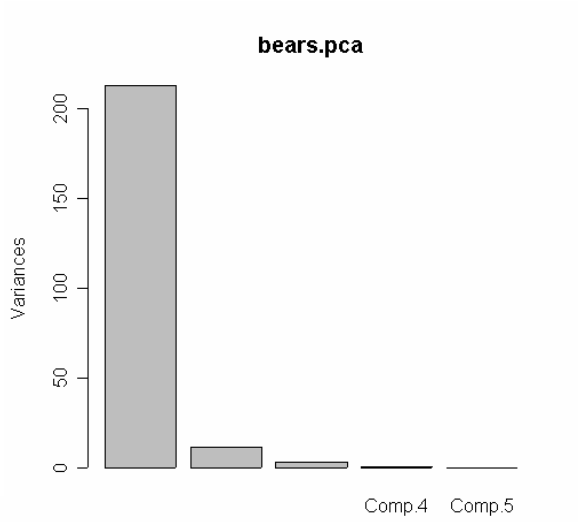
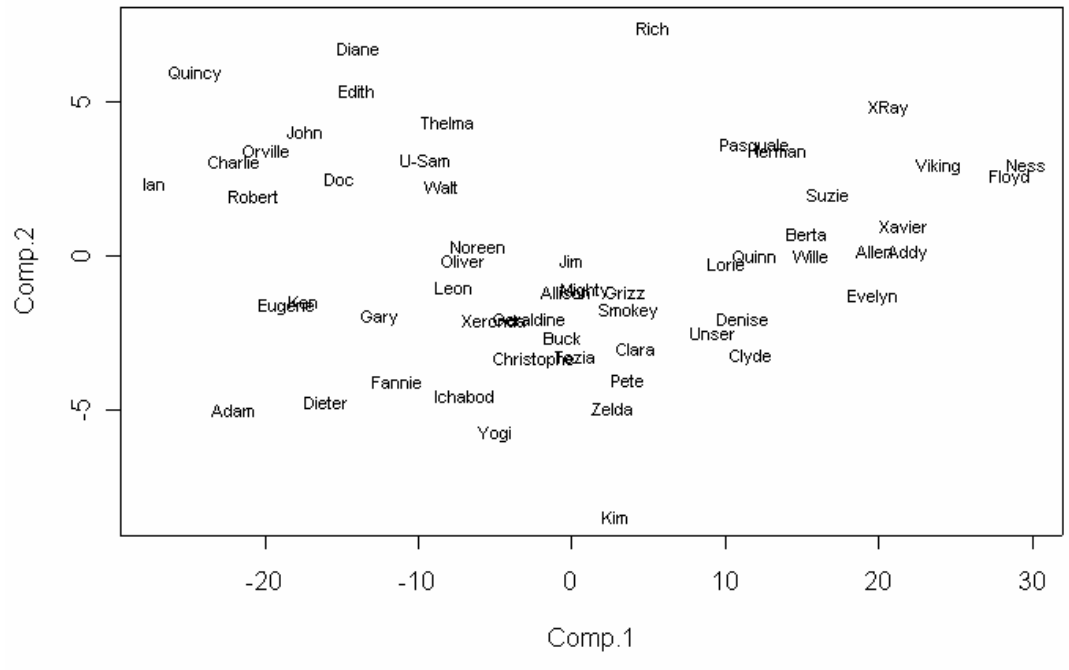
The data set below is described in *Reader's Digest* (April, 1979) and *Sports Afield*, (September, 1981).
The data set consists of several measurements for bears that were captured, measured, and released. (The full data set actually caught several of the bears multiple times over a period of years.) The variables in the data set are: estimated age in months, gender (1=male, 2=female), length of head in inches, width of head in inches, girth of the neck in inches, body length in inches, girth of the chest in inches, weight in pounds, and name. The observations are currently ordered by name.

Age	Sex	Head_L	Head_W	Neck_G	Length	Chest_G	Weight	Name
70	1	15.0	6.5	28.0	78.0	45.0	334	Adam
8	2	10.0	4.5	10.0	43.5	24.0	29	Addy
19	1	10.0	5.0	15.0	45.0	23.0	65	Allen
45	2	13.0	6.5	21.0	60.0	34.5	182	Allison
19	2	11.0	6.5	20.0	47.5	24.0	70	Berta
21	1	14.5	5.5	20.0	61.0	34.0	150	Buck
115	1	17.0	10.0	31.5	72.0	49.0	348	Charlie
22	1	13.0	6.0	20.0	63.0	35.0	172	Christophe
53	2	12.5	6.0	18.0	58.0	31.0	144	Clara
19	1	11.0	5.5	16.0	53.0	26.0	80	Clyde
17	2	11.5	5.0	15.0	52.5	28.0	76	Denise
82	2	13.5	6.5	28.0	64.0	48.0	356	Diane
56	1	15.0	7.5	26.5	73.5	41.0	262	Dieter
55	1	16.5	9.0	28.0	67.5	45.0	344	Doc
70	2	14.5	6.5	26.0	65.0	48.0	316	Edith
68	1	16.0	9.0	29.0	73.0	44.0	332	Eugene
17	2	11.0	4.5	13.0	46.0	23.0	48	Evelyn
100	2	13.0	7.0	21.0	70.0	41.0	220	Fannie
8	1	9.0	4.5	13.0	37.0	19.0	34	Floyd
31	1	15.5	6.0	23.0	69.0	42.5	289	Gary
104	2	15.5	6.5	22.0	62.0	35.0	166	Geraldine
18	1	12.5	8.5	18.0	57.3	32.8	140	Grizz
10	1	11.5	5.0	17.0	47.0	29.5	86	Herman
70	1	15.5	7.0	28.0	76.5	55.0	446	Ian
32	1	14.0	5.0	21.5	67.0	37.0	180	Ichabod
34	1	13.0	7.0	21.0	59.0	35.0	150	Jim
51	1	13.5	8.0	27.0	68.5	49.0	360	John
34	1	16.5	6.5	27.0	72.0	44.5	270	Ken
44	2	12.5	4.5	10.5	63.0	32.0	140	Kim
34	1	14.0	5.5	24.0	65.0	39.0	202	Leon
20	2	11.5	5.0	17.5	52.0	29.0	105	Lorie
32	1	13.0	8.0	21.5	59.0	33.0	166	Mighty
9	2	9.0	4.5	12.0	36.0	19.0	26	Ness
58	2	13.5	6.5	21.5	63.0	40.0	202	Noreen
45	1	13.5	7.0	24.0	64.0	39.0	204	Oliver
58	1	15.5	7.0	28.0	70.5	50.0	365	Orville
11	1	11.5	6.0	16.5	48.0	31.0	79	Pasquale
21	1	13.0	6.0	19.0	59.0	30.0	120	Pete
81	1	15.5	8.0	31.0	72.0	54.0	416	Quincy
17	1	11.5	5.0	17.0	50.5	28.0	90	Quinn
23	1	12.0	6.5	19.0	50.0	38.0	148	Rich
177	1	16.0	9.5	30.0	72.0	48.0	436	Robert
57	2	12.5	5.0	19.0	57.5	32.0	125	Smokey
11	2	9.0	5.0	15.0	46.0	27.0	62	Suzie
83	2	14.5	7.0	23.0	61.5	44.0	236	Thelma
81	2	13.0	5.0	20.0	61.0	33.0	132	Tozia
21	1	13.0	5.0	17.0	54.0	28.0	90	Unser
35	1	13.5	8.5	23.0	63.5	44.0	212	U-Sam
9	1	10.0	4.0	13.0	40.0	23.0	40	Viking
45	1	16.0	6.0	24.0	63.0	42.0	220	Walt
16	1	10.0	4.0	15.5	48.0	26.0	60	Wille
9	1	10.0	4.0	13.5	43.0	23.0	46	Xavier
57	2	13.5	7.0	20.0	64.0	38.0	204	Xeronda
16	1	10.0	5.0	15.0	41.0	26.0	64	XRay
33	1	13.5	6.0	22.0	66.5	34.0	154	Yogi
57	2	13.0	5.5	17.5	60.5	31.0	116	Zelda

```
> bears<-read.table("http://www.stat.sc.edu/~habing/courses/data/bears.txt",head=T)
> beardat<-bears[,3:7]


> var(beardat)
           Head.L    Head.W    Neck.G    Length   Chest.G
Head.L   4.445455  2.177273 10.145455  20.45182 16.806364
Head.W   2.177273  2.097078  6.338474  10.49146  9.860455
Neck.G  10.145455  6.338474 29.943425  50.00849 46.110682
Length  20.451818 10.491461 50.008490 112.75018 86.570682
Chest.G 16.806364  9.860455 46.110682  86.57068 84.013182

> bears.pca<-princomp(beardat,cor=F)
> summary(bears.pca)

Importance of components:
                           Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
Standard deviation     14.5837167 3.39161961 1.90683041 0.840558881 0.744133809
Proportion of Variance  0.9284134 0.05021335 0.01587191 0.003084191 0.002417169
Cumulative Proportion   0.9284134 0.97862673 0.99449864 0.997582831 1.000000000

> loadings(bears.pca)

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Head.L  -0.133        -0.129 -0.108  0.976
Head.W                -0.207 -0.961 -0.141
Neck.G  -0.347  0.328 -0.833  0.253 -0.117
Length  -0.703 -0.701                -0.119
Chest.G -0.602  0.627  0.495

> plot(bears.pca)
```




```
> bears.pred<-predict(bears.pca)
> plot(bears.pred[,1:2],type="n")
> text(bears.pred[,1:2],as.character(bears[,9]),cex=0.7)
```



1) How many principal components do you think are needed to make a good summary of this data set?

2) The first principal component seems to be an overall measure of size. Based on the information provided in the covariance matrix for the original data, why it is no surprise that Length receives the most weight in the first principal component and that Head.W receives the least?

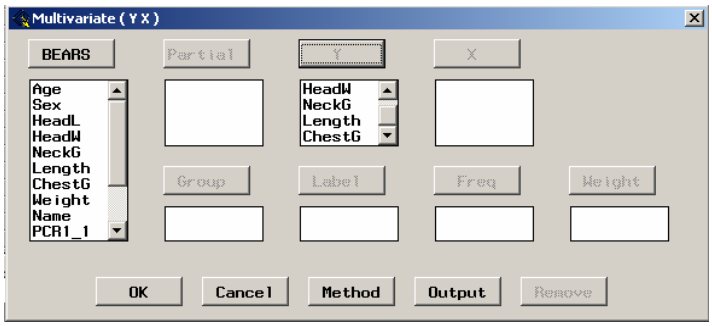
3) The first principal component seems to be an overall measure of size of the bear. Choose either the second or third component and give an interpretation of what that component seems to measure. Try to give your answer in terms of the shape of the bear or the relative sizes of its various parts.

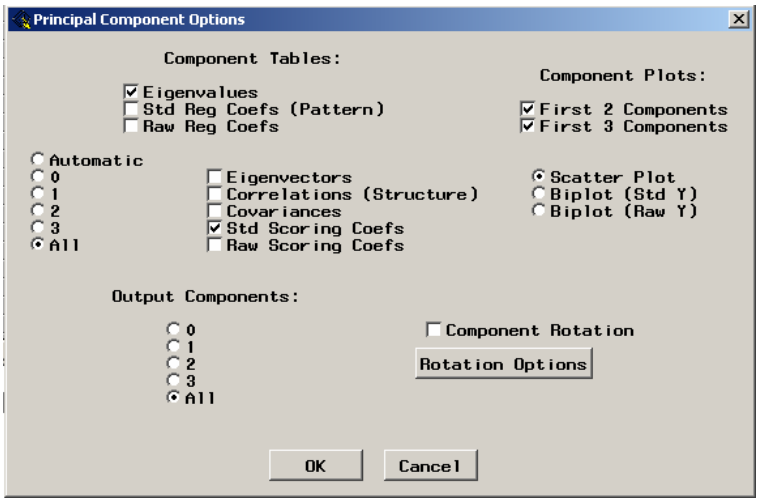
4) What percent of the total variation in the data is explained by the first two principal components combined?

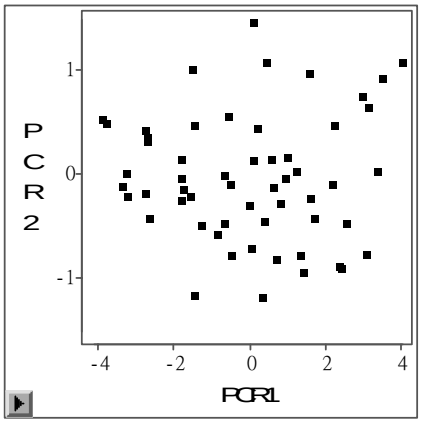
5) What is the correlation between the first and second principal component?

6) Interpret the plot of the first two principal components by explaining how Rich, Kim, Ian, and Ness's positions relate to the original data set.

```
DATA bears;
INPUT Age    Sex    $ HeadL     HeadW NeckG Length      ChestG      Weight      Name $;
CARDS;
70    1     15.0  6.5   28.0  78.0  45.0  334   Adam
<insert rest of data here>
57    2     13.0  5.5   17.5  60.5  31.0  116   Zelda
;
```







Eigenvalues (CORR)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	4.316997	3.939320	0.8634	0.8634
2	0.377678	0.230994	0.0755	0.9389
3	0.146684	0.050944	0.0293	0.9683
4	0.095739	0.032836	0.0191	0.9874
5	0.062903	.	0.0126	1.0000

Std Scoring Coefs

Variable	PCR1	PCR2	PCR3	PCR4	PCR5
HeadL	0.454613	-0.289714	0.514953	-0.530885	0.402962
HeadW	0.405120	0.859530	0.253963	0.170097	0.060471
NeckG	0.462742	0.054542	-0.449194	-0.503602	-0.572281
Length	0.451773	-0.393619	0.326079	0.583050	-0.441237
ChestG	0.459322	-0.139155	-0.601850	0.309301	0.558363

1) How many principal components do you think are needed to make a good summary of this data set?

2) Interpret what the first two principal components seem to be measuring?

3) How would you get SAS to tell you which bears were which on the scatter plot?

Charles Spearman studied the test scores of boys in a preparatory school. Each student had six scores: Classics, French, English, Mathematics, Determining the Pitch of a Note, and Music. The correlation matrix of these six test scores were:

	Classics	French	English	Math	Pitch	Music
Classics	1.00	0.83	0.78	0.70	0.66	0.63
French	0.83	1.00	0.67	0.67	0.65	0.57
English	0.78	0.67	1.00	0.64	0.54	0.51
Math	0.70	0.67	0.64	1.00	0.45	0.51
Pitch	0.66	0.65	0.54	0.45	1.00	0.40
Music	0.63	0.57	0.51	0.51	0.40	1.00

A principal components analysis using the correlation matrix returned the following results:

```
> summary(spear.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.0255297 0.7868352 0.7153546 0.59753636 0.52000750
Proportion of Variance 0.6837951 0.1031850 0.0852887 0.05950828 0.04506797
Cumulative Proportion  0.6837951 0.7869801 0.8722688 0.93177704 0.97684501
                           Comp.6
Standard deviation     0.37273307
Proportion of Variance 0.02315499
Cumulative Proportion  1.00000000


> loadings(spear.pca)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Classics -0.462                -0.128  0.266  0.835
French   -0.441 -0.119         0.228  0.734 -0.448
English  -0.416        -0.342 -0.765 -0.191 -0.298
Math     -0.397  0.255 -0.558  0.567 -0.380
Pitch    -0.367 -0.712  0.388  0.157 -0.425
Music    -0.356  0.643  0.648        -0.171
```

1) Why can we do principal components even if we don't have the raw data?

2) Give the formula for determining the fifth principal component in terms of the original standardized variables.

3) Give the mean, variance, and standard deviation of the students' fifth principal component.

4) The correlation between the students' standardized classical and French scores scores is 0.83. Give the correlation between the students' first two principal components.

5) Some of the coefficients are not shown in the output because their values were smaller than 0.10. What could we use (e.g. ask R for) to find these missing values?

6) Briefly interpret the first three principal components.