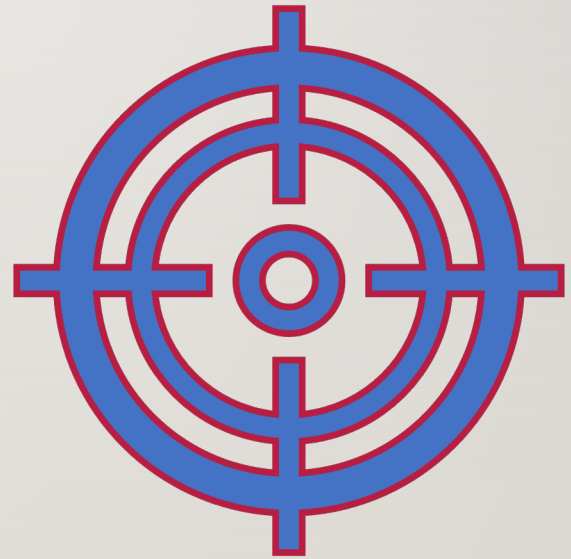
- Distribution
- Bar Charts
- Count Plots
- Histplots

Bivariate Analysis

- Scatter Plots
- Bar Charts
- Correlation Analysis
- Heatmap

BUSINESS OBJECTIVE

- To identify the variable(features) that impact the chances of default on loans by customers



CUSTOM FUNCTIONS

missing_data

- To find the missing values in a data frame and return a both quantity and percentage.

col_check

- To check the columns for unique values and value counts

drop_col

- Drop columns with missing values and display the statistics of the data set

drop_rows

- To drop rows with missing values and display dataset statistics .

univariate_bar_plot_1

- Plots the variable against percentage of defaulted loans

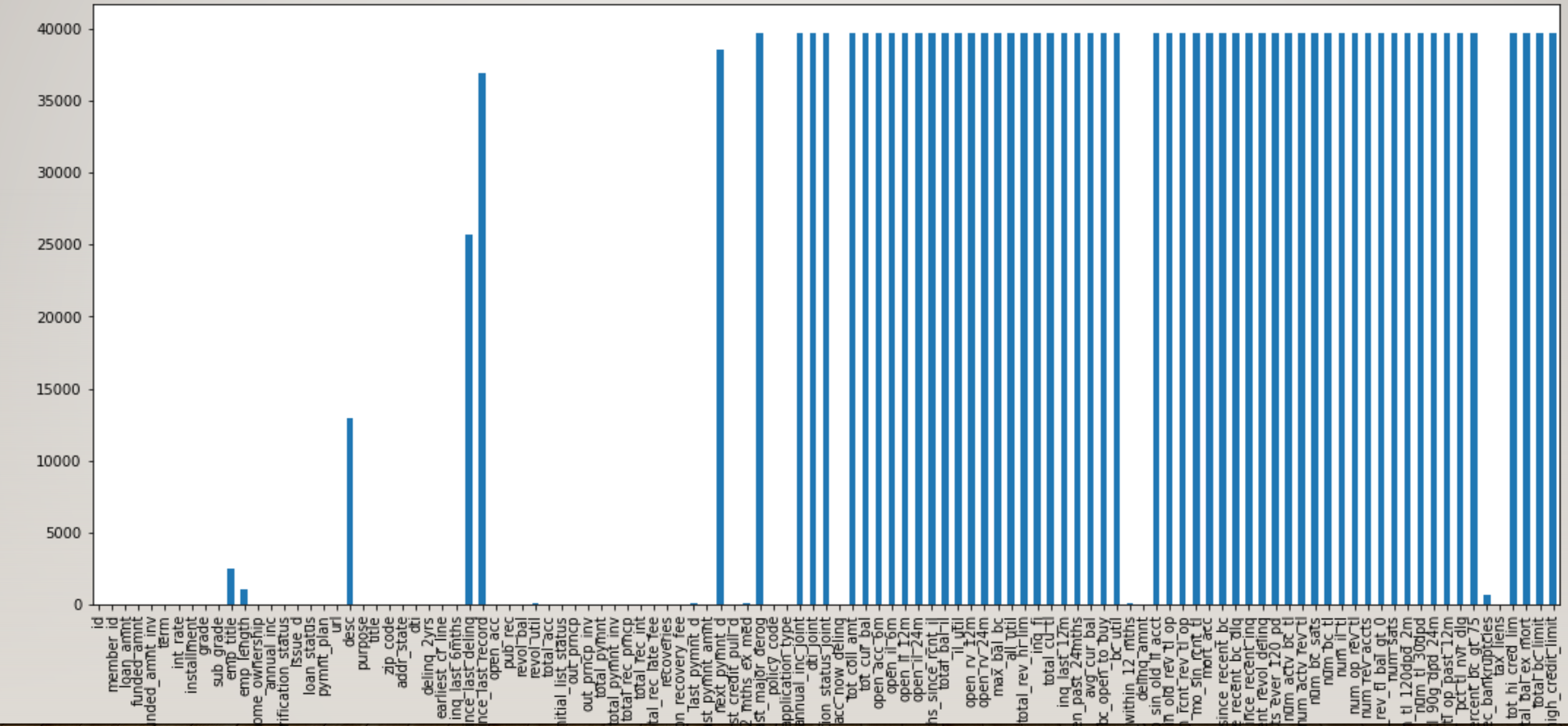
DATA LOADING

- Loading Data
  - Data Set uploaded to GitHub Repository.
  - Path to the dataset is then used in pandas to create a data frame.
- Basic Characteristics
  - Columns : 111
  - Rows : 39717
- Data Types
  - Object : 24
  - Float : 74
  - Int : 13
  - Boolean : 00

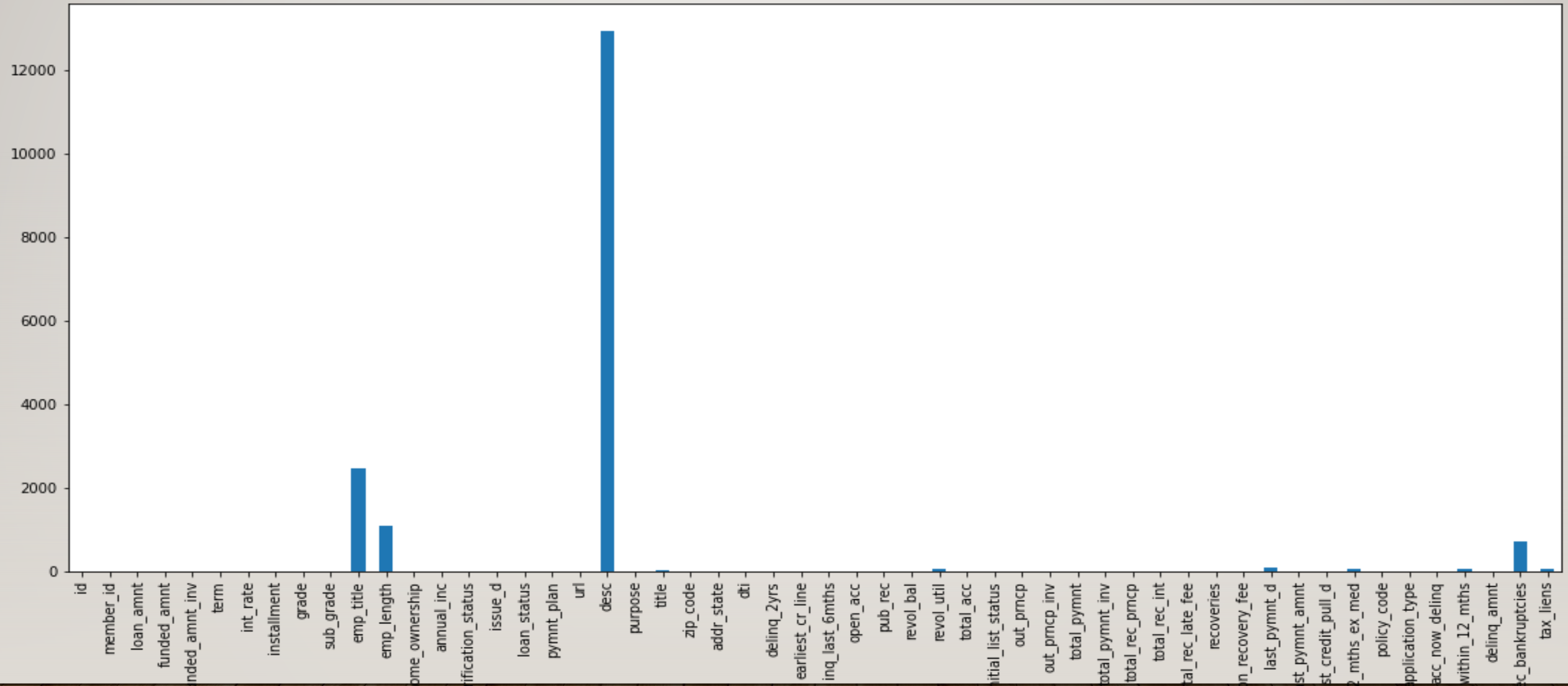
DATA CLEANING

- Missing Values
  - 57 columns have more than 40% data missing
  - Path to the dataset is then used in pandas to create a data frame.
  - Total Rows deleted 1892
  - Total Columns deleted 61 including 57 above
  - Balance data : Rows : 37825, Columns : 50
- Unique Values
  - Checked for single unique values in the columns post dealing with missing values
  - 07 columns have only one unique value and hence they were deleted
  - Total columns remaining is 43
- Redundant Data
  - 04 columns has identity variable
  - 15 columns contain data not useful for analysis
  - In loan status, current loans do not provide the desired insight and hence deleted

MISSING DATA: INITIAL DATA



MISSING DATA: AFTER REMOVING COL WITH MORE THAN 40%



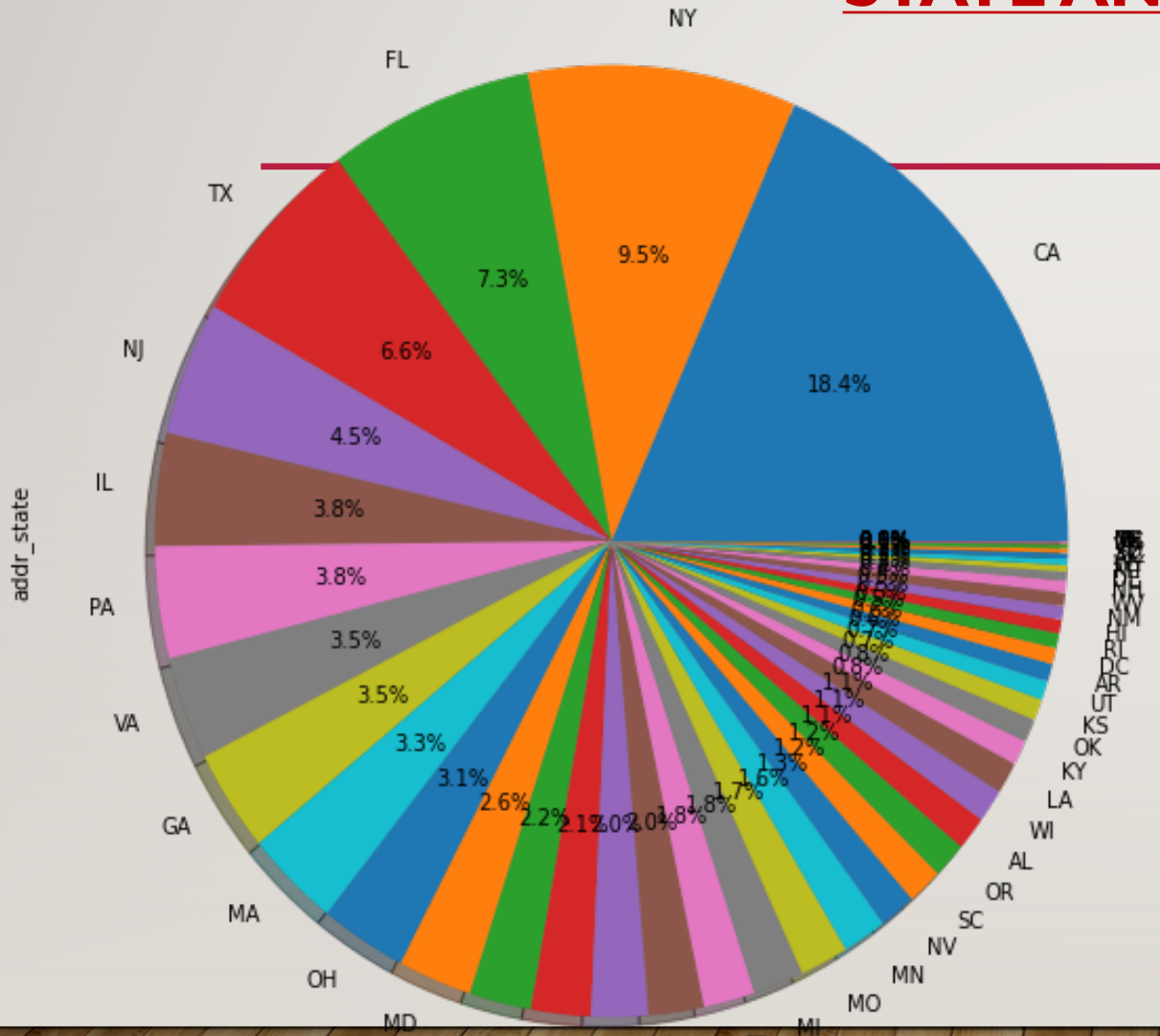
DATA CLEANING

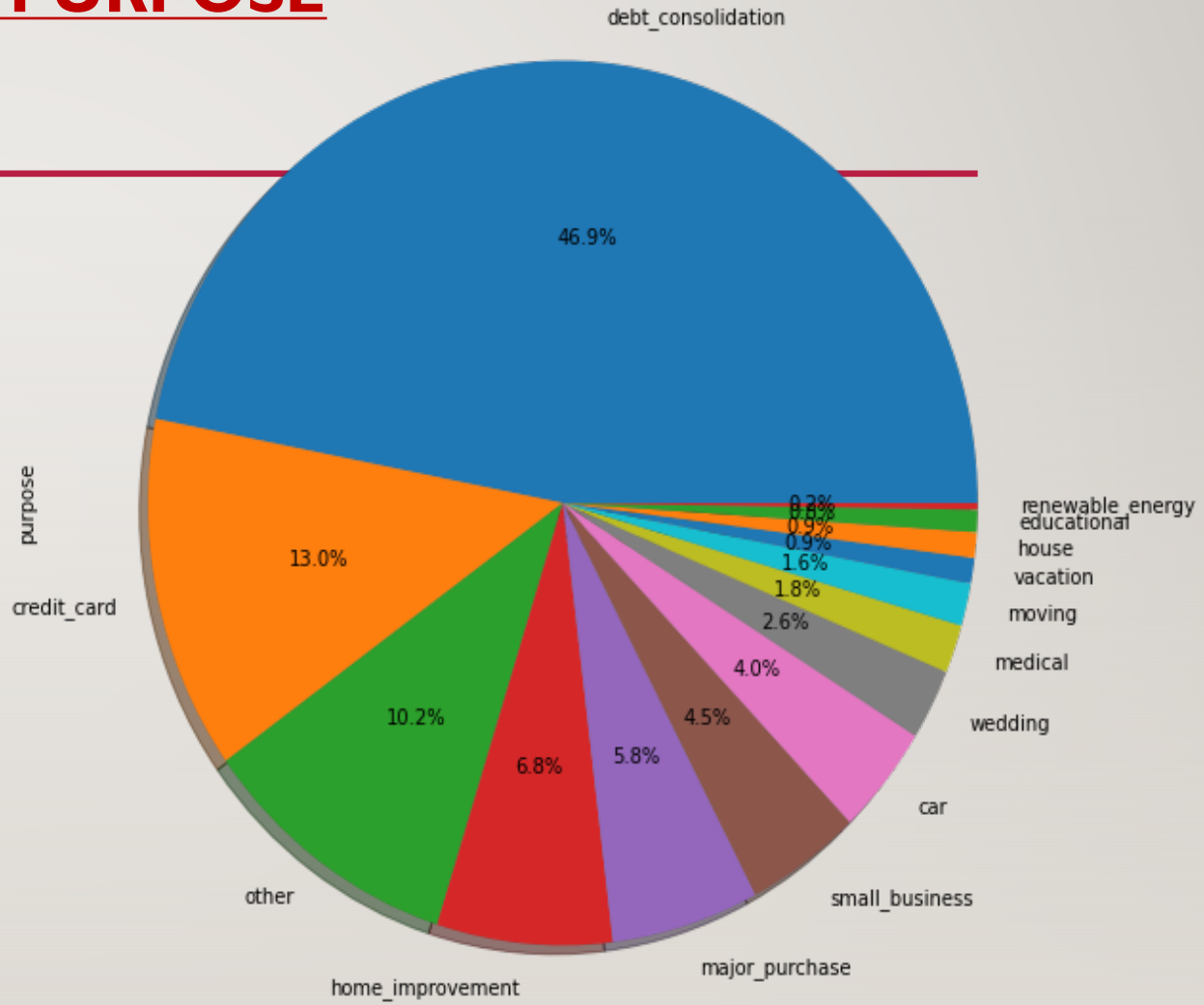
- Dates
  - Two date columns, both are initially object type
  - Converted both to datetime
  - Derived Columns
    - Months and year from date columns
- Outlier Detection
  - Outliers were detected in most of the numerical columns
  - Data above 95% limit were removed from following columns with major outliers
    - Annual Income
    - Open Accont
    - Total Account

- Binning of Data
  - Create bins with appropriate ranges for
    - Annual Income
    - Interest Rate
    - Debt to Income
    - Installments

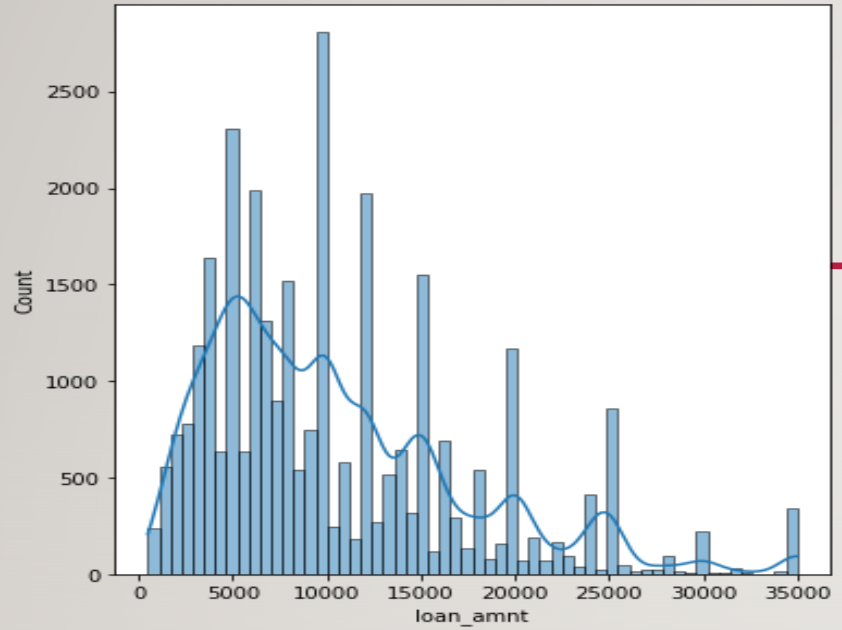
DISTRIBUTION OF DATA

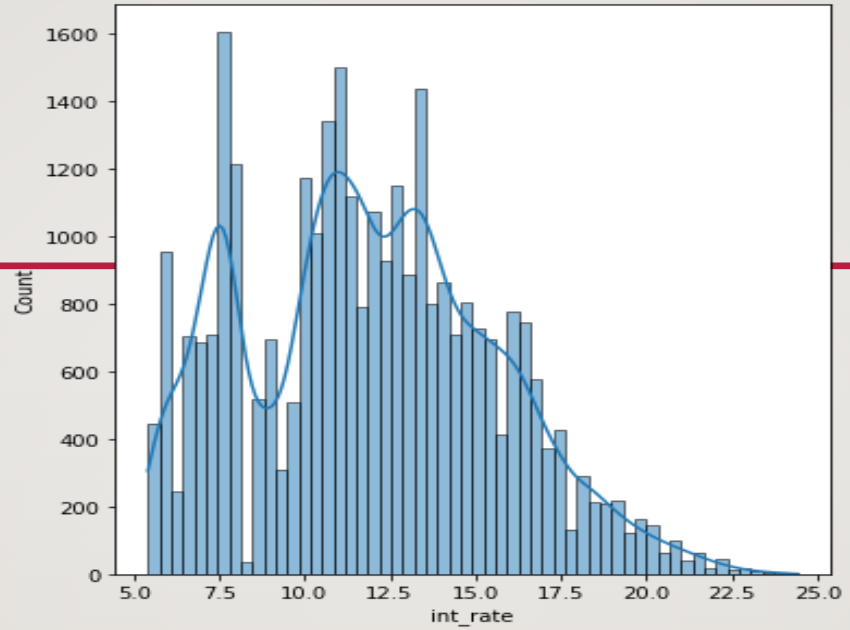
STATE AND PURPOSE

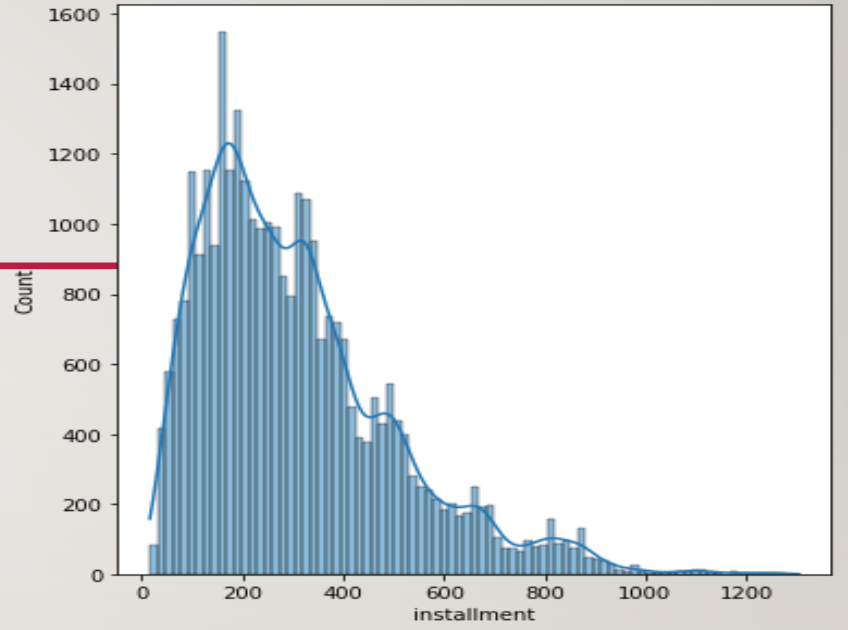


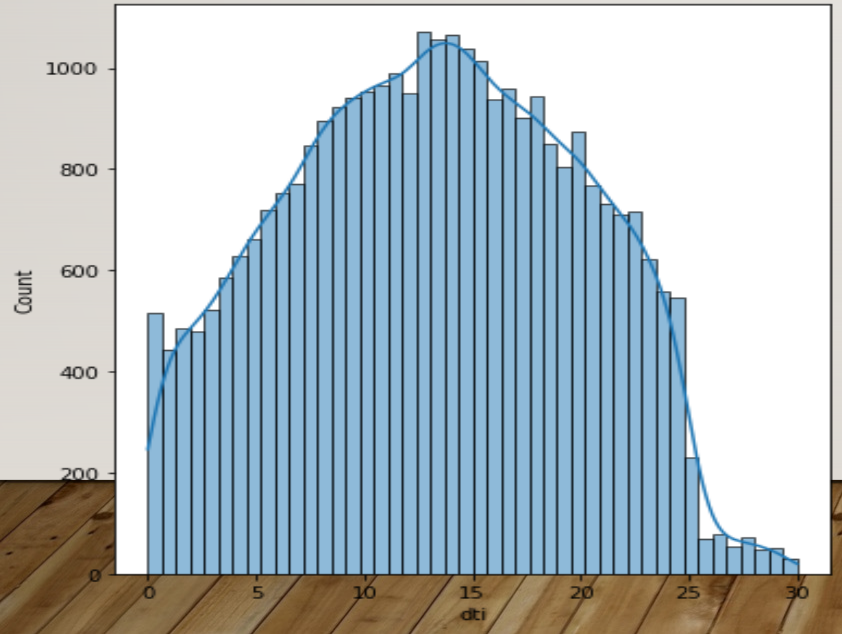


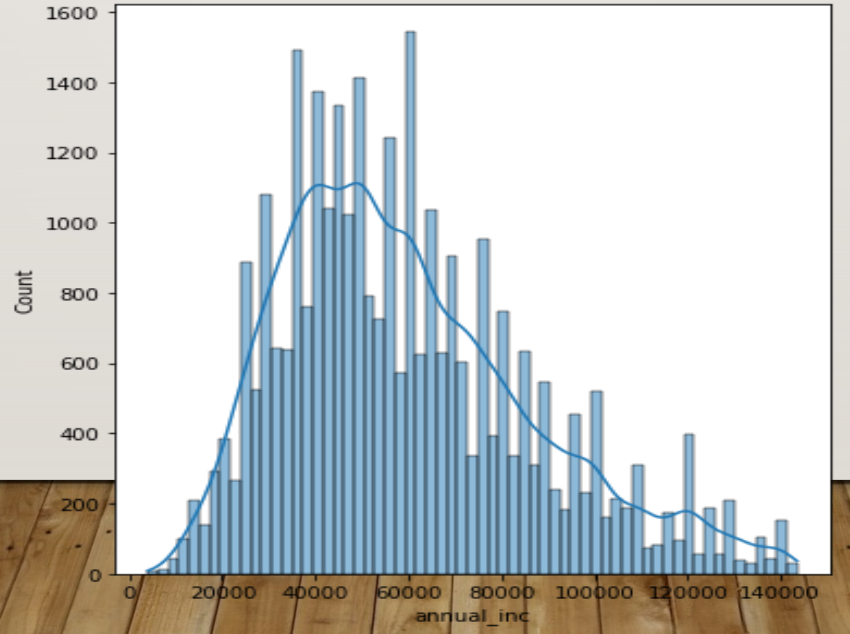
DISTRIBUTION OF DATA

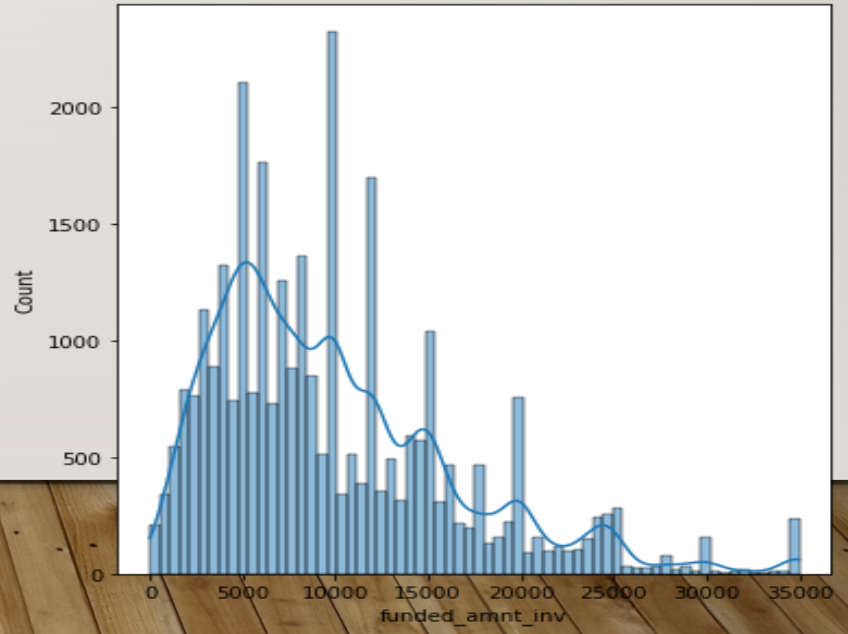










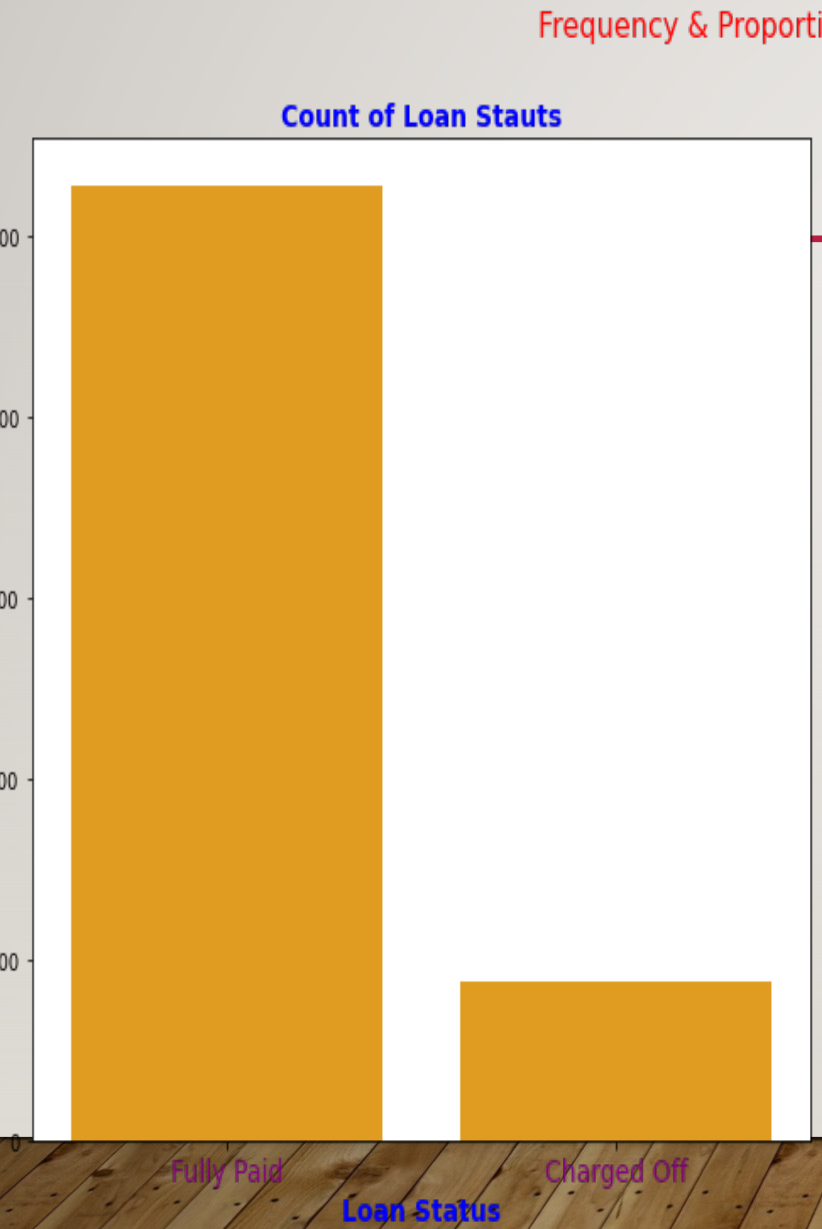


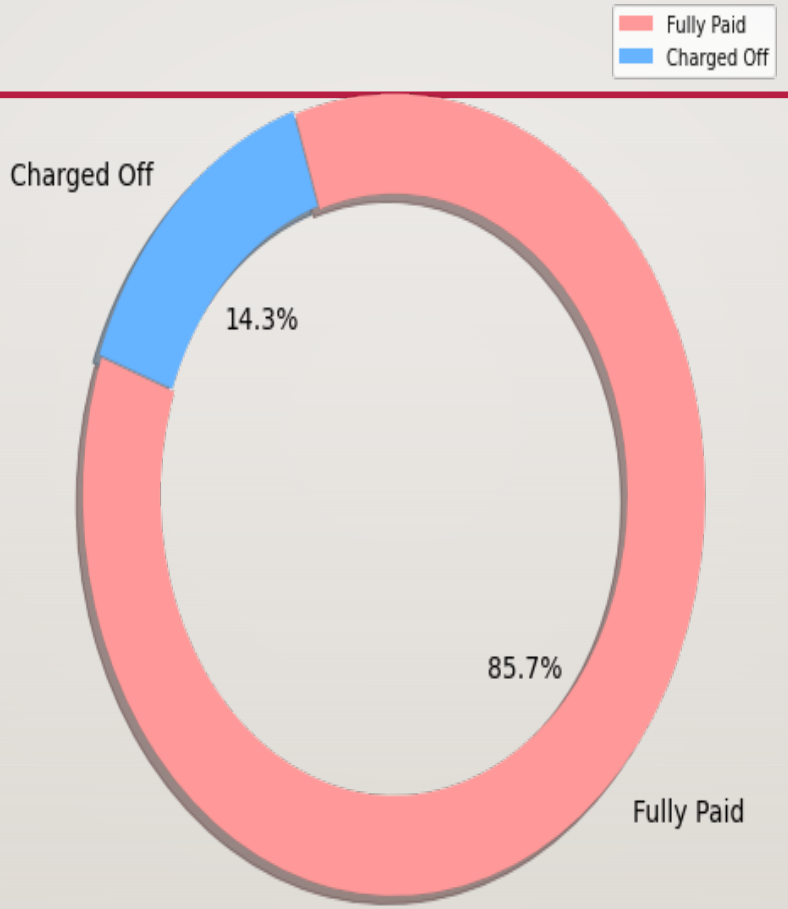
OBSERVATIONS: DISTRIBUTION OF DATA

- **Loan amount**
  - Max number of loans are in the range of 10000 and dips down towards 35000.
  - Spikes as expected can be seen in rounded figures of 10,000 and 5000's.
  - The distribution is non uniform with heavy right skew.
- **Interest Rate**
  - Definite spikes around 6% and 10 % is visible.
  - Higher interest rates are probably for riskier loans
  - Some of the rates between 6 and 10% could be for specific customers and higher loans.
  - Distribution is roughly uniform with a slight right skew.
- **Instalments**
  - Instalment generally follow the trend of loan amount.
  - Peak is seen in the vicinity of 200
  - Spikes in instalments seems to be commensurate with both interest rates and loan amount.
  - **Debt to Income Ration****
  - Peaks between 10 to 15%
  - See a trend of sharp increase and then sharp decrease around the peak.

- **Debt to Income Ration**
  - Peaks between 10 to 15%
  - See a trend of sharp increase and then sharp decrease around the peak.
- **Annual Income**
  - The distribution is near uniform.
  - People with higher income are less likely to apply for loans and hence their count will be reduced
  - As seen, its the middle income group who have both desire and capability for loans for improving the lifestyle.
- **Investor Funded Amount**
  - Its seen that investors mostly fund lesser value loans
  - Higher value loans in some cases for particular applicant could also be funded by investors.

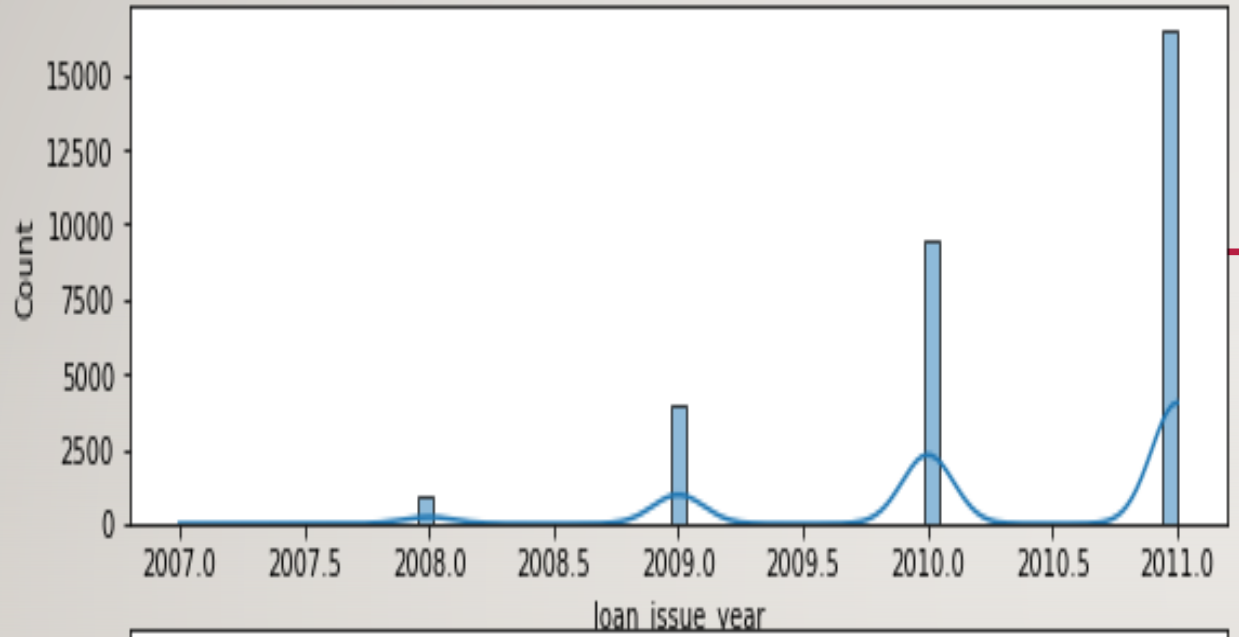
DISTRIBUTION OF DATA

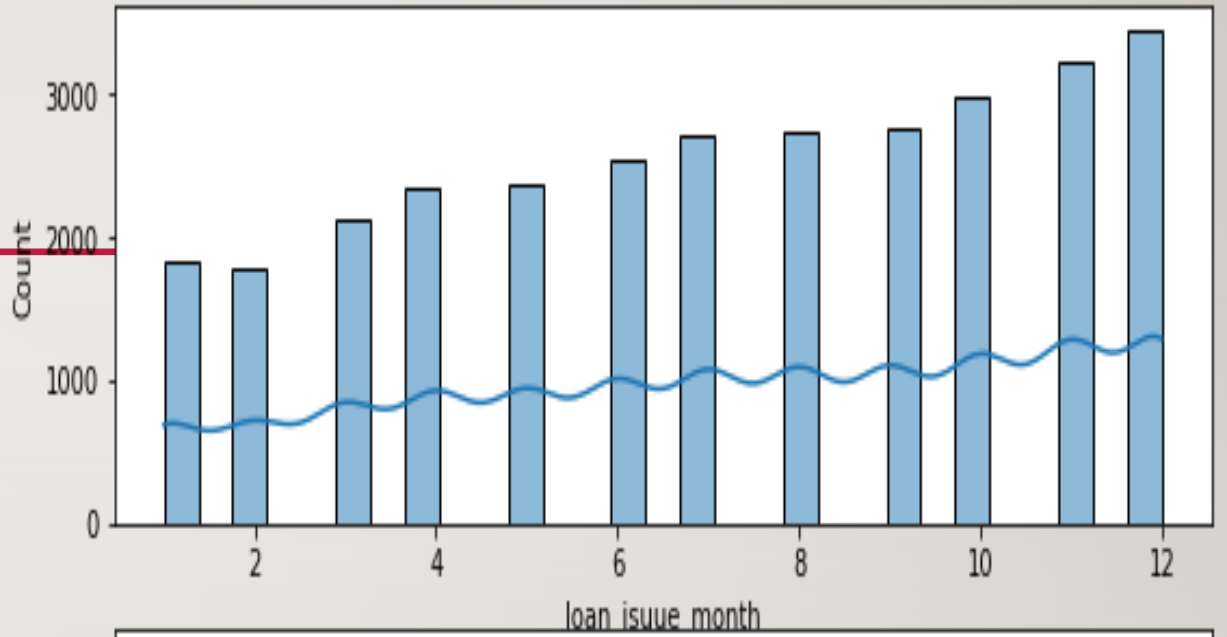


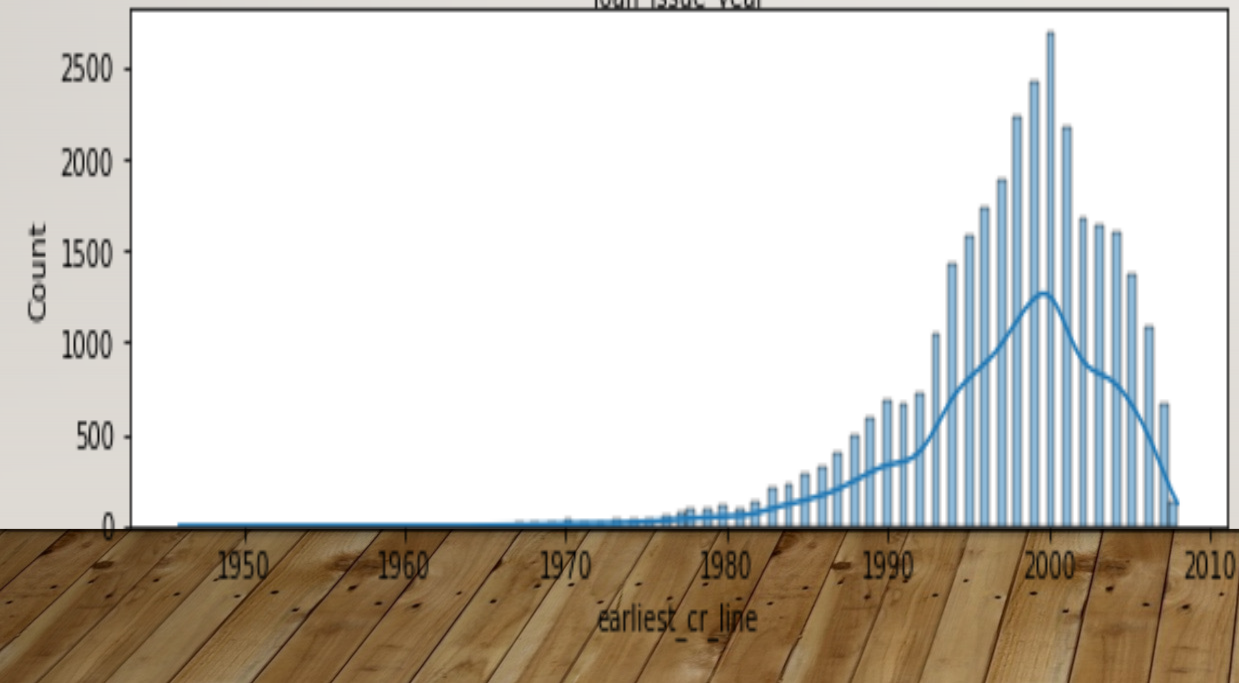


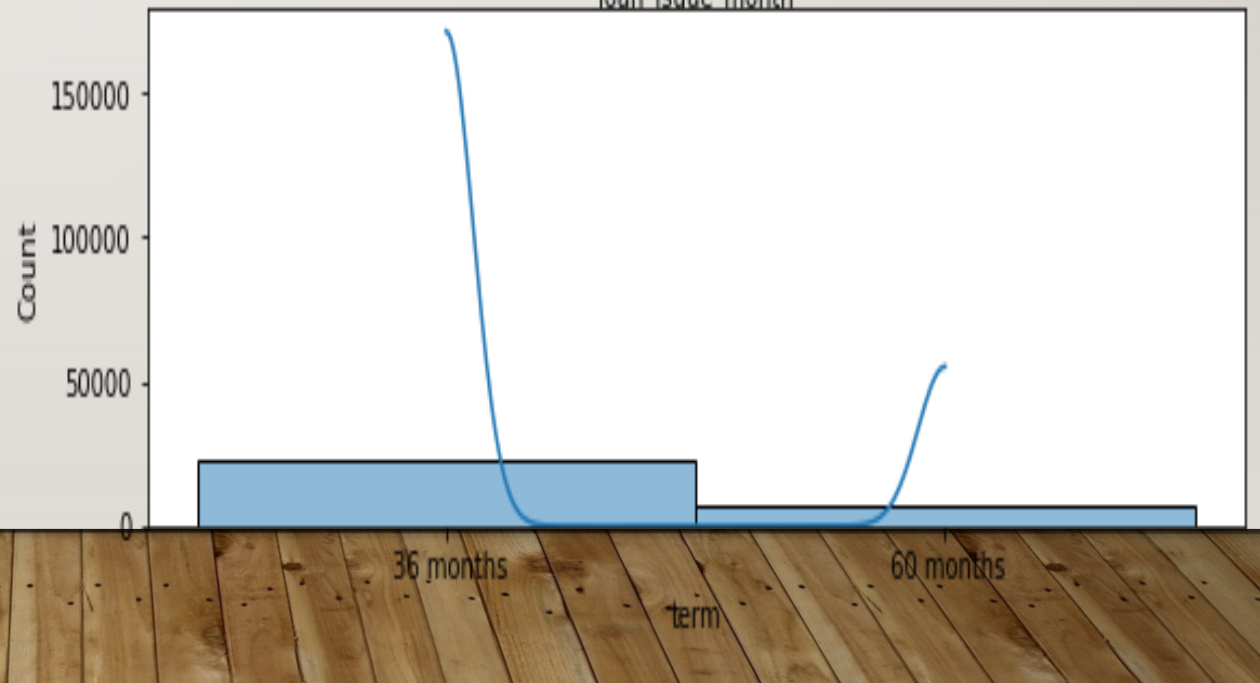
- Only 14.2 % of the loans had to be charged off.
- However, depending upon the value of the loans, the NPA can be very high or bearable.

DISTRIBUTION OF DATA





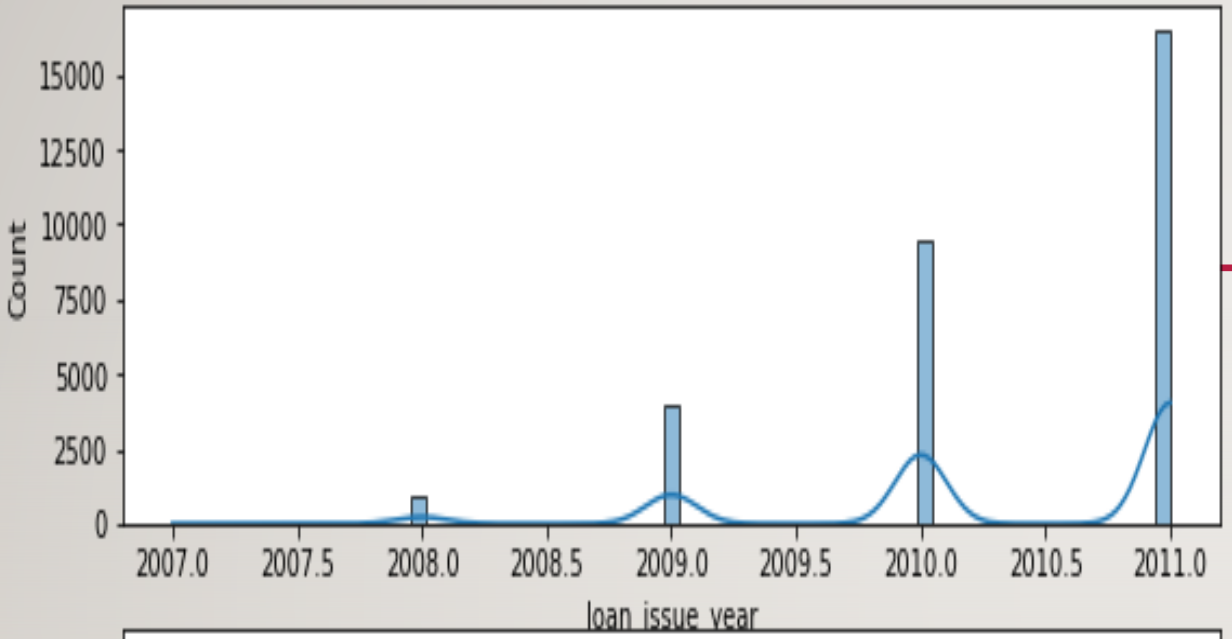


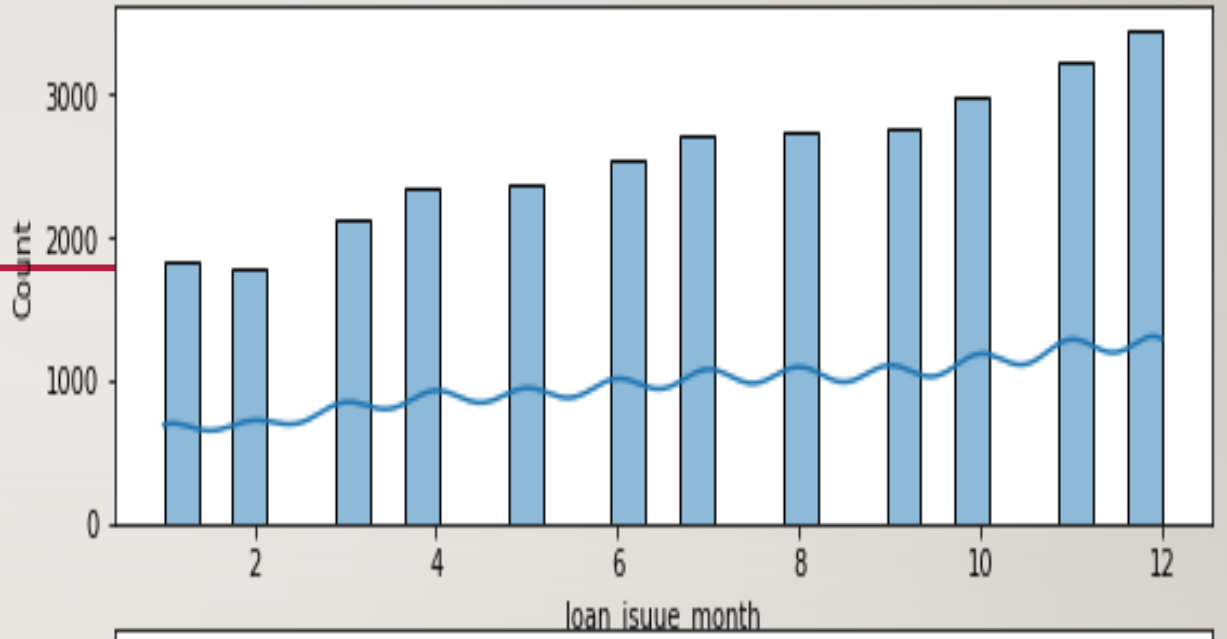


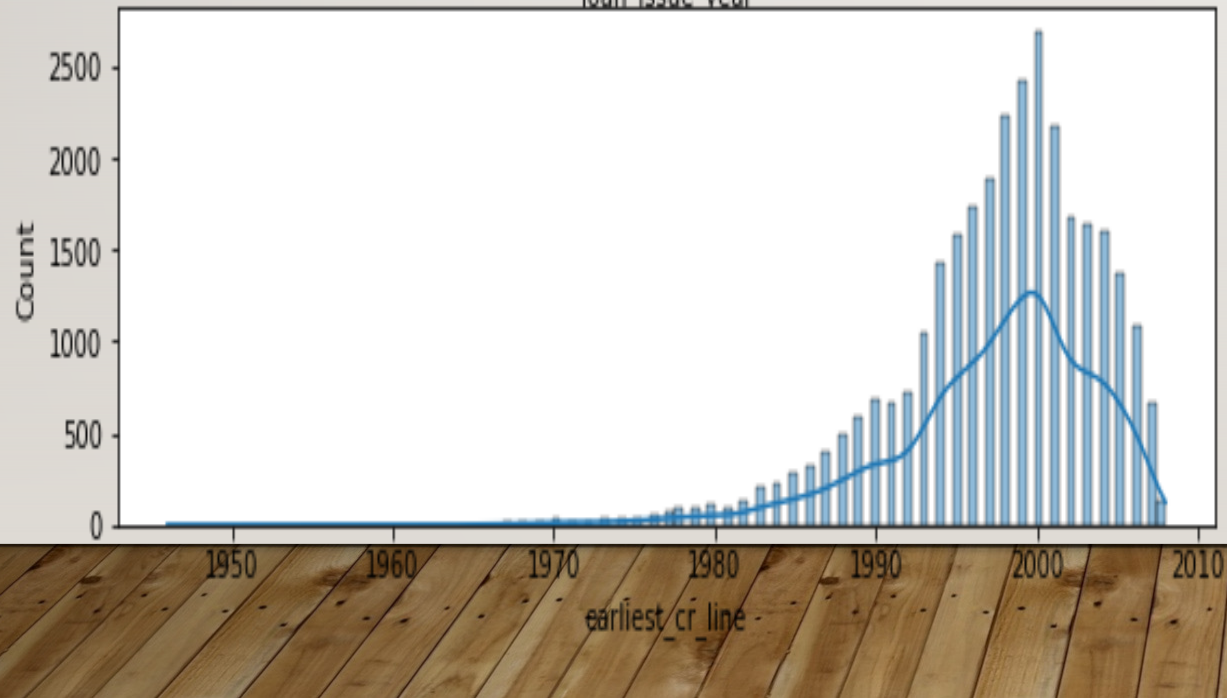
OBSERVATIONS: DISTRIBUTION OF DATA

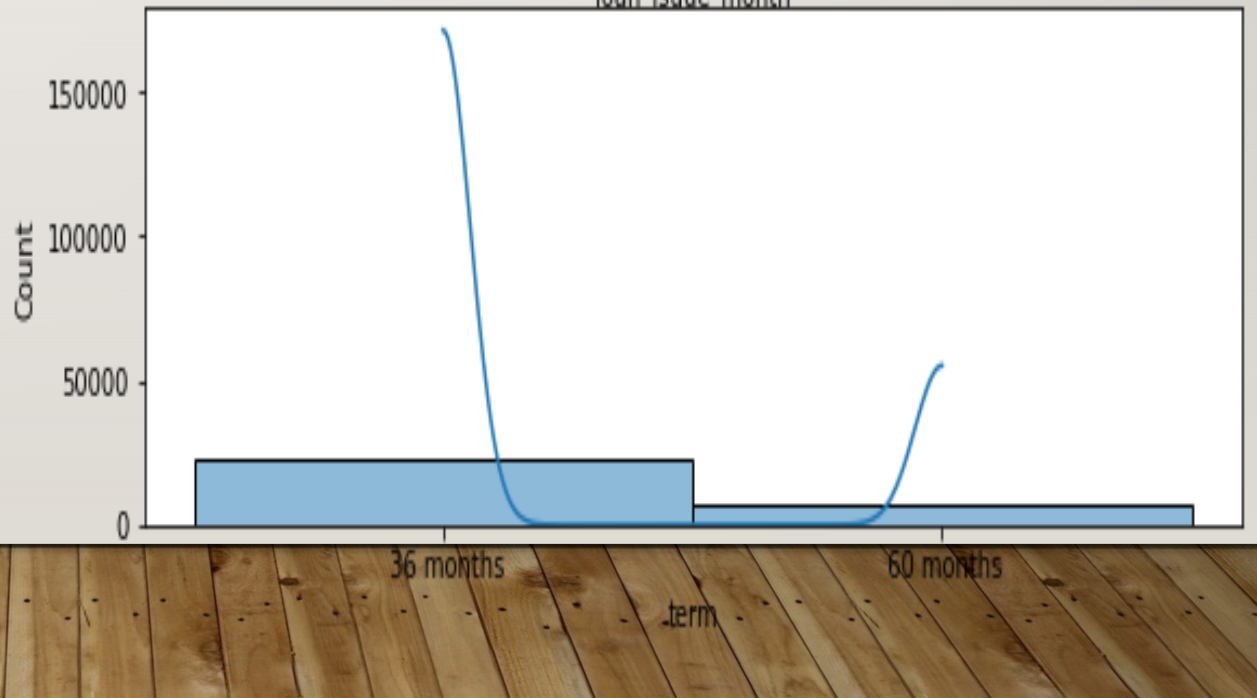
- Preferred loan term is 36 months
- Loans charged off are almost similar for loan terms
- B grade has the highest no of loans.
- A Grade loans have the least no of charged off loans in comparison to total loans
- Maximum loans are taken for debt consolidation followed by credit cards payment
- Max loans charged off are for debt consolidation.

DISTRIBUTION OF DATA





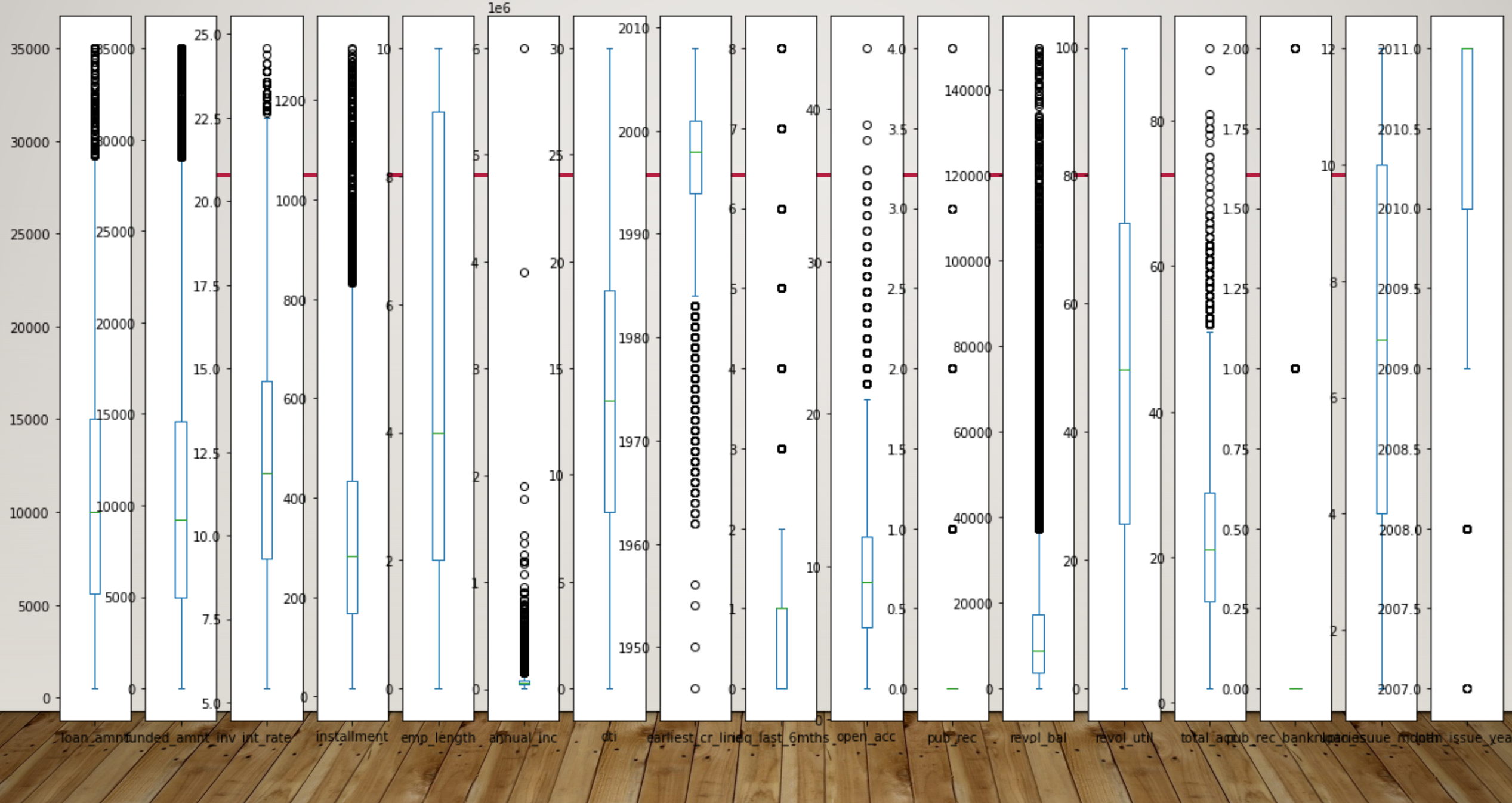




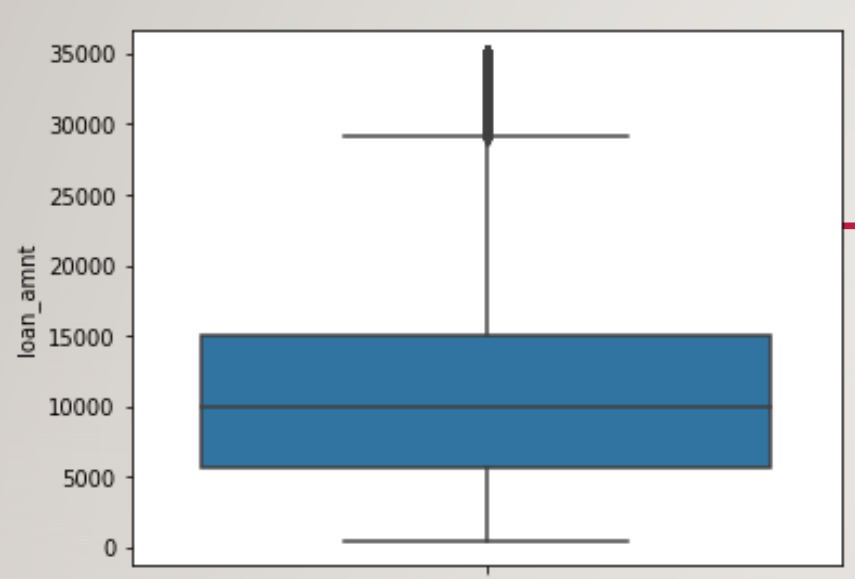
OBSERVATIONS: DISTRIBUTION OF DATA

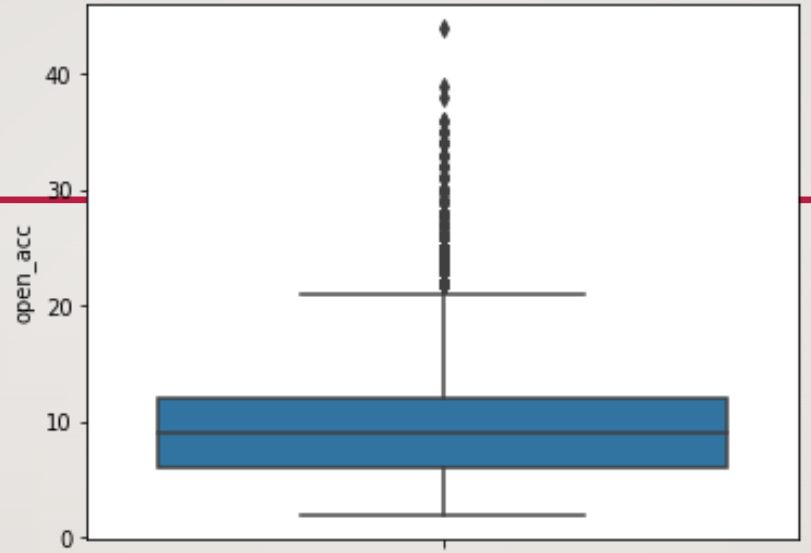
- **Year of loan issue**
  - There has is a steady increase in the number of applicants per year.
  - Its almost doubling every year
- **Month of Loan Issue**
  - Loan issue seems to increase every month from Jan to Dec
  - Probably to meet yearly goals and targets.
- **Earliest Credit Line**
  - Post 1980 there has been a steady increase until 2000.
  - There has been a sharp decline in the first time loan seekers since 2000

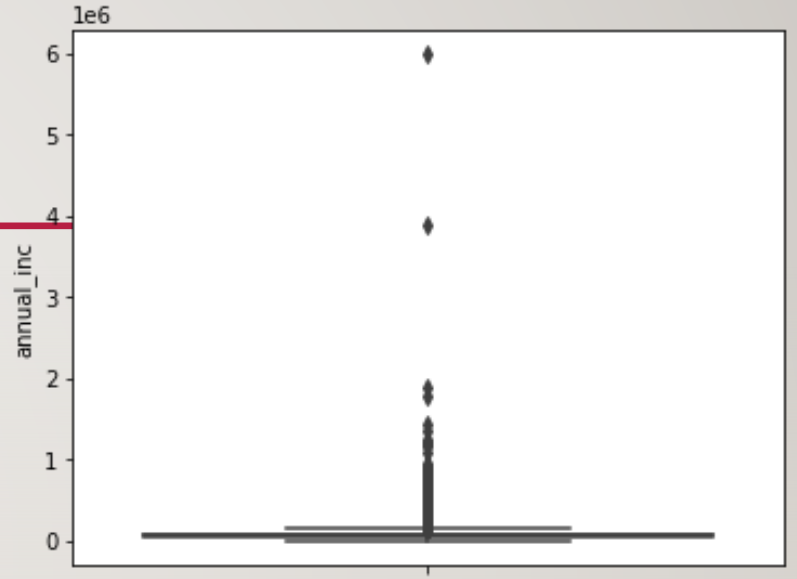
OUTLIER DETECTION

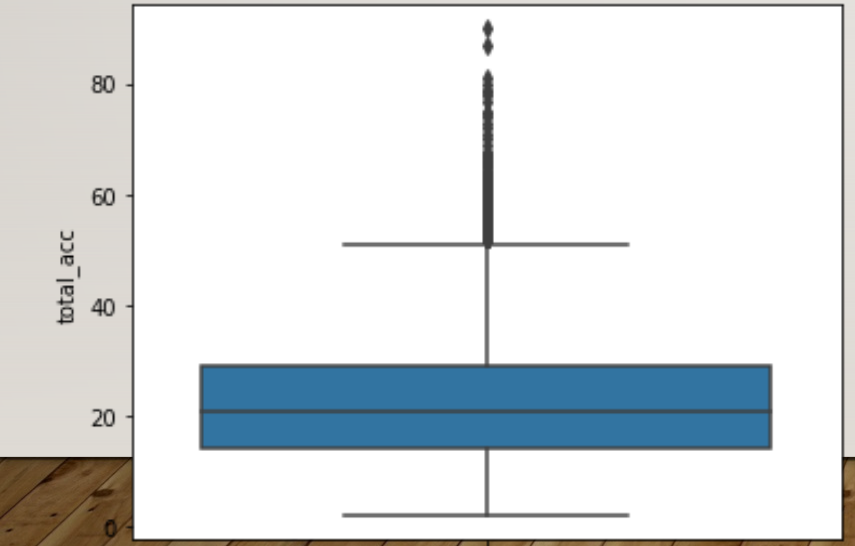


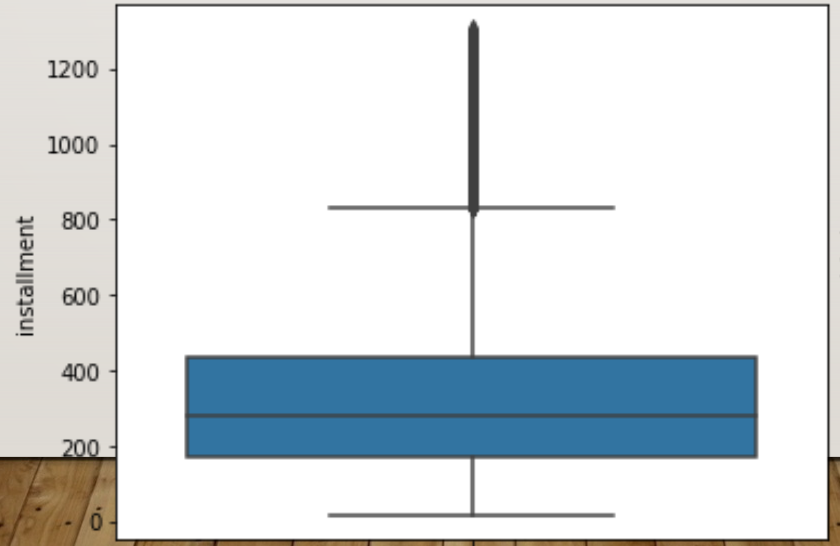
OUTLIER DETECTION

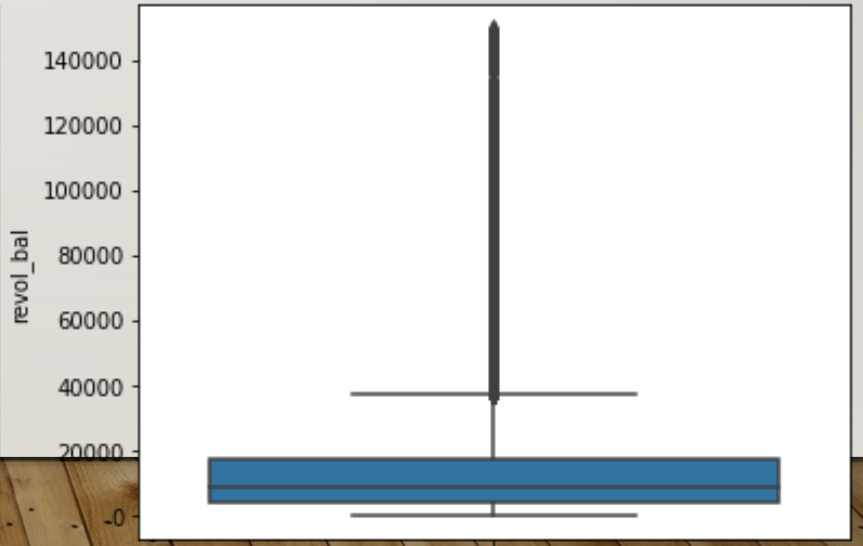






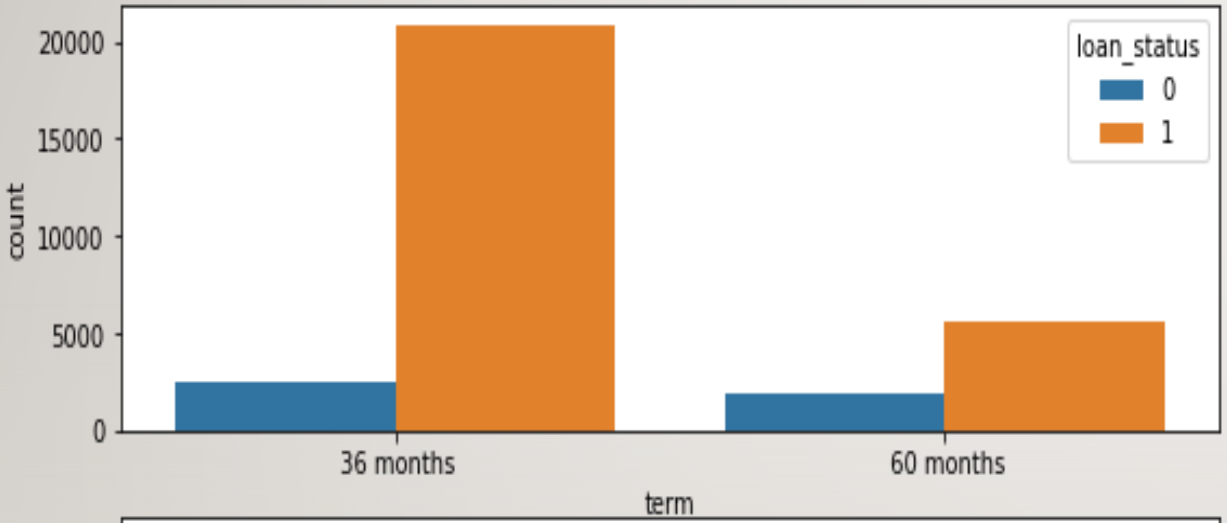


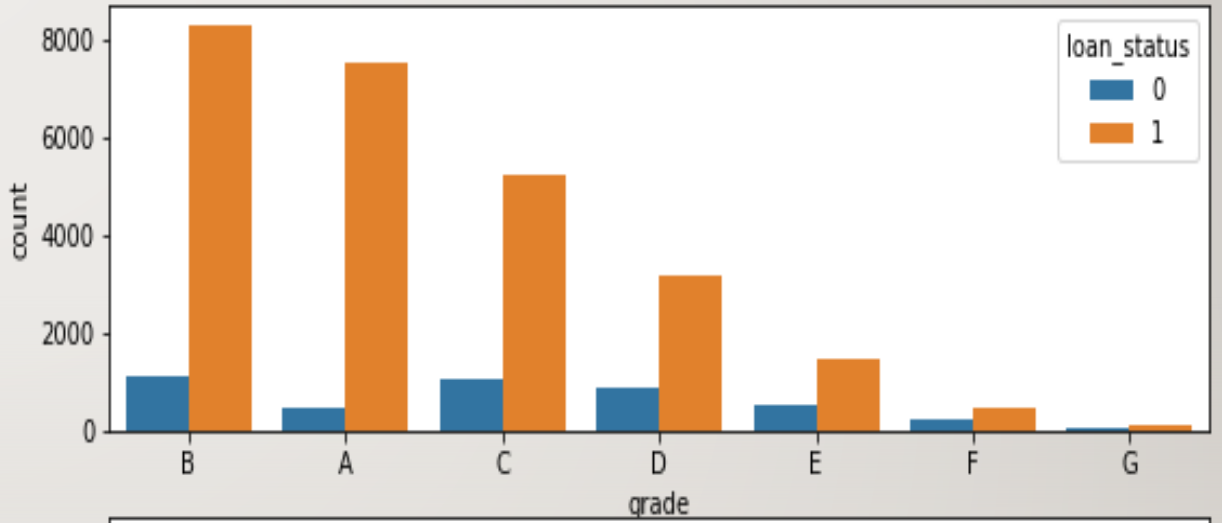


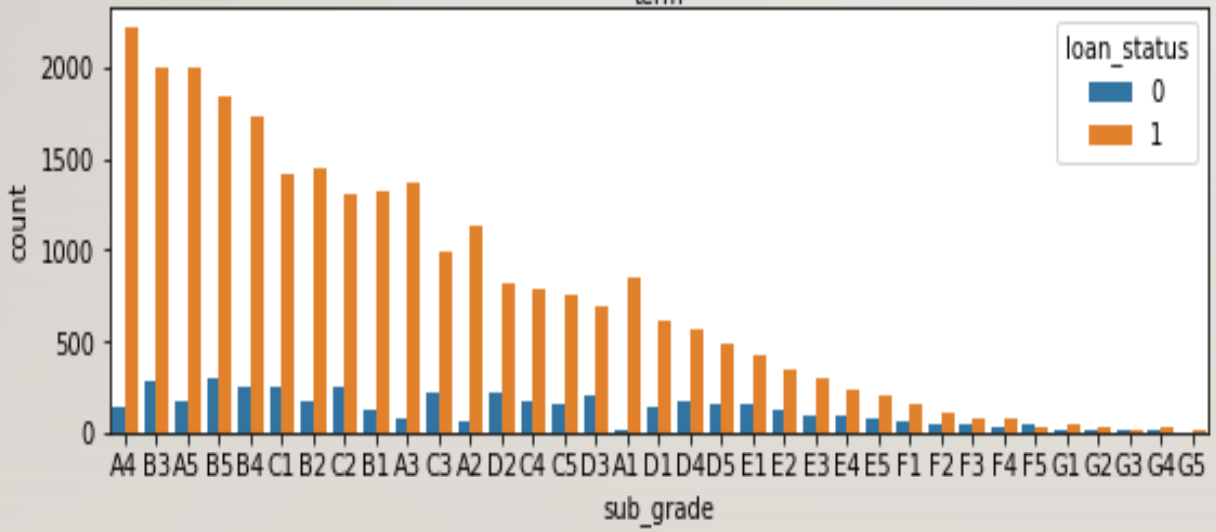


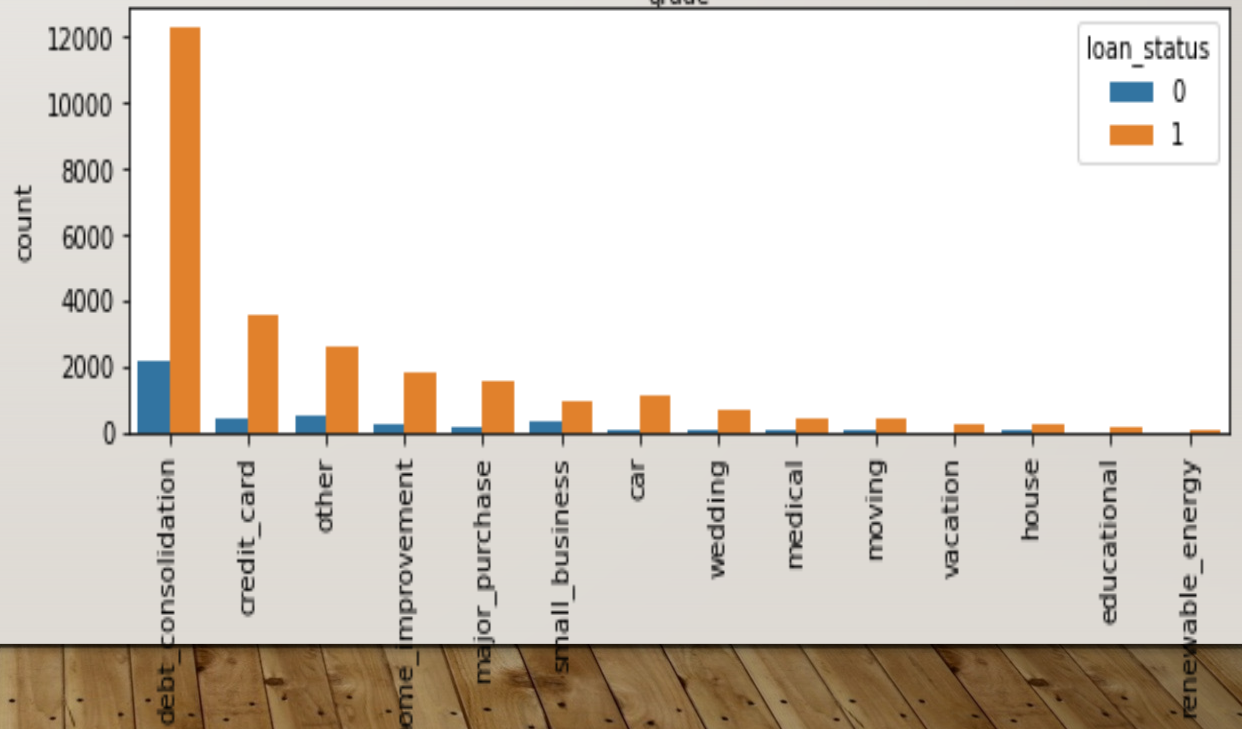
UNIVARIATE ANALYSIS

ANALYSIS OF VARIABLE AGAINST LOAN STATUS

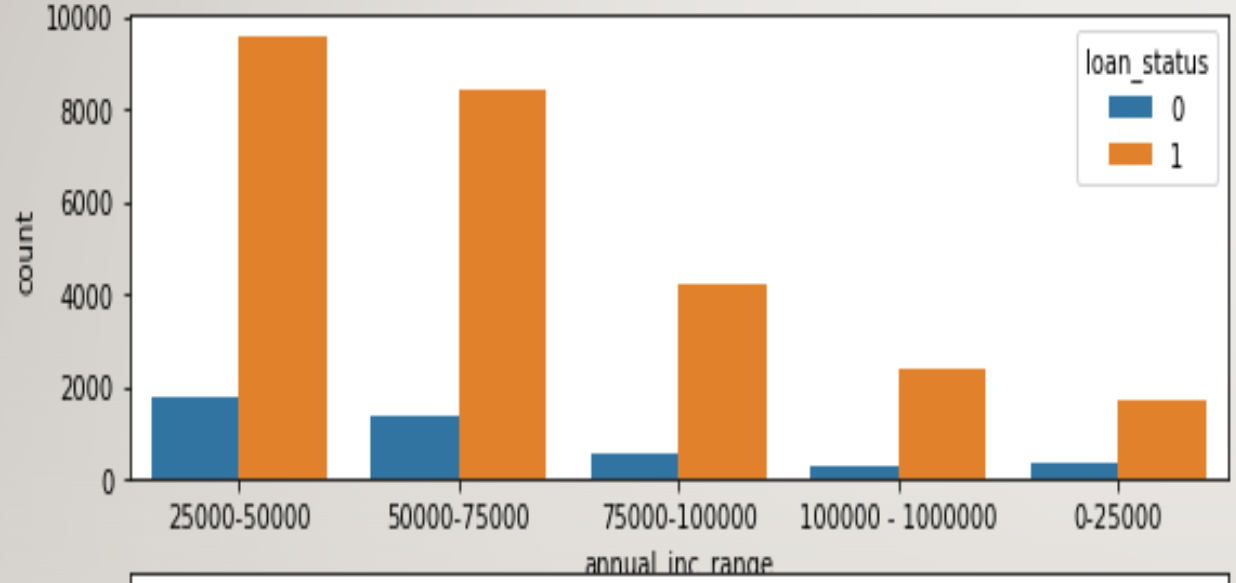


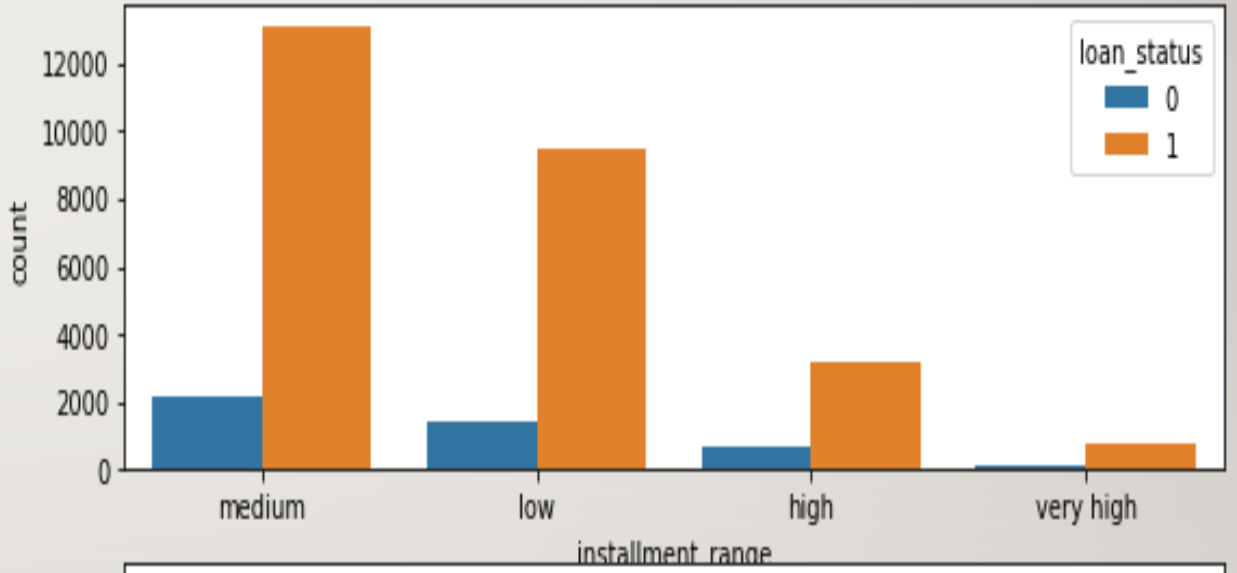


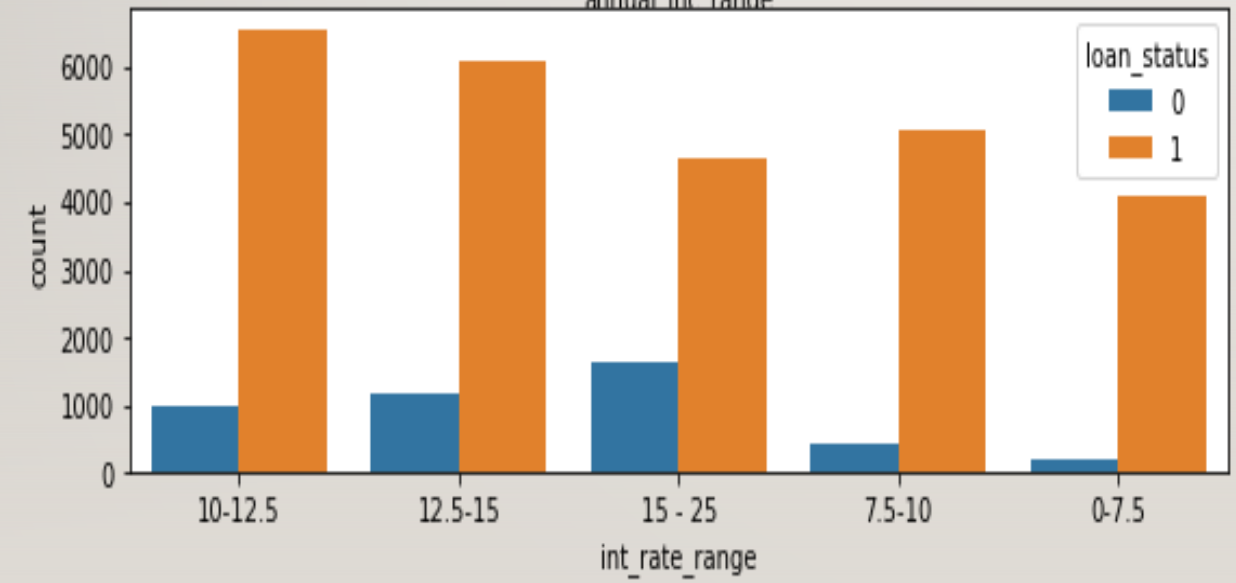


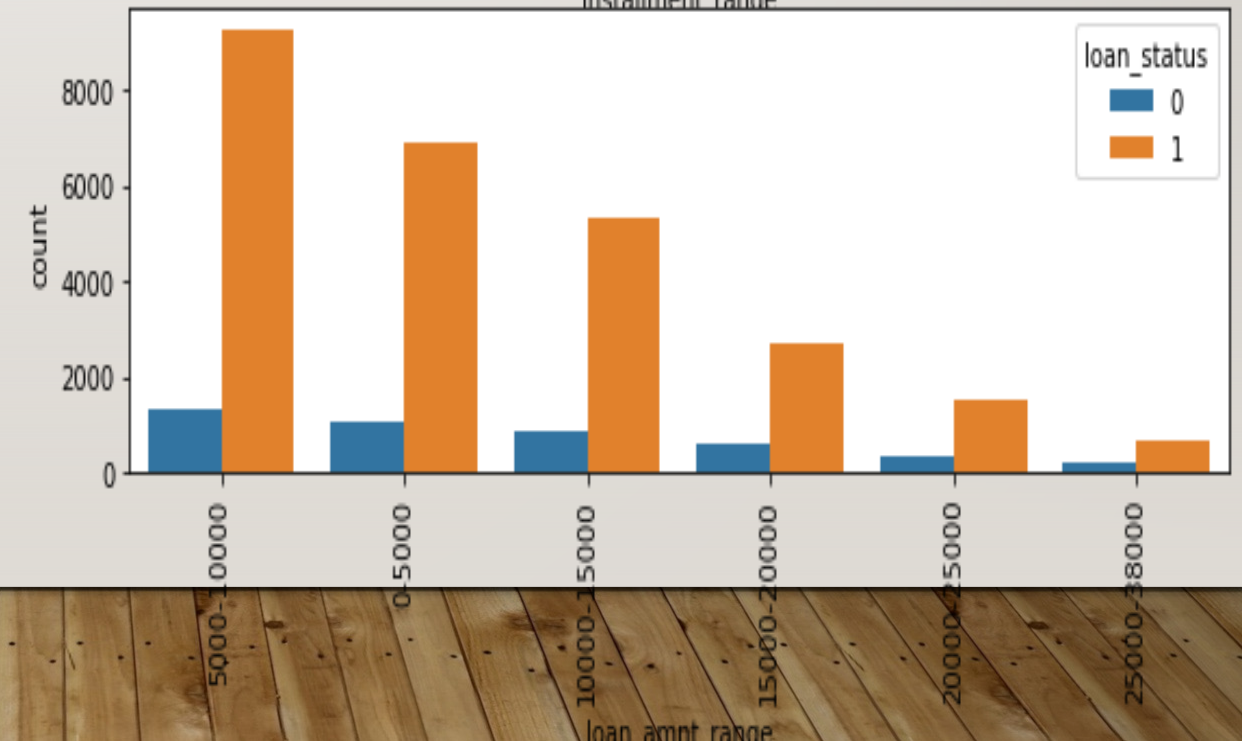


ANALYSIS OF VARIABLE AGAINST LOAN STATUS

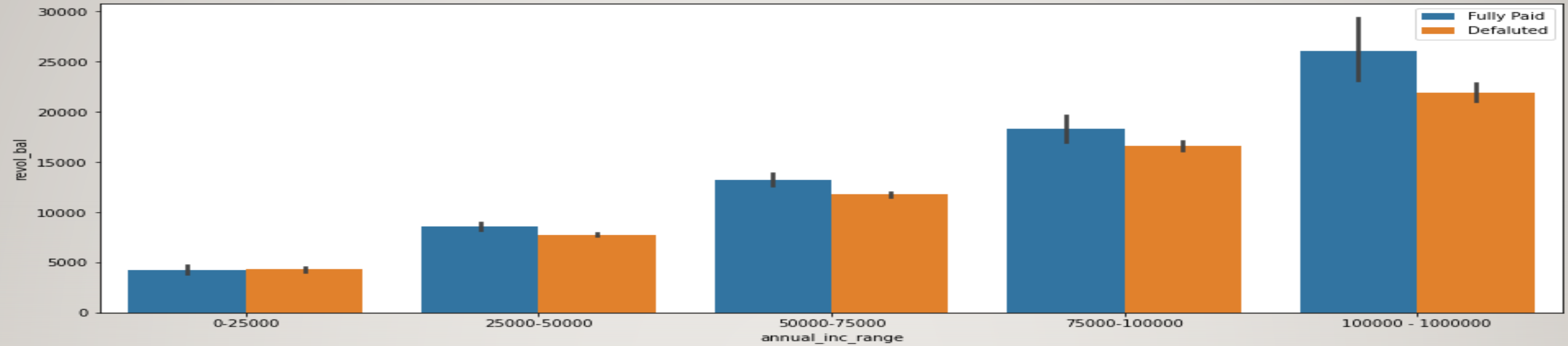


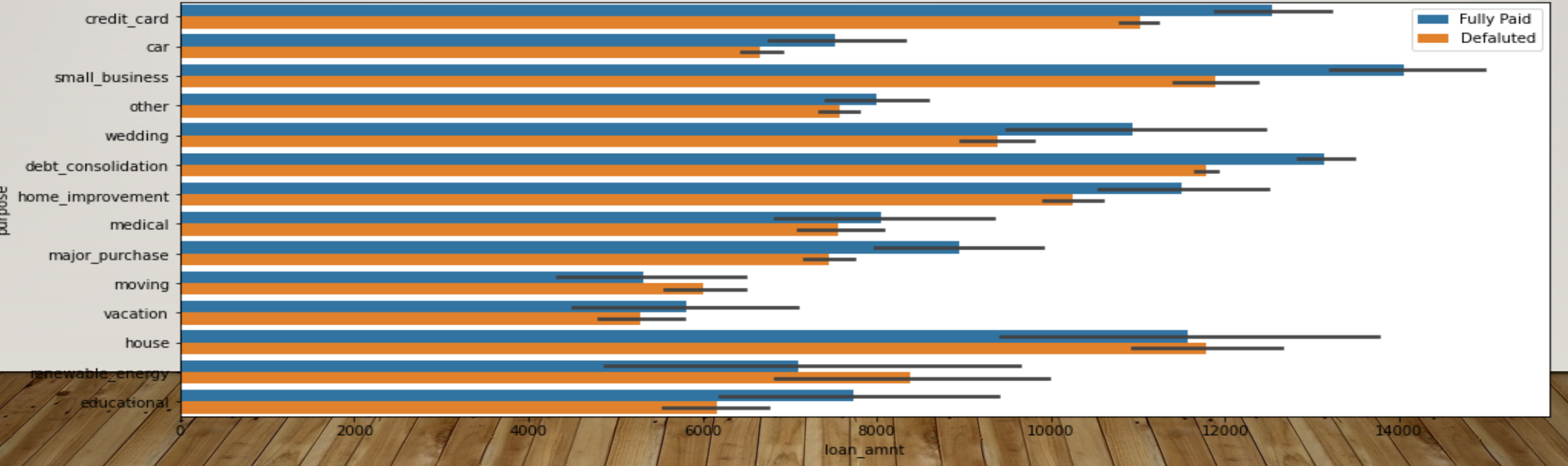




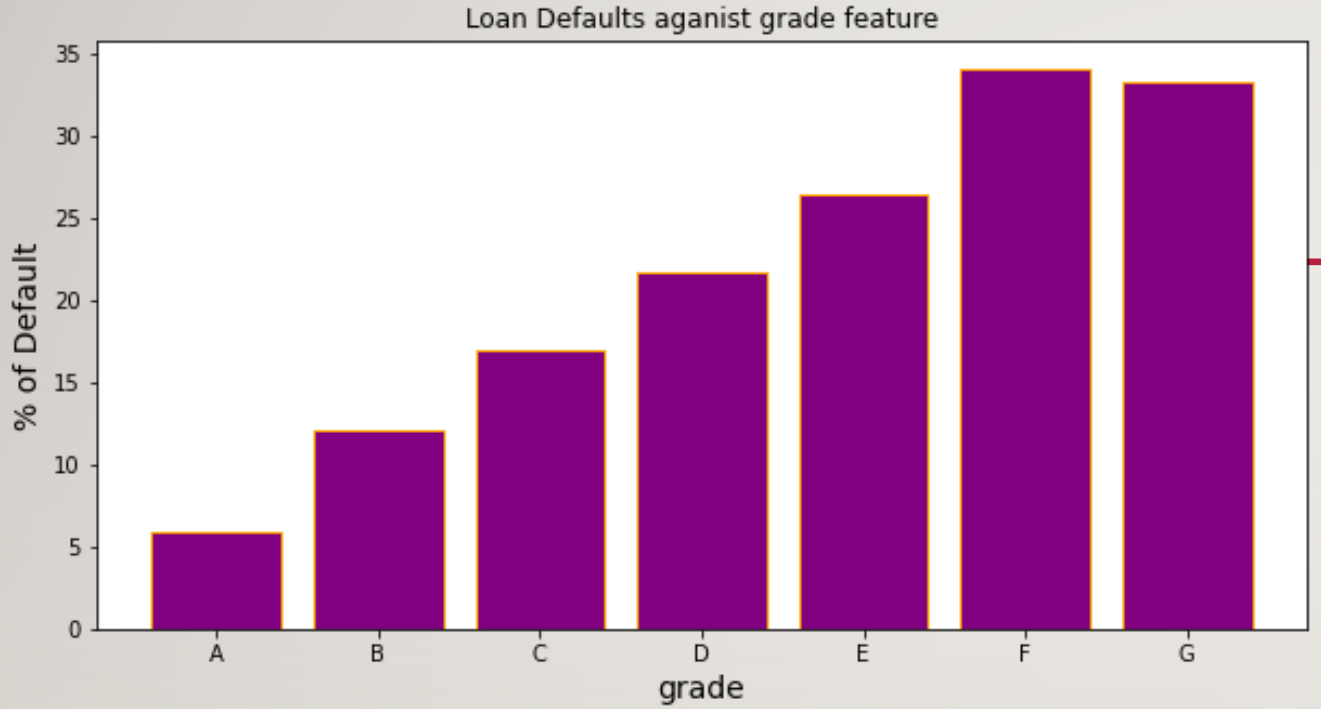


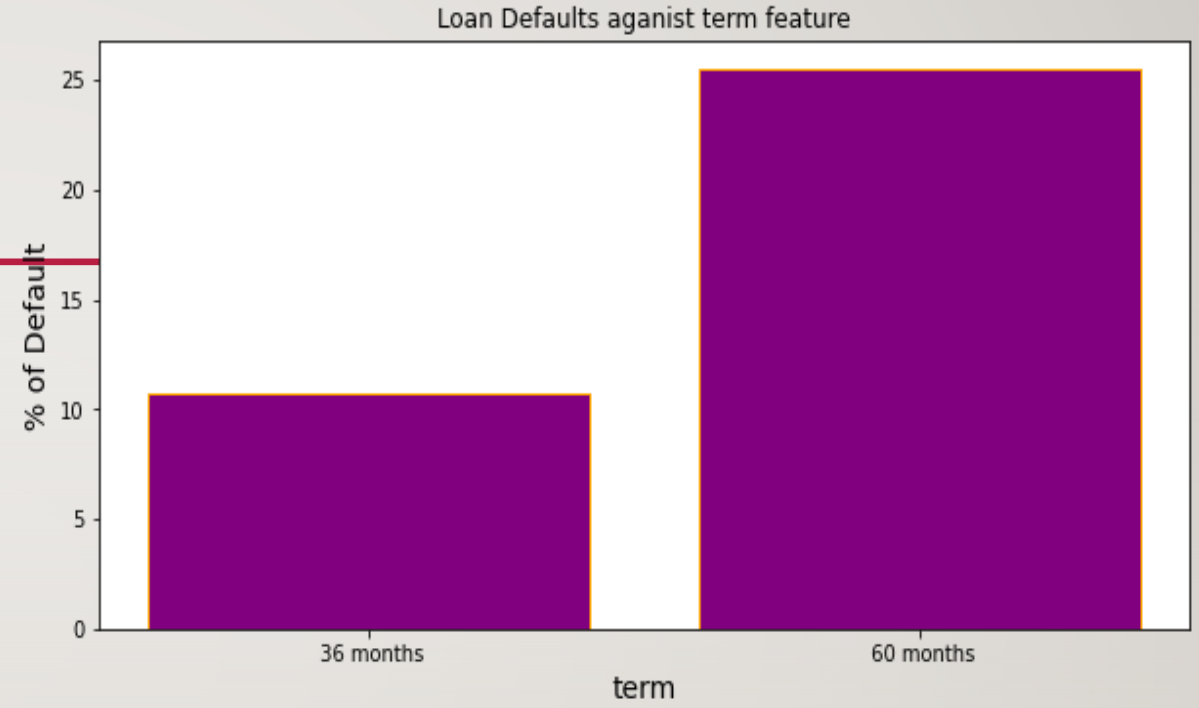
ANALYSIS OF VARIABLE AGAINST LOAN STATUS

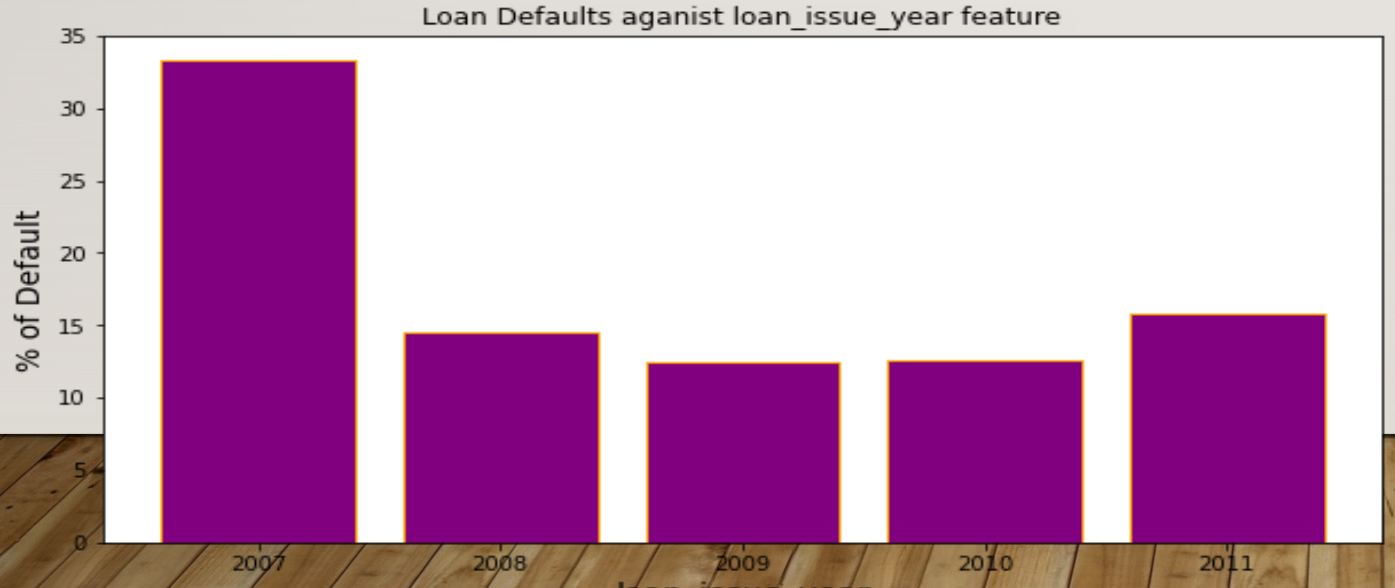


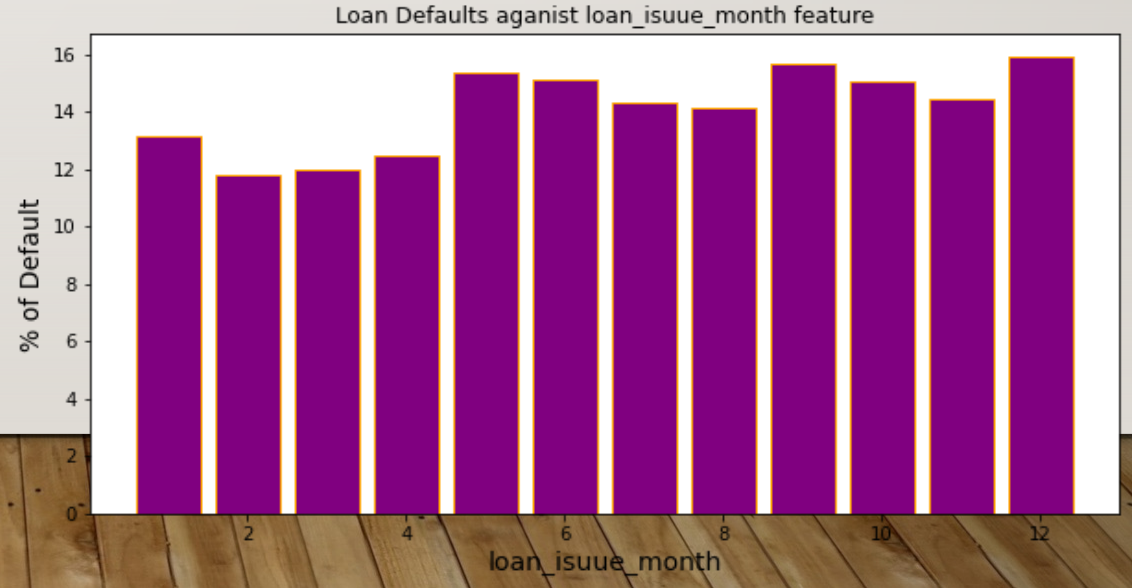


ANALYSIS AGAINST PERCENTAGE OF LOAN DEFAULTS

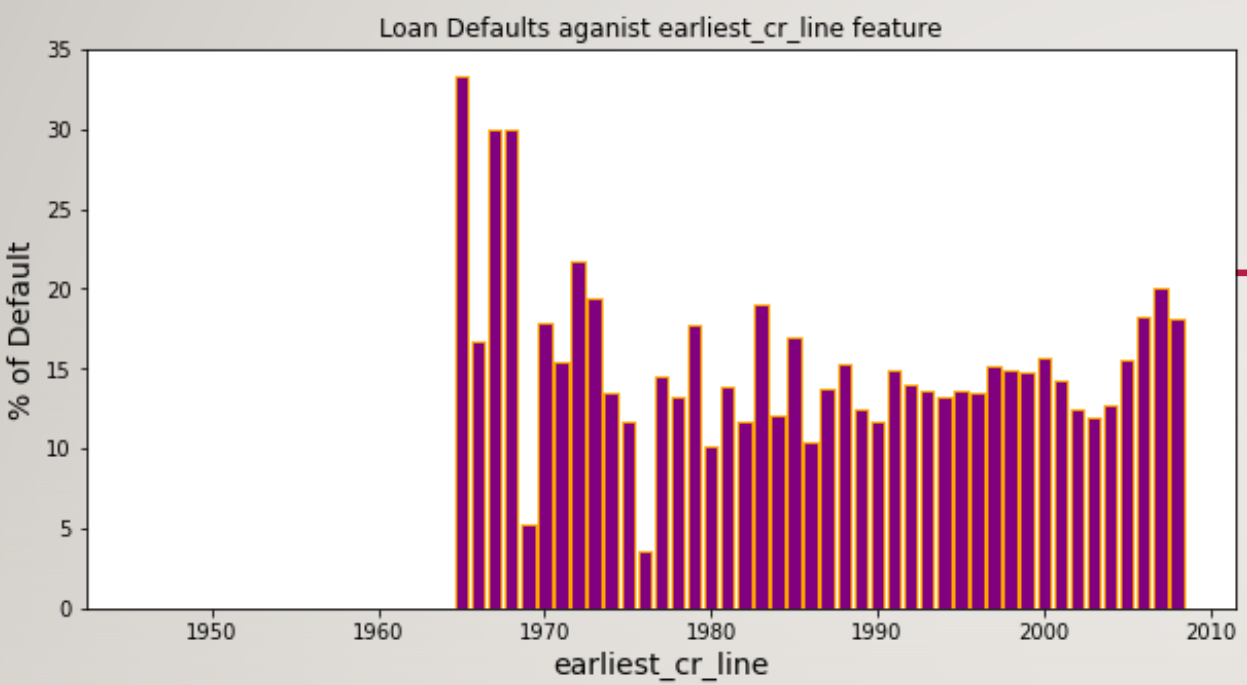


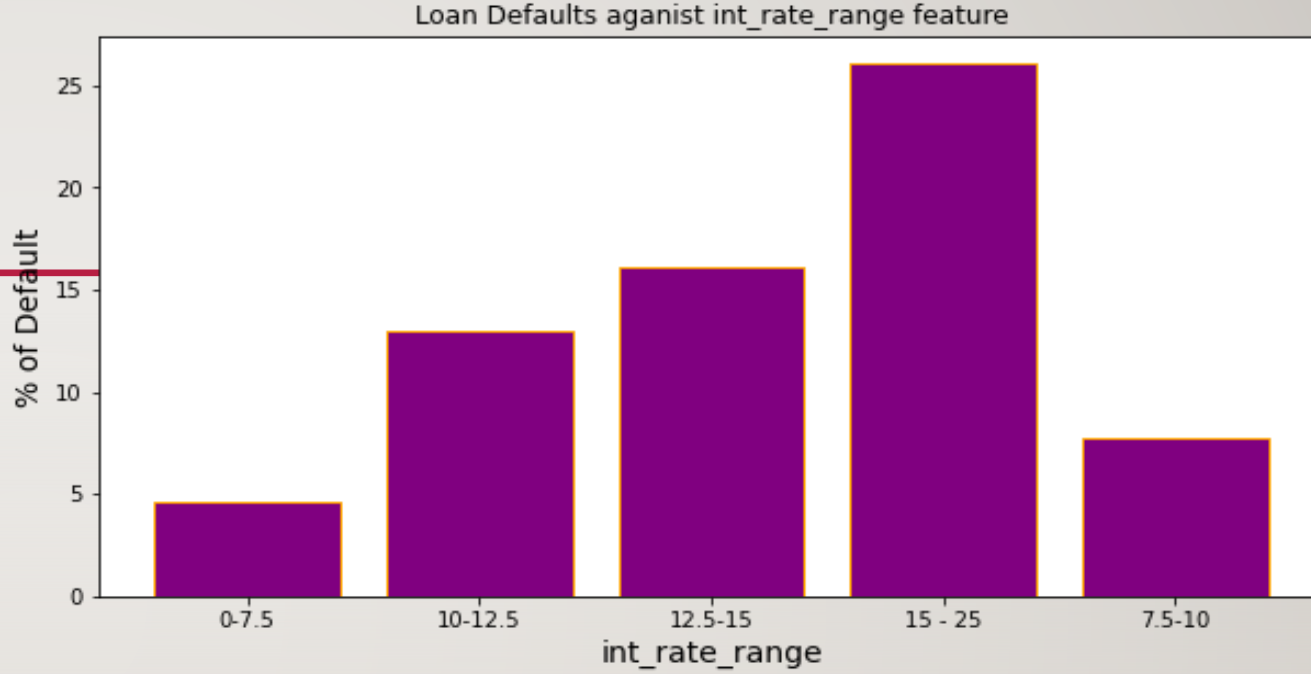


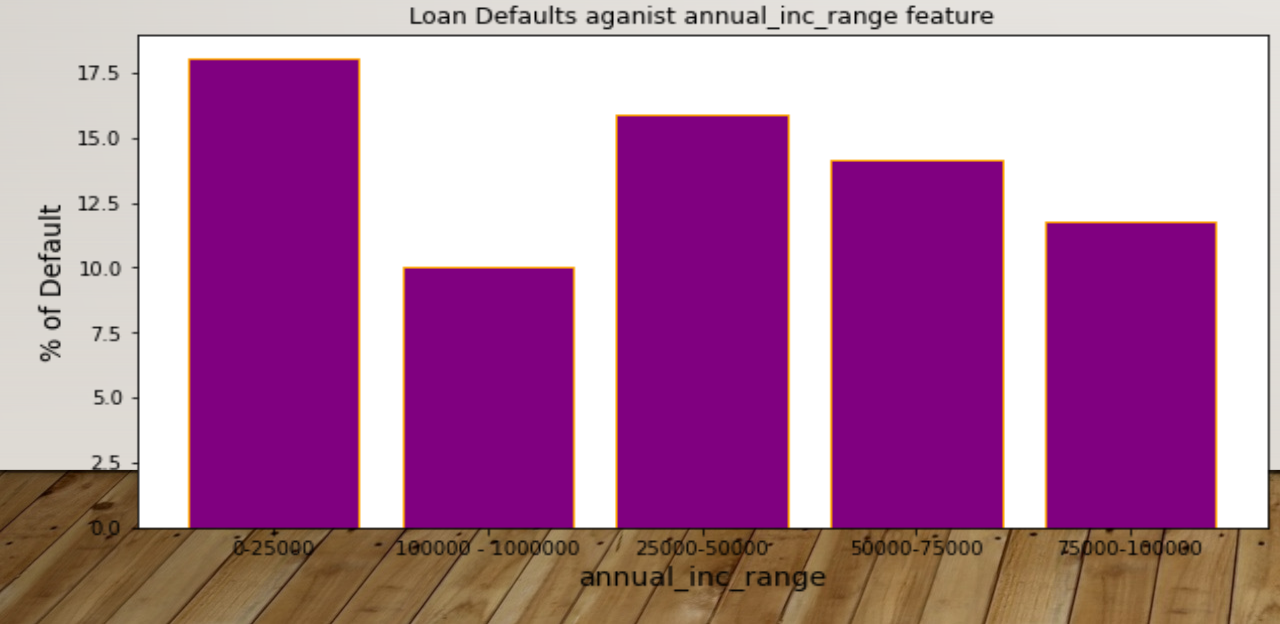


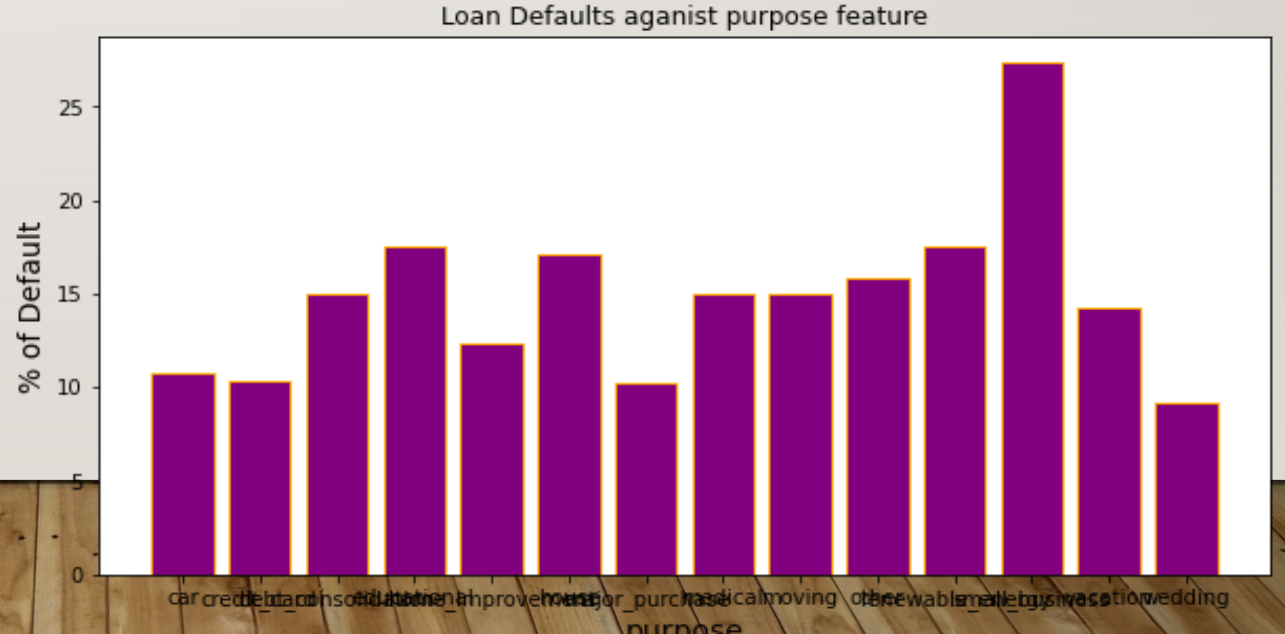


ANALYSIS AGAINST PERCENTAGE OF LOAN DEFAULTS







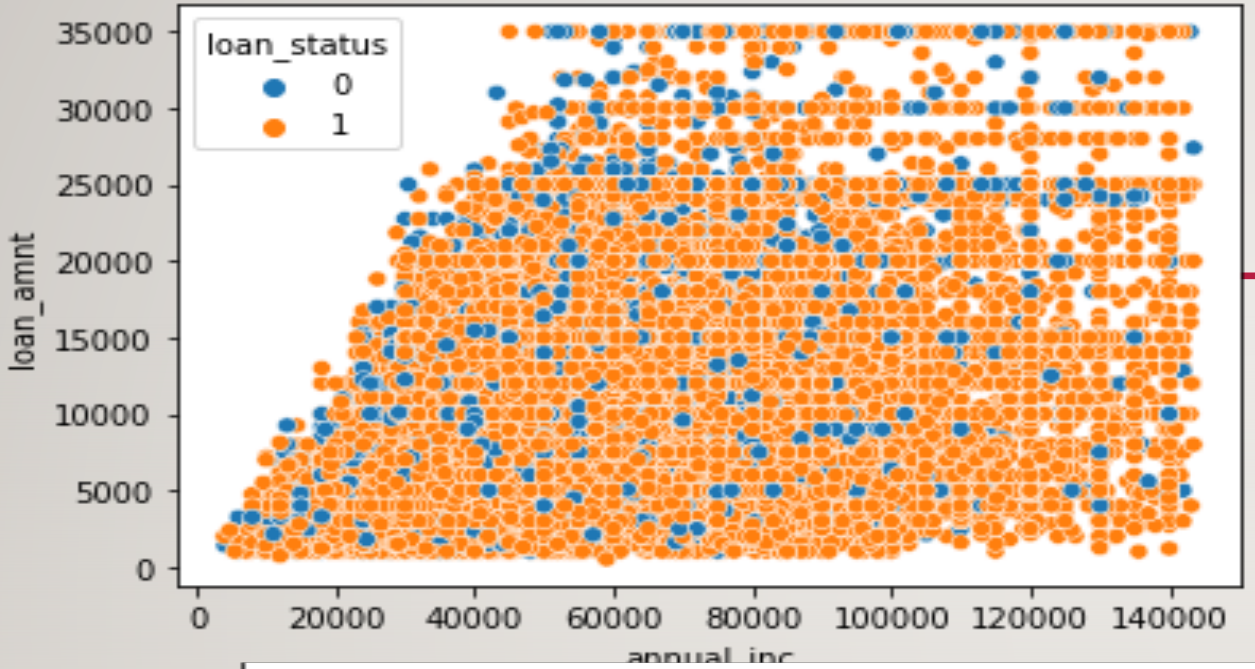


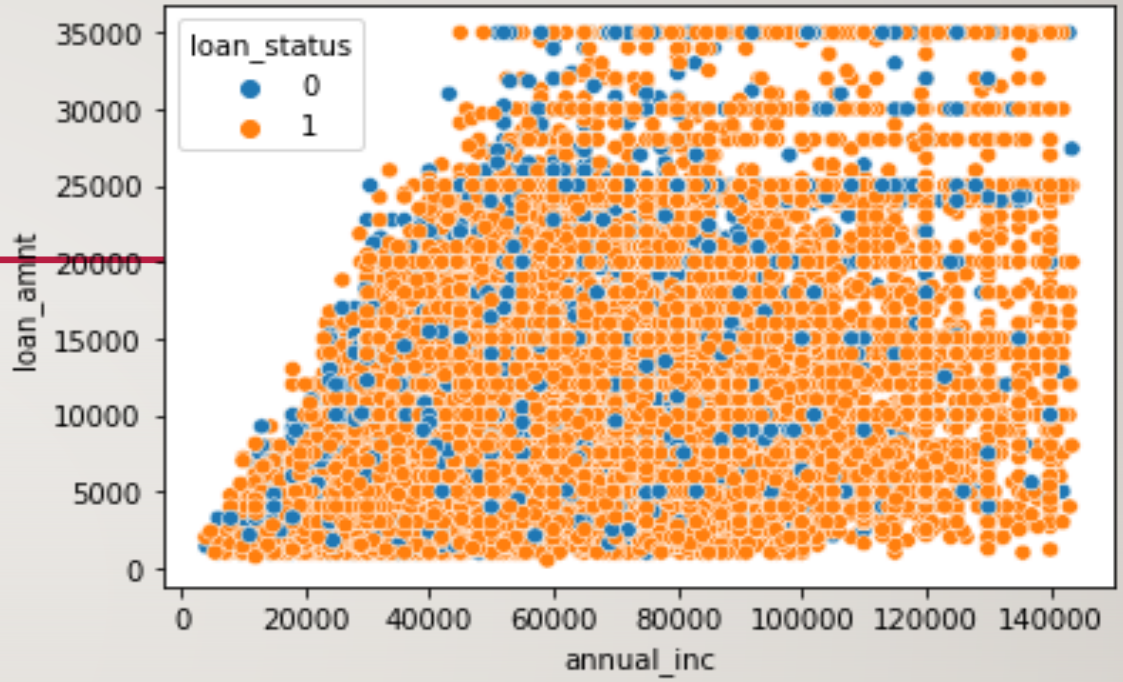
OBSERVATIONS: UNIVARIATE ANALYSIS

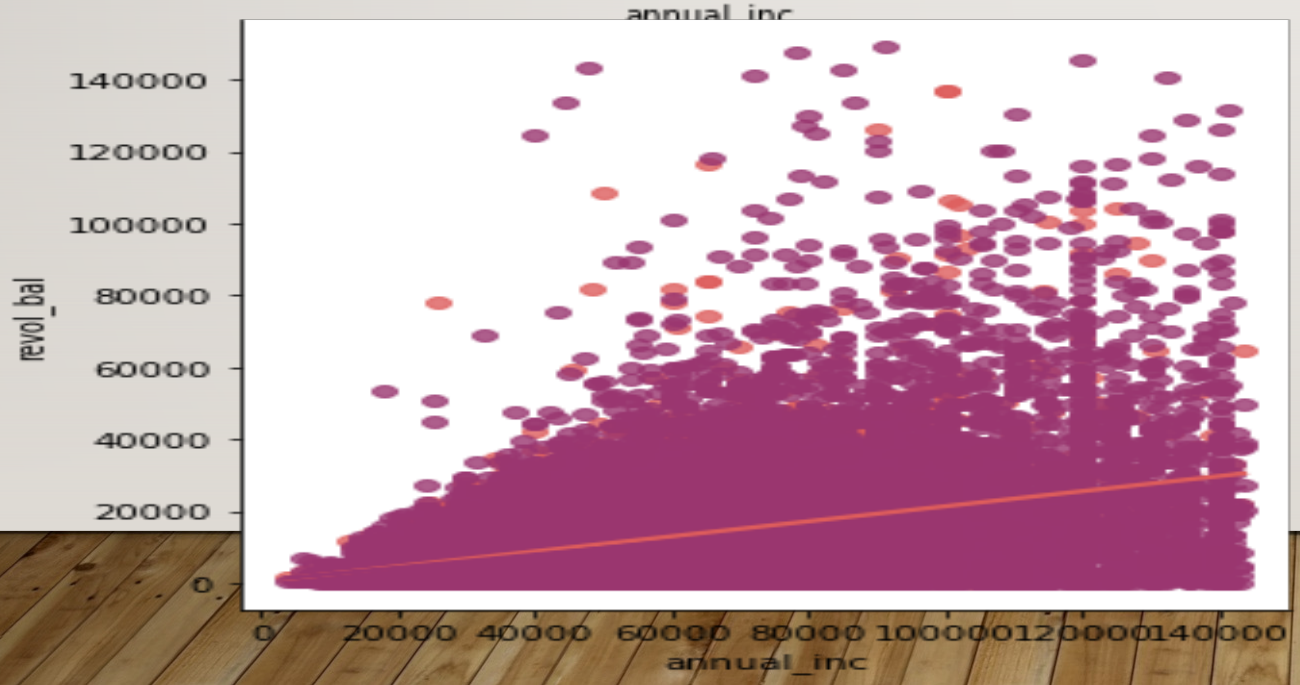
- Univariate analysis shows the importance of each of the against a single parameter
- In this case the percentage of defaults was the parameter.
- Higher variation in percentage of default based on variable values means that the variable is important to the analysis.
- Form the above analysis we can see that following features are of importance
    - Interest rate
    - Term
    - Purpose
    - Annual Income
    - Grade

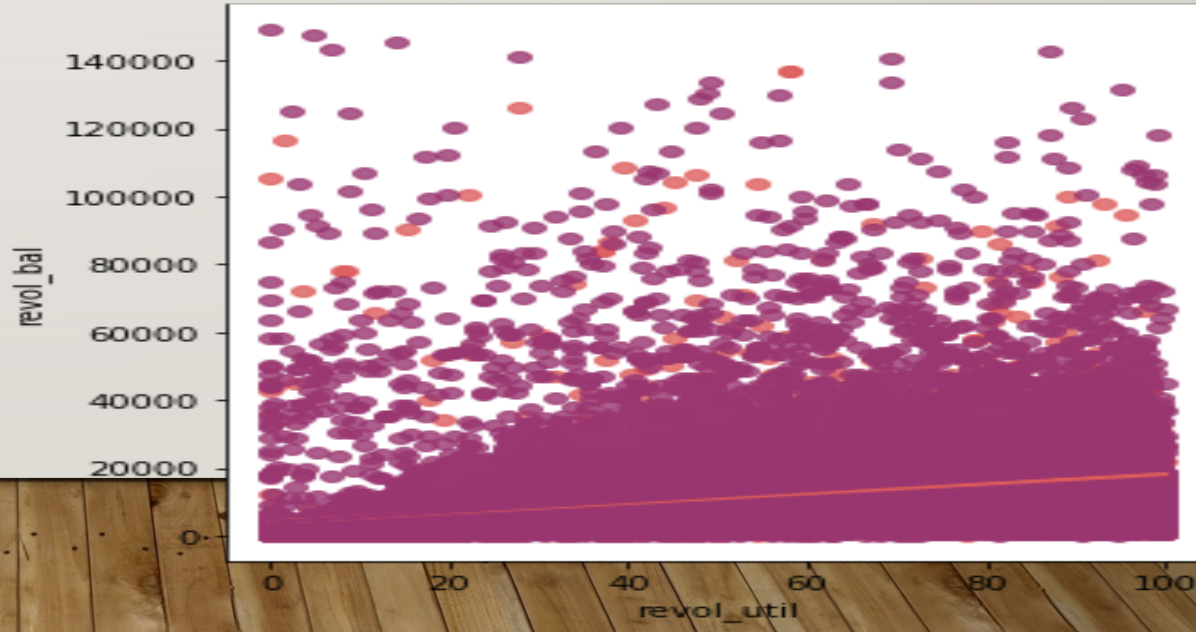
BIVARIATE ANALYSIS

BIVARIATE ANALYSIS

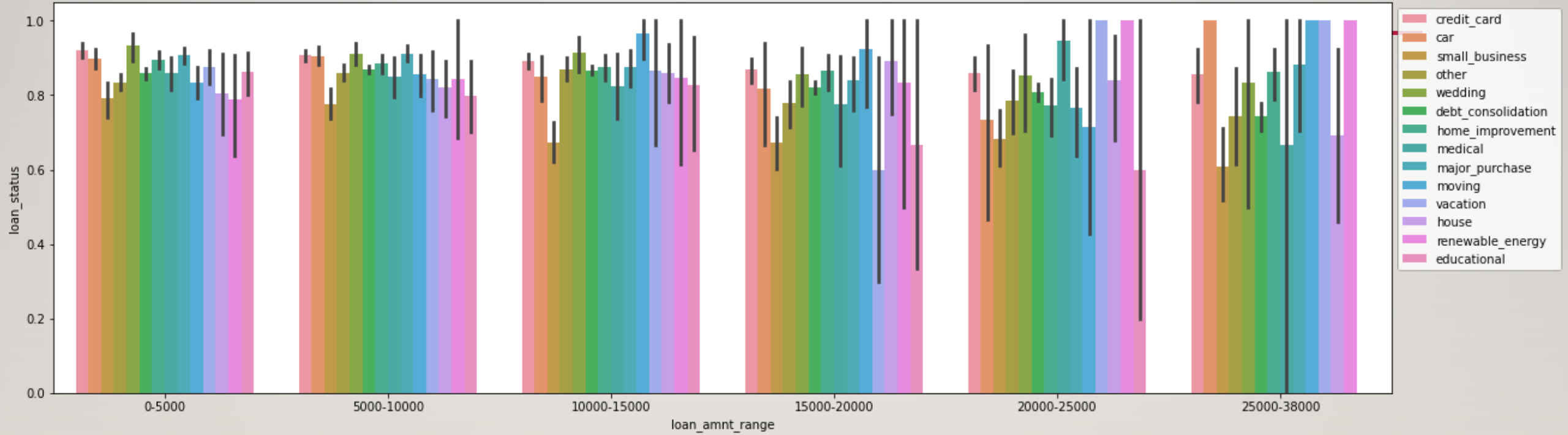




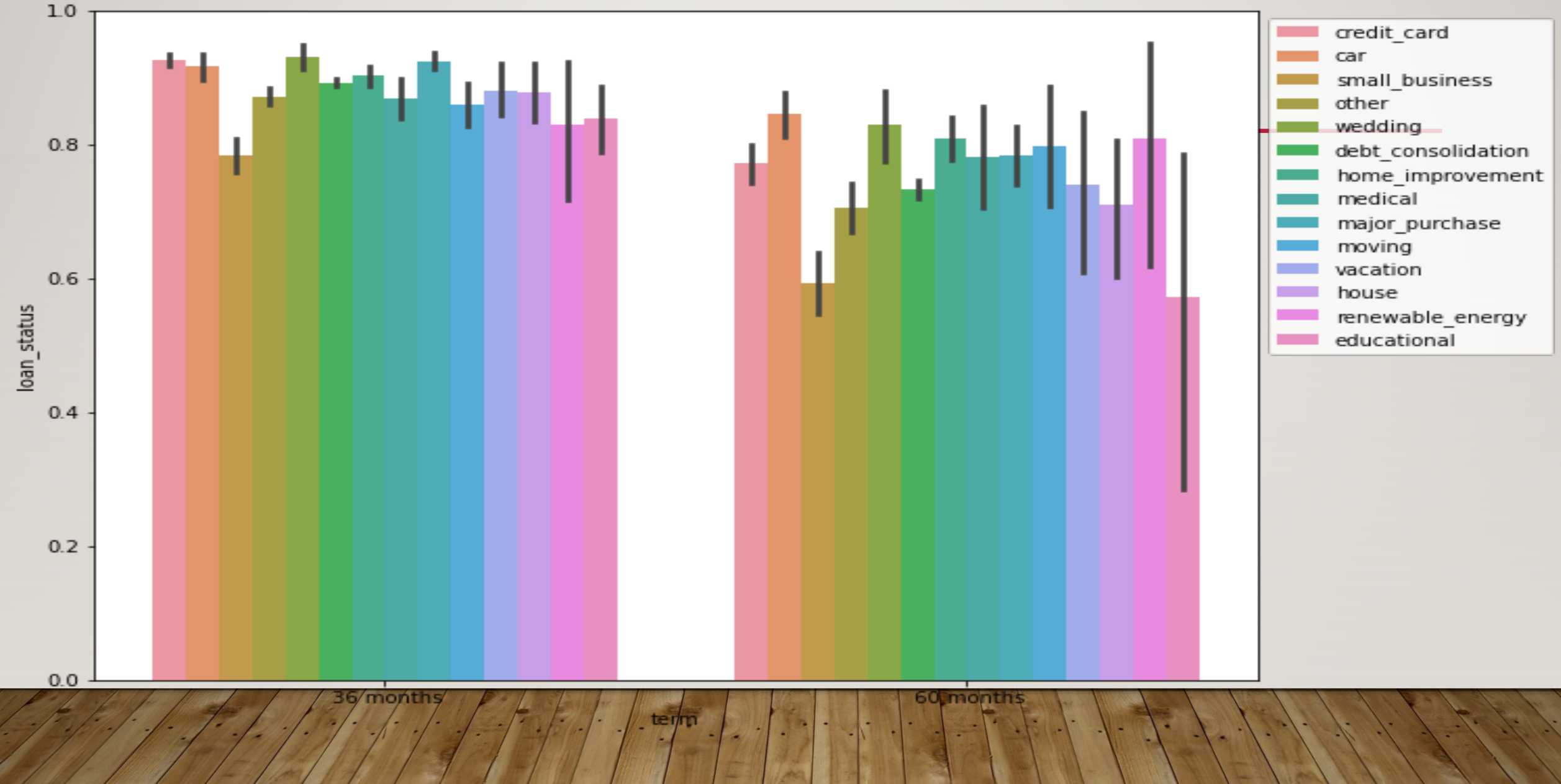




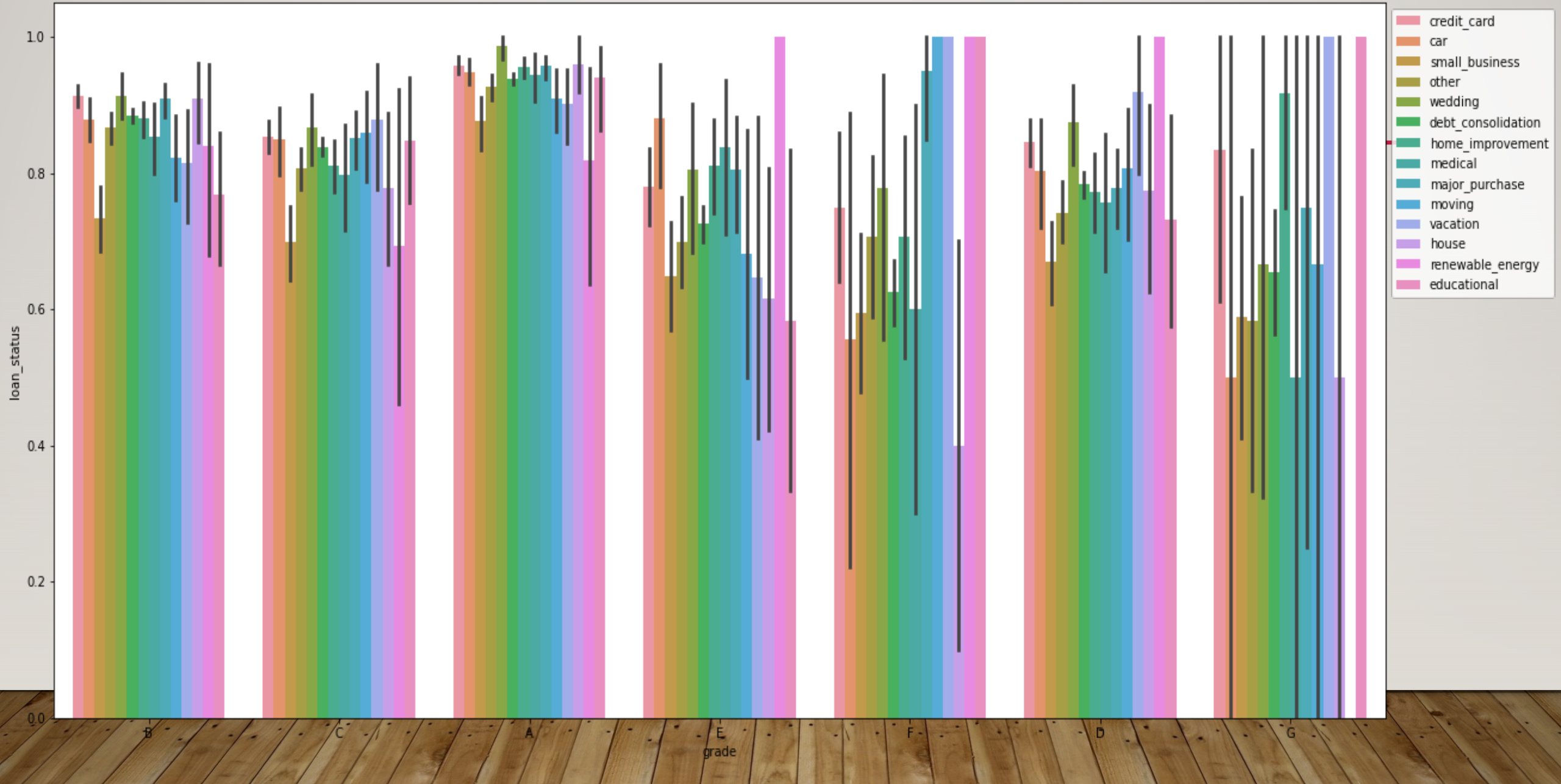
BIVARIATE ANALYSIS



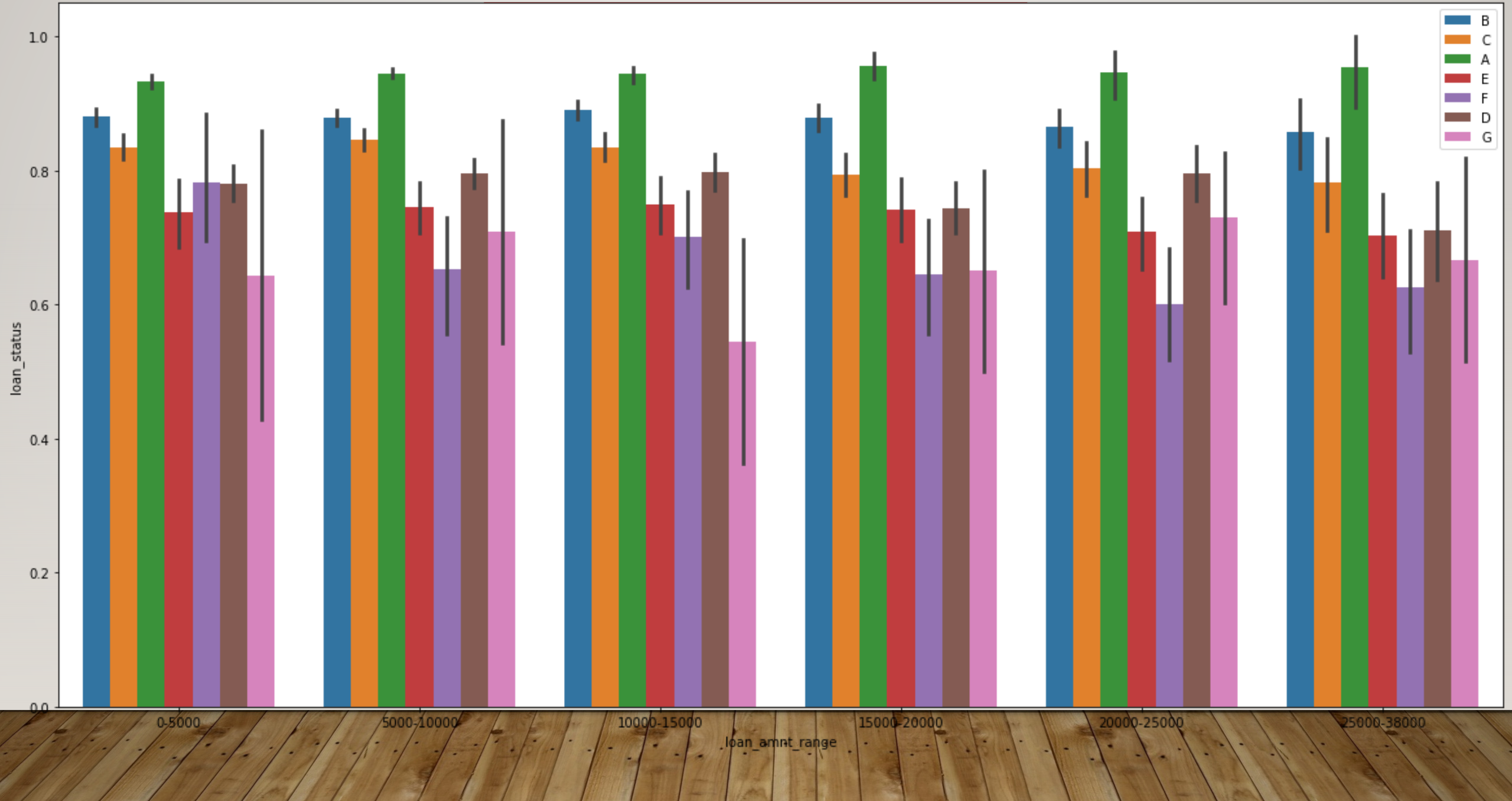
BIVARIATE ANALYSIS



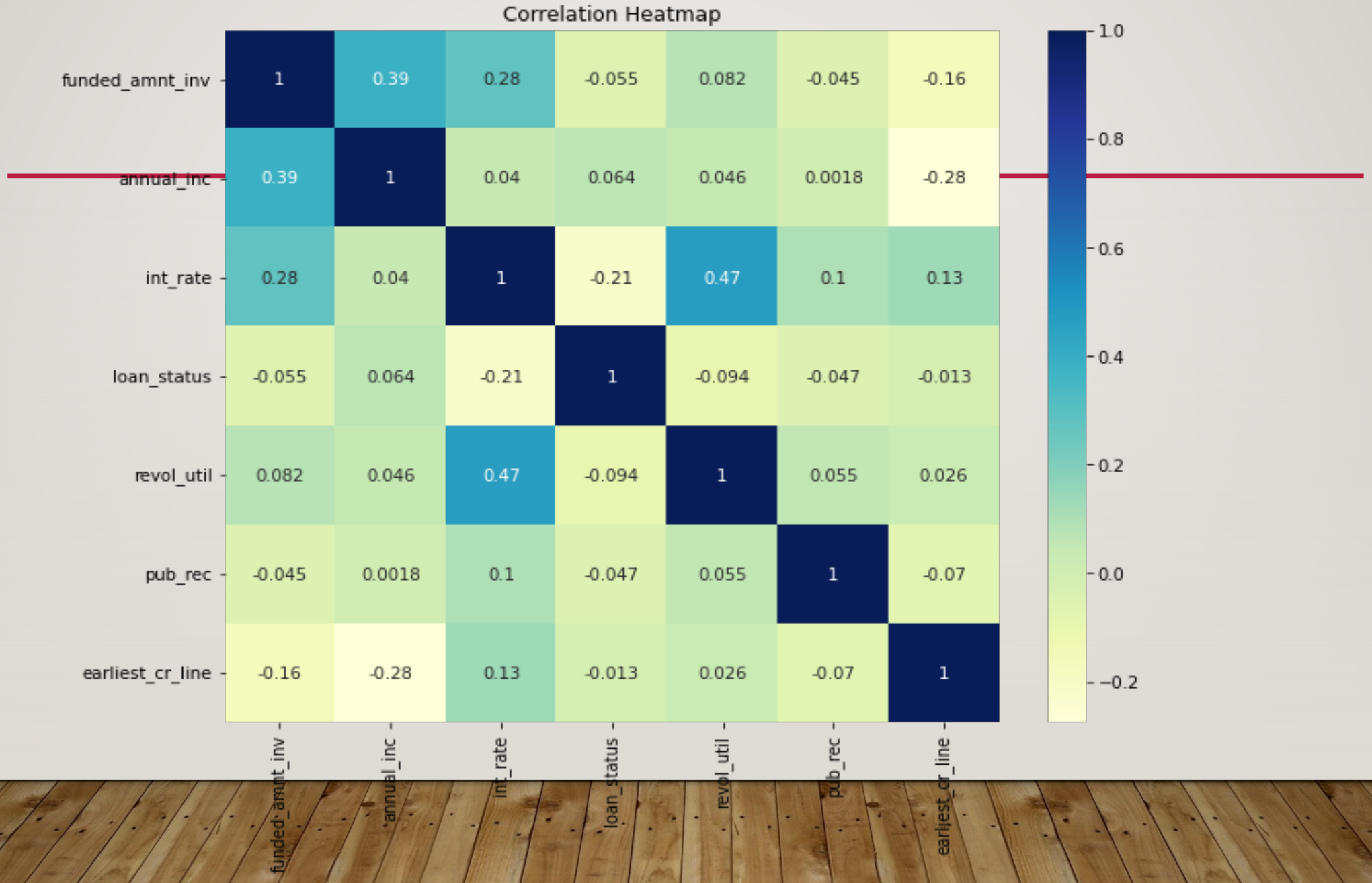
BIVARIATE ANALYSIS



BIVARIATE ANALYSIS



HEAT MAP: CORRELATION OF NUMERIC VARIABLES



OBSERVATIONS: BIVARIATE ANALYSIS

- Once more we clearly see that following are the important variables
  - Annual Income
  - Interest Rate
  - Grade
  - Term
  - Purpose
  - Revolving Credit Utilisation
- Correlation of numeric variables
  - Annual income and loan amount is corelated.
  - Loan Status and interest rate have negative correlation
  - Annual income and investor funded loan has some positive correlation.
  - Interest rate and utilization of revolving credit has some positive corelation

CONCLUSION

- **Summary**
  - Though a large number of features were present in the initial data set, most of the data were redundant and not good for the analysis.
  - Cleaning of the data and removing redundant data brought the features by more than 50%.

- **Recommendations**
- The best driving features for the Loan default analysis are:
  - **Interest Rate** : Higher interest rare leads to more defaults
  - **Term** : Lower terms means larger instalments and hence more chances of defaults, unless borrower is interested in earlier payback.
  - **Grade** : Better the grade lower the chances of default
  - **Purpose** : Payment towards already existing debts are likely to end in default.
  - **Revolving Credit Utilisation** : Higher utilisation of revolving credit may lead to defaults as the borrower's financial ability to service the loan become lesser due to existing liabilities.
  - **Instalment** : Higher the instalment, more chances of default
  - **Annual Income** : Lower the income, chances to default are more. Also, see this in relation with existing credit because high income people with high liabilities can also be defaulters.

THANK YOU

Acknowledgement

- Team UpGrad and III-T Banglore.
- Student Buddy.
- Fellow Data Enthusiasts in GitHub, Kaggle and other online forums
- Thank you for the wonderful opportunity