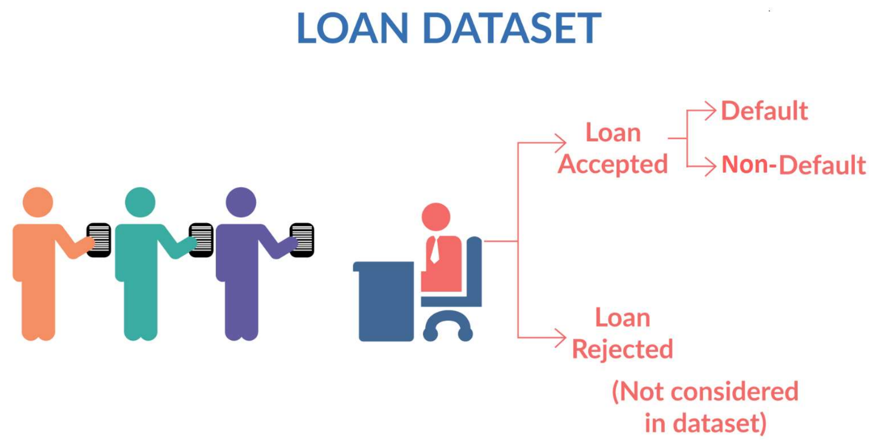
Lending Club Case Study

Ashish Kumar Sahoo

Ravichandra Kolluru

Problem Statement

Using historic loan data to determine the factors contributing to the Charged off for deciding future applicant’s risk for Charge off



Lending Club Case Study Approach

- ❖ Data Analysis – Using the data dictionary and data file understand the data
- ❖ Data Cleansing – Drop irrelevant columns, drop rows not applicable, clear outliers
- ❖ Data Transformation – Transform the data to create variables that can be of use
- ❖ Univariate Analysis – Analyze the key attributes for analysis
- ❖ Bivariate Analysis – Analyze the risk attributes in combination of different attributes
- ❖ Case Study Summary – Identify risk factors and summarize the same

Libraries used

Pandas

Numpy

Matplotlib

Seaborn

Plotly

Data Analysis

- Reviewed data dictionary and identified loan_status is the driver
- There are 111 columns in the data
- 39717 records for 39717 unique members
- The data in the file is for loans granted between 2007 and 2011
- There are more attributes that may not be of relevance for analysis based on the description but will be decided based on further analysis
- Roughly 83% of loans are fully paid, 14% charged off, 3% loans are current

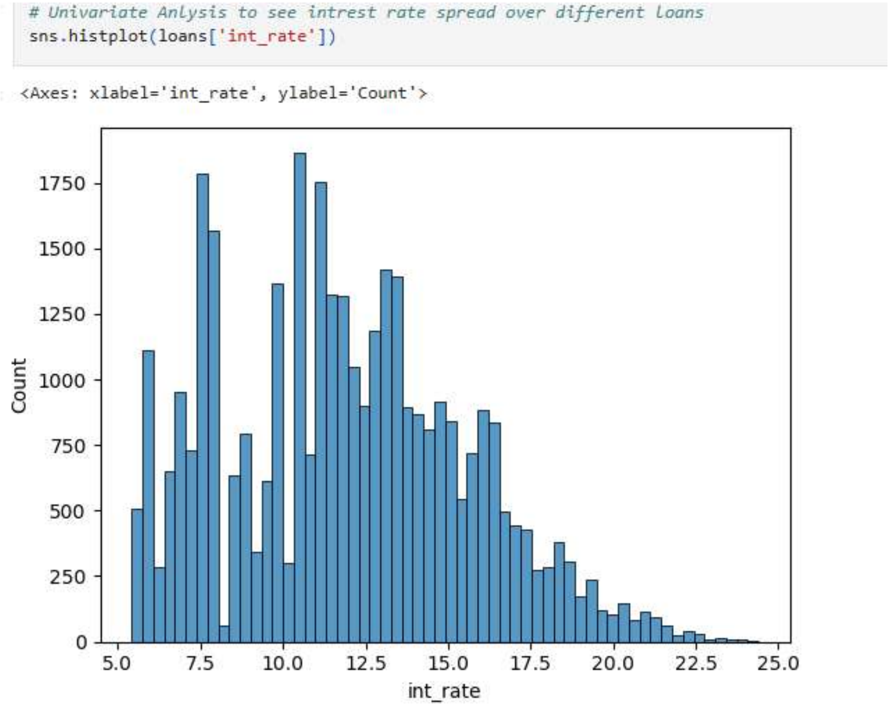
Data Cleanse

- Identified columns that have 100% nulls or have 0 or 1 unique values. These columns will be of no use with analysis and thus dropped
- Dropped columns with Null percentage more than 30%
- Dropped data with columns having unique values (i.e. each row has different unique value)
- Removed columns that are of no relevance based on descriptions (e.g. title, emp_titile, delinq_2yrs, earliest_cr_line,….)
- Total 18 base columns are used for final analysis
- Dropped rows with loan_status as Current as they can’t contribute to the risk analysis
- Duplicate Rows are dropped (no duplicates found)
- Remove outliers in the data

Data Transformation

- Changed data types for columns like loan_amt, funded_amt to floats
- Removed percentage from int_rate and revol_util columns and converted to floats
- Filled with zero values for revol_util is na (not available)
- Date column issue_dt is converted to dates, created year and month column for Issue Date
- Experience columns is standardized
- Defined functions CountPlot and PercentageBarPlot to analyze different attributes easily

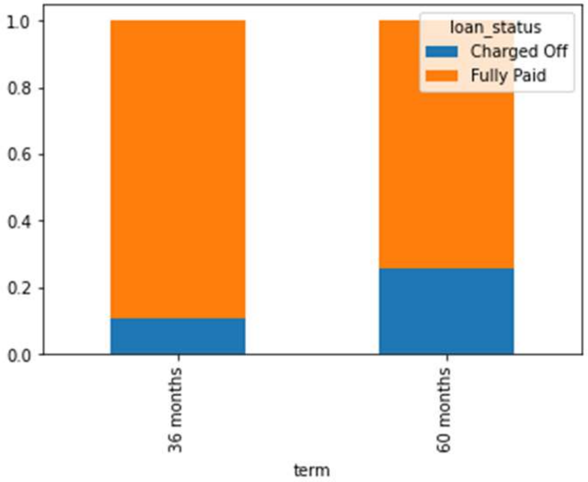
Univariate Analysis

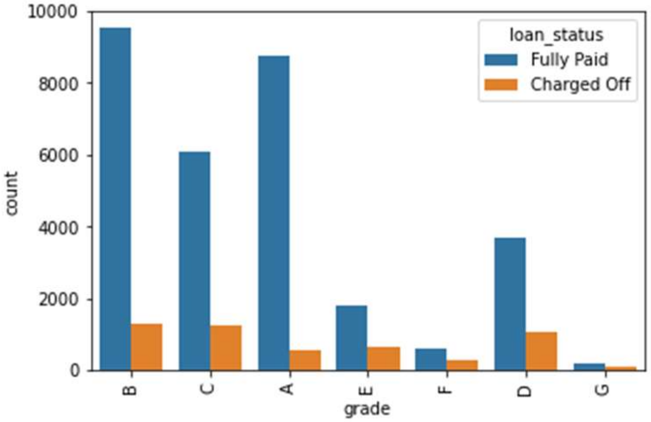


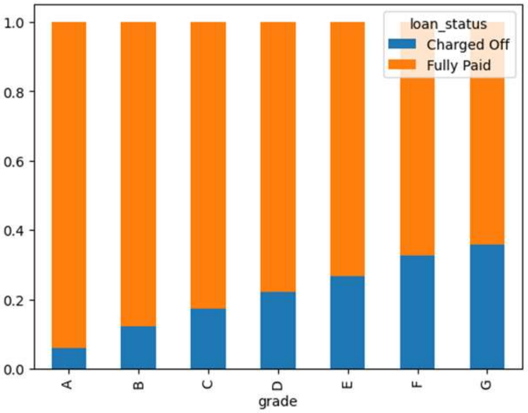
```
# Univariate Anlysis to see intrest rate spread over different loans
sns.histplot(loans['int_rate'])
```

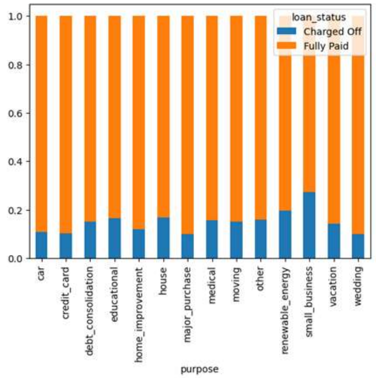
```
<Axes: xlabel='int_rate', ylabel='Count'>
```

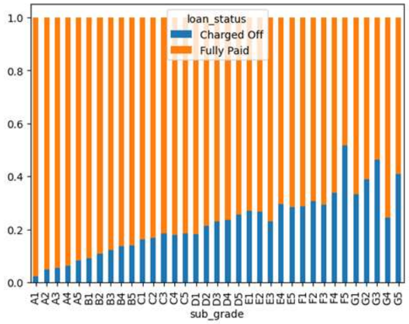
Bivariate Analysis – Visual Analysis

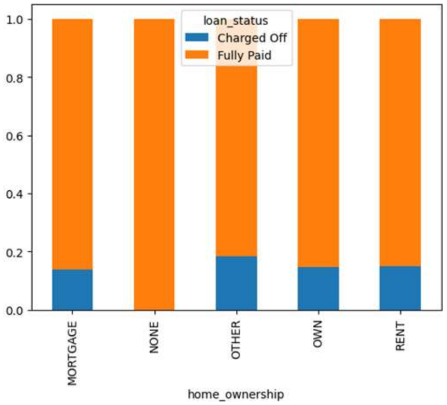




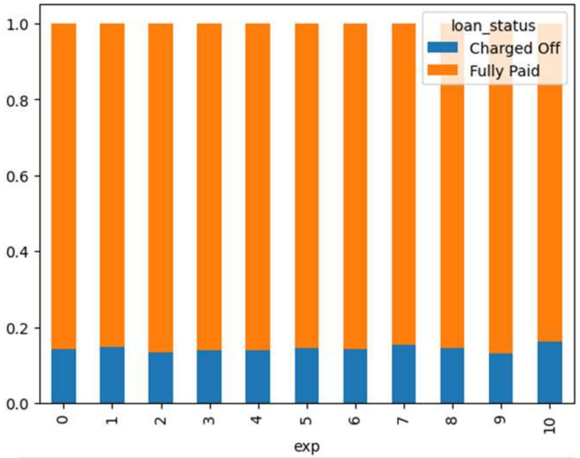


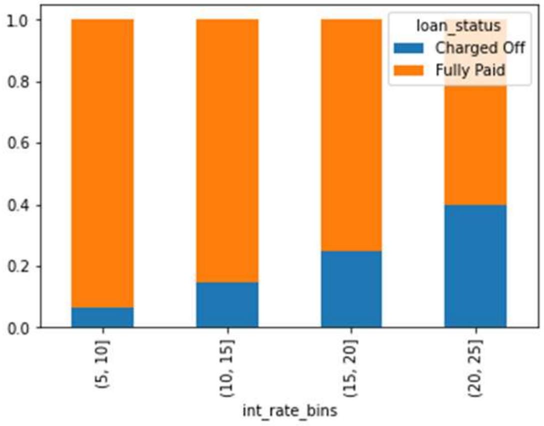


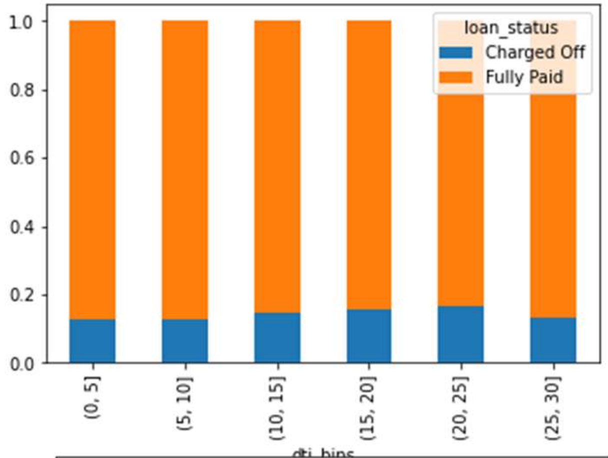


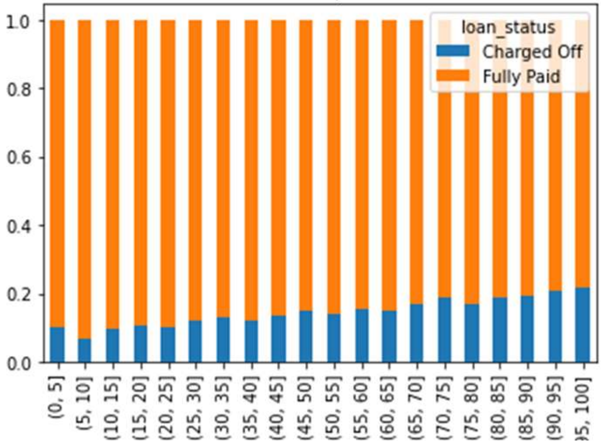


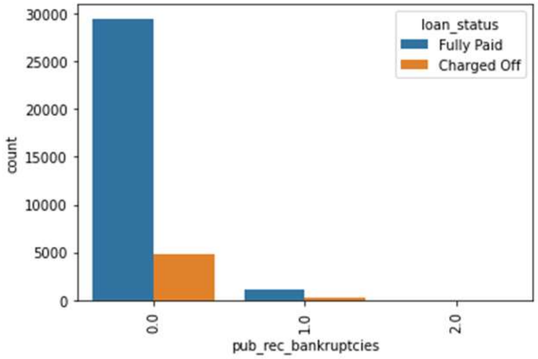
Bivariate Analysis – Visual Analysis

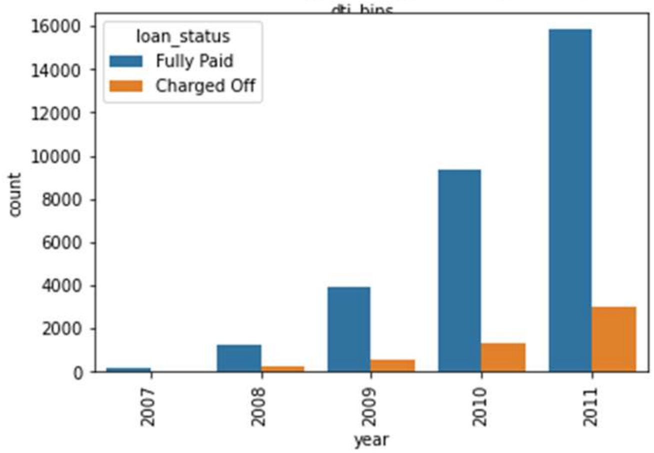












Case Study Summary

Below attributes are identified as key risk factors for loan

- Small Business seem to have highest risk of Charge off
- Higher the revolving credit more change of Charge Off
- Grade/SubGrade has higher impact with Charge off Loans, but most of the loans are for higher grade customers with low charge off
- The more the percent of Revolving Credit used the risk for charge off
- Increase in interest rate have adverse impact on Loan Status
- Most of the Small Business Loans have higher interest rates

Case Study Summary

Below Attributes did not seem to have considerable impact on the loan status

- Experience has no impact on the Charge Off vs Fully Paid
- Verification Status has no impact on the status
- DTI does not have an impact on the status

Case Study Summary

Observations

- Majority of Attributes have no data and removed
- May attributes with Unique or 1 value are removed from Analysis
- Some variables with no much impact are removed
- No correlation found between different continuous variables

Thank you