# LENDING CLUB CASE STUDY

BY – ANKIT CHANDRE

Problem Statement :

You work for a **consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for 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** 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

Constraint:

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

**1.Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
3. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

**2.Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Objective:

Explore the impact of consumer attributes and loan attributes on default tendency through Exploratory Data Analysis (EDA).

# Data Source :

Loan.csv contains 39717 rows and 111 columns and has loan and customer attributes.

In [3]:
```python
#Display first few rows
loan_data.head(5)
```

Out[3]:

| | id | member_id | loan_amnt | funded_amnt | funded_amnt_inv | term | int_rate | installment | grade | sub_grade | ... | num_tl_90g_dpd_24m | num_tl_op_past_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1077501 | 1296599 | 5000 | 5000 | 4975.0 | 36 months | 10.65% | 162.87 | B | B2 | ... | NaN | |
| 1 | 1077430 | 1314167 | 2500 | 2500 | 2500.0 | 60 months | 15.27% | 59.83 | C | C4 | ... | NaN | |
| 2 | 1077175 | 1313524 | 2400 | 2400 | 2400.0 | 36 months | 15.96% | 84.33 | C | C5 | ... | NaN | |
| 3 | 1076863 | 1277178 | 10000 | 10000 | 10000.0 | 36 months | 13.49% | 339.31 | C | C1 | ... | NaN | |
| 4 | 1075358 | 1311748 | 3000 | 3000 | 3000.0 | 60 months | 12.69% | 67.79 | B | B5 | ... | NaN | |

5 rows × 111 columns

In [4]:
```python
#Display last few rows
loan_data.tail(5)
```

Out[4]:

| | id | member_id | loan_amnt | funded_amnt | funded_amnt_inv | term | int_rate | installment | grade | sub_grade | ... | num_tl_90g_dpd_24m | num_tl_op_pa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39712 | 92187 | 92174 | 2500 | 2500 | 1075.0 | 36 months | 8.07% | 78.42 | A | A4 | ... | NaN | |
| 39713 | 90665 | 90607 | 8500 | 8500 | 875.0 | 36 months | 10.28% | 275.38 | C | C1 | ... | NaN | |
| 39714 | 90395 | 90390 | 5000 | 5000 | 1325.0 | 36 months | 8.07% | 156.84 | A | A4 | ... | NaN | |
| 39715 | 90376 | 89243 | 5000 | 5000 | 650.0 | 36 months | 7.43% | 155.38 | A | A2 | ... | NaN | |
| 39716 | 87023 | 86999 | 7500 | 7500 | 800.0 | 36 months | 13.75% | 255.43 | E | E2 | ... | NaN | |

5 rows × 111 columns

In [5]:
```python
# check dimensions of dataframe , no. of rows and column
loan_data.shape
```

Out[5]: (39717, 111)

In [23]:
```python
loan_data.shape
```

Out[23]: (38577, 50)

In [24]:
```python
# List of columns containing behavioral data
behavioural_cols = [
    'delinq_2yrs', 'earliest_cr_line', 'last_pymnt_amnt', 'inq_last_6mths',
    'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc',
    'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv',
    'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries',
    'collection_recovery_fee', 'application_type', 'last_pymnt_d', 'last_credit_pull_d'
]

# Remove columns containing behavioral data
loan_data.drop(columns=behavioural_cols, inplace=True)

```

In [25]:
```python
loan_data.nunique().sort_values(ascending=True)
```

Out[25]:
```
pymnt_plan                       1
delinq_amnt                      1
chargeoff_within_12_mths         1
acc_now_delinq                   1
policy_code                      1
collections_12_mths_ex_med       1
initial_list_status              1
tax_liens                        1
loan_status                      2
term                             2
verification_status              3
pub_rec_bankruptcies             3
home_ownership                   5
grade                            7
emp_length                      11
purpose                         14
addr_state                      50
issue_d                         55
mths_since_last_delinq          95
mths_since_last_record         111
int_rate                       370
zip_code                       822
loan_amnt                      870
funded_amnt                   1019
dti                           2853
annual_inc                    5215
funded_amnt_inv               8050
installment                  15022
id                           38577
dtype: int64
```

# Data Cleaning

**1.Absence of Identifiable Headers and Footers:**

1. Initially, the dataset lacked distinct headers and footers, making it challenging to discern the structure and organization of the data.

**2.Data Type Verification and Handling Missing Values:**

1. A comprehensive examination was conducted to verify the data types across all columns and assess the presence of null or missing values within each column and row.

**3.Analysis of Loan Status Column:**

1. A specific focus was directed towards analysing the loan status column to understand the distribution of different loan statuses.
2. Rows corresponding to "current loan" status were identified and subsequently removed from further analysis to maintain data integrity.

**4.Identification and Removal of Columns with Unique Values:**

1. Columns containing unique values, which provide limited or redundant information, were identified and deemed unnecessary for the analysis.
2. These columns were then removed to streamline the dataset and improve its relevance to the analytical objectives.

**5.Exclusion of Behavioural Attribute Columns:**

1. Certain columns representing behavioural attributes were identified, which are typically not relevant during the loan approval process and may introduce bias into the analysis.
2. To ensure the analysis focuses solely on pertinent loan attributes, these behavioural attribute columns were excluded from consideration.

**6.Elimination of Columns with Single Unique Value:**

1. Columns featuring only one unique value across all records were identified, indicating minimal variability and limited significance for analysis purposes.
2. Consequently, these columns were eliminated to enhance the quality and relevance of the remaining dataset.

**7.Removal of Columns with High Null Value Percentage:**

1. Columns with a significant proportion of null or missing values, exceeding a predefined threshold (e.g., 50%), were identified.
2. Given the substantial data gaps in these columns, they were deemed unsuitable for meaningful analysis and subsequently excluded from further consideration.

# Data Conversion

1. Leading and trailing white spaces were eliminated from the 'term' column, and unique values were identified.
    1. This process involved removing any unnecessary spaces from the beginning and end of each term value.
    2. Additionally, we identified all unique values present in the 'term' column for further analysis.
2. Frequency counts were performed for each value in the 'term' column, and string values were converted to integers.
    1. We tabulated the occurrence of each term value to understand its distribution.
    2. String representations of term durations were converted to integer format for consistency and numerical analysis.
3. The 'int_rate' column underwent a transformation from string to integer format.
    1. Initially, the interest rates were stored as string data types.
    2. To facilitate mathematical operations and analysis, these string representations were converted to integers.
4. Extraneous '%' symbols were removed from the 'int_rate' column to ensure data uniformity and accuracy.
    1. Some entries in the 'int_rate' column included percentage symbols, which were unnecessary for numerical analysis.
    2. Thus, these symbols were removed to standardize the data and avoid errors in calculations.
5. The columns 'loan_funded_amnt' and 'funded_amnt' were converted to floating-point format.
    1. Initially, these columns likely contained numerical values but were stored as strings.
    2. To enable arithmetic operations and precise numerical representation, they were converted to floating-point data types.
6. Several columns, including 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', and 'dti', had their values rounded off to two decimal points.
    1. Rounding off numerical values to a specific decimal precision aids in clarity and consistency.
    2. This process ensures that numerical data maintains a uniform format throughout the dataset.
7. The 'issue_d' column was converted to the date-time data type, and 'loan_amt' and 'funded_amnt' were converted to float64 data types.
    1. 'issue_d' likely represents the date of loan issuance, making it essential to convert it to a date-time format for chronological analysis.
    2. Similarly, 'loan_amt' and 'funded_amnt' were converted to float64 data types for accurate numerical representation.
8. Finally, a few columns were rounded off to two decimal places for consistency and precision.
    1. Rounding numerical values to a specific decimal place reduces clutter and enhances readability, facilitating easier analysis and interpretation.

# Analysis

## Distribution of Loan Amount

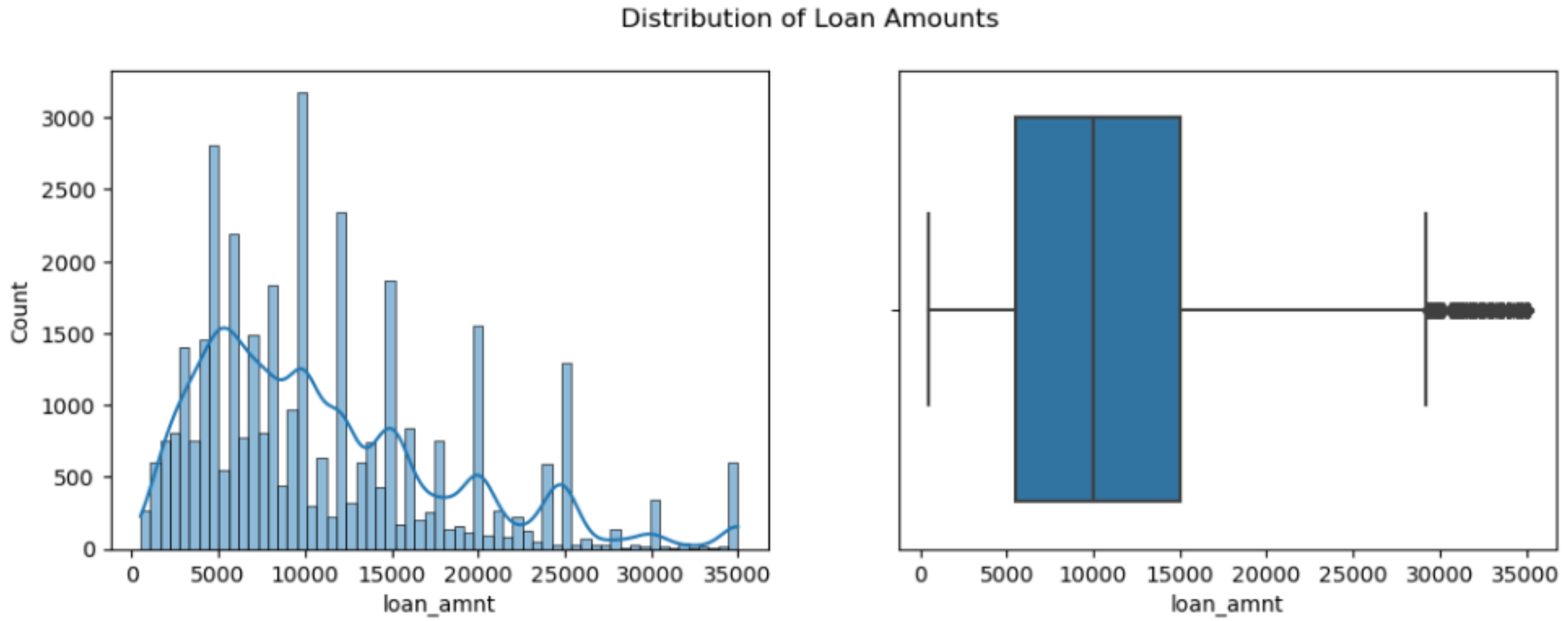
Observation: The distribution of loan amounts indicates that a significant number of individuals opted for a loan amount of "10,000", with the median loan amount also being "10,000" Relatively few individuals chose loan amounts exceeding "30,000"

#Univariate Analysis

Loan

In [49]:

```
# Set the figure size
plt.figure(figsize=(12, 4))

# Plot the histogram
plt.subplot(1, 2, 1)
sns.histplot(data=loan_data, x='loan_amnt', kde=True)  # Removed rug=True

# Plot the boxplot
plt.subplot(1, 2, 2)
sns.boxplot(data=loan_data, x='loan_amnt')

# Set a single title for both subplots
plt.suptitle('Distribution of Loan Amounts')

# Show the plot
plt.show()
```



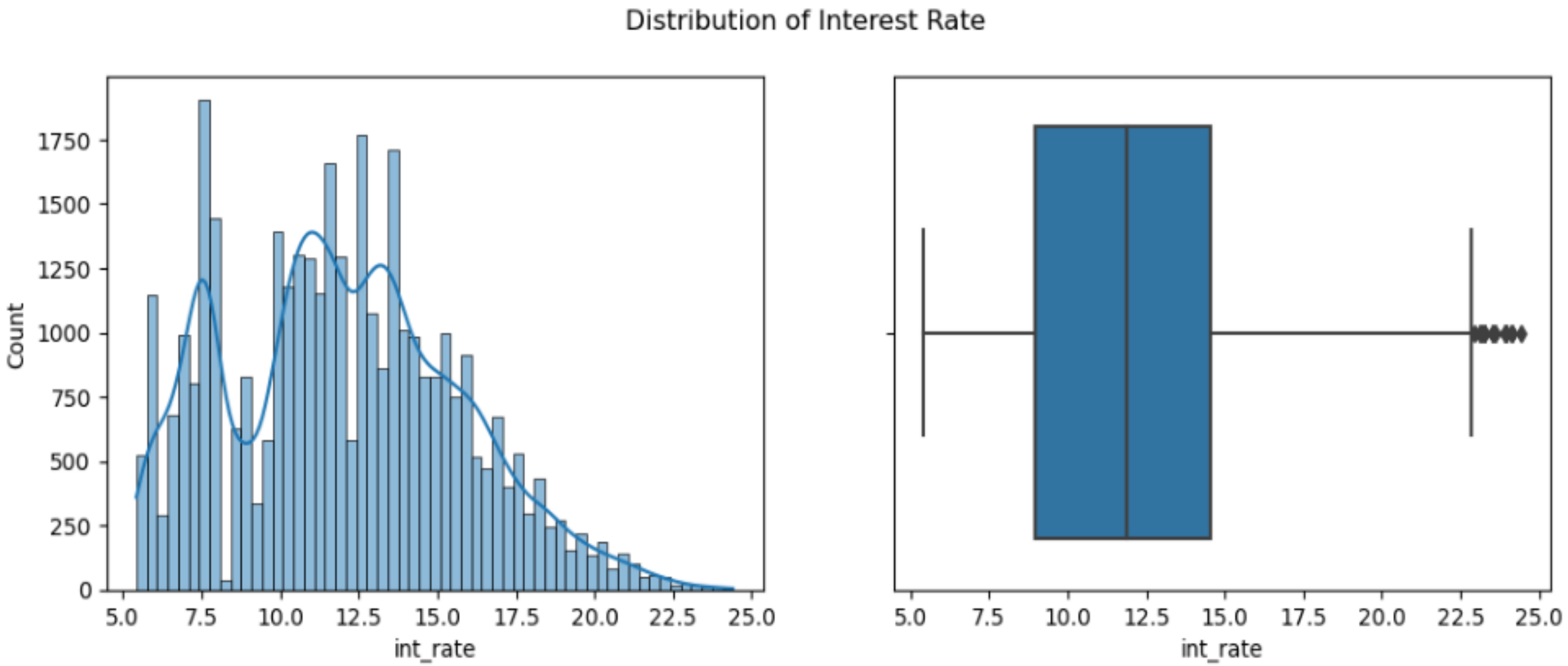
# Distribution of Interest Rate

Observation: The majority of applicants have interest rates ranging from 8% to 14%, with an average interest rate of 11.7%. The distribution of interest rates indicates that most fall between 9% and 14.5%, while some individuals have opted for higher rates, reaching up to 22.5%.

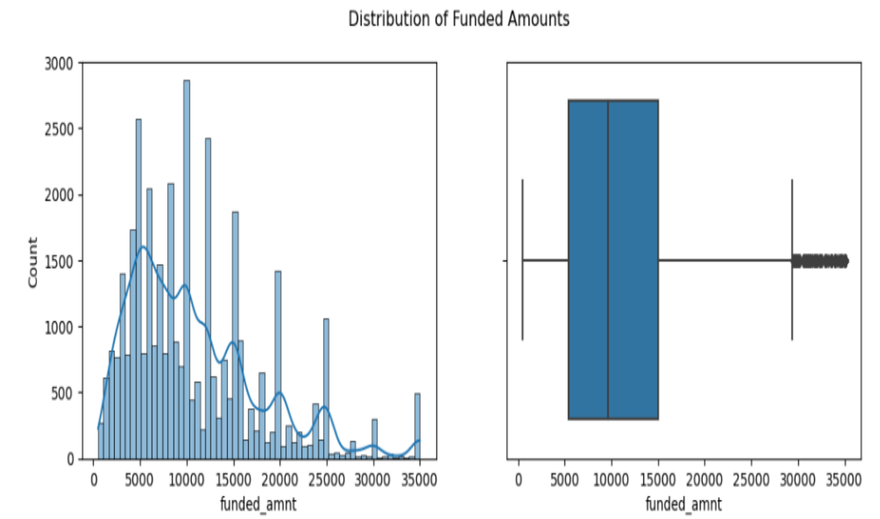
Interest Rate

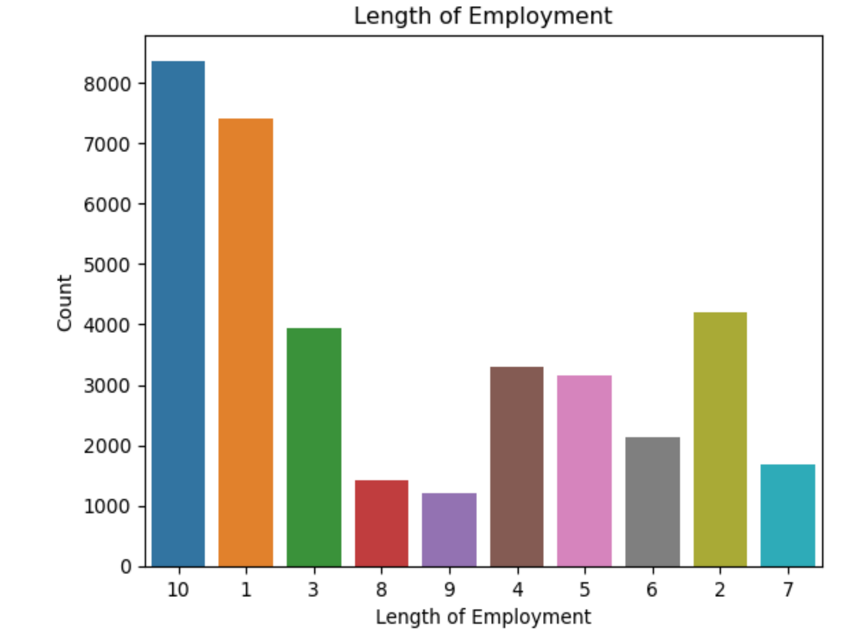
In [53]:

```
# Set the figure size
plt.figure(figsize=(12, 4))

# Plot the histogram using histplot
plt.subplot(1, 2, 1)
sns.histplot(data=loan_data, x='int_rate', kde=True)

# Plot the boxplot
plt.subplot(1, 2, 2)
sns.boxplot(data=loan_data, x='int_rate')

# Set a single title for both subplots
plt.suptitle('Distribution of Interest Rate')

# Show the plot
plt.show()
```

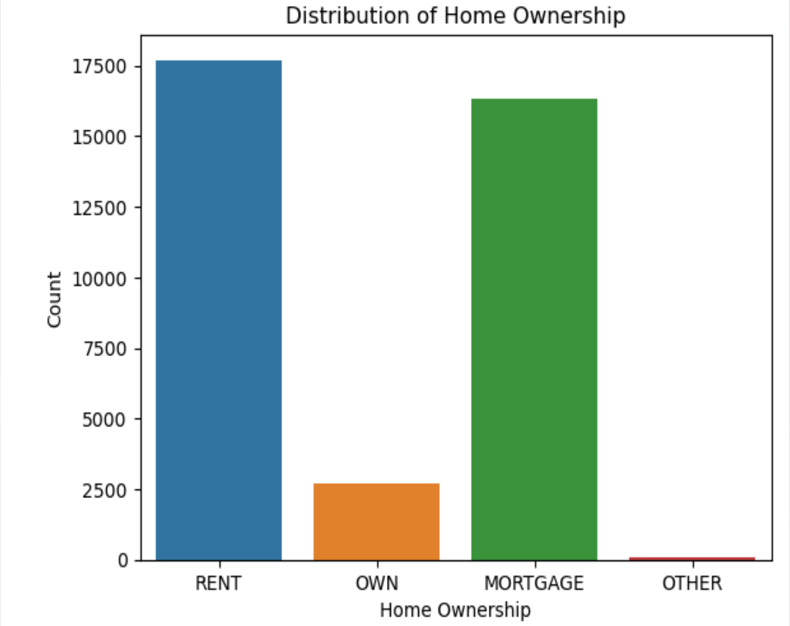


# Analysis

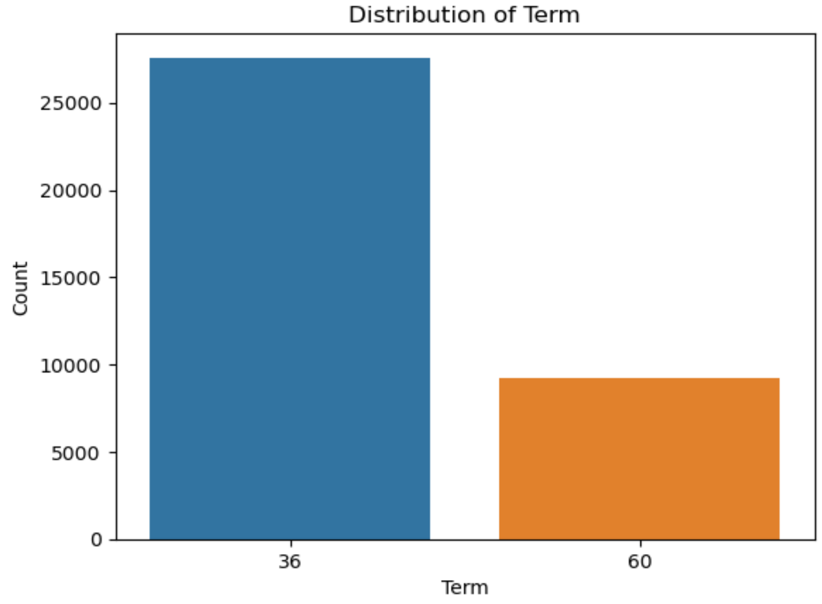


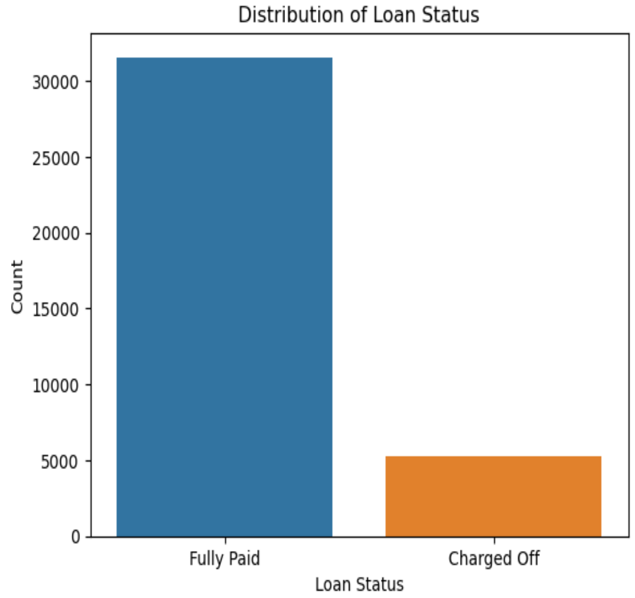

observation: Emplyment length is 10+ for most of the borrowers



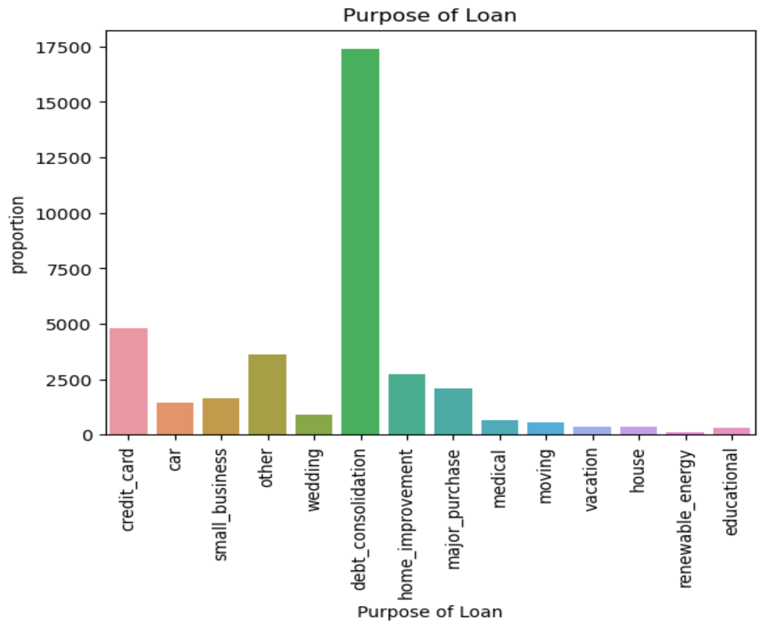

Observation: Most of the home owners are living on Rent or on Mortgage



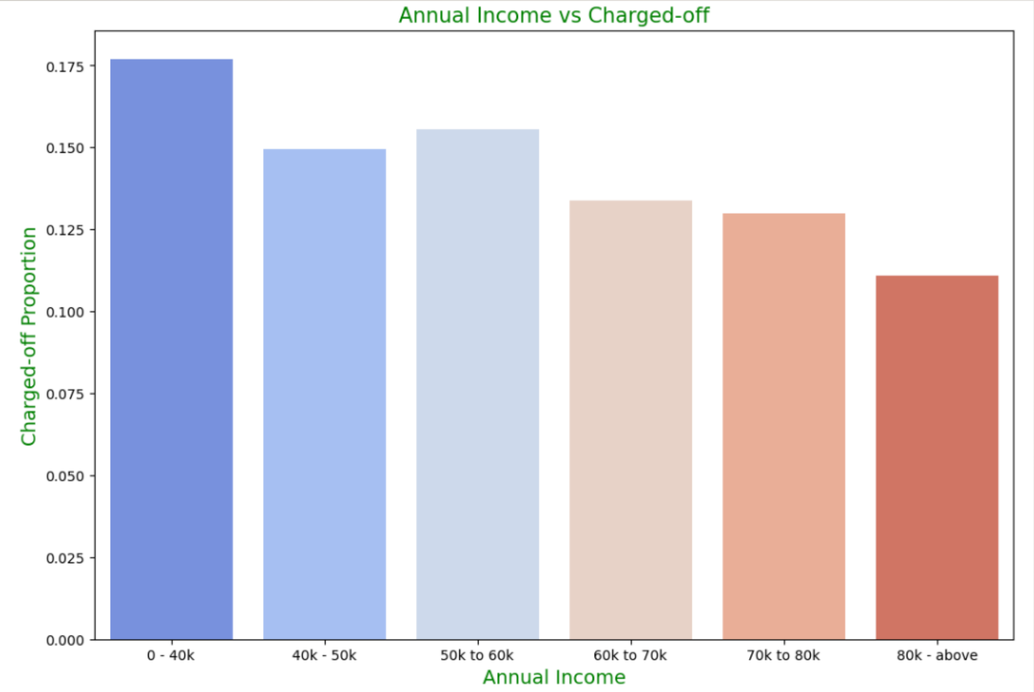

Observation: term is more of 36 months

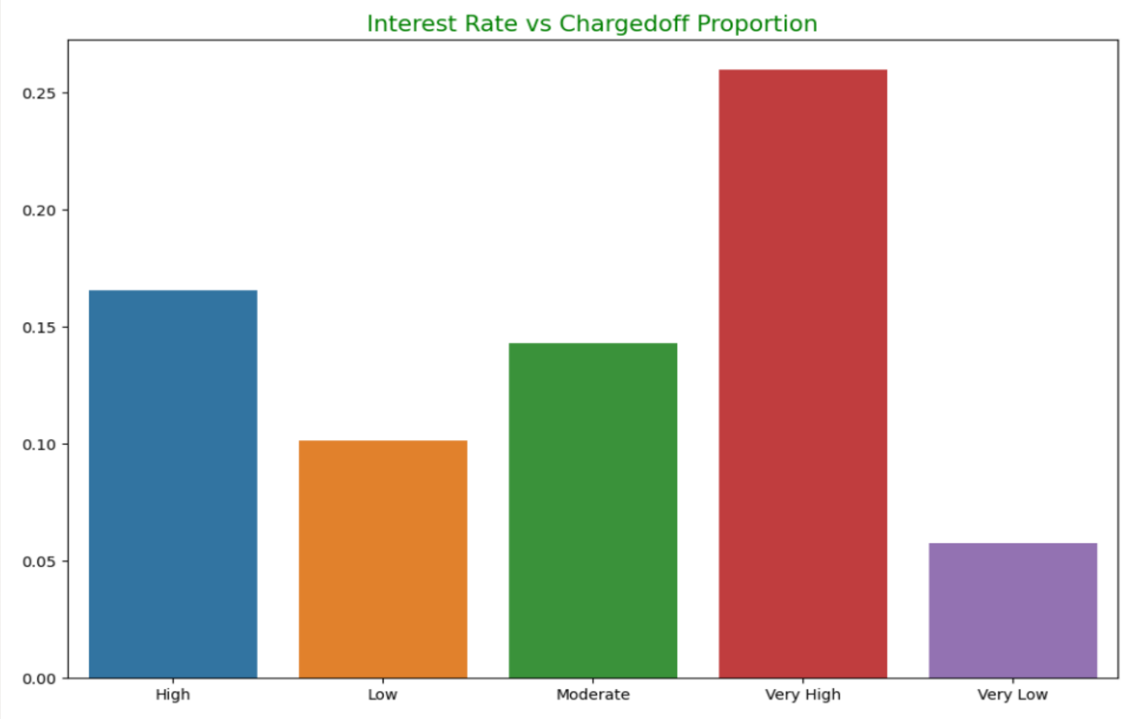


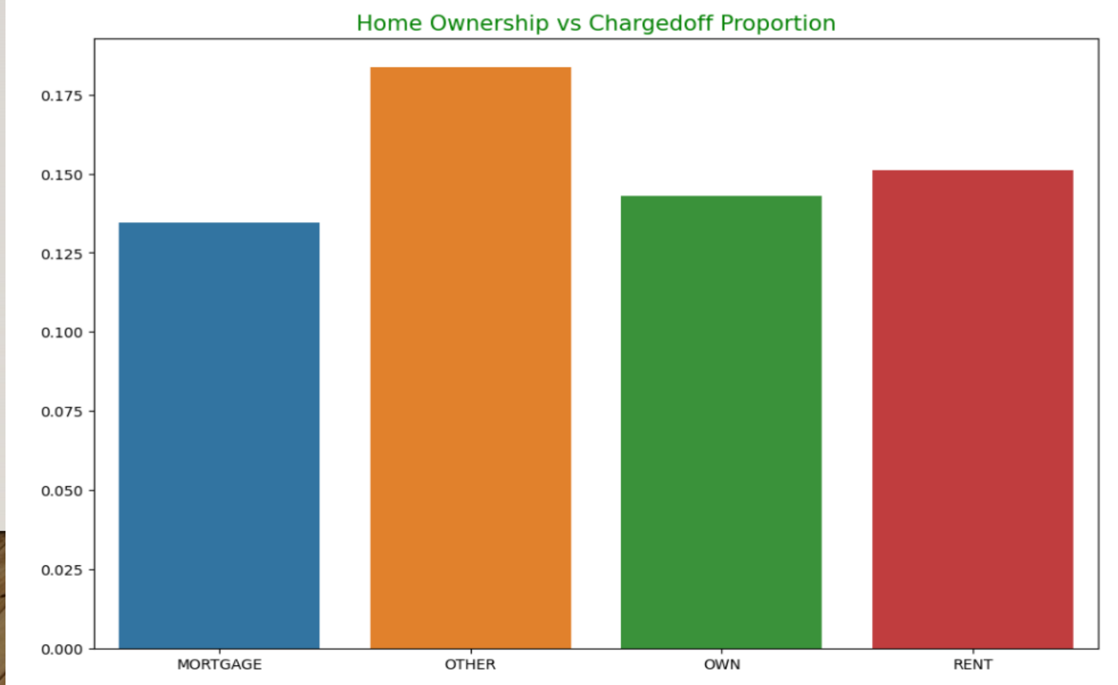

Observations: The data reveals that 85% of borrowers have fully paid off their loans, while 14% have defaulted on their loans

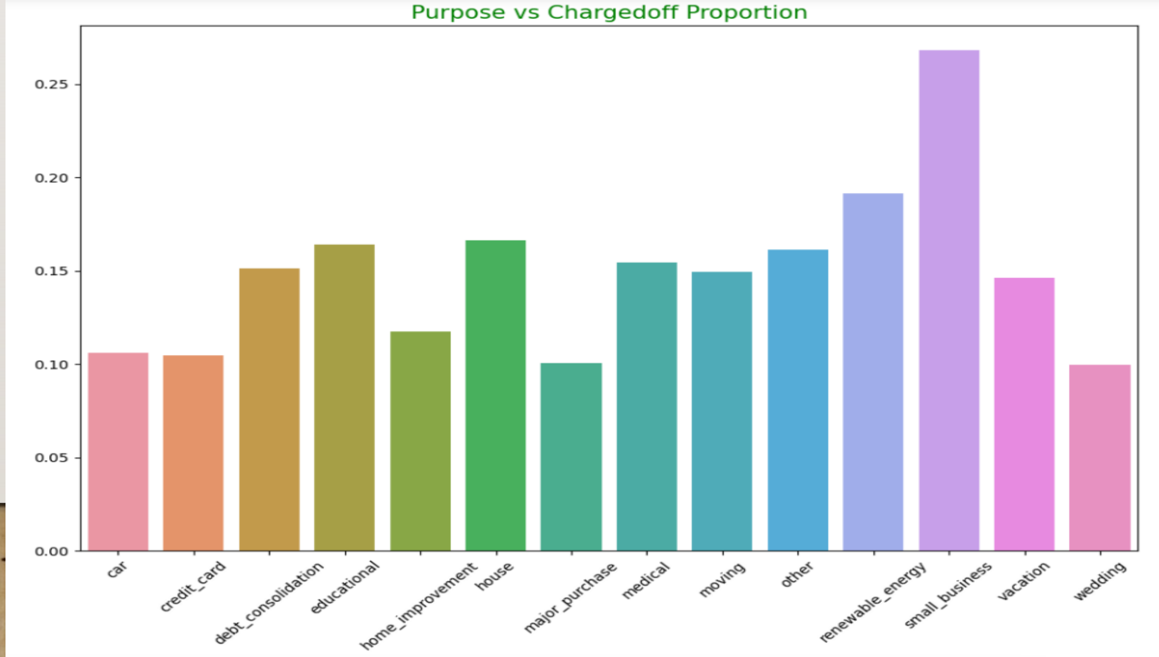


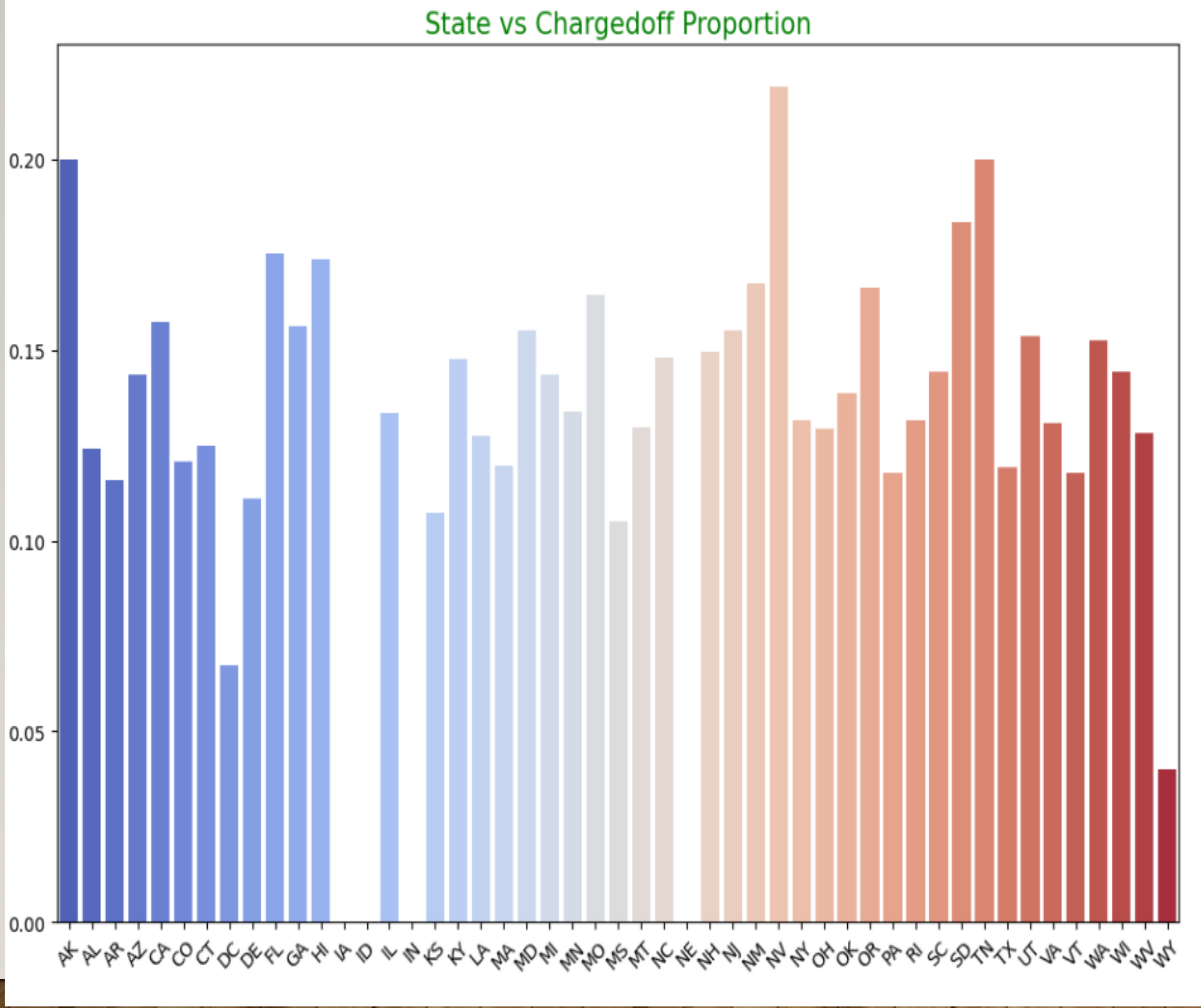

Obersvation: The majority of loan applicants are seeking debt consolidation.

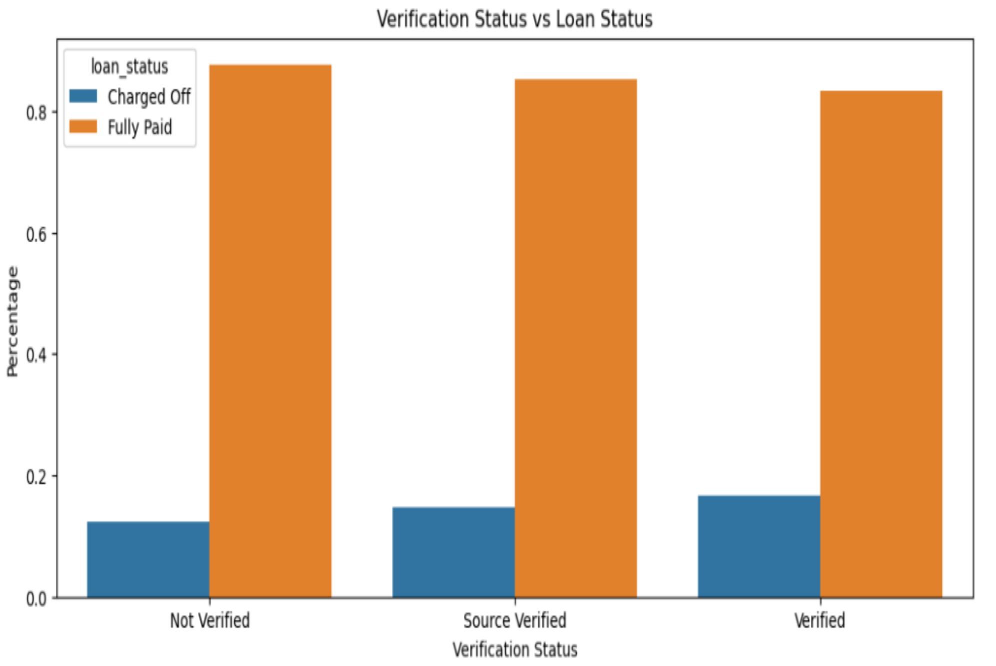

Annual Income vs Charged-off
Charged-off Proportion
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0 - 40k
40k - 50k
50k to 60k
60k to 70k
70k to 80k
80k - above
Annual Income


Interest Rate vs Chargedoff Proportion
0.00
0.05
0.10
0.15
0.20
0.25
High
Low
Moderate
Very High
Very Low

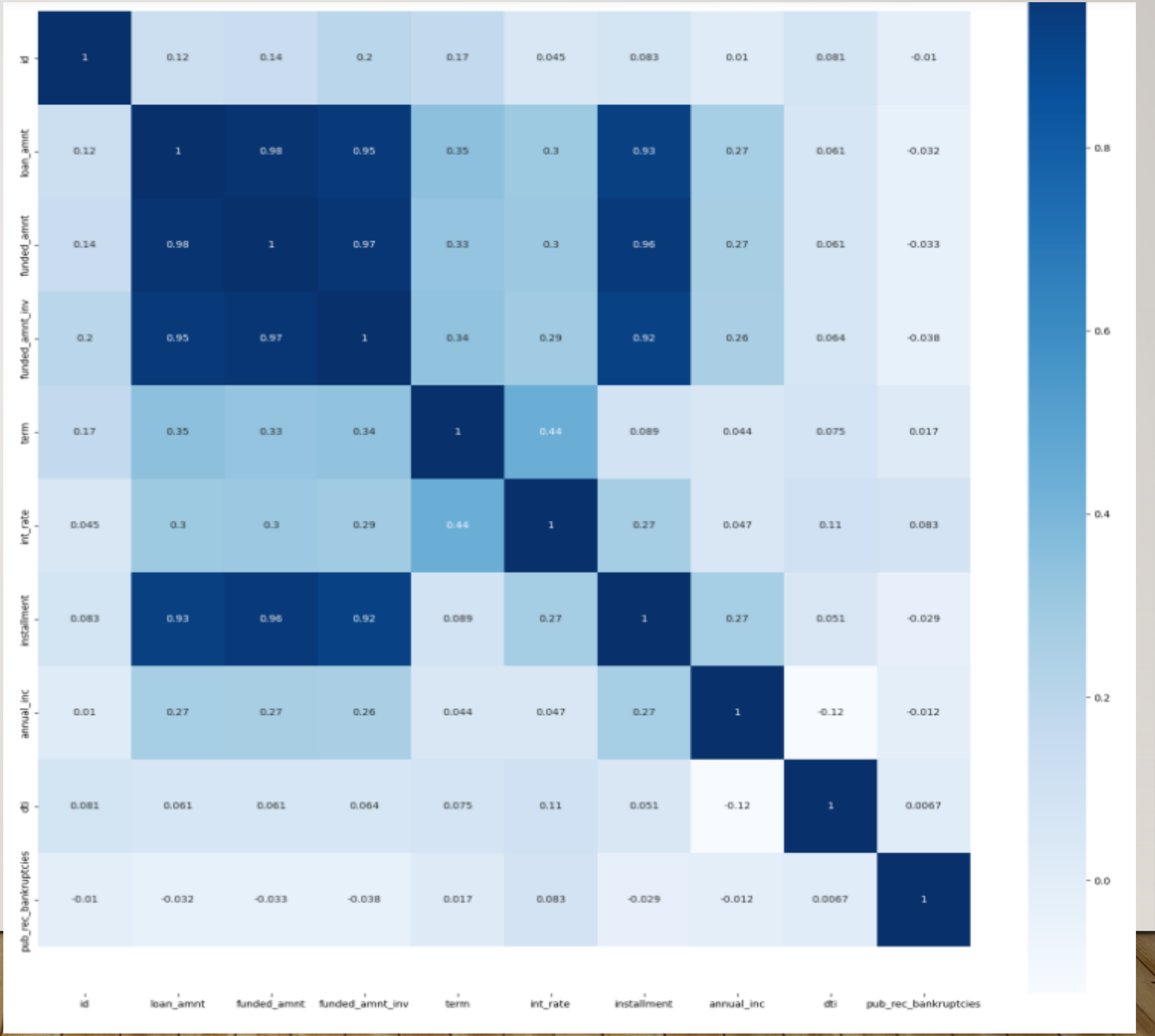

Home Ownership vs Chargedoff Proportion
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
MORTGAGE
OTHER
OWN
RENT

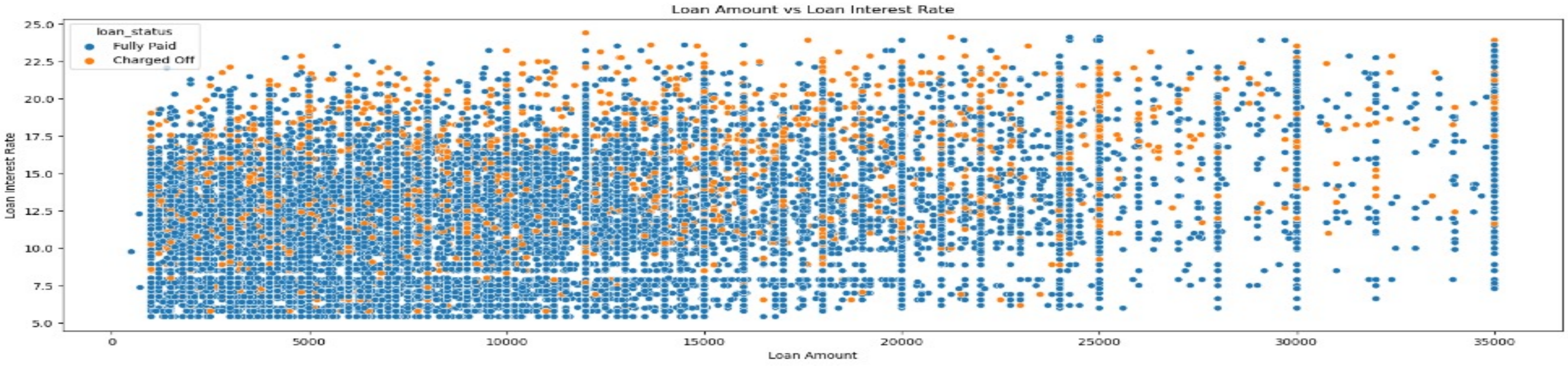

Purpose vs Chargedoff Proportion
0.00
0.05
0.10
0.15
0.20
0.25
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding

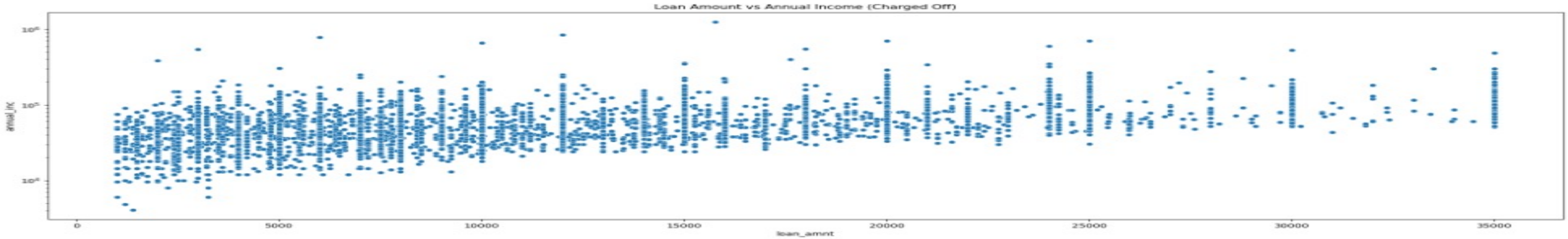

State vs Chargedoff Proportion
0.20
0.15
0.10
0.05
0.00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD MI MN MO MS MT NC NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY

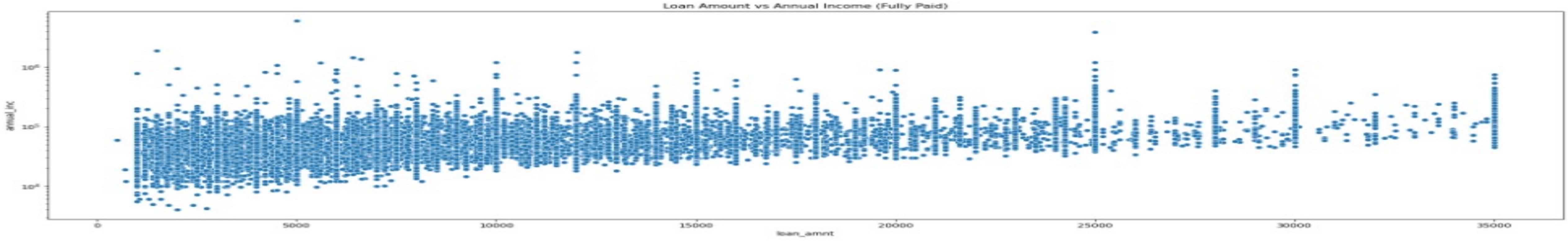

Verification Status vs Loan Status
loan_status
Charged Off
Fully Paid
Percentage
0.0
0.2
0.4
0.6
0.8
Not Verified
Source Verified
Verified
Verification Status

# Correlation between Numeric Columns




Loan Amount vs Loan Interest Rate
loan_status
Fully Paid
Charged Off
Loan Interest Rate
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
0
5000
10000
15000
20000
25000
30000
35000
Loan Amount


Loan Amount vs Annual Income (Charged Off)
annual_inc
0
5000
10000
15000
20000
25000
30000
35000
loan_amnt


Loan Amount vs Annual Income (Fully Paid)
annual_inc
0
5000
10000
15000
20000
25000
30000
35000
loan_amnt

## Conclusion:

- Income ranges between 0 and 20000 exhibit a higher likelihood of defaulting on loans.
- Interest rates exceeding 16% present a significant risk of default compared to other interest rate categories.
- Non-homeowners face an elevated risk of loan default.
- Applicants seeking loans for small businesses are at a heightened risk of default.
- High debt-to-income (DTI) ratios correlate with an increased risk of default.
- Higher numbers of bankruptcies records are associated with a greater likelihood of loan defaults.
- States like Delaware (DE) show the highest number of loan defaults.
- Applicants with loan Grade G demonstrate the highest default rates.

Key factors for predicting default likelihood and mitigating credit losses:

- Debt-to-income ratio (DTI)
- Loan grades
- Verification status
- Annual income
- Public recorded bankruptcies
- Other factors to consider regarding defaults:
- Borrowers residing outside large urban cities such as California, New York, Texas, Florida, etc.
- Borrowers with annual incomes ranging from 50,000 to 100,000.
- Borrowers with public recorded bankruptcies.
- Borrowers assigned lower grades (E, F, G), indicating higher risk.
- Borrowers with very high debt-to-income ratios.
- Borrowers with over 10 years of work experience.