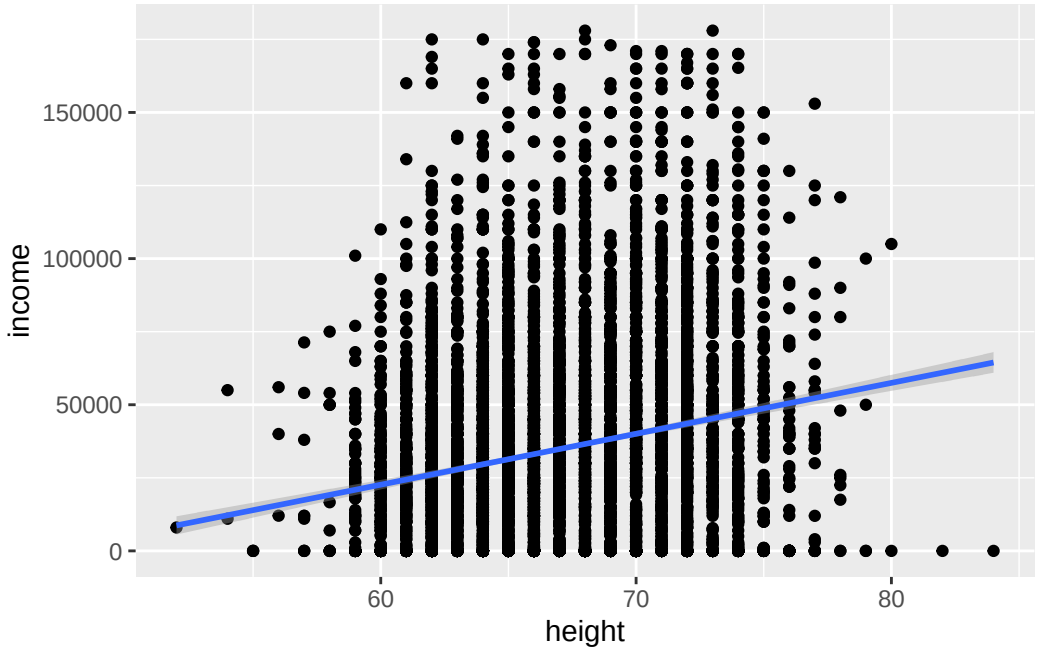
EDA

Se på variabler:

```
heights <- modelr::heights
```

```
heights |>
  select(income, height) |>
  filter(income < 300000) |>
  ggplot(mapping = aes(x = height, y = income)) +
  geom_point() +
  geom_smooth(method = "lm")
```

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




```
summary(heights)
```

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

NA i heights:

```
# NAs in heights?
heights %>%
  apply(MARGIN = 2, FUN = is.na) %>%
  apply(MARGIN = 2, FUN = sum)
```

```
income    height    weight       age   marital       sex education      afqt
     0         0        95         0         0         0        10       262
```

Får akkurat samme svar ved å bruke komandoen:

```
# NAs in heights?
heights %>%
  is.na() %>%
  apply(MARGIN = 2, FUN = sum)
```

```
income    height    weight       age   marital       sex education      afqt
     0         0        95         0         0         0        10       262
```

Her får vi bare opp de variablene som faktisk har NA verdier.

- Punktum betyr her dataene i pipen. Legger det inn i firkanklammer for å gi beskjed om hvilke verdier jeg vil ha med fra dataframen.

```
# number of NAs in each variable
# drop variables with no NA
heights %>%
  is.na() %>%
  colSums() %>%
  .[. > 0]
```

```
weight education      afqt
    95        10       262
```