Opgaver 2

1. **CH index**
   Implementer i `R` CH indexet som diskuteret i forelæsningen

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)} = \frac{SS_B/(K-1)}{SS_W/(n-K)}.$$

   Konstruer funktionen så den tager et `kmeans`-objekt som input og returnerer CH indexet:

```
ch_index <- function(kmeans_obj){
  ...
  ch
}
```

2. **Simuler data**

   Brug fx. `mvrnorm` fra fx. `mvtnorm` pakken til at simulere multivariat normalfordelte variable. Variér antallet af clustre samt deres form, `Sigma`, og placering, `mu`.

   i) Undersøg hvorledes k-means, `kmeans`, k-mediods, `pam`, og et par hierarkiske clusterings algoritmer (fra pakken `cluster`) identificere clustrene:
   Agglomerative: `hclust` (med forskellige link) og `agnes`. *Divisive:* `diana`

3. **drengenavne.csv**

   Indlæs `drengenavne.csv`. For at undgå at navnene bliver indlæst som faktorer, skal I bruge argumentet `stringsAsFactors=FALSE` i `read.csv` (eller `readr::read_csv`).

   Brug `adist`-funktionen til at bestemme afstanden Levenshtein mellem navnene.

   i) Brug `agnes` og `hclust` til at lave hierarkisk clustering.

   ii) Brug `cutree` til at opdele i tre clustre Er denne opdeling den samme som ved brug af k-mediods, `pam`?

4. **vote.csv**

   Start med at blive bekendt med funktionen `daisy` i `cluster`-pakken vha. `?daisy`

   Bl.a. kan en dataframe med forskellige typer håndteres (numeriske, kategoriske, ordinale) samt forskellige vægte (`weights`) kan benyttes for forskellige attributter.

   i) Lav cluster analyse af `vote.csv` hvor `party`-variablen udelades. Benyt efterfølgende denne til at vurdere performance af forskellige cluster metoder (beregn fejlrate).

5. **wine.csv**

   i) Analyser vha. af en passende cluster algoritme data i `wine.csv` (data og variable er beskrevet i `wine.names`) og udelad `Type` fra analysen. Benyt denne til at vurdere performance som i **vote.csv**.

   ii) Alle observationer (attributter) er numeriske. Er der fordel i at lave PCA på data før clustering (i forhold til fejlrate)?

   iii) Plot resultaterne - fx. ved brug af PCA.

   iv) Plot dissimilarity matrix vha. `heatmap`

6. **aims_freq.csv**

   i) Lav en hierarkisk cluster analyse af allel frekvenserne i `aims_freq.csv`. Hvilken `dist` metode og linkage giver bedst overensstemmelse med de geografiske grupperinger (se plot kommando i forrige selv studie).

I øvrigt...
Pakken `dendextend` giver nogle ret fede funktioner til at lege og plotte med dendrogrammer. Se eksempler på dendextend vignette