

Technische Universität Dortmund

Modul: Wissenschaftliches Arbeiten

Explorative Analyse des Titanic-Datensatzes

Gruppenarbeit

Gruppe B

Verfasser:

Adem Doggaz

Sebastian Koriath

Gruppenmitglieder:

Rica Lungu

Mariem Ben Fadhel

Dozent:

Steffen Maletz

Abgabedatum: 08.02.2026

Inhaltsverzeichnis

1 Einleitung	2
2 Datensatzbeschreibung	2
3 Methodik	2
4 Deskriptive Analyse	3
4.1 Überlebensrate	3
4.2 Altersverteilung	3
5 Analyse von Einflussfaktoren	4
5.1 Geschlecht und Überleben	4
5.2 Ticketklasse	4
5.3 Ticketpreis	5
6 Diskussion	6
7 Fazit	6
8 Literatur	6

1 Einleitung

Der Untergang der RMS Titanic im Jahr 1912 stellt eines der bekanntesten Schiffsunglücke der Geschichte dar. Die Tragödie führte zum Tod von mehr als 1500 Menschen und hatte weitreichende Auswirkungen auf maritime Sicherheitsstandards.

Der Titanic-Datensatz enthält umfangreiche Informationen über die Passagiere des Schiffes. Neben dem Überlebensstatus umfasst der Datensatz verschiedene demographische sowie sozioökonomische Variablen. Diese Daten bieten eine geeignete Grundlage zur Untersuchung von Faktoren, die die Überlebenswahrscheinlichkeit beeinflusst haben könnten.

Ziel dieser Arbeit ist die Durchführung einer explorativen Datenanalyse des bereinigten Titanic-Datensatzes. Dabei werden statistische Methoden sowie Visualisierungstechniken verwendet, um Zusammenhänge zwischen Variablen zu identifizieren. Die Analyse erfolgt unter Verwendung selbst entwickelter Funktionen in der Programmiersprache R.

2 Datensatzbeschreibung

Der verwendete Datensatz basiert auf historischen Passagierdaten der Titanic. Jede Beobachtung entspricht einer einzelnen Person an Bord des Schiffes.

Die wichtigsten Variablen sind:

- **Survived:** Überlebensstatus
- **Pclass:** Ticketklasse
- **Sex:** Geschlecht
- **Age:** Alter
- **Fare:** Ticketpreis
- **Embarked:** Einschiffungshafen

Vor der Analyse wurde eine Datenbereinigung durchgeführt. Fehlende Werte wurden identifiziert und entsprechend behandelt. Zusätzlich wurden Variablenformate angepasst, um eine konsistente Analyse zu gewährleisten.

3 Methodik

Die Analyse wurde mit der Programmiersprache R durchgeführt. Zur strukturierten Durchführung der Untersuchung wurden selbst entwickelte Funktionen verwendet. Zur Untersuchung kategorialer Zusammenhänge wurde der Chi-Quadrat-Test verwendet, während für numerische Unterschiede ein Welch-t-Test zur Anwendung kam.

Die Analyse umfasst folgende Schritte:

1. Explorative Datenanalyse
2. Berechnung statistischer Kennzahlen

3. Visualisierung von Verteilungen
4. Untersuchung von Zusammenhängen zwischen Variablen

Zur Visualisierung wurden Histogramme, Balkendiagramme und Boxplots verwendet. Diese ermöglichen eine intuitive Darstellung der Datenstruktur.

4 Deskriptive Analyse

4.1 Überlebensrate

Die Analyse der Überlebensrate stellt den Ausgangspunkt der Untersuchung dar. Sie zeigt den Anteil der Passagiere, die das Unglück überlebt haben.

Tabelle 1: Überlebensrate nach Geschlecht

Status	Anzahl Passagiere	
	Frauen	Männer
Nicht überlebt	81	233
Überlebt	468	109

Die Ergebnisse zeigen, dass ein erheblicher Teil der Passagiere das Unglück nicht überlebt hat. Dies verdeutlicht die dramatischen Auswirkungen der Katastrophe.

4.2 Altersverteilung

Die Altersstruktur der Passagiere wurde mithilfe eines Histogramms analysiert.

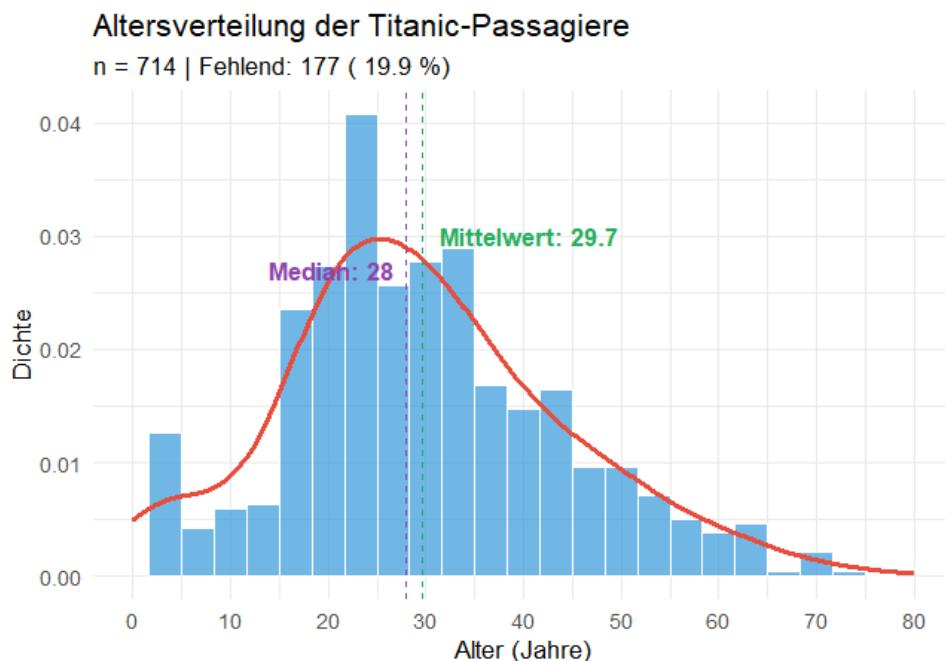


Abbildung 1: Altersverteilung der Passagiere

Die Verteilung zeigt, dass die Mehrheit der Passagiere im jungen Erwachsenenalter lag. Extremwerte sind ebenfalls vorhanden und können auf einzelne Sonderfälle hinweisen.

5 Analyse von Einflussfaktoren

5.1 Geschlecht und Überleben

Die Analyse zeigt deutliche Unterschiede zwischen männlichen und weiblichen Passagieren.

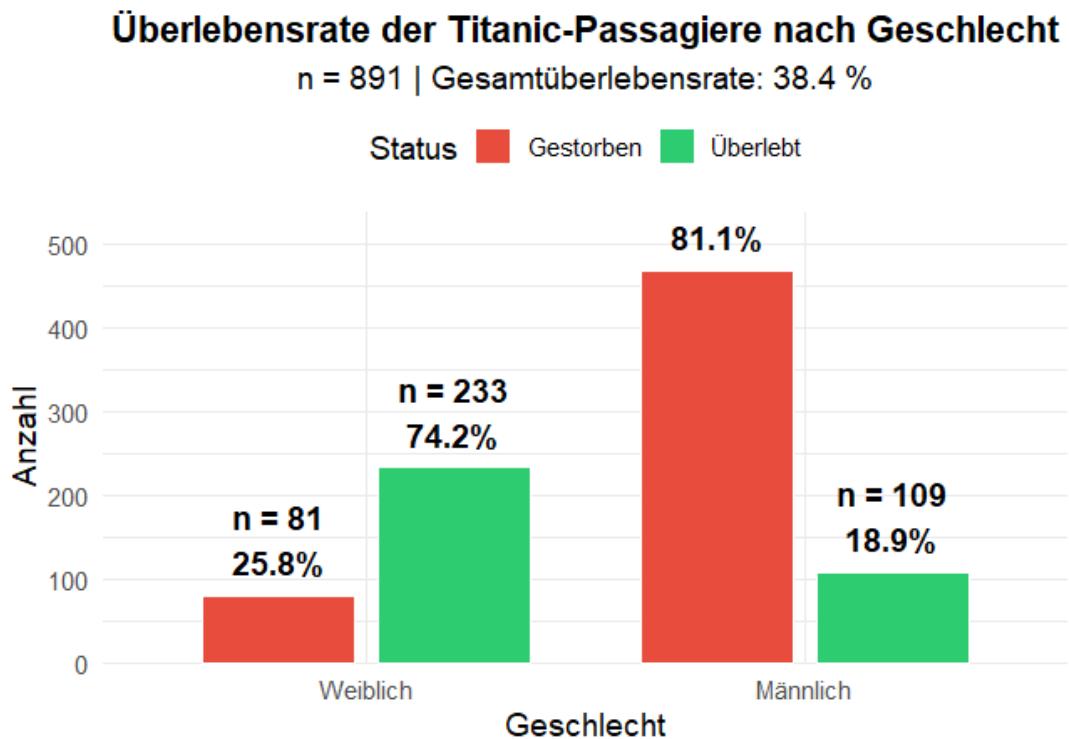


Abbildung 2: Überlebensrate nach Geschlecht

Während nur ein vergleichsweise geringer Anteil männlicher Passagiere überlebte, lag die Überlebensrate weiblicher Passagiere deutlich höher. Dieses Ergebnis stimmt mit historischen Berichten über Evakuierungsprioritäten überein.

5.2 Ticketklasse

Die Ticketklasse stellt einen wichtigen sozioökonomischen Indikator dar.

Überlebensraten nach Passagierklasse und Geschlecht

Gesamtüberlebensrate: 38.4 % | n = 891

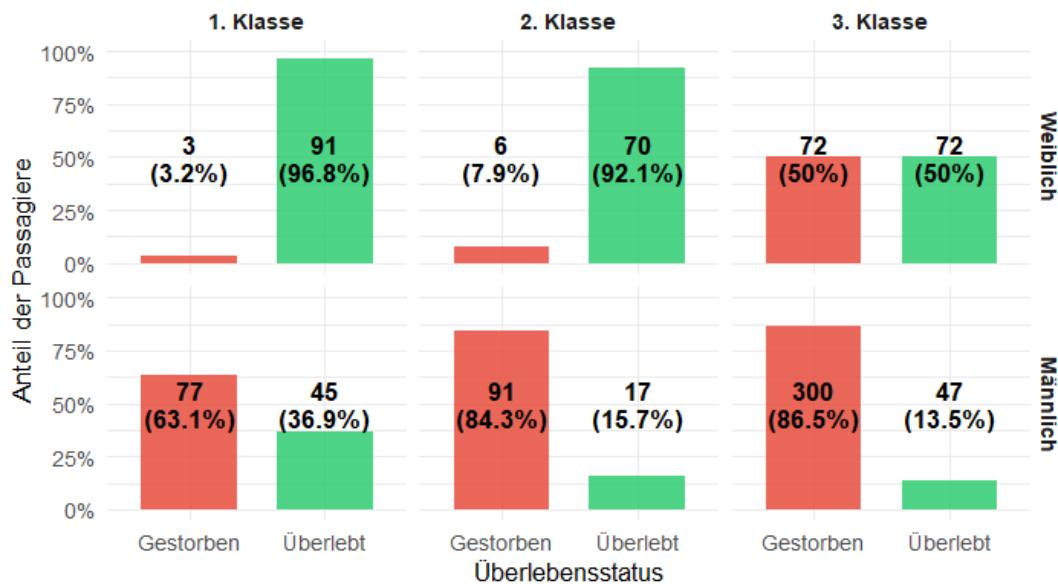


Abbildung 3: Überlebensrate nach Ticketklasse

Passagiere höherer Klassen hatten bessere Überlebenschancen. Dies kann durch räumliche Nähe zu Rettungsbooten sowie organisatorische Vorteile erklärt werden.

5.3 Ticketpreis

Ticketpreis in Abhängigkeit vom Überlebensstatus

Violin = Dichte | Box = Quartile | Gelbe Raute = Mittelwert

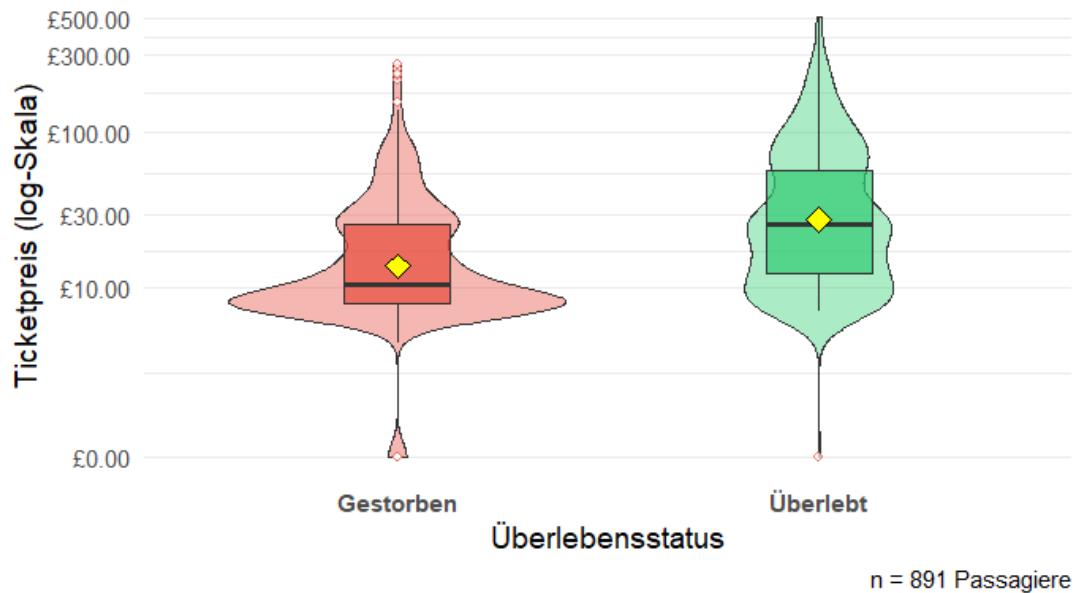


Abbildung 4: Ticketpreis in Abhängigkeit vom Überlebensstatus

Es zeigt sich ein positiver Zusammenhang zwischen Ticketpreis und Überlebenswahrscheinlichkeit. Dies deutet auf einen Zusammenhang zwischen sozialem Status und Überlebenswahrscheinlichkeit hin.

6 Diskussion

Die Ergebnisse der Analyse verdeutlichen den Einfluss sozialer und demographischer Faktoren auf die Überlebenswahrscheinlichkeit. Geschlecht und Ticketklasse zeigen die stärksten Zusammenhänge mit dem Überlebensstatus.

Darüber hinaus deutet der Zusammenhang zwischen Ticketpreis und Überleben auf soziale Ungleichheiten hin. Passagiere mit höherem sozioökonomischem Status hatten tendenziell bessere Chancen, die Katastrophe zu überleben.

Trotz der aufschlussreichen Ergebnisse weist die Analyse Einschränkungen auf. Fehlende Werte sowie mögliche Verzerrungen im Datensatz können die Interpretation beeinflussen. Mögliche Wechselwirkungen zwischen Variablen wurden nicht modelliert, sodass sich die Ergebnisse auf bivariate Zusammenhänge beschränken. Da es sich um explorative Analysen handelt, erlauben die dargestellten Zusammenhänge keine kausalen Schlussfolgerungen, sondern zeigen ausschließlich statistische Assoziationen.

7 Fazit

Diese Arbeit zeigt, dass mehrere Faktoren in Zusammenhang mit der Überlebenswahrscheinlichkeit der Titanic-Passagiere stehen. Besonders Geschlecht, Ticketklasse und Ticketpreis weisen deutliche Zusammenhänge mit dem Überlebensstatus auf.

Die Verwendung selbst entwickelter Funktionen ermöglichte eine strukturierte und reproduzierbare Analyse der Daten. Zukünftige Untersuchungen könnten multivariate statistische Modelle oder maschinelle Lernverfahren einsetzen, um komplexere Zusammenhänge systematisch zu analysieren.

8 Literatur

- Kaggle Titanic Datensatz: <https://www.kaggle.com/c/titanic>
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

Aufgabenverteilung

Aufgabe / Abschnitt	Verfasser
Einleitung	Adem Doggaz
Datensatzbeschreibung	Sebastian Koriath
Methodik	Adem Doggaz
Deskriptive Analyse	Sebastian Koriath
Analyse von Einflussfaktoren	Adem Doggaz
Diskussion	Sebastian Koriath
Fazit	Adem Doggaz
Literatur und Formatierung	Sebastian Koriath