In this assignment we will run the Kmean clustering in Spark, and make some observations for the following quiz.

Download and review the following script. This script creates some normally distributed random 2D data points, centered around 3 different coordinates, runs Kmeans, and prints out the sum-squared-error (SSE).

dokmeans.py

(note that the random data is generated with seed values set so it should be the same across runs)

1. Execute the kmeans script as follows:

```
>>> exec(open('dokmeans.py').read())
```

The script is set up to run Kmeans for k=1 clusters. Observe the SSE that is printed out by the script.

(Note: In fact if you look at the cluster centers, using my_kmmodel.clusterCenters, you get almost the same values used to create the data. You can use the cluster center coordinate points as a kind of summary of the data. In some cases the cluster centers could be labeled to serve as descriptions of the data.)

2. Change the number of clusters to k=4 and rerun the training command and get the SSE (or rerun the whole script).

Observe the approximate SSE for k=4.

3. Try getting summary statistics on the RDD of random data. Enter the following:

```
>>> my_data.stats()
```

Take note of standard deviation values. If you square each value you get the variance for each dimension.

Now compare these stats results to Kmeans SSE when k=3.

Observe the relationship between SSE when k=3 and the variance of each dimension. It's not obvious but write down the numbers for the quiz.