62_SOK-1004 H24 - Case 4

Kandidat 62

Instruksjoner

Denne oppgaven er laget av Even S. Hvinden og oppdatert av Derek J. Clark. Sistnevnte er ansvarlig for eventuelle feil og mangler.

Oppgaven skal løses interaktivt i RStudio ved å legge inn egen kode og kommentarer. Det ferdige dokumentet lagres med kandidatnummeret som navn `[kandidatnummer]_SOK1004_C4_H24.qmd` og lastes opp på deres GitHub-side. Hvis du har kandidatnummer 43, så vil filen hete `43_SOK1004_C4_H22.qmd`. Påse at koden kjører og at dere kan eksportere besvarelsen til pdf. Lever så lenken til GitHub-repositoriumet i Canvas.

Bakgrunn, læringsmål

Innovasjon er en kilde til økonomisk vekst. I denne oppgaven skal vi se undersøke hva som kjennetegner bedriftene som bruker ressurser på forskning og utvikling (FoU). Dere vil undersøke FoU-kostnader i bedriftene fordelt på næring, antall ansatte, og utgiftskategori. Gjennom arbeidet vil dere repetere på innhold fra tidligere oppgaver og øve på å presentere fordelinger av data med flere nivå av kategoriske egenskaper.

Last inn pakker

```
# output | false
rm(list=ls())
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(rjstat)
```

```
Attaching package: 'rjstat'

The following object is masked from 'package:dplyr':

    id
```

```
library(gdata)
```

```
Attaching package: 'gdata'

The following objects are masked from 'package:dplyr':

    combine, first, last, starts_with

The following object is masked from 'package:purrr':

    keep

The following object is masked from 'package:tidyr':

    starts_with

The following object is masked from 'package:stats':

    nobs
```

```
The following object is masked from 'package:utils':

    object.size

The following object is masked from 'package:base':

    startsWith
```

```
library(httr)
```

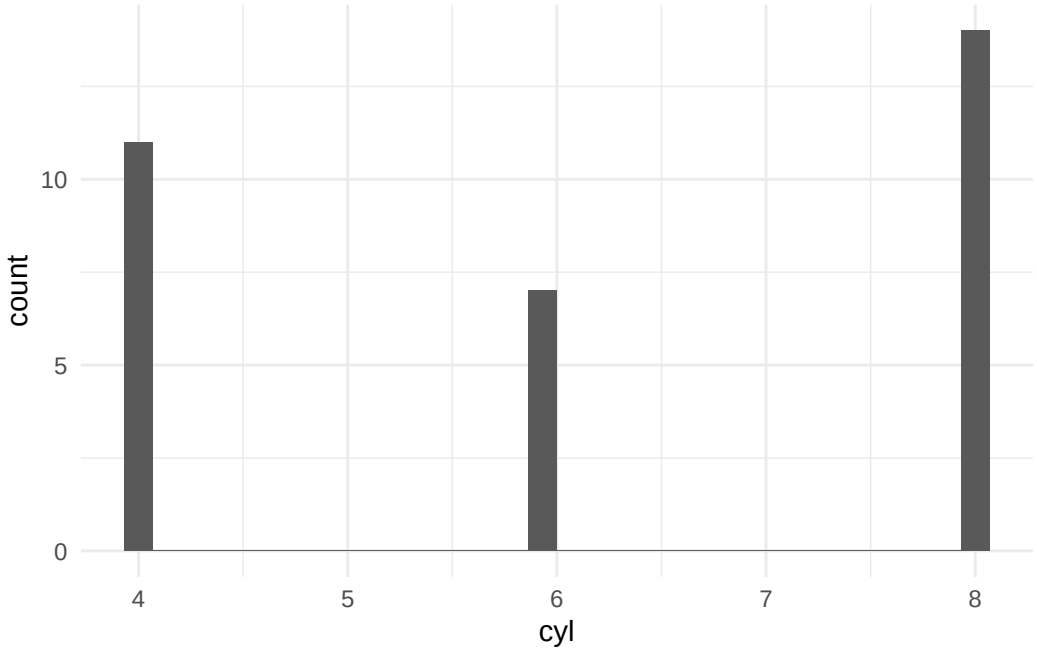
Oppgave I: Introduksjon til histogram

Et histogram eller frekvensfordeling er en figur som viser hvor ofte forskjellige verdier oppstår i et datasett. Frekvensfordelinger spiller en grunnleggende rolle i statistisk teori og modeller. Det er avgjørende å forstå de godt. En kort innføring følger. Du kan lese om histogram i R for Data Science, kap 1.4

La oss se på et eksempel. I datasettet `mtcars` viser variabelen `cyl` antall sylindere i motorene til kjøretøyene i utvalget.

```
data(mtcars)
mtcars %>%
  ggplot(aes(cyl)) +
  geom_histogram() +
  theme_minimal()
```

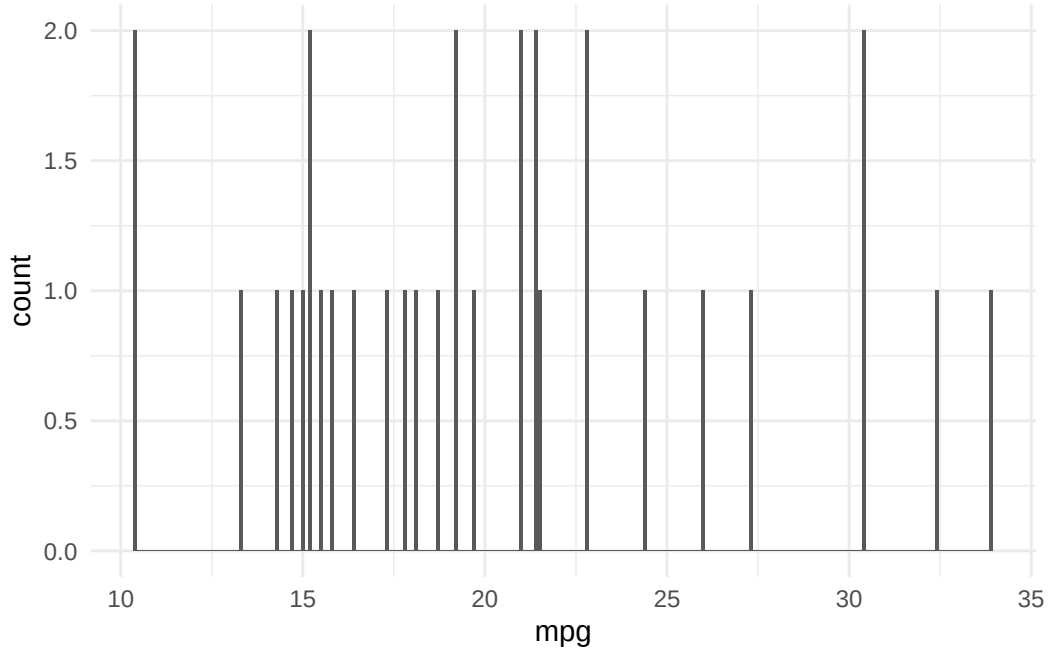
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Verdiene av variabelen er er gitt ved den horisontale aksen, antall observasjoner på den vertikale aksen. Vi ser at det er 11, 7, og 14 biler med henholdsvis 4, 6, og 8 sylindere.

La oss betrakte et eksempel til. Variabelen `mpg` i `mtcars` måler gjennomsnittlig drivstoffforbruk i amerikanske enheter. Variabelen er målt med ett desimal i presisjon.

```
data(mtcars)
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram(binwidth=0.1) +
  theme_minimal()
```



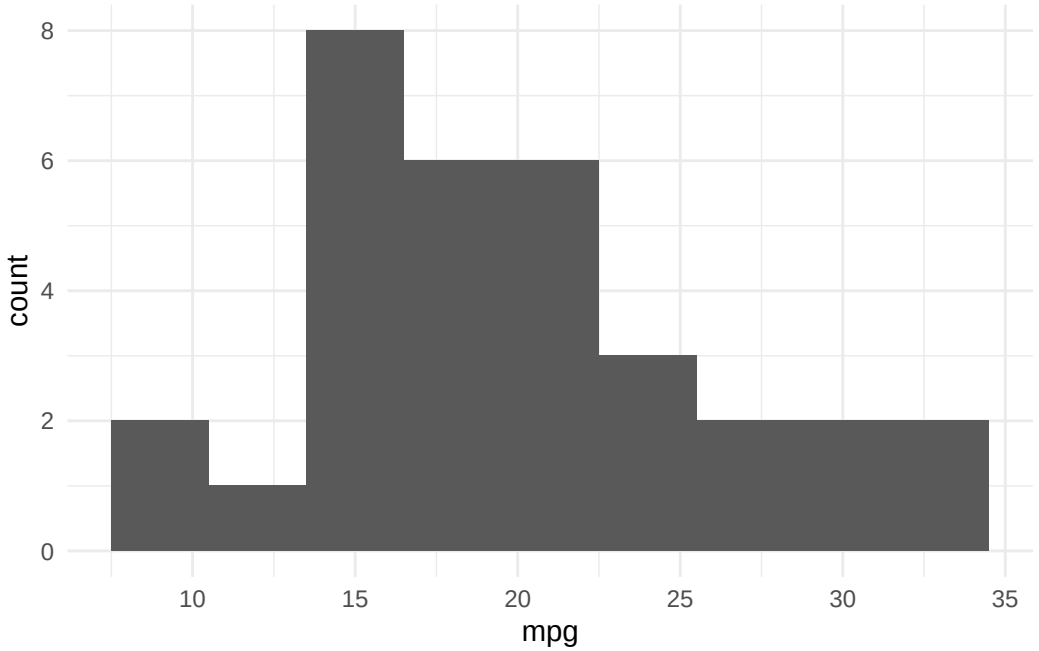
Datasettet inneholder mange unike verdier, hvilket gir utslag i et flatt histogram, noe som er lite informativt. Løsningen da er å gruppere verdier som ligger i nærheten av hverandre. Kommandoen `binwidth` i `geom_histogram()` bestemmer bredden av intervallene som blir slått sammen. Kan du forklare hvorfor alle unike verdier blir telt ved å å bruke `binwidth = 0.1`?

Eksperimenter med forskjellige verdier for `binwidth` og forklar hva som kjennetegner en god verdi.

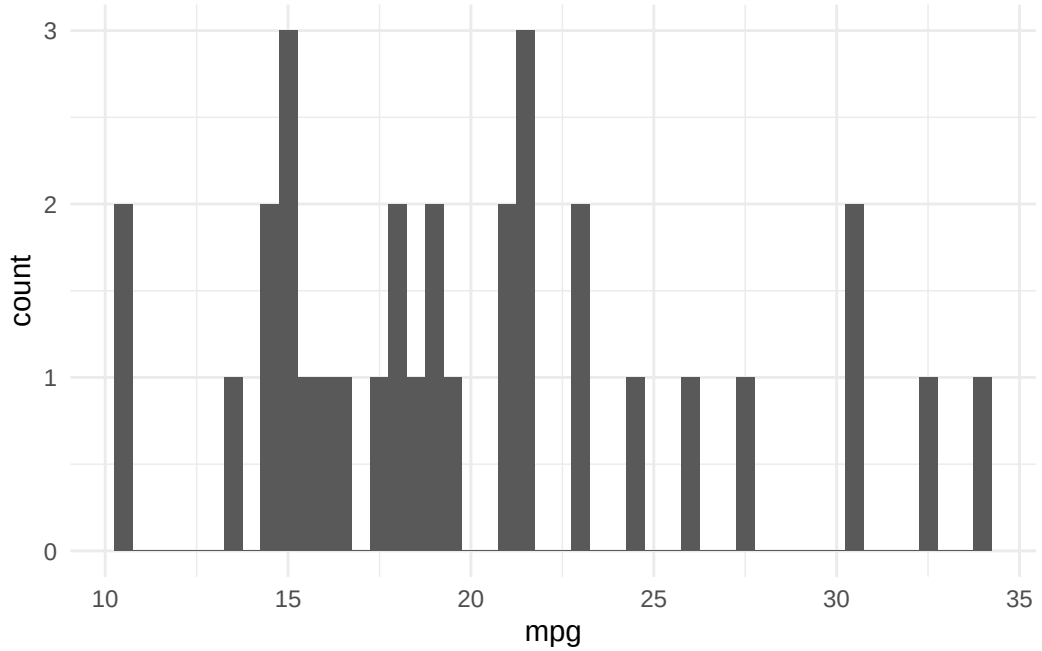
Svar:

Alle unike verdier blir telt ved binwidth 0.1 fordi da er søylene så små at ingen slås sammen. Dette er bra for detaljnivå, men kan være uoversiktlig og bli for mye informasjon. En god binwidth verdi finner balansen mellom detalj og oversiktlighet, slik at det blir enkelt å tyde, men ikke mister for mye detaljer.

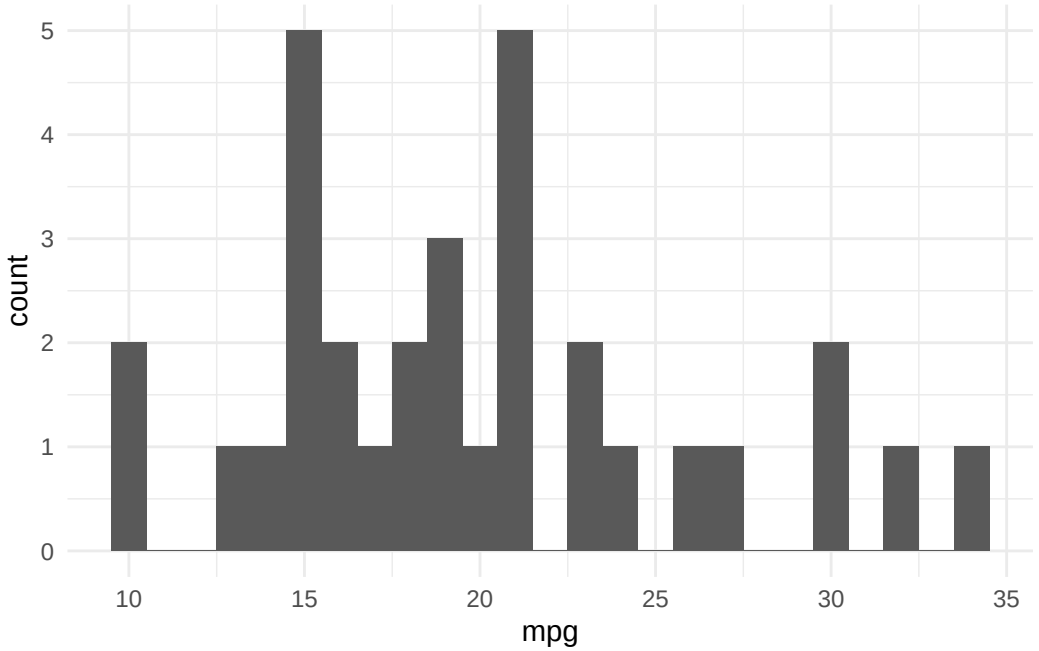
```
# løs oppgave I her

# Tester binwidth 3
data(mtcars)
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram(binwidth=3) +
  theme_minimal()
```




```
# Tester binwidth 0.5
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram(binwidth=0.5) +
  theme_minimal()
```




```
#Teskter binwidth 1.5
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram(binwidth=1) +
  theme_minimal()
```



Oppgave II: Last ned og rydd i data

Vi skal nå undersøkene dataene i Tabell 07967: Kostnader til egenutført FoU-aktivitet i næringslivet, etter næring (SN2007) og sysselsettingsgruppe (mill. kr) 2007 - 2022 SSB. Dere skal laster de ned ved hjelp av API. Se brukerveiledningen her.

Bruk en JSON-spørring til å laste ned alle statistikkvariable for alle år, næringer, og sysselsettingsgrupper med 10-19, 20-49, 50-99, 100-199, 200 - 499, og 500 eller flere ansatte. Lagre FoU-kostnader i milliarder kroner. Sørg for at alle variabler har riktig format, og gi de gjerne enklere navn og verdier der det passer.

Hint. Bruk lenken til SSB for å hente riktig JSON-spørring og tilpass koden fra case 3.

```
# besvar oppgave II her

#Henter url fra ssb
url <- "https://data.ssb.no/api/v0/no/table/07967/"

#Setter inn JSON
query <- '{
  "query": [
    {
```

```
      "code": "NACE2007",
      "selection": {
        "filter": "item",
        "values": [
          "A-N",
          "C",
          "G-N",
          "A-B_D-F"
        ]
      }
    },
    {
      "code": "SyssGrp",
      "selection": {
        "filter": "item",
        "values": [
          "10-19",
          "20-49",
          "10-49",
          "50-99",
          "100-199",
          "200-499",
          "500+"
        ]
      }
    }
  ],
  "response": {
    "format": "json-stat2"
  }
}'

hent_indeks.tmp <- url %>%
  POST(body = query, encode = "json")

df <-  hent_indeks.tmp %>%
  content("text") %>%
  fromJSONstat() %>%
  as_tibble()
```

Oppgave III: Undersøk fordelingen

Vi begrenser analysen til bedrifter med minst 20 ansatte og tall fra 2015 - 2022. Lag en figur som illustrerer fordelingen av totale FoU-kostnader fordelt på type næring (industri, tjenesteyting, andre) og antall ansatte i bedriften (20-49, 50-99, 100-199, 200-499, 500 og over). Tidsdimensjonen er ikke vesentlig, så bruk gjerne histogram.

Merknad. Utfordringen med denne oppgaven er at fordelingene er betinget på verdien av to variable. Kommandoen `facet_grid()` kan være nyttig til å slå sammen flere figurer på en ryddig måte.

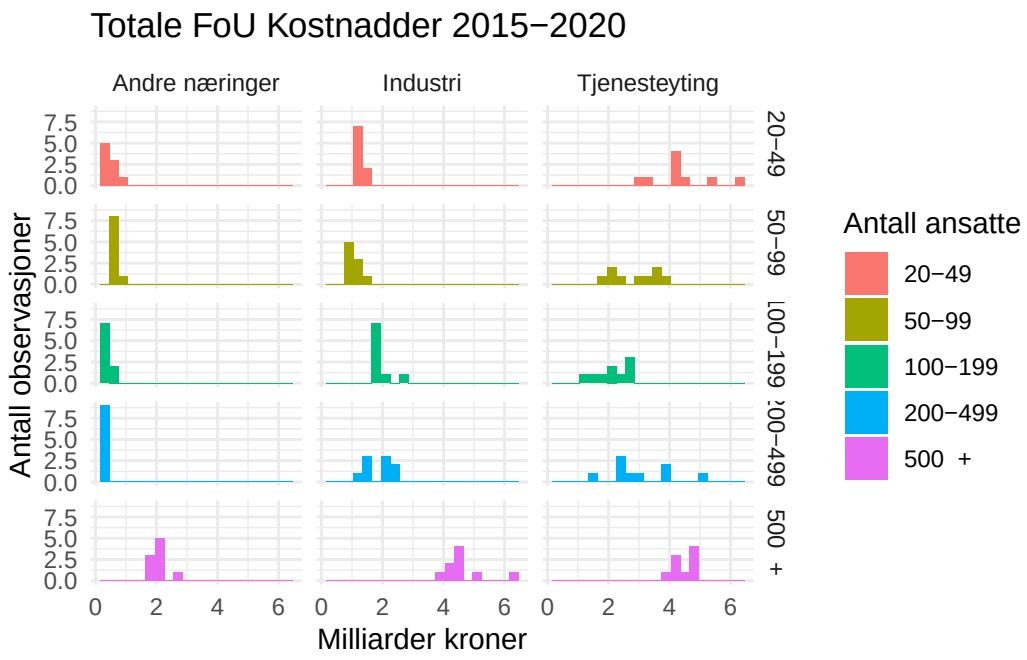
```
# besvar oppgave III her

# Setter til UTF-8 for å kunne bruke norske bokstaver
Sys.setlocale("LC_CTYPE", "nb_NO.UTF-8")
```

```
[1] "nb_NO.UTF-8"
```

```
# Gjør årstallene til integers
df$år<- df$år %>%
  as.integer()


# Navngir kolonnene på nytt
df <- df %>%
  rename(næring = `næring (SN2007)`,
         gruppe = sysselsettingsgruppe,
         var = statistikkvariabel,
         verdi = value) %>%

  # Deler verdien på 1000 for å gjøre det mer oversiktlig
  mutate(verdi = verdi/1000)

df <- df %>%

  # Fjerner sysselsatte
  mutate(gruppe = str_replace(gruppe, "sysselsatte", "")) %>%

  # Endrer "og over" til +
  mutate(gruppe = str_replace(gruppe, "og over", "+")) %>%

  # Endrer "lønnskostnader" til "lønn"
```

```r
  mutate(var = str_replace(var, "Lønnskostnader", "Lønn"))

# Ekstraherer gruppenavnene
gruppe_navn <- df$gruppe %>%
  unique()

# Ekstraherer næringsnavnene
næring_navn <- df$næring %>%
  unique()

# Ekstraherer variabelnavnene
var_navn <- df$var %>%
  unique()



df %>%

  # Filtrerer på år fra 2015 og senere
  filter(år >= 2015) %>%

  # Velger grupper
  filter(gruppe %in% c(gruppe_navn[2], gruppe_navn[4:7])) %>%

  # Velger næringer
  filter(næring %in% c(næring_navn[2:4])) %>%

  # Velger variabler
  filter(var %in% var_navn[1]) %>%

  # Sorterer gruppene i stigende rekkefølge
   mutate(gruppe = factor(gruppe, levels = gruppe_navn[2:7])) %>%

  # Genererer et plot
  ggplot(aes(x = verdi, fill = gruppe),alpha(0.5))+

  # Lager et histogram
  geom_histogram(binwidth = 0.3)+

  # Bruker facet_grid som anbefalt i oppgaveteksten
  facet_grid(gruppe ~ næring)+
  theme_minimal()+
```

```
  labs(x = "Milliarder kroner", y = "Antall observasjoner", title = "Totale FoU Kostnadder 20
```



Oppgave IV: Undersøk fordelingen igjen

Kan du modifisere koden fra oppgave III til å i tillegg illustrere fordelingen av FoU-bruken på lønn, innleie av personale, investering, og andre kostnader?

Merknad. Kommandoen `fill = [statistikkvariabel]` kan brukes i et histogram.

```
# besvar oppgave IV her

df %>%

  # Filtrerer på år fra 2015 og senere
  filter(år >= 2015) %>%

  # Velger grupper
  filter(gruppe %in% c(gruppe_navn[2], gruppe_navn[4:7])) %>%

  # Velger næringer
  filter(næring %in% c(næring_navn[2:4])) %>%
```

```r
# Velger variabler
filter(var %in% c(var_navn[3:6])) %>%
 mutate(gruppe = factor(gruppe, levels = gruppe_navn[2:7])) %>%

# Generer et plot
ggplot(aes(x = verdi, fill  = var))+

# Lager et histogram
geom_histogram(binwidth = 0.3)+
facet_grid(gruppe ~ næring, scales = "free_y")+
theme_minimal() +
labs(x = "Milliarder av kroner", y = "Antall observasjoner", fill = "Kostnadskategori")
```