2. Praktikum Data Science
Korrelation und Wahrscheinlichkeit
WiSe 23/24

Prof. Dr. Marina Tropmann-Frick & Sabrina Göllner

14. November 2023

Wichtig:

1. Die Aufgaben 1 bis 3 sollen auf einem Zettel per Hand ohne Nutzung von Python oder Excel erledigt werden (Klausurrelevant), andernfalls wird die Lösung nicht akzeptiert!

Aufgabe 1: Korrelation zwischen Videoaufrufen und Likes

Ein YouTube-Analyse-Team hat eine Stichprobe von 15 zufällig ausgewählten Gaming-Videos aus den deutschen Trends analysiert. Sie möchten herausfinden, ob es einen Zusammenhang zwischen der Anzahl der Videoaufrufe und der Anzahl der Likes gibt. Die Ergebnisse ihrer Analyse sind in der folgenden Tabelle dargestellt:

Video	**Videoaufrufe (in Tsd.)**	**Likes (in Tsd.)**
1	734	25
2	609	25
3	679	13
4	242	2
5	885	39
6	813	40
7	757	24
8	409	47
9	59	25
10	773	18
11	327	38
12	804	26
13	854	14
14	649	9
15	120	10

Tabelle 1: Daten für zufällig ausgewählte Gaming-Videos

(a) Welche sind die statistischen Einheiten dieser Untersuchung, und welche Merkmale wurden an ihnen gemessen? Geben Sie das Skalenniveau der Merkmale an.

(b) Zeichnen Sie ein Streudiagramm der Daten und interpretieren Sie das resultierende Bild. Welche Art von Beziehung scheint zwischen den beiden Variablen zu bestehen?

(c) Berechnen Sie den Pearson-Korrelationskoeffizienten zwischen den Videoaufrufen und Likes. Was sagt dieser Wert über die Stärke und Richtung der Beziehung zwischen den beiden Variablen aus?

(d) Auf Basis Ihres Ergebnisses: Wenn ein Video eine außergewöhnlich hohe Anzahl an Aufrufen hat, erwarten Sie dann auch eine außergewöhnlich hohe Anzahl an Likes? Begründen Sie Ihre Antwort.

Aufgabe 2: Einfluss der Hintergrundmusik auf die Zuschauerzufriedenheit

YouTube möchte untersuchen, inwieweit die Art der Hintergrundmusik in einem Video die Zufriedenheit der Zuschauer beeinflusst. In einer Pilotstudie wurden 9 Zuschauer gebeten, ein bestimmtes Video anzusehen. Das Video wurde mit drei verschiedenen Hintergrundmusikarten gezeigt: drei Zuschauern mit klassischer Musik, drei mit Popmusik und drei mit Ambient-Klängen. Den Zuschauern wurde nicht mitgeteilt, dass es Unterschiede in der Hintergrundmusik gibt, sondern sie wurden lediglich gebeten, ihre Zufriedenheit mit dem Video anhand einer 5-Punkte-Skala zu bewerten:

5-Punkte-Skala:

1. sehr unzufrieden
2. eher unzufrieden
3. neutral
4. zufrieden
5. sehr zufrieden

Die Bewertungen sind gemeinsam mit der Art der Hintergrundmusik in der folgenden Tabelle wiedergegeben:

Zuschauer-Nr.	Hintergrundmusik	Zufriedenheitsbewertung
1	*Ambient*	1
2	*Ambient*	2
3	*Ambient*	2
4	*Pop*	2
5	*Pop*	2
6	*Pop*	2
7	*Rock*	3
8	*Rock*	1
9	*Rock*	5

Für die Zwecke dieser Analyse ordnen Sie bitte die Musikarten in Bezug auf ihre Intensität wie folgt: Ambient (am wenigsten intensiv) - Rang 1, Pop (mittel) - Rang 2 und Rock (am intensivsten) - Rang 3.

(a) Berechnen Sie die den Rangkorrelationskoeffizienten nach Spearman (einfache Formel) und interpretieren Sie das Resultat.

(b) Welche potenziellen Probleme könnten beim Verwenden der Formel in Bezug auf diese Daten auftreten?

(c) Wie könnten solche Probleme in einer umfangreicheren oder komplexeren Analyse angegangen oder vermieden werden?

Aufgabe 3: Wahrscheinlichkeit

Sie sind der Datenanalyst eines großen Werbeunternehmens, das mit YouTube-Influencern zusammenarbeitet. Ein Unternehmen möchte wissen, wie oft bestimmte Influencer in den YouTube-Empfehlungen auftauchen. Sie haben den Auftrag, dies anhand von 1.000 zufälligen YouTube-Nutzern zu erheben. Hierzu bitten Sie Ihre Mitarbeiter, zufällige Nutzer zu kontaktieren und nach ihren Empfehlungen zu fragen. Ein Mitarbeiter schafft es im Schnitt pro Stunde, 10 Nutzer zu befragen.

(a) Erläutern Sie, welches Wahrscheinlichkeitsmodell die beschriebene Situation am besten abbilden würde. Welche Parameter benötigen Sie für dieses Modell und welchen Wert würden Sie den Parametern hier zuweisen?

(b) Bestimmen Sie unter Rückgriff auf das von Ihnen gewählte Modell die Wahrscheinlichkeit, dass ein Mitarbeiter innerhalb einer Stunde genau 8 Nutzer befragt.

(c) Wie groß ist die Wahrscheinlichkeit, dass ein Mitarbeiter innerhalb einer Stunde höchstens 8 Nutzer befragt?

(d) Wie groß ist die Wahrscheinlichkeit, dass ein Mitarbeiter innerhalb einer Stunde mindestens 8 Nutzer befragt?

Tipps und Anregungen:

- Handelt es sich um einen diskreten oder einen kontinuierlichen Prozess?
- Zählen wir die Anzahl der Ereignisse in einem bestimmten Zeit- oder Raumintervall, oder messen wir die Zeit oder den Raum bis zum ersten Auftreten eines Ereignisses?
- Sind die Ereignisse unabhängig und treten sie mit einer konstanten durchschnittlichen Rate auf?

Aufgabe 4: Korrelationsanalyse der YouTube Daten mit Python

- Suchen sie sich 2 Features (Spalten) aus den Daten aus
- Führen Sie die Korrelationsanalyse mit dem passenden Algorithmus in Python durch
- Plotten Sie das Ergebnis mit einer geeigneten Visualisierung
- Verwenden Sie das Jupyter Notebook "02_Korrelation"