Breast Cancer Project Part 1

The **brca** dataset contains information about breast cancer diagnosis biopsy samples for tumors that were determined to be either benign (not cancer) and malignant (cancer). The **brca** object is a list consisting of:

- **brca$y**: a vector of sample classifications ("B" = benign or "M" = malignant)
- **brca$x**: a matrix of numeric features describing properties of the shape and size of cell nuclei extracted from biopsy microscope images

For these exercises, load the data by setting your options and loading the libraries and data as shown in the code here:

```
options(digits = 3)
library(matrixStats)
library(tidyverse)
library(caret)
library(dslabs)
data(brca)
```

The exercises in this assessment are available to Verified Learners only and are split into four parts, all of which use the data described here.

IMPORTANT: Some of these exercises use **dslabs** datasets that were added in a July 2019 update. Make sure your package is up to date with the command `update.packages("dslabs")`. You can also update all packages on your system by running `update.packages()` with no arguments, and you should consider doing this routinely.

Question 1: Dimensions and properties

2.0/2.0 points (graded)

How many samples are in the dataset?

569 ✔ **Answer:** 569

569

Explanation
The number of samples can be found using the following code:

```
dim(brca$x)[1]
```

How many predictors are in the matrix?

30 ✔ **Answer:** 30

30

Explanation
The number of predictors can be found using the following code:

```
dim(brca$x)[2]
```

What proportion of the samples are malignant?

0.373 ✔ **Answer:** 0.373

0.373

Explanation
The proportion of samples can be found using the following code:

```
mean(brca$y == "M")
```

Which column number has the highest mean?

24 ✔ **Answer:** 24

24

Explanation
The column number can be found using the following code:

```
which.max(colMeans(brca$x))
```

Which column number has the lowest standard deviation?

20 ✔ **Answer:** 20

20

Explanation
The column number can be found using the following code:

```
which.min(colSds(brca$x))
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 2: Scaling the matrix

2.0/2.0 points (graded)
Use `sweep` two times to scale each column: subtract the column mean, then divide by the column standard deviation.

After scaling, what is the standard deviation of the first column?

1 ✔ **Answer:** 1

1

Explanation
The standard deviation can be found using the following code:

```
x_centered <- sweep(brca$x, 2, colMeans(brca$x))
x_scaled <- sweep(x_centered, 2, colSds(brca$x), FUN = "/")

sd(x_scaled[,1])
```

After scaling, what is the median value of the first column?

-0.215 ✔ **Answer:** -0.215

−0.215

Explanation
The median value can be found using the following code:

```
median(x_scaled[,1])
```

Submit You have used 2 of 10 attempts

Answers are displayed within the problem

Question 3: Distance

2.0/2.0 points (graded)
Calculate the distance between all samples using the scaled matrix.

What is the average distance between the first sample, which is benign, and other benign samples?

4.4 ✔ **Answer:** 4.41

4.4

Explanation
The average distance can be found using the following code:

```
d_samples <- dist(x_scaled)
dist_BtoB <- as.matrix(d_samples)[1, brca$y == "B"]
mean(dist_BtoB[2:length(dist_BtoB)])
```

What is the average distance between the first sample and malignant samples?

7.12 ✔ **Answer:** 7.12

7.12

Explanation
The average distance can be found using the following code:

```
dist_BtoM <- as.matrix(d_samples)[1, brca$y == "M"]
mean(dist_BtoM)
```

Submit You have used 2 of 10 attempts

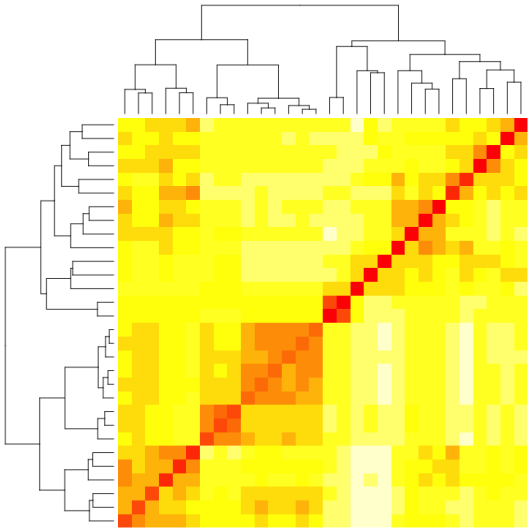
Answers are displayed within the problem

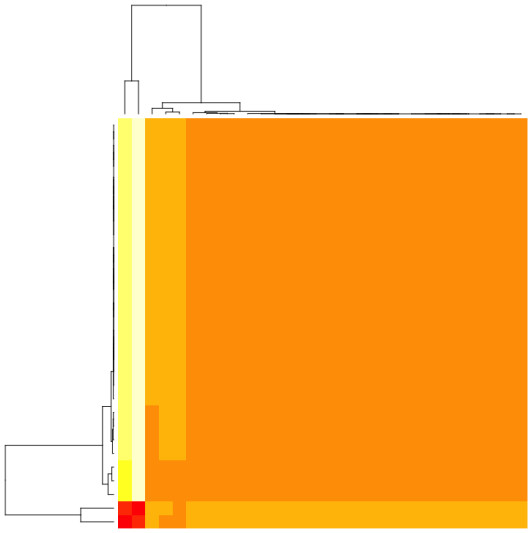
Question 4: Heatmap of features

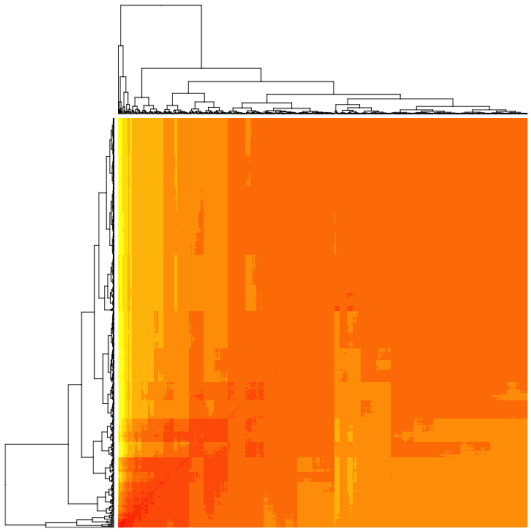
0.0/1.0 point (graded)
Make a heatmap of the relationship between features using the scaled matrix.

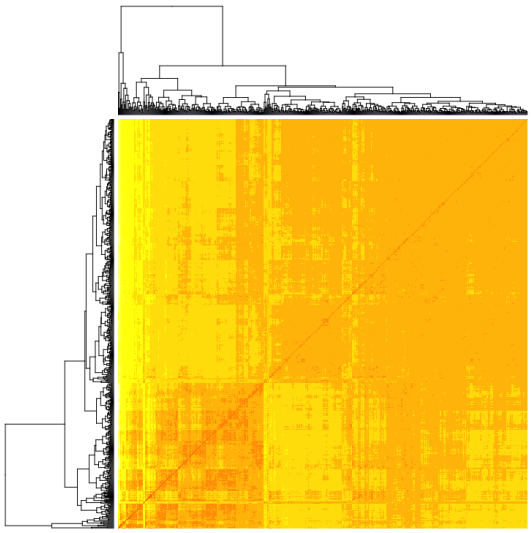
Which of these heatmaps is correct?

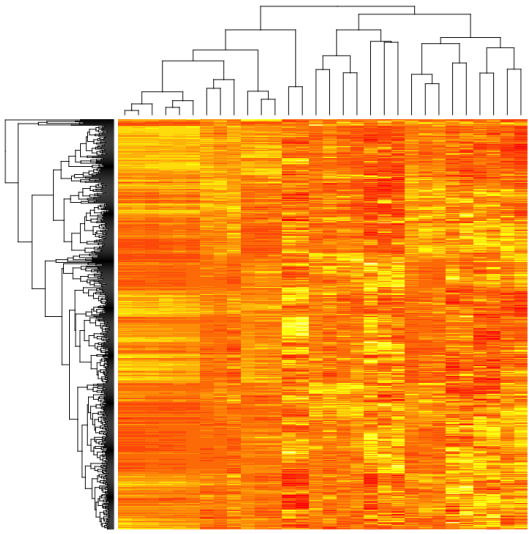
To remove column and row labels like the images below, use `labRow = NA` and `labCol = NA`.











✖

Answer
Incorrect: Try again. This heatmap is not scaled or transformed.

Explanation
The correct heatmap can be generated using the following code:

```
d_features <- dist(t(x_scaled))
heatmap(as.matrix(d_features), labRow = NA, labCol = NA)
```

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Question 5: Hierarchical clustering

1.0/1.0 point (graded)

Perform hierarchical clustering on the 30 features. Cut the tree into 5 groups.

All but one of the answer options are in the same group.

Which is in a different group?

- `smoothness_mean`
- `smoothness_worst`
- `compactness_mean`
- `compactness_worst`
- **`concavity_mean`** ✔
- `concavity_worst`

Explanation

The hierarchical clustering can be done using the following code:

```
h <- hclust(d_features)
groups <- cutree(h, k = 5)
split(names(groups), groups)
```

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem