Er det høyde som bestemmer inntekt?

Assignment 2 i MSB105 Data Science

Kevin Ha - 571821 Ola Andre Olofsson - 170745

Innledning

Dette er oppgave 2 i kurset MSB105 Data Science. I den følgende artikkelen anvendes datasettet **heights** fra pakken **modelr** for å besvare følgende problemstilling; **Er det høyde som bestemmer inntekt?**

En kort litteraturgjennomgang på ca. 1 side

Analyse med egen versjon av datasettet

I henhold til oppgaveteksten, angir vi datasettet for *hoyde*.

```
# Vi selekterer ut dataene for heights fra pakken modelr, og angir deretter benevnelsen "hoyde"

data('heights', package = 'modelr')
hoyde <- heights

# Vi rydder videre opp i benevnelsene ved å slik at de blir enklere å jobbe med. Vi oversetter dem til

hoyde$inntekt <- hoyde$income*8.5
hoyde$height_cm <- hoyde$height*2.54
kable(summary(hoyde[,9:10]))
```

inntekt	height_cm
Min. : 0	Min. :132.1
1st Qu.: 1407	1st Qu.:162.6
Median : 251511	Median :170.2
Mean : 350234	Mean :170.4
3rd Qu.: 467500	3rd Qu.:177.8
Max. :2922555	Max. :213.4

```
# Til slutt kan vi oppsummere de interessante variablene i metrisk form, samt oversatt.
```

Beskrivende statistikk (beskrivelse av dataer)

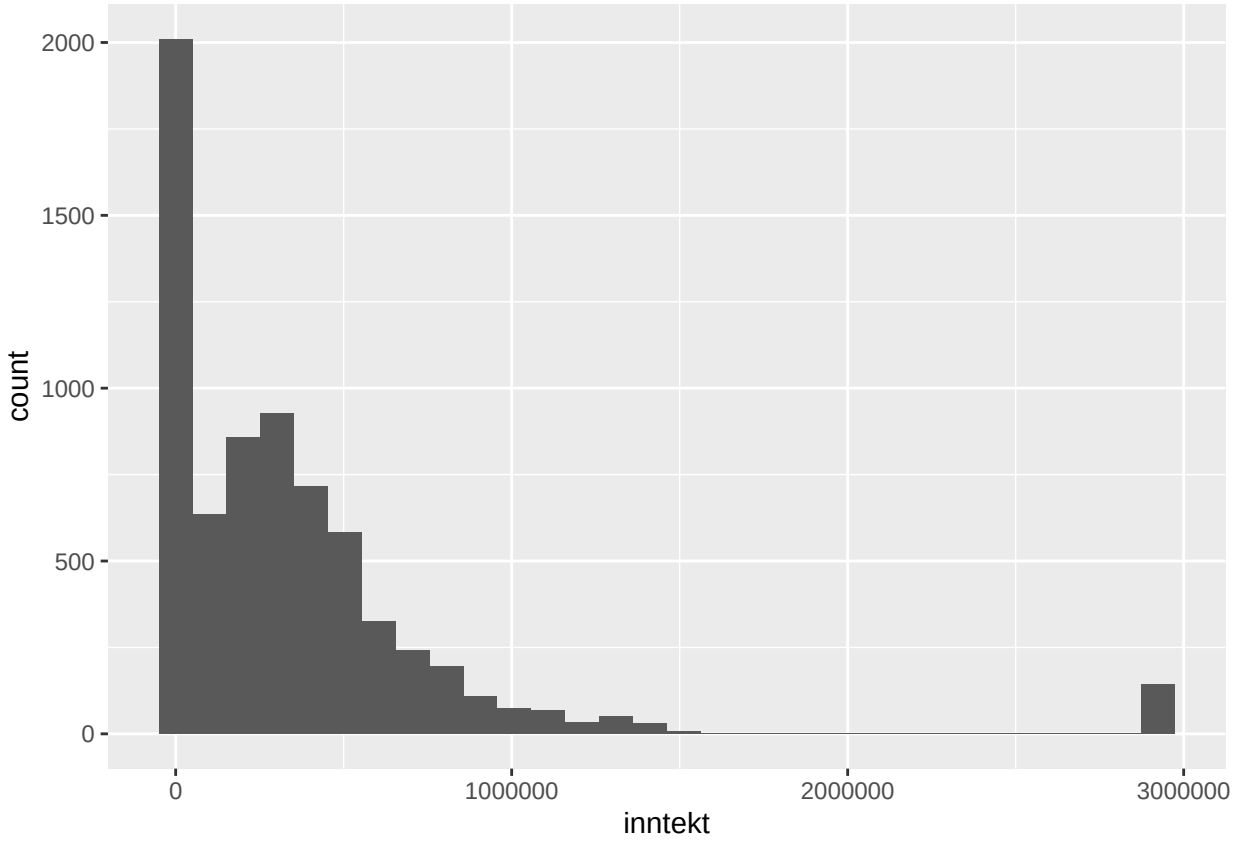
Datasettet vi bruker, *modelr* er hentet fra National Longitudinal Study, som er sponset av U.S. Bureau of Labor Statistics. Dataene stammer fra 2012. Følgende er forklaringene på variablene:

- *height* = høyde i tommer
- *weight* = vekt i pund
- *age* = alder mellom 47 og 56
- *marital* = sivilstatus
- *sex* = kjønn
- *education* = år med utdanning
- *afqt* = prosentskår på test for militær egnethet

Exploratory Data Analysis (EDA) vha. ggplot

```r
# Her har vi laget et histogram av variablene income (også kalt inntekt)
ggplot(data = hoyde,
       aes(x = inntekt)) +
  geom_histogram()
```

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



Her ser vi noen utliggere på høyresiden. Dette er 143 observasjoner av personer som tjener rett under 3MNOK. De skiller seg fra resten av observasjonene i histogrammet grunnet at både median- og snittlønn er langt lavere.

Vi har også personer *uten* inntekt i datasettet.

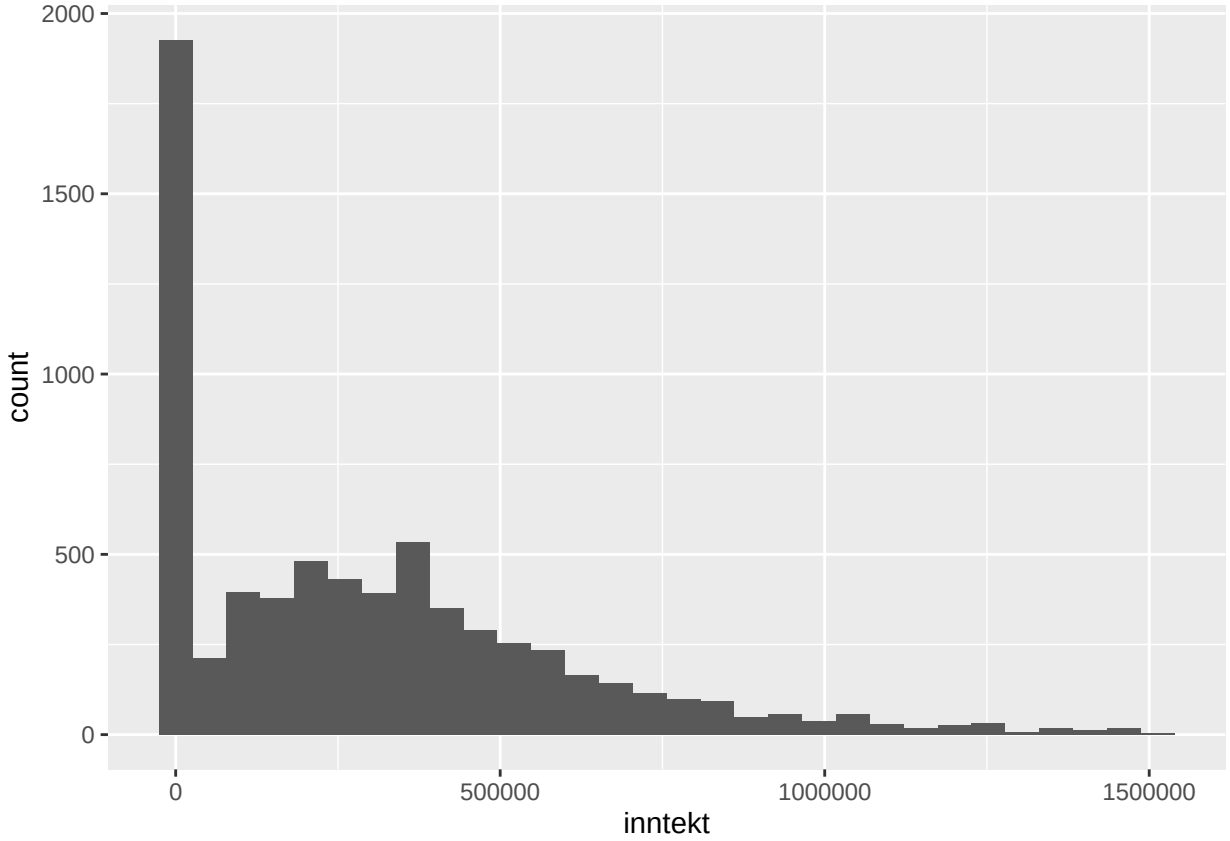
Regresjonsanalyse

```
##
## Call:
## lm(formula = inntekt ~ height_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 <0.0000000000000002 ***
## height_cm       9978.5      534.3   18.68 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Her ser vi at en økning i høyden på 1 cm, gir 9978.5 kr mer i årlig inntekt. La oss prøve med datasett uten de 2% med toppinntekt, og uten de med inntekt = 0.

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



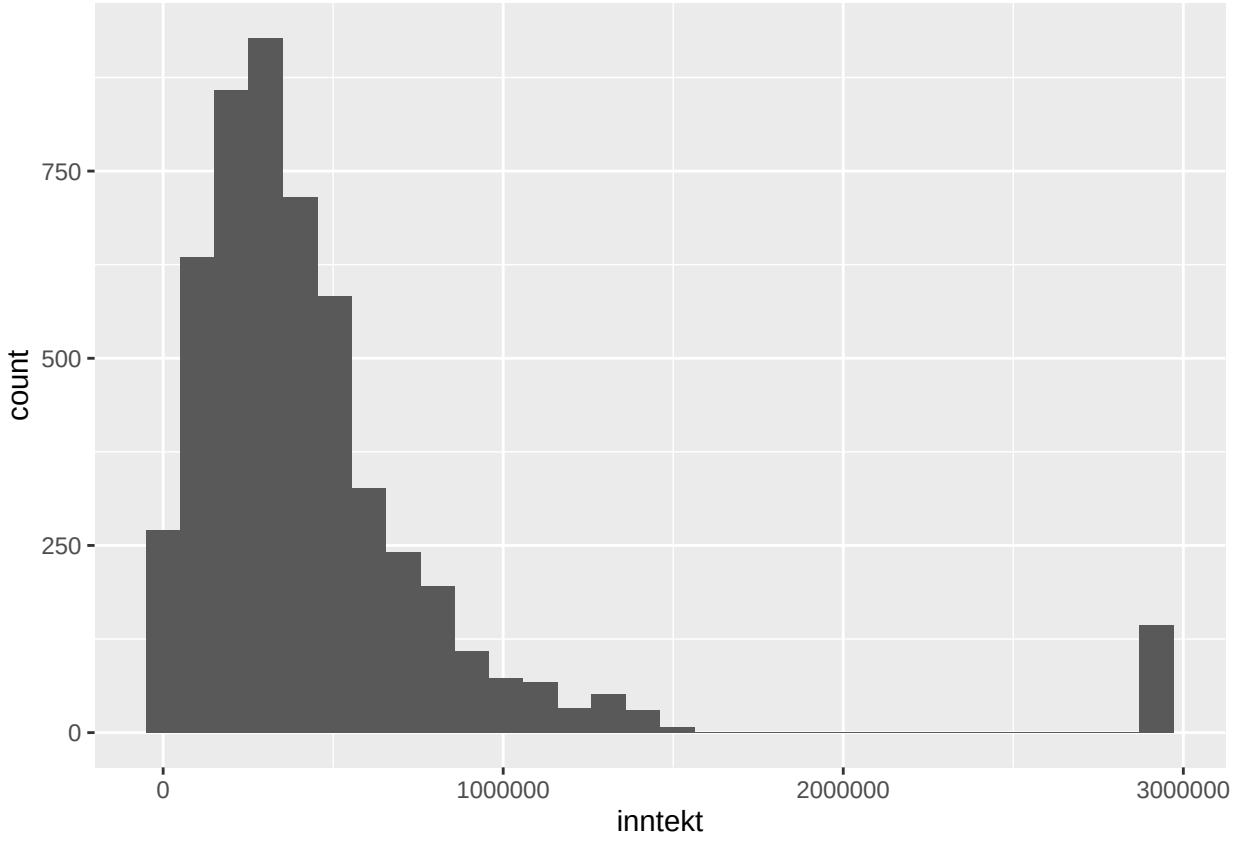
Her ser vi at utliggerne forsvinner, ettersom den vannrette aksen kun viser observasjoner hvor inntekt er lavere enn 1.600.000.

```
##
## Call:
## lm(formula = inntekt ~ height_cm, data = hoyde_max_inntekt)
##
## 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 <0.0000000000000002 ***
## height_cm      5828.4      342.5   17.02 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Her ser vi at en økning i høyden på 1 cm, gir 5828.4 kr mer i årlig inntekt.

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




```
##
## Call:
## lm(formula = inntekt ~ height_cm, data = hoyde_min_inntekt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -714128 -253106 -103101   95637 2634963
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -1435793.6   110687.8  -12.97 <0.0000000000000002 ***
```

```
## height_cm      11122.9      646.2   17.21 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 483000 on 5264 degrees of freedom
## Multiple R-squared:  0.05328,    Adjusted R-squared:  0.0531
## F-statistic: 296.3 on 1 and 5264 DF,  p-value: < 0.00000000000000022
```

Her ser vi at en økning i høyden på 1 cm, gir 11122.9 kr mer i årlig inntekt.