Lending Club Case Study

Submitted By
Group Facilitator : Venkata Sairam Gajarampalli
Team Member : Rohit Kini

Batch:- ML67

Problem Statement

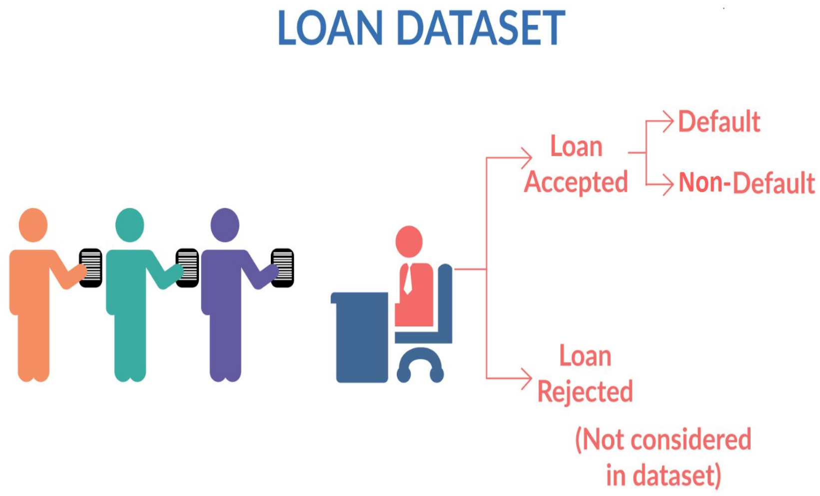
Business Understanding

You work for a **consumer finance company** which specializes 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

In this case study, you will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
- **Fully paid:** Applicant has fully paid the loan (the principal and the interest rate)
- **Current:** Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
- **Charged-off:** Applicant has not paid the installments in due time for a long period of time, i.e. he/she has defaulted on the loan
- **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)



Business Objectives

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

1. Data Loading and Basic Analysis

1. Import necessary packages.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# To ignore warnings
import warnings
warnings.filterwarnings('ignore')
pd.set_option("display.max_columns",300)
pd.set_option("display.max_rows",500)
```

2. Loads the loan dataset from a CSV file named "loan 2.csv."
3. Displays the first 5 rows of the dataset using the `head()` function.

Load the dataset

```
[4]: loan_data = pd.read_csv('loan 2.csv',low_memory=False)
```

Display the first 5 rows

```
[6]: print("First 5 rows of loan_data:")
loan_data.head()
```

First 5 rows of loan_data:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownershir
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	NaN	10+ years	RENT
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT

4. Prints the column names and their data types using the `info()` function.

Print the column names and their data types

```
print("\nColumn names and their data types for loan_data:")
loan_data.info(verbose=True)
```

```
Column names and their data types for loan_data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 111 columns):
 #    Column                          Dtype
---   ------                          -----
 0    id                              int64
 1    member_id                       int64
 2    loan_amnt                       int64
 3    funded_amnt                     int64
 4    funded_amnt_inv                 float64
 5    term                            object
 6    int_rate                        object
 7    installment                     float64
 8    grade                           object
 9    sub_grade                       object
 10   emp_title                       object
 11   emp_length                      object
 12   home_ownership                  object
 13   annual_inc                      float64
```

2. Data Cleaning

- Calculates and displays the percentage of null values in eacl
- Removes columns with more than 50% null values.
- Prints the shape (number of rows and columns) of the cleane
- Displays the first 5 rows of the cleaned dataset.
- Prints summary statistics of all columns in the cleaned datas
- Fills null values in specific columns with appropriate values (
- Displays the first 5 rows of the dataset after handling null val

Get the number of null values for each column ¶

+ 1 cell hidden

To find and display the percentage of the null values in all columns

```
missingvaluepercentage = 100*loan_data.isnull().sum()/len(loan_data)
missingvaluepercentage.sort_values(ascending=False)
```

```
verification_status_joint      100.000000
annual_inc_joint               100.000000
mo_sin_old_rev_tl_op           100.000000
mo_sin_old_il_acct             100.000000
bc_util                        100.000000
bc_open_to_buy                 100.000000
avg_cur_bal                    100.000000
```

Removing the columns whose percentage of null values is greater than 50 %

```
loan_data_clean=loan_data.loc[:,missingvaluepercentage < 50]
loan_data_clean.isnull().sum()
```

```
id                  0
member_id           0
loan_amnt           0
funded_amnt         0
funded_amnt_inv     0
term                0
int_rate            0
```

Lets check the data after removing columns

```
loan_data_clean.shape
```

```
(39717, 54)
```

```
loan_data_clean.head()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2

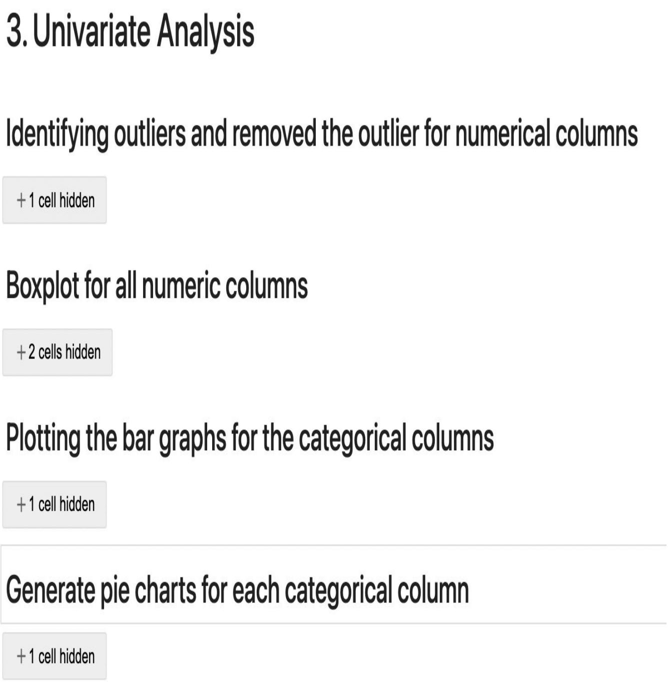
3. Univariate Analysis

Univariate analysis is an application of statistical analysis that is used to investigate and
single variable in a dataset. It is used in understanding the patterns, central tendencies,
regard to its relation with other variables.

Descriptive Statistics: This involves descriptive statistics which include mean, median,
representation of the general data.

Visualizations: This involves lots of visualizations such as histograms, box plots, densit

Purpose: To comprehend the central tendency, variability, and shape of distribution of or
any other unusual pattern; to better understand characteristics of the data.



Observations

1. Most of the loan applications are from CA state.

2. Most cases the purpose of loan is for devt_consolidation.

3. Most of the applications are having 10+ yrs of Exp.

4. Most of loan applicants are either living on Rent or on Mortgage.

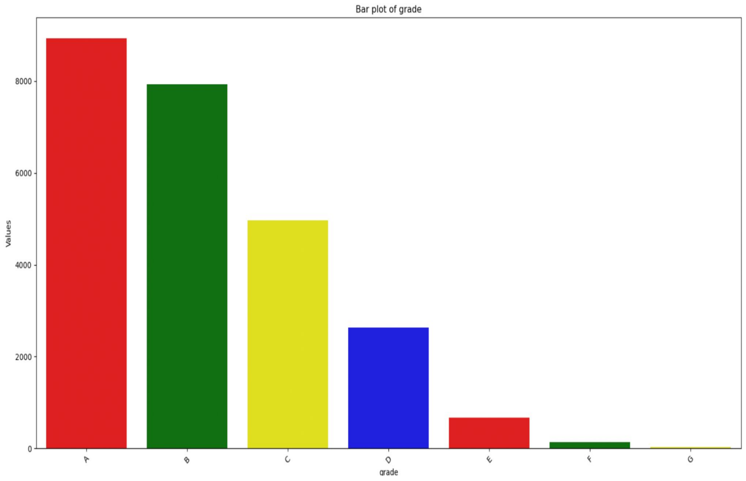
5. Majority of the customers are from grade B.

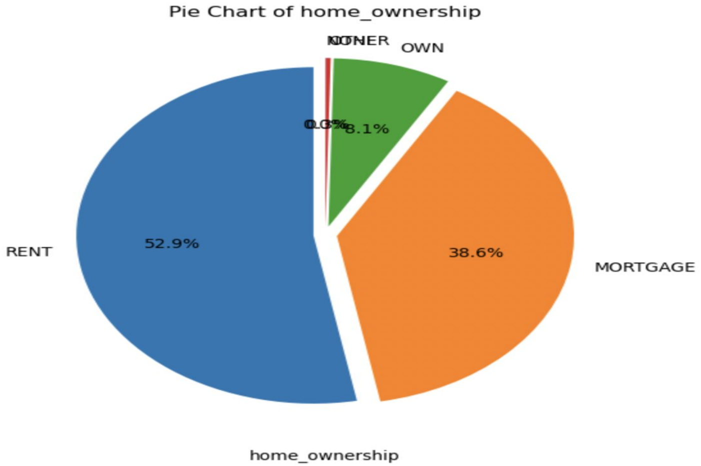
6. Avg interest rate falls near to 11.3%.

7. Majority of the customers are not verified.

8. Majority of the customers has paid the loan.

9. Avg annual income of the customer falls in between 40000 and 600





4. BiVariate Analysis

Bivariate analysis is a statistical method that explains the relationship between two variables. It shows us how changes in one variable can tend to be associated with changes in another. It attempts to establish some patterns or correlations or dependencies that are observed between pairs of variables beyond merely describing individual variables alone, which is known as univariate analysis.

Main attributes of bivariate analysis :

Two variables: It involved the study of two variables together, such as the relationship between the amount borrowed and the interest rate, as well as years on the job and loan status.

Exploring the relationship: It tries to identify and quantify the nature and strength of the relationship of both variables.

annual_inc vs loan_status (Numeric vs Categorical)

Observations

1. Avg annual income of the fully paid and charged off custom

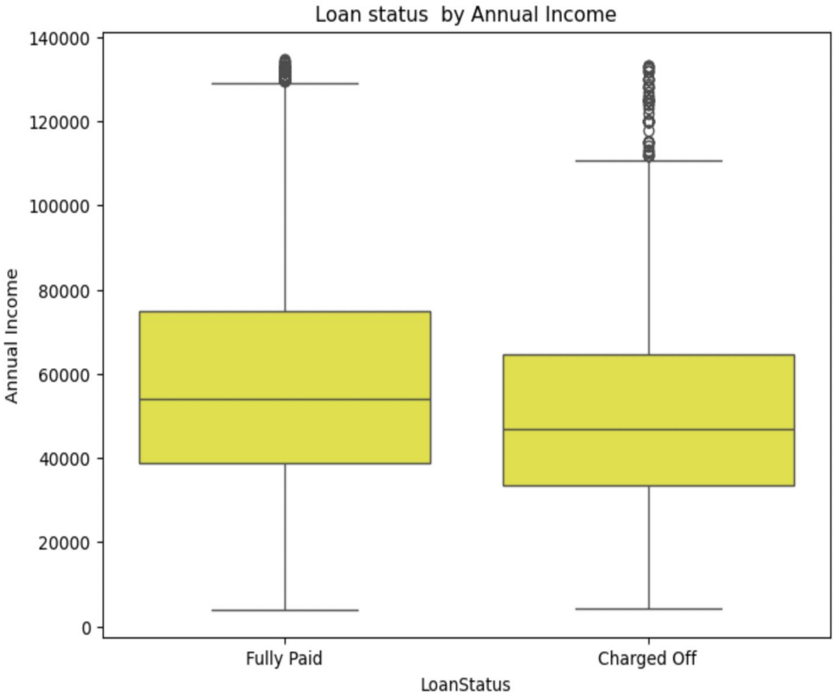
2. Annual income less than 40000 has high chance of charge

Total Distribution of Income:

Mean Median Income Similar: Both "Fully Paid" and "Charg
meaning that the average client in each group earns approxin

Income Outliers for Charged Off Loans: The "Charged Off"
whisker and has more income outliers at the higher end. Ther
than most others but still defaulted on their loans. This could s
a loan.

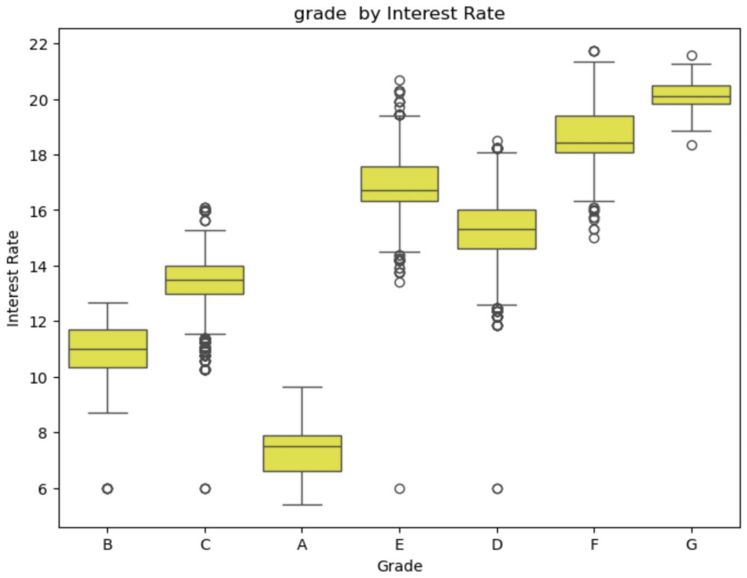
Similar Interquartile Range: The boxes for both groups indic
of each group of borrowers have a comparable spread of inco



Interest rate vs grade (Numeric vs Categorical)

Observations

1. Strong Positive Correlation: In fact, there is a very tight positive correlation between Grade and Interest Rate. Since Grade ranges from A to G (one might expect, hence, from most creditworthy to least), interest rate appears to be rising. There is intuitive appeal to this result, as it makes charge on higher interest rates for borrowers perceived
2. The lowest interest rates among all loans have the Grade A, which has th interquartile range (IQR) and less rate variability amongst it.
3. The median interest rate increases gradually from Grade A to G.
4. Grade G has the Highest Interest Rates: Grade G loans have the highest variability in rates for this group.
5. Outliers: There are some outliers in nearly all grades, particularly in the hi extremely high interest rates compared to others in the same grade.
6. Overlapping IQRs: Although the medians indicate a strong trend, there is therefore, implies that there might be some overlap on the interest rates b



Installment vs int rate (Numeric vs Numeric)

There's a clear positive correlation between installment and int_rate. This means that as the installment amount increases, the interest rate also tends to increase. This is somewhat expected, as higher inte

Observations:

1. Wide range of values: Both installment and int_rate exhibit a wide

2. Density: The plot is quite dense, indicating many data points. This

3. Potential outliers: There might be some outliers, especially for high investigating these further.

4. No clear clusters: There aren't any obvious distinct clusters in the interest rate is fairly continuous.

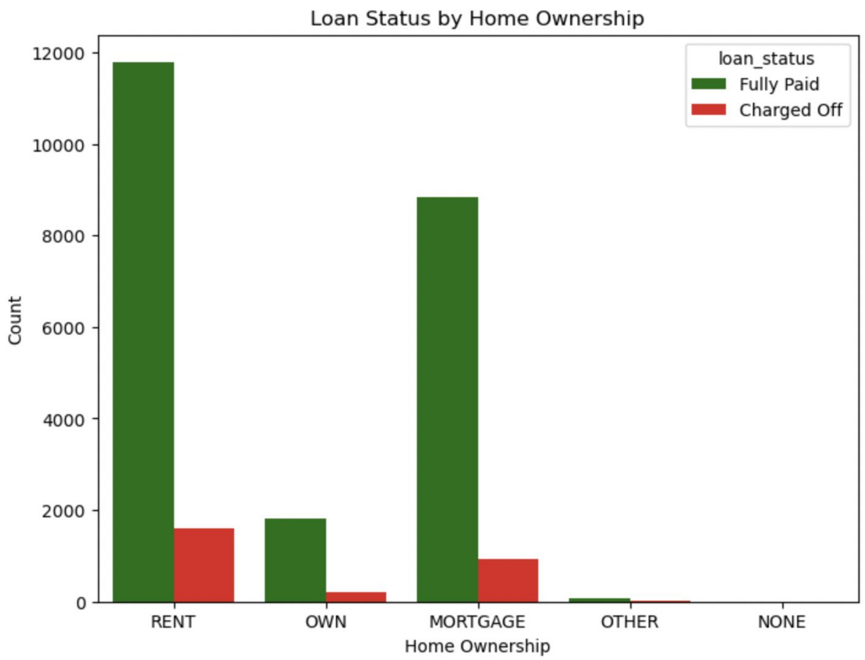
Home ownership vs loan status (Categorical vs Categorical)

Observations

1. Renters Have High Default Rates: This category, RENT,
("Charged Off") for RENT, however. I guess that means ren
mortgage holders.

2. Mortgage Holders Have the Highest Loan Counts and Be
is much higher than that of the red ("Charged Off") bars for
behavior of mortgage holders is better than those under oth

3. Fewer Loans for Other Categories : The other three cate
than MORTGAGE and RENT. It's difficult to make much of a
sizes are so small, but they look like they are all very differe



Purpose and the loan status **(Categorical vs Categorical)**

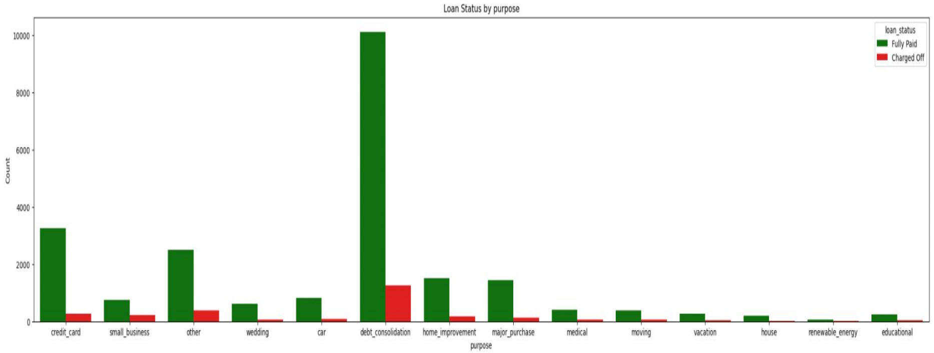
Observations

1. Consolidation Debts Takes the Top Slot: The loans pertaining to debt consolidation figure at the highest count, so it is likely one of the most pressing needs for availing loans.

2. Charge-offs Are Quite High in Debt Consolidation: The debt_consolid
sizeable amount of charged-off loans as well (red bars). This means th

3. Other Purposes Have Lower Volumes: All the other loan types, exce

4. Charge-Off Rates by Loan Purpose: credit_card is the following mos
charge-offs visible.home_improvement, major_purchase also have a re
categories such as renewable_energy, vacation, wedding, etc. have ve
considerable extent.



Verification status and the loan status **(Categorical vs Categorical)**

Observations

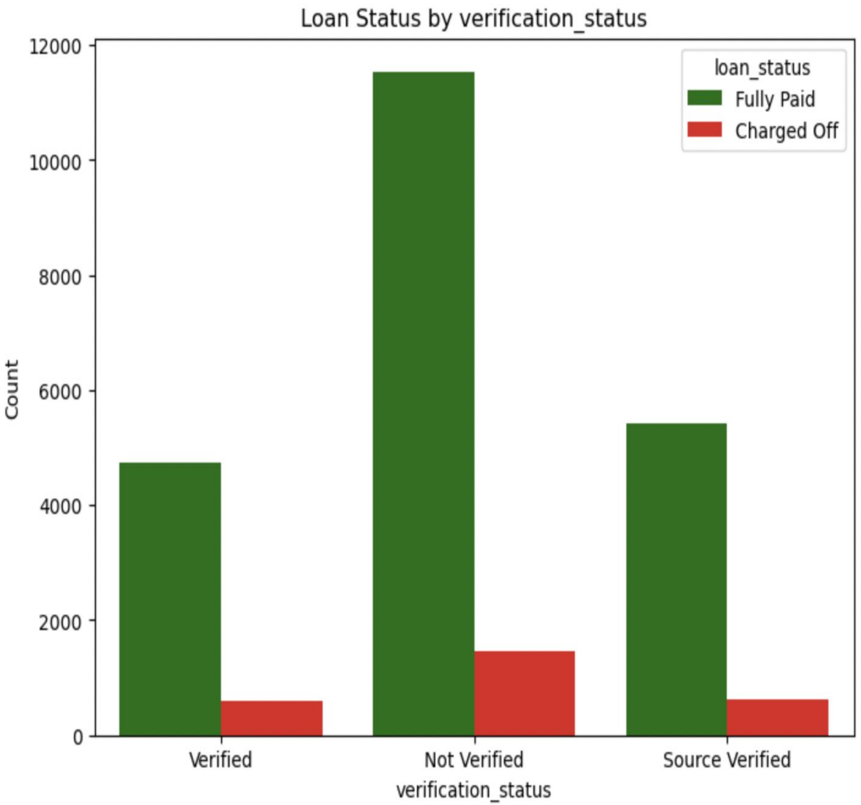
1. **Verification Status Impacts Loan Performance**

Loans Verified Work Best: The above graphically demonstrates that loans c
('Verified' status), do indeed have the highest percentage of 'Fully Paid' loans
(red bars).

Not Verified loans are the riskiest: Not verified status loans, where a lende
'Charged Off' loans. These loans bear much more risk of defaulting.

Source Verified falls somewhere in the middle: In loans whose informatior
the performance falls somewhere in between. A 'Source Verified' loan charge

2. Majority of Loans are Either Verified or Source Verified :This plot also s
Verified, and 'Not Verified' are more low frequency loans.



5. Multivariate Analysis

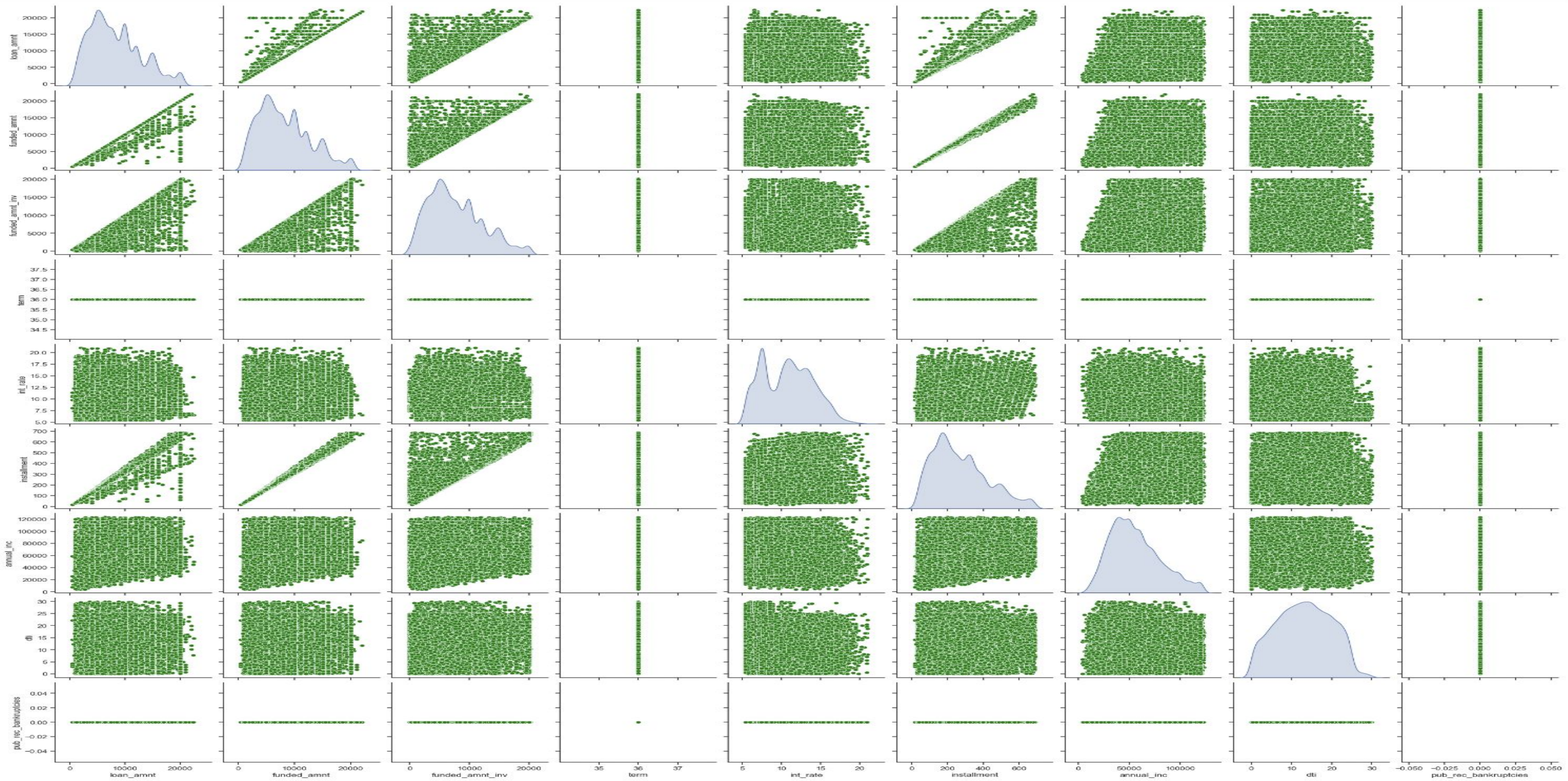
Multivariate analysis, being a statistical technique, explores and understands relationships among multiple variables simultaneously. It goes beyond univariate and bivariate analysis because it recognizes complex relationships and interdependencies between three or more variables. This helps find patterns, trends, and relationships that one may fail to notice when analyzing variables individually or in pairs.

It involves more than one variable, handling three or more; both independent and dependent variables are involved.

Complex relationship: It is developed to bring out the complex interaction and interdependence of various variables involved.

Statistical methods: It makes use of advanced statistical techniques such as multiple regression, factor analysis, cluster analysis, discriminant analysis, and principal component analysis.

Purpose It helps in understanding the combined effect of multiple variables on an outcome, identifying underlying patterns or structures within the data, and making predictions or classifications based on multiple variables.



loan_amnt
funded_amnt
funded_amnt_inv
term
int_rate
installment
annual_inc
dti
pub_rec_bankruptcies

Above pairplot shows the interactions of many numeric variables in your dataset. The plot is quite handy to take an overview of the distributions (histograms on the diagonal) and pairwise relations (scatterplots) between those variables.

Color code: Uniform green color for the points in the scatterplots. It may also decrease the readability of distinguishing differently colored groups or trends if any.

Specific Observations - Distribution (Diagonal)

loan_amnt: Almost like a right-skewed curve, showing more loans for lesser amount and fewer loans for higher amount.

funded_amnt: The distribution is almost the same as loan_amnt. Therefore, most of these loan requests would have been fully funded.

funded_amnt_inv: right-skewed also, which may represent the proportion of the investments in these loans.

int_rate: The spread of the distribution is large which might imply a variety of interest rates offered.

installment: Right-skewed, most installments are for small amount.

Annual Inc: Highly Right-skewed, meaning most of the borrowers' incomes is low and the distribution skew out at high values.

DTI: It is nearing a uniform distribution with possible concentration by lower value

Total Pymnt: Highly Right-skewed, means most loans have lower total payment.

Total Pymnt Inv: Distribution is nearly the same as total_pymnt

Specific Observations - Relationships (Scatterplots)

loan_amnt vs funded_amnt & funded_amnt_inv: Strong positive linear relationship, which, in general, loans are mostly funded either at full or very close to the requested amount also on the investor funding end.

loan_amnt/funded_amnt vs installment: Positive linear, as expected - the bigger the loans, the higher the installments.

loan_amnt/funded_amnt vs int_rate: This seems to be a very weakly positive relationship so that large loans correspond to high interest rates. The trend, however is not really discernible because of so many points.

int_rate vs installment: Probably positive and makes a high interest rate equivalent to a large installment.

annual_inc vs loan_amnt/funded_amnt: It is difficult to say if it is there or not, but looks kinda positive. Only due to skewness, I could not infer much from the densely packed scatterplot. High income earners might borrow higher sums of loans.

dti vs loan_amnt/funded_amnt: It appears very slightly negative, and sure enough that people having higher dti get smaller loans; yet that is kinda vague in this case.

Total pymnt/total pymnt inv vs loan amnt/funded amnt: Pos. Relation, this shows that the greater the loan is the bigger is total payments for both sides, lender and investors

6. Correlation

Observations

High Positive Correlations:

1.Loan Amount, Funded Amount, Funded Amount Investor: These three variables are highly indicates they move almost lockstep; with one increase, others tend to increase nearly by the sa money funded and invested is directly proportional to the original loan amount.

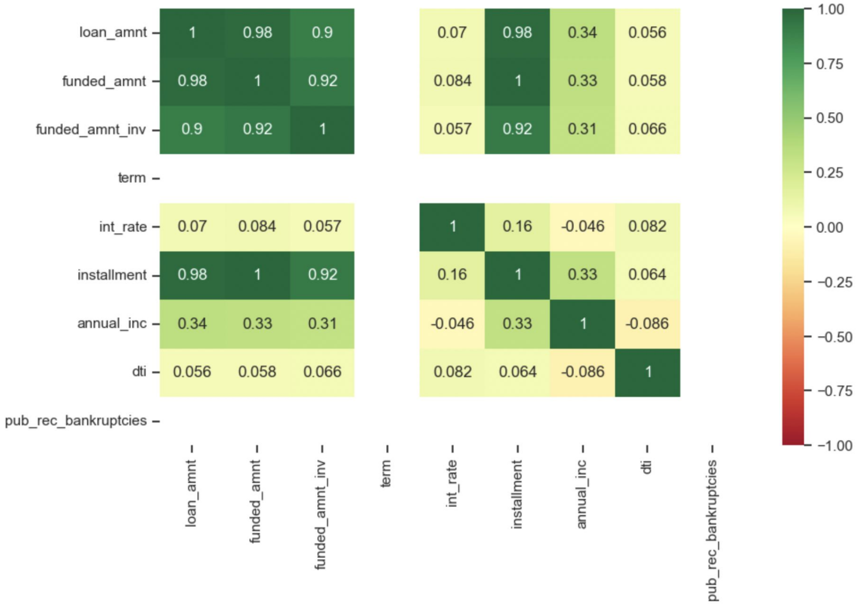
2.Installation and Loan/Funded Amounts: Installation also shares a high positive correlation w the amount borrowed, the higher will be its installment.

Moderate Positive Correlation:

1.Annual Income and Loan/Funded Amounts: There is also a moderate positive correlation b individuals earning higher incomes tend to borrow a larger amount.

Low Correlations:

1.Interest Rate and others: Interest rate is weakly correlated with most of the variables. This co



2.Debt-to-Income Ratio (DTI) and Annual Income: DTI and annual income have a weak negative correlation. This would imply that people with higher incomes possibly have lower DTI, thus possibly meaning that they have a better financial health.

7. Insights

Loan Purpose Matters: There are many debt consolidation loans but at greater danger of default. The other purposes of loans bring varying risks.

Verification is Everything: The verification of consumer's information reduces the risk of default as much as possible. Loans that are labeled 'Verified' are the best; those labeled 'Not Verified' have the highest charge-off rates.

Income Does Not Isolate: Income alone does not predict the performance of a loan. Rich-income consumers can still default, and so such factors as credit history and debt-to-income are more important to know.

Interest Rates Are in Line with the Credit Grade: Interest rates are positively correlated with the credit grade, and this tells about the risk-based pricing. High-risk consumers have poor grades and the interest rates are higher.

States Differ from Each Other with Respect to Volume and Defaults: States would vary with regards to the volume of loan as well as default rates. This may be due to regional economic condition and demographics.

Interdependence Between Features: The pairplot and the correlation heatmap reveal high interdependence among various attributes. Analysis relating to such dependency would be quite intricate.

With this knowledge and guidance, the company will be in a better position to make informed decisions regarding lending, reduce potential losses from bad credit, and improve overall performances in portfolios.

8. Recommendations:

1. Risk Assessment Models be made more accurate by incorporating loan purpose and verification status into risk assessment models.

2. **Purpose-Specific Underwriting Policies:** Develop underwriting policies that include only high-risk loan purposes, such as debt consolidation. Such policies are bound to have higher qualification cut-offs or interest rates.

3. **Verify First:** Make the borrowers realize the importance of verification and realise that complete and verifiable information will help them get lower interest rates.

4. **Track Local Trends:** Review state-by-state loan performance and economic conditions to spot local trends in the emergence of risks or opportunities.

5. **More in-depth Analysis:** Analyze loan attributes and default risk more deeply to explore complex relationships. This may involve advanced statistical techniques or machine learning models.

Thank You