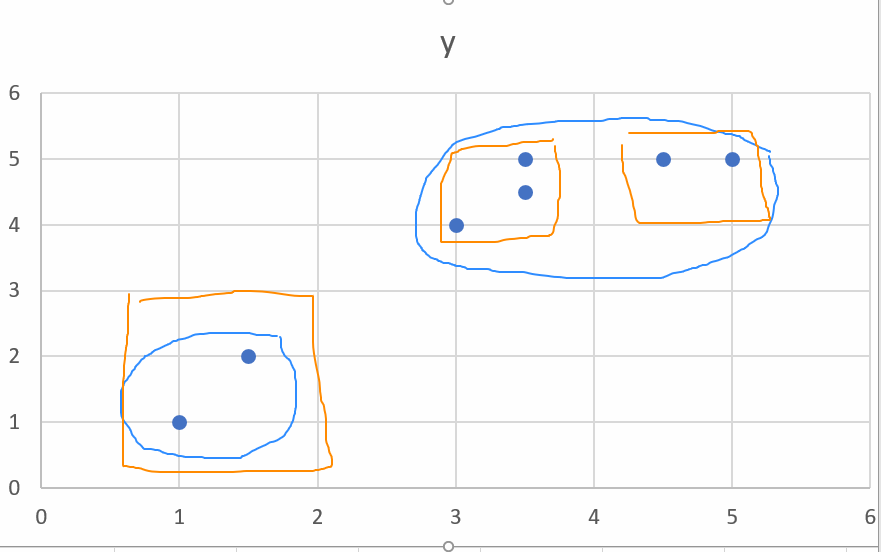
Clustering is unsupervised machine learning approach/model where the goal is to figure out how to group data. The outcomes from this step should include the following:

1. How many groups do we have?
2. Which observations are in which group?
   1. We tried to minimize the variance within a group and maximize the distances between observations in different groups



**The challenges:**

* We can group the same data in different ways (as shown in the example below). We can divide it into two groups (blue ellipses) or three (orange rectangles)
* Clustering leads to groups 🡪 try to figure out what are the commonalities within groups and what are the differences across the groups you have created.
  + **We do not have ground truth**
  + **The choice of clustering method, distance measure and variables will have an impact**
* We do not know how to weigh the different variables. So we typically assume that they have the same weight:
  + Rescale the variables (z-transformation [-3,3], (x-min)/range [0-1], ….)
  + The flip side is that some of these variables may not be important (essentially, you are assuming that all the variables are equally important)

**Questions:**

1. How do you measure distance?
   1. One approach is to take the Euclidean Distance = . **Why did I say one approach is to take the Euclidean Distance?**
      1. It would fail when one or more of the variables are non-numeric (i.e., categorical, binary)
      2. In 401/491, you only consider Euclidean Distances. With a lot of our cybersecurity data, this is not going to cut it (because we have categorical and binary variables)
2. How do we decide on the number of groups? (e.g. is it 2 or 3 for the picture above)
   1. Approach 1: You decide it in advance (kmeans) – go back and reconsider your choice of k
   2. Approach 2: Hierarchical Clustering – typically involves a dendogram
      1. Bottom up (each observation is in its own cluster and you will split it based on a distance calculation)
      2. Top down
   3. Approach 3**: We will vote (use many of the indices that are suggested in research and we will decide on the number of clusters based on a majority vote)**
      1. 26 indices – we will choose the number clusters based on what the majority of these indices indicated
3. What is the impact of the values for the variables we are looking at?