1 Assignment 1

Instructions: *This exercise should be done "by hand", that is, not using Python. 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 Assignment 2

For this assignment, we are going to 'scrape' data on Welsh soccer players from the EA sports FIFA games website. It is important that you know how to scrape data from the web that is not available in a convenient form such as a CSV file.

1. Explore manually the website http://sofifa.com. Under the tab 'All', press on the any of the Welsh flags (e.g. the flag of G. Bale). Notice how the URL of the opened webpage changes to http://sofifa.com/players?na=50. Scrolling down, notice that not all players fit in one page. If you press 'Next', the new URL is http://sofifa.com/players?na=52&offset=60.

Can you see the pattern? Next, select an individual player and notice how the URL changes. We want to download the numerical attributes available for the first 660 Argentinian players (as appearing on the website).

2. Download the Python script `footballscrape.py`. This code worked for a previous version of the website, but recent changes means that it does not correctly scrape the website any longer. Fix the code, and explain the code as well as your fix. In order to better understand the code, you may want to look at the following websites:

   - https://www.crummy.com/software/BeautifulSoup/
   - http://www.aivosto.com/vbtips/regex.html
   - https://docs.python.org/2/library/re.html

3. How would you change the code to download the first 360 English players instead?

4. Use the `sklearn.cluster.KMeans` Python class to cluster the 660 Argentinian players from point 1 into 4 clusters.

5. By inspecting the clusters and looking up individual players online, try to assign meaningful labels to the clusters.

6. For a new and unknown player, the following attributes are available:

Crossing	45
Sprint Speed	40
Long Shots	35
Aggression	45
Marking	60
Finishing	40
GK_Handling	15

   For each of your 4 clusters from Step 4, compute the cluster centroid. Assign the new player to the nearest cluster based on the distance to the cluster centroids, using only the available attributes.