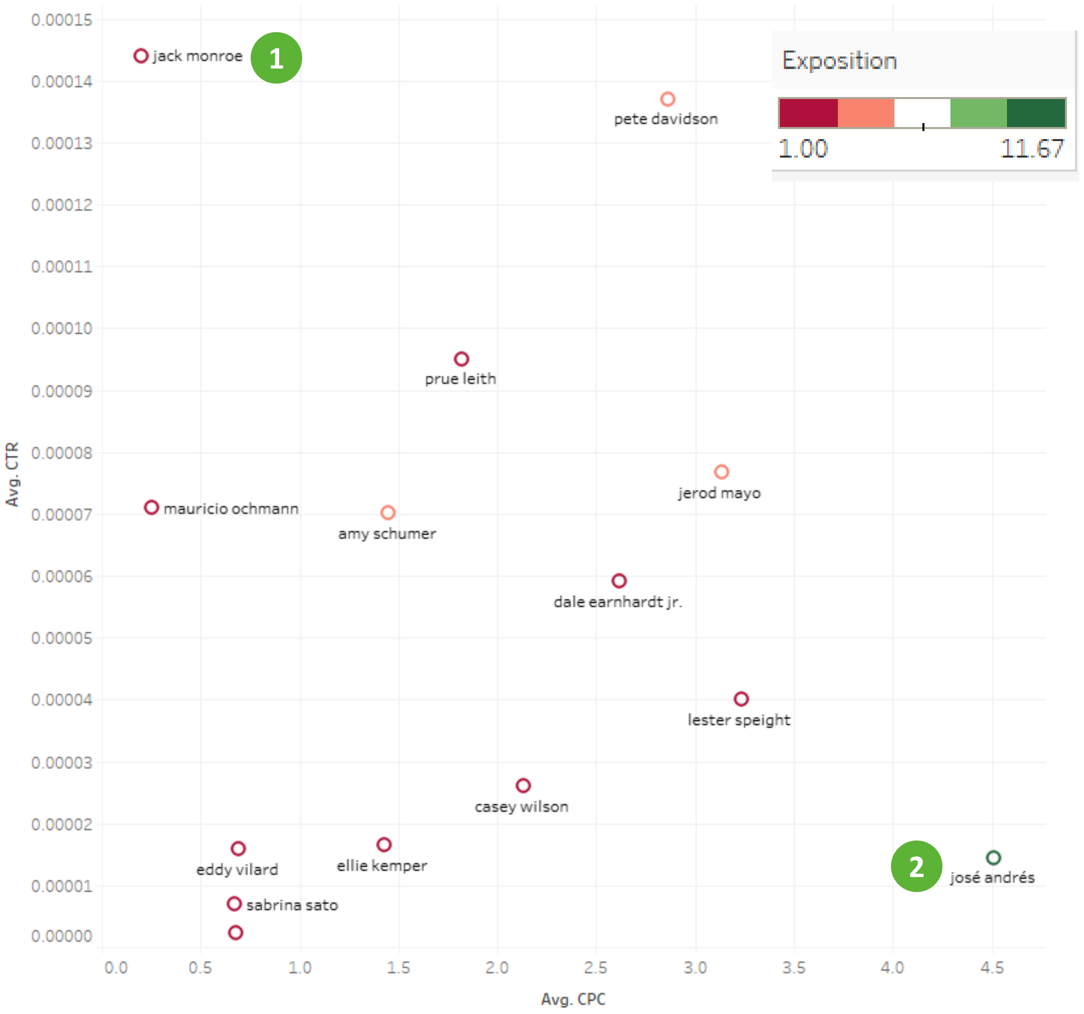
Analysis and Conclusions

Technical Assessment: Data Analyst

Candidate: Maximiliano Vedoya R

Data Preliminary analysis

Before modeling it is necessary to create some plots to help understand relationship between each predictor and the performance metrics.



In this graph we can observe CPC and CTR average metrics for each creative element (in this case celebrities)

1. Jack Monroe is the celebrity that on average has the highest CTR at the Lowest CPC, yet the frequency of usage of its image on our assets is low

2. Jose Andres is the celebrity that on average has the highest CPC at the Lowest CTR, yet the frequency of usage of its image on our assets is high

Both points suggest that **there is value on exploring how different creative elements correlates with the overall performance** of the asset, since we are looking into a suboptimal configuration already

However, this could be **misleading** since we are only taking average values and not considering other factors that contribute to this results. E.g Jack Monroe background was disturbing and that is the real. To analyze how each creative element correlates with the overall performance of the tool we need to use more advance analytics.

Overall Procedure

1. Use a non-parametric model (GBM) to determine the overall importance of each independent variable
2. Based on the previous model, chose only the most relevant predictors to reduce model complexity
3. Train a multivariant linear model using the previous selected features.
4. Assess linear model coefficients as a measure to directionally determine the contribution of each selected feature

Predictors preliminary analysis

Considering the main goals of this exercise, I will mainly leverage Tag_name and Tag_type to build the model, following my main points on the remaining predictors

- **Ad_id:** not useful information, it is the id of each element in the tables

- **Asset_id:** this something important if we were to measure performance of each asset, however, here we are trying to explain performance of elements present across all assets. An alternative approach would be to use this as a predictor to understand which asset perform better for each performance metric
- **Asset_type:** relevant to understand frequency of usage for each asset, in case of video, we should take exposition time into account when comparing usage frequency of certain elements
- **Confidence**: only used to filter out elements with confidence below the indicated threshold 0.9
- **Date_captured:** there is definitely relevant information here, different dates might play a massive role on cpc or ctr and vvr, this could be used to understand what time is better to display certain elements within our assets. However, for the scope of this analysis I will exclude this categorical variable from the model. This decision is based on the fact that the product we are trying to sell (mayonnaise) has no strong seasonality and therefore changes in CPC/CTR across time should not be attribute to changes in demand. For the scope of the available data, changes in performance across time should be related to the use of certain assets (that work better) during that time. Under these assumptions time is strongly correlated to tag_name and tag_type and should not be used
- **Market:** in more detailed analysis this could be used to understand how each element of different assets perform in different regions, for the scope of this analysis I will focus on understand this concept at a global level
- **Start_timestamp- End_timestamp :** only useful on videos as an indirect measure of exposition time, used to compute frequency of usage and not as an independent predictor
- **Start_frame-End_frame :** only useful on videos as an indirect measure of exposition time (noticed that we could use this metric in combination with the previous one to approximate FPS, a quality measure of our asset)

Results

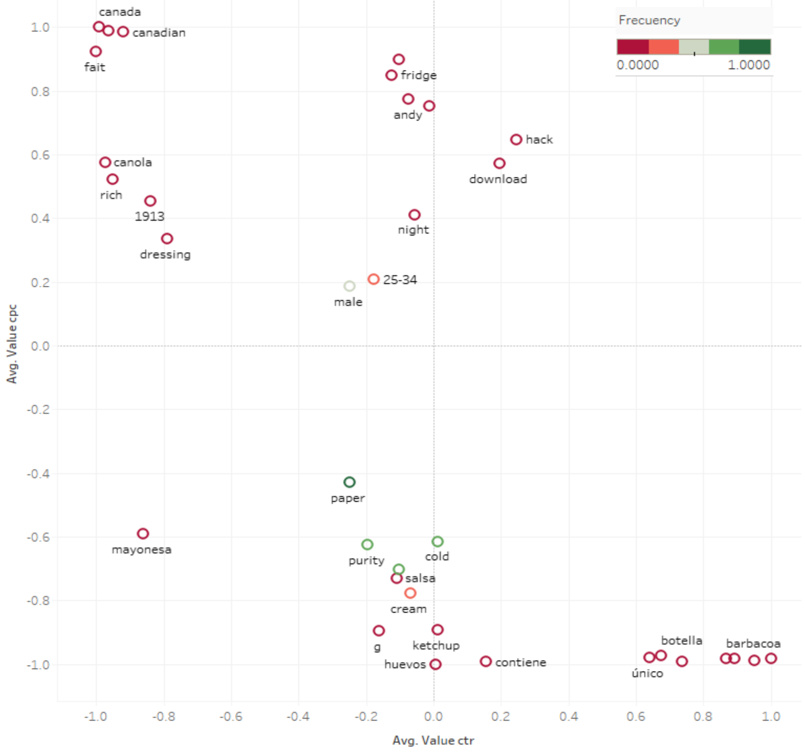
Importance of each creative feature by tag_type and tag_name

A full list of the most important creative elements can be found at

- *Results_Importance_CPC.xlsx*
- *Results_Importance_CTR.xlsx*

In addition, I plotted each feature based on values for CPC and CTR

- Performance will be determine based on CPC and CTR (VVR would be excluded since it only provides guidance for video assets)
- For CPC positive coefficients indicates more cost per click therefore a bad performance
- For CTR positive coefficients indicates more people clicked on the ad therefore good performance



* GMB capacity to explain performance through the usage of creative assets is better than linear model but, in both cases, R^2 is moderate.

Results

Which are the creative elements that drive higher ad performance?

Creative elements that perform better are the ones that have low cpc and high ctr values. In the plot , the ones placed on quadrat 2

Which creative elements should be included more in the client's ads and which should be used less?

In this case, we can leverage the same plot and considering the frequency color code. As previously mentioned, best case scenario are creative elements placed on quadrant 2, in this case we can see the words like **barbacoa/barbecue** tend to perform well but are seldom used (we should increase the use of these elements)

On the other side, the worst performance create elements are place on quadrant 3, we should reduce its usage, spcially if their frecuency is relatively high such as **male** presence on the ad