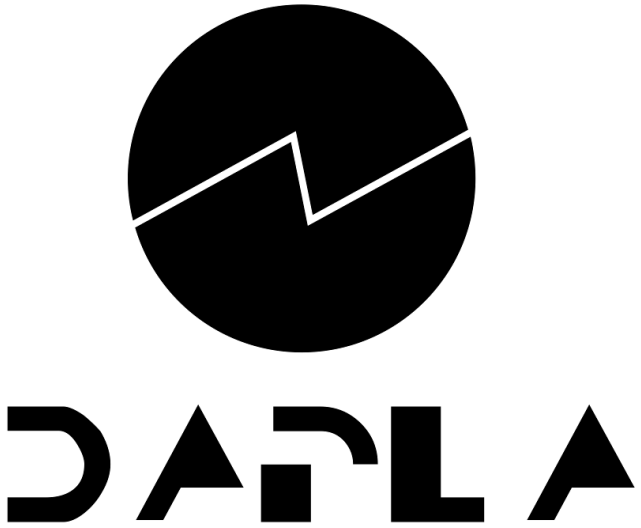
TALL SOM FORTELLER

på



GRUNNKURS

SUSIE JENTOFT

DAG 2

Statistisk sentralbyrå
Statistics Norway

Mål

- Kjennskap til programmet R og RStudio
- Åpne RStudio og kjøre enkle beregninger
- Lese inn data
- Behandle data
- Lage tabeller og oppsummere
- Lage figurer

Agenda

Dag 2: 2. februar

14:00 – Oppsummering og løsninger

14:10 – Databehandling og tabeller

14:30 – Øvelse 3

15:00 – Kobling og figurer

15:30 – Øvelse 4

15:50 – Oppsummering

Øvelser gjennomgang

Data behandling med tidyverse

- Gjør koden ryddigere
- Pipelines %>%

Base R:

```
leave_house(get_dressed(get_out_of_bed(wake_up(me))))
```

tidyverse:

```
me %>%
  wake_up() %>%
  get_out_of_bed() %>%
  get_dressed() %>%
  leave_house()
```

Data behandling med tidyverse

Funksjon	Hva den gjøre
mutate()	Lage nye variable
rename()	Gi en variabel et nytt navn
filter()	Selektere noen enheter/rad
select()	Selektere noen variable
summarise()	Oppsummering av en variabel
group_by()	Gjøre prosesser innenfor grupper

Lage nye variabler: mutate()

- Kan brukes som en del av en pipeline

```
datanavn %>%
  mutate(nyvariabel = 1000)
```

Gir variabel et navn

```
datanavn %>%
  mutate(nyvariabel = oldvariabel * 1000)
```

Gir variabel et navn

Eksisternende variabel

+ hva skal gjøres

Endre variabelnavn: rename()

```
datanavn %>%
  rename(nyttnavn = gammeltnavn)
```

Velg noen rader: filter()

- For å velge ut noen rader bruker vi filter()
- Skriv logiske setning inn i parentes.
- Flere logiske setninger kan brukes sammen (skille med ,)

```
datanavn %>%
  filter(condition)
```

- Igjen: Ingenting lagres uten <-

Velg ut noen variabler: select()

- Brukes med pipelines
- Skriv variabelnavn i parentes
- En eller flere variabler (skille med ,)
- Brukes sammen med andre funksjoner (for eks. filter( ))

```
datanavn %>%
  select(variabelnavn)
```

```
datanavn %>%
  filter(condition) %>%
  select(variabelnavn)
```

Oppsummering/aggregering: summarise()

- Ta oppsummering (summen, gjennomsnitt, median, antall) av en variabel med summarise()

```
datanavn %>%
  summarise(oppsummeringsnavn = mean(variabelnavn))
```

Gir oppsummering et navn

Hva skal gjøres:
- mean()
- median()
- sum()
- n()

Eksisternende variabel

Gruppering: group_by()

- Gjøre alle prosesser etterpå innen hver gruppe

```
datanavn %>%
  group_by(grupperingsvariabel) %>%
  summarise(oppsummeringsnavn = mean(variabelnavn))
```

Eksempler på dapla

Øvelser 3

- Øvelsene til oppgavesett 3 er på fil: **øvelser_dag2.R**

DataA

ID	Kommune	Omsetning
1	0301	212400
2	0301	75000
3	5011	620000

DataB

ID	Antall_ansatte
2	8
3	22
4	3

left_join(DataA, DataB)

ID	Kommune	Omsetning	Antall ansatte
1	0301	212400	NA
2	0301	75000	8
3	5011	620000	22

Koble to datasett

right_join(DataA, DataB)

ID	Kommune	Omsetning	Antall ansatte
2	0301	75000	8
3	5011	620000	22
4	NA	NA	3

DataA

ID	Kommune	Omsetning
1	0301	212400
2	0301	75000
3	5011	620000

DataB

ID	Antall_ansatte
2	8
3	22
4	3

inner_join(DataA, DataB)

ID	Kommune	Omsetning	Antall ansatte
2	0301	75000	8
3	5011	620000	22

Koble to datasett

full_join(DataA, DataB)

ID	Kommune	Omsetning	Antall ansatte
1	0301	212400	NA
2	0301	75000	8
3	5011	620000	22
4	NA	NA	3

Koble to datasett

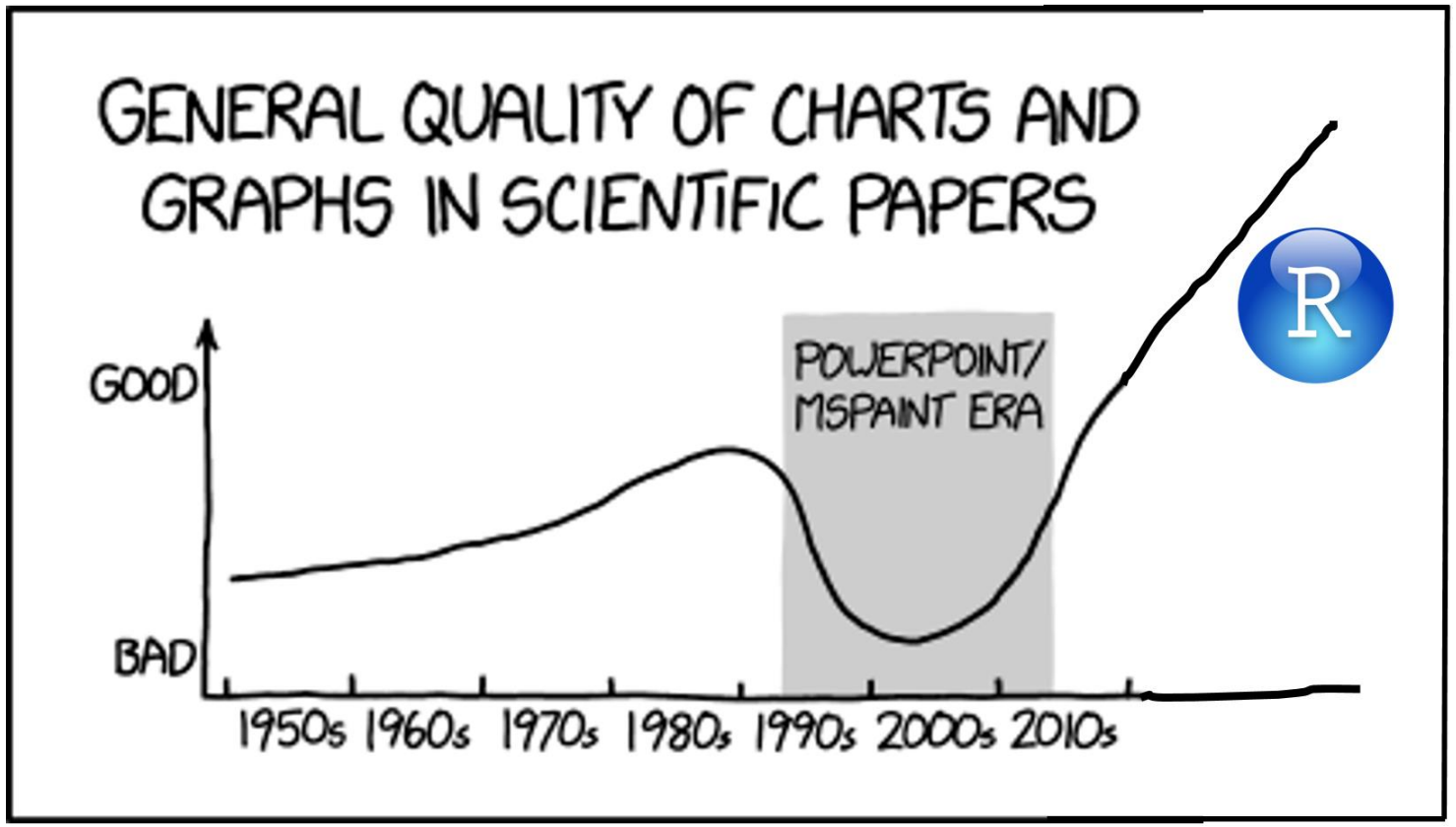
- Bruk **`by =`** for å spesifisere nøkkel variabel til å koble på

by = c("aar" = "year")

```
kobletdata <- left_join(datanavn1, datanavn2, by = variabelnavn)
```

- Flere variabler kan brukes for å koble på (som en vektor)

Plotting


GENERAL QUALITY OF CHARTS AND GRAPHS IN SCIENTIFIC PAPERS
GOOD
BAD
POWERPOINT/ MSPAINT ERA
R
1950s 1960s 1970s 1980s 1990s 2000s 2010s

Plotting med ggplot()

- **aes** : aesthetics, hvilke variabler
- **geom_** : hva slags figur
- **stat** : hva slags statistisk aggregat å presentere

Table 18-1 A Selection of Geoms and Associated Default Stats

Geom	*Description*	*Default Stat*
`geom_bar()`	Bar chart	`stat_bin()`
`geom_point()`	Scatterplot	`stat_identity()`
`geom_line()`	Line diagram, connecting observations in order by *x*-value	`stat_identity()`
`geom_boxplot`	Box-and-whisker plot	`stat_boxplot()`
`geom_path`	Line diagram, connecting observations in original order	`stat_identity()`
`geom_smooth`	Add a smoothed conditioned mean	`stat_smooth()`
`geom_histogram`	An alias for `geom_bar()` and `stat_bin()`	`stat_bin()`

Søylediagram

```
ggplot(aes(variabelnavn)) +
  geom_bar()
```

Bruke + for å legge til figurtype

Spesifisere variabelen

Spesifisere søylediagram

```
ggplot(aes=c(x = variabelnavn1, y = variabelnavn2)) +
  geom_bar(stat="identity")
```

Spesifisere x og y variablene

Spesifisere å bruke verdi

Eksempler på dapla

Øvelse 4

- Øvelsene til oppgavesett 1 er på fil: **øvelser_4.R**

Videre

- sparklyr: https://spark.rstudio.com/

Oppsummering

- Husk library( )
- Les inn filer: read_csv2( ) read_sas( )
- Ny variabel: mutate( )
- Velg noen linje: filter( )
- Aggregere/oppsummere: summarise( )
- Figur: ggplot( ), aes( ), geom_...( )

- https://wiki.ssb.no/display/s880/For+R+brukere
- Yammer: R i SSB
- Google