Lending Club Case Study

Explorative Data Analysis (EDA)

Agenda

1. Business Understanding
2. Business Problem Statement
3. Data Understanding & Assumptions
4. Data Quality Issues
5. Data Cleaning
6. Analysis & Insights
7. Key Indicators of Default
8. Recommendations

Business Understanding

- Lending Club or LC is a fintech marketplace bank operating out of San Francisco, California since 2007 .
- LC offers its members various types of loans to meet their financial needs
- It gets the funding from investors by offering a return for funding these loans
- The process of granting loan involves the following:
  - The borrower applies for a certain amount of loan to LC
  - LC reviews the application and may or may not approve the loan application
  - If approved, LC finds out if any investors are interested in funding the loan
  - If investors are interested, they fund the loan amount (funded amount could be less than the applied amount)
- In order to analyze the data, it is important to understand as to what practices are followed in approving a loan application from a domain perspective

Business Understanding – Lender's Perspective

- Factor lenders consider for loan eligibility (source: www.lendingclub.com):
  - Credit Score
  - Payment History
  - Income
  - Current debt
- Factors That Influence Credit Score (source: www.lendingclub.com):
  - payment history
  - Age and type of credit
  - Percentage of credit limit used (revolving utilization)
  - Total balances or debt
  - Recent credit behavior and inquiries
  - Available credit
- This information will provide guidance as to what data should be considered for analysis and what data is not relevant.

Business Problem Statement

- The key to success and growth of LC depends on its ability to minimize its risk in lending, i.e.
  - Identify risky borrowers to avoid bad credit, thereby preventing financial loss
  - Identify borrowers who are likely to fully pay back, thereby not missing the opportunity to make financial gain for itself & its investors
- It is required to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default, so that a suitable decision can be made when a new applicant (member) applies for loan
- The objective of the assignment is to
  - Analyze the data and identify key influencers behind default
  - Provide recommendations to avoid bad credit & still not miss good investment opportunities

Data Understanding & Assumptions

- Dataset has been provided for loans issued between 2007 and 2011 in CSV format along with a Data Dictionary in Excel file
- The dataset has information on 39717 loans with 111 attributes
- The data contains information about the loans which were approved, with the status
  - Fully Paid - applicant has fully paid the loan (the principal and the interest)
  - Charged Off - Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
  - Current - Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'
- The data does not contain the rejected applications. Rejected loan applications are not relevant to the objective of this analysis as it is not known whether they would have defaulted or not.
- By the same logic the loans with "Current" loan status are also not relevant to the objective of this analysis as it is not known whether they will default or not. Accordingly, the data for loans where loan status="Current", has been excluded from the analysis.

Data Understanding & Assumptions

- The data primarily contains the following information
  - Data Captured from the Loan Application
    - Borrower's personal information e.g., annual income, employer, length of employment etc.
    - Loan requirements e.g., loan amount, term, purpose etc.
  - Data Captured from applicant's Credit Reports (assumption)
    - Borrower's Credit history e.g. , % of credit limit utilized, credit amount that goes unpaid from month to month, public records of defaults, past bankruptcies filed etc.
  - Data Captured from LC database
    - Information mostly about the existing loan accounts e.g. last payment date, last payment amount, next payment date etc.
- To understand the meaning of the Variables, LC's website (www.lendingclub.com) and other public websites (www.balance.com, www.investopedia.com etc.) were referred. The meaning of the variables as understood have been updated on the Data Dictionary.

Data Quality Issues

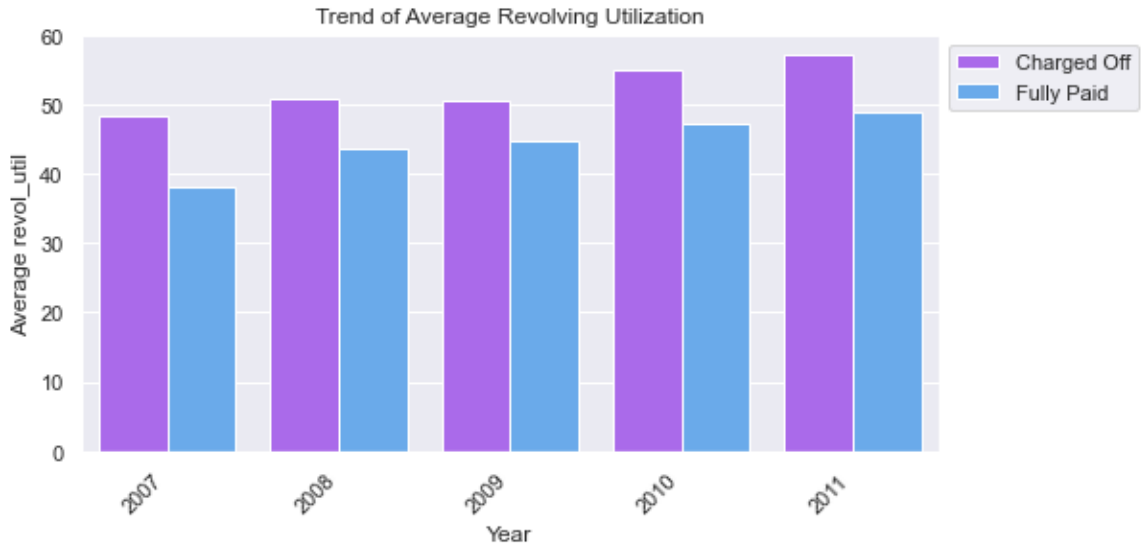
- Data Quality Issues
  - 55 nos. out of 111 (~50%) columns had no values - LC needs to check the relevance of these attributes and reasons for not capturing the same
  - Redundant information in dataset e.g., columns “title” and “desc” are expected to capture the same information as “purpose”,
  - Interest rate & Revolving line utilization rate, which should be decimal numbers, are suffixed with “%”character and stored as string
  - The number of payments on the loan (term), whose value can be 36 or 60, is suffixed with “months” and stored as string
  - Data of “Current” loans are not relevant to this analysis, but present in the dataset
- There are extreme values (e.g. outliers in case of annual income) present in the data, but they seem to be genuine data
- Data Quality Good Aspects
  - There were no duplicate data, neither in rows nor in columns
  - After removing unnecessary columns, at least 95% data was filled-in in all rows

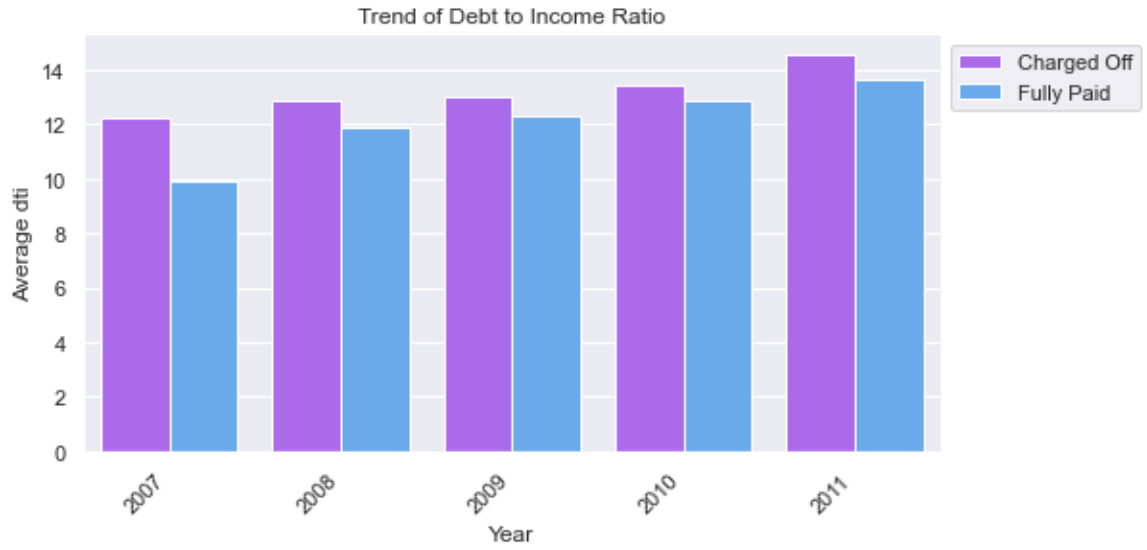
Data Cleaning

- Columns needed to be Removed from dataset
  - 55 nos. columns with no data
  - 22 nos. columns, either redundant or not relevant based on business understanding
  - 2 nos. columns having more than 25% missing values
  - 5 nos. columns having constant values
  - 3 nos. columns with quasi-constant values
- Data Manipulated / Transformed
  - Removal of prefix, suffix e.g. “<”, “year”, “years”, “%”, “ months”, “xx” etc. and changing datatypes for columns where number is expected
  - Converted columns to appropriate date format as needed
- Value Imputation
  - Missing values have been imputed with mean/median/mode value after checking that the imputation does not significantly alter the data distribution
  - Value imputation was not done for one column as imputing values could have altered the data distribution in a significant way

Analysis & Insights

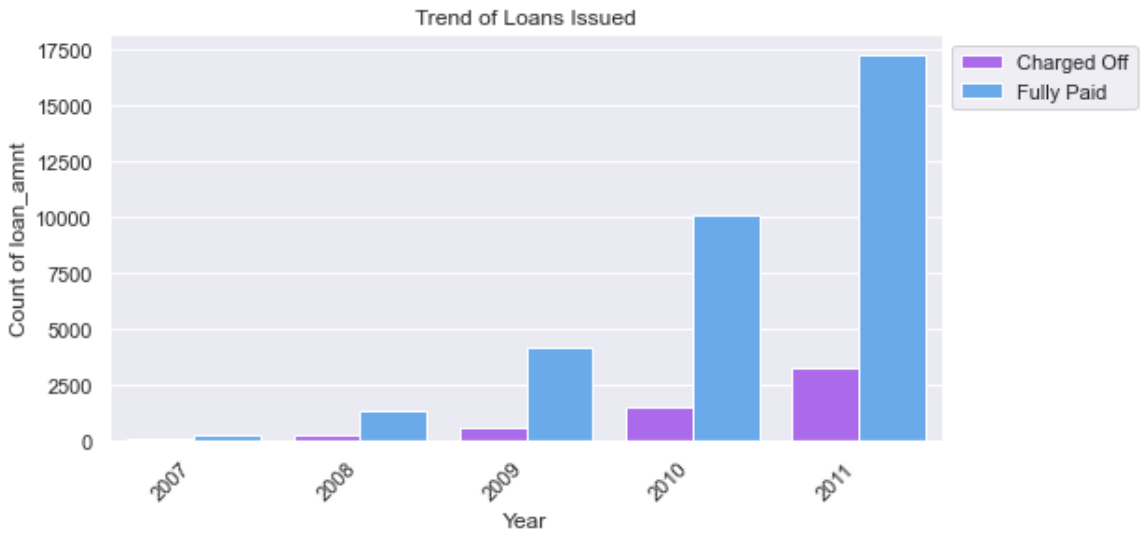
- Trend Analysis - Borrowing Habits
  - Borrowing tendency (Revolving Utilization or amount borrowed in % of credit limit) has increased by approx. 10% (may be due to inflation), that means more debt
  - Average annual income of borrower has not changed significantly (chart not included in slide)
  - Increase in Debt-to-Income Ratio as seen in the slide can be explained based on the above
  - Borrowers, who have defaulted, have higher average revolving utilization and higher average debt-to-income ratio

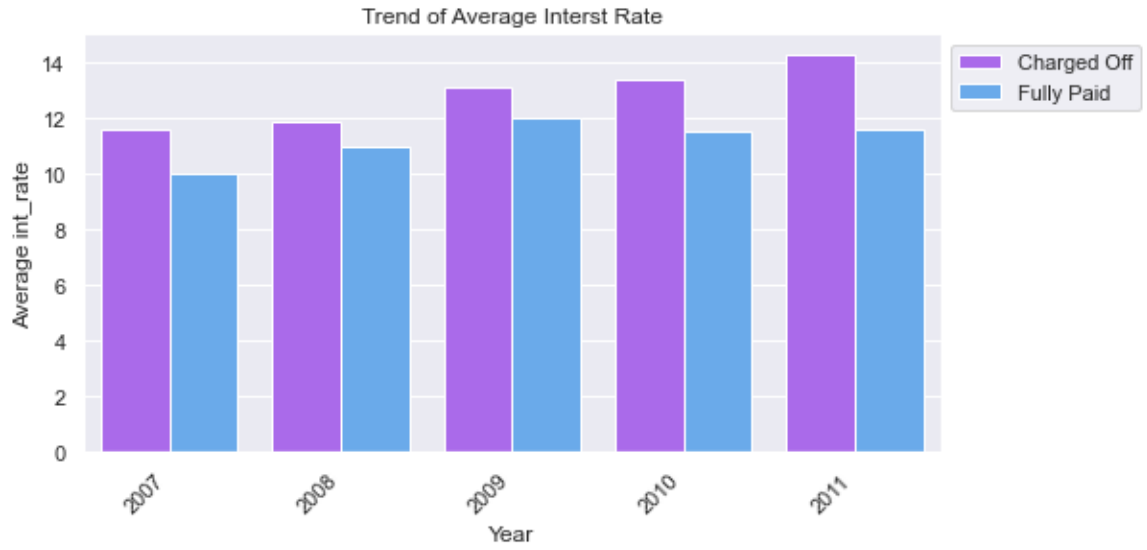




Analysis & Insights

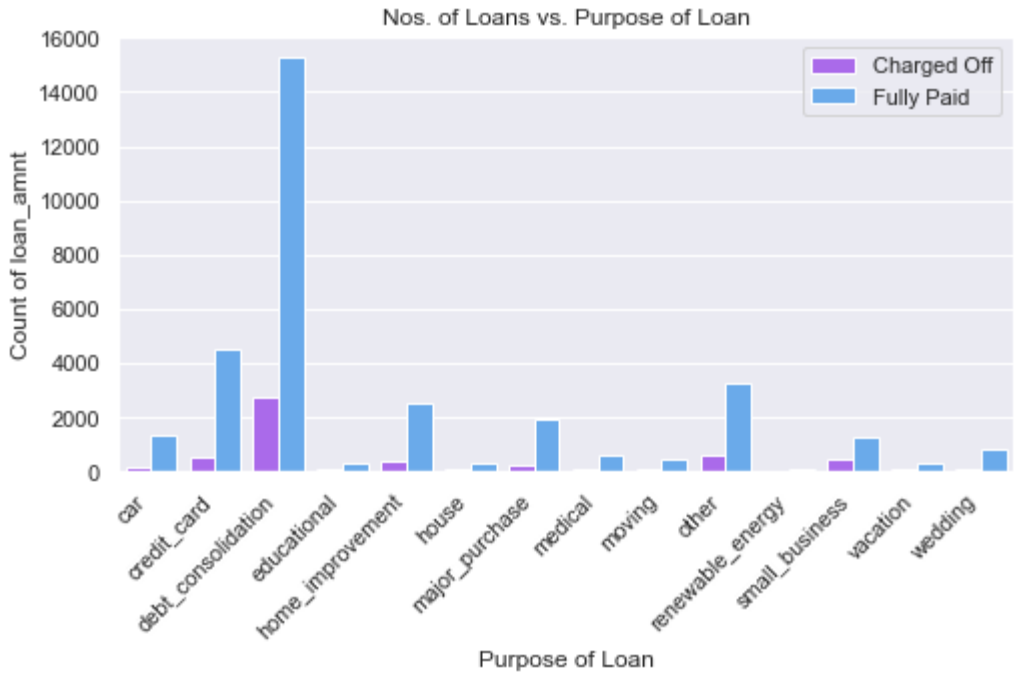
- Trend Analysis - Demand for Loans
  - Number of Loans issued have grown exponentially
  - Average Interest Rate has grown by approx. 2%
  - Demand of Loans has increased despite increase in Interest Rate
  - Increase in demand of loans can be explained in terms of the change in borrowing habits (previous slide)
  - Borrowers who are defaulting have (or may be because of) higher average Interest Rates

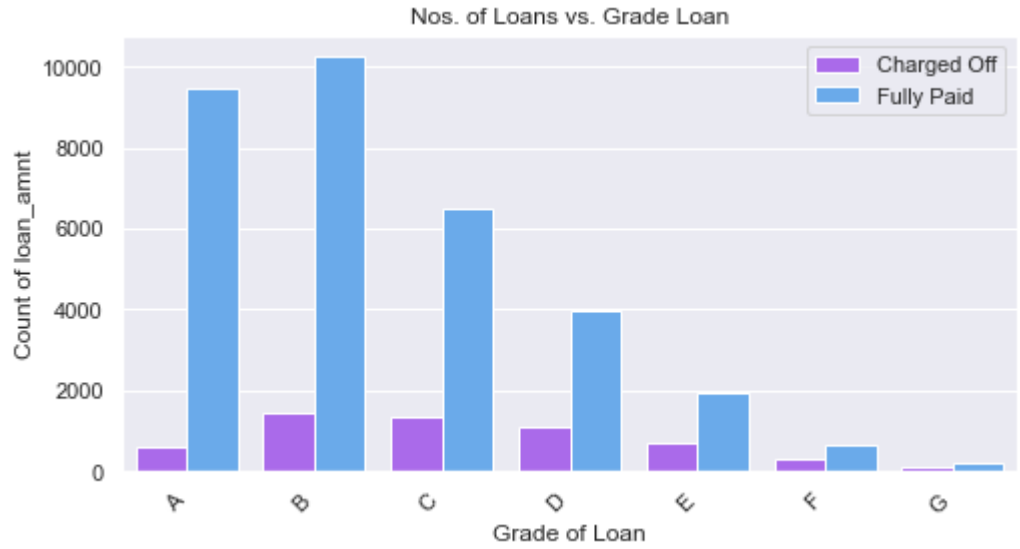




Analysis & Insights

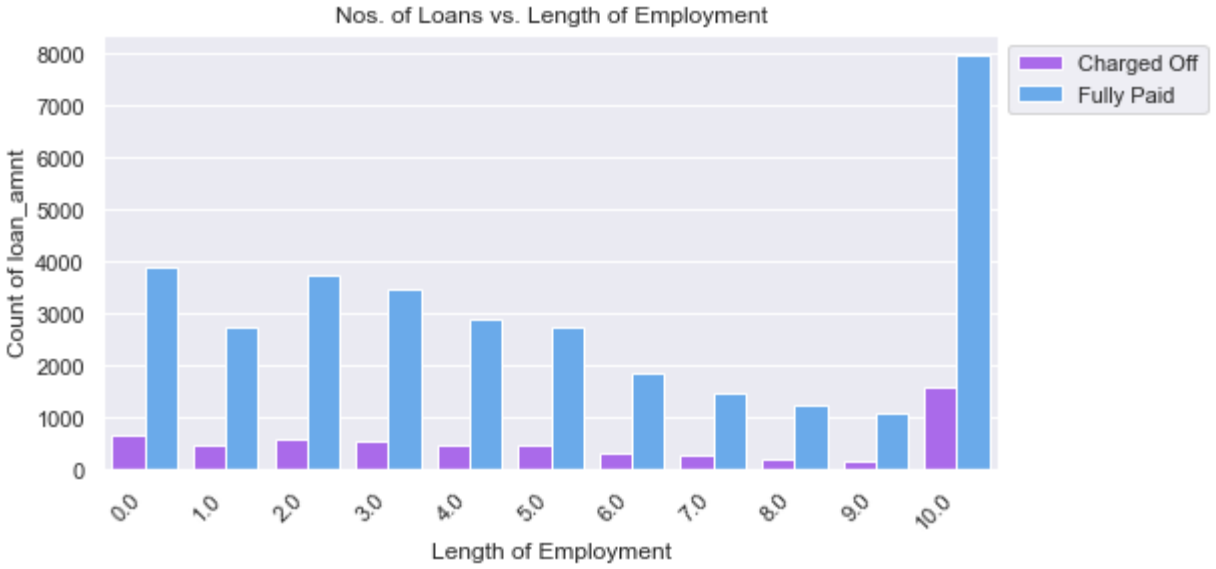
- ## Category Wise Analysis
  - Very Large nos. (approx. 30%) loans are taken for Debt Consolidation
  - Followed by Credit card Payment and Other Miscellaneous Purpose
  - Nos. of loans granted in Grades E, F & G are less, probably these are high risk loans
  - Lending risk appears to be in increasing order from Grade B to G

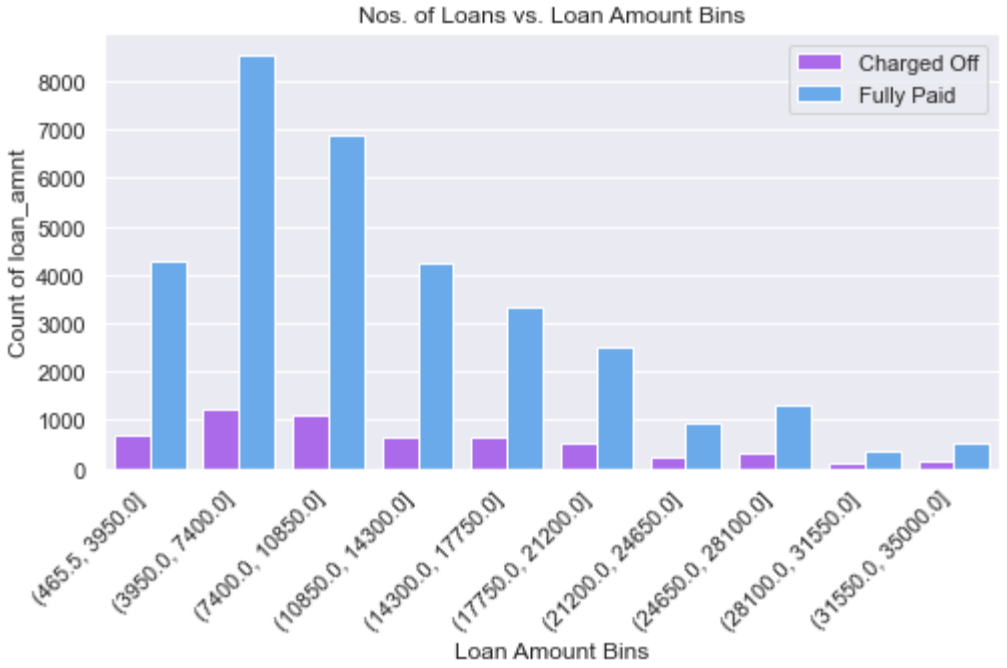




Analysis & Insights

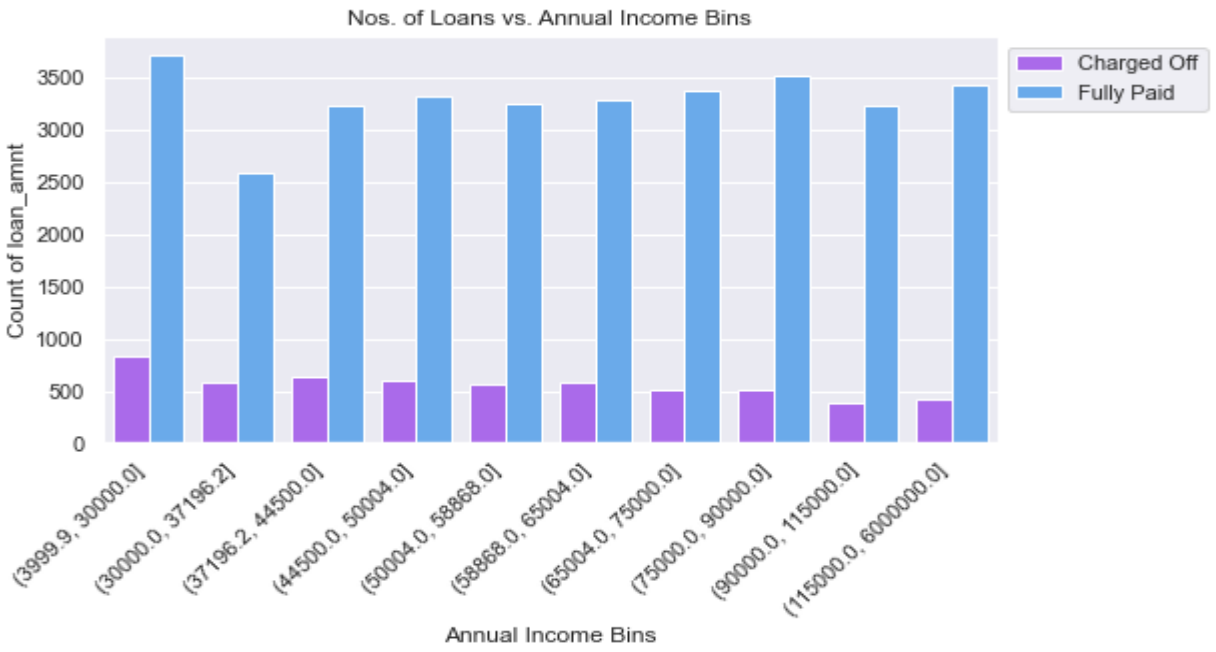
- Category Wise Analysis
  - Borrowers with employment length 10 or more years are the biggest chunk (approx. 25%)
  - It may be because 10+ are clubbed together, so the total count is bigger compared to others
  - Approx. 50% loans are under $10,000

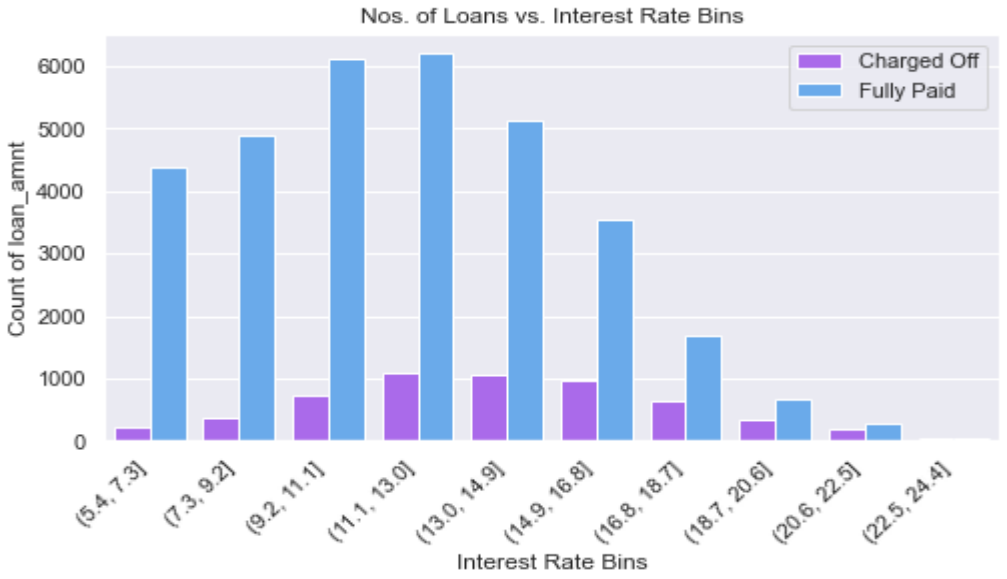




Analysis & Insights

- Category Wise Analysis
  - Very low income (<= $30,000) borrower's take the highest nos. of loans
  - Nos. of defaults is also highest for very low income (<= $44,000) borrower's and low at the other end of the spectrum
  - Highest nos. of loans have been granted with interest rates between 11.1% & 13%
  - Very few (< 2000) loans have interest rates 20% or more

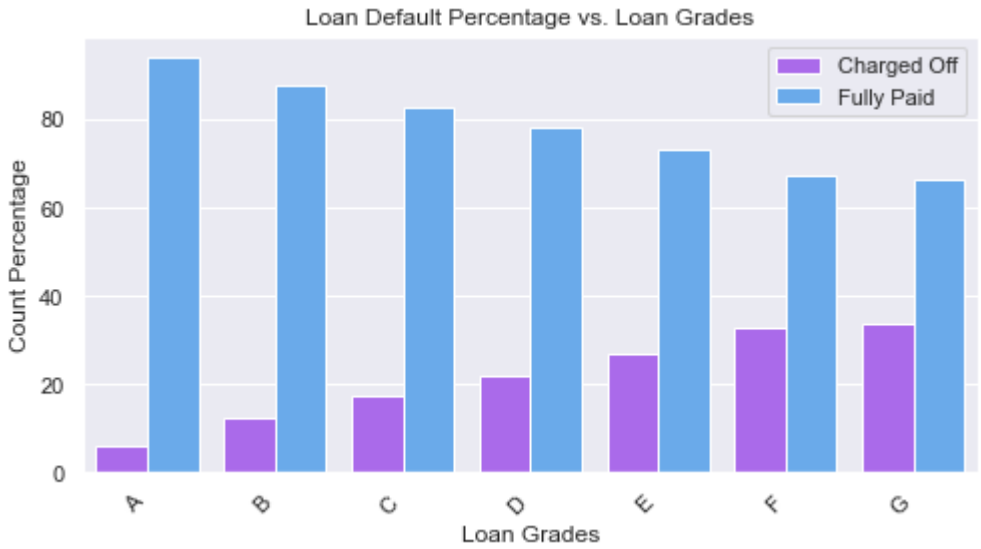




Analysis & Insights

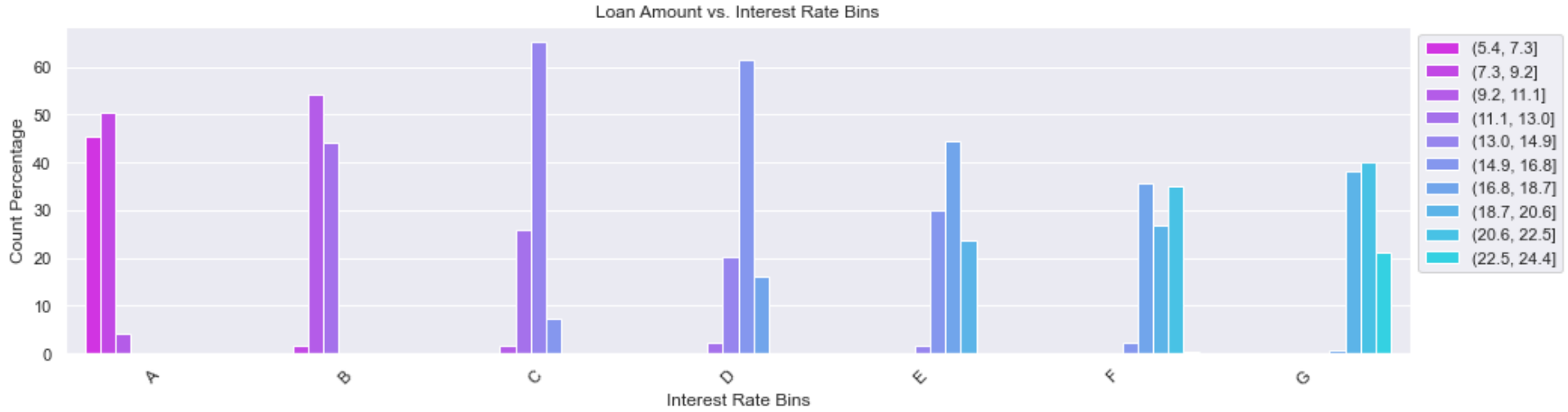
- Category Wise Analysis
  - Rate of Default has an upward Trend from Grade A to G, i.e. lowest in Grade A and highest in Grade G
  - Clearly indicates the Grade A to G indicates low risk to high risk
  - Interest rates are in low range in Grade A loans and are in the highest band for Grade G loans (next slide)
  - Number of low income borrowers are very less in Grade E F & G (next slide)
  - Grade E, F & G have very high income borrowers with high loan amount with high interest rates (next slide)
  - Inference: Grade E, F & G are High Risk, High Return credits to High Income Borrowers

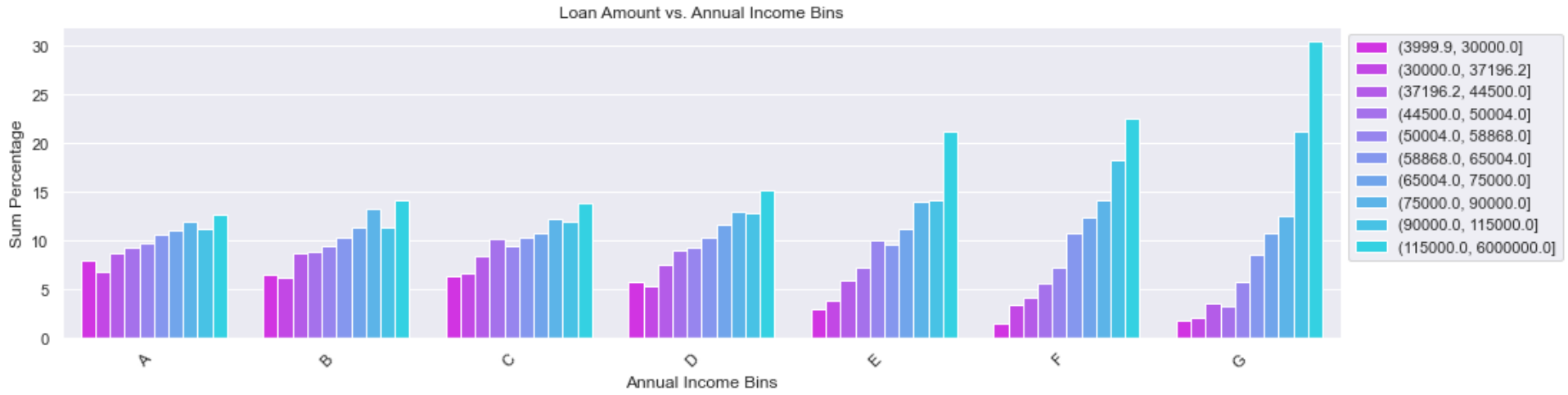
Charts for Interest Rates & Annual Income vs. Grade are in the next Slide



Analysis & Insights

- Charts for Interest Rates & Annual Income vs. Loan Amount by Grade

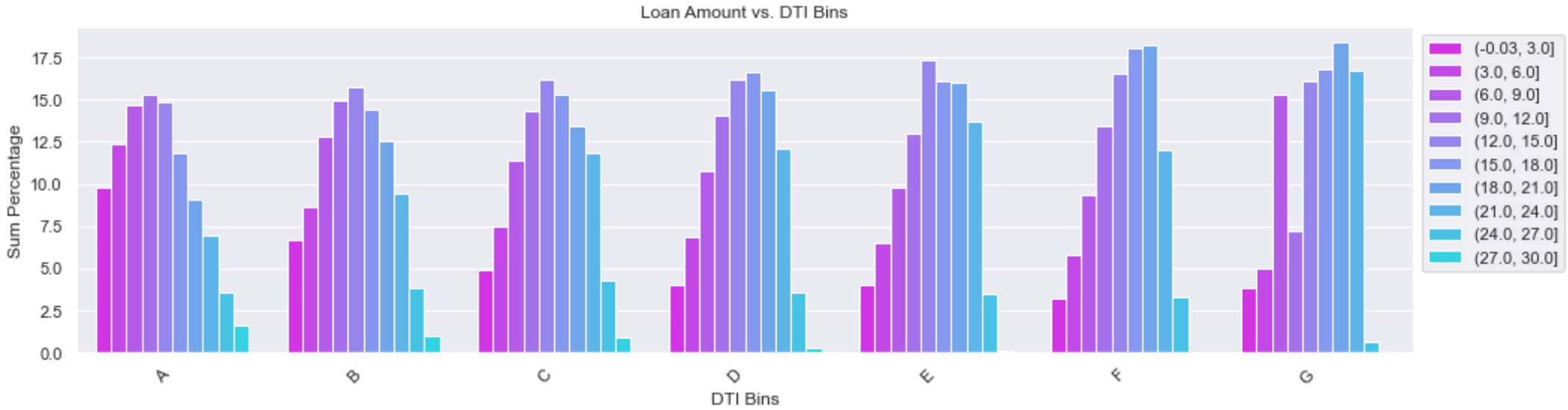




Analysis & Insights

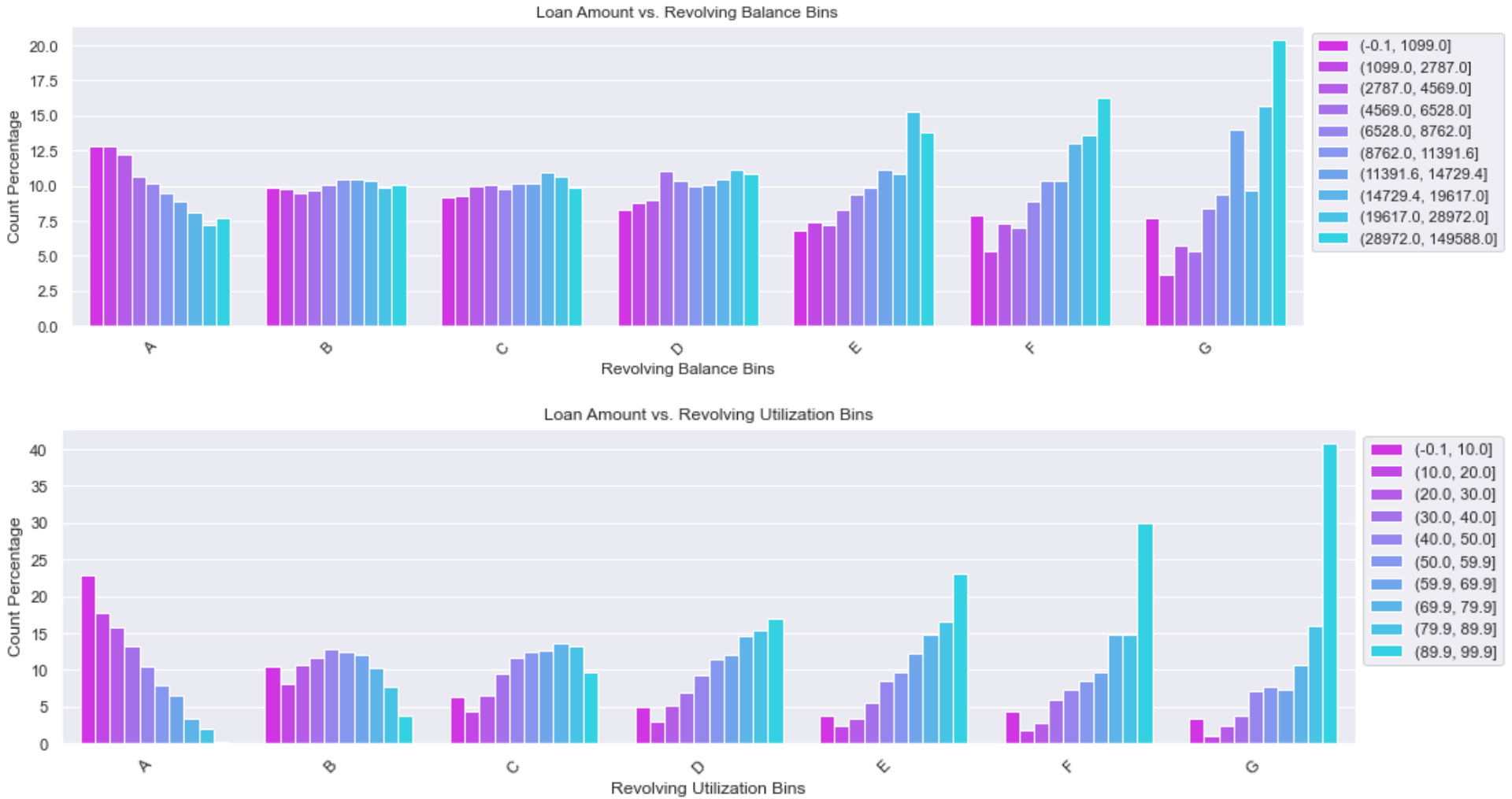
- ## Category Wise Analysis
  - The Debt-to-Income ratio for Grade E, F & G is little higher than Grade A, B & C, high DTI means high risk
  - Check on DTI threshold for loan eligibility may reduce some risk for Grade E, F & G loans
  - Revolving Balance (part of the credit that goes unpaid from month to month) is high for Grade E, F & G
  - Check on Revolving Balance threshold for loan eligibility may reduce some risk for Grade E, F & G loans
  - Revolving Utilization (part of the credit in % that is utilized by borrower) is high for Grade E, F & G
  - Check on Revolving Utilization threshold for loan eligibility may reduce some risk for Grade E, F & G loans

Charts for Revolving balance & Revolving Utilization are in the next slide



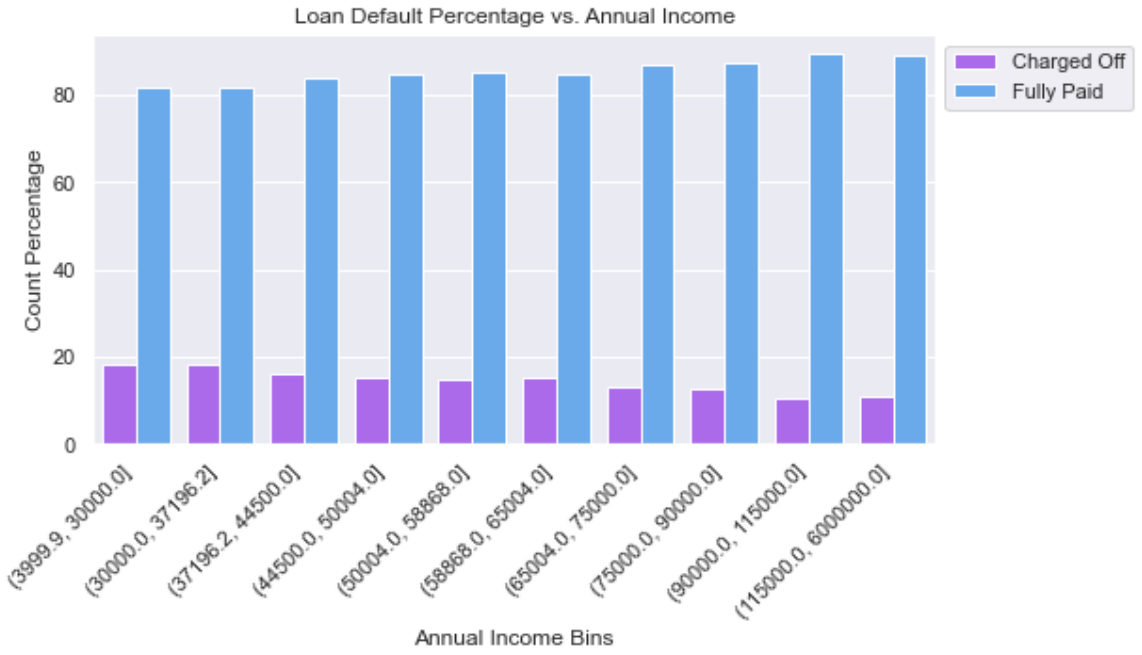
Analysis & Insights

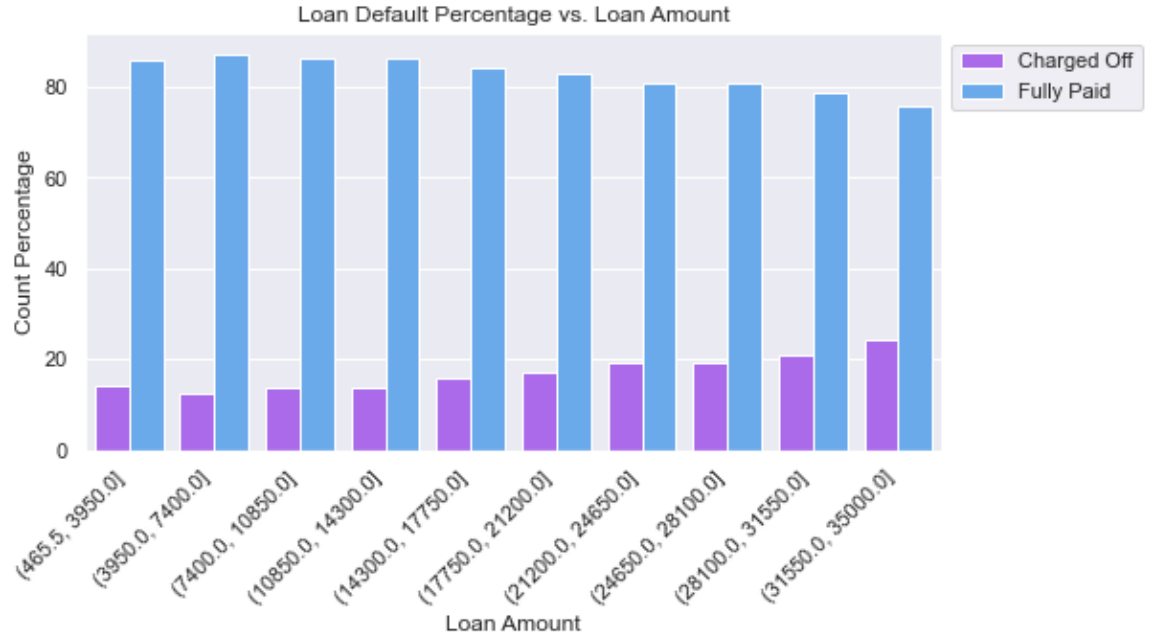
- Charts for Revolving Balance & Revolving Utilization by Grade



Analysis & Insights

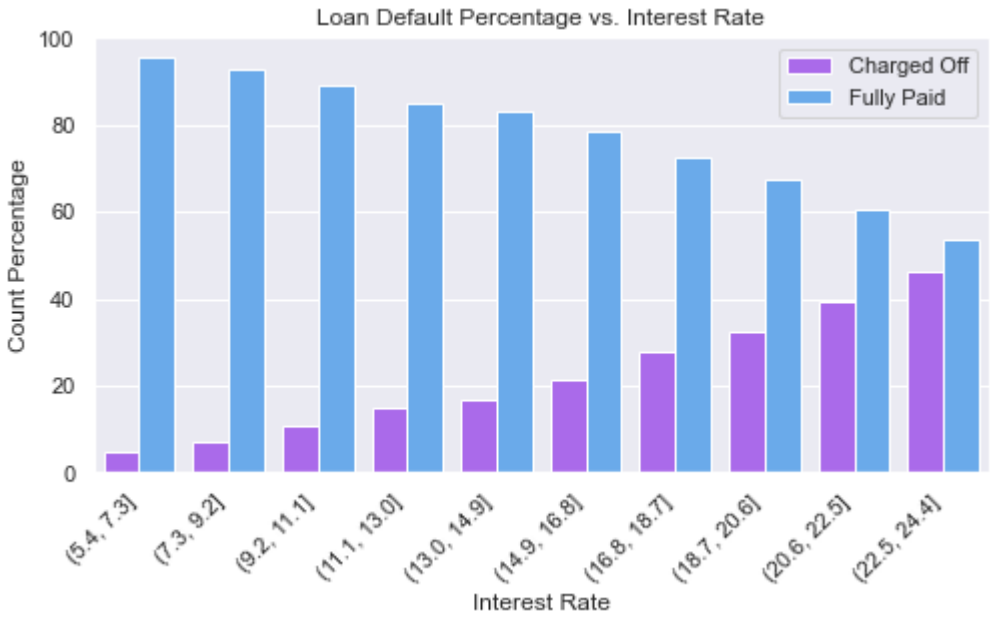
- Loan Default Analysis – Key Indicators
  - Higher chances of loan default if the borrower's annual income is low
  - Higher chances of loan default if the loan amount is high

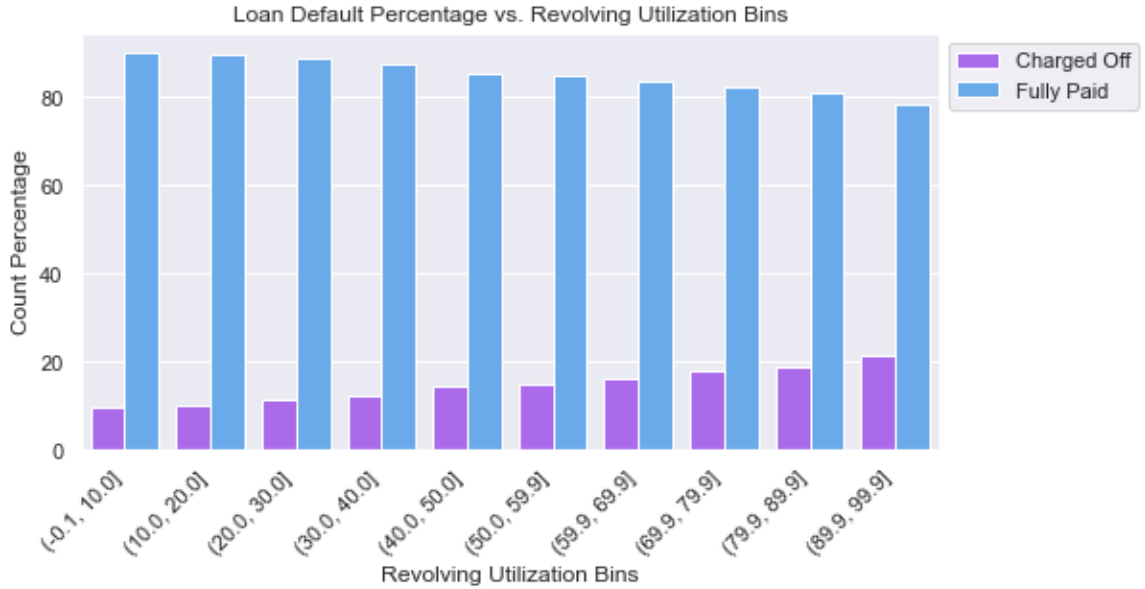




Analysis & Insights

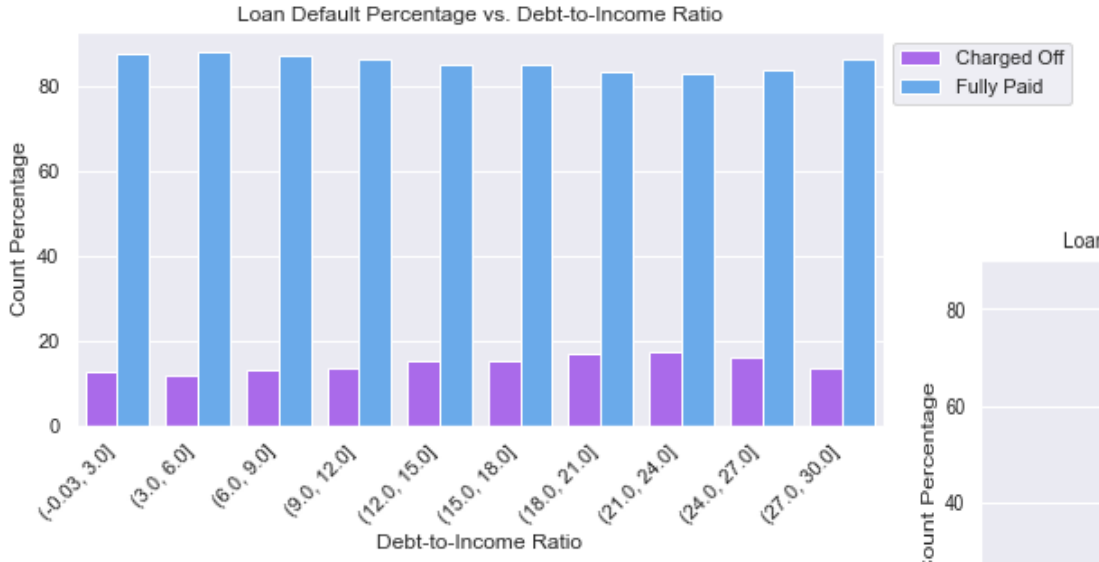
- Loan Default Analysis – Key Indicators
  - Higher chances of loan default if the is interest rate is high
  - Higher chances of loan default if the borrower's revolving utilization is high

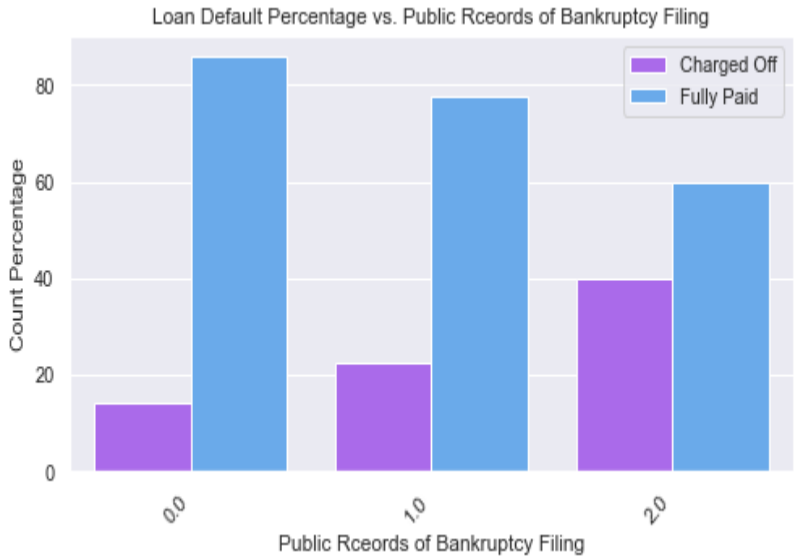


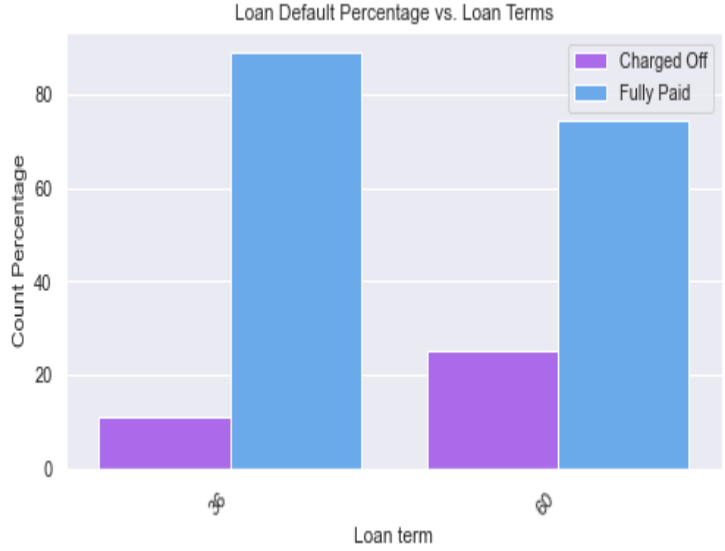


Analysis & Insights

- Loan Default Analysis – Key Indicators
  - Higher chances of loan default if the borrower has high Debt-to-Income Ratio
  - Higher chances of loan default if the borrower has history of filing bankruptcy
  - Higher chances of loan default for loans with longer payback period







Key Indicators of Default – Any of the Following

- Loan Requirements

Purpose is Small Business

High Loan Amount > $21,000

Longer Term (5 Years)

High Interest Rate > 15%

- Borrower Profile

Low Annual Income < $40,000

Credit Limit Utilization > 60%

History of filing bankruptcy

High DTI > 18

- Others

Grade G/ F/ E Loans

Recommendations

- Cap the loan amount to low income borrowers to reduce loss on account of default
- Avoid 5 year term, if possible, they have the worse default rate
- Avoid Small Business Loans, they have the worst default rate against loans for any other purpose
- Review the process of granting loans in Grade E, F & G, they have high default rate
- Exercise caution while granting loan amount above USD 21,000, default rate is high
- Avoid borrowers with past bankruptcy filing, chances of default are more
- Review cases where interest rates is above 15%, they tend to default more
- Keep max limit for DTI at 18 (or less) and Revolving Utilization at 60% (or less) , especially for Grade E, F & G loans to reduce risks of default
- Review the income verification process – the data does not show that the default rate is lower when income / source is verified compared to not-verified
- Encourage Car, Vacation & Medical Loans to very high income borrowers (annual income > USD 180,000), there are no defaults
- Suggested to consider applicant's Credit Score and set minimum score for eligibility. Credit Score is obtained separately and not available on the Credit Report

Thank You !