Haensel AMS

Data Quality Challenge

-Nitin Nandan Singh

Introduction

This report is part of my application for the Business Intelligence & Data Analyst position at Haensel AMS. It outlines my approach to the data quality challenge and provides solutions to the questions posed. The challenge and questions are summarized first, followed by a description of my approach and solutions.

Data Quality Challenge

The HAMS Data Quality Challenge involves analyzing a database for Company X to identify data quality issues. The database includes tables for e-commerce purchases (conversions), user sessions on the company's website (session_sources), backend conversion data (conversions_backend), AdWords campaign costs (api_adwords_costs), and the customer journey linking sessions to conversions (attribution_customer_journey).

Approach

The first step in this case study was to understand the goals and expectations. I thoroughly read the challenge details multiple times, analyzed the questions, and brainstormed ideas. For instance, to address the stability of conversions over time, I visualized a line graph with conversion values on the y-axis and time units on the x-axis.

I researched unfamiliar terms like ihc and then focused on the technical details, including the database schema and table relationships. I set up a local and remote repository, added the database file, and connected to it using Python. I queried all tables into data frames to leverage Pandas for exploration, examining row counts, data types, missing values, and any peculiarities before tackling the questions. The observations from solving the questions are provided below, and the notebooks can be found in the github repo.

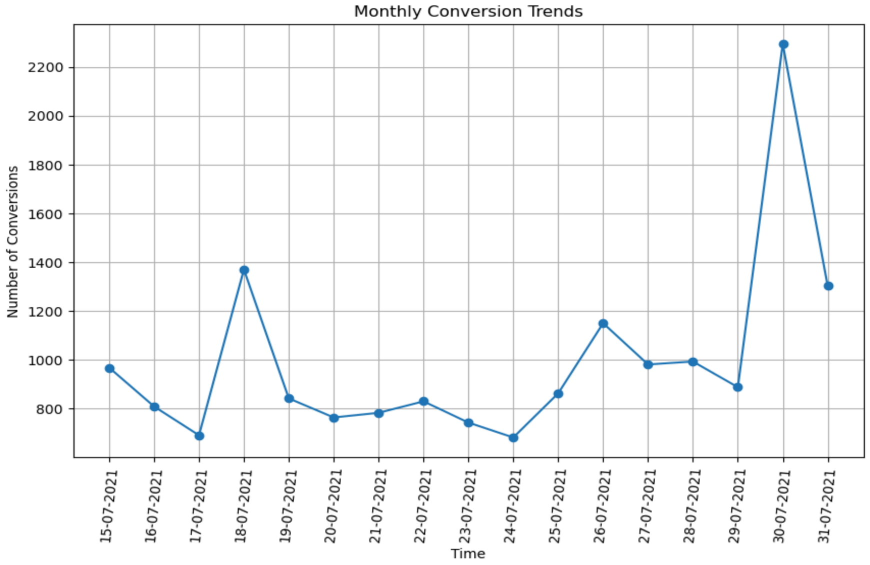
Solutions

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

   To check if the costs in the ‘api_adwords_costs’ table are fully covered in the ‘session_sources’ table, I grouped both tables by event date and campaign ID, summing the costs and CPC respectively. I then took a difference of the totals for discrepancies. There were mismatches in **108 campaigns** between the ‘api_adwords_costs’ and ‘session_sources’ tables.

2. **Are the conversions in the 'conversions' table stable over time? Any pattern?**

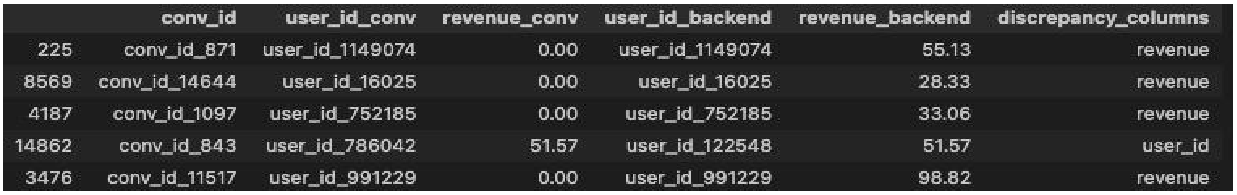
   For this, I chose to plot a bar chart of the number of conversions over time.



   It can be observed that the conversions kept fluctuating overtime going as high as 2294 (may be due to sale day, for example Prime day for Amazon).

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

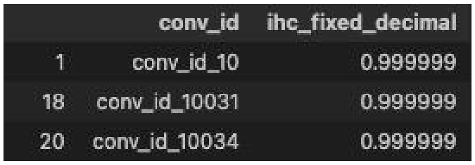
   To perform a comprehensive matching between the two tables I compared the corresponding rows between the tables. I found that not only are the number of records different between the tables but also there are discrepancies in the actual values in columns such as revenue and user_id. A snippet of such conversions can be seen below.



	conv_id	user_id_conv	revenue_conv	user_id_backend	revenue_backend	discrepancy_columns
225	conv_id_871	user_id_1149074	0.00	user_id_1149074	55.13	revenue
8569	conv_id_14644	user_id_16025	0.00	user_id_16025	28.33	revenue
4187	conv_id_1097	user_id_752185	0.00	user_id_752185	33.06	revenue
14862	conv_id_843	user_id_786042	51.57	user_id_122548	51.57	user_id
3476	conv_id_11517	user_id_991229	0.00	user_id_991229	98.82	revenue

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

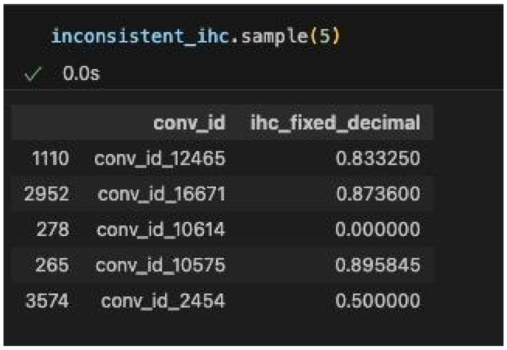
I focused on verifying that the total ihc value for each conversion equals 1. I grouped the data in attribution_customer_journey by conv_id and summed the ihc values, then checked for sums not equal to 1. Initially, sums like 1.000000 were incorrectly flagged due to varying decimal precision. To resolve this, I set ihc values to six decimal places for consistency, which corrected the issue and eliminated the false discrepancies.



	conv_id	ihc_fixed_decimal
1	conv_id_10	0.999999
18	conv_id_10031	0.999999
20	conv_id_10034	0.999999

Values like 0.999999 were flagged as inconsistent ihc values despite being close to 1. Based on a business decision, I chose to use a tolerance level of 10e-6 to consider them consistent.

There are 161 conversions with inconsistent ihc values.

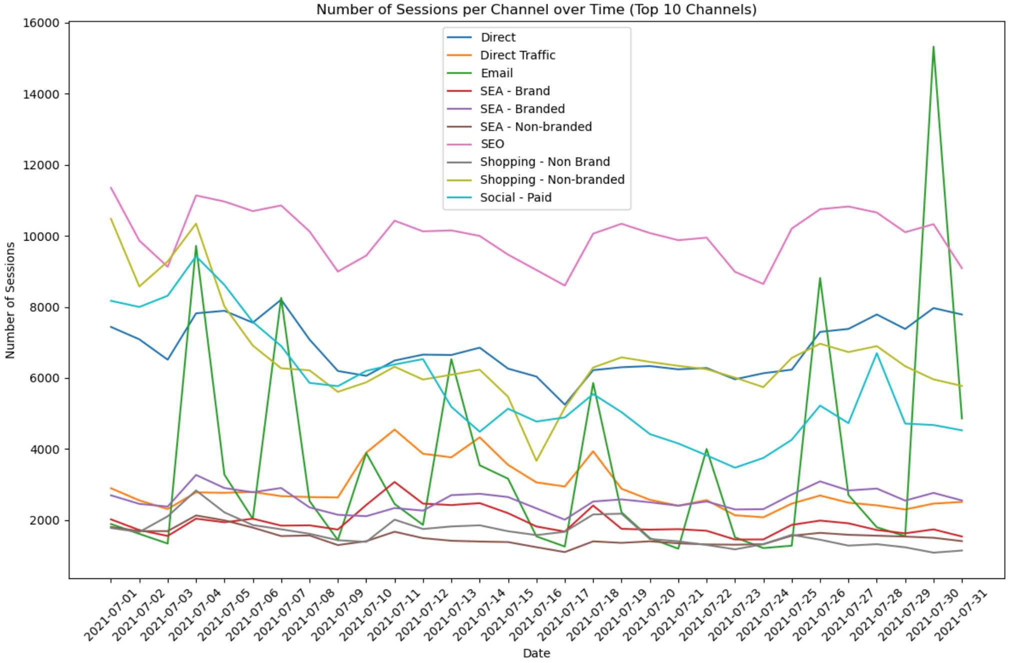


```
inconsistent_ihc.sample(5)
✓ 0.0s
```

	conv_id	ihc_fixed_decimal
1110	conv_id_12465	0.833250
2952	conv_id_16671	0.873600
278	conv_id_10614	0.000000
265	conv_id_10575	0.895845
3574	conv_id_2454	0.500000

Bonus Question

1. **Do we have an issue with channeling? Are the number of sessions per channel stable over time?**



The plot shows the number of sessions over time for the top 10 channels. There is significant fluctuation, notably a spike on July 30, 2021, in the email channel. This spike, along with a higher number of conversions on the same day, suggests a promotional email likely led to increased conversions on that date.

2. **Any other issues?**

   In the conversions table, there are some conversions for which the revenue is 0. This can be a data quality issue. But there can be other reasons for this such as i) A user used a 100% discount code while purchasing the item, ii) Problem while performing the purchase, maybe the conversion was recorded but due to issue with payment the revenue was not processed.