Pizza Nutrition PCA & Multicollinearity Analysis

The data contained in the pizza.csv file which contains 300 records of information about nutritional info on different samples of pizza from 10 different brands labeled A-J. A brief description of each variable is found below. In total there will be two main parts to this assignment.

brand -- Pizza brand (class label)
id -- Sample analyzed
mois -- Amount of water per 100 grams in the sample
prot -- Amount of protein per 100 grams in the sample
fat -- Amount of fat per 100 grams in the sample
ash -- Amount of ash per 100 grams in the sample
sodium -- Amount of sodium per 100 grams in the sample
carb -- Amount of carbohydrates per 100 grams in the sample
cal -- Amount of calories per 100 grams in the sample

Part 1

a) Using what you have learned so far import the pizza.csv file into the SAS Work library so that you can use it in this assignment.

b) Using the variables mois, prot, fat, ash, sodium, carb, and cal run a precorrelation test and evaluate if any of these variables should be flagged for removal. Report your findings.

c) With the variables present in your model run a Principal Components Analysis and generate a corresponding scree plot. Using the Kaiser criterion identify how many components should be extracted from the data and state the individual eigenvalues of each, as well as the cumulative percent of variance explained by the components. Does the scree plot look as it should? Did the number of significant components SAS identify match your own?

d) Discuss the factor patterns and eigenvectors of each variable to the significant components, what do these values say about the variable's relationships to each factor? Using this information remove one variable from this model to use for part 2. Justify your decision.

Part 2

a) Using the variables in your reduced model run another Principal Components Analysis. With the Kaiser criterion in mind, check if the number of components that should be extracted from the data has changed and state the individual eigenvalues of each, as well as the cumulative percent of variance explained by the components.

b) Calculate the difference in communality estimates between the initial and final estimates for both the full and reduced model, based off this calculation and the prior information gathered in 2a do you believe that the reduced model is better than the full? Explain your answer.