Principal Component Analysis on Bull Data with SAS

Qiang Zhang

Problem Statement:

Consider the data on bulls in Table 1. Utilize the seven variables YrHgt, FtFrBody, PrctFFB, Frame, BkFat, SaleHt, and SaleWt, perform a principal component analysis using the covariance matrix S and correlation matrix R respectively. The analysis include the following:
(a) Determine the appropriate number of components to effectively summarize the sample variability. Construct a scree plot to aid determination.
(b) Interpret the sample principal components.
(c) Do you think it is possible to develop a “body size” or “body configuration” index from the data on the seven variables above? Explain.
(d) Using the values for the first two principal components, plot the data in a two-dimensional space with y1 along the vertical axis and y2 along the horizontal axis. Can you distinguish groups representing the three breeds of cattle? Are there any outliers?
(e) Construct a Q-Q plot using the first principal component. Interpret the plot.

Solution:

Part one: do analysis with covariance matrix S.
(a) Determine the appropriate number of components to effectively summarize the sample variability. Construct a scree plot to aid determination.

Observations	76
Variables	7

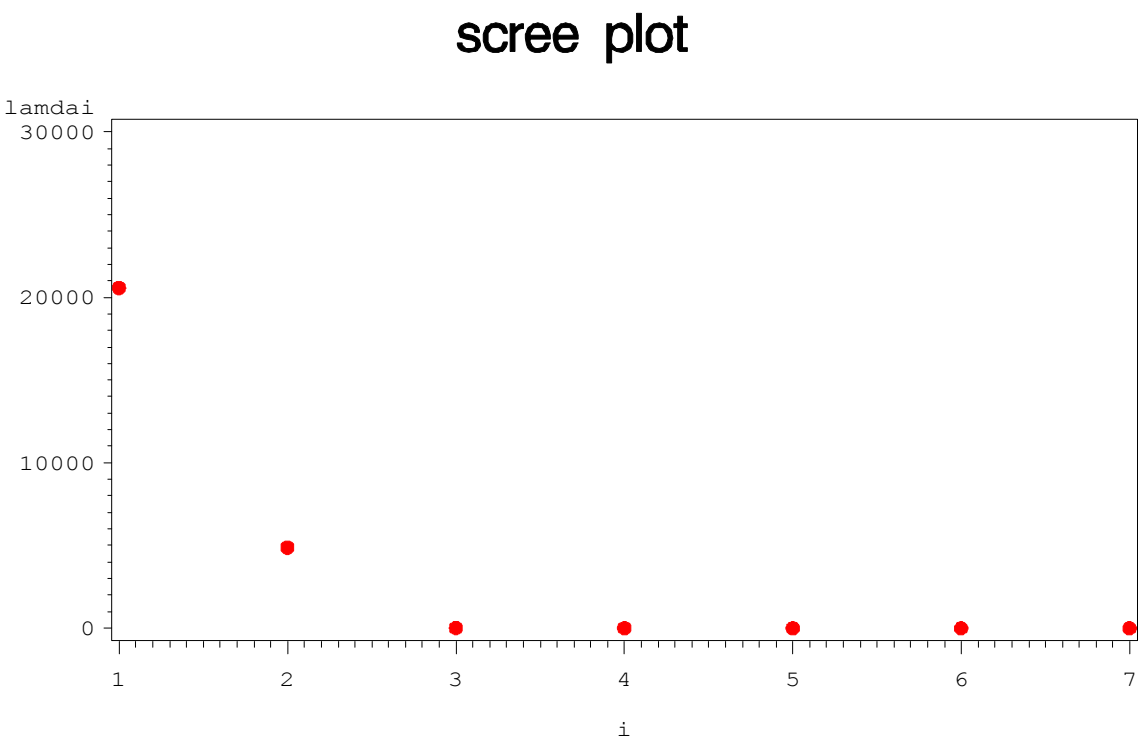
Simple Statistics							
	YrHgt	**FtFrbody**	**PrctFFB**	**Fram**	**BkFat**	**SaleHt**	**SaleWt**
Mean	50.52236842	995.9473684	70.88157895	6.315789474	0.1967105263	54.12631579	1555.289474
StD	1.73148096	92.7056841	3.26980980	0.926794135	0.0895676751	2.00448620	129.810099

Covariance Matrix							
	YrHgt	FtFrbody	PrctFFB	Fram	BkFat	SaleHt	SaleWt
YrHgt	2.99803	100.13053	2.96002	1.50884	-0.05339	2.98314	82.81077
FtFrbody	100.13053	8594.34386	209.50435	51.95018	-1.39818	129.94007	6680.30877
PrctFFB	2.96002	209.50435	10.69166	1.45923	-0.14299	3.41422	83.92540
Fram	1.50884	51.95018	1.45923	0.85895	-0.02161	1.48758	44.32070
BkFat	-0.05339	-1.39818	-0.14299	-0.02161	0.00802	-0.05065	2.41296
SaleHt	2.98314	129.94007	3.41422	1.48758	-0.05065	4.01796	147.28961
SaleWt	82.81077	6680.30877	83.92540	44.32070	2.41296	147.28961	16850.66175

Total Variance	25463.580231

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	20579.6126	15704.9378	0.8082	0.8082
2	4874.6748	4869.2456	0.1914	0.9996
3	5.4292	2.1129	0.0002	0.9998
4	3.3163	2.8475	0.0001	1.0000
5	0.4688	0.3948	0.0000	1.0000
6	0.0741	0.0695	0.0000	1.0000
7	0.0045		0.0000	1.0000

Eigenvectors							
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**
YrHgt	0.005887	0.009680	0.286337	0.608787	0.535569	-.509727	0.024592
FtFrbody	0.487047	0.872697	-.034277	-.003227	0.000444	-.000457	-.000253
PrctFFB	0.008526	0.029196	0.904389	-.425175	0.008388	0.010389	0.014293
Fram	0.003112	0.004886	0.133267	0.311194	0.390573	0.855204	-.037984
BkFat	0.000069	-.000493	-.018864	-.005278	0.011906	0.043786	0.998778
SaleHt	0.009330	0.008577	0.284215	0.593037	-.748598	0.082331	0.013820
SaleWt	0.873259	-.487193	0.004847	-.005597	0.002665	-.000341	-.000256



(b) Interpret the sample principal components.

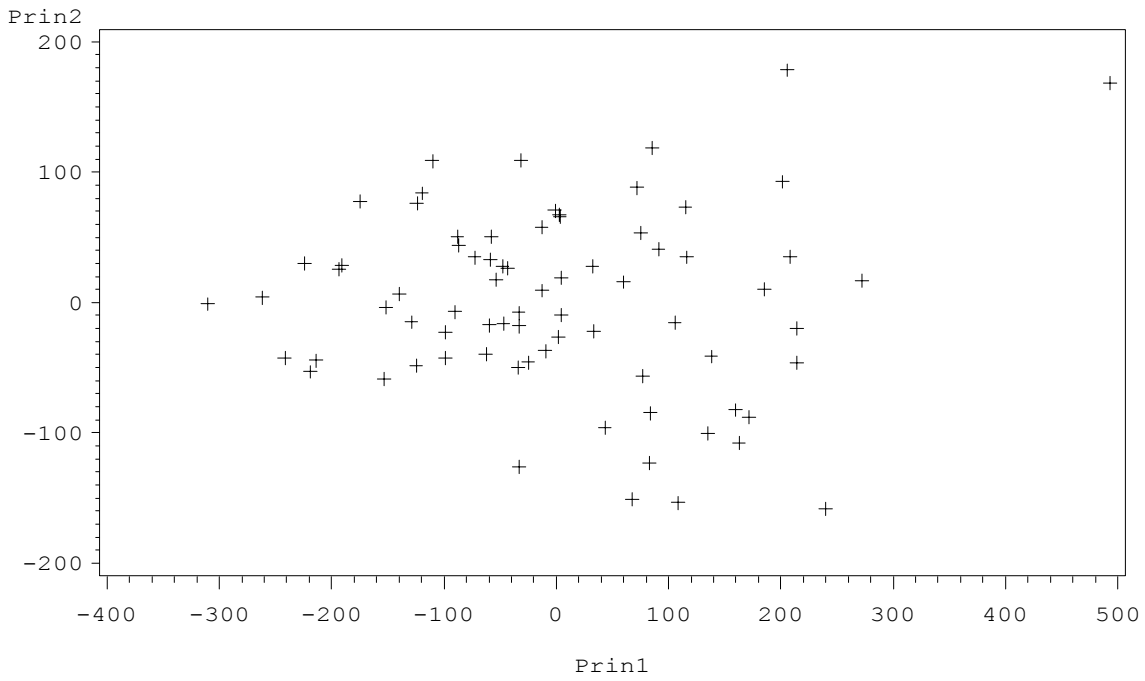
The first principal component is dominant by SaleWt, so it can be interpreted as the body-weight. And the second principal component dominant by FfFrbody, which can be interpreted as fat-free-weight.

(c) Do you think it is possible to develop a “body size” or “body configuration” index from the data on the seven variables above? Explain.

When the body size is big, the body weight is big too. So the body-size can represent the body weight. So the first component can be taken as “body size” and the second can be served as “body configuration” index.

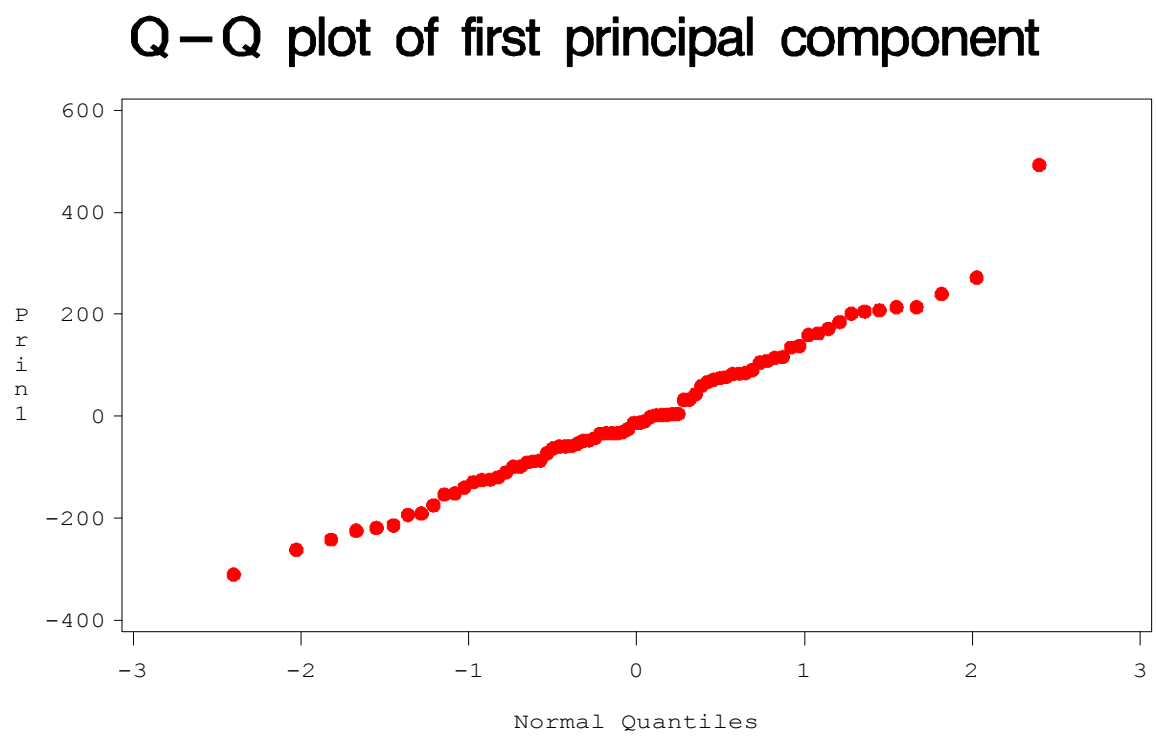
(d) Using the values for the first two principal components, plot the data in a two-dimensional space with y1 along the vertical axis and y2 along the horizontal axis. Can you distinguish groups representing the three

breeds of cattle? Are there any outliers?



From the graph above we cannot distinguish groups representing the three breeds of cattle.

(e) Construct a Q-Q plot using the first principal component. Interpret the plot.



The plot indicates clear normality.

Part two: do analysis with correlation matrix R.

(a) Determine the appropriate number of components to effectively summarize the sample variability. Construct a scree plot to aid determination.

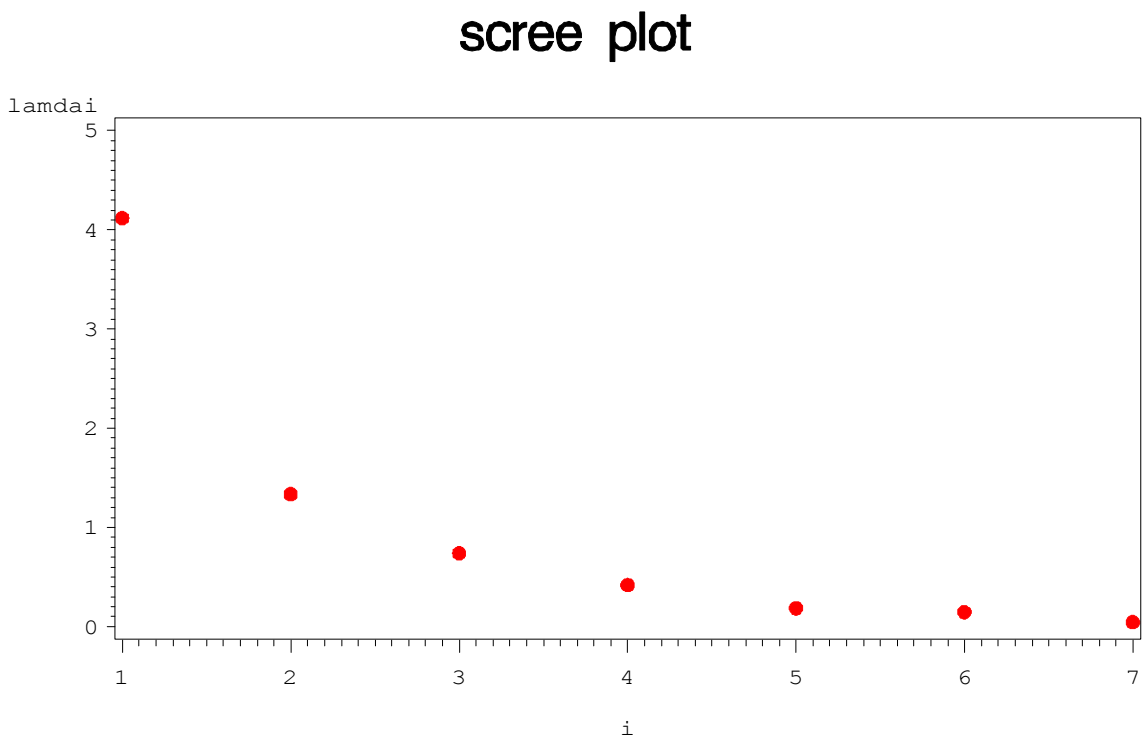
Observations	76
Variables	7

Simple Statistics							
	YrHgt	**FtFrbody**	**PrctFFB**	**Fram**	**BkFat**	**SaleHt**	**SaleWt**
Mean	50.52236842	995.9473684	70.88157895	6.315789474	0.1967105263	54.12631579	1555.289474
StD	1.73148096	92.7056841	3.26980980	0.926794135	0.0895676751	2.00448620	129.810099

Correlation Matrix							
	YrHgt	**FtFrbody**	**PrctFFB**	**Fram**	**BkFat**	**SaleHt**	**SaleWt**
YrHgt	1.0000	0.6238	0.5228	0.9402	-.3443	0.8595	0.3684
FtFrbody	0.6238	1.0000	0.6911	0.6046	-.1684	0.6993	0.5551
PrctFFB	0.5228	0.6911	1.0000	0.4815	-.4883	0.5209	0.1977
Fram	0.9402	0.6046	0.4815	1.0000	-.2604	0.8007	0.3684
BkFat	-.3443	-.1684	-.4883	-.2604	1.0000	-.2821	0.2075
SaleHt	0.8595	0.6993	0.5209	0.8007	-.2821	1.0000	0.5661
SaleWt	0.3684	0.5551	0.1977	0.3684	0.2075	0.5661	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.12069793	2.78356863	0.5887	0.5887
2	1.33712930	0.59574675	0.1910	0.7797
3	0.74138255	0.31995733	0.1059	0.8856
4	0.42142522	0.23561929	0.0602	0.9458
5	0.18580593	0.03930356	0.0265	0.9723
6	0.14650237	0.09944566	0.0209	0.9933
7	0.04705670		0.0067	1.0000

Eigenvectors							
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**
YrHgt	0.449931	-.042790	-.415709	0.113356	0.065871	-.072234	0.774926
FtFrbody	0.412326	0.129837	0.450292	0.247479	-.719343	-.177061	0.017768
PrctFFB	0.355562	-.315508	0.568273	0.314787	0.579367	0.127800	-.002397
Fram	0.433957	0.007728	-.452345	0.242818	0.142995	-.434144	-.582337
BkFat	-.186705	0.714719	-.038732	0.618117	0.160238	0.208017	0.042442
SaleHt	0.452854	0.101315	-.176650	-.215769	-.109535	0.799288	-.236723
SaleWt	0.269947	0.600515	0.253312	-.582433	0.290547	-.276561	0.047036



(b) Interpret the sample principal components.

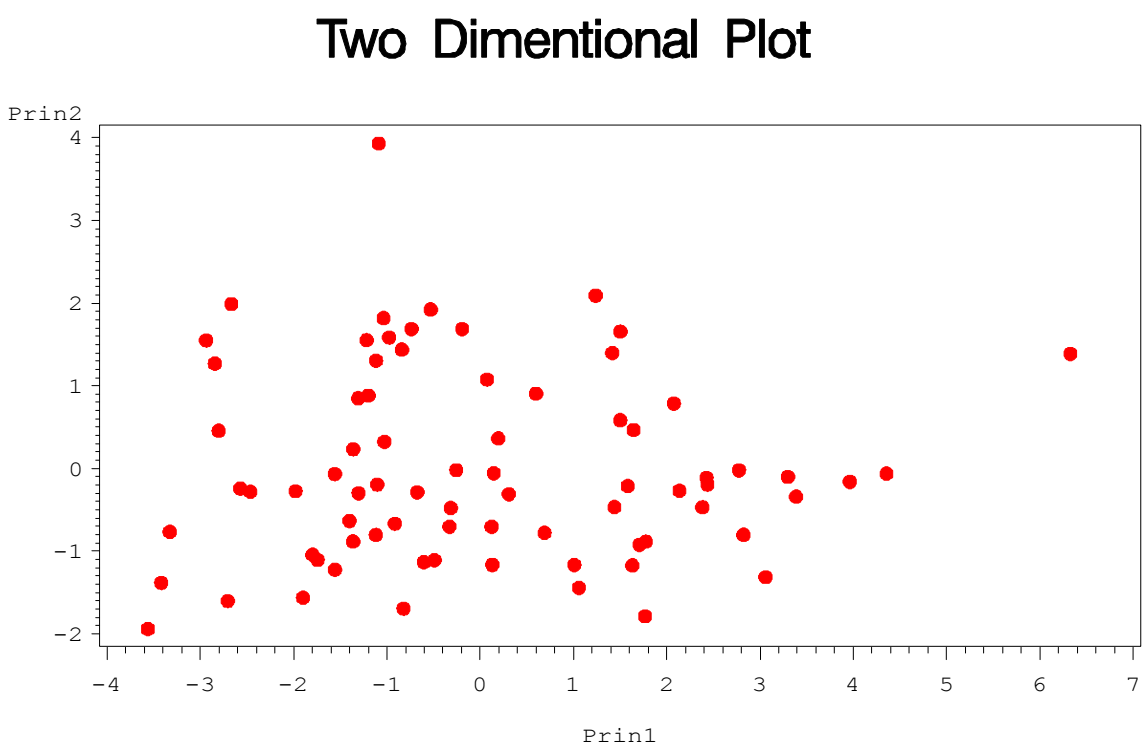
The first principal component loads the average body's weight. Second and the third one are related to body fat.

(c) Do you think it is possible to develop a "body size" or "body configuration" index from the data on the seven variables above? Explain.

Yes. The first principal component is body size. And the second and the third principal components can be classified as body configuration index.

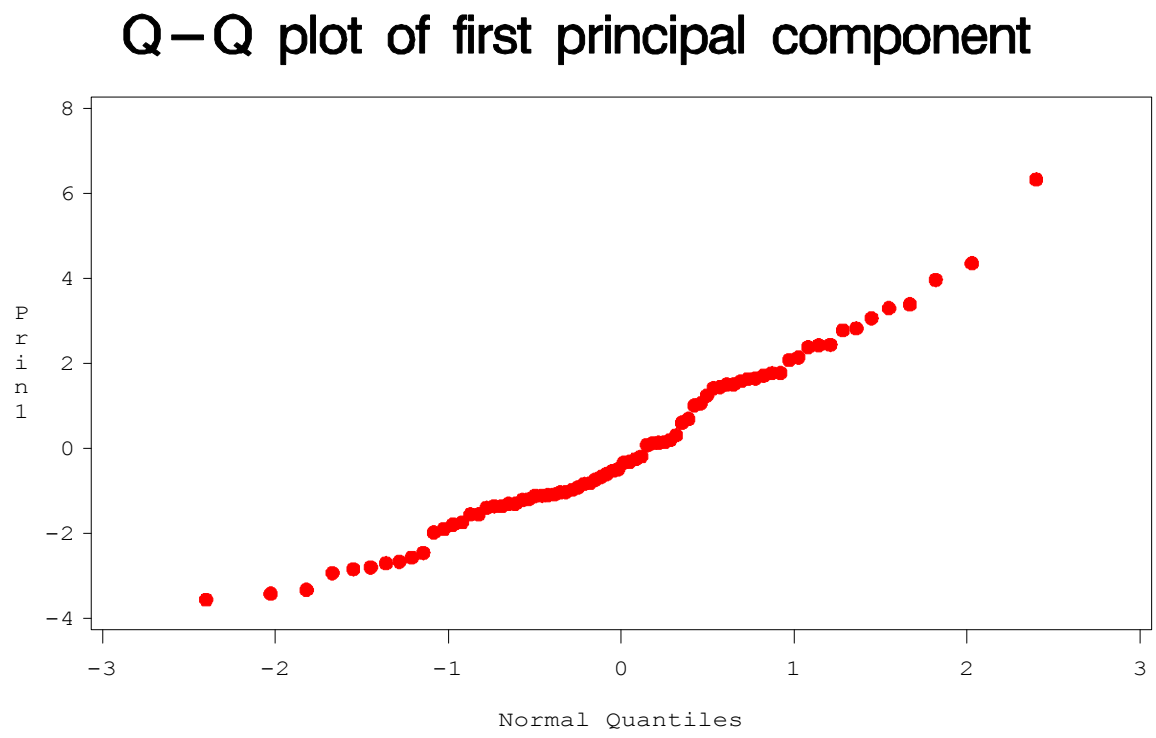
(d) Using the values for the first two principal components, plot the data in a two-dimensional space with y1 along the vertical axis and y2 along the horizontal axis. Can you distinguish groups representing the three

breeds of cattle? Are there any outliers?



From this plot we cannot tell the groups representing the three breeds of cattle. There are two outliers.

(e) Construct a Q-Q plot using the first principal component. Interpret the plot.



The QQ-plot indicates clear normality.

SAS Code:

```
/*PCA Bull*/
options nodate;

data onbull;infile 'D:\SAS\MultiVariateCD\T1-10.dat';
input breed SalePr YrHgt FtFrbody PrctFFB Fram BkFat SaleHt SaleWt;
run;

proc print;run;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_1.rtf';
proc princomp cov data=onbull out=pctrack2;
var YrHgt FtFrbody PrctFFB Fram BkFat SaleHt SaleWt;
run;
proc princomp data=onbull out=pctrack3;
var YrHgt FtFrbody PrctFFB Fram BkFat SaleHt SaleWt;
run;
ods rtf close;run;quit;


proc sort data=pctrack2;
by breed;
run;
proc sort data=pctrack2;
by prin1 prin2;
run;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_graph1.rtf';
proc gplot data=pctrack2;
plot prin2*prin1=2;
symbol2 c=red i=none value=dot mode=include;run;quit;
ods rtf close;run;quit;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_graph2.rtf';
proc capability data=pctrack2;
var prin1;
qqplot;
title 'Q-Q plot of first principal component';
run;
ods rtf close;run;quit;

proc sort data=pctrack3;
by breed;
```

```
run;
proc sort data=pctrack3;
by prin1 prin2;
run;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_graph3.rtf';
proc gplot data=pctrack3;
plot prin2*prin1=3;
symbol3 c=red i=none value=dot mode=include;
title "Two Dimentional Plot " ;run;quit;
ods rtf close;run;quit;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_graph4.rtf';
proc capability data=pctrack3;
var prin1;
qqplot;
title 'Q-Q plot of first principal component';
run;
ods rtf close;run;quit;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_2.rtf';
data d1;
input lamdai i;
cards;
20579.6126  1
4874.6748   2
5.4292      3
3.3163      4
0.4688      5
0.0741      6
0.0045      7
;
run;


symbol1 c=blue i=join l=1  w=2 value=p mode=include;

proc gplot data=d1;
plot lamdai * i = 2;
title "scree plot";
run;
ods rtf close;run;quit;

ods rtf file='D:\SAS\Project\PCA\PCA_bull_3.rtf';
```

```
data d2;
input lamdai i;
cards;
4.12069793   1
1.33712930   2
0.74138255   3
0.42142522   4
0.18580593   5
0.14650237   6
0.04705670   7

;
run;

symbol1 c=blue i=join l=1  w=2 value=none mode=include;
proc gplot data=d2;
plot lamdai * i = 2;
title "scree plot";
run;
ods rtf close;run;quit;
```