RPTU Kaiserslautern

Projekt Datenanalyse

Grundlagen und Anwendung der
Wahrscheinlichkeitstheorie

Gruppe 2: Sara Soheili

WS 2024/2025
Dozent: Christian De Schryver

Quellenverzeichnis:

Datensätze 1-3: www-genesis.destatis.de

Datensatz 1

R1.1 - Beschreibung des Datensatzes

Der Datensatz besteht aus einer UTF-8-kodierten CSV-Datei, die eine Tabelle mit fünfzig Zeilen und zwei Spalten enthält. In der ersten Spalte sind die verschiedenen Bodennutzungsarten aufgeführt, während in der zweiten Spalte die entsprechende Anzahl landwirtschaftlicher Betriebe angegeben ist. Die Daten stammen aus der Statistik „Landwirtschaftliche Betriebe, Fläche: Deutschland, Jahre, Bodennutzungsarten“ für das Jahr 2023, die auf der Webseite www-genesis.destatis.de veröffentlicht wurde.

R1.2 - Skalenvariante

Die Bodennutzungsarten sind eine Nominalskala, da es keine natürliche Rangordnung oder Reihenfolge zwischen den verschiedenen Kategorien gibt. Jede Kategorie steht für eine eigenständige, unterscheidbare Klasse.
Die Anzahl landwirtschaftlicher Betriebe ist eine Absolutskala, da es sich um numerische Werte handelt, bei denen sowohl Abstände als auch Verhältnisse zwischen den Werten sinnvoll interpretiert werden können. Es gibt einen natürlichen Nullpunkt, der das Fehlen von landwirtschaftlichen Betrieben darstellt.

R1.3 - Benutzte Softwares und Funktionen

Die Bearbeitung der Daten erfolgt mit der Programmiersprache Python in der Entwicklungsumgebung PyCharm, sowie mit den Programmen Numbers und Google Docs. Zur Datenverarbeitung und Analyse wurden die folgenden Python-Bibliotheken verwendet: math, scipy, matplotlib, csv und numpy.

R1.7 - Modus, Mittelwert und Median

Für den gegebenen Datensatz können der Median, der Modus und der arithmetische Mittelwert nur für die Variable „landwirtschaftliche Betriebe” berechnet werden, da die Variable „Bodennutzungsarten” nur aus nominalen Werten (Namen) besteht und daher keine numerischen Berechnungen zulässt.

- Median: 15410
- Modus: Alle Werte kommen gleich oft vor.
- Arithmetischer Mittelwert: 42977,755

R1.8 - Spannweite

Die Spannweite kann ausschließlich für die Variable „landwirtschaftliche Betriebe” berechnet werden, da die Variable „Bodennutzungsarten” lediglich nominale Werte enthält und somit keine numerische Spannweite ermöglicht. Für die Variable „landwirtschaftliche Betriebe” beträgt die Spannweite 254.800. Dieser Wert ergibt sich aus der Differenz zwischen dem kleinsten Wert (Anzahl der Betriebe für Heil-, Duft-, Gewürzpflanzen sowie Glas-/andere begehbare Schutzabdeckungen: 210) und dem größten Wert (Anzahl der Betriebe für Betriebsflächen: 255.010).

R1.9 - Mittlere Abweichung vom Median

Die mittlere Abweichung vom Median kann nur für die Variable „landwirtschaftliche Betriebe” berechnet werden, da die Variable „Bodennutzungsarten” ausschließlich nominale Werte enthält und daher keine Berechnung der mittleren Abweichung vom Median zulässt. Für die Variable „landwirtschaftliche Betriebe” beträgt die mittlere Abweichung vom Median MAD = 39.455,102.

R1.10 - Stichprobenvarianz

Die Stichprobenvarianz kann nur für die Variable „landwirtschaftliche Betriebe” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher keine Stichprobenvarianz berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe” beträgt die Stichprobenvarianz 62.723,865.

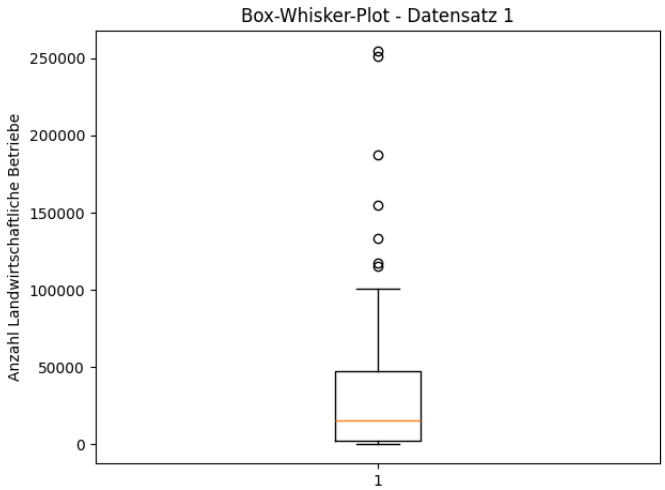
R1.11 - Variationskoeffizient

Der Variationskoeffizient kann nur für die Variable „landwirtschaftliche Betriebe” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher kein Variationskoeffizient berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe” beträgt der Variationskoeffizient 1,459.

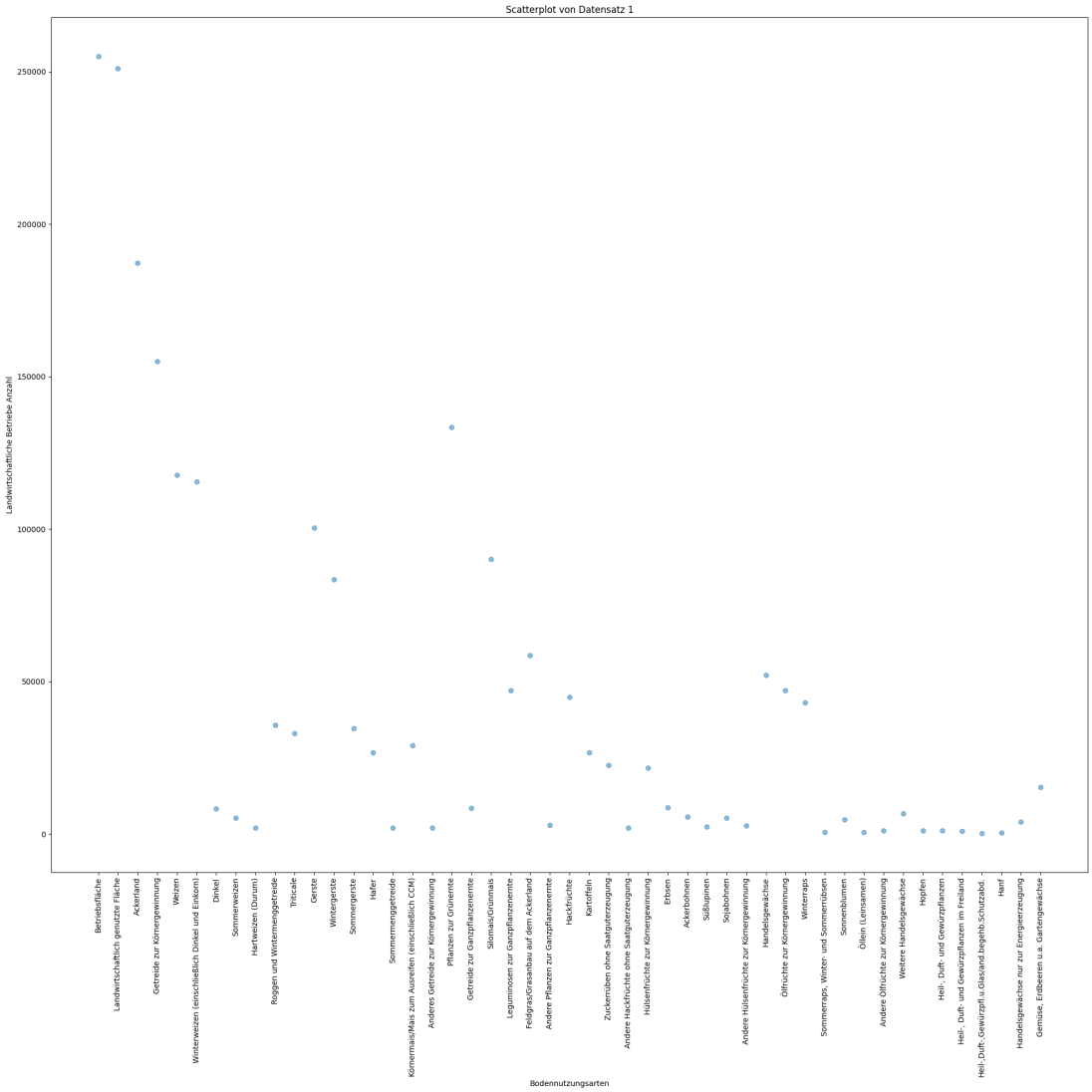
Wenn der Variationskoeffizient größer als 1 ist, bedeutet dies, dass die Standardabweichung mehr als das Einfache des Mittelwerts ausmacht. In anderen Worten: Die Werte im Datensatz variieren erheblich im Vergleich zum Durchschnittswert. Das deutet auf eine hohe Unsicherheit oder Streuung der Daten hin, was in vielen Fällen auf eine heterogene oder unregelmäßige Verteilung der Werte hinweist.

Zusammengefasst signalisiert ein Variationskoeffizient größer als 1 eine größere Variabilität der Daten im Verhältnis zum Mittelwert und wird oft als Indikator für eine größere Unsicherheit oder Unbeständigkeit in den Messungen interpretiert.

R1.12 Box-Whisker-Plot



R1.13 Scatterplot



R1.15 - Quartile und Dezibel

Quartile:
Q1(0,25) = 2070
Q2(0,5) = 15410
Q3(0,75) = 47085

Dezile:
D(0,1) = 920
D(0,2) = 1920
D(0,3) = 2840
D(0,4) = 5570
D(0,5) = 15410
D(0,6) = 29030
D(0,7) = 44870
D(0,8) = 83470
D(0,9) = 133460

R1.16 - Quartilsabstand RQ0.5

Der Quartilsabstand kann nur für die Variable „landwirtschaftliche Betriebe” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher kein Quartilsabstand berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe” beträgt der Quartilabstand 45015.

R1.17 - Kovarianz

Die Kovarianz ist 0 bzw. nicht sinnvoll bestimmbar, da die Variable „Bodennutzungsarten” nominale Werte (Namen) enthält, die keine numerische Auswertung ermöglichen. Mathematisch wäre es möglich, die Namen zu kodieren und eine Kovarianz zu berechnen, jedoch wäre das Ergebnis technisch und inhaltlich bedeutungslos, da eine Kodierung keine echte numerische Beziehung zwischen den Kategorien herstellt.

R1.18 - Korrelationskoeffizient

Der Korrelationskoeffizient ist 0 bzw. nicht sinnvoll bestimmbar, da die Variable „Bodennutzungsarten” nominale Werte (Namen) enthält, die keine numerische Auswertung erlauben. Mathematisch könnte man die Namen kodieren und einen Korrelationskoeffizienten berechnen, allerdings wäre das Ergebnis technisch und inhaltlich bedeutungslos, da eine Kodierung keine echte numerische Beziehung zwischen den Kategorien widerspiegelt.

R1.19 - Klassifizierung

Klassifizierung nach „Bodennutzungsarten”:

KLASSE 1: FLÄCHENNUTZUNG

Betriebsfläche; Landwirtschaftlich genutzte Fläche; Ackerland

KLASSE 2: GETREIDE

Getreide zur Körnergewinnung; Weizen; Winterweizen (einschließlich Dinkel und Einkorn); Sommerweizen; Dinkel; Hartweizen (Durum); Roggen und Wintermenggetreide; Triticale; Gerste; Wintergerste; Sommergerste; Hafer; Sommermenggetreide; Körnermais/Mais zum; Ausreifen (einschließlich CCM); Anderes Getreide zur Körnergewinnung; Pflanzen zur Grünernte; Getreide zur Ganzpflanzenernte; Silomais/Grünmais; Leguminosen zur Ganzpflanzenernte; Feldgras/Grasanbau auf dem Ackerland; Andere Pflanzen zur Ganzpflanzenernte

KLASSE 3: HACKFRÜCHTE

Hackfrüchte; Kartoffeln; Zuckerrüben ohne Saatguterzeugung; Andere Hackfrüchte ohne Saatguterzeugung

KLASSE 4: HÜLSENFRÜCHTE (ZUR KÖRNERGEWINNUNG)

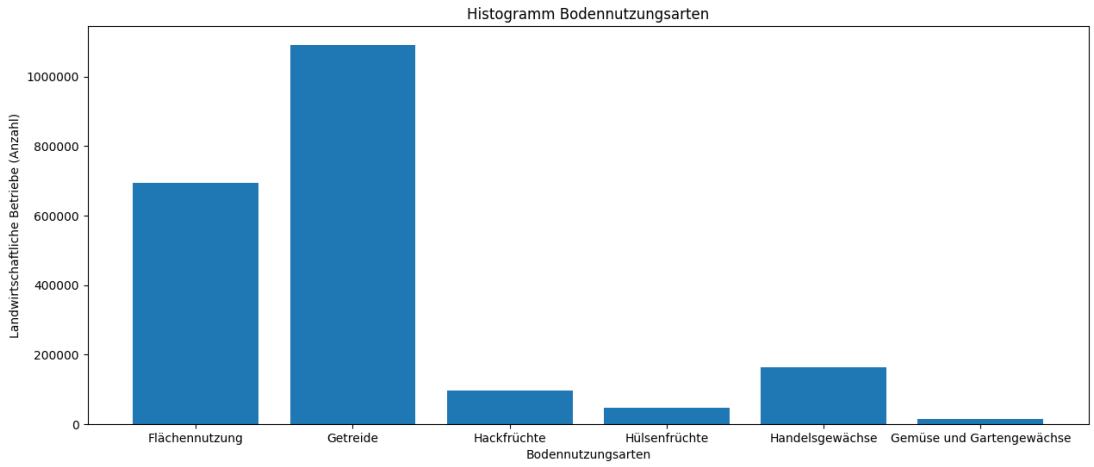
Erbsen; Ackerbohnen; Süßlupinen; Sojabohnen; Andere Hülsenfrüchte zur Körnergewinnung

KLASSE 5: HANDELSGEWÄCHSE

Ölfrüchte zur Körnergewinnung; Winterraps; Sommerraps, Winter- und Sommerrübsen, Sonnenblumen; Öllein (Leinsamen), Andere Ölfrüchte zur Körnergewinnung; Weitere Handelsgewächse; Hopfen; Heil-, Duft- und Gewürzpflanzen; Heil-, Duft- und Gewürzpflanzen im Freiland; Heil-, Duft-, Gewürzpfl. u. Glas/and. begehb. Schutzabd.; Hanf; Handelsgewächse nur zur Energieerzeugung

KLASSE 6: GEMÜSE UND GARTENGEWÄCHSE

Gemüse, Erdbeeren und andere Gartengewächse



Klassifizierung nach „ Anzahl landwirtschaftliche Betriebe”:

KLASSE 1: KLEINE BETRIEBE (BETRIEBSFLÄCHE < 10.000):

Beispiele:

"Heil-, Duft- und Gewürzpflanzen" (1.040)
"Heil-, Duft- und Gewürzpflanzen im Freiland" (920)
"Heil-, Duft-, Gewürzpflanzen u. Glas/and. begehb. Schutzabdeckungen" (210)
Hanf (410)
"Sommerraps, Winter- und Sommerrübsen" (500)
"Gemüse, Erdbeeren u.a. Gartengewächse" (15.410)

KLASSE 2: MITTLERE BETRIEBE (BETRIEBSFLÄCHE ZWISCHEN 10.000 UND 100.000):

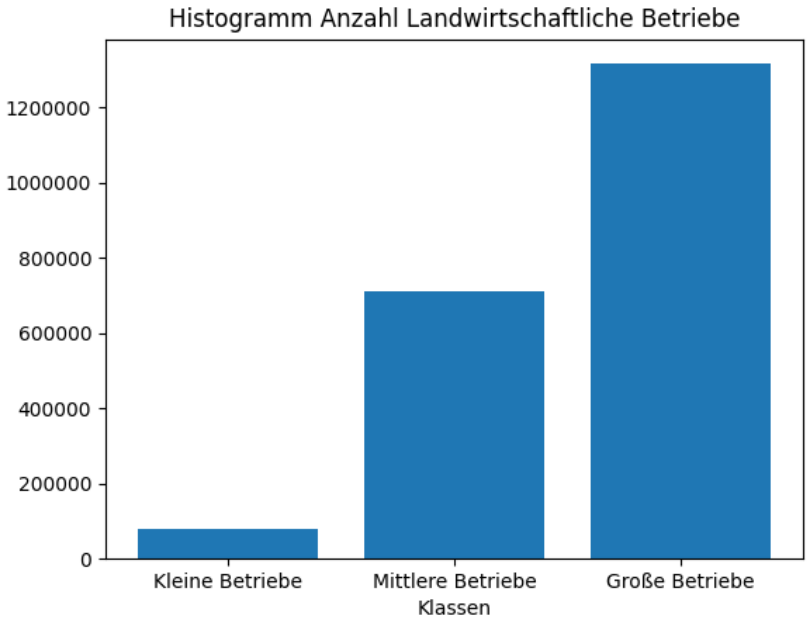
Beispiele:

Getreide zur Körnergewinnung (155.080)
Weizen (117.750)
Winterweizen (inkl. Dinkel und Einkorn) (115.570)
Gerste (100.470)
Silomais/Grünmais (90.160)
Leguminosen zur Ganzpflanzenernte (47.060)
Ölfrüchte zur Körnergewinnung (47.110)

KLASSE 3: GROßE BETRIEBE (BETRIEBSFLÄCHE > 100.000):

Beispiele:

Betriebsfläche (255.010)
Landwirtschaftlich genutzte Fläche (251.130)
Ackerland (187.300)



R1.14 - Fazit

Der analysierte Datensatz gibt einen detaillierten Überblick über die Verteilung landwirtschaftlicher Betriebe in Deutschland nach Bodennutzungsarten im Jahr 2023. Die Daten wurden in einer CSV-Datei gespeichert und mit Python sowie weiteren Software-Tools verarbeitet.

Die statistische Analyse ergab, dass die Anzahl der landwirtschaftlichen Betriebe eine erhebliche Streuung aufweist, wie der hohe Variationskoeffizient (>1) zeigt. Dies deutet auf eine große Heterogenität in der Betriebsgröße und -verteilung hin. Während einige Bodennutzungsarten nur von wenigen Betrieben genutzt werden, gibt es andere, die eine sehr hohe Anzahl an Betrieben aufweisen.

Die berechneten Kenngrößen, wie Median (15.410), arithmetischer Mittelwert (42.977,755) und Spannweite (254.800), verdeutlichen diese Ungleichverteilung. Zudem zeigt die Klassifizierung der Daten, dass landwirtschaftliche Betriebe hinsichtlich ihrer Fläche und Nutzung stark variieren, von kleinen spezialisierten Betrieben bis hin zu großflächigen Landwirtschaftsbetrieben.

Zusammenfassend liefert der Datensatz wertvolle Einblicke in die Struktur und Verteilung landwirtschaftlicher Betriebe in Deutschland. Die hohe Variabilität der Daten sollte bei weiteren Analysen und Interpretationen berücksichtigt werden, insbesondere wenn es um wirtschaftliche oder politische Entscheidungen im Agrarsektor geht.

Datensatz 2

R2.1 - Beschreibung des Datensatzes

Der Datensatz besteht aus einer UTF-8-kodierten CSV-Datei, die eine Tabelle mit 53 Zeilen und zwei Spalten enthält. In der ersten Spalte sind die verschiedenen Bodennutzungsarten aufgeführt, während in der zweiten Spalte die entsprechende Anzahl landwirtschaftlicher Betriebe mit ökologischem Landbau angegeben ist. Die Daten stammen aus der Statistik „Landwirtschaftliche Betriebe mit ökologischem Landbau, Fläche, Ökologisch bewirtschaftete Fläche: Deutschland, Jahre, Bodennutzungsarten“ für das Jahr 2023, die auf der Webseite www-genesis.destatis.de veröffentlicht wurde.

R2.3 - Maßnahmen zur Datenbereinigung

Die zweite Zeile wurde zur Bereinigung der Daten entfernt, da sie lediglich die Einheit der zweiten Spalte wiederholt, die bereits in der ersten Zeile angegeben ist. Zudem enthält diese Zeile in der ersten Spalte keinen Wert und trägt somit nicht zur Analyse bei. Darüber hinaus wurde die Zeile mit dem Eintrag „Baumobstanlagen“ entfernt, da in der zweiten Spalte der Wert „.“ steht, der für die Auswertung nicht verwendet werden kann. Des Weiteren wurden die Zeichen „/“ und „\“ entfernt, da sie keinen Mehrwert für die Tabelle bieten und die Umwandlung in das Excel-Format erschweren. Diese Zeichen wurden durch Kommas ersetzt. Nach der Bereinigung enthält die Tabelle 51 Zeilen.

R2.4 - Benutzte Softwares und Funktionen

Die Bearbeitung der Daten erfolgt mit der Programmiersprache Python in der Entwicklungsumgebung PyCharm, sowie mit den Programmen Numbers und Google Docs. Zur Datenverarbeitung und Analyse wurden die folgenden Python-Bibliotheken verwendet: math, scipy, matplotlib, csv und numpy.

R2.8 - Median, Modus und Mittelwert

Für den gegebenen Datensatz können der Median, der Modus und der arithmetische Mittelwert nur für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” berechnet werden, da die Variable „Bodennutzungsarten” nur aus nominalen Werten (Namen) besteht und daher keine numerischen Berechnungen zulässt.

- Median: 930
- Modus: Die Zahl 90 und die Zahl 1470 kommen jeweils zweimal vor

  90: „GARTENBAUSÄMEREIEN, JUNGPFLANZENERZ. ZUM VERKAUF” UND „REBFLÄCHEN FÜR TAFELTRAUBEN”
  1470: „ERBSEN” UND „ANDERE HÜLSENFRÜCHTE ZUR KÖRNERGEWINNUNG”

- Arithmetischer Mittelwert: 2523

R2.9 - Spannweite

Die Spannweite kann nur für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher keine Spannweite berechnet werden kann.

Die Spannweite beträgt 26070 und ergibt sich aus der Differenz zwischen dem kleinsten Wert (Anzahl der Betriebe für Sommerraps, Winter- und Sommerrübsen: 20) und dem größten Wert (Anzahl der Betriebe für Dauergrünland: 26090).

R2.10 - Mittlere Abweichung vom Median

Die mittlere Abweichung vom Median kann nur für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher keine mittlere Abweichung vom Median berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” beträgt die mittlere Abweichung vom Median 2185.

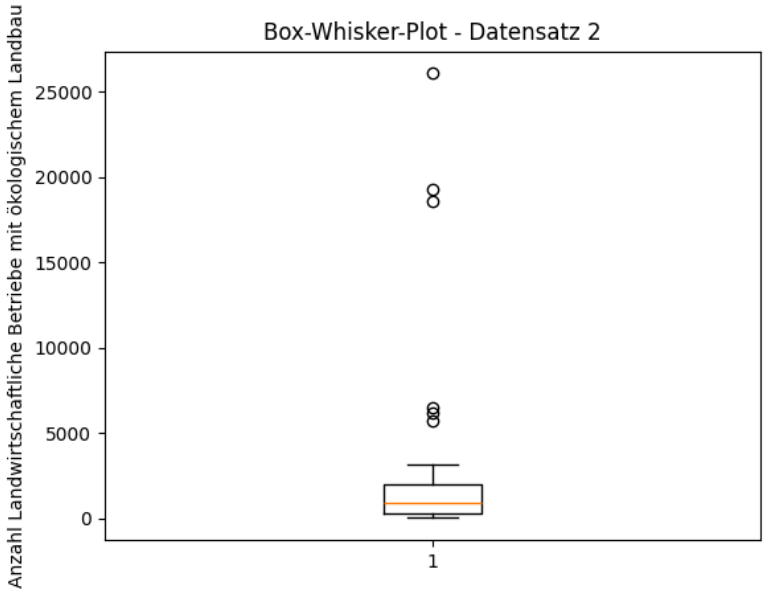
R2.11 - Stichprobenvarianz

Die Stichprobenvarianz kann nur für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher keine Stichprobenvarianz berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” beträgt die Stichprobenvarianz 5098,808.

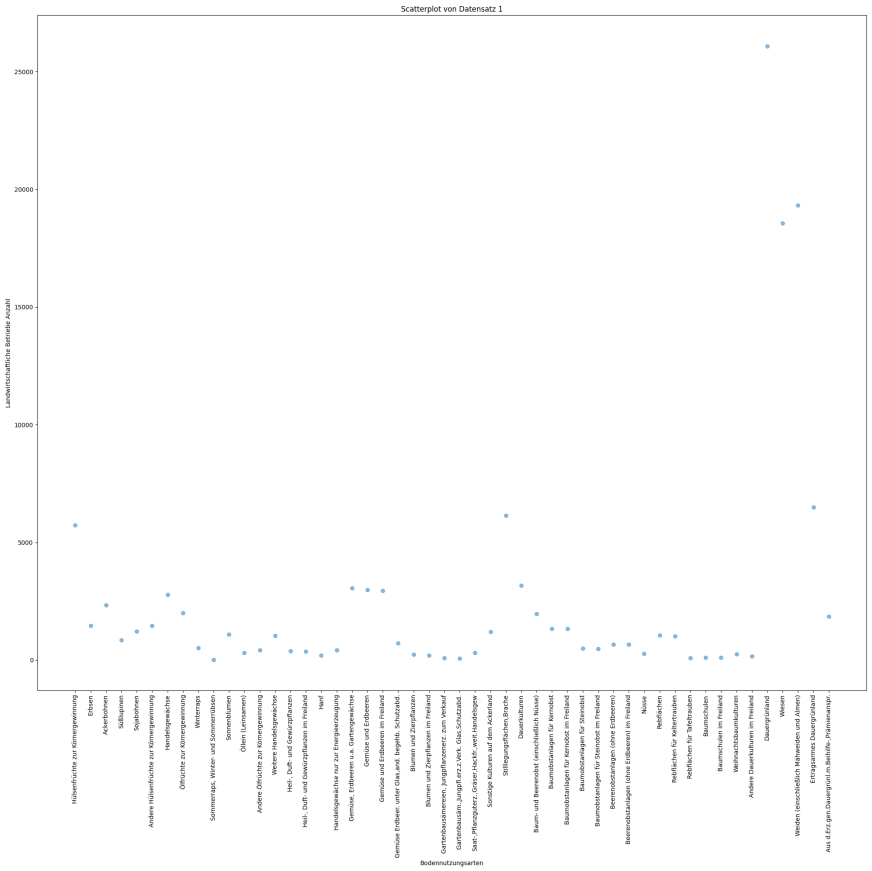
R2.12 - Variationskoeffizient

Der Variationskoeffizient kann nur für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” berechnet werden, da die Variable „Bodennutzungsarten” nur nominale Werte enthält und daher kein Variationskoeffizient berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau” beträgt der Variationskoeffizient 2,021.
Wenn der Variationskoeffizient größer als 1 ist, bedeutet dies, dass die Standardabweichung mehr als das Einfache des Mittelwerts ausmacht. In anderen Worten: Die Werte im Datensatz variieren erheblich im Vergleich zum Durchschnittswert. Das deutet auf eine hohe Unsicherheit oder Streuung der Daten hin, was in vielen Fällen auf eine heterogene oder unregelmäßige Verteilung der Werte hinweist.
Zusammengefasst signalisiert ein Variationskoeffizient größer als 1 eine größere Variabilität der Daten im Verhältnis zum Mittelwert und wird oft als Indikator für eine größere Unsicherheit oder Unbeständigkeit in den Messungen interpretiert.

R2.13 - Box-Whisker-Plot



R2.14 - Scatterplot



R2.17 - Quartile und Dezile

Quartile:
Q1(0,25) = 310
Q2(0,5) = 930
Q3(0,75) = 2010

Dezile:
D(0,1) = 110
D(0,2) = 260
D(0,3) = 390
D(0,4) = 520
D(0,5) = 1010
D(0,6) = 1220
D(0,7) = 1850
D(0,8) = 2940
D(0,9) = 6150

R2.18 - Quartilsabstand RQ0.5

Der Quartilsabstand kann nur für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau" berechnet werden, da die Variable „Bodennutzungsarten" nur nominale Werte enthält und daher kein Quartilabstand berechnet werden kann.
Für die Variable „landwirtschaftliche Betriebe mit ökologischem Landbau" beträgt der Quartilsabstand 1700.

R2.19 - Kovarianz

Die Kovarianz ist 0 bzw. nicht sinnvoll bestimmbar, da die Variable „Bodennutzungsarten" nominale Werte (Namen) enthält, die keine numerische Auswertung ermöglichen. Mathematisch wäre es möglich, die Namen zu kodieren und eine Kovarianz zu berechnen, jedoch wäre das Ergebnis technisch und inhaltlich bedeutungslos, da eine Kodierung keine echte numerische Beziehung zwischen den Kategorien herstellt.

R2.20 - Korrelationskoeffizient

Der Korrelationskoeffizient ist 0 bzw. nicht sinnvoll bestimmbar, da die Variable „Bodennutzungsarten" nominale Werte (Namen) enthält, die keine numerische Auswertung erlauben. Mathematisch könnte man die Namen kodieren und einen Korrelationskoeffizienten berechnen, allerdings wäre das Ergebnis technisch und inhaltlich bedeutungslos, da eine Kodierung keine echte numerische Beziehung zwischen den Kategorien widerspiegelt.

R2.16 - Fazit

Der vorliegende Datensatz enthält Informationen zu landwirtschaftlichen Betrieben mit ökologischem Landbau in Deutschland und umfasst 51 Zeilen, nachdem die erforderliche Bereinigung durchgeführt wurde. Die Bereinigung betraf hauptsächlich das Entfernen von überflüssigen Zeilen und Zeichen, um eine saubere Datenstruktur zu gewährleisten.
Die Analyse der numerischen Variablen, insbesondere der Anzahl der landwirtschaftlichen Betriebe mit ökologischem Landbau, zeigt eine hohe Variabilität. Der Variationskoeffizient von 2,021 weist auf eine erhebliche Streuung der Werte im Vergleich zum Mittelwert hin, was auf eine heterogene Verteilung der Betriebe in den verschiedenen Bodennutzungsarten hindeutet. Die Spannweite der Betriebe reicht von 20 bis 26.090, was eine deutliche Bandbreite in der Anzahl der Betriebe je Bodennutzungsart widerspiegelt.
Der Median (930) und der arithmetische Mittelwert (2523) zeigen ebenfalls eine größere Verteilung der Werte. Der Quartilsabstand von 1700 bestätigt die Streuung der Daten, und die berechneten Quartile und Dezile geben weitere Einblicke in die Verteilung der landwirtschaftlichen Betriebe.
Die Variabilität der Daten und die hohe Streuung deuten darauf hin, dass die landwirtschaftlichen Betriebe in verschiedenen Bodennutzungsarten sehr unterschiedlich vertreten sind. Eine Korrelation oder Kovarianz zwischen den Bodennutzungsarten und der Anzahl der Betriebe kann aufgrund der nominalen Natur der Bodennutzungsarten nicht sinnvoll berechnet werden.
Insgesamt liefert der Datensatz wertvolle Informationen zur Verteilung landwirtschaftlicher Betriebe im ökologischen Landbau, wobei die Streuung und Variabilität der Betriebe in den verschiedenen Kategorien auf interessante Muster hinweisen, die möglicherweise durch unterschiedliche landwirtschaftliche Praktiken oder geografische Gegebenheiten beeinflusst werden.

Datensatz 3

R3.1 - Beschreibung des Datensatzes

Struktur und Inhalt des Datensatzes
Der Datensatz besteht aus zwei UTF-8-kodierten CSV-Dateien, die jeweils eine Tabelle enthalten:

Erste Tabelle:
Anzahl der Zeilen: 57
Spalten:
- Erste Spalte: Schlüssel (Identifikatoren)
- Zweite Spalte: Unternehmensdemografien

Zweite Tabelle:
Anzahl der Zeilen: 57
Spalten:
- Erste Spalte: Schlüssel (Identifikatoren)
- Zweite Spalte: Anzahl überlebender Unternehmen (vor zwei Jahren)
- Dritte Spalte: Überlebensrate der Unternehmen (vor zwei Jahren)

Herkunft der Daten
Der Datensatz ist Teil der Statistik „Überlebende Unternehmen: Deutschland, Jahre, Wirtschaftszweige, Zeitpunkt der Unternehmensgründung, Beschäftigtengrößenklassen“ und wurde von der Webseite GENESIS-Online Datenbank des Statistischen Bundesamtes im Jahr 2022 bezogen.

R3.4 - Maßnahmen zur Datenbereinigung

Die Maßnahmen zur Datenbereinigung wurden sorgfältig durchgeführt. Dabei wurden Anführungszeichen entfernt, leere Zeilen gelöscht und unnötige Zeilenumbrüche in der ersten Zeile (Tabellenbeschriftungen) korrigiert. Dies gewährleistet, dass der Inhalt jeder Zeile vollständig und korrekt in einer einzigen Zeile dargestellt ist. Ansonsten kann die Zeile mit „MU574731,Erziehung und Unterricht,17,106.3“ entfernt werden, da Überlebensrate >1 nicht sinnvoll ist. Um Zeit zu sparen und mögliche Fehler zu vermeiden, wurden alle Schritte manuell ausgeführt.

R3.6 - Software und Funktionen

Für die Datenbearbeitung kam die Programmiersprache Python in der Entwicklungsumgebung PyCharm zum Einsatz, ergänzt durch die Programme Numbers und Google Docs. Zur Datenverarbeitung und Analyse wurden folgende Python-Bibliotheken verwendet: math, scipy, matplotlib, csv, numpy und collections.

R3.9 - Modus, arithmetischer Mittelwert und Median

Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können kein modus, median und Mittelwert berechnet werden, da sie nur nominale werte enthalten.
Anzahl überlebender Unternehmen
- Modus: 1 (vier mal)
- Median: 23
- Mittelwert: 65,91

Überlebensrate der Unternehmen
- Modus: 100 (13 mal)
- Median: 90,1%
- Mittelwert: 88,63%

R3.10 - Spannweite

Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können keine Spannweite berechnet werden, da sie nur nominale werte enthalten.
Spannweite zur Anzahl überlebender Unternehmen: 313
Spannweite zur Überlebensrate der Unternehmen: 50%

R3.11 - Mittlere Abweichung vom Median

Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können keine Median berechnet werden, da sie nur nominale werte enthalten.
Mittlere Abweichung vom Median zur Anzahl überlebender Unternehmen: 57,673
Mittlere Abweichung vom Median zur Überlebensrate der Unternehmen: 7.235

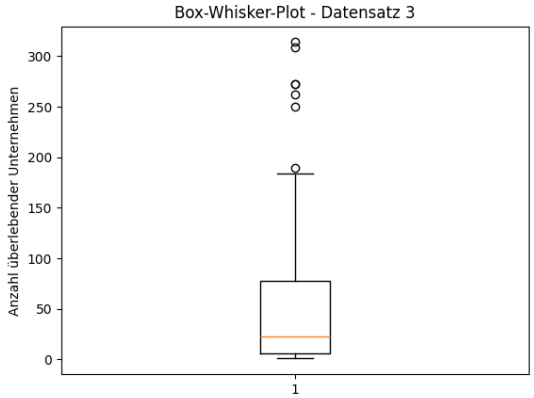
R3.12 - Stichprobenvarianz

Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können keine Stichprobenvarianz berechnet werden, da sie nur nominale werte enthalten.
Stichprobenvarianz zur Anzahl überlebender Unternehmen: 90,16
Stichprobenvarianz zur Überlebensrate der Unternehmen: 10,808

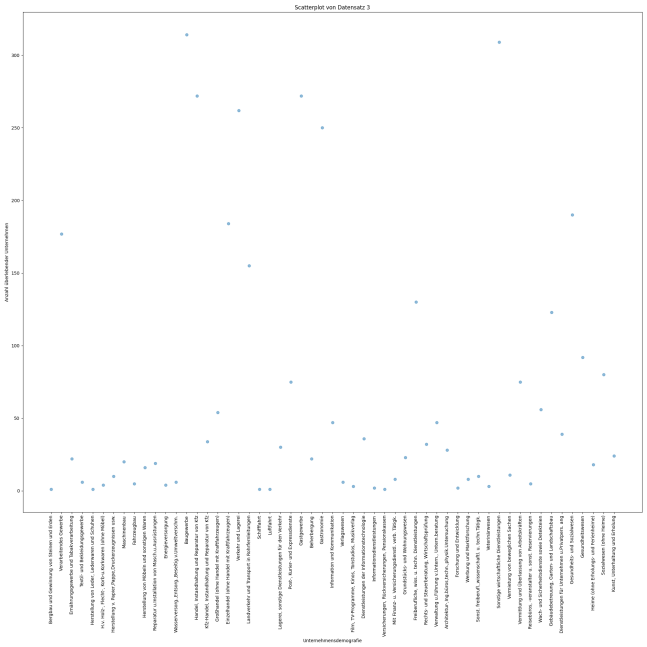
R3.13 - Variationskoeffizient

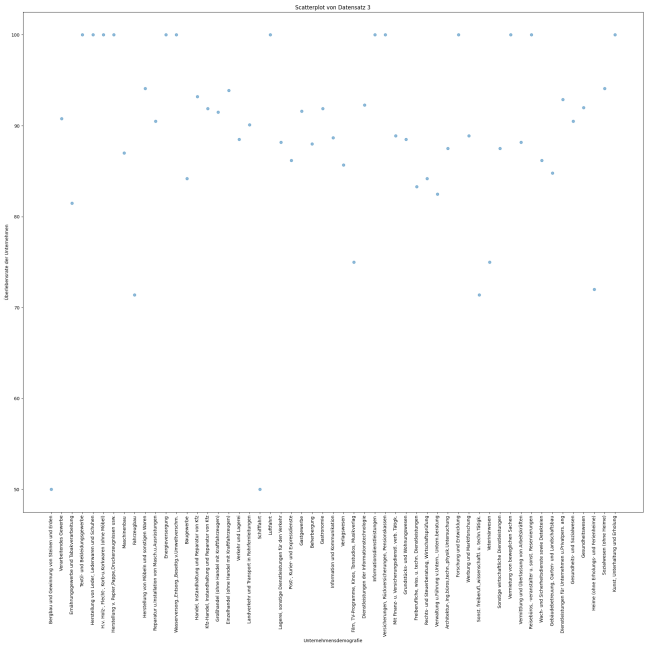
Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können keine Variationskoeffizient berechnet werden, da sie nur nominale werte enthalten.
Variationskoeffizient zur Anzahl überlebender Unternehmen: 1,368
→ Dies weist auf eine Streuung hin, die größer als der Mittelwert ist. Das bedeutet, dass der Mittelwert keine zuverlässige Darstellung der Daten bietet.
Variationskoeffizient zur Überlebensrate der Unternehmen: 0,122
→ Dies deutet auf eine geringe Streuung hin, wodurch der Mittelwert brauchbar und aussagekräftig ist.

R3.14 - Box-Whisker-Plot



R3.15 - Scatterplot





R3.20 - Quartile und Dezile

Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können keine Quartile und Dezile berechnet werden, da sie nur nominale werte enthalten.
Quartile und Dezile zur Anzahl überlebender Unternehmen:

Quartile:
Q1(0,25) = 5,5
Q2(0,5) = 23
Q3(0,75) = 75

Dezile:
D(0,1) = 2
D(0,2) = 5
D(0,3) = 8
D(0,4) = 18
D(0,5) = 23
D(0,6) = 36
D(0,7) = 56
D(0,8) = 130
D(0,9) = 250

Quartile und Dezile zur Überlebensrate der Unternehmen:
Quartile:
Q1(0,25) = 85,25
Q2(0,5) = 90,1
Q3(0,75) = 94,1

Dezile:
D(0,1) = 75
D(0,2) = 84,2
D(0,3) = 87
D(0,4) = 88,5
D(0,5) = 90,1
D(0,6) = 91,9
D(0,7) = 93,2
D(0,8) = 100
D(0,9) = 100

R3.21 - Quartilsabstand RQ0.5

Zu den Variablen „Schlüssel” und „Unternehmensdemografien” können keinen Quartilsabstand berechnet werden, da sie nur nominale werte enthalten.
Quartilsabstand zur Anzahl überlebender Unternehmen: 96,5
Quartilsabstand zur Überlebensrate der Unternehmen: 8,85

R3.22 - Kovarianz

Die Kovarianz zwischen Anzahl überlebender Unternehmen und Überlebensrate der Unternehmen berechnet lässt sich berechnen. Diese Beträgt 44,875.

R3.23 - Korrelationskoeffizient

Der Korrelationskoeffizient zwischen Anzahl überlebender Unternehmen und Überlebensrate der Unternehmen berechnet lässt sich berechnen. Diese Beträgt 0,046.

R3.19 - Fazit

Der vorliegende Datensatz besteht aus zwei Tabellen, die Informationen über Unternehmensdemografien und die Überlebensraten von Unternehmen in Deutschland enthalten. Er umfasst insgesamt 57 Zeilen pro Tabelle und enthält sowohl nominale als auch numerische Variablen.
Im Rahmen der Datenbereinigung wurden notwendige Anpassungen wie das Entfernen von Anführungszeichen und das Löschen leerer Zeilen durchgeführt, um eine fehlerfreie Datenbasis sicherzustellen. Die verwendeten Tools und Software, darunter Python mit verschiedenen Bibliotheken, ermöglichten eine gründliche Analyse der Daten.
Die statistischen Auswertungen des Datensatzes zeigen eine geringe Streuung bei den Überlebensraten der Unternehmen, während die Anzahl der überlebenden Unternehmen eine größere Variabilität aufweist. Der Korrelationskoeffizient von 0,046 zwischen der Anzahl der überlebenden Unternehmen und deren Überlebensrate deutet auf eine sehr schwache positive lineare Beziehung hin. Dies lässt darauf schließen, dass andere Faktoren möglicherweise stärker mit der Überlebensrate der Unternehmen korrelieren, jedoch keine klare Abhängigkeit zwischen der Anzahl der überlebenden Unternehmen und ihrer Überlebensrate besteht.
Insgesamt zeigt der Datensatz interessante Einsichten in die Überlebensraten von Unternehmen, aber die schwache Korrelation legt nahe, dass eine detailliertere Untersuchung weiterer Faktoren notwendig ist, um die Ursachen für den Unternehmenserfolg besser zu verstehen.

Datensatz 4

R4.3 - Maßnahmen zur Datenbereinigung

Die Maßnahmen zur Datenbereinigung umfassten zwei wesentliche Schritte. Zunächst wurden alle negativen Werte entfernt, da die MTTF (Mean Time to Failure) keinen Wert kleiner als 1 annehmen kann. Anschließend wurde die Werte 10.0371 und 9.1166 ausgeschlossen, da sie sehr hohe Abweichung von den restlichen Daten aufweisen und somit wahrscheinlich fehlerhaft sind. Diese Bereinigungsschritte stellen sicher, dass die Analyse auf plausiblen und realistischen Daten basiert.

R4.4 - Software und Funktionen

Für die Datenbearbeitung kam die Programmiersprache Python in der Entwicklungsumgebung PyCharm zum Einsatz, ergänzt durch die Programme Numbers und Google Docs. Zur Datenverarbeitung und Analyse wurden folgende Python-Bibliotheken verwendet: math, scipy, matplotlib, csv, numpy und collections.

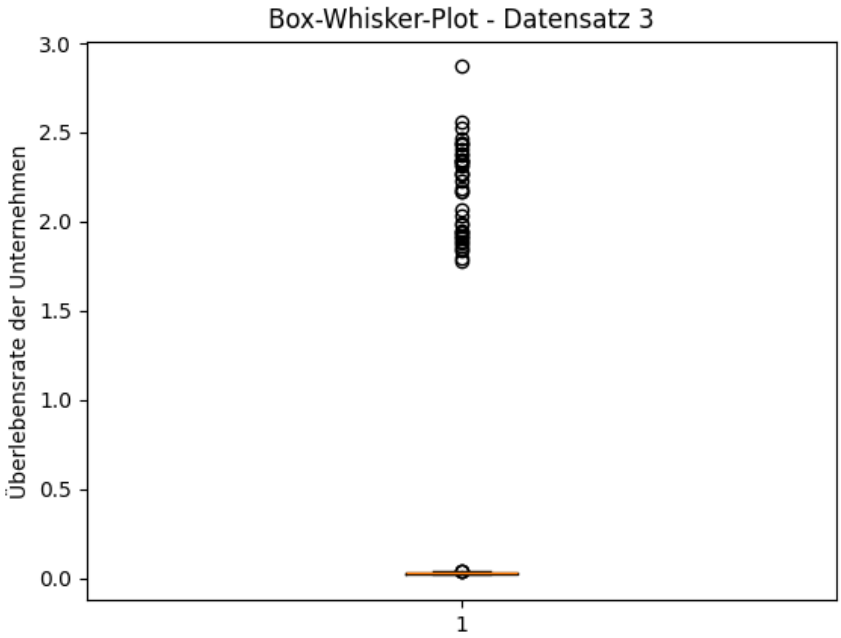
R4.5 - Modus, Median und Mittelwert

Modus: 0,0281
Median: 0,0279
Mittelwert: 0,365

R4.6 - Stichprobenvarianz

Stichprobenvarianz beträgt: s = 0,791.

R4.7 - Box-Whisker-Plot



R4.8 - Fazit

Die durchgeführten Datenbereinigungsmaßnahmen haben dazu beigetragen, fehlerhafte oder unplausible Werte zu entfernen und die Datenqualität zu verbessern. Durch den Ausschluss negativer Werte sowie der auffällig abweichenden Werte 10.0371 und 9.1166 wurde sichergestellt, dass die Analyse auf einer konsistenten und aussagekräftigen Datengrundlage basiert.

Die Berechnungen der zentralen Lageparameter zeigen, dass der Median (0,0279) und der Modus (0,0281) nahe beieinanderliegen, während der Mittelwert (0,365) aufgrund möglicher Ausreißer deutlich höher ist. Dies weist auf eine asymmetrische Verteilung der Daten hin. Die relativ hohe Stichprobenvarianz (0,791) bestätigt zudem eine erhebliche Streuung der Werte.

Durch den Einsatz von Python und verschiedenen statistischen Bibliotheken konnte eine fundierte Analyse der Daten durchgeführt werden, wodurch wertvolle Erkenntnisse zur Verteilung und Streuung der MTTF-Werte gewonnen wurden.