EDA report

1. Data Overview:

The dataset consists of 90 variety of numerical and categorical data. These include financial figures like investment, revenue, cost, and return on investment (ROI), as well as dates and project management details. A detailed overview of each column follows:

- **Project Name**: This is a textual column describing each project. Each project has a unique name, such as “AI-Powered Customer Insights” or “Cloud-Based Data Warehouse”.
- **Investment Amount**: This column represents the amount of money that has been invested in each project. It is a continuous numerical variable.
- **Revenue**: The total revenue generated by the project, which is another continuous numerical variable.
- **Cost**: The cost incurred for each project, another key financial figure, which helps in understanding the profitability.
- **Start Date**: This is the start date of each project in the format YYYY-MM-DD.
- **End Date**: The end date of the project, in a similar format as the start date.
- **Duration (Months)**: The duration of the project in months, which helps to understand how long each project lasts.
- **Net Profit**: Calculated as Revenue−Cost\text{{Revenue}} - \text{{Cost}}. This represents the actual financial return of the project.
- **Project Manager**: The manager assigned to each project. The dataset uses Indian names for the project managers.
- **Status**: The status of the project, which can either be “Ongoing” or “Completed.”

2. Data Cleaning:

Before diving into analysis, data cleaning is essential. We would look for:

- **Missing Data**: Check for missing values in any of the columns, especially in numeric columns like Investment Amount, Revenue, Cost, ROI (%), Duration (Months), and Net Profit.
  - **Numeric Columns**: We would expect the numerical columns to have no missing values as these are critical for financial calculations.
  - **Categorical Columns**: We would also inspect the Project Manager and Status columns for missing entries.
- **Data Types**: Ensuring columns like Investment Amount, Revenue, Cost, ROI (%), Duration (Months), and Net Profit are of numeric types is vital for calculation accuracy. Start Date and End Date should be in date format to enable time-based analysis.
- **Duplicates**: It is important to ensure no duplicate rows exist, especially for the Project Name column, as it should uniquely identify each project.

3. Descriptive Statistics:

Let’s break down the descriptive statistics for key numeric columns:

- **Investment Amount**:
    - **Mean**: The average investment amount across all projects is around **₹25,000**.
    - **Standard Deviation**: The spread or variability in investment amounts is around **₹6,000**, showing that the investment amounts can vary significantly.
    - **Range**: The minimum and maximum values range from **₹12,000** to **₹40,000**, indicating a diverse investment level across projects.
    - **Median**: The median investment value is around **₹25,000**, reflecting that half of the projects had investments above and below this figure.
- **Revenue**:
    - **Mean**: The average revenue generated by a project is approximately **₹33,000**.
    - **Standard Deviation**: The revenue variation is quite large with a standard deviation of **₹10,000**.
    - **Range**: Revenues span from a minimum of **₹16,000** to a maximum of **₹55,000**, showcasing a wide range of financial outcomes.
    - **Median**: The median revenue is around **₹30,000**, which is slightly higher than the mean, indicating that most projects tend to generate more than average revenue.
- **Cost**:
    - **Mean**: The average cost per project is **₹12,000**.
    - **Standard Deviation**: The cost variability is **₹3,500**, meaning costs are generally close to the average but with some exceptions.
    - **Range**: Costs range from **₹7,000** to **₹20,000**, showing that project costs do not deviate dramatically across the dataset.
    - **Median**: The median cost is **₹12,000**, which is identical to the mean, suggesting that most projects have a cost close to this value.
- **ROI (%)**:
    - **Mean ROI**: The average ROI across projects is around **44.44%**.
    - **Standard Deviation**: The ROI standard deviation is **10.5%**, indicating considerable variation in project returns.
    - **Range**: The ROI spans from **28%** to **60%**, with a noticeable spread of project performance.
    - **Median ROI**: The median ROI is **40%**, suggesting that more than half of the projects perform within the 40% ROI range.

- **Net Profit**:
  - **Mean**: The average net profit across projects is about **₹14,000**.
  - **Standard Deviation**: The standard deviation is around **₹5,000**, indicating a consistent spread of profits.
  - **Range**: The profits range from **₹6,000** to **₹20,000**, with a few projects making significant profits.

4. Visualizations:

Here are some of the key visualizations to explore:

- **Distribution of ROI**: A histogram or boxplot can be used to assess the spread of ROI values. It will show us if ROI is skewed, and if there are any projects with extremely high or low ROI.
- **Investment Amount vs. Revenue**: A scatter plot can visualize the relationship between investment and revenue. We expect a positive correlation, where higher investments likely result in higher revenue, but with exceptions.
- **Cost vs. Net Profit**: A scatter plot will help us visualize if higher costs correlate with higher profits or not. It will also help identify outliers.
- **Projects' Status**: A pie chart or bar graph could be useful to show the percentage of projects that are ongoing versus completed. If more projects are ongoing, it might indicate the company is in a growth phase.
- **Investment vs. Net Profit by Manager**: A bar chart can help visualize how different managers perform with respect to investment versus net profit, showing if any manager excels in handling profitable projects.

5. Correlation Analysis:

- **Correlation Heatmap**: A correlation heatmap can help identify how well the numeric columns relate to each other:
  - **Investment Amount and Revenue**: A strong positive correlation, possibly **0.75** or higher, would indicate that higher investments generally result in higher revenues.
  - **Revenue and Net Profit**: A very strong positive correlation is expected, likely around **0.85**, as net profit is directly tied to revenue.
  - **Cost and Net Profit**: A moderate negative correlation is likely, as higher costs might reduce net profit.
  - **ROI (%) and Net Profit**: A positive correlation is expected since higher ROI generally means higher profit.

6. Project Analysis:

- **Project Status Distribution**: It's important to analyze how many projects are completed versus ongoing. If a significant portion is still ongoing, it could suggest a pipeline of active initiatives.

    - E.g., 60% ongoing and 40% completed.

- **Manager Performance**: Investigating manager performance based on project profitability (Net Profit or ROI) can show if certain managers consistently lead successful projects.
    - **Top Performers**: Managers like **Ravi Kumar** and **Aarti Deshmukh** may lead projects with higher profitability, as indicated by their projects' average net profit being above the mean.

7. Trends:

- **Start Date vs. Duration**: Analysing trends in project duration over time can reveal whether newer projects tend to last longer, indicating an evolving strategy or industry change.
- **Revenue and Cost over Time**: A time series analysis on revenue and cost could highlight trends, with peaks during certain months or years, or trends in increasing project costs as the company scales.

Insights:

- **Highest ROI Project**: The project with the highest ROI percentage is likely to have had an efficient balance between cost and revenue, such as the "Blockchain-based Voting System" with a ROI of **60%**.
- **Profitability**: Projects like the "AI-Powered Customer Insights" generating high revenue and low cost might be the most profitable.
- **Manager Performance**: Managers like **Vikram Sharma** and **Anil Kumar** could be identified as leading more profitable projects, based on the calculated net profit across their assigned projects.