Übung: Variablen erkunden

Am Beispiel der Marktforschungs-Musterstudie

Prof. Dr. Nicolas Meseth

In dieser Übung arbeitet ihr mit einem Umfragedatensatz aus einer Marktforschungs-Musterstudie zu Schokolade und Milchalternativen. Der Fokus liegt nicht auf komplexen Modellen, sondern auf den Grundlagen guter Datenanalyse: Welche Variablen gibt es, was messen sie, welche Datentypen liegen vor, welche Wertebereiche und Missing-Codes wurden verwendet, und welche Zusammenfassungen und Visualisierungen passen zu unterschiedlichen Variablentypen?

Schritt 1: 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` und `skimr`. 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. Lest den Datensatz `mds12_schoko_milch.csv` in R ein. Prüft dabei das Dateiformat, zum Beispiel Trennzeichen, Kodierung und Missing-Value-Darstellung. Speichert den Einlesecode in eurer Skriptdatei und legt direkt ein Objekt an, mit dem ihr in den

folgenden Schritten weiterarbeitet.

Schritt 2: Erster Blick auf den Datensatz

6. Verschafft euch einen ersten Überblick über den Datensatz. Wie viele Beobachtungen und wie viele Variablen sind enthalten? Welche Datentypen wurden beim Einlesen zugewiesen? Welche ersten Hinweise geben euch die Variablennamen über den Aufbau des Fragebogens?

7. Überlegt, wie in diesem Datensatz eine Beobachtung eindeutig identifizierbar gemacht werden kann. Gibt es eine bereits vorhandene Variable, die offensichtlich als Primärschlüssel taugt? Wenn nicht, wie würdet ihr das Problem praktisch lösen?

8. Prüft, welche Informationen ihr allein aus dem Datensatz rekonstruieren könnt und welche ihr nur mit einem Fragebogen oder Codebook sicher interpretieren könnt. Schaut euch insbesondere die Variablennamen für die Fragen `1, 2, 3, 4, 5, 8, 11, 12, 21, 22, 23, 26, 29` und `216` an.

Schritt 3: Einfache Einzelvariablen analysieren

In diesem Abschnitt betrachtet ihr zunächst einfache Einzelvariablen. Ziel ist, für jede Variable den Typ, den Wertebereich, Missing Values und sinnvolle Zusammenfassungen sauber zu beschreiben.

9. Analysiert die Variablen zu den Fragen `1, 2, 3, 4` und `5`, also `q001hheinkauf`, `Q002alter`, `q003land`, `Q004geschlecht` und `q005os`. Prüft jeweils:

- Skalenniveau
- Datentyp in R
- Wertebereich und Füllgrad
- passende Kennzahlen
- eine sinnvolle Visualisierung

Legt für jede dieser Fragen einen kleinen Analyse-Tibble an, der mindestens die `respondent_id` und die jeweils betrachtete Variable enthält.

Schritt 4: Mehrfachnennungen und Itembatterien

Viele Fragen liegen nicht als einzelne Variable vor, sondern als Block aus mehreren dichotomen Variablen oder als Itembatterie. Diese Struktur verlangt eine andere Vorgehensweise als bei einer einfachen Einzelvariable.

10. Analysiert Frage `8`, also den Block `v008ort_*`. Beschreibt zuerst, warum es sich hier nicht um eine einzelne kategoriale Variable, sondern um eine Mehrfachantwort-Struktur handelt. Ermittelt anschließend:

- welche Variablen zum Block gehören
- wie hoch der Füllgrad je Item ist
- wie häufig die einzelnen Antwortoptionen gewählt wurden
- eine geeignete Visualisierung

11. Analysiert die Fragen `11` und `12`, also die Blöcke `p011regio_*` und `p012neo_*`. Prüft, welche Codes als echte Antwortkategorien und welche eher als Sonder- oder Missing-Codes zu interpretieren sind. Bereitet die Daten für die Analyse so auf, dass sich die inhaltlichen Bewertungen sinnvoll zusammenfassen und visualisieren lassen.

12. Analysiert die Fragen `21`, `22` und `23`, also die Blöcke `v021pack_*`, `v022kenn_*` und `v023frei_*`. Identifiziert die Struktur dieser Blöcke und vergleicht, wie häufig einzelne Optionen genannt werden. Welche Visualisierungsform eignet sich, um die wichtigsten Nennungen pro Frageblock übersichtlich zu zeigen?

Schritt 5: Rangfolgen, Imagebatterien und Mediennutzung

13. Analysiert Frage `26`, also die Blöcke `p026gericht1_*` und `p026gericht2_*`. Beschreibt zunächst, was die Struktur dieser Variablenform nahelegt. Vergleicht anschließend die Verteilungen ausgewählter Items und erstellt mindestens eine Visualisierung, die zeigt, wie unterschiedlich die Gerichte bewertet oder eingeordnet wurden.

14. Analysiert Frage `29`, also den Block `p029mik_*`. Untersucht, ob es sich um eine Itembatterie handelt, welche Antwortwerte vorkommen und wie stark die einzelnen Aussagen im Mittel ausfallen. Visualisiert das Ergebnis so, dass die wichtigsten Unterschiede zwischen den Aussagen schnell erkennbar sind.

15. Analysiert Frage `216`, also den Block `d216medien_*`. Prüft zunächst den Typ dieser Variablenstruktur und beschreibt anschließend, welche Medienkanäle im Datensatz am häufigsten genannt oder genutzt werden. Wählt eine Visualisierung, die die wichtigsten Kanäle klar vergleichbar macht.

Schritt 6: Reflexion

16. Fasst zusammen, welche grundsätzlichen Variablentypen euch in diesem Datensatz begegnen. Nennt für jeden Typ mindestens ein Beispiel aus der Übung und erläutert kurz, welche Kennzahlen und Visualisierungen jeweils gut dazu passen.

17. Diskutiert, warum Sondercodes wie `-1` und `-2` in Umfragedaten analytisch problematisch sind, wenn man sie unreflektiert wie normale Werte behandelt. Was würdet ihr in einer dokumentierten Analyse tun, um mit solchen Codes sauber umzugehen?

18. Überlegt abschließend, was euch für eine wirklich publikationsfähige Analyse dieser Variablen noch fehlt. Welche zusätzlichen Informationen aus Fragebogen, Feldbericht oder Codebook würdet ihr unbedingt anfordern, bevor ihr inhaltliche Schlussfolgerungen veröffentlicht?