LENDING CLUB CASE STUDY

Project work for Masters in Machine Learning and Artificial Intelligence
LJMU + IIIT Bangalore

done by **Himanshu S (@himanshuxd)**

Table of Contents :

1. Problem Statement
2. Business Understanding
3. Data Sourcing
4. Data Cleaning
5. Data Analysis Approach
6. Analysis
7. Observations
8. Conclusion

Problem Statement

Conduct an Exploratory Data Analysis (EDA) to gain insights into how consumer and loan attributes affect the likelihood of loan defaults.

We will leverage a provided dataset containing historical information on loan applicants and their default status. The objective is to uncover discernible patterns that can help predict whether an individual is at risk of defaulting on a loan. This information can inform various actions, such as loan denial, reducing loan amounts, or offering loans to higher-risk applicants at elevated interest rates.

By identifying applicants with a higher likelihood of default through EDA, we can mitigate credit losses by reducing the number of such loans. The primary goal of this case study is to pinpoint the critical variables that strongly influence loan defaults, serving as valuable insights for portfolio management and risk assessment within the company.

Business Understanding

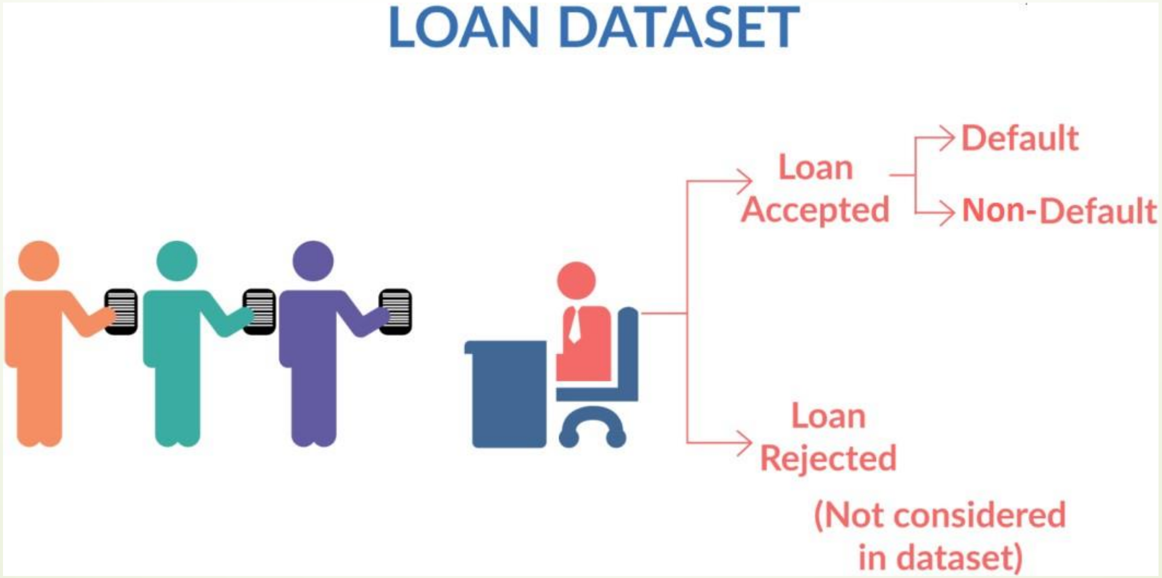
For a consumer finance company specializing in urban lending, critical loan approval decisions are driven by applicant profiles, considering two key risks:

1. **Risk of Not Approving a Loan**:
    1. If the applicant is likely to repay the loan but the loan is not approved, the company loses potential business.

2. **Risk of Default**:
    1. If the applicant is not likely to repay the loan (i.e., a high risk of default) and the loan is approved, it may lead to financial losses for the company.
    2. The largest source of financial loss, known as credit loss, occurs when loans are extended to 'risky' applicants who ultimately default.
    3. Credit loss represents the money lost by the lender when borrowers refuse to pay or abscond with the owed money.
    4. Customers labeled as 'charged-off' are categorized as 'defaulters' and contribute significantly to this credit loss.

Data Sourcing

The dataset "loan.csv" is included as a component of the case study. This dataset encompasses comprehensive loan information for all loans granted from 2007 to 2011. Additionally, a data dictionary, denoted as "Data_Dictionary.xlsx," is furnished in conjunction with the dataset. This data dictionary serves to elucidate the significance of the column names and variables within the dataset.



Data Cleaning

Data cleaning is a crucial phase in data analysis, involving the identification and resolution of quality issues in the dataset. It is often the most time-consuming aspect of data analysis. The goal is to address formatting errors, missing values, repeated rows, spelling inconsistencies, and other data anomalies. Working with clean data is essential to prevent errors and ensure meaningful and relevant results in subsequent analysis.

The following are the general steps to be followed for data cleaning:

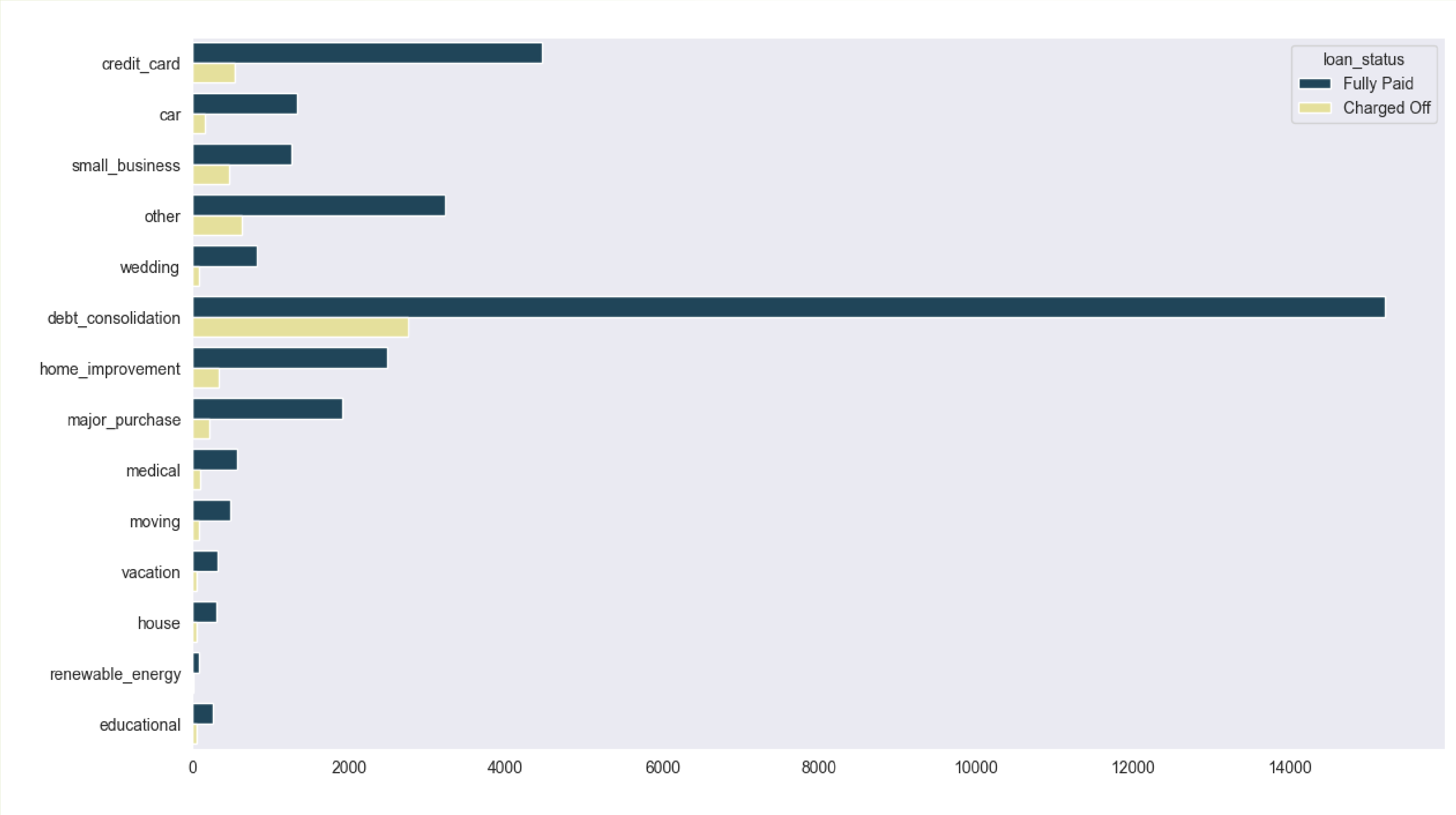
1. **Fix Rows and Columns**:
   Address issues related to the structure of the dataset, such as removing duplicate rows or redundant columns that do not contribute to the analysis.
2. **Fix Missing Values**:
   Handle missing data by either imputing missing values or deciding how to deal with them based on the nature of the data and the analysis objectives.
3. **Standardize Values**:
   Standardize data values to ensure consistency, such as converting data types, units, or formats to a uniform standard.
4. **Fix Invalid Values**:
   Identify and rectify invalid data entries that do not conform to the expected range or format.
5. **Filter Data**:
   Filter and select relevant columns and indicators that are essential for identifying characteristics of defaulters, discarding any irrelevant information.

Data Analysis Approach

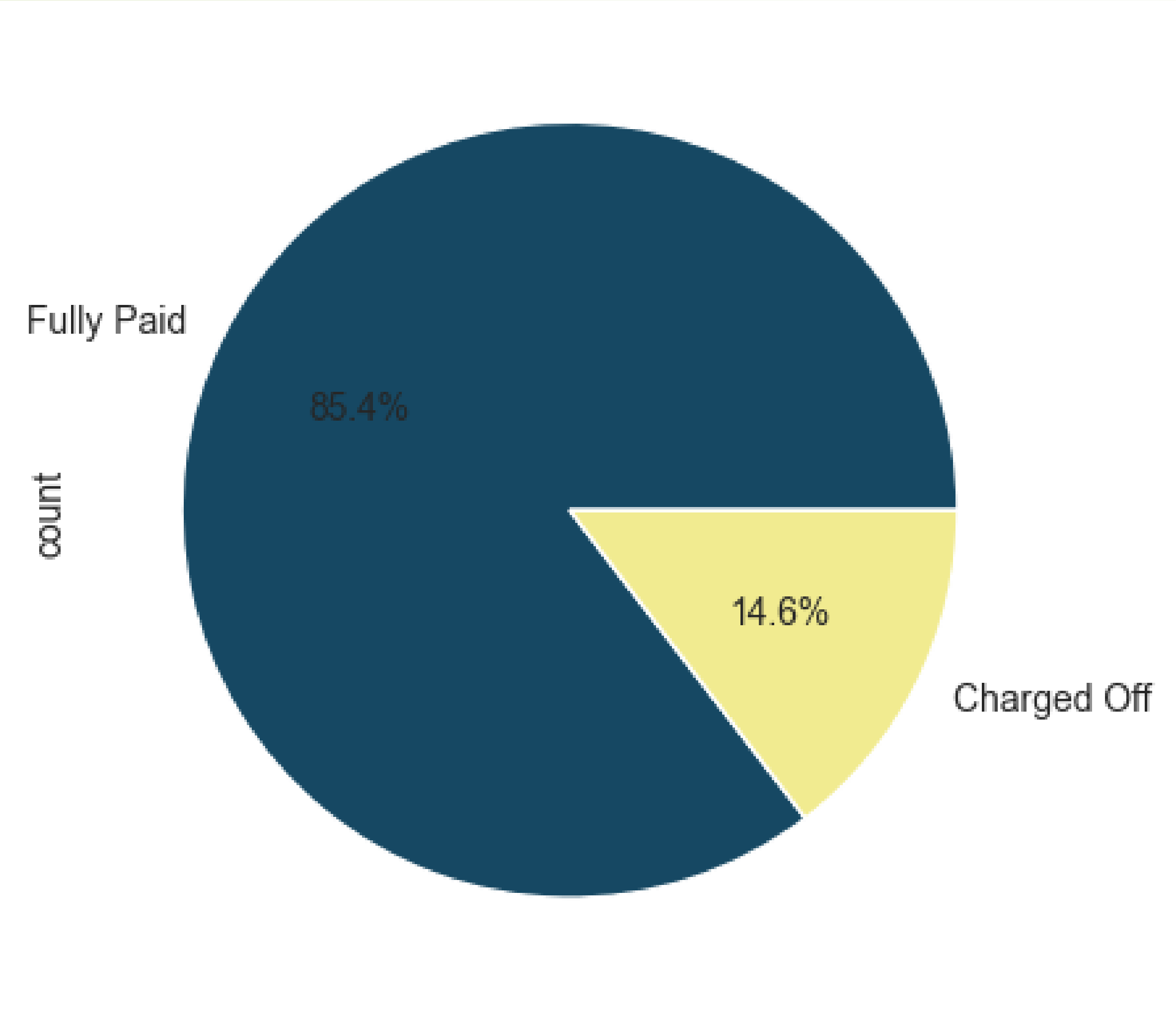
Based on the principles of exploratory data analysis (EDA), I plan to execute the following steps in the specified sequence to analyze the provided dataset:

1. **Data Sourcing**:
   Obtain and gather the dataset, ensuring that it is accessible and ready for analysis.
2. **Data Cleaning**:
   Clean and preprocess the data to handle missing values, data inconsistencies, and any other data quality issues.
3. **Outliers Detection**:
   Identify and address outliers in the dataset, which are values significantly different from the majority and might skew the analysis.
4. **Univariate Analysis**:
   Analyze individual variables (features) in the dataset to understand their distributions, central tendencies, and characteristics.
5. **Bivariate Analysis**:
   Examine relationships between pairs of variables to uncover correlations, dependencies, and insights that emerge when considering two variables together.

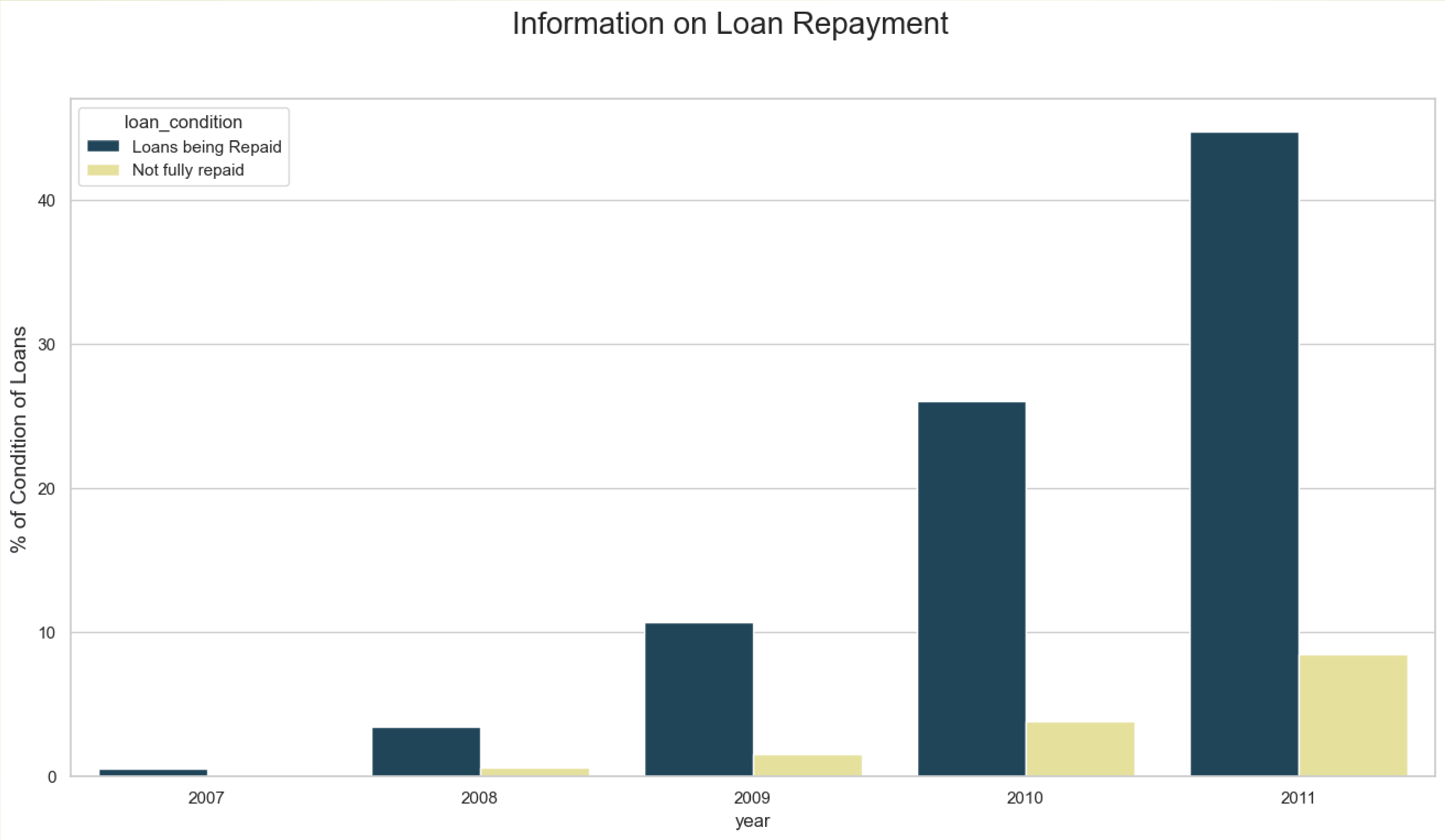
Analysis



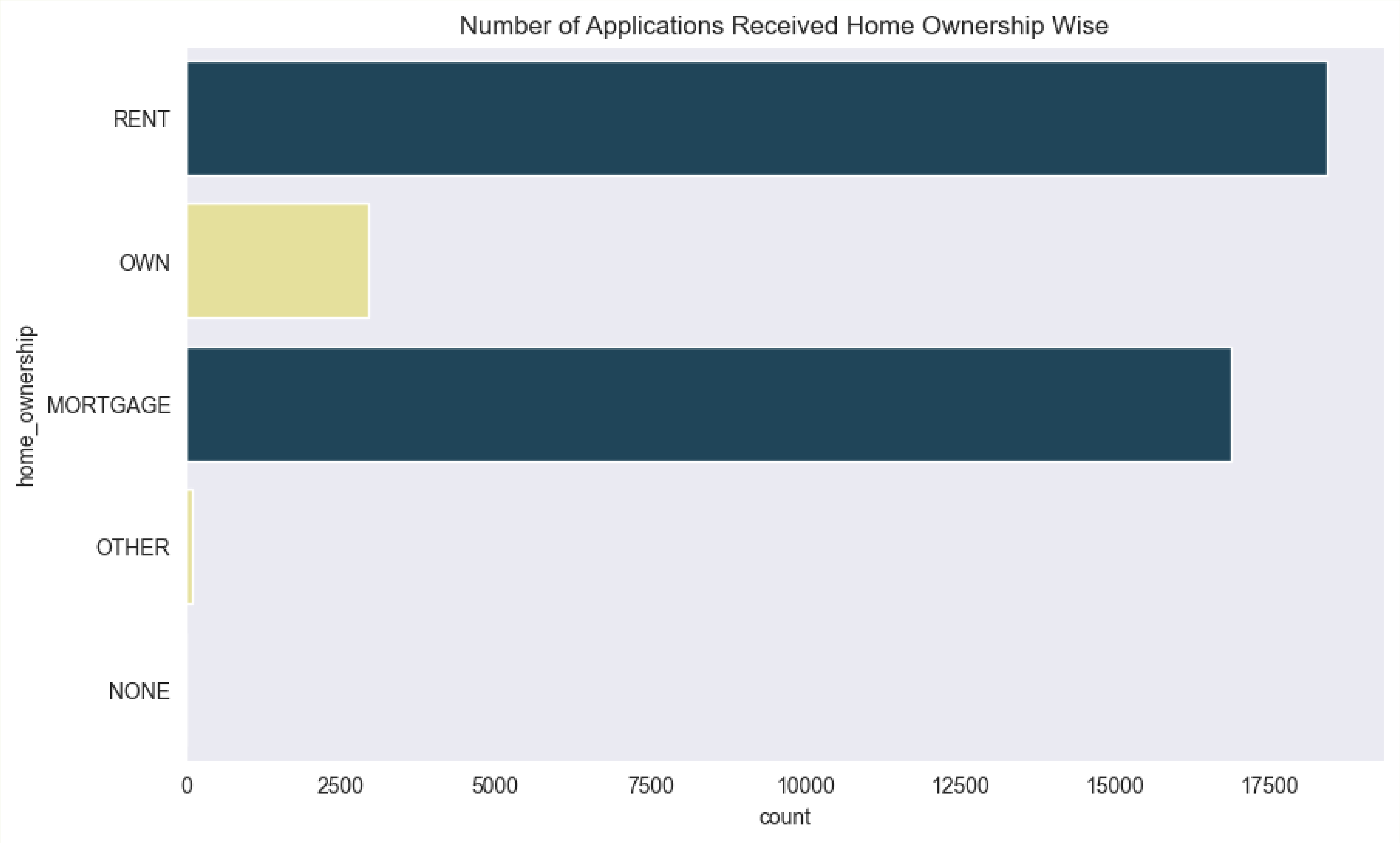
What we can observe is that most loans are being taken for the purpose of repayment of Debt followed by loans taken for Credit Card repayment.



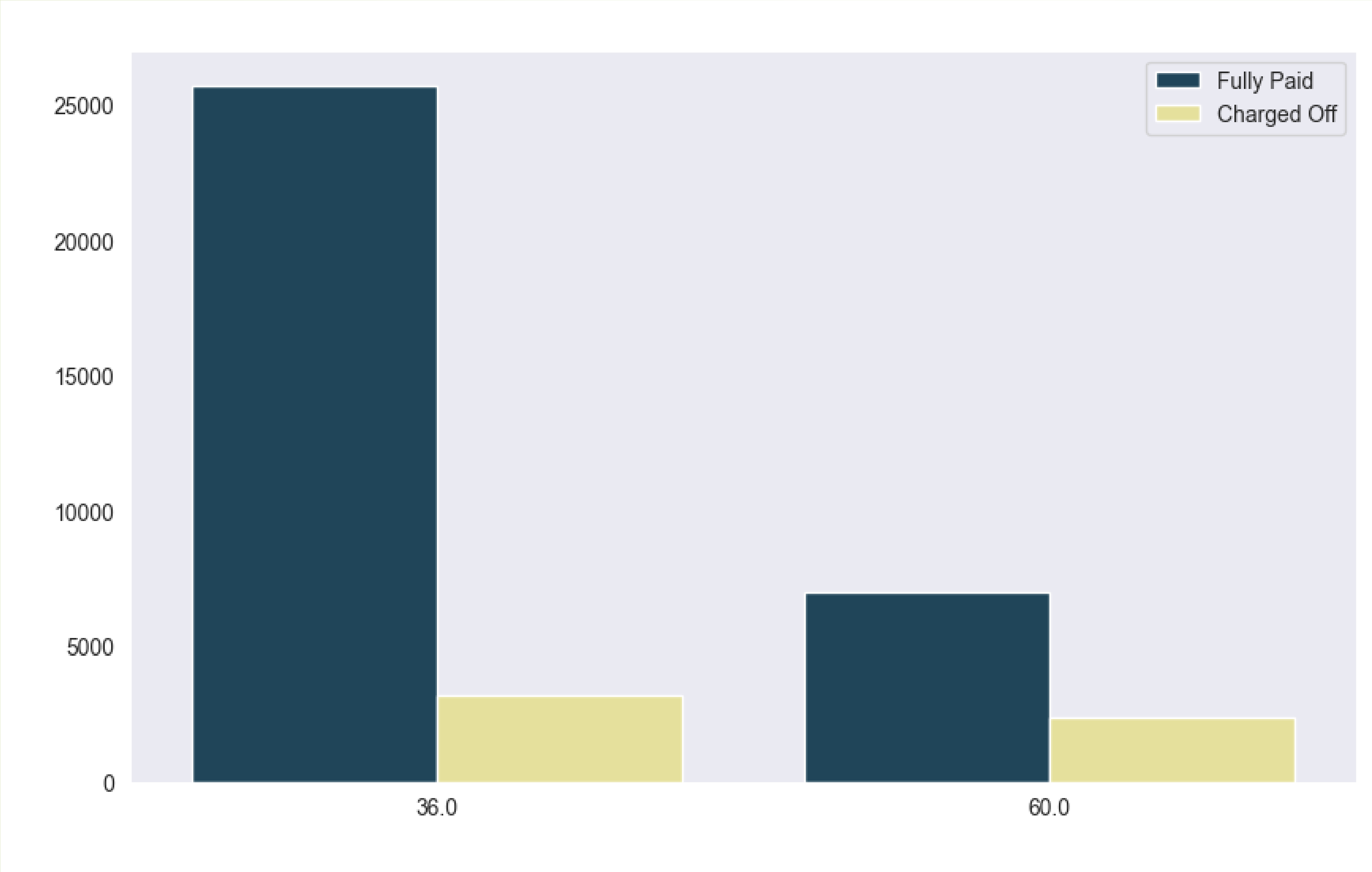
We see that 85.4% that is majority of the loans are being fully repaid rather than being Charged off



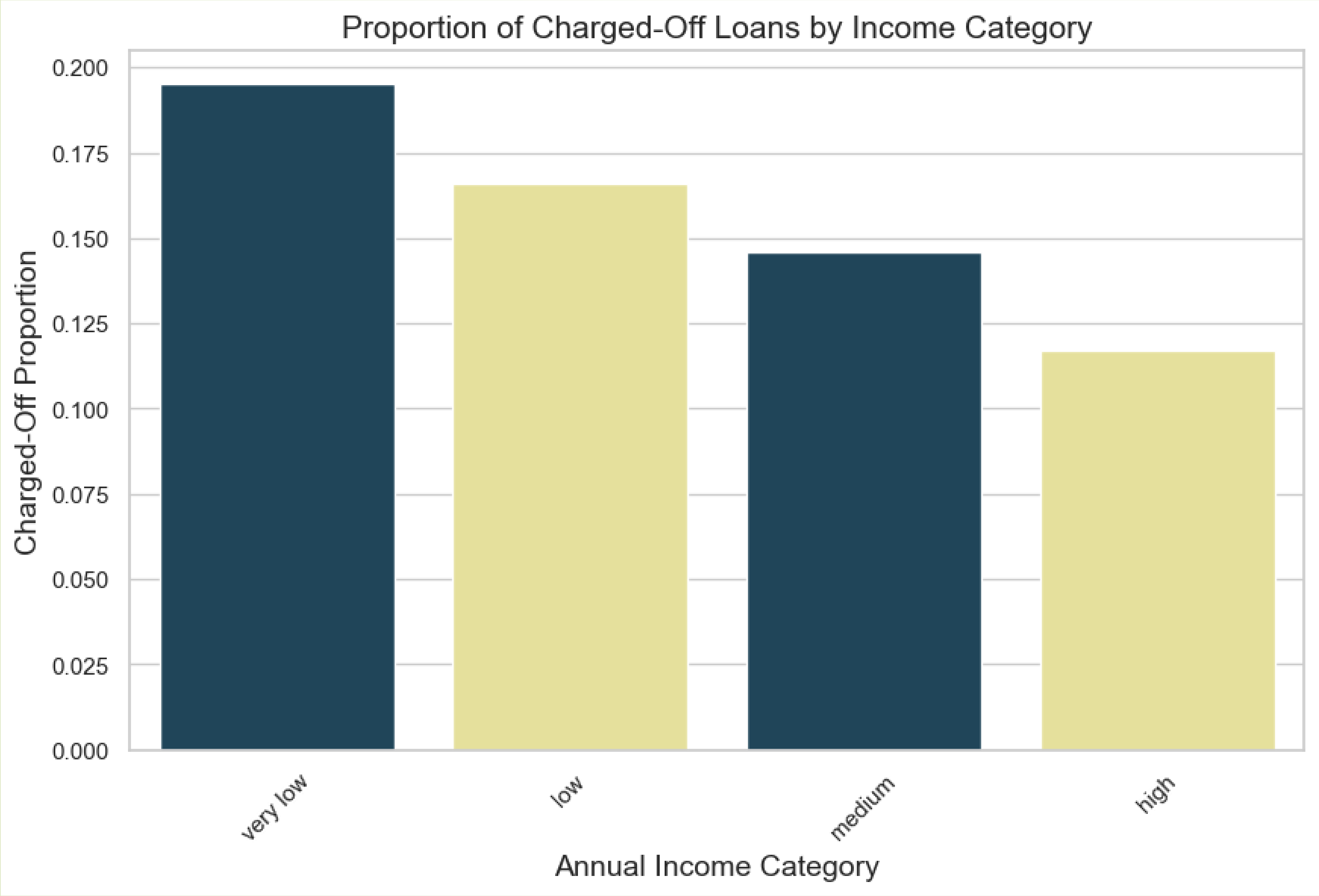
We observe that loans are getting repaid more and more over the years



Most loans are being taken with those who are living in a rented house or paying mortgage so it should not affect their capability to secure a loan



We observe that those opting for a 60 month repayment schedule are more likely of getting charged off compared to those paying the loan amount in 36 months. This is crucial when setting term limit.



Those with lower incomes are much more likely of getting charged off compared to those with higher incomes, it is observed that risk decreases as payee annual income increases.

Observations :

1.Loan Purpose: The majority of loans are taken for debt repayment and credit card payoff, suggesting that borrowers are using loans to manage existing financial obligations.

2.Loan Repayment: A significant 85.4% of loans are fully repaid, while only a minority are charged off. This indicates that Lending Club borrowers have a strong repayment history.

3.Loan Trends: Over the years, there is an increasing trend in loan repayments, suggesting that the Lending Club platform is gaining trust and popularity among borrowers.

4.Housing Status: Most borrowers live in rented houses or have mortgages. Housing status doesn't seem to be a significant factor affecting loan approval, indicating that it doesn't impair their ability to secure a loan.

5.Loan Term: Borrowers opting for a 60-month repayment schedule have a higher likelihood of being charged off compared to those choosing a 36-month term. This implies that shorter loan terms may be a safer choice for borrowers and lenders alike.

6.Income Influence: Borrowers with lower incomes have a higher likelihood of being charged off, while risk decreases as payee annual income increases. This emphasizes the importance of considering income when assessing creditworthiness.

Conclusion :

1. Lending Club is fulfilling its role as a platform for helping borrowers manage existing debt and credit card payments.
2. Lending Club's business model appears to be sound, with a high percentage of loans being successfully repaid, which is essential for its sustainability and growth.
3. The increasing trend in loan repayments over the years is a positive sign, indicating growing trust in the platform.
4. Housing status does not appear to be a major factor affecting loan approval, which is beneficial for a wide range of potential borrowers.
5. Lending Club should carefully consider loan term options, as shorter terms seem to result in lower default rates, benefiting both borrowers and the platform's financial health.
6. Income is a significant determinant of loan repayment risk, highlighting the importance of thorough income assessment when evaluating loan applications.