Oppgaver R-seminar 4
STV1020 – VÅREN 2021

Louisa Boulaziz

February 16, 2021

1 OPPGAVER

1. Last inn datasettet "VALGDATA.Rdata"
2. Hva er navnene på variablene i datasettet?
3. Hvor mang missing er det totalt i datasettet?
4. Hvor mange missing er det på hver enkelt variabel?
5. Lag et subset av dataene hvor du fjerner NA på variabelen valg.
6. Hvor mange kvinner og hvor mange menn er det i datasettet?
7. Gjør variabelen alder og kjønn til numerisk.
8. Finn gjennomsnittsalderen.
9. Finn gjennomsnittsalderen til henholdsvis menn og kvinner. Regn ut forskjellen.
10. Få oversikt over variablen valg. Lag en oversikt som viser den univariate fordelingen – hvor mange har stemt på hvert av partiene?
11. Finn korrelasjonen mellom alder og rik.
12. Lag et spredningsdiagram mellom alder og rik med støttelinje. Endre navnene på x-aksen, y-aksen, og gi diagrammet en tittel. Tolk form, retning og styrke.

Variabler

1. rik – hvor viktig er det å være rik, ha penger og dyre ting? (1-6), 1 = helt enig, 6 = helt uenig
2. alder – i antall år

3. kjonn – dikotom, 1 = mann, 2 = kvinne
4. tillit – tillit til politikere, (1-10), 1 = ikke noe tillit, 10 = full tillit
5. valg – hvilket parti stemte du ved forrige stortingsvalg
6. redusere – Er du enig i at regjeringen skal omfordele goder. (1-5), 1 = veldig enig, 5 = helt uenig.

2 FASIT

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

```
load("VALGDATA.Rdata")

# Bytter navn

data <- nyedata

# Fjerner den andre

rm(nyedata)

# Pakke

library(tidyverse)

## -- Attaching packages ---------------------------------------
tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.6     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()
```

2. Hva er navnene på variablene i datasettet?

```
show(variable.names(data))

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

3. Hvor mang missing er det totalt i datasettet?

```
sum(is.na(data))

## [1] 377
```

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

```
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
```

```
# Eventuelt en for en
```

```
sum(is.na(data$redusere))
```

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

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

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

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

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

```
df <- data %>%
  drop_na(valg)

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

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

```
1406-323 # regner ut
```

```
## [1] 1083
```

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

```
table(data$kjonn)

##
##   1   2
## 777 629
```

7. Gjør variabelen alder og kjønn til numerisk.

```
class(df$alder)

## [1] "haven_labelled"

df$alder <- as.numeric(df$alder)

class(df$kjonn)

## [1] "haven_labelled"

df$kjonn <- as.numeric(df$kjonn)
```

8. Finn gjennomsnittsalderen.

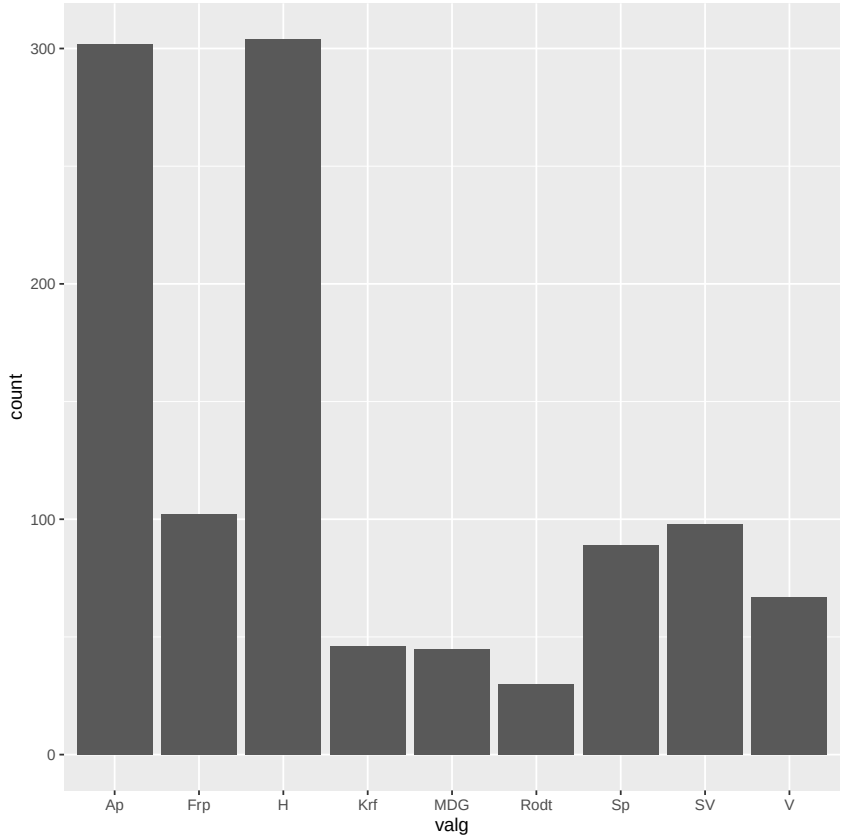
```
mean(df$alder, na.rm = TRUE)

## [1] 50.45677
```

9. Finn gjennomsnittsalderen til henholdsvis menn og kvinner. Regn ut forskjellen.

```
menn <- mean(df$alder[df$kjonn == 1], na.rm = T)
menn

## [1] 51.2381

kvinner <- mean(df$alder[df$kjonn == 2], na.rm = T)
kvinner

## [1] 49.4916

diff<- mean(df$alder[df$kjonn == 1], na.rm = T) - mean(df$alder[df$kjonn == 2], na.rm =

diff

## [1] 1.746499
```

10. Få oversikt over variablen valg. Lag en oversikt som viser den univariate fordelingen – hvor mange har stemt på hvert av partiene?

```
table(df$valg)

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

ggplot(df, aes(valg)) +
  geom_bar()
```



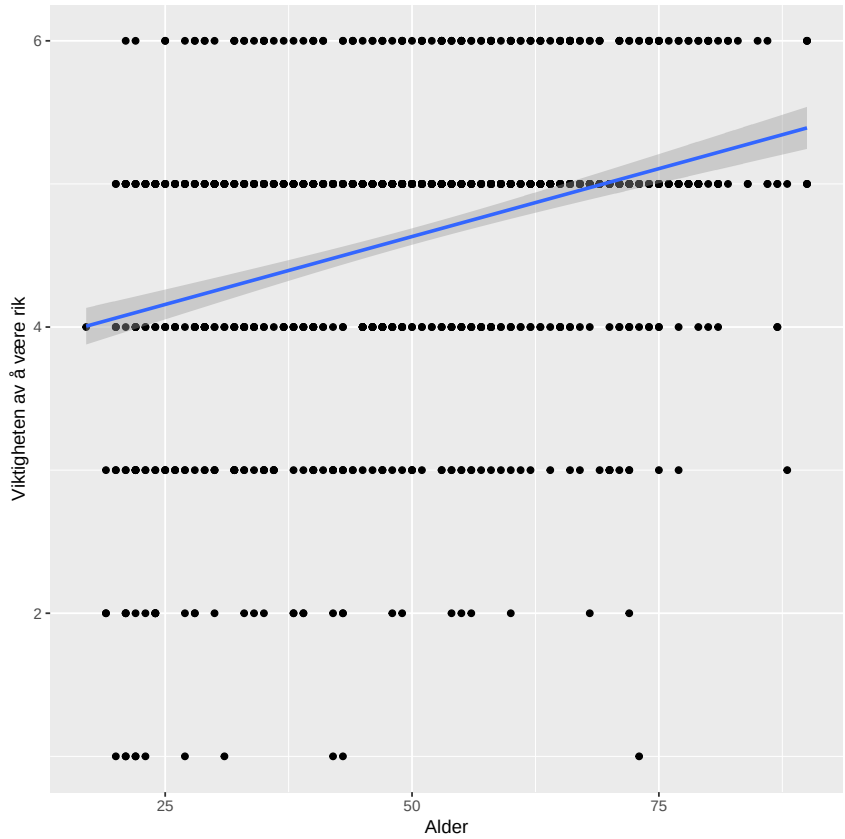
11. Finn korrelasjonen mellom alder og rik.

```
cor(x= df$rik,
    y= df$alder,
    use = "pairwise.complete.obs")
```

```
## [1] 0.3177088
```

12. Lag et spredningsdiagram mellom alder og rik med 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) +
  xlab("Alder") +
  ylab("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).
```




```
# Ingen svært tydelig gruppering langs en rett linje
# Positiv sammenheng. (0.32)
# Jo eldre man blir, jo mindre viktig blir det å a være rik,
# ha penger og dyre ting.
# Men svak sammenheng. 0.32.
```