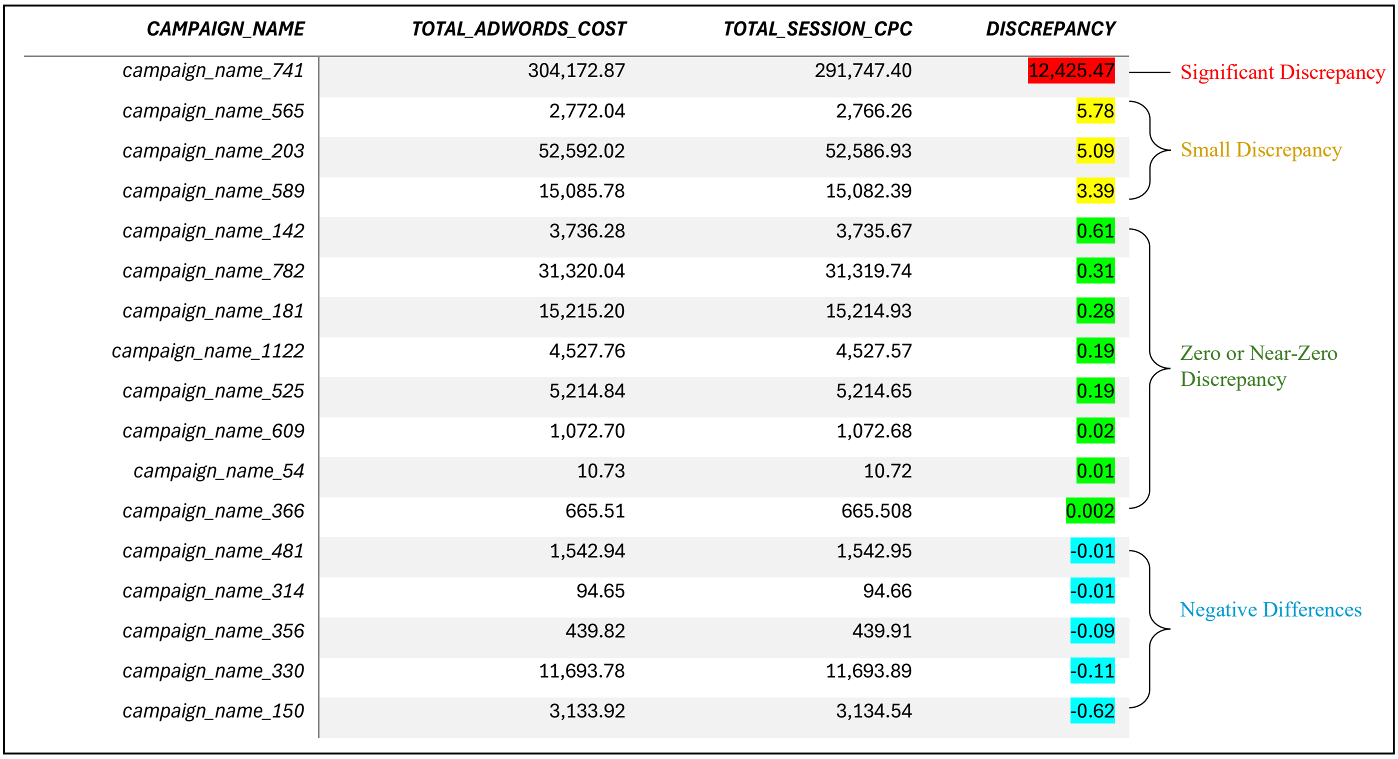
DATA QUALITY

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

```
-- 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
campaign_name_565	2,772.04	2,766.26	5.78
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
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
campaign_name_330	11,693.78	11,693.89	-0.11
campaign_name_150	3,133.92	3,134.54	-0.62

Significant Discrepancy

Small Discrepancy

Zero or Near-Zero Discrepancy

Negative Differences

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:

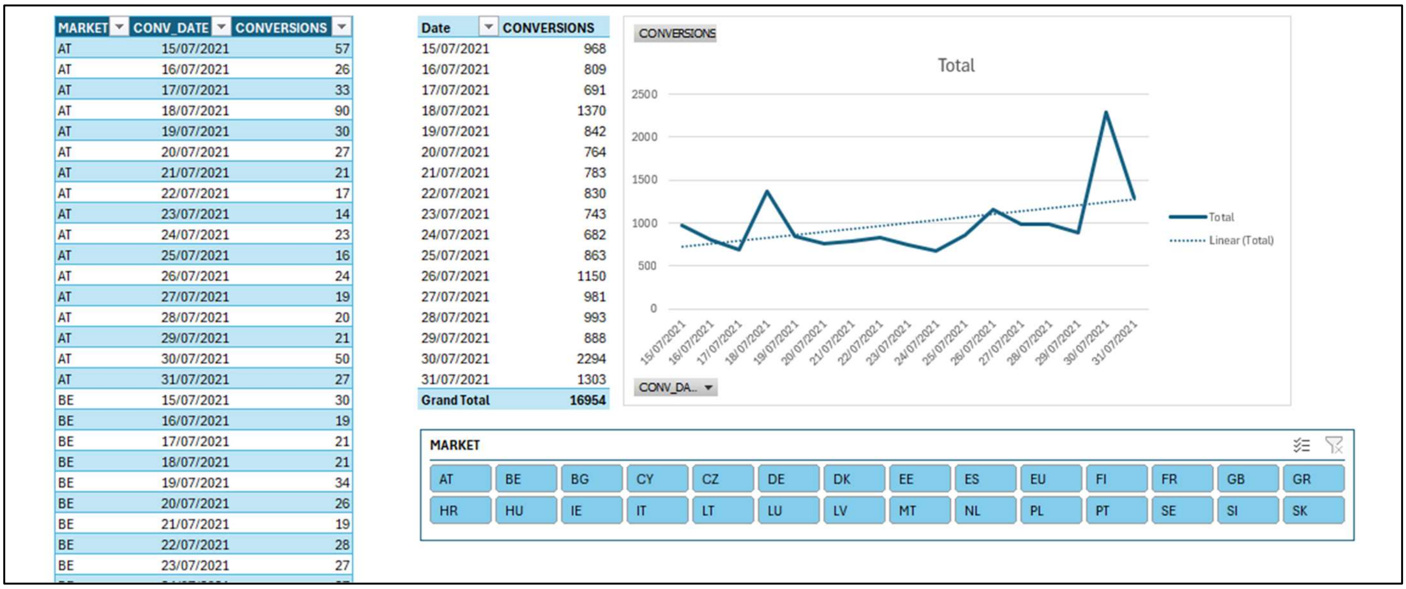
- Rounding errors.
- Slight timing discrepancies.
- Minor data entry inconsistencies.

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?



MARKET	CONV_DATE	CONVERSIONS
AT	15/07/2021	57
AT	16/07/2021	26
AT	17/07/2021	33
AT	18/07/2021	90
AT	19/07/2021	30
AT	20/07/2021	27
AT	21/07/2021	21
AT	22/07/2021	17
AT	23/07/2021	14
AT	24/07/2021	23
AT	25/07/2021	16
AT	26/07/2021	24
AT	27/07/2021	19
AT	28/07/2021	20
AT	29/07/2021	21
AT	30/07/2021	50
AT	31/07/2021	27
BE	15/07/2021	30
BE	16/07/2021	19
BE	17/07/2021	21
BE	18/07/2021	21
BE	19/07/2021	34
BE	20/07/2021	26
BE	21/07/2021	19
BE	22/07/2021	28
BE	23/07/2021	27

Date	CONVERSIONS
15/07/2021	968
16/07/2021	809
17/07/2021	691
18/07/2021	1370
19/07/2021	842
20/07/2021	764
21/07/2021	783
22/07/2021	830
23/07/2021	743
24/07/2021	682
25/07/2021	863
26/07/2021	1150
27/07/2021	981
28/07/2021	993
29/07/2021	888
30/07/2021	2294
31/07/2021	1303
Grand Total	**16954**

Takeaways:

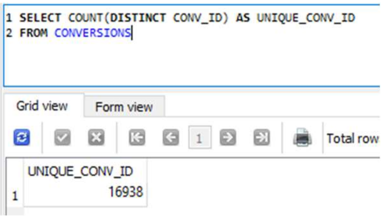
The e-commerce conversions in July 2021 show considerable **instability** with **large fluctuations**. There are no clear, stable patterns, suggesting that external factors, promotional activities, or specific events may have caused these variations.

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:

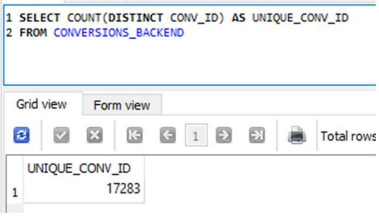
- Count of Days: **17**
- Mean (Average) Conversions: **997**
- Standard Deviation: **388.43** (High SD suggests a considerable fluctuation in daily conversions.)
- Minimum Conversions: **682** (on July 24, 2021)
- 25th Percentile (Q1): **783**
- Median (Q2): **863**
- 75th Percentile (Q3): **993**
-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

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



```
SELECT COUNT(DISTINCT CONV_ID) AS UNIQUE_CONV_ID
FROM CONVERSIONS
```

	UNIQUE_CONV_ID
1	16938



```
SELECT COUNT(DISTINCT CONV_ID) AS UNIQUE_CONV_ID
FROM CONVERSIONS_BACKEND
```

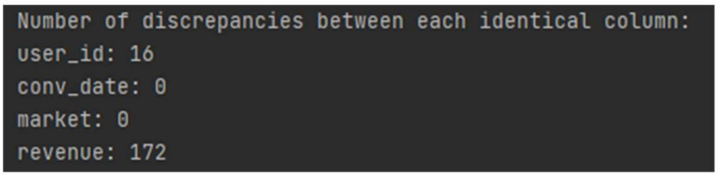
	UNIQUE_CONV_ID
1	17283

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:

```
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:



```
Number of discrepancies between each identical column:
user_id: 16
conv_date: 0
market: 0
revenue: 172
```

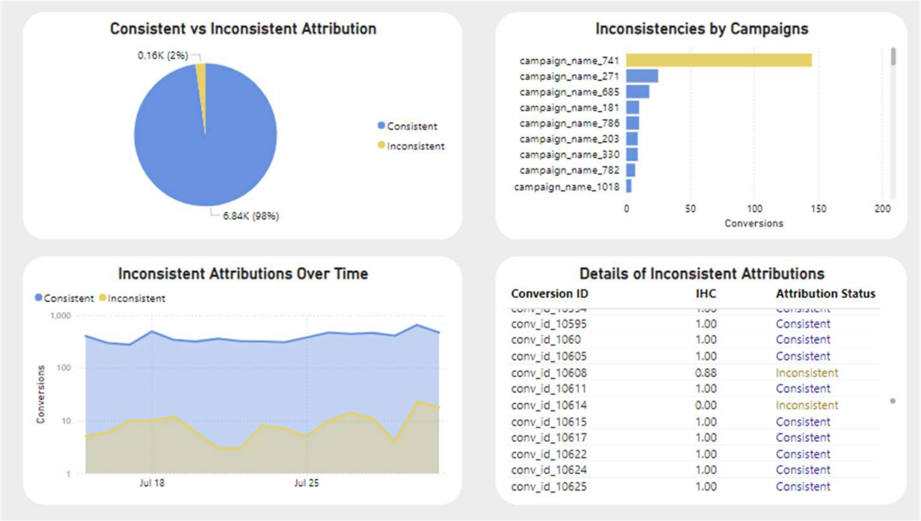
Question 4) Are attribution results consistent? Do you find any conversions where the 'ihc' values don't make sense?

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:



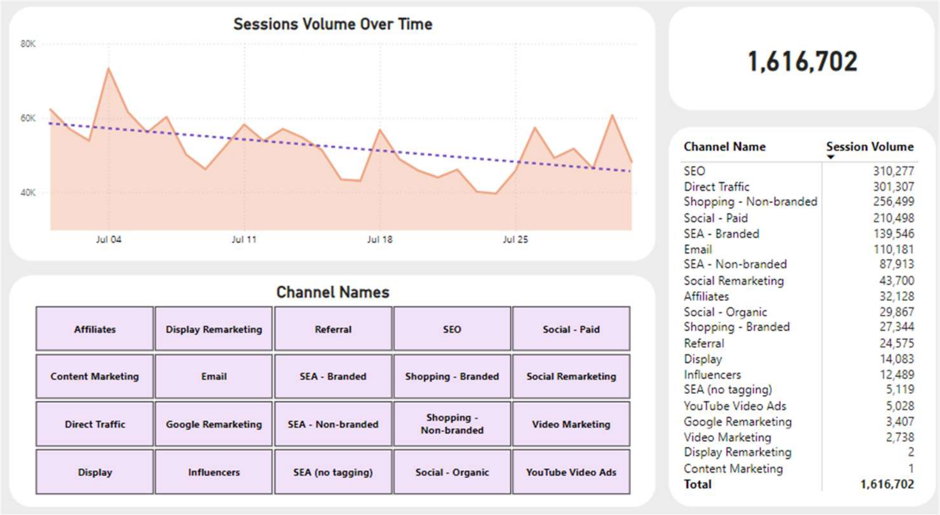
Details of Inconsistent Attributions

Conversion ID	IHC	Attribution Status
conv_id_10595	1.00	Consistent
conv_id_1060	1.00	Consistent
conv_id_10605	1.00	Consistent
conv_id_10608	0.88	Inconsistent
conv_id_10611	1.00	Consistent
conv_id_10614	0.00	Inconsistent
conv_id_10615	1.00	Consistent
conv_id_10617	1.00	Consistent
conv_id_10622	1.00	Consistent
conv_id_10624	1.00	Consistent
conv_id_10625	1.00	Consistent

Question 5) 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.



1,616,702

Channel Names

Affiliates	Display Remarketing	Referral	SEO	Social - Paid
Content Marketing	Email	SEA - Branded	Shopping - Branded	Social Remarketing
Direct Traffic	Google Remarketing	SEA - Non-branded	Shopping - Non-branded	Video Marketing
Display	Influencers	SEA (no tagging)	Social - Organic	YouTube Video Ads

Channel Name	Session Volume
SEO	310,277
Direct Traffic	301,307
Shopping - Non-branded	256,499
Social - Paid	210,498
SEA - Branded	139,546
Email	110,181
SEA - Non-branded	87,913
Social Remarketing	43,700
Affiliates	32,128
Social - Organic	29,867
Shopping - Branded	27,344
Referral	24,575
Display	14,083
Influencers	12,489
SEA (no tagging)	5,119
YouTube Video Ads	5,028
Google Remarketing	3,407
Video Marketing	2,738
Display Remarketing	2
Content Marketing	1
Total	**1,616,702**

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) 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.