3.2 Training Data Set Creation

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

Create the final training dataset

Our training data set is almost ready. At this stage we can remove the 'userId' from each row, since 'userId' is a computer generated random number assigned to each user. It does not capture any behavioral aspect of a user. One way to drop the 'userId', is to select the other two columns.

```
In [37]: training_df = combined_df[['totalAdClicks', 'totalGameClicks','revenue']]
         training_df.head(5)
```

Out[37]:

	totalAdClicks	totalGameClicks	revenue
0	44	716	21.0
1	10	380	53.0
2	37	508	80.0
3	19	3107	11.0
4	46	704	215.0

Display the dimensions of the training dataset

Display the dimension of the training data set. To display the dimensions of the training_df, simply add , shape as a suffix and hit enter.

```
In [38]: training_df.shape
Out[38]: (543, 3)
```

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

# of clusters created: 3