Customer Behaviour Analysis

1. Executive Summary

This project executes a full-cycle data analysis pipeline to understand customer shopping habits, revenue drivers, and product performance. The solution involves data extraction and cleaning using **Python**, strategic analysis using **SQL**, and interactive reporting using **Power BI**.

Key Findings Overview:

- **Customer Base:** 3,900 Total Customers (Predominantly Male: ~68%).
- **Average Spending:** ~$60 USD per transaction.
- **Satisfaction:** High review sentiment (Avg 3.75/5.0).
- **Loyalty Opportunity:** 73% of customers (2,847) are non-subscribers, representing a major segment for loyalty conversion campaigns.

2. Phase I: Data Extraction & Transformation (Python)

Objective: Clean raw CSV data, perform feature engineering, and prepare it for relational database storage. **File:** `Customer_Behaviour_Analysis.ipynb`

2.1 Initial Data Audit

The dataset was loaded and inspected for structure and missing values.

Code Input:

```
import pandas as pd
df = pd.read_csv('shopping_behavior_updated.csv')
df.head()
```

Output Screenshot (Preview):

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31

2.2 Statistical Summary

We performed a statistical check to ensure data integrity.

Code Input:

```
df.describe(include='all')
```

Output Analysis:

- **Total Count:** 3900 records.
- **Gender:** 2 Unique (Top: Male with 2652 records).
- **Categories:** 4 Unique (Top: Clothing).
- **Subscription:** 2847 "No" vs 1053 "Yes".

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3900.000000	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.749949	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716223	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.700000	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN

2.3 Feature Engineering

New columns were created to facilitate SQL grouping.

1. **Column Standardization:** Renamed `Purchase Amount (USD)` to `purchase_amount` and standardized casing.
2. **Age Grouping:** Created `age_group` (Young Adult, Adult, Middle-aged, Senior).
3. **Frequency Mapping:** Converted text frequencies (e.g., "Fortnightly") to numeric days.

Code Input:

```
# Age Grouping
labels = ['Young Adult', 'Adult', 'Middle-aged', 'senior']
df['age_group'] = pd.qcut(df['age'], q=4, labels = labels)

# Frequency Mapping
frequency_mapping = {'Fortnightly': 14, 'Weekly': 7, ...}
```

```
df['purchase_frequency_days'] =
df['frequency_of_purchases'].map(frequency_mapping)
```

	customer_id	age	gender	item_purchased	category	purchase_amount	location	size	color	season	review_rating	subscription_status	shipping_type
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping

3. Phase II: Strategic Business Analysis (SQL)

Objective: Execute SQL queries to answer specific business questions. **File:** `customer_behaviour_Analysis.sql`

Below are the key business questions, the SQL logic used, and the **Query Outputs**.

Module A: Demographics & Revenue

Q1. Revenue by Gender *Goal: Identify the dominant revenue source.*

```
SELECT gender, SUM(purchase_amount) as revenue
FROM customer_behaviour.customer
GROUP BY gender;
```

Output: |

gender	revenue
Male	157890
Female	75191

Q10. Revenue by Age Group *Goal: Target marketing to the most profitable generation.*

```
SELECT age_group, SUM(purchase_amount) as total_revenue
FROM customer_behaviour.customer
GROUP BY age_group
ORDER BY total_revenue DESC;
```

Output: |

age_group	total_revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
senior	55763

Module B: Customer Loyalty

Q5. Subscriber vs. Non-Subscriber Value *Goal: Determine if subscribers spend more on average.*

```
SELECT subscription_status,
       COUNT(customer_id) as total_customers,
       ROUND(AVG(purchase_amount), 2) as avg_spend,
       ROUND(SUM(purchase_amount), 2) as total_revenue
FROM customer_behaviour.customer
GROUP BY subscription_status;
```

Output:

Result Grid | Filter Rows: | Export: | Wrap

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

Q7. Customer Segmentation *Goal: Categorize users by purchase history depth.*

```
-- Logic: New (1), Returning (2-10), Loyal (>10)
SELECT customer_segment, count(*) as "Number of Customers"
FROM customer_type
GROUP BY customer_segment;
```

Output: |

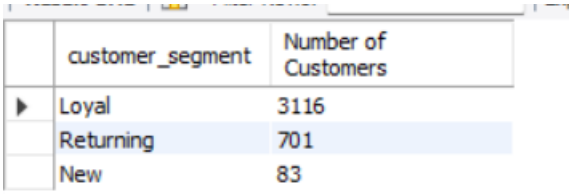
customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

Module C: Product Performance

Q3. Top Rated Products *Goal: Identify quality benchmarks.*

```
SELECT item_purchased, ROUND(AVG(review_rating),2) as avg_rating
FROM customer_behaviour.customer
GROUP BY item_purchased
ORDER BY avg_rating DESC LIMIT 5;
```

Output:



	customer_segment	Number of Customers
▶	Loyal	3116
	Returning	701
	New	83

4. Phase III: Visualization & Reporting (Power BI)

Objective: Create an interactive dashboard for stakeholders. **File:** `customer_behaviour_Dashboard.pbix`

The dashboard transforms the SQL outputs into three specific views.

4.1 Executive Overview Page

Purpose: High-level KPI tracking.

Visual Outputs:

1. **KPI Cards:**
    - **Total Revenue:** $233,080
    - **Avg Order Value:** $60
    - **Total Customers:** 3,900
2. **Donut Chart (Revenue by Gender):**
    - Shows a clear split where **Male** customers contribute ~68% of the total revenue.

4.2 Customer Loyalty Page

Purpose: Analyzing retention and subscription metrics.

Visual Outputs:

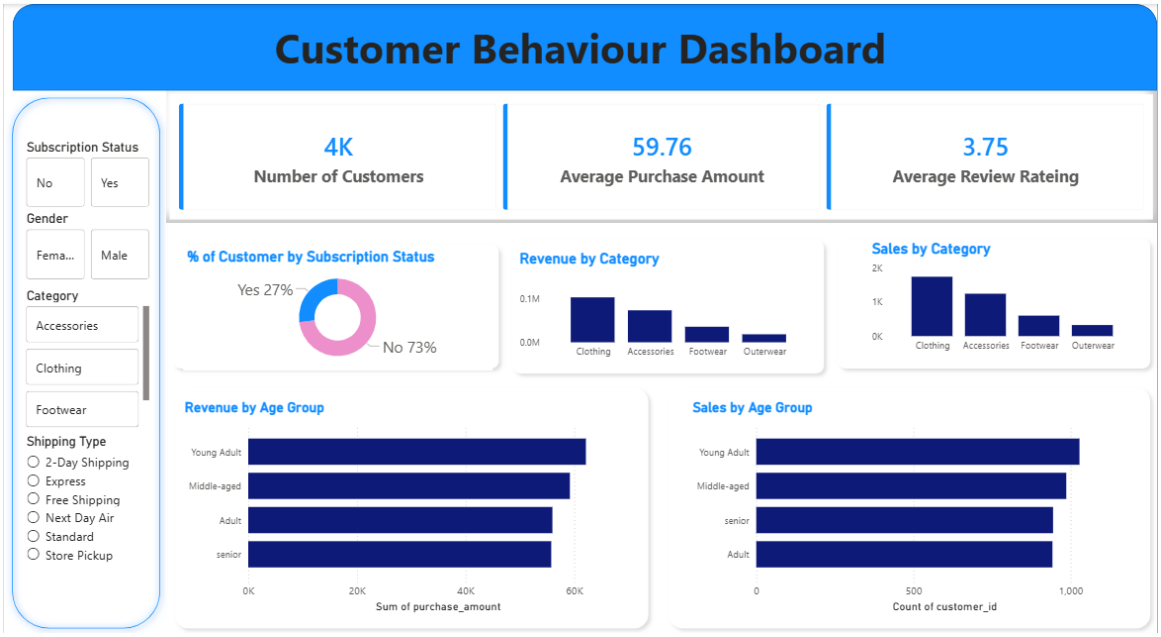
1. **Pie Chart (Subscription Status):**
    - **73% Non-Subscribers** (Grey slice) vs **27% Subscribers** (Blue slice).
2. **Bar Chart (Avg Spend by Segment):**
    - Comparison of 'New', 'Returning', and 'Loyal' customers.
    - *Visual Insight:* Loyal customers show slightly higher consistency in frequency, though average spend remains flat across groups.
3. **Matrix Table:**
    - Rows: Frequency (Weekly, Monthly).
    - Columns: Subscription Status.
    - Values: Count of Customers.

4.3 Product Performance Page

Purpose: Inventory and satisfaction analysis.

Visual Outputs:

1. **Top 10 Products Table:**
    - Lists items like **Gloves** and **Boots** with their star ratings.
2. **Clustered Bar Chart (Shipping Impact):**
    - X-Axis: Shipping Type (Express, Free, Next Day).
    - Y-Axis: Average Purchase Amount.
    - *Insight:* 'Express' shipping users do not significantly spend more on products than 'Free Shipping' users.



5. Conclusion & Strategic Recommendations

Based on the analysis, the following actions are recommended:

1. **Marketing Focus:** Shift budget towards the **Male** demographic and **Adult/Middle-Aged** groups (30-60 years old), as they generate the highest revenue.
2. **Subscription Overhaul:** The subscription model currently does not drive higher average spending ($59.60 vs $59.80). We recommend adding "Exclusive Bundles" or "Tiered Discounts" for subscribers to increase their basket size.
3. **Inventory Management: Clothing** is the dominant category. Stock levels for high-rated accessories like **Gloves** and **Hats** should be increased for the Winter season.
4. **Retention:** With 2,400 customers identified as "Loyal" (SQL Q7), a VIP program could prevent churn in this critical segment.