Seminar 2

Lise Rødland

March 15, 2021

På seminar to skal vi se på det følgende:

1. Organisering av arbeidet
2. Pakker
3. Laste inn data
4. Organisering av data
5. Klasser og målenivå
6. Deskriptiv statistikk
7. Plotting

Organisering av arbeidet

Når vi jobber R så kan vi organisere arbeidet vårt på to måter: 1. Bruke prosjekter 2. Sette working directory

Working directory angir den mappen vi ønsker å hente og lagre filer til. Du kan tenke på det som den mappen du lagret prosjektfilen din i på første seminar. Når du åpner prosjektfilen din så husker R hvilken mappe dette er, men dersom du åpner et vanlig script må du fortelle R det.

Her kan man enten bruke `setwd("filbane")` eller så kan man trykke seg frem via verktøylinjen. Da velger du Session -> Set Working Directory -> Choose Directory og klikker deg frem til mappen din.

Her er et eksempel på bruk av `setwd()`:

```
setwd("~/Dokumenter/STV1020")
```

I dette seminaret skal vi bruke et datasett fra European Social Survey Round 9 (2018), og dettedatasettet inneholder svarene fra norske respondenter. Filen heter "ESS9NO.dta" og ligger i Canvas.

Pass på at du har lagret datasettet vi skal bruke i dag i samme mappe som den du har satt som working directory.

Overskrifter og tekst

Hvordan man organiserer et R-script kommer an på hva man selv synes er mest oversiktlig, men det er viktig at man klarer å holde oversikt over hva man har kodet og forstår hva man har gjort når man kommer tilbake til et script.

Det er lurt å lage overskrifter for å huske hva du tenker at koden din skal gjøre. Dersom du velger overskrift-formatet "# Overskrift —-" så vil R automatisk gi deg muligheten til å gjemme koden under overskriften.

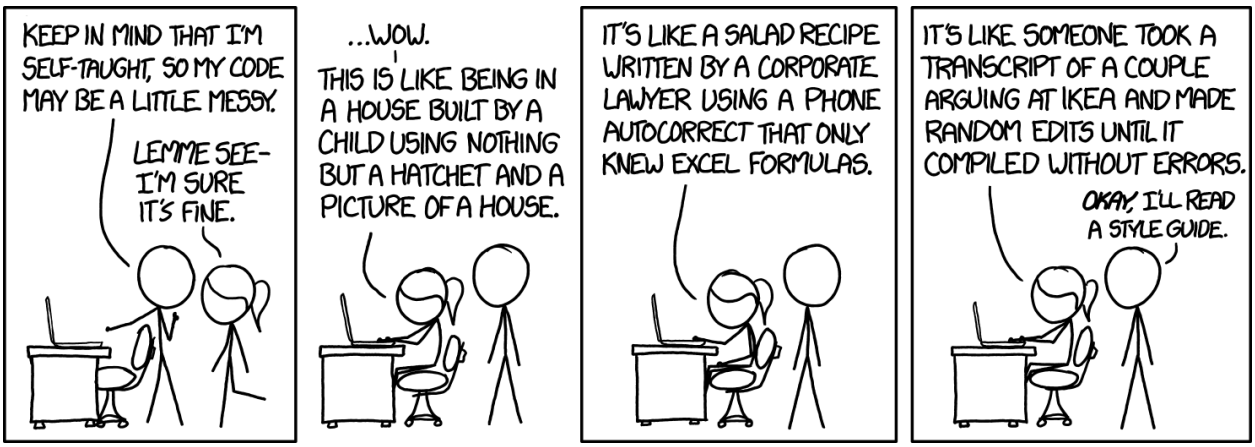


Figure 1: Ryddig kode gjør det enklere for deg selv og andre når du står fast. Tegneserie fra XKCD: https://xkcd.com/1513/

Om du trykker på "Show document outline" i menylinjen til høyre over scriptet ditt kan du også få opp en innholdsfortegnelse basert på overskriftene dine.

Det kan også være lurt å inkludere kommentarer i scriptet ditt som forklarer hva du gjør. Tekst som ikke skal leses av R skriver man etter emneknagg (#). Vi glemmer fort så det er en god idé å tenke at du skal kunne se tilbake på dette scriptet om et år igjen og skjønne hva du har gjort.

Til sist så er det lurt å ikke skrive for mange tegn før du bytter linje. The tidyverse style guide anbefaler at en begrenser antall tegn til 80 per linje. R teller for hvor mange tegn du har per linje til venstre nedenfor scriptet ditt. Tidyverse sin stilguide inneholder også flere tips til hvordan man kan skrive lettleselig kode.

Pakker

R-pakker er utvidelser til programmeringsspråket R. De inneholder kode, data, og dokumentasjon som gir oss tilgang til funksjoner som løser ulike problemer og gjør koding enklere. Første gang man skal bruke en pakke må man installere den. Etter at vi har installert pakken så må vi "hente den fra biblioteket" for å fortelle R at vi ønsker å bruke pakken. Dette må vi gjøre hver gang vi åpner R på nytt og ønsker å bruke pakken.

Den første pakken vi skal installere er Tidyverse. Tidyverse er et sett med pakker som gjør databehandling mye, mye enklere. Føst installerer vi pakken. Om dere har gjort dette på forhånd trenger dere ikke gjøre dette på nytt. Å installere gjør vi kun en gang. Vi installerer pakker ved hjelp av funksjonen `install.packages("pakkenavn")`:

```
install.packages("tidyverse")
```

Det neste vi gjør er å laste inn pakken ved hjelp av `library()`:

```
library(tidyverse)
```

```
## -- Attaching packages ---------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Merk at pakkenavnet ikke står i hermetegn når vi bruker `library()`. Hermetegn rundt pakkenavnet er bare nødvendig når vi bruker `install.packages()`.

Laste inn data

Dersom dere skal gjøre statistisk analyse, er som regel den første seksjonen import og forberedelse av data. En styrke ved R, er at det er mulig å importere mange ulike filtyper, både fra en mappe på pcen din og fra en url på internett. Det er også mulig å ha flere datasett oppe i R samtidig. Jeg går gjennom import av filer fra excel, stata, spss og R, men vit at det finnes mange andre muligheter. Hvis man lurer på hvordan man skal laste inn en bestemt filtype og har glemt hvordan man gjør det så er dette veldig lett å finne på internett

Når du skal laste inn eller lagre noe lokalt på pc-en så vil R til enhver tid forvente at filnavn du refererer til befinner seg i working directory. Som vi husker så er working directory en mappe på pcen din; enten den du har lagret prosjektet ditt i eller den du har satt som working directory ved hjelp av `setwd()`. For å sjekke hva nåværende working directory er, og hvilke filer som finnes i den mappen, kan du skrive følgende kode (jeg har gjemt egen output):

```
getwd()
```

```
## [1] "C:/Users/liserod/OneDrive - Universitetet i Oslo/Projects/Undervisning/STV1020/doc/seminar2"
```

```
list.files()
```

```
## [1] "seminar2.md"    "seminar2.pdf"   "seminar2.Rmd"   "seminar2_files"
```

Datasett kommer i mange ulike filformater. Noen vanlige formater er csv, dta (Stata-datasett), sav (SPSS-datasett) og Rdata. Hvilket format dataene dine har bestemmer hvilken funksjon du må bruke for å laste inn datasettet. For det meste så følger funksjonene dette formatet:

```
# Laster inn og lagrer datasettet som et objekt:
datasett <- read_filtype("filbane/filnavn.filtype")
```

For eksempel så er datasettet vi skal bruke i dag en dta-fil. Pakken `foreign` inneholder funksjoner for å lese dta-filer og sav-filer. Det første vi gjør er derfor å installere og laste inn `foreign`. Et alternativ til `foreign` er `haven`.

```
install.packages("foreign")
```

```
library(foreign)

ess <- read.dta("../../data/ESS9NO.dta")
```

Her er eksempler på noen andre funksjoner for å laste inn data:

```
# For .csv:
read_csv("data/filnavn.csv")

# For filer i R-format:
load("data/filnavn.Rdata")
```

For å laste inn en excel-filer bruker vi pakken `readxl`:

```
install.packages("readxl")

library(readxl)

df <- read_excel("data/filnavn.xlsx")
```

Organisering av data

Når man bruker større datasett som ESS, så inneholder datasettet ofte mange flere variabler enn de vi ønsker å bruke i våre analyser og variablene har navn som kan være vanskelig å huske, f.eks. nwspol. For å finne ut hvilken informasjon variablene inneholder så kan vi slå opp i kodeboken. Kodeboken til ESS finner dere her. ESS inneholder også mange landspesisifkke variabler og kodeboken er derfor veldig lang.

Når vi skal jobbe videre med data så kan det være lurt å fjerne de variablene vi ikke skal bruke og gi variablene navn som er lette for oss å forstå og huske. For å gjøre dette skal vi benytte oss av funksjoner i `tidyverse`. Først bruker vi `select()` til å velge de variablene vi vil beholde og så bruker vi `rename` til å endre navnene. Vi bruker en pipe `%>%` mellom funksjonene, som tar outputen til et utsagn og gjør det til inputen til det neste utsagnet. Pipen kan sees på som ordet "så". Rename-funksjonen lar oss forandre navnet til variabler og bruker syntaksen nytt_navn = gammelt_navn.

```
ess_subset <- ess %>%
  select(nwspol, polintr, vote, yrbrn) %>%
  rename(
        news = nwspol,
        interest = polintr,
        year_born = yrbrn
  )
```

Dersom du har sjekket kodeboken på forhånd så kan du også endre variablenavnene når du bruker `select()`:

```
ess_subset <- ess %>%
  select(
    vote,
    news = nwspol,
    interest = polintr,
    year_born = yrbrn
    )
```

Vi kan regne oss frem til alder ved å bruke variabelen `year_born`og informasjonen om at undersøkelsen ble gjennomført i 2018:

```
ess_subset$age <- 2018 - ess_subset$year_born
```

Klasser og målenivå

I forelesning gikk dere gjennom tre ulike målenivåer; kategorisk, ordinalt og kontinuerlig. Disse sammen faller til en viss grad med noen av klassene i R. Men det er viktig å huske at R ikke vet hvilken klasse en variabel har så dere kan ikke nødvendigvis bruke dette til å sjekke variabelens målenivå. Det viktige er at dere gir R riktig informasjon om hvilket målenivå en variabel har.

Kategorisk

Når variabler er kategoriske så kan egenskapen deles i to eller flere gjensidig utelukkende kategorier. I ESS datasettet vårt er variabelen "vote" kategorisk; man har enten stemt, ikke stemt, eller så er man ikke berettiget til å stemme. Dette kan vi se i utklippet fra kodeboken.

vote - Voted last national election

Type	Code
vote	Some people don't vote nowadays for one reason or another. Did you vote in the last [country] national election in [month/year]?
Location	B13
Question	Some people don't vote nowadays for one reason or another. Did you vote in the last [country] national election in [month/year]?

1	Yes
2	No
3	Not eligible to vote

Figure 2: Utdrag fra ESS sin kodebok for variabelen vote

Vi kan sjekke hvilken klasse variabelen har ved hjelp av `class()`

```
class(ess_subset$vote)
```

```
## [1] "factor"
```

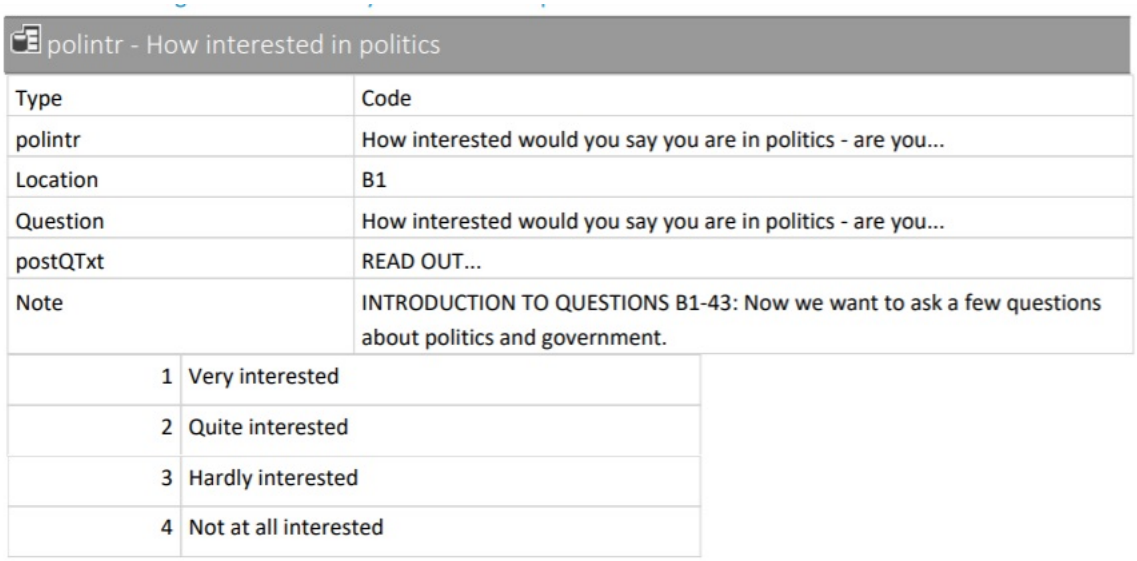
Her får vi opp klassen "factor". Klassen i R samsvarer altså med variabelens målenivå. Vi kan bruke `levels()` til å sjekke hvilke nivåer eller kategorier som er registrert:

```
levels(ess_subset$vote)
```

```
## [1] "Yes"                  "No"                   "Not eligible to vote"
## [4] "Refusal"              "Don't know"           "No answer"
```

Ordinalnivå

Når variabler er på ordinalnivå kan de deles i to eller flere gjensidig utelukkende kategorier som kan rangeres, men vi kan ikke si noe om avstanden mellom verdiene og en enhets økning har ikke samme betydning. I ESS datasettet vår så er variabelen `interest` et eksempel på en variabel på ordinalnivå; i utdraget fra kodeboken ser vi at man kan være ikke interessert, lite interessert, ganske interessert, eller veldig interessert i politikk.



polintr - How interested in politics

Type	Code
polintr	How interested would you say you are in politics - are you...
Location	B1
Question	How interested would you say you are in politics - are you...
postQTxt	READ OUT...
Note	INTRODUCTION TO QUESTIONS B1-43: Now we want to ask a few questions about politics and government.

1	Very interested
2	Quite interested
3	Hardly interested
4	Not at all interested

Figure 3: Utdrag fra ESS sin kodebok for variabelen interest (opprinnelig navn polintr)

```
class(ess_subset$interest)
```

```
## [1] "factor"
```

```
is.numeric(ess_subset$interest)
```

```
## [1] FALSE
```

Som vi ser er også denne variabelen registrert som en faktor av R. Igjen så kan vi bruke levels:

```
levels(ess_subset$interest)
```

```
## [1] "Very interested"       "Quite interested"      "Hardly interested"
## [4] "Not at all interested" "Refusal"               "Don't know"
## [7] "No answer"
```

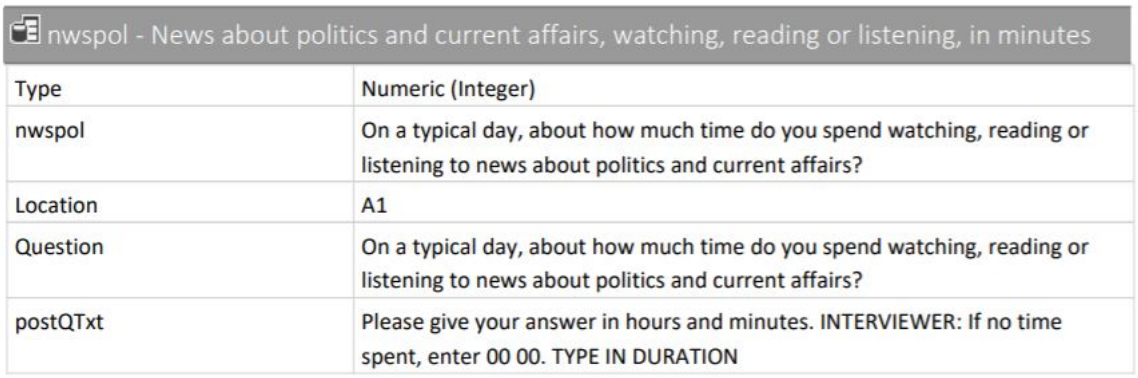
Kontinuerlig

Kontinuerlige variabler kan rangeres, har samme avstand mellom alle verdier og en enhets økning betyr alltid det samme. Her er det altså snakk om variabler med faktiske tallverdier. I ESS datasettet vårt så er variabelen "news" kontinuerlig. Som vi kan se i utdraget fra kodeboken så måler variabelen hvor mange minutter man bruker på nyheter hver dag. Det er et minutts avstand mellom hver verdi, og en økning på en enhet vil alltid bety en økning på et minutt.

Vi kan sjekke klassen her også:

```
class(ess_subset$news)
```

```
## [1] "integer"
```



nwspol - News about politics and current affairs, watching, reading or listening, in minutes

Type	Numeric (Integer)
nwspol	On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs?
Location	A1
Question	On a typical day, about how much time do you spend watching, reading or listening to news about politics and current affairs?
postQTxt	Please give your answer in hours and minutes. INTERVIEWER: If no time spent, enter 00 00. TYPE IN DURATION

Figure 4: Utdrag fra ESS sin kodebok for variabelen news (opprinnelig navn nwspol)

```
is.numeric(ess_subset$news)
```

```
## [1] TRUE
```

Denne variabelen er numerisk og skal være det så her er alt i orden.

Jeg vil oppfordre dere til å være obs og alltid sjekke at klassen på en variabel dere skal bruke stemmer overens med målenivået. I mange datasett får kategorisk variabler ofte tall istedenfor kategorinavn som verdier og lastes inn som klassen numeric. Dette gjør at kategoriske variabler kan fremstå som at de har et høyere målenivå enn de faktisk har i R. Derfor er det alltid viktig å også sjekke kodeboken for å se hvilket målenivå variabelen faktisk har. Det kommer ikke til å stå "denne variabelen har kategorisk målenivå" så dere må gjøre en selvstendig vurdering basert på hvilke verdier variabelen har.

Utforske data

Det er mange ulike måter å utforske datasett og variabler på. Vi skal se på funksjonene: `summary()`, `str()`, `head()` og `tail()`.

For å få et deskriptivt sammendrag av et objekt kan vi bruke `summary()` eller str().

```
summary(ess_subset)
```

```
##                    vote           news                          interest
##  Yes                 :1156   Min.   :   0.0   Hardly interested    :568
##  No                  : 124   1st Qu.:  30.0   Quite interested    :564
##  Not eligible to vote: 125   Median :  60.0   Very interested     :184
##  Refusal             :   0   Mean   : 104.1   Not at all interested: 89
##  Don't know          :   0   3rd Qu.: 120.0   Refusal             :  0
##  No answer           :   0   Max.   :1109.0   (Other)             :  0
##  NA's                :   1   NA's   :34       NA's                :  1
##    year_born         age
##  Min.   :1928   Min.   :15.00
##  1st Qu.:1957   1st Qu.:32.00
##  Median :1971   Median :47.00
##  Mean   :1971   Mean   :46.54
##  3rd Qu.:1986   3rd Qu.:61.00
```

```
##  Max.   :2003   Max.   :90.00
##  NA's   :32     NA's   :32
```

```
str(ess_subset)
```

```
## 'data.frame':    1406 obs. of  5 variables:
##  $ vote     : Factor w/ 6 levels "Yes","No","Not eligible to vote",..: 1 1 1 1 1 1 1 3 1 1 ...
##  $ news     : int  60 60 540 30 60 120 60 90 120 60 ...
##  $ interest : Factor w/ 7 levels "Very interested",..: 3 2 2 2 2 1 2 3 2 2 ...
##  $ year_born: int  1961 1960 1956 1967 1972 1964 1959 2000 1950 1975 ...
##  $ age      : num  57 58 62 51 46 54 59 18 68 43 ...
```

Hvis man vil se de første eller siste radene i et datasett, kan man bruke henholdsvis head- og tail-funksjonene. Man kan også velge for eksempel å bare se på de første eller siste verdiene til en bestemt variabel.

```
head(ess_subset)
```

```
##   vote news          interest year_born age
## 1  Yes   60 Hardly interested      1961  57
## 2  Yes   60  Quite interested      1960  58
## 3  Yes  540  Quite interested      1956  62
## 4  Yes   30  Quite interested      1967  51
## 5  Yes   60  Quite interested      1972  46
## 6  Yes  120   Very interested      1964  54
```

```
tail(ess_subset)
```

```
##                      vote news          interest year_born age
## 1401                  Yes   90   Very interested      1955  63
## 1402 Not eligible to vote   20  Quite interested      2003  15
## 1403                  Yes   30 Hardly interested      1994  24
## 1404                  Yes   60  Quite interested      1984  34
## 1405                   No   NA Hardly interested      1974  44
## 1406                  Yes   30 Hardly interested      1988  30
```

Alle disse kan også bruker på enkeltvariabler.

Deskriptiv statistikk

Som dere husker fra forelesning og fra kapittel seks i Kellsted og Whitten så er det variabelens målenivå som avgjør hvilken deskriptiv statistikk som er fornuftig.

Kategoriske variabler

R har ingen innebygd funksjon for å finne modusverdien. Ved å søke på internett så finner du fort mange ulike funksjoner du kan bruke, men for å gjøre det enkelt bruker vi bare `table()`. Funksjonen `table()` gir oss en frekvenstabell, mens `prop.table` gjør om frekvenstabellen til andeler. ESS datasettet mangler data for noen observasjoner. ved å ta med `useNA = "always"` så får vi ogås denne informasjonen i tabellen:

```r
table(ess_subset$vote, useNA = "always")
```

```
##
##                  Yes                   No Not eligible to vote
##                 1156                  124                  125
##              Refusal          Don't know           No answer
##                    0                    0                    0
##                 <NA>
##                    1
```

```r
prop.table(table(ess_subset$vote))
```

```
##
##                  Yes                   No Not eligible to vote
##           0.82277580           0.08825623           0.08896797
##              Refusal          Don't know           No answer
##           0.00000000           0.00000000           0.00000000
```

```r
prop.table(table(ess_subset$vote, useNA = "always"))
```

```
##
##                  Yes                   No Not eligible to vote
##         0.8221906117         0.0881934566         0.0889046942
##              Refusal          Don't know           No answer
##         0.0000000000         0.0000000000         0.0000000000
##                 <NA>
##         0.0007112376
```

Kontinuerlige variabler

```r
# Finner minimumsverdi (det laveste antall minutter brukt på nyheter)
min(ess_subset$news, na.rm = TRUE) # na.rm = TRUE sier at missing skal droppes i beregningen
```

```
## [1] 0
```

```r
# Finner maksimumsveriden (den høyeste antall minutter brukt på nyheter)
max(ess_subset$news, na.rm = TRUE)
```

```
## [1] 1109
```

```r
# Finner gjennomsnittlig antall minutter
mean(ess_subset$news, na.rm = TRUE)
```

```
## [1] 104.1006
```

```r
# Finner median
median(ess_subset$news, na.rm = TRUE)
```

```
## [1] 60
```

```r
# Finner standardavviket
sd(ess_subset$news, na.rm = TRUE)
```

```
## [1] 155.5571
```

```r
# Finner varians
var(ess_subset$news, na.rm = TRUE)
```

```
## [1] 24198.01
```

```r
# Finner kvantilverdiene
quantile(ess_subset$news, na.rm = TRUE)
```

```
##   0%  25%  50%  75% 100%
##    0   30   60  120 1109
```

```r
# Finner forskjellig deskriptiv statistikk for en variabel
summary(ess_subset$news)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    30.0    60.0   104.1   120.0  1109.0      34
```

Plotting

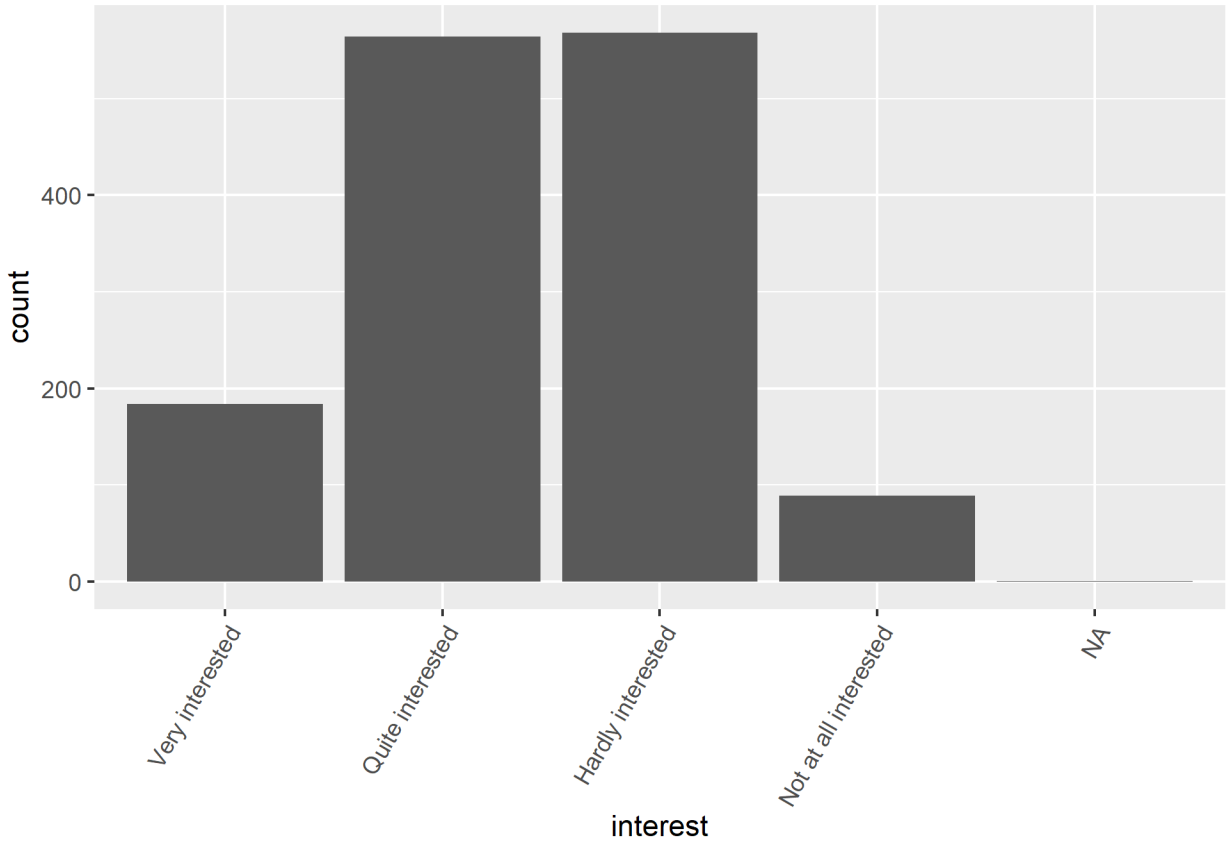
Vi skal kort introdusere hvordan man kan visualisere data i dette seminaret, og så vil dere få en mer grundig gjennomgang neste seminar. Det er gøy å kunne visualisere dataene våre, både for vår egen del, men også for de som skal lese oppgavene våre. For å få fine grafer kan man bruke funksjonen `ggplot()`.

Kategoriske variabeler

Søylediagram og kakediagram med en variabel

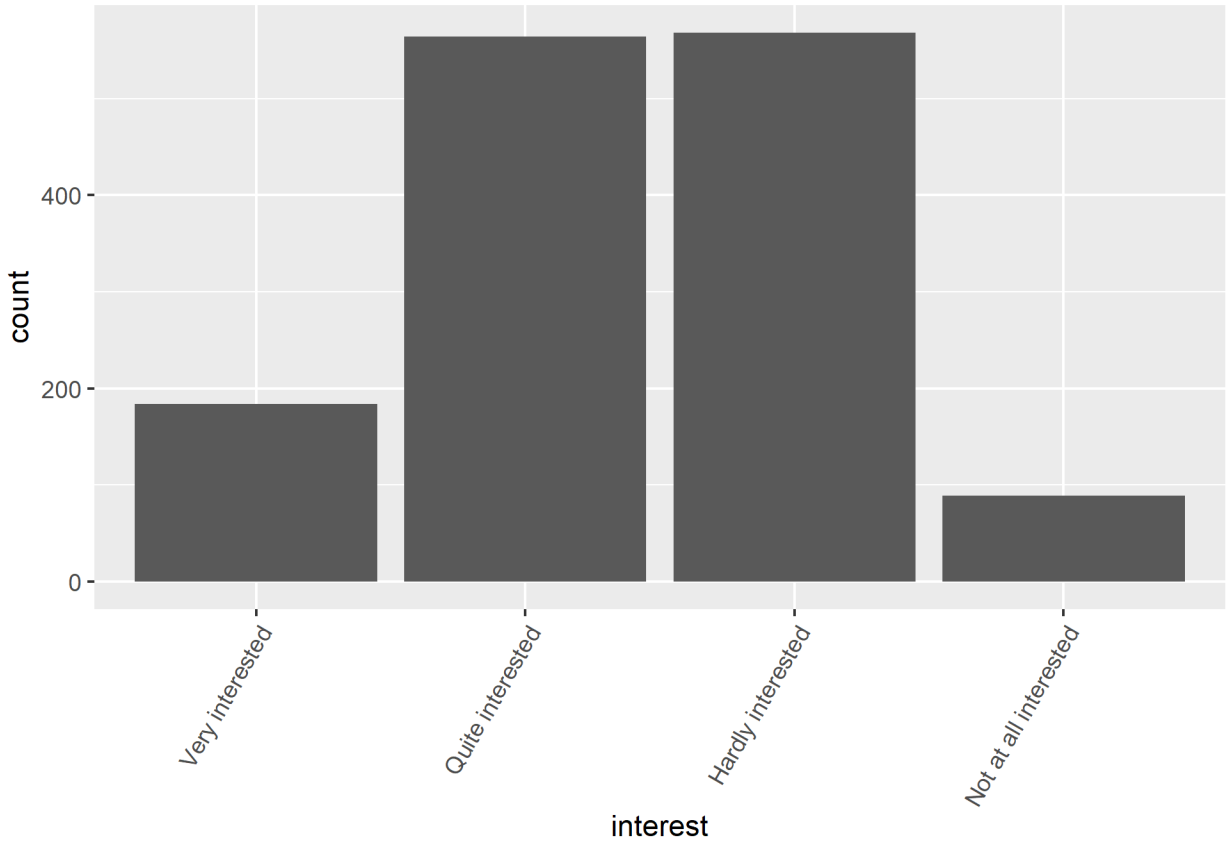
Hvordan kan vi visualisere hvordan fordelingen av politisk interesse er? Her kan vi bruke `geom_bar` til å lage et søylediagram (bar chart). Et søylediagram viser antall observasjoner av hver verdi.

```r
ggplot(data = ess_subset, aes(x = interest)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



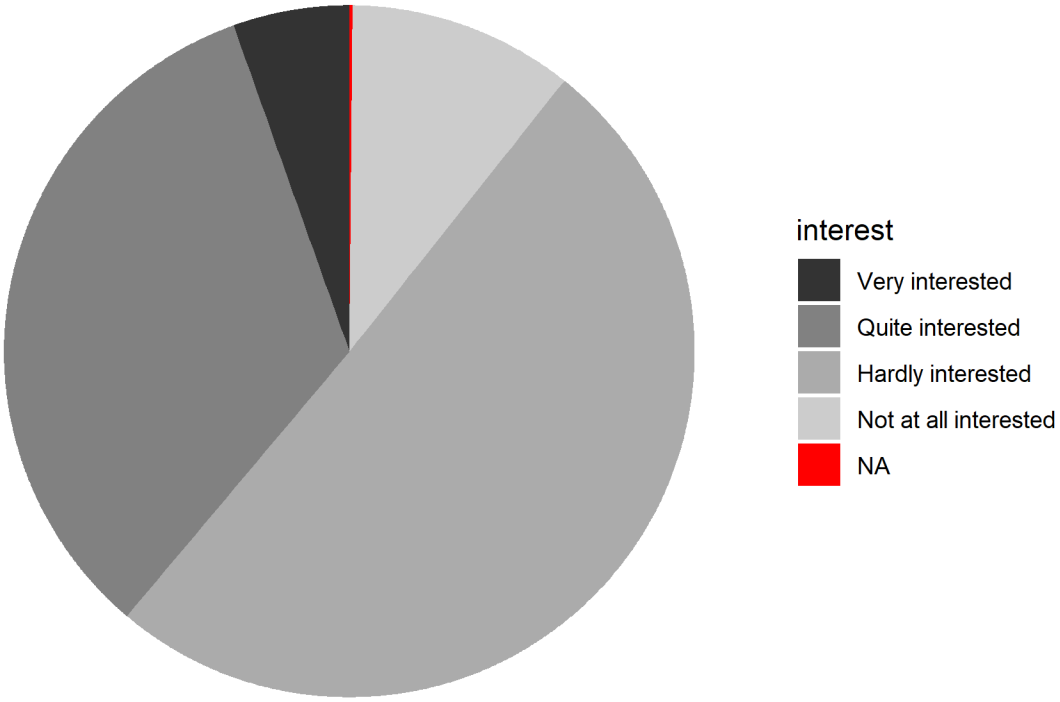
Dersom vi ikke ønsker å gi missingverdiene (NA) en egen søyle så kan vi bruke `filter()` til å fjerne disse:

```
ggplot(data = ess_subset %>%
         filter(!is.na(interest)),
       aes(x = interest)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Et alternativ til søylediagram er kakediagram (pie chart):

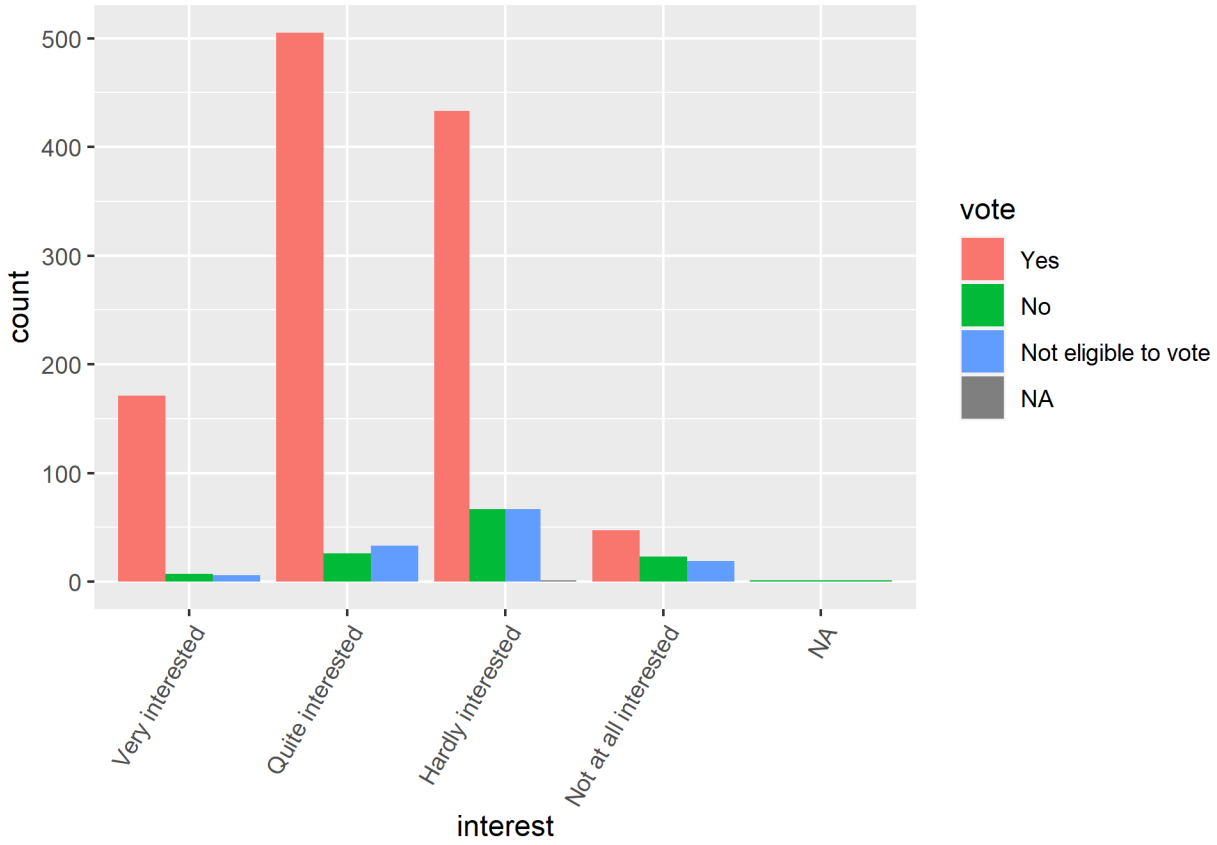
```
ggplot(ess_subset, aes(x = "", y = interest, fill = interest)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  theme_void() +
  scale_fill_grey()
```



Søylediagram med to variabler

Hvor mange innenfor hvert nivå av politisk interesse stemte? Vi kan bruke `geom_bar()` igjen, men vi sier at vi også vil se fordelingen av hvordan respondentene stemte innenfor hvert nivå av politisk interesse med `(aes(fill = vote2))`. Så sier vi at vi vil at det skal være en søyle for de ulike alternativene for vote med `position = "dodge"`.

```
ggplot(data = ess_subset,
       aes(x = interest)) +
  geom_bar(aes(fill=vote),
           position = "dodge") +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

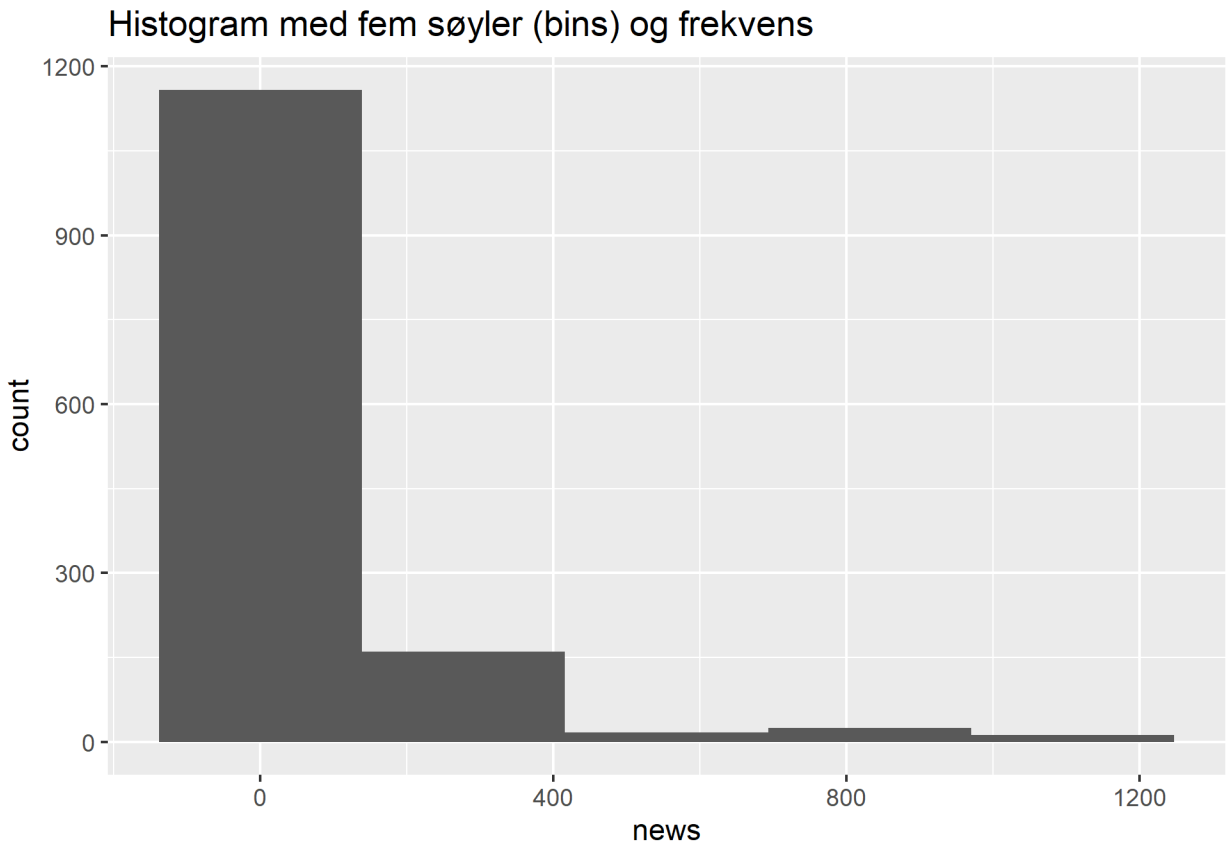


Kontinuerlige variabler

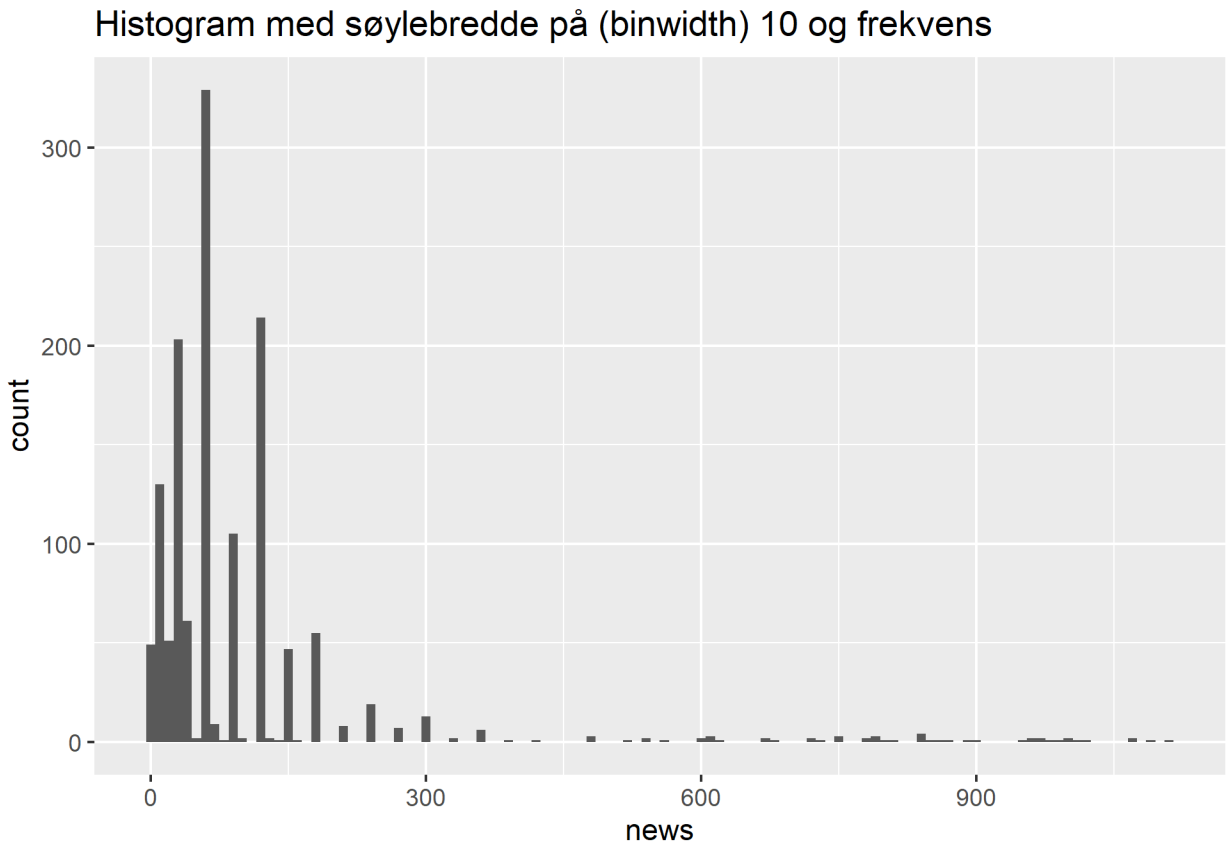
Histogram

Hvordan fordeler respondentenes alder og tiden de bruker på nyheter seg? Disse variablene er kontinuerlige, så vi kan bruke `geom_histogram` for å lage et histogram. Her gjør jeg det med variabelen news.

```
ggplot(data = ess_subset, aes(x = news)) +
      geom_histogram(bins = 5) +
  ggtitle("Histogram med fem søyler (bins) og frekvens")
```

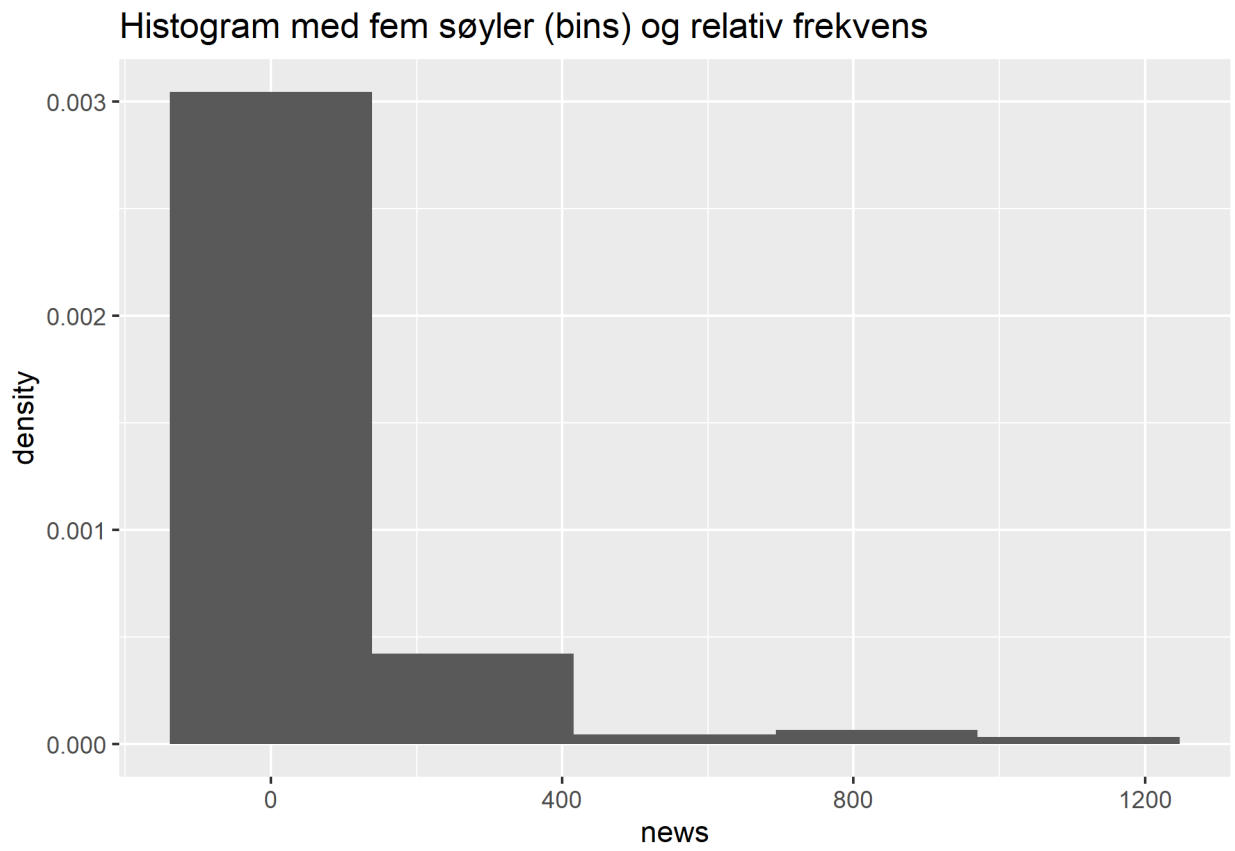



```
ggplot(data = ess_subset, aes(x = news)) +
  geom_histogram(binwidth = 10) +
  ggtitle("Histogram med søylebredde (binwidth) på 10 og frekvens")
```

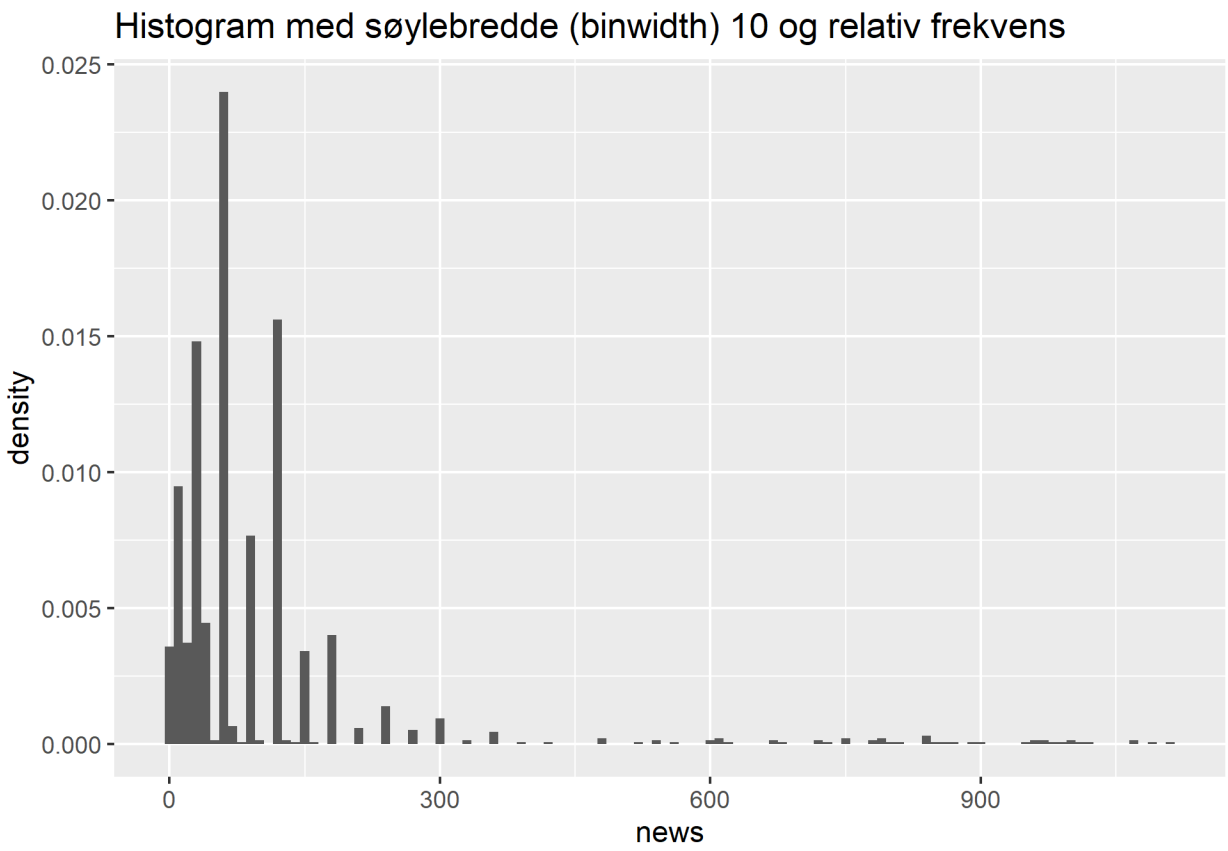


Et histogram viser hvor mange enheter det er i hver kategori. Vi kan enten spesifisere hvor mange søyler vi vil ha (bins) eller hvor stor hver søyle skal være (bindwidth). Vi kan også velge å plotte density fremfor count. Da får vi histogrammer tilsvarende figur 6.5 i Kellsted og Whitten:

```
ggplot(data = ess_subset, aes(x = news, y = ..density..)) +
      geom_histogram(bins = 5) +
  ggtitle("Histogram med fem søyler (bins) og relativ frekvens")
```



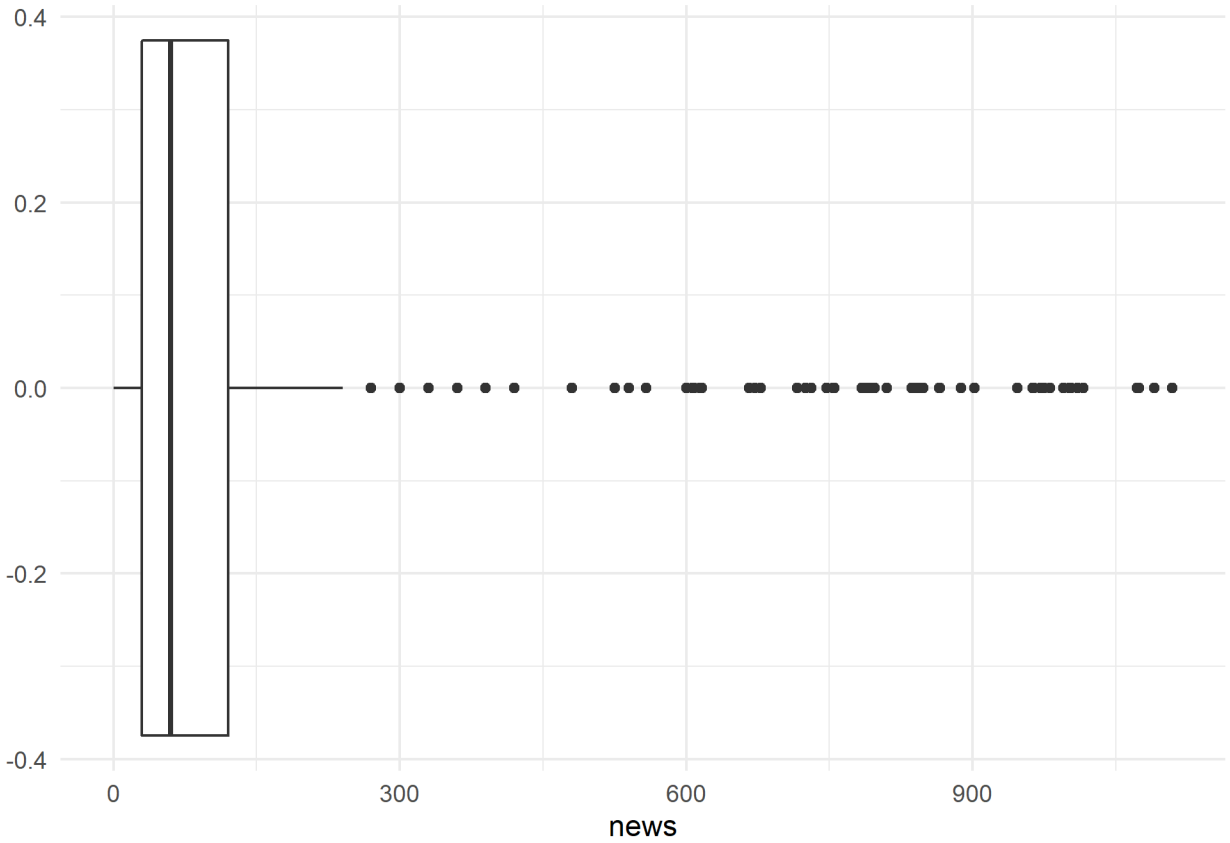

```
ggplot(data = ess_subset, aes(x = news, y = ..density..)) +
  geom_histogram(binwidth = 10) +
  ggtitle("Histogram med søylebredde (binwidth) 10 og relativ frekvens")
```



Boksplott

Hvordan fordeler alder seg på interesse? Vi kan lage et boksplott med `geom_boxplot`. Et boksplott kan vise hvordan en kontinuerlig variabel (alder) er fordelt innenfor en annen kategorisk variabel (politisk interesse). Boksen i midten representerer spennet til de 50% vanligste verdiene (andre og tredje kvartil), mens strekene viser spennet til verdiene i nedre og øvre kvartil.

```
ggplot(data = ess_subset, aes(x = news)) +
  geom_boxplot() +
  theme_minimal()
```



Hvis dere vil utforske hvordan man kan tilpasse de ulike diagrammene vi har sett på og mange andre, kan denne siden være nyttig: https://www.r-graph-gallery.com/index.html