upGrad

Lending Club Case Study
(Assignment submission)

Venkata Prakash Reddy A
Ganga Gowthami

10-05-2023

Index

- Problem Statement
- Step 1: Data Sourcing
    - Dataset information analyis
- Step 2: Data Cleaning
    - Manual inspection and columns cleaning
    - Rows Cleaning
    - Treating missing values
    - prepared clean data set without missing values
- Step 3: Derived Columns
- Step 4: Univariate Analysis
    - Quantittative variables distribution
    - Outliers detection and correction
    - Categorical variables univariate analysis
    - Inferences from categorical variables univariate analysis
- Step-5: Bivariate Analysis
    - Categorical variables Bivariate analysis
    - Inferences from categorical variables bivariate analysis
    - Quantitative variables Bivariate analysis
    - Inferences from quantitative variables bivariate analysis
- Conclusion

Problem statement

The data given 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.

Step 1: Data Sourcing

The data provided in this case study is private data. all the data is provided in one file loan.csv

We need to perform the

Step2-Data Cleaning

upGrad

Manual inspection and columns cleaning –

- Below mentioned columns will be removed as they donot have valid data.
- Some of these columns are customer behaviour columns for which we will not be having data at the time of loan application hence deleting them.
- 26 columns are left after cleaning

Rows Cleaning –

- Removing the rows with loan status current as they will not be usefull for loan default analysis

Treating missing values –

- Checking the percentage of missing values for each column.
- Removing the columns having high percentage of missing values
- Imputing missing values for columns with less percentage of missing values

Prepared clean data set without missing values

- After performing the above mentioned steps we got a cleaned data set without missing values and 22 columns.

Step3- Derived Columns

- Derived 2 new columns issue_month and issue_year from column issue_d

Step4- Univariate Analysis

upGrad

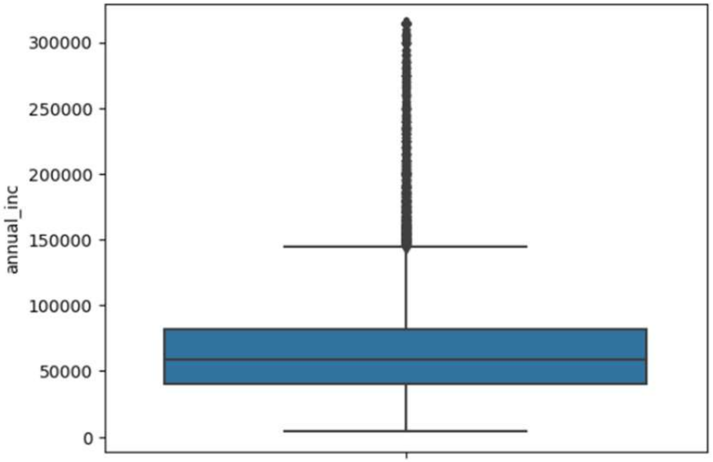
- Splitting the columns into categorical and quantitative columns
- This splitting is done based on manual inspection and the number of unique values present in each column

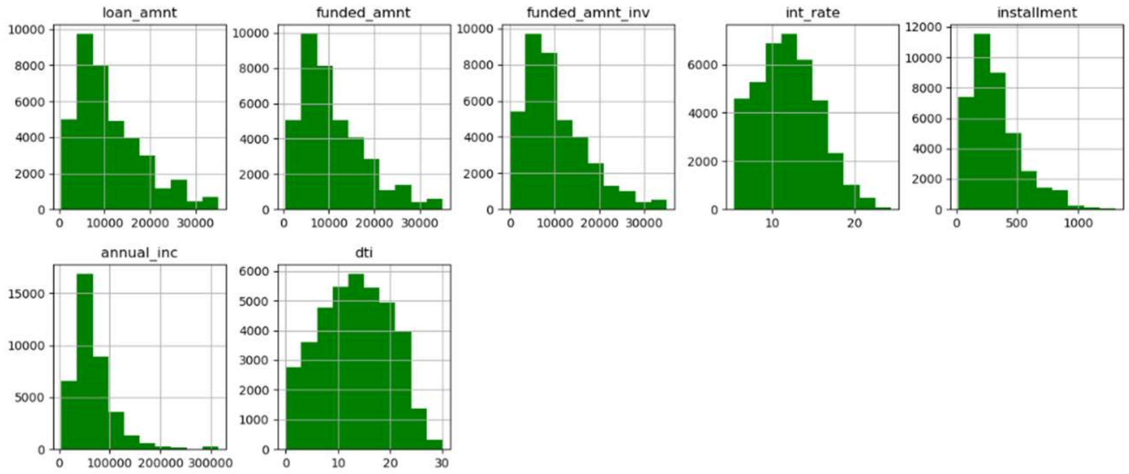
Quantitative variables univariate analysis

- Plotted Histograms for understanding distribution of Quantitative variables
- All the distributions look fine except annual_inc.

Outliers detection and correction

- As the histogram of Annual_inc is not distributed properly there might be chances of outliers.
- Used box plot to understand the outliers of Annual_inc.
- Created function to check outliers for all the columns.
- while treating outliers considered 5th percentile as Q1 and 95th percentile as Q2 as this is financial data
- verified if there are any outliers in other columns using the function.





Step4- Univariate Analysis

upGrad

Categorical variables univariate analysis

- Plotted bar graph for all the categorical variables to understand idstribution of data with respect to categorical variables.

Inferences from categorical variables univariate analysis

- Initial_list_status feature can be ignored as it has only one value.
- The number of loans gradually increases year by year.
- More loans are issued at the year end in the month of december. It might be because of festival season.
- Most of the borrowers houses are uner mortgage or they are renting the house.

Step5- Bivariate Analysis

upGrad

- Bivariate analysis is used to find the relation between 2 features of the dataset.

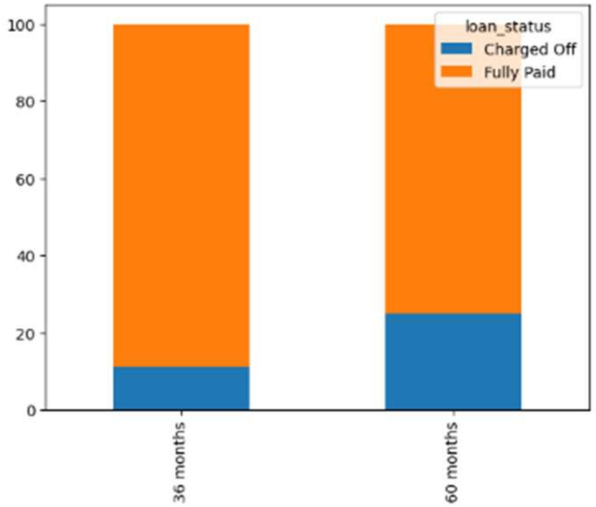
Categorical variables Bivariate analysis

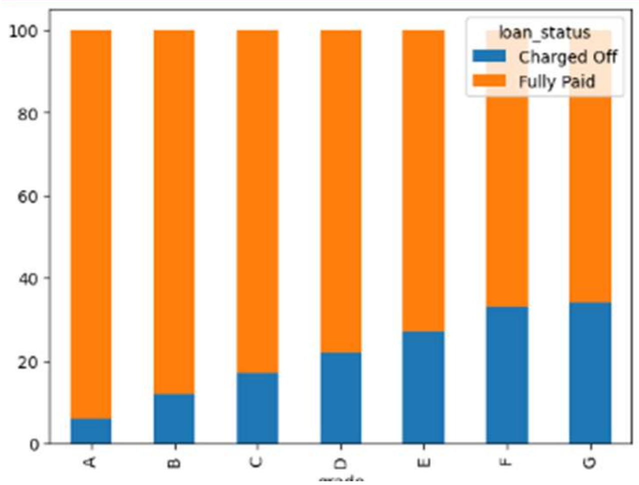
- Bivariate analysis of categorical variables is performed against the target loan status with stacked bar charts.

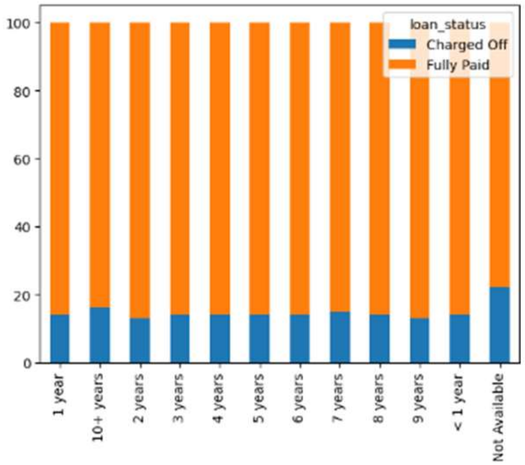
Inferences from categorical variables bivariate analysis

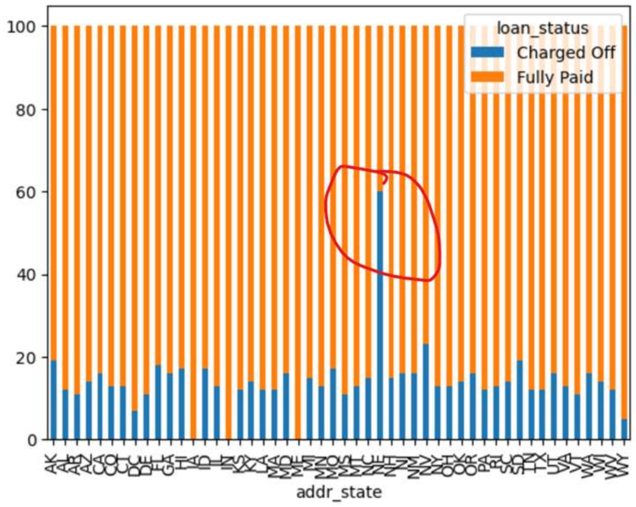
- Term - 60 Month term loans have higher percentage of defaulters
- Grade - The percentage of defaulters increases gradually as the grade changes from A to G
- Sub_grade - sub grade F5 has highest number of Defaulters between 45 to 50 %
- Emp_length - There is no impact of emp length on default percentages.
- Home_ownership - there is no specific trend of defaulters based on home ownership.
- Verification_status - surprisingly verified customer have slightly more percentage of defaulters than non verified
- Purpose - Loans taken for small business have high default rate.
- State - The state NE(Nebraska) has very high number of defaulters, also in this state defaulters are more than fully paid borrowers
- Pub_rec_bankrupcies - higher the number of bankrupcies higher the chances of defaulting the loan.
- Issue_month - no specific trend of defaulters percentage acroos the months.
- Issue_year - no specific trend of defaulters percentage acroos the years

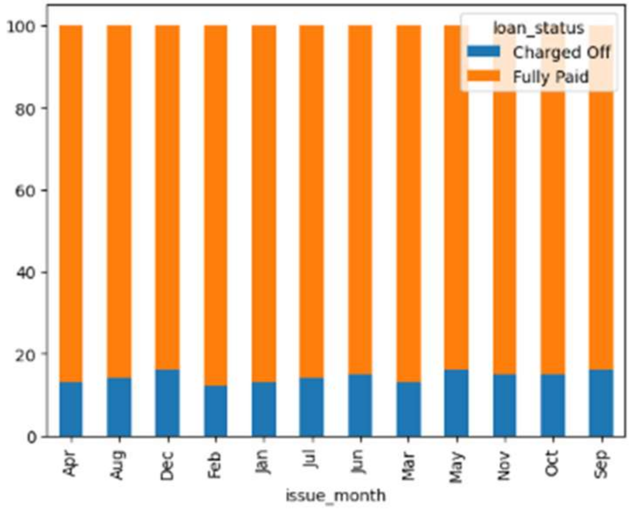
Step5- Bivariate Analysis

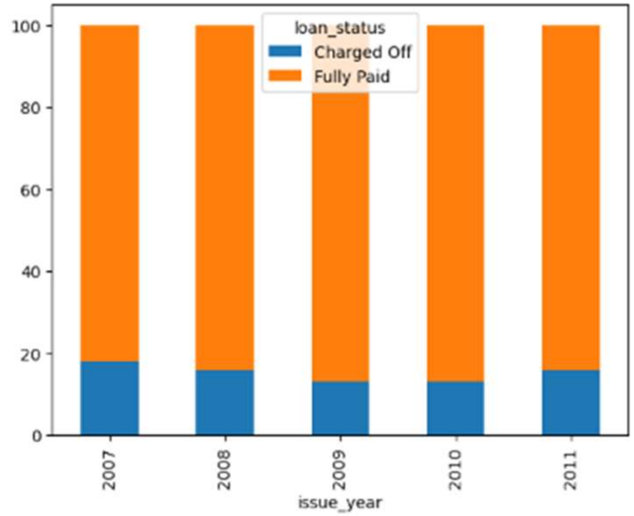












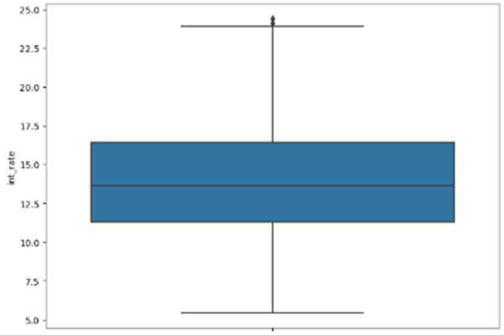
Step5- Bivariate Analysis

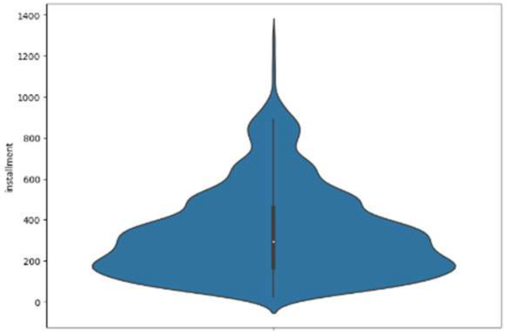
Quantitative variables Bivariate analysis

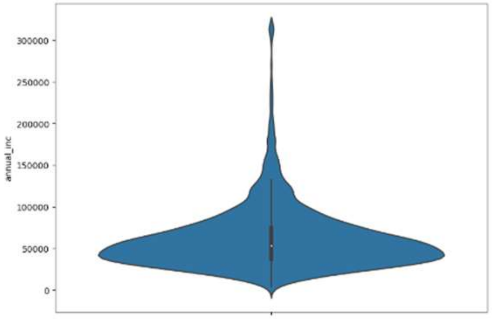
- For this analysis we will consider the data with loan_status Charged Off only.
- With this data we can check how the defaulted loans are spread with respect to quantitative variables.

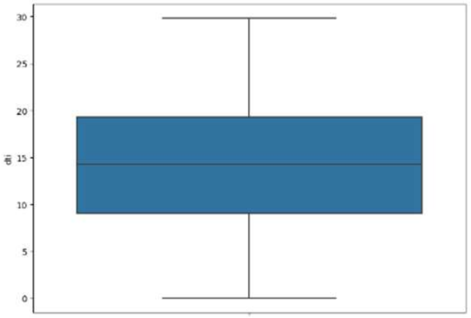
Inferences from quantitative variables bivariate analysis

- Int_rate - most of the loans with interest rate between 11 to 16 % are defaulted
- Installment - most of the defaulted loans have installements less than 1000
- Annual_inc - most of the defaulted loans have annual income less than 16000
- Dti - most of the defaulted loans have dti between 9 to 20.
- Loan_amnt - the loan amount of most of the defaulted loans is between 5000 to 16000
- Funded_amnt - same behaviour as loan_amt
- Funded_amnt_inv - same behaviour as loan_amt









Conclusion

Below are the high level conclusions. for more details check inferences sections of different analysis.

- Loans take in the state of NE(Nebraska) has very high very high risk of getting defaulted.
- Loans with smaller amounts have high risk of getting defaulted.
- Loans given to the customer with history of bancrupcy have high chances of defaulting.

upGrad

Thank You!