```
> options(digits = 3)
> library(matrixStats)
> library(tidyverse)
> library(caret)
> library(dslabs)
> data(brca)
>
>
> dim(brca$x)[1]
[1] 569
> dim(brca$x)[2]
[1] 30
> mean(brca$y == "M")
[1] 0.373
> which.max(colMeans(brca$x))
area_worst
        24
> which.min(colSds(brca$x))
[1] 20
>
>
>
> x_centered <- sweep(brca$x, 2, colMeans(brca$x))
> x_scaled <- sweep(x_centered, 2, colSds(brca$x), FUN = "/")
> sd(x_scaled[,1])
[1] 1
> median(x_scaled[,1])
[1] -0.215
>
>
>
> d_samples <- dist(x_scaled)
> dist_BtoB <- as.matrix(d_samples)[1, brca$y == "B"]
> mean(dist_BtoB[2:length(dist_BtoB)])
[1] 4.41
> dist_BtoM <- as.matrix(d_samples)[1, brca$y == "M"]
> mean(dist_BtoM)
[1] 7.12
>
>
>
> d_features <- dist(t(x_scaled))
> heatmap(as.matrix(d_features), labRow = NA, labCol = NA)
>
>
>
> h <- hclust(d_features)
> groups <- cutree(h, k = 5)
> split(names(groups), groups)
$`1`
 [1] "radius_mean"       "perimeter_mean"    "area_mean"
 [4] "concavity_mean"    "concave_pts_mean"  "radius_se"
 [7] "perimeter_se"      "area_se"           "radius_worst"
[10] "perimeter_worst"   "area_worst"        "concave_pts_worst"

$`2`
[1] "texture_mean"  "texture_worst"

$`3`
[1] "smoothness_mean"   "compactness_mean"  "symmetry_mean"
[4] "fractal_dim_mean"  "smoothness_worst"  "compactness_worst"
[7] "concavity_worst"   "symmetry_worst"    "fractal_dim_worst"

$`4`
[1] "texture_se"    "smoothness_se" "symmetry_se"
```

```
$`5`
[1] "compactness_se" "concavity_se"   "concave_pts_se" "fractal_dim_se"

>
>
>
> pca <- prcomp(x_scaled)
> summary(pca)
Importance of components:
                          PC1   PC2    PC3   PC4   PC5    PC6    PC7    PC8
Standard deviation     3.644 2.386 1.6787 1.407 1.284 1.0988 0.8217 0.6904
Proportion of Variance 0.443 0.190 0.0939 0.066 0.055 0.0403 0.0225 0.0159
Cumulative Proportion  0.443 0.632 0.7264 0.792 0.847 0.8876 0.9101 0.9260
                          PC9   PC10   PC11    PC12    PC13    PC14    PC15
Standard deviation     0.6457 0.5922 0.5421 0.51104 0.49128 0.39624 0.30681
Proportion of Variance 0.0139 0.0117 0.0098 0.00871 0.00805 0.00523 0.00314
Cumulative Proportion  0.9399 0.9516 0.9614 0.97007 0.97812 0.98335 0.98649
                          PC16    PC17    PC18    PC19    PC20  PC21    PC22
Standard deviation     0.28260 0.24372 0.22939 0.22244 0.17652 0.173 0.16565
Proportion of Variance 0.00266 0.00198 0.00175 0.00165 0.00104 0.001 0.00091
Cumulative Proportion  0.98915 0.99113 0.99288 0.99453 0.99557 0.997 0.99749
                          PC23   PC24    PC25    PC26    PC27    PC28    PC29
Standard deviation     0.15602 0.1344 0.12442 0.09043 0.08307 0.03987 0.02736
Proportion of Variance 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005 0.00002
Cumulative Proportion  0.99830 0.9989 0.99942 0.99969 0.99992 0.99997 1.00000
                         PC30
Standard deviation     0.0115
Proportion of Variance 0.0000
Cumulative Proportion  1.0000
>
>
>
> data.frame(pca$x[,1:2], type = brca$y) %>%
+   ggplot(aes(PC1, PC2, color = type)) +
+   geom_point()
>
>
>
> data.frame(type = brca$y, pca$x[,1:10]) %>%
+     gather(key = "PC", value = "value", -type) %>%
+     ggplot(aes(PC, value, fill = type)) +
+     geom_boxplot()
>
>
```