Data Visualization/Exploration

Data Cleanup & Preparation

A dataset with two tables was provided: ***orders*** and ***products***, both were uploaded to BigQuery. Before starting with the analyses described in the task, it was important to validate and normalize the data. All fields were checked for data formats, nulls, ranges/outliers, both tables were checked for duplicates.

orders table:

- *Completeness of key fields*: Essential fields (*customer_id*, *created_at*, *product_items*) contain no null values.
- *Country encoding issue*: Encoding issue with “Côte d'Ivoire” in *billing_address_country* on 17 occurrences, normalized to “Ivory Coast” to avoid non-English characters and punctuation.
- *Timestamp normalization*: *created_at*, *processed_at*, *first_date_order* are STRINGS, successfully parsed to TIMESTAMP, all dates were valid.
- *Exact duplicate remova*l: 277 full exact duplicates identified and removed.
- *Remaining order_number anomalies*: After removing exact duplicates, 53 order_numbers were still duplicated, each appearing twice. All 106 affected rows occurred on **2021-03-03**, indicating a one-off system error. To avoid counting accidental duplicates, rows were removed: when the same customer (*customer_id*) had the same order content more than once (but with different *order_number*) **AND** the *order_number* also appeared for another customer, so that the most plausible version of each order was retained.
- *Product list normalization*: *product_items* is a comma-separated list of product titles, orders were unnested, so that every row contained only 1 product item.

products table:

- *Completeness of key fields***:** Essential fields (*product_type, product_category, product_title, product_price*) contain no null values.
- *Duplicate products*: 2 exact duplicates (Golf_balls_200, Golf_balls_100) identified and removed.
- *Referential integrity*: Every product item referenced in orders exists in the products table.
- *Price validation*: All price values are valid, no outliers.
- *Text format consistency***:** *product_title* values followed a consistent underscore-based naming pattern with no irregular casing, spacing, or punctuation issues.

The results of data cleaning, normalization, and the join with the products table were saved to a new table, ***orders_products_enriched***. Source tables remain unchanged.

2. a) Customer Cohort Analysis

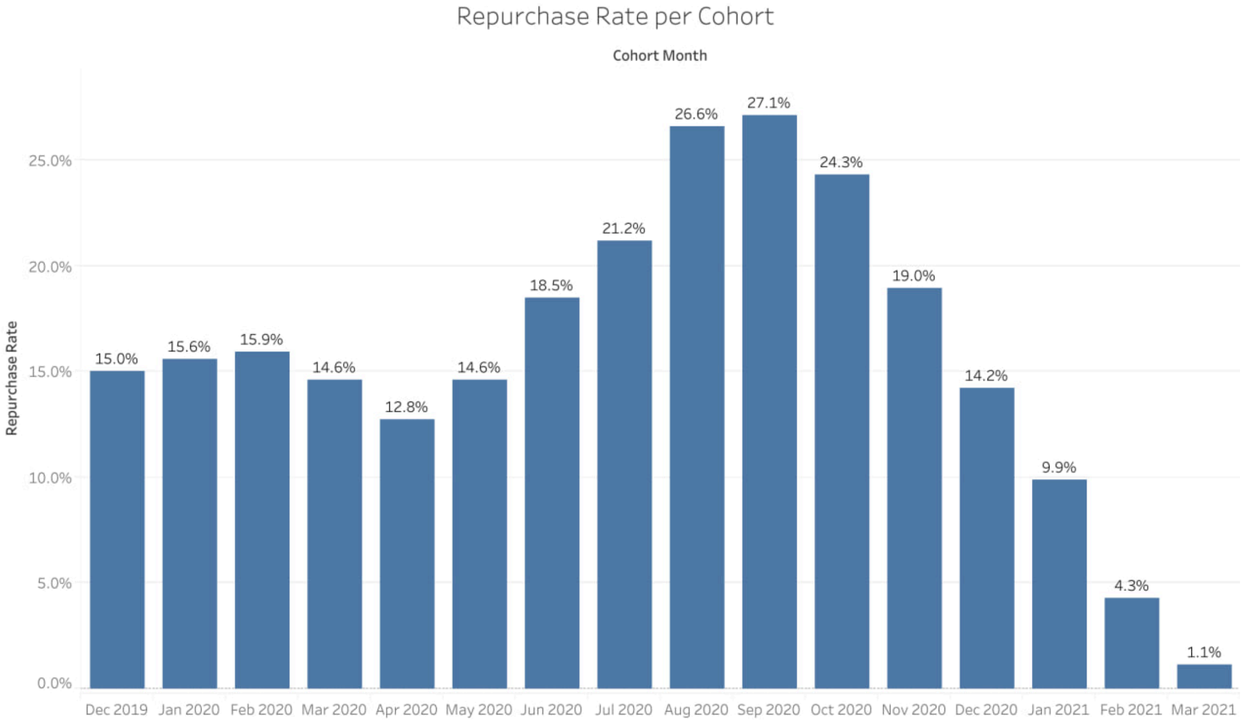
Tools: Data was processed using GoogleSQL in BigQuery. Visualisations were created in Tableau.

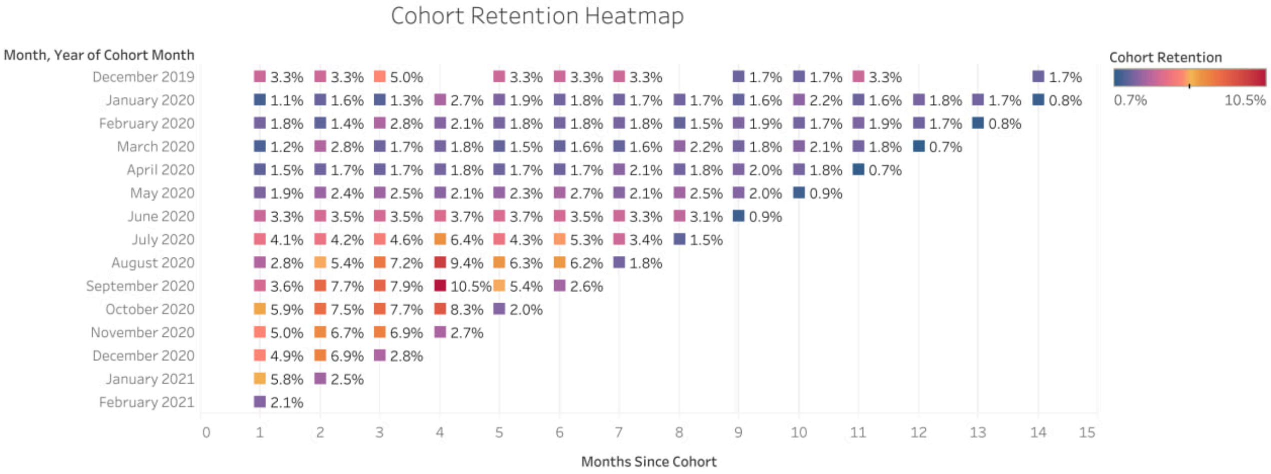
Time of acquisition is the most common characteristic to group customers into cohorts. The data covers customer orders over 16 months, which is too short for yearly cohorts, so monthly cohorts were defined. Thus, each customer was assigned to a cohort based on the month of their first recorded purchase.

Three metrics were calculated for each order-item:

- *Months Since Cohort*: Indicates the number of months between the first purchase and a subsequent order.
- *Cohort Retention*: Share of customers within a cohort who made at least one additional purchase for every month following the month of acquisition.
- *Repurchase Rate*: Share of customers who completed two or more purchases at any point in time.

Results





Cohort Retention Heatmap

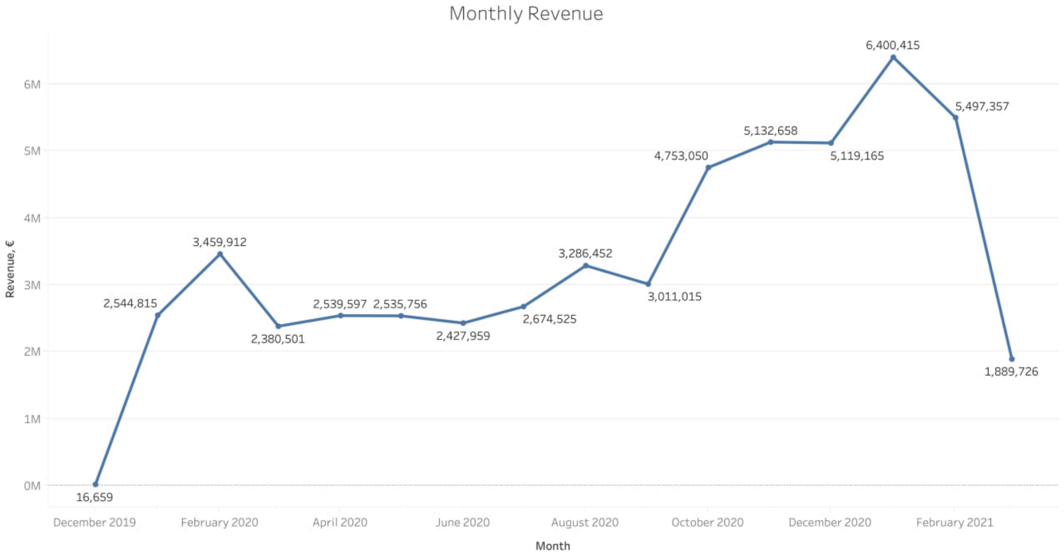
Month, Year of Cohort Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
December 2019	3.3%	3.3%	5.0%		3.3%	3.3%	3.3%		1.7%	1.7%	3.3%			1.7%
January 2020	1.1%	1.6%	1.3%	2.7%	1.9%	1.8%	1.7%	1.7%	1.6%	2.2%	1.6%	1.8%	1.7%	0.8%
February 2020	1.8%	1.4%	2.8%	2.1%	1.8%	1.8%	1.8%	1.5%	1.9%	1.7%	1.9%	1.7%	0.8%	
March 2020	1.2%	2.8%	1.7%	1.8%	1.5%	1.6%	1.6%	2.2%	1.8%	2.1%	1.8%	0.7%		
April 2020	1.5%	1.7%	1.7%	1.8%	1.7%	1.7%	2.1%	1.8%	2.0%	1.8%	0.7%			
May 2020	1.9%	2.4%	2.5%	2.1%	2.3%	2.7%	2.1%	2.5%	2.0%	0.9%				
June 2020	3.3%	3.5%	3.5%	3.7%	3.7%	3.5%	3.3%	3.1%	0.9%					
July 2020	4.1%	4.2%	4.6%	6.4%	4.3%	5.3%	3.4%	1.5%						
August 2020	2.8%	5.4%	7.2%	9.4%	6.3%	6.2%	1.8%							
September 2020	3.6%	7.7%	7.9%	10.5%	5.4%	2.6%								
October 2020	5.9%	7.5%	7.7%	8.3%	2.0%									
November 2020	5.0%	6.7%	6.9%	2.7%										
December 2020	4.9%	6.9%	2.8%											
January 2021	5.8%	2.5%												
February 2021	2.1%													

- Earlier cohorts (Dec 2019 - May 2020) demonstrate moderate but stable behaviour.
- Customer behaviour improved significantly between June and October 2020 (strongest repurchase and retention rates).
- Although this improvement could reflect successful organisational or marketing initiatives, it may also be linked to the overall eCommerce boost at the beginning of COVID-related restrictions.
- The growth trend may have continued, as the Nov 2020 - Jan 2021 cohorts also show strong retention within the limited observation window.
- Retention within cohorts is stable over time. For example, the March 2020 cohort maintains retention levels between 1.2% and 2.8% across all fully observable months, with no significant decline. This pattern repeats across earlier and mid-2020 cohorts.
- The apparent decline of the repurchase rate in late 2020 and early 2021, as well as the decline of the retention rates in the last month across all cohorts, is therefore not a behavioural trend but a consequence of limited follow-up time, as the dataset ends in March 2021.

2b) High level Business Development

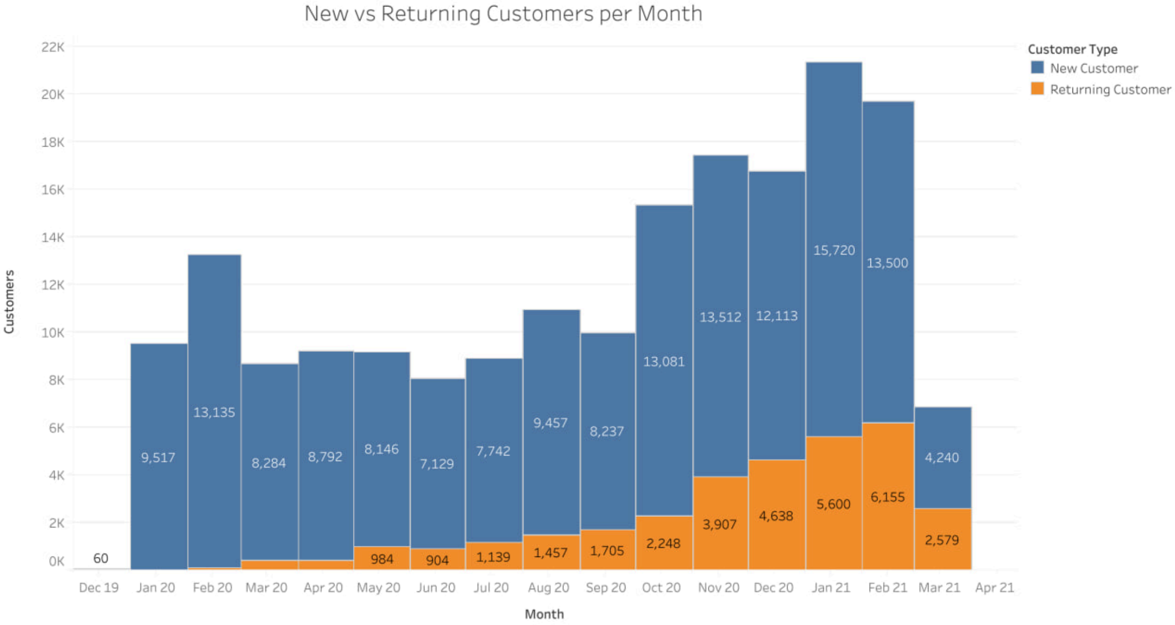
1. Revenue Trend

Revenue grows strongly throughout the period and peaks in January 2021, showing seasonal demand in the winter. The drop in March 2021 can be ignored, since the data for that month is incomplete.



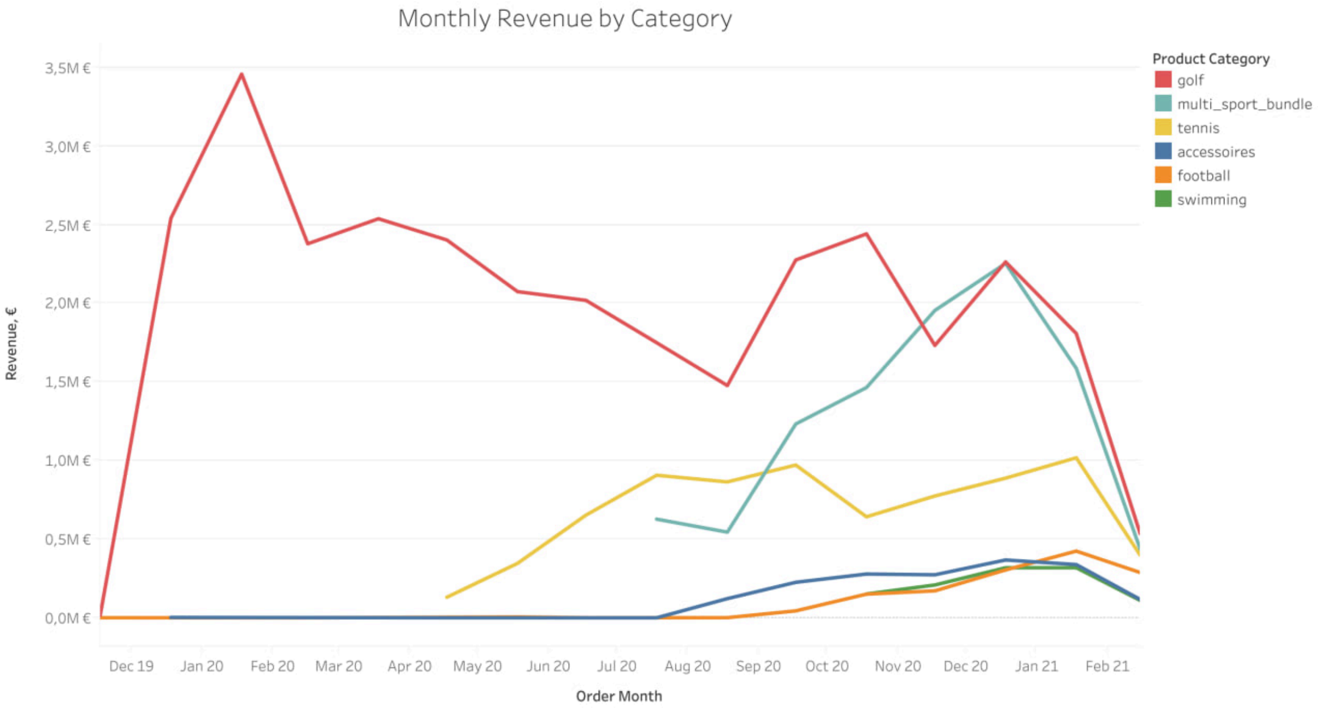
2. New vs Returning Customers

New customers drive most volume, but returning customers increase sharply from late 2020, that indicates improving loyalty and higher LTV.



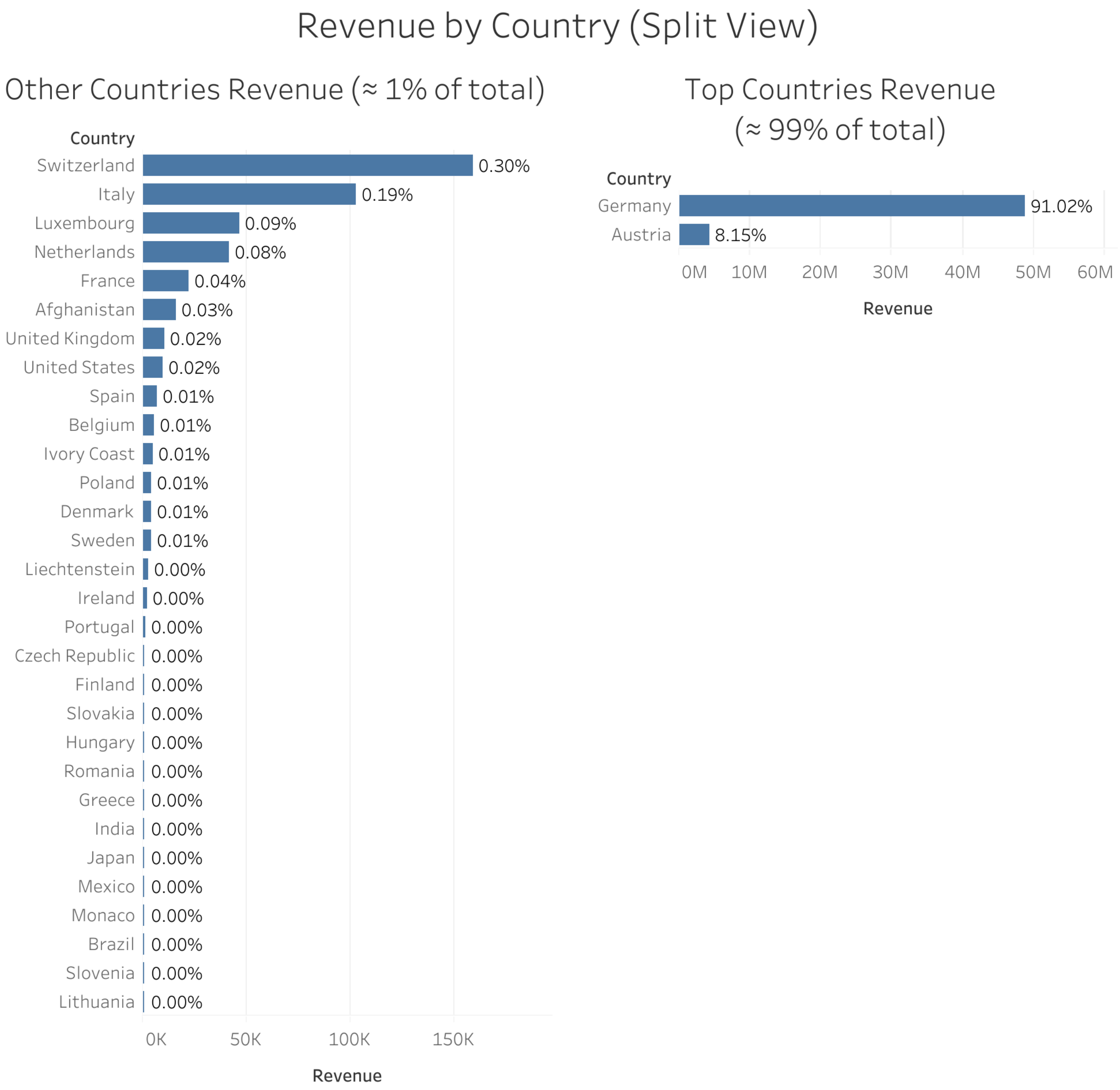
3. Product Category Performance

Golf items bring most revenue, multi-sport bungles grow rapidly in late 2020 catching up with golf in December. Other categories remain minor but stable contributors.



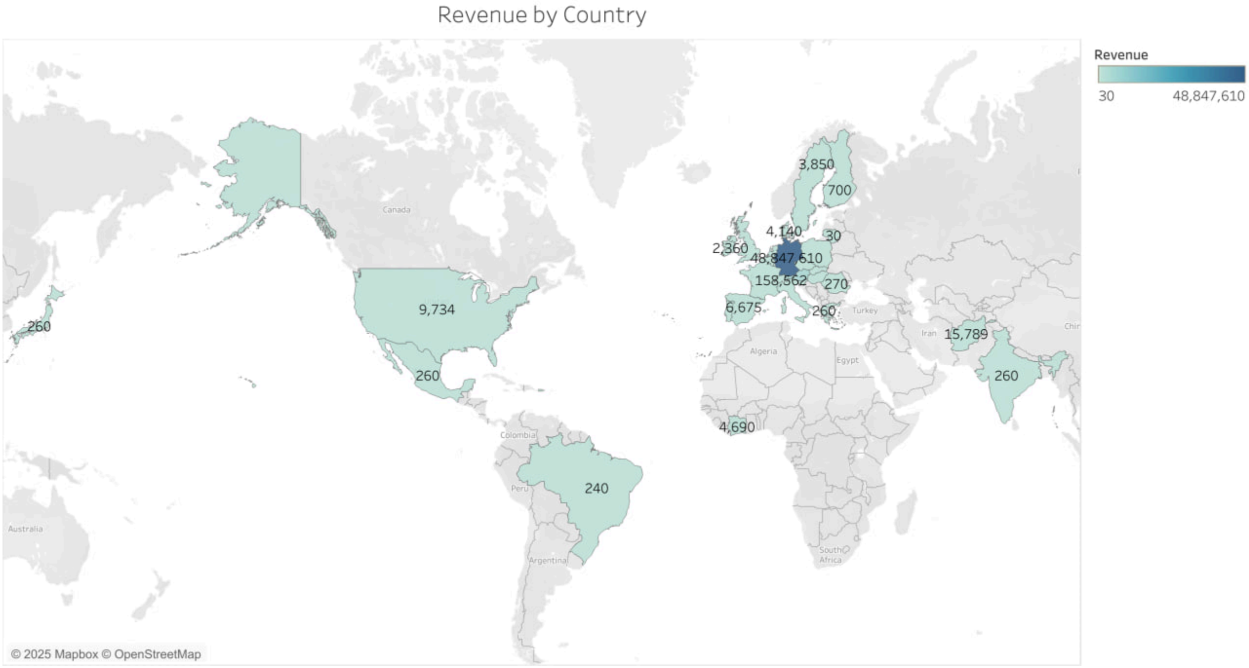
4. Revenue by Country

Germany is responsible for 91% of the revenue, Austria with 8% is the only notable secondary market. This shows strong core-market penetration but also international growth potential.



Bonus

While the country analysis could have ended on the SQL query showing 91% for Germany and 8% for Austria, I couldn't resist the temptation to build a map visualization in Tableau.


Revenue by Country
Revenue
30
48,847,610
3,850
700
4,140
30
2,360
48,847,610
158,562
270
6,675
260
260
9,734
260
15,789
260
4,690
240
Canada
Turkey
Iran
Algeria
Egypt
Colombia
Peru
Argentina
Australia
South Africa
© 2025 Mapbox © OpenStreetMap