Large Retail Sales Dataset Analysis – 301,010 Rows:

Data Cleaning and Exploratory Data Analysis (EDA)

Project Background

The following is an analysis project undertaken based on retail transactional data that encompasses:

- Five (5) countries (USA, Germany, Australia, UK and Canada)
- One (1) year worth of data
- 159,335 customers (ranging from 18-70 years of age; 62% male, 38% female) and
- 301,010 transactions

This analysis was conducted by myself individually on the below dataset:
https://www.kaggle.com/datasets/sahilprajapati143/retail-analysis-large-dataset/data

This dataset contains numerous data points from:

- customer demographic information (name, email address, country, phone number address, gender etc) to
- purchasing information (order quantity and dollar figure totals) through to
- product category and transaction information

Page 3 of this analysis takes the reader through the data structure defined in more depth.

This large dataset was not without its errors; the pre-uploaded dataset to MySQL had numerous NULL values and blank fields, inconsistencies in column headers as well as errors in the data fields themselves more generally and formatting issues that I have addressed with the use of Excel and SQL queries to correct and standardise.

The project was planned in two (2) key phases:

1. **Data Cleaning**:
   - Conducted various cleaning and standardisation tasks in Excel and SQL (more detail on page 4) to ensure the best possible accuracy, authenticity and reliability of the analysis, in order to best practically apply its insights.

2. **Exploratory Data Analysis (EDA)**:
   This phase uses specific SQL Queries and Tableau visualisation techniques to take a deeper dive into what insights can be gleaned from the data in alignment with business metrics from a sales, customer and product standpoint (more detail page 6).

The intention is to provide sales professionals/managers and/or employees operating within a retail context, actionable recommendations that they can use to focus their efforts to be more profitable.

The analysis is intentionally focussed on these retail industry metrics/key areas of interest:

Sales Performance Insights

- What are the top revenue generating products?
- Which countries have the highest numbers of purchases?
- What is the average order value (AOV)?
- What is the average order quantity?

Customer Insights

- Do customers follow a particular: gender, age and/or income pattern or trend?
- What is the percentage split of customers by gender, income level?
- What about the average age of customers?
- How much revenue is driven by repeat vs one-off/new customers?

Product (Market) Trends

- Which products are underperforming?
- Which product categories are performing? and which are not?
- Are there particular brands performing better/worse than others?

Data Structure

Schema and Table name: retail_sales_australia.new_retail_data	
Column	**Data Types**
Transaction_ID	text
Customer_ID	text
Full_Name	text
Email	text
Phone	int
Address	text
City	text
State	text
Zipcode	int
Country	text
Age	tinyint
Gender	text
Income	text
Customer_Segment	text
Date	text
Year	text
Month	text
Time	text
Total_Purchases	int
Amount	int
Total_Amount	int
Product_Category	text
Product_Brand	text
Product_Type	text
Feedback	text
Shipping_Method	text
Payment_Method	text
Order_Status	text
Ratings	tinyint
Products	text

Executive Summary

This project achieves its intention to glean from this large data set, as many actionable recommendations and insights it can for businesses either in, or adjacent to, the retail industry to utilise in tailoring their own approaches to their marketing and sales efforts. Post cleaning the data to ensure blank fields, NULLs, inconsistencies with IDs, date data and other fields were dealt with, the appropriate SQL queries detailed below were applied with their results also shown for reporting purposes. The project has interesting findings from a sales performance, customer behaviour and product performance perspective.

From a **sales performance** perspective, the results vary from country to country however, if we take the USA and Australia as examples, a significant majority of sales come directly from bottled water grocery categorised items as well as other healthcare essentials like sports clothing, footwear or refrigerators – e.g. Pepsi branded/owned bottled water variants making up 80% of the top 20 revenue generating products in the USA and 32% in Australia.

Interestingly the project uncovers 55% of the top 20 (and 43% of *all*) overall highest spending **customers** are middle income earners and customers in the age brackets 18-25 and 41-70 represent 74% collectively of overall customers with 1 out of every 3 customers being a new customers vs a repeat customer.

Products in the grocery items and electronics categories are clearly very popular with Pepsi, Whirlpool, Coca-Cola and Adidas from an Australian perspective and Pepsi, Home Depot and Nike from a US perspective for example, being front runners as best-selling product brands.

So as retail (or adjacent to retail) organisation, they can take this data tailored specifically to their region and begin to build a profile of the kinds of customers who are grossing the most revenue: being 18-25 and 41-70 year olds and begin to understand what they are spending most on: showing a generalised trend toward grocery items and health related electronics (refrigerators) and clothing (shoes, shorts, other clothing). This information can be used to either specifically target these brands, demographics or products to continue their furthered success or acknowledge the need to address unperforming products such as niche, specific types of clothes and coffee brands that the retailers can use to cross sell other, better performing brands/items in order to increase their sell through rate or consider discontinuing these specific lines and saving on the logistics and inventory costs of these items associated with them.

Data Cleaning Tasks – Pre-Exploratory Data Analysis (EDA) Phase

Task	Notes	Tool Used
Blank fields; addressed with "Not Provided" or "N/A" tags	Overwrote the fields to minimise NULLs wherever possible	Excel filters, flash fill and lookup functions
Assigned New Unique Customer_ID To Each Full Name in order of date of transaction	Were numerous duplications - multiple customer names assigned to the one Customer_ID across the original dataset	Excel filters, flash fill and lookup functions
Assigned new Transaction_ID to Each Row in order of date of transaction	Duplications - multiple Customer IDs and names were assigned to same Transaction_ID in original dataset	Excel filters, flash fill and lookup functions
Removed suffixes to names such as MD, PHD etc	These were irrelevant to the dataset I feel and causing issues searching for names	Excel Text to columns and filters
Assigned "0/0/0000" value to all rows without a date recorded	360 rows without a date recorded for the transaction	Excel flash fill
Reformatted Phone Number Field	For better readability	Excel flash fill
Fixing of blank Amount fields		With excel calculations i.e. if no Total Amount but quantity and Amount provided made simple calculations to fill missing values Quantity * Amount = Filled Total Amount Field OR Total Amount / Quantity = Filled Blank Amount field etc
Fixing mismatched email addresses	Email addresses were not accurately aligned to names The emails could just be for example purposes, or this could also be a data security consideration, however the dataset is in the public domain so I am not sure however, nevertheless this was inconsistent so I decided to use a partial match VLOOKUP function to fix and re-align VLOOKUP(C2&"*",F:G,1,0) and now the full names and emails align	Excel - partial match lookup function and filters
Created the database table first in MySQL	Wrote a CREATE TABLE script <link>	MySQL

Used MySQL's Parallel Table Import Utility to Upload the table data via a javascript command in MySQL Shell	As the table import wizard was far too slow and inefficient while the Parallel Table Import Utility Uploaded the Data in Seconds :) i.e.: *util.importTable("C:/Users/Shane/Documents/DataPortfolioProjects2025/SalesDataProject/new_retail_data_original - Edited_ForUpload.csv", {schema: "all_retail_sales_data", table: "all_retail_sales", dialect: "csv-unix", skipRows: 1, showProgress: true})*	MySQL
Created a staging copy of the database to preserve original data state in case roll back is required	*CREATE TABLE all_retail_data_staging AS SELECT * FROM all_retail_sales;*	MySQL

Key Insights Obtained

This section will detail:

- The key insights uncovered in alignment with each business metric
- The different SQL queries used to obtain their associated reporting and
- Their resulting tables of information to be used in the Tableau dashboard visualisation

Business recommendations are included on page 25 at the conclusion of this project report.

Sales Performance Insights

- Bottled water products make up over 80% of the top 20 best-selling products in 2023 **and** 2024
- The next highest selling product not in the bottled water grocery category has over 65% less spend than its preceding bottled water product on the top 20 rankings for 2024 and over 76% less spend that it's bottled water predecessor in 2024!
- From an Australian perspective, products are more evenly split however, bottled water products still make up over 32% of the top 20 revenue generating products in the country

Top 20 Selling Products in 2024

```
# Top 20 Selling Products in 2024
SELECT
RANK() OVER(ORDER BY SUM(Total_Amount) DESC) as "Rank",
`Year`, Products, Country, Product_Type, Product_Brand, Product_Category,
FORMAT(SUM(Total_Amount), 'N','en-us') AS "Total_Spent_$"
FROM all_retail_sales
WHERE Year = 2024
GROUP BY Products, Country, Product_Type, Product_Brand, Product_Category, "Total_Spent_$"
ORDER BY SUM(Total_Amount) DESC
LIMIT 20;
```

Rank	Year	Products	Country	Product_Type	Product_Brand	Product_Category	Total_Spent_$
1	2024	Spring water	USA	Water	Pepsi	Grocery	291,554
2	2024	Alkaline water	USA	Water	Pepsi	Grocery	290,604
3	2024	Sparkling water	USA	Water	Pepsi	Grocery	289,284
4	2024	Purified water	USA	Water	Pepsi	Grocery	280,426
5	2024	Flavored water	USA	Water	Pepsi	Grocery	268,846
6	2024	Bottled water	USA	Water	Pepsi	Grocery	260,946
7	2024	Mineral water	USA	Water	Pepsi	Grocery	254,992
8	2024	Coconut water	USA	Water	Pepsi	Grocery	248,266
9	2024	Distilled water	USA	Water	Pepsi	Grocery	245,215
10	2024	Artesian water	USA	Water	Pepsi	Grocery	214,636
11	2024	Picture frames	USA	Decorations	Home Depot	Home Decor	73,549
12	2024	Cycling shorts	USA	Shorts	Nike	Clothing	67,180
13	2024	Health	USA	Non-Fiction	Random House	Books	65,912
14	2024	Gaming headphones	USA	Headphones	Sony	Electronics	64,844
15	2024	Dystopian	USA	Fiction	Random House	Books	64,428
16	2024	Curtains	USA	Decorations	IKEA	Home Decor	64,373
17	2024	Tomato juice	USA	Juice	Pepsi	Grocery	63,859
18	2024	Sony Xperia Tablet	USA	Tablet	Samsung	Electronics	63,746
19	2024	Long-sleeve tee	USA	T-shirt	Nike	Clothing	62,038
20	2024	Samsung Galaxy Tab	USA	Tablet	Apple	Electronics	60,800

Top 20 Products: Australian Context in 2024

```
# Top 20 Selling Products in Australia 2024
SELECT
    RANK() OVER(ORDER BY SUM(Total_Amount) DESC) as "Rank",
    `Year`, Products, Country, Product_Type, Product_Brand, Product_Category,
    FORMAT(SUM(Total_Amount), 'N','en-us') AS "Total_Spent_$"
FROM all_retail_sales
    WHERE Year = 2024
    AND Country = "Australia"
    GROUP BY Products, Country, Product_Type, Product_Brand, Product_Category, "Total_Spent_$"
    ORDER BY SUM(Total_Amount) DESC
    LIMIT 20;
```

Rank	Year	Products	Country	Product_Type	Product_Brand	Product_Category	Total_Spent_$
1	2024	Bottled water	Australia	Water	Pepsi	Grocery	51,990
2	2024	Coconut water	Australia	Water	Pepsi	Grocery	47,772
3	2024	Spring water	Australia	Water	Pepsi	Grocery	46,705
4	2024	French door refrigerator	Australia	Fridge	Whirlpool	Electronics	41,630
5	2024	Compact refrigerator	Australia	Fridge	Whirlpool	Electronics	41,555
6	2024	Grape juice	Australia	Juice	Coca-Cola	Grocery	41,074
7	2024	Boots	Australia	Shoes	Adidas	Clothing	41,048
8	2024	Distilled water	Australia	Water	Pepsi	Grocery	38,307
9	2024	Huawei P	Australia	Smartphone	Samsung	Electronics	37,330
10	2024	Literary fiction	Australia	Fiction	Random House	Books	36,988
11	2024	Nokia	Australia	Smartphone	Samsung	Electronics	36,111
12	2024	Crime	Australia	Thriller	HarperCollins	Books	36,105
13	2024	Romance	Australia	Fiction	Random House	Books	35,846
14	2024	Split AC	Australia	Mitsubishi AC	Mitsubhisi	Electronics	35,795
15	2024	Sparkling water	Australia	Water	Pepsi	Grocery	35,327
16	2024	Artesian water	Australia	Water	Pepsi	Grocery	35,298
17	2024	Horror	Australia	Fiction	HarperCollins	Books	35,286
18	2024	Counter-depth refrigerator	Australia	Fridge	Whirlpool	Electronics	35,235
19	2024	Varsity jacket	Australia	Jacket	Adidas	Clothing	34,750
20	2024	Cappuccino	Australia	Coffee	Nestle	Grocery	34,695

Top 20 Selling Products in 2023

```
# Top 20 Selling Products in 2023
SELECT
RANK() OVER(ORDER BY SUM(Total_Amount) DESC) as "Rank",
`Year`, Products, Country, Product_Type, Product_Brand, Product_Category,
FORMAT(SUM(Total_Amount), 'N','en-us') AS "Total Spent $"
FROM all_retail_sales
WHERE Year = 2023
GROUP BY Products, Country, Product_Type, Product_Brand, Product_Category
ORDER BY SUM(Total_Amount) DESC
LIMIT 20;
```

Rank	Year	Products	Country	Product_Type	Product_Brand	Product_Category	Total Spent $
1	2023	Artesian water	USA	Water	Pepsi	Grocery	1,401,552
2	2023	Bottled water	USA	Water	Pepsi	Grocery	1,358,759
3	2023	Flavored water	USA	Water	Pepsi	Grocery	1,318,893
4	2023	Distilled water	USA	Water	Pepsi	Grocery	1,313,325
5	2023	Sparkling water	USA	Water	Pepsi	Grocery	1,304,600
6	2023	Spring water	USA	Water	Pepsi	Grocery	1,285,632
7	2023	Coconut water	USA	Water	Pepsi	Grocery	1,271,289
8	2023	Mineral water	USA	Water	Pepsi	Grocery	1,245,027
9	2023	Alkaline water	USA	Water	Pepsi	Grocery	1,243,508
10	2023	Purified water	USA	Water	Pepsi	Grocery	1,223,470
11	2023	Grape juice	USA	Juice	Pepsi	Grocery	287,531
12	2023	Techno-thriller	USA	Thriller	HarperCollins	Books	286,033
13	2023	Top-freezer refrigerator	USA	Fridge	Whirlpool	Electronics	279,264
14	2023	Chino shorts	USA	Shorts	Nike	Clothing	272,886
15	2023	Distilled water	UK	Water	Pepsi	Grocery	270,361
16	2023	History	USA	Non-Fiction	HarperCollins	Books	265,958
17	2023	Horror	USA	Fiction	Random House	Books	264,150
18	2023	Noise-cancelling headphones	USA	Headphones	Sony	Electronics	263,167
19	2023	Orange soda	USA	Soft Drink	Pepsi	Grocery	262,352
20	2023	Orange juice	USA	Juice	Coca-Cola	Grocery	259,994

Top 20 Products by Revenue Across 2023 and 2024

```
# Top 20 Products by Revenue Across 2023 and 2024
SELECT
RANK() OVER(ORDER BY SUM(Total_Amount) DESC) as "Rank",
`Year`, Products, Product_Type, Product_Brand, Product_Category,
FORMAT(SUM(Total_Purchases), 'N', 'en-us') AS "Total Amount Purchased",
FORMAT(SUM(Total_Amount), 'N', 'en-us') AS "Total Spent $"
FROM all_retail_sales
GROUP BY `Year`, Products, Product_Type, Product_Brand, Product_Category
ORDER BY SUM(Total_Amount) DESC
LIMIT 20;
```

Rank	Year	Products	Product_Type	Product_Brand	Product_Category	Total Amount Purchased	Total Spent $
1	2023	Distilled water	Water	Pepsi	Grocery	8,577	2,189,158
2	2023	Artesian water	Water	Pepsi	Grocery	8,571	2,188,387
3	2023	Bottled water	Water	Pepsi	Grocery	8,345	2,162,330
4	2023	Flavored water	Water	Pepsi	Grocery	8,264	2,107,468
5	2023	Sparkling water	Water	Pepsi	Grocery	7,769	2,060,669
6	2023	Coconut water	Water	Pepsi	Grocery	8,136	2,042,088
7	2023	Spring water	Water	Pepsi	Grocery	8,075	2,036,179
8	2023	Alkaline water	Water	Pepsi	Grocery	7,938	2,030,635
9	2023	Mineral water	Water	Pepsi	Grocery	7,903	2,021,635
10	2023	Purified water	Water	Pepsi	Grocery	7,904	1,986,477
11	2023	French door refrigerator	Fridge	Whirlpool	Electronics	3,612	888,901
12	2023	Top-freezer refrigerator	Fridge	Whirlpool	Electronics	3,421	886,267
13	2023	Bottom-freezer refrigerator	Fridge	Whirlpool	Electronics	3,336	874,953
14	2023	Compact refrigerator	Fridge	Whirlpool	Electronics	3,466	857,336
15	2023	Window AC	Mitsubishi AC	Mitsubhisi	Electronics	3,403	854,697
16	2023	Central AC	Mitsubishi AC	Mitsubhisi	Electronics	3,297	844,992
17	2023	Stainless steel refrigerator	Fridge	Whirlpool	Electronics	3,248	844,618
18	2023	Counter-depth refrigerator	Fridge	Whirlpool	Electronics	3,349	838,452
19	2023	Side-by-side refrigerator	Fridge	Whirlpool	Electronics	3,328	828,979
20	2023	Mini fridge	Fridge	Whirlpool	Electronics	3,185	823,620

Total Products Sold Per Country 2024

```
# Total Products Sold Per Country 2024
SELECT `Year`, Country,
FORMAT(SUM(Total_Purchases), 'N', 'en-us') AS "Total Amount of Products Purchased 2023"
FROM all_retail_sales
WHERE `Year` = 2023
GROUP BY Country
ORDER BY 3 DESC;
```

Year	Country	Total Amount of Products Purchased 2023
2023	USA	425,119
2023	UK	282,192
2023	Germany	235,691
2023	Australia	202,653
2023	Canada	202,114

Total Products Sold Per Country 2023

```
# Total Products Sold Per Country 2023
SELECT `Year`, Country,
FORMAT(SUM(Total_Purchases), 'N', 'en-us') AS "Total Amount of Products Purchased 2023"
FROM all_retail_sales
WHERE `Year` = 2024
GROUP BY Country
ORDER BY 3 DESC;
```

Year	Country	Total Amount of Products Purchased 2023
2024	USA	83,202
2024	UK	56,039
2024	Germany	46,809
2024	Australia	39,892
2024	Canada	39,794

NOTE: More data is needed in order to conduct an *accurate* year on year comparison but (for example purposes) if it were to be done this is how; these numbers are reflective of not having enough of the 2024 data to hand however, further analysis of future data would be interesting to consider in follow up:

YOY Comparison – Total Amount of Products Purchased and Percentage Change

```
# Year on Year (YOY) Comparison - Total Amount of Products Purchased and Percentage Change
WITH cte1 AS (
    SELECT
        `Year`,
        Country,
        FORMAT(SUM(Total_Purchases), 'N', 'en-us') AS Total_Purchases_2023
    FROM
        all_retail_sales
    WHERE
        `Year` = 2023
        AND
        Country <> "N/A"
    GROUP BY `Year`, Country
    ),
    cte2 AS (
    SELECT
        `Year`,
        Country,
        FORMAT(SUM(Total_Purchases),'N', 'en-us') AS Total_Purchases_2024
    FROM
         all_retail_sales
    WHERE
         `Year` = 2024
         AND
         Country <> "N/A"
    GROUP BY `Year`, Country
    )
    SELECT
    cte1.`Year`,
    cte1.Country,
    cte1.Total_Purchases_2023,
    cte2.`Year`,
    cte2.Country,
    cte2.Total_Purchases_2024,
    (SELECT FORMAT(SUM(Total_Purchases),'N', 'en-us') FROM all_retail_sales) AS Overall_Total_Purchases,
    (SELECT FORMAT(((cte2.Total_Purchases_2024 - cte1.Total_Purchases_2023) /
    cte1.Total_Purchases_2023 * 100), 'N','en-us')) AS Percentage_Change_YOY
FROM
    cte1
LEFT JOIN
    cte2
ON cte1.Country = cte2.Country
ORDER BY
    cte1.Total_Purchases_2023 DESC;
```

Year	Country	Total_Purchases_2023	Year	Country	Total_Purchases_2024	Overall_Total_Purchases	Percentage_Change_YOY
2023	USA	425,119	2024	USA	83,202	1,616,757	-80
2023	UK	282,192	2024	UK	56,039	1,616,757	-80
2023	Germany	235,691	2024	Germany	46,809	1,616,757	-80
2023	Australia	202,653	2024	Australia	39,892	1,616,757	-81
2023	Canada	202,114	2024	Canada	39,794	1,616,757	-81

Average Order Value (AOV)

```
# Average Order Value (AOV)
SELECT FORMAT(AVG(Total_Amount), 'N', 'en-us') AS "Average Order Value (AOV) $"
FROM all_retail_sales;
```

Average Order Value (AOV) $
1,368

Average Order Quantity

```
# Average Order Quantity
SELECT FORMAT(AVG(Total_Purchases), 'N', 'en-us') AS "Average Order Quantity"
FROM all_retail_sales;
```

Average Order Quantity
5

Customer Insights

- 55% of the top 20 (and 43% of *overall*) spending customers are medium income earners and 43% of overall customers are medium income earners
- Customers in the age ranges of 18-25 and 41-70 represent 37% and 37% = 74% of the overall customers
- 2 out of every 3 customers do not repeat purchase

Top 20 Spending Customers Overall

```
# Top 20 Spending Customers Overall
SELECT
ROW_NUMBER() OVER(ORDER BY Total_Amount DESC) as "Rank",
Customer_ID, Full_Name, Gender, Age, Income, FORMAT((Total_Amount),'N', 'en-us') AS Total_Customer_Spend
FROM all_retail_sales
GROUP BY Customer_ID, Full_Name, Gender, Age, Income, Total_Amount
ORDER BY Total_Amount DESC
LIMIT 20;
```

Rank	Customer_ID	Full_Name	Gender	Age	Income	Total_Customer_Spend
1	C135087	Sarah Smith	Female	19	Medium	5,000
2	C051669	Erica Green	Male	26	High	4,999
3	C111697	Morgan Oconnor	Male	46	Low	4,999
4	C070579	Jesse Hughes	Male	34	Medium	4,999
5	C025510	Casey Herrera	Female	23	Medium	4,999
6	C025394	Carrie Webb	Female	69	Medium	4,998
7	C028302	Cheryl Decker	Female	43	Medium	4,998
8	C147844	Theresa Smith	Female	22	Medium	4,998
9	C146642	Teresa Hogan	Female	20	Low	4,998
10	C042443	Derek Peters MD	Female	27	Medium	4,998
11	C157493	William Carroll	Male	20	High	4,998
12	C064741	Janet Madden	Female	22	Medium	4,998
13	C072886	Joanna Reyes	Male	26	Low	4,998
14	C071634	Jessica West	Female	24	Medium	4,997
15	C065355	Jasmin Lucero	Female	20	Low	4,997
16	C074216	John Hull	Male	32	High	4,997
17	C038908	David Carey	Male	20	Medium	4,997
18	C048897	Edward Wade	Female	54	High	4,997
19	C062248	Jacob Romero	Female	55	Low	4,997
20	C081748	Karen Miller	Male	48	Medium	4,997

Percentage Split of Customers by Gender

```
# Percentage Split of Customers by Gender
SELECT
# Overall percentage of Male Customers
ROUND(((SELECT COUNT(Gender) FROM all_retail_sales WHERE Gender = 'Male')
/
(SELECT COUNT(Gender) FROM all_retail_sales) * 100),0) AS "% of Male Customers Overall",

# Overall percentage of Female Customers
ROUND(((SELECT COUNT(Gender) FROM all_retail_sales WHERE Gender = 'Female')
/
(SELECT COUNT(Gender) FROM all_retail_sales) * 100),0) AS "% of Female Customers Overall"

FROM all_retail_sales
GROUP BY Customer_ID, Transaction_ID, Gender, Age, Income, Total_Amount
ORDER BY Total_Amount DESC
LIMIT 1;
```

% of Male Customers Overall	% of Female Customers Overall
62	38

Percentage Split of Customers By Income Level

```
# Percentage Split of Customers By Income Level
SELECT
# Overall percentage of Low Income Customers
ROUND(((SELECT COUNT(Income) FROM all_retail_sales WHERE Income = 'Low')
/
(SELECT COUNT(Income) FROM all_retail_sales) * 100),0) AS "% of Customers At Low Income Bracket",

# Overall percentage of Medium Income Customers
ROUND(((SELECT COUNT(Income) FROM all_retail_sales WHERE Income = 'Medium')
/
(SELECT COUNT(Income) FROM all_retail_sales) * 100),0) AS "% of Customers At Medium Income Bracket",

# Overall percentage of High Income Customers
ROUND(((SELECT COUNT(Income) FROM all_retail_sales WHERE Income = 'High')
/
(SELECT COUNT(Income) FROM all_retail_sales) * 100),0) AS "% of Customers At High Income Bracket"

FROM all_retail_sales
GROUP BY Customer_ID, Transaction_ID, Gender, Age, Income, Total_Amount
ORDER BY Total_Amount DESC
LIMIT 1;
```

% of Customers At Low Income Bracket	% of Customers At Medium Income Bracket	% of Customers At High Income Bracket
32	43	25

Average Age of Customer

```
# Average Age of Customer
SELECT ROUND(AVG(Age),0) AS "Average Age of Customer"
FROM all_retail_sales;
```

Average Age of Customer
35

Split of Customers By Recorded Age

```
# Split of Customers By Recorded Age

(SELECT
    '18-25' AS Age_Ranges,
    FORMAT(COUNT(Customer_ID),'N', 'en-us') AS Customer_Count
FROM
    all_retail_sales
WHERE
    Age BETWEEN 18 AND 25)
UNION ALL
(SELECT
    '26-40' AS Age_Ranges,
    FORMAT(COUNT(Customer_ID),'N', 'en-us') AS Customer_Count
FROM
    all_retail_sales
WHERE
    Age BETWEEN 26 AND 40)
UNION ALL
(SELECT
    '41-70' AS Age_Ranges,
    FORMAT(COUNT(Customer_ID),'N', 'en-us') AS Customer_Count
FROM
    all_retail_sales
WHERE
    Age BETWEEN 41 AND 70
ORDER BY
    Customer_Count DESC);
```

Age_Ranges	Customer_Count
18-25	111,897
26-40	78,920
41-70	111,020

Revenue Split – New vs Repeat Customers

Total Transactions From Repeat Customers

```
# Total Transactions From Repeat Customers (More Than One Transaction)

SELECT
FORMAT((COUNT(Customer_ID) OVER())  ,'N','en-us') AS Total_Transaction_Count_From_Repeat_Customers
FROM all_retail_sales
GROUP BY Customer_ID, Full_Name, Customer_ID
HAVING COUNT(Customer_ID) > 1
ORDER BY COUNT(Customer_ID) DESC
LIMIT 1;
```

Total_Transaction_Count_From_Repeat_Customers
52,712

Total Transactions From One-Off/New Customers

```
# Total Transactions From One-Off/New Customers
SELECT
FORMAT((COUNT(Customer_ID) OVER())  ,'N','en-us') AS Total_Transaction_Count_From_One_Off_Customers
FROM all_retail_sales
GROUP BY Customer_ID, Full_Name, Customer_ID
HAVING COUNT(Customer_ID) = 1
ORDER BY COUNT(Customer_ID) DESC
LIMIT 1;
```

Total_Transaction_Count_From_One_Off_Customers
106,622

Percentages of Once Off vs Repeat Customers Over Total Customers

```
SELECT
(SELECT
FORMAT((COUNT(Customer_ID) OVER())  ,'N','en-us')
FROM all_retail_sales
GROUP BY Customer_ID, Full_Name, Customer_ID
HAVING COUNT(Customer_ID) > 1
ORDER BY COUNT(Customer_ID) DESC LIMIT 1) AS "Total Transaction Count From Repeat Customers",
(
SELECT
FORMAT((COUNT(Customer_ID) OVER())  ,'N','en-us')
FROM all_retail_sales
GROUP BY Customer_ID, Full_Name, Customer_ID
HAVING COUNT(Customer_ID) = 1
ORDER BY COUNT(Customer_ID) DESC
LIMIT 1) AS "Total Transaction Count From One Off Customers",
FORMAT( (SELECT(52712 / 159334) * 100), 'N','en-us') AS "Repeat Customers Percentage",
FORMAT( (SELECT(106622 / 159334) * 100), 'N', 'en-us') AS "Once Off Customers Percentage"
FROM all_retail_sales
LIMIT 1;
```

Total Transaction Count From Repeat Customers	Total Transaction Count From One Off Customers	Repeat Customers Percentage	Once Off Customers Percentage
52,712	106,622	33	67

Product (Market) Trends

- As a brand, Pepsi is an outlier, most pronounced in a USA context, being the:
  - Highest selling brand per unit: 172% higher than the next most purchased brand by units sold
  - Highest revenue generating brand: 168% higher than the next highest brand
- Products in the *grocery* and *electronics* categories together make up 60% of the top 20 revenue generating products
- Products with less than 10 units sold and less than 5 customer rating tend to be very niche products such as very specific decorations, types of coffee and very specific types of clothing

Unperforming Products

Products That Have Sold Less Than < 10 Units and Have Customer Rating Less Than < 5

```
# Products That Have Sold Less Than 10 Units and With Customer Rating Less Than 5
SELECT
    Products, Product_Brand, Country, Product_Type, Ratings, SUM(Total_Purchases) "Total Purchases Of Product"
FROM all_retail_sales
GROUP BY Products, Product_Brand, Country, Product_Type, Ratings, "Total Purchases"
HAVING SUM(Total_Purchases) < 10 AND Ratings < 5
ORDER BY Product_Type;
```

Products	Product_Brand	Country	Product_Type	Ratings	Total Purchases Of Product
Shower	Bed Bath & Beyond	Australia	Bathroom	0	2
Toilet paper holder	N/A	UK	Bathroom	4	5
Toothbrush holder	Bed Bath & Beyond	Australia	Bathroom	0	8
Towel rack	N/A	Canada	Bathroom	4	4
Sink	Bed Bath & Beyond	N/A	Bathroom	3	8
Blanket	Bed Bath & Beyond	N/A	Bedding	4	7
Pillowcase set	N/A	UK	Bedding	1	6
Mattress topper	Bed Bath & Beyond	Australia	Bedding	0	8
Throw pillow	Bed Bath & Beyond	N/A	Bedding	2	4
Mattress topper	N/A	Germany	Bedding	2	3
Mattress topper	N/A	USA	Bedding	2	4
Pillowcase set	N/A	USA	Bedding	2	7
Mattress topper	Bed Bath & Beyond	USA	Bedding	0	3
Cassette AC	BlueStar	Germany	BlueStar AC	3	9
Inverter AC	BlueStar	Germany	BlueStar AC	3	5
Window AC	BlueStar	Canada	BlueStar AC	3	8
Central AC	BlueStar	Australia	BlueStar AC	3	2
Mini-split AC	BlueStar	N/A	BlueStar AC	4	3
Mini-split AC	BlueStar	UK	BlueStar AC	3	2
Central AC	BlueStar	Germany	BlueStar AC	3	1

Which product categories are performing? and which are not?

```
# Top 20 Selling Categories
SELECT
    Product_Category, Country, ROUND(AVG(Ratings),0) AS "Average Product Rating",
    FORMAT(SUM(Total_Purchases), 'N', 'en-us') "Total Purchases In Product Category"
FROM all_retail_sales
GROUP BY Product_Category, Country, Ratings
ORDER BY SUM(Total_Purchases) DESC
LIMIT 20;
```

Product_Category	Country	Average Product Rating	Total Purchases In Product Category
Grocery	USA	4	49,527
Grocery	USA	3	37,540
Electronics	USA	4	34,823
Electronics	UK	4	28,232
Books	USA	4	24,816
Electronics	Germany	4	24,754
Clothing	USA	4	24,735
Home Decor	USA	4	24,406
Electronics	Canada	4	22,961
Electronics	USA	2	22,669
Grocery	USA	2	22,637
Electronics	Australia	4	22,606
Books	USA	2	22,441
Home Decor	USA	2	22,260
Clothing	USA	2	21,732
Grocery	UK	4	21,098
Clothing	UK	4	19,806
Books	UK	4	19,168
Home Decor	UK	4	18,800
Grocery	Germany	4	18,337

20 Lowest Selling Product Categories

```sql
# 20 Lowest Selling Product Categories
SELECT
    Product_Category, Country, ROUND(AVG(Ratings),0) "Average Product Rating" ,
    FORMAT(SUM(Total_Purchases), 'N', 'en-us') "Total Purchases In Product Category"
FROM all_retail_sales
GROUP BY Product_Category, Country, Ratings
ORDER BY SUM(Total_Purchases) ASC
LIMIT 20;
```

Product_Category	Country	Average Product Rating	Total Purchases In Product Category
Books	Canada	0	7
Grocery	N/A	3	8
Grocery	Canada	0	10
Clothing	Australia	0	11
Grocery	Australia	0	11
Clothing	Germany	0	12
Electronics	Canada	0	15
Books	N/A	1	19
Grocery	N/A	2	19
N/A	Germany	3	20
Grocery	Germany	0	20
N/A	Australia	1	21
N/A	Canada	3	21
N/A	USA	3	21
Books	Australia	0	24
Home Decor	UK	0	25
N/A	Canada	1	26
Home Decor	Canada	0	26
Books	N/A	5	27
Books	N/A	3	27

Are there particular brands performing better/worse than others?

Best Selling Brands By Country

```
# Best Selling Brands By Country
SELECT
    Product_Brand, Country, `Year`, FORMAT(SUM(Total_Purchases), 'N','en-us')
    "Total Purchases Of Brand In Country"
FROM all_retail_sales
GROUP BY Product_Brand, Country, `Year`
ORDER BY SUM(Total_Purchases) DESC;
```

Product_Brand	Country	Year	Total Purchases Of Brand In Country
Pepsi	USA	2023	67,422
Samsung	USA	2023	24,747
Sony	USA	2023	24,700
HarperCollins	USA	2023	24,624
Zara	USA	2023	24,546
Nestle	USA	2023	24,514
Nike	USA	2023	24,492
Penguin Books	USA	2023	24,384
Coca-Cola	USA	2023	24,341
Home Depot	USA	2023	24,307
Random House	USA	2023	24,288
Bed Bath & Be...	USA	2023	24,192
Adidas	USA	2023	23,919
Apple	USA	2023	23,684
IKEA	USA	2023	23,680
Pepsi	UK	2023	19,940
Samsung	UK	2023	18,264
Random House	UK	2023	17,970
Zara	UK	2023	17,938
Apple	UK	2023	17,911
Adidas	UK	2023	17,796
Home Depot	UK	2023	17,724
IKEA	UK	2023	17,701
Penguin Books	UK	2023	17,647
Nestle	UK	2023	17,611

Brands With Just One Product Sold and Less Than < 5 Customer Rating

```
# Brands That Have Sold Just One Unit And Have Rating Lower Than 5
SELECT
    Product_Brand, Country, Ratings, SUM(Total_Purchases) "Total Purchases Of Product"
FROM all_retail_sales
WHERE Country <> "N/A"
AND Product_Brand <> "N/A"
GROUP BY Products, Product_Brand, Country, Product_Type, Ratings, "Total Purchases"
HAVING SUM(Total_Purchases) = 1 AND Ratings < 5;
```

Product_Brand	Country	Ratings	Total Purchases Of Product
BlueStar	Germany	3	1
BlueStar	USA	3	1
Pepsi	Germany	0	1
Samsung	USA	0	1
Coca-Cola	UK	0	1
IKEA	Australia	0	1
BlueStar	Australia	3	1
Penguin Books	Germany	0	1
Zara	Australia	0	1
Nestle	UK	0	1
IKEA	USA	0	1
IKEA	USA	0	1
Random House	UK	0	1
Samsung	UK	0	1
Zara	Canada	0	1
HarperCollins	USA	0	1
Sony	Canada	0	1
IKEA	Australia	0	1
Apple	UK	0	1
Penguin Books	Australia	0	1
Apple	USA	0	1

Business Recommendations

The analysis of this data brings with it many new actionable suggestions and strategies that this analyst can now make to businesses either in, or adjacent to, other businesses in the retail industry. These include:

- Doubling down/focusing on popular brands such as Pepsi, Whirlpool, Coca-Cola and Adidas further investing in holding more stock of these items due to their clear, outlying popularity
- Utilising the popularity of these brands in cross sell campaigns with other less popular brands to increase the less popular brands' sell through rates.
  For example: US retailers could experiment with leveraging the popularity of Pepsi as a brand in selling lesser popular brands such as Nestle by doing targeted cross sell campaigns offering Pepsi products bundled with Nestle products and trial these campaigns for a measurable period, to evaluate their effectiveness and whether they will be continued
- Focusing marketing efforts on male, medium income earners in the 18-25 and 41-70 age brackets as they represent the majority of customers
- Or conversely tailoring marketing efforts to sell more to less represented markets (in this reporting) such as females and the 26–40 year age range
- Given that 2 out of every 3 customers is just a once-off buyer – retailers should also focus on campaigns around repeat purchasing e.g. offering rewards/incentives for repeat purchases in follow up transactions
- Retailers can also re-evaluate those products with less than 10 units sold and less than 5/10 customer rating e.g. those niche products such as very specific decorations, types of coffee and very specific types of clothing – and evaluate whether they can discontinue these items and reduce costs in holding inventory around these that can be freed up for re-investment back into the business or to focus on more popular products