Digital Causality Lab

Collider Bias

Universität Hamburg
Fakultät für Betriebswirtschaft
Lehrstuhl für Mathematik und Statistik in den Wirtschaftswissenschaften

Gliederung

1. Einleitung

- Ausgangssituation und Problemstellung
- Motivation

2. Statistischer Hintergrund

- Collider - DAG

3. Implementierung in R

- Quellcode

4. Fazit

1. Einleitung
Ausgangssituation und Problemstellung

Beispiel Friseursalon

Datensatz

- 101 Beobachtungen (Friseursalons)

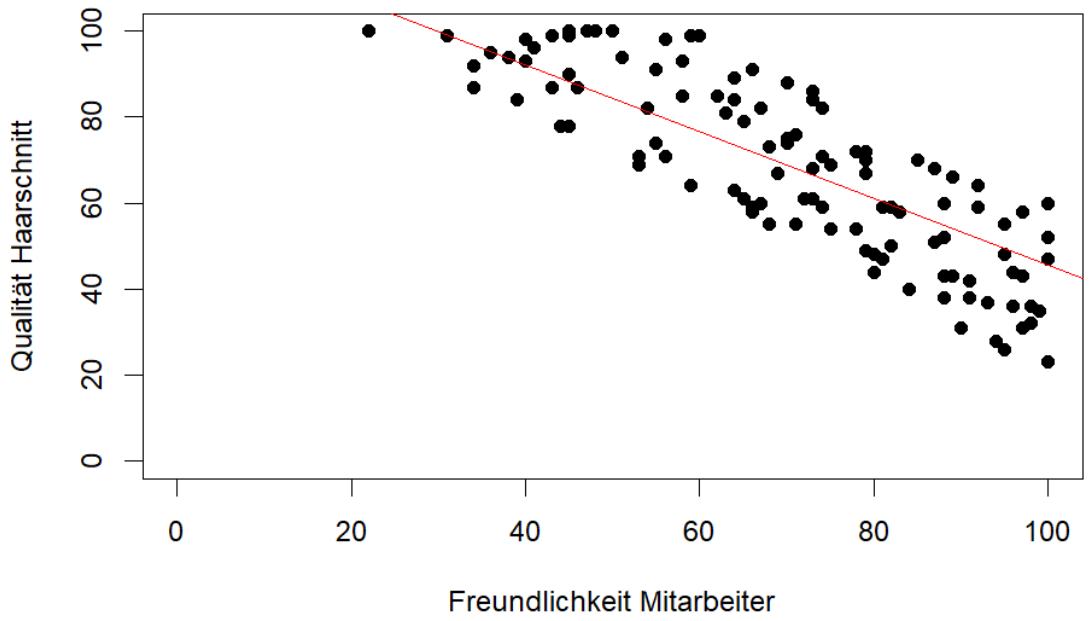
- 2 Variablen:
  - Freundlichkeit Mitarbeiter (F)
  - Qualität Haarschnitt (Q)

  F und Q jeweils in Punkten gemessen (0 bis 100 Punkte)

Problemstellung

Unsere Studie resultiert in folgendem Diagramm:

→ „Um einen qualitativ hochwertigen Haarschnitt zu erhalten, muss man einen unfreundlichen Mitarbeiter aufsuchen“



→ Was ist hier schiefgelaufen?

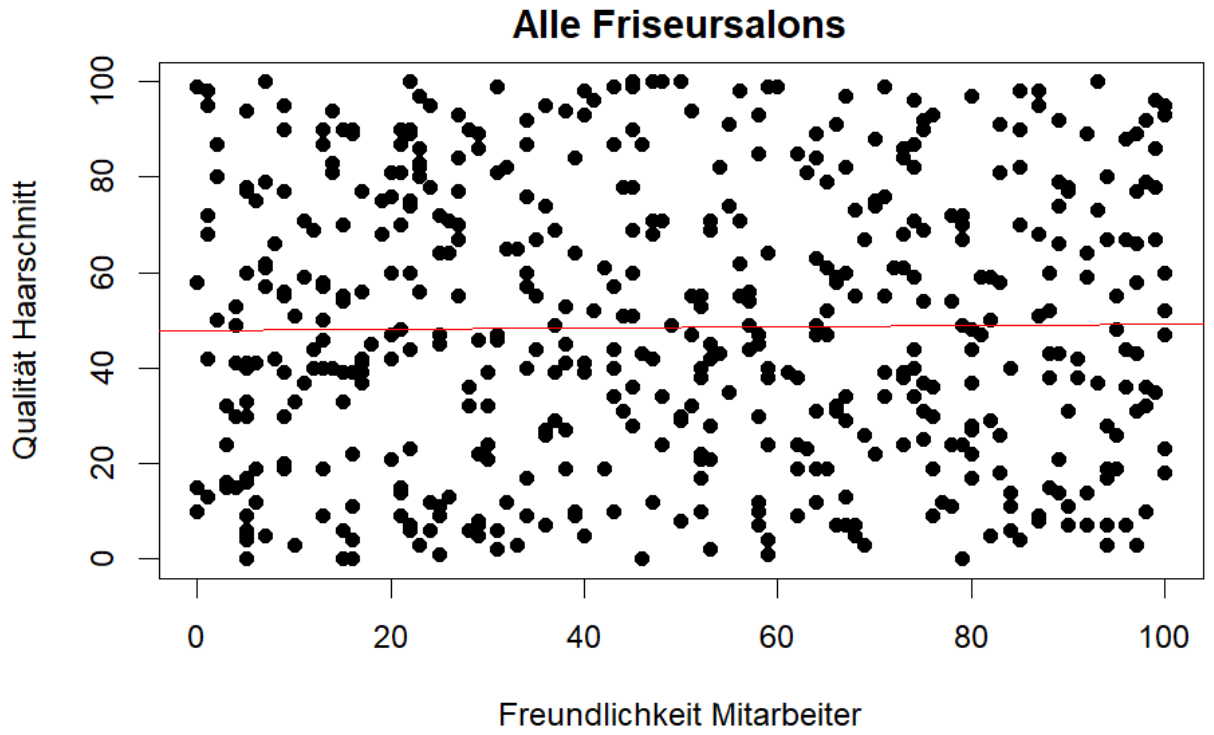
Erweiterung

Problem: Es wurden versehentlich nur 4-Sterne-Friseursalons betrachtet!

Erweiterung des Datensatzes

- Aufnahme von 1, 2, 3 und 5-Sterne-Friseursalons in die Studie
- Neuer Datensatz mit 500 Beobachtungen
  - Freundlichkeit Mitarbeiter (F)
  - Qualität Haarschnitt (Q)
  - Neu: Sternebewertung (C)

C gemessen auf Skala von 1 bis 5

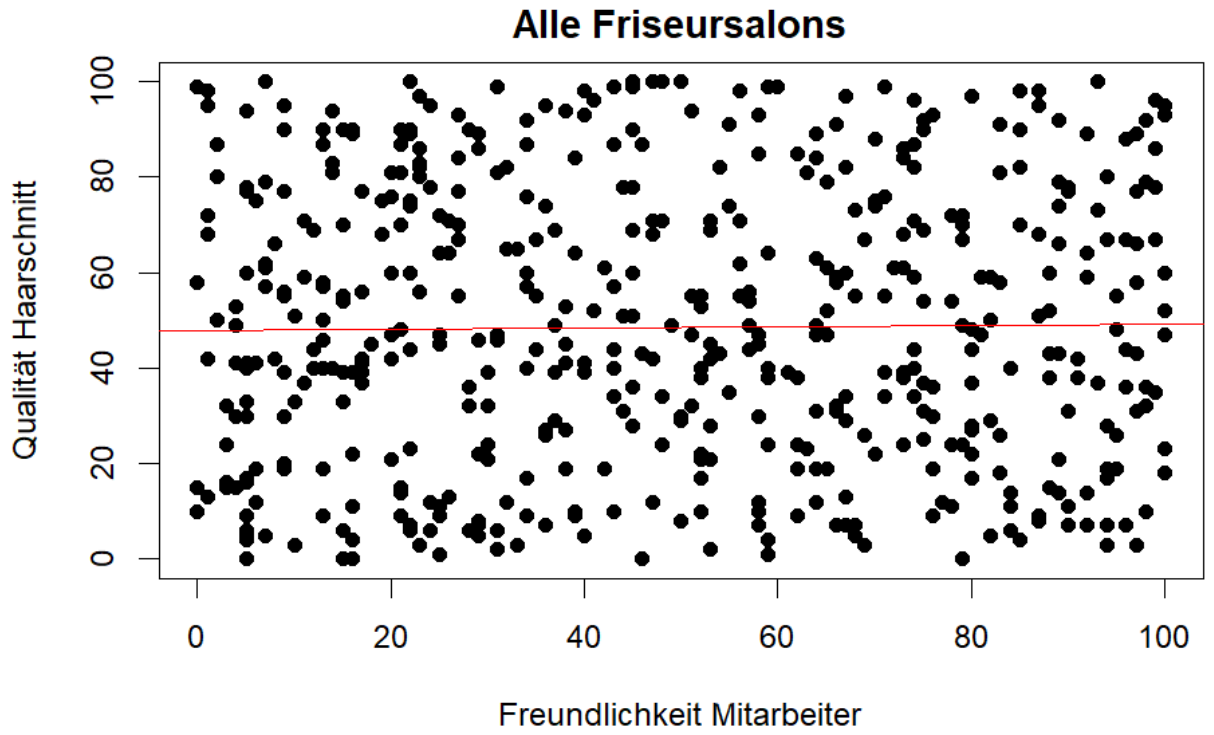


Erweiterung

Auf Populationsebene besteht keine Korrelation!

→ Qualität Haarschnitt und Freundlichkeit der Mitarbeiter sind unabhängig

$$F \perp\!\!\!\perp Q$$



1. Einleitung
Motivation

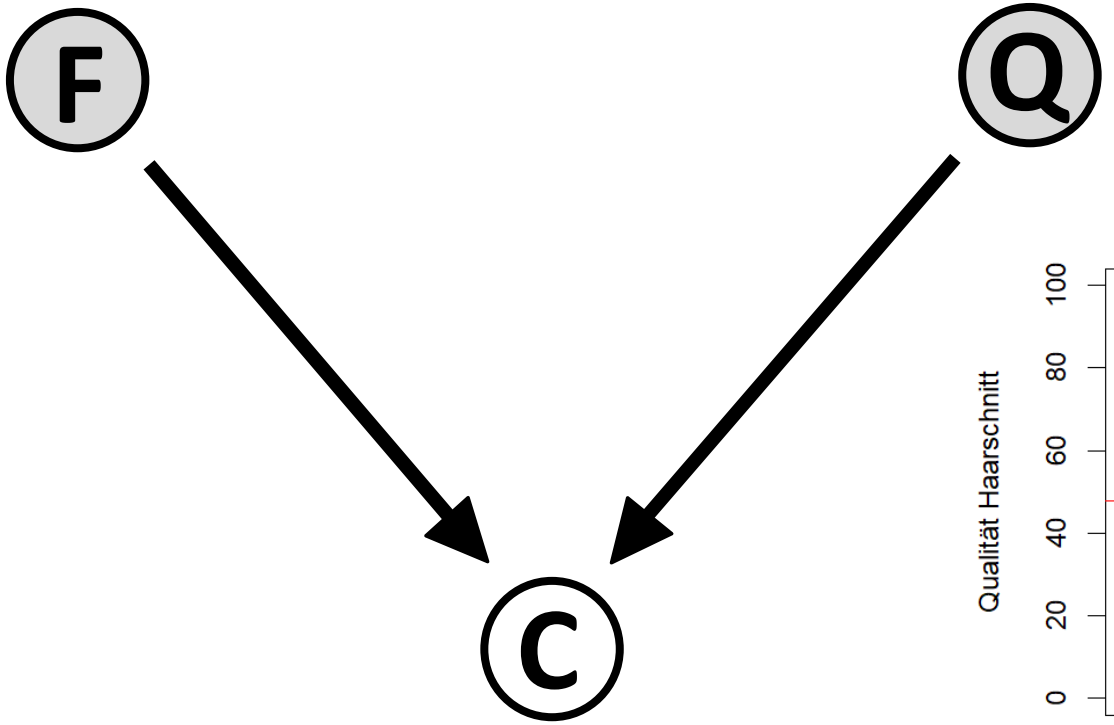
Motivation

- Datenerhebung ist kritischer Schritt, bereits hier können Verzerrungen entstehen
- verzerrter Datensatz → falsche kausale Schlussfolgerungen
- Plausibilität der gewonnenen Resultate muss geprüft werden
- Ursache für Friseursalon-Beispiel: **Collider Bias**

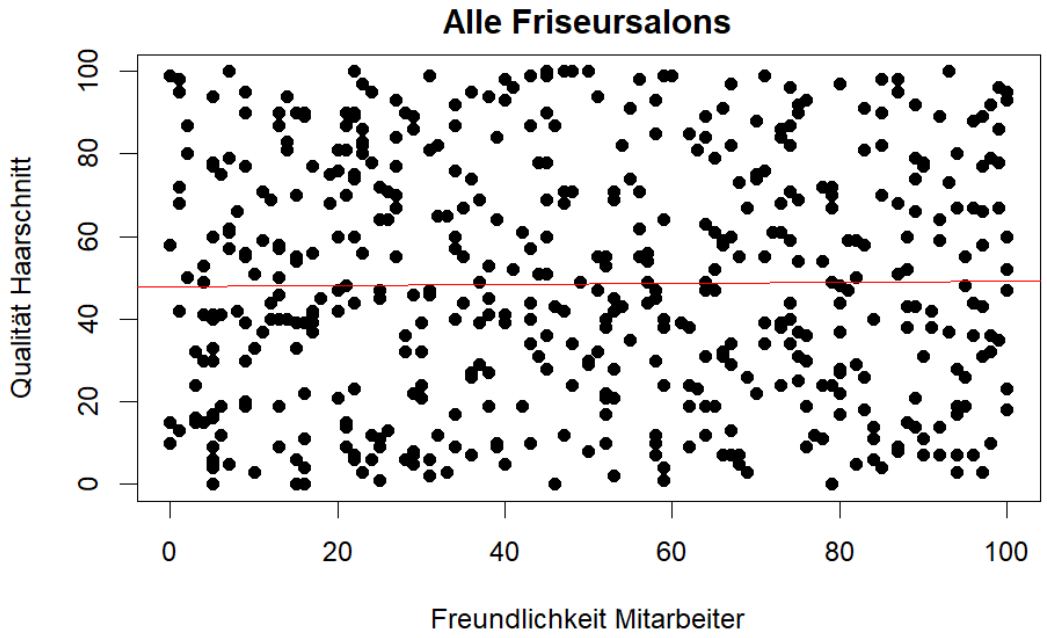
Beseitigung des Problems: Ordnungsgemäße Behandlung des Colliders „Sternebewertung“

2. Statistischer Hintergrund
Collider - DAG

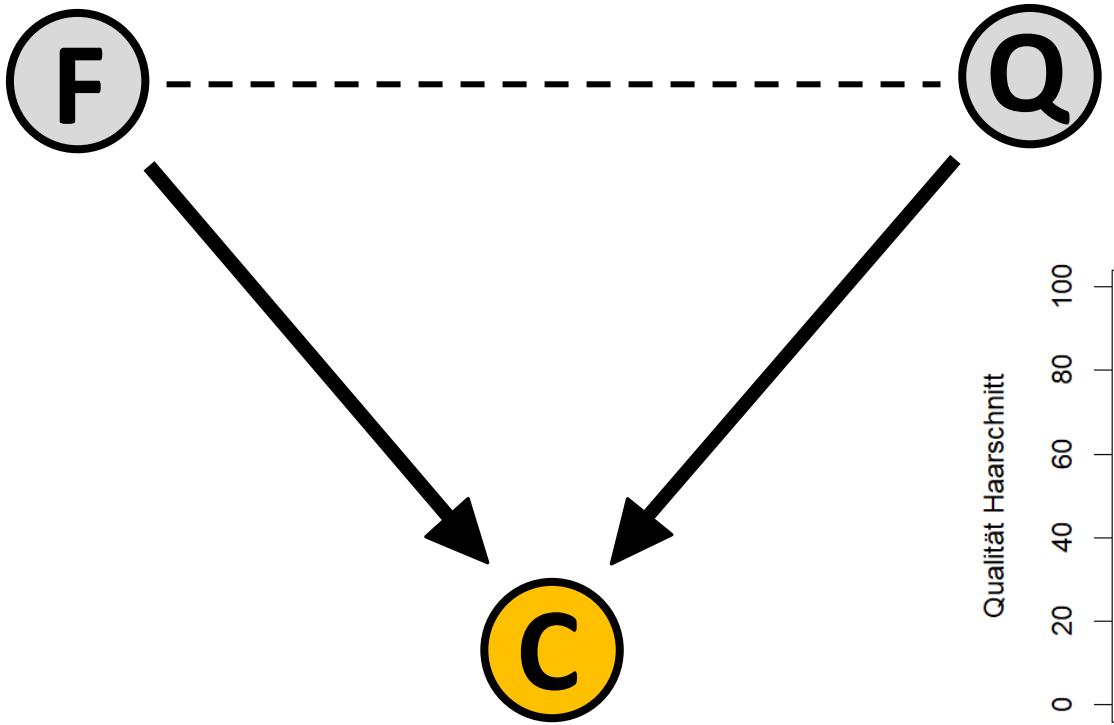
Collider - DAG



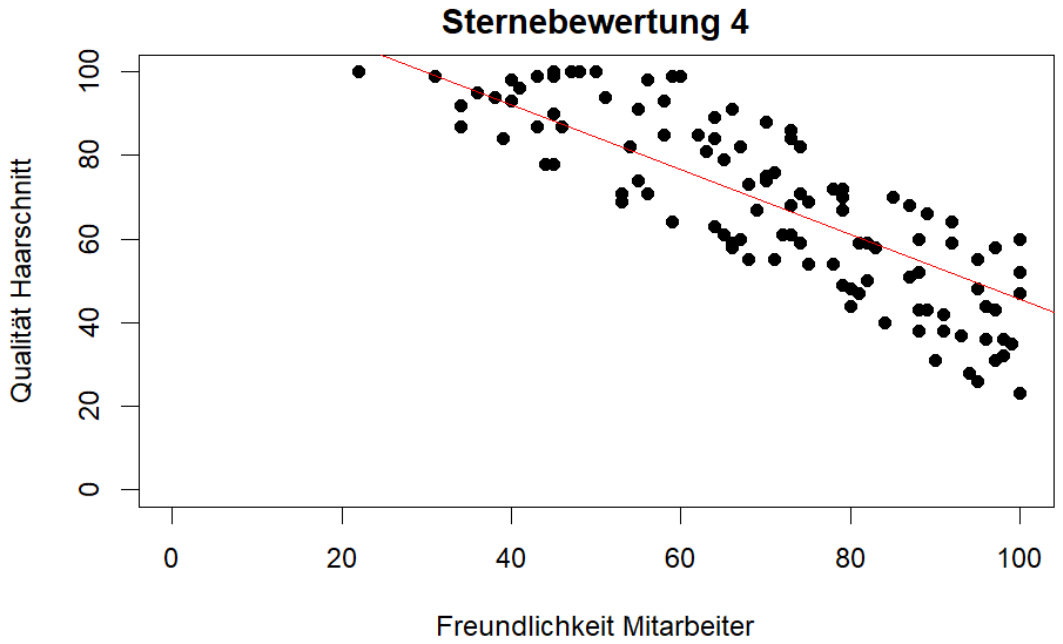
F: Freundlichkeit Mitarbeiter
Q: Qualität Haarschnitt
C: 4-Sterne-Bewertung



Collider - DAG



F: Freundlichkeit Mitarbeiter
Q: Qualität Haarschnitt
C: 4-Sterne-Bewertung



3. Implementierung in R
Quellcode

Quellcode

```
FreundlichkeitMitarbeiter = sample(0:100,size=500,replace = T)
QualitätHaarschnitt       = sample(0:100,size=500,replace = T)

EinStern   = ifelse((FreundlichkeitMitarbeiter+QualitätHaarschnitt) <= 40, EinStern <- 1,
EinStern <-  0)
ZweiSterne = ifelse((FreundlichkeitMitarbeiter+QualitätHaarschnitt) > 40 &
(FreundlichkeitMitarbeiter+QualitätHaarschnitt) <= 80 , ZweiSterne <- 2, ZweiSterne <- 0)
DreiSterne = ifelse((FreundlichkeitMitarbeiter+QualitätHaarschnitt) > 80 &
(FreundlichkeitMitarbeiter+QualitätHaarschnitt) <= 120 , DreiSterne <- 3, DreiSterne <- 0)
VierSterne = ifelse((FreundlichkeitMitarbeiter+QualitätHaarschnitt) > 120 &
(FreundlichkeitMitarbeiter+QualitätHaarschnitt) <= 160 , VierSterne <- 4, VierSterne <- 0)
FünfSterne = ifelse((FreundlichkeitMitarbeiter+QualitätHaarschnitt) > 160 &
(FreundlichkeitMitarbeiter+QualitätHaarschnitt) <= 200 , FünfSterne <- 5, FünfSterne <- 0)


Sternebewertung = abs(EinStern-ZweiSterne-DreiSterne-VierSterne-FünfSterne)
```

Quellcode

```
Datensatz =
data.frame(FreundlichkeitMitarbeiter,
QualitätHaarschnitt,Sternebewertung)
```

	FreundlichkeitMitarbeiter	QualitätHaarschnitt	Sternebewertung
1	97	51	4
2	53	1	2
3	94	60	4
4	75	11	3
5	48	26	2
6	89	55	4
7	73	33	3
8	52	58	3
9	85	49	4
10	75	71	4
11	47	39	3
12	36	67	3
13	19	37	2
14	3	77	2
15	66	48	3
16	2	24	1
17	80	84	5

Quellcode

```
Teilmenge1 = subset(Datensatz,Sternebewertung == 1)
Teilmenge2 = subset(Datensatz,Sternebewertung == 2)
Teilmenge3 = subset(Datensatz,Sternebewertung == 3)
Teilmenge4 = subset(Datensatz,Sternebewertung == 4)
Teilmenge5 = subset(Datensatz,Sternebewertung == 5)
```

	FreundlichkeitMitarbeiter	QualitätHaarschnitt	Sternebewertung
1	97	51	4
3	94	60	4
6	89	55	4
9	85	49	4
10	75	71	4
20	94	33	4
23	88	36	4
26	100	58	4

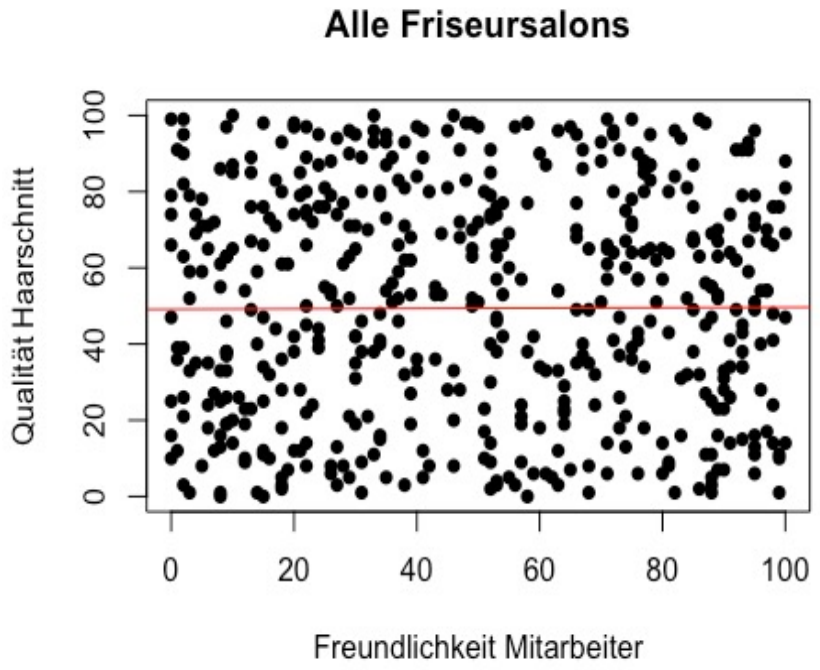
Quellcode

```
RegressionDatensatz  = lm(FreundlichkeitMitarbeiter~QualitätHaarschnitt, Datensatz)
RegressionTeilmenge1 = lm(FreundlichkeitMitarbeiter~QualitätHaarschnitt,Teilmenge1)
RegressionTeilmenge2 = lm(FreundlichkeitMitarbeiter~QualitätHaarschnitt,Teilmenge2)
RegressionTeilmenge3 = lm(FreundlichkeitMitarbeiter~QualitätHaarschnitt,Teilmenge3)
RegressionTeilmenge4 = lm(FreundlichkeitMitarbeiter~QualitätHaarschnitt,Teilmenge4)
RegressionTeilmenge5 = lm(FreundlichkeitMitarbeiter~QualitätHaarschnitt,Teilmenge5)
```

Quellcode

```
par(pch=16)

plot(Datensatz$FreundlichkeitMitarbeiter,Datensatz$QualitätHaarschnitt,xlab="Freund
lichkeit Mitarbeiter",ylab="Qualität Haarschnitt",main = "Alle Friseursalons")
abline(RegressionDatensatz, col="red")
```

→ Ergebnis:



Quellcode

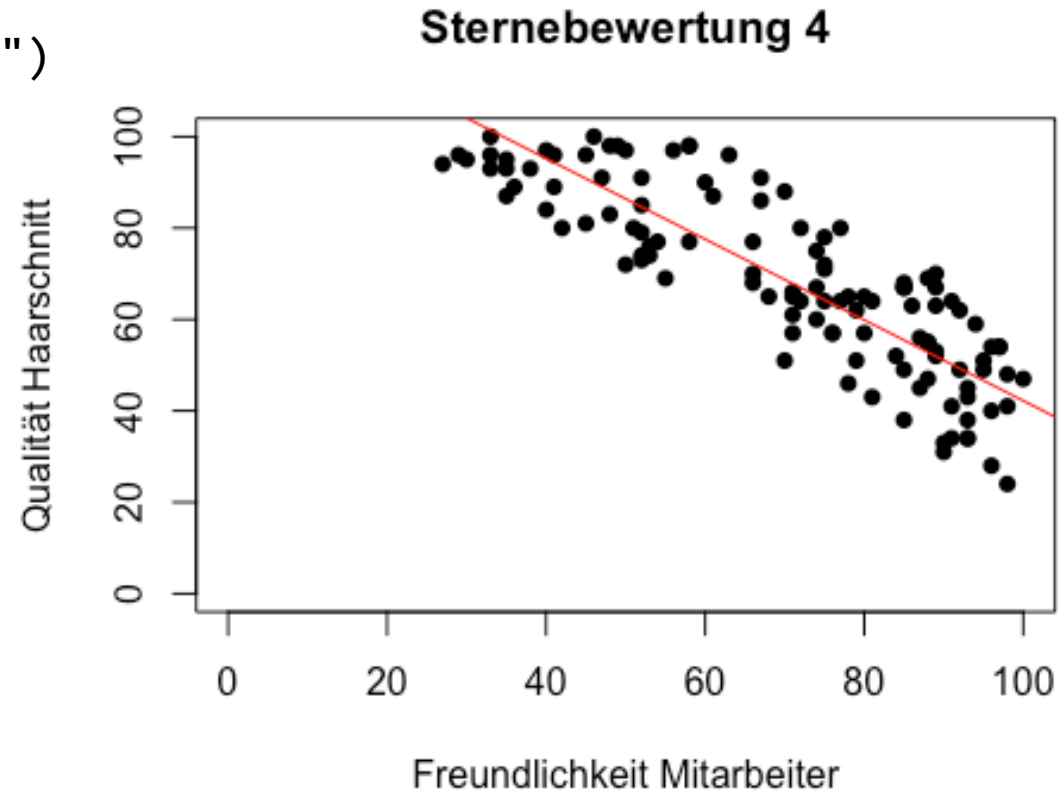
```
plot(Teilmenge1$FreundlichkeitMitarbeiter,Teilmenge1$QualitätHaarschnitt,xlab="Freu
ndlichkeit Mitarbeiter",ylab="Qualität Haarschnitt",xlim = c(0,100),ylim = c(0,100)
,main="Sternebewertung 1")

abline(RegressionTeilmenge1, col="red")

plot(Teilmenge2$FreundlichkeitMitarbeiter,Teilmenge2$QualitätHaarschnitt,xlab="Freu
ndlichkeit Mitarbeiter",ylab="Qualität Haarschnitt",xlim = c(0,100),ylim = c(0,100)
,main="Sternebewertung 2")

abline(RegressionTeilmenge2, col="red")

plot(Teilmenge3$FreundlichkeitMitarbeiter,Teilmenge3$QualitätHaarschnitt,xlab="Freu
ndlichkeit Mitarbeiter",ylab="Qualität Haarschnitt",xlim = c(0,100),ylim = c(0,100)
,main="Sternebewertung 3")

abline(RegressionTeilmenge3, col="red")
```

Quellcode

```
plot(Teilmenge4$FreundlichkeitMitarbeiter,Teilmenge4$QualitätHaarschnitt,xlab="Freu
ndlichkeit Mitarbeiter",ylab="Qualität Haarschnitt",xlim = c(0,100),ylim = c(0,100)
,main="Sternebewertung 4")

abline(RegressionTeilmenge4, col="red")

plot(Teilmenge5$FreundlichkeitMitarbeiter,Teilmenge5$QualitätHaarschnitt,xlab="Freu
ndlichkeit Mitarbeiter",ylab="Qualität Haarschnitt",xlim = c(0,100),ylim = c(0,100)
,main="Sternebewertung 5")

abline(RegressionTeilmenge5,col="red")
```

Quellcode

```
plot(Teilmenge4$FreundlichkeitMitarbeiter,Teilmenge4$QualitätHaarschnitt,xlab="Freu
ndlichkeit Mitarbeiter",ylab="Qualität Haarschnitt",xlim = c(0,100),ylim = c(0,100)
,main="Sternebewertung 4")

abline(RegressionTeilmenge4, col="red")
```



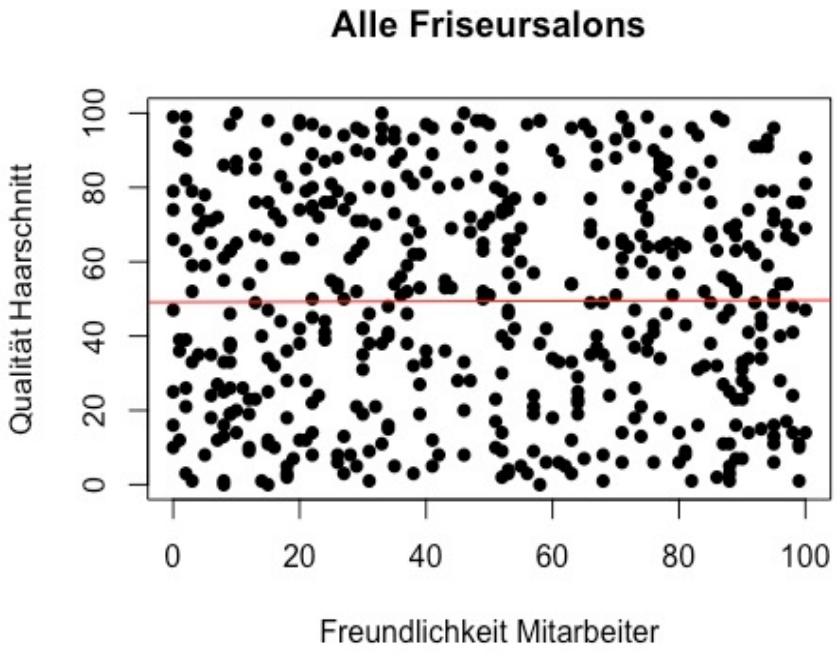
Quellcode

```
summary(RegressionDatensatz)

summary(RegressionTeilmenge1)

summary(RegressionTeilmenge2)

summary(RegressionTeilmenge3)

summary(RegressionTeilmenge4)

summary(RegressionTeilmenge5)
```

Quellcode

```
summary(RegressionDatensatz)
```

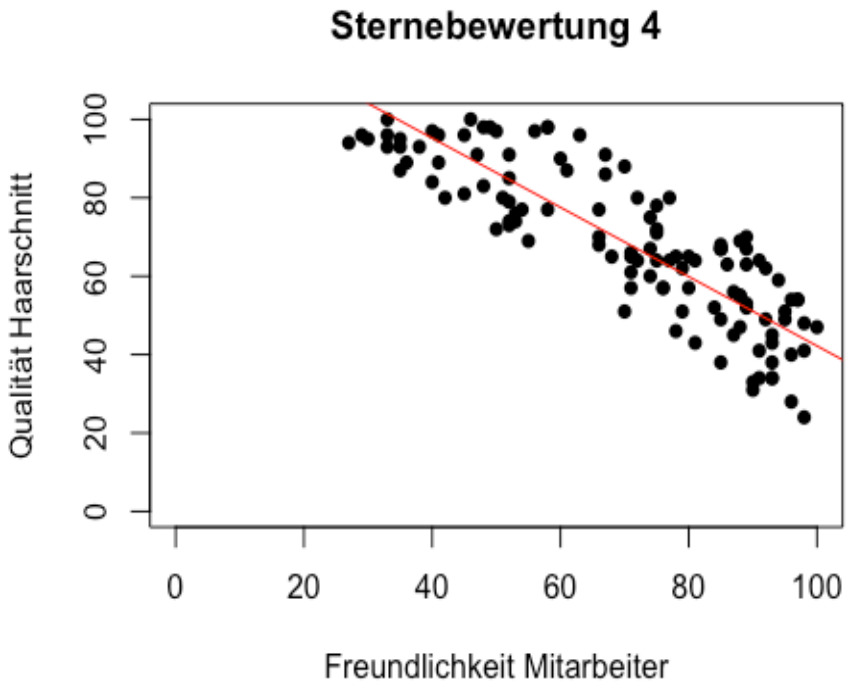
```
## Call:
## lm(formula = FreundlichkeitMitarbeiter ~ QualitätHaarschnitt,
##     data = Datensatz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.643 -27.410  -0.011  27.421  50.797
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         49.130909   2.701680  18.185   <2e-16 ***
## QualitätHaarschnitt  0.005176   0.046225   0.112    0.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.51 on 498 degrees of freedom
## Multiple R-squared:  2.517e-05,  Adjusted R-squared:  -0.001983
## F-statistic: 0.01254 on 1 and 498 DF,  p-value: 0.9109
```



Quellcode

```
summary(RegressionTeilmenge4)
```

```
## Call:
## lm(formula = FreundlichkeitMitarbeiter ~ QualitätHaarschnitt,
##     data = Teilmenge4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4720  -9.1914  -0.6513   8.7643  20.3128
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         130.56488    3.60464   36.22   <2e-16 ***
## QualitätHaarschnitt  -0.88397    0.05073  -17.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.78 on 116 degrees of freedom
## Multiple R-squared:  0.7236, Adjusted R-squared:  0.7212
## F-statistic: 303.6 on 1 and 116 DF,  p-value: < 2.2e-16
```



4. Fazit

Fazit

Im Rahmen der Studie können sehr schnell **Scheinkorrelationen** entstehen!

Collider-Bias bei Datenerhebung beachten, besonders bei unrealistischen Forschungsergebnissen

Wenn man sich über einen möglichen Collider-Bias bewusst ist, lassen sich korrekte kausale Schlussfolgerungen ziehen

Bei kausalen Analysen lohnt es sich also, bereits bei der Datenerhebung über mögliche Collider nachzudenken!

Vielen Dank für Eure Aufmerksamkeit!

Gibt es noch Fragen?