Shopping Analysis of a Customers — Final Project Report

1. Project Overview

Business Problem Statement

A leading retail company wants to better understand its customers' shopping behavior in order to improve sales, customer satisfaction, and long-term loyalty. The management team has observed changes in purchasing patterns across demographics, product categories, and sales channels (online vs. offline). They are particularly interested in identifying which factors — such as discounts, reviews, seasons, or payment preferences — influence consumer decisions and repeat purchases.

Overarching Business Question

How can the company leverage consumer shopping data to identify trends, improve customer engagement, and optimize marketing and product strategies?

Project Objective

Perform a complete analytics workflow — from data cleaning in Python, to insight extraction using SQL, and visualization in Power BI — to uncover meaningful customer behavior patterns and business opportunities.

Deliverables

1. Data Preparation & Modeling (Python): Clean and transform the raw dataset for analysis.

2. Data Analysis (SQL): Organize data into a structured database, simulate transactions, and extract insights about customer segments, loyalty, and purchase drivers.

3. Visualization & Insights (Power BI): Build an interactive dashboard to highlight key patterns and trends, helping stakeholders make data-driven decisions.

4. Report & Presentation: Summarize findings and actionable recommendations in a business-style report and Power BI presentation.

5. GitHub Repository: Host all Python scripts, SQL queries, and dashboard files in a well-structured repository.

2. Dataset Summary

Records: ~3,900
Columns: 18

Data Categories

Customer details: customer_id, age, gender, location, subscription_status
Transaction data: product_name, category, purchase_amount, review_rating, shipping_type
Behavior metrics: discount_applied, previous_purchases, purchase_frequency, season

Missing Values: 37 missing ratings imputed using median by product category.
Source: Synthetic dataset designed for data analytics learning and business simulation.

Tools Used

Tool	Purpose
Jupyter Notebook (Python)	Data cleaning, preprocessing, and feature creation
PostgreSQL (SQL)	Data modeling, structured analysis, business queries
Excel	Data validation and tabular exports
Power BI	Dashboard visualization and storytelling

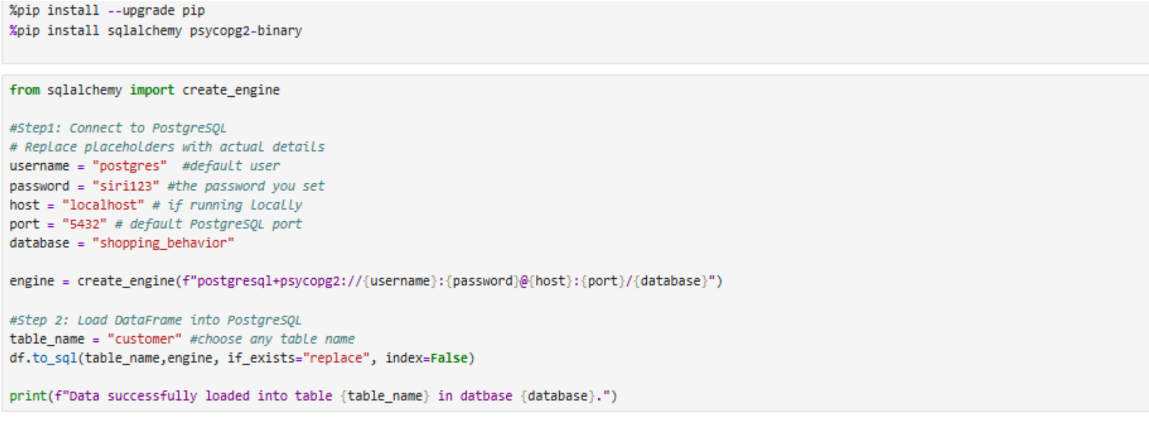
3. Data Cleaning using Python

Performed in Jupyter Notebook using pandas and NumPy.

Steps:
- Renamed inconsistent column names to snake_case.
- Handled missing values using category-level medians.
- Created derived columns such as age_group and purchase_frequency_days.
- Removed duplicates and irrelevant columns.
- Exported final cleaned dataset for SQL upload.

Python to SQL Connectivity

After cleaning, the dataset was pushed to PostgreSQL using SQLAlchemy. Example code:



```
%pip install --upgrade pip
%pip install sqlalchemy psycopg2-binary
```

```
from sqlalchemy import create_engine

#Step1: Connect to PostgreSQL
# Replace placeholders with actual details
username = "postgres"  #default user
password = "siri123" #the password you set
host = "localhost" # if running locally
port = "5432" # default PostgreSQL port
database = "shopping_behavior"

engine = create_engine(f"postgresql+psycopg2://{username}:{password}@{host}:{port}/{database}")

#Step 2: Load DataFrame into PostgreSQL
table_name = "customer" #choose any table name
df.to_sql(table_name,engine, if_exists="replace", index=False)

print(f"Data successfully loaded into table {table_name} in datbase {database}.")
```

Key points:

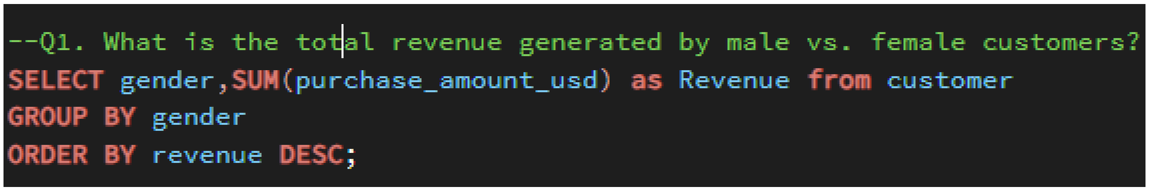
- Install: pip install sqlalchemy psycopg2-binary
- Avoid hardcoding credentials in production.
- For large tables use chunksize and method='multi' for performance.

4. Exploratory Data Analysis (EDA) using SQL

EDA was performed via SQL queries in PostgreSQL to understand customer spending patterns, loyalty behavior and purchase trends.

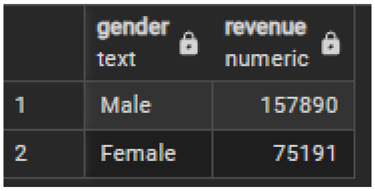
SQL Business Questions and Insights

1. Total Revenue by Gender



```
--Q1. What is the total revenue generated by male vs. female customers?
SELECT gender,SUM(purchase_amount_usd) as Revenue from customer
GROUP BY gender
ORDER BY revenue DESC;
```

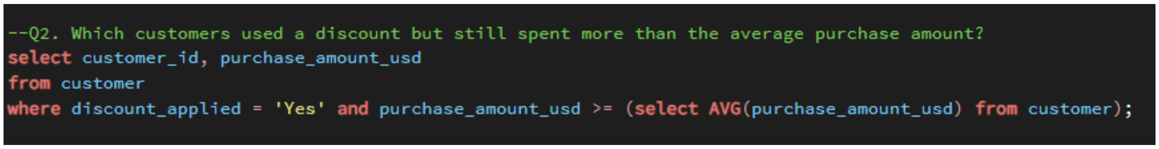
- SQL output here.



	gender text	revenue numeric
1	Male	157890
2	Female	75191

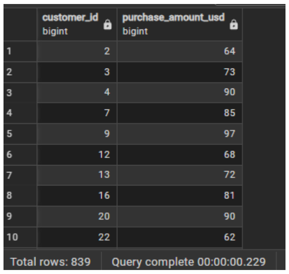
- Male customers generated significantly higher revenue (157,890) compared to females (75,191).
- This shows males contribute a larger share to total sales — suggesting room to improve female engagement.

2. High-Spending Discount Users



```
--Q2. Which customers used a discount but still spent more than the average purchase amount?
select customer_id, purchase_amount_usd
from customer
where discount_applied = 'Yes' and purchase_amount_usd >= (select AVG(purchase_amount_usd) from customer);
```

- SQL output here.

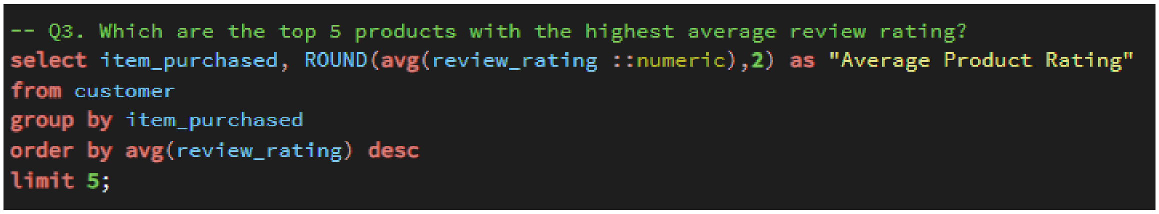


	customer_id bigint	purchase_amount_usd bigint
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62

Total rows: 839 Query complete 00:00:00.229

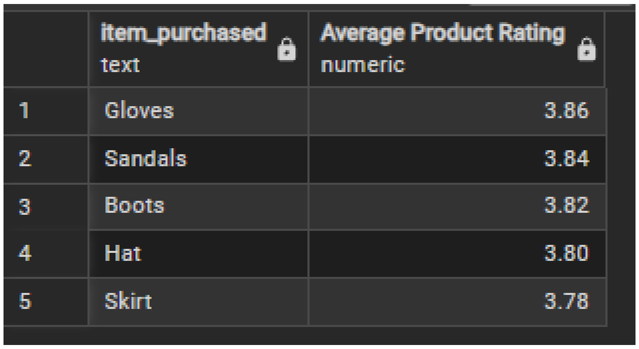
- Customers who availed discounts but spent above the average purchase value show high buying intent despite price sensitivity. These users are valuable deal-seekers — personalized discount offers can help convert them into loyal, repeat buyers.

3. Top 5 Products by Review Rating



```
-- Q3. Which are the top 5 products with the highest average review rating?
select item_purchased, ROUND(avg(review_rating ::numeric),2) as "Average Product Rating"
from customer
group by item_purchased
order by avg(review_rating) desc
limit 5;
```

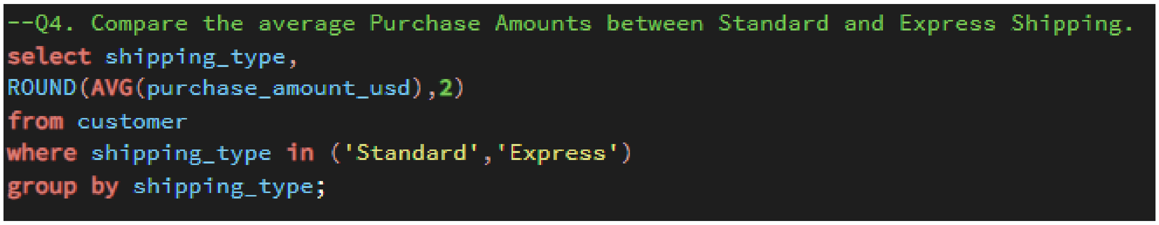
- SQL output here.



	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

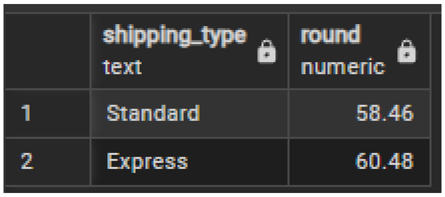
- Products like Gloves, Sandals, and Boots received the highest average ratings (≈3.8–3.9), showing strong customer satisfaction.
- These items can be featured in promotions or marketing campaigns to attract new customers and reinforce product trust.

4. Average Purchase by Shipping Type



```
--Q4. Compare the average Purchase Amounts between Standard and Express Shipping.
select shipping_type,
ROUND(AVG(purchase_amount_usd),2)
from customer
where shipping_type in ('Standard','Express')
group by shipping_type;
```

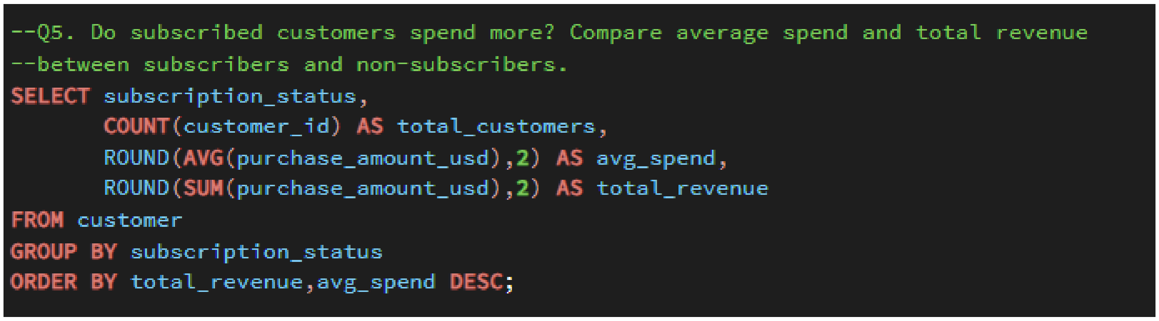
- SQL output here.



	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

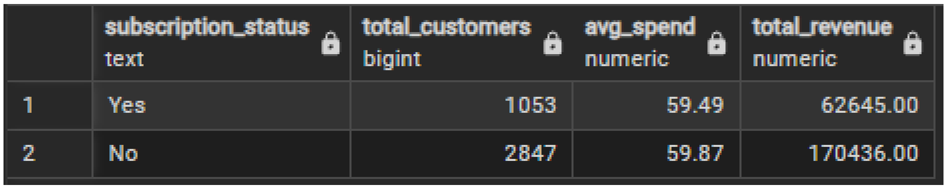
Express shipping customers spent slightly more on average (₹60.48) than Standard shipping users (₹58.46).
This indicates that premium shipping users are higher-value buyers, and offering express shipping incentives could increase revenue

5. Subscriber Spending Analysis



```
--Q5. Do subscribed customers spend more? Compare average spend and total revenue
--between subscribers and non-subscribers.
SELECT subscription_status,
       COUNT(customer_id) AS total_customers,
       ROUND(AVG(purchase_amount_usd),2) AS avg_spend,
       ROUND(SUM(purchase_amount_usd),2) AS total_revenue
FROM customer
GROUP BY subscription_status
ORDER BY total_revenue,avg_spend DESC;
```

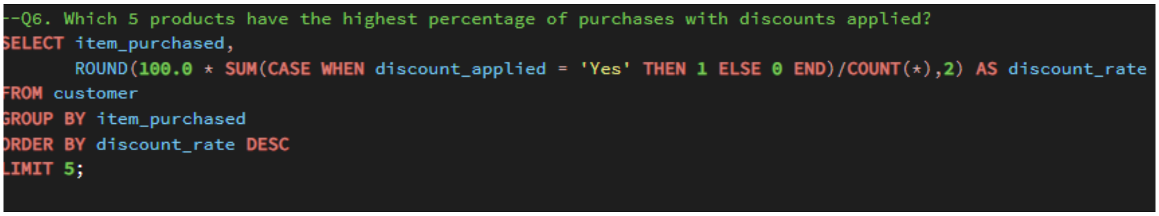
- SQL output here.



	subscription_status text	total_customers bigint	avg_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

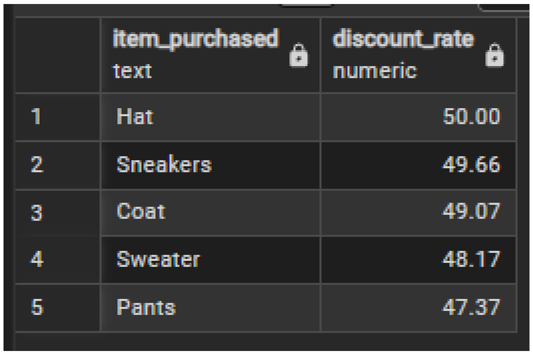
- Subscribers, though fewer in number, contribute a higher total revenue (₹62,645) and maintain a strong average spend.
- This shows that subscription programs boost customer loyalty and spending consistency — worth expanding further.

6. Top 5 Discounted Products



```
--Q6. Which 5 products have the highest percentage of purchases with discounts applied?
SELECT item_purchased,
       ROUND(100.0 * SUM(CASE WHEN discount_applied = 'Yes' THEN 1 ELSE 0 END)/COUNT(*),2) AS discount_rate
FROM customer
GROUP BY item_purchased
ORDER BY discount_rate DESC
LIMIT 5;
```

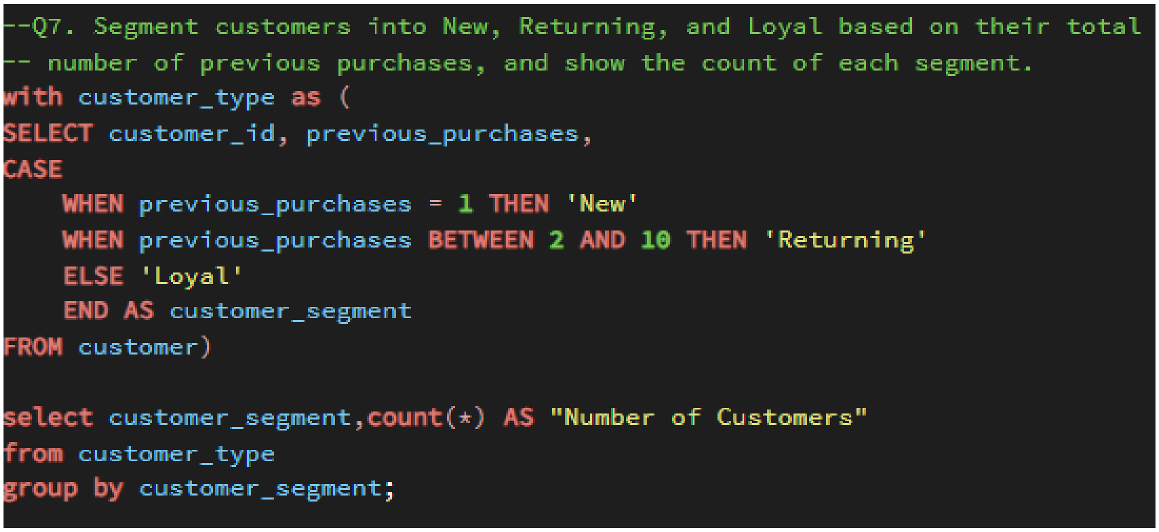
- SQL output here.



	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

- Products like Hats, Sneakers, and Coats have the highest discount rates (≈47–50%), showing they are frequently promoted to drive sales.
- These items likely attract deal-seekers — optimizing discount levels could maintain demand while protecting profit margins.

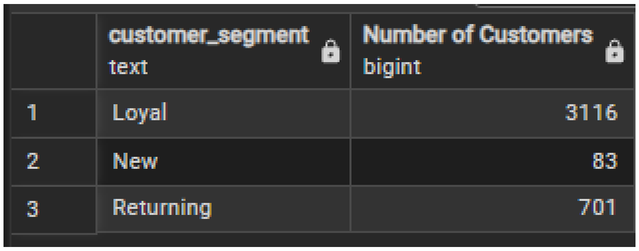
7. Customer Segmentation by Loyalty



```
--Q7. Segment customers into New, Returning, and Loyal based on their total
-- number of previous purchases, and show the count of each segment.
with customer_type as (
SELECT customer_id, previous_purchases,
CASE
    WHEN previous_purchases = 1 THEN 'New'
    WHEN previous_purchases BETWEEN 2 AND 10 THEN 'Returning'
    ELSE 'Loyal'
    END AS customer_segment
FROM customer)

select customer_segment,count(*) AS "Number of Customers"
from customer_type
group by customer_segment;
```

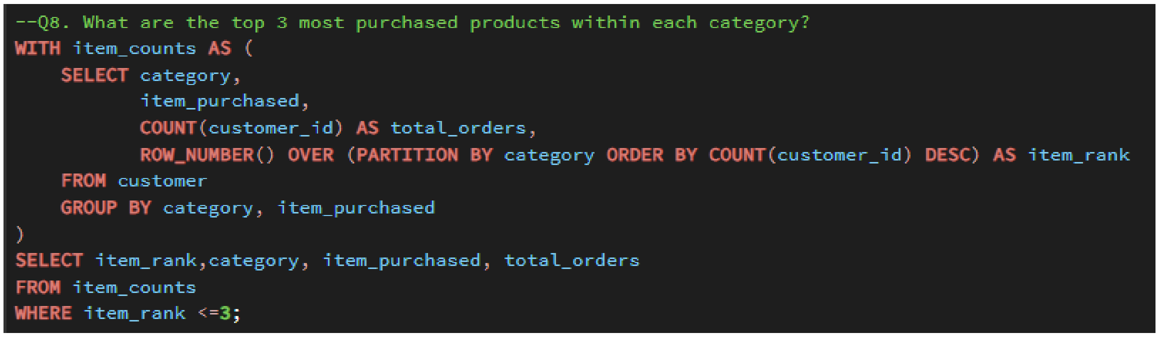
- SQL output here.



	customer_segment text	Number of Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

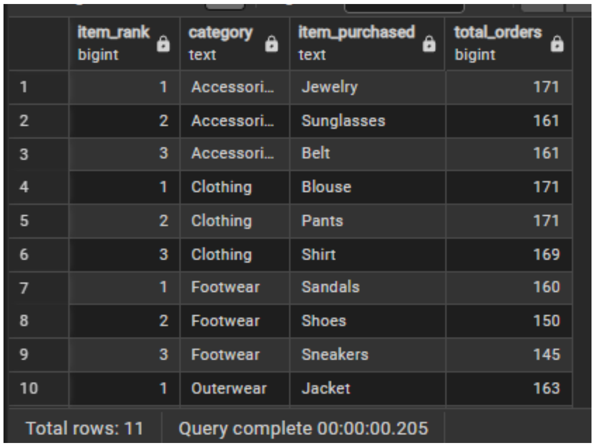
- Loyal customers (3,116) form the largest segment and are the main revenue drivers, while new (83) and returning (701) customers are smaller groups.
- This highlights the need to retain loyal customers and nurture new ones through personalized engagement strategies.

8. Top 3 Products by Category



```sql
--Q8. What are the top 3 most purchased products within each category?
WITH item_counts AS (
    SELECT category,
           item_purchased,
           COUNT(customer_id) AS total_orders,
           ROW_NUMBER() OVER (PARTITION BY category ORDER BY COUNT(customer_id) DESC) AS item_rank
    FROM customer
    GROUP BY category, item_purchased
)
SELECT item_rank,category, item_purchased, total_orders
FROM item_counts
WHERE item_rank <=3;
```

- SQL output here.

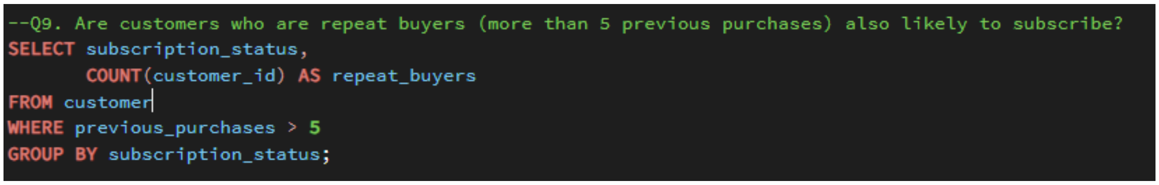


	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	Belt	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163

Total rows: 11 Query complete 00:00:00.205

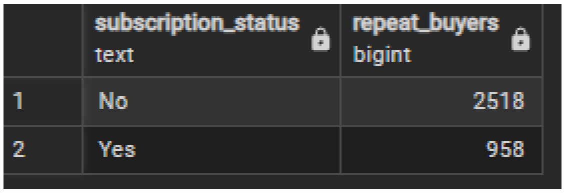
- Across categories, items like Jewelry, Pants, and Sandals appear among the most purchased, showing consistent demand.
- This indicates category concentration, suggesting a focus on cross-selling related products to boost overall sales..

9. Repeat Buyers vs Subscription Correlation



```sql
--Q9. Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?
SELECT subscription_status,
       COUNT(customer_id) AS repeat_buyers
FROM customer
WHERE previous_purchases > 5
GROUP BY subscription_status;
```

- SQL output here.

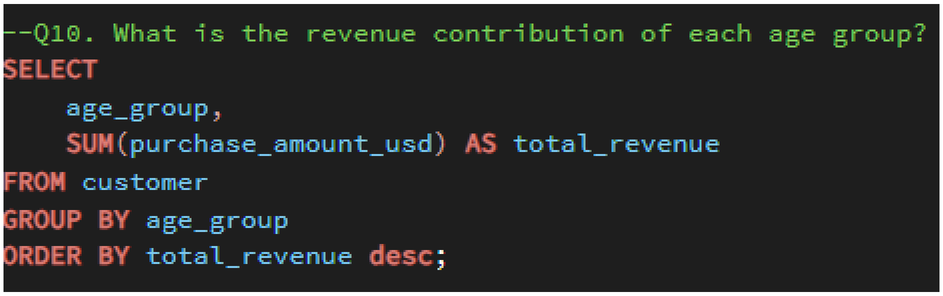


	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

- Among repeat buyers, 958 are subscribed while 2,518 are non-subscribed, showing a partial but positive correlation.

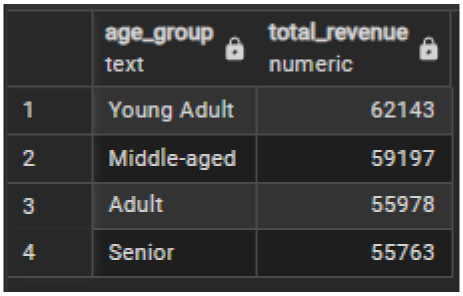
- Encouraging frequent buyers to subscribe can strengthen loyalty and increase long-term retention.

10. Revenue Contribution by Age Group



```
--Q10. What is the revenue contribution of each age group?
SELECT
    age_group,
    SUM(purchase_amount_usd) AS total_revenue
FROM customer
GROUP BY age_group
ORDER BY total_revenue desc;
```

- SQL output here.



	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

- Young Adults generate the highest revenue (₹62,143), followed by Middle-aged and Adult groups.
- This indicates that younger shoppers are the most active buyers, and targeted marketing toward them can further boost sales..

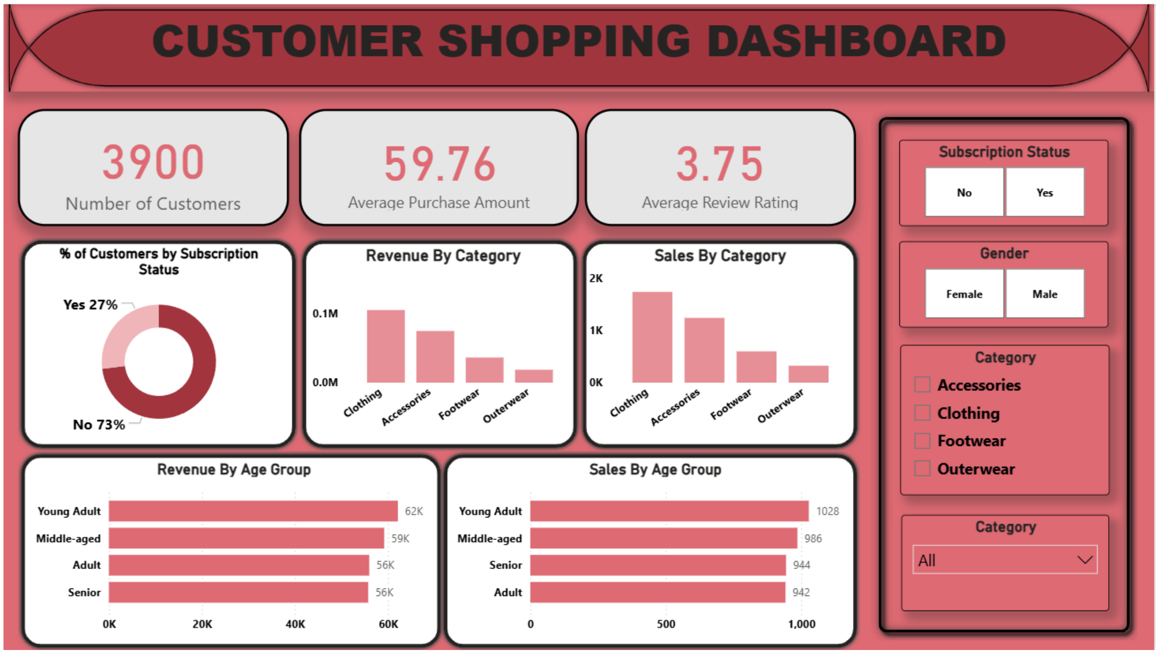
PostgreSQL to Power BI Connectivity

Steps:
1. Open Power BI Desktop → Get Data → PostgreSQL Database.
2. Enter server (host:port) and database name.
3. Choose Import or DirectQuery.
4. Authenticate and select tables or supply native SQL queries/views.
5. Build visuals and publish to Power BI Service.
Tip: Install and configure the Npgsql ADO.NET driver if prompted.

Power BI Dashboard

Created interactive dashboard showcasing:
- Revenue breakdown by gender and region.
- Discount effectiveness and SKU performance.
- Top-rated products and customer loyalty segmentation.

- Slicers for age group, subscription status, and shipping type.



CUSTOMER SHOPPING DASHBOARD

3900 Number of Customers

59.76 Average Purchase Amount

3.75 Average Review Rating

Category: All

Business Recommendations

- Increase subscription campaigns targeted at frequent and high-value buyers.
- Focus discount activity on product categories with healthy margin profiles.
- Promote Express shipping options to customers with higher average order values.
- Expand loyalty programs to reduce churn and increase lifetime value.
- Use review ratings to market high-performing products and improve lower-rated items.

Conclusions

Built an end-to-end analytics pipeline (Python → PostgreSQL → Power BI). Derived actionable insights into customer preferences, loyalty behavior, discount effectiveness, and demographic contributions.

Author

Mukesh Gopi Nandh

Email: mukeshudatha7@gmail.com

LinkedIn: https://linkedin.com/in/mukesh-gopi-nandh

GitHub: https://github.com/Mukeshgn