L02 Projekt – Datenanalyse

Modul Grundlagen und Anwendung der Wahrscheinlichkeitstheorie WS 22/23

Gruppe 07

Nico Denkiewicz

Lauraine Maheva Tagakou Signe

Jakob Bund

Aufgabe 1

Der Datensatz beinhaltet die Entwicklung des insgesammten Verbraucherpreisindex mit Beginn im Januar 2017 bis zum September 2022, in monatlicher Auflösung. “Der Verbraucherpreisindex misst monatlich die **durchschnittliche Preisentwicklung** aller Waren und Dienstleistungen, die private Haushalte in Deutschland für Konsumzwecke kaufen.” Die Daten wurden durch das statistische Bundesamt erhoben und stellen jeweils die Veränderungsrate des Verbrauchspreisindex für Deutschland gegenüber des Vorjahresmonats in % dar. Die Daten liegen im .csv Format vor wobei als Trennzeichen das Semikolon und als Dezimaltrennzeichen das “,” verwendet wird. Allgemein ist bekannt, dass der Verbraucherpreis in den Jahren 2020 bis 2022 durch Faktoren, wie die Coronapandemie und den russischen Angriffskrieg in der Ukraine starken Schwankungen unterlagen und zuletzt stark gestiegen sind.

https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Tabellen/Verbraucherpreise-12Kategorien.html#236118 (30.Januar2023)

Zur Auswertung der Daten wurde ein Python-Skript (Aufgabe1.py) inklusive der Module pandas, numpy, scipy und matplotlib verwendet. Weiterhin fand die Software LibreOffice Calc in der Version 7.4.5.1 zur Bereinigung des Datesatzes Anwendung.

Die Betrachtung der, in der unteren Tabelle aufgeführten Eigenschaften der Monats-, und Jahresdaten ist in unserem Fall uninteressant, da die Datenpunkte über den betrachteten Zeitraum Gleichverteilt sind. Interessant ist lediglich der arithmetische Median der Zeitdaten im November 2019, also der Halbzeit des betrachteten Zeitraumes und die Spannweite von fünf Jahren und neun Monaten, also die Länge des betrachteten Zeitraumes.

Aufgabe	Eigenschaft	Verbraucherpreisindex
R1.7	Modus	1,4
	Arithmetisches Mittel	2,401
	Median	1,6
R1.8	Spannweite	10,3
R1.9	Mittlere Abweichung vom Median	0,4
R1.10	Stichprobenvarianz	5,052
R1.11	Variationskoeffizient	0,929

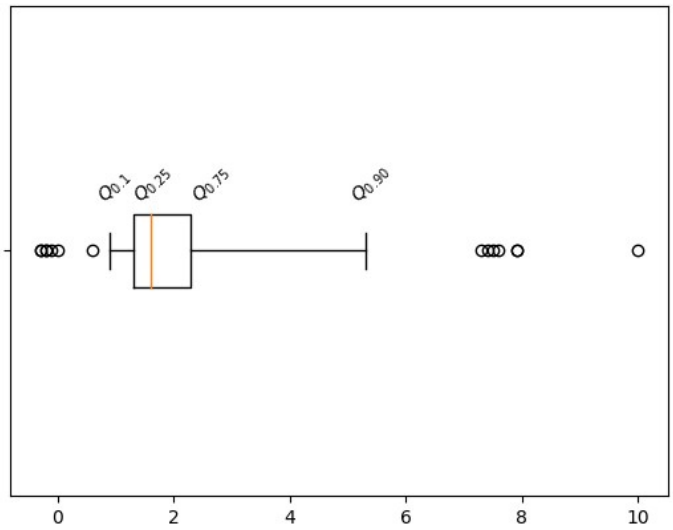

Figure 1: R1.12 Box-Whisker-Plot

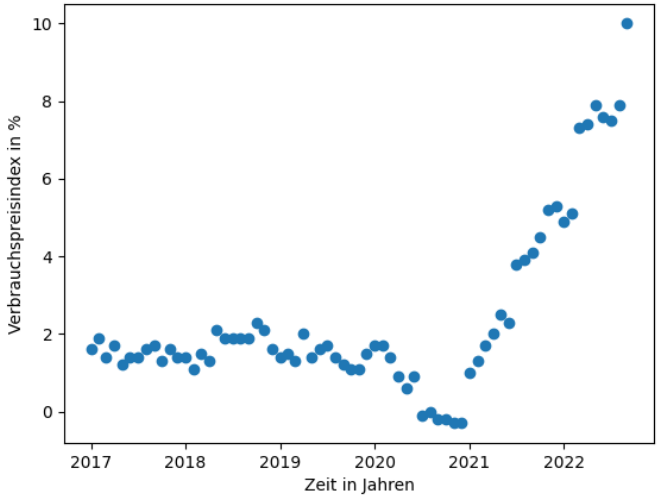

Figure 2: R1.13 Scatterplot

Quantil	Wert
0.1	0.84
0.2	1.2
0.25	1.3
0.3	1.4
0.4	1.42
0.5	1.6
0.6	1.7

0.7	2.0
0.75	2.3
0.8	3.84
0.9	5.7

Zusammenfassung Datensatz 1

Die Daten bilden den Verbraucherpreisindex der Bundesrepublik Deutschland über einen Zeitraum von Januar 2017 bis September 2022 , in monatlicher Auflösung ab. Wie im Boxplot erkennbar ist streuen die Daten stark nach oben, was Folge der jüngst stark gestiegenen Verbraucherpreisindexes ist. Im Scatterplot sind außerdem die starken Schwankungen erkennbar, denen der Verbraucherpreisindex seit Anfang 2020 (Beginn der Corona Pandemie). Während er vor der Pandemie im Bereich des Median von 1,6 lag ist er jüngst auf 10 (September 2022) gestiegen, in Folge der globalen Lieferengpässe und des russischen Angriffs auf die Ukraine.

Aufgabe 2

Der Datensatz 2 liegt als einzelne .csv Datei vor wobei als Trennzeichen das Semikolon und als Dezimaltrennzeichen das “,” verwendet wird. Inhaltlich entspricht er dem Datensatz eins, welcher die Entwicklung des Verbraucherpreisindex zwischen Januar 2017 und September 2022 abbildet. Der Unterschied besteht in der Veränderung des Wertes einiger Datenpunkte, diese enthalten teilweise nicht näher definierte Strings oder den definierten Wert nan (“not a number”). Die Quelle entspricht der, des ersten Datensatzes, wobei die Urheberschaft der veränderten Datenpunkte nicht näher beschrieben wird.

Zur Bereinigung der Daten wurden im Datensatz händisch die, strings enthaltenden, Datenpunkte mit dem Python-lesbaren Wert ‘nan’ ersetzt. Weiterhin wurde das Semikolon als Trennzeichen mit Kommata ersetzt und der ‘.’ Als Dezimaltrennzeichen eingefügt. Der bereinigte Datensatz liegt als ‘data.csv’ vor. Alle beschriebenen Arbeiten wurden in der Software LibreOffice Calc in der Version 7.4.5.1 durchgeführt. Zur Auswertung der Daten wurde ein Python-Skript (Aufgabe2.py) inklusive der Module pandas, numpy, scipy und matplotlib verwendet. Um die Datensätze auswerten zu können werden die ‘nan’-Datenpunkte mit der Dataframe-Funktion ‘dropna()’ ignoriert.

Aufgabe	Eigenschaft	Verbraucherpreisindex
R2.8	Modus	1,4
	Arithmetisches Mittel	2,255
	Median	1,6
R2.9	Spannweite	10,3
R2.10	Mittlere Abweichung vom Median	0,4
R2.11	Stichprobenvarianz	4,377
R2.12	Variationskoeffizient	0,920

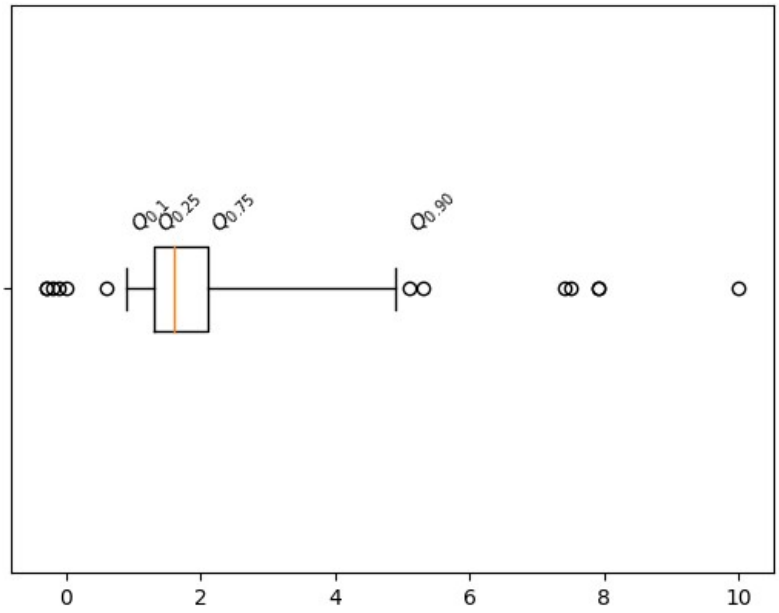

Figure 3: R2.13 Box-Whisker-Plot

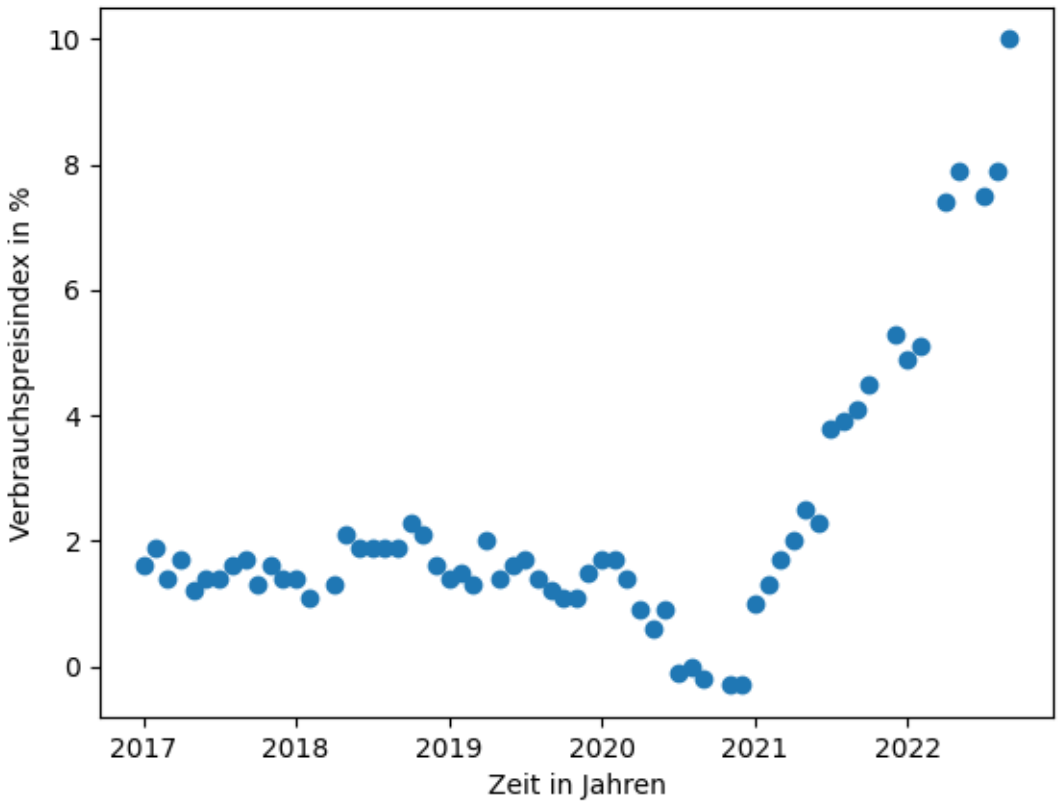

Figure 4: R2.14 Scatterplot

Wie der erste Datensatz auch bildet der zweite die Entwicklung des Verbraucherpreisindex Deutschlands zwischen Januar 2017 und September 2022, in % ab. Lediglich fünf Datenpunkte wurden verändert, weswegen die beobachteten Metriken auf dem veränderten Datensatz nur wenige, bis keine Abweichungen aufzeigen.

Quantil	Wert
0.1	0.9
0.2	1.2
0.25	1.3
0.3	1.4
0.4	1.4
0.5	1.6
0.6	1.7
0.7	1.91
0.75	2.1

0.8	2.38
0.9	5.04

Aufgabe 3

Der Datensatz 3 liegt in zwei getrennten .csv Dateien vor, wobei in Datensatz data-3-a.csv die Monate inklusive der Jahre und in Datensatz data-3-b.csv der Verbraucherpreisindex gespeichert sind. Beiden Datensätzen gemein ist die Spalte ‘Key’, welche, beginnend bei 1, den Datenpunkten aufsteigende integer Werte zuweist. Der zugrundeliegende Datensatz und die Codierung der .csv Dateien entsprechen den Aufgabenteilen eins und zwei. Der Unterschied besteht in der Veränderung des Wertes einiger Datenpunkte, diese enthalten teilweise nicht näher definierte Strings oder den definierten Wert nan (“not a number”). Die Quelle entspricht der, des ersten Datensatzes, wobei die Urheberschaft der veränderten Datenpunkte nicht näher beschrieben wird.

Wie bereits in Aufgabe 2 wurden die Daten in LibreOffice Calc bereinigt, die Codierung umgestellt und die fehlenden Datenpunkte durch ‘nan’ ersetzt. Die Zusammenführung der zwei Datensätze geschah auch hier.

Zur weiteren Auswertung der Datei ‘data.csv’ wurde ein Python-Skript (Aufgabe3.py) inklusive der Module pandas, numpy, scipy und matplotlib verwendet.

Aufgabe	Eigenschaft	Verbraucherpreisindex
R3.9	Modus	1,4
	Arithmetisches Mittel	2,255
	Median	1,6
R3.10	Spannweite	10,3
R3.11	Mittlere Abweichung vom Median	0,4
R3.12	Stichprobenvarianz	4,377
R3.13	Variationskoeffizient	0,921

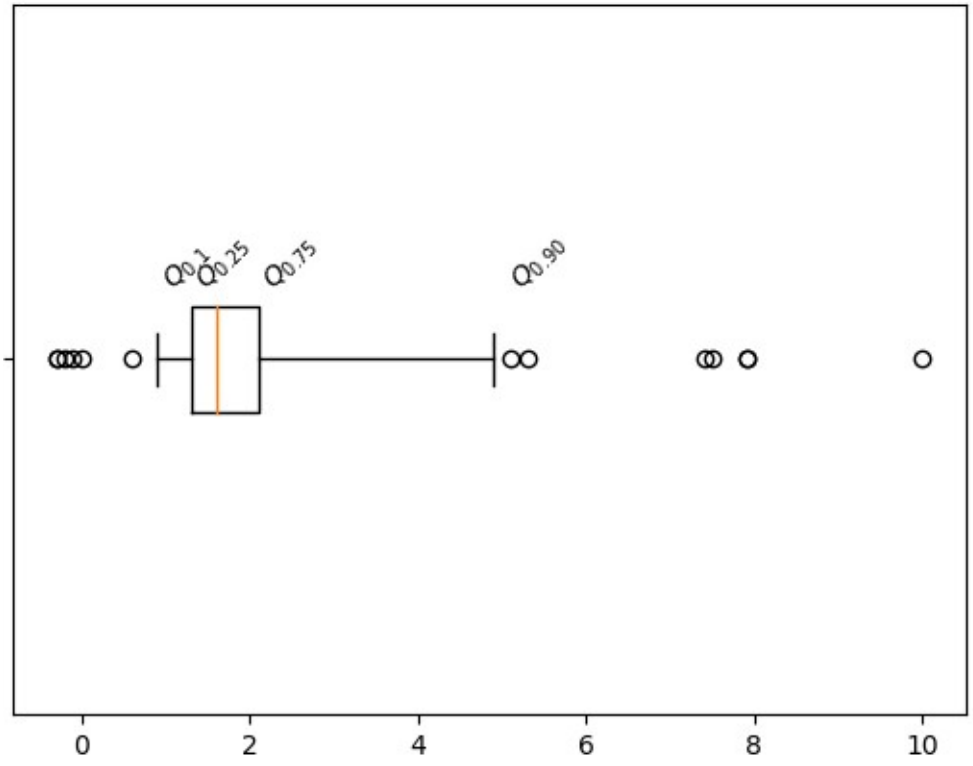

Figure 5: R3.14 Box-Whisker-Plot

Quallangabe:

Datensatz: https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Tabellen/Verbraucherpreise-12Kategorien.html#236118 (30.Januar2023)