# Customer Purchase Data Analysis

Nabeel Kamal Shamsi

2025-12-14

## Introduction

This report presents an analysis of customer purchase data. The objective is to explore patterns in revenue, average purchase amounts, sales volume, and customer demographics such as gender and country.
The analysis will include data cleaning, summarization, and visualization to generate actionable insights for business decision-making.

### Import libraries

```
knitr::opts_chunk$set(echo = TRUE)
# Data import & summarization
library(readr)
library(skimr)

# Data manipulation & cleaning
library(tidyverse)
library(janitor)
library(lubridate)

# Visualization helpers
library(scales)
```

### Overview of the dataset

Table 1: Data summary

| | |
|---|---|
| Name | data |
| Number of rows | 15000 |
| Number of columns | 7 |
| Column type frequency: | |
| character | 4 |
| numeric | 3 |
| Group variables | None |

### Variable type: character

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| Gender | 0 | 1 | 0 | 6 | 750 | 4 | 0 |
| Country | 0 | 1 | 0 | 9 | 1200 | 7 | 0 |
| Purchase.Date | 0 | 1 | 0 | 10 | 1050 | 1461 | 0 |
| Product.Category | 0 | 1 | 0 | 14 | 900 | 7 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| User.ID | 0 | 1.00 | 7500.50 | 4330.27 | 1.00 | 3750.75 | 7500.50 | 11250.25 | 15000.00 |
| Age | 1500 | 0.90 | 43.40 | 14.93 | 18.00 | 31.00 | 43.00 | 56.00 | 69.00 |
| Purchase.Amount | 1800 | 0.88 | 253.22 | 143.11 | 5.05 | 130.33 | 253.64 | 378.59 | 499.95 |

## Data cleaning

**Before performing any analysis, the dataset was carefully cleaned to ensure data quality and consistency.**

- The following steps were applied:
- Handling missing values: All blank entries in the dataset were converted to NA, and rows containing any missing values were removed. This ensures that all remaining data is complete and reliable for analysis.
- Standardizing column names: Column names were cleaned and standardized to a consistent format using clean_names(). This makes the dataset easier to work with and reduces the risk of errors due to inconsistent naming.
- Checking date consistency: The purchase_date column was converted to a proper Date format. Any illogical dates (earlier than 2019-01-01 or later than today) were identified to ensure date-related analysis is accurate.
- Verifying age ranges: The maximum and minimum age values were checked to confirm that age data is valid and there are no extreme or erroneous entries.

```
##          User.ID              Age           Gender          Country
##                0             1500              750             1200
##  Purchase.Amount    Purchase.Date Product.Category
##             1800             1050              900
```

```
## [1] 9076
```

```
## [1] user_id          age              gender           country
## [5] purchase_amount  purchase_date    product_category
## <0 rows> (or 0-length row.names)
```

Maximum and minimum age

```
## [1] 69
```

```
## [1] 18
```

## Cleaned Dataset Overview

The dataset has been cleaned to ensure accuracy and consistency.
Missing or blank values have been removed, column names have been standardized, dates have been checked for validity, and the age column has been verified for reasonable values.

Below is a preview of the cleaned data showing customer demographics, purchase amounts, dates, and product categories.

```
##   user_id age gender   country purchase_amount purchase_date product_category
## 1       1  56 Female       Usa          331.79    2021-11-21           Sports
## 2       2  69   Male Australia          335.72    2022-03-05   Home & Kitchen
## 4       4  32   Male   Germany           80.97    2023-06-08           Sports
## 7       7  38 Female    Canada          222.20    2022-02-23           Beauty
## 8       8  56   Male       Usa          217.27    2021-09-08           Sports
## 9       9  36   Male Australia          116.53    2022-06-04         Clothing
```

**Total revenue per country with percentage**

| country | total_revenue | avg_purchase_amount | sales_volume | percentage |
|---|---|---|---|---|
| France | 413149.8 | 251.7671 | 1641 | 18% |
| Canada | 381937.2 | 249.9589 | 1528 | 16.7% |
| Germany | 378364.6 | 251.2381 | 1506 | 16.5% |
| Usa | 375926.8 | 258.1915 | 1456 | 16.4% |
| Australia | 373731.3 | 252.8628 | 1478 | 16.3% |
| Uk | 368060.9 | 250.8936 | 1467 | 16.1% |

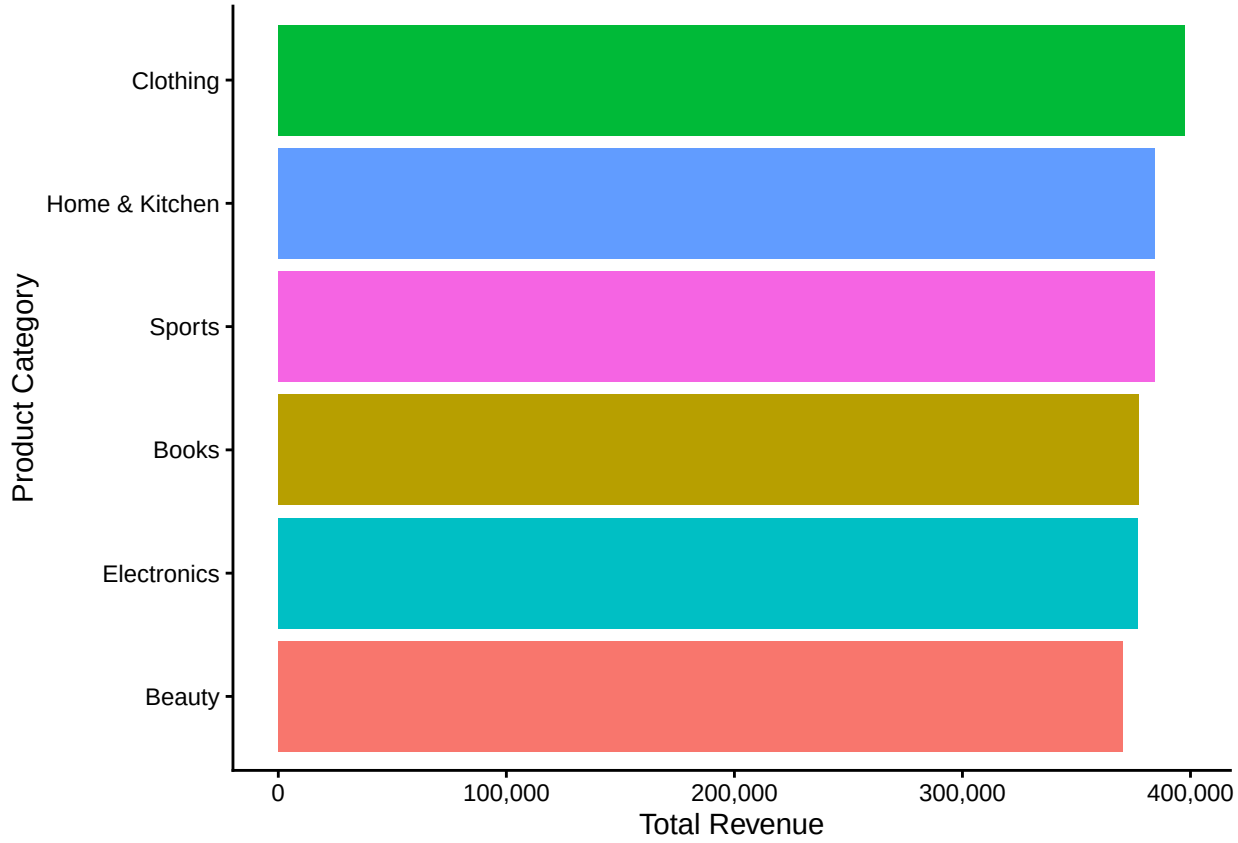
**Interpretation:**
France has the highest total revenue at 18% of overall sales, followed closely by Canada (16.7%) and Germany (16.5%).
The remaining countries (USA, Australia, UK) each contribute around 16% of the total revenue, showing a fairly balanced distribution among top markets.

**Average spend per gender with percentage**

| gender | avg_purchase_amount | total_revenue | sales_volume | percentage |
|---|---|---|---|---|
| Female | 255.5703 | 787667.7 | 3082 | 34.37840 |
| Male | 249.7634 | 747541.7 | 2993 | 32.62707 |
| Other | 251.9031 | 755961.1 | 3001 | 32.99454 |

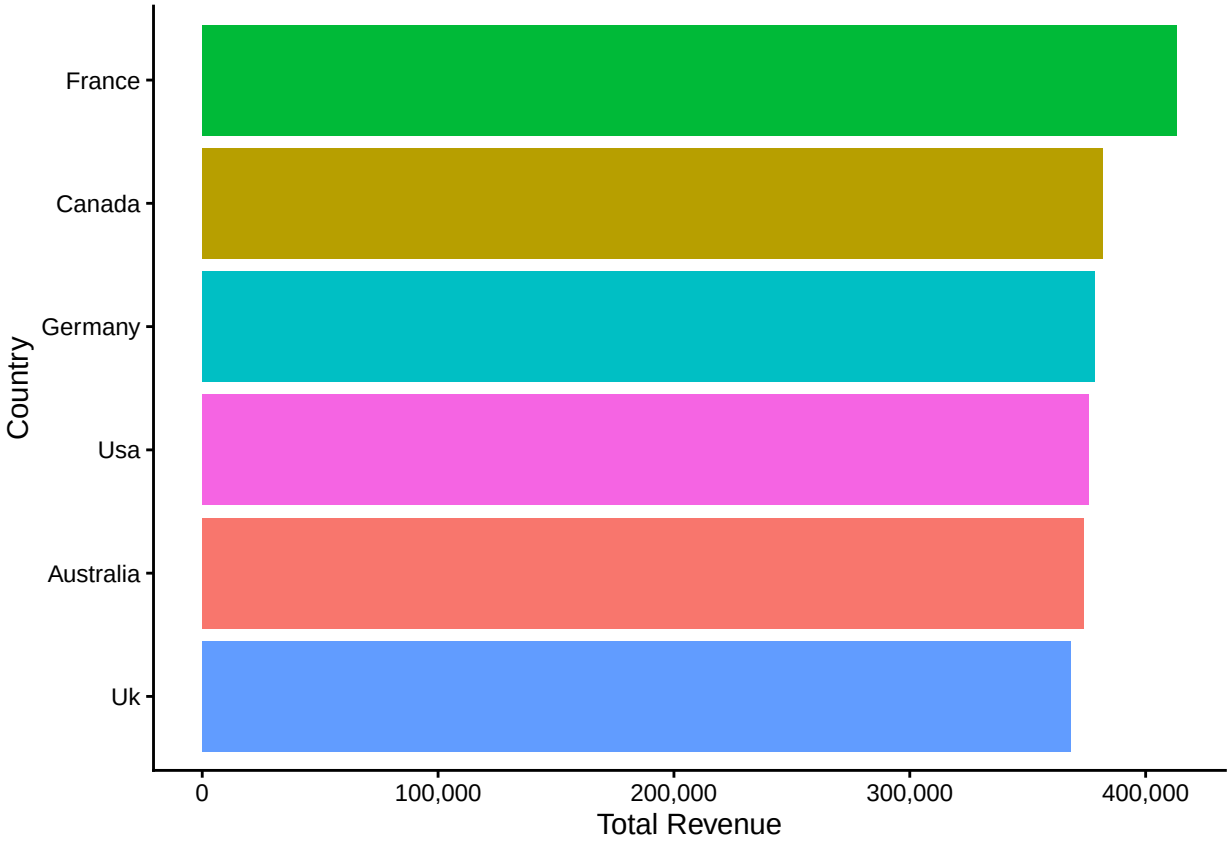
**Interpretation:**
Female customers contribute the highest share of total revenue at 34.4%, with an average purchase amount of 255.57.
Male and Other customers are fairly close, contributing 32.6% and 33.0% of revenue respectively, showing a relatively balanced spending pattern across genders.

**Total revenue per product category**



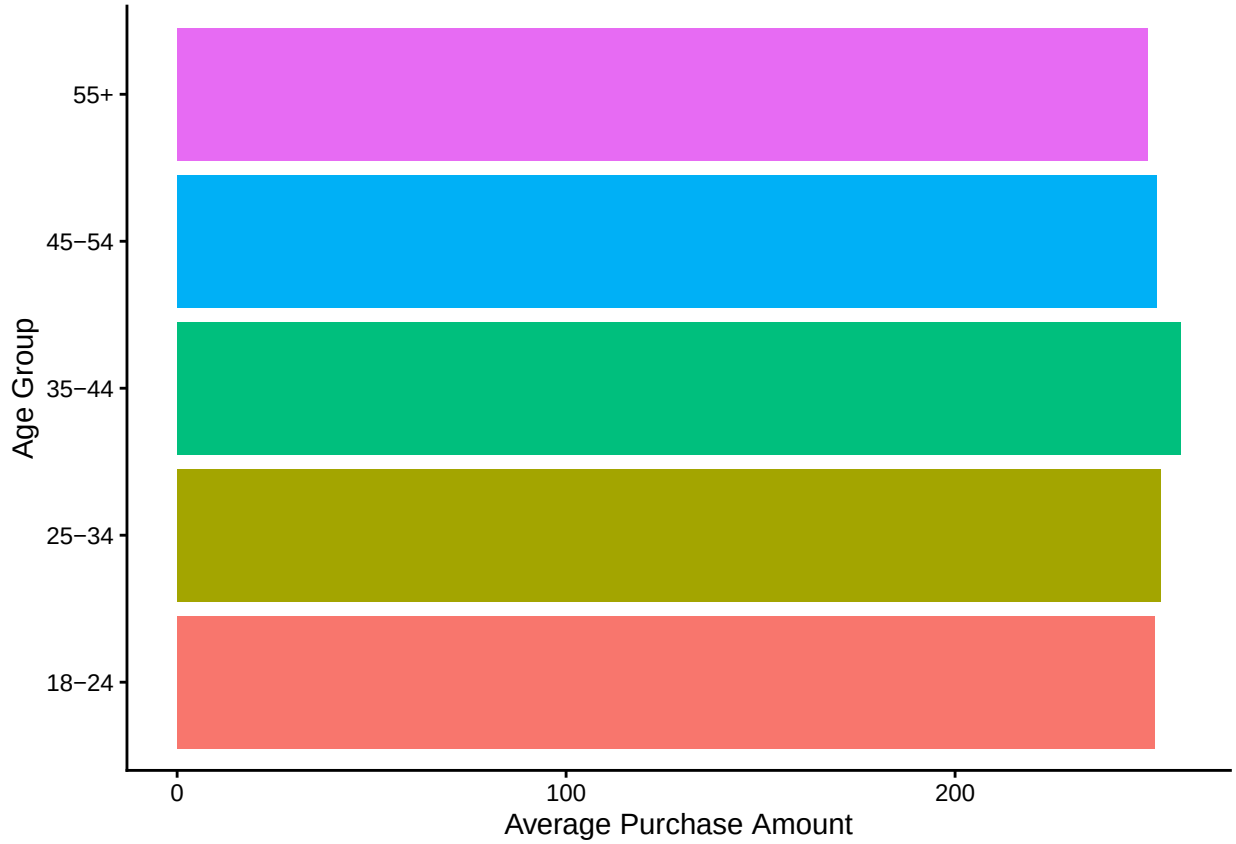

**Interpretation:**
The bar chart shows total revenue for each product category. Clothing is the top revenue generator, approaching $400,000.
Home & Kitchen, Sports, Books, and Electronics are closely clustered between $360,000 and $380,000, showing a highly competitive middle tier.
Beauty has the lowest revenue, just below the middle tier, at around $350,000. Overall, revenue is fairly diversified across categories.

**Total Revenue per Country**



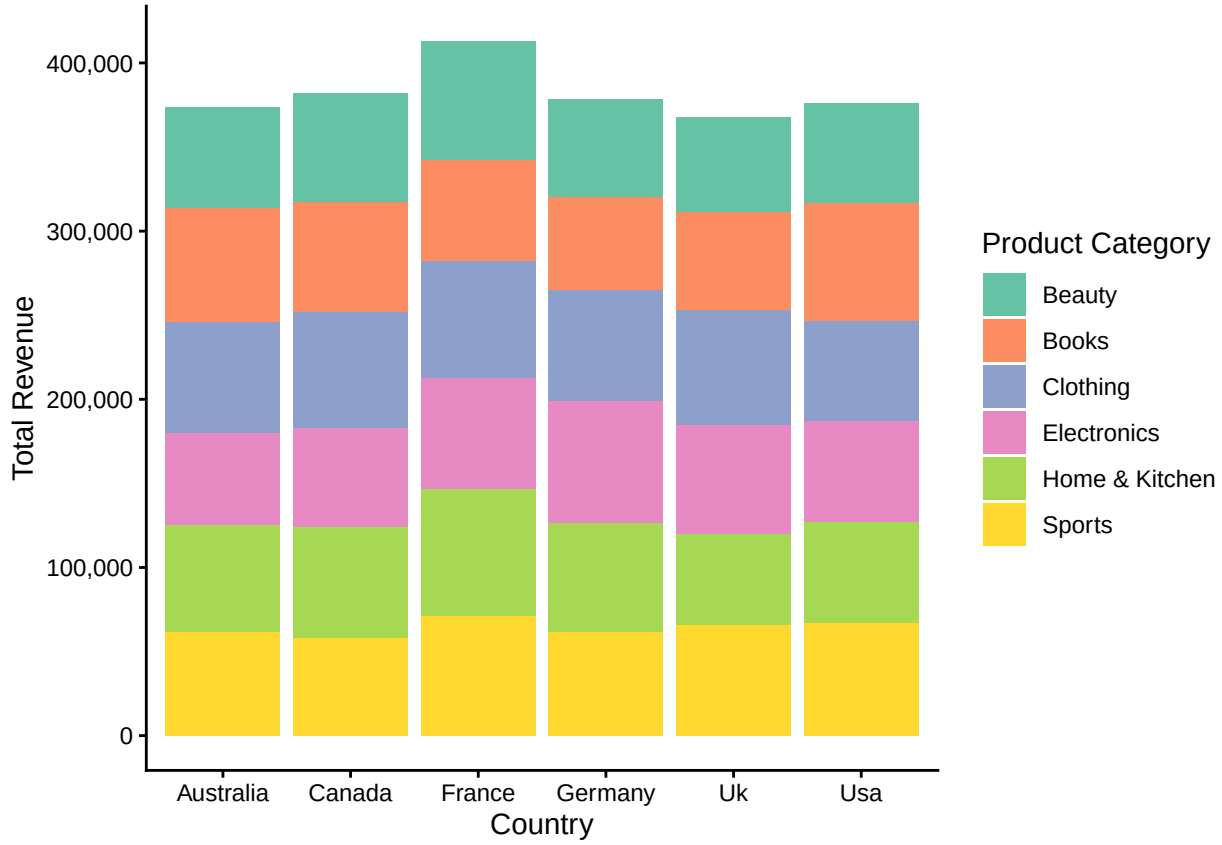

**Interpretation:**
France generates the highest total revenue, slightly above 400,000, while the UK records the lowest, at around 375,000.
Canada, Germany, USA, and Australia have very similar revenue levels, clustered between approximately 380,000 and 390,000, with no clear ranking among them.
Overall, France leads clearly, the UK trails, and the remaining countries form a tightly grouped mid-tier.

**Average Purchase Amount by Age Group**



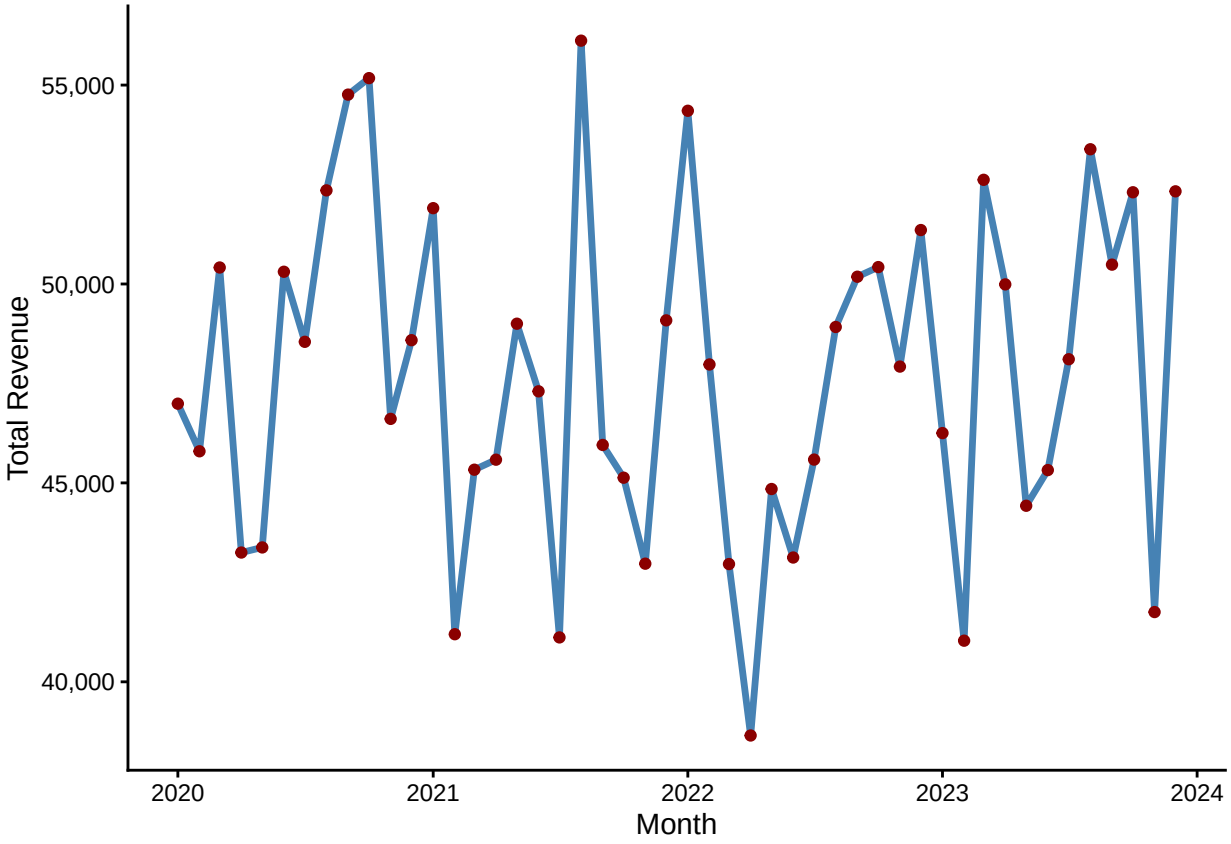

**Interpretation:**
The bar chart shows average purchase amount by age group. Customers aged **35–44** have the highest average spending, making them the top-spending group.
The **18–24**, **25–34**, and **55+** age groups show very similar purchase levels, forming a mid-range cluster.
The **45–54** age group has the lowest average purchase amount. Overall, spending peaks in the middle-age segment and is slightly lower among older customers.

**Product Mix by Country**



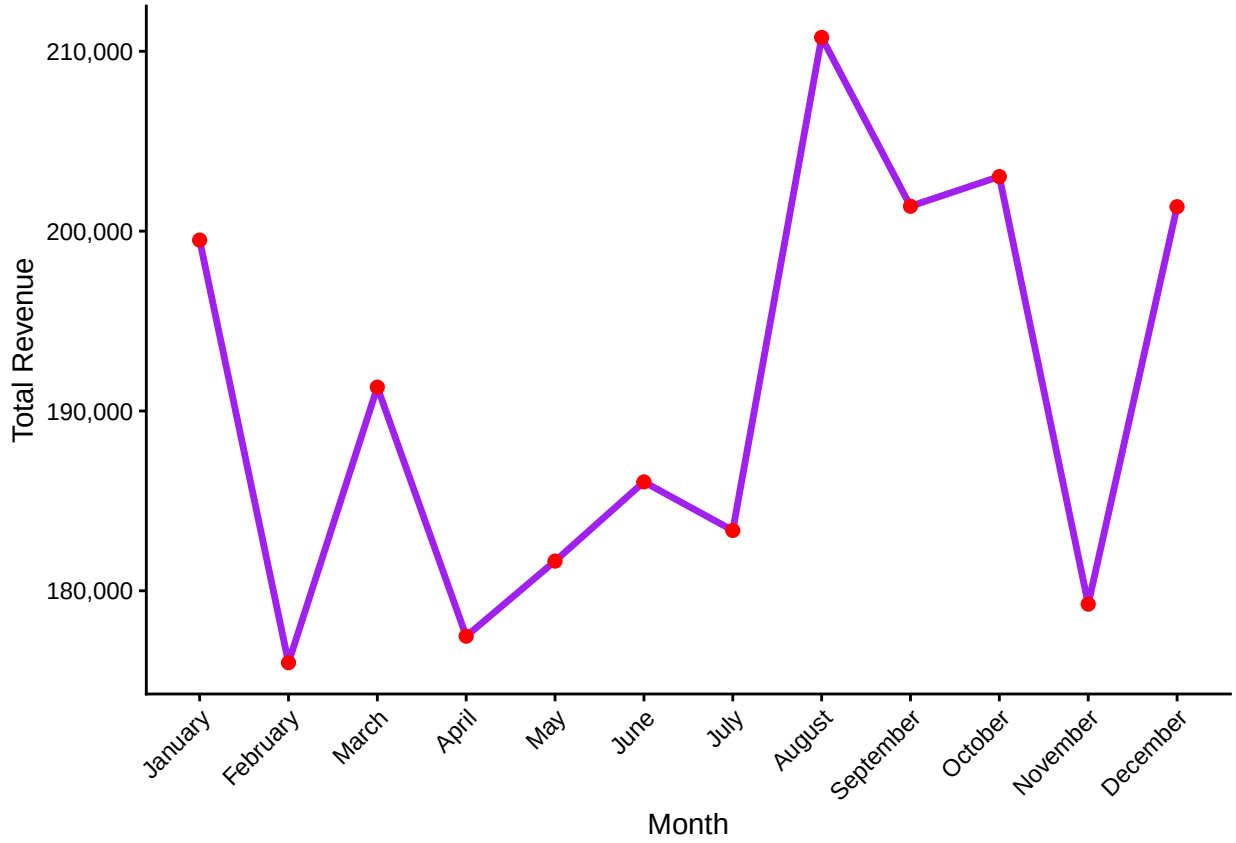

**Interpretation:**
France emerges as the market leader, driven by exceptionally strong performance, particularly in the Clothing category.
While most countries display a relatively balanced revenue mix, differences in the Clothing, Books, and Beauty segments appear to be the main contributors to variation in total revenue.
These category-level differences help explain the gap between the top-performing country (France) and the others, such as the UK and USA.

**Sales Trend Over Time**



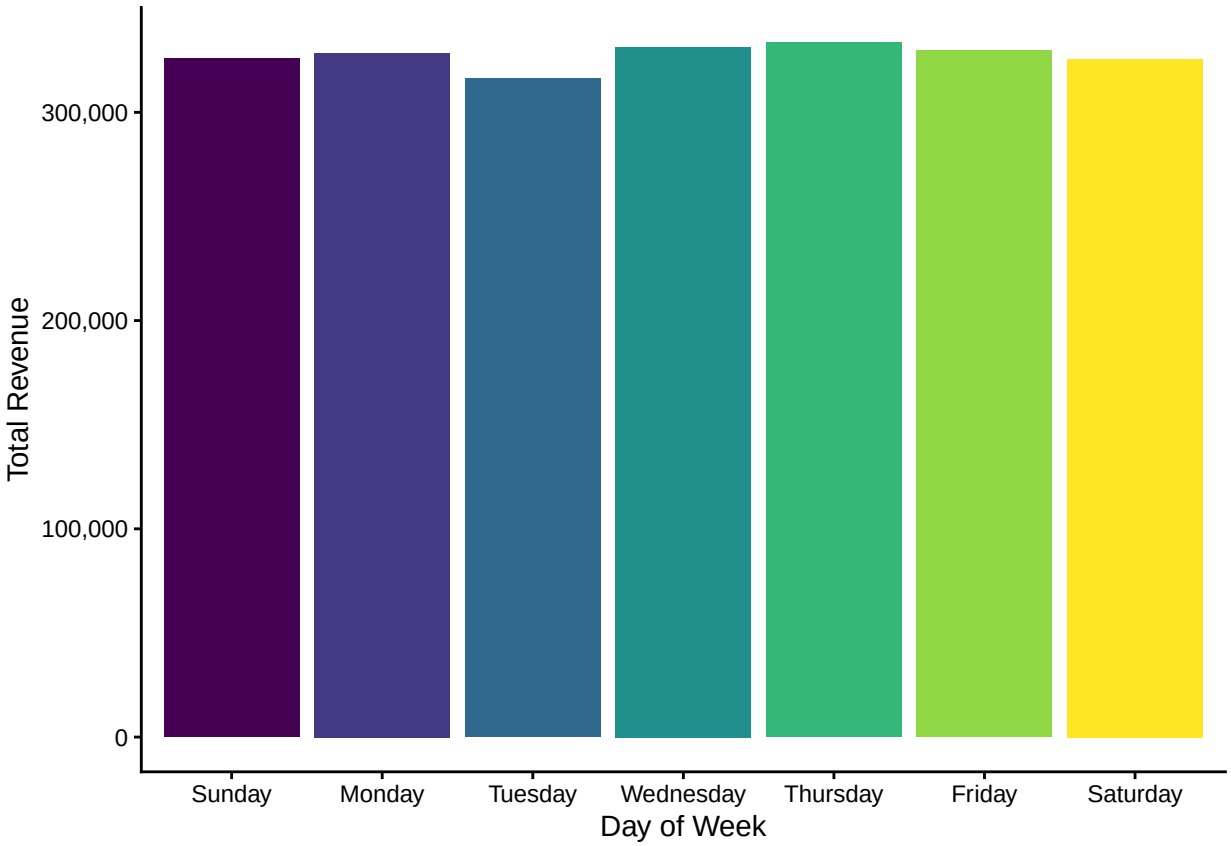

**Interpretation:**
The line chart shows monthly total revenue over a four-year period with strong fluctuations. Revenue exhibits clear seasonal patterns, with recurring peaks and troughs each year, indicating high volatility.
Despite these short-term swings, there is no clear long-term upward or downward trend, suggesting that overall revenue has remained relatively stable over time rather than experiencing sustained growth or decline.

**Plot monthly seasonality**




**Interpretation:**
From 2020–2024, revenue shows high volatility with a flat long-term trend. Peaks occur in August and year-end (Sep–Dec), while troughs appear in February, April, and November. Management should leverage the August peak and address early-spring and late-autumn dips.

## Total Revenue by Day of the Week




**Interpretation** Revenue remains stable throughout the week, mostly ranging from $315K–$325K, indicating consistent demand

## Recommendations for Business Strategy

### 1. Market and Product Strategy

- **Reinforce France's Clothing Segment:** France leads revenue, driven by Clothing (18% of total). Allocate marketing and inventory to strengthen this core market.
- **Investigate Mid-Tier Markets:** Canada, Germany, USA, and Australia generate ~16% each. Analyze categories or demographics limiting growth to unlock potential.
- **Boost Low-Performing Categories:** Beauty generates the lowest revenue. Use targeted campaigns or bundles, especially in the UK and USA, to increase contribution.

### 2. Time-Based Strategy

- **Capitalize on Peak Season:** Prepare inventory and marketing for August peaks and Q4 (Sep–Dec) to maximize revenue.

- **Mitigate Revenue Troughs:** Implement promotions or inventory strategies during low months (Feb, Apr, Nov) to smooth volatility.

- **Maintain Uniform Daily Operations:** Revenue is consistent across weekdays. Focus on operational efficiency rather than shifting resources by day.

**3. Customer Focus**

- **Prioritize 35–44 Age Group:** This segment spends the most. Offer premium products and loyalty programs targeting them.

- **Ensure Gender Balance:** Revenue is fairly balanced across Female (34.4%), Male (32.6%), and Other (33%). Maintain inclusive marketing to appeal broadly.