Sales Insights

E-Commerce customers & transactions dataset

Exploratory data analysis using MySQL

by ks.milda.ks@gmail.com

Table of contents

01 Introduction

02 Number of Orders

03 Average Order Value

04 Total Revenue

05 Customers Demographic

06 Product & Category

07 Churn Rate

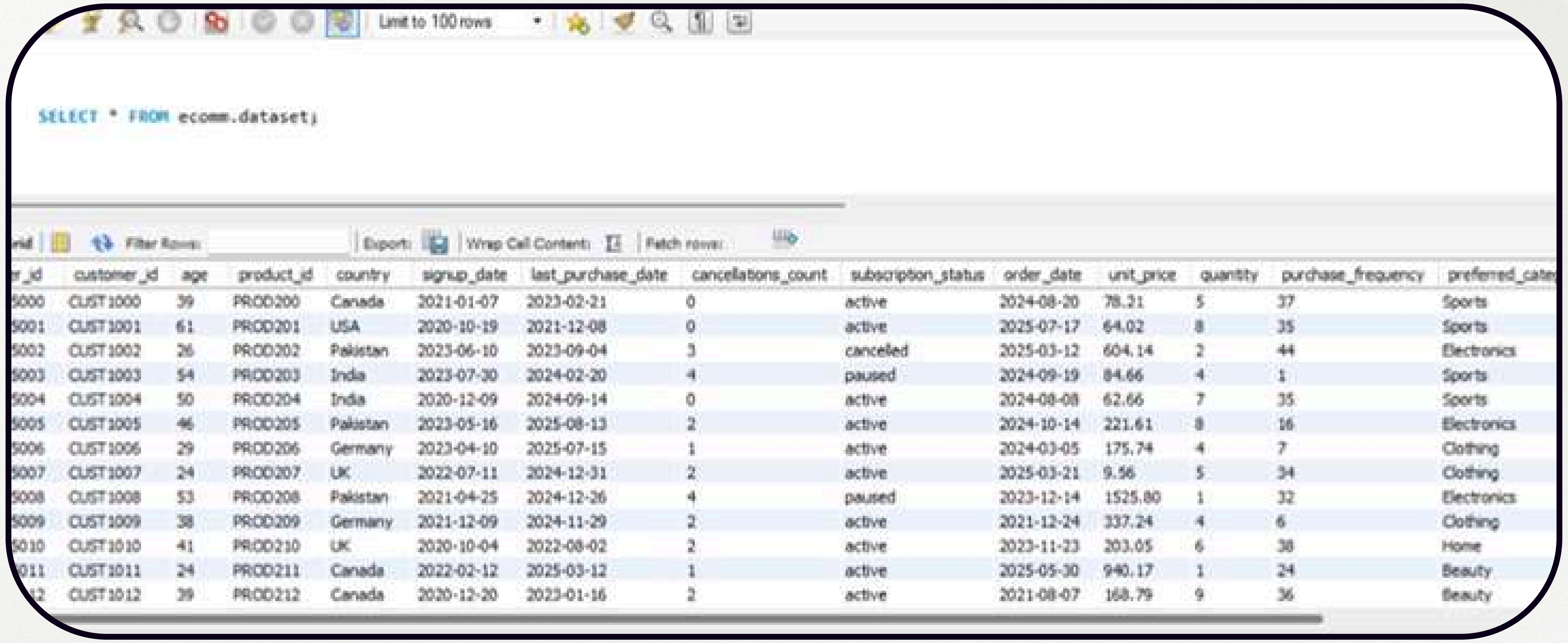
08 Dashboard

09 Conclusion

Dataset Introduction

This dataset contains 2,000 customer-level and order-level records from a simulated e-commerce platform (from kaggle). The customers ranging from several countries, age, gender, and order date.

Before the exploration begins, the dataset is checked and cleaned to ensure there are no missing values nor duplicate rows, as documented in this file. The process include formatting the data type (normalizing) and setting the column order_id as the primary key.



```
SELECT * FROM ecomm.dataset;
```

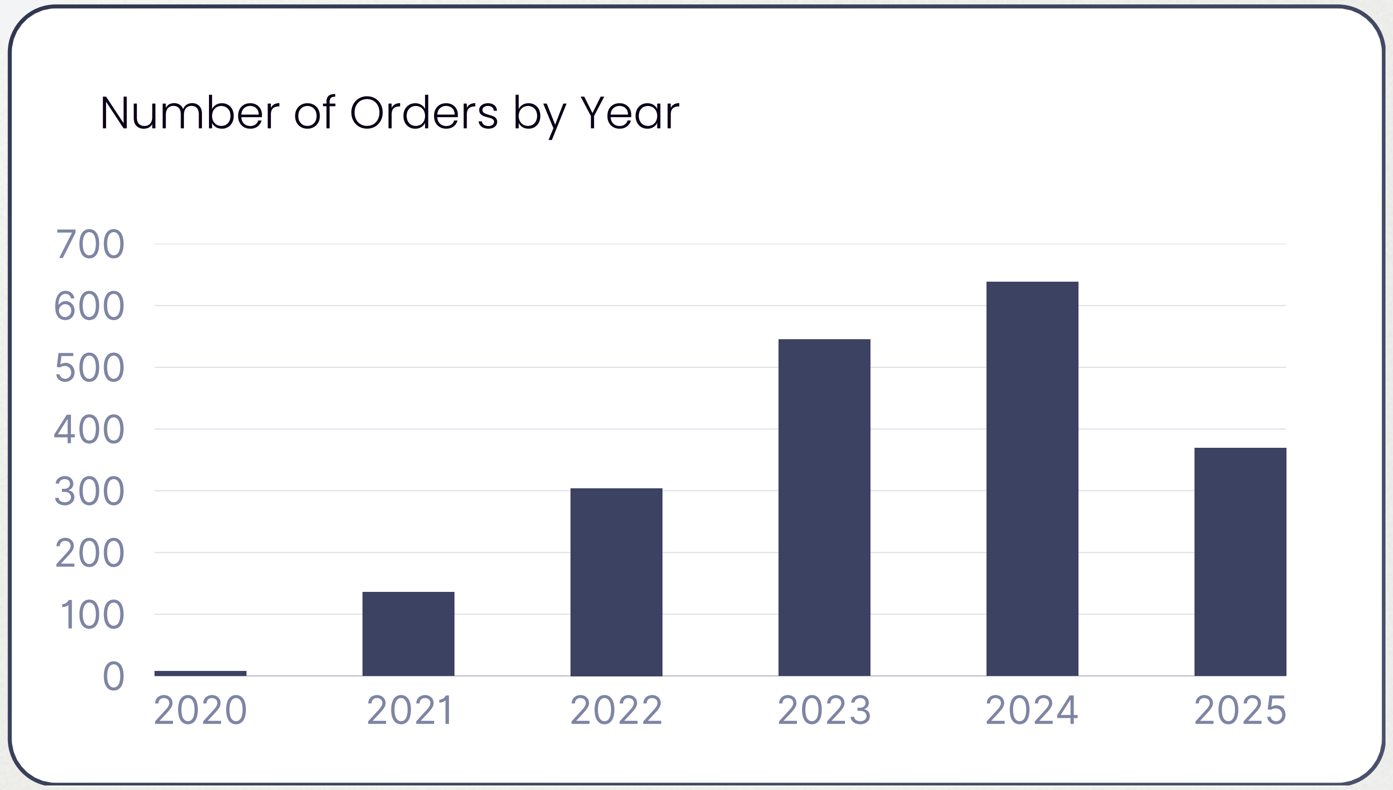
r_id	customer_id	age	product_id	country	signup_date	last_purchase_date	cancellations_count	subscription_status	order_date	unit_price	quantity	purchase_frequency	preferred_cate
5000	CUST1000	39	PROD200	Canada	2021-01-07	2023-02-21	0	active	2024-08-20	78.21	5	37	Sports
5001	CUST1001	61	PROD201	USA	2020-10-19	2021-12-08	0	active	2025-07-17	64.02	8	35	Sports
5002	CUST1002	26	PROD202	Pakistan	2023-06-10	2023-09-04	3	cancelled	2025-03-12	604.14	2	44	Electronics
5003	CUST1003	54	PROD203	India	2023-07-30	2024-02-20	4	paused	2024-09-19	84.66	4	1	Sports
5004	CUST1004	50	PROD204	India	2020-12-09	2024-09-14	0	active	2024-08-08	62.66	7	35	Sports
5005	CUST1005	46	PROD205	Pakistan	2023-05-16	2025-08-13	2	active	2024-10-14	221.61	8	16	Electronics
5006	CUST1006	29	PROD206	Germany	2023-04-10	2025-07-15	1	active	2024-03-05	175.74	4	7	Clothing
5007	CUST1007	24	PROD207	UK	2022-07-11	2024-12-31	2	active	2025-03-21	9.56	5	34	Clothing
5008	CUST1008	53	PROD208	Pakistan	2021-04-25	2024-12-26	4	paused	2023-12-14	1525.80	1	32	Electronics
5009	CUST1009	38	PROD209	Germany	2021-12-09	2024-11-29	2	active	2021-12-24	337.24	4	6	Clothing
5010	CUST1010	41	PROD210	UK	2020-10-04	2022-08-02	2	active	2023-11-23	203.05	6	38	Home
011	CUST1011	24	PROD211	Canada	2022-02-12	2025-03-12	1	active	2025-05-30	940.17	1	24	Beauty
12	CUST1012	39	PROD212	Canada	2020-12-20	2023-01-16	2	active	2021-08-07	168.79	9	36	Beauty

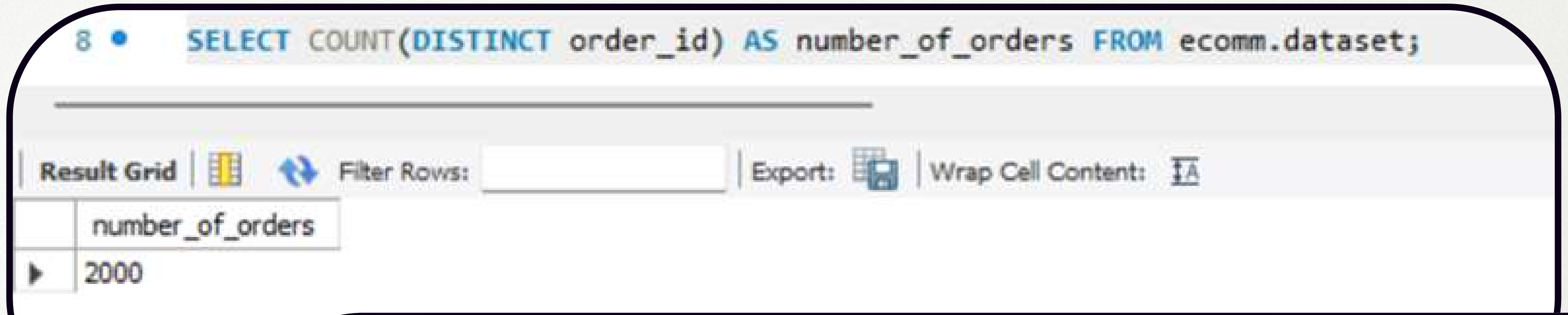
Number of Orders

The number of order were increase every year, which may indicate growing customer engagement.

Total orders

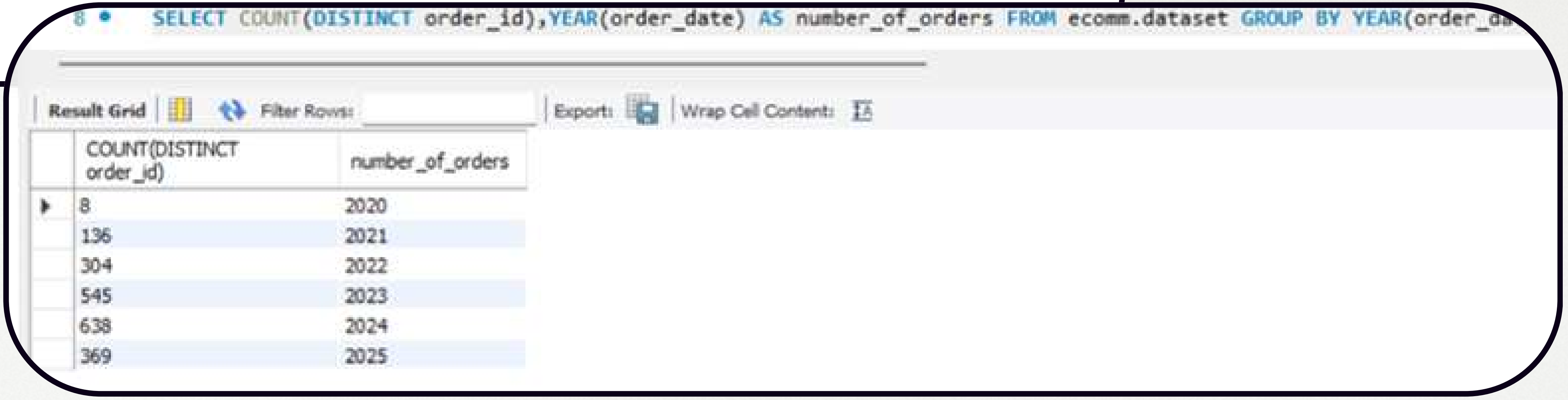
2000





```
SELECT COUNT(DISTINCT order_id) AS number_of_orders FROM ecomm.dataset;
```

number_of_orders
2000



```
SELECT COUNT(DISTINCT order_id),YEAR(order_date) AS number_of_orders FROM ecomm.dataset GROUP BY YEAR(order_da
```

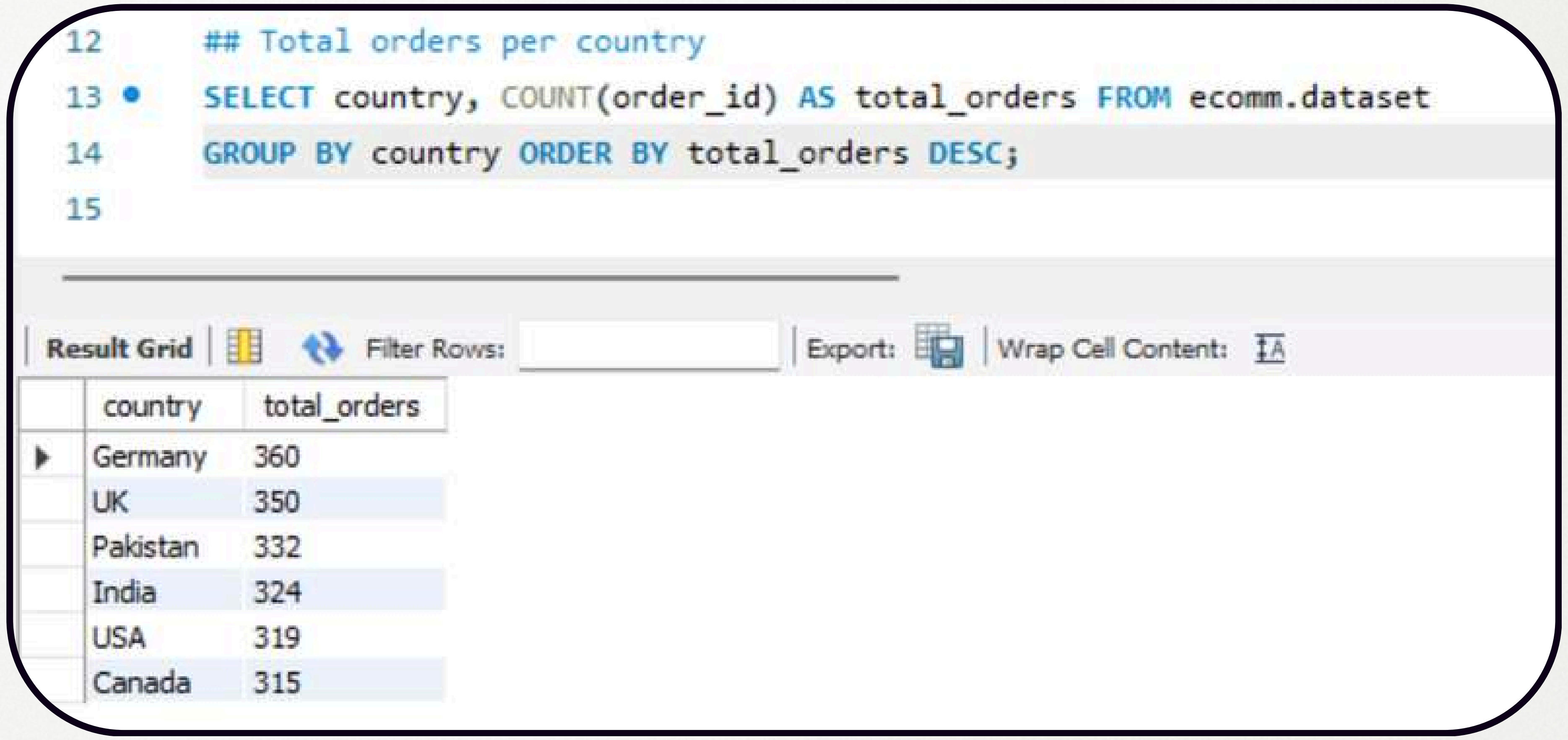
COUNT(DISTINCT order_id)	number_of_orders
8	2020
136	2021
304	2022
545	2023
638	2024
369	2025

Number of Orders (2)

The number of orders based on each country. Picture below showed that country with the most number of orders were Germany with 360 orders.

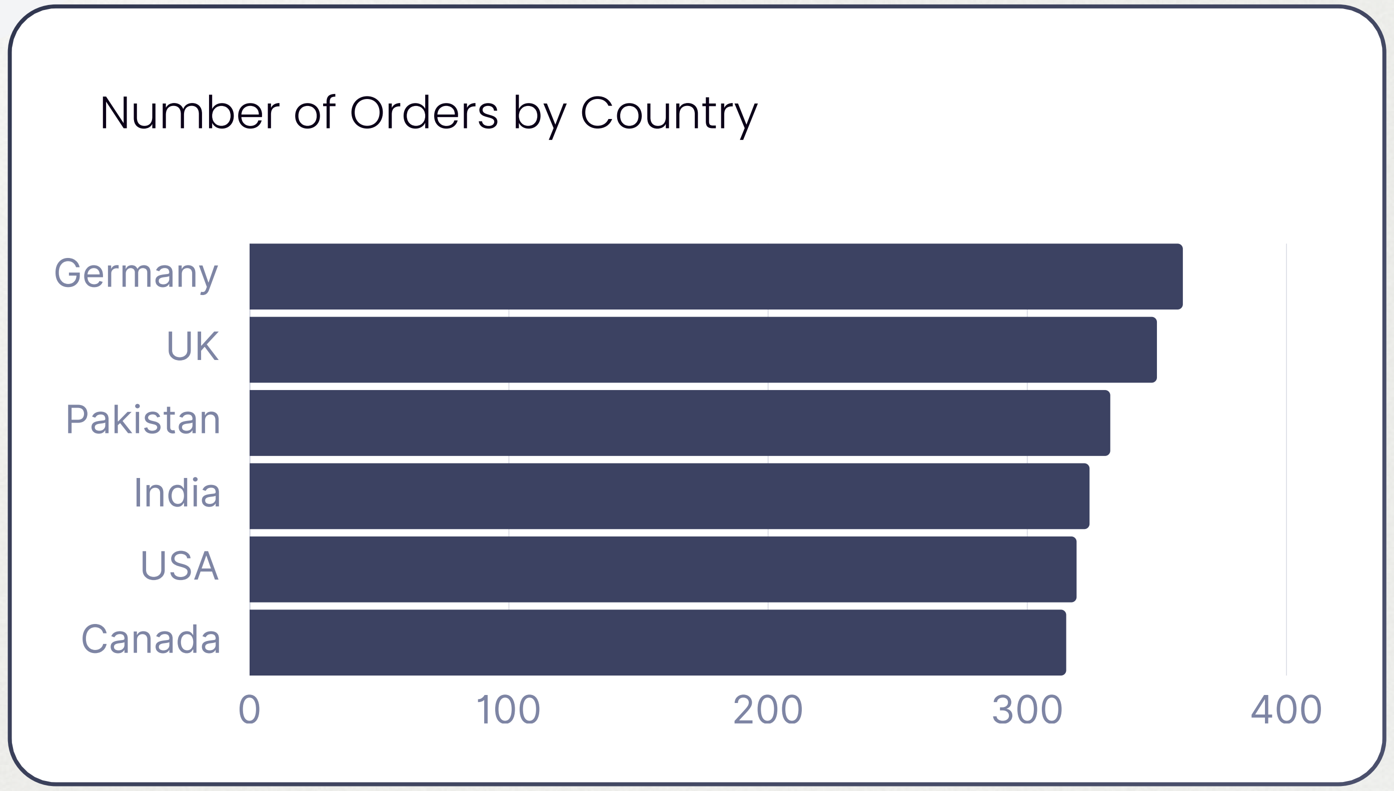
Total orders

2000



```
## Total orders per country
SELECT country, COUNT(order_id) AS total_orders FROM ecomm.dataset
GROUP BY country ORDER BY total_orders DESC;
```

country	total_orders
Germany	360
UK	350
Pakistan	332
India	324
USA	319
Canada	315



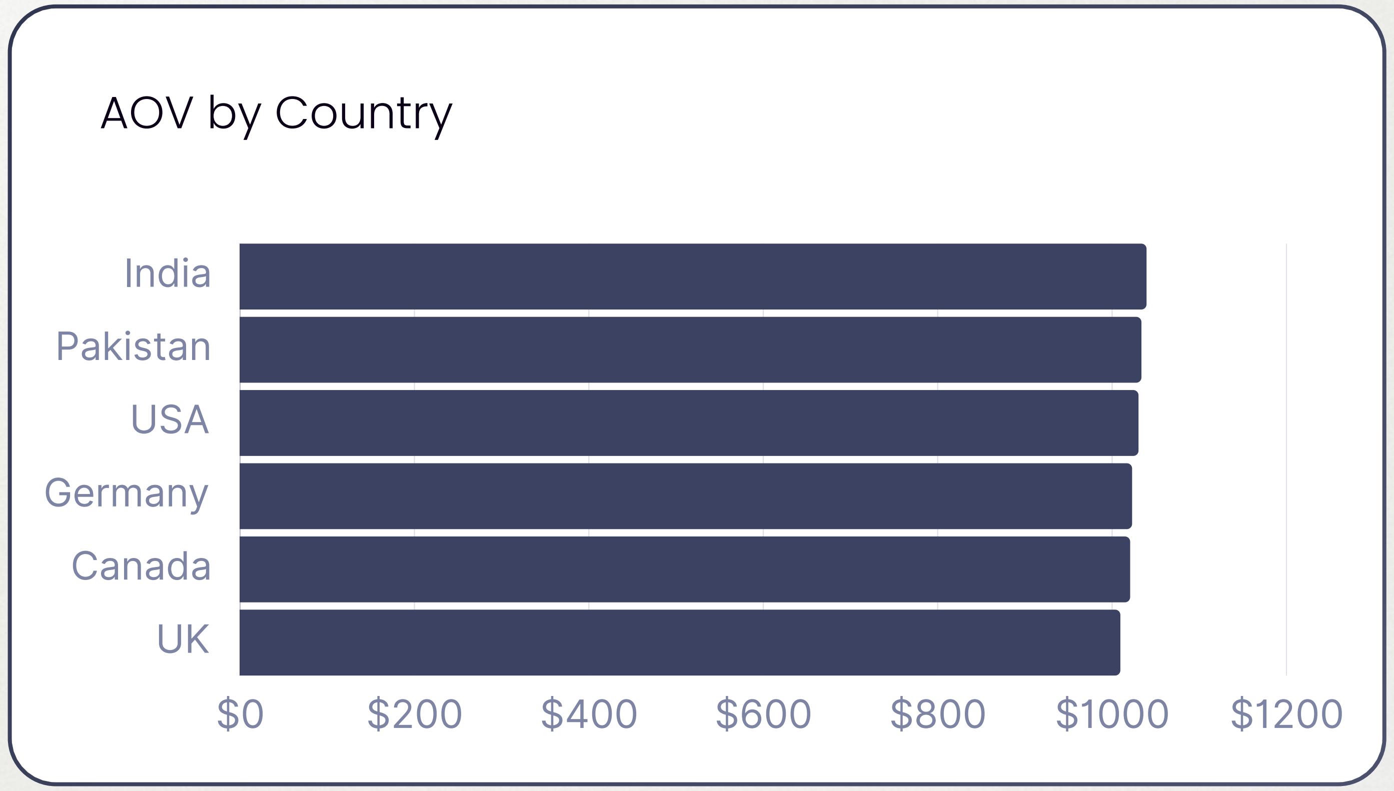
Number of Orders by Country

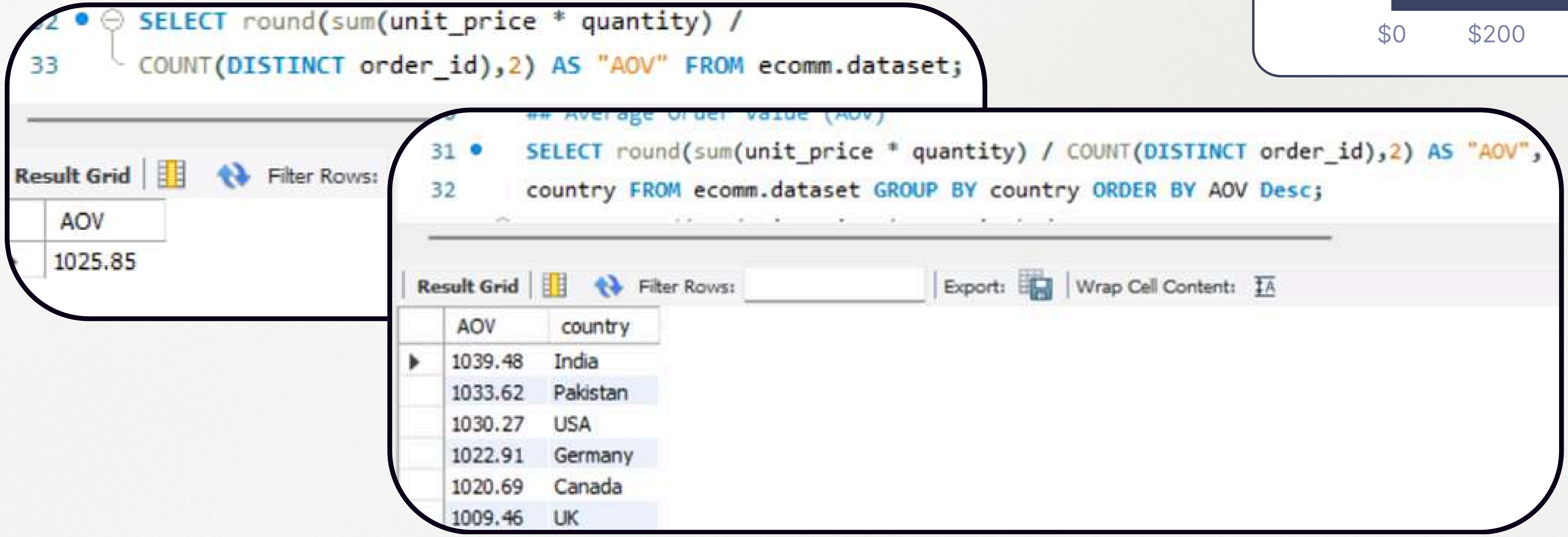
Average Order Value

The AOV for every country means on average each transaction contributes as number below in revenue. Countries like India and Pakistan show higher AOVs, while the UK has lower AOV.

AOV

$1025.85


AOV by Country



```
SELECT round(sum(unit_price * quantity) /
    COUNT(DISTINCT order_id),2) AS "AOV" FROM ecomm.dataset;
```

```
AOV
1025.85
```

```
## Average Order Value (AOV)
SELECT round(sum(unit_price * quantity) / COUNT(DISTINCT order_id),2) AS "AOV",
country FROM ecomm.dataset GROUP BY country ORDER BY AOV Desc;
```

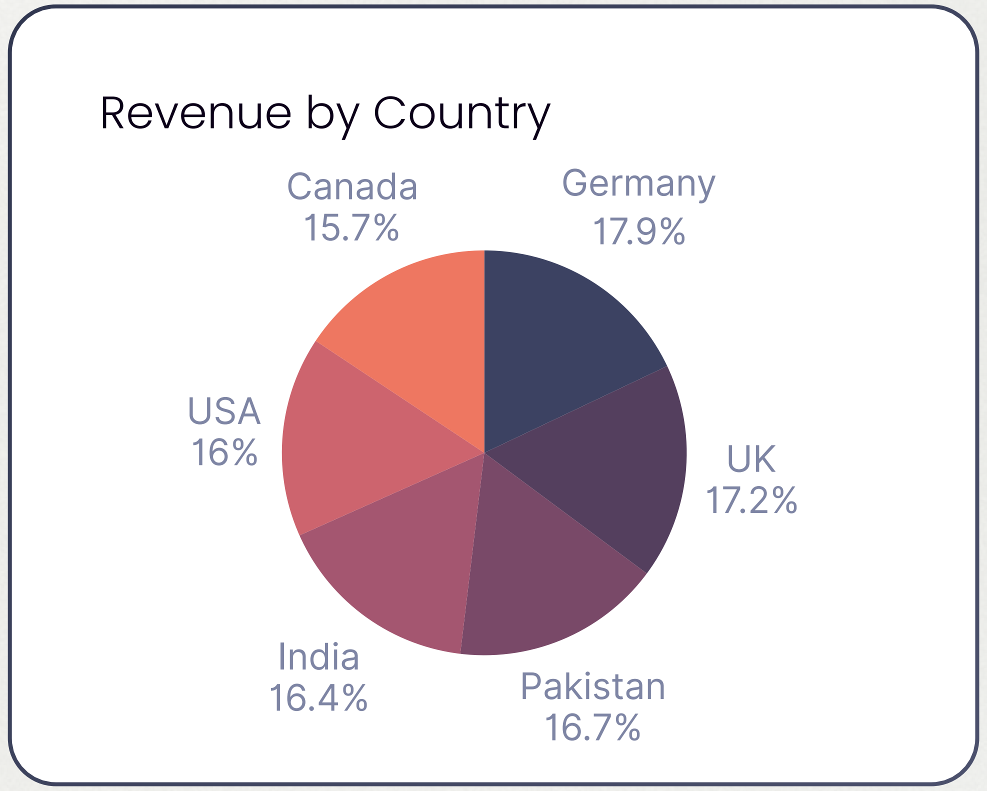
```
AOV      country
1039.48  India
1033.62  Pakistan
1030.27  USA
1022.91  Germany
1020.69  Canada
1009.46  UK
```

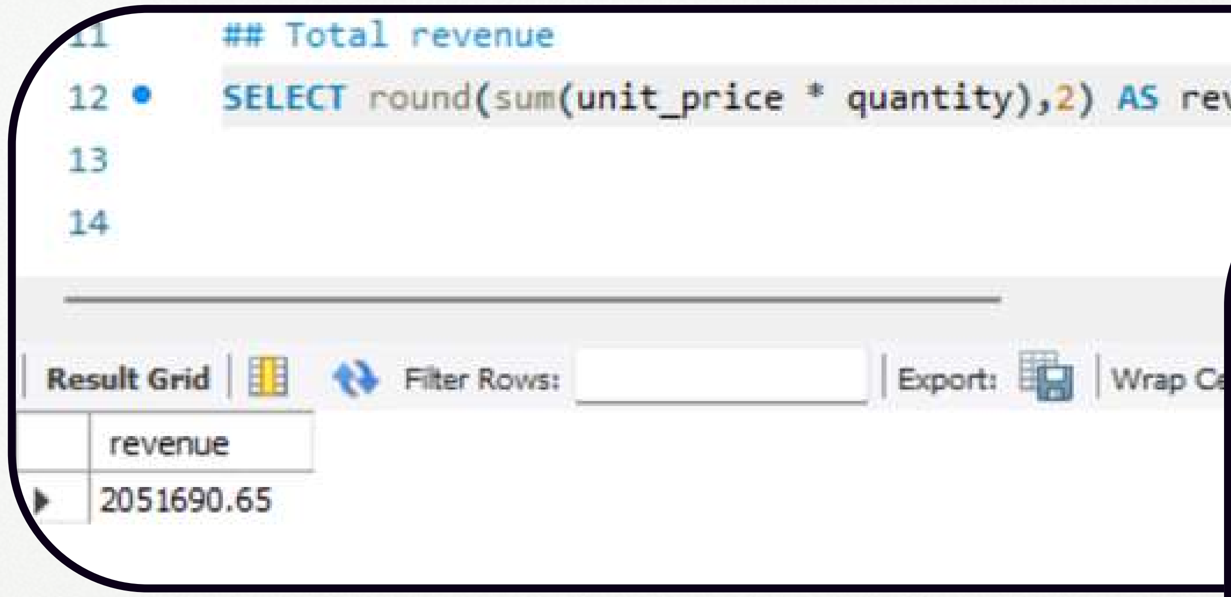
Total Revenue

The total revenue for 5 years based on country as attached below, show that the percentage revenue for each country is relatively balanced, with no single country accounting for more than one-fifth of total sales. This suggests the business has a diversified customer base across markets.

Total Revenue

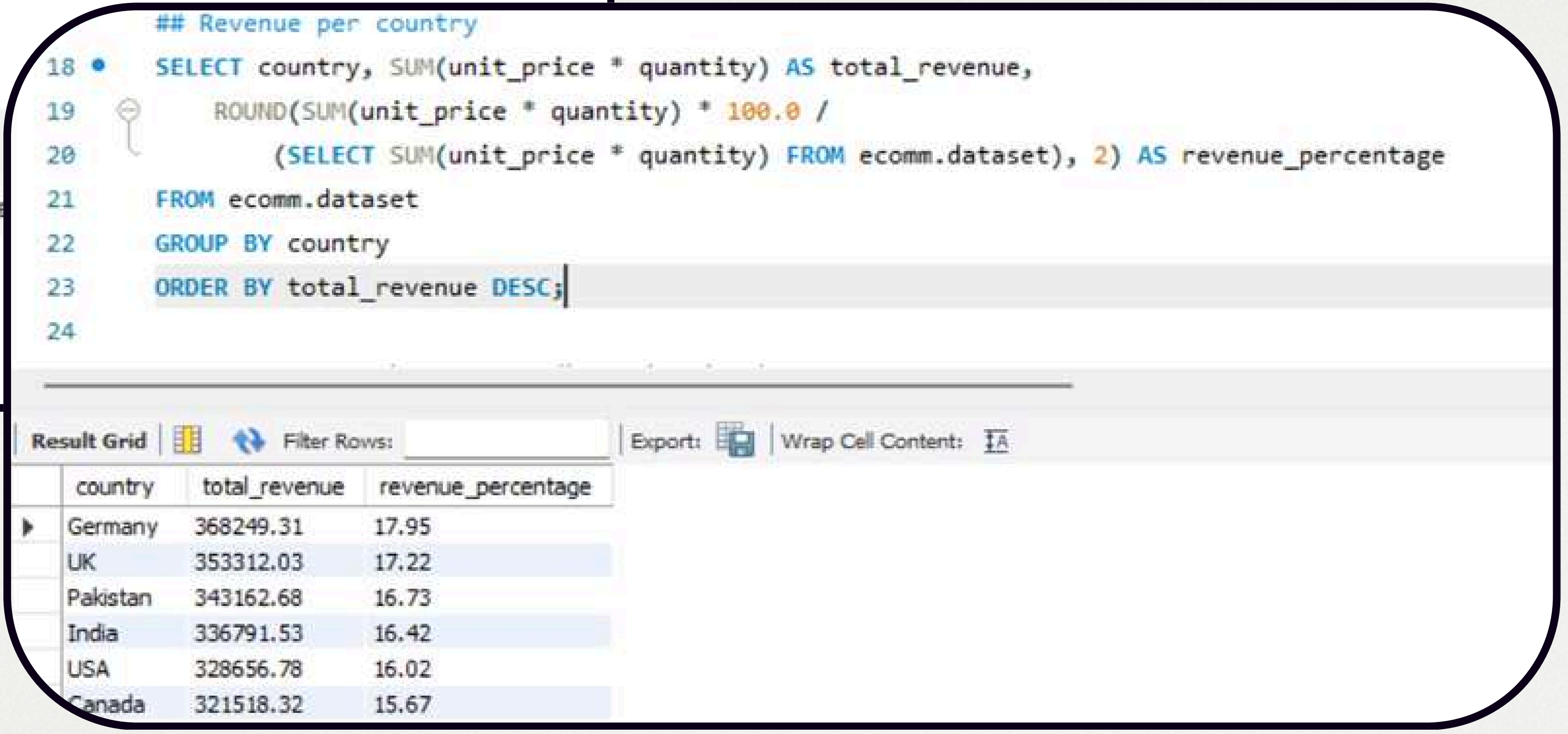
$2.051.690,65.





```
## Total revenue
SELECT round(sum(unit_price * quantity),2) AS revenue FROM ecomm.dataset;
```

revenue
2051690.65



```
## Revenue per country
SELECT country, SUM(unit_price * quantity) AS total_revenue,
    ROUND(SUM(unit_price * quantity) * 100.0 /
        (SELECT SUM(unit_price * quantity) FROM ecomm.dataset), 2) AS revenue_percentage
FROM ecomm.dataset
GROUP BY country
ORDER BY total_revenue DESC;
```

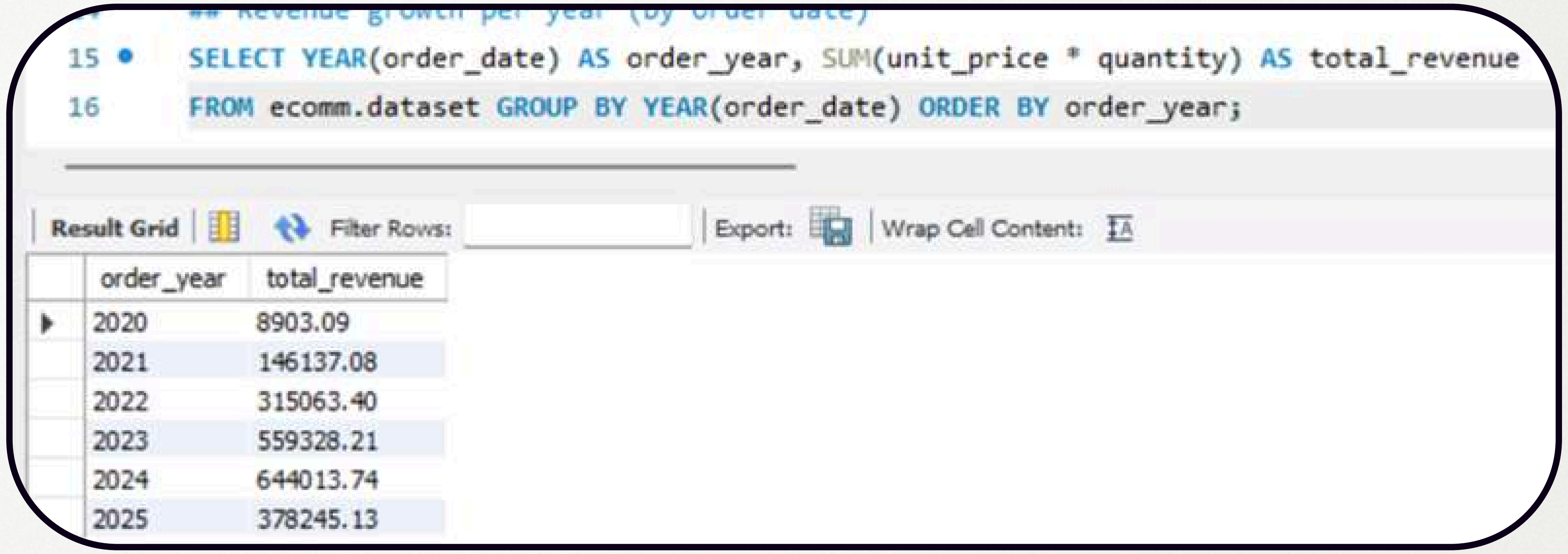
country	total_revenue	revenue_percentage
Germany	368249.31	17.95
UK	353312.03	17.22
Pakistan	343162.68	16.73
India	336791.53	16.42
USA	328656.78	16.02
Canada	321518.32	15.67

Total Revenue (2)

The revenue growth were increased every year, along with sales growth. This indicates that the growth is being driven by both higher transaction counts and consistent customer spending.

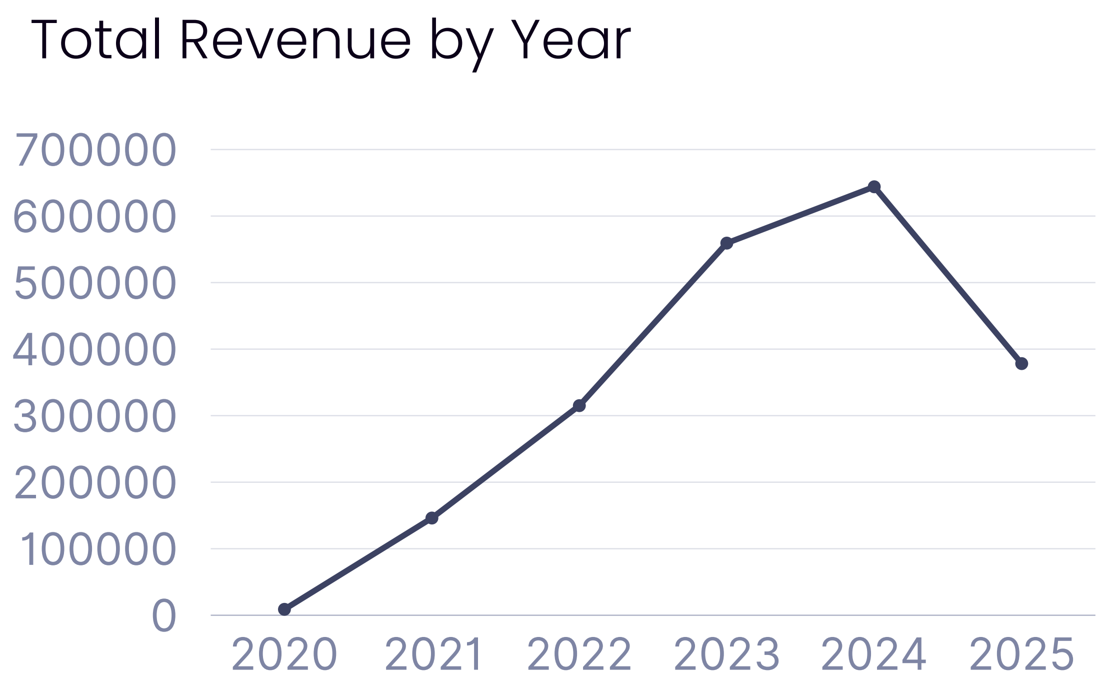
Total Revenue

$2.051.690,65.



```
## Revenue growth per year (by order date)
SELECT YEAR(order_date) AS order_year, SUM(unit_price * quantity) AS total_revenue
FROM ecomm.dataset GROUP BY YEAR(order_date) ORDER BY order_year;
```

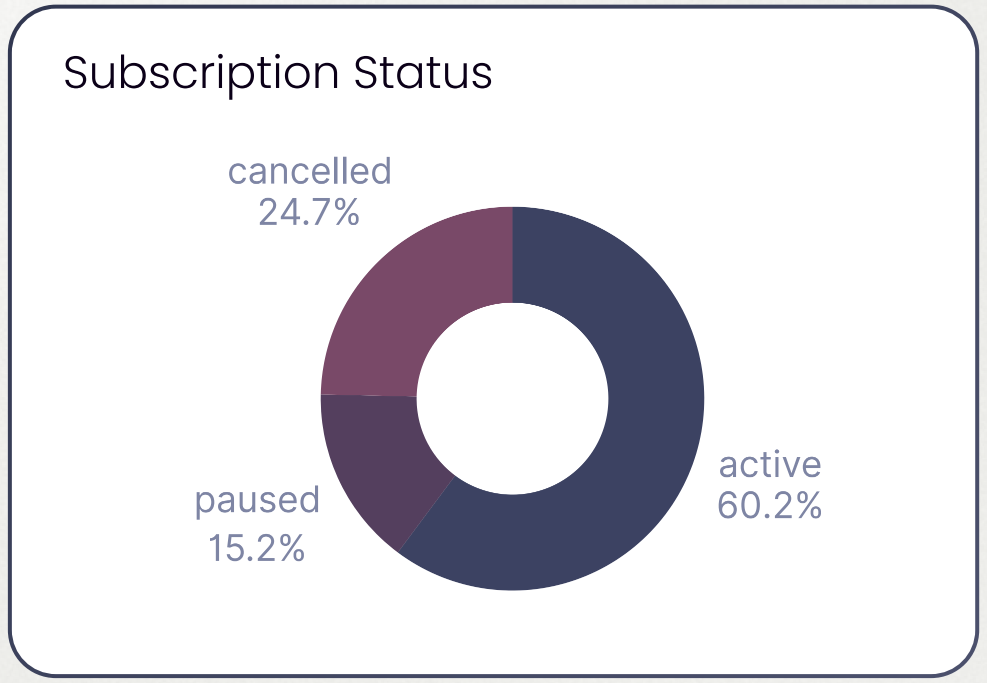
```
order_year  total_revenue
2020        8903.09
2021        146137.08
2022        315063.40
2023        559328.21
2024        644013.74
2025        378245.13
```

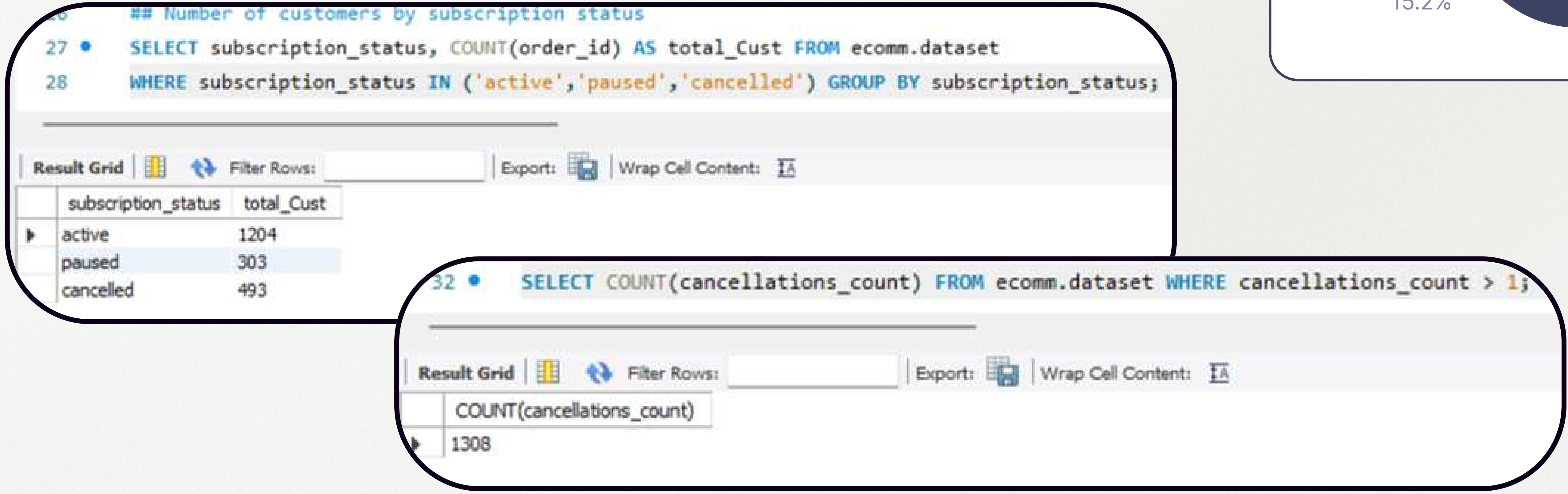


Total Revenue by Year

Customers Demographic

There are 2000 customers. By the time this dataset used (early second half of 2025), based on subscription status there is 1204 active customers and there number of customers with more than 1 cancellation orders are 1308 cust.





```
## Number of customers by subscription status
SELECT subscription_status, COUNT(order_id) AS total_Cust FROM ecomm.dataset
WHERE subscription_status IN ('active','paused','cancelled') GROUP BY subscription_status;
```

subscription_status	total_Cust
active	1204
paused	303
cancelled	493

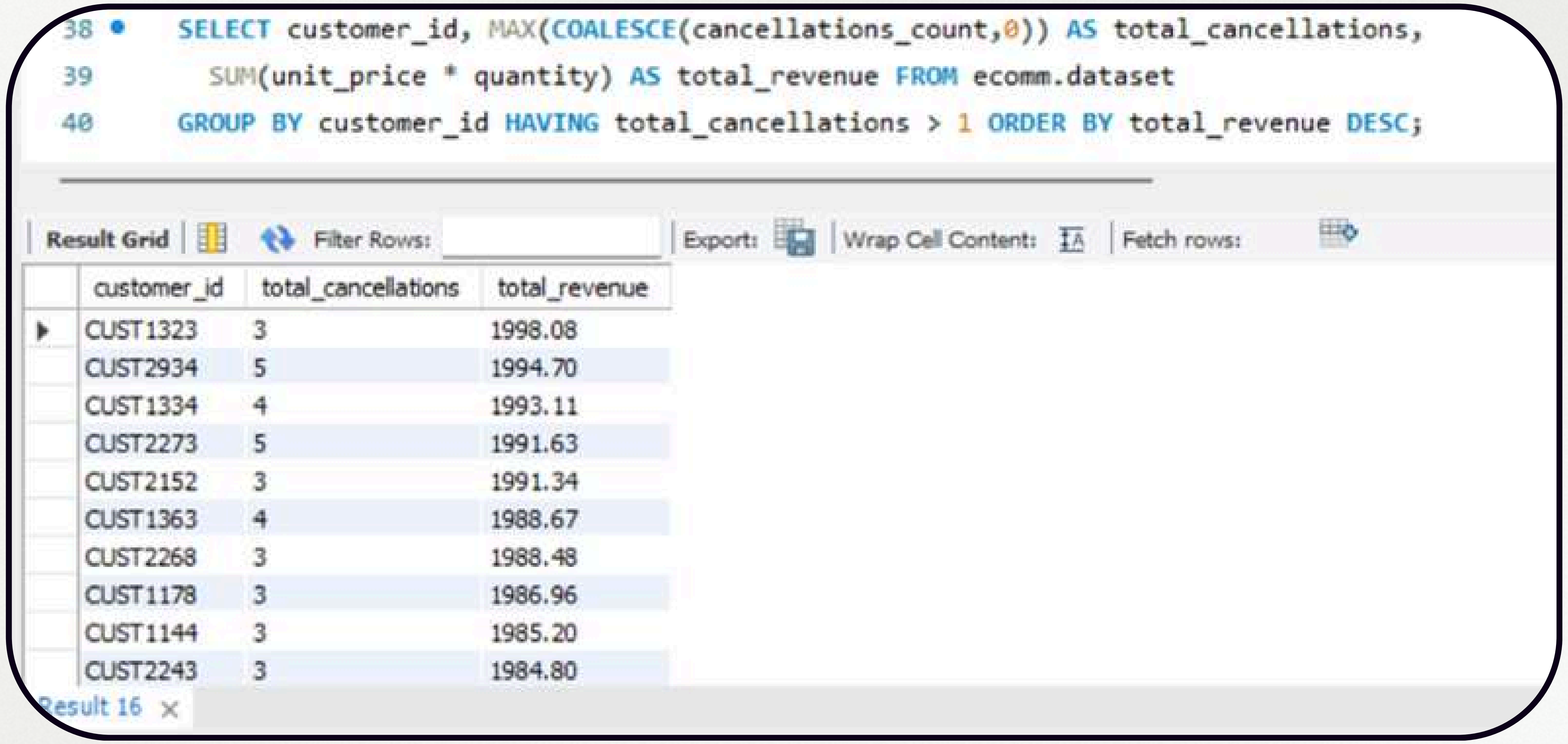
```
SELECT COUNT(cancellations_count) FROM ecomm.dataset WHERE cancellations_count > 1;
```

COUNT(cancellations_count)
1308

The numbers of customers with more than one cancellation are 1308 customers.

Customers Demographic (2)

There are several high-value customers doing repeated cancellations (3–5 times each) while still generating close to $2,000 in revenue. This suggests that cancellations do not necessarily equate to lost customers — instead, they highlight friction points in the purchase journey for otherwise loyal and profitable users.

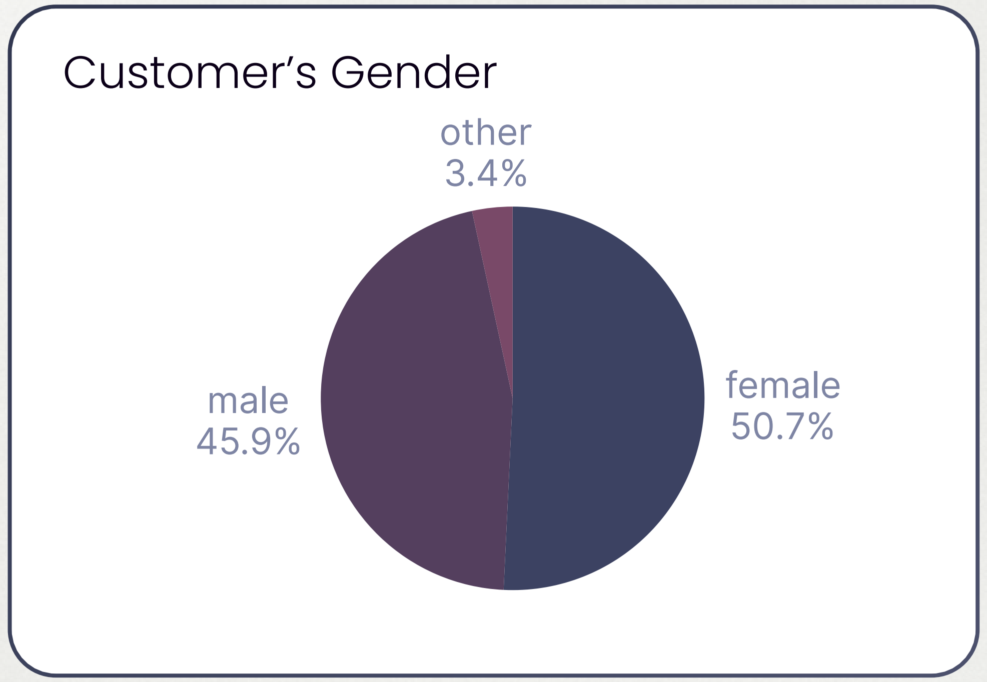


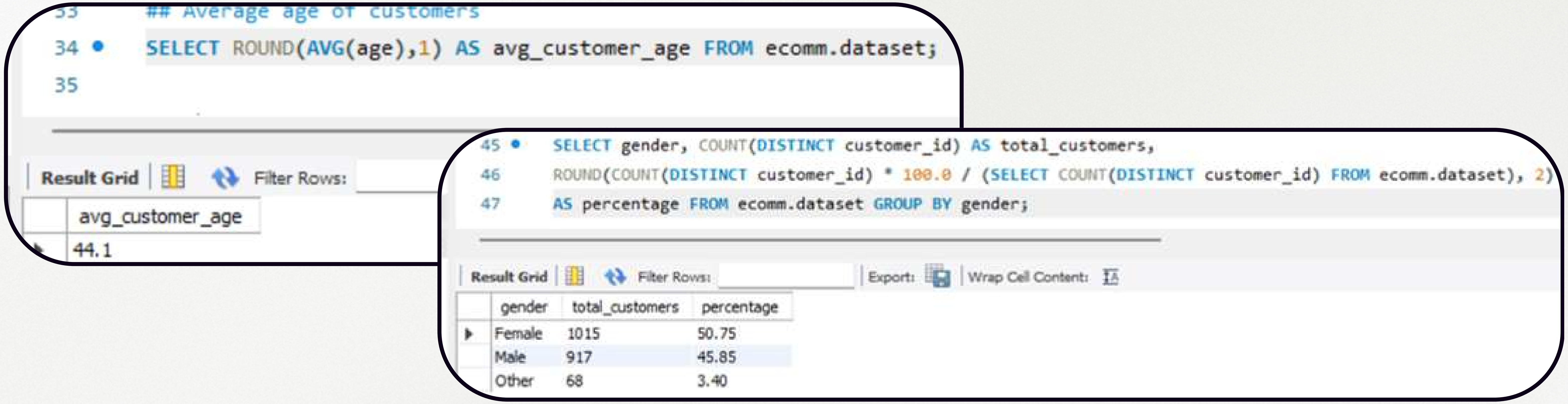
```
SELECT customer_id, MAX(COALESCE(cancellations_count,0)) AS total_cancellations,
  SUM(unit_price * quantity) AS total_revenue FROM ecomm.dataset
GROUP BY customer_id HAVING total_cancellations > 1 ORDER BY total_revenue DESC;
```

customer_id	total_cancellations	total_revenue
CUST1323	3	1998.08
CUST2934	5	1994.70
CUST1334	4	1993.11
CUST2273	5	1991.63
CUST2152	3	1991.34
CUST1363	4	1988.67
CUST2268	3	1988.48
CUST1178	3	1986.96
CUST1144	3	1985.20
CUST2243	3	1984.80

Customers Demographic (3)

The average age of customers is 44 years, indicating that most of them are middle-aged individuals with an assumably stable income so they may be more capable of purchasing mid- to high-priced products. The gender distribution is nearly equal. Since both genders are equally represented, segmentation by gender may not be necessary — other factors (like age or income) might provide better targeting opportunities.





```
## Average age of customers
SELECT ROUND(AVG(age),1) AS avg_customer_age FROM ecomm.dataset;
```

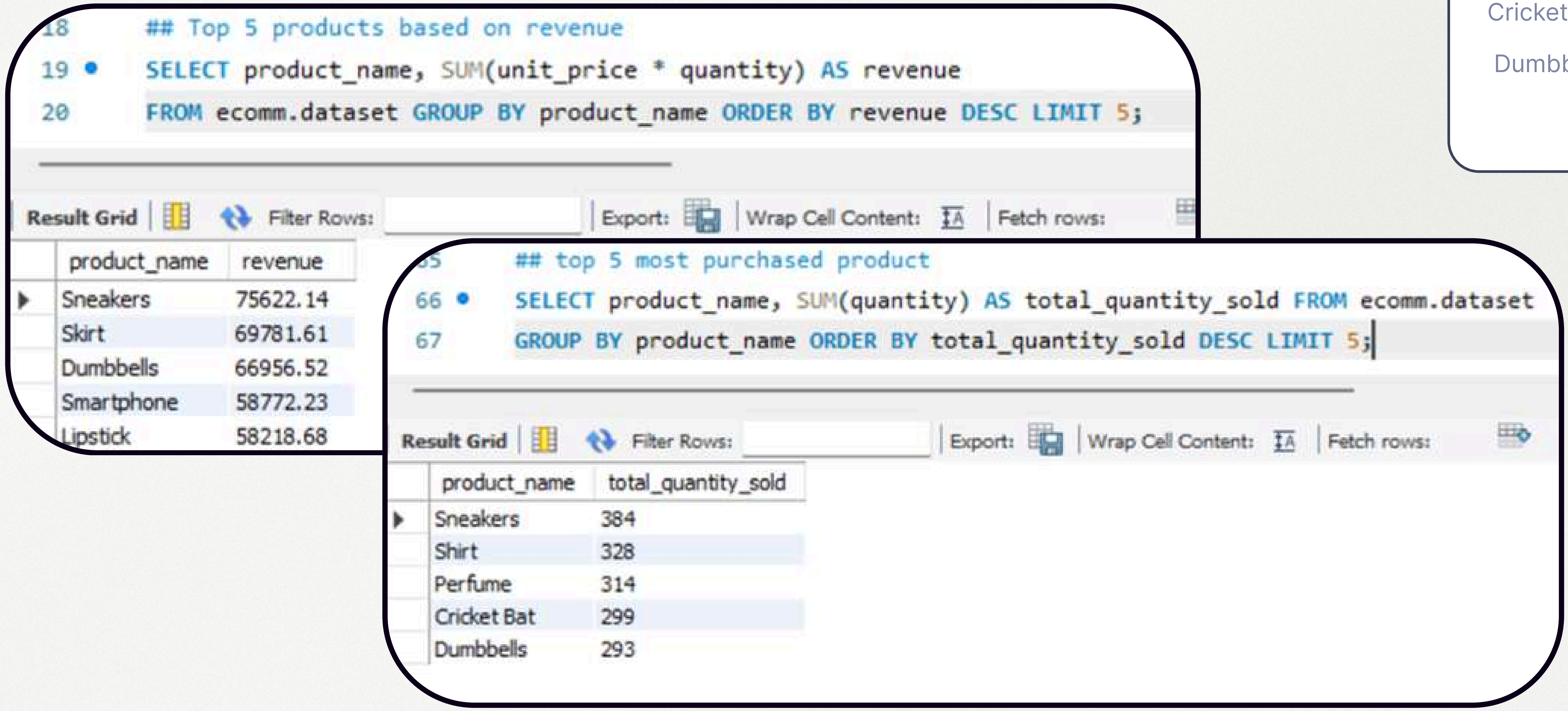
avg_customer_age
44.1

```
SELECT gender, COUNT(DISTINCT customer_id) AS total_customers,
ROUND(COUNT(DISTINCT customer_id) * 100.0 / (SELECT COUNT(DISTINCT customer_id) FROM ecomm.dataset), 2)
AS percentage FROM ecomm.dataset GROUP BY gender;
```

gender	total_customers	percentage
Female	1015	50.75
Male	917	45.85
Other	68	3.40

Product & Category

The top 5 products based on revenue show that “Sneakers” earned the highest revenue of $75,622.14 and were the most sold product with 384 units sold.

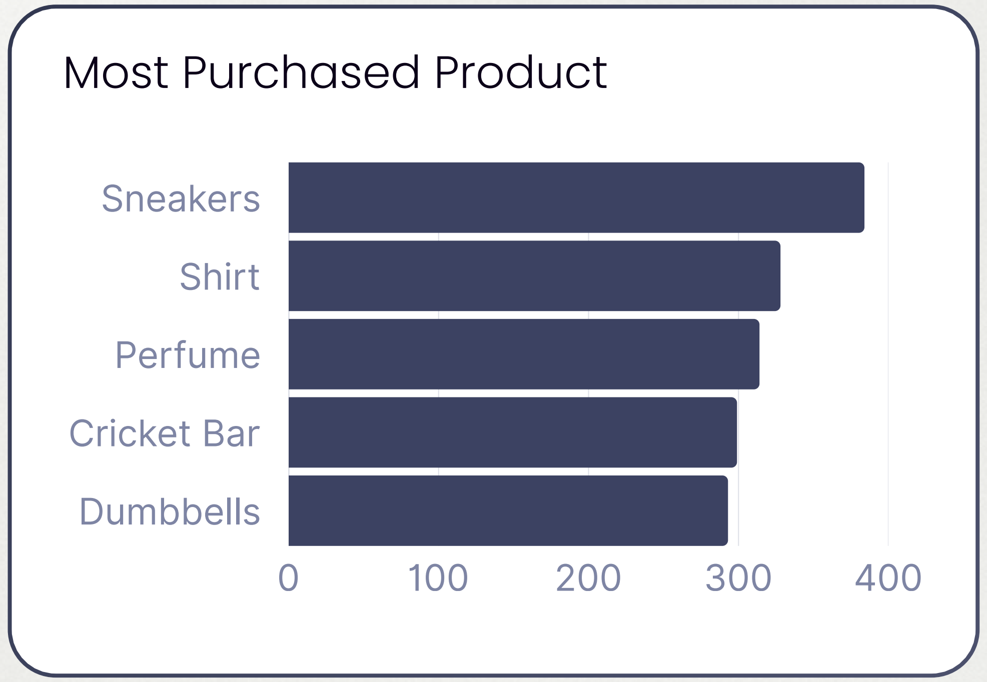


```
## Top 5 products based on revenue
SELECT product_name, SUM(unit_price * quantity) AS revenue
FROM ecomm.dataset GROUP BY product_name ORDER BY revenue DESC LIMIT 5;
```

product_name	revenue
Sneakers	75622.14
Skirt	69781.61
Dumbbells	66956.52
Smartphone	58772.23
Lipstick	58218.68

```
## top 5 most purchased product
SELECT product_name, SUM(quantity) AS total_quantity_sold FROM ecomm.dataset
GROUP BY product_name ORDER BY total_quantity_sold DESC LIMIT 5;
```

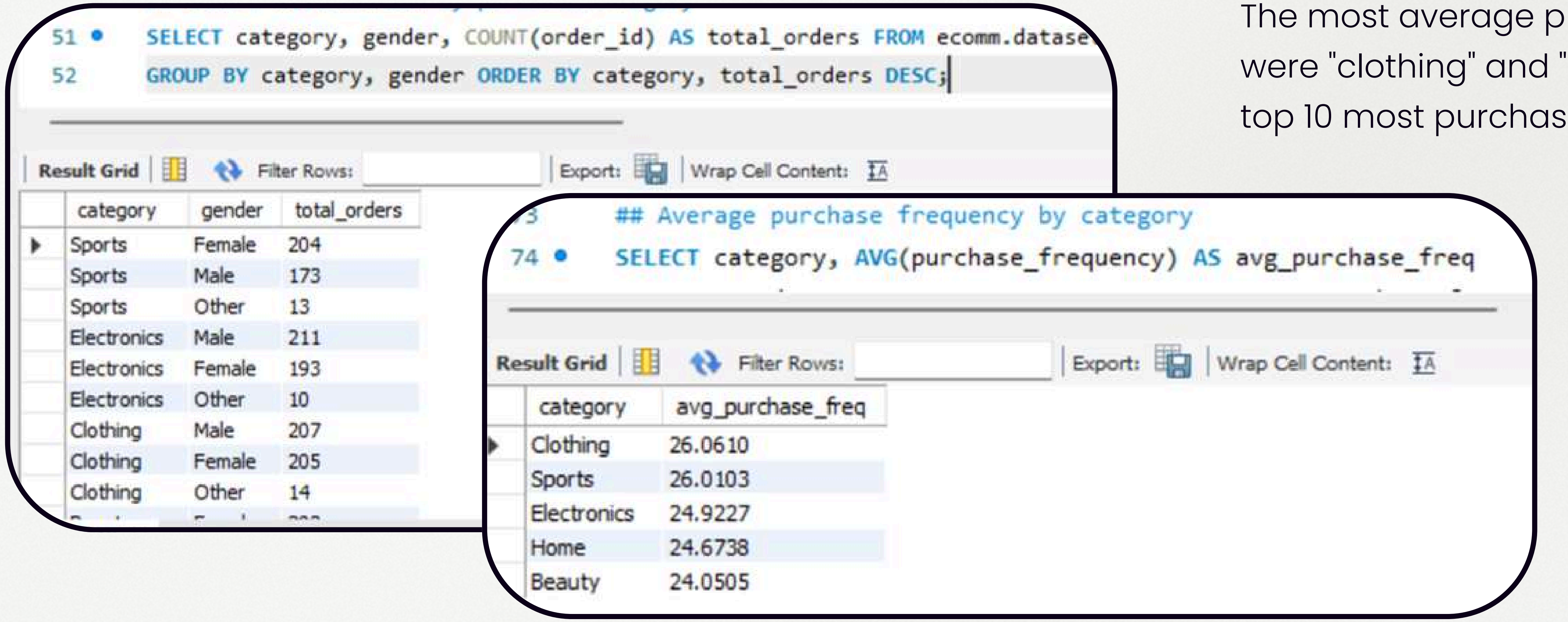
product_name	total_quantity_sold
Sneakers	384
Shirt	328
Perfume	314
Cricket Bat	299
Dumbbells	293



Most Purchased Product

Product & Category (2)

The purchase category by gender distribution showed the most purchased were "electronics" and "clothing" by male, and "home" and "clothing" by female. The "clothing" category being one of the most purchase category were align with product "shirt" as in top 5 most purchased product.



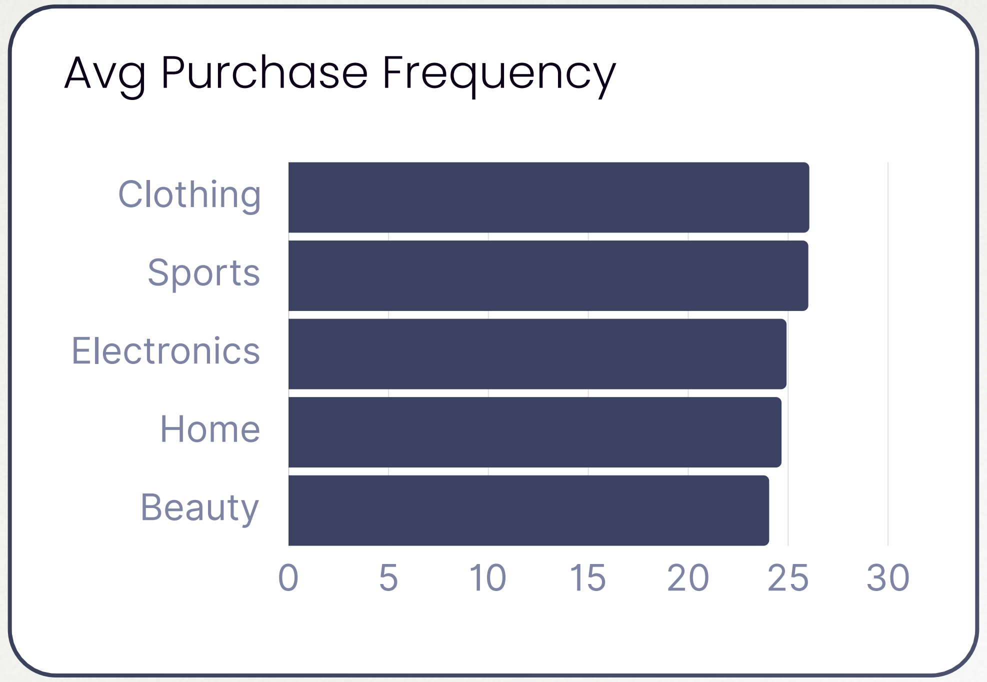
```sql
SELECT category, gender, COUNT(order_id) AS total_orders FROM ecomm.datase
GROUP BY category, gender ORDER BY category, total_orders DESC;
```

category	gender	total_orders
Sports	Female	204
Sports	Male	173
Sports	Other	13
Electronics	Male	211
Electronics	Female	193
Electronics	Other	10
Clothing	Male	207
Clothing	Female	205
Clothing	Other	14

```sql
## Average purchase frequency by category
SELECT category, AVG(purchase_frequency) AS avg_purchase_freq
```

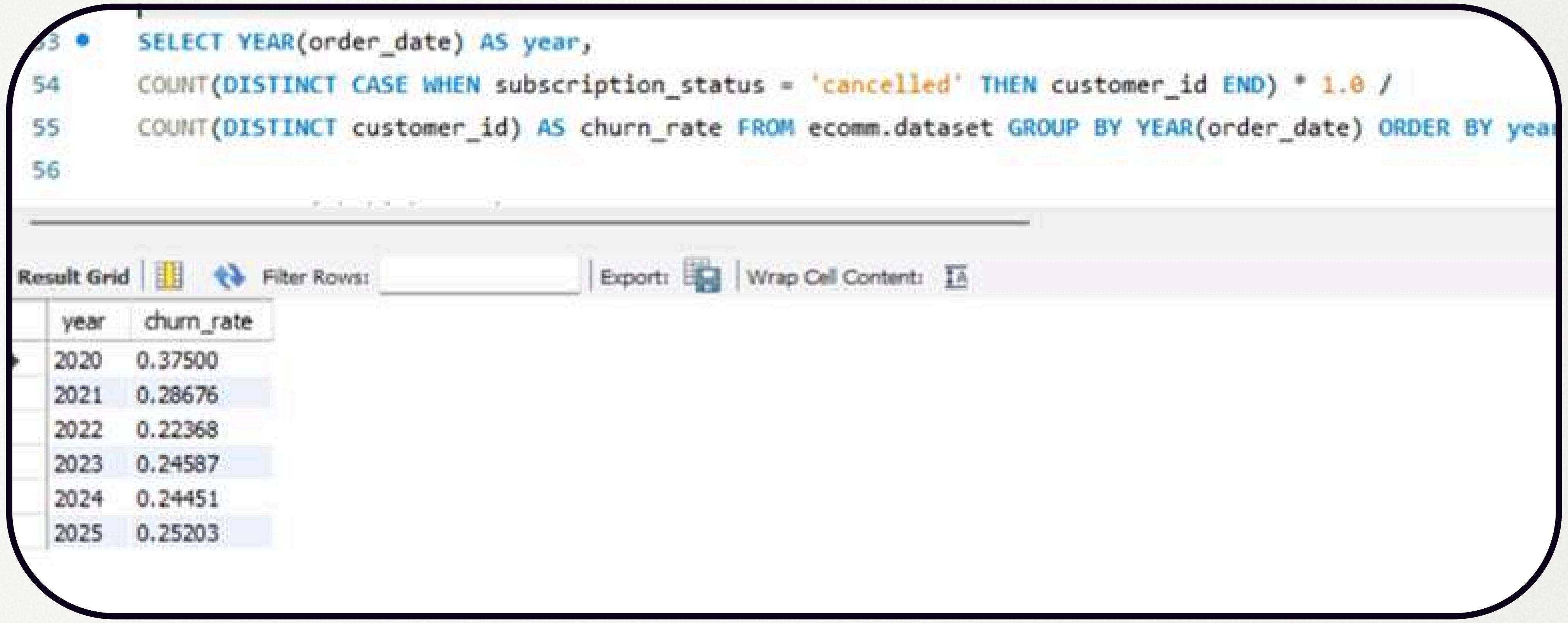
category	avg_purchase_freq
Clothing	26.0610
Sports	26.0103
Electronics	24.9227
Home	24.6738
Beauty	24.0505

The most average purchase frequency by category were "clothing" and "sports" as both of them leading in top 10 most purchase category.



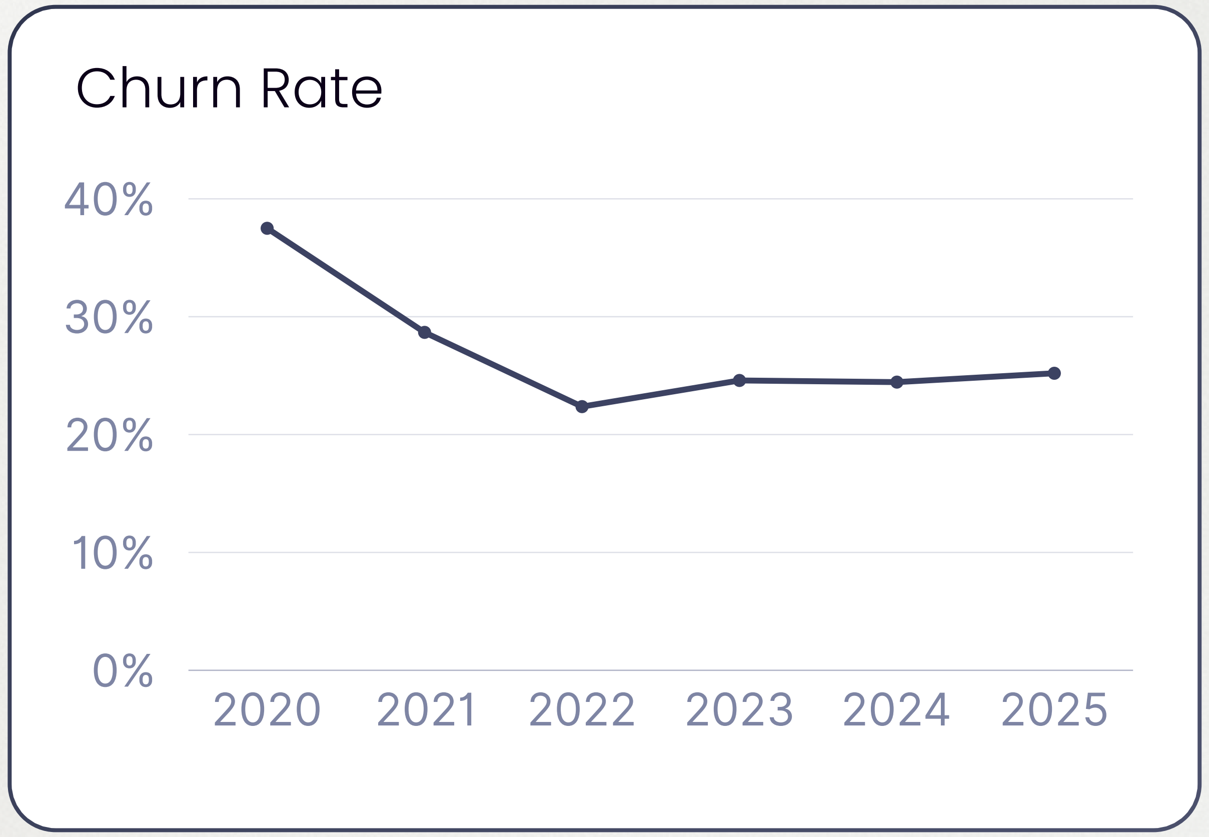
Churn Rate

The churn rate incredibly declined from 37% in 2020 to 22% in 2022 indicating some improvement in customer experience. However, the churn rate start to increased in 2023, suggesting that retention efforts weakened as the customer base grew.



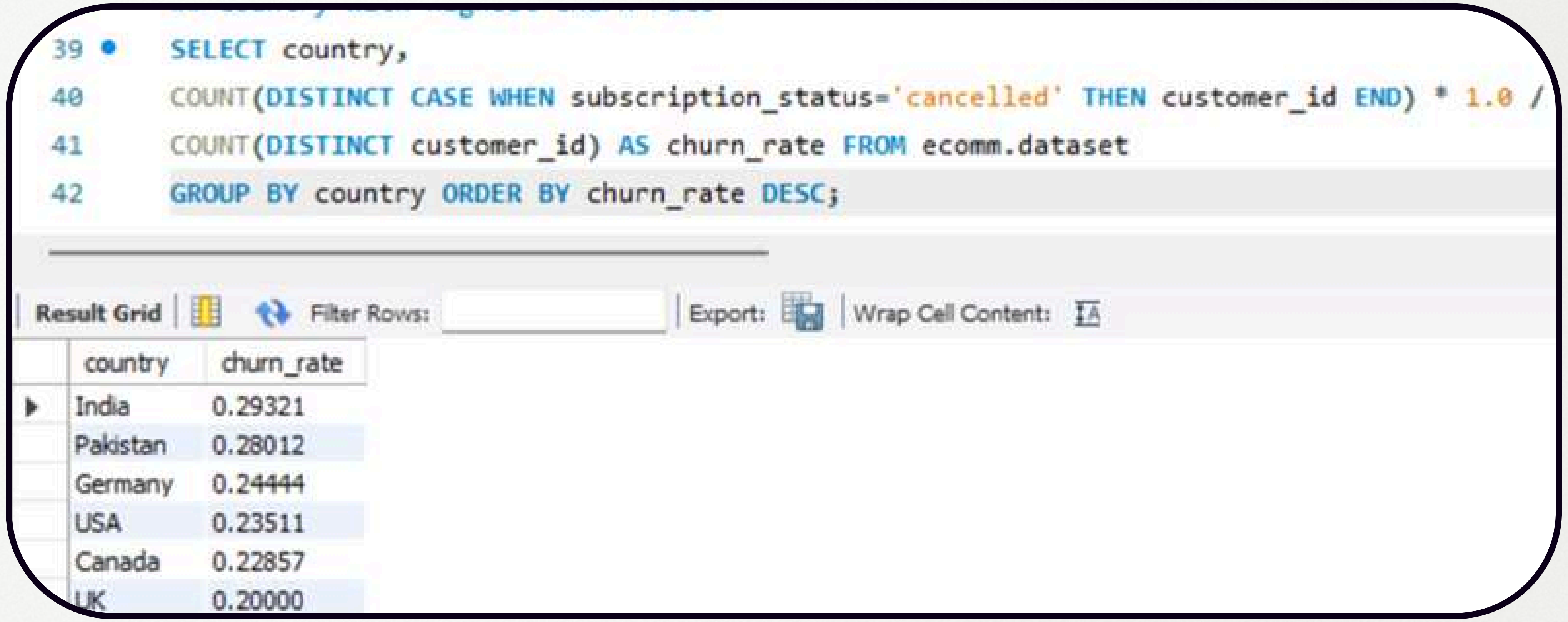
```
SELECT YEAR(order_date) AS year,
COUNT(DISTINCT CASE WHEN subscription_status = 'cancelled' THEN customer_id END) * 1.0 /
COUNT(DISTINCT customer_id) AS churn_rate FROM ecomm.dataset GROUP BY YEAR(order_date) ORDER BY yea
```

year	churn_rate
2020	0.37500
2021	0.28676
2022	0.22368
2023	0.24587
2024	0.24451
2025	0.25203



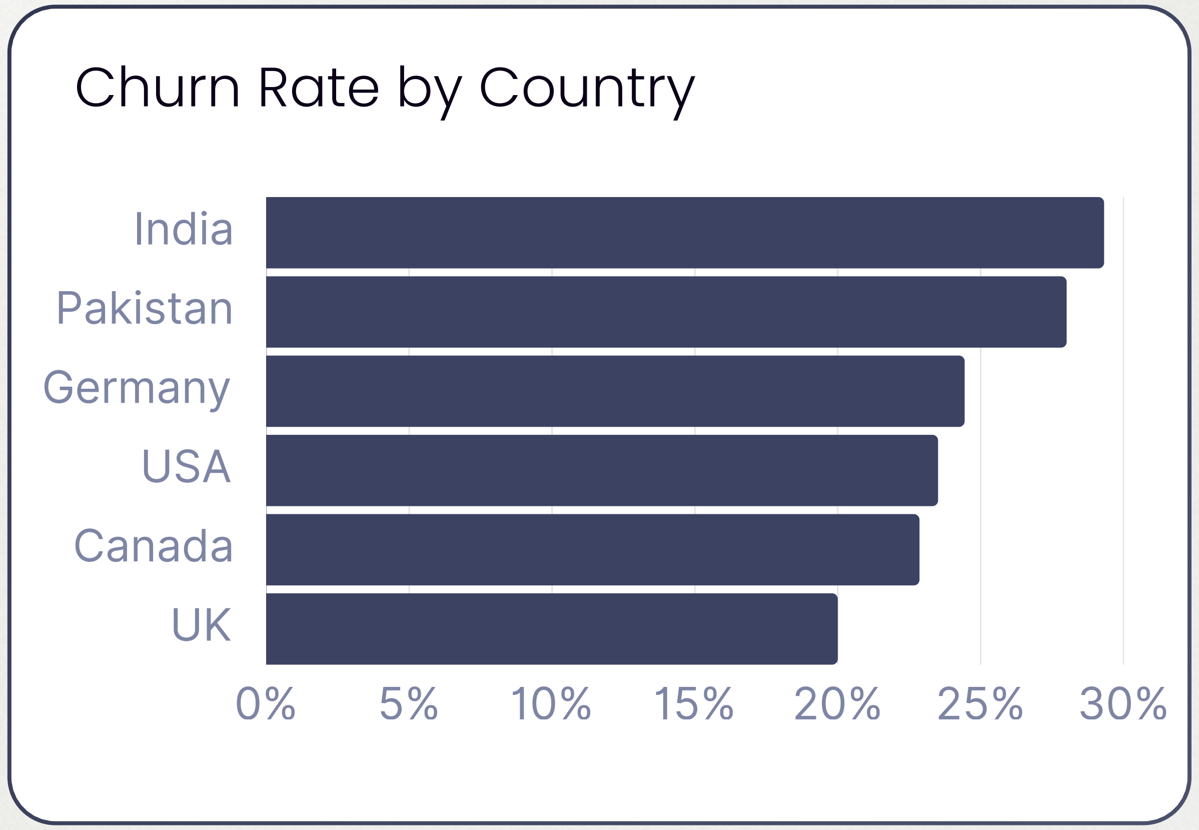
Churn Rate (2)

India has the highest churn rate (0.293), indicating that nearly 29% of customers in India canceled their subscriptions. This suggests a possible issue with customer satisfaction, pricing, or local engagement strategies compared to other countries.



```sql
SELECT country,
COUNT(DISTINCT CASE WHEN subscription_status='cancelled' THEN customer_id END) * 1.0 /
COUNT(DISTINCT customer_id) AS churn_rate FROM ecomm.dataset
GROUP BY country ORDER BY churn_rate DESC;
```

```
country    churn_rate
India      0.29321
Pakistan   0.28012
Germany    0.24444
USA        0.23511
Canada     0.22857
UK         0.20000
```



Churn Rate by Country

Overview

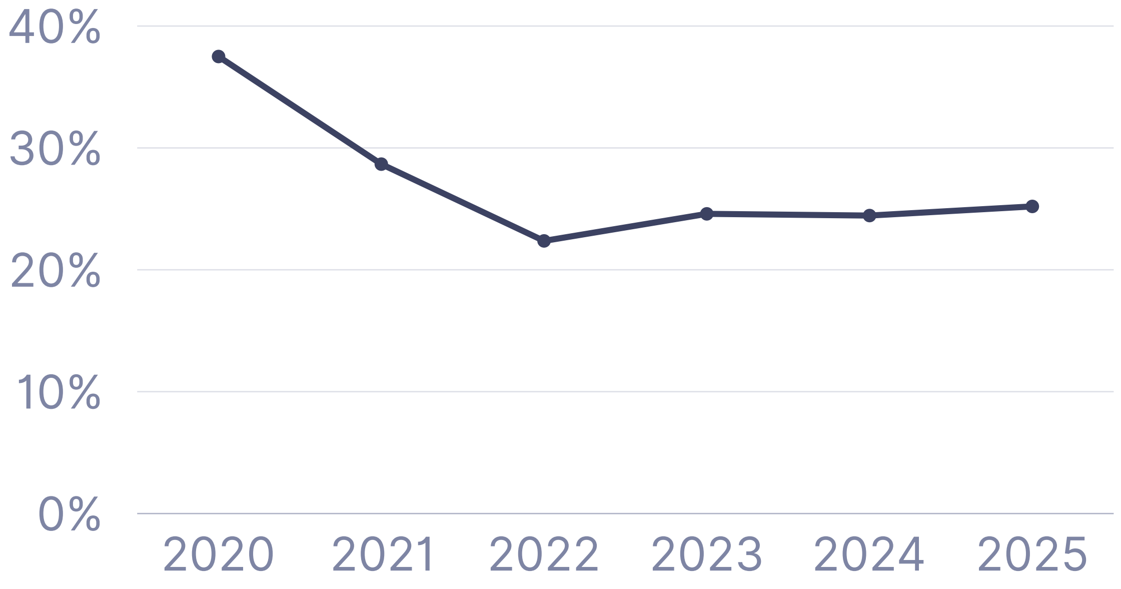
Total Orders
2000

Total Revenue
$2.051.690,65.

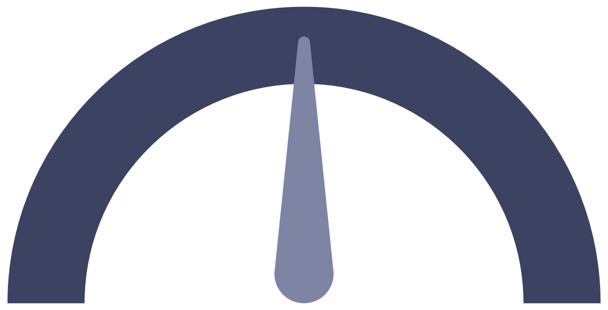
Average Order Value
$1025.85

Churn Rate

+0.7% From last year



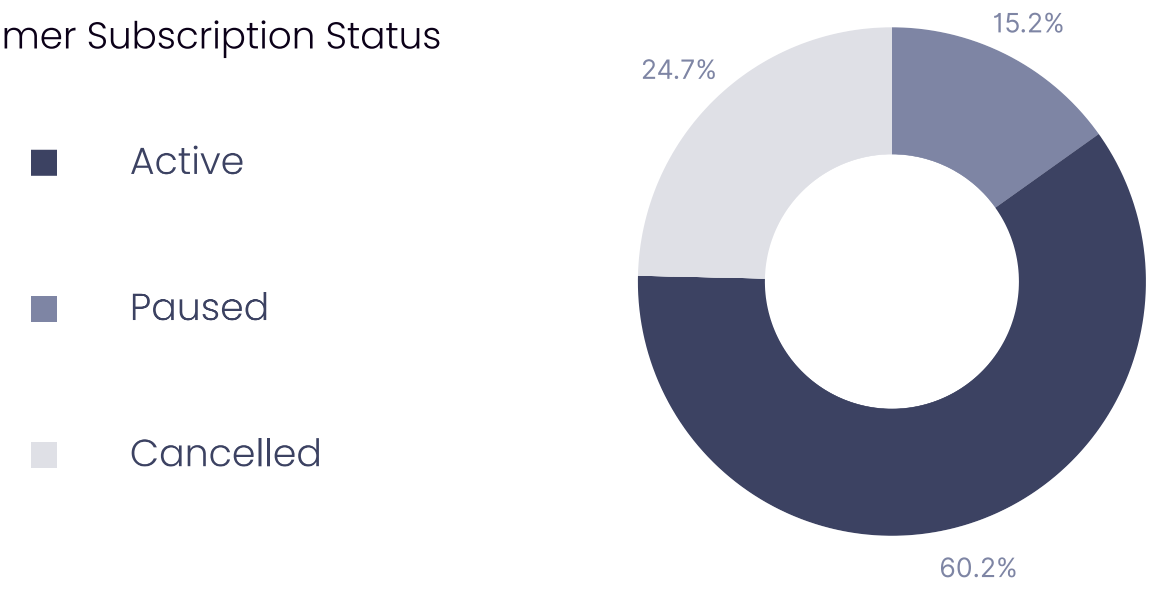
Customer Retention Rate



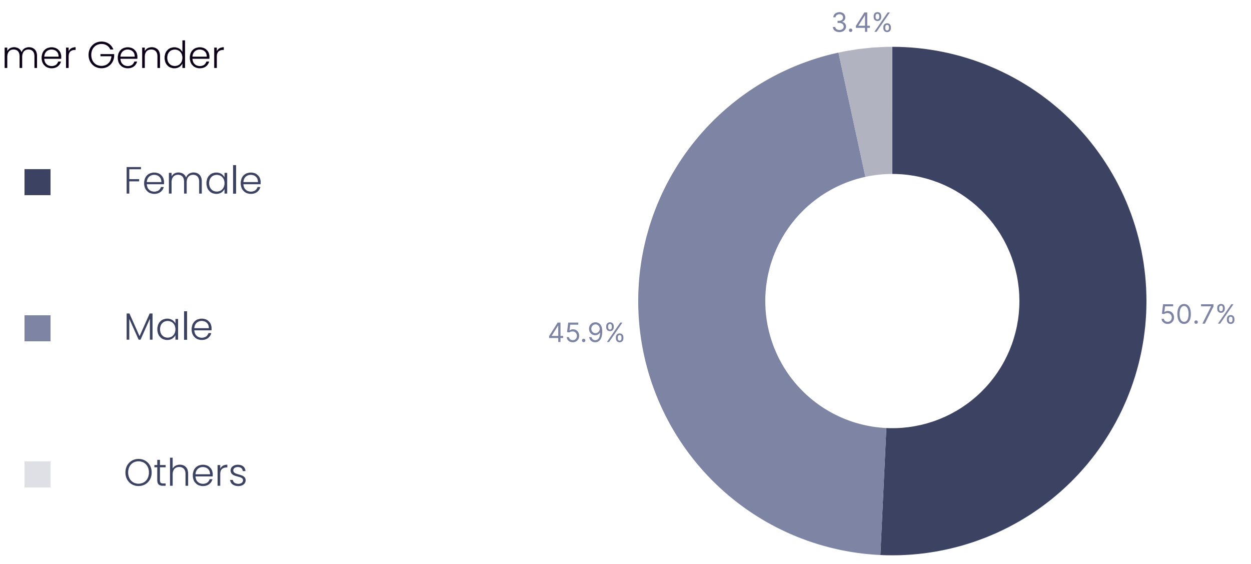
74.8%

-0.75% From last year

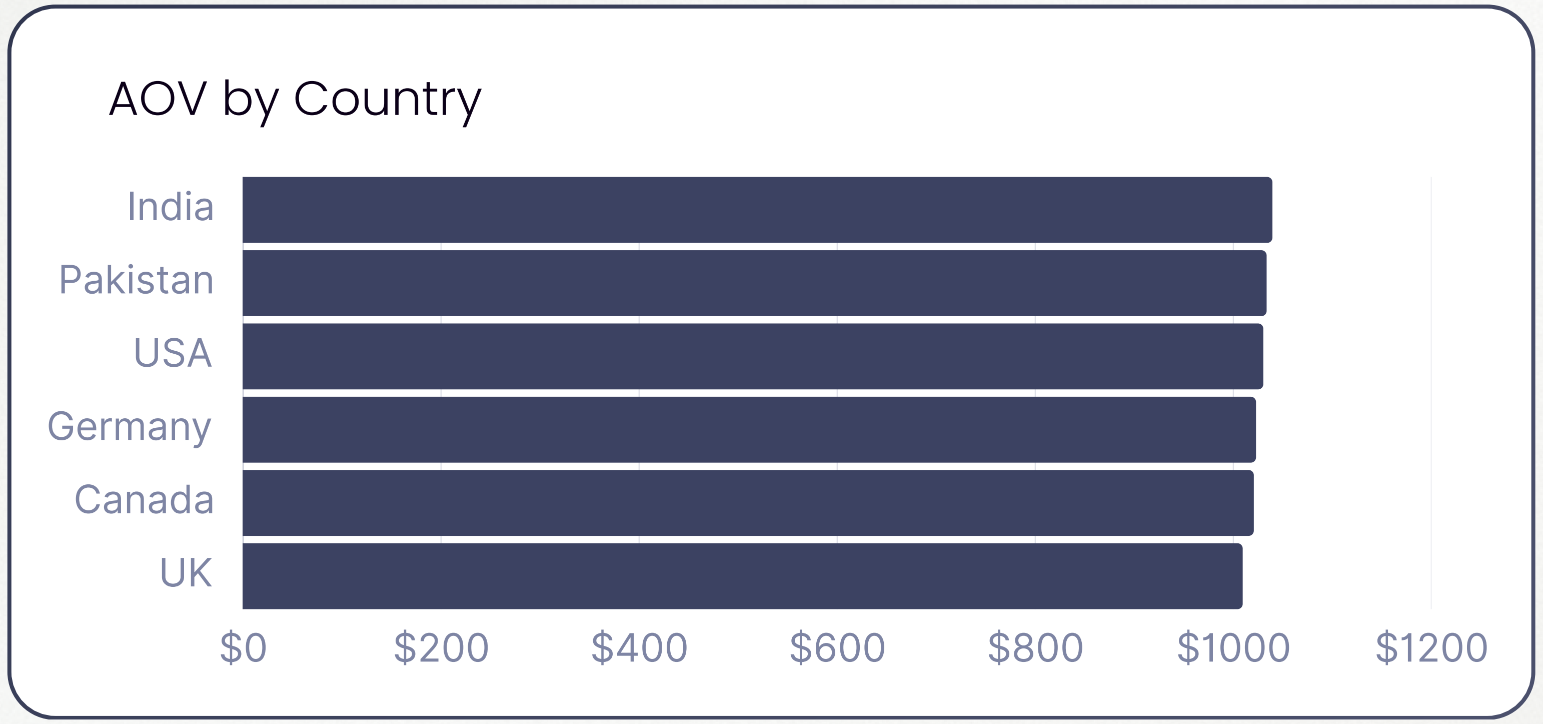
Customer Subscription Status

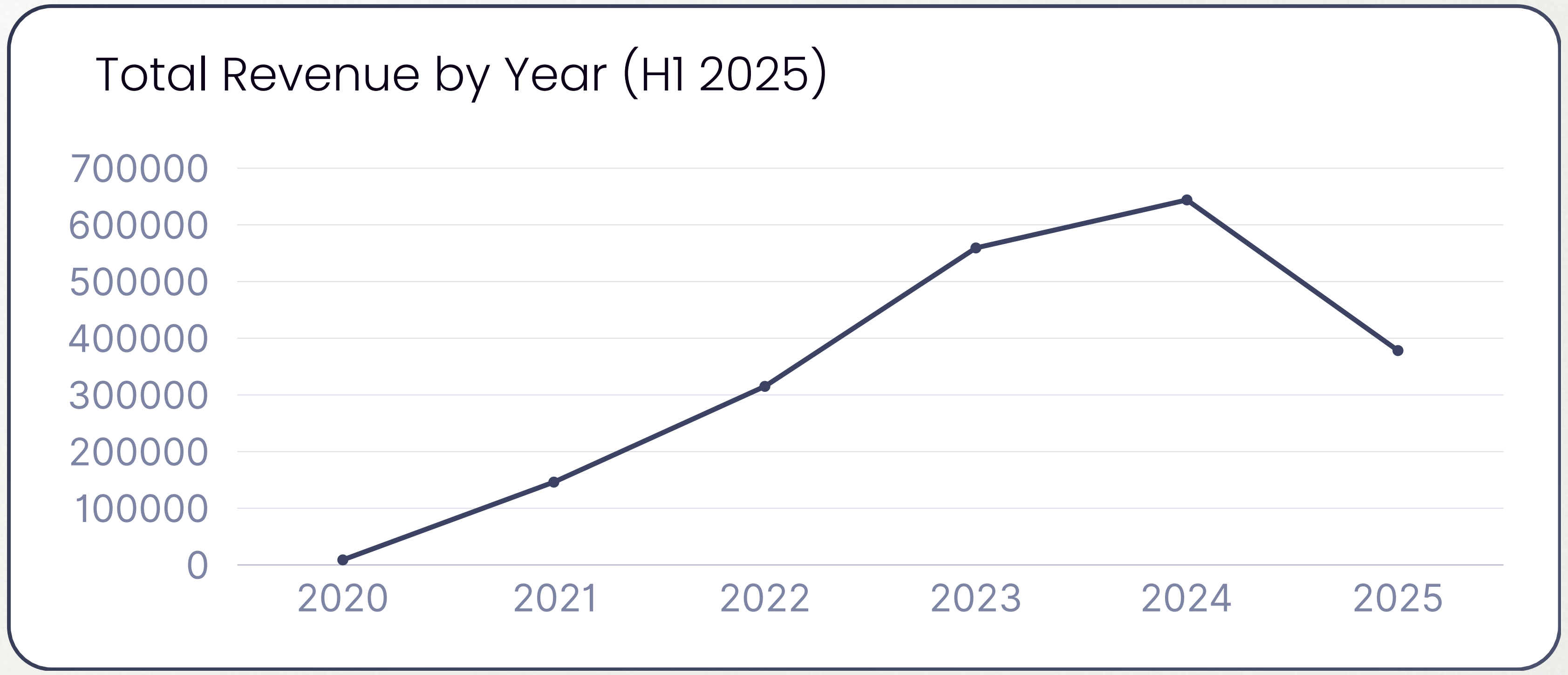


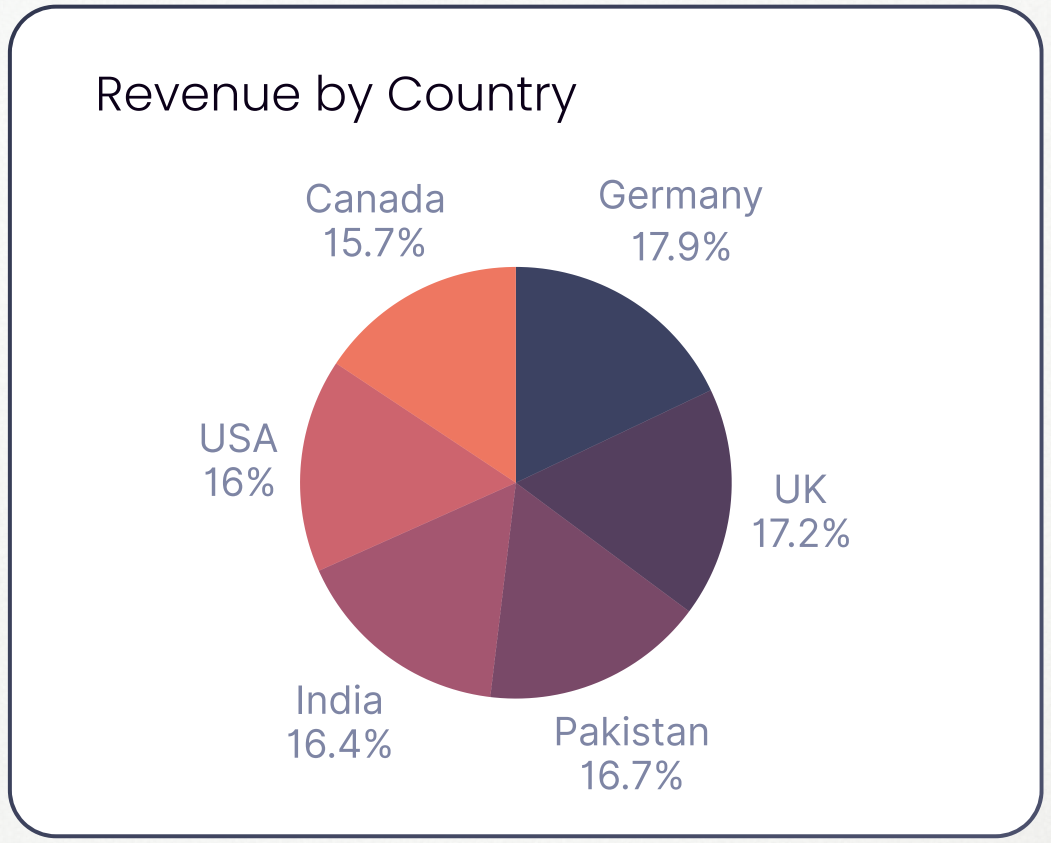
Customer Gender

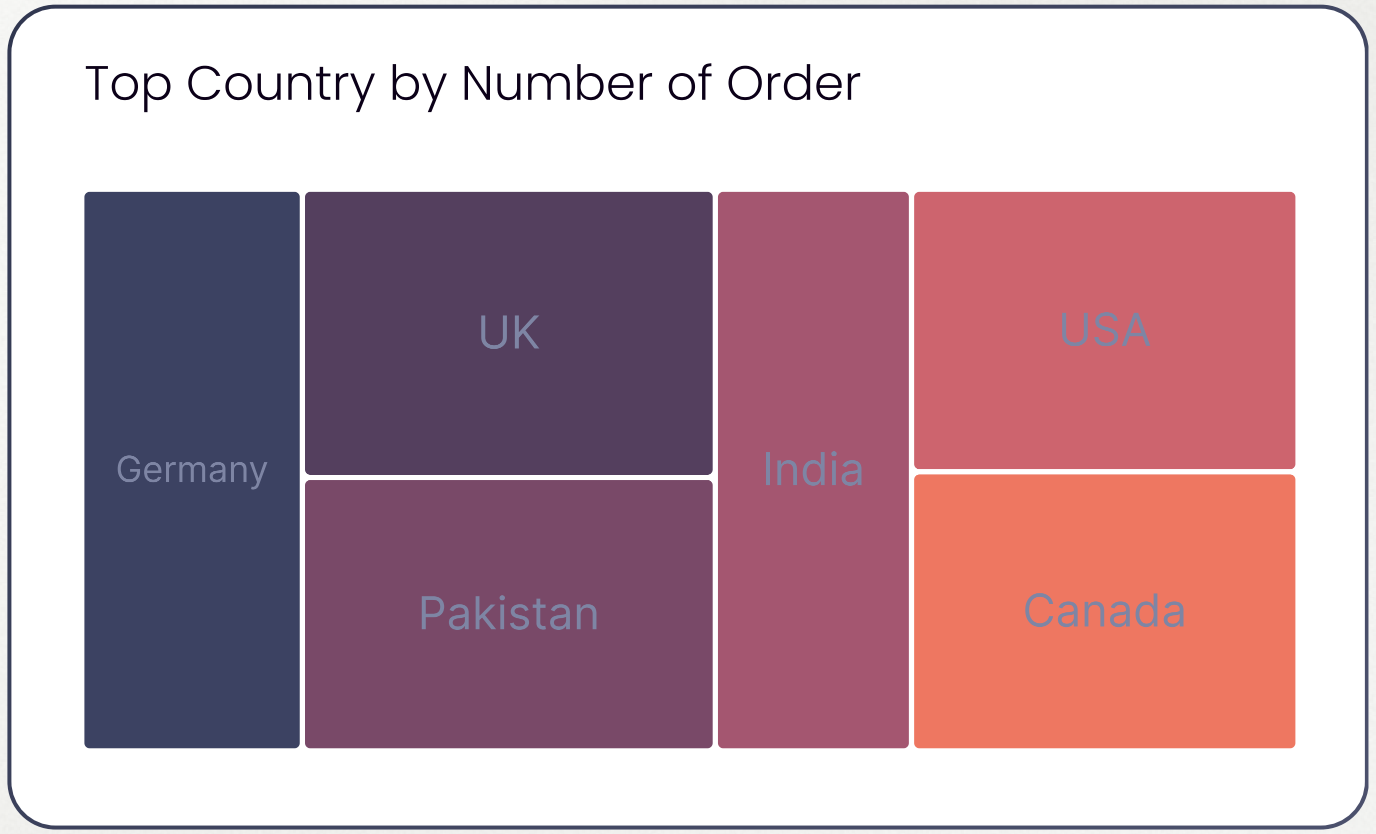


Sales Performance









Insights

Germany contributes the most orders; India and Pakistan show high AOV.

Conclusion

Exceeding Expectations

The business demonstrates strong overall performance with consistent growth in both orders and total revenue each year, reflecting increasing customer engagement. Revenue is well distributed across multiple countries, reducing dependency on a single market, while India and Pakistan stand out for their high Average Order Values (AOV). The customer base is diverse and stable, with an average age of 44 years and nearly equal gender distribution. Product performance is also promising, with “Sneakers” leading in both sales and revenue, supported by the popularity of categories such as Clothing and Sports.

Areas Needing Attention

Despite earlier improvements, the churn rate began to rise again after 2023, suggesting that retention strategies have weakened as the customer base expanded. Some high-value customers were found to cancel multiple orders despite contributing significant revenue, indicating potential friction in the purchase or delivery process. Additionally, churn varies by country, highlighting inconsistencies in customer satisfaction and retention efforts across regions. These areas require closer examination to prevent future customer loss.

Focus Areas

To sustain growth, the company should prioritize strengthening customer retention and loyalty programs while addressing issues behind repeated cancellations. Improving user experience, especially during checkout and after purchase, will be essential to maintain long-term satisfaction. Marketing strategies can be refined through demographic segmentation, leveraging insights from age, gender, and product preferences. Furthermore, focusing on high-performing product categories like Clothing and Sports while exploring underperforming ones will help balance growth. Continuous monitoring of churn and retention by country and customer group will ensure data-driven decision-making and consistent improvement.

Thank You

☎ +62-851-7990-6131

Linkedin milda-s

ks.milda.ks@gmail.com

Yogyakarta, Indonesia