Haensel AMS

**Data Quality Challenge**

-Nitin Nandan Singh

horizontal line

# 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](https://github.com/nitinnandansingh/data_quality_challenge/tree/main).

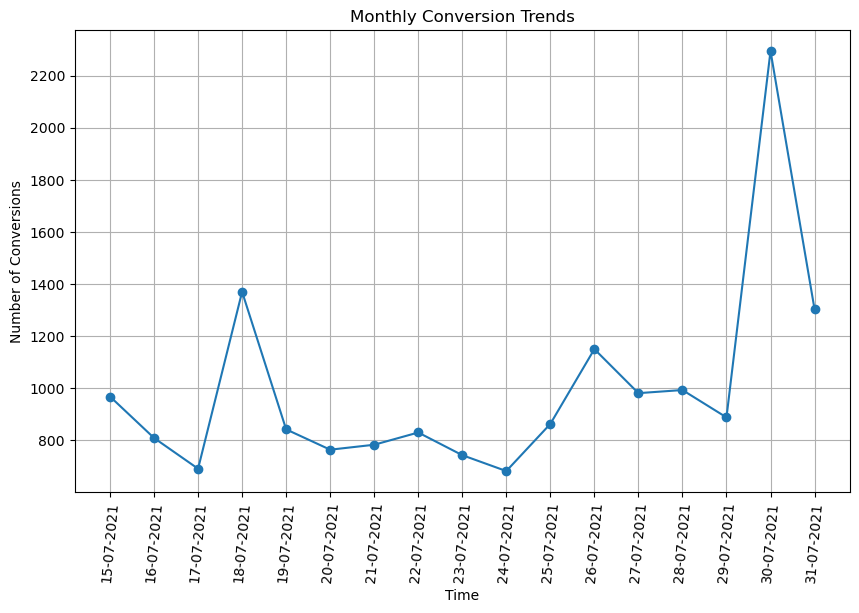
# Solutions

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.

1. **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).

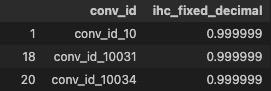
1. **Double check conversions ('conversions' table) with backend ('conversions\_backend' table), any issues?**

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.

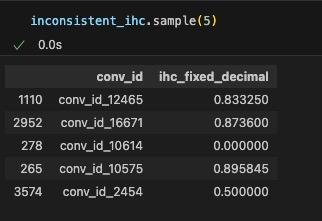


1. **Are attribution results consistent? Do you find any conversions where the 'ihc' values don't make sense?**

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.

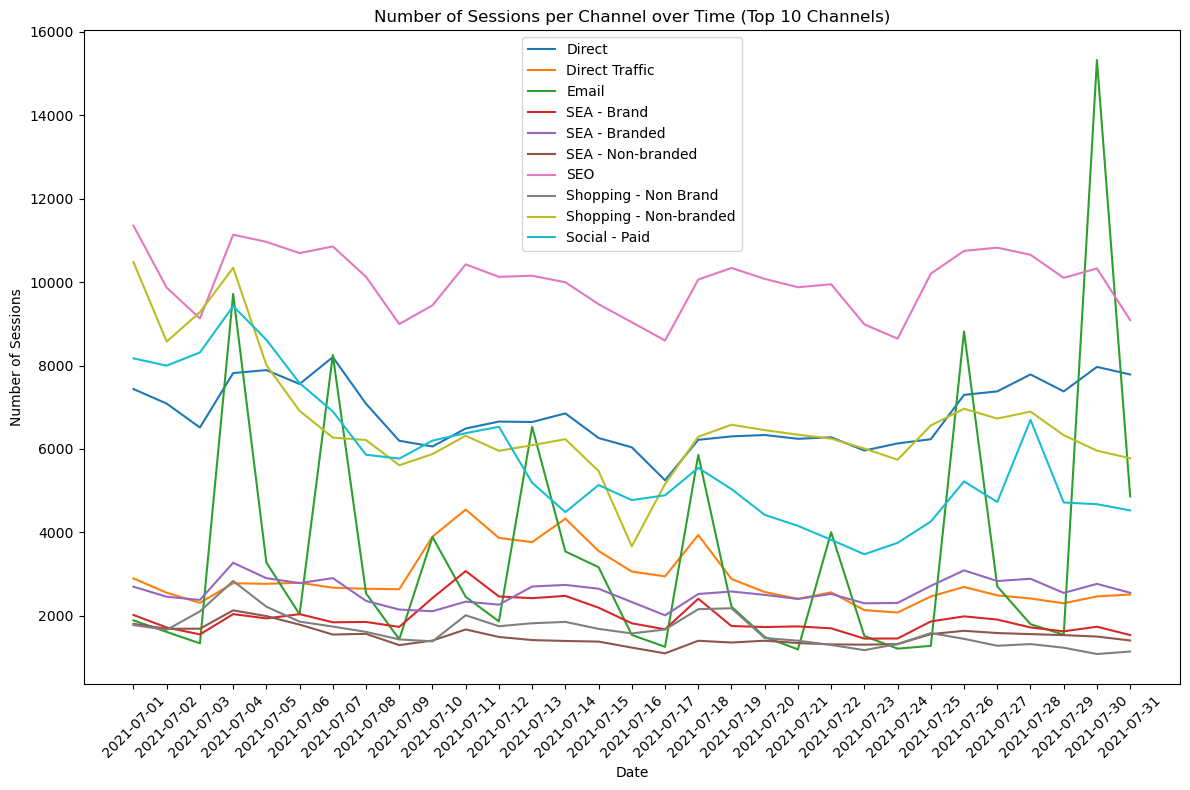


Values like 0.999999 were flagged as inconsistent ihc values despite being close to 1. Based on a business decision, I chose to use a tolerance level of 10e-6 to consider them consistent.

**There are 161 conversions with inconsistent ihc values.**

**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.

1. **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.