

Übung: Variablen eines Datensatzes erkunden

Am Beispiel der Marktforschungs-Musterstudie

Aufgaben

Projekt aufsetzen und Daten einlesen

1. Erstellt ein neues Projekt für diese Übung und öffnet es in der Entwicklungsumgebung Positron. Stellt sicher, dass ihr eine *virtuelle R-Umgebung* für das Projekt erstellt und aktiviert habt.
2. Installiert die folgenden Pakete in eurer virtuellen Umgebung: `tidyverse`, `janitor`, `skimr`, `summarytools`. Verwendet das Metapaket `pacman`, damit ihr die Pakete mit einem einzigen Befehl installieren und laden könnt!
3. Ladet den Datensatz der Marktforschungs-Musterstudie auf euren Computer herunter und speichert ihn in einem Unterordner `data/` in eurem Projektverzeichnis.
4. Erzeugt einen neuen Ordner `scripts/` in eurem Projektverzeichnis und erstellt eine neue R-Skriptdatei `explore_variables.R` in diesem Ordner.
5. Der Datensatz liegt als CSV-Datei vor. Schaut euch die Datei genauer an und prüft, wie genau das Format aussieht (z. B. Trennzeichen, Dezimalzeichen, Kodierung). Findet einen geeigneten Weg, um die Daten mit R einzulesen. Speichert den Code für das Einlesen in dieser Skriptdatei.

Erster Blick auf den Datensatz

6. Welche Kontextinformationen habt ihr über die Daten? Findet möglichst viele Informationen darüber, wie der Datensatz entstanden ist, welche Variablen enthalten sind und welche Bedeutung sie haben.
7. Versucht euch einen ersten Überblick über den Datensatz zu verschaffen. Welche Variablen gibt es? Wie viele Beobachtungen sind enthalten? Welche Datentypen haben die Variablen?

8. Informiert euch darüber, warum eine Beobachtung immer eindeutig identifizierbar sein muss. Überlegt euch, welche Variable in diesem Datensatz als eindeutiger Identifikator dienen könnte. Wie könnt ihr das Problem sonst lösen?

Was messen die Variablen?

Öffnet den Fragebogen der Marktforschungs-Musterstudie inklusive der Kodierungshinweise als PDF-Datei. Betrachtet die Fragen mit den folgenden Nummern:

1, 2, 3, 4, 5, 8, 11, 12, 13, 16, 17, 21, 22, 23, 26, 29, 216

Bearbeitet für jede dieser Fragen die folgenden Aufgaben:

- a. Findet die zugehörige Variable bzw. die zugehörigen Variablen im Datensatz. Erstellt einen neuen Tibble, der nur (i) die eindeutige ID-Variable (siehe Aufgabe 8) und (ii) die Variable(n) zur jeweiligen Frage enthält. Benennt den Tibble nach dem folgenden Muster: `s_q001` für Frage 1, `s_q002` für Frage 2 usw.
- b. Ermittelt das Skalenniveau der Variable(n).
- c. Prüft den Datentyp, der den Variable(n) im Datensatz zugewiesen ist. Passt der Datentyp zum Skalenniveau? Falls nicht, ändert ihn im neuen Tibble, den ihr für die jeweilige Frage erstellt habt.
- d. Ermittelt den Wertebereich und den Füllgrad der Variable(n).
- e. Ermittelt passende Kenngrößen zur Zusammenfassung der Variable(n) (abhängig von deren Art).
- f. Stellt die Verteilung der Werte der Variable(n) visuell dar. Wählt eine passende Visualisierungsform für den jeweiligen Variablentyp. Versucht, die Visualisierung so aufzubereiten, dass sie möglichst publikationsfähig ist.