In [1]:

```
import pandas as pd
```

In [16]:

```
store_df = pd.read_csv('optimized_gstore_data.csv', low_memory = False)
```

In [17]:

```
store_df.head()
```

Out[17]:

	channelGrouping	date	fullVisitorId	visitId	visitNumber	visitStartTime	device.browser	device.operatingSystem
0	Organic Search	20160902	1131660440785968503	1472830385	1	1472830385	Chrome	Windows
1	Organic Search	20160902	377306020877927890	1472880147	1	1472880147	Firefox	Macintosh
2	Organic Search	20160902	3895546263509774583	1472865386	1	1472865386	Chrome	Windows
3	Organic Search	20160902	4763447161404445595	1472881213	1	1472881213	UC Browser	Linux
4	Organic Search	20160902	27294437909732085	1472822600	2	1472822600	Chrome	Android

5 rows × 35 columns

In [18]:

```
type(store_df['date'][0])
```

Out[18]:

```
numpy.int64
```

In [19]:

```
store_df['date'] = store_df['date'].astype(str)
```

In [20]:

```
type(store_df['date'][0])
```

Out[20]:

```
str
```

In [21]:

```
# storing date in correct format

store_df['date'] = store_df['date'].apply(lambda x : x[:4] + '-' + x[4:6] + '-' + x[6:])
```

In [22]:

```
store_df['date']
```

Out[22]:

```
0         2016-09-02
1         2016-09-02
2         2016-09-02
3         2016-09-02
4         2016-09-02
             ...
903648    2017-01-04
```

```
903649    2017-01-04
903650    2017-01-04
903651    2017-01-04
903652    2017-01-04
Name: date, Length: 903653, dtype: object
```

In [23]:

```
store_df['date'] = pd.to_datetime(store_df['date'])
```

In [24]:

```
store_df.head()
```

Out[24]:

	channelGrouping	date	fullVisitorId	visitId	visitNumber	visitStartTime	device.browser	device.operatingSystem	de
0	Organic Search	2016-09-02	1131660440785968503	1472830385	1	1472830385	Chrome	Windows	
1	Organic Search	2016-09-02	377306020877927890	1472880147	1	1472880147	Firefox	Macintosh	
2	Organic Search	2016-09-02	3895546263509774583	1472865386	1	1472865386	Chrome	Windows	
3	Organic Search	2016-09-02	4763447161404445595	1472881213	1	1472881213	UC Browser	Linux	
4	Organic Search	2016-09-02	27294437909732085	1472822600	2	1472822600	Chrome	Android	

5 rows × 35 columns

In [25]:

```
store_df.isnull().sum()
```

Out[25]:

```
channelGrouping                      0
date                                 0
fullVisitorId                        0
visitId                              0
visitNumber                          0
visitStartTime                       0
device.browser                       0
device.operatingSystem               0
device.isMobile                      0
device.deviceCategory                0
geoNetwork.continent                 0
geoNetwork.subContinent              0
geoNetwork.country                   0
geoNetwork.region                    0
geoNetwork.metro                     0
geoNetwork.city                      0
geoNetwork.networkDomain             0
totals.hits                          0
totals.pageviews                   100
totals.bounces                  453023
totals.newVisits                200593
totals.transactionRevenue       892138
trafficSource.campaign               0
trafficSource.source                 0
trafficSource.medium                 0
trafficSource.keyword           502929
trafficSource.isTrueDirect      629648
trafficSource.referralPath      572712
trafficSource.adwordsClickInfo.page     882193
trafficSource.adwordsClickInfo.slot     882193
trafficSource.adwordsClickInfo.gclId    882092
trafficSource.adwordsClickInfo.adNetworkType    882193
trafficSource.adwordsClickInfo.isVideoAd        882193
trafficSource.adContent         892707
trafficSource.campaignCode      903652
```

```
dtype: int64
```

In [28]:

```
store_df['totals.pageviews'].value_counts()
```

Out[28]:

```
1.0      452522
2.0      143770
3.0       73835
4.0       45192
5.0       33411
          ...
309.0         1
162.0         1
197.0         1
327.0         1
249.0         1
Name: totals.pageviews, Length: 213, dtype: int64
```

In [29]:

```
store_df['totals.pageviews'].fillna(1,inplace=True)
```

In [30]:

```
store_df['totals.pageviews'] = store_df['totals.pageviews'].astype(int)
```

In [31]:

```
# now for totals.bounces

store_df['totals.bounces'].value_counts()
```

Out[31]:

```
1.0    450630
Name: totals.bounces, dtype: int64
```

In [33]:

```
store_df['totals.bounces'].fillna(0,inplace=True)
```

In [35]:

```
# The null values can be zero as the user hasn't made any purchases

store_df['totals.transactionRevenue'].fillna(0.0,inplace = True)
```

In [36]:

```
store_df['trafficSource.isTrueDirect'].value_counts()
```

Out[36]:

```
True    274005
Name: trafficSource.isTrueDirect, dtype: int64
```

In [37]:

```
# since this is directed revenue hence the nulls must be false

store_df['trafficSource.isTrueDirect'].fillna(False, inplace = True)
```

In [39]:

```
# filling all null values


store_df['totals.newVisits'] = store_df['totals.newVisits'].fillna(0).astype(int)
```

In [42]:

```
store_df['trafficSource.keyword'] = store_df['trafficSource.keyword'].fillna('(not provided)')
```

In [43]:

```
# The null values are 0 since ad clicks haven't been recorded

store_df['trafficSource.adwordsClickInfo.page'] = store_df['trafficSource.adwordsClickInfo.page'].f
illna(0).astype(int)
```

In [44]:

```
#The null values haven't been recorded since there weren't any ads

store_df['trafficSource.adwordsClickInfo.slot'] = store_df['trafficSource.adwordsClickInfo.slot'].f
illna('NoAds')
```

In [45]:

```
#  The null values haven't been recorded since there weren't any ads

store_df['trafficSource.adwordsClickInfo.adNetworkType'] =
store_df['trafficSource.adwordsClickInfo.adNetworkType'].fillna('NoAds')
```

In [46]:

```
# The null values haven't been recorded since there weren't any ads

store_df['trafficSource.adwordsClickInfo.isVideoAd'] =
store_df['trafficSource.adwordsClickInfo.isVideoAd'].fillna('NoAds')
```

In [47]:

```
# again checking the null value

store_df.isnull().sum()
```

Out[47]:

```
channelGrouping                         0
date                                    0
fullVisitorId                           0
visitId                                 0
visitNumber                             0
visitStartTime                          0
device.browser                          0
device.operatingSystem                  0
device.isMobile                         0
device.deviceCategory                   0
geoNetwork.continent                    0
geoNetwork.subContinent                 0
geoNetwork.country                      0
geoNetwork.region                       0
geoNetwork.metro                        0
geoNetwork.city                         0
geoNetwork.networkDomain                0
totals.hits                             0
totals.pageviews                        0
totals.bounces                          0
totals.newVisits                        0
totals.transactionRevenue               0
trafficSource.campaign                  0
trafficSource.source                    0
trafficSource.medium                    0
trafficSource.keyword                   0
```

```
trafficSource.isTrueDirect                          0
trafficSource.referralPath                     572712
trafficSource.adwordsClickInfo.page                 0
trafficSource.adwordsClickInfo.slot                 0
trafficSource.adwordsClickInfo.gclId           882092
trafficSource.adwordsClickInfo.adNetworkType        0
trafficSource.adwordsClickInfo.isVideoAd            0
trafficSource.adContent                        892707
trafficSource.campaignCode                     903652
dtype: int64
```

Dropping rest columns as they are very sparse.

In [49]:

```
store_df.drop(['trafficSource.referralPath','trafficSource.adwordsClickInfo.gclId','trafficSource.a
dContent','trafficSource.campaignCode'], axis=1,inplace = True)
```

In [50]:

```
# checking the null values again


store_df.isnull().sum()
```

Out[50]:

```
channelGrouping                                0
date                                           0
fullVisitorId                                  0
visitId                                        0
visitNumber                                    0
visitStartTime                                 0
device.browser                                 0
device.operatingSystem                         0
device.isMobile                                0
device.deviceCategory                          0
geoNetwork.continent                           0
geoNetwork.subContinent                        0
geoNetwork.country                             0
geoNetwork.region                              0
geoNetwork.metro                               0
geoNetwork.city                                0
geoNetwork.networkDomain                       0
totals.hits                                    0
totals.pageviews                               0
totals.bounces                                 0
totals.newVisits                               0
totals.transactionRevenue                      0
trafficSource.campaign                         0
trafficSource.source                           0
trafficSource.medium                           0
trafficSource.keyword                          0
trafficSource.isTrueDirect                     0
trafficSource.adwordsClickInfo.page            0
trafficSource.adwordsClickInfo.slot            0
trafficSource.adwordsClickInfo.adNetworkType   0
trafficSource.adwordsClickInfo.isVideoAd       0
dtype: int64
```

In [51]:

```
store_df.to_csv('cleaned_gstore_data.csv',header=True,index=False)
```

In [52]:

```
store_df.info(memory_usage = "deep")
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 903653 entries, 0 to 903652
Data columns (total 31 columns):
```

```
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   channelGrouping                         903653 non-null  object
 1   date                                    903653 non-null  datetime64[ns]
 2   fullVisitorId                           903653 non-null  object
 3   visitId                                 903653 non-null  int64
 4   visitNumber                             903653 non-null  int64
 5   visitStartTime                          903653 non-null  int64
 6   device.browser                          903653 non-null  object
 7   device.operatingSystem                  903653 non-null  object
 8   device.isMobile                         903653 non-null  bool
 9   device.deviceCategory                   903653 non-null  object
 10  geoNetwork.continent                    903653 non-null  object
 11  geoNetwork.subContinent                 903653 non-null  object
 12  geoNetwork.country                      903653 non-null  object
 13  geoNetwork.region                       903653 non-null  object
 14  geoNetwork.metro                        903653 non-null  object
 15  geoNetwork.city                         903653 non-null  object
 16  geoNetwork.networkDomain                903653 non-null  object
 17  totals.hits                             903653 non-null  int64
 18  totals.pageviews                        903653 non-null  int32
 19  totals.bounces                          903653 non-null  float64
 20  totals.newVisits                        903653 non-null  int32
 21  totals.transactionRevenue               903653 non-null  float64
 22  trafficSource.campaign                  903653 non-null  object
 23  trafficSource.source                    903653 non-null  object
 24  trafficSource.medium                    903653 non-null  object
 25  trafficSource.keyword                   903653 non-null  object
 26  trafficSource.isTrueDirect              903653 non-null  bool
 27  trafficSource.adwordsClickInfo.page     903653 non-null  int32
 28  trafficSource.adwordsClickInfo.slot     903653 non-null  object
 29  trafficSource.adwordsClickInfo.adNetworkType  903653 non-null  object
 30  trafficSource.adwordsClickInfo.isVideoAd      903653 non-null  object
dtypes: bool(2), datetime64[ns](1), float64(2), int32(3), int64(4), object(19)
memory usage: 1.1 GB
```

Hence in first notebook the data was optimized to 1.2 GB from 2.6 GB and now in this notebook it is further optimized to 1.1 GB. Hence we have cleaned and optimized the data provided.

In [ ]: