## 1. Understanding of Data

The data itself is a random sample of a bigger dataset which is Google's e-commerce dataset. The content of the dataset itself is record of every user's session and actions within those sections in each day ranging from 1 August 2016 to 1 August 2017. The data is typical of what you would see for an ecommerce website. It includes the following kinds of information:

- **Traffic source data:** information about where website visitors originate. This includes data about organic traffic, paid search traffic, display traffic, etc.
- **Content data:** information about the behavior of users on the site. This includes the URLs of pages that visitors look at, how they interact with content, etc.
- **Transactional data:** information about the transactions that occur on the Google Merchandise Store website.

## 2. Data Quality Issues

I indeed noticed that there are data quality issues in the provided random sample dataset that is named data-to-insights.ecommerce.all_sessions. The issues of the dataset are the followings:

a. There are typos in the product name which both have the same product SKU number, here are the example of that.

| | | | | |
|---|---|---|---|---|
| 3 | GGOEGBMC056599 | Waterproof Gear Bag | 159.98 | No Refund |
| 4 | GGOEGBMC056599 | Waterpoof Gear Bag | 79.99 | No Refund |

This makes the revenue analysis harder to analyze since there are 2 different products with the same product SKU number even though it should be the same product.

b. There are different products that are coded by the same product SKU number, here are the example of that.

| | | | | |
|---|---|---|---|---|
| 5 | GGOEGBJR018199 | Chevron Shopper | 90.3 | No Refund |
| 6 | GGOEGBJR018199 | Reusable Shopping Bag | 90.3 | No Refund |

This can prove to be fatal since it could heavily affect inventory management, order fulfillment, system integration issues, potential customer confusion because of wrong order, and many more.

c. There are many transactions without the transactionId. Transactions without the transactionId makes the transaction analysis on the dataset more difficult since the revenue can't be traced down to the transaction specificity.

## 3. Test Case Data Analysis

On this section I will be providing data analysis based on all of the test case queries. I will first show the output of the query, then put analysis after it.

### a. Test Case 1: Channel Analysis

| | country character varying (255) | channelgrouping character varying (255) | totalrevenue double precision |
|---|---|---|---|
| 1 | Canada | Referral | 3204.46 |
| 2 | Canada | Affiliates | 0 |
| 3 | Canada | Organic Search | 4719.76 |
| 4 | Canada | Paid Search | 0 |
| 5 | Canada | Display | 0 |
| 6 | Canada | Social | 0 |
| 7 | Canada | (Other) | 0 |
| 8 | Canada | Direct | 218.34 |
| 9 | Curaçao | Organic Search | 208.33 |
| 10 | Taiwan | Affiliates | 0 |
| 11 | Taiwan | Referral | 797.1 |
| 12 | Taiwan | Organic Search | 0 |
| 13 | Taiwan | Direct | 0 |
| 14 | Taiwan | Social | 0 |
| 15 | Taiwan | Paid Search | 0 |
| 16 | United States | Referral | 55966.54 |
| 17 | United States | Affiliates | 0 |
| 18 | United States | Paid Search | 4190.91 |
| 19 | United States | Direct | 13878.2 |
| 20 | United States | Social | 464.66 |
| 21 | United States | Display | 8497.86 |
| 22 | United States | Organic Search | 27443.82 |
| 23 | Venezuela | Affiliates | 0 |
| 24 | Venezuela | Direct | 92.5 |
| 25 | Venezuela | Social | 0 |
| 26 | Venezuela | Organic Search | 9952.16 |

Based on the output I can gain the following **insights**:

- Canada has the most diverse channel source with 8 different channel grouping, followed by United States with 7, and Taiwan with 6
- Organic search channel have been the most profitable for Canada with Referall and Direct channel followed and the rest of the channel group contribute none for profit

- Customers from Curacao only originate from organic search channels with $208.33 revenue.
- Referral have been the most and only profitable channel source in Taiwan, there are 5 other channels but none of them bring any profit.
- Referral have been the most profitable channel source in the United States, followed by Organic Search, Direct, Display, Paid Search, and Social followed with the rest of the channel contribute nothing to profit.
- Organic have been the most profitable channel source in Venezuela, followed by Direct channel with the rest of the channel contribute nothing to profit.

Based on the insights I have the following **recommendations**:

**Optimize Marketing Strategies:**

For Canada, since organic search is the most profitable channel, consider investing more in SEO and content marketing to enhance organic search visibility. In Taiwan, focus efforts on referral strategies since it's the only profitable channel. Evaluate and improve referral programs to maximize their impact. The rest should be treated with similarity like Canada and Taiwan

**Review and Adjust Channel Allocation:**

In countries where specific channels are not contributing to profit, evaluate whether resources allocated to those channels can be better used elsewhere. Consider reallocating budget and efforts to more profitable channels.

b. **Test Case 2: User Behavior Analysis**

| | | | | | |
|---|---|---|---|---|---|
| 46 | 479908255003687540 | 1452 | 9 | 0 | Flagged |
| 47 | 7298336494740484350 | 3529 | 107 | 0 | Not Flagged |
| 48 | 4007203671996544263 | 256 | 16 | 0 | Not Flagged |
| 49 | 3343749549580907329 | 0 | 1 | 0 | Not Flagged |
| 50 | 2437897518770727158 | 1658 | 31 | 0 | Not Flagged |
| 51 | 3271421739508344638 | 184 | 8 | 0 | Not Flagged |
| 52 | 5757867978141494279 | 132 | 12 | 0 | Not Flagged |
| 53 | 3885196846909777802 | 186 | 5 | 0 | Not Flagged |
| 54 | 1629521602007682337 | 0 | 1 | 0 | Not Flagged |
| 55 | 5756725854221787 | 536 | 13 | 0 | Not Flagged |
| 56 | 1748978919655397239 | 2364 | 22 | 0 | Flagged |
| 57 | 3061338030780455732 | 65 | 4 | 0 | Not Flagged |
| 58 | 9932791336038034651 | 97 | 4 | 0 | Not Flagged |
| 59 | 8983098566666686138 | 300 | 10 | 0 | Not Flagged |
| 60 | 5432060363704066851 | 275 | 28 | 0 | Not Flagged |
| 61 | 5073655045946074433 | 310 | 4 | 0 | Not Flagged |
| 62 | 919292755482493022 | 340 | 8 | 0 | Not Flagged |
| 63 | 855265669670534854 | 329 | 16 | 0 | Not Flagged |
| 64 | 5803052831879556332 | 106 | 8 | 0 | Not Flagged |
| 65 | 1404718434462154192 | 1143 | 14 | 0 | Flagged |

(snippet of the ouput of the query)

Based on the snippet of the ouput of the query, I can gain the following **insights**:

- There are users who spends a lot of time idling in the website proven by their above average timeOnSite but lower than average website interactions
- Many of sessionQualityDim columns are not calculated and results in 0 values even though sessionQualityDim can be a measure of user session's quality

Based on the insights I have the following **recommendations**:

**User Experience Enhancement:**

Improve the overall user experience to increase engagement. This could include optimizing website navigation, simplifying the user interface, and ensuring that content is easily accessible and appealing.

**A/B Testing:**

Conduct A/B testing on different engagement strategies to identify the most effective approaches. Test variations in content, calls-to-action, and user interface elements to optimize for higher engagement.

**Session Quality Calculation:**

Investigate why many sessionQualityDim columns are not calculated and result in 0 values. Ensure that the calculation methodology is accurate and aligns with the definition of session quality. If needed, update the calculation method to provide more meaningful insights into user sessions.

c. **Test Case 3: Product Performance**

| | productsku character varying (255) | v2productname character varying (255) | totalquantitysold double precision | netrevenue double precision | totalrefund double precision | refundflag text |
|---|---|---|---|---|---|---|
| 1 | GGOEGOCD078399 | Google Leather Perforated Journal | 27 | 989.73 | 0 | No Refund |
| 2 | GGOEGBMC056599 | Waterpoof Gear Bag | 3 | 239.96999999999997 | 0 | No Refund |
| 3 | GGOEGBMC056599 | Waterproof Gear Bag | 3 | 239.96999999999997 | 0 | No Refund |
| 4 | GGOEGBMJ013399 | Sport Bag | 44 | 219.56000000000003 | 0 | No Refund |
| 5 | GGOEGBJR018199 | Chevron Shopper | 26 | 90.3 | 0 | No Refund |
| 6 | GGOEGBJR018199 | Reusable Shopping Bag | 26 | 90.3 | 0 | No Refund |

(top 5 high performing products)

Based on the sample (top 5 high performing products) of the output I can gain the following **insights**:

- The top 5 high performing products are Google Leather Perforated Journal, Waterproof Gear Bag, Sport Bag, Chevron Shopper, and Reusable Shopping Bag
- There are no recorded product refunds in the data

- There are data quality issues where 2 different products have the same product SKU number and there are supposedly the same product counted to be different because of typo in the name

Based on the insights I have the following **recommendations**:

**Promotion and Upselling:**

Leverage top performing product's popularity to implement targeted promotions and upselling strategies. Consider bundling these products with complementary items or offering exclusive deals to encourage customers to purchase more.

**Refund Tracking:**

Although there are no recorded product refunds in the data, it's important to have a robust refund tracking system in place. Implement a process to record and analyze product refunds.

**Quality Assurance Checks:**

Implement regular quality assurance checks for product data to catch and correct any discrepancies early. This includes validating SKU uniqueness, checking for consistent naming conventions, and ensuring data accuracy.