Assignment 4

Innledning

I denne siste oppgaven som er «svennestykket» vårt skal vi boltre oss i data fra SSB. Vi skal lese inn 4 moderat store tabeller

1. 10540 Registrerte arbeidsledige i % fordelt på tre aldersgrupper per kommune i perioden januar 1999 til mai 2014. Månedlige data.
2. 04471 Registrert arbeidsløse i % fordelt på kjønn per kommune i perioden januar 1999 til mai 2014. Månedlige data.
3. 10594 Registrert antall arbeidsløse fordelt på kjønn per kommune i perioden januar 1999 til mai 2014. Månedlige data.
4. 07459 Befolkning fordelt på 106 årsklasser fra 000-105+ (totalt 106 årsklasser) fordelt på kommuner. Årlige data (antar start av året).

Tanken er at vi skal få en følelse med hvordan det er å jobbe med virkelige data. Oppgaven består av tre deler. Første del dreier seg om å lese data direkte inn fra SSB via en api (`get-ssb-data.Rmd`). I andre del behandler vi dataene og får dem i et **tidy** format (`make-data-tidy.Rmd`) før vi så i tredje del (`model.Rmd`) benytter teknikkene vi lærte i slutten av kurset til å kjøre lineære modeller for hver enkelt kommune.

Vi vil bruke R pakken `PxWebApiData` som gir direkte tilgang til SSBs statstikkbank. Pakken kan installeres på vanlig vis og lastes vha. ‘library(`PxWebApiData`)‘ i første code-chunk kalt `setup`.

Arbeidet vi skal gjøre skal fordeles på 3 R Notebooks. På øverste nivå skal det ligge en Notebook kalt `model.Rmd` under dette nivået har vi en `Folder` kalt `data`. Mappen data skal inneholde to Notebooks kalt hhv. `get-ssb-data.Rmd` (dere skal få en mal for denne) og `make-data-tidy.Rmd`. I `get-ssb-data.Rmd` skal vi samle koden som henter dataene fra SSB og skriver dataene ut i 5 `.csv` filer. Når vi er ferdig med denne setter vi `knitr::opts_chunk$set(eval = FALSE)` i setup chunk-en og skriver inn i dokumentet datoen vi hentet dataene. Dette gjør vi for å hindre at vi ved et uhell henter inn dataene på nytt (som i prinsippet kan ha endret seg hvis SSB oppdaterer statistikken).

`.csv` filene som vi skrev ut til slutt i `get-ssb-data.Rmd` leser vi inn i `make-data-tidy.Rmd` og gjør eventuelle endringer på rådataene for å få dem i et **tidy** format. Når vi er ferdige skriver vi ut nye **tidy** `.csv` filer som vil være input for vår modellering i filen `model.Rmd`.

For å hjelpe dere i gang har jeg laget en mal for `get-ssb-data.Rmd`, `make-data-tidy.Rmd` og `model.Rmd`. Instruksjoner er gitt i filene. Noe kode ligger der allrede, men dere må fylle inn resten. Begynn med `get-ssb-data.Rmd` fortsett så med `make-data-tidy.Rmd`. Husk at begge disse skal ligge i en undermappe kalt data sammen med datafilene dere genererer. Når dere har ferdig de to «tidy» datafilene “al9914m.csv” og “bef9914MK.csv” kan dere forsette med `model.Rmd` som skal ligge i mappen ovenfor data.

Start med å lage en repository på github for assignment-4. Når dere har laget prosjektet i RStudio legger dere inn malen mal-model.Rmd i denne mappen. Endre navnet fra mal-model.Rmd til model.Rmd. Opprett en undermappe kalt data. I denne legger dere mal-get-ssb-data.Rmd og mal-make-data-tidy. Rmd. Endre navnene til get-ssb-data.Rmd og make-data-tidy.Rmd.

Dere er nå klare til å starte arbeidet med get-ssb-data.Rmd (kanskje en commit er på sin plass).

Jeg har lagt ut filene get-ssb-data.pdf, make-data-tidy.pdf og model.pdf. Der echo=FALSE er satt for den koden dere skal skrive selv. Filene viser imidlertid hvilke variabelnavn jeg har brukt og størrelsen på de ulike tibble/dataframene. Jeg vil be dere om å bruke samme variabelnavn som meg. Det gjør det enklere hvis dere skulle trenge hjelp. Størrelsen kan dere bruke for å sjekke at dere er på rett veg. Lykke til!