EDA Lending Group Case Study

Welcome to our EDA lending case study. In this presentation, we will explore the importance of EDA lending and its impact on loan approval decisions.

by Nishanth A.V

Introduction

EDA lending, or Exploratory Data Analysis lending, is a crucial process in the financial industry. It involves analyzing patterns in customer attributes and loan data to make informed loan approval decisions.

- A consumer finance company faces the challenge of balancing the risk of lost business with the risk of financial losses due to loan defaults.
- EDA techniques can be used to uncover patterns and relationships between borrower attributes and loan performance, reducing the risk of non-repayment while still capturing profitable lending opportunities.
- The goal is to identify patterns in consumer attributes and loan attributes associated with loan defaults, understand the relationship between these attributes and loan defaults, and develop insights that can be used to predict whether a new loan applicant is likely to default..

Background

The lending industry is constantly evolving, facing numerous challenges such as increasing default rates and changing customer preferences. Understanding these challenges is essential for successful loan approval processes.

- Loan lending is a crucial aspect of the financial industry, enabling individuals and businesses to access credit.
- Loan defaults occur when borrowers fail to make scheduled payments, leading to financial losses for lenders.
- Borrower-related factors influencing loan defaults include credit history, debt-to-income ratio, and employment stability.
- Loan-related factors influencing loan defaults include loan amount, interest rate, and loan terms.
- EDA plays a vital role in loan lending by identifying patterns, trends, and correlations between borrower attributes, loan attributes, and loan defaults.
- Key objectives of EDA in loan lending are to identify patterns, understand relationships, and develop predictive insights for new loan applicants.

Problem Statement

A consumer finance company is looking for patterns in customer and loan attributes that are associated with loan defaults. By identifying these patterns, they can improve their loan approval decisions and mitigate risks.

Through comprehensive EDA, we aim to gain a deeper understanding of the following aspects:

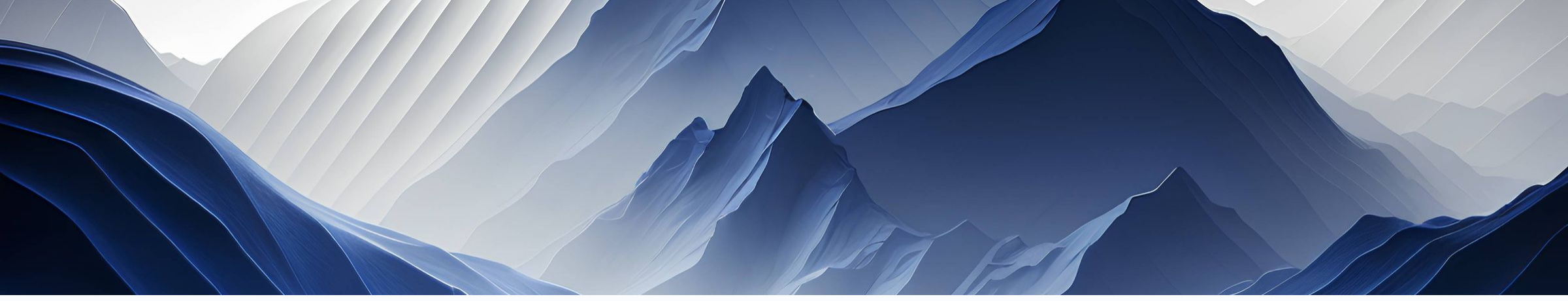
- Identifying Patterns in Consumer Attributes and Loan Attributes Associated with Loan Defaults: Uncovering the specific characteristics of borrowers and loan terms that are more closely linked to defaults.
- Understanding the Relationship Between Consumer Attributes and Loan Attributes with Loan Defaults: Exploring the interplay between borrower characteristics, such as credit history, income levels, and debt-to-income ratios, and loan attributes, such as loan amounts, interest rates, and repayment terms, in influencing default rates.
- Developing Insights that Can Be Used to Predict Whether a New Loan Applicant is Likely to Default: Utilizing the insights gained from EDA to develop predictive models that can assess the creditworthiness of new loan applicants, enabling more informed loan approval decisions.

Exploratory Data Analysis

Data analysis plays a crucial role in EDA lending. By leveraging data, financial institutions can gain valuable insights into customer behavior, loan performance, and market trends, enabling them to make informed decisions.The goal of this analysis is to develop insights that can be used to predict whether a new loan applicant is likely to default.

These are the following steps needed to perform Data Analysis

1. **Data Extraction**
2. **Data Cleaning**
3. **Exploratory Data Analysis (EDA):-**
    - **Univariate Analysis**
    - **Segmented Univariate Analysis**
    - **Bivariate Analysis**
    - **Derived Metrics**



Data Extraction & Cleaning

Loan lending data has been extracted and convertes to csv file. It contains the complete loan data for all loans issued through the time period 2007 to 2011.We have imported the Data to EDA on it.

Data Cleaning

1. **Identify and handle missing values in the data**:-
   - In the dataset, we encountered missing values, unwanted columns, and null values. We dropped unwanted columns, removed missing value rows, and imputed values to make the dataframe concise .
   - As we can see in the following images,The dataframe does not contain null or missing values.

```
Out[33]: id                          0.0
         member_id                   0.0
         loan_amnt                   0.0
         funded_amnt                 0.0
         funded_amnt_inv             0.0
         term                        0.0
         int_rate                    0.0
         installment                 0.0
         grade                       0.0
         sub_grade                   0.0
         emp_length                  0.0
         home_ownership              0.0
         annual_inc                  0.0
         verification_status         0.0
         issue_d                     0.0
         loan_status                 0.0
         purpose                     0.0
         zip_code                    0.0
         addr_state                  0.0
         dti                         0.0
         delinq_2yrs                 0.0
         earliest_cr_line            0.0
         inq_last_6mths              0.0
         open_acc                    0.0
         pub_rec                     0.0
         revol_bal                   0.0
         revol_util                  0.0
         total_acc                   0.0
         out_prncp                   0.0
         out_prncp_inv               0.0
         total_pymnt                 0.0
         total_pymnt_inv             0.0
```

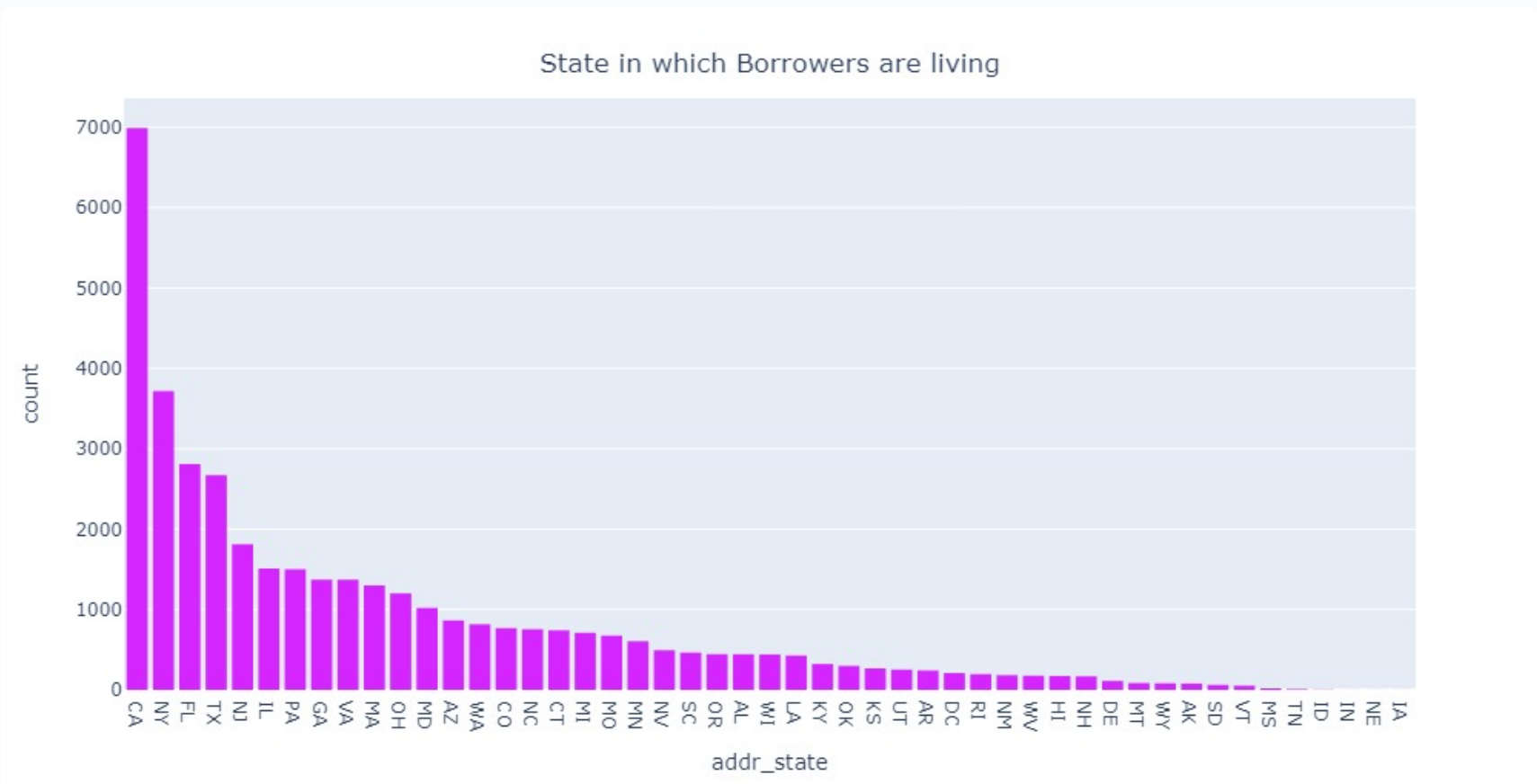
2. **Standardize and normalize numerical variables if necessary**:-
   - In the dataframe, For column employee length(work experiance)we observed Unstandardized value.Since work experiance value should be in integer,we removed unstandardized values and converted the column to integer
3. **Check for outliers and inconsistencies in the data**:-
   - In the Dataframe, we observed outliers in the annual income column. To identify these outliers, we plotted a bar plot and removed them from the dataset.
   - We also removed inconsistant column Description (desc) since they were added by the borrower and were not relevant to the analysis
   - In following set of images, we can see how the annual income column looks before(Left) and after(right) removing outliers

Exploratory Data Analysis (EDA)

1. **Univariate Analysis:-**It is a statistical technique that involves the examination of a single variable at a time. It is a useful tool for understanding the basic characteristics of a dataset and for identifying patterns and trends.

To apply univariate Analaysis for the given dataset,we have identified the following variables

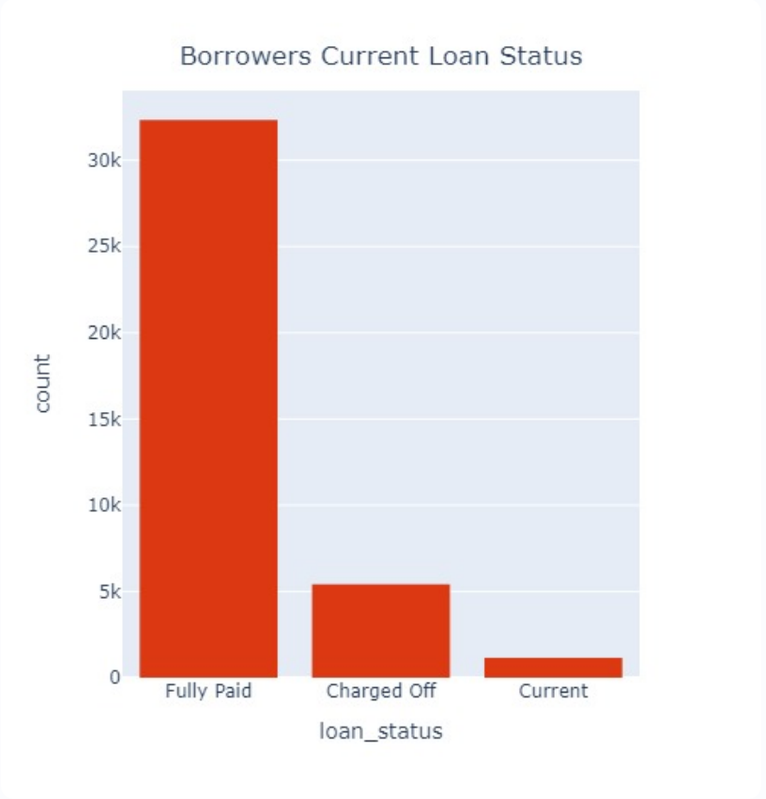
- Term, Grade, Sub-Grade, Home-Ownernship , Verification Status , Loan-Status , Address-State etc.
- We have used Data Visualization to plot graph for same variables.
- The Following are the some of the graph that we have plotted using single variable .

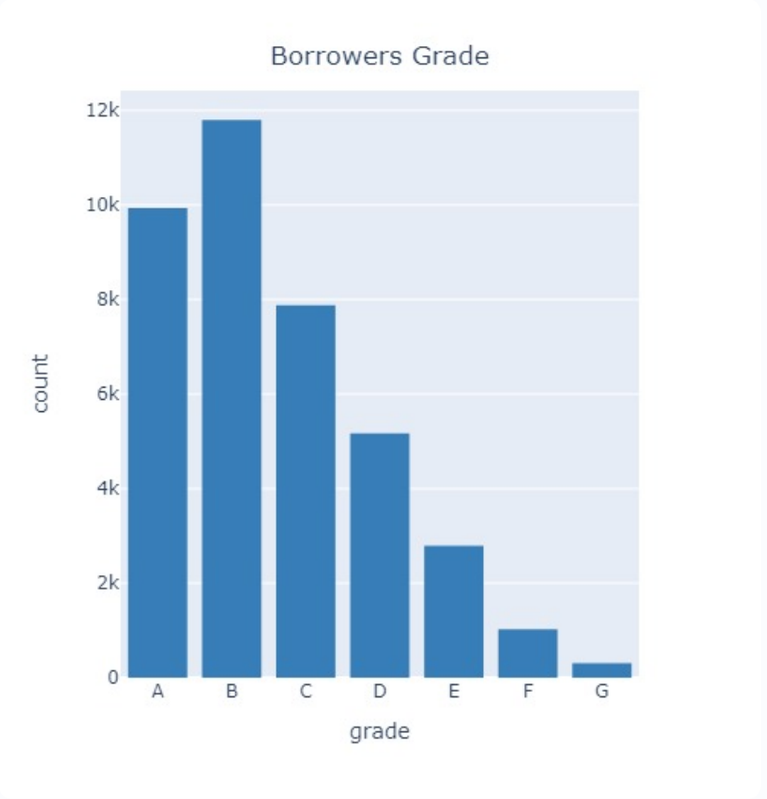


Address State Histogram

- The Histogram Shows the spread of the borrowers in the states.We can see the state CA has the Highest number of borrowers and NY has Second Highest number of borrowers etc.
- Using this plot, the bank can give loyalty/discount program to retain the number of borrowers in higher borrower states while they can give newer/updated programs for lower borrowers states to boost number of borrowers.

Univariate Analysis Graph





Loan Status Histogram

The graph shows borrowers current loan status .

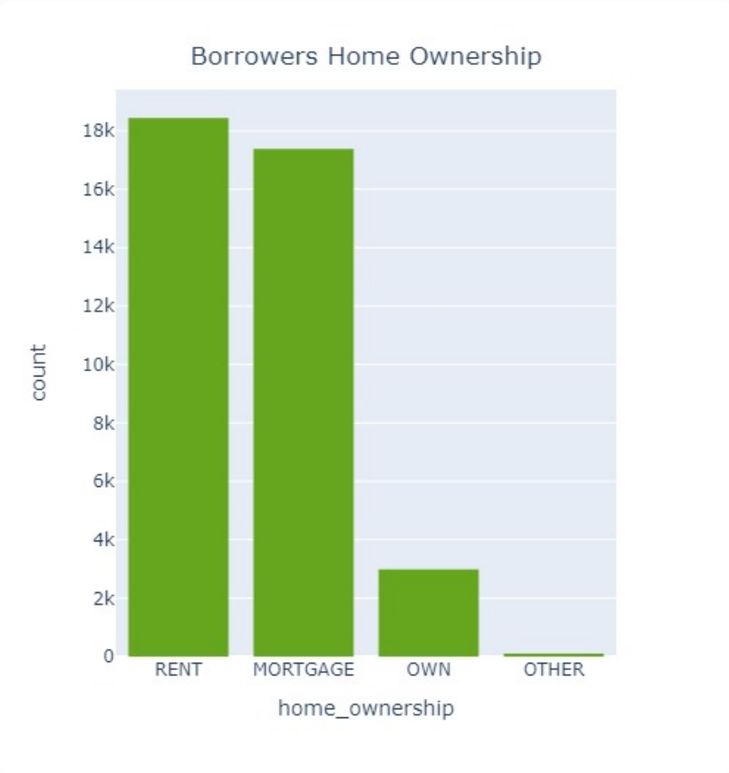
As we can see, very small number of borrowers are in processs of paying of the loan while sum percentages of borrowers has been charged off

Grade Histogram

. The graph shows the grade given to borrowers based on loan merit

As we can see grade B has highest number of borrowers and we can see the trend decreasing number of borrowers as the grade increases

Univariate Analysis Graph



Verification Status Histogram

The graph shows borrowers verification status.

From the fig we can observe huge amount of borrowers are not verified when compared to verified and source verified borrowers.

This can help the bank to provide loan to ideal borrower who have been verified

Home Ownership Histogram

The graph shows borrowers home ownership.

From the fig we can observe that borrowers who are associated with rent and mortgage homes are more likelyto take loan compared to others.

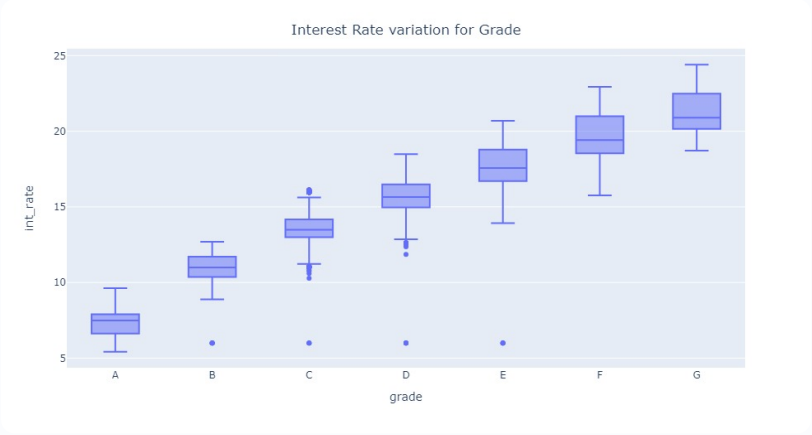
Segmented Univariate Analysis

Segmented univariate analysis is an extension of univariate analysis that involves dividing a dataset into segments based on a categorical variable and then performing univariate analysis on each segment. This allows you to compare and contrast the characteristics of a variable across different groups or categories

To apply univariate Analaysis for the given dataset,we have identified the following variables

- Interest_rate Grade, Term,Purpose, Home-Ownernship , Verification Status , Loan-Amount , Address-State etc.
- We have used Data Visualization to plot graph for same variables.
- The Following are the some of the graph that we have plotted.

Segmented Univariate Analysis Graph



Grade bar plot for interest rate

The plot shows borrowers interest rate range for the grade

As we can observe in graph ,the interest rate increases rapidly as grade for which borrowers have taken loan increases

Term bar plot for loan amount

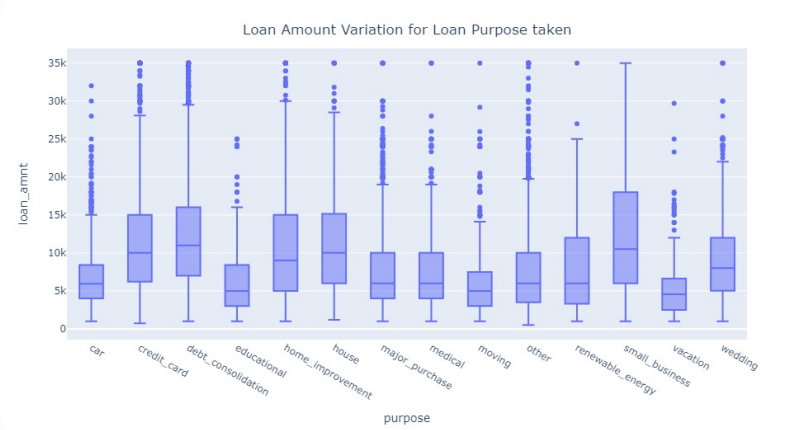
The plot shows borrowers   loan amount range for the selected term

As we can see in the graph,the loan amount has more variations for term 36 months than 60 months

Segmented Univariate Analysis Graph

Verification Status Bar plot for loan amount

The plot shows we can that the borrowers are verified have loan amount accepted for loan while source verified and not verified borrowers have less amount of loan



Purpose Bar plot for loan amount

The plot we can see that the borrowers have taken loan for less amount for educational purpose while they have taken more loan amount for debt consolidation and small business purpose

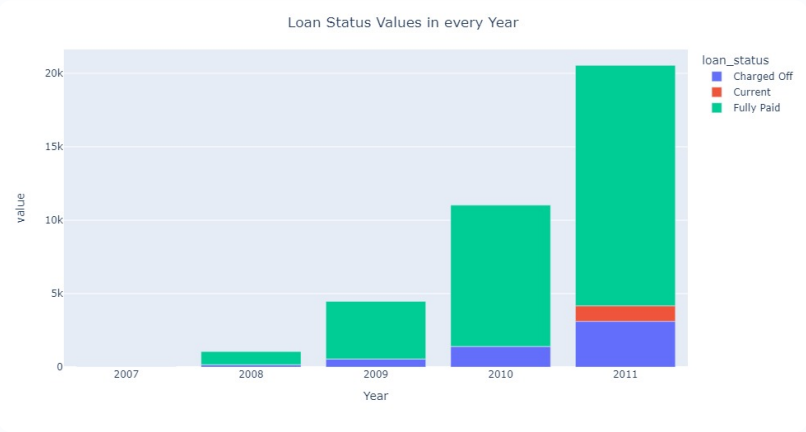
Bivariate Analysis

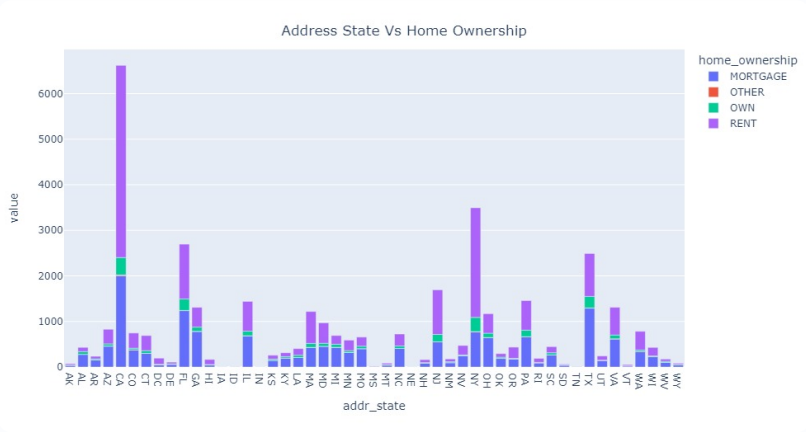
Bivariate analysis explores the relationship between two variables. It helps us understand how changes in one variable affect the other, revealing patterns, correlations, and potential associations.

To apply Bivariate Analaysis for the given dataset,we have identified the following variables:-

- Grade, Purpose, Home Ownernship , Verification Status, Loan Amount ,Loan Status,Funded Amount etc.
- We have used Data Visualization to plot graph for Two different variables.
- The Following are the some of the graph that we have plotted.

Bivariate Analysis Graph





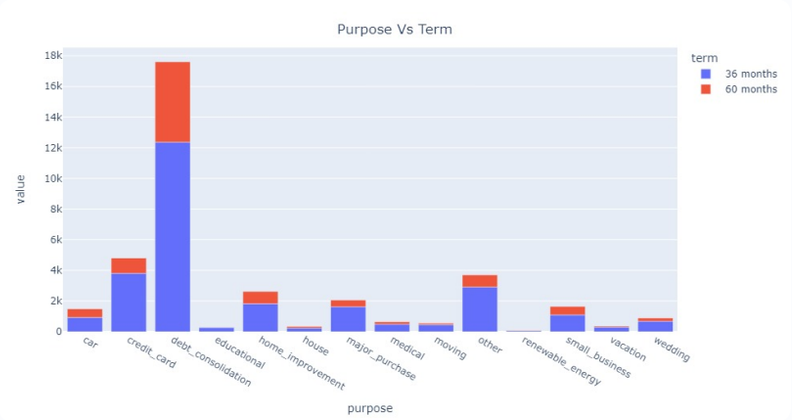
Bar graph of loan status for every year

- In the Bar graph we can see highest amount of the loans taken in year 2011.
- We can also see consistest number of loans being charged off in every year

Bar graph for home ownership for address state

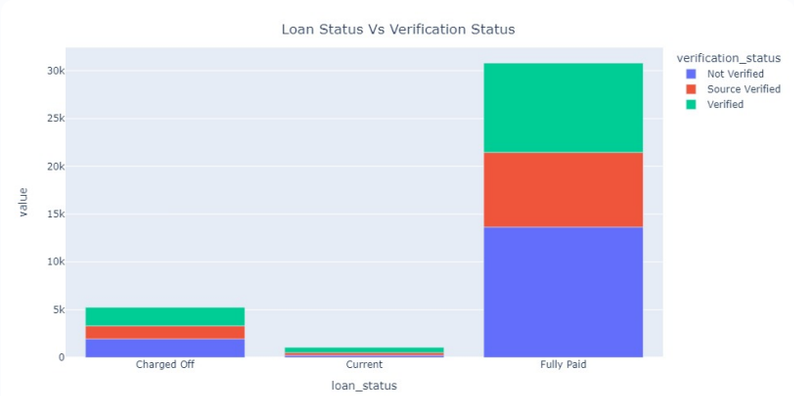
- In the bar graph we can see highest number of loans taken in CA state
- Most number of borrowers in CA state are in rented home
- Same analysis is true for NY state as well.

Bivariate Analysis Graph



Bar graph of purpose for loan term

- In the bar graph we can see most number of loans taken by the borrowers are for debt consolidation.And Most of them are for 36 months terms.
- While we saw debt consolidation and small business borrowers had similar loan amount spread in segmented univariate analysis graph we can see here that number of borrowers are disproportionate.



Bar graph of Loan Status for Verification Status

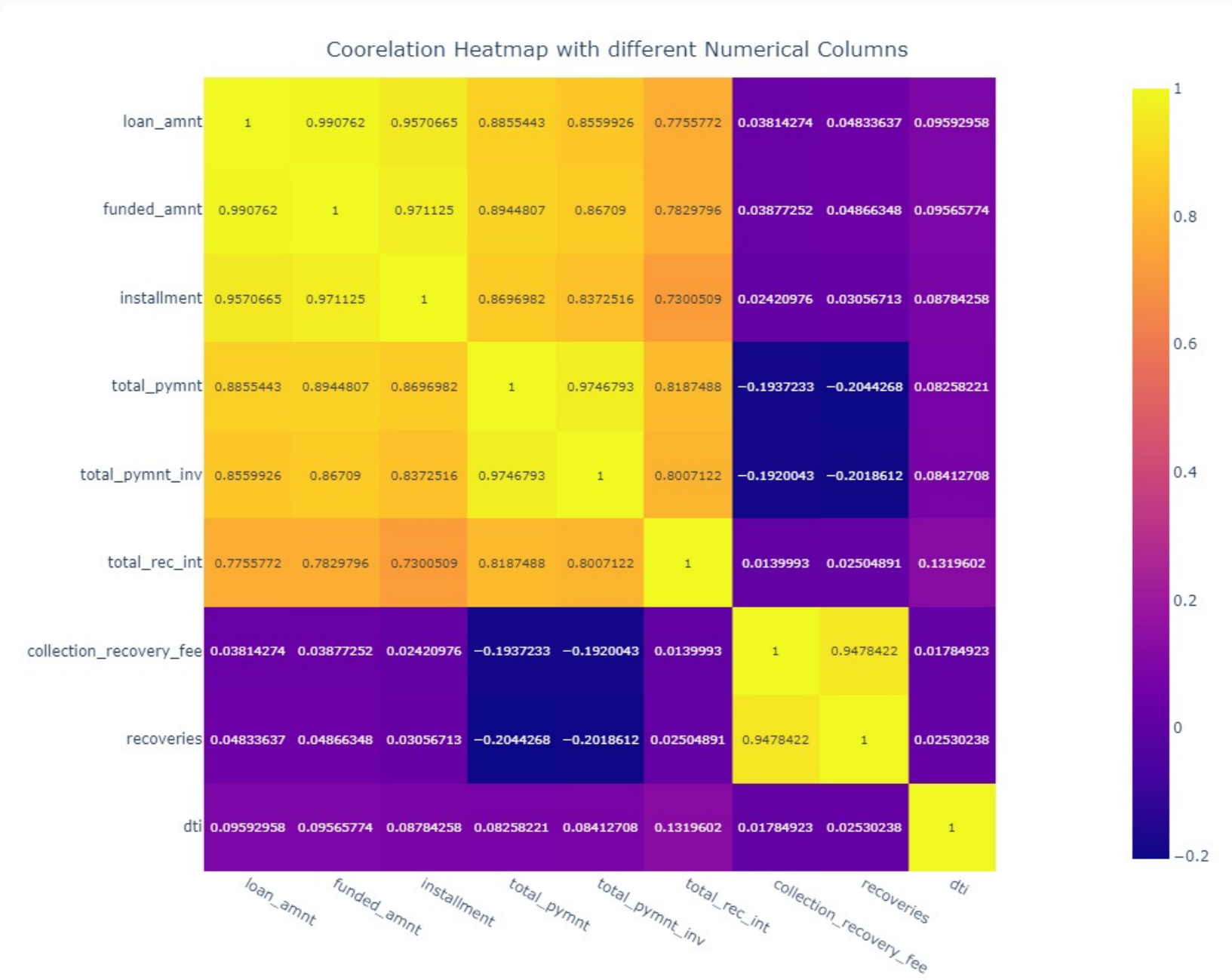
- In the graph we can see the borrowers have who been charged off belong to all 3 verification status group.As a result verification status has no relation for borrowers not paying back loan amount.
- We can also see the borrowers who have not been verified are the ones who have fully paid the loan amount more than the verified ones.

Bivariate Analysis Graph

The Following Heat map shows correlation between the different numerical columns in the dataset .

This shows how on columns value changes while another columns values changes.

- [0→1] value denotes the columns changes positively i.e. if one column value increases the other also increases.
- [0→-1]value denotes the columns changes negatively i.e if one column value increases the other also increases
- [0] value denotes the columns are not correlated



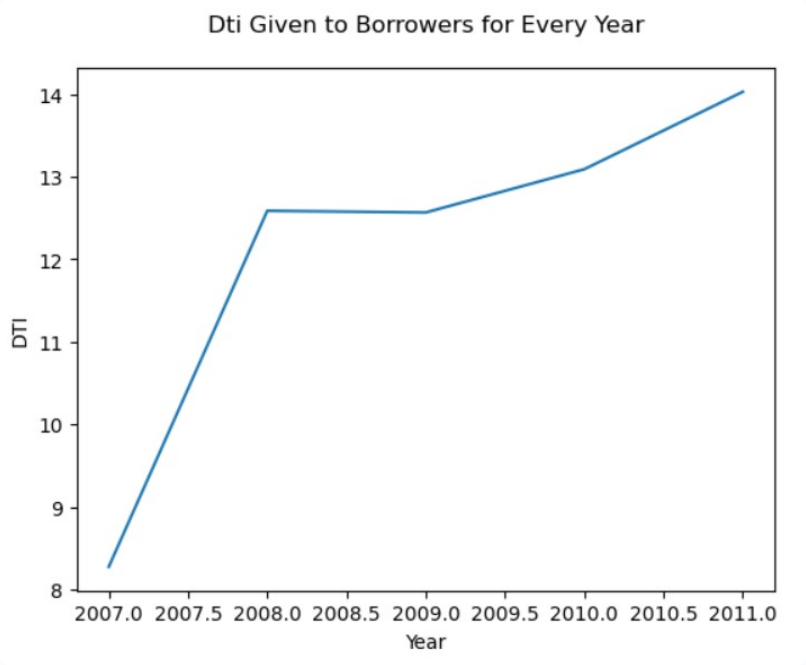
Derived Metrics

A derived metric is a new measurement created by combining or transforming existing variables in a dataset. Derived metrics are often used to gain a deeper understanding of the data and uncover hidden patterns or relationships that might not be apparent from the original variables.

To applyDerived metrics for the given dataset,we have identified the following variables:-

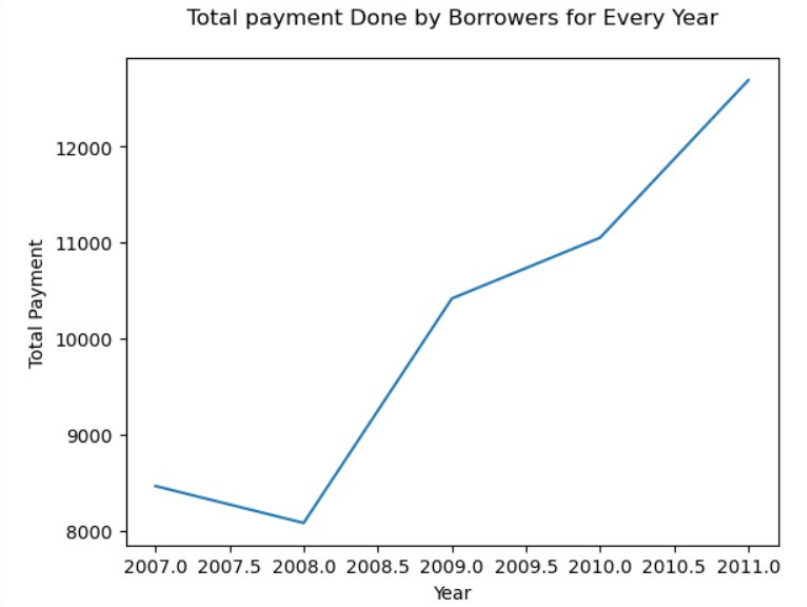
- We have extracted year and month from issue_d column and used the same for plot creation
- We have used Data Visualization to plot graph for Year and Month grouped for different variables .
- The Following are the some of the graph that we have plotted.

Derived Metrics graph



Line Graph of dti for every year

- In the graph we can the dti increases exponentially from 2007 to 2008 and then stays the same for 2009 and then starts to increase gradually till 2011.



Line Graph of Total Payment for every year

- In the graph we can see the total payment done decreases slowly from 2007-2008 and then increases exponentially.

Results

While EDA lending offers significant benefits, there are also challenges and limitations to consider. It requires thorough data collection, processing, and interpretation. Careful analysis is essential to avoid biased decision-making.

- Exploratory Data Analysis (EDA) plays a crucial role in identifying patterns and relationships between borrower attributes, loan attributes, and loan defaults.
- The goal of EDA in loan lending is to develop insights that can be used to predict whether a new loan applicant is likely to default.
- Key outcomes of EDA for loan lending include identifying patterns in borrower attributes, understanding the relationship between borrower and loan attributes, and refining loan approval criteria.
- EDA can also help in identifying high-risk borrowers, benchmarking against industry standards, and guiding marketing campaigns.
- EDA can inform decision-making, reduce risk, and enhance the profitability of lending operations.

Conclusion

In summary, EDA lending is a powerful tool for improving loan approval decisions. By understanding customer attributes and loan data, financial institutions can mitigate risks and make more informed lending choices.

Prepared By

Sandeep B [Group Facilitator]

Nishanth A V[Group member]