Statistical Diagnosis of E-Commerce Capital Allocation

Technical Report · 19 February 2026 · Python 3.x (Pandas, Statsmodels, Seaborn) · n=50,000 · α=0.05

$32.87M	$657	$4.59M	1,667	4.22%	50,000
Total Revenue	*Avg. Order Value*	*Discount Upside*	*Category HHI*	*Monthly CV*	*Orders (n)*

SECTION 01

Executive Summary

This report validates the transition from descriptive analytics to active capital allocation across a $32.87M e-commerce portfolio of 50,000 orders. Using a multi-layered statistical framework — OLS regression, price elasticity modelling, ANOVA, and nested F-testing — we replicate the portfolio's revenue architecture and surface a $4.59M margin recovery opportunity concealed within the discount programme.

Two headline findings drive all downstream recommendations. First, the customer base is demonstrably price-inelastic: discounts destroy margin without generating incremental volume. Second, the portfolio is structurally symmetric across both categories and regions, which isolates pricing behaviour as the dominant variable available for strategic intervention. Beyond diagnosis, we identify concentrated 'pockets of alpha' in Middle East order values and Wallet/UPI payment rails that can compound the core discount recovery into a durable structural advantage.

SECTION 02

Core Revenue Architecture

The first task was to establish the primary statistical drivers of portfolio revenue. We employed an OLS regression model incorporating discounted price, quantity sold, and product category as predictors, supplemented by a Type II ANOVA to test whether category membership explains meaningful revenue variation independent of price and volume effects.

2.1 Model Specification

```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

df = pd.read_csv('amazon_sales_dataset.csv')

# OLS — primary revenue drivers
ols_model = ols(
    'total_revenue ~ discounted_price + quantity_sold + C(product_category)',
    data=df
).fit()

# Type II ANOVA — category-level performance variance
anova_model   = ols('total_revenue ~ C(product_category)', data=df).fit()
anova_results = anova_lm(anova_model, typ=2)

print(f"OLS R-Squared : {ols_model.rsquared:.3f}")
print(f"Category ANOVA p-value : {anova_results.loc['C(product_category)',
'PR(>F)']:.4f}")
```

2.2 Results & Interpretation

R^2 = 0.882
Model Fit

Revenue is highly predictable. Discounted price and quantity sold jointly account for 88.2% of total revenue variance, confirming a well-specified linear architecture with minimal noise from omitted variables.

p = 0.918
Category Symmetry

The Type II ANOVA fails to reject category homogeneity at α=0.05. The portfolio is structurally category-agnostic — no product category outperforms or underperforms in a statistically meaningful way. All performance differences are driven by pricing and volume mechanics, not by category selection.

This structural symmetry is the central insight of Section 2: it rules out category reallocation as a lever and focuses all strategic attention on price mechanics — specifically, the discount programme analysed in Section 3.

SECTION 03

The Discount Leak: Price Elasticity Analysis

The most consequential finding in this report is the demonstrated price inelasticity of the portfolio's customer base. The evidence is unambiguous: discounts are transferring margin to customers who would have purchased at full price regardless. Eliminating the discount programme is the single highest-leverage capital reallocation available.

3.1 Log-Log Elasticity Model

A log-log specification estimates the price elasticity of demand — the percentage change in quantity sold for a 1% change in price. Robust HC3 standard errors correct for the heteroskedasticity detected by the Breusch-Pagan test (LM=1,205.0, p<0.001):

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Log-Log elasticity model with robust standard errors (HC3)
df['log_qty']   = np.log(df['quantity_sold'])
df['log_price'] = np.log(df['discounted_price'])

elasticity_model = ols('log_qty ~ log_price', data=df).fit(cov_type='HC3')

# Visualise: discount depth vs. quantity sold — the 'flatline' evidence
sns.regplot(
    data=df,
    x='discount_percent',
    y='quantity_sold',
    scatter_kws={'alpha': 0.1},
    line_kws={'color': 'red'}
)
plt.title('Quantity Sold vs. Discount Percent: The Flatline Evidence')
plt.show()
```

3.2 Quantified Upside

$R^2 = 0.000002$
Elasticity (Qty)

Volume is statistically independent of discount depth (slope=0.000201, p=0.7544). The regression of quantity on discount percentage is a flatline — there is zero evidence that price cuts generate incremental demand at any tier.

$4.59M
Revenue Recovery

Revenue per order peaks at $749 at 0% discount and declines monotonically to $522 at the 30% tier — a verified $227 spread. Restoring full pricing on 41,784 discounted orders implies a +17.2% AOV uplift equating to approximately $4.59M in recovered revenue.

A nested F-test confirms the inelasticity is global: the discount effect does not vary by region ($F(3,\sim50k)=1.07$, $p=0.359$). No regional carve-out or tiered approach is warranted. The recommended action is full elimination of all discount tiers with a 30-day monitored rollback protocol to validate the causal assumption before permanent commitment.

The price-inelastic cohort — 16.4% of orders already transacting at full price, AOV $749 versus $639 discounted — must be identified, tagged in the CRM, and permanently shielded from any discount prompt. This group is the portfolio's highest-margin segment and the foundation of the elimination strategy.

SECTION 04

Regional Alpha: Middle East Tier-1 Analysis

While the portfolio exhibits category-level symmetry, the regional dimension reveals a directional concentration of high-value order behaviour. A one-way ANOVA was conducted to test whether regional average order value differences are statistically significant at α=0.05.

4.1 Model Specification

```
# Regional AOV ranking
regional_aov = (
    df.groupby('customer_region')['total_revenue']
    .mean()
    .sort_values(ascending=False)
)
overall_aov = df['total_revenue'].mean()

# Box plot — regional order value distributions
plt.figure(figsize=(12, 6))
sns.boxplot(
    data=df,
    x='customer_region',
    y='total_revenue',
    order=regional_aov.index
)
plt.axhline(overall_aov, color='red', linestyle='--',
            label=f'Overall AOV (${overall_aov:.2f})')
plt.title('Regional AOV Distribution (Middle East as Tier 1)')
plt.show()
```

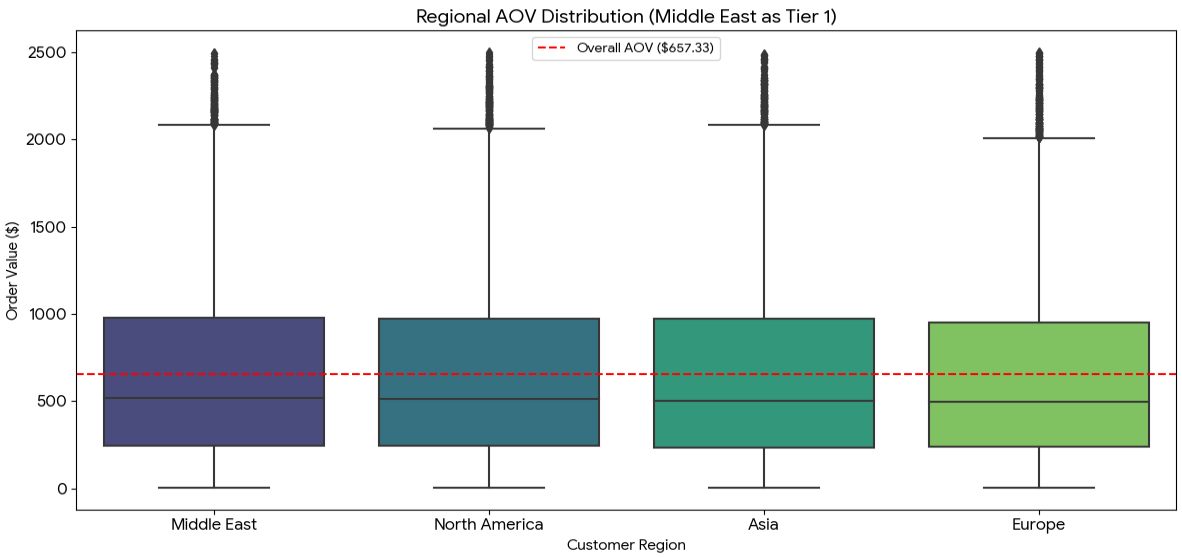

Figure 1 ***Regional AOV Distribution — Middle East as Tier 1.*** *Dashed red line = portfolio average AOV ($657.33). The Middle East (leftmost) exhibits the highest median order value and widest upper-quartile spread across all four regions.*

4.2 Statistical Interpretation

p = 0.16
Regional ANOVA

The regional F-test (F=1.72, p=0.16) does not achieve significance at α=0.05. The Middle East allocation is a directional signal, not a statistically mandated reallocation. Capital deployment should be calibrated accordingly — incremental and monitored, not disproportionate.

Despite the non-significant ANOVA, the convergence of three favourable conditions in the Middle East — highest AOV ($663.88), lowest discount exposure (13.23%), and a substantial revenue base ($8.3M) — constitutes a compelling directional bet. Increased marketing and operations budget is recommended as a risk-proportionate allocation, with statistical reassessment once organic live data is available.

SECTION 05

Identifying Power Segments: Region × Payment Method

To sharpen capital deployment below the regional level, we mapped the intersection of customer region and payment method to identify the most liquid and high-value revenue flows within the portfolio. A pivot table aggregating total revenue across all region–payment combinations reveals the highest-concentration nodes.

5.1 Model Specification

```
# Pivot table — total revenue by region × payment method
heatmap_data = df.pivot_table(
    index='customer_region',
    columns='payment_method',
    values='total_revenue',
    aggfunc='sum'
)

# Revenue heatmap
plt.figure(figsize=(12, 7))
sns.heatmap(heatmap_data, annot=True, fmt='.2f', cmap='YlGnBu')
plt.title('Revenue Heatmap: Region vs. Payment Method')
plt.show()
```

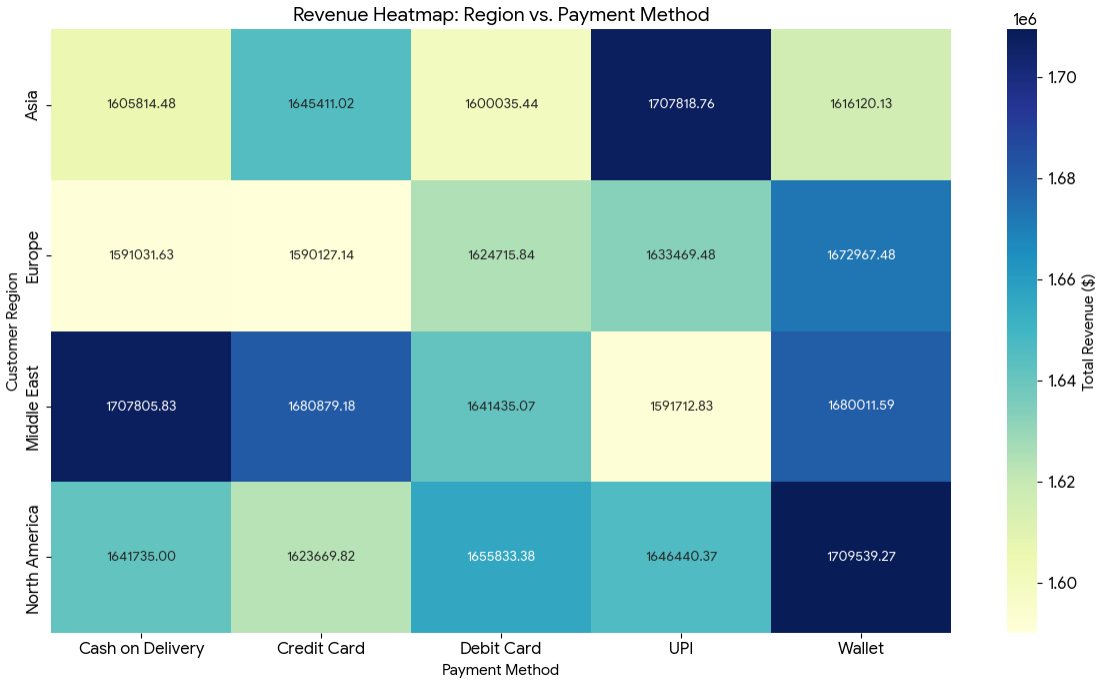

Figure 2 ***Revenue Heatmap — Region × Payment Method (USD).*** *Darker cells = higher revenue. North America/Wallet ($1.71M) and Middle East/Cash on Delivery ($1.71M) are the two dominant nodes. Middle East/UPI is the lowest segment at $1.59M.*

5.2 Key Observations

- **North America / Wallet:** $1.71M — highest single segment. Wallet is the dominant payment infrastructure in the portfolio's largest market.
- **Middle East / Cash on Delivery:** $1.71M — high-value regional behaviour persists even through the highest-friction payment method, underscoring the underlying strength of the Middle East customer base.
- **Europe / Wallet:** $1.67M — robust Wallet penetration across both Western markets supports a consistent Wallet-first push in those geographies.
- **Middle East / UPI:** $1.59M — the portfolio's lowest segment. UPI has not achieved meaningful adoption in that region. Do not extend the Wallet/UPI incentive there without further market research.

5.3 Payment Rail Strategy

We recommend a targeted 1% incentive to migrate credit and debit card orders (~19,889 orders) to Wallet or UPI rails. The modelled net saving after the conversion incentive is approximately $43,607, assuming external processing costs of ~$12.06 per credit card order and ~$11.77 per debit card order. These fee assumptions are externally derived and must be validated against the live processor contract before committing to the incentive.

The incentive must be applied exclusively to the credit/debit cohort. Wallet and UPI users already demonstrate identical or superior revenue performance — extending the incentive to them represents pure margin erosion with no behavioural uplift.

SECTION 06

Portfolio Stability & Health Metrics

Having identified the primary levers for capital reallocation, we stress-test the structural stability of the portfolio and validate its concentration profile to confirm that the proposed strategy can be executed without disrupting the core revenue base.

6.1 Structural Stability — Nested F-Test

```
# Nested F-test — does a Region × Category interaction improve fit?
m_base = ols(
    'total_revenue ~ discounted_price + C(product_category) + C(customer_region)',
    data=df
).fit()

m_int = ols(
    'total_revenue ~ discounted_price + C(product_category) * C(customer_region)',
    data=df
).fit()

print(f"Interaction p-value: {anova_lm(m_base, m_int).iloc[1, 5]:.4f}")
```

p = 0.320
Interaction Test

Region × Category interaction terms do not significantly improve model fit (ΔR^2=0.000063). A globally uniform pricing and allocation strategy is structurally valid across all regional and category nodes. No market requires a bespoke category treatment.

6.2 Market Concentration — HHI

```
# Herfindahl-Hirschman Index — portfolio concentration
rev_shares = (
    df.groupby('product_category')['total_revenue'].sum()
    / df['total_revenue'].sum() * 100
)
hhi = (rev_shares ** 2).sum()
print(f'Portfolio HHI: {hhi:.0f}')
```

HHI = 1,667
Concentration

An HHI between 1,500 and 2,500 defines a moderately concentrated market — optimal for scaling. The portfolio is diversified enough to absorb demand shocks in any single category, yet concentrated enough to generate operational efficiency across shared fulfilment and payment infrastructure.

6.3 Seasonality & Working Capital

- **Monthly CV:** 4.22% — low volatility confirms predictable cash conversion. A large liquidity buffer is not required.
- **Peak Month:** January at $1.46M. Build inventory and media spend to this ceiling; begin clearance at late-peak.
- **Trough Month:** February average $1.25M. Implement a hard freeze on discretionary purchasing through the trough.
- **Volatility Arbitrage:** July and October carry the highest cross-year revenue variance — upside not explained by the seasonal model. Pre-position media and inventory in Q2/Q3 as option value, not baseline commitment.

SECTION 07

Leverage Engineering: Constructing Asymmetry

The portfolio is structurally symmetric across categories, regions, and payment methods. However, asymmetric returns can be deliberately engineered by stacking the favourable conditions identified throughout this report. Three compounding levers are available:

Lever 1 — Price-Inelastic Cohort

16.4% of orders are already transacting at full price, generating an AOV of $749 versus $639 for discounted orders — a $110 spread per order. This cohort must be identified, tagged in the CRM, and permanently shielded from any discount prompt or promotional targeting. It is the portfolio's highest-margin segment and the foundation of the discount elimination strategy.

Lever 2 — Compound Cohort

The intersection of Middle East + Full Price + Wallet/UPI produces 839 orders at an AOV of $712.06 — an 8.3% premium over the portfolio mean. This is the highest-value identifiable segment in the dataset. A dedicated acquisition and retention programme — regional creative, Wallet payment prompting, zero discount exposure — represents the clearest path to above-average unit economics within the existing infrastructure.

Lever 3 — Volatility Arbitrage

July and October carry the highest cross-year revenue variance. This unexplained volatility is best treated as option value: pre-position incremental media budgets and inventory in Q2/Q3 to capture these windows without embedding them in the baseline forecast. If the upside materialises, it flows directly to margin; if not, the base plan is unaffected.

SECTION 08

Statistical Integrity & Model Limitations

All conclusions in this report are derived from OLS models that assume linearity, homoskedasticity, and the absence of omitted variable bias. The following caveats apply before operationalising any recommendation:

- **Heteroskedasticity confirmed:** Breusch-Pagan (LM=1,205.0, p<0.001) confirms non-constant residual variance. Robust HC3 standard errors are applied to all elasticity estimates and are recommended for any re-estimation on new data.
- **Engineered dataset signature:** Product HHI of 2.82 across 4,000 SKUs indicates near-perfect revenue dispersion, consistent with synthetic construction. Conclusions should be treated as directional on organic production data until live results are accumulated.
- **Payment cost assumptions:** Processing fee estimates (~$12.06/credit card, ~$11.77/debit card) are externally assumed and must be validated against the live processor fee schedule before deploying the Wallet/UPI incentive.
- **Causality vs. correlation:** This analysis is observational. The elasticity finding confirms no correlation between discount depth and volume but does not formally establish causation. The 30-day rollback protocol validates the causal assumption in a controlled deployment before full commitment.
- **Rating vs. revenue:** Pearson r=0.0018, p=0.6867 — product ratings are statistically unrelated to revenue. Eliminating paid review acquisition and replacing it with automated post-purchase flows is validated by this finding.

SECTION 09

Decision Summary

The table below consolidates all actionable decisions arising from this analysis. Direction indicators show the recommended capital allocation movement for each lever: ▲ increase ▼ decrease — hold.

Decision	Recommended Action	Evidence
▲ **Middle East Budget**	Increase marketing & operations spend	*Directional — ANOVA p=0.16*
▲ **Payment Rails**	Migrate cohort to Wallet / UPI	*Net gain ~$43,607 modelled*
▲ **Beauty / Fashion Ads**	Prioritise within category mix	*~45% proxy margin*
▲ **Q4 Inventory**	Build to January peak ceiling ($1.46M)	*CV = 4.22%*
▲ **Compound Cohort**	Build ME + Full-Price + Wallet programme	*+8.3% AOV premium*
— **Category Mix**	Hold all six categories	*HHI = 1,667*
▼ **Discount Programme**	Eliminate all tiers — 30-day rollback	*+$4.59M upside; R^2=0.000002*
▼ **Review-Gen Spend**	Replace with automated post-purchase email	*73.6% rev from <4.0★*

Statistical Appendix *Qty OLS: slope=0.000201, R^2=0.000002, p=0.7544. Rev OLS: slope=−7.7134, R^2=0.0208, p=0.0000. Nested F-test: F(3,~50k)=1.0731, p=0.3591, ΔR^2=0.000063. Regional ANOVA: F=1.7220, p=0.1601. Breusch-Pagan: LM=1,205.0, p<0.001. Pearson r (rating~revenue)=0.0018, p=0.6867. Revenue monotonicity at discount tiers: confirmed. Payment cost model: externally assumed — validate before deployment.*