LENDING CLUB EDA CASE STUDY

BY

SUJEET KUMAR GOIT

VENUGOPAL PAMIDIPATI

PROBLEM STATEMENT

- Financial companies faces the critical task of assessing loan applications of urban customers to make informed decisions. When evaluating loan applications, the company confronts two primary risks: denying a loan to a creditworthy applicant, resulting in a loss of business, or approving a loan to a borrower likely to default, thereby risking financial losses.

- The objective is to analyze the past loan applicant's data 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.

ANALYSIS APPROACH

The approach is to perform the following activities in sequence to identify the patterns for defaults:-

- Data Cleaning
- Univariate Analysis
- Bivariate Analysis

DATA CLEANING

Following Data Cleaning sequence applied on the data

1. Analyse dataset
   - Check columns for missing values
   - Process Data - Drop columns where more than 40% data is missing
2. Handling Incorrect data types
   - Remove special characters (%) from data
   - Convert data type to integer or float
3. Impute Missing Values
   - Remove rows where data is missing (less than 1% data) - As removing this will have no impact on the data set
4. Sanity checks
   - Check if Data is within the range as expected.
5. Outlier finding for numerical Data
   - Check for outlier data and process dataset by removing
6. Derived metrics
   - Create derived metrics from dataset, such as based on Annual Income, Interest rate, Loan amount

UNIVARIATE ANALYSIS

APPROACH

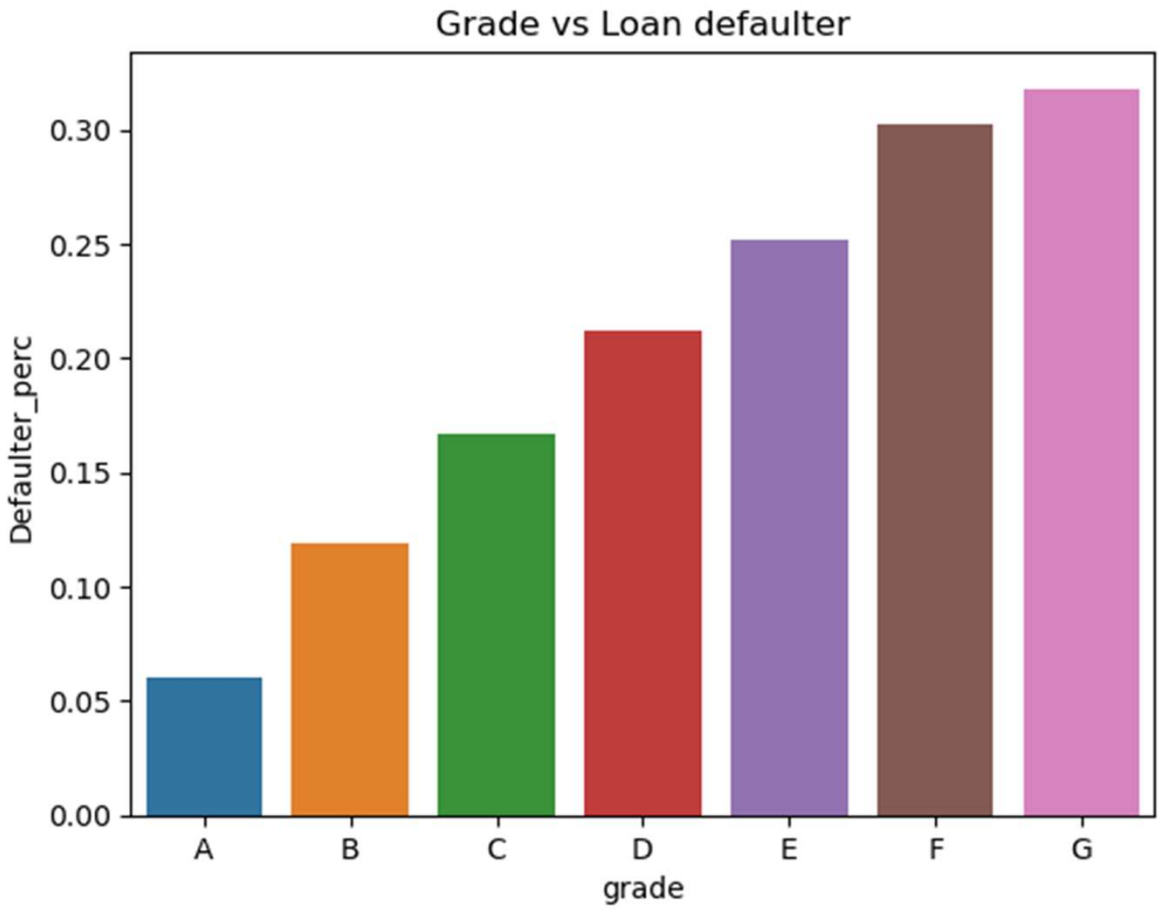
Grouping is done for Grade and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Grade. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented grades.

x-axis : grade

y-axis : Defaulter_perc (Percentage od defaulters in each Grade)

INFERENCE

In the bar plot, as the grade moves along the x-axis from A to G, the probability of default increases.



APPROACH

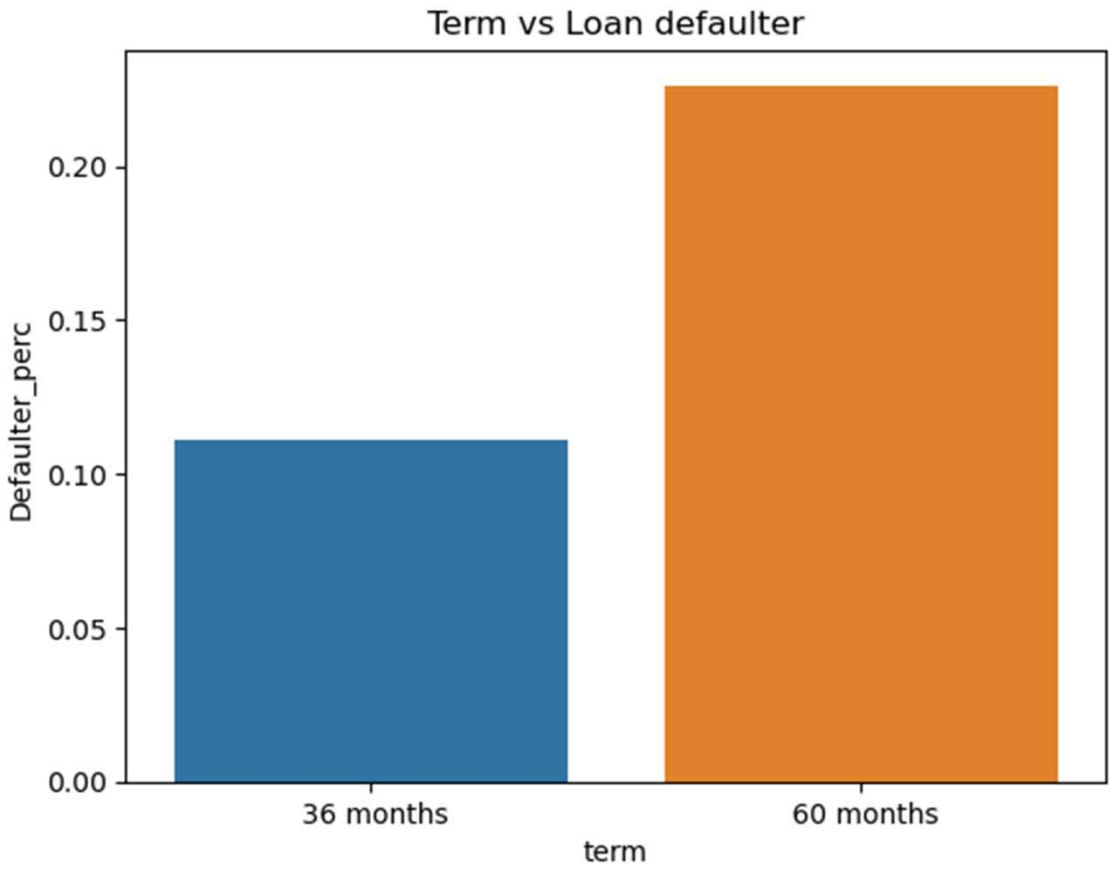
Grouping is done for term and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Term. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented terms.

x-axis : term

y-axis : Defaulter_perc (Percentage od defaulters in each Term)

INFERENCE

In the bar plot, the probability of default increases for loan having 60 months term.



APPROACH

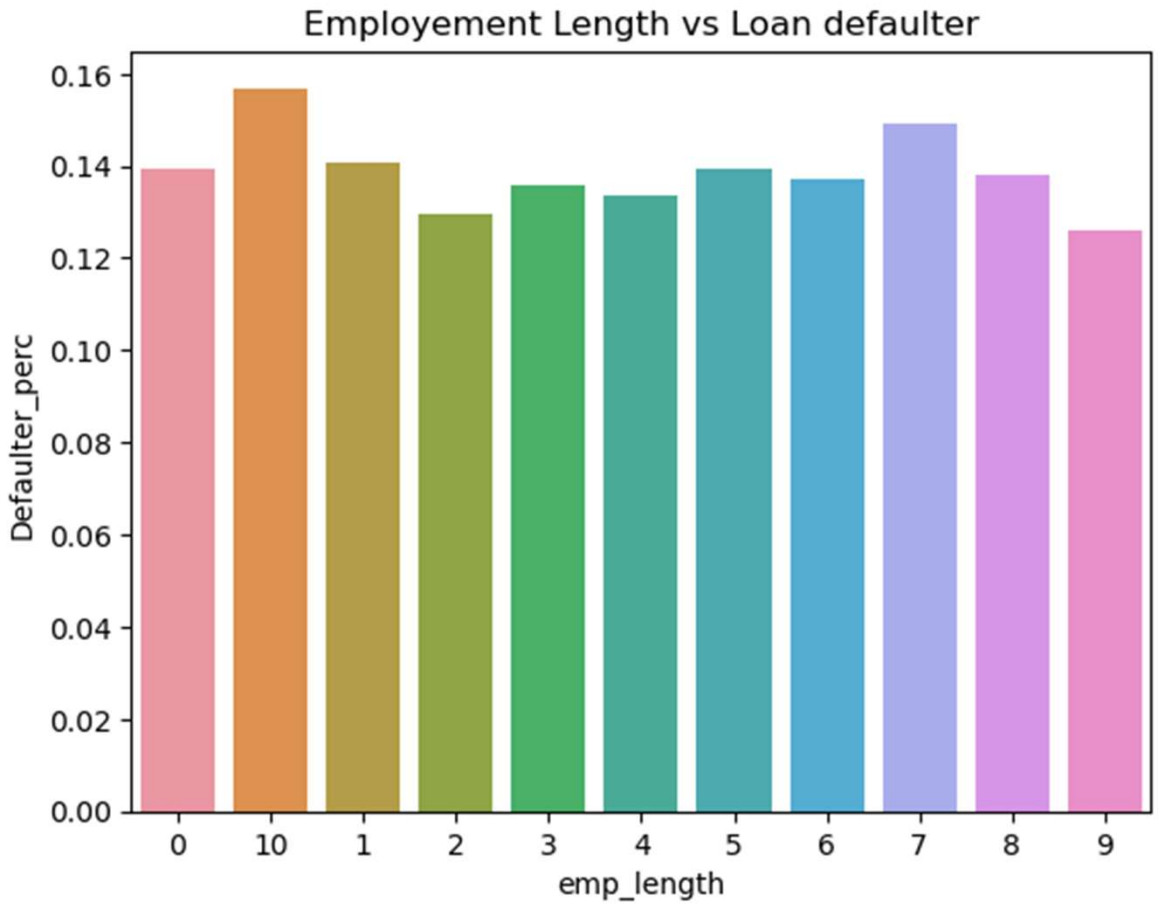
Grouping is done for Employee Length and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Employee Length. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Employee Length.

x-axis : emp_length (Employee Length)

y-axis : Defaulter_perc (Percentage od defaulters in each Employee Length)

INFERENCE

In the bar plot, No inference is concluded as all are almost in equal height.



APPROACH

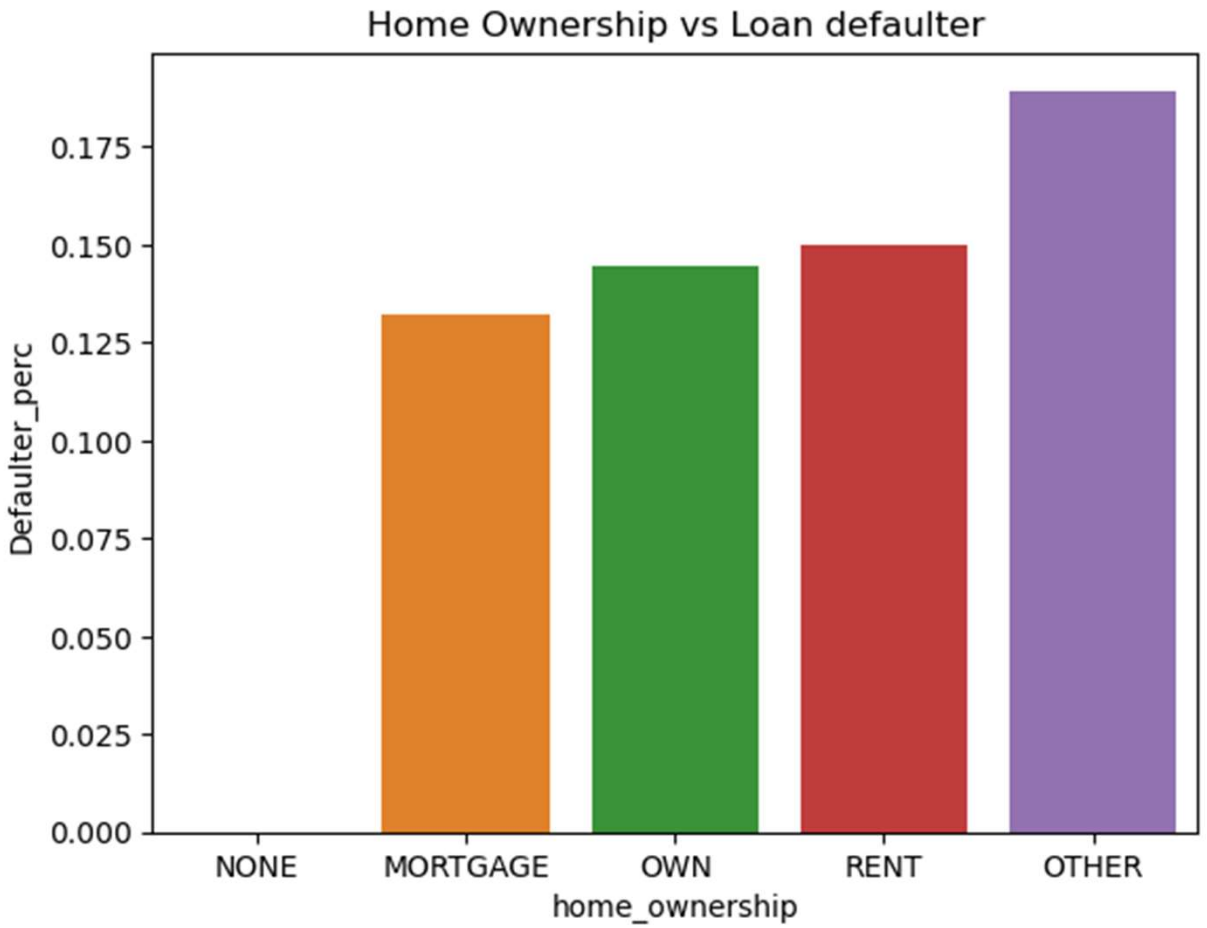
Grouping is done for Home Ownership and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Home Ownership. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Home Ownership.

x-axis : home_ownership (Home Ownership)

y-axis : Defaulter_perc (Percentage od defaulters in each Home Ownership)

INFERENCE

In the bar plot, The probability of default increases for Home Ownership as OTHER.



APPROACH

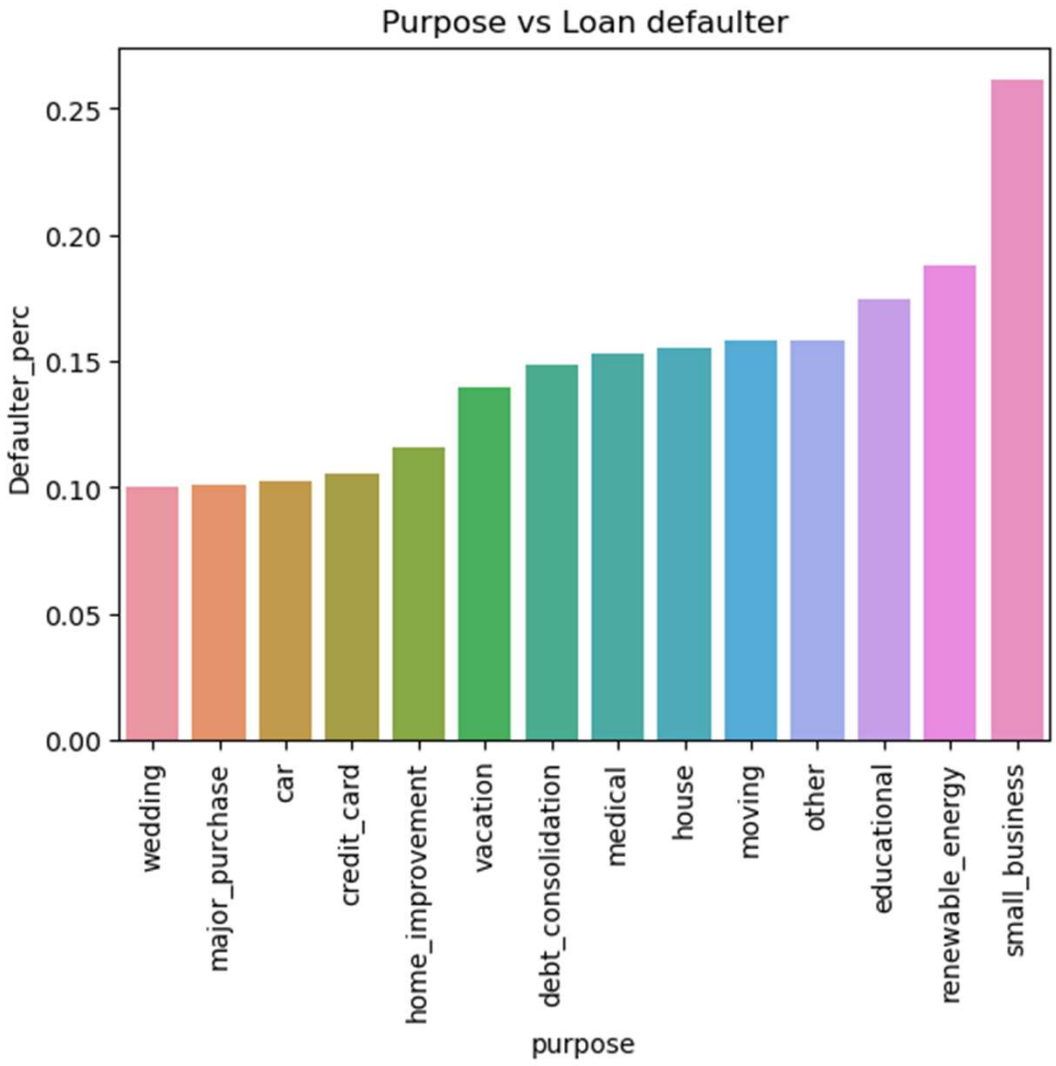
Grouping is done for Purpose and Loan Status variables. Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Purpose. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Purpose.

x-axis : purpose

y-axis : Defaulter_perc (Percentage od defaulters in each Purpose)

INFERENCE

In the bar plot, Small businesses are considered riskier, as their default percentage is notably high.



APPROACH

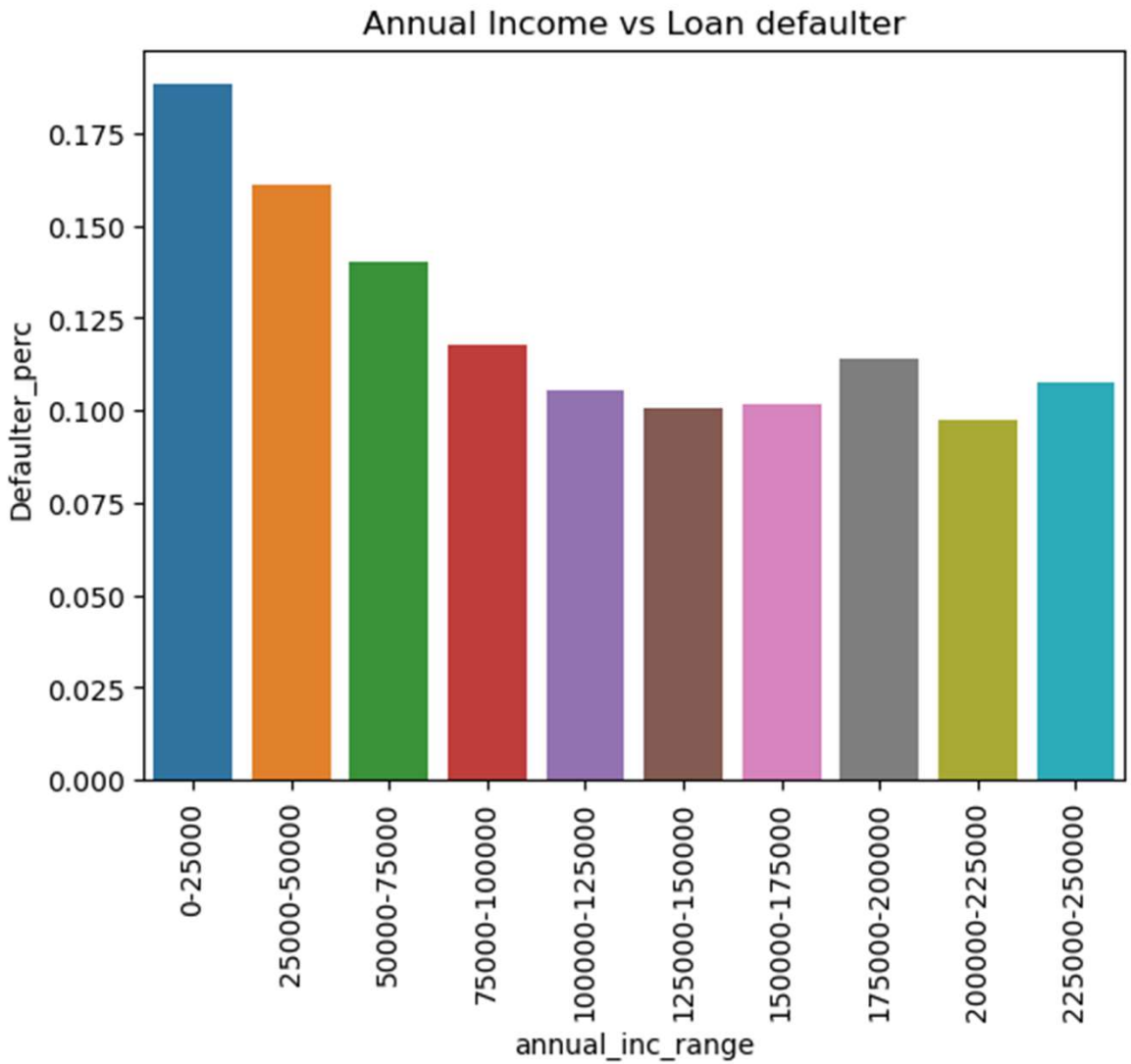
Grouping is done for Annual Income Range and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Annual Income Range. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Annual Income Range.

x-axis : annual_inc_range (Annual Income Range)

y-axis : Defaulter_perc (Percentage od defaulters in each Annual Income Range)

INFERENCE

In the bar plot, There are almost 18% defaulters whose Annual Income Range is less than or equal to 25000 and 16% defaulters for Annual Income Range between 25000 – 50000. Bank should be more vigilant on giving loan to customers whose Annual Income Range is less than or equal to 50000.



APPROACH

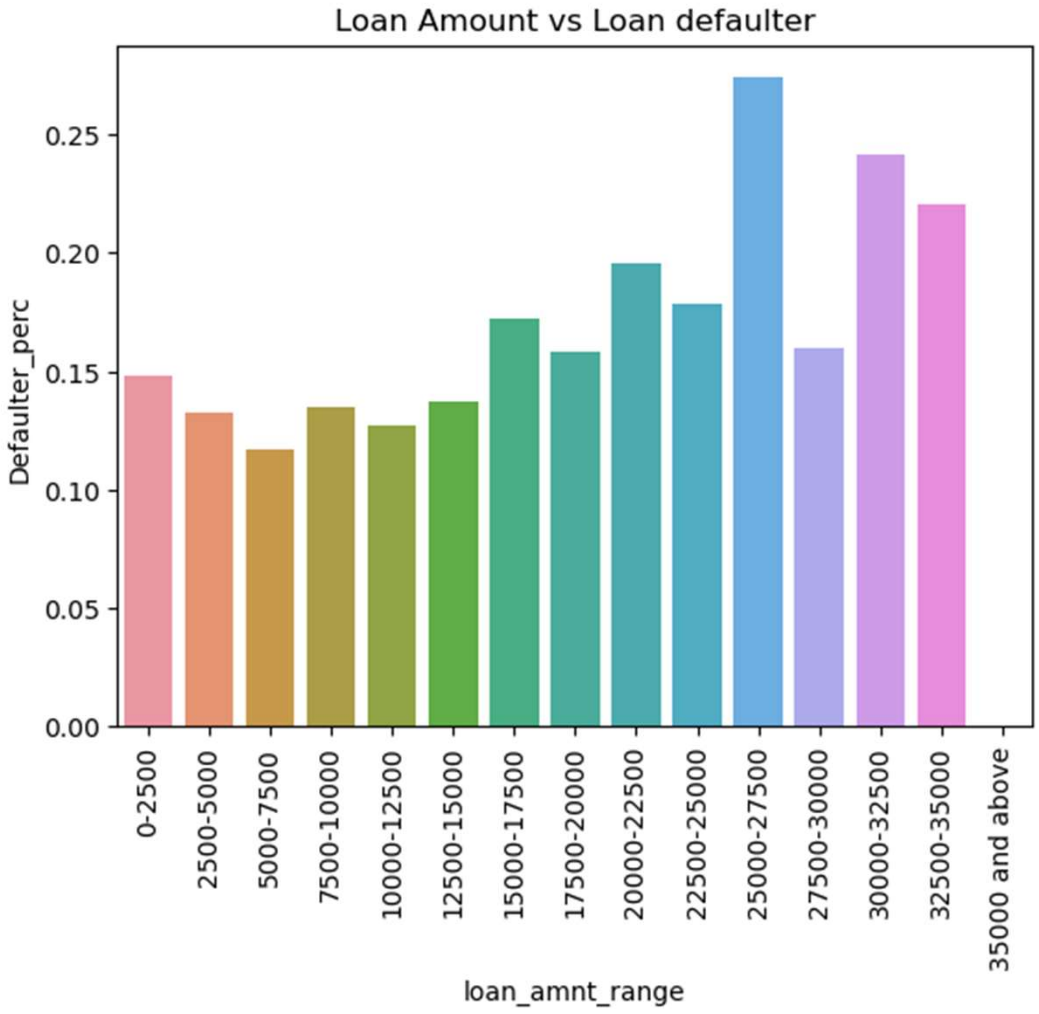
Grouping is done for Loan Amount and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Loan Amount Range. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Loan Amount Range.

x-axis : loan_amnt_range (Loan Amount Range)

y-axis : Defaulter_perc (Percentage od defaulters in each Loan Amount Range)

INFERENCE

In the bar plot, More defaulters are observed in the loan amount range of 25000 – 27500 followed by 30000 and above.



APPROACH

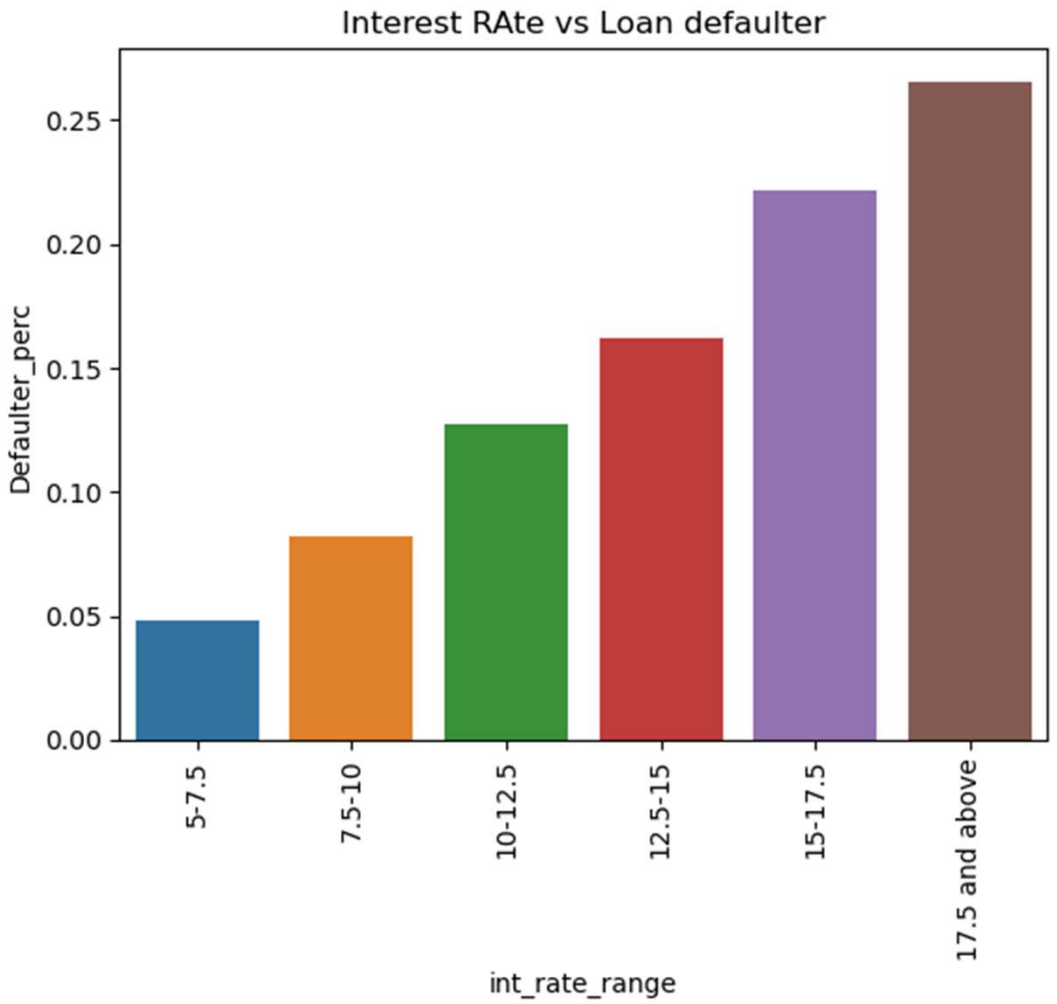
Grouping is done for Interest Rates and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Interest Rates Range. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Interest Rates Range.

x-axis : int_rate_range (Interest Rates Range)

y-axis : Defaulter_perc (Percentage od defaulters in each Interest Rates Range)

INFERENCE

In the bar plot, Its likely that the number of defaulters will increase as interest rates rise.



APPROACH

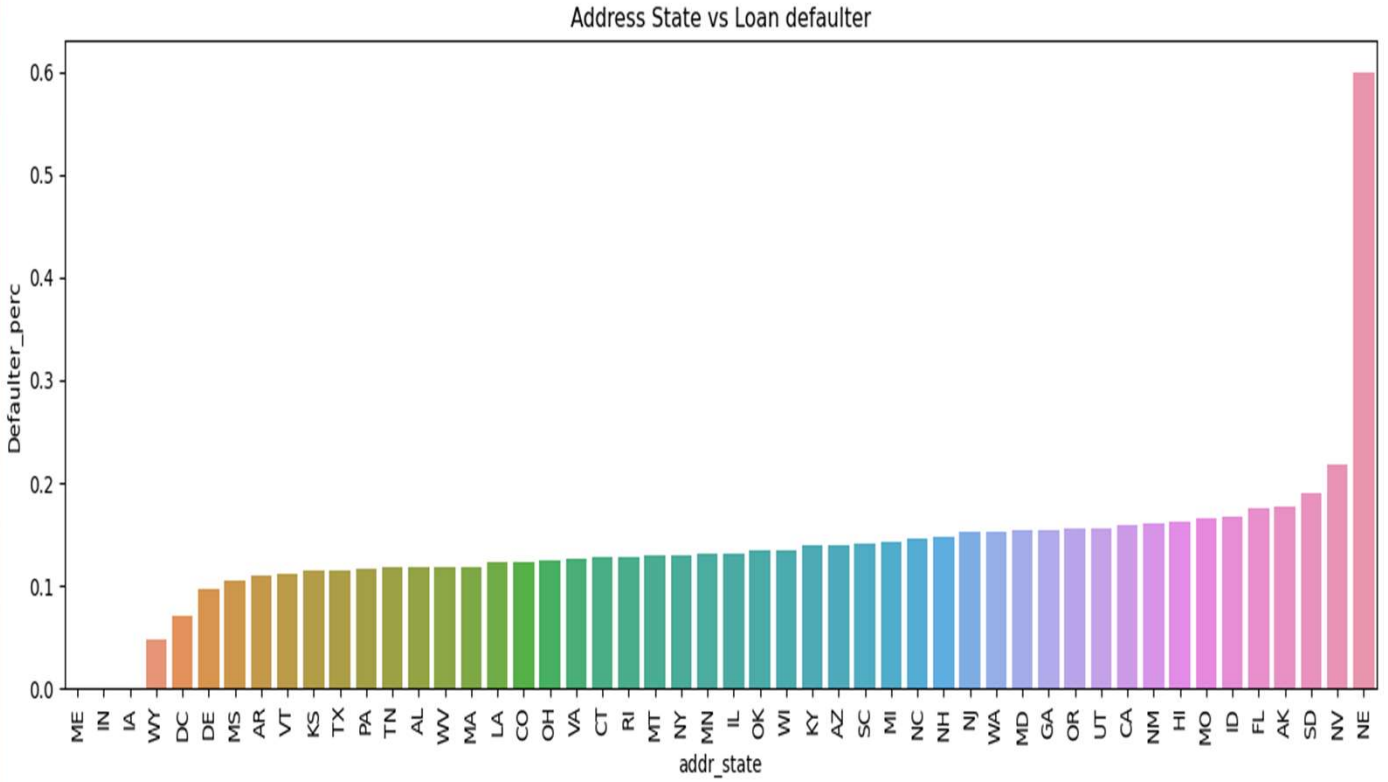
Grouping is done for Address State and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each Address State. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented Address State.

x-axis : addr_state (Address State)

y-axis : Defaulter_perc (Percentage od defaulters in each Address State)

INFERENCE

In the bar plot, NE state is showing high percentage of defaulters, but the total customers count is very minimal , so its not considerable. The NV state can be taken into consideration for having a high number of defaulters , as this state has reasonable amount of loan customers.



APPROACH

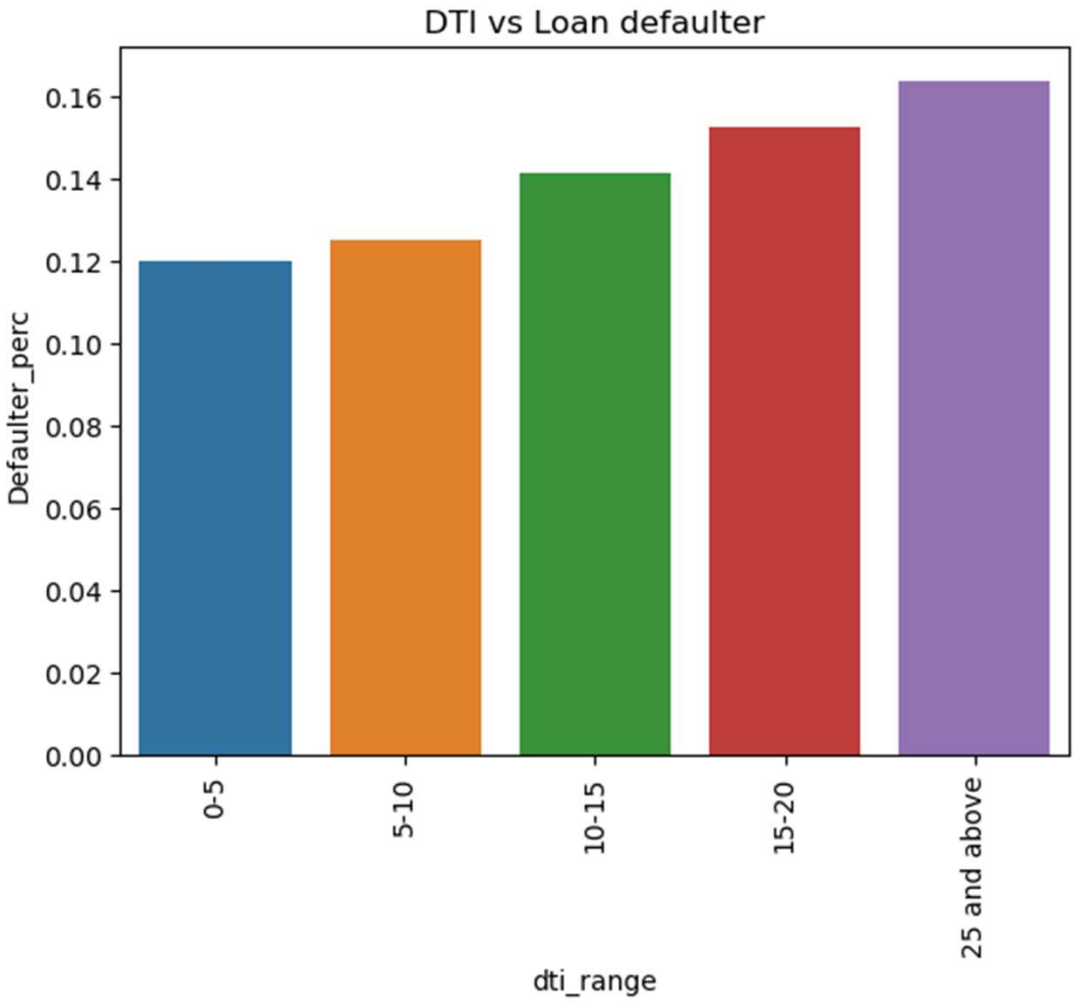
Grouping is done for DTI and Loan Status variables . Derived Defaulter_perc as new variable in the group that shows percentages of defaulters for each DTI range. By using bar plot, Segmented univariate analysis is done on defaulters across different segmented DTI Range.

x-axis : dti_range (dti range)

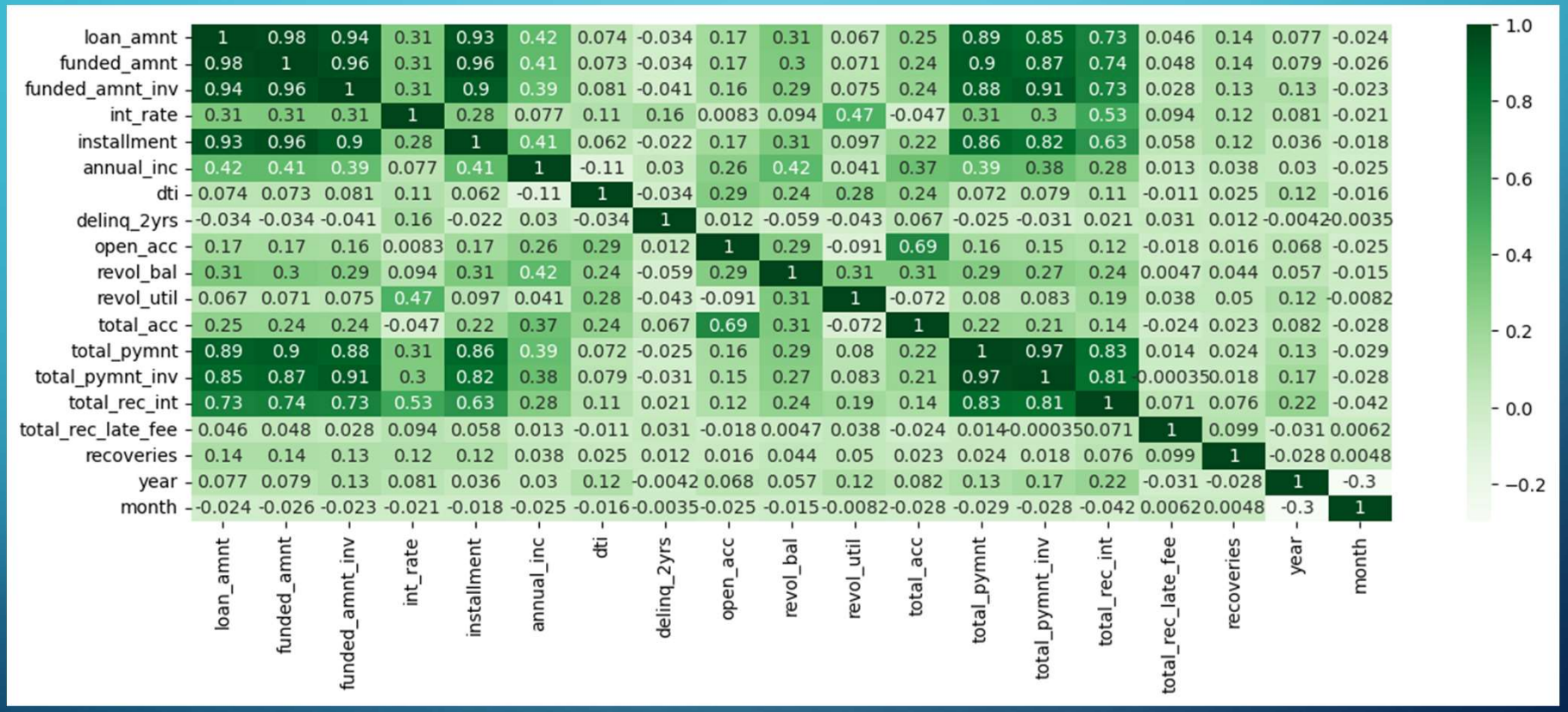
y-axis : Defaulter_perc (Percentage od defaulters in each dti range)

INFERENCE

In the bar plot, Its likely that number of defaulters will increase as DTI value rise.



BIVARIATE ANALYSIS AMONG NUMERICAL VARIABLES


loan_amnt 1 0.98 0.94 0.31 0.93 0.42 0.074 -0.034 0.17 0.31 0.067 0.25 0.89 0.85 0.73 0.046 0.14 0.077 -0.024
funded_amnt 0.98 1 0.96 0.31 0.96 0.41 0.073 -0.034 0.17 0.3 0.071 0.24 0.9 0.87 0.74 0.048 0.14 0.079 -0.026
funded_amnt_inv 0.94 0.96 1 0.31 0.9 0.39 0.081 -0.041 0.16 0.29 0.075 0.24 0.88 0.91 0.73 0.028 0.13 0.13 -0.023
int_rate 0.31 0.31 0.31 1 0.28 0.077 0.11 0.16 0.0083 0.094 0.47 -0.047 0.31 0.3 0.53 0.094 0.12 0.081 -0.021
installment 0.93 0.96 0.9 0.28 1 0.41 0.062 -0.022 0.17 0.31 0.097 0.22 0.86 0.82 0.63 0.058 0.12 0.036 -0.018
annual_inc 0.42 0.41 0.39 0.077 0.41 1 -0.11 0.03 0.26 0.42 0.041 0.37 0.39 0.38 0.28 0.013 0.038 0.03 -0.025
dti 0.074 0.073 0.081 0.11 0.062 -0.11 1 -0.034 0.29 0.24 0.28 0.24 0.072 0.079 0.11 -0.011 0.025 0.12 -0.016
delinq_2yrs -0.034 -0.034 -0.041 0.16 -0.022 0.03 -0.034 1 0.012 -0.059 -0.043 0.067 -0.025 -0.031 0.021 0.031 0.012 -0.0042 -0.0035
open_acc 0.17 0.17 0.16 0.0083 0.17 0.26 0.29 0.012 1 0.29 -0.091 0.69 0.16 0.15 0.12 -0.018 0.016 0.068 -0.025
revol_bal 0.31 0.3 0.29 0.094 0.31 0.42 0.24 -0.059 0.29 1 0.31 0.31 0.29 0.27 0.24 0.0047 0.044 0.057 -0.015
revol_util 0.067 0.071 0.075 0.47 0.097 0.041 0.28 -0.043 -0.091 0.31 1 -0.072 0.08 0.083 0.19 0.038 0.05 0.12 -0.0082
total_acc 0.25 0.24 0.24 -0.047 0.22 0.37 0.24 0.067 0.69 0.31 -0.072 1 0.22 0.21 0.14 -0.024 0.023 0.082 -0.028
total_pymnt 0.89 0.9 0.88 0.31 0.86 0.39 0.072 -0.025 0.16 0.29 0.08 0.22 1 0.97 0.83 0.014 0.024 0.13 -0.029
total_pymnt_inv 0.85 0.87 0.91 0.3 0.82 0.38 0.079 -0.031 0.15 0.27 0.083 0.21 0.97 1 0.81 -0.00035 0.018 0.17 -0.028
total_rec_int 0.73 0.74 0.73 0.53 0.63 0.28 0.11 0.021 0.12 0.24 0.19 0.14 0.83 0.81 1 0.071 0.076 0.22 -0.042
total_rec_late_fee 0.046 0.048 0.028 0.094 0.058 0.013 -0.011 0.031 -0.018 0.0047 0.038 -0.024 0.014 -0.00035 0.071 1 0.099 -0.031 0.0062
recoveries 0.14 0.14 0.13 0.12 0.12 0.038 0.025 0.012 0.016 0.044 0.05 0.023 0.024 0.018 0.076 0.099 1 -0.028 0.0048
year 0.077 0.079 0.13 0.081 0.036 0.03 0.12 -0.0042 0.068 0.057 0.12 0.082 0.13 0.17 0.22 -0.031 -0.028 1 -0.3
month -0.024 -0.026 -0.023 -0.021 -0.018 -0.025 -0.016 -0.0035 -0.025 -0.015 -0.0082 -0.028 -0.029 -0.028 -0.042 0.0062 0.0048 -0.3 1
loan_amnt
funded_amnt
funded_amnt_inv
int_rate
installment
annual_inc
dti
delinq_2yrs
open_acc
revol_bal
revol_util
total_acc
total_pymnt
total_pymnt_inv
total_rec_int
total_rec_late_fee
recoveries
year
month
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

CORRELATION AMONG NUMERICAL VARIABLES

Points to be concluded from the graph presented before.

- Loan Amount, Funded Amount, Funded Amount Inv, Total Payment, Total Payment Inv, Total Received Interest and Installments are formed a cluster of positive correlation. Means, all these variables have same kind of relation with Loan Status.

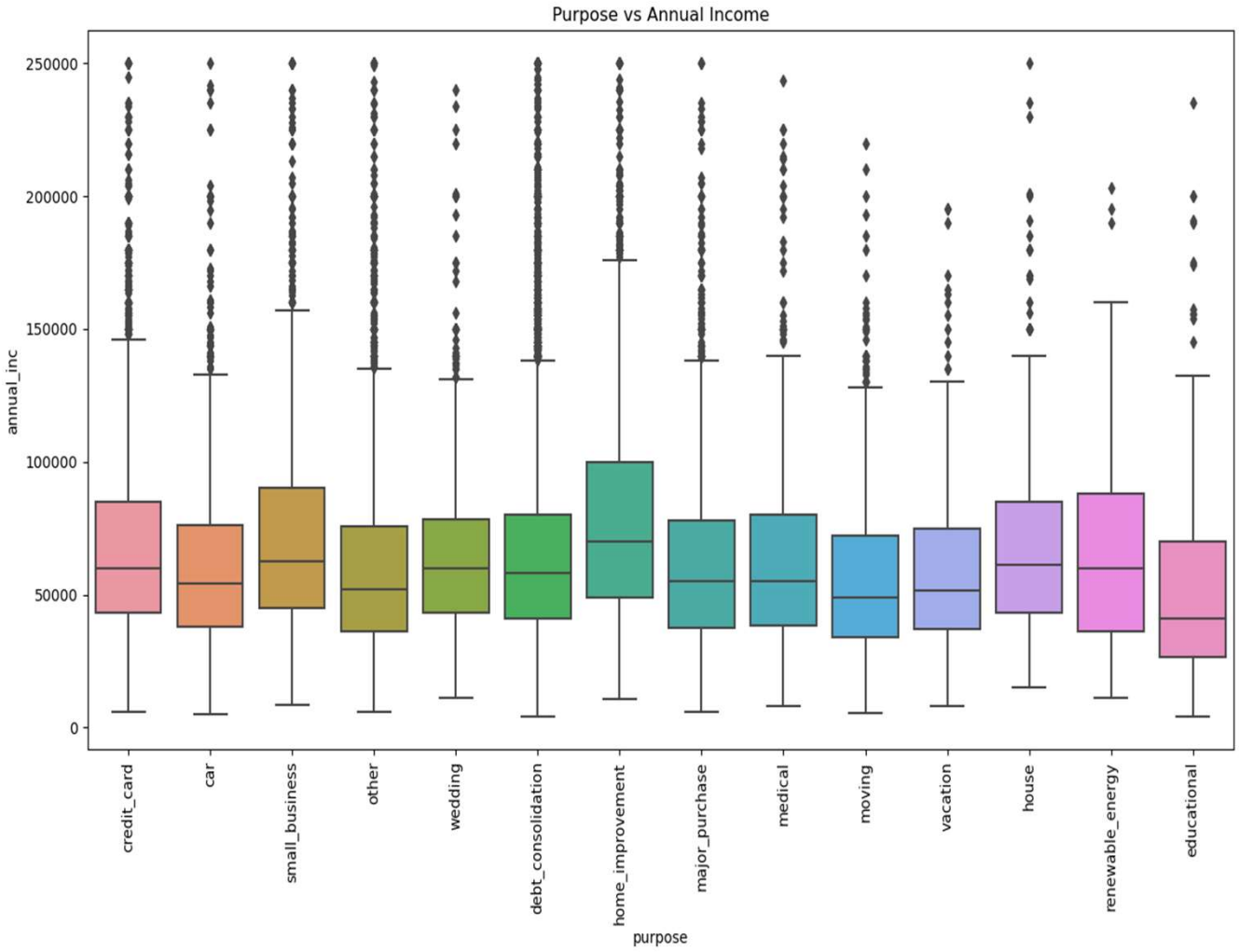
BIVARIATE ANALYSIS BETWEEN CATEGORICAL VARIABLES

APPROACH

IN the previous slides, Segmented Univariate Analysis showcased Small Business purpose has high default percentage.
This Box Plot is used for Bivariate Analysis to showcase the relation between Purpose and other numerical variable like Annual Income.

INFERENCE

IN the Box Plot, Its likely that Highly annual income people having small business purpose loan will fall into more defaults compare to lower annual income customers

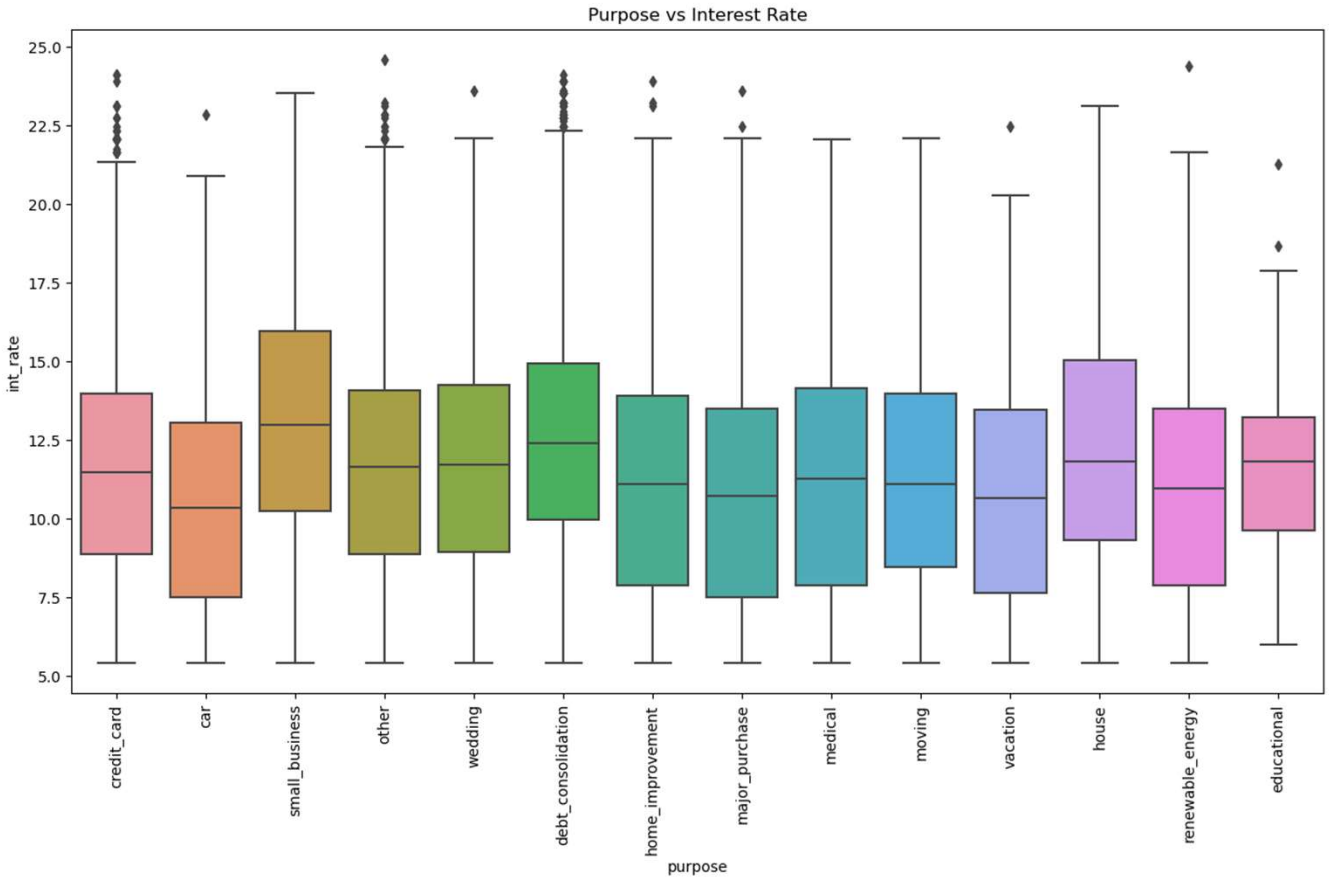


APPROACH

IN the previous slides, Segmented Univariate Analysis showcased Small Business purpose has high default percentage.
This Box Plot is used for Bivariate Analysis to showcase the relation between Purpose and other numerical variable like Interest rate.

INFERENCE

IN the Box Plot, Its likely that Small Business purpose loans having High interest rates would fall into more defaults.

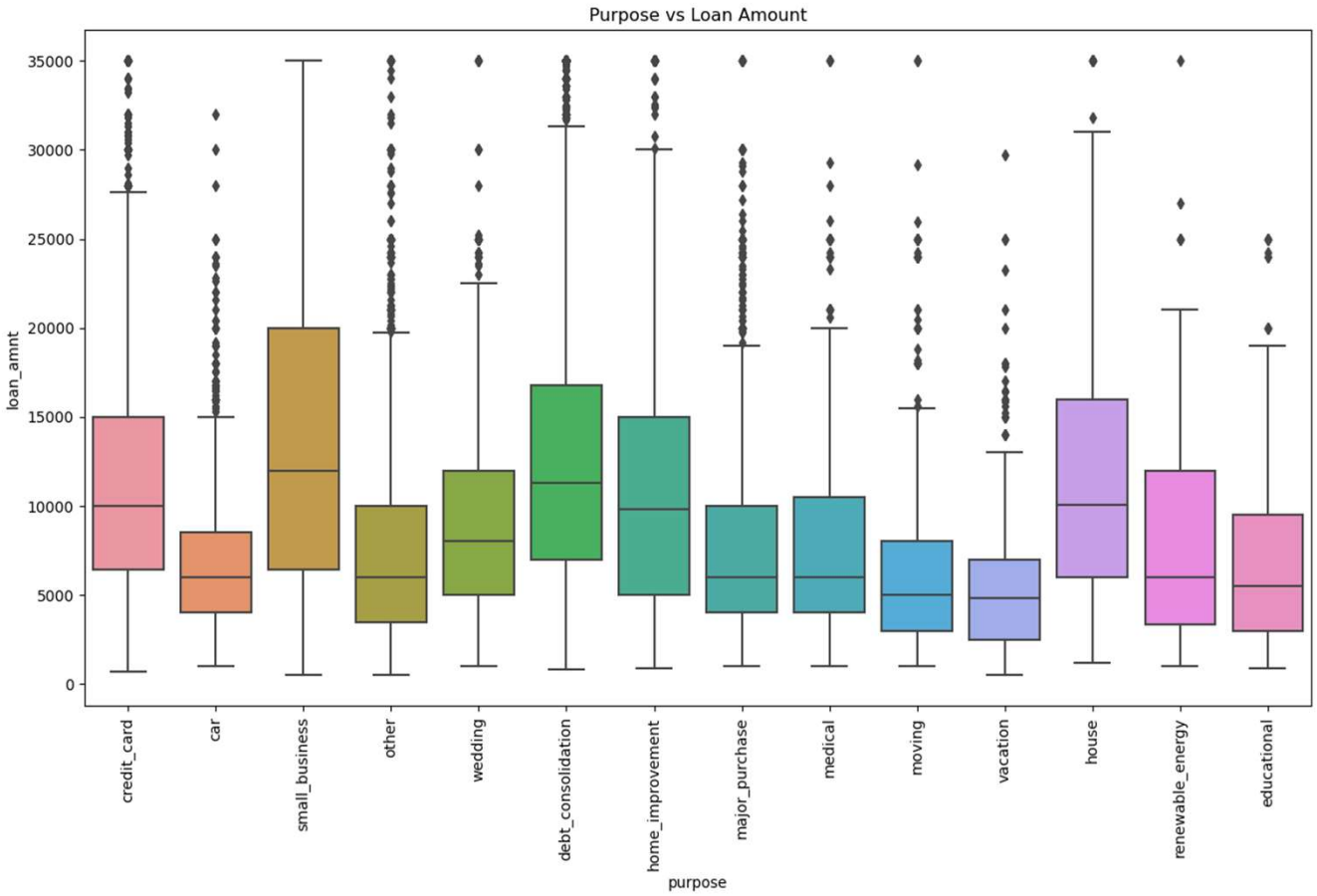


APPROACH

IN the previous slides, Segmented Univariate Analysis showcased Small Business purpose has high default percentage.
This Box Plot is used for Bivariate Analysis to showcase the relation between Purpose and other numerical variable like Loan Amounts.

INFERENCE

IN the Box Plot, Its likely Small Business purpose loans having High loan amounts would fall into more defaults.



CONCLUSION

1. Small business are the riskiest to give loan to.
2. Small business should be given loan at lower Interest Rate or less Loan amount, as these will reduce the risk of default.
3. As the Grades move from A to G, default is increasing. Bank can focus of this parameter.
4. Shorter duration loans(36 months) are paid more often

THANK YOU