Lending Club Case Study

Sireesha Sunkara

Suraj Kumar

Introduction:

This case study involves performing Exploratory Data Analysis (EDA) to understand the factors influencing loan default in a consumer finance company. By analyzing historical loan application data, we aim to identify patterns that can help minimize financial losses from loan defaults.

Problem Statement:

To identify risky loan applicants and reduce credit loss through EDA.

Business Understanding :

The finance company faces two types of risks:

Loss of Business: Not approving a loan for an applicant who is likely to repay.
Financial Loss: Approving a loan for an applicant who is likely to default.

The objective is to identify attributes of applicants that predict default risk, allowing the company to make informed lending decisions.

Data Understanding:

The dataset includes information about past loan applicants (2007-2011), detailing various consumer and loan attributes along with a label indicating whether they defaulted (charged-off) or not.

Key Terms:

Charged-off: Applicant has defaulted.
Fully Paid: Applicant has repaid the loan.
Current: Applicant is still in repayment.

Analysis Approach:

Data Importing and Understanding:

- Imported the necessary libraries for the analysis, including Pandas, Seaborn, Matplotlib, and NumPy.
- Loaded the loan dataset using pd.read_csv().
- Checked the top 5 records with head().
- Obtained basic information (total records, columns, data types) using info().
- Reviewed statistical details (average, minimum, maximum) with describe().

Data Cleaning Process:

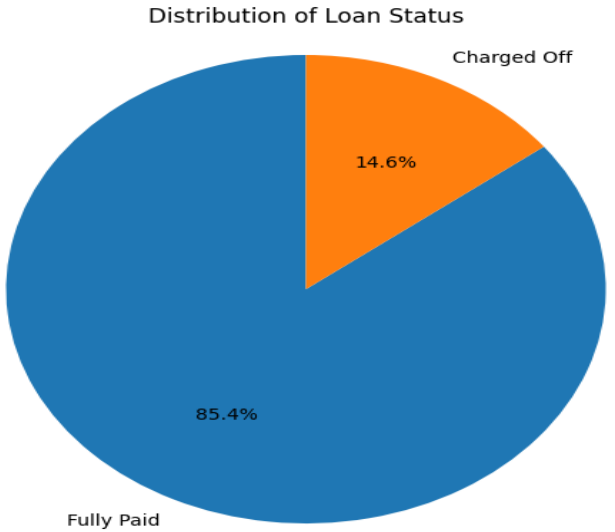
- Removed all columns with no values (all entries as NA).
- Calculated the percentage of missing values for each column.
- Dropped columns with more than 30% missing values.
- Filled missing values in numerical columns with the median.
- Filled missing values in categorical columns with the mode (most common value).
- Changed the case of specific columns (emp_title, title, purpose) to title case for consistency.
- Removed rows where the loan status was 'Current' to focus on completed loans.
- Re-evaluated the percentage of missing values after cleaning.

Univariate Analysis :

1) Loan Status Distribution Pie Chart:

Loan Status: Created a pie chart showing the distribution of loan statuses.
The chart reveals the percentage of each loan status category (e.g., Fully Paid, Charged-off).

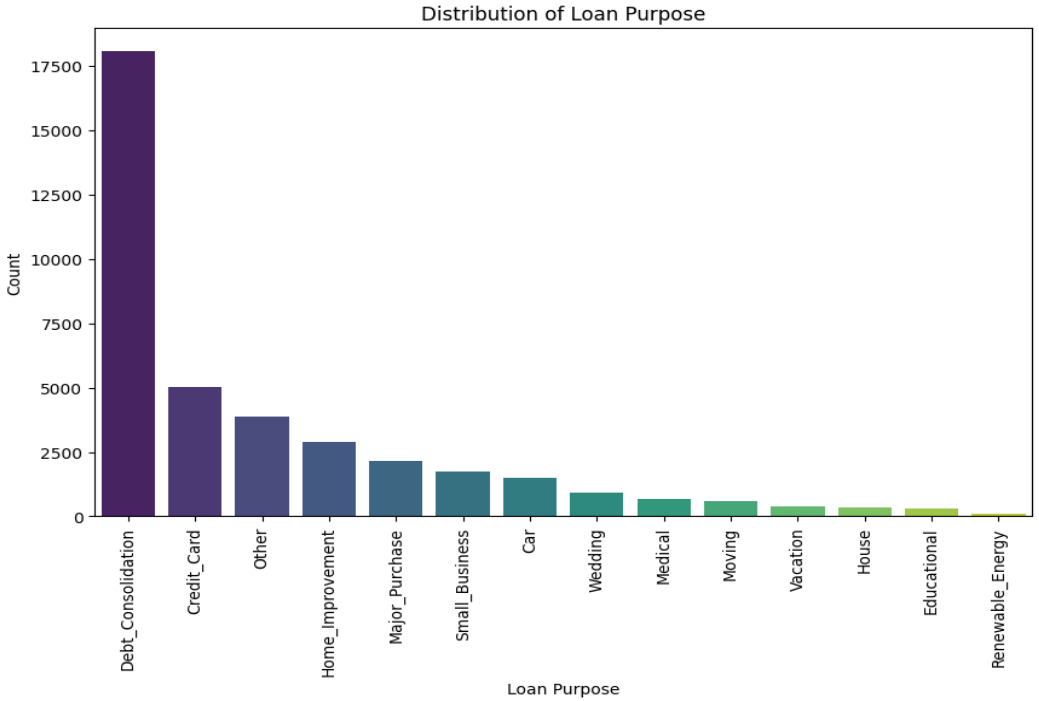
Conclusion: The pie chart shows that around 14.6% of borrowers have defaulted on their loans. This highlights a significant portion of borrowers who are at risk of not repaying.



2) Loan Purpose Distribution Bar Plot:

Loan Purpose: Generated a bar plot to illustrate the distribution of loan purposes.
The plot shows the count of loans for different purposes (e.g., debt consolidation, home improvement).

Conclusion: The bar chart shows that most borrowers take loans for debt consolidation, credit card repayment, and car purchases, with debt consolidation being the most common purpose.



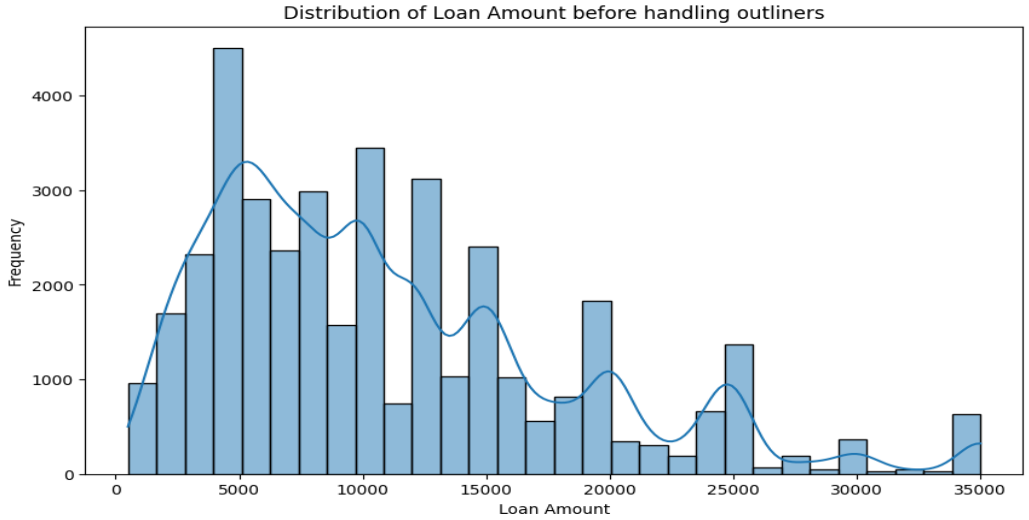
Analyzing Loan Amount and Annual Income :

1) Loan Amount Distribution with outliners: Histogram showing the frequency of loan amounts before handling outliers.

Conclusion: The distribution of Histogram is right-skewed, means longer tail on the right side. This indicates that there are some larger loan amounts that are less frequent compared to the smaller loan amounts.

The presence of outliers indicating that there are some unusually large loan amounts that may be affecting the overall distribution.

The peak of the distribution appears to be around the 5000-10000 loan amount range, indicating that this is the most common loan amount. repaying

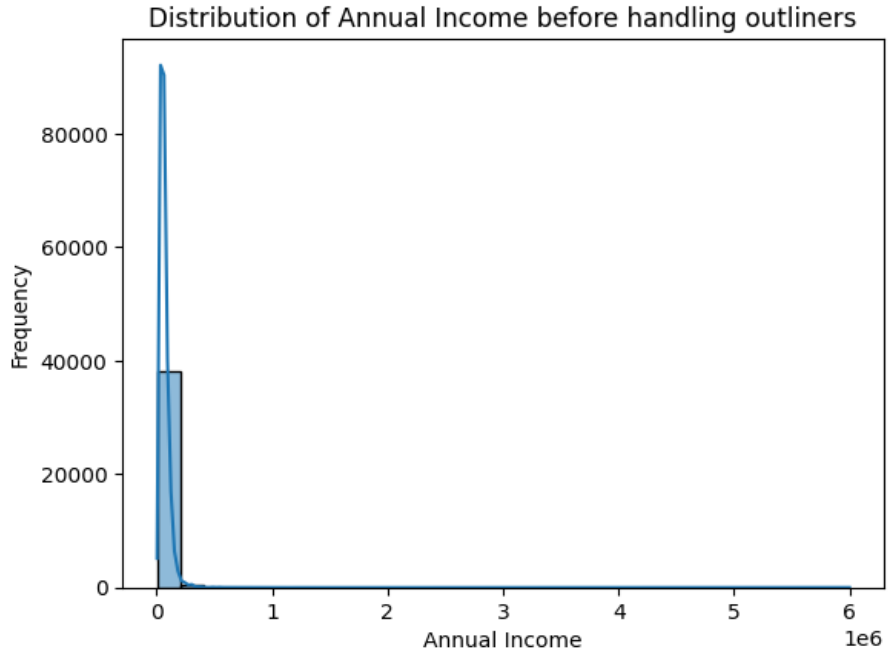


2) Annual Income Distribution with outliners: Histogram illustrating the frequency of annual incomes before handling outliers.

Conclusion: The distribution is right-skewed, longer tail on the right side. This indicates that there are some very high incomes that are less frequent compared to the lower incomes.

The presence of outliers indicating that there are some extremely high incomes that may be skewing the distribution.

The peak of the distribution appears to be around the 40000-60000 income range, indicating that this is the most common income level.



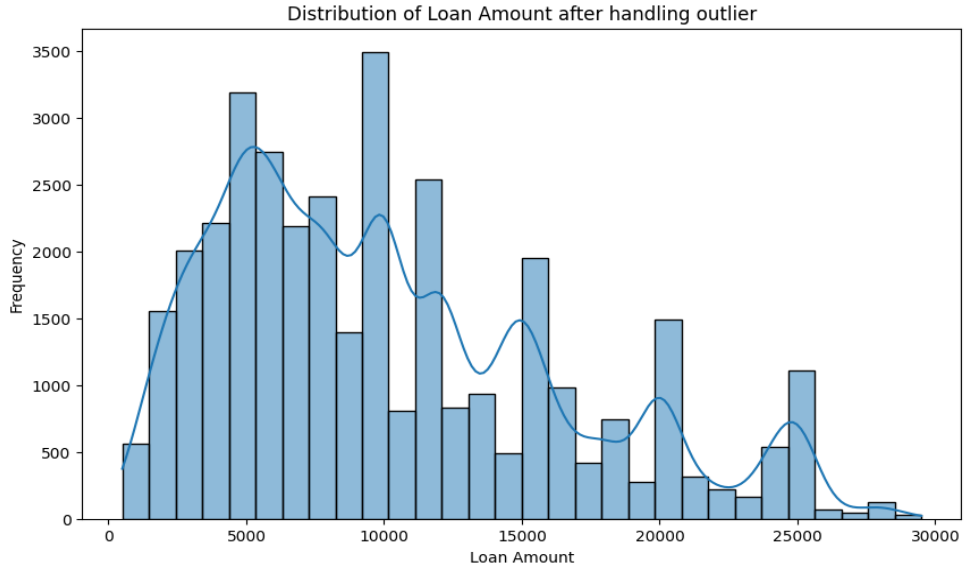
Analyzing Loan Amount and Annual Income after handling outliers :

1) Loan Amount Distribution without outliners: Histogram showing the frequency of loan amounts after handling outliers.

Conclusion: The distribution of Histogram is right-skewed, means longer tail on the right side. This indicates that there are some larger loan amounts that are less frequent compared to the smaller loan amounts.

The distribution is approximately bell-shaped, indicating a normal distribution. The data is centered around a particular loan amount with a relatively symmetrical spread.

The peak of the distribution appears to be around the 5000-10000 loan amount range, indicating that this is the most common loan amount.

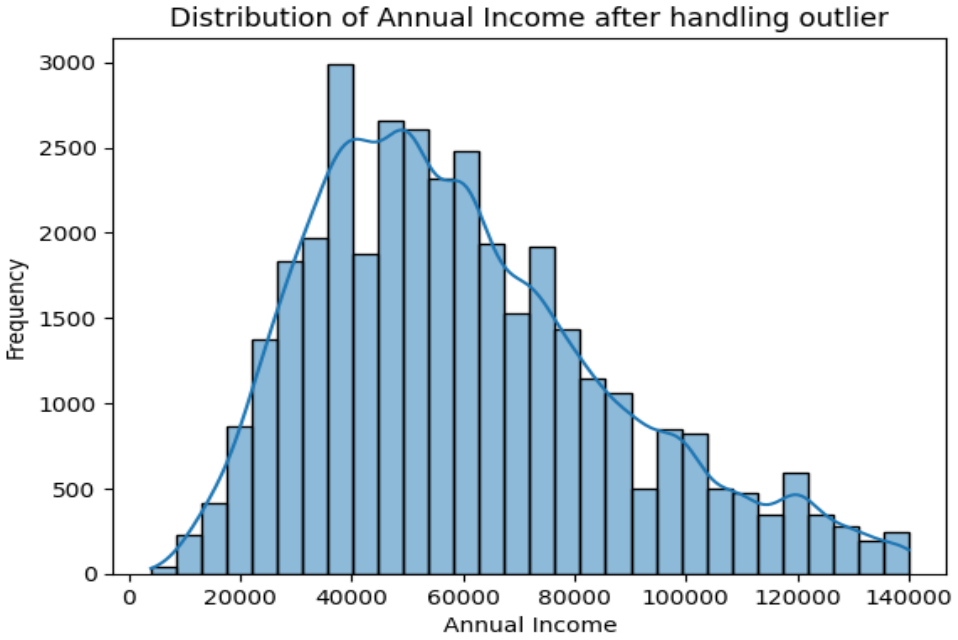


2) Annual Income Distribution without outliners: Histogram illustrating the frequency of annual incomes after handling outliers.

Conclusion: The distribution is right-skewed, longer tail on the right side. This indicates that there are some very high incomes that are less frequent compared to the lower incomes.

After removing outliers, the distribution is now more symmetrical, with a clear bell-curve shape. This indicates that the data is more normally distributed after handling outliers.

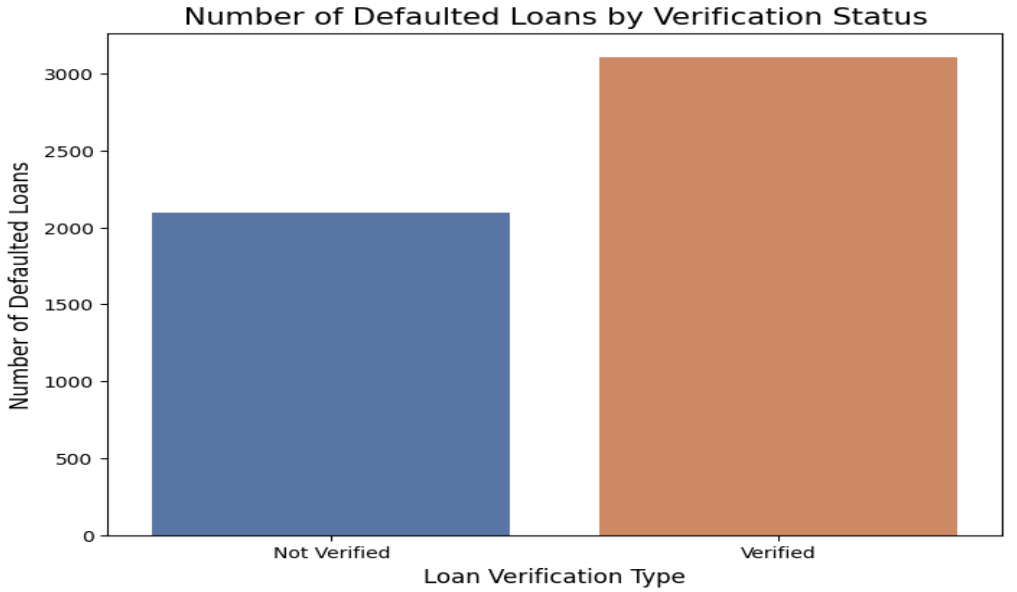
The peak of the distribution appears to be around the 40000-60000 income range, indicating that this is the most common income level.



Bivariate Analysis:

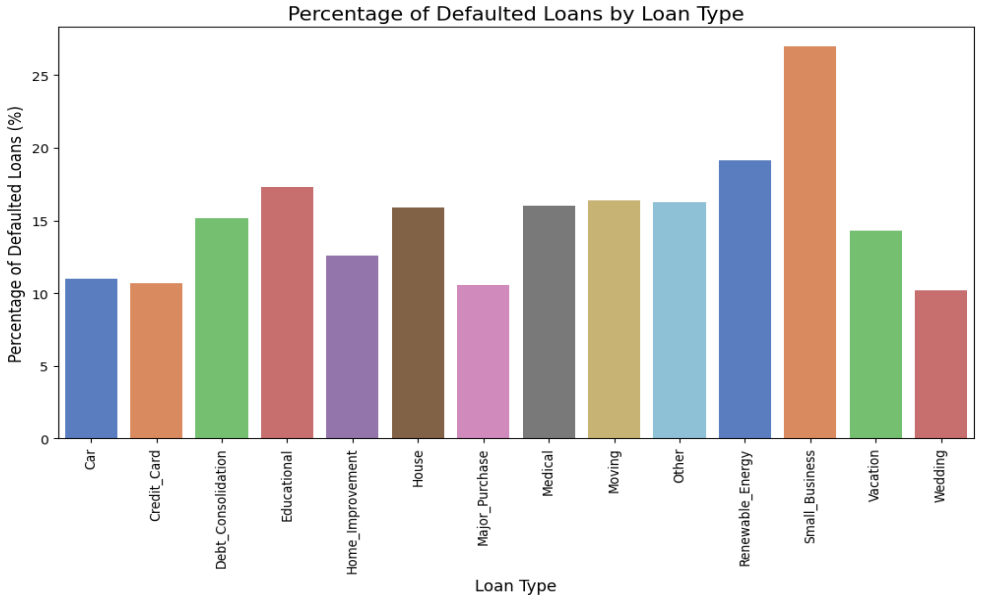
1) Analysis of Loan Default by Verification Status:
The bar plot shows the number of defaulted loans categorized by loan verification status, with 'Verified' and 'Source Verified' combined into a single 'Verified' category. The chart compares the number of defaulted loans for two verification statuses "Not Verified" and "Verified."

Conclusion: The number of defaulted loans is higher for "Verified" loans compared to "Not Verified" loans. that



2) Analysis of Loan Defaults by Loan Purpose: The bar plot shows the percentage of loans that defaulted for each loan purpose. This helps us see which types of loans are more likely to result in defaults. Here we group the data by loan purpose and counts the number of defaults, then normalizes these counts to get percentages.

Conclusion: Based on our analysis the loan types, such as "Small Business" and "Wedding," have significantly higher default rates compared to others.



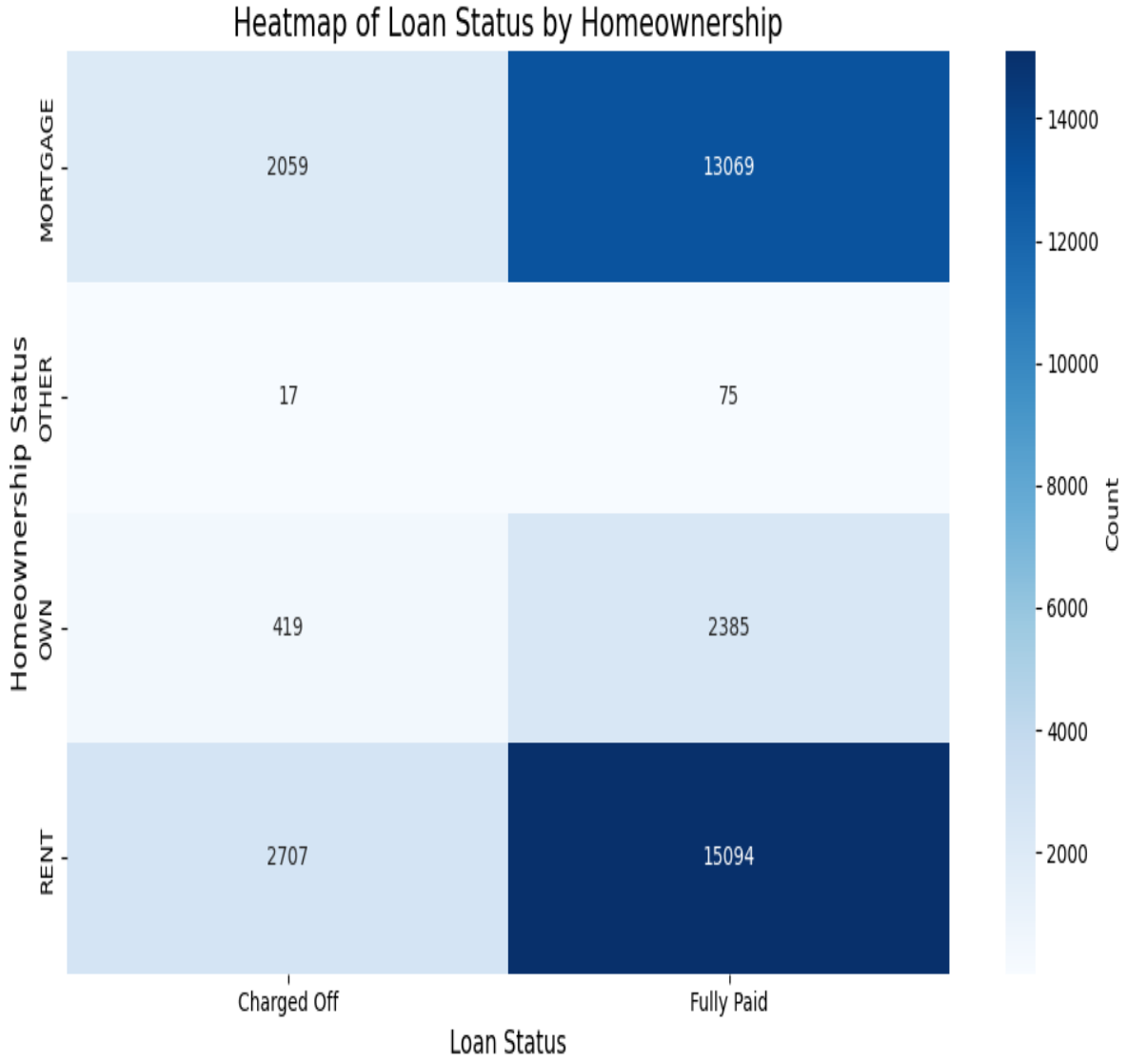
Analysis of Loan Status by Homeownership:

- First, we cleaned out the entries with "NONE" as the homeownership status.
- Grouped the data by homeownership and loan status
- Created Heat map with the result data.

The heatmap provides a visual representation of the relationship between loan status (Charged Off or Fully Paid) and homeownership status (MORTGAGE, OTHER, OWN, RENT). The color intensity in each cell indicates the count of loans within that specific combination of loan status and homeownership. number

Conclusion:
- Based on the analysis and heatmap data, there is a strong association between homeownership status and loan status.

Borrowers with a MORTGAGE are most likely to fully pay off their loans, while renters are more likely to default. This suggests that homeownership may be a positive indicator of creditworthiness and repayment ability.



Multivariate analysis:

Analysis of Loan Amount Distribution by Purpose and Loan Status :

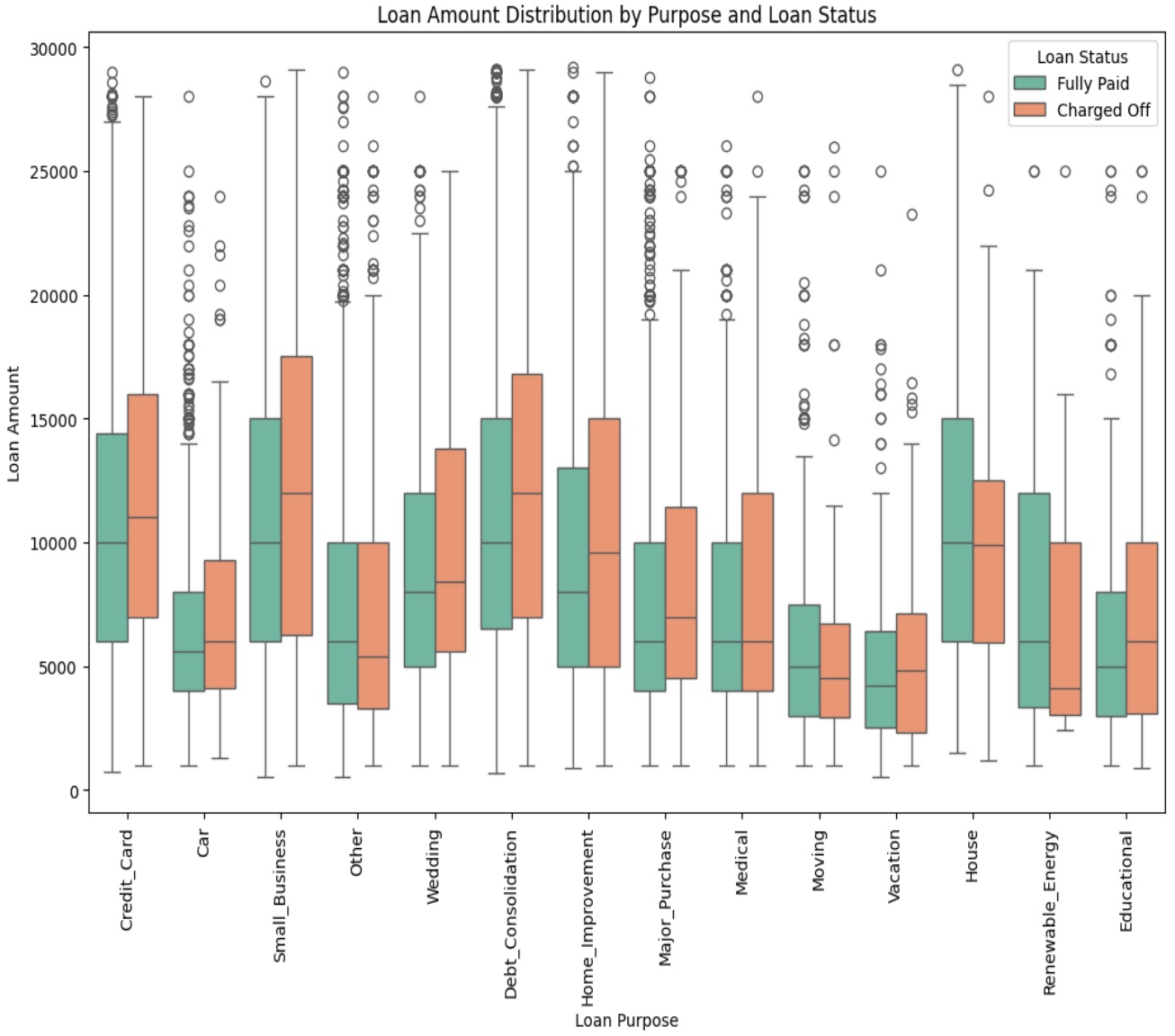
- Calculated the 25th (Q1) and 75th (Q3) percentiles of loan amounts to determine the spread of the data.
- Lower and upper bounds are calculated using the interquartile range (IQR) to define outlier limits.
- Created a new DataFrame by excluding the loan amounts outside of these bounds.
- Created a box plot to show the distribution of loan amounts by purpose and loan status, using different colors for each status.

All this analysis helped us to understand how loan amounts vary by purpose and the differences between fully paid loans and defaults.
The rows removed for outliers are **836**.

Conclusion:
- The box plots for "Small Business" and "Wedding" shows a wider interquartile range, which indicates greater variability in loan amounts for these loan purposes.

- The median loan amounts for "Small Business" and "Wedding" are higher than other loan purposes.

Based on analysis Charged-Off Loans are associated with Higher Amounts. The borrowers who default on loans tend to have taken out larger loans, especially for categories like small business and wedding.



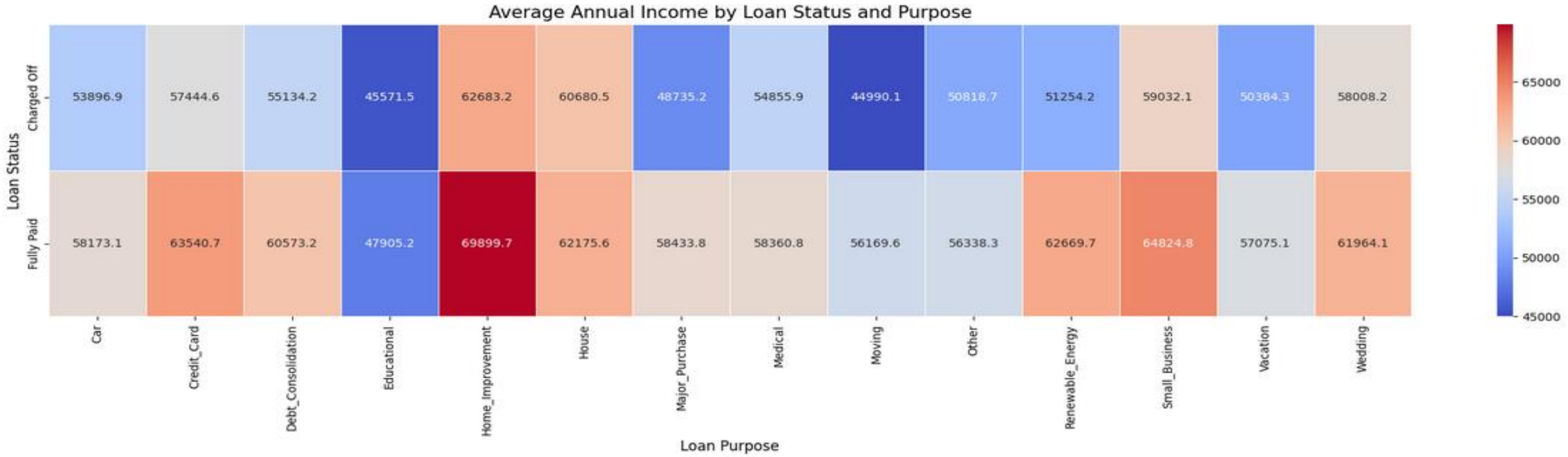
Analysis of Average Annual Income by Loan Status and Purpose:

- First, we cleaned out the entries with "NONE" as the homeownership status.
- Grouped the data by homeownership and loan status and created Heat map with the result data.

The heatmap provides a visual representation of the relationship between loan status (Charged Off or Fully Paid) and homeownership status (MORTGAGE, OTHER, OWN, RENT). The color intensity in each cell indicates the count of loans within that specific combination of loan status and homeownership.

Conclusion:
- Based on the heatmap, there is a strong association between annual income and loan status. Borrowers with higher annual incomes are more likely to fully pay off their loans, while those with lower incomes are more likely to default.

- The loan purpose significantly influences the average annual income of borrowers. Loans for Small Business, Debt Consolidation, and Home Improvement typically involve borrowers with higher incomes, while loans for Credit Card, Car, and Medical purposes often involve borrowers with lower incomes.
removed



Key Findings:

Loan Status Distribution Pie Chart: The pie chart shows that around 14.6% of borrowers have defaulted on their loans.

Loan Purpose Distribution Bar Plot: The bar chart shows that most borrowers take loans for debt consolidation, credit card repayment, and car purchases, with debt consolidation being the most common purpose.

Loan Amount Distribution: The loan amount Histogram distribution is right-skewed, indicating a concentration of smaller loans. The peak of the distribution appears to be around the 5000-10000 loan amount range, indicating that this is the most common loan amount.

Annual Income Distribution: The annual income distribution is also right-skewed. The peak of the distribution appears to be around the 40000-60000 income range, indicating that this is the most common annual income.

Loan Status and Annual Income: Borrowers with higher annual incomes are more likely to fully pay off their loans, while those with lower incomes are more likely to default.

Loan Purpose and Annual Income: Certain loan purposes (e.g., Small Business, Debt Consolidation) are associated with higher average annual incomes, while others (e.g., Credit Card, Car) are associated with lower incomes.

Loan Purpose and Default Rates: The default rate varies across different loan purposes, with Small Business and Wedding loans having higher default rates. Loans for Small Business, Debt Consolidation, and Home Improvement have higher default rates, suggesting that these loan purposes may involve greater financial risks.

Homeownership and Loan Status: Borrowers with a mortgage are more likely to fully pay off their loans compared to renters. Renters are generally more likely to default removed.

Loan Amount Distribution by loan status and purpose : Borrowers with larger loan amounts are more likely to be charged off.

Average Annual Income by loan status and purpose : Borrowers with higher income are more likely to repay their loans.

Overall Conclusions:

Loan Performance is Influenced by Multiple Factors: Loan status is influenced by a combination of factors, including annual income, loan purpose, and homeownership.

Income plays a Significant Role: Borrowers with higher incomes are more likely to repay their loans on time.

Loan Purpose Matters: The type of loan can also affect repayment outcomes. Loans for Small Business and Wedding are associated with higher risk.

Loan Amount : Borrowers with larger loan amounts are more likely to be charged off.

Homeownership is a Positive Indicator: Being a homeowner, especially with a mortgage, is generally associated with better loan performance.

Thank you