Er det høyde som bestemmer inntekt?

```
library(modelr)
library(ggplot2)
library(tinytex)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

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

```
library(ggpubr)
library(huxtable)
```

```
##
## Attaching package: 'huxtable'
```

```
## The following object is masked from 'package:ggpubr':
##
##     font
```

```
## The following object is masked from 'package:dplyr':
##
##     add_rownames
```

```
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
options(scipen = 999)
```

Introduksjon

I denne oppgaven skal vi finne ut om høyde bestemmer inntekten våres.

Kort litteraturgjennomgang

Beskrivende statistikk

Analyse:

For analyse delen lager vi først et histogram med variabelen inntekt. Vi har gjort om inntekt, høyde og vekt til metrisk standard. Som vil si at inntekten blir gjort om til norske kroner, høyde i cm og vekten i kg.

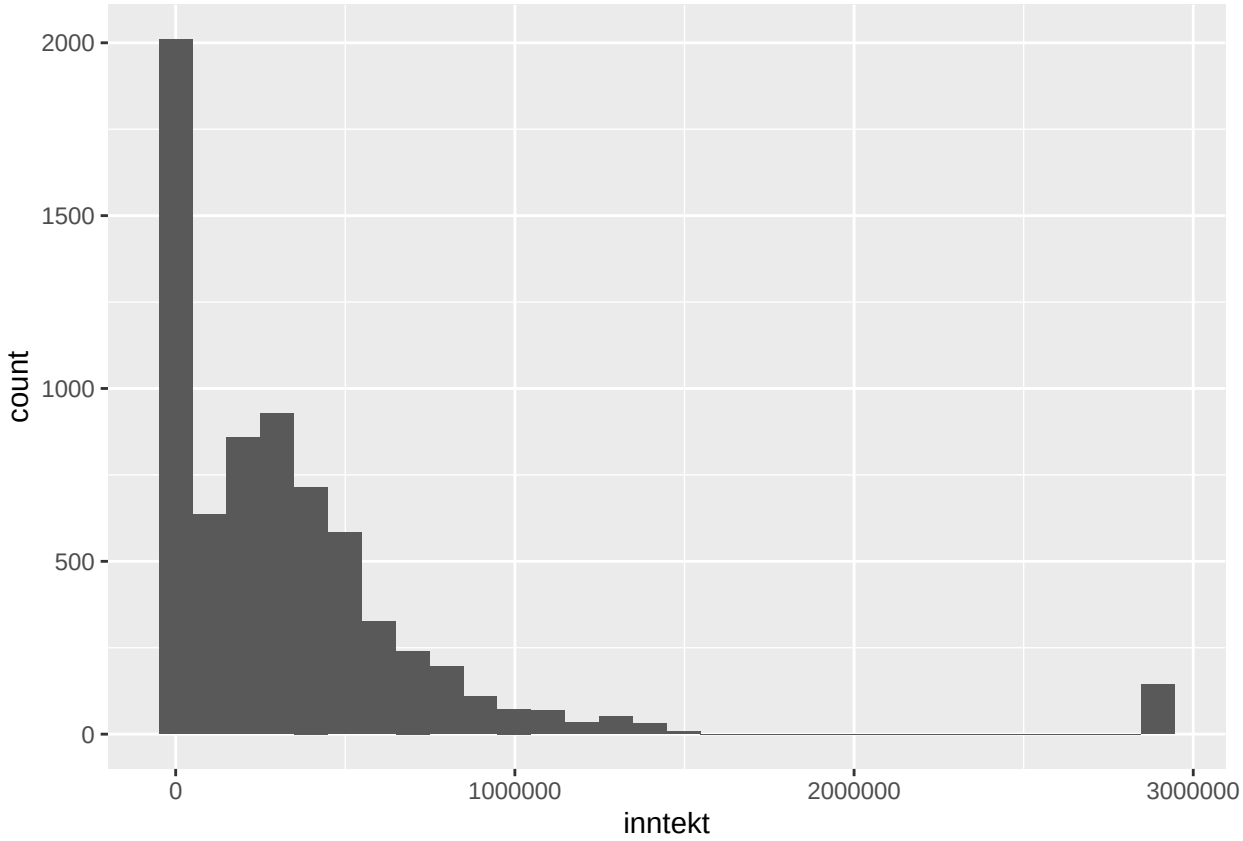
```
hoyde <- heights
```

Her har man gjort om variablene til metrisk standard. En har også lagt til tre nye variabler.

```
hoyde <- hoyde %>%
  mutate(inntekt = income * 8.42,
         hoyde_cm = height * 2.54,
         vekt_kg = weight * 0.454,
         BMI = vekt_kg/(hoyde_cm/100)^2)
```

```
ggplot(data = hoyde, aes(x = inntekt)) +
  geom_histogram()
```

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




```
  geom_histogram(bins = 30)
```

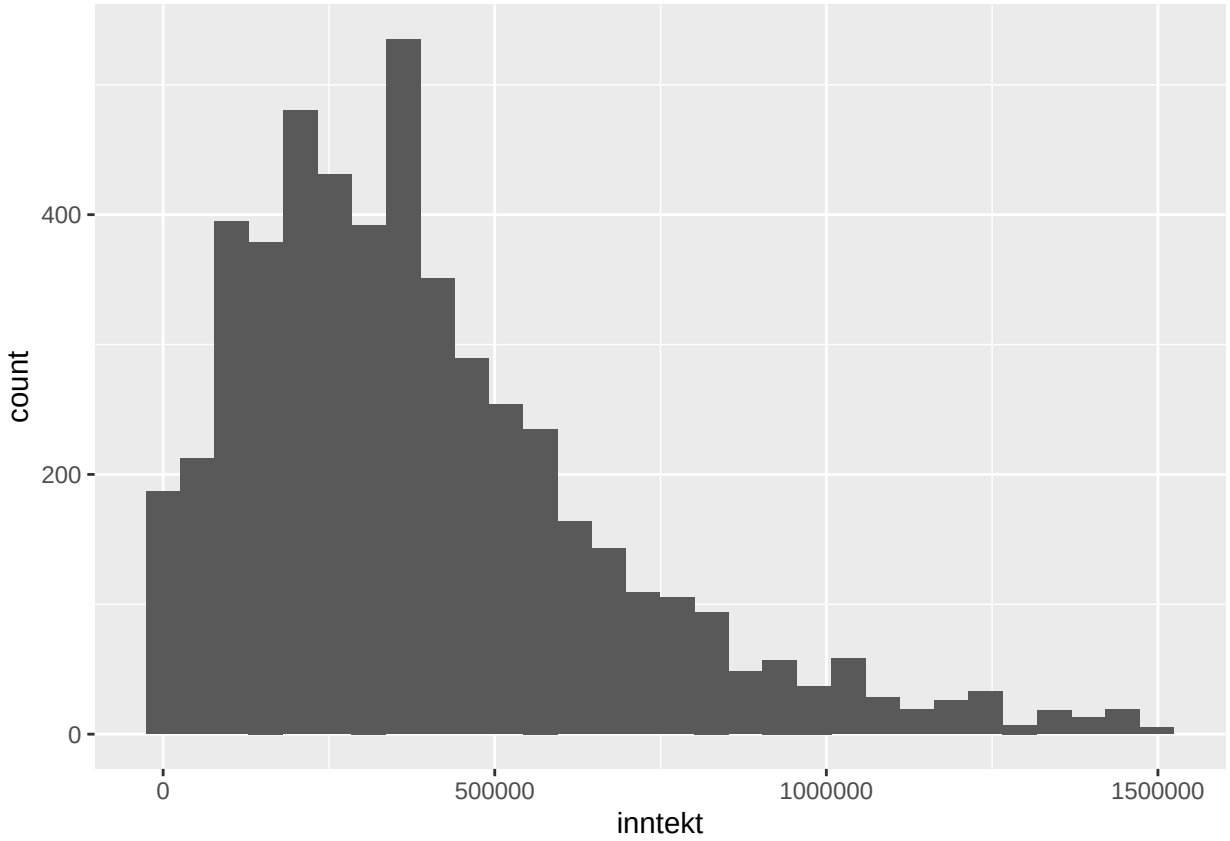
```
## geom_bar: na.rm = FALSE, orientation = NA
## stat_bin: binwidth = NULL, bins = 30, na.rm = FALSE, orientation = NA, pad = FALSE
## position_stack
```

I histogrammet ovenfor ser en at utliggerne ligger veldig langt til høyre. Grunnen for dette er at de har funnet gjennomsnittet av topp 2% inntekt.

```
hoyde_begr <- hoyde %>%
  filter(inntekt < 1500000,
         inntekt > 1)
```

```
ggplot(data = hoyde_begr, aes(x = inntekt)) +
  geom_histogram()
```

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



Som man kan se så er personer uten inntekt tatt med i datasettet. Og summen er 1740 personer uten inntekt.

```
sum(hoyde$income == 0)
```

```
## [1] 1740
```

Regresjonsanalyse

```
mod1 <- "inntekt ~ hoyde_cm"
lm1 <- lm(mod1, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm1)
```

```
##
## Call:
## lm(formula = mod1, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -782810 -267359  -94513  123099 2699234
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
```

```
## (Intercept) -1361001.0    94430.0  -14.41 <0.0000000000000002 ***
## hoyde_cm       10047.9      552.8   18.18 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 467300 on 6643 degrees of freedom
## Multiple R-squared:  0.04737,    Adjusted R-squared:  0.04723
## F-statistic: 330.3 on 1 and 6643 DF,  p-value: < 0.00000000000000022
```

```
-1361001.0 + (10047.9 * 173)
```

```
## [1] 377285.7
```

```
-1361001.0 + (10047.9 * 161)
```

```
## [1] 256710.9
```

Man øker inntekten sin med 10047.9 kr per cm en øker i høyde.

```
mod2 <- "inntekt ~ hoyde_cm + vekt_kg"
lm2 <- lm(mod2, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = mod2, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -843668 -263322  -92573  125798 2715000
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -1466873.6    96890.5 -15.139 < 0.0000000000000002 ***
## hoyde_cm       11430.3      624.3  18.308 < 0.0000000000000002 ***
## vekt_kg        -1518.4      320.5  -4.737           0.00000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 466600 on 6642 degrees of freedom
## Multiple R-squared:  0.05058,    Adjusted R-squared:  0.05029
## F-statistic: 176.9 on 2 and 6642 DF,  p-value: < 0.00000000000000022
```

```
-1466873.6 + (11430.3 * 173) + (-1518.4 * 70)
```

```
## [1] 404280.3
```

```
-1466873.6 + (11430.3 * 161) + (-1518.4 * 65)
```

```
## [1] 274708.7
```

Når høyden øker så går inntekten opp, mens når vekten økes går lønnen ned. Men en kombinasjon av disse gir økt inntekt.

```
mod3 <- "inntekt ~ hoyde_cm + vekt_kg + BMI"
lm3 <- lm(mod3, data = hoyde, subset = complete.cases(hoyde))
```

```
summary(lm3)
```

```
##
## Call:
## lm(formula = mod3, data = hoyde, subset = complete.cases(hoyde))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -886295 -261634  -93597  124905 2709981
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)
## (Intercept) -2015890     447005  -4.510 0.0000066012 ***
## hoyde_cm       14669       2649   5.537 0.0000000319 ***
## vekt_kg        -4723       2567  -1.840       0.0658 .
## BMI             9224       7332   1.258       0.2084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 466600 on 6641 degrees of freedom
## Multiple R-squared:  0.05081,    Adjusted R-squared:  0.05038
## F-statistic: 118.5 on 3 and 6641 DF,  p-value: < 0.00000000000000022
```

```
hoyde <- hoyde %>%
  mutate(
    married = factor(
      case_when(
        marital == 'married' ~ TRUE, TRUE ~ FALSE
      )
    )
  )
```

```
huxreg(
   list("mod1" = lm1, "mod2" = lm2, "mod3" = lm3),
   error_format = "[{statistic}]",
   note = "Regresjonstabell 3: {stars}. T statistics in brackets."
   )
```

```
mod4 <- "inntekt ~ sex*(hoyde_cm + vekt_kg + I(vekt_kg^2)) + BMI + I(BMI^2)"
lm4 <- lm(mod4, data = hoyde)
summary(lm4)
```

```
##
## Call:
## lm(formula = mod4, data = hoyde)
```

	mod1	mod2	mod3
(Intercept)	-1361000.990 ***	-1466873.555 ***	-2015889.845 ***
	[-14.413]	[-15.139]	[-4.510]
hoyde_cm	10047.860 ***	11430.259 ***	14669.413 ***
	[18.175]	[18.308]	[5.537]
vekt_kg		-1518.381 ***	-4722.577
		[-4.737]	[-1.840]
BMI			9224.408
			[1.258]
N	6645	6645	6645
R2	0.047	0.051	0.051
logLik	-96177.211	-96166.004	-96165.212
AIC	192360.423	192340.008	192340.424

Regresjonstabell 3: *** p < 0.001; ** p < 0.01; * p < 0.05. T statistics in brackets.

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -864444 -245100  -91019  126362 2681172
##
## Coefficients:
##                          Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)           -2821666.91  1904365.52  -1.482   0.13847
## sexfemale              1181398.44   293082.63   4.031 0.0000562 ***
## hoyde_cm                 17091.78    10627.73   1.608   0.10783
## vekt_kg                  -4749.34    17977.28  -0.264   0.79164
## I(vekt_kg^2)               -17.95       42.26  -0.425   0.67109
## BMI                      34177.41    57584.98   0.594   0.55286
## I(BMI^2)                  -190.52      435.11  -0.438   0.66150
## sexfemale:hoyde_cm       -4729.20     1812.91  -2.609   0.00911 **
## sexfemale:vekt_kg        -9825.85     5200.88  -1.889   0.05890 .
## sexfemale:I(vekt_kg^2)      45.96       27.06   1.699   0.08941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458300 on 6901 degrees of freedom
##   (95 observations deleted due to missingness)
## Multiple R-squared:  0.06165,    Adjusted R-squared:  0.06043
## F-statistic: 50.38 on 9 and 6901 DF,  p-value: < 0.00000000000000022
```

```
linearHypothesis(lm4, c("sexfemale = 0", "sexfemale:hoyde_cm = 0"))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6.9e+03	1.45e+15				
6.9e+03	1.45e+15	2	3.42e+12	8.13	0.000297

Referanser