Assignment 4

Machine Learning
MSc Business Analytics

Xiaocheng Li

1 Individual Assignment

Instructions: *This exercise should be done "by hand", that is, not using Python built-in functions (you can use spreadsheets or your own Python scripts). All necessary calculations should be included in the submission, as well as brief explanations of what you do.*

Consider the following 7 two-dimensional observations:

	Observation i						
	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$
x_{i1}	1	1	1	5	2	6	4
x_{i2}	4	3	2	1	3	2	1

1. Plot the observations in a two-dimensional graph.

2. Perform K-means clustering with $K = 2$ using the Euclidean norm. Toss a coin 7 times to initialise the algorithm.

3. Cluster the data using hierarchical clustering with complete linkage and the Euclidean norm. Draw the resulting dendrogram.

2 Group Assignment

For this assignment, you are given two datasets of customer data. The first dataset, *customers.csv*, contains 200 samples and four features: gender, age, income and score. The second dataset, *customers_noisy.csv*, contains the same features plus four additional ones which have been corrupted by noise.
Perform the following steps:

1. Load the *customers.csv* dataset. Apply a Z-score normalisation on the numerical features, i.e. age, income and score.

2. Perform K-means with different k values, $k = 2, 3, .., 10$. In computing the distance, use only the normalised features from the previous point. Use a heuristic measure to find the best k.

3. Cluster the samples using K-means with the best k. Plot the clusters and centroids (in 3D with denormalised axes). Can you find any meaningful results? Can you identify customer segments?

4. Create three different datasets, each with two out of the three features you normalised, i.e. (age,income), (age, score), (income, score). Perform K-means and find the best k for each of them.

5. Plot the clusters and distinguish the data points based on the 'Gender' categorical feature. (You can use the visualisation method you prefer, e.g scatter plots with different markers, bar plots, etc.)

   Can you recognise customer subsegments that you did not identify earlier? Is it worth converting 'Gender' into a binary variable and including it in the K-means?

6. Load the *customers_noisy.csv* dataset. This dataset contains the four original features ('Gender' is now a binary feature) plus four noisy new ones. Perform hierarchical clustering on all features, plot the dendrogram and explain what you find.