Lending Club Case Study

Table of Contents

Problem Statement

A consumer finance company specializing in lending various loans to urban customers. When the company receives a loan application, the company has to decide on loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

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

Approach

Load Data

First of all, we are reading the data file and loading the data in the memory. The data is given in a .csv file.
File name: loan.csv

Data Structure Analysis For Data Clean-up

Then we analyzed its data structure from different angles. Below are the details:

- Analyzed the number of rows and columns in the given data.
- Data types of the different fields in the data.
- Analyzed the required fields for the analysis of the given problem statement.
- Found columns that are completely empty or very less data or not useful for the analysis and drop them as part of data cleaning.
- Identified the columns required to clean data or impute missing data to use it for analysis.

Data Cleaning

- Dropped 54 empty columns and some more columns having very fewer data and columns which doesn't add value to the analysis.
- Corrected some column data to convert them to int or float data type from object type to make them suitable for the numerical analysis.
- Analyzed key fields for outliers, found it and removed/dropped such data as it could negatively influence the analysis.

Data Analysis Techniques

We used following data analysis techniques for Lending Club project.

- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis

Derived Columns

We derived new columns from the existing columns/fields for Bivariate Analysis as it required during the analysis.

Custom Function

We created a custom function to find the percentage of Charged off loans against different parameters for Bivariate Analysis.

Data Cleaning and Results

After loading the data file, initial data structure was (39717, 111). Below are the data cleaning results:

- Found **54** empty columns and dropped them.
- **Result:** After that the data structure was (39717, 57).
- Then we removed the columns having fewer data and columns having no contribution to the analysis. As there is not enough data to participate in the analysis.
- **Result:** After that the data structure was (39717, 48)
- Corrected the **emp_length** column by imputing "**0**" where it was missing data. Also extracted the numeric data in each row and trimmed extra characters in the column. Then changed its type to integer.
  Result: emp_length column don't have any null value and its data type is integer now.
- Corrected the **int_rate** (interest rate) column by trimming/removing "%" from all the rows. Then converted its data type to float.
  Result: int_rate column's data type is float now.

During univariate analysis some rows were dropped to remove outliers. We will see that during univariate analysis and its results.

During Bivariate analysis some columns were derived and added to the dataframe. We will see that during bivariate analysis and its results.

Data Analysis Techniques, Results and Observations

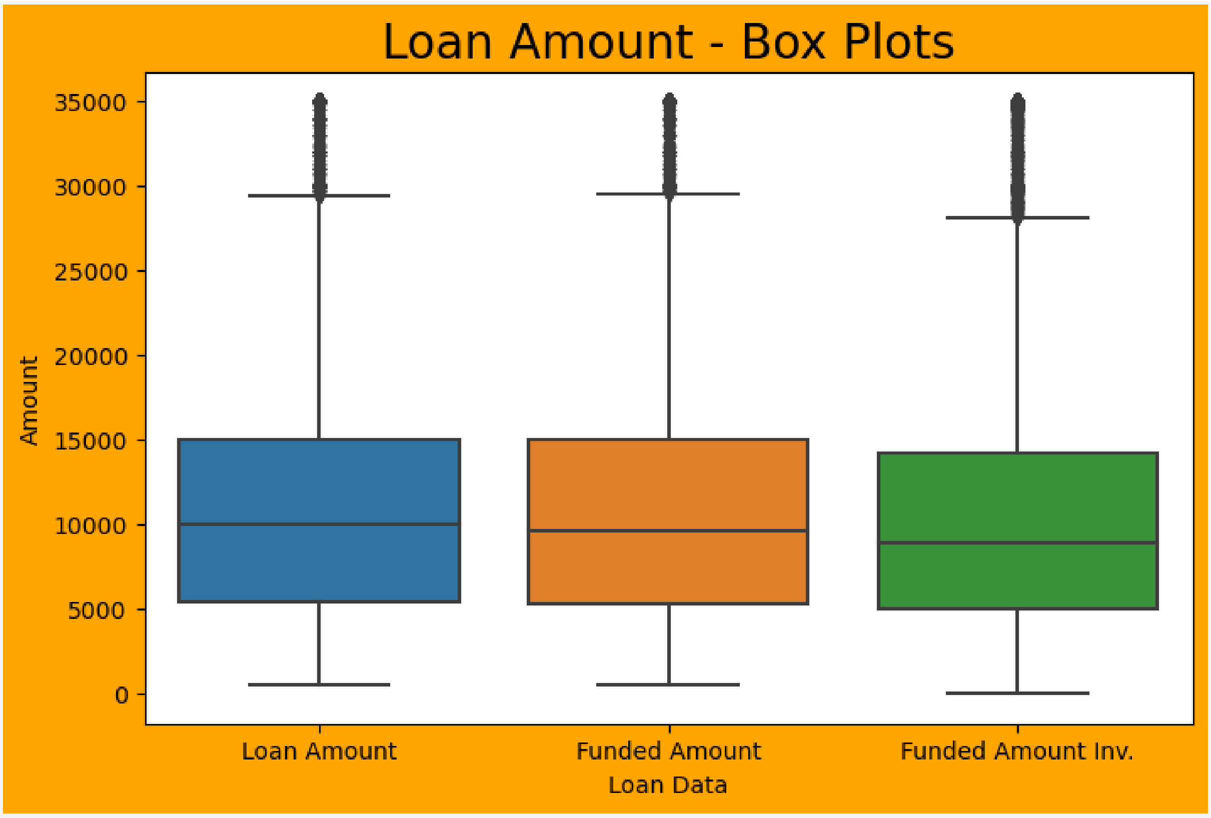
Univariate Analysis

Below are the important results and observations:

1. Univariate analysis - Loan amount, Funded amount and Funded amound inv.

   Data Dictionary:
   Loan amount - The listed amount of the loan applied for by the borrower.
   Funded amount - The total amount committed to that loan at that point in time.
   Funded amount inv. - The total amount committed by investors for that loan at that point in time.

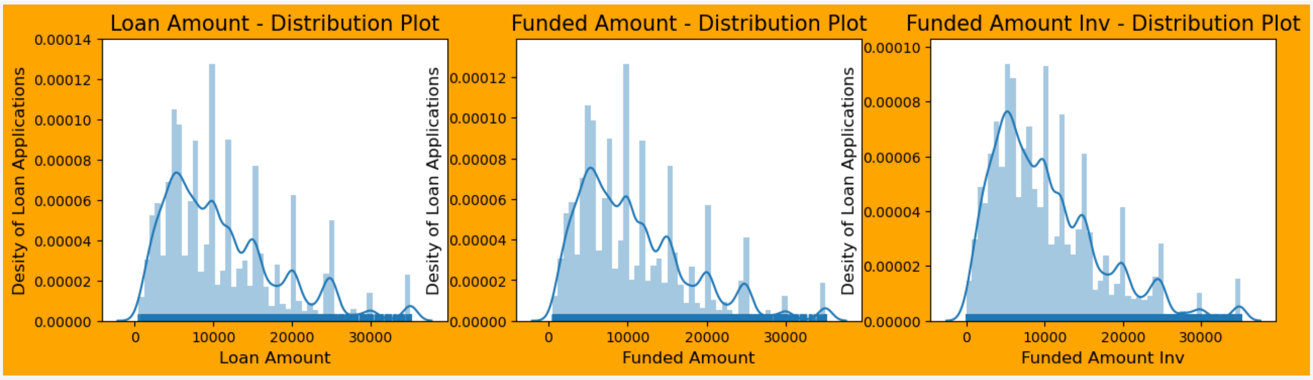
   Below are the box plots of loan amount, funded amound and funded amount inv. to identify the outliers.



   Observations:

- Loan amount, funded amount and funded amount inv. shows almost the same pattern of data spread in box plots.
- Data near whisker line are also closely placed to the line.
- We can conclude taht there are no outliers here.

Below are the distribution plots of Loan amount, Funded amount and Funded amount inv. after outliers removed from annual income analysis.

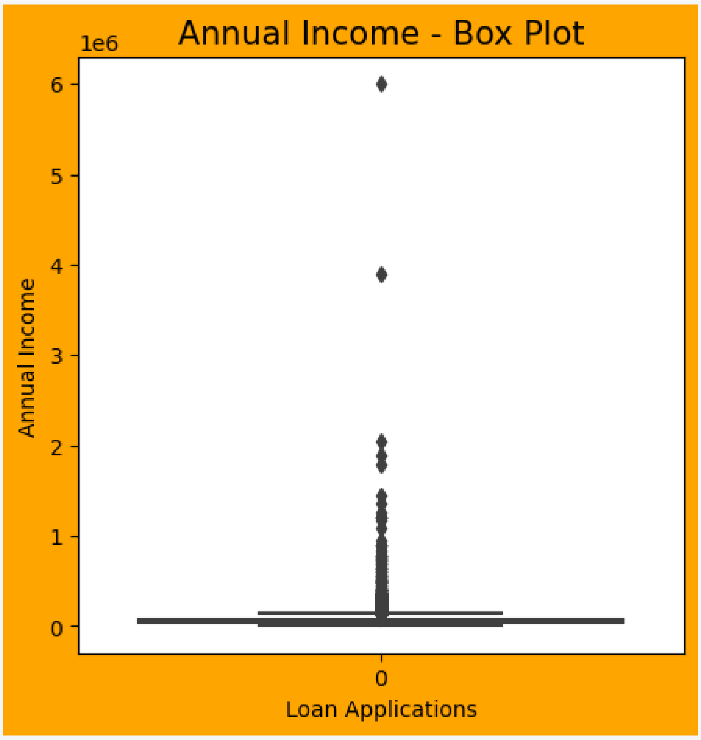


Observations:

- There is almost similar pattern of data spread for all the three parameters. Funded Amount Inv shows a little higher curve near 10000 but it can be ignored.
- We can consider Loan amount field for further analysis against loan amount.

2. Univariate analysis - Annual income

Annual Income box plot before removing outliers



Observations:

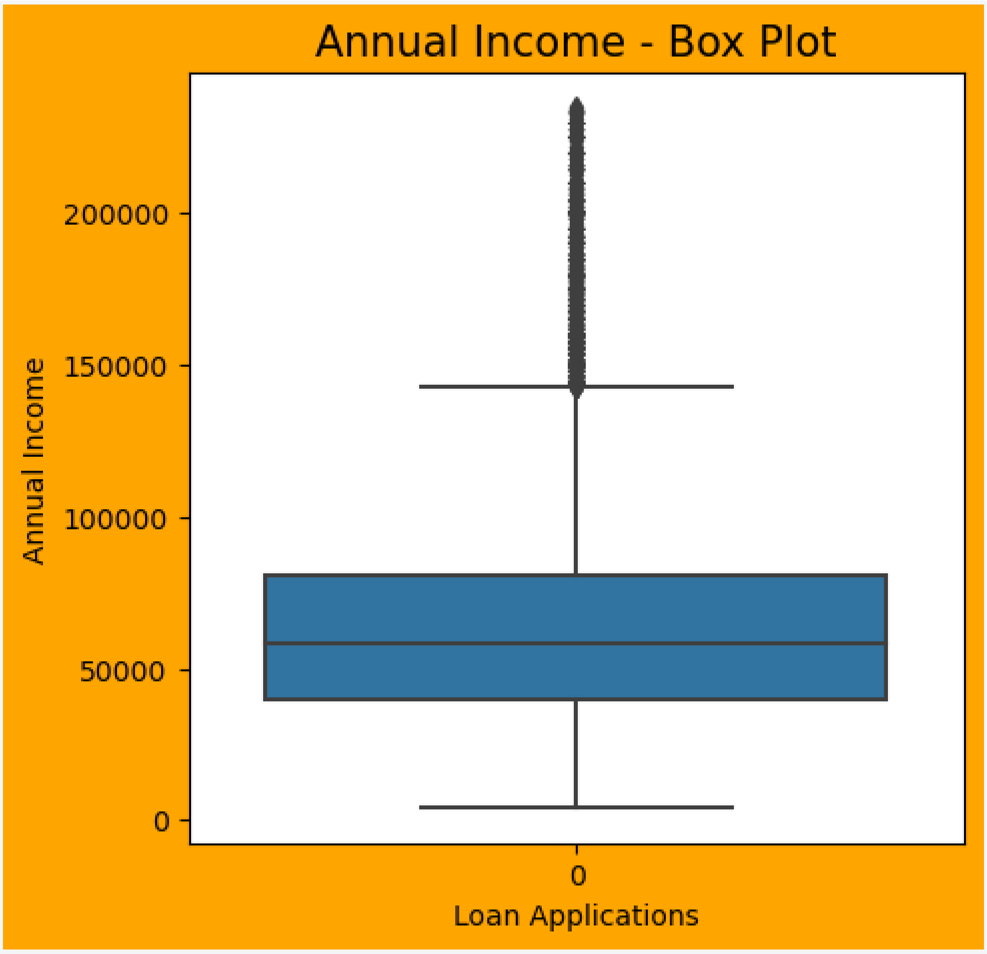
- From the above box plot we can say that there are outliers in annual income. Some applicant's annual income is shown unexpectedly high.
- This can impact the analysis adversely. So we need to remove outliers.

Data cleaning:

We dropped outlier rows. As part of it we have dropped 398 rows.

After data cleaning our new dataframe structure is **(39319, 48)**

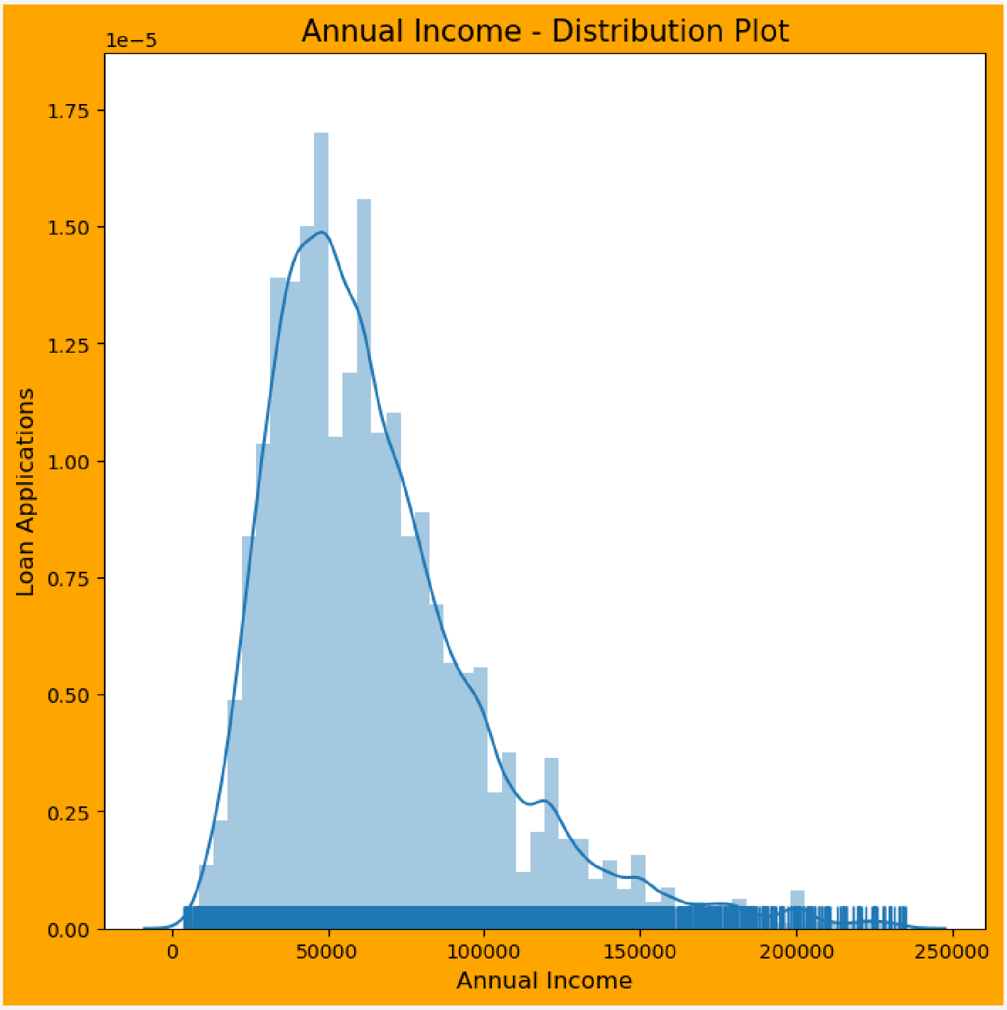
Annual Income box plot after removing outliers



Observations:

- After removing outliers we can see that most of the loan applicant's annual income is between 5000 - 10000

Annual Income - Distribution Pot

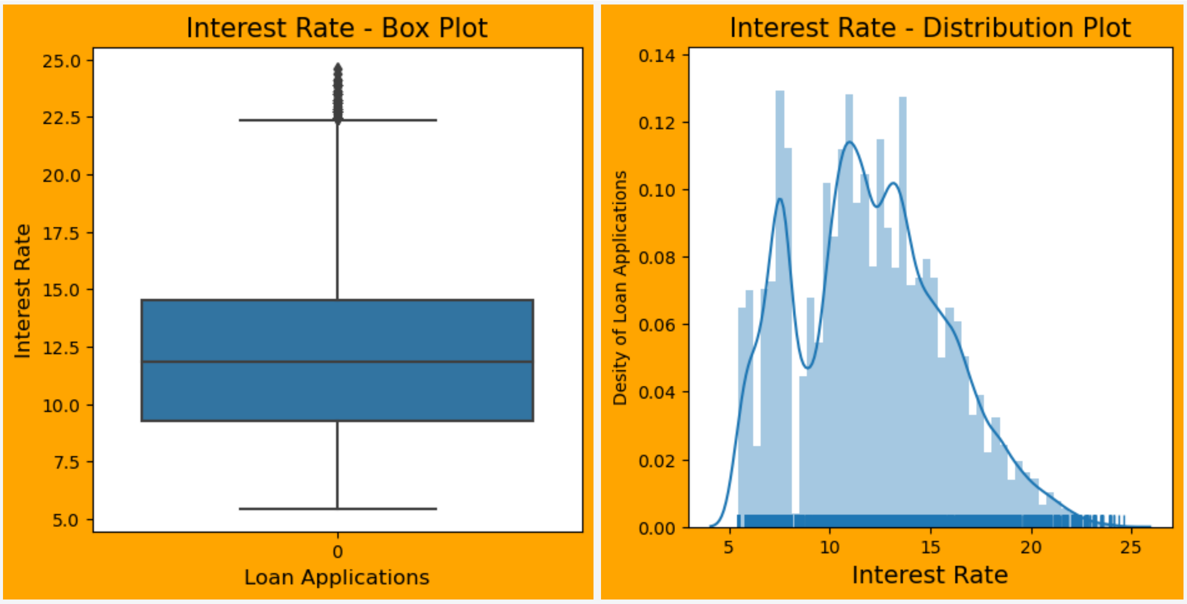


Observations:

- As we observed in box plot, most of the loan applicant's annual income is between 5000 - 10000.
- As the annual income is increasing, it shows the drop in loan applications or say requirements of the loan.

3. Univariate Analysis for Interest Rate

Below are the **Interest Rate** box plot and distribution plots:

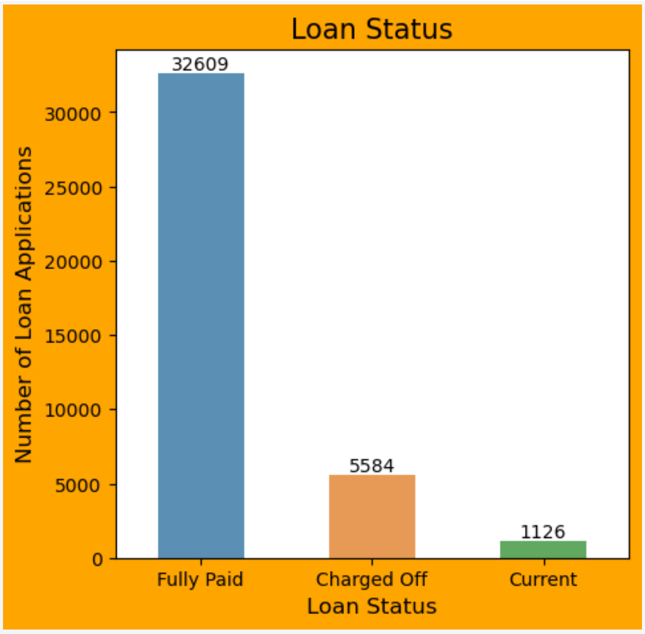


Observations:

- From box plot we can see there are some loans at higher interest rate between 22% - 25%. But we will not consider them outliers. It depends on the applicant's profile.
- Form box plot and distribution plot, we can see most of the loans interest rate lies between 9% - 15%.

4. Univariate Analysis - Loan Status

Loan Status - Bar chart

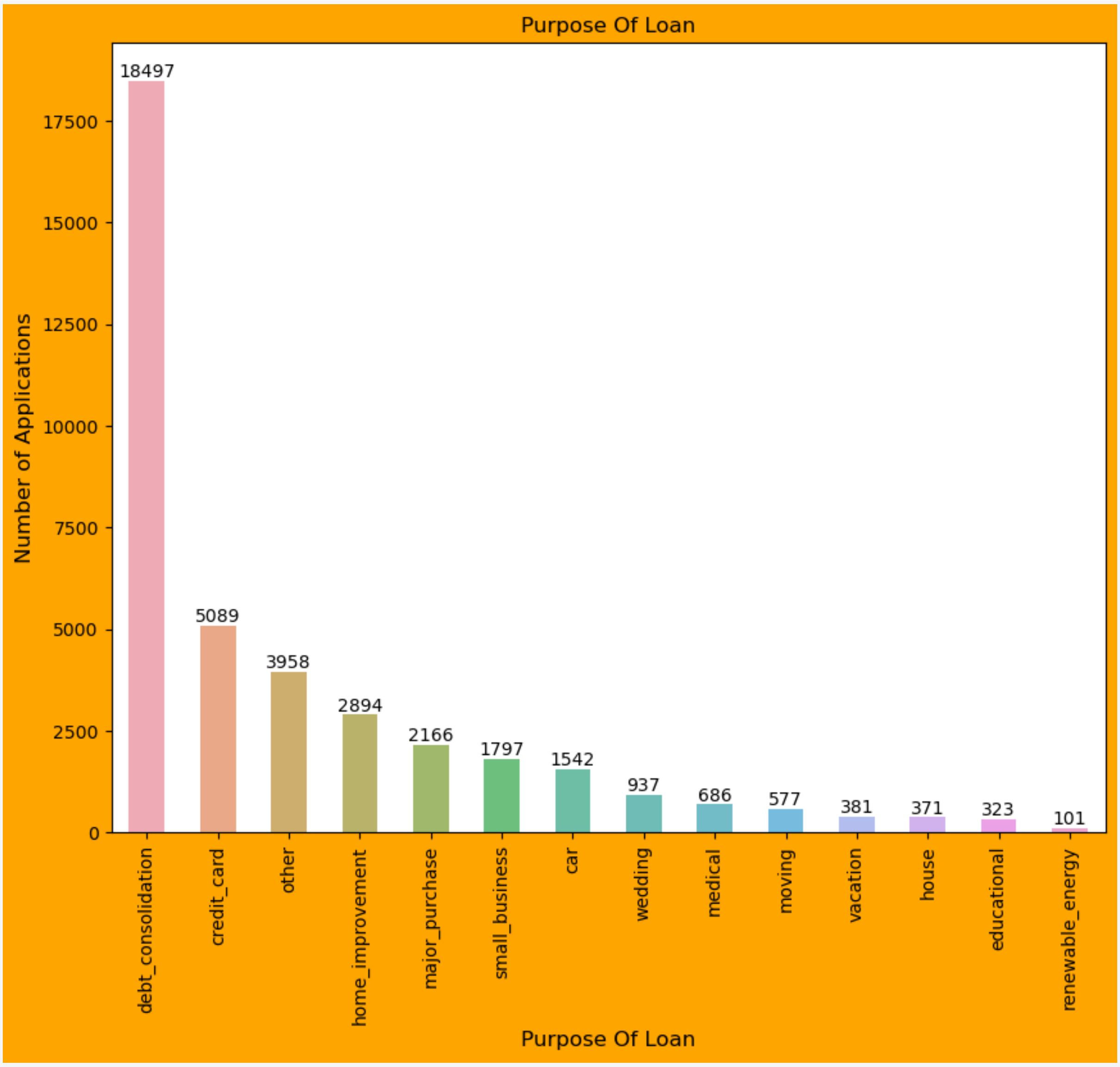


Observations:

- There are three types of loan statuses can be seen throughout the data.
- Maximum loans are fully paid, but there is a significant number of charged off loans also.
- Our analysis will be based around the behaviour of charged off loans vs different parameters. It will help in identifying the risky applicants who are more likely to default.

5. Univariate Analysis - Purpose (categorical unordered variable)

Below is the bar chart for **Purpose** of the loan:

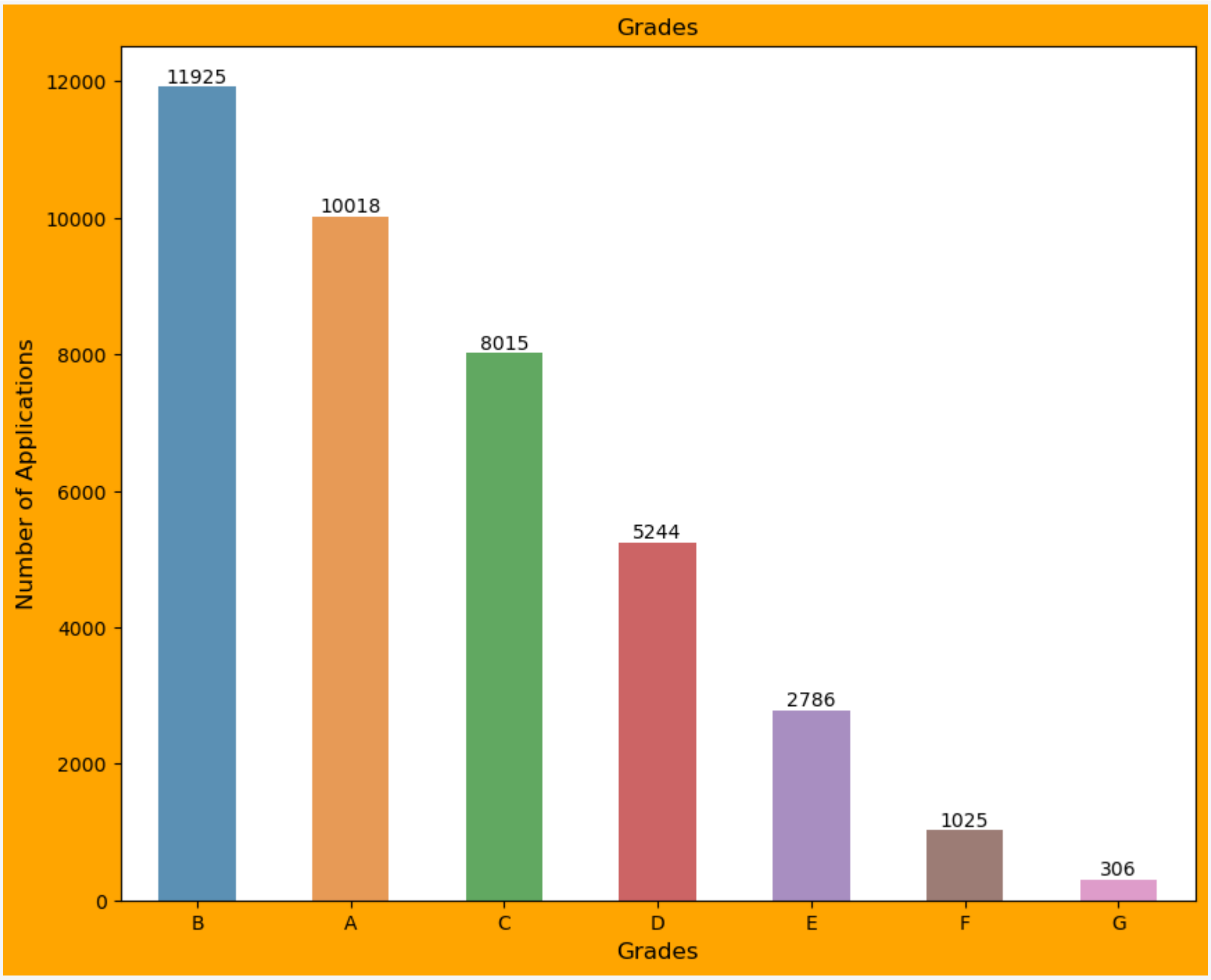


Observations:

- Maximum loans are taken for debt consolidation and credit cards bills repayment.

6. Univariate Analysis - Grade (categorical ordered variable)

Grades - Bar Chart

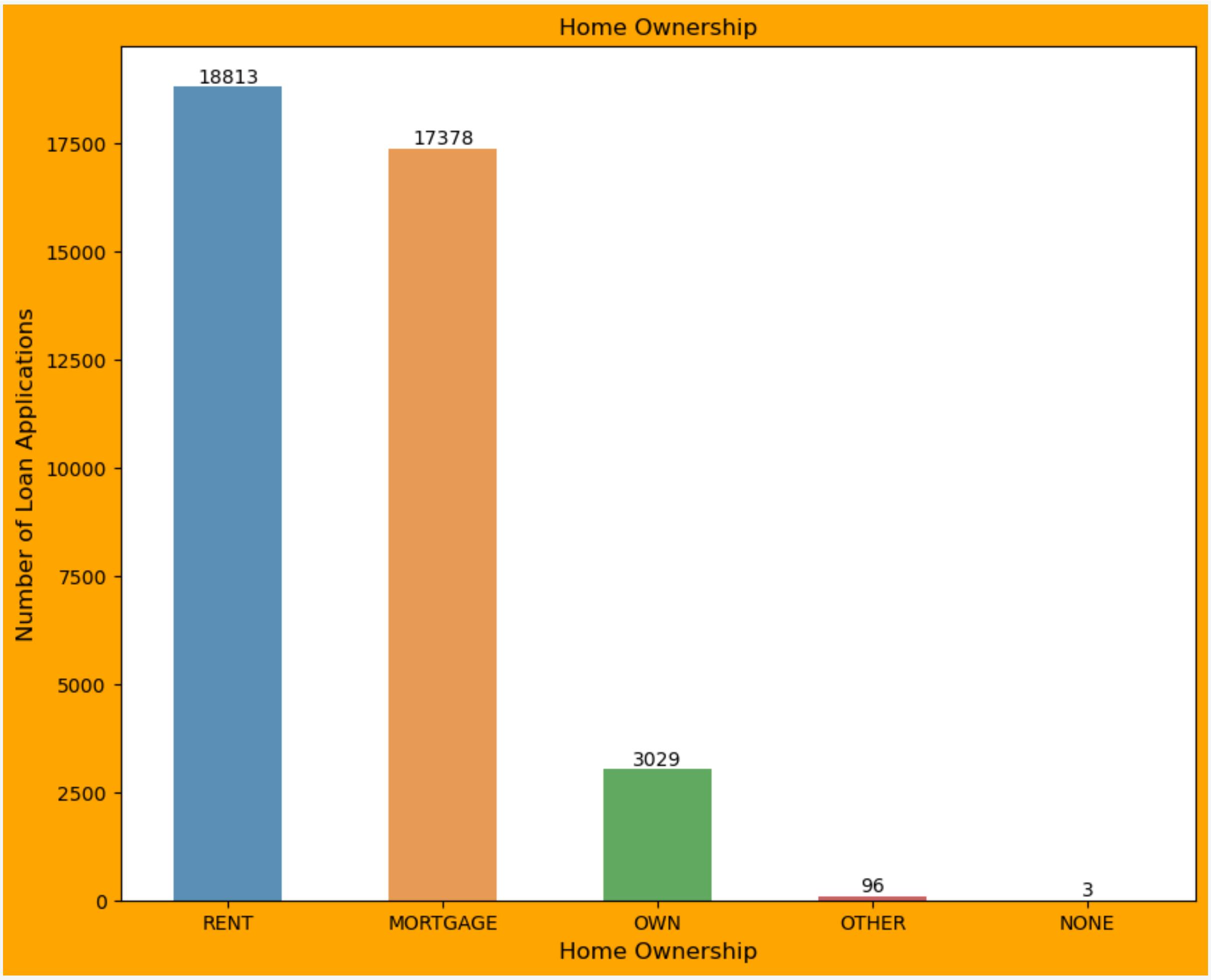


Observation:

- Grades are directly related to the interest rate. As you move from A to G, interest rate also increases.
- It is exactly aligned with the distribution chart of interest rate, lwhere loans are less for less than than 9% rates and greater than 15% rates.
- So we can say Grade A applies to lower interest rates, B is for 9% -15% interest rates and so on.

7. Univariate Analysis - Home Ownership (Categorical Unordered Variable)

Home Ownership - Bar chart



Observations:

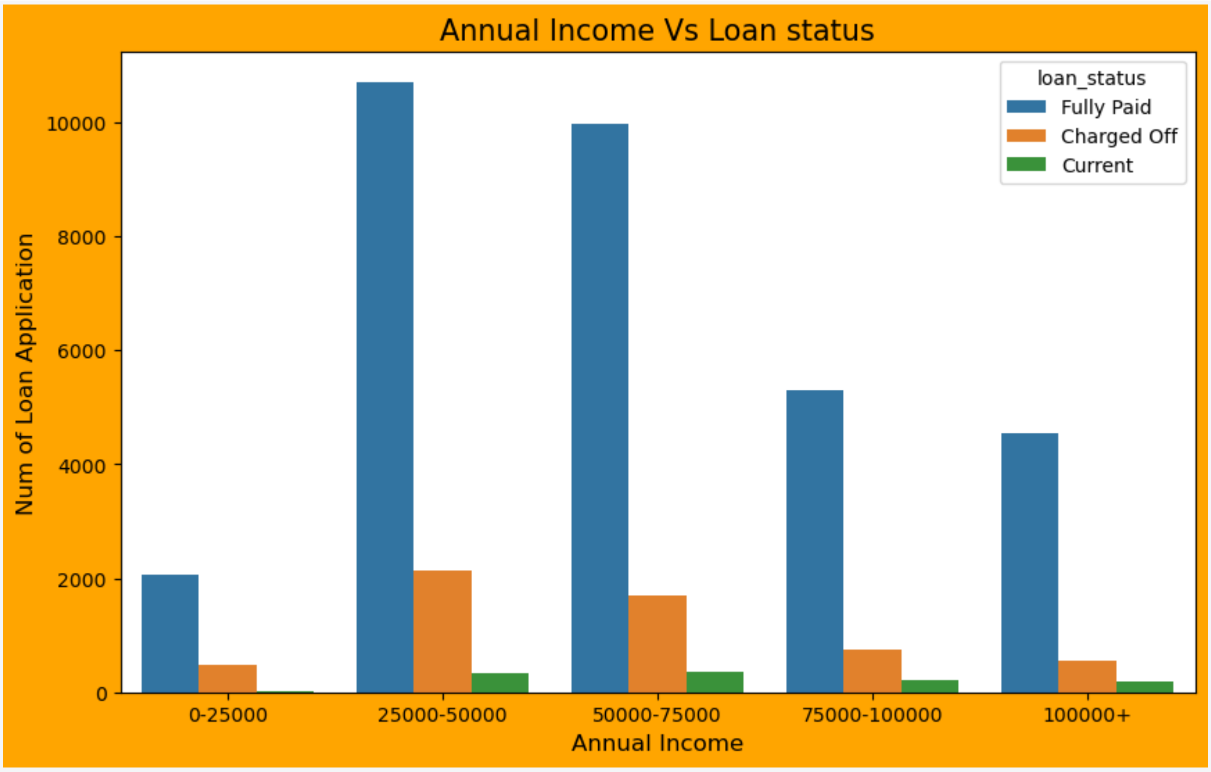
- Maximum applicants are staying in rented or mortgazed houses.

Segmented Univariate Analysis

Below are the important results of Segmented Univariate Analysis

1. Segmented Univariate Analysis- Annual Income vs. Loan Status

Annual income is segmented in buckets to analyse against Loan Status

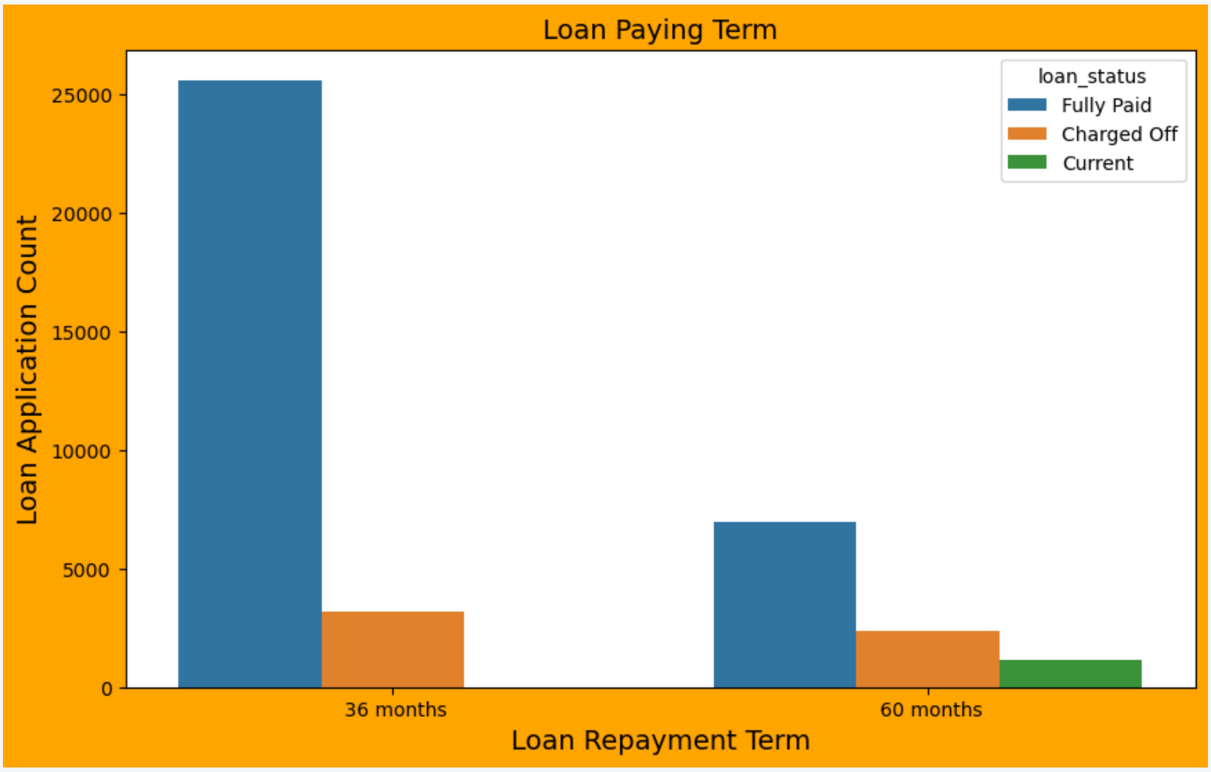


Observations:

- Compared to distribution chart, it is more clear here that very few loan are approved for lower income applicants.
- Maximum loans are approved for the income range 25000-50000 and 50000-75000.
- As the income increases, it can be assumed that those applicants are less.

2. Segmented Univariate Analysis - Term vs. Loan Status

Term is segmented into two buckets to analyze against Loan Status

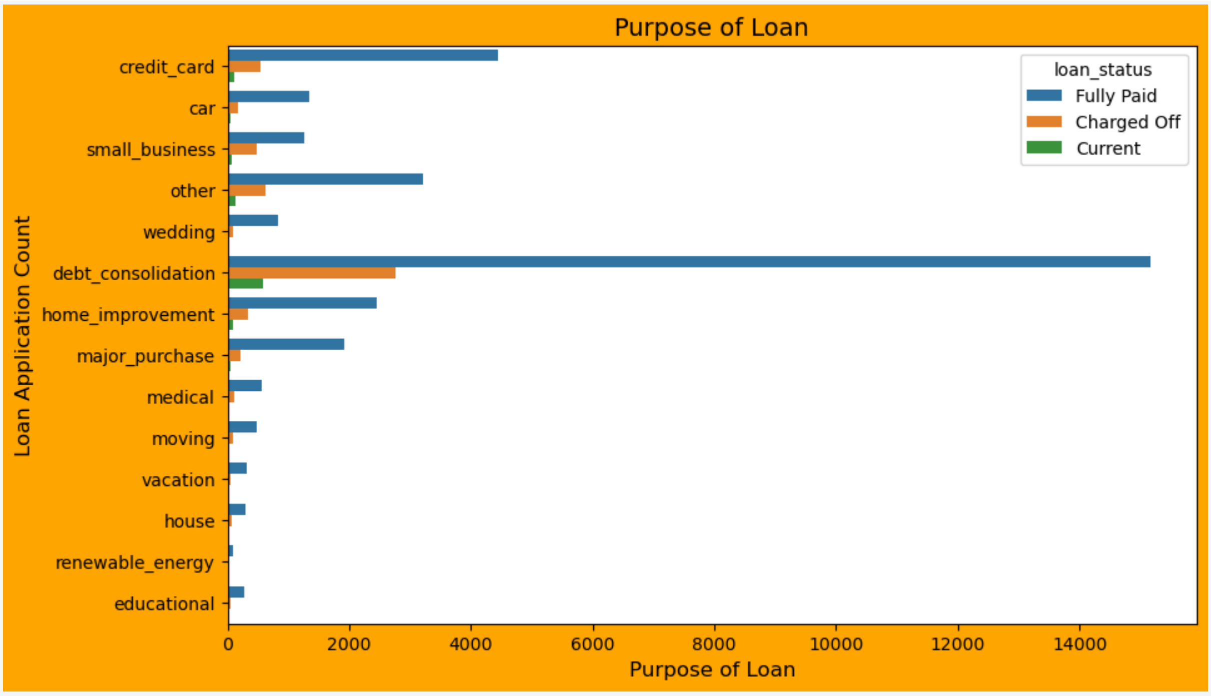


Observations:

- It shows longer the term for repayment, more are the chances to default the loan.
- Here with 60 months of repayment duration has more Chareged Off laons compared to the total loans in that segment.

3. Segmented Univariate Analysis - Purpose vs. Loan Status
4.

Purpose is segmented by its categories and analyzed against Loan Status



Observations:

- Maximum loans are taken for debt consolidation and creadi card bills payment.
- So as the numbe of charged-off loans are also high for both.
- If loan is taken for settling other dues or debts, there are high chances to be charged-off.

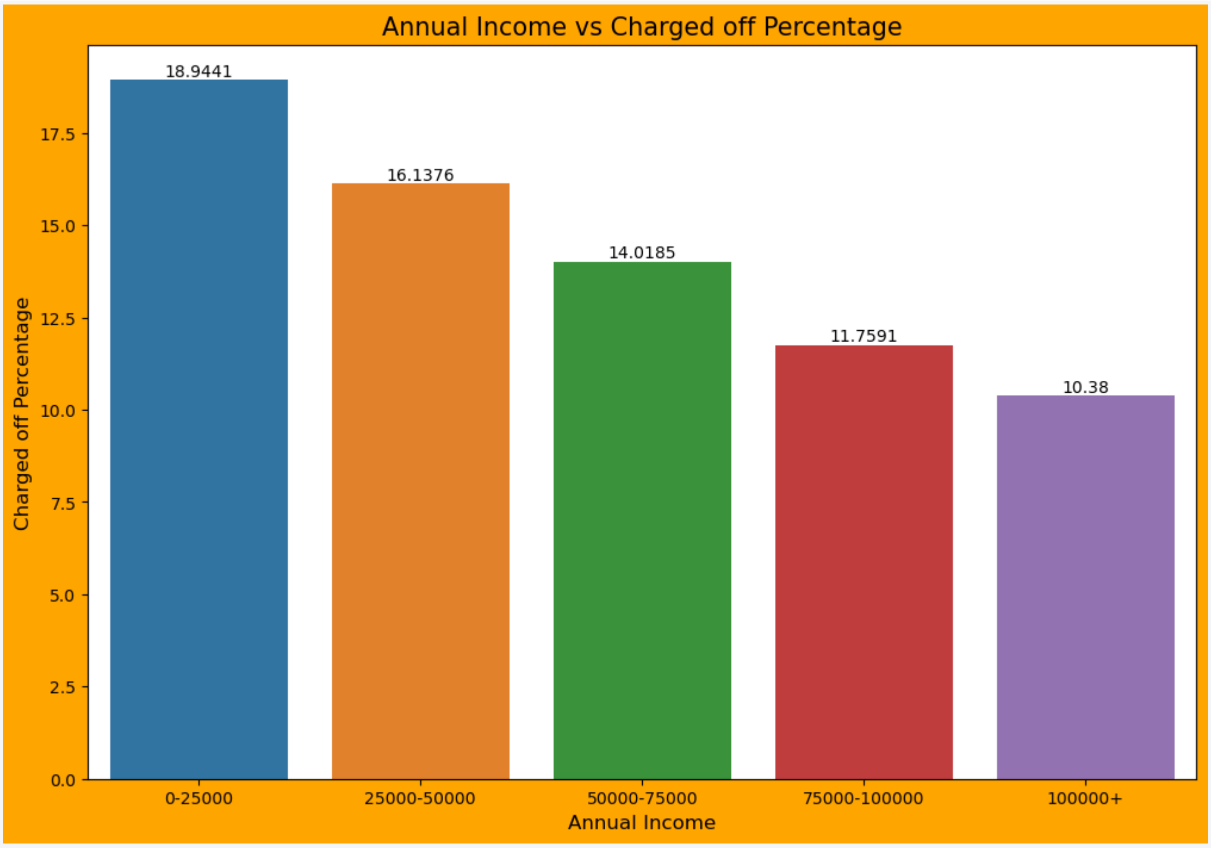
Bivariate Analysis

Note: Derived columns/fields
For bivaraite analysis we derived some new fiends from existing fields to divide them in smaller buckets and labled them to analyze them against Charged-off percentage. We derived these columns from Annual Income, Loan Amount and Interest Rate columns.

Below are the important results of Bivariate Analysis:

1. Bivariate Analysis - Annual Income vs. Charged-off Percentage

Annual Income vs. Charged-off Percentage

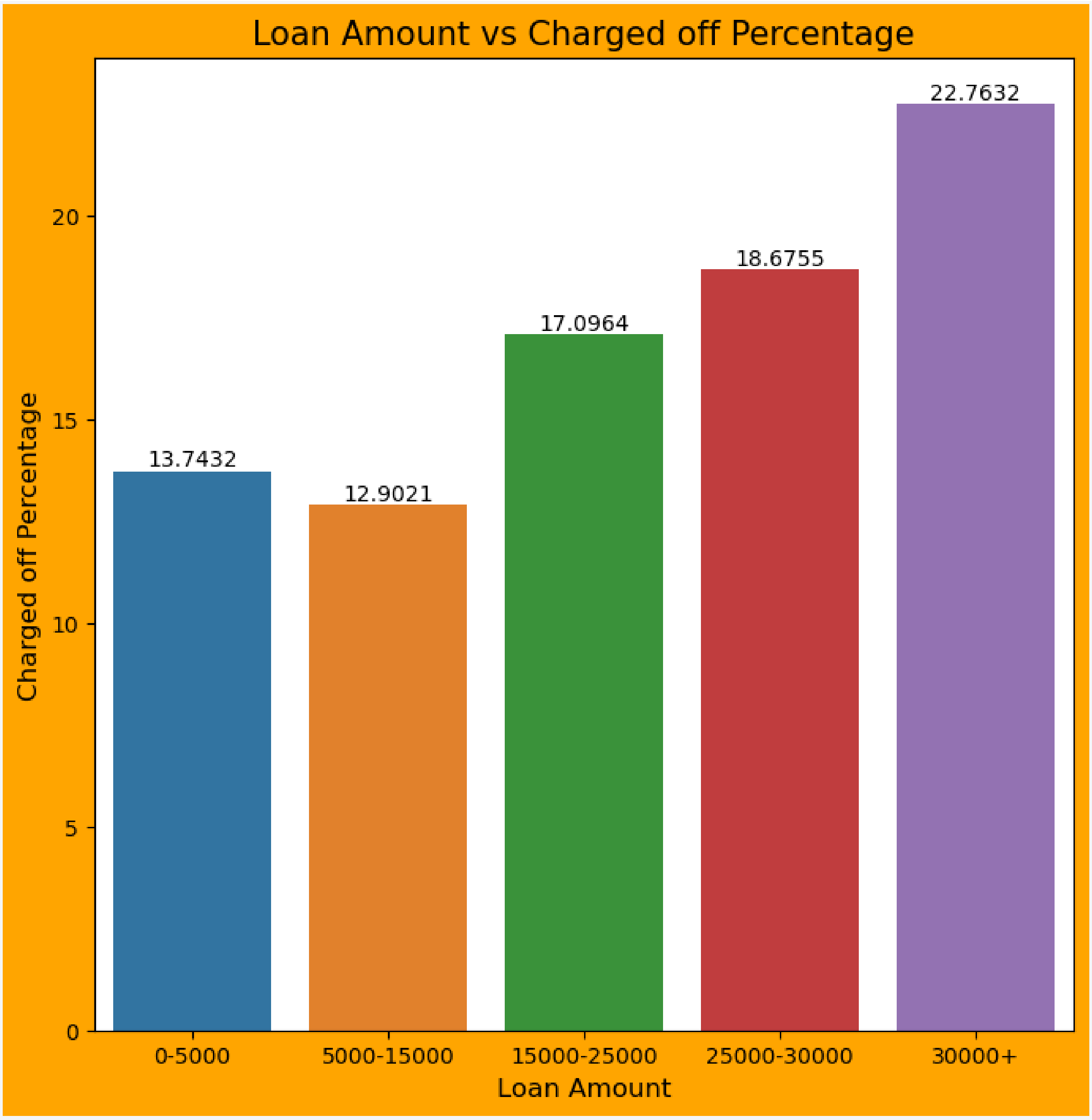


Observations:

- It is clearly visible that there are more applicants in the income range of 0-25000 (there is no zero income for any applicant) who default to repay the loan compared to higher income ranges.
- As the annual income increases the probability of defaulting the loan is decreasing.
- It is safe to lend money to applicants with higer annual income

2. Bivaraite Analysis - Loan Amount vs. Charged-off Percentage

Loan Amount vs. Charged-off Percentage



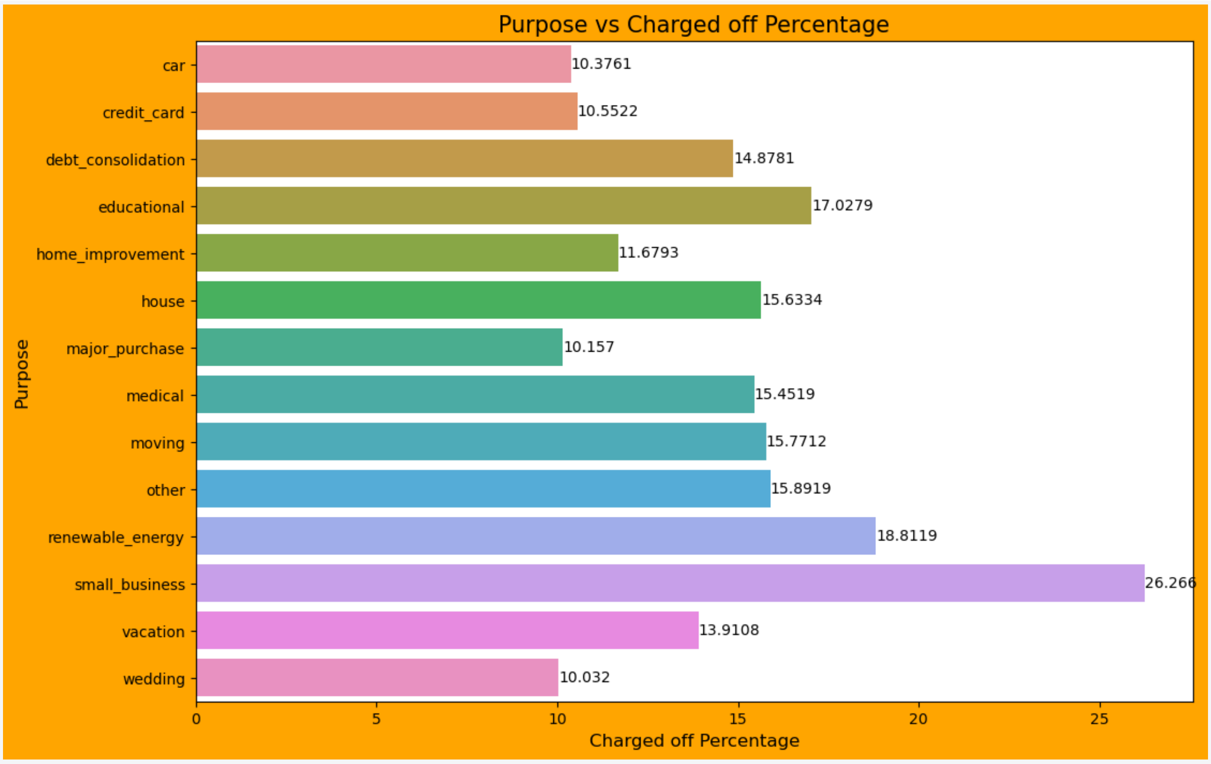
Observations:

- As the loan amount is increasing the charged-off rate is also increasing. For now data is very less on higher loan amount like 30000+ loan amount to conclude.
- But if we compare '5000-15000' and '15000-25000' above observation stands true.

- 0-5000 segment is showing more charged off percentage than 5000-15000, which is significant. These loan would be given to the applicants with low annual income and it is charged-off for low loan amount also.

3. Bivariate Analysis - Purpose vs. Charged-off Percentage

Purpose vs. Charged-off Percentage

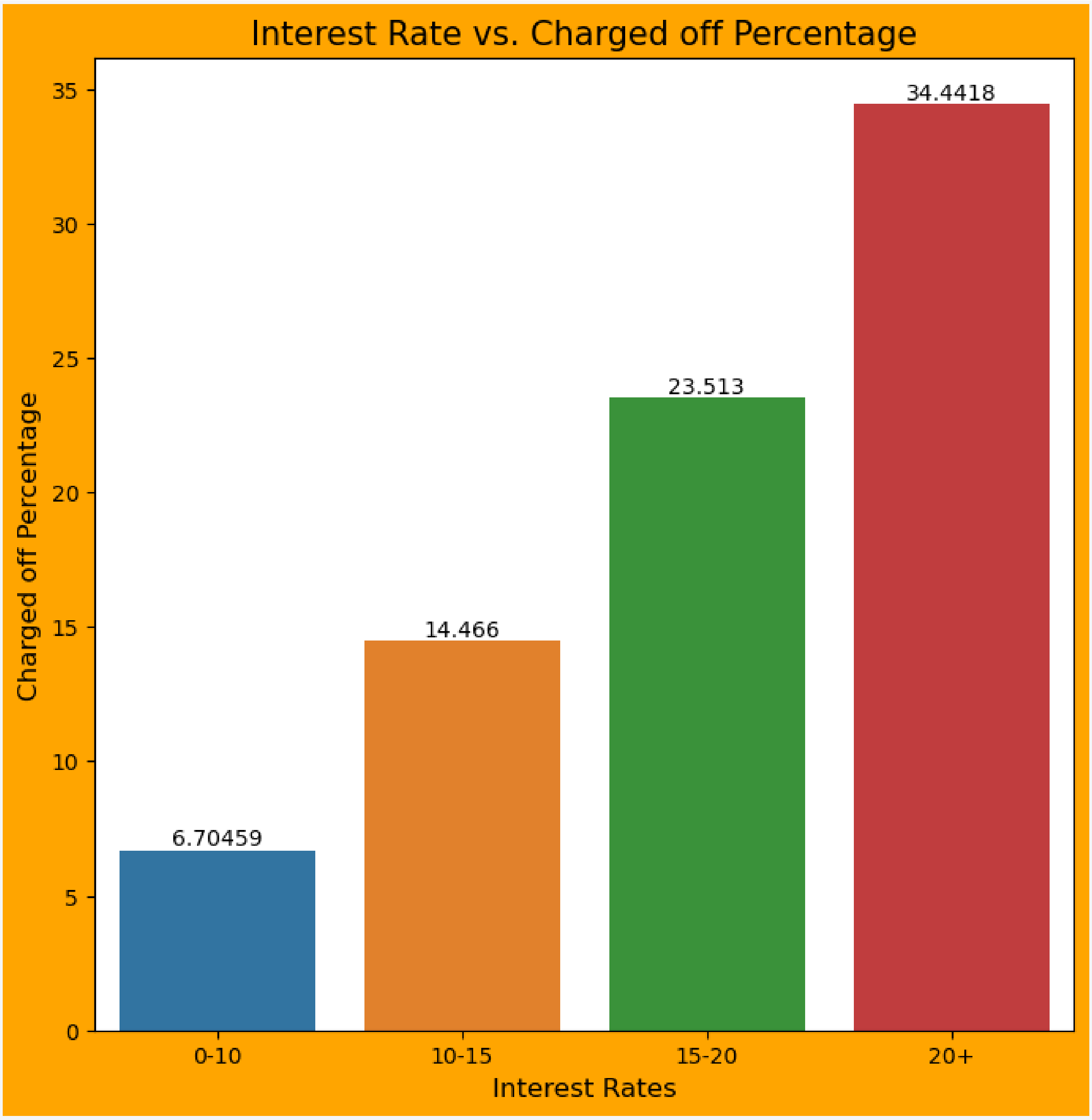


Observations:

- Small business loans are charged off more. That means small businesses taking loans are at higher risk to be charged-off due to not being able to make immediate returns on their investments or other reasons. But business loan data is too small to conclude.
- Compared to other reasons, "debt_consolidation", "Other" are more applicants and their percentage of charged off are close to 15%, which shows more risk than other purposes. Also the purpose is to pay off the existing debts which is shows the concern about their repayment capability.
- Car, credit card, major purchases and wedding looks more genuine reasons as their percentage of getting charged off is less.

4. Bivariate Analysis - Interest Rate vs. Charged-off Percentage

Interest Rate vs. Charged-off Percentage

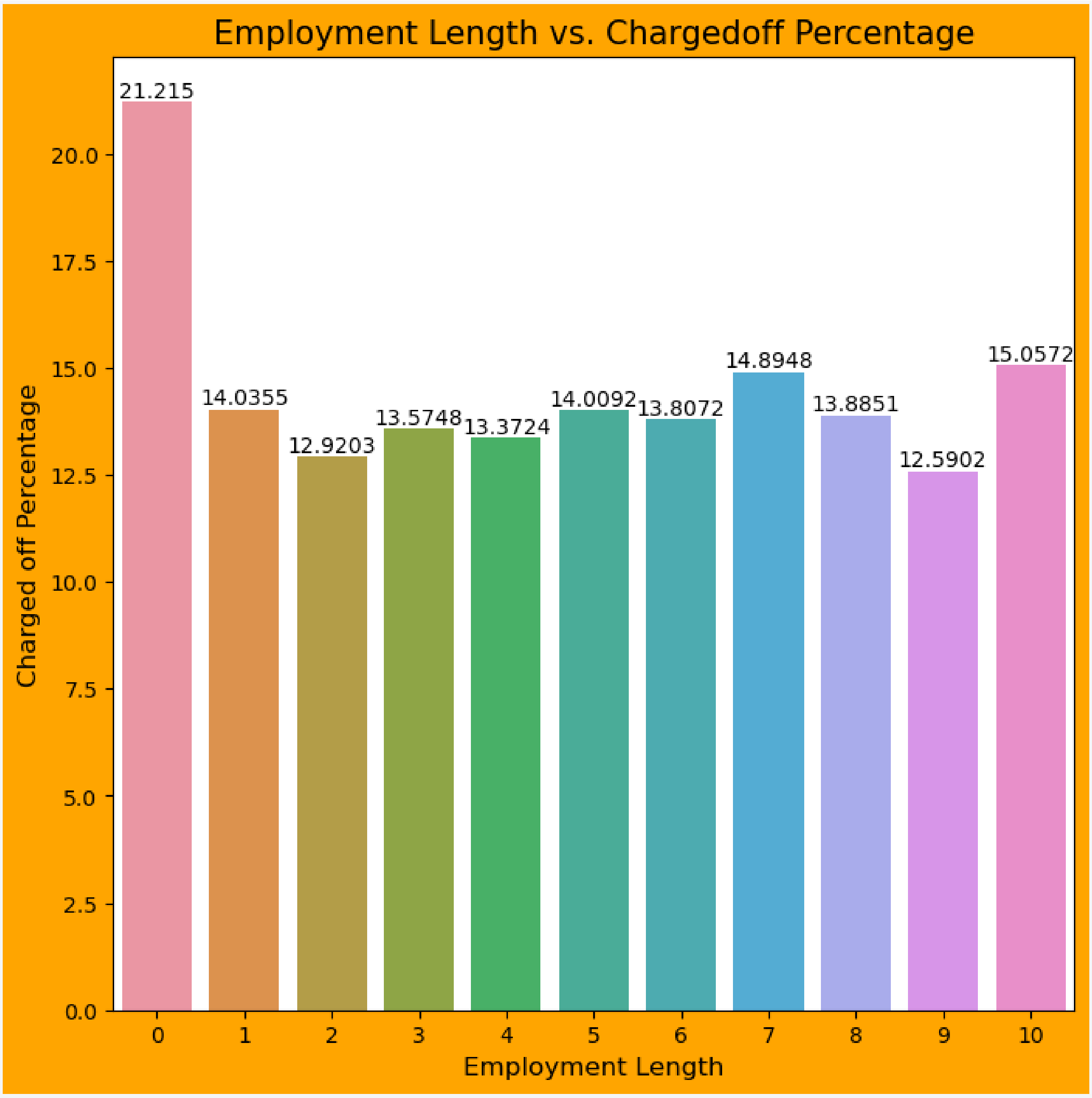


Observations:

- It clearly shows that with the increase in interest rate, the risk of charge-off is also increasing.
- Above 15% interest rate, charged off loans rate is drastically increasing.

5. Bivaraite Analysis - Employment Length vs. Charged-off Percentage

Employment Length vs. Charged-off Percentage

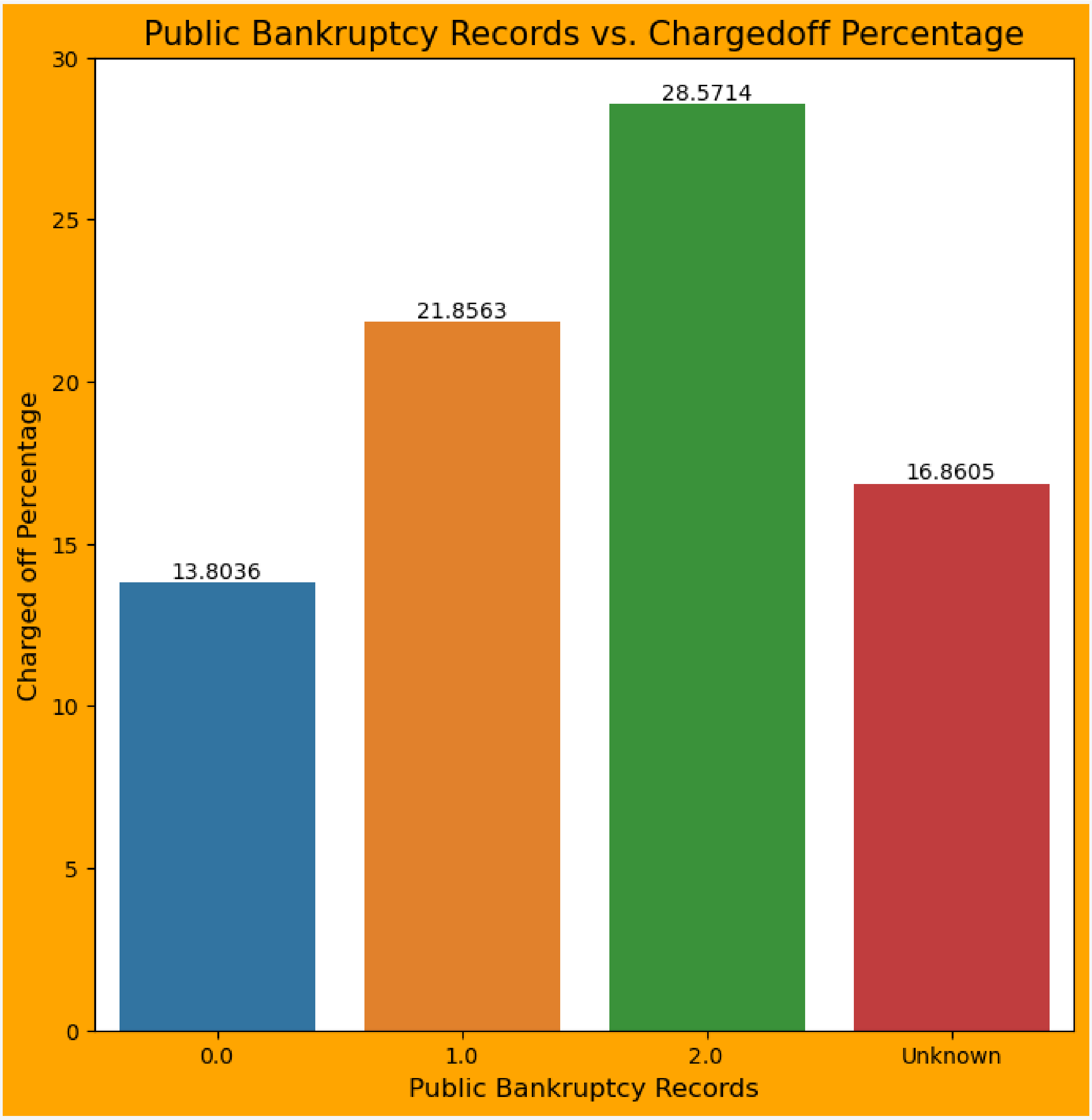


Observations:

- If applicant's employment time zero or less than one month, he/she would not be earning and so the chances of charged-off the loan are more.
- Here it is clear that charged off rate is high for those applicants with less than one month of employment.
- For others it is more or less equal chances of charged-off.

6. Bivariate Analysis - Public Bankruptcy Records vs. Charged-off Percentage

Bankruptcy Records vs. Charged-off Percentage



Observations:

- People who has 0 bankruptcy also defaults the loan. Number of loans are more with 0.
- People with 1 and 2 are having higher percentage of default the laon but numbers are low to conclude.
- But there are high chances of default for those who has record of bankruptcy.

Overall Observations

Important observations to approve/reject the loans applications:

1. Lower the annual income, higher the chances of charged-off .
2. Higher the loan amount, higher the chances of charged-off.
3. Higher the interest rate, higher the chances of charged-off.
4. Higher the repayment term, higher the chances of charged-off.
5. When purpose is to settle the other debts then chances are high that the loan will be charged-off.
6. For new businesses also, possibilities are high to default as ROI is not guaranteed on new businesses.
7. Applicants with no employment history may have more chances of default the loan.
8. Applicants having bankruptcy record are having more chances to default the loan.

Glossary

Deatails of Data Used In Analysis:

Source of Data: loan.csv
Format: csv
Number of Rows: 39717
Each row is: giving the details of the existing loan applican't characteristics.
Sampling Method: All the loan applications between Apr-2008 and Sep-2011

Technologies Used For Analysis:

Programming Language: Python, version: 3.10.9
Supporting Libraries: Numpy, verson: 1.23.5
Pandas, version: 1.5.3
Charting Tools: Matplotlib, version: 3.7.0
Seaborn: 0.12.2