In-class worksheet 9

Feb 19, 2019

In this worksheet, we will use the libraries tidyverse, egg, and grid:

```
library(tidyverse)
theme_set(theme_bw(base_size=12)) # set default ggplot2 theme
library(egg) # required to arrange plots side-by-side
library(grid) # required to draw arrows
library(ggthemes) # for colorblind color scale
```

1. PCA of the iris data set

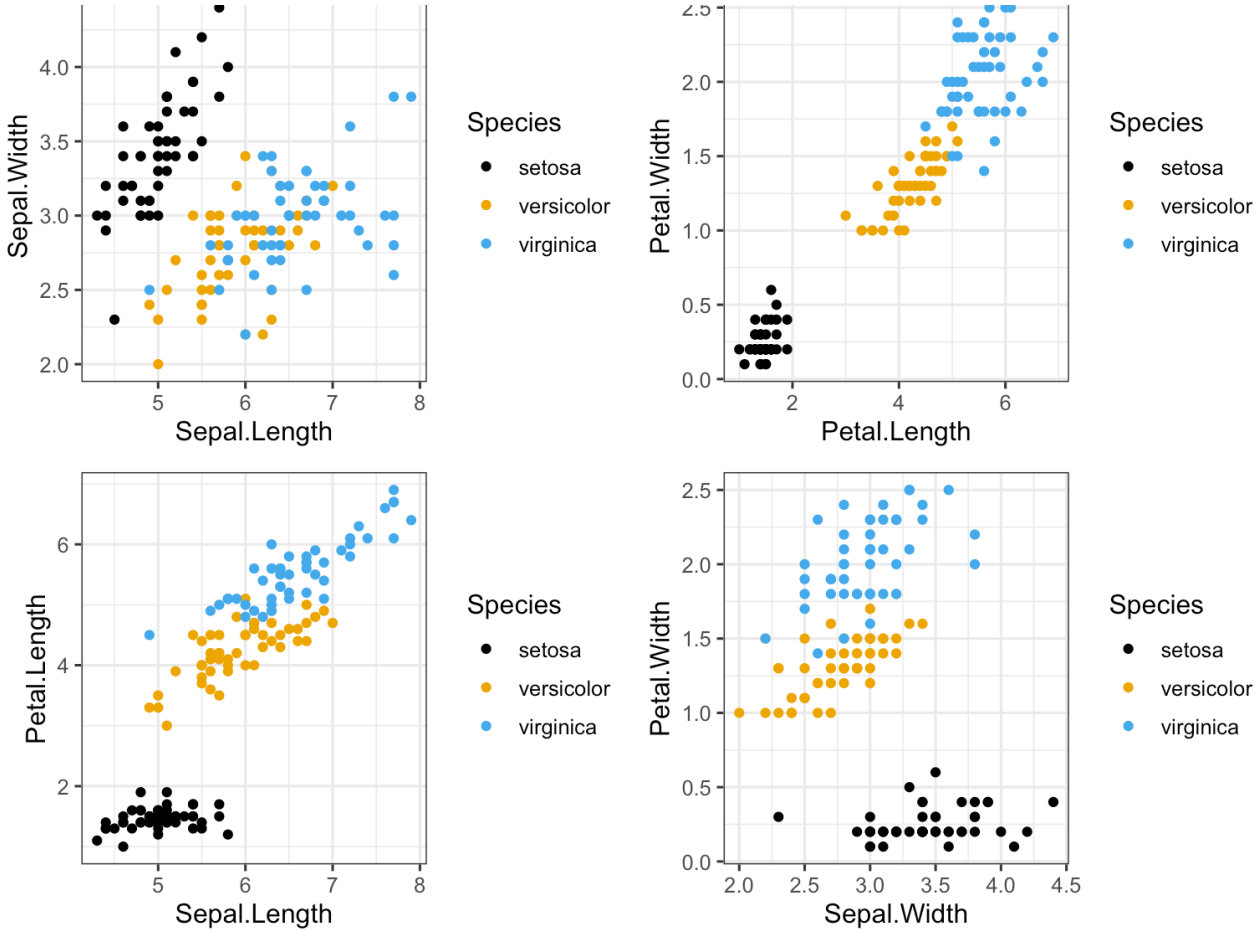
The `iris` dataset has four measurements per observational unit (iris plant):

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

If we want to find out which characteristics are most distinguishing between iris plants, we have to make many individual plots and hope we can see distinguishing patterns:

```
p1 <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point() +
  scale_color_colorblind()
p2 <- ggplot(iris, aes(x = Petal.Length, y = Petal.Width, color = Species)) +
  geom_point() +
  scale_color_colorblind()
p3 <- ggplot(iris, aes(x = Sepal.Length, y = Petal.Length, color = Species)) +
  geom_point() +
  scale_color_colorblind()
p4 <- ggplot(iris, aes(x = Sepal.Width, y = Petal.Width, color = Species)) +
  geom_point() +
  scale_color_colorblind()
ggarrange(p1, p2, p3, p4, ncol = 2) # arrange in a grid
```



In this particular case, it seems that petal length and petal width are most distinct for the three species. Principal Components Analysis (PCA) allows us to systematically discover such patterns, and it works also when there are many more variables than just four.

The basic steps in PCA are to (i) prepare a data frame that holds only the numerical columns of interest, (ii) scale the data to 0 mean and unit variance, and (iii) do the PCA with the function `prcomp()`:

```
iris %>%
  select(-Species) %>%   # remove Species column
  scale() %>%            # scale to 0 mean and unit variance
  prcomp() ->            # do PCA
  pca                    # store result as `pca`

# now display the results from the PCA analysis
pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.7083611 0.9560494 0.3830886 0.1439265
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2        PC3        PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

The main results from PCA are the standard deviations and the rotation matrix. We will talk about them below. First, however, let's plot the data in the principal components. Specifically, we will plot PC2 vs. PC1. The rotated data are available as `pca$x` :

```
head(pca$x)
```

```
##            PC1        PC2         PC3          PC4
## [1,] -2.257141 -0.4784238  0.12727962  0.024087508
## [2,] -2.074013  0.6718827  0.23382552  0.102662845
## [3,] -2.356335  0.3407664 -0.04405390  0.028282305
## [4,] -2.291707  0.5953999 -0.09098530 -0.065735340
## [5,] -2.381863 -0.6446757 -0.01568565 -0.035802870
## [6,] -2.068701 -1.4842053 -0.02687825  0.006586116
```

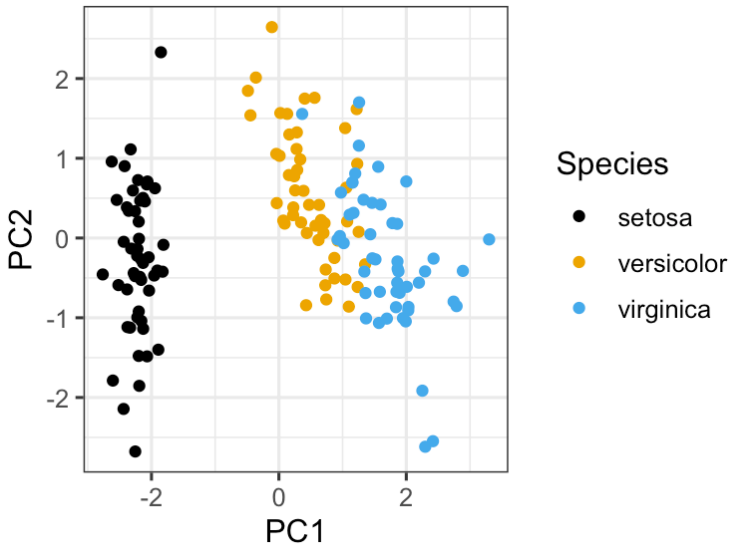
As we can see, these data don't tell us to which species which observation belongs. We have to add the species information back in:

```
# add species information back into PCA data
pca_data <- data.frame(pca$x, Species = iris$Species)
head(pca_data)
```

```
##         PC1        PC2         PC3          PC4 Species
## 1 -2.257141 -0.4784238  0.12727962  0.024087508  setosa
## 2 -2.074013  0.6718827  0.23382552  0.102662845  setosa
## 3 -2.356335  0.3407664 -0.04405390  0.028282305  setosa
## 4 -2.291707  0.5953999 -0.09098530 -0.065735340  setosa
## 5 -2.381863 -0.6446757 -0.01568565 -0.035802870  setosa
## 6 -2.068701 -1.4842053 -0.02687825  0.006586116  setosa
```

Now we can plot as usual:

```
ggplot(pca_data, aes(x = PC1, y = PC2, color = Species)) +
  geom_point() +
  scale_color_colorblind()
```



In the PC2 vs PC1 plot, versicolor and virginica are much better separated.

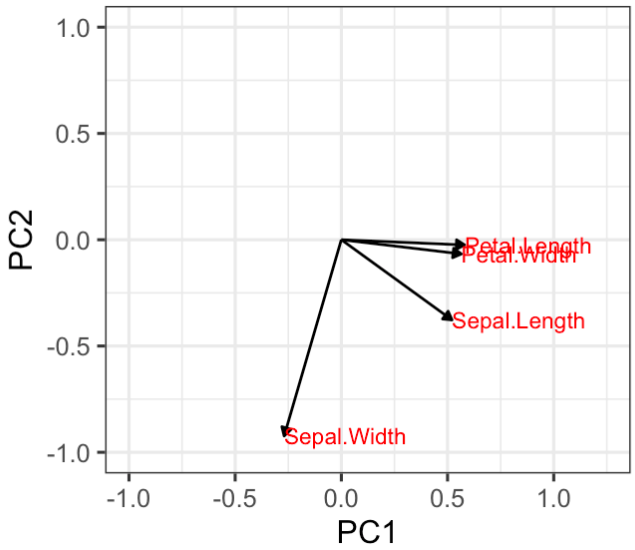
Next, let's look at the rotation matrx:

```
pca$rotation
```

```
##                    PC1         PC2        PC3        PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

It tells us how much each variable contributes to each principal component. For example, `Sepal.Width` contributes little to PC1 but makes up much of PC2. Often it is helpful to plot the rotation matrix as arrows. This can be done as follows:

```
# capture the rotation matrix in a data frame
rotation_data <- data.frame(
  pca$rotation,
  variable = row.names(pca$rotation)
)

# define a pleasing arrow style
arrow_style <- arrow(
  length = unit(0.05, "inches"),
  type = "closed"
)

# now plot, using geom_segment() for arrows and geom_text() for labels
ggplot(rotation_data) +
  geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style)
+
  geom_text(aes(x = PC1, y = PC2, label = variable), hjust = 0, size = 3, color
= "red") +
  xlim(-1., 1.25) +
  ylim(-1., 1.) +
  coord_fixed() # fix aspect ratio to 1:1
```



We can now see clearly that `Petal.Length`, `Petal.Width`, and `Sepal.Length` all contribute to PC1, and `Sepal.Width` dominates PC2.

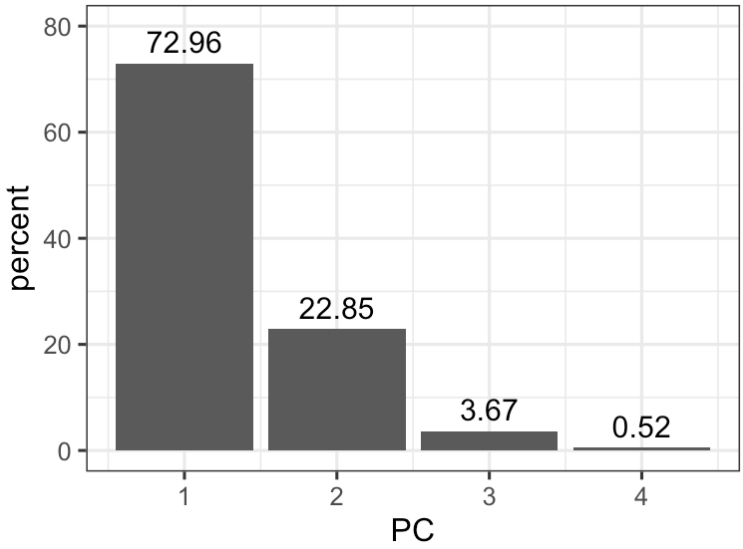
Finally, we want to look at the percent variance explained. The `prcomp()` function gives us standard deviations (stored in `pca$sdev`). To convert them into percent variance explained, we square them and then divide by the sum over all squared standard deviations:

```
percent <- 100*pca$sdev^2 / sum(pca$sdev^2)
percent
```

```
## [1] 72.9624454 22.8507618  3.6689219  0.5178709
```

The first component explains 73% of the variance, the second 23%, the third 4% and the last 0.5%. We can visualize these results nicely in a bar plot:

```
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 80)
```



2. Now do it yourself: The biopsy data set

The biopsy data set contains data from 683 patients who had a breast biopsy performed. Each tissue sample was scored according to 9 different characteristics, each on a scale from 1 to 10. Also, for each patient the final outcome (benign/malignant) was known:

```
biopsy <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/biopsy.csv")
head(biopsy)
```

```
##   clump_thickness uniform_cell_size uniform_cell_shape marg_adhesion
## 1               5                 1                 1             1
## 2               5                 4                 4             5
## 3               3                 1                 1             1
## 4               6                 8                 8             1
## 5               4                 1                 1             3
## 6               8                10                10             8
##   epithelial_cell_size bare_nuclei bland_chromatin normal_nucleoli mitoses
## 1                    2           1               3               1       1
## 2                    7          10               3               2       1
## 3                    2           2               3               1       1
## 4                    3           4               3               7       1
## 5                    2           1               3               1       1
## 6                    7          10               9               7       1
##     outcome
## 1    benign
## 2    benign
## 3    benign
## 4    benign
## 5    benign
## 6 malignant
```

Use PCA to predict the outcome (benign/malignant) from the scored characteristics.

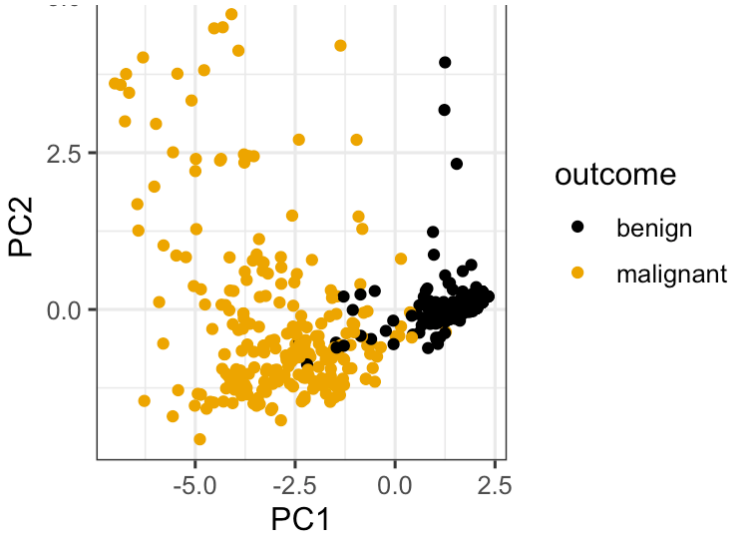
First we do the PCA:

```
biopsy %>%
  select(-outcome) %>%    # remove outcome column
  scale() %>%             # scale to 0 mean and unit variance
  prcomp() ->             # do PCA
  pca                     # store result as `pca`

# now display the results from the PCA analysis
pca
```

```
## Standard deviations (1, .., p=9):
## [1] 2.4288885 0.8808785 0.7343380 0.6779583 0.6166651 0.5494328 0.5425889
## [8] 0.5106230 0.2972932
##
## Rotation (n x k) = (9 x 9):
##                               PC1         PC2          PC3         PC4
## clump_thickness      -0.3020626 -0.14080053  0.866372452 -0.10782844
## uniform_cell_size    -0.3807930 -0.04664031 -0.019937801  0.20425540
## uniform_cell_shape   -0.3775825 -0.08242247  0.033510871  0.17586560
## marg_adhesion        -0.3327236 -0.05209438 -0.412647341 -0.49317257
## epithelial_cell_size -0.3362340  0.16440439 -0.087742529  0.42738358
## bare_nuclei          -0.3350675 -0.26126062  0.000691478 -0.49861767
## bland_chromatin      -0.3457474 -0.22807676 -0.213071845 -0.01304734
## normal_nucleoli      -0.3355914  0.03396582 -0.134248356  0.41711347
## mitoses              -0.2302064  0.90555729  0.080492170 -0.25898781
##                               PC5         PC6          PC7         PC8
## clump_thickness       0.08032124 -0.24251752 -0.008515668  0.24770729
## uniform_cell_size    -0.14565287 -0.13903168 -0.205434260 -0.43629981
## uniform_cell_shape   -0.10839155 -0.07452713 -0.127209198 -0.58272674
## marg_adhesion        -0.01956898 -0.65462877  0.123830400  0.16343403
## epithelial_cell_size -0.63669325  0.06930891  0.211018210  0.45866910
## bare_nuclei          -0.12477294  0.60922054  0.402790095 -0.12665288
## bland_chromatin       0.22766572  0.29889733 -0.700417365  0.38371888
## normal_nucleoli       0.69021015  0.02151820  0.459782742  0.07401187
## mitoses               0.10504168  0.14834515 -0.132116994 -0.05353693
##                               PC9
## clump_thickness      -0.002747438
## uniform_cell_size    -0.733210938
## uniform_cell_shape    0.667480798
## marg_adhesion         0.046019211
## epithelial_cell_size  0.066890623
## bare_nuclei          -0.076510293
## bland_chromatin       0.062241047
## normal_nucleoli      -0.022078692
## mitoses               0.007496101
```

Now we plot PC2 vs PC1, colored by outcome:

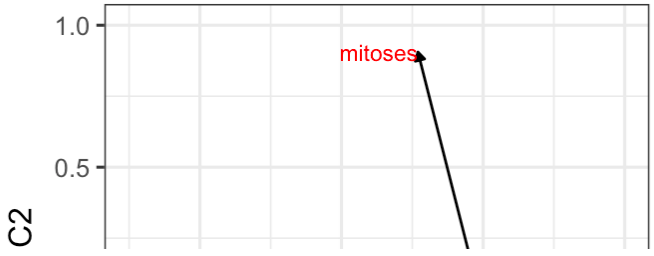
```
# add outcome information back into PCA data
pca_data <- data.frame(pca$x, outcome = biopsy$outcome)

# and plot
ggplot(pca_data, aes(x = PC1, y = PC2, color = outcome)) +
  geom_point() +
  scale_color_colorblind()
```

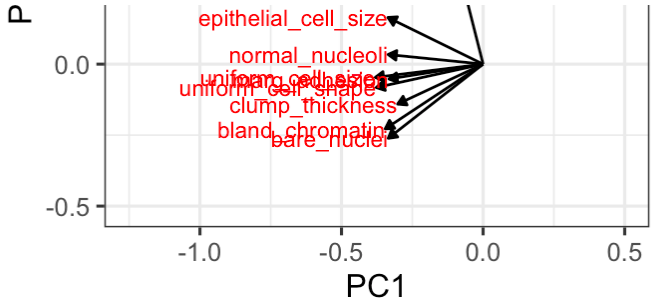


PC1 seems to separate benign from malignant.

Now we'll take a closer look at the rotation matrix:

```
# capture the rotation matrix in a data frame
rotation_data <- data.frame(
  pca$rotation,
  variable = row.names(pca$rotation)
)

# define a pleasing arrow style
arrow_style <- arrow(
  length = unit(0.05, "inches"),
  type = "closed"
)

# now plot, using geom_segment() for arrows and geom_text for labels
ggplot(rotation_data) +
  geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style)
+
  geom_text(aes(x = PC1, y = PC2, label = variable), hjust = 1, size = 3, color
= "red") +
  xlim(-1.25, .5) +
  ylim(-.5, 1) +
  coord_fixed() # fix aspect ratio to 1:1
```





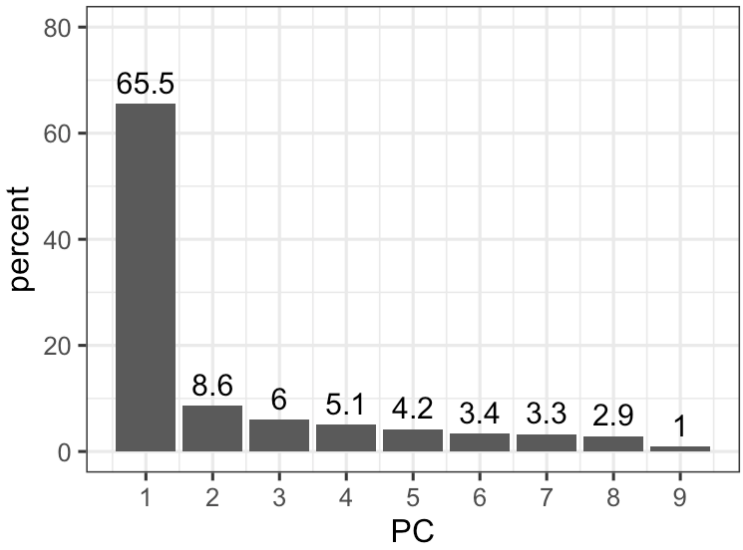
Nearly all indicators point into the direction of PC1. PC2, which doesn't predict malignancy, consists mostly of "mitoses."

Finally, the percent variance explained:

```
percent <- 100*pca$sdev^2 / sum(pca$sdev^2)
percent
```

```
## [1] 65.5499928  8.6216321  5.9916916  5.1069717  4.2252870  3.3541828
## [7]  3.2711413  2.8970651  0.9820358
```

```
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  geom_text(aes(label = round(percent, 1)), size = 4, vjust = -.5) +
  ylim(0, 80) +
  scale_x_continuous(breaks = 1:9) # make sure each PC gets an axis tick
```



The first PC carries the vast majority of variance explained.

3. If this was easy

The pottery data set contains the chemical composition of ancient pottery found at four sites in Great Britain:

```
pottery <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/pottery.csv")
head(pottery)
```

```
##        Site   Al   Fe   Mg   Ca   Na
## 1 Llanedyrn 14.4 7.00 4.30 0.15 0.51
## 2 Llanedyrn 13.8 7.08 3.43 0.12 0.17
## 3 Llanedyrn 14.6 7.09 3.88 0.13 0.20
## 4 Llanedyrn 11.5 6.37 5.64 0.16 0.14
## 5 Llanedyrn 13.8 7.06 5.34 0.20 0.20
## 6 Llanedyrn 10.9 6.26 3.47 0.17 0.22
```

Use PCA to see whether pottery found at different sites has different chemical composition.

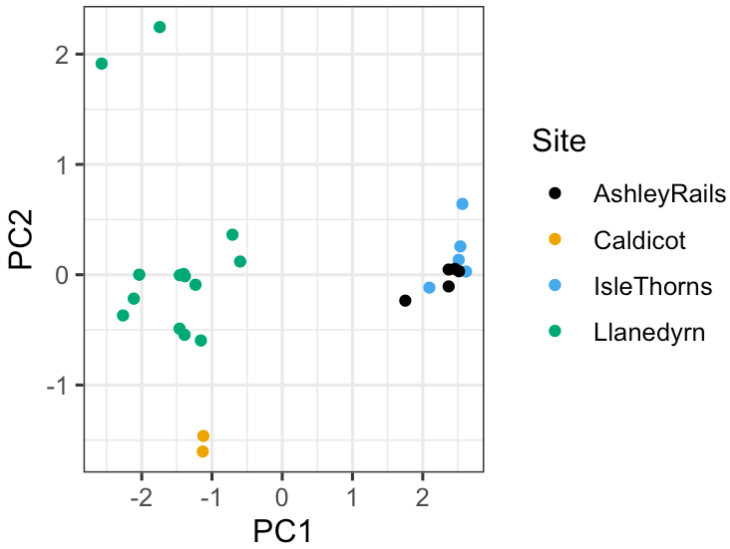
First we do the PCA:

```
pottery %>%
  select(-Site) %>%    # remove Site column
  scale() %>%          # scale to 0 mean and unit variance
  prcomp() ->          # do PCA
  pca                  # store result as `pca`

# now display the results from the PCA analysis
pca
```

```
## Standard deviations (1, .., p=5):
## [1] 1.9692123 0.7802627 0.4941581 0.4300110 0.2903301
##
## Rotation (n x k) = (5 x 5):
##           PC1         PC2          PC3        PC4         PC5
## Al  0.4454340  0.35652382 -0.694985231  0.4360290 -0.03678997
## Fe -0.4781318  0.04117517  0.157338238  0.6584787 -0.55798300
## Mg -0.4865413 -0.04960723 -0.158559572  0.3509063  0.78264954
## Ca -0.4490540 -0.34414647 -0.683443409 -0.3722000 -0.27255454
## Na -0.3668876  0.86619726 -0.002043385 -0.3385505 -0.02179880
```

Now we plot PC2 vs PC1, colored by site:

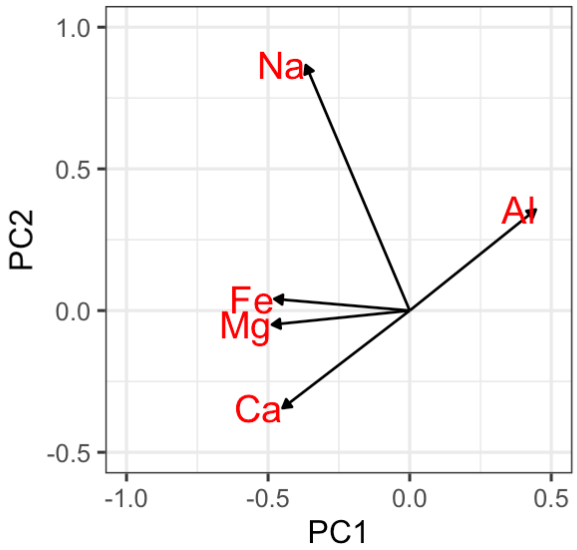
```
# add outcome information back into PCA data
pca_data <- data.frame(pca$x, Site = pottery$Site)

# and plot
ggplot(pca_data, aes(x = PC1, y = PC2, color = Site)) +
  geom_point() +
  scale_color_colorblind()
```



Pottery from Llanedyrn is clearly distinct from the rest. Pottery from the other locations does not separate.

Now we'll take a closer look at the rotation matrix:

```
# capture the rotation matrix in a data frame
rotation_data <- data.frame(
  pca$rotation,
  variable = row.names(pca$rotation)
)

# define a pleasing arrow style
arrow_style <- arrow(
  length = unit(0.05, "inches"),
  type = "closed"
)

# now plot, using geom_segment() for arrows and geom_text for labels
ggplot(rotation_data) +
  geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style)
+
  geom_text(aes(x = PC1, y = PC2, label = variable), hjust = 1, size = 5, color
= "red") +
  xlim(-1., .5) +
  ylim(-.5, 1) +
  coord_fixed() # fix aspect ratio to 1:1
```



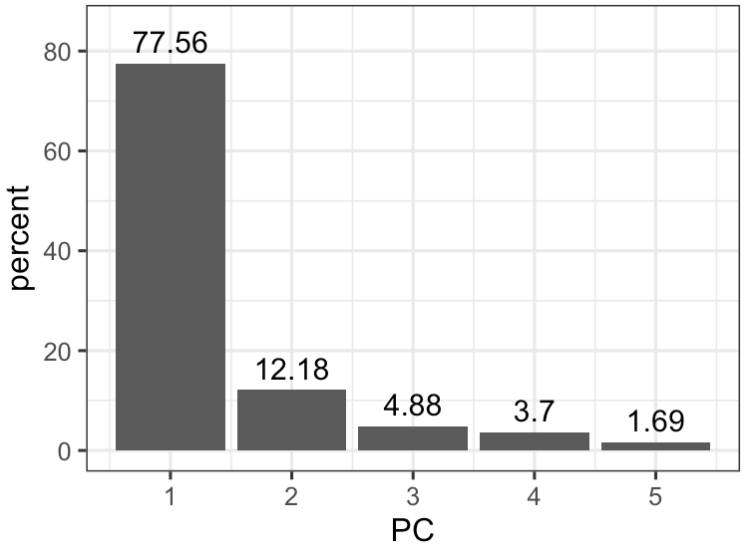
Fe and Mg contribute only to PC1, and the other three elements contribute partially to both PCs.

Finally, the percent variance explained:

```
percent <- 100*pca$sdev^2 / sum(pca$sdev^2)
percent
```

```
## [1] 77.555940 12.176197  4.883844  3.698188  1.685831
```

```
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -.5) +
  ylim(0, 85) +
  scale_x_continuous(breaks = 1:5) # make sure each PC gets an axis tick
```



Again, the first PC carries the vast majority of variance explained.