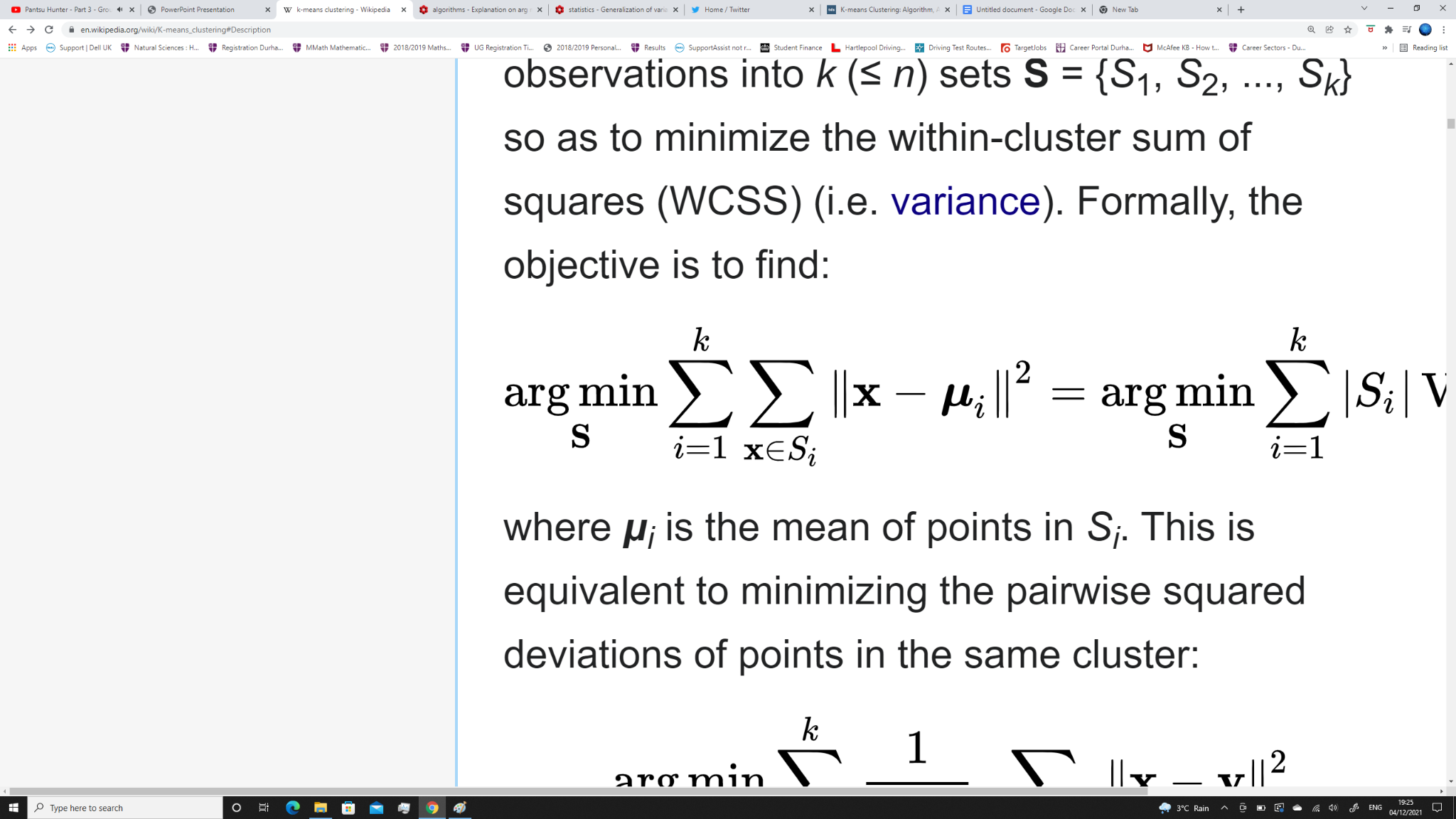
Week 6 HLT - Summary about k-means Machine Learning Algorithm

The k-means machine learning algorithm is a type of clustering algorithm that takes training data and partitions the data into ‘similar’ groups or clusters. This algorithm is an example of unsupervised machine learning because none of the initial training data have any labels assigned to the data. The clusters are determined purely by patterns in the data.

Mathematically speaking, ‘similar’ groups or clusters are defined as the set of k clusters labelled S1, S2, …, Sk that minimise the following double sum:



In this function, **μi** is the centre of the ith cluster and **x** is a data point that belongs to the ith cluster. This function is the sum of all of the squares of the (Euclidean) distance between the data point and the centre of the cluster it belongs to. The k-means algorithm seeks to find these clusters.

We can think of this function as the total deviation from a perfect set of clusters. For example, suppose we want to divide all the n observations in training data in which each data point has a value of x,y or z and we want to divide this data into 3 clusters. Since there are only 3 possible values (x,y,z) for the training data, we can get a perfect set of 3 clusters if we let the centres of the clusters be x,y and z. Substituting these centres into the above double sum gives a value of 0. This shows that a perfect set of clusters has the minimum possible value of the double sum (the argument in the summand is squared so it is always greater than or equal to 0). The larger the distance the data points are from the centre of the clusters, the larger the above sum will be.

There are many approaches to find these clusters Si, one approach is the naive k-means which starts with a good guess of what the centres of the clusters will be, and then uses an iterative process to update the centres. The algorithm stops when there is little change in the centres when updating.

The k-means algorithm is most useful when the clusters are separable and form a circular/spherical shape due to the use of the Euclidean metric to measure distance. This allows us to understand what is the best number of clusters to use when applying this algorithm.

An application of this algorithm is image compression. Each pixel in an image has three 8 bit RGB values. We can compress the image by selecting a small number of colours to only be included in the image (less colours for the majority of cases implies less data to store). The pixels are plotted in 3 dimensional space with their red, green and blue values as the axes and the number of colours is the number of clusters. K-means is applied here and the centres of the clusters are the colour that appear in the image (an ‘averaging’ of colours).