Exploratory Data Analysis

E-Commerce Transaction Dataset

Analysis Summary & Key Insights

Dataset: transactions.csv
5,050 Transactions | 8 Features | 2023-2024

Executive Summary

This exploratory data analysis examines an e-commerce transaction dataset containing 5,050 records across 8 variables. The analysis reveals key insights about customer behavior, sales patterns, and data quality issues that require attention before further modeling.

Key Metrics at a Glance

Metric	Value	Note
Total Transactions	**5,050**	2-year period
Unique Customers	**~1,850**	Avg 2.73 orders/customer
Average Price	**₺2,524**	Median: ₺714
Data Completeness	**99.6%**	Minimal missing values

1. Dataset Overview

Data Structure

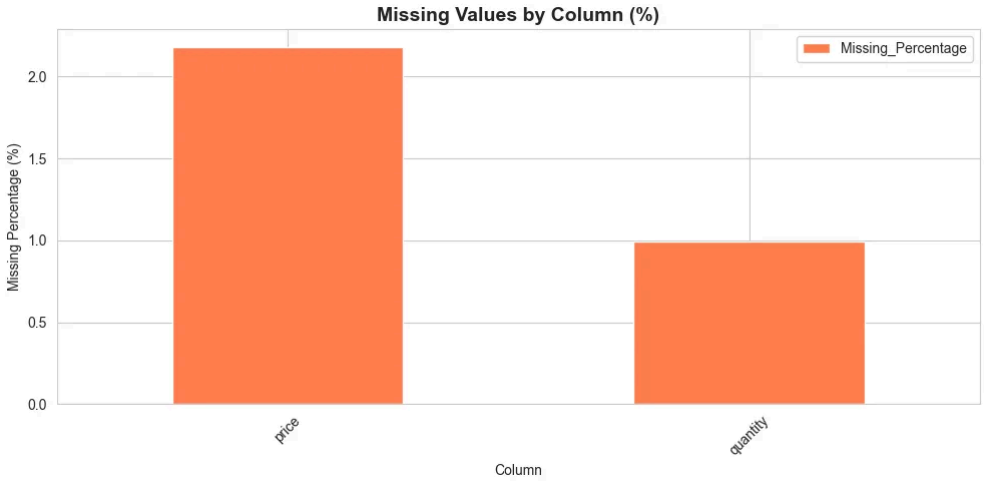
Column	Data Type	Description
customer_id	int64	Unique customer identifier
order_id	int64	Unique order identifier
order_date	datetime64	Transaction date (2023-01-01 to 2024-12-31)
product_category	object	7 categories: Fashion, Electronics, Grocery, etc.
channel	object	Sales channel: mobile, desktop, web
city	object	Customer city location (9 unique values)
price	float64	Transaction price (₺)
quantity	float64	Items per transaction (0-5)

2. Data Quality Assessment

Missing Values

The dataset shows minimal missing data with overall 99.6% completeness:

- **Price:** 2.18% missing (~110 records)
- **Quantity:** 0.99% missing (~50 records)



Data Quality Issues Identified

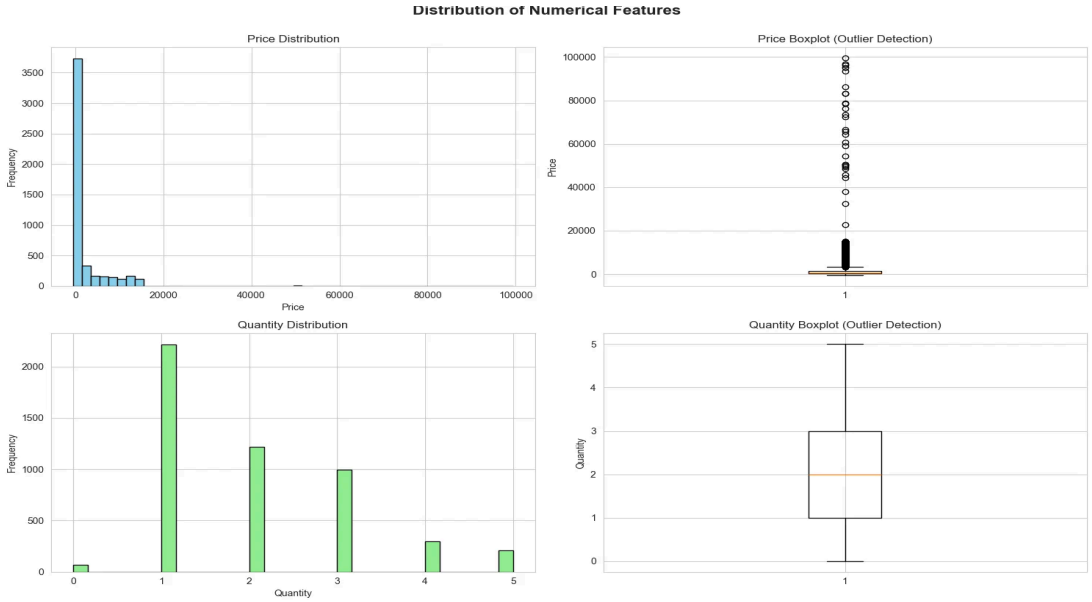
Issue	Count	Severity	Recommended Action
Negative Prices	26	HIGH	Investigate - may be refunds
Duplicate Order IDs	50	MEDIUM	Deduplicate or investigate
Zero Quantities	68	MEDIUM	Remove or flag as invalid
City Name Typos	164	LOW	Standardize encoding

Note: City encoding issues include: Istanbul vs İstanbul, Izmir vs İzmir, 'antlaya' typo

3. Distribution Analysis

Numerical Features

Both Price and Revenue show **highly right-skewed distributions** with significant outliers. Log transformation is recommended for modeling.



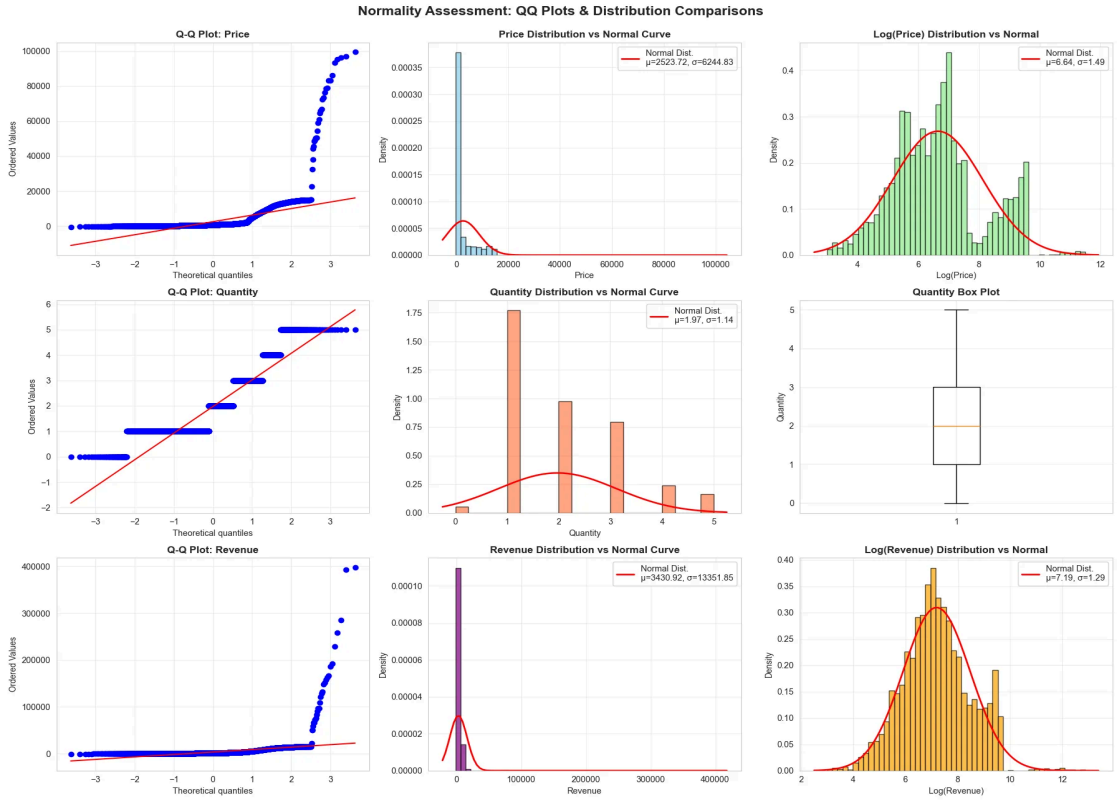
Key Statistics

Variable	Mean	Median	Std Dev
Price	₺2,524	₺714	₺6,245
Quantity	1.97	2.00	1.14

⚠️ **Important:** Mean >> Median for Price indicates significant right skew. Use Median for central tendency reporting.

Normality Assessment

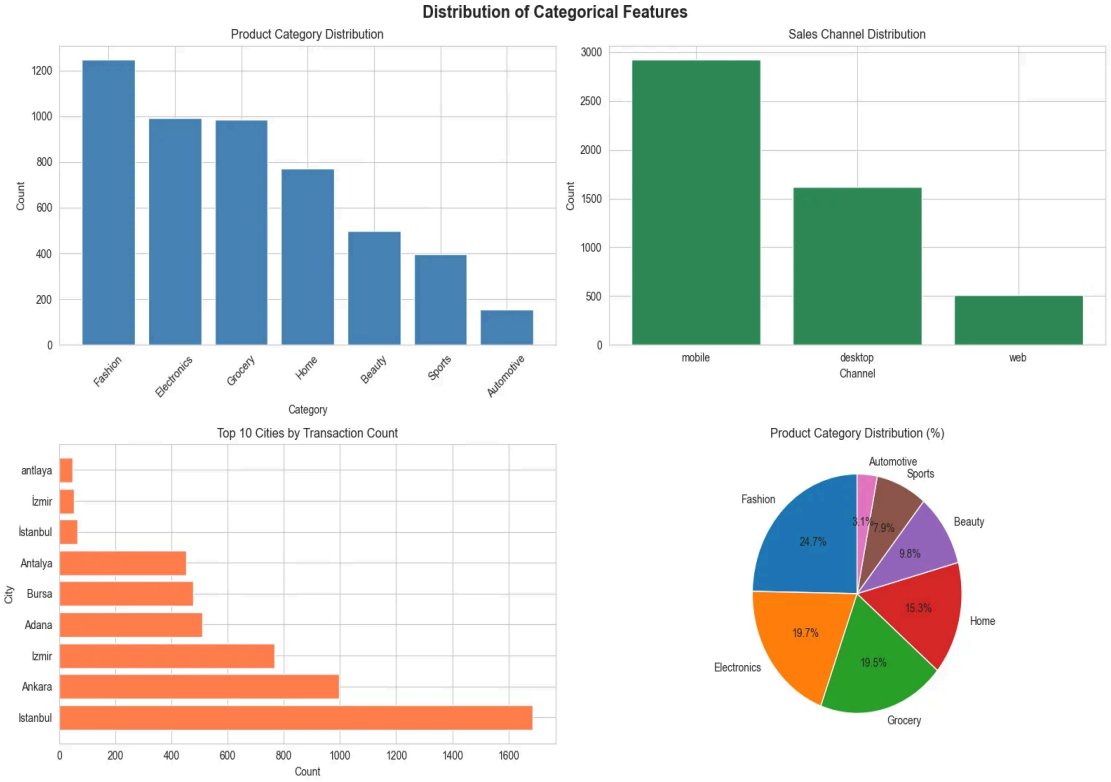
Q-Q plots and distribution comparisons reveal that Price and Revenue deviate significantly from normal distribution. Log transformation produces more bell-shaped distributions suitable for parametric modeling.



Modeling Implications

- **Linear Models:** Use log-transformed Price and Revenue
- **Tree-based Models:** Can use raw values (naturally resistant to skew)
- **Count Data (Quantity):** Consider Poisson regression for prediction

4. Categorical Features Analysis



Key Findings

Product Categories

- **Fashion** leads with 24.7% of transactions (1,247)
- **Electronics** and **Grocery** follow closely at ~19.7% each
- **Automotive** is smallest segment (3.1%)

Sales Channels

- **Mobile dominates** with 57.9% of transactions - Mobile-first customer base!
- Desktop: 32.0% | Web: 10.1%

Geographic Distribution

- **Istanbul** leads with 33.4% of transactions
- **Top 3 cities** (Istanbul, Ankara, Izmir) account for ~68% of business

5. Revenue Analysis

Volume vs Value Strategy

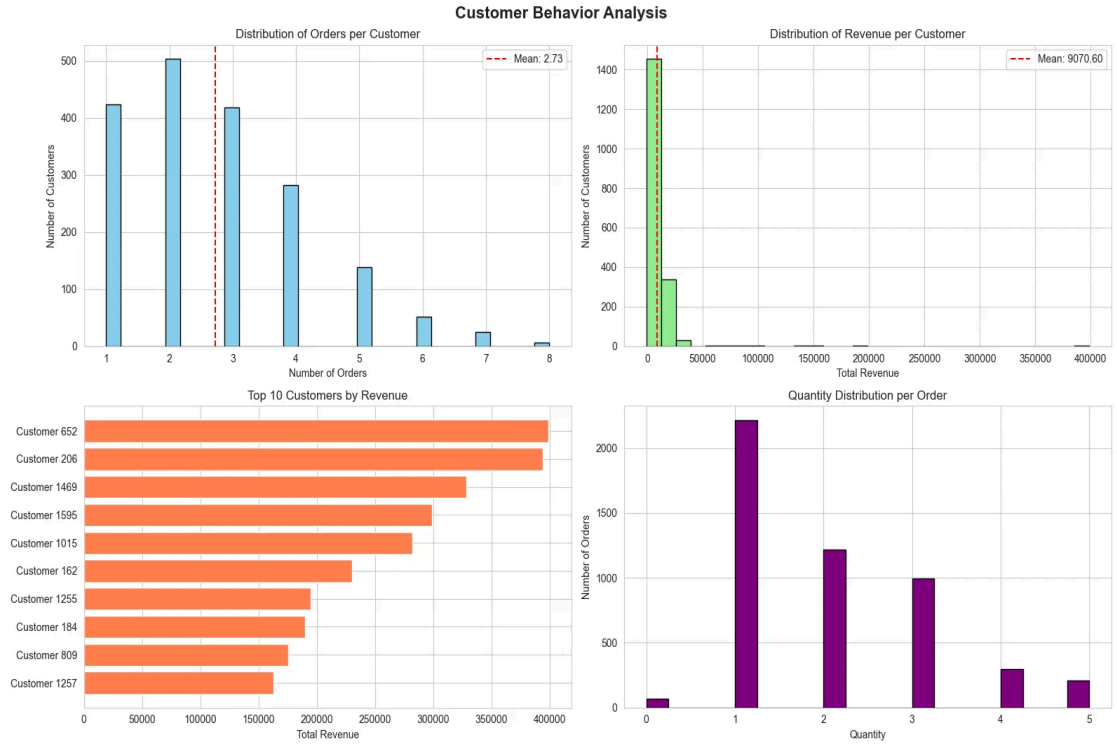
A critical insight emerges: **Fashion has most transactions but Electronics generates highest revenue.** This reveals two distinct business strategies at play:

High-Volume / Low-Price	Low-Volume / High-Price
Fashion, Grocery, Beauty **Strategy:** Volume-driven growth **Avg Price:** ₺200-900	Electronics, Home **Strategy:** Value-driven premium **Avg Price:** ₺2,000-8,500

Channel Performance

- **Desktop users** purchase higher-value items (avg ₺3,900 vs ₺1,800 mobile)
- **Mobile generates most revenue** due to volume (₺8M vs ₺7M desktop)

6. Customer Behavior Analysis



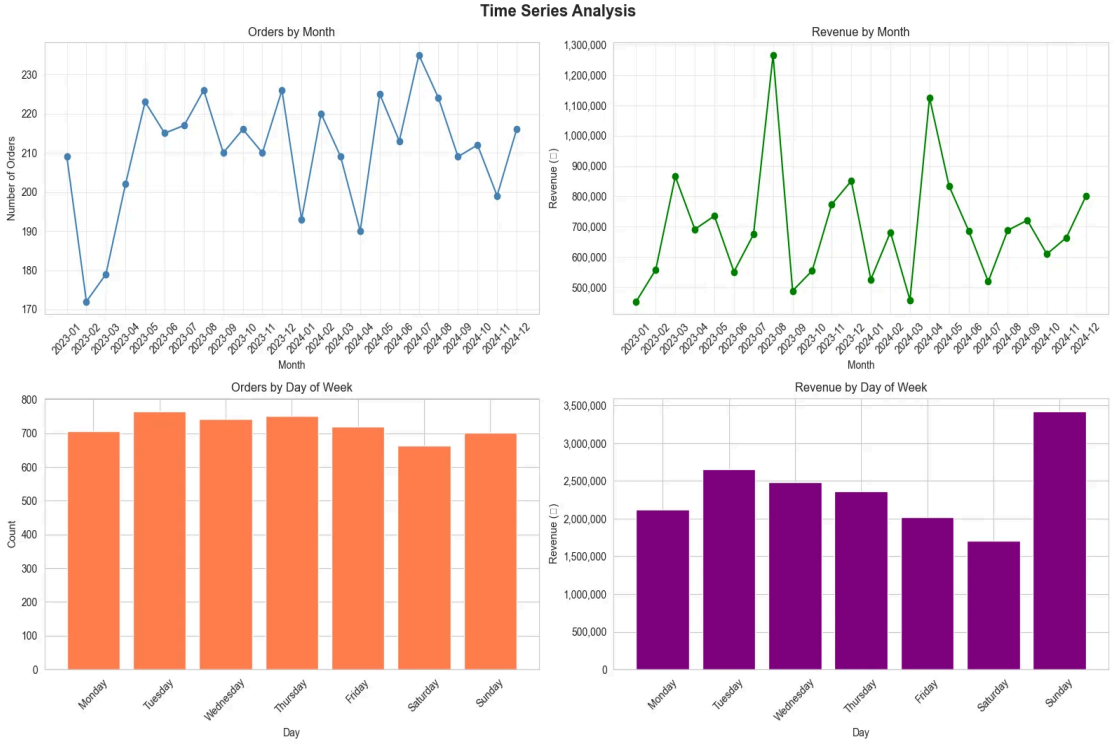
Customer Metrics

- **Average Orders per Customer:** 2.73
- **Average Revenue per Customer:** ₺9,071
- **Most Common Quantity:** 1 item per order (45% of orders)

Top Customers

The top 10 customers generate disproportionately high revenue, with Customer 652 leading at ~₺400,000. This represents a potential **concentration risk** and opportunity for VIP customer programs.

7. Time Series Analysis



Monthly Trends

- Orders fluctuate between 170-235 per month with no clear seasonal pattern
- **Peak Revenue:** August 2023 (~₺1.3M) and March 2024 (~₺1.1M)
- Revenue volatility higher than order count volatility (price mix effect)

Day of Week Patterns

- **Tuesday-Thursday** shows highest order count (~750/day)
- **Sunday** generates highest revenue despite moderate order count (higher AOV)
- **Saturday** shows lowest activity (opportunity for promotions)

8. Key Findings & Recommendations

Major Findings

1. **Mobile-First Business:** 58% of transactions from mobile - prioritize mobile experience optimization
2. **Volume vs Value Split:** Fashion drives volume; Electronics drives revenue - different strategies needed
3. **Geographic Concentration:** Top 3 cities = 68% of business - logistics hub opportunity
4. **Customer Concentration:** Top customers generate disproportionate revenue - VIP program recommended
5. **Data Skewness:** Use Median over Mean for reporting; Log transform for modeling

Data Cleaning Recommendations

1. Impute missing prices using category-specific median values
2. Investigate 26 negative prices (possible refunds) before removal
3. Standardize city names: Istanbul/İstanbul → Istanbul, fix 'antlaya' → Antalya
4. Remove/flag 68 zero-quantity records as invalid transactions
5. Deduplicate 50 duplicate order_ids after investigation

Next Steps

1. **Data Cleaning:** Execute cleaning pipeline before further analysis
2. **Feature Engineering:** Create customer RFM segments, time-based features
3. **Predictive Modeling:** Build customer churn and CLV prediction models
4. **A/B Testing:** Test Saturday promotions to address low weekend activity

— End of Report —