Discussion 6

Chris Conley

February 14, 2017

Exercise 8.28

Read in the data.

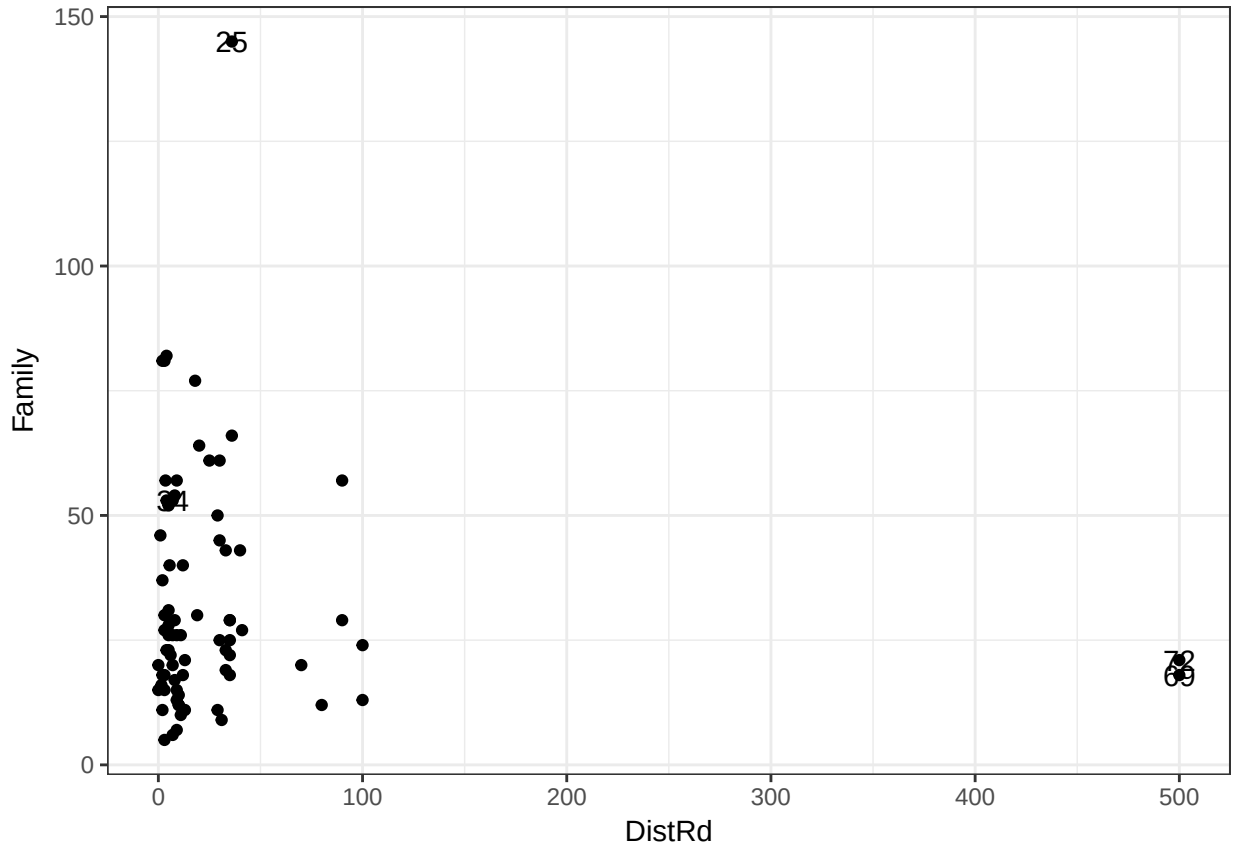
```
mali <- read.table(file = "~/../Box Sync/sta135-winter17/textbook_data/T8-7.DAT",
                   col.names = c("Family", "DistRd", "Cotton",
                                 "Maize", "Sorg", "Millet", "Bull",
                                 "Cattle", "Goats"))
kable(head(mali))
```

Family	DistRd	Cotton	Maize	Sorg	Millet	Bull	Cattle	Goats
12	80	1.5	1.0	3	0.25	2	0	1
54	8	6.0	4.0	0	1.00	6	32	5
11	13	0.5	1.0	0	0.00	0	0	0
21	13	2.0	2.5	1	0.00	1	0	5
61	30	3.0	5.0	0	0.00	4	21	0
20	70	0.0	2.0	3	0.00	2	0	3

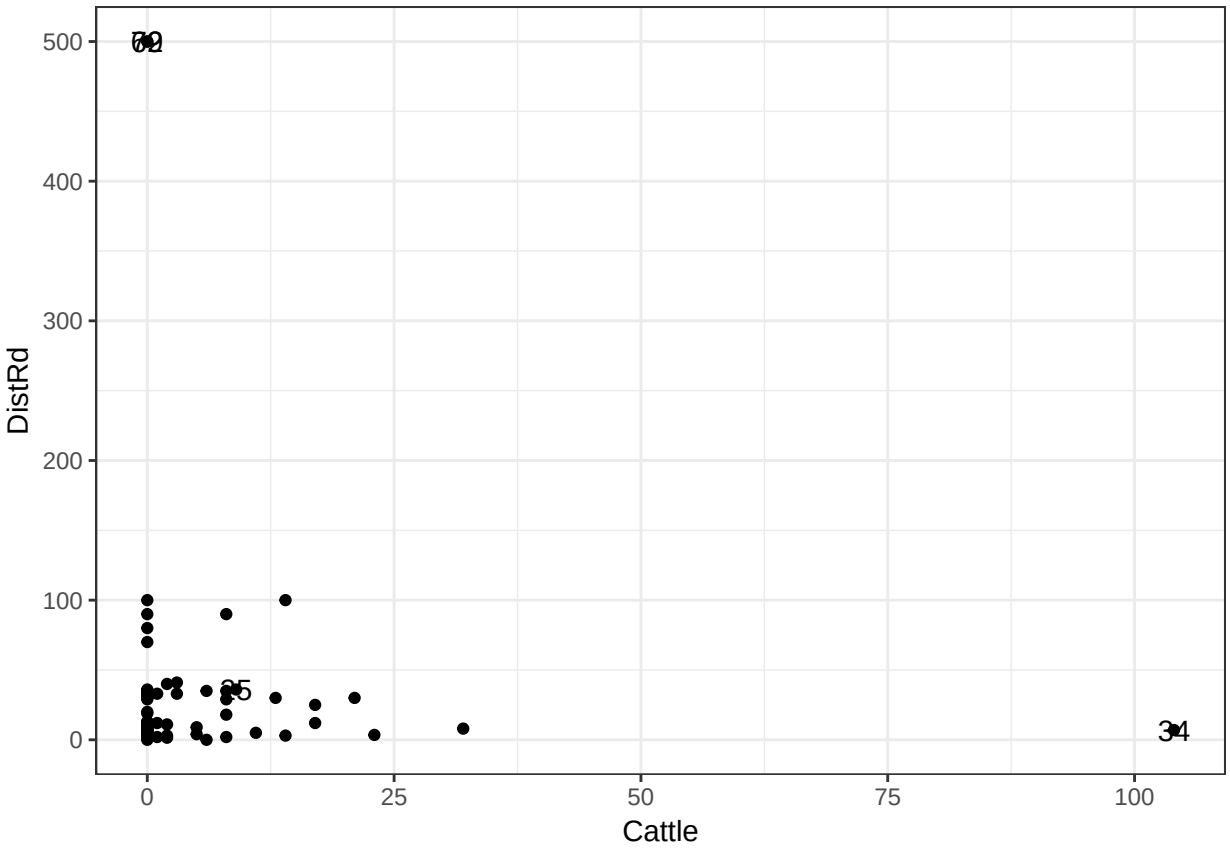
(a) Outlier removal

Remove the outliers through scatter plots.

```
#identify outliers through subsetting the data.
outliers <- subset(mali, DistRd > 400 | Family  >100 | Cattle > 80)
#obtain the indices of outlying values
outliers$index <- as.integer(row.names(outliers))
suppressPackageStartupMessages(library(ggplot2))
#scatter plots
ggplot(data = mali, aes(x = DistRd, y = Family)) + geom_point() +
   geom_text(data=outliers,
             aes(x = DistRd, y = Family, label = index)) +
  theme_bw()
```




```
ggplot(data = mali, aes(y = DistRd, x = Cattle)) + geom_point() +
   geom_text(data=outliers,
             aes(y = DistRd, x = Cattle, label = index)) +
  theme_bw()
```



Dropping observations 25, 34, 69, and 72 from the Mali Farm Data.

```
mali2 <- mali[-outliers[,"index"],]
```

(b)

The `stats::prcomp` function comes with base R and performs a numerically accurate PCA through the SVD of the data matrix instead of evaluating `eigen` on the covariance matrix. It is highly recommended to evaluate PCA under unit variance for all variables through the `scale.` parameter. We also elect to return the PC scores $\{y_{ij} : i, \ldots, n, j = 1, \ldots, p\}$ with the `retx` option.

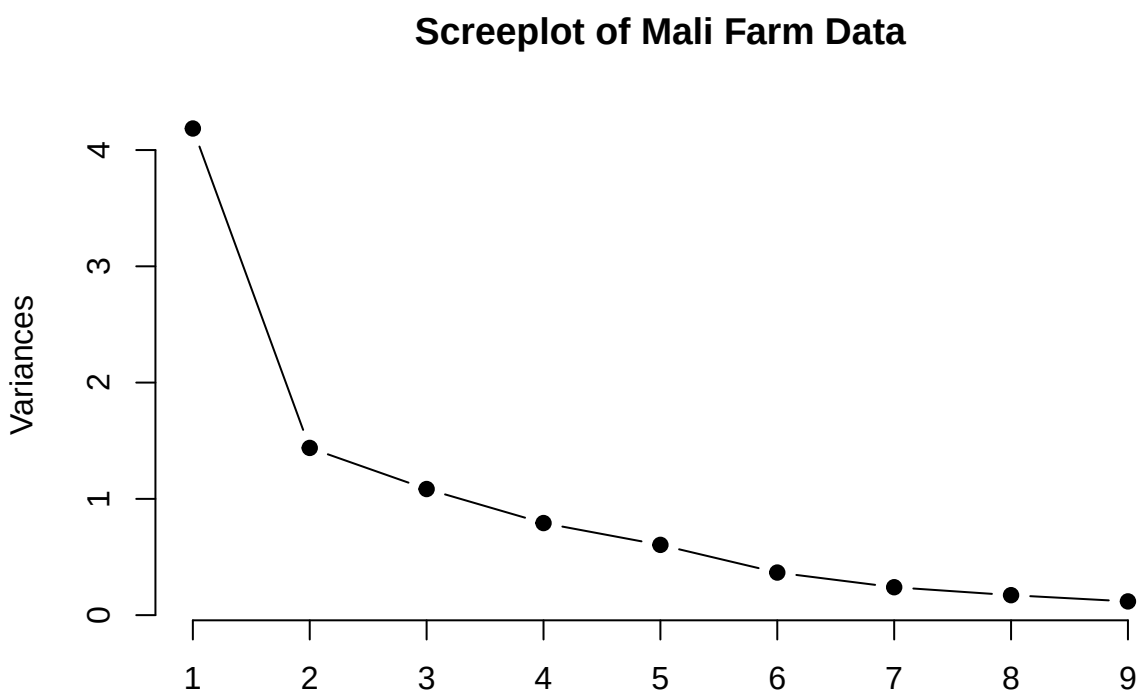
```
pca_mali <- prcomp(x = mali2, retx = TRUE, center = TRUE, scale. = TRUE)
summary(pca_mali)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.046 1.1992 1.0414 0.88984 0.77738 0.60509 0.48992
## Proportion of Variance 0.465 0.1598 0.1205 0.08798 0.06715 0.04068 0.02667
## Cumulative Proportion  0.465 0.6248 0.7453 0.83328 0.90043 0.94111 0.96778
##                           PC8     PC9
## Standard deviation     0.41452 0.34374
## Proportion of Variance 0.01909 0.01313
## Cumulative Proportion  0.98687 1.00000
```

Note that the standard deviations $\sqrt{\hat{\lambda}_i)}$ for $i = 1 \ldots, p$ are reported from this function. If we retain 5 principle components, approx. 90% of the variation is maintained. That is $\frac{1}{p}\sum_{i=1}^{5} \hat{\lambda}_i = .90043$. The scree

plot confirms that this is a suitable number of PCs to maintain due to limited total variance contribution offered by PCs 6-9.

```
screeplot(pca_mali, type = "lines", pch = 19,
          main = "Screeplot of Mali Farm Data")
```



(c)

The eigen vectors (or PC loadings) yield limited insight into the correlation structure of the data. Note that the assignment of a particular sign (+/-) to the PC loadings are arbitrary and are mostly helpful to identify contrasting variables. PC1 has roughly equal positive loadings for most variables related to farm size. Hence, PC1 is a suitable farm size index. PC2 has a contrast between Maize and DistRd vs. Sorghum and Millet. PC3 is basically a DistRd and Goats component due to their relatively large loadings. PC4 is another contrast between Goat and Cattle vs. DistRd and Millet. PC5 is a contrast between DistRd, Cotton, and Sorghum vs. Millet, Cattle, and Goat. PC2-5 are not easily interpreted and reveal the complex correlation structure of the Mali Farm data.

```
kable(pca_mali$rotation, digits = 3)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Family	0.434	-0.065	0.098	-0.171	0.011	-0.040	-0.797	-0.263	-0.249
DistRd	0.008	0.497	-0.569	-0.496	-0.378	0.187	0.021	-0.048	-0.065
Cotton	0.446	0.009	0.132	0.027	-0.219	-0.200	0.361	0.329	-0.675
Maize	0.352	0.353	0.388	-0.240	-0.079	-0.273	-0.024	0.363	0.574
Sorg	0.204	-0.604	-0.111	0.059	-0.645	0.246	-0.021	0.126	0.293
Millet	0.240	-0.415	-0.116	-0.616	0.527	0.181	0.241	0.077	0.048

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Bull	0.445	0.068	-0.030	0.146	-0.028	-0.134	0.396	-0.751	0.190
Cattle	0.355	0.284	0.014	0.373	0.218	0.759	-0.011	0.169	0.038
Goats	0.255	-0.049	-0.687	0.351	0.249	-0.402	-0.131	0.274	0.149

Note that we can obtain the correlation coefficients between each PC and the variables, but there is disagreement among statisticians of their overall value when it comes to interpretation of PCs. Basically, the correlation coeffificent ρ_{Y_i,X_k} measure the contribution of X_k to the component Y_i without accounting for the other $X_{j\neq k}$. Interpretation of the e_{ik} do account for all variables, which may be the best way to rank the importance of each variables' contribution to a component. See pgs 433-434 of your textbook for more discussion.

Exercise 8.10

```
stocks <- read.table(file = "~/../Box Sync/sta135-winter17/textbook_data/T8-4.DAT",
                     col.names = c("JPMorgan", "Citibank", "WellsFargo", "RoyalDutchSchell", "ExxonMobil"))
kable(head(stocks))
```

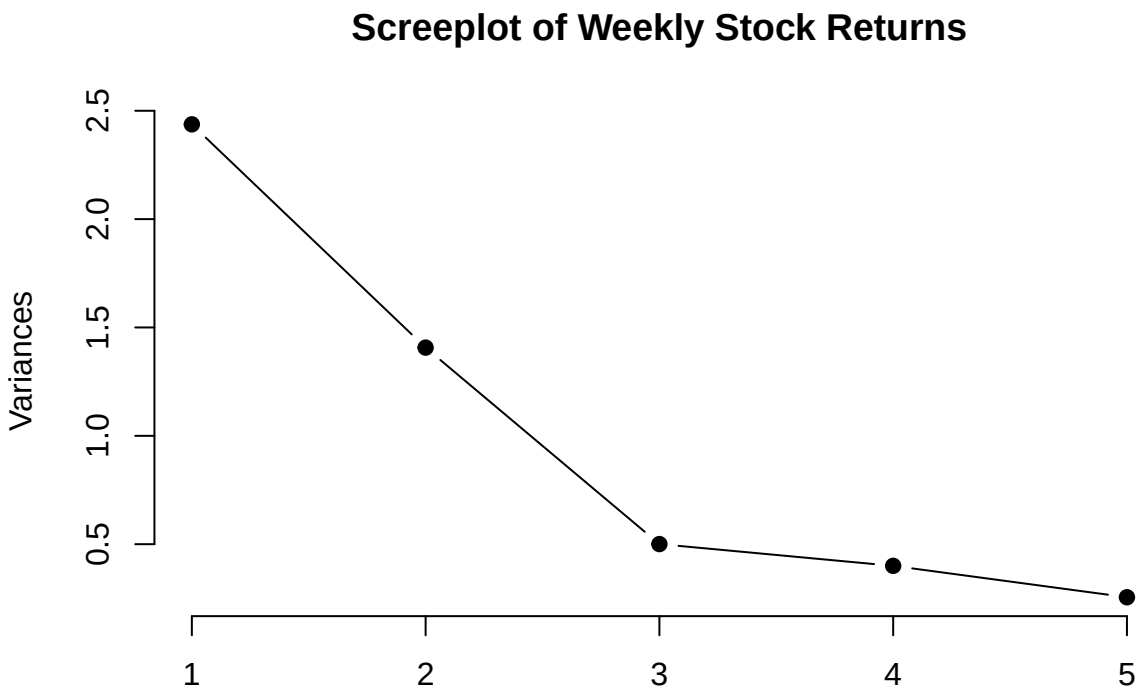
JPMorgan	Citibank	WellsFargo	RoyalDutchSchell	ExxonMobil
0.0130338	-0.0078431	-0.0031889	-0.0447693	0.0052151
0.0084862	0.0166886	-0.0062100	0.0119560	0.0134890
-0.0179153	-0.0086393	0.0100360	0.0000000	-0.0061428
0.0215589	-0.0034858	0.0174353	-0.0285917	-0.0069534
0.0108225	0.0037167	-0.0101345	0.0291900	0.0409751
0.0101713	-0.0121978	-0.0083768	0.0137083	0.0029895

Perform PCA under the correlation matrix in spite of them telling you otherwise in the exercise.

```
pca_stocks <- prcomp(x = stocks, retx = TRUE, center =  TRUE, scale. = TRUE)
summary(pca_stocks)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.5612 1.1862 0.7075 0.63248 0.50514
## Proportion of Variance 0.4874 0.2814 0.1001 0.08001 0.05103
## Cumulative Proportion  0.4874 0.7689 0.8690 0.94897 1.00000
```

```
screeplot(pca_stocks, type = "lines", pch = 19,
          main = "Screeplot of Weekly Stock Returns")
```



(b)

The elbow of the screeplot is at the thrird component. So we keep the first PCs.

$\frac{1}{p}\sum_{i=1}^{5}\hat{\lambda}_i = 0.8690$

(c)

A large sample $100(1-\alpha)\%$ percentile (formula 8-33) when m variances are being estimated is:

$$\frac{\hat{\lambda}_i}{1+z(\alpha/2m)\sqrt{2/n}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1-z(\alpha/2m)\sqrt{2/n}}$$

And in code it is:

```
ci_lambda <- function(hatlambda, alpha, m, n) {
  z <- qnorm(p = 1 - (alpha/(2*m)))
  c(hatlambda/(1 + z*sqrt(2/n)), hatlambda/(1 - z*sqrt(2/n)))
}
```

We evaluate the 95% simultaneous confidence intervals for the first three eigenvalues (variances of the PC1-3).

```
hatlambdas <- pca_stocks$sdev[1:3]^2
cil <- t(sapply(hatlambdas, ci_lambda, alpha = 0.05,
                m = length(hatlambdas), n = nrow(stocks)))
```

```r
colnames(cil) <- c("lower", "upper")
rownames(cil) <- paste0("lambda",1:3)
kable(cil)
```

	lower	upper
lambda1	1.8275990	3.6573340
lambda2	1.0550541	2.1113412
lambda3	0.3753115	0.7510616

(c)

PC1 represents a market index since all the loadings are roughly equally distributed across the variables. The PC scores $\hat{\mathbf{y}}_\mathbf{1} = (\mathbf{X} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1}\mathbf{1}^\mathbf{T}\mathbf{X})\hat{\mathbf{e}}_\mathbf{1}$ would give you a weekly index for these companies market's performance. Note the negative loadings in PC1 are totally arbitrary. PC2 represents an industry-specific component, independent of the market component. The negative loadings are on the oil companies and the positive loadings are on the banks. PC3 is beyond my interpretation. Most of the variation is due to general market forces and also industry-specific trends.

```r
kable(pca_stocks$rotation, digits = 3)
```

	PC1	PC2	PC3	PC4	PC5
JPMorgan	-0.469	0.368	-0.604	0.363	0.384
Citibank	-0.532	0.236	-0.136	-0.629	-0.496
WellsFargo	-0.465	0.315	0.772	0.289	0.071
RoyalDutchSchell	-0.387	-0.585	0.093	-0.381	0.595
ExxonMobil	-0.361	-0.606	-0.109	0.493	-0.498

PCA diagnostics

Often there can be categorical variables in addition to multivariate continuous variables. When classification is not the goal, PCA can be a useful dimension reduction technique to see how the categories (i.e. classes) cluster without any constraint coming from the classes. This is an especially helpful diagnostic in large-scale experiments when we want to confirm that samples have been appropriately labeled and cluster with their respective treatment. If we see problematic samples that appear to cluster with the wrong treatment, we ought to be suspicious. When classification is the goal, partial least squares regression (PLS) is a more appropriate dimension reduction technique because it maximizes the variation while accounting for the class label structure.

Let's take a look at the classic iris data set to illustrate this diagnostic application of PCA.

```r
data("iris")
kable(head(iris))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Perform PCA on the iris measurements. Notice if we don't use the correlation matrix, then the first PC dominates almost all of the variation.

```
pca_iris <- prcomp(iris[,1:4], center = T, scale. = F)
summary(pca_iris)
```

```
## Importance of components:
##                           PC1     PC2    PC3     PC4
## Standard deviation     2.0563 0.49262 0.2797 0.15439
## Proportion of Variance 0.9246 0.05307 0.0171 0.00521
## Cumulative Proportion  0.9246 0.97769 0.9948 1.00000
```

Whereas the PCA under the correlation matrix balances out the variation across more PCs.

```
pca_iris <- prcomp(iris[,1:4], center = T, scale. = T)
summary(pca_iris)
```

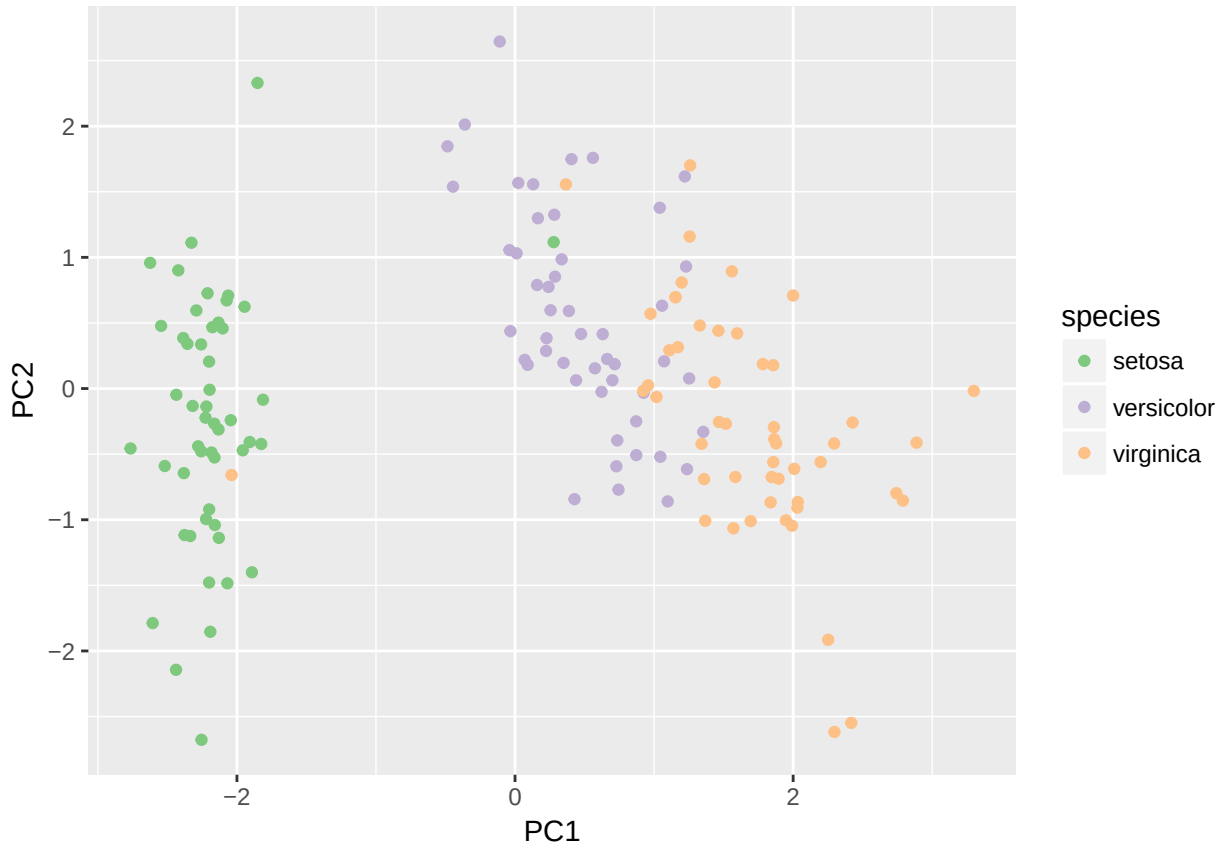
```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

The PCA shows that PC1-2 capture the majority of the variation. Let's look at the PC scores by species (class label). But first, let's misslabel a few of the class labels to see if we can pick them out in a PC score scatter plot.

```
ggiris <- cbind(as.data.frame(pca_iris$x), species = iris$Species)
ggiris$species[91] <- "setosa"
ggiris$species[37] <- "virginica"
```

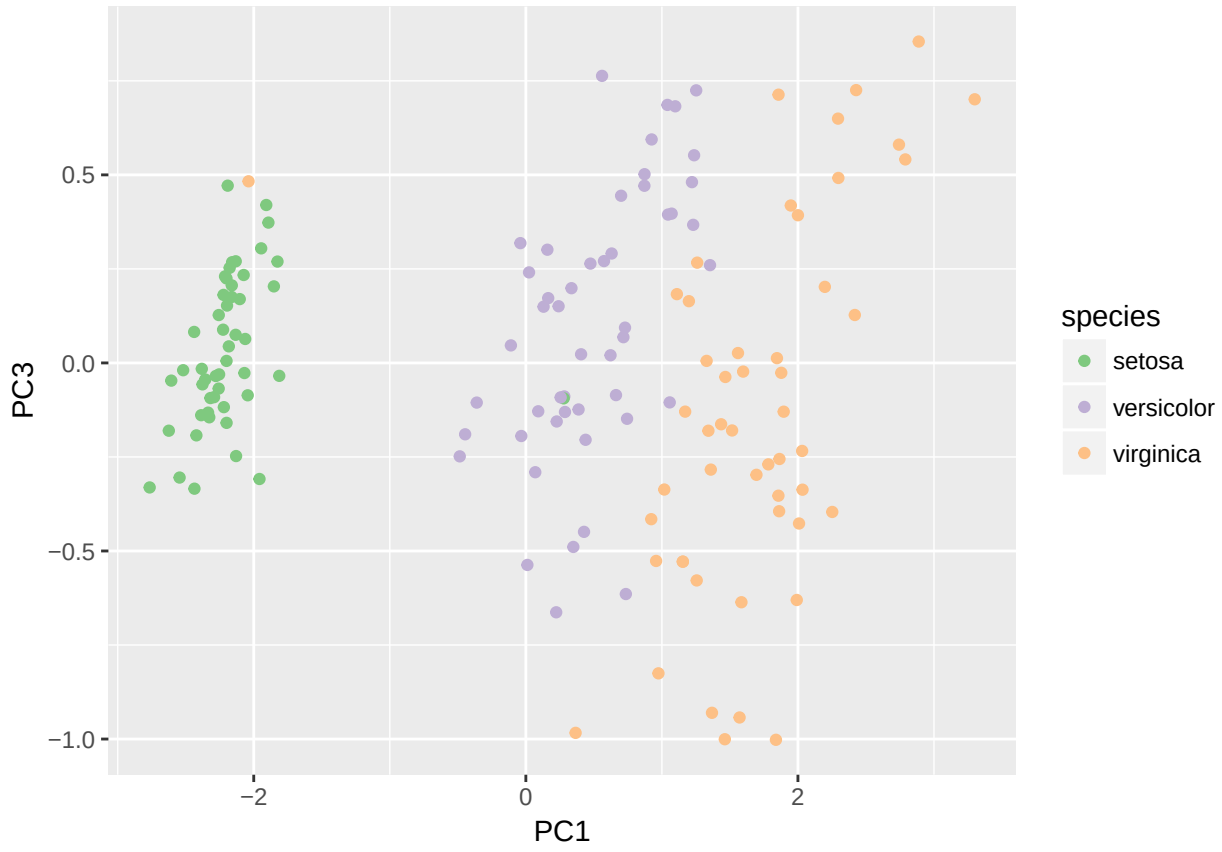
Sure enough, we see mislabel in the first scatter plot very clearly by virtue of the strange grouping of one verginica sample with the setosa samples and one setosa sample with the versicolor samples.

```
library(ggplot2)
ggplot(data = ggiris, aes(x=PC1, y=PC2, group = species, colour = species)) + geom_point() + scale_colo
```



Same thing in this projection of PC1 and PC3.

```
ggplot(data = ggiris, aes(x=PC1, y=PC3, group = species, colour = species)) + geom_point() + scale_colo
```


0.5
0.0
−0.5
−1.0
PC3
−2
0
2
PC1
species
setosa
versicolor
virginica