Har en persons høyde innvirkning på lønnen?

Assignment 2 i MSB105 Data Science - innleveringsfrist 12.10.20

Katrine Hope Karl-Gunnar Severinsen

Contents

Introduksjon

Vi har fått i oppgave å se på om det kan være en sammenheng mellom høyde og inntekt. Stemmer det egentlig at man tjener mer dess høyere man er?

Vi bruker datasettet ***heights*** til *National Longitudinal Study* (U.S. Bureau of Labor Statistics) (finnes i `modelr`-pakken), for å utføre statistiske analyser for å se om vi kan finne noen momenter som kan påvirke inntektsnivå opp mot høyde, samt se om det er andre faktorer som har en påvirkningskraft.

I analysedelen vil vi benytte oss av forskjellige typer *plots* og gjerne knytte datasettet opp mot enkle regresjonsmodeller for å prøve å svare på spørsmålet vårt.

Analyse

Oversikt

Vi starter med å lage en kolonne der høyden er vist i centimeter og inntekten i norske kroner. Vi gjør dette for å kunne få en bedre og mer forståelig analyse, da vi vil få en bedre forståelse ved å benytte kjente verdier. Deretter sorterer vi utvalget i datasettet inn i 10 intervaller med sammendragsstatistikk, for å gi en kjapp oversikt:

```
##      weight           age             marital         sex         education
##  Min.   : 76.0   Min.   :47.00   single   :1124   male  :3402   Min.   : 1.00
##  1st Qu.:157.0   1st Qu.:49.00   married  :3806   female:3604   1st Qu.:12.00
##  Median :184.0   Median :51.00   separated: 366                 Median :12.00
##  Mean   :188.3   Mean   :51.33   divorced :1549                 Mean   :13.22
##  3rd Qu.:212.0   3rd Qu.:53.00   widowed  : 161                 3rd Qu.:15.00
##  Max.   :524.0   Max.   :56.00                                  Max.   :20.00
##  NA's   :95                                                     NA's   :10
```

```
##       afqt           inntekt          height_cm        height_cmInt
##  Min.   :  0.00   Min.   :      0   Min.   :132.1   (163,173]:2298
##  1st Qu.: 15.12   1st Qu.:   1490   1st Qu.:162.6   (173,183]:1957
```

```
##  Median : 36.76   Median : 266306   Median :170.2   (152,163]:1778
##  Mean   : 41.21   Mean   : 370835   Mean   :170.4   (183,193]: 628
##  3rd Qu.: 65.24   3rd Qu.: 495000   3rd Qu.:177.8   (142,152]: 285
##  Max.   :100.00   Max.   :3094470   Max.   :213.4   (193,203]:  48
##  NA's   :262                                        (Other)  :  12
```

Her ser vi statistikk på blant annet vekt (i lbs), alder, sivilstatus, kjønn og utdanning inntekt i NOK og høyde (i cm).

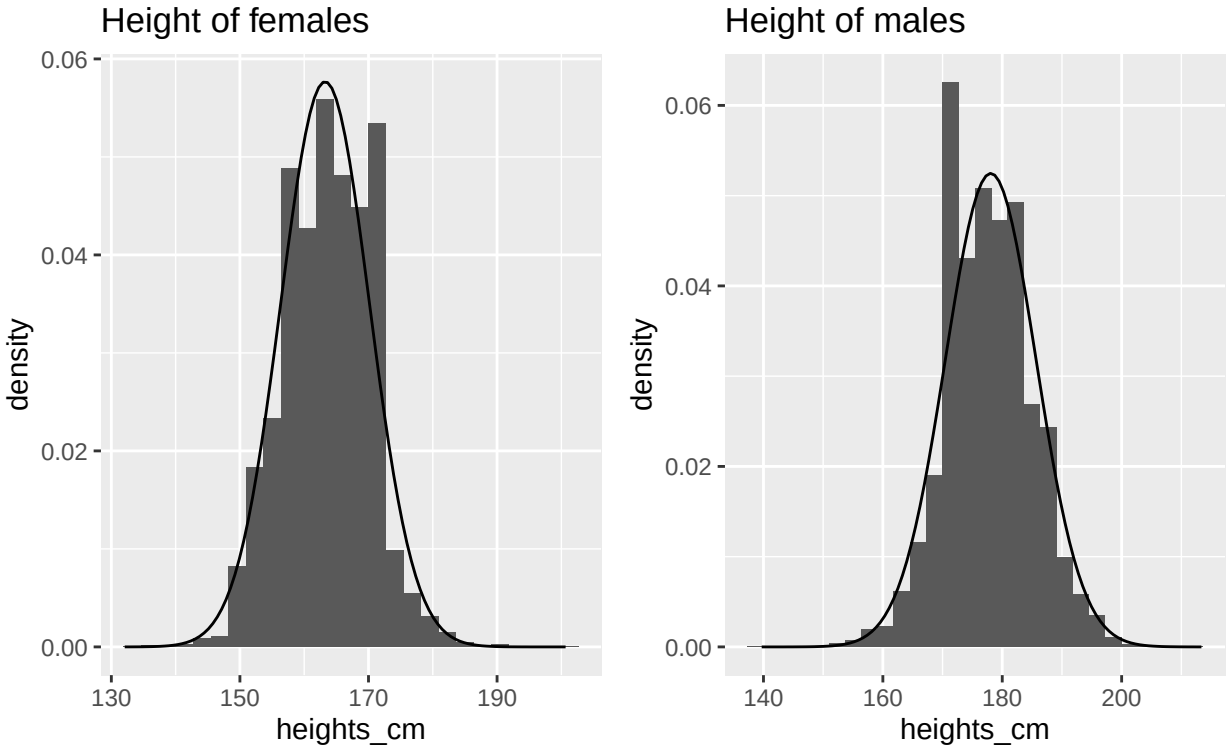
Under kjører vi samme oversikten, der vi ønsker se på hvordan den ser ut dersom vi begrenser inntekten til 1.600.000 NOK. Vi ønsker å gjøre dette får å utelukke de høye variablene som i følge *help-funksjonen* er beregnet gjennomsnittsinntekt av de 2 prosentene med høyest lønn.

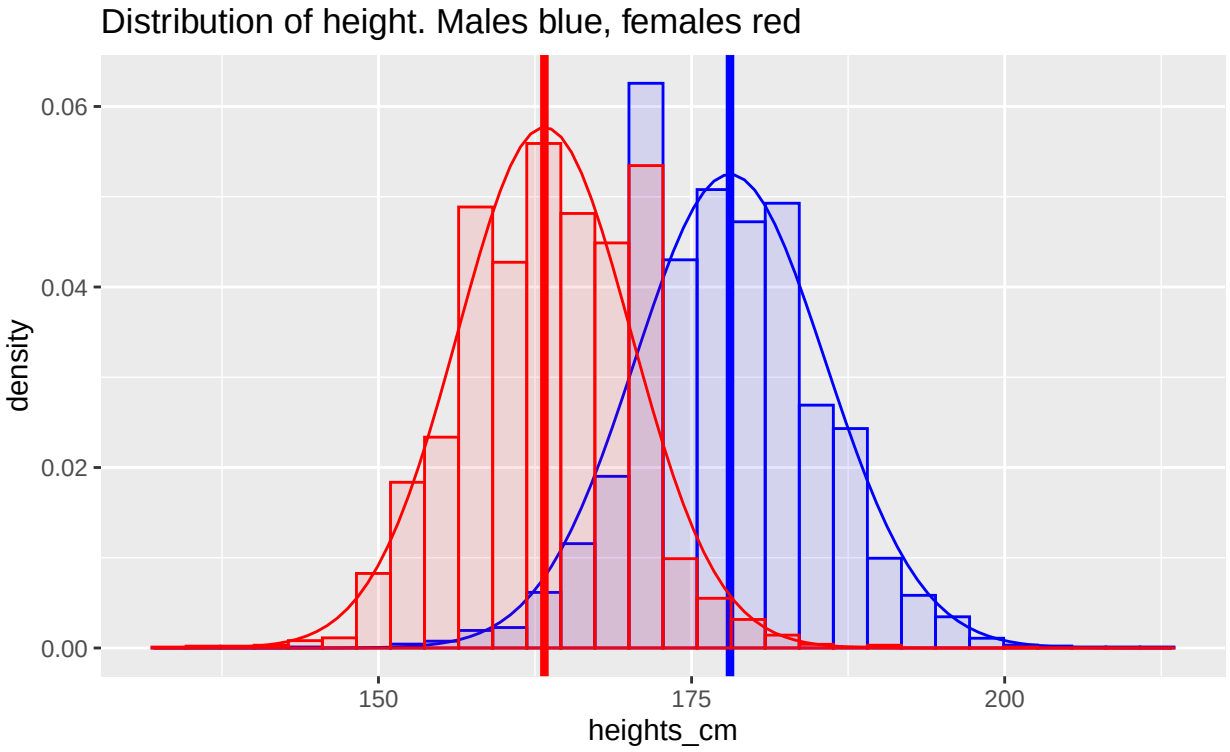
weight	age	marital	sex	education
Min. : 76.0	Min. :47.00	single :1113	male :3269	Min. : 1.00
1st Qu.:156.0	1st Qu.:49.00	married :3686	female:3592	1st Qu.:12.00
Median :183.0	Median :51.00	separated: 366	NA	Median :12.00
Mean :188.2	Mean :51.33	divorced :1536	NA	Mean :13.14
3rd Qu.:212.0	3rd Qu.:53.00	widowed : 160	NA	3rd Qu.:14.00
Max. :524.0	Max. :56.00	NA	NA	Max. :20.00
NA's :95	NA	NA	NA	NA's :10

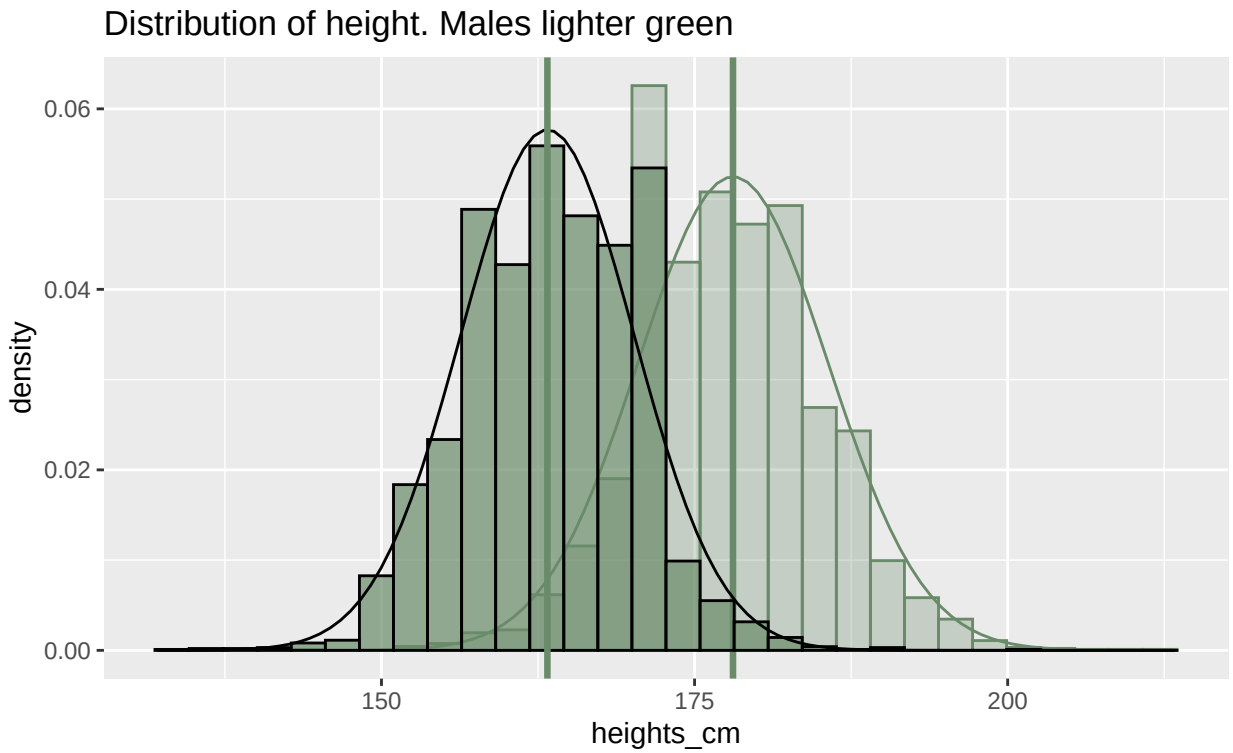
```
##       afqt           inntekt          height_cm       height_cmInt
##  Min.   :  0.00   Min.   :      0   Min.   :132.1   (163,173]:2270
##  1st Qu.: 14.75   1st Qu.:      0   1st Qu.:162.6   (173,183]:1877
##  Median : 35.66   Median : 252000   Median :170.2   (152,163]:1773
##  Mean   : 40.40   Mean   : 313709   Mean   :170.3   (183,193]: 600
##  3rd Qu.: 63.72   3rd Qu.: 479592   3rd Qu.:177.8   (142,152]: 285
##  Max.   :100.00   Max.   :1575000   Max.   :213.4   (193,203]:  45
##  NA's   :262                                        (Other)  :  11
```

1-variabels analyser

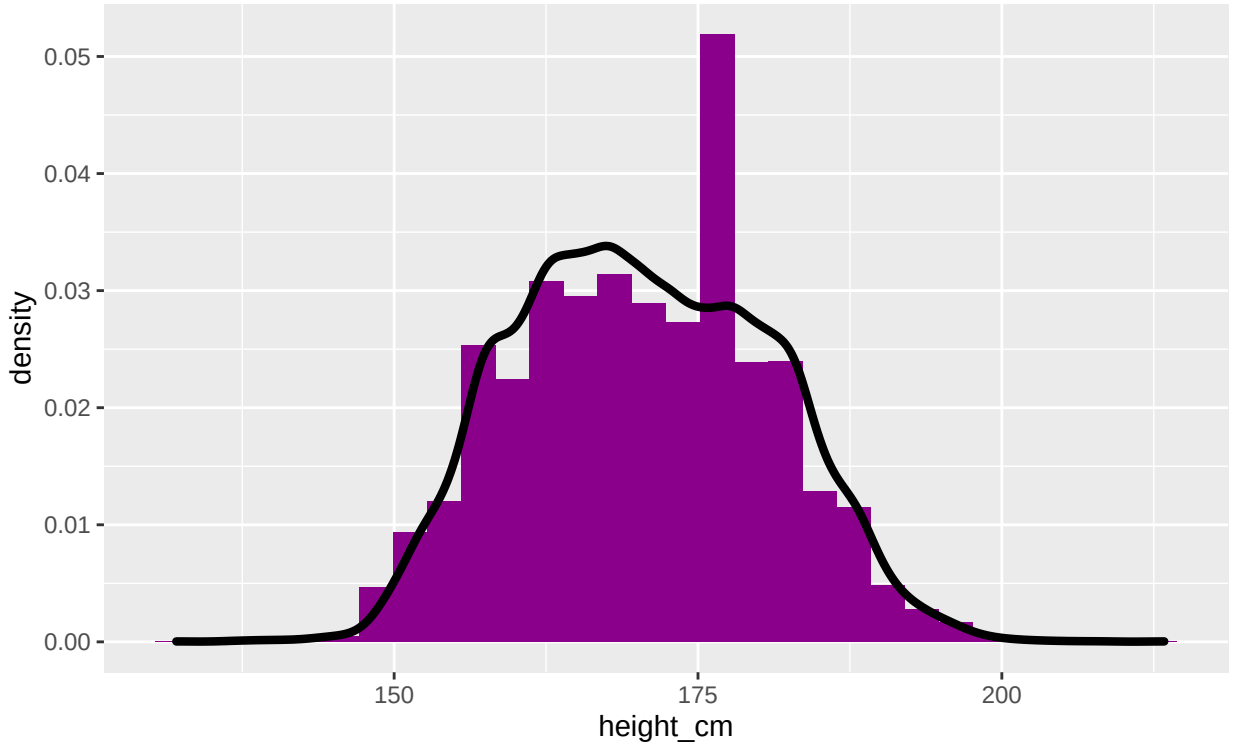
Videre har vi laget noen forskjellige oversikter for variablene *height_cm* og *inntekt* for å kunne vurdere om variablene er normalfordelte eller ikke. Vi starter med et histogram.



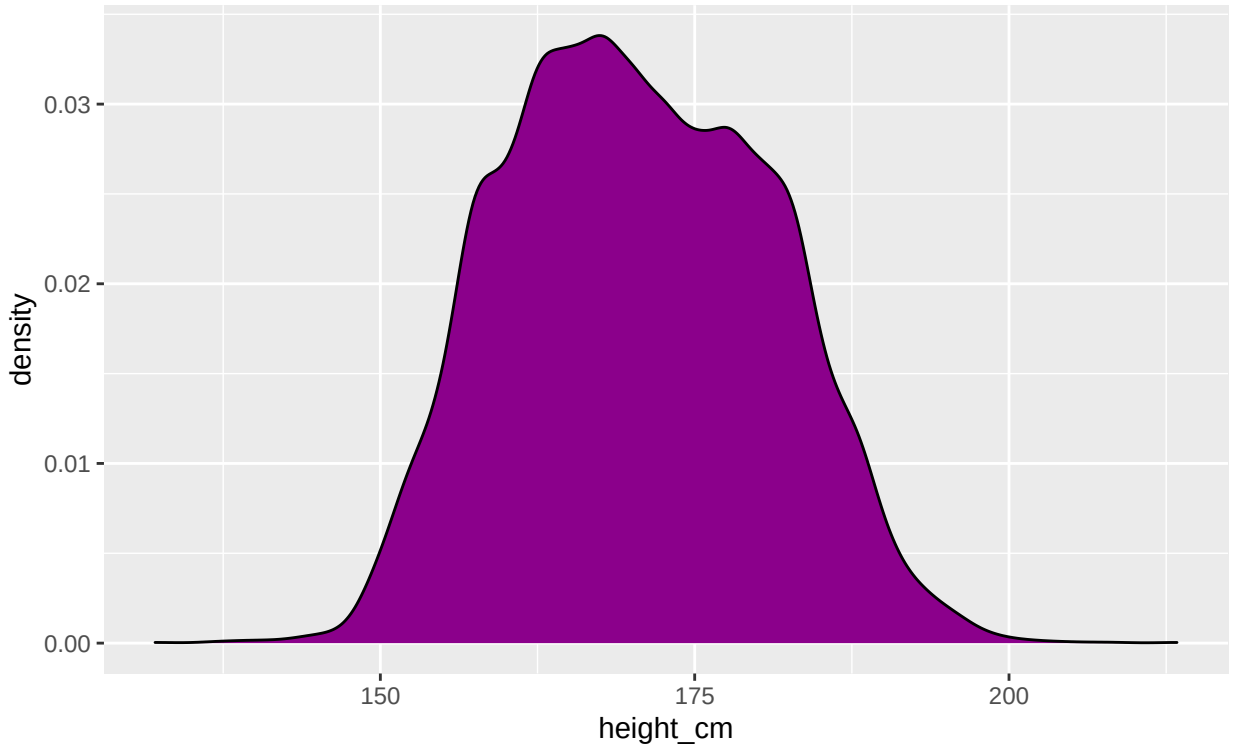




I histogrammet over ser vi høyde i centimeter opp mot frekvens. Histogrammet reflekterer *height_cmInt* fra oversikten i forrige kapittel. Vi kan også se at fordelingen er tilnærmet normalfordelt, med hovedvekten av observasjonene ligger mellom 160 til 180 centimeter.

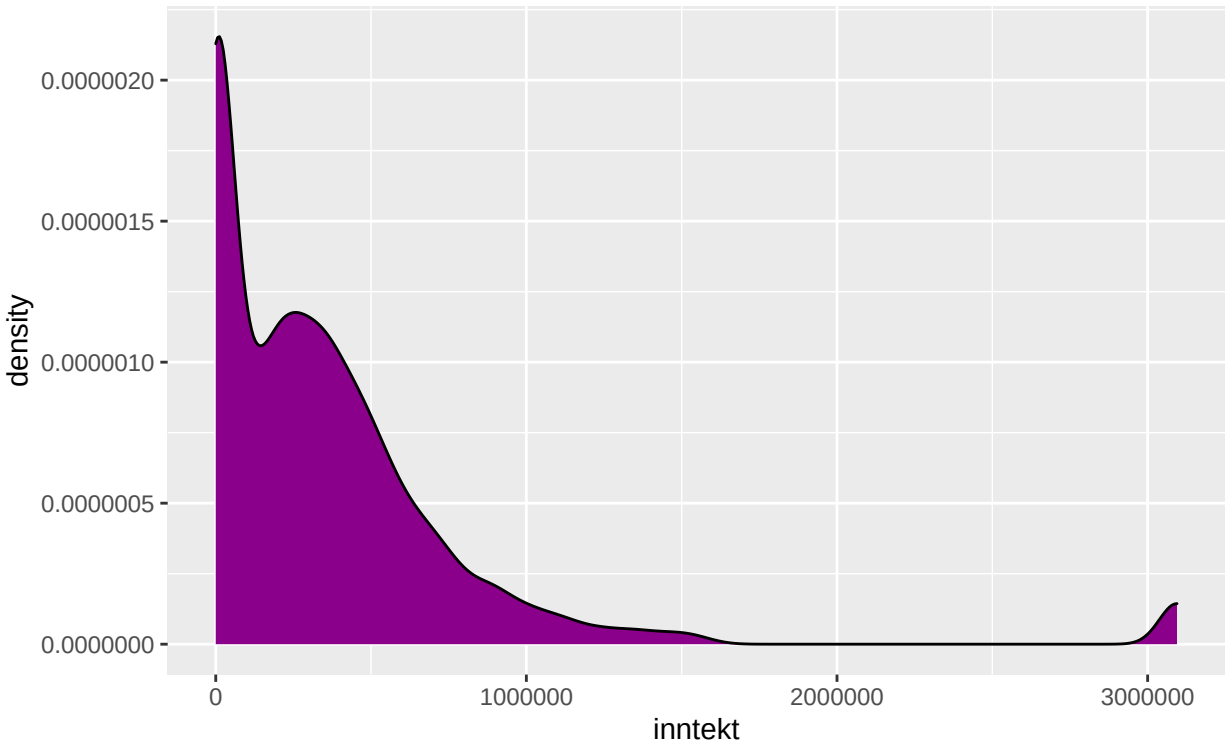


Over har vi et forsøk på å vise normalfordelingskurven sammen med histogrammet. Kurven forsvinner ut av grafen, men gir likevel en god indikasjon på at høyden i centimeter er tilnærmet normalfordelt.



Over har vi et density plot, som i grunn viser det samme som histogrammet over, men ved hjelp av en jevn kurve som viser tettheten av observasjonene. Igjen ser vi at fordelingen er tilnærmet normalfordelt.

Vi ønsker også å se på fordelingen i inntekt.

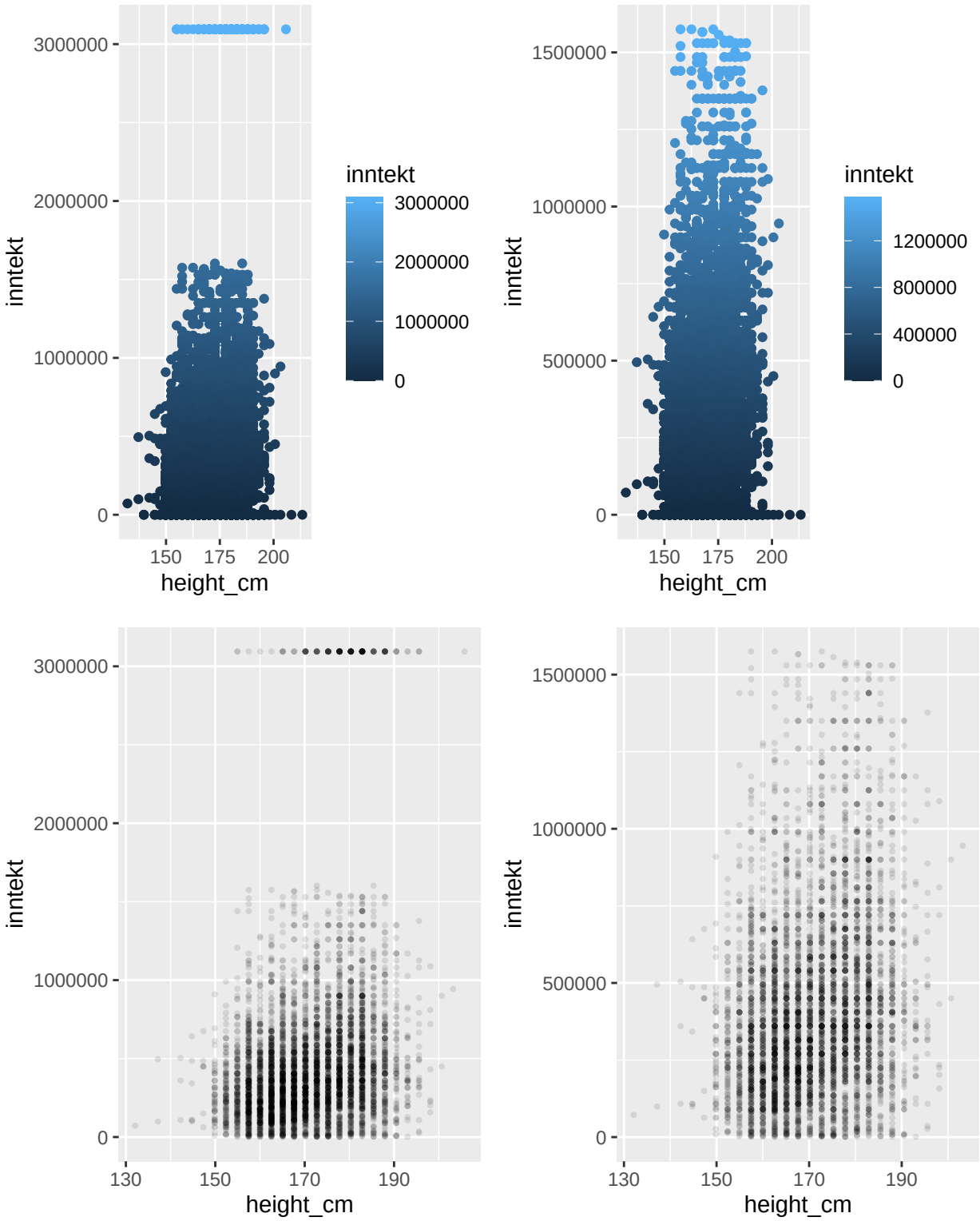


Her ser vi tydelig at inntekten ikke er normalfordelt. Dette har sammenheng med at det er mange observasjoner rundt 0, samt de 2% med høyest lønn er vist ved et gjennomsnitt av deres lønn. Dette medfører null observasjoner mellom 1.600.000 og 3.000.000, og en stor tetthet rundt 3-400.000. Vi så også dette tidligere i oversikten, der vi får en median på 266.000 og et gjennomsnitt på 370.000.

2-variablers analyse

Vi har til nå sett på variablene *height_cm* og *inntekt* hver for seg. For å kunne vurdere om de har noe sammenheng, må vi putte dem inn i samme plot.

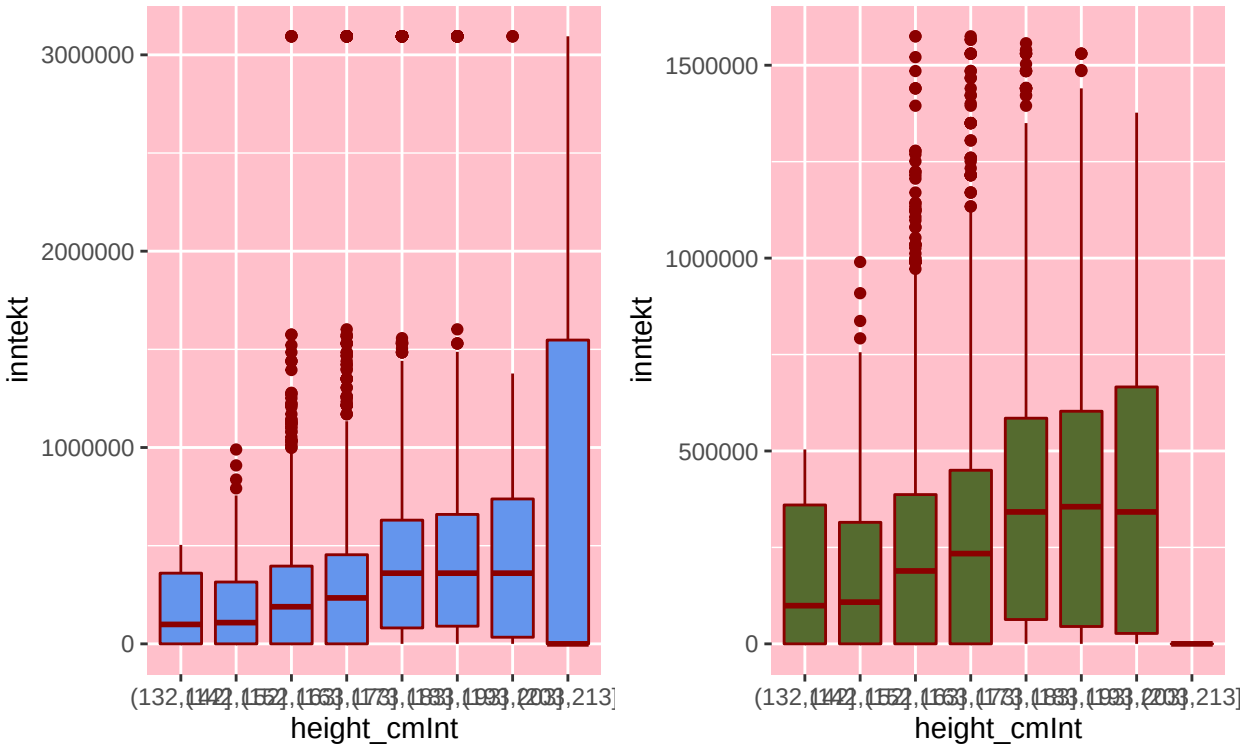
Vi velger å kjøre to *plots* side ved side nedover. Plottet til venstre vil inneholde alle observasjonene fra datasettet, som da inkluderer *outliers*. På høyresiden har vi valgt å begrense inntekten til 1.600.000 NOK, da dette vil ekskludere de øverste 2% som kan være forstyrrende for å få et korrekt bilde av analysen.

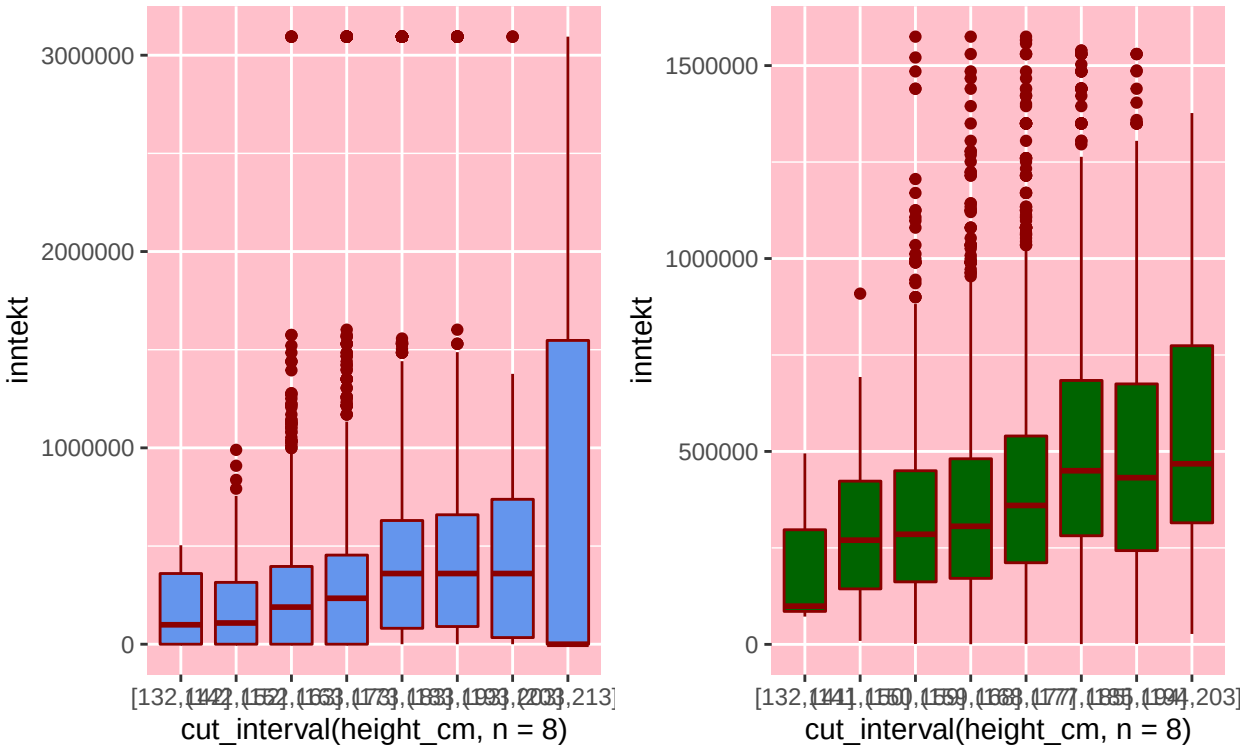


Til venstre ser vi ulempen med å inkludere de med høyest lønn, da det ikke er noen observasjoner overhodet mellom 1.600.000 og 3.000.000. Vi kan også se at de på toppen er fordelt gjennom hele høydespekteret.

I begge plottene kan vi for øvrig se at alle observasjonene er jevnt fordelt over hele høyde- og inntektsspekteret. Dette kan være en indikasjon på at høyde ikke har noe relevans for hvor mye en person tjener.

Vi kan også vise dette ved hjelp av et *boxplot*, der vi grupperer observasjonene i høydeintervaller på 10cm per boks.

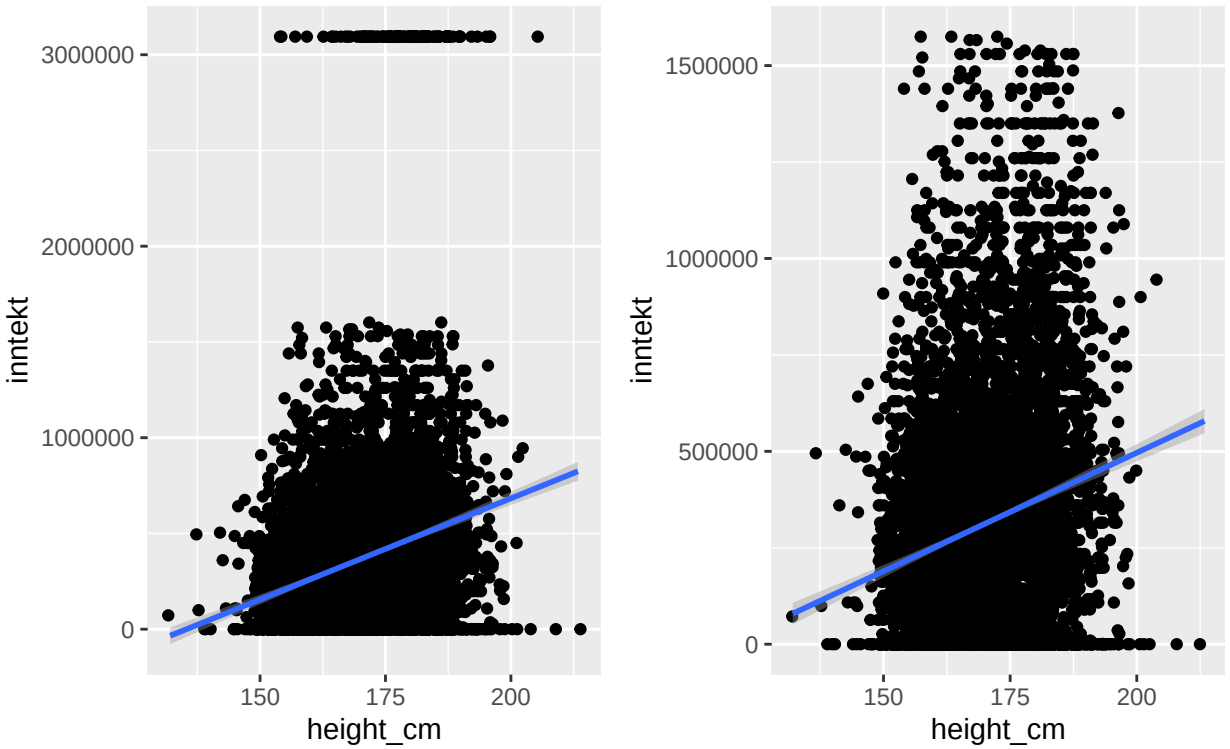




I *boxplottet* der selve boksen representerer første og tredje kvantil, i tillegg ser vi medianen i form av streken som går gjennom hver boks. De tynne strekene over/under boksene, kalles *whiskers*, og inkluderer observasjoner inntil 1.5% utover boksene.

Vi får her i stor grad det samme bilde som i plottet over. I dette tilfellet får vi *outliers* i begge grafene, noe som har en sammenheng med at både median- og gjennomsnittslønn er relativt lav i forhold til alle observasjonene. Vi ser at vi kunne redusert *outliers* ytterliggere ved å begrense datasettet til å kun inkludere de med inntekt opp til 1.000.000, men vi føler ikke dette vil gi et like riktig bilde.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



I modellene over har vi valgt å bruke *jitter*-funksjonen som viser resultatet mer spredt slik at vi får litt mer "*støy*". Dette gjør at dataene smelter mer sammen, slik at det blir tydeligere å se hvor hovedkonsentrasjonen av observasjonene ligger.

Når vi da velger å legge inn kommandoen `geom_smooth` så får vi en regresjonslinje. Vi ser at helningen/stigningstallet til regresjonslinjen til venstre er høyere enn regresjonslinjen i modellen til høyre. Dette kommer av at når vi begrenser maks inntekt til kr 1.600.000 unngår vi gjennomsnittsinntekten av de 2 prosentene med høyest lønn som påvirker datasettet mye, siden det ikke er noen observasjoner mellom 1.600.000 og 3.000.000. Regresjonslinen til høyre ligger nærmere hovedvekten av observasjonene.

I alle observasjonene over ser vi at høyde ikke er en tydelig forklaring på inntekten til observasjonene. Vi kan også vise dette ved å kjøre en enkel lineær regresjonsmodell, og gjør dette for begge datasettene vi har benyttet over.

```
##
## Call:
## lm(formula = inntekt ~ height_cm, data = heights)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -824251 -283597  -98035  133939 2887452
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -1429992.5    96603.8  -14.80 <0.0000000000000002 ***
## height_cm      10565.4      565.7   18.68 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 491000 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: < 0.00000000000000022

##
## Call:
## lm(formula = inntekt ~ height_cm, data = heights_liminc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -578427 -250796  -56999  166790 1339805
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -732097.1    61736.8  -11.86 <0.0000000000000002 ***
## height_cm      6142.3      361.9   16.97 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

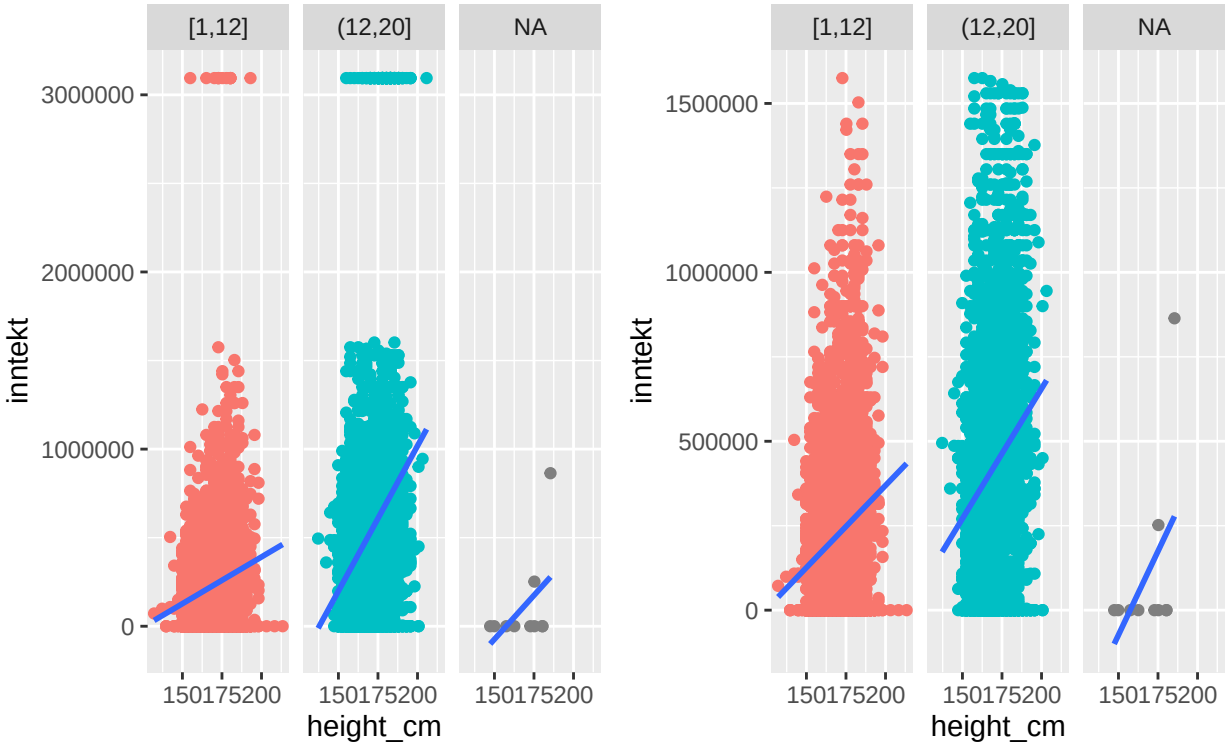
```
## Residual standard error: 309900 on 6859 degrees of freedom
## Multiple R-squared:  0.0403, Adjusted R-squared:  0.04016
## F-statistic:   288 on 1 and 6859 DF,  p-value: < 0.00000000000000022
```

Ved første øyekast kan det se ut som at én ekstra centimeters høyde, vil gi henholdsvis 10.565 eller 6.142 NOK ekstra i årslønn (avhengig av datasett). Men samtidig ser vi også tydelig at høyde ikke er en særlig relevant faktor for inntekten. Dette som følge av at i den første modellen ser vi av *R-squared* er på 0.0477, noe som tilsvarer en forklaringsgrad på kun 4.77%. I modellen under, der vi har begrenset inntektsnivået til 1.600.000, ser vi at høyden faktisk forklarer enda mindre med en forklaringsgrad på kun 4.02%.

Analyse med flere variabler

I delkapittelet om 2-variablers analyse så vi at høyde ikke hadde noe påvirkning på inntektsnivå, vi velger derfor å studere om andre faktorer kan være med å ha en påvirkningskraft.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Vi ser at når vi trekker inn utdanning som en ekstra variabel blir helningen/stigningstallet på regresjonslinjene høyere når man har utdanning mellom 12 til 20 år. Utdanning vil derfor ha en påvirkningskraft på inntektsnivået, som gir mening siden man gjerne har høyere lønn når man har høyere utdanning.

Vi kan også vise dette ved hjelp av en enkel regresjonsanalyse.

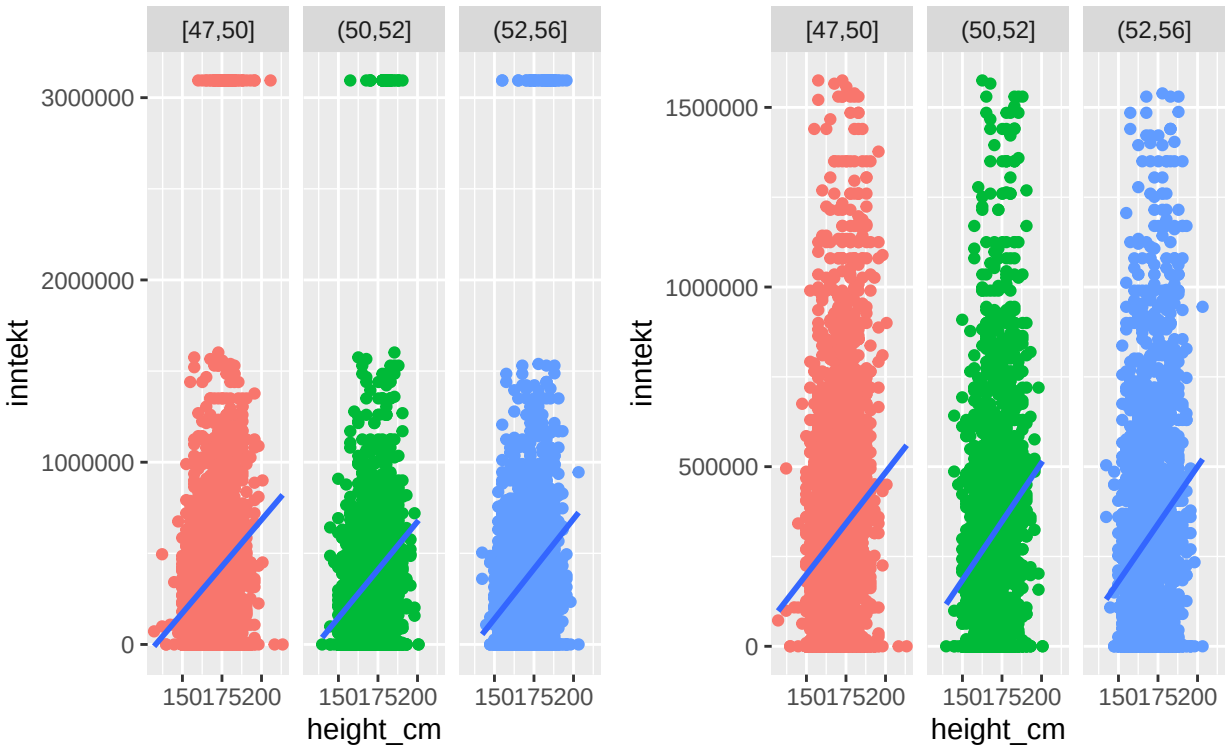
```
##
## Call:
## lm(formula = inntekt ~ education, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -887536 -278466  -70733  130119 3044406
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -635140      28666  -22.16 <0.0000000000000002 ***
## education      76134       2128   35.78 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462700 on 6994 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.1547, Adjusted R-squared:  0.1546
## F-statistic:  1280 on 1 and 6994 DF,  p-value: < 0.00000000000000022

##
## Call:
## lm(formula = inntekt ~ education, data = heights_liminc)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -653583 -221304  -41374  152074 1317696
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -337114      18368  -18.35 <0.0000000000000002 ***
## education      49535       1372   36.11 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 289900 on 6849 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.1599, Adjusted R-squared:  0.1598
## F-statistic:  1304 on 1 and 6849 DF,  p-value: < 0.00000000000000022
```

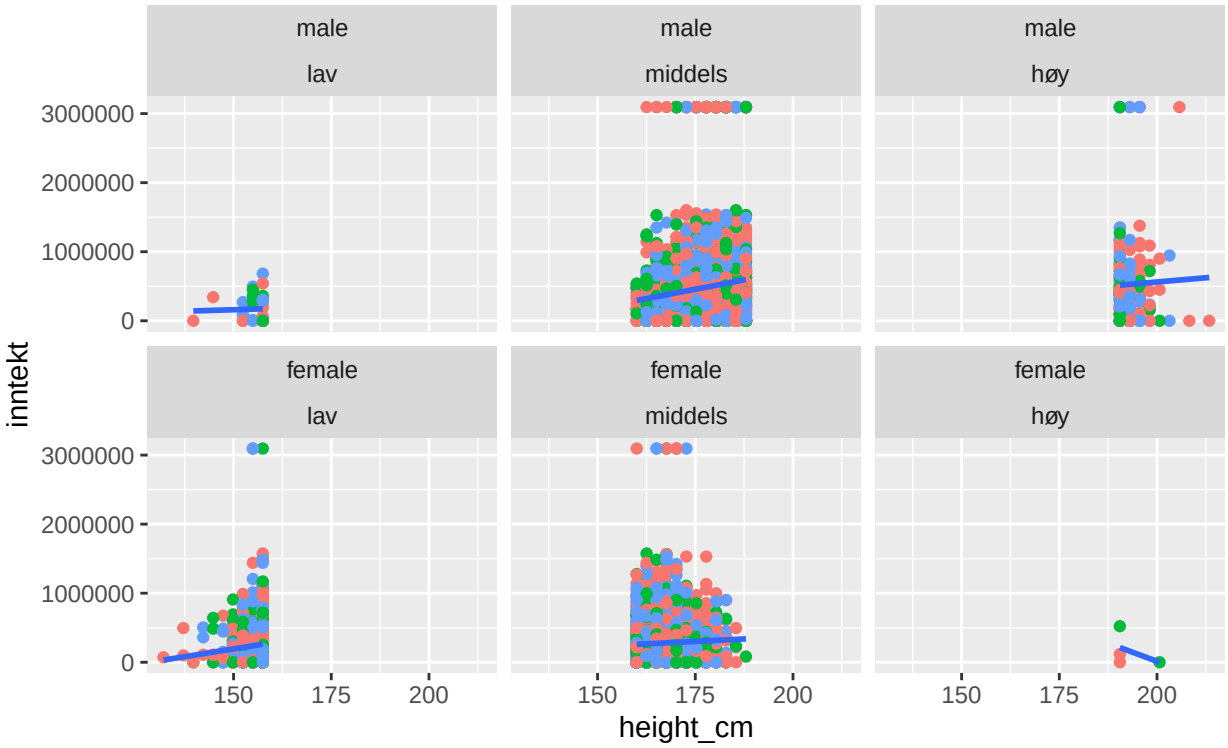
Her ser vi at utdanning har en forklaringsgrad på 15.46% og 15.98%, noe som kan sies å være rimelig forklarende sett for seg selv. Det gjenspeiler også den større endringen i plottet over, der vi nå ser en tydeligere differanse i regresjonslinjene.

Vi ønsker også å se på hvordan alder spiller inn på inntekten, da det vil være naturlig å anta at eldre gjerne tjener mer enn yngre.

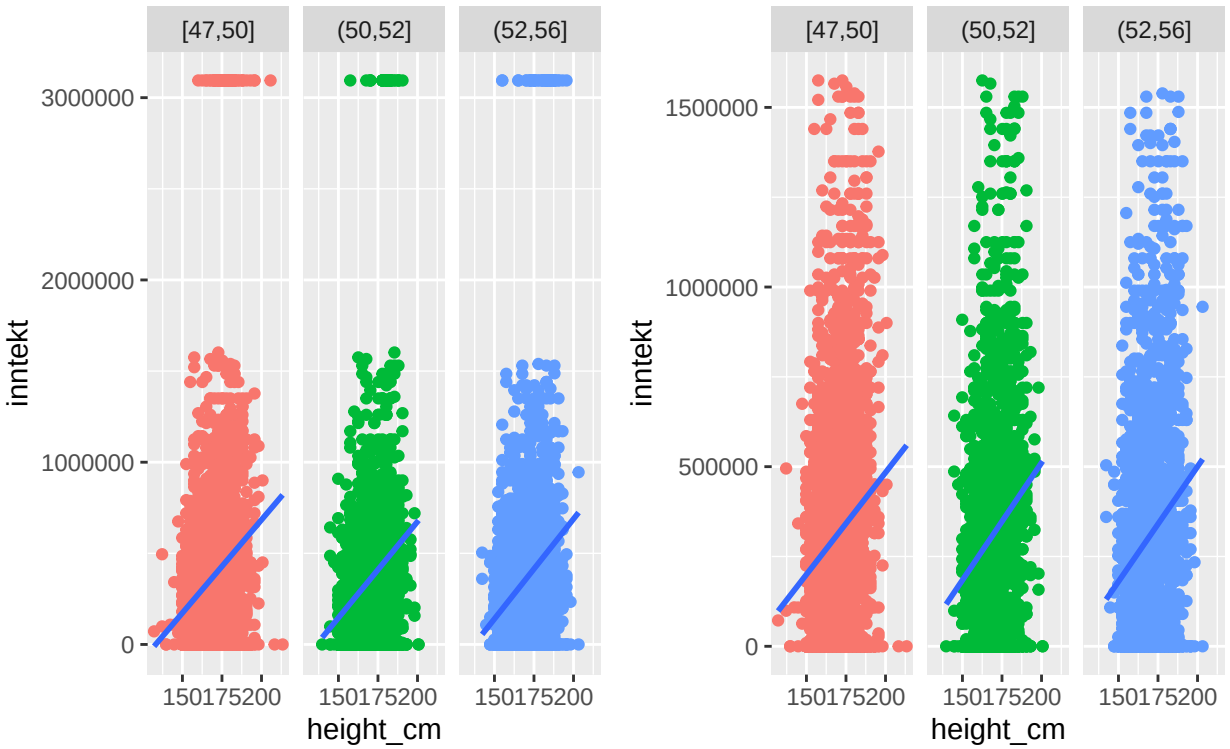
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
```




```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Her ser vi at det er veldig lite forskjell ved å inkludere alder som en faktor. Dette kan begrunnes med at datasettet har et relativt kort aldersspekter som kun går fra 47 til 56 år. Vi ser også at vi finner individer i alle aldre jevnt fordelt på forskjellig høyde og inntektsnivå.

Dette kan vi også vise ved å kjøre en enkel regresjon som viser relasjonen mellom inntekt og alder i våre aktuelle datasett.

```
##
## Call:
## lm(formula = inntekt ~ age, data = heights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -393156 -351871 -111513  127317 2742599
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   635705     138121   4.603 0.00000425 ***
```

```
## age            -5161       2688  -1.919      0.055 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 502900 on 7004 degrees of freedom
## Multiple R-squared:  0.0005258,  Adjusted R-squared:  0.0003831
## F-statistic: 3.684 on 1 and 7004 DF,  p-value: 0.05496

##
## Call:
## lm(formula = inntekt ~ age, data = heights_liminc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -325231 -303944  -62588  166090 1260412
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   450294      87829   5.127 0.000000303 ***
## age            -2661       1709  -1.557        0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316300 on 6859 degrees of freedom
## Multiple R-squared:  0.0003531,  Adjusted R-squared:  0.0002074
## F-statistic: 2.423 on 1 and 6859 DF,  p-value: 0.1196
```

Her ser vi akkurat det samme vi så i plottene over, at alder har en **veldig** lav forklaringsgrad på henholdsvis 0.038% og 0.02%.

Til nå har vi sett at de variablene som vi har sett på ikke egentlig forklarer så mye av inntekten til observasjonene i datasettene *heights* og *heights_liminc*. Helt til slutt ønsker vi

derfor å kjøre en regresjon, der vi inkluderer alle variablene sett opp mot inntekt. Dette gjør vi for å vurdere om hele datasettet kanskje er for mangelfullt eller har et for snevert spekter blant observasjonene.

```
##
## Call:
## lm(formula = inntekt ~ education + height_cm + sex + weight +
##     afqt + marital, data = heights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1056802  -225461   -46493   133032  2946379
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      -584323.5   137135.7  -4.261 0.000020640835651005 ***
## education          53478.9     2602.4  20.549 < 0.0000000000000002 ***
## height_cm           1036.6      807.1   1.284                0.1991
## sexfemale        -223653.6    15701.1 -14.244 < 0.0000000000000002 ***
## weight              -202.6      138.7  -1.460                0.1442
## afqt                3512.6      238.6  14.720 < 0.0000000000000002 ***
## maritalmarried    127179.1    15787.2   8.056 0.000000000000000928 ***
## maritalseparated   30191.3    27500.8   1.098                0.2723
## maritaldivorced    50193.8    17917.6   2.801                0.0051 **
## maritalwidowed     93263.4    38570.7   2.418                0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 441900 on 6635 degrees of freedom
##   (361 observations deleted due to missingness)
## Multiple R-squared:  0.2554, Adjusted R-squared:  0.2544
```

```
## F-statistic: 252.8 on 9 and 6635 DF,  p-value: < 0.00000000000000022

##
## Call:
## lm(formula = inntekt ~ education + height_cm + sex + weight +
##     afqt + marital, data = heights_liminc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -764427 -181465  -27554  147279 1266341
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      -303337.42   85436.87  -3.550             0.000387 ***
## education          34555.74    1631.41  21.182 < 0.0000000000000002 ***
## height_cm            409.08     502.65   0.814             0.415758
## sexfemale        -131130.71    9787.32 -13.398 < 0.0000000000000002 ***
## weight               -19.10      86.22  -0.222             0.824678
## afqt                2373.44     148.57  15.975 < 0.0000000000000002 ***
## maritalmarried    104372.28    9806.18  10.644 < 0.0000000000000002 ***
## maritalseparated   25624.56   16979.00   1.509             0.131298
## maritaldivorced    55390.35   11105.79   4.988          0.000000627 ***
## maritalwidowed     57159.40   23873.89   2.394             0.016684 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272500 on 6490 degrees of freedom
##   (361 observations deleted due to missingness)
## Multiple R-squared:  0.2665, Adjusted R-squared:  0.2655
## F-statistic: 262.1 on 9 and 6490 DF,  p-value: < 0.00000000000000022
```

Konklusjon

Som mistenkt, ser vi ovenfor at ved å kjøre en regresjon som inkluderer alle 6 variablene i datasettene, så får vi kun en forklaringsgrad på relativt lave 25.44% når vi inkluderer hele settet og 26.55% når vi begrenser inntekten til 1.600.000. Dette betyr at det fremdeles er ukjente variabler som spiller en vesentlig større rolle på inntekten enn de variablene som er observert. Samtidig så kan selve datasettet være hentet inn på dårlig grunnlag, noe som bidrar til at det er vanskeligere å trekke noe tydelig konklusjon på hva som faktisk har noe betydning for inntektsnivået.

Det vi kan konkludere med er at ***høyde har ingen innvirkning på inntekten***.

Appendiks

```
library(modelr)
library(ggplot2)
library(knitr)
library(tinytex)
library(tidyverse)
library(ggpubr)
library(dplyr)
options(scipen = 999)
# Jeg vil foreslå
# options(scipen = 2)
# som et kompromiss
heights$inntekt <- heights$income*9
heights$height_cm <- heights$height*2.54
heights$height_cmInt <- cut(heights$height_cm,
                            breaks = 8)
heights[,3:7] %>%
  summary()# %>% kable()
heights[,8:11] %>%
  summary()# %>% kable()
#skrevet som pipes
heights %>%
  filter(inntekt < 1600000) %>%
  select(3:7) %>%
  summary() %>%
  kable()

heights %>%
  filter(inntekt < 1600000) %>%
```

```
  select(8:11) %>%
  summary() # %>% kable()

# Vil foreslå å heller først lage en egen kopi av heights
# så benytte mutate() til å lage en ny variable.
my_heights <- heights
my_heights <- my_heights %>%
  mutate(
    heights_cm = heights$height*2.54,
    heights_cm_int = cut_width(heights_cm, width = 2*2.54)
  )
# mean og sd male heights
 h_male <-  my_heights %>%
  filter(sex == "male") %>%
  summarise(
    mu = mean(height_cm),
    sd = sd(height_cm)
    )
# mean og sd female heights
  h_female <-  my_heights %>%
  filter(sex == "female") %>%
  summarise(
    mu = mean(height_cm),
    sd = sd(height_cm)
    )
# Dette er ikke så lett. Hovedårsaken er at stat_function() er designet
# til å tegne **samme** funksjon over plotene. Vi trenger å tegne en funksjon for kvin
# female plot
 fe_plot <- my_heights %>%
```

```
   filter(sex == "female")  %>%
  ggplot() +
  geom_histogram(mapping = aes(x = heights_cm,
                               y = ..density..),
                 binwidth = 2.72
                 ) +
   stat_function(fun = dnorm,
                 args = list(mean = h_female$mu,
                             sd = h_female$sd)
                 ) +
   ggtitle("Height of females")
# male plot
 m_plot <- my_heights %>%
   filter(sex == "male")  %>%
  ggplot() +
  geom_histogram(mapping = aes(x = heights_cm,
                               y = ..density..
                               ),
                 binwidth = 2.72) +
   stat_function(fun = dnorm,
                 args = list(mean = h_male$mu,
                             sd = h_male$sd)
                 ) +
   ggtitle("Height of males")
# put the two histograms together
ggarrange(fe_plot, m_plot, ncol = 2, nrow = 1)
# Both curves and histogram in same figure
# garish colours
my_heights %>%
```

```
ggplot() +
### male ###
geom_histogram(data = filter(my_heights, sex == "male"),
               mapping = aes(x = heights_cm, y = ..density..),
               binwidth = 2.72,
               alpha = 0.1,
               colour = 'blue',
               fill = 'blue') +
# mark the mean
geom_vline(xintercept = h_male$mu,
           colour = 'blue',
           lwd = 1.5) +
# draw normal distribution for male mean height and sd
 stat_function(fun = dnorm,
              args = list(mean = h_male$mu, sd = h_male$sd),
              colour = 'blue') +
### female ###
geom_histogram(data = filter(my_heights, sex == "female"),
               mapping = aes(x = heights_cm, y = ..density..),
               binwidth = 2.72,
               alpha = 0.1,
               colour = 'red',
               fill = 'red') +
stat_function(fun = dnorm,
              args = list(mean = h_female$mu, sd = h_female$sd),
              colour = 'red') +
geom_vline(xintercept = h_female$mu,
           colour = 'red',
           lwd = 1.5) +
```

```
  ggtitle("Distribution of height. Males blue, females red")
# nicer colour scheme. One colour different alpha
my_heights %>%
  ggplot() +
  ### male ###
  geom_histogram(data = filter(my_heights, sex == "male"),
                 mapping = aes(x = heights_cm, y = ..density..),
                 binwidth = 2.72,
                 alpha = 0.3,
                 colour = 'darkseagreen4',
                 fill = 'darkseagreen4'
                 ) +
  geom_vline(xintercept = h_male$mu,
             colour = 'darkseagreen4',
             lwd = 1.15
             ) +
   stat_function(fun = dnorm,
                 args = list(mean = h_male$mu, sd = h_male$sd),
                 colour = 'darkseagreen4'
                 ) +
  geom_histogram(data = filter(my_heights, sex == "female"),
                 mapping = aes(x = heights_cm, y = ..density..),
                 binwidth = 2.72,
                 alpha = 0.7,
                 colour = 'black',
                 fill = 'darkseagreen4'
                 ) +
   stat_function(fun = dnorm,
                 args = list(mean = h_female$mu, sd = h_female$sd),
```

```
                colour = 'black'
                ) +
  geom_vline(xintercept = h_female$mu,
             colour = 'darkseagreen4',
             lwd = 1.15
             ) +
  ggtitle("Distribution of height. Males lighter green")
# NOTE! The option "Show output inline for all R MArkdown documents"
# under Preferences > Markdown > Basic MUST be turned off for the
# following to work. Then output wiil be sent to console and the
# Plots tab in the lower right corner.
# Code will work in normal scripts and when whole document is knit whitout
# this change in preferences. This is a problem concerning rmarkdown and
# R Notebooks
# height_cm <- heights$height * 2.54
# hist(height_cm,
#      breaks = 20,
#      freq = FALSE,
#      main = "Høyde i centimeter",
#      xlab = "Centimeter",
#      ylab = "Frekvens",
#      col = "darkmagenta")
# lines(density(height_cm), lwd = 3)
####
####
# Easy to do the same in ggplot and we don't have to change our prefs
mean_height_cm <- mean(heights$height)
sd_height_cm <- sd(heights$height)

heights %>%
```

```
  mutate(
    height_cm = height * 2.54
  ) %>%
  ggplot(mapping = aes(x = height_cm, y = ..density..)) +
  geom_histogram(bins = 30, fill = "darkmagenta") +
  geom_density(lwd = 1.5)
ggplot(data = heights) +
  geom_density(aes(x = height_cm),
                   fill = "darkmagenta")
ggplot(data = heights) +
  geom_density(aes(x = inntekt),
                   fill = "darkmagenta")
# Bruk fullt navn på parametrene, colour = ikke col =
# Gir fargekoding av inntekt egentlig noen ekstra info?
# Inntekt blir jo målt på vertikal akse og kan leses ut fra
# punktets plassering. Jeg ville heller lagt vekk på å få
# frem hvor vi har mange observasjoner og hvor vi har få.
# Se forslag til figur nedenfor. Ideen er at ved lav opasitet
# vil punktet blir mørkere når vi har mange oppå hverandre,
# dvs mørke punkter der vi har mange observasjoner.
heights_liminc <- heights %>%
  filter(inntekt < 1600000)
m1 <- ggplot(heights,
             mapping = aes(x = height_cm,
                           y = inntekt,
                           col = inntekt)) +
  geom_point()

m2 <- ggplot(heights_liminc,
```

```
            mapping = aes(x = height_cm,
                          y = inntekt,
                          col = inntekt)) +
  geom_point()

ggarrange(m1,m2)
m1 <- heights %>%
  filter(inntekt > 0) %>%
  ggplot(mapping = aes(x = height_cm,
                       y = inntekt
                       )
         ) +
  geom_point(size = 0.7, alpha = 0.1)

m2 <- heights %>%
  filter(inntekt > 0) %>%
  filter(inntekt < 1600000) %>%
  ggplot(mapping = aes(x = height_cm,
                       y = inntekt
                       )
         ) +
  geom_point(size = 0.7, alpha = 0.1)

ggarrange(m1,m2)
# Fin figur. Igjen vil jeg anbefale å jobbe med heights, eventuelt en
#  egen kopi, og så bruke filter, mutate, select etc på dette datasettet.
#  Mye mindre å holde styr på og gir ofte muligheter for copy-paste der
#  bare små endringer må gjøres. Se forslag nedenfor
m3 <- ggplot(heights,
```

```
                mapping = aes(
                  x = height_cmInt,
                  y = inntekt)) +
        geom_boxplot(colour= "darkred",
                     fill = "cornflowerblue") +
  theme(panel.background = element_rect(fill = "pink"))

m4 <- ggplot(heights_liminc,
               mapping = aes(
                 x = height_cmInt,
                 y = inntekt)) +
        geom_boxplot(colour = "darkred",
                     fill = "darkolivegreen") +
  theme(panel.background = element_rect(fill = "pink"))

ggarrange(m3, m4)
m3 <- heights %>%
  ggplot(mapping = aes(
                  x = cut_interval(height_cm, n = 8),
                  y = inntekt)
         ) +
        geom_boxplot(colour = "darkred",
                     fill = "cornflowerblue") +
  theme(panel.background = element_rect(fill = "pink"))

m4 <- heights %>%
  filter(inntekt > 0) %>%
  filter(inntekt < 1600000) %>%
  ggplot(mapping = aes(
```

```
                  x = cut_interval(height_cm, n = 8),
                  y = inntekt)
         ) +
        geom_boxplot(colour = "darkred",
                     fill = "darkgreen") +
  theme(panel.background = element_rect(fill = "pink"))

ggarrange(m3, m4)
m5 <- ggplot(heights,
             mapping = aes(x = height_cm,
                           y = inntekt)) +
  geom_point(position = "jitter") +
  geom_smooth(method = 'lm')

m6 <- ggplot(heights_liminc,
             mapping = aes(x = height_cm,
                           y = inntekt)) +
  geom_point(position = "jitter") +
  geom_smooth(method = 'lm')

ggarrange(m5, m6)
summary(lm(inntekt ~ height_cm,
           data = heights))

summary(lm(inntekt ~ height_cm,
           data = heights_liminc))
m7 <- ggplot(data = heights,
             mapping = aes(x = height_cm,
                           y = inntekt)
```

```
            ) +
    facet_wrap(~cut_number(education, n = 2)) +
  geom_point(aes(colour = cut_number(education,
                                     n = 2)),
             show.legend = F) +
  geom_smooth(method = "lm",
              se = FALSE)

m8 <- ggplot(data = heights_liminc,
             mapping = aes(x = height_cm,
                           y = inntekt)) +
    facet_wrap(~cut_number(education,
                           n = 2)) +
  geom_point(aes(colour = cut_number(education,
                                     n = 2)),
             show.legend = F) +
  geom_smooth(method = "lm",
              se = FALSE)

ggarrange(m7, m8)
summary(lm(inntekt ~ education,
           data = heights))

summary(lm(inntekt ~ education,
           data = heights_liminc))
m9 <- ggplot(data = heights,
             mapping = aes(x = height_cm,
                           y = inntekt)) +
    facet_wrap(~cut_number(age,
```

```
                            n = 3)) +
  geom_point(aes(colour = cut_number(age,
                                     n = 3)),
             show.legend = F) +
  geom_smooth(method = "lm",
              se = FALSE)

m10 <- ggplot(data = heights_liminc,
              mapping = aes(x = height_cm,
                            y = inntekt)) +
    facet_wrap(~cut_number(age,
                            n = 3)) +
  geom_point(aes(colour = cut_number(age,
                                     n = 3)),
             show.legend = F) +
  geom_smooth(method = "lm",
              se = FALSE)

ggarrange(m9, m10)
#Mitt forslag
heights %>%
  group_by(sex) %>%
  mutate(
    height_cmInt = cut_interval(height_cm, breaks = c(132, 160, 188, 216), labels = c("l
  ) %>%
  ungroup() %>%
  ggplot(mapping = aes(x = height_cm,
                            y = inntekt
                       )
```

```
        ) +
    facet_wrap(sex ~ height_cmInt) +
  geom_point(mapping = aes(colour = cut_number(age,
                                    n = 3)),
             show.legend = F) +
  geom_smooth(method = "lm",
              se = FALSE)

m10 <- ggplot(data = heights_liminc,
              mapping = aes(x = height_cm,
                            y = inntekt)) +
    facet_wrap(~cut_number(age,
                           n = 3)) +
  geom_point(aes(colour = cut_number(age,
                                    n = 3)),
             show.legend = F) +
  geom_smooth(method = "lm",
              se = FALSE)

ggarrange(m9, m10)
summary(lm(inntekt ~ age,
           data = heights))

summary(lm(inntekt ~ age,
           data = heights_liminc))
summary(lm(inntekt ~ education + height_cm + sex + weight + afqt + marital,
           data = heights))

summary(lm(inntekt ~ education + height_cm + sex + weight + afqt + marital,
```

```
            data = heights_liminc))
```