Analyse av populasjonsdata

Introduksjon

I denne mappeoppgaven skal vi se på populasjonsdata. Vi skal benytte oss av data fra NCHS ("National Center for Health Statistics") som registrerer alle fødsler i USA.

Data

Her finner du dataene du skal bruke. Vi skal benytte data fra den øverste 'boksen', merket "Birth Data Files". Du trenger "U.S. Data (.zip files)" og "User's Guide (.pdf files)".

Alle skal laste ned og benytte dataene fra 2022. Dette er et stort datasett på over 200 MB, med nesten 3,7 millioner obeservasjoner og over 250 variabler. Du skal derfor ikke benytte hele datasettet. Du skal deretter laste ned et tilsvarende datasett, som er mellom 10 og 30 år gammelt.

Lese fast filformat

I "User's Guide (.pdf files)" finner du informasjon om strukturen/posisjonen på dataene. Dataene er i fast filformat. Du skal dermed benytte en funksjon som leser .zip-filer og fast filformat.

Du kan f.eks lese fast filformat med `read_fwf` fra pakken `readr`. Benytt funksjonen for å spesifisere posisjonene til kolonnene. Videre kan du spesifisere kolonnetypene med `col_types`. "User's Guide (.pdf files)" gir deg informasjon om posisjon og navnet på kolonnen/variabelen.

Dataoppgaver

Du skal trekke ut data på barnets fødselsår, måned, tidspunkt, ukedag, kjønn og vekt. Du skal beholde data på mor og fars alder og utdannelse, samt sivilstatus (gift eller ikke). Du skal også beholde informasjon om dette er den førstefødte. I tillegg skal du registrere hvor mange sigaretter mor røykte per dag under svangerskapet (flere varaibler). Røykevaner ble iallefall registrert tilbake i 1992 (30 år siden). Du skal deretter slå sammen disse datasettene til ett datasett.

Rensking

Du skal rense dataene for evt manglende og feilaktige verdier. Skriv korte kommentarer i koden din som forklarer prosessen. *Analysene skal kun gjennomføres for mødre som er førstefødende.* Når du er ferdig med å rense og slå sammen datasettene, skal du lagre det nye datasettet som du gjør dine analyser på som en "feather" fil. Dette er et binært filformat som er raskt å lese og skrive til. Du kan lese mer om feather her.

Analyse

Analysen består av en rekke ggplot-figurer og tabeller. Som tekst/kommentar skal du skrive en kort begrunnelse hvorfor du valgte en spesifikk type figur/layout. Alle figurer skal ha selvforklarende "labler", tekst på akser og titler. En figur trenger ikke å være bare et plot, men kan bestå av flere plot ved siden av eller over hverandre. Alle tabeller skal ha selvforklarende kategorier og titler. Alle figurer og tabeller har til hensikt å sammenligne 2022 med det eldre datasettet.

Figur 1

- Du skal lage en figur som viser fordelingen av fødsler per måned. Du skal også legge til et informasjonslag som viser den relative fordelingen av fødsler per måned.

Figur 2

- Du skal lage en figur som viser fordelingen av fødsler per ukedag. Du skal også legge til et informasjonslag som viser den relative fordelingen av fødsler per ukedag.

Figur 3

- Du skal lage en figur som viser fordelingen av fødsler per tidspunkt på døgnet. Du skal også legge til et informasjonslag som viser den relative fordelingen av fødsler per tidspunkt på døgnet.

Figur 4

- Du skal lage en figur som viser fordelingen av fødsler per alder på mor.

Figur 5

- Du skal lage en figur som viser fordelingen av fødsler per alder på far.

Figur 6

- Du skal lage en figur som viser mors og fars alder ved fødsel. Du skal også legge til et informasjonslag som viser gjennomsnittet av mors og fars alder ved fødsel.

Figur 7

- Du skal lage en figur som viser gjennomsnittlig fødselsvekt per kjønn.

Du skal nå lage tabeller. Du kan du bruke `knitr::kable` og `kableExtra` for å lage tabellene. Et annet godt alternative er `gt` fra pakken `gt`. Her finner du dem.

Tabell 1

- Du skal lage en tabell som viser fordelingen av fødsler per utdannelse på mor.

Tabell 2

- Du skal lage en tabell som viser fordelingen av fødsler per utdannelse på far.

Tabell 3

- Du skal lage en tabell som viser fordelingen av fødsler per sivilstatus på mor.

Tabell 4

- Du skal lage en tabell som viser hvor mye mor røykte før og under svangerskapet.

Tabell 5

- Du skal lage en tabell som viser gjennomsnittlig fødselsvekt per antall sigaretter mor røykte per dag under svangerskapet.