In [7]:
```
import os
import json
import pickle
import numpy as np
import pandas as pd
from tqdm import tqdm_notebook
import matplotlib.pyplot as plt
from catboost import CatBoostRegressor

# Displaying pd Dataframe options
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
pd.set_option('display.max_colwidth', 1000)
```

```
/anaconda3/lib/python3.6/importlib/_bootstrap.py:219: RuntimeWarning: numpy.dtype size changed, may indicate binary incompatibility. Expected 96, got 88
  return f(*args, **kwds)
/anaconda3/lib/python3.6/importlib/_bootstrap.py:219: RuntimeWarning: numpy.dtype size changed, may indicate binary incompatibility. Expected 96, got 88
  return f(*args, **kwds)
```

In [8]:
```
%%time
# Get datas from pickle
train_df = pd.read_pickle('train_df.pickle')
train_df['date'] = pd.to_datetime(train_df['date'], format='%Y%m%d')
train_df['weekday'] = train_df['date'].dt.weekday_name
train_df['day_off'] = (train_df['date'].dt.dayofweek > 4).astype(int)
```

```
CPU times: user 2.96 s, sys: 855 ms, total: 3.81 s
Wall time: 3.86 s
```

In [9]:
```
%%time
test_df = pd.read_pickle('test_df.pickle')
test_df['date'] = pd.to_datetime(test_df['date'], format='%Y%m%d')
test_df['weekday'] = test_df['date'].dt.weekday_name
test_df['day_off'] = (test_df['date'].dt.dayofweek > 4).astype(int)
```

```
CPU times: user 2.54 s, sys: 792 ms, total: 3.33 s
Wall time: 3.38 s
```

In [10]:
```
def check_diff_in_dfs():
    print('train_df \ test_df', set(train_df.columns).difference(set(test_df.columns)))
    print('test_df \ train_df', set(test_df.columns).difference(set(train_df.columns)))
```

In [11]:
```
# Data revision
check_diff_in_dfs()
```

```
train_df \ test_df {'transactionRevenue'}
test_df \ train_df set()
```

на пример adContent, где очень мало заполнено данных создать рядом новую фичу 'заполнена ли adContent'. Может быть сам adContent дропуть, поскольку она не несет информацию. Отступаем от обучения, работаем с данными.

In [12]:
```
'channelGrouping'
```

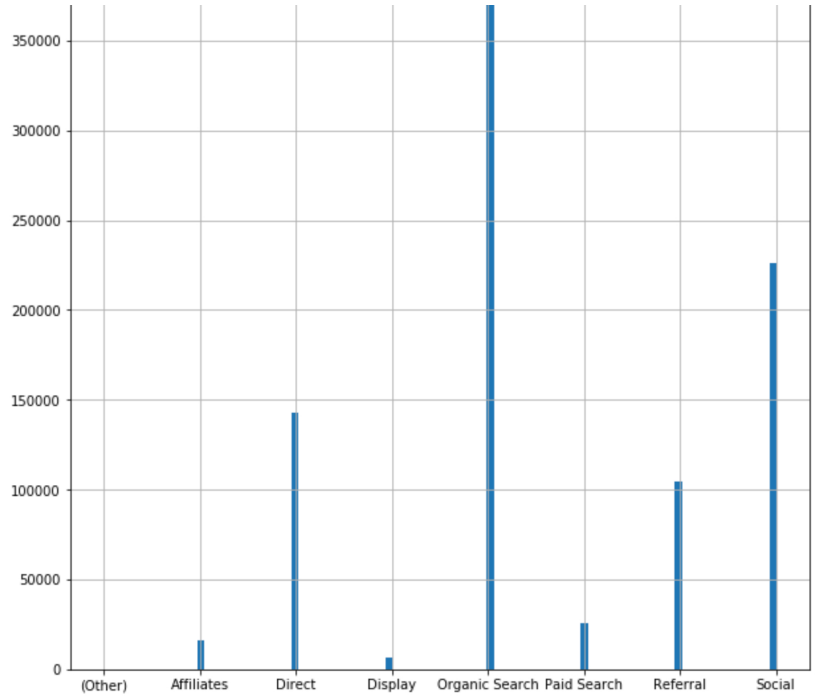
Out[12]:
```
'channelGrouping'
```

In [13]:
```
train_df['channelGrouping'].value_counts()
```

Out[13]:
```
Organic Search    381561
Social            226117
Direct            143026
Referral          104838
Paid Search        25326
Affiliates         16403
Display             6262
(Other)              120
Name: channelGrouping, dtype: int64
```

In [14]:
```
%matplotlib inline
train_df.channelGrouping.hist(bins=100, figsize=(10,10))
```

Out[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1158702e8>
```



In [ ]:
```
'date'
```

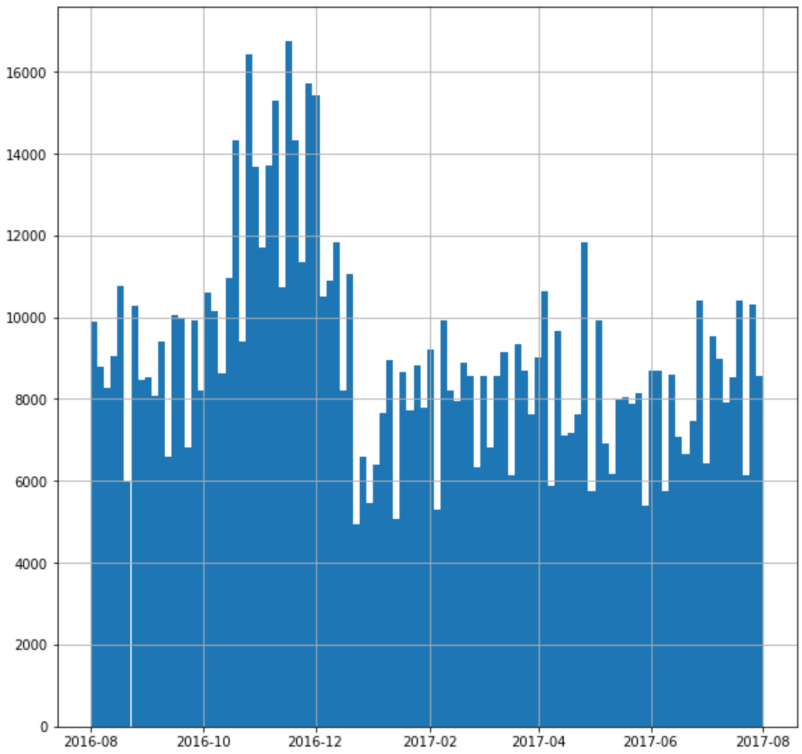
In [15]:
```
train_df['date'].value_counts()
```

Out[15]:
```
2016-11-28    4807
2016-11-15    4685
2016-11-14    4466
2016-11-30    4435
2016-10-26    4375
2016-11-29    4337
2016-11-16    4334
2016-10-04    4322
2016-12-05    4265
2017-04-26    4224
2016-12-01    4200
2016-10-27    4162
2016-11-21    4143
2016-11-17    4074
2016-10-24    4063
2016-11-10    4055
2016-11-03    4014
2016-11-02    3960
2016-11-22    3942
2016-11-08    3899
2016-10-25    3842
2016-11-07    3832
2016-10-31    3827
2016-11-23    3805
2016-10-28    3791
2016-10-05    3770
2016-11-24    3770
2016-11-25    3759
2016-10-20    3755
2016-11-09    3752
              ...
2016-08-13    1596
2017-04-30    1594
2016-10-01    1589
2016-12-26    1586
2017-01-15    1576
2017-04-29    1566
2017-06-11    1555
2017-02-04    1549
2017-04-23    1548
2017-06-10    1545
2017-06-04    1534
2017-01-14    1526
2017-02-05    1522
2017-06-24    1510
2017-04-16    1507
2017-04-15    1506
2017-05-27    1502
2017-05-28    1463
2017-06-18    1432
2017-05-07    1400
2017-06-03    1399
2017-06-17    1391
```

```
2016-12-25    1386
2017-05-06    1383
2017-01-01    1364
2017-05-14    1290
2017-05-13    1251
2016-12-30    1232
2016-12-24    1231
2016-12-31    1211
Name: date, Length: 366, dtype: int64
```

In [16]:
```
%matplotlib inline
train_df.date.hist(bins=100, figsize=(10,10))
```

Out[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x19520f978>
```



In [3]:
```
'fullVisitorId' # ???
```

Out[3]:
```
'fullVisitorId'
```

In [17]:
```
train_df['fullVisitorId'].value_counts()
```

Out[17]:
```
1957458976293878100    278
8248397261184852 74    205
3608475193341679870    201
1856749147915772585    199
3269834865385146569    155
7634897085866546110    148
4038076683036146727    138
3694234028523165868    129
720311197761340948     121
6254908847172458133    117
3525537916960843419    115
3148617623907142276    112
6018775317735347795    111
9801276214964695322    110
2194592743396253647    107
232377434237234751     105
7498695963354635199    104
9609104828919391966     99
3937673380007666721     94
949718915643445721      91
4913801338365738862     90
2082625651279391786     89
7813149961404844386     86
7445235885559107095     83
4578640586284138624     81
6588500311054802771     76
3041133261614133977     74
5208937953046059083     72
9292327473106748702     71
9534537897118577344     71
                       ...
84628143632300127        1
5260117902407993152      1
6574240968048973826      1
```

```
9756885146234423606      1
0035338492120820250      1
873475769283421870       1
2582631781060890714      1
6409334789258306643      1
278370119809134195       1
13112254498225448        1
755478055824777498       1
3870411131466688891      1
1981255727278868759      1
6352068296250840290      1
1374669736121576302      1
567693922642677756       1
6335270223693537138      1
0999162464548654721      1
594164222613551627       1
7761147609779274483      1
5063485259887289626      1
9064845677734789328      1
8533784760176859006      1
2762862025874979842      1
1371773997701402132      1
8416529278258295030      1
7928189135460156192      1
5993502009410362381      1
4965853732342736067      1
8241106111400388501      1
Name: fullVisitorId, Length: 716924, dtype: int64
```

In [18]:
```
%matplotlib inline
#????
```

In [ ]:
```
'sessionId'
```

In [19]:
```
train_df['sessionId'].value_counts()
```

Out[19]:
```
4686096939646334160_1482911798    2
2656385077897383189_1473403798    2
5580260050052139637_1470380267    2
4483741101095290795_1471157564    2
2788824330199765490_1489560554    2
9926456995105457200_1472626397    2
5135775780419153299_1477205290    2
4163776473564568453_1496299236    2
9847307838447511234_1472713116    2
5961536486573634219_1476080711    2
0064585292262671874_1493620320    2
8596239481909424450_1487318035    2
684380858040966297_1477465101     2
7849236513149694148_1491461947    2
7047642838587768548_1497163538    2
419524103074603391_1471589669     2
1876136855853663327_1493881074    2
6563911580191009917_1471589823    2
9096282942669500075_1472106044    2
1505413932671276932_1495090533    2
6187281368753177854_1471935527    2
932491977900282496_1478156313     2
0821104645538474369_1478764781    2
357023552915423846_1500964877     2
5980617790873815212_1493276349    2
4682115121818915623_1473058728    2
5539996708481290888_1482306425    2
1468326551501903507_1480492763    2
8795944303443767835_1481011145    2
8721068465074288178_1473404368    2
                                 ..
8927187548900118849_1480345214    1
4889512885884147913_1492614517    1
7428846393390011907_1486953393    1
5907666072487495293_1472781396    1
9480230640899140010_1499613170    1
7022531201626857911_1475641353    1
1702969702621864203_1473636520    1
3784758895707507381_1494794421    1
253454431071066526_1492065322     1
5389062134224524315_1490875830    1
0070407945682650508_1493042107    1
220177451040378591_1476392328     1
1424355572088327868_1501013364    1
5711582711607566547_1501062354    1
114917623082779326_1487345064     1
213003977022600203_1470361175     1
2190889485144456356_1485306203    1
9949171640993720841_1490132642    1
7956806160537425O2_1486987859     1
3234319072209247431_1480424068    1
0874035486462890331_1497909899    1
0354112508038877105_1499502404    1
```

```
5719161159863692412_1493672881    1
0420156595432368551_1496921060    1
0136570880626144528_1501317315    1
9393123630317318368_1481116366    1
3645563003369778834_1497856006    1
4240165280056468872_1498273524    1
6792687366845509283_1499642527    1
754524501110741865_1496166469     1
Name: sessionId, Length: 902755, dtype: int64
```

In [20]:
```
%matplotlib inline
#????
```

In [ ]:
```
'visitId'
```

In [21]:
```
train_df['visitId'].value_counts()
```

Out[21]:
```
1493146175    8
1478345904    6
1481369525    6
1484649802    6
1500856602    5
1495031359    5
1494374199    5
1473406997    4
1478942540    4
1490264449    4
1499778118    4
1489390306    4
1473406994    4
1490255839    4
1485823856    4
1490899032    4
1481365730    4
1478243627    4
1500856594    4
1500856591    4
1490274628    4
1475719086    4
1490384962    4
1484649800    4
1474613955    4
1489390296    4
1493406494    4
1479668456    4
1481874207    4
1500880897    4
              ..
1500254858    1
1482755722    1
1476466313    1
1474518720    1
1478710977    1
1472415427    1
1470258912    1
1474465530    1
1474490102    1
1476577011    1
1500606546    1
1491209962    1
1480728296    1
1474455271    1
1478651622    1
1470265061    1
1491347172    1
1495500507    1
1487107781    1
1490718365    1
1470361302    1
1476654805    1
1485035217    1
1478698703    1
1497432928    1
1478704844    1
1487085256    1
1500456028    1
1480814278    1
1497372675    1
Name: visitId, Length: 886303, dtype: int64
```

In [22]:
```
%matplotlib inline
#????
```

In [23]:
```
'visitNumber'
```

Out[23]:
```
'visitNumber'
```

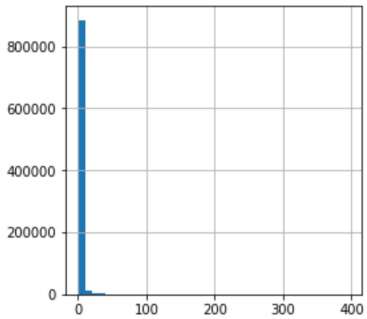
In [24]:
```
train_df['visitNumber'].value_counts()
```

In [24]: train_df['visitNumber'].value_counts()

Out[24]:
```
1      703060
2       92548
3       35843
4       19157
5       11615
6        7677
7        5413
8        4031
9        3084
10       2415
11       1936
12       1573
13       1292
14       1092
15        928
16        809
17        699
18        611
19        541
20        497
21        446
22        399
23        355
24        340
25        304
26        272
27        245
28        223
29        203
30        188
        ...
310         1
304         1
313         1
314         1
316         1
290         1
394         1
327         1
328         1
329         1
333         1
334         1
335         1
337         1
352         1
338         1
339         1
340         1
341         1
342         1
343         1
344         1
345         1
346         1
347         1
306         1
349         1
350         1
351         1
395         1
Name: visitNumber, Length: 384, dtype: int64
```

In [32]:
```
%matplotlib inline
train_df.visitNumber.hist(bins=40, figsize=(4,4))
```

Out[32]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x147ee9358>
```



In [ ]:
```
'visitStartTime'
```

In [33]:
```
train_df['visitStartTime'].value_counts()
```

Out[33]:
```
1483146175    8
```

Out[33]:
```
1493146175    8
1481369525    6
1478345904    6
1484649802    6
1500856602    5
1495031359    5
1494374199    5
1475719086    4
1485823856    4
1499778118    4
1478243627    4
1477470782    4
1490384962    4
1500856594    4
1478942540    4
1489390296    4
1484649800    4
1498823774    4
1490264767    4
1500856591    4
1500856605    4
1479300875    4
1473406994    4
1481365730    4
1473406997    4
1488904421    4
1479668456    4
1481874207    4
1500880897    4
1490255839    4
              ..
1473763415    1
1482147925    1
1496825940    1
1500600455    1
1496391816    1
1475422345    1
1487937704    1
1498642627    1
1481859264    1
1485875385    1
1500567735    1
1496371382    1
1485891761    1
1475352751    1
1471156398    1
1481652395    1
1481650346    1
1481658534    1
1494290574    1
1477462181    1
1500528804    1
1485863075    1
1479567521    1
1490131103    1
1488027804    1
1475457178    1
1481756822    1
1496445074    1
1498540177    1
1497372675    1
Name: visitStartTime, Length: 887159, dtype: int64
```

In [34]:
```
%matplotlib inline
#????
```

In [ ]:
```
'operatingSystem'
```

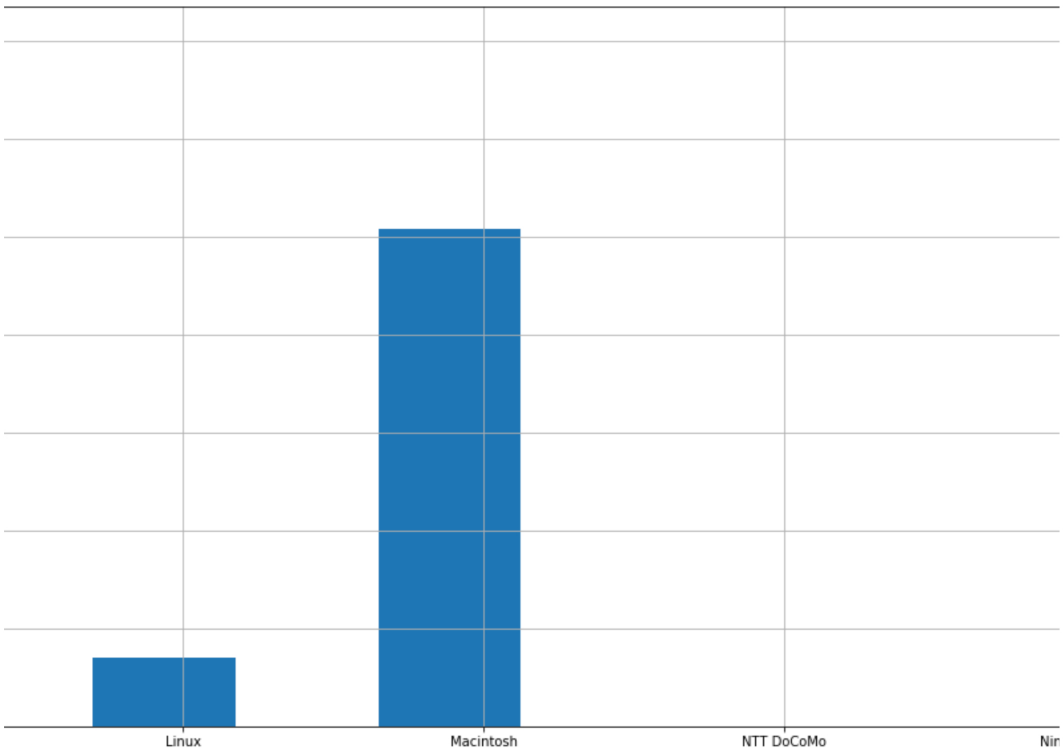
In [35]:
```
train_df['operatingSystem'].value_counts()
```

Out[35]:
```
Windows          350072
Macintosh        253938
Android          123892
iOS              107665
Linux             35034
Chrome OS         26337
(not set)          4695
Windows Phone      1216
Samsung             280
BlackBerry          218
Nintendo Wii        100
Firefox OS           89
Xbox                 66
Nintendo WiiU        35
FreeBSD               9
Nokia                 2
OpenBSD               2
Nintendo 3DS          1
NTT DoCoMo            1
```

```
SunOS                    1
Name: operatingSystem, dtype: int64
```

In [53]:
```
%matplotlib inline
train_df.operatingSystem.hist(bins=40, figsize=(85,10))
```

Out[53]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1b28522b0>
```



In [ ]:
```
'isMobile'
```

In [54]:
```
train_df['isMobile'].value_counts()
```

Out[54]:
```
False    664530
True     239123
Name: isMobile, dtype: int64
```

In [57]:
```
%matplotlib inline
#????
```

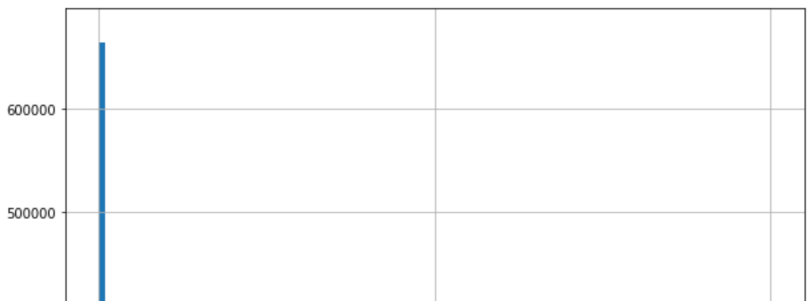
In [ ]:
```
'deviceCategory'
```

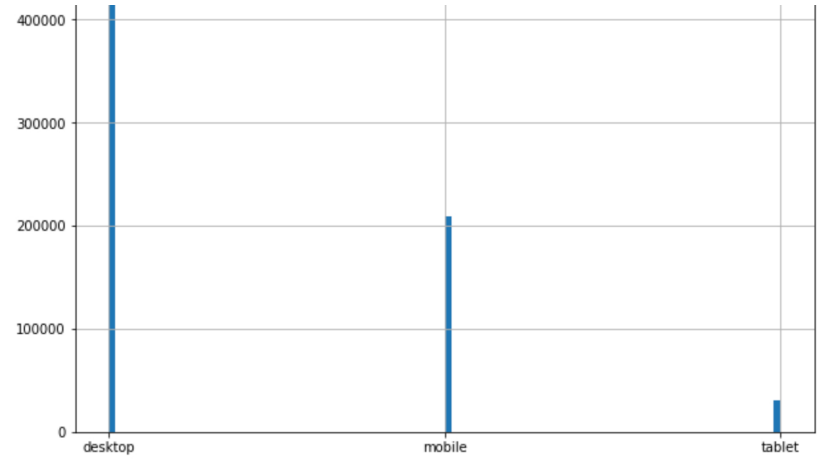
In [58]:
```
train_df['deviceCategory'].value_counts()
```

Out[58]:
```
desktop    664479
mobile     208725
tablet      30449
Name: deviceCategory, dtype: int64
```

In [59]:
```
%matplotlib inline
train_df.deviceCategory.hist(bins=100, figsize=(10,10))
```

Out[59]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1cbfc8400>
```





In [ ]:
```
'browser'
```

In [60]:
```
train_df['browser'].value_counts()
```

Out[60]:
```
Chrome                                          620364
Safari                                          182245
Firefox                                          37069
Internet Explorer                                19375
Edge                                             10205
Android Webview                                   7865
Safari (in-app)                                   6850
Opera Mini                                        6139
Opera                                             5643
UC Browser                                        2427
YaBrowser                                         2096
Coc Coc                                            727
Amazon Silk                                        561
Android Browser                                    553
Mozilla Compatible Agent                           374
MRCHROME                                           263
Maxthon                                            246
BlackBerry                                         184
Nintendo Browser                                   140
Puffin                                              93
Nokia Browser                                       67
Iron                                                33
LYF_LS_4002_12                                      21
SeaMonkey                                           15
Seznam                                              11
Mozilla                                             11
Apple-iPhone7C2                                      9
(not set)                                            8
Nichrome                                             7
0                                                    7
Lunascape                                            5
osee2unifiedRelease                                  5
DASH_JR_3G                                           4
LYF_LS_4002_11                                       3
ThumbSniper                                          3
no-ua                                                3
NokiaE52-1                                           2
YE                                                   2
MQQBrowser                                           2
Android Runtime                                      2
Konqueror                                            1
HTC802t_TD                                           1
TCL P500M                                            1
ADM                                                  1
Changa 99695759                                      1
M5                                                   1
CSM Click                                            1
User Agent                                           1
subjectAgent: NoticiasBoom                           1
Reddit                                               1
Hisense M20-M_LTE                                    1
DoCoMo                                               1
IE with Chrome Frame                                 1
[Use default User-agent string] LIVRENPOCHE          1
Name: browser, dtype: int64
```

In [61]:
```
%matplotlib inline
#?
```

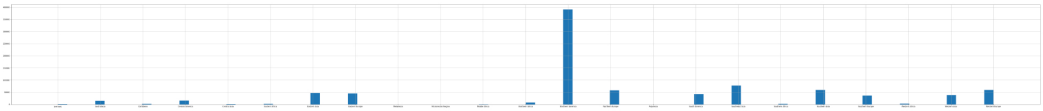
In [ ]:
```
'subContinent'
```

In [62]:
```
train_df['subContinent'].value_counts()
```

Out[62]:
```
Northern America      390657
Southeast Asia         77800
Southern Asia          59321
Western Europe         59114
Northern Europe        58168
Eastern Asia           46919
Eastern Europe         45249
South America          41731
Western Asia           38443
Southern Europe        35780
Central America        15583
Australasia            14893
Northern Africa         7683
Western Africa          2573
Caribbean               2406
Southern Africa         2169
Eastern Africa          1927
(not set)               1468
Central Asia            1215
Middle Africa            393
Melanesia                 81
Micronesian Region        55
Polynesia                 25
Name: subContinent, dtype: int64
```

In [64]:
```
%matplotlib inline
train_df.subContinent.hist(bins=100, figsize=(100,10))
```

Out[64]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e4174b70>
```



In [ ]:
```
'country'
```

In [65]:
```
train_df['country'].value_counts()
```

Out[65]:
```
United States               364744
India                        51140
United Kingdom               37393
Canada                       25869
Vietnam                      24598
Turkey                       20522
Thailand                     20123
Germany                      19980
Brazil                       19783
Japan                        19731
France                       15832
Mexico                       13225
Taiwan                       12996
Australia                    12698
Russia                       11662
Spain                        11658
Netherlands                  11453
Italy                        11332
Poland                        9693
Indonesia                     9273
Philippines                   9244
Singapore                     7172
Ireland                       6493
Malaysia                      6439
Romania                       6428
Ukraine                       5577
Israel                        5563
Peru                          5546
Sweden                        5315
South Korea                   5237
                             ...
Timor-Leste                      7
Liechtenstein                    7
Lesotho                          6
Turkmenistan                     5
Antigua & Barbuda                5
Greenland                        5
British Virgin Islands           5
Caribbean Netherlands            4
Guinea-Bissau                    4
Equatorial Guinea                4
Central African Republic         4
San Marino                       4
Vanuatu                          3
Isle of Man                      3
Seychelles                       3
Dominica                         3
Congo - Brazzaville              3
Cook Islands                     3
Marshall Islands                 2
```

```
Marshall Islands                2
Comoros                         2
St. Barthélemy                  1
St. Pierre & Miquelon           1
São Tomé & Príncipe             1
American Samoa                  1
Norfolk Island                  1
St. Martin                      1
Åland Islands                   1
Samoa                           1
Eritrea                         1
Anguilla                        1
Name: country, Length: 222, dtype: int64
```

In [66]:
```
%matplotlib inline
#????
```

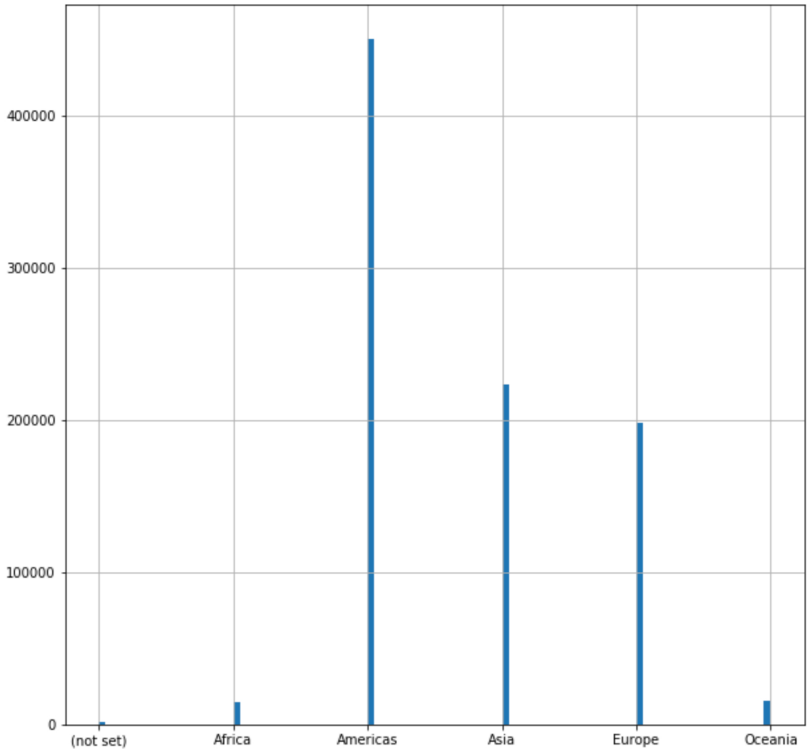
In [ ]:
```
'continent'
```

In [67]:
```
train_df['continent'].value_counts()
```

Out[67]:
```
Americas     450377
Asia         223698
Europe       198311
Oceania       15054
Africa        14745
(not set)      1468
Name: continent, dtype: int64
```

In [68]:
```
%matplotlib inline
train_df.continent.hist(bins=100, figsize=(10,10))
```

Out[68]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e44f3b70>
```



In [ ]:
```
'region'
```

In [69]:
```
train_df['region'].value_counts()
```

Out[69]:
```
not available in demo dataset    508229
California                       107495
(not set)                         27827
New York                          26433
England                           13198
Texas                              8749
Bangkok                            7709
Washington                         7642
Illinois                           7585
Ho Chi Minh                        7250
Istanbul                           6330
Maharashtra                        6184
Ontario                            5997
Taipei City                        5789
Karnataka                          5476
Delhi                              5278
```

```
Hanoi                                5141
New South Wales                      4932
Tamil Nadu                           4632
Ile-de-France                        4338
State of Sao Paulo                   4189
Telangana                            3955
County Dublin                        3650
Tel Aviv District                    3633
Tokyo                                3341
Jakarta                              3338
Masovian Voivodeship                 3288
Virginia                             3101
Community of Madrid                  2937
Victoria                             2741
                                    ...
Baja California                         6
Puebla                                  6
Erzurum                                 6
Central Java                            6
Manitoba                                6
Phu Tho Province                        6
Beirut Governorate                      6
Bremen                                  6
Ha Tinh Province                        6
Oran Province                           6
State of Espirito Santo                 6
Aydin Province                          6
Oklahoma                                6
Lopburi                                 6
Leiria District                         6
Abruzzo                                 6
Maha Sarakham                           6
Okinawa Prefecture                      6
Sonora                                  6
Assam                                   6
Queretaro                               6
Binh Phuoc                              6
Islamabad Capital Territory             6
Nord-Pas-de-Calais                      6
Overijssel                              6
Guanajuato                              6
Hradec Kralove Region                   6
San Salvador Department                 5
Kumamoto Prefecture                     5
Montana                                 3
Name: region, Length: 376, dtype: int64
```

In [71]:
```
%matplotlib inline
#?
```

In [ ]:
```
'networkDomain'
```

In [72]:
```
train_df['networkDomain'].value_counts()
```

Out[72]:
```
(not set)                244881
unknown.unknown          146034
comcast.net               28743
rr.com                    14827
verizon.net               13637
ttnet.com.tr              13228
comcastbusiness.net        9985
hinet.net                  7919
virginm.net                6414
3bb.co.th                  6046
prod-infinitum.com.mx      5960
cox.net                    5812
sbcglobal.net              5388
btcentralplus.com          5304
att.net                    5230
google.com                 5035
optonline.net              4972
totbb.net                  4895
vnpt.vn                    4508
asianet.co.th              4374
pldt.net                   4008
rima-tde.net               3963
amazonaws.com              3769
t-ipconnect.de             3656
telecomitalia.it           3571
qwest.net                  3534
airtelbroadband.in         3389
virtua.com.br              3318
bell.ca                    2904
ztomy.com                  2845
                          ...
costcom.ru                    1
orbuscompany.com              1
dupre-groenprojecten.nl       1
nielsenapms.com               1
```

```
vianovatelecom.com.br          1
feinewerkzeuge.de              1
hastwood.net                   1
alandick.ro                    1
rusneftekhim.ru                1
hi-p.com                       1
serverspace.co.uk              1
centuryshpg.com.hk             1
uol.com.sg                     1
thebiltmore.net                1
infowest.net                   1
hs-rm.de                       1
shamrockelectric.com           1
sertao.net                     1
dcs.co.zm                      1
dsbrown.com.sg                 1
stlghana.com                   1
ksnetworkbd.net                1
beingmate.com                  1
tamdistrict.org                1
state.mi.us                    1
getronics.com                  1
tecnicscorp.com                1
icm.edu.pl                     1
hatteland.com                  1
novotrucks.ro                  1
Name: networkDomain, Length: 28064, dtype: int64
```

In [ ]:

In [ ]: `'city'`

In [73]:
```
train_df['city'].value_counts()
```

Out[73]:
```
not available in demo dataset    508229
Mountain View                     40884
(not set)                         34262
New York                          26371
San Francisco                     20329
Sunnyvale                         13086
London                            12607
San Jose                          10295
Los Angeles                        8670
Bangkok                            7709
Chicago                            7444
Ho Chi Minh City                   7342
Istanbul                           6330
Bengaluru                          5468
Toronto                            5223
Hanoi                              5032
Seattle                            5025
Sydney                             4926
Dublin                             4877
Sao Paulo                          4106
Mumbai                             4099
Chennai                            4090
Paris                              4013
Hyderabad                          3934
Austin                             3790
Tel Aviv-Yafo                      3542
Hong Kong                          3508
Jakarta                            3338
Singapore                          3299
Warsaw                             3288
                                   ...
Netanya                               6
Thane                                 6
Beirut                                6
Islamabad                             6
Wollongong                            6
Annecy-le-Vieux                       6
Talence                               6
Sandy                                 6
Newark                                6
Vadodara                              6
Surat                                 6
Hradec Kralove                        6
Orem                                  6
Marlboro                              6
Amherst                               6
Vila Velha                            6
Kuwait City                           6
Tallahassee                           6
New Westminster                       6
Winnipeg                              6
Muntinlupa                            6
Clermont-Ferrand                      6
San Salvador                          5
Campbell                              5
```

```
Campbell                                3
Kumamoto                                5
Douglasville                            5
Deep River                              5
Daly City                               4
Bozeman                                 3
Boise                                   3
Name: city, Length: 649, dtype: int64
```

In [ ]:

In [ ]: `'metro'`

In [74]:
```
train_df['metro'].value_counts()
```

Out[74]:
```
not available in demo dataset                508229
(not set)                                    201766
San Francisco-Oakland-San Jose CA             95913
New York NY                                   26917
London                                        12571
Los Angeles CA                                 9995
Seattle-Tacoma WA                              7642
Chicago IL                                     7585
Austin TX                                      3790
Washington DC (Hagerstown MD)                  3380
Boston MA-Manchester NH                        2628
Houston TX                                     2475
Atlanta GA                                     2463
Detroit MI                                     2403
Roanoke-Lynchburg VA                           2227
Dallas-Ft. Worth TX                            2012
San Diego CA                                   1364
Portland OR                                    1319
Pittsburgh PA                                  1076
Denver CO                                       877
Philadelphia PA                                 800
Phoenix AZ                                      537
Charlotte NC                                    525
Columbus OH                                     517
La Crosse-Eau Claire WI                         396
San Antonio TX                                  393
Orlando-Daytona Beach-Melbourne FL              360
JP_KANTO                                        279
North West                                      271
Minneapolis-St. Paul MN                         197
                                               ...
Utica NY                                         23
North Scotland                                   21
Honolulu HI                                      21
HTV Wales                                        20
Albany-Schenectady-Troy NY                       19
Madison WI                                       17
Colorado Springs-Pueblo CO                       16
Greenville-Spartanburg-Asheville-Anderson        16
Wheeling WV-Steubenville OH                      14
HTV West                                         13
Charlottesville VA                               13
JP_OTHER                                         12
Abilene-Sweetwater TX                            10
Memphis TN                                       10
Mankato MN                                        9
Panama City FL                                    8
Tri-Cities TN-VA                                  8
Springfield MO                                    8
Lexington KY                                      8
Tucson (Sierra Vista) AZ                          7
New Orleans LA                                    7
Syracuse NY                                       7
Augusta GA                                        7
Springfield-Holyoke MA                            6
Tallahassee FL-Thomasville GA                     6
Rochester-Mason City-Austin,IA                    6
Chattanooga TN                                    6
Providence-New Bedford,MA                         6
Boise ID                                          3
Butte-Bozeman MT                                  3
Name: metro, Length: 94, dtype: int64
```

In [ ]:

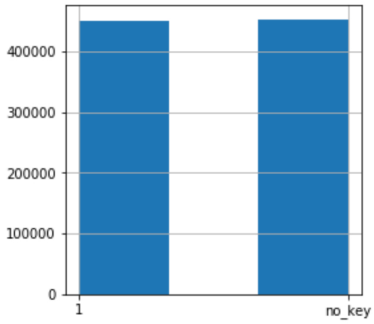
In [ ]: `'bounces'`

In [75]:
```
train_df['bounces'].value_counts()
```

Out[75]:
```
no_key    453023
1         450630
Name: bounces, dtype: int64
```

In [84]:
```
%matplotlib inline
train_df.bounces.hist(bins=3, figsize=(4,4))
```

Out[84]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ae701908>
```



In [ ]:
```
'transactionRevenue'
```

In [85]:
```
train_df['transactionRevenue'].value_counts()
```

Out[85]:
```
0.000000     892138
16.648135       256
16.759423       189
17.329739       187
17.617495       170
16.424845       135
17.840684       122
16.810743       116
16.587474        98
16.536148        93
16.769900        92
16.212496        84
17.909688        81
17.033986        77
18.197412        65
17.118360        64
17.117992        62
17.341283        54
17.504140        51
18.022886        46
14.503645        44
16.906098        40
17.175191        40
17.398474        39
17.229624        39
16.682841        38
17.229295        37
17.147001        37
16.379690        37
17.280934        37
                ...
16.886350         1
19.134415         1
21.799500         1
19.408063         1
18.221986         1
17.337745         1
17.756510         1
20.937611         1
17.130060         1
19.109368         1
18.827611         1
18.270207         1
17.577408         1
16.922337         1
19.833130         1
16.959663         1
19.327729         1
19.941069         1
17.902831         1
17.777656         1
16.921442         1
19.790973         1
18.336429         1
19.536445         1
17.932432         1
19.028597         1
17.194121         1
18.112524         1
20.434650         1
19.926157         1
Name: transactionRevenue, Length: 5333, dtype: int64
```

In [ ]:

In [ ]:
```
'hits'
```

In [86]:
```
train_df['hits'].value_counts()
```

Out[86]:
```
1      446754
2      137952
3       70402
4       42444
5       30939
6       23918
7       19518
8       15484
9       12959
10      10640
11       9264
12       7879
13       6881
14       6194
15       5384
16       4716
17       4130
18       3755
19       3291
20       3064
21       2798
22       2477
23       2190
24       2126
25       1925
26       1718
27       1504
28       1406
29       1291
30       1289
        ...
361         1
386         1
205         1
483         1
237         1
204         1
251         1
216         1
291         1
273         1
224         1
307         1
259         1
195         1
181         1
178         1
437         1
203         1
247         1
199         1
292         1
347         1
246         1
283         1
239         1
328         1
262         1
180         1
278         1
353         1
Name: hits, Length: 274, dtype: int64
```

In [ ]:
```

```

In [ ]:
```
'pageviews'
```

In [87]:
```
train_df['pageviews'].value_counts()
```

Out[87]:
```
1      452522
2      143770
3       73835
4       45192
5       33411
6       24688
7       19476
8       15272
```

```
9       12585
10      10104
11       8671
12       7097
13       6197
14       5291
15       4720
16       4010
17       3511
18       3150
19       2682
20       2409
21       2211
22       1949
23       1722
24       1567
25       1413
26       1318
27       1143
28       1091
29        875
30        873
        ...
358         1
431         1
162         1
136         1
466         1
340         1
249         1
155         1
164         1
174         1
323         1
469         1
333         1
208         1
197         1
429         1
309         1
144         1
154         1
400         1
190         1
195         1
182         1
151         1
275         1
351         1
220         1
186         1
141         1
305         1
Name: pageviews, Length: 214, dtype: int64
```

In [ ]:

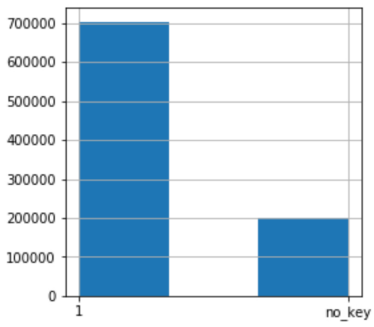
In [ ]:
```
'newVisits'
```

In [88]:
```
train_df['newVisits'].value_counts()
```

Out[88]:
```
1         703060
no_key    200593
Name: newVisits, dtype: int64
```

In [89]:
```
%matplotlib inline
train_df.newVisits.hist(bins=3, figsize=(4,4))
```

Out[89]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1ae84aa90>
```



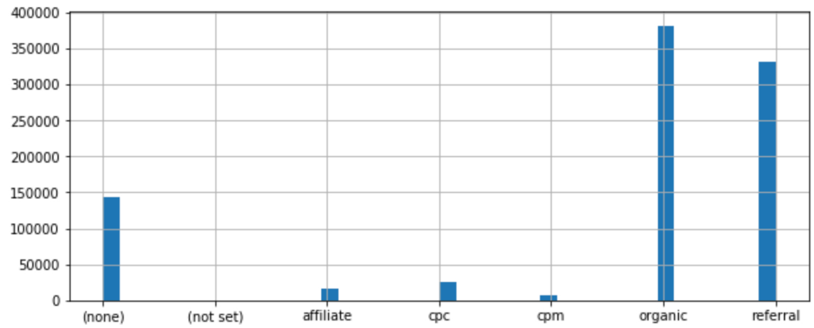
In [ ]:
```
'medium'
```

In [90]:
```
train_df['medium'].value_counts()
```

Out[90]:
```
organic      381561
referral     330955
(none)       143026
cpc           25326
affiliate     16403
cpm            6262
(not set)      120
Name: medium, dtype: int64
```

In [95]:
```
%matplotlib inline
train_df.medium.hist(bins=40, figsize=(10,4))
```

Out[95]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x147163ba8>
```



In [ ]:
```
'campaign'
```

In [96]:
```
train_df['campaign'].value_counts()
```

Out[96]:
```
(not set)                                            865347
Data Share Promo                                      16403
AW - Dynamic Search Ads Whole Site                    14244
AW - Accessories                                       7070
test-liyuhz                                             392
AW - Electronics                                         96
Retail (DO NOT EDIT owners nophakun and tianyu)          50
AW - Apparel                                             46
All Products                                              4
Data Share                                                1
Name: campaign, dtype: int64
```

In [102]:
```
%matplotlib inline
train_df.campaign.hist(bins=20, figsize=(40,1))
```

Out[102]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1d9064128>
```

In [ ]:
```
'referralPath'
```

In [103]:
```
train_df['referralPath'].value_counts()
```

```
                  1
/ads/richmedia/studio/pv2/47369756/20170126162720541/index.html

                  1
/mail/mu/mp/766/

                  1
/a/google.com/nestlabs/the-nest-pulse

                  1
/from=1019023b/bd_page_type=1/ssid=0/uid=0/pu=usm%403%2Csz%401320_1001%2Cta%40iphone_2_6.0_3_
537/baiduid=81B5E0170C3376788049DE4747F04F78/w=0_10_google+app+store/t=iphone/l=3/tc
                  1
/from=844b/bd_page_type=1/ssid=0/uid=0/pu=usm%400%2Csz%40320_1001%2Cta%40iphone_2_6.0_3_537/b
aiduid=FC6AC1F20A63A3F54B1225A33FC9B60F/w=0_10_/t=iphone/l=3/tc
                  1
/from=1017188c/bd_page_type=1/ssid=0/uid=0/pu=usm%400%2Csz%40320_1001%2Cta%40iphone_2_5.1_3_5
37/baiduid=E0CA0A7FD6C9064042D282172A4F3883/w=0_10_/t=iphone/l=3/tc
                  1
Name: referralPath, Length: 1476, dtype: int64
```

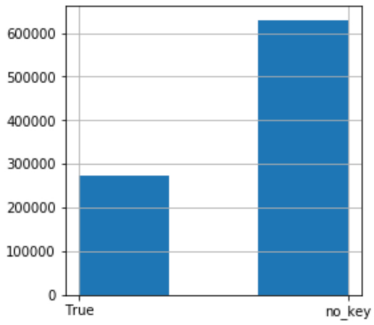
In [ ]:
```
'isTrueDirect'
```

In [104]:
```
train_df['isTrueDirect'].value_counts()
```

Out[104]:
```
no_key    629648
True      274005
Name: isTrueDirect, dtype: int64
```

In [105]:
```
%matplotlib inline
train_df.isTrueDirect.hist(bins=3, figsize=(4,4))
```

Out[105]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e26ec710>
```



In [ ]:
```
'keyword'
```

In [106]:
```
train_df['keyword'].value_counts()
```

```
                                1
youtube notebook buy online
                                1
google chrome shirt
                                1
Merchandise store
                                1
lava floor lamp
                                1
sports t shirts for kids
                                1
google pens
                                1
www.google.com mens tshirt
                                1
Mechandise
                                1
how to get youtube 5shirt
                                1
Name: keyword, Length: 3660, dtype: int64
```

In [ ]:
```
'adContent'
```

In [107]:
```
train_df['adContent'].value_counts()
```

Out[107]:
```
no_key                                          892707
Google Merchandise Collection                     5122
Google Online Store                               1245
Display Ad created 3/11/14                         967
Full auto ad IMAGE ONLY                            822
Ad from 12/13/16                                   610
Ad from 11/3/16                                    489
Display Ad created 3/11/15                         392
{KeyWord:Google Brand Items}                       251
{KeyWord:Google Merchandise}                       155
Ad from 11/7/16                                    123
First Full Auto Template Test Ad                    87
Google Merchandise                                  87
20% discount                                        75
{KeyWord:Google Branded Gear}                       67
{KeyWord:Looking for Google Bags?}                  65
Swag with Google Logos                              64
Display Ad created 11/17/14                         50
{KeyWord:Want Google Stickers?}                     42
JD_5a_v1                                            41
{KeyWord:Google Drinkware}                          32
{KeyWord:Google Men's T-Shirts}                     30
LeEco_1a                                            25
{KeyWord:Google Branded Kit}                        16
Full auto ad TEXT ONLY                              16
{KeyWord:Google Branded Apparel}                    10
Want Google Sunglasses                               8
Full auto ad TEXT/NATIVE                             7
Google Paraphernalia                                 7
{KeyWord:Google Branded Outerwear}                   5
{KeyWord:Want Google Pet Toys?}                      4
Google Store                                         4
Official Google Merchandise - Fast Shipping          4
free shipping                                        3
Full auto ad with Primary Color                      3
url_builder                                          3
Ad from 2/17/17                                      3
Full auto ad NATIVE ONLY                             3
Google store                                         2
Free Shipping!                                       2
```

```
google store                                        1
visit us again                                      1
Swag w/ Google Logos                                1
Men's-Outerwear Google Apparel                      1
GA Help Center                                      1
Name: adContent, dtype: int64
```

In [ ]:
```
```

In [ ]:
```
'source'
```

In [108]:
```
train_df['source'].value_counts()
```

Out[108]:
```
google                                          400788
youtube.com                                     212602
(direct)                                        143028
mall.googleplex.com                              66416
Partners                                         16411
analytics.google.com                             16172
dfa                                               5686
google.com                                        4669
m.facebook.com                                    3365
baidu                                             3356
sites.google.com                                  2983
facebook.com                                      2296
siliconvalley.about.com                           2097
reddit.com                                        2022
qiita.com                                         1813
quora.com                                         1546
bing                                              1530
t.co                                              1529
yahoo                                             1480
mail.google.com                                   1457
gdeals.googleplex.com                             1063
groups.google.com                                 1025
l.facebook.com                                     795
blog.golang.org                                    742
dealspotr.com                                      528
plus.google.com                                    524
moma.corp.google.com                               419
docs.google.com                                    388
productforums.google.com                           364
google.co.jp                                       356
                                                  ...
0.shared.bow.cat2.ads-bow.yw.borg.google.com:9895    1
0.shared.bow.cat2.ads-bow.lf.borg.google.com:9860    1
x20web.corp.google.com                               1
fr.yhs4.search.yahoo.com                             1
search.snapdo.com                                    1
lmgtfy.com                                           1
0.shared.bow.cat2.ads-bow.yw.borg.google.com:9898    1
adwords-prod-west.qa.adz.google.com                  1
google.com.ar                                        1
web.whatsapp.com                                     1
biztools.corp.google.com                             1
hosted.verticalresponse.com                          1
gsuite.google.com                                    1
good.barkpost.com                                    1
0.shared.bow.cat2.ads-bow.lf.borg.google.com:9817    1
s7-eu4.ixquick.com                                   1
allo.corp.google.com                                 1
meetup.com                                           1
google.se                                            1
hk.search.yahoo.com                                  1
google.sk                                            1
0.shared.bow.cat2.ads-bow.qk.borg.google.com:9831    1
google.lk                                            1
0.shared.bow.cat2.ads-bow.vw.borg.google.com:9891    1
collaborate.northwestern.edu                         1
us.wow.com                                           1
google.bg                                            1
0.shared.bow.cat2.ads-bow.lf.borg.google.com:9879    1
gdeals-stg.googleplex.com                            1
0.shared.bow.cat2.ads-bow.yw.borg.google.com:9839    1
Name: source, Length: 380, dtype: int64
```

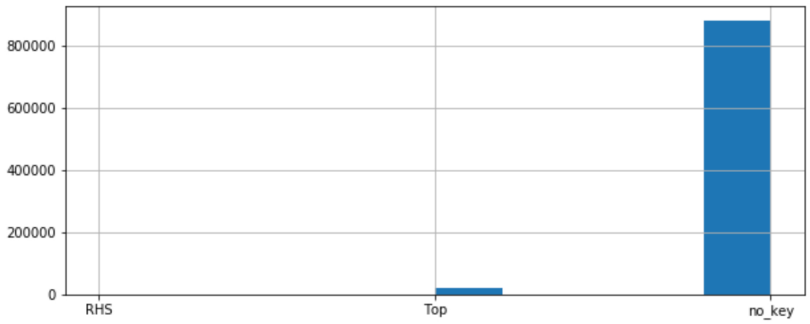
In [ ]:
```
'slot'
```

In [109]:
```
train_df['slot'].value_counts()
```

Out[109]:
```
no_key    882193
Top        20956
RHS          504
Name: slot, dtype: int64
```

In [113]:
```
%matplotlib inline
train_df.slot.hist(bins=10, figsize=(10,4))
```

Out[113]: <matplotlib.axes._subplots.AxesSubplot at 0x1e2da1860>



In [ ]:
```
'adNetworkType'
```

In [114]:
```
train_df['adNetworkType'].value_counts()
```

Out[114]:
```
no_key             882193
Google Search       21453
Search partners         7
Name: adNetworkType, dtype: int64
```

In [115]:
```
%matplotlib inline
train_df.adNetworkType.hist(bins=10, figsize=(10,4))
```

Out[115]: <matplotlib.axes._subplots.AxesSubplot at 0x1e2f946d8>

In [ ]:
```
'gclId'
```

In [116]:
```
train_df['gclId'].value_counts()
```

```
1
EAIaIQobChMIwarr5PTh1AIVz7bACh3TEgXgEAAYASAAEgKT0PD_BwE
1
CL6-7rabxNACFYEVgQodQa4K9g
1
CjwKEAjwoLfHBRD_jLW93remyAQSJABIygGpL0o8ujACMO_8DtJyWuvnj-mFEZDJdL0dG0j9NY7RxxoC3Bzw_wcB
1
CJHl_c_9iNECFUo9gQod9doLKg
1
Cj0KEQjw_9-9BRCqpZeZhLeOg68BEiQAOviWAgo5ufeoUwXlTTiTT1hryF4tjc-5Gky09iPKZODMKKgaAgov8P8HAQ
1
Cj0KEQiAkO7CBRDeqJ_ahuiPrtEBEiQAbYupJX1gU3VVQVhX7O1Hw5ONP0WppOESCTFbj6x6euvP5GgaAhND8P8HAQ
1
Cj0KEQjwyJi_BRDLusby7_S7z-IBEiQAwCVvn3dKgYE24s3C15u7NMqRdHaUdRos3jbas8QRThcZlogaArj98P8HAQ
1
CjwKCAjwzMbLBRBzEiwAfFz4gZKUfhmAo3wmL66EjqQ1UM5gkBZ8l4Q_juaHlBvCOMCbMIGeW6GxohoCeJ4QAvD_BwE
1
Cj0KCQjwkN3KBRCuARIsADT_flq0QKwopsitZuJ1SHPy6wuaZL7hps7UOjQHVsUDVVVbUlQfg0lDd3EaAqe7EALw_wcB
1
Cj0KEQiAw_DEBRChnYiQ_562qsEBEiQA4Lcssj9sLIjirTAntwVrpRzSxbdr7aCTR_y73752PKLWH-4aAh7B8P8HAQ
```

In [ ]:
```

```

In [ ]:
```
'page'
```

In [117]:
```
train_df['page'].value_counts()
```

Out[117]:
```
no_key    882193
1          21362
2             73
3             10
5              7
7              3
4              2
9              2
14             1
Name: page, dtype: int64
```

In [ ]:

In [ ]:
```
'targetingCriteria'
```

In [118]:
```
train_df['targetingCriteria'].value_counts()
```

Out[118]:
```
no_key        902193
empty_dict      1460
Name: targetingCriteria, dtype: int64
```

In [119]:
```
#!!!!
```

In [ ]:
```
'isVideoAd'
```

In [121]:
```
train_df['isVideoAd'].value_counts()
```

Out[121]:
```
no_key    882193
False      21460
Name: isVideoAd, dtype: int64
```

In [ ]:
```
#!!!!
```

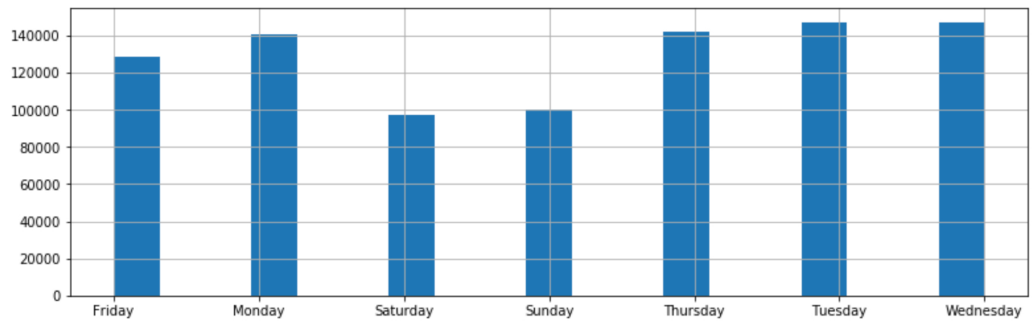
In [ ]:
```
'weekday'
```

In [122]:
```
train_df['weekday'].value_counts()
```

Out[122]:
```
Tuesday      147279
Wednesday    146733
Thursday     142399
Monday       140905
Friday       128331
Sunday       100360
Saturday      97646
Name: weekday, dtype: int64
```

In [126]:
```
%matplotlib inline
train_df.weekday.hist(bins=19, figsize=(13,4))
```

Out[126]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1aee7af28>
```



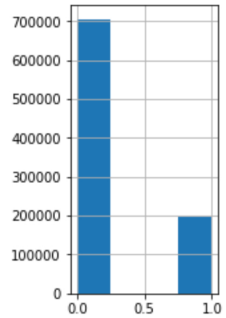
In [ ]:
```
'day_off'
```

In [127]:
```
train_df['day_off'].value_counts()
```

Out[127]:
```
0    705647
1    198006
Name: day_off, dtype: int64
```

In [132]:
```
%matplotlib inline
train_df.day_off.hist(bins=4, figsize=(2,4))
```

Out[132]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1b045acf8>
```



In [ ]: