Lending Club

CASE STUDY

Understanding Risk Analytics in Banking and Financial Services using Exploratory Data Analysis (EDA)

- Venkat Lata and Sunil Bhairi

Agenda Overview

01 Background of the Study

02 Problem Statement

03 Methodology

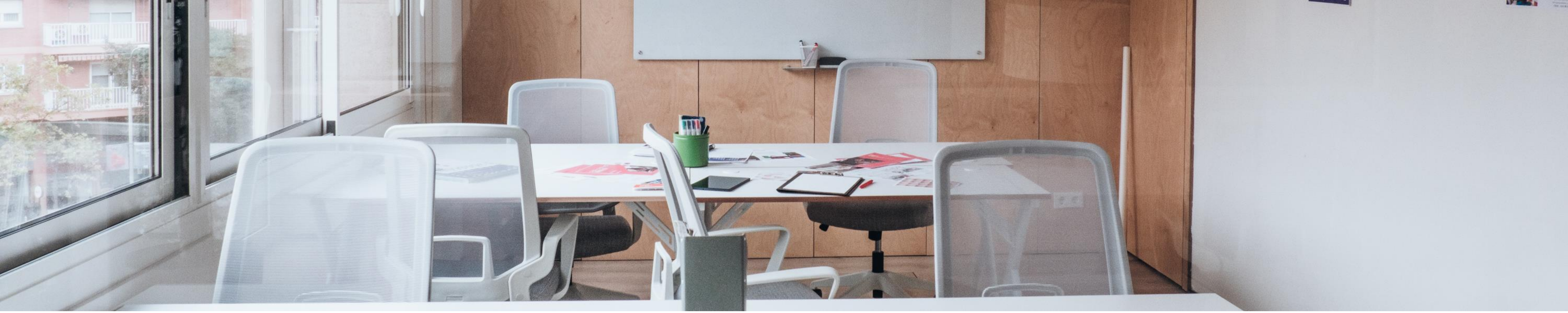
04 Univariate Analysis

05 Segmented Univariate Analysis

06 Bivariate Analysis

07 Multivariate Analysis

08 Recommendations



Background of the Study

Lending Club, a **consumer finance company,** specializes in offering various types of loans to urban customers, facilitating access to lower interest rate loans such as personal loans, business loans, and financing for medical procedures through a fast online interface.

Like most other lending companies, not approving the loan to 'potential' applicants and lending loans to 'risky' applicants are the largest sources of financial loss (called credit loss).

In this case study, we used **Exploratory Data Analysis (EDA)** to identify patterns and understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Problem Statement

We analyzed the loan data for all loans issued through the time period 2007 to 2011 by the Lending Club Company.
Our goal was to understand the **driving factors** behind loan default.
The company can utilise this knowledge for its portfolio and risk assessment.

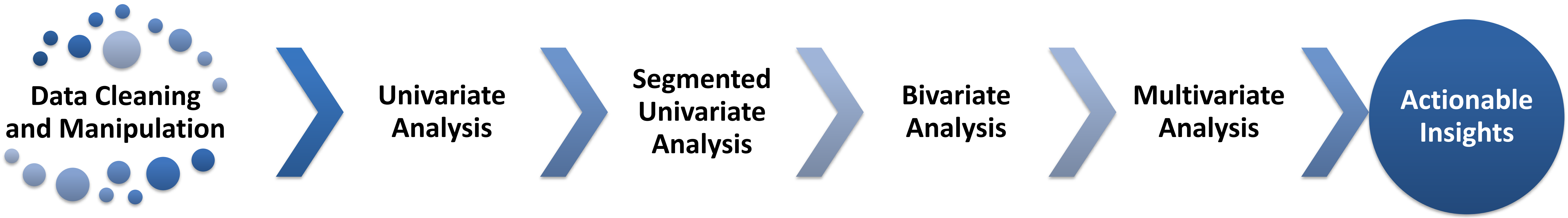
Technologies used
Data Cleaning and Manipulation, and Exploratory Data Analysis using **Python**, **NumPy**, and **Pandas.**
Data Visualization using **Matplotlib, Seaborn** and **Plotly.**

Target Variable : loan_status
Loan status has three unique categories: Fully Paid, Charged Off, and Current. The Current category is not relevant as the tenure of the loan is not yet complete.

Research Question
Which variables have the strongest correlation with the risk of default? How can we segment the loan applicants based on their risk profiles ?

Methodology

Overview



Data Cleaning and Manipulation

- ✓ Handling Missing Values and Dropping Duplicates
- ✓ Data Type Casting and Data Transformation
- ✓ Performing Sanity Checks and Handling Outliers
- ✓ Derived Metrics

Univariate Analysis

- ✓ Plot 1
- ✓ Plot 2
- ✓ Plot 3
- ✓ Plot 4

Segmented Univariate Analysis

- ✓ Plot 5
- ✓ Plot 6
- ✓ Plot 7
- ✓ Plot 8

Bivariate Analysis

- ✓ Plot 9
- ✓ Plot 10
- ✓ Plot 11
- ✓ Plot 12

Multivariate Analysis

- ✓ Plot 13
- ✓ Plot 14

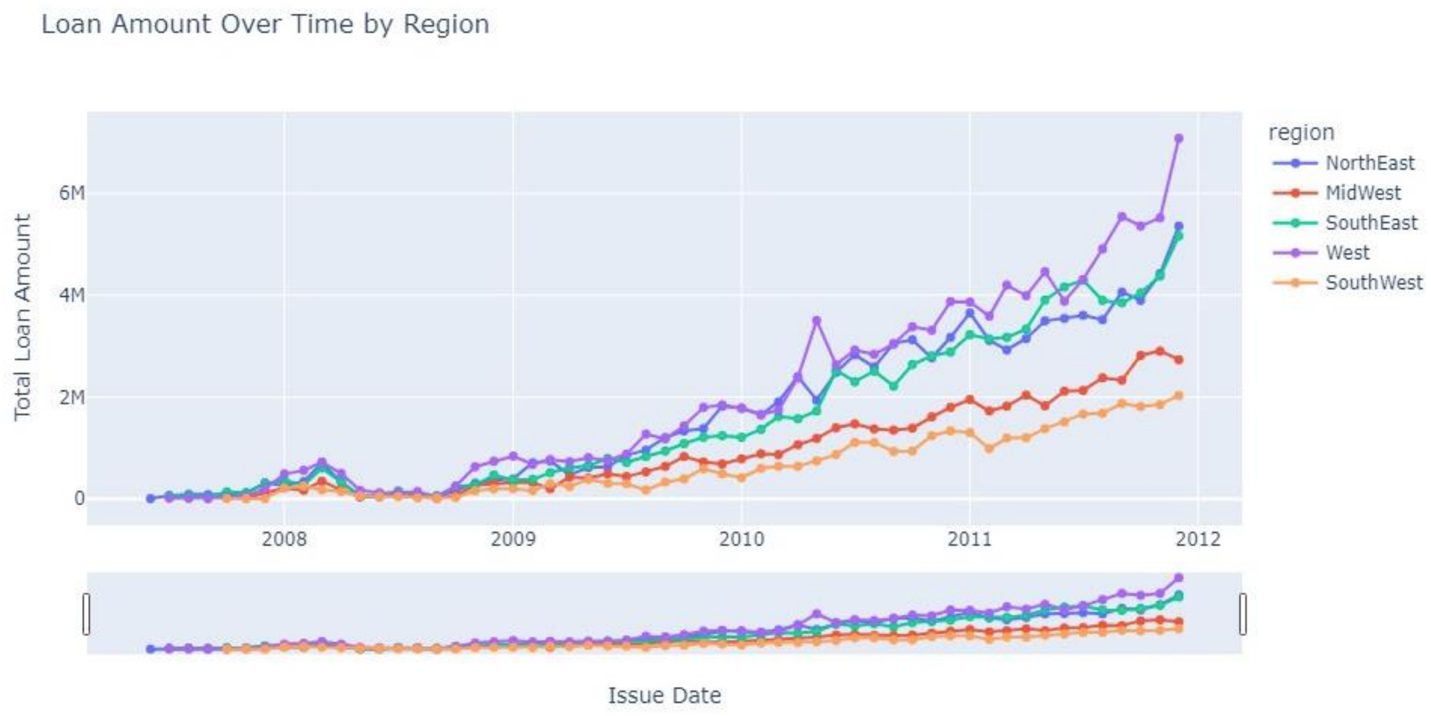
Actionable Insights

- ✓ Data Inference
- ✓ Major Takeaways
- ✓ Suggested Actions
- ✓ Way Forward

Business Overview

Business Targets and Achievements

- **Loan Amounts:** States with the highest mean loan amounts include Arkansas, Iowa, and District of Columbia. These states have significantly larger average loans compared to others.
- **Regional Trends:** States with high mean loan amounts often show higher average interest rates, annual incomes, and default rates (Only Exception is the Midwest Region). This suggests that higher loan amounts might be associated with greater financial risks and opportunities in these regions.
- **Geographic Variations:** Some states, like Rhode Island and Louisiana, have lower mean loan amounts despite having high average incomes, indicating a potential divergence between income levels and loan sizes in those areas.
- **Default Rates:** States with higher mean loan amounts, e.g. Arkansas, also tend to have higher default rates, highlighting a possible risk factor related to larger loan sizes.
- **Interest Rates:** There is a noticeable correlation between higher mean loan amounts and elevated average interest rates in certain states (Exception: Iowa), pointing to regional differences in lending practices.
- **Loan Amount Trends:** The total loan amount has varied over time across different regions. Notable increases or decreases in loan amounts can be observed in specific periods, indicating regional economic fluctuations or changes in lending practices.
- **Regional Variations:** Different regions display distinct trends in total loan amounts, reflecting regional differences in loan demand or economic conditions. Loan Amounts seem to fork into two branches between 2010 and 2011, which West, Southeast and Northeast receiving the most loan amounts as compared to Midwest and Southwest.
- **Impact of Economic Changes:** The fluctuations in total loan amounts and loan characteristics by region between 2007 and 2011 could be possibly due to a recovery in the US economy after the 2008 Financial Crisis.



Univariate >> Numerical Variables

1. ***Loan Amount Distribution:***
- Most loans were issued within the $5,000 to $14,000 range, with $10,000 being the most frequently issued amount.
- However, the most popular issued amount among investors was $5,000.

2. ***Loan Amounts and Default Status:***
- Borrowers who defaulted received higher loan amounts on average, with a median loan of $10,000 compared to $9,000 for those who fully paid their loans.

3. ***Loan Application and Funding:***
- The distributions of loan applications, amounts issued, and amounts funded by investors are closely aligned. This suggests that borrowers who meet the qualifications are likely to receive the full amount they apply for.

4. ***Interest Rate Distribution and Default Status:***
- Most interest rates were between 9% and 14%. The most common rate for defaults was 11.5%, whereas the most common rate for fully paid loans was 11%.
- There is a noticeable gap in the frequency data between 8% and 9%, indicating a potential under-utilization of this low interest rate range, because fewer defaulters chose the lower interest rates compared to those who fully repaid.

5. ***Installment Distribution and Default Status:***
- Most installments ranged from $170 to $390. The most common installment amount for defaults was $275, whereas the most common installment amount for fully paid loans was $310. This could be due to differences in terms.

6. ***Annual Income Distribution and Default Status:***
- Most annual incomes were between $40,000 and $79,000. The average annual income of charged-off borrowers was $57,000, whereas the average annual income for fully paid loans was $65,000.

7. ***Debt-to-Income (DTI) Ratio Distribution and Default Status:***
- Defaults had a higher average DTI of 14% compared to 13% for fully paid loans. Most DTI ratios ranged from 8% to 18% for fully paid loans, while defaults had DTI ratios between 9% and 19%.

8. ***Revolving Utilization Rate Distribution and Default Status:***
- Defaults had a higher average revolving utilization rate of 56% compared to 47% for fully paid loans. Most revolving utilization rates ranged from 24% to 71% for fully paid loans, while defaults had rates between 35% and 79%.

9. ***Earliest Credit Line:***
- On average, the number of months since the earliest credit line for borrowers was roughly over 3 years (160 months).

10. ***Open and Total Credit Lines:***
- The most common number of open credit lines was 7, and the most common number of total credit lines was 14-16.

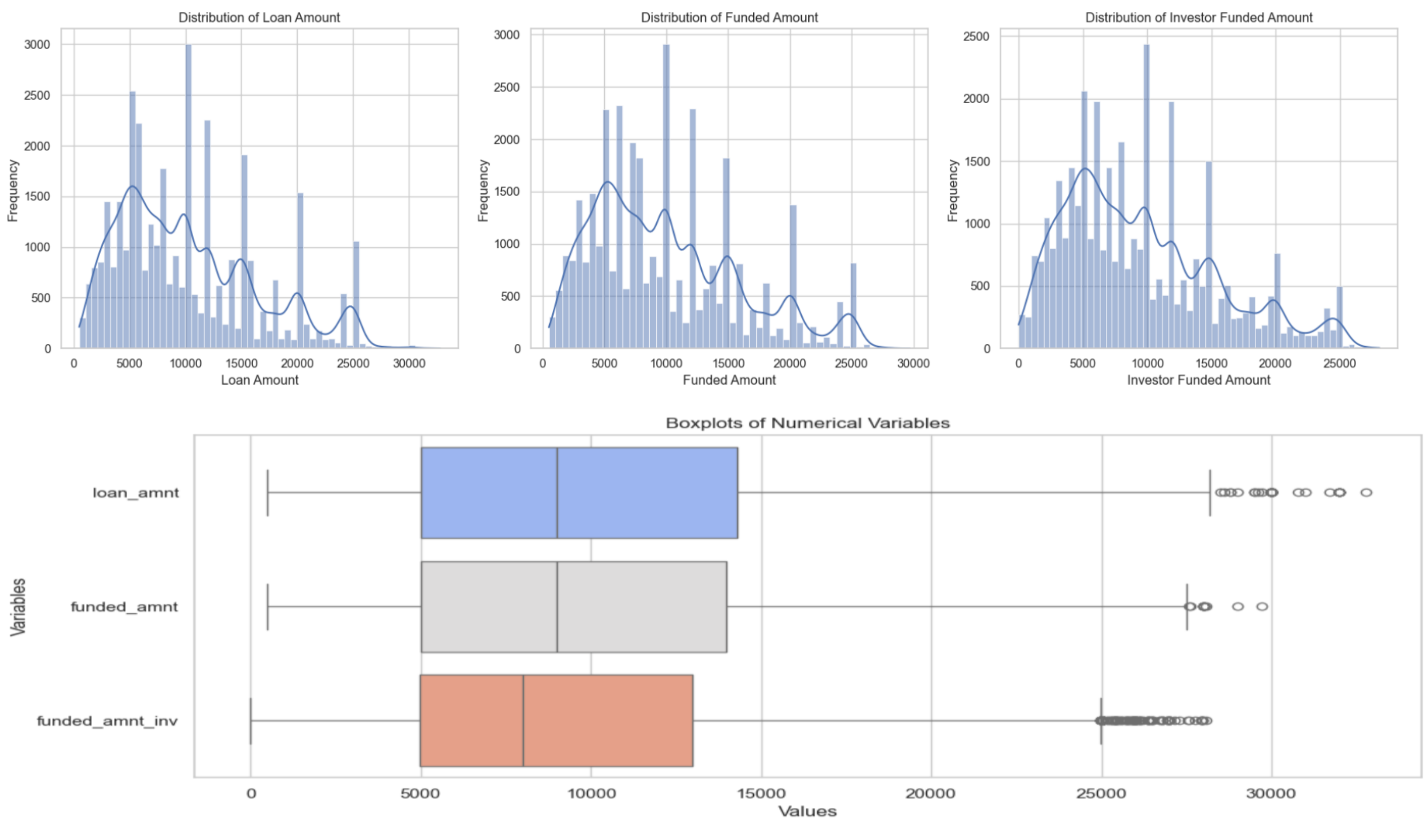
OUTCOME:

Based on the Univariate and Segmented Univariate Analysis, we can conclude that the following quantitative attributes may be the **numerical indicators** for charged-off loans:

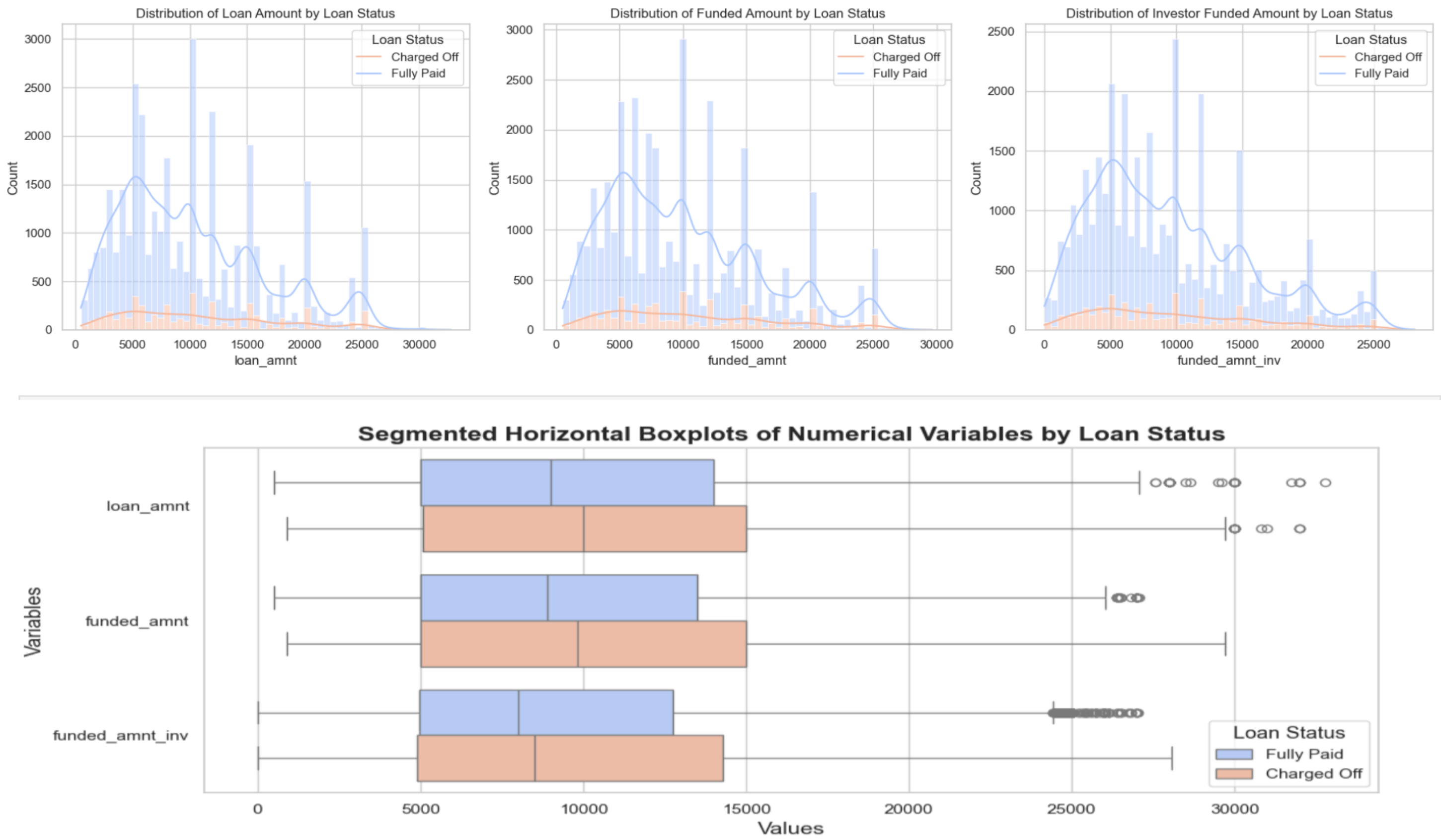
1. **Loan Amount**
2. **Interest Rate**
3. **Annual Income**
4. **Debt-to-Income (DTI) Ratio**
5. **Revolving Utilization Rate**

Univariate & Segmented Univariate >> Numerical Variables

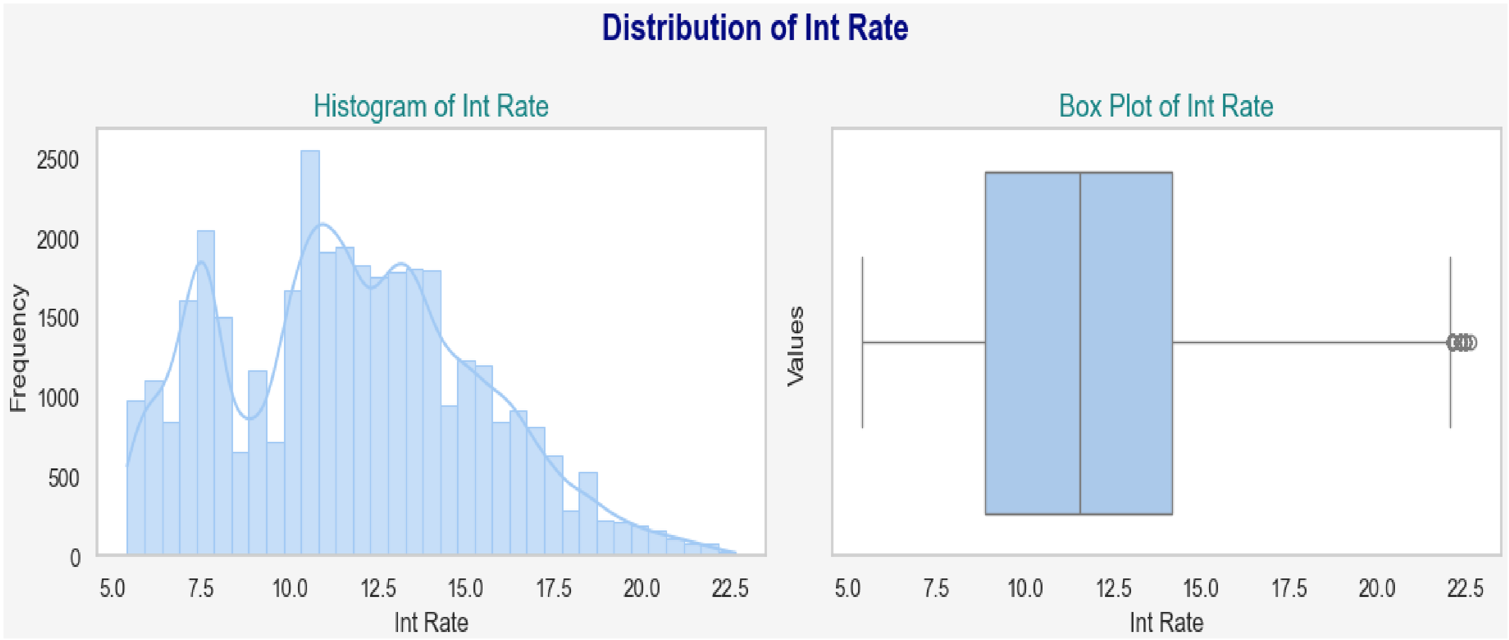
Loan Amount Univariate Analyses



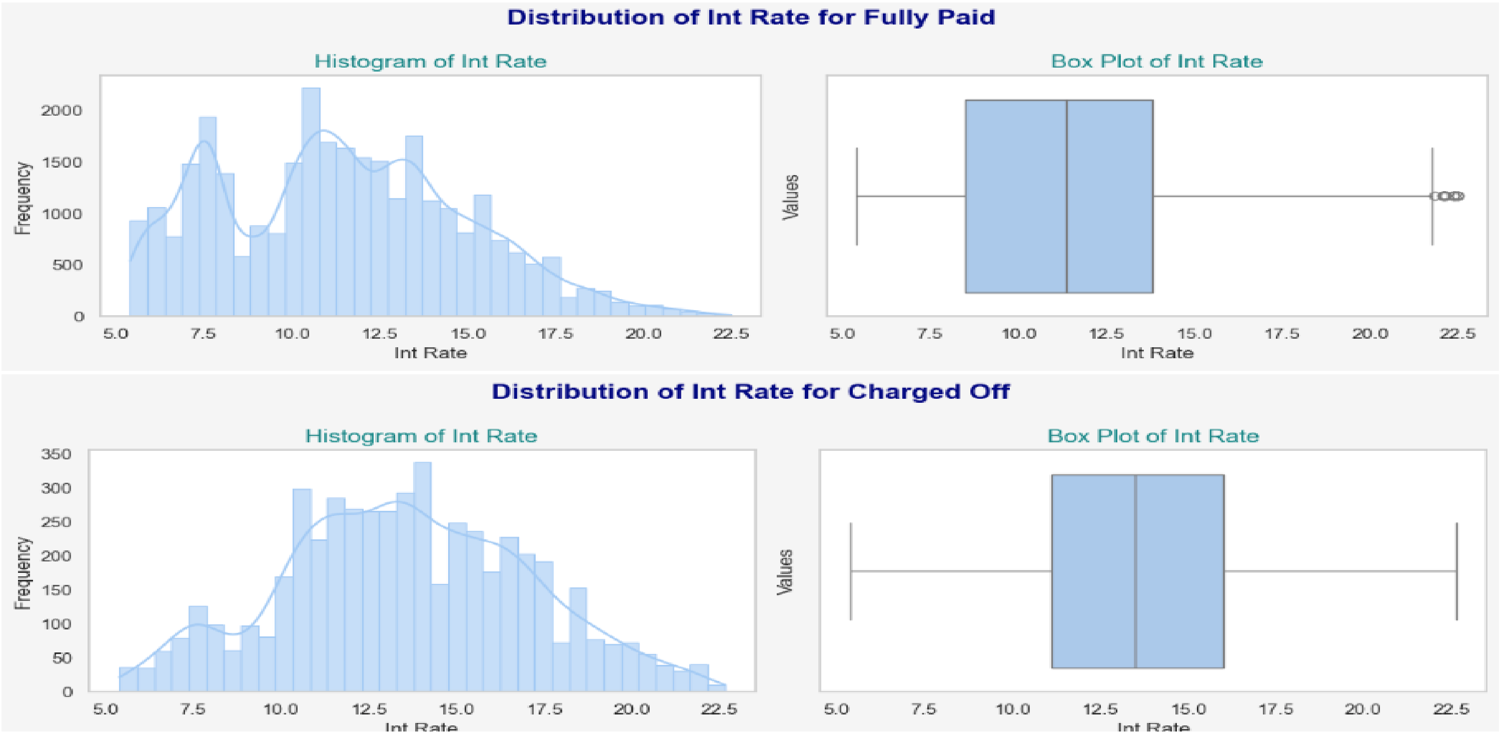
Loan Amount segmented Univariate analyses



Interest rate univariate Analyses

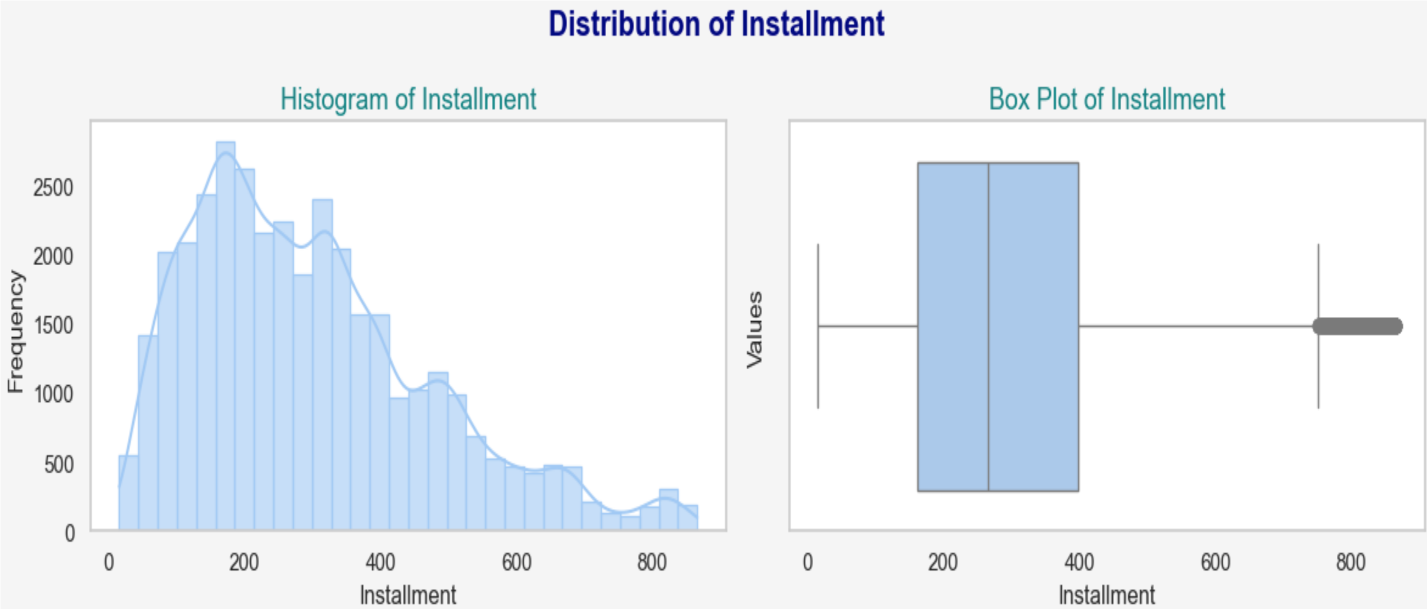


Interest rate segmented univariate Analyses

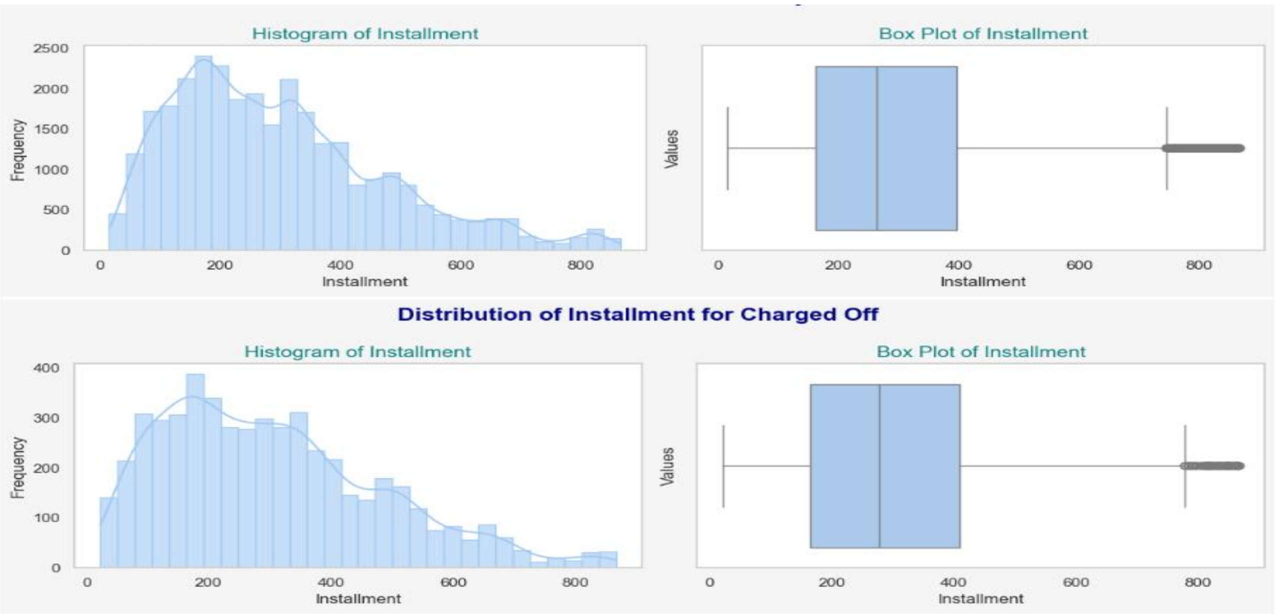


Univariate & Segmented Univariate Analyses - Numerical Variables

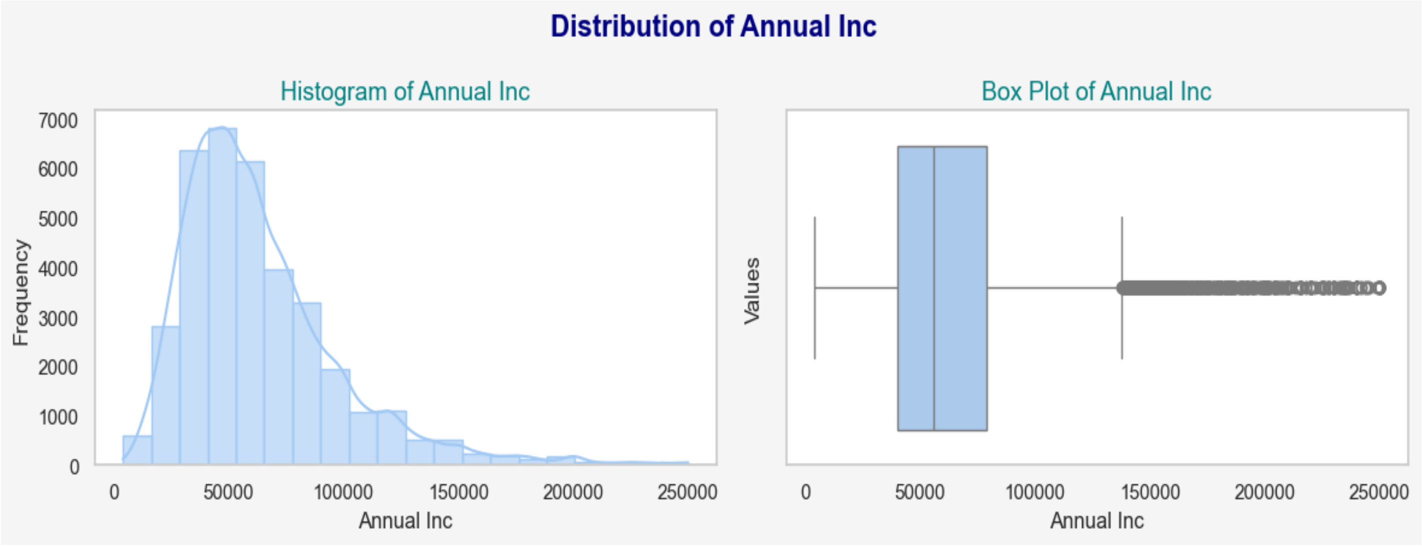
installment Univariate Analyses



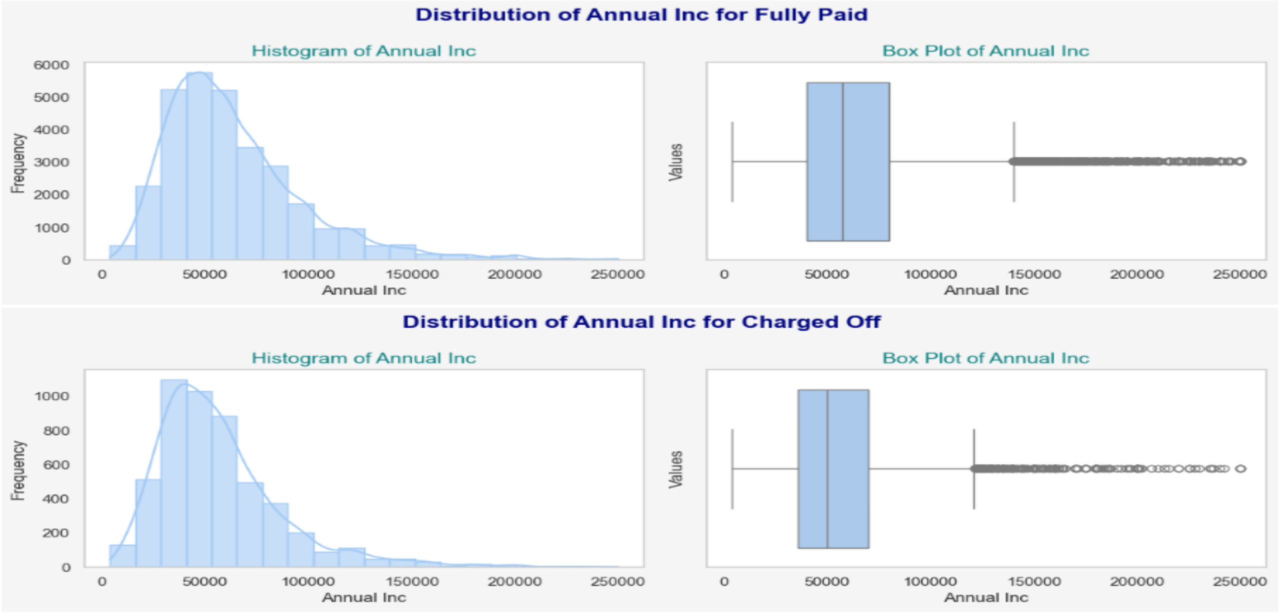
Segmented installment Univariate Analyses



Annual income Univariate Analyses



Segmented Annual income Univariate Analyses



Univariate >> Categorical Variables

1. Term:

- Most loans had a term of 36 months(77%), i.e. 3 years, and it was common among both defaulters and fully paid borrowers.
- However, proportions differs for fully paid borrowers, the ratio of 36-month to 60-month terms is 80:20, while for defaulters it is 60:40.

2. Loan Grade:

- Loans with Grade B were the most frequent (31%) among both defaulters and fully paid borrowers, followed by A(27%) and C(21%). - However, loans with Grades A(30%), B(31%) and C(20%) were the most popular among fully paid borrowers whereas Lower grades B(26%), C(25%), D(20%) were more common among defaulters.

3. Sub Grade:

- Sub-Grades like A4, A5 and B3 were the most frequently issued.
- Defaulters tended to have loans with lower sub-grades, i.e. B3, B4, B5, C1, C2 compared to those who fully repaid their loans who had loans of sub-grades, A4, A5 and B3.

4. Purpose:

- The most common loan purpose was debt consolidation(46.5%).
- Borrowers using loans for this purpose had a higher likelihood of default compared to other purposes like home improvement or medical expenses.
- 48% defaulters and 46 fully paid borrowers specified debt consolidation as their loan purpose.

5. Employment Length:

- Most borrowers had an employment length of more than 10 years(24%).
- Borrowers with higher employment lengths had a higher rate of default compared to those with shorter employment durations.
- 27% defaulters and 23% fully paid borrowers more than 10 years of employment lengths.

6. Income Category:

- Most borrowers fell into the Medium income category(46%), followed by Low-income category(42%).
- Defaulters were more likely to be in the Low-income category(50%), whereas fully paid borrowers were predominantly in the Medium income category(46.5%).

7. Home Ownership:

- Most borrowers were either renters(49%) or had mortgages(43%). Defaulters were more likely to be renters(52%) compared to those who fully repaid their loans(49%).

8. Verification Status:

- Most borrowers were Not-Verified(45%). 40% defaulters were not-verified and 34% defaulters were verified, whereas 46% fully paid borrowers were not-verified and 29% were verified. This suggests that verification status may not effectively identify those who default.

9. Region:

- The majority of loans originated from the western regions such as the West(30%) and the Mid West(14%). Proportion of Loans emerging in these regions are similar for both Defaulters and Fully Paid Borrowers.
- 10. Address State:
- Loans were most issued in states like California, (CA: 18%), New York (NY: 9.5%), Texas(TX: 7%), Florida (FL: 7%). Proportion of Loans emerging in these states are similar for both Defaulters and Fully Paid Borrowers.

11. Inquiries in the Past 6 Months:

- Borrowers with one(28%) or no(49%) inquiries in the past 6 months were most common.
- It is difficult to identify default rate patterns because 50.5% of the defaulters had no inquiries in the past 6 months.

12. Issue Year:

- Loans issued had been increase subsequently year on year between 2007 and 2011 in recent years, and year 2011 saw 20% increase in loan issuance and with it we saw a higher incidence of default (20% increase) compared to 2010.

13. Issue Month:

- The majority of loans were issued increases month-on-month every year, peaking in December. Defaults varied slightly by month, sharing a similar trend with the average.

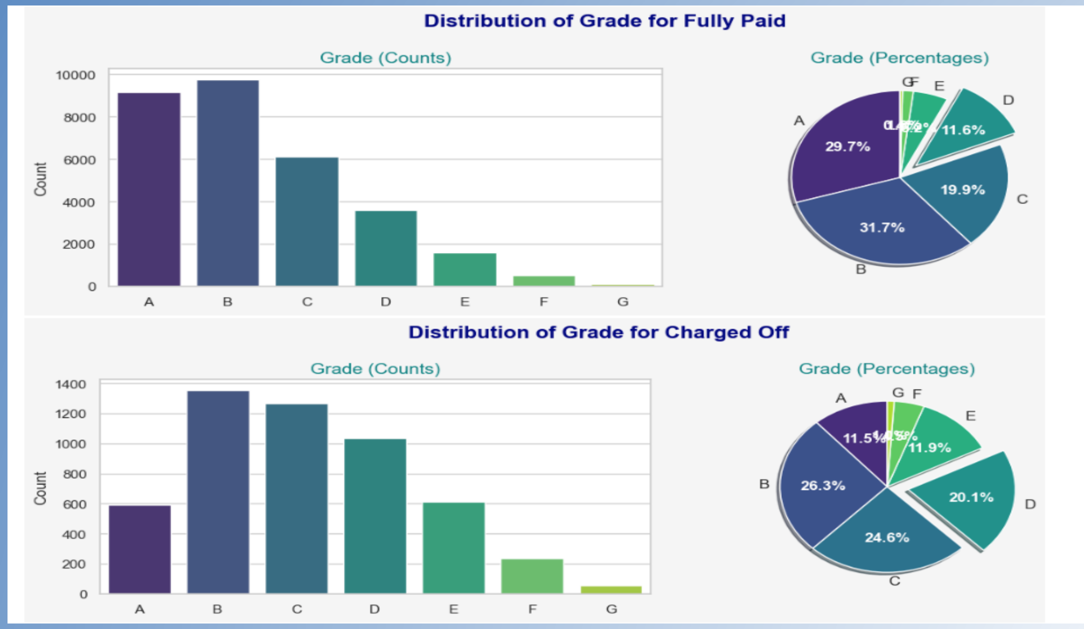
14. Issue Quarter:

- Loans issued in the fourth quarter had a higher rate of default compared to loans issued in other quarters, but similar trend is seen in the loans issued to fully paid borrowers as well.

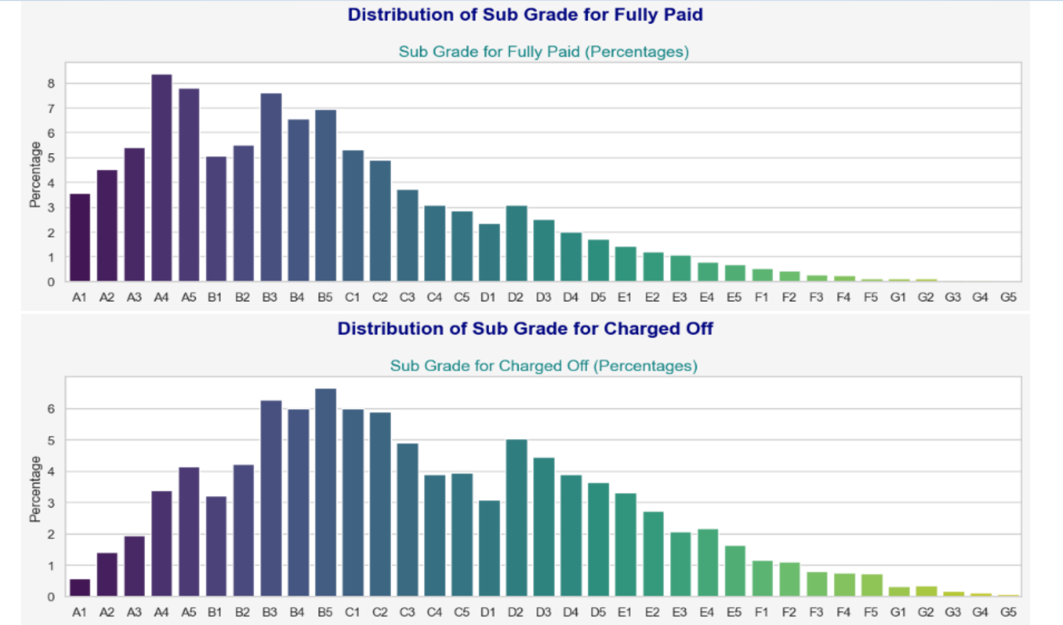
OUTCOME:

Based on the Univariate and Segmented Univariate Analysis, we can conclude that the following qualitative attributes may be the categorical indicators for charged-off loans:

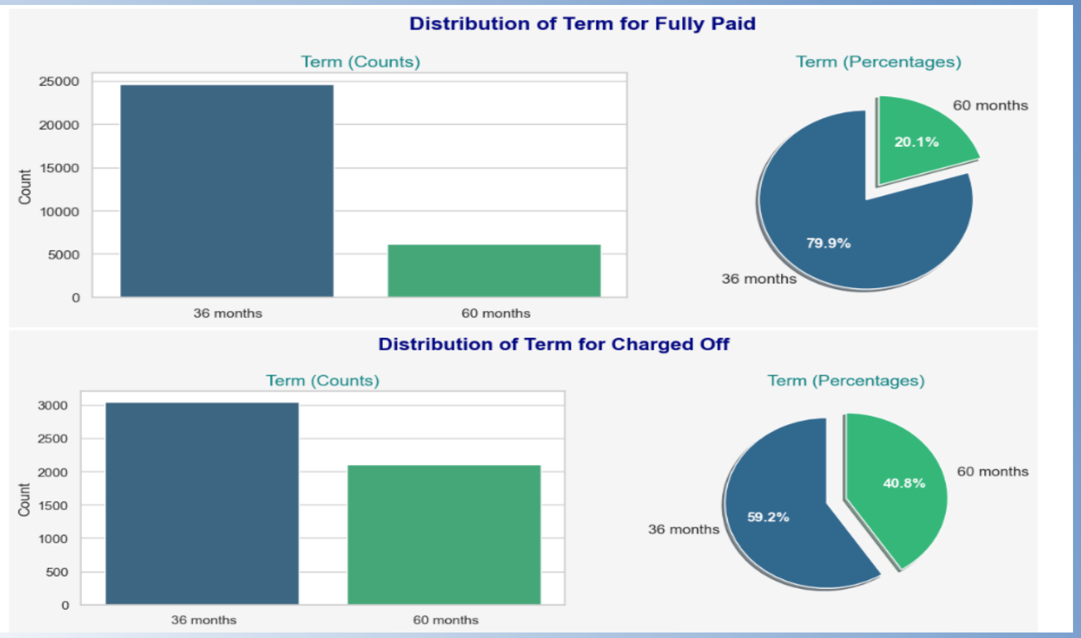
1.Term
2.Loan Grade
3.Purpose
4.Employment Length
5.Income Category
6.Home Ownership
7.Verification Status
8.Region


Distribution of Grade for Fully Paid

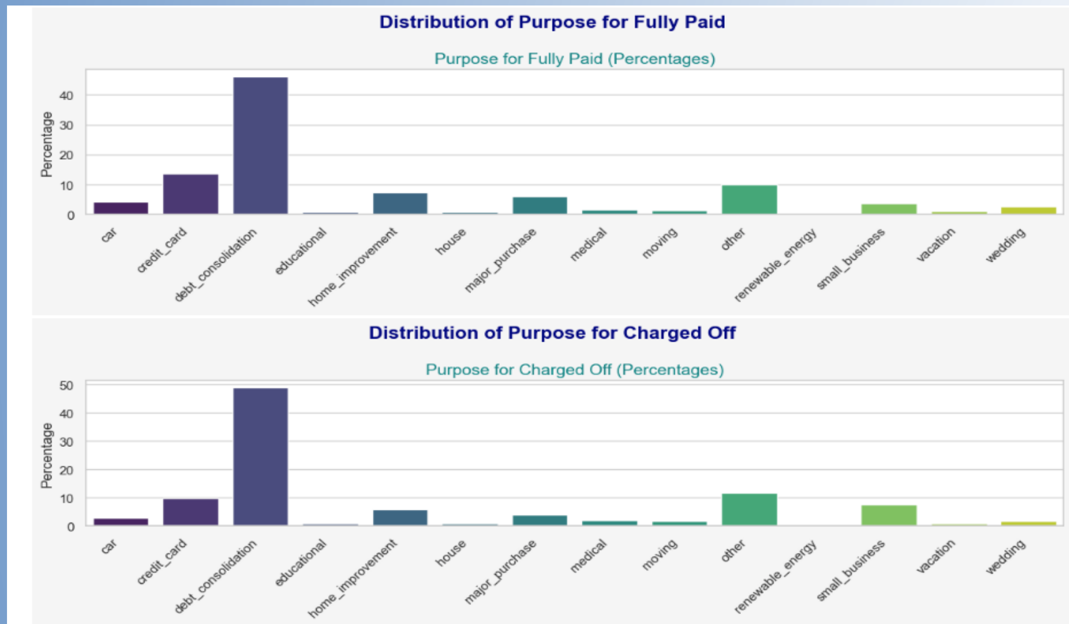
Distribution of Grade for Charged Off


Distribution of Sub Grade for Fully Paid

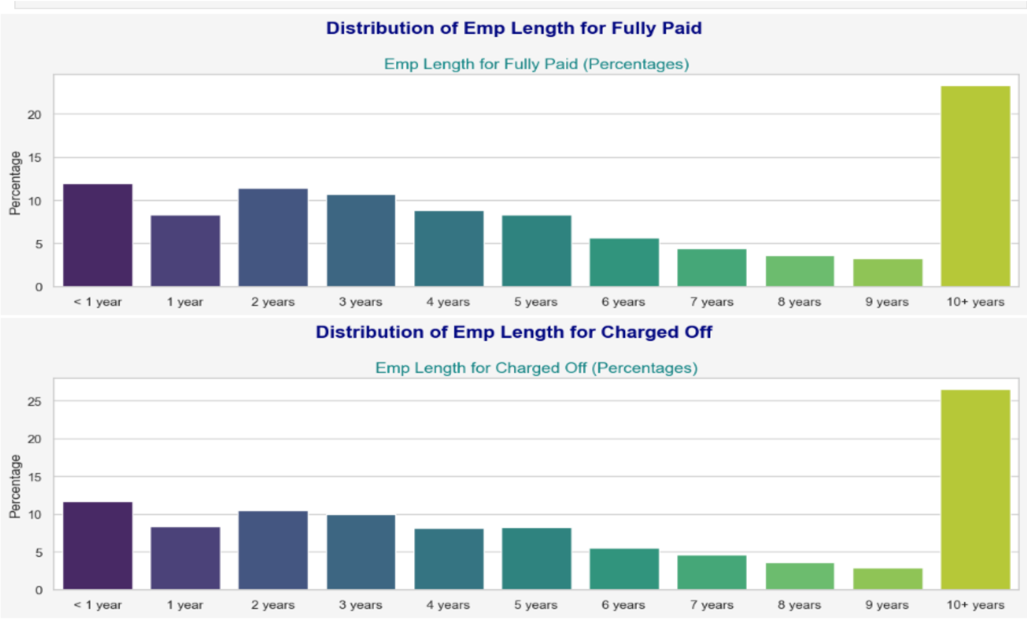
Distribution of Sub Grade for Charged Off


Distribution of Term for Fully Paid

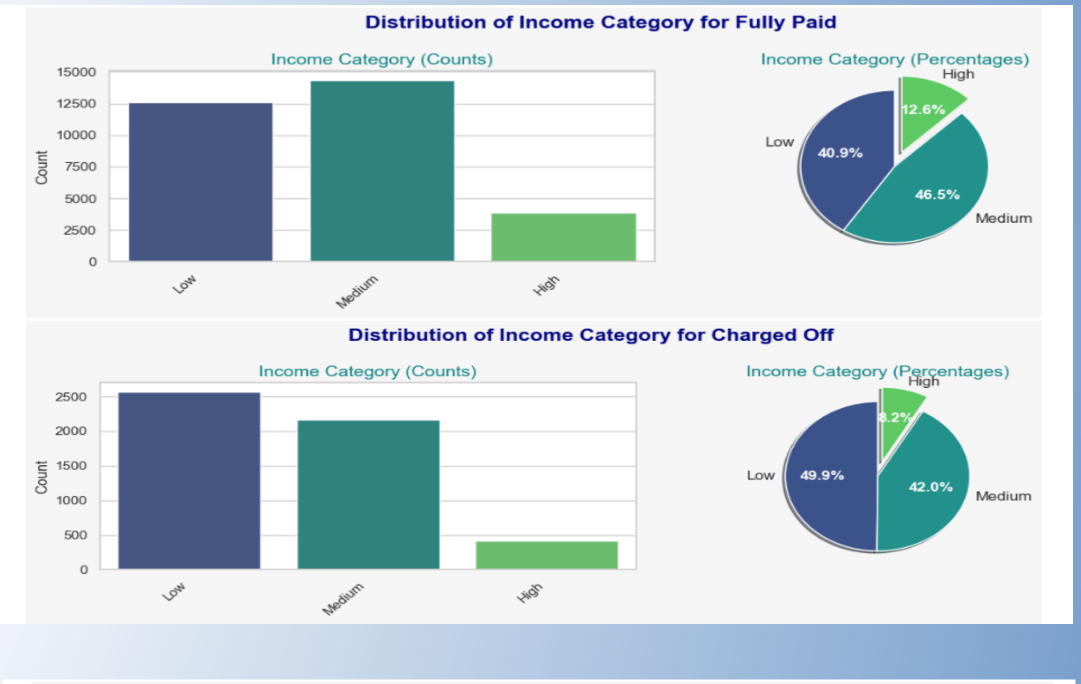
Distribution of Term for Charged Off


Distribution of Purpose for Fully Paid

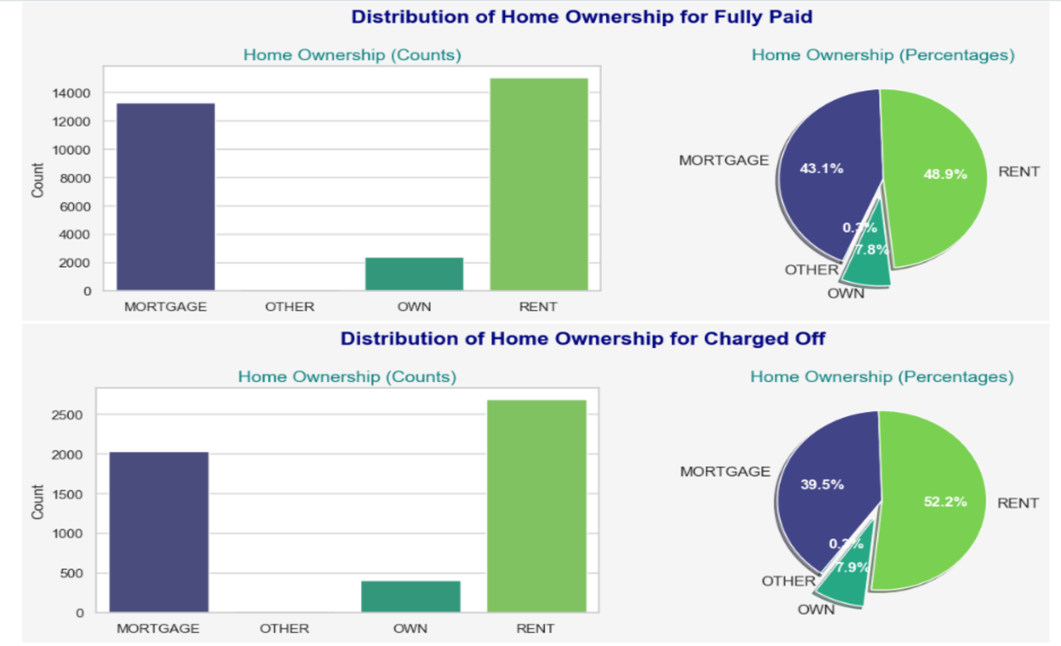
Distribution of Purpose for Charged Off


Distribution of Emp Length for Fully Paid

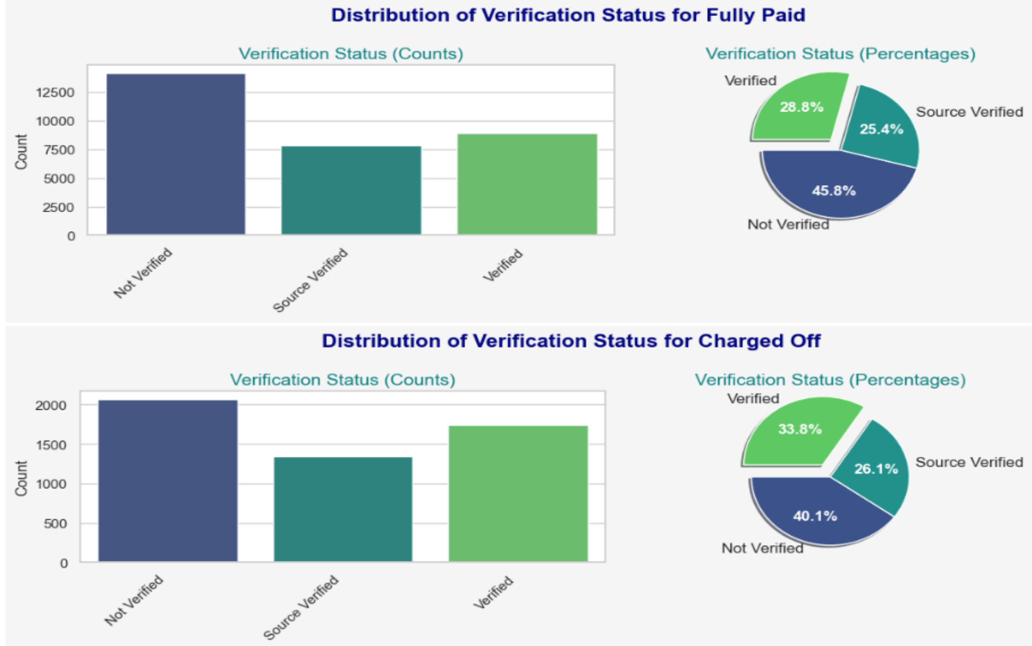
Distribution of Emp Length for Charged Off


Distribution of Income Category for Fully Paid

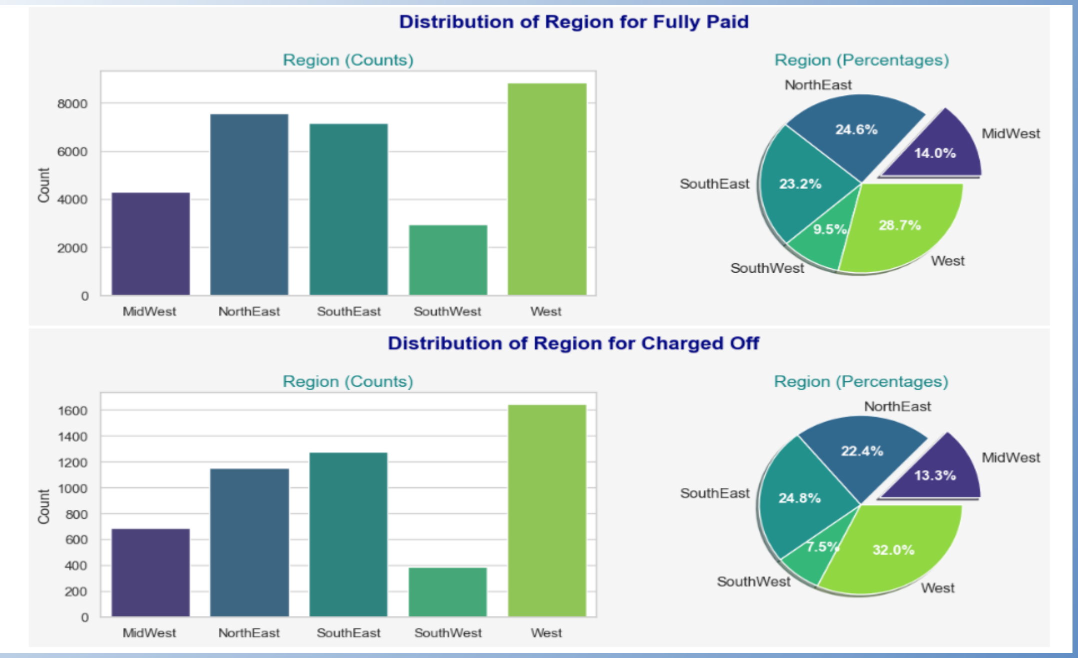
Distribution of Income Category for Charged Off


Distribution of Home Ownership for Fully Paid

Distribution of Home Ownership for Charged Off


Distribution of Verification Status for Fully Paid

Distribution of Verification Status for Charged Off


Distribution of Region for Fully Paid

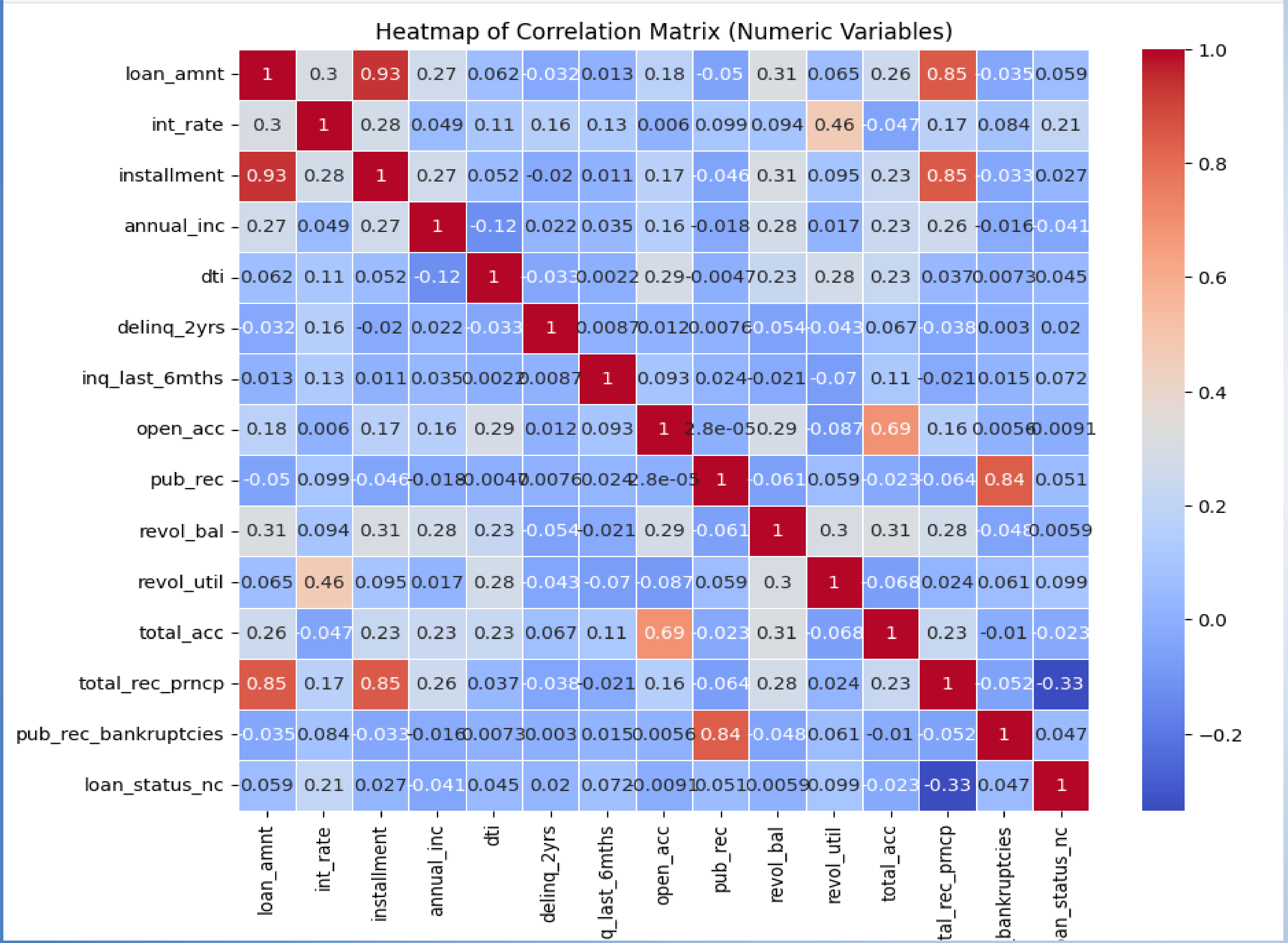
Distribution of Region for Charged Off

Bivariate Analyses

with respect to the charged off loans for each variable suggests the following. There is a more probability of defaulting when :

- Applicants having house ownership as 'RENT'
- Applicants who use the loan to clear other debts
- Applicants who receive interest at the rate of 13-17%
- Applicants who have an income of range 31201 - 58402
- Applicants who have 10-19 open acc
- Applicants with employment length of 10 or more years
- Loan amount is between 5429 - 10357
- DTI is between 12-18
- When monthly instalments are between 145-274
- Term of 36 months
- When the loan status is Not verified
- When the no of enquiries in last 6 months is 0
- When the number of derogatory public records is 0
- When the purpose is 'debt consolidation'
- Grade is 'B'
- And a total grade of 'B5' level.
- Applicants last credit pulled month falls under May -2016
- Applicants loan issued in the month of December 2011
- Applicants from the region CA - 23%
- Applicants revol util is between 60-80%
- Applicants with source not verified
- Borrowers with Charged Off Loans most likely belonged to the West Region by count but is almost uniform across all regions by percentage.
- Borrowers with Charged Off Loans mostly belongs to the **Low-Income Category** by both count and percentages.
- Borrowers with Charged Off Loans most likely had no inquiries by count and 3 inquiries by percentages, but there is only slight variation across the dataset.
- Borrowers with Charged Off Loans in general increased year-on-year, with slight variations over months and quarters by count. And by percentages, there is only slight variations in loans in years, quarters and months.

Bivariate Analyses >> Numerical vs Numerical columns



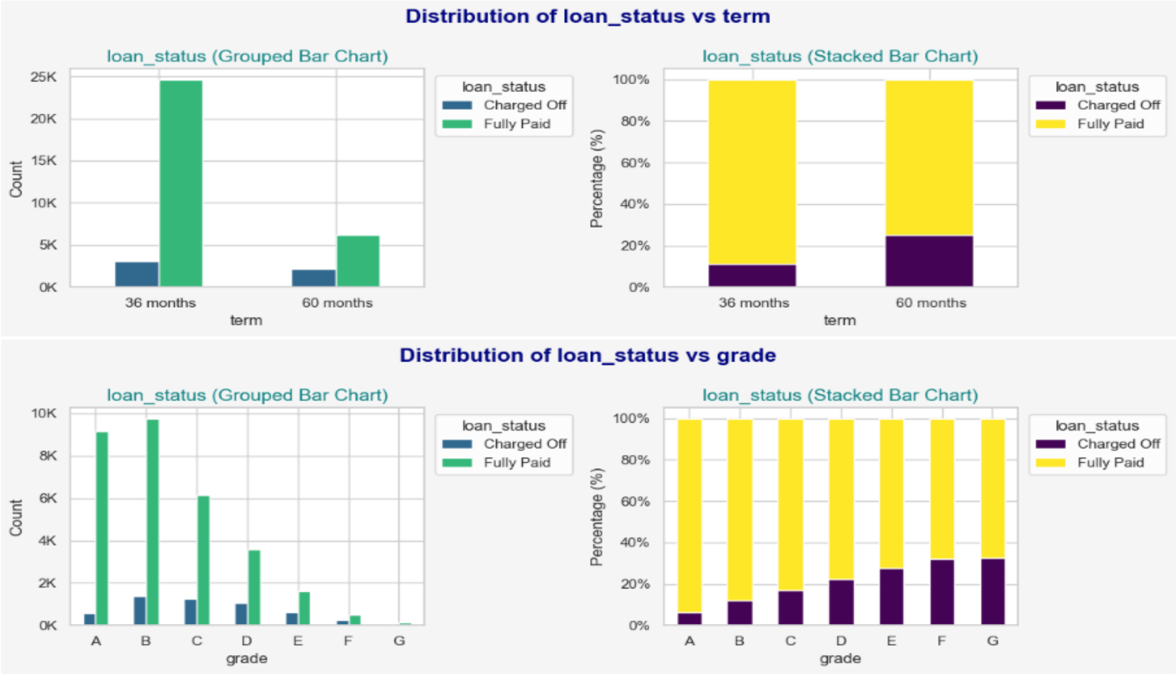
Heatmap of Correlation Matrix (Numeric Variables)

	loan_amnt	int_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	total_acc	total_rec_prncp	pub_rec_bankruptcies	loan_status_nc
loan_amnt	1	0.3	0.93	0.27	0.062	-0.032	0.013	0.18	-0.05	0.31	0.065	0.26	0.85	-0.035	0.059
int_rate	0.3	1	0.28	0.049	0.11	0.16	0.13	0.006	0.099	0.094	0.46	-0.047	0.17	0.084	0.21
installment	0.93	0.28	1	0.27	0.052	-0.02	0.011	0.17	-0.046	0.31	0.095	0.23	0.85	-0.033	0.027
annual_inc	0.27	0.049	0.27	1	-0.12	0.022	0.035	0.16	-0.018	0.28	0.017	0.23	0.26	-0.016	-0.041
dti	0.062	0.11	0.052	-0.12	1	-0.033	0.0022	0.29	-0.0047	0.23	0.28	0.23	0.037	0.0073	0.045
delinq_2yrs	-0.032	0.16	-0.02	0.022	-0.033	1	0.0087	0.012	0.0076	-0.054	-0.043	0.067	-0.038	0.003	0.02
inq_last_6mths	0.013	0.13	0.011	0.035	0.0022	0.0087	1	0.093	0.024	-0.021	-0.07	0.11	-0.021	0.015	0.072
open_acc	0.18	0.006	0.17	0.16	0.29	0.012	0.093	1	2.8e-05	0.29	-0.087	0.69	0.16	0.0056	0.0091
pub_rec	-0.05	0.099	-0.046	-0.018	-0.0047	0.0076	0.024	2.8e-05	1	-0.061	0.059	-0.023	-0.064	0.84	0.051
revol_bal	0.31	0.094	0.31	0.28	0.23	-0.054	-0.021	0.29	-0.061	1	0.3	0.31	0.28	-0.048	0.0059
revol_util	0.065	0.46	0.095	0.017	0.28	-0.043	-0.07	-0.087	0.059	0.3	1	-0.068	0.024	0.061	0.099
total_acc	0.26	-0.047	0.23	0.23	0.23	0.067	0.11	0.69	-0.023	0.31	-0.068	1	0.23	-0.01	-0.023
total_rec_prncp	0.85	0.17	0.85	0.26	0.037	-0.038	-0.021	0.16	-0.064	0.28	0.024	0.23	1	-0.052	-0.33
pub_rec_bankruptcies	-0.035	0.084	-0.033	-0.016	0.0073	0.003	0.015	0.0056	0.84	-0.048	0.061	-0.01	-0.052	1	0.047
loan_status_nc	0.059	0.21	0.027	-0.041	0.045	0.02	0.072	0.0091	0.051	0.0059	0.099	-0.023	-0.33	0.047	1

Based on the Bivariate Correlation of Numerical Variables with Loan_Status, we can conclude that the following quantitative attributes may be the numerical indicators for charged-off loans:

1. ***Loan Amount***
2. ***Interest Rate***
3. ***Annual Income***
4. ***Debt-to-Income (DTI) Ratio***
5. ***Revolving Utilization Rate***

Bivariate Analyses >> categorical vs categorical columns


Distribution of loan_status vs term

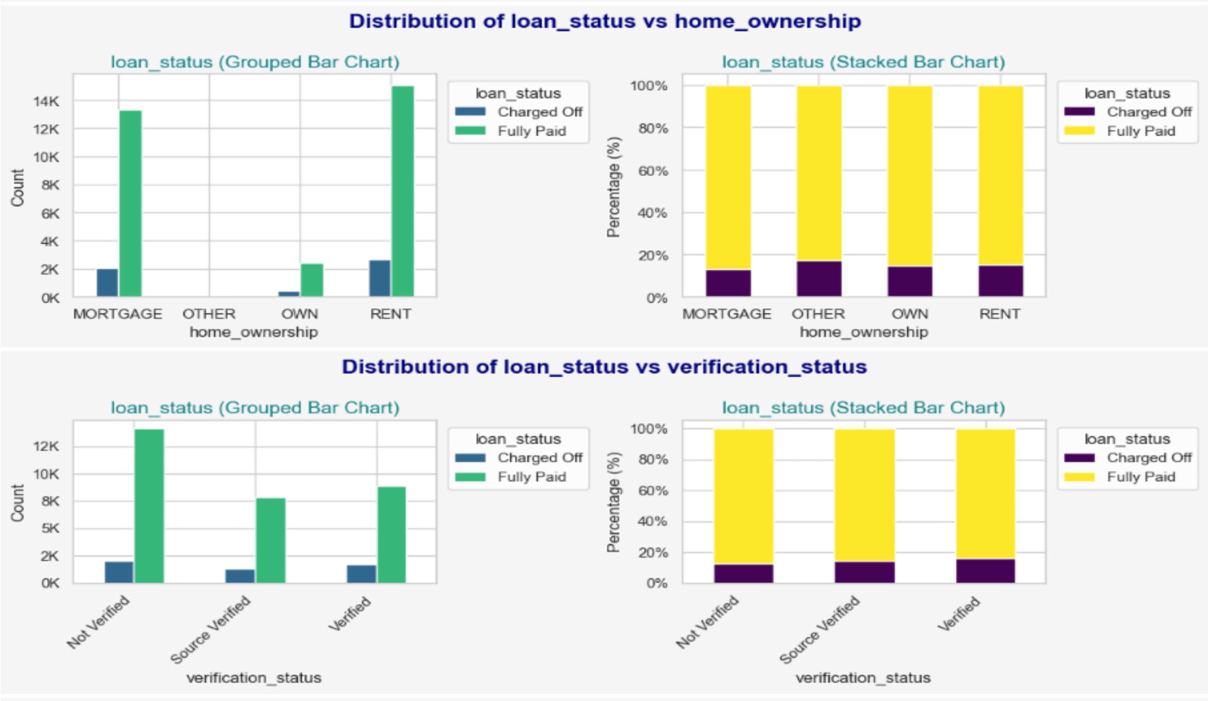
loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Distribution of loan_status vs grade

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)


Distribution of loan_status vs home_ownership

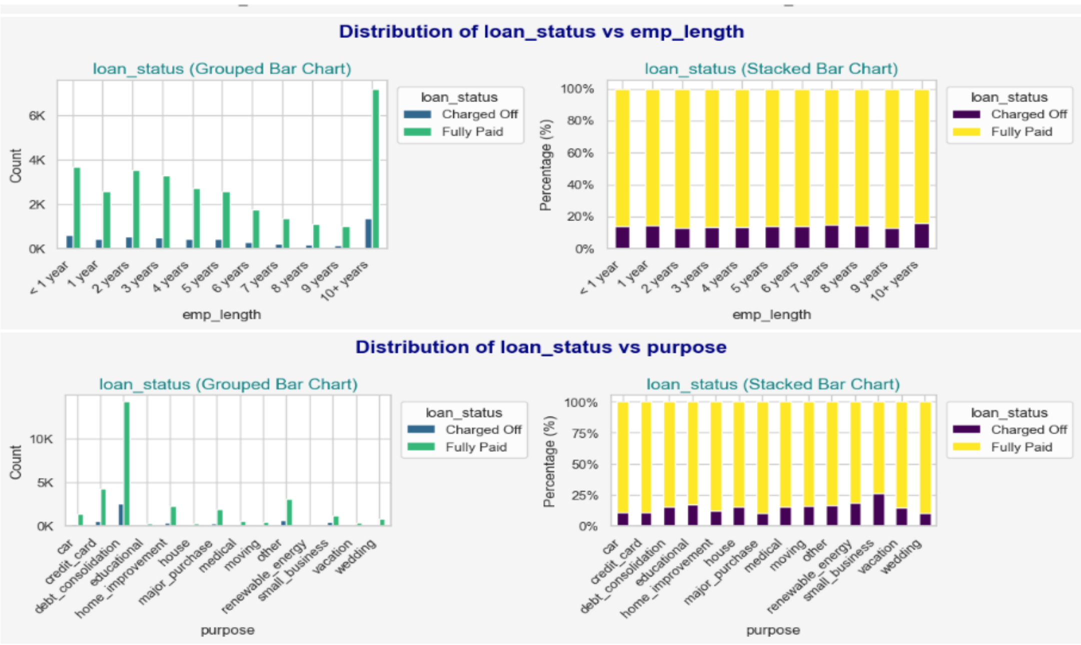
loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Distribution of loan_status vs verification_status

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)


Distribution of loan_status vs emp_length

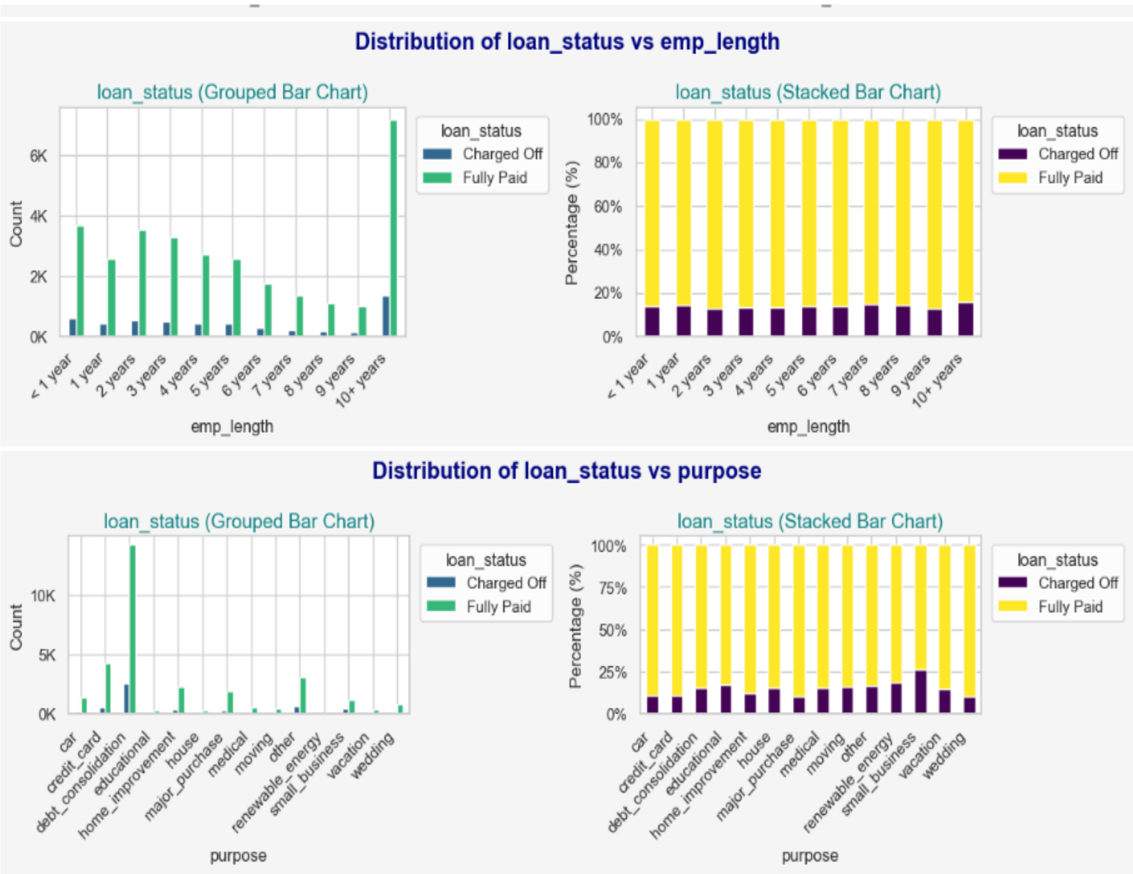
loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Distribution of loan_status vs purpose

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)


Distribution of loan_status vs emp_length

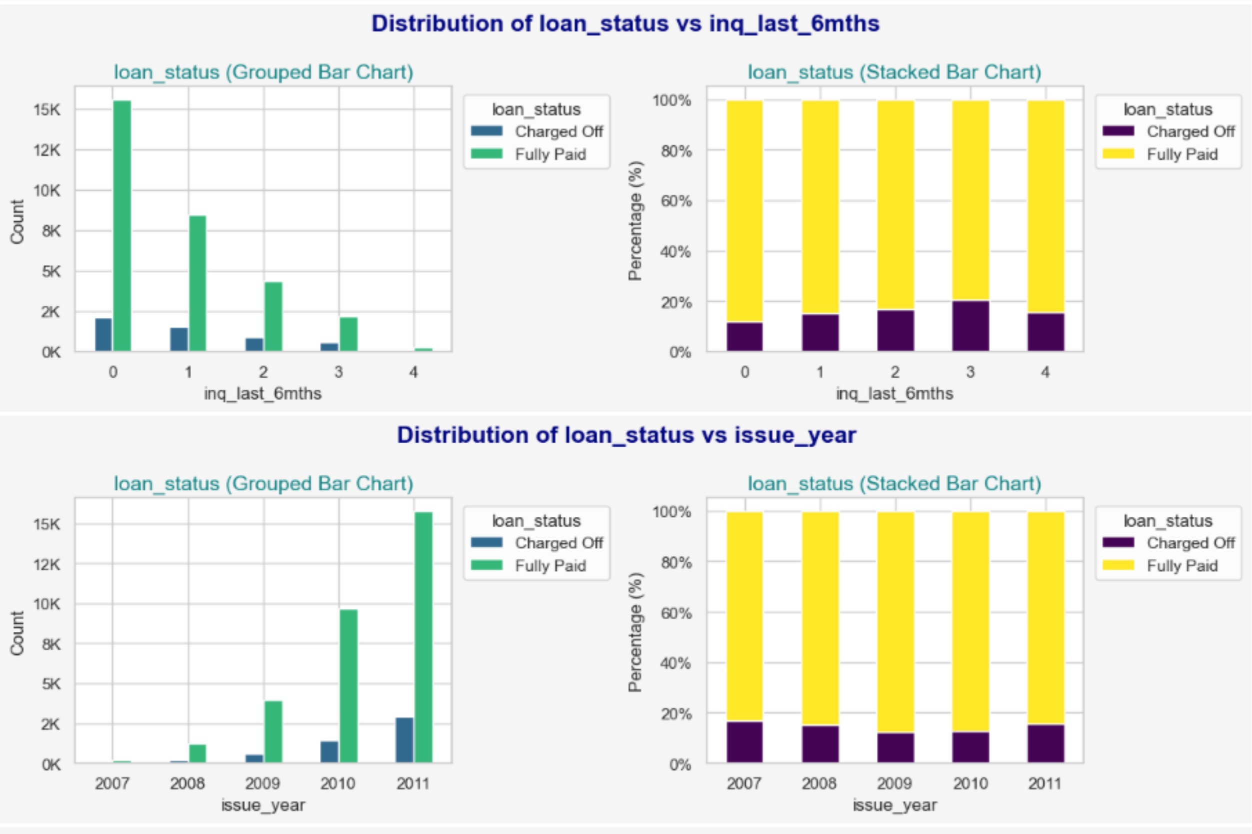
loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Distribution of loan_status vs purpose

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)


Distribution of loan_status vs inq_last_6mths

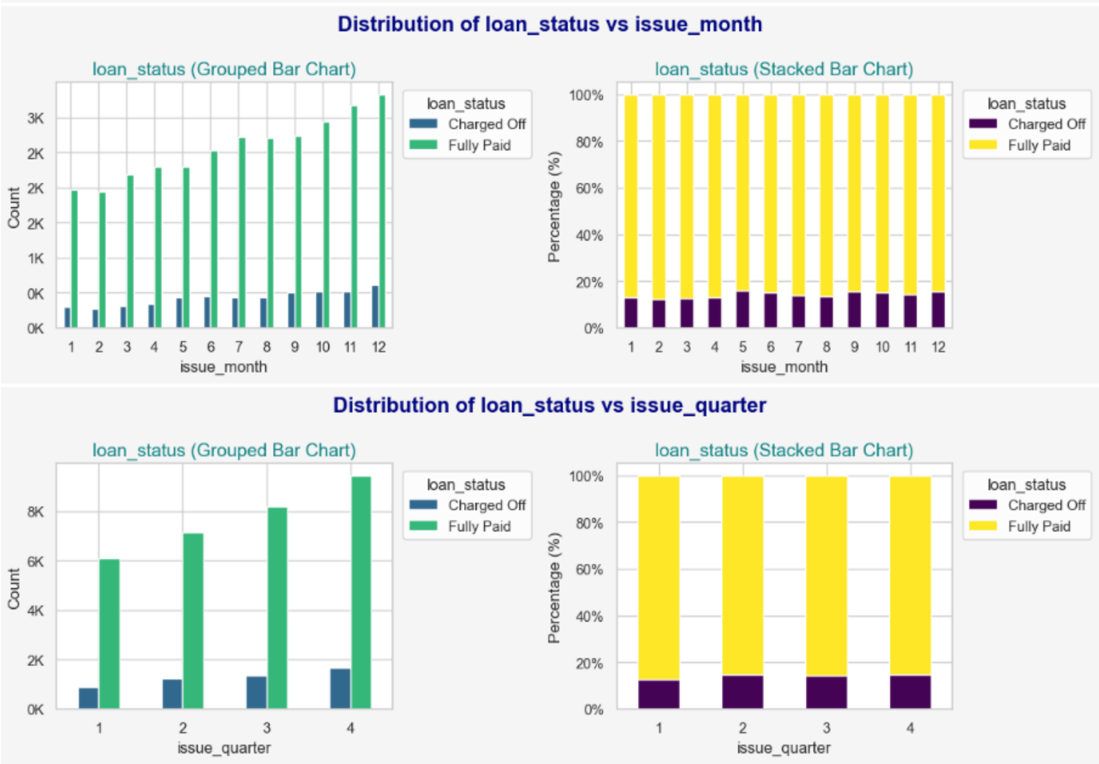
loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Distribution of loan_status vs issue_year

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)


Distribution of loan_status vs issue_month

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Distribution of loan_status vs issue_quarter

loan_status (Grouped Bar Chart)

loan_status (Stacked Bar Chart)

Bivariate Analyses >> Numerical vs Categorical

Loan Amount vs. Categorical Variables:

- **Term:** Boxplots and violin plots show a higher median loan amount for longer-term loans (' 60 months') compared to shorter-term loans (' 36 months'). The distribution for longer-term loans is also more spread out.
- **Grade:** Higher loan amounts are generally associated with lower grades ('E' to 'G'). Violin plots indicate that 'G' grades tend to have higher loan amounts, with a wide distribution in higher grades.
- **Purpose:** Boxplots reveal that loan amounts do not vary significantly based on the loan purpose. Loans for 'credit card', 'debt consolidation', 'home improvement' and 'small business' show higher median amounts compared to others.
- **Income Category:** The loan amount increases with the income category. Higher income categories show a higher median loan amount, with broader distribution in the 'High' income category.

Interest Rate vs. Categorical Variables:

- **Term:** Interest rates are higher for long-term loans ('60 months') compared to shorted-term loans ('36 months'). The distribution is similar for both the term.
- **Grade:** Interest rates are higher for lower grades. For instance, 'G' grades tend to have higher interest rates compared to 'A' grades.
- **Purpose:** Different loan purposes have varied interest rates. Loans for 'house' and 'small business' have higher interest rates compared to other loans.
- **Income Category:** Interest rates is almost similar for all income categories.

Annual Income vs. Categorical Variables:

- **Term, Grade, Purpose:** Annual income shows no significant trend in central tendency and distribution with respect to Term, Grade and Purpose.
- **Income Category:** The annual income increases with the income category. 'High' income category shows the highest annual income with the most significant variability. This mainly because the Income category is a derivative of Annual Incomes.

Debt-to-Income Ratio vs. Categorical Variables:

- **Term, Grade, Purpose:** DTI ratios show no significant trend in central tendency and distribution with respect to Term, Grade and Purpose.
- **Income Category:** DTI ratios are higher in lower income categories. 'Low' income category shows higher median DTI ratios compared to 'High' income category.

Revolving Utilization vs. Categorical Variables:

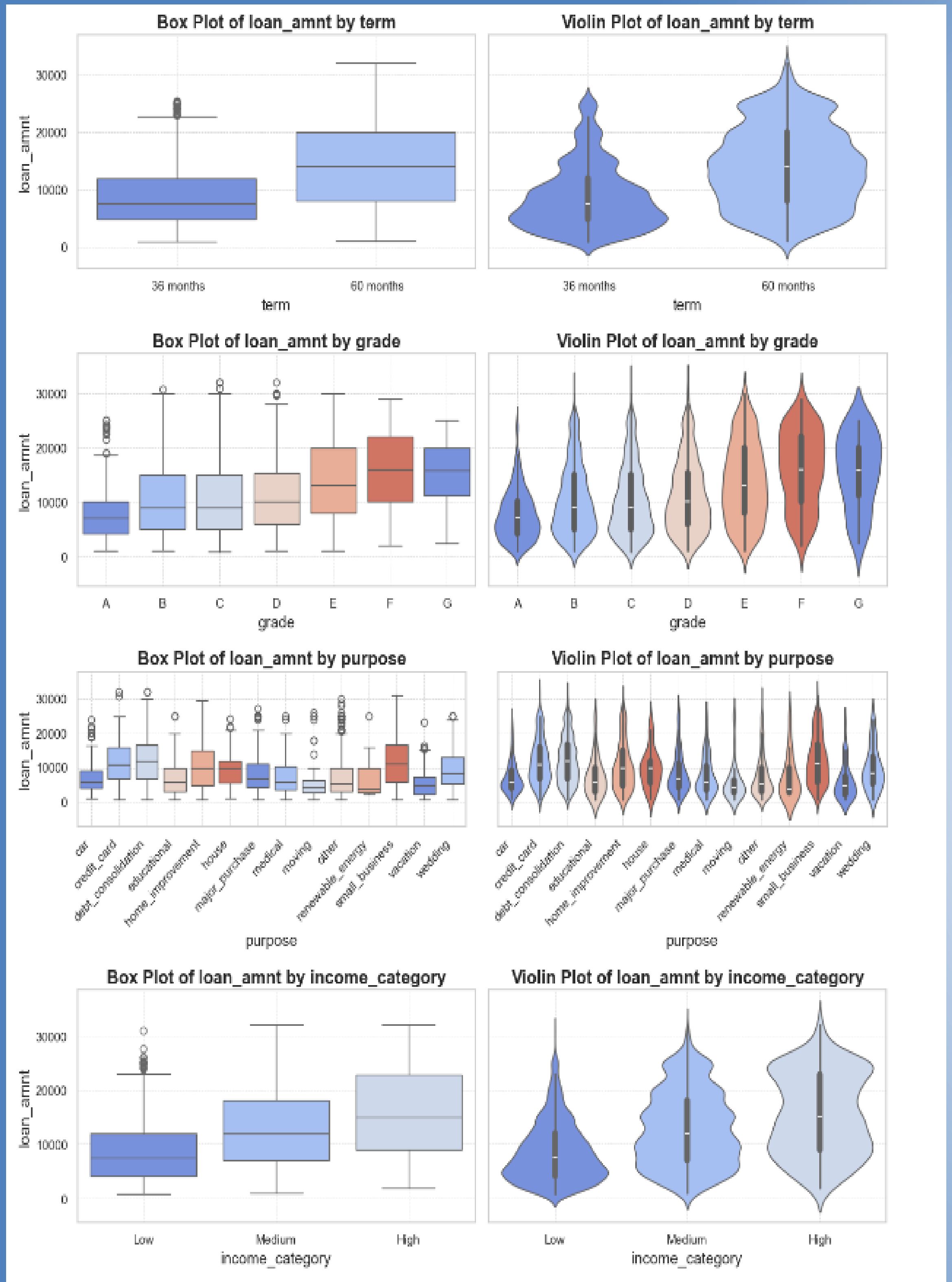
- **Term:** Revolving utilization is similar for both ' 36 months' loans and ' 60 months.
- **Grade:** Lower grades exhibit higher revolving utilization. 'F' grade borrowers, followed by G, tend to have higher utilization compared to all the other grade borrowers, but 'G' grades have a wider distribution.
- **Purpose:** Revolving utilization varies with loan purpose. 'Credit card' loans have a higher revolving utilization compared to 'debt consolidation' loans.
- **Income Category:** Higher income categories show higher revolving utilization. The 'Low' income category generally exhibits wider distribution compared to 'High' income.

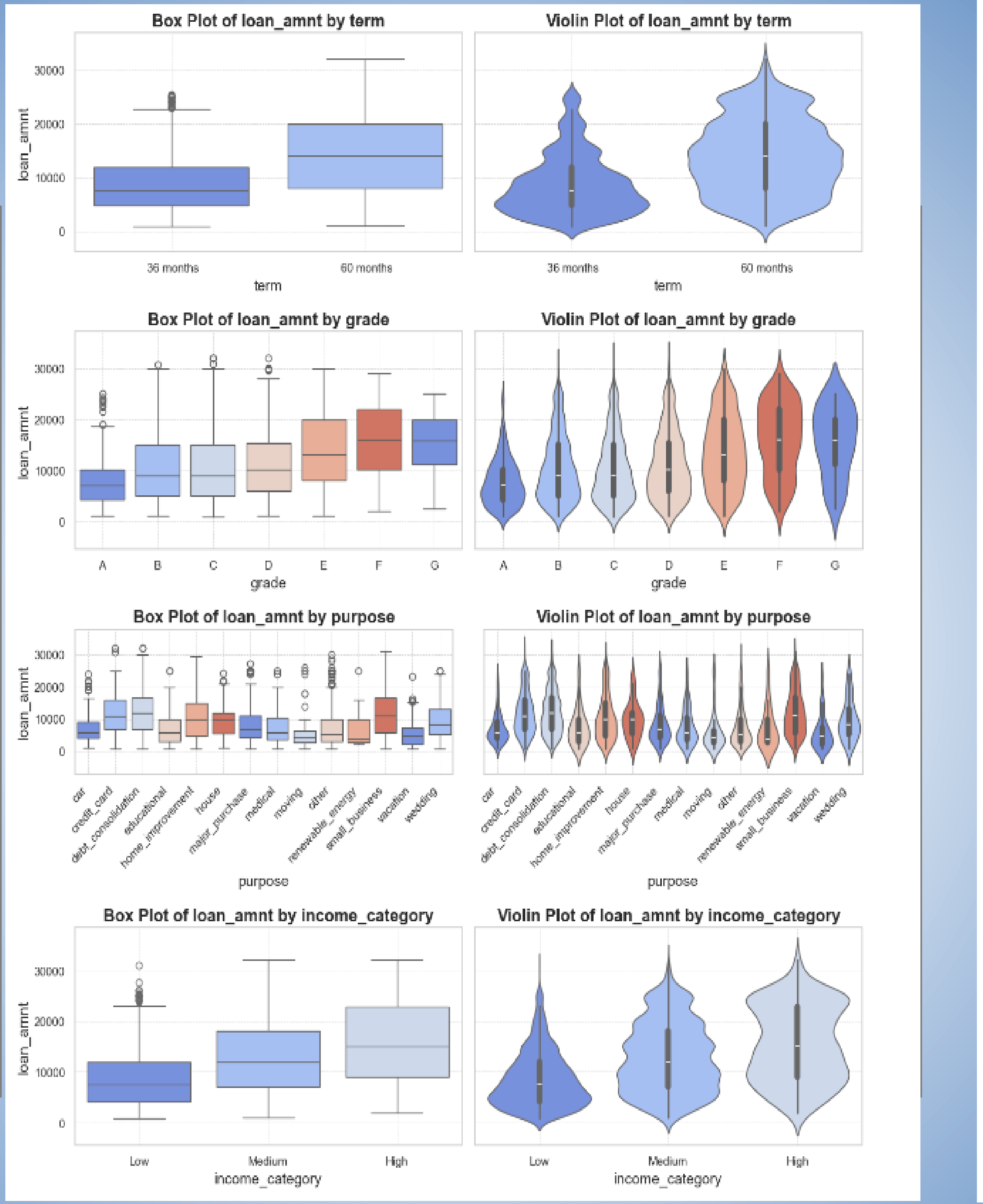
OUTCOME:

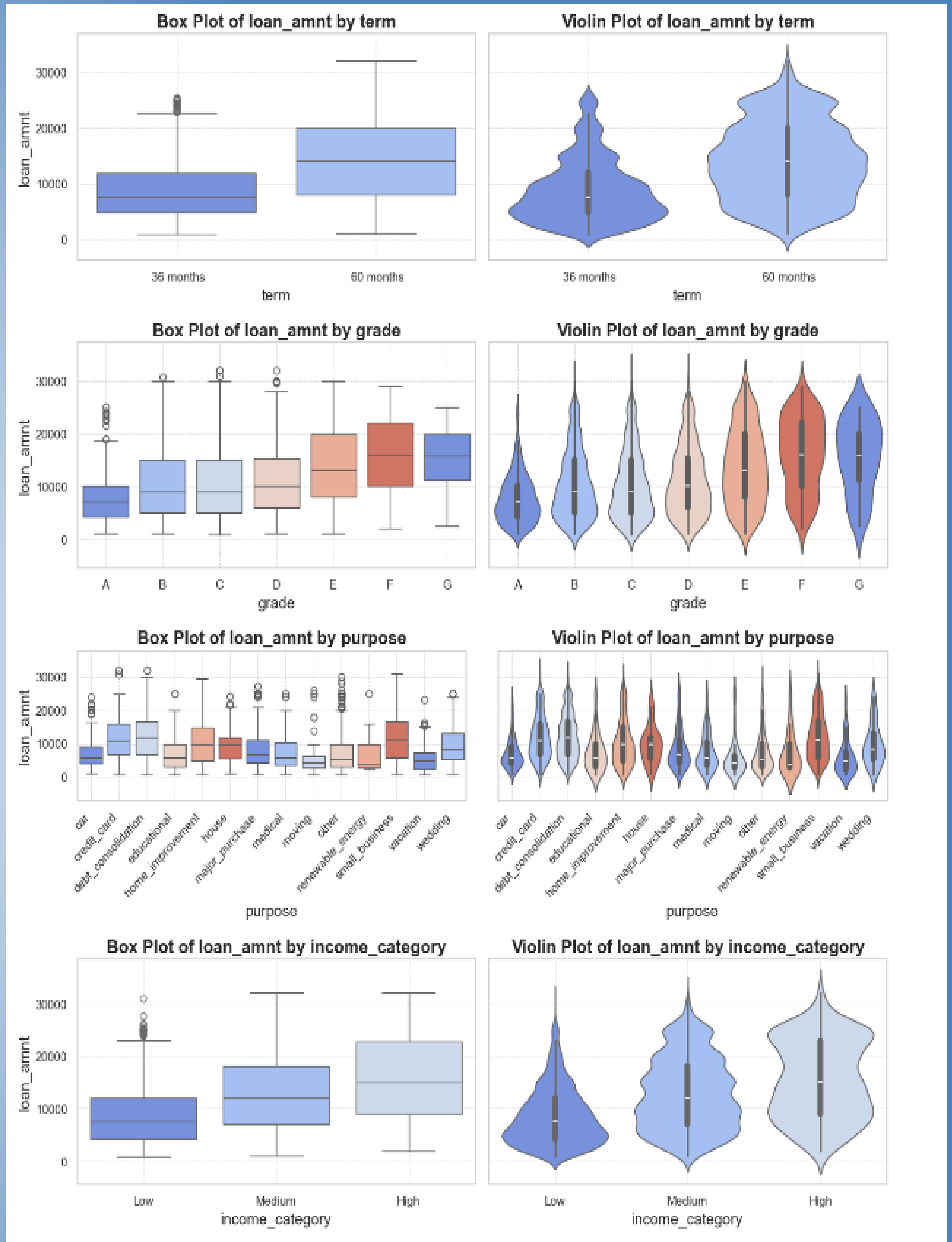
Based on the Bivariate Correlation of Numerical Variables vs Categorical Variables, we can conclude that for Charged Off Loans:

1.Loan Amount is positively correlated to Term, Grade and Income Category
2.Interest Rate is strongly correlated to Term and Grade
3.Annual Income is only correlated to Income Category which is expected
4.DTI is weakly correlated to Income Category
5.Revolving Utilization Rate is correlated to Grade and Income Category

Bivariate Analyses >> Numerical vs Categorical







Multivariate Analyses

- **High Revolving Utilization:** Applicants with high annual income are more likely to default when their revolving utilization rate is between 80% and 100%.
- **Home Ownership:** Defaults are more common among high-income applicants with a mortgage.
- **Loan Purpose:** High-income defaulters often take loans for home improvement, renewable energy, or housing.
- **Verification Status:** High-income defaulters are frequently those with verified verification status.
- **Loan Grade:** Defaults are associated with high-income applicants who have a Grade G loan.
- **Sub Grade:** High-income defaulters often have a Sub Grade of G4.
- **Loan Amount:** High-income applicants default more often when the loan amount is between $25,000 and $30,000.
- **Geographic Location:** High-income defaults are observed for loans applied in Arkansas

Loan Amount vs. Other Columns:

- **Home Ownership Status:** High loan amounts are linked to defaults when the home ownership status is unspecified.
- **Loan Purpose:** Defaults occur more frequently with high loan amounts for purposes such as small business and housing.
- **Verification Status:** High loan amounts are associated with defaults when the verification status is verified.
- **Loan Grade:** Defaults are common among high loan amounts with Grade G.
- **Sub Grade:** High loan amounts are tied to defaults when the Sub Grade is G4.
- **Interest Rates:** Defaults are prevalent for high loan amounts with interest rates between 21% and 24%.
- **Geographic Location:** High loan amounts lead to defaults more frequently in Wyoming.

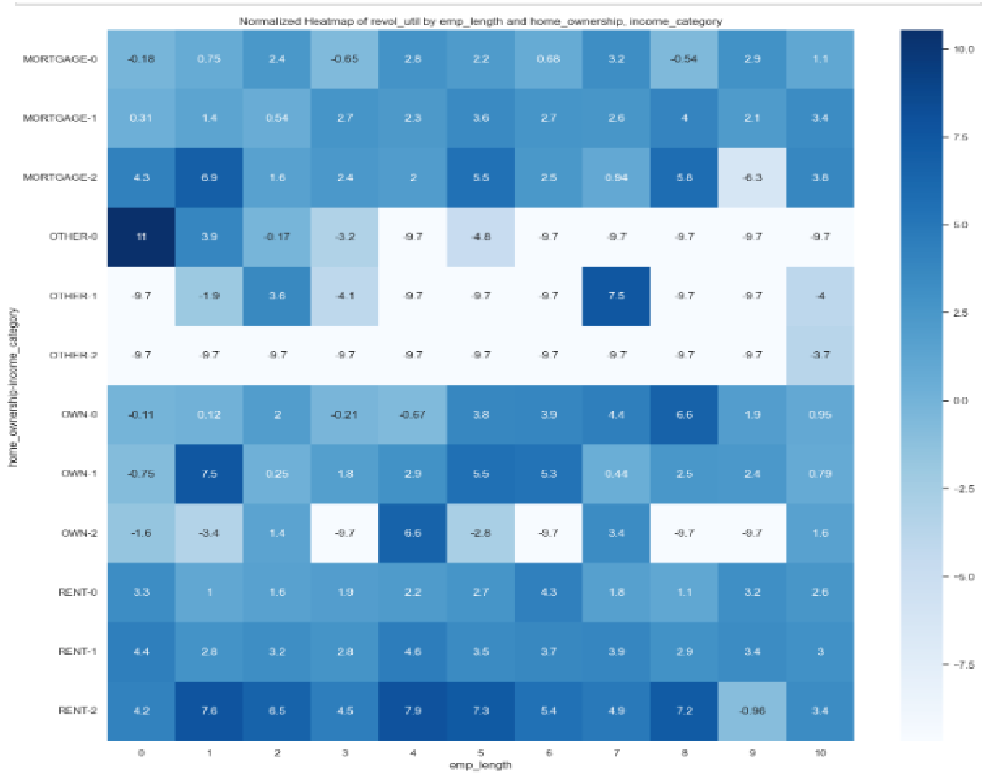
Understanding the Demographic of the people whose loans were Charged Off

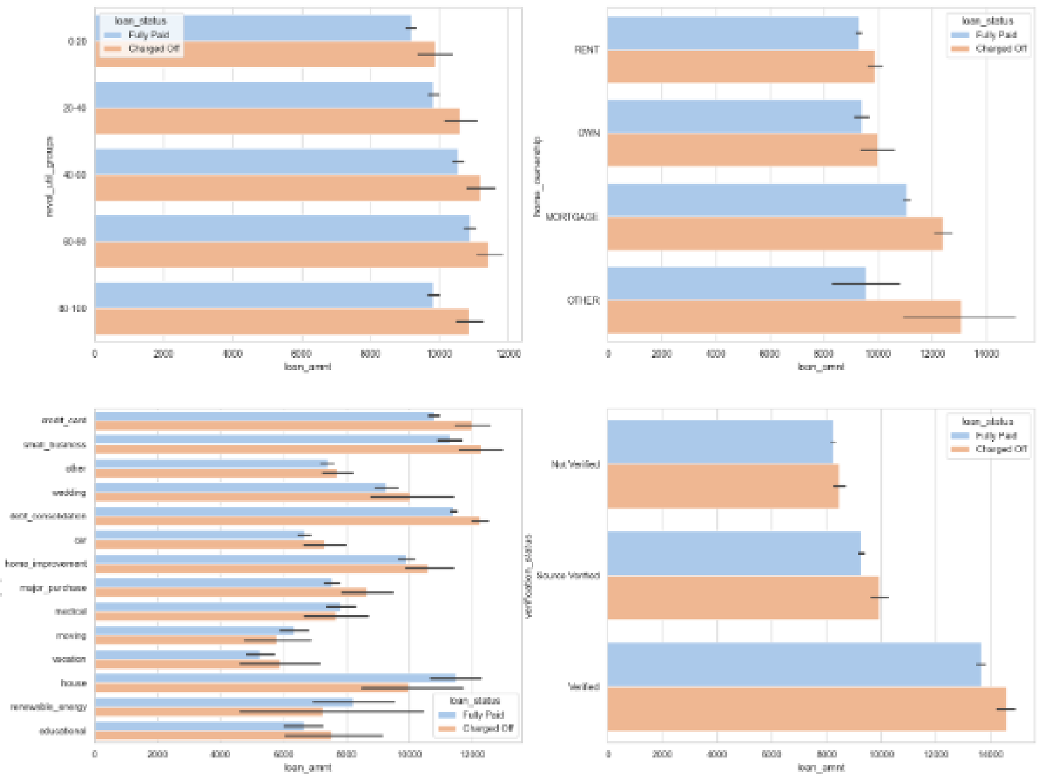
- **Variations in Loan Amounts:**
  - The largest charged-off loans were given to individuals in the High-Income category who took loans of Grade F for 'Major Purchases' or 'Other' purposes.
  - The Loan Amounts were most likely smaller for the purpose of 'Renewable Energy' across all grades and income categories.
- **Variations in Annual Income:**
  - The highest annual incomes among charged-off loan recipients were associated with those who had mortgages, took Grade F loans, and had 8-9 years of employment.
  - The Annual incomes were most likely smallest for those whose home ownership status was 'Other'.
- **Variations in Interest Rates:**
  - The highest interest rates among charged-off loan recipients were most associated with Grade F and G loans, spanning all income categories, and were linked to various loan purposes. Exceptions in purpose include moving, vacation, education, and renewable energy, where the interest rates were lowest for Grades F and G.
- **Variations in DTI ratios:**
  - DTI ratios vary widely among individuals with charged-off loans. The highest DTI ratio was observed for borrowers with 2 years of employment, a high-income category, and loans taken for major purchases.
- **Variations in Revol_Util:**
  - The Revolving Utilization Rate, a key indicator, highlights potential misuse of the Lending Club platform. High rates are linked to low-income individuals with less than one year of employment and 'Other' home ownership. In contrast, those with longer employment and varied home ownership show lower utilization rates.

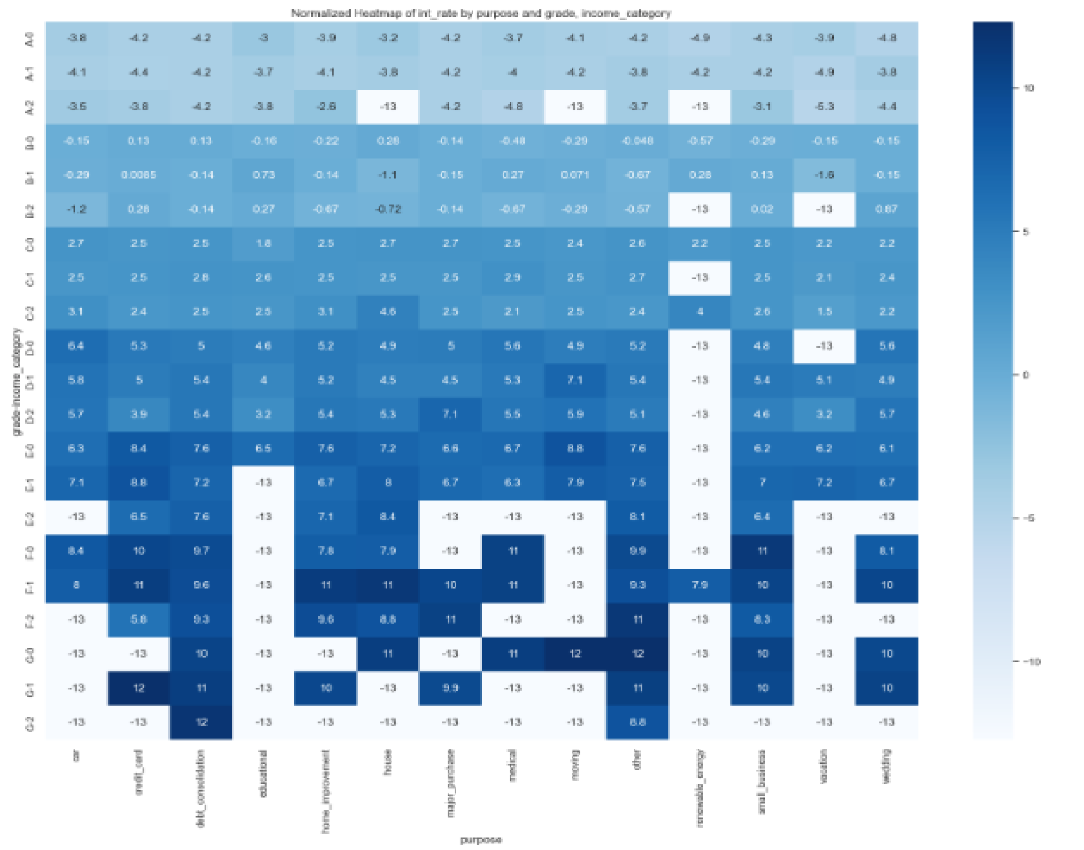
OUTCOME:
Based on the Multivariate Analysis, we can conclude that for Charged Off Loans:

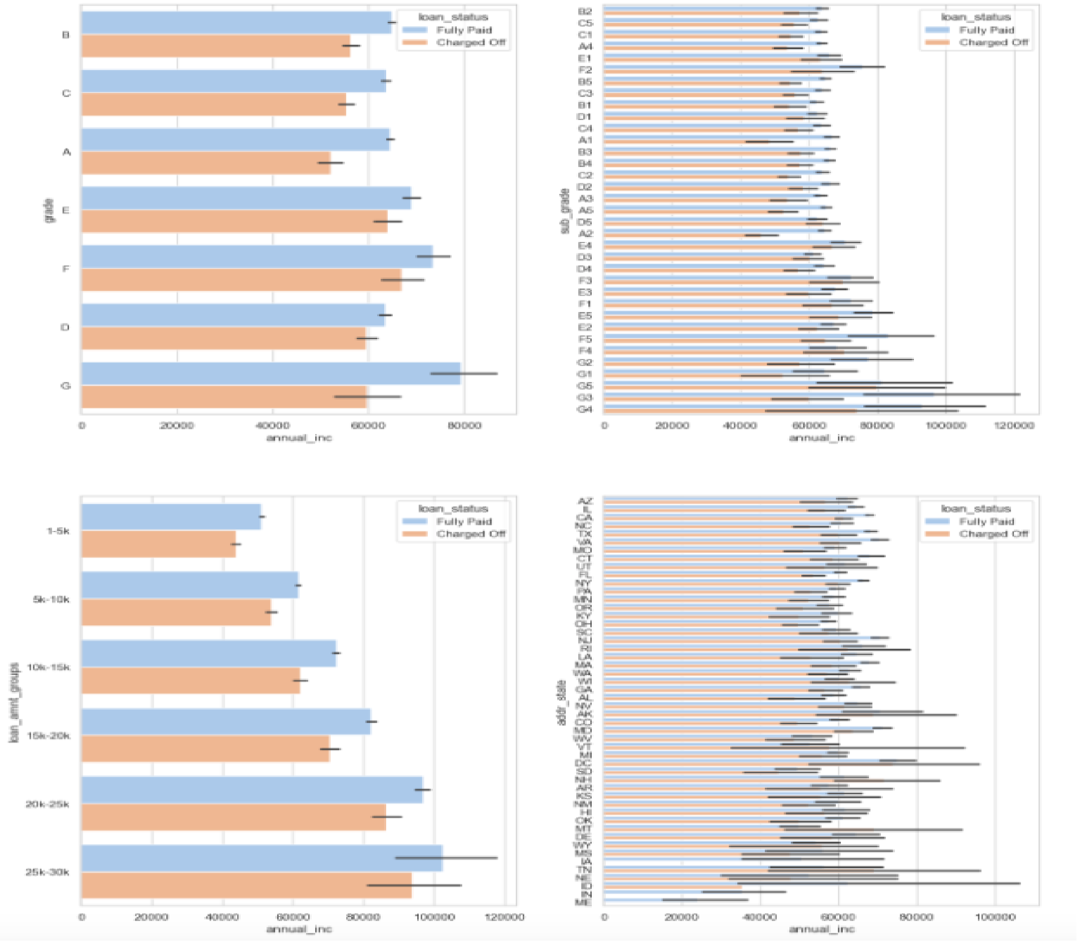
•**Revolving Utilization Rate is the most important indicator of Default behaviour.**
•The other Key Indicators include, Loan Amount, Interest Rate, Annual Income, DTI, Term, Grade, Purpose, Employment Length, Income Category, Verification Status, Home Ownership, Sub Grade and Address State

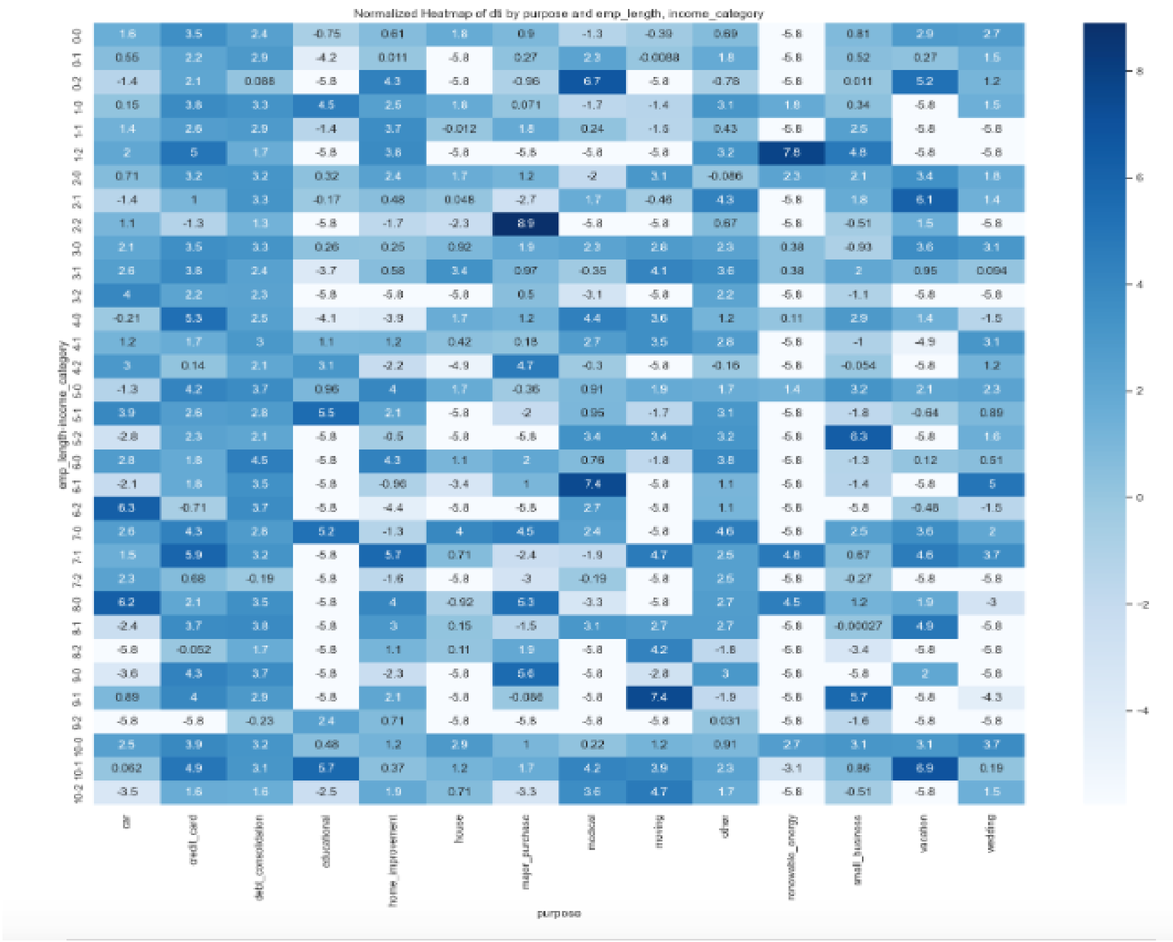
Multivariate Analyses

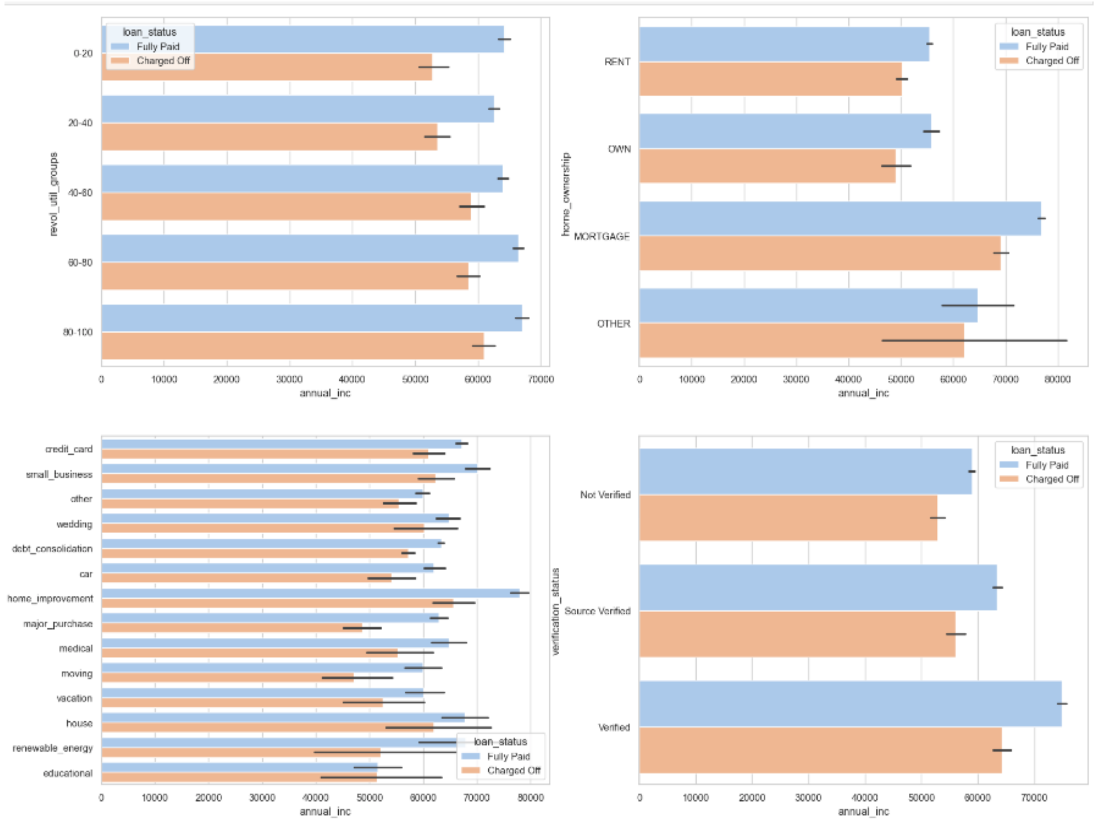












Conclusions

- **Quantitative Indicators:**
  - **Loan Amount:** Higher loan amounts are associated with increased default rates. This trend is evident in both the general dataset and specifically among defaulted loans.
  - **Interest Rate:** Higher interest rates are moderately associated with higher default rates, especially in defaulted loans.
  - **Annual Income:** Higher annual incomes are weakly correlated with higher loan amounts and slightly higher default rates.
  - **Debt-to-Income Ratio (DTI):** A higher DTI ratio is weakly associated with defaulted loans, indicating that higher debt burdens may contribute to defaults.
  - **Revolving Utilization Rate:** Higher revolving utilization rates are moderately associated with defaults, suggesting that higher credit usage may be a risk factor.

- **Qualitative Indicators:**
  - **Term:** Loans with a term of 36 months are more common among both defaulters and fully paid borrowers, but defaulters have a higher proportion of 60-month terms.
  - **Grade:** Lower loan grades (e.g., B, C, D) are more frequent among defaulted loans compared to fully paid loans.
  - **Purpose:** Loans taken for debt consolidation show a higher likelihood of default compared to other purposes.
  - **Income Category:** Defaulters are predominantly in the Low-income category, while fully paid loans are more common among Medium-income borrowers.
  - **Home Ownership:** Renters are slightly more likely to default compared to homeowners.

- **Bivariate Relationships:**
  - **Loan Amount vs. Other Attributes:** Higher loan amounts are linked with higher interest rates and annual incomes but show weak correlations with DTI and revolving utilization rates.
  - **Interest Rate vs. Other Attributes:** Interest rates have weak to moderate correlations with loan amount and revolving utilization rates but minimal correlation with annual income and DTI.
  - **Annual Income vs. Other Attributes:** Annual income is weakly associated with higher loan amounts and revolving utilization rates, showing minimal impact on DTI.

- **Multivariate Relationships:**
  - **Loan Amount, Interest Rate, and Default Risk:** The interaction between loan amount and interest rate significantly impacts default risk. Higher loan amounts combined with higher interest rates amplify the probability of default.
  - **Income, DTI, and Default Risk:** For borrowers with low income and high DTI ratios, the likelihood of default increases. This multivariate effect suggests that both financial stress (high DTI) and limited income are critical in predicting default.
  - **Loan Purpose and Loan Grade Interaction:** The purpose of the loan combined with its grade creates a nuanced risk profile. For instance, loans for debt consolidation with lower grades show a higher propensity for default compared to other loan purposes and grades.
  - **Home Ownership and Income Category:** The combination of home ownership status and income category influences default risk. Renters in lower income brackets are more likely to default, whereas homeowners with higher income levels show a lower risk.

- Overall, the analysis highlights that larger loan amounts, higher interest rates, and higher revolving credit utilization are significant indicators of loan defaults. Multivariate interactions, such as the combination of loan amount and interest rate or income and DTI, provide deeper insights into the complex nature of default risk. These findings can guide lenders in refining their risk assessment models and improving lending practices.

Recommendations

Loan Amounts

1. **Observation:** States with higher average loan amounts, like Arkansas and the District of Columbia, show higher default rates. For example, Arkansas shows a significant increase in default rates with higher loan amounts.
2. **Recommendation:** Implement stricter lending criteria or enhanced risk assessment protocols in states with higher average loan amounts to mitigate default risks.

Interest Rates

1. **Observation:** A moderate positive correlation (0.291) exists between loan amounts and interest rates for defaulted loans. Higher loan amounts are often associated with higher interest rates for defaulters.
2. **Recommendation:** Review and adjust interest rate settings to align with the risk profiles of borrowers, especially for larger loan amounts.

Annual Income

1. **Observation:** Defaulting borrowers have slightly lower average annual incomes ($57,000) compared to those who fully repaid their loans ($65,000). Higher annual incomes are also associated with higher loan amounts in defaults (correlation of 0.391).
2. **Recommendation:** Enhance income verification processes and consider introducing income-based lending limits to lower default risk.

Debt-to-Income Ratio (DTI)

1. **Observation:** Defaults are associated with higher average DTI ratios (14%) compared to fully paid loans (13%). Higher DTI ratios are a significant factor in loan defaults.
2. **Recommendation:** Lower acceptable DTI ratios for loan approvals or implement additional checks for borrowers with higher DTI ratios.

Revolving Utilization Rate

1. **Observation:** Defaulted loans have higher average revolving utilization rates (56%) compared to fully paid loans (47%). There is also a moderate positive correlation (0.375) between revolving utilization and interest rates.
2. **Recommendation:** Monitor borrowers' revolving credit utilization more closely, and adjust credit limits or terms to address high utilization rates.

Qualitative Indicators

1. **Observation:** Factors such as loan term (36 months), lower loan grades (B, C), and purposes like debt consolidation are more common among defaulters. For instance, 48% of defaulters cited debt consolidation as the loan purpose.
2. **Recommendation:** Re-evaluate loan terms, grades, and purposes to reduce default risk. Additional scrutiny may be needed for certain loan categories.

Regional Variations

1. **Observation:** Regional trends show varied loan amounts and default rates, with noticeable differences between the West, Southeast, and Midwest regions. A significant increase in total loan amounts occurred post-2008 financial crisis.
2. **Recommendation:** Tailor regional lending strategies and risk management approaches based on local economic conditions and historical default trends.

Verification Status

1. **Observation:** Verification status has minimal impact on default rates, with verified and non-verified borrowers showing similar patterns.
2. **Recommendation:** Strengthen verification processes or explore alternative methods for assessing borrower reliability beyond verification status.

Sunil Bhairi: https://github.com/SunilBhairi/

Venkat Lata: https://github.com/svenkatlata

Thank You

for your patience...