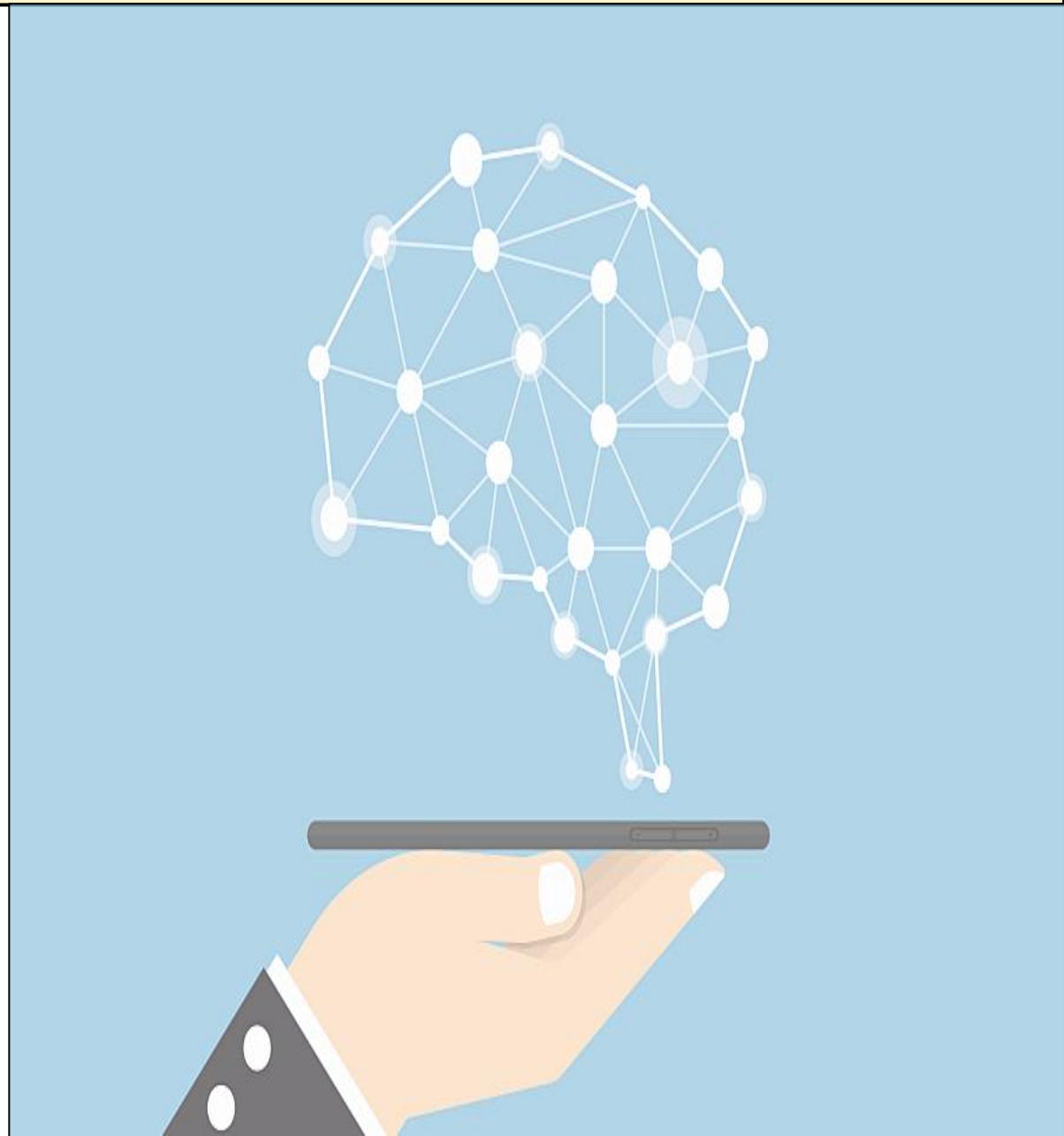
Lending Club Case Study

By Akhil Suresh and Irfan Khan Mohammed

Key Findings

- Most of the customers who defaulted had taken loans for running **small business ,debt consolidation and credit cards**
    - **Low annual income** and **higher interest rates** were among the reasons for customers belonging to this category to default
- Factors like **Higher DTI** , Higher number of **open credit accounts** have an impact on defaulting
- Factors like **history of bankruptcies** and **public derogatory records** also influence the customers chances of defaulting , higher the numbers greater than chance
- Customers with annual income of around $ 0-20000 , has higher chances of charge off/defaulting
- There's a higher charge off percentage for loans taken on smaller tenure **(term=36 months)** mostly due to higher installment amounts

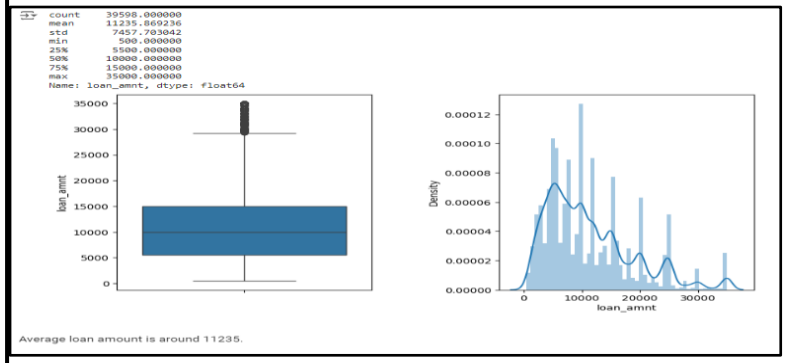


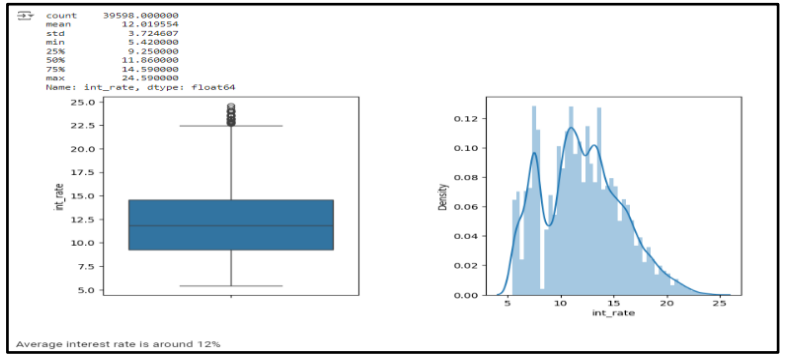
Details on Coding, Data points and Summary

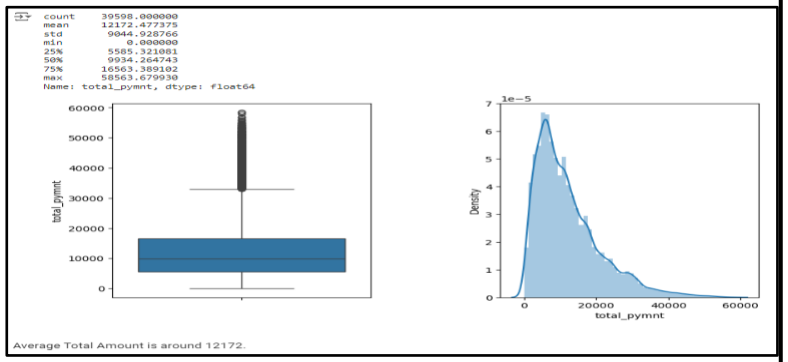
1. Initial Phase
   a. Imported all Libraries to be used for the Project.
   b. Set the book to display all rows and columns.
   c. Loading and reading loan.csv file for Analysis.
   d. Displaying the first 5 rows from the loan.csv file.
   e. Looking the Dimension of the data loaded.
2. Data Preprocessing
   a. Checked and Validated Data Types
      i. Data has 39717 rows and 111 Columns.
      ii. Consist of DTypes : float64, int64 and Object.
3. Data Cleaning
   a. Validated Data for null values - Using mean values and multiplied by 100 and in descending order.
   b. Found 111 Columns 55 Columns had 100% rows with null values, 43 columns with Zero rows with null values, 13 columns between 0.1% to 99.9% null values.
   c. Selected the columns wherein all rows are null.
   d. Validated Rows for all null values.
   e. Dropping all columns with null values and drop columns those are not relevant.
   f. In 111 columns 63 columns were dropped.
   g. Analysis to perform on 48 columns and 39598 rows.
   h. Added Columns Year, Month, Loan_to_Approve and Defaulter.
4. Data Transformation
   a. Removed strings in data for columns.
   b. Treated the missing values.
   c. Filled na with 0 and unknown.
   d. Validated for duplicate entries.
   e. Added columns year and month from issue date.
   f. Separated Numerical and Categorical Columns.
   g. Added a new column named Defaulter and Loan_to_Accept from column loan_Status.
5. Data Analysis
   a. Univariate Analysis
      i. Found Outlier using box plot for all columns in numerical category.
      ii. Checked the distribution of the numeric columns using Distribution plot.
      iii. Performed analysis on Loan Amount, Interest rate, Total Amount, Annual Income using box and distribution plot.
      iv. Removed outliers from the Annual Income column kept it at 95% quantile with box and distribution plot.
      v. Performed Univariate Categorical Analysis on Loan Status, Purpose, Ownership Public Record Bankruptcies, , LC Grade using plot graph.
   b. Bivariate Analysis
      i. Ratio between loan amount and funded amount.
      ii. Reaggregation b/w Funded amount and Annual Income.
      iii. Minimum estimator of loan amount and customer grade.
      iv. Loan amount wherein we see most defaulters.
      v. Loan Acceptance chances in listed Customers.
      vi. Correlation of funded amount every annually.
      vii. Highest and Lowest funding on monthly basis
      viii. Heat & Plot Maps.
      ix. Loan taken What purpose has higher charge off.
      x. Higher Charge Off % multiple columns.
      xi. Avg variations for the all the imp numeric columns
      xii. Purpose why higher loan amount is taken?
      xiii. Loan Status against other columns was performed.
      xiv. Purpose on higher loan amount is taken and interest.
      xv. The highest dept to income ratio against Loan Status and interest rate.

Univariate Analysis

1. Performed Analysis on Loan Amount, Interest Rate, Total Amount, Annual Income distribution using box plot and distribution plot:







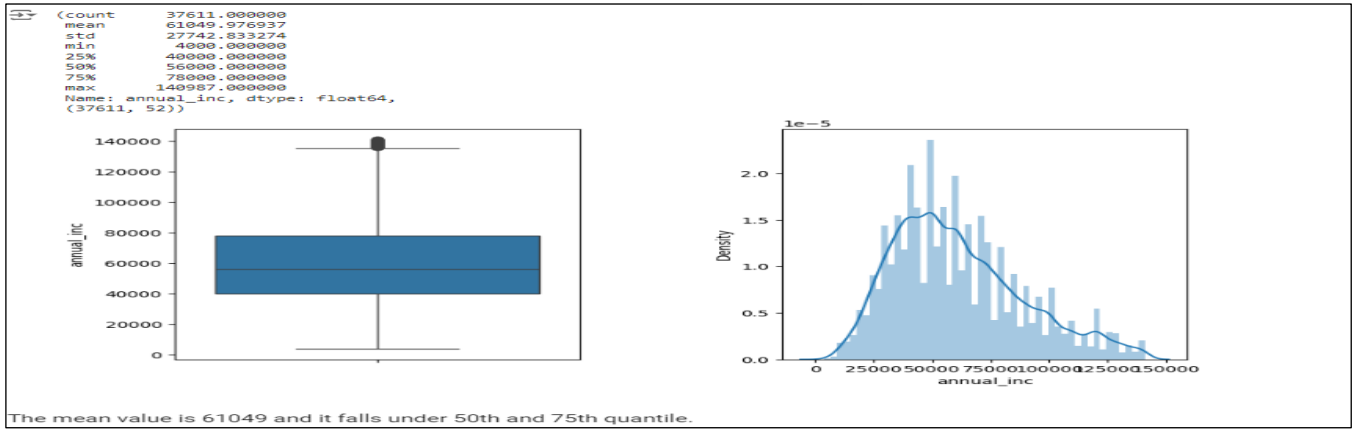
Conclusion

a. Average loan amount is around 11235.
b. Average interest rate is around 12%
c. Average Total Amount is around 12172.
d. Average Annual Income is around 69000.

2. Removed outliers from the Annual Income column kept it at 95% quantile with boxplot and distribution plot

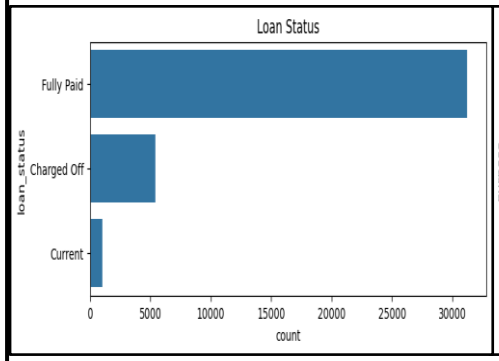
Conclusion

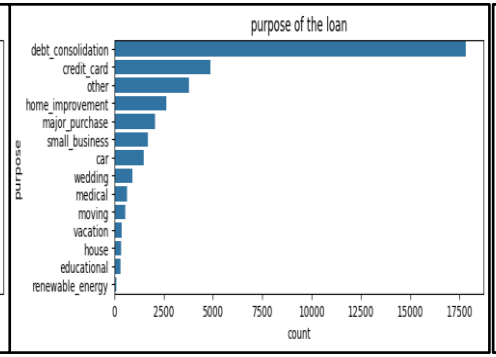
The mean value is 61049 and it falls under 50th and 75th quantile.

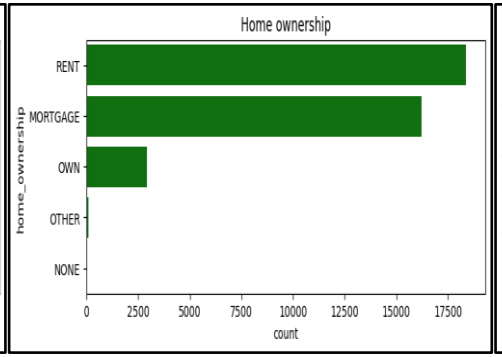


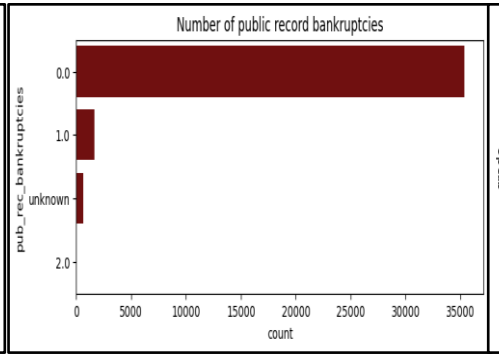
Univariate Analysis - Continued

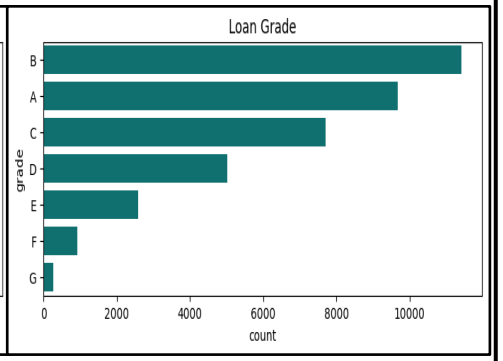
1. Performed Univariate Categorical Analysis on Loan Status, Purpose, Ownership Public Record Bankruptcies, , LC Grade using plot graph.







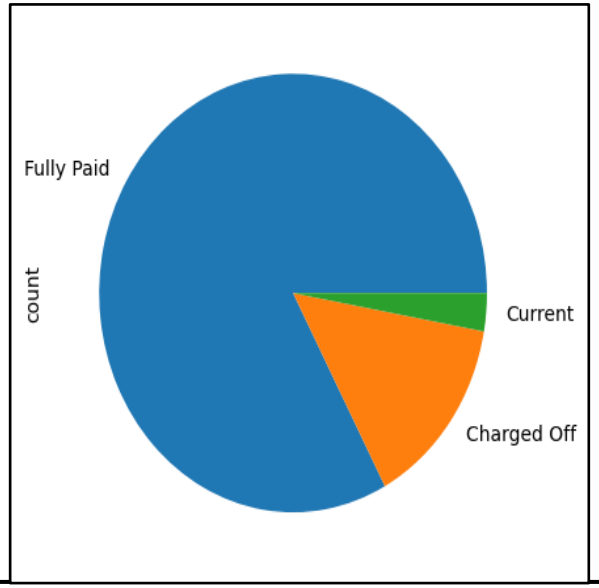




Conclusion

a. Around 83% loan are fully paid, 14% are Charged off (delinquent) and 3% are Current..
b. Top 5 Loan Purpose round off are as follows:
   i. Debt_consolidation = 47%
   ii. credit_card = 13%
   iii. Other = 10%
   iv. home_improvement = 7%
   v. major_purchase = 6%
c. Top 3 Home Ownership round off are as follows:
   i. Rent = 49%
   ii. Mortgage = 43%
   iii. Own = 8%.
d. Close to 96% have 0 public record bankruptcies.
e. Top 3 LC Grade category are as follows:
   i. B = 30%
   ii. A = 26%
   iii. C = 20%

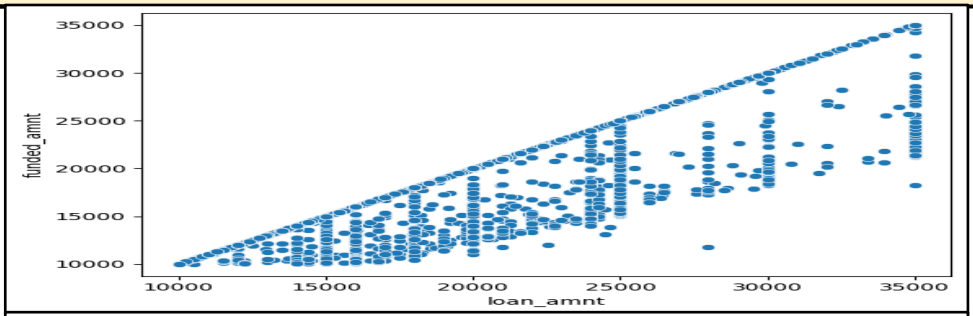
2. Pie Chart representing Loan Status.



Bivariate Analysis

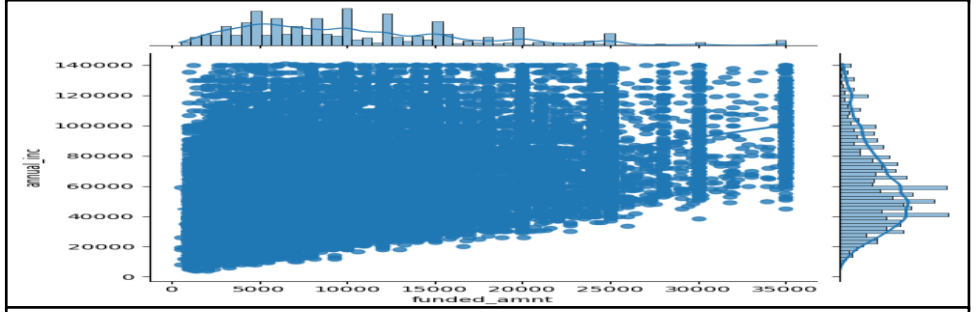
1. What is ratio between loan amount and funded amount >=10000.

From the scatter plot we noticed:
Loan amount and Funding are not equal.
Most likely Customers with lower loan amount to receive higher funding (within the loan amount).



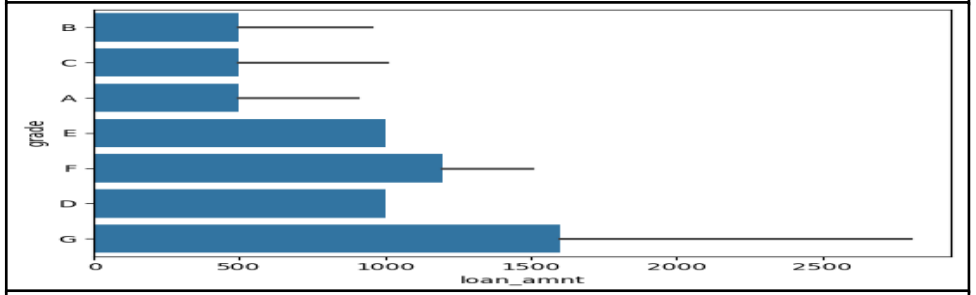
2. What is the Reaggregation between Funded amount and Annual Income?

Seems like the higher the income lower is customer opting for loan.



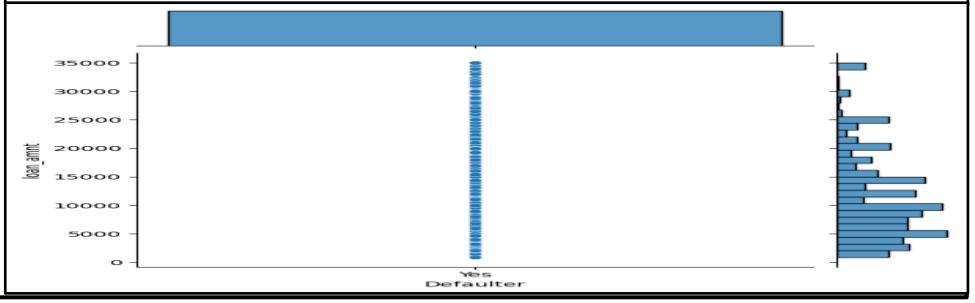
3. What is the minimum estimator of loan amount and customer grade?

Seems like Grade G customers has highest minimum estimator.



4. What is loan amount wherein we see most defaulters.

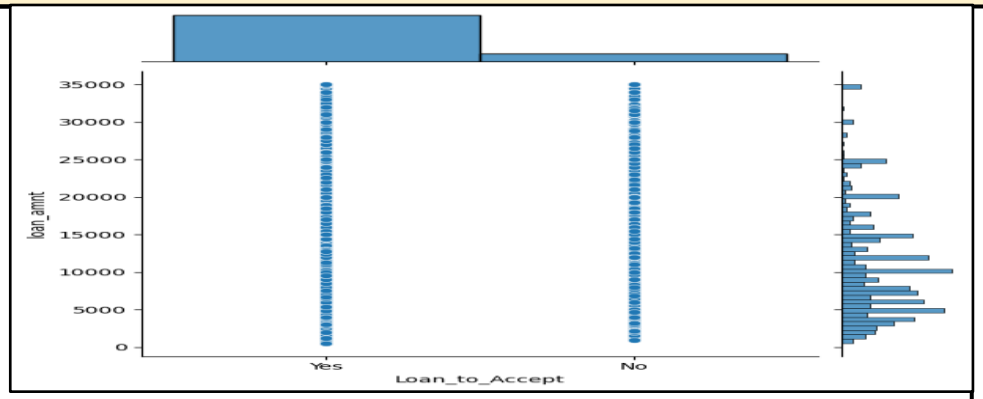
From above joint plot we noticed:
The correction between Loan amount and Defaulters is picked at 5000, 10000 and 15000.



Bivariate Analysis - Continued

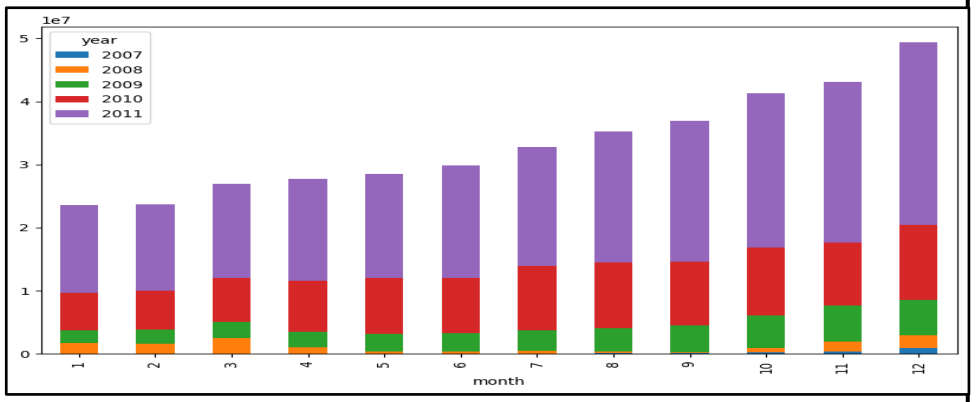
5. What is Loan Acceptance chances in listed Customers?

From above joint plot we noticed:
We noticed higher acceptance compared to rejection.



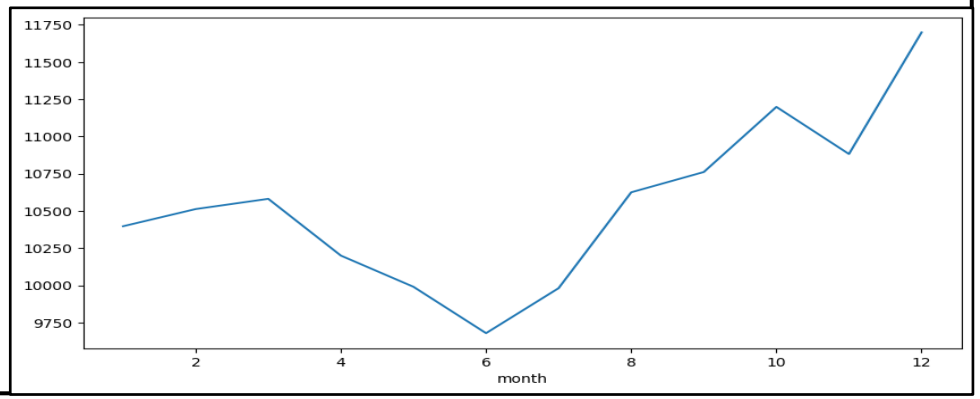
6. What is the correlation of funded amount every month on yearly basis?

From above stacked bar plot we noticed:
Year on year on monthly basis funding amount has been increasing.



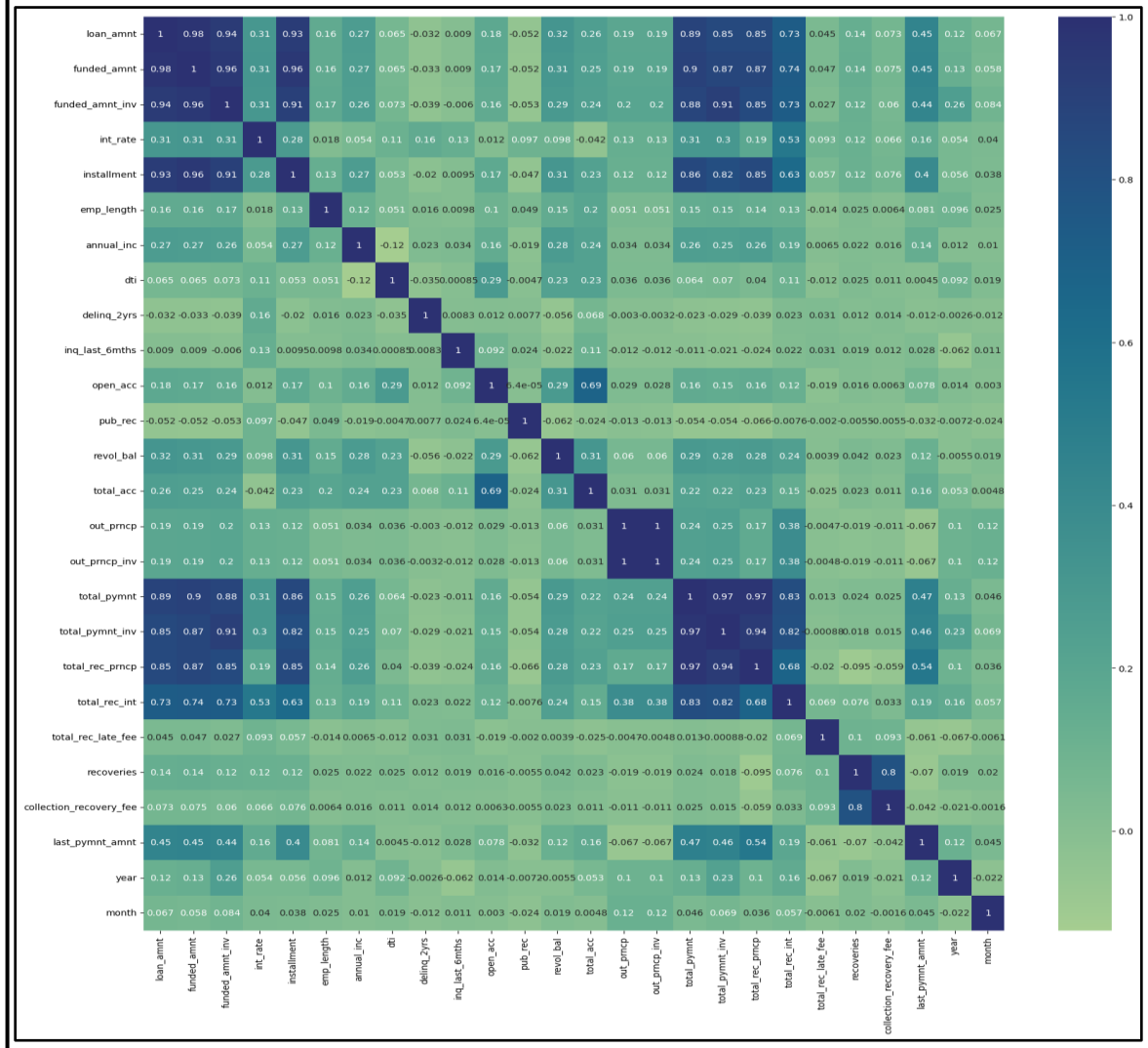
7. What is highest and lowest funding on monthly basis?

From above Line plot we noticed:
Funding is highest in December and lowest in June.

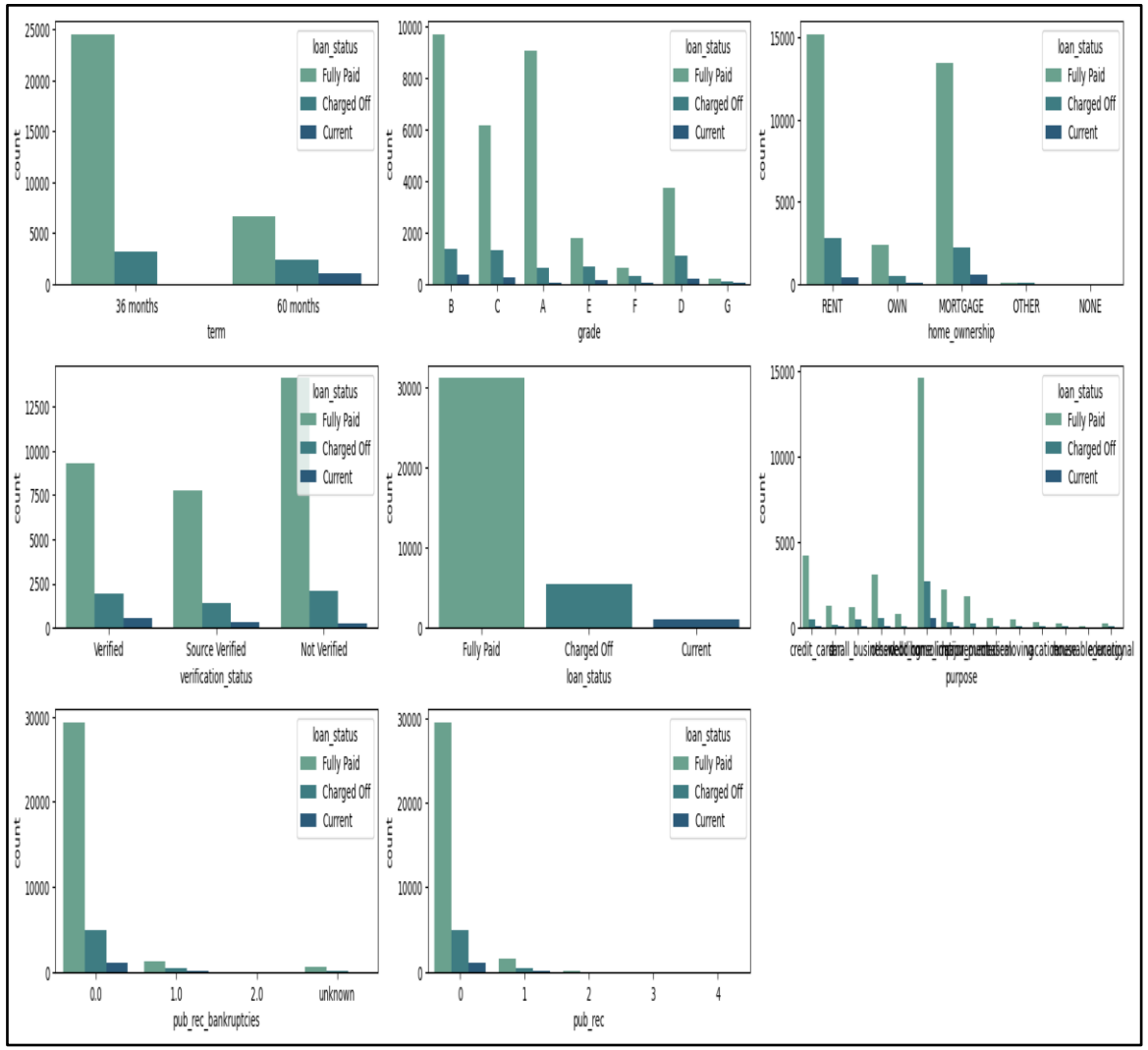


Bivariate Analysis - Continued

8. Heatmap for all numerical values under table loan_df2



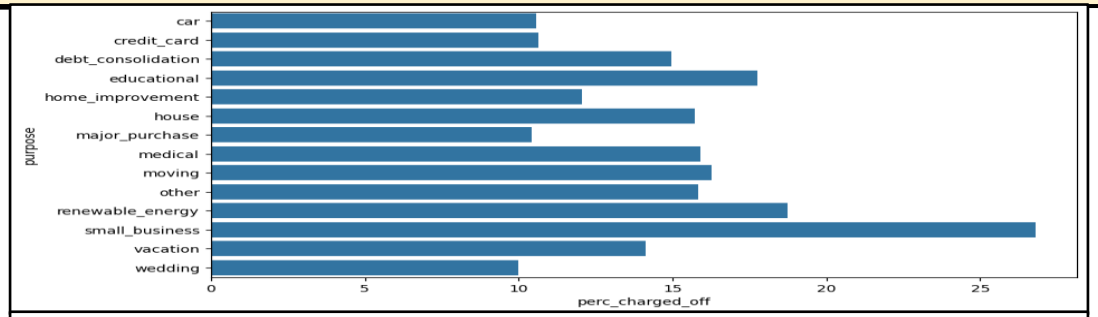
9. Plot graph showing term, grade, home_ownership, verification_status, purpose, pub_rec_bankruptcies, pub_rec in hue with loan status



Bivariate Analysis - Continued

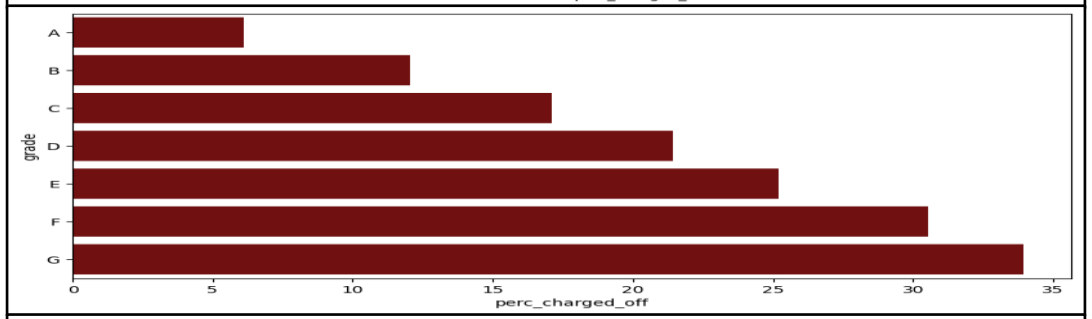
10. Loan taken What purpose has higher charge off?.

Conclusion: Loan taken for small business purposes see higher charge off.



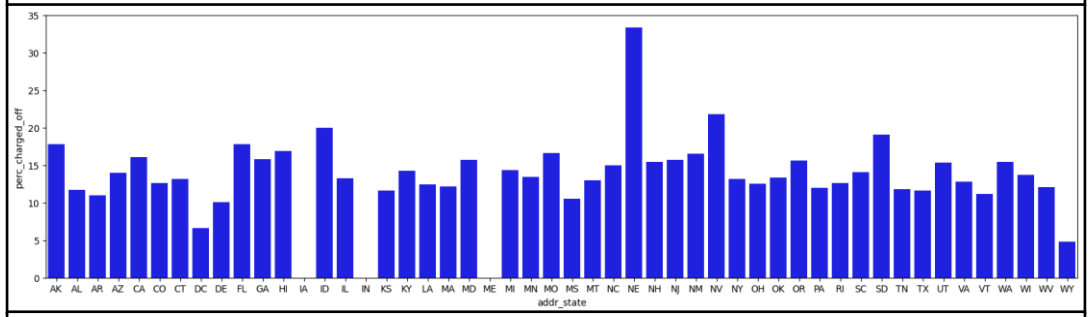
11. What is loan application with higher charge off percentage?

Conclusion: Loan application with Grade G has higher charge off percentage.



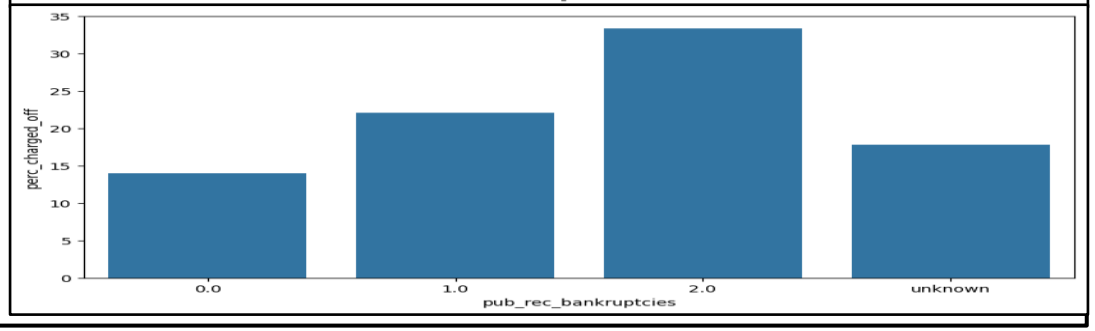
12. Which state has significantly higher charge off percentage?

Conclusion: NV(Nevada) state has significantly higher charge off percentage and NE(Nebraska)'s higher charge off due to lower total.



13. Which customers are with higher public bankruptcies charge off %?

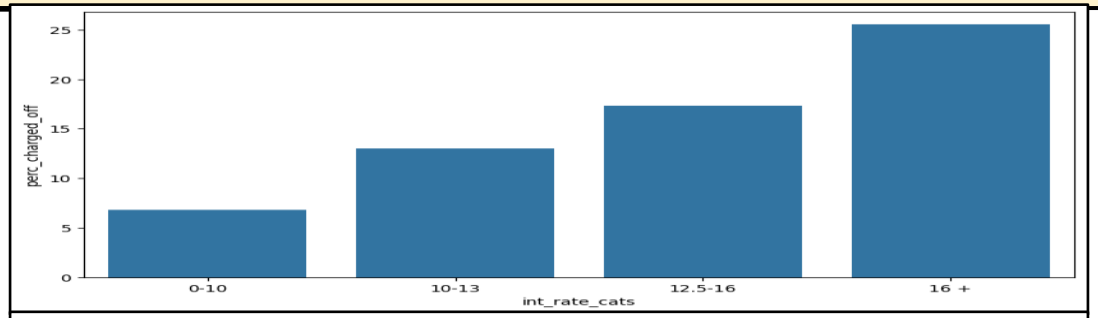
Conclusion: Customers with 2 public bankruptcies have higher charge off percentage.



Bivariate Analysis - Continued

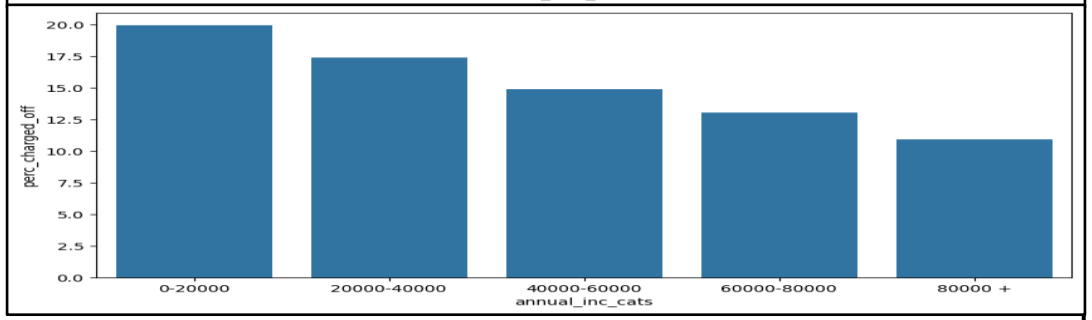
14. What is the highest interest rate against the % Charged Off?.

Conclusion: Its 16%+ against 25% Charged off customers.



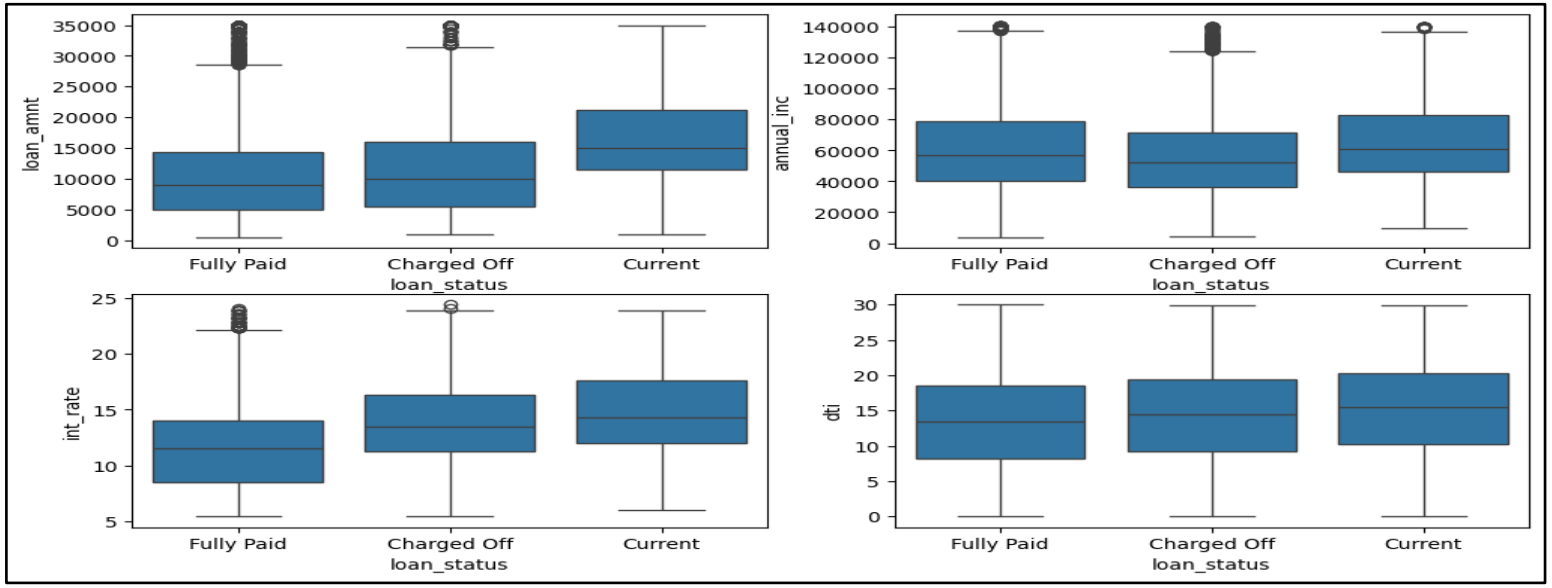
15. What is the highest Annual income Category against the Percentage Charged Off??

Conclusion: Its 0 to 20000 against 20% Charged off customers.



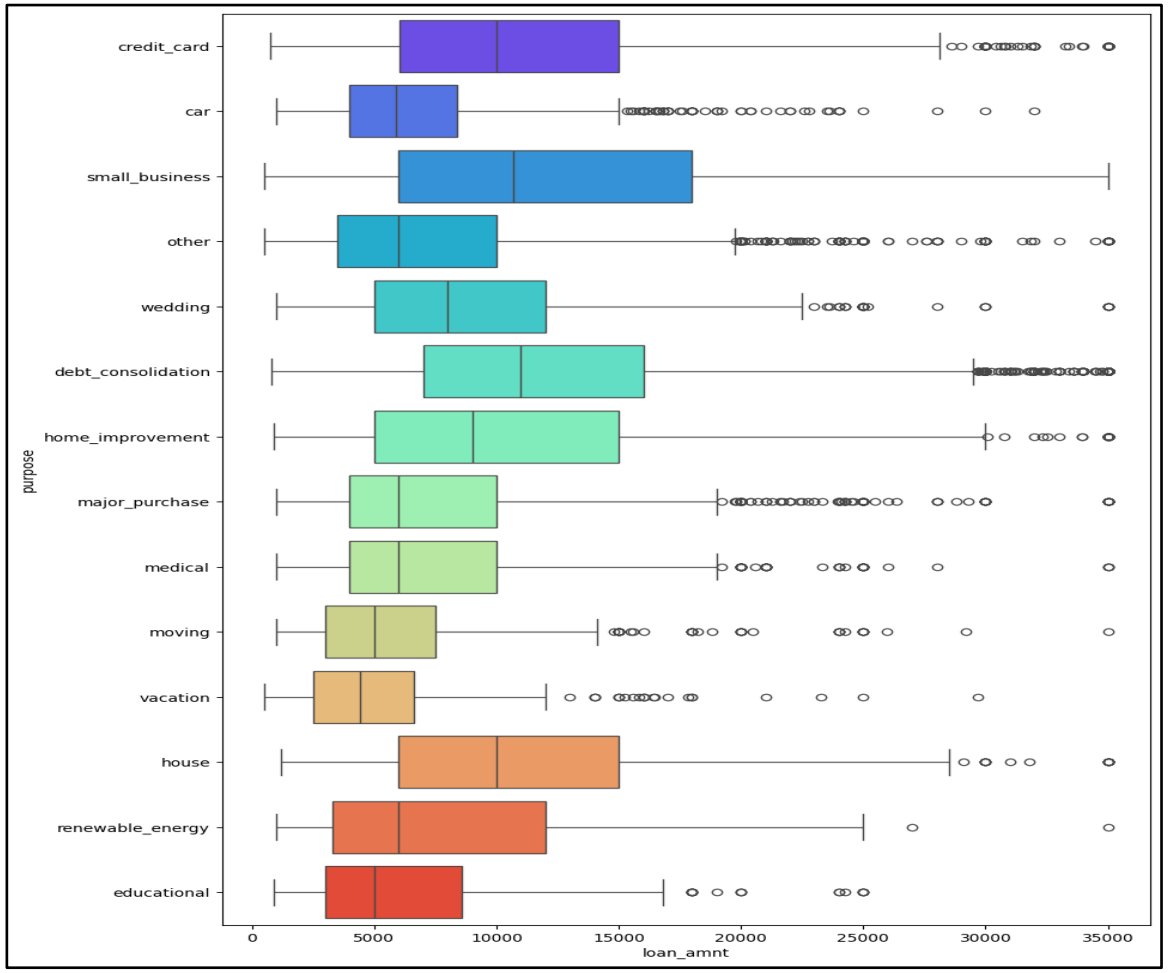
16. What is the average variations for the all the imp numeric columns?

Conclusion: No huge variations in the averages for the all the imp numeric column.



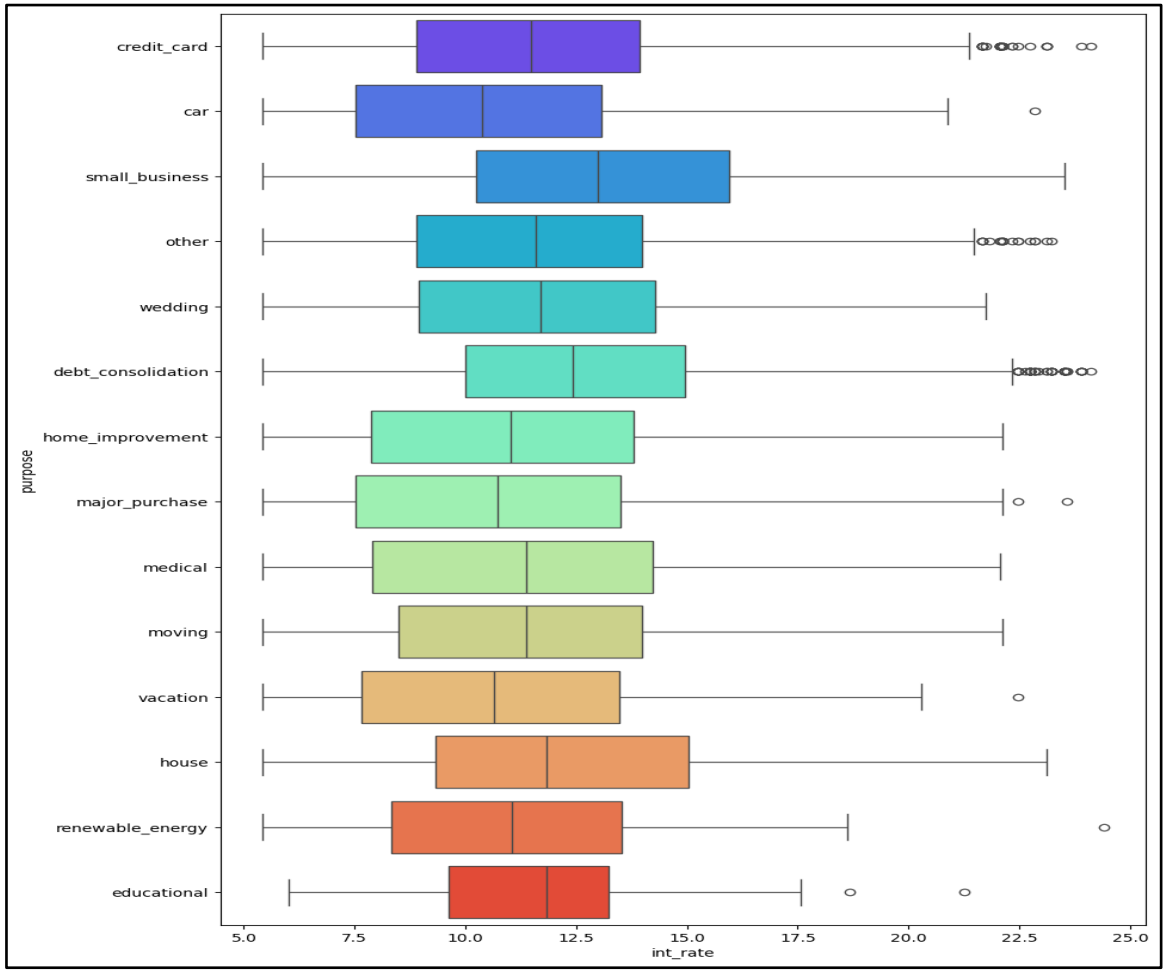
Bivariate Analysis - Continued

17. For what purpose higher loan amount is taken?



Conclusion: For small business purpose has higher average loan amount.

18. For what purpose highest interest rate is charged?

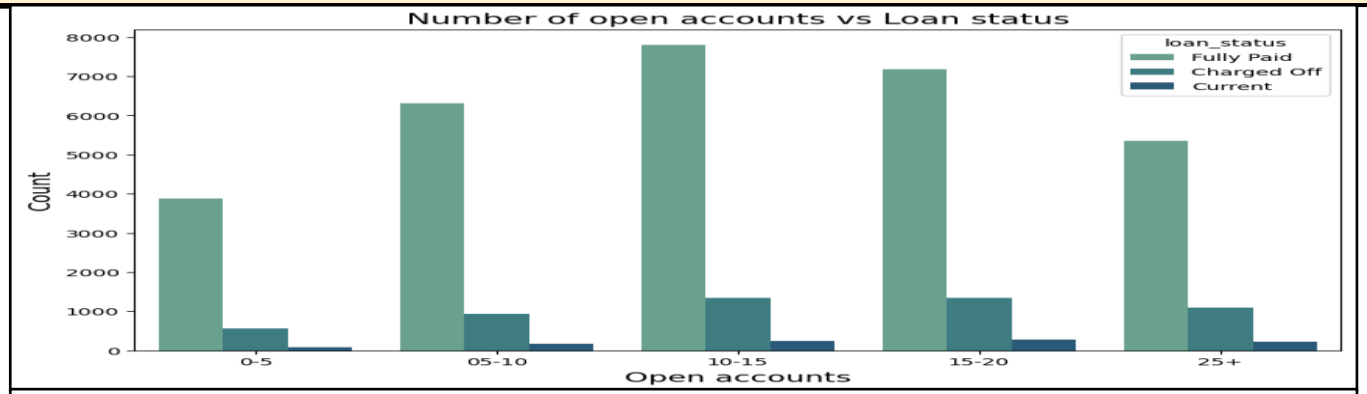


Conclusion: For small business purpose has higher average loan amount.

Bivariate Analysis - Continued

19. What is the highest range under which accounts are opened against Loan Status?

Conclusion: 10 to 15 highest range under which accounts are opened against Fully Paid Loan status.

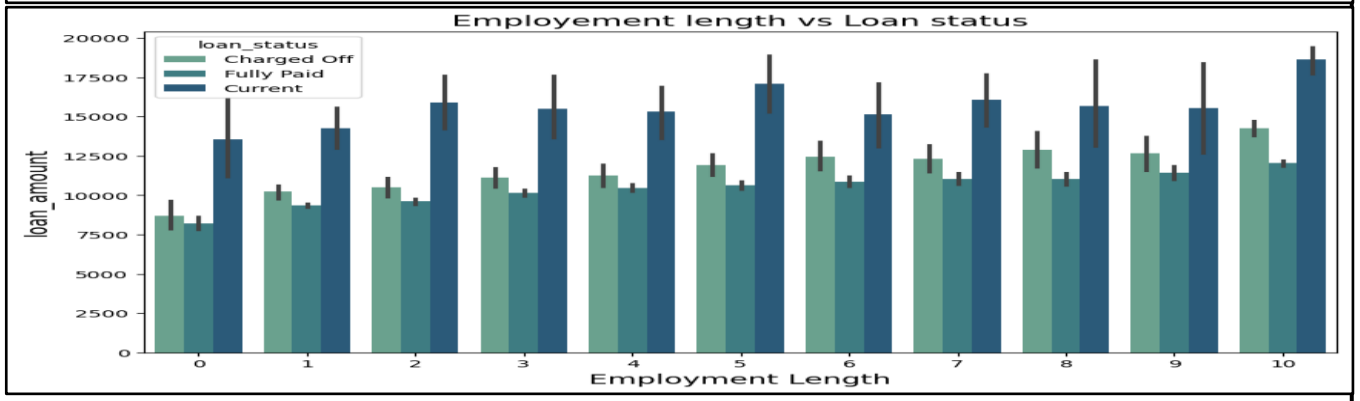


20. What is the highest range under Employment Length opened against Loan Status?

Conclusion: Highest Loans are taken in the 10th year followed by 1st year.

21. What is the highest range under Employment Length opened against Loan Status and Loan Amount?

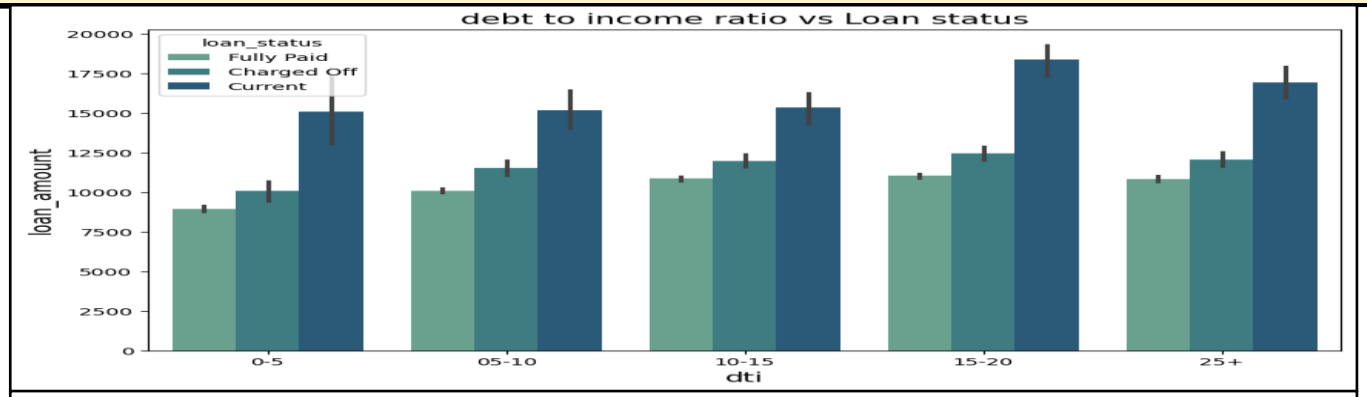
Conclusion: The highest range under Employment Length Opened against Loan Status and Loan Amount is 10th.



Bivariate Analysis - Continued

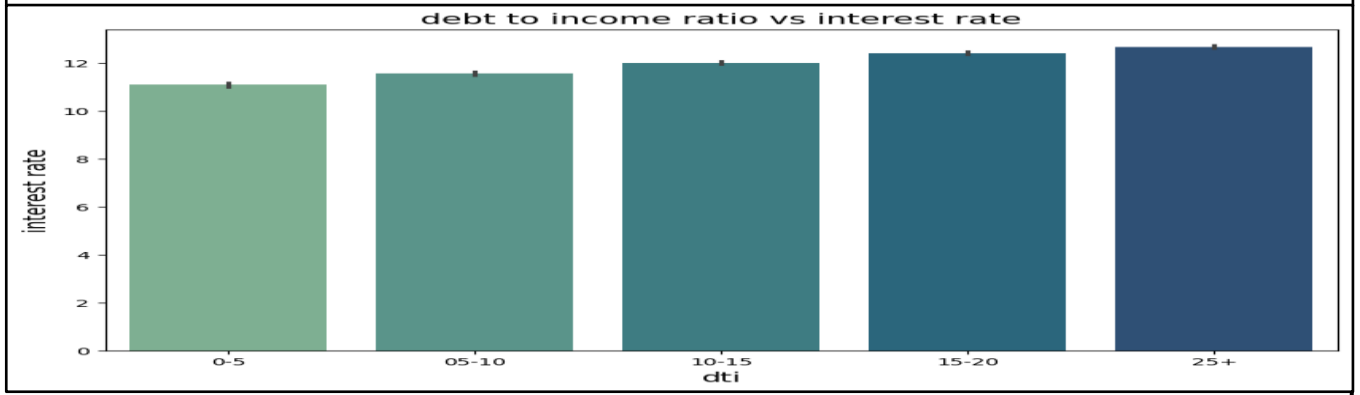
22. What is the highest dept to income ratio against Loan Status?.

Conclusion: The highest dept to income ratio against Loan Status is between 15 to 20.



23. What is the dept to income ratio against interest rate?

Conclusion: The highest dept to income ratio against interest rate is 25%+.



Conclusion Summary on Lending Club Case Study

1. Data was preprocessed to investigate.
2. Data was cleaned to transform.
3. Multiple Analysis on Univariate and Bivariate categories were performed.
4. Achieved the result for the Data Analyzed.
5. PPT was prepared on outcome.