# Sales Analysis

This project performs an in-depth analysis of e-commerce sales data, focusing on data quality, comprehensive exploratory data analysis (EDA), visualization, and applying a statistical model (Simple Linear Regression) to test a specific economic hypothesis.

## Project Goals

1. **Data Clean-up and Detailed EDA:** Validate data quality, handle missing values, and generate detailed descriptive statistics using Pandas and NumPy.
2. **Data Visualization:** Create meaningful charts to illustrate sales trends, product performance, and variable distributions.
3. **Statistical Analysis & Hypothesis Testing:**
   * Formulate a testable hypothesis regarding the relationship between core sales variables.
   * Use correlation and p-value analysis to statistically test the hypothesis.
   * Build a Simple Linear Regression model to model the relationship and interpret its significance.

## Files in this Project

* Sales\_Data.csv.csv: The raw sales transaction data.
* sales\_analysis\_enhanced.py: The Python script containing all the data cleaning, EDA, visualization, and statistical modeling logic.
* README.md (This file): Project overview and execution guide.

## Prerequisites

* Python 3.x
* The following Python libraries:
  + pandas
  + numpy
  + matplotlib
  + seaborn
  + statsmodels (for statistical modeling and hypothesis testing)

You can install the required libraries using pip:

pip install pandas numpy matplotlib seaborn statsmodels

## How to Run the Analysis

1. Ensure the Sales\_Data.csv.csv file is in the same directory as sales\_analysis\_enhanced.py.
2. Execute the Python script from your terminal:  
   python sales\_analysis\_enhanced.py
3. The script will print the detailed EDA, correlation results, the summary of the Linear Regression model, and display the generated plots.

## Analysis Steps & Explanation

The sales\_analysis\_enhanced.py script executes the following stages:

### 1. Data Loading and Clean-up

* **Load Data:** The CSV file is loaded into a Pandas DataFrame.
* **Initial Inspection:** df.info() and df.head() are used to check data types and initial structure.
* **Missing Data Handling:** Check for and remove/impute any missing values. (In this dataset, we drop rows with any missing data for simplicity).
* **Type Conversion:** Ensure Date is a datetime object.
* **Feature Engineering:** A crucial new column, Total Sales (Quantity \* Price), is calculated using NumPy array multiplication for efficiency.

### 2. Exploratory Data Analysis (EDA)

* **Descriptive Statistics:** df.describe() is used on all numerical columns (Quantity, Price, Total Sales) to find mean, median, standard deviation, and quartiles.
* **Categorical Analysis:** Unique counts of Product and Category are inspected.
* **Time Series Analysis:** Monthly sales are aggregated to identify overall trends and seasonality.

### 3. Data Visualization

* **Monthly Sales Trend:** A line plot to show sales performance over time, identifying peak seasons.
* **Price Distribution:** A histogram of the Price column to understand product pricing structure.
* **Sales by Category:** A bar plot showing which product categories generate the most revenue.
* **Regression Plot:** A scatter plot with a linear fit line is generated to visualize the relationship between the two variables used in the hypothesis.

### 4. Statistical Analysis & Linear Regression

#### Hypothesis Formulation

Based on the economic **Law of Demand**, the hypothesis is:

**H0 (Null Hypothesis):** There is **no significant linear relationship** between the Price of a product and the Quantity sold per transaction.

**H1 (Alternative Hypothesis):** There **is a significant negative linear relationship** between the Price of a product and the Quantity sold per transaction. (i.e., as price increases, quantity sold decreases).

#### Correlation and Test

* **Pearson Correlation Coefficient:** Calculated between Price and Quantity to measure the strength and direction of the linear relationship.
* Linear Regression Model: A Simple Linear Regression model is built:  
  $$\text{Quantity} = \beta\_0 + \beta\_1 \cdot \text{Price} + \epsilon$$

#### Interpreting Results

* **P-value of** $\beta\_1$ **(Price Coefficient):** If the p-value is less than $0.05$, we reject the Null Hypothesis ($\text{H}\_0$), suggesting that the relationship is statistically significant.
* **Coefficient (**$\beta\_1$**):** A negative coefficient supports the Law of Demand (H1).
* **R-squared:** Measures the proportion of the variance in Quantity that is predictable from Price.

## Key Insights from Analysis

*(The actual values depend on the specific data, but the script is designed to extract these insights)*

* **Seasonality:** Peak sales are generally observed in [**Month X**] and [**Month Y**], indicating a potential holiday or promotional window.
* **Demand:** The Linear Regression analysis confirmed that the relationship between Price and Quantity is [**statistically significant/not significant**] and the correlation is [**positive/negative**], providing evidence [**for/against**] the Law of Demand in this specific dataset.