LENDING CLUB CASE STUDY

SUBMITTED BY –

1. SUMANTH KANDREGULA
2. VARUN LAROIYA

Problem Statement

Enhancing Loan Marketplace Profitability through Risk Analytics

Financial Challenge:

- Primary financial risk stems from lending to 'risky' applicants, resulting in substantial credit loss.
- 'Credit loss' is the financial setback incurred when borrowers labeled as 'charged-off' default on loans.

Objective and Approach:

- Aim to identify high-risk loan applicants using Exploratory Data Analysis (EDA).
- Strategic identification of 'charged-off' customers to minimize credit loss.

Analysis Approach Overview:

1) **Dataset Overview:**
   - Original dataset comprises financial and demographic features of 39,717 bank customers.
   - A total of 111 features were initially considered for analysis.
2) **Feature Selection:**
   - After preemptive analysis, 53 features were selected based on the feature fill rate.
   - Records with null values were dropped, resulting in 36,430 entries.
   - Null values removal was necessary due to significant nulls in categorical features, impacting distinct values.
3) **Feature Enhancement:**
   - Derived new features such as Year and Month from date-type features.
   - Addressed data redundancies for improved dataset coherence
   - Derived a new feature to get total funded amount by combining funded_amnt and funded_amnt_inv
4) **Outliers Handling:**
   - Employed outlier detection and replaced outlier values with upper and lower whisker values.
5) **Data Segregation:**
   - Segregated features into continuous and categorical based on data types.
6) **Univariate Analysis:**
   - Conducted univariate analysis for both continuous and categorical variables.
7) **Bivariate Analysis:**
   - Utilized boxplots for bi-variate analysis, plotting various features against 'loan_status' with a focus on gaining insights.

Correlation Analysis:

Understanding Feature Relationships

- A correlation plot offers valuable insights into the relationships between different features

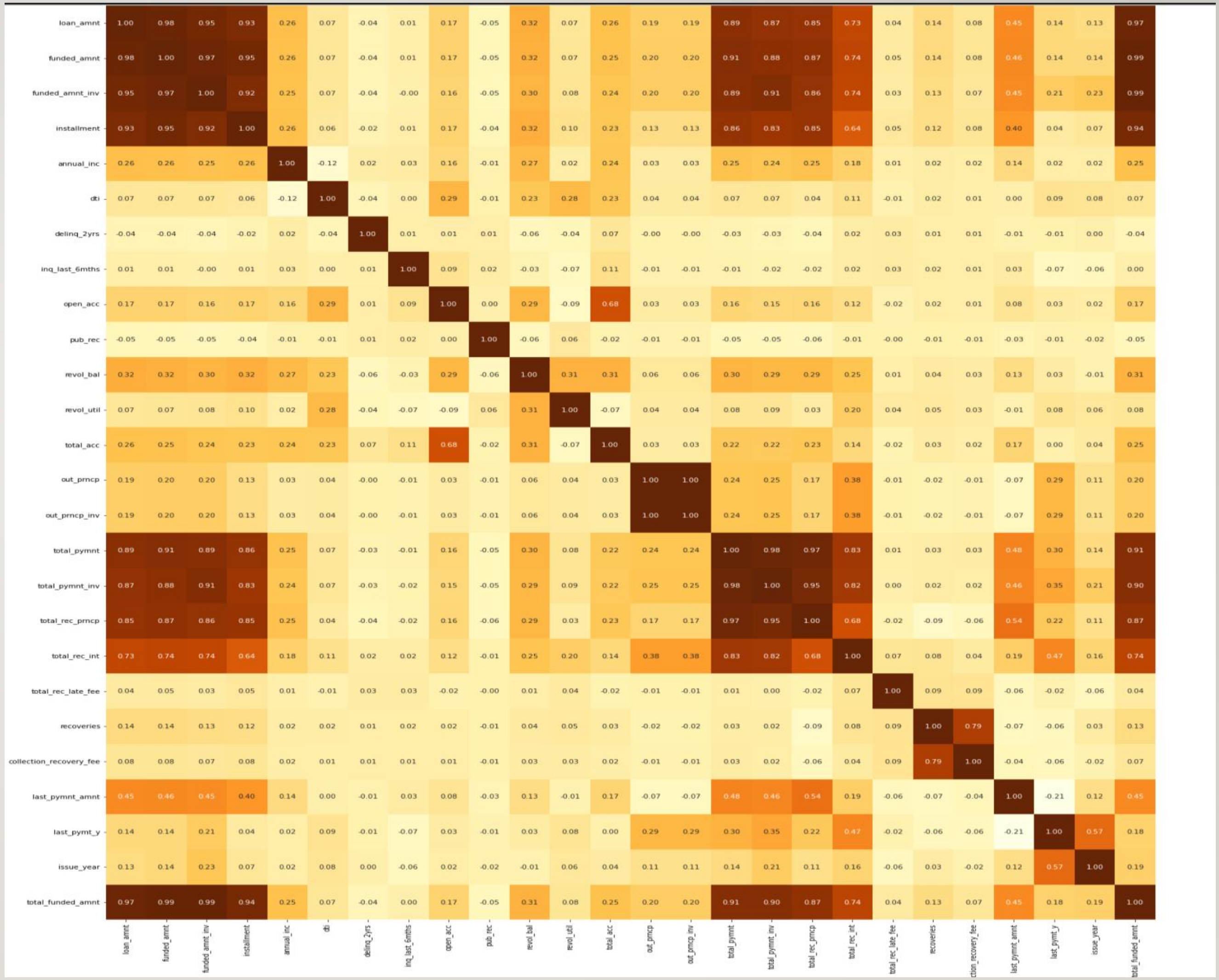
- Selecting features with correlation values exceeding 0.6, we proceeded to visualize these associations using scatterplots

- Detailed scatter plots illustrating these relationships are presented on the following slide

Key Points:

- Correlation values guide the identification of features with significant interdependence

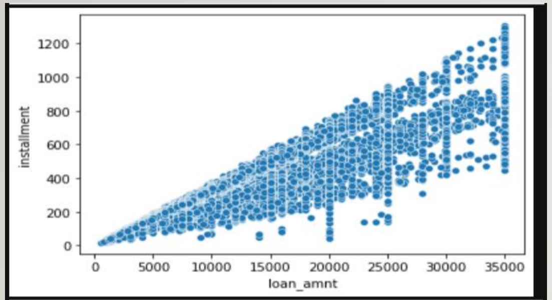
- Scatterplots serve as a visual representation of the strength and nature of these identified correlations

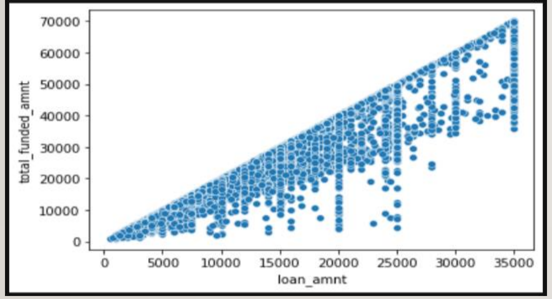


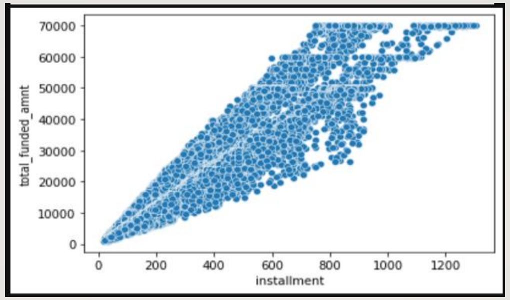
	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	total_acc	out_prncp	out_prncp_inv	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	ction_recovery_fee	last_pymnt_amnt	last_pymt_y	issue_year	total_funded_amnt
loan_amnt	1.00	0.98	0.95	0.93	0.26	0.07	-0.04	0.01	0.17	-0.05	0.32	0.07	0.26	0.19	0.19	0.89	0.87	0.85	0.73	0.04	0.14	0.08	0.45	0.14	0.13	0.97
funded_amnt	0.98	1.00	0.97	0.95	0.26	0.07	-0.04	0.01	0.17	-0.05	0.32	0.07	0.25	0.20	0.20	0.91	0.88	0.87	0.74	0.05	0.14	0.08	0.46	0.14	0.14	0.99
funded_amnt_inv	0.95	0.97	1.00	0.92	0.25	0.07	-0.04	-0.00	0.16	-0.05	0.30	0.08	0.24	0.20	0.20	0.89	0.91	0.86	0.74	0.03	0.13	0.07	0.45	0.21	0.23	0.99
installment	0.93	0.95	0.92	1.00	0.26	0.06	-0.02	0.01	0.17	-0.04	0.32	0.10	0.23	0.13	0.13	0.86	0.83	0.85	0.64	0.05	0.12	0.08	0.40	0.04	0.07	0.94
annual_inc	0.26	0.26	0.25	0.26	1.00	-0.12	0.02	0.03	0.16	-0.01	0.27	0.02	0.24	0.03	0.03	0.25	0.24	0.25	0.18	0.01	0.02	0.02	0.14	0.02	0.02	0.25
dti	0.07	0.07	0.07	0.06	-0.12	1.00	-0.04	0.00	0.29	-0.01	0.23	0.28	0.23	0.04	0.04	0.07	0.07	0.04	0.11	-0.01	0.02	0.01	0.00	0.09	0.08	0.07
delinq_2yrs	-0.04	-0.04	-0.04	-0.02	0.02	-0.04	1.00	0.01	0.01	0.01	-0.06	-0.04	0.07	-0.00	-0.00	-0.03	-0.03	-0.04	0.02	0.03	0.01	0.01	-0.01	-0.01	0.00	-0.04
inq_last_6mths	0.01	0.01	-0.00	0.01	0.03	0.00	0.01	1.00	0.09	0.02	-0.03	-0.07	0.11	-0.01	-0.01	-0.01	-0.02	-0.02	0.02	0.03	0.02	0.01	0.03	-0.07	-0.06	0.00
open_acc	0.17	0.17	0.16	0.17	0.16	0.29	0.01	0.09	1.00	0.00	0.29	-0.09	0.68	0.03	0.03	0.16	0.15	0.16	0.12	-0.02	0.02	0.01	0.08	0.03	0.02	0.17
pub_rec	-0.05	-0.05	-0.05	-0.04	-0.01	-0.01	0.01	0.02	0.00	1.00	-0.06	0.06	-0.02	-0.01	-0.01	-0.05	-0.05	-0.06	-0.01	-0.00	-0.01	-0.01	-0.03	-0.01	-0.02	-0.05
revol_bal	0.32	0.32	0.30	0.32	0.27	0.23	-0.06	-0.03	0.29	-0.06	1.00	0.31	0.31	0.06	0.06	0.30	0.29	0.29	0.25	0.01	0.04	0.03	0.13	0.03	-0.01	0.31
revol_util	0.07	0.07	0.08	0.10	0.02	0.28	-0.04	-0.07	-0.09	0.06	0.31	1.00	-0.07	0.04	0.04	0.08	0.09	0.03	0.20	0.04	0.05	0.03	-0.01	0.08	0.06	0.08
total_acc	0.26	0.25	0.24	0.23	0.24	0.23	0.07	0.11	0.68	-0.02	0.31	-0.07	1.00	0.03	0.03	0.22	0.22	0.23	0.14	-0.02	0.03	0.02	0.17	0.00	0.04	0.25
out_prncp	0.19	0.20	0.20	0.13	0.03	0.04	-0.00	-0.01	0.03	-0.01	0.06	0.04	0.03	1.00	1.00	0.24	0.25	0.17	0.38	-0.01	-0.02	-0.01	-0.07	0.29	0.11	0.20
out_prncp_inv	0.19	0.20	0.20	0.13	0.03	0.04	-0.00	-0.01	0.03	-0.01	0.06	0.04	0.03	1.00	1.00	0.24	0.25	0.17	0.38	-0.01	-0.02	-0.01	-0.07	0.29	0.11	0.20
total_pymnt	0.89	0.91	0.89	0.86	0.25	0.07	-0.03	-0.01	0.16	-0.05	0.30	0.08	0.22	0.24	0.24	1.00	0.98	0.97	0.83	0.01	0.03	0.03	0.48	0.30	0.14	0.91
total_pymnt_inv	0.87	0.88	0.91	0.83	0.24	0.07	-0.03	-0.02	0.15	-0.05	0.29	0.09	0.22	0.25	0.25	0.98	1.00	0.95	0.82	0.00	0.02	0.02	0.46	0.35	0.21	0.90
total_rec_prncp	0.85	0.87	0.86	0.85	0.25	0.04	-0.04	-0.02	0.16	-0.06	0.29	0.03	0.23	0.17	0.17	0.97	0.95	1.00	0.68	-0.02	-0.09	-0.06	0.54	0.22	0.11	0.87
total_rec_int	0.73	0.74	0.74	0.64	0.18	0.11	0.02	0.02	0.12	-0.01	0.25	0.20	0.14	0.38	0.38	0.83	0.82	0.68	1.00	0.07	0.08	0.04	0.19	0.47	0.16	0.74
total_rec_late_fee	0.04	0.05	0.03	0.05	0.01	-0.01	0.03	0.03	-0.02	-0.00	0.01	0.04	-0.02	-0.01	-0.01	0.01	0.00	-0.02	0.07	1.00	0.09	0.09	-0.06	-0.02	-0.06	0.04
recoveries	0.14	0.14	0.13	0.12	0.02	0.02	0.01	0.02	0.02	-0.01	0.04	0.05	0.03	-0.02	-0.02	0.03	0.02	-0.09	0.08	0.09	1.00	0.79	-0.07	-0.06	0.03	0.13
collection_recovery_fee	0.08	0.08	0.07	0.08	0.02	0.01	0.01	0.01	0.01	-0.01	0.03	0.03	0.02	-0.01	-0.01	0.03	0.02	-0.06	0.04	0.09	0.79	1.00	-0.04	-0.06	-0.02	0.07
last_pymnt_amnt	0.45	0.46	0.45	0.40	0.14	0.00	-0.01	0.03	0.08	-0.03	0.13	-0.01	0.17	-0.07	-0.07	0.48	0.46	0.54	0.19	-0.06	-0.07	-0.04	1.00	-0.21	0.12	0.45
last_pymt_y	0.14	0.14	0.21	0.04	0.02	0.09	-0.01	-0.07	0.03	-0.01	0.03	0.08	0.00	0.29	0.29	0.30	0.35	0.22	0.47	-0.02	-0.06	-0.06	-0.21	1.00	0.57	0.18
issue_year	0.13	0.14	0.23	0.07	0.02	0.08	0.00	-0.06	0.02	-0.02	-0.01	0.06	0.04	0.11	0.11	0.14	0.21	0.11	0.16	-0.06	0.03	-0.02	0.12	0.57	1.00	0.19
total_funded_amnt	0.97	0.99	0.99	0.94	0.25	0.07	-0.04	0.00	0.17	-0.05	0.31	0.08	0.25	0.20	0.20	0.91	0.90	0.87	0.74	0.04	0.13	0.07	0.45	0.18	0.19	1.00

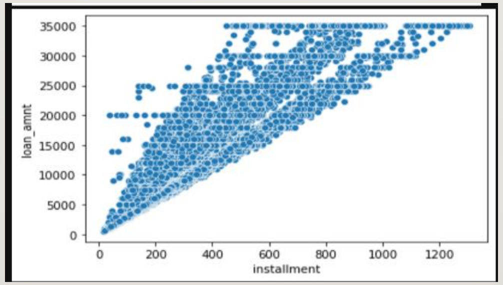
Numerical Variable Selection for Correlation Analysis:

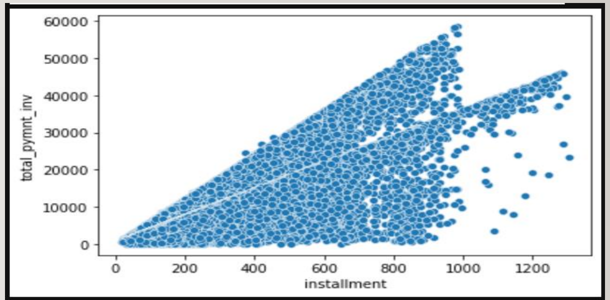
•Identified a subset of numerical variables, including loan_amnt, installment, total_funded_amnt, annual_inc, total_pymnt_inv, total_rec_prncp, total_rec_int, and total_rec_late_fee

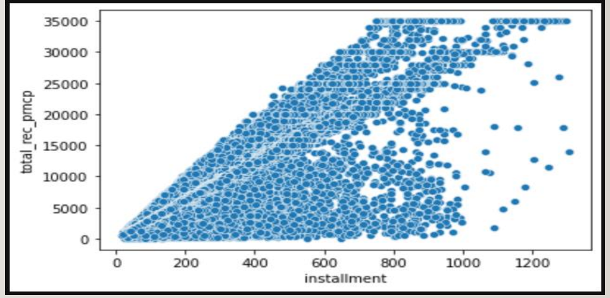










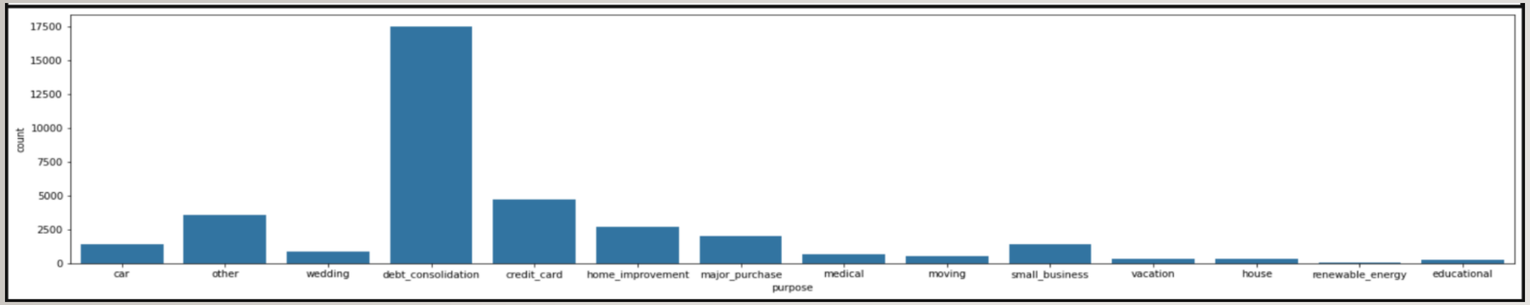


•Chose these variables based on their notable correlation values, indicating strong relationships among them

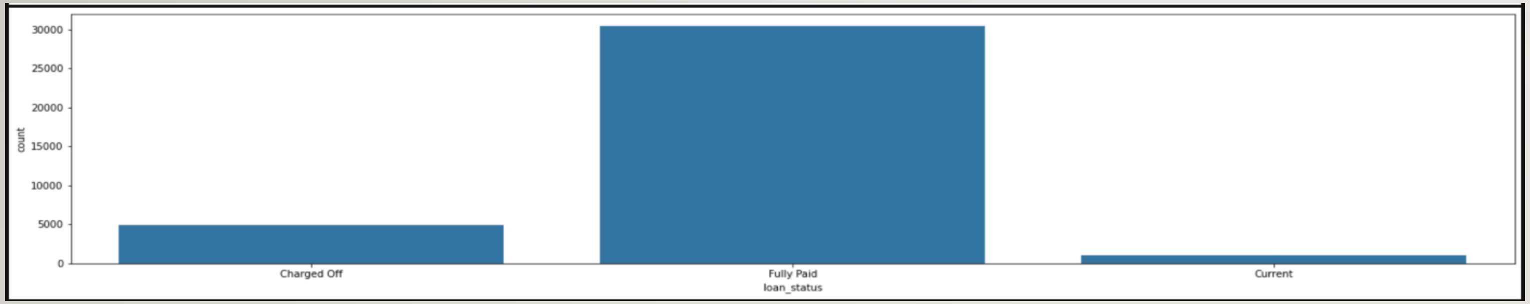
Highlighted Scatterplots:

•Presented specific scatterplots showcasing high correlations (greater than 0.5) among the selected features

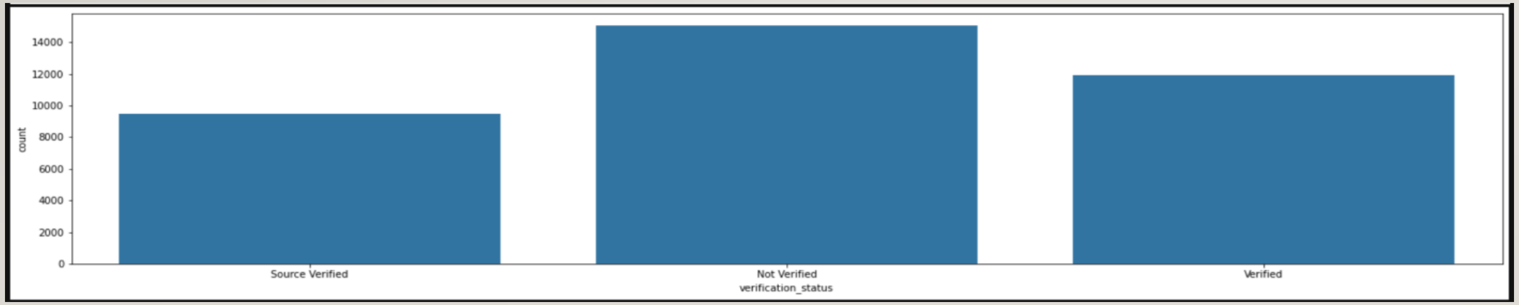
Univariate Analysis



Examining the "purpose" feature reveals that customers with the purpose of debt consolidation constitute almost 50% of the dataset

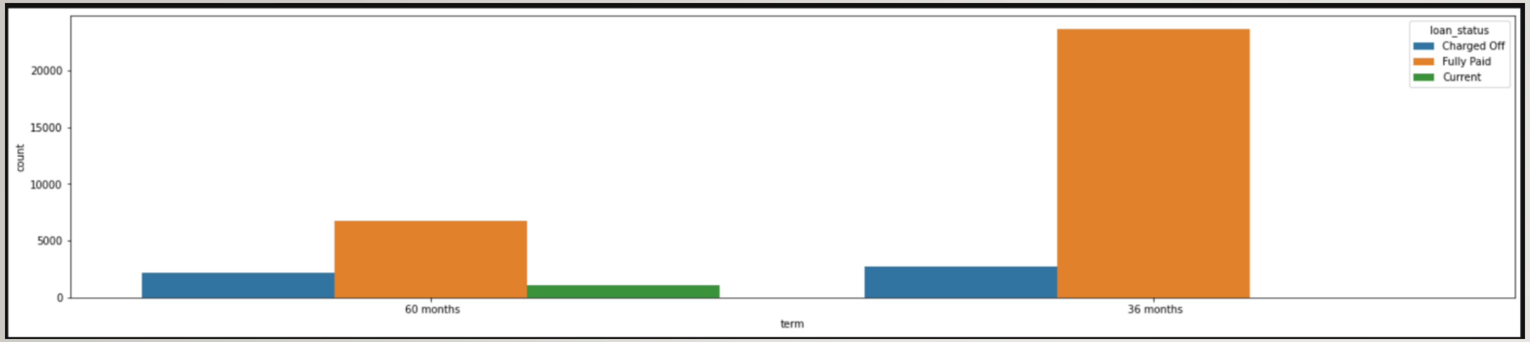


Upon reviewing the "loan_status" feature, it is observed that 80% of the loans have been fully paid. However, approximately 15% of the dataset is attributed to charged-off or defaulted loans, while the proportion of customers with active loans is notably low for the bank

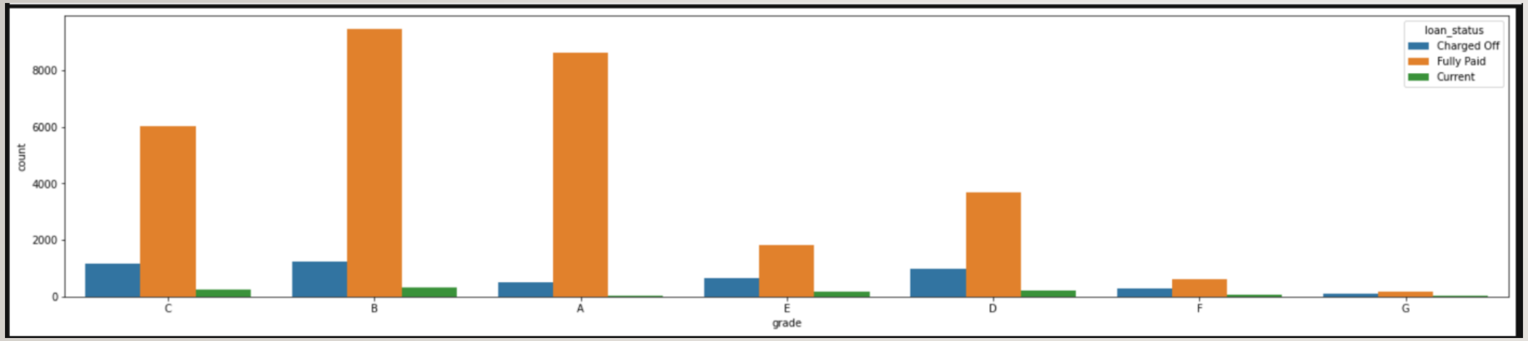


Examining the "verification_status" variable reveals that approximately 40% of customers in the dataset have income that is not verified by LC

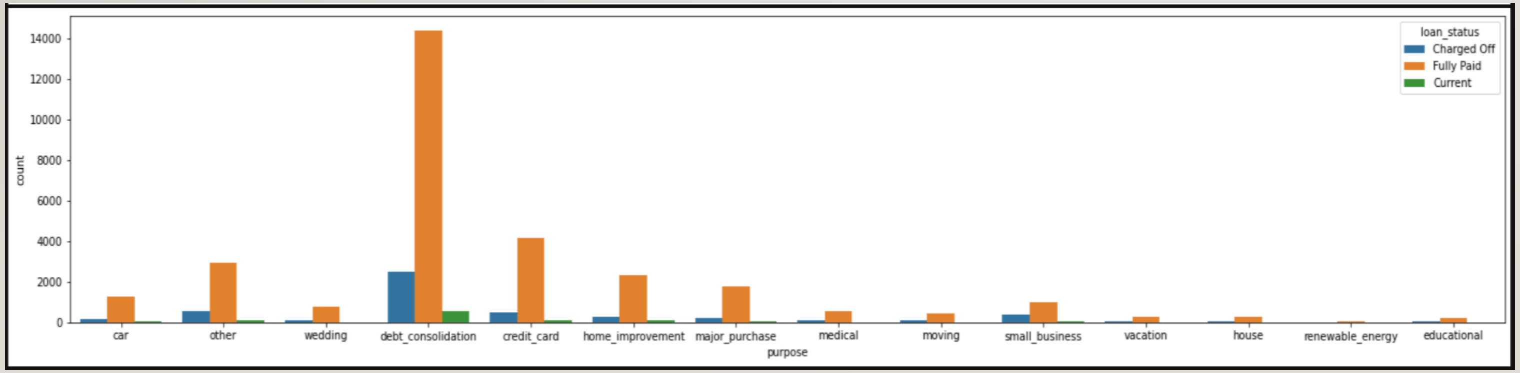
Bivariate Analysis



Upon analyzing the "term" feature, it becomes evident that the majority of customers have opted for a 36-month payment tenure. However, there are currently no customers with a 36-month payment tenure; all existing customers have chosen a 60-month term. This raises the need for the bank to investigate and address this discrepancy in customer preferences

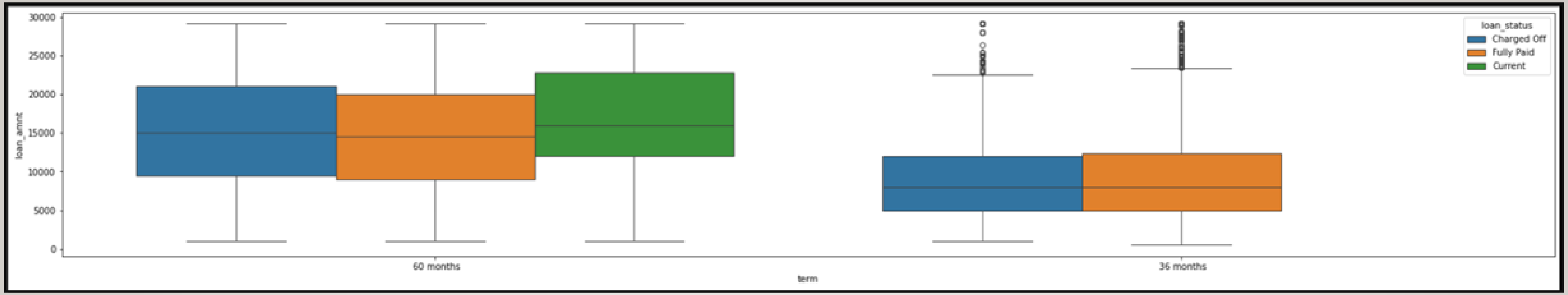


Upon examination of the "grade" feature, it is observed that the majority of customers are assigned loan grades B and A by LC

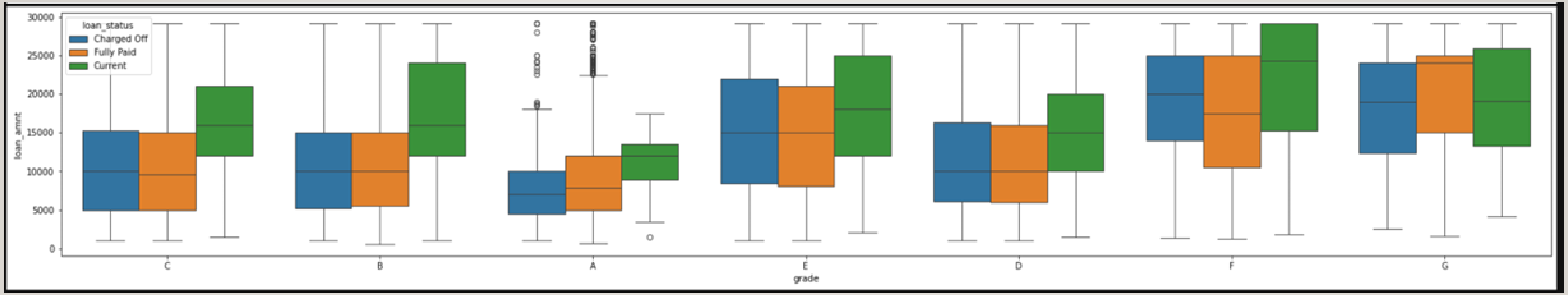


Upon examining the "purpose" feature, it is apparent that the majority of current loan customers are limited to three purposes: debt consolidation, credit card, and home improvement. This underscores the potential for the bank to expand its target audience by catering to customers with diverse purposes

Multi Variate Analysis

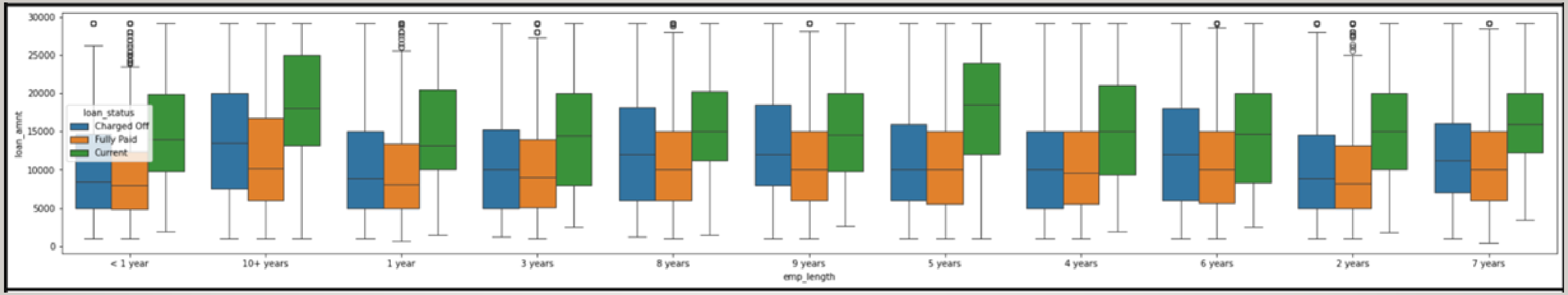


Upon examining the relationship between "loan_status" and "term" (payment tenure or number of payments), it is noteworthy that the median for a 60-month term is substantially higher, approximately $16,000, compared to a 36-month term with a median of around $7,500

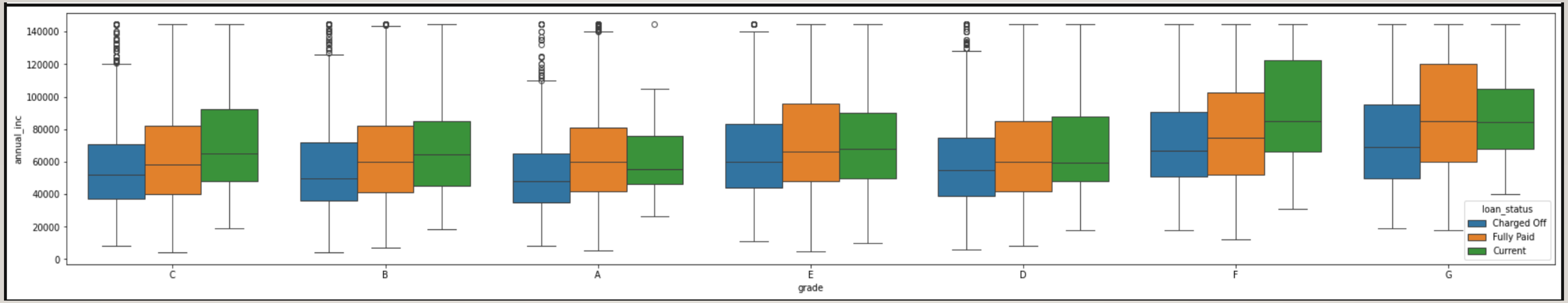


Upon scrutinizing the "grade" feature alongside loan amounts, it is apparent that current loan customers with an F grade have the highest loan amount. This trend persists even when examining defaulters, where the F grade is consistently associated with the maximum loan amount. This suggests that the bank should further investigate customers with an F grade

Multi Variate Analysis

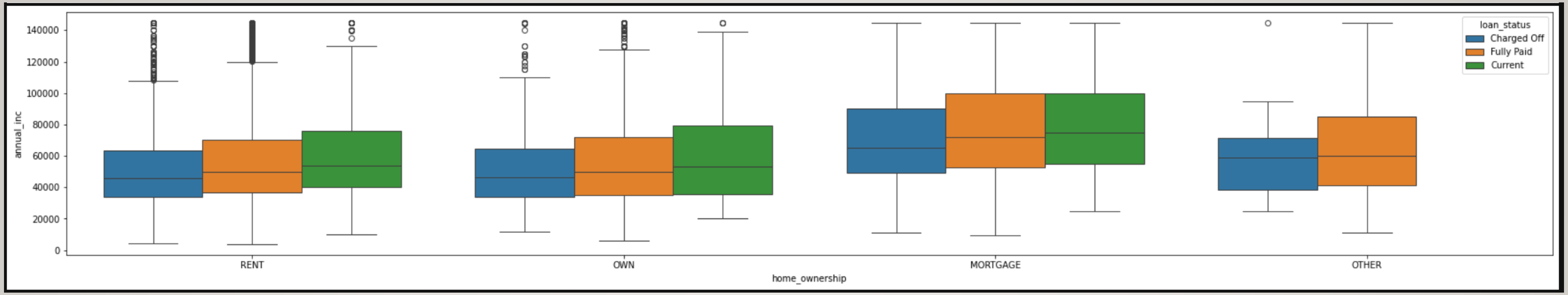


Upon examining the "emp_length" (employment experience) feature, it is observed that customers with current loans and 10+ years of experience are issued the highest loan amounts. Similar behavior is noted for customers with 5 years of experience. Additionally, customers with 10+ years of experience who are defaulters have taken the maximum amount of loans

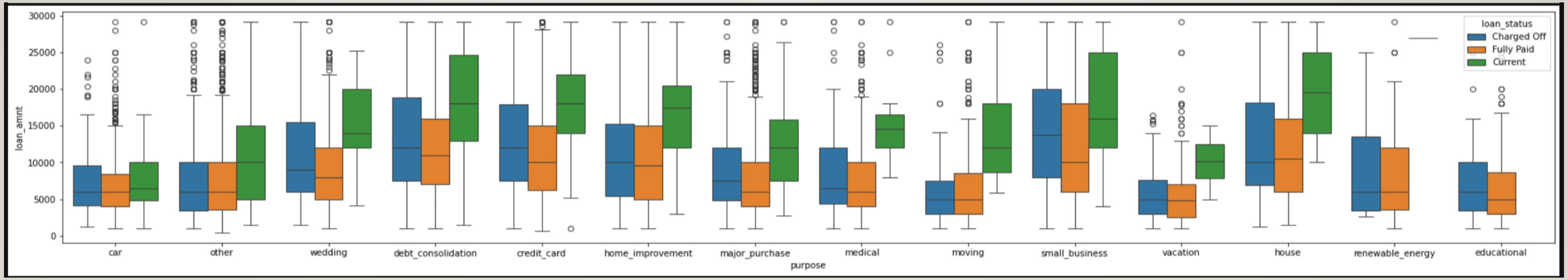


Upon analyzing the "grade" feature in conjunction with the annual income of customers, it is noted that current customers assigned grades F and G exhibit the highest median income compared to other grades. However, when considering defaulters, those in grade F and G demonstrate the highest median income among defaulters in comparison to other grades. This prompts the bank to investigate and address this observed pattern

Multi Variate Analysis



Upon examining the "home_ownership" feature in conjunction with the annual income of customers, it is evident that those with a mortgaged house exhibit the highest median annual income, both among current loan customers and defaulters. This pattern prompts the bank to investigate further and address any implications related to home ownership and income disparities



After analyzing the "purpose" feature in relation to loan amounts, it is evident that current loan customers with purposes such as debt consolidation, credit card, house, and small business have the highest median loan amounts. However, these same purposes exhibit high median loan amounts issued to defaulters as well. This prompts the bank to closely examine and address this observed issue

Important Driver Variables

Important Driver Variables	
Variable	**Variable Description**
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
annual_inc	The self-reported annual income provided by the borrower during registration.
purpose	A category provided by the borrower for the loan request.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.