R-seminar 4: Løsningsforslag

STV1020 Vår 2021

Uke 14

Variabler

- rik – Hvor viktig er det å være rik, ha penger og dyre ting? (1-6), 1 = helt enig, 6 = helt uenig
- alder – I antall år
- kjonn – Dikotom, 1 = mann, 2 = kvinne
- tillit – Tillit til politikere, (1-10), 1 = ikke noe tillit, 10 = full tillit
- valg – Hvilket parti stemte du ved forrige stortingsvalg
- redusere – Er du enig i at regjeringen skal omfordele goder. (1-5), 1 = veldig enig, 5 = helt uenig.

1. Last inn datasettet "VALGDATA.Rdata" og last inn pakken "tidyverse".

```
# Laster inn datasettet fra working directory
load("VALGDATA.Rdata")

# Bytter navn til data
data <- nyedata

# Fjerner den andre
rm(nyedata)

# Henter pakke fra biblioteket
library(tidyverse)
```

2. Hva er navnene på variablene i datasettet?

```
# Henter ut navnene på variablene i datasettet
names(data)
```

```
## [1] "rik"      "alder"    "kjonn"    "tillit"   "valg"     "redusere"
```

3. Hvor mang missing er det totalt i datasettet?

```r
# Sjekker antall missing i datasettet
sum(is.na(data))
```

```
## [1] 377
```

4. Hvor mange missing er det på hver enkelt variabel?

```r
# Sjekker antall missing på alle variablene i datasettet
summary(data)
```

```
##       rik            alder          kjonn           tillit
##  Min.   :1.000   Min.   :15.0   Min.   :1.000   Min.   : 0.00
##  1st Qu.:4.000   1st Qu.:32.0   1st Qu.:1.000   1st Qu.: 4.00
##  Median :5.000   Median :47.5   Median :1.000   Median : 5.00
##  Mean   :4.558   Mean   :47.1   Mean   :1.447   Mean   : 5.28
##  3rd Qu.:5.000   3rd Qu.:61.0   3rd Qu.:2.000   3rd Qu.: 7.00
##  Max.   :6.000   Max.   :90.0   Max.   :2.000   Max.   :10.00
##  NA's   :7       NA's   :32                     NA's   :8
##      valg             redusere
##  Length:1406        Min.   :1.000
##  Class :character   1st Qu.:2.000
##  Mode  :character   Median :2.000
##                     Mean   :2.194
##                     3rd Qu.:3.000
##                     Max.   :5.000
##                     NA's   :7
```

```r
# Eventuelt en og en
sum(is.na(data$redusere))
```

```
## [1] 7
```

```r
sum(is.na(data$valg))
```

```
## [1] 323
```

```r
# osv...
```

Variabler som har målenivå "character" viser ikke NA i funksjonen "summary(data)"

5. Lag et subset av datasettet hvor du fjerner NA på variabelen valg.

```r
# Lager subset av data uten NA på valg
df <- data %>%
  drop_na(valg)

sum(is.na(data$valg)) # Sjekker at det blir riktig
```

```
## [1] 323

# Antall observasjoner i df - NA i dataÆvalg
1406-323 # regner ut

## [1] 1083
```

6. Hvor mange kvinner og hvor mange menn er det i datasettet?

```
# Tabell over kjønn
table(data$kjonn)

##
##   1   2
## 777 629

# Prøv med df også
table(df$kjonn) # Færre fordi færre observasjoner i df grunnet fjerning

##
##   1   2
## 600 483

# av missing
```

7. Gjør variabelen alder til en numerisk variabel og kjønn til en factor variabel.

```
# Gjør om til numerisk og factor
df$alder <- as.numeric(df$alder)
df$kjonn <- as.factor(df$kjonn)

# Sjekker at det ble riktig
class(df$alder)

## [1] "numeric"

class(df$kjonn)

## [1] "factor"
```

8. Finn gjennomsnittsalderen til menn og deretter kvinner. Regn ut forskjellen.

```
# Gjennomsnittsalder til menn, menn har kategorien 1
menn <- mean(df$alder[df$kjonn == 1], na.rm = TRUE)
menn
```

```
## [1] 51.2381
```

```
# Gjennomsnittsalder til kvinner, kvinner har kategorien 2
kvinner <- mean(df$alder[df$kjonn == 2], na.rm = TRUE)
kvinner
```

```
## [1] 49.4916
```

```
# Regner ut forskjellen, lagrer i objektet "diff"
diff <- mean(df$alder[df$kjonn == 1], na.rm = TRUE) -
  mean(df$alder[df$kjonn == 2], na.rm = TRUE)
diff
```

```
## [1] 1.746499
```

```
# Eventuelt:
diff_2 <- menn-kvinner
diff_2
```

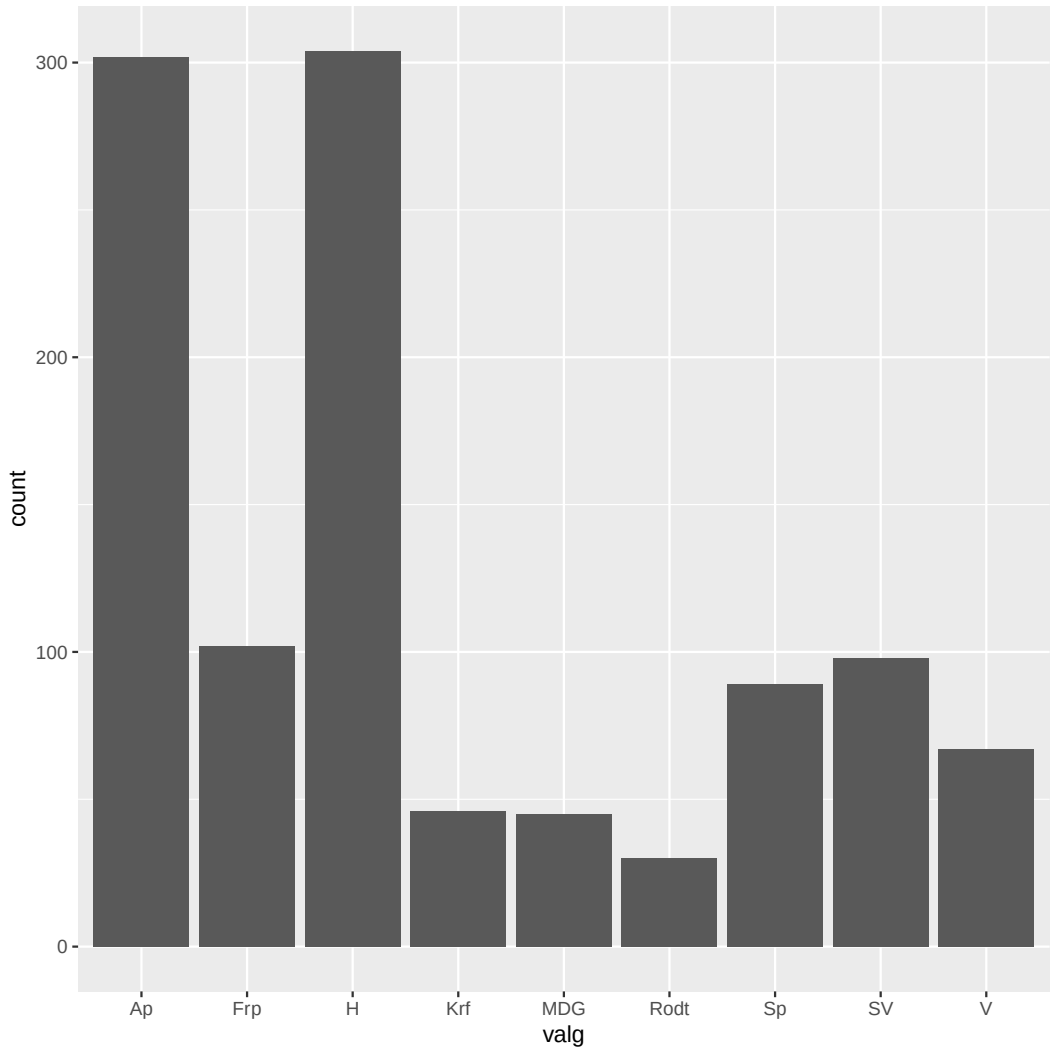
```
## [1] 1.746499
```

9. Få oversikt over variablen valg. Lag et plot som viser den univariate fordelingen. Hvor mange har stemt på hvert av partiene?

```
# Tabell over partivalg
table(df$valg)
```

```
##
##   Ap  Frp    H  Krf  MDG Rodt   Sp   SV    V
##  302  102  304   46   45   30   89   98   67
```

```
# Plot over partivalg-variabelen
ggplot(df, aes(valg)) +
  geom_bar()
```



10. Finn korrelasjonen mellom alder og rik.

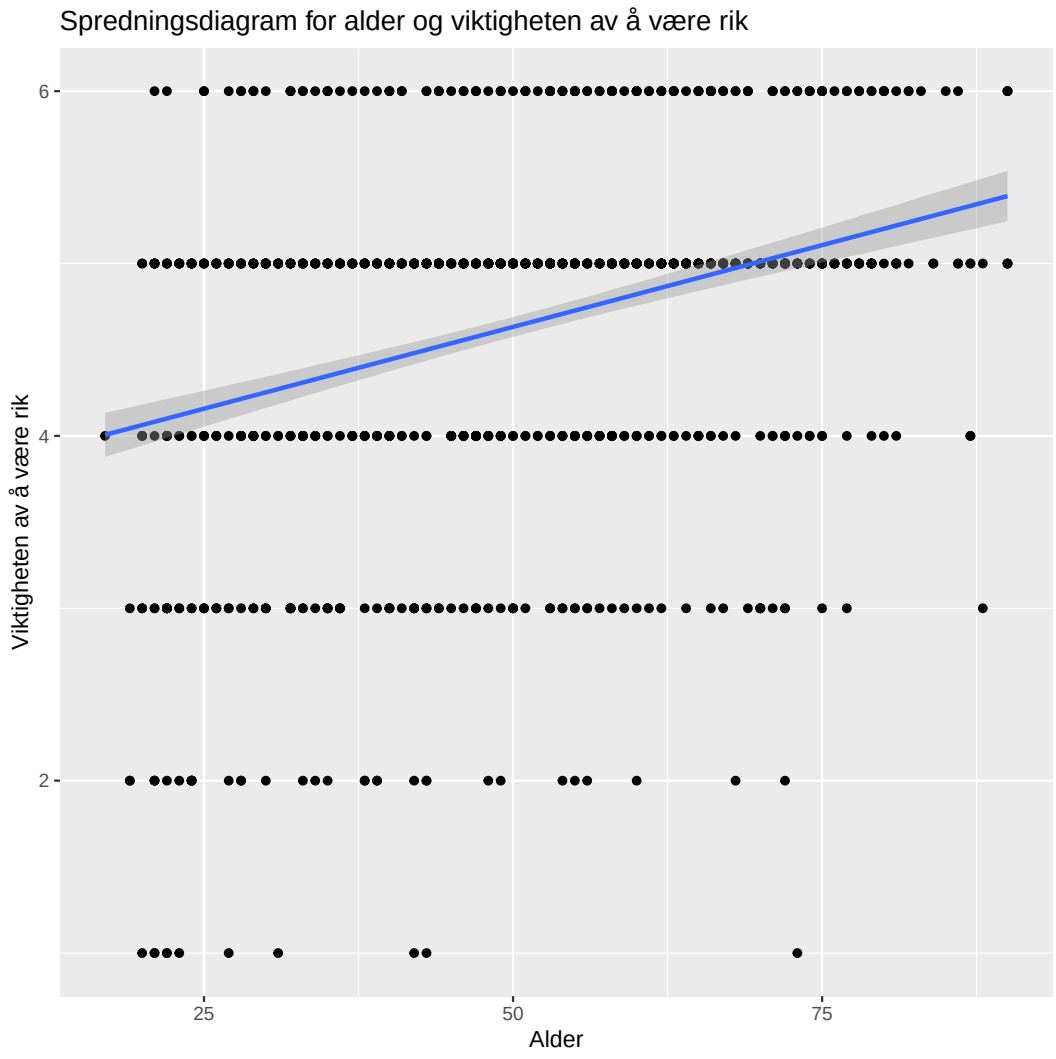
```
# Korrelasjon mellom alder og hvor viktig det er å være rik
r <- cor(x = df$rik,
         y = df$alder,
         use = "complete.obs",
         method = "pearson")

r


## [1] 0.3177088
```

11. Lag et spredningsdiagram mellom alder og rik med lineær støttelinje. Endre på navnene på x-aksen, y-aksen, og gi diagrammet en tittel. Tolk form, retning og styrke.

```
# Lager plot med støttelinje
ggplot(df, aes(alder, rik)) +
```

```
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Alder",
       y = "Viktigheten av å være rik",
       title = "Spredningsdiagram for alder og viktigheten av å være rik")
```

```
## Don't know how to automatically pick scale for object of type haven_labelled.
Defaulting to continuous.
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 23 rows containing non-finite values (stat_smooth).
## Warning: Removed 23 rows containing missing values (geom_point).
```




```
# Jo eldre man blir, jo mindre viktig blir det å være rik, ha penger og
# dyre ting. (Tenk på skalaretningen til rik)
# Obs: Én kategorisk (rik) og én kontinuerlig (alder) variabel, derfor noe rar
# gruppering av punkter.
```