Explore and Transform REWE Products

Practice Exercise

Prof. Dr. Nicolas Meseth

Preparation

Ladet den Datensatz mit den Produkten des REWE-Onlineshops mit den folgenden Zeilen:

```
library(tidyverse)
rewe <- read_csv("data/rewe_products.csv")
```

Aufgabe 1: Datensatz erkunden

1. Wie viele Spalten und Zeilen sind im Datensatz enthalten?
2. Gebt alle Spaltennamen auf der Konsole aus!
3. Welche und wie viele Spalten haben einen numerischen Datentyp?
4. Lasst euch die ersten 20 Produktnamen ausgeben! Wie könnt ihr alle Produkte sehen?
5. Gebt die ersten 30 Hersteller aus! Was fällt euch auf? Wie könntet ihr das lösen?
6. Fasst die Wertebereiche der Spalten `vegan` und `vegetarian` zusammen. Was sagt ihr zu der Datenqualität der beiden Spalten?
7. Erstellt eine Zusammenfassung aller Spalten, die einen Wert in Gramm enthalten. Welche verschiedenen Möglichkeiten findet ihr, das zu erreichen?

Aufgabe 2: Spalten auswählen mit `select`

Findet Lösungen für die folgenden Aufgaben, in denen ihr jeweils eine Teilmenge der Spalten auswählen sollt:

1. Erstellt einen neuen Dataframe, der den Produktnamen, die Produktkategorie und den Verkaufspreis enthält!

2. Erstellt einen neuen Dataframe mit allen Nährwertangaben sowie dem Produktnamen und der Produktkategorie!
3. Erstellt einen neuen Dataframe, der nur numerische Spalten enthält. Prüft die Spalten und schreibt auf, was euch auffällt!

Aufgabe 3: Zeilen einschränken mit `filter`

Findet eine Lösung, um die Zeilen des Datensatzes wie beschrieben einzuschränken:

1. Filtert die Daten, sodass nur Produkte aus Deutschland enthalten sind! Behaltet nur den Produktnamen und das Herkunftsland im Ergebnis!
2. Findet all veganen Bioprodukte!
3. Welche Biersorten werden im REWE-Onlineshop angeboten?
4. Findet alle Rotweine unter 2 EUR!
5. Findet alle Produkte, die das Allergen Soja enthalten! Schaut euch dafür die Funktion `str_detect()` an.

Aufgabe 4: Neue Spalten verändern oder erzeugen mit `mutate` oder `transmute`

1. Ändert den Datentyp der beiden Spalten `productId` und `gtin` in Zeichenketten (`chr`) um.
2. Erstellt eine neue Spalte `sum_nutrition`, in der ihr die Summe aller Nährwertangaben bildet. Belasst im Ergebnis nur die neue Summenspalte und die einzelnen Nährwertspalten.
3. Berechnet, ob ein Produkt mehr als 90% Fett enthält und speichert diese Information auf einer neuen Spalte `high_fat`! Zeigt zur Überprüfung nur Zeilen an, bei denen der Wert `TRUE` ist. Die neue Spalte soll vor der Spalte `productDescription` eingefügt werden.
4. Erstellt eine Spalte `imported_bio`, die `TRUE` enthalten soll, wenn das Produkt ein Bioprodukt ist und gleichzeitig nicht aus Deutschland stammt.

Aufgabe 5: Daten zusammenfassen mit `group_by` und `summarize`

1. Wie viele Produkte befinden sich im Datensatz?
2. Wie viele Produkte hat jede Produktkategorie?
3. Gebt zusätzlich zur Produktkategorie und der Anzahl der Produkte nun auch den durchschnittlichen Verkaufspreis mit an!
4. Listet alle Produktkategorien nach dem durchschnittlichen Fettgehalt ihrer Produkte auf! Auf welches Problem stoßt ihr dabei und wie könnt ihr das lösen?
5. Verwendet das vorherige Ergebnis und behaltet nur die Top 5 der Kategorien mit dem höchsten durchschnittlichen Fettgehalt ihrer Produkte!
6. Welche Marken (`brand`) haben die Produkte mit dem höchsten Proteingehalt im Sortiment? Listet die Top 10!