Projektarbeit – Grundlagen und Anwendungen der Wahrscheinlichkeitstheorie

Statistische Analyse landwirtschaftlicher Betriebe nach Bodennutzungsarten

Veranstaltung:

Grundlagen und Anwendungen der Wahrscheinlichkeitstheorie

Semester:

Wintersemester 2025/2026

Studierende:

Alex Straßburger, Matrikelnummer: 432248

Paul Gib, Matrikelnummer: 432002

Hochschule:

Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)

1. Einleitung

Ziel dieser Projektarbeit ist die deskriptive statistische Analyse mehrerer Datensätze im Kontext landwirtschaftlicher Betriebe und ihrer Bodennutzungsarten. Dabei werden grundlegende Verfahren der Wahrscheinlichkeitstheorie und Statistik angewendet, um die Struktur, Verteilung und Streuung der vorliegenden Daten zu untersuche.

Der Schwerpunkt der Analyse liegt auf der Berechnung und Interpretation von Lage- und Streuungsparametern sowie auf der grafischen Darstellung der Daten. Die Auswertung erfolgt ausschließlich mit Hilfe von Python und gängigen Bibliotheken zur Datenanalyse. Die Ergebnisse werden systematisch dokumentiert und interpretiert

2. Methodik und verwendete Software

Die statistische Auswertung der Datensätze erfolgt mit der Programmiersprache Python. Zur Datenverarbeitung und -analyse werden die Bibliotheken pandas und numpy verwendet. Die grafische Darstellung der Ergebnisse erfolgt mit matplotlib. Ergänzende statistische Kennzahlen werden mithilfe von scipy berechnet.

Alle Auswertungen werden in Jupyter Notebooks durchgeführt. Die Rohdaten, der vollständige Quellcode sowie alle erzeugten Ergebnisdateien liegen in elektronischer Form vor. Sämtliche verwendeten CSV-Dateien sind im UTF-8- Format mit Komma als Trennzeichen und Punkt als Dezimaltrennzeichen gespeichert.

3. Datensatz 1: Landwirtschaftliche Betriebe und Bodennutzungsarten

3.1 Beschreibung des Datensatzes

Der vorliegende Datensatz umfasst 49 Beobachtungen und enthält zwei Variablen. Die Variable Bodennutzungsarten beschreibt verschiedene landwirtschaftliche Nutzungsformen in kategorialer Form. Die Variable Landwirtschaftliche Betriebe Anzahl gibt die Anzahl der landwirtschaftlichen Betriebe pro Bodennutzungsart an. Die Daten liegen als CSV-Datei vor und werden mit Python (pandas) ausgewertet.

3.2 Variablen und Skalenniveaus

- **Bodennutzungsarten:** Nominalskala Begründung: Die Werte stellen Kategorien ohne natürliche Reihenfolge dar.
- **Landwirtschaftliche Betriebe Anzahl:** Metrische Skala (Verhältnisskala, diskret) Begründung: Es handelt sich um absolute Häufigkeiten mit sinnvollem Nullpunkt, bei denen Abstände und Verhältnisse interpretierbar sind.

3.3 Lage- und Streuungsparameter

Lageparameter – Landwirtschaftliche Betriebe Anzahl

Für die metrische Variable Landwirtschaftliche Betriebe Anzahl werden Modus, arithmetischer Mittelwert und Median berechnet.

	Lageparameter	Wert
0	Modus	kein eindeutiger Modus
1	Arithmetischer Mittelwert	42977.755102
2	Median	15410.0

Abbildung 1Lageparameter Betriebe

Interpretation: Der arithmetische Mittelwert beschreibt die durchschnittliche Anzahl landwirtschaftlicher Betriebe pro Bodennutzungsart. Der Median teilt die Verteilung in zwei gleich große Hälften und ist robust gegenüber Ausreißern. Der Modus gibt den am häufigsten vorkommenden Wert an.

Erläuterung zum Modus: Ein eindeutiger Modus existiert für die Variable Landwirtschaftliche Betriebe Anzahl nicht, da jeder beobachtete Wert nur einmal im Datensatz vorkommt. Der Modus ist daher statistisch nicht aussagekräftig.

Lageparameter – Bodennutzungsarten

Da es sich bei den Bodennutzungsarten um eine nominalskalierte Variable handelt, sind arithmetischer Mittelwert und Median nicht sinnvoll definiert. Der Modus gibt die am häufigsten vorkommende Kategorie an.

	Lageparameter	Wert
0	Modus	kein eindeutiger Modus

Abbildung 2 Lageparameter Bodennutzung

Erläuterung zum Modus: Da jede Bodennutzungsart im Datensatz genau einmal vorkommt, existiert kein eindeutiger Modus. Der Modus kann daher nicht zur Charakterisierung der Verteilung genutzt werden.

Spannweite – Landwirtschaftliche Betriebe Anzahl

	Kennzahl	Wert
0	Minimum	210
1	Maximum	255010
2	Spannweite	254800

Abbildung 3 Spannweite Betriebe

Interpretation: Die Spannweite zeigt den gesamten Wertebereich der Anzahl landwirtschaftlicher Betriebe über alle Bodennutzungsarten hinweg. Sie reagiert empfindlich auf Extremwerte und sollte daher stets in Kombination mit weiteren Streuungsmaßen betrachtet werden.

Hinweis: Für die nominalskalierte Variable Bodennutzungsarten ist die Spannweite nicht definiert und wird daher nicht berechnet.

Streuungsmaß: Mittlere Abweichung vom Median

Die mittlere Abweichung vom Median beschreibt die durchschnittliche absolute Abweichung der einzelnen Beobachtungen vom Median der Verteilung. Sie ist weniger empfindlich gegenüber Ausreißern als die Spannweite.

	Kennzahl	Wert
0	Median	15410.000000
1	Mittlere Abweichung vom Median	39455.102041

Abbildung 4 Mittlere Abweichung Betriebe

Interpretation: Die mittlere Abweichung vom Median gibt an, wie stark die Anzahl landwirtschaftlicher Betriebe pro Bodennutzungsart im Durchschnitt vom Medianwert abweicht. Dieses Streuungsmaß ist robuster gegenüber Extremwerten als die Spannweite und eignet sich daher gut zur Beschreibung der Datenverteilung.

Hinweis: Für die nominalskalierte Variable Bodennutzungsarten ist die mittlere Abweichung vom Median nicht definiert und wird daher nicht berechnet.

Streuungsmaß: Stichprobenvarianz

Die Stichprobenvarianz ist ein Maß für die Streuung der Daten um ihren arithmetischen Mittelwert. Sie berücksichtigt die quadrierten Abweichungen der einzelnen Werte vom Mittelwert und ist ein zentrales Streuungsmaß in der Statistik. Für die metrische Variable Landwirtschaftliche Betriebe Anzahl wird die Stichprobenvarianz berechnet.

	Kennzahl	Wert
0	Arithmetischer Mittelwert	4.297776e+04
1	Stichprobenvarianz	3.934283e+09

Abbildung 5 Stichprobenvarianz Betriebe

Interpretation: Die Stichprobenvarianz beschreibt die durchschnittliche quadrierte Abweichung der Anzahl landwirtschaftlicher Betriebe vom arithmetischen Mittelwert. Ein hoher Varianzwert weist auf eine starke Streuung der Daten hin.

Hinweis: Für die nominalskalierte Variable Bodennutzungsarten ist die Stichprobenvarianz nicht definiert und wird daher nicht berechnet.

Lage- und Streuungsparameter: Quartile und Dezile

Zur detaillierten Beschreibung der Verteilung werden die Quartile und Dezile der metrischen Variable berechnet.

	Quantil	Wert
0	0.25	2430.0
1	0.50	15410.0
2	0.75	47110.0
3	0.10	1016.0
4	0.20	1956.0
5	0.30	3304.0
6	0.40	5796.0
7	0.50	15410.0
8	0.60	28564.0
9	0.70	44122.0
10	0.80	68542.0
11	0.90	120892.0

Abbildung 6 Quartile und Dezile

Interpretation: Die Quartile und Dezile verdeutlichen die starke Konzentration der Daten im unteren Wertebereich. Ein Großteil der Beobachtungen liegt deutlich unterhalb des arithmetischen Mittelwertes, was die rechtsschiefe Verteilung bestätigt.

Streuungsmaß: Quartilsabstand

Der Quartilsabstand beschreibt die Spannweite der mittleren 50 % der Daten und ist robust gegenüber Ausreißern.

	Kennzahl	Wert
0	1. Quartil (Q1)	2430.0
1	3. Quartil (Q3)	47110.0
2	Quartilsabstand	44680.0

Abbildung 7 Quartilsabstand

Interpretation: Der Quartilsabstand zeigt, dass die mittleren 50 % der Werte vergleichsweise eng beieinander liegen, während die Gesamtstreuung stark durch Ausreißer beeinflusst wird.

3.4 Grafische Darstellung

Box-Whisker-Plots – Landwirtschaftliche Betriebe Anzahl:

Zur Analyse der Verteilung der Anzahl landwirtschaftlicher Betriebe werden Box-Whisker Plots verwendet. Aufgrund der stark rechtsschiefen Verteilung werden zwei Darstellungen betrachtet: eine vollständige Darstellung inklusive Ausreißern sowie eine ergänzende Darstellung ohne Ausreißer zur besseren Sichtbarkeit des zentralen Datenbereichs.

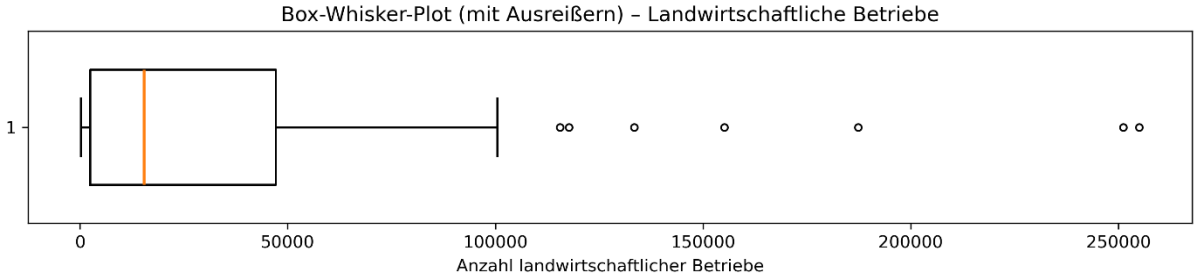

Abbildung 8 Box-Whisker-Plot mit Ausreißern

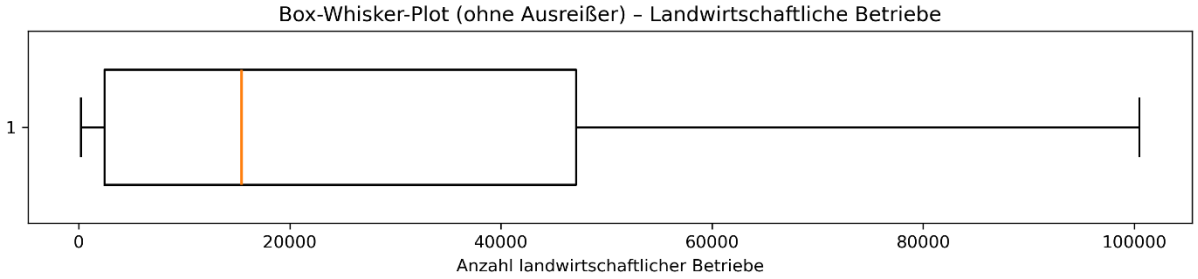

Abbildung 9 Box-Whisker-Plot ohne Ausreißer

Interpretation: Der vollständige Box-Whisker-Plot zeigt eine stark rechtsschiefe Verteilung mit mehreren ausgeprägten Ausreißern nach oben. Die ergänzende Darstellung ohne Ausreißer ermöglicht eine bessere Beurteilung der Lage und Streuung des zentralen Datenbereichs. Beide Darstellungen zusammen liefern ein vollständiges und differenziertes Bild der Datenverteilung.

Hinweis: Für die nominalskalierte Variable Bodennutzungsarten kann kein Box-Whisker-Plot erstellt werden, da keine numerische Ordnung der Kategorien existiert.

Scatterplot – Landwirtschaftliche Betriebe Anzahl:

Der Scatterplot zeigt die Anzahl landwirtschaftlicher Betriebe in Abhängigkeit von der Beobachtungsnummer.

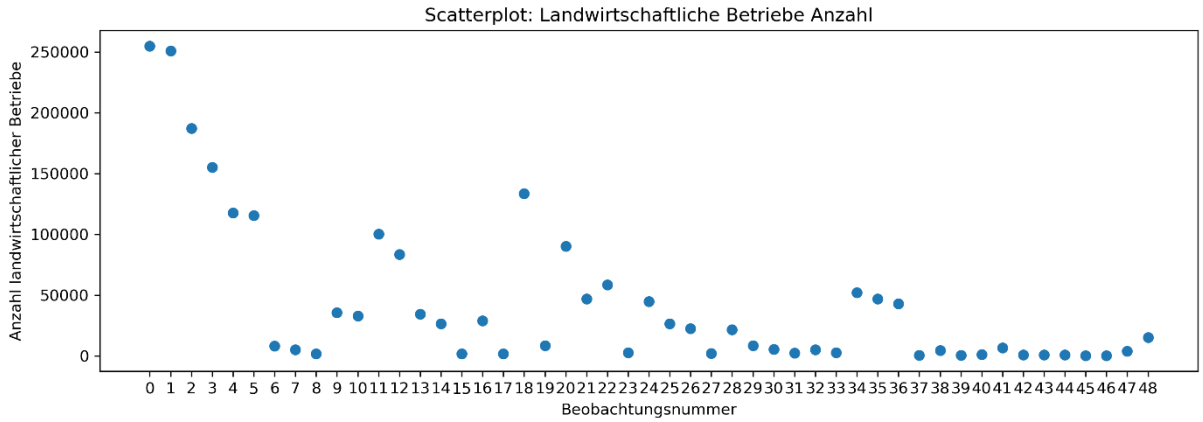

Abbildung 10 Scatterplot Betriebe

Interpretation: Der Scatterplot verdeutlicht die starke Streuung der Anzahl landwirtschaftlicher Betriebe über die einzelnen Beobachtungen hinweg. Ein klarer funktionaler Zusammenhang ist nicht erkennbar, was aufgrund der Verwendung der Beobachtungsnummer als unabhängige Variable zu erwarten ist. Die Darstellung dient primär der Visualisierung der Werteverteilung.

Ergänzende Darstellung: Landwirtschaftliche Betriebe nach Bodennutzungsarten:

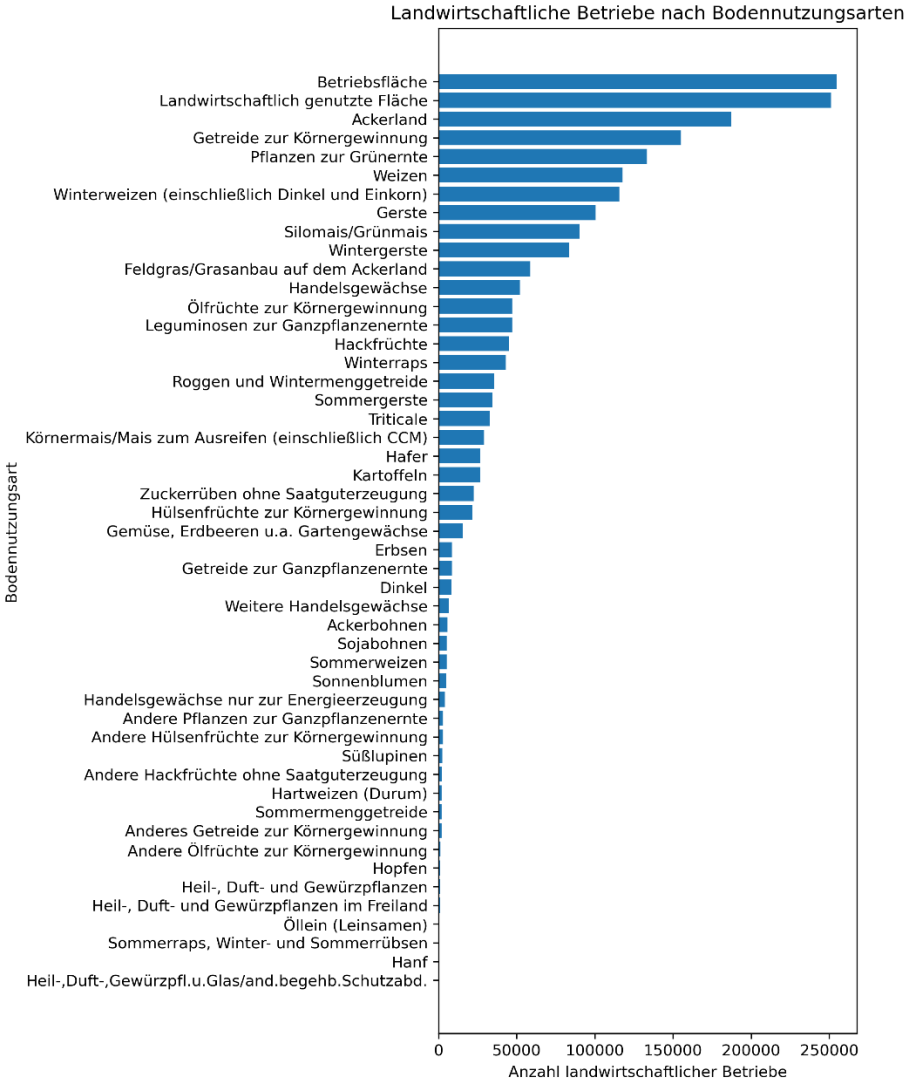

Abbildung 11 Betriebe und Bodennutzungsart

Interpretation: Das Balkendiagramm verdeutlicht die starken Unterschiede in der Anzahl landwirtschaftlicher Betriebe zwischen den einzelnen Bodennutzungsarten. Insbesondere wenige Nutzungsarten weisen sehr hohe Werte auf, während der Großteil der Kategorien deutlich niedrigere Fallzahlen besitzt.

3.5 Erweiterte statistische Auswertung

Kovarianz

Die Kovarianz beschreibt den gemeinsamen linearen Zusammenhang zwischen zwei metrischen Variablen.

Hinweis: Da der vorliegende Datensatz nur eine metrische Variable enthält, kann keine Kovarianz berechnet werden.

Korrelationskoeffizient

Der Korrelationskoeffizient nach Pearson misst Stärke und Richtung eines linearen Zusammenhangs zwischen zwei metrischen Variablen.

Hinweis: Mangels einer zweiten metrischen Variable ist die Berechnung eines Korrelationskoeffizienten im vorliegenden Datensatz nicht möglich.

Klasseneinteilung und Histogramm

Für die metrische Variable Landwirtschaftliche Betriebe Anzahl wird eine sinnvolle Klasseneinteilung definiert und ein Histogramm erstellt. Da die Daten stark rechtsschief verteilt sind (viele kleine Werte, wenige sehr große Werte), werden im unteren Wertebereich kleinere Klassenbreiten gewählt und im oberen Wertebereich größere Klassenbreiten, um sowohl den zentralen Datenbereich als auch die Ausreißer sinnvoll darzustellen.

	Klasse	Häufigkeit
0	[0, 5000)	17
1	[5000, 10000)	7
2	[10000, 20000)	1
3	[20000, 50000)	12
4	[50000, 100000)	4
5	[100000, 150000)	4
6	[150000, 200000)	2
7	[200000, 260000)	2

Abbildung 12 Klassentabelle

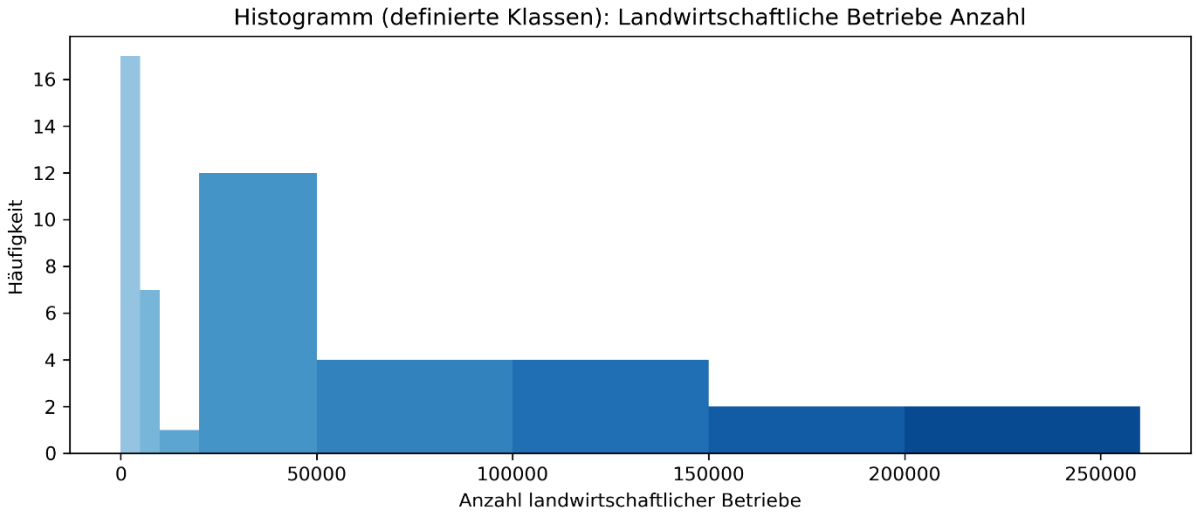

Abbildung 13 Histogramm

Begründung der Klasseneinteilung: Die gewählten Klassen sind im unteren Wertebereich enger (0–5k, 5–10k, 10–20k), da dort der Großteil der Beobachtungen liegt. Für höhere Werte wurden breitere Klassen (z. B. 50k–100k, 100k–150k, ...) gewählt, da dort deutlich weniger Werte vorliegen und ansonsten viele Klassen leer wären. Dadurch bleibt das Histogramm interpretierbar und zeigt sowohl den zentralen Bereich als auch die Ausreißer.

Interpretation: Das Histogramm bestätigt die stark rechtsschiefe Verteilung: viele Beobachtungen liegen in den unteren Klassen, während nur wenige Bodennutzungsarten sehr hohe Betriebszahlen aufweisen.

Kontingenztabelle

Eine Kontingenztabelle dient zur Darstellung der gemeinsamen Häufigkeitsverteilung zweier kategorialer Variablen. Im vorliegenden Datensatz ist jedoch nur eine kategoriale Variable (Bodennutzungsarten) vorhanden. Da keine zweite kategoriale Variable existiert, kann keine Kontingenztabelle erstellt werden.

Rangkorrelationskoeffizient nach Spearman

Der Rangkorrelationskoeffizient nach Spearman misst den monotonen Zusammenhang zwischen zwei ordinalen oder metrischen Variablen auf Basis ihrer Ränge. Da der vorliegende Datensatz nur eine metrische Variable enthält und keine zweite geeignete Variable für einen Rangvergleich vorhanden ist, kann der Rangkorrelationskoeffizient nach Spearman nicht berechnet werden.

3.6 Zusammenfassung

Der vorliegende Datensatz beschreibt die Anzahl landwirtschaftlicher Betriebe in Abhängigkeit von verschiedenen Bodennutzungsarten. Insgesamt umfasst der Datensatz 49 Beobachtungen mit einer nominalskalierten Variablen (Bodennutzungsarten) sowie einer metrischen Variablen (Landwirtschaftliche Betriebe Anzahl). Die Auswertung der Lageparameter zeigt, dass für die Variable Landwirtschaftliche Betriebe Anzahl kein eindeutiger Modus existiert, da alle Werte nur einmal im Datensatz vorkommen. Der arithmetische Mittelwert liegt deutlich über dem Median, was auf eine stark rechtsschiefe Verteilung der Daten hinweist. Diese Einschätzung wird durch die

grafische Analyse mittels Box-Whisker-Plot bestätigt. Die Spannweite der Daten ist sehr groß, was auf erhebliche Unterschiede in der Anzahl landwirtschaftlicher Betriebe zwischen den einzelnen Bodennutzungsarten schließen lässt. Auch die mittlere Abweichung vom Median sowie die berechnete Stichprobenvarianz weisen auf eine starke Streuung der Werte hin. Besonders einige Bodennutzungsarten stellen ausgeprägte Ausreißer mit sehr hohen Betriebszahlen dar. Der Scatterplot verdeutlicht die Verteilung der Anzahl landwirtschaftlicher Betriebe über die einzelnen Beobachtungen hinweg, ohne dass ein funktionaler Zusammenhang erkennbar ist. Dies ist aufgrund der Verwendung der Beobachtungsnummer als unabhängige Variable erwartungsgemäß. Eine ergänzende grafische Darstellung in Form eines Balkendiagramms zeigt die Anzahl landwirtschaftlicher Betriebe je Bodennutzungsart und macht die starken Unterschiede zwischen den Kategorien deutlich. Insgesamt lässt sich festhalten, dass die Daten durch eine stark ungleichmäßige Verteilung mit wenigen sehr großen und vielen vergleichsweise kleinen Werten geprägt sind.

4. Datensatz 2: Landwirtschaftliche Betriebe und Bodennutzungsarten (ökologisch)

4.1 Beschreibung des Datensatzes

Der vorliegende Datensatz umfasst **52 Beobachtungen** und enthält **zwei Variablen**. Die Variable **Bodennutzungsarten** beschreibt unterschiedliche landwirtschaftliche Nutzungsformen in kategorialer Form. Die Variable **Landwirtschaftliche Betriebe mit ökologischem Landbau Anzahl** gibt die Anzahl landwirtschaftlicher Betriebe an, die ökologischen Landbau betreiben, bezogen auf die jeweilige Bodennutzungsart.

Die Daten liegen in tabellarischer Form als CSV-Datei vor und wurden im Rahmen dieser Projektarbeit mit der Programmiersprache Python unter Verwendung der Bibliothek *pandas* ausgewertet.

4.2 Bereinigte Daten

Nach dem Einlesen der Rohdaten wurden diese bereinigt und liegen anschließend in konsistenter Form vor. Der Datensatz enthält nach der Bereinigung keine fehlenden Werte mehr und die metrische Variable liegt in numerischer Form vor.

Die bereinigten Daten bilden die Grundlage für alle weiteren statistischen Auswertungen.

4.3 Dokumentation der Datenbereinigung

Im Zuge der Datenbereinigung wurden mehrere Schritte durchgeführt. Zunächst wurden Zeilenumbrüche aus den Spaltennamen entfernt, um eine einheitliche Weiterverarbeitung zu gewährleisten. Darüber hinaus enthielt der Datensatz eine zusätzliche Kopfzeile, die keine Beobachtungen darstellte und daher entfernt wurde. Fehlende Werte waren durch ein Sonderzeichen („.“) gekennzeichnet. Diese Einträge wurden als fehlende Werte interpretiert und aus dem Datensatz entfernt.

Abschließend wurde der Datentyp der metrischen Variable in einen numerischen Datentyp umgewandelt. Nach diesen Maßnahmen liegt ein vollständig bereinigter Datensatz vor.

Die bereinigten Daten wurden als konsolidierte Excel-Datei im .xlsx-Format gespeichert. Diese Datei stellt di verbindliche Datenbasis für alle weiteren statistischen Auswertungen dar.

4.4 Variablen und Skalenniveaus

Bodennutzungsarten

Skalenniveau: Nominalskala

Begründung: Die Ausprägungen stellen unterschiedliche Kategorien landwirtschaftlicher Nutzungsformen dar, zwischen denen keine natürliche Reihenfolge besteht.

Landwirtschaftliche Betriebe mit ökologischem Landbau Anzahl

Skalenniveau: Metrische Skala (Verhältnisskala, diskret)

Begründung: Es handelt sich um absolute Häufigkeiten mit einem sinnvollen Nullpunkt. Abstände und Verhältnisse zwischen den Werten sind interpretierbar.

4.5 Ur- und Ranglisten

Für beide Variablen des Datensatzes wurden Urlisten erstellt. Die Urlisten enthalten die beobachteten Werte in ihrer ursprünglichen Reihenfolge, ohne Sortierung oder statistische Verarbeitung, und wurden jeweils als separate CSV-Dateien gespeichert.

Auf Grundlage der Urlisten wurden Ranglisten erstellt. Die nominalskalierte Variable wurde alphabetisch sortiert, während die metrische Variable aufsteigend nach ihrer Größe geordnet wurde. Die Ranglisten wurden als CSV-Dateien gespeichert.

4.6 Lage- und Streuungsparameter

Lageparameter – Landwirtschaftliche Betriebe mit ökologischem Landbau Anzahl

Für die metrische Variable werden Modus, arithmetischer Mittelwert und Median berechnet und tabellarisch dargestellt

	Lageparameter	Wert
0	Modus	kein eindeutiger Modus
1	Arithmetischer Mittelwert	2523.0
2	Median	930.0

Abbildung 14 Tabelle Betriebe

Interpretation:
Der arithmetische Mittelwert beschreibt die durchschnittliche Anzahl landwirtschaftlicher Betriebe mit ökologischem Landbau pro Bodennutzungsart. Der Median teilt die Verteilung in zwei gleich große Hälften und ist robuster gegenüber Ausreißern.

Erläuterung zum Modus:

Ein eindeutiger Modus existiert nicht, da jeder beobachtete Wert nur einmal im Datensatz vorkommt. Der Modus ist daher statistisch nicht aussagekräftig.

Lageparameter-Bodennutzungsarten

Da es sich bei den Bodennutzungsarten um eine nominalskalierte Variable handelt, sind arithmetischer Mittelwert und Median nicht sinnvoll definiert.

Erläuterung zum Modus:

Da jede Bodennutzungsart im Datensatz genau einmal vorkommt, existiert kein eindeutiger Modus. Eine Charakterisierung der Verteilung über den Modus ist daher nicht möglich.

Spannweite – Landwirtschaftliche Betriebe mit ökologischem Landbau Anzahl

Die Spannweite beschreibt den gesamten Wertebereich der metrischen Variable und ergibt sich aus der Differenz zwischen dem größten und dem kleinsten beobachteten Wert.

	Kennzahl	Wert
0	Minimum	20
1	Maximum	26090
2	Spannweite	26070

Abbildung 15 Spannweite Betriebe

Interpretation:
Die Spannweite zeigt, dass erhebliche Unterschiede in der Anzahl landwirtschaftlicher Betriebe zwischen den einzelnen Bodennutzungsarten bestehen. Sie reagiert empfindlich auf Extremwerte und sollte daher stets in Kombination mit weiteren Streuungsmaßen interpretiert werden.

Hinweis:
Für die nominalskalierte Variable Bodennutzungsarten ist die Spannweite nicht definiert.

Streuungsmaß: Mittlere Abweichung vom Median

Die mittlere Abweichung vom Median beschreibt die durchschnittliche absolute Abweichung der Beobachtungen vom Median der Verteilung.

	Kennzahl	Wert
0	Median	930.0
1	Mittlere Abweichung vom Median	2185.0

Abbildung 16 Mittlere Abweichung Betriebe

Interpretation:
Dieses Streuungsmaß ist robuster gegenüber Ausreißern als die Spannweite und eignet sich gut zur Beschreibung der Streuung um den zentralen Bereich der Daten.

Hinweis:
Für die nominalskalierte Variable Bodennutzungsarten ist die mittlere Abweichung vom Median nicht definiert.

Streuungsmaß: Stichprobenvarianz

Die Stichprobenvarianz misst die durchschnittliche quadrierte Abweichung der Werte vom arithmetischen Mittelwert.

	Kennzahl	Wert
0	Arithmetischer Mittelwert	2.523000e+03
1	Stichprobenvarianz	2.599784e+07

Abbildung 17 Stichprobenvarianz Betriebe

Interpretation:
Ein hoher Varianzwert weist auf eine starke Streuung der Anzahl landwirtschaftlicher Betriebe mit ökologischem Landbau zwischen den einzelnen Bodennutzungsarten hin.

Hinweis:
Für die nominalskalierte Variable Bodennutzungsarten ist die Stichprobenvarianz nicht definiert.

Variationskoeffizient

	Kennzahl	Wert
0	Standardabweichung	5098.807884
1	Arithmetischer Mittelwert	2523.000000
2	Variationskoeffizient	2.020931

Abbildung 18 Variationskoeffizient

Der Variationskoeffizient ist ein dimensionsloses Streuungsmaß und wird als Quotient aus der Standardabweichung und dem arithmetischen Mittelwert berechnet.

Er gibt an, wie groß die Streuung im Verhältnis zum Mittelwert ist, und eignet sich insbesondere zum Vergleich der Streuung verschiedener Datensätze oder Variablen.

Für nominalskalierte Variablen ist der Variationskoeffizient nicht definiert.

Lage- und Streuungsparameter: Quartile und Dezile

Zur detaillierten Beschreibung der Verteilung werden die Quartile und Dezile der metrischen Variable berechnet.

	Quantil	Wert
0	Q1 (25 %)	312.5
1	Q2 (Median)	930.0
2	Q3 (75 %)	2000.0
3	D1 (10 %)	109.0
4	D2 (20 %)	254.0
5	D3 (30 %)	381.0
6	D4 (40 %)	508.0
7	D5 (50 %)	930.0
8	D6 (60 %)	1214.0
9	D7 (70 %)	1584.0
10	D8 (80 %)	2820.0
11	D9 (90 %)	5781.0

Abbildung 19 Streuungsparamter

Interpretation:
Die Quartile und Dezile verdeutlichen eine starke Konzentration der Werte im unteren Bereich der Verteilung. Ein Großteil der Beobachtungen liegt deutlich unterhalb des arithmetischen Mittelwertes, was auf eine rechtsschiefe Verteilung hinweist.

Streuungsmaß: Quartilsabstand

Der Quartilsabstand beschreibt die Spannweite der mittleren 50 % der Daten und ist robust gegenüber Ausreißern.

	Kennzahl	Wert
0	Q1 (25 %)	312.5
1	Q3 (75 %)	2000.0
2	Quartilsabstand	1687.5

Abbildung 20 Quartilsabstand

Interpretation:
Die mittleren 50 % der Werte liegen vergleichsweise eng beieinander, während die Gesamtstreuung stark durch einzelne hohe Werte beeinflusst wird.

Grafische Darstellung

Box-Whisker-Plots – Landwirtschaftliche Betriebe mit ökologischem Landbau Anzahl

Zur Analyse der Verteilung werden zwei Box-Whisker-Plots betrachtet: eine vollständige Darstellung inklusive Ausreißern sowie eine ergänzende Darstellung ohne Ausreißer zur besseren Sichtbarkeit des zentralen Datenbereichs.

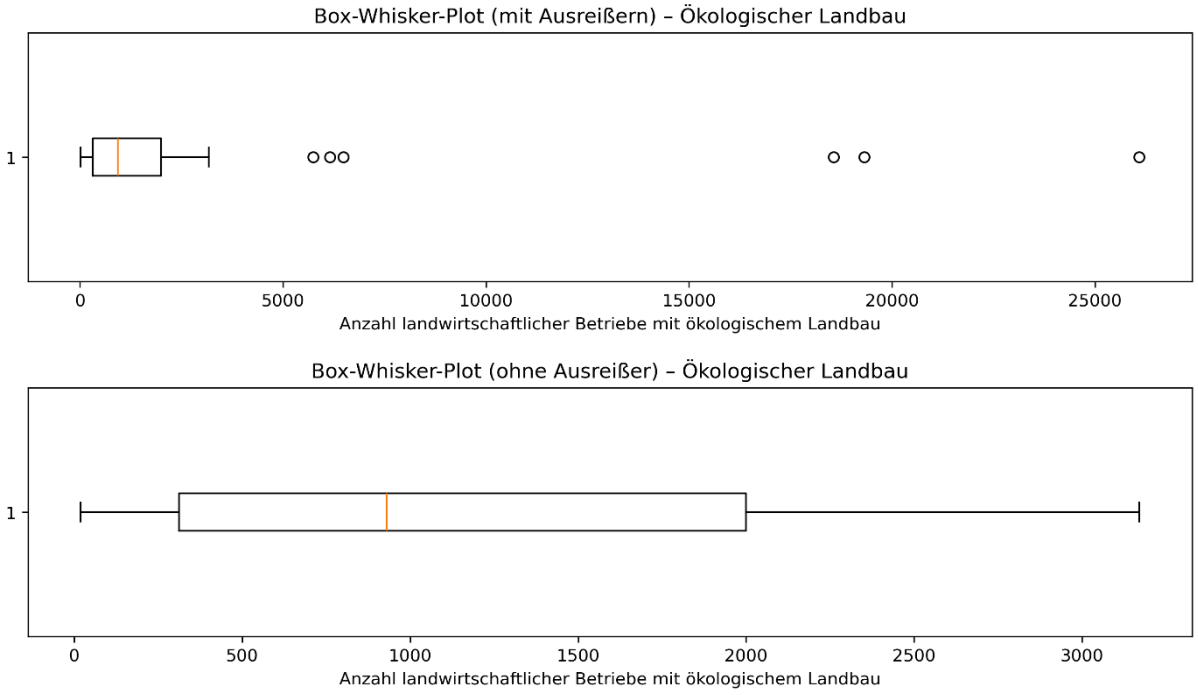

Abbildung 21Boxplots Betriebe

Interpretation:
Die vollständige Darstellung zeigt eine stark rechtsschiefe Verteilung mit mehreren ausgeprägten Ausreißern. Die ergänzende Darstellung ohne Ausreißer ermöglicht eine bessere Beurteilung von Lage und Streuung des zentralen Datenbereichs.

Scatterplot – Landwirtschaftliche Betriebe mit ökologischem Landbau Anzahl

Der Scatterplot zeigt die Anzahl landwirtschaftlicher Betriebe in Abhängigkeit von der Beobachtungsnummer.

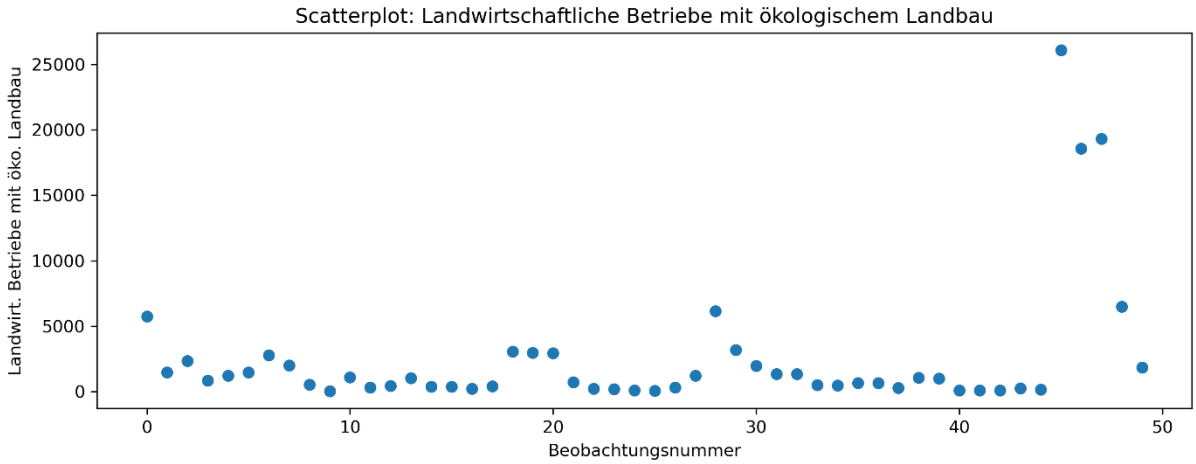

Abbildung 22 Scatterplott Betriebe

Interpretation:
Der Scatterplot verdeutlicht die starke Streuung der Werte. Ein funktionaler Zusammenhang ist nicht erkennbar, was aufgrund der Verwendung der Beobachtungsnummer als unabhängige Variable erwartungsgemäß ist.

Ergänzende Darstellung: Landwirtschaftliche Betriebe nach Bodennutzungsart

Zur inhaltlichen Einordnung der Daten wird ein horizontales Balkendiagramm erstellt, das die Anzahl landwirtschaftlicher Betriebe mit ökologischem Landbau pro Bodennutzungsart darstellt.

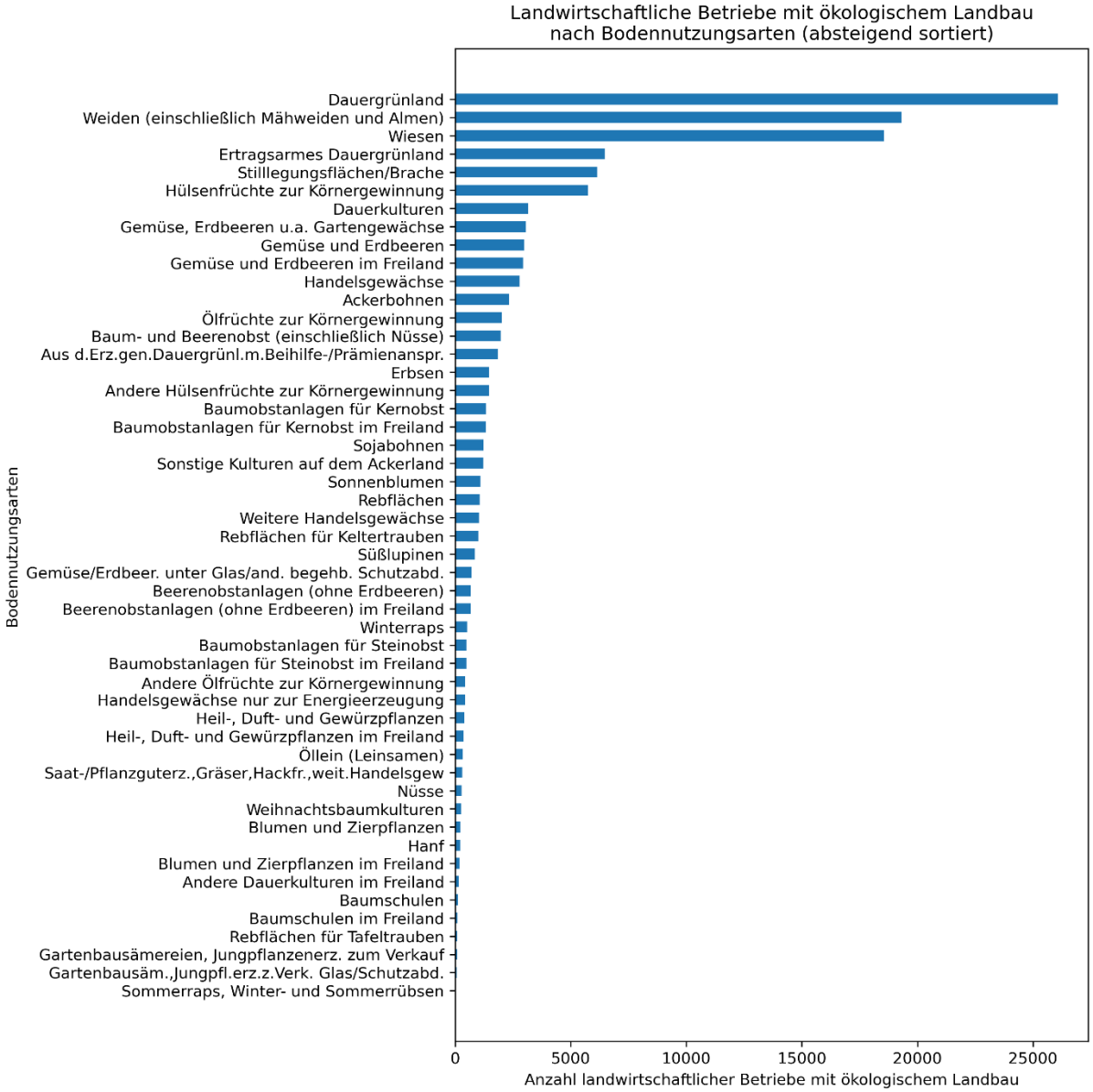

Abbildung 23 Balkendiagramm

Interpretation:
Das Balkendiagramm macht die starken Unterschiede zwischen den einzelnen Bodennutzungsarten deutlich. Wenige Nutzungsarten weisen sehr hohe Werte auf, während der Großteil der Kategorien deutlich niedrigere Fallzahlen besitzt.

Zusammenfassung

Der vorliegende Datensatz beschreibt die Anzahl landwirtschaftlicher Betriebe mit ökologischem Landbau in Abhängigkeit von verschiedenen Bodennutzungsarten. Insgesamt umfasst der Datensatz 52 Beobachtungen mit einer nominalskalierten und einer metrischen Variable.

Die Auswertung der Lageparameter zeigt, dass kein eindeutiger Modus existiert und dass der arithmetische Mittelwert deutlich über dem Median liegt. Dies weist auf eine stark rechtsschiefe Verteilung hin, die durch die grafische Analyse bestätigt wird.

Die Streuungsmaße zeigen eine hohe Heterogenität der Daten. Insbesondere wenige Bodennutzungsarten stellen ausgeprägte Ausreißer mit sehr hohen Betriebszahlen dar. Insgesamt ist der Datensatz durch eine stark ungleichmäßige Verteilung mit vielen kleinen und wenigen sehr großen Werten geprägt.

5. Datensatz 3: Überlebensraten von Unternehmen nach Wirtschaftszweigen

5.1 Beschreibung des Datensatzes

Der vorliegende Datensatz beschreibt die Anzahl überlebender Unternehmen sowie deren Überlebensrate in Prozent, differenziert nach verschiedenen Wirtschaftszweigen.

Die Daten liegen in zwei separaten CSV-Dateien vor, die jeweils eine gemeinsame Identifikationsvariable enthalten. Eine Datei enthält die Bezeichnungen der Wirtschaftszweige, die zweite Datei die zugehörigen numerischen Angaben. Im Rahmen dieser Projektarbeit wurden die beiden Datensätze anhand der gemeinsamen Identifikationsvariable zusammengeführt und gemeinsam ausgewertet.

Der Datensatz liegt in tabellarischer Form vor. Jede Zeile beschreibt einen Wirtschaftszweig. Die Daten liegen zunächst in Rohform vor und enthalten formatierungsbedingte Besonderheiten, die im weiteren Verlauf bereinigt werden.

5.2 Datenaufbereitung

Die Daten liegen in zwei separaten CSV-Dateien vor. Beide Dateien wurden mithilfe der Programmiersprache Python eingelesen und anschließend anhand der gemeinsamen Identifikationsvariable zusammengeführt. Dadurch entsteht ein konsolidierter Datensatz, der sowohl die Wirtschaftszweige als auch die zugehörigen numerischen Angaben enthält.

Nach dem Einlesen und Zusammenführen der Daten wurden diese bereinigt. Eine Prüfung auf fehlende Werte ergab, dass im bereinigten Datensatz keine fehlenden Einträge vorhanden sind.

Im Rahmen der Datenbereinigung wurden folgende Maßnahmen durchgeführt:

- Vereinheitlichung der Spaltennamen,
- Korrektur von Zeichenkodierungsproblemen (z. B. Umlaute),
- Umwandlung der numerischen Variablen in geeignete Datentypen,
- Anpassung des Dezimaltrennzeichens zur korrekten numerischen Verarbeitung.

Alle Bereinigungsschritte wurden nachvollziehbar im Notebook dokumentiert.

Der bereinigte und zusammengeführte Datensatz wurde zusätzlich als konsolidierte Excel-Datei im .xlsx-Format gespeichert.

Für die Analyse wurde die Programmiersprache Python in einer Jupyter-Notebook-Umgebung verwendet. Zur Datenverarbeitung und -analyse kamen insbesondere die Bibliotheken pandas und numpy zum Einsatz. Die grafischen Darstellungen wurden mithilfe der Bibliothek matplotlib erstellt.

Für beide metrischen Variablen wurden jeweils Urlisten und Ranglisten erstellt. Die Urlisten enthalten die unveränderten Messwerte in der Reihenfolge ihres Auftretens, während die Ranglisten die Werte in aufsteigender Reihenfolge darstellen.

Die Urlisten und Ranglisten wurden mithilfe von Python erzeugt und jeweils als separate CSV-Dateien gespeichert. Sie dienen als Grundlage für die weiteren statistischen Auswertungen.

5.3 Lagekennwerte

Für die beiden metrischen Variablen wurden die Lagekennwerte Modus, arithmetischer Mittelwert und Median bestimmt. Die Anzahl überlebender Unternehmen weist einen Modus von 1 auf. Der arithmetische Mittelwert beträgt 65,04, während der Median bei 22,5 liegt. Der deutlich höhere Mittelwert im Vergleich zum Median deutet auf eine rechtsschiefe Verteilung hin, bei der wenige Wirtschaftszweige eine sehr hohe Anzahl überlebender Unternehmen aufweisen.

Die Überlebensrate der Unternehmen besitzt einen Modus von 100,0 %. Der arithmetische Mittelwert liegt bei 88,94 %, der Median bei 90,3 %. Die Nähe von Median und Mittelwert deutet auf eine vergleichsweise homogene Verteilung der Überlebensraten hin.

5.4 Streuungsmaße

Zur Beschreibung der Streuung der beiden metrischen Variablen wurden die Spannweite, die mittlere Abweichung vom Median, die Stichprobenvarianz sowie der Variationskoeffizient bestimmt.

Die Anzahl überlebender Unternehmen weist eine Spannweite von 313 auf. Die mittlere Abweichung vom Median beträgt 56,75, die Stichprobenvarianz 8023,74. Der Variationskoeffizient von 1,38 zeigt eine sehr hohe relative Streuung der Werte. Dies bestätigt, dass sich die Anzahl überlebender Unternehmen stark zwischen den Wirtschaftszweigen unterscheidet.

Die Überlebensrate der Unternehmen besitzt eine Spannweite von 56,3. Die mittlere Abweichung vom Median beträgt 7,39, die Stichprobenvarianz 120,27. Der Variationskoeffizient von 0,12 weist auf eine vergleichsweise geringe relative Streuung der Überlebensraten hin.

5.5 Grafische Darstellungen

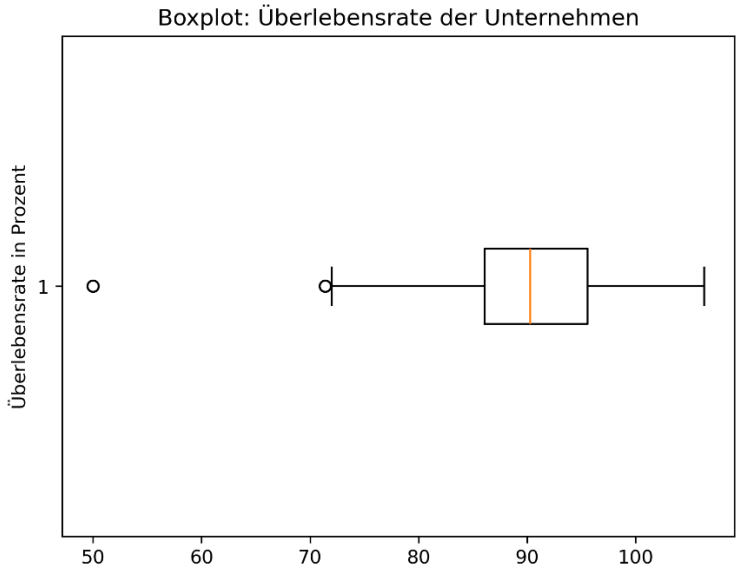

Abbildung 24 Anzahl überlebender Unternehmen

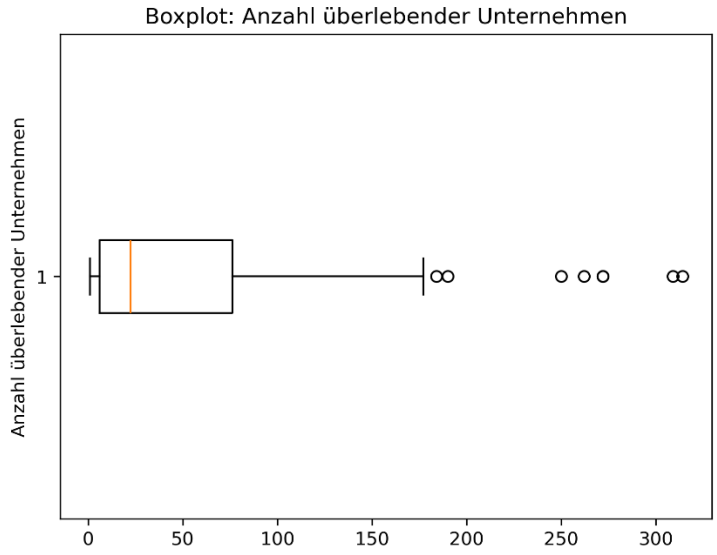

Abbildung 25 Überlebensrate in Prozent

Die Box-Whisker-Plots verdeutlichen die unterschiedlichen Verteilungseigenschaften der beiden Variablen. Die Anzahl überlebender Unternehmen zeigt eine stark asymmetrische Verteilung mit großer Streuung. Die Überlebensrate weist hingegen eine kompaktere Verteilung mit geringerer Streuung auf.

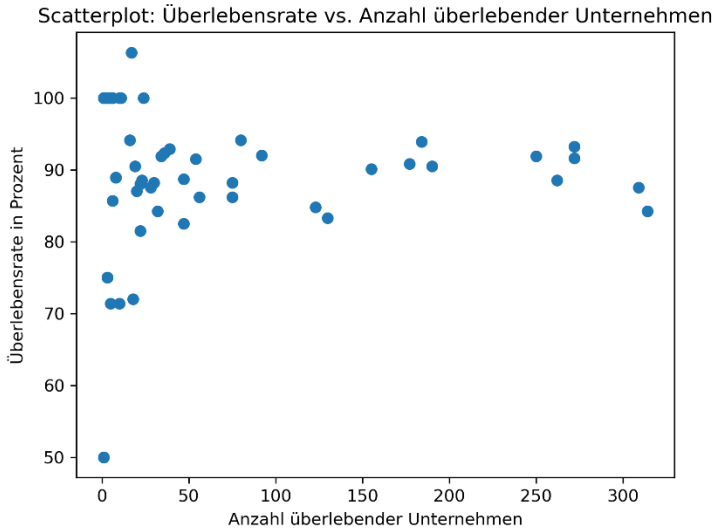

Abbildung 26 Scatterplot – Rohdaten

Der Scatterplot zeigt den Zusammenhang zwischen der Anzahl überlebender Unternehmen und der Überlebensrate. Es ist kein klarer funktionaler Zusammenhang erkennbar, da die Datenpunkte stark streuen.

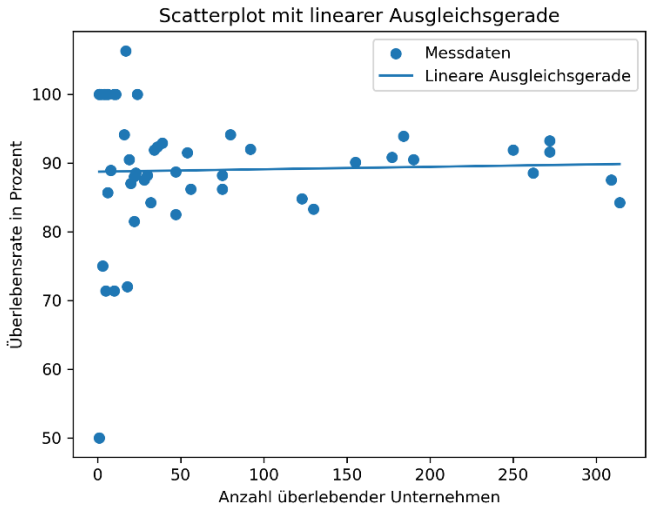

Abbildung 27 Scatterplot mit linearer Ausgleichsgerade und Legende

Zur Beschreibung eines möglichen Trends wurde ein lineares Curve Fitting durchgeführt. Die lineare Ausgleichsgerade weist nur eine sehr geringe Steigung auf, was auf einen schwachen linearen Zusammenhang zwischen der Anzahl überlebender Unternehmen und der Überlebensrate hinweist.

Die im Diagramm dargestellten Messdaten sowie die lineare Ausgleichsgerade sind durch eine Legende eindeutig gekennzeichnet.

5.6 Zusatzkennwerte

Zur weiterführenden Beschreibung der Verteilung wurden für beide metrischen Variablen Quartile und Dezile bestimmt.

Die Anzahl überlebender Unternehmen weist eine stark rechtsschiefe Verteilung auf. Das erste Quartil liegt bei 6, der Median bei 22,5 und das dritte Quartil bei 76,25. Die oberen Dezile erreichen deutlich höhere Werte, was auf einzelne Wirtschaftszweige mit sehr hohen Unternehmenszahlen hinweist.

Die Überlebensrate der Unternehmen ist deutlich homogener verteilt. Bereits das erste Quartil liegt über 86 %, und ab dem achten Dezil werden Überlebensraten von 100 % erreicht.

Quartilsabstand

Anzahl überlebender Unternehmen

- Quartilsabstand: 70,25

Überlebensrate in Prozent

- Quartilsabstand: 9,50

Der Quartilsabstand der Anzahl überlebender Unternehmen beträgt 70,25 und zeigt eine starke Streuung der mittleren 50 % der Werte. Der Quartilsabstand der Überlebensrate liegt bei 9,50 und ist damit deutlich geringer.

Kovarenz

Kovarianz = 28,63

Die Kovarianz zwischen der Anzahl überlebender Unternehmen und der Überlebensrate ist positiv, jedoch von geringem Betrag. Dies deutet auf einen sehr schwachen gemeinsamen linearen Trend der beiden Variablen hin.

Korrelationskoeffizient

$$r \approx 0{,}03$$

Der berechnete Pearson-Korrelationskoeffizient beträgt etwa 0,03 und liegt damit nahe bei null. Es besteht somit kein nennenswerter linearer Zusammenhang zwischen der Anzahl überlebender Unternehmen und der Überlebensrate.

Zusammenfassung

Der Datensatz 3 wurde aus zwei zusammengehörigen CSV-Dateien erstellt, zusammengeführt und bereinigt. Die bereinigten Daten bilden die Grundlage für die statistische Auswertung der Anzahl überlebender Unternehmen sowie deren Überlebensrate in verschiedenen Wirtschaftszweigen.

Die Analyse der Lagekennwerte zeigt, dass sich die Anzahl überlebender Unternehmen stark zwischen den Wirtschaftszweigen unterscheidet. Der arithmetische Mittelwert liegt deutlich über dem Median, was auf eine rechtsschiefe Verteilung hinweist. Die Überlebensrate der Unternehmen ist hingegen vergleichsweise homogen verteilt und weist insgesamt ein hohes Niveau auf.

Die Streuungsmaße bestätigen diese Beobachtungen. Während die Anzahl überlebender Unternehmen eine hohe absolute und relative Streuung aufweist, ist die Streuung der Überlebensrate deutlich geringer. Auch die grafischen Darstellungen unterstützen diese Ergebnisse.

Der Scatterplot sowie die durchgeführte Zusammenhangsanalyse zeigen keinen ausgeprägten linearen Zusammenhang zwischen der Anzahl überlebender Unternehmen und der Überlebensrate. Das lineare Curve Fitting ergibt eine nahezu horizontale Ausgleichsgerade, und auch der Korrelationskoeffizient liegt nahe bei null.

Zusammenfassend lässt sich feststellen, dass sich die Überlebensrate von Unternehmen zwischen den Wirtschaftszweigen weniger stark unterscheidet als die absolute Anzahl überlebender Unternehmen. Ein signifikanter linearer Zusammenhang zwischen beiden Variablen ist nicht erkennbar.

6. Datensatz 4: Simulierte Temperaturmessung

6.1 Beschreibung des Datensatzes

Der Datensatz 4 beschreibt eine eindimensionale Messreihe von Temperaturwerten, die eine angenommene Messung der Außentemperatur auf einem Balkon über den Zeitraum eines Jahres darstellen. Die Messwerte repräsentieren Tagesmittelwerte der Temperatur und wurden in gleichmäßigen zeitlichen Abständen erfasst.

Die Daten wurden im Rahmen dieser Projektarbeit selbst generiert und orientieren sich an realistischen jahreszeitlichen Temperaturverläufen. Ziel dieses Datensatzes ist die statistische Analyse einer einzelnen Messgröße anhand geeigneter Lage- und Streuungskennwerte sowie einer grafischen Darstellung.

6.2 Datenaufbereitung

Der Datensatz wurde nach der Erzeugung auf fehlende Werte überprüft. Dabei wurde festgestellt, dass keine fehlenden Einträge vorhanden sind. Die Daten liegen somit in bereinigter Form vor.

Da die Messwerte bereits strukturiert und in einem geeigneten Formatvorlagen, waren keine weiteren Bereinigungsmaßnahmen erforderlich. Die durchgeführten Prüfschritte wurden im zugehörigen Notebook dokumentiert.

Für die Erzeugung und Analyse der Daten wurde die Programmiersprache Python in einer Jupyter-Notebook-Umgebung verwendet. Zur Datenerzeugung und -verarbeitung kamen insbesondere die Bibliotheken numpy und pandas zum Einsatz.

6.3 Lage- und Streuungskennwerte

Für den Datensatz wurden die Lagekennwerte Modus, arithmetischer Mittelwert und Median sowie die Stichprobenvarianz bestimmt.

Der arithmetische Mittelwert der Temperaturwerte beträgt etwa 12,02 °C, der Median liegt bei etwa 12,34 °C. Die Nähe von Mittelwert und Median weist auf eine weitgehend symmetrische Verteilung der Temperaturwerte hin. Ein eindeutiger Modus ist nicht vorhanden, was bei stetigen Messgrößen wie Temperatur zu erwarten ist.

Die Stichprobenvarianz beträgt etwa 52,59 und beschreibt die Streuung der Temperaturwerte über den betrachteten Zeitraum. Die vergleichsweise hohe Varianz ist durch den jahreszeitlichen Temperaturverlauf mit niedrigen Winter- und hohen Sommertemperaturen erklärbar.

6.4 Grafische Darstellung

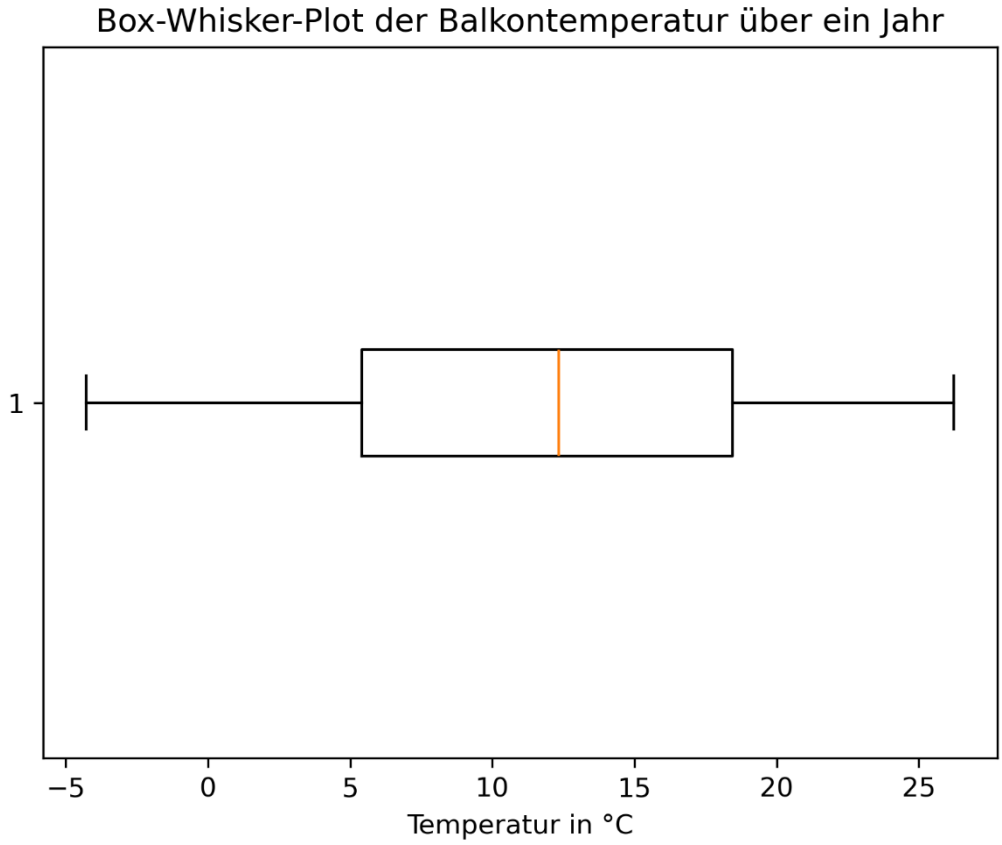

Abbildung 28 Box-Whisker-Plot der Balkontemperatur über den Zeitraum eines Jahres

Der Box-Whisker-Plot zeigt die Verteilung der Temperaturwerte über den betrachteten Zeitraum eines Jahres. Der Median liegt annähernd in der Mitte der Box, was auf eine weitgehend symmetrische Verteilung der Temperaturwerte hindeutet.

Die große Spannweite zwischen Minimum und Maximum ist durch die jahreszeitlichen Temperaturunterschiede zwischen Winter- und Sommermonaten erklärbar.

6.5 Zusammenfassung

Der Datensatz 4 beschreibt eine eindimensionale Messreihe von Temperaturwerten, die eine angenommene Balkonmessung über den Zeitraum eines Jahres darstellen. Insgesamt umfasst der Datensatz 365 tägliche Messwerte.

Die statistische Auswertung zeigt, dass der arithmetische Mittelwert der Temperaturwerte bei etwa 12,02 °C liegt und der Median mit etwa 12,34 °C nahe beim Mittelwert liegt. Dies deutet auf eine weitgehend symmetrische Verteilung der Temperaturwerte hin. Die Stichprobenvarianz von etwa 52,59 verdeutlicht die Streuung der Temperaturwerte über den Jahresverlauf.

Der Box-Whisker-Plot bestätigt diese Ergebnisse und zeigt eine große Spannweite der Temperaturwerte, die auf die jahreszeitlichen Temperaturunterschiede zwischen Winter- und Sommermonaten zurückzuführen ist. Der Datensatz eignet sich damit gut zur Demonstration der deskriptiven statistischen Analyse einer einzelnen Messgröße.

7. Gesamtfazit

Im Rahmen dieser Projektarbeit wurden mehrere Datensätze mit unterschiedlicher Struktur und Zielsetzung statistisch analysiert. Ziel der Arbeit war es, grundlegende Methoden der deskriptiven Statistik anzuwenden, Daten aufzubereiten, geeignete Kennwerte zu bestimmen und die Ergebnisse sowohl numerisch als auch grafisch darzustellen.

Die untersuchten Datensätze unterschieden sich deutlich hinsichtlich ihrer Dimensionalität und Aussagekraft. Während die ersten Datensätze primär zur Analyse von Lage- und Streuungskennwerten dienten, wurde im Datensatz 3 zusätzlich der Zusammenhang zwischen zwei Variablen mithilfe von Scatterplots, Curve Fitting und Korrelationsmaßen untersucht. Der Datensatz 4 stellte eine eindimensionale Messreihe dar und eignete sich besonders zur Demonstration der statistischen Auswertung einer einzelnen Messgröße über einen längeren Zeitraum.

Die Ergebnisse zeigen, dass statistische Kennwerte stets im Kontext der zugrunde liegenden Daten interpretiert werden müssen. Insbesondere wurde deutlich, dass ein statistischer Zusammenhang nicht zwangsläufig auf einen kausalen Zusammenhang schließen lässt. Grafische Darstellungen erwiesen sich als wichtige Ergänzung zur numerischen Analyse, da sie Verteilungen, Ausreißer und Trends anschaulich verdeutlichen.

Zusammenfassend hat die Projektarbeit gezeigt, dass eine sorgfältige Datenaufbereitung sowie die bewusste Auswahl geeigneter statistischer Methoden entscheidend für eine aussagekräftige Analyse sind. Die verwendeten Methoden bilden eine solide Grundlage für weiterführende statistische und datenanalytische Fragestellungen.

8. Quellenverzeichnis

- Dataset 1: www-genesis.destatis.de

- Dataset 2: www-genesis.destatis.de

- Dataset 3: www-genesis.destatis.de