BHARATIYA VIDYA BHAVANS
SARDAR PATEL INSTITUTE OF TECHNOLOGY
(Empowered Autonomous Institute Affiliated to University of Mumbai)
[Knowledge is Nectar]

Advance Data Visualization

UID	2021700030
Name	Khushi Jain
Batch	Batch L
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: Loan Dataset

Link: https://www.kaggle.com/datasets/mirzahasnine/loan-data-set

About Dataset:

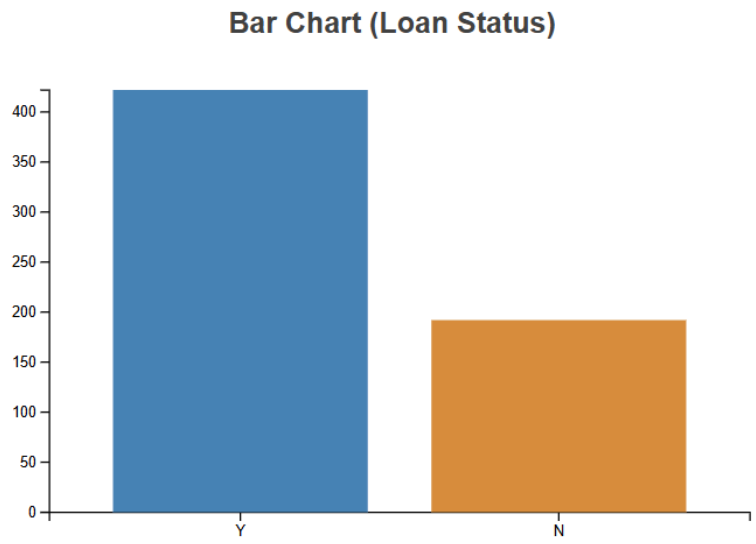
This dataset provides a mix of categorical (e.g., Gender, Married, Education) and numerical variables (e.g., ApplicantIncome, LoanAmount) about applicants seeking loans. It includes information about applicants' demographics, financial status, and loan details, along with whether or not

their loan was approved. This data structure is typically used for loan approval analysis, credit risk assessment, or customer segmentation studies.

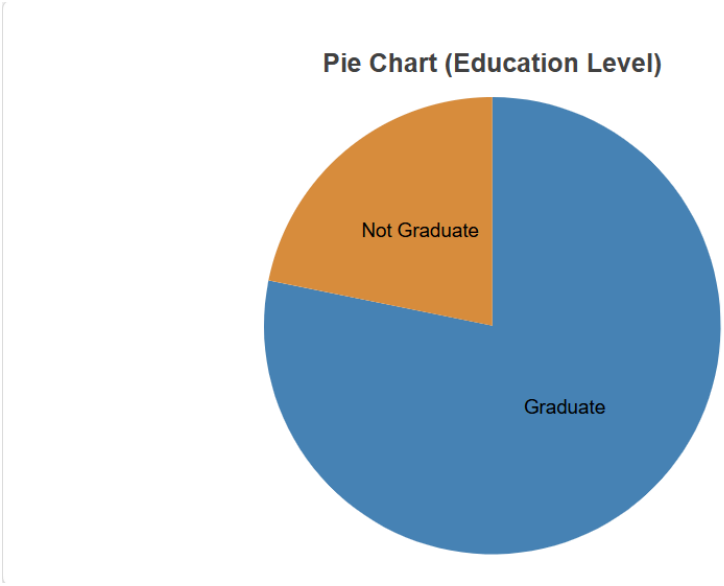
Column Attributes:

- **Loan_ID**: A unique identifier for each loan applicant.
- **Gender**: The gender of the applicant (e.g., Male, Female).
- **Married**: Marital status of the applicant (e.g., Yes, No).
- **Dependents**: Number of dependents the applicant has (e.g., 0, 1, 2, 3+).
- **Education**: Educational background of the applicant (e.g., Graduate, Not Graduate).
- **Self_Employed**: Whether the applicant is self-employed (e.g., Yes, No).
- **ApplicantIncome**: Income of the applicant.
- **CoapplicantIncome**: Income of the co-applicant, if any.
- **LoanAmount**: The loan amount requested by the applicant.
- **Loan_Amount_Term**: The term (duration) of the loan in months.
- **Credit_History**: Indicator of whether the applicant has a credit history (1 for "Yes," 0 for "No").
- **Property_Area**: The type of area where the property is located (e.g., Urban, Rural, Semiurban).
- **Loan_Status**: Whether the loan was approved or not (e.g., Y for "Yes," N for "No").

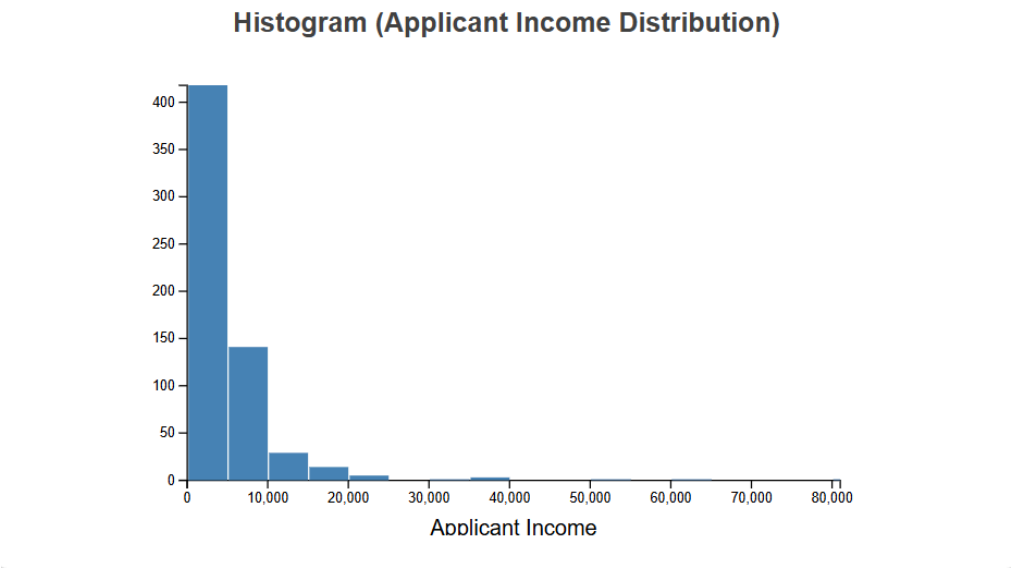
Basic Visualizations:



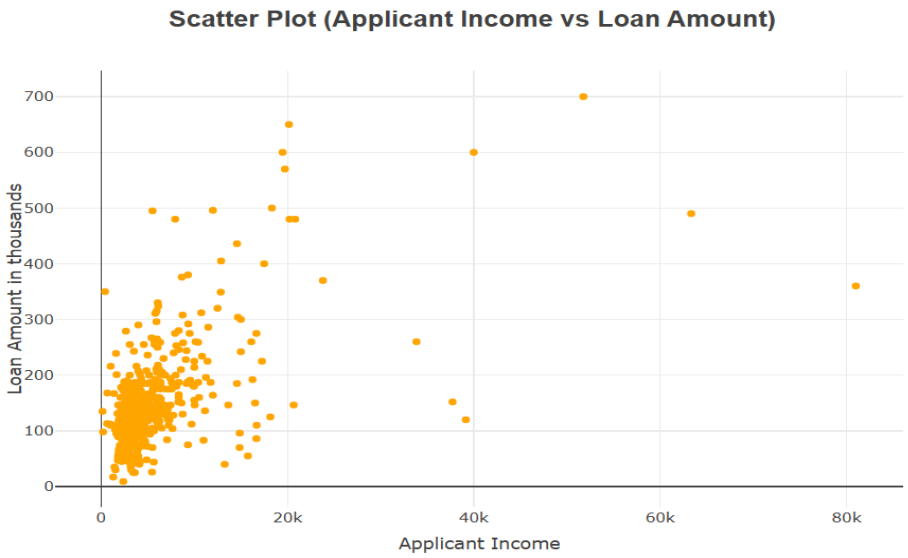
Positive Loan Status Dominates: The "yes" category has a significantly higher count compared to the "no" category, indicating that there are many more positive loan status.



The visual disparity between the "graduate" and "not graduate" segments highlights the significant difference in education level within the loan applicants.



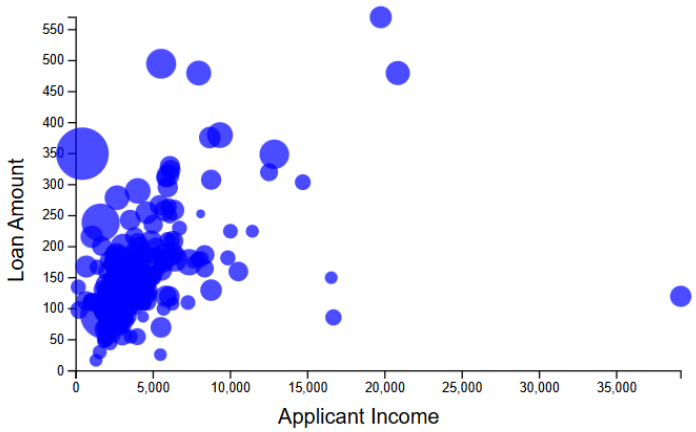
Most applicants have an income of less than 10,000, with a few applicants having much higher incomes, creating a right-skewed distribution.



Most data points are clustered at lower income and loan amounts, indicating that applicants with lower incomes tend to apply for smaller loan amounts.

There is a positive correlation between Applicant Income and Loan Amount, meaning that as Applicant Income increases, the Loan Amount also tend to increase.

Bubble Plot (Applicant Income, Loan Amount and Coapplicant Income)



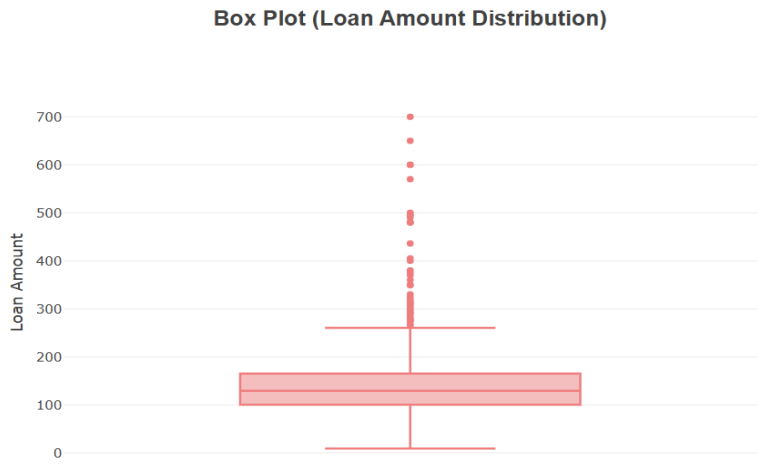
This plot helps identify the relationship between applicant income, loan amount, and coapplicant income. For instance, applicants with higher incomes seem to request larger loans, though not all of them have high coapplicant incomes.

The size variation suggests that **coapplicant income is not strongly correlated** with either applicant income or loan amount, as large and small bubbles appear across the spectrum.

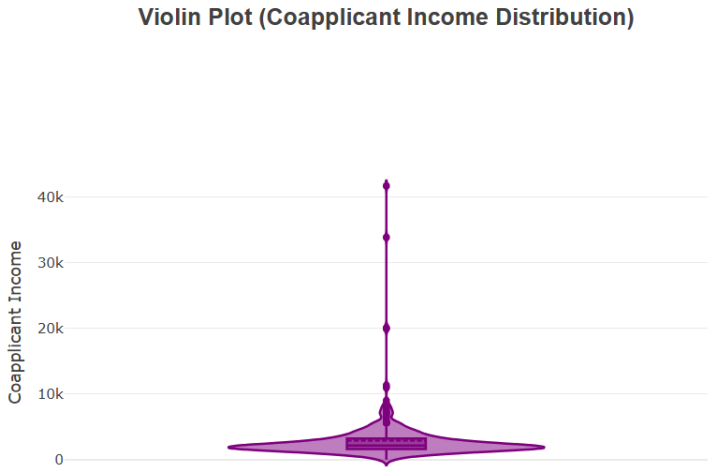
Advanced Visualizations:

Word Cloud (Property Area)

Word Cloud shows that most applicants live in Semi urban regions and least in Rural areas.

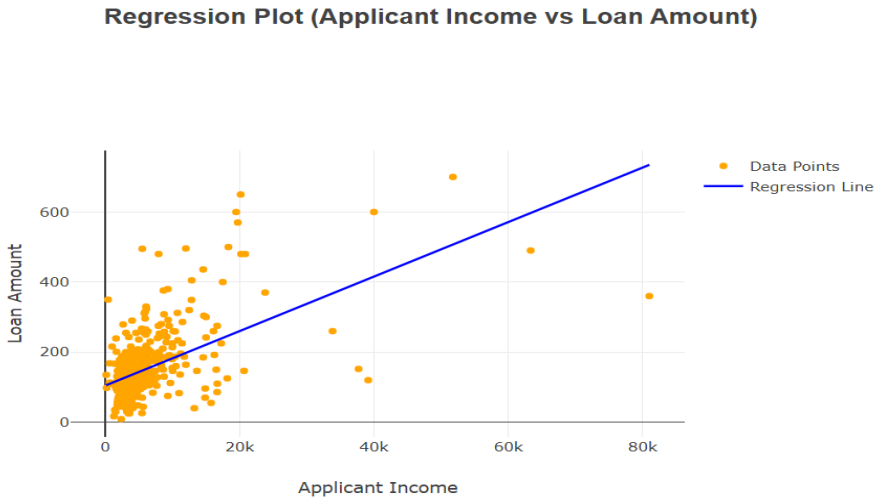


The plot shows that the majority of loan amounts are clustered around the lower values, with most between approximately 100 and 200.There is a significant number of high-value outliers, which could indicate special cases where larger loans were requested, perhaps by applicants with higher incomes or other qualifying criteria.

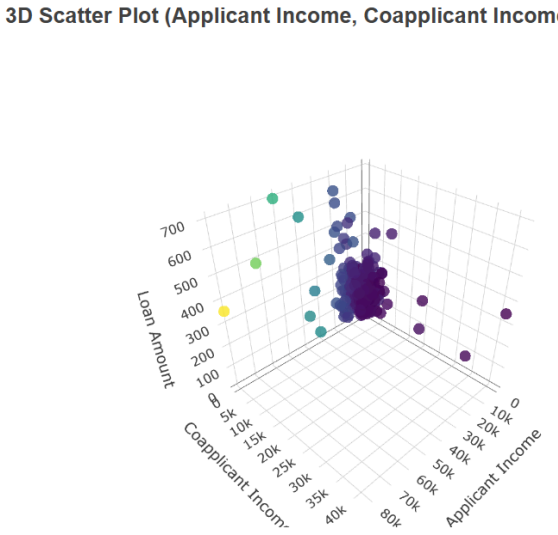


The plot is denser around the lower income range, suggesting that most coapplicants have lower incomes. This concentration indicates that a significant portion of the loan applications are from individuals with low to moderate coapplicant incomes. The plot shows a long, thin tail extending upward, representing high-income outliers. This suggests that while most

coapplicant incomes are relatively low, there are a few cases with much higher incomes.

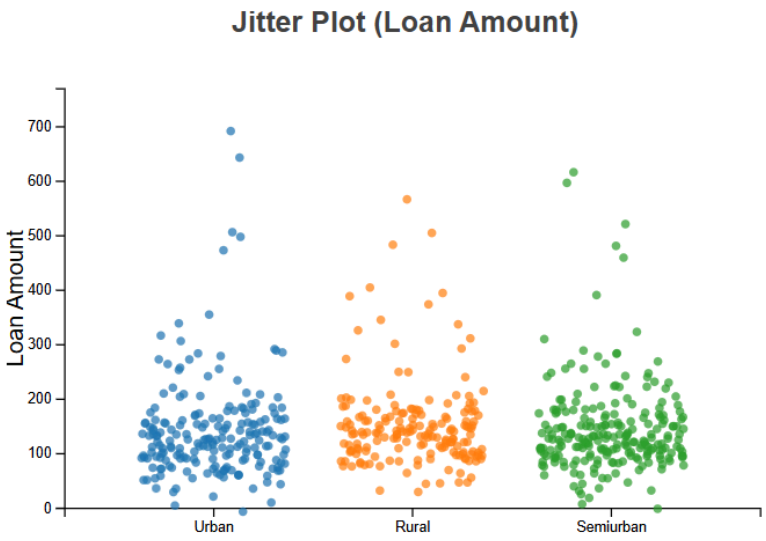


There is a positive correlation (as seen by the regression line) between Applicant Income and Loan Amount, meaning that as Applicant Income increases, the Loan Amount also tend to increase.



Most of the data points cluster in the lower ranges of all three axes, indicating that many applicants have low to moderate incomes and are

requesting smaller loan amounts. The scatter plot suggests positive correlations, as higher applicant and coapplicant incomes are generally associated with higher loan amounts.



The distribution patterns appear relatively similar across the three property areas, suggesting no major differences in loan amount requests based on property area.

A significant concentration of loan amounts is in the range of 100–200 across all categories, indicating a common loan amount preference regardless of location.

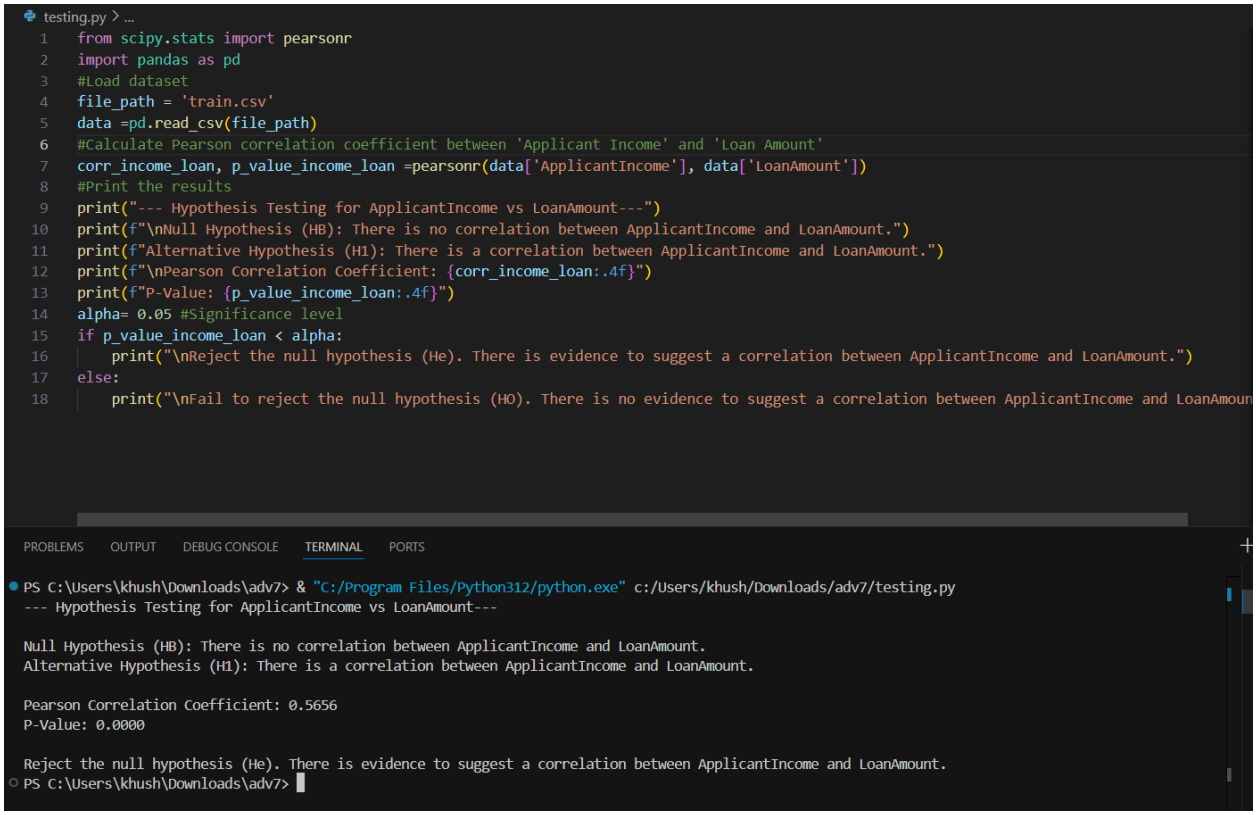
Hypothesis Testing

Perform hypothesis testing to evaluate the correlation between Applicant Income and Loan Amount in the dataset.

Hypothesis

1. Null Hypothesis (H0): There is no correlation between Applicant Income and Loan Amount.
2. Alternative Hypothesis (H1): There is a correlation between Applicant Income and Loan Amount.

CODE AND OUTPUT:



```
testing.py > ...
1  from scipy.stats import pearsonr
2  import pandas as pd
3  #Load dataset
4  file_path = 'train.csv'
5  data =pd.read_csv(file_path)
6  #Calculate Pearson correlation coefficient between 'Applicant Income' and 'Loan Amount'
7  corr_income_loan, p_value_income_loan =pearsonr(data['ApplicantIncome'], data['LoanAmount'])
8  #Print the results
9  print("--- Hypothesis Testing for ApplicantIncome vs LoanAmount---")
10 print(f"\nNull Hypothesis (HB): There is no correlation between ApplicantIncome and LoanAmount.")
11 print(f"Alternative Hypothesis (H1): There is a correlation between ApplicantIncome and LoanAmount.")
12 print(f"\nPearson Correlation Coefficient: {corr_income_loan:.4f}")
13 print(f"P-Value: {p_value_income_loan:.4f}")
14 alpha= 0.05 #Significance level
15 if p_value_income_loan < alpha:
16     print("\nReject the null hypothesis (He). There is evidence to suggest a correlation between ApplicantIncome and LoanAmount.")
17 else:
18     print("\nFail to reject the null hypothesis (HO). There is no evidence to suggest a correlation between ApplicantIncome and LoanAmoun

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS C:\Users\khush\Downloads\adv7> & "C:/Program Files/Python312/python.exe" c:/Users/khush/Downloads/adv7/testing.py
--- Hypothesis Testing for ApplicantIncome vs LoanAmount---

Null Hypothesis (HB): There is no correlation between ApplicantIncome and LoanAmount.
Alternative Hypothesis (H1): There is a correlation between ApplicantIncome and LoanAmount.

Pearson Correlation Coefficient: 0.5656
P-Value: 0.0000

Reject the null hypothesis (He). There is evidence to suggest a correlation between ApplicantIncome and LoanAmount.
PS C:\Users\khush\Downloads\adv7>
```

Conclusion: Using the loan dataset, visualizations created with D3.js revealed significant insights. These interactive and dynamic visualizations, such as regression plots and box plots, facilitated deeper exploration of data patterns and trends. These tools are invaluable for understanding complex relationships within the dataset and enhancing data analysis capabilities.