BHARATIYA VIDYA BHAVAN’S

SARDAR PATEL INSTITUTE OF TECHNOLOGY

(Empowered Autonomous Institute Affiliated to University of Mumbai)

[Knowledge is Nectar]

Advance Data Visualization

UID	2021300010
Name	Kunal Bhatia
Batch	Batch G
Aim	Experiment Design for Creating Visualizations using D3.js on a Finance Dataset

Objectives:

- To explore and visualize a dataset related to Finance/Banking/Insurance/Credit using D3.js.
- To create basic visualizations (Bar chart, Pie chart, Histogram, Timeline chart, Scatter plot, Bubble plot) to understand data distribution and trends.
- To create advanced visualizations (Word chart, Box and Whisker plot, Violin plot, Regression plot, 3D chart, Jitter) for deeper insights and complex relationships.
- To perform hypothesis testing using the Pearson correlation coefficient to evaluate relationships between numerical variables in the dataset.

Dataset: Health Insurance Dataset

Link: Credit Risk Dataset

About Dataset:

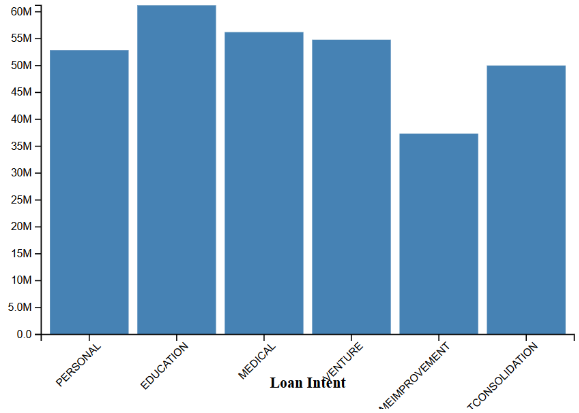
The dataset contains loan application records, detailing borrower demographics and loan characteristics, including loan purpose, amounts, income percentages for repayment, ages, annual incomes, and interest rates, useful for analyzing borrowing trends and borrower profiles.

1. **loan_intent**: Indicates the purpose of the loan, such as personal, business, or education.
2. **loan_amnt**: Represents the total amount of the loan requested by the borrower.
3. **loan_percent_income**: Shows the percentage of the borrower’s income used to repay the loan.
4. **person_age**: Indicates the age of the borrower.

5. **person_income**: Represents the annual income of the borrower.
6. **interest_rate**: Shows the annual interest rate applied to the loan.

Basic Visualizations:

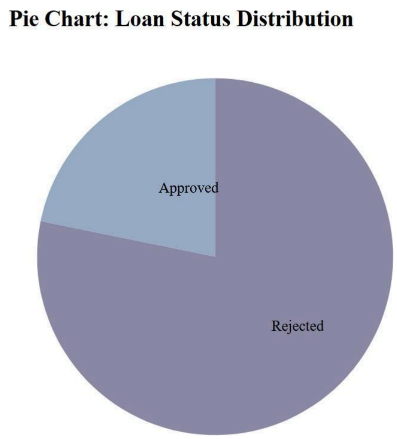
Bar Plot: Loan Amount by Loan Intent



Interpretation of the Bar Plot (Loan Amount by Loan Intent)

Key Points:

- **Loan Intent:** The plot compares the loan amounts for different loan intents.
- **Highest Loan Amounts:** The "PERSONAL" and "EDUCATION" categories have the highest loan amounts, exceeding 50 million.
- **Lowest Loan Amounts:** The "HOME IMPROVEMENT" and "TECH CONSOLIDATION" categories have the lowest loan amounts, both below 40 million.
- **Comparison:** The plot allows for easy comparison of the loan amounts across different intents.
- **Overall:** The plot provides insights into the distribution of loan amounts based on the purpose of the loan.
-

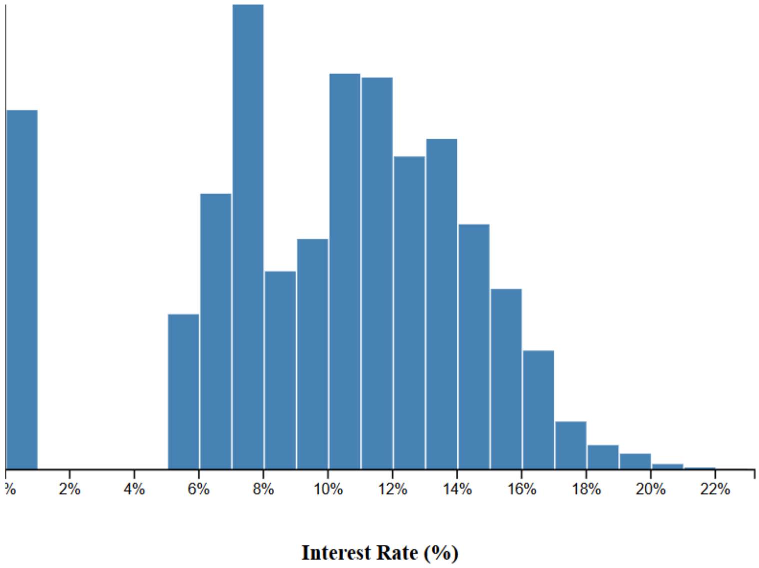

Pie Chart: Loan Status Distribution

Interpretation of the Pie Chart (Loan Status Distribution)

Key Points:

- **Loan Status:** The pie chart represents the distribution of loan statuses, categorized as "Approved" and "Rejected."
- **Dominant Category:** The "Rejected" category occupies the majority of the pie chart, indicating that a larger proportion of loan applications were rejected.
- **Comparison:** The chart visually compares the relative sizes of the two categories, providing a clear understanding of the approval and rejection rates.
- **Overall:** The pie chart offers a concise overview of the loan status distribution, highlighting the dominance of rejected applications.

Histogram: Loan Interest Rate Distribution

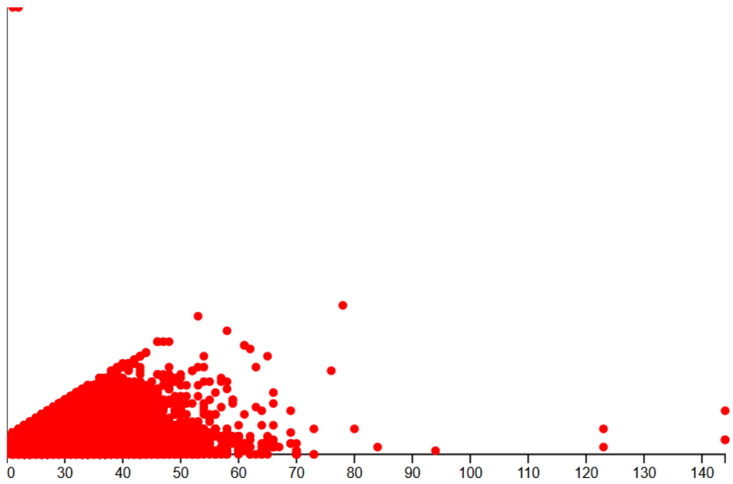


Interpretation of the Histogram (Loan Interest Rate Distribution)

Key Points:

- **Distribution:** The histogram shows the distribution of loan interest rates.
- **Skewness:** The distribution is **skewed to the right**, indicating a longer tail on the higher interest rate side. This means that there are a few loans with significantly higher interest rates compared to the majority.
- **Mode:** The **mode** (most frequent interest rate) appears to be around 10%.
- **Range:** The interest rates range from approximately 2% to 22%.
- **Shape:** The histogram has a **bell-shaped** curve, suggesting a normal distribution with some skewness.
- **Overall:** The histogram provides a visual representation of the loan interest rate distribution, highlighting the concentration of rates around 10% and the presence of higher interest rates.

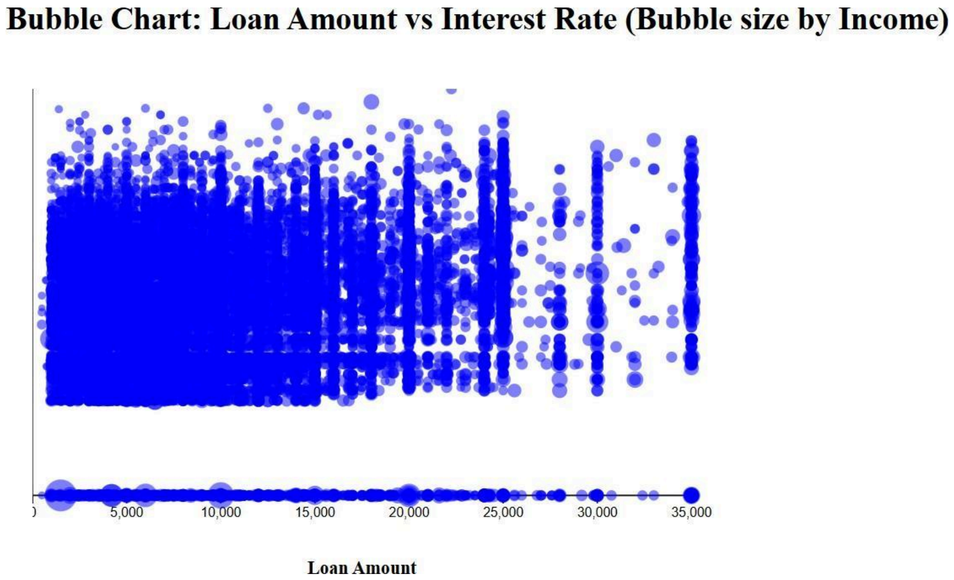
Timeline Chart: Employment Length over Age



Interpretation of the Timeline Chart (Employment Length over Age)

Key Points:

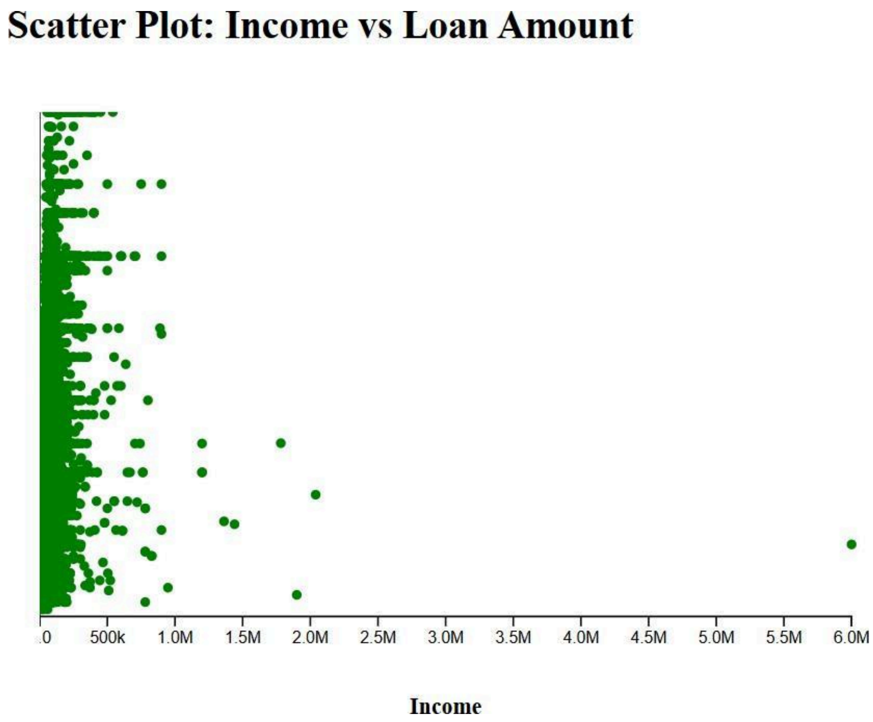
- **Relationship:** The chart illustrates the relationship between employment length and age.
- **Clustering:** The data points cluster in the lower left corner, indicating that many individuals have shorter employment lengths at younger ages.
- **Trend:** As age increases, there's a general trend towards longer employment lengths. This is evident in the concentration of data points towards the upper right corner.
- **Outliers:** A few outliers can be seen, such as individuals with very long employment lengths at relatively young ages or individuals with shorter employment lengths at older ages.
- **Overall:** The chart suggests a positive relationship between age and employment length, with most individuals having longer tenures as they get older. However, there are exceptions and individual variations.



Interpretation of the Bubble Chart (Loan Amount vs Interest Rate)

Key Points:

- **Relationships:** The chart shows how loan amount, interest rate, and income are connected.
- **Clusters:** Groups of data points suggest relationships between the variables.
- **Loan Amount and Interest Rate:** Larger loans often have higher interest rates, but there are exceptions.
- **Income and Bubble Size:** Bigger bubbles mean higher income. People with higher incomes tend to borrow more.
- **Outliers:** Some data points are unusual and don't follow the general trends.
- **Overall:** The chart shows that loan amount and interest rate are related, and income influences how much people borrow.

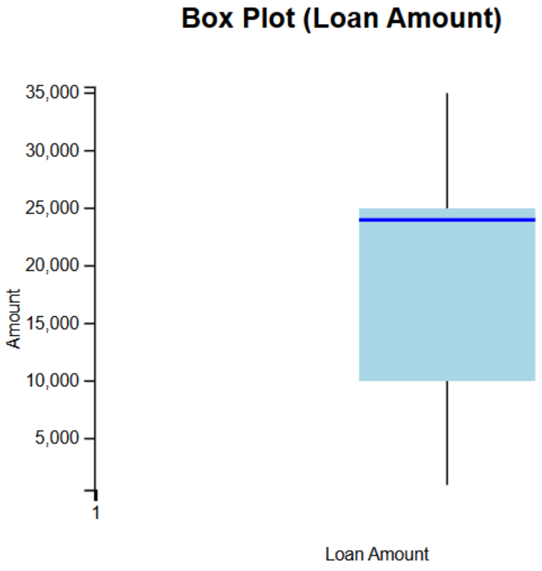


Interpretation of the Scatter Plot (Income vs Loan Amount)

Key Points:

- **Relationship:** The plot shows the relationship between income and loan amount.
- **Clustering:** The data points cluster heavily in the lower left corner, indicating that most individuals have lower incomes and borrow smaller amounts.
- **Positive Correlation:** There appears to be a general positive correlation between income and loan amount, meaning that as income increases, loan amounts tend to increase as well. However, this relationship is not perfectly linear, as there is some scattering of data points.
- **Outliers:** A few outliers can be observed, such as individuals with high incomes but relatively small loan amounts, or individuals with low incomes but very large loan amounts.
- **Overall:** The plot suggests a general positive relationship between income and loan amount, but with significant variability and the presence of outliers.

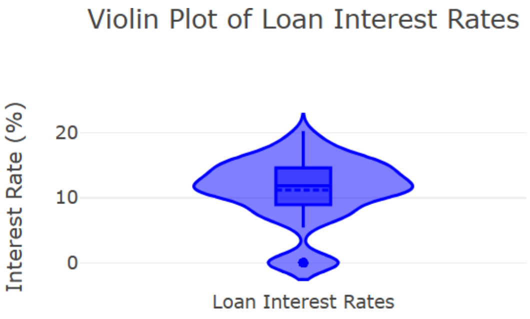
Advanced Visualizations:



Interpretation of the Box Plot (Loan Amount):

- **Median:** The middle 50% of loan amounts are centered around 25,000.
- **IQR:** The range of the middle 50% of loan amounts is approximately 5,000.
- **Outlier:** There is a significant outlier on the high end, suggesting the presence of very large loan amounts.
- **Distribution:** The overall distribution of loan amounts appears to be skewed to the right, with a longer tail on the high end.

Violin Plot:

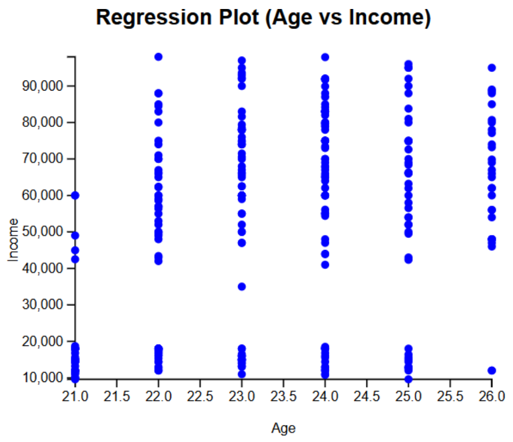


Interpretation of the Violin Plot (Loan Interest Rates)

Key Points:

- **Distribution:** The data is skewed to the right, indicating a longer tail on the higher interest rate side.
- **Median:** The median interest rate is approximately 10%.
- **IQR:** The interquartile range (IQR) is relatively small, suggesting a tight clustering of the middle 50% of rates around the median.
- **Outliers:** A potential outlier on the lower end suggests the presence of unusually low interest rates.
- **Overall:** The plot shows a spread in interest rates, with a majority around 10% but with some loans having significantly higher or lower rates

Regression Plot:

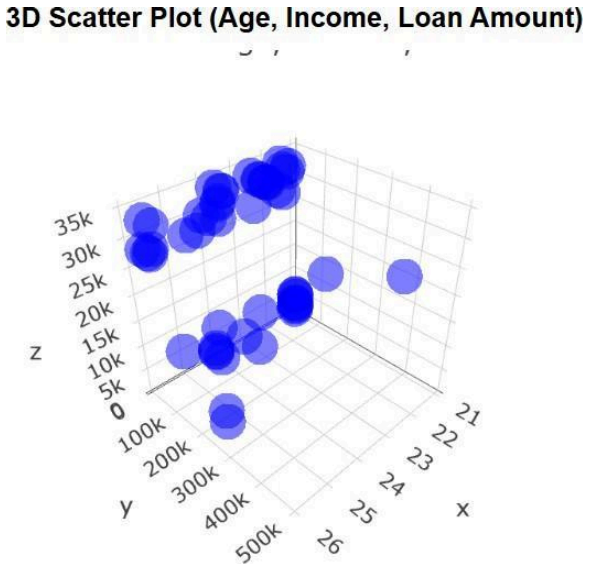


Interpretation of the Regression Plot (Age vs Income)

Key Points:

- **Positive Correlation:** There appears to be a **positive correlation** between age and income. This means that as age increases, income tends to increase as well.
- **Scatter:** The data points are somewhat **scattered** around the trend line, indicating that age is not the sole determinant of income. Other factors may also influence income.
- **Outliers:** There are a few **outliers** visible in the plot, particularly on the lower end of the age range. These are data points that deviate significantly from the overall trend.
- **Linear Relationship:** The trend line appears to be **approximately linear**, suggesting a linear relationship between age and income. However, the scatter indicates that the relationship may not be perfectly linear.

Scatter Plot 3D:



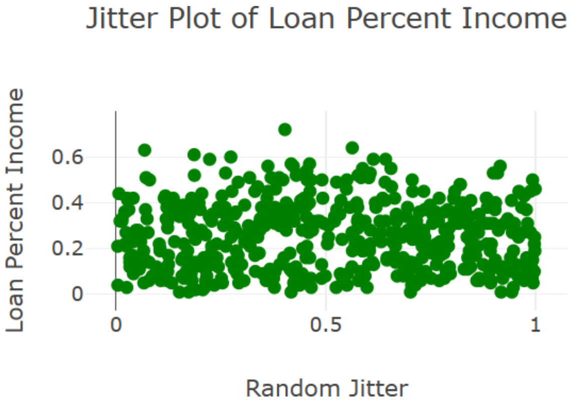
3D Scatter Plot (Age, Income, Loan Amount)

Interpretation of the 3D Scatter Plot (Age, Income, Loan Amount)

Key Points:

- **Relationships:** The plot visualizes the relationships between age, income, and loan amount.
- **Clustering:** There appears to be some clustering of data points, suggesting potential relationships or dependencies between the variables.
- **Income and Loan Amount:** There seems to be a general trend where individuals with higher incomes tend to have larger loan amounts, although there is some variability.
- **Age and Loan Amount:** The relationship between age and loan amount is less clear. While there might be a slight trend, it's not as pronounced as the relationship between income and loan amount.
- **Outliers:** A few outliers can be observed, indicating data points that deviate significantly from the general trends.

Jitter Plot



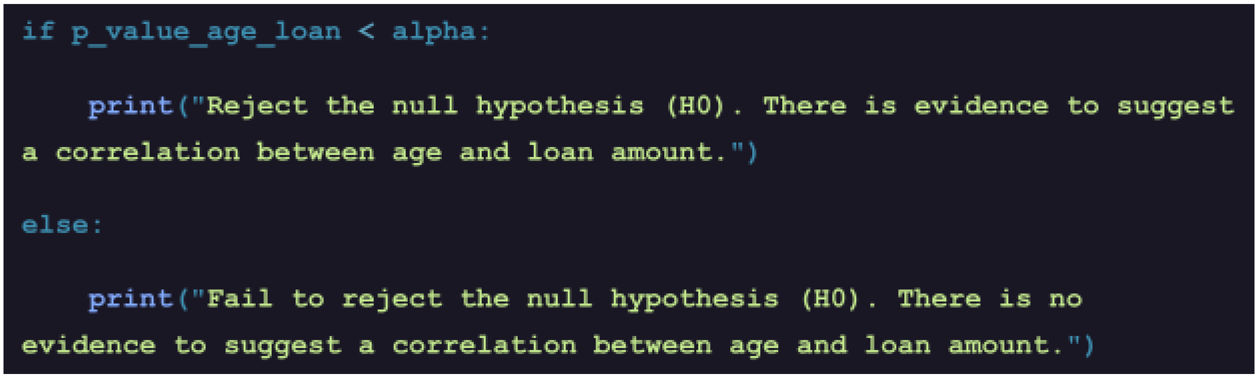
Interpretation of the Jitter Plot (Loan Percent Income)

Key Points:

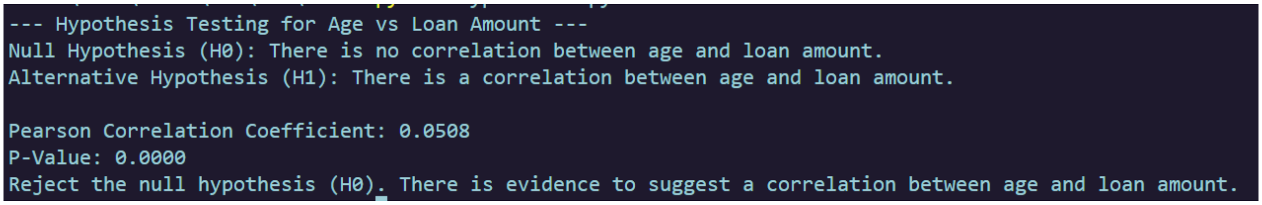
- **Distribution:** The jitter plot shows the distribution of loan percent income, which is the ratio of loan amount to income.
- **Clustering:** The data points appear to cluster around the middle of the x-axis, suggesting that a majority of loans are around 50% of income.
- **Spread:** There is a moderate spread of data points, indicating some variability in the loan percent income.
- **Outliers:** A few outliers can be observed, particularly on the higher end, indicating loans that are a much larger percentage of income.
- **Overall:** The plot suggests a general clustering of loan percent income around 50%, with some variation and the presence of outliers.

Hypothesis Testing:

```
from scipy.stats import pearsonr

import pandas as pd


# Load dataset

file_path = 'finance.csv'

data = pd.read_csv(file_path)


# Assuming the dataset has columns named 'person_age' and 'loan_amnt'

# Calculate Pearson correlation coefficient between 'person_age' and
'loan_amnt'

corr_age_loan, p_value_age_loan = pearsonr(data['person_age'],
data['loan_amnt'])


# Print the results

print(f"--- Hypothesis Testing for Age vs Loan Amount ---")

print(f"Null Hypothesis (H0): There is no correlation between age and loan
amount.")

print(f"Alternative Hypothesis (H1): There is a correlation between age
and loan amount.")


print(f"\nPearson Correlation Coefficient:

{corr_age_loan:.4f}") print(f"P-Value: {p_value_age_loan:.4f}")


# Set significance level

alpha = 0.05
```



```
if p_value_age_loan < alpha:

    print("Reject the null hypothesis (H0). There is evidence to suggest
a correlation between age and loan amount.")

else:

    print("Fail to reject the null hypothesis (H0). There is no
evidence to suggest a correlation between age and loan amount.")
```



```
--- Hypothesis Testing for Age vs Loan Amount ---
Null Hypothesis (H0): There is no correlation between age and loan amount.
Alternative Hypothesis (H1): There is a correlation between age and loan amount.

Pearson Correlation Coefficient: 0.0508
P-Value: 0.0000
Reject the null hypothesis (H0). There is evidence to suggest a correlation between age and loan amount.
```

- **Null Hypothesis (H0):** There is no correlation between age and loan amount.
- **Alternative Hypothesis (H1):** There is a correlation between age and loan amount.
- **Pearson Correlation Coefficient:** 0.0508, indicating a weak positive correlation.
- **P-Value:** 0.0000, which is significantly smaller than the typical alpha level of 0.05.
- **Conclusion:** Since the p-value is less than the alpha level, we **reject the null hypothesis**. This means there is sufficient evidence to suggest a **correlation between age and loan amount**.

In summary, the statistical analysis provides evidence that age and loan amount are correlated, although the correlation is weak.

CONCLUSION:

From the visualizations and analysis of the loan dataset, we learnt about D3.js, which allows creating interactive, dynamic visualizations like regression plots and custom box plots, enabling in-depth exploration of data patterns