

Mini Project Week 7 - SQL Customer Segmentation

22nd August 2025

Astrid Restu

Data

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice	InvoiceMonth
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-1 08:26:00	2.55	17850	United Kingdom	15.3	2010-12-1
536365	71053	WHITE METAL LANTERN	6	2010-12-1 08:26:00	3.39	17850	United Kingdom	20.34	2010-12-1
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-1 08:26:00	2.75	17850	United Kingdom	22	2010-12-1
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-1 08:26:00	3.39	17850	United Kingdom	20.34	2010-12-1
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-1 08:26:00	3.39	17850	United Kingdom	20.34	2010-12-1
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-1 08:26:00	7.65	17850	United Kingdom	15.3	2010-12-1
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-1 08:26:00	4.25	17850	United Kingdom	25.5	2010-12-1
536366	22633	HAND WARMER UNION JACK	6	2010-12-1 08:28:00	1.85	17850	United Kingdom	11.1	2010-12-1
536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-1 08:28:00	1.85	17850	United Kingdom	11.1	2010-12-1
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-1 08:34:00	1.69	13047	United Kingdom	54.08	2010-12-1
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-1 08:34:00	2.1	13047	United Kingdom	12.6	2010-12-1
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-1 08:34:00	2.1	13047	United Kingdom	12.6	2010-12-1
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-1 08:34:00	3.75	13047	United Kingdom	30	2010-12-1

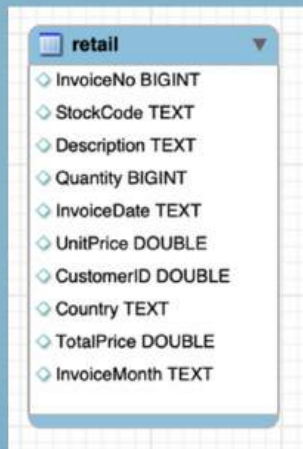
- 10 Columns
- 250,000 Rows (max display of 50,000 on MySQLWorkbench)

22nd August 2025

2

Originally over 500,000 rows in the Kaggle dataset

Schema



A screenshot of a database schema for a table named 'retail'. The table contains 10 columns with their respective data types: InvoiceNo (BIGINT), StockCode (TEXT), Description (TEXT), Quantity (BIGINT), InvoiceDate (TEXT), UnitPrice (DOUBLE), CustomerID (DOUBLE), Country (TEXT), TotalPrice (DOUBLE), and InvoiceMonth (TEXT). Each column is preceded by a small blue diamond icon.

retail	
InvoiceNo	BIGINT
StockCode	TEXT
Description	TEXT
Quantity	BIGINT
InvoiceDate	TEXT
UnitPrice	DOUBLE
CustomerID	DOUBLE
Country	TEXT
TotalPrice	DOUBLE
InvoiceMonth	TEXT

- Kaggle Link: [Online Retail Customer Segmentation Project](#)
- retail_cleaned.csv

22nd August 2025

3

Business Questions / Insights to Explore

- Geographical patterns: Which regions generate the most revenue?
- Which regions generate the least revenue?
- Seasonal trends: Identify peak purchase periods per region
- Bonus: Pricing Strategy Question: How does unit price correlate with purchase volume (Quantity)?
- Analysis: Scatter plots or regression to explore price elasticity
- RFM analysis (Recency, Frequency, Monetary):
- Identify loyal vs. at-risk customers.

22nd August 2025

4

Objective

- Learn about Customer Segmentation
- Practice SQL joins, aggregations
- Learn and Practice CTEs, Window functions, and Stored procedures in a realistic business context.

22nd August 2025

5

Geographical Patterns: Which Regions Generate the Most Revenue?

```
SELECT Country, ROUND(SUM(TotalPrice), 2)
FROM retail
GROUP BY Country
ORDER BY ROUND(SUM(TotalPrice), 2) DESC
LIMIT 10;
```

Country	ROUND(SUM(TotalPrice), 2)
United Kingdom	4212825.49
Netherlands	139844.57
EIRE	119997.15
Germany	111182.61
France	93793.28
Australia	81567.75
Spain	32624.14
Switzerland	24640.08
Japan	23041.77
Sweden	19698.88

22nd August 2025

6

Geographical Patterns: Which Regions Generate the Least Revenue?

```
SELECT Country, ROUND(SUM(TotalPrice), 2)
FROM retail
GROUP BY Country
ORDER BY ROUND(SUM(TotalPrice), 2) ASC
LIMIT 10;
```

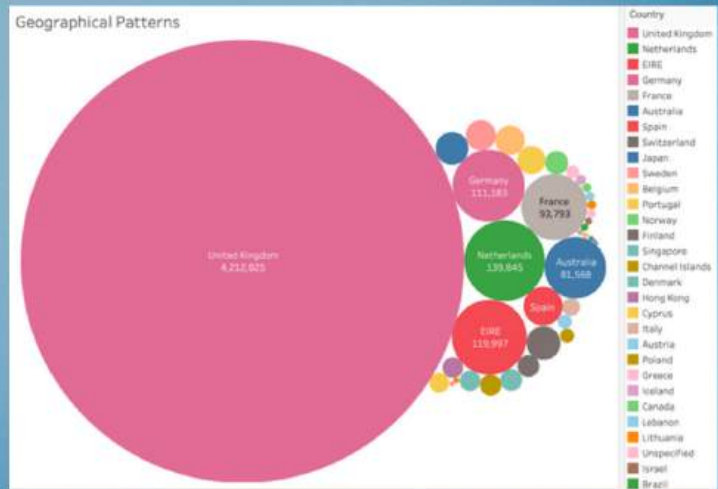
Country	ROUND(SUM(TotalPrice), 2)
Saudi Arabia	145.92
USA	383.95
Czech Republic	549.26
European Commu...	623.45
Bahrain	754.14
Malta	863.16
United Arab Emira...	889.24
Brazil	1143.6
Israel	1268.94
Unspecified	1540.75

22nd August 2025

7

Geographical Patterns: Revenue per Region

37 countries on MySQLWorkbench



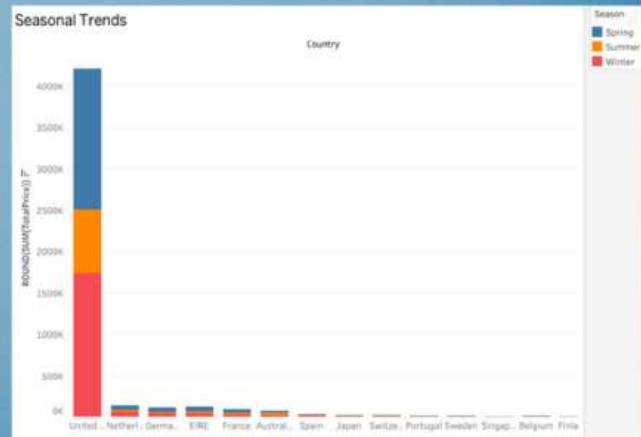
22nd August 2025

8

Seasonal Trends: Identify Peak Purchase Periods per Region

Exported the CSV to Tableau to create a stacked bar chart.

```
SELECT Country, ROUND(SUM(TotalPrice)),
CASE
WHEN MONTH(InvoiceMonth) IN (12, 1, 2) THEN 'Winter'
WHEN MONTH(InvoiceMonth) IN (3, 4, 5) THEN 'Spring'
WHEN MONTH(InvoiceMonth) IN (6, 7, 8) THEN 'Summer'
ELSE 'Fall'
END AS season
FROM retail
GROUP BY Country, season
ORDER BY Country, ROUND(SUM(TotalPrice)) DESC;
```



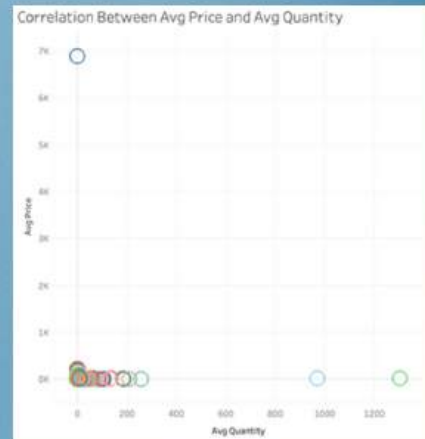
22nd August 2025

9

Bonus: Pricing Strategy Question: How Does Unit Price Correlate With Purchase Volume (Quantity)?

Calculated Pearson's correlation manually on MYSQLWorkbench

```
SELECT (
  (COUNT(*) * SUM(UnitPrice * Quantity) - SUM(UnitPrice) * SUM(Quantity)) /
  (SQRT(COUNT(*) * SUM(UnitPrice * UnitPrice) - POWER(SUM(UnitPrice), 2)) *
   SQRT(COUNT(*) * SUM(Quantity * Quantity) - POWER(SUM(Quantity), 2)))
) AS correlation_price_quantity
FROM retail;
```



22nd August 2025

10

Main Takeaways & Future Steps

- U.K. dominates purchases
- Trend of Summer downtime
- No "Fall" data
- Pearson Correlation: -0.004. Close to 0 so no clear relationship between Unit Price and Volume (Quantity)
- Scatter plots confirms this on Tableau
- RFM
- Problems - size of the kaggle datasets
- Work with product categories
- Work with attributes on customer

22nd August 2025

II

Problems - size of the kaggle datasets not having the whole picture for the seasonal trends as we only had 8 months of data, not including "Fall/Autumn".

Questions?

22nd August 2025

Astrid Restu