

Projekt Datenanalyse

Grundlagen und Anwendungen der Wahrscheinlichkeitstheorie
Wintersemester 2025/2026

Gruppe 01

Dippel, Markus — Matr.Num.: 428623
Pantleon, Andrea — Matr.Num.: 432016
Schäfer, Marc — Matr.Num.: 432324

Januar 2026

Inhaltsverzeichnis

I	Datensatz 1	2
1	Beschreibung	2
2	Grafische Darstellung	3
3	Auswertung der Daten	6
II	Datensatz 2	10
1	Beschreibung	10
2	Grafische Darstellung	11
3	Auswertung der Daten	14
III	Datensatz 3	17
1	Beschreibung	17
2	Grafische Darstellung	18
3	Auswertung der Daten	21
IV	Datensatz 4	25
1	Beschreibung	25
2	Grafische Darstellung	26
3	Auswertung der Daten	30
V	Programmcode und verwendete Funktionen	33
VI	Dateienverzeichnis	35
VII	Quellenverzeichnis	36

Verwendete Dateien und Quellen werden im folgenden mit D oder Q respektiv angegeben, mitsamt einer Nummerierung. Diese verweisen entsprechend auf Dateien- oder Quellenverzeichnis.

I Datensatz 1

1 Beschreibung

1.1 Struktur und Inhalt

In dem gegebenen Datensatz 1 wird die Elektrizitätserzeugung aus Steinkohle in Deutschland von 2002 bis 2023 angegeben. Es handelt sich im Bezug auf die Quelle um einen Ausschnitt der Rohdaten, welche unbereinigt (vollständig, aber unsortiert) vorliegen. Die Daten sind in vier, von Semicolon getrennten Spalten unterteilt. Hierbei werden die Daten in Jahr (Spalte A), Monat (Spalte B), Energieträger (Spalte C) und Netto Energieerzeugung (Spalte D) geteilt. Während der Energieträger konstant als Steinkohle angegeben ist, wird die Energieerzeugung in Megawattstunden (MWh) angegeben. In jedem Jahr ist jeweils ein Energiewert aus den Monaten April, August und Dezember gegeben.

1.2 Datenursprung

Die Daten stammen vom statistischen Bundesamt ([Q1](#)). Ihr Stand ist vom 22.10.2024.

1.3 Format

Die Daten wurden uns in Form einer Datei ([D1](#)) zur Verfügung gestellt. Sie besitzt die Codierung "UTF-8".

1.4 Datenaufbereitung

Direkte Maßnahmen zum Bereinigen waren hier nicht vonnöten, da das Ordnen der Datenpunkte im hierrauflgenden Schritt von unserem Programm übernommen wurde. Wir haben die uns gegebenen Daten von einem selbstgeschriebenen Programm in Python auswerten lassen, welches auch im nachfolgenden Punkt 1.5 genauer referenziert wird. Aus dieser Auswertung ergaben sich für uns mehrere Dateien: einzeln aufgetrennte Ur- sowie Ranglisten für die Variablen "Elektrizitätserzeugung in TWh" ([D2](#)), "Jahreszahl" ([D3](#)) sowie "Monat" ([D4](#)); Box-Whisker-Plots für jede Variable, welche im Teil 'Grafische Darstellung' sichtbar sind; ein Histogramm für die durchschnittliche Energieerzeugung pro Jahr; ein Scatter-Plot für die Datenpunkte der Energieerzeugung nach Jahren.

1.5 Verwendete Software und Funktionen

Für die Aufbereitung und Darstellung der uns gegebenen Daten haben wir ein Programm in Python geschrieben. Dieses lässt sich vollständig im Git Repository unserer Abgabe finden; ein Link zu dieser sowie eine Auflistung aller verwendeten Funktionen sind auch im Punkt 'Programmcode und verwendete Funktionen' zu finden.

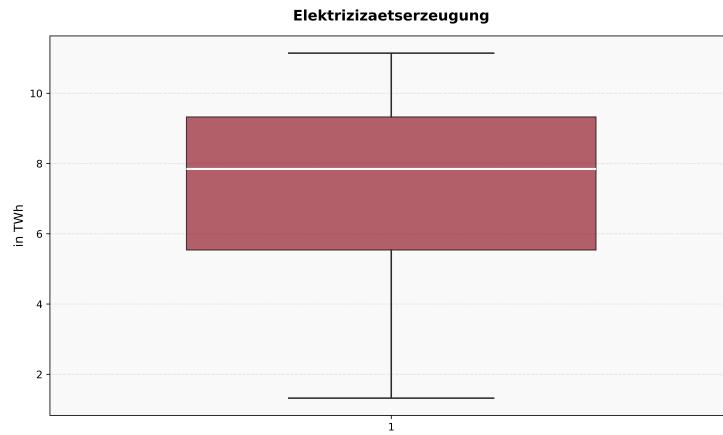
2 Grafische Darstellung

2.1 Skalenvarianten

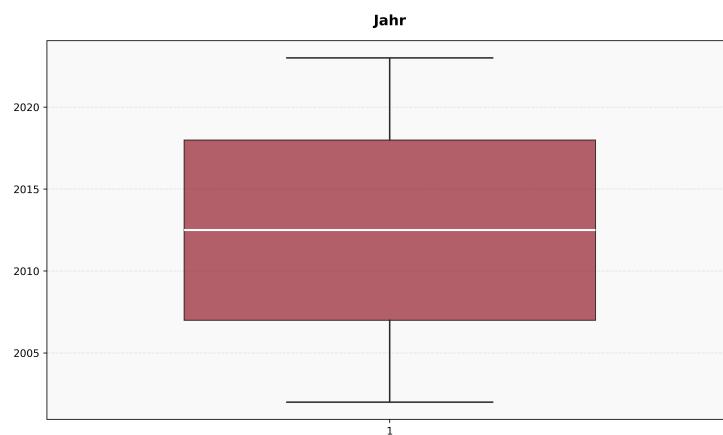
Die gewählten Skalenvarianten der einzelnen Variablen lauten wie folgt:

- Elektrizitätserzeugung: Verhältnisskala (in TWh)
- Jahr: Intervallskala
- Monat: Intervallskala (nutzt umgewandelten Zahlenwert der Monate)

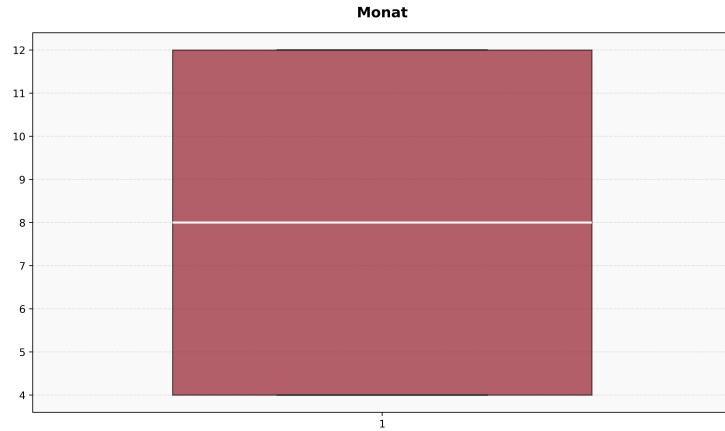
2.2 Box-Whisker-Plots



Grafik 1: Elektrizitätserzeugung in TWh

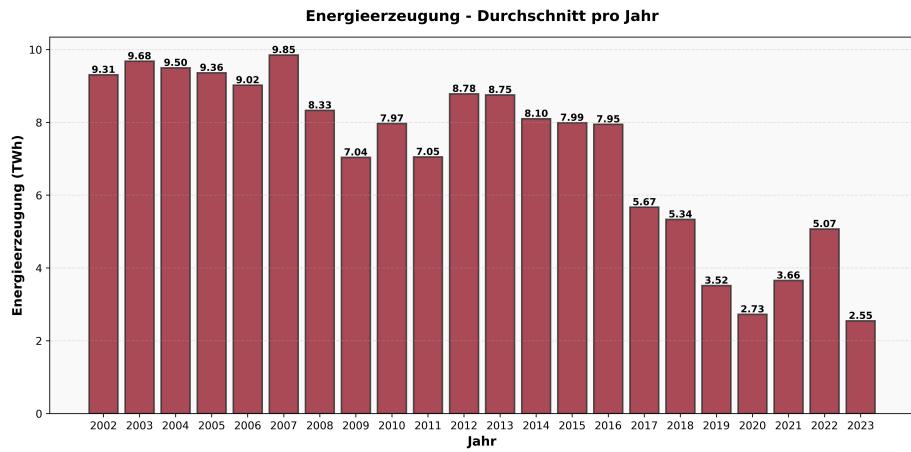


Grafik 2: Jahre

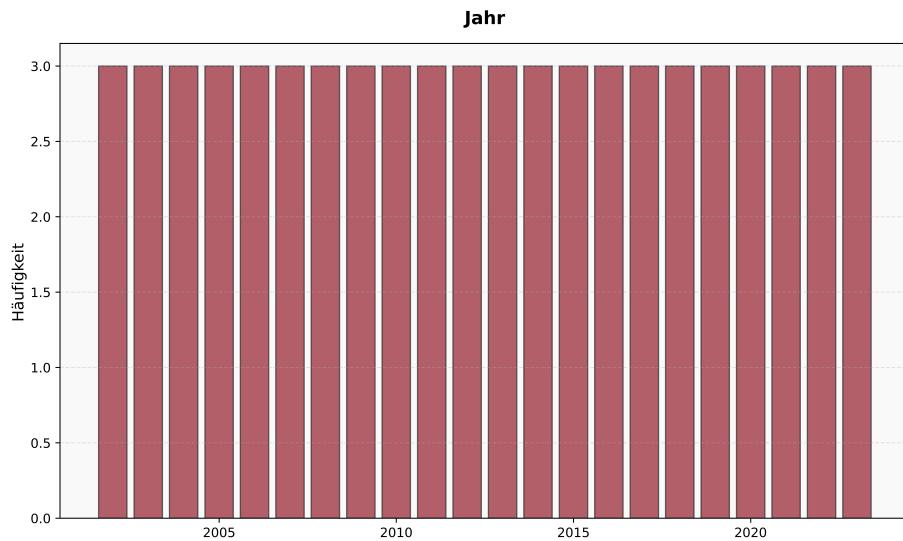


Grafik 3: Monate

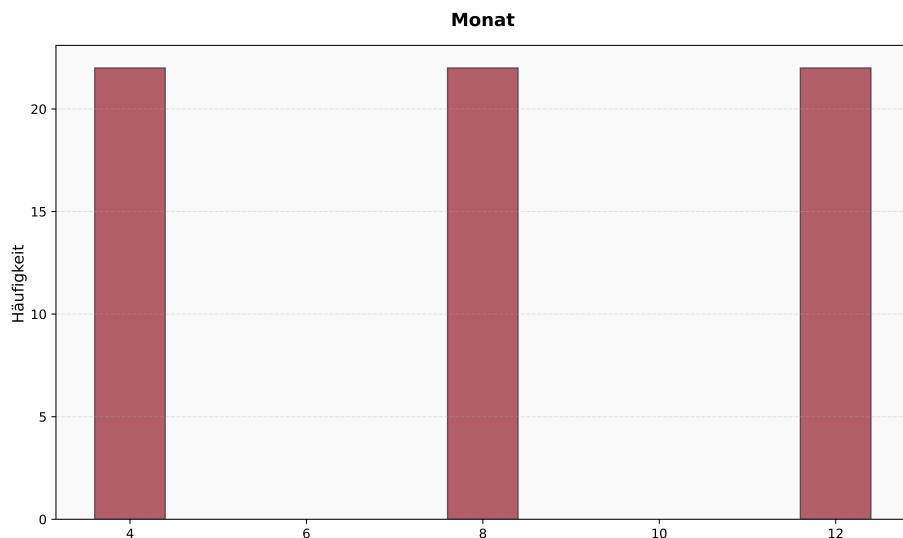
2.3 Histogramme



Grafik 4: Energieerzeugung in TWh, 2002 bis 2023

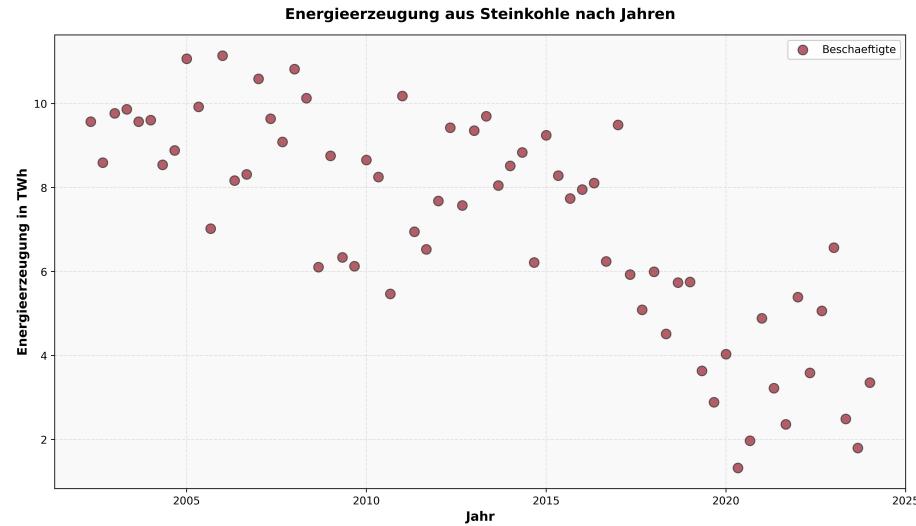


Grafik 5: Energieerzeugung in TWh, 2002 bis 2023



Grafik 6: Jahre

2.4 Scatter-Plot



Grafik 7: Monate

3 Auswertung der Daten

3.1 Variablenassoziierte Werte

Variable: Elektrizitätserzeugung netto in MWh

Kenngröße	Wert
Modus	9 567 448,000
Arithmetischer Mittelwert	7 145 764,364
Median	7 844 663,000
Spannweite	9 821 158,000
Abweichung Median	2 129 084,939
Variationskoeffizient	0,358
Quartilsabstand	3 791 441,250

Tabelle 1: Statistische Maße der Variable Elektrizitätsproduktion

Quartile: • **q25:** 5 536 304,5 • **q50:** 7 844 663
 • **q75:** 9 327 745,75 • **q100:** 11 142 512

Kenngroße	Wert
Varianz	$6,528 \times 10^{12}$
Kovarianz	$6,629 \times 10^{12}$

Dezile:	• d1: 3 286 622,5	• d6: 8 315 228
	• d2: 5 062 261	• d7: 8 860 691
	• d3: 5 960 593	• d8: 9 567 448
	• d4: 6 526 260	• d9: 9 892 590,5
	• d5: 7 844 663	• d10: 11 142 512

Variable: Jahr

Kenngroße	Wert
Modus	2 002,000
Arithmetischer Mittelwert	2 012,500
Median	2 012,500
Spannweite	21,000
Abweichung Median	5,500
Varianz	40,250
Variationskoeffizient	0,003
Kovarianz	40,869
Quartilsabstand	11,000

Tabelle 2: Statistische Maße der Variable Jahr

Quartile: • q25: 2007 • q50: 2012,5
 • q75: 2018 • q100: 2023

Dezile: • d1: 2 004 • d6: 2 015
 • d2: 2 006 • d7: 2 017
 • d3: 2 008 • d8: 2 019
 • d4: 2 010 • d9: 2 021
 • d5: 2 012,5 • d10: 2 023

Variable: Monat

Quartile: • q25: 4 • q50: 8
 • q75: 12 • q100: 12

Kenngröße	Wert
Modus	4,000
Arithmetisches Mittelwert	8,000
Median	8,000
Spannweite	8,000
Abweichung Median	2,667
Varianz	10,667
Variationskoeffizient	0,408
Kovarianz	10,831
Quartilsabstand	8,000

Tabelle 3: Statistische Maße der Variable Monat

- | | |
|--|--|
| Dezile: <ul style="list-style-type: none"> • d1: 4 • d2: 4 • d3: 4 • d4: 8 • d5: 8 | <ul style="list-style-type: none"> • d6: 8 • d7: 12 • d8: 12 • d9: 12 • d10: 12 |
|--|--|

3.2 Rangkorrelationskoeffizienz

Der Rangkorrelationskoeffizient nach Spearman zwischen den Jahren und der durchschnittlichen Energieerzeugung lautet: $-0,892$.

3.3 Fazit

Insgesamt ist aus dem Datensatz herauszulesen, dass die netto Elektrizitätserzeugung in der Zeitspanne von April 2002 bis Dezember 2023 gesunken ist.

In der durchschnittlichen Energieerzeugung pro Jahr ist zunächst in den Jahren 2002-2007 keine signifikante Änderung zu sehen, wobei im Dezember 2005 der Höhepunkt der Energieerzeugung (mit 11 142 512 MWh erreicht wurde. Diesem voraus geht ein Sprung von August auf Dezember 2005. Hier nimmt die Energieerzeugung um circa 4,12 Millionen MWh zu. Nach einem Jahresschnittsmaximum in 2007 ist dann ein globaler Trend gegen die Energieerzeugung aus Steinkohle zu erkennen.

Während 2009 und 2011 noch lokale Minima im Jahresschnitt darstellen, sinkt dieser in den Jahren 2017-2022 recht stark, wobei er 2020 sogar einen dem Graphen globalen Tiefpunkt erreicht. In diesem Jahr wurde der anfängliche Durchschnitt von 2002 auf ein Drittel dessen reduziert.

Erwähnenswert ist ebenso der April 2020, welcher am Graphen betrachtet das absolute Minimum der Monatsbilanz darstellt. Hier wurden nur etwas 11 Pro-

zent des Höhepunktes von 2005 erzeugt.

Der Abwärtstrend zwischen 2008 und 2011 ist unter anderem mit dem Kyoto-Protokoll ([Q2](#)) zu begründen; hier verspflchtete sich auch Deutschland, seinen Treibhausgas-Ausstoß der sechs wichtigsten Treibhausgase bis 2012 zu begrenzen. Dessen Nachfolger, das Pariser Klimaabkommen ([Q3](#)), ist indes auch eine potentiell kausale Verbindung zum Verhalten des Jahresdurchschnitts nach 2015.

Ebenso ist eine deutliche Senkung im Jahr 2023 zu erkennen. Hier halbiert es sich annähernd im Vergleich zum Vorjahr, und ist somit in etwa auf dem selben Jahresdurchschnitt wie 2020. Der Anstieg in 2022 könnte auf gewisse politische Ereignisse zurückzuführen sein. Koks (Kohlenstoff-Rest, welcher durch Steinkohle entsteht) spielt unter anderem in der Stahlproduktion eine große Rolle. Diese stieg in den Jahren 2020 bis 2022 insgesamt an. ([Q4](#))

Um nun auf die Jahre jeweils einzugehen, lässt sich ein "VTrend erkennen. So wurde im August weniger Energie aus Steinkohle gewonnen als jeweils in den Monaten April und Dezember. Dies lässt sich unter anderem mit dem Nutzen der Energie erklären, da man sich in Deutschland im August mehr auf erneuerbare Energien beziehungsweise Solarenergie stützen kann.

II Datensatz 2

1 Beschreibung

1.1 Struktur und Inhalt

Die uns gegebene Liste enthält Datenpunkte einer Statistik zu Beschäftigten im Einzelhandel. Es existiert jeweils eine Spalte für die Jahreszahl (Spalte A), den Monat des zugehörigen Datums (Spalte B), sowie eine Spalte für Beschäftigte (Spalte C).

Spalte A enthält pro Jahr zwei Einträge, obwohl dies in kleinen Einzelfällen durch Datenverlust oder Fehler gestört wird. Die angegebenen Datenwerte umfassen die Jahre 1994 bis 2024. Spalte B enthält wiederum für jedes Jahr zwei Monatsangaben, jeweils einmal Januar und einmal Dezember. In Spalte C lässt sich dann entsprechend die Zahl der Beschäftigten finden, ausgedrückt in prozentualen Wert mit dem Vergleichsjahr 2015.

1.2 Datenursprung

Die Daten stammen vom statistischen Bundesamt ([Q5](#)). Ihr Stand ist vom 22.10.2024.

1.3 Format

Die Daten wurden uns in Form einer Datei ([D5](#)) zur Verfügung gestellt. Sie besitzt die Codierung "ISO-8859-1".

1.4 Datenaufbereitung

Zum Bereinigen der Daten haben wir die Liste mit der Originalliste vom statistischen Bundesamt abgeglichen und entsprechende Fehlerstellen abgeändert beziehungsweise ergänzt. Wir haben die uns gegebenen Daten von einem selbstgeschriebenen Programm in Python auswerten lassen, welches auch im nachfolgenden Punkt 1.5 genauer referenziert wird. Aus dieser Auswertung ergaben sich für uns mehrere Dateien: einzeln aufgetrennte Ur- sowie Ranglisten für die Variablen "Beschäftigte" ([D6](#)), "Jahreszahl" ([D7](#)) sowie "Monat" ([D8](#)); Box-Whisker-Plots für jede Variable, welche im Teil 'Grafische Darstellung' sichtbar sind; ein Scatter- sowie ein Line-Plot für die prozentuale Anzahl an Beschäftigten im Einzelhandel nach Jahren, welche im Teil 'Grafische Darstellung' sichtbar sind.

1.5 Verwendete Software und Funktionen

Für die Aufbereitung und Darstellung der uns gegebenen Daten haben wir ein Programm in Python geschrieben. Dieses lässt sich vollständig im Git Repository unserer Abgabe finden; ein Link zu dieser sowie eine Auflistung aller

verwendeten Funktionen sind auch im Punkt 'Programmcode und verwendete Funktionen' zu finden.

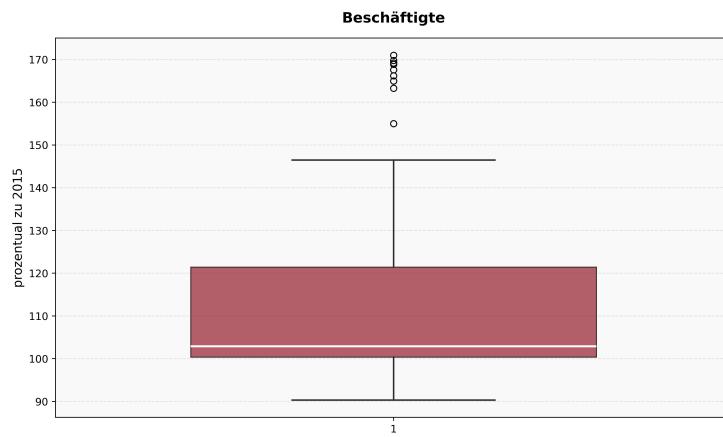
2 Grafische Darstellung

2.1 Skalenvarianten

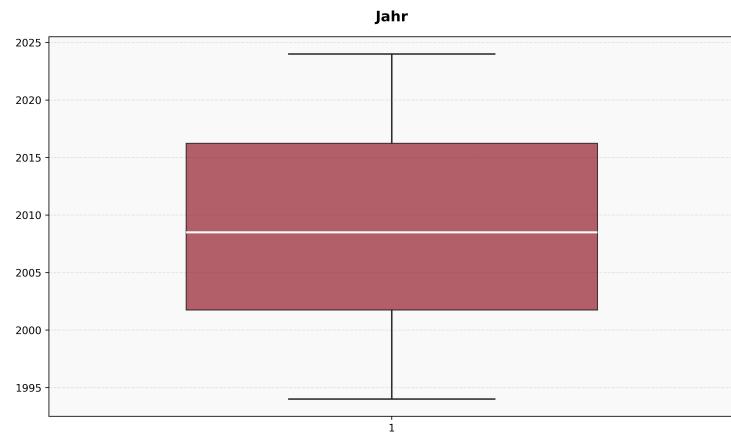
Die gewählten Skalenvarianten der einzelnen Variablen lauten wie folgt:

- Beschäftigte: Verhältnisskala (prozentual)
- Jahr: Intervallskala
- Monat: Intervallskala (nutzt umgewandelten Zahlenwert der Monate)

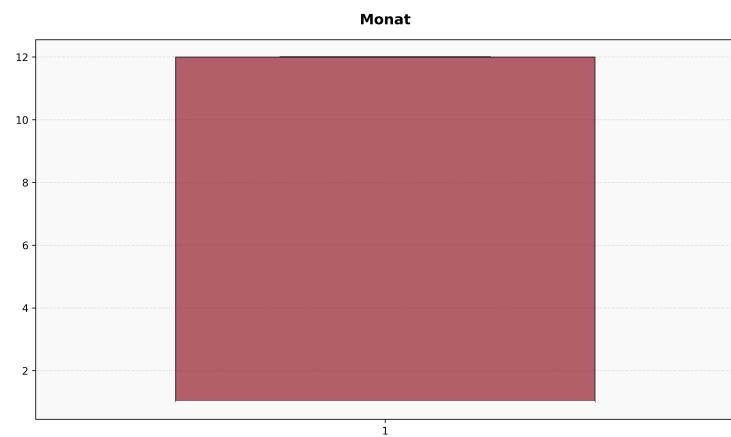
2.2 Box-Whisker-Plots



Grafik 8: Beschäftigte

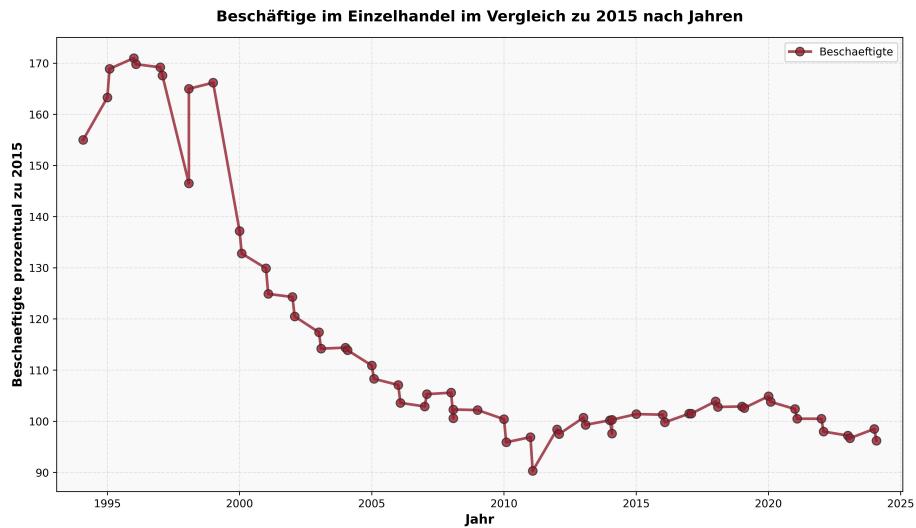


Grafik 9: Jahre



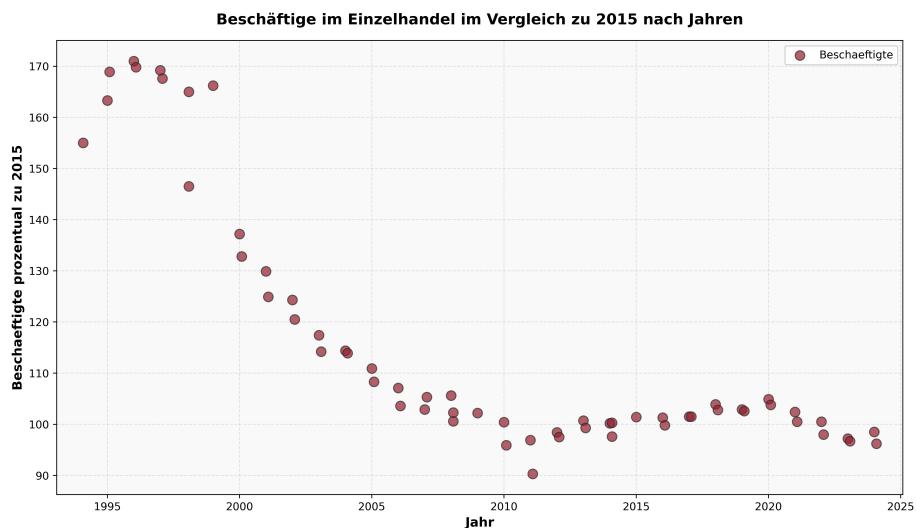
Grafik 10: Monate

2.3 Line-Plot



Grafik 11: Beschäftigte im Einzelhandel nach Jahren

2.4 Scatter-Plot



Grafik 12: Beschäftigte im Einzelhandel nach Jahren

3 Auswertung der Daten

3.1 Variablenassoziierte Werte

Variable: Beschäftigte prozentual zu 2015

Kenngröße	Wert
Modus	102,900
Arithmetisches Mittelwert	115,693
Median	102,900
Spannweite	80,700
Abweichung Median	15,828
Varianz	581,489
Variationskoeffizient	0,208
Kovarianz	591,690
Quartilsabstand	22,925

Tabelle 4: Statistische Maße der Variable Beschäftigte prozentual zu 2015

Quartile: • q25: 100,425 • q50: 102,9
 • q75: 123,35 • q100: 171

Dezile: • d1: 97,57 • d6: 105,36
 • d2: 99,96 • d7: 114,35
 • d3: 100,61 • d8: 131,64
 • d4: 102,06 • d9: 165,36
 • d5: 102,9 • d10: 171

Variable: Jahr

Kenngröße	Wert
Modus	1 998,000
Arithmetischer Mittelwert	2 008,966
Median	2 008,500
Spannweite	30,000
Abweichung Median	7,690
Varianz	78,654
Variationskoeffizient	0,004
Kovarianz	80,034
Quartilsabstand	15,500

Tabelle 5: Statistische Maße der Variable Jahr

-
- Quartile:**
- q25: 2001,25
 - q50: 2008,5
 - q75: 2016,75
 - q100: 2024
- Dezile:**
- | | | | |
|-------|--------|--------|--------|
| • d1: | 1996,7 | • d6: | 2012,2 |
| • d2: | 2000 | • d7: | 2014,9 |
| • d3: | 2003,1 | • d8: | 2018 |
| • d4: | 2006 | • d9: | 2021 |
| • d5: | 2008,5 | • d10: | 2024 |

Variable: Monat

Kenngröße	Wert
Modus	1,000
Arithmetischer Mittelwert	6,317
Median	6,500
Spannweite	11,000
Abweichung Median	5,500
Varianz	30,250
Variationskoeffizient	0,846
Kovarianz	30,781
Quartilsabstand	11,000

Tabelle 6: Statistische Maße der Variable Monat

- Quartile:**
- q25: 1
 - q50: 6,5
 - q75: 12
 - q100: 12
- Dezile:**
- | | | | |
|-------|-----|--------|----|
| • d1: | 1 | • d6: | 12 |
| • d2: | 1 | • d7: | 12 |
| • d3: | 1 | • d8: | 12 |
| • d4: | 1 | • d9: | 12 |
| • d5: | 6,5 | • d10: | 12 |

3.2 Fazit

Generell lässt sich den Graphen und Daten die Aussage entnehmen, dass die Zahl der Angestellten im Einzelhandel erst einen starken Abfall von 1999 bis 2015 erlebt hat, gefolgt von einem leichten Anstieg bis 2020 und dann darauf wieder einen leichten Abfall eingeleitet hat. Vor allem der starke Abstieg von 1999 bis 2015 ist sehr herausstechend, obwohl man erkennen kann, dass er bei

Annäherung an 2015 abflacht. Um 2015 bildet sich dann der globale Tiefpunkt der Funktion.

Da es ein prozentualer Vergleich mit 2015 ist, sollte hier nicht von absoluten Zahlen an angestellten ausgegangen werden.

Mit diesen begrenzten Daten können wir natürlich nicht wirklich kausale Schlussfolgerungen treffen, jedoch lassen sich Annahmen für korrelierende Ereignisse für die einzelnen Intervalle aufstellen.

Zum Einen das erste Intervall von 1999 bis 2015, in welchem der starke Abstieg zu verzeichnen ist. Seit der Jahrhundertwende wächst die Wichtigkeit des Onlinehandels im Vergleich zum Einzelhandel. Durch den demographischen Wandel der Bevölkerung stirbt ebenfalls langsam ein Teil der Bevölkerung, welcher noch einen vergleichsweise kleinen Anteil am Online-Handel hat. Zum anderen sind gesuchte Arbeitskräfte im Online-Handel zum größeren Teil Helfer, was für viele neue Berufseinsteiger die Attraktivität erhöhen kann. (Q6) All diese Veränderungen könnten einen Zusammenhang mit dem Abstieg haben.

Im folgenden Intervall von 2015 bis 2024 erlebt der Graph erst einen leichten Anstieg bis 2020, wobei danach wieder ein stetiger but langsamer Abfall erfolgt. Dies lässt sich möglicherweise mit der Corona-Pandemie erklären, da der Knick im Graphen um das Jahr 2020 recht auffällig erscheint.

III Datensatz 3

1 Beschreibung

1.1 Struktur und Inhalt

Der Datensatz 3 gibt die Anzahl der Ankünfte und Übernachtungen in Beherbergungsbetrieben in Deutschland wieder. Er wurde ursprünglich in zwei Dateien aufgeteilt und muss zunächst richtig zusammengefügt werden.

Die aufgenommenen Werte für Ankünfte (Spalte C) und Übernachtungen (Spalte D) sind für jeden Monat (Spalte B) eines Jahres (Spalte A) vorhanden, und erstrecken sich von Januar 1992 bis Juni 1999.

1.2 Datenursprung

Die Daten stammen vom statistischen Bundesamt ([Q7](#)). Ihr Stand ist vom 22.10.2024.

1.3 Format

Die Daten wurden uns in Form zweier Dateien ([D9](#)) zur Verfügung gestellt. Sie besitzen die Codierung "ISO-8859-1".

1.4 Datenaufbereitung

Die beiden Dateien mussten erst konsolidiert werden, damit sie ordentlich ausgewertet werden können. Dies erfolgte ebenfalls über das unten genannte Programm. Wir haben die uns gegebenen Daten von einem selbstgeschriebenen Programm in Python auswerten lassen, welches auch im nachfolgenden Punkt 1.5 genauer referenziert wird.

Zuerst ließen wir die beiden getrennten Datensätze konsolidieren, danach ließen wir das Programm diesen auswerten.

Aus dieser Auswertung ergaben sich für uns mehrere Dateien: einzeln aufgetrennte Ur- sowie Ranglisten für die Variablen "Anzahl Ankünfte"([D10](#)), "Anzahl Übernachtungen"([D11](#)), "Jahreszahl"([D12](#)) sowie "Monat" ([D13](#)); Box-Whisker-Plots für jede Variable, welche im Teil 'Grafische Darstellung' sichtbar sind; Scatter-Plots, einmal ohne und einmal mit Referenzlinien zur besseren Erkennbarkeit der Periodizität, welche im Teil 'Grafische Darstellung' sichtbar sind.

1.5 Verwendete Software und Funktionen

Für die Aufbereitung und Darstellung der uns gegebenen Daten haben wir ein Programm in Python geschrieben. Dieses lässt sich vollständig im Git Repository unserer Abgabe finden; ein Link zu dieser sowie eine Auflistung aller verwendeten Funktionen sind auch im Punkt '[Programmcode und verwendete Funktionen](#)' zu finden.

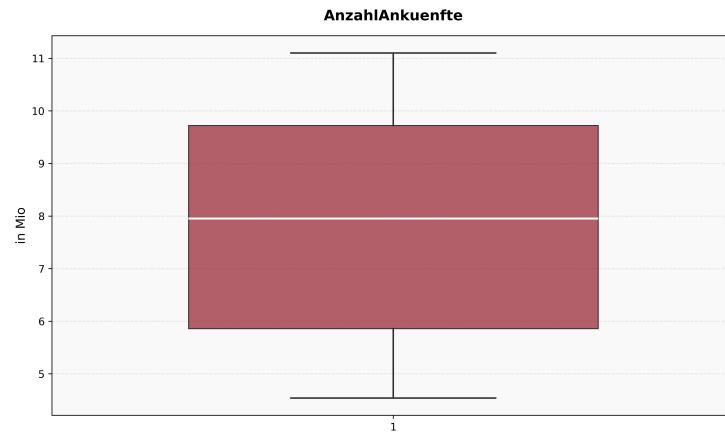
2 Grafische Darstellung

2.1 Skalenvarianten

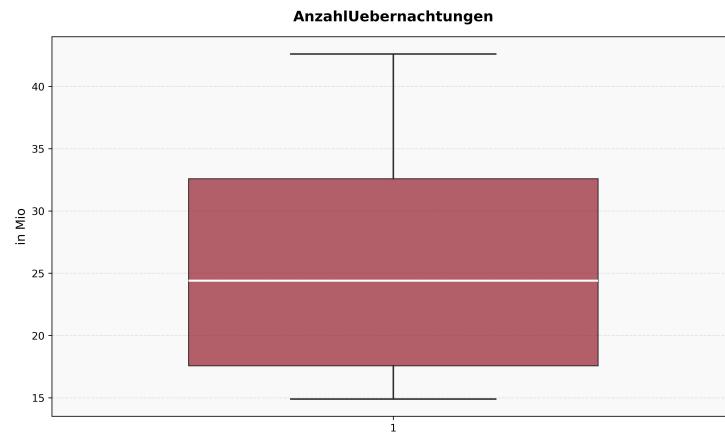
Die gewählten Skalenvarianten der einzelnen Variablen lauten wie folgt:

- Anzahl Ankünfte: Verhältnisskala
- Anzahl Übernachtungen: Verhältnisskala
- Jahr: Intervallskala
- Monat: Intervallskala (nutzt umgewandelten Zahlenwert der Monate)

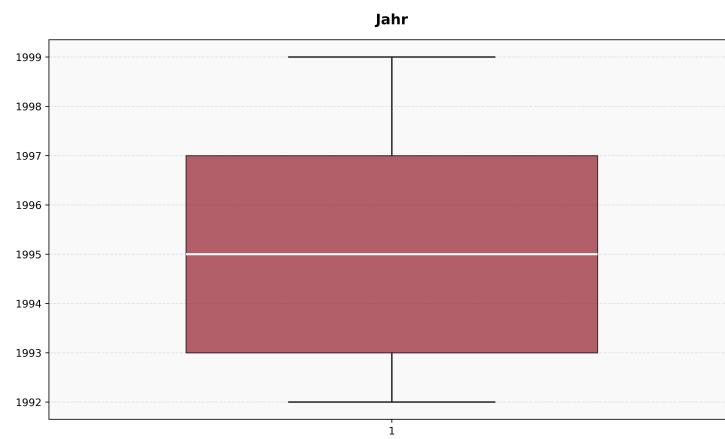
2.2 Box-Whisker-Plots



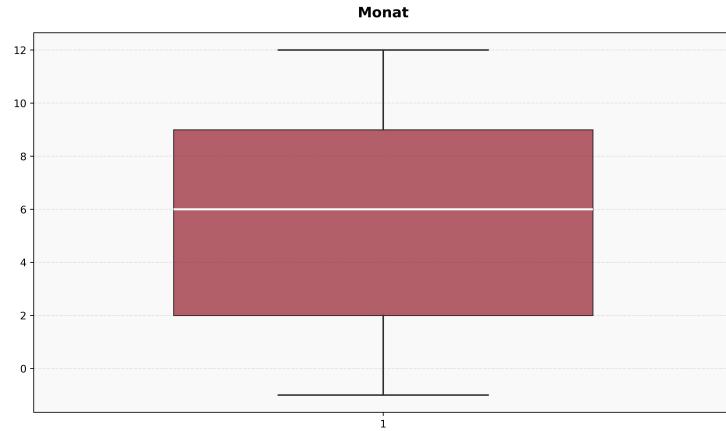
Grafik 13: Anzahl Ankünfte



Grafik 14: Anzahl Übernachtungen

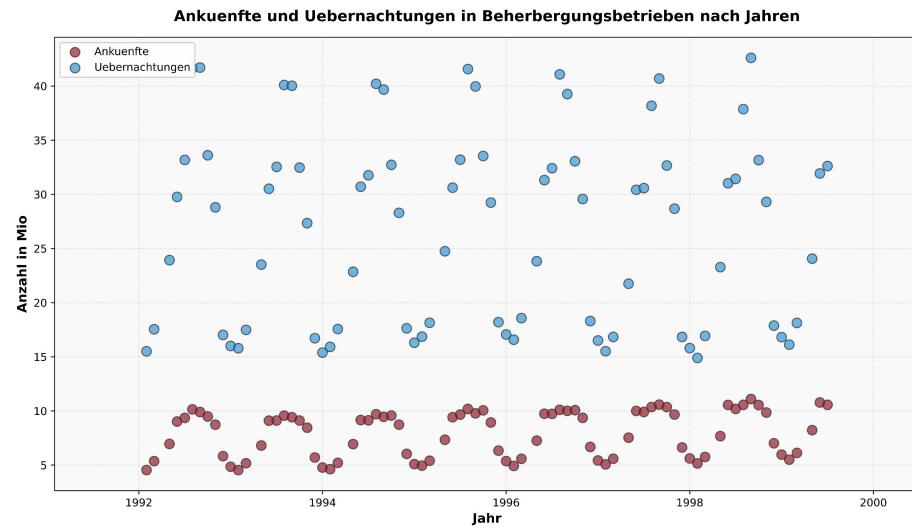


Grafik 15: Jahre

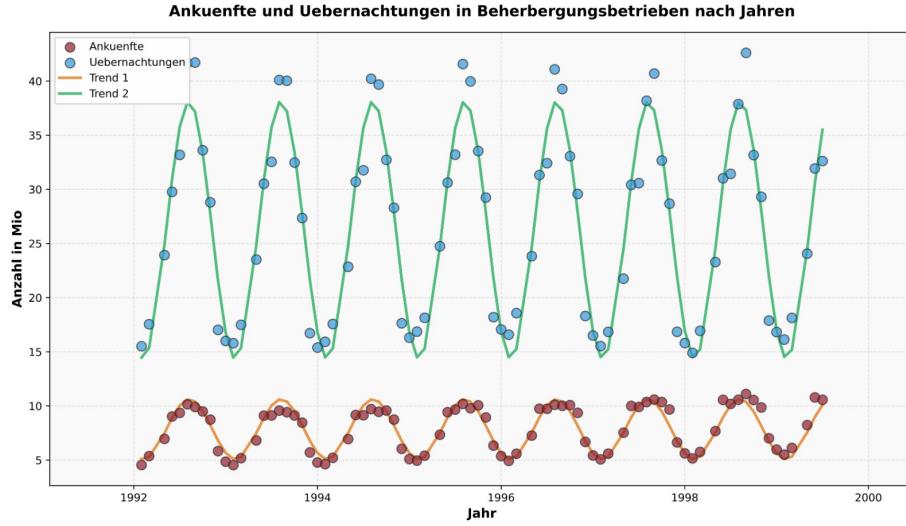


Grafik 16: Monate

2.3 Scatter-Plots



Grafik 17: Ankünfte und Übernachtungen in Beherbergungsbetrieben nach Jahren



Grafik 18: Ankünfte und Übernachtungen in Beherbergungsbetrieben nach Jahren (Scatter-Plot mit Referenzlinien)

Sinus-Funktion für das Curvefitting der Werte für Ankünfte:

$$2,765 * \sin(6,287x - 588,678) + 26,263 \quad (1)$$

Sinus-Funktion für das Curvefitting der Werte für Übernachtungen:

$$-11,879 * \sin(6,277x - 1068,677) + 26,264 \quad (2)$$

Werte, die dem Scatterplot als Legende dienen, sind in der konsolidierten Datei 'data-3_konsolidiert.csv' im Git Repository der Abgabe zu finden.

3 Auswertung der Daten

3.1 Variablenassoziierte Werte

Variable: Anzahl Ankünfte

Quartile: • q25: 5 101 327,8 • q50: 8 838 283,5
 • q75: 9 773 489,5 • q100: 11 100 085

Dezile:	• d1: 5 101 327,8	• d6: 9 365 677,2
	• d2: 5 520 768,2	• d7: 9 670 734,3
	• d3: 6 063 990,7	• d8: 9 983 344,6
	• d4: 7 112 657,8	• d9: 10 342 139,8
	• d5: 8 838 283,5	• d10: 11 100 085

Kenngroße	Wert
Modus	4 545 298,000
Arithmetisches Mittelwert	7 963 443,829
Median	8 838 283,500
Spannweite	6 560 767,000
Abweichung Median	1 884 820,367
Variationskoeffizient	0,262
Quartilsabstand	4 058 848,500

Tabelle 7: Statistische Maße der Variable Anzahl Ankünfte

Kenngroße	Wert
Varianz	$4,347 \times 10^{12}$
Kovarianz	$4,401 \times 10^{12}$

Variable: Anzahl Übernachtungen

Kenngroße	Wert
Modus	15 509 574,000
Arithmetisches Mittelwert	267 556 379
Median	28 740 142,500
Spannweite	27 705 308,000
Abweichung Median	7 800 643,902
Variationskoeffizient	0,329
Quartilsabstand	15 205 252,000

Tabelle 8: Statistische Maße der Variable Anzahl Übernachtungen

Quartile: • q25: 17 505 306,5 • q50: 28 740 142,5
 • q75: 32 710 558,5 • q100: 42 603 257

• d1:	16 143 207,5	• d6:	30 602 688,0
• d2:	16 872 959,6	• d7:	32 455 232,0
Dezile:	• d3: 17 953 961,1	• d8:	33 199 623,6
	• d4: 23 374 143,4	• d9:	40 026 292,4
	• d5: 28 740 142,5	• d10:	42 603 257

Kenngröße	Wert
Varianz	$77,602 \times 10^{12}$
Kovarianz	$78,560 \times 10^{12}$

Variable: Jahr

Kenngröße	Wert
Modus	1 992,000
Arithmetischer Mittelwert	1 995,244
Median	1 995,000
Spannweite	7,000
Abweichung Median	1,854
Varianz	4,672
Variationskoeffizient	0,001
Kovarianz	4,730
Quartilsabstand	4,000

Tabelle 9: Statistische Maße der Variable Jahr

Quartile: • q25: 1 993 • q50: 1 995
 • q75: 1 997 • q100: 1 999

Dezile:	• d1: 1 992	• d6: 1 996
	• d2: 1 993	• d7: 1 997
	• d3: 1 994	• d8: 1 997
	• d4: 1 994,4	• d9: 1 998
	• d5: 1 995	• d10: 1 999

Variable: Monat

Quartile: • q25: 4 • q50: 7
 • q75: 9,75 • q100: 12

Dezile:	• d1: 2	• d6: 8
	• d2: 4	• d7: 9
	• d3: 5	• d8: 10
	• d4: 6	• d9: 11
	• d5: 7	• d10: 12

Kenngröße	Wert
Modus	1
Arithmetischer Mittelwert	6,622
Median	7,000
Spannweite	11,000
Abweichung Median	2,939
Varianz	11,869
Variationskoeffizient	0,520
Kovarianz	12,016
Quartilsabstand	5,750

Tabelle 10: Statistische Maße der Variable Monat

3.2 Fazit

Insgesamt lässt sich sagen, dass die Anzahl der Ankünfte und Übernachtungen über den gegebenen Zeitraum weder stark ansteigt oder abnimmt.

Aus den Daten wird jedoch ein im Durchschnitt annehmbar direktes Verhältnis zwischen Ankünften und Übernachtungen deutlich. Die Anzahl der Übernachtungen ist doch deutlich höher, weshalb man die Vermutung aufstellen könnte, dass Anreisende Personen im Schnitt mehrere Tage in den Betrieben verbringen. Eine mögliche Korrelation zwischen den beiden Ereignissen Ankunft und Übernachtungen wird vor allem durch die Betrachtung der Scatter-Plots suggeriert, da die Funktion der Übernachtungen einen Abwärtstrend aufweist, sobald auch die Anzahl der Ankünfte abnimmt. Die Übernachtungen steigen dann auch parallel zur Anzahl der Ankünfte an.

Ebenso ist das periodische Schwanken der Funktionen möglicherweise damit zu erklären, dass je nach Jahreszeit ein größerer Urlaubsdrang vorhanden ist als in anderen Monaten.

Das globale Minimum an Übernachtungen und Ankünften stellt der August 1998 dar; aufgeteilt auf die separaten Werte gibt es die wenigsten Übernachtungen im Januar 1998, die wenigsten Ankünfte hingegen im Januar 1993.

IV Datensatz 4

1 Beschreibung

1.1 Struktur und Inhalt

Der vorliegende Datensatz umfasst die tägliche Schrittzahl eines Studenten über einen Zeitraum von einem Jahr (2025). Die Zeiteinheit der einzelnen Datenpunkte entspricht hier jeweils einem Kalendertag (24 Stunden), und es wird die Anzahl der an diesem Tag zurückgelegten Schritte dargestellt.

Die Daten wurden fortlaufend und regelmäßig von einer Apple Watch 2 Ultra erhoben, sodass keine zeitlichen Lücken im Datensatz vorliegen. Die Schrittzahlen variieren von Tag zu Tag und spiegeln unerschiedliche Aktivitätsniveaus wider; diese können unter anderem durch Alltag, Vorlesungen, Freizeitaktivitäten, Wochenenden, Feiertagen oder Krankheit beeinflusst werden.

Datum des Exportes ist der 27.01.2026. Somit ist das Erhebungsjahr zum Zeitpunkt der Datenexportierung abgeschlossen.

1.2 Datenursprung

Der Datensatz stammt von einem Studenten der Gruppe.

1.3 Format

Die Daten wurden uns in Form einer Datei ([D14](#)) zur Verfügung gestellt. Sie besitzt die Codierung "UTF-8".

1.4 Datenaufbereitung

Wir haben die uns gegebenen Daten von einem selbstgeschriebenen Programm in Python auswerten lassen, welches auch im nachfolgenden Punkt 1.5 genauer referenziert wird.

Aus dieser Auswertung ergaben sich für uns mehrere Dateien: einzeln aufgetrennte Ur- sowie Ranglisten für die Variablen "Schritte" ([D15](#)), "Monat" ([D16](#)), "Tag" ([D17](#)) sowie "Wochentag" ([D18](#)); Box-Whisker-Plots für jede Variable, welche im Teil 'Grafische Darstellung' sichtbar sind; eine Verteilung von Scatter-Plots jeweils Verteilung von Schritten pro Monat sowie die Verteilung schon Schritten pro Monatstag; Histogramme für die durchschnittlichen Schritte pro Monat sowie die durchschnittlichen Schritte pro Wochentag; ein Line-Plot über die durchschnittlichen Schritte pro Monatstag.

1.5 Verwendete Software und Funktionen

Für die Aufbereitung und Darstellung der uns gegebenen Daten haben wir ein Programm in Python geschrieben. Dieses lässt sich vollständig im Git Repository unserer Abgabe finden; ein Link zu dieser sowie eine Auflistung aller

verwendeten Funktionen sind auch im Punkt 'Programmcode und verwendete Funktionen' zu finden.

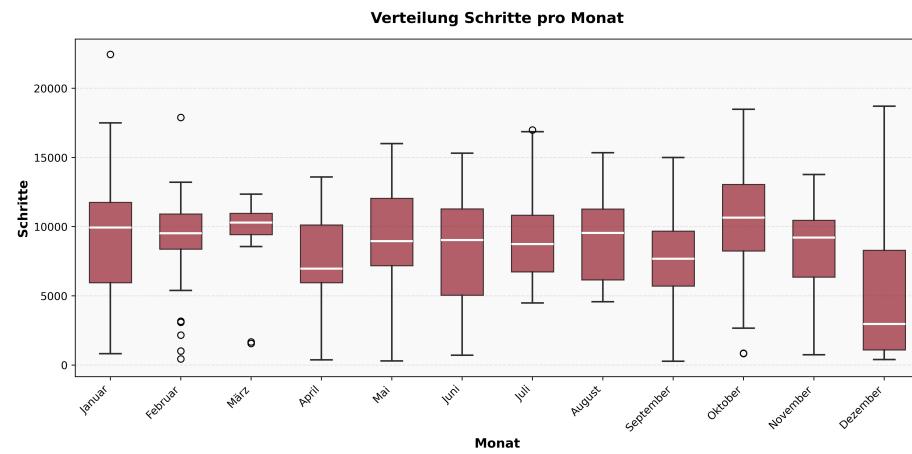
2 Grafische Darstellung

2.1 Skalenvarianten

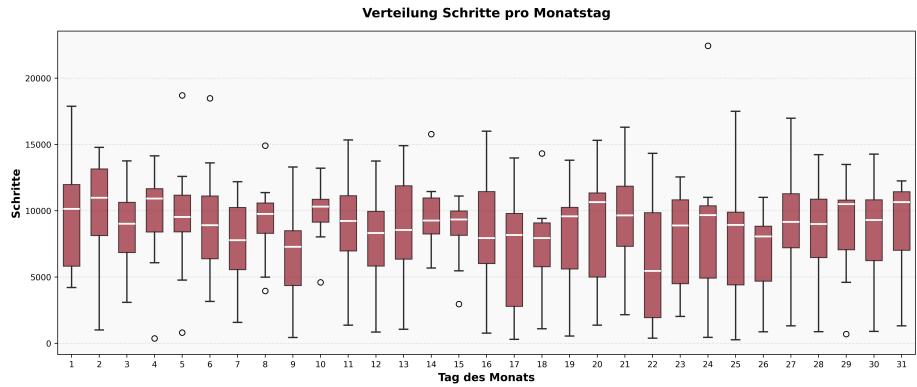
Die gewählten Skalenvarianten der einzelnen Variablen lauten wie folgt:

- Schritte: Verhältnisskala
- Monat: Intervallskala (nutzt umgewandelten Zahlenwert der Monate)
- Tag: Intervallskala

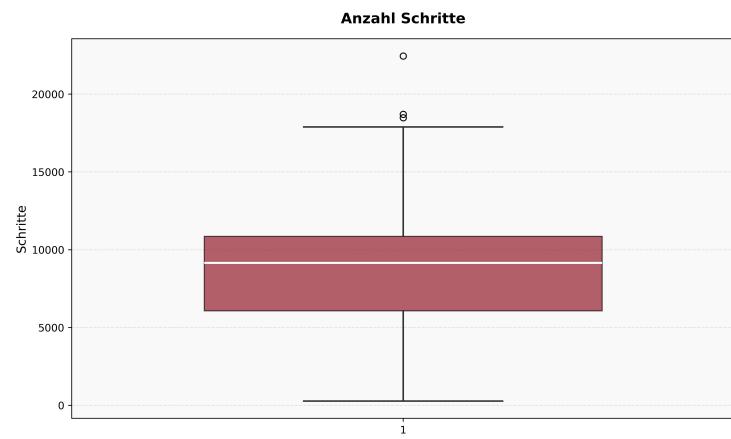
2.2 Box-Whisker-Plots



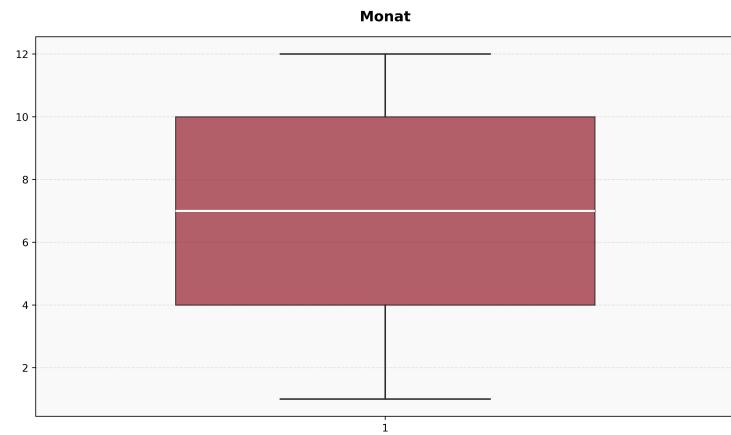
Grafik 19: Schritte pro Monat



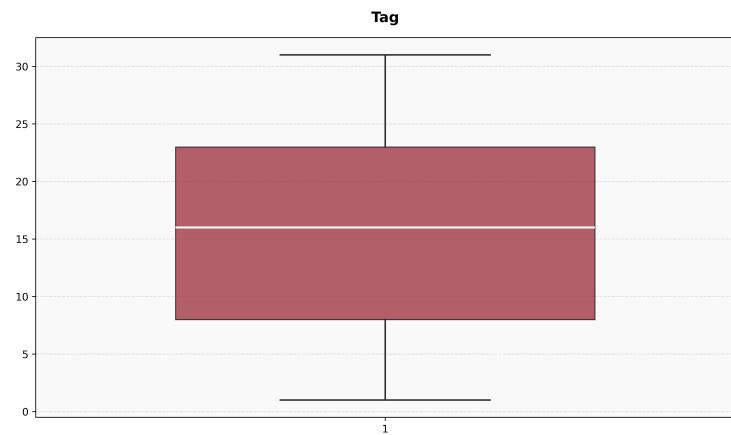
Grafik 20: Schritte pro Monatstag



Grafik 21: Anzahl Schritte

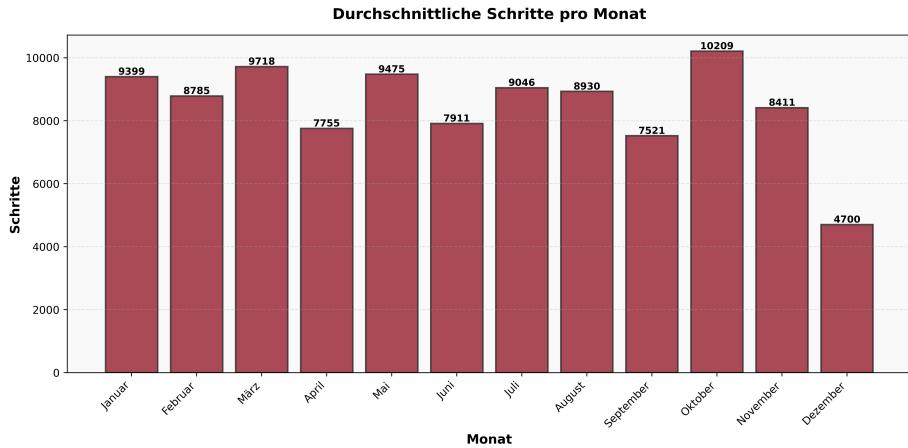


Grafik 22: Monate

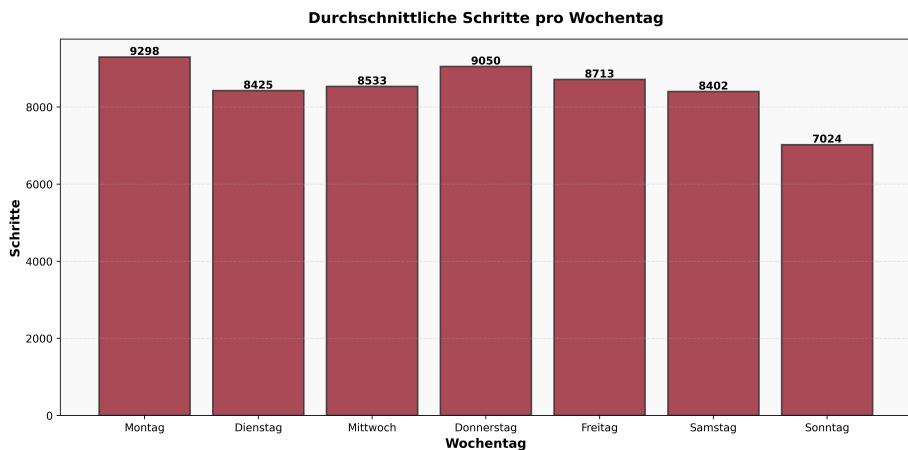


Grafik 23: Tage

2.3 Histogramme

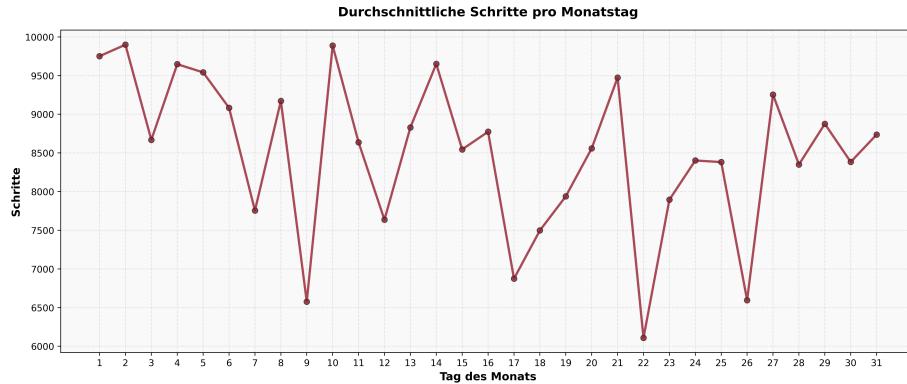


Grafik 24: Durchschnittliche Schritte pro Monat



Grafik 25: Durchschnittliche Schritte pro Wochentag

2.4 Line-Plot



Grafik 26: Durchschnittliche Schritte pro Monatstag

3 Auswertung der Daten

3.1 Variablenassoziierte Werte

Variable: Anzahl Schritte

Kenngröße	Wert
Modus	8 788,000
Arithmetischer Mittelwert	8 492,378
Median	9 156,000
Spannweite	22 177,000
Abweichung Median	3 047,014
Variationskoeffizient	0,461
Quartilsabstand	4 787,000

Tabelle 11: Statistische Maße der Variable Anzahl Schritte

Kenngröße	Wert
Varianz	$15,341 \times 10^6$
Kovarianz	$15,383 \times 10^6$

Quartile: • q25: 6 080 • q50: 9 156
 • q75: 10 867 • q100: 22 448

	• d1: 2 626,2	• d6: 9 700,8
	• d2: 5 039,59	• d7: 10 600,8
Dezile:	• d3: 6 402	• d8: 11 329,8
	• d4: 8 351,4	• d9: 13 097
	• d5: 9 156	• d10: 22 448

Variable: Monat

KenngroÙe	Wert
Modus	1,000
Arithmetischer Mittelwert	6,526
Median	7,000
Spannweite	11,000
Abweichung Median	2,989
Varianz	11,888
Variationskoeffizient	0,528
Kovarianz	11,920
Quartilsabstand	6,000

Tabelle 12: Statistische MaÙe der Variable Monat

Quartile:	• q25: 4	• q50: 7
	• q75: 10	• q100: 12

	• d1: 2	• d6: 8
	• d2: 3	• d7: 9
Dezile:	• d3: 4	• d8: 10
	• d4: 5	• d9: 11
	• d5: 7	• d10: 12

Variable: Tag

Quartile:	• q25: 8	• q50: 16
	• q75: 23	• q100: 31

	• d1: 4	• d6: 19
	• d2: 7	• d7: 22
Dezile:	• d3: 10	• d8: 25
	• d4: 13	• d9: 28
	• d5: 16	• d10: 31

Kenngröße	Wert
Modus	1,000
Arithmetisches Mittelwert	15,721
Median	16,000
Spannweite	30,000
Abweichung Median	7,611
Varianz	77,374
Variationskoeffizient	0,560
Kovarianz	77,587
Quartilsabstand	15,000

Tabelle 13: Statistische Maße der Variable Tag

3.2 Fazit

Insgesamt lässt sich aus dem Datensatz erkennen, dass ein konstantes Aktivitätsniveau über das gesamte Jahr gehalten wurde.

Wenn die durchschnittlichen Schritte pro Wochentag betrachtet werden, so fallen Maxima an Montag und Donnerstag auf, während das Wochenende die niedrigsten Werte durchschnittlichen Schritte aufweist.

Besonders ist dies am Sonntag zu erkennen, welcher im Vergleich zu dem globalen Maximum (Montag) eine Senkung von über 2000 Schritten verzeichnet. Interessant hierbei ist auch der schwingungsähnliche Verlauf des gesamten Graphen.

Auf den Monat ausgeweitet stechen wenige besonderer Tag hervor, da die Verteilung generell stark zwischen lokalen Minima und Maxima schwankt. Dies ist wahrscheinlich der Fall, da für jeden Tag einzeln immer maximal zwölf Datenpunkte zur Verfügung stehen.

Weiter hinausskaliert fällt bei den durchschnittlichen Schritten pro Tag in den ersten neun Monaten des Jahres ein nicht allzu schwankender Wert ins Auge, während Oktober das globale Maximum darstellt. Nach diesem folgt ein recht starker Abfall über die verbleibenden zwei Monate. Im Dezember haben sich die durchschnittlichen Schritte im Vergleich zum Oktober nahezu halbiert. Hierbei ist ein Krankheitsfall am Ende des Jahres die Ursache für diese starke Anomalie, da dieser Monat nunmal im Vergleich einen starken Ausreißer darstellt.

Der größte Ausreißer mit Bezug auf die Gesamtmenge der Datenpunkte ist im Januar zu finden. Ihm ist ein Wert von über 20000 Schritten zugeordnet. Sonstige Tage des Monats sind jedoch zum Teil recht unterdurchschnittlich, wordurch der Mittelwert des Januars im Vergleich mit den anderen Monaten nur recht mittelmäßig beziehungsweise leicht über dem Durchschnitt abfällt.

V Programmcode und verwendete Funktionen

0.1 Code

Das Python Programm ist unter dem Dateinamen 'Auswertung_neu.py' im Git Repository der Abgabe zu finden.

0.2 Verwendete Funktionen im Code

Die folgende Aufzählung enthält vollständig alle im Programm verwendeten Anweisungen.

Imports

- import os
- import matplotlib.pyplot as plt
- import matplotlib.dates as mpl dates
- import matplotlib.ticker as mpl.tick
- from statistics import mean, median, mode, multimode
- import numpy as np
- import csv
- from scipy.optimize import curve_fit

Konstanten

- MONATE
- MONATE_ZAHLEN

Funktionen

- monthToInt(month)
- inttoMonth(month)
- abweichungMedian(data)
- quartile(data)
- dezile(data)
- variationsKoeffizient(data)
- korrelationsKoeffizient(data, data2)

-
- `readFromDataToArray(data, array, dataSet)`
 - `clean(array)`
 - `outputToFile(fileName, array, beschreibung)`
 - `urlist(filePath, fileName, data)`
 - `ranglist(filePath, fileName, data)`
 - `boxWhiskerPlot(filePath, fileName, data, yLabel)`
 - `histogram(data, fileName)`
 - `sine_function(x, A, B ,C, D)`
 - `_fit_sine_curve(zeit, werte, initial_guess)`
 - `scatterPlot(fileName, data, title, yLabel)`
 - `scatterPlotNoLine(fileName, data, title, yLabel)`
 - `stepsPerWeekday(data, fileName)`
 - `stepsPerMonth(data, fileName)`
 - `stepsPerDay(data, fileName)`
 - `stepsPerDayBoxPlot(data, fileName)`
 - `stepsPerMonthBoxPlot(data, fileName)`

VI Dateienverzeichnis

Alle aufgelisteten Dateien sind als solche (jeweils mit passendem Dateityp) im GIT-Repository der Abgabe zu finden.

D1	data-1.csv
D2	data-1_urliste-Elektrizitaetserzeugung.csv
–	data-1_rangliste-Elektrizitaetserzeugung.csv
D3	data-1_urliste-Jahr.csv / data-1_rangliste-Jahr.csv
D4	data-1_urliste-Monat.csv / data-1_rangliste-Monat.csv
D5	data-2.csv
D6	data-2_urliste-Beschaeftigte.csv
–	data-2_rangliste-Beschaeftigte.csv
D7	data-2_urliste-Jahr.csv / data-2_rangliste-Jahr.csv
D8	data-2_urliste-Monat.csv / data-2_rangliste-Monat.csv
D9	data-3-a.csv / data-3-b.csv
D10	data-3_urliste-Ankuenfte.csv / data-3_rangliste-Ankuenfte.csv
D11	data-3_urliste-Uebernachtungen.csv
–	data-3_rangliste-Uebernachtungen.csv
D12	data-3_urliste-Jahr.csv / data-3_rangliste-Jahr.csv
D13	data-3_urliste-Monat.csv / data-3_rangliste-Monat.csv
D14	data-4.csv
D15	data-4_urliste-Anzahl Schritte.csv
–	data-4_rangliste-Anzahl Schritte.csv
D16	data-4_data-4_urliste-Monat.csv / data-4_rangliste-Monat.csv
D17	data-4_urliste-Tag.csv / data-4_rangliste-Tag.csv
D18	data-4_urliste-Wochentag.csv
–	data-4_rangliste-Wochentag.csv

VII Quellenverzeichnis

- Q1:** <https://www-genesis.destatis.de/datenbank/online/table/43311-0002>
(28.01.26, 10:37 Uhr)
- Q2:** <https://share.google/x2zkz0Dal0ZvvHoLG> (24.01.26, 13:49 Uhr)
- Q3:** <https://share.google/6T1bHOUUMX2M0ivD1H> (24.01.26, 13:49 Uhr)
- Q4:** <https://www.bundeskirtschaftsministerium.de/Redaktion/DE/Artikel/Branchenfokus-stahl-und-metall-01.html>
(24.01.26, 14:01 Uhr)
- Q5:** <https://www-genesis.destatis.de/datenbank/online/table/45212-0002>
(31.01.26, 14:51 Uhr)
- Q6:** https://statistik.arbeitsagentur.de/DE/Statischer-Content/Statistiken/Themen-im-Fokus/Transformation/generische-Publikationen/AM-kompakt-Onlinehandel.pdf?__blob=publicationFile (29.01.26 19:02 Uhr)
- Q7:** <https://www-genesis.destatis.de/datenbank/online/table/45412-0014>
(31.01.26, 14:52 Uhr)