

Bericht L02 Projekt Datenanalyse

Gruppe 03, Franck Borel Kana Zambou, Andreas Naujok

Veranstaltung: Grundlagen und Anwendung der

Wahrscheinlichkeitstheorie

Wintersemester 2023/2024

## Datensatz 1

### R1.1

Der Datensatz liegt im .csv-Format vor (Trennzeichen','; Dezimaltrennzeichen'.';UTF-8 Kodierung). Jedes Element des Datensatzes beinhaltet eine Jahreszahl zwischen 1991 und 2021, alle Jahreszahlen in diesem Intervall sind genau einmal vorhanden. Jedes Element beinhaltet einen Verbraucherpreisindex. Das Jahr 2015 stellt den Referenzwert dar und entspricht dem Wert 100. Die Daten stammen vom statistischen Bundesamt (Destatis). Stand der Daten: 10.10.2022 / 10:26:07.

### R1.2

Jahreszahlen: Intervallskala

Verbraucherpreisindex: Intervallskala

### R1.3

Verwendete Software/Funktionen: Python, Jupyter, Pandas

Funktionen: pandas.read\_csv, pandas.DataFrame.sort\_values, pandas.DataFrame.mode, pandas.DataFrame.to\_csv, pandas.DataFrame.mean, pandas.DataFrame.median

### R1.4

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigt.csv"
DataZiel1="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1UrlisteJahre.csv"
DataZiel2="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1UrlisteVerbraucherpreisindex.csv"
df = pandas.read_csv(Data1)
df.to_csv(path_or_buf = DataZiel1, columns=['Jahr'], index = False)
df.to_csv(path_or_buf = DataZiel2, columns=['Verbraucherpreisindex'], index = False)
```

### R1.5

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigt.csv"
DataZiel1="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1RanglisteJahre.csv"
```

```
DataZiel2="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1RanglisteVerbraucherpreisindex.csv"
df = pandas.read_csv(Data1)
df.sort_values(by = ["Jahr"])
df.to_csv(path_or_buf = DataZiel1, columns=['Jahr'], index = False)
df.sort_values(by=["Verbraucherpreisindex"])
df.to_csv(path_or_buf = DataZiel2, columns=['Verbraucherpreisindex'], index = False)
```

R1.7

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigt.csv"
df = pandas.read_csv(Data1)
df.median()
df.mean()
df.mode()
```

Kein Modus vorhanden, da keine Dopplungen.

Arithmetischer Mittelwert: Jahr: 2006; Verbraucherpreisindex: 88.251613

Median: Jahr: 2006, Verbraucherpreisindex: 87.6

R1.8

Spannweite der Variable Verbraucherpreisindex:  $109,1 - 65,5 = 43,6$

Spannweite der Variable Jahr:  $2021 - 1991 = 30$

R1.9

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-
1/data-1bereinigt.csv"
df = pandas.read_csv(Data1)
abs(df - df.median()).mean()
```

Mittlere Abweichung vom Median der Variable Verbraucherpreisindex: 10,297

Mittlere Abweichung vom Median der Variable Jahr: 7.741935

R1.10

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-
1/data-1bereinigt.csv"
df = pandas.read_csv(Data1)
df.var()
```

Stichprobenvarianz Verbraucherpreisindex: 144.859914  
Stichprobenvarianz Jahr: 82.666667

R1.11

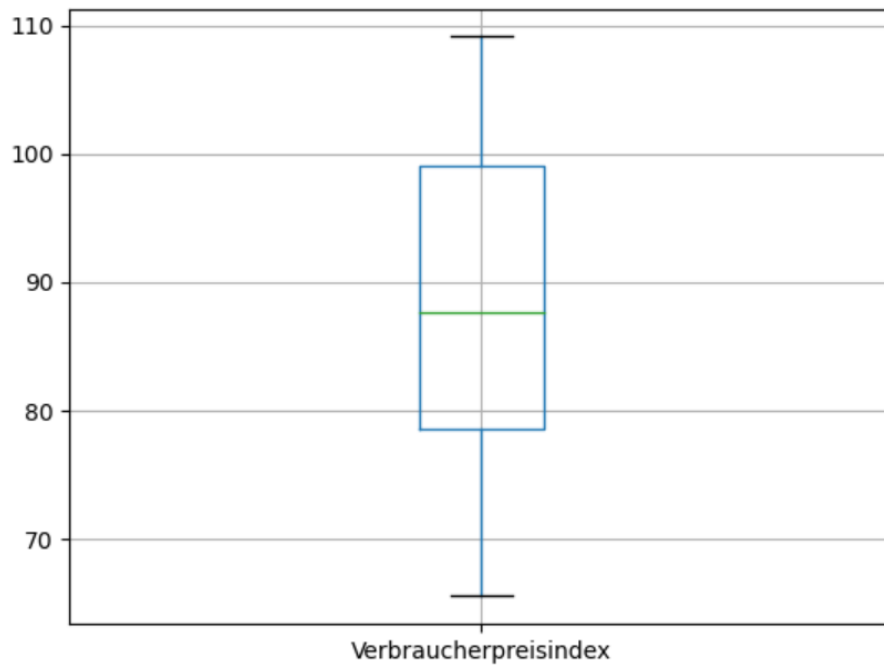
```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-
1/data-1bereinigt.csv"
df = pandas.read_csv(Data1)
df.std()/df.mean()
```

Variationskoeffizient Verbraucherpreisindex: 0.136380

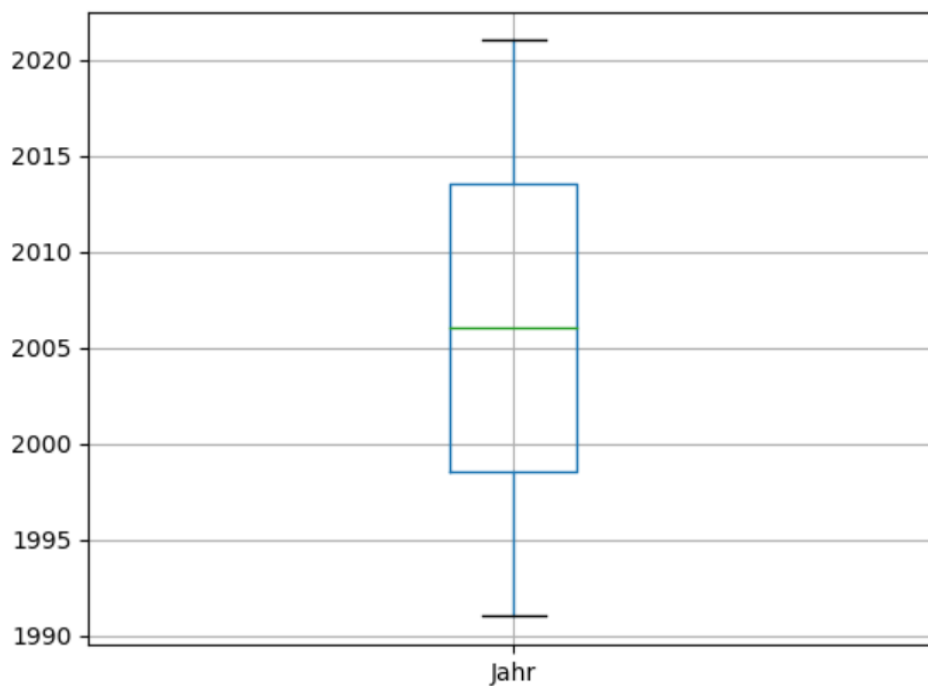
Variationskoeffizient Jahr: 0.004532

R1.12

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-
1/data-1bereinigt.csv"
df = pandas.read_csv(Data1)
df.boxplot(column="Verbraucherpreisindex")
plt.show()
```



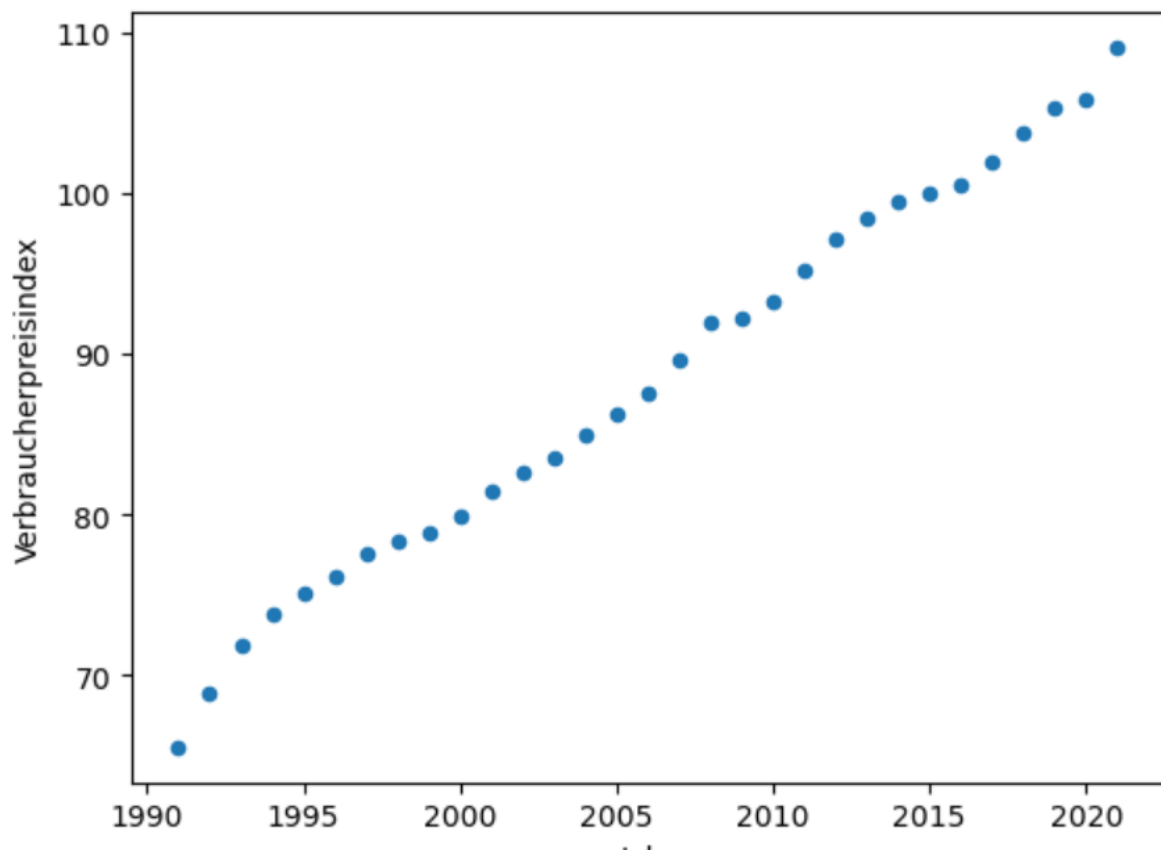
```
df.boxplot(column="Jahr")
plt.show()
```





R1.13

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigt.csv"
df = pandas.read_csv(Data1)
df.plot.scatter(x='Jahr', y='Verbraucherpreisindex')
plt.show()
```



R1.14

Vor der Bearbeitung lag der Datensatz im .csv-Format vor, Trennzeichen wurden angepasst, sowie jeweilige Rang- und Urlisten erstellt. Im Datensatz sind Werte für den Verbraucherpreisindex der Jahre 1991 bis 2021 aufgelistet. Der Verbraucherpreisindex des Jahres 2015 stellt den Referenzwert von 100 dar. Die Daten stammen vom Statistischen Bundesamt. Die Variablen 'Jahr' und 'Verbraucherpreisindex' sind dem Skalenniveau Intervallskala zuzuordnen. Es existiert kein Modus, da bei beiden Variablen die Werte jeweils nur einfach vorkommen. Der Arithmetische Mittelwert der Variable Jahr ist 2006, der der Variable Verbraucherpreisindex ist 88,25. Der Median der Variable Jahr ist ebenfalls 2006, bei der Variable Verbraucherpreisindex ist 87,6 der Wert des Median. Am Scatterplot erkennt man, dass der Verbraucherpreisindex relativ gleichmäßig steigt, trotzdem sind Zeiträume unterscheidbar, in denen der Verbraucherpreisindex mal mehr und mal weniger steigt.

R1.15

```
import pandas
```



```

import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigt.csv"
df = pandas.read_csv(Data1)
df.quantile(q=[0.25, 0.5, 0.75], axis=0, numeric_only=False, interpolation='nearest',
method='single')

```

Quartil	Jahr	Verbraucherpreisindex
0.25	1999	78.8
0.50	2006	87.6
0.75	2013	98.5

```

df.quantile(q=[.1, .2, .3, .4, .5, .6, .7, .8, .9 ], axis=0, numeric_only=False,
interpolation='nearest', method='single')

```

Dezil	Jahr	Verbraucherpreisindex
0.1	1994	73.8
0.2	1997	77.6
0.3	2000	79.9
0.4	2003	83.5
0.5	2006	87.6
0.6	2009	92.2
0.7	2012	97.1
0.8	2015	100.0
0.9	2018	103.8

R1.16

Quartilsabstand:

Jahr: 14;

Verbraucherpreisindex: 19.7

R1.17

```

import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigt.csv"
df = pandas.read_csv(Data1)
df.cov()

```

Kovarianz der Variablen Jahr und Verbraucherpreisindex: 109.093333

R1.18

```

import pandas
import matplotlib.pyplot as plt

```

```
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02  
Projekt/gawtl02.dataset.d03-main/gawtl02.dataset.d03-main/Dataset-1/data-  
1bereinigt.csv"  
df = pandas.read_csv(Data1)  
df.corr()
```

Korrelationskoeffizient der Variablen Jahr und Verbraucherpreisindex: 0.996917

R1.19

Die Variable Verbraucherpreisindex wurde in 4 etwa gleich große Intervalle aufgeteilt:

1= {65<x<76}

2= {76<=x<87}

3={87<=x<98}

4={98<=x<110}

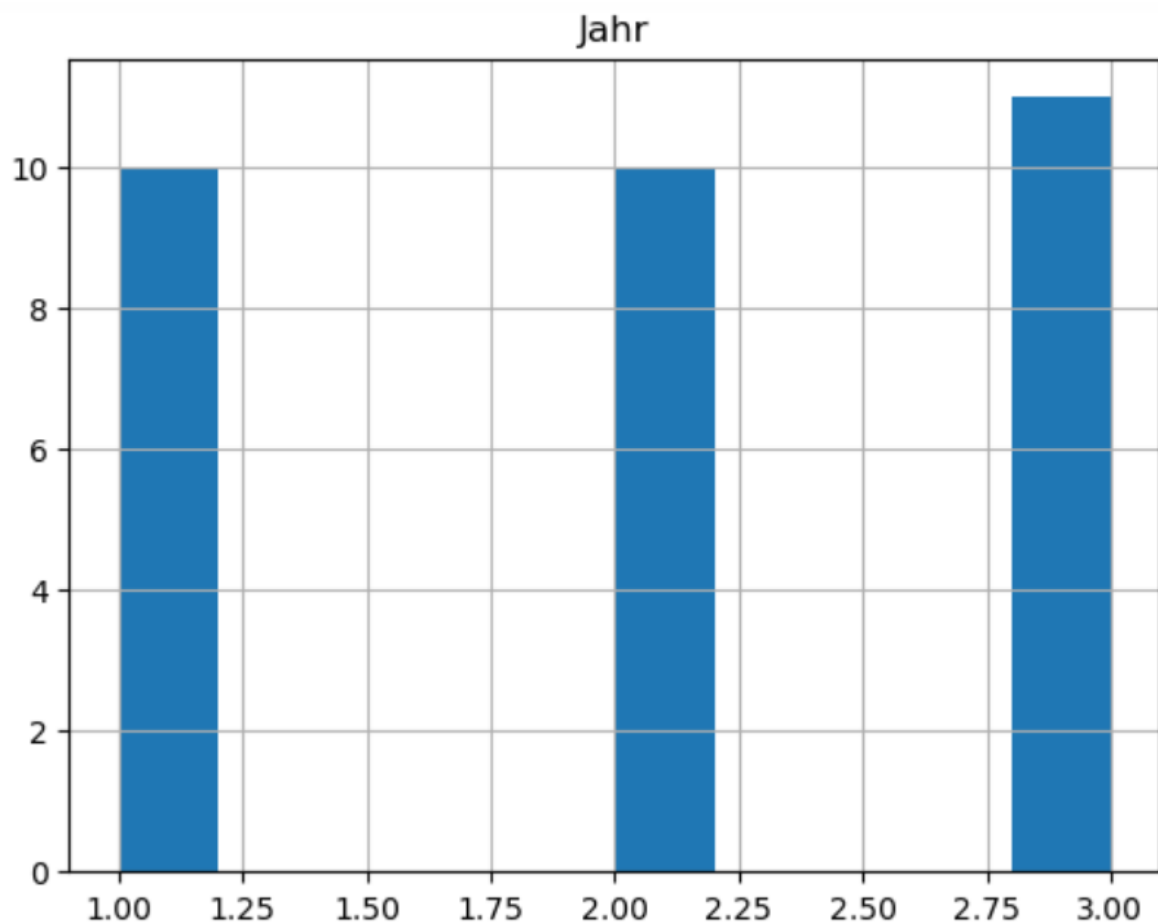
Die Variable Jahr wurde in 3 etwa gleich große Intervalle aufgeteilt:

1=[1991<x<2000]

2=[2001<x<2010]

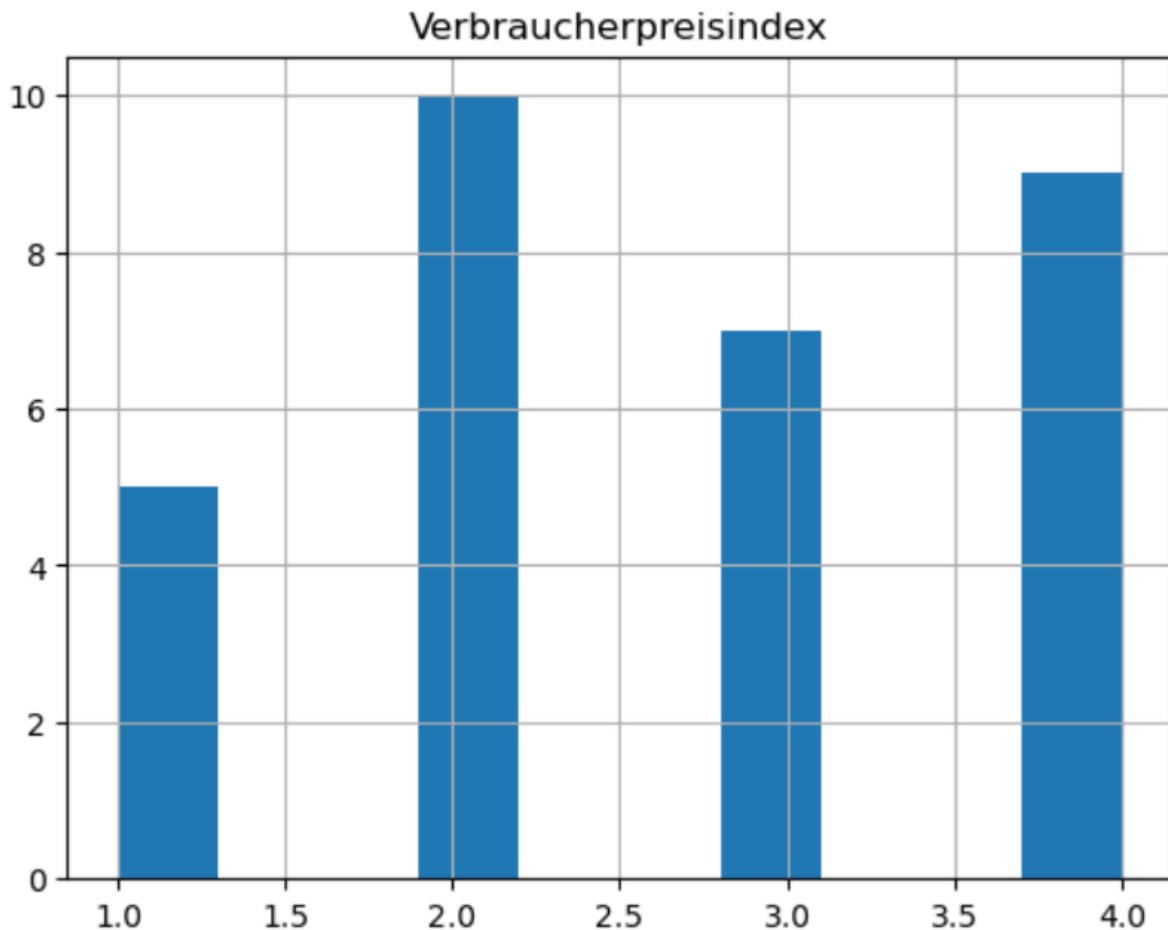
3=[2011<x<2021]

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-1/data-
1bereinigtKlassifiziertR1_19.csv"
df = pandas.read_csv(Data1)
df.hist(column='Jahr')
plt.show()
```



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

```
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawtl02.dataset.d03-main/gawtl02.dataset.d03-main/Dataset-1/data-
1bereinigtKlassifiziertR1_19.csv"
df = pandas.read_csv(Data1)
df.hist(column='Verbraucherpreisindex')
plt.show()
```



## R2.1

Der Datensatz liegt im .csv-Format vor (Trennzeichen','; Dezimaltrennzeichen'.';UTF-8 Kodierung). Jedes Element des Datensatzes beinhaltet eine Jahreszahl zwischen 1991 und 2021, alle Jahreszahlen in diesem Intervall sind genau einmal vorhanden. Die meisten Elemente beinhalten einen Verbraucherpreisindex, bei manchen Datenpunkten fehlt der Verbraucherpreisindex oder ist nicht plausibel (keine Zahl oder NaN). Das Jahr 2015 stellt den Referenzwert dar und entspricht dem Wert 100. Die Daten stammen vom statistischen Bundesamt (Destatis). Stand der Daten: 10.10.2022 / 10:26:07.

## R2.3

Es existieren zwei Arten fehlerhafter Datenpunkte: Fehlender Wert der Variable Verbraucherpreisindex, nicht plausibler Wert der Variable Verbraucherpreisindex (keine Zahl, Ausreißer in Jahreszahl). Die fehlerhaften Datenpunkte wurden beseitigt durch Löschen. Der Ausreißer in Jahreszahlen wurde angepasst: 1795;2105.

Folgende Datenpunkte wurden gelöscht:

1991,  
1994,Peter  
1997,  
2000,  
2002,  
2008,Hans  
2014,NaN  
2020,

R2.4

Verwendete Software/Funktionen: Python, Jupyter, Pandas.

R2.5

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
DataZiel="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.xlsx"
df = pandas.read_csv(Data1)
df.to_excel(DataZiel)
```

R2.6

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
DataZiel1="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2UrlisteJahre.csv"
DataZiel2="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2UrlisteVerbraucherpreisindex.csv"
df = pandas.read_csv(Data1)
df.to_csv(path_or_buf = DataZiel1, columns=['Jahr'], index = False)
df.to_csv(path_or_buf = DataZiel2, columns=['Verbraucherpreisindex'], index = False)
```

R2.7

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
```

```

Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
DataZiel1="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2RanglisteJahre.csv"
DataZiel2="C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2RanglisteVerbraucherpreisindex.csv"
df = pandas.read_csv(Data1)
df.sort_values(by = ["Jahr"])
df.to_csv(path_or_buf = DataZiel1, columns=['Jahr'], index = False)
df.sort_values(by=["Verbraucherpreisindex"])
df.to_csv(path_or_buf = DataZiel2, columns=['Verbraucherpreisindex'], index = False)

```

R2.8

```

import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.median()
df.mean()
df.mode()

```

Kein Modus vorhanden, da keine Dopplungen in den Werten.

Arithmetischer Mittelwert: Jahr: 2006.956522; Verbraucherpreisindex: 89.530435

Median: Jahr: 2007.0; Verbraucherpreisindex: 89.6

R2.9

Spannweite Variable Jahr: 2021-1992=29

Spannweite Variable Verbraucherpreisindex: 109.1 - 68.8 = 40.3

R2.10

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
abs(df - df.median()).mean()
```

Mittlere Abweichung vom Median der Variable Verbraucherpreisindex: 9.747826

Mittlere Abweichung vom Median der Variable Jahr: 7.347826

R2.11

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.var()
```

Stichprobenvarianz Jahr: 76.952569

Stichprobenvarianz Verbraucherpreisindex: 132.671304

R2.12

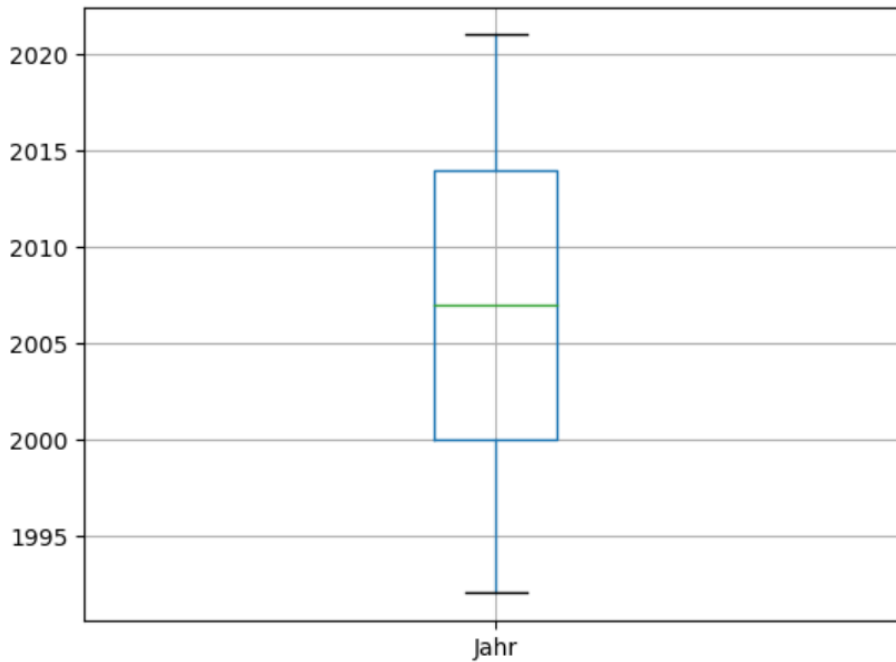
```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.std()/df.mean()
```

Variationskoeffizient Jahr: 0.004371

Variationskoeffizient Verbraucherpreisindex: 0.128652

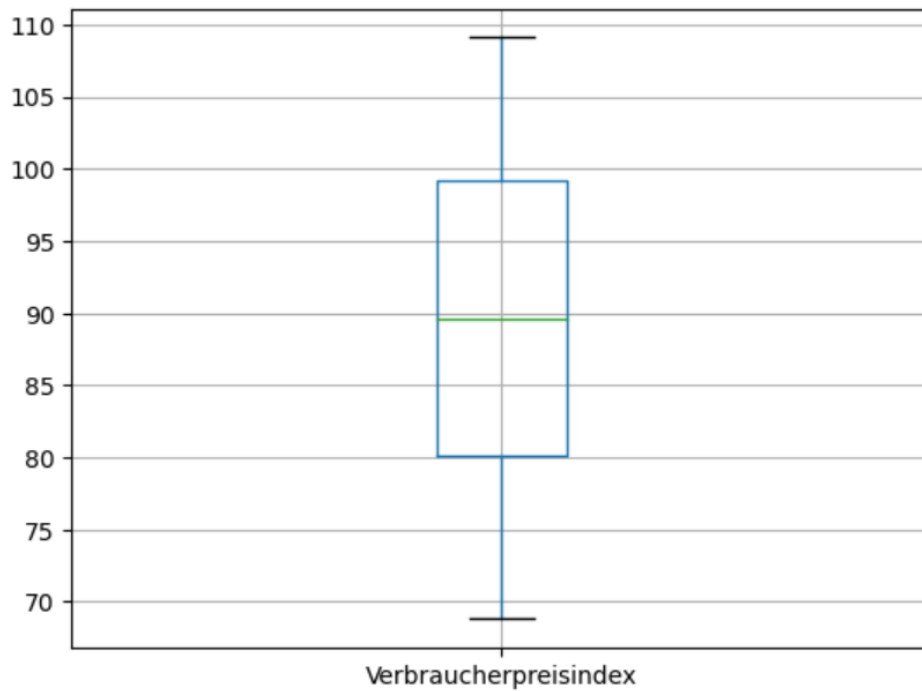
R2.13

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.boxplot(column="Jahr")
plt.show()
```



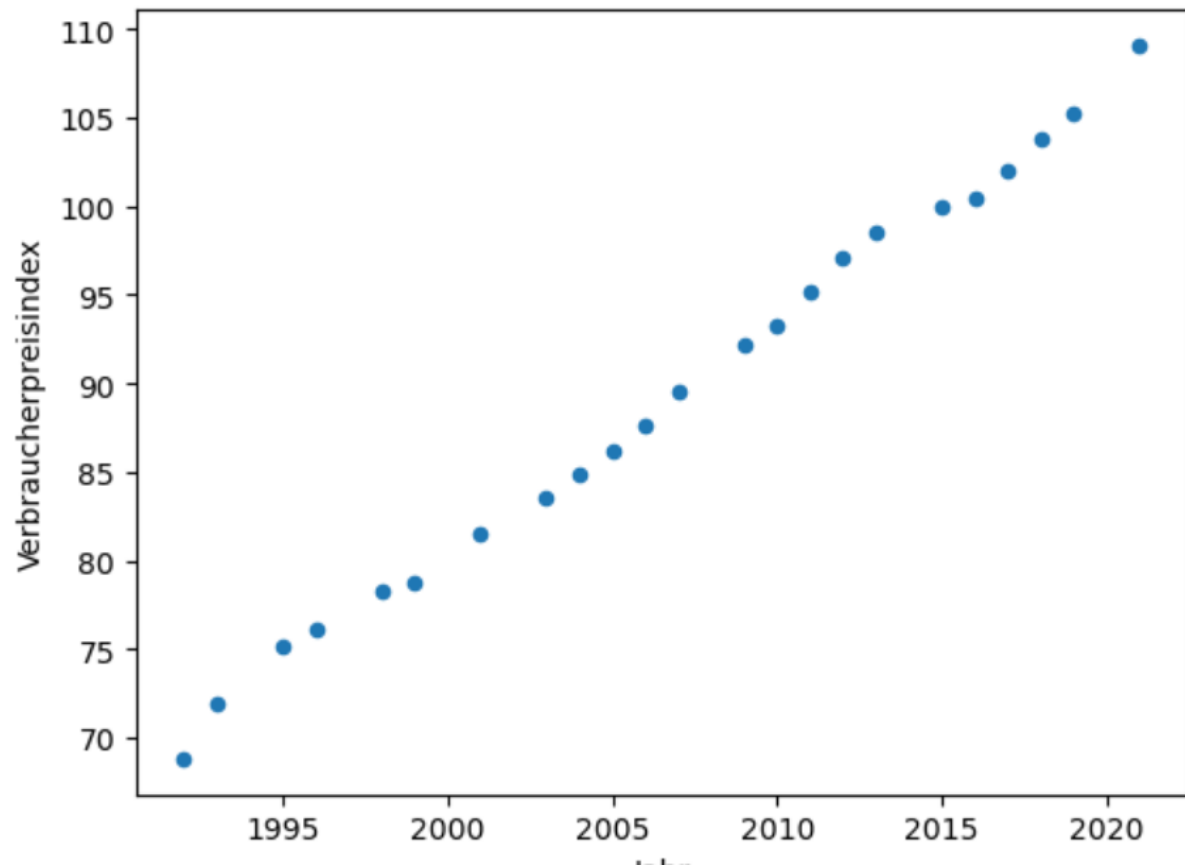
```
df.boxplot(column="Verbraucherpreisindex")
plt.show()
```





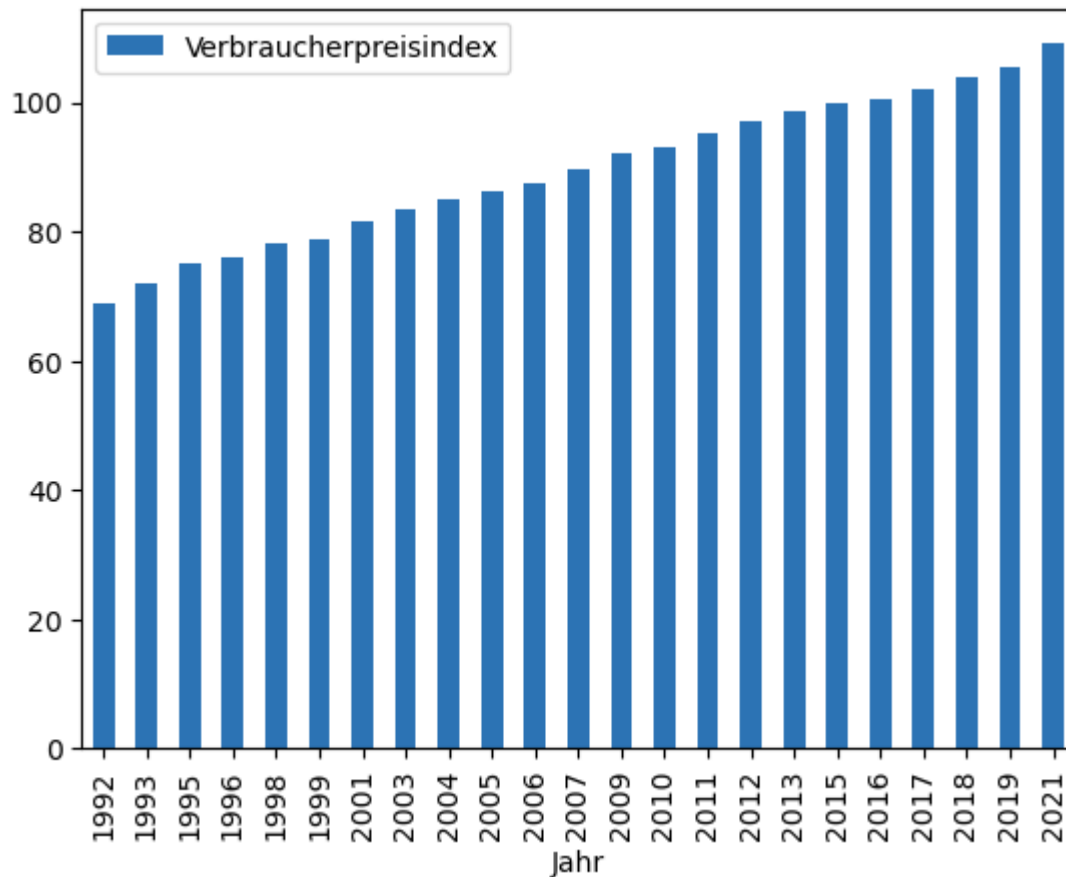
R2.14

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawtl02.dataset.d03-main/gawtl02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.plot.scatter(x='Jahr', y='Verbraucherpreisindex')
plt.show()
```



R2.15

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.plot.bar(x='Jahr', y='Verbraucherpreisindex')
plt.show()
```



Die gezeigte grafische Darstellung ist ein Säulendiagramm, welches den Verbraucherpreisindex in den verschiedenen Jahren visualisiert. Jede Säule entspricht einem anderen Jahr, abzulesen an der X-Achse. Die Höhe der jeweiligen Säule entspricht dem Verbraucherpreisindex, abzulesen an der Y-Achse.

R2.17

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.quantile(q=[0.25, 0.5, 0.75], axis=0, numeric_only=False, interpolation='nearest',
method='single')
```

Quartil	Jahr	Verbraucherpreisindex
0.25	2001	81.5
0.50	2007	89.6
0.75	2013	98.5

```
df.quantile(q=[.1, .2, .3, .4, .5, .6, .7, .8, .9 ], axis=0, numeric_only=False,
interpolation='nearest', method='single')
```

Dezil	Jahr	Verbraucherpreisindex
0.1	1995	75.1
0.2	1998	78.3
0.3	2003	83.5
0.4	2005	86.2
0.5	2007	89.6
0.6	2010	93.2
0.7	2012	97.1
0.8	2016	100.5
0.9	2018	103.8

R2.18

Quartilsabstand:

Jahr: 12

Verbraucherpreisindex: 17

R2.19

```
import pandas
import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.cov()
```

Kovarianz der Variablen Jahr und Verbraucherpreisindex: 100.837747

R2.20

```
import pandas
```

```

import matplotlib.pyplot as plt
Data1 = "C:/Users/andre/OneDrive/Dokumente/Uni/GAWT/L02
Projekt/gawt.l02.dataset.d03-main/gawt.l02.dataset.d03-main/Dataset-2/data-
2bereinigt.csv"
df = pandas.read_csv(Data1)
df.corr()

```

Korrelationskoeffizient der Variablen Jahr und Verbraucherpreisindex: 0.997983

### R3.1

Es gibt 02 (zwei) Datensätze : 3a und 3b

Der Datensatz 3a Der Datensatz liegt im .csv-Format vor (Trennzeichen','; Dezimaltrennzeichen',';UTF-8 Kodierung). Jedes Element des Datensatzes beinhaltet ein Zahl oder ein Key zwischen 1 und 31 entsprechend jeweils den Jahren von 1991 bis 2021, alle Key in diesem Intervall ist genau einmal vorhanden. Die meisten Elemente beinhalten einen Verbraucherpreisindex, bei manchen Datenpunkten fehlt der Verbraucherpreisindex oder ist nicht plausibel (keine Zahl, ein Name oder NaN). Das Jahr 2015 stellt den Referenzwert dar und entspricht dem Wert 100. Die Daten stammen vom statistischen Bundesamt (Destatis). Stand der Daten: 10.10.2022 / 10:26:07.

Der Datensatz 3b liegt im .csv-Format vor (Trennzeichen','; Dezimaltrennzeichen',';UTF-8 Kodierung). Jedes Element des Datensatzes beinhaltet eine Jahreszahl zwischen 1991 und 2021, alle Jahreszahlen in diesem Intervall sind genau einmal vorhanden. Die alle Elemente beinhalten eine Zahl, die das "Key" darstellt. Der Key reicht von einem bis 1 bis 31 und entspricht damit der Anzahl der Jahre zwischen 1991 und 2021. Der Key für 1991 "1" ist und der Key für 2021 "31" ist.

### R3.4

Zu 3a : Es existieren zwei Arten fehlerhafter Datenpunkte: Fehlender Wert der Variable Verbraucherpreisindex, nicht plausibler Wert der Variable Verbraucherpreisindex (keine Zahl,name Ausreißer in Jahreszahl). Die fehlerhaften Datenpunkte wurden beseitigt durch Löschen.

Der Ausreißer in Jahreszahlen wurde angepasst: 1795;2105.

Folgende Datenpunkte wurden gelöscht:

1991,  
 1994,Peter  
 1997,  
 2000,  
 2002,  
 2008,Hans  
 2014,NaN  
 2020,

Für 3b braucht man eigentlich keine Maßnahmen zu nehmen, da die Daten stellen kein fehlerhaft form hier dar .

### R3.5

### R3.6

Verwendete Software/Funktionen: Python, Jupyter, Pandas.

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
Data1 = "C:/Users/kanaf/Downloads/data-3-a.csv"
DataZiel="C:/Users/kanaf/Downloads/data-3-a.csv.xlsx"
df = pandas.read_csv(Data1)
df.to_excel(DataZiel)
```

```
import pandas
import matplotlib.pyplot as plt
pandas.__version__
Data1 = "C:/Users/kanaf/Downloads/data-3-b.csv"
DataZiel="C:/Users/kanaf/Downloads/data-3-b.csv.xlsx"
df = pandas.read_csv(Data1)
df.to_excel(DataZiel)
```