Lending Club Case Study

Manish Kumar Prajapati

Manjunath B Nagarajaiah

Problem statement / Benefits of the Case Study

Practical EDA Application: Learn how EDA is applied to real-life business challenges.

Risk Analytics Understanding: Gain a foundational knowledge of risk analytics in banking and finance.

Financial Decision-Making: Discover how data helps minimize financial losses when lending to clients.

Improved Visualization Skills: Enhance your ability to choose and create effective visualizations for real-world data.

Objective

The goal of this case study is to use Exploratory Data Analysis (EDA) to tackle a real-world problem, uncover valuable insights, and present them in a way that's easy for business stakeholders to understand.

Business Understanding:

The business goal is to make informed decisions on loan application whether to approve or reject the based on specific variables.

Dataset Details

- Content: The dataset contains information about past loan applicants and indicates whether they defaulted on their loans.
- Scope: The data includes details about approved loans, not rejected ones.
- Loan Statuses: The loans have three statuses: Fully Paid, Current, and Charged-Off.

Approach

Data Cleaning

- There were no header, footers, summary or Total rows found.
- There were 57 columns which is having more than 50% rows values as null/blank and doesn’t participate in analyse has been removed.
- We excluded 21 columns of behavioral data from the analysis, as this information is captured but will not be available during the loan approval process
- 8 columns whose values were 1, and is uniqueness in nature has been dropped from analysis.
- Deleted all the columns which value is unique in nature.
- Deleted 'member_id', and 'url' as it doesn't count in EDA. Will keep the id columns as represent the row data unique for further analysis
- Deleted descriptive or textual informations columns as doesn't participate in EDA analysis.
- Removed funded_amnt_inv as it is a internal data and is calculated after loan approval thus cannot be used as input for the
- Removed zip_code as it is a masked data and cannot be used as input for the analysis
- After all the Data cleaning process we are left with 39717 rows and 18 columns.

Data Enrichment

Deleting and fixing the null values

- Removed the null values from the emp_length and pub_rec_bankruptcies column
- Filtered the completed and defaulted loan entries as ‘current’ does not participate in analysis.
- There were no duplicates rows found.

Data Enrichment

Correcting Data Types and Deriving New Columns

- Converted term column from string to int. Additional ‘months’ has been trimmed.
- ‘int_rate’ has been converted from string to flot. Additional ‘%’ has been trimmed.
- issue_d has been converted to datatype.
- Creating a derived columns for ‘issue_year’ and ‘issue_month ’ from ‘issue_d’ for analysis.
- Clean 'emp_length' column and convert to numeric.

Removing the outliers

- Outliers exits for numeric data 'loan_amnt','int_rate', ‘annual_inc'.
- Outliers treatment has been done for above fields using quantile mechanism.

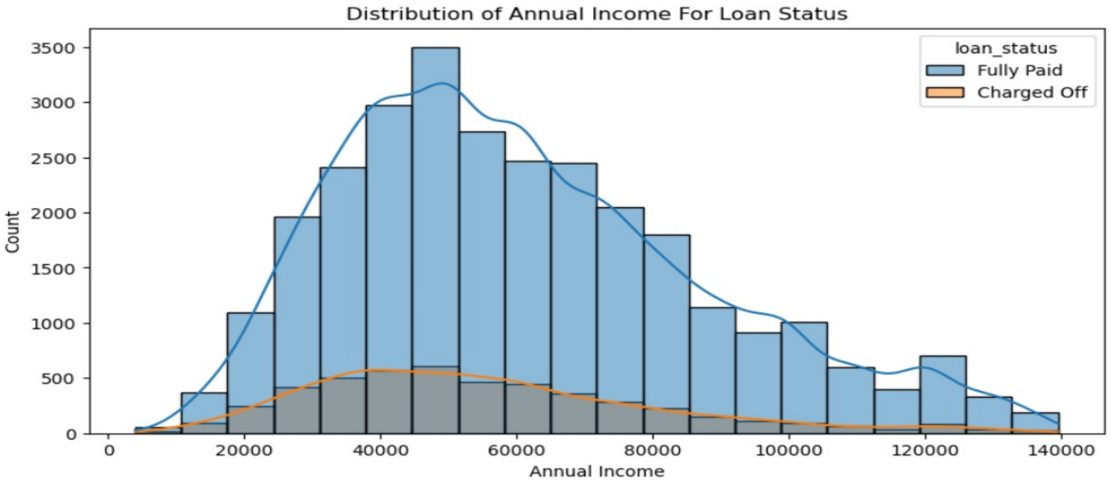
Data Analysis – Univariate

- Loan Status : Defaulted loan are low in numbers compared to Fully Paid.
- Loan Amount: Most of the loan amount applied was in the range of 5k-14k
- Term: More than half of the loan taken has term of 36 months compared to 60 months.
- Interest Rate: The interest rate is more crowded around 5-10 and 10-15 with a drop near 8-10
- Grade: A large amount of loans are with grade 'A' and 'B' compared to other grade (C, D, E, F, G)
- Employment Duration: Majority of borrowers have working experience greater than 10 years
- Home Ownership: Majority of borrowers don't possess property and are on mortgage or rent.
- Verification Status: About 50% of the borrowers are verified by the company or have source verified.
- Annual Income: Majority of burrowers have very low annual income compared to rest.
- Purpose: A large percentage of loans are taken for debt consolidation followed by credit card.
- Address: Majority of the borrowers are from the large urban cities like California, new York, Texas, Florida
- DTI: Majority of the borrowers have very large debt compared to the income registered, concentrated in the 10-15 DTI ratio.
- Public Record Bankruptcies: Majority of the borrowers have no record of Public Recorded Bankruptcy.
- Month: Majority of the loans are given in last quarter of the year
- Year: The number of loans approved increases with the time at exponential rate, thus we can say that the loan approval rate is increasing with the time

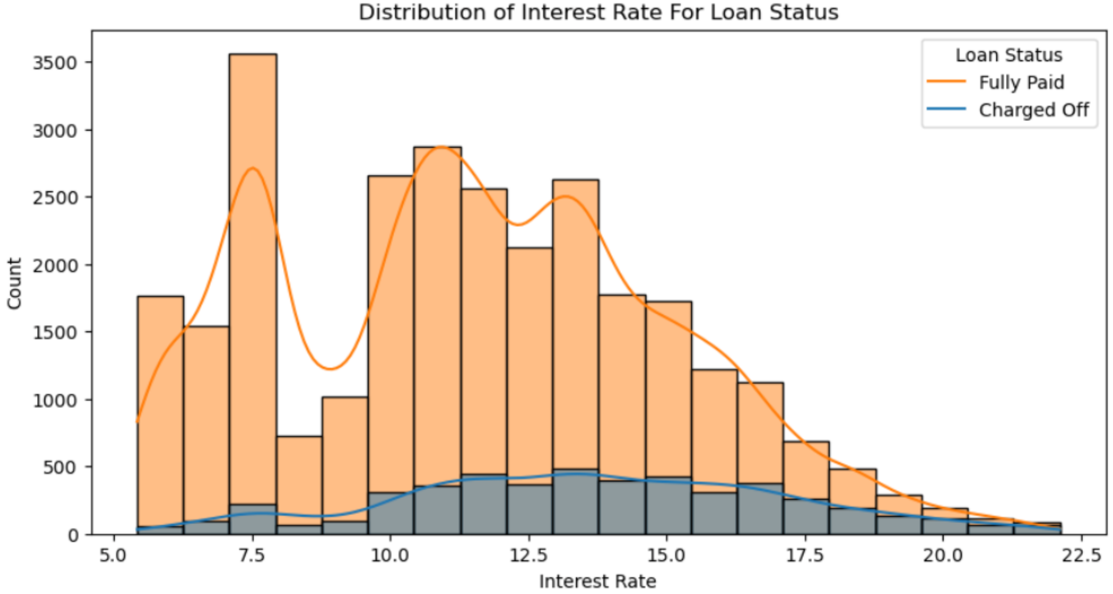
Data Analysis – Segmented Univariate

- Charged off loans have a larger IQR and a higher median, suggesting that defaults are more common among larger loans.
- Debt Consolidation is the most common loan purpose, also defaulter larger in this area.
- The 60 months term has higher chance of defaulting than 36 months term whereas the 36 months term has higher chance of fully paid loan.
- The Loan Status varies with the DTI ratio. We can see that the loans with a DTI ratio in the range of 10-20 have a higher number of defaulted loans. Generally, a higher DTI ratio indicates a higher chance of defaulting.
- Defaulters are most likely from RENT or Mortgage categories
- Analysis shows that defaulter most likely have low income.
- The default loan amount increases with interest rate and shows are decline after 17.5 % interest rate.
- The Employees with 10+ years of experience are likely to more defaulter and also have higher chance of fully paying the loan.

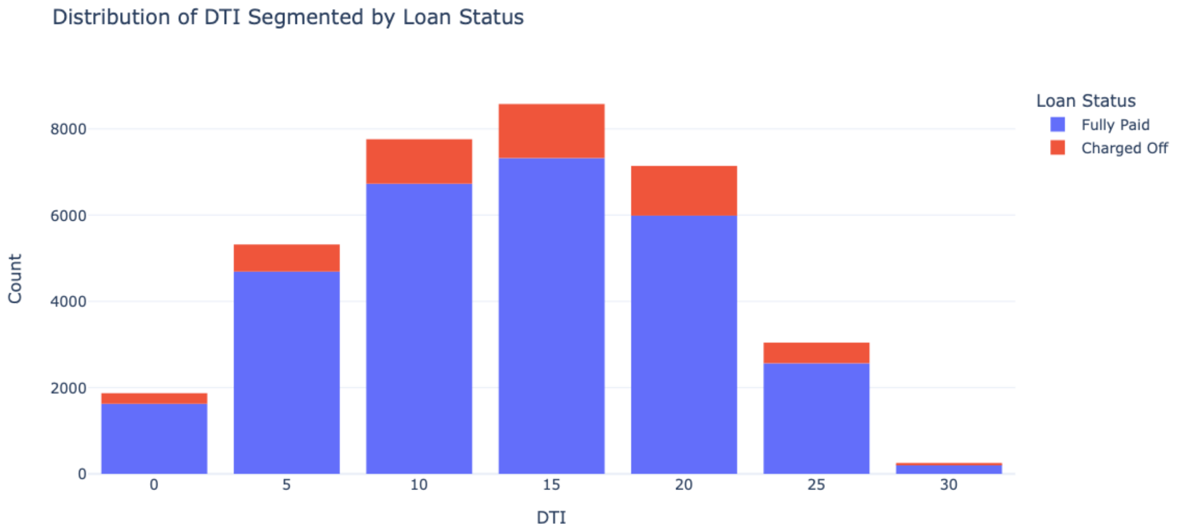
Data Analysis – Segmented Univariate Graphs



Distribution of Annual Income For Loan Status



Distribution of Interest Rate For Loan Status



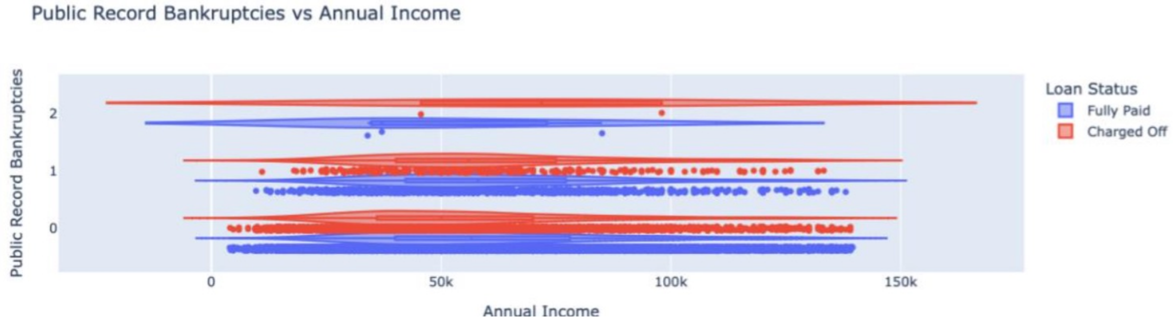
Distribution of DTI Segmented by Loan Status

Data Analysis – Bivariate

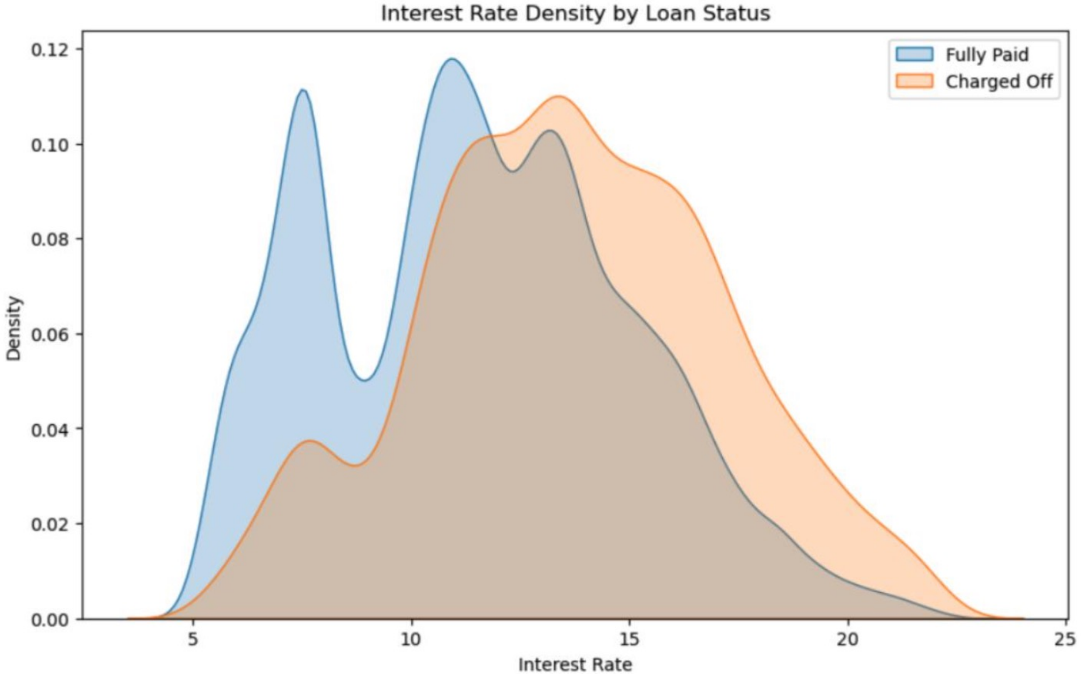
- Loans with grade A have the lowest median interest rate, around 7%. This indicates that A grade is a lower interest rate category.
- Loans with grade G have the highest median interest rate, around 20%. This indicates that G grade is a higher interest rate category.
- The median interest rate increases steadily from grade A to grade G, reflecting the increasing risk associated with the Grade.
- With grade C, D, E, F, G higher chance of defaulter
- Charged Off loans have consistently higher interest rates across all grades. This suggests that higher interest rates are a strong indicator of default risk
- Grades with (E, F, G) show higher median interest rates and more variability, indicating a higher risk of default. Defaulters are more likely found in these grades with high-interest rates
- Borrowers with higher annual incomes are more likely to fully repay their loans, particularly those with no public record bankruptcies.
- Borrowers with one or more public record bankruptcies are more likely to default on their loans, especially if they have lower annual incomes
- Charged Off loans have consistently higher interest rates across all grades. This suggests that higher interest rates are a strong indicator of default risk.
- Lower grades (E, F, G) show higher median interest rates and more variability, indicating a higher risk of default. Defaulters are more likely found in these lower grades with high-interest rates
- Fully Paid Loans (Blue):
  - These loans are more densely populated at higher annual income levels, indicating that borrowers with higher incomes are more likely to fully repay their loans.
  - The density of blue points is highest at the lower end of public record bankruptcies (0), indicating that borrowers with no bankruptcies and higher incomes tend to fully repay their loans.
- Charged Off Loans (Red):
  - These loans show a higher density at lower annual income levels, indicating a higher risk of default among lower-income borrowers.
  - The density of red points increases as the number of public record bankruptcies increases, indicating that borrowers with more bankruptcies and lower incomes are more likely to default.
  - Conclusion
  - Borrowers with higher annual incomes are more likely to fully repay their loans, particularly those with no public record bankruptcies.

Data Analysis – Bivariate Graphs

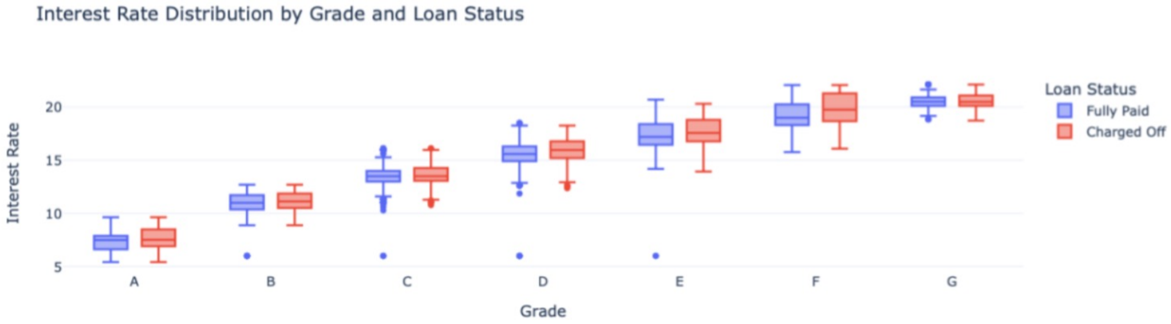
Public Record Bankruptcies vs Annual Income



Interest Rate Distribution by Loan Status

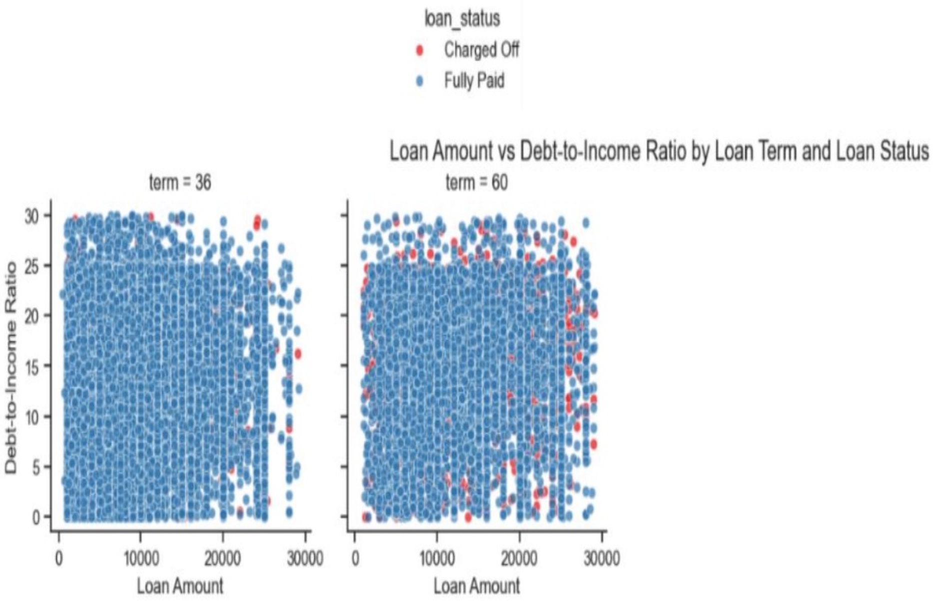


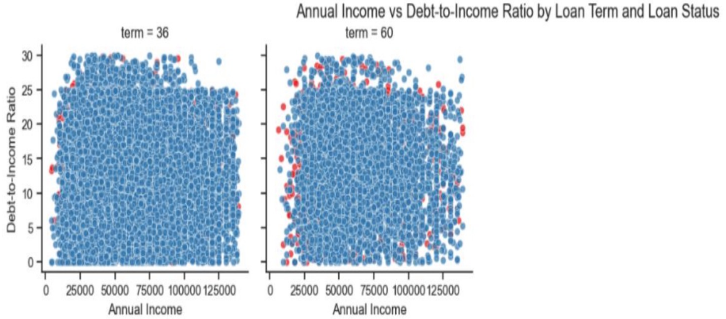
Interest Rate Distribution by Grade and Loan Status

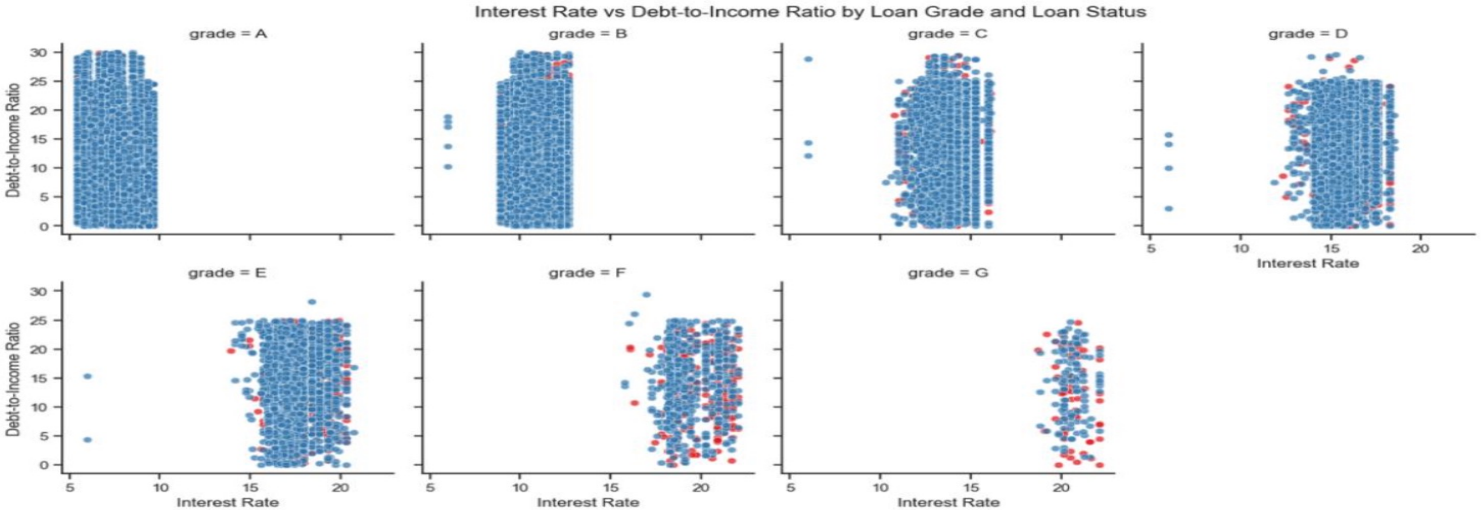


Data Analysis – Multivariate

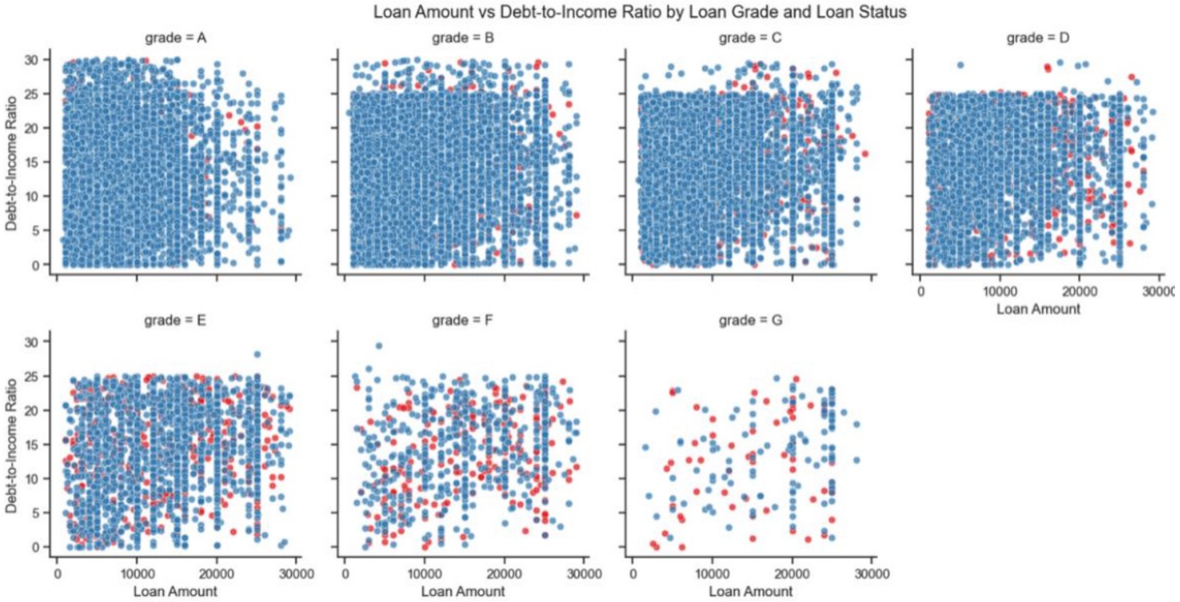
- A lower annual income combined with a higher interest rate is a strong indicator of a potential defaulter.
- DTI combined with a higher interest rate is a strong indicator of a potential defaulter than someone has no DTI.
- A higher loan amount combined with a higher interest rate is a strong indicator of a potential defaulter.
- Defaulter is likely, If they have high DTI and Large amount and they have taken the loan for 60 months term.
- Defaulter is likely, If they have some DTI and low annual income and they have taken the loan for 60 months term.
- Defaulter is likely from the Grade F and Garde G with higher Interest rate.

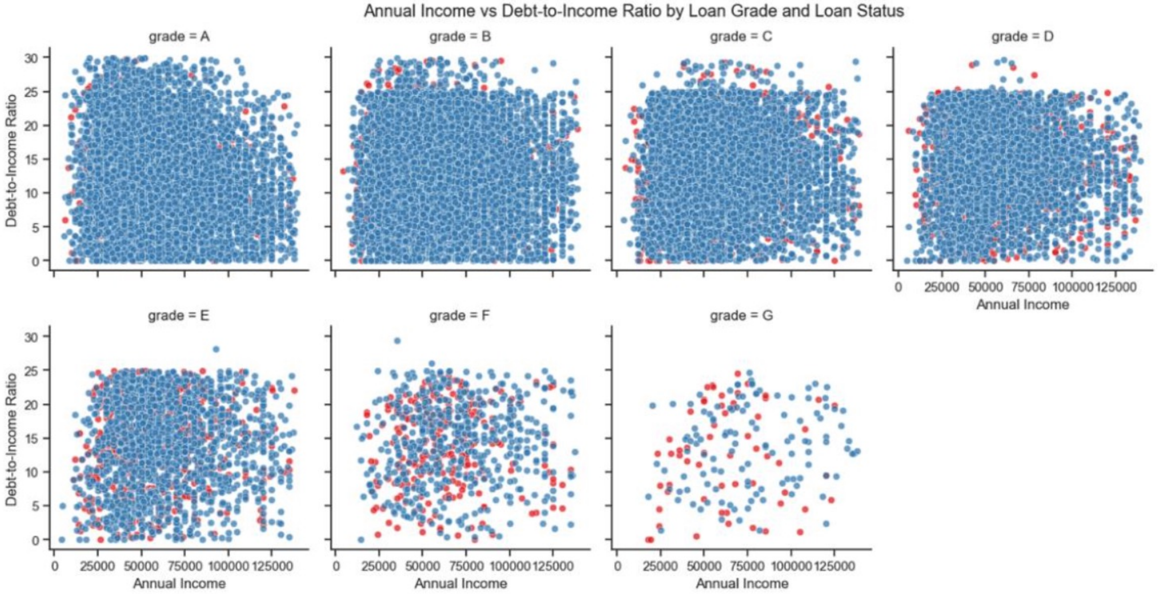

Loan Amount vs Debt-to-Income Ratio by Loan Term and Loan Status


Annual Income vs Debt-to-Income Ratio by Loan Term and Loan Status

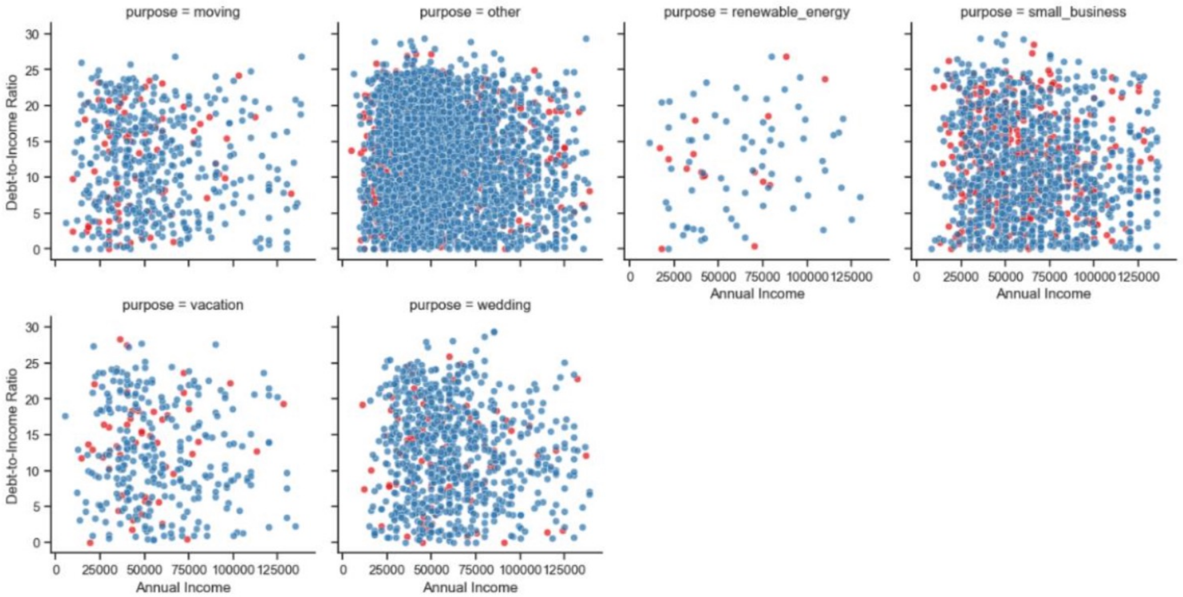

Interest Rate vs Debt-to-Income Ratio by Loan Grade and Loan Status

Analysis of Loan Amount vs DTI by Loan status and Grade





Analysis of Annual Income vs DTI by Purpose and Loan Status



As per the above analysis, defaulter is most likely, if they have taken a large amount loan for small business and they have low annual income.

Probability of Loan Default

Major combined driving factors that can be used to predict the probability of default and avoid credit loss

- DTI with Large amount and term
- DTI with low Annual Income and term
- Grade with higher Interest rate
- Purpose of loan - business
- Lower annual Income with higher interest
- Public record bankruptcies with lower annual income

More Major driving factors that can be used to predict the probability of default and avoid credit loss

- DTI
- Grade (C to G), Increasing Risk, with Increasing Grade
- Higher interest rates
- Low annual Income (Between 20K to 100K)
- Public record bankruptcies