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:

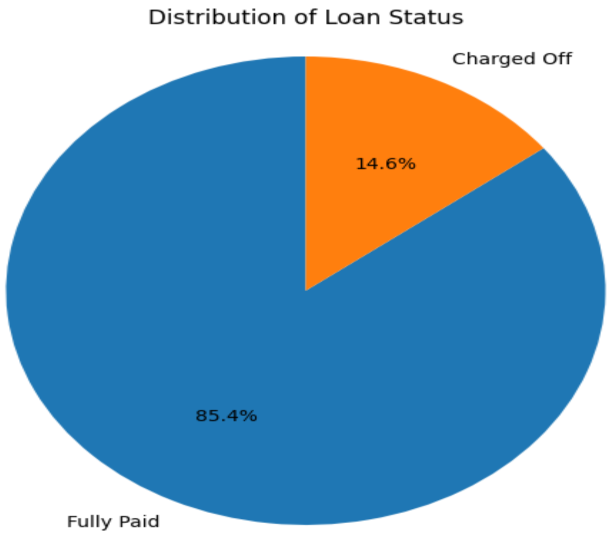
- Removed all columns with no values (all entries as NA).
- Calculated the percentage of missing values for each column.
- Dropped columns with more than 30% missing values.
- Filled missing values in numerical columns with the median.
- Filled missing values in categorical columns with the mode (most 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).

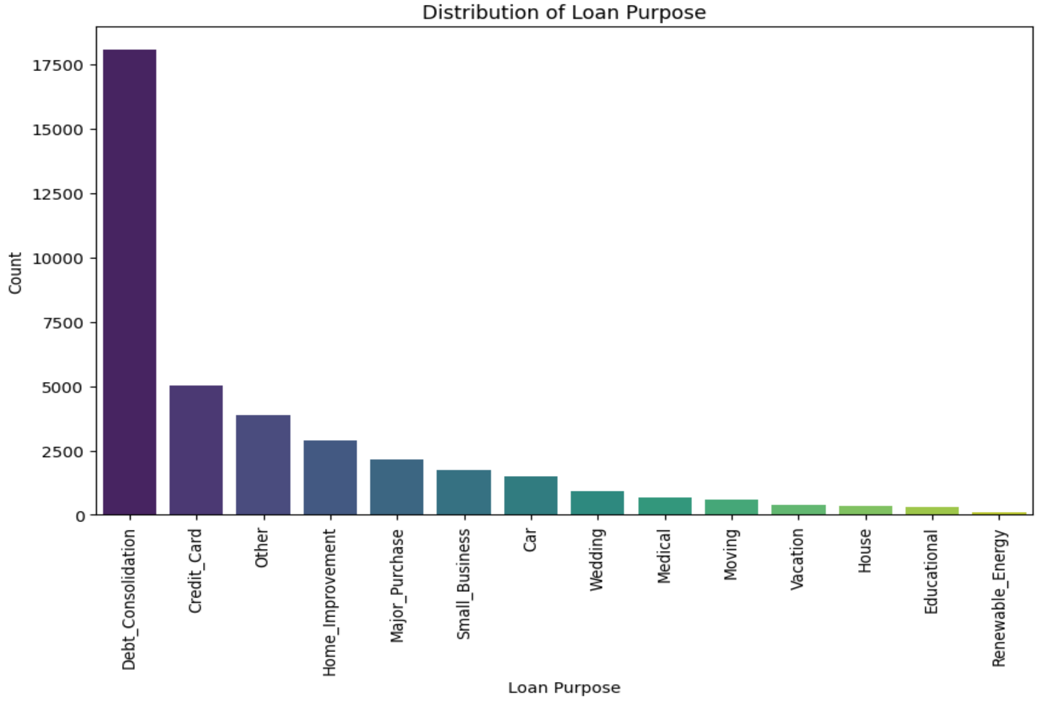
Conclusion: The pie chart shows that around 14.6% of borrowers have defaulted on their loans. This highlights a significant portion of borrowers who are at risk of not repaying.



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.



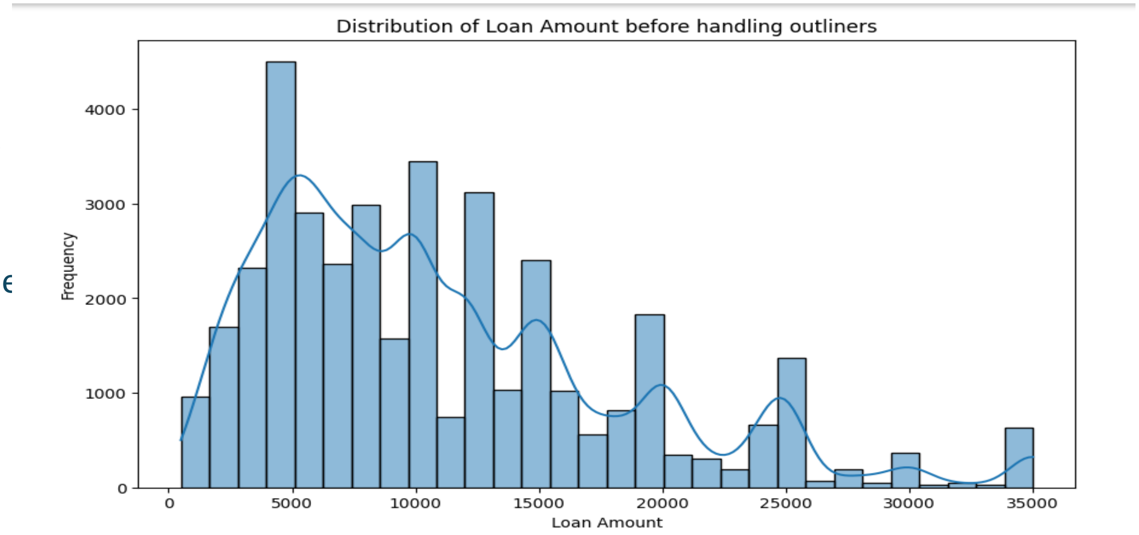
Analyzing Loan Amount and Annual Income :

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

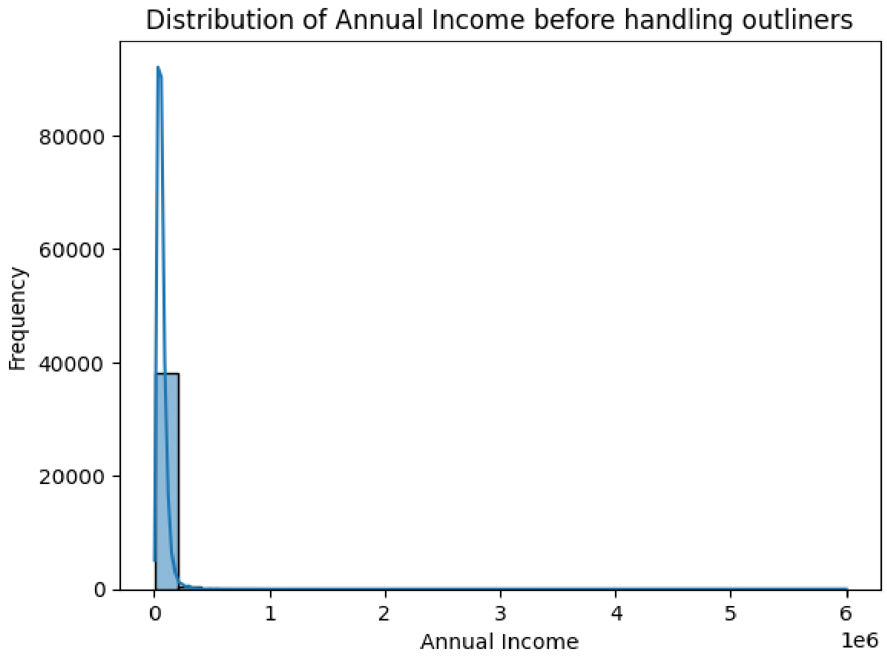


2) Annual Income Distribution with outliners: Histogram illustrating the frequency of annual incomes before handling outliers.

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.

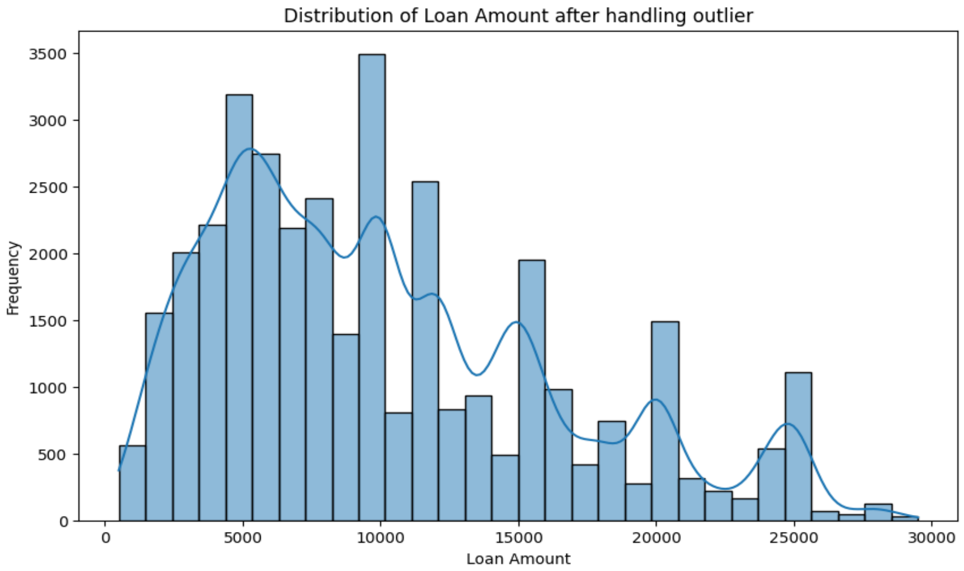
The peak of the distribution appears to be around the 5000-10000 loan amount range, indicating that this is the most common loan amount.

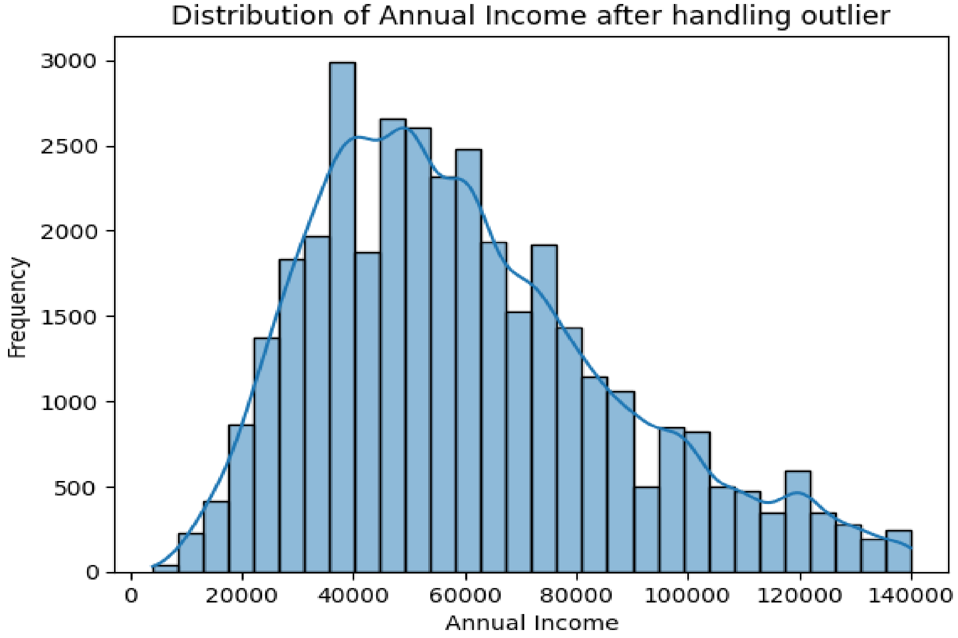
2) Annual Income Distribution without outliners: Histogram illustrating the frequency of annual incomes after handling outliers.

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.

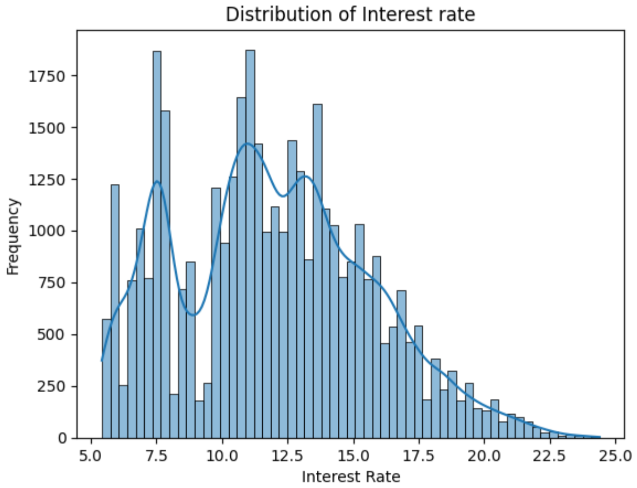




1) Distribution of Interest rate The provided histogram shows the distribution of Interest rate.

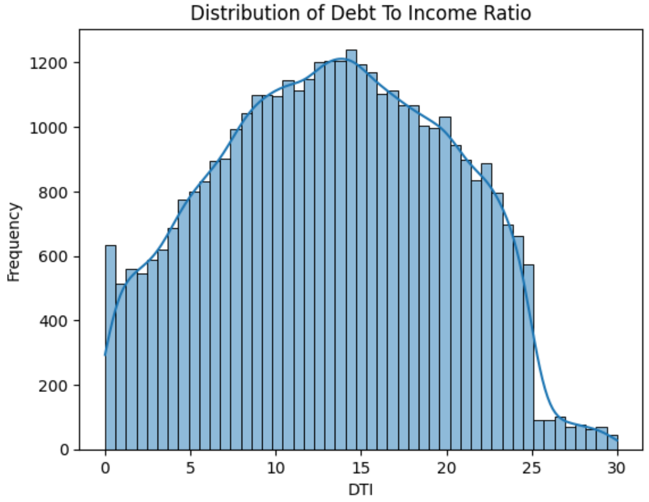
Conclusion: The histogram shows a right-skewed distribution of interest rates with a peak around 10-12.5%, indicating that most interest rates fall within this range, but there is a significant portion with higher rates.

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) Distribution of Debt To Income Ratio: The provided histogram shows the distribution of Debt-to-Income (DTI) ratios.

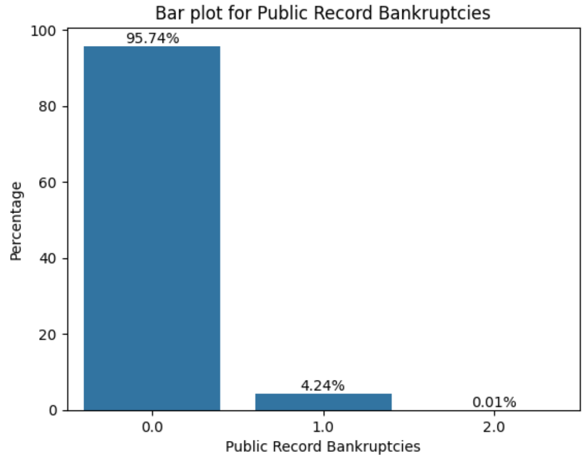
Conclusion: The histogram shows a right-skewed distribution of DTI ratios with a peak around 15-20, indicating that most individuals have DTI ratios in this range, but there is a significant portion with higher debt burdens.



1) Bar plot for Public Record Rankruptcies: The provided histogram shows the distribution of Interest rate.

Conclusion:

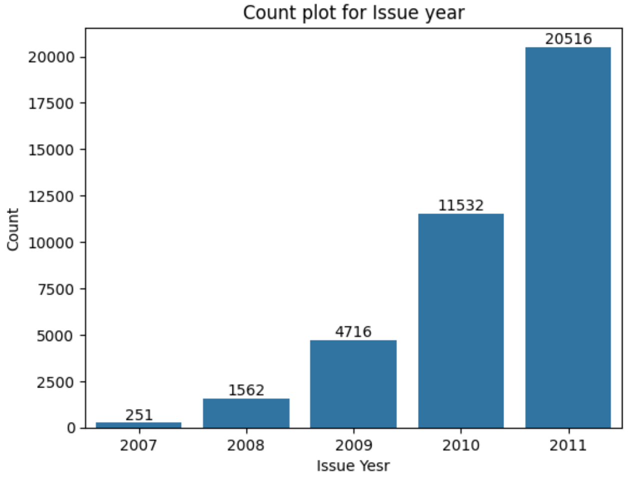
The bar plot shows the distribution of public record bankruptcies. The majority of individuals (95.74%) have no public record bankruptcies, while a small percentage (4.24%) have one bankruptcy, and an extremely small percentage (0.01%) have two or more bankruptcies. This suggests that most individuals have a clean credit history with respect to bankruptcies.



2) Count plot for Issue Year: The provided histogram shows the distribution of Debt-to-Income (DTI) ratios.

Conclusion:

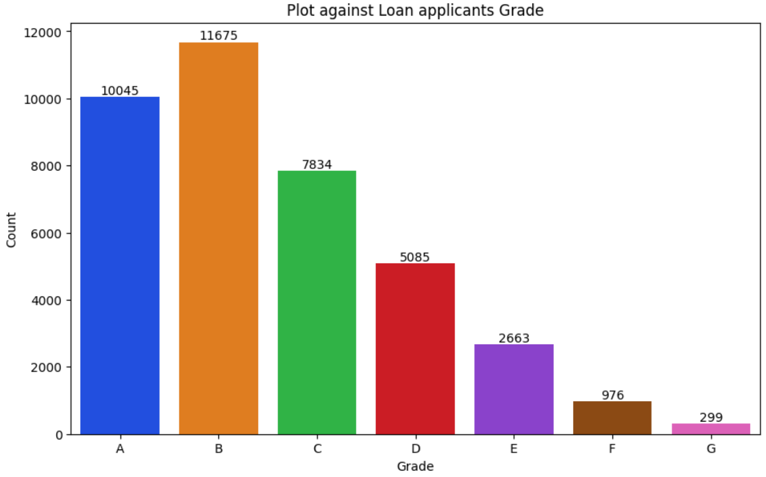
The count plot shows the number of loans issued each year from 2007 to 2011. The number of loans issued has generally increased over the years, with a significant jump from 2010 to 2011. This suggests that the demand for loans has grown over this period.



1) Bar plot for Loan applicants Grade: The provided histogram shows the distribution of Interest rate.

Conclusion:

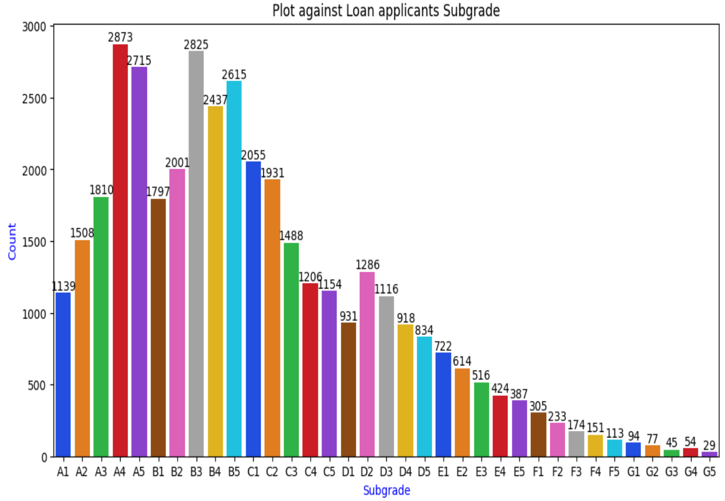
The bar plot shows the distribution of loan applicants by their grade. The majority of applicants have grades A, B, or C, indicating a relatively high creditworthiness. Grades D, E, F, and G represent lower creditworthiness, with significantly fewer applicants falling into these categories. This suggests that most loan applicants have good or fair credit histories.



2) Bar plot for loan applicants subgrade: The provided histogram shows the distribution of Debt-to-Income (DTI) ratios.

Conclusion:

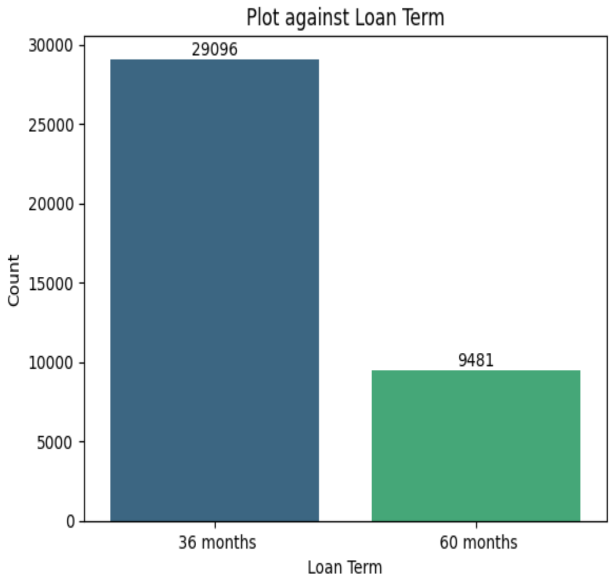
The bar plot shows the distribution of loan applicants by their subgrade. The subgrades are arranged from A1 to G5, representing decreasing creditworthiness. The plot reveals that most applicants fall into the A and B subgrades, indicating a relatively high creditworthiness. As the subgrades progress from C to G, the number of applicants decreases significantly, reflecting a lower creditworthiness. This suggests that most loan applicants have good or fair credit histories.



1) Bar plot for Loan Term:

Conclusion:

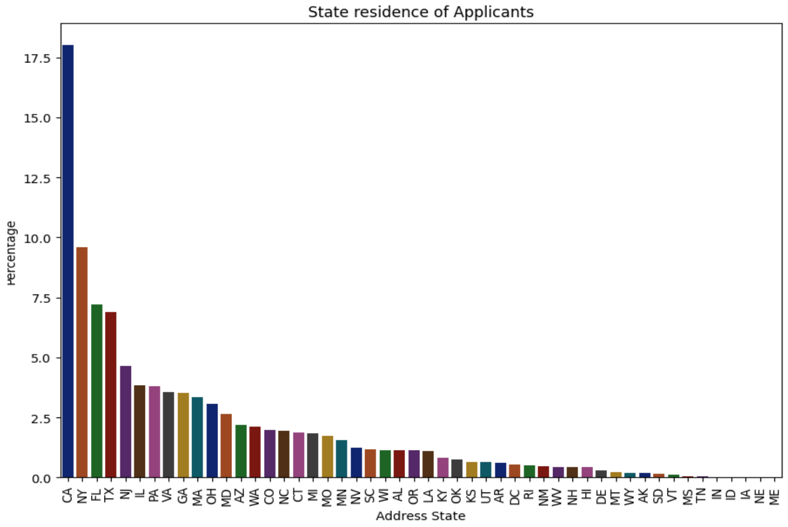
The bar plot shows the distribution of loan terms. The majority of loans have a term of 36 months (29,096 loans), while a smaller number have a term of 60 months (9,481 loans). This suggests that most borrowers prefer shorter loan terms.



2) Bar plot for state residence of applicants:

Conclusion:

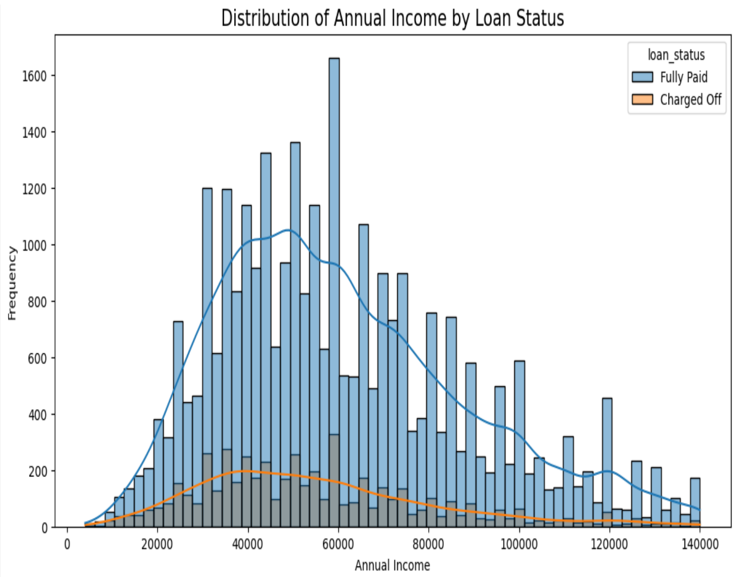
The bar plot shows the distribution of loan applicants by their state of residence. California has the highest percentage of applicants, followed by New York and Florida. The number of applicants from other states gradually decreases, with states like Wyoming, North Dakota, and South Dakota having the lowest percentages. This suggests that a significant portion of loan applicants reside in California, New York, and Florida.



1) Distribution of Annaul Income by loan status:

Conclusion:

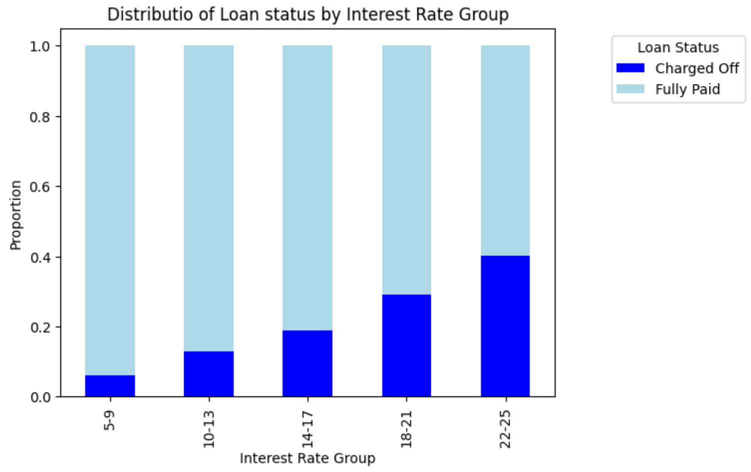
The plot shows the distribution of annual income for fully paid and charged-off loans. Higher income borrowers are more likely to fully pay their loans, while lower income borrowers are more likely to default. However, other factors also influence loan repayment.



2) Distribution of loan status by Interest rate:

Conclusion:

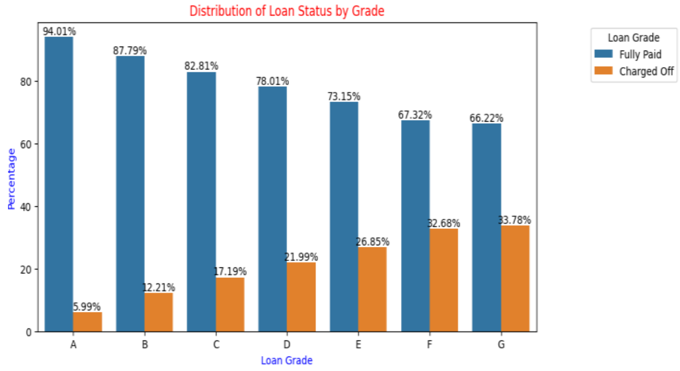
The plot shows the proportion of fully paid and charged-off loans within different interest rate groups. As the interest rate increases, the proportion of charged-off loans also increases, while the proportion of fully paid loans decreases. This suggests that borrowers with higher interest rates are more likely to default on their loans.



1) Distribution of Loan status by Grade:

Conclusion:

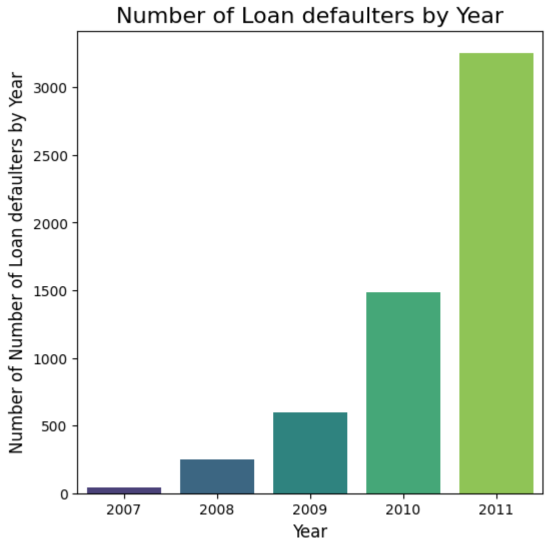
The plot shows the proportion of fully paid and charged-off loans for each loan grade. As the loan grade decreases from A to G, the proportion of charged-off loans increases, while the proportion of fully paid loans decreases. This suggests that borrowers with lower loan grades are more likely to default on their loans..



2) Number of loan defaulters by year:

Conclusion:

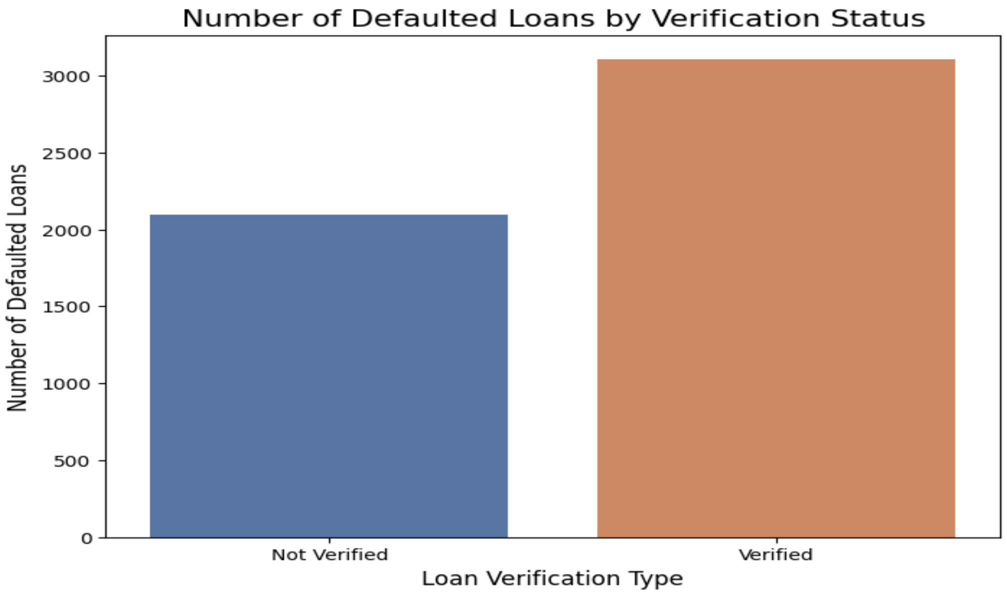
The bar plot shows the number of loan defaulters by year. The number of defaulters has increased from 2007 to 2011, with a significant jump from 2010 to 2011. This suggests that the rate of loan defaults has been rising over this period.



Bivariate Analysis:

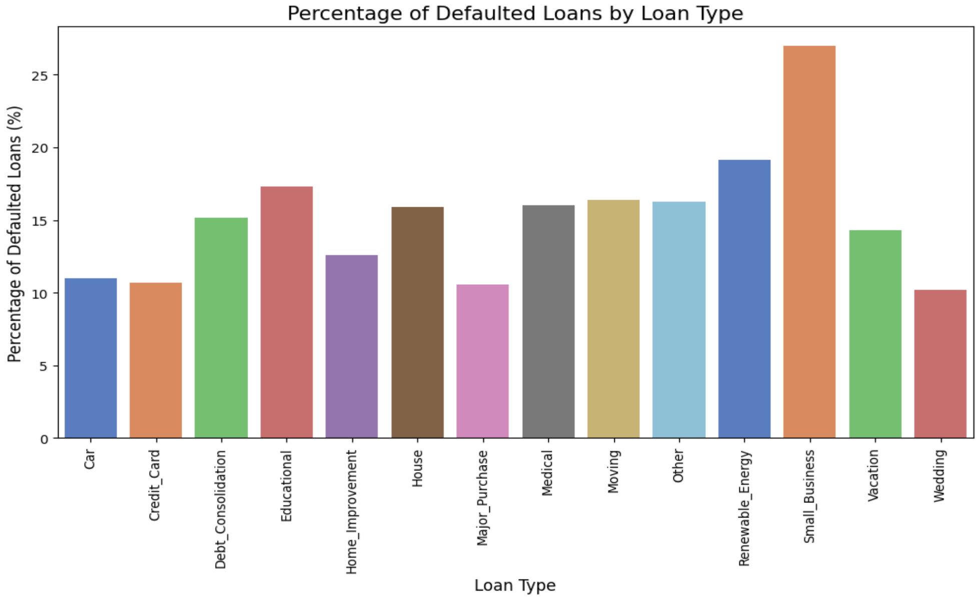
1) Analysis of Loan Default by Verification Status:
The bar plot shows the number of defaulted loans categorized by loan verification status, with 'Verified' and 'Source Verified' combined into a single 'Verified' category. The chart compares the number of defaulted loans for two verification statuses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

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



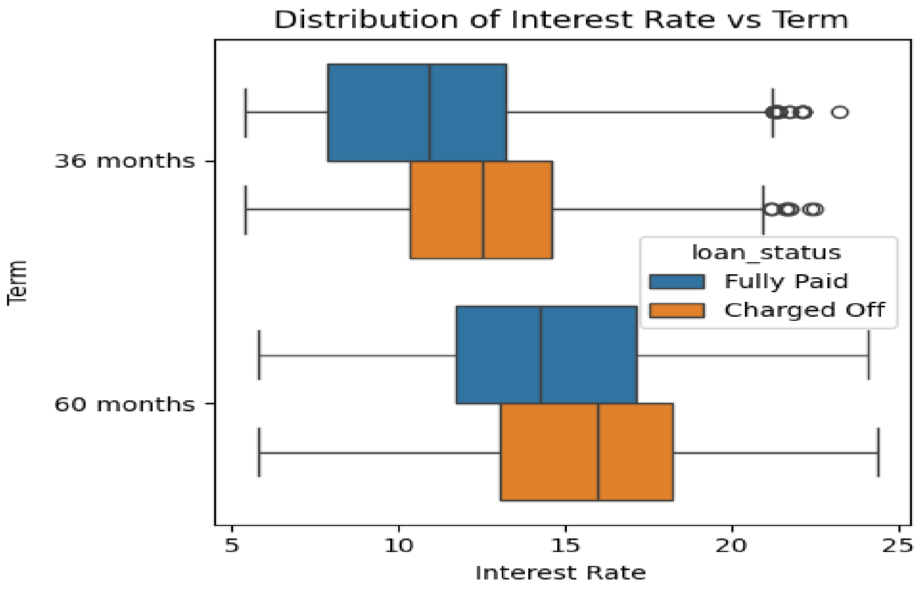
2) Analysis of Loan Defaults by Loan Purpose: The bar plot shows the percentage of loans that defaulted for each loan purpose. This helps us see which types of loans are more likely to result in defaults. Here we group the data by loan purpose and counts the number of defaults, then normalizes these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

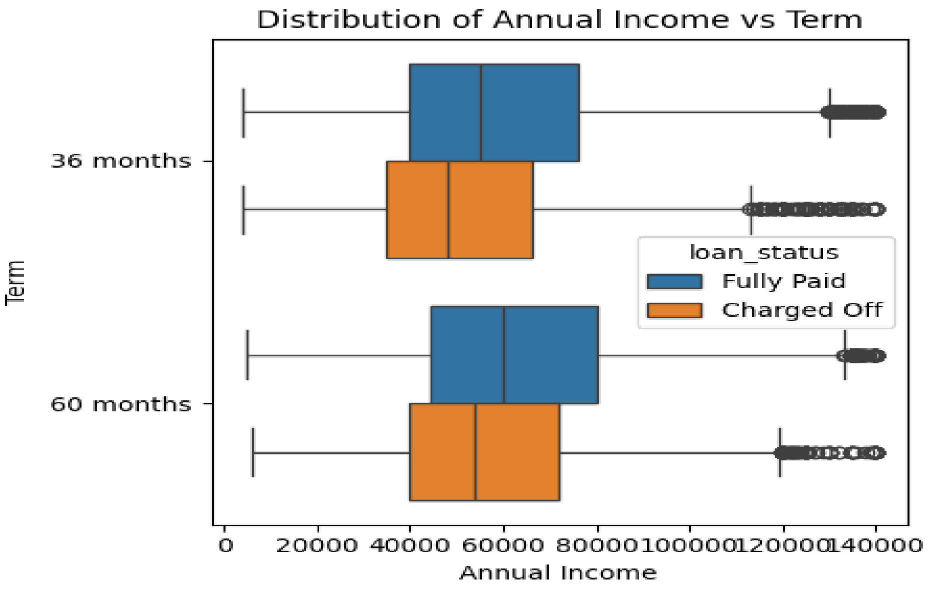
Conclusion: Based on our analysis the loan types, such as "Small Business" and "Wedding," have significantly higher default rates compared to others.



1) Distribution of IR Vs Term and Annual Income Vs Term:
Conclusion:

The box plots show that longer loan terms are associated with higher interest rates and lower incomes and are more likely to be charged off. Shorter loan terms tend to have lower interest rates, higher incomes, and are less likely to default. However, other factors also influence loan repayment.





Analysis of Loan Status by Homeownership:

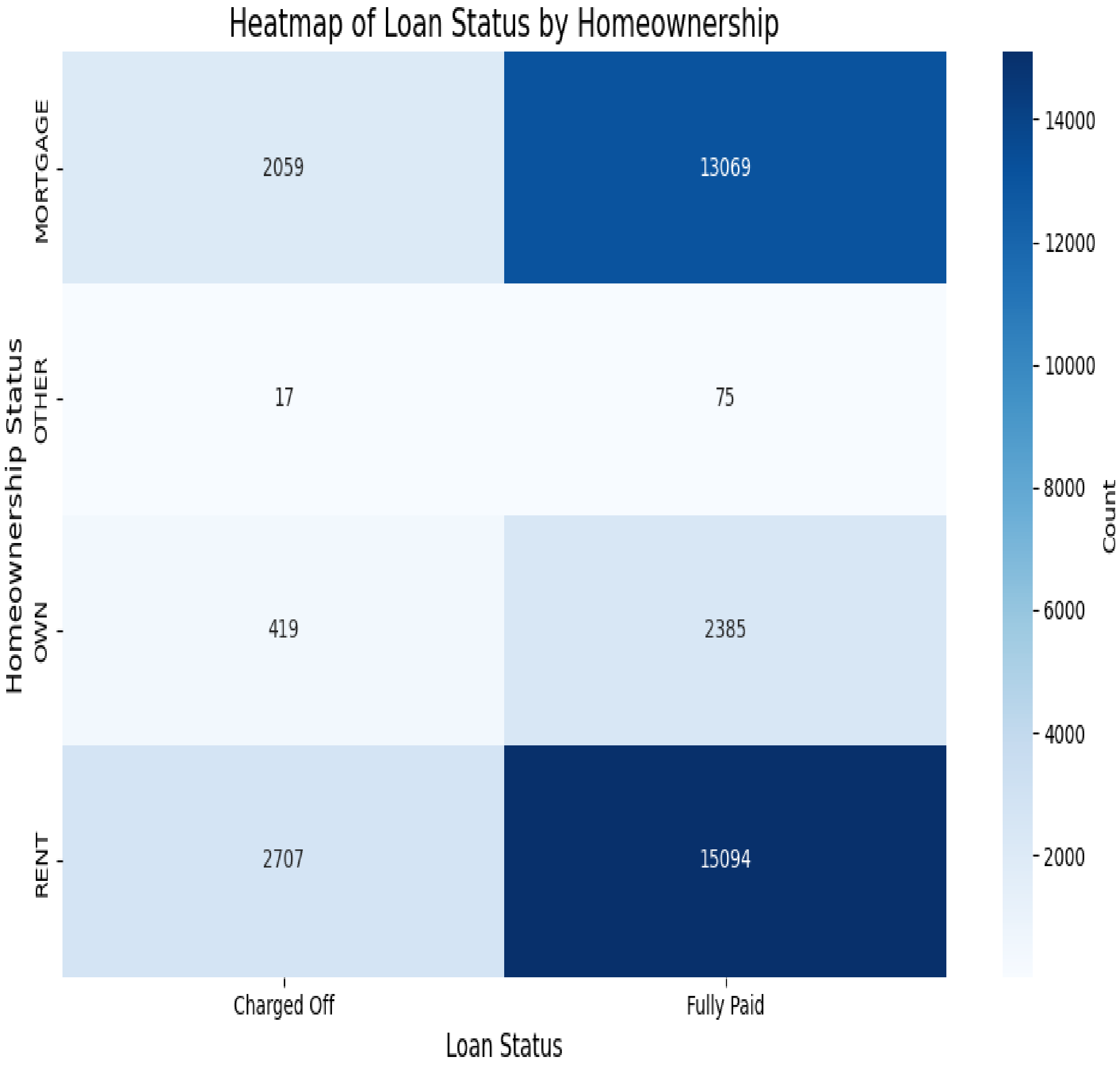
- First, we cleaned out the entries with "NONE" as the homeownership status.
- Grouped the data by homeownership and loan status
- Created Heat map with the result data.

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.

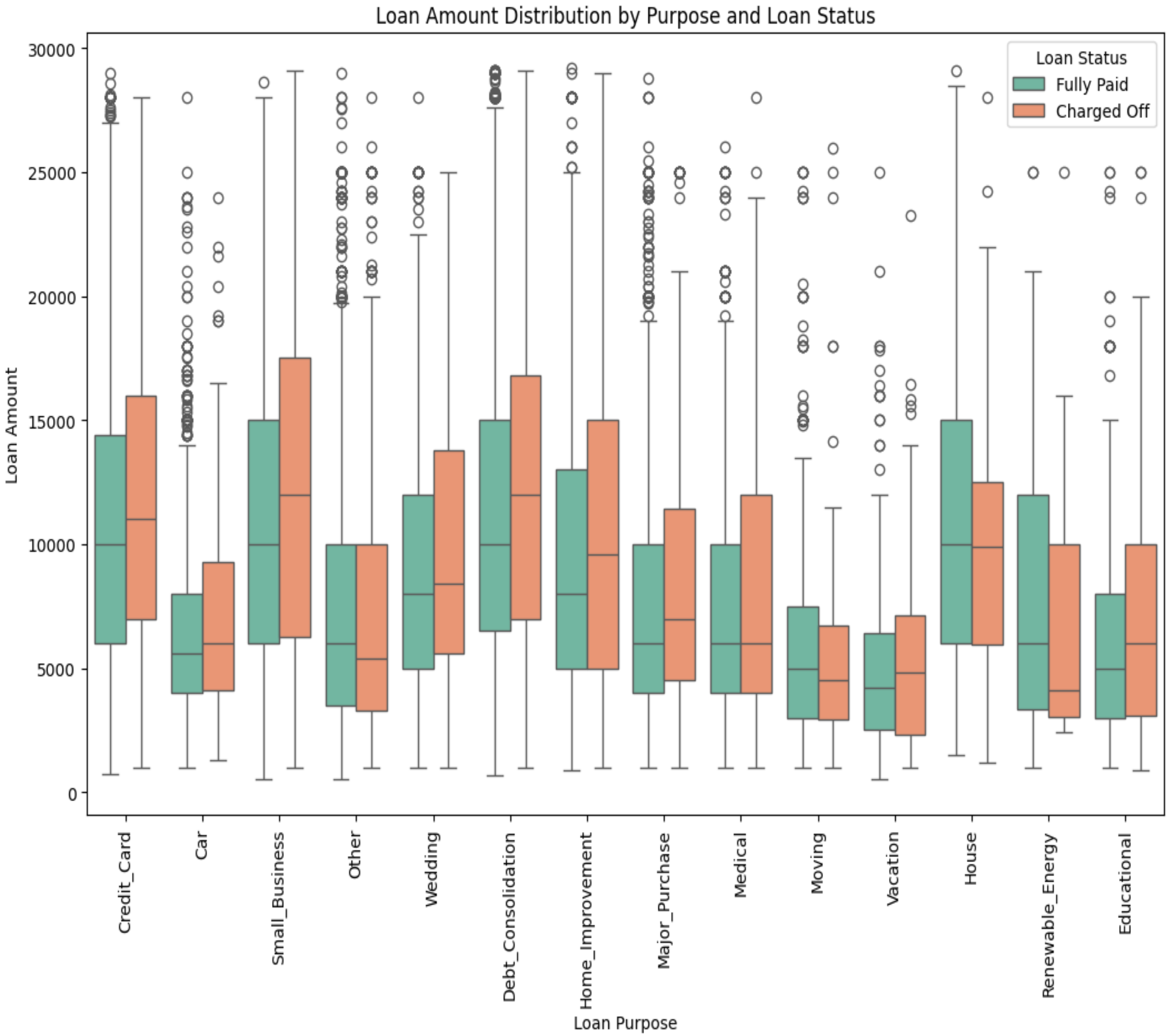
All this analysis helped us to understand how loan amounts vary by purpose and the differences between fully paid loans and defaults.
The rows removed for outliers are **836**.

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.

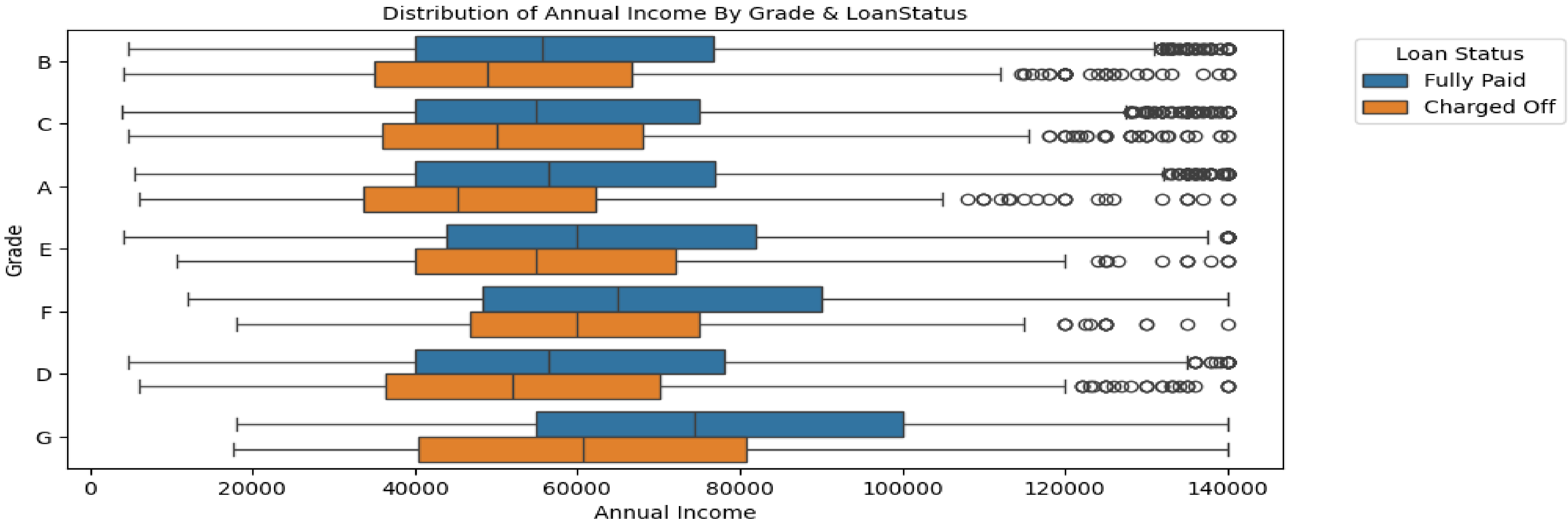
Based on analysis Charged-Off Loans are associated with Higher Amounts. The borrowers who default on loans tend to have taken out larger loans, especially for categories like small business and wedding.



Distribution of Annual Income by Grade & loan status:

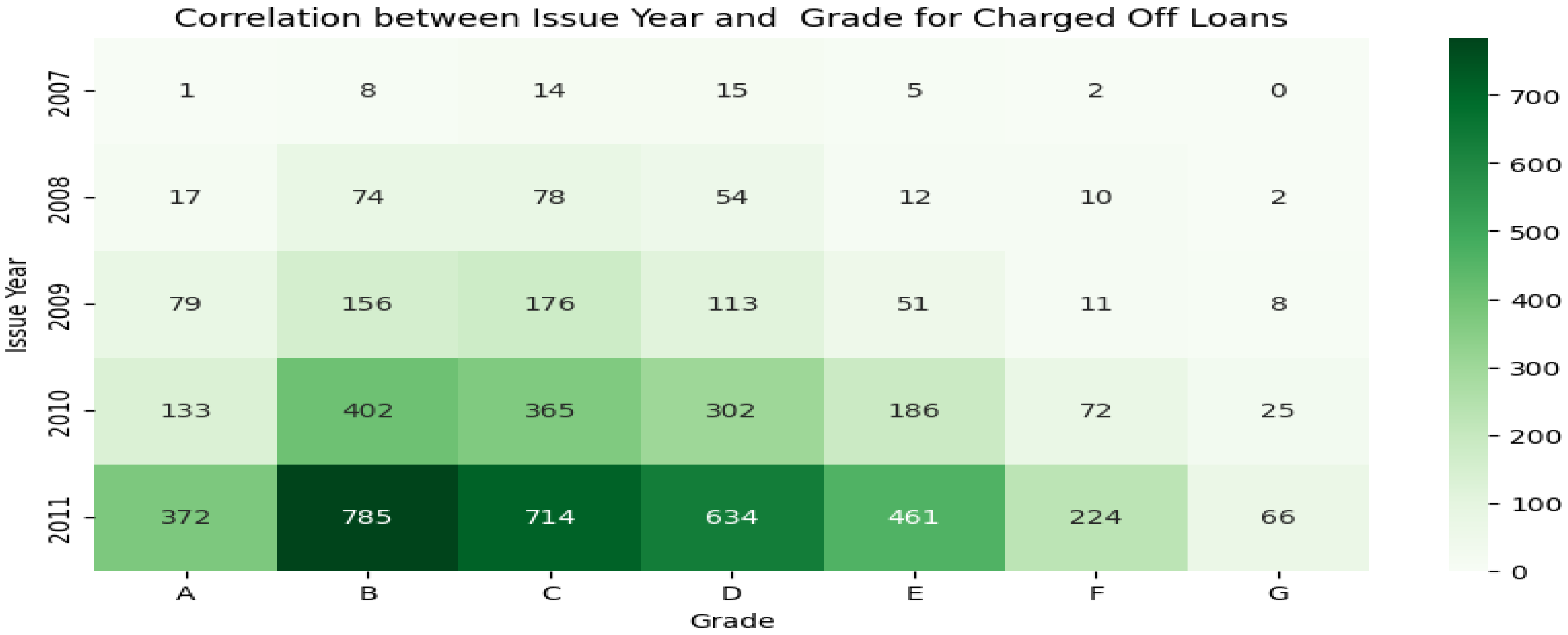
Conclusion:

The box plot shows the distribution of annual income by loan grade and loan status. Higher grades are associated with higher incomes and lower default rates. Lower grades tend to have lower incomes and higher default rates. However, other factors also influence loan repayment.



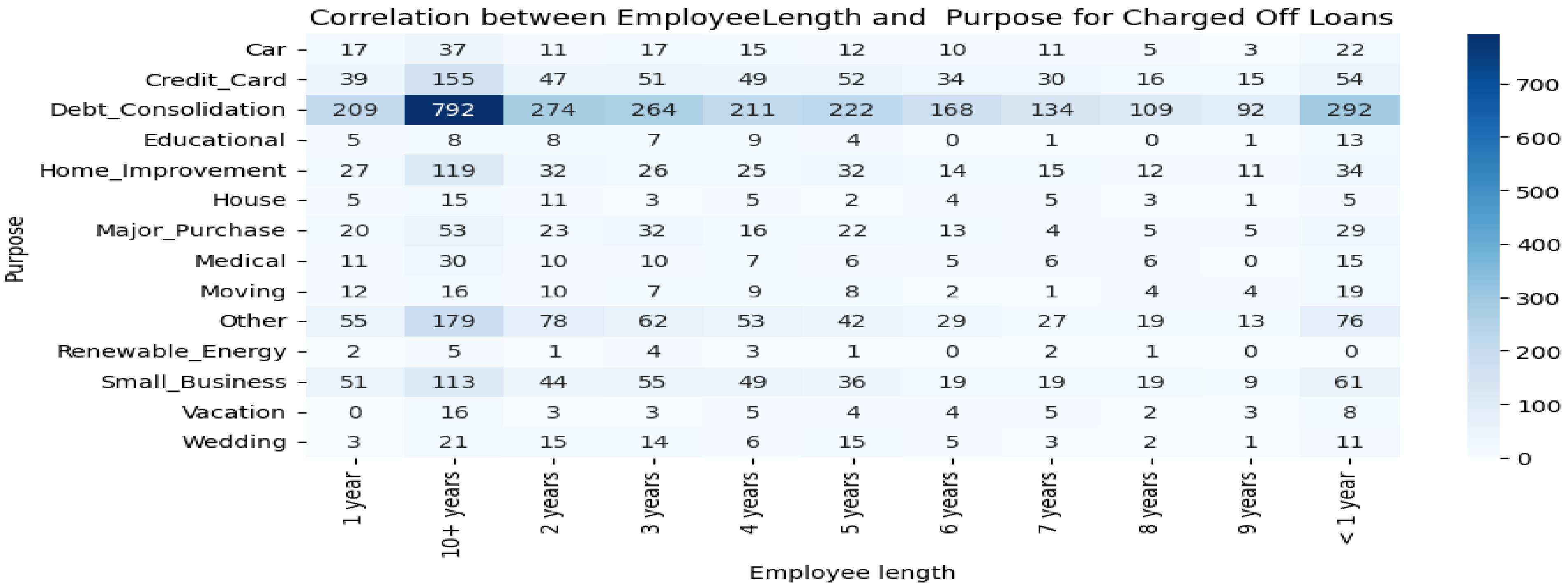
Correlation between Issue Year and Grade for Charged Off Loans:

The heatmap shows the correlation between issue year and grade for charged-off loans. Lower grades and later issue years are associated with higher default rates. However, other factors also influence loan repayment.



Correlation between Employee Length and Purpose for Charged Off Loans:

The heatmap shows the correlation between employee length and loan purpose for charged-off loans. Debt consolidation and credit card loans are associated with longer employee lengths and higher default rates. Other loan purposes, such as education and house, have lower default rates and are often associated with shorter employee lengths.



Correlation between EmployeeLength and Purpose for Charged Off Loans

Purpose	1 year	10+ years	2 years	3 years	4 years	5 years	6 years	7 years	8 years	9 years	< 1 year
Car	17	37	11	17	15	12	10	11	5	3	22
Credit_Card	39	155	47	51	49	52	34	30	16	15	54
Debt_Consolidation	209	792	274	264	211	222	168	134	109	92	292
Educational	5	8	8	7	9	4	0	1	0	1	13
Home_Improvement	27	119	32	26	25	32	14	15	12	11	34
House	5	15	11	3	5	2	4	5	3	1	5
Major_Purchase	20	53	23	32	16	22	13	4	5	5	29
Medical	11	30	10	10	7	6	5	6	6	0	15
Moving	12	16	10	7	9	8	2	1	4	4	19
Other	55	179	78	62	53	42	29	27	19	13	76
Renewable_Energy	2	5	1	4	3	1	0	2	1	0	0
Small_Business	51	113	44	55	49	36	19	19	19	9	61
Vacation	0	16	3	3	5	4	4	5	2	3	8
Wedding	3	21	15	14	6	15	5	3	2	1	11

Analysis of Average Annual Income by Loan Status and Purpose:

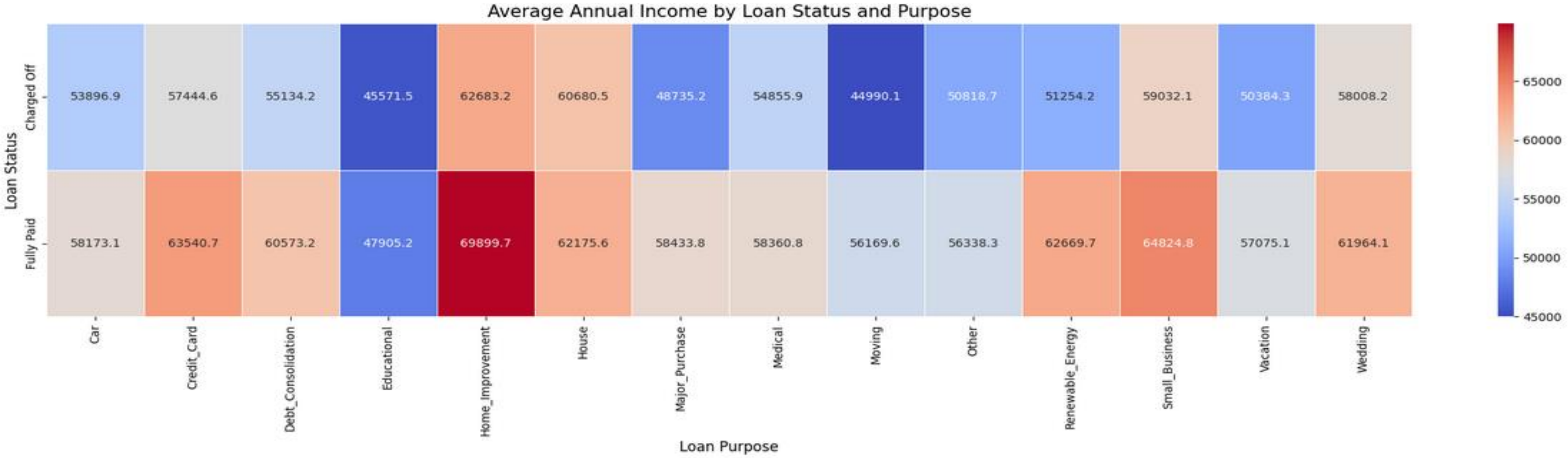
- First, we cleaned out the entries with "NONE" as the homeownership status.
- Grouped the data by homeownership and loan status and created Heat map with the result data.

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:

Interest rate is a strong predictor of loan default risk. Higher interest rates are associated with significantly increased charge-off rates.

Applicants income is a critical factor to do Loan payment. Lower-income individuals exhibit higher default rates.

Loan grade is a significant determinant of loan risk. Lower-grade loans (D, E, F, G) have a substantially higher defaults compared to higher-grade loans (A, B).

Debt-to-income (DTI) ratio is positively correlated with loan default risk. Higher DTI ratios indicate increased chances of charge-off.

Bankruptcy history is a strong indicator of elevated loan default risk.

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.

Loan Applicants with shorter employment lengths having higher interest rates are more likely to default on their loans.

Economic conditions can influence loan performance. In year 2011, we saw an increase in both loan volume and default rates

Loan grade is a strong indicator of loan performance. Higher-grade loans have lower default rates.

Bankruptcy history is likely correlated with loan default, as indicated by the relationship between grade and charge-off rates

Thank you