

Training Data Set Creation

The training data set used for this analysis is shown below (first 5 lines):

```
In [33]: trainingDF = trainingDF[['totalAdClicks', 'totalGameClicks', 'totalPurchase']]
trainingDF.show(5)
trainingDF.count()
```

```
+-----+-----+-----+
|totalAdClicks|totalGameClicks|totalPurchase|
+-----+-----+-----+
|          19|          365|          63.0|
|          39|          683|          20.0|
|          37|          287|          28.0|
|          34|          794|          95.0|
|          41|          644|          53.0|
+-----+-----+-----+
only showing top 5 rows
```

```
Out[33]: 543
```

Dimensions of the training data set (rows x columns) : **543 x 3**

of clusters created: **2**