

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:

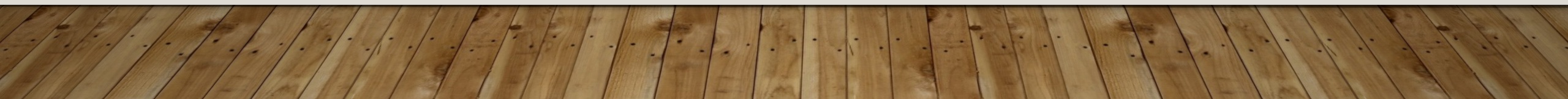
1.Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

- 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 :

In [3]:

```
1 #Display first few rows
2 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]:

```
1 #Display last few rows
2 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]:

```
1 # check dimensions of dataframe , no. of rows and column
2 loan_data.shape
```

Out[5]:

(39717, 111)

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

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 analyzing 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 Behavioral Attribute Columns:

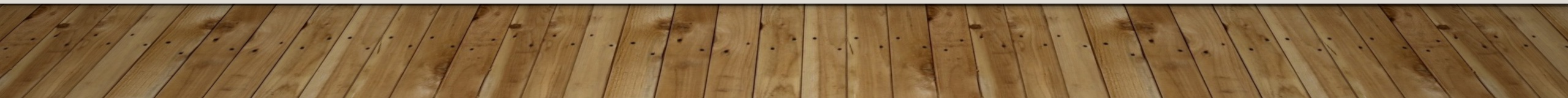
1. Certain columns representing behavioral attributes were identified, which are typically not relevant during the loan approval process and may introduce bias into the analysis.
2. To ensure the analysis focuses solely on pertinent loan attributes, these behavioral attribute columns were excluded from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

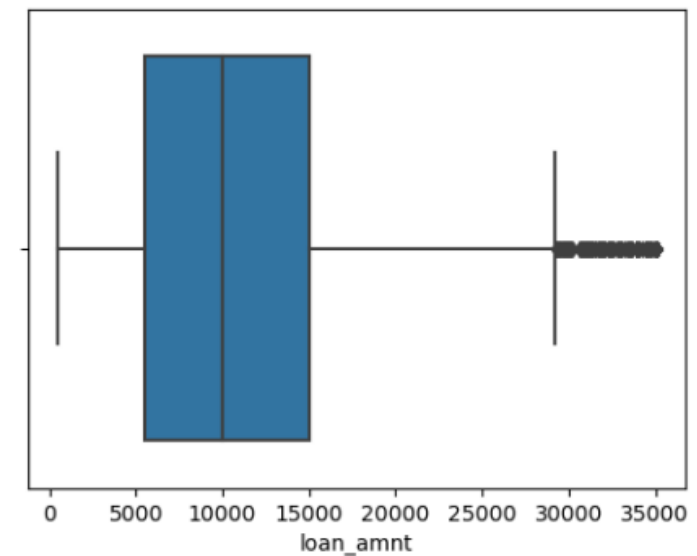
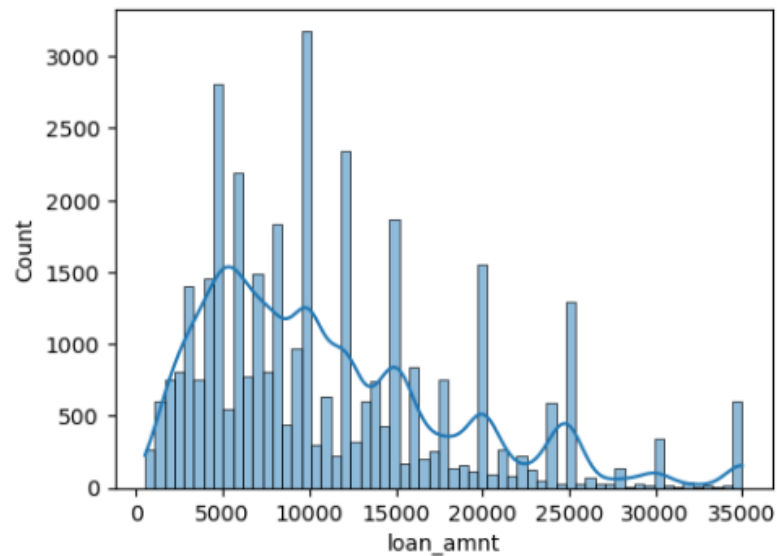
Observation: The distribution of loan amounts indicates that a significant number of individuals opted for a loan amount of "10,000", with the median loan amount also being "10,000". Relatively few individuals chose loan amounts exceeding "30,000".

#Univariate Analysis

Loan

```
In [49]: 1 # Set the figure size
2 plt.figure(figsize=(12, 4))
3
4 # Plot the histogram
5 plt.subplot(1, 2, 1)
6 sns.histplot(data=loan_data, x='loan_amnt', kde=True) # Removed rug=True
7
8 # Plot the boxplot
9 plt.subplot(1, 2, 2)
10 sns.boxplot(data=loan_data, x='loan_amnt')
11
12 # Set a single title for both subplots
13 plt.suptitle('Distribution of Loan Amounts')
14
15 # Show the plot
16 plt.show()
```

Distribution of Loan Amounts



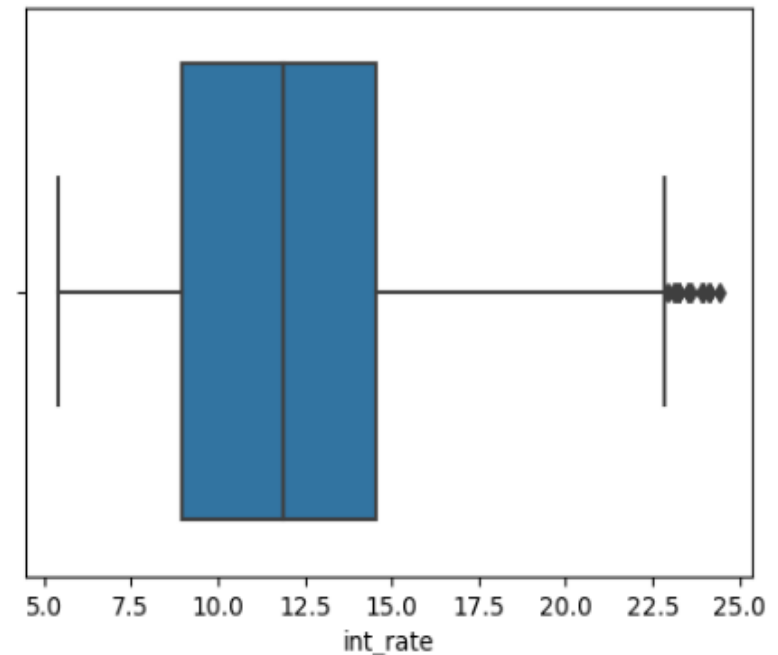
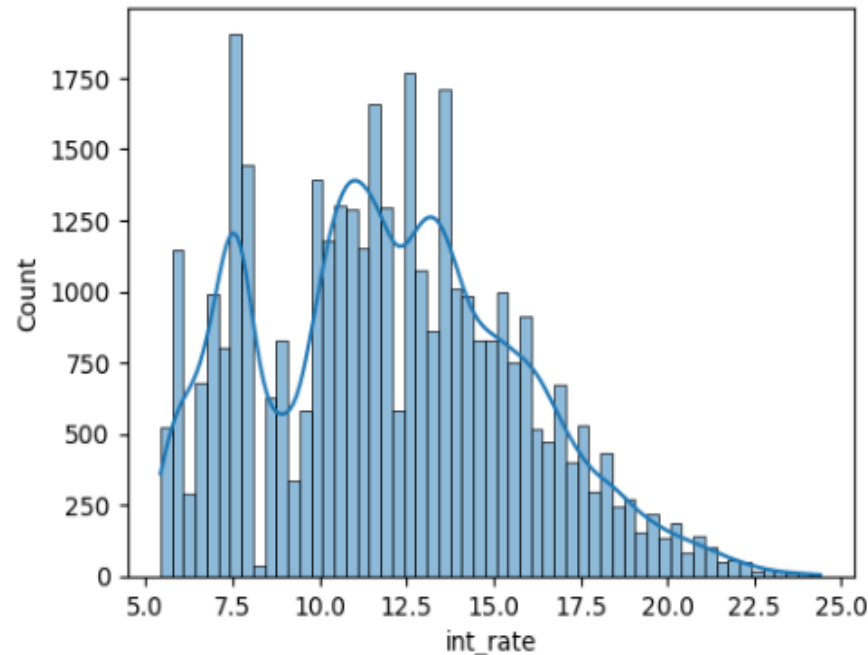
Distribution of Interest Rate

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%.

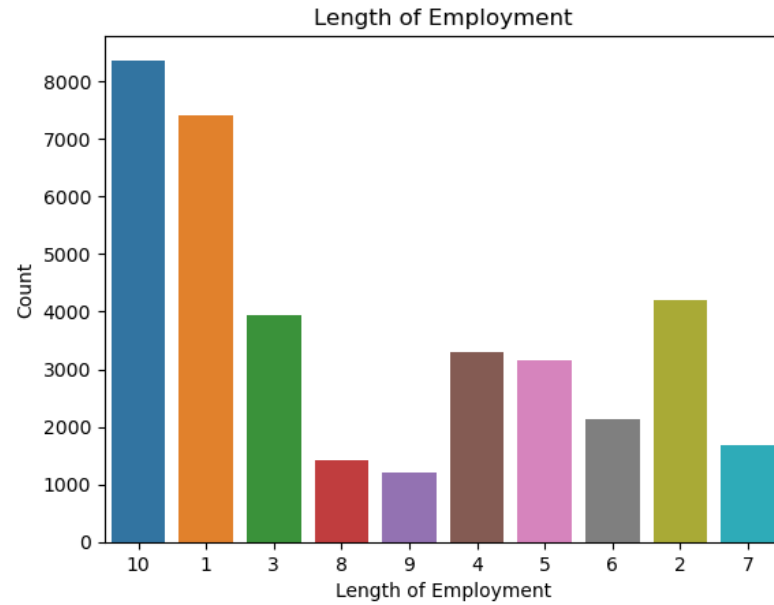
Interest Rate

```
In [53]: 1 # Set the figure size
2 plt.figure(figsize=(12, 4))
3
4 # Plot the histogram using histplot
5 plt.subplot(1, 2, 1)
6 sns.histplot(data=loan_data, x='int_rate', kde=True)
7
8 # Plot the boxplot
9 plt.subplot(1, 2, 2)
10 sns.boxplot(data=loan_data, x='int_rate')
11
12 # Set a single title for both subplots
13 plt.suptitle('Distribution of Interest Rate')
14
15 # Show the plot
16 plt.show()
```

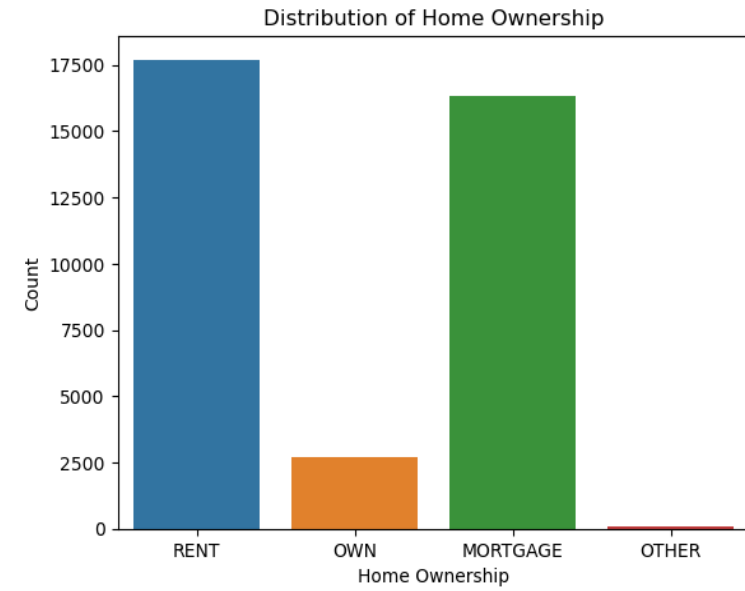
Distribution of Interest Rate



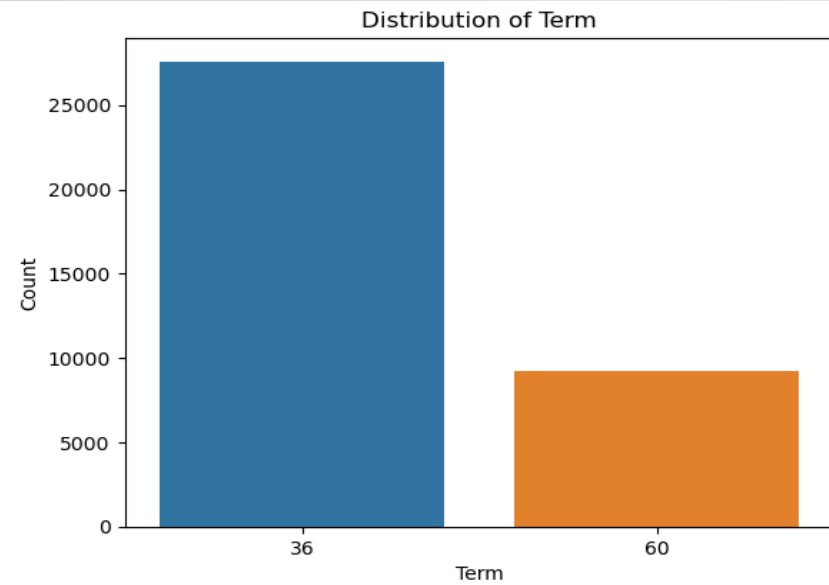
Analysis



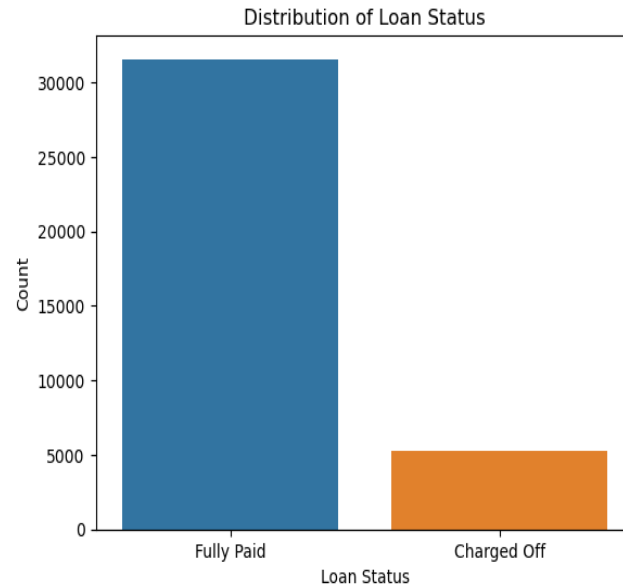
observation: Emplment length is 10+ for most of the borrowers



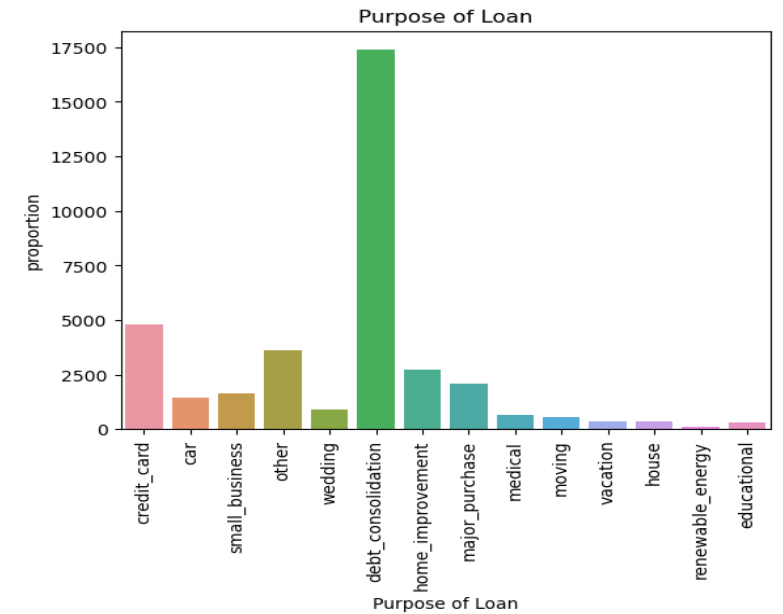
Observation: Most of the home owners are living on Rent or on Mortgage



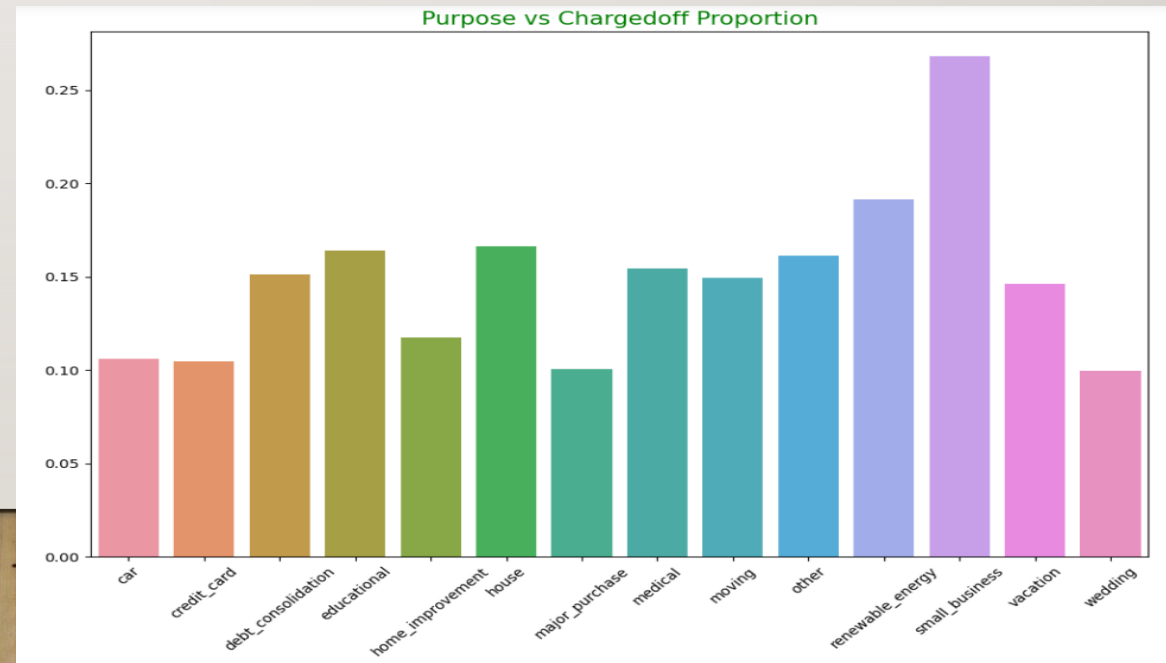
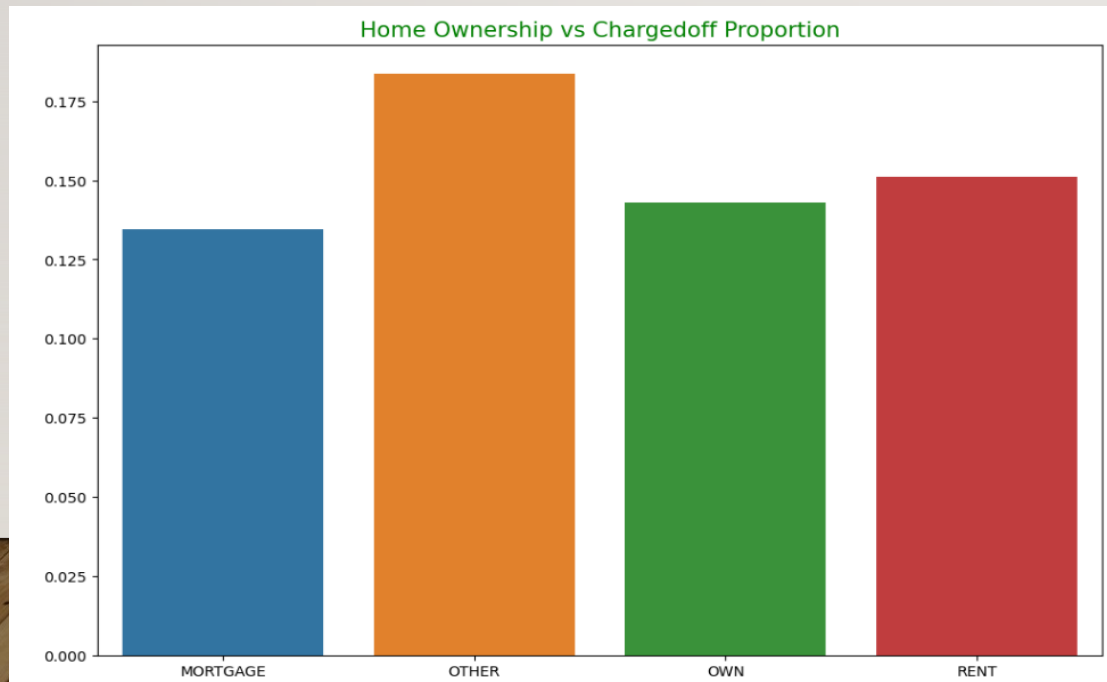
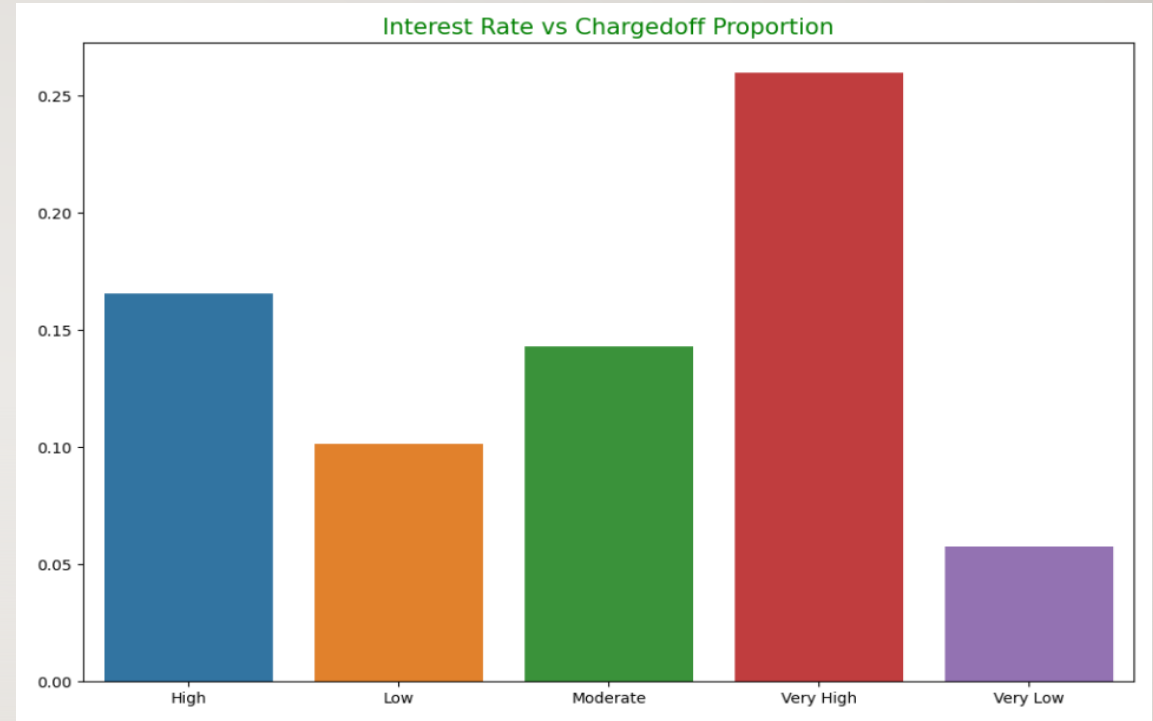
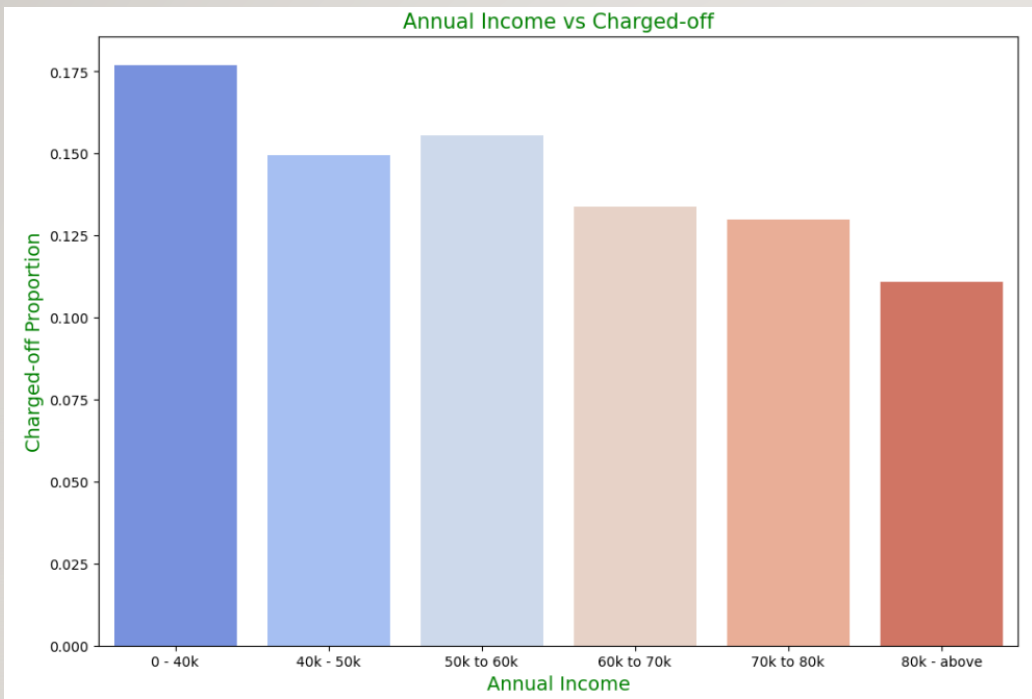
Observation: term is more of 36 months



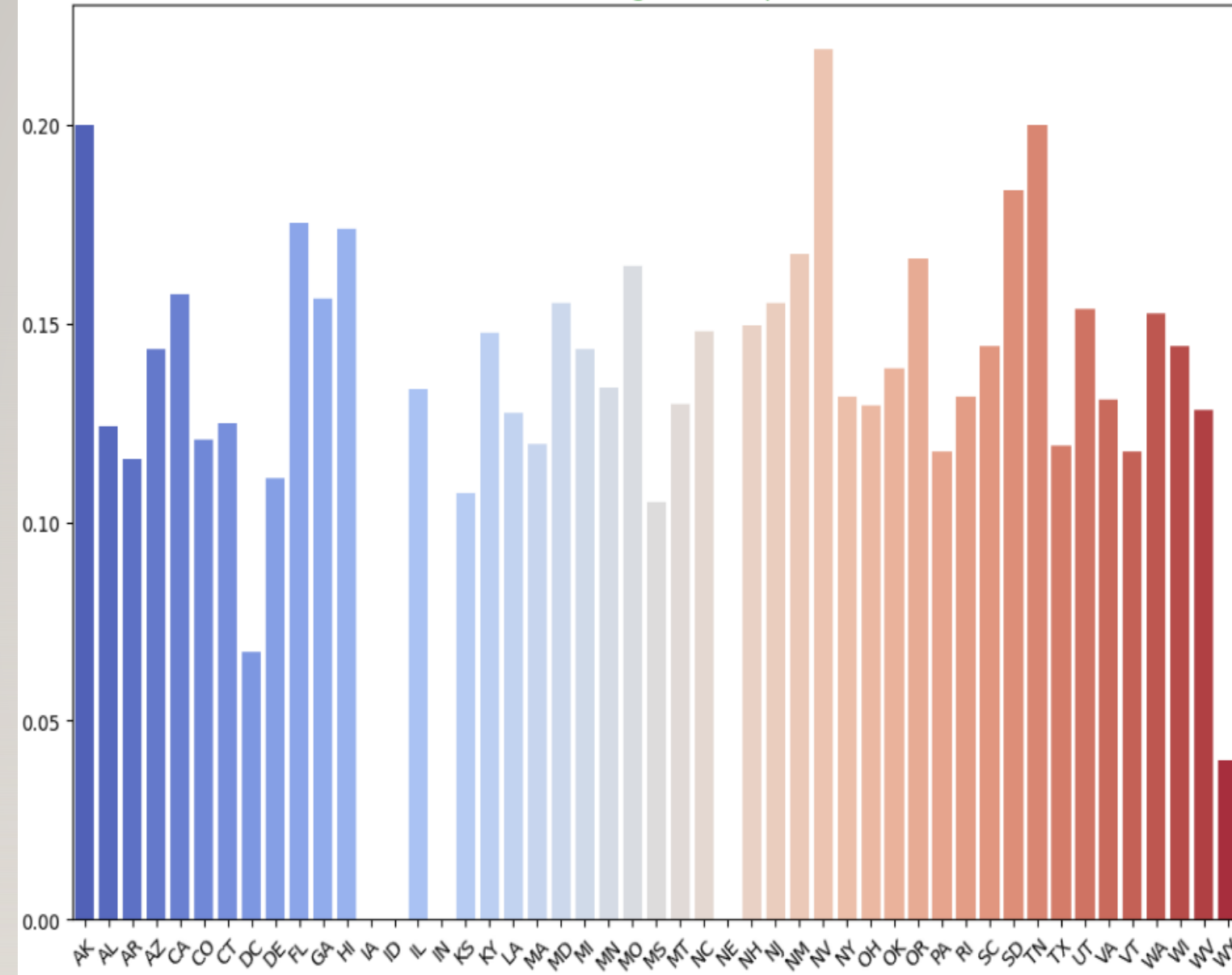
Observations: The data reveals that 85% of borrowers have fully paid off their loans, while 14% have defaulted on their loan



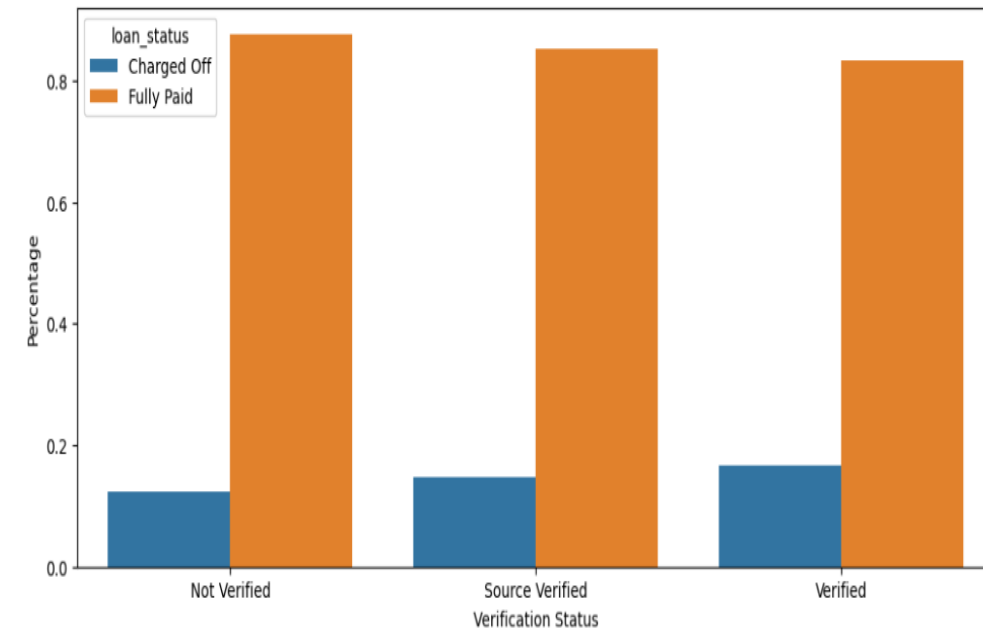
Observation: The majority of loan applicants are seeking debt consolidation.



State vs Chargedoff Proportion



Verification Status vs Loan Status



Correlation between Numeric Columns

