

Datamics präsentiert: Frei zugängliche Datensammlungen und Online-Ressourcen

© Datamics

Inhalt

STARTPUNKTE	4
Google Public Data.....	4
Quora Public Data.....	4
FlowingData: Resources to Find the Data You Need, 2016 Edition	4
Duke University Statistics Lab.....	4
DATENSAMMLUNGEN	5
Amerikanische Regierungsquellen:	5
National Centers for Environmental Information (NCEI)	5
Bildungsdatenbanken.....	5
UC Irvine Machine Learning Repository	5
Datenjournalismus:	5
FiveThirtyEight	5
FlowingData	5
App Built-ins:	6
Plotly Datasets	6
The R Datasets Package	6
Seaborn Datasets	6
SPEZIFISCHE DATENSÄTZE.....	7
Chicago Data Portal - Motor Vehicle Theft.....	7
CO2 PPM - Trends in Atmospheric Carbon Dioxide.....	7
Federal Housing Finance Agency - FHFA House Price Indexes (HPIs)	7
The Reddit Button.....	7
DATENSÄTZE, DIE IN DIESEM KURS VERWENDET WURDEN.....	8
U.S. Census Bureau	8
2018 Winter Olympics in PyeongChang, South Korea.....	8
mpg.csv	8
Mark Twain and the Quintus Curtius Snodgrass Letters	8
Fremont Bridge Bicycle Traffic.....	8
National Oceanographic and Atmospheric Administration (NOAA), U.S. Climate Reference Network (USCRN).....	9

Iris Dataset	9
Abalone Dataset	9
Old Faithful Geyser Eruptions.....	9
Seaborn Flights Data	10
Dash gapminderDataFiveYear	10
NASDAQ Companies and Stock Symbols	11
Arrhythmia.....	11
WEITERE OPTIONEN.....	12
Dash - Chris Parmer's indicators.csv.....	12

Dies ist eine Liste der online verfügbaren Datensätze und Datensammlungen. Die Ressourcen in dieser Liste sind öffentlich und erfordern keine Mitgliedschaften oder Abonnements, um Zugriff zu erhalten. Aus diesem Grund wurden Kaggle-Datensätze nicht in die Liste aufgenommen.

- Der erste Abschnitt listet die allgemeinen Datensatz Quellen auf
- Der zweite Abschnitt enthält Organisationswebsites und Datensammlungen.
- Der dritte Abschnitt enthält spezifische Datensätze, die du für verschiedene Zwecke herunterladen kannst.
- Der vierte Abschnitt enthält Datensätze, die speziell in diesem Kurs verwendet werden.
- Im letzten Abschnitt sind noch weitere Optionen aufgeführt



STARTPUNKTE

Google Public Data

<https://www.google.com/publicdata/directory>

Quora Public Data

<https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public>

FlowingData: Resources to Find the Data You Need, 2016 Edition

<https://flowingdata.com/2016/11/10/find-the-data-you-need-2016-edition/>

Es kann frustrierend sein, wenn du speziell benötigte Daten suchst, daher gibt es hier einige Tipps, wie du sie findest (zumindest die frei verfügbaren) und einige themenspezifische Quellen für den Beginn der Suche.

Duke University Statistics Lab

<http://stat.duke.edu/resources/datasets> und <http://www2.stat.duke.edu/courses/Fall03/sta290/datasets.html>



DATENSAMMLUNGEN

Amerikanische Regierungsquellen:

National Centers for Environmental Information (NCEI)

<https://www.ncei.noaa.gov/access>

<https://www.ncdc.noaa.gov/data-access>

Ursprünglich das “National Climatic Data Center” (NCDC), heute “NOAA's National Centers for Environmental Information” (NCEI), welches die Daten hostet und öffentlichen Zugang zu einem der bedeutendsten Archive für Umweltdaten weltweit bietet. Über das „Center for Weather and Climate“ und das „Center for Coasts, Oceans, and Geophysics“ liefert es über 25 Petabyte umfassende atmosphärische, küsten-, ozeanische und geophysikalische Daten.

Bildungsdatenbanken

UC Irvine Machine Learning Repository

<http://mlr.cs.umass.edu/ml/>

Wir führen derzeit 22 Datensätze als Service für die Machine Learning Community

Datenjournalismus:

FiveThirtyEight

<http://fivethirtyeight.com/>

Nachrichten-Outlet von Nate Silver, Datensätze verfügbar auf Github

FlowingData

<http://flowingdata.com/category/statistics/data-sources/>

Viel Spaß beim Rumspielen!



App Built-ins:

Plotly Datasets

<https://github.com/plotly/datasets>

The R Datasets Package

<http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>

Seaborn Datasets

<https://github.com/mwaskom/seaborn-data>

Datenarchiv für Seaborn-Beispiele (Grafik-Bibliothek). *Dies ist kein allgemeines Datenarchiv.*

Dieses Archiv dient nur dazu, schnell und einfach Beispieldatensätzen für die `seaborn.load_dataset` Funktion zum Herunterladen bereitzustellen. Es dient dazu, Seaborn einfach zu dokumentieren, ohne Zeit darauf zu verschwenden, Daten zu laden oder gar zusammenzusuchen. Die Datensätze können sich jedoch jederzeit ändern oder entfernt werden, wenn sie für die Seaborn-Dokumentation nicht mehr nützlich sind.

SPEZIFISCHE DATENSÄTZE

Chicago Data Portal - Motor Vehicle Theft

<https://data.cityofchicago.org/Public-Safety/motor-vehicle-theft/7ac4-d9tk>

Dieser Datensatz spiegelt die gemeldeten Fälle von Kfz-Diebstählen wider, die in der Stadt Chicago von 2001 bis heute stattfanden, abzüglich der letzten sieben Tage. Die Daten werden aus dem CLEAR-System (Citizen Law Enforcement Analysis and Reporting) des Chicago Police Departments extrahiert. Um die Privatsphäre der Opfer von Straftaten zu schützen, werden die Adressen nur auf Blockebene angezeigt, es werden daher keine spezifischen Orte bereitgestellt.

9 Felder, 307644 Datensätze (und Zähler)

CO2 PPM - Trends in Atmospheric Carbon Dioxide

<http://datahub.io/core/co2-ppm>

Die Daten stammen vom „Earth System Research Laboratory“ der US-Regierung, Abteilung Globales Monitoring. Es werden zwei Hauptinformationen angeboten: die Mauna Loa-Serie (die längste durchgehende Serie seit 1958) und eine Global Average-Serie (ein globaler Durchschnitt mariner Oberflächengewässer).

Federal Housing Finance Agency - FHFA House Price Indexes (HPIs)

<https://catalog.data.gov/dataset/fhfa-house-price-indexes-hpis>

Der FHFA House Price Index (HPI) ist ein breites Messinstrument für Preisentwicklung von Einfamilienhäusern. Der HPI ist ein gewichteter Wiederverkaufs-Index, d.h. er misst die durchschnittliche Preisänderungen bei Wiederverkäufen oder Refinanzierungen derselben Immobilien. Diese Informationen erhält er durch eine Überprüfung der wiederkehrenden Hypothekengeschäfte von Einfamilienhäusern, deren Hypotheken seit Januar 1975 von Fannie Mae oder Freddie Mac erworben oder verbrieft wurden.

The Reddit Button

Ein großer (wenn auch etwas unseriöser) Satz von Zeitreihendaten, der 2015 aus einem sozialen Experiment auf Reddit abgeleitet wurde. Am 1. April postete Reddit einen einfachen Button mit einem 60-Sekunden-Timer, der auf null heruntergezählt wurde. Bei jedem Drücken der Taste durch einen beliebigen Reddit-Benutzer wird der Timer auf 60 Sekunden zurückgesetzt. Nach mehr als zwei Monaten und einer Million Klicks schaffte es der Timer schließlich, ohne einen weiteren Klick auf Null Sekunden runter zu laufen.

Diskussion: <https://redditblog.com/2015/06/08/the-button-has-ended/>

Daten: https://github.com/reddit/thebutton-data/blob/master/thebutton_presses.csv

4 Felder, 1008316 Datensätze, 44MB

DATENSÄTZE, DIE IN DIESEM KURS VERWENDET WURDEN

U.S. Census Bureau

<https://www.census.gov/data/datasets/2017/demo/popest/nation-total.html#ds>

<https://www2.census.gov/programs-surveys/popest/datasets/2010-2017/national/totals/nst-est2017-alldata.csv>

Komplette Datensätze der Bundesstaaten und des Commonwealth von Puerto Rico zu: Bevölkerung, Bevölkerungsschwankungen und geschätzter Zusammensetzung der Schwankungen: 1. April 2010 bis 1. Juli 2017
121 Felder, 57 Datensätze

2018 Winter Olympics in PyeongChang, South Korea

<http://time.com/5143796/winter-olympic-medals-by-country-2018/>

<https://www.pyeongchang2018.com/en/game-time/results/OWG2018/en/general/medal-standings.htm>

Medaillenwertung nach Land bei den Olympischen Winterspielen 2018

6 Felder, 30 Datensätze

mpg.csv

<https://gist.github.com/omarish/5687264>

Meilen pro Gallone und andere Statistiken für Fahrzeuge, die von 1970-1982 produziert wurden.

9 Felder, 399 Datensätze

Mark Twain and the Quintus Curtius Snodgrass Letters

Angepasst aus <https://www.math.utah.edu/~treiberg/M3074TwainEg.pdf> unter Zitierung der Daten von:

Brinegar, C. S. (1963) Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship.


Journal of the American Statistical Association. 58, 85-96

2 Felder, 18 Datensätze

Fremont Bridge Bicycle Traffic

<https://data.seattle.gov/Transportation/Grouped-by-Hour/7mre-hcut>

Der „Fremont Bridge Bicycle Counter“ erfasst die Anzahl der Fahrräder, die die Brücke auf den Fußgänger-/Radwegen überqueren. Induktive Messschleifen an den östlichen und westlichen Enden zählen die durchfahrenden Fahrräder unabhängig von der Fahrtrichtung. Die Daten bestehen aus einem Datums- und Zeitfeld: Datum, Zähler der Überquerung der östlichen Messschleife (Fremont Bridge NB) und Zähler der Überquerung der westlichen Messstelle (Fremont Bridge SB). Die Felder Durchfahrtszähler stellen die Gesamtzahl der in der angegebenen Stunde festgestellten Fahrräder dar. Die Fahrtrichtung ist nicht angegeben, aber im Allgemeinen ist der größte Verkehr im Feld „Fremont Bridge NB“ in Richtung Norden und der größte Verkehr im



Feld „Fremont Bridge SB“ in Richtung Süden.
3 Felder, 47400 Datensätze (und Zähler)

National Oceanographic and Atmospheric Administration (NOAA),
U.S. Climate Reference Network (USCRN)

Seite: <https://www.ncdc.noaa.gov/crn/qcdatasets.html>

Datensätze: <https://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02/2010/>

Das USCRN ist ein systematisches und nachhaltiges Netzwerk von Klimaüberwachungsstationen mit Standorten in den USA, Alaska und Hawaii. Diese Stationen verwenden hochwertige Instrumente zur Messung von Temperatur, Niederschlag, Windgeschwindigkeit, Bodenbeschaffenheit und mehr.

Iris Dataset

<https://archive.ics.uci.edu/ml/machine-learning-databases/iris/>

Der [Iris flower data set](#) (Iris Flower-Datensatz) oder Fisher's Iris-Datensatz ist ein multivariater Datensatz, der 1936 von Sir Ronald Fisher als Beispiel für eine Diskriminanzanalyse eingeführt wurde. Der Datensatz besteht aus 50 Proben von jeweils drei Irisarten (*Iris setosa*, *Iris virginica* und *Iris versicolor*) also insgesamt 150 Proben. An jeder Probe wurden vier Merkmale gemessen: Länge und Breite der Kelchblätter, sowie Blütenblätter in cm.
5 Felder, 150 Datensätze

Abalone Dataset

<http://mlr.cs.umass.edu/ml/datasets/Abalone>

Das Abalone-Dataset wird beim Machine Learning verwendet und bietet 8 Funktionen, mit denen das Alter von Abalonen (Muschelart) vorhergesagt werden kann. Das Alter wird bestimmt, indem die Schale am Kegel angeschnitten, eingefärbt und die Anzahl der Ringe durch ein Mikroskop gezählt werden - eine langweilige und zeitraubende Aufgabe. Um das Alter vorherzusagen, können andere Messwerte, die leichter zu erhalten sind, verwendet werden.


Wir haben die folgenden Spaltennamen hinzugefügt:

‘sex’, ‘length’, ‘diameter’, ‘height’, ‘whole_weight’, ‘shucked_weight’, ‘viscera_weight’, ‘shell_weight’, ‘rings’
(‘Geschlecht’, ‘Länge’, ‘Durchmesser’, ‘Höhe’, ‘Gesamtgewicht’, ‘Gewicht ohne Schale’, ‘Gewicht der Innereien’, ‘Schalengewicht’, ‘Jahresringe’)

9 Felder, 4177 Datensätze

Old Faithful Geyser Eruptions

Old Faithful ist ein Kegel-Geysir. Seit 2000 variieren die Intervalle zwischen 44 und 125 Minuten, mit einem Durchschnitt von etwa 90 bis 92 Minuten. Die Dauer beträgt 1 1/2 bis 5 Minuten und die Höhe beträgt 90 bis 184 Fuß. Es ist nicht möglich, im Voraus mehr als einen Ausbruch vorherzusagen. Old Faithful ist derzeit bimodal. Er hat zwei Ausbruchslängen, entweder eine lange (über 4 Minuten) oder etwas seltener eine kurze Dauer (etwa 2 1/2 Minuten). Kurze Ausbrüche führen zu einem Intervall von ca. einer Stunde und lange Ausbrüche zu einem



Intervall von etwa 1 1/2 Stunden.

Rohdaten (elektronische Aufzeichnungen von Eruptionen und tabellierte Beobachtungsprotokolle der Eruptionsdauer) findest du unter:

<http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFaithful> und
<http://www.geyserstudy.org/ofvclogs.aspx>

Der für diese Übung verwendete Datensatz stammt aus dem Statistiklabor der Duke University unter:
<http://www2.stat.duke.edu/courses/Fall03/sta290/datasets.html>

Er besteht aus 3 Feldern (D, Y, X) wobei

D = Aufnahmedaten pro Monat (im August),

X = Dauer des aktuellen Ausbruchs in Minuten (bis 0,1 Minuten)

Y = Zeit bis zum nächsten Ausbruch in Minuten (zur nächsten Minute).

Die ersten 107 Datensätze wurden vom 1. bis 8. August 1978 und die weiteren 115 Datensätze ein Jahr später, vom 16. bis 23. August 1979, aufgezeichnet.

Seaborn Flights Data

Zu den Optionen gehören das Herunterladen des Datensets von <https://github.com/mwaskom/seaborn-data> oder das direkte Importieren aus dem Seaborn-Modul mit

```
import seaborn as sns
df = sns.load_dataset("flights")
```

3 Felder, 144 Datensätze

Dash gapminderDataFiveYear

Seite: <https://github.com/plotly/datasets>

Datensätze: <https://raw.githubusercontent.com/plotly/datasets/master/gapminderDataFiveYear.csv>

Dieser Datensatz wird in Online-Tutorials von Dash verwendet.

Er besteht aus 6 Feldern:

Land

Jahr

Pop Population (Bevölkerung)

Kontinent [Asien, Europa, Afrika, Amerika, Ozeanien]

lifeExp Lebenserwartung

gdpPercap Bruttoinlandsprodukt pro Kopf

und enthält 1704 Datensätze



NASDAQ Companies and Stock Symbols

Datensatz:

<http://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ&render=download>

Auswahl der an der NASDAQ-Börse notierten Unternehmen. 3299 Originaleinträge wurden händisch auf 256 reduziert, basierend auf der jeweiligen Marktkapitalisierung der Unternehmen.

Symbol, Name, LastSale, MarketCap, IPOyear, Sector, Industry (Stand: 06.04.2014)

7 Felder, 256 Datensätze

Arrhythmia

Quelle:

<https://archive.ics.uci.edu/ml/datasets/arrhythmia>

420 Datensätze wurden aus dem Arrhythmie-Datensatz entfernt. Für unser Beispiel-Histogramm haben wir nur Spalten für Alter, Geschlecht (0 = männlich, 1 = weiblich) und Größe (in Zentimeter) verwendet. Wir löschen jeden unter 20 und jeden der weniger als 30 kg wiegt.

3 Felder, 420 Datensätze



WEITERE OPTIONEN

Dash - Chris Parmer's indicators.csv

Datensatz:

<https://gist.githubusercontent.com/chridyp/cb5392c35661370d95f300086accea51/raw/8e0768211f6b747c0db42a9ce9a0937dafcbd8b2/indicators.csv>

5 Felder [(unbenannter Index), Ländername, Indikatorname, Jahr, Wert], 85.536 Datensätze, 36.616 Werte