**Exploratory Data Analysis (EDA) for Big Mart Outlets Sales Data**

The data scientists at BigMart have collected 2013 sales data for 1,559 products across 10 stores in different cities. Additionally, they have defined various attributes for each product and store.

**GOAL:**The goal is to perform an Exploratory Data Analysis (EDA) to gain insights into the data and to understand the underlying patterns.

**OBJECTIVES:**

1. **Data Understanding:**
   * **Overview:** Get a general sense of the dataset, including the number of records, features, and types of attributes.
   * **Attributes:** Identify and understand the attributes of products and stores, such as sales, discounts, and store types.
2. **Data Cleaning:**
   * **Missing Values:** Detect and handle missing values in the dataset.
   * **Outliers:** Identify and address outliers that could impact analysis.
   * **Data Types:** Ensure that data types are correct for each feature (e.g., categorical vs. numerical).
3. **Descriptive Statistics:**
   * **Summary Statistics:** Calculate summary statistics for numerical features (mean, median, standard deviation).
   * **Distribution:** Analyze the distribution of numerical attributes (e.g., sales, discounts).
4. **Data Visualization:**
   * **Histograms and Boxplots:** Visualize the distribution of numerical features and detect outliers.
   * **Bar Charts:** Compare categorical variables such as store types and product categories.
   * **Correlation Heatmap:** Identify correlations between numerical features.

**First, we imported the required libraries:**

import numpy as np

import pandas as pd

import seaborn as sns

from matplotlib import pyplot as plt

Next, we loaded the dataset and checked the available data. We conducted exploratory data analysis by examining the shape and info of the dataset. We also computed the statistical properties of the dataset and identified missing values, calculating their percentage.

We then plotted the histogram of the distribution of Item\_Weight and Outlet\_Size. For univariate analysis, we explored the distribution of the target variable and analyzed categorical variables. We used a count plot to visualize categorical values.

For bivariate analysis, we created a correlation heatmap of numerical data, a scatter plot of Item\_MRP versus Outlet\_Sales, and a box plot for categorical values.

Finally, for multivariate analysis, we generated a pair plot of Item\_Weight, Item\_Visibility, Item\_MRP, and Item\_Outlet\_Sales.

**Insights from the EDA:**

1. **General Data Overview:**
   * **Dataset Size:** Understanding the number of rows and columns can give you an idea of the dataset's complexity and scope.
   * **Data Types:** Checking the data types ensures that attributes are in the correct format for analysis.
2. **Missing Values:**
   * **Missing Data:** Identifying and quantifying missing values helps in deciding how to handle them (e.g., imputation, removal).
   * **Percentage of Missing Values:** This metric indicates the extent of missing data and its potential impact on analysis.
3. **Distribution Analysis:**
   * **Item Weight Distribution:** The histogram for Item\_Weight can reveal the spread and central tendency of item weights. It might show if most items fall within a certain weight range or if there are any skewed distributions.
   * **Outlet Size Distribution:** Understanding the distribution of Outlet\_Size helps in assessing the variety and size of stores.
4. **Univariate Analysis:**
   * **Target Variable Distribution:** Analyzing the distribution of the target variable (Item\_Outlet\_Sales) shows its range, central tendency, and potential skewness.
   * **Categorical Variables:** The count plot for categorical values helps in understanding the frequency of different categories, which can indicate the distribution of products and stores.
5. **Bivariate Analysis:**
   * **Correlation Heatmap:** This visualizes the relationships between numerical features, helping to identify strong correlations (e.g., between Item\_MRP and Item\_Outlet\_Sales). It can reveal patterns such as whether higher item prices are associated with higher sales.
   * **Scatter Plot of Item MRP vs. Outlet Sales:** This plot shows how item prices relate to sales figures, which can highlight trends, clusters, or outliers.
   * **Box Plot for Categorical Variables:** This visualizes the distribution of sales across different categories, helping to identify which categories perform better or worse.
6. **Multivariate Analysis:**
   * **Pair Plot of Key Variables:** The pair plot for Item\_Weight, Item\_Visibility, Item\_MRP, and Item\_Outlet\_Sales provides a comprehensive view of relationships between these features. It helps in spotting patterns, correlations, and potential interactions between variables.

**Key Insights:**

* **Sales Trends:** You may discover that higher item prices generally lead to higher sales, or that certain stores or categories are performing better.
* **Impact of Weight and Visibility:** Insights into how item weight and visibility influence sales can be useful for optimizing product placements and marketing strategies.
* **Store Performance:** Differences in sales across various store sizes or types can help in identifying which store formats are more successful.
* **Outliers and Anomalies:** Outliers in sales data or item weight might suggest unusual patterns or data quality issues that need further investigation.

These insights can guide decision-making and strategic planning for optimizing sales and store performance.