Useful Commands:

Drop all the null values, dropna()
Fill null values with a constant number, train.fillna(-1).show()
Drop Duplicates: dropDuplicates()
Count Null, empty & NA values in a column

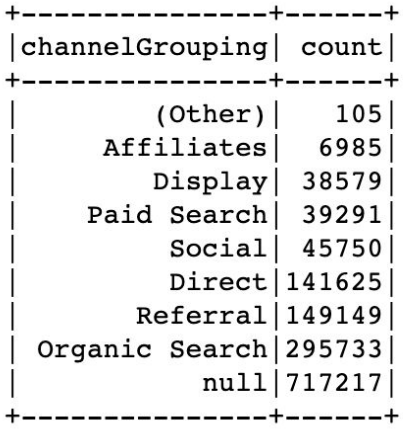
For continuous features

display(diabetes_df.select(fn.count('*'), fn.avg('age'), fn.min('age'), fn.max('age'), fn.stddev('age')))
df.describe().show()

For categorical features

We can calculate pair-wise frequences: train.crosstab(‘age’, ‘gender’).show()
df.cube("x").count().show()

https://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html#module-pyspark.sql.functions
PySpark official guide on SQL functions



```
+---------------+------+
|channelGrouping| count|
+---------------+------+
|        (Other)|   105|
|     Affiliates|  6985|
|        Display| 38579|
|    Paid Search| 39291|
|         Social| 45750|
|         Direct|141625|
|       Referral|149149|
| Organic Search|295733|
|           null|717217|
+---------------+------+
```

|-- channelGrouping[1]: string (nullable = true)

Affiliates:

Affiliates of Google, medium exactly matches affiliates

Display

Interactions with a medium of "display" or "cpm". Also includes Google Ads interactions with ad distribution network set to "content".

Paid Search

Traffic from the Google Ads Search Network or other search engines, with a medium of "cpc" or "ppc"

Direct Link:

Sessions in which the user typed the name of your website URL into the browser or came to your site via a bookmark (i.e., source="(direct)" and medium="(not set)" or "(none)").

Social Network:

Traffic from any of approximately 400 social networks (that are not tagged as ads).

[1] https://support.google.com/analytics/answer/3297892?hl=en
https://support.google.com/analytics/answer/1191184?hl=en
https://support.google.com/analytics/answer/3437719?hl=en This is the most detailed explanation for all variables

Organic Search

Traffic from unpaid search on any search engine (i.e., medium="organic").

Referral

Traffic from websites that are not social networks.

*Each category is pretty important, we'll put null with others, and them create dummy variables for all the them

|-- date: integer (nullable = true)

Drop the date column|-- fullVisitorId: decimal(20,0) (nullable = true)

|-- fullVisitorId: decimal(20,0) (nullable = true)

This is basically the ID of the dataset

|-- socialEngagementType: string (nullable = true)

All not socially engaged, dropped

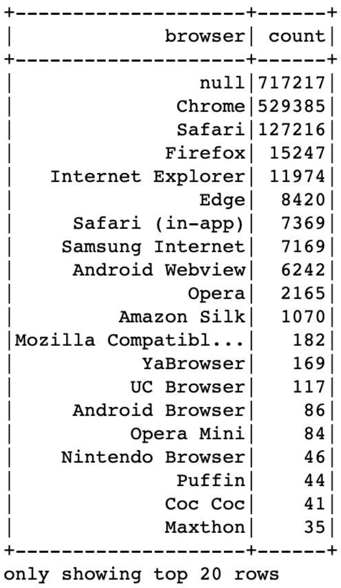
|-- visitId: integer (nullable = true)

|-- visitNumber: integer (nullable = true)

|-- browser: string (nullable = true)

We saved 3 types, Chrome, Safari, & Firefox

The new variable name is **Nbrowers**

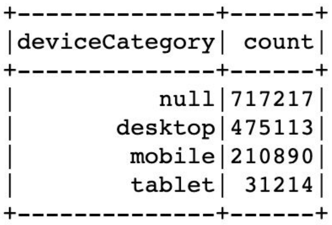


```
+--------------------+------+
|             browser| count|
+--------------------+------+
|                null|717217|
|              Chrome|529385|
|              Safari|127216|
|             Firefox| 15247|
|   Internet Explorer| 11974|
|                Edge|  8420|
|     Safari (in-app)|  7369|
|    Samsung Internet|  7169|
|     Android Webview|  6242|
|               Opera|  2165|
|         Amazon Silk|  1070|
|Mozilla Compatibl...|   182|
|           YaBrowser|   169|
|          UC Browser|   117|
|     Android Browser|    86|
|          Opera Mini|    84|
|    Nintendo Browser|    46|
|              Puffin|    44|
|             Coc Coc|    41|
|             Maxthon|    35|
+--------------------+------+
only showing top 20 rows
```

|-- deviceCategory: string (nullable = true)

Here, there are 3 categories

Desktop, mobile, and tablet

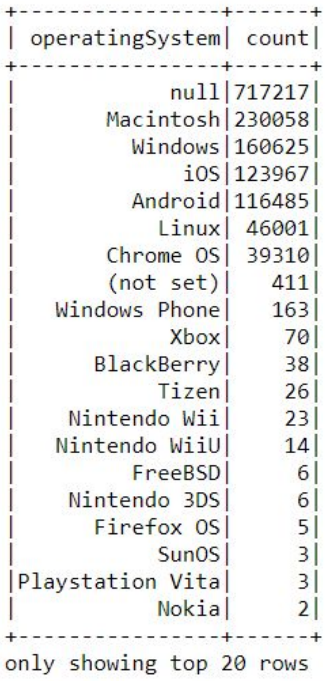


```
+--------------+------+
|deviceCategory| count|
+--------------+------+
|          null|717217|
|       desktop|475113|
|        mobile|210890|
|        tablet| 31214|
+--------------+------+
```

~~|-- isMobile: boolean (nullable = true)~~

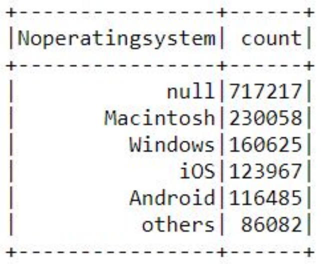
This is overlapped with device

|-- operatingSystem: string (nullable = true)



```
+-----------------+------+
|  operatingSystem| count|
+-----------------+------+
|             null|717217|
|        Macintosh|230058|
|          Windows|160625|
|              iOS|123967|
|          Android|116485|
|            Linux| 46001|
|        Chrome OS| 39310|
|        (not set)|   411|
|    Windows Phone|   163|
|             Xbox|    70|
|       BlackBerry|    38|
|            Tizen|    26|
|     Nintendo Wii|    23|
|    Nintendo WiiU|    14|
|          FreeBSD|     6|
|     Nintendo 3DS|     6|
|       Firefox OS|     5|
|            SunOS|     3|
|Playstation Vita|     3|
|            Nokia|     2|
+-----------------+------+
only showing top 20 rows
```

The new varaible’s name is **Noperatingsystem**



```
+----------------+------+
|Noperatingsystem| count|
+----------------+------+
|            null|717217|
|       Macintosh|230058|
|         Windows|160625|
|             iOS|123967|
|         Android|116485|
|          others| 86082|
+----------------+------+
```

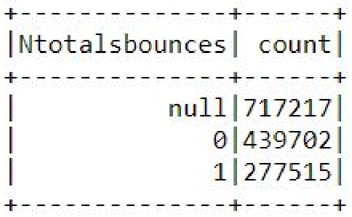
|-- country: string (nullable = true)

Dataset is already filtered by the United States

|-- totalsbounces: integer (nullable = true)

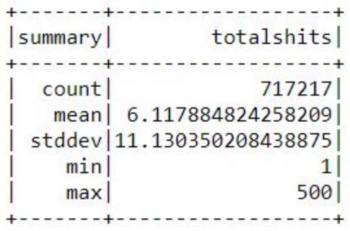
There are Nulls in the data, replaced with 0

The new feature is: **Ntotalsbounces**



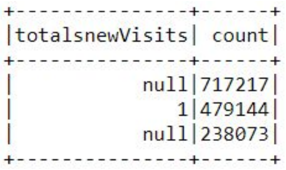
```
+--------------+------+
|Ntotalsbounces| count|
+--------------+------+
|          null|717217|
|             0|439702|
|             1|277515|
+--------------+------+
```

|-- totalshits: integer (nullable = true)



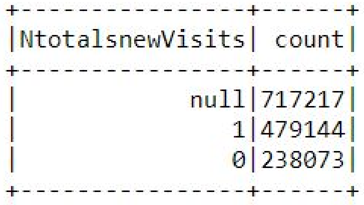
```
+-------+-----------------+
|summary|        totalshits|
+-------+-----------------+
|  count|           717217|
|   mean| 6.117884824258209|
| stddev|11.130350208438875|
|    min|                1|
|    max|              500|
+-------+-----------------+
```

|-- totalsnewVisits: integer (nullable = true)



```
+---------------+------+
|totalsnewVisits| count|
+---------------+------+
|           null|717217|
|              1|479144|
|           null|238073|
+---------------+------+
```

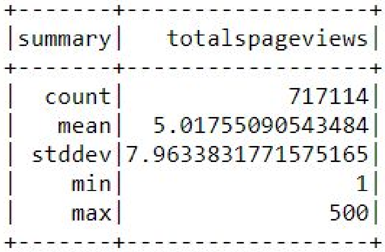
After replacement,



```
+----------------+------+
|NtotalsnewVisits| count|
+----------------+------+
|            null|717217|
|               1|479144|
|               0|238073|
+----------------+------+
```

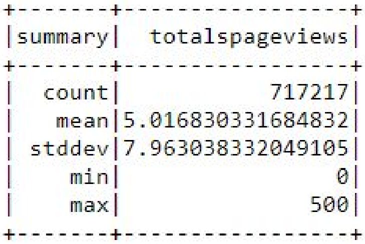
|-- totalspageviews: integer (nullable = true)

103 Nulls



```
+-------+------------------+
|summary|   totalspageviews|
+-------+------------------+
|  count|            717114|
|   mean|  5.01755090543484|
| stddev|7.9633831771575165|
|    min|                 1|
|    max|               500|
+-------+------------------+
```

After replace all Nulls with 0

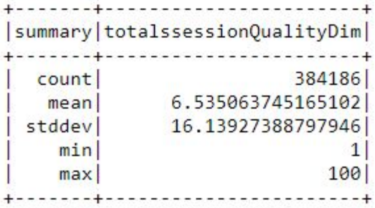


```
+-------+-----------------+
|summary|  totalspageviews|
+-------+-----------------+
|  count|           717217|
|   mean|5.016830331684832|
| stddev|7.963038332049105|
|    min|                0|
|    max|              500|
+-------+-----------------+
```

|-- totalssessionQualityDim: integer (nullable = true)

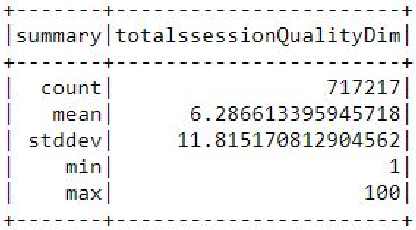
This feature could be highly correlated with previous features

On the right is the state before making any changes, 333031 Nulls into average



```
+-------+-----------------------+
|summary|totalssessionQualityDim|
+-------+-----------------------+
|  count|                 384186|
|   mean|      6.535063745165102|
| stddev|      16.13927388797946|
|    min|                      1|
|    max|                    100|
+-------+-----------------------+
```

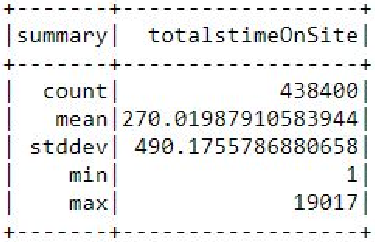
Below is the stat after adjusting Nulls



```
+-------+-----------------------+
|summary|totalssessionQualityDim|
+-------+-----------------------+
|  count|                 717217|
|   mean|      6.286613395945718|
| stddev|     11.815170812904562|
|    min|                      1|
|    max|                    100|
+-------+-----------------------+
```

|-- totalstimeOnSite: integer (nullable = true)

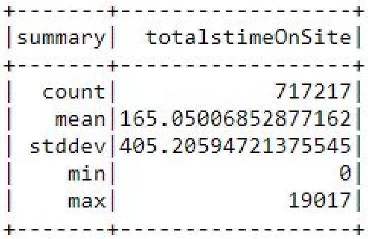
Fill Nulls with 0, on the right is the stat before the adjustment.



```
+-------+------------------+
|summary|  totalstimeOnSite|
+-------+------------------+
|  count|            438400|
|   mean|270.01987910583944|
| stddev| 490.1755786880658|
|    min|                 1|
|    max|             19017|
+-------+------------------+
```

278817 Nulls

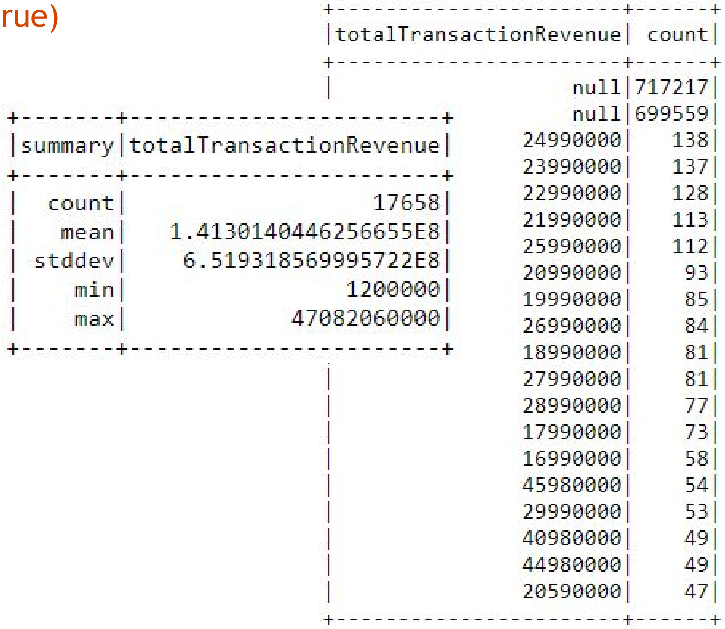
Below is the stat that’s after the Null adjustment



```
+-------+------------------+
|summary|  totalstimeOnSite|
+-------+------------------+
|  count|            717217|
|   mean|165.05006852877162|
| stddev|405.20594721375545|
|    min|                 0|
|    max|             19017|
+-------+------------------+
```

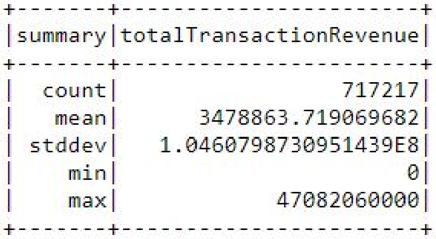
|-- totalTransactionRevenue: long (nullable = true)

There are 699559 Nulls, 17658 with values
0.02462016377 =17658 /699559
On the right, it's the stats without adjustment.
It's reasonal to assume 0 for Null values
Aslo, the top totaltransaction revenue is extremely high, it could be caused by internal pruchases of Google Swags by Google.
Below is the stat after the change



```
+--------------------------+------+
|totalTransactionRevenue| count|
+--------------------------+------+
|                     null|717217|
|                     null|699559|
|                 24990000|   138|
|                 23990000|   137|
|                 22990000|   128|
|                 21990000|   113|
|                 25990000|   112|
|                 20990000|    93|
|                 19990000|    85|
|                 26990000|    84|
|                 18990000|    81|
|                 27990000|    81|
|                 28990000|    77|
|                 17990000|    73|
|                 16990000|    58|
|                 45980000|    54|
|                 29990000|    53|
|                 40980000|    49|
|                 44980000|    49|
|                 20590000|    47|
+--------------------------+------+
```

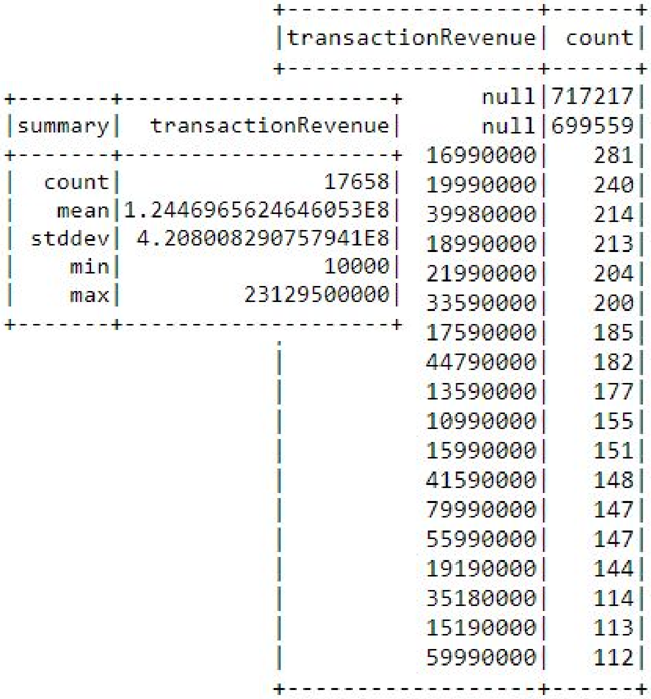
```
+-------+-----------------------+
|summary|totalTransactionRevenue|
+-------+-----------------------+
|  count|                  17658|
|   mean|   1.413014044625665E8|
| stddev|   6.519318569995722E8|
|    min|                1200000|
|    max|            47082060000|
+-------+-----------------------+
```



```
+-------+-----------------------+
|summary|totalTransactionRevenue|
+-------+-----------------------+
|  count|                 717217|
|   mean|      3478863.719069682|
| stddev|  1.046079873095143E8|
|    min|                      0|
|    max|            47082060000|
+-------+-----------------------+
```

|-- transactionRevenue: long (nullable = true)

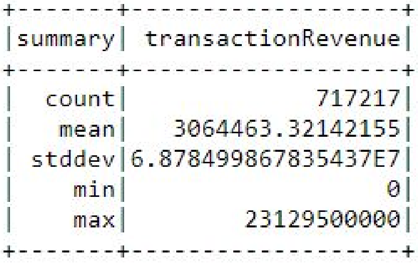
The similar logic applies to transactionRevenue, here we replace Null with 0s.
Over the right is the state before adjustment for Null Values



```
+------------------+------+
|transactionRevenue| count|
+------------------+------+
|              null|717217|
|              null|699559|
|          16990000|   281|
|          19990000|   240|
|          39980000|   214|
|          18990000|   213|
|          21990000|   204|
|          33590000|   200|
|          17590000|   185|
|          44790000|   182|
|          13590000|   177|
|          10990000|   155|
|          15990000|   151|
|          41590000|   148|
|          79990000|   147|
|          55990000|   147|
|          19190000|   144|
|          35180000|   114|
|          15190000|   113|
|          59990000|   112|
+------------------+------+
```

```
+-------+--------------------+
|summary|  transactionRevenue|
+-------+--------------------+
|  count|               17658|
|   mean|1.2446965624646053E8|
| stddev| 4.208008290757941E8|
|    min|               10000|
|    max|         23129500000|
+-------+--------------------+
```

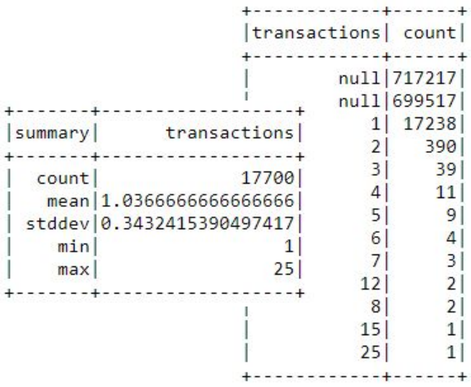
Below is the stat after the adjustment



```
+-------+------------------+
|summary|transactionRevenue|
+-------+------------------+
|  count|            717217|
|   mean|  3064463.32142155|
| stddev|6.878499867835437E7|
|    min|                 0|
|    max|       23129500000|
+-------+------------------+
```

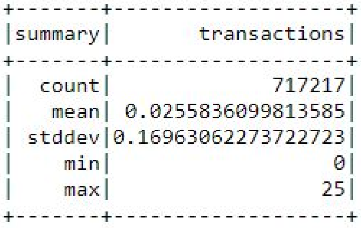
|-- transactions: integer (nullable = true)

Replace Nulls with 0. Stat before the change over the right
New Stat



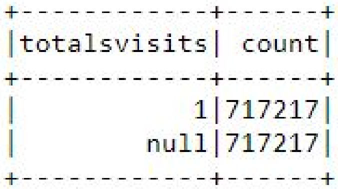
```
+------------+------+
|transactions| count|
+------------+------+
|        null|717217|
|        null|699517|
|           1| 17238|
|           2|   390|
|           3|    39|
|           4|    11|
|           5|     9|
|           6|     4|
|           7|     3|
|          12|     2|
|           8|     2|
|          15|     1|
|          25|     1|
+------------+------+
```

```
+-------+------------------+
|summary|      transactions|
+-------+------------------+
|  count|             17700|
|   mean|1.0366666666666666|
| stddev|0.3432415390497417|
|    min|                 1|
|    max|                25|
+-------+------------------+
```



```
+-------+-------------------+
|summary|       transactions|
+-------+-------------------+
|  count|             717217|
|   mean| 0.0255836099813585|
| stddev|0.16963062273722723|
|    min|                  0|
|    max|                 25|
+-------+-------------------+
```

|-- totalsvisits: integer (nullable = true)



```
+------------+------+
|totalsvisits| count|
+------------+------+
|           1|717217|
|        null|717217|
+------------+------+
```

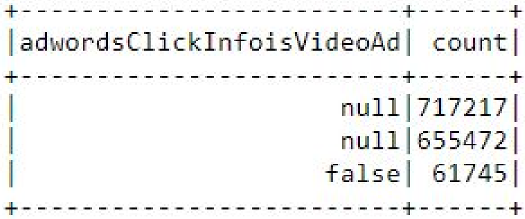
|-- adwordsClickInfoisVideoAd: boolean (nullable = true)

*Here information about video advertisement is perserved, by passing this into the model, it could improve acurracy.

1 -> Video has affects

0 -> Video has no affects

Detail info/interpretation about this feature is lost.



```
+-------------------------+------+
|adwordsClickInfoisVideoAd| count|
+-------------------------+------+
|                     null|717217|
|                     null|655472|
|                    false| 61745|
+-------------------------+------+
```

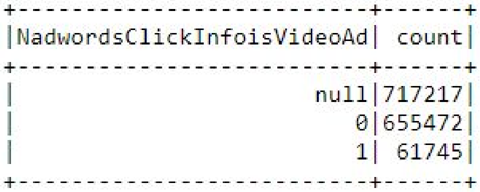
Assign 0 to Nulls, 1 to “False”.

Over the right is the stat before any change.

The new feature: **NadwordsClickInfoisVideoAd**

Below is the new stat after the adjustment

I.E.



```
+--------------------------+------+
|NadwordsClickInfoisVideoAd| count|
+--------------------------+------+
|                      null|717217|
|                         0|655472|
|                         1| 61745|
+--------------------------+------+
```

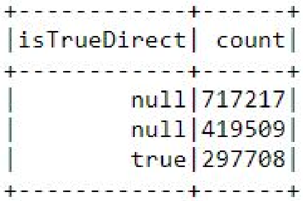
|-- isTrueDirect: boolean (nullable = true)

We'll assign 0 to Null, 1 to False

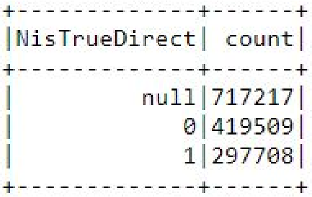
419509 Nulls

Over the right is the stat before the adjustment

*(From Google)True if the source of the session was Direct (meaning the user typed the name of your website URL into the browser or came to your site via a bookmark), This field will also be true if 2 successive but distinct sessions have exactly the same campaign details. Otherwise NULL.

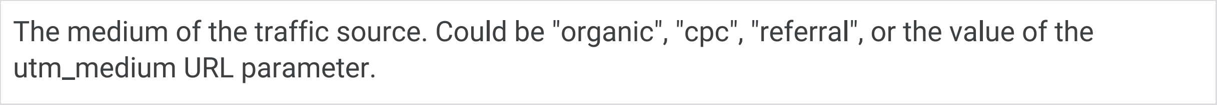


```
+------------+------+
|isTrueDirect| count|
+------------+------+
|        null|717217|
|        null|419509|
|        true|297708|
+------------+------+
```



```
+-------------+------+
|NisTrueDirect| count|
+-------------+------+
|         null|717217|
|            0|419509|
|            1|297708|
+-------------+------+
|-- trafficSourcemedium: string (nullable = true)
```

NtrafficSourcemedium is the new feature.


The medium of the traffic source. Could be "organic", "cpc", "referral", or the value of the utm_medium URL parameter.

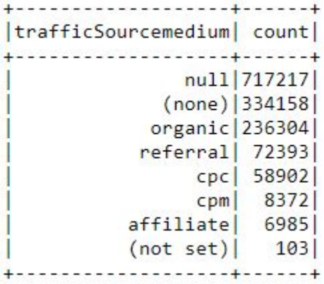
“organic” (unpaid search)

“cpc” (cost per click, i.e. paid search)

“referral” (referral)

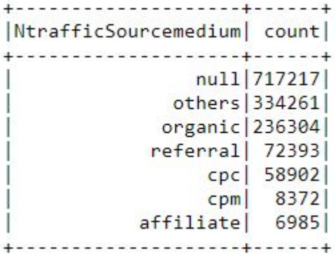
CPM, Definition: Cost Per Thousand Impressions – A way to bid where you pay per one thousand views (impressions) on the Google Display Network.

Cpm & cpc are two different ways of bitting google ad services


```
+-------------------+------+
|trafficSourcemedium| count|
+-------------------+------+
|               null|717217|
|             (none)|334158|
|            organic|236304|
|           referral| 72393|
|                cpc| 58902|
|                cpm|  8372|
|          affiliate|  6985|
|          (not set)|   103|
+-------------------+------+
```

Over the right is the stat before making any change.

Below is the stat that’s after the change


```
+--------------------+------+
|NtrafficSourcemedium| count|
+--------------------+------+
|                null|717217|
|              others|334261|
|             organic|236304|
|            referral| 72393|
|                 cpc| 58902|
|                 cpm|  8372|
|           affiliate|  6985|
+--------------------+------+
```

```
|-- trafficSourcesource: string (nullable = true)
```

NtrafficSourcesource is the new feature.

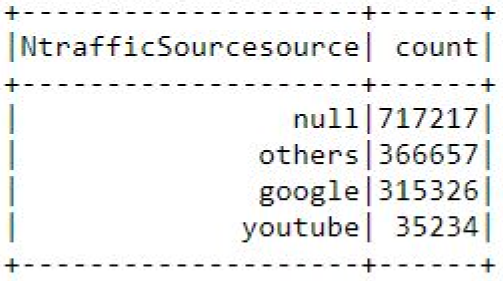
Here, we plan to put

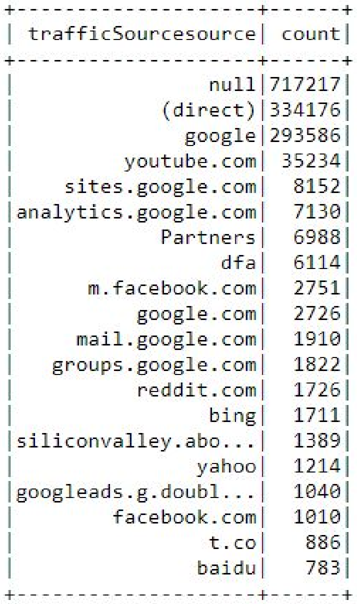
Googles -> google

Youtube.com -> youtube

Others -> others

Below is the stat after the adjustment


```
+--------------------+------+
|NtrafficSourcesource| count|
+--------------------+------+
|                null|717217|
|              others|366657|
|              google|315326|
|             youtube| 35234|
+--------------------+------+
```


```
+--------------------+------+
| trafficSourcesource| count|
+--------------------+------+
|                null|717217|
|            (direct)|334176|
|              google|293586|
|         youtube.com| 35234|
|    sites.google.com|  8152|
|analytics.google.com|  7130|
|            Partners|  6988|
|                 dfa|  6114|
|      m.facebook.com|  2751|
|          google.com|  2726|
|     mail.google.com|  1910|
|   groups.google.com|  1822|
|          reddit.com|  1726|
|                bing|  1711|
|siliconvalley.abo...|  1389|
|               yahoo|  1214|
|googleads.g.doubl...|  1040|
|        facebook.com|  1010|
|                t.co|   886|
|               baidu|   783|
+--------------------+------+
```