Lending Club Case Study

Submitted By : K Lalithabai Sobha
Ashwani Singh

Submission Date: 23-10-2024

Table of Contents

Introduction

- ## Overview

  This case study is to do basic risk analytics in financial services and understand how data is used to minimize the risk of losing money while lending loans to customers by applying EDA techniques.

- ## Purpose

  The Purpose of this document is to describe the various steps followed in Exploratory Data Analysis in detail and to provide a recommendations to the Lending Company in order to reduce the risks while issuing a new loan to the customers.

- ## Scope

  The scope of this case study is to conduct Exploratory Data Analysis, starting from Data Cleaning, Data Correction, Data Classification and Derived Data Metrics to conduct Univariate Analysis, Segmented Univariate Analysis, Bi Variate Analysis and Multi Variate Analysis to observe, infer and finally derive a conclusion to the given problem statement.

Problem Statement

Problem

The **Consumer Finance Company** which specializes in lending various types of loans to urban customers, is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss).

Two types of risks are associated with the company's decision:

- If the applicant is ***likely to repay*** the loan, then **not approving** the loan results in **a loss of business** to the company
- If the applicant is ***not likely to repay*** the loan, i.e. he/she is likely to default, then **approving** the loan may lead to a **financial loss** for the company.

Business Objectives

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Aim

To identify the patterns in terms of consumer attributes and loan attributes which indicate a loan applicant is likely to default or not by using EDA

Data Understanding

Data Source

Loan dataset in csv(Loan.csv) format which contains the complete loan data for all loans issued through the time period 2007 t0 2011 are provided along with a data dictionary file.

Data Summary

Loan.csv file contains **39717** rows and **111** columns/attributes.

There are two types of attributes Loan Attribute and Customer attributes.

Target Variable

The **Target** Variable in the Loan data set is ‘**Loan Status’** attribute which contains three distinct values as ‘Fully Paid’, ‘Charged Off’ and ‘Current’. For the analysis purpose we are considering only the records having ‘Fully Paid’ or ‘Charged Off’ loan status.

Data Cleaning

- No header/footer/summary rows found
- No Duplicate rows found
- There were 111 columns/attributes present in the dataset.
- There were 54 attributes which contains only null values. Hence dropped them.
- There were 9 attributes which are filled with only single value. They cannot be used for analysis. Hence dropped them.
- 'id','member_id','emp_title','url','desc','title’ are of no use in analysis. Hence dropped them.
- 'zip_code’ and ‘addr_state’, both will lead to same analysis. Hence dropped 'zip_code’ and kept ‘addr_state’ for analysis.
- Investor related attributes are dropped as they are of much use in the analysis
- Some of the column entries are entered once the loan is active. Hence at the application stage analysis, these are not useful. Hence dropped them
- ‘mths_since_last_record’ and ‘mths_since_last_delinq’ attributes are dropped as they are having more than 50% of null values
- Total 87 columns/attributes were removed from the dataset.
- After all the Data cleaning process we are left with 39717 rows and 24 columns.

Data Imputing & Data Type Correction

- There are 50 records of 'revol_util' attribute with null value. As most of the records are filled with 0%, replaced the null values with the mode value of 0%
- There are 697 records of ‘pub_rec_bankruptcies' attribute with null value. As 98% of the ‘pub_rec_bankruptcies’ records are filled with 0, replaced the null values with the mode value of 0
- There are 1075 records of 'emp_length' attribute with null value. As it is a categorical variable, replaced the null values with the mode value with mode value.
- Replaced the % sign from ‘int_rate’ and 'revol_util’ values and converted them into float type
- Removed +,<,’years’ from ‘emp_length’ values and converted it into int type with values ranging from 0 to 10.
- Removed ’months’ from ‘term’ attribute values and converted it into int type with values as 36 and 60
- Column ‘loan_amnt’ and ‘funded_amnt’ converted to float.
- All the floating column values are rounded to two decimals.
- Replaced the ‘NONE’ value to ‘OTHER’ in ‘home_ownership’ attribute
- Replaced 'Source Verified’ as 'Verified’ in ‘verification_status’ attribute
- Converted ‘issue_d’ from object type to ‘datetime’
- Filtered the data set by removing the records containing Loan-status as ‘Current’ .
- After all the Data cleaning process we are left with 38577 rows and 24 columns.

Derived Metrics

- Derived issue_Year, issue_Month and issue_Qtr coulmns from issue_d attribute to save the Year, Month and Quarter values for the analysis purpose and then dropped the column issue_d.
- Derived the following bucket columns from the corresponding quantitative/numerical attributes with appropriate bucket cuts.
  - loan_amnt → loan_amnt_bucket ('0 - 5K', '5K - 10K', '10K - 15K', '15K - above’)
  - funded_amnt → funded_amnt _bucket ('0 - 5K', '5K - 10K', '10K - 15K', '15K - above’)
  - int_rate → int_rate_bucket ('Below 9%', '9%-11%','11%-13%','13%-15%','Above 15%’)
  - installment → installment_bucket('0 - 200', '200 - 400', '400 - 600','600 - 800', 'Above 800 ‘)
  - annual_inc → annual_inc_bucket('0 - 40k', '40k - 50k', '50k - 60k', '60k - 70k', '70k - 80k', 'Above 80k’)
  - dti → dti_bucket('Very Low', 'Low', 'Medium', 'High', 'Very High’)
  - revol_util → revol_util_bucket('Below 25%', '25%-50%','50%-75%','Above 75%’)
  - total_acc → total_acc_bucket('Below 13','13-18','18-23','23-28','Above 28’)
  - revol_bal → revol_bal_bucket('Very Low','Low','Medium','High','Very High’)
  - open_acc → open_acc_bucket('Below 6','6 - 8','8 - 10','10 - 12','Above 12')

Data Categorization

The final cleaned up data can be divided into the following groups for visualization and analysis

1. Ordered categorical data

    1. grade 2. sub_grade 3. term 4. emp_length 5. issue_Year 6. issue_Month 7. issue_Qtr 8. pub_rec_bankruptcies 9. pub_rec 10. inq_last_6mths 11. delinq_2yrs

2. Unordered categorical data

    1. addr_state 2. purpose 3. home_ownership 4. verification_status

3. Numerical data

    1. loan_amnt 2. funded_amnt 3. int_rate 4. installment

    5. annual_inc 6. dti 7. open_acc 8. revol_util 9. revol_bal 10. total_acc

4. Numerical Derived data

    1. loan_amnt_bucket 2. funded_amnt_bucket 3. int_rate_bucket

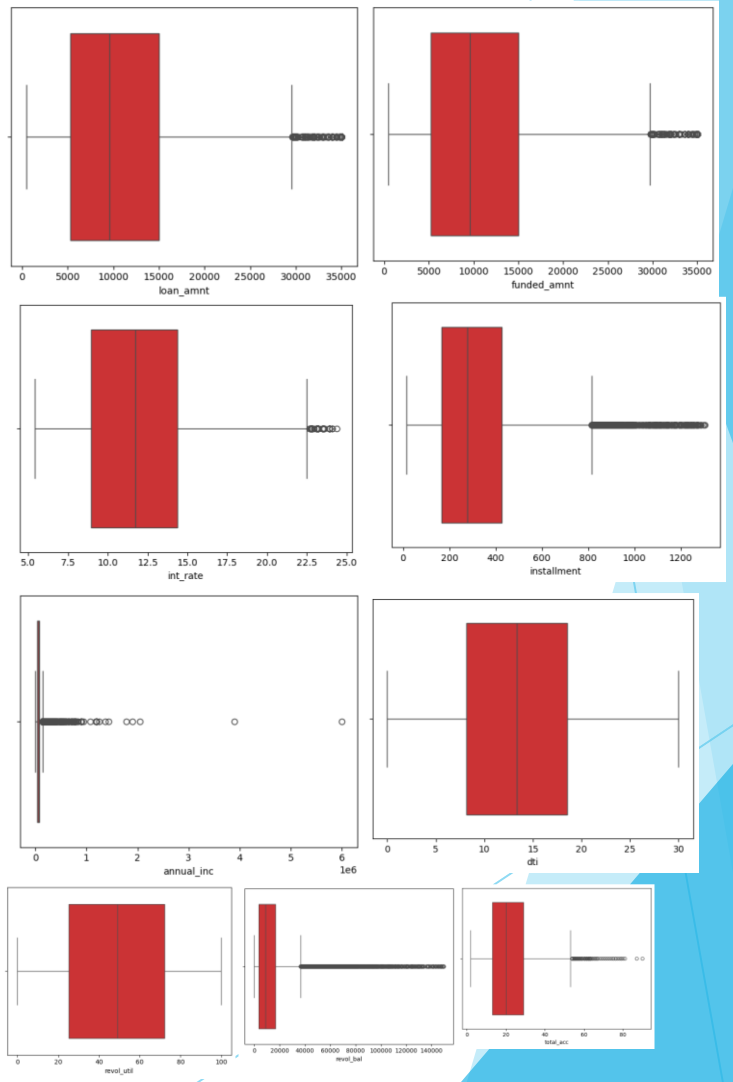
    4. installment_bucket 5. annual_inc_bucket 6. dti_bucket 7. open_acc_bucket

    8. revol_util_bucket 9. revol_bal_bucket 10. total_acc_bucket

Data Outlier Analysis

Outlier Observation

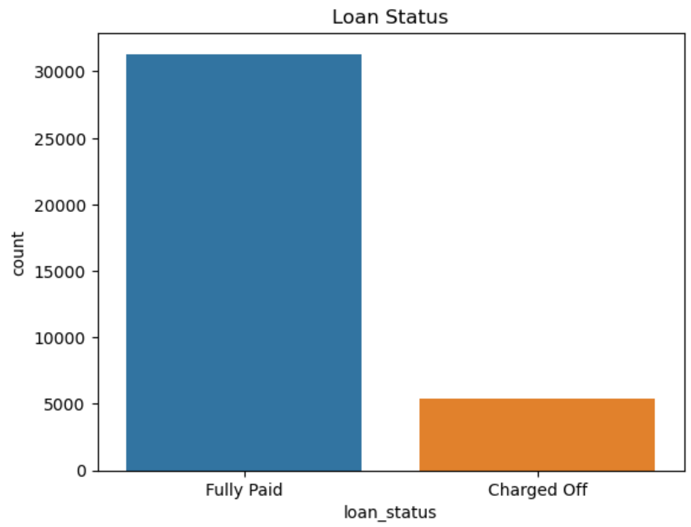
- **loan_amnt** - Outliers are present, but seems to be continuous. Hence No need to remove them
- **funded_amnt** - Outliers are present, but seems to be continuous. Hence No need to remove them
- **int_rate** - Outliers are present, but seems to be continuous. Hence No need to remove them
- **installment** - Outliers are present, but seems to be continuous. Hence No need to remove them
- **annual_inc** - Outliers present. Removed the outliers
- **dti** - No Outliers
- **open_acc** - Outliers are present, but seems to be minor. Hence No need to remove them
- **revol_util** - No Outliers
- **total_acc** - Outliers are present, but seems to be continuous. Hence No need to remove them

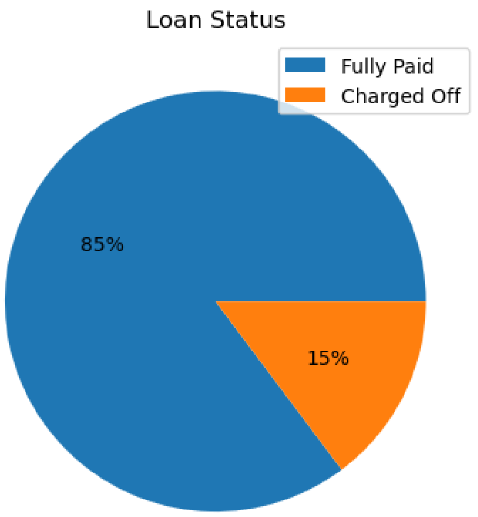


Univariate Analysis

Loan Status

Among the total loan records between fully paid and charged-off, around 15% of loan applicants are likely to be defaulters



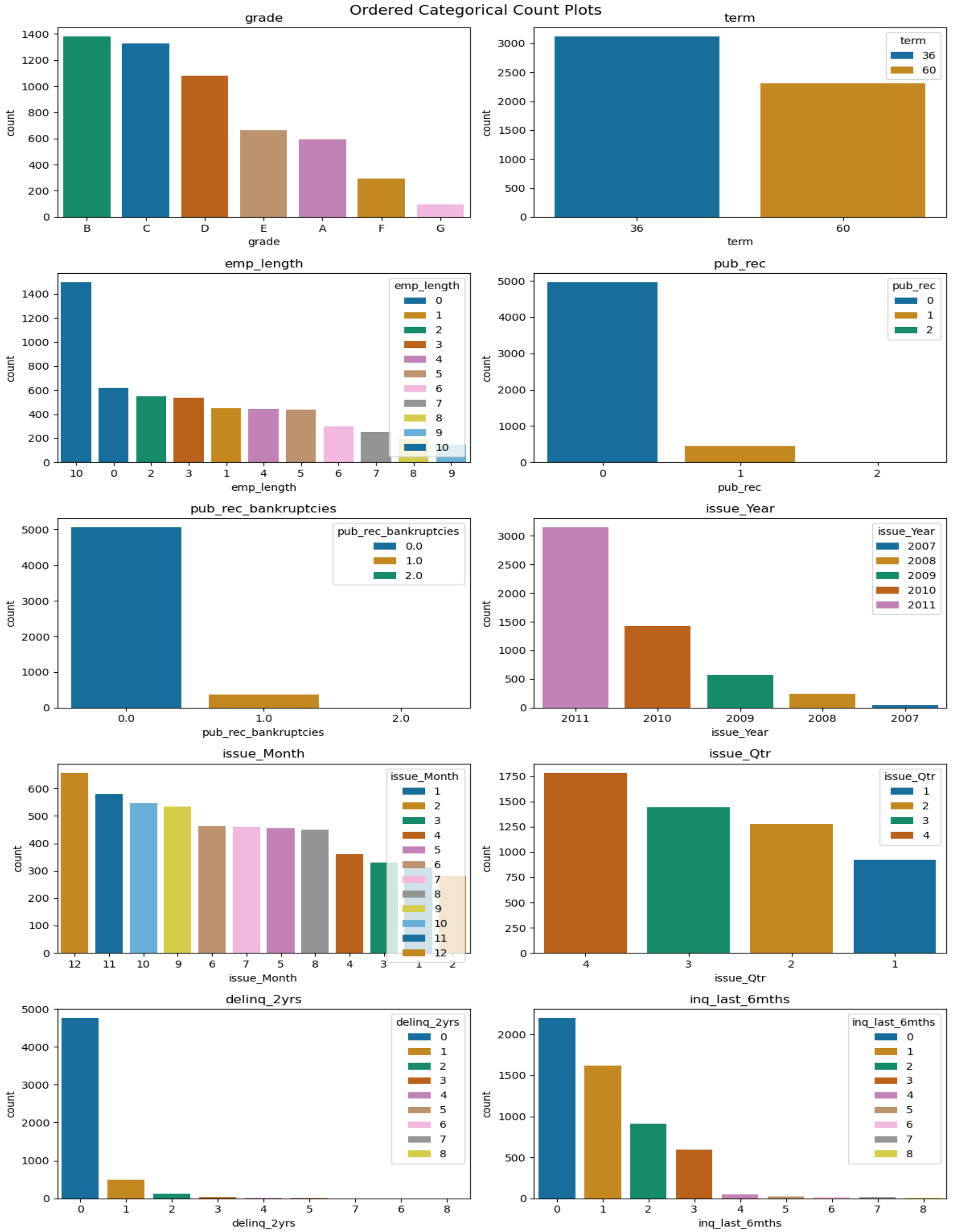


Univariate Analysis

Ordered Categorical Data

Observations

- **Grade** - B & C had the highest number of "Charged off" loan applicants, indicating that applicants with these credit grade faced challenges in repaying their loans.
- **Term** - Short-term loans with a duration of 36 months having more number of defaulters(57%). This suggests that a significant portion of applicants who experienced loan default chose shorter repayment terms.
- **Employment Length-** Applicants who had been employed for more than 10 years accounted for the highest number of "Charged off" loans(28%). This indicates that long-term employment history did not necessarily guarantee successful loan repayment.
- **Loan Year** - The year 2011 recorded the highest number of "Charged off" loan applications. This could be indicative of economic or financial recession during that year.
- **Loan Taken Month and Quarter** - '"Charged off" loans were predominantly taken during the 4th quarter, primarily in December. This peak in loan applications during the holiday season might suggest that financial pressures during the holidays contributed to loan defaults.
- **Derogatory Public Records-** There is a significant number of defaulters(92%) do not have any derogatory public record. Having no derogatory record doen't indicate a non-defaulter.
- **Public Record Bankruptcie** - There is a significant number of defaulters do not have any bankruptcy record. Having no bank ruptcy record is not an indication of a non-defaulter.
- **Past-due incidences of delinquency -** There is a significant number of defaulters do not have any Past-due incidences of delinquency for 2 years. Having no delinquency record doen't indicate a non-defaulter.
- **number of inquiries in past 6 months-** There is a significant number of defaulters do not have made any inquiries in past 6 months . Having not made any inquiry doen't indicate non-defaulter.



Univariate Analysis
Un-Ordered Categorical Data

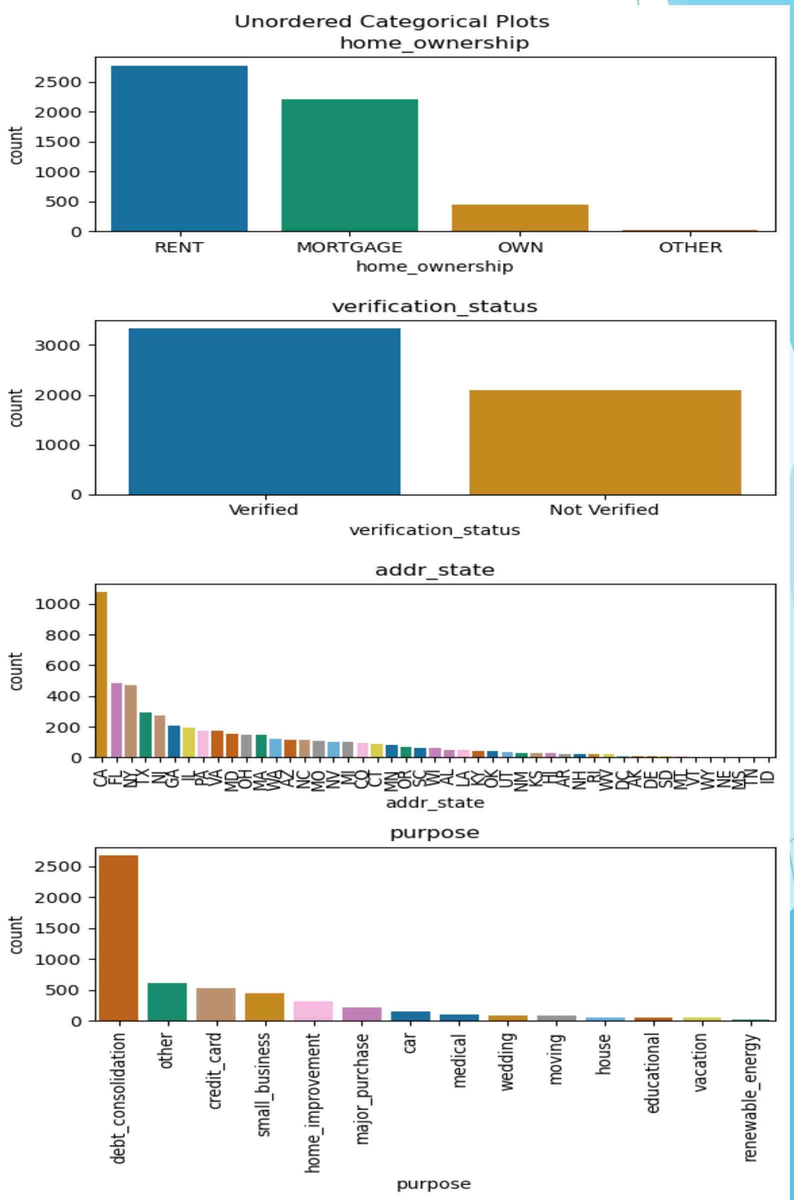
Observations

Address State - The highest number of "Charged off" loan applicants are from the state California. Hence more precautions need to be taken while assessing the loan applications.

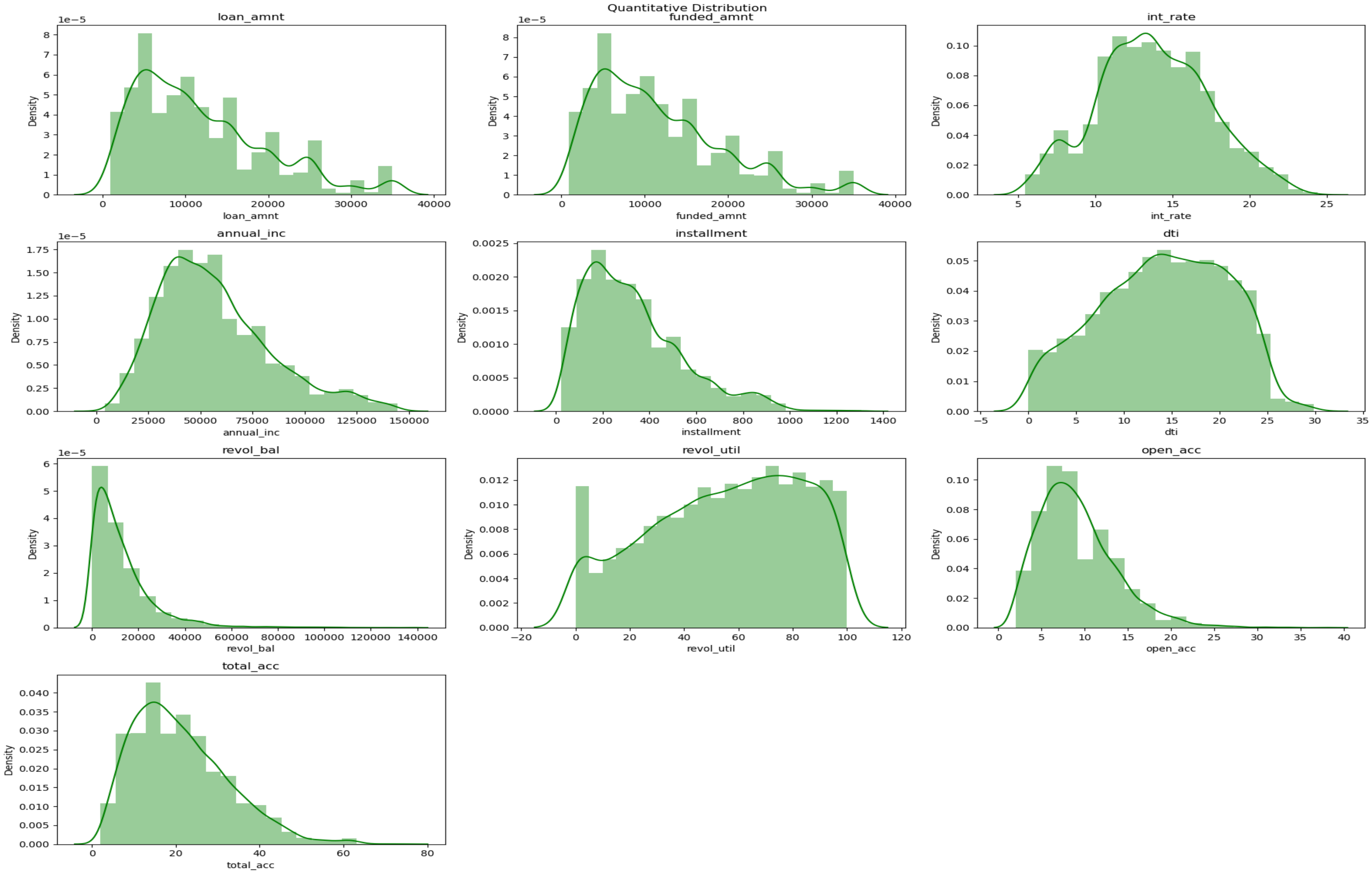
Loan Purpose - Debt consolidation was the primary loan purpose for most of the "Charged off" loan applicants. Hence needs to exercise more caution while approving loans for debt consolidation purposes.

Home Ownership - The majority of "Charged off" loan applicants were living in rented houses. Hence needs to exercise more caution in assessing the financial condition of applicants living in rented houses.

Verifiation Staus - For the majority of "Charged off" loan applicant's income were verified before issuing the loan. This indicates the verification of the income alone is not sufficient. Hence the lending company should exercise more caution while analysing the financial stability of the applicants.



Univariate Analysis
Numerical Data



Univariate Analysis
Numerical Data

Observations

Annual Income - Most of the defaulters are having an annual income of 25000 - 75000 USD. For higher income group, loan defaulting is less.

Loan Amount - When loan amount is high, charged-off density is low. This implies most of the loan defaulters are having relatively small loan amount

Interest Rate - Loan Defaulting is more when interest rate increases.

dti - The loan defaulting is more when debt to income is between 10 and 20.

Installment Amount - Charged-Off is more for lesser amount of installment. May be due to poor financial stability

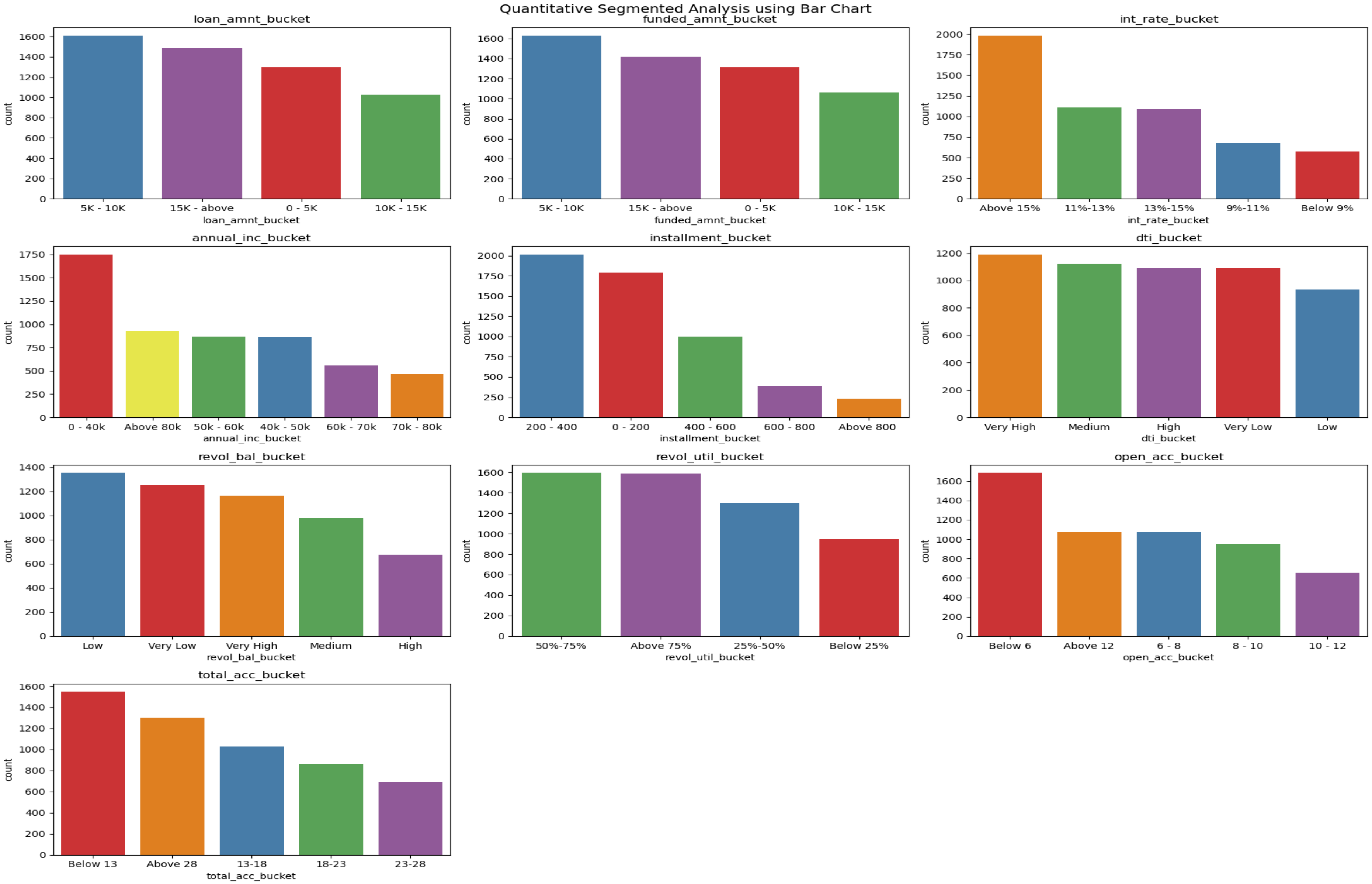
Revolving Balance - Charged off density is very high when revolving Balance is very low

Revolving Line Utilization - Charged off density is almost uniform across revolving line Utilization.

Open Credit Lines - Charged off density is high when number of open credit lines are in the range of 5-15

Total Credit Lines - Charged off density is high when number of total credit lines are in the range of 15-25

Univariate Analysis
Numerical Segmented Data



Univariate Analysis
Numerical Segmented Data

Observations

Annual Income - Most of the defaulters are having an annual income of less than 40000 USD. The lending company should exercise caution when lending to individuals with low annual salaries. They should implement rigorous income verification and assess repayment capacity more thoroughly for applicants in this income bracket.

Loan Amount - Most of the defaulters took a loan amount in the range of 5K-10K. Still a considerable defaulters are there for loan amount is above 15K. Hence extra care is needed when lending higher loan amount.

Interest Rate - Loan Defaulting is more when interest rate is above 15%. When interest rate is less, defaulting is also less. Hence lending company may consider giving loans at a lower interest rate.

Debt to Income Ratio (dti) - The loan defaulting increase with increase in debt to income ratio.

Installment Amount - Charged-Off is more for lesser amount of installment ie below 400 USD. May be due to poor financial stability

Revolving Balance - Charged of is very high when revolving Balance is very low. At the same time there are considerable number of loan defaulters when revolving Balance is high. No remarkable pattern can be observed here. Hence this parameter is not much of a deciding factor for Loan approval

Revolving Line Utilization - Loan defaulting is less when revolving line Utilization is below 25%.

Open Credit Lines - Charged off density is high when number of open credit lines are Below 6

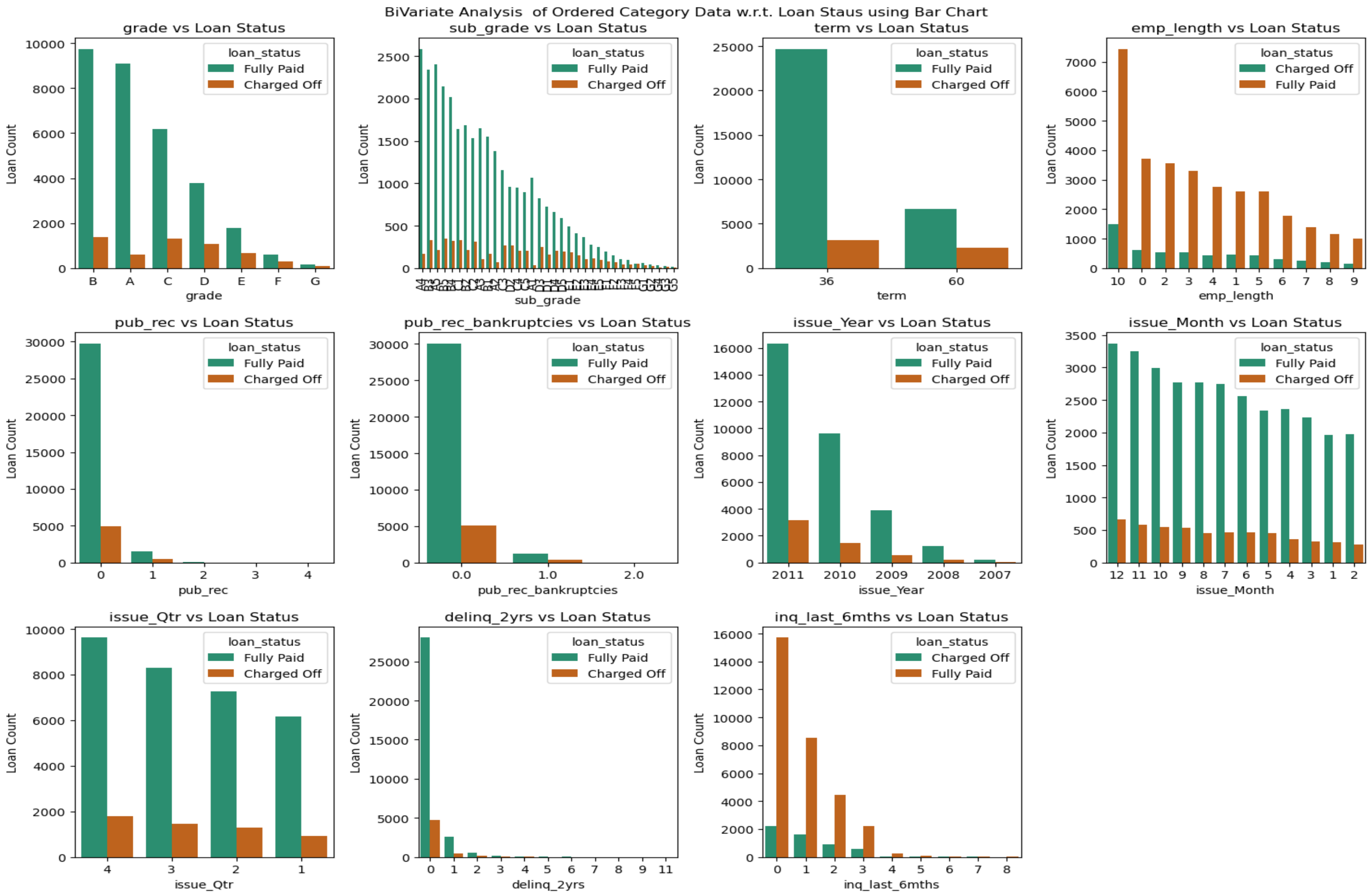
Total Credit Lines - Charged off density is high when number of total credit lines are Below 13. Hence care should be taken when there is not much of alternate credit lines are available for the applicant.

Bivariate Analysis

Approach

Categorical (both Ordered and Unordered) and Numerical data need to be analyzed as part of bivariate analysis against Loan Status (`loan_status`) attribute.

Bivariate Analysis of Ordered Categorical Data



Bivariate Analysis of Ordered Categorical Data

Observations

Grade - The loan applicants belonging to Grades B, C and D having the most number of defaulters

Sub Grade - Loan applicants belonging to Sub Grades B3, B5, and B4 are more likely to default

Term - Loan applicants applying loan for 60 months are more likely to default than applying for 36 months

Employment Length - Most of the loan applicants are having 10 or more years of experience. They also are the ones who are most likely to default

Pub-rec / Pub-rec-bankruptcies - Most of the loan applicants don't have any derogatory public records or bankruptcy records. The one having a public record are more likely to default.

Issue Year - The loan applicants have increased steadily from 2007 to 2011 showcasing positive trend in the upcoming years. The year 2011 have maximum number of loan applications and also the highest number of defaulters. Could be because of the economic recession experienced at the year

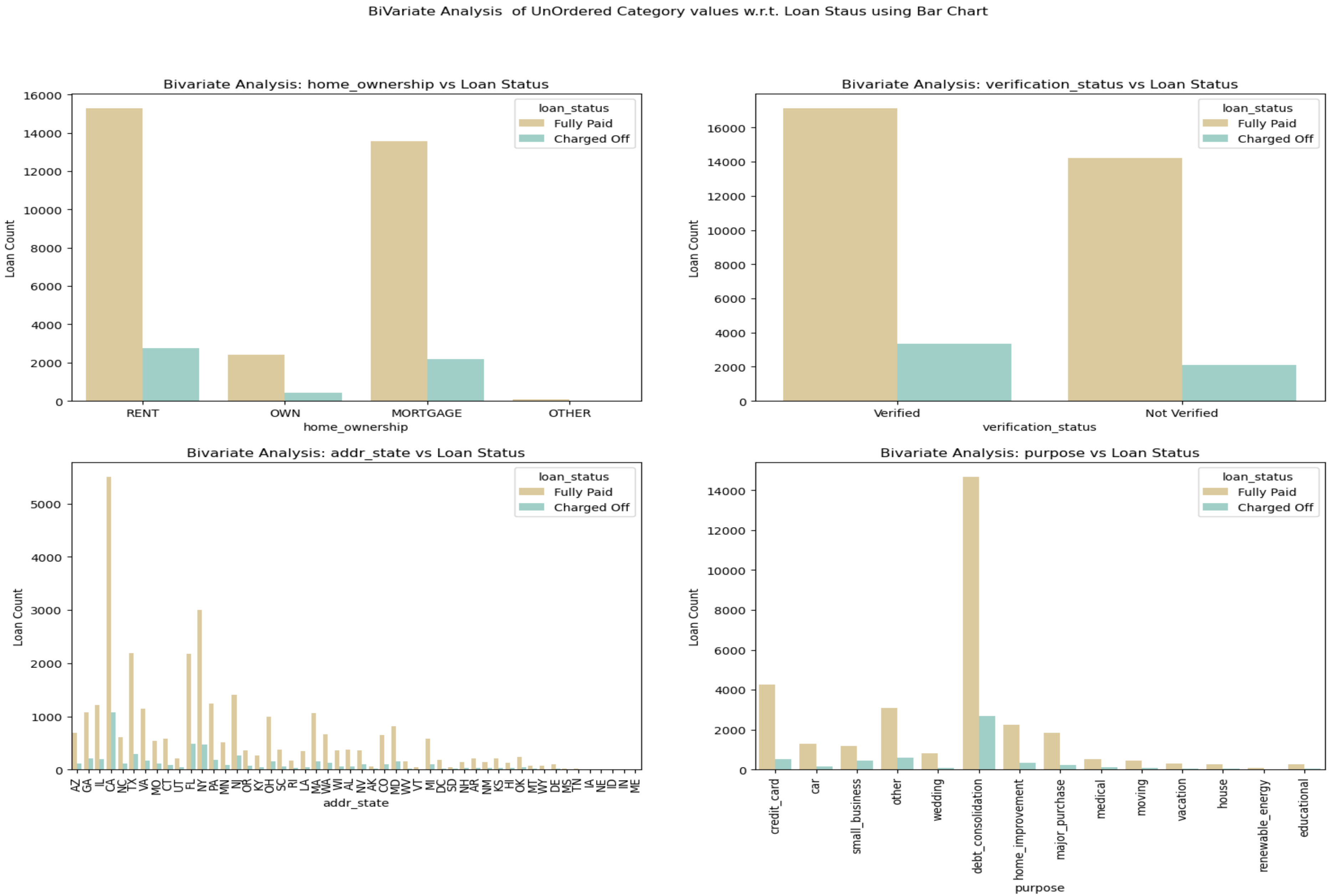
Issue Month - The month of December is the most preferred month of taking loans. This may be due to the holiday season.

Issue Quarter - Maximum number of loans are applied in the 4th Quarter. This is mainly due to the holiday season coming up

delinq_2yrs - Most of the loan applicants don't have any Past-due incidences of delinquency for 2 years. The one having a delinquency are more likely to default.

inq-last-6mths - Most of the loan applicants don't have made any inquiries in past 6 months. The one who have made inquiries are more likely to default.

Bivariate Analysis of UnOrdered Categorical Data



Bivariate Analysis of UnOrdered Categorical Data

Observations

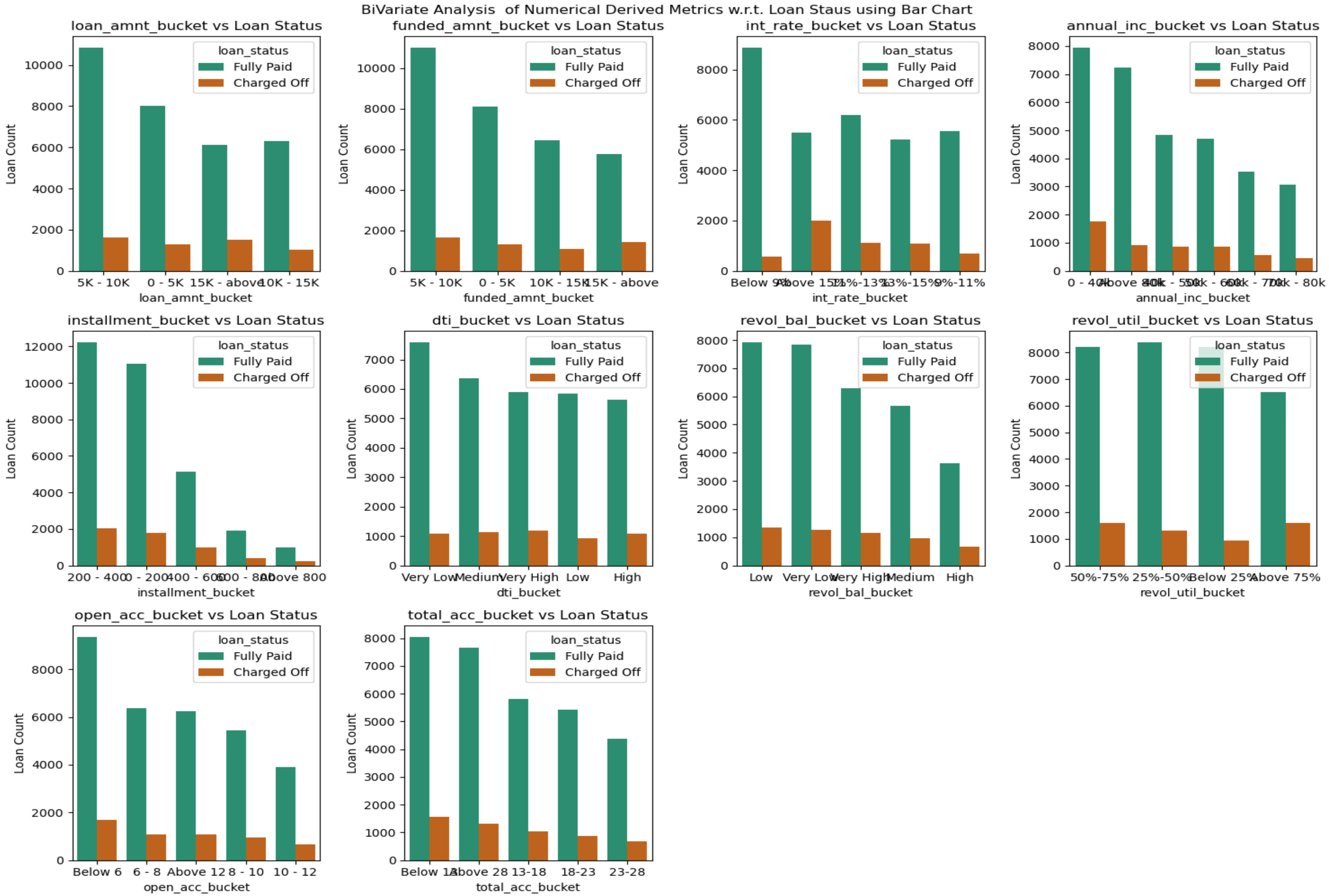
Address State - The States California, New York and Florida are having more number of Loan applicants and correspondingly more number of defaulters. Hence more precautions need to be taken while assessing the loan applications from these states.

Loan Purpose - Most of the loans are taken for the purpose of debt_consolidation and correspondingly the number of defaulters are also more Hence needs to exercise more caution while approving loans for debt consolidation purposes.

Home Ownership - The majority of loan applicants were living in rented houses and are having the highest count of defaulters. Hence needs to exercise more caution in assessing the financial condition of applicants living in rented houses.

Verification Status - For the majority of loan applicant's income were verified before issuing the loan. Still there are more number of applicants defaulted as compared to non verified applicants. This indicates the verification of the income alone is not sufficient. Hence the lending company should exercise more caution while analyzing the financial stability of the applicants.

Bivariate Analysis of Numerical Derived Metrics



Bivariate Analysis of Numerical Derived Metrics

Observations

Annual Income - Most of the loan applicants are having an annual income of less than 40000 USD and correspondingly the charged-off count among them are also the highest. At the same time, considerable number of loan applicants are these from high income group of above 80K and there are some defaulters at this income level too. The proportionate analysis is needed to make any conclusion.

Loan Amount - Most of the loan applicants took a loan amount in the low range of 5K-10K. Still considerable defaulters are there for this range of loan. At the same time, considerable number of loan applicants took high amount of loan of above 15K and there are a significant count of defaulters are there for this high loan amount too. Hence proportionate analysis is needed to make any conclusion.

Interest Rate - Most of the loans are disbursed with very low interest rate and correspondingly the number of defaulters are less compared to higher interest rate. It is also observed that the Loan Defaulting is more when interest rate is above 15%. Hence Lending Company should exercise extra care while issuing loans at higher interest rate.

debt to income ratio (dti) - The count of defaulters are more or less same across different levels of debt to income ratio, even though more number of loans are issued for low dti value. This means having a higher value of dti increases the chances of defaulting.

Installment Amount - Majority of loans are having low installment amount , still Charged-Off is more for lesser amount of installment ie below 400 USD. May be due to poor financial stability

Revolving Balance - Loans are usually issued for very Low to Medium revolving balance. Charged off is high when revoving Balance is very low. At the same time there are considerable number of loan defaulters when revolving Balance is high. Hence proportionate analysis is needed to make any conclusion.

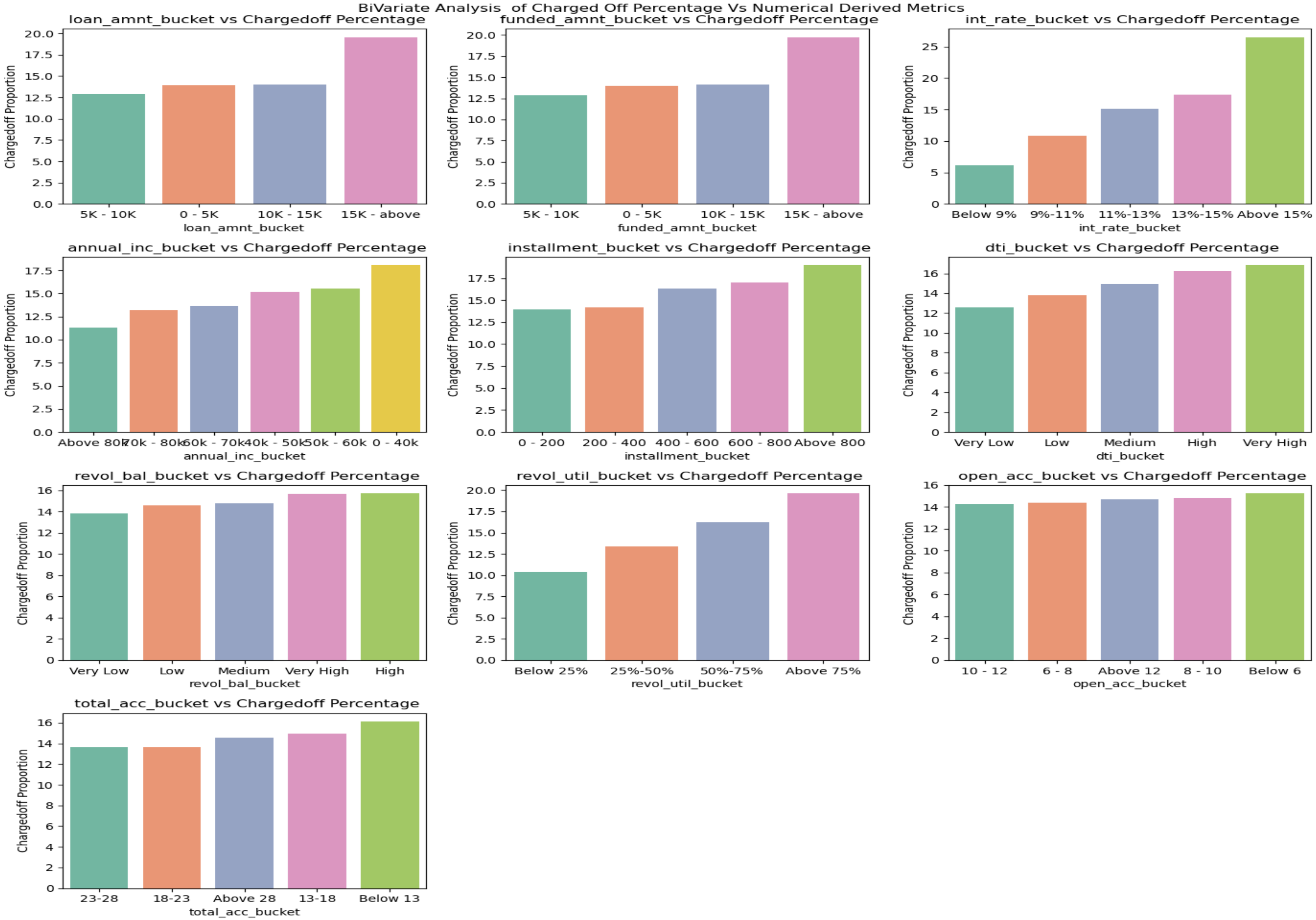
Revolving Line Utilization - When revolving utilization is more, the applicant is likely to default

Open Credit Lines - More number of loans are issued for low levels (Below 6) of open credit lines. But no of defaulters are more at this low open credit lines. proportionate analysis is needed to make any conclusion.

Total Credit Lines - More loans are issued for a very low total credit lines as well as very high credit lines. The defaulters in both the cases are also high as compared to other levels of total credit lines. Hence proportionate analysis is needed to make any conclusion.

BiVariate Analysis of Charged Off Proportion vs Various Categories

Bivariate Analysis of Derived Buckets vs Chargedoff_Proportion



BiVariate Analysis of Charged Off Proportion vs Various Categories

Bivariate Analysis of Derived Buckets vs Chargedoff_Proportion

Observations

Annual Income - When annual income increases, charged off percentage decreases.Hence the loan applicant whose annual income is Below 40k is more likely to default. Higher the annual income, the less chances of default.

Interest Rate - Charged Off Percentage increases as the interest rate increases. For interest rate above 15% has good chances of getting charged off as compared to other category interest rates.

Loan Amount - More likely to default when Loan Amount is High (ie Above 15K)

Funded Amount - More likely to default when Funded Amount is High (ie Above 15K)

Installment - More likely to default when installment is High (ie.Above 800)

debt to income ratio (dti) - When DTI increases, charged off percentage increases and vice-versa. Hence Debt to Income Ration should be as low as possible

Revolving Balance - When revol-bal is more there is a slightly more chances of charged-off

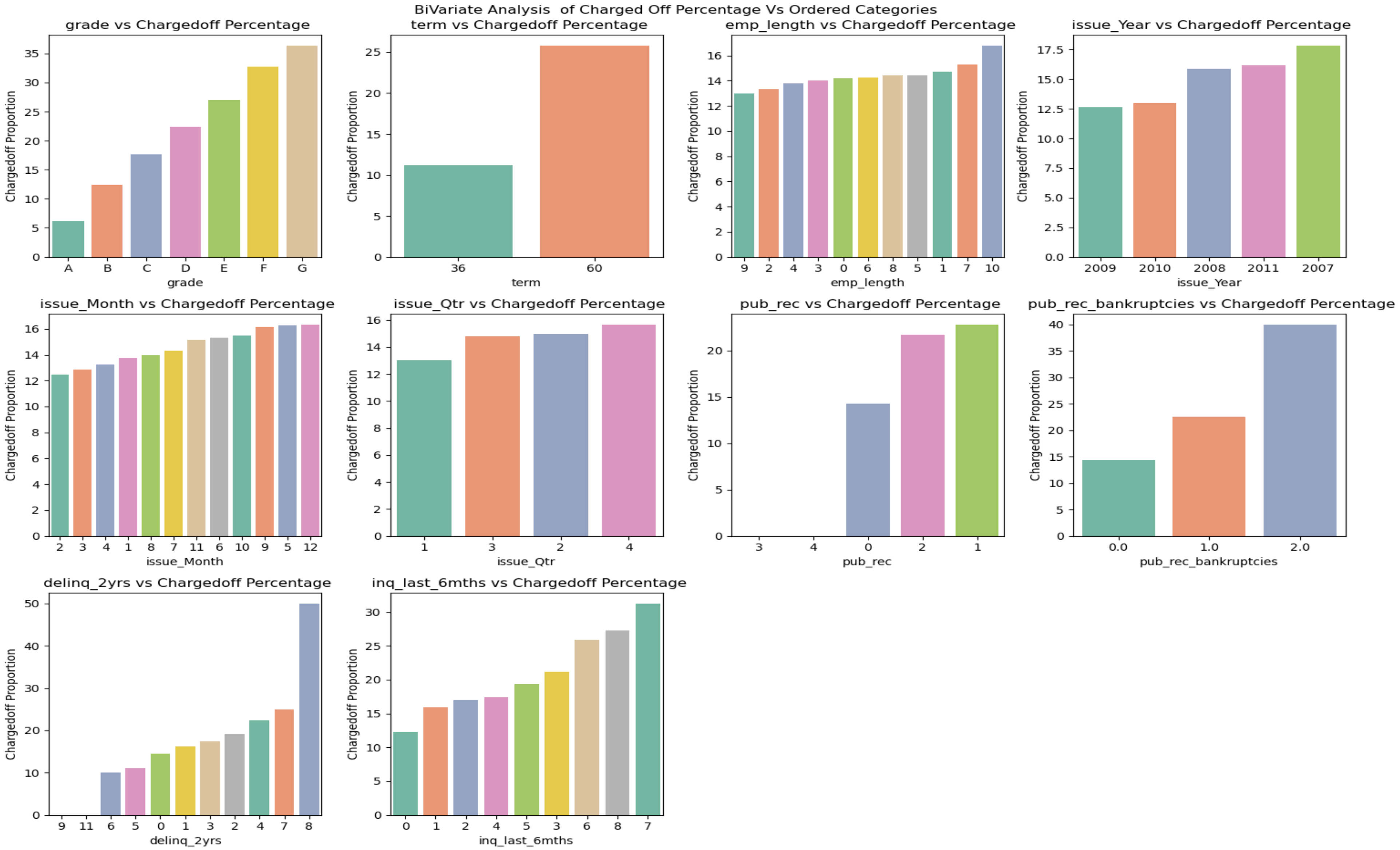
Revolving Line Utilization - When revolv-util percentage increases, chances of defaulting also increases. When revolv-util is below 25%, there is less chances default and when it is above 75%, there is a high chances of default

Open Credit Lines - There is no significant variation in default percentage for various categories of open-acc. Hence not a deciding factor

Total Credit Lines - There is no significant variation in default percentage for various categories of open-acc. Hence not a deciding factor

BiVariate Analysis of Charged Off Proportion vs Various Categories

Bivariate Analysis of Ordered Categories vs Chargedoff_Proportion



BiVariate Analysis of Charged Off Proportion vs Various Categories

Bivariate Analysis of Ordered Categories vs Chargedoff_Proportion

Observations

Loan applicants applying loan for **grade G** are more likely to default than the one taking loan in any other category and grades **A and B** are less likely to default

Loan applicants applying loan for **60 months** are more likely to default than the one taking loan for 36 months

Loan applicants applying loan having employment length **10+ years** are more likely to default than the one having lesser years

Charged-off percentage is high in 2007 and low in 2009

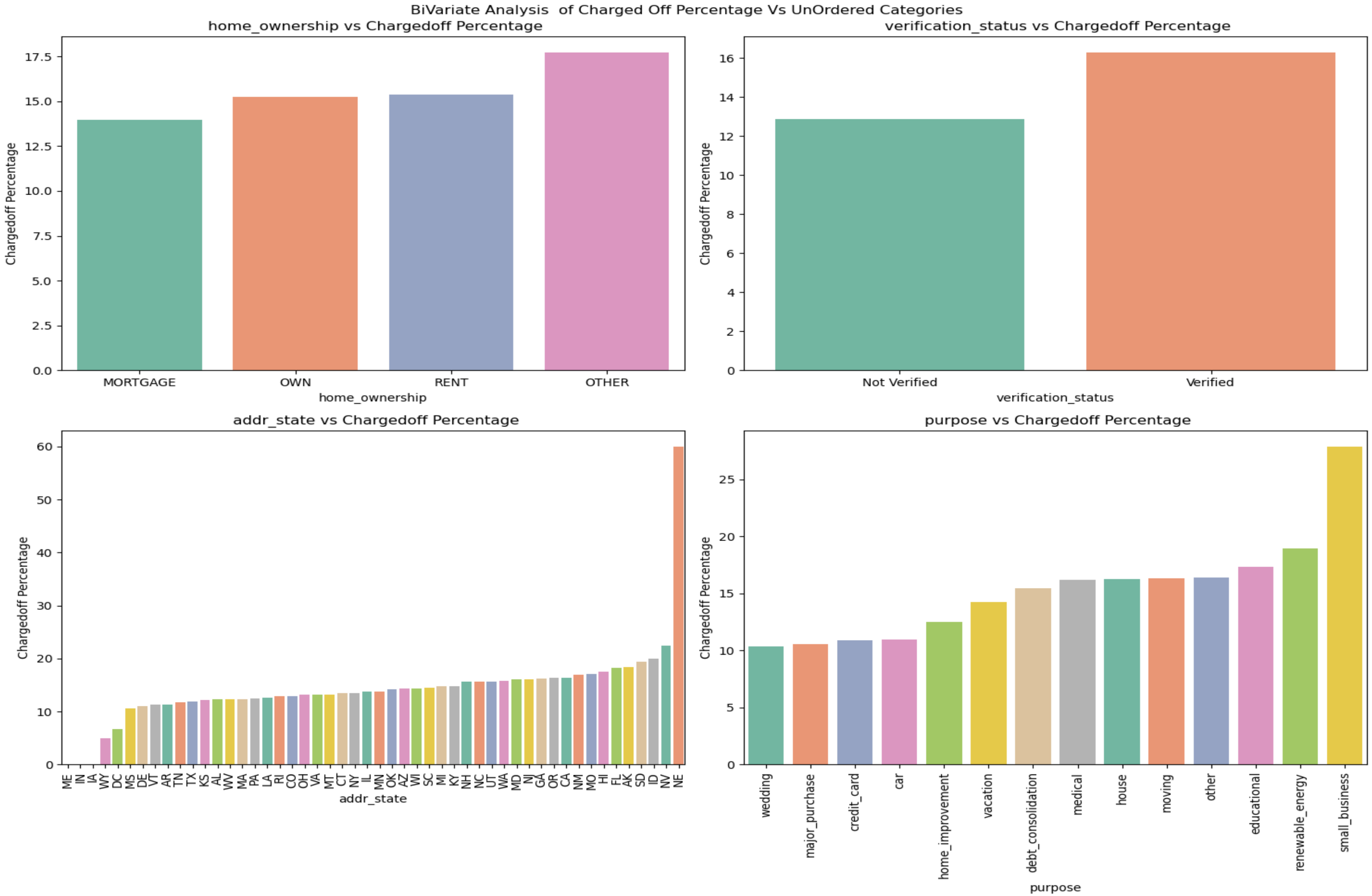
Charged-off percentage is high in the month of **May, September and December** and low in the month of February

Charged-off percentage is high in the **last quarter** of the year

Loan applicants who are having **high public record of bankrupticies** are more likely to default. Lower the Bankruptcies lower the risk.

BiVariate Analysis of Charged Off Proportion vs Various Categories

Bivariate Analysis of UnOrdered Categories vs Chargedoff_Proportion



BiVariate Analysis of Charged Off Proportion vs Various Categories

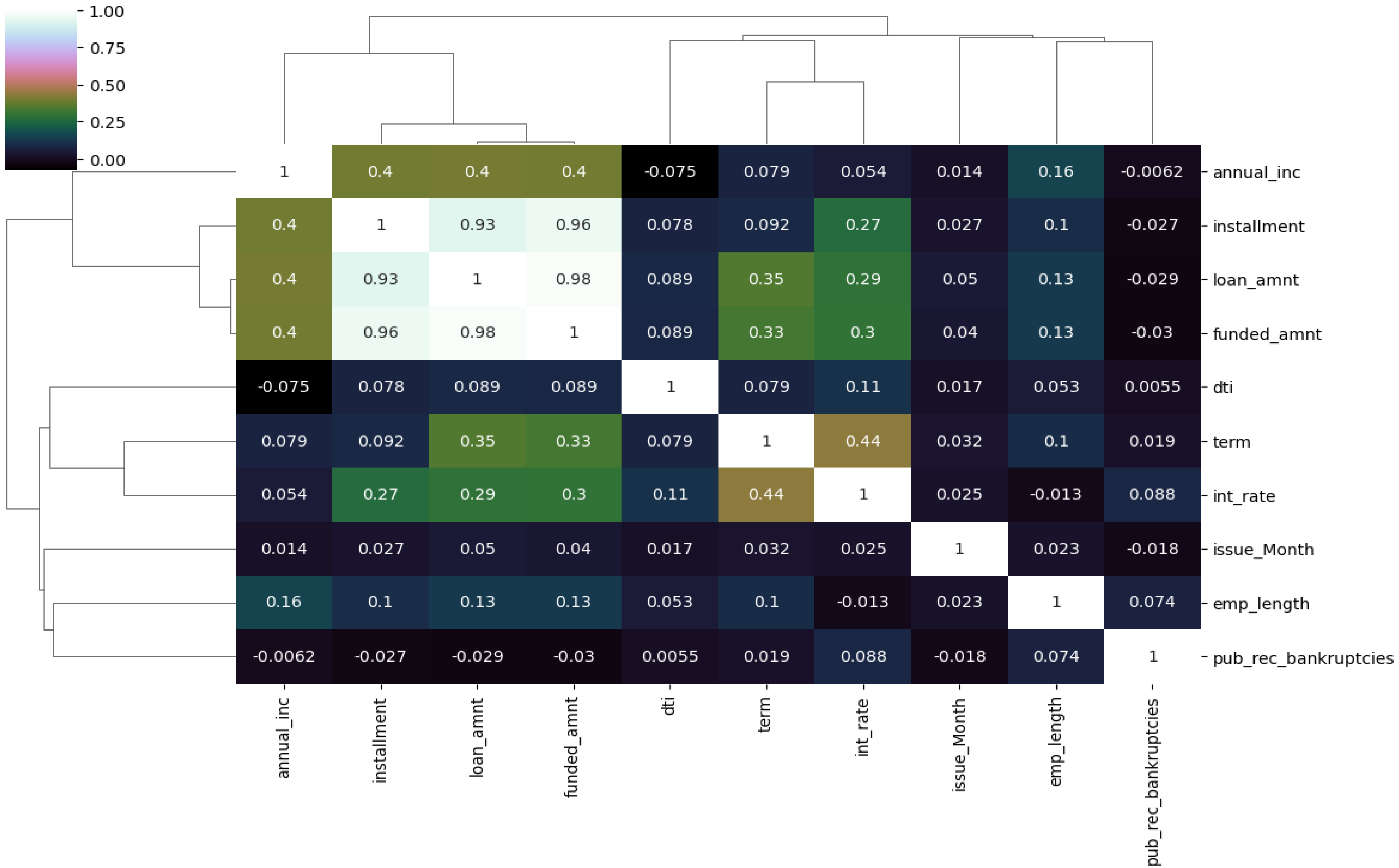
Bivariate Analysis of UnOrdered Categories vs Chargedoff_Proportion

Observations

- Those who are **not owning the home** is having high chances of loan defaults.
- Those applicants who is having mortgage is having low chances of loan defaults.
- Those applicants having loan taken for **small business** is having high chances of defaults.
- Those applicants having loan for wedding is less likely to do loan defaults.
- **Florida** States is having high number of loan defaults   DC is having low number of loan defaults
- The loans which are in ‘Verified’ status is having more number of defaults than in ‘not-verified’ status. Hence income verification alone is not sufficient to identify the defaulters.

Multivariate Analysis

Correlation Analysis



Multivariate Analysis

Correlation Analysis

Observations

Strong/Moderate Correlation

- **Installment** has a strong correlation with **Loan Amount & Funded Amount**
- **Annual Income** has a moderate correlation with **Loan Amount** and **Installment**
- **Term** has a moderate correlation with Interest **Rate** and **Loan Amount**
- **Interest rate** has a moderate correlation with **Loan Amount**

Weak Correlation

- **dti** has weak correlation with most of the fields
- **emp_length** has weak correlation with most of the fields
- **issue_Month** has weak correlation with most of the fields

Negative Correlation

- **Annual income** has a weak negative correlation with **dti**
- **Loan Amount** has a weak negative correlation with **Public Bankruptcies**

Conclusion

Key Findings

- **Lower the Annual income** (below 40k), more likely to default. Higher the annual income, the less chances of default.
- **Higher the interest rate**(above 15%), more likely to default.
- **Higher the Loan Amount** (ie Above 15K) More likely to default
- **Higher the Installment** High (ie.Above USD 800) more likely to default
- **Higher the debt to income ratio (dti)** %), more likely to default
- When **Revolving Balance** is more there is a slightly more chances of charged-off
- **Revolving Line Utilization** - When revolv-util percentage increases, chances of defaulting also increases. When revolv-util is below 25%, there is less chances default and when it is above 75%, there is a high chances of default
- Loan applicants applying loan for **grade G** are **more likely** to default than any other category and grades **A and B** are **less likely** to default
- Loan applicants applying loan for **60 months** are more likely to default than the one taking loan for 36 months
- Loan applicants applying loan having employment length **10+ years** are more likely to default than the one having lesser years
- Loan applicants who are having **high public record of bankruptcies/derogatory records** are more likely to default. Lower the Bankruptcies lower the risk.
- Those who are **not owning the home** is having high chances of loan defaults.
- Those applicants who are having **mortgage is less likely** to default.
- Those applicants having loan taken for **small business** is having high chances of defaults.
- Those applicants having loan for **wedding** is **less likely** to do loan defaults.
- **Florida** States is having high number of loan defaults **DC** is having low number of loan defaults

Conclusion

Recommendations

Underwriting Criteria:

- Consider stricter criteria for lower grades and sub-grades
- Prioritize applicants with lower DTIs and higher incomes

Loan Terms

- Offer shorter loan terms to mitigate risk, especially for higher-risk applicants

Interest Rates

- Implement more targeted pricing strategies based on borrower risk factors

Verification

- Strengthen income verification processes to reduce information asymmetry