In this homework, you are guided step by step to perform the real example in Lecture 4.

1. Use Pandas to read the HW2.xlsx (provided) into Dataframes ‘equity’ and ‘factor’. Convert them into numpy arrays and construct the corresponding simple returns.

2. Set the following parameters: The required explanatory power ‘reqExp’ as 0.8; the required minimum correlation for the factor with the eigen portfolio ‘reqCorr’ as 0.4; the maximum allowed between-factor correlation ‘reqFcorr’ as 0.7.

3. Perform a PCA analysis on the factors using numpy linalg.eig, find the minimum number of principal components to cover the required explanatory power (0.8).

4. This is a good looping exercise: This is to find the most relevant factors to represent the principal components. You will need to set up some empty lists to store the factor names, factor correlation etc.
   The algorithm as follow:
   PC1: run each factor correlation with PC1 (pearsonr from scipy.stats). For the first factor, if the correlation (absolute) is greater than ‘reqCorr’, keep it. For the 2nd factor onward, the correlation needs to be greater than ‘reqCorr’ but less than the ‘reqFcorr’ to be kept.
   After PC1, you must have some factors in the list already, go on for PC2 and then PC3: For each factor, keep those with correlation greater than ‘reqCorr’ but less than the ‘reqFcorr’.

5. With the list of factors from Q4, normalize (standardize) their returns. Standardize the return for the equity indexes as well.

6. Run a for loop for each equity index over the standardized factors from Q4: the OLS (statsmodels.api) with intercept (‘add_constant’ function), retrieve the beta, t-value and the R-square and keep them into 3 different list. Output all of them into beta.csv, tvalue.csv, Rsq.csv.