



# Training Data Set Creation

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

 week3 Last Checkpoint: 26 minutes ago (unsaved changes) 

File Edit View Insert Cell Kernel Help Python 2

Code CellToolbar

```
Out[115]:
```

	totalAdClicks	revenue	hitsPerHour
0	44	21.0	0.219371
1	10	53.0	0.475000
2	37	80.0	0.164894
3	19	11.0	1.333333
4	46	215.0	0.221420

```
In [116]: trainingDF.shape
Out[116]: (529, 3)

In [118]: pDF = sqlContext.createDataFrame(trainingDF)

In [119]: parsedData = pDF.rdd.map(lambda line: array([line[0], line[1], line[2]])) #totalAdClicks, revenue
```

### Train KMeans model

```
In [123]: my_kmmodel = KMeans.train(parsedData, 3, maxIterations=10, runs=10, initializationMode="random")

In [124]: print(my_kmmodel.centers)
[array([ 26.93714286, 16.68, 0.46511615]), array([ 41., 142.10204082, 0.3021974 ]), array([ 34.35384615, 64.4, 0.39701866])]
```

Dimensions of the training data set (rows x columns) : (529,3)

# of clusters created: 3