Gift Wonder Capstone Project Final Report

Hio Wa Mak

02/10/2025

Background

Gift Wonder (pseudonym), an online retail company specializes in selling unique gift items for various occasions, with a significant portion of its customers being wholesalers.

Problem Statement

The central question for this project is: How can Gift Wonder effectively identify customer segments based on their spending behaviors to optimize targeted marketing strategies and maximize revenue?

Data Source

The dataset for this project is sourced from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/502/online+retail+ii). This dataset includes all transactions made by a UK-based, non-store online retailer between December 1, 2009, and December 9, 2011. The dataset contains 8 columns and over 1 million rows, with each row representing an item of a transaction. Some customers made multiple transactions over the two-year period.

Data Wrangling

To prepare the data for analysis, the following data wrangling steps were performed. First, I examined whether the data in December 2010 in the first dataset overlaps with the data in December 2010 in the second dataset. I found that there were 22,523 duplicated records, which were then removed. Second, the two worksheets were combined into a single dataset, resulting in a dataset with 8 columns and 1,044,848 observations, where each observation represents a transaction item. Second, rows with missing CustomerID values were removed, reducing the dataset by approximately 23%. Third, rows with negative Quantity values, which appeared to represent invalid transactions, were dropped (n = 18,446). Similarly, rows with zero UnitPrice (n = 70), which do not contribute to revenue, were also removed. Afterward, I reviewed the distribution of variables and identified some outliers. However, these outliers were retained, as there was no evidence suggesting they were invalid.

Additional variables were created to enhance the dataset for analysis. The *InvoiceDateTime* variable was broken down into *InvoiceYear*, *InvoiceMonth*, *InvoiceDay*, and *InvoiceTime*. A new variable *Period* was created to indicate which of the two data worksheets the record

came from. Additionally, *TotalPrice*, was computed by multiplying *UnitPrice* with *Quantity*. Finally, the columns were reordered for better organization.

The cleaned final analytic dataset contains 14 columns and 791,045 rows.

Exploratory Data Analysis

Exploratory data analysis suggests that most revenue comes from the United Kingdom (Figure 1). Revenue is highest during September, October, and November, which are likely a peak season for buying gifts (Figure 2). Furthermore, the quantity and number of purchases (invoices) are also highest in those months. Lastly, the number of customers across months show similar patterns as revenue, quantity, and invoices

Although the number of customers is higher during the peak season, the average invoice revenue is relatively stable (Figure 3). This indicates that customers are not making bigger purchases, just due to more customers during those months.

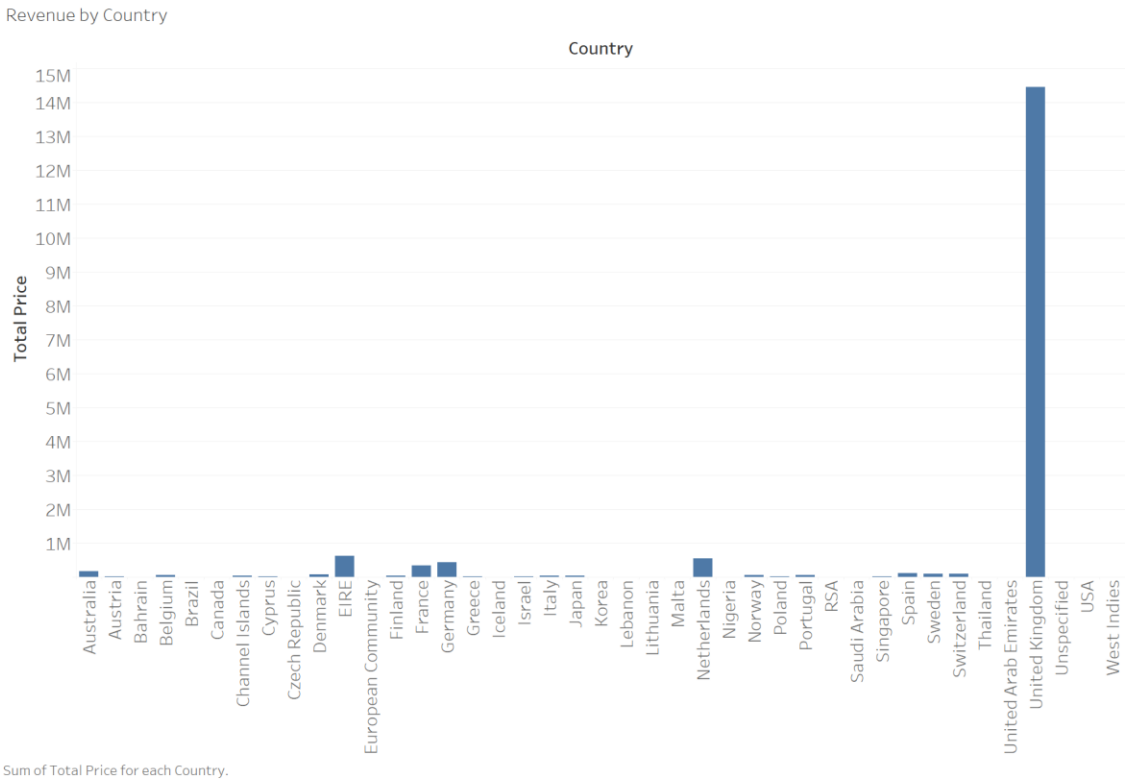

Figure 1. Total Revenue by Country

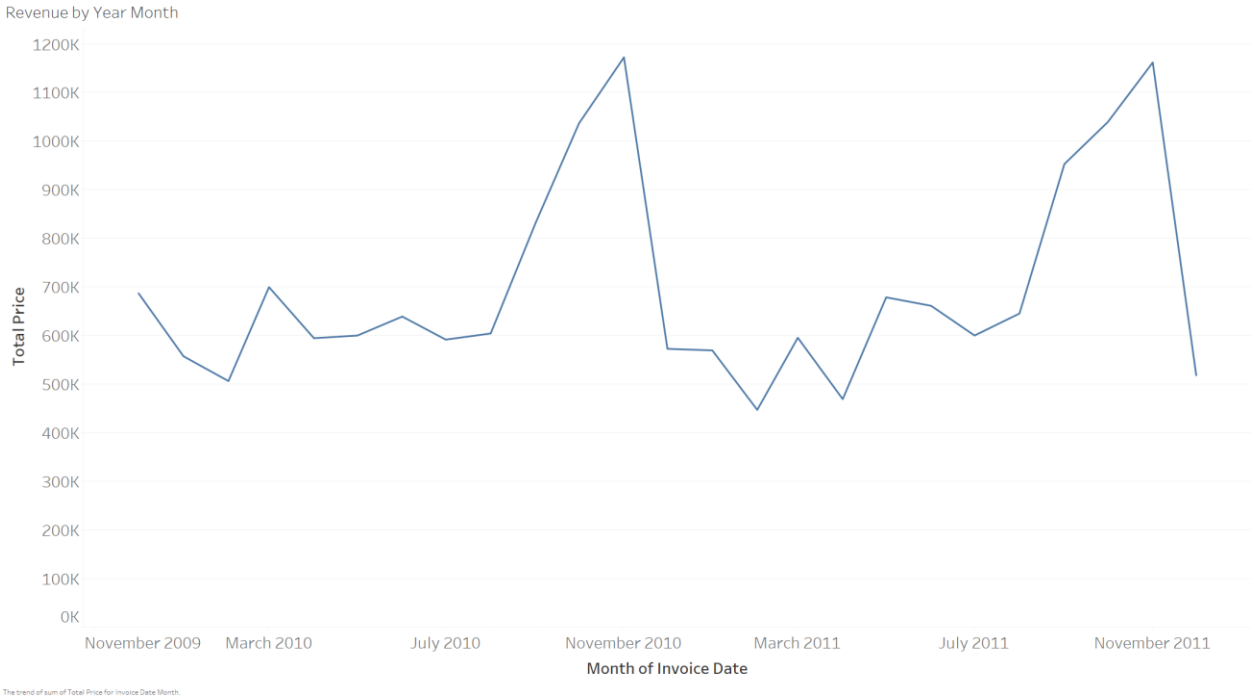

Figure 2. Total Revenue by Month.

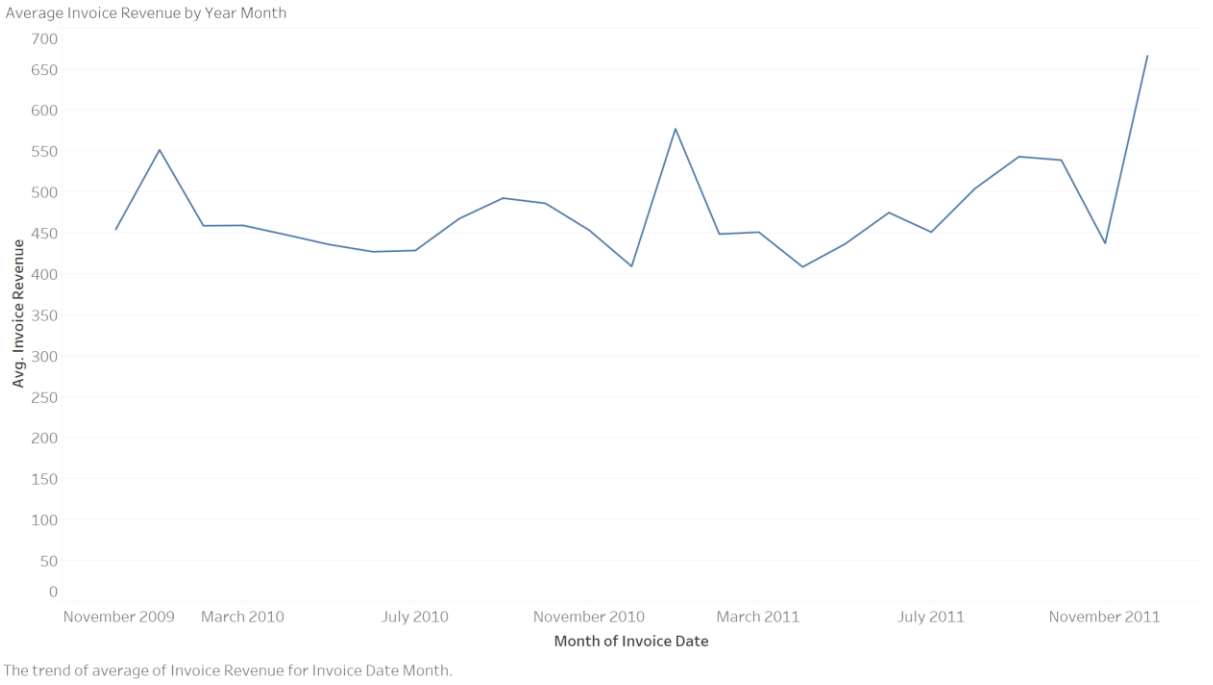

Figure 3. Average Invoice Revenue by Month

Data Preprocessing and Feature Engineering

In the preprocessing stage, the primary goal is to prepare the dataset for K-Means clustering. A crucial step in this process is generating recency, frequency, and monetary value (RFM) variables for customer segmentation:

Recency -- the number of days passed since the last purchase during the two-year period
Frequency -- the total number of purchases made during the two-year period
Monetary Value -- the total amount of money a customer spent during the two-year period

One major challenge is that the data is in a long format, where each row represents an item in an invoice, and customers can have multiple invoices. In other words, items are nested within invoices, which are further nested within customers. The RFM variables are at the customer level. I calculated the three RFM variables and standardized them using z-scores to ensure comparability when applying distance-based K-Means clustering.

Modeling

I determined the optimal number of clusters using the elbow method based on the sum of squared errors, identifying five as the optimal number (Figure 4). A K-Means model with five clusters was then fitted to the data, and the resulting clusters were merged with the RFM dataset for further analysis.

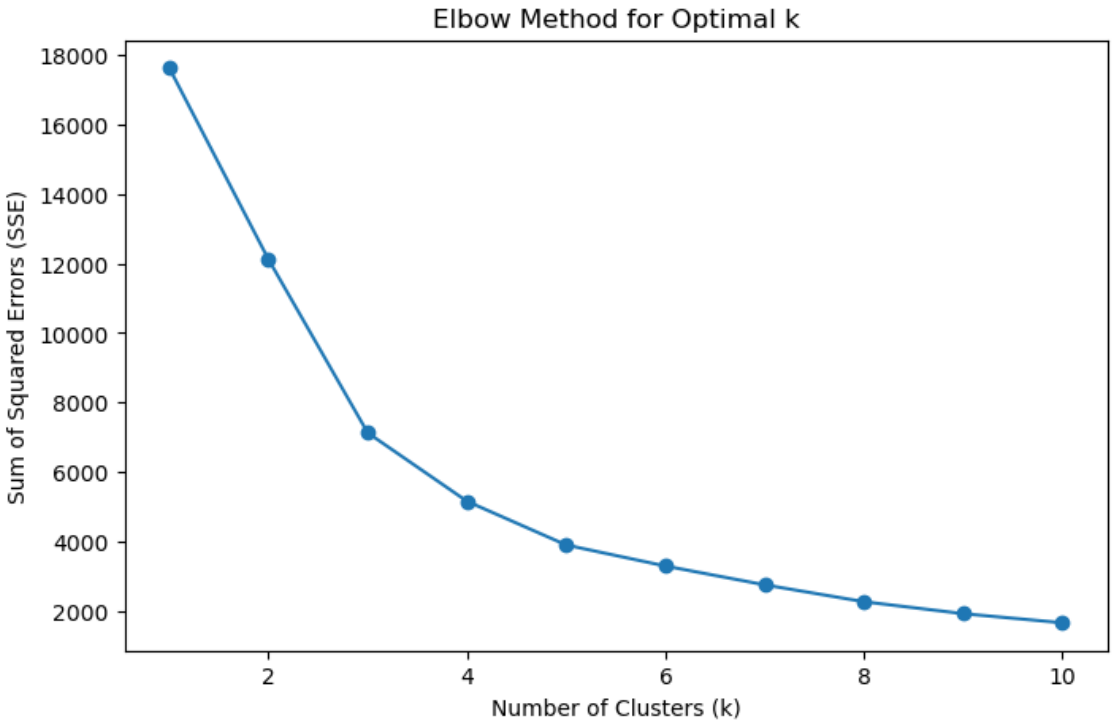

Figure 4. Optimal Number of K using Elbow Method.

Spending Behaviors by Customer Segments

Then, I examined the RFM characteristics of the 5 clusters. The five clusters identified are:

- **Cluster 0:** Lost Low-Value Customers (N = 1914)
- **Cluster 1:** Recent Low-Value Recurrent Customers (N = 380)
- **Cluster 2:** Recent Moderate-Value Recurrent Customers (N = 24)
- **Cluster 3:** Very Recent High-Value Recurrent Customers (N = 4)
- **Cluster 4:** Infrequent Low-Value Customers (N = 3556)

The cluster sizes are shown in Figure 5, with Custer 4 being the largest cluster, and Cluster 3 being the smallest cluster.

\

Figure 5. Customer Segment Sizes.

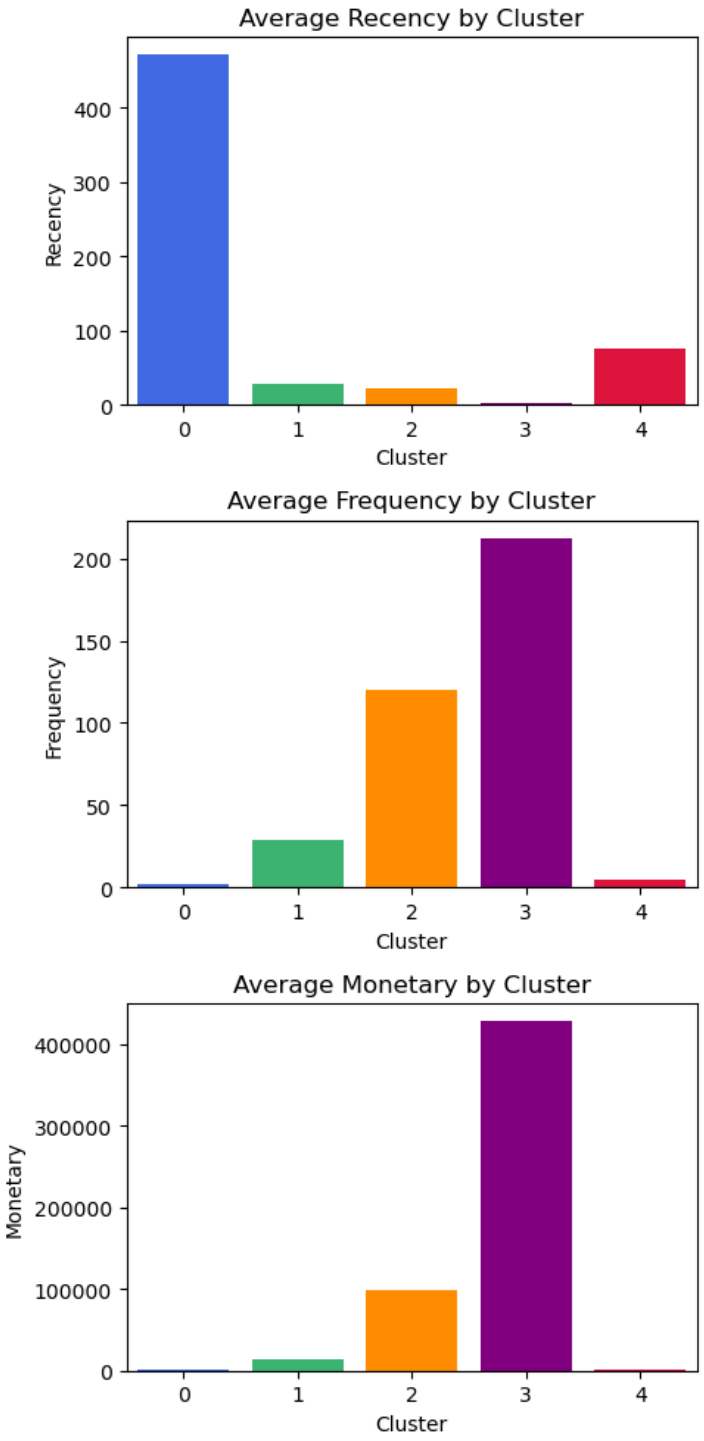

Figure 6. Recency, Frequency, and Momentary Value by Customer Segments.

Note: More recent customers have lower recency value.

Beyond customer-level RFM analysis, I integrated the clusters into the original long-format dataset to gain deeper insights.

Revenue Contribution by Customer Segments

I examined Gift Wonder's total revenue by customer segments across the two-year period. I found that:

- **High- and Moderate-Value Customers (Clusters 2 & 3)**, despite comprising just 28 of 5,878 customers (0.47%), contributed 23.4% (13.6% + 9.8%) of total revenue.
- **Recent & Infrequent Low-Value Customers (Clusters 1 & 4)**, making up 67% of customers (3,936 total), generated 68.4% (29.9% + 38.5%) of total revenue.
- **Lost Low-Value Customers (Cluster 0)** accounted for 8.2% of total revenue.

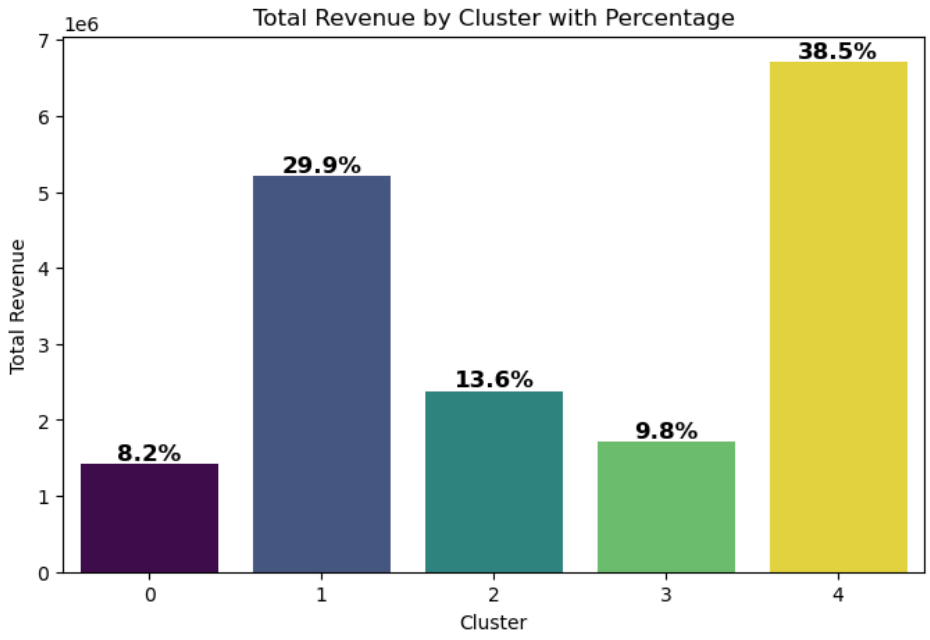

Figure 7. Revenue Contribution by Customer Segments

This customer segmentation highlights the disproportionate impact of high-value customers on revenue and provides strategic insights for targeted customer retention and marketing efforts.

Recommendations

Based on the RFM-based K-Means clustering, I identified five distinct customer segments, each with unique spending behaviors. Below are specific recommendations tailored to each group to optimize customer retention, targeted marketing, and revenue maximization.

High-Priority Segments: Maximizing Revenue from High-Value Customers

Cluster 3: Very Recent High-Value Recurrent Customers (N = 4, 0.07%)

These customers have very high spending and recent purchases, making them the most valuable. They make frequent purchases, and retaining them is critical for long-term profitability.

Recommendations:

- Exclusive VIP Program: Offer personalized incentives, early access to new products, and exclusive discounts to maintain their loyalty.
- Personalized Outreach: Assign a dedicated account manager or send personalized emails with customized product recommendations.

Cluster 2: Recent Moderate-Value Recurrent Customers (N = 24, 0.40%)

This segment spends a significant amount but is smaller in size. They have made recent purchases, showing potential for long-term retention.

Recommendations:

- Targeted Upselling & Bundling: Offer product bundles and discounts on bulk purchases to encourage increased spending.
- Post-Purchase Engagement: Send follow-up emails with discounts on their next purchase and personalized gift recommendations.
- Subscription-Based Model: Encourage them to sign up for a VIP membership or subscription box with exclusive perks.

Growth Opportunity Segments: Strengthening Engagement & Frequency

Cluster 1: Recent Low-Value Recurrent Customers (N = 380, 6.47%)

They have made recent purchases, but their spending is relatively low. This group shows engagement potential if nurtured properly.

Recommendations:

- Increase Purchase Frequency: Send personalized discounts and limited-time offers to encourage repeat purchases.
- Cross-Selling & Product Discovery: Highlight complementary products and feature trending items in targeted email campaigns.
- Rewards: Introduce a loyalty program where repeat purchases unlock higher discounts or exclusive perks.

Cluster 4: Infrequent Low-Value Customers (N = 3,556, 60.5%)

This is the largest customer segment, contributing a significant portion of revenue due to sheer volume. Their spending is low and infrequent, indicating a need for engagement strategies.

Recommendations:

- Email Re-Engagement Campaigns: Send personalized reminders and highlight seasonal deals to encourage them to return.
- Limited-Time Offers & Free Shipping: Provide discounts or free shipping for their next purchase to reduce hesitation.
- Social Media & Retargeting Ads: Use social media retargeting to display ads featuring products based on their past browsing history.

Low-Priority Segment: Reactivating Lost Customers

Cluster 0: Lost Low-Value Customers (N = 1,914, 32.5%)

These customers have not purchased recently and contribute minimal revenue. Reactivating them requires cost-effective strategies since their overall value is low.

Recommendations:

- Send "We Miss You" emails with a small discount to encourage another purchase.
- Identify Lapsed Reasons: Use customer surveys to understand why they stopped purchasing and adjust marketing strategies accordingly.

Additional Strategic Recommendations

- Seasonal & Wholesale Focus
    - Capitalize on Peak Seasons (Sept-Nov): Since revenue is highest during these months, create targeted promotions early to maximize sales.

    - Wholesale Customer Strategy: Since a significant portion of customers are wholesalers, create bulk purchase incentives and a wholesale membership program.
- Refined Marketing Strategies
    - Email Segmentation & Automation: Use segmented email marketing to send different offers based on cluster behaviors.
    - Referral & Social Media Engagement: Encourage referrals by offering discounts to customers who bring in new shoppers.