DataQuest Recommender System Analysis

# Introduction

Before building my first recommender system, I conducted a thorough review of the raw dataset and identified several issues that could impact the accuracy of my analysis. To ensure clean and reliable data, I applied the following preprocessing steps:

1. Removed rows with missing CustomerID values to maintain customer integrity.
2. Excluded returns and cancellations, as they do not contribute to meaningful purchase patterns.
3. Filtered out erroneous pricing entries (where UnitPrice ≤ 0) to prevent distortions in item popularity calculations.

After cleaning the data, I converted InvoiceDate to a standardized date format (ymd\_hms) for better time-based analysis. With this refined dataset, I proceeded to develop a recommender system that identifies the most frequently purchased items, laying the foundation for personalized product suggestions.

# Popularity-Based Recommender-System

With this refined dataset, I proceeded to develop a recommender system that identifies the most frequently purchased items, laying the foundation for personalized product suggestions.

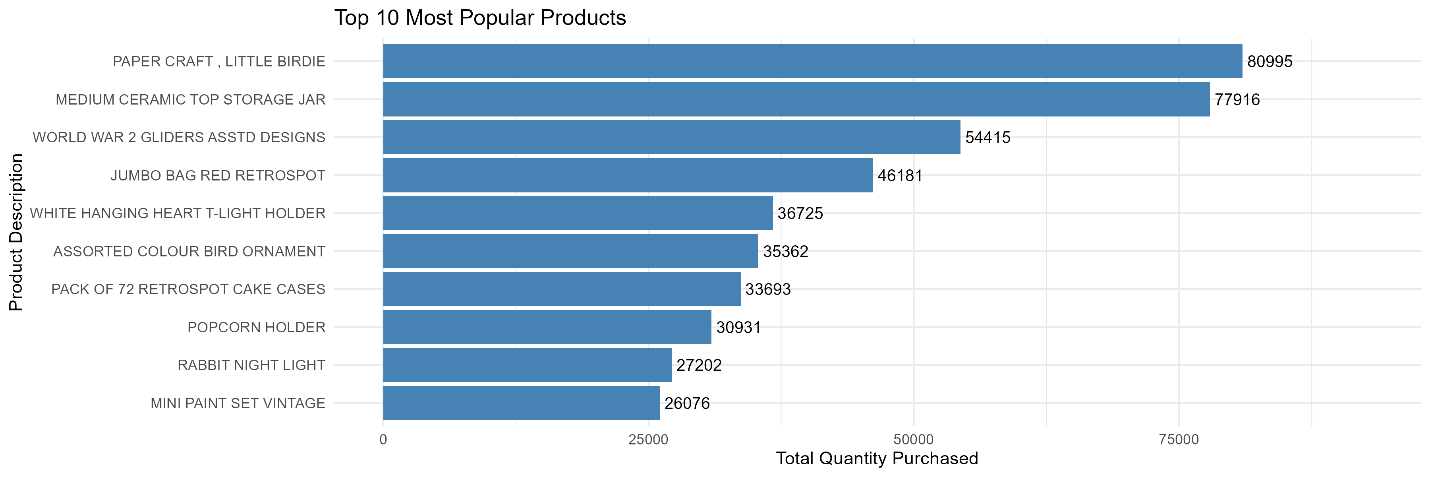


Figure : Most popular Items

Popularity-based recommendations serve as a **strong baseline**, especially for new customers (cold start scenarios) or when behavioral data is sparse. They are computationally efficient and often highlight products with broad appeal.

Initially, I considered using the most frequently purchased items as the foundation for a recommender system. However, upon closer analysis the dataset revealed that a massive portion of transactions originated from the United Kingdom. This data imbalance posed a challenge: global popularity rankings risk being overly influenced by one region, potentially reducing the relevance of recommendations for customers in other countries.

To address this, I expanded the popularity-based system into a group-specific model, incorporating both **regional** (country-level) and temporal (seasonal/monthly) popularity to increase relevance and personalization.

## Group-Based Popularity Recommender

This grouped bar chart shows how the top globally popular products vary in total purchase quantity across different countries. It helps explain why group-specificpopularity recommenders can outperform global recommenders in regional contexts. While global popularity captures general demand, personalization begins by recognizing that not all customers are the same — and neither are their shopping contexts.

### Popular by Country

This grouped bar chart illustrates how the top 10 globally popular products perform across different countries. The massive skew toward the UK reflects both its purchasing volume and its dominance in the dataset — but also highlights why **a** global recommender may not generalize well to smaller markets.

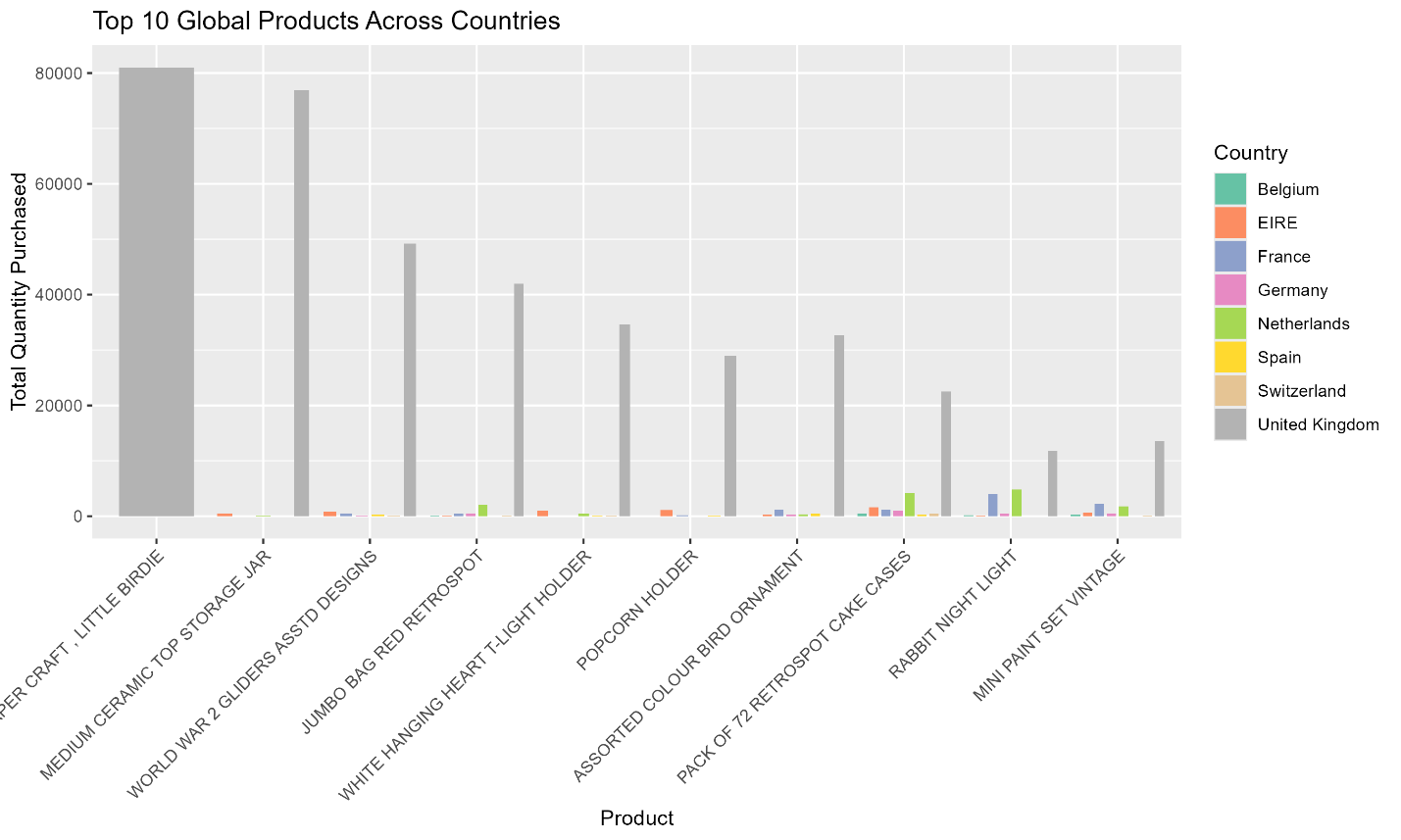


Figure .1: Popular products per country

As the graph above illustrates, there is a massive difference between the quantity of some products bought in other countries and that there is in others and some countries have a higher number of overall purchases globally. Considering regional purchasing trends, I recognized that product popularity varies across different countries.

One particularly striking observation is the overwhelming dominance of the United Kingdom in product purchases. While this reflects the underlying distribution of the dataset, it also underscores the importance of not letting global popularity overrepresent trends from dominant groups. A global recommender trained on this data without adjustment would heavily bias toward UK preferences, which may not serve customers in smaller countries effectively.

Some products that are only moderately popular globally show notable popularity in specific countries — for instance, *RABBIT NIGHT LIGHT* or *MINI PAINT SET VINTAGE* sees relatively higher purchases in countries like Germany or the Netherlands. These long-tail regional signals can be valuable in surfacing products that might otherwise be buried by global trends.

Based on these patterns, I created a country-specific popularity recommender that isolates and leverages the most frequently purchased items within each country, rather than across the entire customer base. This allows the system to serve customers with recommendations that reflect their regional preferences, leading to potentially higher engagement and relevance.

### Country-Month Popularity

While working on the country-based system, I realized that there are some things I tend to buy more during certain times of the year, for example, festive items closer to holidays, or home decor during seasonal changes. I wanted to test whether similar patterns existed in this dataset. That’s what led me to create the visualizations below, showing monthly purchase trends for the top 5 global products across the top 5 countries.

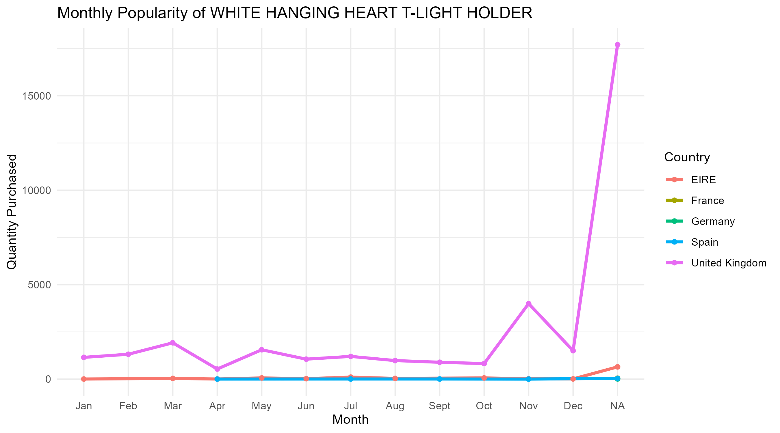
A graph of different colored lines

AI-generated content may be incorrect.

Figure 2.2: Seasonal Popularity

The graphs clearly validate that assumption. None of the top products are purchased in equal amounts throughout the year. For instance, several items show strong surges in November and December, possibly due to holiday shopping or end-of-year gifting. Meanwhile, other products may have mid-year spikes or drop significantly during certain months.

These trends reinforce the idea that seasonality plays a role in product popularity. By only using global averages or country-wide trends without considering time, recommendations may miss key patterns tied to holidays, weather, or other seasonal behaviors.

To improve recommendation relevance, I built a month-aware popularity recommender, which recommends items based on what’s trending in a customer’s region during the current month. This adds a layer of temporal personalization, helping the system stay more dynamic and responsive to seasonal shifts in customer preferences.

For instance, the WHITE HANGING HEART T-LIGHT HOLDER sees significantly higher purchases in **November and December**, particularly in the UK and Germany, which may reflect holiday or winter-themed home decor. Similarly, the RABBIT NIGHT LIGHT shows higher demand in **March and April**, which might align with seasonal events like Easter or spring gifting.

Figure 2.3

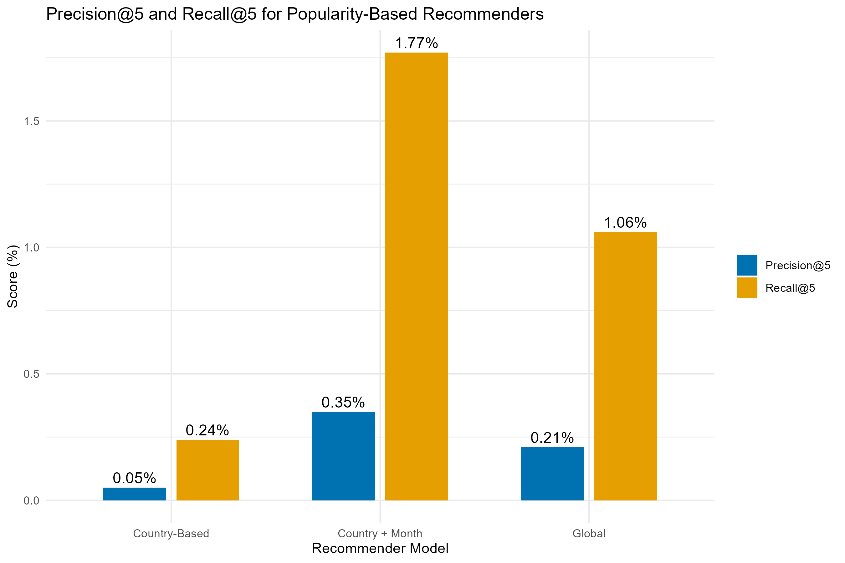
## Comparison Between the Three Systems

Now observing these findings, we see a clear progression in performance across the three popularity-based systems. As more context is considered, first geographical (country), then temporal (month), the recommendations become more relevant and personalized to the customer’s environment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recommender Type | Context Considered | Example Logic | Precision@5 (Sample/Estimated) | Recall@5 *(Sample/Estimated)* |
| Global Popularity | None | Top products overall | 0.0021 | 0.0106 |
| Country-Based Popularity | Geographic (customer's country) | Top products in customer’s country | 5e-04 | 0.0024 |
| Month + Country-Based | Geographic + Temporal | Top products in customer's country during current month | 0.0035 | 0.0177 |

Table : Performance Comparison of the 3 popularity Systems

Precision@5 measures how many of the 5 globally popular items were actually purchased by the customer in the test set and Recall@5 measures ow many of the products the customer actually bought were included in the top 5 recommendations.

The Global Popularity Recommender, which simply recommends the five most purchased products overall, performs poorly on both precision and recall.

* Only 0.21**%** of recommended items were actually relevant to a customer (Precision@5 = 0.0021)
* Only 1.06**%** of what a customer actually bought was captured by these recommendations (Recall@5 = 0.0106)

This low performance highlights the limitation of using global averages, especially in datasets with strong regional or seasonal variations. These results form a baseline for comparison against more context-aware systems.

The Country-Based Popularity Recommender considers where the customer is located and recommends the top items purchased in that country. Surprisingly, its performance was even slightly lower than the global model:

* Only 0.05% of recommended items were actually relevant to a customer (Precision@5 = 0.0005)
* Only 0.24% of what a customer actually bought was captured by these recommendations (Recall@5 = 0.0024)

This may be due to countries with fewer customers or limited product variety, where local trends are too sparse or inconsistent to outperform the global average. Still, the system offers better regional personalization, especially in larger countries like the UK.

The Country + Month Popularity Recommender takes both location and timing into account, recommending products that are popular in the customer’s country during the current month. This model performs the best of the three:

* Only 0.35% of recommended items were actually relevant to a customer (Precision@5 = 0.0035)
* Only 1.77% of what a customer actually bought was captured by these recommendations (Recall@5 = 0.0177)

This improvement shows that seasonal trends matter, customers are more likely to buy items that are trending locally and seasonally. For example, decorative items or gifts may spike during holidays, which a time-aware model is better able to capture.

From this observation, we can conclude that Global Popularity performs modestly, but lacks personalization, Country-Based Popularity reflects regional preferences, but may struggle in smaller markets while Country + Month Popularity adds valuable temporal context, resulting in more relevant and timely recommendations.

These findings show that the more context a system understands, the better it can recommend, which sets the stage for even more personalized techniques like collaborative filtering and hybrid models.

# Collaborative Filtering