12/24/2024

ANALYZING LENDING CLUB DATA TO PREDICT LOAN DEFAULTS

Using Python for Exploratory Data Analysis

Project By: Vijay Kumar & Nagammai Shanmugham

Problem Statement

Objective:

- To understand the factors that contribute to loan defaults in the Lending Club dataset and develop insights for risk assessment.

Key Questions:

- What are the key characteristics of borrowers who default on their loans?
- Are there specific loan purposes, loan amounts, or borrower demographics associated with higher default rates?
- How do factors like income, credit history, and debt-to-income ratio influence loan performance?

Data Set Understanding

Data Understanding

Data Source: Lending Club Dataset

Key Variables:

- **Target Variable:** Loan Status (e.g., "Fully Paid", "Charged Off")
- **Independent Variables:**
- loan_amnt, int_rate, installment, annual_inc, dti, home_ownership, verification_status, purpose, addr_state, emp_length, grade, sub_grade, revol_bal, total_acc, pub_rec_bankruptcies (and other 111 variables)

	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_12m	pct_tl_nvr_dlq
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	NaN	NaN
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	NaN	NaN
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	NaN	NaN
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	NaN	NaN
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	NaN	NaN

5 rows × 111 columns

Checking the row and column of data

```
df.shape
```

```
(39717, 111)
```

```
df1.isna().sum()
```

```
loan_amnt                0
funded_amnt              0
funded_amnt_inv          0
term                     0
int_rate                 0
installment              0
grade                    0
sub_grade                0
emp_length            1033
home_ownership           0
annual_inc               0
verification_status      0
issue_d                  0
loan_status              0
purpose                  0
addr_state               0
dti                      0
pub_rec_bankruptcies   697
dtype: int64
```

```
Emp length has years so we are cleaning the data and conv
```

Cap extreme values in the 'annual income' column at the upper bound

```
annual_income_upper_bound = 145000.0
df1['annual_inc'] = np.where(df1['annual_inc'] > annual_income_upper_bound, annual_income_upper_bound, df1['annual_inc'])

# Confirm changes to the 'annual income' column
df1['annual_inc'].describe()
```

```
count     38577.000000
mean      65044.917784
std       32652.937415
min        4000.000000
25%       40000.000000
50%       58868.000000
75%       82000.000000
max      145000.000000
Name: annual_inc, dtype: float64
```

Calculate Loan-to-Income Ratio (lti)

```
df1['lti'] = df1['loan_amnt'] / df1['annual_inc']
df_charged_off['lti'] = df_charged_off['loan_amnt'] / df_charged_off['annual_inc']
```

Python

Numerical Variables

- loan_amnt
- funded_amnt
- funded_amnt_inv
- term
- int_rate
- installment
- annual_inc
- dti

Categorical Variables

- grade
- emp_length
- home_ownership
- verification_status
- issue_d
- loan_status (target variable)
- purpose
- addr_state
- pub_rec_bankruptcies

Data Preprocessing

Handling Missing Values: (e.g., imputation, removal):

- Removing columns with more than 40% values – Reduced column count: 54
- Removing columns with duplicated data and where borrowers are active – New column count: 45
- Dropping additional columns that are with text data, irrelevant information – New column count: 18

Data Cleaning: (e.g., outlier detection, data type conversions):

- Emp_length,term, int_rate, & Pub_rec_bankruptcies – Converted object data type to integer and corrected the data
- **Feature Engineering:** (e.g., creating new features like loan_to_income_ratio)
- **Encoding Categorical Variables:** (e.g., one-hot encoding for home_ownership)
- Outlier detection: Capped upper bound of annual income

Univariate Analysis – Loan Status

Visualizations:

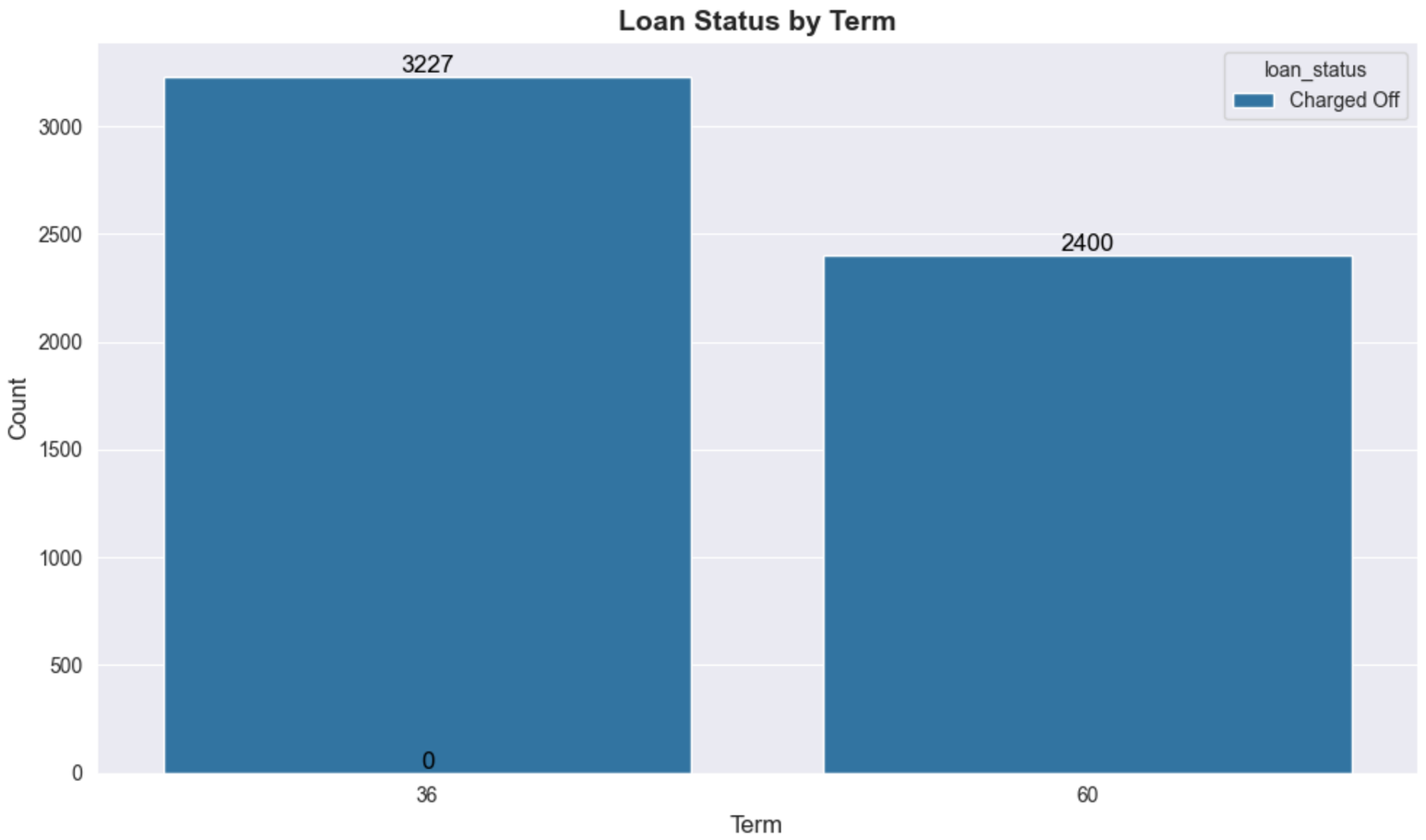
- Bar chart of loan status distribution (proportion of "Fully Paid" vs. "Charged Off" loans)

Insights:

- Overall distribution of loan statuses (e.g., percentage of defaults)

Overview:

- 57% of individuals with a term of 36 are exhibiting higher default.



Univariate Analysis – Numerical Variables

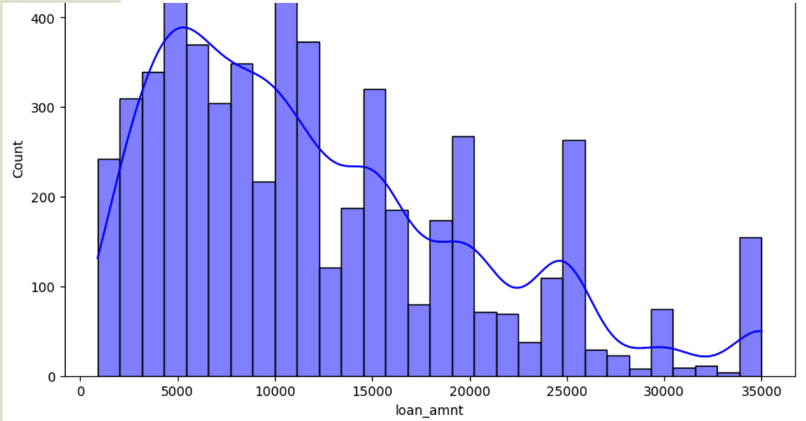
Visualizations:

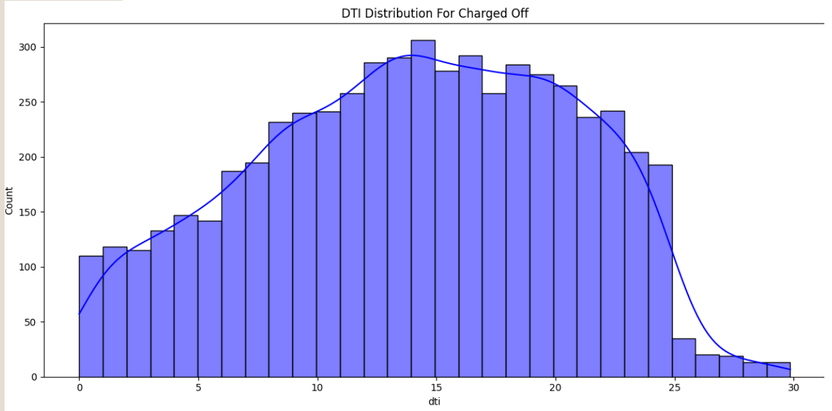
- Histograms for loan_amnt, annual_inc, int_rate, dti

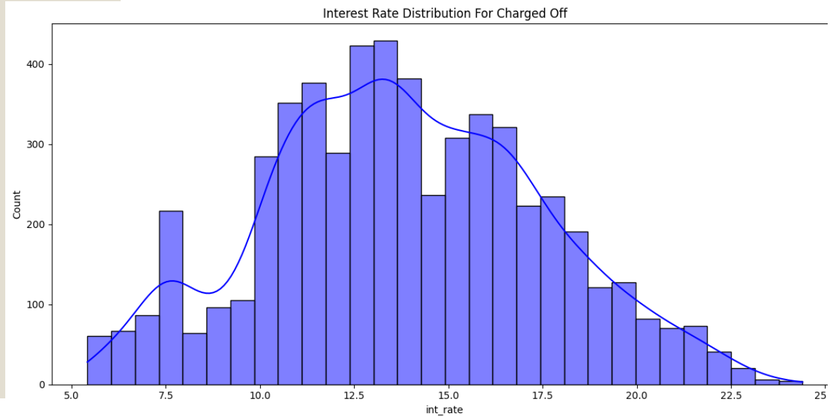
Insights:

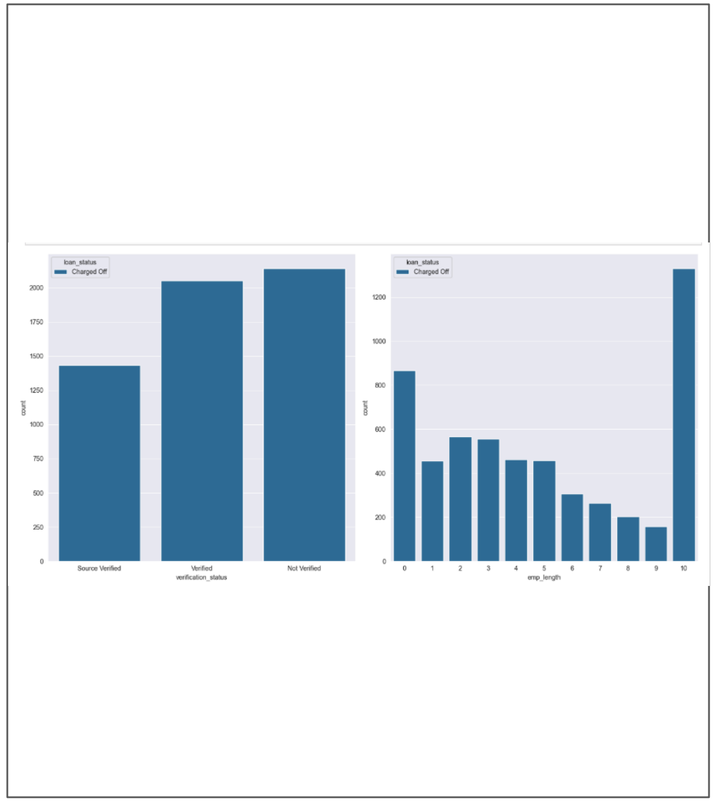
- Distributions of key continuous variables
- Identification of potential outliers

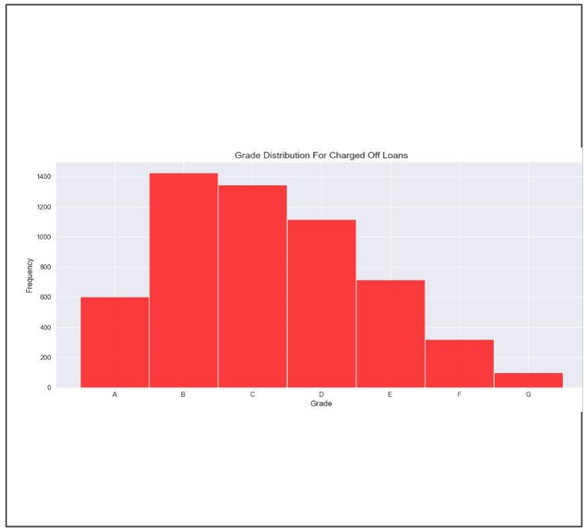
- Individuals who have borrowed amounts ranging from 5,000 to 16,000 are more likely to default.
- There is a noticeable decline in charged-off loans for larger loan amounts, with a significant drop after $20,000. This suggests that larger loan amounts may be associated with a lower likelihood of being charged off.
- The distribution shows a clear peak in the frequency of charged-off loans within the interest rate range of 10-16%.
- People whose DTI is between 9 and 20 are more likely to not pay back their loans.

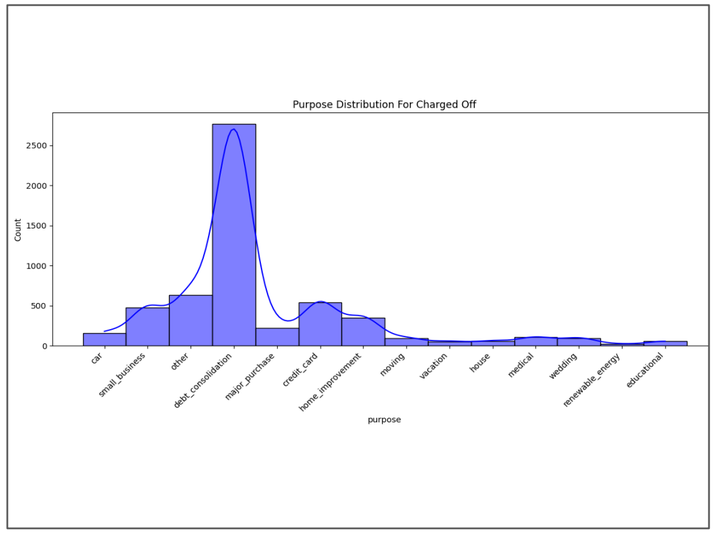


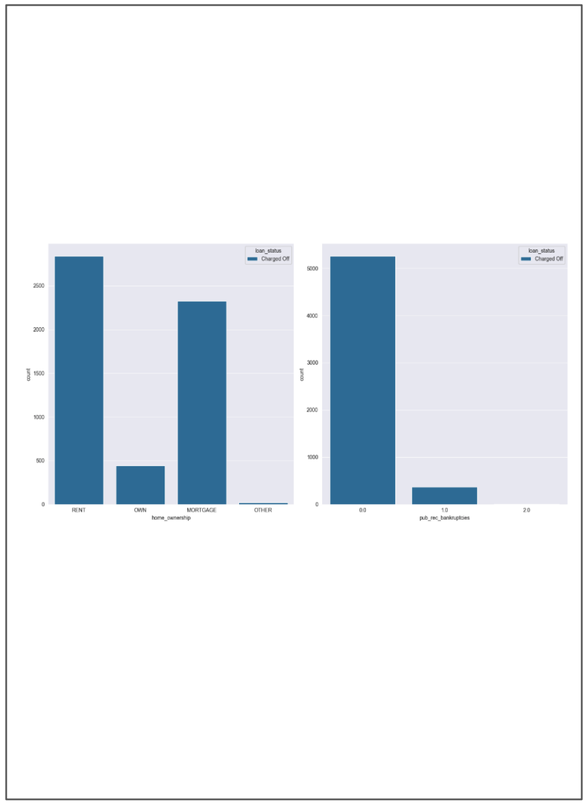












Univariate Analysis – Categorical Variables

- **Visualizations:**
  - Bar charts for home_ownership, emp_length,purpose, pub_rec_bankruptcies, grade
- **Insights:**
  - Loans with LC verified status and not verified, both pose higher risk of charged-off instances
  - Grade B has the highest default rate, and defaulters are trending downward from B to G. Out of all grades, Grade G has the least defaulter
  - Borrower with shorter employment length less than year and 10 or 10+ year employment have higher likelihood of default
  - Borrowers with rented ownership and mortgaged homes, have highest number of charged off loans
  - Borrowers who have not filed for bankruptcy in the public record have a significantly higher number of charged-off loans (approximately 93 percent).

Bivariate Analysis - Correlations

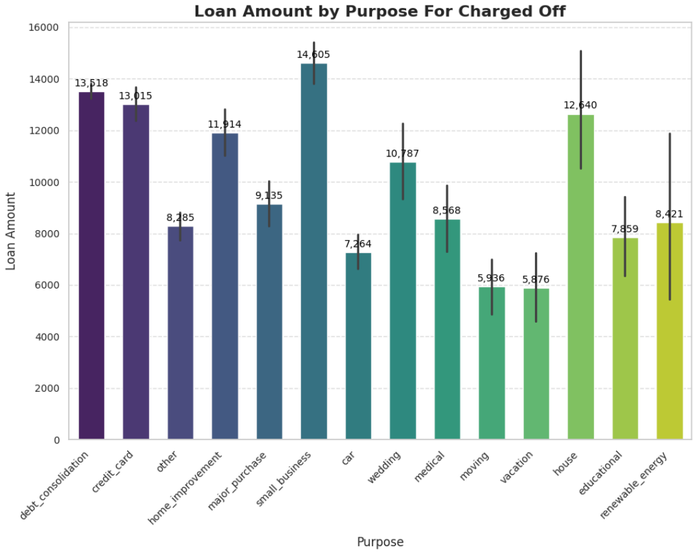
Visualizations:

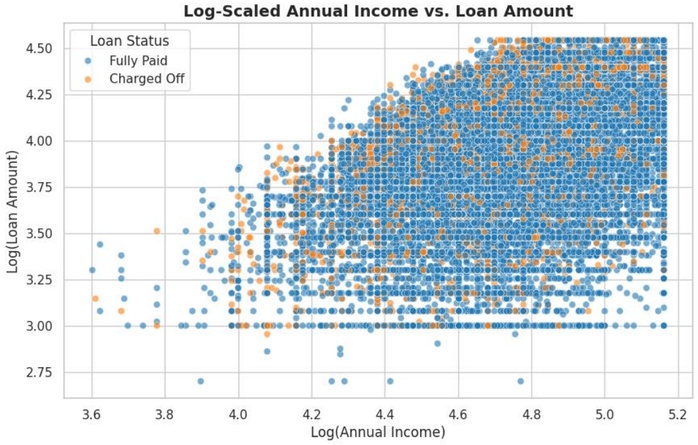
- Bar pot to understand the density and distribution of loan purpose.
- Scatter plots for relevant variable pairs (e.g., loan_amnt vs. annual_inc, int_rate vs. grade)

Insights:

- Borrower with purpose 'debt_consolidation', 'credit_card', 'small_business' and 'house' have taken highest average loan amount andthey were also the most likely to not pay back their loans, so lending case can be more cautious while proving loan for these loan purpose.

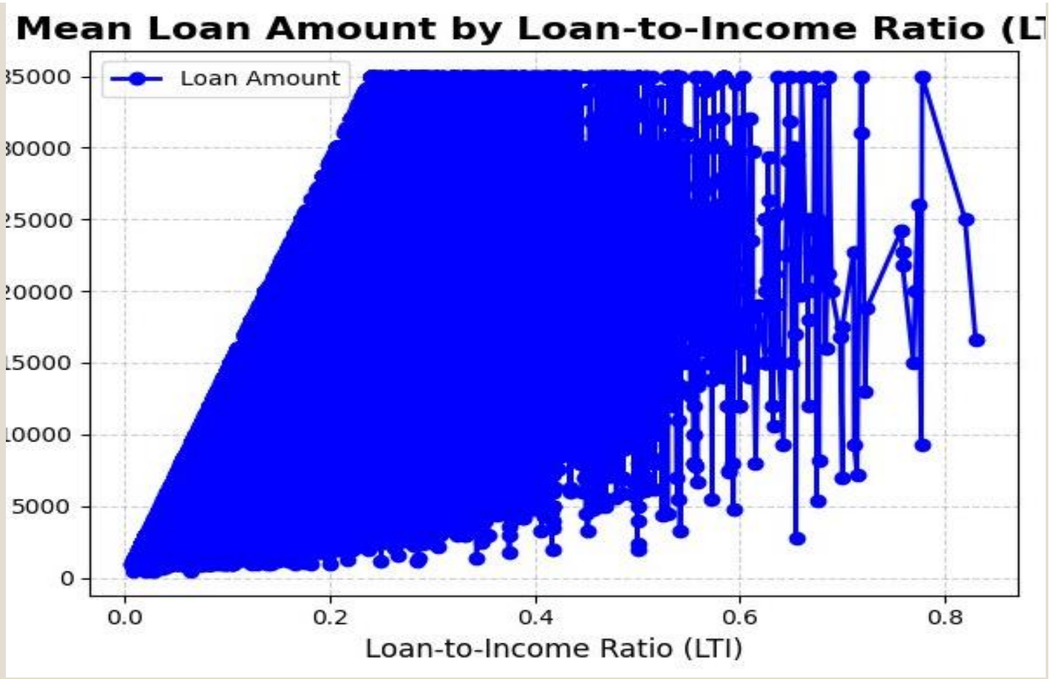
- Diversification of loan purposes might be beneficial. Lending club could consider diversifying their loan portfolio to include more loans for purposes with lower default rates, such as "Education" and "Renewable energy" loans.





Visualizations:

- Line plot for Loan to income – Derived variable.
- Histogram to plot continuous variables – Annual income and loan amount.



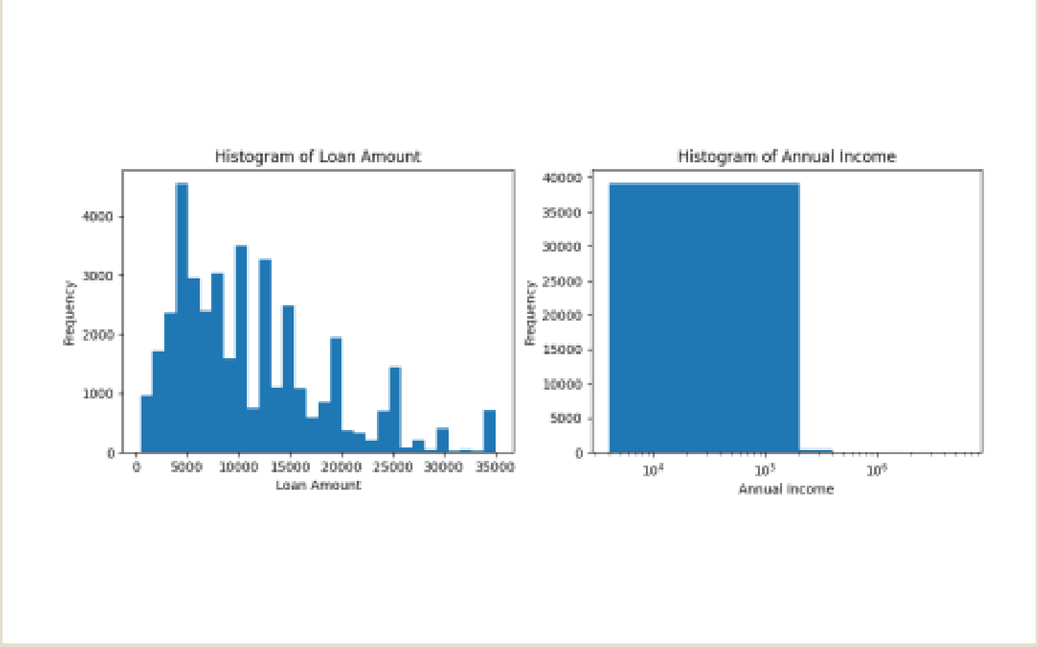
Insights:

- There appears to be a positive correlation between the Loan-to-Income Ratio (LTI) and the Mean Loan Amount. LC may be more willing to provide larger loans to borrowers with higher LTI ratios, assuming they can comfortably afford the debt burden.

- The majority of loans are issued to borrowers with lower LTI ratios, suggesting that lenders might be more cautious when approving loans to borrowers with higher debt-to-income ratios.

- People with LTI ratio less than 0.6 are more likely to default.

- Borrower with annual income between 37000 to 75000 have high changes of defaulting



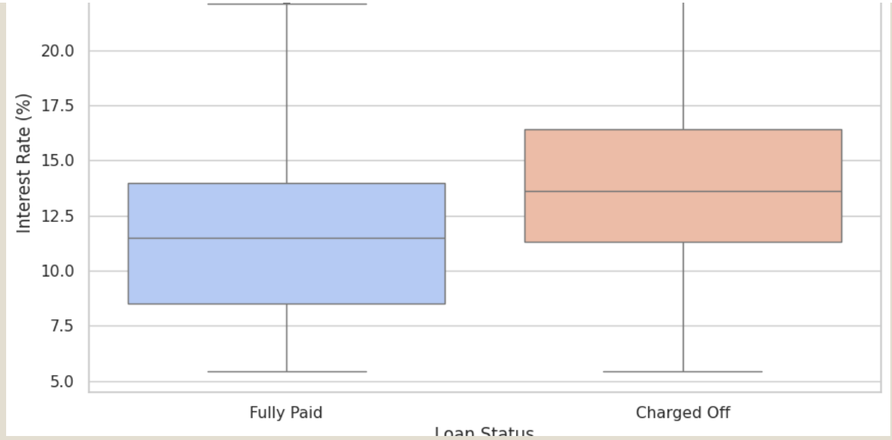
Bivariate Analysis - Loan Status vs. Other Variables

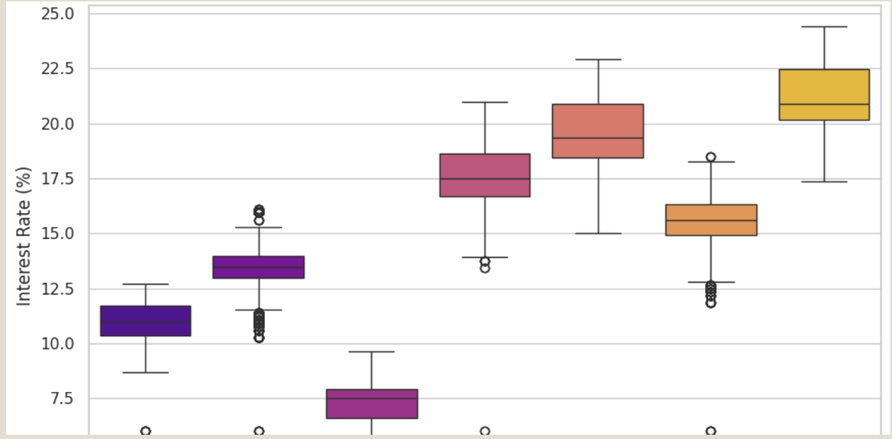
Visualizations:

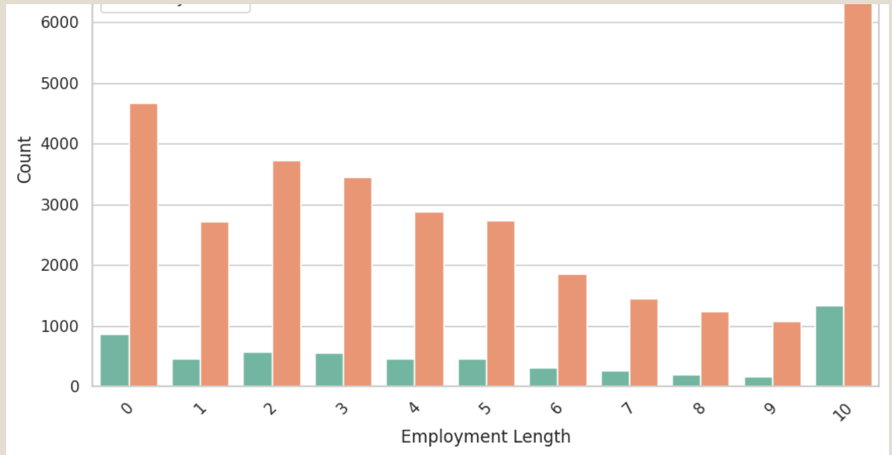
- Box plots: Interest_rate vs. loan_status, grade vs. loan_status
- Bar charts: emp_length vs. loan_status, grade vs. loan_status

Insights:

- Borrowers that have defaulted incur a high interest rate, where the charged-off loans at the 25th percentile are similar to the 50th percentile interest rates of fully paid borrowers.
- Grades E, F, D, and G have the greatest interest rates. Univariate analysis reveals that the defaulter rate is decreasing from B to G, with G having the lowest default rate. Conversely, Grades A, B, and C, which have lower interest rates, display the largest default counts.
- The 75th percentile for charged-off borrowers regarding loan amounts is greater compared to fully paid borrowers, indicating that lending cases should refrain from granting substantial loan amounts to charged-off individuals.
- Borrowers with longer employment history(>=10 years) and less than 1 year experience fall under higher default rate compared to the other employment length category. Hence, while approving loans to those borrowers, need to exercise more caution.







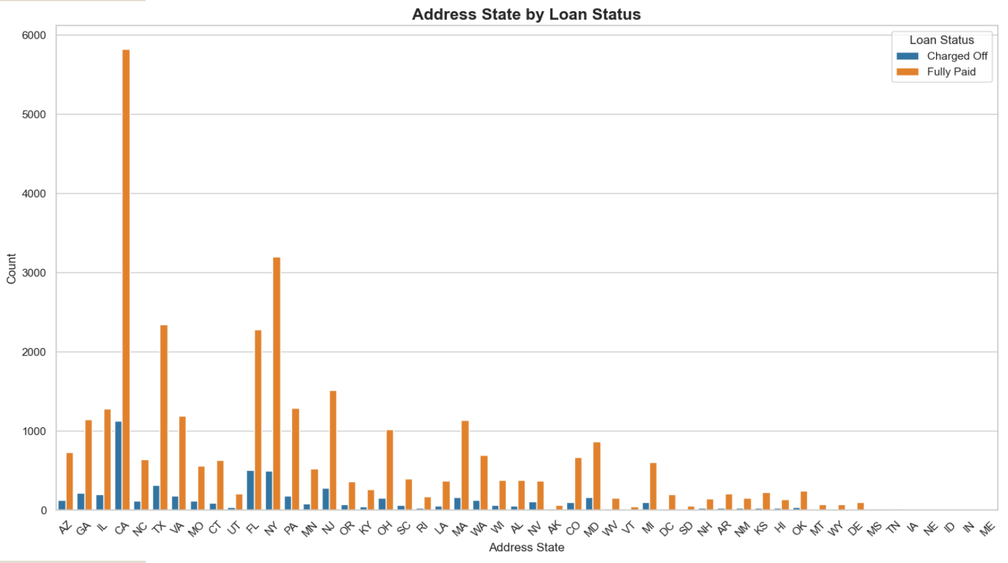
Multivariate Analysis

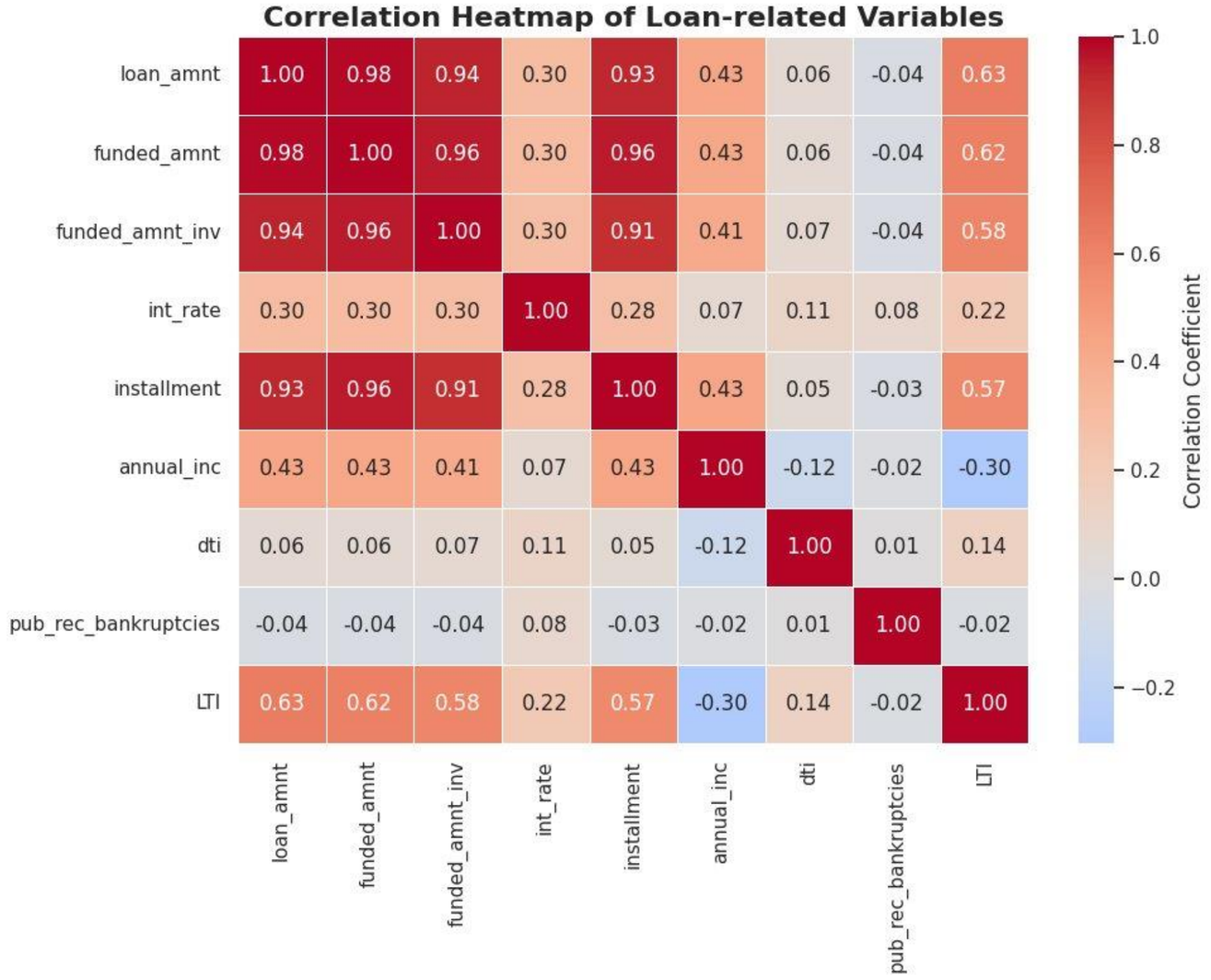
- Borrowers with a 60-month term possess a higher loan amount; yet, univariate research indicates that borrowers with a 36-month term exhibit a greater incidence of default. A larger loan amount correlates with a longer duration, hence mitigating the risk of default.

- As the transition from A to B occurs, the interest rate also escalates, and from the univariate analysis, we noted that Grades D to G have a lower incidence of defaults.

- Borrowers with mortgages possess higher loan amounts compared to other homeownership statuses, while borrowers classified as 'Renters' with smaller amounts exhibit a greater likelihood of defaulting.

- The three states with the highest default rates are California (16%), Florida (18%), and New York (13%). Lending Club should exercise increased caution and conduct additional checks when providing loans to these borrowers.





Correlation Heatmap

- Loan Amount (loan_amnt) and Installment (installment) [0.93]:
  - A strong correlation indicates that higher loan amounts are directly linked to higher installment values. For defaulters, this suggests that larger loans with high installment burdens may contribute significantly to the default risk.

- Loan Amount (loan_amnt) and Loan-to-Income Ratio (LTI) [0.65]:
  - This shows that larger loans tend to have a higher loan-to-income ratio, reflecting a higher financial burden relative to the borrower's income. High LTI is a potential indicator of financial stress and default risk
- Installment (installment) and Loan-to-Income Ratio (LTI) [0.58]:
  - Larger installments also correlate with a higher loan-to-income ratio, further supporting the idea that financial overcommitment could be a significant driver of defaults.
- Loan Amount (loan_amnt) and Annual Income (annual_inc) [0.48]:
  - Borrowers with higher incomes tend to take larger loans. However, higher income alone might not reduce default risk, as these individuals may still struggle if they over-leverage their finances.
- Loan-to-Income Ratio (LTI) and Annual Income (annual_inc) [-0.25]:
  - Higher-income borrowers tend to have relatively lower loan-to-income ratios. This reflects better financial capacity but may not always safeguard against defaults if loan sizes are too large.

Key Business Implications for Lending Club:

1. **Risk Mitigation**: - Lending Club should apply stricter vetting processes for loans with higher default risk characteristics (e.g., debt consolidation, credit card, small business). Introduce cautious loan approval for high DTI or LTI borrowers and loans within the $5,000-$16,000 range.

2. **Portfolio Diversification**: - Promote loans for low-risk purposes such as education and renewable energy to reduce default risks and increase portfolio balance.

3. **Tailored Loan Products**: - Create custom loan offerings for high-risk groups, such as those with shorter employment histories or higher DTIs, to mitigate risk without excluding potential borrowers.

4. **Geographic Focus**: - Conduct additional checks in high-default states (CA, FL, NY) and establish stricter lending criteria.

5. **Interest Rate Strategy**: - Analyze interest rate bands to adjust lending terms for better alignment with repayment capabilities, especially for Grades B and C.

Conclusion