DATA QUALITY

Question 1) Are the costs in the 'api_adwords_costs' table fully covered in the 'session_sources' table? Any campaigns where you see issues?

```sql
-- This query outputs a report on campaign_name, total_adwords_cost, total_session_cpc & the discrepancy between the two
tables' costs. Running each section of the query will allow for more in dept analysis and finding issues' root causes.
SELECT
CAMPAIGN_NAME,
ROUND(SUM(TOTAL_ADWORDS_COST), 2) AS "TOTAL_ADWORDS_COST",
ROUND(SUM(TOTAL_SESSION_CPC), 2) AS "TOTAL_SESSION_CPC",
ROUND(SUM(DIFF), 2) AS "DISCREPANCY"

FROM
-- This section outputs costs and CPC from API & session table, grouped by event_date & campaign_date
(SELECT
A.EVENT_DATE,
A.CAMPAIGN_ID,
A.TOTAL_ADWORDS_COST,
S.TOTAL_SESSION_CPC,
(A.TOTAL_ADWORDS_COST - S.TOTAL_SESSION_CPC) AS "DIFF",
S.CAMPAIGN_NAME

FROM
-- This section summarises costs from api_adwords_cost
(SELECT
EVENT_DATE,
CAMPAIGN_ID,
SUM(COST) AS TOTAL_ADWORDS_COST

FROM API_ADWORDS_COSTS
GROUP BY EVENT_DATE, CAMPAIGN_ID) A

LEFT JOIN
-- This section summarises CPC in session_sources
(SELECT
EVENT_DATE,
CAMPAIGN_ID,
CAMPAIGN_NAME,
SUM(CPC) AS TOTAL_SESSION_CPC

FROM SESSION_SOURCES
GROUP BY EVENT_DATE, CAMPAIGN_ID) S

ON A.EVENT_DATE = S.EVENT_DATE AND A.CAMPAIGN_ID = S.CAMPAIGN_ID)

WHERE DIFF <> 0
GROUP BY CAMPAIGN_NAME
ORDER BY DIFF DESC
```

CAMPAIGN_NAME	TOTAL_ADWORDS_COST	TOTAL_SESSION_CPC	DISCREPANCY	
campaign_name_741	304,172.87	291,747.40	12,425.47	Significant Discrepancy
campaign_name_565	2,772.04	2,766.26	5.78	Small Discrepancy
campaign_name_203	52,592.02	52,586.93	5.09	
campaign_name_589	15,085.78	15,082.39	3.39	
campaign_name_142	3,736.28	3,735.67	0.61	
campaign_name_782	31,320.04	31,319.74	0.31	
campaign_name_181	15,215.20	15,214.93	0.28	
campaign_name_1122	4,527.76	4,527.57	0.19	Zero or Near-Zero Discrepancy
campaign_name_525	5,214.84	5,214.65	0.19	
campaign_name_609	1,072.70	1,072.68	0.02	
campaign_name_54	10.73	10.72	0.01	
campaign_name_366	665.51	665.508	0.002	
campaign_name_481	1,542.94	1,542.95	-0.01	
campaign_name_314	94.65	94.66	-0.01	
campaign_name_356	439.82	439.91	-0.09	Negative Differences
campaign_name_330	11,693.78	11,693.89	-0.11	
campaign_name_150	3,133.92	3,134.54	-0.62	

Takeaways:

1 - Significant Discrepancy:
campaign_name_741 has a large positive difference of 12,425.47 which indicates a substantial discrepancy. This could be due to several possible reasons such as:
- Over-reporting in the api_adwords_costs table.

- Under-reporting or missing session data in the session_sources table.
- Timing issues where costs are recorded differently in each table.

2- Small Discrepancy:
campaign_name_565, **campaign_name_203** and **campaign_name_589** have relatively smaller discrepancies which can be due to:
- Rounding errors.
- Slight timing discrepancies.
- Minor data entry inconsistencies.

3 - Zero or Near-Zero Discrepancy:
There are campaigns (ex: campaign_name_142) with minimal discrepancies which suggest a good data consistency between the two tables.

4 - Negative Differences:
Finally, campaigns with negative differences in costs suggest the "session_sources" table has slightly higher CPC costs than the "api_adwords_costs" table. This could be due to:
- Over-reporting in the session_sources table.
- Under-reporting or missing cost data in the api_adwords_costs table.
- Timing issues or slight mismatches in how costs are aggregated.

Question 2) Are the conversions in the 'conversions' table stable over time? Any pattern?
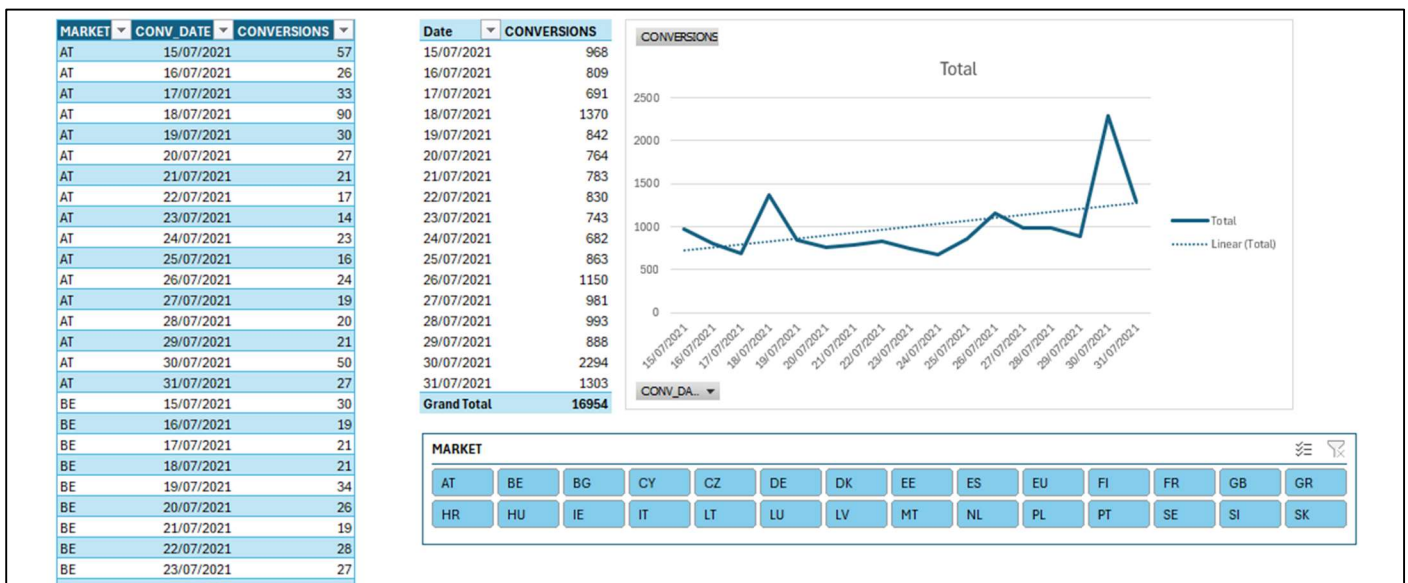
Takeaways:

The e-commerce conversions in July 2021 show considerable **instability** with **large fluctuations**. There are no clear, stable patterns, suggesting that external factors, promotional activities, or specific events may have caused these variations.

However, on **July 30, 2021**, a spike can be observed which could suggest a special event, promotion, or campaign. This increment is also visible across major markets such as Germany, United Kingdom and France. During this period, there is a noticeable upward trend in conversions, represented by a linear line. This trend indicates an increase in conversions as we approach the end of the month.

Statistical Summary:

- Count of Days: **17**
- Mean (Average) Conversions: **997**
- Standard Deviation: **388.43** (High SD suggests a considerable fluctuation in daily conversions.)
- Minimum Conversions: **682** (on July 24, 2021)
- 25th Percentile (Q1): **783**
- Median (Q2): **863**
- 75th Percentile (Q3): **993**
- Maximum Conversions: **2,294** (on July 30, 2021)

Questions 3) Double check conversions ('conversions' table) with backend ('conversions_backend' table), any issues?

Upon initial inspection, the conversions table contains **345** fewer unique Conversion IDs compared to the conversions_backend table which indicates discrepancy between the two tables.

Additionally, to simplify the process of cross-checking tables, we can develop a Python script to identify discrepancies and provide a summary:

```python
import sqlite3
import pandas as pd

# Connect to the SQLite database
db_path = r'C:\Users\Satar\OneDrive\Python Projects\challenge.db'
conn = sqlite3.connect(db_path)

# Load the conversions and conversions_backend tables into pandas DataFrames
conversions_df = pd.read_sql_query("SELECT * FROM conversions", conn)
conversions_backend_df = pd.read_sql_query("SELECT * FROM conversions_backend", conn)

# Close the database connection
conn.close()

# Ensure the data types are consistent between the two DataFrames
conversions_df['revenue'] = conversions_df['revenue'].astype(float)
conversions_backend_df['revenue'] = conversions_backend_df['revenue'].astype(float)

# Merge the two DataFrames on conv_id to compare them
merged_df = conversions_df.merge(conversions_backend_df, on='conv_id', suffixes=('_conv', '_backend'))

# Create a dictionary to store the number of discrepancies for each column
discrepancy_count = {
    'user_id': 0,
    'conv_date': 0,
    'market': 0,
    'revenue': 0
}

# Identify discrepancies between the two tables
for column in discrepancy_count.keys():
    discrepancies = merged_df[merged_df[f'{column}_conv'] != merged_df[f'{column}_backend']]
    discrepancy_count[column] = discrepancies.shape[0]

# Print the number of discrepancies for each column
print("Number of discrepancies between each identical column:")
for column, count in discrepancy_count.items():
    print(f"{column}: {count}")

# List all discrepancies
all_discrepancies = merged_df[(merged_df['user_id_conv'] != merged_df['user_id_backend']) |
                              (merged_df['conv_date_conv'] != merged_df['conv_date_backend']) |
                              (merged_df['market_conv'] != merged_df['market_backend']) |
                              (merged_df['revenue_conv'] != merged_df['revenue_backend'])]

print("\nDiscrepancies found:")
print(all_discrepancies)

# Save the discrepancies to a CSV file
all_discrepancies.to_csv('discrepancy_summary.csv', index=False)
```

Result:

Question 4) Are attribution results consistent? Do you find any conversions where the 'ihc' values don't make sense?

After applying a threshold of **0.001** to account for floating-point precision errors, the analysis revealed that only **2%** of the conversions have inconsistent ihc values, where the sum of ihc values does not fall within the acceptable range of 0.999 to 1.001. These minor discrepancies are expected due to the nature of floating-point arithmetic in computer calculations.

The remaining inconsistencies suggest potential issues in the attribution data, likely because of errors in data collection or flaws in the attribution model.

To address this, it is recommended to implement precise data validation rules, review and correct the attribution model, and conduct a root cause analysis to identify and rectify the sources of these inconsistencies.

From the Power BI dashboard report, we can also observe this analysis by referring to visualizations of the data:
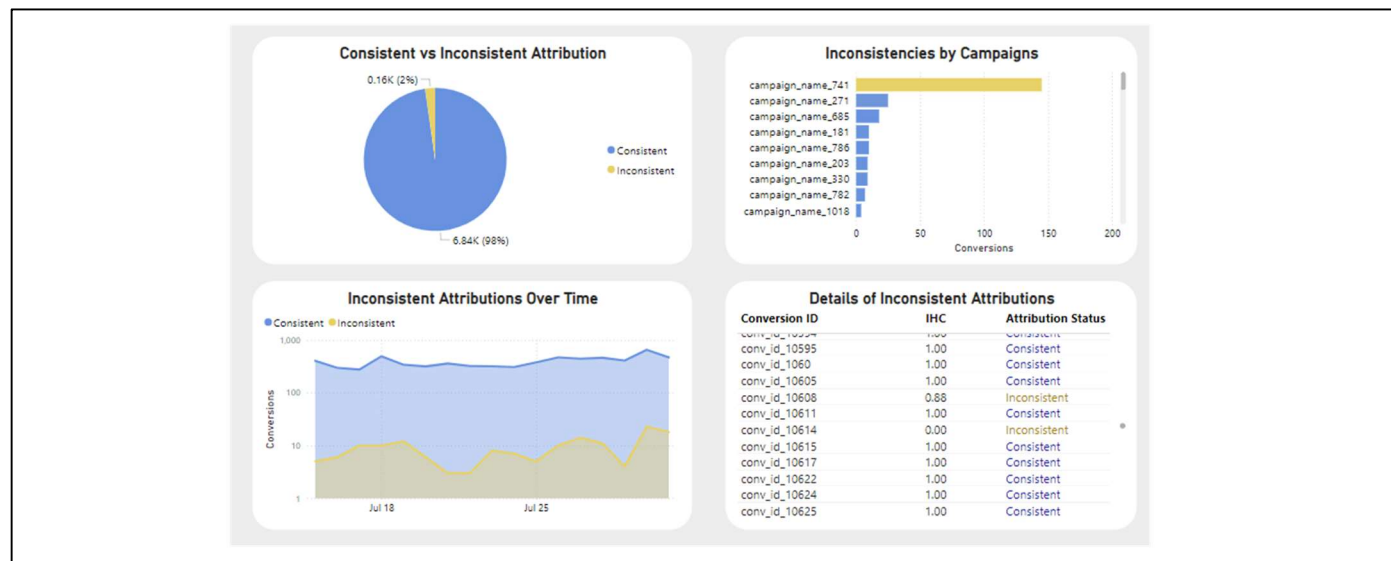
Question 5) (Bonus) Do we have an issue with channeling? Are the number of sessions per channel stable over time?

The analysis revealed that while most channels showed stable trends with minor fluctuations, a few channels exhibited significant variability or declining trends. These outliers indicate potential issues with specific channels that may require further investigation to ensure a balanced and effective channeling strategy.

Direct Traffic, **Google Remarketing**, **Referral**, **SEA – Branded** and **SEO** channels are among the **Stable Trends with Weekly Fluctuations**, while **Email**, **Influencers**, **SEA - Non-branded** are among the Unstable/Downward Trends with Significant Fluctuations. A deep dive on the Influencers channel, for instance, suggests changes in partnerships and campaign effectiveness, necessitating data verification and strategic adjustments.
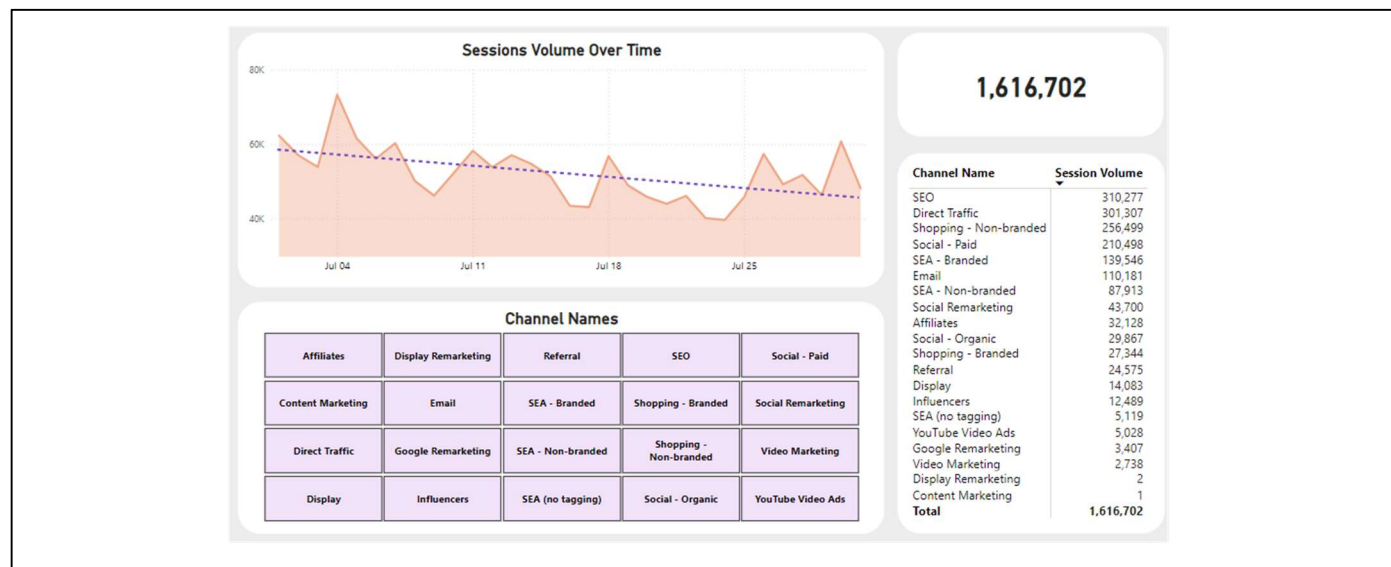
Additionally, there seems to be an issue with data entry for the channel names, in which they are not consistent and it's possible to combine some channel names together to better represent the channel name. for example, "Affiliate" & "Affiliates" could be combined as a single channel. (Already adjusted in above report.)

Question 6) (Bonus) Any other issues?

Other issues identified include potential data consistency and integrity problems such as missing values, duplicate records, and discrepancies in revenue data between conversions and conversions_backend. Additionally, there may be inconsistencies in the date and time fields.

To improve the system, it's recommended to enhance data validation rules, conduct regular audits, adopt more advanced attribution models, and ensure accurate real-time data synchronization. Also, it's a good idea to have monitoring tools (such as PBI dashboards and python scripts) to regularly look for any potential issues.