Customer Dataset

In [1]:

```
import pandas as pd
```

In [2]:

```
# Reading the csv file

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

In [3]:

```
store_df.head()
```

Out[3]:

	channelGrouping	date	fullVisitorId	sessionId	socialEngagementType	visitId	visitNumb
0	Organic Search	20160902	1131660440785968503	1131660440785968503_1472830385	Not Socially Engaged	1472830385	
1	Organic Search	20160902	377306020877927890	377306020877927890_1472880147	Not Socially Engaged	1472880147	
2	Organic Search	20160902	3895546263509774583	3895546263509774583_1472865386	Not Socially Engaged	1472865386	
3	Organic Search	20160902	4763447161404445595	4763447161404445595_1472881213	Not Socially Engaged	1472881213	
4	Organic Search	20160902	27294437909732085	27294437909732085_1472822600	Not Socially Engaged	1472822600	

5 rows × 55 columns

In [4]:

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 903653 entries, 0 to 903652
Data columns (total 55 columns):
 #   Column                                 Non-Null Count   Dtype
---  ------                                 --------------   -----
 0   channelGrouping                        903653 non-null  object
 1   date                                   903653 non-null  int64
 2   fullVisitorId                          903653 non-null  object
 3   sessionId                              903653 non-null  object
 4   socialEngagementType                   903653 non-null  object
 5   visitId                                903653 non-null  int64
 6   visitNumber                            903653 non-null  int64
 7   visitStartTime                         903653 non-null  int64
 8   device.browser                         903653 non-null  object
 9   device.browserVersion                  903653 non-null  object
 10  device.browserSize                     903653 non-null  object
 11  device.operatingSystem                 903653 non-null  object
 12  device.operatingSystemVersion          903653 non-null  object
 13  device.isMobile                        903653 non-null  bool
 14  device.mobileDeviceBranding            903653 non-null  object
 15  device.mobileDeviceModel               903653 non-null  object
 16  device.mobileInputSelector             903653 non-null  object
 17  device.mobileDeviceInfo                903653 non-null  object
 18  device.mobileDeviceMarketingName       903653 non-null  object
 19  device.flashVersion                    903653 non-null  object
 20  device.language                        903653 non-null  object
 21  device.screenColors                    903653 non-null  object
 22  device.screenResolution                903653 non-null  object
 23  device.deviceCategory                  903653 non-null  object
 24  geoNetwork.continent                   903653 non-null  object
```

```
 25  geoNetwork.subContinent                              903653 non-null  object
 26  geoNetwork.country                                   903653 non-null  object
 27  geoNetwork.region                                    903653 non-null  object
 28  geoNetwork.metro                                     903653 non-null  object
 29  geoNetwork.city                                      903653 non-null  object
 30  geoNetwork.cityId                                    903653 non-null  object
 31  geoNetwork.networkDomain                             903653 non-null  object
 32  geoNetwork.latitude                                  903653 non-null  object
 33  geoNetwork.longitude                                 903653 non-null  object
 34  geoNetwork.networkLocation                           903653 non-null  object
 35  totals.visits                                        903653 non-null  int64
 36  totals.hits                                          903653 non-null  int64
 37  totals.pageviews                                     903553 non-null  float64
 38  totals.bounces                                       450630 non-null  float64
 39  totals.newVisits                                     703060 non-null  float64
 40  totals.transactionRevenue                            11515 non-null   float64
 41  trafficSource.campaign                               903653 non-null  object
 42  trafficSource.source                                 903653 non-null  object
 43  trafficSource.medium                                 903653 non-null  object
 44  trafficSource.keyword                                400724 non-null  object
 45  trafficSource.adwordsClickInfo.criteriaParameters    903653 non-null  object
 46  trafficSource.isTrueDirect                           274005 non-null  object
 47  trafficSource.referralPath                           330941 non-null  object
 48  trafficSource.adwordsClickInfo.page                  21460 non-null   float64
 49  trafficSource.adwordsClickInfo.slot                  21460 non-null   object
 50  trafficSource.adwordsClickInfo.gclId                 21561 non-null   object
 51  trafficSource.adwordsClickInfo.adNetworkType         21460 non-null   object
 52  trafficSource.adwordsClickInfo.isVideoAd             21460 non-null   object
 53  trafficSource.adContent                              10946 non-null   object
 54  trafficSource.campaignCode                           1 non-null       object
dtypes: bool(1), float64(5), int64(6), object(43)
memory usage: 2.6 GB
```

In [5]:

```
store_df.nunique( dropna = False )
```

Out[5]:

```
channelGrouping                                    8
date                                             366
fullVisitorId                                 714167
sessionId                                     902755
socialEngagementType                               1
visitId                                       886303
visitNumber                                      384
visitStartTime                                887159
device.browser                                    54
device.browserVersion                              1
device.browserSize                                 1
device.operatingSystem                            20
device.operatingSystemVersion                      1
device.isMobile                                    2
device.mobileDeviceBranding                        1
device.mobileDeviceModel                           1
device.mobileInputSelector                         1
device.mobileDeviceInfo                            1
device.mobileDeviceMarketingName                   1
device.flashVersion                                1
device.language                                    1
device.screenColors                                1
device.screenResolution                            1
device.deviceCategory                              3
geoNetwork.continent                               6
geoNetwork.subContinent                           23
geoNetwork.country                               222
geoNetwork.region                                376
geoNetwork.metro                                  94
geoNetwork.city                                  649
geoNetwork.cityId                                  1
geoNetwork.networkDomain                       28064
geoNetwork.latitude                                1
geoNetwork.longitude                               1
geoNetwork.networkLocation                         1
totals.visits                                      1
totals.hits                                      274
```

```
totals.hits                                          274
totals.pageviews                                     214
totals.bounces                                         2
totals.newVisits                                       2
totals.transactionRevenue                           5333
trafficSource.campaign                                10
trafficSource.source                                 380
trafficSource.medium                                   7
trafficSource.keyword                               3660
trafficSource.adwordsClickInfo.criteriaParameters      1
trafficSource.isTrueDirect                             2
trafficSource.referralPath                          1476
trafficSource.adwordsClickInfo.page                    9
trafficSource.adwordsClickInfo.slot                    3
trafficSource.adwordsClickInfo.gclId                17775
trafficSource.adwordsClickInfo.adNetworkType           3
trafficSource.adwordsClickInfo.isVideoAd               2
trafficSource.adContent                               45
trafficSource.campaignCode                             2
dtype: int64
```

Since columns with single value throughout, doesn't helps in data classification. We drop such columns

In [6]:

```
single_value_col = []


for x in store_df.columns:
    if store_df[x].nunique(dropna = False) == 1:
        single_value_col.append(x)
```

In [7]:

```
single_value_col
```

Out[7]:

```
['socialEngagementType',
 'device.browserVersion',
 'device.browserSize',
 'device.operatingSystemVersion',
 'device.mobileDeviceBranding',
 'device.mobileDeviceModel',
 'device.mobileInputSelector',
 'device.mobileDeviceInfo',
 'device.mobileDeviceMarketingName',
 'device.flashVersion',
 'device.language',
 'device.screenColors',
 'device.screenResolution',
 'geoNetwork.cityId',
 'geoNetwork.latitude',
 'geoNetwork.longitude',
 'geoNetwork.networkLocation',
 'totals.visits',
 'trafficSource.adwordsClickInfo.criteriaParameters']
```

In [9]:

```
# dropping these columns

store_df.drop(single_value_col ,axis=1, inplace = True)
```

In [10]:

```
store_df.head()
```

Out[10]:

	channelGrouping	date	fullVisitorId	sessionId	visitId	visitNumber	visitStartTime	devic
0	Organic Search	20160902	1131660440785968503	1131660440785968503_1472830385	1472830385	1	1472830385	
1	Organic Search	20160902	377306020877927890	377306020877927890_1472880147	1472880147	1	1472880147	
2	Organic Search	20160902	3895546263509774583	3895546263509774583_1472865386	1472865386	1	1472865386	
3	Organic Search	20160902	4763447161404445595	4763447161404445595_1472881213	1472881213	1	1472881213	U
4	Organic Search	20160902	27294437909732085	27294437909732085_1472822600	1472822600	2	1472822600	

5 rows × 36 columns

Also the session is randomly generated hence it's of no use.

In [11]:

```
store_df.drop(['sessionId'],axis=1,inplace=True)
```

In [12]:

```
store_df.head()
```

Out[12]:

	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 [13]:

```
store_df.nunique(dropna = False)
```

Out[13]:

```
channelGrouping                  8
date                           366
fullVisitorId                714167
visitId                      886303
visitNumber                     384
visitStartTime               887159
device.browser                   54
device.operatingSystem           20
device.isMobile                   2
device.deviceCategory             3
geoNetwork.continent              6
geoNetwork.subContinent          23
geoNetwork.country              222
geoNetwork.region               376
geoNetwork.metro                 94
geoNetwork.city                 649
geoNetwork.networkDomain      28064
totals.hits                     274
totals.pageviews                214
totals.bounces                    2
totals.newVisits                  2
totals.transactionRevenue      5333
trafficSource.campaign           10
trafficSource.source            380
trafficSource.medium              7
trafficSource.keyword          3660
trafficSource.isTrueDirect        2
trafficSource.referralPath     1476
trafficSource.adwordsClickInfo.page          9
```

```
trafficSource.adwordsClickInfo.slot                  3
trafficSource.adwordsClickInfo.gclId             17775
trafficSource.adwordsClickInfo.adNetworkType         3
trafficSource.adwordsClickInfo.isVideoAd             2
trafficSource.adContent                             45
trafficSource.campaignCode                           2
dtype: int64
```

In [14]:

```
# Again checking the memory usage

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 903653 entries, 0 to 903652
Data columns (total 35 columns):
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   channelGrouping                           903653 non-null  object
 1   date                                      903653 non-null  int64
 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                          903553 non-null  float64
 19  totals.bounces                            450630 non-null  float64
 20  totals.newVisits                          703060 non-null  float64
 21  totals.transactionRevenue                 11515 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                     400724 non-null  object
 26  trafficSource.isTrueDirect                274005 non-null  object
 27  trafficSource.referralPath                330941 non-null  object
 28  trafficSource.adwordsClickInfo.page       21460 non-null   float64
 29  trafficSource.adwordsClickInfo.slot       21460 non-null   object
 30  trafficSource.adwordsClickInfo.gclId      21561 non-null   object
 31  trafficSource.adwordsClickInfo.adNetworkType  21460 non-null  object
 32  trafficSource.adwordsClickInfo.isVideoAd  21460 non-null   object
 33  trafficSource.adContent                   10946 non-null   object
 34  trafficSource.campaignCode                1 non-null       object
dtypes: bool(1), float64(5), int64(5), object(24)
memory usage: 1.2 GB
```

On comparing [Output 4] & [Output 5] we conclude that, we've made our data concise.

- comparing ooutput 4 & output 14

In [17]:

```
#Exporting the optimised csv file

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

In [ ]:

In [ ]: