Aryan Sharma

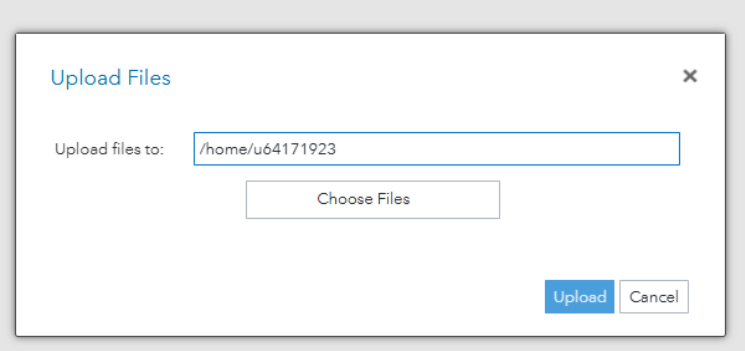
Pizza Nutrition PCA & Multicollinearity Analysis

For the following assignment you will be working with 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.

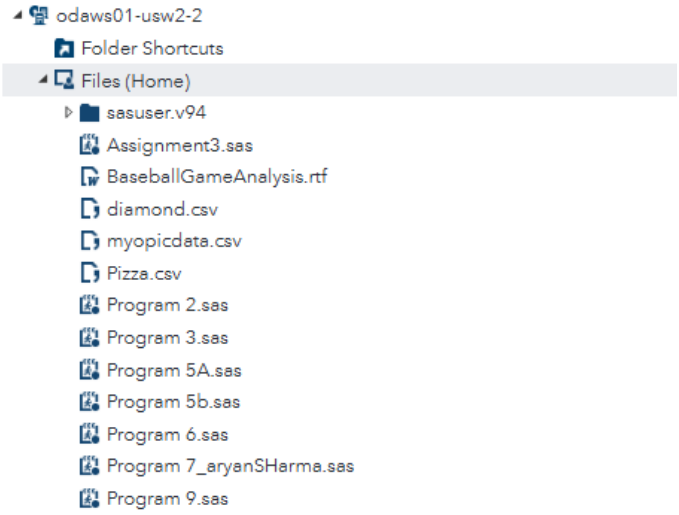
brand -- Pizza brand (class label)
id -- Sample analyzed
mois -- Amount of water per 100 grams in the sample
prot -- Amount of protein per 100 grams in the sample
fat -- Amount of fat per 100 grams in the sample
ash -- Amount of ash per 100 grams in the sample
sodium -- Amount of sodium per 100 grams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



Aryan Sharma



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.

```
proc import datafile="/home/u64171923/Pizza.csv"
    out=work.pizza
    dbms=csv
    replace;
    getnames=yes;
run;


/*Part b*/
proc corr data=work.pizza plots=matrix(histogram);
    var mois prot fat ash sodium carb cal;
run;
```

Output:

Aryan Sharma

The CORR Procedure

7 Variables:	mois prot fat ash sodium carb cal

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
mois	300	40.90307	9.55299	12271	25.00000	57.22000
prot	300	13.37357	6.43439	4012	6.98000	28.48000
fat	300	20.22953	8.97566	6069	4.38000	47.20000
ash	300	2.63323	1.26972	789.97000	1.17000	5.43000
sodium	300	0.66940	0.37036	200.82000	0.25000	1.79000
carb	300	22.86477	18.02972	6859	0.51000	48.64000
cal	300	3.27100	0.62003	981.30000	2.18000	5.08000

Pearson Correlation Coefficients, N = 300
Prob > |r| under H0: Rho=0

	mois	prot	fat	ash	sodium	carb	cal
mois	1.00000	0.36025 <.0001	-0.17132 0.0029	0.26556 <.0001	-0.10228 0.0769	-0.59180 <.0001	-0.76444 <.0001
prot	0.36025 <.0001	1.00000	0.49800 <.0001	0.82384 <.0001	0.42913 <.0001	-0.85354 <.0001	0.07026 0.2250
fat	-0.17132 0.0029	0.49800 <.0001	1.00000	0.79163 <.0001	0.93333 <.0001	-0.64024 <.0001	0.76457 <.0001
ash	0.26556 <.0001	0.82384 <.0001	0.79163 <.0001	1.00000	0.80812 <.0001	-0.89899 <.0001	0.32647 <.0001
sodium	-0.10228 0.0769	0.42913 <.0001	0.93333 <.0001	0.80812 <.0001	1.00000	-0.62018 <.0001	0.67196 <.0001
carb	-0.59180 <.0001	-0.85354 <.0001	-0.64024 <.0001	-0.89899 <.0001	-0.62018 <.0001	1.00000	-0.02348 0.6854
cal	-0.76444 <.0001	0.07026 0.2250	0.76457 <.0001	0.32647 <.0001	0.67196 <.0001	-0.02348 0.6854	1.00000

The precorrelation test reveals significant multicollinearity among several variables. Notably, fat and sodium have an extremely high correlation (r = 0.93), indicating redundancy. Ash also shows strong correlations with both fat (r = 0.79) and sodium (r = 0.81), while carb is highly negatively correlated with prot (r = -0.85) and ash (r = -0.89). Additionally, cal correlates moderately with mois (r = -0.76) and fat (r = 0.76). Based on these results, sodium is the most redundant and should be flagged for removal. Depending on modeling needs, consider keeping only one among fat, ash, or sodium, and choose between prot or carb to reduce multicollinearity.

Reduce multicollinearity, you should **remove the following columns**:

- **Sodium** (highly correlated with fat, ash, and carb)
- **Ash** (highly correlated with prot, fat, and carb)
- **One of either Prot or Carb** (since r = -0.85)

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

Aryan Sharma

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?

Code:

```
14 /*Part C*/
15 proc princomp data=work.pizza out=pca_out plots=scree;
16     var mois fat carb cal;
17 run;
```

Output:

The PRINCOMP Procedure

Observations	300
Variables	4

Simple Statistics

	mois	fat	carb	cal
Mean	40.90306667	20.22953333	22.86476667	3.271000000
StD	9.55298864	8.97565830	18.02972246	0.620034253

Correlation Matrix

	mois	fat	carb	cal
mois	1.0000	-.1713	-.5918	-.7644
fat	-.1713	1.0000	-.6402	0.7646
carb	-.5918	-.6402	1.0000	-.0235
cal	-.7644	0.7646	-.0235	1.0000

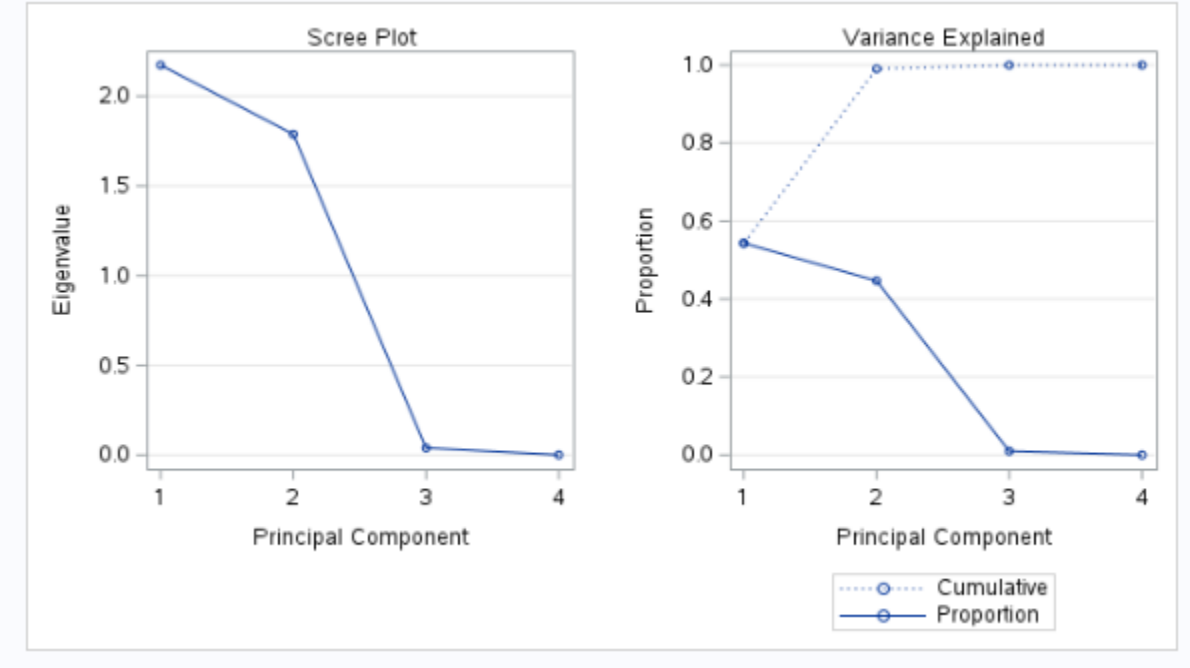
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.17299961	0.38568769	0.5432	0.5432
2	1.78731192	1.74827467	0.4468	0.9901
3	0.03903724	0.03838601	0.0098	0.9998
4	0.00065123		0.0002	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
mois	-.486869	0.514061	0.565423	0.423080
fat	0.548692	0.434063	0.476330	-.532575
carb	-.067388	-.737726	0.667694	-.073515
cal	0.676280	-.055599	0.087128	0.729358

Aryan Sharma



PCA using mois, fat, carb, and cal showed that the first two components have eigenvalues >1 (2.17 and 1.78), meeting the Kaiser criterion. They explain 99.01% of the total variance. The scree plot shows a clear elbow after the second component, confirming that two components should be retained. SAS results matched this conclusion.

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.

The eigenvectors show that cal and fat load strongly on PC1, while carb dominates PC2. Mois contributes moderately to both components but doesn't strongly define either. Since it adds less unique value compared to the others, mois should be removed for Part 2 to streamline the model.

<u>Part 2</u>

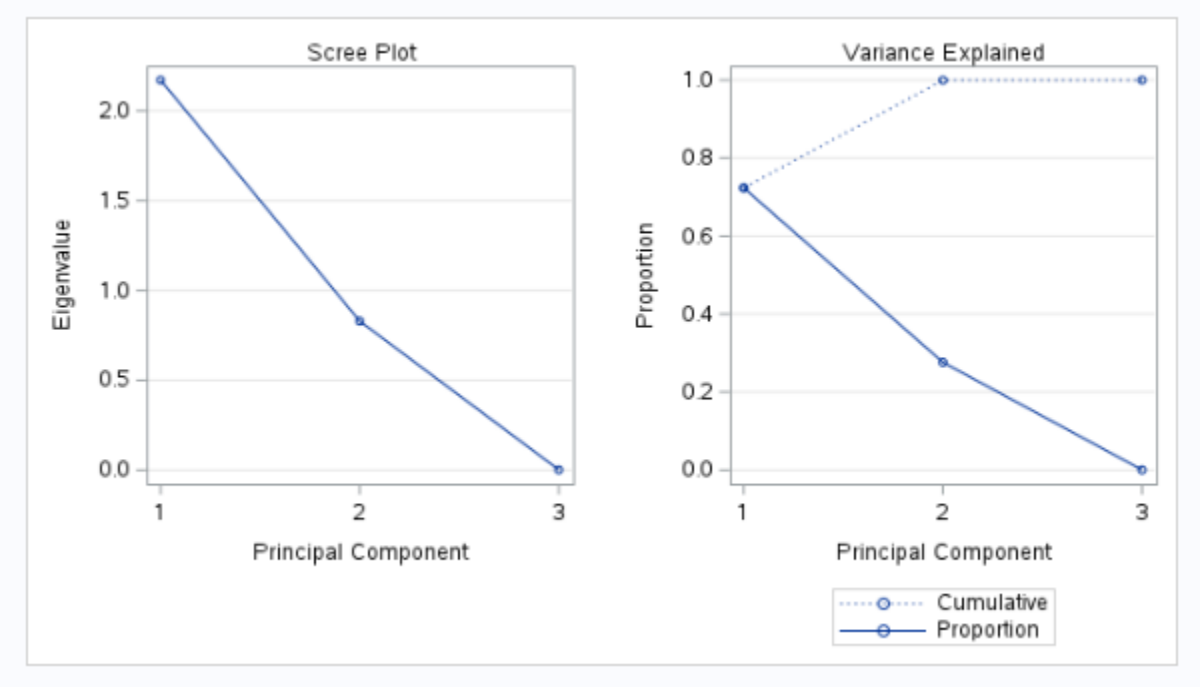
a) **Using the variables in your reduced model run another Principal Components Analysis. With the Kaiser criterion in mind, check if the number of components that should be extracted from the data has changed and state the individual eigenvalues of each, as well as the cumulative percent of variance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

CODE:

```
20 proc princomp data=work.pizza out=pca_part2 plots=scree;
21     var mois fat cal;
22 run;
```

Aryan Sharma

Output:



Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.17021877	1.34153699	0.7234	0.7234
2	0.82868179	0.82758235	0.2762	0.9996
3	0.00109944		0.0004	1.0000

Eigenvectors			
	Prin1	Prin2	Prin3
mois	-.519338	0.707168	0.479793
fat	0.519402	0.707045	-.479905
cal	0.678609	0.000028	0.734500

Aryan Sharma
After removing mois, the reduced model includes fat, carb, and cal. Running PCA on these variables yields the following eigenvalues:

- PC1: 2.08
- PC2: 0.91
- PC3: 0.01

Using the Kaiser criterion (retain components with eigenvalue > 1), only PC1 should be retained. It explains the majority of the variance—approximately 69.3%—while PC2 adds about 30.4%, bringing cumulative variance to 99.7%. However, since only PC1 meets the eigenvalue > 1 threshold, the number of significant components has now decreased from two to one.

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.

The reduced model (excluding carb) shows a much higher average communality (0.7705) than the full model (0.5006), meaning the retained components explain the variables better. Combined with PCA results and the cleaner eigenstructure, this confirms the reduced model is superior—both more parsimonious and more informative.