# 2 Deskriptive Statistik (mehrere Merkmale)

Bis jetzt haben wir uns auf die Beschreibung von einzelnen Merkmalen beschränkt. Bei vielen Fragestellungen sind jedoch verschiedene Merkmale bei einem Merkmalsträger interessant. Teilweise ist auch das Ziel der Untersuchungen, Zusammenhänge zwischen diesen Merkmalen zu finden. Solche Daten werden als bivariate Daten bei zwei Merkmalen und allgemein als multivariate Daten bei mehreren Merkmalen bezeichnet.
Wenn wir mit multivariaten Daten arbeiten, bietet sich das Python Modul «pandas» an, da wir dort nicht nur mit Matrizen, sondern vielmehr mit DataFrames arbeiten. Darin können einfacher auch kategorielle und metrische Daten miteinander kombiniert werden.

## 2.1 Bivariate Daten

### 2.1.1 Grafische Darstellung

Wir zeigen am folgenden Beispiel wie man multivariate Daten grafisch präsentieren kann indem man Auswahlen von 2 Merkmalen trifft und diese individuell betrachtet. Zu jeder getroffenen Auswahl von 2 Merkmalen werden die Daten gruppiert und in einem 2-dimensionalen Plot dargestellt. Die Art der Grafik richtet sich dabei nach der Merkmalsausprägung der ausgewählten Merkmale.

**Beispiel 1: Bewertung der Kaufkraft**
Bei einem Dienstleister, der das Management von Kundenbindungsprogrammen durchführt, steht eine Tabelle mit den Informationen über total 16’881 Kunden zur Verfügung. In der unten abgebildeten Tabelle sind die Eigenschaften der ersten 10 Kunden dargestellt.

| ID | Alter | Zivilstand | Geschlecht | Personen im Haushalt | Kaufkraft | Einkaufs-betrag |
|---|---|---|---|---|---|---|
| 1 | 50 | Ledig | Mann | 1 | hoch | 10951 |
| 2 | 86 | Verheiratet | Frau | 2 | sehr hoch | 23091 |
| 4 | 65 | Ledig | Mann | 1 | sehr hoch | 18866 |
| 6 | 58 | Ledig | Mann | 1 | mittel | 2243 |
| 7 | 58 | Partnerschaft | Frau | 3 | hoch | 7669 |
| 9 | 54 | Verheiratet | Mann | 2 | sehr hoch | 23643 |
| 10 | 43 | Verheiratet | Mann | 5 | mittel | 5993 |
| 12 | 84 | Partnerschaft | Frau | 2 | sehr hoch | 22379 |
| 13 | 52 | Partnerschaft | Mann | 4 | sehr hoch | 12499 |
| 14 | 53 | Verheiratet | Mann | 2 | mittel | 7439 |
| … | … | … | … | … | … | … |

Zusätzlich zur Personen-ID sind sechs Merkmale vorhanden: Das Alter, die Personenzahl im Haushalt und der Einkaufbetrag sind metrische Merkmale. Der Zivilstand und das Geschlecht sind kategoriell-nominale Merkmale mit drei bzw. zwei Ausprägungen. Die Kaufkraft ist eine kategoriell-ordinale Grösse mit vier Ausprägungen («tief», «mittel», «hoch» und «sehr hoch»).

### *Zwei kategorielle Merkmale: Kaufkraft und Zivilstand*

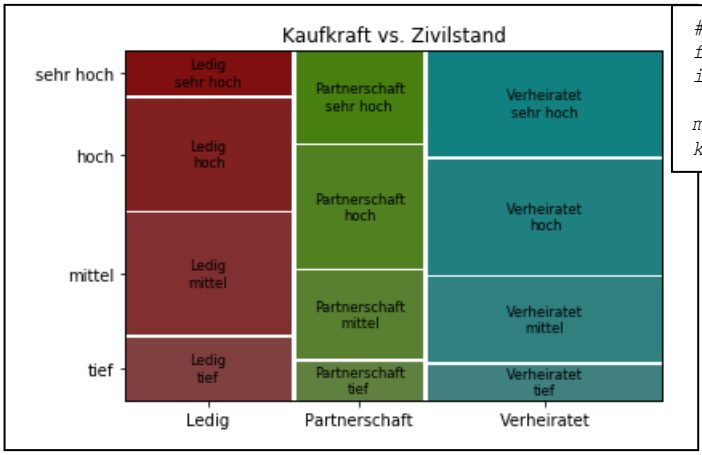
Möchten wir zwei kategorielle Merkmale vergleichen,
bieten sich sogenannte Kontingenztabellen an.

In einer solchen Häufigkeitstabelle für zwei Merkmale $A$ und $B$ beziehen sich die Zeilen auf Merkmal $A$ und die Spalten auf Merkmal $B$. Der Eintrag in Zeile $x$ und Spalte $y$ ist die (absolute oder relative) Zahl der Merkmalsträger, welche Ausprägung $x$ von Merkmal $A$ und die Ausprägung $y$ von Merkmal $B$ aufweisen. Dabei hilft es, bei ordinalen Merkmalen, deren Ordnung zu berücksichtigen. Unten ist eine Kontingenztabelle für die Merkmale Zivilstand und Kaufkraft zu sehen.

| Kaufkraft | tief | mittel | hoch | sehr hoch |
|---|---|---|---|---|
| Zivilstand | | | | |
| Ledig | 978 | 1885 | 1748 | 695 |
| Partnerschaft | 457 | 1044 | 1452 | 1091 |
| Verheiratet | 781 | 1870 | 2550 | 2330 |

```
#Zwei kategorielle Merkmale
import pandas as pd

df=pd.read_excel('Bsp_2_1.xlsx')
print (df.columns) #Bezeichnung der
Datenreihen

#Kontingenztabelle
df['Kaufkraft'] =
df['Kaufkraft'].astype('category').cat.se
t_categories(['tief','mittel','hoch','seh
r hoch'])
table=pd.crosstab(df['Zivilstand'],df['Ka
ufkraft'])
print(table)
```

Neben der tabellarischen Darstellung kann dieselbe Information auch als
Mosaikplot dargestellt werden




```
#Mosaikplot
from statsmodels.graphics.mosaicplot
import mosaic

mosaic(table.stack(),gap=0.01,title='Kauf
kraft vs. Zivilstand')
```

### *Ein kategorielles und ein metrisches Merkmal: Kaufkraft und Einkaufsbetrag*

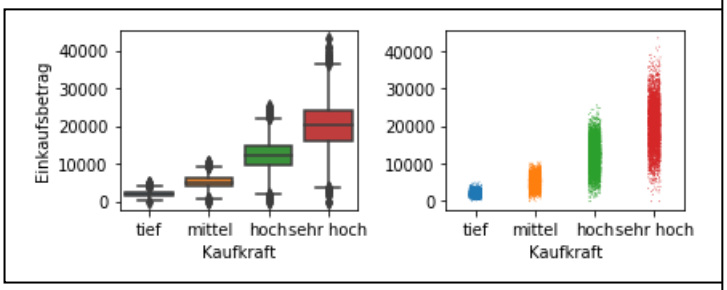
Möchten wir ein metrisches Merkmal $A$ zusammen mit einem kategoriellen Merkmal $B$ präsentieren, können wir für jede Ausprägung von Merkmal $B$ eine eigene Statistik nach den Ausprägungen von Merkmal $A$ bilden, zum Beispiel in Form eines Histogramms, eines Boxplots oder einer Aufstellung der Kennwerte.
Unten im Bild sind die Kennwerte (Mittelwert und Standardabweichung) der verschiedenen Ausprägungen tief, mittel hoch, sehr hoch des Merkmals Kaufkraft in Bezug zum metrischen Merkmal Einkaufsbetrag zu sehen.

| | Mittelwert | Standardabweichung |
|---|---|---|
| tief | 2055.965253 | 738.448383 |
| mittel | 5059.689519 | 1577.749078 |
| hoch | 12053.706609 | 3715.787543 |
| sehr hoch | 20197.653061 | 6075.097007 |

```
#Kategorielles und metrisches Merkmal
import pandas as pd

#Auswertung für alle drei getrennt
t=df.loc[df['Kaufkraft'] == 'tief',:]
m_t=t['Einkauf'].mean()
s_t=t['Einkauf'].std()
m=df.loc[df['Kaufkraft'] == 'mittel',:]
m_m=m['Einkauf'].mean()
s_m=m['Einkauf'].std()
h=df.loc[df['Kaufkraft'] == 'hoch',:]
m_h=h['Einkauf'].mean()
s_h=h['Einkauf'].std()
s=df.loc[df['Kaufkraft'] == 'sehr hoch',:]
m_s=s['Einkauf'].mean()
s_s=s['Einkauf'].std()
KW=pd.DataFrame(([m_t,s_t],[m_m,s_m],[m_h,
s_h],[m_s,s_s]),index=['tief','mittel','ho
ch','sehr hoch'],
columns=['Mittelwert','Standardabweichung'
])
print(KW)
```

Einfacher ist es die eine Auswertung des
metrischen Merkmals für die verschiedenen kategoriellen Ausprägungen
als Boxplots oder Stripcharts zu machen.

Hierfür existiert das Modul «seaborn», das etwas mehr Optionen bei der grafischen Darstellung von Dataframes bietet.

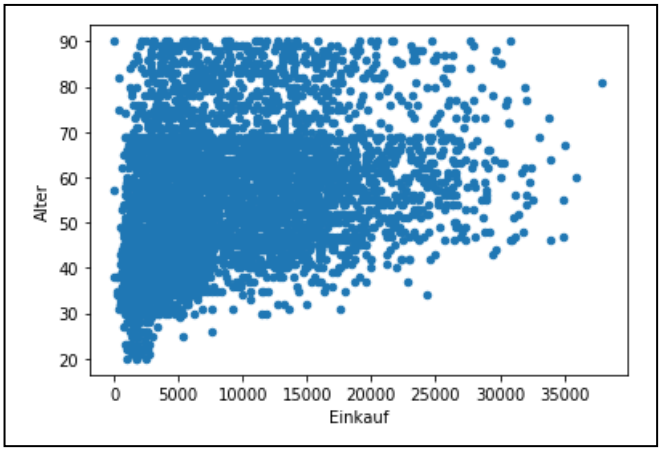



```
import matplotlib.pyplot as plt
import seaborn as sns

#Boxplot
plt.figure()
plt.subplot(2,2,1)
df_=pd.DataFrame(([t['Einkauf'],m['Ei
nkauf'],h['Einkauf'],s['Einkauf']]),i
ndex=['tief', 'mittel', 'hoch','sehr
hoch'])
df_=df_.transpose()
sns.boxplot( data=df_)
#df_.boxplot() Alternative mit Pandas
plt.ylabel('Einkaufsbetrag')
plt.xlabel('Kaufkraft')

#Stripplots
plt.subplot(2,2,2)
sns.stripplot( data=df_,size=0.8)
plt.xlabel('Kaufkraft')
plt.tight_layout()
```

In den Kennwerten können wir ebenso wie bei den beiden Plots erkennen, dass bei höherer Kaufkraft im Mittel der Einkaufsbetrag höher ist, aber auch eine grössere Streuung besitzt. Dies lässt sich intuitiv begründen: Denn auch wenn man viel Geld hat, kann man durchaus auch wenig kaufen.

### *Zwei metrische Merkmale: Einkaufsbetrag und Alter*

Die Stichprobenwerte zweier metrischer Merkmale können wie oben beschrieben oder graphisch als Punkte in der $(x, y)$-Ebene dargestellt werden.
Die so entstehende Punktwolke nennt man
Streudiagramm (bzw. Scatterplot)




```
#Zwei metrische Merkmale
import pandas as pd

#Scatterplot mit einem Subset der Daten
df_=df.loc[df['Zivilstand'] == 'Ledig',:]
df_.plot.scatter(x='Einkauf', y='Alter')
```

Aus dem Scatterplot kann man die gemeinsame Verteilung der beiden Merkmale erkennen. Auch eventuelle Ausreisser werden sichtbar.
Bei einer grossen Stichprobengrösse fällt es allerdings meist schwer, über die Graphik Zusammenhänge zu erkennen. Es bedarf dazu weiterer Kenngrössen, die wir in den folgenden Beispielen mit kleineren Stichproben motivieren und im nachfolgenden Abschnitt erklären werden.

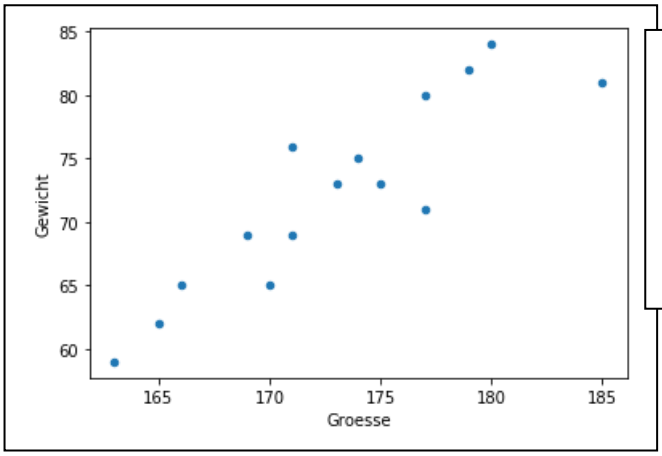
**Beispiel 2: Streudiagramme verschiedener Zusammenhänge**
Von 15 zufällig ausgewählten erwachsenen Personen werden Körpergrösse $x$ und Gewicht $y$ gemessen. Es ergeben sich folgende Wertepaare $(x_i,y_i)$ in cm bzw. kg:

(163,59), (165,62), (166,65), (169,69), (170,65), (171,69), (171,76), (173,73),
(174,75), (175,73) , (177,80) , (177,71), (179,82),(180,84), (185,81)

Wir tragen dafür in $x$-Richtung die Körpergrösse auf und in $y$-Richtung das Gewicht.
Die Daten der ersten Person werden also durch den Punkt (163,59) dargestellt, usw.

……

Man erhält das folgende Streudiagramm:




```
#Verschiedene Scatterplots
import pandas as pd

#Scatterplot gleichsinniger Zusammenhang
df1=pd.DataFrame({
'Groesse':[163,165,166,169,170,171,171,173,
174,175,177,177,179,180,185],
'Gewicht':[59,62,65,69,65,69,76,73,75,73,80
,71,82,84,81]})
df1.plot.scatter(x='Groesse',y='Gewicht')
```

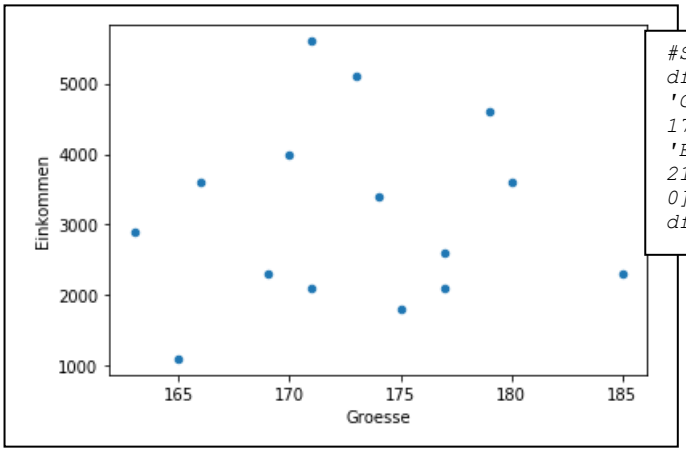
Aus der Abbildung kann man die Tendenz ablesen:
«je grösser, umso schwerer»

……

Nun werden für diese Personen deren Körpergrösse *x* und ihr monatliches Einkommen *y* ermittelt. Es ergeben sich folgende Wertepaare in cm bzw. CHF:

(163,2900), (165,1100), (166,3600), (169,2300), (170,4000),
(171,5600), (171,2100), (173,5100), (174,3400), (175,1800),
(177,2100), (177,2600), (179,4600), (180,3600), (185,2300)

Das Streudiagramm ist in der folgenden Abbildung dargestellt:




```
#Scatterplot ohne Zusammenhang
df2=pd.DataFrame({
'Groesse':[163,165,166,169,170,171,171,173,
174,175,177,177,179,180,185],
'Einkommen':[2900,1100,3600,2300,4000,5600,
2100,5100,3400,1800,2100,2600,4600,3600,230
0]})
df2.plot.scatter(x='Groesse',y='Einkommen')
```

Die Punkte liegen ziemlich regellos verteilt.

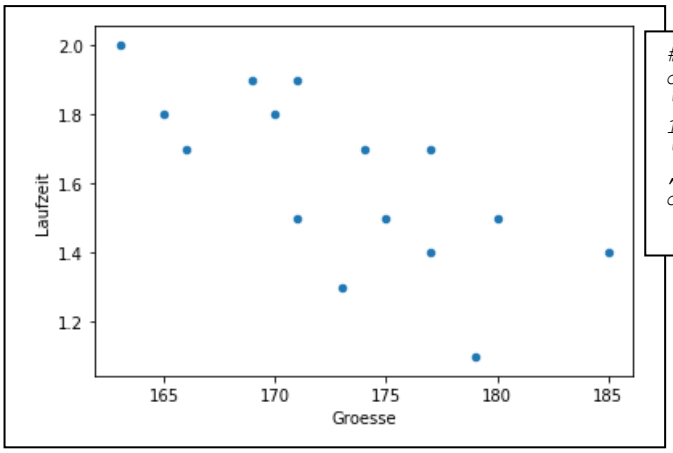
.....

Es unterstützt die Vermutung, dass es zwischen Grösse und Einkommen einer Person keinen Zusammenhang gibt.

Nun werden für diese Personen deren Körpergrösse *x* und ihre Laufzeiten *y* bei einem 400m Lauf ermittelt. Es ergeben sich folgende Wertepaare in cm bzw. min:

(163,2.0), (165,1.8), (166,1.7), (169,1.9), (170,1.8), (171,1.5), (171,1.9), (173,1.3),
(174,1.7), (175,1.5), (177,1.4), (177,1.7), (179,1.1), (180,1.5), (185,1.4)

Das Streudiagramm ist in der folgenden Abbildung dargestellt:




```
#Scatterplot gegensinniger Zusammenhang
df3=pd.DataFrame({
'Groesse':[163,165,166,169,170,171,171,173,
174,175,177,177,179,180,185],
'Laufzeit':[2.0,1.8,1.7,1.9,1.8,1.5,1.9,1.3
,1.7,1.5,1.4,1.7,1.1,1.5,1.4]})
df3.plot.scatter(x='Groesse',y='Laufzeit')
```

Es lässt sich eine gewisse Tendenz erkennen, dass
mit grösser werdenden x-Werten die y-Werte kleiner werden.

.....

Ein weiteres Beispiel für einen solchen gegensinnigen Zusammenhang wären Flugzeit und Kerosinverbrauch.

Man kann aus der grafischen Darstellung der beiden Merkmale also bereits die Form, Richtung und Stärke des Zusammenhangs erkennen.

**Form:**
Bei einem linearen Zusammenhang streuen die Punkte um eine Gerade herum. Es kann aber auch sein, dass eine gekrümmte Kurve, bzw. eine beliebige Funktion die Form des Zusammenhangs besser als eine Gerade beschreibt. Es ist auch möglich, dass mehrere separate Punktwolken vorhanden sind.

**Richtung**
Bei einem positiven Zusammenhang nehmen die Werte des einen Merkmals bei zunehmenden Werten des anderen Merkmals tendenziell zu. Bei einem negativen Zusammenhang würden kleinere $y$-Werte mit grösseren $x$-Werten auftreten.

**Stärke**
Der Zusammenhang im zweiten Beispiel ist nicht sehr stark. Wenn hingegen nur wenig Streuung vorliegt (wie im ersten Beispiel), und der Wert des einen Merkmals mit dem anderen recht genau vorhergesagt werden kann, so liegt ein starker Zusammenhang vor.

Diese drei Eigenschaften werden wir im Folgenden quantitativ bewerten.

### 2.1.2 Korrelation

Der Zusammenhang zwischen zwei metrischen Merkmalen kann bei grösseren Stichprobe mit einem Scatterplot meist nicht beurteilt werden. Mit speziellen Kennzahlen, sogenannten Korrelationskoeffizienten
……,
werden diese Zusammenhänge quantifiziert. Wir werden im folgenden zwei solche Kennwerte besprechen: Der Pearson-Korrelationskoeffizient misst die Stärke und Richtung des linearen Zusammenhangs und der Spearman Rangkorrelationskoeffizient misst die Stärke und Richtung des monotonen Zusammenhangs zwischen den $x$ und $y$ Werten in der Stichprobe.

#### *Pearson-Korrelationskoeffizient*

Dieser Kennwert misst die Stärke und die Richtung des *linearen Zusammenhangs*. Beurteilt wird, ob die Wertepaare nahe um eine Gerade herum liegen, oder stark darum streuen. Weil eine Gerade die einfachste Beschreibung des Zusammenhangs zwischen zwei metrischen Merkmalen darstellt, ist der Pearson-Korrelationskoeffizient (auch Bravais-Pearson-Korrelation oder Produkt-Moment-Korrelation) das am häufigsten benutzte Mass.

**Definition**

Gegeben seien die Wertepaare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
Die (*empirischen*) *Standardabweichungen* der $x_i$ bzw. der $y_i$-Werte lauten

$$\tilde{s}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \text{bzw.} \qquad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\tilde{s}_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} \qquad\qquad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Die (*empirische*) sogenannte *Kovarianz* lautet

$$\tilde{s}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) \qquad \text{bzw.} \qquad s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})$$

mit den *arithmetischen Mittelwerten* $\overline{x}, \overline{y}$.
Der *Korrelationskoeffizient* (nach Bravais-Pearson) wird dann definiert als:

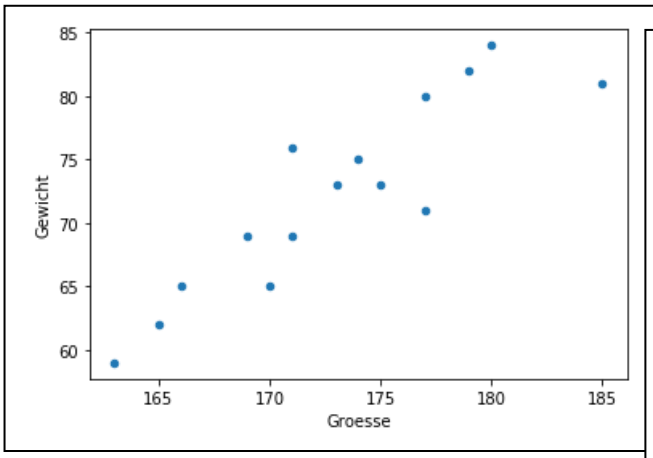
$$r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y} \qquad \text{bzw.} \qquad r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

**Bemerkung**

Bei der Definition des Korrelationskoeffizienten kommt es nicht darauf an, welche der empirischen Kovarianzen und Standardabweichungen (korrigiert oder nicht) verwendet werden.

**Beispiel 3: Korrelationskoeffizient (zu Beispiel 2)**

Berechnen Sie den Korrelationskoeffizienten der Stichproben aus Beispiel 2.
**Für Grösse und Gewicht:**



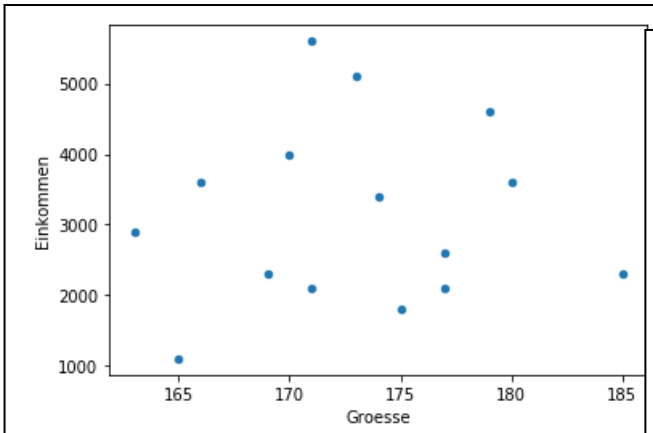

```
#Verschiedene Scatterplots
import pandas as pd

#Scatterplot gleichsinniger Zusammenhang
df1=pd.DataFrame({
'Groesse':[163,165,166,169,170,171,171,173,
174,175,177,177,179,180,185],
'Gewicht':[59,62,65,69,65,69,76,73,75,73,80
,71,82,84,81]})
df1.plot.scatter(x='Groesse',y='Gewicht')

#Kennwerte korr.
df=pd.concat([df1,df2.loc[:,'Einkommen'],df
3.loc[:,'Laufzeit']],axis=1)
m_x=df.mean()[0]
m_y1=df.mean()[1]
s_x=df.std()[0]
s_y1=df.std()[1]
s_xy1=df.cov().values[0,1]
r_xy1=df.corr().values[0,1]
```

| | |
|---|---|
| Mittelw.x | 173.000000 |
| Mittelw.y | 72.266667 |
| St.abw.x | 6.047432 |
| St.abw.y | 7.563698 |
| Kovarianz | 41.071429 |
| Korr.Koeff. | 0.897914 |

Die Python-Funktion gibt jeweils die korrigierten Werte für Standardabweichung und Kovarianz aus.

Es ergibt sich der Korrelationskoeffizient $r_{xy} = 0.898$. Der Korrelationskoeffizient ist nahe
ein starker positiver linearer Zusammenhang
bei 1, was besagt, dass ……
zwischen den beiden Merkmalen vorliegt.
Das positive Vorzeichen von $r_{xy}$ bedeutet, dass tendenziell
mit wachsenden $x_i$-Werten auch die $y_i$-Werte zunehmen.
……

**Für Grösse und Einkommen:**

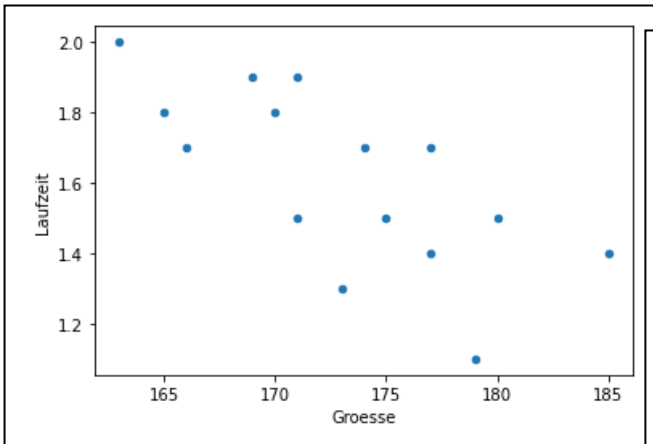



```
#Verschiedene Scatterplots
import pandas as pd

#Scatterplot ohne Zusammenhang
df2=pd.DataFrame({'Groesse':[163,165,166,16
9,170,171,171,173,174,175,177,177,179,180,1
85],
'Einkommen':[2900,1100,3600,2300,4000,5600,
2100,5100,3400,1800,2100,2600,4600,3600,230
0]})
df2.plot.scatter(x='Groesse',y='Einkommen')

#Kennwerte korr.
df=pd.concat([df1,df2.loc[:,'Einkommen'],df
3.loc[:,'Laufzeit']],axis=1)
m_x=df.mean()[0]
m_y2=df.mean()[2]
s_x=df.std()[0]
s_y2=df.std()[2]
s_xy2=df.cov().values[0,2]
r_xy2=df.corr().values[0,2]
```

| | |
|---|---|
| Mittelw.x | 173.000000 |
| Mittelw.y | 3140.000000 |
| St.abw.x | 6.047432 |
| St.abw.y | 1285.523795 |
| Kovarianz | 471.428571 |
| Korr.Koeff. | 0.060641 |

Es ergibt sich der Korrelationskoeffizient $r_{xy} = 0.061$. Der Korrelationskoeffizient liegt nahe bei 0. Das weist darauf hin, dass die beiden Merkmale nicht linear korreliert sind.

**Für Grösse und Laufzeit:**




```
#Verschiedene Scatterplots
import pandas as pd

#Scatterplot gegensinniger Zusammenhang
df3=pd.DataFrame({'Groesse':[163,165,166,16
9,170,171,171,173,174,175,177,177,179,180,1
85],
'Laufzeit':[2.0,1.8,1.7,1.9,1.8,1.5,1.9,1.3
,1.7,1.5,1.4,1.7,1.1,1.5,1.4]})
df3.plot.scatter(x='Groesse',y='Laufzeit')

#Kennwerte korr.
df=pd.concat([df1,df2.loc[:,'Einkommen'],df
3.loc[:,'Laufzeit']],axis=1)
m_x=df.mean()[0]
m_y3=df.mean()[3]
s_x=df.std()[0]
s_y3=df.std()[3]
s_xy3=df.cov().values[0,3]
r_xy3=df.corr().values[0,3]
```

| | |
|---|---|
| Mittelw.x | 173.000000 |
| Mittelw.y | 1.613333 |
| St.abw.x | 6.047432 |
| St.abw.y | 0.253170 |
| Kovarianz | -1.078571 |
| Korr.Koeff. | -0.704474 |

Es ergibt sich der Korrelationskoeffizient $r_{xy} = -0.704$. Der Wert ist betragsmässig nahe bei 1, was einen gewissen, aber nicht starken linearen Zusammenhang der beiden Merkmale bedeutet. Das negative Vorzeichen von $r_{xy}$ sagt aus, dass tendenziell
mit zunehmenden $x_i$-Werten die $y_i$-Werte fallen.

.....

**Bemerkung**
Der Korrelationskoeffizient $r_{xy}$ ist so definiert, dass seine Werte immer zwischen –1 und +1, liegen, also $-1 \leq r_{xy} \leq +1$. Je näher $r_{xy}$ bei –1 oder bei 1 liegt, umso besser liegen die Punkte $(x_i, y_i)$ um eine Gerade konzentriert. Dabei bedeutet:

- $r_{xy} > 0$: Die Punkte liegen tendenziell um eine Gerade mit positiver Steigung (gleichsinniger linearer Zusammenhang, positive Korrelation).
- $r_{xy} < 0$: Die Punkte liegen tendenziell um eine Gerade mit negativer Steigung (gegensinniger linearer Zusammenhang, negative Korrelation).
- $r_{xy}$ nahe bei 0: kein linearer Zusammenhang zwischen den beiden Merkmalen.

Die Werte +1 bzw. –1 nimmt $r_{xy}$ an, wenn alle Punkte $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ genau auf einer Geraden mit positiver bzw. negativer Steigung liegen, und zwar ist $r_{xy} = +1$ genau dann, wenn alle Punkte auf einer Geraden mit positiver Steigung liegen und $r_{xy} = -1$, wenn alle Punkte auf einer Geraden mit negativer Steigung liegen.

**Beispiel 4: Empirische Kenngrössen (Teil 1)**
Bestimmen Sie für die Stichproben alle oben definierten Kenngrössen.

$$x = [1,2,3] \text{ und } y = [4,-1,2]$$

Die arithmetischen Mittel lauten:

$$\bar{x} = \frac{1}{3}[1+2+3] = 2$$
$$\bar{y} = \frac{1}{3}[4-1+2] = \frac{5}{3}$$

Die Varianzen lauten:

$$\tilde{s}_x{}^2 = \frac{1}{3}[(1-2)^2 + (2-2)^2 + (3-2)^2] = \frac{2}{3}$$
$$\tilde{s}_y{}^2 = \frac{1}{3}[(4-\frac{5}{3})^2 + (-1-\frac{5}{3})^2 + (2-\frac{5}{3})^2] = \frac{1}{3}[(\frac{7}{3})^2 + (\frac{8}{3})^2 + (\frac{1}{3})^2] = \frac{38}{9}$$

Die Standardabweichungen lauten:

$$\tilde{s}_x = \sqrt{\frac{2}{3}} \qquad \tilde{s}_y = \frac{\sqrt{38}}{3}$$

Hinweis an Studierende, dass wir hier mit den nicht-korrigierten Werten arbeiten

Die Kovarianz lautet:

$$\tilde{s}_{xy} = \frac{1}{3}[(1-2)(4-\frac{5}{3}) + (2-2)(-1-\frac{5}{3}) + (3-2)(2-\frac{5}{3}) = \frac{-2}{3}$$

Der Korrelationskoeffizient lautet:

$$r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \tilde{s}_y} = -\sqrt{\frac{3}{19}}$$

```
#Empirische Kenngrössen
import pandas as pd

df=pd.DataFrame({'x':[1,2,3],'y':[4,-1,2]})

#arithm.Mittel
m_x,m_y=df.mean()

#Varianzen (korr.),(nicht korrigiert ddof=0)
v_x,v_y=df.var()

#Standardabweichungen (korr.)
s_x,s_y=df.std()

#Kovarianz (korr.)
#Es wird eine Matrix erstellt mit den Elementen
#v_x,s_xy,
#s_xy,v_y
s_xy=df.cov().values[0,1]

#Korr.koeffizient
r_xy_1=s_xy/(s_x*s_y)
#Es wird eine Matrix erstellt mit den Elementen
#r_xx,r_xy,
#r_xy,r_yy
r_xy_2=df.corr().values[0,1]
```

Zur systematischen und vereinfachten Berechnung der Kovarianz und des Korrelationskoeffizienten verwendet man gerne die folgenden Regeln**:**

**Definition**

Gegeben seien die Wertepaare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Die (*empirischen*) *Varianzen* der $x_i$ bzw. der $y_i$-Werte lauten

$\tilde{s}_x{}^2 = \overline{x^2} - \bar{x}^2$ und $\tilde{s}_y{}^2 = \overline{y^2} - \bar{y}^2$

Die (*empirische*) *Kovarianz* lautet

$\tilde{s}_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$

Der *Korrelationskoeffizient* (nach Bravais-Pearson) ergibt sich damit zu:

$$r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$$

Mit den folgenden arithmetischen Mittelwerten:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ und } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ und } \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \text{ und } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

**Bemerkung**

Bei der Berechnung von $r_{xy}$ können die Vorfaktoren $\frac{1}{n}$ alle durch $\frac{1}{n-1}$ ersetzt oder alle weggelassen werden.

**Beispiel 4: Empirische Kenngrössen (Teil 2)**

Bestimmen Sie für die Stichproben alle oben definierten Kenngrössen.

$$x = [1,2,3] \text{ und } y = [4,-1,2]$$

Es hilft hierbei mit einer Tabelle zu arbeiten:

| $i$ | 1 | 2 | 3 | Arithmetische Mittel |
|---|---|---|---|---|
| $x_i$ | 1 | 2 | 3 | $\bar{x} = \frac{1}{3}[1+2+3] = 2$ |
| $y_i$ | 4 | –1 | 2 | $\bar{y} = \frac{1}{3}[4-1+2] = \frac{5}{3}$ |
| $x_i^2$ | 1 | 4 | 9 | $\overline{x^2} = \frac{1}{3}(1+4+9) = \frac{14}{3}$ |
| $y_i^2$ | 16 | 1 | 4 | $\overline{y^2} = \frac{1}{3}(16+1+4) = 7$ |
| $x_i y_i$ | 4 | –2 | 6 | $\overline{xy} = \frac{1}{3}(4+(-2)+6) = \frac{8}{3}$ |

Die (empirischen) Varianzen lauten:

$$\tilde{s}_x{}^2 = \overline{x^2} - \bar{x}^2 = \frac{14}{3} - 2^2 = \frac{2}{3}$$

$$\tilde{s}_y{}^2 = \overline{y^2} - \bar{y}^2 = 7 - \left(\frac{5}{3}\right)^2 = \frac{38}{9}$$

Die (*empirische*) *Kovarianz* lautet:

$$\tilde{s}_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = \frac{8}{3} - 2 \cdot \frac{5}{3} = -\frac{2}{3}$$

Der Korrelationskoeffizient lautet:

$$r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \, \tilde{s}_y} = -\sqrt{\frac{3}{19}}$$

**Interpretation der empirischen Grössen**

Die *Kovarianz* und der *Korrelationskoeffizient* sind Masse für den *linearen Zusammenhang* zwischen zwei Merkmalen in der Stichprobe. Je grösser die Kovarianz ist (im Vergleich zu den Varianzen), desto mehr sind die Merkmale voneinander (linear) abhängig, man sagt *positiv korreliert*. In anderen Worten, grosse Werte des einen Merkmals treten dann gehäuft mit grossen Werten des anderen Merkmals auf. Der Korrelationskoeffizient ist in diesem Fall nahe beim Wert 1.

Ist die Kovarianz nahe beim Wert Null, gibt es keinen linearen Zusammenhang zwischen den Merkmalen. Im Streudiagramm der Merkmalspaare sind diese gleichmässig um den Schwerpunkt $(\bar{x}, \bar{y})$ der Punkte $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ verteilt. Der Korrelationskoeffizient ist in diesem Fall nahe bei null und man sagt, die Merkmale sind *unkorreliert*.

Hat die Kovarianz ein negatives Vorzeichen und ist sehr klein, so treten eher grosse Werte des einen Merkmals mit kleinen Werten des anderen Merkmals auf. Der Zusammenhang ist wieder linear, allerdings mit negativem Vorzeichen. Die Merkmale sind damit *negativ korreliert* und der Korrelationskoeffizient ist nahe bei -1.

Am besten sieht man die Bedeutung von Kovarianz und Korrelationskoeffizient im folgenden Beispiel, wo die verschiedenen Fälle anhand von 6 Stichproben dargestellt sind.

**Beispiel 5: Kovarianz und Korrelationskoeffizient**

Gegeben ist die Stichprobe $x = [1,2,3,4,5]$. Dazu werden die folgenden 6 weiteren Stichproben $y$ betrachtet:

| | | |
|---|---|---|
| a) $y = [5,7,9,11,13]$ | b) $y = [5,7,8,11,12]$ | c) $y = [6,3,8,5,6]$ |
| d) $y = [6,3,8,5,4]$ | e) $y = [9,8,7,4,2]$ | f) $y = [10,8,6,4,2]$ |

Die nachfolgenden Streudiagramme zeigen die Stichproben in den jeweiligen Fällen mit den Werten für die empirische Kovarianz $s_{xy}$ und dem empirischen Korrelationskoeffizienten $r_{xy}$.

In den Beispielen a), b) und c) sind die Kovarianzen positiv.
Die Merkmale $x$ und $y$ sind positiv korreliert.

……

In Beispiel a) ist der Korrelationskoeffizient gleich 1. Alle Stichprobenwerte liegen auf einer steigenden Geraden. Sie sind *positiv korreliert*.

In Beispiel b) ist die lineare Abhängigkeit deutlich zu sehen. Die geringfügigen Abweichungen beruhen möglicherweise auf Messungenauigkeiten.

In den Beispielen c) und d) streuen die Werte um den Schwerpunkt der Punktemenge gleichmässig.
Die Merkmale sind nahezu unkorreliert..

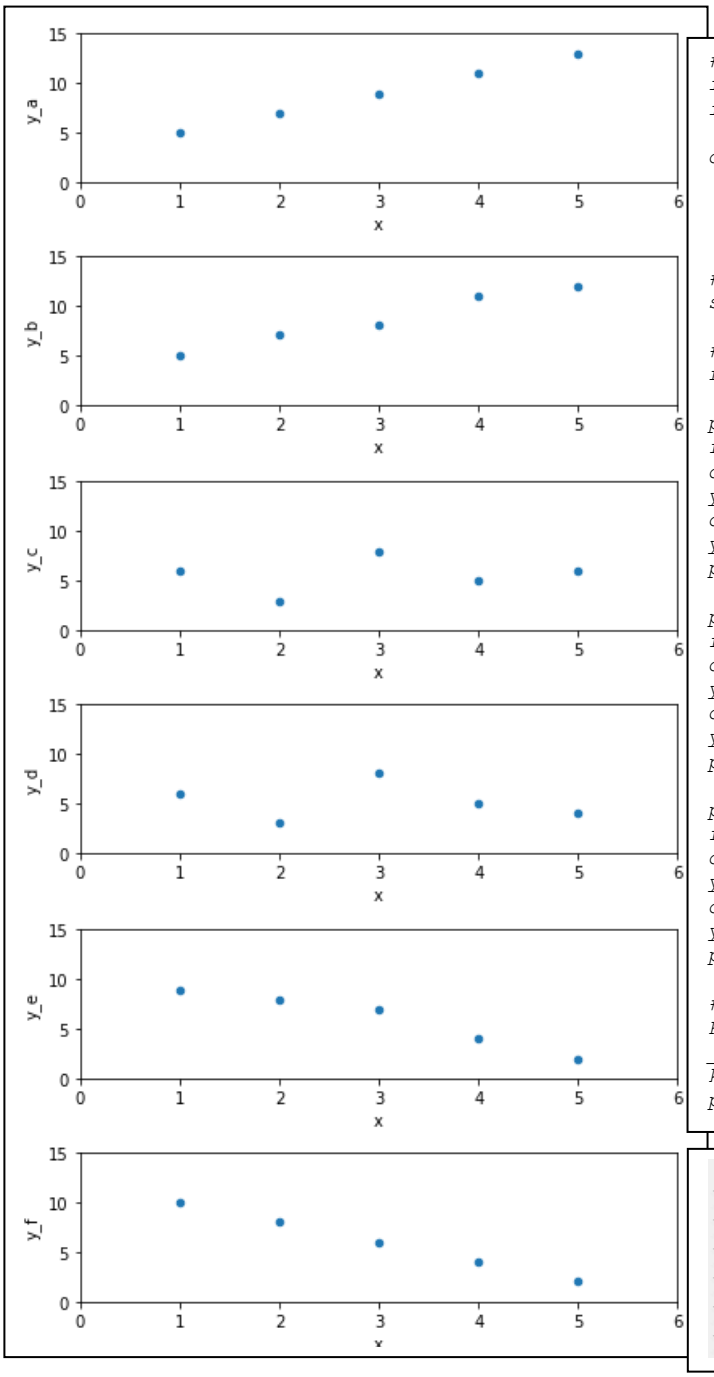
……

In den Beispielen d), e), f) sind die Kovarianzen negativ.
Die Merkmale $x$ und $y$ sind negativ korreliert.

……

In Beispiel f) ist der Korrelationskoeffizient -1. Die Merkmalspaare liegen auf einer fallenden Geraden. Sie sind *negativ korreliert*. Grosse Werte von $x$ treten gehäuft mit kleinen Werten von $y$ auf.

In Beispiel e) ist die lineare Abhängigkeit deutlich zu sehen. Die geringfügigen Abweichungen beruhen auch hier möglicherweise auf Messungenauigkeiten.




```
#Kovarianz und KorrKoeff
import pandas as pd
import matplotlib.pyplot as plt

df=pd.DataFrame({'x':[1,2,3,4,5],
'y_a':[5,7,9,11,13],'y_b':[5,7,8,11,12],
'y_c':[6,3,8,5,6],'y_d':[6,3,8,5,4],
'y_e':[9,8,7,4,2],'y_f':[10,8,6,4,2]})

#Kovarianz (korr.)
s_xy=df.cov()

#Korr.koeffizient
r_xy=df.corr()

plt.figure(1)
fig, axs = plt.subplots(2)
df.plot.scatter(x='x',
y='y_a',ax=axs[0],xlim=[0,6],ylim=[0,15])
df.plot.scatter(x='x',
y='y_b',ax=axs[1],xlim=[0,6],ylim=[0,15])
plt.tight_layout()

plt.figure(2)
fig, axs = plt.subplots(2)
df.plot.scatter(x='x',
y='y_c',ax=axs[0],xlim=[0,6],ylim=[0,15])
df.plot.scatter(x='x',
y='y_d',ax=axs[1],xlim=[0,6],ylim=[0,15])
plt.tight_layout()

plt.figure(3)
fig, axs = plt.subplots(2)
df.plot.scatter(x='x',
y='y_e',ax=axs[0],xlim=[0,6],ylim=[0,15])
df.plot.scatter(x='x',
y='y_f',ax=axs[1],xlim=[0,6],ylim=[0,15])
plt.tight_layout()

#Kennwerte
KW=pd.concat([s_xy.loc['y_a':'y_f','x'],r
_xy.loc['y_a':'y_f','x']],axis=1,
keys=['s_xy', 'r_xy'])
print(KW)
```

| | s_xy | r_xy |
|---|---|---|
| y_a | 5.0 | 1.000000 |
| y_b | 4.5 | 0.987878 |
| y_c | 0.5 | 0.174078 |
| y_d | -0.5 | -0.164399 |
| y_e | -4.5 | -0.976187 |
| y_f | -5.0 | -1.000000 |

**Bemerkung:**
Der Pearson-Korrelationskoeffizient ist nicht robust. Bereits durch einen einzelnen Ausreisser kann der Kennwert vor allem bei kleinerer Stichprobengrösse sehr stark beeinflusst werden.

Der Korrelationskoeffizient misst nur den Grad der linearen Abhängigkeit. Wenn $r_{xy} = 0$, so kann trotzdem ein anderer Zusammenhang zwischen den Wertepaaren bestehen.
Betrachten wir zum Beispiel $(-2{,}5), (-1{,}2), (0{,}1), (1{,}2), (2{,}5)$. Obwohl die Wertepaare die Beziehung $y = x^2 + 1$ erfüllen, ist $r\_xy = 0$.
Der Zusammenhang ist allerdings nichtlinear.

.....

### *Spearman-Rangkorrelationskoeffizient*

Die *Spearman-Rangkorrelation* misst Stärke und Richtung des *streng monotonen Zusammenhangs* zwischen zwei Merkmalen $x$ und $y$, d.h. wie eng die Punkte um eine Kurve streuen, welche durch eine beliebige, monotone Funktion definiert ist.
Bei der Spearman-Korrelation wird nicht davon ausgegangen, dass die Daten aus einer bestimmten Verteilung stammen, es handelt sich um ein sogenanntes nichtparametrisches Korrelationsmass.

**Beispiel 6: Vergleich Korrelationskoeffizienten**
Wir führen für die Daten aus Beispiel 2 einen Vergleich der beiden Korrelationskoeffizienten durch.

```
#Verschiedene Scatterplots
import pandas as pd

#Scatterplot gleichsinniger Zusammenhang
df1=pd.DataFrame({'Groesse':[163,165,166,169,170,171,171,173,174,175,177,177,179,180,185],
'Gewicht':[59,62,65,69,65,69,76,73,75,73,80,71,82,84,81]})

#Scatterplot ohne Zusammenhang
df2=pd.DataFrame({'Groesse':[163,165,166,169,170,171,171,173,174,175,177,177,179,180,185],
'Einkommen':[2900,1100,3600,2300,4000,5600,2100,5100,3400,1800,2100,2600,4600,3600,2300]})

#Scatterplot gegensinniger Zusammenhang
df3=pd.DataFrame({'Groesse':[163,165,166,169,170,171,171,173,174,175,177,177,179,180,185],
'Laufzeit':[2.0,1.8,1.7,1.9,1.8,1.5,1.9,1.3,1.7,1.5,1.4,1.7,1.1,1.5,1.4]})

df=pd.concat([df1,df2.loc[:,'Einkommen'],df3.loc[:,'Laufzeit']],axis=1)
```

| | Groesse | Gewicht | Einkommen | Laufzeit |
|---|---|---|---|---|
| Groesse | 1.000000 | 0.897914 | 0.060641 | -0.704474 |
| Gewicht | 0.897914 | 1.000000 | 0.084040 | -0.692065 |
| Einkommen | 0.060641 | 0.084040 | 1.000000 | -0.434115 |
| Laufzeit | -0.704474 | -0.692065 | -0.434115 | 1.000000 |

```
#Korrelationskoeffizient Pearson
pandas-Funktion
r_xy_p=df.corr()
print(r_xy_p)
```

| | Groesse | Gewicht | Einkommen | Laufzeit |
|---|---|---|---|---|
| Groesse | 1.000000 | 0.899552 | 0.057399 | -0.746185 |
| Gewicht | 0.899552 | 1.000000 | 0.028725 | -0.658245 |
| Einkommen | 0.057399 | 0.028725 | 1.000000 | -0.336356 |
| Laufzeit | -0.746185 | -0.658245 | -0.336356 | 1.000000 |

```
#Korrelationskoeffizient Spearman
pandas-Funktion
r_xy_s=df.corr(method='spearman')
print(r_xy_s)
```

| | Pearson | Spearman |
|---|---|---|
| Gewicht | 0.897914 | 0.899552 |
| Einkommen | 0.060641 | 0.057399 |
| Laufzeit | -0.704474 | -0.746185 |

```
#Gegenüberstellung der Korrelationen vs. Grösse
r_xy=pd.concat([r_xy_p.loc['Gewicht':'Laufzeit','Groesse'
],r_xy_s.loc['Gewicht':'Laufzeit','Groesse']],keys=['Pear
son','Spearman'],axis=1)
print(r_xy)
```

In diesem Beispiel liegen die Werte der beiden Korrelationskoeffizienten sehr nah beeinander, da es entweder keinen monotonen Zusammenhang
……
oder einen annähernd linearen Zusammenhang gibt.
……

Mit der Spearman-Korrelation misst man also wie mit der Pearson-Korrelation einen bestimmten Zusammenhang zwischen zwei Merkmalen. Die Spearman-Korrelation nimmt als ein Spezialfall des Pearson Koeffizienten ebenfalls Werte von -1 (perfekte negative Korrelation) bis +1 (perfekte positive Korrelation) an; und ist nahe bei 0, falls gar keine strenge Monotonie vorliegt.

Der Spearman-Korrelationskoeffizient $r_{Sp}$ wird auch Rangkorrelationskoeffizient genannt, weil er im Gegensatz zum klassischen Pearson-Korrelationskoeffizienten die Korrelation nicht zwischen den Datenpunkten selbst,
sondern zwischen ihren Rängen berechnet.

.....

**Beispiel 7: Berechnung Spearman-Korrelation**
Wir betrachten den Zusammenhang zwischen dem Alter einer Person und ihrer Performance beim 100-Meter-Lauf. Von 6 Personen bestimmen wir das Alter in Jahren, und die Zeit für 100 Meter in Sekunden

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| $x_i$ | 59 | 35 | 43 | 23 | 42 | 27 | |
| $y_i$ | 14.6 | 11.8 | 14.3 | 13.0 | 14.2 | 11.0 | |
| $\mathrm{rg}(x_i)$ | 6 | 3 | 5 | 1 | 4 | 2 | $\overline{\mathrm{rg}(x)} = 3.5$ |
| $\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}$ | 2.5 | -0.5 | 1.5 | -2.5 | 0.5 | -1.5 | |
| $\mathrm{rg}(y_i)$ | 6 | 2 | 5 | 3 | 4 | 1 | $\overline{\mathrm{rg}(y)} = 3.5$ |
| $\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}$ | 2.5 | -1.5 | 1.5 | -0.5 | 0.5 | -2.5 | |

Dabei wird dieser Koeffizient berechnet wie die Pearson-Korrelation mit dem Unterschied, dass die Ränge statt der Originaldaten verwendet werden.

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

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

$$= \sqrt{2.5^2 + (-0.5)^2 + 1.5^2 + (-2.5)^2 + 0.5^2 + (-1.5)^2} = \sqrt{17.5}$$

$$\sum_{i=1}^{n}\left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}\right)$$

$$= 2.\dot{5} \cdot 2.5 + (-0.5) \cdot (-1.5) + 1.5 \cdot 1.5 + (-2.5) \cdot (-0.5) + 0.5 \cdot 0.5 + (-1.5) \cdot (-2.5)$$

$$= 14.5$$

$$r_{Sp} = \frac{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}\right)\left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(x_i) - \overline{\mathrm{rg}(x)}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(\mathrm{rg}(y_i) - \overline{\mathrm{rg}(y)}\right)^2}} = \frac{14.5}{\sqrt{17.5} \cdot \sqrt{17.5}} \approx 0.83$$

Als Interpretation könnte man nun vermuten, dass mit steigendem Rang des Alters auch der Rang des Platzes ansteigt.
Da bei der Spearman-Korrelation die Ränge verwendet werden, sind also dort die tatsächlichen Abstände zwischen z.B. Platz 1 und Platz 2 egal. Die Spearman-Korrelation ist immer dann 1, wenn der niedrigste Wert für $x$ gepaart ist mit dem niedrigsten Wert von $y$, etc. und dabei streng monoton steigend.

Es kann passieren, dass z.B. zwei oder mehr Werte für $x$ denselben Wert annehmen. In diesem Fall wird diesen *verbundenen Rängen* der Durchschnittsrang zugewiesen. Hierzu zwei Beispiele

**Beispiel 8: Verbundene Ränge**
Eine andere Personengruppe hätte evtl. ein anderes Alter gehabt:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $x_i$ | 23 | 27 | 35 | 35 | 42 | 59 |
| $\mathrm{rg}(x_i)$ | 1 | 2 | (3+4)/2 =3.5 | (3+4)/2 =3.5 | 5 | 6 |

Bzw.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $x_i$ | 23 | 27 | 35 | 35 | 35 | 59 |
| $\mathrm{rg}(x_i)$ | 1 | 2 | (3+4+5) /3=4 | (3+4+5) /3=4 | (3+4+5) /3=4 | 6 |

Dies ist ein aufwändiges Berechnungsverfahren. Da immer Ränge zwischen 1 und der Stichprobenlänge vorkommen, gibt es weniger rechenaufwändige Formeln. Zum Beispiel gilt, wenn die Ränge alle verschieden sind:

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

### 2.1.3 Grenzen der Korrelation

Aber auch wenn zwischen zwei Grössen eine Korrelation besteht (d.h. $r$ bzw. $r_{Sp}$ ist signifikant von Null verschieden), so muss das noch lange nicht einen *kausalen Zusammenhang* bedeuten. Man spricht in einem solchen Fall von einer Scheinkorrelation.

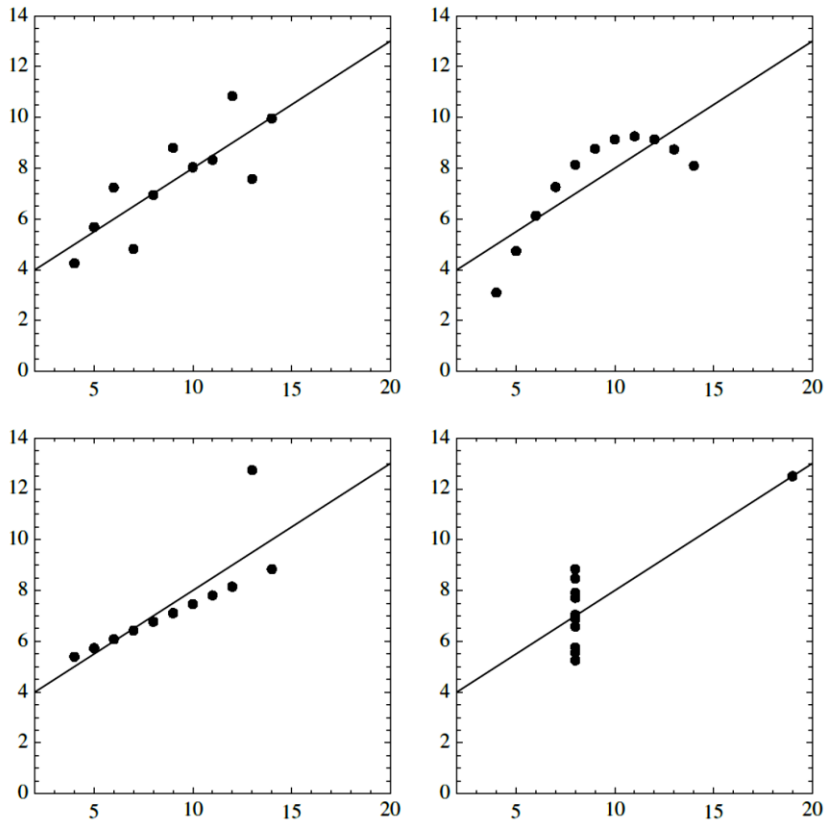
……

Das bedeutet, dass man eine Korrelation zwischen zwei Merkmalen beobachtet, die aber inhaltlich nicht sinnvoll ist.
So ein scheinbarer Zusammenhang kann z.B. dadurch entstehen,
dass ein drittes Merkmal übersehen wird,
……
das aber mit den beiden anderen Merkmalen inhaltlich im Zusammenhang steht.

Das bekannteste Beispiel ist das von Störchen und Babys:
In einer Studie wurde für verschiedene Regionen untersucht, wie viele Störche dort zu Hause sind und wie hoch die Geburtenrate ist. Es zeigte sich eine signifikante positive Korrelation zwischen der Anzahl der Störche und der Anzahl der Babys. Das heisst, je mehr Störche eine Region hat, umso mehr Babys gibt es dort. Heisst das nun, dass der Storch die Babys bringt?

Nein, in diesem Fall ist eher ein drittes Merkmal für diesen scheinbaren Zusammenhang verantwortlich. Dieses dritte Merkmal ist die Industrialisierung. Die Regionen, die stark industrialisiert sind, haben weniger Störche, da es sich um eher städtische Regionen handelt. Aus dem gleichen Grund haben sie auch weniger Babys, da Familien sich eher im ländlichen Bereich ansiedeln.

Zum Ende dieses Abschnitts folgen warnende Beispiele des englischen Statistikers Francis Anscombe, die in der folgenden Abbildung dargestellt sind.




Alle Datensätze haben den gleichen Korrelationskoeffizienten von 0.816, aber nur beim ersten Datensatz ist ein linearer Zusammenhang erkennbar und das Berechnen der Regressionsgerade sinnvoll (wie man eine solche Gerade bestimmt, betrachten wir später, s. Abschnitt 6.2). Beim zweiten liegt offensichtlich ein nichtlinearer Zusammenhang vor, beim dritten führt ein

Ausreisser zu einer verfälschten Regressionsgeraden, und beim letzten führt wiederum ein Ausreisser zu einem zu hohen Korrelationskoeffizienten.
Diese Beispiele zeigen, wie wichtig es ist,
Daten zu visualisieren und
……
keine voreiligen Schlussfolgerungen zu ziehen.
……

## 2.2 Mehrere Merkmale

In manchen Fällen soll die Abhängigkeit nicht allein von einer, sondern von mehreren Merkmalen betrachtet werden. Hier ist das Vorgehen vergleichbar (dann sinnvollerweise mit Rechnerunterstützung zu lösen).

### *Kategorielle Merkmale*

Für rein kategorielle Merkmale und/oder eine Kombination von kategoriellen und metrischen Merkmalen können dieselben Diagrammen verwendet werden. Die dritte Dimension kann in diesen Fällen z.B. in Form einer farblichen Codierung eingebracht werden.

### *Nur metrische Merkmale*

Bei rein metrischen Merkmalen kann man die Darstellungsformen für bivariate Daten erweitern und nicht nur Einzelwerte oder Plots, sondern Matrizen dieser Darstellungen nutzen.

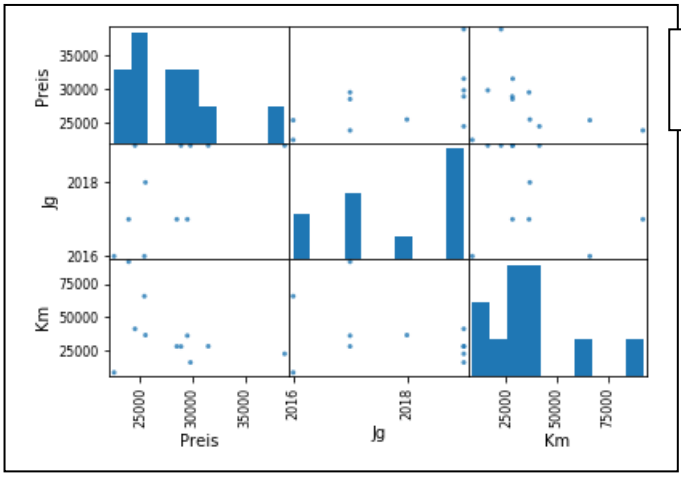
**Beispiel 9: Occasionsfahrzeuge**
Der «Preis» eines Occasionsautos soll durch den Jahrgang «Jg» und die abgelaufenen «Km» modelliert werden.

| Preis | Jahrgang | Kilometerstand |
|---|---|---|
| 28500 | 2017 | 28100 |
| 31500 | 2019 | 28200 |
| 28900 | 2019 | 28000 |
| 29500 | 2017 | 36200 |
| 24500 | 2019 | 41300 |
| 25500 | 2018 | 36600 |
| 38800 | 2019 | 22500 |
| 25400 | 2016 | 66000 |
| 23900 | 2017 | 92000 |
| 29800 | 2019 | 16000 |
| 22500 | 2016 | 84000 |

```
Occasionsfahrzeuge
import pandas as pd
import matplotlib.pyplot as plt

df=pd.DataFrame({'Preis':[28500,31500,28900,29500
,24500,25500,38800,25400,23900,29800,22500],
'Jg':[2017,2019,2019,2017,2019,2018,2019,2016,201
7,2019,2016],
'Km':[28100,28200,28000,36200,41300,36600,22500,6
6000,92000,16000,8400],})
```

Ein dreidimensionales Streudiagramm ist jetzt schwieriger zu zeichnen und bietet eine schlechtere Übersicht.

Besser nutzt man eine Matrix von Streudiagrammen.




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

Zusätzlich kann ergänzend die *Korrelationsmatrix* genutzt werden. Sie ist so aufgebaut, dass für jedes Paar von Merkmalen der Pearson-Korrelationskoeffizient ausgegeben wird.

| | Preis | Jg | Km |
|---|---|---|---|
| Preis | 1.000000 | 0.560559 | -0.367787 |
| Jg | 0.560559 | 1.000000 | -0.323916 |
| Km | -0.367787 | -0.323916 | 1.000000 |

```
#Korr.koeffizient
r_xy=df.corr()
```

Zudem ist wie bereits erklärt zu beachten, dass Korrelationskoeffizienten nie ohne das Prüfen der zugehörigen Scatterplots angegeben werden sollten. Die Korrelationsmatrix ist somit als Ergänzung zur Scatterplot-Matrix zu sehen. Zudem gilt auch hier weiterhin, dass ein betragsmässig hoher Wert des Korrelationskoeffizienten noch nicht unmittelbar eine Kausalität zwischen den Merkmalen beschreibt.

## 2.3 Lernziele für dieses Kapitel

- ☐ Ich kann die Begriffe multivariate und bivariate Datenmenge erklären.
- ☐ Ich kann eine multivariate Datenmenge in bivariate Daten zerlegen und diese visualisieren.
- ☐ Ich kenne die Bedeutung von Scatterplots und kann diese für kleine Datenmengen per Hand anfertigen.
- ☐ Ich kann die Kovarianz von bivariaten Stichproben berechnen und als Mass für den linearen Zusammenhang der Merkmale richtig interpretieren.
- ☐ Ich kann den Pearson Korrelationskoeffizienten von bivariaten Stichproben berechnen und als Mass für den linearen Zusammenhang der Merkmale richtig interpretieren.
- ☐ Ich kann die Berechnung von Varianz, Kovarianz und Korrelationskoeffizient per Hand mit Hilfe einer geeigneten Tabelle, wie in Beispiel 4, an kleinen Stichproben durchführen.
- ☐ Ich kenne den Unterschied zwischen der empirischen Varianz und der empirischen korrigierten Varianz und kann die eine in die andere umrechnen.
- ☐ Ich kenne den Unterschied zwischen der empirischen Standardabweichung und der empirischen korrigierten Standardabweichung und kann die eine in die andere umrechnen.
- ☐ Ich kann den Spearman Korrelationskoeffizienten von bivariaten Stichproben berechnen und als Mass für den monotonen Zusammenhang der Merkmale richtig interpretieren.
- ☐ Ich kann den Unterschied zwischen Pearson und Spearman Korrelationskoeffizient erklären.