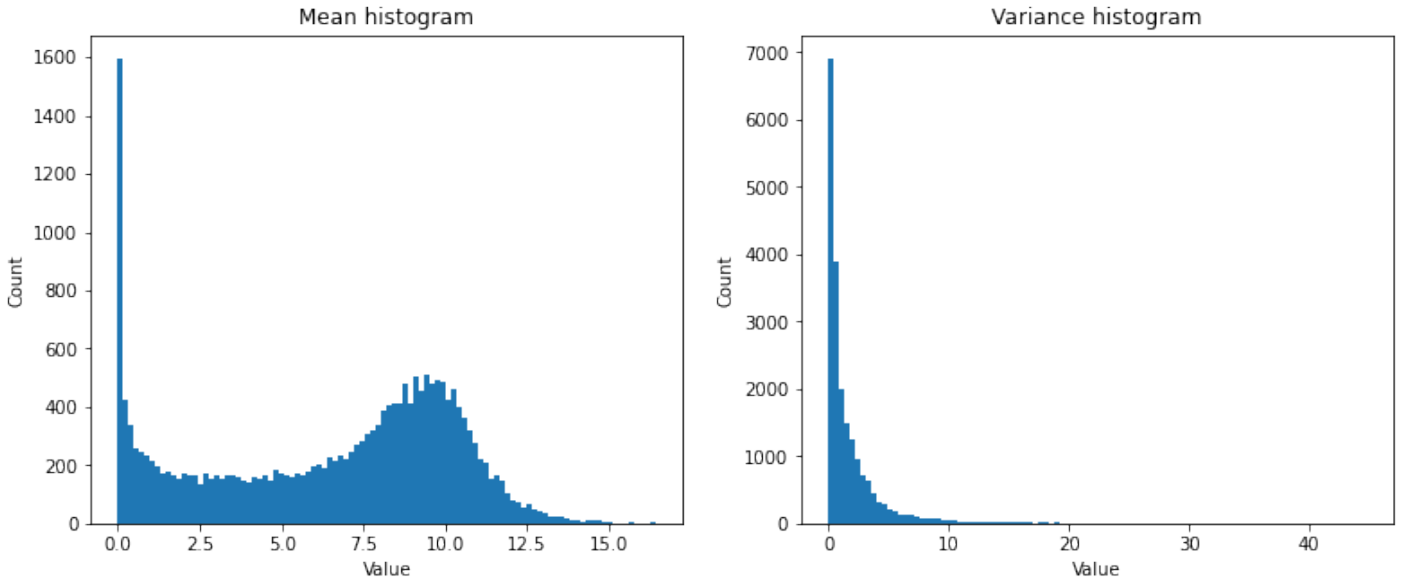
Project 1 - Classification

David Tonderski, Fredrik Meisingseth

23/4 - 2021

Task 1

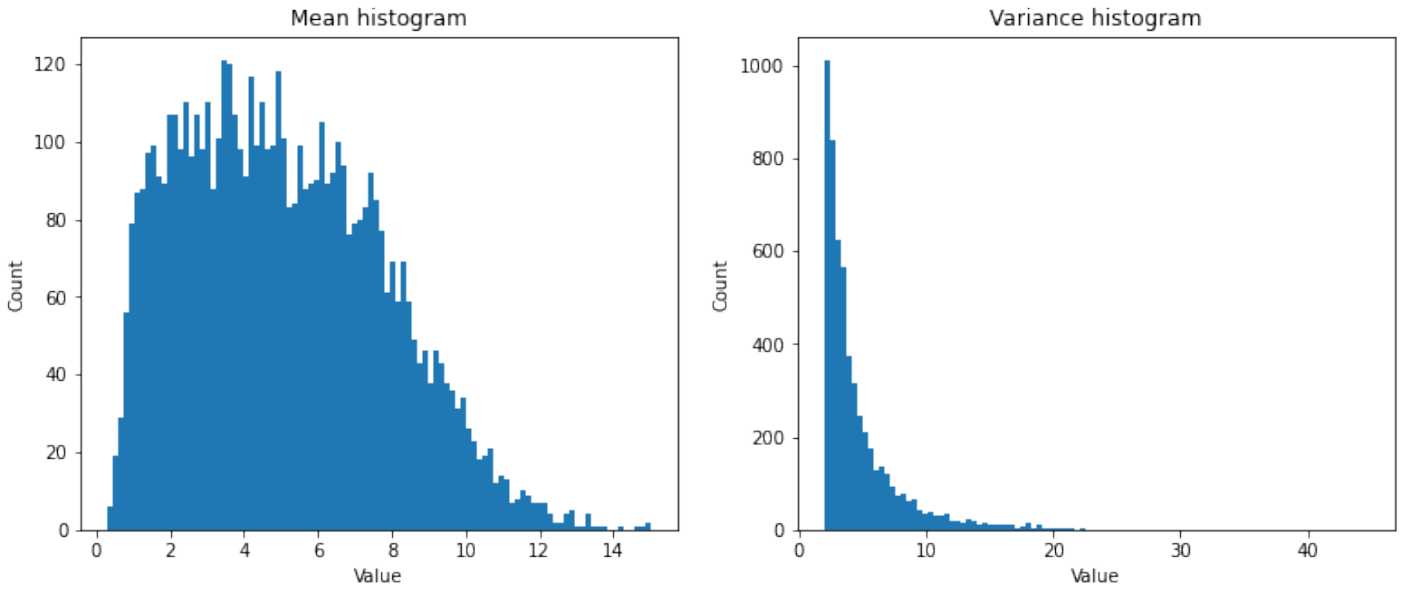
- The dataset contains no missing values. Histograms of the mean and the variance are shown below. Most variances are below 5, but some go up to over 40. The means vary from 0 to about 17.



- There are 267 constant features.

Task 2

- Variance filtering was performed by removing all features with a variance lower than 2, which leads to a reduced feature count of 5579. The new data is visualised in the histograms below.

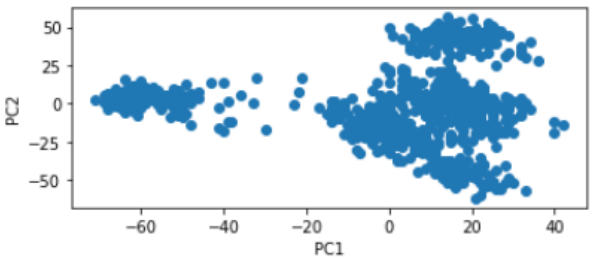


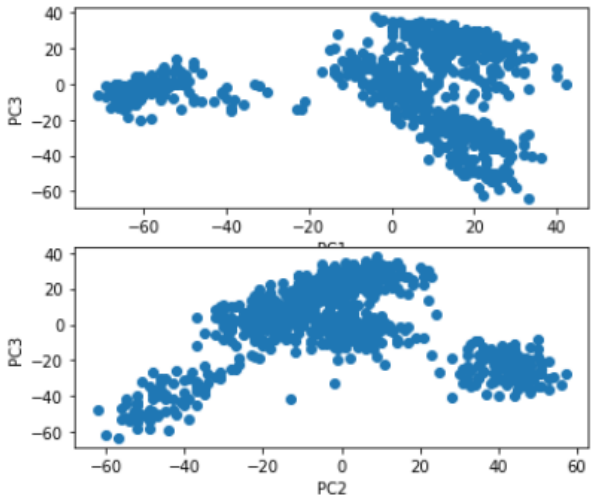
- Note that removing low variance data also removed low mean data, which might have unintended consequences. The variances and means still vary widely, so the data is centered and normalized.

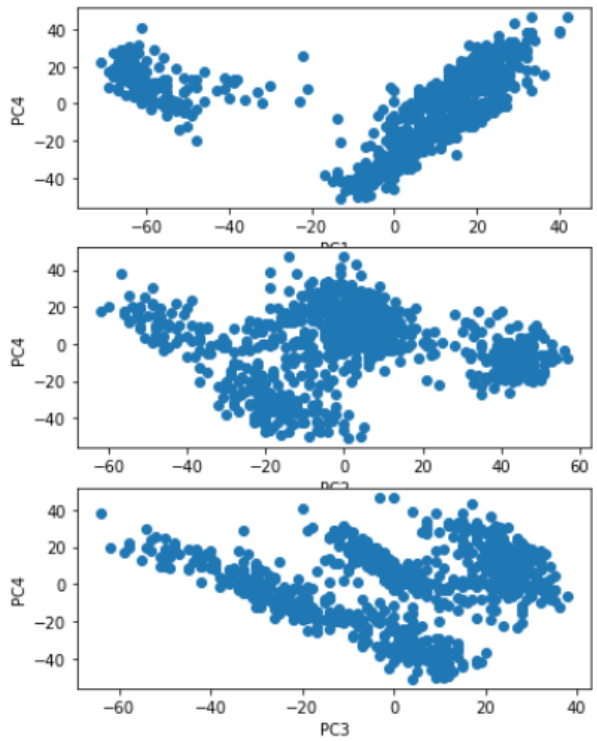
Task 3

- When plotting the first four PCs against each other (see next slide), there definitely are some patterns. Perhaps the clearest patterns, not surprisingly, may be seen in the plot of PC1 against PC2, where we can se that the data points are quite clearly divided in two according to PC1. The group with the higher PC1-values also seems to be divided into three or four groups according to their PC2.
- When first observing the scree plot of the eigenvalues of the covariance matrix, it is tempting to discard all but the first 20 or so PCs. This view is skewed by the magnitude of the first few eigenvalues, and if one ignores the first 100 or so PCs, one may see that there still occurs significant change in the eigenvalues until there is a sharp drop-off at the 800th one ($v_{799} \approx 0.34$, $v_{800} \approx 2 \cdot 10^{-14}$).
- Therefore we chose to perform a dimension reduction and only use the first 800 PCs.

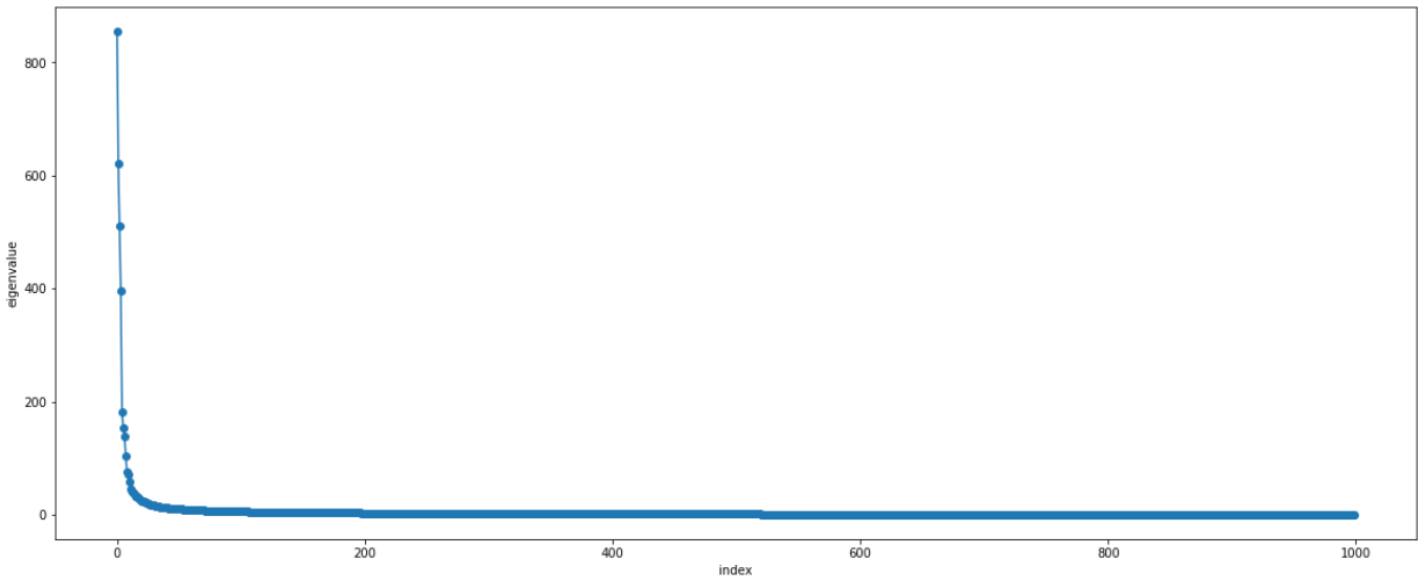
Task 3 - Pairwise PCA plots

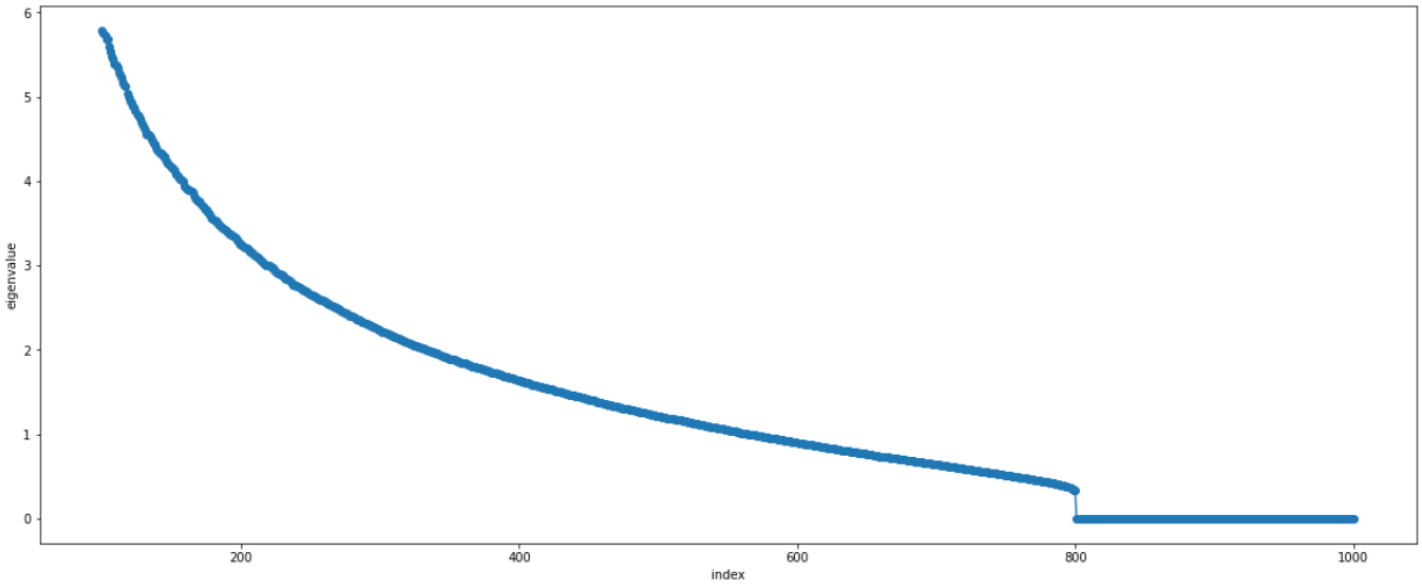






Task 3 - scree plots

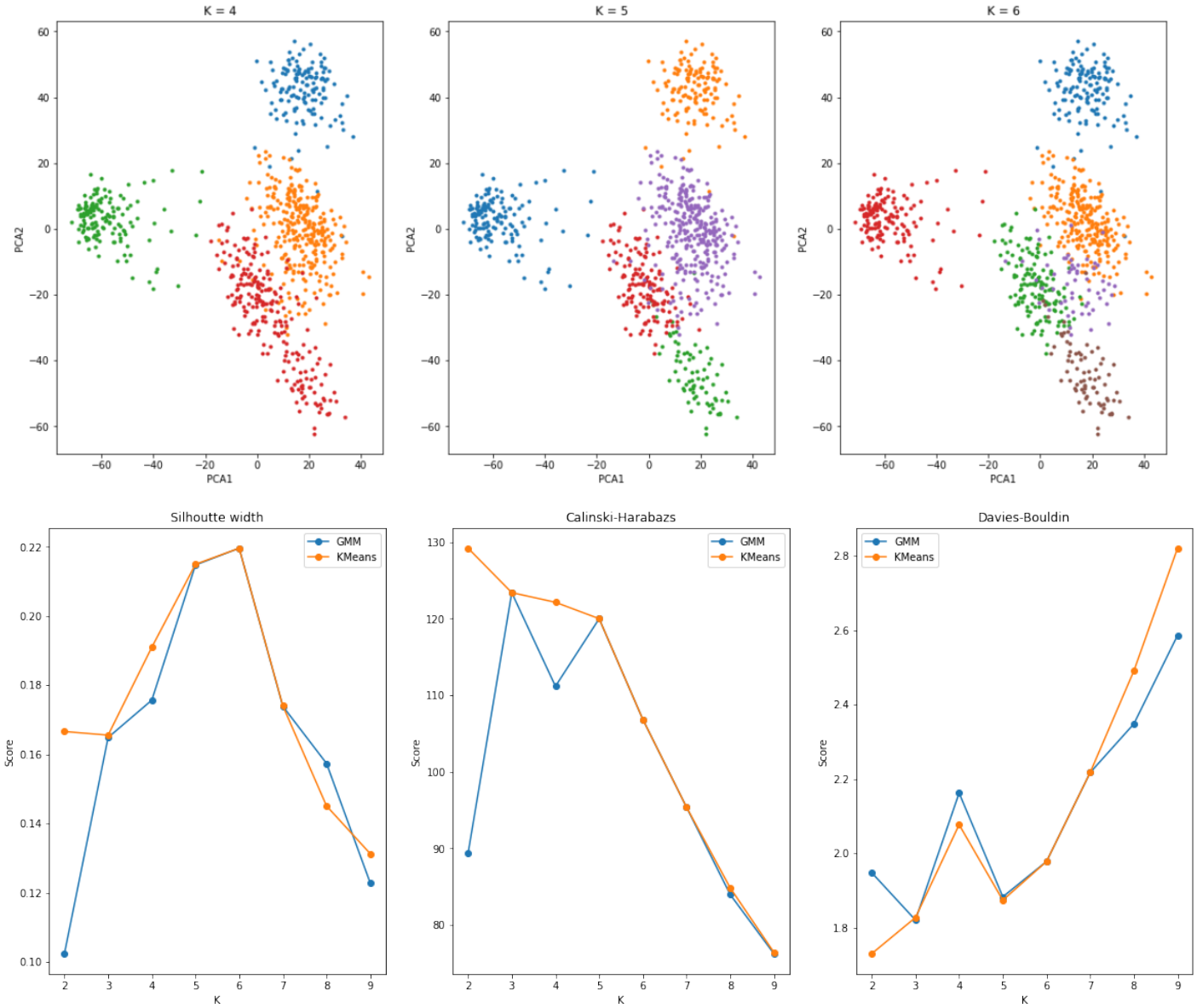




Task 4 - k-means

- The pair-plot of PCA1 and PC2 looks quite clear for K=4 and even for K = 5 but for K = 6 the patterns become significantly less clear. This suggests that the upper limit for K probably is not much above 6, and one might conjecture an upper limit of perhaps 8 clusters.
- The plot of the Silhuette widths suggests that K=5 or K = 6 are superior choices and then a quick loss of suitability as distance from those choices increase. The plot of the Calinski-Harabasz index scores suggests that a choice of $K \leq 5$ is suitable and the plot for the Davies-Bouldin index scores suggests that a choice of $K \leq 6$ is suitable, with the somewhat surprising result that K=4 seems significantly worse than K=3 or K=5.

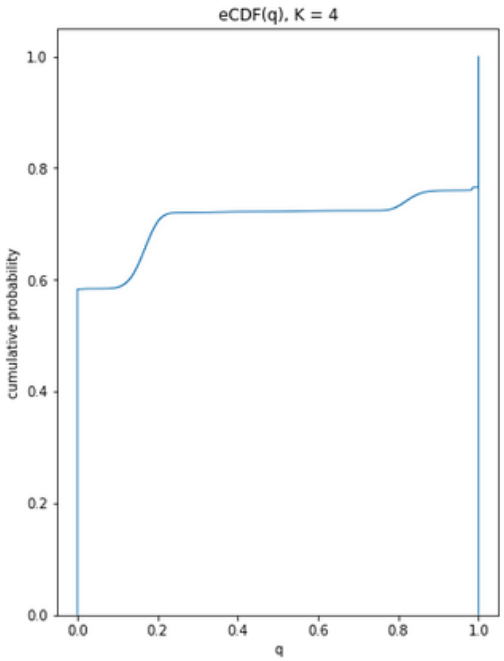
Task 4 - PCA pair cluster plots and internal index plots

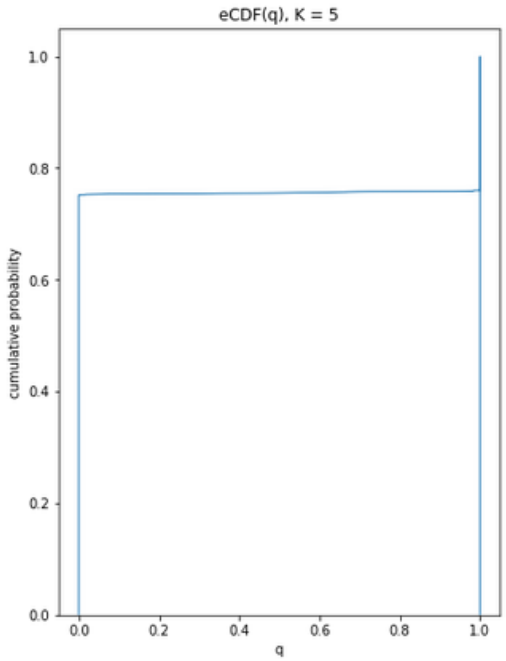


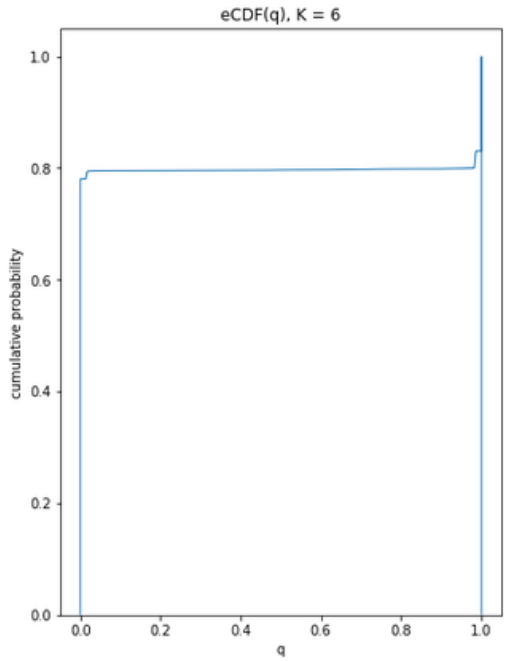
Task 5

- From the eCDF plots we see that, non-surprisingly, most entries in the consensus matrix are either 0 or 1 for all choices of K. We also see that there is a quite clear reduction of number of elements inbetween 0 and 1 for $K = 5$ compared to the choices around it, with the exception that $K = 2$ seems to give a practically optimal curve.
- The PAC curve reflects the observations of the eCDF curve, that $K = 5$ seems to be a suitable choice but $K = 2$ is optimal.
- **Discussion:** The results from the PAC curve corresponds roughly to the results from the internal indices, with the exception that $K = 2$ also seems like a suitable choice. All in all, without further knowledge of the exact context in which the clustering is to be used, we would argue for the cluster count choice of $K = 5$. This is because it scores amongst the best choices, if not the best, for all of our measures. Also, although dependent on the context of the data, in the choice of K=5 vs $K = 2$, we find it intuitively more sound to cluster to finely and then realize some clusters belong together rather than risking combining two clusters incorrectly.

Task 5 - eCDF plots







Task 5 - PAC plot