Textbasierte Daten

Bisher haben wir vor allem zahlenbasierte Daten betrachtet, die man beispielsweise in einem Diagramm darstellen und auch empirisch auswerten kann. Für die meisten Anwendungen reicht das vollkommen aus, aber im Internet und auch im Alltag begegnen uns vor allem textbasierte Daten, das heißt Wörter, Sätze, Dokumente und ganze Sammlungen an Texten. In dieser Einheit wagen wir einen Exkurs hin zu dieser Art von Daten, mit denen wir ganz andere Untersuchungen anstellen können als mit Zahlen allein. Zum Beispiel können wir untersuchen:

- Welche Wörter in einer Sorte Text am häufigsten vorkommen und sie mit anderen Texten vergleichen
- Wie die Stimmung (das *Sentiment*) dieser Texte ist oder
- Ob es bei Immobilien einen Zusammenhang zwischen Wortwahl und Kaufpreis gibt.

Um diesen Zielen näher zu kommen, benutzen wir die Datei angebote_1000.csv. Sie enthält Daten zu über 200.000 Immobilienangeboten aus den Jahren 2018 und 2019. Diese Daten wurden entnommen aus diesem Kaggle-Datensatz und stammen von der Seite immobilienscout24.de.