Lending Club Case Study: Exploratory Data Analysis

Submitted by

Himanshu Agrawal

Introduction

- The dataset contains 39,717 rows and 111 columns.
- Goal: Identify factors leading to customers being charged off and gain insights into loan defaults.
- Target Variable: 'loan_status' (Fully Paid, Charged Off, Current).

Loan Status	Count	Percentage
Fully Paid	32950	82.96
Charged Off	5627	14.17
Current	1140	2.87

Steps Followed

1. Null Values Treatment
2. Data Type Treatment
3. Univariate, Bivariate, and Multivariate Analysis
4. Insights

Null Values Treatment

- **Dropped Columns with High Null Values**:
  Removed columns with more than 30% missing values.
- **Imputed Remaining Null Values**:
  Used median for numerical columns and mode for categorical columns.
- **Removed Low Variance Columns**:
  Dropped columns containing only a single unique value.

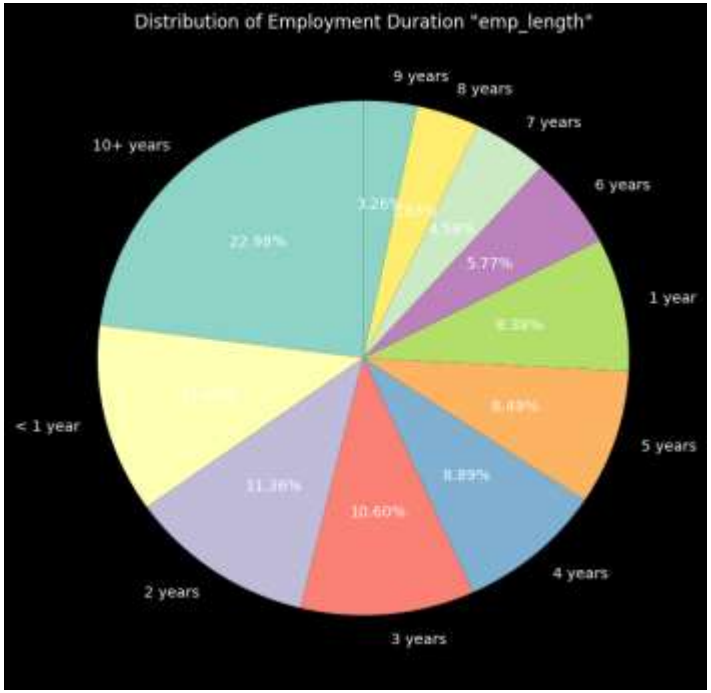
Data Type treatment

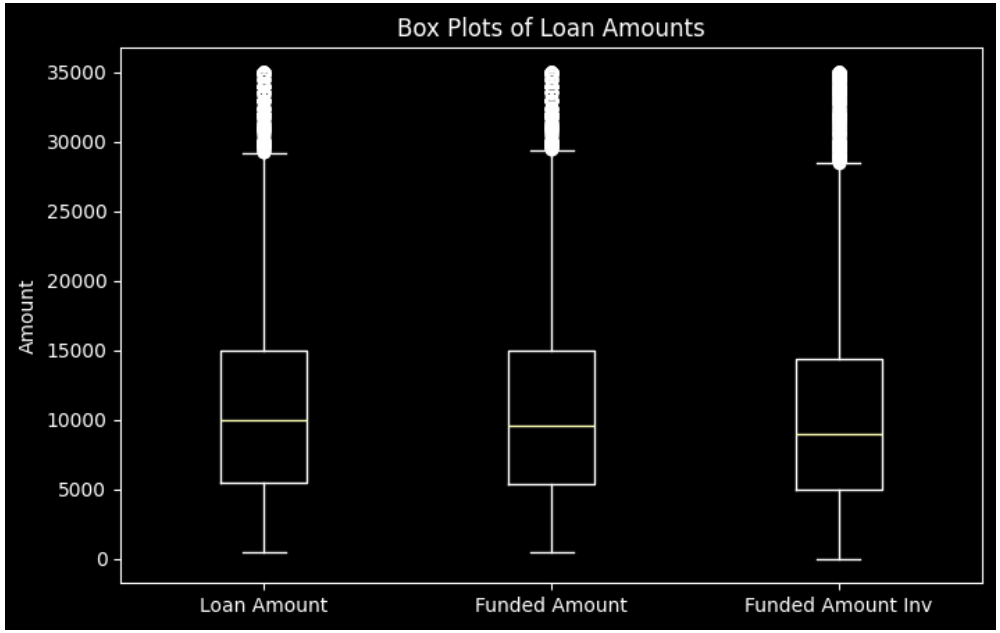
- **Converted Date Columns**:
  Transformed object columns representing dates into datetime objects.
- **Separated Numerical and Categorical Columns**:
  Categorized columns as numerical or categorical for further analysis.
- **Validated Data Types**:
  Ensured numerical columns are correctly typed.
  Rechecked if categorical columns are appropriately labeled.
- **Removed High-Cardinality Columns**:
  Dropped categorical columns with more than 50% unique values.
- **Created a Clean Dataset**:
  Combined refined numerical and categorical columns to form the final clean dataset.

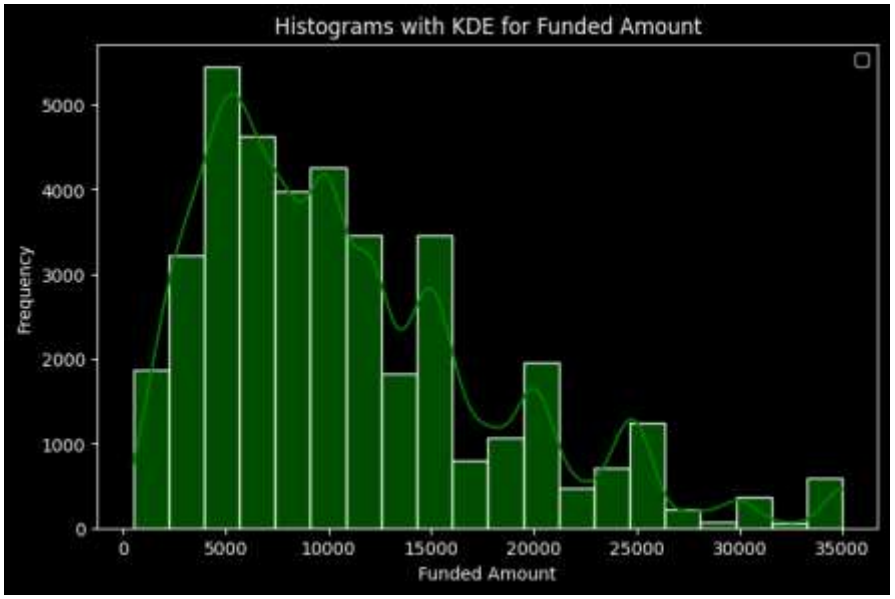
- **Univariate Analysis**:
  Explored individual columns for distribution and trends.
  Plotted histograms, boxplots, and pie charts to identify patterns and outliers.

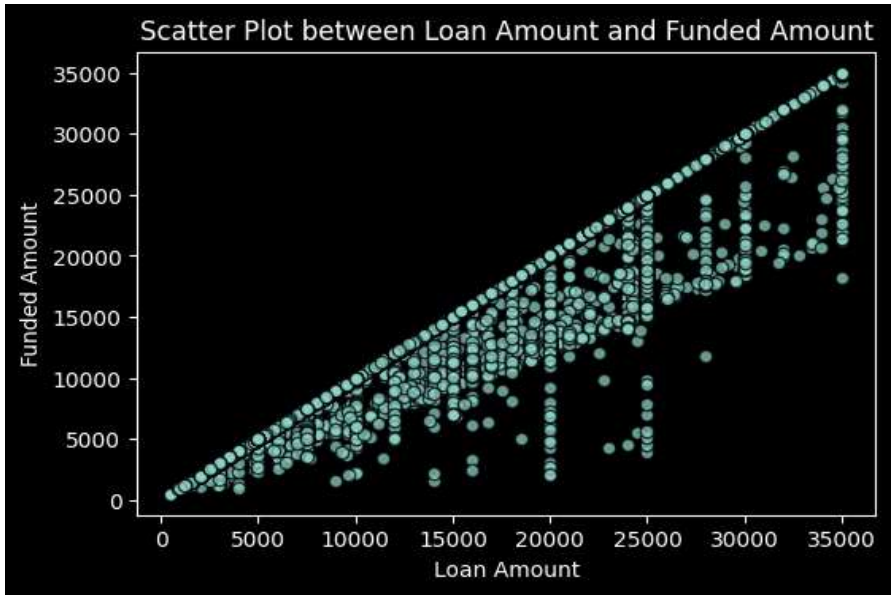
- **Bivariate Analysis**:
  Analyzed relationships between two variables.
  Used scatter plots, bar charts, and pivot tables to identify correlations.
  Compared "Charged Off %" across different categories of key columns.

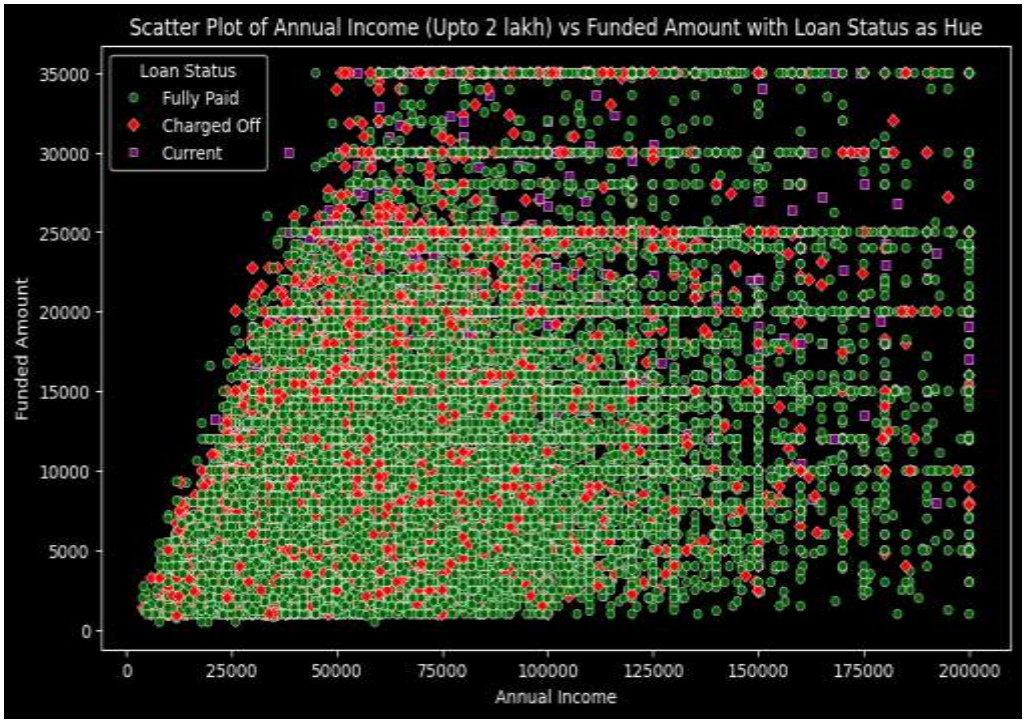
- **Multivariate Analysis**:
  Studied interactions among multiple variables.
  Combined visualizations like scatter plots for deeper insights.
  Identified key drivers influencing the "Charged Off" status.

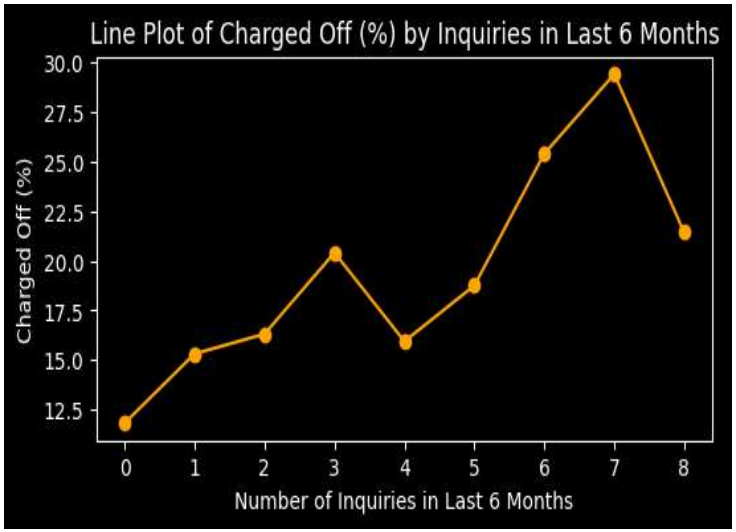

Distribution of Employment Duration "emp_length"
9 years
8 years
7 years
6 years
1 year
5 years
4 years
3 years
2 years
< 1 year
10+ years
22.98%
5.77%
10.60%
8.89%

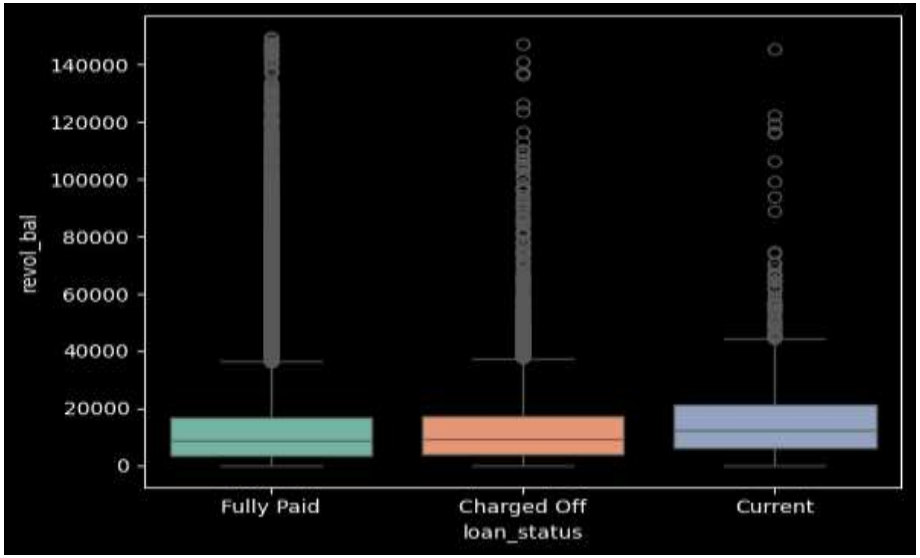

Box Plots of Loan Amounts
Amount
35000
30000
25000
20000
15000
10000
5000
0
Loan Amount
Funded Amount
Funded Amount Inv

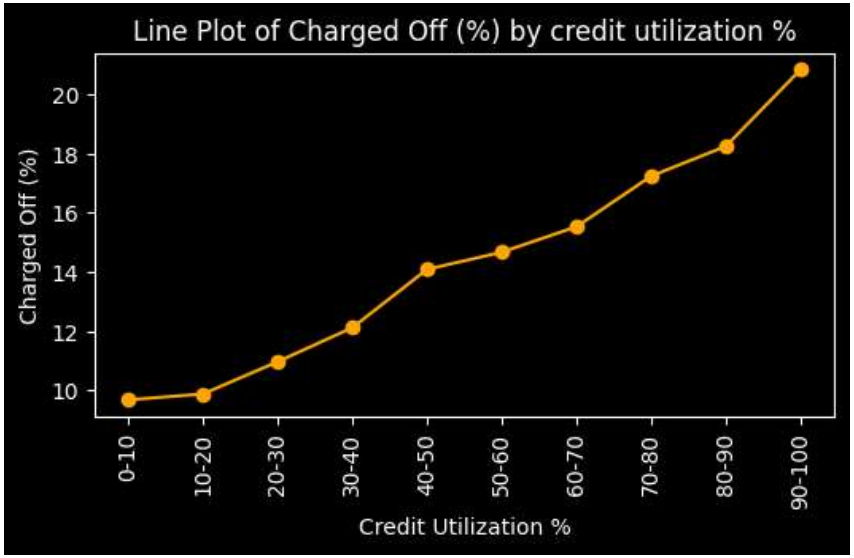

Histograms with KDE for Funded Amount
Frequency
Funded Amount
0
1000
2000
3000
4000
5000
0
5000
10000
15000
20000
25000
30000
35000

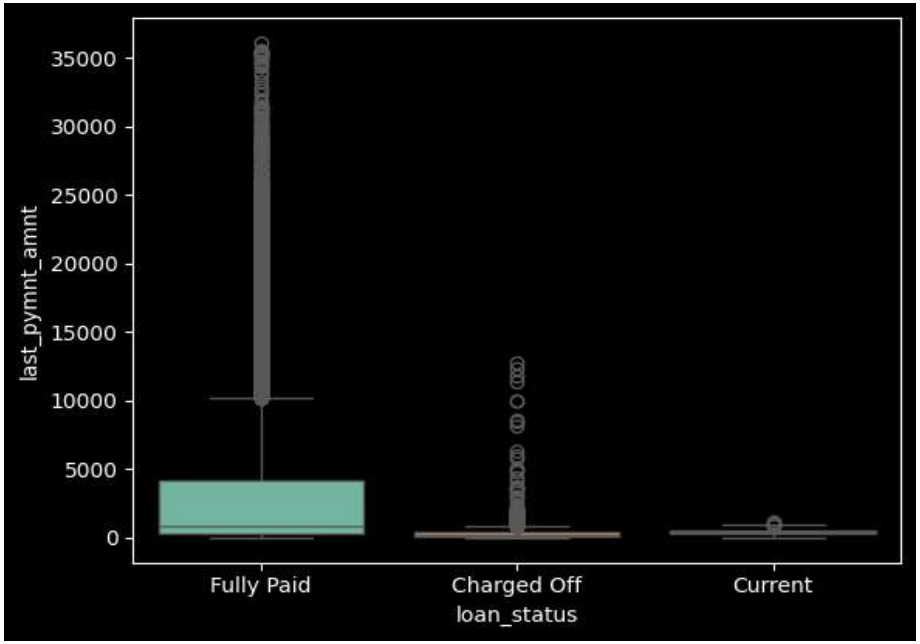

Scatter Plot between Loan Amount and Funded Amount
Funded Amount
Loan Amount
0
5000
10000
15000
20000
25000
30000
35000

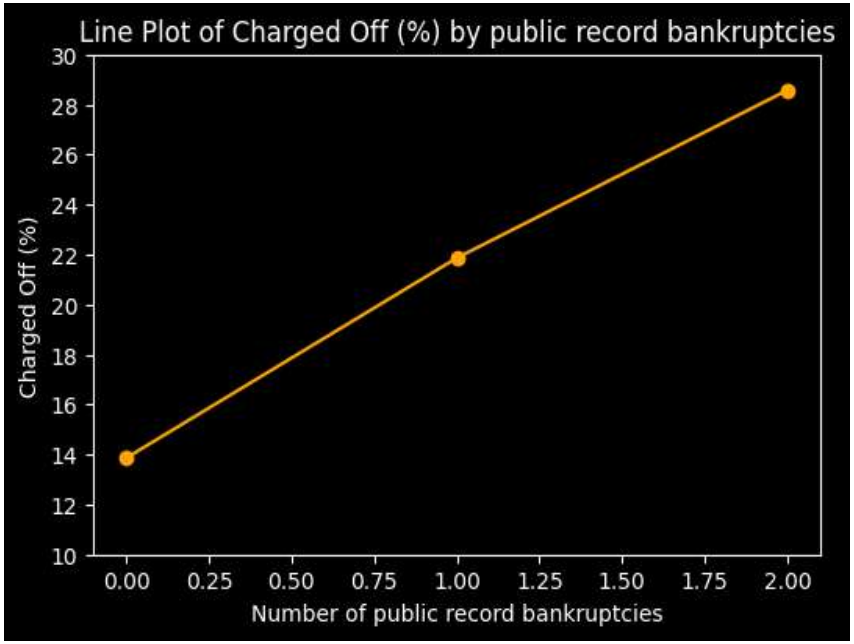

Scatter Plot of Annual Income (Upto 2 lakh) vs Funded Amount with Loan Status as Hue
Loan Status
Fully Paid
Charged Off
Current
Funded Amount
Annual Income
0
5000
10000
15000
20000
25000
30000
35000
0
25000
50000
75000
100000
125000
150000
175000
200000

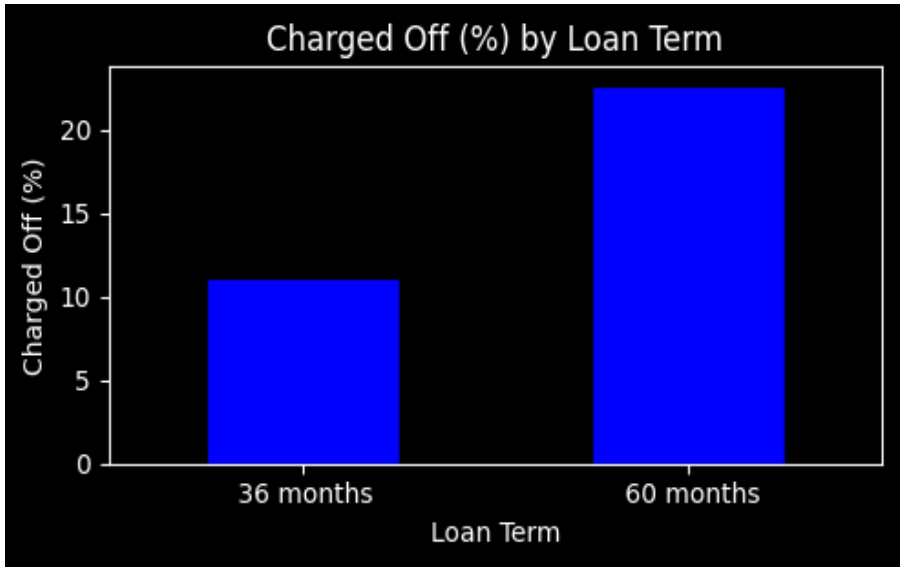

Line Plot of Charged Off (%) by Inquiries in Last 6 Months
Charged Off (%)
Number of Inquiries in Last 6 Months
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0
0
1
2
3
4
5
6
7
8

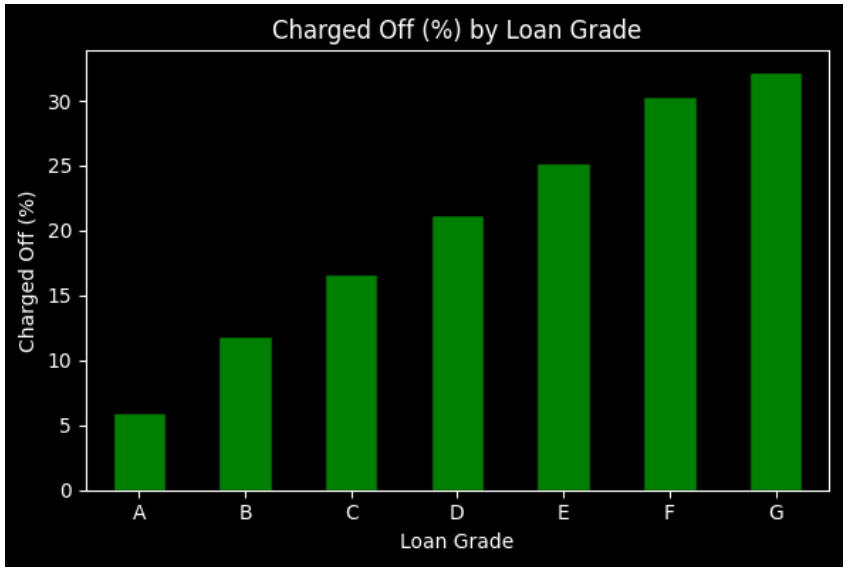

revol_bal
140000
120000
100000
80000
60000
40000
20000
0
Fully Paid
Charged Off
Current
loan_status

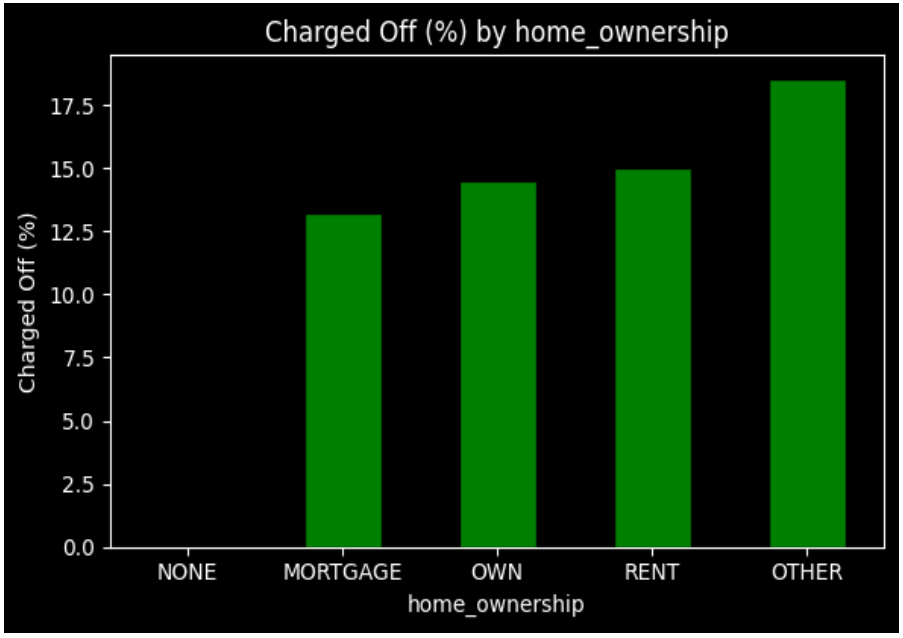

Line Plot of Charged Off (%) by credit utilization %
Charged Off (%)
20
18
16
14
12
10
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Credit Utilization %

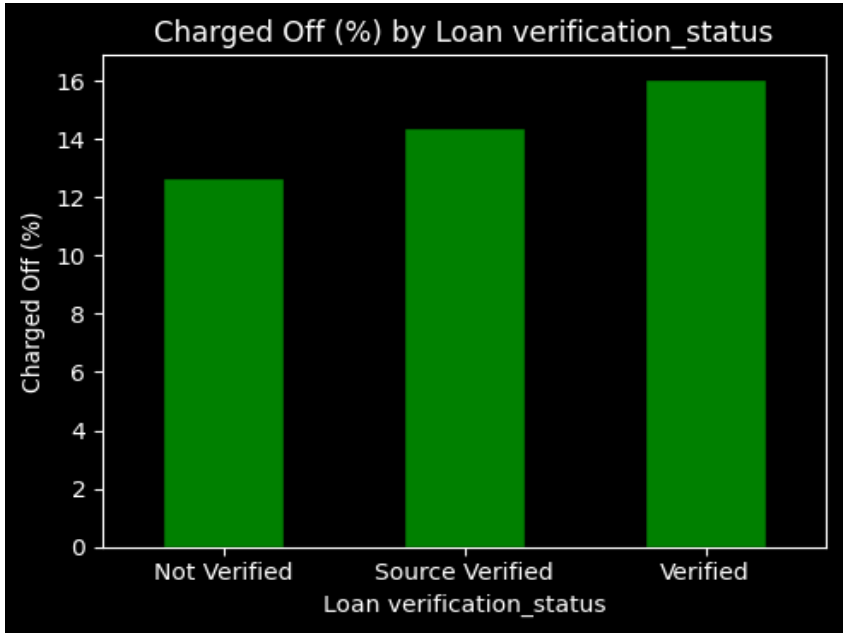

last_pymnt_amnt
35000
30000
25000
20000
15000
10000
5000
0
Fully Paid
Charged Off
Current
loan_status


Line Plot of Charged Off (%) by public record bankruptcies
Charged Off (%)
30
28
26
24
22
20
18
16
14
12
10
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Number of public record bankruptcies


Charged Off (%) by Loan Term
Charged Off (%)
0
5
10
15
20
36 months
60 months
Loan Term


Charged Off (%) by Loan Grade
Charged Off (%)
0
5
10
15
20
25
30
A
B
C
D
E
F
G
Loan Grade


Charged Off (%) by home_ownership
Charged Off (%)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
NONE
MORTGAGE
OWN
RENT
OTHER
home_ownership


Charged Off (%) by Loan verification_status
Charged Off (%)
0
2
4
6
8
10
12
14
16
Not Verified
Source Verified
Verified
Loan verification_status

Key Insights

1. As Annual income is rising, installment amount is also rising till annual_inc < 1 Lakh. After that it is random.
2. As annual income is increasing, funded amount has also increased but this is limited to annual income < 1 Lakh, after that it is random
3. With increasing number of inquiries in last 6 months, there are more chances of being charged off
4. Higher the credit utilization %, higher the chances of being charged off.
5. Those who have paid (total payment) equal to or less than 100% of the funded amount, they are charged off
6. Those who have paid almost 80% or lesser of the principal amount, are 'Charged Off'
7. Last payment amounts are way higher in case of 'Fully Paid' than the other two categories - Charged Off, Current

8. With increasing number of public record bankruptcies, there are more chances of being charged off.
9. Loans with a 60-month term have twice the likelihood of being charged off compared to those with a 36-month term.
10. From A to G in Loan Grades, likelihood of being charged off is also increasing.
11. As we go from A1 to G5 in Loan subgrade, the % of being charged off is increasing.
12. There are slight variations of being charged off in case of home_ownership which is OTHER > RENT > OWN > MORTGAGE. OTHER has the highest chances of being charged off.
13. Surprisingly, the loans which are verified have higher chances of being charged off than those which are not verified.
14. Loans for small business purpose is at the most risk and chances are high of being charged off than other categories.

Conclusion

- 1. Longer loan terms (60 months) carry higher default risk.
- 2. High credit utilization and low last payment amounts signal risk.
- 3. Loan verification processes need reevaluation.

Thank You