Name: Rashmi Rajeshirke
UID: 2021700051
Class: CSE-DS BE
Batch: A

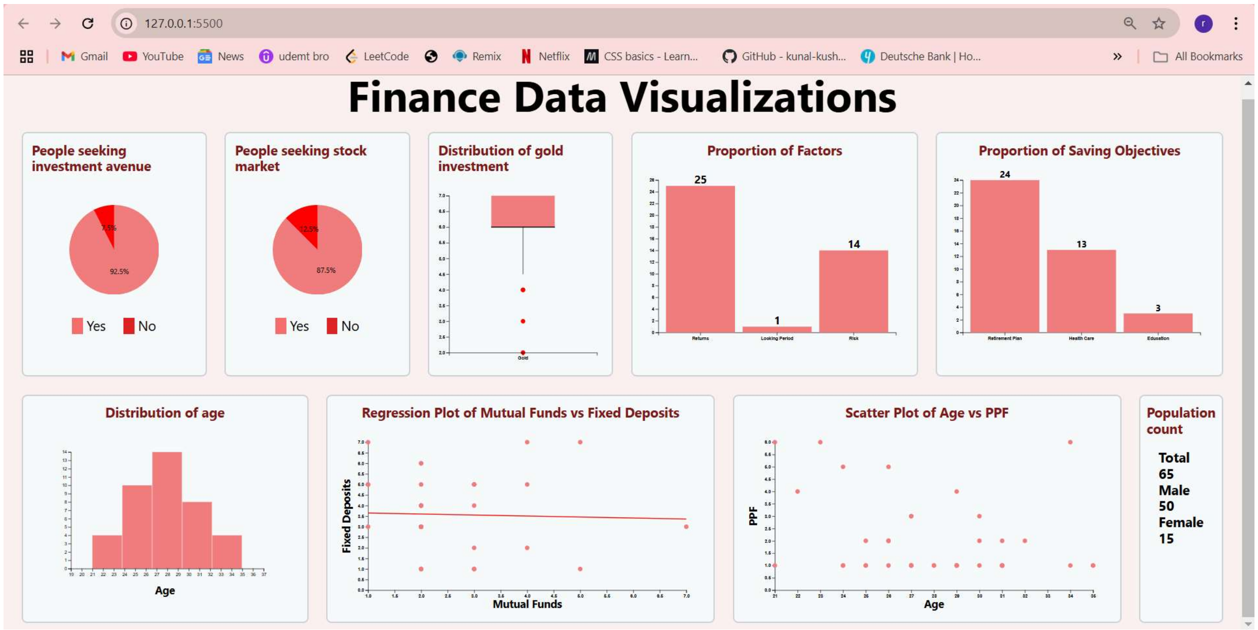
Lab7

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.

Dashboard:



Total
65
Male
50
Female
15

Observations:

1. People Seeking Investment Avenues: This pie chart shows that 92.5% of people are seeking various investment avenues, while 7.5% are not. This suggests that the majority of the population is actively looking for ways to grow their wealth, which could reflect an increased awareness of financial planning or a desire to achieve long-term financial

goals, such as retirement or wealth accumulation.

2. People Seeking Stock Market Investments: The pie chart reveals that 87.5% of respondents are interested in investing in the stock market, while 12.5% are not. This indicates a strong inclination toward stock market investments among the majority, which could be driven by the potential for higher returns, especially as more people become aware of equity investments as a viable option for wealth growth.
3. Distribution of Gold Investment: The box plot visualizes the distribution of gold investments. The median investment is relatively high, with some lower outliers indicating that a few individuals invest significantly less. This suggests that gold is a popular investment option for many due to its perceived safety and stability, especially in uncertain economic conditions. The spread in the box plot indicates a variation in how much individuals are willing to allocate to gold.
4. Proportion of Factors Influencing Investment: This bar chart reveals that returns are the most significant factor driving investment decisions, with 25 respondents prioritizing it. The next important factor is risk, mentioned by 14 respondents, while the locking period is only a concern for 1 individual. This shows that while people are primarily driven by the potential for high returns, they also take risks into consideration, and the locking period (the time money is tied up in an investment) is not a major concern for most.
5. Proportion of Saving Objectives: This bar chart shows the different saving objectives among respondents. Retirement planning is the leading objective, cited by 24 people, followed by health care (13), and education (3). This indicates that most individuals are focused on long-term financial security, particularly for retirement, while others are also considering healthcare needs. Education savings appear to be a lower priority in comparison.
6. Distribution of Age: The histogram shows the distribution of participants' ages, with most individuals being in their late 20s to early 30s. The peak occurs around the age of 27-29. This age group appears to be the most actively engaged in investment and savings decisions, likely due to being in the early or middle stages of their careers, when financial planning becomes crucial.
7. Regression Plot of Mutual Funds vs Fixed Deposits: The scatter plot with a regression line shows the relationship between mutual funds and fixed deposit investments. The flat regression line suggests that there is no significant correlation between the amounts people invest in mutual funds and fixed deposits. This could indicate that individuals diversify their portfolios and may invest in these two options independently based on their risk tolerance, rather than one influencing the other.
8. Scatter Plot of Age vs PPF (Public Provident Fund): This scatter plot explores the relationship between age and investments in PPF. The data points are scattered, indicating no clear correlation between age and the amount invested in PPF. This suggests that people across different age groups invest in PPF, likely due to its tax-saving benefits and guaranteed returns, making it appealing to individuals regardless of their age.
9. Population Count: The small box on the right displays the total population count of 65 respondents, with 50 males and 15 females. This gender distribution suggests that the dataset is male-dominated, which could be reflective of the sample population’s financial engagement or participation in investment activities.

Hypothesis testing:

Step 1: Formulate Hypotheses

- Null Hypothesis (H_0): There is no correlation between age and the type of savings objectives pursued by respondents. This implies that changes in age do not influence the savings objectives individuals select.
- Alternative Hypothesis (H_1): There is a correlation between age and the type of savings objectives pursued by respondents. This suggests that age influences the savings objectives that individuals consider.

Step 2: Choose a Significance Level (α)

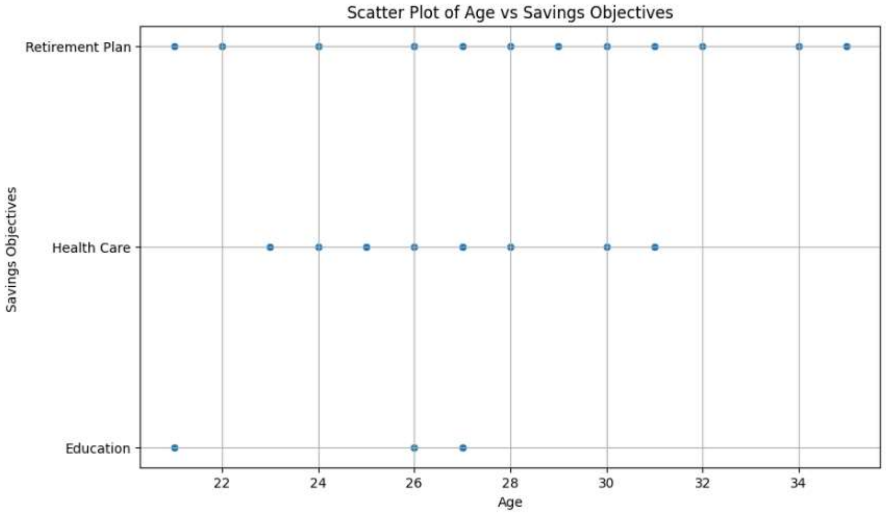
- The significance level (α) is set at 0.05. This threshold indicates that we are willing to accept a 5% chance of incorrectly rejecting the null hypothesis when it is true, which is a commonly accepted level in hypothesis testing.

Step 3: Select the Appropriate Statistical Test

- The appropriate statistical test for this analysis is the Pearson correlation test. This test is specifically designed to assess the strength and direction of the linear relationship between two continuous variables—in this case, age and the numeric representation of savings objectives.

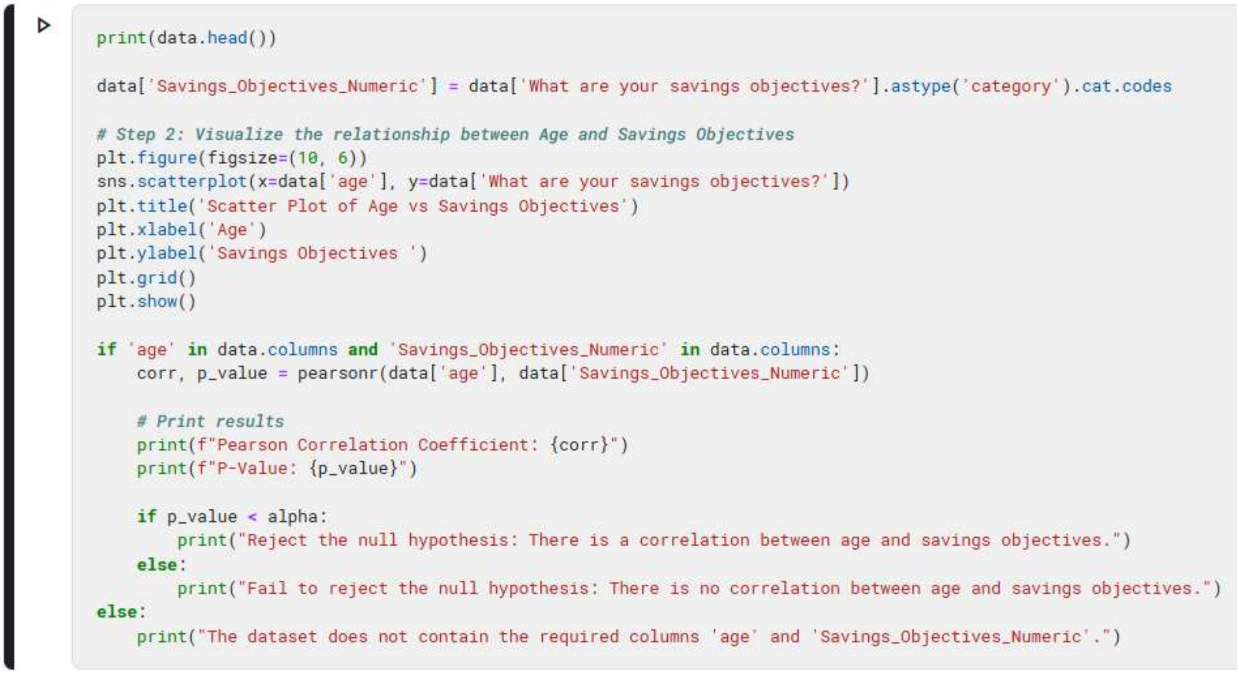
Step 4: Collect and Visualize Data

- Before performing the statistical test, the data was visualized to observe trends, relationships, and any potential outliers. A scatter plot was used to illustrate the relationship between age and the numeric representation of savings objectives. This visual representation helps to assess whether a linear relationship may exist between the two variables.

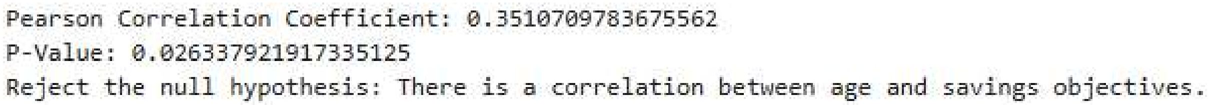


Step 5: Perform the Hypothesis Test

1. Compute the Test Statistic: The Pearson correlation coefficient was calculated to quantify the degree of correlation between age and savings objectives.
2. Calculate the P-Value: The p-value obtained from the Pearson correlation test indicates the probability of observing the data (the calculated correlation coefficient) under the assumption that the null hypothesis is true.
3. Compare the P-Value to α:
   - The computed Pearson Correlation Coefficient is approximately 0.351.
   - The P-Value is approximately 0.026.
   - Since the p-value (0.026) is less than the significance level (0.05), we reject the null hypothesis (H_0).



```
print(data.head())

data['Savings_Objectives_Numeric'] = data['What are your savings objectives?'].astype('category').cat.codes

# Step 2: Visualize the relationship between Age and Savings Objectives
plt.figure(figsize=(10, 6))
sns.scatterplot(x=data['age'], y=data['What are your savings objectives?'])
plt.title('Scatter Plot of Age vs Savings Objectives')
plt.xlabel('Age')
plt.ylabel('Savings Objectives ')
plt.grid()
plt.show()

if 'age' in data.columns and 'Savings_Objectives_Numeric' in data.columns:
    corr, p_value = pearsonr(data['age'], data['Savings_Objectives_Numeric'])

    # Print results
    print(f"Pearson Correlation Coefficient: {corr}")
    print(f"P-Value: {p_value}")

    if p_value < alpha:
        print("Reject the null hypothesis: There is a correlation between age and savings objectives.")
    else:
        print("Fail to reject the null hypothesis: There is no correlation between age and savings objectives.")
else:
    print("The dataset does not contain the required columns 'age' and 'Savings_Objectives_Numeric'.")
```



```
Pearson Correlation Coefficient: 0.3510709783675562
P-Value: 0.026337921917335125
Reject the null hypothesis: There is a correlation between age and savings objectives.
```

Step 6: Interpret Results

- Reject H_0: The p-value is less than the significance level, indicating that there is sufficient evidence to support the alternative hypothesis (H_1). This suggests that there is a statistically significant correlation between age and the type of savings objectives pursued by respondents.
- Conclusion: The analysis indicates that as individuals age, their savings objectives are likely to change, reflecting different financial priorities or strategies. This finding may provide valuable insights into how age-related factors influence financial planning and decision-making.