CS306 Data Analysis and Visualization

Lab. 6 Principal Component Analysis (PCA) and visualization of multivariate data.

In this practical you are given data file "New_York_Neighborhoods.xlsx". In the programming language of your choice read the file and get the numerical values of the features. The file contains 50 observations of 12 variables/features which are used to define neighbourhood of some suburbs in New York. Carry out the following data analysis and visualizations.

1. Compute, display and interpret the Pearson correlation matrix for the data.
2. Carry out PCA for the data. Python users may use sklearn implementation of PCA. Use n_components to be same as the number of variables in the data.
    a. Visualize the percentage variance explained by each principal component.
    b. Scatter plot each individuals/samples on the x & y axis as PC1 and PC2.
    c. Graph the variables as unit vector using their projection values on PC1 and PC2.
    d. Biplot both individuals/observations and the variables. Biplot is a combine scatter plots of the samples (b) and variable vectors (c).

3. Introduce following two outliers in the data.

70	70	700	80	83	71	600	70	65	900	45	800
77	600	72	82	800	73	65	900	62	75	-500	80

   Examine how the biplot changes. You can also experiment with scaling the variables and not scaling the variables.

You may use Python/R for this exercise