Innlevering 2

Innlevering 2 i Data Science 2021 - Maren Sognefest og Daniel Karstad

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     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

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

##
## 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
```

Er det høyde som bestemmer inntekt?

Denne artikkelen skal ta for seg om det er en, og eventuelt hvilken, sammenheng det er mellom høyde og inntekt. Ved hjelp av flere analyser, skal vi bruke datasettet *heights* (fra (**National?**) Longitudinal Study) som vi finner i pakken *modelr* skal vi prøve å finne ut om det er en sammenheng. Kan det stemme at høye personer tjener mest?

I analyse-delen av artikkelen vil vi bruke ulike **plots** for å analysere spørsmålet, og komme frem til en konklusjon.

Litteraturgjennomgang

At høyde har påvirkning på karrieresuksess er ingen ny tanke, men det er gjort få studier på dette området. En av de viktigste indikatorene for suksess i arbeidslivet er lønn, og i 2004 skrev (**judge?**) og Cabel en artikkel som omhandlet hvorvidt høyde bestemmer inntekt. De nevner at utsagnet «høyde påvirker inntekten» kanskje er mer sant enn man først skulle tro. Tidligere studier har vist at høye mennesker fremstår som mer overbevisende ((**Young?**) & French, 1996), mer attraktive som partnere (Freedman, 1979; Harrison & Saeed, 1977; Lerner & Moore, 1974) og fremstår mer naturlig som en lederskikkelse (Higham & Carment, 1992; Stogdill, 1948). Bakgrunnen for dette kan ligge i biologien vår og i evolusjonsteorien. I dyreverden brukes nemlig høyde som et mål på styrke i en risikovurdering i en fight-or-flight-situasjon.

Det nevnes at høyde kan gi mennesker bedre selvillit og bedre sosial selvtillit, noe som i seg selv kan være en faktor for å yte bedre - og følgelig ha høyere inntekt. Videre i artikkelen inkluderer de fire forskjellige studier. I tillegg til de to åpenbare variablene i studien (høyde og inntket), inkluderte de tre kontrollvariabler; kjønn, alder og vekt.

Fordi menn generelt er høyere enn kvinner, og at det er kjent at menn tjener bedre enn kvinner kan dette være med på å påvirke resultatet av studiene. Alder er relevant, fordi et gjennomsnittsenneske vil "krympe" 5 cm i høyde i løpet av livet sitt. Vekt henger naturligvis sammen med høyde, men de kan dra i ulike retninger. Det er i følge Judge og Cabel flere grunner til å tro at høyde har positiv innvirkning, kan vekt ha det motsatte.

For å gjøre resultatene så generaliserbare som mulig, ble sammenhengen mellom høyde og inntekt sett på gjennom fire unike prøver. Dataene er begrenset til enkeltpersoner som jobber minst 20 timer i snitt ukentlig, med unntak av studie 3 hvor dataene i analysen er begrenset til å kun omhandle personer som var hovedinntektskilden i husholdningen.

Dataene til de fire studiene, ble hentet inn fra fire ulike kilder - en for hvert studie. Judge og Cabel konkluderte med at det er sammenheng mellom en persons høyde og inntekt, noe som har vært diskutert i ettertid. Det er dette vi skal se på nå, ved bruk av et helt annet datasett enn det som ble brukt i 2004 (**Judge2004?**).

Judge and Cable (2004)

Datasett

income	height	weight	age	marital	sex	education	afqt
Min. : 0.0	Min. :52.0	Min. : 76.0	Min. :47.00	single :1124	male :3402	Min. : 1.00	Min. : 0.00
1st Qu.: 165.5	1st Qu.:64.0	1st Qu.:157.0	1st Qu.:49.00	married :3806	female:3604	1st Qu.:12.00	1st Qu.: 15.12
Median : 29589.5	Median :67.0	Median :184.0	Median :51.00	separated: 366	NA	Median :12.00	Median : 36.76
Mean : 41203.9	Mean :67.1	Mean :188.3	Mean :51.33	divorced :1549	NA	Mean :13.22	Mean : 41.21
3rd Qu.: 55000.0	3rd Qu.:70.0	3rd Qu.:212.0	3rd Qu.:53.00	widowed : 161	NA	3rd Qu.:15.00	3rd Qu.: 65.24
Max. :343830.0	Max. :84.0	Max. :524.0	Max. :56.00	NA	NA	Max. :20.00	Max. :100.00
NA	NA	NA's :95	NA	NA	NA	NA's :10	NA's :262

Over er sammendraget av statistikken i det originale datasettet "heights." Man har kolonner

med inntekt i dollar, høyde i tommer, vekt i pound, alder, sivilstatus, kjønn, utdannelse og score på Armed Forces Qualitication Test.

Under har vi kopiert datasettet "heights" og kalt det "hoyde." Her har vi regnet dataen om til europeiske standarderer; vi bruker høyde i cm og vekt i kg, i tillegg til at inntekten er i norske kroner, at vi har laget en egen kolonne for BMI og at vi har laget en forenklet utgave av marital (married - not married). Det er vedlagt et sammendrag av dette under.

```
hoyde <-   heights %>%
mutate(heights, hoyde_cm = height*2.54, #høyde i cm = hoyde i tommer * 2,54 fordi 1 tom
       vekt_kg = weight/2.2, #vekt i kg = vekt i pound / 2.2 fordi 2,2 kg = 1 pound
       inntekt_nok = income*8.5,  #inntekt i nok = inntekt i dollar * 8,5 fordi 1 dollo
        married = factor(
          case_when(
            marital == 'married' ~ TRUE,
            TRUE ~ FALSE)
          )
       )
hoyde$bmi <- hoyde$vekt_kg/(hoyde$hoyde_cm/100)/(hoyde$hoyde_cm/100) #bmi = vekt i kg /
hoyde$weight <- NULL #fjerner vekt i pounds fra datasett
hoyde$height <- NULL #fjerner høyde i tommer fra datasett
hoyde$income <- NULL #fjerner inntekt i dollar fra datasett


knitr::kable(summary(hoyde[1:5]))
```

age	marital	sex	education	afqt
Min. :47.00	single :1124	male :3402	Min. : 1.00	Min. : 0.00
1st Qu.:49.00	married :3806	female:3604	1st Qu.:12.00	1st Qu.: 15.12
Median :51.00	separated: 366	NA	Median :12.00	Median : 36.76
Mean :51.33	divorced :1549	NA	Mean :13.22	Mean : 41.21

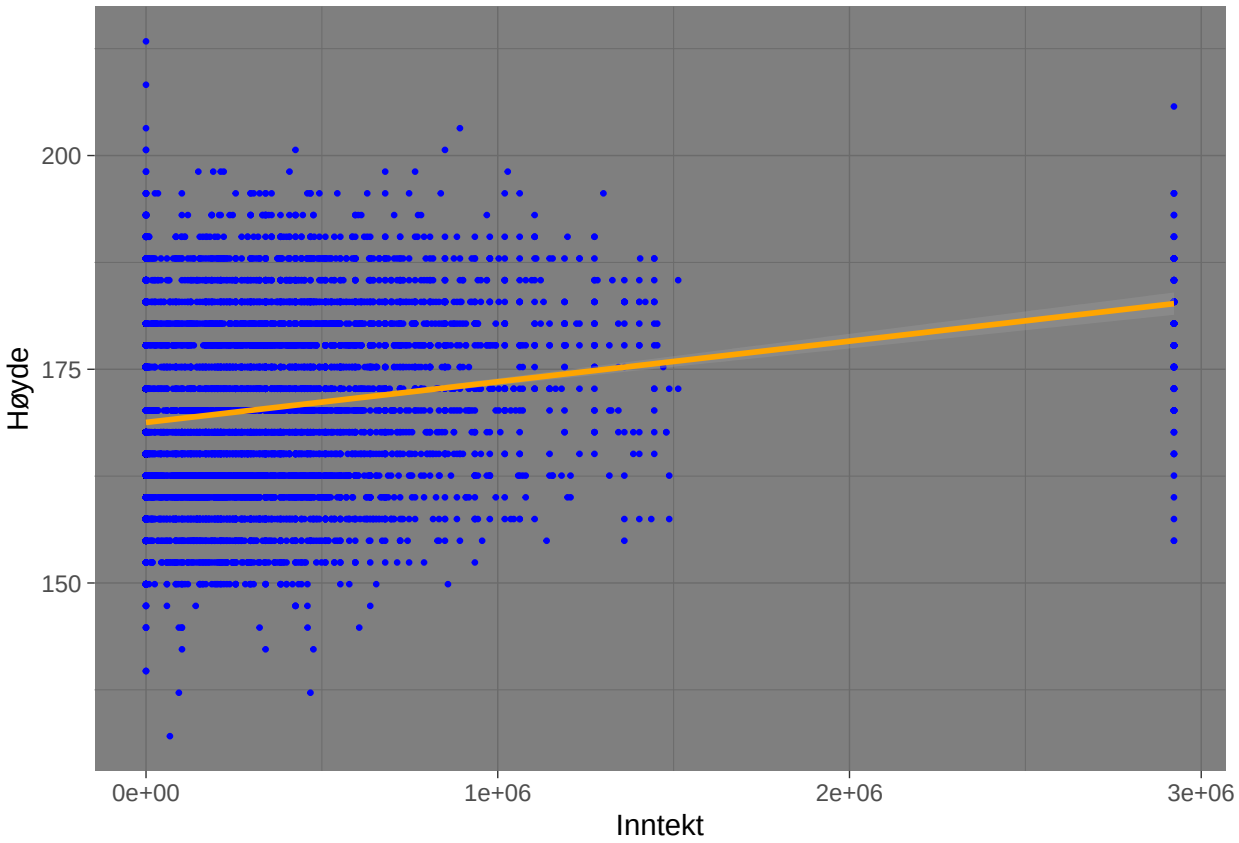
age	marital	sex	education	afqt
3rd Qu.:53.00	widowed : 161	NA	3rd Qu.:15.00	3rd Qu.: 65.24
Max. :56.00	NA	NA	Max. :20.00	Max. :100.00
NA	NA	NA	NA's :10	NA's :262

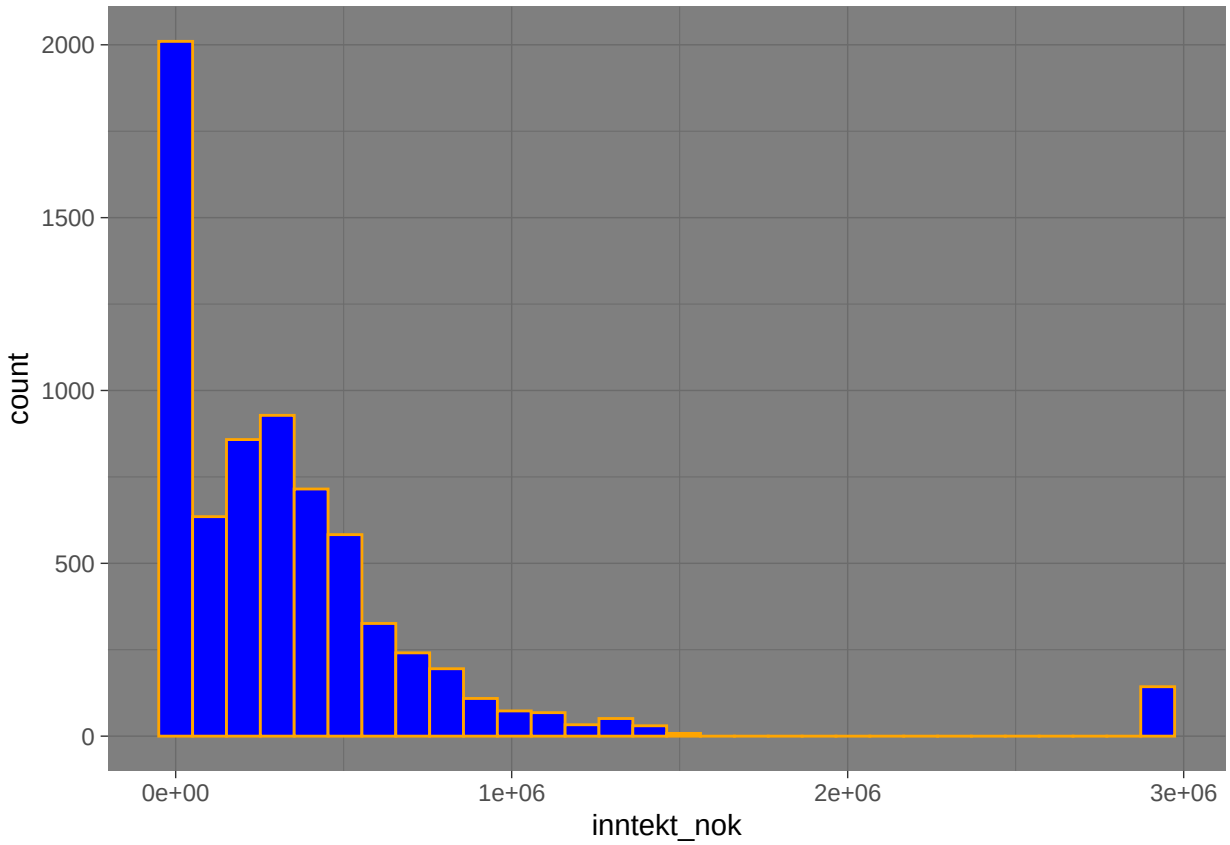
```
knitr::kable(summary(hoyde[6:10]), "pipe")
```

hoyde_cm	vekt_kg	inntekt_nok	married	bmi
Min. :132.1	Min. : 34.55	Min. : 0	FALSE:3200	Min. :12.90
1st Qu.:162.6	1st Qu.: 71.36	1st Qu.: 1407	TRUE :3806	1st Qu.:25.14
Median :170.2	Median : 83.64	Median : 251511	NA	Median :28.38
Mean :170.4	Mean : 85.59	Mean : 350234	NA	Mean :29.37
3rd Qu.:177.8	3rd Qu.: 96.36	3rd Qu.: 467500	NA	3rd Qu.:32.35
Max. :213.4	Max. :238.18	Max. :2922555	NA	Max. :75.15
NA	NA's :95	NA	NA	NA's :95

EDA

Under ser vi hvordan datasettet ser ut ved hjelp av punkter og et histogram. Utliggerne er der fordi det er beregnet gjennomsnittsinntekt av de to prosentene med høyest inntekt. Som nevnt, er dette et gjennomsnittet som er brukt, og har erstattet alle de øverste verdiene.


Høyde
200
175
150
0e+00
1e+06
2e+06
3e+06
Inntekt



Som man kan se utfra modellene over er det med flere uten inntekt i datasettet. Det kan man også se slik:

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

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

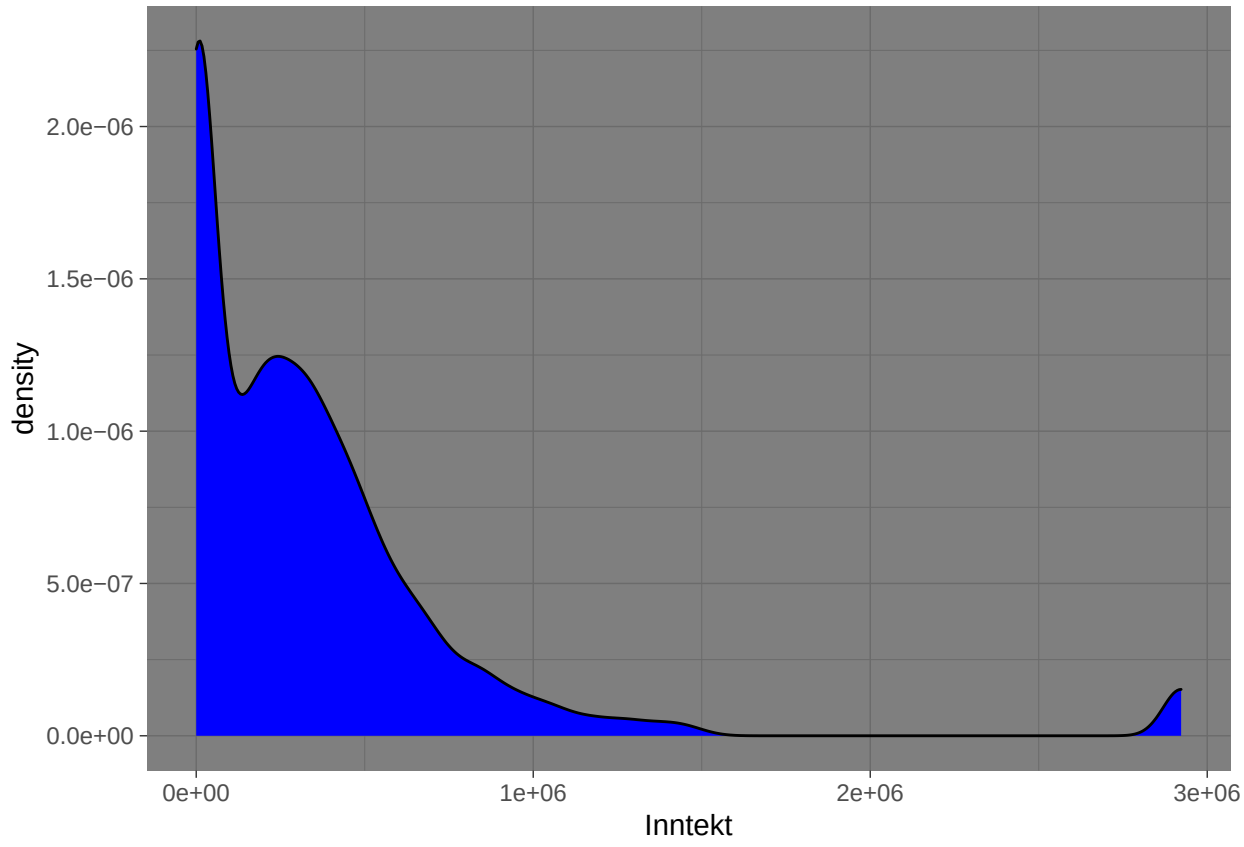
Det er altå 1740 personer uten inntekt med i datasettet.

```
max(hoyde$inntekt_nok)
```

```
## [1] 2922555
```

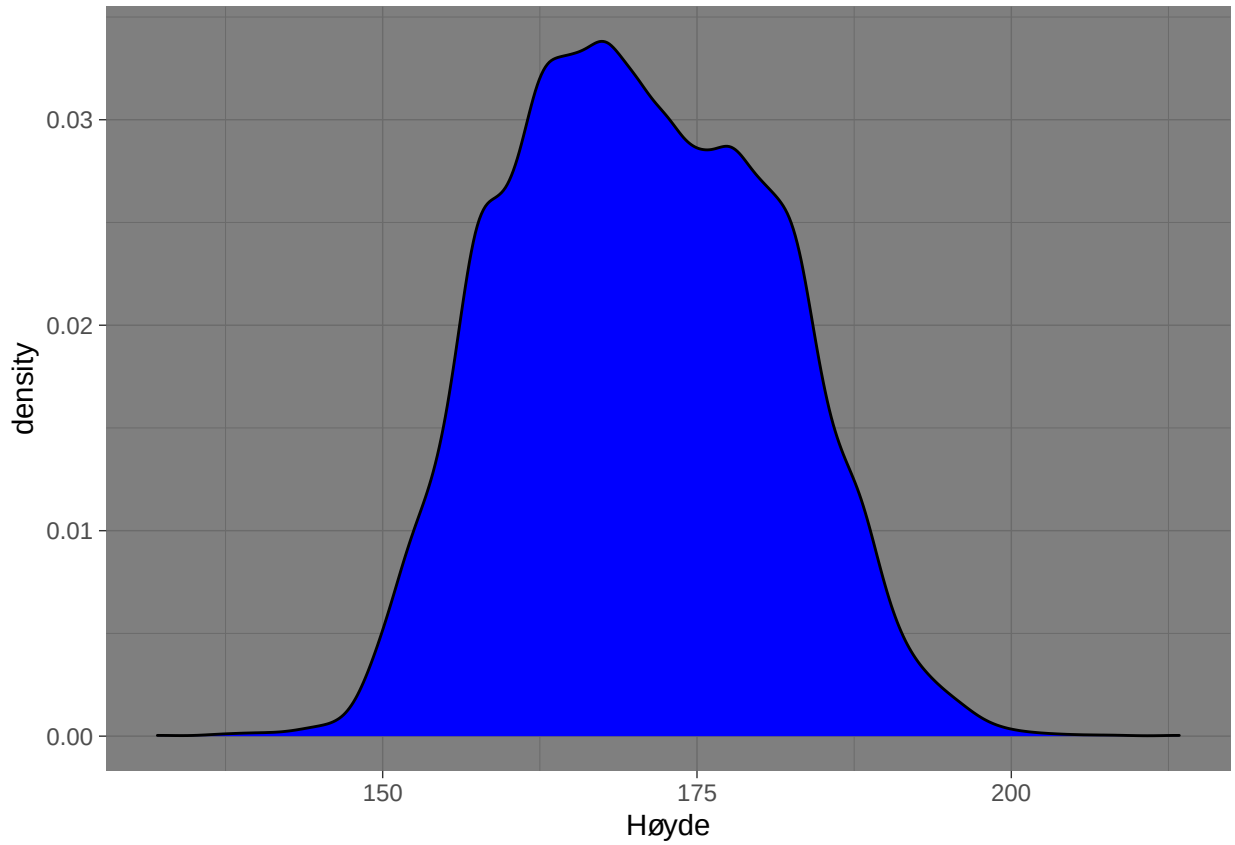
Gjennomsnittet som er målt av de 2% med høyest lønn, er kroner 292 2555.

```
Inntekt <- hoyde$inntekt_nok
ggplot(hoyde = Inntekt) + theme_dark() +
  geom_density(aes(x = Inntekt), fill = "blue")
```



Grafen over illustrerer inntekt, og her ser man ved første øyekast at det ikke er normalfordelt. Dette fordi man har mange observasjoner tilnærmet 0, og hovedvekten er <500,000.00 NOK. I andre enden av skalaen, de med 2% høyest inntekt er illustrert ved et gjennomsnitt av lønnen. Derfor foreligger det ingen observasjoner mellom 1,600,000.00 NOK og 3,000,000.00 NOK.

```
Høyde <- hoyde$hoyde_cm
Inntekt <- hoyde$inntekt_nok
ggplot(hoyde = Høyde) + theme_dark() +
  geom_density(aes(x = Høyde), fill = "blue")
```



Modellen over: Vi merker oss at histogrammet er tilnærmet normalfordelt, der de fleste observasjonene ligger mellom 160-180 cm.

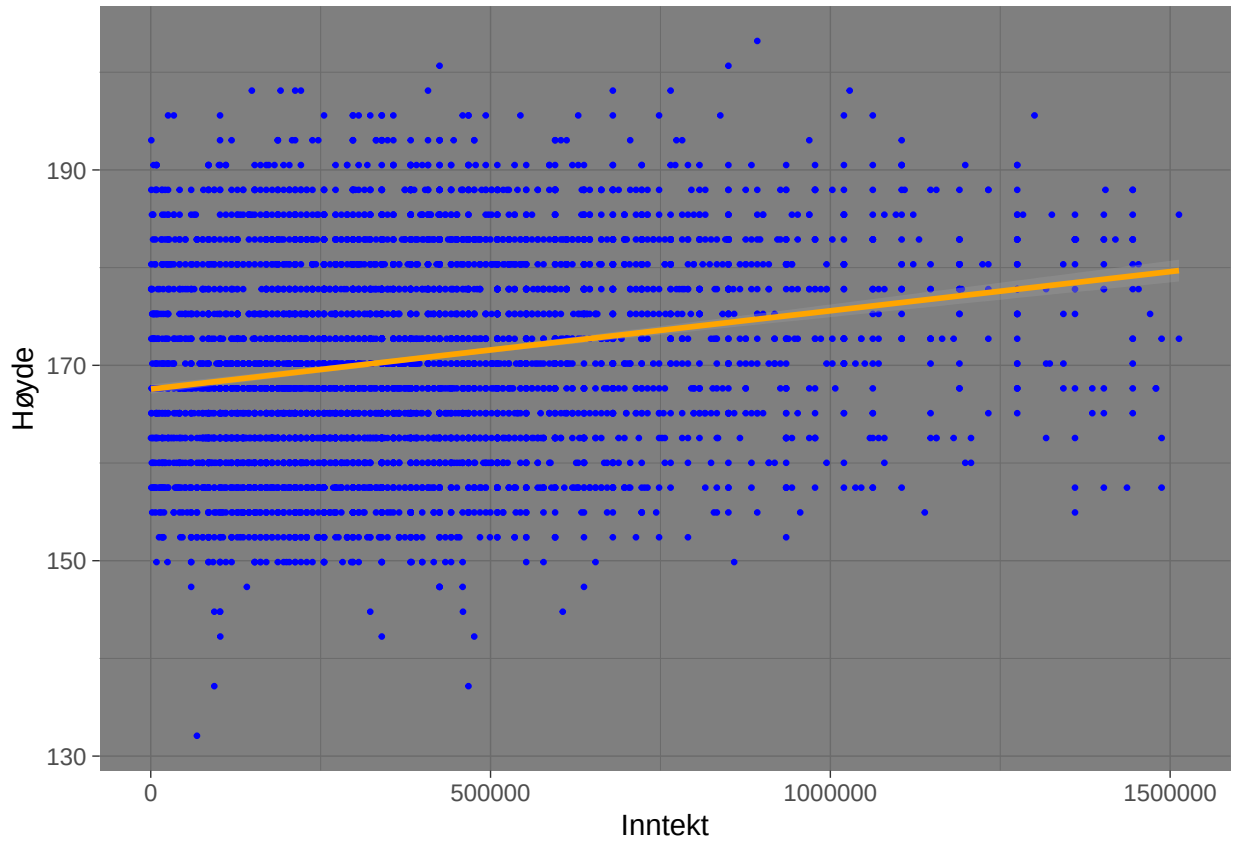
Redusert datasett

Vi fjerner de mest ekstreme; utliggerne til høyre (2% høyeste lønninger) og utliggerne til venstre (inntekt på kr 0). Dette gjør vi ved å lage et nytt datasett, kalt *hoyde_filter*. og dette lager vi ved hjelp av funksjonen *filter*. Vi legger ved de samme illustrasjonene av dette datasettet, som vi har av det originale:

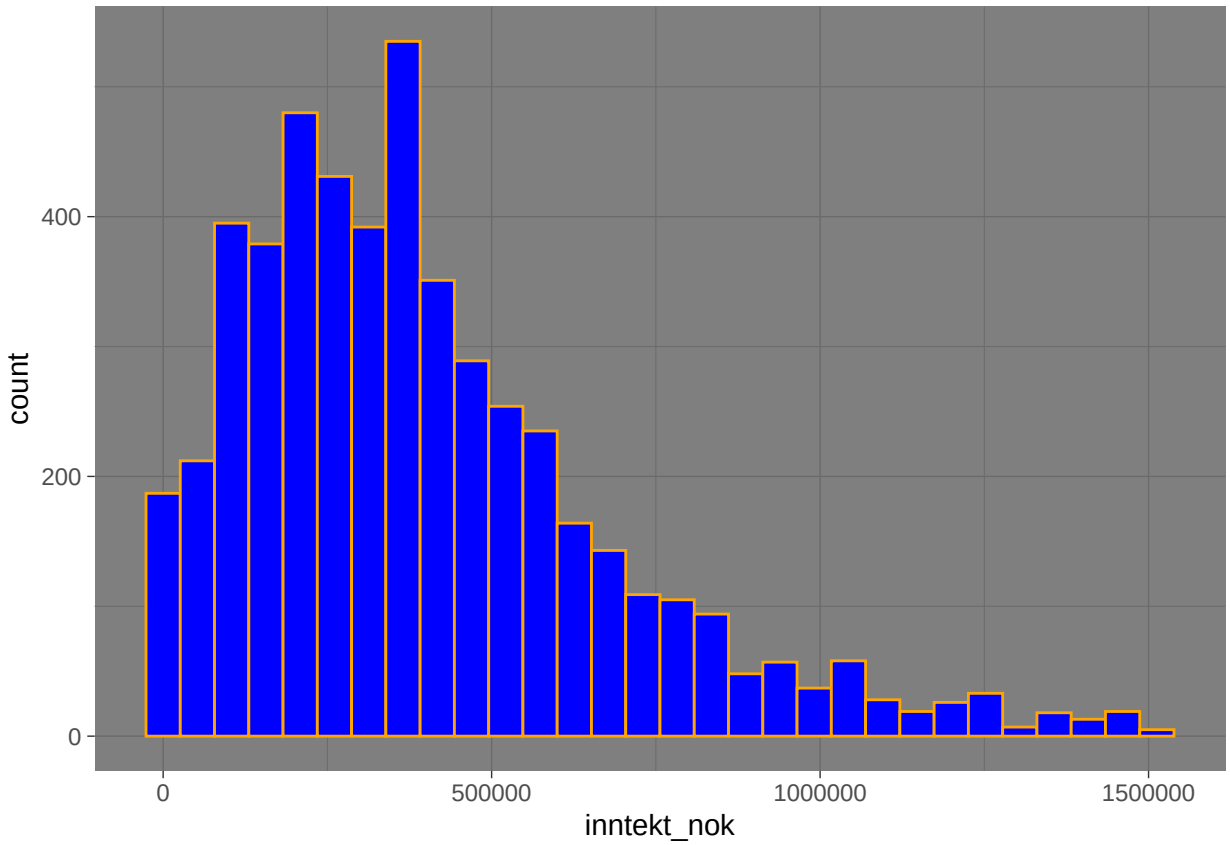
```
library(dplyr)
hoyde_filter = filter(hoyde, inntekt_nok != 0, inntekt_nok != "2922555")

Høyde <- hoyde_filter$hoyde_cm
Inntekt <- hoyde_filter$inntekt_nok
```

```
ggplot(hoyde_filter, aes(Inntekt, Høyde)) +  theme_dark() +
  geom_point(color= "blue", size = 0.5) +
geom_smooth(formula = y ~ x, method = "lm", color= "orange")
```




```
ggplot(data = hoyde_filter,
       aes(x = inntekt_nok)) +
  geom_histogram(bins = 30, col = "orange", fill = "blue") + theme_dark()
```



Regresjonsanalyse

Øverst i grafen kan man se ulempen ved å inkludere de med 2% høyest lønn, da det blir ingen punkter/observasjoner mellom 1.600.000 og 3.000,000 NOK. Det som er interessant er jo at de på toppen er representert gjennom hele høydespekteret. Den oransje regresjonslinjen stiger på x- og y-aksen jo høyere man kommer på høydespekteret, men den stiger ikke nevneverdig. Altså vil det ikke gi sterke indisier på at høyde påvirker inntekt i særlig grad, men man kan se at det er en økning.

```
(lm(inntekt_nok ~ hoyde_cm, data = hoyde)) %>%
  summary()
```

```
##
## Call:
```

```
## lm(formula = inntekt_nok ~ hoyde_cm, data = hoyde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -778460 -267842  -92589  126498 2727038
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1350548.5    91236.9  -14.80   <2e-16 ***
## hoyde_cm        9978.5      534.3   18.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 463700 on 7004 degrees of freedom
## Multiple R-squared:  0.04744,    Adjusted R-squared:  0.0473
## F-statistic: 348.8 on 1 and 7004 DF,  p-value: < 2.2e-16
```

Denne regresjonsanalysen viser at 1 cm ekstra høyde, vi gi 9.978,5 NOK ekstra i årsinntekt. Ved å se på R-squared, ser man også at høyde ikke har særlig innvirkning på inntekt, med en forklaringsgrad på bare 4.74%. Vi gjør tilsvarende analyse av det reduserte datasettet:

```
(lm(inntekt_nok ~ hoyde_cm, data = hoyde_filter)) %>%
  summary()
```

```
##
## Call:
## lm(formula = inntekt_nok ~ hoyde_cm, data = hoyde_filter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532259 -190685  -57109  135445 1170911
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -642281.3    64244.0  -9.998   <2e-16 ***
## hoyde_cm       6088.8      375.6  16.212   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276000 on 5121 degrees of freedom
## Multiple R-squared:  0.04882,    Adjusted R-squared:  0.04863
## F-statistic: 262.8 on 1 and 5121 DF,  p-value: < 2.2e-16
```

Tar man bort alle som tjener 0 kroner, og de 2% som har høyest inntekt, vil vi se at en økning på 1 cm vil gi en økt inntekt på 6088,8 kroner.

Vi skal også gjøre denne analysen av et datasett kun uten de 2% øverste inntektene, altså inkluderer vi de med inntekt på 0 kr:

```
hoyde_semi= filter(hoyde, inntekt_nok != "2922555")


(lm(inntekt_nok ~ hoyde_cm, data = hoyde_semi)) %>%
  summary()
```

```
##
## Call:
## lm(formula = inntekt_nok ~ hoyde_cm, data = hoyde_semi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -547811 -236923  -54031  158327 1265382
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -695742.7    58424.7  -11.91   <2e-16 ***
## hoyde_cm       5828.4      342.5   17.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 293300 on 6861 degrees of freedom
## Multiple R-squared:  0.0405, Adjusted R-squared:  0.04036
## F-statistic: 289.6 on 1 and 6861 DF,  p-value: < 2.2e-16
```

Dersom vi kun fjerner de 2 prosentene med høyest lønn, vil 1 cm økning i høyde gi 5828,4 kroner økt årlig inntekt.

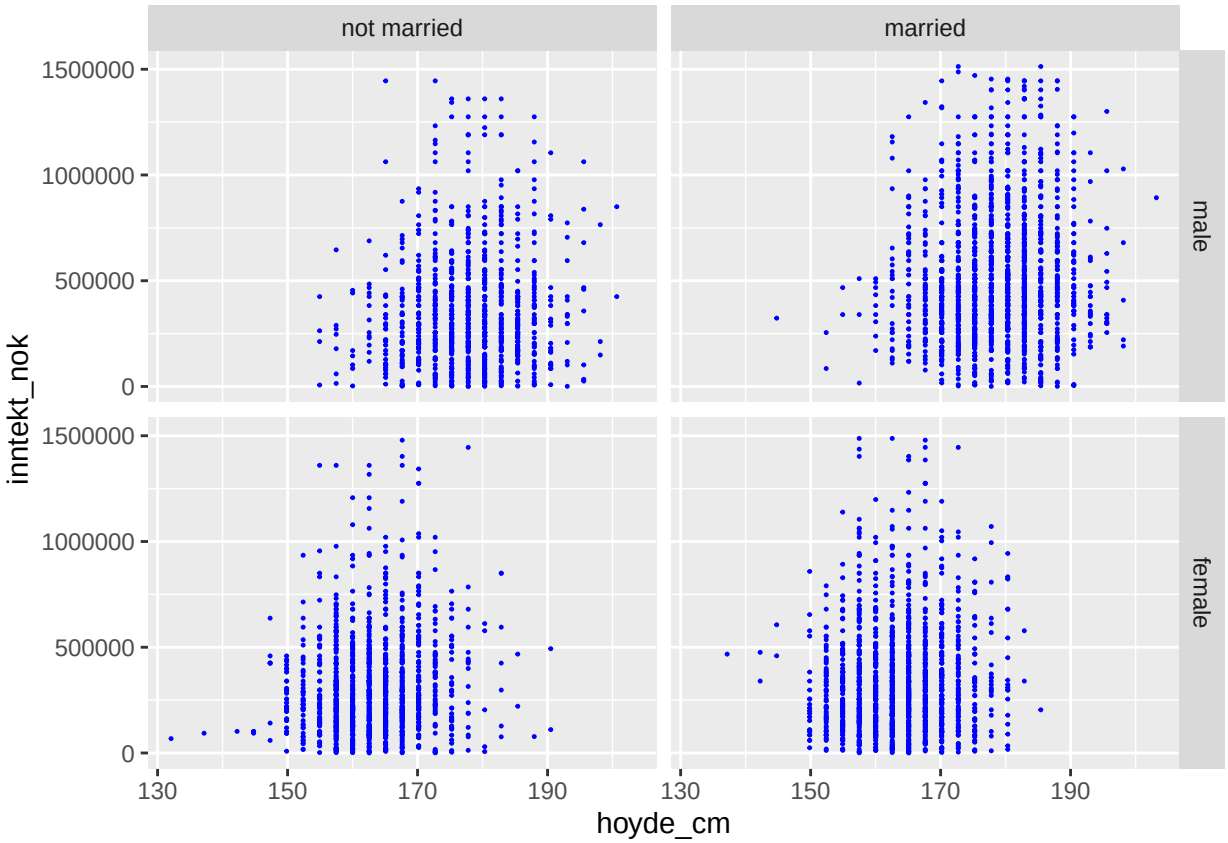
```
model <- lm(inntekt_nok ~ hoyde_cm + education + sex, data = hoyde_filter)
summary(model)
```

```
##
## Call:
## lm(formula = inntekt_nok ~ hoyde_cm + education + sex, data = hoyde_filter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -771654 -163506  -35043  121591 1089553
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -385258      88237  -4.366 1.29e-05 ***
## hoyde_cm        1341        497   2.698    0.007 **
## education      45968       1400  32.827  < 2e-16 ***
## sexfemale    -130227      10262 -12.690  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249800 on 5117 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2209, Adjusted R-squared:  0.2204
## F-statistic: 483.5 on 3 and 5117 DF,  p-value: < 2.2e-16
```

Inntekt = -385.258 + 1.341 * høyde + 45.968 * utdannelse (- 130.227 om du er kvinne). R2 er 0,2209, altså kan 22,09% av variasjonen i inntekt forklares av variablene høyde, utdannelse og kjønn.

```
ggplot(data = hoyde_filter, mapping = aes(x = hoyde_cm, y = inntekt_nok)) +
  geom_point(
    data = hoyde_filter,
    mapping = aes(x = hoyde_cm, y = inntekt_nok),
    colour = "blue",
    size = 0.2
    ) +
  facet_grid(sex ~ factor(married, labels = c("not married", "married")))
```



Her ser man umiddelbart at de som er gift har høyere inntekt, for begge kjønn. Dette er nok trolig fordi personer i et etablert ekteskap har høyere sosioøkonomisk status. Noe som er en variabel for inntekt av flere grunner, t.d utdanning, yrke etc.

Huxreg

Lager to nye datasett, kalt lm_hoyde og lm_hoyde_filter. Disse er laget utfra den naturlige logaritmen til *hoyde* og *hoyde_filter*, og vil derfor inneholde andre variabler.

```
lm_hoyde <- (lm(
  inntekt_nok ~ hoyde_cm + age + vekt_kg + marital + bmi,
                data = hoyde))
lm_hoyde_filter <- (lm(
  inntekt_nok ~ hoyde_cm + age + vekt_kg + marital + bmi,
```

```
                data = hoyde_filter))
```

```
huxreg(
  list("Hoyde"=lm_hoyde, "Hoyde_filter"=lm_hoyde_filter),
  error_format = "[{statistic}]",
  borders = 0.5,
  outer_borders = 0.8,
  error_pos = "same")
```

	Hoyde	Hoyde_filter
(Intercept)	-2053576.057 *** [-4.604]	-536239.880 [-1.620]
hoyde_cm	15603.357 *** [6.174]	5510.633 ** [2.936]
age	-5190.421 * [-2.110]	-720.749 [-0.419]
vekt_kg	-6078.851 * [-2.488]	271.764 [0.150]
maritalmarried	208713.507 *** [13.340]	118060.599 *** [10.241]
maritalseparated	-32269.415 [-1.166]	-30090.445 [-1.448]
maritaldivorced	70094.685 *** [3.881]	56777.105 *** [4.277]
maritalwidowed	37500.959 [0.963]	-1371.838 [-0.048]
bmi	13715.404 * [1.961]	-2500.027 [-0.477]
N	6911	5054
R2	0.088	0.082
logLik	-99855.098	-70408.779
AIC	199730.197	140837.558

*** p < 0.001; ** p < 0.01; * p < 0.05.

Utfra informasjonen i tabellen over kan vi se at det er stor forskjell på datasettene vi har brukt (*hoyde* og *hoyde_filter*). Endringene vi har gjort i datasettet har altså ført til betydelige endringer i resultatet.

Som vi kan lese utfra tabellen har populasjonen i datasettet gått fra 6911 til 5054, R2 har

gått fra 0.088 til 0.082. Dette betyr at i vårt filtrerte datasett kan 8,2% av av variasjonen i y (inntekt) kan forklares av x (høyde). Dette er ikke veldig høyt.

De viktigste variablene i vårt datasett har vært høyde (*hoyde_cm*) og inntekt (*inntekt_nok*). Utfra tabellen kan vi se at ved alle dataene inkludert vil 1 cm øke årsinntekt med 6911 kroner, mens i det filtrerte datasettet vil den øke med 5054 kr. Endringer i signifikansnivå her.

Residualer i datasettet

Nå skal vi legge residualene fra modellen vår inn i datasettet.

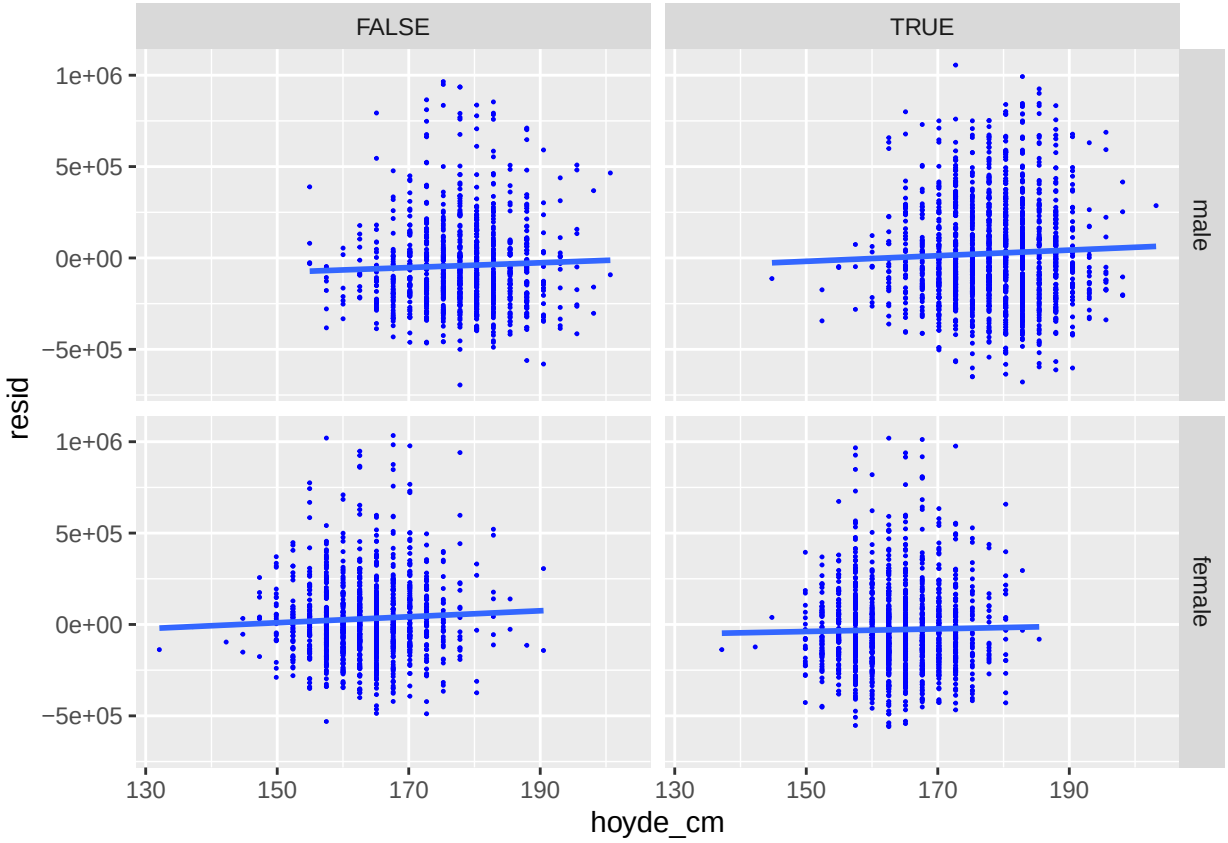
```
# Her legger dere residualene fra lm_hoyde_filter inn i hoyde
# Tror dere mente å legge dem inn i lm_hoyde_filter
# Ønsker å se hvor stor sammenheng det er mellom residualene og høyde når vi
# har korrigert for andre faktorer. Kan da ikke ha høyde som variabel i modellen
# legger derfor heller inn kjønn og utdanning (se løsning for interaksjonsmodell)
mod_ag <- 'inntekt_nok ~ sex + education + age + vekt_kg + marital'
lm_hoyde_filter_ag <- lm(mod_ag, data = hoyde_filter, subset = complete.cases(hoyde_filt
hoyde_filter_ag <- hoyde_filter %>%
  add_residuals(lm_hoyde_filter_ag)
#summary(hoyde)
```

```
ggplot(data = hoyde_filter_ag, mapping = aes(x = hoyde_cm, y = resid)) +
  facet_grid(sex ~ married) +
  geom_point(
    # satt under ggplot() så trengs ikke her
    #data = hoyde_filter,
    # trenger ikke mapping her når du har den samme i ggplot()
    # mapping = aes(x = hoyde_cm, y = inntekt_nok),
    colour = "blue",
    size = 0.2
```

```
    ) +
    geom_smooth(formula = 'y ~ x', method = 'lm', se = FALSE)
```

```
## Warning: Removed 71 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 71 rows containing missing values (geom_point).
```



Med mod_ag er det en svak postiv sammenheng mellom høyde og inntekt. Se løsningsforslag for en enkel interaksjonsmodell som ser ut til å fjerne all sammenheng mellom høyde og inntekt.

Konklusjon: Er det høyde som bestemmer inntekt?

Ved hjelp av analysene og regresjon, med ulike variabler ser vi at får kun en forklaringsgrad på 8,2% når det gjelder variasjonen i inntekt som kan forklares av høyde/x. Dette er relativt

lavt, og betyr at det er andre variabler som er vesentlig mer relevante for inntektsnivå. Vår konklusjon er at høyde ikke har noen innvirkning på inntekt.

Referanser

Judge, T. A., and Cable, D. M. (2004). The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model. *Journal of Applied Psychology*, *89*(3), 428–441.

Appendiks

```
library(tidyverse)
library(tinytex)
library(ISLR) #for the dataset
library(modelr)
library(knitr)
library(ggpubr)
library(huxtable)
data('heights', package = 'modelr')
knitr::kable(summary(heights[1:8]), "pipe")
hoyde <-   heights %>%
mutate(heights, hoyde_cm = height*2.54, #høyde i cm = hoyde i tommer * 2,54 fordi 1 tom
       vekt_kg = weight/2.2, #vekt i kg = vekt i pound / 2.2 fordi 2,2 kg = 1 pound
       inntekt_nok = income*8.5,  #inntekt i nok = inntekt i dollar * 8,5 fordi 1 dolla
        married = factor(
          case_when(
            marital == 'married' ~ TRUE,
            TRUE ~ FALSE)
          )
       )
hoyde$bmi <- hoyde$vekt_kg/(hoyde$hoyde_cm/100)/(hoyde$hoyde_cm/100) #bmi = vekt i kg /
hoyde$weight <- NULL #fjerner vekt i pounds fra datasett
hoyde$height <- NULL #fjerner høyde i tommer fra datasett
```

```r
hoyde$income <- NULL #fjerner inntekt i dollar fra datasett

knitr::kable(summary(hoyde[1:5]))
knitr::kable(summary(hoyde[6:10]), "pipe")

Høyde <- hoyde$hoyde_cm
Inntekt <- hoyde$inntekt_nok

ggplot(hoyde, aes(Inntekt, Høyde)) +  theme_dark() +
  geom_point(color= "blue", size = 0.5) +
geom_smooth(formula = y ~ x, method = "lm", color= "orange")

ggplot(data = hoyde,
       aes(x = inntekt_nok)) +
  geom_histogram(bins = 30, col = "orange", fill = "blue") + theme_dark()

sum(hoyde$inntekt_nok == 0)
max(hoyde$inntekt_nok)

Inntekt <- hoyde$inntekt_nok
ggplot(hoyde = Inntekt) + theme_dark() +
  geom_density(aes(x = Inntekt), fill = "blue")

Høyde <- hoyde$hoyde_cm
Inntekt <- hoyde$inntekt_nok
ggplot(hoyde = Høyde) + theme_dark() +
  geom_density(aes(x = Høyde), fill = "blue")
library(dplyr)
hoyde_filter = filter(hoyde, inntekt_nok != 0, inntekt_nok != "2922555")
```

```r
Høyde <- hoyde_filter$hoyde_cm
Inntekt <- hoyde_filter$inntekt_nok
ggplot(hoyde_filter, aes(Inntekt, Høyde)) +  theme_dark() +
  geom_point(color= "blue", size = 0.5) +
geom_smooth(formula = y ~ x, method = "lm", color= "orange")

ggplot(data = hoyde_filter,
       aes(x = inntekt_nok)) +
  geom_histogram(bins = 30, col = "orange", fill = "blue") + theme_dark()

(lm(inntekt_nok ~ hoyde_cm, data = hoyde)) %>%
  summary()
(lm(inntekt_nok ~ hoyde_cm, data = hoyde_filter)) %>%
  summary()
hoyde_semi= filter(hoyde, inntekt_nok != "2922555")

(lm(inntekt_nok ~ hoyde_cm, data = hoyde_semi)) %>%
  summary()
model <- lm(inntekt_nok ~ hoyde_cm + education + sex, data = hoyde_filter)
summary(model)
ggplot(data = hoyde_filter, mapping = aes(x = hoyde_cm, y = inntekt_nok)) +
  geom_point(
    data = hoyde_filter,
    mapping = aes(x = hoyde_cm, y = inntekt_nok),
    colour = "blue",
    size = 0.2
    ) +
  facet_grid(sex ~ factor(married, labels = c("not married", "married")))
lm_hoyde <- (lm(
```

```
  inntekt_nok ~ hoyde_cm + age + vekt_kg + marital + bmi,
                data = hoyde))
lm_hoyde_filter <- (lm(
  inntekt_nok ~ hoyde_cm + age + vekt_kg + marital + bmi,
                data = hoyde_filter))
huxreg(
  list("Hoyde"=lm_hoyde, "Hoyde_filter"=lm_hoyde_filter),
  error_format = "[{statistic}]",
  borders = 0.5,
  outer_borders = 0.8,
  error_pos = "same")
# Her legger dere residualene fra lm_hoyde_filter inn i hoyde
# Tror dere mente å legge dem inn i lm_hoyde_filter
# Ønsker å se hvor stor sammenheng det er mellom residualene og høyde når vi
# har korrigert for andre faktorer. Kan da ikke ha høyde som variabel i modellen
# legger derfor heller inn kjønn og utdanning (se løsning for interaksjonsmodell)
mod_ag <- 'inntekt_nok ~ sex + education + age + vekt_kg + marital'
lm_hoyde_filter_ag <- lm(mod_ag, data = hoyde_filter, subset = complete.cases(hoyde_filt
hoyde_filter_ag <- hoyde_filter %>%
  add_residuals(lm_hoyde_filter_ag)
#summary(hoyde)
ggplot(data = hoyde_filter_ag, mapping = aes(x = hoyde_cm, y = resid)) +
  facet_grid(sex ~ married) +
  geom_point(
    # satt under ggplot() så trengs ikke her
    #data = hoyde_filter,
    # trenger ikke mapping her når du har den samme i ggplot()
    # mapping = aes(x = hoyde_cm, y = inntekt_nok),
    colour = "blue",
    size = 0.2
```

```
  ) +
  geom_smooth(formula = 'y ~ x', method = 'lm', se = FALSE)
```