Technische Universität Dortmund

Wintersemester 2025/26

Wissenschaftliches Arbeiten: Bericht über Titanic-Datensatz

Einflussfaktoren auf die Überlebenschancen beim Untergang der Titanic

Verfasser:

Lukas König

Elaha Bahir

John Wilhelm

Annika Homm

08.02.2026

1 Einleitung

Der Untergang der Titanic ist ein weltbekanntes Ereignis in der Geschichte der Menschheit. Es ist Gegenstand zahlreicher historischer und statistischer Forschung. In diesem Bericht soll es darum gehen, durch eine explorative Datenanalyse signifikante Merkmale (wie Geschlecht, Alter, Klasse) der Passagiere auf der Titanic zu identifizieren und Zusammenhänge zum Überleben zu finden.

2 Daten und Methoden

- **Daten:** Zur Datenanalyse wurden der Datensatz titanic.csv und die Pakete "tidyverseünd "ggplot2"verwendet. Vor der Datenbereinigung wies die Variable *Age* 177 NA-Werte (ca. 19,86 %) auf, *Embarked* besaß zwei leere Zellen (ca. 0,22 %) und *Cabin* hatte 687 leere Zellen, was einer Quote von ca. 77,10 % entspricht.
- **Methodik:** Leere Zeichenketten wurden als `NA` kodiert und nicht für die Analyse relevante Variablen (wie *PassengerID, Name, Ticket, Cabin*) wurden nach Merkmalsextraktion entfernt.
  - **Age (Alter in Jahren beim Untergang):** Fehlende Werte wurden durch das arithmetische Mittel der jeweiligen Titelgruppe ersetzt, um die demografische Struktur beizubehalten und keinen Bias zu kreieren.
  - **Embarked (Zustiegshafen):** Zwei fehlende Einträge wurden durch den Modus SSouthhampton"(S) ersetzt.
  - **Cabin (Kabinennummer):** Aus der Kabinennummer wurden das **Deck** (*Deck*) und die **Schiffsseite** (*Side*) mittels Modulo-Operation extrahiert.
  - *Hinweis: Aufgrund der hohen NA-Quote der Variablen Deck und Side und da sie durch Passagiere der dritten Klasse überrepräsentiert (ca. 69,72 %) ist, ist eine Analyse dieser Variablen verzerrt (Selection Bias). Im weiteren Verlauf des Berichts wird daher der Fokus auf die robusteren Variablen wie Geschlecht, Klasse und Alter gelegt.*
- **Datentransformation:** Die Variablen *Sex, Survived, Embarked, Side* und *Deck* wurden als Faktoren kodiert. Die Passagierklasse (*Pclass*) wurde als **ordered factor** definiert, um die Rangfolge abzubilden.

3 Ergebnisse

3.1 Zusammenhang zwischen kategorialen Merkmalen und Überlebensstatus

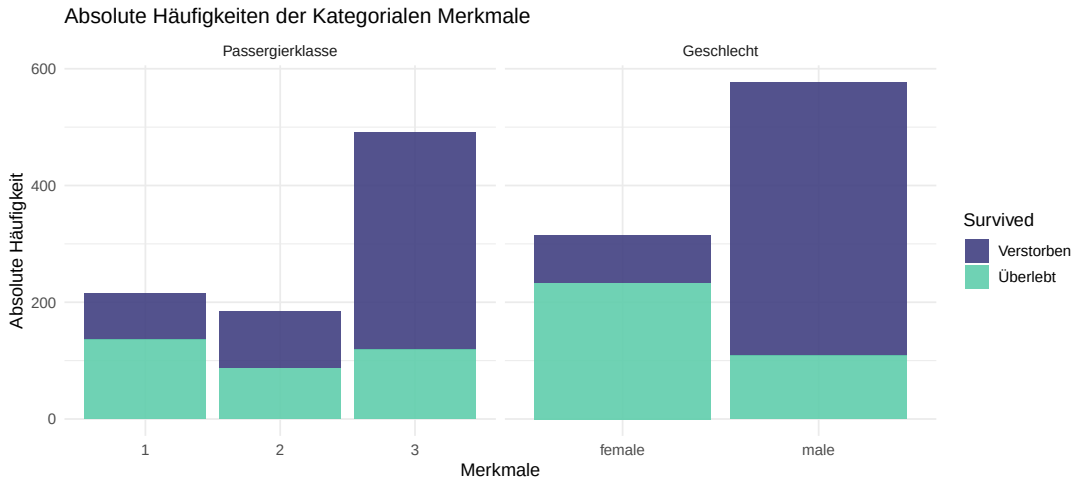

Abbildung 1: Absolute Häufigkeiten kategorialer Daten gruppiert nach Überlebensstatus

Die Demografie der Passagiere an Bord der Titanic bestand zum Großteil aus Männern und der Passagierklasse 3. Mehr als die Hälfte, ca. 61,61 % , haben den Untergang der Titanic nicht überlebt. In der ersten Passagierklasse, hat die Mehrheit überlebt, in der zweiten Passagierklasse war es fast ausgeglichen und in der dritten Klasse ist die große Mehrheit verstorben. Interessanterweise überlebten in jeder Klasse fast gleich viele, doch der Anteil der Menschen in der dritten Klasse ist weitaus größer als in den anderen zwei Klassen. Die Passagierklasse korreliert signifikant mit dem Überleben, wie ein Chi-Quadrat-Test zeigt ($\chi^2(1) = 102.89, p < 2.2e-16$). Mit einem Cramer's V von 0,34 ist der Effekt mittel. Ein weiterer großer Unterschied besteht zwischen den überlebenden Frauen und Männern, denn unter den Überlebenden gab es ca. doppelt so viele Frauen wie Männer. Der Anteil der Überlebenden bestand aus ca. 68,12 % aus Frauen, und das obwohl ca. 64,75 % der Passagiere an Bord männlich waren. Diesen Zusammenhang verdeutlicht ein Chi-Quadrat-Test, welcher einen hochsignifikanten Zusammenhang zwischen dem Geschlecht und dem Überlebensstatus zeigt ($\chi^2(1) = 260.72, p < .001$). Die Effektstärke nach Cramer's V liegt bei 0,54, was auf einen starken Effekt hindeutet.

3.2 Zusammenhang zwischen Ticketpreisen und Überleben

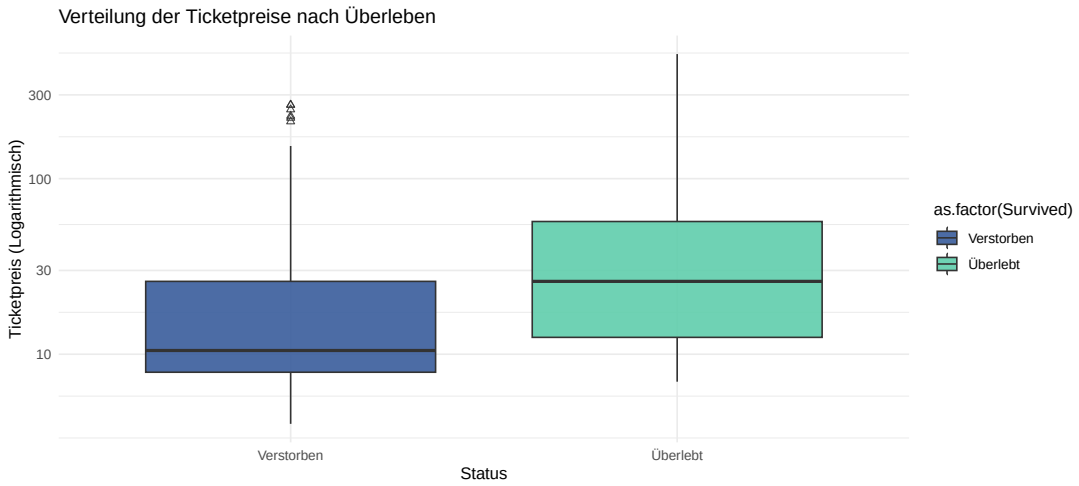

Abbildung 2: Boxplot der Ticketpreise gruppiert nach Überlebensstatus

Allgemein lag der Median der Ticketpreise bei 14.45 Währungseinheiten und das arithmetische Mittel bei 32.20 Währungseinheiten was bereits auf eine Verfälschung des Mittelwerts durch eine Minderheit an extrem teuren Tickets hindeutet. Die Ticketpreise variierten stark zwischen 0 und 512 Währungseinheiten. Es gibt bei den Verstorbenen eine deutliche Tendenz nach unten, der Median orientiert sich an das untere Quartil und beide Quartile liegen unter den Quartilen der Überlebenden. Zudem gibt es bei den Verstorbenen auch viele Ausreißer nach oben, was darauf hindeutet, dass vereinzelt auch Passagiere mit teureren Tickets verstorben sind, allerdings verhältnismäßig wenige. Bei den Überlebenden liegt der mediane Ticketpreis knapp unter 30 Währungseinheiten und somit fast 3 mal teurer als der mediane Ticketpreis der Verstorbenen (ca. 10 Währungseinheiten). Der Median liegt zudem mittig und es gibt keine Ausreißer. Insgesamt zeigt sich eine positive Korrelation ($r_{pb} = 0,257$) zwischen dem Ticketpreis und dem Überlebensstatus.

3.3 Zusammenhang zwischen Familiengröße und Überleben

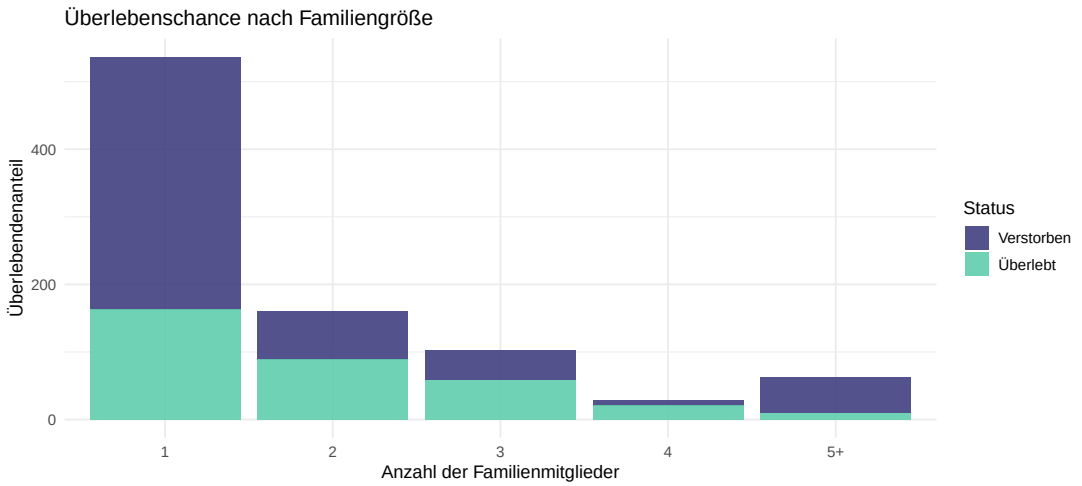

Abbildung 3: Balkendiagramm der Familiengröße gruppiert nach Überlebensstatus

Bei der Mehrheit der Passagiere handelte es sich um Alleinreisende, wobei der Großteil von ihnen nicht überlebte. Auch Familien ab einer Größe von fünf und höher sind zu 85,10 % verstorben, wobei diese eher selten vertreten waren.

3.4 Zusammenhang zwischen Alter und Überleben

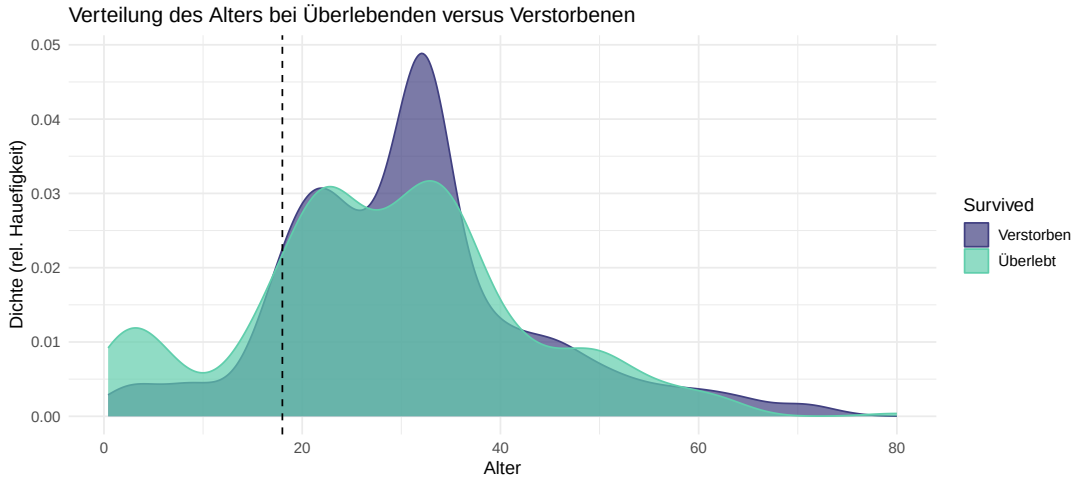

Abbildung 4: Dichteverteilungsfunktion des Alters gruppiert nach Überlebensstatus

Das Durchschnittsalter betrug 29.76 Jahre mit einer Standardabweichung von 13.28. Sowohl unter den Überlebenden, als auch den Verstorbenen liegt das Alter meist zwischen ca. 16-40 Jahren. Auffällig ist das lokale Maximum bei den Überlebenden zwischen etwas 0 und 10 Jahren, welches bei den Verstorbenen nicht vorhanden ist. Dies deutet darauf hin, dass Babys und kleine Kinder deutlich häufiger überlebt haben. Ein zweites lokales

Maximum befindet sich bei den Verstorbenen von etwa 30 Jahren, welches methodisch bedingt ist (Datenimputation der Mittelwerte) und keine natürliche Häufung des Alters repräsentiert. Da die Datenimputation nach dem Mittelwert der jeweiligen Titel-klasse erfolgt ist, ist die Interpretation, dass Kinder (z.B. mit dem Titel Master) häufiger überleben nicht gefährdet. Die punktbiseriale Korrelation ist mit $r_{pb} = -0.089$ zwar schwach, allerdings ist dies auf die nicht-lineare Verteilung zurückzuführen, denn der Überlebensvorteil beschränkt sich primär auf Kinder und hat keinen durchgängigen linearen Trend über alle Altersgruppen hinweg. Ein t-Test bestätigt dennoch die Signifikanz des Altersunterschieds ($p = 0.0075 < .05$).

4 Fazit

Zusammenfassend lässt sich sagen, dass überwiegend Frauen überlebt haben, und das obwohl sie demografisch weniger an Bord vertreten waren. Zusätzlich haben auch Kinder deutlich häufiger überlebt, was auf das Prinzip von "Frauen und Kinder zuerstßchließen lässt. Dies könnte auch eine Erklärung für die Ausreißer in Abbildung 2 sein, da es sich um reiche Männer handeln könnte, die trotz erster Klasse an Bord blieben. Große Familien haben häufig nicht überlebt, vermutlich weil sie nicht alle auf ein Rettungsboot gepasst haben und daher zurückbleiben mussten. Auch Alleinreisende sind zu einem großen Anteil verstorben, eventuell weil ca. 60,33 % von ihnen die dritte Passagierklasse besaß oder weil ca. 76,53 % von ihnen männlich waren. Letztendlich lässt sich sagen, dass vermutlich sowohl die Moral (Frauen und Kinder zuerst), als auch das Geld die Passagiere gerettet hat.