

Package datadepot

Description The **datadepot** library provides a collection of datasets used in the book Data Science Foundations and Machine Learning with Python.

URL <https://github.com/vanraak/datadepot>

Depends Python (≥ 3.8) and Pandas (>2.0)

License GPL (≥ 2)

Repository Pypi

Author Jeroen van Raak

Installation

```
pip install datadepot
```

or

```
conda install conda-forge::datadepot
```

Usage

```
import datadepot
df=datadepot.load("<dataset>")
```

Replace with the name of the dataset, such as “bank”, “house”, or “churn”.

Example

```
df=datadepot.load("bank") # Load the bank dataset.
```

Datasets

adult	3
bank	5
bicycle_counts	7
bike_sharing	8
churn	10
cpu	12
credit	13
credit_card	15
diamonds	16
drug	18
house	19
house_price	20
loan	21
mpg	22
nyc_taxi	23
online_shoppers	24
red_wines	26

adult Adult dataset

Description

The Adult dataset was collected by the US Census Bureau. The primary task is to predict whether a given adult makes more than \$50K per year based on attributes such as education, hours worked per week, and other demographic features. The target variable is income, a factor with levels “<=50K” and “>50K”, and the remaining 15 variables are predictors.

Usage

```
df=datadepot.load(“adult”)
```

Format

The adult dataset contains 48,598 rows and 16 columns.

The 16 variables are:

- **age**: age in years.
- **workclass**: a factor with 6 levels.
- **demogweight**: the demographics to describe a person.
- **education**: a factor indicating the highest level of education attained (16 levels).
- **education_num**: ordinal encoding of education level.
- **marital_status**: a factor with 5 levels.
- **occupation**: a factor with 15 levels.
- **relationship**: a factor with 6 levels.
- **race**: a factor with 5 levels.
- **gender**: a factor with levels “Female”, “Male”.
- **capital_gain**: capital gains.
- **capital_loss**: capital losses.
- **hours_per_week**: number of hours of work per week.
- **native_country**: country of origin.
- **region**: geographic region derived from **native_country**.
- **income**: yearly income as a factor with levels “<=50K” and “>50K”.

Source

This dataset is publicly available from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/2/adult>

Creators

Barry Becker and Ronny Kohavi

License

[Creative Commons Attribution 4.0 International](#)

References

Kohavi, R. (1996, August). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In Kdd (Vol. 96, pp. 202-207).

Remarks

We renamed several columns, corrected typos in the `native_country` column, and added a `region` feature derived from `native_country`.

bank Bank Marketing dataset

Description This dataset contains data from direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable deposit).

Usage

```
df=datadepot.load("bank")
```

Format

The bank dataset contains 4,521 rows (customers) and 17 columns. The 17 variables are explained below.

Bank client data:

- **age**: numeric.
- **job**: type of job; categorical: “admin.”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar,”self-employed”, “retired”, “technician”, “services”.
- **marital**: marital status; categorical: “married”, “divorced”, “single”; note: “divorced” means divorced or widowed.
- **education**: categorical: “secondary”, “primary”, “tertiary”, “unknown”.
- **default**: has credit in default?; binary: “yes”, “no”.
- **balance**: average yearly balance, in euros; numeric.
- **housing**: has housing loan? binary: “yes”, “no”.
- **loan**: has personal loan? binary: “yes”, “no”.

Related with the last contact of the current campaign:

- **contact**: contact: contact communication type; categorical: “unknown”, “telephone”, “cellular”.
- **day**: last contact day of the month; numeric.
- **month**: last contact month of year; categorical: “jan”, “feb”, “mar”, ..., “nov”, “dec”.
- **duration**: last contact duration, in seconds; numeric.

Other attributes:

- **campaign**: number of contacts performed during this campaign and for this client; numeric, includes last contact.

- **pdays:** number of days that passed by after the client was last contacted from a previous campaign; numeric, -1 means client was not previously contacted.
- **previous:** number of contacts performed before this campaign and for this client; numeric.
- **poutcome:** outcome of the previous marketing campaign; categorical: “success”, “failure”, “unknown”, “other”.

Target variable:

- **deposit:** Indicator of whether the client subscribed a term deposit; binary: “yes” or “no”.

Source

This dataset is publicly available from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

Creators

S. Moro, P. Rita and P. Cortez

License

[Creative Commons Attribution 4.0 International](#)

References

Moro, S., Laureano, R. and Cortez, P. (2011) Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference.

bicycle_counts Bicycle counts dataset.

Description The dataset summarizes daily bicycle crossing counts across four New York City bridges (Brooklyn, Manhattan, Williamsburg, and Queensboro), combined with corresponding weather conditions including temperature and precipitation. The data covers the period from April 2017 to October 2017 and contains 183 daily observations with 10 variables.

Usage

```
df=datadepot.load("bicycle_counts")
```

Format

The dataset contains 183 observations (days) and 10 columns. The variables are explained below.

- **date**: Calendar date of the observation.
- **high_temp**: Daily maximum temperature recorded (°F).
- **low_temp**: Daily minimum temperature recorded (°F).
- **precipitation**: Total daily precipitation amount.
- **precipitation_trace**: Indicator of whether a trace (measurable but below the minimum reportable amount) of precipitation was observed (**True**) or not (**False**).
- **brooklyn_bridge**: Daily bicycle crossing count for the Brooklyn Bridge.
- **manhattan_bridge**: Daily bicycle crossing count for the Manhattan Bridge.
- **williamsburg_bridge**: Daily bicycle crossing count for the Williamsburg Bridge.
- **queensboro_bridge**: Daily bicycle crossing count for the Queensboro Bridge.
- **total**: Total daily bicycle crossing count across all bridges.

Source

Department of Transportation (DOT), NYC Open Data: https://data.cityofnewyork.us/Transportation/Bicycle-Counts-for-East-River-Bridges-Historical-/gua4-p9wg/about_data

Creators

NYC Department of Transportation (DOT)

License

[New York City Open Data Terms of Use](#)

bike_sharing Bike Sharing dataset

Description The Seoul Bike Sharing Demand dataset contains hourly records of bicycle rentals in Seoul, South Korea, together with weather and environmental variables such as temperature, humidity, wind speed, visibility, rainfall, snowfall, and solar radiation. The dataset is commonly used to study the relationship between weather conditions and bike rental demand and to develop predictive models for forecasting bicycle usage.

Usage

```
df=datadepot.load("bike_sharing")
```

Format

The bank dataset contains 8,760 rows (customers) and 15 columns. The 14 variables are explained below.

- **date**: Date of the observation.
- **bike_count**: Number of bicycles rented during the hour.
- **hour**: Hour of the day (0–23).
- **temperature**: Air temperature (°C).
- **humidity**: Relative humidity (%).
- **wind_speed**: Wind speed (m/s).
- **visibility**: Visibility distance (10 m units).
- **dew_point_temperature**: Dew point temperature (°C).
- **solar_radiation**: Solar radiation (MJ/m²).
- **rainfall**: Rainfall amount (mm).
- **snowfall**: Snowfall amount (cm).
- **seasons**: Season of the year.
- **holiday**: Indicates whether the day is a holiday.
- **functioning_day**: Indicates whether bike rental services were operating.

Source

This dataset is publicly available from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand>

The dataset originates from Seoul Open Data Plaza: <https://data.seoul.go.kr/>

Creators

Unknown

License

[Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)

References

Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, 153, 353-366. <https://doi.org/10.1016/j.comcom.2020.02.007>

churn Churn dataset for Credit Card Customers

Description

Customer *churn* occurs when customers stop doing business with a company, also known as customer attrition. The dataset contains 10,127 rows (customers) and 21 columns (features). The “churn” column is our target which indicate whether customer churned (left the company) or not.

Usage

```
df=datadepot.load(“churn”)
```

Format

The churn dataset contains 10,127 rows (customers) and 21 columns.

The 21 variables are:

- **customer_id**: Customer ID.
- **gender**: Whether the customer is a male or a female.
- **age**: Customer’s Age in Years.
- **educaton**: Educational Qualification of the account holder (example: high school, college graduate, etc.)
- **marital_status**: Married, Single, Divorced, Unknown
- **income**: Annual Income (in Dollar). Category of the account holder (< \$40K, \$40K - 60K, \$60K - \$80K, \$80K-\$120K, > \$120K).
- **dependent_counts**: Number of dependent counts.
- **card_category**: Type of Card (Blue, Silver, Gold, Platinum).
- **months_on_book**: Period of relationship with bank.
- **relationship_count**: Total number of products held by the customer.
- **months_inactive**: Number of months inactive in the last 12 months.
- **contacts_count_12**: Number of Contacts in the last 12 months.
- **credit_limit**: Credit Limit on the Credit Card.
- **revolving_balance**: Total Revolving Balance on the Credit Card.
- **open_to_buy**: Open to Buy Credit Line (Average of last 12 months).
- **transaction_amount_Q4_Q1**: Change in Transaction Amount (Q4 over Q1).
- **transaction_amount_12**: Total Transaction Amount (Last 12 months).
- **transaction_count**: Total Transaction Count (Last 12 months).
- **transaction_change**: Change in Transaction Count (Q4 over Q1).
- **utilization_ratio**: Average Card Utilization Ratio.
- **churn**: Whether the customer churned or not (yes or no).

Source

This dataset is publicly available from Kaggle: <https://www.kaggle.com/sakshigoyal7/credit-card-customers>

Creators

Unknown

License

[CC0: Public Domain](#)

cpu CPU dataset

Description

This dataset contains detailed specifications and performance metrics for a range of computer processors (CPUs). It includes hardware characteristics such as core and thread counts, clock speeds, cache size, and thermal design power (TDP), along with market price information. The dataset is designed to analyze price-to-performance trade-offs.

Usage

```
df=datadepot.load("cpu")
```

Format

The dataset contains 45 rows and 12 columns.

- **brand**: The brand of the CPU (AMD or Intel)
- **model**: The model of the processor
- **architecture**: The microarchitecture of the CPU
- **p_cores**: Number of performance cores (P-cores)
- **e_cores**: Number of efficiency cores (E-cores)
- **threads**: Number of logical threads the CPU can execute simultaneously
- **base_ghz**: The base operating frequency of the CPU in gigahertz
- **boost_ghz**: The maximum turbo/boost frequency the CPU in gigahertz
- **cache**: Total L3 cache size in MB
- **tdp**: Typical thermal design power (TDP) in watts under standard load conditions
- **price**: Market price of the CPU (in US Dollars)

Source

Hardware specifications are based on publicly available manufacturer data. Price data was collected via Google search during Spring 2026 and reflects approximate retail market prices at that time.

Creators

Unknown

License

[Creative Commons Attribution 4.0 International](#)

credit South German Credit dataset

Description

This dataset classifies people described by a set of attributes as good or bad credit risks.

Usage

```
df=datadepot.load("credit")
```

Format

The dataset contains 1,000 rows and 20 columns (variables).

The 20 variables are:

- status
- duration
- credit_history
- purpose
- amount
- savings
- employment_duration
- installment_rate
- personal_status_sex
- other_debtors
- present_residence
- property
- age
- other_installment_plans
- housing
- number_credits
- job
- people_liable
- telephone
- foreign_worker
- credit_risk

Source

The dataset is publicly available from UC Irvine Machine Learning Repository: <https://doi.org/10.24432/C5QG88>

Creators

Hans Hofman

License

[Creative Commons Attribution 4.0 International](#)

credit_card Credit Card Fraud dataset

Description

Credit Card Transactions

Usage

```
df=datadepot.load("credit_card")
```

Format

The dataset contains 284,807 rows and 31 columns.

The dataset contains the following columns:

- **class**: Indicator of Fraud (1) or Non-Fraud (0)
- **time**: Seconds elapsed between each transaction and the first transaction in the dataset.
- **amount**: Transaction amount
- **v1, v2, ... v28**: Transformed features, obtained through Principal Component Analysis

More information is available at: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

Source

This dataset is publicly available from Kaggle: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

Creators

Machine Learning Group - ULB

License

[Database Contents License \(DbCL\) v1.0](#)

diamonds Diamonds dataset

Description

The diamonds dataset from ggplot2 is a comprehensive collection of data about diamonds, widely used in data science to practice visualization, statistical modeling, and machine learning. It includes detailed characteristics of individual diamonds and their corresponding prices. The dataset contains detailed information on over 50,000 diamonds, including both physical characteristics and pricing.

Usage

```
df=datadepot.load("diamonds")
```

Format

The diamonds dataset contains 53,940 rows and 10 columns.

The 10 variables are:

- **carat**: weight of the diamond (ranging from 0.2 to 5.01),
- **cut**: quality of the cut (Fair, Good, Very Good, Premium, Ideal),
- **color**: color grade, from D (most colorless) to J (least colorless),
- **clarity**: clarity grade, from I1 (least clear) to IF (flawless),
- **depth**: total depth percentage calculated as: $2 * z / (x + y)$,
- **table**: width of the top facet relative to the widest point (43-95%),
- **x**: length in mm
- **y**: width in mm
- **z**: depth in mm
- **price**: price in US dollars (ranging from \$326 to \$18,823).

Source

This dataset is publicly available in the ggplot2 R package: <https://ggplot2.tidyverse.org/reference/diamonds.html>

For more information related to the dataset see:

<https://search.r-project.org/CRAN/refmans/nodbi/html/diamonds.html>

Creators

Unknown

License

[MIT](#)

References

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4

drug Drug Classification dataset

Description

A synthetically generated dataset of 200 patients that includes their age, sodium-to-potassium (Na/K) ratio, and the prescribed drug type.

Usage

```
df=datadepot.load("drug")
```

Format

The drug dataset contains 200 rows and 3 columns.

The 3 variables are:

- **age**: patient age,
- **ratio**: sodium-to-potassium (Na/K) ratio,
- **type**: prescribed drug type with three levels: A, B and C.

Source

liver - R Package

Creators

Reza Mohammadi

License

[GPL-3](#)

house House dataset

Description

The house dataset contains 6 features and 414 records. The target feature is *unit_price* and the remaining 5 variables are predictors.

Usage

```
df=datadepot.load("house")
```

Format

The house dataset contains 414 rows and 6 columns. The 6 variables are:

- **house_age**: house age (numeric, in year).
- **distance_to_mrt**: distance to the nearest MRT station (numeric).
- **stores_number**: number of convenience stores (numeric).
- **latitude**: latitude (numeric).
- **longitude**: longitude (numeric).
- **unit_price**: house price of unit area (numeric).

Source

The data is publicly available from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set>

Creators

I-Cheng Yeh

License

[Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)

house_price House Price dataset

Description

This dataset, created by Dean De Cock, contains 1460 rows and 81 columns (features). The `saleprice` column is the target.

Usage

```
df=datadepot.load("house_price")
```

Format

The `house_price` dataset contains 1460 rows and 81 columns. More information about the variables can be found at: <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>

Source

The data is publicly available from Kaggle: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

Creators

Dean De Cock

License

[MIT](#)

loan Loan Approval

Description

The loan approval dataset is a collection of financial records and associated information used to determine the eligibility of individuals or organizations for obtaining loans from a lending institution.

Usage

```
df=datadepot.load("loan")
```

Format

The dataset contains 4,269 observations and 13 columns:

- **loan_id**
- **no_of_dependents**: Number of dependents of the applicant.
- **education**: Education of the applicant.
- **self_employed**: Employment status of the applicant.
- **income_annum**: Annual income of the applicant.
- **loan_amount**: Loan amount.
- **loan_term**: Loan term in years.
- **cibil_score**: Credit score measuring the applicant's creditworthiness.
- **residential_assets_value**: Value of residential property owned by the applicant.
- **commercial_assets_value**: Value of commercial property owned by the applicant.
- **luxury_assets_value**: Value of luxury assets.
- **bank_assets_value**: Value of the applicant's financial assets held in banks.
- **loan_status**: Outcome of the application: "Approved" or "Rejected".

Source

This dataset is publicly available from Kaggle: <https://www.kaggle.com/architsharma01/loan-approval-prediction-dataset/data>

Creators

Unknown

License

[MIT](#)

mpg Auto MPG dataset

Description

The Auto MPG dataset contains information on various car models from the 1970s and 1980s, with the goal of predicting fuel efficiency (miles per gallon, **mpg**). It includes attributes describing engine characteristics, vehicle weight, performance, model year, origin, and car name.

Usage

```
df=datadepot.load("mpg")
```

Format

The Auto MPG dataset contains 398 observations (cars) and 9 columns:

- **mpg**: miles per gallon, a continuous variable measuring fuel efficiency.
- **cylinders**: number of cylinders in the engine, a discrete factor with typical values 3, 4, 5, 6, 8.
- **displacement**: engine displacement in cubic inches, a continuous variable representing engine size.
- **horsepower**: engine power in horsepower, a continuous variable (may have missing values).
- **weight**: vehicle weight in pounds, a continuous variable.
- **acceleration**: time to accelerate from 0 to 60 mph in seconds, a continuous variable.
- **model_year**: year of the car model, a discrete variable usually coded as two digits (e.g., 70 = 1970).
- **origin**: origin of the car, a factor with 3 levels (1 = USA, 2 = Europe, 3 = Japan).
- **car_name**: car model name, a string variable for identification.

Source

This dataset is publicly available from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/9/auto+mpg>

Creators

R. Quinlan

License

[Creative Commons Attribution 4.0 International](#)

Remarks

Formatted origin labels for display

nyc_taxi NYC Yellow Taxi Zones dataset

Description The dataset summarizes trip characteristics across 260 New York City Taxi and Limousine Commission (TLC) pickup zones using aggregated NYC Yellow Taxi trip records from 2025.

The dataset was generated by aggregating individual trip records by pickup location. To mitigate the influence of extreme outliers and likely data-entry errors, observations were filtered using the 99.9th percentile of trip distance, fare amount, and tip amount. Records with non-positive trip distances or fares were removed. The cleaned dataset can be used to perform clustering analysis.

Usage

```
df=datadepot.load("nyc_taxi")
```

Format

The bank dataset contains 260 observations (zones) and 6 columns. The 6 variables are explained below.

- **zone_id**: Unique identifier of the taxi pickup zone as defined by the NYC Taxi and Limousine Commission.
- **zone_name**: Name of the taxi pickup zone.
- **trips**: Total number of taxi trips originating from the pickup zone during the study period.
- **med_distance**: Median trip distance (miles) for trips originating from the pickup zone.
- **med_fare**: Median fare amount (USD) for trips originating from the pickup zone.
- **med_tip**: Median tip amount (USD) for trips originating from the pickup zone.

Source

New York City Taxi and Limousine Commission (TLC) Trip Record Data: <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

Taxi zone boundaries are derived from the NYC TLC Taxi Zone geographic reference file.

Creators

NYC Taxi & Limousine Commission

License

[NYC.gov Terms of Use](#)

online_shoppers Online Shoppers Purchasing Intention Dataset

Description

This dataset contains information on online shopping session. Each observation represents a single session and contains information about the visitor's browsing behavior, engagement metrics, session timing, technical details (e.g., browser and operating system), visitor type, and whether the session resulted in a purchase. The dataset is imbalanced, with substantially more non-purchasing sessions than purchasing sessions.

Usage

```
df=datadepot.load("online_shoppers")
```

Format

The dataset contains 12,330 rows and 18 columns.

The 18 variables are:

- **administrative**: Number of administrative pages visited (numerical).
- **administrative_duration**: Total time spent on administrative pages (numerical).
- **informational**: Number of informational pages visited (numerical).
- **informational_duration**: Total time spent on informational pages (numerical).
- **product_related**: Number of product-related pages visited (numerical).
- **product_related_duration**: Total time spent on product-related pages (numerical).
- **bounce_rates**: Percentage of sessions leaving after viewing a page without further interaction (numerical).
- **exit_rates**: Percentage of pageviews where the page was the last visited in a session (numerical).
- **page_values**: Average value of a page before a completed purchase (numerical).
- **special_day**: Closeness of the visit to a major shopping-related holiday (numerical).
- **month**: Month in which the session occurred (categorical).
- **operating_systems**: Operating system used by the visitor (categorical).
- **browser**: Browser used during the session (categorical).
- **region**: Geographic region of the visitor (categorical).
- **traffic_type**: Source/type of website traffic (categorical).
- **visitor_type**: Indicates whether the visitor is new or returning (categorical).
- **weekend**: Indicates whether the session occurred on a weekend (binary).
- **revenue**: Indicates whether the session ended with a purchase (binary).

Source

The dataset is publicly available from the UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset>

Creators

C. Sakar and Yomi Kastro

License

[Creative Commons Attribution 4.0 International](#)

References

Sakar, C.O., Polat, S.O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31, 6893 - 6908.

red_wines Red Wines dataset

Description

The red_wines datasets are related to red variants of the Portuguese “Vinho Verde” wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The dataset can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Usage

```
df=datadepot.load("red_wines")
```

Format

The red_wines dataset contains 1599 rows and 12 columns.

The 12 variables are:

Input variables (based on physicochemical tests):

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- ph
- sulphates
- alcohol

Output variable (based on sensory data)

- quality: score between 0 and 10.

Source

The dataset is publicly available from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/186/wine+quality>

Creators

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis

License

[Creative Commons Attribution 4.0 International](#)

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4), 547-553.