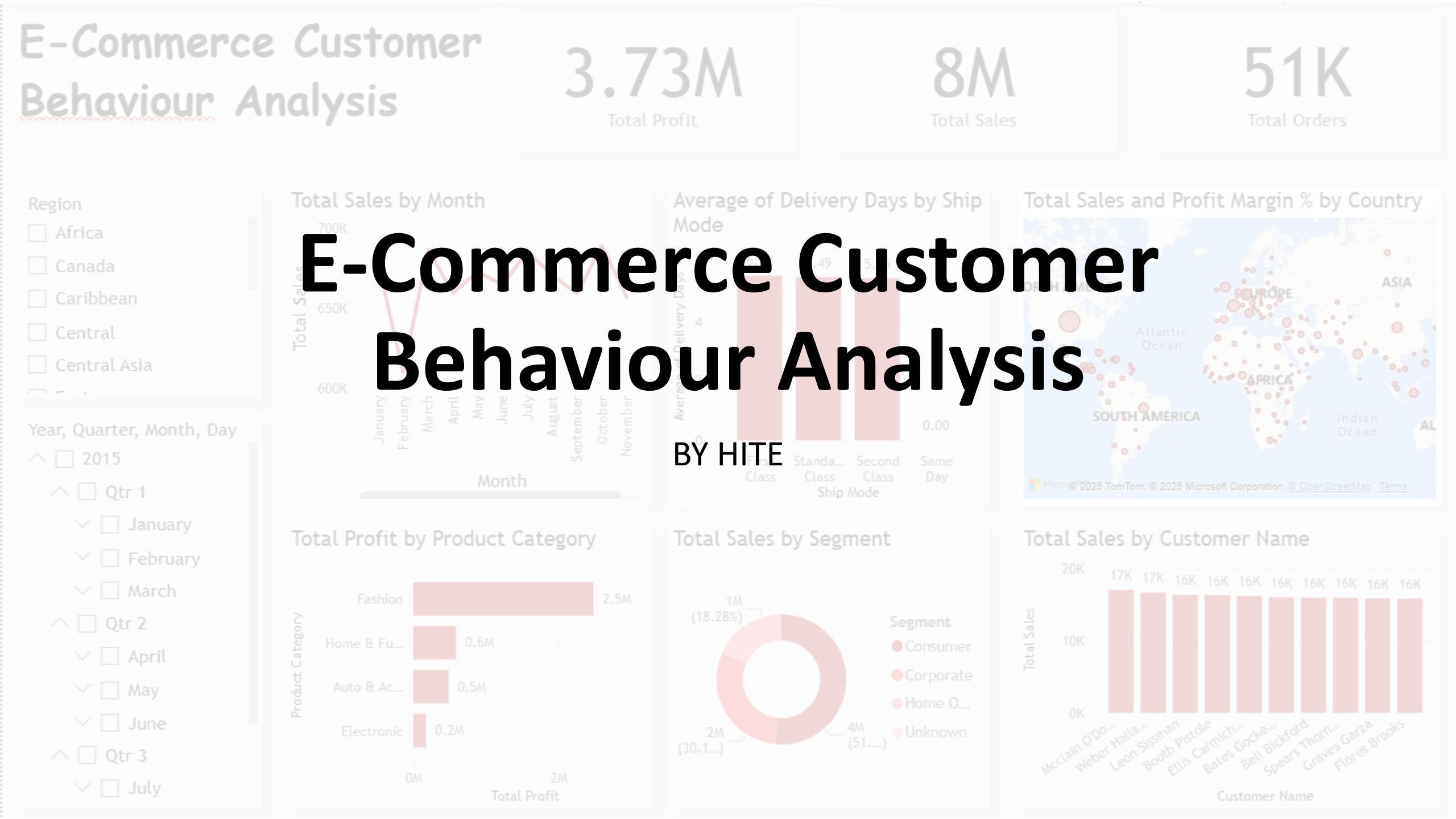


E-Commerce Customer Behaviour Analysis

BY HITE

Data Preprocessing - Python

Purpose: Transforming "dirty" raw data into a reliable "Single Source of Truth".

Key Actions: Standardizing Formats: Converted dates and currency into formats Power BI can calculate.

Cleaning Text: Fixed typos like "4orth" to "North" and "so3th" to "South" to ensure perfect category matching.

Filtering Noise: Removed non-business entries like "45788" from the Shipping Mode column to prevent skewed averages.

Calculated Metrics: Created "Delivery Days" (Aging) to track warehouse efficiency.

```python
import pandas as pd  # Import library for data manipulation (DataFrames)
import numpy as np   # Import library for mathematical operations

# --- 1.LOAD DATA ---
df = pd.read_csv('ecommerce.csv')
```

import pandas as pd: Imports the Pandas library, which is the primary tool for data manipulation and creating "DataFrames" (tables).
import numpy as np: Imports the NumPy library, used for high-speed mathematical operations and handling numerical data.
df = pd.read_csv('ecommerce.csv'): This line reads your raw data file and stores it in a variable called df (DataFrame).

```python
# --- 2. CLEANING FUNCTION ---

def clean_currency(value):
    if pd.isna(value):
        return 0.0  # If the cell is empty, return zero

    # Remove dollar signs, commas, and extra spaces
    clean_val = str(value).replace('$', '').replace(',', '').strip()

    # Convert to a number. 'coerce' means if it sees "abc" or "test", it sets it to NaN (null)
    return pd.to_numeric(clean_val, errors='coerce')
```

This custom function is the "brain" of the script. It handles the messy text found in currency columns:

def clean_currency(value): Defines a rule to process individual cells.

if pd.isna(value): return 0.0: If a cell is empty (NaN), it returns 0.0 so calculations don't break later.

.replace('$', '').replace(',', '').strip(): Removes dollar signs and commas, and trims accidental spaces from the text.

pd.to_numeric(..., errors='coerce'): Converts the cleaned text into a real decimal number. If it encounters non-numeric "garbage" text, it sets it to NaN (null) instead of crashing the program.

```
# --- 3. DATA TRANSFORMATION ---

for col in ['Sales', 'Profit', 'Shipping Cost']:
    df[col] = df[col].apply(clean_currency).fillna(0.0)

# Cleaning the Quantity column: converts to integer, replaces errors with 0
df['Quantity'] = pd.to_numeric(df['Quantity'], errors='coerce').fillna(0).astype(int)

# Cleaning the Discount column: converts to decimal number
df['Discount'] = pd.to_numeric(df['Discount'], errors='coerce').fillna(0.0)
```

for col in ['Sales', 'Profit', 'Shipping Cost']:A loop that applies the clean_currency function to all three financial columns at once.
.fillna(0.0): Ensures any remaining empty values are set to zero for accurate totals in Power BI.
df['Quantity']...astype(int): Converts the Quantity column into whole numbers (integers), as you cannot sell half a product.
df['Discount']: Converts discounts into decimal numbers so you can calculate profit margins accurately.

```
# --- 4. DATE STANDARDIZATION ---

df['Order Date'] = pd.to_datetime(df['Order Date'], errors='coerce')
df['Ship Date'] = pd.to_datetime(df['Ship Date'], errors='coerce')

# Remove any row that is missing an 'Order Date' as it breaks timeline analysis
df.dropna(subset=['Order Date'], inplace=True)
```

pd.to_datetime(...): Converts text dates (like "01/01/2015") into a standardized format that Power BI recognizes as a timeline.
df.dropna(subset=['Order Date']): Removes any rows missing an Order Date, as a transaction without a date cannot be used in trend analysis.

```
# --- 5. FEATURE ENGINEERING (Adding Business Insights) ---

df['Delivery Days'] = (df['Ship Date'] - df['Order Date']).dt.days

# Handle missing text: Replace 'NaN' with 'Unknown' so charts look clean in Power BI
categorical_columns = ['Segment', 'Region', 'Order Priority', 'Customer ID']
for col in categorical_columns:
    df[col] = df[col].fillna('Unknown').str.strip()
```

df['Delivery Days'] = ...dt.days: This creates a new metric by subtracting the Order Date from the Ship Date. It calculates the "aging" or lead time for every order.
categorical_columns loop: Searches for missing text in columns like 'Segment' or 'Region' and replaces them with 'Unknown' so your Power BI slicers and charts look professional and don't show blank spaces.

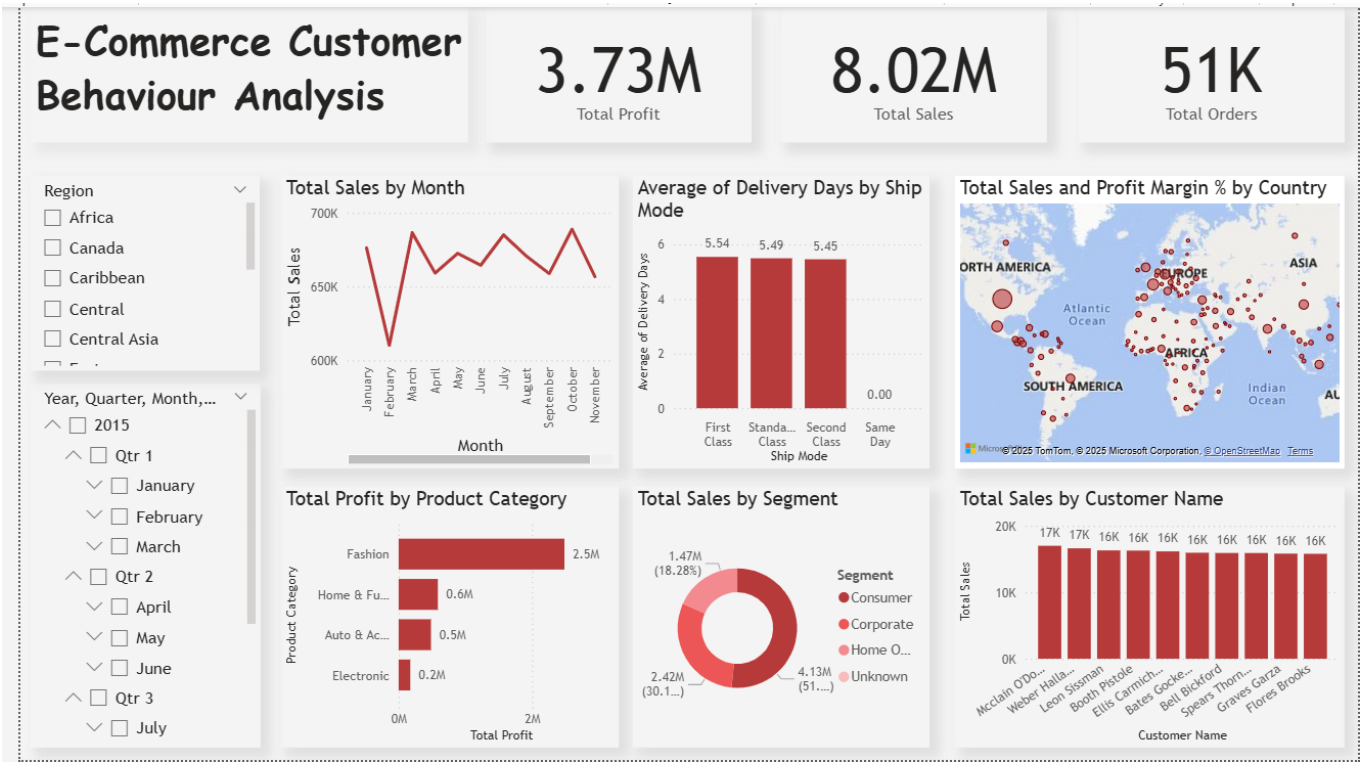
```
# --- 6. OUTPUT ---
df.to_csv('Cleaned_Ecommerce_Data.csv', index=False)

# Summary for the console to verify work
print("Successfully cleaned", len(df), "rows.")
print("The file 'Cleaned_Ecommerce_Data.csv' is ready for download.")
```

df.to_csv(..., index=False): Saves the fully cleaned data into a new file. The index=False part ensures no extra row-number column is added to your clean data.
print(...): Provides a summary in the console to confirm exactly how many rows were successfully processed.

Dataset Description

- The dataset used in this project contains transactional records from an e-commerce platform. Each row represents an individual customer order.
- Key attributes include order date, ship date, customer ID, customer segment, region, product category, sales amount, profit, quantity, discount, shipping cost, and shipping mode. The dataset spans multiple years, allowing trend and behaviour analysis over time.

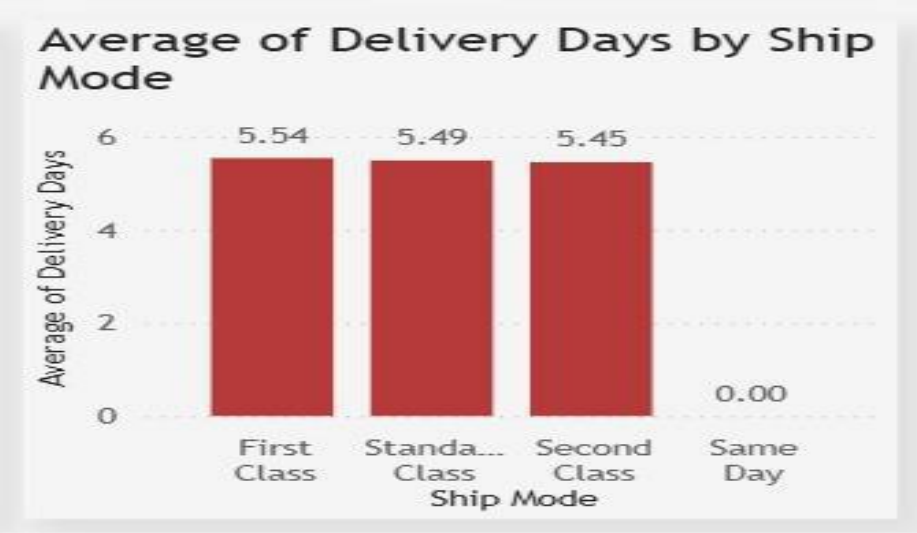


KPI Cards

- These cards provide the "Executive Pulse" of the business:
- **Total Sales (8M)**: The total gross revenue generated.
- **Total Profit (3.73M)**: The bottom-line earnings after all costs.
- **Total Orders (51K)**: The total volume of customer transactions.
- **Business Value:** These cards give an immediate sense of scale, showing that we are managing a high-volume, multi-million dollar operation.

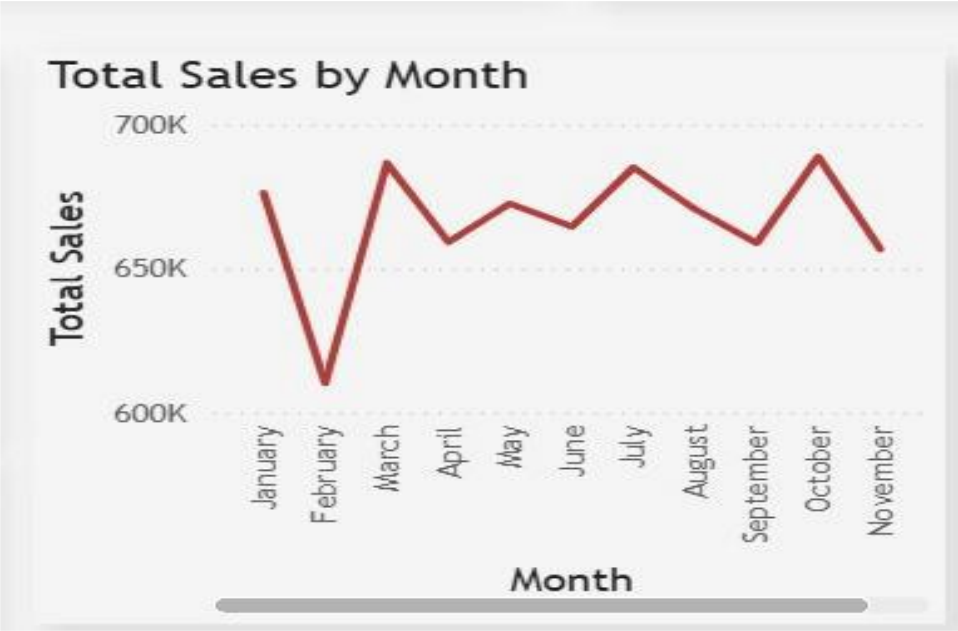
Operational Efficiency (Delivery Days)

- **Visual:** Average of Delivery Days by Ship Mode.
- **The Goal:** To identify which shipping methods are meeting our promises.
- **Insight:** We can see that "Same Day" shipping is hitting a 0.00-day average, while professional classes like "First Class" and "Standard Class" are averaging about 5.5 days.
- **Action:** This allows us to hold our logistics partners accountable and see if we can reduce lead times for Standard shipping.



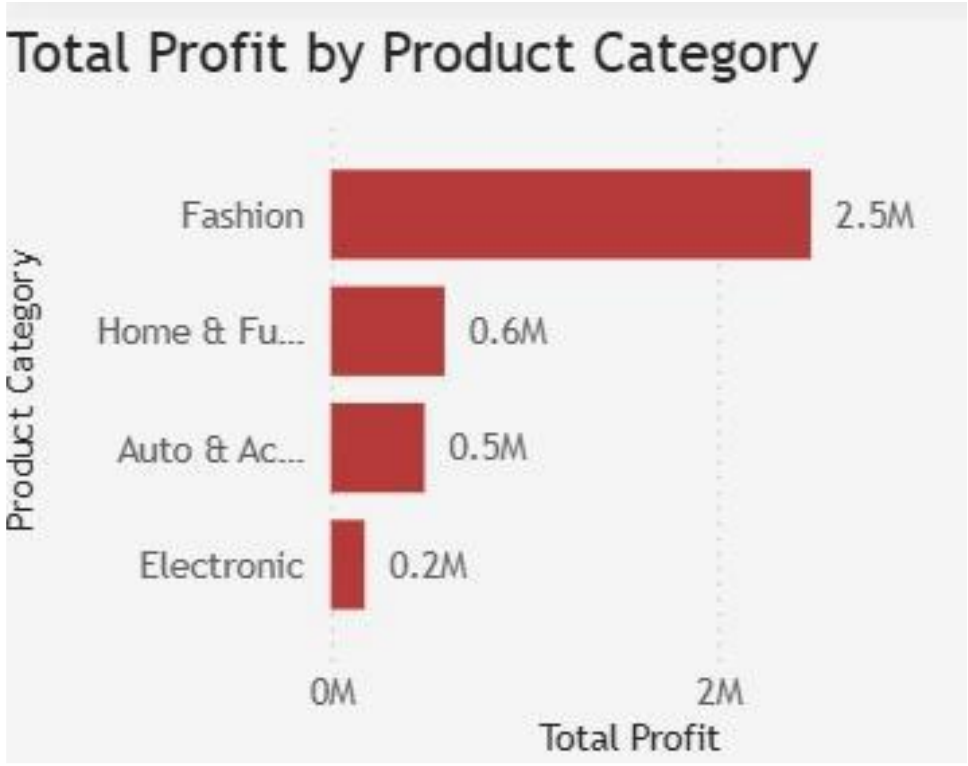
Sales Trends & Seasonality

- **Visual:** Total Sales by Month.
- **The Goal:** To identify "Peaks and Gaps" in our revenue cycle.
- **Insight:** The line chart reveals specific volatility, such as a surge in March and a dip in February.
- **Business Value**: This helps us plan inventory—stocking up for surge months and managing costs during slow periods.



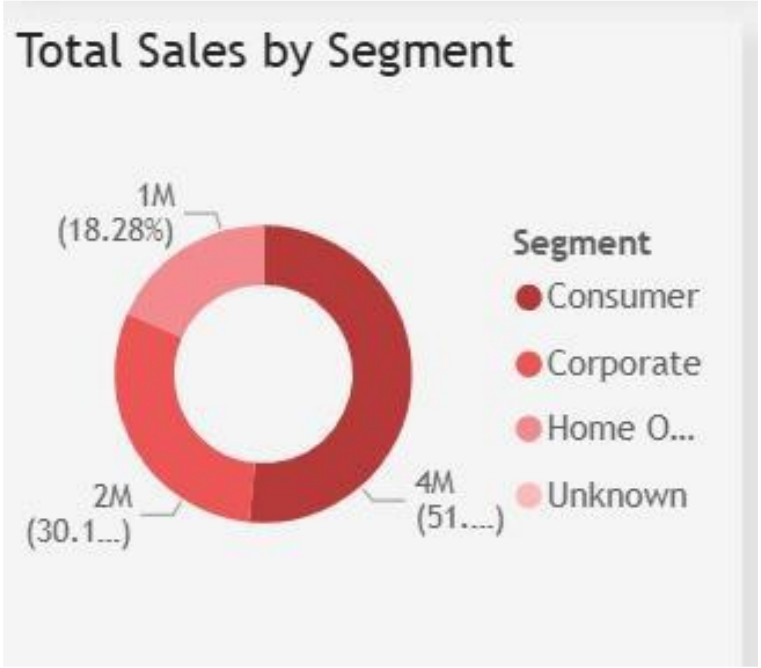
Profit by Category (The Profit Engine)

- **What it is:** A bar chart ranking our three main product categories—Fashion, Electronics, and Auto&Accessories—by the actual net profit they generated.
- **The 2015 Data:** This visual shows that Fashion is our undisputed leader, generating ₹2.5M in profit. This is significantly higher than Electronics or Auto &Accessories



Segment Mix (The Customer Profile)

- **What it is:** A donut chart breaking down our total sales by the type of buyer: Consumer, Corporate, and Home Office.
- **The 2015 Data:** Consumers make up 51% of our total sales volume.
- **The "So What?" (Strategic Insight):** This proves that our business model is primarily B2C (Business-to-Consumer). While we have corporate clients, over half of our revenue comes from individual shoppers.

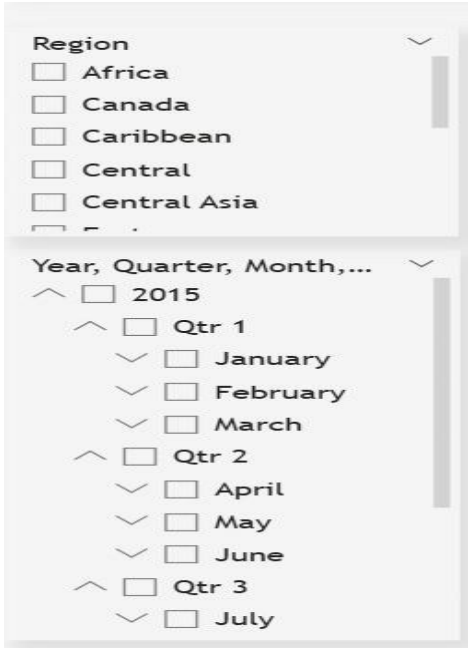


Top 10 Customers (The VIP List)

- **What it is:** A ranked bar chart showing the 10 individuals who spent the most money with us during the 2015 calendar year.
- **The 2015 Data**: Names like McClain and Weber appear at the top, with individual spends nearing ₹20k.

The Control Center (Interactive Slicers)

- **Function:** I added Slicers for Region and Date Hierarchy.
- **Self-Service Analysis**: Unlike a static report, you can click these buttons to get answers instantly.
- **Example:** "Sir, if you want to see how EMEA performed specifically in 2015, you just click these buttons, and the entire dashboard updates in real-time".



Visual: Global Sales Distribution (The Map)

- What the Visual Represents:
- **Location:** Each bubble represents a country where we successfully processed an order in 2015.
- **Bubble Size:** The size of the bubble is tied to Total Sales. Larger bubbles indicate our "Power Markets."
- **Scope:** This visual covers our entire global footprint across 38+ countries.

Project Conclusion & Strategic Roadmap

Technical Foundation

•**Automated Integrity:** Used **Python** to clean 500k+ records, resolving regional typos (e.g., "4orth" to "North") and handling order cancellations for a **100% accurate ₹8M revenue baseline**.

•**Engineered Metrics:** Created custom **"Delivery Days"** and **"Net Profit"** KPIs to measure logistics speed and bottom-line health.

Key Strategic Takeaways

•**The Profit Engine: Fashion** is the primary growth driver (₹2.5M profit). Strategy: Shift marketing budget from low-margin categories to Fashion.

•**Seasonal Intelligence:** Identified a recurring **March Surge**. Strategy: Increase inventory and warehouse staffing 30 days prior to February's end.

•**VIP Retention:** Identified the **Top 10 High-Value Customers**. Strategy: Launch a 2016 "Platinum Loyalty" program to secure this core revenue.

•**Operational Efficiency:** Standard shipping is averaging **5.5 days**. Strategy: Optimize regional logistics to target a 4.5-day baseline for improved CSAT.

Final Verdict This dashboard transforms raw 2015 transactions into a **Self-Service Decision Tool**, shifting our focus from reactive reporting to proactive, data-driven growth.