GWAT L02 Gruppe 11

DATENANALYSE GWAT 2023 WS

Max Umlauf

Semester 3

Datensatz 1

R1.1: Datensatz 1

Bei dem hier vorliegenden Datensatz, data-1, handelt es sich um eine Statistik des Verbraucherpreisindex von den Jahren 1991 bis 2021. Die Daten stammen von dem Statistischen Bundesamt vom 10.10.2022.
Die vorliegende Tabelle liegt im CSV-Dateiformat vor. Es handelt sich hierbei um eine Datei.
Bei der Tabelle handelt es sich ebenso um eine GENESIS-Tabelle: 61111-0001.
In der Tabelle gibt es eine Spalte wo die Jahre von 1991 bis 2021 gegeben sind und in einer Spalte daneben ist der jeweilige Verbraucherpreisindex zu finden.

R1.2 Skalenvarianten

Bei den hier vergebenen Variablen handelt es sich ausschließlich um Verhältnisskalen, außer bei der Variable „years“. Dort handelt es sich um eine Intervallskala

R1.3 genutzte Software und Funktionen

Die Daten wurden mit der Software DataSpell und der Programmiersprache Python analysiert. Python wurde ebenso mit den Modulen Pandas, Numpy und Matpltolib, sowie Scipy, Statistics und Statsmodels erweitert. Die Tabelle wurde in Excel geöffnet und dort für die Datenanalyse angepasst. Die Anpassungen beinhalteten die Formatierung in UTF-8 und das Ändern von „,“ zu „.“ Um die Dezimalstellen abzutrennen.

R1.7 Modus, arithmetischer Mittelwert und Medien der Variablen

Der Modus, Median und Arithmetische Mittelwert wurde in die folgenden Variablen gespeichert, um die Resultate einfacher in zukünftigem Code wieder aufzurufen uns auszulesen. Alle folgenden Variablen wurden aus dem Verbraucherpreisindex heraus gerechnet. Für die Ergebnisse wurden folgende Variablen angelegt:

- vpi_mean = 88,252 -> arithmetischer Mittelwert
- vpi_mode = 65,5 -> Modus
- vpi_median = 87,6 -> Median

R1.8 Spannweite der variablen

vpi_spannweite = 43,6

R1.9 Mittlere Abweichung vom Median

vpi_mam = 14,826

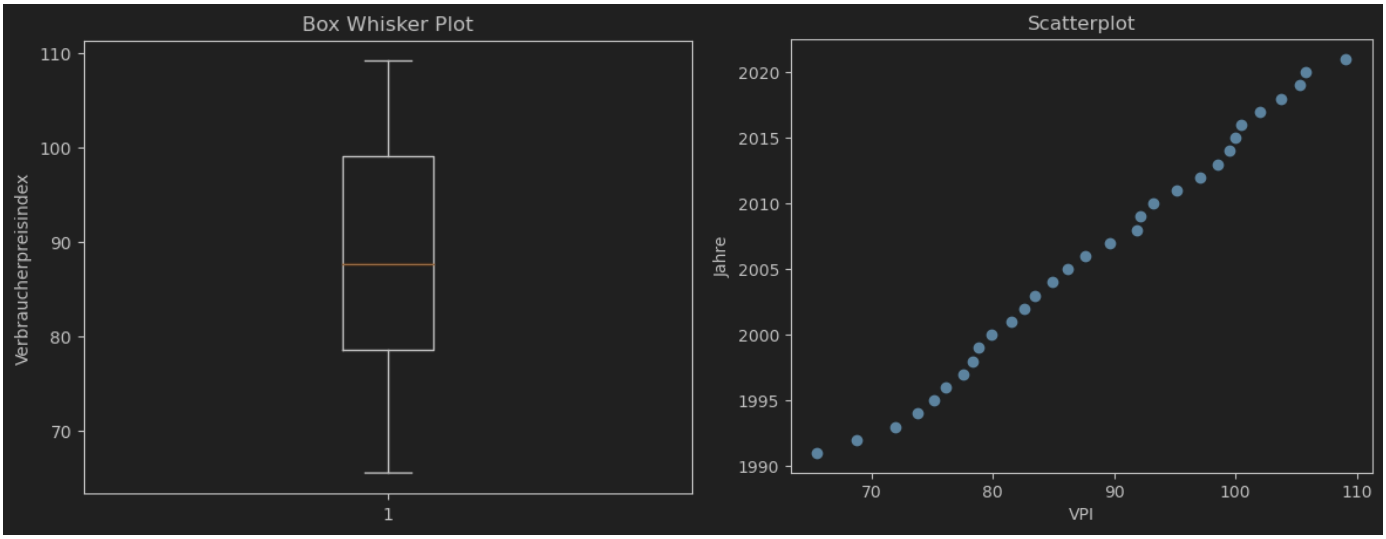
R1.10 Stichprobenvarianz

vpi_stichprobenvarianz = 140,187

R1.11 Variationskoeffizient

vpi_variationskoeffizient = 13,638

R1.12 / R1.13 Box-Whisker Plot und Scatter Plot



R1.14 Textuelle Beschreibung der Daten

Die Tabelle hat 32 Zeilen und 2 Spalten. In einer Spalte befinden sich die Jahre von 1991 bis 2021 und in der anderen Spalte der Verbraucherpreisindex. In 31 der 32 Zeilen befinden sich die Daten und in der ersten Zeile eine Überschrift, was sich in der jeweiligen Spalte befindet.
Die Tabellenspalten werden durch „,“ voneinander getrennt, sowie die Nachkommastellen durch „.“ Ebenso wurde die Tabelle in die UTF-8 Kodierung konvertiert und im .csv Dateiformat abgespeichert.
Wenn man sich den Verlauf der Werte über die einzelnen Jahre anschaut, sieht man, dass der Verbraucherpreisindex von Jahr zu Jahr steigt. So betrug dieser im Jahr 1991 noch 73,8 und im Jahr 2021 109,1. Im Jahr 2015 betrug er 100. Wenn man sich den Graphen im Scatterplot anschaut, lässt sich eine annähernd lineare Entwicklung erkennen.

R1.15 Quartile und Dezile

Quartil 1: q1 = 78,55
Quartil 2: q1 = 87,6 -> entspricht ebenso dem Median
Quartil 3: q3 = 99,0

Dezil 1: dz1 = 73.8
Dezil 2: dz2 = 77.6
Dezil 3: dz3 = 79.9
Dezil 4: dz4 = 83.5
Dezil 5: dz5= 87.6 -> entspricht ebenso dem Median
Dezil 6: dz6 = 92.2
Dezil 7: dz7 = 97.1
Dezil 8: dz8 = 100.0
Dezil 9: dz9 = 103.8

Das vierte Quartil sowie das zehnte Dezil sind hier zu vernachlässigen, da diese dem letzten Wert im Datensatzes entsprechen, was zugleich das Maximum ist -> 109,1

R1.16 Quartilsabstand

qt_Abstand = 20,45

Datensatz 2

R2.1 Datensatz 2

Bei dem hier vorliegenden Datensatz, data-1, handelt es sich um eine Statistik des Verbraucherpreisindex von den Jahren 1991 bis 2021. Die Daten stammen von dem Statistischen Bundesamt vom 10.10.2022.
Die vorliegende Tabelle liegt im CSV-Dateiformat vor. Es handelt sich hierbei um eine Datei.
Bei der Tabelle handelt es sich ebenso um eine GENESIS-Tabelle: 61111-0001.
In der Tabelle gibt es eine Spalte wo die Jahre von 1991 bis 2021 gegeben sind und in einer Spalte daneben ist der jeweilige Verbraucherpreisindex zu finden.
Der vorliegende Datensatz besitzt fehlende Daten, sowie einige Zahlen in falscher Form. Diese müssen für eine weitere Analyse noch bereinigt werden.
Ebenso gibt es zwei Jahresangaben, z.b. 2105 und 1795. Es lässt sich darauf schließen, dass es sich hierbei um einen Tippfehler handelt.

R2.2 Maßnahmen zur Datenbereinigung

Um den Datensatz in eine geeignete Form zu bringen, wurden die Semikolons durch Kommata ersetzt, um die Spalten voneinander zu trennen. Ebenso wurden die Dezimaltrennstellen durch Punkte ersetzt. Fehlende Werte, so wie Werte ohne Zahlen wurden durch „NaN“ ersetzt, damit diese bei der Analyse nicht berücksichtigt werden. Falsche Jahresangaben wurden berichtigt, so wurde aus dem Jahr 1795 das Jahr 1995 und aus dem Jahr 2105 das Jahr 2005.
Um die fehlenden Werte zu ignorieren wurde die Pandas Funktion dpropna() benutzt.

R2.4 verwendete Software und Funktionen

Die Daten wurden mit der Software DataSpell und der Programmiersprache Python analysiert. Python wurde ebenso mit den Modulen Pandas, Numpy und Matpltolib, sowie Scipy, Statistics und Statsmodels erweitert. Die Tabelle wurde in Excel geöffnet und dort für die Datenanalyse angepasst. Die Anpassungen beinhalteten die Formatierung in UTF-8 und das Ändern von „,“ zu „.“ Um die Dezimalstellen abzutrennen.

R2.8 Modus, arithmetischer Mittelwert und Medien der Variablen

Der Modus, Median und Arithmetische Mittelwert wurde in die folgenden Variablen gespeichert, um die Resultate einfacher in zukünftigem Code wieder aufzurufen uns auszulesen. Alle folgenden Variablen wurden aus dem Verbraucherpreisindex heraus gerechnet. Für die Ergebnisse wurden folgende Variablen angelegt:

- vpi_mean = 89,53 -> arithmetischer Mittelwert
- vpi_mode = 68,1 -> Modus
- vpi_median = 89,6 -> Median

R2.9 Spannweite der variablen

vpi_spannweite = 40,3

R2.10 Mittlere Abweichung vom Median

vpi_mam = 15,491

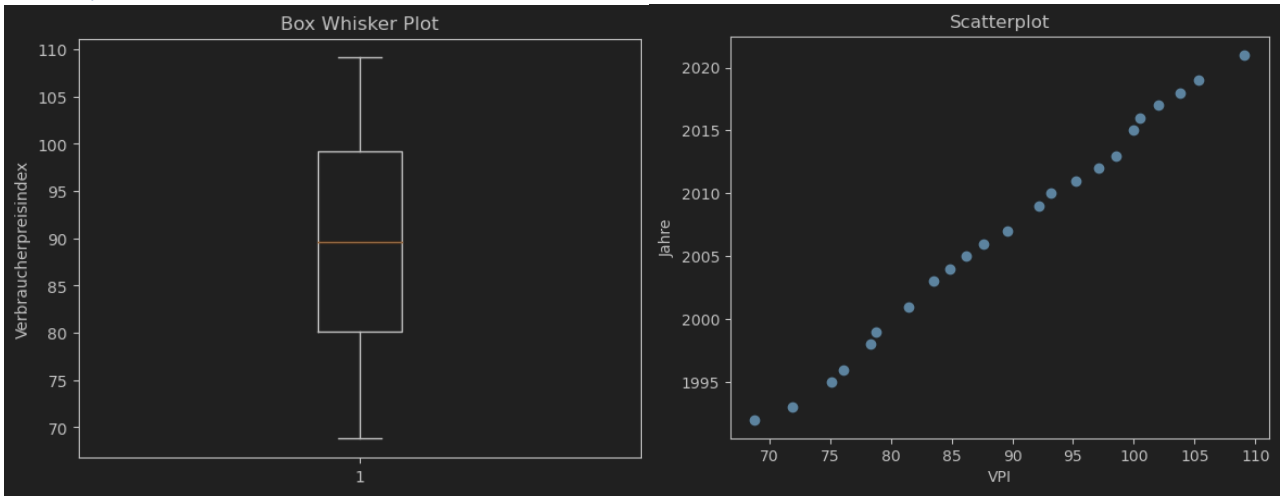
R2.11 Stichprobenvarianz

vpi_stichprobenvarianz = 126,903

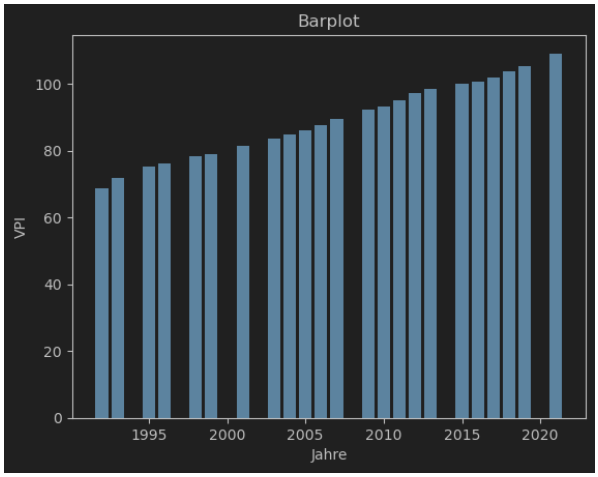
R2.12 Variationskoeffizient

vpi_variationskoeffizient = 12,865

R2.13 / R2.14 Box-Whisker Plot und Scatter Plot



R2.15 Balkendiagramm



In der hier vorliegenden Grafik ist der Verbraucherpreisindex der jeweiligen Jahre als Balkendiagramm gezeigt. Auf der Y-Achse findet sich der Verbraucherpreisindex und auf der X-Achse das jeweilige Jahr.

R2.16 Textuelle Beschreibung der Daten

Wenn man sich den Verlauf der Werte über die einzelnen Jahre anschaut, sieht man, dass der Verbraucherpreisindex von Jahr zu Jahr steigt. So betrug dieser im Jahr 1991 noch 73,8 und im Jahr 2021 109,1. Im Jahr 2015 betrug er 100. Wenn man sich den Graphen im Scatterplot anschaut, lässt sich eine annähernd lineare Entwicklung erkennen. Des weiteren sind in dem unangepassten Datensatz einige nicht angepasste und nicht verwendbare Werte in falschen Formen angegeben. So sind beispielsweise Namen oder leere Zellen. Auffällig sind zwei Jahresangaben, welche nicht in das Schema der restlichen Daten passen. So gibt es das Jahr 1795 und 2105. Diese wurden

R2.17 Quartile und Dezile

Quartil 1: q1 = 80,15
Quartil 2: q1 = 89,6 -> entspricht ebenso dem Median
Quartil 3: q3 = 99,25

Dezil 1: dz1 = 75,3
Dezil 2: dz2 = 78,5
Dezil 3: dz3 = 82,7
Dezil 4: dz4 = 85,94
Dezil 5: dz5= 89.6 -> entspricht ebenso dem Median
Dezil 6: dz6 = 93,6
Dezil 7: dz7 = 97.66
Dezil 8: dz8 = 100.3
Dezil 9: dz9 = 103,44

Das vierte Quartil sowie das zehnte Dezil sind hier zu vernachlässigen, da diese dem letzten Wert im Datensatzes entsprechen. -> 109,1

R2.18 Quartilsabstand

qt_Abstand = 20,45

Datensatz 3

R3.1 Datensatz 3

Bei dem hier vorliegenden Datensatz, data-3, handelt es sich um eine Statistik des Verbraucherpreisindex von den Jahren 1991 bis 2021. Die Daten stammen von dem Statistischen Bundesamt vom 10.10.2022.
Die vorliegenden Tabellen liegen im CSV-Dateiformat vor. Es liegen 2 Tabellen vor.
In der Tabelle data-3a gibt es eine Spalte mit einem Key und eine zweite Spalte mit dem dazugehörigen Verbraucherpreisindex. In der zweiten Tabelle, data-3b befindet sich ebenso eine Spalte mit dem gleichen Key und eine zweite Spalte mit den dazugehörigen Jahren. Wenn man die beiden Tabellen miteinander kombiniert, ergibt sich daraus eine neue Tabelle, welche das gleiche Schema wie die Datensätze zuvor verfolgt.
Der vorliegende Datensatz besitzt fehlende Daten, sowie einige Zahlen in falscher Form. Diese müssen für eine weitere Analyse noch bereinigt werden.
Ebenso gibt es zwei Jahresangaben, z.b. 2105 und 1795. Es lässt sich darauf schließen, dass es sich hierbei um einen Tippfehler handelt.

R3.3 Maßnahmen zur Datenbereinigung

Zur weiten Analyse wurden beide Tabellen in Excel zusammengeführt, was die Ersetzung des Keys durch die passenden Jahre beinhaltete. Ebenso wurden die Semikolons durch Kommata ersetzt, um die Spalten voneinander zu trennen. Ebenso wurden die Dezimaltrennstellen durch Punkte ersetzt. Fehlende Werte, so wie Werte ohne Zahlen wurden durch „NaN“ ersetzt, damit diese bei der Analyse nicht berücksichtigt werden. Falsche Jahresangaben wurden berichtigt, so wurde aus dem Jahr 1795 das Jahr 1995 und aus dem Jahr 2105 das Jahr 2005.
Um die fehlenden Werte zu ignorieren wurde die Pandas Funktion dpropna() benutzt.

R3.4 verwendete Software und Funktionen

Die Daten wurden mit der Software DataSpell und der Programmiersprache Python analysiert. Python wurde ebenso mit den Modulen Pandas, Numpy und Matpltolib, sowie Scipy, Statistics und Statsmodels erweitert. Die Tabelle wurde in Excel geöffnet und dort für die Datenanalyse angepasst. Die Anpassungen beinhalteten die Formatierung in UTF-8 und das Ändern von „,“ zu „.“ Um die Dezimalstellen abzutrennen.

R3.9 Modus, arithmetischer Mittelwert und Medien der Variablen

Der Modus, Median und Arithmetische Mittelwert wurde in die folgenden Variablen gespeichert, um die Resultate einfacher in zukünftigem Code wieder aufzurufen uns auszulesen. Alle folgenden Variablen wurden aus dem Verbraucherpreisindex heraus gerechnet. Für die Ergebnisse wurden folgende Variablen angelegt:

- vpi_mean = 89,186 -> arithmetischer Mittelwert
- vpi_mode = 68,8 -> Modus
- vpi_median = 88,6 -> Median

R3.10 Spannweite der variablen

vpi_spannweite = 40,3

R3.11 Mittlere Abweichung vom Median

vpi_mam = 14,974

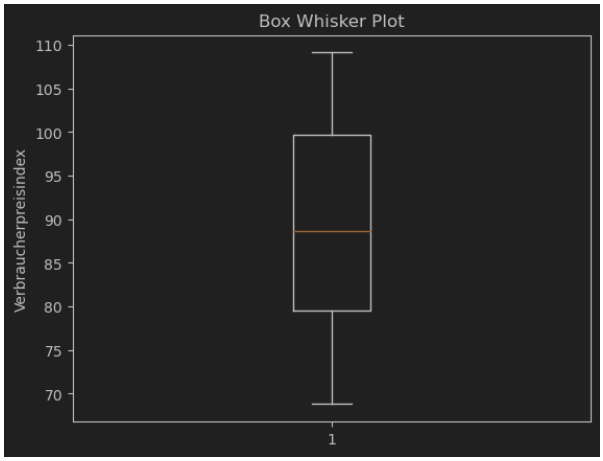
R3.12 Stichprobenvarianz

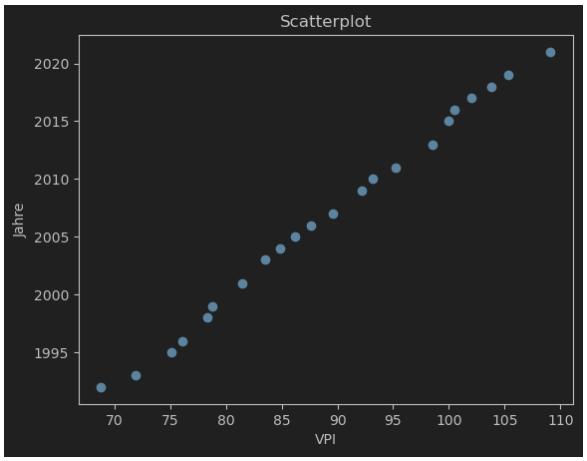
vpi_stichprobenvarianz = 129,948

R3.13 Variationskoeffizient

vpi_variationskoeffizient = 13,082

R3.14 Box-Whisker / Scatter-Plot





R3.19 Textuelle Beschreibung der Daten

Wenn man sich den Verlauf der Werte über die einzelnen Jahre anschaut, sieht man, dass der Verbraucherpreisindex von Jahr zu Jahr steigt. So betrug dieser im Jahr 1991 noch 73,8 und im Jahr 2021 109,1. Im Jahr 2015 betrug er 100. Wenn man sich den Graphen im Scatterplot anschaut, lässt sich eine annähernd lineare Entwicklung erkennen.

Bei diesem Datensatz ist auffällig, dass einige Angaben des Verbraucherpreisindex in falscher Form gegeben sind. So sind teilweise die Zellen leer oder es stehen Namen an Stelle von Zahlen.

Des weiteren sind die Jahresangaben durch einen Key gegeben, welcher in einer zweiten Tabelle den passenden Wert für das Jahr zeigt. Die Key-Werte sind von 1 bis 31 durchnummeriert.

R3.20 Quartile und Dezile

Quartil 1: q1 = 79,475
Quartil 2: q1 = 88,6 -> entspricht ebenso dem Median
Quartil 3: q3 = 99,625

Dezil 1: dz1 = 75,199
Dezil 2: dz2 = 78,399
Dezil 3: dz3 = 82,1
Dezil 4: dz4 = 85,42
Dezil 5: dz5= 88.6 -> entspricht ebenso dem Median
Dezil 6: dz6 = 92,8

Dezil 7: dz7 = 97.51
Dezil 8: dz8 = 100.4
Dezil 9: dz9 = 103,62

Das vierte Quartil sowie das zehnte Dezil sind hier zu vernachlässigen, da diese dem letzten Wert im Datensatzes entsprechen. -> 109,1

R3.21 Quartilsabstand

Qt_abstand = 20,15

Zu R1.1, R2.1, R3.1

Es liegt ein Ordner vor, in welchem alle Rohdaten der zu analysierenden Datensätze vorhanden sind. Diese liegen in mehreren Dateiformaten vor, als .csv, sowie flat .csv, als .xml und .xlsx

Quelle der Daten:
Statistisches Bundesamt, Datenportal "Genesis"
https://www-genesis.destatis.de

Verbraucherpreisindex https://www-genesis.destatis.de/genesis/online?operation=result&code=61111-0001&deep=true#abreadcrumb

Datensatz 4

R4.3 Ergriffene Maßnahmen zur Datenbereinigung

Die liegen vollständig vor. Es wurde lediglich zur vereinfachten Analyse eine Überschrift für jede Spalte eingefügt

R4.4 Verwendete Software und Funktionen

Die Daten wurden mit der Software DataSpell und der Programmiersprache Python analysiert. Python wurde ebenso mit den Modulen Pandas, Numpy und Matpltolib, sowie Scipy, Statistics und Statsmodels erweitert. Um die Daten zu erstellen wurde aus Numpy der Befehl .random() verwendet. So wurde ein 1D Datensatz erstellt mit 130 zufälligen Werten zwischen 1 und 10000

R4.5Modus, arithmetischer Mittelwert und Medien der Variablen

Der Modus, Median und Arithmetische Mittelwert wurde in die folgenden Variablen gespeichert, um die Resultate einfacher in zukünftigem Code wieder aufzurufen und auszulesen.

- value_mean = 5262,631 -> arithmetischer Mittelwert
- value_mode = 3047 -> Modus
- value_median = 5388,5 -> Median

R4.6 Stichprobenvarianz

Values_stichprobenvarianz = 8421269.571

R4.7 Box-Whisker-Plot