Kunal Mehta
50352960
kunalkam

Quiz 4

1. Analysis of principal components that cause most of the total variation in the values of the variables of the dataset reveals outliers since they skew the original variables' variances and co-variances. Visualizing the principal component scores helps find multivariate outliers. Boxplots and scatterplots are two of the ways in which you can do so, as shown in the example in the PCA 4 video. Boxplots are generally easier to visualize since a large number of principal components can be analysed in relatively lesser space compared to scatter plots that require more screen space and are limited to two principal components at once. Once visualizing, there is simply the matter of identifying which observation/s the outliers are representing. In general, it is a good practice to check the first few and the last few principal components for outliers. The last few principal components account for negligible variation, therefore having an outlier here could reveal something interesting about the dataset.
2. Two characteristics of a dataset that would make principal component analysis an attractive option are:
    a. The dataset is all numerical.
    b. The dataset has a high dimensionality. Principal component analysis would reduce the dimensionality of the dataset, making the problem much easier to solve, since it is very difficult to interpret the dataset when you have a very high number of variables. Working with a smaller set of principal components, the ones that contribute the most to the total variances of the original variables is more feasible.
3. a. The variables Illiteracy and Life Exp contribute to the largest source of variation in the dataset. This is because the vectors representing these point in almost the exact opposite directions, indicating independence.

    b. California and New York's separation from the cloud of states is chiefly down to their PC2 scores. Both California and New York have significantly higher populations and incomes than the other states. From the given biplot, it is clear to see that the variable population, followed by the variable income are the most contributing ones to the PC2 score of a state being low.

    High values for the variables Murder and MS Grad (I can't see very clearly, the one under Life Exp) do contribute to lower PC2 scores as well, but not nearly as much as Population and Income do.