Inntekt og Høyde

Tjener høye mennesker mer?

Arian Steen Anne Grete Lilleland

12 10 2020

Har høyde noe å si for inntekt, lønn eller bonuser?

Det er ofte snakket om at ulike attributter har mye å si for en persons karriere,liv og generell livskvalitet. Et av de tingene vi ofte hører er at høye mennesker vil i snitt tjene mer. Vi skal i denne analysen se nærmere på utsagnet "Din høyde har betydning for hvor mye du kan tjene". Vi har installert pakken 'modelr' og skal gå gjennom datasettet 'heights', dette er data som er hentet fra 'Nation longitudinal study' i USA, som er sponset av det amerikanske Bureau of Labour Statistics, som kan sammenlignes med Norges Statistisk Sentralbyrå.

Ulike variabler som er inkludert i analysen

Datasettet vi har brukt inkluderer variablene:

- Høyde
- Inntekt
- Kjønn
- Vekt
- Sivilstatus
- Alder
- Utdanning

Vi gjør oppmerksom på at dette datasettet ikke inkluderer etnisitet, familiens formue eller hvor de utvalgte bor.Datasettet er tatt fra det amerikanske arbeidsmarkedet, det må tas hensyn til dette dersom vi skal bruke dette datasettet som et utgangspunk for andre land og verdensdeler. Analysen vil dermed se på hvordan høyde korrelerer med inntekt når det gjelder det amerikanske arbeidsmarkedet.

Ulike analyse metoder og verktøy brukt

Vi har brukt ulike typer analyser for å tolke datasettet. Dette er gjort for å få kunne tolke datasettet grundigere, de ulike typene analysene vi har brukt inkluderer:

- Pearson korrelasjonskoeffisient
- Spearman's rank korrelasjonskoeffisient
- Histogram distribusjon og Density funksjon
- Korrelasjon matrise/Matrixplot
- Korrelasjons test, to variabler
- Korrelasjon og Regresjonsanalyse
- Økonometrisk modell

Oppsumering av ulike variabler i datasettet

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

Pearson korrelasjonskoeffisient

Pearson korrelasjonskoeffisient måler, samvariasjonen mellom to variabler ved å dele variablenes kovarians på produktet av variablens respektive standardavvik. Rho har verdi mellom +1 og -1. En verdi på +1 er fullstendig positiv lineær korrelasjon, 0 er ingen lineær korrelasjon, og -1 er fullstendig negativ korrelasjon.

```
##
##  Pearson's product-moment correlation
##
## data:  heights$income and heights$height
## t = 18.676, df = 7004, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1953779 0.2399910
## sample estimates:
##       cor
## 0.2177982
```

Siden Rho har en verdi på 0.217 er det ikke en klar positiv lineær korrelasjon mellom inntekt og høyde

Spearman's rank korrelasjonskoeffisient

```
## The following objects are masked from heights (pos = 3):
##
##     afqt, age, education, height, heightInt, income, marital, sex,
##     weight

## Warning in cor.test.default(heights$income, heights$height, method =
## "spearman"): Cannot compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  heights$income and heights$height
## S = 4.5234e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2107673
```

Her har vi også en veldig svak korrelasjon mellom inntekt og høyde, verdien ligger på 0.210.

Pearson korrelasjon

Mellom Inntekt og Utdanning

```
## The following objects are masked from heights (pos = 3):
##
##     afqt, age, education, height, heightInt, income, marital, sex,
##     weight

## The following objects are masked from heights (pos = 4):
##
##     afqt, age, education, height, heightInt, income, marital, sex,
##     weight

##
##  Pearson's product-moment correlation
##
## data:  heights$income and heights$education
## t = 35.779, df = 6994, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3733432 0.4129626
## sample estimates:
##       cor
## 0.3933354
```

Spearman Korrelasjon

Mellom Inntekt og Utdanning

```
## The following objects are masked from heights (pos = 3):
##
##     afqt, age, education, height, heightInt, income, marital, sex,
##     weight

## The following objects are masked from heights (pos = 4):
##
##     afqt, age, education, height, heightInt, income, marital, sex,
##     weight

## The following objects are masked from heights (pos = 5):
##
##     afqt, age, education, height, heightInt, income, marital, sex,
##     weight

## Warning in cor.test.default(heights$income, heights$education, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  heights$income and heights$education
## S = 3.3888e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4061854
```

Konklusjon Pearson og Spearman

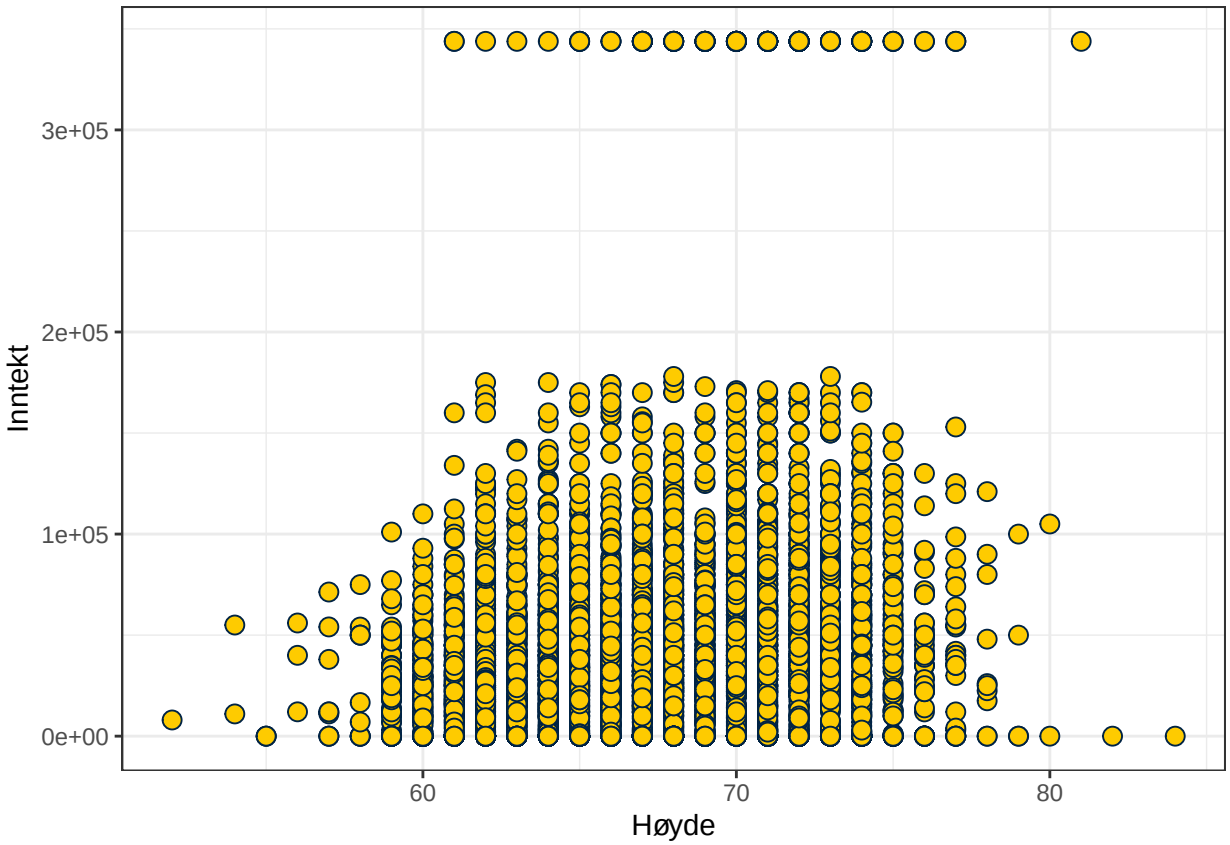
Korrelasjonen mellom høyde og inntekt er ikke klar ifølge Pearson testen fikk vi en verdi på 0.217, ifølge Spearman testen fikk vi en rho på 0.210.

Derimot finnes det en korrelasjon mellom utdanning og inntekt, ifølge Pearson testen fikk vi en verdi på *0.393* Ifølge Spearman testen fikk en *rho på 0.406* dette er en klar korrelasjon

Vi legger til grunn for at Spearman og Pearson koeffisient metoden ikke er den mest nøyaktige måten man kan analysere sammenhengen mellom inntekt og høyde, men det er fortsatt interresant på bakgrunn av klare tall.

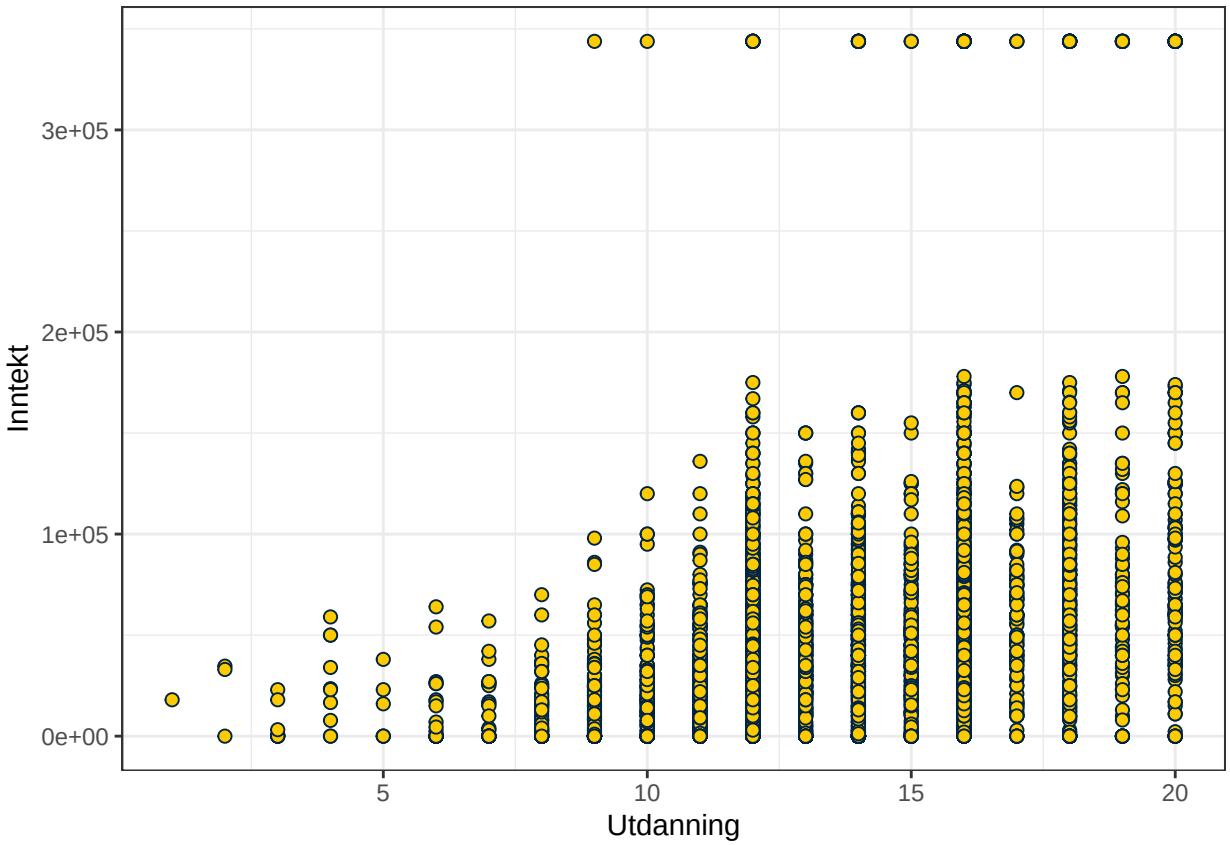
Scatterplot

Vi begynner med et scatterplot. Et spredningsplot vil vise verdien av to variabler i et datasett, vi vil visualisere hvor mye den ene variablen (høyde) vil påvirke den andre variablen (inntekt), Dette er spearman korrelasjonen visualisert grafisk. Man kan se at det ikke er noen sterk sammenheng mellom høyde og inntekt, utenom noen få outliers.



Scatterplot mellom Inntekt og Utdanning

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



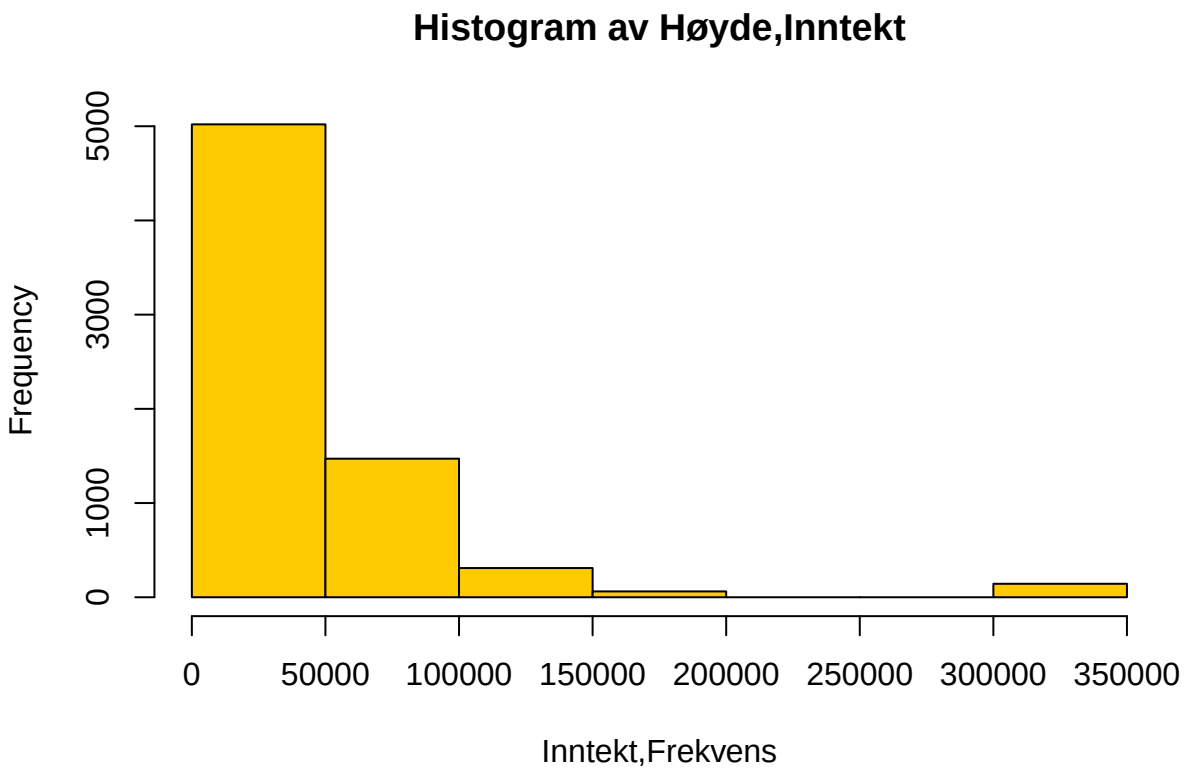
Vi kan se at utdanning er positivt korrelert med inntekt, det vil si at du tjener mer dersom du har høyere utdanning. Dette kan sammenlignes med høyde hvor korrelasjonen ikke er like positiv. Legg også merke til outliers på toppen av grafen.

Regresjonslinje og scatterplot

Histogram og Density plot

Et histogram er en grafisk framstilling, som brukes til å analysere og presentere data, Høyden av en søyle er frekvensen delt på klassebredden. Vi skal bruke et histogram til å analysere våre ulike variabler. Samt trekke en konklusjon basert på funnet.

Densityplot viser oss distribusjonen av numeriske variabler, vi bruker den i denne sammenheng for å se hvordan distribusjonen er i vårt datasett med tanke på Inntekt og Høyde.



Konklusjonen er at de aller fleste observasjoner, ligger trukket sammen uten en klar sammenheng. Vi ser også at det finnes noen få "outliers" i vårt datasett. Vi kan ikke trekke en klar konklusjon om at Høyde påvirker inntekt i noen særlig grad.

Matrise/Matrixplot og korrelasjon mellom Inntekt og Høyde

```
## [1] 0.2177982
```

En enkel korrelasjonstest av datasettet viser oss at korrelasjonen er 0.2177, med andre ord ikke så høy korrelasjon. Vi lager et korrelasjonsplot illustrert grafisk og inkludert samtlige av de syv variablene ved hjelp av pakkene corrplot og ggplot vist nedenfor.

2 grafs-plot

Line chart, linje chart