lab9

YILIN LI

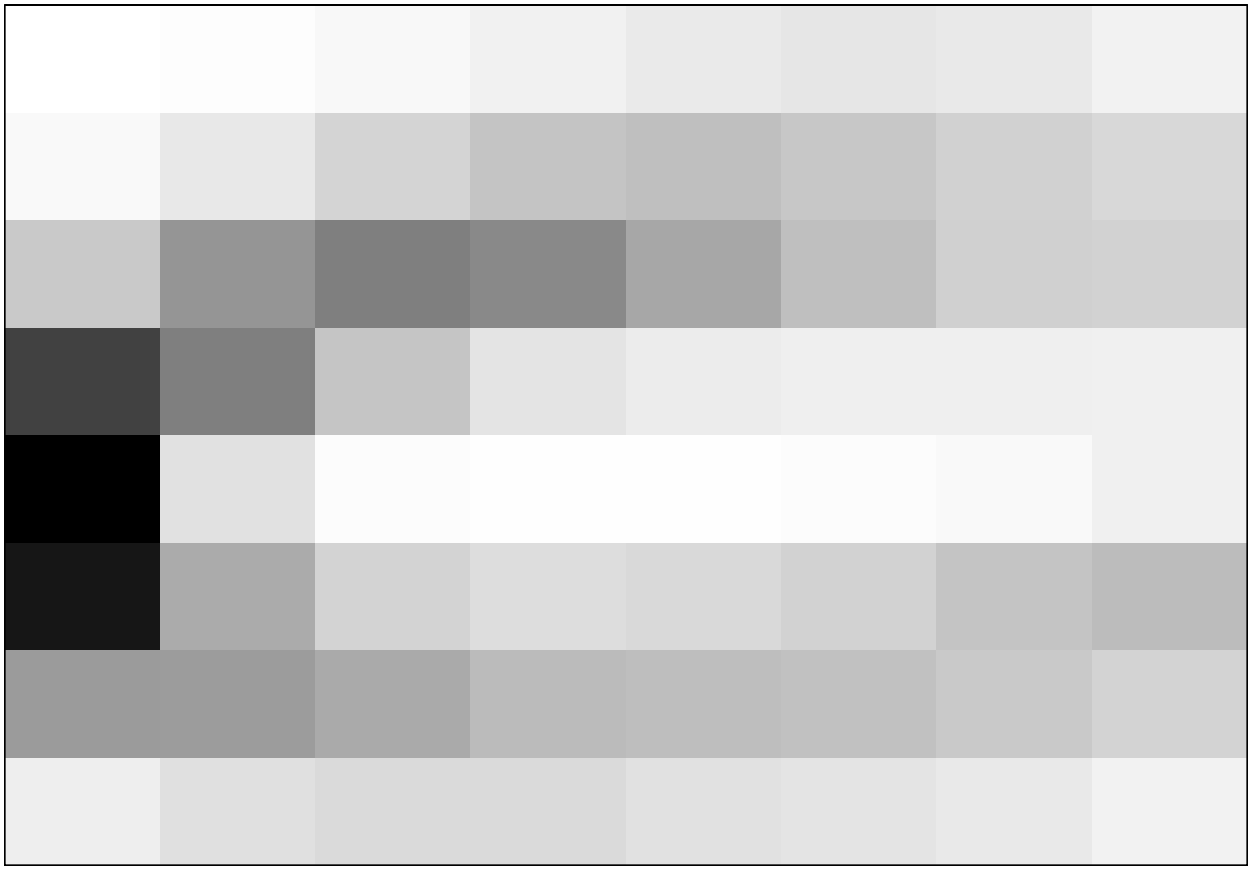
2019/11/20

```r
d <- read.csv("https://raw.githubusercontent.com/stat-learning/course-materials/master/data/handwritten

plot_letter <- function(x, hasletter = TRUE) {
  if(hasletter) {
    a <- as.numeric(x[, -1])
  }else{a <- as.numeric(x)}
  m <- matrix(a, nrow = 8, byrow = TRUE)
  m <- t(apply(m, 2, rev)) # rotate matrix
  par(mar = rep(0, 4))
  image(m, axes = FALSE, col = rev(grey(seq(0, 1, length = 256)))) #this should be a divergent palette
  box()
}

pc_grid <- function(pca, data) {
  d <- data
  grid_points <- as.matrix(expand.grid(seq(-1.5, 1.5, length.out = 5),
                                       seq(-1.5, 1.5, length.out = 5)))
  pc_points <- pca$x[, 1:2]
  nearest_ind <- rep(NA, nrow(grid_points))
  for(i in 1:nrow(grid_points)) {
    gp <- matrix(rep(grid_points[i, ], nrow(pc_points)),
                 ncol = 2, byrow = TRUE)
    nearest_ind[i] <- which.min(rowSums((pc_points - gp)^2))
  }
  nearest_grid <- data.frame(d[nearest_ind, ])
  par(mfrow = c(5, 5))
  regrid <- c(21:25, 16:20, 11:15, 6:10, 1:5)
  for(i in regrid) {
    plot_letter(nearest_grid[i, ])
  }
}

d_c <- d[d["letter"] == "c" ,] # include only letter "c"
mean_c <- colMeans(d_c[,-1])   # not include the first column "letter"
plot_letter(mean_c, hasletter = F)
```



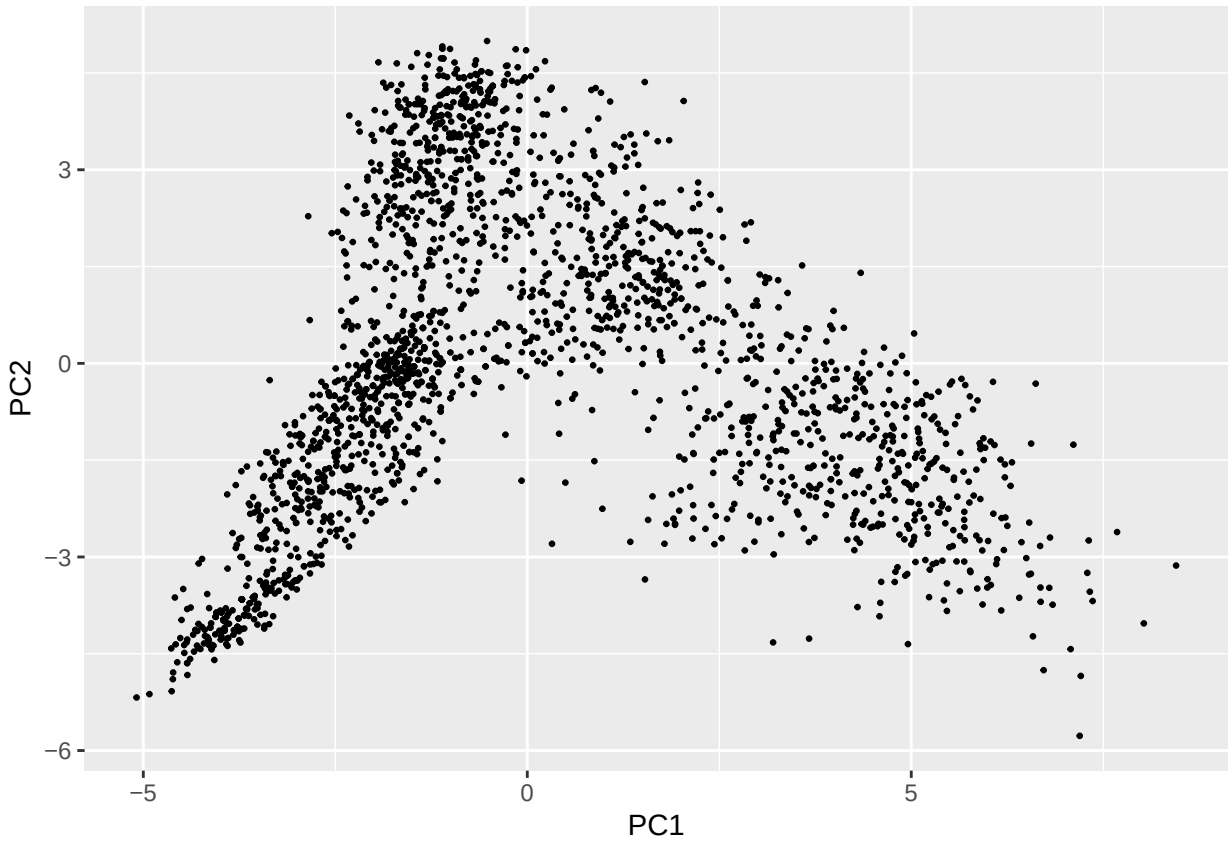
Q1

The columns are pixels, 0 means black, 1 means white. Each row represents an image of a specific letter.

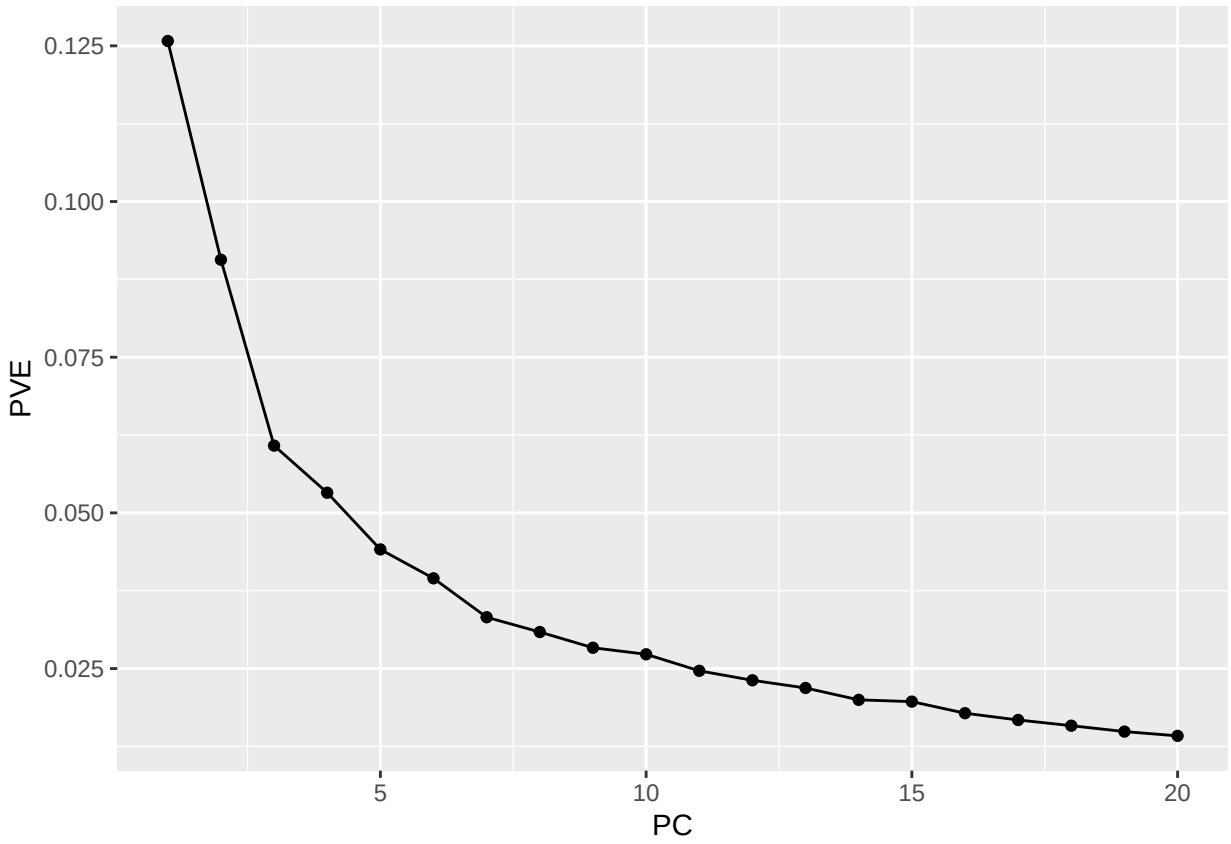
Q4

The mean form of "c" is shown above.

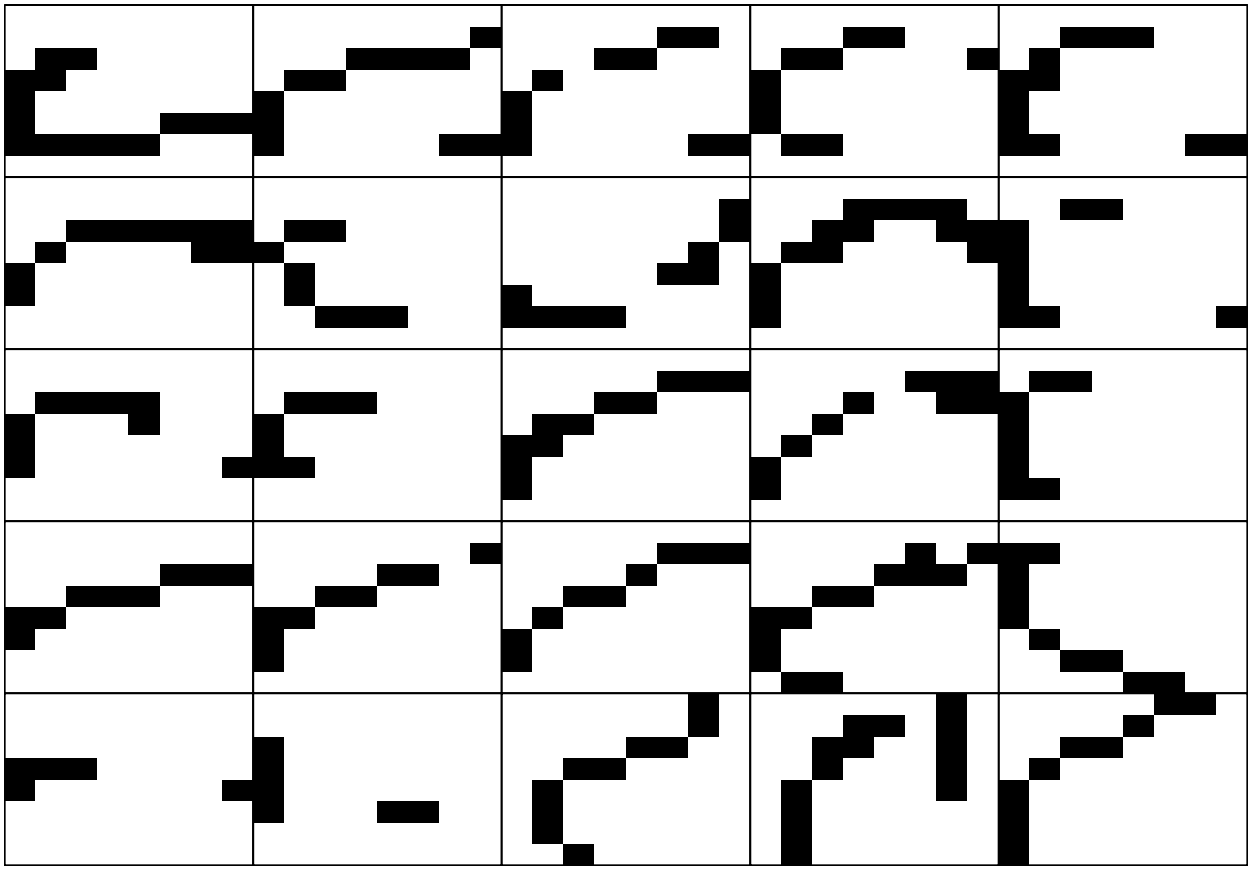
```
library(ggplot2)
pca1 <- prcomp(d_c[,-1], scale = T)
d <- as.data.frame(pca1$x)
p1 <- ggplot(d, aes(x = PC1, y = PC2)) +
  geom_point(size = 0.5)
p1
```




```
d <- data.frame(PC = 1:20,
                PVE = (pca1$sdev^2 /
                  sum(pca1$sdev^2))[1:20])
ggplot(d, aes(x = PC, y = PVE)) +
  geom_line() +
  geom_point()
```




```
pc_grid(pca1, d_c)
```



It seems like pc1 and pc2 represent the general shape of the letter “c”