

## Lösung zu Kapitel 11: Beispiel 1

Eine Untersuchung bei 253 Personen zur Kundenzufriedenheit mit einer Einzelhandelskette im Südosten der USA enthält Variablen mit sozialstatistischen Daten der befragten Person, verschiedene Fragen zur Kundenzufriedenheit und spezifische Fragen, wie die Kundenzufriedenheit verbessert werden könnte.

- Um die Dimensionalität der Kundenzufriedenheit zu erforschen, soll eine Hauptkomponentenanalyse durchgeführt werden (Variablen `perf_1` bis `perf_20` im Datenfile `konsumenten.dat`).
- Zusätzlich sollen Forschungshypothesen formuliert und untersucht werden, die Unterschiede zwischen Konsumentengruppen (gebildet aus den sozialstatistischen Daten, wie z.B. Geschlecht, Alter, Einkommen) bezüglich der neuen Variablen (Komponenten) zum Gegenstand haben.

**R**

```
> konsumenten <- read.table("konsumenten.dat", header = TRUE)
> dat <- konsumenten[, 18:37]
```

Ob die vorhandenen Daten für eine Hauptkomponentenanalyse geeignet sind, wurde vorweg mittels der Kaiser-Meyer-Olkin-Statistik überprüft. Diese berechnet das Maß der Korrelation, welches in den Daten steckt - unter Berücksichtigung des Ausmaßes an partieller Korrelation.

**R**

```
> library("rela")
> test <- paf(as.matrix(dat))
> cat("KMO Statistik:", test$KMO, " Bartlett-Statistik:", test$Bartlett,
+     "\n")
```

KMO Statistik: 0.88835 Bartlett-Statistik: 2116.9

**R**

```
> library("psych")
> bart <- cortest.bartlett(cor(dat), n = nrow(dat))
> unlist(bart)
```

chisq	p.value	df
2.1169e+03	4.2984e-322	1.9000e+02

Um die Dimensionalität der Kundenzufriedenheit zu erforschen, wird eine Hauptkomponentenanalyse mit anschließender Varimax-Rotation durchgeführt, wobei die Faktoren auf Basis ihrer Korrelation erstellt werden. D.h. es werden die Gruppen mit der höchsten Korrelation verwendet.

**R**

```
> pca <- principal(dat, 4, rotate = "none")
> pca$criteria <- NULL
> pca
```

### Principal Components Analysis

Call: principal(r = dat, nfactors = 4, rotate = "none")

Standardized loadings based upon correlation matrix

	PC1	PC2	PC3	PC4	h2	u2
perf_1	0.02	-0.45	0.32	0.56	0.62	0.38
perf_2	0.44	0.61	-0.01	0.22	0.61	0.39
perf_3	0.42	0.66	-0.22	0.11	0.67	0.33
perf_4	0.32	0.42	0.47	-0.34	0.61	0.39
perf_5	0.23	0.48	0.46	-0.32	0.60	0.40
perf_6	0.71	-0.02	0.17	0.02	0.53	0.47
perf_7	0.44	0.45	0.19	-0.06	0.44	0.56
perf_8	0.44	0.17	0.08	0.45	0.43	0.57
perf_9	0.51	0.55	-0.07	0.36	0.69	0.31
perf_10	0.74	-0.12	-0.13	-0.04	0.58	0.42
perf_11	0.78	-0.31	0.03	-0.12	0.71	0.29
perf_12	0.81	-0.21	-0.09	0.07	0.72	0.28
perf_13	0.73	-0.06	-0.02	0.21	0.58	0.42
perf_14	0.70	-0.25	-0.04	-0.14	0.58	0.42
perf_15	0.74	-0.26	0.01	-0.05	0.62	0.38
perf_16	0.73	-0.10	0.07	-0.01	0.55	0.45
perf_17	0.48	-0.13	-0.27	-0.03	0.33	0.67
perf_18	0.63	-0.25	0.17	-0.06	0.50	0.50
perf_19	0.65	-0.30	-0.09	-0.29	0.60	0.40
perf_20	0.25	0.36	-0.64	-0.21	0.64	0.36

	PC1	PC2	PC3	PC4
SS loadings	6.71	2.54	1.21	1.14
Proportion Var	0.34	0.13	0.06	0.06
Cumulative Var	0.34	0.46	0.52	0.58

Man sieht, dass die erste Komponente einen sehr hohen Eigenwert von 10,052 hat und 50 % der Gesamtvarianz ausschöpft. Die zweite Komponente hat einen Eigenwert von 2,095 und hat bereits einen wesentlich geringeren Erklärungswert von 10 %. Ansonsten gibt es noch zwei weitere Hauptkomponenten deren Eigenwerte (gerade noch) größer als 1 ist. Durch diese vier Komponenten können insgesamt ca. 73 % der Gesamtvarianz erklärt werden.

### R

```
> pca.rotate <- principal(dat, 4, scores = TRUE)
> pca.rotate$criteria <- NULL
> print(pca.rotate, cut = 0.5, sort = TRUE, digits = 2)
```

### Principal Components Analysis

Call: principal(r = dat, nfactors = 4, scores = TRUE)

Standardized loadings based upon correlation matrix

	item	RC1	RC2	RC3	RC4	h2	u2
perf_11	11	0.84				0.71	0.29
perf_12	12	0.81				0.72	0.28
perf_15	15	0.78				0.62	0.38
perf_14	14	0.76				0.58	0.42
perf_19	19	0.75				0.60	0.40
perf_10	10	0.72				0.58	0.42
perf_16	16	0.69				0.55	0.45
perf_18	18	0.67				0.50	0.50

```

perf_13  13 0.64          0.58 0.42
perf_6   6 0.62          0.53 0.47
perf_17  17 0.51          0.33 0.67
perf_9   9      0.80      0.69 0.31
perf_2   2      0.72      0.61 0.39
perf_3   3      0.69      0.67 0.33
perf_8   8      0.57      0.43 0.57
perf_4   4      0.76      0.61 0.39
perf_5   5      0.76      0.60 0.40
perf_7   7      0.44      0.44 0.56
perf_20  20      0.73 0.64 0.36
perf_1   1     -0.73 0.62 0.38

```

```

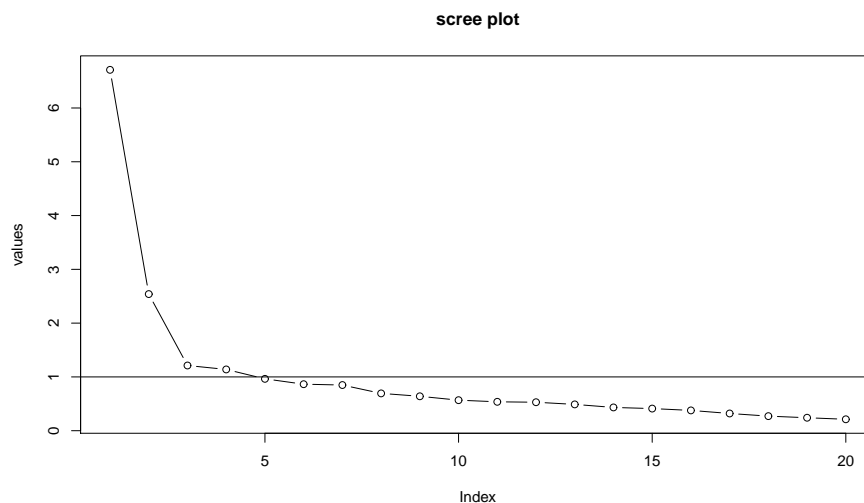
          RC1 RC2 RC3 RC4
SS loadings  5.78 2.68 1.72 1.42
Proportion Var 0.29 0.13 0.09 0.07
Cumulative Var 0.29 0.42 0.51 0.58

```

Um die Hauptkomponenten inhaltlich sinnvoll interpretieren zu können, lässt man sich zudem einen Screeplot ► Abbildung 1 als grafische Unterstützung erstellen. In diesem Fall stimmt die Anzahl der Hauptkomponenten, deren Eigenwert größer als 1 ist, mit der Anzahl der Hauptkomponenten, welche im Scree-Plot links von der Knickstelle liegen, überein.

**R**

```
> VSS.scree(dat)
```



**Abbildung 1:** Screeplot

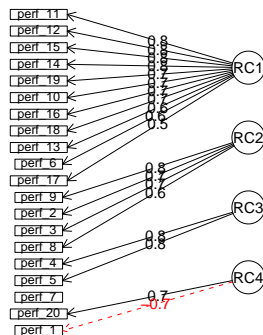
**R**

```
> fa.diagram(pca.rotate, cut = 0.5, cex = 0.8, rsize = 0.5, main = "")
```

Um die Hauptkomponenten zu interpretieren, sucht man jene Variablen, die stark mit einer Komponente korrelieren und vergibt geeignete Namen für die gemeinsamen Eigenschaften der Items, die auf einer Komponente hoch laden.

Nachdem die neu gewonnenen Variablen benannt wurden

■ **FAKTOR 1 ...quality of service**



- FAKTOR 2 ...customer care
- FAKTOR 3 ...promptness
- FAKTOR 4 ...environment

wird mit diesen weitergerechnet.

Im Folgenden sollen nun 2 Forschungshypothesen überprüft werden:

1. Die **quality of service** des Shops wird von Frauen und Männern unterschiedlich beurteilt.
2. Das **customer care** der Angestellten im Shop hat einen Einfluss auf die Treue der Kunden.

Hinweis: Diese dienen nur als Beispiele! Es können auch zahlreiche andere Fragestellungen bzw. Hypothesen formuliert werden.

Um die erste Hypothese zu überprüfen, kann ein Boxplot zur grafischen Darstellung verwendet werden.

Anhand der Grafik lässt sich erkennen, dass die **quality of service** des Shops von den Männern und Frauen sehr ähnlich bewertet wurde, die Bewertungen der Frauen hinsichtlich dieser Komponente jedoch einer größeren Streuung unterliegen.

Da die Boxplots auf eine schiefe Verteilung hinweisen und der Stichprobenumfang (mit 63 gültigen Werten) nicht all zu groß ist, wird zur Überprüfung, ob der Unterschied zwischen den Männern und Frauen in Bezug auf die Hauptkomponente **quality of service** signifikant ist, wird ein t-Test gerechnet.

**R**

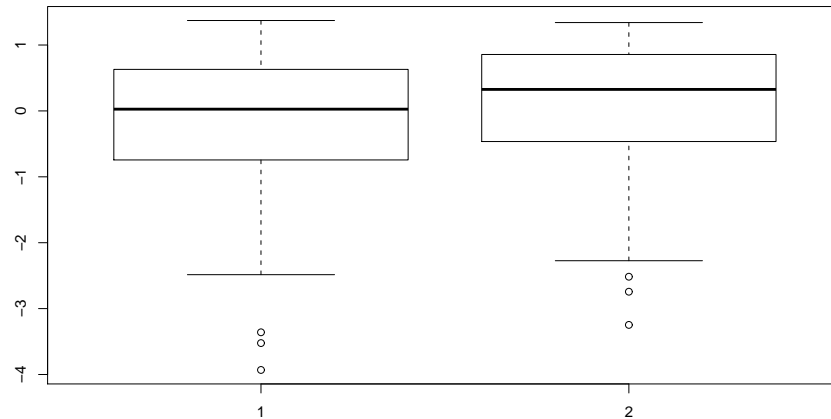
```
> t.test(pca.scores[, 1] ~ konsumenten$gender)
```

Welch Two Sample t-test

```
data:  pca.scores[, 1] by konsumenten$gender
t = -1.9236, df = 161.99, p-value = 0.05615
alternative hypothesis: true difference in means is not equal to 0
```

R

```
> pca.scores <- data.frame(pca.rotate$scores)
> boxplot(pca.scores[, 1] ~ konsumenten$gender)
```



**Abbildung 2:** Boxplot für quality of service und gender

95 percent confidence interval:

-0.5321858 0.0069735

sample estimates:

mean in group 1 mean in group 2

-0.168151 0.094455

Die Höhe des p-Wertes (0.056) weist darauf hin, dass es sich um keinen signifikanten (allerdings nur knapp nicht signifikanten) Unterschied handelt.

Die erste Forschungshypothese wird somit verworfen.

Um zu bewerten, ob das **customer care** der Angestellten im Shop Einfluss auf die Treue der Kunden hat, wird eine Regression auf **length** und **FAKTOR2** gerechnet.

R

```
> reg <- lm(pca.scores[-c(30, 223), 2] ~ konsumenten[-c(30, 223),
+ 10])
> summary(reg)
```

Call:

```
lm(formula = pca.scores[-c(30, 223), 2] ~ konsumenten[-c(30,
223), 10])
```

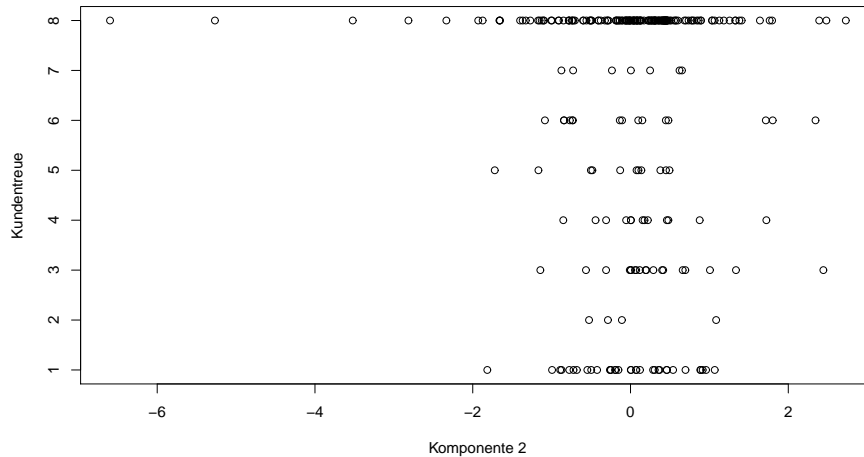
Residuals:

Min	1Q	Median	3Q	Max
-3.5617	-0.4692	0.0455	0.4051	2.6888

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.06872	0.13516	0.51	0.61

```
> plot(pca.scores[, 2], konsumenten$length, xlab = "Komponente 2",
+      ylab = "Kundentreue")
```



```
konsumenten[-c(30, 223), 10] -0.00352    0.02036   -0.17    0.86
```

Residual standard error: 0.851 on 249 degrees of freedom

Multiple R-squared: 0.00012, Adjusted R-squared: -0.0039

F-statistic: 0.0299 on 1 and 249 DF, p-value: 0.863

Aufgrund der Höhe des p-Wertes (0.86) ist auch die zweite Forschungshypothese zu verwerfen. Das bedeutet, dass das **customer care** der Angestellten im Shop keinen signifikanten Einfluss auf die Treue der Kunden hat.