Online Retail Data Python EDA

Michele McCluskey

January 2026

Introduction

This mini project is based on a LinkedIn challenge created by Karina Samsonova, a Data Science & Data Analytics Consultant/Freelancer. The goal of the analysis is to identify, for UK customers, which products are best sellers and when customers are most likely to make purchases.

The Data

The dataset used is the "Online Retail Data Set" by Vijaykumar Ummadisetty, hosted on Kaggle. It contains over 500,000 transaction records and 8 columns: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. After filtering to UK only transactions, the working dataset was reduced to just over 300,000 rows. The data covers online retail transactions from December 2010 through December 2011.

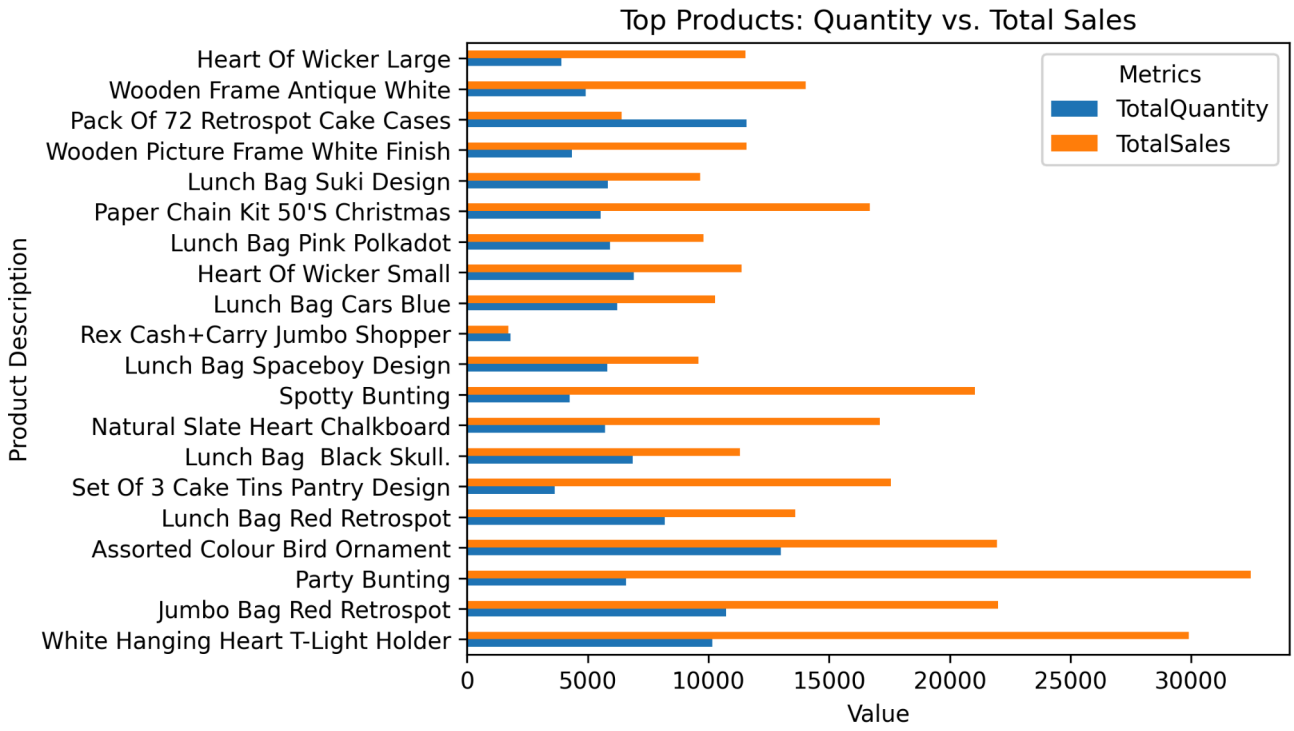
Methodology

After an initial review, I performed basic data cleaning, including trimming whitespace and standardizing column names. The InvoiceDate column was converted from an object type to a datetime type to support temporal analysis. Missing values were concentrated in the Description and CustomerID fields; because these are essential for product and customer level analysis, rows with nulls in either column were removed. Canceled orders, indicated by InvoiceNo values starting with "C", were excluded from the dataset, and 5,192 duplicate rows were identified and dropped. Outliers were also removed to prevent extreme values from skewing results.

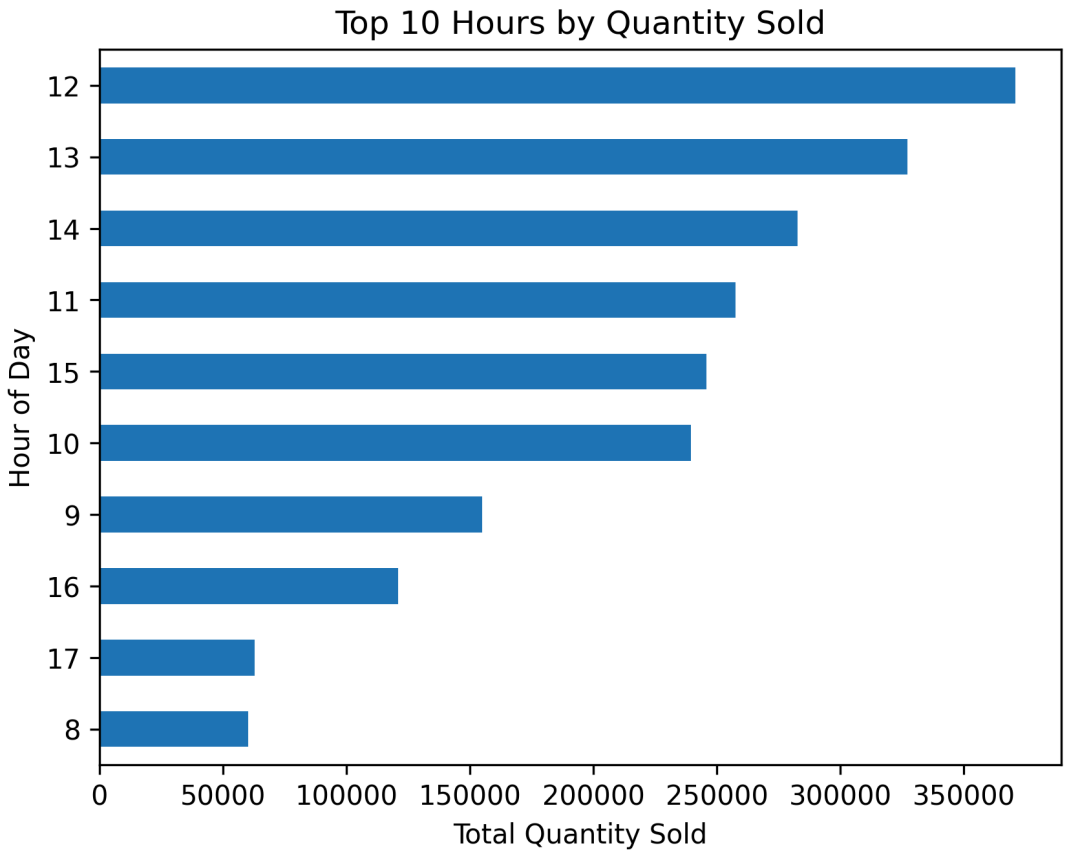
To support time based and revenue based analysis, I generatedthree new columns: hour of purchase, day of week, and total sales per transaction line (Quantity × UnitPrice). Using **groupby()**, I identified the top 10 products by quantity sold as well as the top 10 most profitable products by total sales. I then explored temporal trends by aggregating sales by hour and by day of week. As an additional exercise, I analyzed customer behavior, including top customers by revenue and order count, repeat purchase behavior, and items frequently bought together.

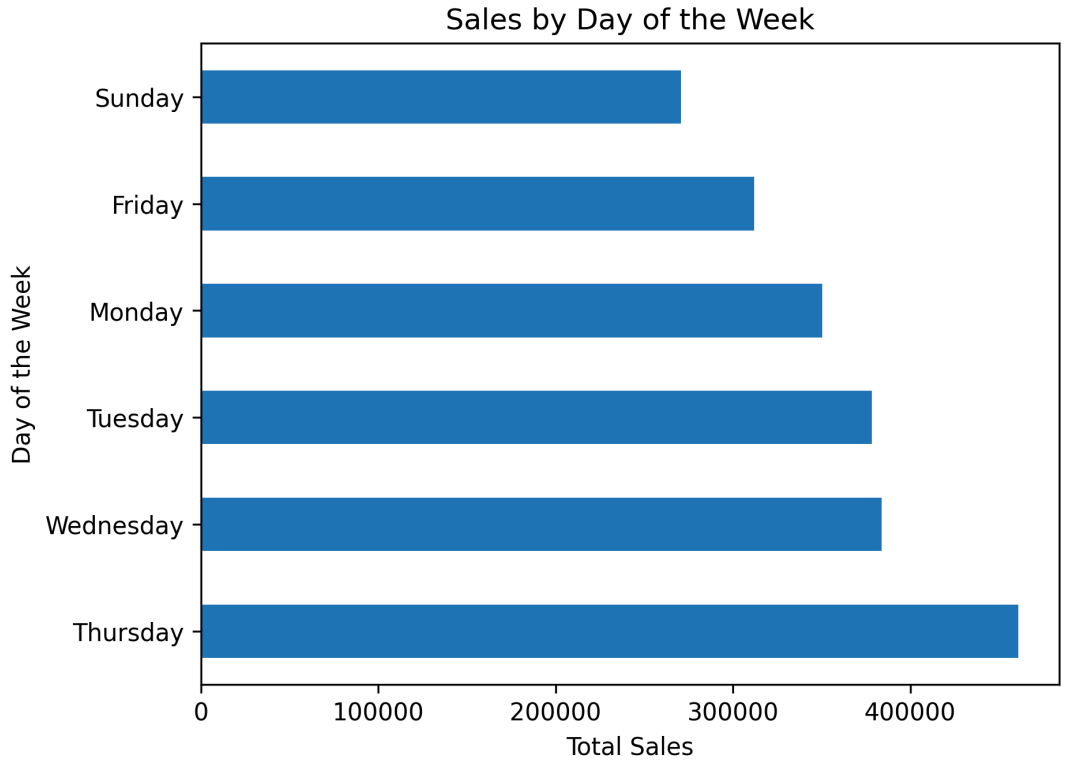
Analysis and Findings

Overall, the sales mix showed a healthy balance between high margin products and products sold in high volume.



Peak purchasing hours fell between 11:00 a.m. and 3:00 p.m., indicating that late morning to midafternoon is the most active sales window. Thursdays emerged as the most profitable day of the week, outperforming other days by approximately 75,000–190,000 in total sales. One notable anomaly in the dataset was the complete absence of Saturday purchases, which is unlikely in real world retail behavior and points to possible gaps or issues in data collection.

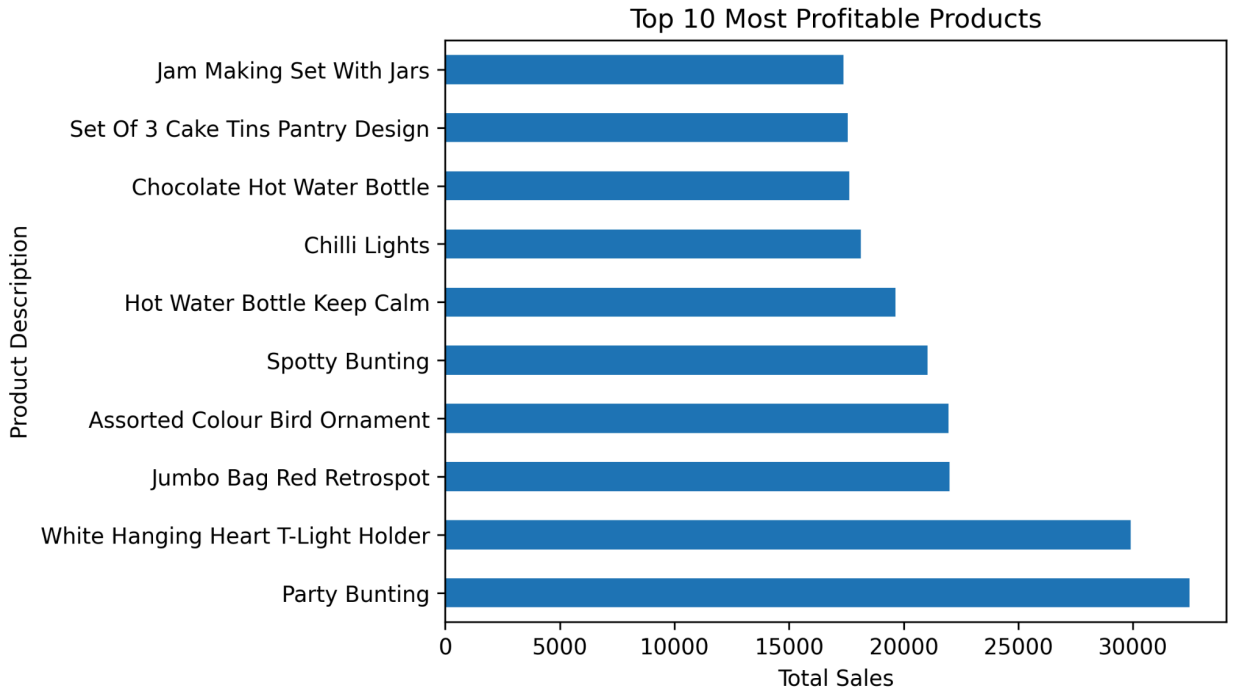

Top 10 Hours by Quantity Sold
Hour of Day
12
13
14
11
15
10
9
16
17
8
0
50000
100000
150000
200000
250000
300000
350000
Total Quantity Sold

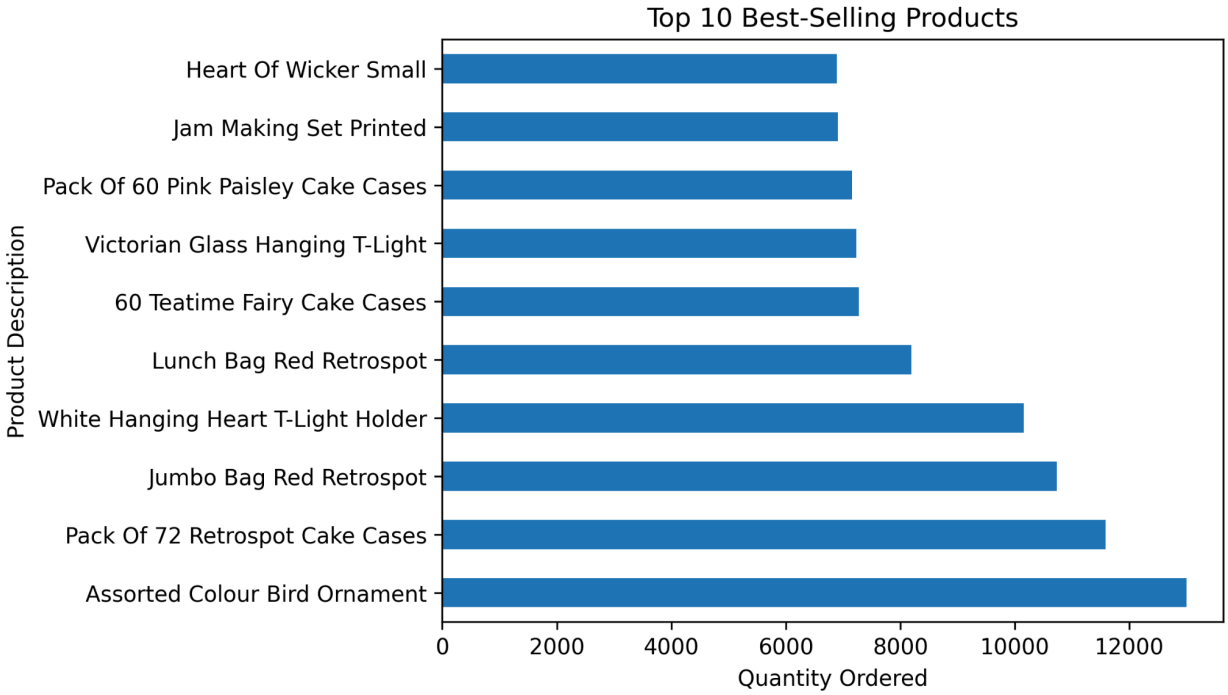

Sales by Day of the Week
Day of the Week
Sunday
Friday
Monday
Tuesday
Wednesday
Thursday
0
100000
200000
300000
400000
Total Sales

Conclusions and Recommendations

This mini project demonstrates how transactional retail data can be transformed into clear, useful insights using fundamental data cleaning, feature engineering, and aggregation techniques. By focusing on product performance and temporal purchasing patterns, the analysis showcases both high impact products and high value time windows for sales activity. The work also surfaces data quality issues (such as missing Saturdays) that would need to be addressed before relying on the dataset for operational decision making.

- The top products should be used for inventory planning and bundling strategies and featured product placements.
- Deepen customer analytics: The analysis of customer data needs to develop into a system which identifies regular customers from first time customers to create targeted loyalty programs and cross sell promotions based on their buying habits of connected products.


Top 10 Most Profitable Products
Jam Making Set With Jars
Set Of 3 Cake Tins Pantry Design
Chocolate Hot Water Bottle
Chilli Lights
Hot Water Bottle Keep Calm
Spotty Bunting
Assorted Colour Bird Ornament
Jumbo Bag Red Retrospot
White Hanging Heart T-Light Holder
Party Bunting
Product Description
0
5000
10000
15000
20000
25000
30000
Total Sales


Top 10 Best-Selling Products
Heart Of Wicker Small
Jam Making Set Printed
Pack Of 60 Pink Paisley Cake Cases
Victorian Glass Hanging T-Light
60 Teatime Fairy Cake Cases
Lunch Bag Red Retrospot
White Hanging Heart T-Light Holder
Jumbo Bag Red Retrospot
Pack Of 72 Retrospot Cake Cases
Assorted Colour Bird Ornament
Product Description
0
2000
4000
6000
8000
10000
12000
Quantity Ordered

References

Samsonova, K. (2026, January 29). *Let's build a mini data analytics project!* LinkedIn.com. Retrieved January 31, 2026, from https://www.linkedin.com/feed/update/urn:li:activity:7422893704379518976/

Ummadisetty, V. (2014). *Online Retail Data Set*. Kaggle.com. Retrieved January 31, 2026, from https://www.kaggle.com/datasets/vijayuv/onlineretail/code

GeeksforGeeks. (2025a, July 23). *Pandas Groupby value counts on the DataFrame*. GeeksforGeeks. https://www.geeksforgeeks.org/python/pandas-groupby-value-counts-on-the-dataframe/

GeeksforGeeks. (2025b, July 28). *Pandas DataFrame.reset_index()*. GeeksforGeeks. https://www.geeksforgeeks.org/pandas/python-pandas-dataframe-reset_index/

GeeksforGeeks. (2025c, September 25). *fstrings in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/python/formatted-string-literals-f-strings-python/

Python, R. (n.d.). *itertools | Python Standard Library – Real Python*. https://realpython.com/ref/stdlib/itertools/

Perplexity AI. (2026, January 31). *Help debugging Python script [Generative AI chat]*. https://www.perplexity.ai