SINCE2008
RED & WHITE
SKILL EDUCATION
SHAPING SKILLS FOR SCALING HIGHER...!!!

Email from Business Analyst:

Subject: Data Analysis and Visualization Tasks for Q1 Sales Dataset

Dear Rahul,

I hope you're doing well. Please find the Q1 Sales dataset attached to this email. Below are the requirements for the data analysis and visualization tasks based on the dataset:

Requirements:

1. **Data Cleaning:**
    - Handle missing values in the dataset. Please decide the best approach (e.g., imputation or removal) based on the column.
    - Ensure that the 'Date' column is in the correct format and extract the month and year as separate columns.
    - Remove any duplicate rows from the dataset.

2. **Summary Statistics:**
    - Provide the basic summary statistics (mean, median, mode, standard deviation) for numerical columns like sales, revenue, and quantity.
    - Calculate the correlation matrix for all numerical columns and visualize it using a heatmap.

3. **Top-performing Products:**
    - Identify the top 5 products with the highest revenue.
    - Plot a bar chart displaying the top 5 products by total revenue.

4. **Sales Performance over Time:**
    - Plot a line chart to visualize the monthly sales trend across Q1.
    - Add annotations to the line chart showing the highest and lowest sales months.

5. **Geographical Insights:**
    - Group sales by region and visualize the total revenue per region using a bar chart.

- Identify the region with the highest sales growth and explain the trend.

6. **Customer Behavior Analysis:**
   - Group customers based on total purchase frequency and average order value.
   - Categorize them manually into three types: Low, Medium, and High Value based on simple thresholds.
   - Visualize this classification using a scatter plot (x-axis: Purchase Frequency, y-axis: Average Order Value) with different colors or markers.

Attachments:

Dataset - **Q1_Sales_Data.csv**

Please feel free to reach out if you need any clarifications. Looking forward to seeing the analysis and visualizations.

Best Regards,
Anupam Shah
DiscoverData Coorp.

Workflow for the Data Analyst:

1. **Load the Dataset:**
   - The analyst starts by loading the dataset from the provided attachment
   - Use `pandas` to load the dataset and inspect the first few rows.

2. **Data Cleaning:**
   - Handle missing values using `fillna()` or `dropna()`.
   - Convert the 'Date' column to datetime format using `pd.to_datetime()`.
   - Extract 'Month' and 'Year' columns from 'Date'.
   - Remove duplicate rows using `drop_duplicates()`.

3. **Summary Statistics:**
   - Calculate the mean, median, mode, and standard deviation for numerical columns.
   - Generate the correlation matrix and visualize it using a heatmap.

4. **Top-performing Products:**
    - ○ Group by 'Product' and sum the 'revenue' to identify the top 5 products by total revenue.
    - ○ Plot a bar chart for the top 5 products.

5. **Sales Performance over Time:**
    - ○ Group by 'Month' and sum the 'sales' to visualize the monthly sales trend.
    - ○ Annotate the highest and lowest sales months.

6. **Geographical Insights:**
    - ○ Group sales by 'Region' and sum the 'revenue'.
    - ○ Plot the total revenue per region.

7. **Customer Behavior Analysis:**
    - ○ Group customers based on total purchase frequency and average order value.
    - ○ Categorize them manually into three types: Low, Medium, and High Value based on simple thresholds.
    - ○ Visualize this classification using a scatter plot (x-axis: Purchase Frequency, y-axis: Average Order Value) with different colors or markers.

In this situation, the data analyst will be performing data cleaning, statistical analysis, and generating visualizations to answer business questions related to sales performance, top-performing products, geographical insights, and customer segmentation. All tasks are accomplished using Python libraries like Pandas, Numpy, Matplotlib, and Seaborn (if needed).