Grundlagen der Statistik – Projektarbeit – Übungsbeispiel WS 2022/23

Betreuung: K. Horneck und J. Hatzl

Die Ausarbeitung erfolgt über R/ Jupyter Notebook bzw. RStudio. Sinnvoll ist die Verwendung des tidyverse package, Ihr könnt aber gerne auch andere packages verwenden.

Erklärungen, Begründungen und Interpretation der Ergebnisse müssen als **Markdown** integriert sein.

Ziel: explorative Bearbeitung eines Datensatzes

- statistische Ausarbeitung eines Datensatzes
- Erklärung, Begründung, Interpretation

Es besteht die Möglichkeit für Online Beratung bei der Lektorin bzw dem Lektor (15 min pro Gruppe).

BEURTEILUNG

- statistische Ausarbeitung
  Methodik (korrekte Wahl der Methode und Begründung), Korrektheit der Ausführung, Vollständigkeit der Ausarbeitung
- Interpretation
  der Ergebnisse, unterstützt durch Diagramme
- Formale Kriterien
  übersichtliche Darstellung, korrekte Sprache, Verwendung von Markdown Elementen

Aufgabenstellung

1. **Datenimport**
    a. Daten einlesen
    b. Überprüfe die Struktur und Variablenbezeichnungen.
    c. Bestimme das Skalenniveau und den Merkmaltyp aller Merkmale
    d. Welche Merkmale liegen als Faktor vor? Warum?
2. **Datenbereinigung**
    a. Gibt es fehlende Werte?
    b. Beschreibe kurz was durch den Aufruf der Funktion `` `na.omit()` `` passiert?
    c. Lösche die Datensätze mit fehlenden Werten.
3. **Häufigkeitstabelle**
    a. Erstelle eine Häufigkeitstabelle (absolut und relativ) für ein ausgesuchtes Merkmal. Interpretiere das Ergebnis.
    b. Ist es sinnvoll für dieses ausgesuchte Merkmal auch eine kumulierte Häufigkeitstabelle zu erstellen?
4. **Häufigkeitstabelle mit Klasseneinteilung**
    a. Erstelle eine Klasseneinteilung für ein ausgesuchtes Merkmal mit gleich breiten Klassen.
    b. Erstelle eine Häufigkeitstabelle (absolut, relativ, kumuliert abs und rel) für das klassierte Merkmal. Interpretiere das Ergebnis.
5. **Daten visualisieren**
    a. Erstelle ein Stabdiagramm für das ausgesuchte Merkmal von Aufgabenstellung 3. Interpretiere das Ergebnis.
    b. Erstelle für dieses Merkmal auch einen Boxplot. Verwende ein zweites Merkmal, um eine sinnvolle Differenzierung der Ausgangsdaten zu zeigen. Interpretiere das Ergebnis.
    c. Erstelle ein Histogramm für das Merkmal aus Aufgabenstellung 4. Wäre es sinnvoll bei diesem Merkmal unterschiedlich breite Klassen zu verwenden? Worauf muss dann geachtet werden?
6. **Kennzahlen**
    a. Die R-Funktion *summary* gibt einen guten Kennzahlenüberblick (Lagemaße) der Variablen. Worin unterscheiden sich die Kennzahlen einer nominalen Variable zu einer metrischen Variable beim Output der Funktion *summary*.
    b. Berechne den arithmetischen Mittelwert und die Standardabweichung für ein ausgesuchtes Merkmal. Wann kann der arithmetische Mittelwert und die Standardabweichung bei einem Merkmal angewendet werden?
    c. Zeigen Sie welche zwei metrischen Variablen den stärksten Zusammenhang aufweisen?
7. **Regressionsanalyse**
    a. Suchen sie zwei sinnvolle metrische Variablen, die für eine einfache lineare Regression geeignet sind. Beurteilen Sie den R-Outputs auf folgende Fragestellungen
        i. Bestimmtheitsmaß (Wert und Interpretation):
        ii. F-Statistik (Wert und Interpretation):
        iii. Standardfehler der Regression (Wert und Interpretation):
        iv. Prüfung des Regressionskoeffizienten (Wert und Interpretation):
    b. Stellen Sie die Regressionsgerade auf und geben Sie an, was die einzelnen Teile inhaltlich bedeuten.
    c. Zeigen Sie an einem Beispiel mit konkreten Werten, wie groß der Fehler bei diesem einen Beispiel ist.