EDA Lending Group Case 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.

Introduction

The Lending Club case study is a popular dataset for data analysis and machine learning. It is a collection of loan collection of loan information from Lending Club, a peer-to-peer lending platform. The dataset includes includes information on loan amounts, interest rates, borrower demographics, and loan performance. performance.

- 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, loan defaults, understand the relationship between these attributes and loan defaults, and develop develop insights that can be used to predict whether a new loan applicant is likely to default..

Background

Gain insights into the lending industry and the challenges faced by financial institutions, such as risk assessment assessment and determining borrower creditworthiness.

- Key objectives of EDA in loan lending are to identify patterns, understand relationships, and develop predictive insights for new loan applicants.
- Loan defaults occur when borrowers fail to make scheduled payments, leading to financial losses for 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 the LendingClub case study by providing a comprehensive understanding of the data and identifying patterns that can inform strategic decisions and improve the company's overall risk management and lending practices.

Problem Statement

LendingClub case study is to develop a predictive model that can identify loan applications that are at a high risk of at a high risk of defaulting. This would help LendingClub to make better decisions about which loans to approve loans to approve and which loans to reject, thereby reducing the company's risk of financial loss.

- Understand the relationship between borrower demographics, loan terms, and loan performance.
- 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.
- 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

Exploratory Data Analysis (EDA) is a crucial step in data analysis, particularly when dealing with large and complex datasets like the LendingClub case study. It involves examining the data to gain insights about its characteristics, patterns, and relationships between variables.

Key Steps in EDA for the LendingClub Data

1. **Data Extraction and Data Cleaning**
2. **Univariate Analysis**
3. **Segmented Univariate Analysis**
4. **Bivariate Analysis**
5. **Derived Metrics**

Data Extraction & Cleaning

Loan lending data has been extracted and converts 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

There are various types of quality issues when it comes to data, and that’s why data cleaning is one of the most time-consuming steps of data analysis. For example, there could be formatting errors (e.g. rows and columns are merged), missing values, repeated rows, spelling inconsistencies ,etc.

- Data cleaning is the process that removes data that does not belong in your dataset.
- Data transformation is the process of converting data from one format or structure into another.
- Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing.

- This article focuses on the processes of cleaning that data.

In the below example we are removing the column which are having all null values and all 0 value in it this will be it this will be first step of cleaning the data.

```
Out[4]: desc                        12940
        emp_title                    2459
        emp_length                   1075
        pub_rec_bankruptcies          697
        last_pymnt_d                   71
        revol_util                     50
        title                          11
        last_credit_pull_d              2
        total_pymnt                     0
        pub_rec                         0
        revol_bal                       0
        total_acc                       0
        initial_list_status             0
        out_prncp                       0
        out_prncp_inv                   0
        total_rec_prncp                 0
        total_pymnt_inv                 0
        inq_last_6mths                  0
        total_rec_int                   0
        total_rec_late_fee              0
        recoveries                      0
        collection_recovery_fee         0
        last_pymnt_amnt                 0
        policy_code                     0
        application_type                0
        open_acc                        0
        id                              0
        earliest_cr_line                0
        delinq_2yrs                     0
        loan_amnt                       0
        funded_amnt                     0
        funded_amnt_inv                 0
        term                            0
        int_rate                        0
        installment                     0
        grade                           0
        sub_grade                       0
        home_ownership                  0
        annual_inc                      0
        verification_status             0
        issue_d                         0
        loan_status                     0
        pymnt_plan                      0
        url                             0
        purpose                         0
        zip_code                        0
        addr_state                      0
        member_id                       0
        dti                             0
        dtype: int64
```

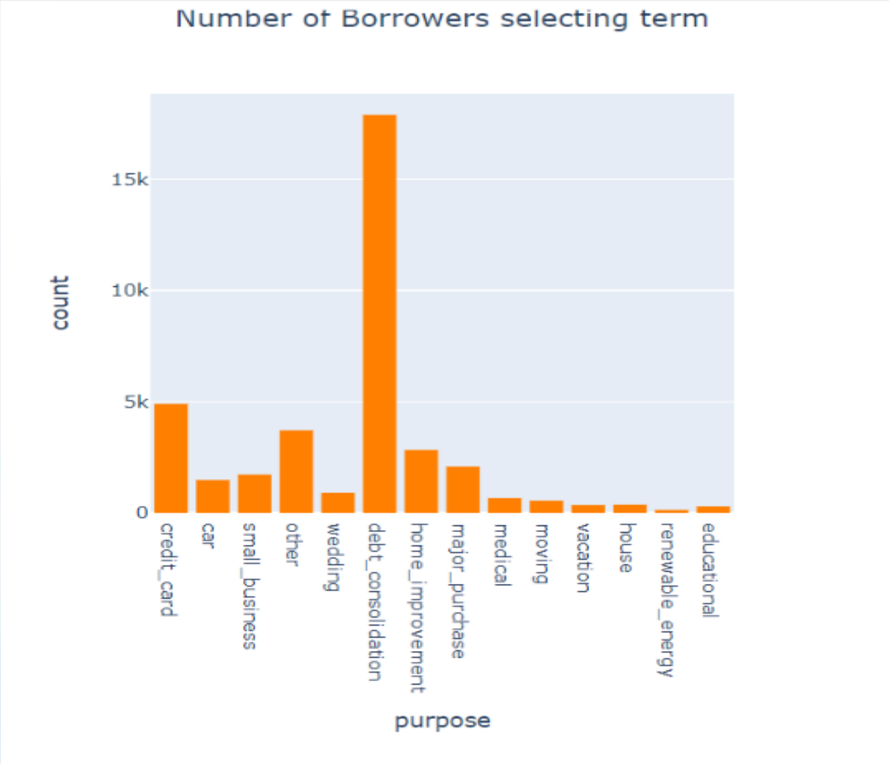
- Removed all the rows with majority Null values
- Removed Columns which are not defined in the Column Definition dataset
- Formatting to the correct type is also part of data cleaning so we formatted the unwanted data and converting to there specific type (type casting).

Univariate Analysis

As the term “univariate” suggests, this session deals with analyzing variables one at a time. It is important to separately understand each variable before moving on to analyzing multiple variables together.

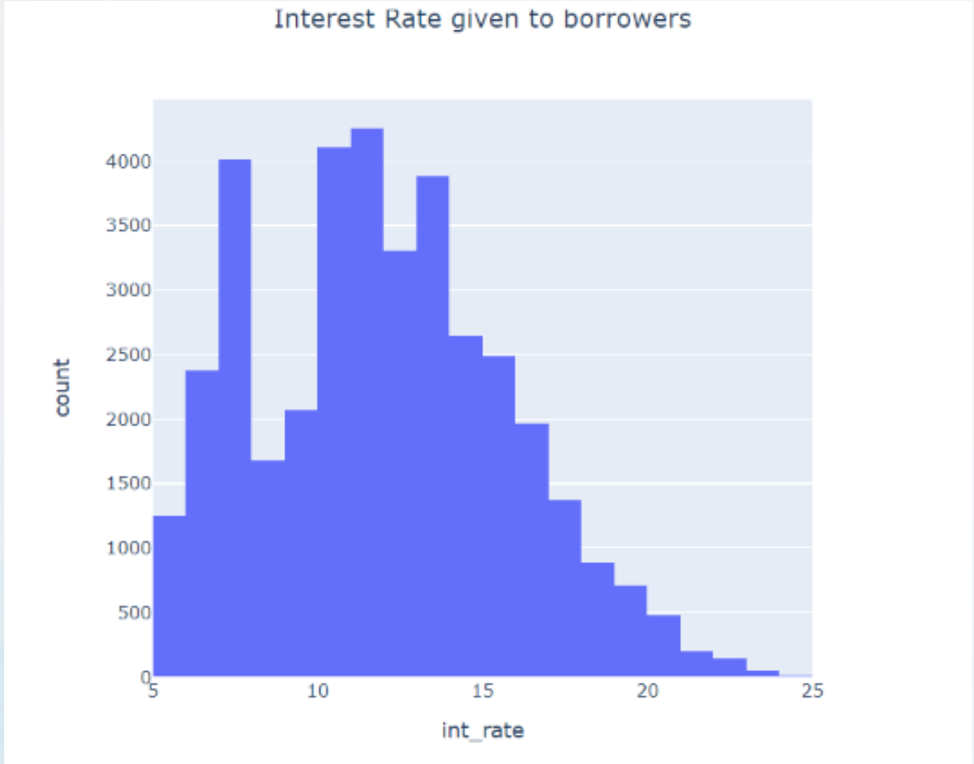
To apply univariate analysis for the given Dataset, we have identified the following variables

- Term, Grade, Sub Grade, Purpose, Interest rate, Loan status.
- We have used Data Visualization to plot graph for same variables.
- Following are the some of the graph used to plot using single variable.



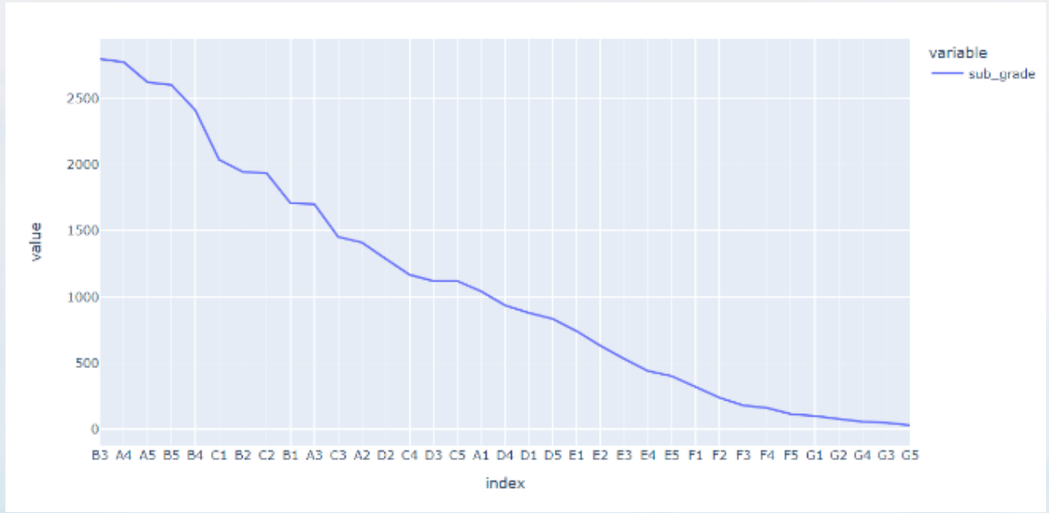
Histogram Graph of Purpose for Borrowers Borrowers

- Graph is plotted to understand the loan purpose for borrowers
- Borrowers of debt_consolidation purpose having having higher number of counts compare to others others



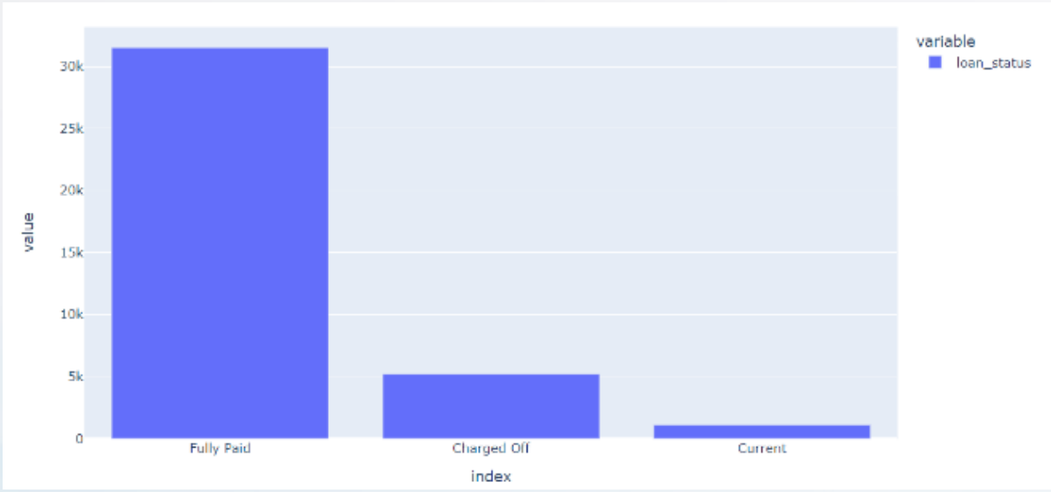
Histogram Graph of interest rate given to to borrowers

- Graph is plotted to see Interest rate given for borrowers



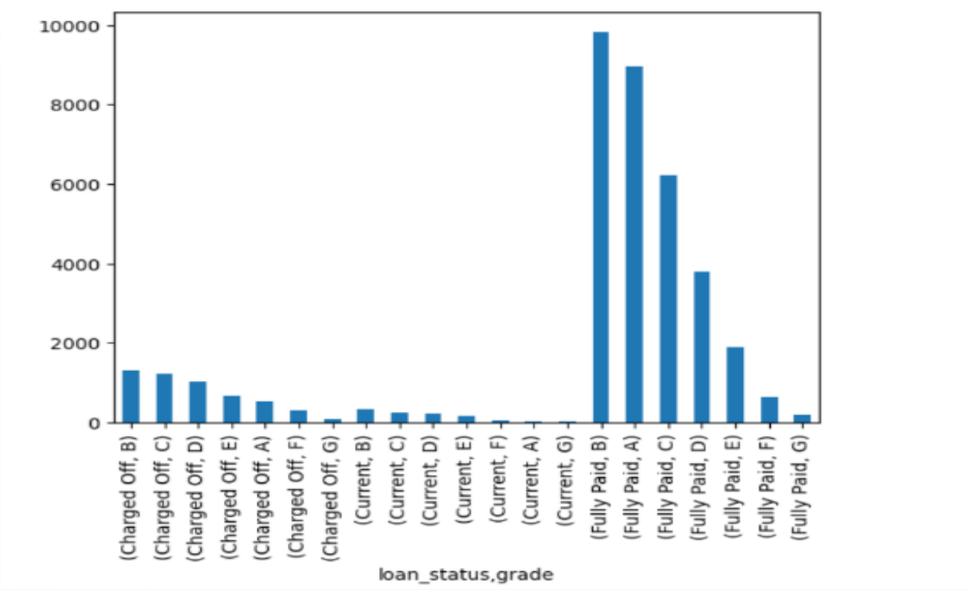
Sub Grade Line Graph

- Plotted the order univariant group that is the sub grade and number of people to apply for a loan.



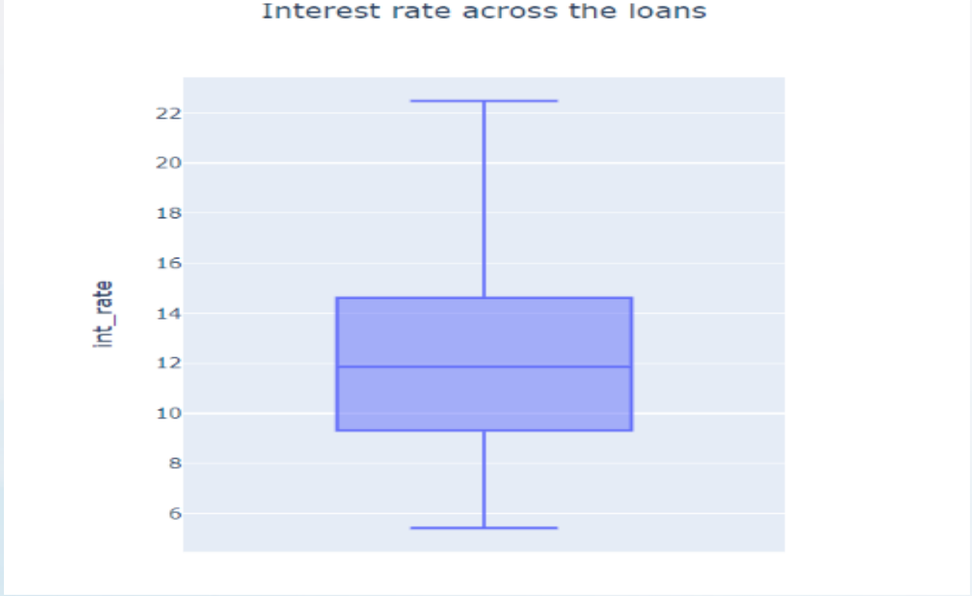
Loan Status Bar Graph

- In the Bar Graph we can see the loan status is status is more for Fully paid .
- We can also see the charged Off loan status is very low



Bar Plot for loan status, Grade

- In the Bar Plot we can see Group B having maximum loan loan taken of there is a chance in Charged off.
- We can also observe Group A having almost the same amount of Charged off



Box Plot for Interest rate

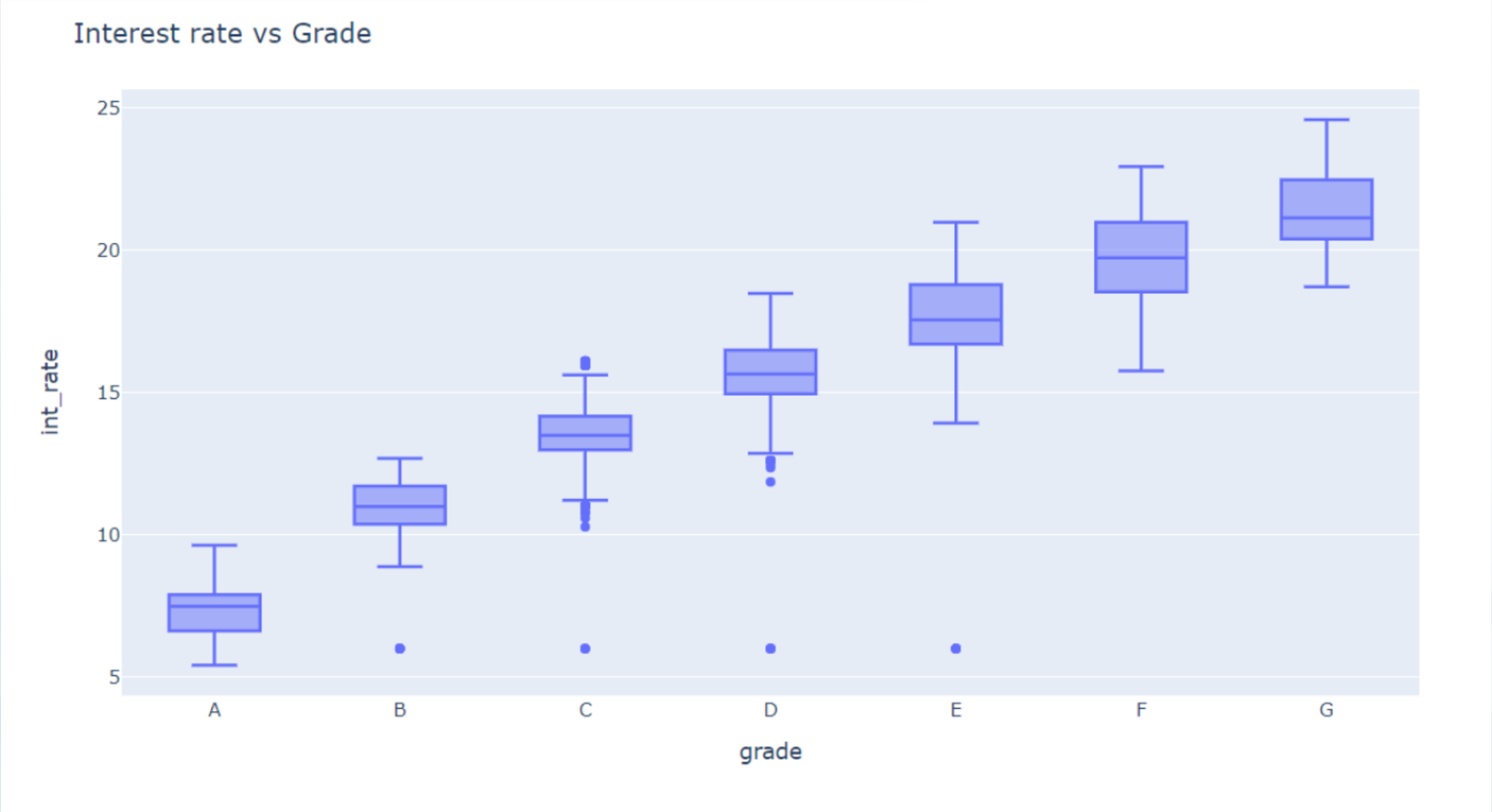
- We have plotted the Bar plot for Interest rate across the loans
- Inter-Quartile Range is almost approximate approximate to 5.2899

Segmented Univariate Analysis

Segmented univariate analysis is a data analysis technique that involves dividing a dataset into segments based on a categorical variable and then analyzing each segment separately using univariate analysis techniques. This technique can be used to identify patterns and trends in the data that would not be visible if the data were analyzed as a whole.

To apply Segmented Univariate analysis for the given Dataset, we have identified the following variables variables

- Grade, Interest rate.
- We have used Data Visualization to plot graph for same variables.
- Following are the some of the graph that we have plotted..



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 have taken loan increases

Bivariate Analysis

Bivariate analysis is a statistical technique used to investigate the relationship between two variables. It involves examining the association between two variables to determine if there is a correlation or not. Bivariate analysis can be used for both categorical and numerical variables.

There are two main types of bivariate analysis:

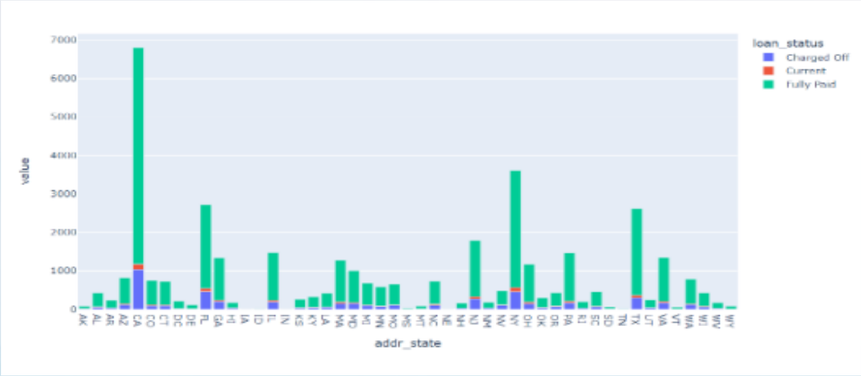
- **Categorical variable analysis**
- **Correlation analysis**

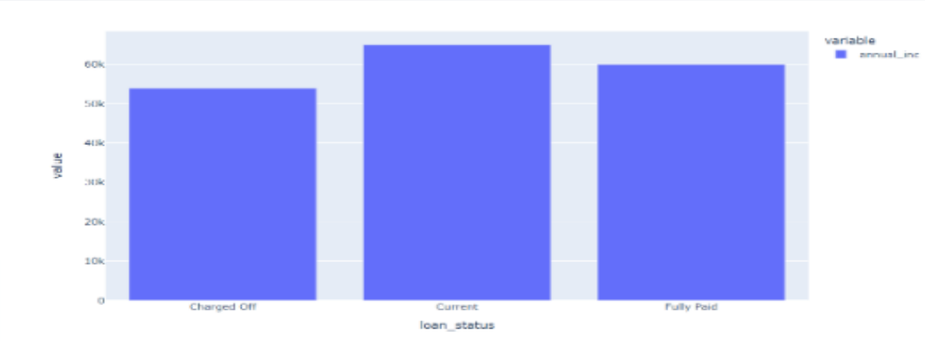
.To apply Bivariate analysis for the given Dataset, we have identified the following variables

- Term, Grade, Interest rate, Loan status, Verification status, Purpose, Funded amount, Home ownership, Address state, Annual income, Investors Funded amount.
- We have used Data Visualization to plot graph for two different variables.
- Following are the some of the graph that we have plotted

Categorical variable

The categorical bivariate analysis is essentially an extension of the segmented univariate analysis to another categorical variable





Bar Graph of home ownership for address address state

- In the bar graph we can conclude that the state CA as the most likely to take loan.
- Most numbers of borrowers in CA state are in rented home.
- Same analysis is true for NY state as well

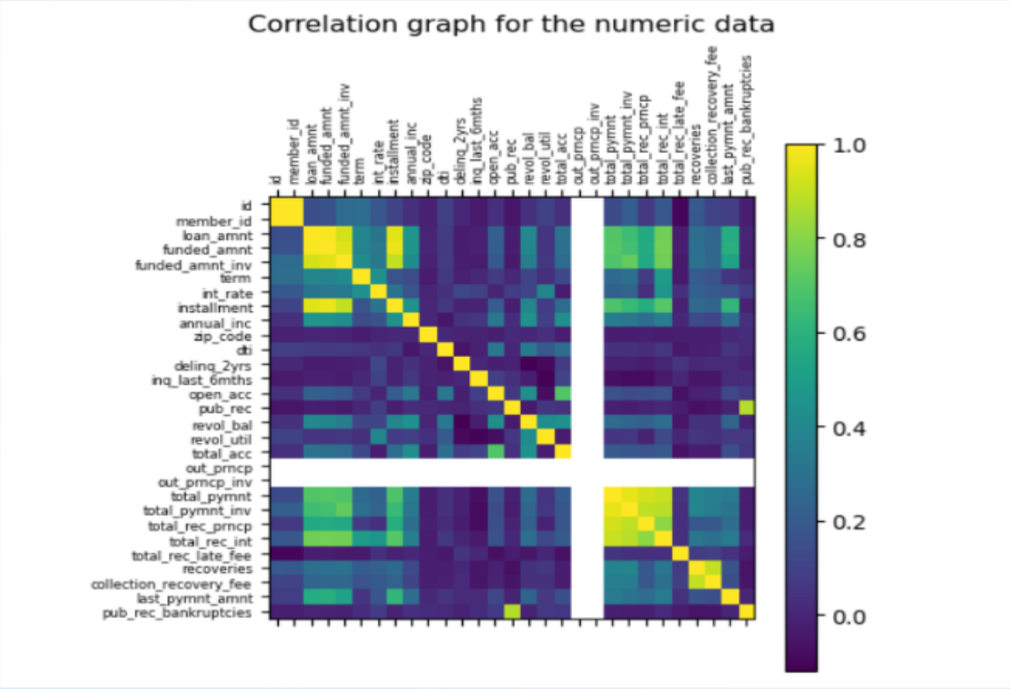
Bar Graph of annual income for loan status status

- We can observe clearly, For current ongoing loan the increment value is high ,so there is chance that it would be paid correctly
- For the Charged off increment value is little little low

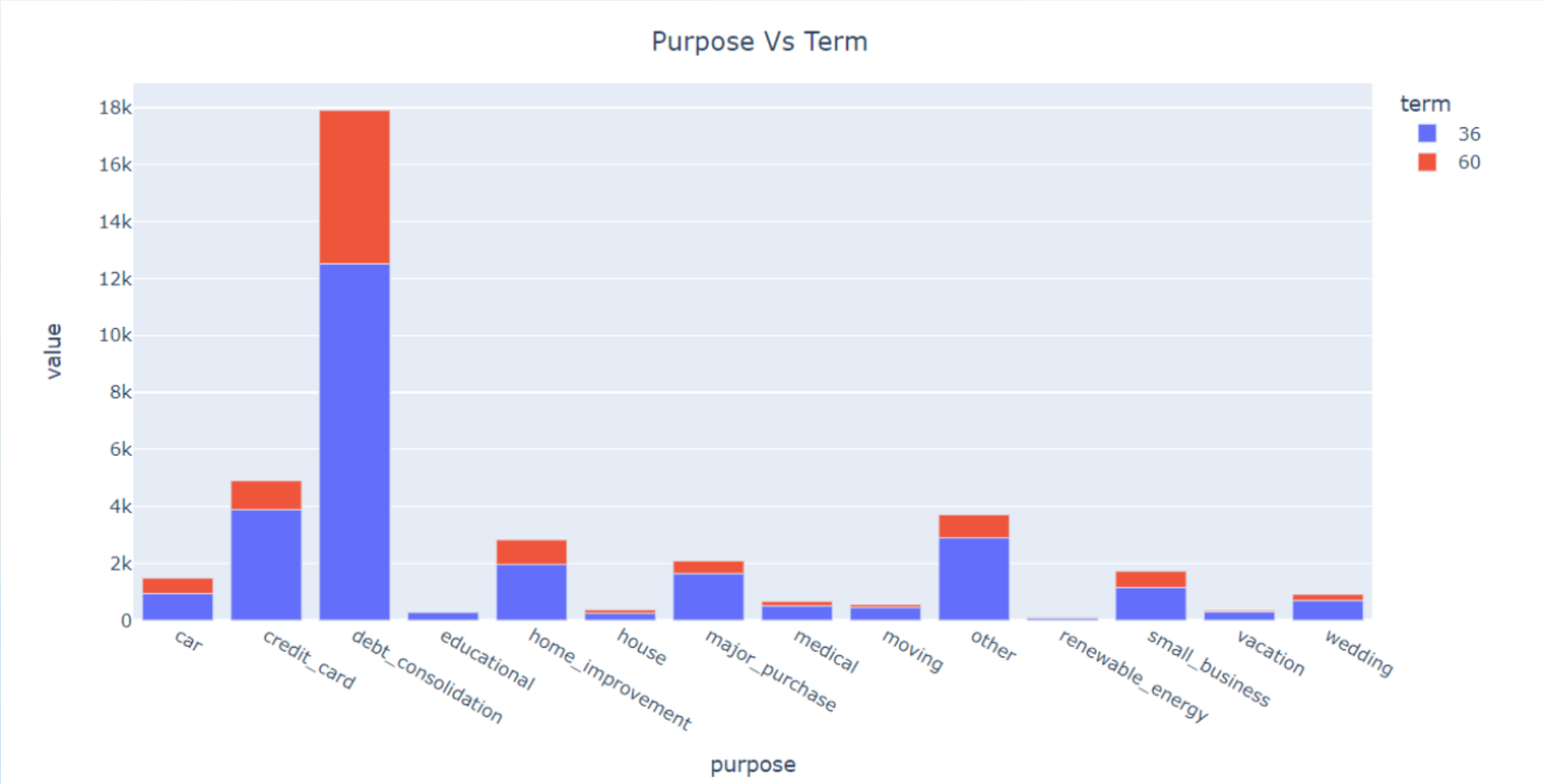
Bivariant Analysis of Correlation data

Bivariate analysis of correlation data involves examining the relationship between two variables to determine if determine if they are correlated and the strength of that correlation. Correlation measures the degree to which degree to which two variables move in tandem, indicating whether they tend to increase or decrease together. decrease together.

Checking any **correlation data in numeric data**



From the above graph we can clearly identify the positive co-relational data and where we can even categorize the columns loan amount, funded amount, funded amount with interest correlate with total payments, total payment with interest and total record principal



Bar Graph of purpose for loan term

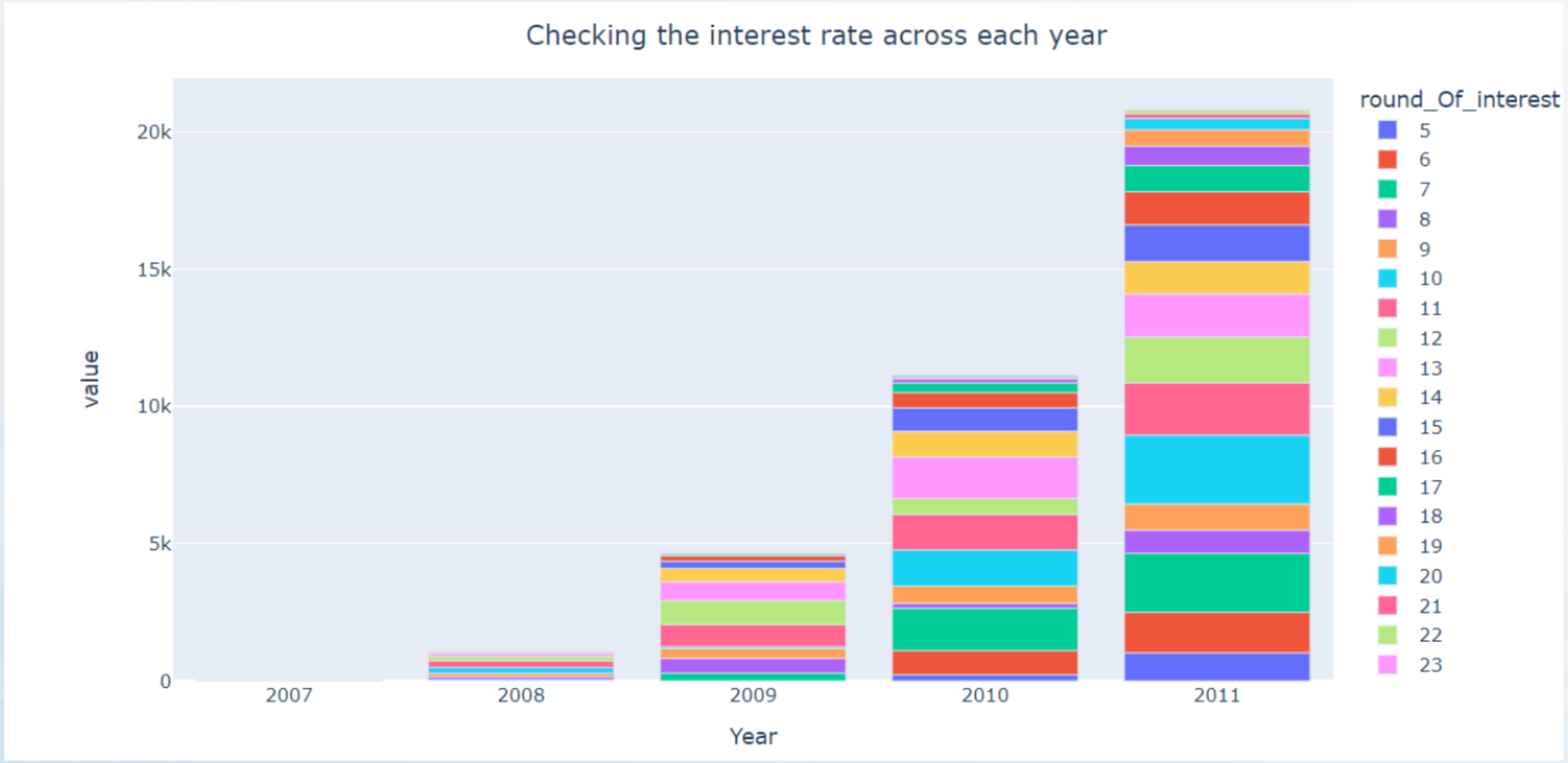
- In the bar graph we can see most number of loans taken by the borrowers are for debt consolidation and consolidation and Most of them are for 36 months terms
- As we can see here that the number of borrowers are disproportionate.

Derived Metrics

Derived metrics are new metrics that are created by combining two or more existing metrics or variables. These variables. These metrics are useful for gaining insights into the underlying patterns and trends in the data and for data and for measuring the impact of various factors on the outcome variable.

To apply Derived metrics for the given Dataset, we have identified the following variables

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



Bar Graph of interest rate for every Year

- Graph clearly says that it generalizing the the interest rate to round of part
- We have used the crosstab for this graph
- In the graph we can see Interest rate increases exponentially from 2007 to 2011.

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 understanding the relationship between borrower and loan attributes, and refining loan approval criteria 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 understanding ,customer attributes and loan data, financial institutions can mitigate risks and make more more informed lending choices

Prepared by

Likhith D Gowda[Group Facilitator]

Rajesh R [Group Member]