LENDING CLUB CASE STUDY

Prathmesh Rahate

PROBLEM STATEMENT

The challenge is to employ Exploratory Data Analysis (EDA) to identify features indicative of 'risky' loan applicants, labeled as 'charged-off' or defaulters. By pinpointing these risk factors, the goal is to minimize credit loss, allowing lenders to make informed decisions and reduce financial liabilities.

APPROACH

Data Pre-processing

- Data Understanding
- Data Cleaning
- Final Preparation

Univariate Analysis

- Outlier Treatments
- Feature Analysis

Bivariate Analysis

- Feature Comparison

Correlation Matrix

- Correlation Analysis

Conclusion

- Final Words

DATA PRE-PROCESSING

Data Understanding

- Checked data format, number of records, and field identities for a comprehensive understanding.
- Executed queries and visualized data to gain insights into its structure and distribution.
- Evaluated data quality, identified unnecessary or dirty data, and took appropriate actions.

DATA PRE-PROCESSING

Data Cleaning

- Deleted unnecessary columns.
- Removed outliers (very high and low values) affecting the analysis.
- Treated missing values using appropriate approaches.
- Removed duplicate data.
- Removed unnecessary strings from numerical data.

DATA PRE-PROCESSING

Final Preparation

- Created new columns based on existing data to enhance analysis.
- Defined data types as either numeric or categorical for accurate representation.

UNIVARIATE ANALYSIS

Outlier treatment involves identifying and addressing extreme values in a dataset. This process includes detecting outliers using statistical methods or visualization, deciding on an appropriate treatment strategy (such as removal, transformation, or imputation), and executing the chosen method to enhance the dataset's reliability for analysis. Outlier Treatment done on :

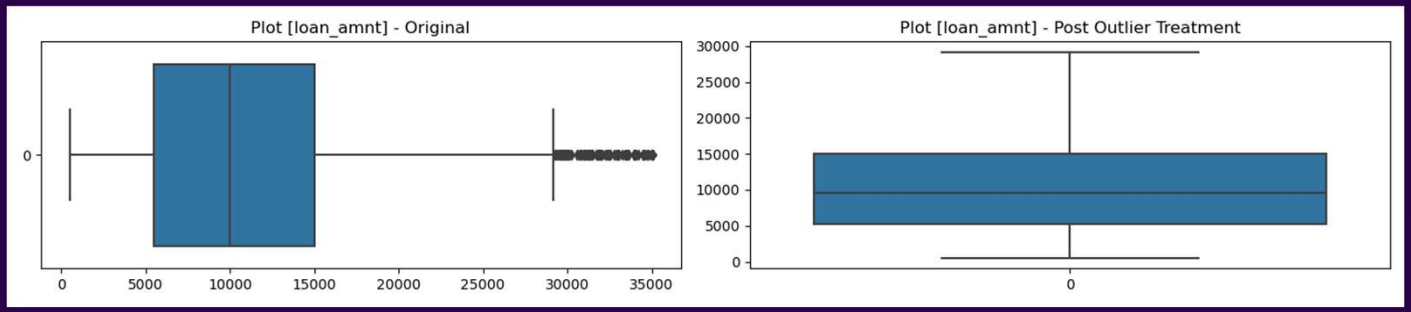
1. loan_amnt
2. funded_amnt
3. funded_amnt_inv
4. int_rate
5. installment
6. annual_inc
7. dti

UNIVARIATE ANALYSIS

Outlier treatment for loan_amnt:

Row Dropped: 1076

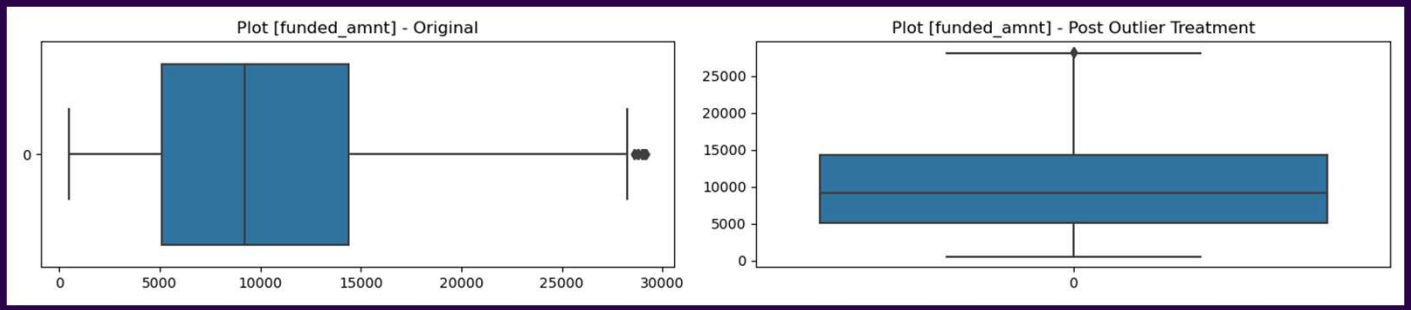
Percentage Rows Dropped: 2.93%



Outlier treatment for funded_amnt:

Row Dropped: 30

Percentage Rows Dropped: 0.08%

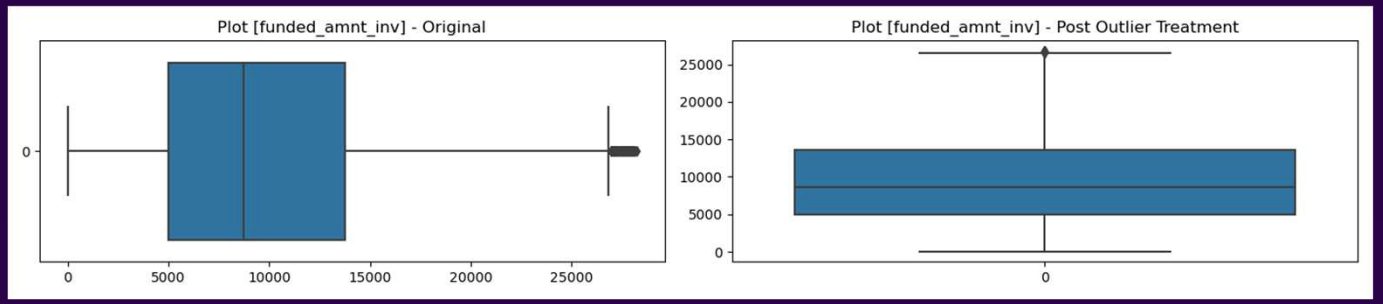


UNIVARIATE ANALYSIS

Outlier treatment for funded_amnt_inv:

Row Dropped: 152

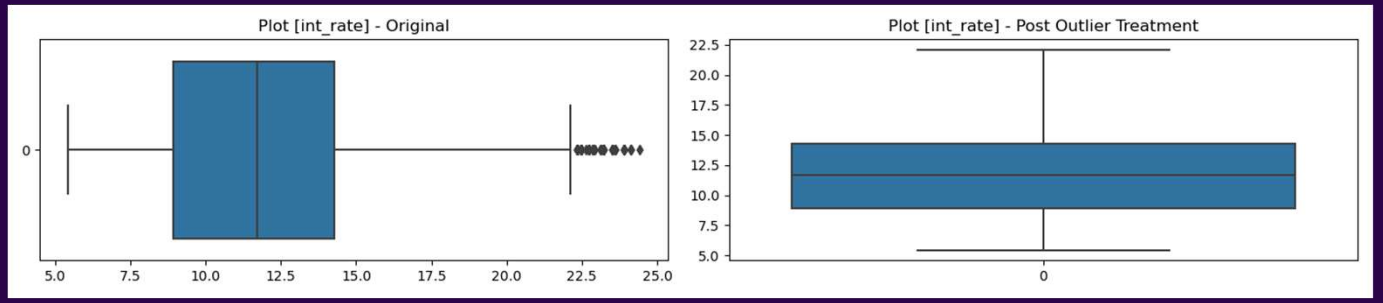
Percentage Rows Dropped: 0.43%



Outlier treatment for int_rate :

Row Dropped: 63

Percentage Rows Dropped: 0.18%

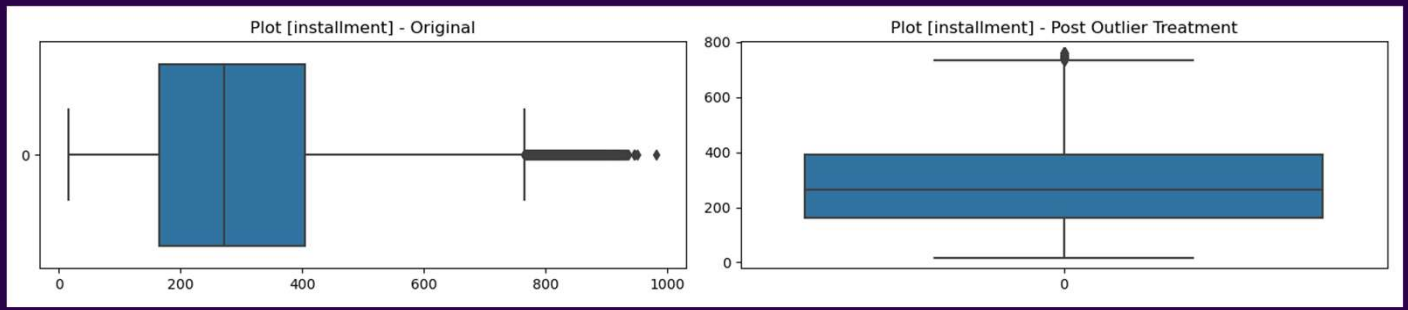


UNIVARIATE ANALYSIS

Outlier treatment for installment:

Row Dropped: 978

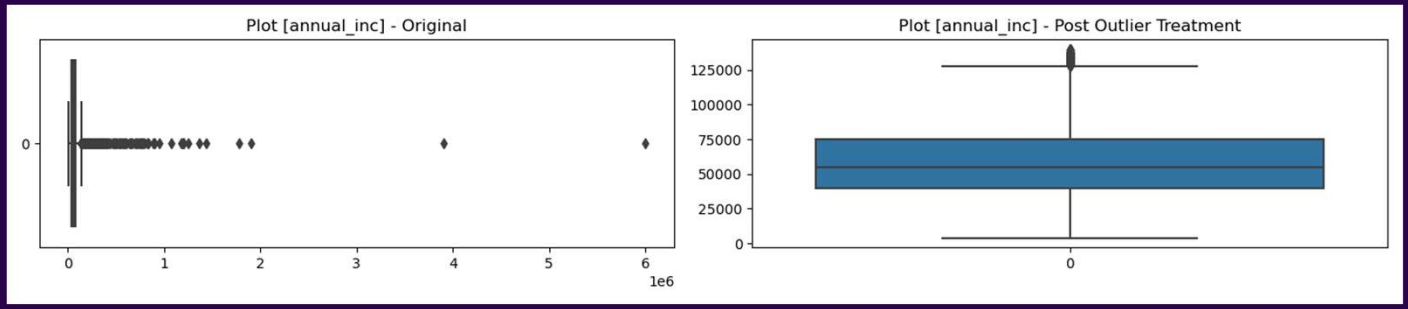
Percentage Rows Dropped: 2.76%

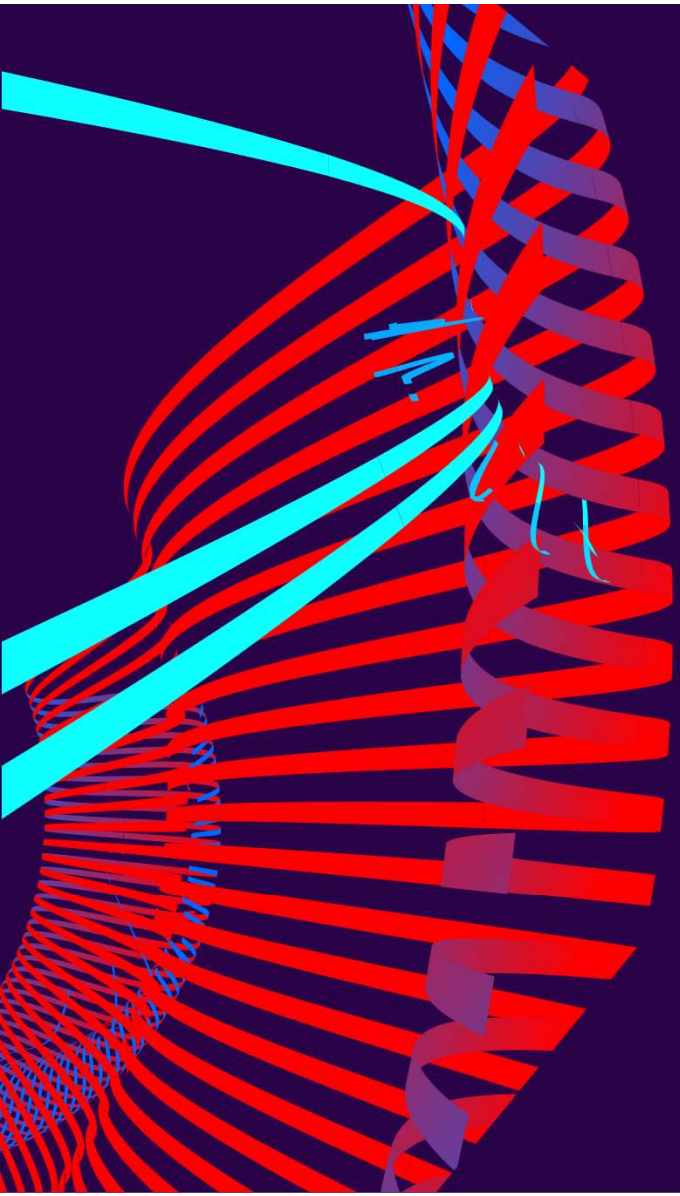


Outlier treatment for annual_inc :

Row Dropped: 1483

Percentage Rows Dropped: 4.31%

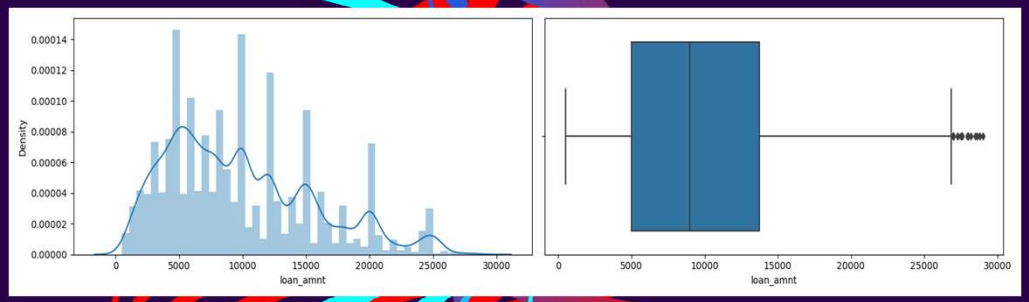


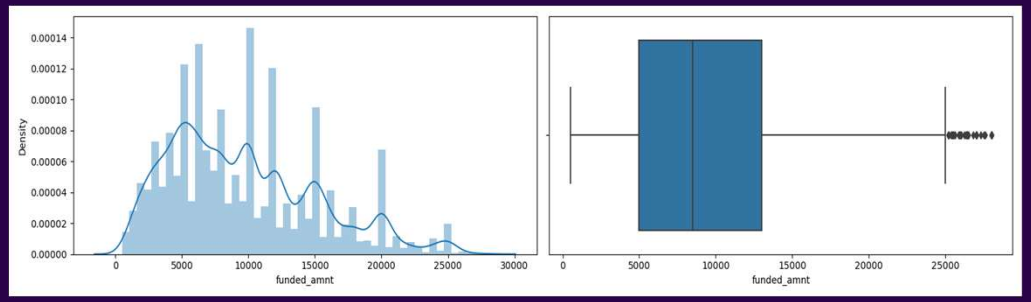


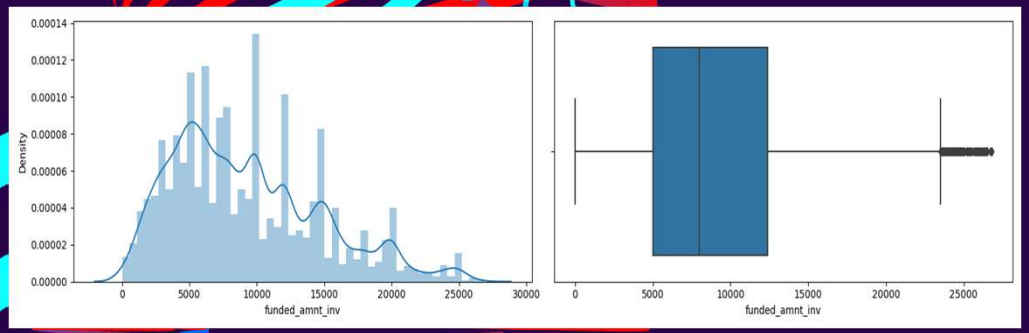
UNIVARIATE ANALYSIS

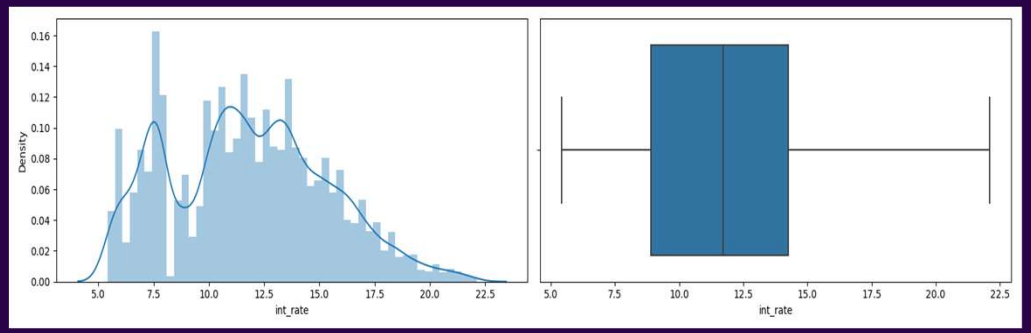
Quantitative Variable Analysis

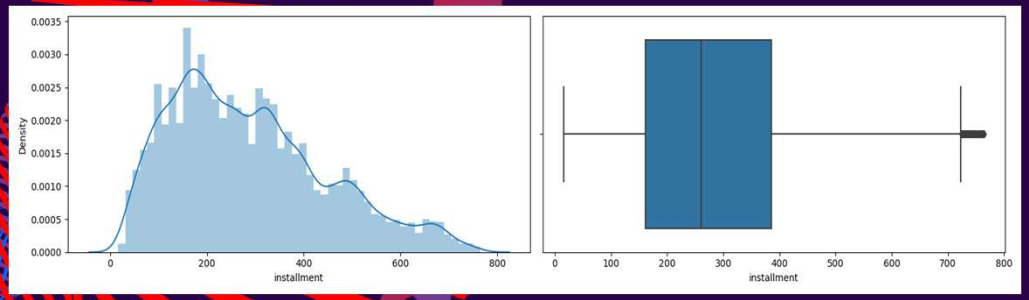
UNIVARIATE ANALYSIS

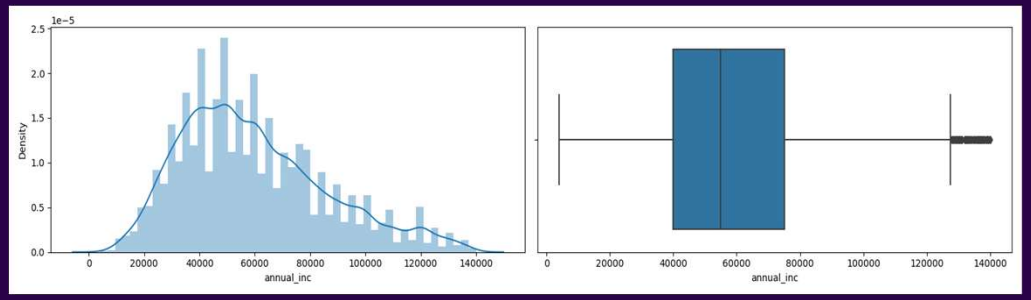


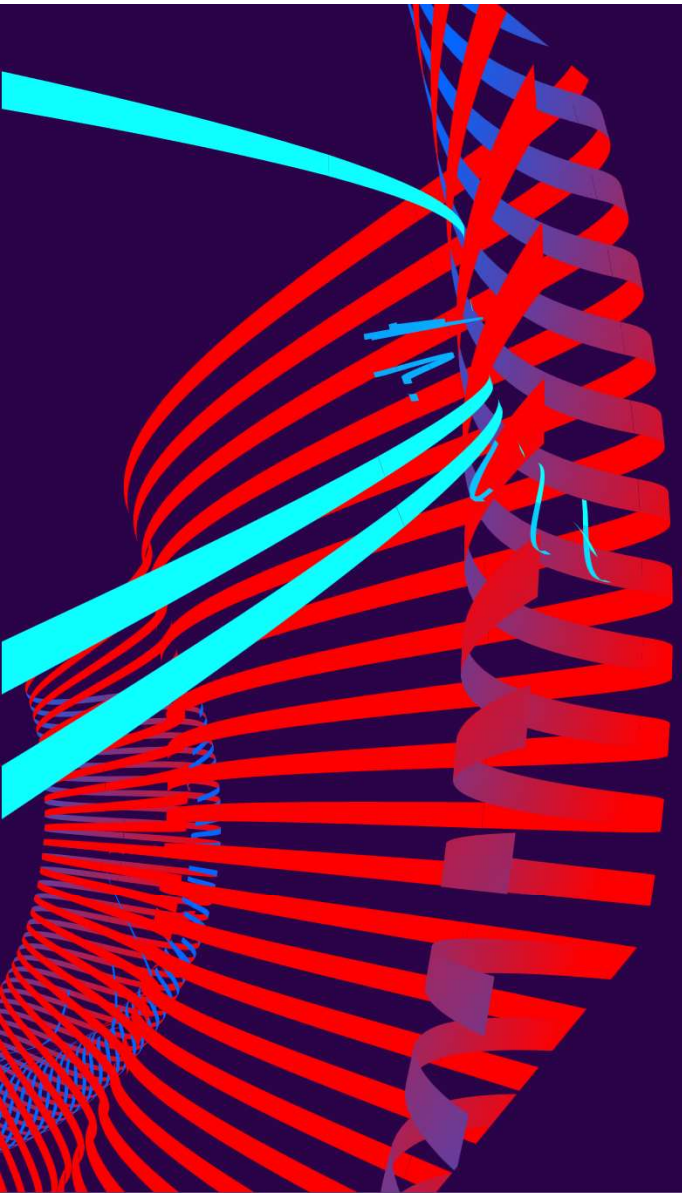












UNIVARIATE ANALYSIS

Quantitative Variable Analysis

- Most of the loan amount is in the range of 5000 to 14000
- Majority of the funded amount is in the range of 5000 to 13000
- Majority of the funded amnt_inv is in the range of 5000 to 12000
- Majority of the interest rate is in the range of 5% to 16% going at the max to 22%
- Majority of the installment is in the range of 20 to 400 going at the max to 700
- Majority of the annual income is in the range of 4000 to 40000 going at the max to 12000. This column required major outlier treatment.
- Majority of the debt to income is in the range of 0 to 20 going at the max to 30

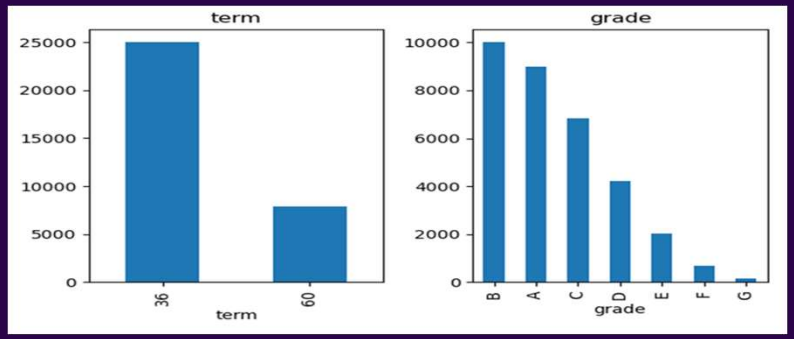
UNIVARIATE ANALYSIS

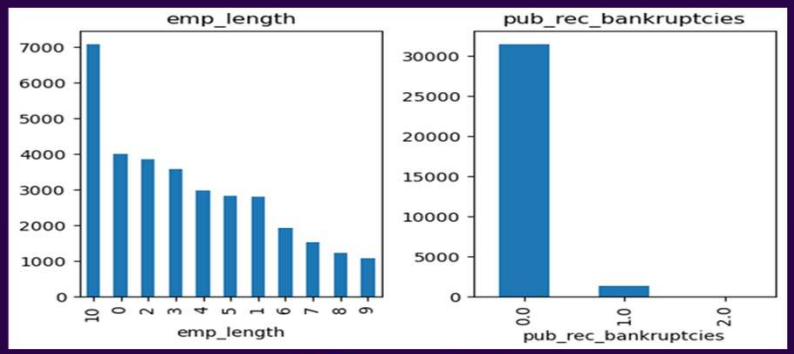
Ordered Categorical Variable Analysis

UNIVARIATE ANALYSIS

Ordered Categorical Variable Analysis

- Majority of the loan applications counts are in the term of 36 months.
- Majority of loan application counts fall under the catogory of Grade B.
- Majority of the employment length of the customers are 10+ years and then in the range of 0-2 years.
- Majority of the loan applicants are in the category of not having an public record of bankruptcies.



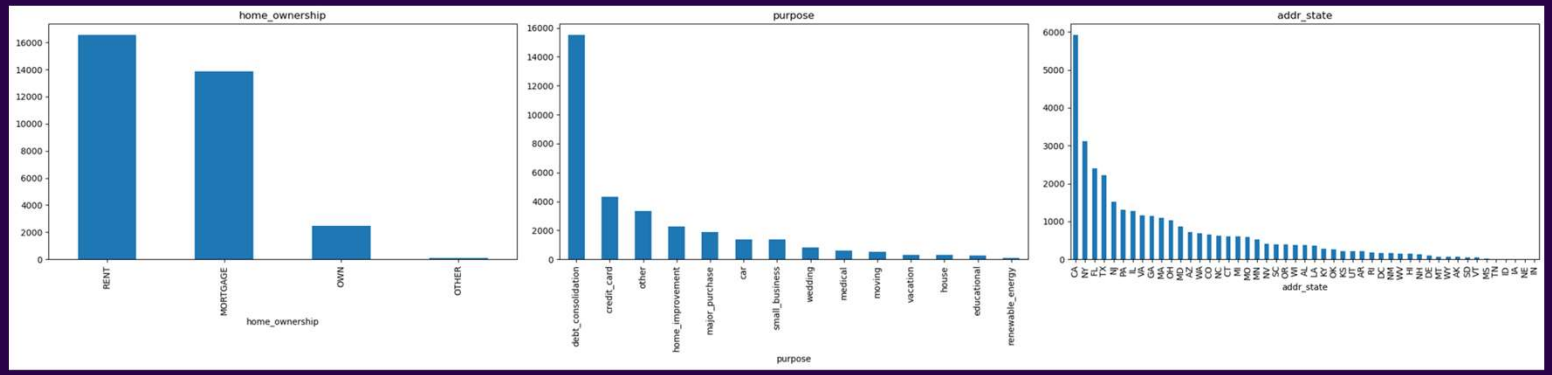


UNIVARIATE ANALYSIS

Unordered Categorical Variable Analysis

UNIVARIATE ANALYSIS

Unordered Categorical Variable Analysis

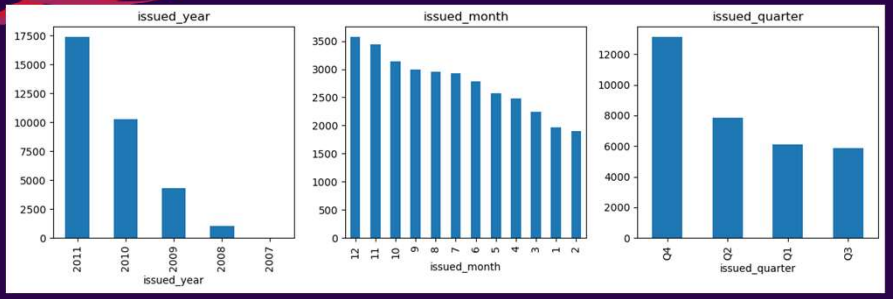


- Majority of the home owner status are in status of RENT and MORTGAGE.
- Majority of loan application are in the category of debt_consolidation.
- CA state has the maximum amount of loan applications.

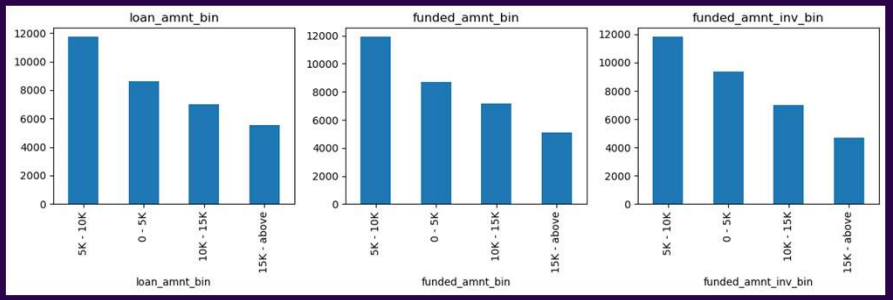
UNIVARIATE ANALYSIS

Derived Variable Analysis

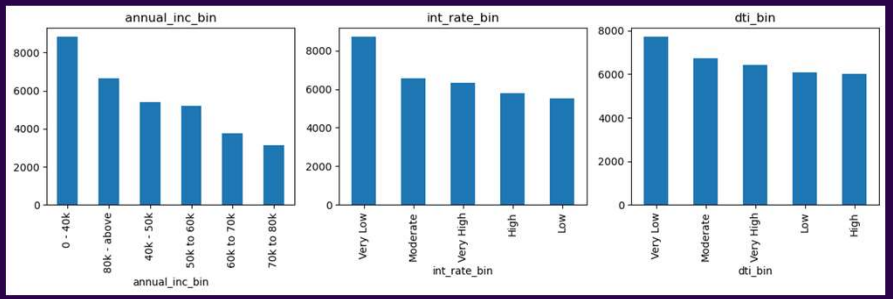
Loan application counts are increasing year over year. Maybe the risk exposure is increasing over the year (un proven hypothesis).The lowest loans application count are in the month of Jan, Feb, March and highest counts are in Oct, Nov, Dec. Highest loan application volume in Quarter 4 of a year.



Highest loan amount applications fall in the range of 5k to 10k.
Highest funded amount applications fall in the range of 5k to 10k
Highest loan amount applications fall in the range of 5k to 10k



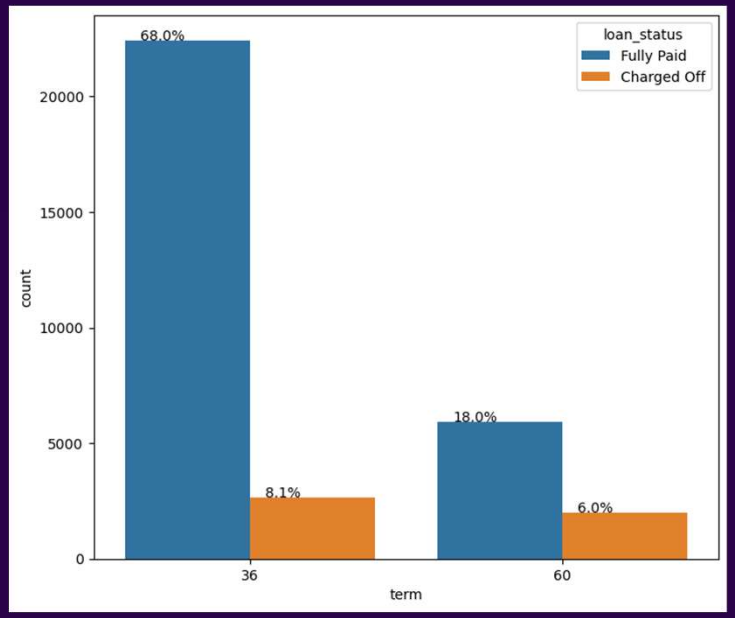
Majority of the loan applocants are in the range of 0 - 40K anual income
Majority of the loan applocations are in the category of Very Low interest rates
Majority of the loan applications are in Moderate debt to income ratio



BIVARIATE ANALYSIS

Analysis of term vs loan_status

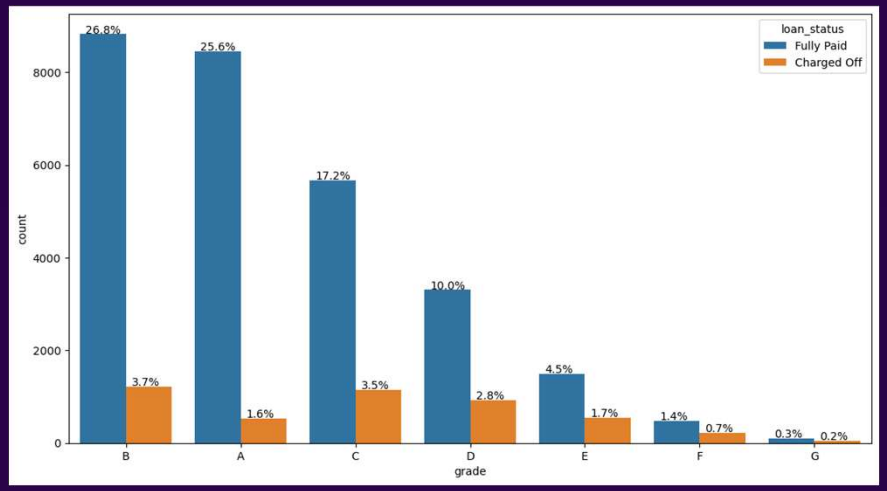
The analysis reveals that loan volume is higher in the term=36 category. Although the overall charge-off percentage is slightly higher in term=36 (8%) compared to term=60 (6%), the ratio of charge-offs within the term=60 category is significantly higher (25% vs. 10%). This suggests that term=60 applications may require more scrutiny, as they are more likely to result in charge-offs.



BIVARIATE ANALYSIS

Analysis of grade vs loan status

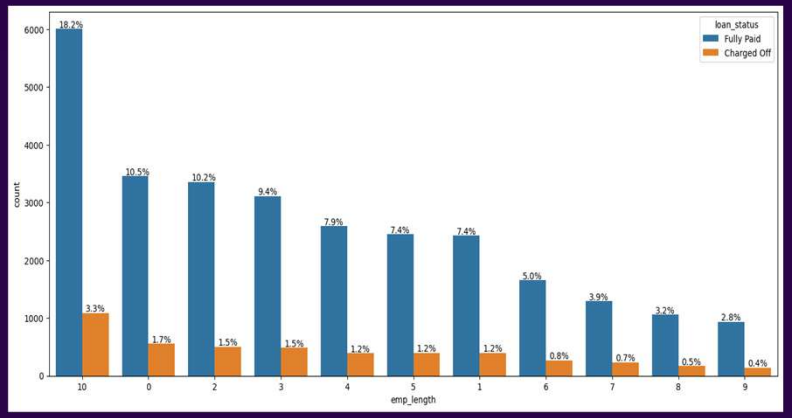
The majority of loan volume is in grade=B. The highest percentage of overall charge-offs is in grades B (3.7%) and C (3.5%). Analyzing the charge-off ratio within categories reveals the highest percentage in grade=G, with the highest cluster in grades G and F (>30%). Although Grade G has low volume , it doesn't significantly contribute to overall risk. Inferences suggest that the highest risk of charge-offs is in grades B and C, while grades F and G have very high chances of charge-offs with low volumes. Grade A has a low probability of charge-offs, and the risk increases from A to G.



BIVARIATE ANALYSIS

Analysis of emp_length vs loan_status

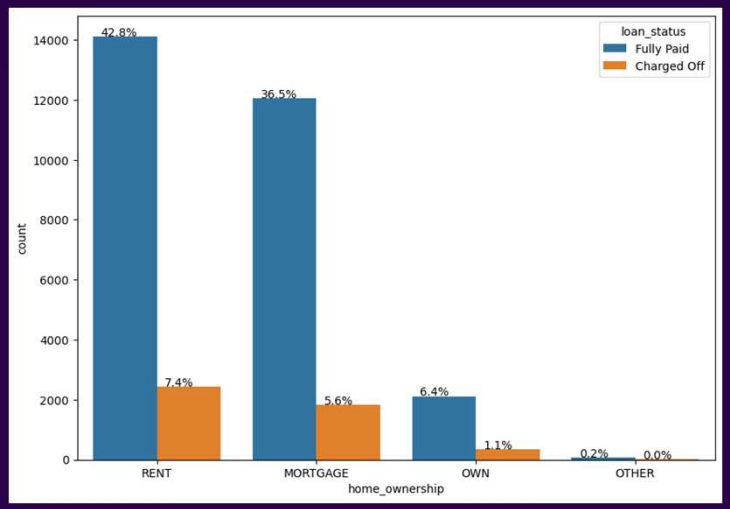
The analysis indicates that the highest charge-offs occur in the employee length category of 10 years and above. However, the charge-off ratio within these categories is similar and inconclusive. Inferences suggest that there is a high probability of charge-offs for individuals with an employee length of 10 years and above. Additionally, there is a higher probability of charge-offs for individuals with an income range of less than 1 year. However, the ratio within these income ranges is inconclusive.



BIVARIATE ANALYSIS

Analysis of home_ownership vs loan_status

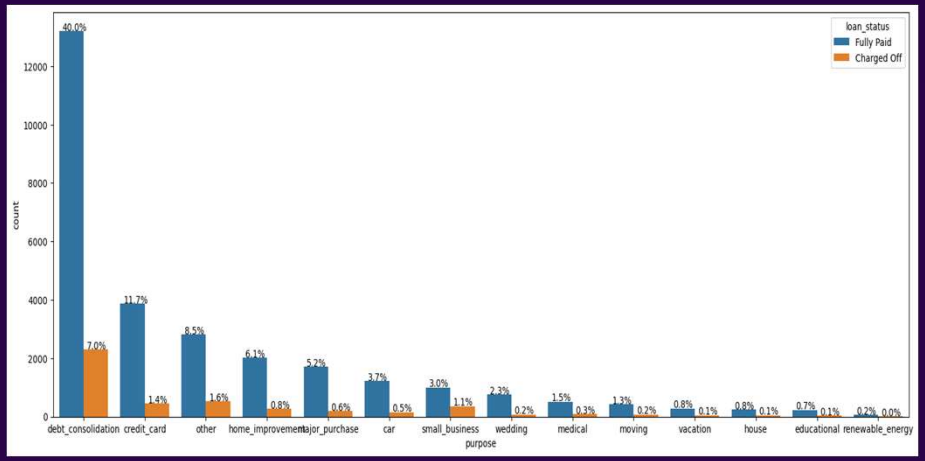
The analysis indicates that the overall highest number of charge-offs occurs in the home ownership categories of RENT and MORTGAGE. Within each home ownership category, the ratio of charge-offs for the "Other" category is higher. Inferences suggest that the home ownership statuses of MORTGAGE and RENT are at the highest risk of charge-offs. The MORTGAGE status also has the highest range of loan amounts, further increasing the risk.



BIVARIATE ANALYSIS

Analysis of purpose vs loan status

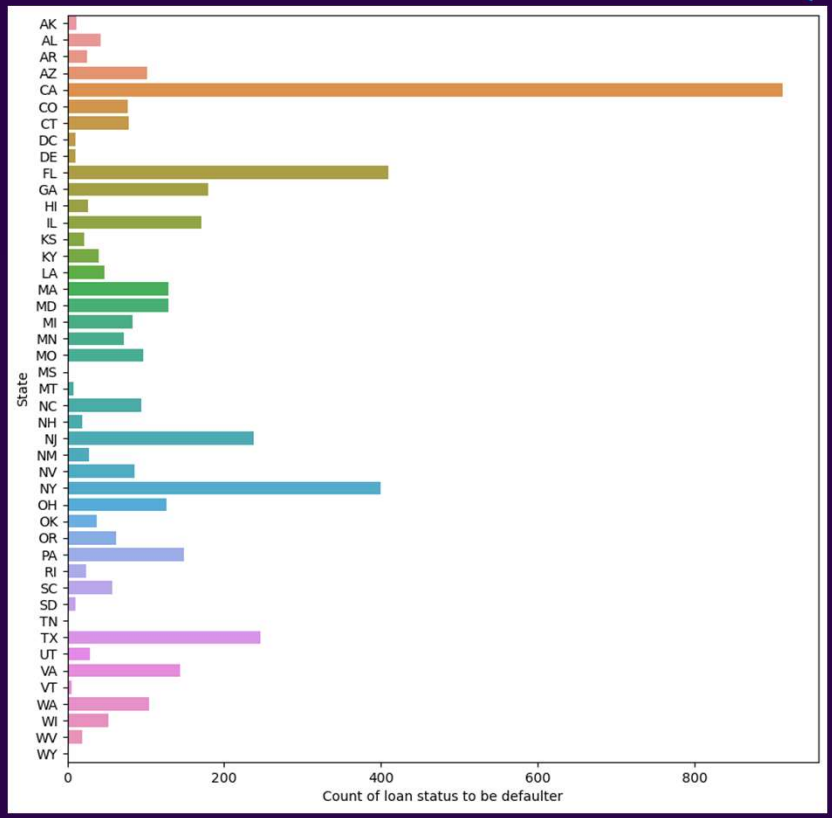
The analysis reveals that the highest risk of charge-offs is associated with the purpose of debt consolidation. Although small business has the highest probability of charge-offs within a category, the volume is extremely low. The highest loan amount ranges are observed in small business, debt consolidation, and house categories. Inferences suggest that debt consolidation poses the highest risk of charge-offs, small business applicants have a high probability of being charged off, and renewable energy has the lowest risk of charge-offs in terms of volume.



BIVARIATE ANALYSIS

Analysis of addr_state vs loan state

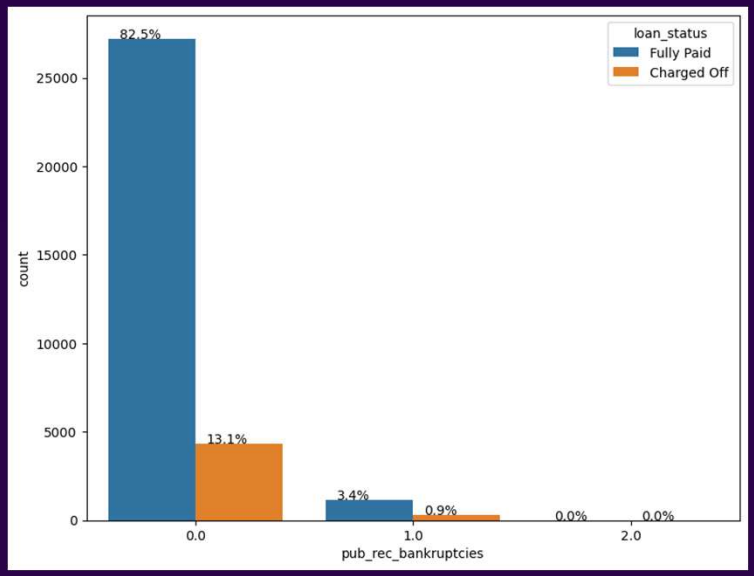
The analysis indicates that California (CA) has the highest volume of loans, and purely based on volumes, the highest charge-offs are also from CA. However, when considering charge-offs within each state, Nebraska (NE) and Nevada (NV) have the highest charge-offs. Since NE has very low volume, it may not be considered significant. The inferences suggest that loan applications from Nevada (NV) have a high risk of charge-offs. Nebraska (NE) has a very high probability of charge-offs, but the volume is too low to be considered. Nevada (NV), California (CA), and Florida (FL) have a high percentage of charge-offs.



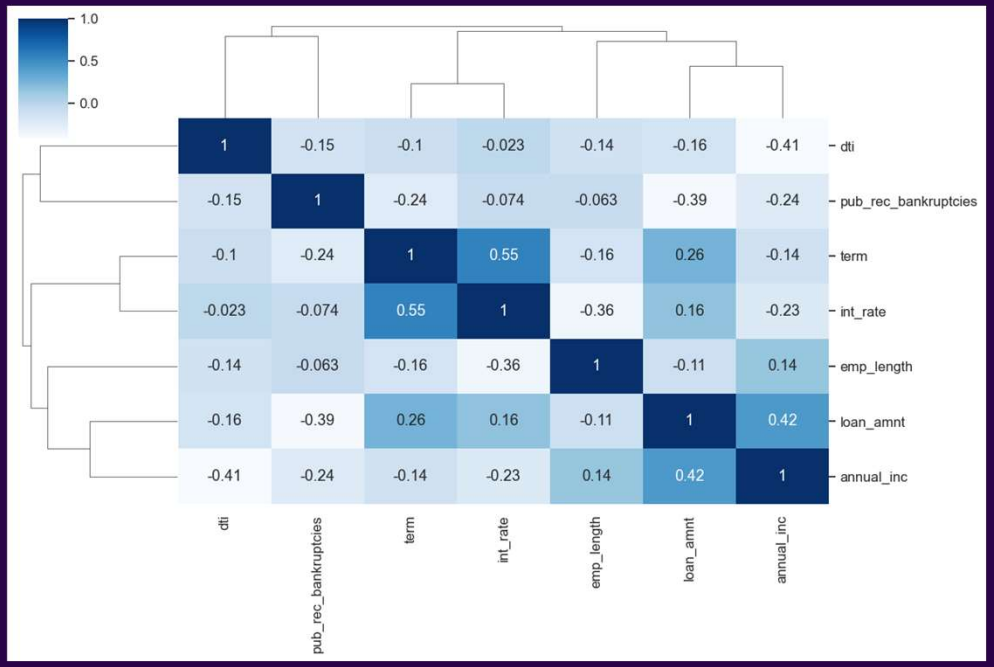
BIVARIATE ANALYSIS

Analysis of pub_rec_bankruptcies vs loan_status

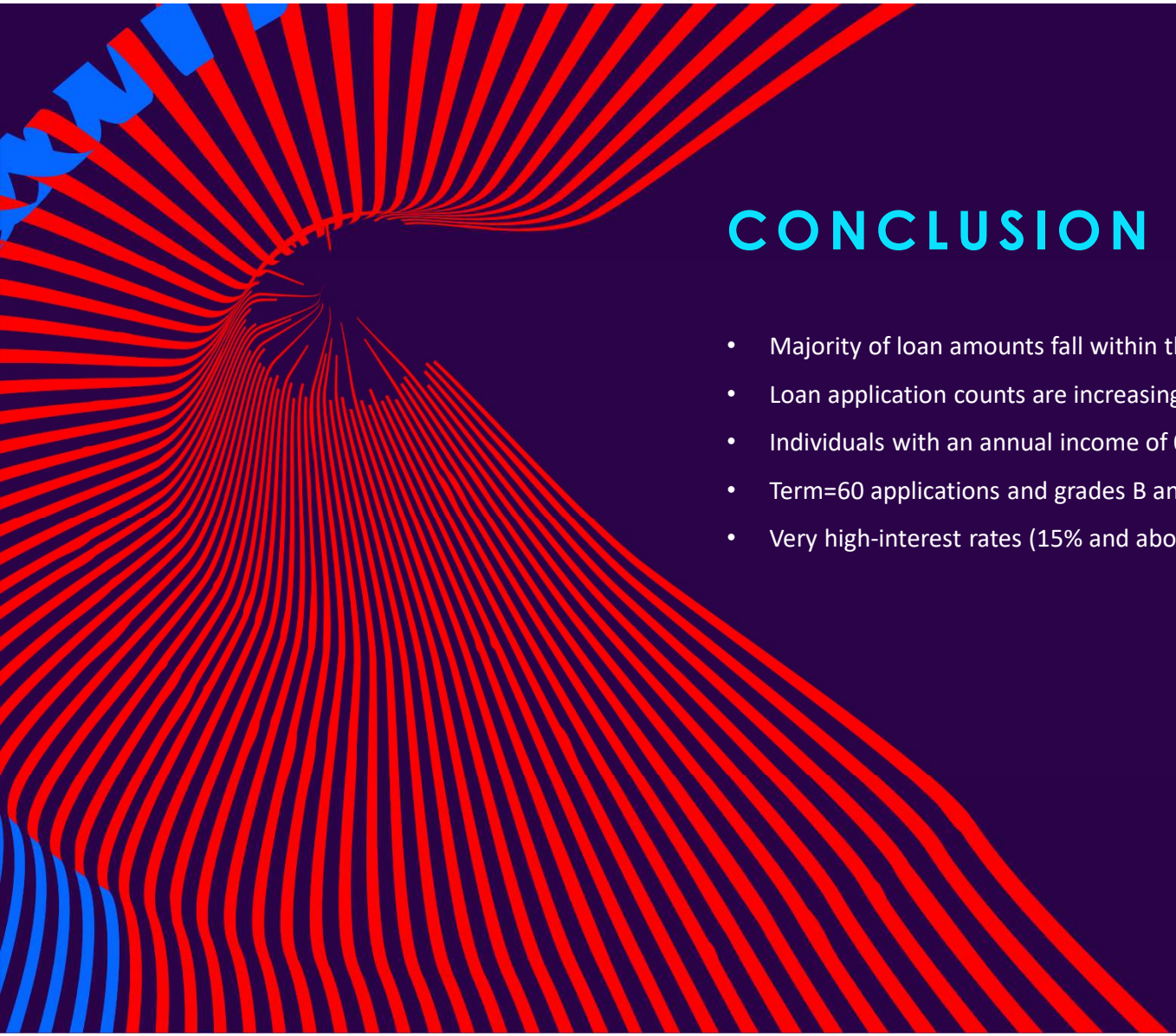
The analysis, based on volumes, shows that the highest number of charge-offs is in the category of 0 (no bankruptcy record). However, when considering ratios within each category, customers with a bankruptcy record have a higher charge-off ratio. Inferences suggest that customers with a bankruptcy record are at a high risk of charge-offs, and those with a pub_rec_bankruptcies count of 2 have an even higher charge-off ratio.



CORRELATION MATRIX



- Negative Correlation :
  - loan_amnt has a negative correlation with pub_rec_bankruptcies.
  - annual_inc has a negative correlation with dti (debt-to-income ratio).
- Strong Correlation:
  - term has a strong correlation with loan_amnt.
  - term has a strong correlation with int_rate.
  - annual_inc has a strong correlation with loan_amnt.
- Weak Correlation:
  - pub_rec_bankruptcies has a weak correlation with most fields.



CONCLUSION

- Majority of loan amounts fall within the range of 5000 to 14000.
- Loan application counts are increasing yearly, with the highest volume in Quarter 4.
- Individuals with an annual income of 0-40K are at the highest risk of charge-offs.
- Term=60 applications and grades B and C pose higher risks of charge-offs.
- Very high-interest rates (15% and above) are associated with a higher risk of charge-offs.

THANK YOU

Prathmesh Rahate

prathmeshvr@gmail.com