Lending Club Case Study

Exploratory Data Analysis

Collaborators:

1. Sameer Ganeshe
2. Saumy Dholu

Index

- Problem Statement
- Data Summary & Libraries used
- Data Understanding & Cleansing
- Date Conversion & Creating Derived Columns
- Performing Outlier analysis to remove outliers
- Univariate, Segmented & Bivariate Analysis
- Correlation Analysis

Problem Statement

Business Understanding:

You are working for a consumer finance company that specializes in providing various types of loans to urban customers. The company faces a critical decision-making process when evaluating loan applications, balancing two primary risks:

- **Loss of Business**: If a loan is not approved for an applicant who is likely to repay, the company misses out on potential business.
- **Financial Loss**: If a loan is approved for an applicant who is likely to default, the company incurs a financial loss.

Objective:

The dataset provided includes information on past loan applicants and their repayment status. The objective is to identify patterns that predict the likelihood of default, which can inform decisions such as loan denial, loan amount adjustment, or offering loans at higher interest rates to riskier applicants.

Perform Exploratory Data Analysis (EDA) to understand how consumer and loan attributes influence default tendencies.

- When a loan is approved, it can result in one of three outcomes:
  - fully paid,
  - currently being paid
  - charged-off (defaulted)
- If a loan is rejected, there is no transactional history available for those applicants.

The goal is to develop a model that helps the company make informed loan approval decisions to minimize financial risks and maximize business opportunities.

Data Summary & Libraries used

- ✓ Used “loan.csv” .csv file containing 39717 rows & 111 columns.

- ✓ Majorly the data consist of 2 types of attributes
  - Loan Attributes
  - Data Attributes

- ✓ Libraries used in this Exploratory Data analysis are:
  - pandas
  - numpy
  - datetime
  - matplotlib
  - seaborn
  - warnings

Data Understanding & Cleansing

- ✓ No header, footer, summary, total, sub-total rows found
- ✓ No duplicate rows found.
- ✓ 54 columns with 100% null values, dropped.
- ✓ 4 more columns with more than 30% null values, dropped, which could adversely impact the data analysis.
- ✓ Additionally, dropped the below 12 columns for the reason mentioned below -
  - ✓ 4 columns with only single value → “pymnt_plan”, “application_type”, “policy_code”, “initial_list_status”
  - ✓ 3 columns with only “0.0” values → “tax_liens”, “chargeoff_with_12_mths”, “collections_12_mths_ex_med”
  - ✓ 2 columns with only “0” values → “delinq_amnt”, “acc_now_delinq”
  - ✓ “url” column → did not provide any significant insight.
  - ✓ “emp_title” → holds 28820/38717 unique values, was not of use in EDA.
  - ✓ “sub_grade” → did not furnish any necessary information, and we already had grade to use for EDA if need be.

Data Understanding & Cleansing

- ✓ Invalid Values, Precision check & fixing incorrect data-types in Columns
  - ✓ Removed “%” and converted to float → “int_rate”, “revol_util”
  - ✓ Removed “months” and converted to int64 → “term”
  - ✓ Rounding off to 2 decimal places for precision → “total_pymnt”, “funded_amnt_inv”
  - ✓ Removed “>”, “+” & “years” and converted to int64 → “emp_length”

- ✓ Dropping / Imputing null values in rows
  - ✓ Columns “emp_length” & “pb_rec_bankruptcies” still had 2.7% & 1.8% resp. null values respectively, dropped those rows as it is less than 5 % of the total dataset.
  - ✓ Data-set count reduced to 37945 records dropping 4.46% of records.
  - ✓ Data-set records with loan status “current” is irrelevant for EDA, as the relevant records will be of status “Fully Paid” & “charged-off”. Hence dropped another 1098 records.

Date Conversion & Creating Derived Columns

- ✓ Changed to date format for below 4 columns
  - ✓ “issue_d”
  - ✓ “earliest_cr_line”
  - ✓ “last_payment_d”
  - ✓ “last_credit_pull_d”
- ✓ Created 2 new derived columns from “issue_d”
  - ✓ “issue_month’
  - ✓ “issue_year”

Performing Outlier analysis to remove outliers

- ✓ 15 Columns used for outlier analysis –
  - ✓ “loan_amnt”, “funded_amnt”, “funded_amnt_inv”, “installment”, “annual_inc”, “dti”, “delinq_2yrs”, “inq_last_6mths”, “revol_bal” “revol_util”, “total_acc”, “total_pymnt”, “total_pymnt_inv”, “recoveries”, “last_pymnt_amnt”

- ✓ No outliers found for “dti”, “revol_util”. Hence performed outlier removal after removing these 2 columns from the list.

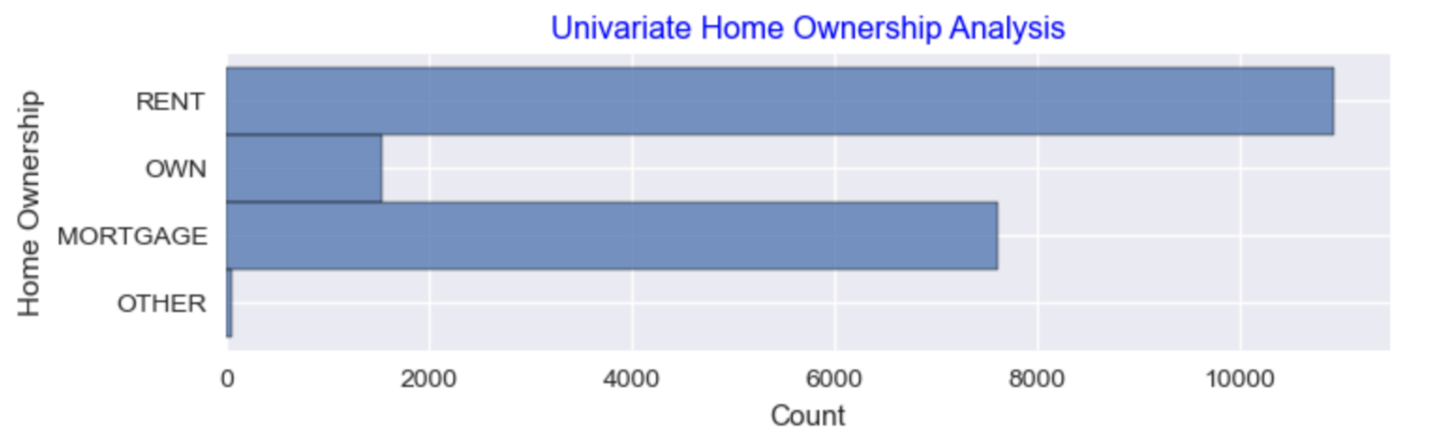
- ✓ Post removing outliers, the remaining dataset have 20129 rows & 43 columns.

- ✓ Next executed Reset index command before starting the univariate, segmented univariate, bivariate and correlation analysis.

Home Ownership Analysis – Univariate & Segmented Univariate

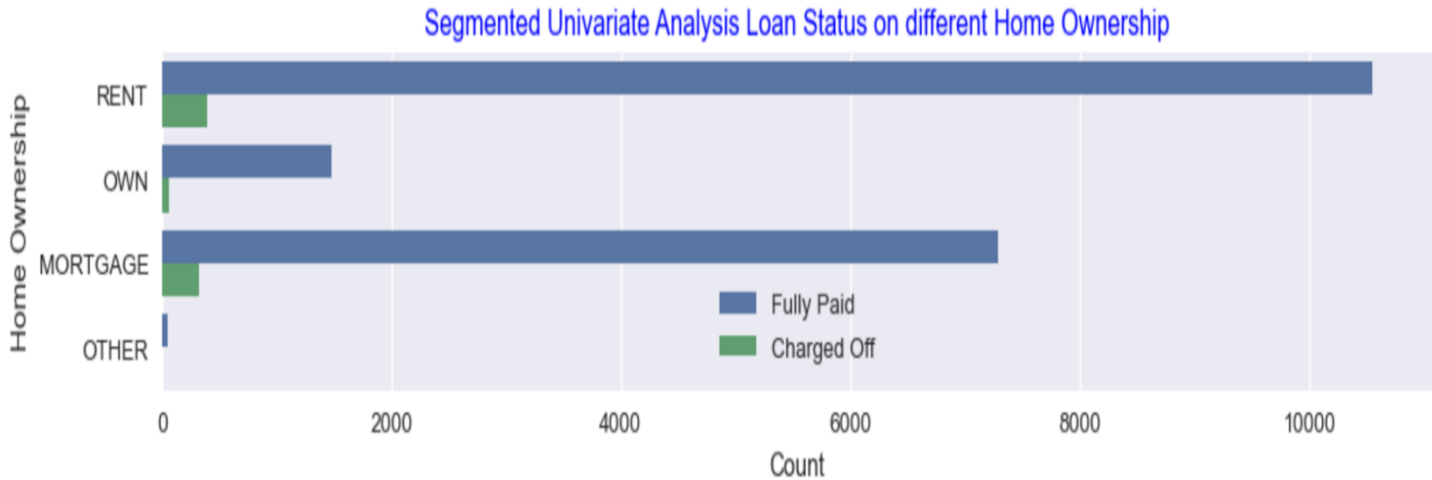
Observation 1:

Majority of Loan Borrowers don't posses property and are either having mortgage or re staying on rent.



Observation 2:

The Charged Off cases are lower incase of loan borrowers who owns a property compared to those who stay on rent or mortgage.

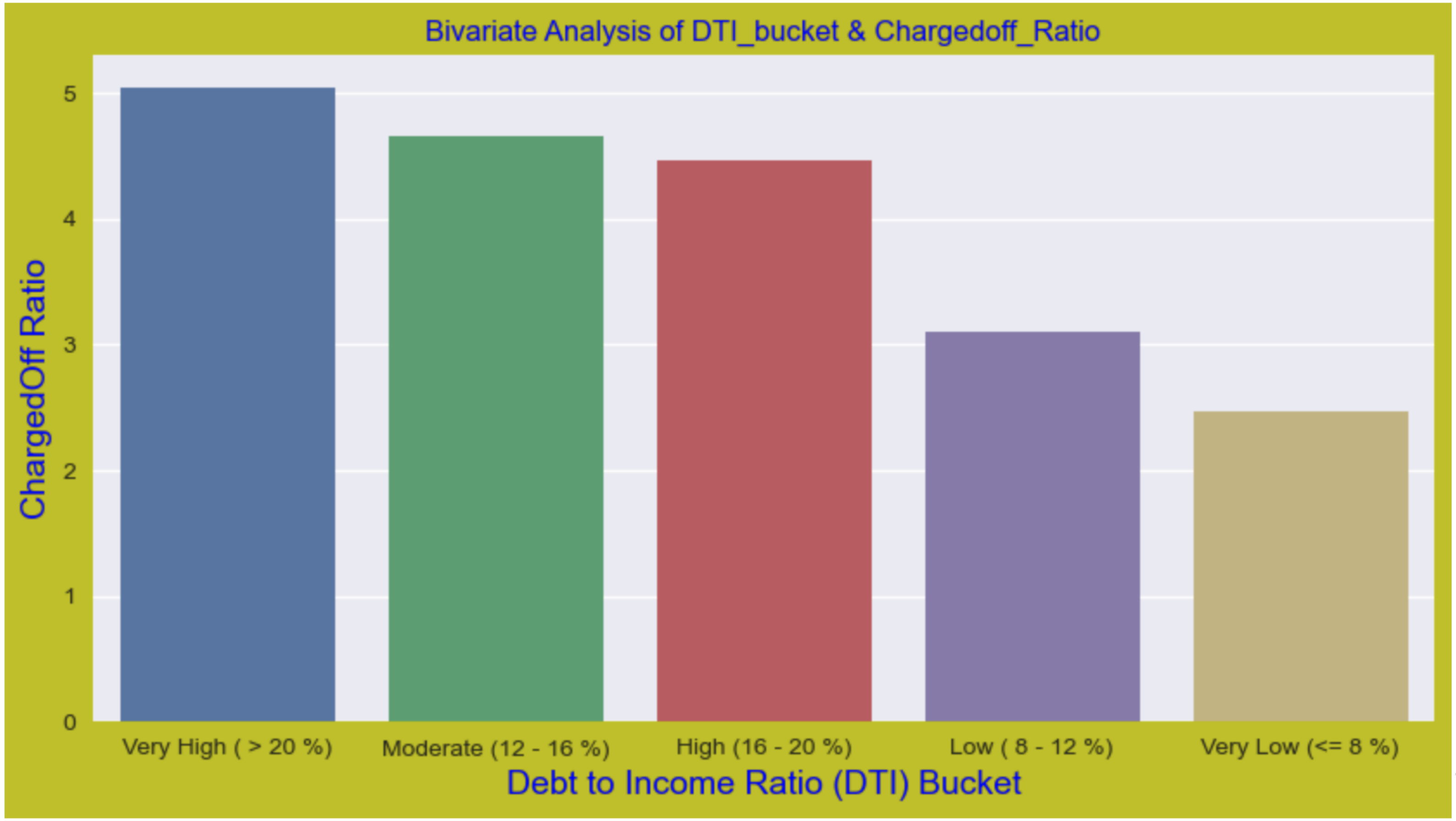


Home Ownership Analysis - Bivariate

Observation

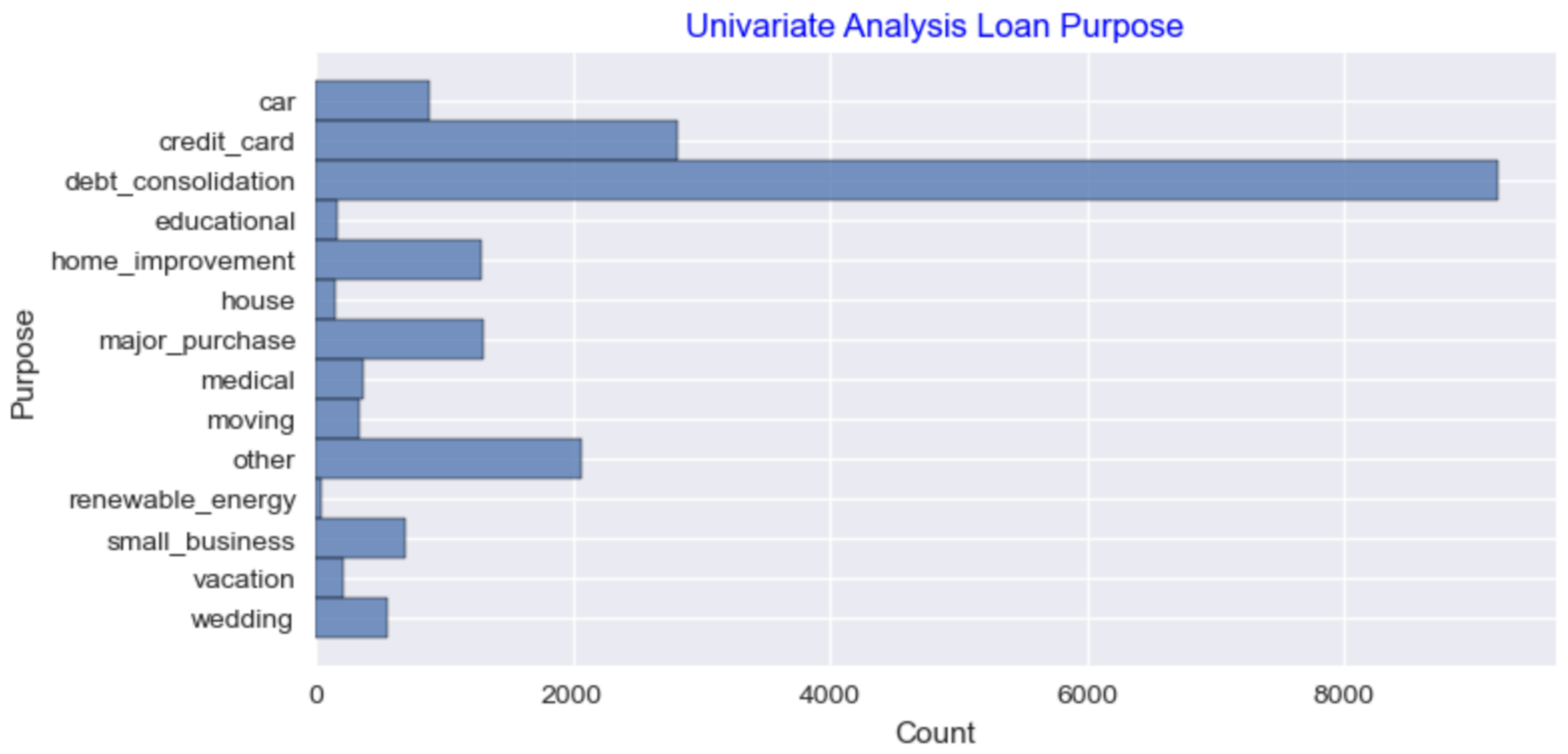
- The chart shows that 'home ownership' has little impact on charged-off ratios, with all categories showing similar default rates.
- Having a mortgage might indicate higher financial obligations, which could increase the risk of default.
- However, homeownership itself doesn't seem to be a strong predictor of default.

home_ownership	Charged Off	Fully Paid	Chargedoff_Ratio
MORTGAGE	326	7285	4.28
OTHER	2	50	3.85
OWN	57	1473	3.73
RENT	390	10546	3.57



Loan Purpose Analysis – Univariate

Observation: A very large percentage of number of loans were taken for "debt consolidation“” followed by credit card". And very fewer loans were taken for “renewable energy”, “house” and “educational”.

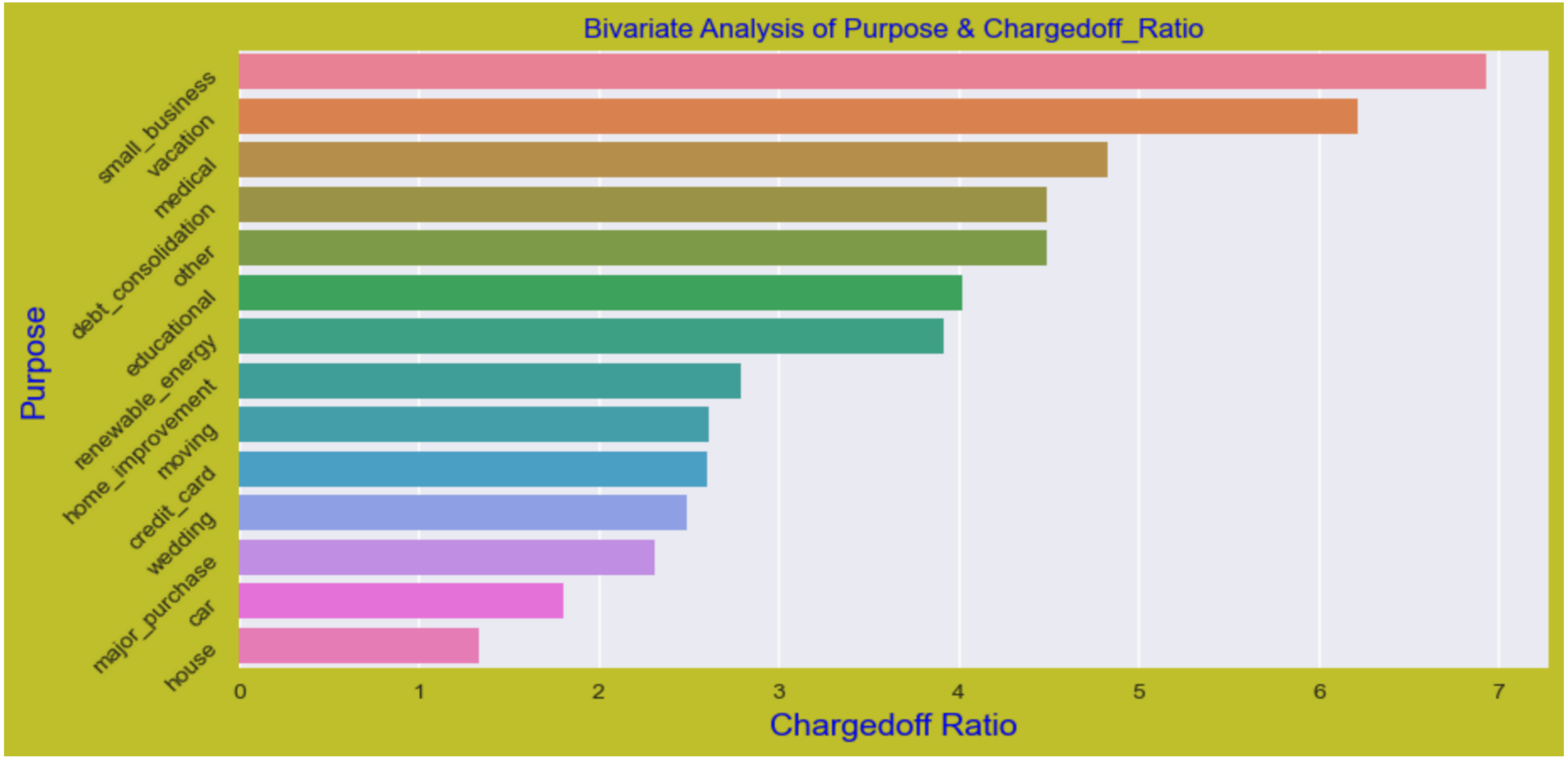


Loan Purpose Analysis – Bivariate

Observation :

- Loan purpose has a significant impact on charged-off ratios.
- Loans for "small_business" and "vacation" have the highest charged-off ratios, indicating a higher risk of default.
- Loans for "house" and "car" have the lowest charged-off ratios, suggesting a lower risk of default.
- Loans for personal expenses may be less secured and have higher risk profiles whereas Loans for assets may have collateral or be secured, reducing the risk of default.

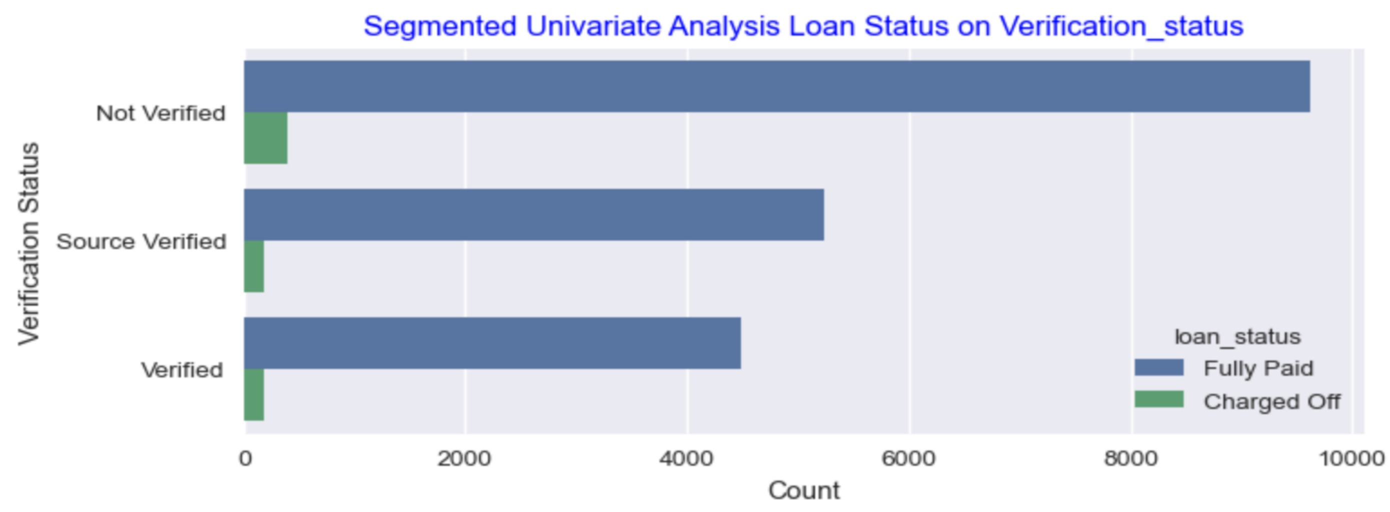
purpose	Charged Off	Fully Paid	Chargedoff_Ratio
small_business	48	645	6.93
vacation	14	211	6.22
medical	18	355	4.83
debt_consolidation	413	8776	4.49
other	93	1980	4.49
educational	7	167	4.02
renewable_energy	2	49	3.92
home_improvement	36	1256	2.79
moving	9	336	2.61
credit_card	73	2740	2.60
wedding	14	548	2.49
major_purchase	30	1269	2.31
car	16	874	1.80
house	2	148	1.33



Loan Purpose Analysis – Segmented Univariate & Bivariate

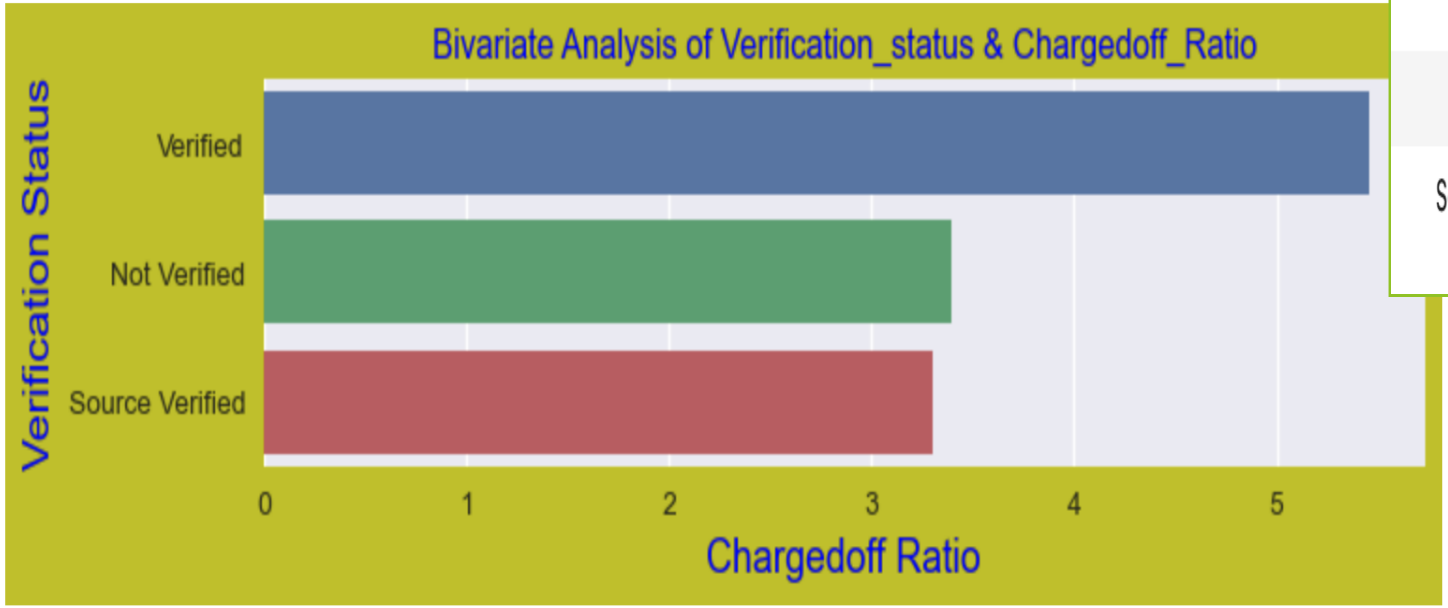
Observation 1:

Only 50% of the Loan borrowers applications are verified by the company or have source verified.



Observation 2:

Though, only 50% of the application are verified or source verified, still the charge-Offs ratio is higher in verified loans. Which infers that the verification process needs to be scrutinized, to be more effective.



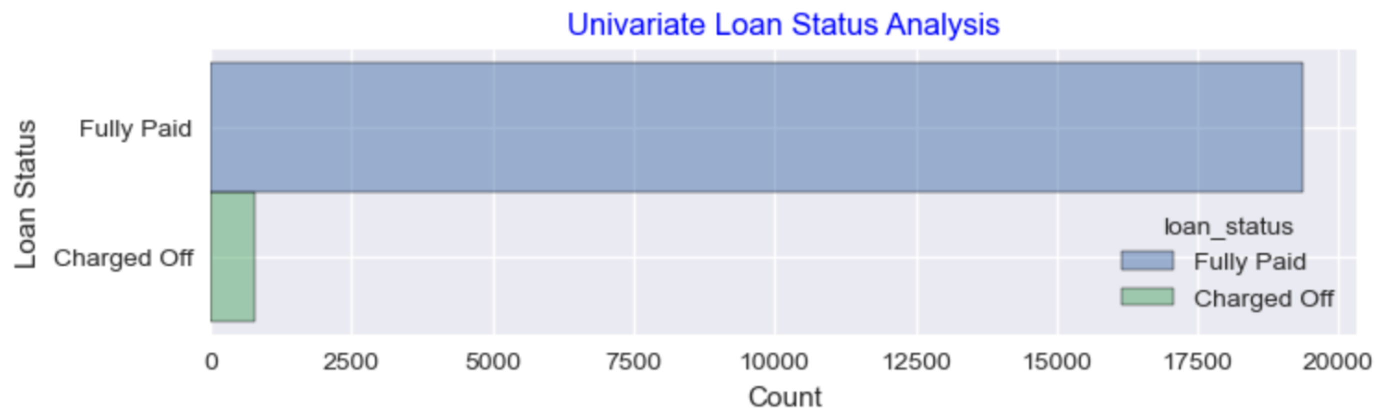
verification_status	Charged Off	Fully Paid	Chargedoff_Ratio
Verified	255	4417	5.46
Not Verified	341	9685	3.40
Source Verified	179	5252	3.30

Loan Status Analysis – Univariate & Segmented Univariate

Observation 1:

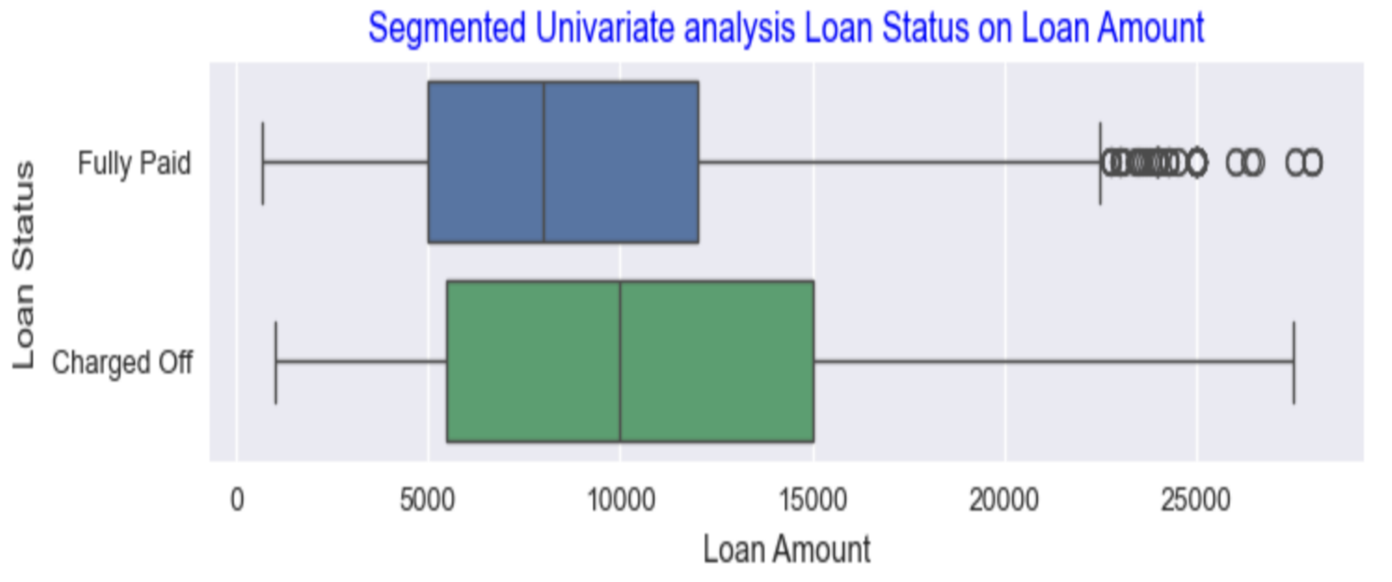
Count of Charged off loan are very low compared to count of Fully Paid loans.

.



Observation:

- Fully paid loans tend to have lower maximum loan amounts compared to charged-off loans.
- Charged-off loans have a longer right whisker, indicating a larger number of borrowers(defaulters) with higher loan amounts.
- Larger loan amounts are associated with a higher risk of default.



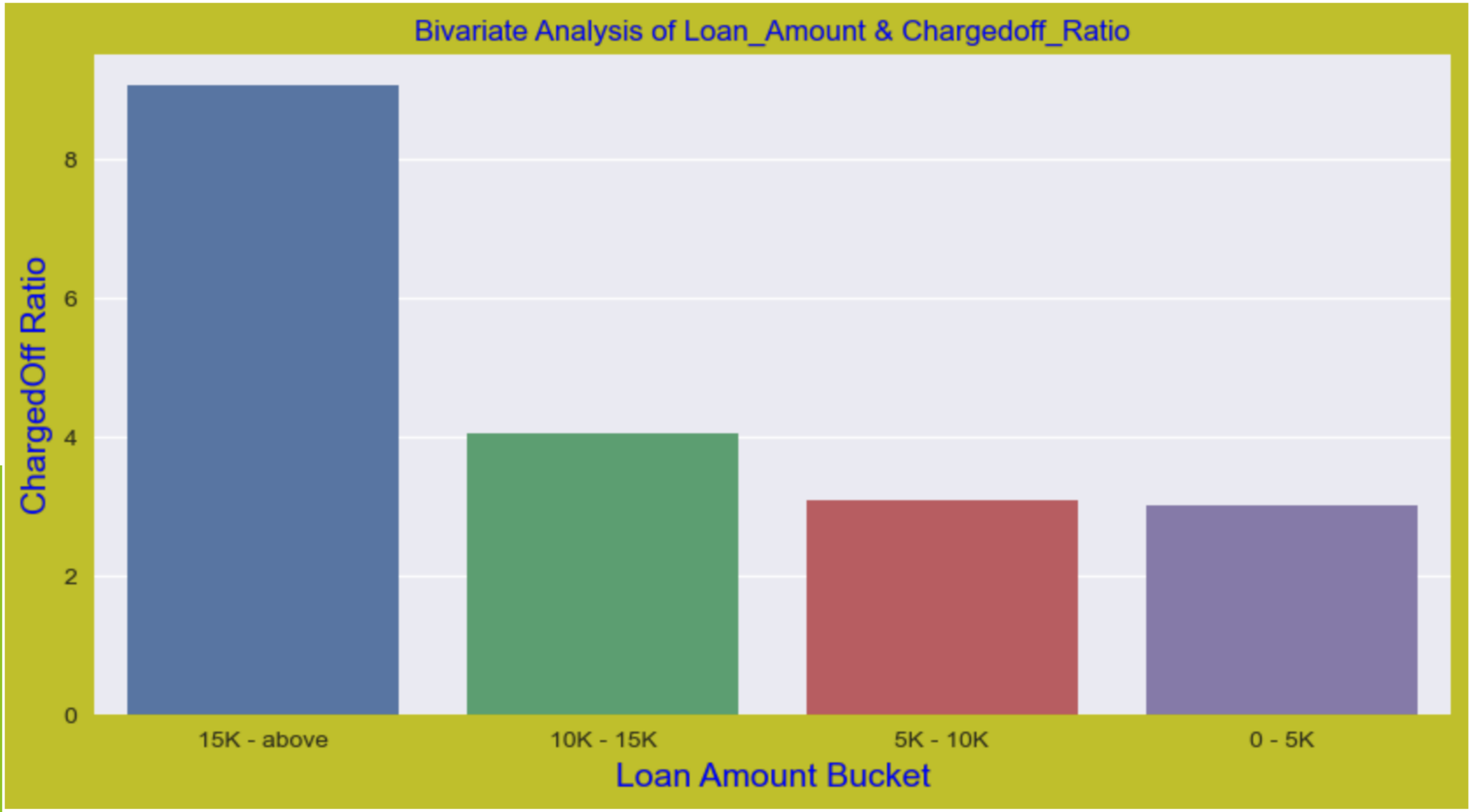
Type to enter a caption

Loan Amount Analysis – Bivariate

Observation:

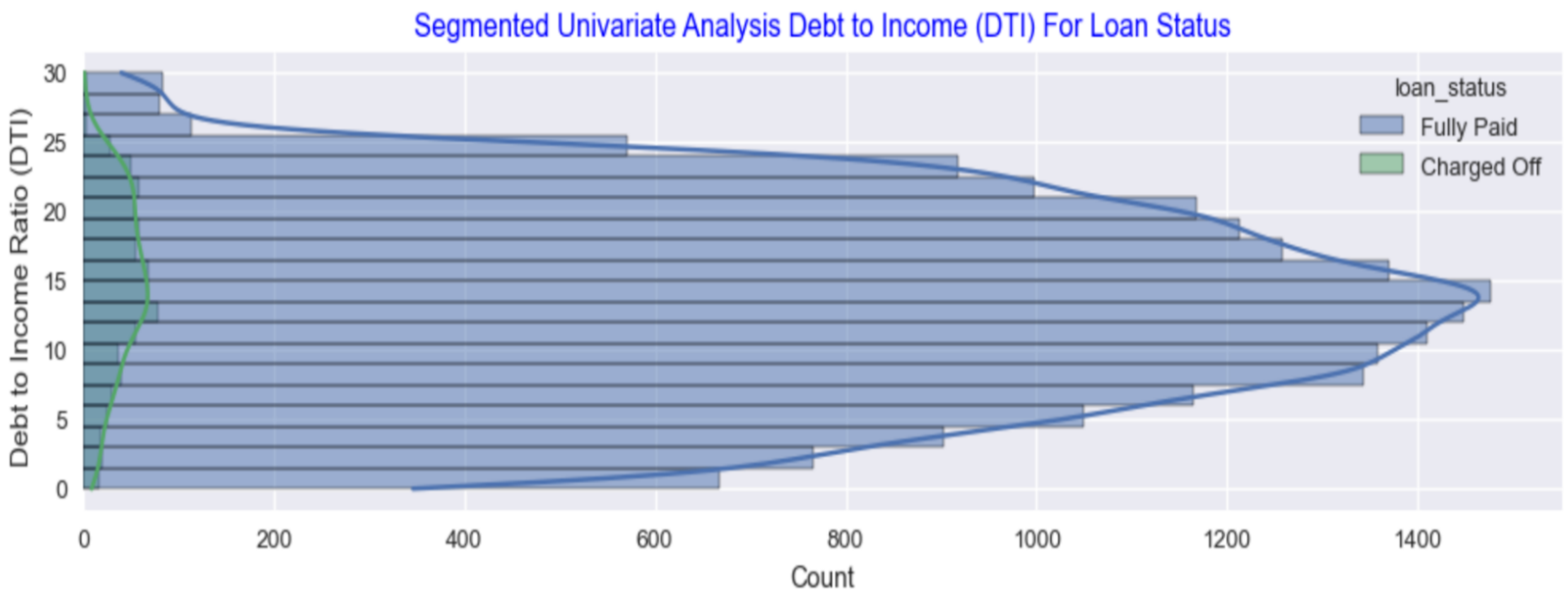
- The Charge Off ratio of all the customers having loan_amount '15K and above' is the highest where as a decreasing trend is seen with decrease in loan amount.
- Larger loan amounts may be associated with riskier borrowers or projects.

loan_amnt_b	Charged Off	Fully Paid	Chargedoff_Ratio
15K - above	181	1816	9.06
10K - 15K	168	3981	4.05
5K - 10K	248	7813	3.08
0 - 5K	178	5744	3.01



Debt to Income (DTI) ratio Analysis – Segmented Univariate

Observation : Charged Off pattern is high incases where DTI ratios are between 5 to 25%.

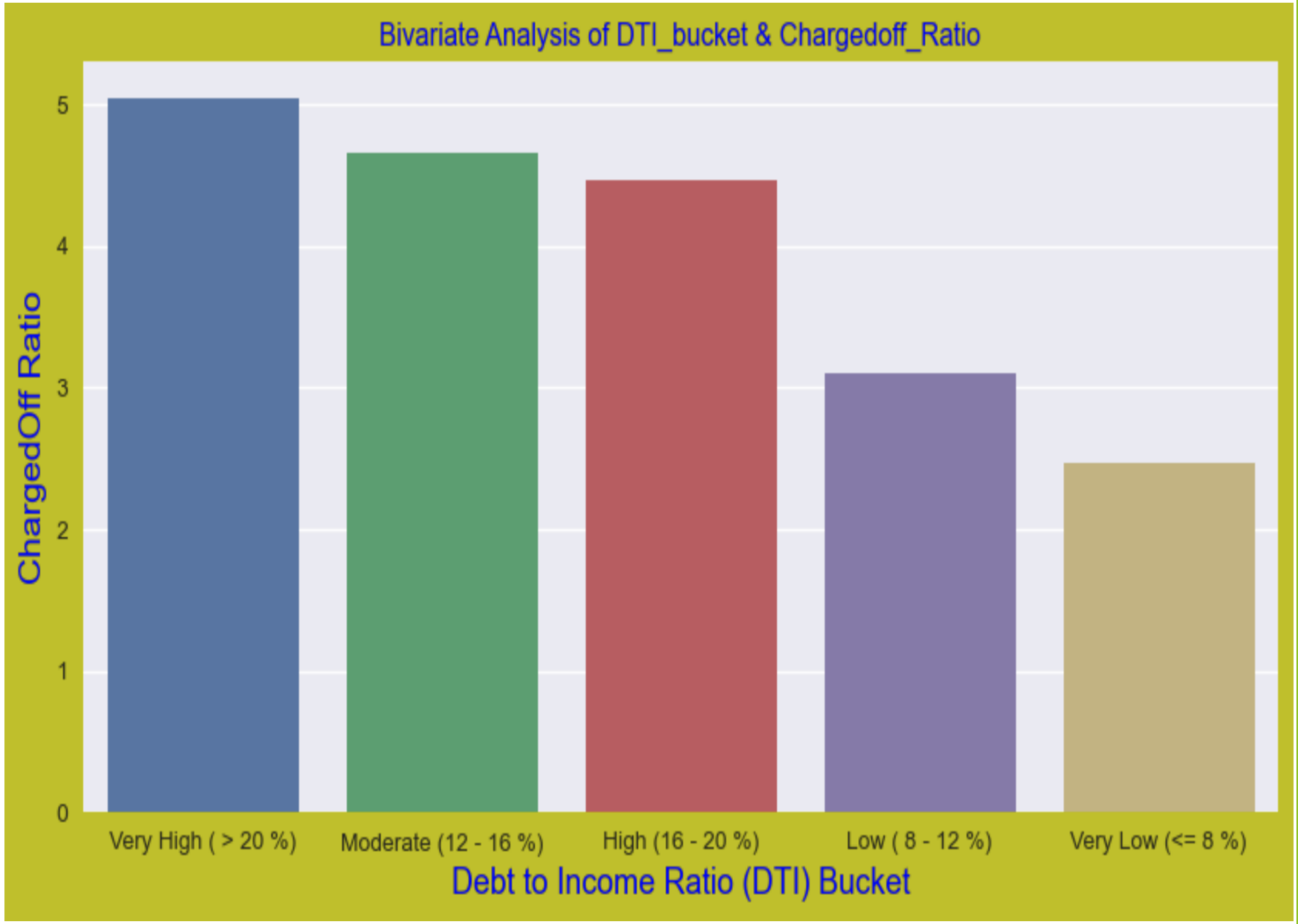


Debt to Income (DTI) ratio Analysis – Bivariate

Observation:

- Trend of increasing charged-off ratio as DTI increases.
- Individuals with higher DTI ratios are more likely to default on their loans.
- Higher DTI ratios may indicate financial stress, making it more difficult for borrowers to meet their loan obligations.

dti_b	Charged Off	Fully Paid	Chargedoff_Ratio
Very High (> 20 %)	186	3498	5.05
Moderate (12 - 16 %)	188	3845	4.66
High (16 - 20 %)	156	3331	4.47
Low (8 - 12 %)	118	3673	3.11
Very Low (<= 8 %)	127	5007	2.47



Annual Income Analysis – Univariate & Segmented Univariate

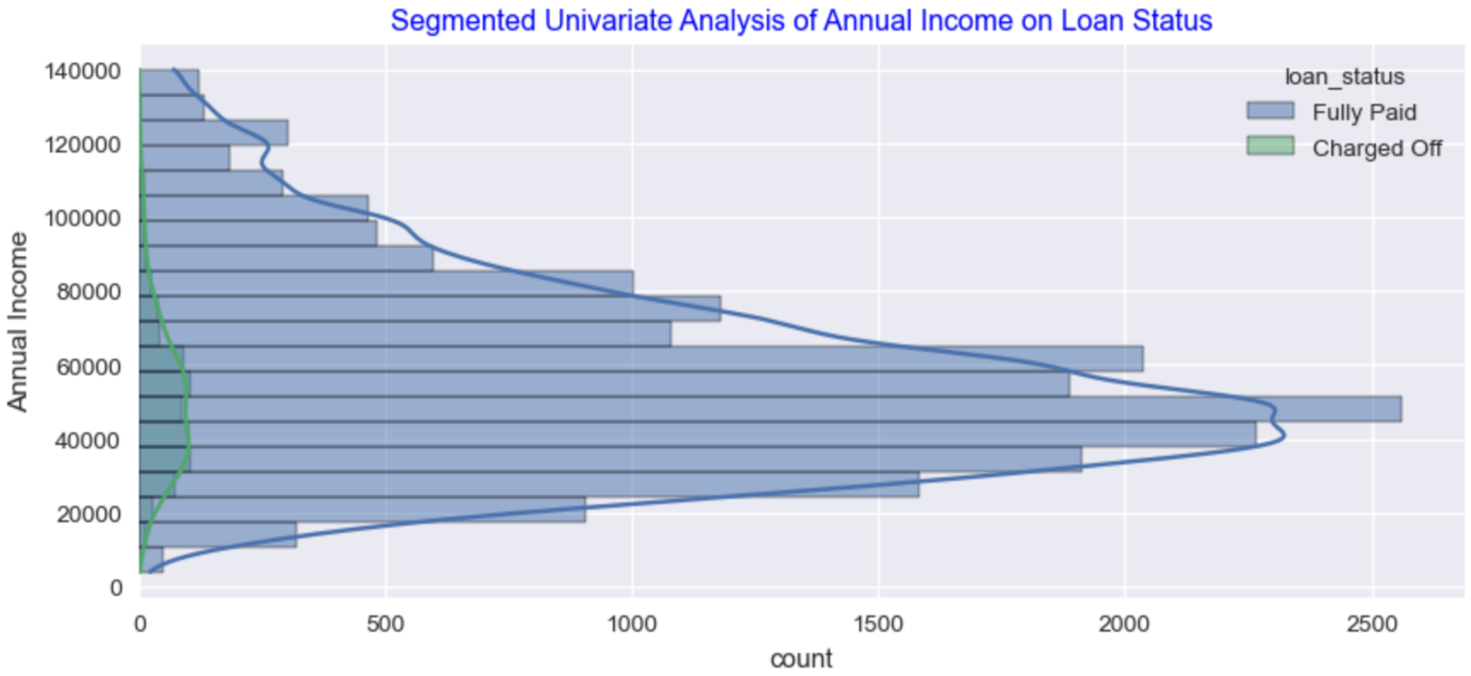
Observation 1:

Majority of the loans Borrowers are in the range of Annual Income range of 35K to 70K, and hence more the number of Charged Off cases in this range of salary too.


```
count    20129.000000
mean     56432.248667
std      25435.811509
min       4000.000000
25%      38000.000000
50%      51996.000000
```

Observation 2:

- Loan Borrowers with annual income between 25K to 65K are more likely to default.
- Higher annual income borrowers > 80K are less likely to default.
- Higher annual income may be a factor associated with a higher likelihood of loan repayment whereas lower annual incomes may be more likely to default on their loans.

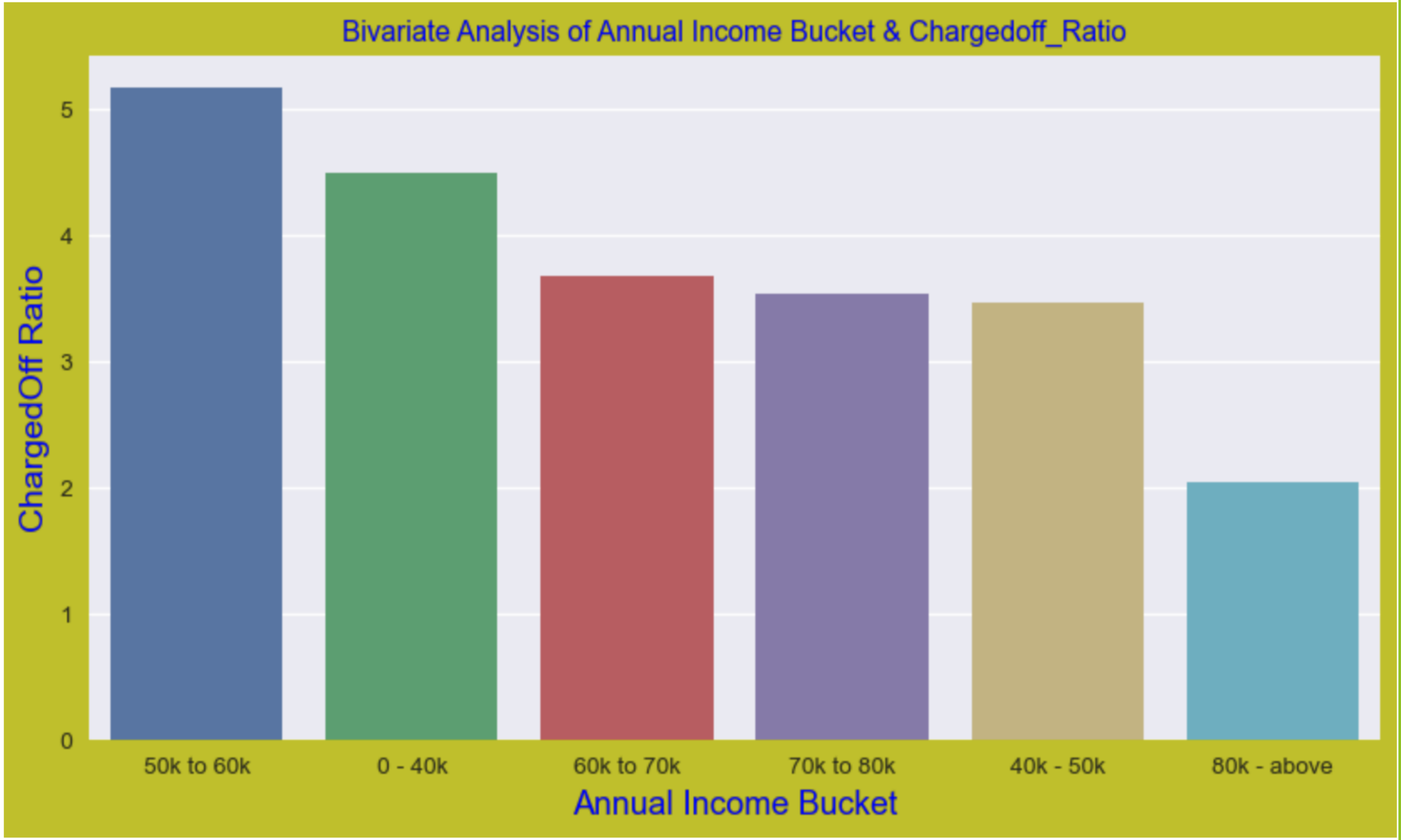


Annual Income Analysis - Bivariate

Observation :

- Income range "80K & above" has the least chance of defaulting.
- Income range 50K-60K has highest chances of defaulting.
- Otherwise, general trend of decreasing charged-off ratio as annual income increases.

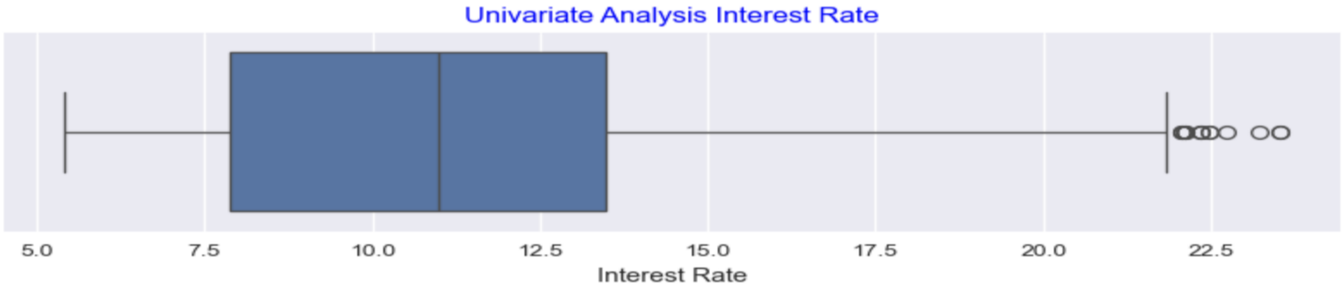
annual_inc_b	Charged Off	Fully Paid	Chargedoff_Ratio
50k to 60k	167	3065	5.17
0 - 40k	276	5869	4.49
60k to 70k	81	2119	3.68
70k to 80k	61	1660	3.54
40k - 50k	123	3422	3.47
80k - above	67	3219	2.04



Annual Income Analysis – Univariate & Segmented Univariate

Observation 1 :

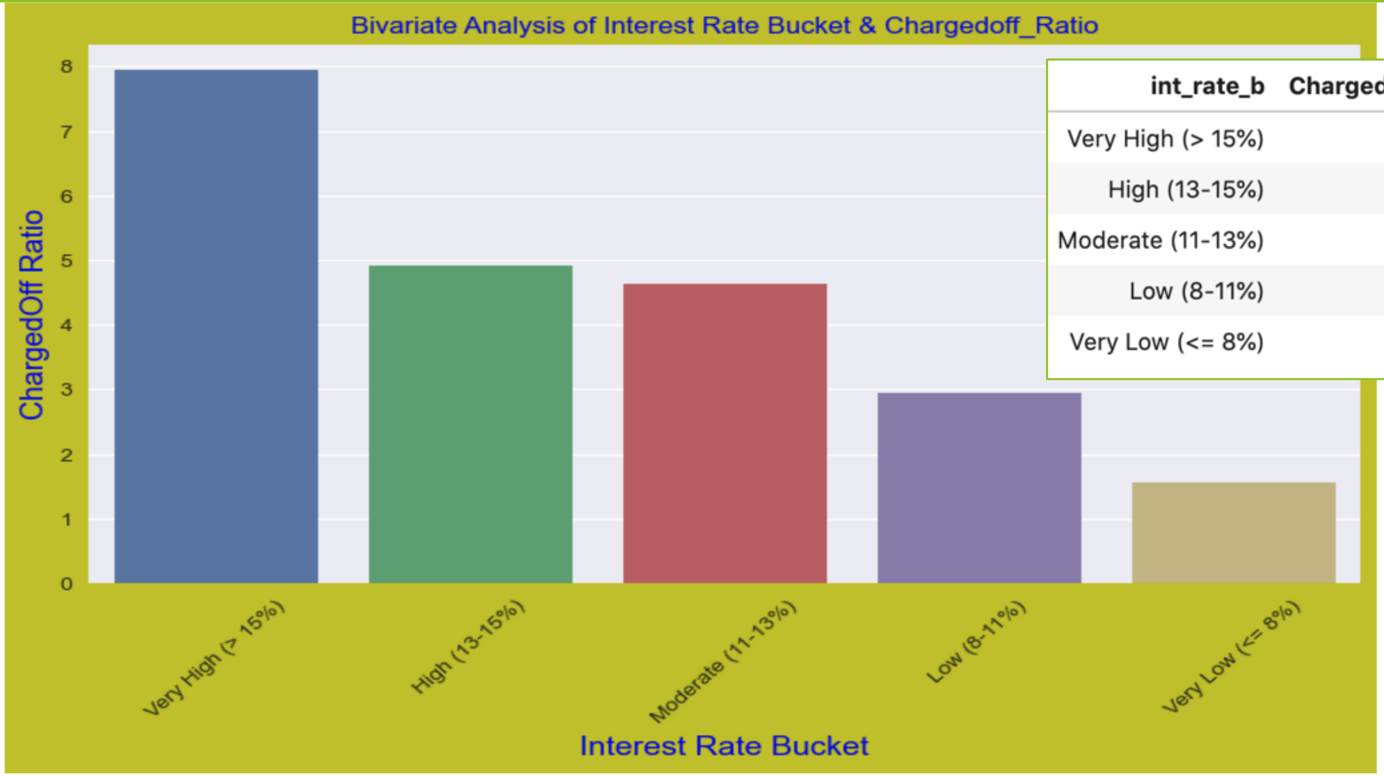
Most of the applicant's rate of interest is between in the range of 7.8 %-13.5%. Average Rate of interest of rate is 11 %.




```
count    20129.000000
mean        11.036188
std          3.399549
min          5.420000
25%          7.880000
50%         10.990000
75%         13.480000
max         23.520000
```

Observation 2 :

- ➢ Maximum number of loans were offered in the very Low & Low interest rates ranges and charge Offs were least in this interest range.
- ➢ As the interest rate started increasing Above 11%, the Defaulting of borrowers increased.

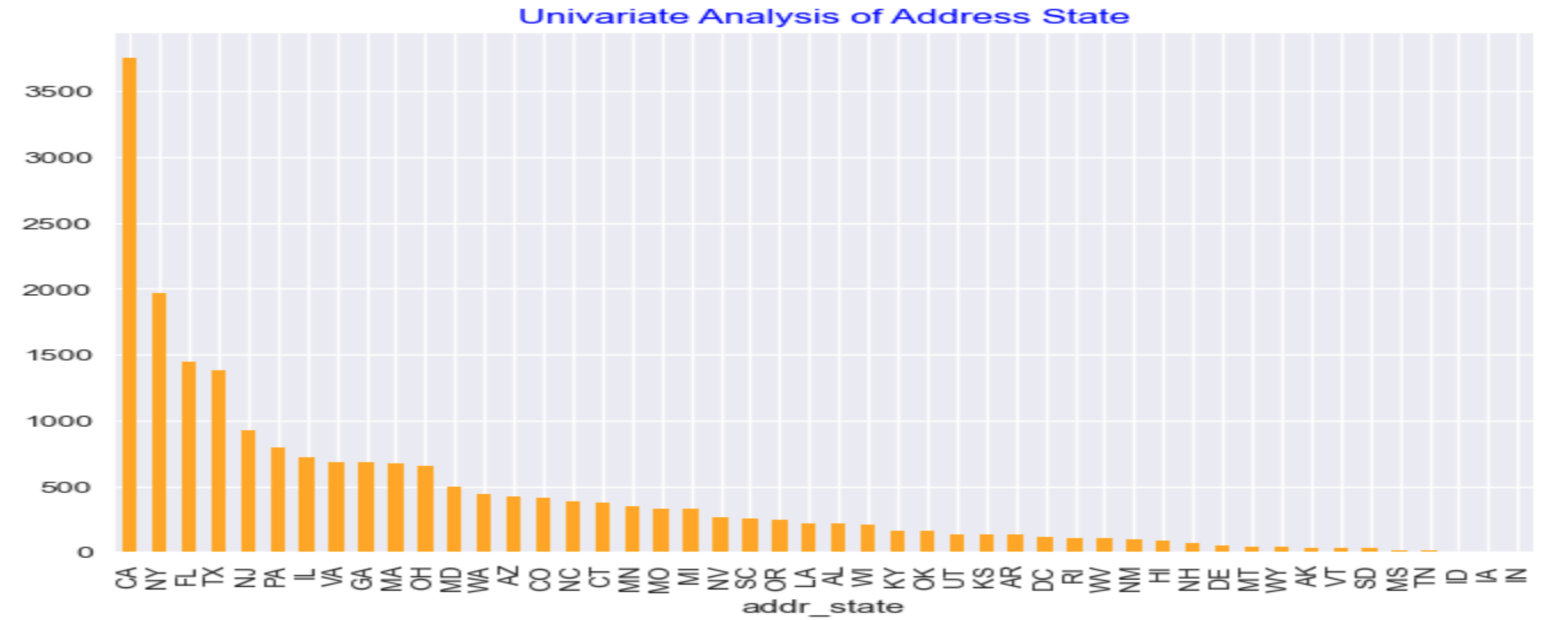


int_rate_b	Charged Off	Fully Paid	Chargedoff_Ratio
Very High (> 15%)	211	2448	7.94
High (13-15%)	157	3037	4.92
Moderate (11-13%)	181	3724	4.64
Low (8-11%)	138	4554	2.94
Very Low (<= 8%)	88	5591	1.55

Address State Analysis – Univariate

Observation:

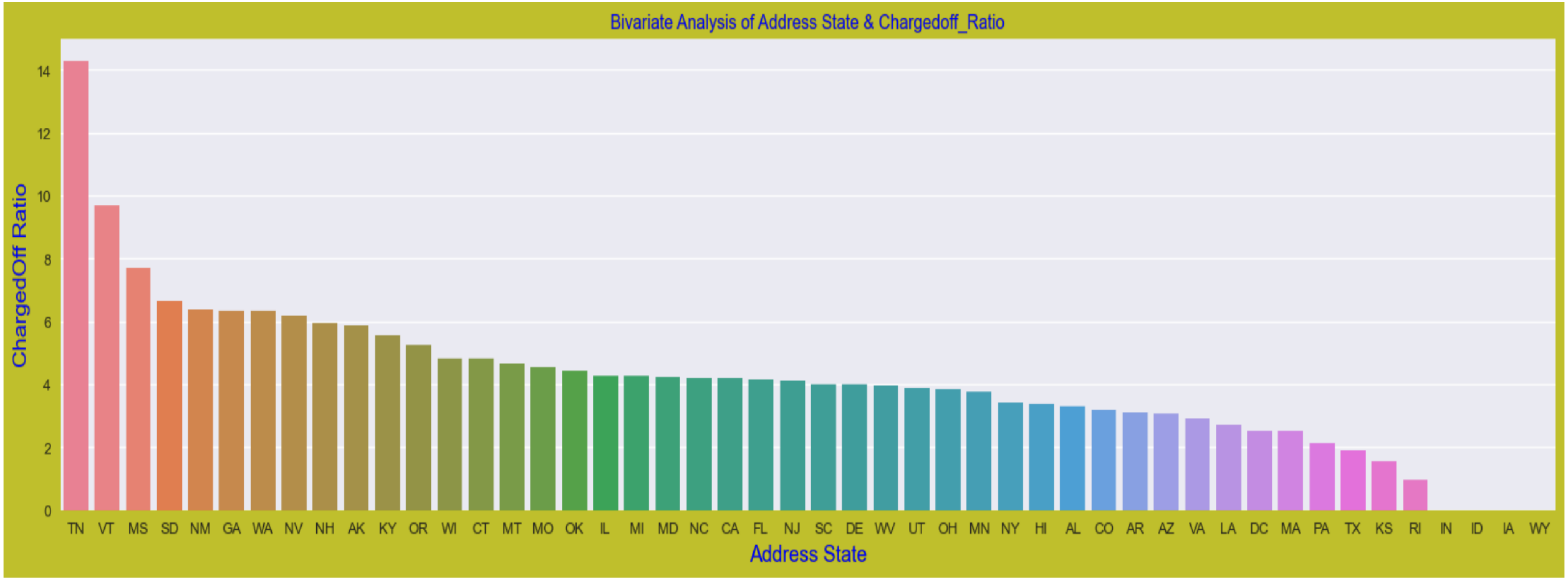
Majority of the Loan borrowers are from the large urban cities like California, New York, Florida, Texas.



Address State Analysis – Bivariate

Observation:

- States "Tennessee (TN)", "Vermont (VT)", "Mississippi(MS)", "South Dakota" contributes to the maximum number of defaulters.
- States "Indiana(IN)", "Idaho(ID)", "Iowa(IA)", "Wyoming(WY)" does not have any defaulters.
- A wide variation in charged-off ratios across different states observed.
- There may be underlying economic, social, or regulatory factors in certain states that contribute to higher default rates.



Correlation Analysis

Strong Positive Correlations :

- "loan_amnt" and "installment": This is expected, as larger loans typically have higher monthly payments.
- "loan_amnt" and "total_pymnt": Again, larger loans will generally have higher total payments made.
- "installment" and "total_pymnt": Higher monthly payments contribute to larger total payments over time.

Moderate Positive Correlations:

- "loan_amnt" and "total_rec_int": Larger loans often incur more interest charges.
- "installment" and "total_rec_int": Higher monthly payments can lead to more interest being paid over time.

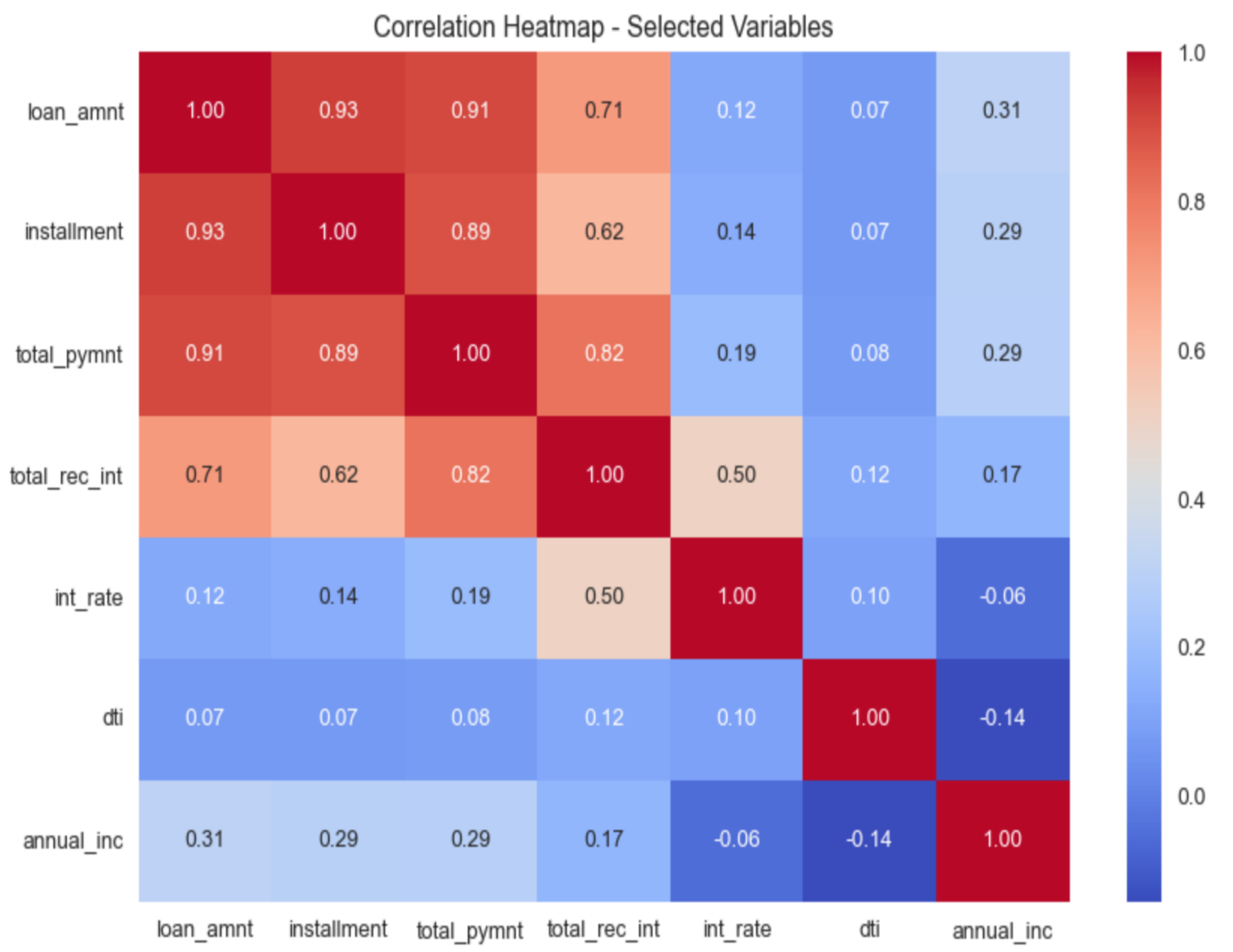
Weak or No Correlation :

- “Interest rate", "debt-to-income ratio", and "annual income" have little to no impact on "loan amount", "installment", and "total payment".

Negative Correlation :

- “dti" and "annual_inc": This suggests that as annual income increases, debt-to-income ratio tends to decrease.

```
Strongly Correlated Pairs (Threshold > 0.7):
installment     loan_amnt        0.93
loan_amnt       installment      0.93
                total_pymnt      0.91
total_pymnt     loan_amnt        0.91
installment     total_pymnt      0.89
total_pymnt     installment      0.89
                total_rec_int    0.82
total_rec_int   total_pymnt      0.82
loan_amnt       total_rec_int    0.71
total_rec_int   loan_amnt        0.71
```



	loan_amnt	installment	total_pymnt	total_rec_int	int_rate	dti	annual_inc
loan_amnt	1.00	0.93	0.91	0.71	0.12	0.07	0.31
installment	0.93	1.00	0.89	0.62	0.14	0.07	0.29
total_pymnt	0.91	0.89	1.00	0.82	0.19	0.08	0.29
total_rec_int	0.71	0.62	0.82	1.00	0.50	0.12	0.17
int_rate	0.12	0.14	0.19	0.50	1.00	0.10	-0.06
dti	0.07	0.07	0.08	0.12	0.10	1.00	-0.14
annual_inc	0.31	0.29	0.29	0.17	-0.06	-0.14	1.00