**avocado\_prices**

**Hintergrund:** Daten zu Preisen und Nachfrage nach Avocados in mehreren US-Städten/Regionen. (Inspiration: <https://www.theguardian.com/lifeandstyle/2017/may/15/australian-millionaire-millennials-avocado-toast-house> In which cities can millenials have their avocado toast AND buy a home?)

**Variablen:** Durchschnittlicher Preis, verkaufte Menge, Zeit, Region, …

**Anwendbarkeit:** Geeignet um eine Regression zu machen (z.B. den Preis vorhersagen) oder durchschnittliche Preise in unterschiedlichen Regionen zu vergleichen und das Data Wrangling aufzuzeigen.

**Quelle:** <https://www.kaggle.com/neuromusic/avocado-prices>

**london\_bike\_shares**

**Hintergrund:** Daten zur Nachfrage nach Leihvelos in London.

**Variablen:** Anzahl neu ausgeliehener Velos, Zeit, Temperatur, Wetter, Windgeschwindigkeit, …

**Anwendbarkeit:** Geeignet um eine einfache Regressionen zu machen. Der Datensatz ist schon ziemlich gut vorbereitet. Man müsste ihn daher wahrscheinlich anpassen, damit die Studierenden beim Data Wrangling noch etwas zu tun haben.

**Quelle:** <https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset>

**grades**

**Hintergrund:** Daten zu schulischen Leistungen von portugiesischen Schülern in Mathematik und Portugiesisch (2 separate Datensätze).

**Variablen:** Noten, demographische variablen, Absenzen, Alkohol-Konsum, …

**Anwendbarkeit:** Geeignet um eine Regression auf die Noten zu machen. Auch geeignet für Data Wrangling (Zusammenführen der Datensätze, Encoding,…).

**Quelle:** <https://archive.ics.uci.edu/ml/datasets/student+performance>

Erklärung der Variablen im txt-file

**diabetes**

**Hintergrund:** Daten zu mehreren medizinischen Prädiktor-Variablen und dem Target Feature (Diabetes; ja/nein) für Pima-Indianerinnen.

**Variablen:** Diabetes, Anzahl Schwangerschaften, Glukose Konzentration, Blutdruck, …

**Anwendbarkeit:** Geeignet um eine logistische Regression zu machen. Der Datensatz ist schon ziemlich gut vorbereitet. Man müsste ihn daher wahrscheinlich anpassen, damit die Studierenden beim Data Wrangling noch etwas zu tun haben.

**Quelle:** <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

**obesity**

**Hintegrund:** Daten zu Übergewichtslevel, Essgewohnheiten und anderen gesundheitlichen Faktoren von Personen aus Mexiko, Peru und Kolumbien.

**Variablen:** Übergewichtslevel, Geschlecht, Altern, Grösse, Gewicht, Raucher, …

**Anwendbarkeit:** Geeignet um logistische Regression zu machen. Auch fürs Data Wrangling geeignet (cleaning, encoding der kategoriellen variablen,…).

**Quelle:**

<https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>+

<https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub> (Erklärung der Variablen).

**forest\_fires**

**Hintergrund:** Daten zu Waldbränden in Algerien (2 verschieden Regionen)

**Variablen:** Tag, Monat, Jahr, Temperatur, Niederschlag, Feuer (Ja/Nein),…

**Anwendbarkeit:** Geeignet um logistische Regression zu machen. Der Datensatz ist schon ziemlich gut vorbereitet. Man müsste ihn daher wahrscheinlich anpassen, damit die Studierenden beim Data Wrangling noch etwas zu tun haben.

**Quelle:** <https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset>++

**used\_cars**

**Hintergrund:** Daten zu Occasion-Autos aus mehreren US-Bundesstaaten

**Variablen:** Preis, Mileage, Marke, Farbe, Bundesstaat, …

**Anwendbarkeit:** Geeignet um eine Regression auf den Preis zu machen oder durchschnittliche Preise zwischen Bundesstaaten zu vergleichen. Auch für Data Wrangling geeignet 🡪 Cleaning, Miles in KM umwandeln, Durchschnittspreise pro Bundesstaat bilden,…

**Quelle:** <https://www.kaggle.com/doaaalsenani/usa-cers-dataset>