ASSIGNMENT -2

UNCOVERING MARKETING INSIGHTS

Dataset : **CRITEO** Live Traffic Data
https://s3-eu-west-1.amazonaws.com/attributiondataset/criteo_attribution_dataset.zip

TEAM 3

Anurag Rachcha	001375637
Gauri Verma	001306996
Shubham Mahajan	001314273

GOAL:

- Analyze and build an analytical dashboard as a proof-of-concept to illustrate the value of data driven analytics.

- To analyze digital marketing dataset using various tools including XCSV, Trifacta, Snowflake, and Salesforce Einstein Analytics as an Algorithmic Marketing Analyst.

ABOUT THE DATASET

- The dataset represents a sample of 30 days of Criteo live traffic data. Each line corresponds to one impression (a banner) that was displayed to a user. For each banner we have detailed information about the context, if it was clicked, if it led to a conversion and if it led to a conversion that was attributed to Criteo or not.
- Criteo's product is a form of display advertising. Crieto’s personalized retargeting solution displays interactive banner advertisements, generated based on the online retail browsing preferences and products for each customer.

XSV

xsv is a command-line program for indexing, slicing, analyzing, splitting and joining CSV files

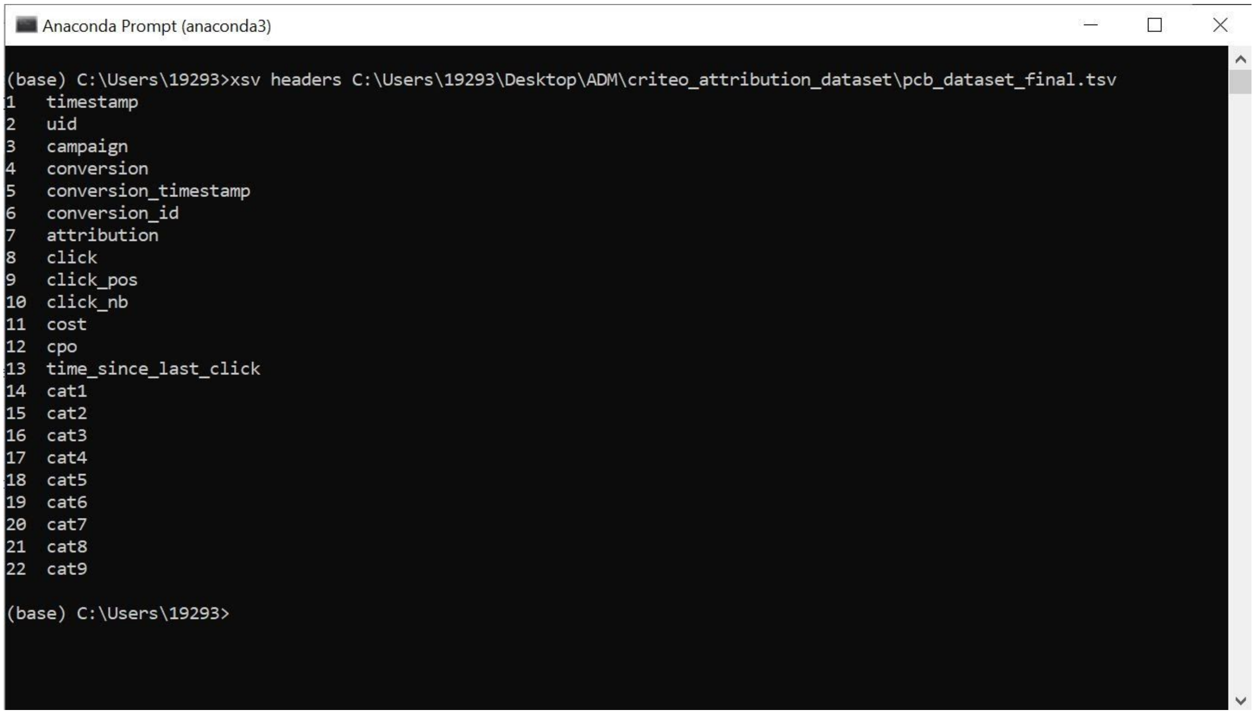
Strengths:

- Commands are simple, fast and composable.
- It has helpful commands such as slice, sample, partition.etc
- The commands are *instantaneous* because they run in time and memory proportional to the size of the slice

Weaknesses:

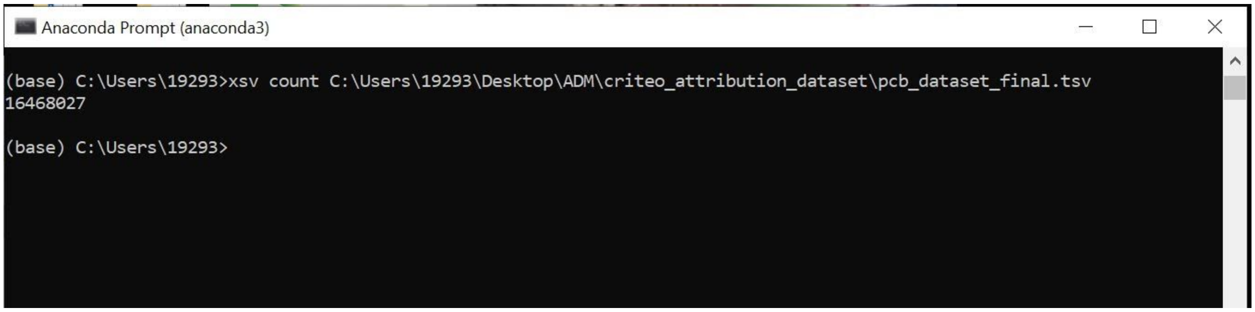
- The user interface is dull. No images or graphics
- Limited number of commands
- Need to be very specific and careful while typing the commands

The headers command indicates all the column names of the dataset



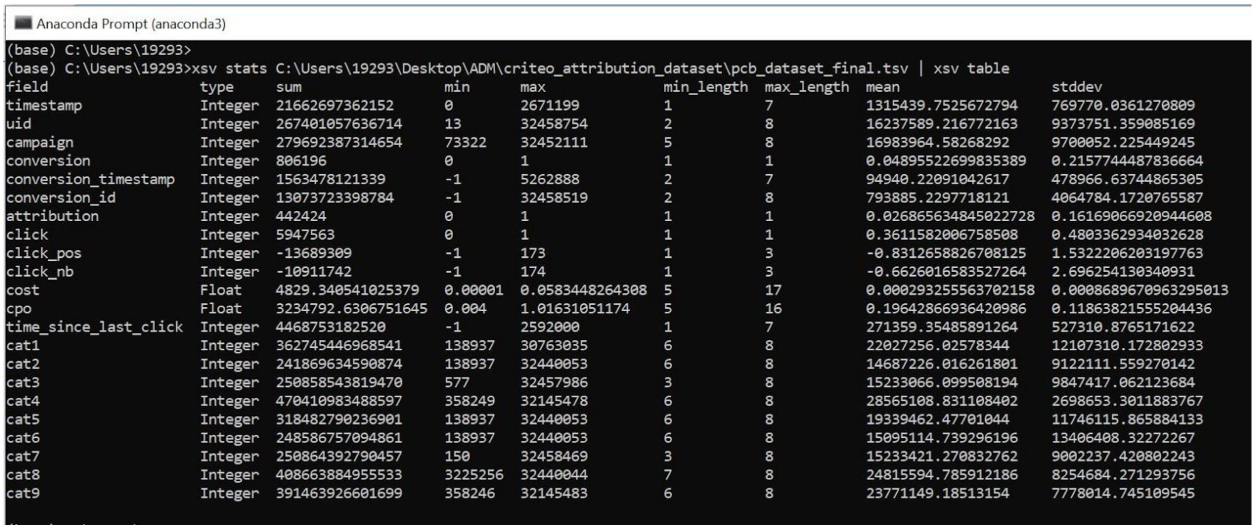
```
(base) C:\Users\19293>xsv headers C:\Users\19293\Desktop\ADM\criteo_attribution_dataset\pcb_dataset_final.tsv
1   timestamp
2   uid
3   campaign
4   conversion
5   conversion_timestamp
6   conversion_id
7   attribution
8   click
9   click_pos
10  click_nb
11  cost
12  cpo
13  time_since_last_click
14  cat1
15  cat2
16  cat3
17  cat4
18  cat5
19  cat6
20  cat7
21  cat8
22  cat9

(base) C:\Users\19293>
```

The count command gives us the count of the number of rows in the dataset - which is 16468027



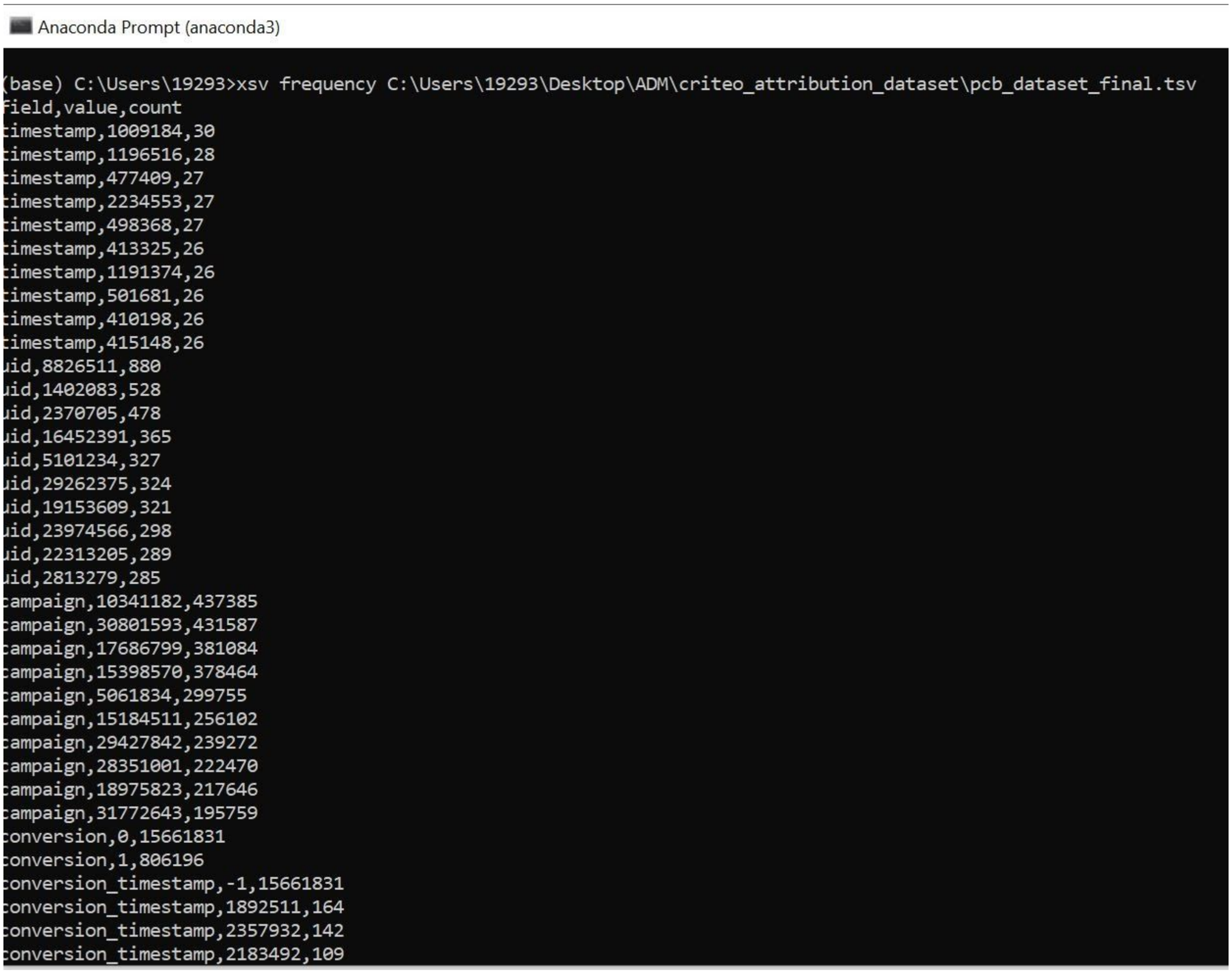
```
(base) C:\Users\19293>xsv count C:\Users\19293\Desktop\ADM\criteo_attribution_dataset\pcb_dataset_final.tsv
16468027

(base) C:\Users\19293>
```

The stats command long with xsv table gives us a tabular representation of the statistics of the data including the data type, min & max value, min & max length, mean and standard deviation of all columns.



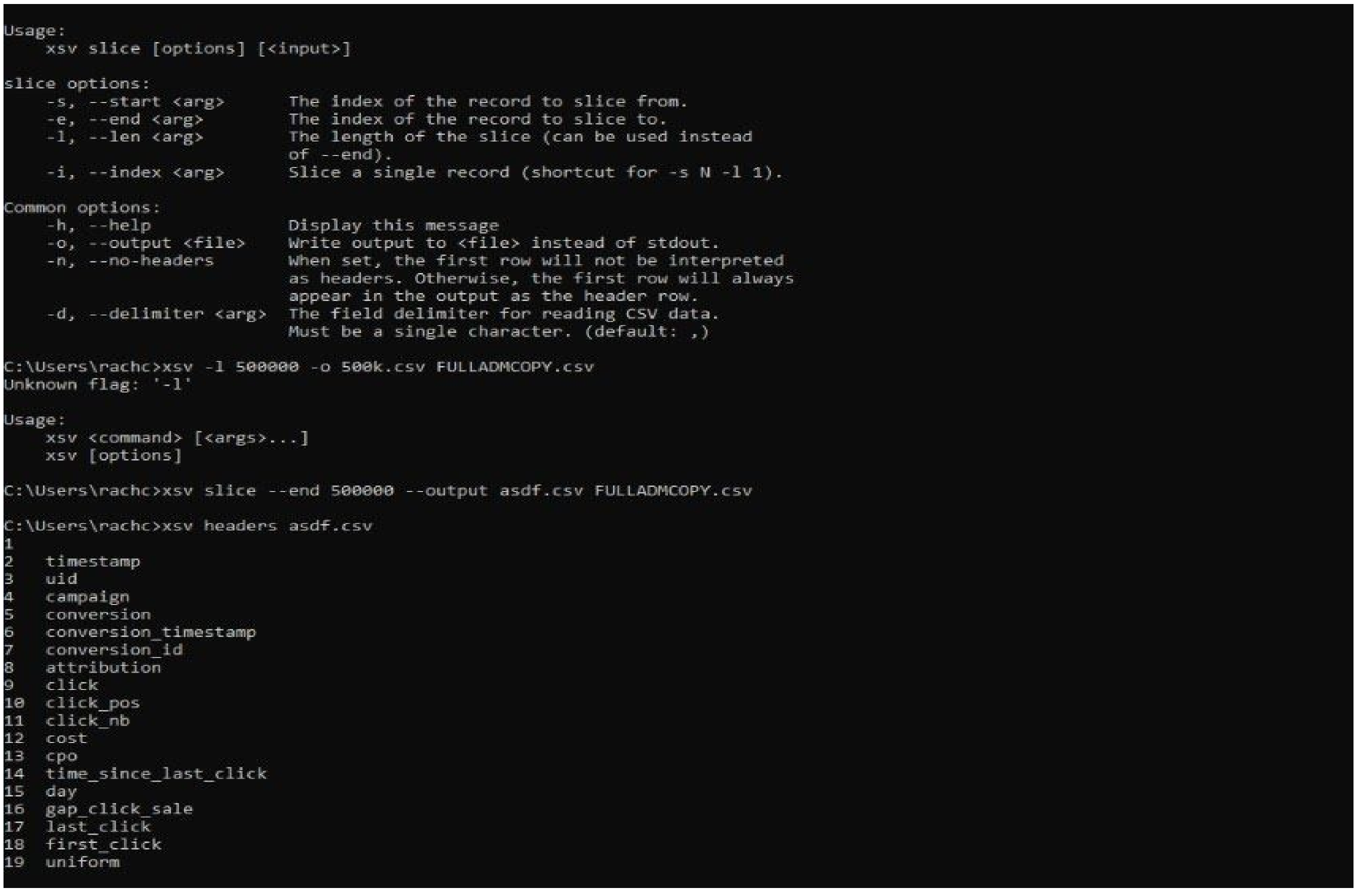
```
(base) C:\Users\19293>
(base) C:\Users\19293>xsv stats C:\Users\19293\Desktop\ADM\criteo_attribution_dataset\pcb_dataset_final.tsv | xsv table
field                  type     sum                 min      max               min_length  max_length  mean                   stddev
timestamp              Integer  21662697362152      0        2671199           1           7           1315439.7525672794     769770.0361270809
uid                    Integer  267401057636714     13       32458754          2           8           16237589.216772163     9373751.359085169
campaign               Integer  279692387314654     73322    32452111          5           8           16983964.58268292      9700052.225449245
conversion             Integer  806196              0        1                 1           1           0.04895522699835389    0.2157744487836664
conversion_timestamp   Integer  1563478121339       -1       5262888           2           7           94940.22091042617      478966.63744865305
conversion_id          Integer  13073723398784      -1       32458519          2           8           793885.2297718121      4064784.1720765587
attribution            Integer  442424              0        1                 1           1           0.026865634845022728   0.16169066920944608
click                  Integer  5947563             0        1                 1           1           0.3611582006758508     0.4803362934032628
click_pos              Integer  -13689309           -1       173               1           3           -0.8312658826708125    1.5322206203197763
click_nb               Integer  -10911742           -1       174               1           3           -0.6626016583527264    2.696254130340931
cost                   Float    4829.340541025379   0.00001  0.0583448264308   5           17          0.000293255563702158   0.0008689670963295013
cpo                    Float    3234792.6306751645  0.004    1.01631051174     5           16          0.19642866936420986    0.11863821555204436
time_since_last_click  Integer  4468753182520       -1       2592000           1           7           271359.35485891264     527310.8765171622
cat1                   Integer  362745446968541     138937   30763035          6           8           22027256.02578344      12107310.172802933
cat2                   Integer  241869634590874     138937   32440053          6           8           14687226.016261801     9122111.559270142
cat3                   Integer  250858543819470     577      32457986          3           8           15233066.099508194     9847417.062123684
cat4                   Integer  470410983488597     358249   32145478          6           8           28565108.831108402     2698653.3011883767
cat5                   Integer  318482790236901     138937   32440053          6           8           19339462.47701044      11746115.865884133
cat6                   Integer  248586757094861     138937   32440053          6           8           15095114.739296196     13406408.32272267
cat7                   Integer  250864392790457     150      32458469          3           8           15233421.270832762     9002237.420802243
cat8                   Integer  408663884955533     3225256  32440044          7           8           24815594.785912186     8254684.271293756
cat9                   Integer  391463926601699     358246   32145483          6           8           23771149.18513154      7778014.745109545
```

The frequency command gives the frequency, i.e, count of occurrence of values in various columns.



```
(base) C:\Users\19293>xsv frequency C:\Users\19293\Desktop\ADM\criteo_attribution_dataset\pcb_dataset_final.tsv
field,value,count
timestamp,1009184,30
timestamp,1196516,28
timestamp,477409,27
timestamp,2234553,27
timestamp,498368,27
timestamp,413325,26
timestamp,1191374,26
timestamp,501681,26
timestamp,410198,26
timestamp,415148,26
uid,8826511,880
uid,1402083,528
uid,2370705,478
uid,16452391,365
uid,5101234,327
uid,29262375,324
uid,19153609,321
uid,23974566,298
uid,22313205,289
uid,2813279,285
campaign,10341182,437385
campaign,30801593,431587
campaign,17686799,381084
campaign,15398570,378464
campaign,5061834,299755
campaign,15184511,256102
campaign,29427842,239272
campaign,28351001,222470
campaign,18975823,217646
campaign,31772643,195759
conversion,0,15661831
conversion,1,806196
conversion_timestamp,-1,15661831
conversion_timestamp,1892511,164
conversion_timestamp,2357932,142
conversion_timestamp,2183492,109
```

Slicing the data as per requirements



```
Usage:
    xsv slice [options] [<input>]

slice options:
    -s, --start <arg>      The index of the record to slice from.
    -e, --end <arg>        The index of the record to slice to.
    -l, --len <arg>        The length of the slice (can be used instead
                           of --end).
    -i, --index <arg>      Slice a single record (shortcut for -s N -l 1).

Common options:
    -h, --help             Display this message
    -o, --output <file>    Write output to <file> instead of stdout.
    -n, --no-headers       When set, the first row will not be interpreted
                           as headers. Otherwise, the first row will always
                           appear in the output as the header row.
    -d, --delimiter <arg>  The field delimiter for reading CSV data.
                           Must be a single character. (default: ,)

C:\Users\rachc>xsv -l 500000 -o 500k.csv FULLADMCOPY.csv
Unknown flag: '-l'

Usage:
    xsv <command> [<args>...]
    xsv [options]

C:\Users\rachc>xsv slice --end 500000 --output asdf.csv FULLADMCOPY.csv

C:\Users\rachc>xsv headers asdf.csv
1
2   timestamp
3   uid
4   campaign
5   conversion
6   conversion_timestamp
7   conversion_id
8   attribution
9   click
10  click_pos
11  click_nb
12  cost
13  cpo
14  time_since_last_click
15  day
16  gap_click_sale
17  last_click
18  first_click
19  uniform
```

TRIFACTA

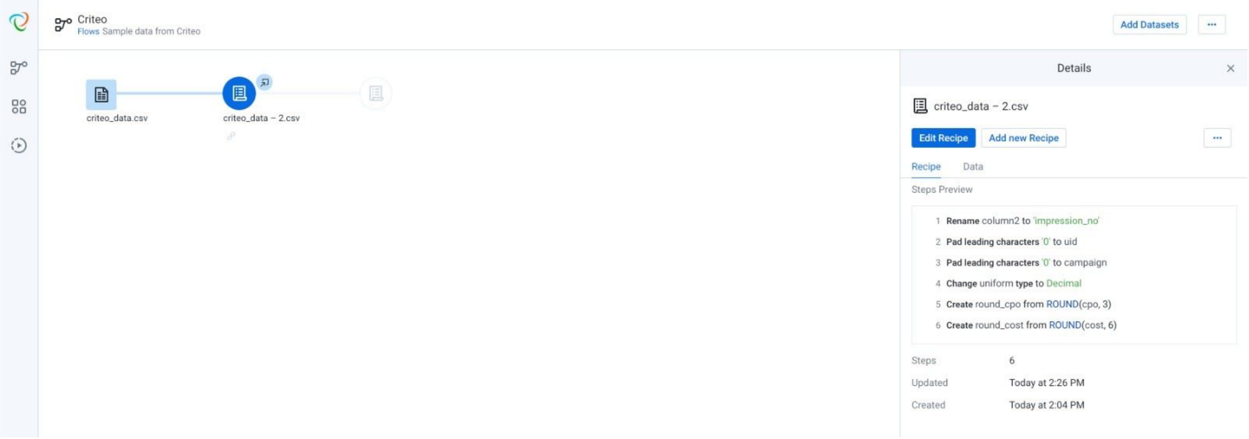
Trifacta develops data wrangling software for data exploration and self-service data preparation for analysis. Trifacta works with cloud and on-premises data platforms. Trifacta is designed for analysts to explore, transform, and enrich raw data into clean and structured formats.

Strengths:

- It has an interactive UI and intelligent execution
- It offers suggestions based on various columns
- The recipe view is very handy and multiple recipe’s can be created on multiple datasets.
- The flow views show an overview of what is going on
- We can generate a report of our dataset

Weaknesses:

- It allows only 100MB of data to be used on the free trial version
- You have to manually select all the data files that you want to combine which can be tedious when handling large number of files

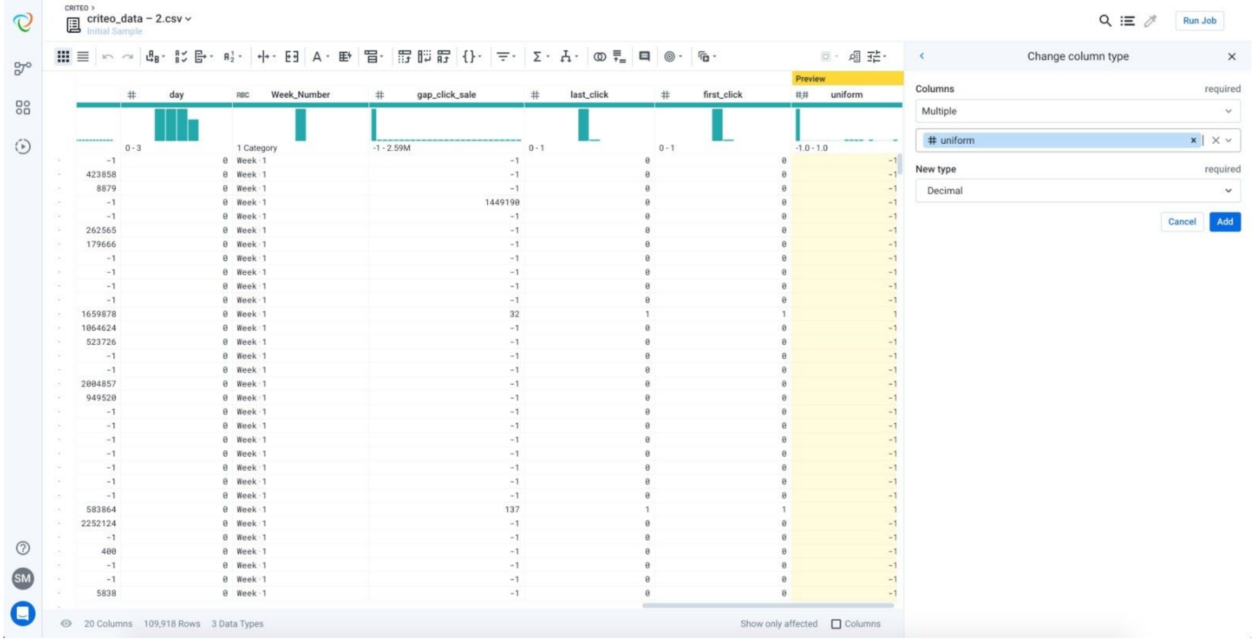


criteo_data – 2.csv

Steps Preview

1 **Rename** column2 to 'impression_no'
2 **Pad leading characters** '0' to uid
3 **Pad leading characters** '0' to campaign
4 **Change** uniform **type** to Decimal
5 **Create** round_cpo from ROUND(cpo, 3)
6 **Create** round_cost from ROUND(cost, 6)

Steps 6
Updated Today at 2:26 PM
Created Today at 2:04 PM



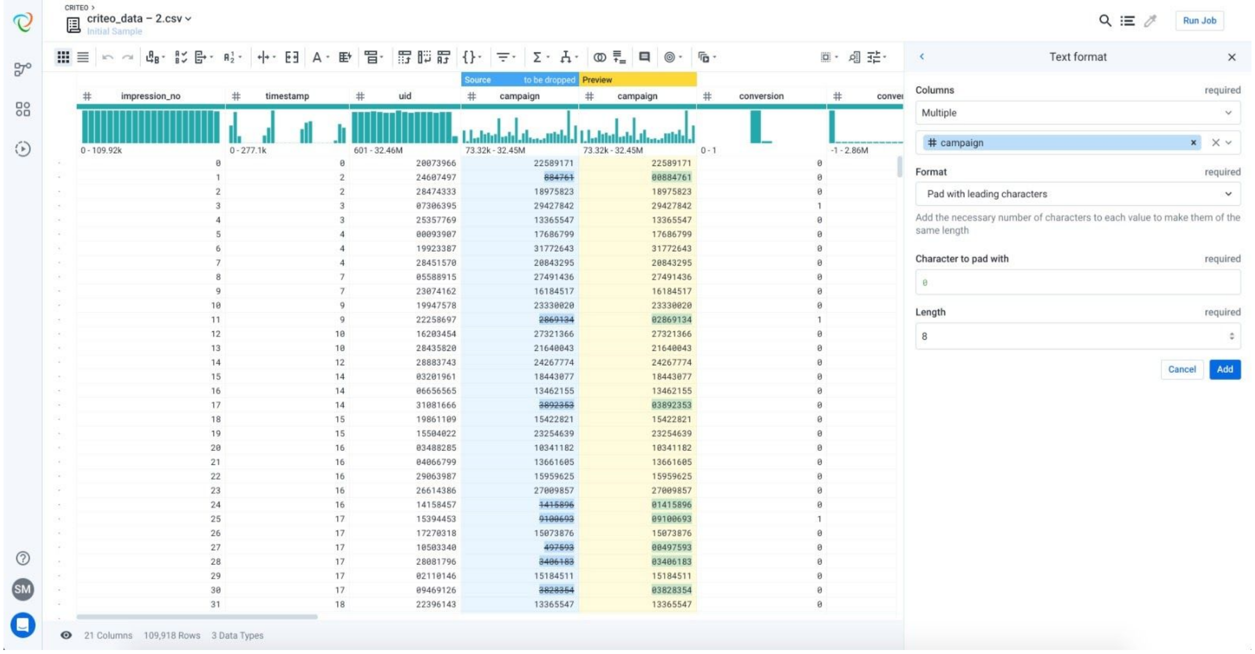
	day	Week_Number	gap_click_sale	last_click	first_click	uniform
-1	0	Week 1	-1	0	0	-1
423858	0	Week 1	-1	0	0	-1
8879	0	Week 1	-1	0	0	-1
-1	0	Week 1	1449190	0	0	-1
-1	0	Week 1	-1	0	0	-1
262565	0	Week 1	-1	0	0	-1
179666	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
1659878	0	Week 1	32	1	1	1
1064624	0	Week 1	-1	0	0	-1
523726	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
2004857	0	Week 1	-1	0	0	-1
949520	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
583864	0	Week 1	137	1	1	1
2252124	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
400	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
-1	0	Week 1	-1	0	0	-1
5838	0	Week 1	-1	0	0	-1

20 Columns 109,918 Rows 3 Data Types

Show only affected □ Columns

Columns required
Multiple
uniform

New type required
Decimal



CRITEO >
criteo_data – 2.csv
Initial Sample

			Source to be dropped	Preview		
# impression_no	# timestamp	# uid	# campaign	# campaign	# conversion	# conver
0 - 109.92k	0 - 277.1k	601 - 32.46M	73.32k - 32.45M	73.32k - 32.45M	0 - 1	-1 - 2.86M
0	0	20073966	22589171	22589171	0	
1	2	24607497	~~884761~~	00884761	0	
2	2	28474333	18975823	18975823	0	
3	3	07306395	29427842	29427842	1	
4	3	25357769	13365547	13365547	0	
5	4	00093907	17686799	17686799	0	
6	4	19923387	31772643	31772643	0	
7	4	28451570	20843295	20843295	0	
8	7	05588915	27491436	27491436	0	
9	7	23074162	16184517	16184517	0	
10	9	19947578	23330020	23330020	0	
11	9	22258697	~~2869134~~	02869134	1	
12	10	16203454	27321366	27321366	0	
13	10	28435820	21640043	21640043	0	
14	12	28883743	24267774	24267774	0	
15	14	03201961	18443077	18443077	0	
16	14	06656565	13462155	13462155	0	
17	14	31081666	~~3892353~~	03892353	0	
18	15	19861109	15422821	15422821	0	
19	15	15504022	23254639	23254639	0	
20	16	03488285	10341182	10341182	0	
21	16	04066799	13661605	13661605	0	
22	16	29063987	15959625	15959625	0	
23	16	26614386	27009857	27009857	0	
24	16	14158457	~~1415896~~	01415896	0	
25	17	15394453	~~9100693~~	09100693	1	
26	17	17270318	15073876	15073876	0	
27	17	10503340	~~497593~~	00497593	0	
28	17	28081796	~~3406183~~	03406183	0	
29	17	02110146	15184511	15184511	0	
30	17	09469126	~~3828354~~	03828354	0	
31	18	22396143	13365547	13365547	0	

21 Columns 109,918 Rows 3 Data Types

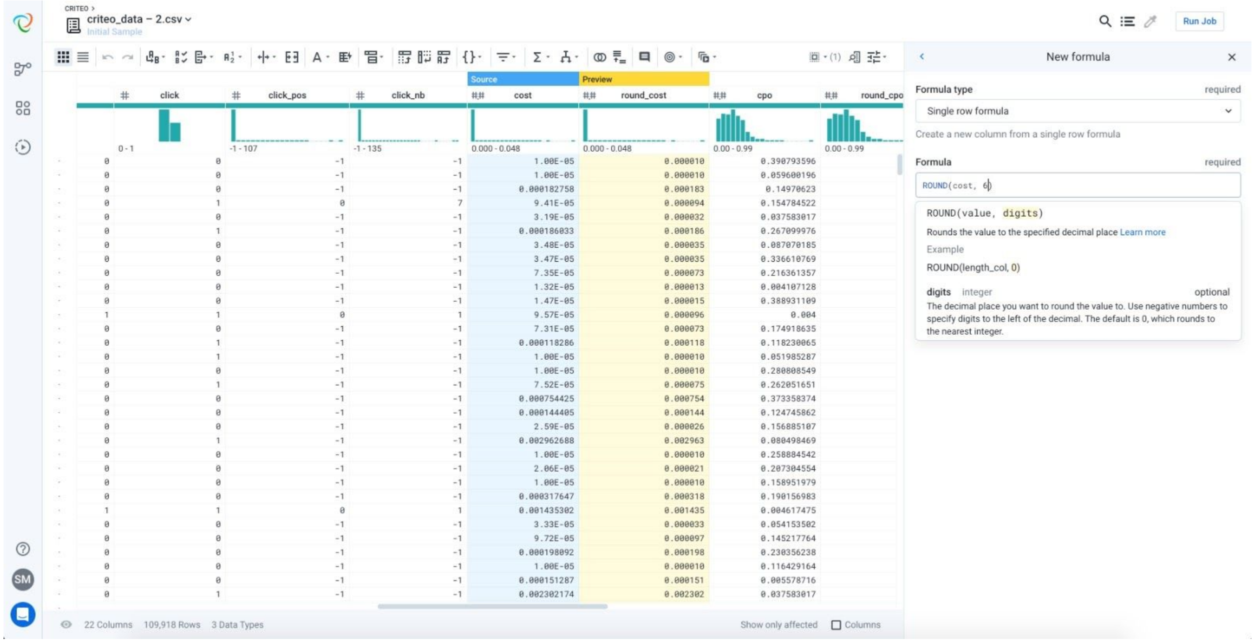
Text format

Columns required
Multiple
# campaign

Format required
Pad with leading characters
Add the necessary number of characters to each value to make them of the same length

Character to pad with required
0

Length required
8



CRITEO >
criteo_data – 2.csv
Initial Sample

				Source	Preview		
	# click	# click_pos	# click_nb	## cost	## round_cost	## cpo	## round_cpo
	0 - 1	-1 - 107	-1 - 135	0.000 - 0.048	0.000 - 0.048	0.00 - 0.99	0.00 - 0.99
0	0	-1	-1	1.00E-05	0.000010	0.390793596	
0	0	-1	-1	1.00E-05	0.000010	0.059600196	
0	0	-1	-1	0.000182758	0.000183	0.14970623	
0	1	0	7	9.41E-05	0.000094	0.154784522	
0	0	-1	-1	3.19E-05	0.000032	0.037583017	
0	1	-1	-1	0.000186033	0.000186	0.267099976	
0	0	-1	-1	3.48E-05	0.000035	0.087070185	
0	0	-1	-1	3.47E-05	0.000035	0.336610769	
0	0	-1	-1	7.35E-05	0.000073	0.216361357	
0	0	-1	-1	1.32E-05	0.000013	0.004107128	
0	0	-1	-1	1.47E-05	0.000015	0.388931109	
1	1	0	1	9.57E-05	0.000096	0.004	
0	0	-1	-1	7.31E-05	0.000073	0.174918635	
0	1	-1	-1	0.000118286	0.000118	0.118230065	
0	1	-1	-1	1.00E-05	0.000010	0.051985287	
0	0	-1	-1	1.00E-05	0.000010	0.280808549	
0	1	-1	-1	7.52E-05	0.000075	0.262051651	
0	0	-1	-1	0.000754425	0.000754	0.373358374	
0	0	-1	-1	0.000144405	0.000144	0.124745862	
0	0	-1	-1	2.59E-05	0.000026	0.156885107	
0	1	-1	-1	0.002962688	0.002963	0.080498469	
0	0	-1	-1	1.00E-05	0.000010	0.258884542	
0	0	-1	-1	2.06E-05	0.000021	0.207304554	
0	0	-1	-1	1.00E-05	0.000010	0.158951979	
0	0	-1	-1	0.000317647	0.000318	0.190156983	
1	1	0	1	0.001435302	0.001435	0.004617475	
0	0	-1	-1	3.33E-05	0.000033	0.054153502	
0	0	-1	-1	9.72E-05	0.000097	0.145217764	
0	0	-1	-1	0.000198092	0.000198	0.230356238	
0	0	-1	-1	1.00E-05	0.000010	0.116429164	
0	0	-1	-1	0.000151287	0.000151	0.005578716	
0	1	-1	-1	0.002302174	0.002302	0.037583017	

22 Columns 109,918 Rows 3 Data Types Show only affected ☐ Columns

New formula

Formula type required
Single row formula
Create a new column from a single row formula

Formula required

```
ROUND(cost, 6)
```

```
ROUND(value, digits)
```

Rounds the value to the specified decimal place Learn more

Example

ROUND(length_col, 0)

digits integer optional
The decimal place you want to round the value to. Use negative numbers to specify digits to the left of the decimal. The default is 0, which rounds to the nearest integer.

PANDAS - PYTHON

Pandas is a software library written for the Python programming language for data manipulation and analysis

Strengths:

- Pandas provide extremely streamlined forms of data representation. This helps to analyze and understand data better.
- Less writing and more work done
- Pandas is very powerful and has an extensive set of features
- Very useful for customizing and editing the data

Weaknesses:

- The syntax can be really tedious sometimes, and remembering the syntax is another task!
- It can take very long to process large data sets
- You need to be known to the programming language to get your hands on python, unlike the other tools.

SNOWFLAKE

Snowflake is a powerful relational database management system. It is offered as an analytic data warehouse for both structured and semi-structured data that follows a Software-as-a-Service (SaaS) model.

Strengths:

- Very simple and easy to run SQL like commands
- The data is stored on cloud
- It is easy to share data between different accounts
- It is very user friendly and is compatible with lots of other technologies

Weaknesses:

- It doesn’t handle on-premise data very well

SALESFORCE EINSTEIN ANALYTICS

Strengths:

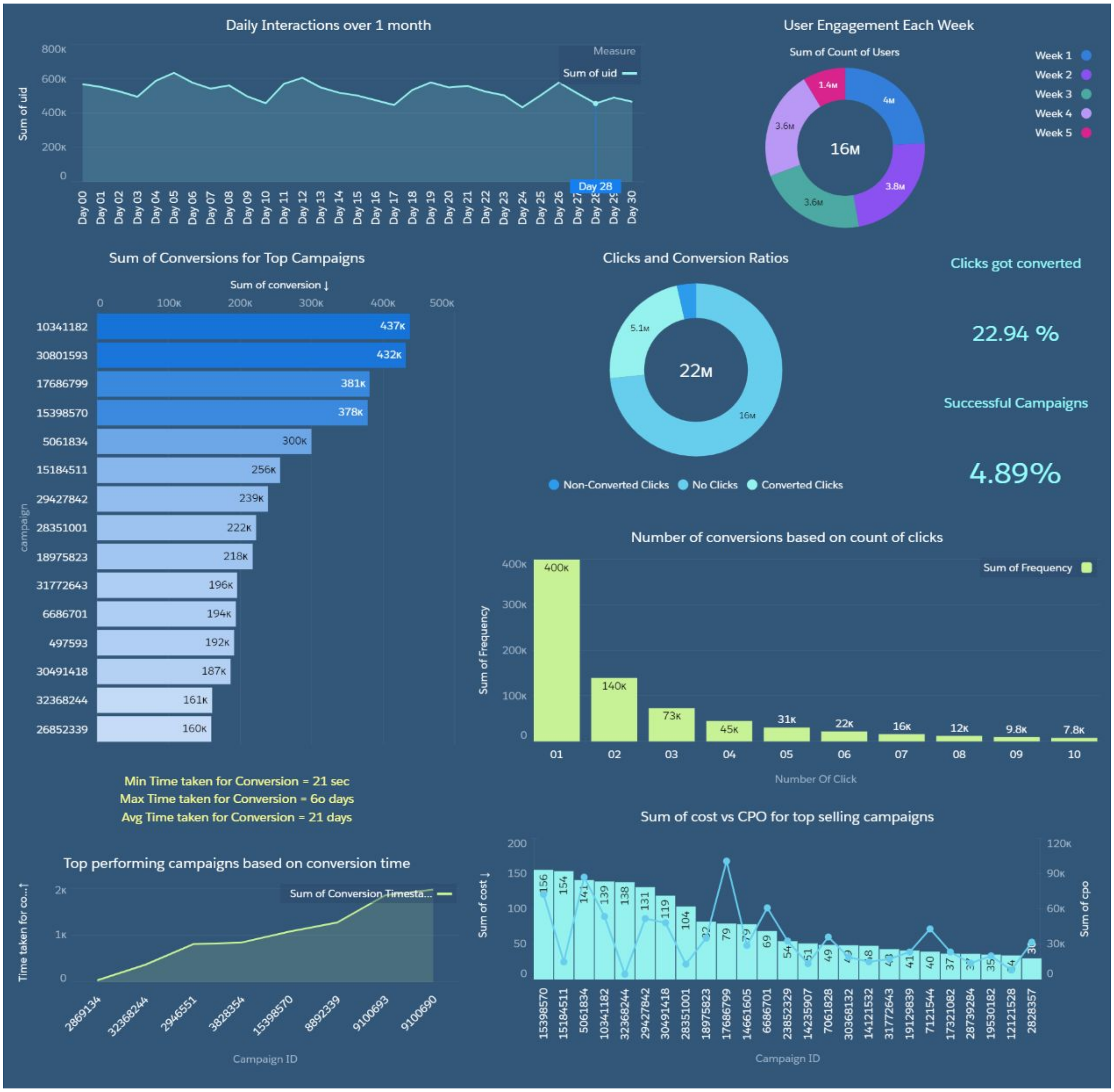
- It helps us connect to various platforms to fetch the data

- The user experience is very elegant and the dashboards have a wonderful design
- It has a lot of useful features such as generating reports feature

Weaknesses:

- It cannot handle large data sets easily
- It has limited support and is pricey

Dashboard:



HOW CRITOE WORKS:

- Deliver the right ad at the right moment in the shopper journey. A custom piece of code placed on your site enables the Criteo AI Engine to see shoppers' engagement and power product recommendations in your ads.

- Gain access to the best ad inventory available. With thousands of the world's top publishers in our open Commerce Marketing Ecosystem, you get better placements across leading sites.

- Drive more sales from visitors who leave your website without making a purchase. Personalized offers, delivered at just the right time and in the right format, can bring this pool of shoppers back.

ANALYSIS

- Daily Interactions : We analyzed over the period of 30 days how each day was performing, what was the user interaction on each day with respect to Criteo’s banner advertisements.

- Best Performing Week/User Interaction each week: It gives us an insight into which week has the most customer interactions based on how many people clicked on Criteo’s dynamic ad services.

- Top performing campaigns : Which campaigns performed the best during the 30 days of Criteo’s live traffic helps us with knowing that the particular campaigns were most successful and how many conversions took place from those campaigns.

- Clicks and Conversions : While analyzing the clicks and conversions , we can conclude that 22.94% of clicks that were made on Criteo banners were converted, which further tells us about the success percent of Criteo Campaigns.

- Approximately 400k user interactions were converted in 0 clicks!