Ecommerce Data Analysis Report

Table of Contents

Executive Summary

The Objective of this project is to analyse and visualise the sample data for an ecommerce business using BigQuery and Looker Studio to draw valuable insights and study any potential trends.

This report showcases the complete process of loading, cleaning, analysing and visualising the dataset in order to achieve the objective.

Key Performance Indicators and Trends observed in the analysis:

i. Highest sales across months for the period of three years: **January 2023**
ii. Highest average sales across months for the period of three years: **October 2024**
iii. Product Category with the highest sales value: **Sports**
iv. Top customer segment by total sales and it's percentage contribution to total sales: **Home Office, 25.4%**
v. Average number of orders per customer: **39**
vi. Region with highest sales: **Asia Pacific**
vii. Region with highest average shipping cost: **Latin America**
viii. Percentage of orders with Shipping Cost higher that $30: **43.4%**
ix. Percentage of customers who have repeated orders: **100%**
x. From the time-series plot of total sales versus months across the three years, it is observed that the total sales increases during Christmas-New Year holidays. Also the total sales decreases steadily from January to November of 2023 and 2024.

Methodology

a. Data Source

The data received in Google Spreadsheet describes the detailed sales of an Ecommerce Business for the period of January 2022 to December 2024. The original data has 14 fields and 50000 rows.

b. Loading

Data was loaded to BigQuery by converting the data file to csv format.

c. Cleaning

- Checked if OrderID, CustomerID and ProductID are null.
- Replaced OrderDate column with a new column OderDate_Cleaned with values from OrderDate converted to DATE from STRING

```
UPDATE `sample_ecommerce_data_50k.sample`
   SET OrderDate_Cleaned =
   CASE
   WHEN REGEXP_CONTAINS(OrderDate, r'[0-9]{1,2}\-[0-
9]{1,2}\-[0-9]{4}') THEN PARSE_DATE('%d-%m-%Y',OrderDate)
   WHEN REGEXP_CONTAINS(OrderDate, r'[0-9]{1,2}\/[0-
9]{1,2}\/[0-9]{4}') THEN
PARSE_DATE('%m/%d/%Y',REGEXP_EXTRACT(OrderDate, r'[0-
9]{1,2}\/[0-9]{1,2}\/[0-9]{4}'))
   ELSE NULL
   END
```

Since I don't have a paid BigQuery account, I am unable to perform DML queries on SandBox. Hence, I fixed the OrderDate column in Excel and over rid the sample table in BigQuery with the cleaned table.

- Cleaned the column ShipDate and replaced with ShipDate_Cleaned with DATE type.
- Removed 14088 records with ShipDate>OrderDate.
- Checked the data for redundant or invalid entries.

d. Tools

Analysis has been performed with the help of SQL queries on BigQuery Sandbox. Cleaning process has been performed using MS Excel.

Visualisation has been achieved on Looker Studio Report linked with BigQuery Table.

Analysis and Findings

A. Sales Performance Analysis:

Calculated the total sales for each month over the past two years. Identified the month with the highest sales. Calculated average order value (AOV) per month.

```
SELECT EXTRACT(YEAR FROM OrderDate_Cleaned) as Year,
EXTRACT(MONTH FROM OrderDate_Cleaned) as Month,
ROUND(SUM(Quantity*UnitPrice),2) as Total_Sales, SUM(Quantity) as
Total_Units_Sold, ROUND(SUM(Quantity*UnitPrice)/SUM(Quantity), 2) as
Avg_Order_Value
FROM sample_ecommerce_data_50k.sample
GROUP BY Year, Month
ORDER BY Total_Sales DESC
```

Year	Month	Total_Sales	Total_Units_Sold	Avg_Order_Value
2023	1	2958147.1	11493	257.39
2023	3	2555001.43	10146	251.82
2023	2	2539971.28	10060	252.48
2024	1	2458690.66	9784	251.3
2023	4	2426342.6	9345	259.64
2023	5	2357819.7	9440	249.77
2024	3	2253125.04	8869	254.04
2024	2	2236089.75	8730	256.14
2023	6	2230253.88	8908	250.37
2024	4	2029364.52	7787	260.61

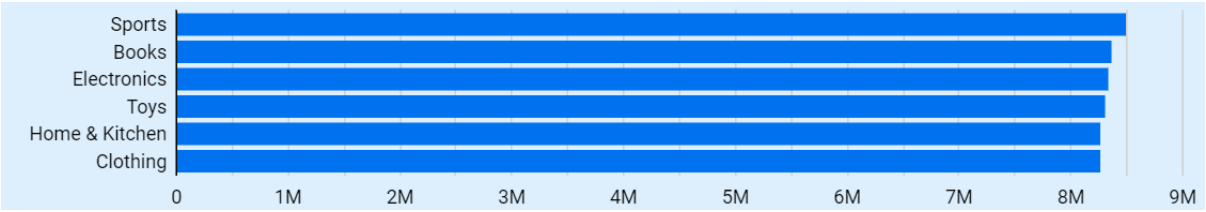
From the above query result and the time-series analysis, it is identified that highest total sales ($2,958,147.10) were observed in January 2023. Although, the highest Average Order Value ($309.48) was observed in October 2024.

B. <u>Product Performance:</u>

Breakdown of sales by product category.

```
SELECT ProductCategory, ROUND(SUM(Quantity*UnitPrice),2) as
Total_Sales, SUM(Quantity) as Total_Units_Sold,
ROUND(SUM(Quantity*UnitPrice)/SUM(Quantity), 2) as
Avg_Order_Value
FROM sample_ecommerce_data_50k.sample
GROUP BY ProductCategory
ORDER BY Total_Sales DESC
```

ProductCategory	Total_Sales	Total_Units_Sold	Avg_Order_Value
Sports	8502385.5	32985	257.77
Books	8367785.7	32958	253.89
Electronics	8338422.91	33021	252.52
Toys	8308325.55	32540	255.33
Home & Kitchen	8274953.21	33098	250.01
Clothing	8267273.25	32735	252.55



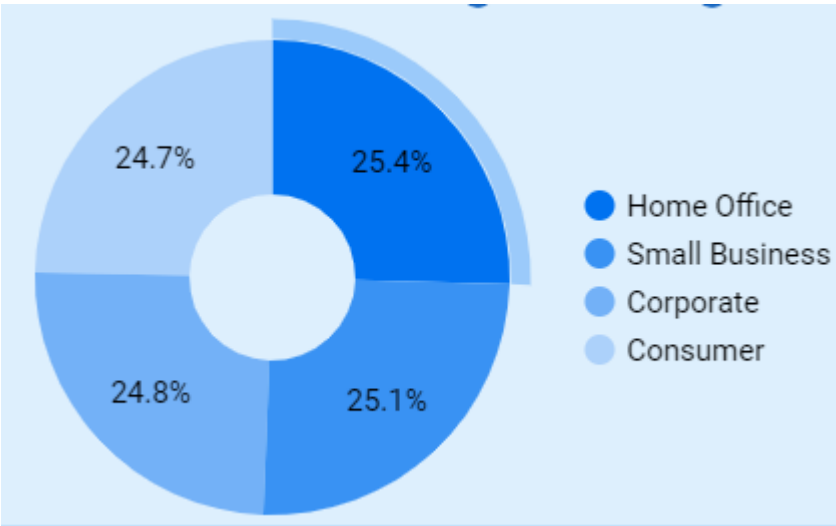
From the above query result and bar graph, it is identified that highest total sales ($8,502,385.50) as well as highest average total sales ($257.77) were observed for the Product Category Sports.

C. Customer Insights:

i. Determined the top 5 customer segments by total sales. Calculated percentage contribution of total sales for each segment.

```
SELECT CustomerSegment, TotalSales, ROUND((TotalSales /
GrandTotalSales) * 100,2) AS SalesPercentage
FROM (
  SELECT CustomerSegment, ROUND(SUM(Quantity * UnitPrice)) AS
TotalSales, SUM(SUM(Quantity * UnitPrice)) OVER () AS
GrandTotalSales
  FROM `sample_ecommerce_data_50k.sample`
  GROUP BY CustomerSegment
)
ORDER BY TotalSales DESC
LIMIT 5;
```

CustomerSegment	TotalSales	SalesPercentage
Home Office	12739650.0	25.45
Small Business	12585890.0	25.14
Corporate	12393413.0	24.76
Consumer	12340194.0	24.65



From the above query result and donut chart, it is identified that the top customer segments by total sales are Home Office, Small Business, Corporate and Consumer with their respective percentage contribution to total sales being 25.4%, 25.1%, 24.8% and 24.7%.

ii. Calculated the average number of orders per customer. Identified customers with the highest repeat order rate.

```
SELECT DISTINCT(CustomerID) as Customer_ID, COUNT(OrderID) as
NumOfOrders
FROM `sample_ecommerce_data_50k.sample`
GROUP BY Customer_ID
HAVING NumOfOrders = 10
ORDER BY NumOfOrders DESC;

SELECT COUNT(Customer_ID) as
NumOfCustomersWithHighestRepeatOrders
FROM(
SELECT DISTINCT(CustomerID) as Customer_ID, COUNT(OrderID) as
NumOfOrders
FROM `sample_ecommerce_data_50k.sample`
GROUP BY Customer_ID
HAVING NumOfOrders = 10
ORDER BY NumOfOrders DESC
)

SELECT ROUND(SUM(Quantity)/COUNT(DISTINCT(CustomerID))) AS
AvgOrdersPerCust

FROM `sample_ecommerce_data_50k.sample`;
```

Row	Customer_ID	NumOfOrders
1	c2f73679-a1c9-4db6-96b7-6ca...	10
2	61d98c0b-aa9e-4726-b559-f1e...	10
3	bbc004d9-7797-4878-80af-9e8...	10
4	5b66827d-7864-461d-abb2-19...	10
5	c056033e-c41b-45ea-bfcf-c34f...	10
6	c154870b-a83f-498c-a934-c37...	10
7	a1c2f808-3a56-44c9-8d81-5e5...	10
8	7ab1ec7d-9c7b-4547-abcf-a4a...	10
9	d6640086-fb4a-479d-8911-e2c...	10
10	1746df3c-85d7-4124-8c87-1a3...	10

NumOfCustomersWithHighestRepeatOrders
168

Row	AvgOrdersPerCust
1	39.0

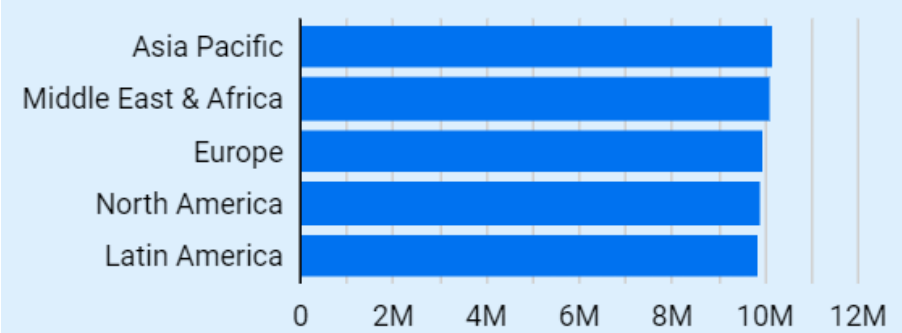
From the query results it is observed that there are 168 customers with highest number of repeated orders, i.e, 10. The average number of orders per customer is 39.

D. <u>Regional Analysis:</u>

i. Analysed sales distribution by region.

```
SELECT Region, ROUND(SUM(Quantity * UnitPrice)) AS TotalSales,
ROUND(SUM((Quantity * UnitPrice)/Quantity)) as AvgSales
FROM `sample_ecommerce_data_50k.sample`
GROUP BY Region
ORDER BY TotalSales DESC
```

Region	TotalSales	AvgSales
Asia Pacific	10192202.0	1833465.0
Middle East & Africa	10096964.0	1834799.0
Europe	9983227.0	1809725.0
North America	9934577.0	1796310.0
Latin America	9852176.0	1819808.0

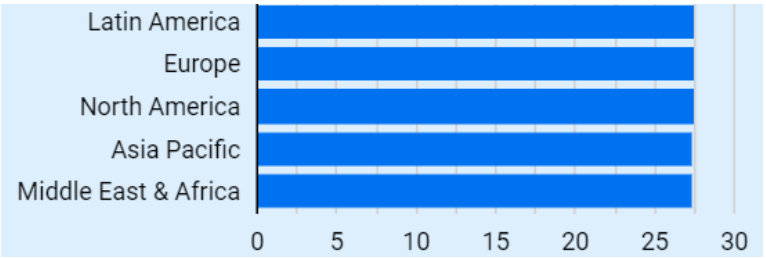


From the above query result and bar graph, it is identified that highest total sales total sales ($10,192,202) were observed for Asia Pacific Region, while the highest average total sales total sales ($1,834,799) were observed for the Middle Eastern Region.

ii. Compared the average shipping cost across different regions.

```
SELECT DISTINCT(Region), ROUND((SUM(ShippingCost)/COUNT(*)),2)
as AvgShipping
FROM `sample_ecommerce_data_50k.sample`
GROUP BY Region
ORDER BY AvgShipping DESC
```

Region	AvgShipping
Latin America	27.56
Europe	27.55
North America	27.53
Asia Pacific	27.44
Middle East & Africa	27.34



From the above query result and bar graph, it is identified that highest average shipping cost ($21.56) was observed for the Latin American Region.

E. <u>Shipping and Fulfilment Analysis:</u>

i. Calculated the average time taken to ship orders (difference between order date and ship date).

```
SELECT MAX(DATE_DIFF(ShipDate_Cleaned, OrderDate_Cleaned, DAY))
as MaxDiffBtwOrderAndShippingDates,
MIN(DATE_DIFF(ShipDate_Cleaned, OrderDate_Cleaned, DAY)) as
MinDiffBtwOrderAndShippingDates, AVG(DATE_DIFF(ShipDate_Cleaned,
OrderDate_Cleaned, DAY)) as AvgDiffBtwOrderAndShippingDates
FROM `sample_ecommerce_data_50k.sample`
```

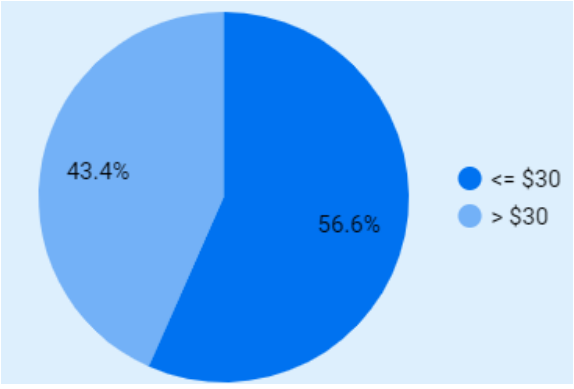
MaxDiffBtwOrderAndShippingDates	MinDiffBtwOrderAndShippingDates	AvgDiffBtwOrderAndShippingDates
353	0	67.1371129427491

It is observed that the maximum, minimum and average number of days to ship orders is 353, 0 and 67.13, respectively.

ii. Determined the percentage of orders with shipping costs greater than $30.

```
SELECT COUNTIF(ShippingCost > 30) / COUNT(*) * 100 AS
PercentageOver30
FROM `sample_ecommerce_data_50k.sample`
```

PercentageOver30
44.46146134996...



As is evident from the query result and the pie chart, 43.4% orders have shipping charges greater than $30.

Due to the data being a random sample, no trends could be observed clearly. But the reasons could be due to holidays, period of the month or region.

F. <u>Customer Retention:</u>

i. Identify customers who have placed more than one order. What percentage of total customers are repeat customers?

```
SELECT (COUNT(Customer_ID)*100/totalCustomers) AS
percentRepeatedCustomers
FROM (
SELECT DISTINCT(CustomerID) as Customer_ID, COUNT(OrderID) as
NumOfOrders, COUNT(CustomerID) OVER() as totalCustomers
FROM `sample_ecommerce_data_50k.sample`
GROUP BY Customer_ID
)
WHERE NumOfOrders > 1
GROUP BY totalCustomers
```

percentRepeatedCustomers
100.0

As is evident from the query result, the percentage of customers with repeated orders is 100%, i.e., all customers included in the given database have placed orders more than once with the ecommerce website.

ii. Calculate the average time between repeat orders for these customers.

```
WITH OrderedCustomerData AS (
    SELECT CustomerID, OrderDate_Cleaned, LAG(OrderDate_Cleaned)
OVER (PARTITION BY CustomerID ORDER BY OrderDate_Cleaned) AS
PreviousOrderDate
    FROM `sample_ecommerce_data_50k.sample`
),
OrderDifferences AS (
    SELECT CustomerID, DATE_DIFF(OrderDate_Cleaned,
PreviousOrderDate, DAY) AS DaysBetweenOrders
    FROM OrderedCustomerData
```

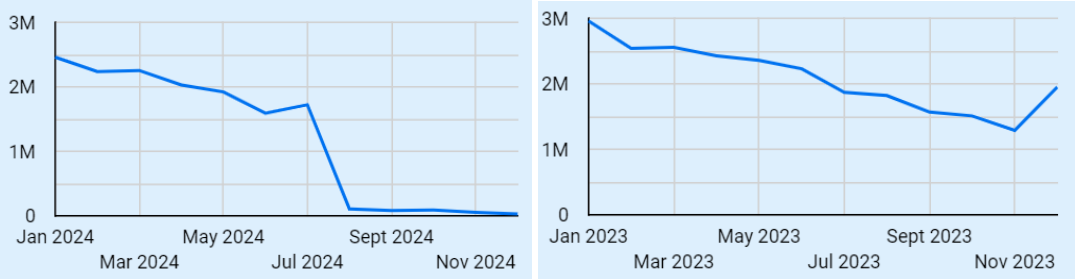
```
    WHERE PreviousOrderDate IS NOT NULL
)
SELECT CustomerID, AVG(DaysBetweenOrders) AS
AvgDaysBetweenOrders
FROM OrderDifferences
GROUP BY CustomerID
ORDER BY AvgDaysBetweenOrders;
```

Row	CustomerID	AvgDaysBetweenOrders
1	96850989-3af1-4228-b24a-1cc...	24.5
2	1ca67af7-9124-45ff-b36e-ecb1...	28.0
3	6ad612fd-3cb1-45d1-85c1-c25...	30.999999999999996
4	93648b9d-e857-478d-ac6e-2bb...	31.0
5	66f214e0-a00a-439f-8b53-7cfb...	31.666666666666664
6	a0223481-9a8f-4b1e-95f5-8c6...	31.75
7	55459bc8-ec3f-4a0d-bc07-743...	31.833333333333332
8	336e96f7-78a0-41c8-913e-9d9...	31.875
9	0d73af65-853f-4832-ab3b-3e2...	32.571428571428577
10	22b193a7-cc11-4ad4-a61e-286...	32.666666666666664

As identified from the query result, the average number of days between orders for customers varies from 24.5 to 443.

G. Time Series Analysis:

Created a time series plot of monthly sales.



Using the plot it has been observed that the highest peaks are achieved in January 2023 and January 2024. This might indicate an increased sales during the Christmas-New Year holidays.

Also, the total sales decreases steadily from January to November of 2023 as well as 2024.

Recommendations

- **Actionable Steps**: Introducing seasonal promotions, flash sales, loyalty programs, discounts on complimenting products, referral programs, marketing campaigns, collaborations, building communities to increase sales.
- **Strategic Implications**: The strategies mentioned above can potentially attract more customers round the year.