**PREDICTING RETAILER PROFITABILITY: USING LINEAR REGRESSION**

**INTRODUCTION**

This data science project is a group project with an aim to build a linear regression model that accurately forecasts operating profit based on key factors like total sales. Through teamwork and knowledge sharing, we strive to unlock data- driven insights and also demonstrate the effectiveness of collaboration in achieving impactful data science solutions.

The data was obtained from Kaggle website, the dataset contains sales information for various beverage brands across different retailers, regions, states, and cities. The sales information in the dataset describes some key financial metrics for the group of retailers.

**PROBLEM STATEMENT**

We would be examining the relationship between sales and profit thereby creating a model to test and predict the operating profit of a retailer based on their total sales. Operating Profit is a crucial metric for understanding the profitability of a particular sale, and its prediction can help businesses optimize their operations, pricing strategies, and sales forecasts.

**OBJECTIVE**

To analyze the relationship between Sales metrics to determine how they influence the operating profit of different retailers. Specifically, the goal is to build a linear regression model to predict the Operating Profit based on the Total Sales for each retailer.

**DESCRIPTION OF COLUMNS**

Retailer: The name of the retailer where the sale occurred (e.g., BevCo, Walmart, Amazon).

Retailer ID: A unique identifier for each retailer in the dataset. This column distinguishes one retailer from another.

Invoice Date: The date on which the sale was made (e.g., 1/1/2022).

Region: The geographical region of the sale (e.g., Northeast, South).

State: The state in which the sale occurred (e.g., New York, Texas).

City: The city in which the sale occurred (e.g., New York, Houston).

Beverage Brand: The brand of the beverage sold (e.g., Coca-Cola, Sprite, Dasani Water).

Price per Unit: The price at which each unit of a product is sold by the retailer. This metric reflects the pricing strategy of the retailer.

Units Sold: The number of units sold (formatted with commas as thousand separators).

Total Sales: The total revenue generated by the retailer from selling its products. This is a measure of the retailer's sales performance.

Operating Profit: The profit earned by the retailer from its core business operations, excluding non-operating expenses like interest and taxes. This metric indicates the retailer's profitability.

Operating Margin: The percentage of revenue that remains as operating profit after accounting for the costs of goods sold and operating expenses.

[ ]

**THE LIBRARIES EMPLOYED**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import sklearn

from sklearn. Model selection import train test split

from sklearn. Linear model import Linear Regression

from sklearn. metrics import mean\_squared\_error, r2\_score

from sklearn. metrics import mean\_absolute\_error