# P8139 HW3

Brian Jo Hsuan Lee

3/28/2022

Load packages

```
library(tidyverse)
library(knitr)
```

Import data

```
data = read.table(file="PCA_data.csv", header=TRUE, row.names=1, sep=",")
```

PCA fitting and exploration

```
data_pca = prcomp(data, retx=TRUE, center=TRUE, scale=TRUE)
scores = data_pca$x
```

a) **Calculate score using only `loadings` and `data`, and compare results with `scores`**

```
# extract loadings from the data_pca object
loadings = data_pca$rotation

# compute mean math, biology and chemistry grades
means =
  data %>%
  summarize(math = mean(math),
            bio = mean(bio),
            chem = mean(chem))

# compute math, biology and chemistry grade variances
vars =
  data %>%
  summarize(math = var(math),
            bio = var(bio),
            chem = var(chem))

# compute the standardized math, biology and chemistry grades
# by subtracting them with their means and dividing by
# their variances
std_data =
  data %>%
  mutate(
    math = (math - means$math)/sqrt(vars$math),
    bio = (bio - means$bio)/sqrt(vars$bio),
    chem = (chem - means$chem)/sqrt(vars$chem)
  ) %>%
  as.matrix()
```

```r
# multiply the matrices to obtain the PCs values for each observation
calc_scores = std_data %*% loadings
```

b) **Percent Variance Explained by each PC**

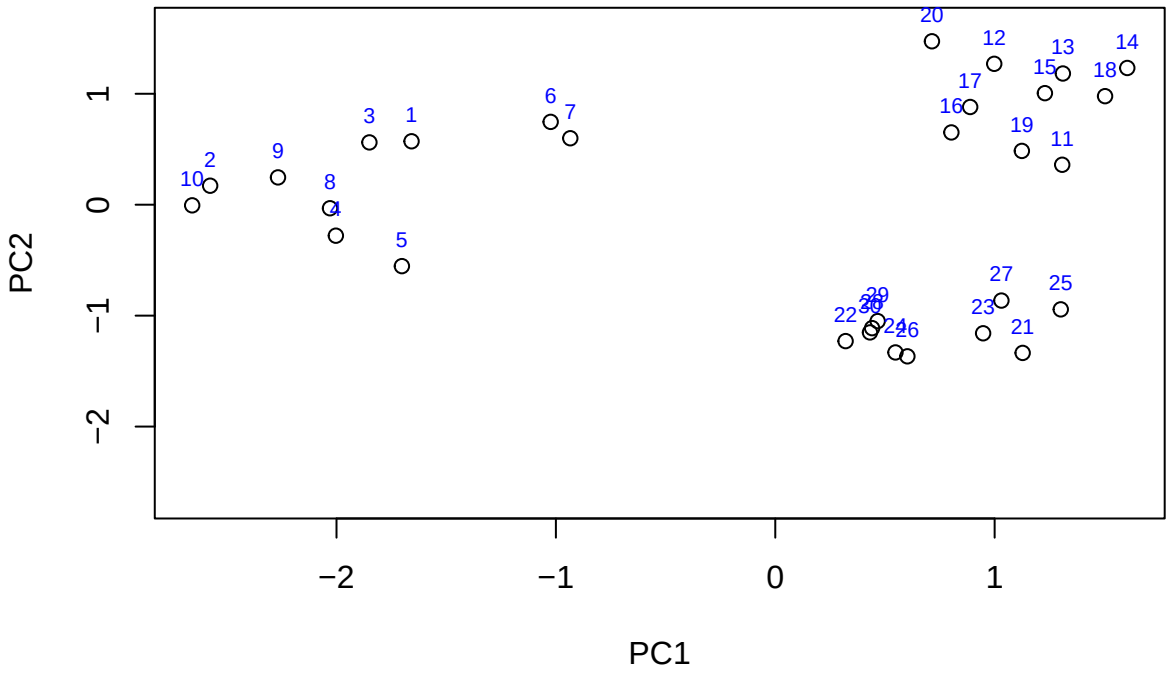
```r
# calculate variance for each PC, get total variance, and divide each by total variance
# to acquire percentage variance
pc_vars =
  data.frame(scores) %>%
  summarize(PC1 = var(PC1),
            PC2 = var(PC2),
            PC3 = var(PC3)) %>%
  mutate(
    tot = sum(PC1, PC2, PC3)
  ) %>%
  summarize(PC1 = PC1/tot,
            PC2 = PC2/tot,
            PC3 = PC3/tot)

# display table
kable(pc_vars, "simple")
```

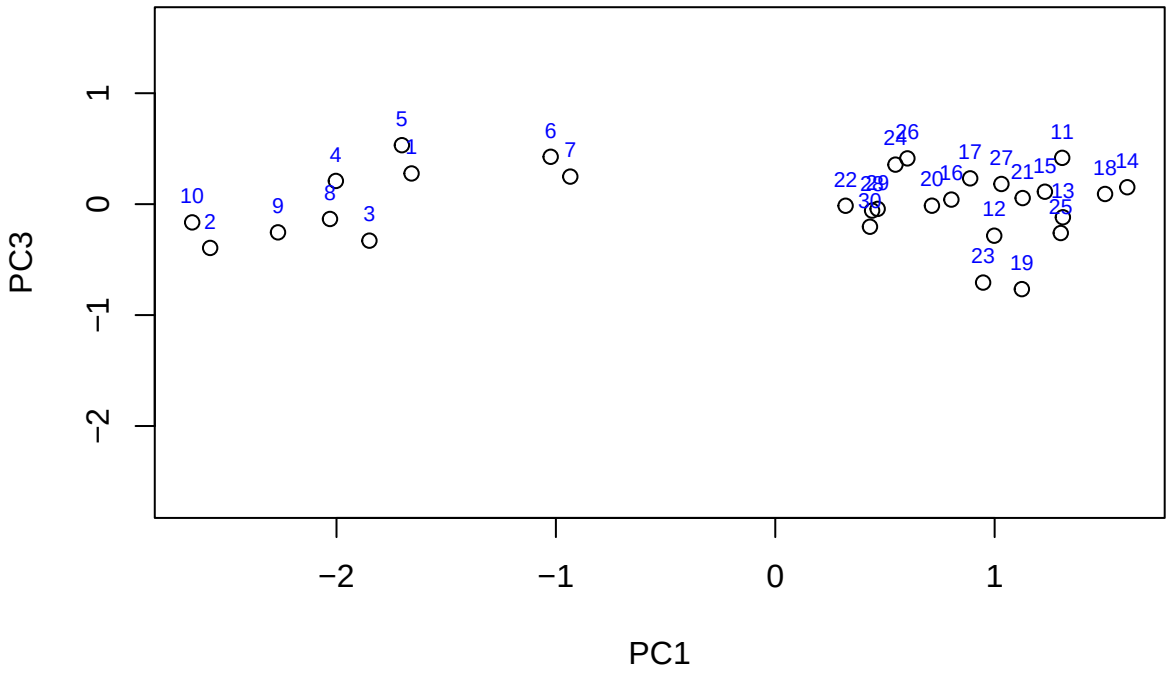
| PC1 | PC2 | PC3 |
|---:|---:|---:|
| 0.6692802 | 0.2971136 | 0.0336061 |

c) **PCA Plots**

```r
# i) PC1 vs PC2
plot(scores[,1:2],
     xlim=c(min(scores[,1:2]),
            max(scores[,1:2])),
     ylim=c(min(scores[,1:2]),
            max(scores[,1:2])))
text(scores[,1], scores[,2], rownames(scores), col="blue", cex=0.7, pos=3)
```




```
# ii) PC1 vs PC3
plot(scores[,c(1,3)],
     xlim=c(min(scores[,c(1,3)]),
            max(scores[,c(1,3)])),
     ylim=c(min(scores[,c(1,3)]),
            max(scores[,c(1,3)])))
text(scores[,1], scores[,3], rownames(scores), col="blue", cex=0.7, pos=3)
```




```
# iii) PC2 vs PC3
plot(scores[,2:3],
     xlim=c(min(scores[,2:3]),
```

```
            max(scores[,2:3])),
     ylim=c(min(scores[,2:3]),
            max(scores[,2:3])))
text(scores[,2], scores[,3], rownames(scores), col="blue", cex=0.7, pos=3)
```