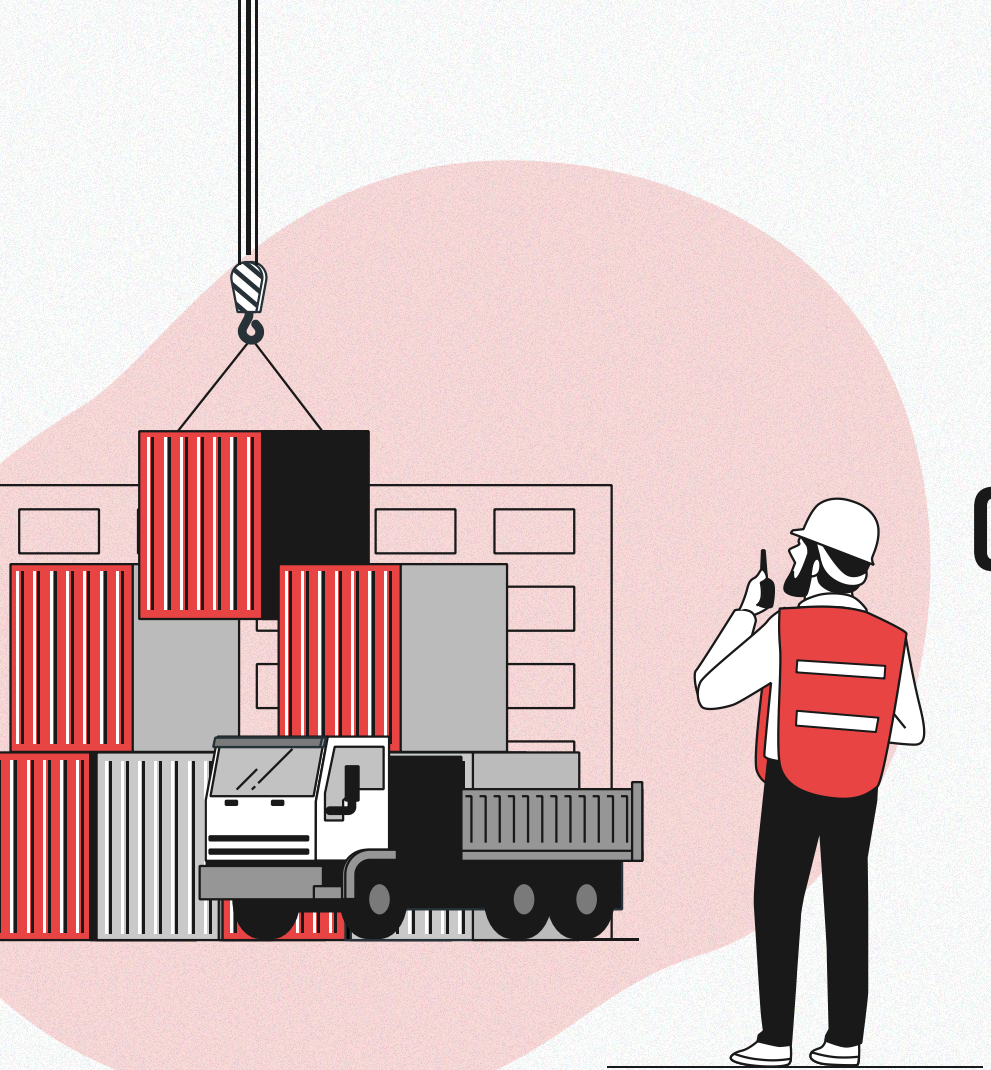


CAPSTONE PROJECT: ONLINE RETAIL CUSTOMER ANALYTICS & SEGMENTATION

Data-Driven Customer Insights & K-Means Clustering

Prepared by: Abdallah

PROJECT OBJECTIVE & DATASET OVERVIEW

•**Objective:** Act as Data Analysts to perform deep Exploratory Data Analysis (EDA), extract actionable business insights, and build data-driven customer segments.

•**Dataset:** UCI Online Retail Dataset (Dec 2010 - Dec 2011).

•**Data Cleaning Decisions:**

1. Converted InvoiceDate to datetime.
2. Dropped transactions missing CustomerID (critical for segmentation).
3. Removed exact duplicates and filtered out cancellations (negative quantities/prices) to ensure pure sales analysis.

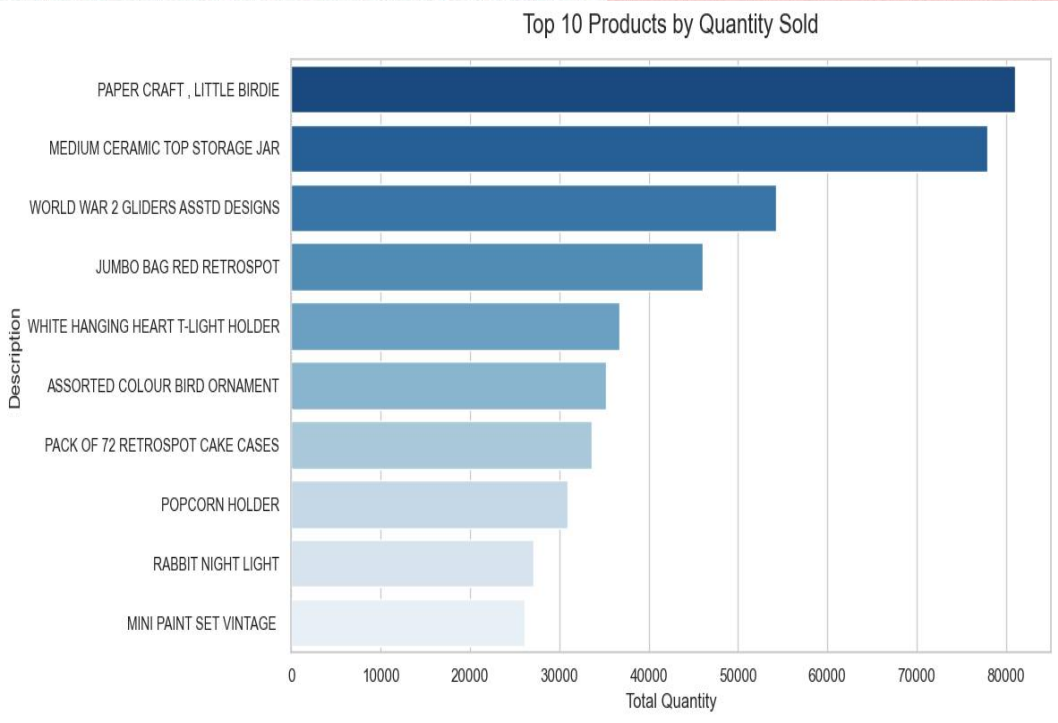
EXPLORATORY DATA ANALYSIS - PRODUCT & MARKET INSIGHTS

TOP PRODUCTS:

* *BY QUANTITY:*

-“PAPER CRAFT ,LITTLE BIRDIE"

-(HIGH-VOLUME, LOW-MARGIN DRIVER).

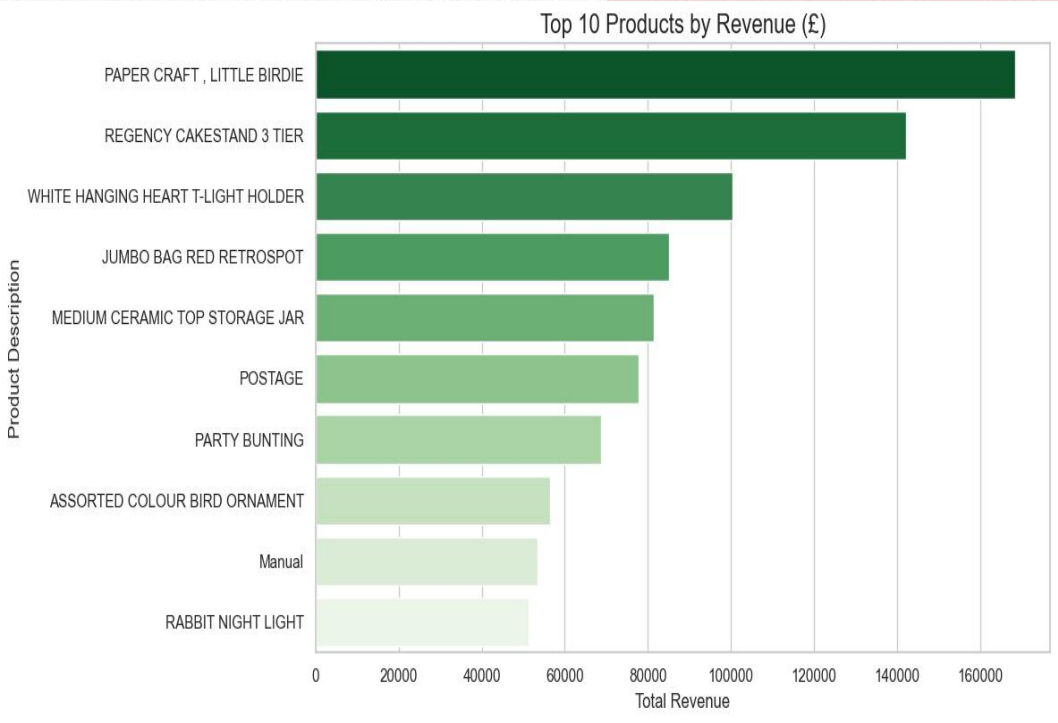


TOP PRODUCTS:

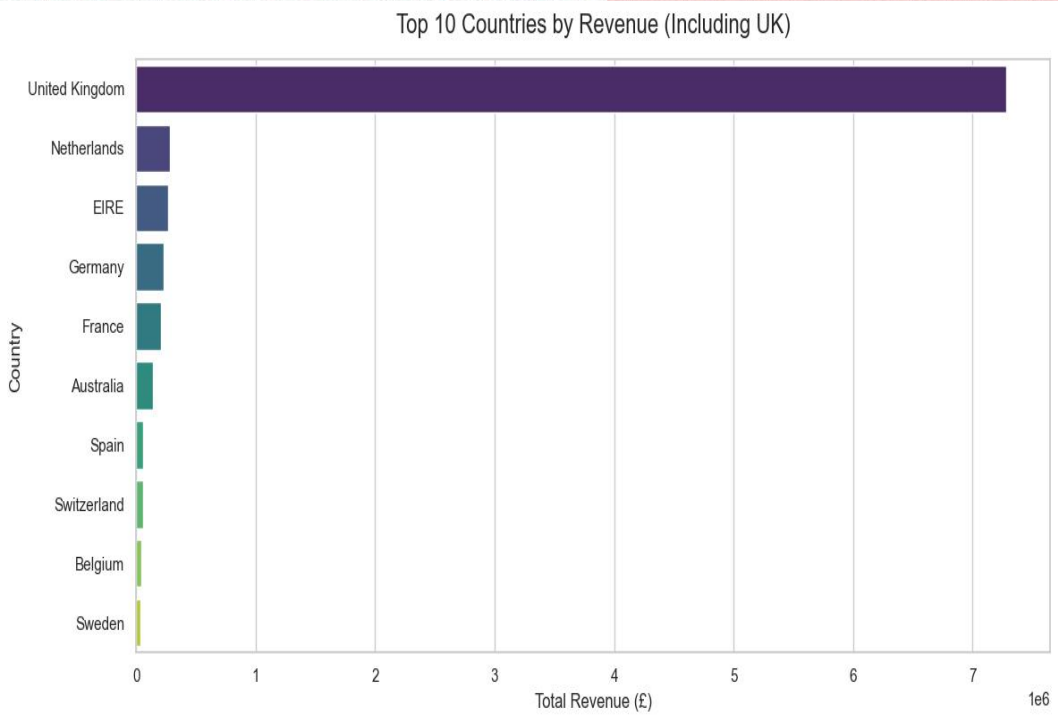
*_BY REVENUE:_

-" PAPER CRAFT ,LITTLE BIRDIE "

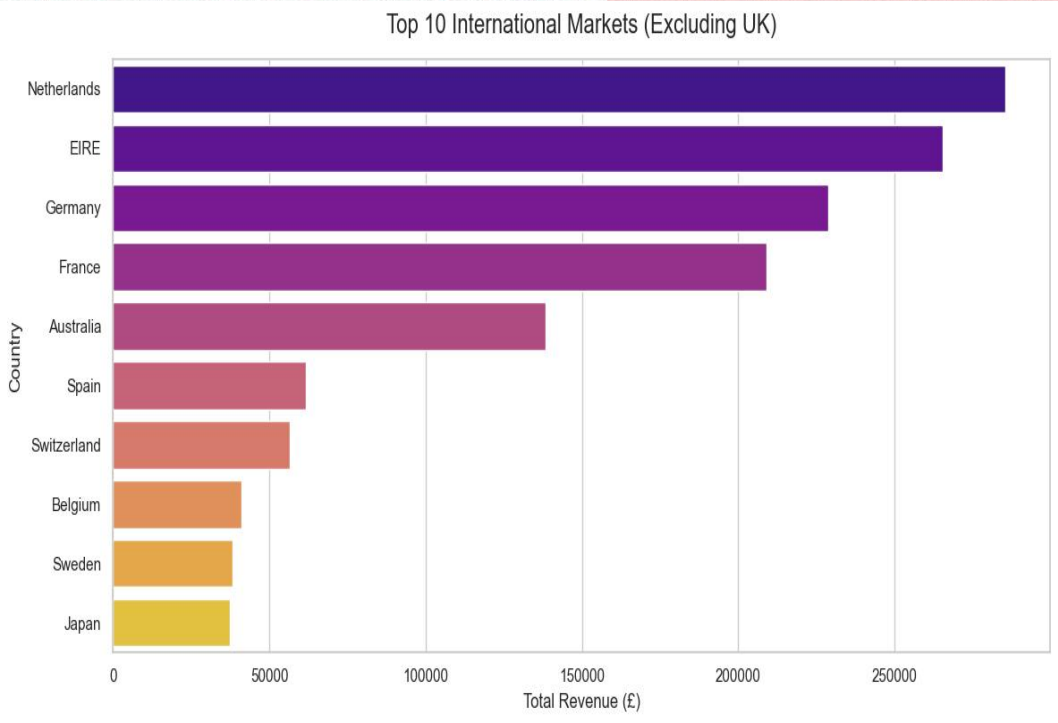
-(THE HERO PRODUCT GENERATING MAXIMUM REVENUE).



- **GEOGRAPHICAL DOMINANCE:**
  THE UK DOMINATES OVER 90% OF TOTAL REVENUE.



- **GEOGRAPHICAL DOMINANCE:** ***INTERNATIONAL MARKETS (EXCLUDING UK):*** **NETHERLANDS ,EIRE AND GERMANY ARE THE TOP 3 REVENUE-GENERATING FOREIGN MARKETS.**

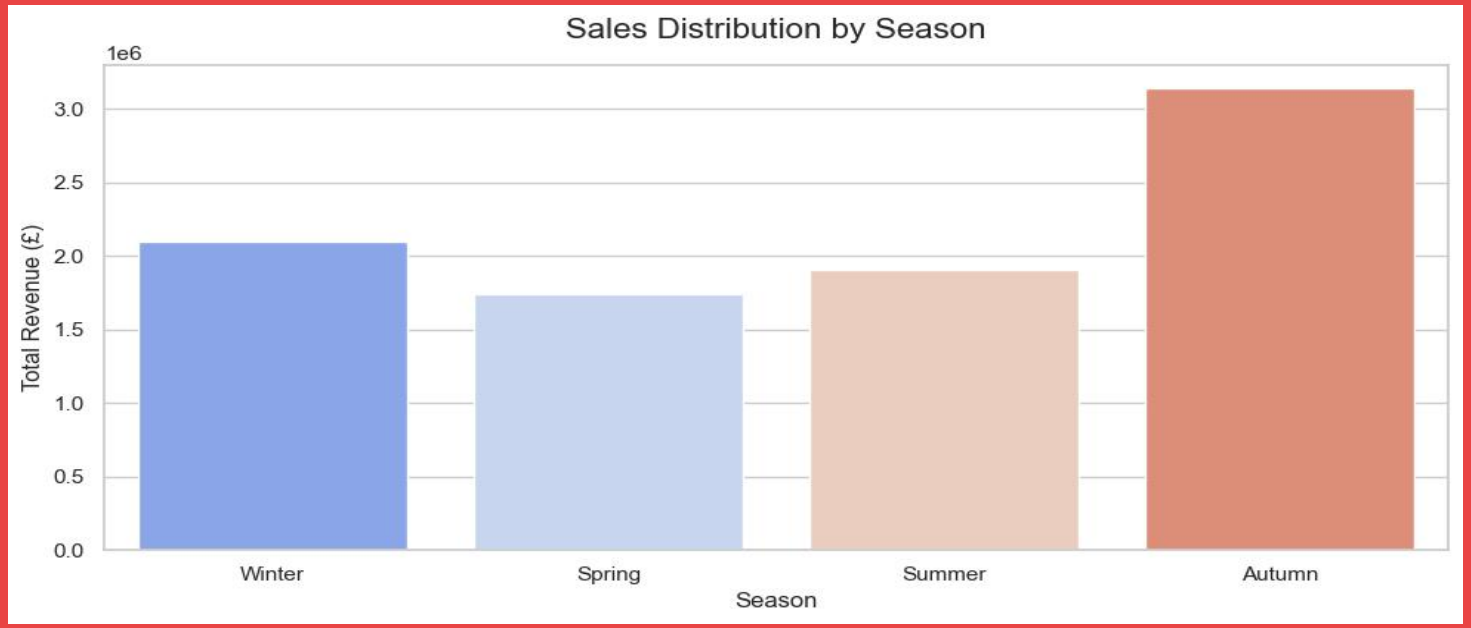


EXPLORATORY DATA ANALYSIS - TEMPORAL SALES TRENDS

SEASONALITY (Q4 PEAK)

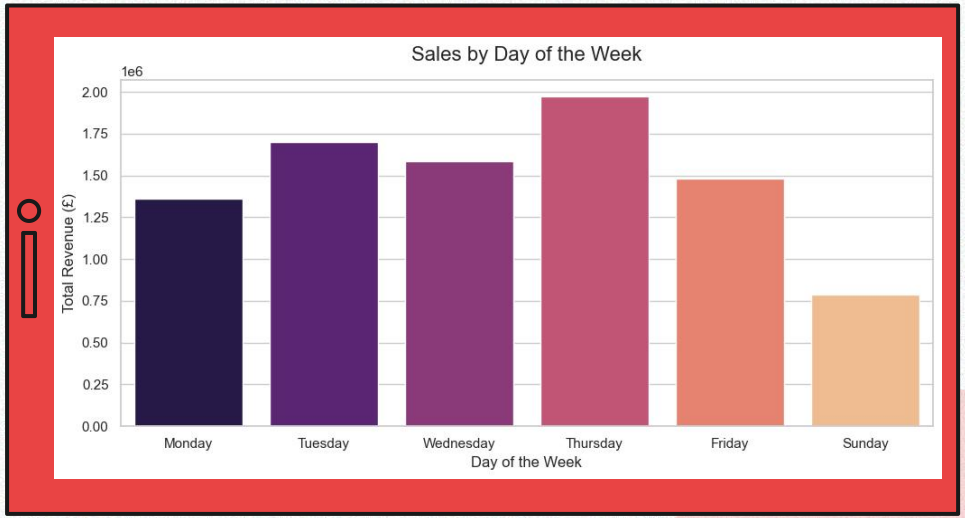
- A massive revenue surge in Septemper, October, November, and December due to Q4 holiday and Christmas shopping

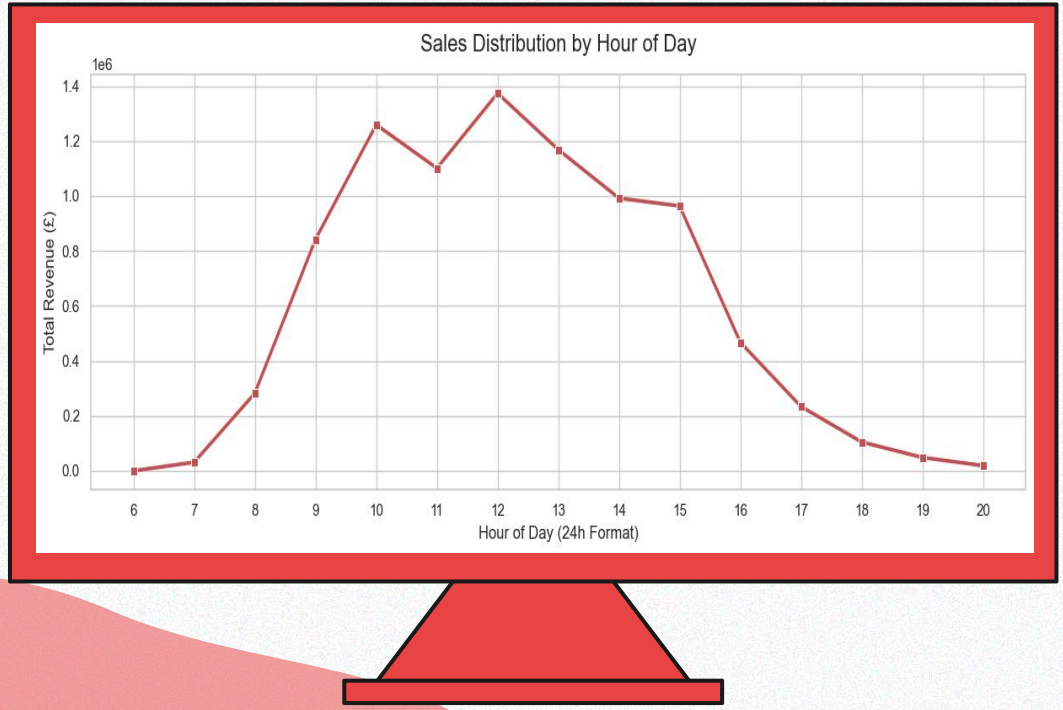
SEASONALITY (Q4 PEAK)



OPERATIONAL BLINDSPOT (THE SATURDAY MYSTERY):

- Zero transactions recorded on Saturdays, indicating a potential revenue loss due to weekend closure.



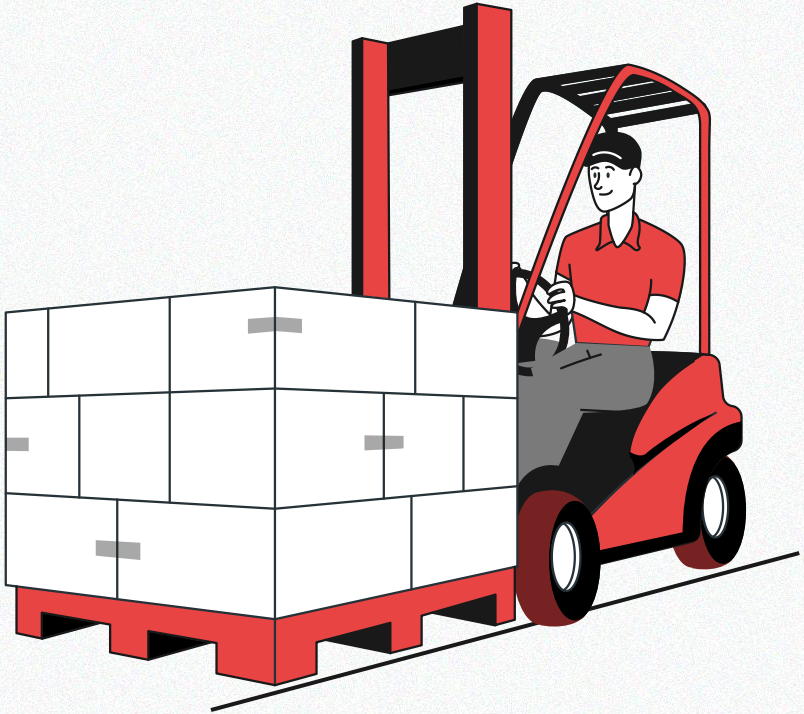


PEAK HOURS:

- Sales hit global maximum between **10:00 AM and 12:00 PM**, defining the golden window for marketing outreach

(RFM)

CUSTOMER SEGMENTATION METHODOLOGY



FEATURE ENGINEERING:

FRAMEWORK BASED ON RFM ANALYSIS

Recency (R)

- Days since the last purchase.

Frequency (F)

- Total number of unique orders.

Monetary (M)

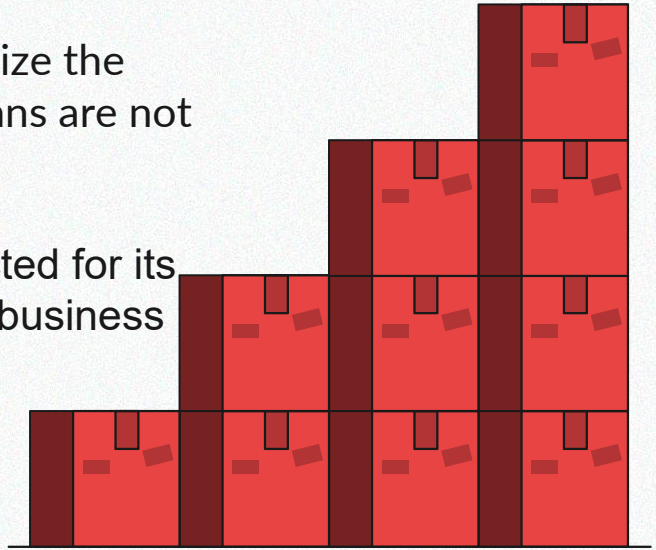
Total spending per customer.

FEATURE ENGINEERING:

FRAMEWORK BASED ON RFM ANALYSIS

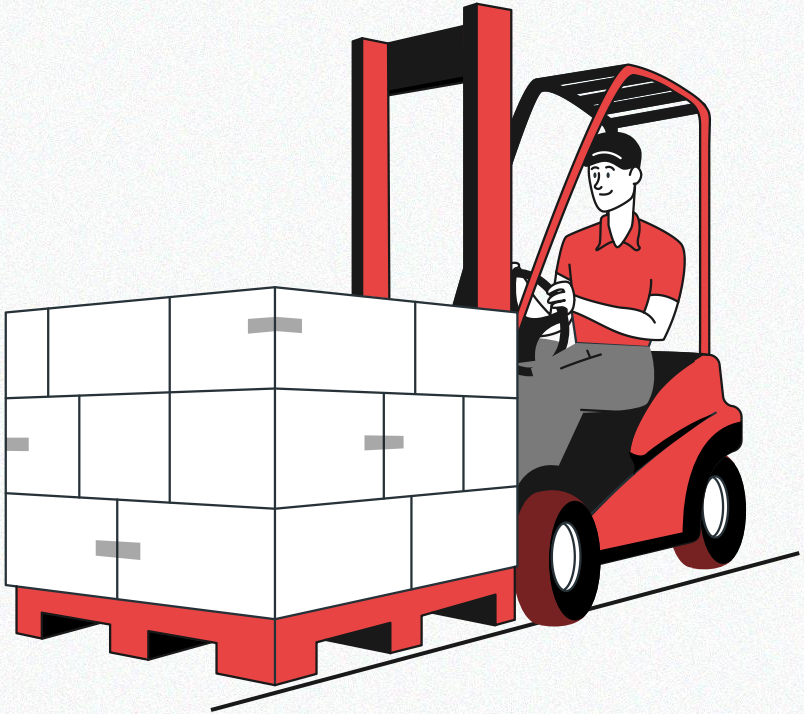
•**Preprocessing:** Applied **StandardScaler** to normalize the RFM metrics, ensuring distance calculations in K-Means are not skewed by different scales.

•**Algorithm Choice: K-Means Clustering** was selected for its high scalability on large transactional data and clear business interpretability.

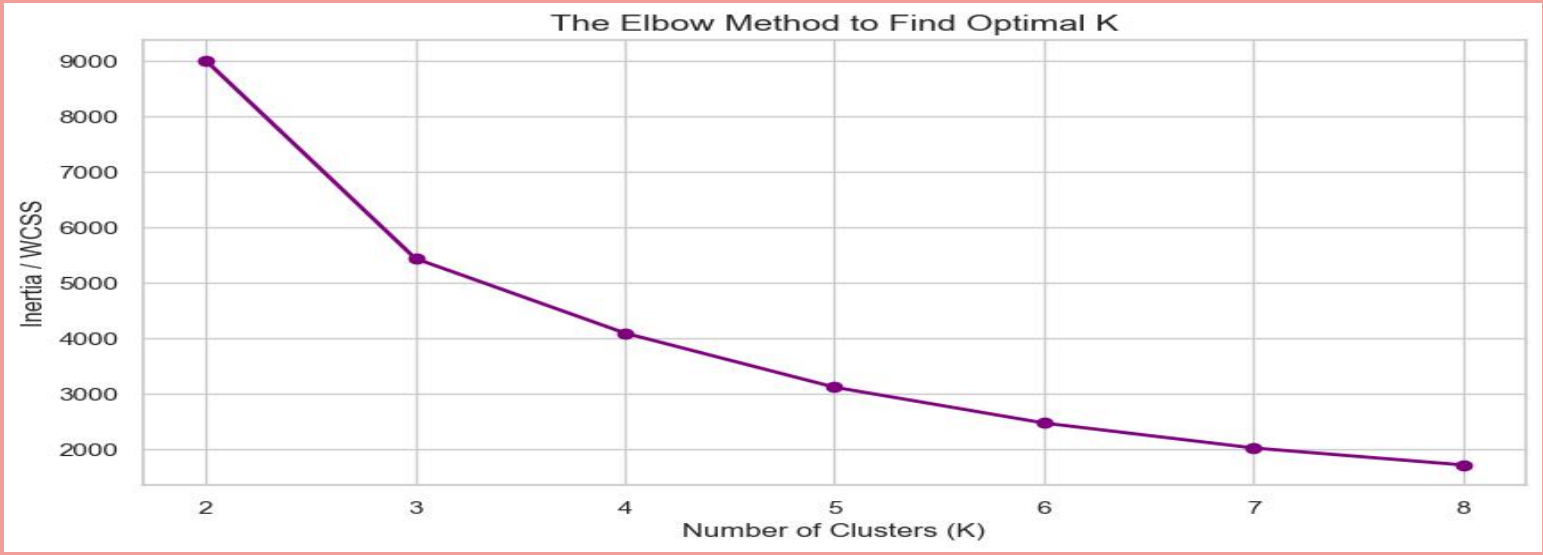


THE ELBOW METHOD

MACHINE LEARNING OPTIMIZATION



THE ELBOW METHOD



- **Optimal K Selection:** Evaluated WCSS (Inertia) from K=2 to K=8.
- **The Elbow Point:** A clear inflection point is observed at **K = 4**, providing the most stable and balanced variance reduction.
- **Stability:** Used n_init=10 to guarantee centroid convergence across multiple random initializations.

THE 4 CUSTOMER SEGMENTS (THE CORE OUTPUT)

01. VIP / CHAMPIONS

Recency: ~7 days | Frequency: ~82 | Monetary: ~£127K. (The highest revenue drivers).

02. LOYAL CUSTOMERS

Recency: ~15 days | Frequency: ~22 | Monetary: ~£12.6K. (Highly frequent, high-trust buyers).

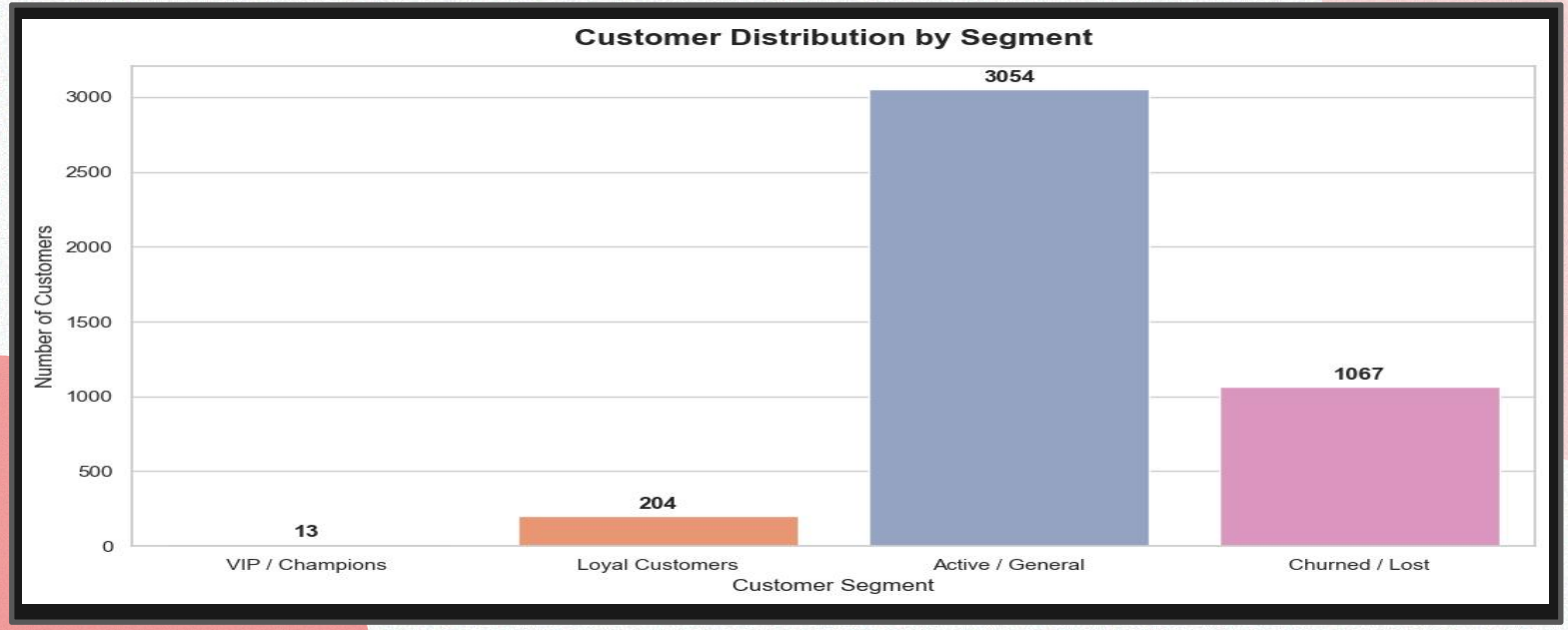
03. ACTIVE / GENERAL MARKET

Recency: ~43 days | Frequency: ~3.6 | Monetary: ~£1.3K. (The core volume/mass market).

04. CHURNED / LOST CUSTOMERS

Recency: ~248 days | Frequency: ~1.5 | Monetary: ~£478. (Inactive for >8 months).

THE 4 CUSTOMER SEGMENTS (THE CORE OUTPUT)



BUSINESS RECOMMENDATIONS & CONCLUSION

VIPS

Implement dedicated account managers and exclusive rewards to prevent churn

LOYAL BUYERS

Earths isLaunch cross-selling campaigns for new inventory arrivals. the Sun’s mass

GENERAL MARKET

Push seasonal discounts and targeted newsletters during peak hours (10 AM - 3 PM) to increase transaction frequency

SUPPLY CHAIN

Secure peak inventory for Hero products before the massive Q4 seasonal surge.

THANKS!

Do you have any questions?

Abdullah