HW3 -Theoretical questions

1. a. K-medoid is more robust than K-means for two reasons:
   1. K-medoid, in contrast to K-means, chooses its medoid, only where there is a sample. Therefore, because of the rarity of the outlier, the medoid almost always will be one of the inliers. In K-means the mean will be influenced by the outlier of the cluster, such that the mean will be far from many inliers, and many dispersed points or points from another cluster will be associated with this mean.
   2. K-medoid is based on the L1 metric, which is less sensitive to noise than K-means metric - L2. The sensitivity can be expressed in the fact that in K-means, the outlier will attract the mean to the outlier more than the outlier in K-medoid will attract the median to the outlier.

   b. derivative and comparison to 0:

$$\frac{d}{d\mu}\sum_{i=1}^{m}(x_i-\mu)^2=0$$

$$-\sum_{i=1}^{m}2(x_i-\mu)=0$$

   Insulation of the centroid value:

$$\sum_{i=1}^{m}x_i=\mu m$$

$$\frac{\sum_{i=1}^{m}x_i}{m}=\mu$$

2. A. 1, B.5, C.3, D.2, E.6, F.4
   Notice, that the dataset is identical for all cases so all the difference is the kernel.
   A and D samples are divided by straight line, which represents linear equation. Meaning in both of them linear regression was in use. In D, the margins are wider than in A, so we can conclude that the regularization is stiffer than in A.
   C and F are the polynomial cases, and we can see that F can't be created with second order polyonomy.
   B and E are created with the BRF kernel, we can see that in case B there is more overfitting, which indicates a higher gamma value.
3. a. "as simple as possible" - The ability of generalization, "but not simpler" – high accuracy.

   b. p- the number of parameters. P needs to be high enough to describe the data (accuracy) but not too high to over-fit (generalization). We want to limit the number of parameters by penalty for big number of parameters. The limitation of parameters makes the model more generalized.

   L-the likelihood should be high but not too high. We want to be awarded on explaining the data well.

   c. p big and L small - complicated model (overfitting).

   L big and p small - low accuracy.

d. We want this term to decrease – have high accuracy with small number of parameters.