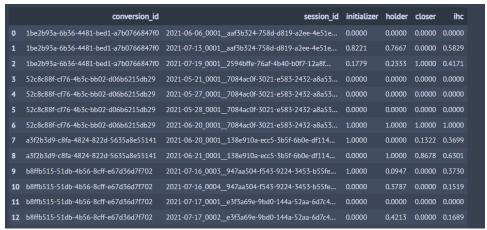
This project aims to demonstrate usage of IHC attribution model API from Haensel AMS GmbH and analyse the data returned by the API and visualize the results.

More information about the IHC attribution model and API can be found here:

https://ihc-attribution.com/ https://ihc-attribution.com/ihc-data-driven-attribution-model/

https://ihc-attribution.com/marketing-attribution-api/

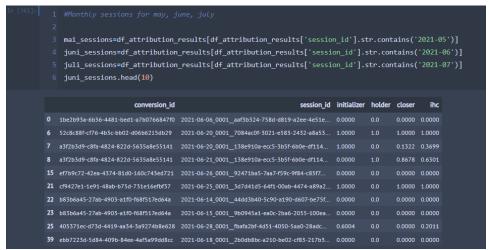
1. Convert received dictionaries from API to dataframe as df_attribution_results: I popped out the dictionaries out of list sent by API which also included statusCode and PartialFailureErrors and converted it to a dataframe.



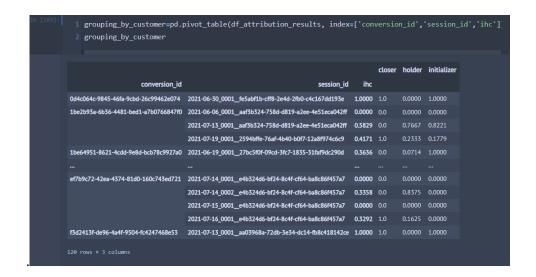
 Assuming that each unique value in customer_id column represents a customer, I found unique customers with customer IDs to eliminate the repetition of customers. Later I counted their total number.

```
Unique Customers with their Customer IDs
['1be2b93a-6b36-4481-bed1-a7b0766847f0'
'52c8c88f-cf76-4b3c-bb02-d06b6215db29'
  a3f2b3d9-c8fa-4824-822d-5635a8e55141'
  'ef7b9c72-42ea-4374-81d0-160c743ed721'
'cf9427e1-1e91-48ab-b75d-731e16efbf37'
  405371ec-d73d-4419-aa34-3a9274b8e628
 '2905fd71-ea4d-4d20-aca6-140a114fef5d'
'ebb7223d-5d84-409b-84ee-4af5a99dd8cc'
  e9985e5c-24b2-4a2c-8567-7b29cbe4db8b
   1d98f466-a32b-4fbd-914d-cd2dd40c821b
  '99182215-acbd-4aed-b428-e6efc783c32b'
'58f939b5-09a1-44f7-a321-9637c29285dd'
  ca59ad35-7cd2-4e6a-9c2c-e65c536328b8'
 'd1bc5ed8-647e-421c-95ce-cadfd46c5efd'
'c10f176e-e7f7-4eaf-baf2-ad95bd4bd906'
  'a7e21a4c-2d77-4ce5-857c-07d6d9299b4d'
'ecd63c4c-3988-4ebb-9527-902c27a0fe27'
  e1b6d544-1f3b-402b-804b-0b6eb7b18f70
  '7260a8e2-8437-48e0-b004-cd7a9b60ba23'
   7ad00a02-712c-4c08-bfe6-2c37bdcb695e
 '6d54814b-4b68-4a3f-857f-d1248f1258a7'
 '964dfa79-b4e3-4769-bbaa-b61aeeaff1da'
'eb55040e-eda4-4bab-aa43-193f213a59bf'
 'b7d6c4f3-8388-45d7-9528-2196db8d3501'
'0d4c064c-9845-46fa-9cbd-26c99462e074'
 '1be64951-8621-4cdd-9e8d-bcb78c9927a0'
'4942aa75-e79d-43da-9fe6-334f109a88d8']
```

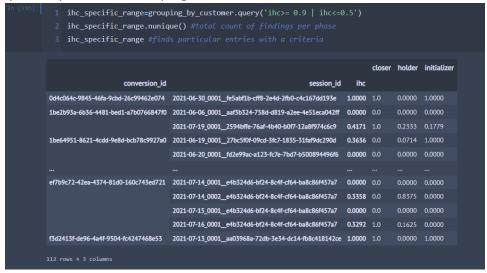
3. The customer data can also be found on a monthly or quarterly basis which shows the results of IHC per month. I found the data for May, June, July by locating the date string in session IDs.



4. I wanted to find all the sessions related to each customer to remove the redundancy, so I used pivot method which shows each customers' sessions and its IHC and initializer, holder and closer phases.

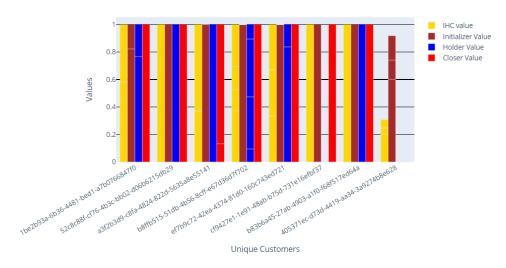


5. If we want to find the data as per any condition, that can be also done using .query() method. I found the customers with all sessions having extreme IHC values below 0.5 or above 0.9. Further, *Average* for all session values for one customer can also be found to compare impact of our campaign over them.

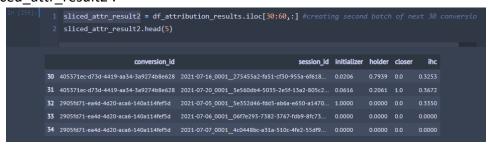


6. Batch processing: Plotting all the conversion_ids and sessions is hard on a single plot so I sliced the 'df_attreibution_results' into first 30 entries as a batch. Then I counted their unique ids and total count.

7. I plotted those unique customers on a bar plot with all the related I/H/C values. Hovering over the bar makes it easy to get good details. *Missing bars* represent zero value.



8. In I created a second batch of the next 30 customers and stored it in var 'sliced attr result2'.



9. The purpose of slicing the next 30 conversion IDs was to compare this batch with the first batch. 'ihc_vs_ihc2' shows the difference of IHC values of both sets of customers. It depicts the pattern of varying IHC values and what batch has better performance with higher scores and what batch has more area for improvement. It is important to note that there can be a difference in the number of conversion IDs on x axis for both plots simply because some of them have a zero IHC score which is not shown on graph.



10. I used the not repeated mean values of I/H/C and IHC columns so that we can have a reference to compare it to the actual value of it as a baseline.



11. I calculated the mean values of whole column for all the values for comparison purposed and to see the performance of one session's I/H/C and IHC values as compared to the category average of other sessions. 'maxwtphase' gives us the maximum of all these values to see the strongest phase of all. This function can also be

modified to give minimum, average, sum or other properties. This function can be called anytime to see the reference mean weights.

```
Average value of initializer phase is 0.28333333333333355

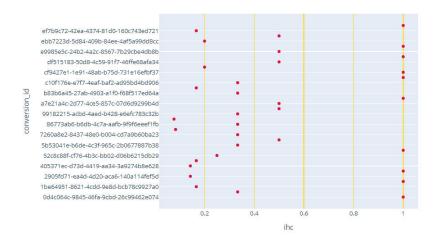
Average value of holder phase is 0.199999999999968

Average value of closer phase is 0.28333333333334

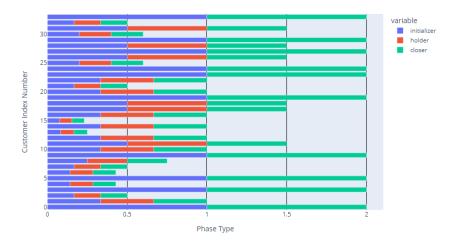
Note: More the value, more of the customers stay in that phase.

The maximun phase value is the 0.283333333333334 for this set of customers.
```

12. To better visualize the IHC column of those 34 unique conversion IDs, I plotted a scatter plot of its mean values against the respective conversion IDs.



13. Next aim was to show all the data of all 34 unique customers into a single plot to save time and space.



14. I computed the averages of all three I/H/C phases for one customer session at a time to find the larger scale comparison of all such customer sessions. This average could also be used to see which session is the most successful session if a customer has multiple sessions. Then we will look at what phase is most impactful in that specific successful session.

```
conversion id
                                     session id
0d4c064c-9845-46fa-9cbd-26c99462e074 2021-06-30_0001__fe5abf1b-cff8-2e4d-2fb0-c4c167dd193e 1.0000
                                                                                                   0.666667
1be2b93a-6b36-4481-bed1-a7b0766847f0 2021-06-06_0001__aaf3b324-758d-d819-a2ee-4e51eca042ff 0.0000
                                                                                                   0.000000
                                                                                                   0.529600
                                    2021-07-19_0001__2594bffe-76af-4b40-b0f7-12a8f974c6c9 0.4171
                                                                                                   0.470400
1be64951-8621-4cdd-9e8d-bcb78c9927a0 2021-06-19_0001__27bc5f0f-09cd-3fc7-1835-31faf9dc290d 0.3636
ef7b9c72-42ea-4374-81d0-160c743ed721 2021-07-14 0001 e4b324d6-bf24-8c4f-cf64-ba8c86f457a7 0.0000
                                                                                                   0.000000
                                     2021-07-14_0002__e4b324d6-bf24-8c4f-cf64-ba8c86f457a7 <u>0</u>.3358
                                                                                                   0.279167
                                    2021-07-15 0001 e4b324d6-bf24-8c4f-cf64-ba8c86f457a7 0.0000
                                                                                                   0.000000
                                                                                                   0.387500
                                                                                                  0.666667
```

15. Then I found the averages of one customer's all session values. For example, if xyz customer has four sessions, then I computed I/H/C values for all four sessions. It was really interesting for me to find that the sum is always 1 for all values of a specific phase scattered in multiple sessions of one customer. Therefore, if one phase has no impact in any of the sessions for that customer, naturally, average impact becomes zero showing that other two phases are dominating the average performance.

```
#Lets calculate averages for one customer's all the sessions.

2 #sum is always 1 for all values of all sessions if only one phase is considered

3 all_sessions_avg= grouping_by_customer.groupby(['conversion_id'], as_index=False).mean()

4 all_sessions_avg.tail(4)

closer holder initializer

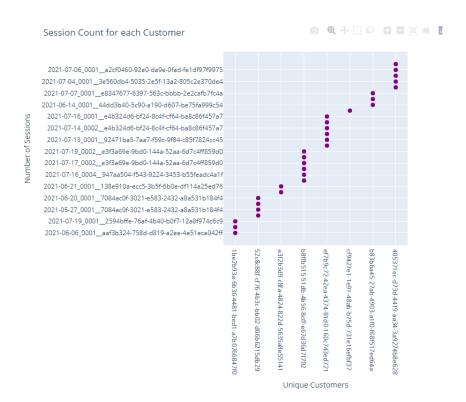
50 0.200000 0.200000 0.200000

31 0.500000 0.500000 0.500000

32 0.166667 0.166667 0.166667

33 1.000000 0.000000 1.000000
```

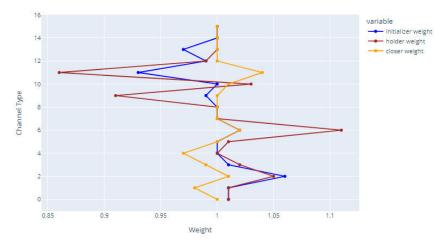
16. Then I plotted a graph showing how many sessions one customer has in that dataset. This will tell us which customer is most engaged to us so that we can focus on closing those customers and putting more efforts into single session customers to increase their future engagement.



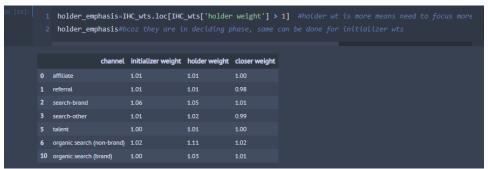
17. I read the downloaded csv file of IHC channel weights. Unfortunately, I could not figure out the way to append the respective IHC weights into this table but I tried to analyze the data at hand efficiently.



18. I plotted the graph showing all the phases attached to their respective channels represented by the index number.



19. I calculated the holder weights that are more than 1. It was interesting to see that all such phases had Initializer and Closer phases with weights more than 1 too. Such high weight phase can be transitioned to same phase but with different channel to continue successful customer acquisition. This loc can be tweaked to get other phases and their channels also.



20. I computed the worst channels in terms of initializing, holding and closing weight assuming that the bigger value represents the stronger phase performance of that channel and vice versa. I considered 1 as the benchmark weight for good performance of a phase and found phase weights less than or equal to 1 for at least one of the three phases. *AND* operator can be used to find weights >=1. Methods with at least one phase with weight less than 1 which demands more work from marketeers on that phase.



21. Related to closer weight, I computed the highest value which derives the most useful channel in terms of finalizing customer purchases. This channel can be weaker in other phases but still stronger in closing a customer journey. Then I summoned the other details of that respective channel for more clarification. Of course, this code can be twisted to get the data of initializer and holder phases too. We can also calculate the min() value of such phases to see the least efficient phases.



22. Next task was to find the average weights of each phase. I distributed those weights over the total channel count which was 16 by dividing those three averages with channel count. Although 'max_avg' gives us the biggest average of all three columns, a distributed average over all channels will give us the granular detailed view of what a phase's contribution is in that channel's performance. In this dataset, all the values are pretty much nearby so the average does not fluctuate much but for a different dataset with large differences in values, the results could have larger differences among them.

```
Following 3 distributed averages tell us how much a phase average has contributed in that channel's overall performence:

1. Initialize phase average per channel is 0.0624609375

2. Holder phase average per channel is 0.0625

3. Closer phase average per channel is 0.062578125

The strongest phase performance(average of a phase) in general is 1.00125

The maximum average of a phase distributed all over all channels is 0.062578125, tells us that phase's contribution in each channel's performance.
```