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 benevnelse

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

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

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:

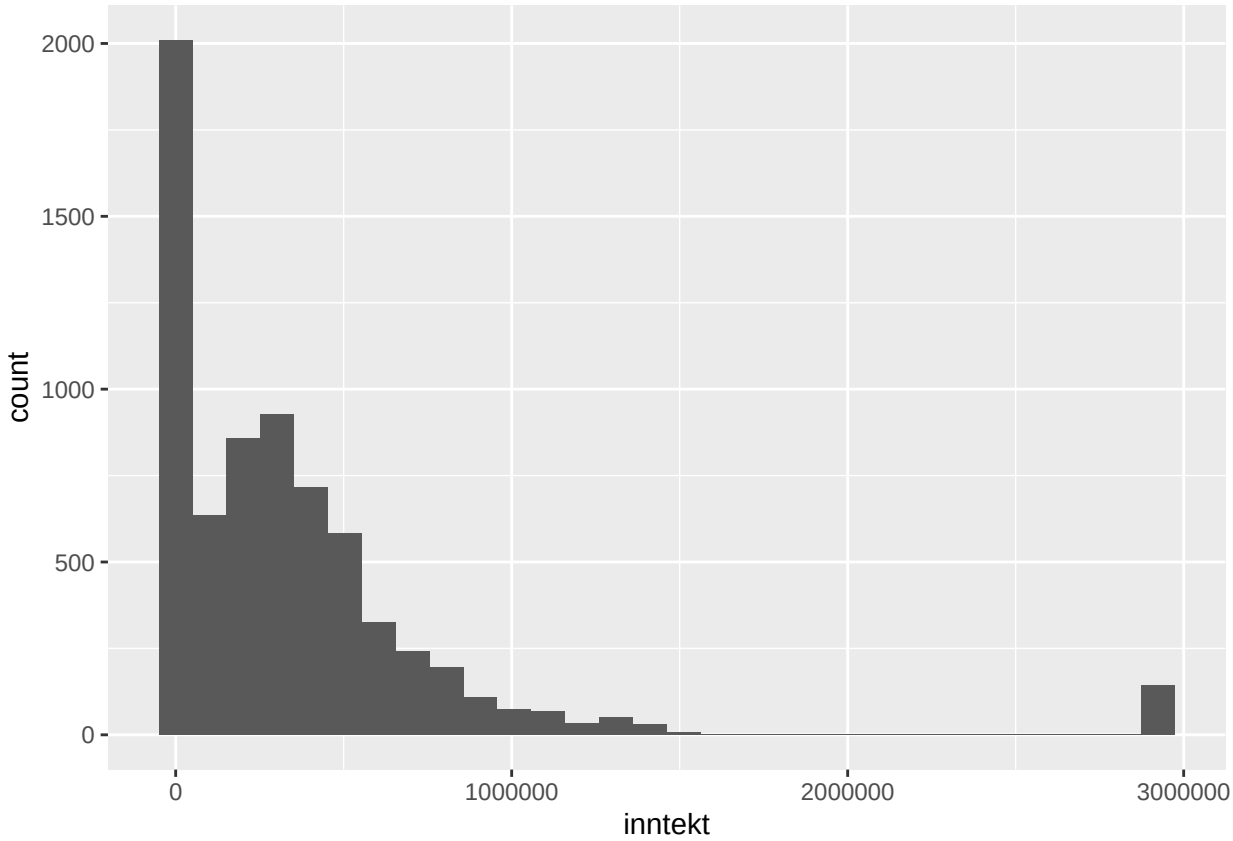
- *height* = høyde i tommer
- *weight* = vekt i pund

- *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

```
# 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.

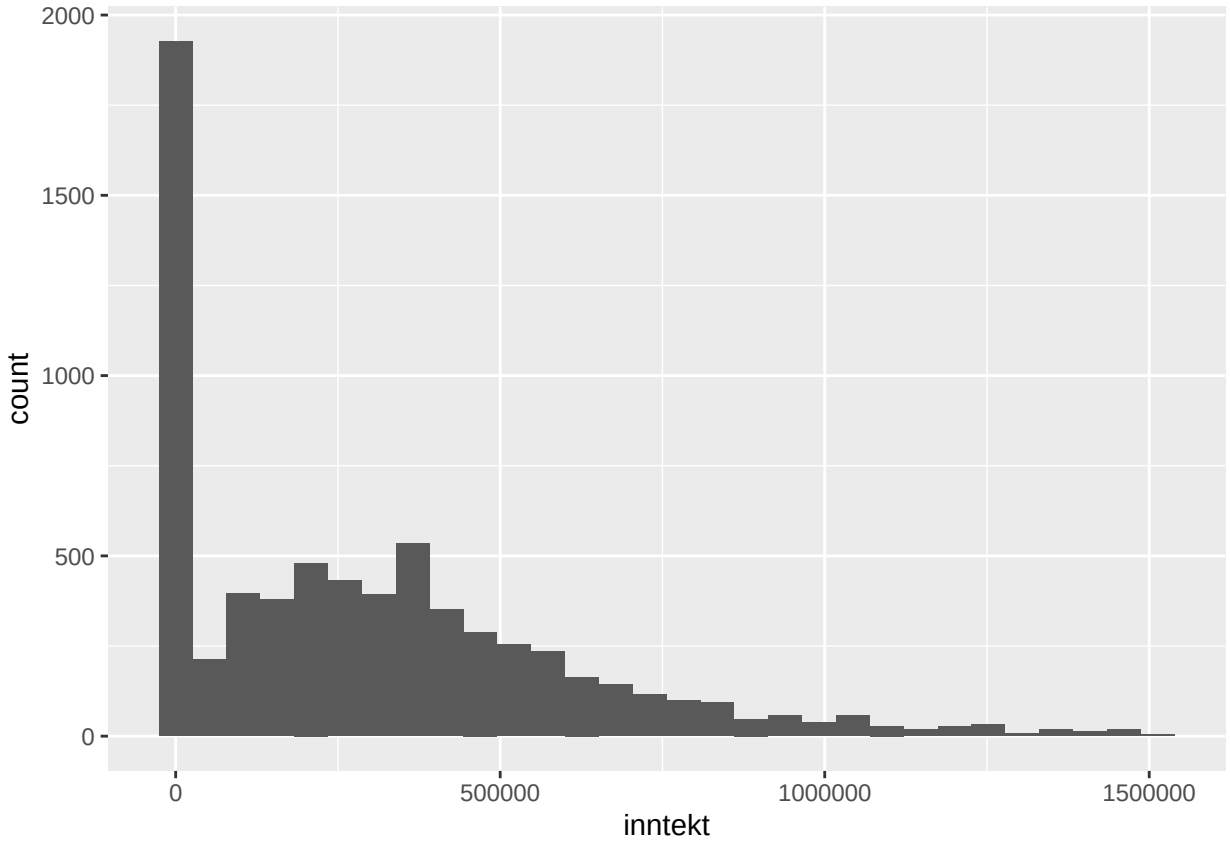
Vi har også personer *uten* inntekt i datasettet.

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`.
```

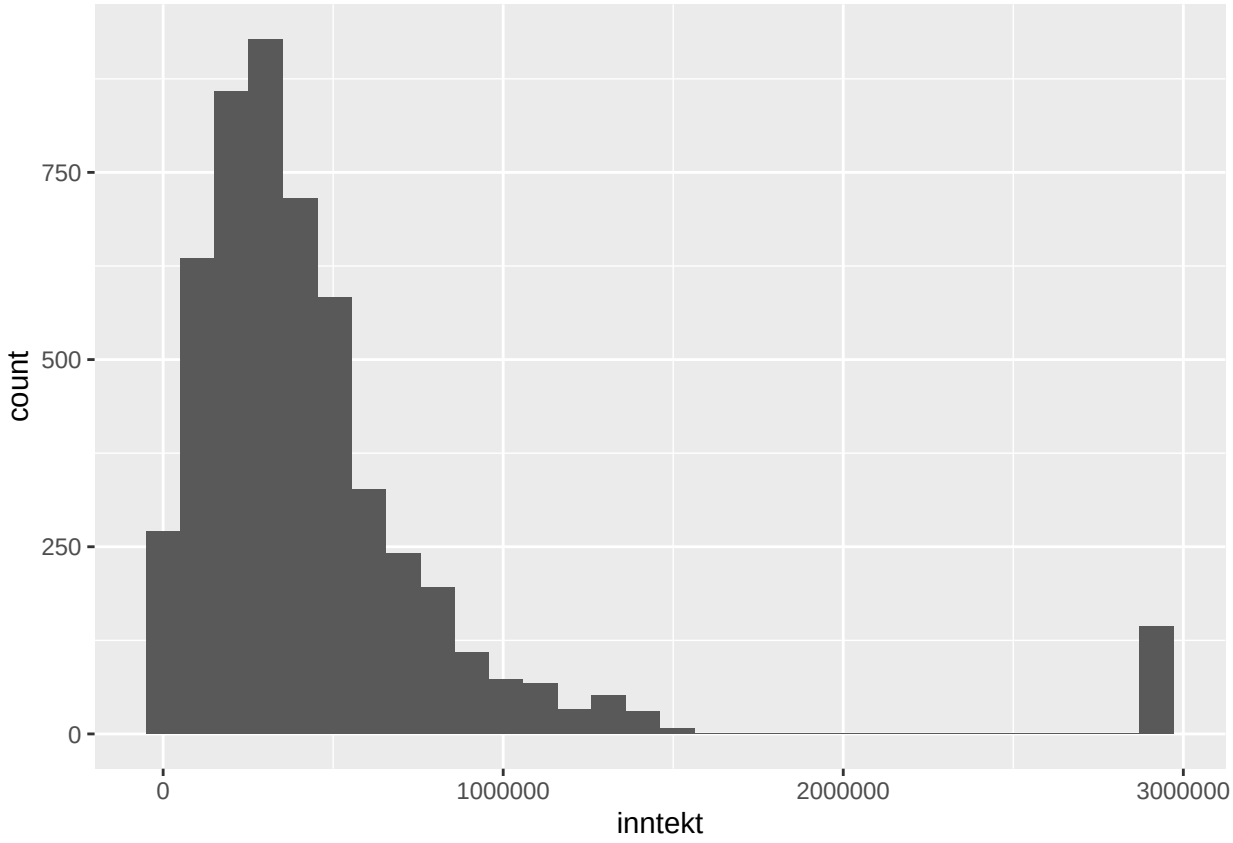


Her ser vi at utliggerne forsvinner, ettersom den vannrette aksen kun viser observasjoner hvor inntekt er lavere enn 1.600.000.

```
##
## 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.

Forklaring til Ytterliggere i *plots* (Prøver å unngå Merge Conflict)

Som vi ser ut fra grafen er det en stor ujevnhet. I datasettet er den største andelen av observasjonene fra ca 700 000 kroner og ned, med en mindre andel over dette. 143 observasjoner har rett i underkant av 3 millioner kroner. Dette er det den høyre ytterliggenheten i datasettet.

Vi har også med observasjoner *uten* lønn. Dette er den venstre ytterliggenheten. Det er 2000 observasjoner der vedkommende ikke har lønn.

Disse ytterliggenhetene påvirker resultatet av analysen. Disse to ekstreme observasjonene resulterer at sammenhengen mellom høyde og snitt- og medianlønn blir feilaktig fremstilt. Vi får dermed feil informasjon ut av dataene vi analyserer. Vi får tilfeller der en lav person er arbeidsledig eller at en høy person har langt høyere inntekt, slik som analysen fra National Longitudinal Study kom frem til.

For å finne et mer reelt resultat må vi se vekk ifra de ekstreme ytterliggenhetene. I dette tilfellet vil resultat blir reelt om vi ser vekk ifra både 0 inntekt og 3 millioner i inntekt. Dette vil vi gjøre **senere/videre** i oppgaven.

Litterature review

I Judge og Cable fra 2004 @judgeEffectPhysicalHeight2004, kommer de med utsagent at "høgde påvirker inntekten" er ved første øyenkast en gammel myte, men at det kanskje er mer til det enn mann først skulle trodd. For å støtte dette utsagnet referer de til Robert & Herman sin forskning som viser til at høyde er ett trekt som er ettertraktet i en sosial sammenheng.

Denne forskningen mener også at mennesker som er høyere er mer overtalenede. De viser også til @highamRiseFallPoliticians1992 som påstår at høye folk er mer sannsynlig til å komme i ein ledelses posisjon. Judge & Cable theoriserer at dette muligens har røter i biologi, ettersom i naturen så er høgde en måling får styrke.

En studie utført av Kurtz @burnsRetailSalespersonsInquiry1993 viste at 78% av ansettelser innen salg var mennesker over gjennomsnittlig høyde. Dette ble argumentert av rekruterene å være fordi mennesker over gjennomsnittlig høyde ville være mer utmerket ovenfor kundene, i forhold til små mennesker.

Jugdge og Cable @judgeEffectPhysicalHeight2004, ville svare på dette med å utføre en studie med tre hovedpunkter. Først fremstille en modell som viser **sammenhengen/forholdet** mellom høyde og karriere suksess. De begrunnet dette med at dette ikke var blitt utført tidlegere. Steg to var å utføre en meta analyse på tidlegere analyser og litteratur for å se etter generelle implikasjoner. Siste steget var å utføre fire nye undersøkelser på **sammenhengen/forholdet** mellom en persons høyde og dens inntekt.

Argumentet for å se på **sammenhengen/forholdet** mellom en persons høyde og inntekt var fordi de anså inntekt som den primære faktoren for karriere suksess. Men i følge @EMPIRICALINVESTIGATIONPREDICTORS og @whitelyRelationshipCareerMentoring1991, så er det nesten ingen forskning på dette.

Jugde og Cable @judgeEffectPhysicalHeight2004, tok i sin undersøkelse utgangspunkt i flere menneskelige faktorer for å produsere sin modell for **sammenhengen/forholdet** mellom en persons høyde og inntekt.

Noen av punktene de så på var selvtillit og sosial aktelse. Judge & Cable @judgeEffectPhysicalHeight2004 mente dette var to viktige punkter i analysen fordi disse to faktorene påvirker en persons arbeidsinnsats og hvordan en person blir behandlet i arbeidslivet av arbeidsgiveren. Dette mente de ville påvirke en persons suksess i arbeidslivet, altså medføre at en høy person ville ha bedre inntekt enn en lav person.

Studien til Judge & Cable judgeEffectPhysicalHeight2004, kom til slutt frem til at det faktisk var en direkte sammenheng mellom en persons høyde og inntekt, men i ettertid har dette resultatet blitt sett på og folk mener at sunn fornuft tilsier at dette ikke kan stemme.