Sales Analysis Project

Steps Performed

1. Data Loading and Merging

- Imported required Python libraries: **pandas, os, matplotlib**.
- Read sales data files for each month from the given directory.
- Concatenated 12 months of data into a single dataframe.
- Exported the merged dataset to `all_data.csv`.

2. Data Cleaning

- **Removed Null Values**: Dropped rows with missing values.
- **Removed Erroneous Entries**: Eliminated rows where the "Order Date" column contained invalid strings such as "Or".
- **Data Type Conversion**: Converted:
  - `Quantity Ordered` → integer
  - `Price Each` → float
  - `Order Date` → datetime

3. Feature Engineering

- **Month Column**: Extracted month from the `Order Date` to facilitate monthly sales analysis.
- **Sales Column**: Created a new column `Sales` by multiplying `Quantity Ordered * Price Each`.
- **City Column**: Extracted `City` and `State` information from the `Purchase Address` field and combined them into a `City` column.

- **Time Features**: Extracted `Hour` and `Minute` from the `Order Date` to analyze order patterns by time.

4. Exploratory Data Analysis (EDA) and Questions Answered

Question 1: What was the best month for sales?

- Grouped data by `Month` and calculated total `Sales`.
- **Result**: December had the highest sales.

Question 2: What city had the highest number of sales?

- Grouped data by `City` and calculated total `Sales`.
- **Result**: San Francisco (CA) had the highest sales.

Question 3: What time should we display advertisements to maximize likelihood of purchases?

- Analyzed order frequency by hour.
- **Result**: Orders peaked at **11 AM** and **7 PM**.
- **Recommendation**: Advertisements should target these hours.

Question 4: What products are most often sold together?

- Identified orders with duplicate `Order ID`s.
- Grouped products bought in the same order.
- Used `itertools combinations` and `Counter` to find frequent pairs.
- **Result**: iPhone and Lightning Charging Cable were the most frequently bought together.

Question 5: What product sold the most? Why?

- Grouped data by `Product` and analyzed `Quantity Ordered`.
- Compared quantity sold with average product price.

- **Result**: Triple A Batteries (4-pack) were the top-selling product.
- **Possible Reasons**:
  - Low price
  - Non-reusable nature
  - High necessity across household appliances

Conclusion

Through systematic data cleaning, feature engineering, and exploratory data analysis, this project identified key business insights from sales data. The findings revealed that:

1. **December was the best-performing month.**
2. **San Francisco was the top city for sales.**
3. **Peak order times occurred at 11 AM and 7 PM**.
4. Product bundling analysis highlighted the importance of accessory sales with premium items.
5. Sales volume analysis emphasized the strong demand for low-cost, essential products like batteries.