Serie 02: Deskriptive Statistik (multivariate Daten)

Aufgabe 1

Gegeben sind die folgenden Daten

i	1	2	3	4	5	6	7	8
x_i	5	10	20	8	4	6	12	15
y_i	27	46	73	40	30	28	47	59

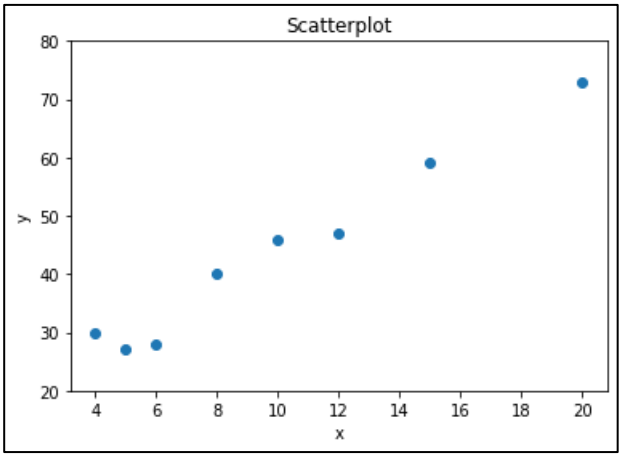
a) Zeichnen Sie das Streudiagramm der Daten.

b) Bestimmen Sie den Korrelationskoeffizienten nach Bravais-Pearson

c) Bestimmen Sie den Korrelationskoeffizienten nach Spearman.

Lösung:

a)




```
import numpy as np
import matplotlib.pyplot as plt

##Aufgabe 2.1
x=np.array([5, 10, 20, 8, 4, 6, 12, 15])
y=np.array([27, 46, 73, 40, 30, 28, 47, 59])
#Streudiagramm
plt.figure(1)
plt.scatter(x,y)
plt.xlabel('x')
plt.ylabel('y')
plt.ylim(20,80)
plt.title('Scatterplot')
```

b)

i	1	2	3	4	5	6	7	8	
x	5	10	20	8	4	6	12	15	$\bar{x} = 10$
y	27	46	73	40	30	28	47	59	$\bar{y} = 43.75$
x_i^2	25	100	400	64	16	36	144	225	$\overline{x_i^2} = 126.25$
y_i^2	729	2116	5329	1600	900	784	2209	3481	$\overline{y_i^2} = 2143.5$
$x_i y_i$	135	460	1460	320	120	168	564	885	$\overline{x_i y_i}$=514

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{514 - 10 \cdot 43.75}{\sqrt{126.25 - 10^2} \cdot \sqrt{2143.5 - 43.75^2}} = 0.986$$

c)

i	1	2	3	4	5	6	7	8
x	5	10	20	8	4	6	12	15
y	27	46	73	40	30	28	47	59
$rg(x_i)$	2	5	8	4	1	3	6	7
$rg(y_i)$	1	5	8	4	3	2	6	7
d_i	1	0	0	0	-2	1	0	0

Es sind alles ungebundene Ränge daher kann man die Vereinfachungsformel nutzen:

$$r_{Sp} = 1 - \frac{6 \cdot \sum_{i=1}^{n} {d_i}^2}{n \cdot (n^2 - 1)} \qquad \text{mit } d_i = \text{rg}(x_i) - \text{rg}(y_i)$$

$$= 1 - \frac{6 \cdot (1^2 + 0 + 0 + 0 + (-2)^2 + 1^2 + 0 + 0)}{8 \cdot (8^2 - 1)} = 1 - \frac{6 \cdot 6}{8 \cdot 63} = 0.929$$

Aufgabe 2

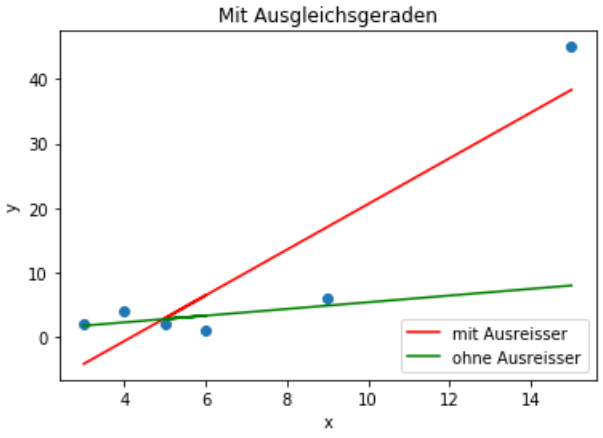
Gegeben die Datenpaare

i	1	2	3	4	5	6
x_i	3	4	6	5	9	15
y_i	2	4	1	2	6	45

a) Zeichnen Sie das Streudiagramm
b) Bestimmen Sie den Korrelationskoeffizienten nach Bravais-Pearson mit und ohne den letzten Wert (Ausreisser). Was fällt Ihnen auf?
c) Bestimmen Sie den Korrelationskoeffizienten nach Spearman mit und ohne den letzten Wert (Ausreisser). Was fällt Ihnen auf?

Lösung:

a)



b)

i	1	2	3	4	5	6	Mittelwerte	Ohne Ausreisser
x_i	3	4	6	5	9	15	7	5.4
y_i	2	4	1	2	6	45	10	3
x_i^2	9	16	36	25	81	225	65.33	33.4
y_i^2	4	16	1	4	36	2025	347.67	12.2
$x_i y_i$	6	16	6	10	54	675	127.83	18.4

Mit Ausreisser:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{127.83 - 7 \cdot 10}{\sqrt{65.33 - 7^2} \cdot \sqrt{347.67 - 10^2}} = 0.909$$

Ohne Ausreisser:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{18.4 - 5.4 \cdot 3}{\sqrt{33.4 - 5.4^2} \cdot \sqrt{12.2 - 3^2}} = 0.597$$

Obwohl grafisch die Daten ohne den Ausreisser eher linear aussehen, liefert der Korrelationskoeffizient für den Fall mit Ausreisser einen höheren Wert.

c)

i	1	2	3	4	5	6	Mittelwerte	Ohne Ausreisser
x_i	3	4	6	5	9	15	7	5.4
y_i	2	4	1	2	6	45	10	3
$rg(x_i)$	1	2	4	3	5	6	3.5	3
$rg(y_i)$	2.5	4	1	2.5	5	6	3.5	3
$\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}$		-1.5	0.5	-0.5	1.5	2.5	Mit Ausr.	
$\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}$	-1	0.5	-2.5	-1	1.5	2.5		
$\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}$		-1	1	0	2		Ohne Ausr.	
$\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}$	-0.5	1	-2	-0.5	2			

Mit Ausreisser:

$$r_{Sp} = \frac{\sum_{i=1}^{n} \left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}\right)}{\sqrt{\sum_{i=1}^{n} \left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} \left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}\right)^2}}$$

$$\sqrt{\sum_{i=1}^{n} \left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}\right)^2} = \sqrt{(-2.5)^2 + (-1.5)^2 + 0.5^2 + (-0.5)^2 + 1.5^2 + 2.5^2} = \sqrt{17.5}$$

$$\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)^2}=\sqrt{(-1)^2+0.5^2+(-2.5)^2+(-1)^2+1.5^2+2.5^2}=\sqrt{17}$$

$$\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)$$
$$=(-2.5)\cdot(-1)+(-1.5)\cdot 0.5+0.5\cdot(-2.5)+(-0.5)\cdot(-1)+1.5\cdot 1.5+2.5\cdot 2.5\ =9.5$$
$$r_{Sp}=\frac{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)^2}}=\frac{9.5}{\sqrt{17.5}\cdot\sqrt{17}}\approx 0.551$$

Ohne Ausreisser:

$$r_{Sp}=\frac{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)^2}}$$

$$\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)^2}=\sqrt{(-2)^2+(-1)^2+1^2+0^2+2^2}=\sqrt{10}$$

$$\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)^2}=\sqrt{(-0.5)^2+1^2+(-2)^2+(-0.5)^2+2^2}=\sqrt{9.5}$$

$$\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)$$
$$=(-2)\cdot(-0.5)+(-1)\cdot 1+1\cdot(-2)+0\cdot(-0.5)+2\cdot 2\ =2$$
$$r_{Sp}=\frac{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i)-\overline{\mathrm{rg}(x)}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(y_i)-\overline{\mathrm{rg}(y)}\right)^2}}=\frac{2}{\sqrt{10}\cdot\sqrt{9.5}}\approx 0.205$$

Obwohl grafisch die Daten ohne den Ausreisser eher linear aussehen, liefert der Korrelationskoeffizient für den Fall mit Ausreisser einen höheren Wert.

Aufgabe 3

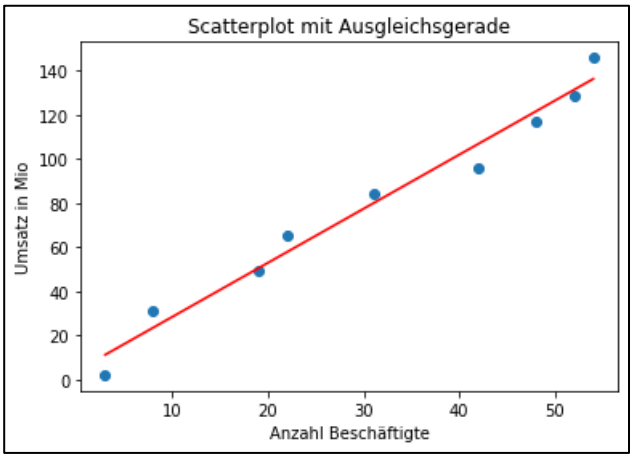
Für ein Unternehmen soll untersucht werden, welcher Zusammenhang zwischen Umsatz und Anzahl Beschäftigten gilt:

i	1	2	3	4	5	6	7	8	9
Anz. Beschäftigte x_i	3	8	19	22	31	42	48	52	54
Umsatz in Mio. y_i	2	31	49	65	84	96	117	129	146

Bewerten Sie den Zusammenhang der beiden Merkmale

Lösung:

Mit Python oder per Hand berechnet ergibt sich der folgende Streuplot (inkl. Gerade)



und der Korrelationskoeffizient (Pearson): $r_{xy} = 0.987$

Damit ist sowohl grafisch als auch über den Korrelationskoeffzienten ein starker linearer positiver Zusammenhang erkennbar. Das heisst, dass es evtl. einen Zusammenhang zwischen der Anzahl an Beschäftigten und dem Umsatz geben mag, allerdings kann auch ein drittes Merkmal, z.B. die Anzahl produzierter Produkte, diesen Zusammenhang eher erklären.

Aufgabe 4

Bei 8 zufällig ausgewählten Arbeitnehmerinnen und Arbeitnehmern wird die Schuhgrösse S und das jährliche Einkommen E erfasst. Es ergeben sich folgende Werte:

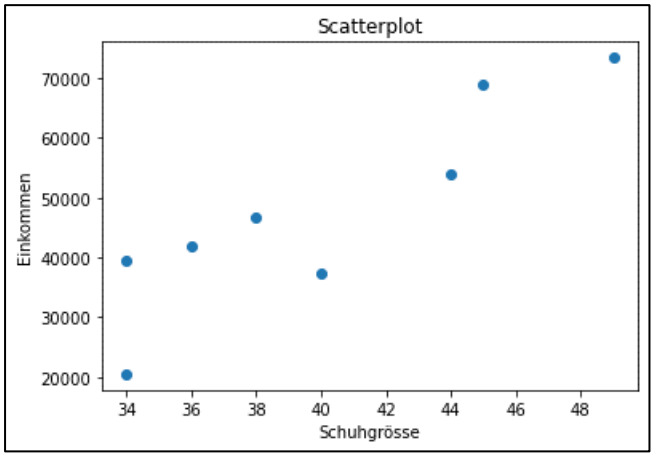
i	1	2	3	4	5	6	7	8
S_i	36	44	40	49	38	34	34	45
E_i	41910	53860	37360	73450	46720	39560	20470	69040

a) Berechnen Sie einen geeigneten Korrelationskoeffizienten zwischen Schuhgrösse und Einkommen. Interpretieren Sie das Ergebnis.

b) Bewerten Sie das Modell bezüglich Kausalität versus Korrelation und geben Sie ein potentielles latentes Merkmal an.

Lösung:

a)



Und der Korrelationskoeffizient (Pearson): $r_{xy} = 0.8945$

Und der Korrelationskoeffizient (Spearman): $r_{xy} = 0.8383$

Damit ist sowohl grafisch als auch über den Korrelationskoeffzienten ein starker linearer positiver Zusammenhang erkennbar.

b) Selbst wenn es eine Korrelation zwischen Schuhgrösse und Einkommen gibt, existiert dabei vermutlich keine direkte Kausalität. Vielmehr spielt ein drittes Merkmal («Geschlecht») eine Rolle. Um dies zu betrachten, müsste man die Auswertung nochmals je Geschlecht durchführen.

Aufgabe 5

In einer Region soll der Zusammenhang zwischen Gasverbrauch ($gasv$), Gaspreis (gpr) und Fernwärmepreis (fpr) untersucht werden.

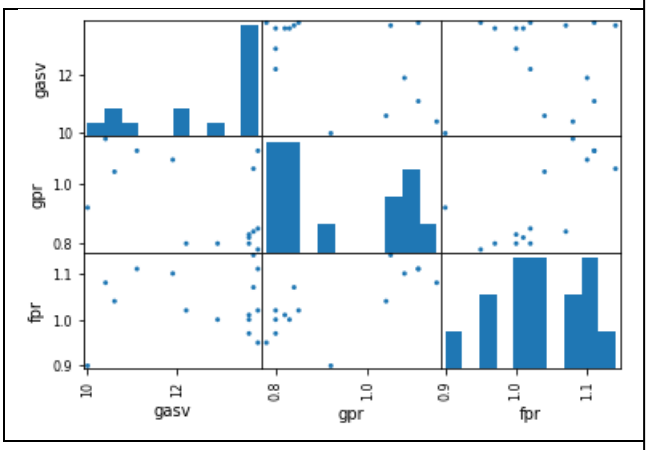
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$gasv$	10	10.6	10.4	11.1	11.9	13.8	13.7	13.7	12.2	12.9	13.6	13.8	13.6	13.6	13.8
gpr	0.92	1.04	1.15	1.11	1.08	1.11	1.05	0.84	0.80	0.80	0.82	0.85	0.83	0.80	0.78
fpr	0.90	1.04	1.08	1.11	1.10	1.11	1.14	1.07	1.02	1.00	1.01	1.02	1.00	0.97	0.95

a) Stellen Sie den Zusammenhang von jeweils zwei Merkmalen grafisch dar

b) Geben Sie die Korrelation der Merkmale an. Wie schätzen Sie den Zusammenhang zwischen den Merkmalen ein.

Lösung:

a)




```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#Aufgabe 2.12
#Gaspreis
gasv=np.array([10, 10.6, 10.4, 11.1, 11.9, 13.8, 13.7,
   13.7, 12.2, 12.9, 13.6, 13.8, 13.6, 13.6, 13.8])
#Gaspreis
gpr=np.array([0.92, 1.04, 1.15, 1.11, 1.08, 1.11,
   1.05, 0.84, 0.80, 0.80, 0.82, 0.85, 0.83, 0.80,
   0.78])
#Fernwärmepreis
fpr=np.array([0.90, 1.04, 1.08, 1.11, 1.10, 1.11,
   1.14, 1.07, 1.02, 1.00, 1.01, 1.02, 1.00, 0.97,
   0.95 ])

#Zusammenführung der Daten in ein DataFrame
df=pd.DataFrame({'gasv':gasv,'gpr':gpr,'fpr':fpr})

#Matrix von Streudiagrammen
plt.figure(1)
pd.plotting.scatter_matrix(df, alpha=1)
plt.show()
```

b)

	gasv	gpr	fpr
gasv	1.000000	-0.506658	0.058703
gpr	-0.506658	1.000000	0.711762
fpr	0.058703	0.711762	1.000000

Es scheint hier sowohl grafisch als auch rechnerisch eine Korrelation zwischen Gaspreis und Fernwärmepreis zu geben. Dieser Zusammenhang mag allerdings ebenfalls durch ein weiteres Merkmal erklärt werden wie z.B. die Aussentemperatur.