Lending Club Case study

Submitted by:
Siddhardha Mudumba
Siju Babu

What is the Lending Club Case Study all about?

Business Objective:

1. The company aims to identify risky loan applicants who are likely to default, thereby reducing credit loss and minimizing financial losses from such defaults.

2. The primary objective is to analyze driving factors behind loan defaults using exploratory data analysis (EDA) to pinpoint variables that are strong indicators of borrower risk.

3. This knowledge will be used by the company to enhance its portfolio and risk assessment processes, allowing for better loan approval decisions.

What are the actions taken along with the key findings?

Data cleaning

1

- Remove rows with loan status as “Current”
- Drop rows/columns with null values and same values.
- Retain only those columns relevant for analysis.
- Remove all unwanted texts from the columns
- Add new columns (Month and Year) from Issue_d.

Outlier detection

2

- Identify and remove outliers in loan amount and annual income using IQR

Key findings from EDA

3

- Bulk of the defaulters occur within the 60 month term.
- A median defaulter earns less but pays a significantly higher interest rates than a fully paid customer.
- There exist a medium positive correlation between loan amount and annual income. Similarly, between term and interest rate as well.

Data Cleaning and Outlier detection

Read loan data from loan.csv file.

```
# Read dataset
loan_data_original = pd.read_csv("loan.csv", low_memory=False)
```

Total rows and columns

```
print("Rows, Columns : ",loan_data_original.shape)
```

```
Rows, Columns :  (39717, 111)
```

1. Total rows and columns in the csv

```
#### What are the different values in loan_status
loan_data_original["loan_status"].value_counts()
```

```
loan_status
Fully Paid     32950
Charged Off     5627
Current         1140
Name: count, dtype: int64
```

2. Remove “Current” loan status

```
Total columns with all null values :  55

loan_data = loan_data.dropna(axis="columns", how="all")
```

3. Remove 55 columns with null

```
# 11 columns are dropped. New dataset will contain the remaining 45 columns.
loan_data = loan_data.loc[:, loan_data.nunique() > 1]
```

4. Remove 11 columns with same values

```
loan_data.columns
```

```
Index(['member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term',
       'int_rate', 'installment', 'grade', 'sub_grade', 'emp_length',
       'home_ownership', 'annual_inc', 'verification_status', 'issue_d',
       'loan_status', 'purpose', 'zip_code', 'addr_state', 'dti', 'pub_rec',
       'total_pymnt', 'pub_rec_bankruptcies'],
      dtype='object')
```

5. Final list of 22 columns after removing irrelevant ones

```
loan_data.dropna(subset = ["emp_length"], inplace=True)
loan_data.dropna(subset = ["pub_rec_bankruptcies"], inplace=True)
loan_data_null_mean = (100* loan_data.isna().mean().sort_values(ascending = False))
print(loan_data_null_mean)
print("Dataset : ", loan_data.shape)
```

6. Remove rows with null values

```
# Remove the % from int rate.
loan_data["int_rate"] = loan_data["int_rate"].str.strip("%").astype(float)

# Remove "months" from term field.
loan_data["term"] = loan_data["term"].str.strip("months").astype(float)
```

7. Remove unwanted texts from the columns

```
# Issue_d is a column with Month-Year. Adding 2 new columns to separate month and year.
loan_data["issued_year"] = pd.to_datetime(loan_data["issue_d"], format='%b-%y').dt.year
loan_data["issued_month"] = pd.to_datetime(loan_data["issue_d"], format='%b-%y').dt.month
```

8. Create Derived Metrics from “issue_d”

```
def calculcateIQR(colName):
    #Calculate IQR
    Q1 = loan_data[colName].quantile(0.25)
    Q3 = loan_data[colName].quantile(0.75)
    IQR = Q3 - Q1

    print("Q1 : {} Q3 : {} IQR : {}".format(Q1,Q3,IQR))

    # lower and upper bounds for outlier
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    return lower_bound,upper_bound
```

9. Function to calculate IQR

```
loan_data = loan_data[(loan_data["loan_amnt"] > lower_bound) & (loan_data["loan_amnt"] < upper_bound)]
loan_data.shape
```

```
(35769, 23)
```

```
# Lets do the outlier removal for annual income.(annual_inc)
# we can write a common function, Since it's only 2 columns,  Duplicated the code.

lower_bound,upper_bound =  calculcateIQR("annual_inc")
loan_data = loan_data[(loan_data["annual_inc"] > lower_bound) & (loan_data["annual_inc"] < upper_bound)]
loan_data.shape
```

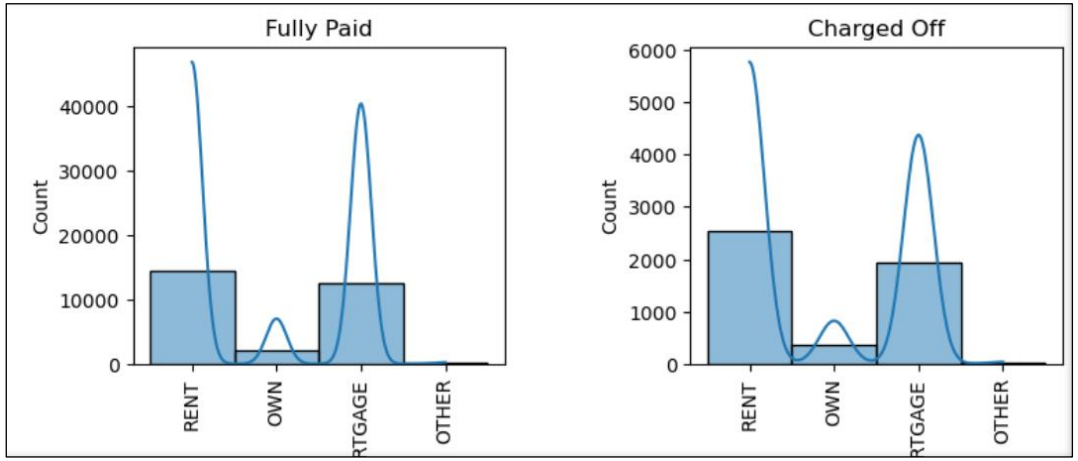
10. Remove outliers based on IQR

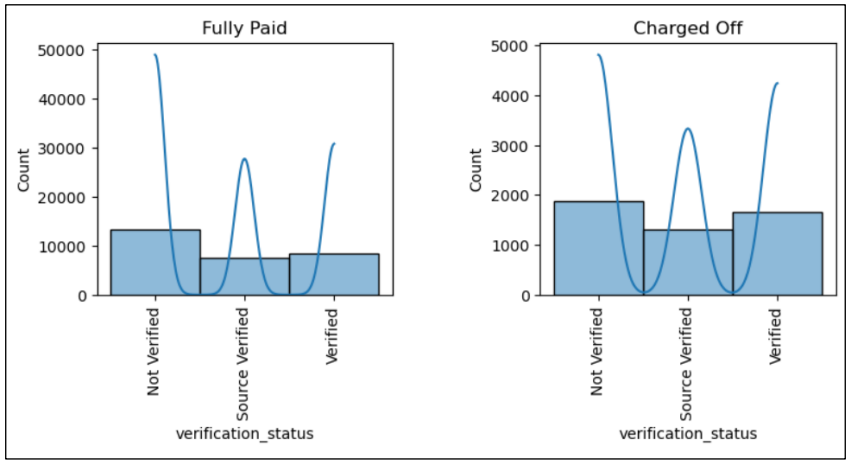
```
Q1 : 40269.0 Q3 : 80000.0 IQR : 39731.0
(34014, 23)
```

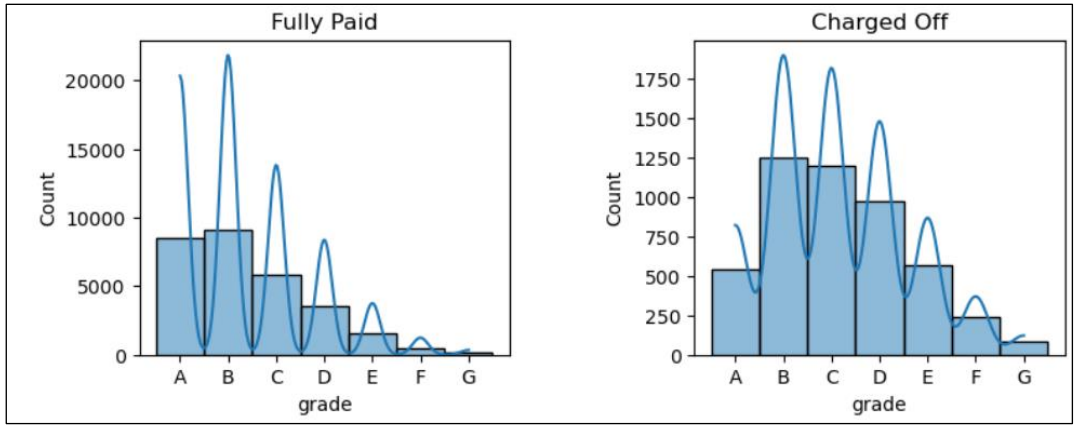
11. Final shape of the data for EDA after data cleaning

Key Insights from EDA - 1

1. While most of the insights from univariate analysis on (un)ordered categorical variables for defaulters and fully paid borrowers are similar it is quite interesting to understand how they are segmented based on house ownership, verification status and grade:





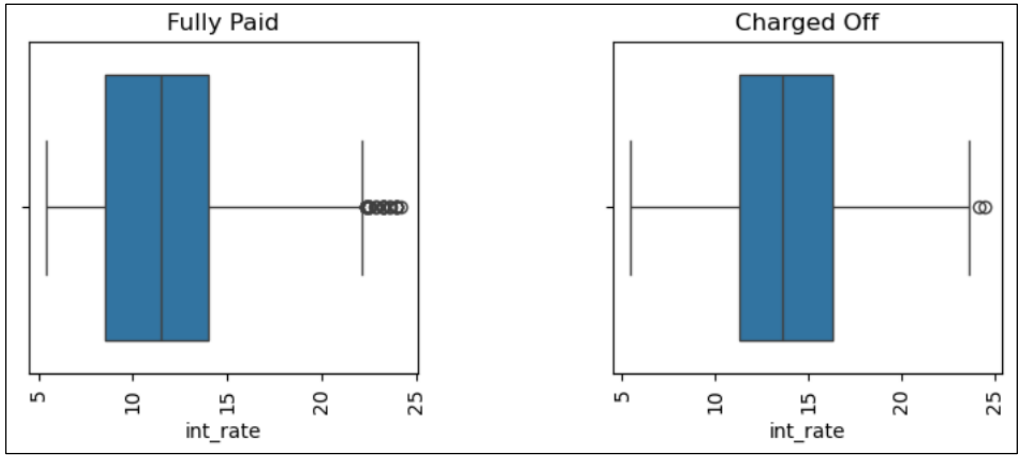


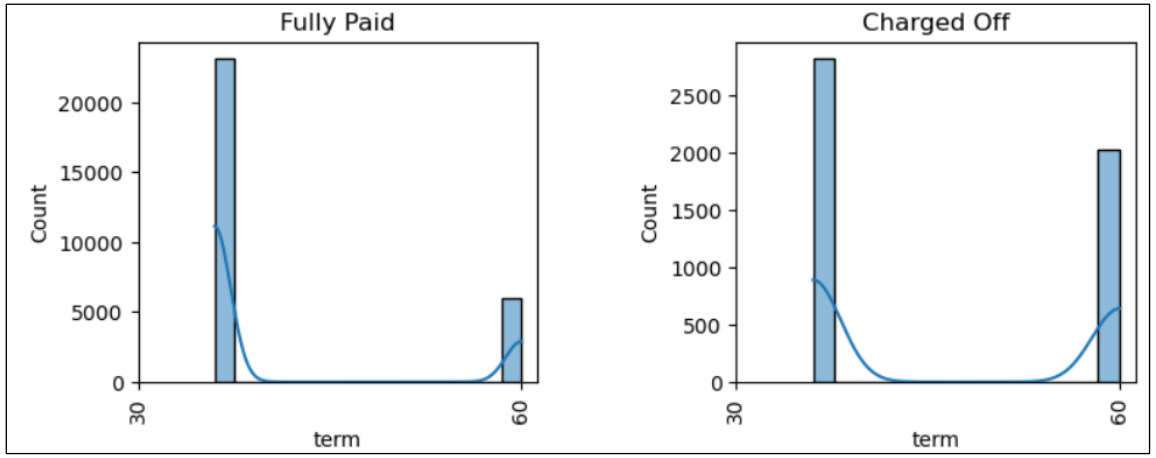
Insights:

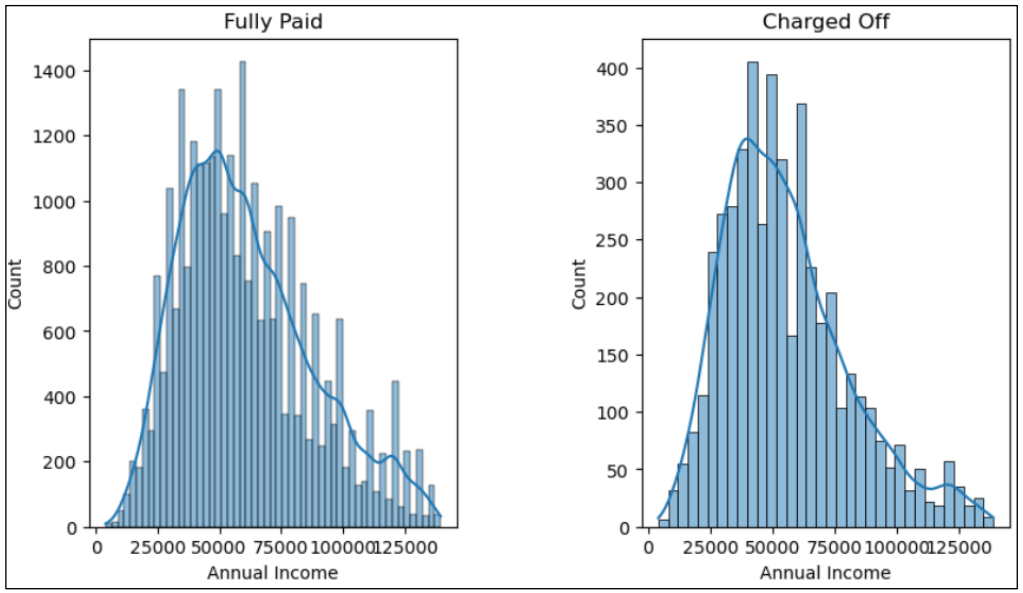
- Over 90% of the borrowers from both groups either have rented or mortgaged their house.
- Interestingly, most borrowers whose income was not verified have either paid fully or have defaulted. While those whose source is verified defaults lesser.
- Surprisingly as the customer grades moves from a higher to lower, the number of defaulters kept decreasing.

Key Insights from EDA - 2

2. We can see some interesting insights on default borrowers when we perform univariate analysis on quantitative data : interest rates, term and annual income





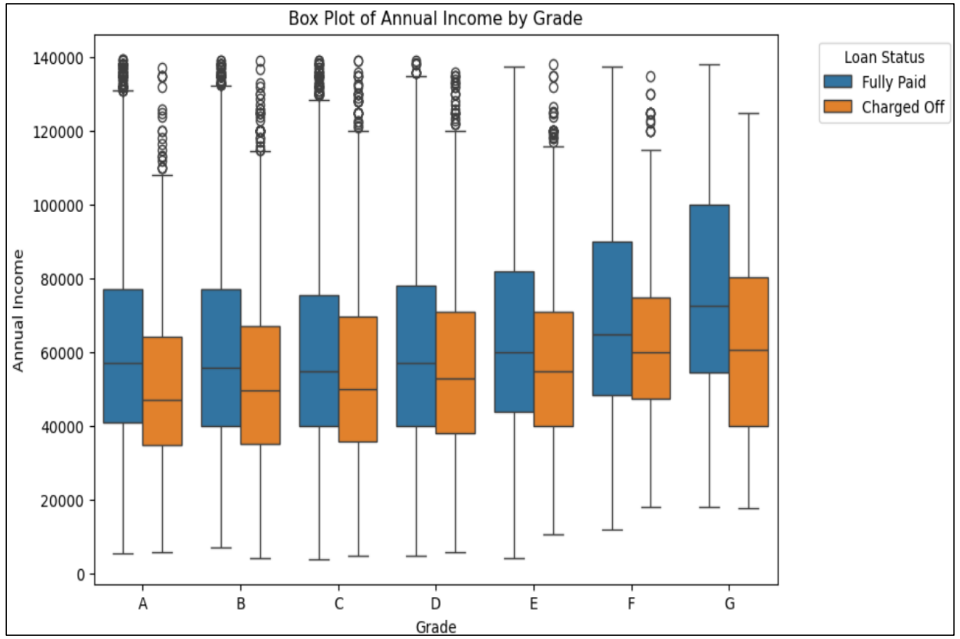


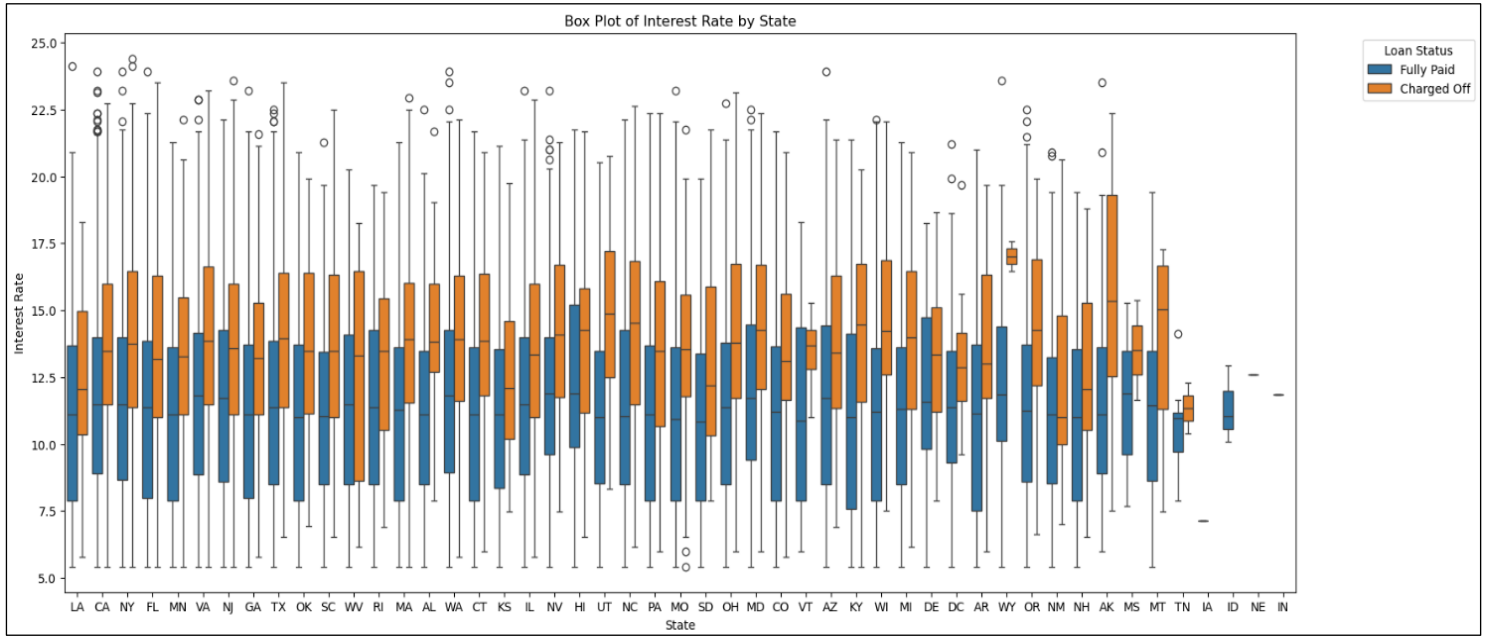
Insights:

- 43% of the defaulters have defaulted when the term is 60 months.
- Probably the reason for that can be the significantly high interest rates paid by the defaulters. The median interest rate of the defaulters is almost the same as the 75th percentile of that of the fully paid borrowers.
- Moreover, we also see that the defaulters have a median annual income ($50,000) which is way less than that of fully paid borrowers.

Key Insights from EDA - 3

3. Next when performing bivariate analysis, that is Annual Income by Grade | Interest Rate by State we further are able to uncover more insights focused on defaulters



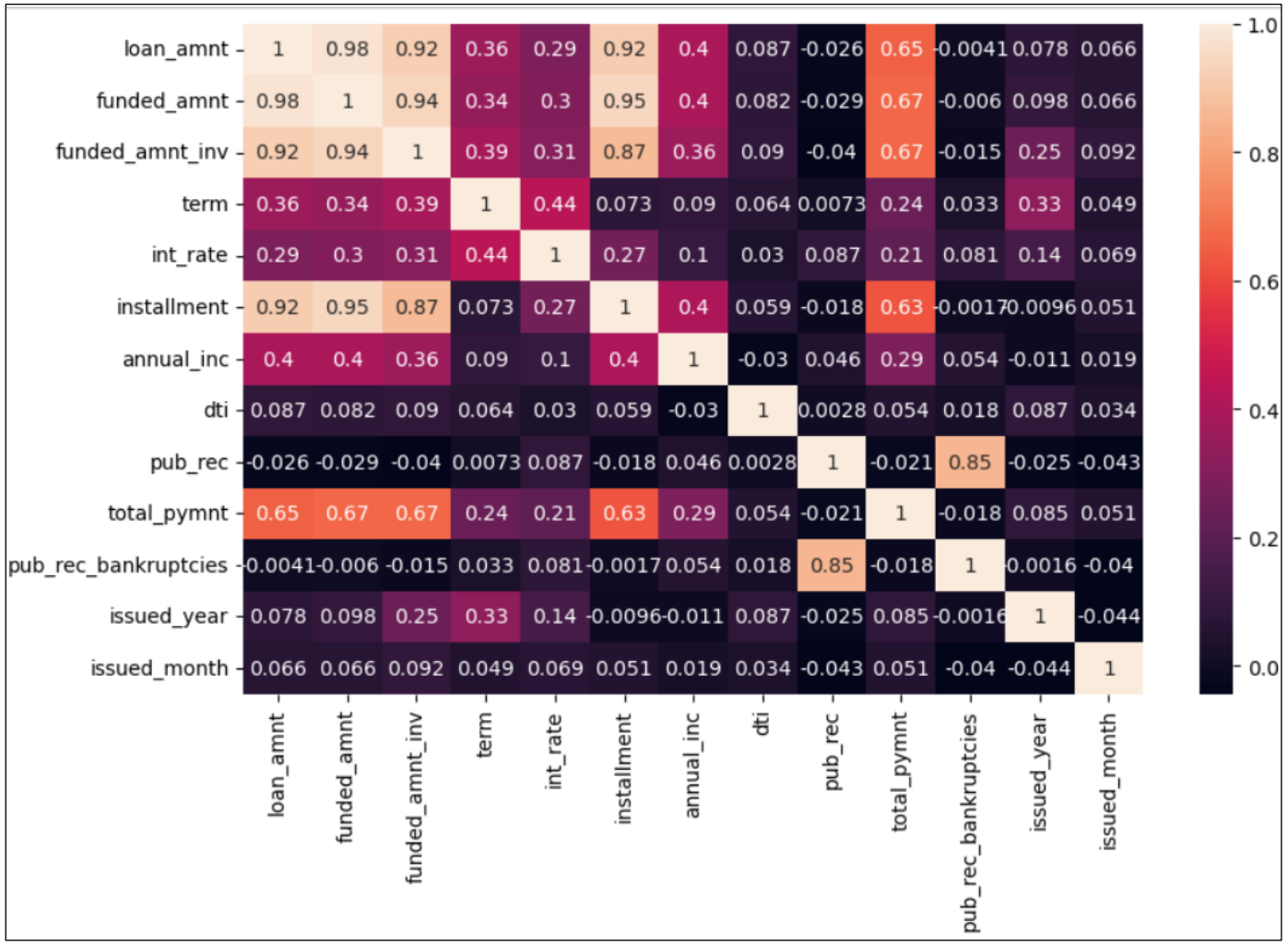


Insights:

- Borrowers under grade G who defaulted have the highest distribution of annual income amongst all.
- Borrowers from Alaska need to be considered more cautiously since the defaulters in this state tend to pay the highest interest rates amongst all the states and hence defaults.

Key Insights from EDA - 4

3. Finally, upon performing correlation analysis on defaulters, we can see some moderate positive correlations as well.



Insights:

- It is obvious to see strong correlations between, loan amount, funded amount, funded amount investors , installments and payments as they are related.
- Interesting to see moderate correlation between, term and interest rates. This clearly explains why defaulters tend to default. As the term increases, the interest rates also increases which trigger the borrowers to default.
- We also see a similar correlation between loan amount and term. That is as loan amount goes up, the term may also go up thereby increasing the interest rates.

Final Recommendation

After analyzing the defaulters behavior and loan performance, we observe several key patterns that can influence defaults. While, this is just the tip of the iceberg we need to perform more extensive research and analysis to provide concrete action items. But, nevertheless this is a good start and below are the summarized insights we can start looking into:

1. **Income Verification -** Borrowers, whose income is not verified either fully pay their loans or default, while those with verified incomes tend to default less. Therefore, strengthening the income verification process could help reduce defaults.
2. **Term Duration and Interest Rates -** A major factor for defaults is the loan term length. 43% of defaulters have a 60-month loan term, and these longer terms are associated with higher interest rates. Borrowers with lower incomes (median of $50,000) are hit hardest by these conditions.
   The correlation between term, loan amount, and interest rates suggests that higher loan amounts with long terms can push interest rates upward, increasing defaults.
3. **Geographical considerations -** Borrowers in Alaska stand out due to the high interest rates they are charged, leading to higher default rates.
4. **Customer Grading System Reassessment –** Surprisingly, borrowers with lower customer grades tend to default less, which is counterintuitive.

Thank You

GitHub link to access all materials