b. [12 Points] As discussed in lecture, principal components analysis (PCA) also has a squared-loss compression interpretation. Explain how these two methods relate to each other, and describe situations and hypothetical data sets in each of them might be appropriate or inappropriate.

PCA and k-means can both be used as feature compression techniques. PCA is used as a dimensionality reduction technique to produce a new set of orthonormal features, called principal components, which are the most salient features of the data. Under the squared-loss compression interpretation, PCA tries to make the projections as close as possible to the original data in as few dimensions as possible and minimize the reconstruction error when carrying out the transformation. Meanwhile k-means attempts to assign data to different clusters while updating the centroid of each cluster according to the data that belong to it, all to attempt to minimize its squared loss objective. If we were to replace each datum with its prototype, this would also be a sort of compression of the data set. However, this kind of compression is not of the dimensionality of the data, but rather of summarizing each individual data point by clumping it with the prototype point in the cluster to which it belongs. \\

K-means requires a parameter in its model, the number of clusters to find, and the choice for this parameter, which may be difficult to make, has a huge effect on the algorithm’s effectiveness. K-means is very effective when a small number of clusters is sufficient to represent the data while it is not very appropriate to use k-means when the number of clusters is large compared to the number of data points. PCA can run on any data set and will have good results on data that can be linearly reduced. PCA works especially well if the data can also be represented well in a small number of dimensions. It would not be appropriate to use PCA on a dataset that cannot be linearly reduced or a dataset that requires a large number of dimensions to represent well. PCA doesn’t do well where there is no linear reduction where the first few components capture enough variance or if there’s low correlation between the different dimensions of the data. \\

An example of a dataset where PCA might not do well but K-means will is on data drawn from a multivariate spherical, or also a sort of diagonal, Gaussian. PCA does not do well because it needs basically all of the dimensions in the data in order to represent the Gaussian well. A low dimensional PCA representation wouldn’t be a good representation of the data. K-means would do well because it can represent spherical things in low numbers of clusters (e.g. 1 cluster). \\

An example where k-means might not do well but PCA would, is data that form many clusters on a line. While PCA would do well because the data can essentially be captured very well in one dimension, k-means would not because it would require a large k to be able to be effective in clustering the large number of clusters on the line. In this case k-means doesn’t do well because it can’t reduce dimensionality and it can’t cluster well with small k in the original dimensions. PCA does well here because it works well as long as you can project the data linearly in a small number of dimensions.