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)

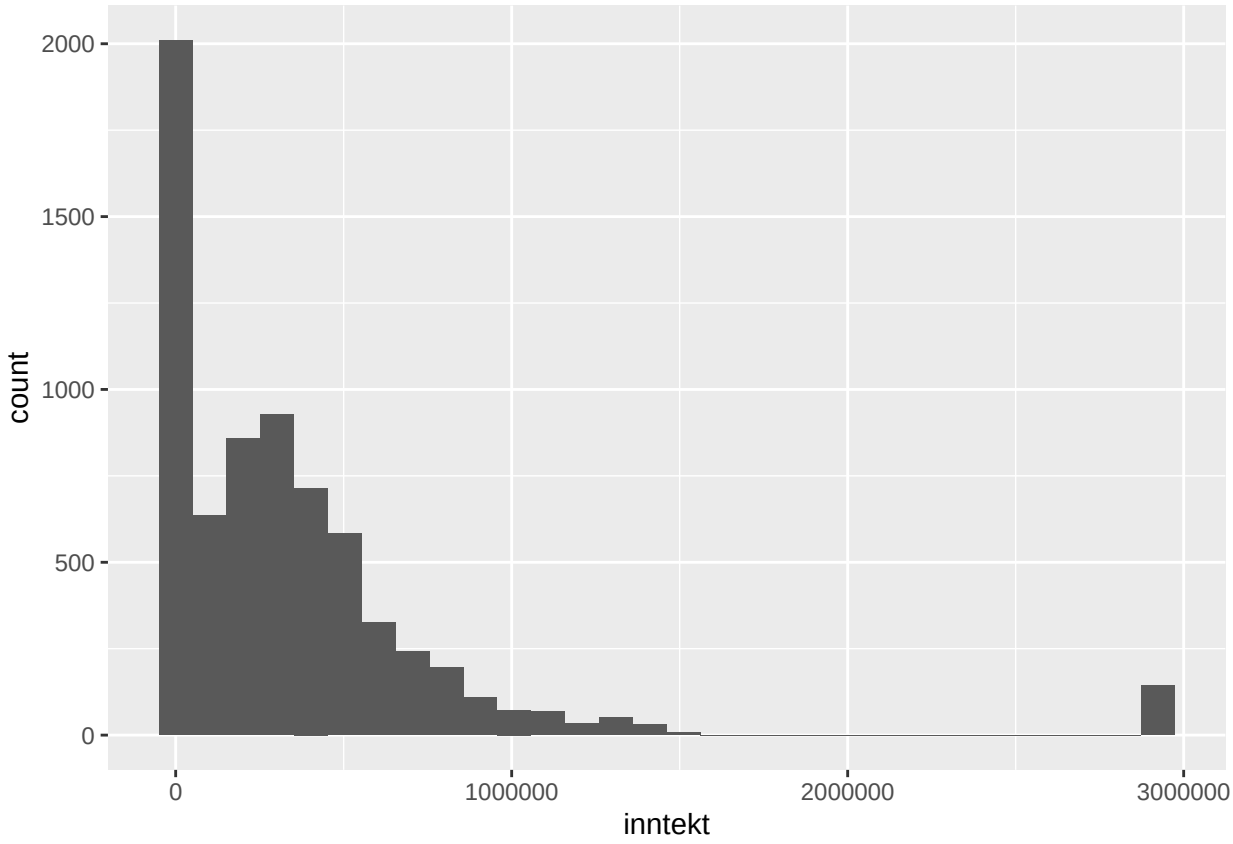
Datasettet vi bruker, *modelr* er hentet fra National Longitudinal Study, som er sponset av U.S. Bureau of Labor Statistics. Dataene stammer fra 2012. Følgende er forklaringene på variablene:

- *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.