SIMPLE LINEAR REGRESSION

Introduction

This project explores pricing consistency within an e-commerce sales dataset sourced from Kaggle's "Unlock Profits with E-Commerce Sales Data". The goal is to understand how reliably Quantity (Qty) predicts Amount in Indian Rupee for individual SKUs, and to identify which products exhibit stable, linear pricing behaviour.

I approached this by cleaning the data, isolating SKUs with meaningful variation in Qty, and running simple linear regressions to evaluate how strongly Amount scales with Quantity. By focusing on SKUs with sufficient data and variation, I could assess pricing patterns and validate whether unit-based pricing influences sales.

Analytics Question: What is the estimated sale amount (INR) for a single-unit purchase of this product?

Data Source: This analysis uses the Unlock Profits with E-Commerce Sales Data dataset from Kaggle, which contains transactional sales records including SKU (product code), quantity, and amount fields for exploratory and regression modelling. Click here to ***Amazon Sale Report.csv***

Linear Regression

Simple linear regression models the relationship between one numeric predictor (X) and one numeric outcome (Y) using a straight-line equation. It estimates how changes in X are associated with changes in Y.

Key requirements

1. Both variables must be continuous numeric
2. The relationship between X and Y should be approximately linear

Variables in this project

- X (independent or input or predictor variable) = Qty: Quantity of the product (integer)
- Y (dependent or outcome or response variable) = Amount: Amount of the sale (float)

What the model answers

1. What is the expected value of Amount when Qty is X?
2. Given a quantity sold, predict the sale amount.

Why it works here

For SKUs with stable pricing, Amount scales almost perfectly with Qty, making linear regression an ideal tool to quantify that relationship.

Data Preparation (Excel)

1. Initial Cleaning

- Remove rows where Qty is blank or zero.
- Remove rows where Amount is blank or zero.

Assumption: each row must represent a valid transaction with at least one unit sold and a non-zero amount.

2. Status Filtering

- Exclude all rows marked Cancelled to avoid non-completed orders.

3. Column Pruning

- Removed columns not required for regression analysis

Data Preparation (R)

- Double check - Removed rows with blank or zero Qty or Amount
- Grouped data by SKU to inspect unique products
- Identified a SKU with sufficient Qty variation
- Filtered the dataset to include only that SKU
- Ran a Simple Linear Regression: Amount ~ Qty

Analysis

- **Simple Linear Regression**

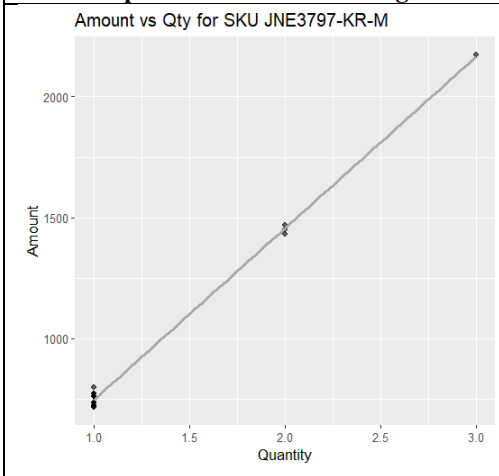
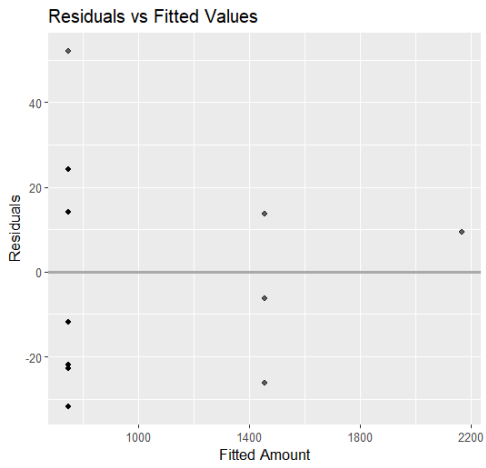
```
SLR_model <- lm(Amount ~ Qty, data = data_file_3)

Summary (SLR_model)

Plot (SLR_model)
```

Plots

- Scatter plot with fitted linear regression line
- Residuals vs Fitted Values diagnostic plot

Scatter plot with fitted linear regression line	Residuals vs Fitted Values diagnostic plot
	

Results:

- Intercept (37.43): Baseline model prediction when Qty = 0; serves as a mathematical anchor rather than a real sales scenario.
- Slope (709.40): Each 1-unit increase in Qty increases Amount by approximately 709.40.
- p-value ($< 2 \times 10^{-16}$): Indicates an extremely strong and statistically significant relationship between Qty and Amount.
- Residuals: Tight residuals show predictions closely match observed values.
- R^2 (0.946): Qty explains 94.6% of the variation in Amount, indicating an excellent model fit.

Analytics Answer: The model estimates a single-unit sale at approximately 709.40 INR

Comments:

The model fits the observed data, but the predictor variable (Qty) shows limited variation, with only three distinct levels, which weakens the regression's explanatory power and limits its ability to generalize beyond the observed range; selecting SKUs with broader Qty variation, ideally five or more distinct levels, would support more robust and generalizable regression models.