Project 1: Analysis Report

UCB Data Analytics Bootcamp

Group 2

Madison Bethke, Weibin He, Gursimran Kaur, Omid Khan, and Evan Wall

June 17, 2024

Project Description

The dataset examines the financial behavior of borrowers, analyzing various factors that influence their loan status.

This project aims to explore the relationships between various financial variables to understand factors influencing credit scores and loan outcomes. Specifically, it addresses two main questions:

1. Is there a significant relationship between annual income and credit score?
2. What factors influence the likelihood of a loan being charged off or fully paid?

Exploratory Visualizations

Histogram: Distribution of Credit Scores

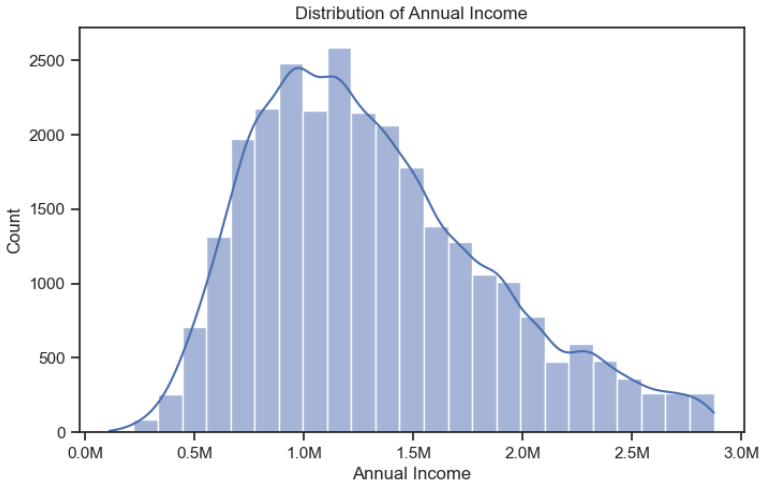
Summary Statistics of Credit Score

count	27981.000000
mean	712.315035
std	28.090851
min	585.000000
25%	699.000000
50%	718.000000
75%	734.000000

The histogram above illustrates the distribution of credit scores within the given sample. As described in the chart above, the data is not normally distributed; it is right-skewed with a long tail. The Kernel Density Estimate (KDE) visualizes the density of the credit score, indicating a smooth distribution to the left. The illustration highlights two peaks around the credit score range of 700-750, explaining two common ranges among the borrowers. The dataset consists of 27,981 credit scores, with an average credit score of 712.32, whereas most scores are between 700 and 735.

To gain a better understanding of the peaks in the histogram above, detailed visualizations will be generated for the credit score range of 710-750 to explore other variables.

Histogram: Distribution of Annual Income



Summary Statistics of Annual Income

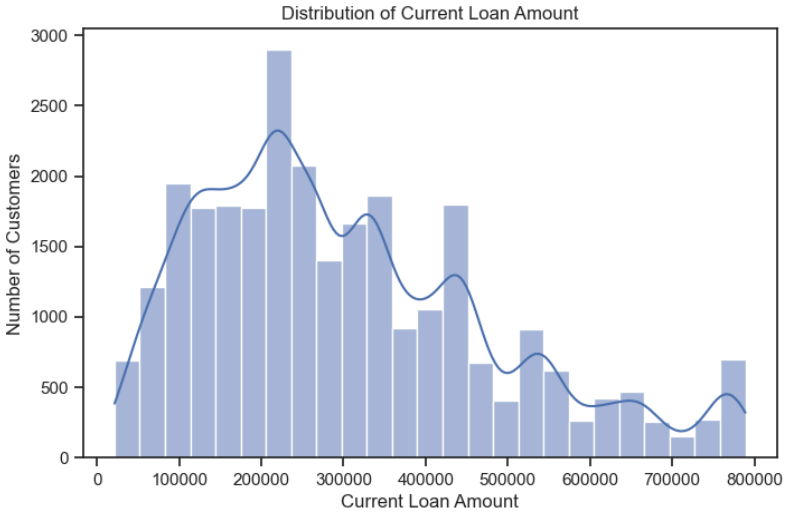
count	2.798100e+04
mean	1.316083e+06
std	5.406290e+05
min	1.112450e+05
25%	9.128170e+05
50%	1.228160e+06
75%	1.643975e+06
max	2.875080e+06

The illustration above illustrates the distribution of annual income, which is slightly right-skewed. This indicates that considerable borrowers with low to mid-range income levels and a small number of borrowers have an income level above two million dollars. The peak is around the one million dollar mark, the common income range for this dataset, with an average income level of 1.3 million dollars.

Histogram: Distribution of Current Loan Amount

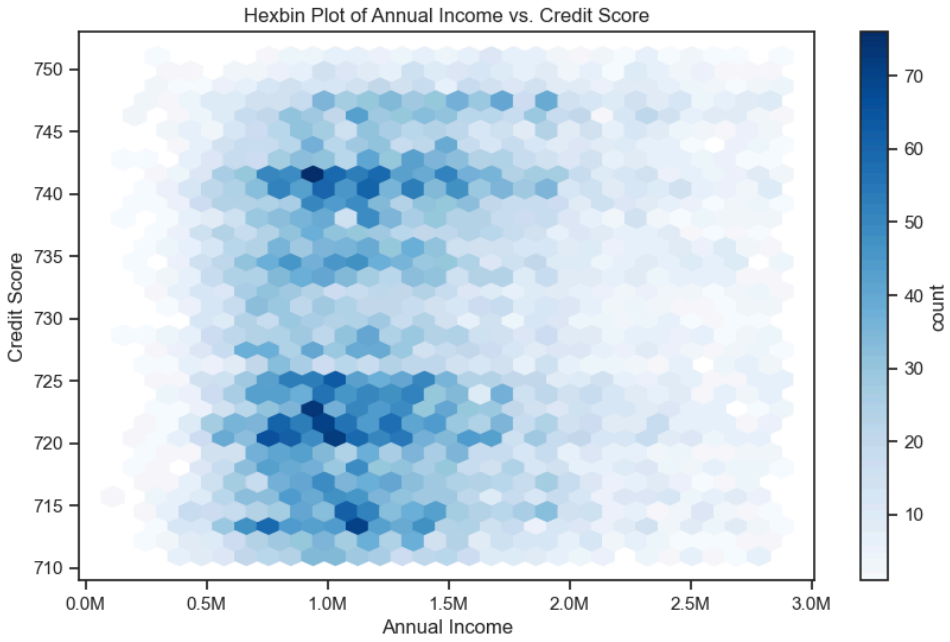
Summary Statistics of Current Loan Amount

count	27981.000000
mean	304442.308852
std	178844.419410
min	21450.000000
25%	171754.000000
50%	265782.000000
75%	417802.000000
max	789096.000000

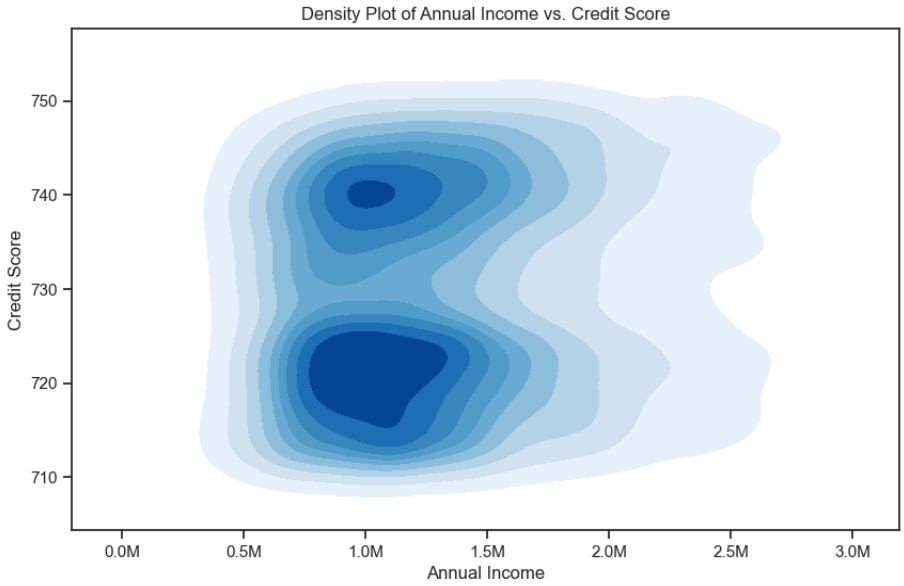


The histogram visualizes the distribution of the current loan amount within the sample dataset. As shown above, the distribution is right-skewed with several peaks, illustrating a high number of loan amounts being relatively low. The loan amount values range from a minimum of $21,450 to a maximum of $789,096. The highest peak is around $200,000, signifying a notable current loan amount around that peak. However, most borrowers have loan amounts between $100,000 and $300,000.

Hexbin Plot of Annual Income vs. Credit Score



Density Plot of Annual Income vs. Credit Score

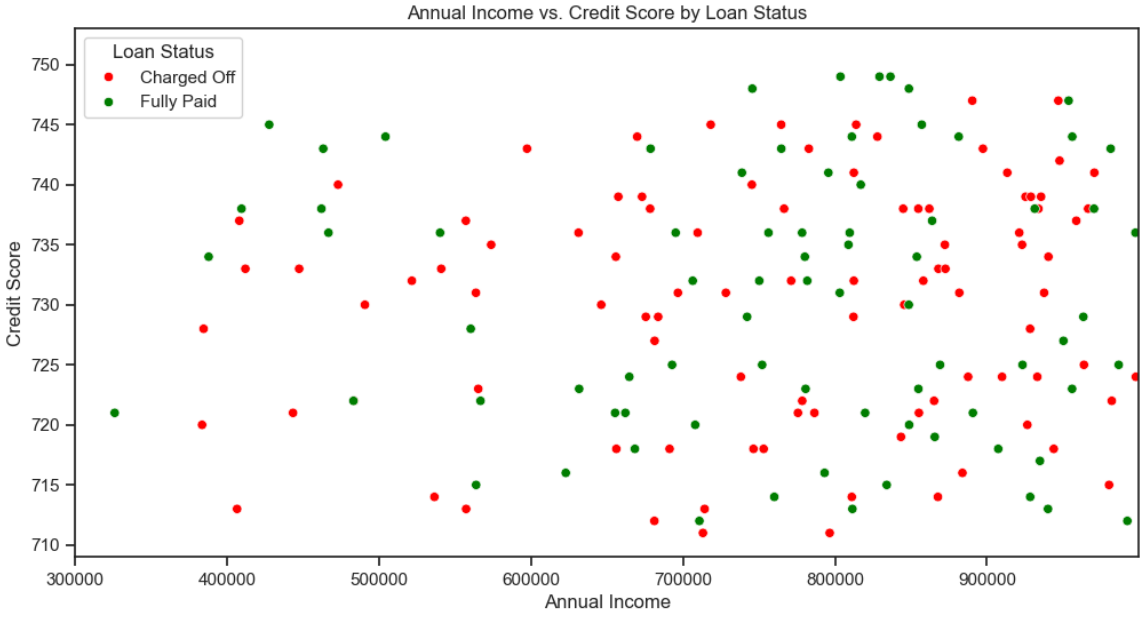


The hexbin and density plots show the relationship between annual income and credit scores. Annual income and credit scores are not uniformly distributed but rather have clustered around high-density areas. The color scale displays the density of the data points where darker areas indicate a higher concentration of data points, while lighter indicates a lower concentration when examining the relationship between the stated variables. Two clusters are visibly present with higher data density, falling between the annual income range of $500,000 to $2,000,000. However, the darkest data point is present between the credit range of 720-740 with a yearly income of **$1,000,000.**

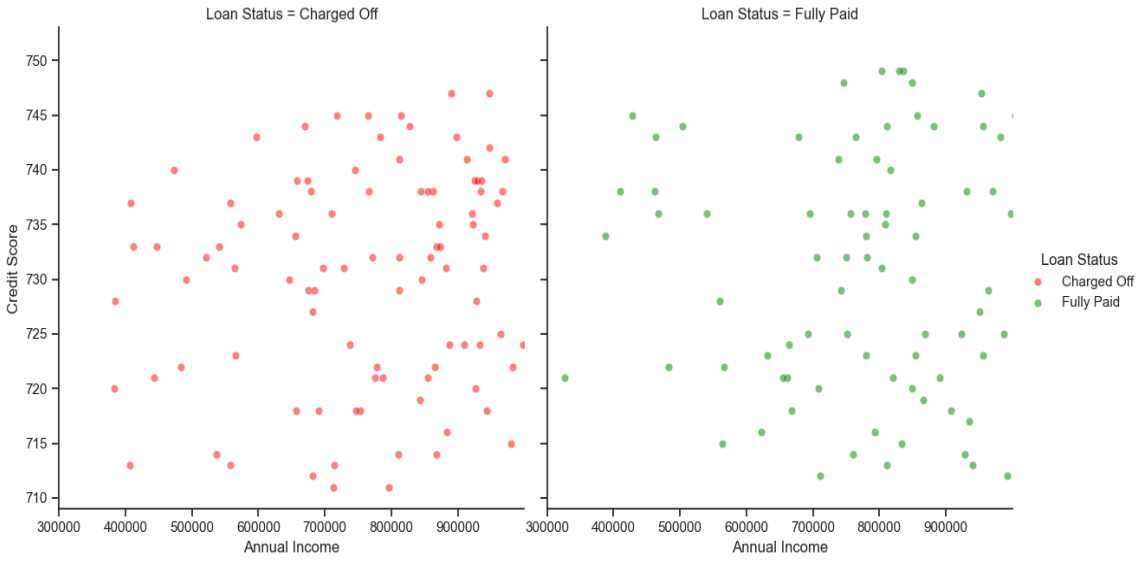
Hypothesis #1:

Null Hypothesis: There is no significant relationship between annual income and credit score.
Alternate Hypothesis: There is a significant relationship between annual income and credit score.

Scatter Plot: Annual Income vs Credit Scores



Scatter Plots: Separating 'Charged Off' and 'Fully Paid' Loan Status Data Points



The scatter plots display a random sample of 250 data points for each loan status category. Red markers represent ‘Charged Off’ data points, while the green represent ‘Fully Paid’ data points, showing no relationship between annual income and credit score. The second image of the scatter plot(s) clearly displays the difference between the different loan status categories. High annual income and credit score does not necessarily ensure that borrowers will make timely payments.

Correlation Analysis:

Charged Off - Pearson correlation coefficient: 0.08378435072441298

Charged Off - P-value: 0.1866912558834643

Fully Paid - Pearson correlation coefficient: 0.060551227688365516

Full Paide - P-value: 0.3403488889013202

The scatter plot shows a relationship between two loan statuses ('Charged Off and 'Fully Paid') with annual income and credit score. The correlation coefficient of 0.083 for 'Charged Off' loan status suggests a weak relationship because high annual income does not result in borrowers' likelihood to pay off their current loan amounts. The p-value of 0.187 is higher than 0.05, highlighting that the data points are not statistically significant.

T-test Analysis:

T-statistic: -0.9290

P-value: 0.3533

The t-test of -0.9290 shows the variation of the dataset where the average of both groups is relatively low. Meanwhile, the p-value of 0.3533 may support the null hypothesis.

Scatter Plot : Annual Income vs Credit Score vs Current Loan Amount

The scatterplot above focuses on a cluster of 'Charged Off' loan status data points from the second scatter plot, represented with red markers. This chart further explains the relationship between credit scores and annual income within the range of $700,000 and $1,000,000 while also considering the current loan amount for the borrowers. The size of the red markers symbolizes the current loan amount. As shown, most of the borrowers in the scatter plot have a current loan amount between $300,000 and $450,000, with only a few exceeding $450,000.

Correlation Analysis:

Correlation between Annual Income and Credit Score: 0.0701, **P-value:** 0.2694

Correlation between Annual Income and Current Loan Amount: 0.0932, **P-value:** 0.1418

Correlation between Credit Score and Current Loan Amount: -0.0022, **P-value:** 0.9727

The correlation coefficient shows a weak relationship between annual amount, credit score, and current loan amount. All p-values are greater than 0.05, signifying no statistical correlation. Therefore, the values have no significant relationship between any of the variables.

T-test Analysis:

T-statistic: -1.2846

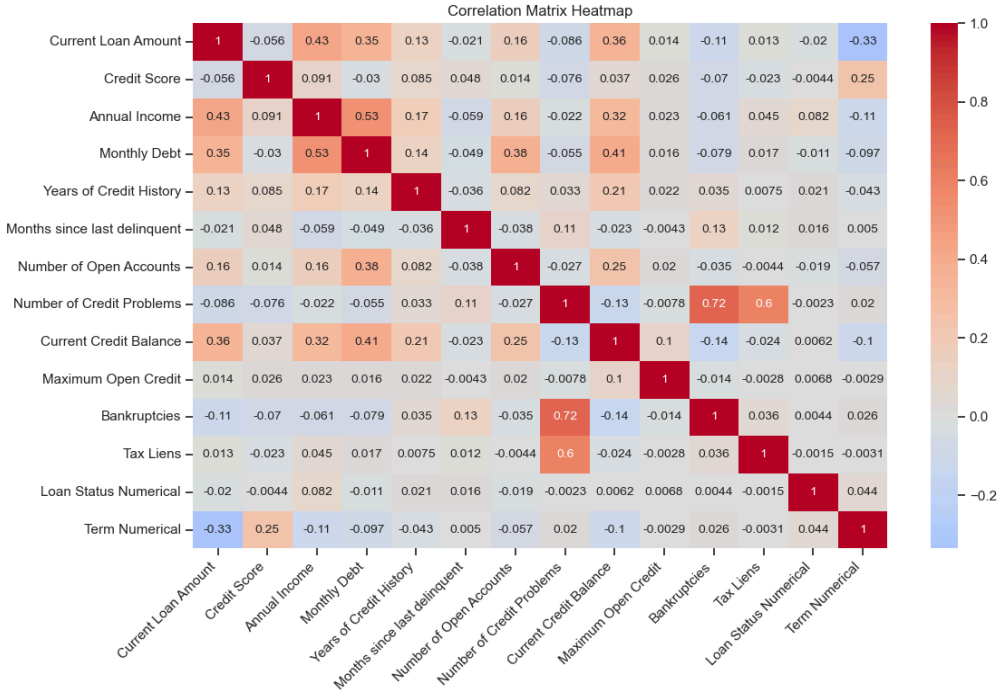
P-value: 0.2038

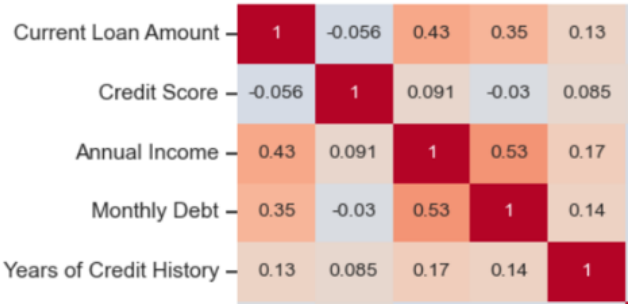
The negative t-test of -1.2846 signifies that the mean of the data point is low. The p-value of 0.2038 is higher than 0.05, meaning there is no statistically significant difference between the variables. Therefore the data **failed to reject the null hypothesis.**

Hypothesis #2:

Null Hypothesis: There is no relationship between annual income and monthly debt.
Alternate Hypothesis: There is a relationship between annual income and monthly debt.

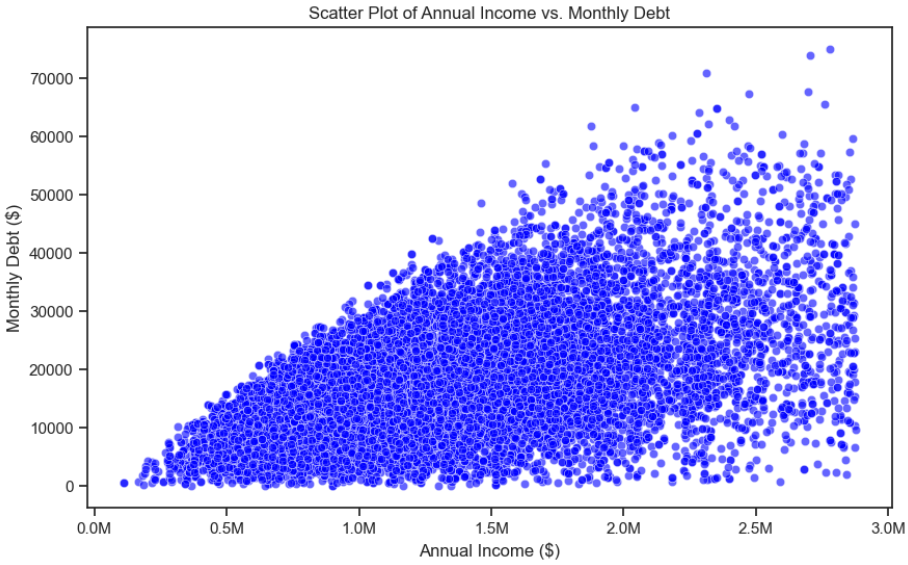
Correlation Matrix Heatmap





The correlation heat map visually represents the strength and direction of relationships between all the dataset variables. Values closer to 1 indicate a positive correlation, while values around -1 indicate a negative correlation. A value of zero indicates no correlation. The second snippet of the heat map exhibits high-density relationships, specifically concentrating on the correlation of 0.53 between monthly debt and annual income for all borrowers, which suggests a strong positive connection between both variables.

Scatter Plot: Annual Income vs Monthly Debt



The scatter plot exhibits the relationship between monthly debt and annual income for all borrowers in the dataset. The blue markers display a clear positive correlation between both variables: monthly debt also tends to increase as yearly income increases. The lower bracket (< $1,500,000) has a higher density of data points, with most borrowers having a monthly debt lower than $40,000. As annual income exceeds $1,500,000, monthly debt consistently becomes larger and more spread out. In conclusion, borrowers with higher annual incomes are more likely to have higher monthly debts.

Scatter Plot: Annual Income vs Monthly Debt (Random Sample of 500)

Correlation Analysis:

Pearson correlation coefficient: 0.483, **P-value:** 1.58e-30

The Pearson correlation coefficient of 0.483 suggests a moderate positive linear relationship between annual income and monthly debt, highlighting that when one variable increases, so does the other, and vice versa. The p-value of 1.58e-30 is extremely small compared to 0.05, explaining that there is no relationship between the variables. Therefore, the data rejected the null hypothesis and accepted the alternative hypothesis, which states that there is a relationship between annual income and monthly debt. In conclusion, borrowers with higher income levels tend to have higher monthly debts.

T-test Analysis:

T-statistic: -8.4779

P-value: 2.62e-16

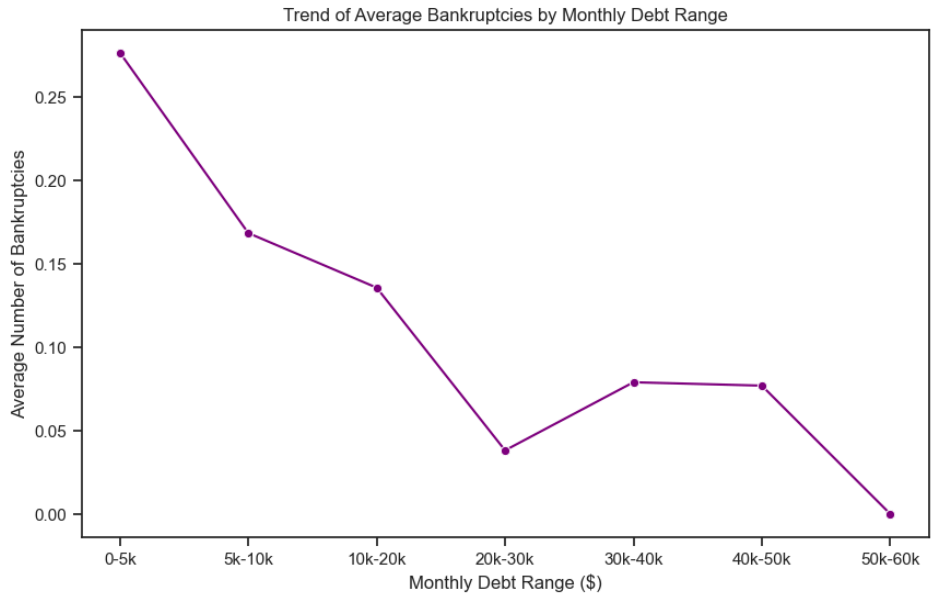
The t-test value of -8.4779 shows a considerable difference between the two variables, which is extremely low. The significantly low p-value of 2.62e-16 emphasizes that the data is statistically significant where the probability of chance is low. Therefore, the data **rejected the null hypothesis** and concluded that there is a positive linear relationship between annual income and monthly debt.

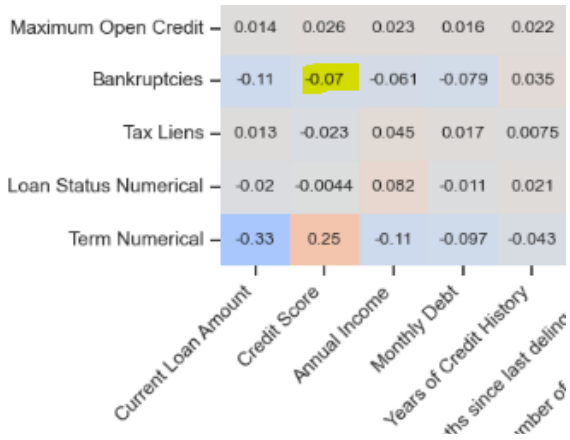
Hypothesis #3

Null Hypothesis: There is no correlation in the average number of bankruptcies and monthly income across different average annual income ranges.

Alternate Hypothesis: There is a correlation in the average number of bankruptcies and monthly income across different average annual income ranges.

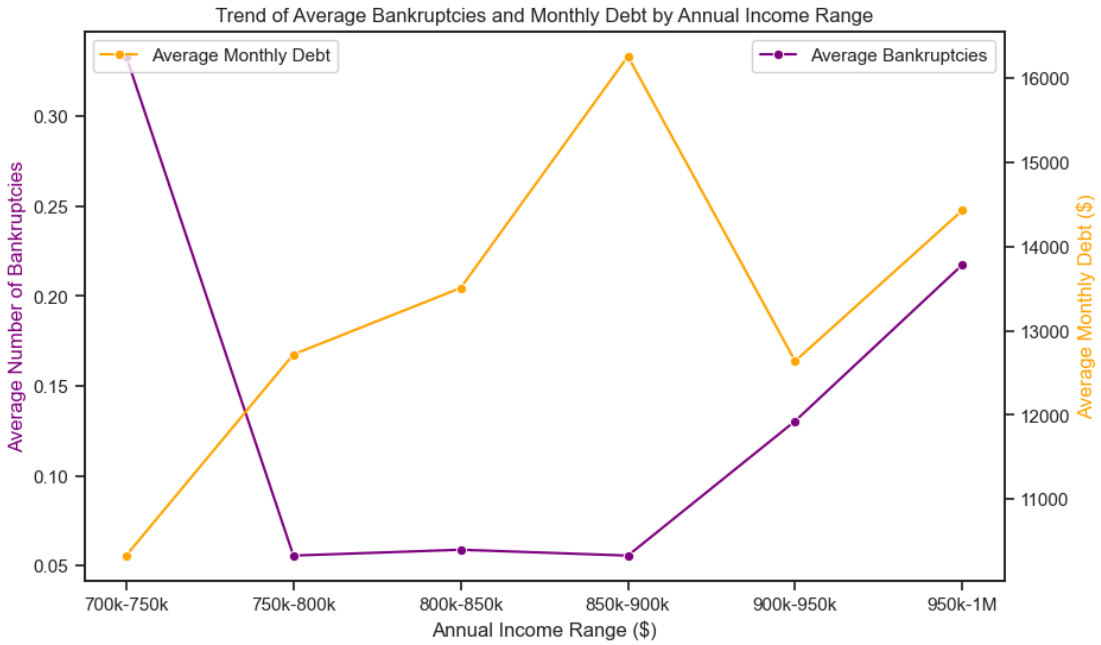
Line Chart: Trend of Average Bankruptcies by Monthly Debt Range





The line chart illustrates the correlation between average bankruptcies across different monthly debt ranges. According to the heat map on the side, the correlation value between both variables is -0.07, signifying a negative correlation. Moreover, the chart shows an inverse trend; as monthly debt increases, the probability of the borrower filing for bankruptcy decreases. On average, borrowers with monthly debt below $50,000 have a high risk for bankruptcy. Borrowers with a monthly debt of $30,000 or higher have a low possibility of bankruptcy. Therefore, borrowers with large amounts of monthly debt have a better sense of financial management or have high income levels (explored in the next visualization).

Multiple Line Chart: Trend of Average Bankruptcies and Monthly Debt by Annual Income Range



ANOVA Test:

ANOVA for Bankruptcies:

F-statistic = 1.6678109788790043 P-value = 0.14880606102897884

ANOVA for Monthly Debt:

F-statistic = 1.5036083272136924 P-value = 0.19496902072674618

The multiple-line chart illustrates the tendency of average bankruptcies and monthly debt across different annual income ranges from a random sample of 500 data points. The purple line symbolizes the average number of bankruptcies, which drastically decreased by the 750-800k income bracket but remained steady until the 850k-900k bracket. However, the orange line symbolizing monthly debt has a positive slope initially, indicating a positive correlation between annual income and monthly debt. From the ANOVA test, bankruptcies and monthly debt have a p-value higher than the alpha value of 0.05, emphasizing no significant difference between the values across annual income ranges. Changes in annual income does not influence the number of bankruptcies borrowers are likely to file or their average monthly debt. However, borrowers with an income of $950,000 or higher are more probable to have high monthly debt and an increased risk of bankruptcy. Therefore, it can be concluded that the data **failed to reject the null hypothesis.**

Conclusion

Question 1: The analysis found no significant relationship between annual income and credit score. High credit scores and annual income do not determine a borrower's ability to repay the loan.

Question 2: Several factors influence the likelihood of a loan being charged off or fully paid. While the correlation analysis did not find significant relationships between the variables studied (annual income, credit score, and loan amount), the analysis of income and debt found a significant positive relationship.