Cognifyz INTERNSHIP PROGRAM BUSINESS ANALYTICS

Task 1: Data Overview:

Objective

Understand the structure and content of the provided dataset before starting data analysis.

Steps Taken:

- Loaded the dataset file "Data_set-2-Copy.csv" into Python using the pandas library.
- Examined the dataset structure by viewing the column names and data types to understand the kind of data available.
- Used the head() method to display the first 5 rows of data and get a sample of the dataset content.
- Checked for missing or null values in all columns using isnull().sum() to assess data quality.
- Verified that the dataset contains both numeric (e.g., age) and categorical (e.g., gender, investment avenues) columns ready for further analysis.
- Documented the findings, including the list of columns, data types, sample data, and missing value status, as the foundation for the full project.

Python Code Used:

Refer the pdf attached below

Summary

All required columns are present and well-labeled. Data includes both numeric (age) and categorical/text fields (gender, investment details, etc.). No critical missing data observed. Dataset structure is clear and ready for further analysis.

Task Output:

- The dataset contains 40 records, no missing values, and 24 columns covering demographics, investment avenues, reasons, durations, expectations, and sources.

Task 2: Gender Distribution

Objective

Analyze and visualize the distribution of genders in the dataset to understand participant composition.

Steps Taken

- Loaded the dataset “Data_set-2-Copy.csv” into Python using the pandas library.
- Extracted the 'gender' column from the dataset to focus analysis on gender distribution.
- Used the value_counts() function in pandas to count the number of male and female participants.
- Printed the count results to understand the composition of participants by gender.
- Created a bar chart visualization using matplotlib to visually illustrate the gender distribution.
- Analyzed and interpreted the results to determine if the dataset is balanced or skewed towards a particular gender group.

Python Code Used:

Refer the attached PDF below

Summary

The dataset contains more female than male participants, as shown in the gender count and bar chart. Understanding the gender distribution sets the foundation for analyzing investment preferences or patterns by gender in later tasks.

Task Output:

- Males (25) outnumber females (15) among participants.

Task 3: Descriptive Statistics

Objective

Present basic statistics for the numerical columns in the dataset—such as age (and income, if available)—to summarize and understand the data's distribution.

Steps Taken

- Loaded the dataset "Data_set-2-Copy.csv" using pandas library.
- Identified numerical columns in the dataset, primarily the "age" column.
- Used pandas describe() method to calculate descriptive statistics:
    - Count (number of valid entries)
    - Mean (average)
    - Standard deviation (spread of data)
    - Minimum and maximum values
    - Quartiles (25%, 50%, 75%)
- Calculated median separately using median() method for central tendency.
- Interpreted the results to understand the distribution of numerical data (e.g., average age, variability).
- (Optional) Created visualizations such as histograms or box plots to better represent the distribution and detect outliers.
- Documented all outputs and explained key takeaways for your report.

Python Code Used:

Refer the attached PDF below

Summary

The basic descriptive statistics (mean, median, and standard deviation) show the average and spread of participant ages. This helps provide a statistical foundation for further analysis in following tasks.

Task Output:

- Participants' age spans 21 to 35 years, averaging 27. Median age is 27.
- Investment avenues like Mutual Funds and Equity Market vary in investment scale.

Task 4: Most Preferred Investment Avenue

Objective

Identify which investment avenue (such as equity, mutual funds, etc.) is the most popular among participants.

Steps to Complete

- **Find the relevant column:**
  Locate the investment avenue column(s) in your dataset (this could be "InvestmentAvenues", or specific fields like "EquityMarket", "MutualFunds", etc.).
- **Count preferences:**
  - ✓ If you have a single column where users select their preferred type, use value_counts() to count how often each type appears.
  - ✓ If your dataset has multiple columns (for each avenue type as Yes/No), count how many times each avenue is marked "Yes".
- **Determine the most preferred:**
  See which investment avenue has the highest count.
- **Visualize (optional but strong):**
  Create a bar chart to show how each avenue compares in popularity.

Python Example (if you have multiple Yes/No columns):

Refer the attached PDF below

Summary

From the count and bar chart, Mutual Funds is the most preferred investment avenue among participants. This kind of insight is crucial for understanding participant preferences and guiding business analytics decisions.

Task Output:

- Equity Market, Mutual Funds, and Fixed Deposits are most popular among participants based on "yes" responses.

Task 5: Reasons for Investment

Objective

To analyze and summarize the key reasons participants provided for their investment choices.

Steps Taken

- Loaded the dataset “Data_set-2-Copy.csv” using pandas.
- Identified columns capturing reasons for investment (Reason_Equity, Reason_Mutual, Reason_Bonds, Reason_FD).
- Extracted and consolidated investment reasons from these columns into a single list.
- Cleaned the data by removing empty or null entries.
- Counted and ranked the frequency of each unique reason.
- Created a summary table showing the most common reasons.
- (Optional) Generated a bar chart visualizing the top reasons.
- Interpreted the results, highlighting major motivations like “Capital Appreciation,” “Better Returns,” etc.

Python Code Used:

Refer the attached PDF below

Summary

This analysis revealed the most common motivations for investment among participants, providing insights into their decision-making priorities.

Task Output:

- Capital Appreciation (30), Assured Returns (26), and Better Returns (24) are top investment reasons.

Task 6: Savings Objectives

Objective

Identify and present the main savings objectives of the participants.

Steps Taken:

- Loaded the dataset “Data_set-2-Copy.csv” using pandas.
- Investigated the column related to participants’ savings objectives.
- Extracted all distinct savings objectives mentioned by participants.
- Counted the frequency of each unique savings objective.
- Created a summary list describing the most common savings objectives participants had.
- (Optional) Visualized the distribution of savings objectives using a bar chart for clarity.

Python Code Used:

Refer the attached PDF below

Summary

The analysis reveals the primary savings goals among participants, which helps understand their financial priorities and plan investment strategies accordingly.

Task Output:

- Retirement Plan (24) and Health Care (13) are the leading savings objectives.

Task 7: Common Information Sources

Objective

Analyze and summarize the most common sources participants rely on for investment information.

Steps Taken:

- Loaded the dataset “Data_set-2-Copy.csv” using pandas.
- Identified the column where participants indicated their sources of investment information (e.g., ‘Source’).
- Counted the frequency of each unique information source mentioned.
- Created a summary table listing the most common sources.
- (Optional) Visualized the counts using a bar chart to better illustrate the distribution of sources participants rely on.

Python Code Used:

Refer the attached PDF below

Summary

The investigation highlights which sources (such as financial advisors, internet, newspapers, peers) are most trusted or frequented by participants for investment-related information, giving valuable insight into information channels influencing investment decisions.

Task Output:

- Financial Consultants (16) and Newspapers & Magazines (14) are the main information sources

Task 8: Investment Duration Analysis

Objective

Calculate the average investment duration among participants to understand their investment time horizons.

Steps Taken:

- Loaded the dataset “Data_set-2-Copy.csv” into Python using pandas.
- Identified the column representing investment duration (e.g., ‘Investment_Duration’).
- Converted the values in this column to numeric format to enable calculations.
- Calculated the average (mean) investment duration among participants to understand typical investment horizons.
- Examined the distribution of investment durations by creating a histogram.
- Interpreted the average duration and distribution to assess how long participants tend to plan their investments.

Python Code Used:

Refer the attached PDF below

Summary

The average investment duration provides insights into how long participants plan to invest their funds. The distribution plot highlights the spread and concentration of durations, helping understand overall participant preferences.

Task Output:

- Average investment duration is 2.98 years.
- Majority invest between 2 to 4 years, with very few below 1 year or more than 5 years.

Task 9: Expectations from Investments

Objective

Summarize and analyze common expectations participants have from their investments.

Steps Taken:

- Loaded the dataset “Data_set-2-Copy.csv” using pandas.
- Reviewed the column where participants provided their expectations from investments (e.g., 'InvestmentExpectations').
- Extracted all entries from this column and combined them into a single series.
- Cleaned the data by handling missing values and standardizing text if needed.
- Counted the frequency of unique expectations to identify common themes.
- Created a summary table listing the most frequent expectations.
- (Optional) Visualized the expectations using a bar chart for better clarity.

Python Code:

Refer the attached PDF below

Summary

Analyzed common investment expectations from participant data, identifying top themes like capital gains, steady income, safety, and liquidity. This highlights key investor priorities and aids understanding of behavior patterns.

Task Output:

- Most expect 20%-30% returns (32 participants), followed by 30%-40% (5) and 10%-20% (3).

Task 10: Correlation and Risk Analysis

Objective

Explore correlations between key factors like age, investment duration, and expected returns, and assess investment risk profile.

Steps Taken:

- Loaded the dataset using pandas.
- Selected relevant columns for analysis: age, investment duration, expected returns.
- Calculated correlation coefficients to identify relationships.
- Interpreted correlation strength and direction to understand dependencies.
- Reviewed risk metrics such as volatility or beta if available to assess risk levels.
- Visualized correlations using scatter plots or heatmaps.

Python Code:

Refer the attached PDF below

Summary

Analyzed relationships between key variables like age, investment duration, and expected returns using correlation coefficients. Found how variables move together, indicating potential dependencies. This helps in understanding risk profiles and aids in portfolio diversification. Visual tools like scatter plots or heatmaps were used to illustrate correlations.

Task Output:

- Minimal correlation (0.05) found between age and duration of investment indicating these factors do not move together significantly.
- The correlation between age and investment duration is very weak, indicating these factors operate independently. This suggests diversification across different factors can effectively reduce overall investment risk.