Task 1: Forecasting the sales of a supermarket

Task 1 Directive: Perform EDA and Predict the sales of the next 7 days from the last date of the Retail dataset of a global superstore for 4 years!

Environment selected for Analysis:

- **Visual Code**
    - **Jupyter Notebook**
        - **Python**
            - **Libraries ->** pandas, numpy, seaborn, matplotlib, calender

Exploratory data analysis (EDA) Steps

Raw Code

1. Imported important libraries.
2. Found the shape of the dataset and printed the first 5 rows.
3. Identified numerical and categorical variables.
4. Checked for missing values and dropped them from the dataset.
5. Ensured data consistency and checked data types.
6. Detected date formats for 'Order Date' and 'Ship Date' columns.
7. Created a new column, 'Order Processing Time', by subtracting 'Order Date' from 'Ship Date'.
8. Identified outliers using 'Order Processing Time'.
9. Checked for negative values.
10. Checked the correlation of columns using a correlation matrix.
11. Plotted histograms of all numerical and categorical columns.
12. Analyzed order trends by year, month, and day.
13. Used an ARIMA model to predict the next 7 days' order count and plotted it.
14. Analyzed segment-wise order distribution and predicted the segment with the highest number of orders for the next 1 week.
15. Analyzed category and sub-category-wise order distribution and predicted the highest order categories for the next 1 week.
16. Analyzed product-wise order distribution and predicted the top products for the next 1 week.
17. Analyzed region, state, and city-wise order distribution and predicted the highest order regions, states, and cities for the next 1 week.
18. Identified and printed the top 5 customers based on their Customer ID, Customer Name, Total Purchase Value, Full Address, and top product preferences.
19. Identified the top 100 potential targetable customers based on their potentiality score, name, and ID.
20. Calculated year-to-date sales and plotted them in a line chart.
21. Calculated month-to-date sales and plotted them in a line chart.
22. Calculated and printed sales by segment, year, category, and sub-category.
23. Performed statistical analysis of sales.
24. Forecast sales for the next 1 week and visualize them in a line chart.
25. Calculated and plotted sales by ship mode.
26. Investigated the relationship between order processing time and sales.
27. Segmented customers into 'Low-Value', 'Mid-Value', and 'High-Value' based on total purchase value, purchase frequency, and recency.
28. Calculated the top 10 customers by CLV and displayed them in a table format.

Performed by: Bishnu Barman

Part of Business Analytics Internship from Bharat Intern

Findings

- Raw Code [https://drive.google.com/file/d/1hG8-nKLuoXRVYsF8h2gfM1rOy49DnOu5/view?usp=sharing]

- ### Findings Summary

1. **Dataset Information:**
    a. Shape of the dataset: (9800, 18)
    b. Numerical variables: ['Row ID', 'Postal Code', 'Sales']
    c. Categorical variables: ['Order ID', 'Ship Mode', 'Customer ID', 'Customer Name', 'Segment', 'Country', 'City', 'State', 'Region', 'Product ID', 'Category', 'Sub-Category', 'Product Name']
    d. Missing values: The Postal Code has 11 missing values.
    e. Data types: The dataset contains a mixture of int64, object, datetime64, and float64 data types.
    f. Size of the cleaned dataset: (9789, 18)

2. **Order Processing Time:**
    a. A new column "Order Processing Time" was added to the dataset, representing the number of days it took to process each order.

3. **Orders by Year:**
    a. The number of orders in each year: 2015 (1953), 2016 (2055), 2017 (2534), 2018 (3258)

4. **Predictions for the Next 1 Week:**
    a. Segment with the highest number of orders: Consumer
    b. Category with the highest number of orders: Office Supplies
    c. Sub-Category with the highest number of orders: Binders
    d. Product Name with the highest number of orders: Staple envelope
    e. Region with the highest number of orders: West
    f. State with the highest number of orders: California
    g. City with the highest number of orders: New York City

5. **Top Customers:**
    a. Top five customers based on their total purchase value:
    b. Customer ID: AB-10105, Full Name: Adrian Barton, Total Purchase Value: $14473.57, Segment: Consumer, Address: Phoenix, Arizona, 85023
    c. Customer ID: RB-19360, Full Name: Raymond Buch, Total Purchase Value: $15117.34, Segment: Consumer, Address: Auburn, New York, 13021
    d. Customer ID: SM-20320, Full Name: Sean Miller, Total Purchase Value: $25043.05, Segment: Home Office, Address: Monroe, North Carolina, ...

6. **Top Product Preferences:**
    a. For each customer, the top five product preferences are listed based on their order history.