Cheat sheet for Classification

Libraries

```
library(tidyverse)
library(dplyr)
```

K-means clustering

We can

```
wcss = vector()
for (i in 1:10){
   wcss[i] = sum(kmeans(dataset), i)$withinss)
}
```

Using the elbow method to find the optimal number of clusters

```
plot(1:10,
     wcss,
     type = 'b',
     main = paste('The Elbow Method'),
     xlab = 'Number of clusters',
     ylab = 'WCSS')
```

After deciding on the number of clusters we can derive the clusters:

```
kmeans = kmeans(x = dataset, centers = 5)
y_kmeans = kmeans$cluster
```

…and add the clusters to the dataset

```
dataset <- data.frame(dataset, y_kmeans)
```

Visualising the clusters

```
library(cluster)
clusplot(dataset[2,11],
         y_kmeans,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste('Clusters of commodity groups'),
         xlab = 'PC1',
         ylab = 'PC2')
```

Hierarchical clustering

Using the dendrogram to find the optimal number of clusters. NB! Specify similarity measures!

```
dendrogram = hclust(d = dist(dataset),
                    method = 'euclidean'),
                    method = 'ward.D')
```

And visualizing

```
plot(dendrogram,
     main = paste('Dendrogram'),
```

```
     xlab = 'Customers',
     ylab = 'Euclidean distances')
```

Cutting the dendrogram with specified number of clusters

```
y_hc = cutree(dendrogram, 5)
```

…and add the clusters to the dataset

```
dataset <- data.frame(dataset, y_hc)
```

Visualising the clusters

```
library(cluster)
clusplot(dataset[,2:11],
         y_hc,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels= 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste('Clusters of commodity groups'),
         xlab = 'PC1',
         ylab = 'PC2')
```