SOK-1004 H24 Forelesning 3

tidyverse og ggplot2

Even S. Hvinden

1. Innledning

- **Mål:** Bruke grunnleggende kommandoer fra dplyr og ggplot2 til å lage fi

  gurer
- **Transformere data:** Fem kommandoer i dplyr: `select()`, `filter()`, `arrange()`, `mutate()`, `summarize()`
- **Visualisere data:** Vi bruker ggplot2 i kombinasjon med dplyr
- **Gjennomføring:** Jeg introduserer kommandoene, dere prøver selv ved hjelp av vedlagt kode

2. Koden til forelesning 3 `SOK-1004_Forelesning_3_h24.R`

2.1 Lagre data i en tibble

En tibble er det anbefalte lagringsformatet i tidyverse. Datasettet kommer i en csv fil, og tidyverse lagrer dette automatisk som tibble (tbl). Skriv `class(co2data)` for å sjekke.

```
# rydd opp
rm(list=ls())

# last inn tidyverse
# sikre at oppstart på pakken ikke vises i pdf
suppressPackageStartupMessages(library(tidyverse))

############################
### data i tibble-format ###
############################
```

```r
# les CO2 data i .csv fra OWID
url <- "https://raw.githubusercontent.com/owid/co2-data/master/owid-co2-data.csv"
co2data <- read_csv(url)
```

```
Rows: 47415 Columns: 79
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): country, iso_code
dbl (77): year, population, gdp, cement_co2, cement_co2_per_capita, co2, co2...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# se: https://github.com/owid/co2-data
# se en beskrivelse av data her:
# https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv

# se på dataene i konsollen
co2data
```

```
# A tibble: 47,415 x 79
   country      year iso_code population   gdp cement_co2 cement_co2_per_capita
   <chr>       <dbl> <chr>         <dbl> <dbl>      <dbl>                  <dbl>
 1 Afghanistan  1850 AFG         3752993    NA         NA                     NA
 2 Afghanistan  1851 AFG         3767956    NA         NA                     NA
 3 Afghanistan  1852 AFG         3783940    NA         NA                     NA
 4 Afghanistan  1853 AFG         3800954    NA         NA                     NA
 5 Afghanistan  1854 AFG         3818038    NA         NA                     NA
 6 Afghanistan  1855 AFG         3835192    NA         NA                     NA
 7 Afghanistan  1856 AFG         3852417    NA         NA                     NA
 8 Afghanistan  1857 AFG         3869714    NA         NA                     NA
 9 Afghanistan  1858 AFG         3887081    NA         NA                     NA
10 Afghanistan  1859 AFG         3904521    NA         NA                     NA
# i 47,405 more rows
# i 72 more variables: co2 <dbl>, co2_growth_abs <dbl>, co2_growth_prct <dbl>,
#   co2_including_luc <dbl>, co2_including_luc_growth_abs <dbl>,
#   co2_including_luc_growth_prct <dbl>, co2_including_luc_per_capita <dbl>,
#   co2_including_luc_per_gdp <dbl>, co2_including_luc_per_unit_energy <dbl>,
#   co2_per_capita <dbl>, co2_per_gdp <dbl>, co2_per_unit_energy <dbl>,
#   coal_co2 <dbl>, coal_co2_per_capita <dbl>, consumption_co2 <dbl>, ...
```

2.2 Velg variabel med `select()`

Bruk `select()` til å velge variable (kolonner) som egen tabell:

```
select(co2data, iso_code)
```

```
# A tibble: 47,415 x 1
   iso_code
   <chr>
 1 AFG
 2 AFG
 3 AFG
 4 AFG
 5 AFG
 6 AFG
 7 AFG
 8 AFG
 9 AFG
10 AFG
# i 47,405 more rows
```

2.3 Velg verdier med `filter()`

Bruk `filter()` til å velge verdier (rader) som egen tabell:

```
filter(co2data, iso_code == "AFG")
```

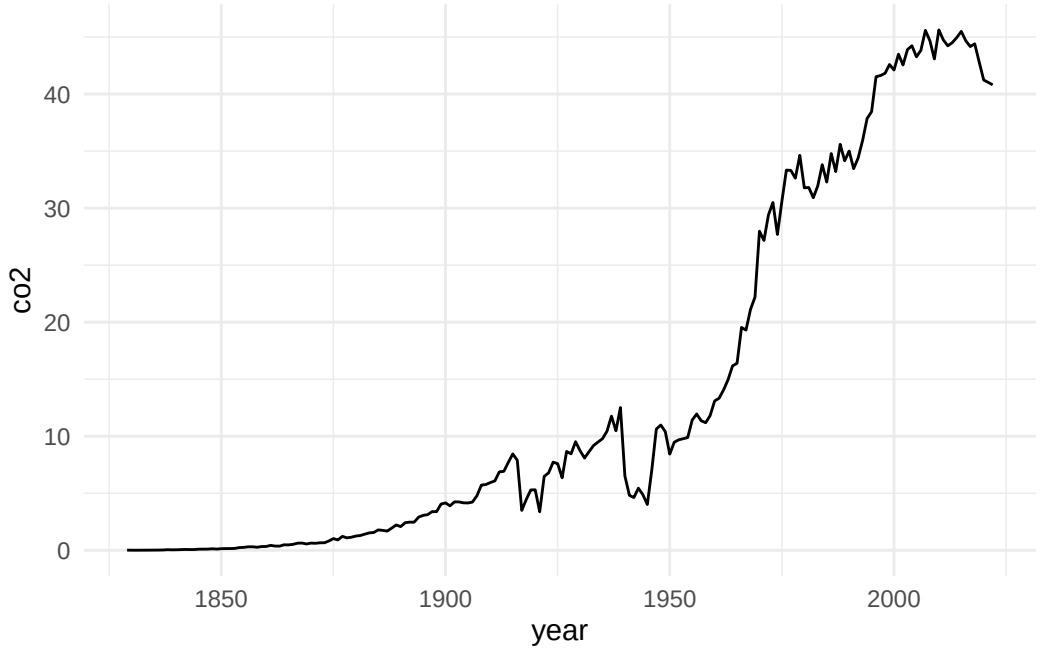
```
# A tibble: 173 x 79
   country      year iso_code population   gdp cement_co2 cement_co2_per_capita
   <chr>       <dbl> <chr>         <dbl> <dbl>      <dbl>                  <dbl>
 1 Afghanistan  1850 AFG         3752993    NA         NA                     NA
 2 Afghanistan  1851 AFG         3767956    NA         NA                     NA
 3 Afghanistan  1852 AFG         3783940    NA         NA                     NA
 4 Afghanistan  1853 AFG         3800954    NA         NA                     NA
 5 Afghanistan  1854 AFG         3818038    NA         NA                     NA
 6 Afghanistan  1855 AFG         3835192    NA         NA                     NA
 7 Afghanistan  1856 AFG         3852417    NA         NA                     NA
 8 Afghanistan  1857 AFG         3869714    NA         NA                     NA
 9 Afghanistan  1858 AFG         3887081    NA         NA                     NA
10 Afghanistan  1859 AFG         3904521    NA         NA                     NA
# i 163 more rows
```

```
# i 72 more variables: co2 <dbl>, co2_growth_abs <dbl>, co2_growth_prct <dbl>,
#   co2_including_luc <dbl>, co2_including_luc_growth_abs <dbl>,
#   co2_including_luc_growth_prct <dbl>, co2_including_luc_per_capita <dbl>,
#   co2_including_luc_per_gdp <dbl>, co2_including_luc_per_unit_energy <dbl>,
#   co2_per_capita <dbl>, co2_per_gdp <dbl>, co2_per_unit_energy <dbl>,
#   coal_co2 <dbl>, coal_co2_per_capita <dbl>, consumption_co2 <dbl>, ...
```

2.4 Bruk `ggplot2` til å lage figurer

- Data - aes() (koordinatsystem) - og geom (visualisering, farger, linjestørrelse) =)
- Råd: Bruk cheat sheet og pipes

```
co2data %>%
  filter(country =="Norway") %>%
  ggplot(aes(x=year, y=co2)) %>%
  + geom_line() %>%
  + theme_minimal()
```



2.5 Gjør oppgave 1-4 i koden til forelesning 3

2.6 Endre rekkefølge på verdier med `arrange()`

Bruk `arrange()` til å endre rekkefølgen på verdier:

```
arrange(co2data, co2)
```

```
# A tibble: 47,415 x 79
   country  year iso_code population   gdp cement_co2 cement_co2_per_capita
   <chr>   <dbl> <chr>         <dbl> <dbl>      <dbl>                  <dbl>
 1 Africa   1750 <NA>       76278087    NA          0                      0
 2 Africa   1751 <NA>             NA    NA          0                     NA
 3 Africa   1752 <NA>             NA    NA          0                     NA
 4 Africa   1753 <NA>             NA    NA          0                     NA
 5 Africa   1754 <NA>             NA    NA          0                     NA
 6 Africa   1755 <NA>             NA    NA          0                     NA
 7 Africa   1756 <NA>             NA    NA          0                     NA
 8 Africa   1757 <NA>             NA    NA          0                     NA
 9 Africa   1758 <NA>             NA    NA          0                     NA
10 Africa   1759 <NA>             NA    NA          0                     NA
# i 47,405 more rows
# i 72 more variables: co2 <dbl>, co2_growth_abs <dbl>, co2_growth_prct <dbl>,
#   co2_including_luc <dbl>, co2_including_luc_growth_abs <dbl>,
#   co2_including_luc_growth_prct <dbl>, co2_including_luc_per_capita <dbl>,
#   co2_including_luc_per_gdp <dbl>, co2_including_luc_per_unit_energy <dbl>,
#   co2_per_capita <dbl>, co2_per_gdp <dbl>, co2_per_unit_energy <dbl>,
#   coal_co2 <dbl>, coal_co2_per_capita <dbl>, consumption_co2 <dbl>, ...
```

2.7 Lag nye variable med `mutate()`

```
mutate(co2data, x= year + 1)
```

```
# A tibble: 47,415 x 80
   country      year iso_code population   gdp cement_co2 cement_co2_per_capita
   <chr>       <dbl> <chr>         <dbl> <dbl>      <dbl>                  <dbl>
 1 Afghanistan  1850 AFG         3752993    NA         NA                     NA
```

```
 2 Afghanistan  1851 AFG         3767956    NA         NA                    NA
 3 Afghanistan  1852 AFG         3783940    NA         NA                    NA
 4 Afghanistan  1853 AFG         3800954    NA         NA                    NA
 5 Afghanistan  1854 AFG         3818038    NA         NA                    NA
 6 Afghanistan  1855 AFG         3835192    NA         NA                    NA
 7 Afghanistan  1856 AFG         3852417    NA         NA                    NA
 8 Afghanistan  1857 AFG         3869714    NA         NA                    NA
 9 Afghanistan  1858 AFG         3887081    NA         NA                    NA
10 Afghanistan  1859 AFG         3904521    NA         NA                    NA
# i 47,405 more rows
# i 73 more variables: co2 <dbl>, co2_growth_abs <dbl>, co2_growth_prct <dbl>,
#   co2_including_luc <dbl>, co2_including_luc_growth_abs <dbl>,
#   co2_including_luc_growth_prct <dbl>, co2_including_luc_per_capita <dbl>,
#   co2_including_luc_per_gdp <dbl>, co2_including_luc_per_unit_energy <dbl>,
#   co2_per_capita <dbl>, co2_per_gdp <dbl>, co2_per_unit_energy <dbl>,
#   coal_co2 <dbl>, coal_co2_per_capita <dbl>, consumption_co2 <dbl>, ...
```

2.8 Få deskriptiv statistikk med `summarize()`

Bruk `summarize()` til å få en tabell med deskriptiv statistikk, eksempelvis sum, gjennomsnitt:

```
summarize(group_by(co2data, year), tot_co2 = sum(co2))
```

```
# A tibble: 273 x 2
    year tot_co2
   <dbl>   <dbl>
 1  1750    55.8
 2  1751    56.4
 3  1752    57.0
 4  1753    57.7
 5  1754    58.4
 6  1755    58.8
 7  1756    59.5
 8  1757    60.6
 9  1758    61.3
10  1759    62.0
# i 263 more rows
```

2.9 Gjør oppgave 5-8 i koden fra forelesning 3