Lending Club Case Study

Presented by

Prafulla Mohadikar

OBJECTIVE

The Lending Club Case Study involves a Consumer Finance Company which specializes in lending various types of loans to urban customers.

The Objective of this case study is to identify risky loan applicants so that loan allotment to such applicants can be avoided leading to lower Credit Loss

This is done using EDA.

Benefits of the case study:

- Gives a fair understanding on how EDA is used in real life business problems.
- Develops a basic understanding of risk analytics in banking and financial services.
- Creates an understanding of how the data is used to minimize loss of money while lending it to clients.
- Improves our understating of data visualization and what charts to use for real life data.

BUSINESS UNDERSTANDING

The business objective is to make a decision about whether a loan should be sanctioned for an applicant based on data of previous defaults.

Dataset Details:

- The given dataset contains information about previous loan applicants and "defaults".
- The data contains details of only approved loans (not the rejected ones).
- It has 3 status of loans - Fully Paid, Current and Charged-Off.

Approach:

Data Cleaning

- Fixing rows and columns
- Standardizing values
- Fixing invalid values
- Filtering the data
- Fixing missing values
- Detecting Outliers

Univariate Analysis

- Unordered Categorial Univariate Analysis
- Ordered Categorial Univariate Analysis
- Quantitative Univariate Analysis
- Segmented Univariate Analysis

Bivariate Analysis

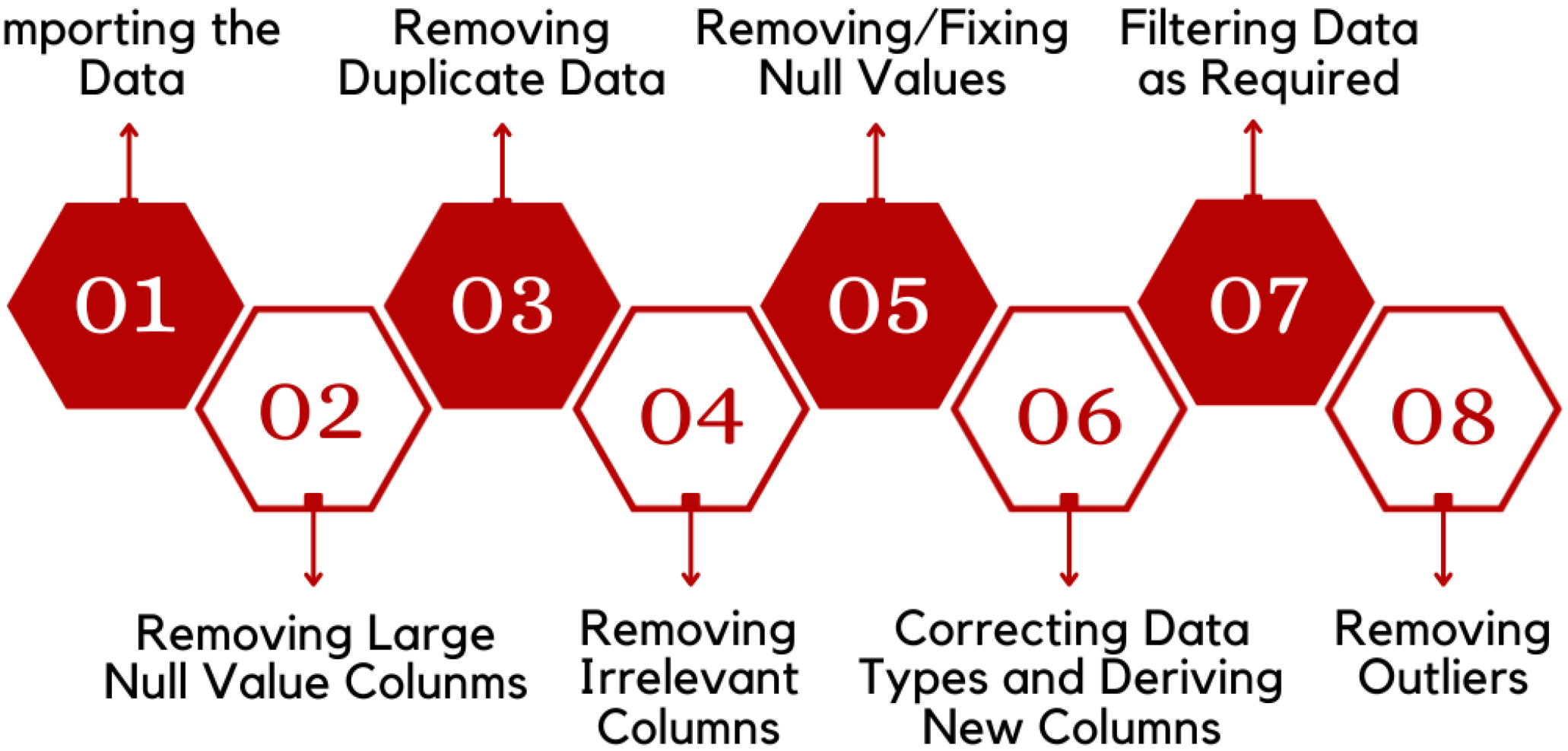
- Unordered Categorial Univariate Analysis
- Ordered Categorial Univariate Analysis
- Quantitative Univariate Analysis
- Segmented Univariate Analysis

Data Cleaning

Data Cleaning is the process of resolving various data quality issues at the source to get useful data for Analysis.

We are using the following Data cleaning techniques to clean up the Loan Dataset.

PROCESS


Importing the Data
01
Removing Large Null Value Colunms
02
Removing Duplicate Data
03
Removing Irrelevant Columns
04
Removing/Fixing Null Values
05
Correcting Data Types and Deriving New Columns
06
Filtering Data as Required
07
Removing Outliers
08

FIXING ROWS AND COLUMNS

Steps:

- Deleting unnecessary columns -We delete all columns with null values or mostly null values
- Renaming columns consistently - We renaming column names like "issue_d" to "issue_date", "last_credit_pull_d" to "last_credit_pull_date" etc
- Deleting incorrect rows, summary rows and extra rows – In this datasheet, none were found
- Adding column names (if missing), Splitting columns for more Data, Merging columns for identifiers and Aligning misaligned columns – This was not required for our dataset Checking the percentage of missing values in each column with more than threshold of 0 percentage (for visibility)
- Dropping Columns with more than 40% of NULL Values based on Empty Column Value Stats

STANDARDIZING VALUES AND FIXING INVALID VALUES

Steps to standardize columns:

- The column ”zip_code” should be standardised by considering only first three characters
- The columns ”int_rate”, ”revol_util” should be standardised by removing percentage suffix.
  Invalid values should be fixed by converting from String to Float
- The column ”verification_status” has values **Verified** and **Source Verified** which are the same, so we replace **Source Verified** with **Verified**
- Columns ”grade”, ”sub_grade”, ”home_ownership”, ”verification_status”, ”purpose” and ”addr_state” should be standardised by converting to datatype category

Steps to fix invalid values:

- The ”issue_date”,” earliest_cr_line” and ”last_credit_pull_date” columns should be fixed by converting them to **datetime**

FILTERING DATA

Dropping irrelevant or duplicate columns:

- Filtering the Columns having non-unique single value which are irrelevant for analysis and dropping them
- Filtering the Columns like ”id”, ”member_id”, ”url” and ”title” which are irrelevant for Analysis and dropping them
- Dropping Columns like ”desc”, ”funded_amnt_inv”, ”out_prncp_inv”, ”total_pymnt_inv” as the data cannot be analysed
- Dropping Column ”emp_title” as the column **purpose** serves the purpose
- Dropping Column ”zip_code” as the column **state** serves the purpose
- Dropping rows from ”loan_status” column which are Current as it is not needed for analysis
- Dropping Column ”funded_amount” as the column **loan_amount** is the same and will be used for analysis

FIXING MISSING VALUES AND DETECTING OUTLIERS

Steps to fix missing values:

- Checking the percentage of missing values in each column after cleaning and there datatypes (for visibility)
- Filling empty cells in column, based on column datatype
- Filling mode in place of missing values for ''emp_length'', ''emp_title'','' last_pymnt_date'', ''last_credit_pull_date'' and ''pub_rec_bankruptcies'' column
- Filling median in place of missing values for ''revol_util ''column

Steps to detect and fix outliers:

An outlier is a point or set of points that are different from other points

- We need to detect and remove outliers because outliers are one of the primary reasons for resulting in a less accurate model
- We get the First Quartile, Third Quartile and Interquartile Range
- We remove the outlier for column **Annual Income** with 99 percentile as there are some outliers beyond this range

Data is now cleaned for Analysis

Univariate Analysis

Univariate Analysis deals with analysing variables one at a time
Univariate analysis is classified into two types:

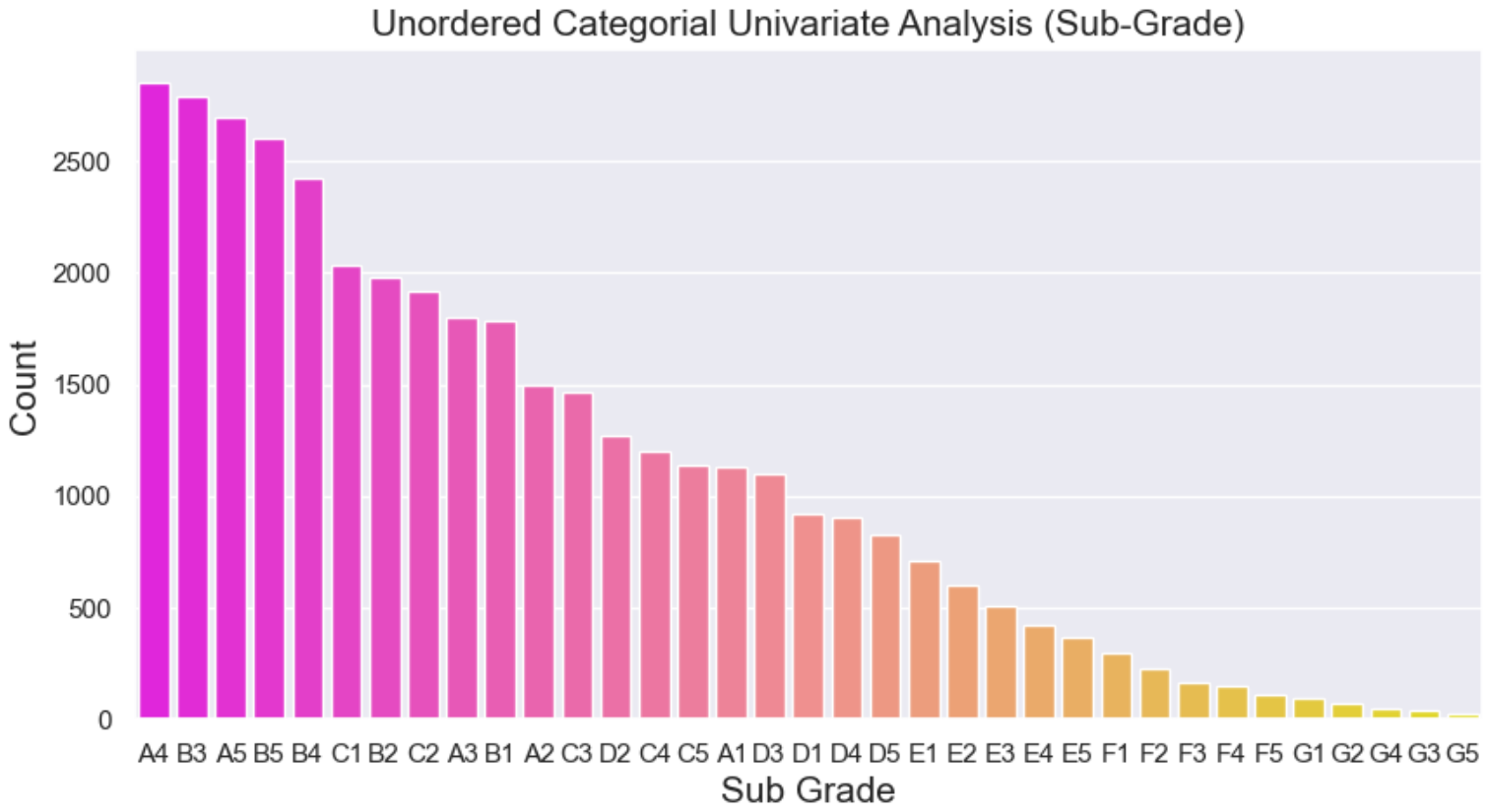
- Categorial Univariate Analysis
  - Unordered Categorial Univariate Analysis
  - Ordered Categorial Univariate Analysis
- Quantitative Univariate Analysis

UNORDERED CATEGORICAL UNIVARIATE ANALYSIS

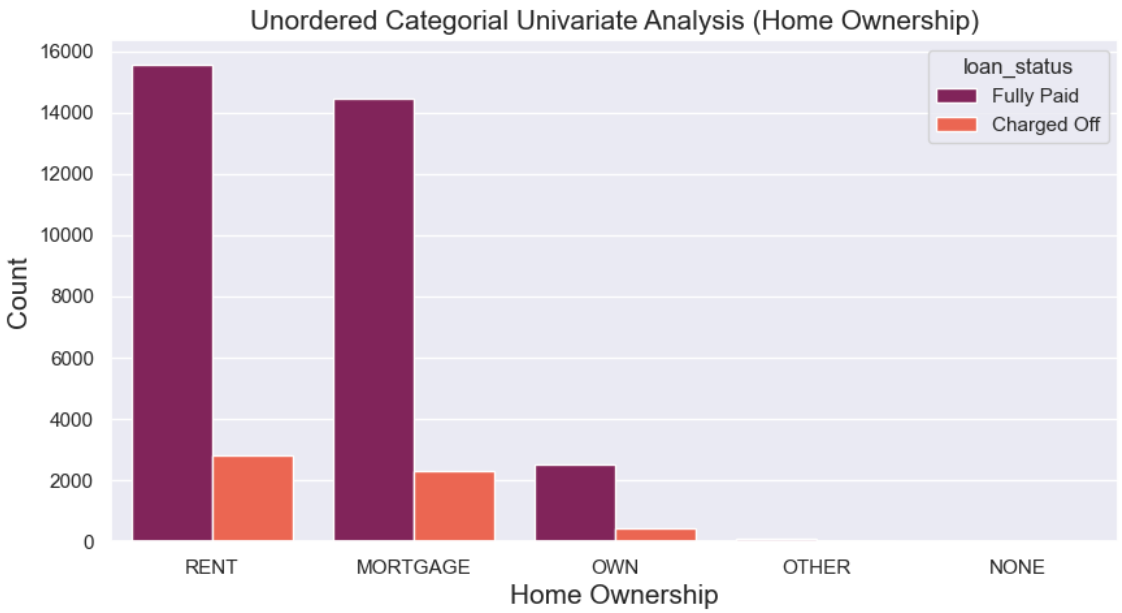
Analysis is done using single variable and its count. The variable involved does not have and sort of ordering

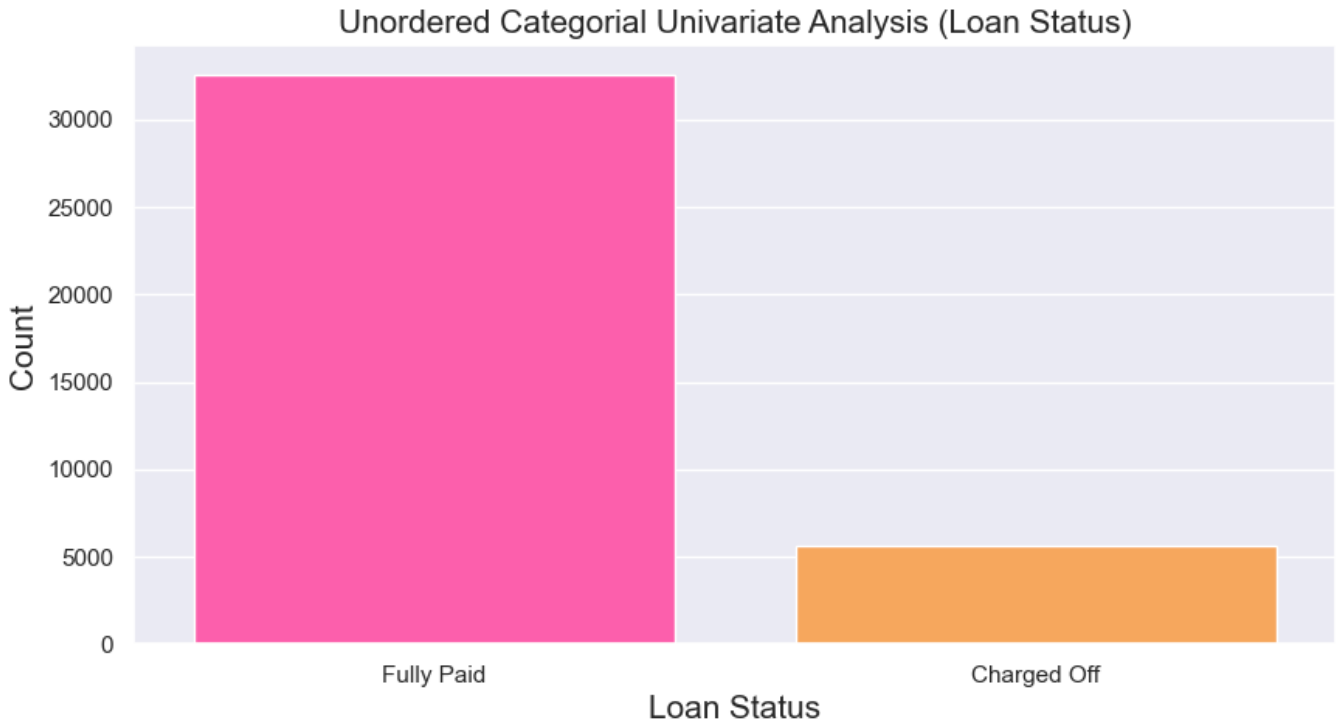
Inference:

The Loan procuring frequency is falling for borrowers having Sub-Grade from E1 and above



UNORDERED CATEGORICAL UNIVARIATE ANALYSIS





- **Inference:** Maximum amont of Loans are taken by borrowers who live in Rented or Mortgage House and majority defaulters lie under same category

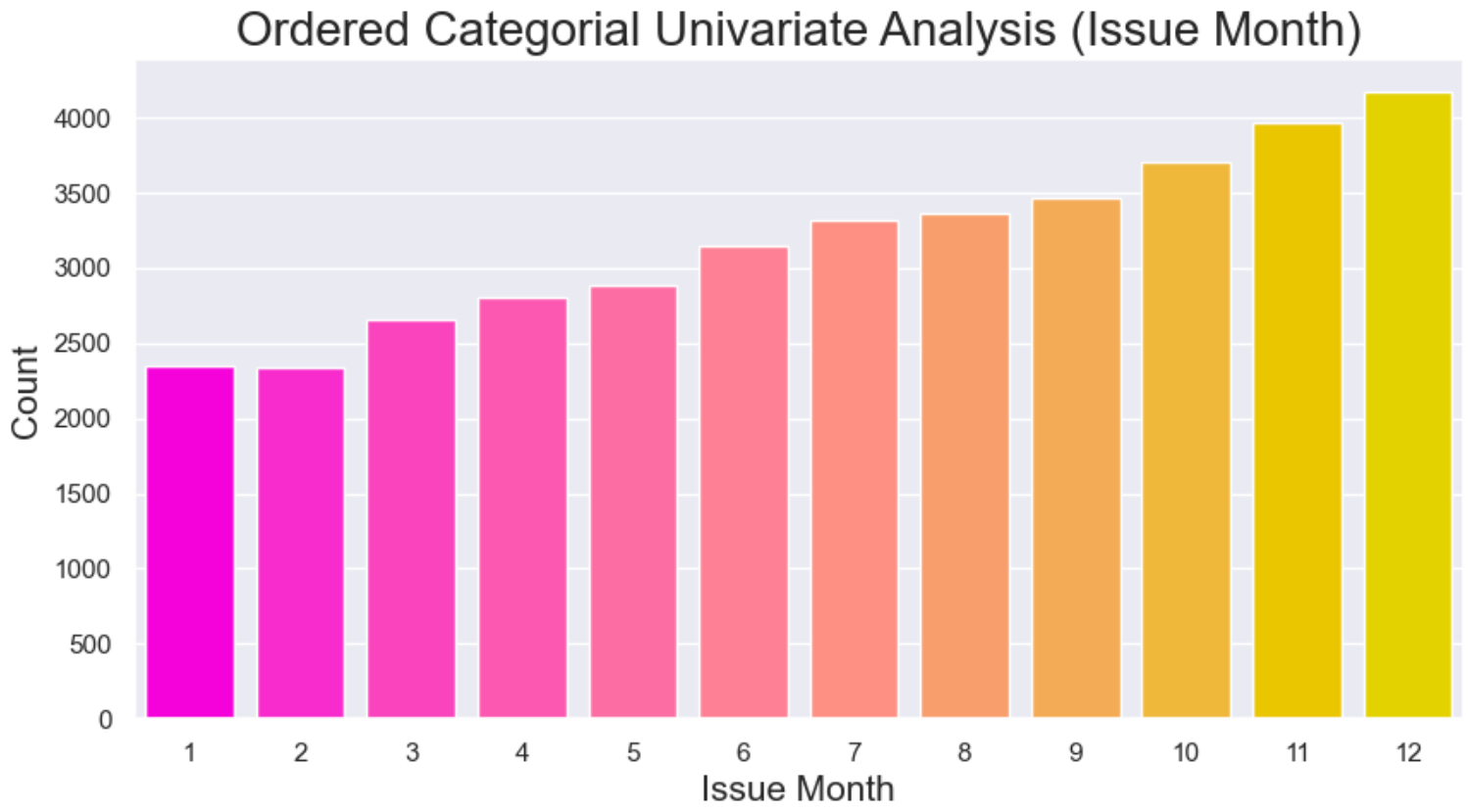
- **Inference:** Almost 13% borrowers have defaulted

ORDERED CATEGORICAL UNIVARIATE ANALYSIS

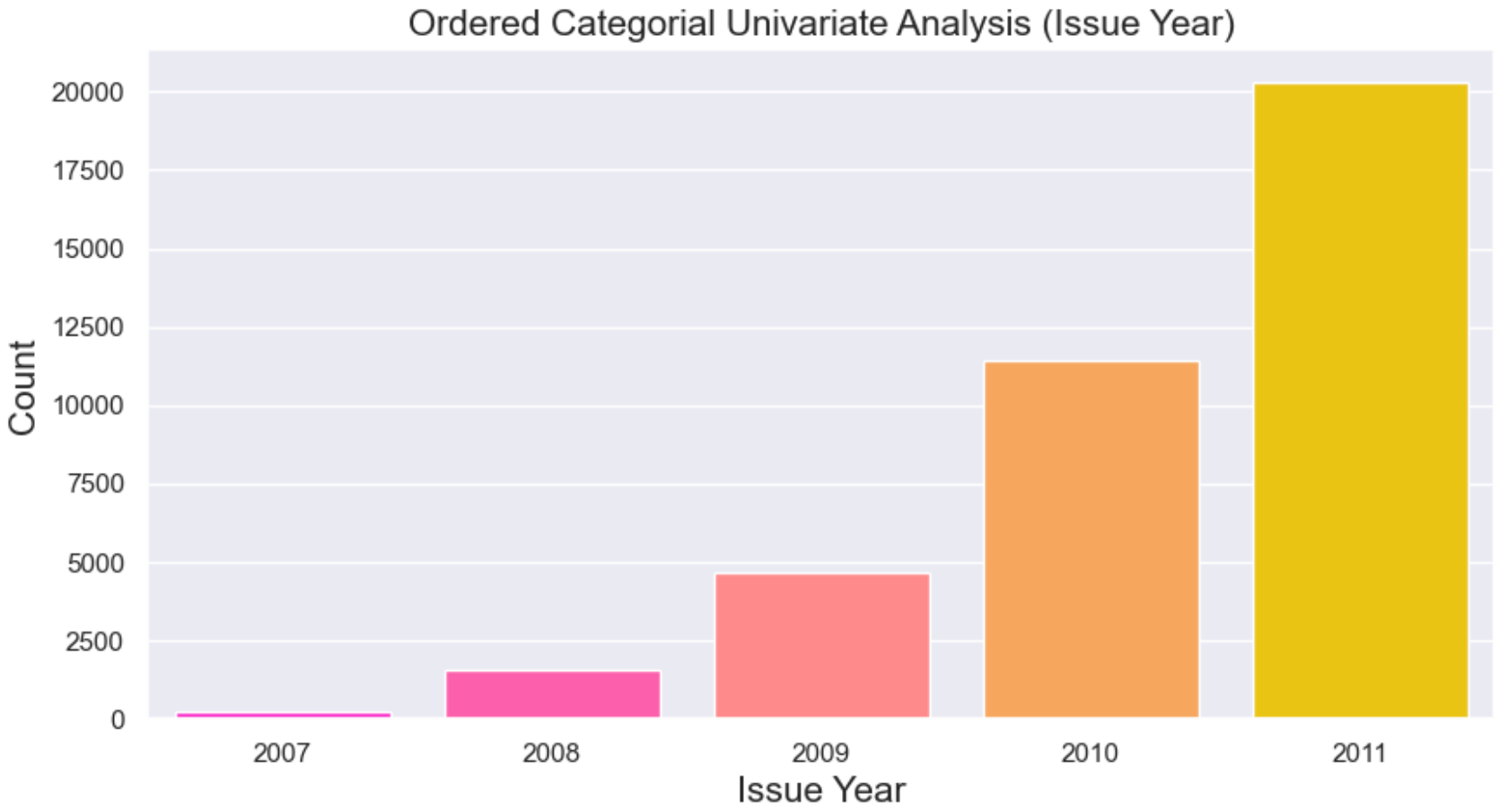
Analysis is done using single variable and its count. The variable involved does have some ordering based on date or numerical values

Inference:

The Volume of Loan borrowing increases at last quarter of the year which indicates borrowers tend to settle there debt consolidations by year end



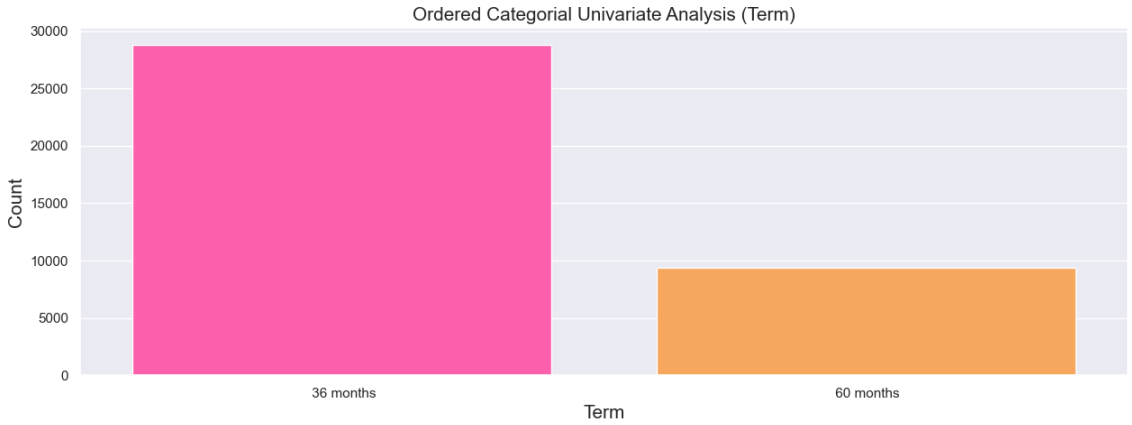
ORDERED CATEGORICAL UNIVARIATE ANALYSIS

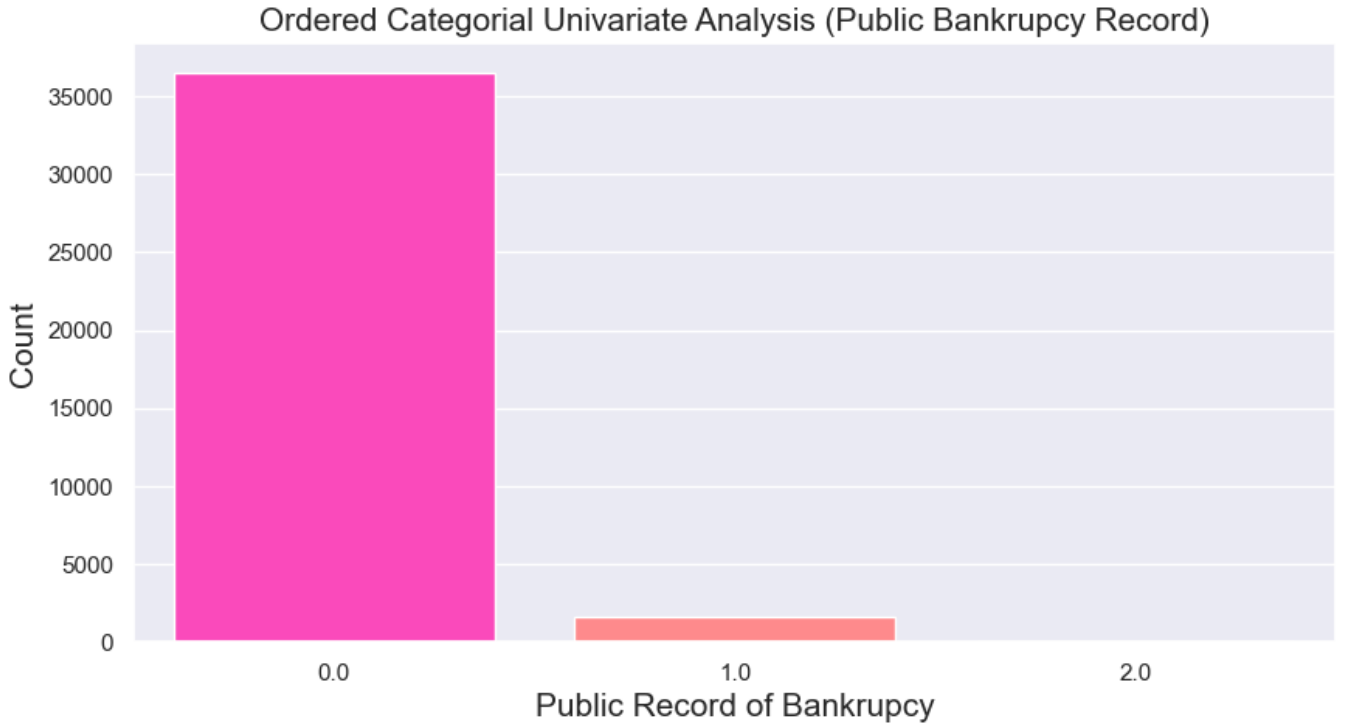


Inference:

Rate of issuing loans increases with time exponentially. Loans borrowed in year 2011 is almost 7 times then year 2008

ORDERED CATEGORICAL UNIVARIATE ANALYSIS

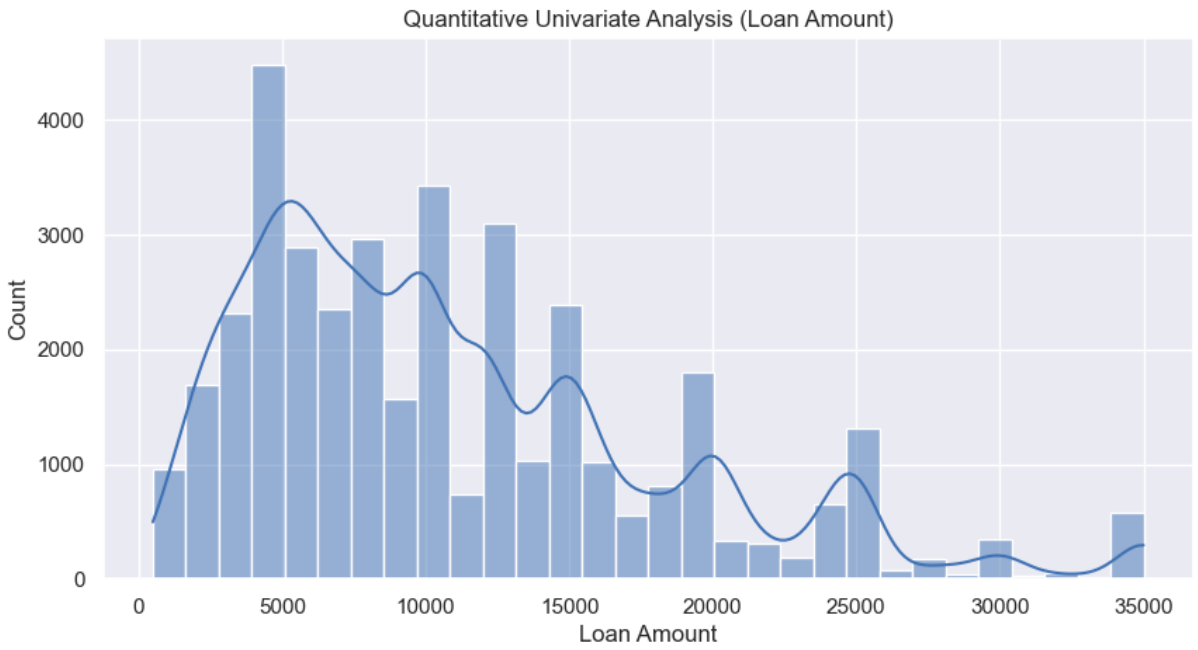




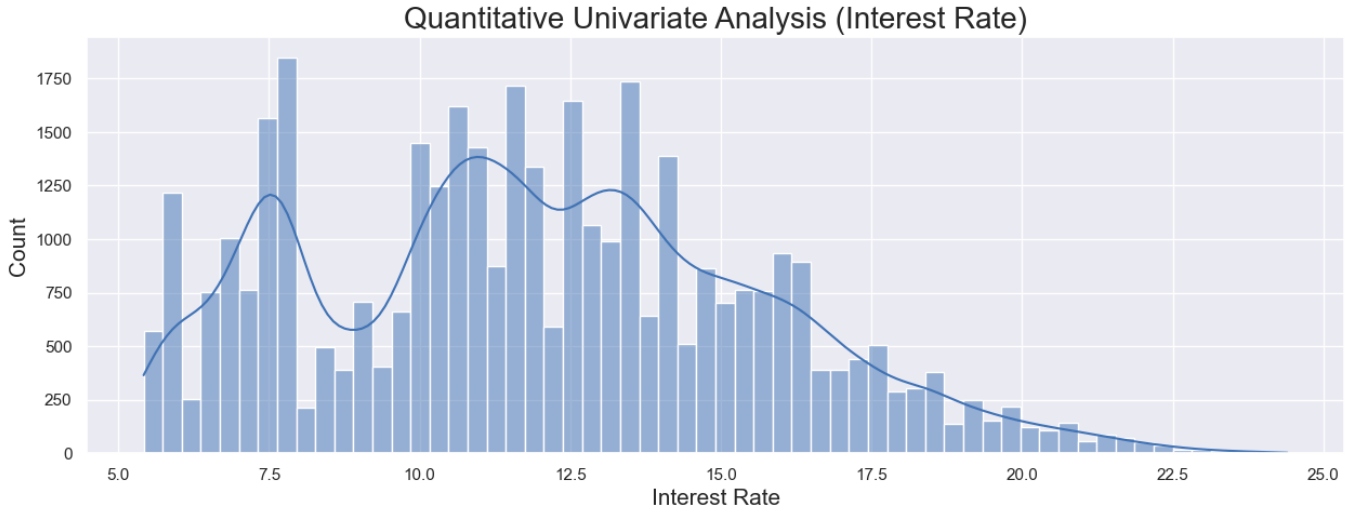
- **Inference:** 75% Loans are borrowed for less duration i.e. 36 Months than 60 Months

- **Inference:** Very small percentage of borrowers have Public Record Bankrupcies so loan default chances are less if there is no public record bankrupcies for an applicant

QUANTITATIVE UNIVARIATE ANALYSIS



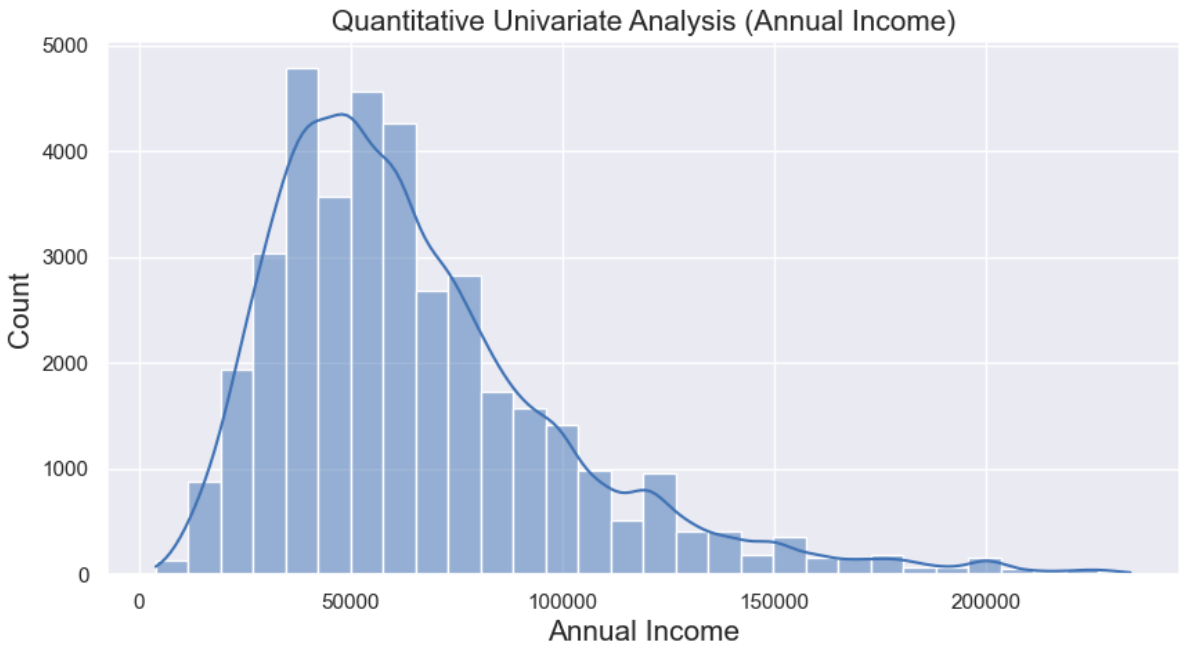
❖**Inference:** Maximum loans are borrowed for amount between 5000 to 15000



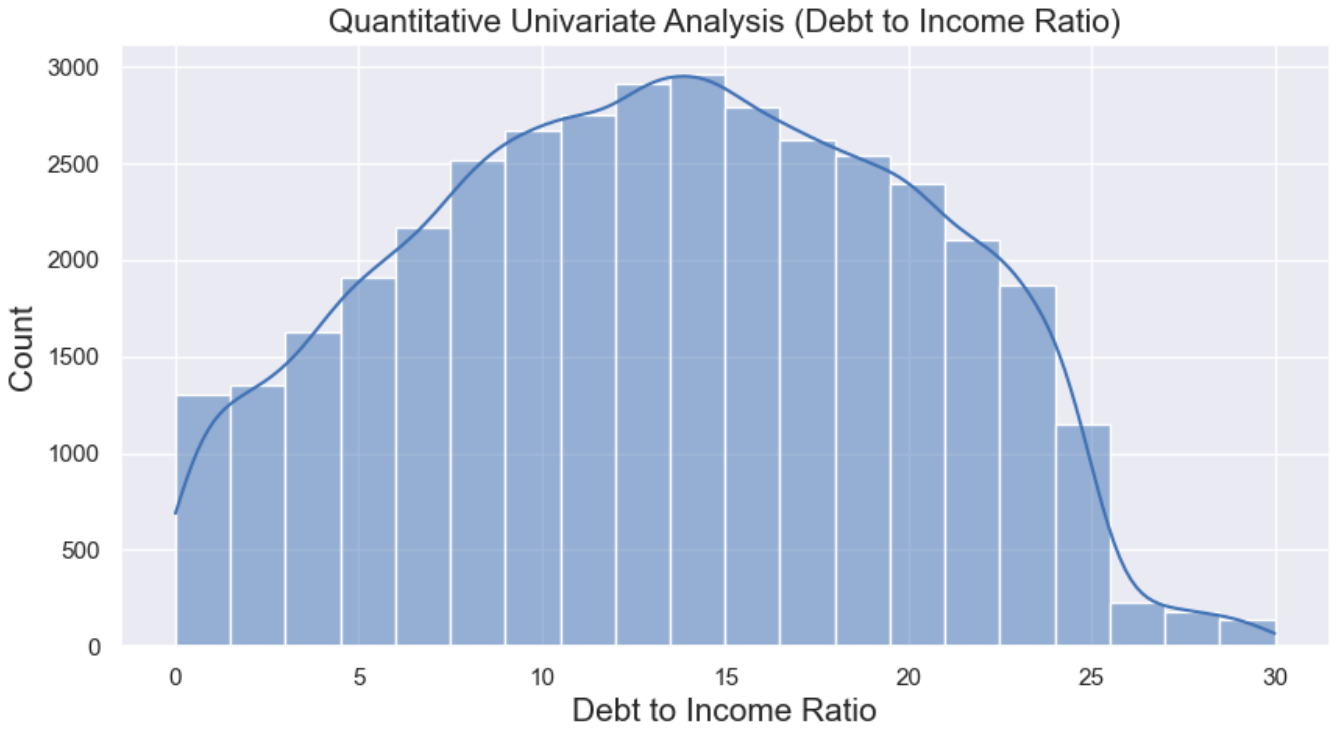
❖**Inference:**

- More Loans are borrowed by interest rates around 5-8% and 10-15%. The quantity of loans borrowed decreases after 15% interest rate.
- No loan is borrowed below 6% and more than 23% interest rate

QUANTITATIVE UNIVARIATE ANALYSIS



❖**Inference:** Majority loans borrowers have less Annual Income as the histogram show left skewed normal distribution

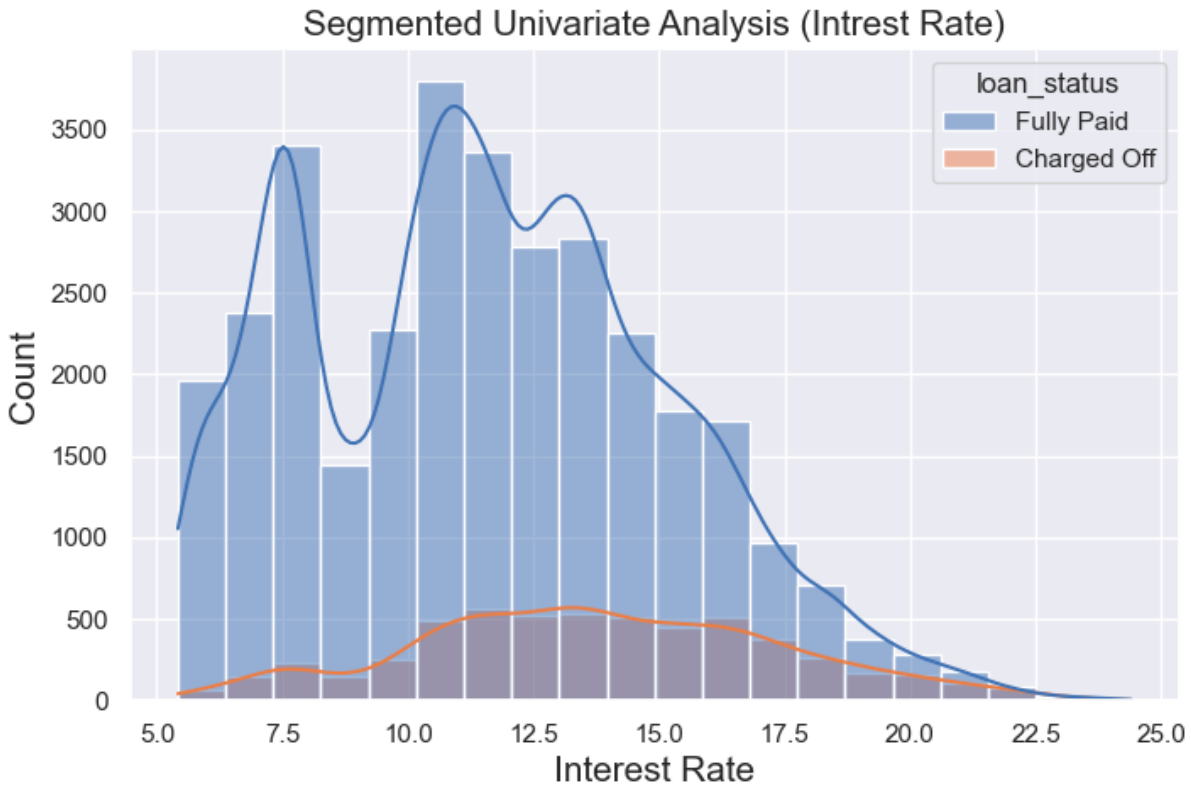


❖**Inference:** Debt to Income ratio is concentrated more between range 10 to 20 of dti

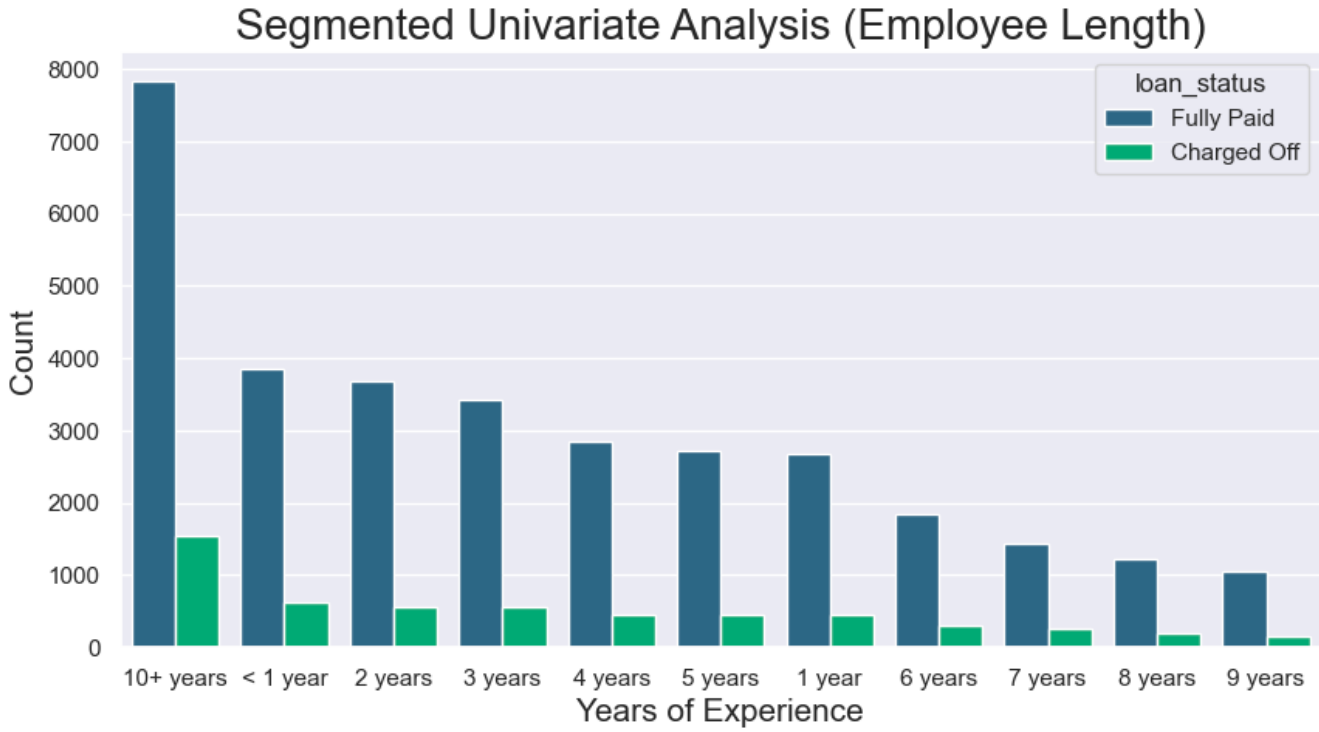
Segmented Univariate Analysis

Analysis is done grouping data on dimensions, comparison of averages and comparison of other metrics

SEGMENTED UNIVARIATE ANALYSIS

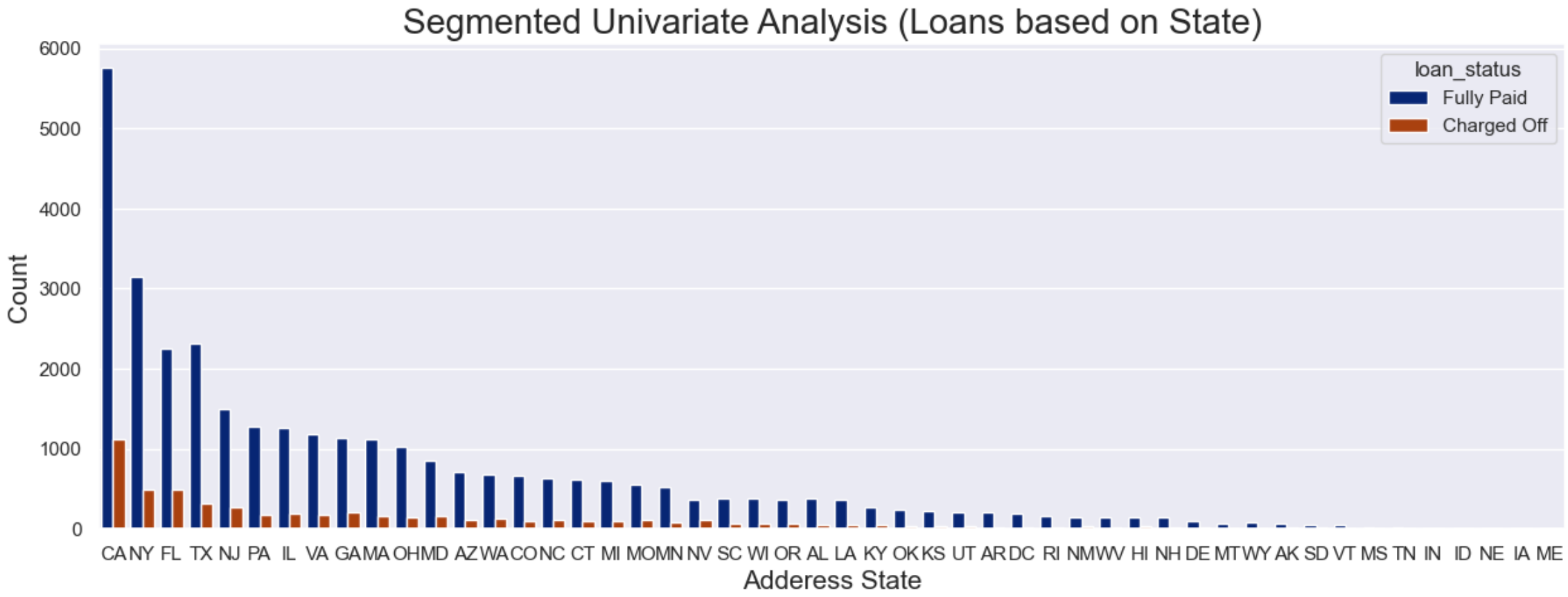


❖**Inference:** Loan default increases with the increase in interest rate



❖**Inference:** Maximum borrowers are having higher experience i.e. greater than 10 years and maximum defaulters are for the same experience range

LOANS BASED ON STATE



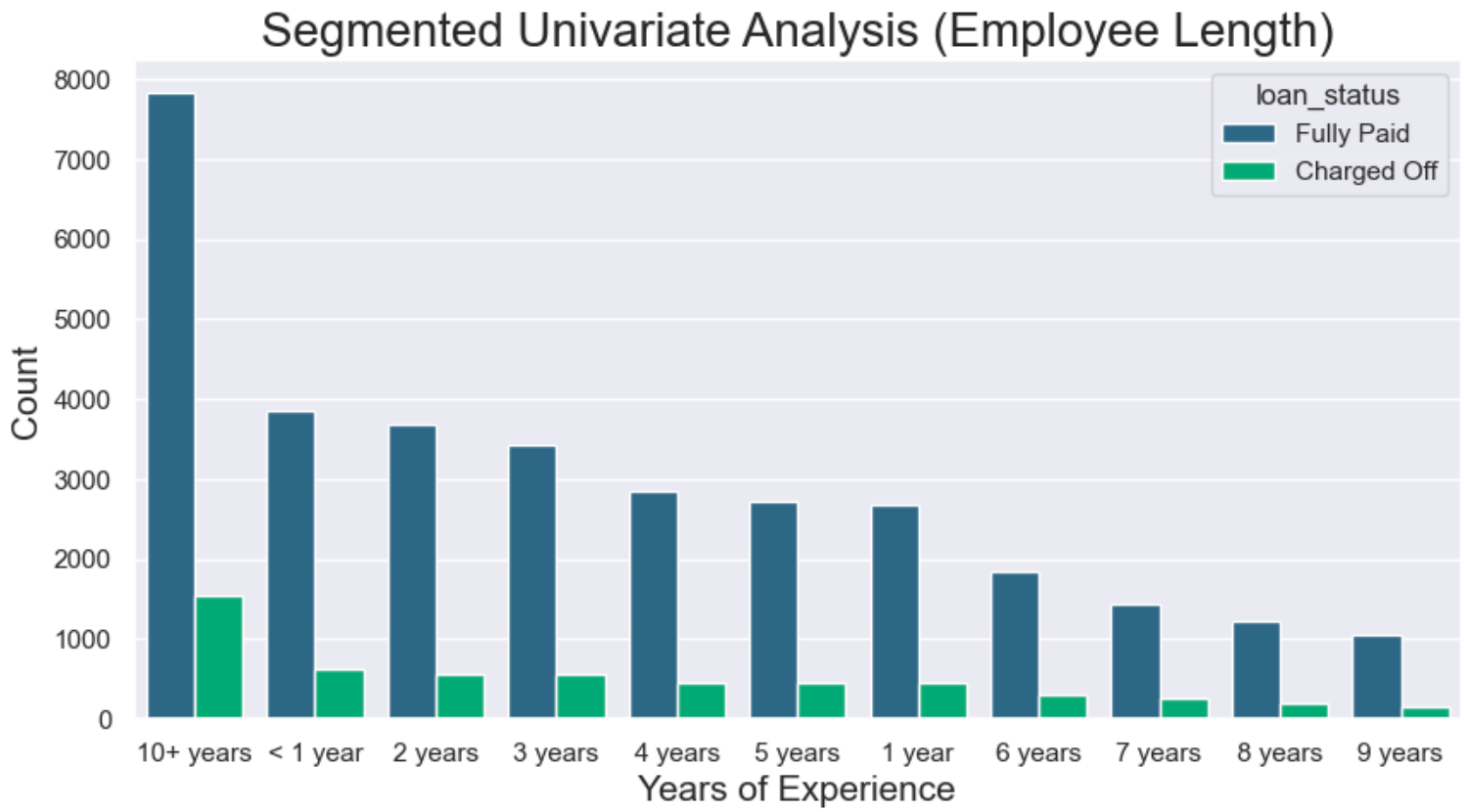
- **Inference:** The volume of loan borrowing increases at last quarter of the year which indicates borrowers tend to settle there debt consolidations by year end

EMPLOYEE LENGTH

Analysis is done using single variable and its count. The variable involved does have some ordering based on date or numerical values

Inference:

Defaults percentage is high for tenure of 60 months than tenure of 36 months. So to lend loan for less duration is recommended

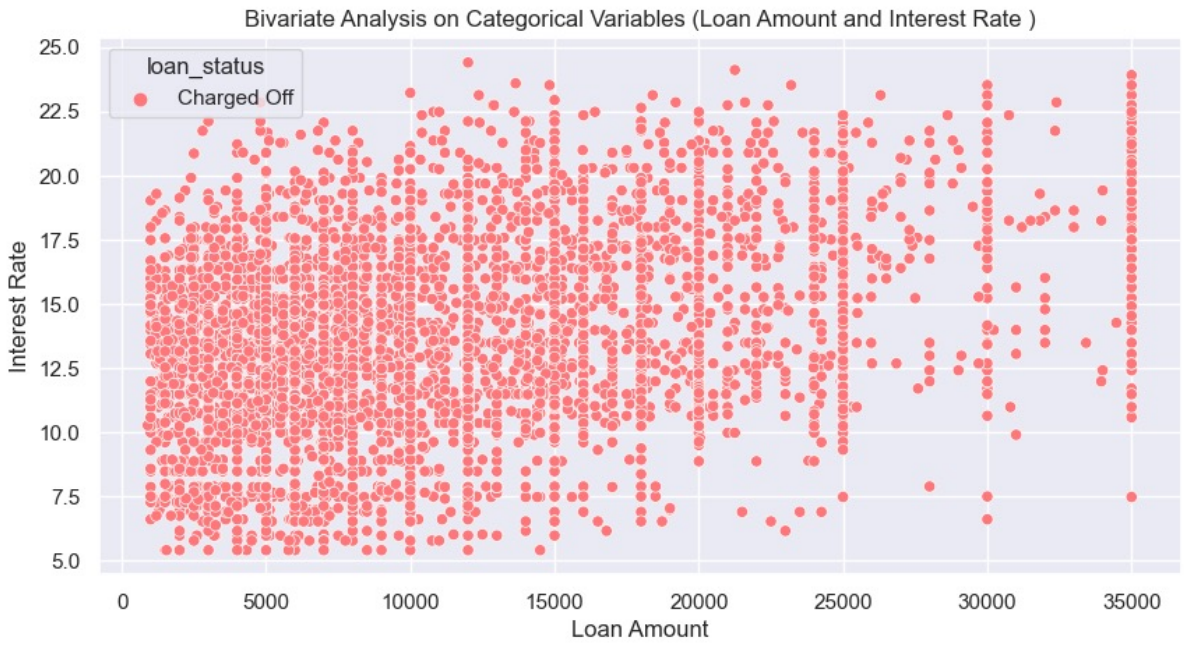


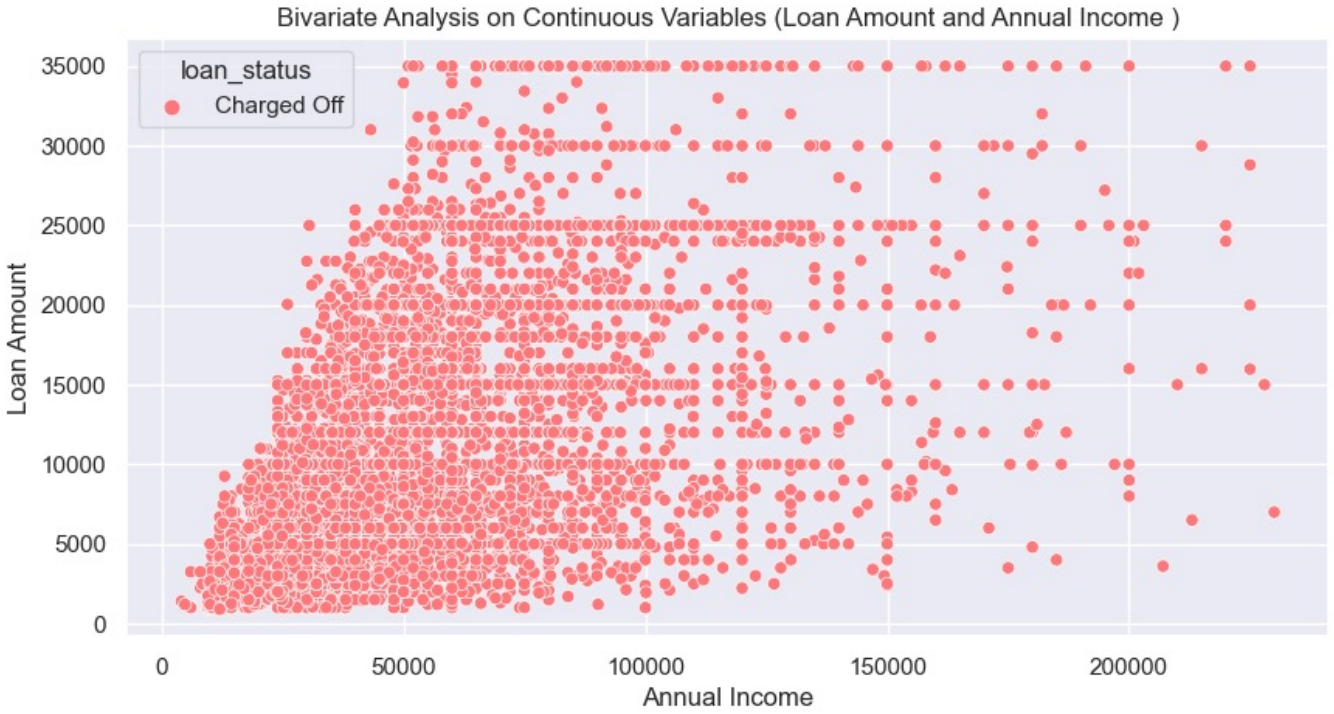
Bivariate Analysis

Analysis between 2 variables is called Bivariate Analysis

- Bivariate Analysis on Continuous Variables
- Bivariate Analysis on Categorial Variables

BIVARIATE ANALYSIS ON CONTINUOUS VARIABLES

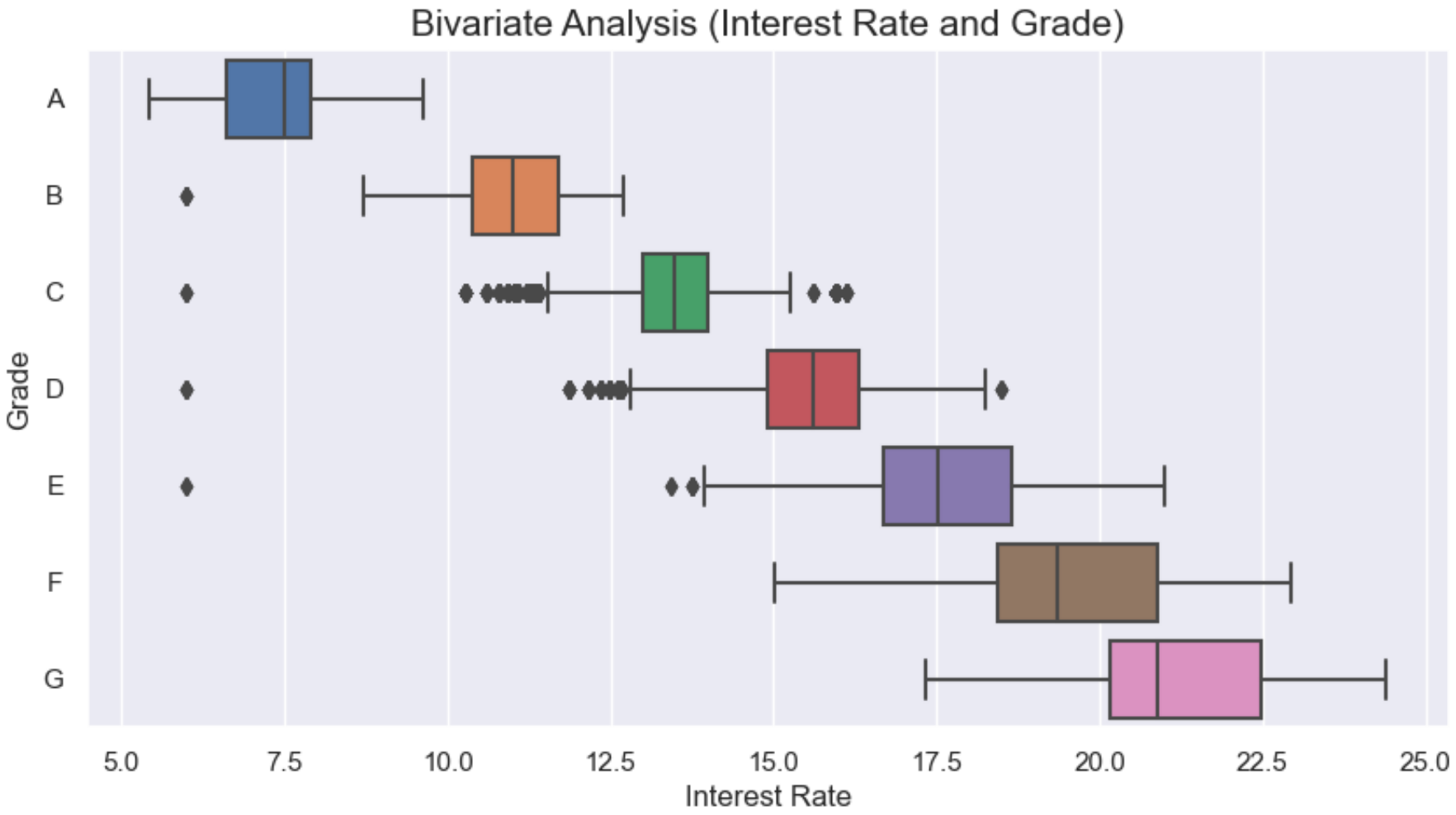




- **Inference:** Loan Defaults increase with increase in Loan Amount or Interest Rate between 10% to 17.5%

- **Inference:** Dense loan defaults is visible with High Loan Amount and Low Annual Income

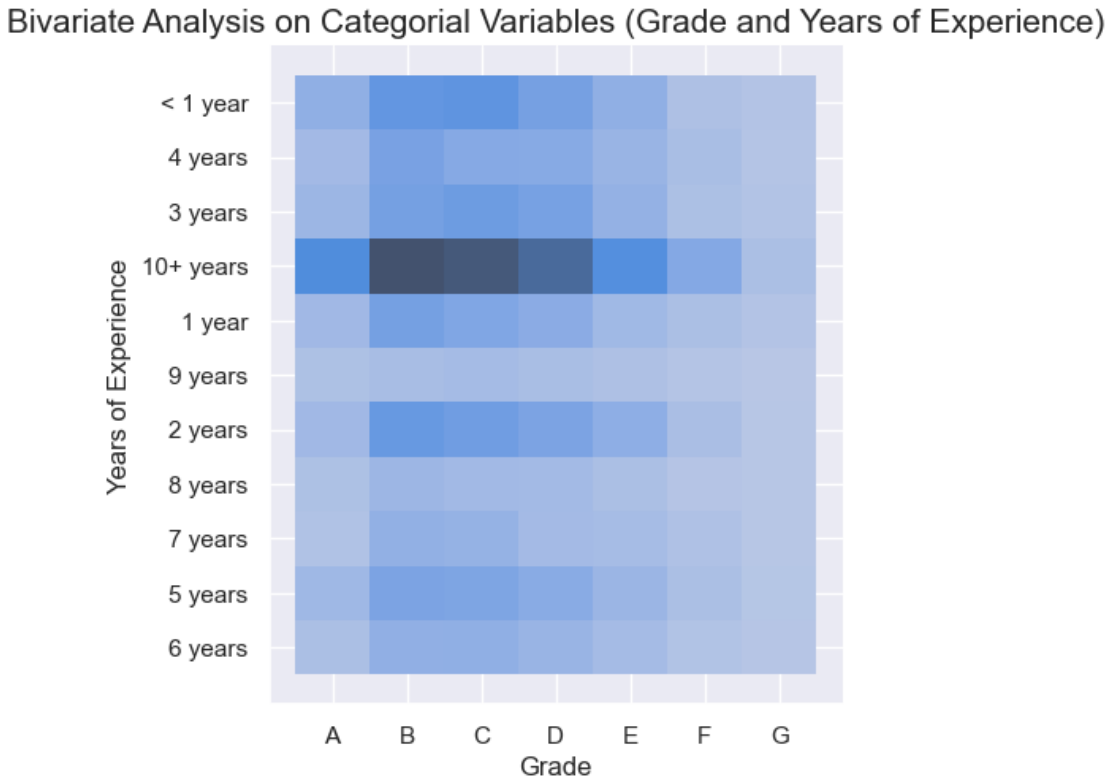
BIVARIATE ANALYSIS ON CONTINUOUS AND CATEGORICAL VARIABLES

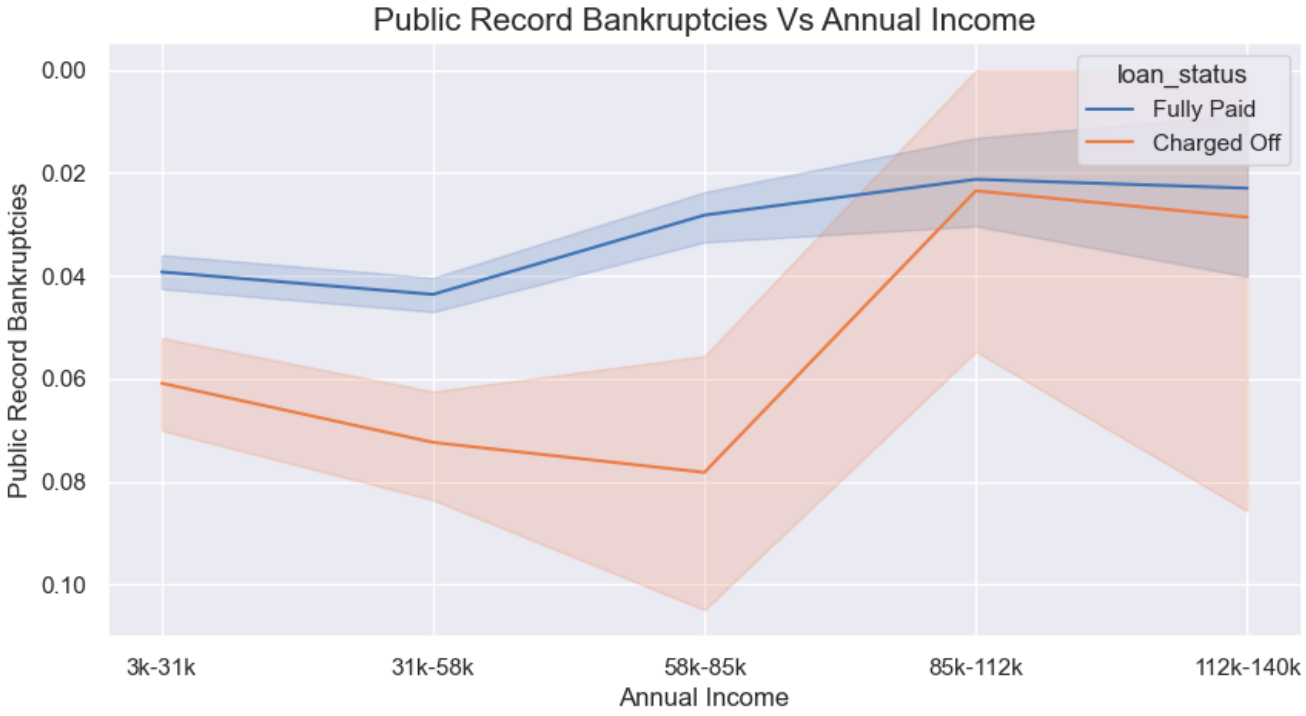


Inference:

Higher Interest Rate are assigned to borrowers with Higher Grade from A - G

BIVARIATE ANALYSIS ON CATEGORICAL VARIABLES





- **Inference**: Higher Interest Rate are assigned to borrowers with Higher Grade from A - G

- **Inference**: Borrowers having no public record bankruptcy usually are not defaulters

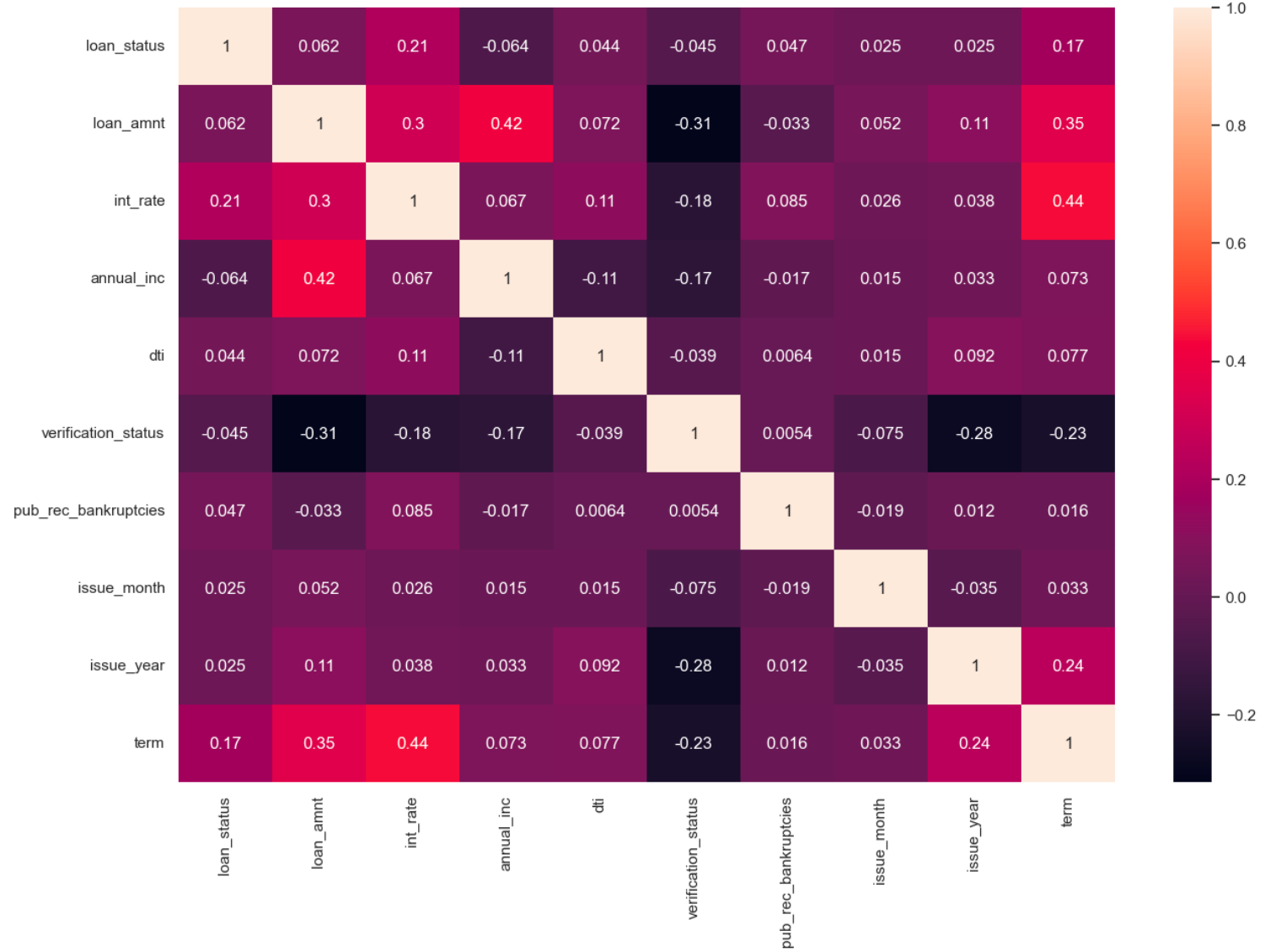
Derived Metrics

Creating variables using co-relationship between variables

DERIVED METRICS

Inference:

- Loan Status is directly proportional to Loan Amount by 0.062, higher the loan amount higher the chances of default
- Loan Status is directly proportional to Interest rate by 0.21, higher the interest higher the chances of default
- Loan Status is inversely proportional to Annual Income by 0.064, higher the annual income less the chances of default
- Loan Status is directly proportional to Debt to Income ratio by 0.044, higher the dti higher the chances of default
- Loan Status is inversely proportional to Verification Status by 0.045, more chances of default if verification is successful which shows flaws in Verification status
- Loan Status is directly proportional to Pub Rec Bankruptcy ratio by 0.047, higher the Public Record for Bankruptcy higher the chances of default
- Loan Status is directly proportional to Issuing months by 0.025, More chances of default at the last quarter
- Loan Status is directly proportional to Issuing year by 0.025, More chances of default at as we proceed with timeline
- Loan Status is directly proportional to Loan Tenure by 0.047, higher the tenure higher the chances of default



Final Observations

DETAILED ANALYSIS

Data Cleaning

- All the non required or empty columns are dropped.
- The null or na values are filled in with appropriate data in the column.
- Removed outliers to increase the efficiency of Analysis.
- Derived columns as per requirement to get clean dataset for analysis.

DETAILED ANALYSIS

Univariate Analysis

- The Loan procuring frequency is falling for borrowers having Sub-Grade from E1 and above
- Maximum amount of Loans are taken by borrowers who live in Rented or Mortgage House and majority defaulters lie under same category
- Almost 13% borrowers have defaulted
- Lesser the duration less the chances of default, 75% Loans are borrowed for less duration i.e. 36 Months than 60 Months
- Very small percentage of borrowers have Public Record Bankruptcies so loan default chances are less if there is no public record bankruptcies for an applicant.
- Rate of issuing loans increases with time exponentially. Loans borrowed in year 2011 is almost 7 times then year 2008
- The Volume of Loan borrowing increases at last quarter of the year which indicates borrowers tend to settle there debt consolidations by year end
- Maximum loans are borrowed for amount between 5000 to 15000
- More Loans are borrowed by interest rates around 5-8% and 10-15%. The quantity of loans borrowed decreases after 15% interest and no loan is borrowed below 6% and more than 23% interest
- Majority loans borrowers have less Annual Income as the histogram show left skewed normal distribution
- Debt to Income ratio is concentrated more between range 10 to 20 of dti

DETAILED ANALYSIS

Segmented Univariate Analysis

- Loan default increases with the increase in loan amount as default has 3rd quartile higher than paid-off.
- Loan default increases with the increase in interest rate.
- Loan default decreases with the increase in annual income. Annual Income is inversely proportional to Loan Defaults.
- Loan defaults are highest with Debt to Income Ratio between 10 to 20.
- Maximum borrowers are having higher experience i.e. greater than 10 years and maximum defaulters are for the same experience range.
- High number of loans are taken by borrowers with grade B and A & Less number of loans are taken by borrowers with grade G and F. High percentage of defaulters are from Grade B, C and D. So person from grade A is healthy grade or borrowers to lend.
- More than 50% borrowers are verified by the companies. But surprisingly more defaulters are verified borrowers this show loop holes with verification process.
- Defaults percentage is high for tenure of 60 months than tenure of 36 months. So to lend loan for less duration is recommended.
- Maximum loans are borrowed for the purpose of Debt Consolidation of multiple loans and defaults are also more for same purpose.
- Maximum borrowers are from state California, NewYork, Texas and Florida and defaulters count is also proportionate.
- Loan Default are increasing from 2009 to 2011 may be due to recession.
- The Volume of Loan borrowing increases at last quarter of the year which indicates borrowers tend to settle there debt consolidations by year end and this results as main purpose for the loan procurement.

DETAILED ANALYSIS

Bivariate Analysis

- Loan default increases with increase in Loan Amount or Interest Rate between 10% to 17.5%.
- Dense loan defaults is visible with High Loan Amount and Low Annual Income.
- Higher Interest Rate are assigned to borrowers with Higher Grade from A - G.
- Borrowers having no public record bankrupcy usually are not defaulters.
- Higher Interest Rate are assigned to borrowers with Higher Grade from A - G.

DETAILED ANALYSIS

Derived Metrics

- Loan Status is directly proportional to Loan Amount by 0.062, higher the loan amount higher the chances of default.
- Loan Status is directly proportional to Interest rate by 0.21, higher the interest higher the chances of default.
- Loan Status is inversely proportional to Annual Income by 0.064, higher the annual income less the chances of default.
- Loan Status is directly proportional to Debt to Income ratio by 0.044, higher the dti higher the chances of default.
- Loan Status is inversely proportional to Verification Status by 0.045, more chances of default if verification is successful which shows flaws in Verification status.
- Loan Status is directly proportional to Pub Rec Bankruptcy ratio by 0.047, higher the Public Record for Bankruptcy higher the - chances of default.
- Loan Status is directly proportional to Issuing months by 0.025, More chances of default at the last quarter.
- Loan Status is directly proportional to Issuing year by 0.025, More chances of default at as we proceed with timeline.
- Loan Status is directly proportional to Loan Tenure by 0.047, higher the tenure higher the chances of default.

RECOMMENDATIONS

Recommendations to decrease the Loan Defaults

- The below Driver variables can be used to avoiding Credit Loss
  - DTI (Debt to Income Ratio)
  - Interest Rate
  - Annual Income
  - Loan Amount
  - Loan Tenure
  - Purpose of Loan
  - Address State
  - Employment Experience
  - Home Ownership
  - Public Record Bankruptcies
  - Month of Issue
- Some suggestions to reduce the loan defaults
  - Verification Process needs to be rectified and flaws should be resolved to get more non-default loan issuance ratio.
  - Lending loan to higher annual income group will reduce loan defaults.
  - Loan off higher amount should be given with through due diligence.
  - Borrowers with high debt to income ration are risk borrowers.
  - Borrowers with more than 10 years of work experience are high risk borrowers.
  - Borrowers with annual income between 25000 to 75000 are more likely to default.