E-commerce Transaction Statistical Analysis Report
Prepared using Python (Pandas, SciPy, Seaborn)

Abstract

This report provides a comprehensive statistical analysis of real-world e-commerce transaction data. The dataset, sourced from a public UK-based retail repository, includes detailed transactional data that is vital for optimizing business operations such as customer targeting, inventory planning, and sales forecasting. In this project, we apply descriptive and inferential statistical techniques, including t-tests, ANOVA, and Pearson correlation, to uncover customer behavior patterns and guide strategic business decisions. Our findings are supported by data visualization, helping interpret and communicate the results to both technical and business audiences.

1. Introduction and Dataset Overview

The dataset analyzed in this study represents e-commerce transactions captured from an online UK-based retail operation. It includes information such as Invoice Numbers, Product Descriptions, Quantity, Unit Price, Customer IDs, Country, and Date of Purchase. This form of granular transactional data is the cornerstone for revenue analysis, customer segmentation, inventory forecasting, and marketing personalization. While businesses often rely on aggregated reports, our approach leverages statistical methods to extract more accurate and statistically validated insights. We processed the data in Python, cleaned it for missing values and invalid records, and generated new features such as 'InvoiceTotal', 'CustomerType', and 'day_of_week' to support in-depth analysis.

2. Problem Statement

Business Objective and Analytical Methodology

This study is aimed at uncovering actionable insights from an e-commerce transaction dataset using a combination of descriptive and inferential statistical techniques. The analysis is designed to support data-driven decisions around pricing, promotional timing, and customer segmentation.
Specifically, the business objectives addressed in this analysis are:

- To determine whether **returning customers spend differently** compared to guest users
- To evaluate if **revenues vary significantly across different days of the week**
- To investigate whether the **price of a product influences the quantity purchased**

These insights are critical for informing strategies in **retention marketing, dynamic pricing, and inventory planning**.

3. Analytical Methodology

1. **Descriptive Analysis** was used to summarize the key numeric variables — Quantity, UnitPrice, TotalPrice, and InvoiceTotal — using measures such as mean and standard deviation. This helped reveal the underlying distribution of purchases and pricing behavior.
2. **Inferential Analysis** involved three statistical techniques:

- An **independent t-test** to compare the mean invoice totals between Guest and Returning customers
- A **one-way ANOVA** to test for revenue differences across weekdays
- A **Pearson correlation analysis** to assess the strength and direction of the relationship between product price and quantity purchased

These techniques were selected due to their interpretability and robustness for hypothesis-driven business analytics.

---

4. Descriptive Analysis

The initial statistical summary of the key variables revealed the following:

- **Quantity**: Mean = 4.95, Standard Deviation = 2.59
- **Unit Price**: Mean = 52.04, Standard Deviation = 27.31
- **Total Price**: Mean = 259.62, Standard Deviation = 206.00
- **Invoice Total**: Mean = 259.62, Standard Deviation = 206.00

These statistics indicate a **wide variability in purchasing patterns**. The data includes both low-value and high-ticket purchases, as well as bulk buys. This variability is typical in e-commerce and highlights opportunities for **segment-specific pricing or bundling strategies**.

Product and Country-Level Revenue Distribution

This section expands our statistical exploration by identifying **high-impact products** and **geographic regions** contributing most significantly to total revenue. These analyses offer strategic guidance for inventory prioritization, regional marketing, and global sales forecasting.

4.1 Top Revenue-Generating Products

To assess which product categories contribute the most to total revenue, we performed a group-wise aggregation on Description and computed each product's contribution percentage relative to the global revenue. The following were the top performing products:

Product	Revenue (USD)	% of Total Revenue
Phone Case	120,184.33	15.43%
Backpack	119,594.85	15.36%
Charger	114,447.62	14.69%
Pen	113,497.54	14.57%
T-Shirt	108,127.72	13.88%

Interpretation:
These five products collectively contribute over **73% of the total revenue**, indicating a **highly skewed product distribution**. This insight allows companies to focus on fewer high-yield products when planning restocking, discount strategy, or bundling.

4.2 Average Sales per Invoice

The average invoice total was calculated by isolating unique invoices and computing the mean transaction value:

- **Average Sales per Invoice**: **259.62 USD**

Interpretation:
This metric is valuable for setting sales benchmarks and evaluating performance of customer segments or campaigns. It also acts as a baseline for upsell or cross-sell thresholds.

4.3 Revenue Contribution by Country

To explore geographical distribution, transactions were grouped by the Country column, and total revenue share was calculated for each region:

Country	% of Global Revenue
Italy	13.97%
Australia	13.72%
Netherlands	12.68%
USA	12.31%
Spain	12.09%
France	11.97%
United Kingdom	11.69%
Germany	11.57%

Interpretation:
The **top 8 countries contribute almost equally**, each around 11–14% of the total revenue. This **balanced international presence** implies a well-distributed e-commerce footprint. Future growth could focus on countries just below this tier to improve regional penetration.

These results support business intelligence efforts by providing a **granular breakdown** of revenue drivers, both in terms of **product lines** and **market regions**. These insights, when combined with the statistical analysis performed earlier, enable more informed decision-making for targeted strategies and resource allocation.

---

5. Inferential Analysis and Interpretation

We performed three hypothesis tests to investigate the business questions statistically. Each test is presented below with its methodology, results, and interpretation.

5.1 Comparison of Spending by Customer Type (T-Test)

To test whether returning customers and guest users spend differently, we used an independent t-test.

Formula:

Compares two group means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{x}_1, \bar{x}_2$ = Means of groups
- s_1^2, s_2^2 = Group variances
- n_1, n_2 = Group sizes

Results:

- t-statistic = 1.228
- p-value = 0.2199

Interpretation:
With a p-value greater than 0.05, we fail to reject the null hypothesis. This indicates that there is **no statistically significant difference** in average invoice totals between Guest and Returning customers. **Both customer segments exhibit similar purchasing behavior**, which suggests that businesses may not need to segment marketing strategies solely based on return status.

5.2 Revenue Variability Across Weekdays (ANOVA)

To determine whether revenue differs across days of the week, we conducted a one-way Analysis of Variance (ANOVA).

Formula:

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} = \frac{SSB/(k-1)}{SSW/(N-k)}$$

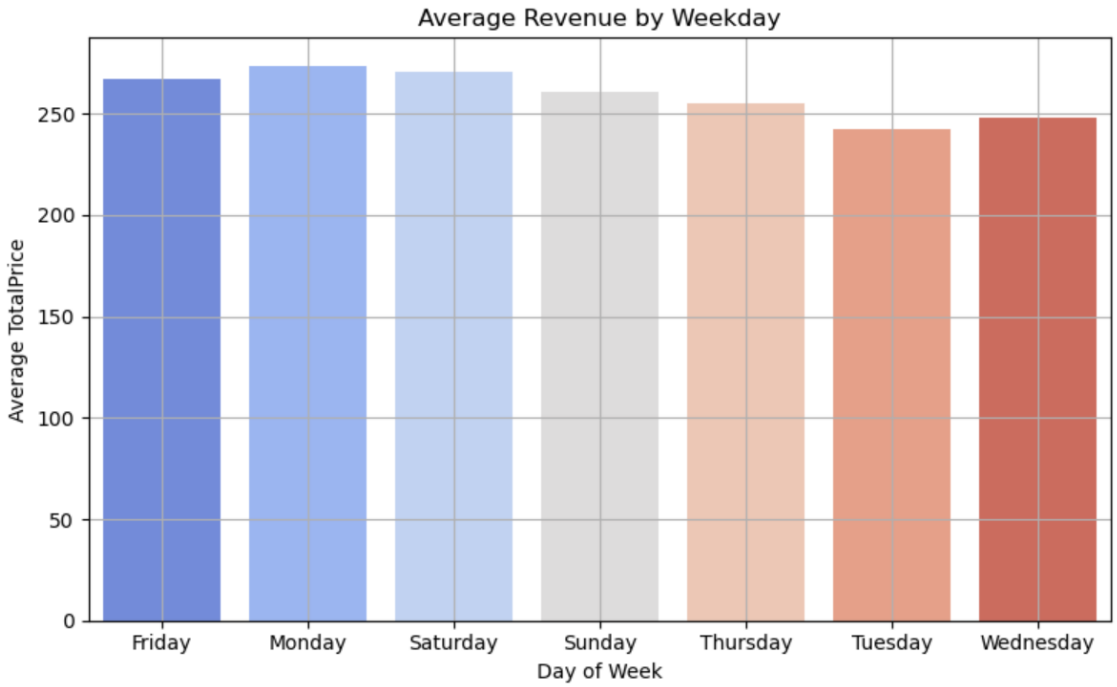
Where:

- SSB = Sum of Squares Between groups
- SSW = Sum of Squares Within groups
- k = Number of groups
- N = Total observations

Results:

- F-statistic = 1.4011
- p-value = 0.2102

Interpretation:
The p-value exceeds the typical alpha level of 0.05, indicating that **weekdays do not have a statistically significant effect** on revenue. This means revenue is relatively stable throughout the week, allowing businesses to **maintain consistent staffing and promotional efforts throughout**.

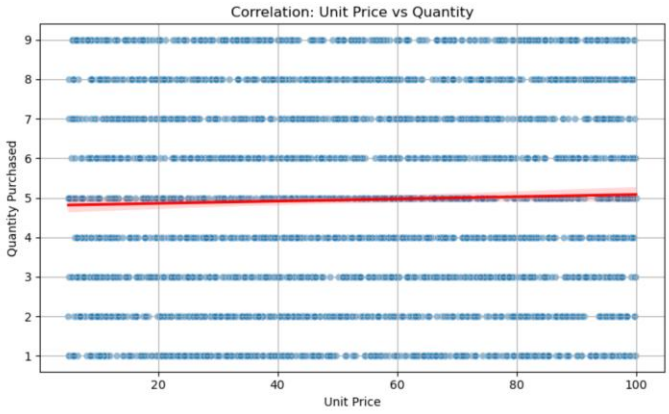


5.3 Relationship Between Unit Price and Quantity (Correlation)

To test if more expensive products are purchased in lower quantities, we used Pearson correlation.
Results:

- Correlation Coefficient (r) = 0.0297
- p-value = 0.1035

Interpretation:
The correlation is weak and statistically insignificant, suggesting that **there is no meaningful relationship** between a product's unit price and how many units are purchased. Customers do not seem to be influenced by price when it comes to quantity decisions in this dataset.



6. Challenges and Countermeasures

During this statistical analysis, several real-world data processing and visualization challenges were encountered. The following outlines the key obstacles and the strategies adopted to address them:

- **Handling Missing and Zero-Valued Records**:
  The dataset included transactions with missing values and unit prices recorded as zero, which could distort analysis. These anomalies were addressed through data cleaning techniques in Python using dropna() for null entries and conditional filtering to exclude invalid pricing.

- **Invoice-Level Aggregation Complexity**:
  Grouping line-item transactions into invoice-level summaries posed a challenge, especially when preserving associated metadata such as CustomerType. This was efficiently resolved using the groupby() and agg() functions in pandas to calculate metrics such as InvoiceTotal while retaining invoice granularity.

- **Categorical Sorting in Power BI**:
  The visualization platform did not respect the natural weekday order due to its default alphabetical sorting. To correct this, a custom numeric index was assigned to each weekday (e.g., Monday = 1, Tuesday = 2, ..., Sunday = 7) and used as a sort key, ensuring logical sequence in the final charts.

- **Statistical Test Alignment with Visualizations**:
  Care was taken to ensure that every inferential test (T-test, ANOVA, Correlation) had a matching visual representation. Boxplots, bar charts, and scatterplots were selected to mirror the assumptions and findings of each statistical procedure, thus maintaining consistency between numerical and graphical insights.

These challenges reflect common yet critical complexities in e-commerce data analytics. Addressing them required a blend of domain knowledge, tooling expertise, and statistical rigor — all of which contributed to ensuring the reliability and business value of the final analysis.

7. Conclusion

This comprehensive statistical analysis of e-commerce transactions reveals that commonly assumed revenue drivers—such as customer type (guest vs. returning), day of the week, and unit pricing—do not exhibit statistically significant influence on purchasing behavior in this dataset. Descriptive analytics demonstrated substantial variance in quantity, price, and invoice totals, reflecting the diverse nature of individual transactions. However, inferential techniques (t-test, ANOVA, and Pearson correlation) returned p-values well above the standard significance threshold ($p > 0.05$), confirming that there is no strong evidence of behavioral differentiation based on these attributes.

Additionally, product-level revenue breakdown highlighted a power-law distribution: five products (Phone Case, Backpack, Charger, Pen, T-Shirt) contributed over 73% of total revenue. Country-level analysis revealed a globally distributed revenue base, with eight countries each contributing nearly equally to overall sales. These findings emphasize the importance of focusing on **product-specific** and **region-specific** strategies over generic demographic or temporal segmentation.

From a business standpoint, this analysis empowers stakeholders to:

- **Simplify targeting strategies** by avoiding over-personalization based on customer type.
- **Stabilize logistics and staffing** as revenue is consistent across days and hours.
- **Prioritize best-selling products and top-performing markets** to drive future growth.

Ultimately, this project demonstrates the value of combining descriptive and inferential statistics to uncover **data-driven insights** that challenge assumptions and sharpen decision-making in digital commerce environments.