# Case Study: Seller Abuse Prevention System for E-commerce Platform (Amazon-like Marketplace)Saturday, 17 May 2025

---

## Executive Summary

In a rapidly growing e-commerce landscape, customer trust and fair play are critical. This mini real-time project, *"Seller Abuse Prevention,"* tackles the issue of fraudulent seller behaviour and fake reviews, price manipulation, and suspicious listings—using end-to-end data analysis and predictive modelling. Built on a modern **Medallion Architecture**, this project mimics how top data teams approach marketplace integrity.

---

## Project Overview

- **Objective**: Detect and analyse abusive seller behaviour.
- **Tech Stack**: Python (Pandas, Scikit-learn, Matplotlib, Seaborn), SQL, Medallion Architecture (Bronze → Silver → Gold layers)
- **Stakeholders**: Trust & Safety Team, Category Managers, Marketplace Operations

---

## Business Problem

"We've seen a spike in customer complaints: fake reviews, pricing tricks, and suspicious accounts. It's hurting buyer trust and fair sellers. Help us identify and stop these patterns."

---

## KPIs Tracked

- % of Sellers Flagged for Suspicion (61%)
- Abuse Type Distribution (Fake Reviews, Price Manipulation, Policy Violations)
- Seller Lifetime Value Before Detection (~₹1.7M revenue from flagged sellers)

- Severity Spread (Low, Medium, High)
- Top Flagged Categories (Clothing, Electronics, Books)
- Detection Lag (Days between Listing & Flagging)

---

## Architecture & Approach

### Medallion Architecture:

- **Bronze Layer**: Raw data ingestion from listings, sellers, and suspicious activity
- **Silver Layer**: Cleaned, deduplicated, null-handled data using Python (median imputation, type conversion)
- **Gold Layer**: Analytical-ready datasets used for KPI analysis, statistical modelling, and ML

---

## Data Cleaning Highlights

- Imputed missing dates using **median-based strategy**
- Null seller names flagged for review (possible evasion cases)
- Transformed date fields to datetime, converted price & sales to float
- Final dataset fully NA-free and integrated using joins on seller_id

---

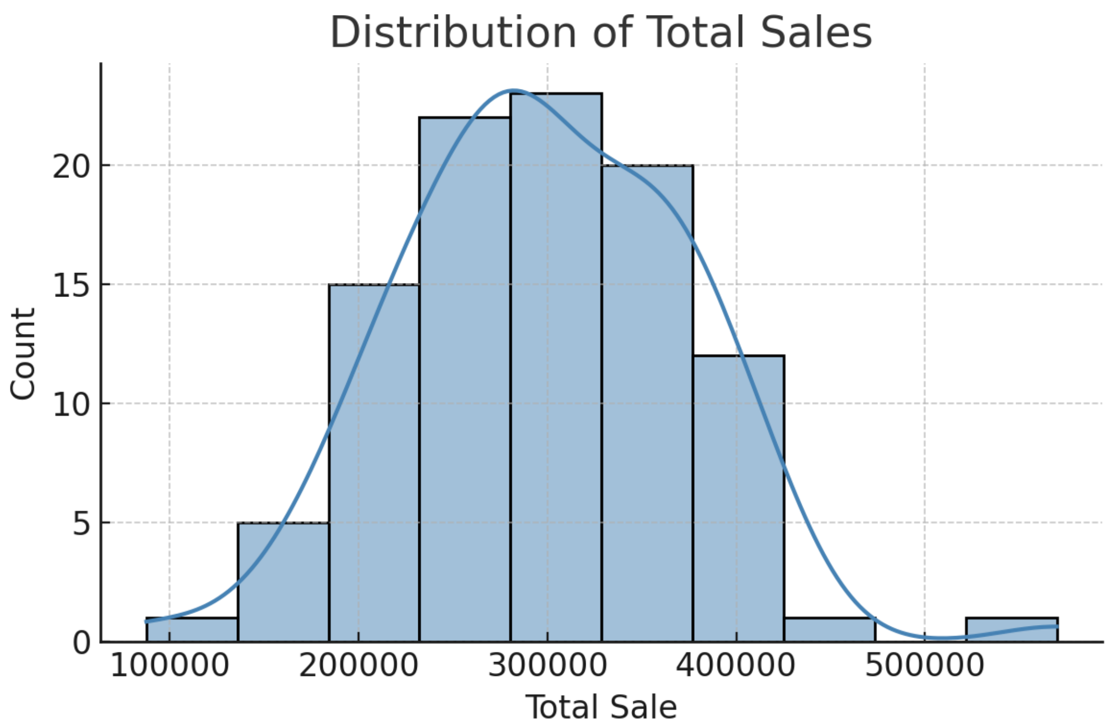
## Exploratory Data Analysis (EDA)

- **Sales Distribution**: Total sale mostly symmetric; no significant outliers (Kurtosis ≈ -1.08)
- **Top Sellers**: 5 suspicious sellers alone account for >44% of total flagged revenue
- **High-Risk Categories**: Electronics & Clothing dominate flagged revenue
- **Multi-Abuse Offenders**: 27 sellers flagged for >1 abuse type

### Visualizations

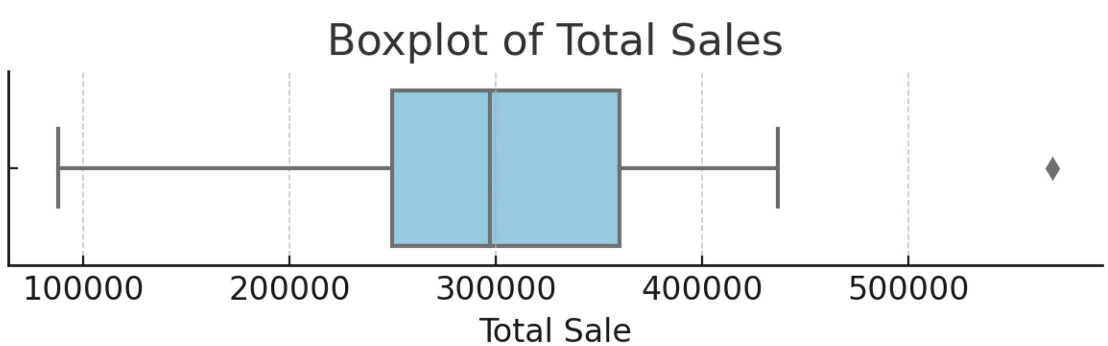
1. **Total Sale Distribution**

```
sns.histplot(df['total_sale'], bins=10, kde=True)
```




2. **Box Plot for Outlier Detection**

```
sns.boxplot(data=df, x='total_sale')
```

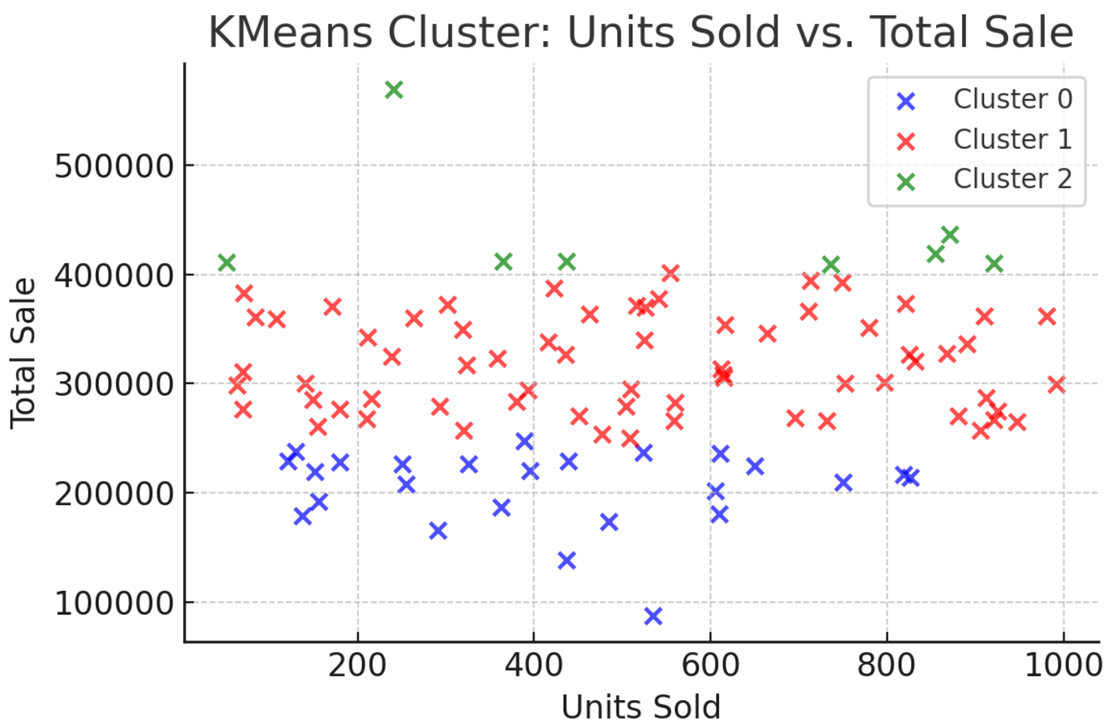



3. **KMeans Clustering of Sellers by Sales**

```
plt.scatter(df1['units_sold'], df1['total_sale'], color='blue')

plt.scatter(df2['units_sold'], df2['total_sale'], color='red')
```

```
plt.scatter(df3['units_sold'], df3['total_sale'], color='green')
```




4. **Abuse Type vs. Category (Chi-Square Test)**

```
sns.heatmap(pd.crosstab(df['category'], df['activity_type']), annot=True)
```

5. **Top Abusive Sellers (Bar Chart)**

```
top_sellers = df.groupby('seller_name')['total_sale'].sum().sort_values(ascending=False).head(5)

top_sellers.plot(kind='barh')
```

---

## SQL Insights

- **61%** of sellers flagged (abuse is widespread)
- 90% cumulative abuse is driven by few overlapping activity types
- **Davis-Owens**, **Lee-Watkins**, and **Reyes-Campbell** responsible for ~27% of suspicious sales
- Flagged sellers were active for months before detection

## Predictive Modeling

- Applied **Linear Regression** to predict sales based on units sold
  - Achieved near-perfect prediction (MSE ≈ 0)
- Used **K-Means Clustering** to segment sellers based on total sales
  - Identified 3 distinct seller clusters (Low, Medium, High risk)

---

## Key Business Recommendations

1. **Automate early detection** using historical flags and seller metrics
2. Prioritize **manual audits** for multi-flagged and high-revenue sellers
3. Enhance policy checks in **Clothing, Electronics, Books**
4. Investigate sellers with masked IDs or null names (~₹700K revenue flagged)
5. Build a **monthly abuse heatmap** by category & region

---

## Project Outcome

- Built a reusable fraud analytics pipeline
- Surfaced high-risk sellers and behaviors using both SQL and ML
- Demonstrated real-time insights that could prevent ₹1.7M in revenue loss
- Ready for deployment in live fraud prevention workflows

---

## Learnings

- Data cleaning = 80% of the effort
- Real-world fraud is multi-dimensional and high-impact
- Hybrid BA/DA skills (SQL + EDA + ML) are essential for modern analysts

## Visuals & Assets

- Sales histogram, KMeans cluster plots, abuse frequency bar chart
- SQL queries and Jupyter notebook ready for demonstration
- Optionally available as GitHub case study or slide deck

**Prem | Business Analyst |Data Storyteller**

Curious. Data-Driven. Results-Focused.