Samling av data

```
library(tidyverse)
library(rvest)
library(xml2)
library(kableExtra)
```

Laste ned SSB's forside

Først laster vi ned hele forsiden til SSB. Deretter laster vi ned de første linkene til hovedkategoriene på forsiden til SSB. Det er 23 linker vi laster ned.

```
# Laste ned SSB's forside
ssb <- download_html(
  url = "https://www.ssb.no/",
  file = "./Data/SSB.html"
)
```

Chunken over er satt til eval=FALSE for at den ikke skal kjøre hver gang vi Knit'er dokumentet.

Laste ned hovedkategoriene

```
# Lese forsiden og finne hovedkategoriene i teksten
hovedkategorier <- read_html("./Data/SSB.html") %>%
  html_nodes('a[class="ssb-link with-icon"]') %>% # klassenavnet for 'hovedkategoriene',
  # vi leser ut hver node som oppfyller dette kravet
  html_attr("href") %>% #forteller at det er en link vi vil ha
  str_c("https://ssb.no", .) #setter sammen hva vi henter med html_nodes med dette
```

Laster ned underkategoriene

Innenfor hver av (/de fleste?) hovedkategoriene finnes en del underkategorier. Vi itererer gjennom alle 23 html-filene med hovedkategoriene lasted ned lokalt for å finne underlinkene.

```
# Step 2: Download all the webpages from the links you just made
hovedkategori_navn <- str_remove(hovedkategorier, "https://ssb.no")
```

```
for(i in 1:length(hovedkategorier)) {
  download.file(hovedkategorier[[i]],
                destfile = str_c("./Data/hovedkategorier/",
                                 hovedkategori_navn[i], ".html"))
  # Download one html-file after another into the folder hovedkategorier
```

```
    Sys.sleep(2)
  # Setting a timer of two seconds each time we download a webpage.
}
```

Har satt eval=FALSE på siste chunk fordi vi ikke trenger å kjøre denne koden mer enn én gang.

Strukturere underkategorier

Lager en liste for hver hovedkategori med alle underkategori html-pathene (url-linkene som vi skal bruke til å laste ned html-filene), nå kan vi ha en løkke som går gjennom listen i listen for å laste ned sidene.

```
underkategorier <- list()

tmp <- list()

for (i in 1:length(hovedkategori_navn)) {
# For each i in every element from place number one to the last place in hovedkategorier_navn
  # (given by length(hovedkategorier_navn))
  filnavn = paste("Data/hovedkategorier", hovedkategori_navn[i], ".html", sep="")
  # Read the html-page for each i
    tmp <- read_html(filnavn)   %>% # create a tmp list with the undercategories
    html_nodes('a[class="ssb-category-link"]') %>% #again with the nodes
    html_attr("href") %>%
    str_c("https://ssb.no", .)
  # insert list of undercategory in list: structure
   underkategorier[[i]] <- tmp
}
```

Laster ned html-filene for alle underkategoriene. Det er i disse html-filene statistikken ligger. Vi må nå inn i hver statistikk, og i hver statistikk for vi kjørt program som samler inn 'om-statistikken'-delen.

```
for(i in 1:length(underkategorier)) { # 23 elementer
  liste = underkategorier[[i]]
# We download the underkategori html-files
  string_som_skal_fjernes <- paste("https://ssb.no", hovedkategori_navn[i], sep="")
  underkategori_navn <- str_remove(liste, string_som_skal_fjernes)
  for(j in 1:length(liste)){
    download.file(liste[j],
                  destfile = str_c("./Data/underkategorier/", underkategori_navn[j], ".html"))
    # Download one html-file after another into the folder hovedkategorier
      Sys.sleep(2)
    # Setting a timer of two seconds each time we download a webpage.
  }
}
```

Satt eval=FALSE fordi chunken over kun trenger å kjøres én gang.

Statistikken hentes ut

Metode som lager url'en til alle statistikkene, liste i liste for hver underkategori.

```
statistikk <- list()

for (i in 1:length(hovedkategori_navn)){
  liste <- underkategorier[[i]]
  tmp <- list()
  string_som_skal_fjernes <- paste("https://ssb.no", hovedkategori_navn[i], sep="")
  underkategori_navn <- str_remove(liste, string_som_skal_fjernes)
  for (j in 1:length(liste)){
    filnavn <- paste("./Data/underkategorier", underkategori_navn[j], ".html", sep="")
    statistikk_2 <- read_html(filnavn) %>%
    html_nodes('div[class="ssb-card"]') %>%
    html_elements('a') %>% html_attr('href') %>%
    str_c('https://ssb.no', .)
    #again with the nodes
    tmp[[j]] <- statistikk_2
    #tester bare ut
  }
  statistikk <- append(statistikk, tmp)
}
```

Se på objektene tmp og statistikk:

```
tmp %>%
  head(n = 1) # Ser kun på [[1]]
```

```
## [[1]]
## [1] "https://ssb.no/nasjonalregnskap-og-konjunkturer/finansregnskap/statistikk/finansielle-sektorregn
## [2] "https://ssb.no/bank-og-finansmarked/finansielle-indikatorer/statistikk/kredittindikator"
## [3] "https://ssb.no/virksomheter-foretak-og-regnskap/konkurser/statistikk/opna-konkursar"
## [4] "https://ssb.no/bank-og-finansmarked/finansielle-indikatorer/statistikk/pengemengde"
```

```
statistikk %>%
  head(n = 3) # Ser kun på [[1]] til og med [[3]]
```

```
## [[1]]
## [1] "https://ssb.no/arbeid-og-lonn/sysselsetting/statistikk/arbeidskraftundersokelsen"
## [2] "https://ssb.no/arbeid-og-lonn/sysselsetting/statistikk/ledige-stillingar"
##
## [[2]]
## [1] "https://ssb.no/arbeid-og-lonn/arbeidsmiljo-sykefravaer-og-arbeidskonflikter/statistikk/arbeidsmi
## [2] "https://ssb.no/helse/helseforhold-og-levevaner/statistikk/arbeidsulykker"
## [3] "https://ssb.no/arbeid-og-lonn/arbeidsmiljo-sykefravaer-og-arbeidskonflikter/statistikk/fagforein
## [4] "https://ssb.no/arbeid-og-lonn/arbeidsmiljo-sykefravaer-og-arbeidskonflikter/statistikk/sykefrava
##
## [[3]]
## [1] "https://ssb.no/arbeid-og-lonn/sysselsetting/statistikk/antall-arbeidsforhold-og-lonn"
## [2] "https://ssb.no/arbeid-og-lonn/lonn-og-arbeidskraftkostnader/statistikk/arbeidskraftkostnader"
## [3] "https://ssb.no/arbeid-og-lonn/lonn-og-arbeidskraftkostnader/statistikk/arbeidskraftkostnadsindel
## [4] "https://ssb.no/arbeid-og-lonn/lonn-og-arbeidskraftkostnader/statistikk/grunnlag-for-arbeidsgiver
## [5] "https://ssb.no/arbeid-og-lonn/lonn-og-arbeidskraftkostnader/statistikk/lonn"
```

Hvor mange statistikker?

```
counter <- 0

for (i in 1:length(statistikk)){
  liste <- statistikk[[i]]
  for (j in 1:length(liste)){
    counter <- counter + 1
  }
}

counter
```

```
## [1] 574
```

Vi ser at vi har 574 statistikk-linker.

Strukturere for nedlastning

Bytter ut / med _ i url linkene for å kunne lagre som fil:

```
statistikk_filnavn <- list()

for (i in 1:length(statistikk)){
  liste <- statistikk[[i]]
  for (j in 1:length(liste)){
    liste[j] <- str_remove(liste[j], "https://ssb.no/")
    statistikk_filnavn <- append(statistikk_filnavn, gsub("/", "_", liste[j]))
  }
}
```

Laste ned statistikken

Laste ned html-filene til alle statistikkene

((Ikke ferdig))

```
getwd()

#underkategori_navn_liste <- list()

for(i in 1:length(statistikk)) { #23 elementer
  liste <- statistikk[[i]]
# We download the underkategori html files
  string_som_skal_fjernes <- paste("https://ssb.no", hovedkategori_navn[i], sep="")
  statistikk_navn <- str_remove(liste, string_som_skal_fjernes)
  for(j in 1:length(liste)){
    download.file(liste[j],
                  destfile = str_c("./Data/underkategorier/", underkategori_navn[j], ".html"))
    # Download one html-file after another into the folder hovedkategorier
```

```
      Sys.sleep(2)
    # Setting a timer of two seconds each time we download a webpage.
  }
}
```