Political Data Science

Lektion 3

R Workshop I: Import, tidy, transform

Opgave 1

1. Indlæs pakken `tidyr` og gem et objekt med datasættet `who` i pakken, som indeholder data fra WHO om nye tilfælde af tuberkulose siden 1980.
2. Hvad er dimensionerne på datasættet?
3. Undersøg datasættet. Er det tidy?
4. Hvad er logikken i kolonnernes navne fra `new_sp_m014` til `newrel_f65`? Få evt hjælp med `?who`.
5. Brug `gather()` til at samle værdierne i kolonnerne til variable. Kald de nye kolonner `key` og `cases` indtil videre.
6. Ledende spørgsmål: Er værdierne i kolonnen `key` navngivet konsistent? Ret `newrel` til `new_rel` [hint: `str_replace()`]
7. Brug `separate()` til at opdele værdierne i kolonnen `key`, så hver værdi har sin egen kolonne og bliver sin egen variabel [hint: tænk på logikken i variablene fra opgave 1.4].
8. Er dit datasæt tidy nu? Og hvad er der sket med dimensionerne på datasættet?

Opgave 2

1. `tidyr` indeholder også datasættet `population`. Tjek det ud. Hvis du vil joine det på, hvilke variable vil så være primary keys? Og hvad betyder det?
2. Join `population` på din tidy udgave af `who` via `left_join()` og gem den resulterende dataframe i et nyt objekt.
3. Hvilke lande - hvis nogen - finder ikke et match når du joiner `population` på `who`? [hint: `anti_join()`]
4. Hvilke(n) variable indeholder NAs? Hvad tror du de skyldes? Gem et nyt objekt, hvor du filtrerer NAs fra.
5. Hvor mange nye tilfælde af tuberkulose var der i hvert af årene 2000 - 2013? Kommentér på tallene.
6. Plot hvert lands udvikling i tuberkulose-tilfælde i årene 2000 - 2013 ved siden af hinanden [hint: `facet_wrap()`]. Hvilke lande ser ud til at drive den udvikling, vi fandt i spørgsmål 5?
7. Lav en variabel med tilfælde af tuberkulose per 100.000 indbyggere (population / 100.000), og plot den for alle landene ved siden af hinanden. Hvilke lande stikker ud nu?

Bonus-opgave

Hent data fra Verdensbanken om den andel af bruttonationalprodukterne, som verdens lande bruger på deres sundhedssystemer. Find data [her]. Undersøg datasættet og gør det tidy. Join data på de tidligere data fra WHO. Illustrér med et plot, om antallet af tuberculose-tilfælde for et enkelt år hænger sammen med den andel af BNP, som landene bruger på sundhed for et enkelt år.