

RPTU Kaiserslautern

DATENPROJEKT

Datensatz 3

verfasst von

Jens Bleymehl

427741 (Matrikelnummer)

Isaac Lee Zen Xue

427008 (Matrikelnummer)



Grundlagen und Anwendungen der Wahrscheinlichkeitstheorie

Wintersemester 2023 / 2024

Dozent: Christian De Schryver

Quellenverzeichnis

Videos

- <https://studyflix.de/statistik/korrelationskoeffizient-22>
-  Boxplot erstellen, Median, unteres/oberes Quartil, Minimum, Maximum | Mathe ...
-  Quantile, Quartile berechnen - Statistik
- <https://youtu.be/cmhVP8g442s?si=ermcGS5mnjy4fmyv>
- <https://youtu.be/2rJn0KWolE8?si=ygOUWDgmPFInrxhz>

Webseiten

- <https://rechneronline.de/korrelation/varianz-standardabweichung.php>
- <https://www.superchart.io/blog/how-to-make-a-scatter-plot-in-google-spreadsheet>

R3.1 - Übersicht

Der Datensatz hat seinen Ursprung in der Webseite "Climate Reanalyzer", einer Ressource des Climate Change Institute der University of Maine und konzentriert sich darauf, die durchschnittliche Meeresspiegeltemperatur weltweit in einem Breitengradbereich von 60°N bis 60°S, zu analysieren. Diese Temperaturwerte beziehen sich hier jeweils auf den 25. August des betrachteten Jahres, werden von 1981 bis 2023 zusammengefasst und bilden so die Grundlage dieses Datensatzes. Das Ziel besteht darin, Einblicke in die Temperaturentwicklung der Erde zu erhalten und zu ermitteln, ob diese Entwicklung jahresabhängig sein könnte. Der Datensatz liegt als Kombination zweier Einzeldateien (jeweils im Tabellenformat) mit einer zusätzlichen Zuordnungsgröße „Key“ (Schlüssel) vor. Die Datenquelle befindet sich unter https://climatereanalyzer.org/clim/sst_daily/.

R3.4 - Maßnahmen der Datenbereinigung

Die implementierten Bereinigungsmaßnahmen der Dateien des Datensatzes 3 sind identisch zu jenen, die in Datensatz 2 durchgeführt wurden. Dementsprechend werden einige Werte der Temperatur präsentiert, die sich 'unrealistisch' zu den anderen Werten verhalten (2013°C (1999) und 220°C (2009)), sowie die Werten -1°C (1981, 2018, 2019) und 0°C (2017), die entfernt wurden. Zunächst jedoch haben wir beide Dateien mit den Werten und deren entsprechenden „Keys“ (Schlüsseln) zusammengefügt, sodass eine einzige Datei daraus entsteht. Wir haben uns dazu entschieden, die Keys aus der Analyse auszuschließen, da sie nur als Zuordnungsgröße dienen und eine Variablenbetrachtung der Keys von uns als unnütz erachtet wurde.

R3.6 - Benutzte Softwares und Funktionen

Die Analyse dieses Datensatzes wurde unter Verwendung der Software GitHub, zusammen mit der Programmiersprache Python und -umgebung Visual Studio Code, sowie verschiedenen Tools von Google (wie Google Sheets und Google Docs) durchgeführt.

R3.9 - Modus, Mittelwert und Median

Modus: 20.14°C aus den Jahren 1983, 1988, 1989 und 1990

Mittelwert: Unter Berücksichtigung der verfeinerten Tabelle, folgt daraus der Mittelwert der Jahreszahlen: 2001, und der Temperaturen: $752.94^{\circ}\text{C} / 37 \approx 20.35^{\circ}\text{C}$

Median: 2001 (Jahreszahl), 20.33°C (Temperatur)

R3.10 - Spannweite

Der Zeitraum, in dem die Temperaturen aufgezeichnet wurden, ist zwischen 1982 und 2023, wobei 1981, 1999, 2009, und 2017-2019 ausgeschlossen werden. Die Temperaturen variieren zwischen einem Minimum von 19.91°C und einem Maximum von 21.09°C .

R3.11 - Mittlere Abweichung vom Median

Die mittlere Abweichung vom Median der Jahreszahlen beträgt 10.32.

Die mittlere Abweichung vom Median der Temperaturen beträgt 0.21°C.

R3.12 - Stichprobenvarianz

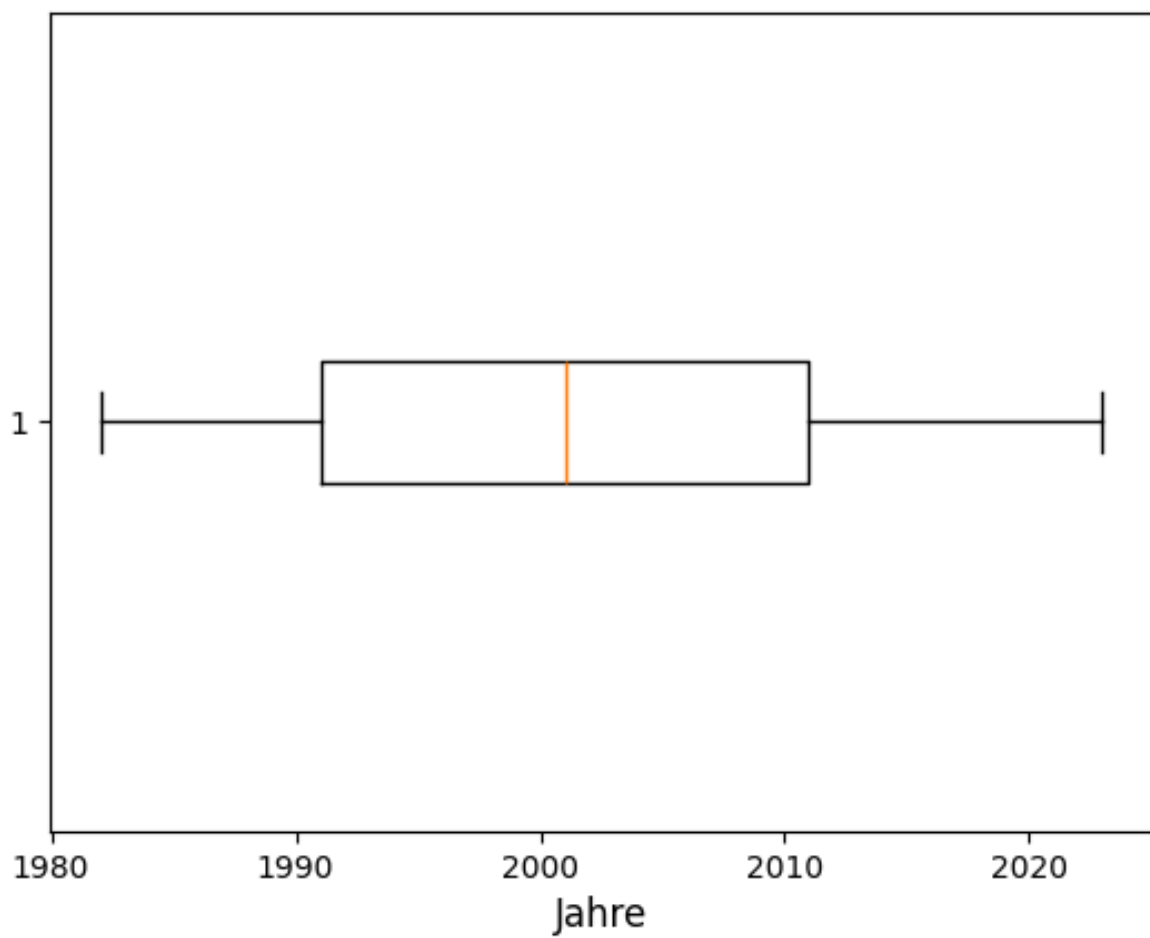
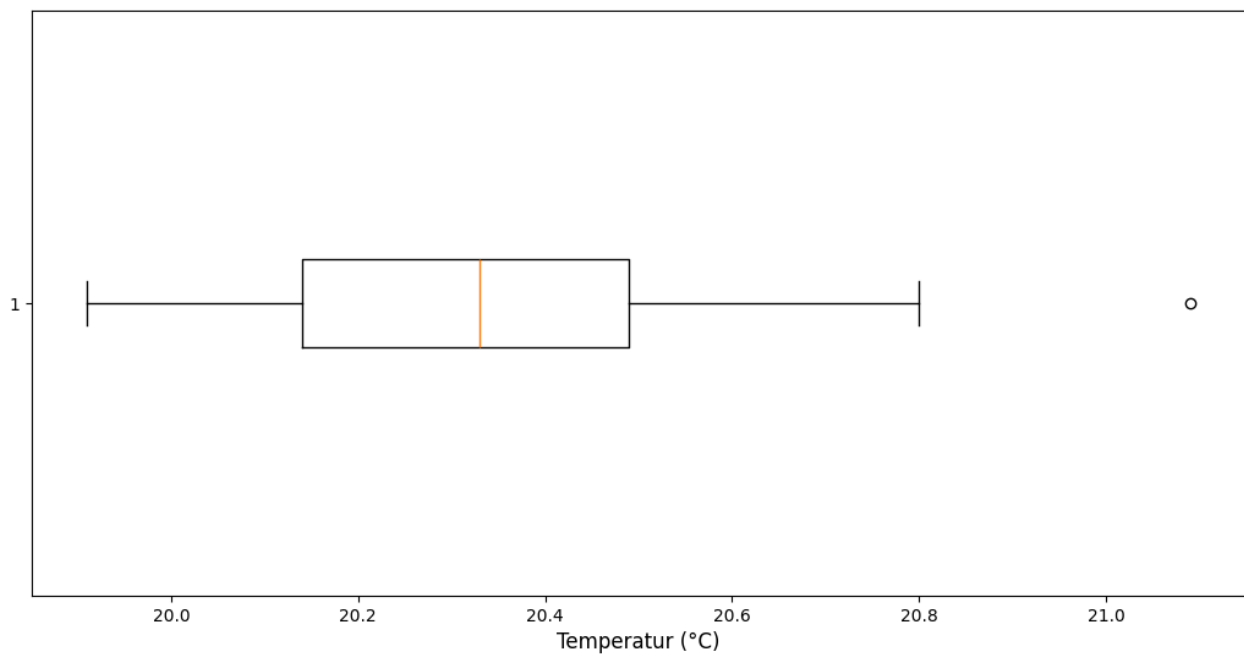
Die Stichprobenvarianz der Jahreszahlen beträgt $s^2 \approx 147.9730 \Rightarrow$ Standardabweichung $s \approx 12.16$.

Die Stichprobenvarianz der Temperaturen beträgt $s^2 \approx 0.0686 \Rightarrow$ Standardabweichung $s \approx 0.262$.

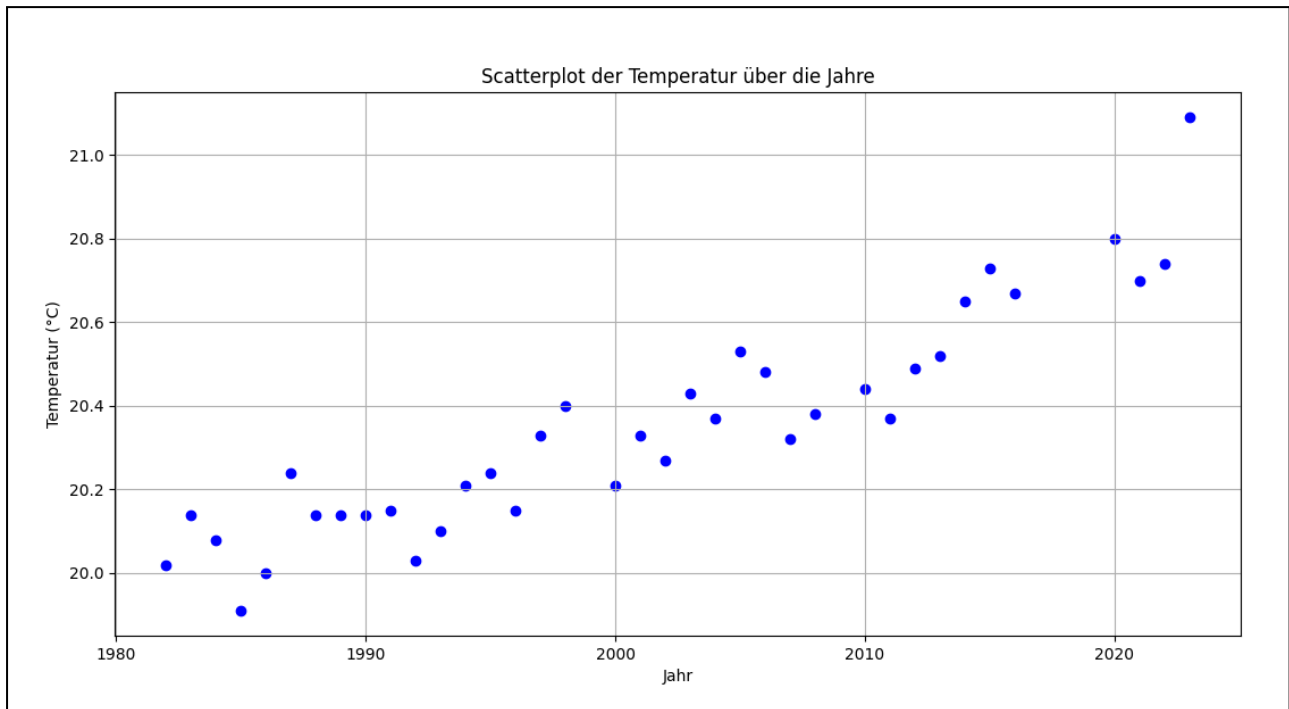
R3.13 - Variationskoeffizient

Der Variationskoeffizient der Jahreszahlen beträgt ≈ 0.006079 ($\approx 0.61\%$) und der Temperaturen ≈ 0.01287 ($\approx 1.28\%$).

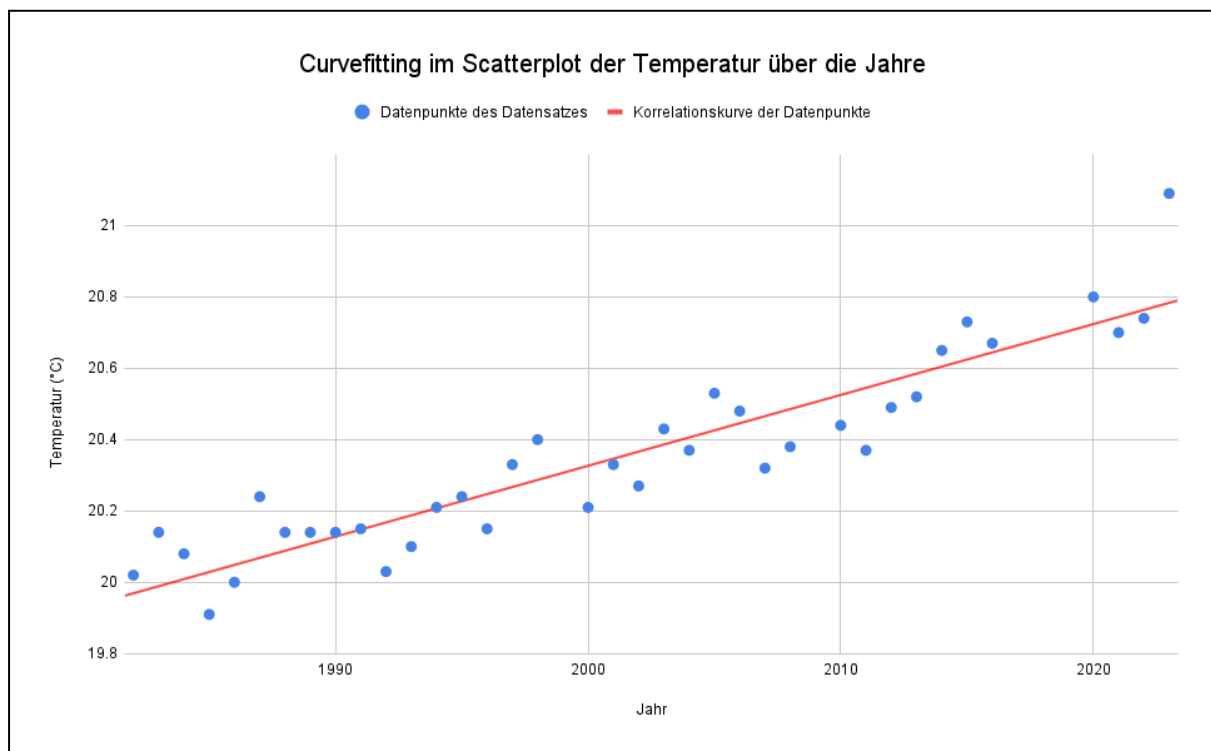
R3.14 - Box-Whisker-Plot der Jahreszahlen und Temperaturen



R3.15 - Scatterplot



R3.16 / R3.17 / R3.18 - Curvefitting im Scatterplot



Mathematische Funktion der Korrelationskurve (linear): $0.0199x + 19.4$

R3.20 - Quartile und Dezile (mit linearer Interpolation)

Jahreszahlen:

$$Q1 (00.25) = 1990.5$$

$$Q2 (00.50) = 2001 \text{ (Median)}$$

$$Q3 (00.75) = 2011.5$$

$$D1 (00.10) = 1984.5$$

$$D2 (00.20) = 1988.5$$

$$D3 (00.30) = 1992.5$$

$$D4 (00.40) = 1996.5$$

$$D5 (00.50) = 2001 \text{ (Median)}$$

$$D6 (00.60) = 2004.5$$

$$D7 (00.70) = 2009$$

$$D8 (00.80) = 2013.5$$

$$D9 (00.90) = 2020.5$$

Temperaturen:

$$Q1 (00.25) = 20.14^{\circ}\text{C}$$

$$Q2 (00.50) = 20.33^{\circ}\text{C (Median)}$$

$$Q3 (00.75) = 20.505^{\circ}\text{C}$$

$$D1 (00.10) = 20.025^{\circ}\text{C}$$

$$D2 (00.20) = 20.14^{\circ}\text{C}$$

$$D3 (00.30) = 20.15^{\circ}\text{C}$$

$$D4 (00.40) = 20.24^{\circ}\text{C}$$

$$D5 (00.50) = 20.33^{\circ}\text{C (Median)}$$

$$D6 (00.60) = 20.375^{\circ}\text{C}$$

$$D7 (00.70) = 20.46^{\circ}\text{C}$$

$$D8 (00.80) = 20.59^{\circ}\text{C}$$

$$D9 (00.90) = 20.735^{\circ}\text{C}$$

R3.21 - Quartilsabstand $R_{Q0.5}$

Es gilt: $R_{Q0.5} = Q_3 - Q_1$. Somit beträgt der Quartilsabstand der Jahreszahlen: $R_{Q0.5}$: 21 und der Temperaturen: 0.365°C .

R3.22 - Kovarianz

Die Kovarianz zwischen Temperaturdaten und Jahreszahlen beträgt: 3.1006.

R3.23 - Korrelationskoeffizient

Der Korrelationskoeffizient zwischen Temperaturdaten und Jahreszahlen beträgt: 0.98822.

R3.19 - Fazit

Die Daten weisen einen sehr hohen Korrelationskoeffizienten auf (sehr knapp an 1), was dafür spricht, dass Temperatur und Jahreszahlen stark miteinander korrelieren. Es lässt sich hierdurch außerdem folgende Aussage treffen: je höher die Jahreszahl, desto höher wird im Mittel wahrscheinlich die Temperatur sein. Diese ist von 1982 bis 2023 um ca. 1°C angestiegen.