1 Oil

Brodnjak-Voncina et al. (2005) describe a set of data where seven fatty acid compositions were used to classify commercial oils as either pumpkin (labeled A), sunflower (B), peanut (C), olive (D), soybean (E), rapeseed (F) and corn (G). There were 96 data points contained in their Table 1 with known results. The breakdown of the classes is given in below:

```
data(oil)
dim(fattyAcids)
[1] 96  7
table(oilType)
```

```
oilType
 A  B  C  D  E  F  G
37 26  3  7 11 10  2
```

As a note, the paper states on page 32 that there are 37 unknown samples while the table on pages 33 and 34 shows that there are 34 unknowns.

Using the data from the Examples section of `caret::createFolds`

```
library(caret)
data(oil)
part <- createDataPartition(oilType, 2)
fold <- createFolds(oilType, 2)

length(Reduce(intersect, part))
# [1] 27
```

```
length(Reduce(intersect, fold))
#[1] 0
```

Looks like `createDataPartition` split your data into smaller pieces, but allows for the same example to appear in different splits.

`createFolds` doesn't allow different examples to appear in different splits of the folds.

Basically, createDataPartition is used when you need to make one or more simple two-way splits of your data. For example, if you want to make a training and test set and keep your classes balanced, this is what you could use. It can also make multiple splits of this kind (or leave-group-out CV aka Monte Carlos CV aka repeated training test splits).

createFolds is exclusively for k-fold CV. Their usage is simular when you use the returnTrain = TRUE option in createFolds.