In this project Measurements were made on 24 prehistoric goblets from Thailand. I have been asked to help organize the goblets according to their similarities. It is believed that different cultures will likely produce pottery with different to very different characteristics. The scientist has measured the mouth width (X1), total width (X2), total height (X3), base width (X4), stem width (X5), and stem height (X6) on each of the 25 goblets.

PCA on Goblet Shape Characteristics

Obs	Obs	mouth_width	total_width	total_height	base_width	stem_width	stem_height
1	1	13	21	21	14	7	8
2	2	14	14	24	19	5	9
3	3	19	23	24	20	6	12
4	4	17	18	16	16	11	8
5	5	19	20	16	16	10	7

Principal Component Computation

The first step is to compute the eigenvalues of the Correlation Matrix

proc factor data=goblets method=prin priors=one;

var mouth_width total_width total_height base_width stem_width stem_height;

run;

The SAS System

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.29691408	3.26571341	0.7162	0.7162
2	1.03120067	0.63372816	0.1719	0.8880
3	0.39747251	0.24156253	0.0662	0.9543
4	0.15590998	0.08654261	0.0260	0.9802
5	0.06936737	0.02023197	0.0116	0.9918
6	0.04913540		0.0082	1.0000

The first eigenvalue, approximately 4.3, accounts for about 71.62% of the variance, indicating that Factor 1 captures most of the variability in the data. This makes Factor 1 quite strong. On the other hand, the second eigenvalue of 1.0 accounts for an additional 17.19% of the variance. Combined, Factors 1 and 2 account for a total of 88.80% of the variability in the data. The remaining eigenvalues are small, making them less relevant in explaining the variance in the data. The first two factors together suggest that they might be sufficient to describe the structure of the data. Therefore, two factors were selected.

The second step is to compute the Principal Component Analysis (PCA) to reduce dimensionalities and find hidden patterns in the dataset.

proc factor data=goblets method=prin priors=one n=2 rotate=varimax;

var mouth_width total_width total_height base_width stem_width stem_height;

run;

Rotated Factor Pattern		
	Factor1	Factor2
mouth_width	0.35981	0.83051
total_width	0.80632	0.46647
total_height	0.96483	0.07211
base_width	0.85033	0.44326
stem_width	0.14479	0.92225
stem_height	0.91557	0.27520

The Rotated Factor Pattern Table reveals that total_width (0.80), total_height (0.96), base_width (0.85), and stem_height (0.91) contribute significantly to Factor 1. We look for variables that have high loadings; typically, absolute values greater than 0.5 are considered significant. The rotation's outcome indicates that Factor 1 might represent the **overall size** or scale of the goblets. On the other hand, Factor 2 shows a strong connection with mouth_width (0.83) and stem_width (0.92), indicating that this factor could be related to the **shape of the goblets**.

Factor Analysis Computation

Next, we compute Factor Analysis (FA) with the same goal as the PCA.

proc factor data=goblets method=principal priors=smc ;

var mouth_width total_width total_height base_width stem_width stem_height;

run;

The SAS System

The FACTOR Procedure
Initial Factor Method: Principal Factors

Eigenvalues of the Reduced Correlation Matrix: Total = 5.0079243 Average = 0.83465405				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.15252448	3.35085769	0.8292	0.8292
2	0.80166679	0.60218320	0.1601	0.9893
3	0.19948358	0.18765620	0.0398	1.0291
4	0.01182738	0.08461683	0.0024	1.0315
5	-.07278945	0.01199904	-0.0145	1.0169
6	-.08478849		-0.0169	1.0000

The Eigenvalues of the Reduced Correlation Matrix Table reveal that the first factor is very strong; its eigenvalue of approximately 4.2 accounts for a significant proportion of the variance—82.92%. The second factor, with an eigenvalue of approximately 0.80, is much smaller, contributing less to the variance explanation. Factor 2 adds 16.01%, totaling 98.93% of the variance explained. The results suggest that Factor 1 and Factor 2 combined summarize much of the information contained in the six measurements (mouth_width, total_width, total_height, base_width, stem_width, stem_height) about the goblets.

```
proc factor data=goblets method=principal priors=smc n=2 rotate=varimax ;

var mouth_width total_width total_height base_width stem_width stem_height;

run;
```

Rotated Factor Pattern		
	Factor1	Factor2
mouth_width	0.33674	0.79061
total_width	0.76877	0.49945
total_height	0.94744	0.10568
base_width	0.81857	0.48509
stem_width	0.16446	0.80581
stem_height	0.88473	0.31936

The Rotated Factor Pattern Table of the Principal Factor reveals that total_width (0.77), total_height (0.95), base_width (0.82), and stem_height (0.88) have a strong relationship with Factor 1, similar to the Principal Component Analysis. As in PCA, we search for variables with high loadings, where typically absolute values greater than 0.5 are considered significant. As in the PCA case, mouth_width (0.79) and stem_width (0.81) are strongly linked to Factor 2. In summary, both the Principal Component and Principal Factor results indicate that Factor 1 is associated with the **overall size** or scale of the goblets, while Factor 2 relates to the **shape of the goblets**.

Comparing PCA's and FA's results

```
proc factor data=goblets method=prin n=2 out=scored_data (rename=(Factor1=Size Factor2=Shape));

var mouth_width total_width total_height base_width stem_width stem_height;

run;

proc print data=scored_data;

run;
```

The SAS System

Obs	Obs	mouth_width	total_width	total_height	base_width	stem_width	stem_height	Size	Shape
1	1	13	21	21	14	7	8	0.18140	0.01849
2	2	14	14	24	19	5	9	0.19193	-0.78301
3	3	19	23	24	20	6	12	1.26136	-0.08856
4	4	17	18	16	16	11	8	0.47549	2.25790
5	5	19	20	16	16	10	7	0.55695	2.33116

The Principal Component and Factor Analysis approaches reduce the number of variables by creating new ones that capture crucial information from the dataset. Specifically, they identify patterns that explain similarities and differences among the goblets. For example, our analysis reveals that the Factor Score for Goblet 5 is Factor1 (Size)=0.55695 and Factor2 (Shape)=2.33116. This indicates that Goblet 5 has considerable total width, total height, base width, and stem height, suggesting it likely has a relatively large base, is tall, and has a significant stem height based on its Factor1 score. Conversely, with Factor2 scores (Mouth Width: 0.49162, Stem Width: 0.68720), Goblet 5 has a significant but not overwhelming mouth width and a wider stem width.

Principal components and Factor Analysis simplify the task of identifying key characteristics that vary among the goblets. For instance, goblets scoring high on the size component from specific archaeological sites or regions may suggest that the cultures associated with these sites preferred taller pottery. These approaches effectively summarize and highlight the most significant physical features.

```
proc factor data=goblets method=principal priors=smc n=2 rotate=varimax score out=scored_data_FA (rename=(Factor1=Size Factor2=Shape));

var mouth_width total_width total_height base_width stem_width stem_height;

run;

proc print data=scored_data_FA;

run;
```

The SAS System

Obs	Obs	mouth_width	total_width	total_height	base_width	stem_width	stem_height	Size	Shape
1	1	13	21	21	14	7	8	0.06819	0.23769
2	2	14	14	24	19	5	9	0.65937	-0.55759
3	3	19	23	24	20	6	12	0.82406	1.08752
4	4	17	18	16	16	11	8	-0.73757	1.74630
5	5	19	20	16	16	10	7	-0.90224	2.16515

One key difference between Factor Analysis and Principal Component Analysis is the scores they assign to observations. For example, in the case of Goblet 5, FA scores Size as -0.902 and Shape as 2.165, indicating a smaller size but distinctive design or shape characteristics. While the pattern remains consistent in Factor 2, there is a notable difference in Factor 1, which can be attributed to the specific characteristics of FA. Factor Analysis is often considered an extension of PCA, aimed at understanding the underlying structure and identifying latent factors that explain observed patterns.

Challenging problem

Are there any goblets that are particularly unusual? Two goblets that are almost of the same shape are really similar but may have very different sizes.

We divided the goblet measurements by the total height of the body of the goblet to remove the effects of size as goblets may differ in shape rather than in size. Such an approach helps ensure that the data values are similar for two goblets with the same shape but with different sizes. We used PCA to answer the question.

```
data goblets_transformed;

   set goblets;

   total_measurements = mouth_width + total_width + total_height + base_width + stem_width + stem_height;

   mouth_width_ratio = mouth_width / total_height;
```

total_width_ratio = total_width / total_height;

total_height_ratio = total_height / total_height;

base_width_ratio = base_width / total_height;

stem_width_ratio = stem_width / total_height;

stem_height_ratio = stem_height / total_height;

Obs	mouth_width	total_width	total_height	base_width	stem_width	stem_height	total_measurements	mouth_width_ratio	total_width_ratio	total_height_ratio	base_width_ratio	stem_width_ratio	stem_height_ratio
1	13	21	21	14	7	8	84	0.61905	1.00000	1	0.66667	0.33333	0.38095
2	14	14	24	19	5	9	85	0.58333	0.58333	1	0.79167	0.20833	0.37500
3	19	23	24	20	6	12	104	0.79167	0.95833	1	0.83333	0.25000	0.50000
4	17	18	16	16	11	8	86	1.06250	1.12500	1	1.00000	0.68750	0.50000
5	19	20	16	16	10	7	88	1.18750	1.25000	1	1.00000	0.62500	0.43750

proc factor data=goblets_transformed method=prin n=2 rotate=varimax out=scored_pca;

var mouth_width_ratio total_width_ratio base_width_ratio stem_width_ratio stem_height_ratio;

title "PCA on Goblet Shape Characteristics";

run;

PCA on Goblet Shape Characteristics

The FACTOR Procedure
Rotation Method: Varimax

Rotated Factor Pattern		
	Factor1	Factor2
mouth_width_ratio	0.92572	0.03905
total_width_ratio	0.92406	0.07645
base_width_ratio	0.52004	0.79649
stem_width_ratio	0.81333	0.35344
stem_height_ratio	-0.03949	0.95304

Since the size effects were removed, the PCA of the new variables shows that Factor 1 captures the width elements of the goblet's shape relative to height: mouth_width_ratio (0.93), total_width_ratio (0.92), base_width_ratio (0.52), and stem_width_ratio (0.81). Thus, Factor 1 represents the **goblet's width profile**. In contrast, Factor 2 focuses on the vertical dimensions, such as stem_height_ratio (0.95) and base_width_ratio (0.80), relative to the goblet's overall height, making it the **goblet's vertical profile** factor.

```
proc univariate data=scored_pca;
   var Factor1;
         run;
proc univariate data=scored_pca;
   var Factor2;
         run;
```

PCA on Goblet Shape Characteristics

The UNIVARIATE Procedure
Variable: Factor1

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-1.282495	2	0.763332	18
-1.188658	11	1.367553	17
-1.157603	20	1.461101	23
-0.996961	21	1.725937	4
-0.987671	19	2.284661	5

PCA on Goblet Shape Characteristics

The UNIVARIATE Procedure
Variable: Factor2

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-2.737775	24	0.956184	20
-1.296293	23	1.050184	3
-1.009822	8	1.172582	16
-0.832976	9	1.284316	17
-0.786135	22	1.486461	4

The table for Factor 1 shows that Goblet 2 has an extremely low Factor1 score of -2.81181, which is significantly lower than most other scores. This score suggests that the characteristics of Goblet 2, particularly in terms of width size ratio, make it unusual compared to other goblets. Regarding Factor 2, Goblet 4, with a score of 1.486461, exhibits a distinctive stem height and base width ratio, distinguishing it from other goblets. In conclusion, PCA can be used to identify unusual goblets.