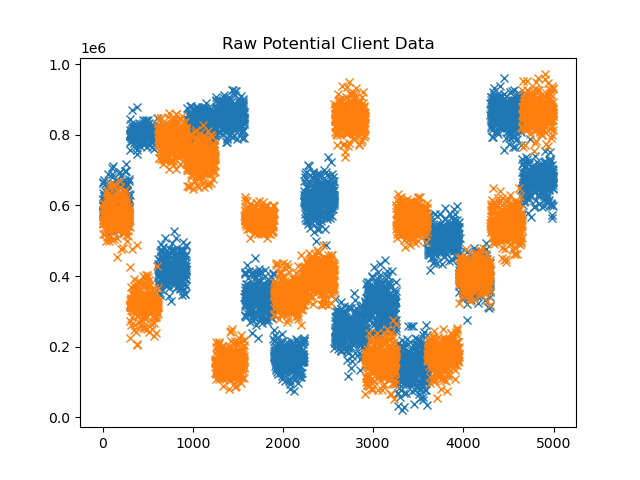
Clustering Exercise

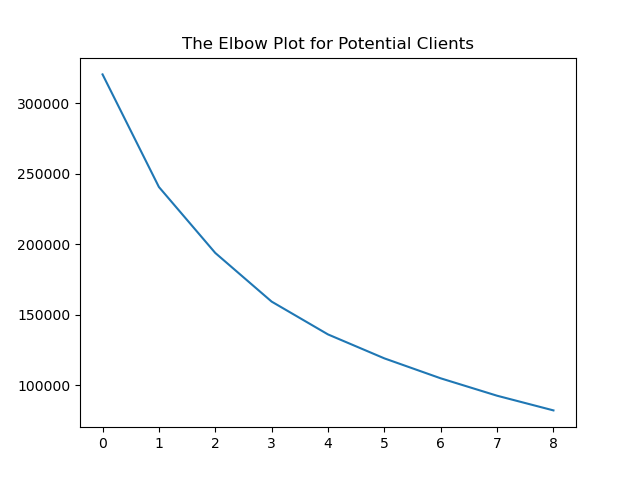
# Setup

In this problem, we are asked to take a collection of data points for potential investors. A variety of features have been derived into two scores that can describe each investor. Our job is to use k-means clustering in order to determine the ideal number of clusters for the data, and then map the data to those clusters. This is what our raw data looked like to start:



# Choosing k

Now that we had our raw data, we could determine how many clusters to use. To do this we will implement an elbow plot. An elbow plot is created by running the k-means algorithm multiple times at varying k values. We then plot the similarity scores for the clusters at each interval. The result allows us to get a feel for how good our clustering is being performed as we increase k. Eventually, there will be a point where the value begins to diminish, and an elbow can be seen in this graph. This ultimately is the best choice for k, as it offers us the best scores without overfitting the data. In our elbow plot, we have a rather curved line, indicating a gradual loss in diminishing returns. As a result, 4 clusters is likely our best choice, although a case could be made for 5, as the decrease in marginal return is still small. Here is the plot of our data:



# Clustering the Data

Now that we have determined the ideal number of clusters. We can simply run the algorithm a final time with our k value set at 4. We can then use the cluster designations generated to plot the data again, but this time color coded by cluster. Here are the results of our clustering exercise:

