Criteo Dataset – Data Viz - Dauphine M2

Attached are 2 datasets:

* matched\_sales.tsv – Sales of an e-commerce website that are attributed to Criteo, using a post-click 30D attribution rule
* campaigns\_delivery.tsv – Delivery metrics for the online advertising campaigns that the e-commerce website is running with Criteo

**Guided questions**

1. Analyze matched\_sales.tsv
   1. Use data visualization techniques to explore the dataset and identify data issue. Clean the dataset accordingly.   
      *Tips: Transactions with an Order Value of zero should be removed; Transactions with abnormally high Order Value should be removed; Duplicated Transactions should be removed.*
   2. Can you estimate the time zone of the client?
   3. What is the evolution of the Criteo attributed revenue over time? Can you pinpoint and explain particular events during the period?
   4. Propose a vizualization for the evolution of the average OV per Os.  
      *Tips: Watch out for transactions with empty Transaction ID*
2. Analyze campaigns\_delivery.tsv
   1. Compute the average Criteo margin
   2. Plot the margin over time (day and then week). Which graph is more insightful? How would you qualify Criteo margin?
   3. Plot the margin by Criteo product. Please comment/investigate the App Install margin.
3. Join the 2 datasets
4. Analyze Criteo Campaigns Performance from the advertiser’s perspective
   1. Identify the most performing (using performance and scale criteria) campaigns using a viz.
   2. The previous viz made it challenging to see patterns. Can you adapt it? Which Criteo campaign types perform the best?  
      *Tips: use Campaign Type instead*
   3. Is the COS/ROAS a relevant metric to look at for all Criteo campaign types? What would you recommend looking at instead?

**Dataset description**

*Matched Sales*

Each row in the dataset refers to a single transaction attributed to Criteo; each transaction is characterized by all the following fields:

* **Timestamp:** Date and time of the transaction expressed according to the Unix time system
* **Environment:** Type of environment on which the transaction took place (web or app)
* **Os:** Type of OS on which the transaction took place (Android, Windows, Mac OS, Other, iOS)
* **Campaign ID:** Identifier for a Criteo campaign
* **User ID:** Cookie-centric identifier of the user (for each browser or device used there is a different user ID)
* **Context ID:** A classification system the Criteo engine uses to identify users based on their interaction with an advertiser's website. The context is cookie-centric and evolves over time.

|  |  |
| --- | --- |
| context\_id | description |
| 0 | Criteo doesn't know the visitor on the current merchant. |
| 1 | Criteo knows the visitor through other Partner such as Bluekai |
| 2 | The visitor has seen a non specific page = homepage |
| 10 | The visitor has seen a listing page |
| 4 | The visitor has seen 1 product page |
| 5 | The visitor has seen 2 or 3 product pages |
| 6 | The visitor has seen 4 or more product pages |
| 7 | The visitor has made one purchase (1 or several products bought) |
| 8 | The visitor has made 2 or 3 purchases (1 or several products bought per purchase) |
| 9 | The visitor has made 4 or more purchases (1 or several products bought per purchase) |

* **Transaction ID:** Identifier for a transaction
* **Order Value:** Amount of the transaction in local currency

*Campaign Delivery*

The dataset contains delivery stats for each campaign x day x environment x os x device. Below is the description of each field.

* **Day:** Day in UTC
* **Environment:** Type of environment
* **Os:** Type of OS
* **Device:** Classification of devices
* **Campaign ID:** Identifier for a Criteo campaign
* **Campaign Optimization:** Type of events (e.g., Visit, Sales, App Installs) that the Criteo campaign is set to optimize for
* **Campaign Type:** Type of campaigns. Available values are: PROSPECTING, INAPP, MID FUNNEL CUSTOM, LOWER FUNNEL CUSTOM, LOWER FUNNEL CUSTOM, APP INSTALL.
* **Context IDs Eligible:** Subset of contexts that can be exposed by the campaign
* **Number of displays:** Total number of displays that were served by Criteo for this campaign x day x environment x os x device
* **Number of clicks:** Total number of clicks that were served by Criteo for this campaign x day x environment x os x device
* **Criteo Revenue:** Total revenue that Criteo generated (for itself) for this campaign x day x environment x os x device. This is the same as the total advertiser spend for this campaign x day x environment x os x device
* **Criteo Cost:** Total inventory cost that Criteo incurred for this campaign x day x environment x os x device.