Deskriptive Analyse des Titanic-Datensatzes in R

Furkan Koca, Jonas Ratke, Niclas Rösener

8. Februar 2026

In diesem Projekt wird der Titanic-Datensatz im Rahmen des Moduls „Wissenschaftliches Arbeiten“ deskriptiv analysiert. Zunächst werden der Datensatz und die Variablen für die Analyse aufbereitet. Anschließend werden deskriptive Statistiken und Visualisierungen erstellt, um Zusammenhänge zwischen dem Überleben und einzelner Variablen zu untersuchen. Die Ergebnisse zeigen deutliche Unterschiede in den Überlebenswahrscheinlichkeiten zwischen unterschiedlichen Gruppen.

1 Einleitung und Zielsetzung

Das Ziel der Analyse ist eine nachvollziehbare und reproduzierbare Beschreibung der Überlebensmuster der Passagiere auf der Titanic. Im Mittelpunkt steht, wie das Überleben von Merkmalen wie u.a. dem Geschlecht oder dem Alter zusammenhängt.

2 Daten und Vorverarbeitung

Der Datensatz titanic.csv enthält u. a. Informationen zu Überleben, Klasse, Name, Geschlecht, Alter, Familienbeziehungen, Ticketpreis, Kabine und Einschiffungshafen.

Die Aufbereitung des Datensatzes erfolgte über das R Skript „data_cleaning.R“. Folgende Schritte durchläuft der Datensatz in dem Skript, um die einwandfreie Arbeit an dem Datensatz zu gewährleisten:

- Es extrahiert die Anrede aus Namen
- Unterschiedliche Variablen werden umcodiert:
  - „Pclass“ als ordered-factor
  - „Survived“, „Sex“, „Embarked“ als factor

- Spalten, die für die Analyse nicht benutzt werden, werden gedroppt („PassengerID“,„Name“,„Ticket“,„Cabin“)
- Ableitung der Kabinenmerkmale anhand der Kabinennummern, um das Deck und die Schiffsseite zu ermitteln, wobei unbekannte Einträge als NA gesetzt worden sind

Der Datensatz wird anschließend als „titanic_clean.csv“ gespeichert, um in der Analyse weiterverwendet zu werden.

3 Methoden der Analysen

Für die Analyse verwenden wir zwei Skripte, welche einmal den Datensatz analysieren („functions_1“) und Hilfsfunktionen dazu enthalten („functions_2“), damit die Analysefunktion sichergestellt werden kann.

In „functions_1“ befinden sich Funktionen für deskriptive Statistiken (metrisch / kategorial / bivariat) und die Visualisierung über die Erweiterung Tidyverse.

Die gewählten Ansätze für die univariaten metrischen Variablen sind jeweils ein Histogramm, ein Boxplot sowie eine vektorielle Übersicht. Für die univariaten kategorialen ist der Ansatz der, dass für die unterschiedlichen Variablen jeweils die absolute und relative Häufigkeit berechnet wurde, sowie die Anzahl der fehlenden Werte. Für die bivariaten kategorialen Variablen gibt es zur Berechnung einerseits Korrelationskoeffizienten wie Cramér's V und Chi-Quadrat-Test, sowie Häufigkeiten, wie das Überleben mit den einzelnen Variablen zusammenhängt. Der Ansatz für die bivariate metrisch x dichotome sowie die allgemeinere metrische x kategoriale Analyse ist ein Scatterplot mit unterschiedlichen Farben, gemeinsame Histogramme, ein gemeinsames Boxplot sowie eine vektorielle Übersicht. Für die Visualisierung mehrerer kategorialer Variablen nutzen wir ein Balkendiagramm, welches farblich für die Geschlechter die Überlebensanteile, basierend auf dem Abfahrthafen („Embarked“) und der Passagierklasse („Pclass“) anzeigt.

Die Auswertung erfolgt in einem separaten Skript („analysing.R“), in dem jede der beschriebenen Funktionen genutzt wird.

4 Ergebnisse

4.1 Univariate Ergebnisse

Die Auswertung der metrischen Variablen (R plot i) zeigt zum einen, dass die Gäste durchschnittlich etwa 30 Jahre alt waren, wobei ein Großteil (mehr als die Hälfte) von ihnen zwischen 20 und 40 Jahre alt waren. Während das jüngste Kind nicht einmal ein halbes Jahr alt war, war die älteste Person an Bord 80 Jahre alt. Zum anderen wird sichtbar, dass der durchschnittliche Ticketpreis etwa 32 (einer nicht angegebenen Währung) beträgt. Allerdings fällt auf, dass ein Ticketpreis bei über 500 liegt, während sich ein sehr großer Teil unterhalb von 100 befindet. Entweder ist also ein Ticket sehr teuer

verkauft worden oder es liegt ein Fehler in den Daten vor, der den Durchschnittspreis in die Höhe zieht. Der mediane Ticketpreis liegt nämlich sogar nur bei etwa 14.

Die Auswertung der univariaten kategorialen Variablen zeigen eine ungleiche Verteilung in den Daten des Datensatzes ($n = 891$) bei u.a. dem Geschlecht der Passagiere („Sex“), wo etwa 65% der Passagiere männlich waren. Weiter gibt es eine Menge an fehlenden Daten von ungefähr 77%, die anzeigen, wer welche Kabinen belegt hat, wodurch Aussagen nur eingeschränkt getroffen werden können. Die Daten zeigen weiter, dass die meisten Reisenden in der dritten Klasse („Pclass“) ($\approx 55\%$) gereist sind und in Southampton (S) ($\approx 72\%$) eingeschifft haben, wobei es hier nur 2 fehlende Werte gibt.

4.2 Bivariate Ergebnisse

Die bivariaten Analysen (metrisch $\times$ dichotom) (R plot iv) zeigen keinen erkennbaren Zusammenhang zwischen dem Überleben und dem Alter, da alle Altersklassen sowohl überlebt als auch nicht überlebt haben. Die Grafiken unter Betrachtung des Überlebens sehen den Grafiken des Alters allgemein nämlich sehr ähnlich. Beim Ticketpreis scheint es allerdings eher einen Zusammenhang zum Überleben zu geben. Denn unter den Nicht-Überlebenden sind die Ticketpreise nur ungefähr halb so teuer (Median 10,5) wie unter den Überlebenden (Median 26).

Die bivariaten Analysen (kategorial $\times$ kategorial) zeigen einen starken Zusammenhang zwischen den Überlebenschancen und Geschlecht („Sex“) und Title an mit jeweils einem Wert bei Cramers v bei über 0,5. Dabei gibt es einen schwachen Zusammenhang zwischen dem Abfahrtshafen („Embarked“) und der Überlebenschance mit einem Wert bei Cramers v unter 0,2.

Weitere bivariate Analysen (metrisch $\times$ kategoriell) (R plot vi) zeigen einen starken Zusammenhang zwischen Zustiegshafen und Ticketpreis. So sind die Tickets am günstigsten, wenn man in Queenstown zugestiegen ist und am teuersten bei Zustieg in Cherbourg. Ebenfalls ein starker Zusammenhang ist zwischen Ticketpreis und Passagierklasse erkennbar, da die bessere Klasse jeweils einen teureren Ticketpreis hat. Auch das Alter hat einen Zusammenhang mit der Passagierklasse. Denn Gäste höherer Passagierklasse sind durchschnittlich älter als die niedrigerer Klassen.

4.3 Visualisierung für 4 kategoriale Variablen

Die Visualisierung (Rplot v) zeigt die Überlebensanteile nach Passagierklasse („Pclass“) und Geschlecht unter Berücksichtigung des Abfahrthafens („Embarked“). Die Grafik zeigt, dass Frauen deutlich höhere Überlebenschancen hatten als Männer. Diese Muster sind dabei nicht auf einzelne Passagierklassen oder Abfahrthäfen beschränkt, sondern spiegeln sich in jeder Variable wider.

5 Diskussion

Die Ergebnisse, die wir durch die Analyse des Datensatzes erhalten haben, stimmen mit historischen Erwartungen überein: Die Überlebensrate von Frauen und Kindern,

sowie die der Passagierklassen (1>2>3), erhöht ist. Auch höhere Ticketpreise für bessere Passagierklassen waren zu erwarten. Durch die Verzerrungen durch nicht vorhandene Daten ist es nicht sinnvoll möglich, eine Aussage über die Beeinflussung zwischen dem Überleben und dem Ort der Kabine auf dem Schiff zu treffen. Auf Grund dessen, dass hier nur deskriptive Methoden angewandt wurden, ist es nicht möglich daraus abzuleiten, warum genau es zu den entsprechenden Ergebnissen gekommen ist.

6 Fazit

Die Analyse zeigt deutliche Unterschiede zwischen der Überlebenswahrscheinlichkeit nach Geschlecht, Alter und Passagierklasse. Für Methoden könnten nun modellbasierte Verfahren benutzt werden, um vertiefende und robustere Aussagen treffen zu können.