Clustering

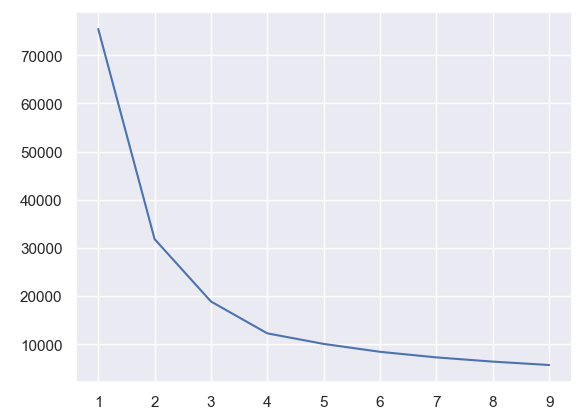
Challenge: Applying Clustering

Clustering is a procedure often used before creating a model and analyzing the properties of specific models. The point is to divide the data into specific groups based on their parameters, to simplify and clarify further research.

In this case, clustering is applicable to products.

Having information about the dimensions of products, we can divide them into K-categories.

To determine the number K, I used the so-called Elbow-Method (elbow method), based on the sum of squared distances to the centroid within one cluster. Having plotted these distances (wcss) at 1<=k<=10 I got:



As we can see, these distances stop changing much after the value of K=4.

This way I got the number of clusters I needed.

I carried out clustering according to two parameters: the estimated volume of the product and its mass.

It is important to take into account that due to the different numerical ranges of each characteristic, the data should be standardized so that the expected value is 0 and the variance is 1.

Result after clustering: