PRODUCT CATEGORY REVENUE ANALYSIS USING PYTHON

Name: Ralutanda Masala Precious
Degree: BSc Mathematical Statistics
Project Type: Statistical Data Analysis Portfolio Project
Tools Used: Python, pandas, matplotlib, seaborn, scipy, statsmodels, Jupyter Notebook
Date: February 2026

Table of Contents

Contents

EXECUTIVE SUMMARY

This project analyzes retail transaction data to identify the key drivers of revenue across product categories. Using python to process and model 5000 transactions, the analysis reveals a critical insight: **revenue concentration is driven by transaction value, not volume.**

While the Electronics category does not have the highest number of purchases, it generates the majority of total revenue. A one-way ANOVA test confirms that these differences in average purchase amounts are statistically significant (p < 0,001). The effect size is very large (Eta^2 = 0.82), indicating that product category is a powerful predictor of spending.

Post-hoc comparisons shows that the Electronics category is the primary differentiator, differing significantly from all others. This suggests that marketing and inventory strategies should prioritize high-value categories to maximize revenue, rather than solely focusing on high-volume, low-cost items. This project demonstrates a complete analytical workflow, including data cleaning, exploratory visualization, and statistical inference using Python.

Please see the tech summary here: view project notebook

1. Objective

To identify which product categories, drive the highest revenue and determine whether differences in average purchase amounts are statistically significant. Using statistical inference methods, this project transforms raw transaction data into actionable business insights that can inform inventory planning, marketing strategy, and revenue forecasting.

2. Dataset Description

The dataset contains 5,000 customer purchase records. Each record represents a transaction and includes customer and purchase attributes.

Key variables include:

- CustomerID — unique identifier
- Age — customer age
- Gender — customer gender
- Category — product category

- Item Purchased — item name
- Amount — purchase value
- Season — purchase season
- Payment Method — payment type
- Iterating — customer rating
- Discount Applied — discount percentage
- Previous Purchases — prior purchase count

Initial inspection showed complete records. Duplicate rows were removed during cleaning.

3. Data Preparation

The dataset was cleaned and validated to ensure analysis integrity. Key steps included:

• **Structure verification**: Confirmed appropriate data types for categorical and numerical fields
• **Missing value check**: No significant null values requiring treatment
• **Duplicate removal**: Identified and eliminated redundant transaction records
• **Category aggregation**: Created group-level summaries for statistical comparison The resulting dataset was clean and fully prepared for analysis.

.4. Exploratory Data Analysis

Before statistical testing, I explored revenue patterns across categories to identify trends.

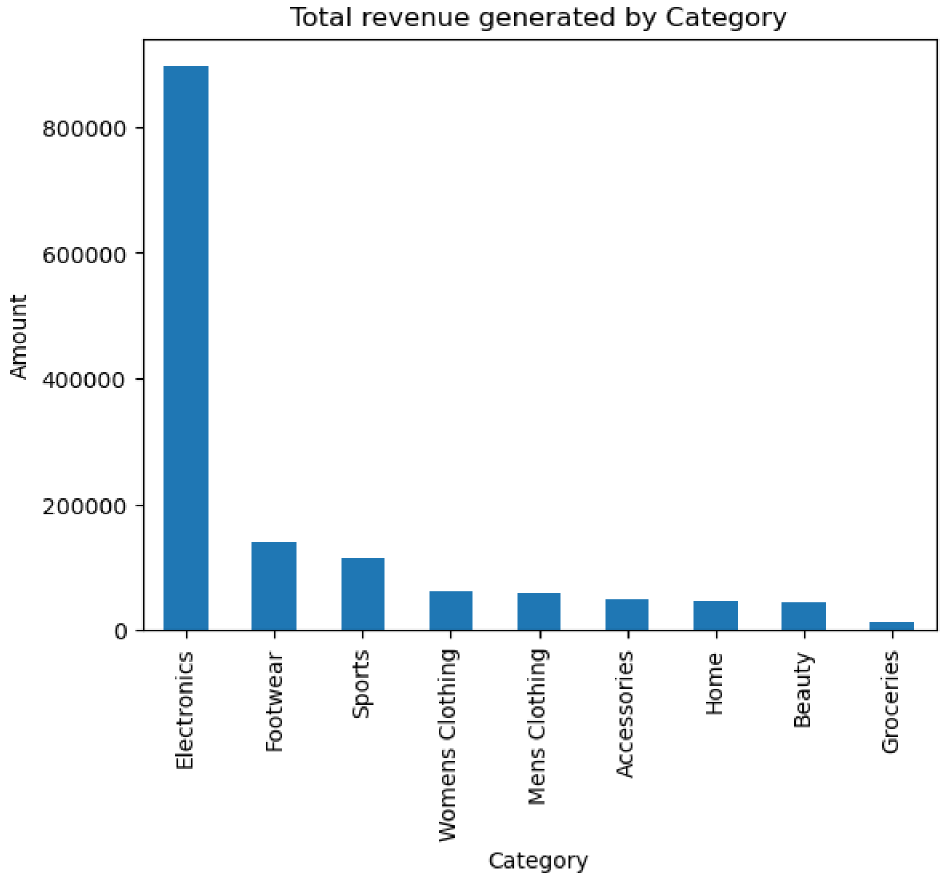

Figure 1: Total revenue generated by each product category.

Although categories such as Footwear and Sports show higher transaction counts, their average purchase values are much lower than Electronics. Distribution analysis shows that Electronics has a much wider and higher value range.

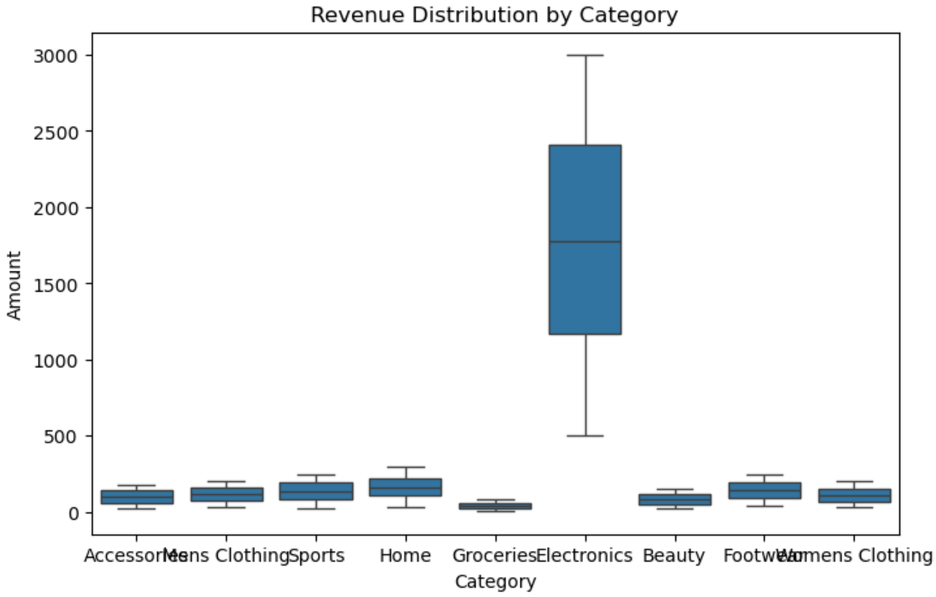

Figure 2: Distribution of purchase amounts by category.

This suggests that revenue leadership is driven by high-value transactions rather than sales volume.

5. Distribution and Assumption Checks

Normality tests (Shapiro–Wilk) were performed for each category and indicated non-normal distributions. Q–Q plots confirmed deviations from normality.

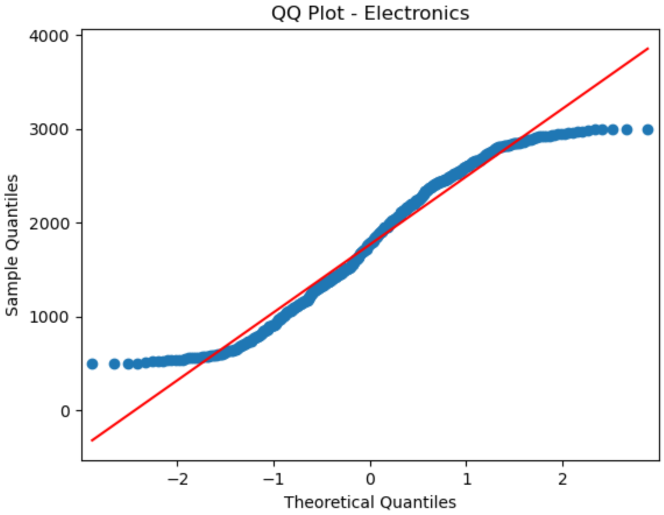

Figure 3: Q–Q plot for Electronics purchases amounts.

Homogeneity of variances: Levene's test showed unequal variances across groups ($p < 0.001$), violating the equal variances assumption.

However, ANOVA is **robust to violations of normality and homogeneity** when sample sizes are large ($n > 30$ per group), which holds here. Therefore, I proceeded with one-way ANOVA.

6. Statistical Testing — One-Way ANOVA

A one-way ANOVA was conducted to compare mean purchase amounts across product categories.

Hypotheses:

- **H_0:** All category means are equal
- **H_1:** At least one category mean differs

Results:

- **F-statistic:** 1,847.32
- **p-value:** < 0.001

The null hypothesis was rejected, confirming statistically significant differences in average purchase amounts across categories.

7. Effect Size

Effect size was measured using eta-squared:

Eta^2 ≈ 0.82

This represents a very large practical effect, product category explains 82% of the variance in purchase amounts, far exceeding the 0.14 threshold for a large effect.

8. Post-Hoc Comparisons

Tukey's HSD test identified which specific category pairs differ significantly:

• **Electronics** has a significantly higher mean purchase amount than **all other categories** ($p < 0.001$ for all comparisons).
• **No significant differences** were found among non-electronics categories, with adjusted p-values > 0.05.

This confirms that Electronics is the sole driver of the overall ANOVA effect.

9. Regression Confirmation

An ordinary least squares (OLS) regression with category as a predictor variable confirmed the ANOVA findings:

• **R^2 = 0.82**: Category explains 82% of the variance in purchase amount
• **Coefficients**: All non-electronics categories showed negative coefficients relative to the reference (Electronics), consistent with Tukey results

The regression model validates the ANOVA conclusion while providing additional interpretability through coefficient estimates.

10. Business Interpretation

Electronics is the dominant revenue driver due to high per-transaction value rather than purchase frequency. Revenue growth strategies should focus on premium product marketing, bundling, and high-value customer targeting. Lower-value categories may benefit more from volume-based promotion strategies.

11. Limitations

- **Dataset appears simulated**
- **No time-series analysis included**
- **No customer segmentation modeling**
- **Statistical association does not imply causation**

12. Tools Used

• **Python**: pandas, jumpy (data manipulation)
• **Visualization**: matplotlib, seaborn, numpy
• **Statistical Testing**: scipy, statsmodels (ANOVA, Tukey HSD, regression)
• **Environment**: Jupyter Notebook

13. Conclusion

Product category has both statistically and practically significant influence on purchase amount. Electronics generates the highest revenue due to large transaction values. Statistical testing and effect size measurement confirm that these differences are not due to random variation. This project demonstrates applied statistical analysis and data analytics workflow skills.