# Project 1: Analysis Report

Group: 2

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
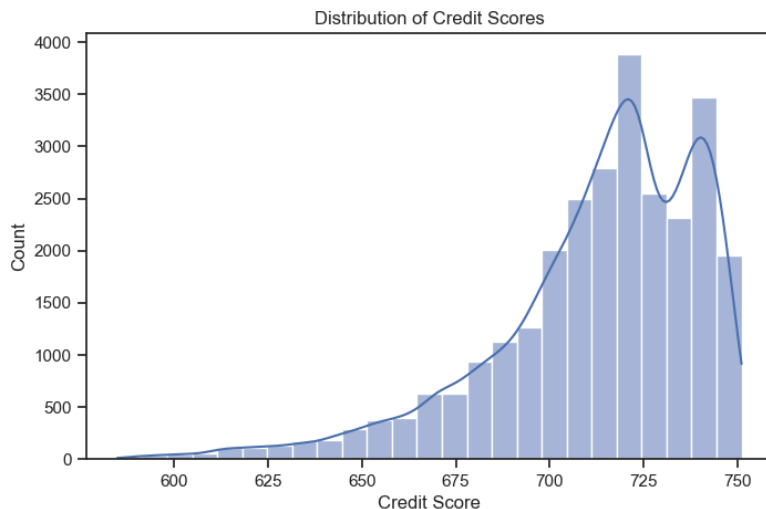
Summary Statistics of Credit Score

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.

Summary Statistics of Annual Income

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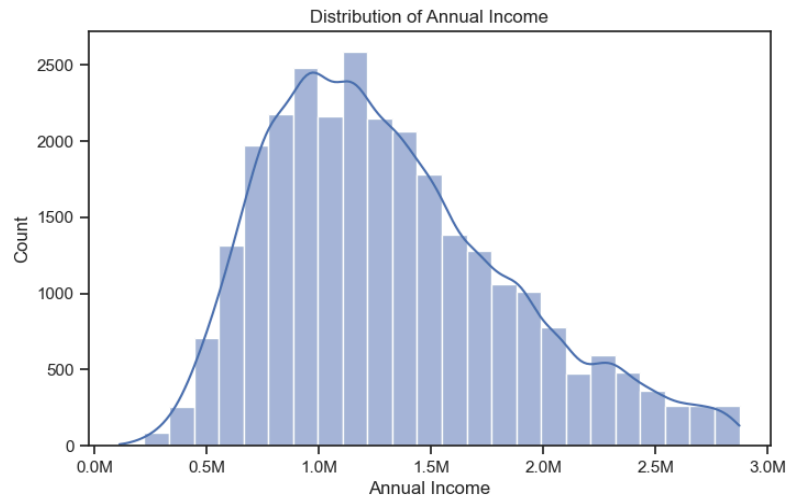
The illustration above illustrates the distribution of annual income, which is slightly right-skewed. This indicates that considerable borrowers with low to mid-range income levels and a small number of borrowers have an income level above two million dollars. The peak is around the one million dollar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

Summary Statistics of Current Loan Amount

Distribution 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
```
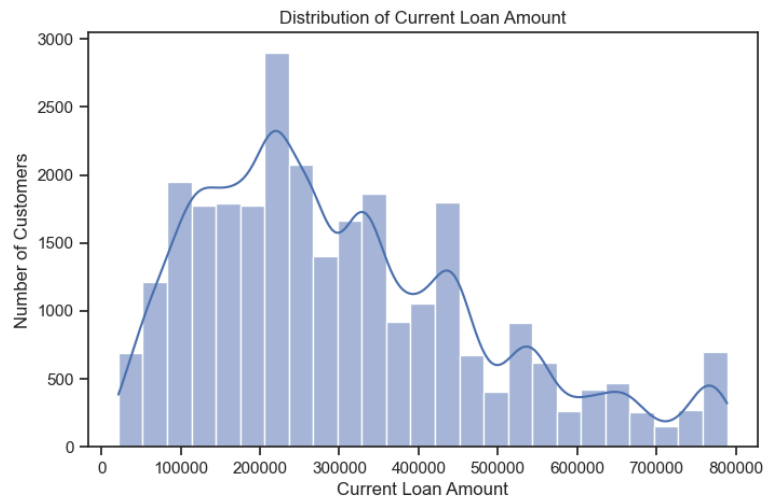
The histogram visualizes the distribution of the current loan amount within the sample dataset. As shown above, the distribution is right-skewed with several peaks, illustrating a high number of loan amounts being relatively low. The loan amount values range from a minimum of $21,450 to a maximum of $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.
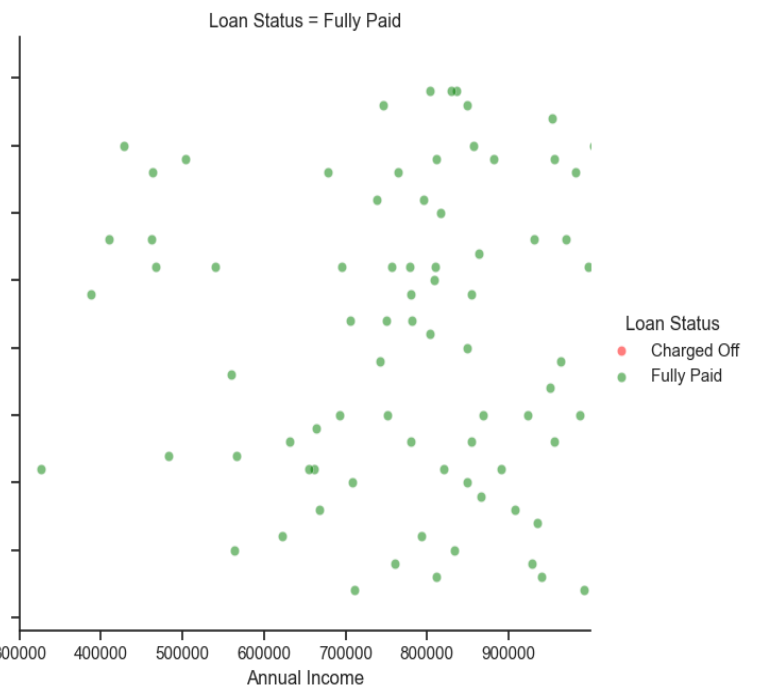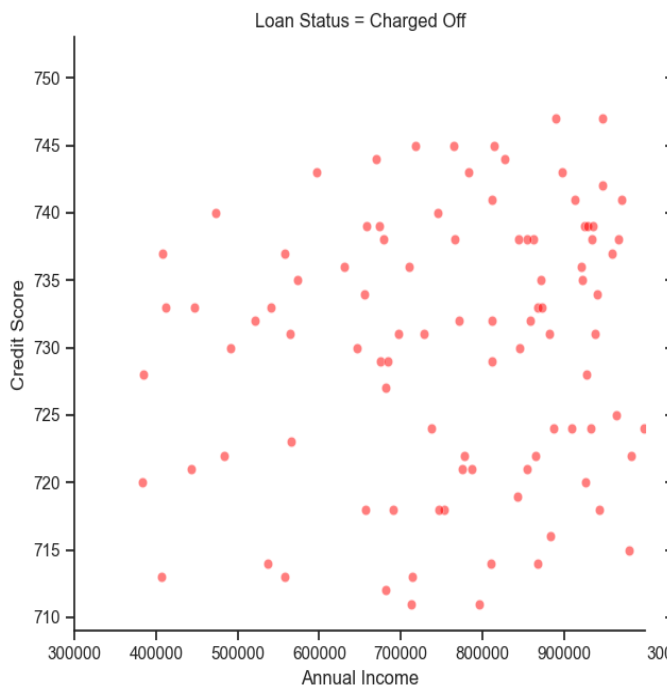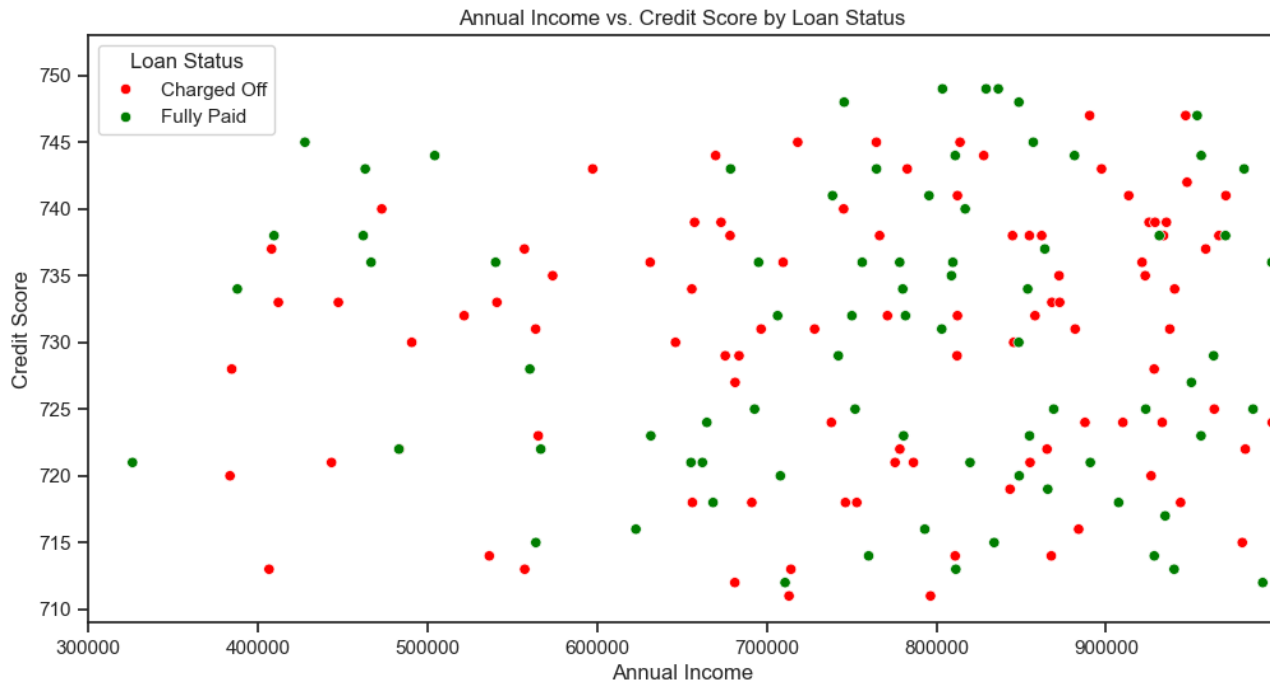
Hexbin Plot of Annual Income vs. Credit Score
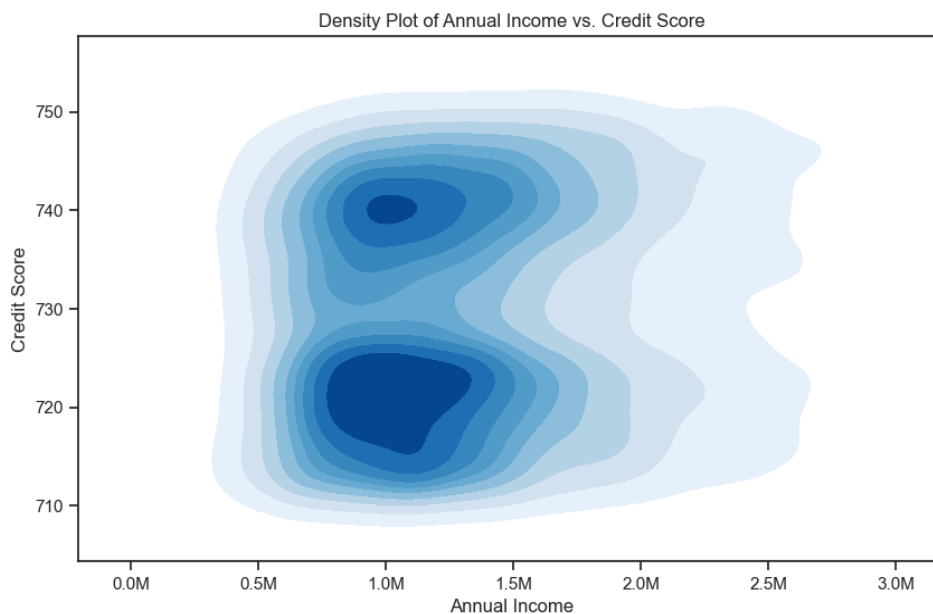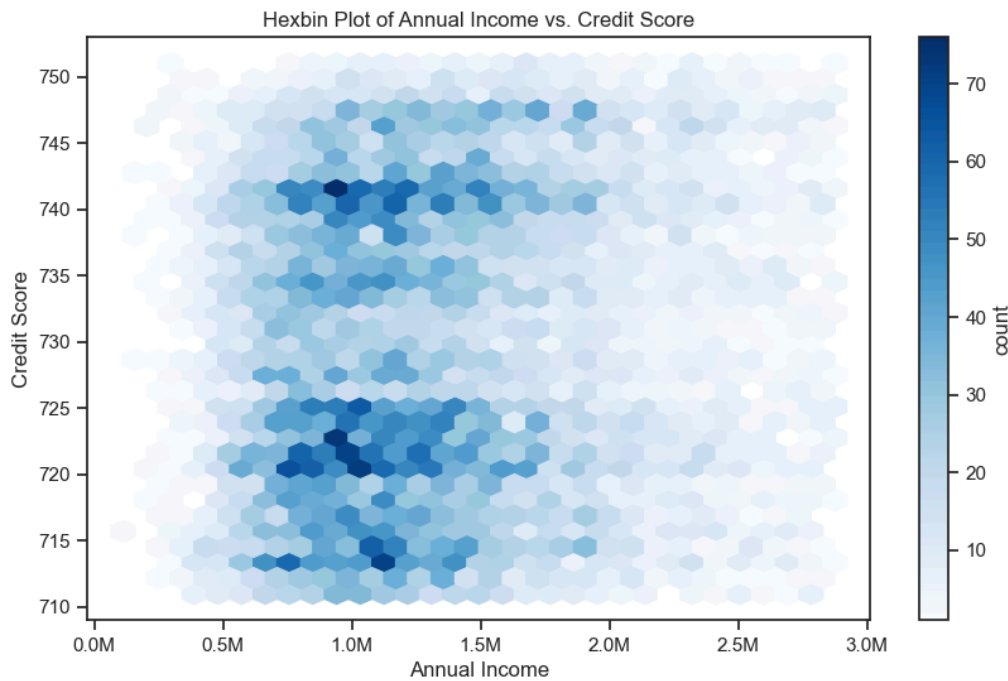
The Hexbin and density plots show the relationship between annual income and credit scores. Annual income and credit scores are not uniformly distributed but rather have clustered around high-density areas. The color scale displays the density of the data points where darker areas indicate a higher concentration 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 0.5 million dollars and 2 million dollars. However, the darkest data point is present between the credit range of 720-740 with a yearly income of **one million dollars**.

Density Plot of Annual Income vs. Credit Score

Hypothesis #1:

Null Hypothesis (H0): There is no significant relationship between annual income and credit score.

Alternate Hypothesis (H1): There is a significant relationship between annual income and credit score.

Annual Income vs. Credit Score by Loan Status

Annual Income vs Credit Score Scatter Plots

Correlation Analysis:

- Charged Off - Pearson correlation coefficient: 0.08378435072441298
- Charged Off - P-value: 0.1866912558834643
- Fully Paid - Pearson correlation coefficient: 0.060551227688365516
- Fully Paid - 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.

Annual Income vs. Credit Score by Loan Status and Loan Amount

Annual Income vs Credit Score vs Current Loan Amount Scatter Plot

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.

Hypothesis #2:

Null Hypothesis (H0): There is no relationship between annual income and monthly debt.
Alternate Hypothesis (H1): There is a relationship between annual income and monthly debt.
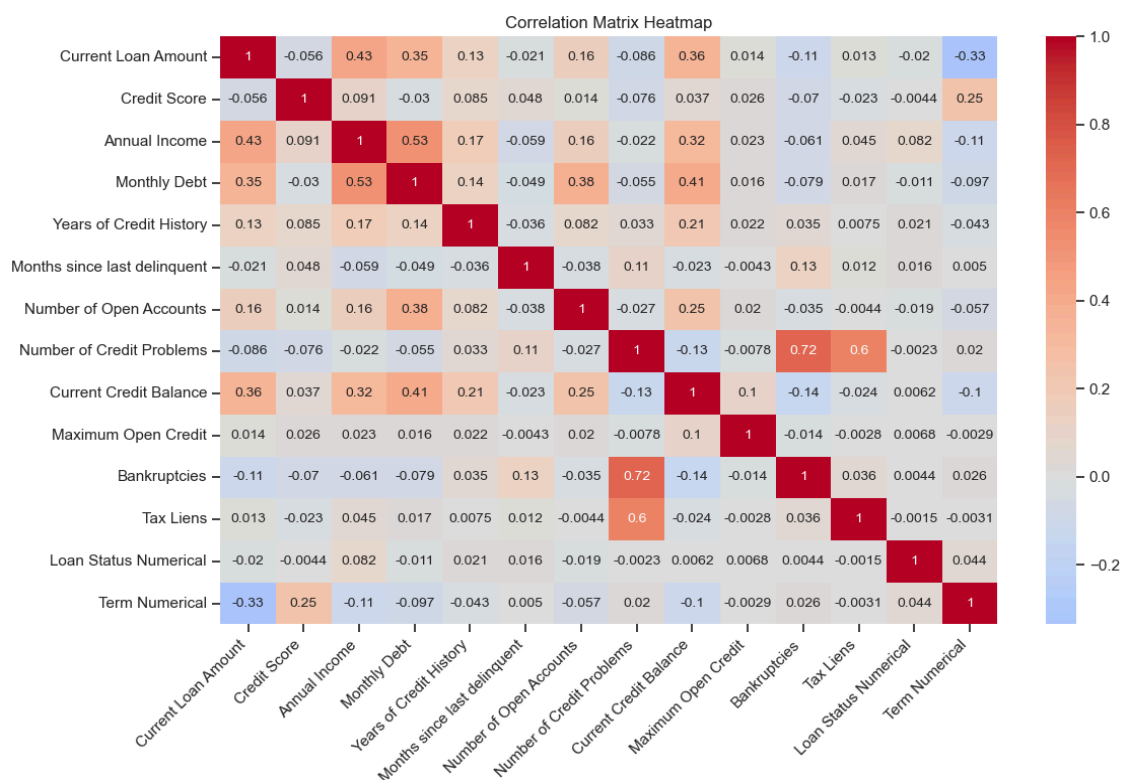
Scatter Plot of Annual Income vs. Monthly Debt

Annual Income vs Monthly Debt Scatter Plot

<u>Correlation Analysis:</u>

- **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.

Conclusion

Question 1: The analysis found no significant relationship between annual income and credit score.

Question 2: The likelihood of a loan being charged off or fully paid is influenced by several factors. 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.