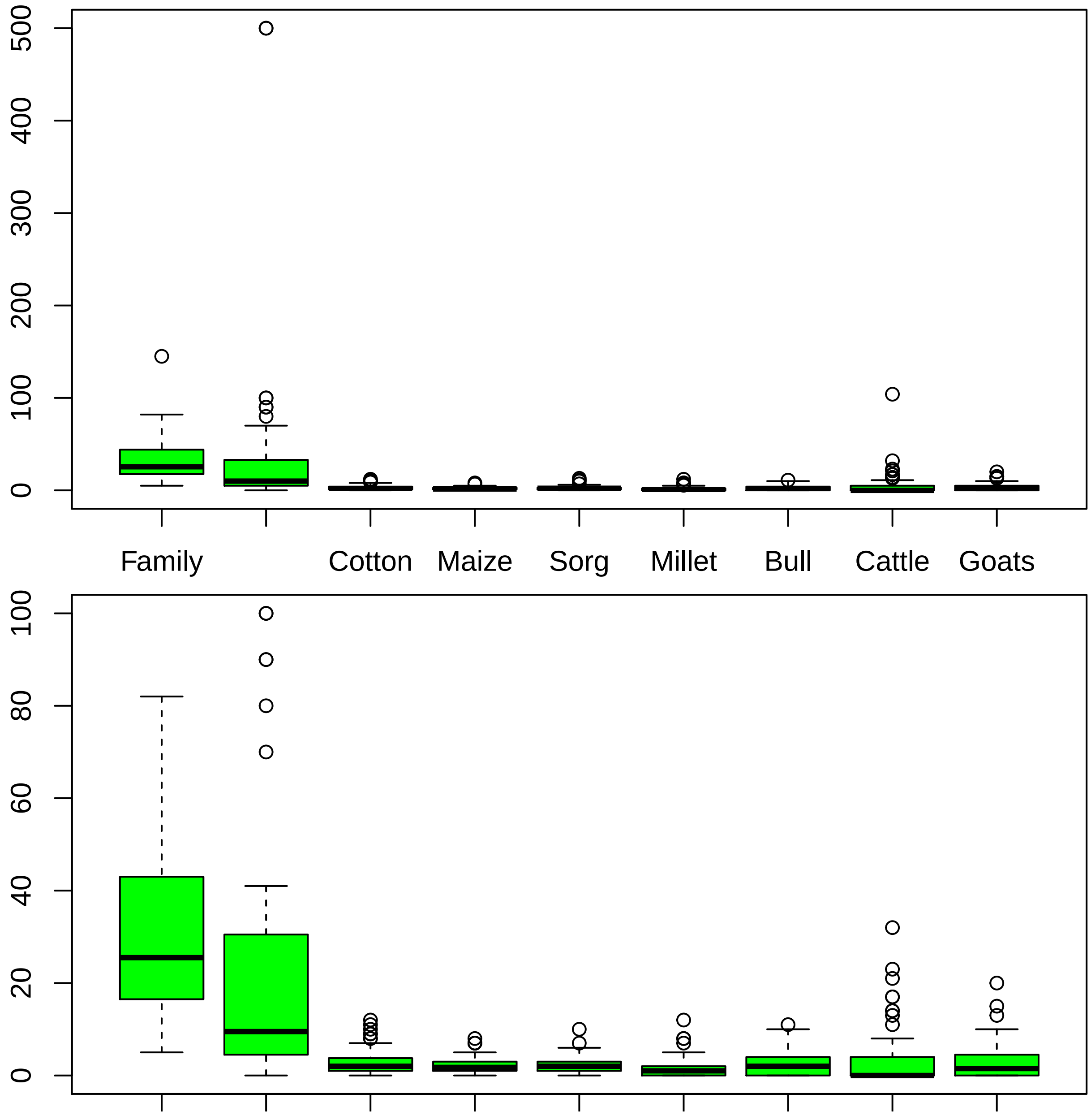
Stat 460 HW3

Wulf Novak

March 24, 2019

Mali Farm Data PCA



The first boxplot illustrates severe outliers in the Family Variable, DistRD Variable, and Cattle Variable. I removed 1 Family observation (145), 2 DistRD observations (Both 500), and 1 Cattle observation (109) - The second boxplot features these omissions and shows arguably less severe outliers.

Correlation Matrix

	Family	DistRD	Cotton	Maize	Sorg	Millet	Bull	Cattle	Goats
Family	1.0000	-0.0315	0.7568	0.6553	0.3880	0.4965	0.7334	0.5533	0.3617
DistRD	-0.0315	1.0000	-0.0311	0.1098	-0.2184	-0.0832	0.0284	0.0601	0.1710
Cotton	0.7568	-0.0311	1.0000	0.7157	0.4069	0.3521	0.8213	0.5987	0.3726
Maize	0.6553	0.1098	0.7157	1.0000	-0.0299	0.1756	0.6290	0.5299	0.0509
Sorg	0.3880	-0.2184	0.4069	-0.0299	1.0000	0.3638	0.3182	0.0602	0.2371
Millet	0.4965	-0.0832	0.3521	0.1756	0.3638	1.0000	0.3362	0.1253	0.2498
Bull	0.7334	0.0284	0.8213	0.6290	0.3182	0.3362	1.0000	0.6698	0.5038
Cattle	0.5533	0.0601	0.5987	0.5299	0.0602	0.1253	0.6698	1.0000	0.3819
Goats	0.3617	0.1710	0.3726	0.0509	0.2371	0.2498	0.5038	0.3819	1.0000

Here I have chosen to use the correlation matrix. This is due to the variables having both different scales and also different units. For example, the DistRD variable ranges from 0 to 500, whereas the Millet variable ranges from 0 to 12 - each with unknown units. Also, since it's not clear what the data represents, it may be the 'safer' option to use the correlation matrix.

Eigenvalues

```
## [1] 4.1851310 1.4380868 1.0845001 0.7918176 0.6043248 0.3661359 0.2400236
## [8] 0.1718252 0.1181550
```

```
## [1] 0.4650146 0.6248020 0.7453020 0.8332817 0.9004289 0.9411107 0.9677800
## [8] 0.9868717 1.0000000
```

The last 4 eigenvalues are quite small (0.5 and smaller). The larger the eigenvalue, the larger the proportion of variation that eigenvalue accounts for. The first eigen value accounts for roughly 46% of variation. The first 2 eigen values account for roughly 62% of the variation, first 3 for 74%, first 4 for 83% of the variation, etc.

PCA Individual Variances and Eigenvalue Comparisons

```
std.mali <- scale(Malidt)
pcaMalidt <- t(t(eig$vectors) %*% t(std.mali))
round(sum(diag(cov(pcaMalidt))) - sum(eig$values),5)
```

```
## [1] 0
```

The sum of individual variances perfectly equals the sum of the eigenvalues - This is a good sign.

```
PCAmali <- princomp(Malidt, cor = TRUE)
sum(diag(cov(PCAmali$scores))) - sum(eig$values)
```

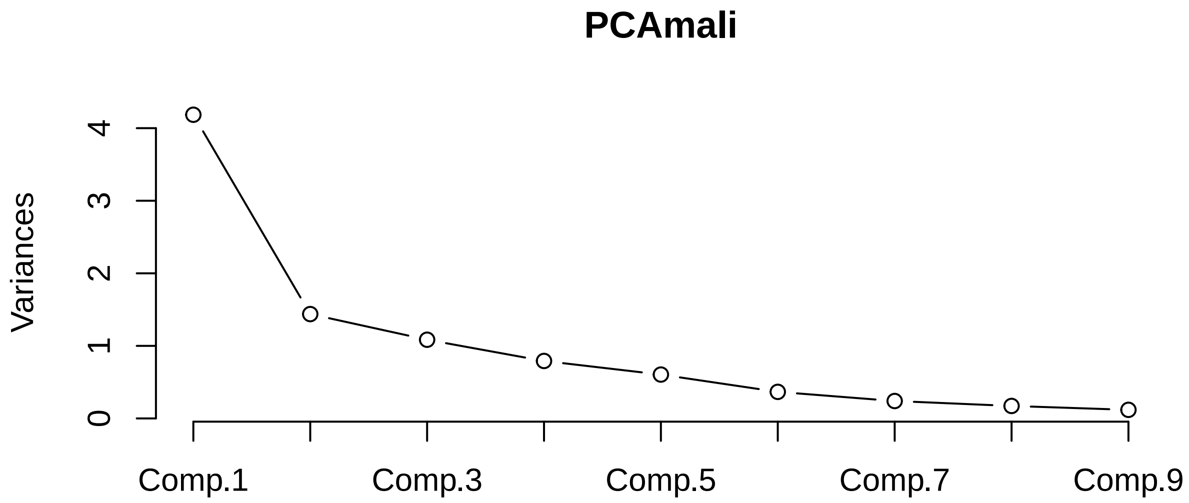
```
## [1] 0.1267606
```

Notably, when using the PCA function the sum of the individual variances does NOT equal the sum of the eigenvalues. :(
However, the PCA function output was used to create the following screeplot

PCA Selection

```
## [1] 1
```

```
## [1] 4.1851310 1.4380868 1.0845001 0.7918176 0.6043248 0.3661359 0.2400236
## [8] 0.1718252 0.1181550
```

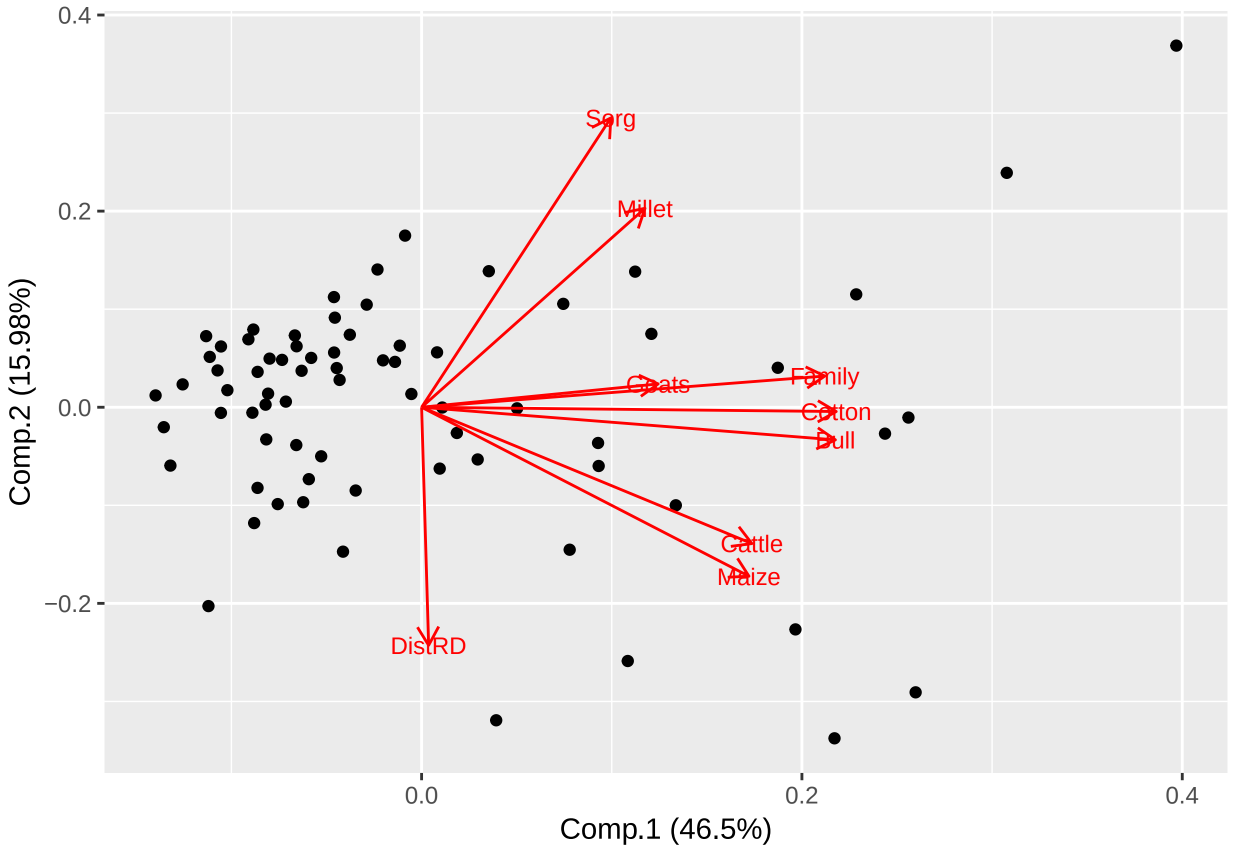


Without having previous knowledge about the level of variation we want our PCs to explain, PC selection can be determined by a scree plot and the elbow method, or by choosing eigenvalues greater than the average eigenvalues - (or by googling more methods). The above screeplot shows an elbow at 2 PCs. The 'average method' would suggest using 3 PCs, because those PC eigenvalues are above the average eigenvalue of 1. I am going to use the first 2 PCs.

Eigenvectors for chosen PCs

```
##              [,1]         [,2]
##  [1,] 0.433842713 -0.065088695
##  [2,] 0.007587031  0.496670914
##  [3,] 0.446140316  0.008917253
##  [4,] 0.352228405  0.352571495
##  [5,] 0.203622111 -0.603667416
##  [6,] 0.240361102 -0.415159516
##  [7,] 0.445273680  0.068042477
##  [8,] 0.355411548  0.284473439
##  [9,] 0.254549533 -0.048668251
```

PC Plot



The most significant variables for PC 1 were Family, Cotton, and Bull. The most significant variables for PC 2 were DistRD. Also, cattle and maize are closely grouped and effect both PCs, but primarily PC 1.

Flea Beetle Data PCA

```
Fleadt <- Fleadt[-c(20), 2:9]
```

I have chosen to remove observation 20 due to the NAs for 4 of the variables. This is done in order to produce a proper correlation matrix without NAs. In addition, I have removed the experiment column.

Covariance Matrix

	x1...2	x2...3	x3...4	x4...5	x1...6	x2...7	x3...8	x4...9
x1...2	187.5965	176.8626	48.3713	113.5819	-3.4269	11.3655	59.0146	0.8743
x2...3	176.8626	345.3860	75.9795	118.7807	-43.6462	-131.5643	-11.8070	-58.6374
x3...4	48.3713	75.9795	66.3567	16.2427	4.9211	20.7164	35.7398	17.2047
x4...5	113.5819	118.7807	16.2427	239.9415	-62.9094	-139.8801	-59.0263	-48.9737
x1...6	-3.4269	-43.6462	4.9211	-62.9094	89.6901	122.7281	25.9825	38.7398
x2...7	11.3655	-131.5643	20.7164	-139.8801	122.7281	401.9181	165.4123	102.6433
x3...8	59.0146	-11.8070	35.7398	-59.0263	25.9825	165.4123	167.2632	73.4035
x4...9	0.8743	-58.6374	17.2047	-48.9737	38.7398	102.6433	73.4035	186.7076

Notably, some of the pre-requisites for PCA, such as all of the variables being similarly correlated, does not exist with this data set. Having looked at the pairs plot (Not included to save space), and looking at the correlation/covariance matrices, it is clear that some of the variables have little correlation, and others have either positive or negative correlation. This indicates that this data is not ideal for PCA. Despite this, I have chosen to use the covariance matrix to continue the analysis because each column has data in a similar numerical range.

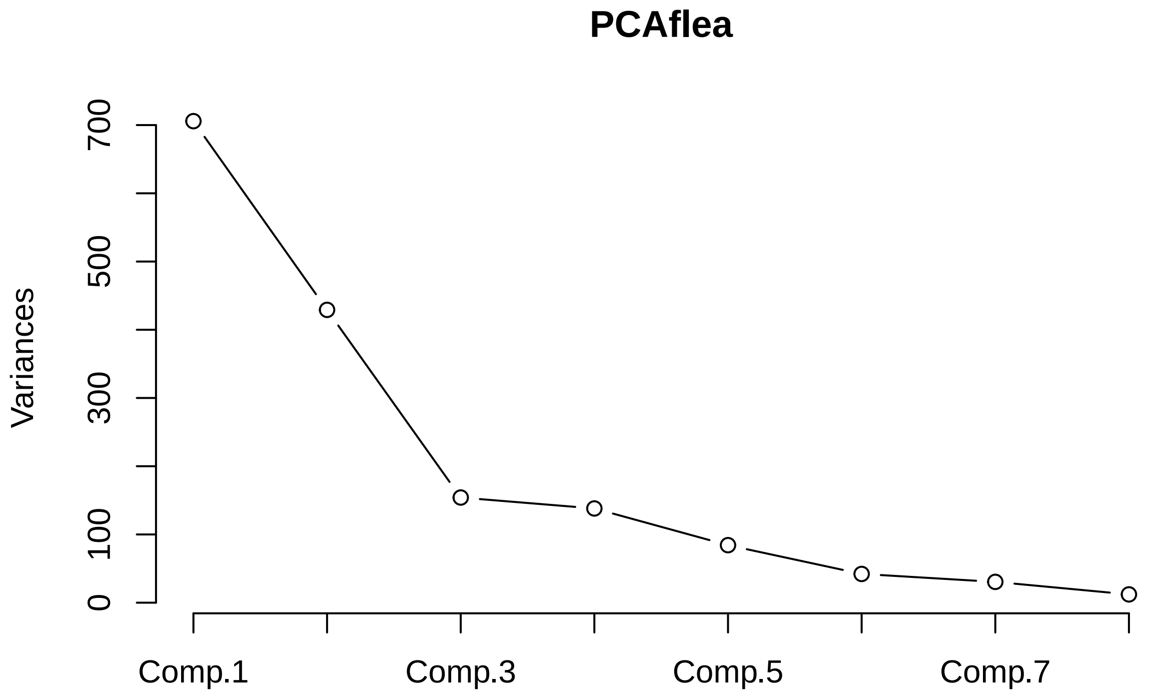
Eigenvalues

```
## [1] 744.93751 453.01212 162.67202 145.76294  88.98040  44.47269  32.17586
## [8]  12.84612
```

```
## [1] 0.4421362 0.7110086 0.8075579 0.8940713 0.9468830 0.9732785 0.9923756
## [8] 1.0000000
```

The first 2 eigenvalues for this data set are ‘large’ (745 and 453), whereas the remaining eigenvalues are relatively small (below 162 and smaller). Nearly 71% of the variation is explained by the first 2 eigenvalues, and then the sharp drop off in eigenvalue size results in increasingly small amounts of variation explantion from the remaining eigenvalues.

PCA Selection



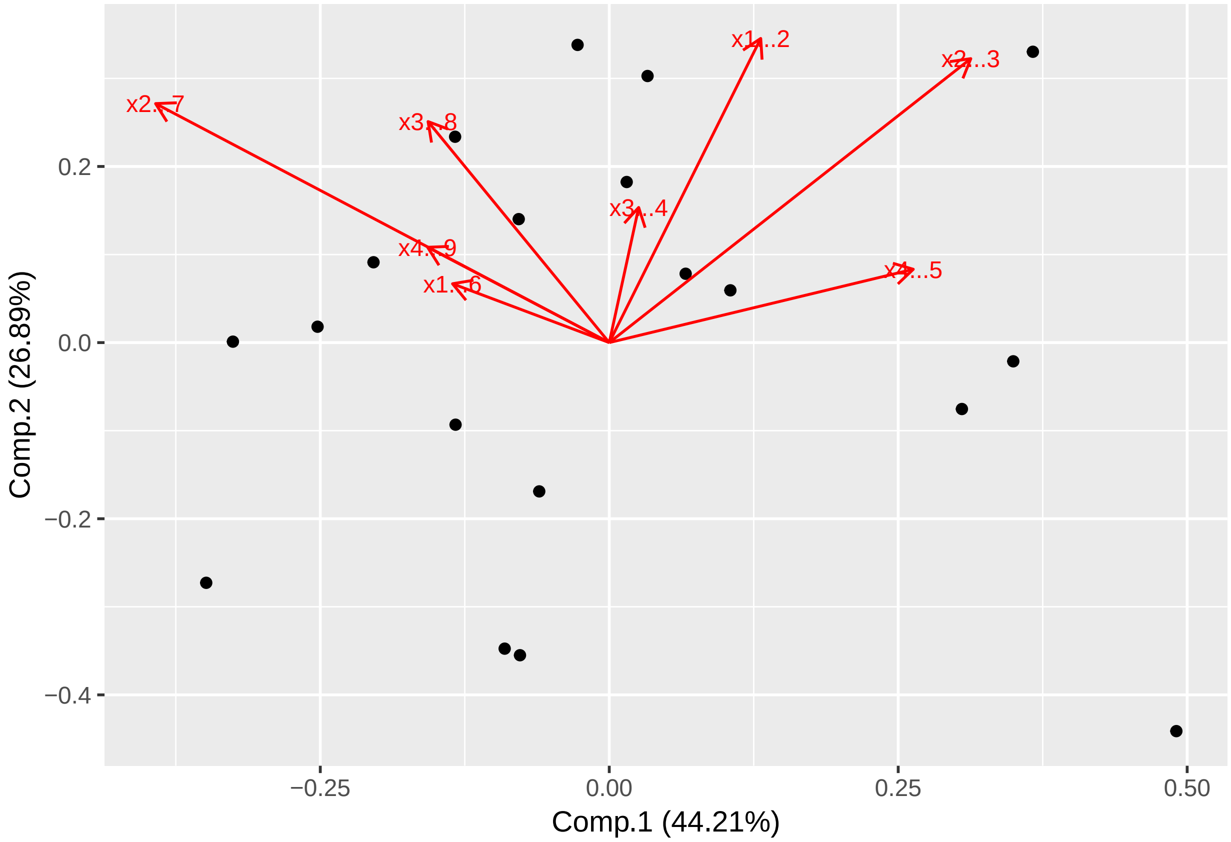
Looking at the screeplot, a noticable elbow exists after the first 3 PCs. Because of this, only the first 3 PCs will be reported.

Eigenvectors for chosen PCs

```
eig$vectors[,1:3]
```

```
##             [,1]       [,2]         [,3]
## [1,] -0.20554772 -0.5416663 -0.168452340
## [2,] -0.49057219 -0.5056867  0.476644905
## [3,] -0.03992397 -0.2406940  0.083659346
## [4,] -0.41253559 -0.1305739 -0.718357671
## [5,]  0.21267134 -0.1048153  0.105394152
## [6,]  0.61555335 -0.4256178  0.007015253
## [7,]  0.24584927 -0.3936236 -0.031772054
## [8,]  0.24643994 -0.1697622 -0.457416888
```

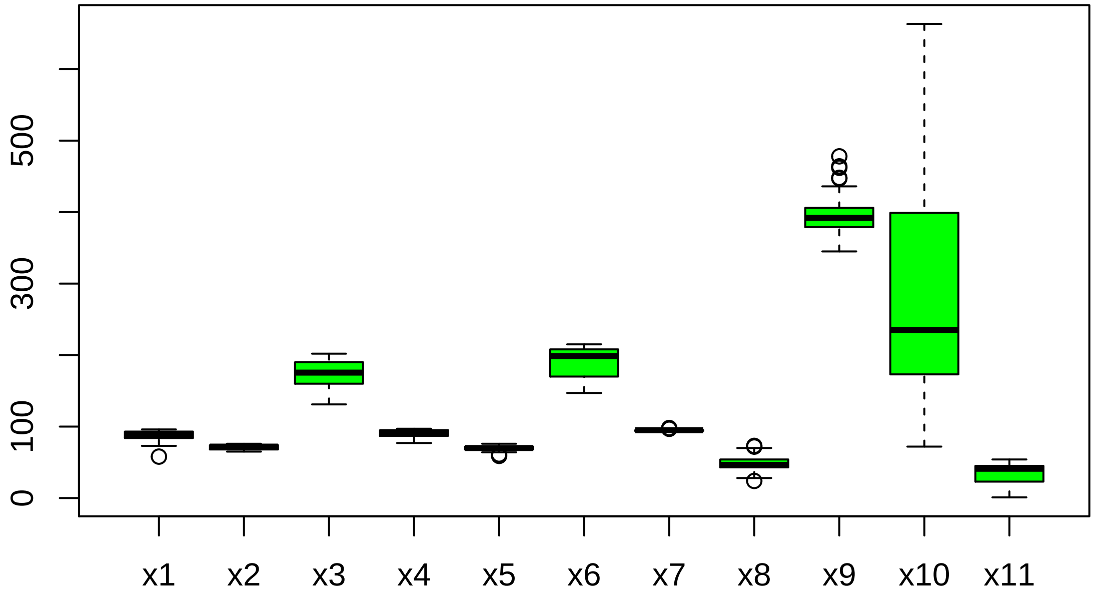
PC Plot



Unfortunately, I haven't been able to produce multiple PC graphs for each combination of the PCs as I haven't found a way to do so with the princomp output. Because of this, I will only describe the variable contribution to the first to PCs explained by the above graph. The variables the contribute the most to PC 1 are x2. . . 3 and x2. . . 7 - however, those vectors are nearly at a 45 degree angle and therefore also contribute a reasonable amount to PC 2. x1. . . 2, the last variable that contributes a lot to PC 1 and 2 mostly contributes to PC 2.

Temperature Data PCA

```r
boxplot(Tempdt, col = 'green')
```



This boxplot accentuates how certain variables have wildly different levels of variation. Notably, X1 and X4 seem to very similarly distributed data, and X2 and X5 seem to have very similarly distributed data.

Correlation Matrix

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
x1	1.0000	0.6705	0.7850	0.7136	0.3796	0.6256	-0.1733	-0.5508	-0.5776	-0.1880	0.5621
x2	0.6705	1.0000	0.9324	0.8400	0.6809	0.8185	-0.1657	-0.3179	-0.4527	0.0304	0.5389
x3	0.7850	0.9324	1.0000	0.9143	0.5907	0.8695	-0.1597	-0.5169	-0.6492	-0.0985	0.6876
x4	0.7136	0.8400	0.9143	1.0000	0.5705	0.8751	-0.1034	-0.5109	-0.6257	-0.0899	0.7233
x5	0.3796	0.6809	0.5907	0.5705	1.0000	0.7808	-0.1215	0.2240	-0.0400	0.4109	0.3278
x6	0.6256	0.8185	0.8695	0.8751	0.7808	1.0000	-0.0406	-0.2779	-0.5032	0.1239	0.7087
x7	-0.1733	-0.1657	-0.1597	-0.1034	-0.1215	-0.0406	1.0000	0.1532	0.2781	-0.1469	-0.1838
x8	-0.5508	-0.3179	-0.5169	-0.5109	0.2240	-0.2779	0.1532	1.0000	0.8858	0.3890	-0.6424
x9	-0.5776	-0.4527	-0.6492	-0.6257	-0.0400	-0.5032	0.2781	0.8858	1.0000	0.2188	-0.8162
x10	-0.1880	0.0304	-0.0985	-0.0899	0.4109	0.1239	-0.1469	0.3890	0.2188	1.0000	0.0434
x11	0.5621	0.5389	0.6876	0.7233	0.3278	0.7087	-0.1838	-0.6424	-0.8162	0.0434	1.0000

The correlation matrix was chosen due to the variables being in very different numerical ranges, which may indicated different units of measurement.

Eigenvalues

```
R <- cor(Tempdt)
eig <- eigen(R)
eig$values
```

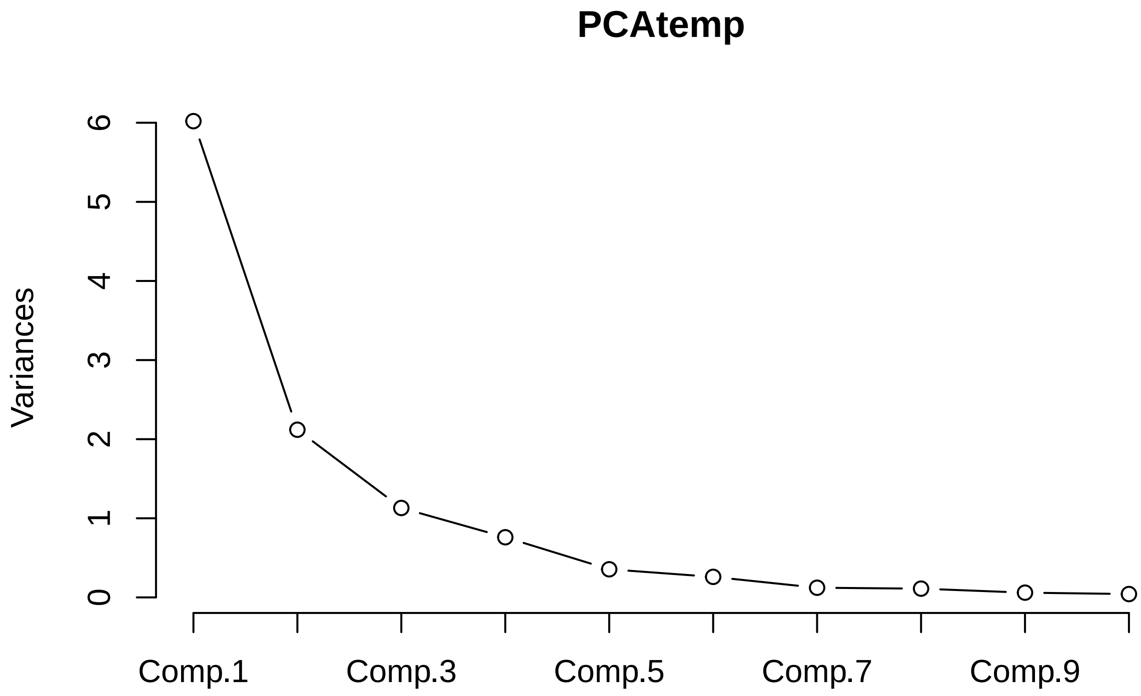
```
##  [1] 6.02024515 2.11933612 1.13029100 0.76001708 0.35535540 0.25934244
##  [7] 0.12207563 0.11048840 0.05980829 0.04218515 0.02085533
```

```
cumsum(eig$values)/sum(eig$values)
```

```
##  [1] 0.5472950 0.7399619 0.8427157 0.9118081 0.9441132 0.9676897 0.9787875
##  [8] 0.9888319 0.9942690 0.9981041 1.0000000
```

The first eigenvalue is very large (6). The first 3 eigenvalues are above 1, and the remaining 8 are below 1. The first eigenvalue explains the majority of the variation at 54%, first 2 explain 74%, first 3 explain 84% and the first 4 eigenvalues explain 91%.

PCA Selection



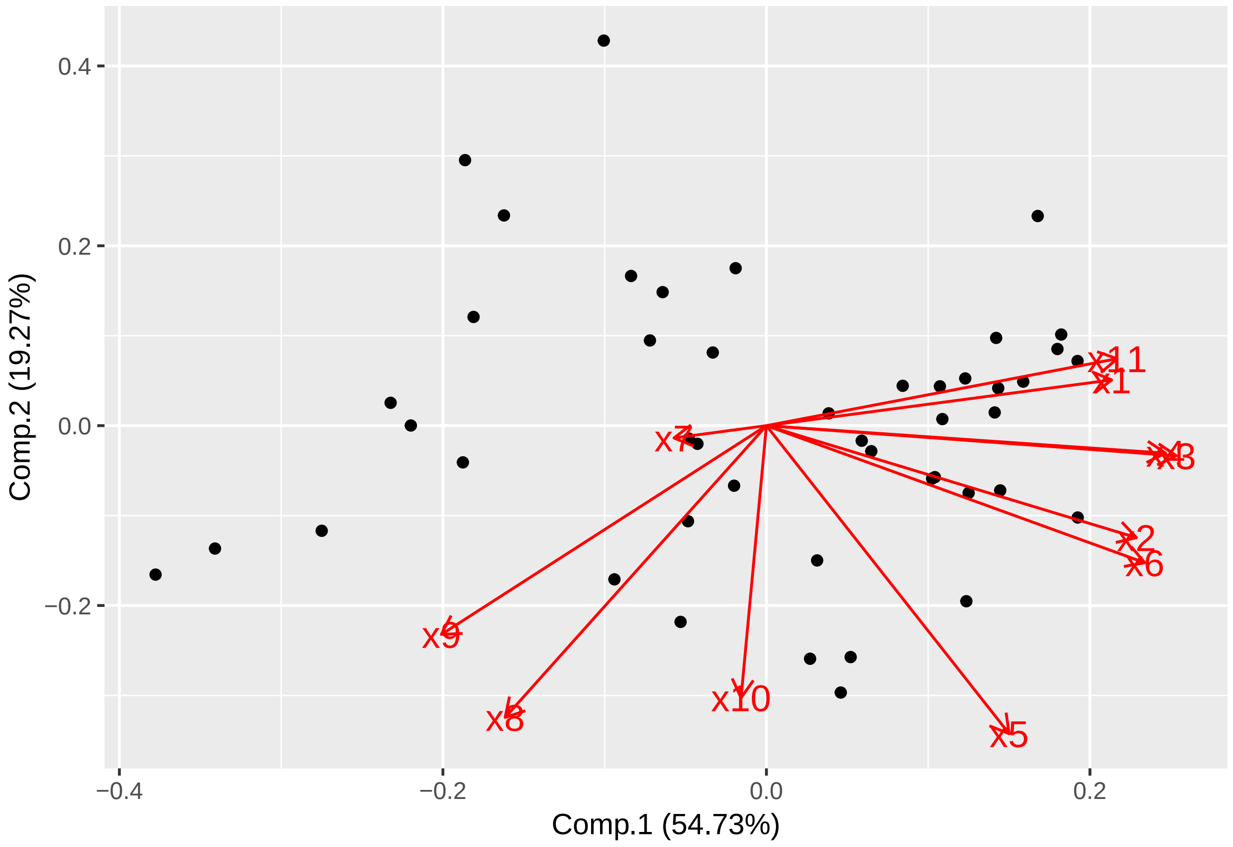
The scree plot does not show a clear elbow. However, I will use the first 3 PCs because their eigenvalues are all above 1.

Eigenvectors for chosen PCs

```
eig$vectors[,1:3]
```

```
##              [,1]        [,2]        [,3]
##  [1,] -0.33042817 -0.07872408 -0.08800766
##  [2,] -0.35415881  0.19280098 -0.10705532
##  [3,] -0.39232582  0.05181668 -0.11048102
##  [4,] -0.38204564  0.04738017 -0.13335408
##  [5,] -0.23230571  0.53031822 -0.01542079
##  [6,] -0.36212305  0.23605654 -0.11982646
##  [7,]  0.08843948  0.02126463 -0.79460449
##  [8,]  0.25005597  0.50229576 -0.08261299
##  [9,]  0.31110797  0.35947297 -0.21358474
## [10,]  0.02426425  0.46848762  0.46693016
## [11,] -0.33568563 -0.11526346  0.18532362
```

PC Plot



Numerous variables appear to make up PC 1, however, the ones the most make up PC 1 are x3 and x4. The variable that most makes up PC 2 is x10. All of the other variables aside from x7 have an impact on both PC 1 and PC2 - x2, x6, x1, and x11 are slightly more influenced by PC1 then PC2.