Lending Club Case Study

ML-C63 Batch

Soumendu Dasgupta
Sonal Jatav

Overview

EDA(Exploratory Data Analysis) on loan dataset to analyse factors predominant in customers who default loans

Analyzing Factors Predominant in Defaulting Loan Repayment

Problem Statement

A consumer finance company that lends different loans to urban customers wants to be able to assess the risks related to grarting loan to a customer.

- Assess risks related to granting loans to customers.
- Identify high-risk applicants to reduce credit loss.
- Determine factors influencing loan defaults.

Outcome: Based on `risk assessment` company should be able to identify and `accept` or `reject` loan application.
Factors/ variables would determine:

1. How many factors does a person suffice to default a loan and is a bad candidate to grant loan
2. If a customer is high risk then should the interest rate be increased
3. Is the customer a good candidate to lend with all variables in their favor
4. Is the customer a good candidate to lend with a few impactful variables in their favor

Assumptions

1. Customers with Loan status **Current** are not considered in deriving the metrics.
   Since they can either turn defaulters or fully pay the loan, hence rendering the data useless for deriving defaulters
2. Following columns are removed
   a. id: redundant column to member_id
   b. url, desc, earliest_cr_line, revol_bal, title, emp_title, collection_recovery_fee: not required for credit loss analysis
   c. zip_code: masked therefore cannot be of help in analysis

Data Analysis

Data Overview

- Dataset: Loan data with 39,717 records and 111 columns.

- Key Columns: loan_amnt, int_rate, annual_inc, dti, loan_status, etc.

- Key Records: Records where loan_status = ‘Charged Off’ OR ‘Fully Paid’

Data Cleanup

- Remove columns with all null values
- Remove columns with more than 60% null values
- Remove columns with non-unique value count as 1
- Remove columns from the list [url, desc, earliest_cr_line, revol_bal, title, emp_title, collection_recovery_fee, id, zip_code]: *as they did not have viable data for analysis*
- Convert columns to appropriate data types: *category & dateTime*
- Derive columns: [issue_d_year, issue_d_month, open_acc_groups, total_acc_groups, annual_inc_groups, loan_amnt_groups, int_rate_groups]
- Remove outliers *annual_inc > 95th percentile*
- Remove records with loan_status = ‘Current’

Data Analysis

- Univariate Analysis
Based on categorical and quantitative variables

- Bivariate Analysis
Based on correlation between continuous and categorical variables

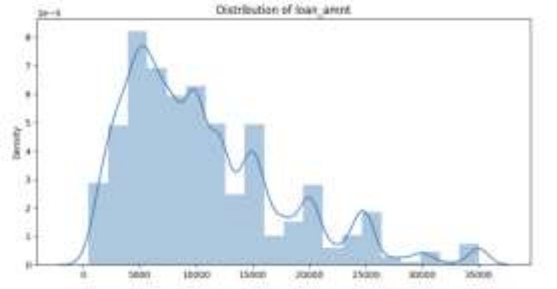
- Multivariate Analysis
Based on quantitative variables derive correlation or driving factors

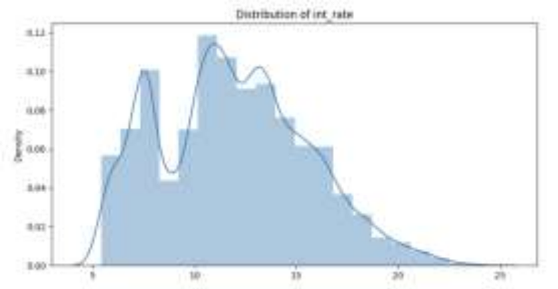
Results

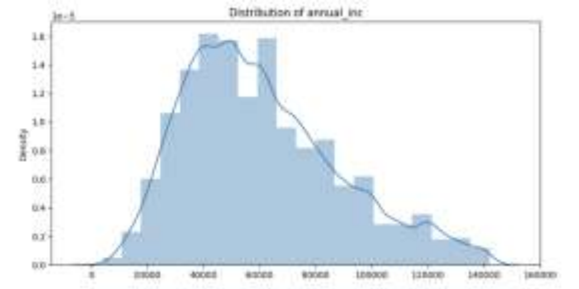
- Identified key factors influencing loan defaults.
- Recommendations for assessing loan applicants to reduce credit loss

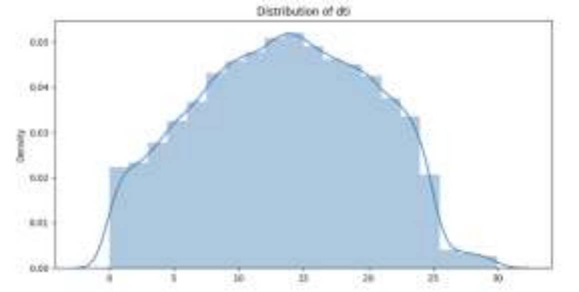
— Univariate Analysis Results

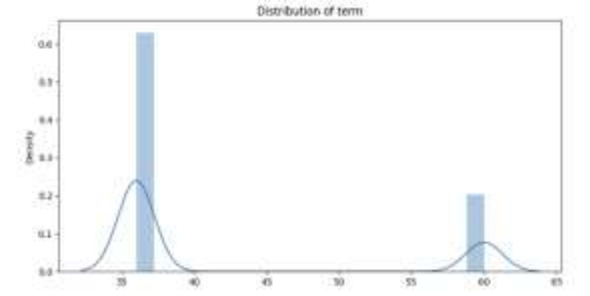
- **Loan Amount**: Most loans are small to moderate amounts with a few large loans.
- **Interest Rate**: Interest rates vary, but there might be common rates around certain values.
- **Annual Income**: Most borrowers have moderate incomes; a few have very high incomes.
- **Debt-to-Income Ratio**: Most borrowers have a manageable Debt To Income ratio, but some might have higher values indicating more debt.
- **Term**: A significant proportion of loans are of term 36 months





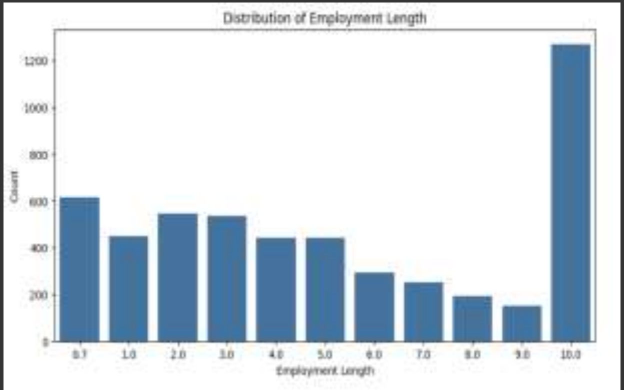


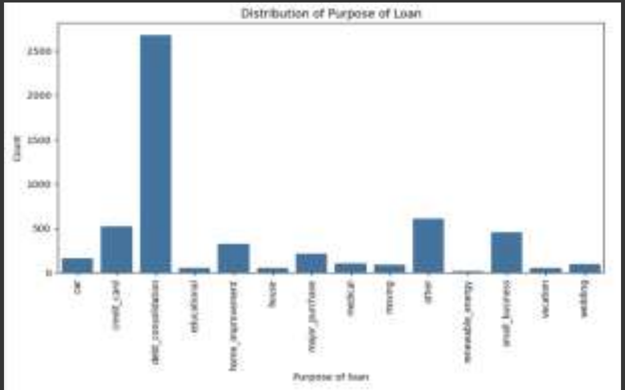


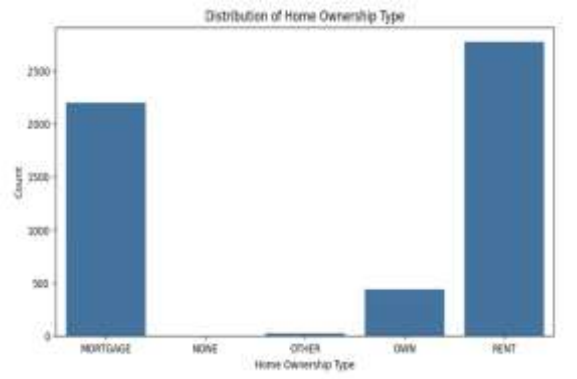


— Univariate Analysis Results

- Majority of the Charged off loans belong to customers with employment length around **10 and above years**
- Majority of the Charged off loans belong to customers who took loan for purpose of another **debt consolidation**
- Majority of the Charged off loans belong to customers who live in a **rented place** and second to it who **already have mortgage** on their homes



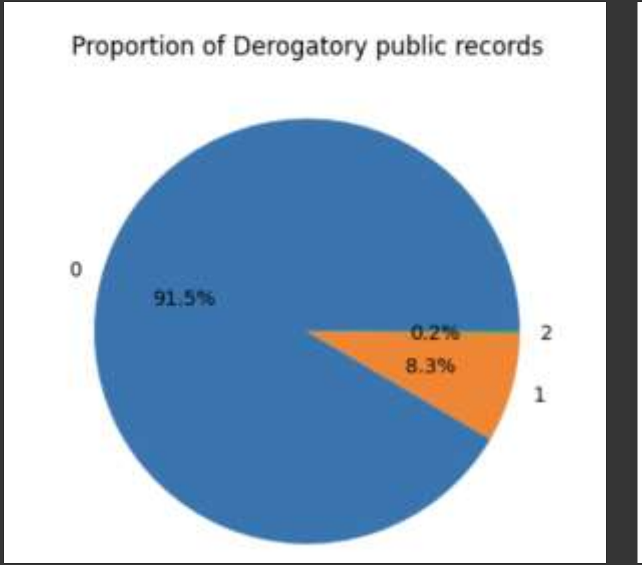


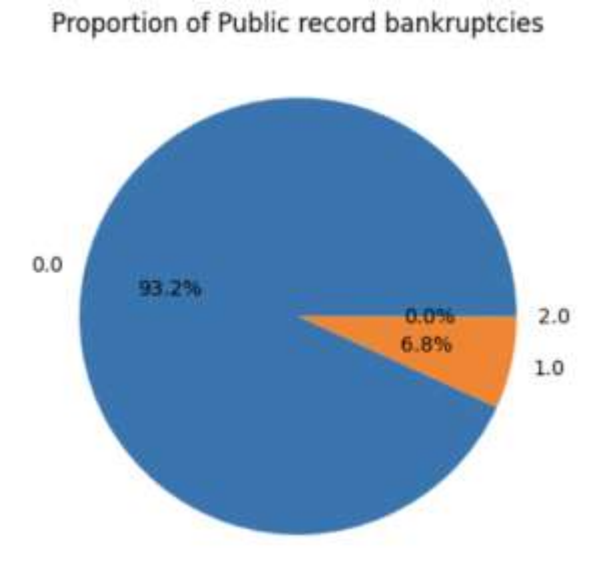


— Univariate Analysis Results

- **Public record bankruptcies:** A very small percentage of loan defaulters have filed bankruptcies.
- **Derogatory public records:** A very small percentage of loan defaulters have derogatory public records.

Although these can be high risk customers to lend

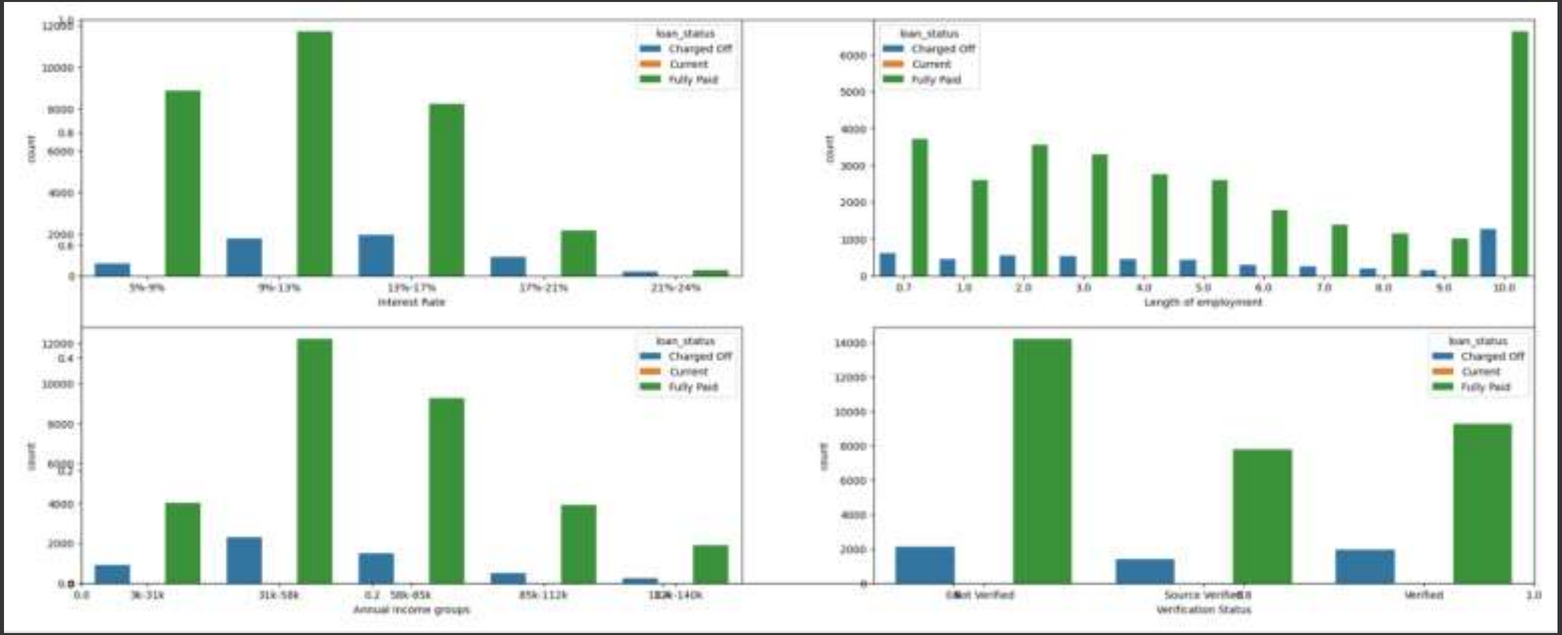




— Bivariate Analysis Results

Factors showing higher charged off loans

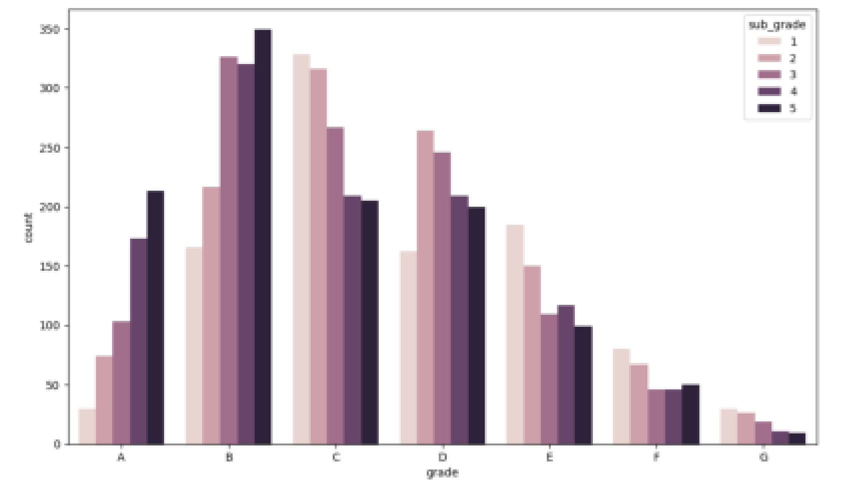
- **Interest Rate:** Higher interest rates correlate with a higher risk of loan default.
- **Length of Employment:** Stable, long-term employment is associated with better loan repayment, though it's not a strong standalone predictor.
- **Annual Income Groups:** Higher annual incomes (above 58k) are associated with a higher likelihood of fully repaying loans. Lower income groups (3k-31k) show a higher risk of default, suggesting that income level is an important factor in assessing loan repayment capability.
- **Verification Status:** Verified income and employment significantly reduce the risk of loan default.

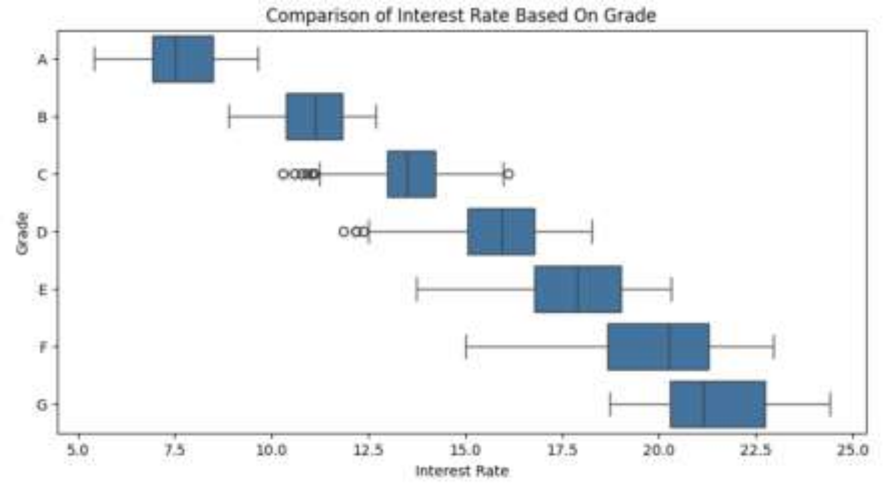


— Bivariate Analysis Results

Loan Grades and SubGrades

- **Higher Grades (A, B):** Likely to have fewer charged-off loans due to lower risk.
- **Moderate Grades (C, D):** Higher likelihood of charged-off loans compared to higher grades.
- **Lower Grades (E, F, G):** Higher risk categories with more charged-off loans.

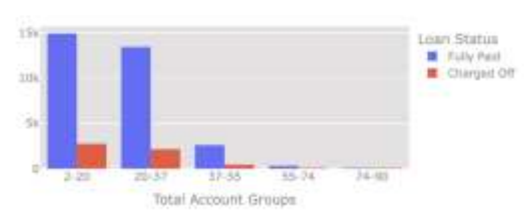


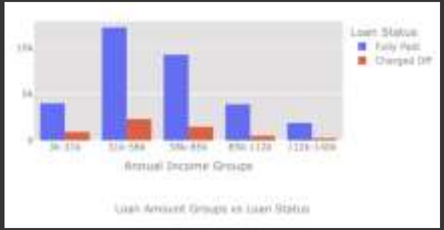


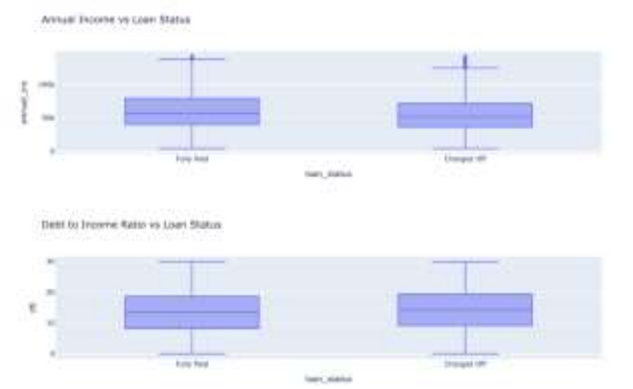
— Bivariate Analysis Results

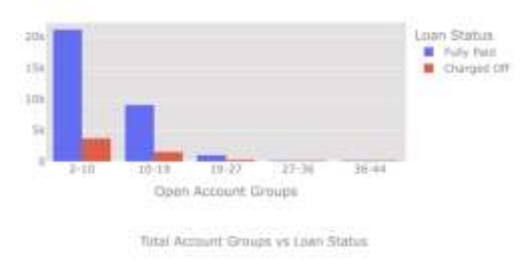
Maximum number of loan defaulters(Charged Off loan_status) were in the following ranges:

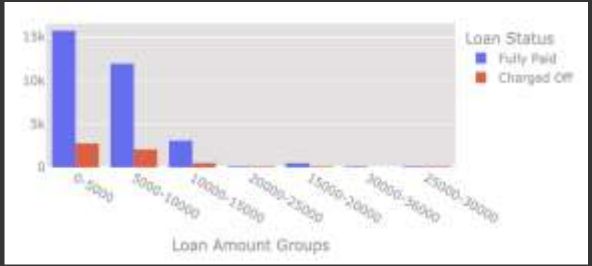
- Annual Income - **36.3k - 71.67k**
- Debt to Income Ratio - **9.18 - 19.40**
- **Interest rates - 9%-17%**
- **Open credit lines - 2-10**
- **Total credit lines currently in the borrower's credit file - 2-37**
- **Loan Amount - 5k - 10k**







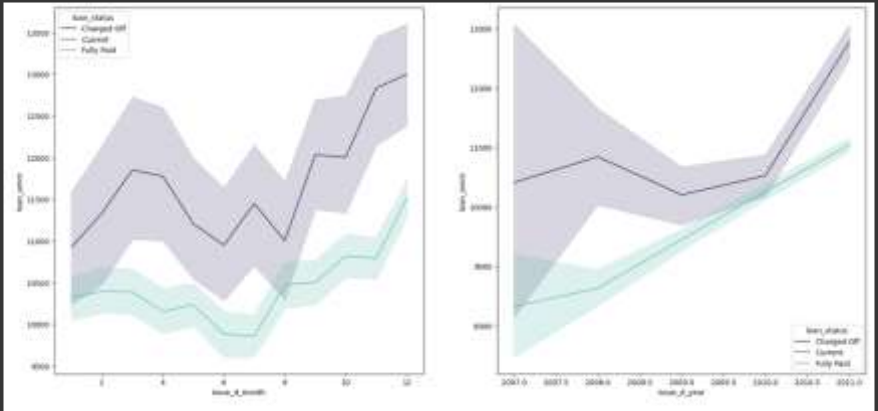


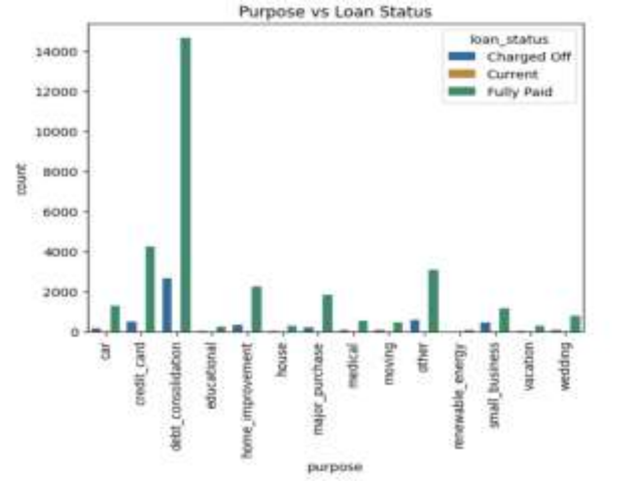


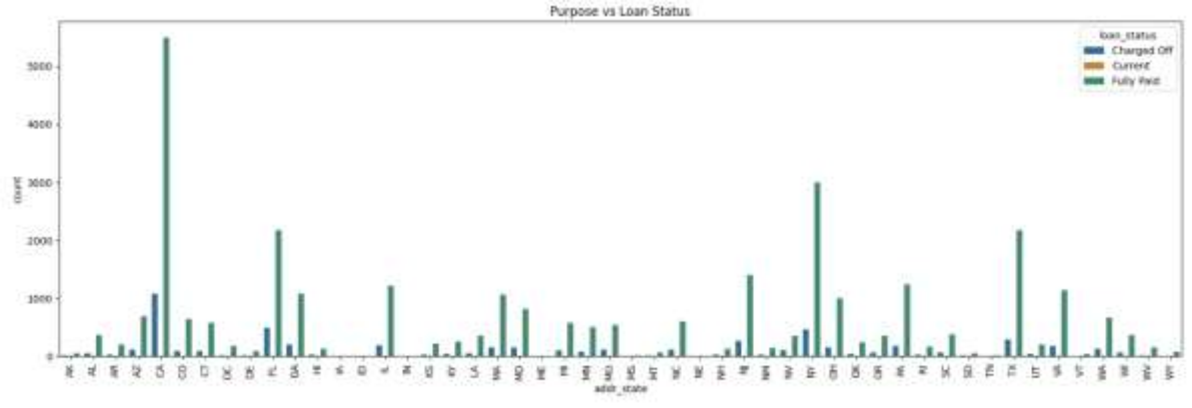
— Bivariate Analysis Results

Highest number loan defaulters(Charged Off loan_status) are as follows:

- Requested loan during the month - **December**
- Requested loan during the year - **2011**
- Requested loan for Purpose - **Debt Consolidation**
- Requested loan from State - **CA - California**

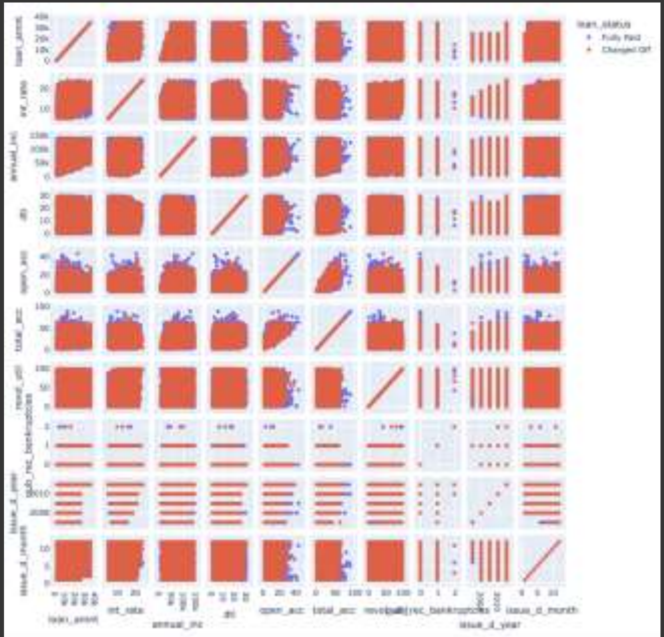


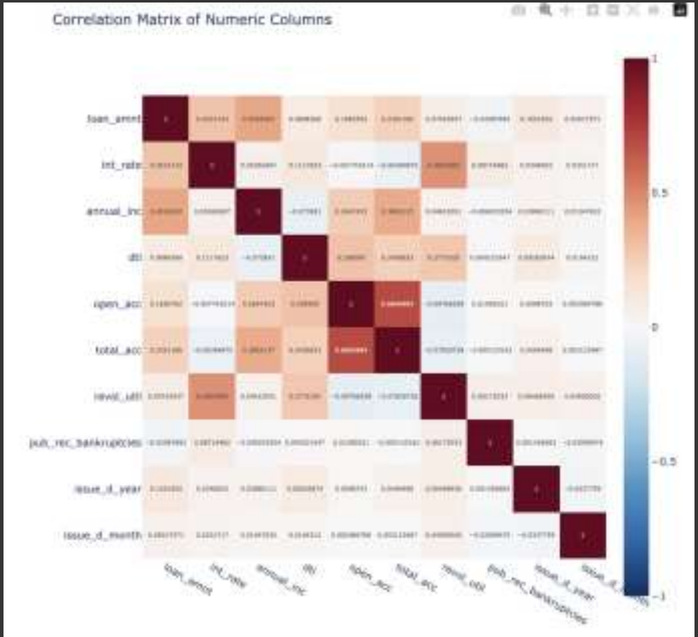




— Multivariate Analysis Results

- **Interest Rate (int_rate):** Higher rates are associated with higher default risk.
- **Annual Income (annual_inc):** Lower incomes are linked to higher default rates.
- **Debt-to-Income Ratio (dti):** Higher DTI ratios are indicative of higher default risk.
- **Revolving Utilization (revol_util):** Higher utilization rates correlate with higher default risk.
- **Public Record Bankruptcies (pub_rec_bankruptcies):** History of bankruptcies increases default likelihood.





Recommendations

1. Key Driving Factors:
    a. DTI Ratio
    b. Annual Income
    c. Loan Grade/ Subgrade
    d. Public Record Bankruptcies
    e. Verification Status
    f. Purpose

2. Driving factors in conjunction with above are:
   Higher chunk of **customers** who defaulted loan had the following traits too
    a. Annual Income - **36.3k - 71.67k - customers belonged to average to lower income ranges**
    b. Purpose: **Most defaulters borrowed for consolidating another debt which can be a high risk pattern for defaulting loan**
    c. Public Record Bankruptcies - **These are high risk customers which can also relate to taking loan for debt consolidation**
    d. DTI Ratio: **Customers with higher DTI and higher open credit lines**
    e. Interest Rate: **Higher interest rates with average income can be another factor for defaulting**
    f. High Revolving Utilisation: **High revolving utilisation with higher interest rate and average income also can lead to defaulting lon**
    g. Lower Grades (E, F, G): **Higher risk categories with more charged-off loans**
    h. Month of the year: **December being holiday season may be a cause for people to borrow**