2. Software Stack Used

The following are the tech stacks which I used to complete the project during the internship

2.1 Python

Fig : Python

Business analysis using Python makes task easier since Python Programming language has many advantages over any other programming language. It has prominent features like being a high-level programming language (the codes are in human readable form) it is easy to understand and use by any programmer or user. Many libraries and functions for statistical, numerical analysis are available in Python. Moreover, the source code is freely available to anyone (free and open source).

Libraries of Python:

1. Numpy:

o Definition: NumPy is a Python library that provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

o Key Features:

- Efficient numerical computations.
- Core data structure: N-dimensional array (ndarray).
- Broadcasting for element-wise operations.
- Linear algebra, Fourier transforms, and more.

o Use Case: NumPy is the backbone for scientific computing and data manipulation.

2. Pandas:

- Definition: Pandas is a powerful data manipulation library for Python. It provides data structures (such as DataFrames and Series) and tools for cleaning, transforming, and analyzing data.
- Key Features:
  - Tabular data representation (like spreadsheets).
  - Data cleaning, filtering, and aggregation.
  - Integration with databases and CSV files.
  - Time series analysis.
- Use Case: Pandas simplifies data wrangling and exploration.

3. Matplotlib:

- Definition: Matplotlib is a versatile data visualization library built on NumPy arrays. It allows you to create various types of plots, including line charts, scatter plots, histograms, and more.
- Key Features:
  - Customizable plots and charts.
  - Supports 2D and 3D visualizations.
  - Publication-quality graphics.
- Use Case: Matplotlib helps you convey insights through visual representations.

4. Seaborn:

o Definition: Seaborn is a statistical data visualization library based on Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics.

o Key Features:

- Simplified syntax for complex plots.
- Built-in color palettes.
- Specialized plots (e.g., violin plots, pair plots).

o Use Case: Seaborn enhances Matplotlib by adding statistical context to your visualizations.

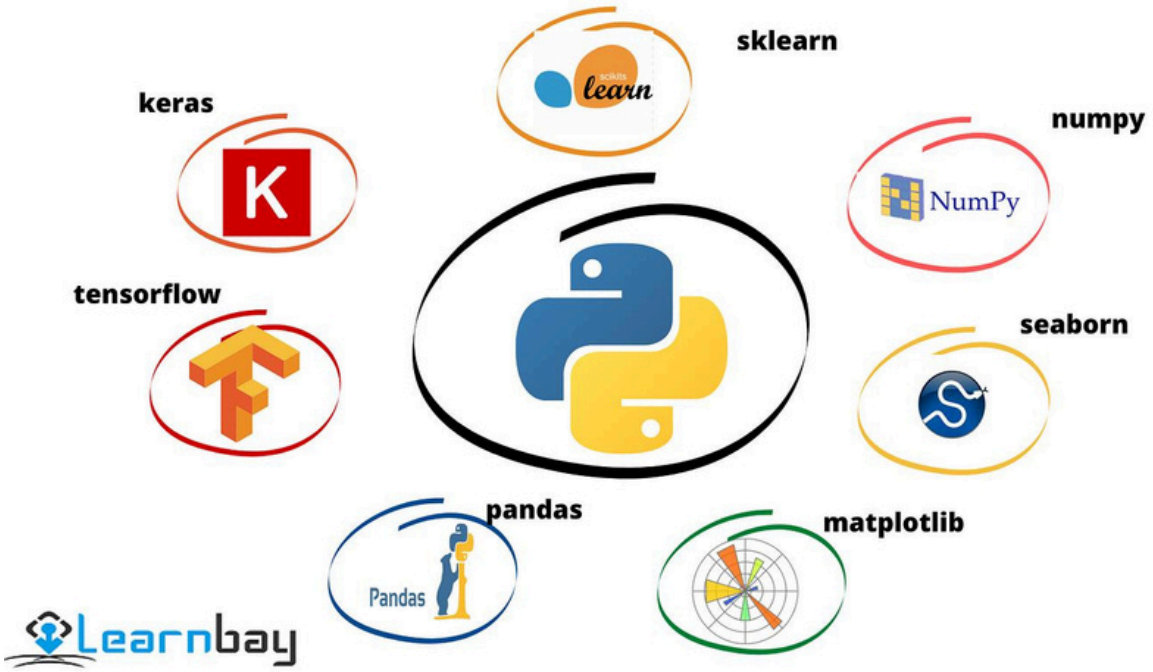


Fig : Python Libraries

Resources

Jupyter Notebook:(Interactive computing environment)

Dataset:(The data used for analysis)
Anaconda:(Managing Python environments and packages)

Tools Used

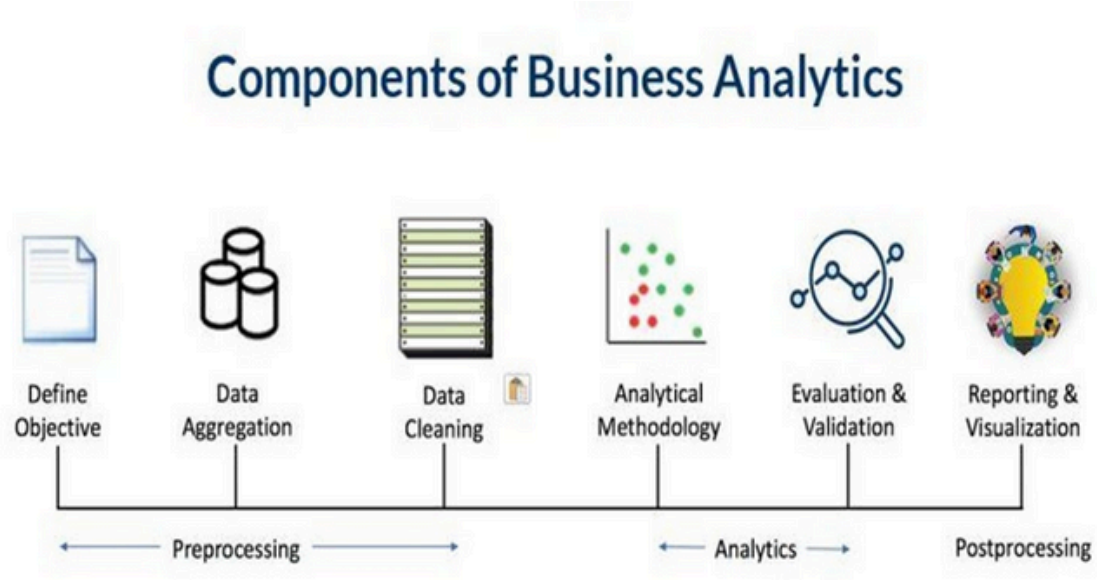
POWER BI

Power BI is a comprehensive suite of software services, apps, and connectors developed by Microsoft. It seamlessly transforms disparate data sources into coherent, visually engaging, and interactive insights. Whether your data originates from an Excel spreadsheet or a combination of cloud-based and on-premises hybrid data warehouses, Power BI allows you to:

1. Connect: Easily link to your data sources.
2. Visualize: Create compelling visualizations and explore what matters most.
3. Share: Collaborate by sharing insights with colleagues or stakeholders.

MAIN PHASES IN BUSINESS/DATA ANALYSIS

- DefineObjective
- DataAggregation
- DataCleaning
- AnalyticalMethodology
- EvaluationandValidation
- ReportingandDataVisualisation



Analysis of Investment Patterns and Preferences

Why using Python?

Python is a high-level, interpreted, multi-purpose programming language. Many programming paradigms like procedural programming language, object-oriented programming is supported in python. It can be used for many applications, that includes statistical computing with various packages and functions. Moreover, it is easy to learn. It can be picked up by anyone including those who has less programming skills.

Some features of Python are as listed below:

- Open source and free
- Interpreted language
- Dynamic typesetting
- Portable
- Numerous IDE

Packages used:

- Numpy
- Pandas
- Seaborn

Packages used:

- Numpy
- Pandas
- Seaborn

Dataset used:

- Diwali Sales Dataset

Working with dataset :

- Importing libraries: Libraries that would be used in the process of analysis are to be imported first.
  - Here are the codes to import the libraries.
  - import pandas as pd
  - import numpy as np
  - import matplotlib.pyplot as plt import
  - seaborn as sns

CHALLENGES FACED

Over the course of my internship, I encountered a diverse range of challenges, each offering unique insights and opportunities for growth. Some of the notable challenges I confronted included:

- **Data Quality Issues:** One of the primary challenges was dealing with data quality. The datasets often contained missing values, duplicates, and inconsistencies, which required significant time for cleaning and preprocessing. Handling these issues was crucial to ensure accurate analysis and insights.

- **Handling Large Datasets:** Working with large datasets posed computational challenges. Performing operations on millions of rows often led to slow processing times, making it necessary to optimize code and use efficient data handling techniques, such as vectorization in Pandas.

- **Understanding Complex Data Structures:** Interpreting complex data structures and relationships within the datasets was challenging. It required a deep understanding of the domain and often involved exploratory data analysis (EDA) to uncover hidden patterns and correlations.

- **Choosing the Right Visualization Techniques:** Selecting appropriate visualization techniques to represent the data insights effectively was another challenge. Ensuring that visualizations were not only informative but also easy to understand for stakeholders required careful consideration of the audience and the message to be conveyed.

- **Balancing Multiple Tools and Technologies:** The internship required proficiency in various tools and technologies such as Python, Pandas, Seaborn, and Matplotlib. Balancing the learning curve of these tools while delivering project outcomes on time was a significant challenge.

- **Interpreting Analytical Results:** Interpreting the results of data analysis in a meaningful way that could drive business decisions was a complex task. It involved translating technical findings into actionable insights that could be easily understood by non-technical stakeholders.

- **Time Management:** Balancing the internship workload with other commitments, such as coursework or personal projects, required effective time management. Prioritizing tasks and managing deadlines was essential to ensure the successful completion of the internship.

LEARNING AND OUTCOMES

My internship proved to be invaluable, providing me with extensive hands-on experience while collaborating on a real-world project with my team. This practical exposure has significantly enhanced my project development skills and allowed me to apply my newfound knowledge effectively. Among the key takeaways from this experience are:

- **Enhanced Data Cleaning and Preprocessing Skills:** Through the internship, I gained a deeper understanding of the importance of data quality. I learned various techniques for data cleaning, such as handling missing values, removing duplicates, and dealing with outliers. These skills are crucial for ensuring the reliability of analytical results.

- **Proficiency in Python and Data Analysis Libraries:** My proficiency in Python improved significantly, especially in using data analysis libraries such as NumPy, Pandas, Seaborn, and Matplotlib. I became adept at using these tools to manipulate large datasets, perform statistical analysis, and create compelling visualizations.

- **Improved Problem-Solving Abilities:** Facing challenges such as optimizing code for large datasets and selecting appropriate visualization techniques helped me develop strong problem-solving skills. I learned to approach problems systematically, break them down into manageable parts, and find efficient solutions.

- **Experience with Real-World Datasets:** Working with real-world datasets provided insights into the complexity and messiness of actual data. This experience helped me understand the gap between theoretical knowledge and practical application, preparing me for future roles in data analytics.

- **Effective Communication of Insights:** I learned how to communicate complex analytical findings effectively to non-technical stakeholders. This involved creating clear and concise reports and visualizations that highlighted key insights and actionable recommendations.

- **Time Management and Project Planning:** The need to balance multiple tasks and meet deadlines improved my time management and organizational skills. I learned to prioritize tasks, plan projects effectively, and work efficiently under pressure.

- **Collaboration and Teamwork:** Working with a team taught me the importance of collaboration and communication. I gained experience in coordinating with team members, sharing responsibilities, and providing constructive feedback, which are essential skills in any professional setting.

- **Adaptability to New Tools and Techniques:** The internship required me to quickly learn and adapt to new tools and analytical techniques. This experience has made me more adaptable and open to continuous learning, which is essential in the rapidly evolving field of data analytics.

PROJECT RESULT

Task 1: Data Overview

Objective: Understand the dataset structure.

Steps: Load the dataset:

Import the dataset into a data analysis tool such as Python with pandas or spreadsheet software.

1. Descriptive Statistics: Used descriptive functions (e.g., info() in pandas) to gather information about the number of entries, columns, and data types.

```
RangeIndex: 40 entries, 0 to 39
Data columns (total 24 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   gender                                40 non-null     object
 1   age                                   40 non-null     int64
 2   Investment_Avenues                    40 non-null     object
 3   Mutual_Funds                          40 non-null     int64
 4   Equity_Market                         40 non-null     int64
 5   Debentures                            40 non-null     int64
 6   Government_Bonds                      40 non-null     int64
 7   Fixed_Deposits                        40 non-null     int64
 8   PPF                                   40 non-null     int64
 9   Gold                                  40 non-null     int64
 10  Stock_Marktet                         40 non-null     object
 11  Factor                                40 non-null     object
 12  Objective                             40 non-null     object
 13  Purpose                               40 non-null     object
 14  Duration                              40 non-null     float64
 15  Invest_Monitor                        40 non-null     object
 16  Expect                                40 non-null     float64
 17  Avenue                                40 non-null     object
 18  What are your savings objectives?     40 non-null     object
 19  Reason_Equity                         40 non-null     object
 20  Reason_Mutual                         40 non-null     object
 21  Reason_Bonds                          40 non-null     object
 22  Reason_FD                             40 non-null     object
 23  Source                                40 non-null     object
dtypes: float64(2), int64(8), object(14)
memory usage: 7.6+ KB
```

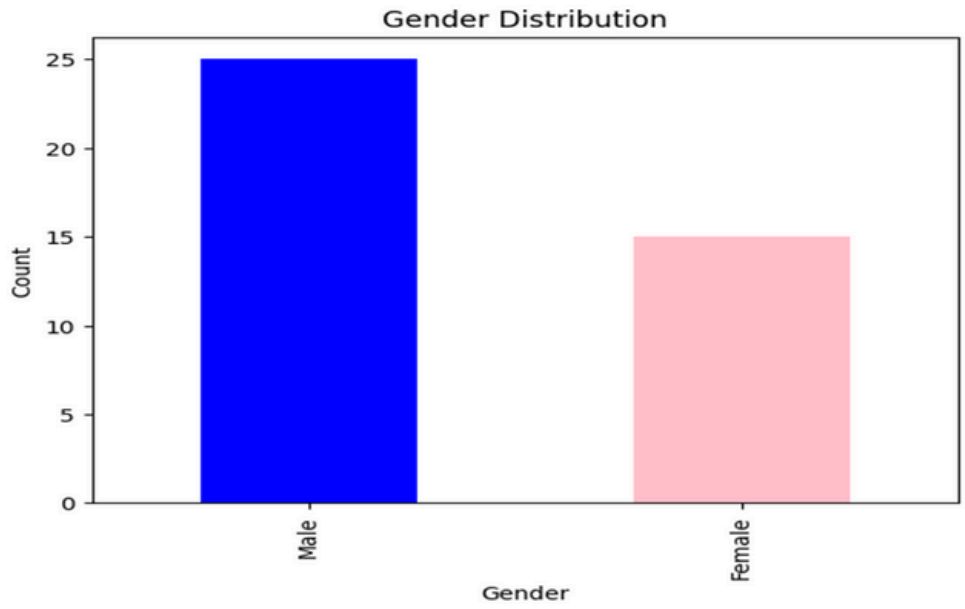
Task 2: Visualize gender distribution in the dataset.

Steps:

Extract Gender Information:

Identify and extract the gender column from the dataset

1. Visualization: Create a simple visualization, such as a bar chart or pie chart, to represent the distribution of genders in the dataset.

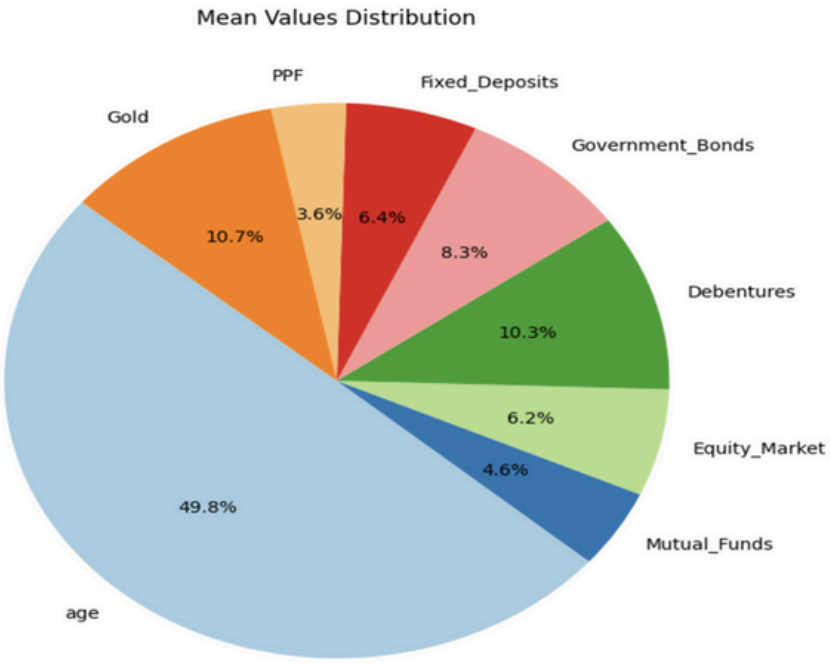


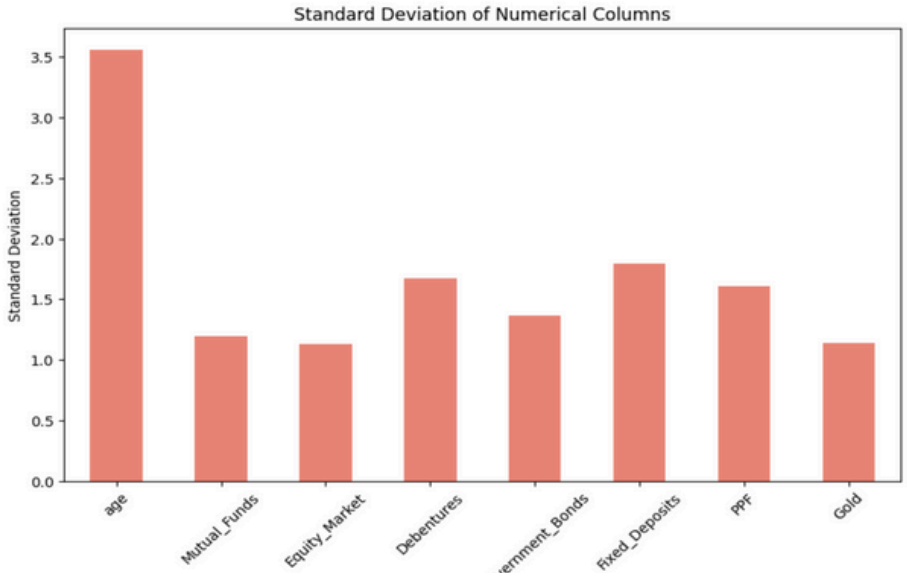
Task 3:Descriptive Statistics Objective: Present basic statistics for numerical columns.

Steps:

Identify Numerical Columns: Reviewed the dataset to identify columns containing numerical data (e.g., age, income).

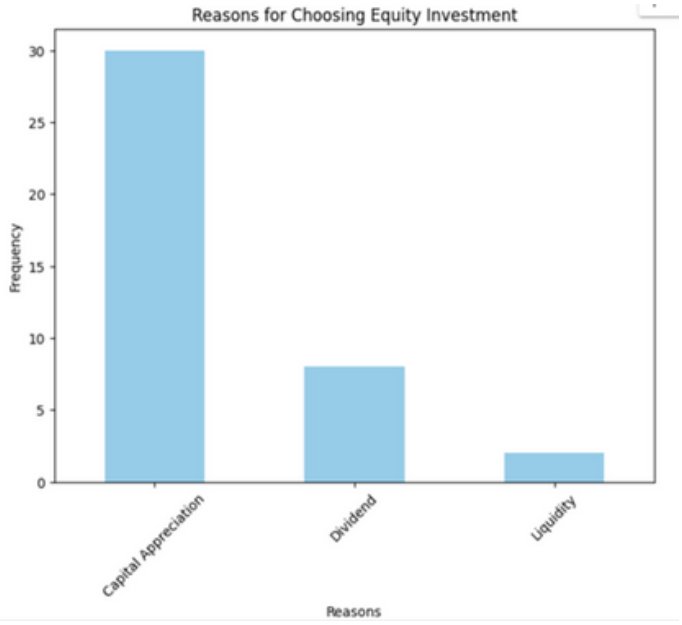
1. Calculations: Used statistical functions (e.g., mean(), median(), std()) to calculate the mean, and standard deviation for each numerical column

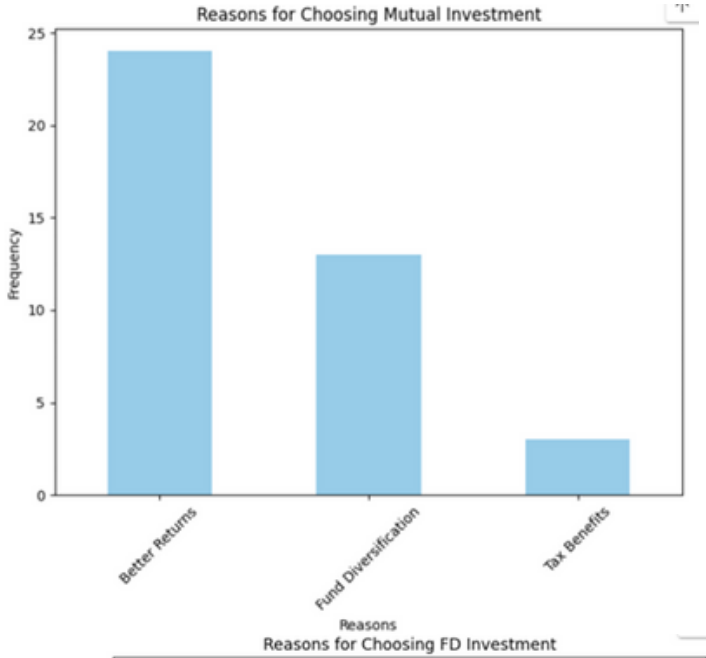


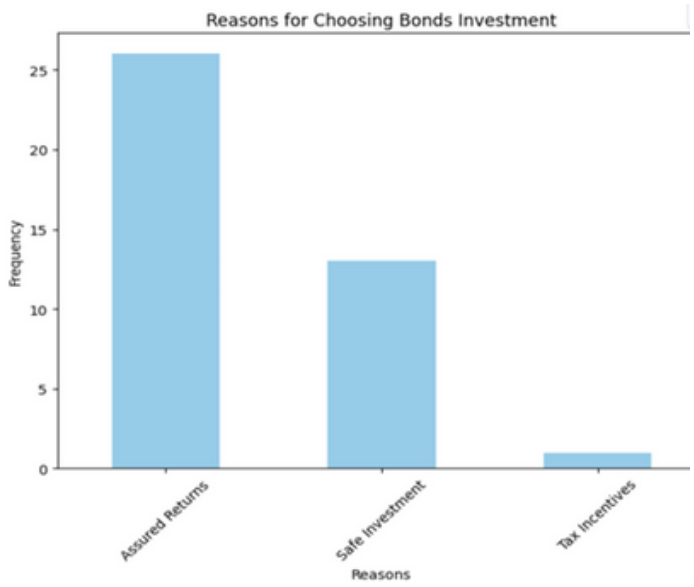


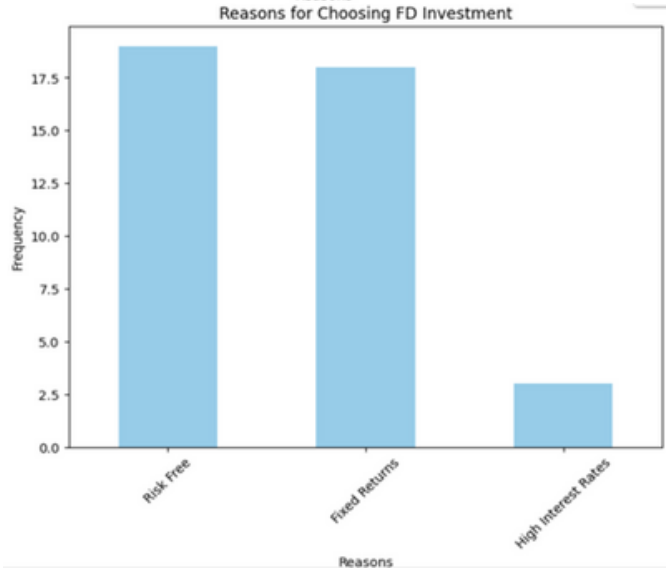
TASK 4: Reasons for Investment

Analyzed the reasons behind choosing different investment avenues.
Summarized the reasons for equity, mutual funds, bonds, and fixed deposits.

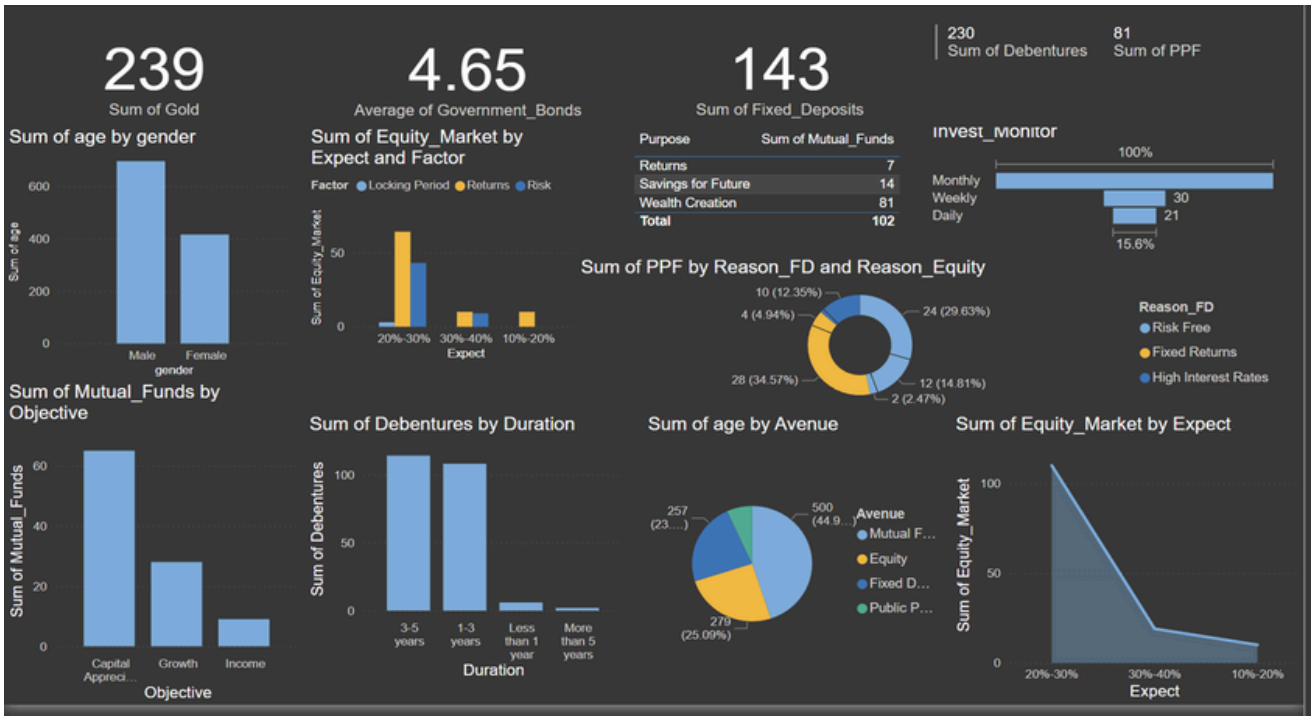








Investment Patterns and Savings Objectives Analysis by using PowerBI



Top Left - "Sum of age by gender":

Explanation: "This chart illustrates the sum of ages distributed across male and female participants, indicating a higher participation from males.“

Top Middle - "Average of Government_Bonds":

Explanation: "The average amount invested in government bonds, showing the stability and preference for secure investments.“

Top Right - "Sum of Fixed_Deposits":

Explanation: "The overall sum invested in fixed deposits, reflecting the preference for low-risk savings options.“

Second Row, Left - "Sum of Mutual_Funds by Objective":

Explanation: "This chart shows the objectives behind investing in mutual funds, with a significant focus on capital appreciation.“

Second Row, Right - "Sum of Debentures by Duration":

Explanation: "The duration of debenture investments, with most participants choosing medium-term investments of 3-5 years."