Datenmanipulation und -auswahl

by Woche 8

In diesem Kapitel befassen wir uns mit einem Aspekt, der in der Praxis oft mehr Zeit und Aufwand erfordert als man denkt. Mit “man” ist hier sowohl man selbst gemeint, als auch die Vorgesetzten bzw. Auftraggeber. Die Rede ist von der Datenmanipulation und -auswahl, auch **Data Wrangling**[1] genannt.

Natürlich ist viel vom Projekt bzw. der Qualität der bereitgestellten Daten abhängig, aber in der Regel ist es so, dass die Daten nicht in der Form vorliegen, in der man sie schließlich auswerten möchte. Demnach ist es notwendig, die Daten zu manipulieren, um sie in die gewünschte Form zu bringen. Dieser Schritt zwischen Import und Analyse ist oft der zeitaufwändigste und wird ggf. als der unangenehmste empfunden. Erfahrungsgemäß lohnt es sich aber sehr, hier Zeit zu investieren, da eine gute Datenbasis die spätere Analyse erheblich erleichtert.

In diesem Kapitel beschäftigen wir uns mit folgenden Aspekten:

- Spalten selektieren/sortieren
- Zeilen filtern/sortieren
- Spalten verändern/hinzufügen
- Daten transponieren

Vorbereitung

Import

Dabei werden wir stets mit einem öffentlich verfügbaren AirBnB Datensatz arbeiten, den wir wie unten gezeigt über die URL importieren können, oder auch hier herunterladen können um ihn statdessen lokal zu importieren.

```
import pandas as pd

csv_url = 'https://github.com/SchmidtPaul/ExampleData/raw/main/airbnb_open_
data/Airbnb_Open_Data.csv'
df = pd.read_csv(csv_url)

df
```

[1] Weitere Begriffe, die in diesem Zusammenhang fallen, sind Datenbereinigung/Data Cleaning, Datenaufbereitung/Data Preprocessing und Data Munging.

```
<string>:2: DtypeWarning: Columns (25) have mixed types. Specify dtype option
on import or set low_memory=False.
             id  ... license
0       1001254  ...     NaN
1       1002102  ...     NaN
2       1002403  ...     NaN
3       1002755  ...     NaN
4       1003689  ...     NaN
...         ...  ...     ...
102594  6092437  ...     NaN
102595  6092990  ...     NaN
102596  6093542  ...     NaN
102597  6094094  ...     NaN
102598  6094647  ...     NaN

[102599 rows x 26 columns]
```

Beim Import dürfte die Warnung `DtypeWarning: Columns (25) have mixed types.Specify dtype option on import or set low_memory=False.` erscheinen. Vereinfacht ausgedrückt bedeutet dies, dass Pandas nicht sicher ist, welchen Datentyp die Spalte 25 haben soll. Tatsächlich versucht die Funktion `pd.read_csv()` ja für jede einzelne Spalte zu “erraten” was für ein Datentyp da eigentlich vorliegt. Bei allen Imports in diesem Kurs bisher hat das auch immer gut funktioniert, aber in diesem Fall gibt es offensichtlich eine problematische Spalte und wir betrachten die Spalte auch gleich genauer. Zur Lösung des Problems werden in der Warnung auch direkt zwei Vorschläge gemacht: Entweder geben wir manuell die Datentypen der Spalte an, oder wir setzen `pd.read_csv(csv_url, low_memory=False)`. Letzteres bedeutet vereinfacht ausgedrückt, dass Pandas mehr Arbeitsspeicher für den Import verwendet, um die Datentypen besser zu erraten. Wir entscheiden uns hier für die erste Variante und geben den Datentyp an - es ist nämlich eine Spalte, die wir als Text/String (`str`) behandeln wollen, sodass wir wie folgt importieren können ohne, dass die Warnung erscheint:

```
df = pd.read_csv(csv_url, dtype={25: str})
```

Ausgabeeinstellungen

Es fällt auf, dass (zum Glück) nur ein Teil der Tabelle gezeigt wird. Unten steht dann `[102599 rows x 26 columns]`, es werden aber nur die ersten und letzten Zeilen und Spalten gezeigt und dazwischen steht `...`. Das ist eine Voreinstellung beim Arbeiten mit Pandas und wie viele Zeilen/Spalten maximal angezeigt werden sollen, können wir manuell einstellen. So können wir mit `pd.set_option()` z.B. bestimmen, dass wir immer nur maximal 4 Spalten und 6 Zeilen angezeigt bekommen wollen und, dass jede Spalte maximal 24 Zeichen breit sein soll (ansonsten wird der Inhalt abgeschnitten).

```
pd.set_option('display.max_columns', 4)
pd.set_option('display.max_rows', 6)
pd.set_option('display.max_colwidth', 24)

df
```

```
             id                     NAME  ...              house_rules license
0       1001254  Clean & quiet apt ho...  ...  Clean up and treat t...     NaN
1       1002102    Skylit Midtown Castle  ...  Pet friendly but ple...     NaN
2       1002403  THE VILLAGE OF HARLE...  ...  I encourage you to u...     NaN
...         ...                      ...  ...                     ...     ...
102596  6093542  Comfy, bright room i...  ...                     NaN     NaN
102597  6094094  Big Studio-One Stop ...  ...                     NaN     NaN
102598  6094647     585 sf Luxury Studio  ...                     NaN     NaN

[102599 rows x 26 columns]
```

Dies sind demnach Einstellungen, die vor allem für euch und euren Arbeitsfluss wichtig sind, da die eigentlichen Daten und Ergebnisse ja nicht verändert werden und - gegeben, dass ihr die Analyse alleine durchführt - niemand außer euch diese ausgegebenen Zwischenergebnisse sehen wird. Stelle euch diese (und ggf. weitere) dieser Pandas Optionen also so ein wie ihr es mögt.

In den folgenden Unterkapiteln gehen wir mehr oder weniger separat auf die einzelnen Aspekte der Datenmanipulation ein.