Author: Daoxing Zhang
UID: dz2479

1. Purpose of this cheatsheet

This cheatsheet summarized some R built-in data sets to help people find their interested data sets to play with. For each data set, this cheatsheet listed all attributes of it with some brief explanations. When people do not know where to find a data set to practice exploratory data analysis and visualization, they can go through this cheatsheet to find some nice pre-loaded R data sets.

2. Data sets in package 'datasets'

mtcars: The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles.

mpg ..... Miles/(US) gallon
cyl ..... Number of cylinders
disp ..... Displacement
hp ..... Gross horsepower
drat ..... Rear axle ratio
wt ..... Weight (1000 lbs)
qsec ..... 1/4 mile time
vs ..... Engine (0 = V-shaped, 1 = straight)
am ..... Transmission (0 = automatic, 1 = manual)
gear ..... Number of forward gears
carb ..... Number of carburetors

iris: This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.

sepal length ..... cm
sepal width ..... cm
petal length ..... cm
petal width ..... cm
class ..... Setosa, Versicolour, Virginica

ToothGrowth: The Effect of Vitamin C on Tooth Growth in Guinea Pigs

len ..... numeric Tooth length
supp ..... factor Supplement type (VC or OJ)
dose ..... numeric Dose in milligrams/day

USArrests: This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Murder ..... numeric Murder arrests (per 100,000)
Assault ..... numeric Assault arrests (per 100,000
UrbanPop ..... numeric Percent urban population
Rape ..... numeric Rape arrests (per 100,000)

3. Data sets in package 'openintro'

fastfood: This data set contains nutrition amounts in 515 fast food items. restaurant ..... Name of restaurant
item ..... Name of item
calories ..... Number of calories
cal_fat ..... Calories from fat
total_fat ..... Total fat
sat_fat ..... Saturated fat
trans_fat ..... trans fat
cholesterol ..... Cholesterol
sodium ..... Sodium
total_crab ..... Total carbs
fiber ..... Fiber
sugar ..... Sugar
protein ..... Protein
vit_a ..... Vitamin A
vit_c ..... Vitamin C
calcium ..... Calcium
salad ..... Salad or not

mtl: The data are from a convenience sample of 25 women and 10 men who were middle-aged or older. The purpose of the study was to understand the relationship between sedentary behavior and thickness of the medial temporal lobe (MTL) in the brain.

subject ..... ID for the individual
sex ..... Gender, which takes values F (female) or M (male)
ethnic ..... Ethnicity, simplified to Caucasian and Other
educ ..... Years of educational
e4grp ..... APOE-4 status, taking a value of E4 or Non-E4
age ..... Age, in years
mmse Score from the Mini-Mental State Examination, which is a global cognition evaluation
ham_a ..... Score on the Hamilton Rating Scale for anxiety
ham_d ..... Score on the Hamilton Rating Scale for depression
sitting ..... Self-reported time sitting per day, averaged to the nearest hour
met_minwk ..... Metabolic equivalent units score (activity level). A score of 0 means "no activity" while 3000 is considered "high activity"
ipa_gqrp ..... Classification of METminwk into Low or High
aca1 ..... Thickness of the CA1 subregion of the MTL
aca23dg ..... Thickness of the CA23DG subregion of the MTL
ae_cort ..... Thickness of a subregion of the MTL
a_fusi_cort .... Thickness of the fusiform gyrus subregion of the MTL
a_ph_cort ... Thickness of the perirhinal cortex subregion of the MTL
a_pe_cort ... Thickness of the entorhinal cortex subregion of the MTL
asubic ..... Thickness of the subiculum subregion of the MTL
total ..... Total MTL thickness.

ames: This data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. See here for detailed variable descriptions. This data set has 82 parameters so that this cheatsheet will only list first 15 of them. For

more information, please first import "openintro" package and then type "?ames".

Order ..... Observation number
PID .. Parcel identification number - can be used with city web site for parcel review
area ..... Above grade (ground) living area square feet
price ..... Sale price in USD
MS.SubClass Identifies the type of dwelling involved in the sale
MS.Zoning Identifies the general zoning classification of the sale
Lot.Frontage ..... Linear feet of street connected to property
Lot.Area ..... Lot size in square feet
Street ..... Type of road access to property
Alley ..... Type of alley access to property
Lot.Shape ..... General shape of property
Land.Contour ..... Flatness of the property
Utilities ..... Type of utilities available
Lot.Config ..... Lot configuration
Land.Slope ..... Slope of property

seattlepets This data set contains names of registered pets in Seattle, WA, between 2003 and 2018, provided by the city's Open Data Portal.

license_issue_date .. Date the animal was registered with Seattle
license_number ..... Unique license number
animal_name ..... Animal's name
species ..... Animal's species (dog, cat, goat, etc.)
primary_breed ..... Primary breed of the animal
secondary_breed ..... Secondary breed if mixed
zip_code ..... Zip code animal is registered in

acs12: This data set contains Results from the US Census American Community Survey, 2012.

income ..... Annual income
employment ..... Employment status
hrs_work ..... Hours worked per week
race ..... Race
age ..... Age, in years
gender ..... Gender
citizen ..... Whether the person is a U.S. citizen
time_to_work ..... Travel time to work, in minutes
lang ..... Language spoken at home
married ..... Whether the person is married
edu ..... Education level
disability ..... Whether the person is disabled
birth_qrtr ... The quarter of the year that the person was born, e.g. Jan thru Mar.

cpu: This data set contains data on computer processors released between 2010 and 2020.

company ..... Manufacturer of the CPU
name ..... Model name of the processor.
codename ... Name given by manufacturer to all chips with this architecture
cores ..... Number of compute cores per processor
threads ..... The number of threads represents the number of simultaneous calculations that can be ongoing in the processor.
base_clock ..... Base speed for the CPU in GHz
boost_clock ..... Single-core max speed for the CPU in GHz
socket ..... Specifies the type of connection to the motherboard
process ..... Size of the process node used in production in nm
l3_cache ..... Size of the level 3 cache on the processor in MB
tdp ..... Total draw power of the processor
released ... Date which the processor was released to the public

loans_full_schema This data set represents thousands of loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals. This data set has 55 variables so that this cheatsheet will only list first 14 of them. For more information, please first import "openintro" package and then type "?loans_full_schema".

emp_title ..... Job title
emp_length ..... Number of years in the job, rounded down. If longer than 10 years, then this is represented by the value 10
state ..... Two-letter state code.
home_ownership ..... The ownership status of the applicant's residence
annual_income ..... Annual income
verified_income ... Type of verification of the applicant's income
debt_to_income ..... Debt-to-income ratio
annual_income_joint ..... If this is a joint application, then the annual income of the two parties applying
verification_income_joint Type of verification of the joint income
debt_to_income_joint .. Debt-to-income ratio for the two parties
delinq_2y .... Delinquencies on lines of credit in the last 2 years
months_since_last_delinq ..... Months since the last delinquency
earliest_credit_line . Year of the applicant's earliest line of credit
inquiries_last_12m ... Inquiries into the applicant's credit during the last 12 months

4. Data sets in package 'agridat'

australia.soybean: This data set contains Yield and other traits of 58 varieties of soybeans, grown in four locations across two years in Australia. This is four-way data of Year x Loc x Gen x Trait.

env ... environment, 8 levels, first character of location and last two characters of year
loc ..... location
year ..... year
gen ..... genotype of soybeans, 1-58
yield ..... yield, metric tons / hectare
height ..... height (meters)
lodging ..... lodging
size ..... seed size, (millimeters)
protein ..... protein (percentage)
oil ..... oil (percentage)

nass.barley, nass.corn, nass.cotton, nass.hay, nass.rice, nass.sorghum, nass.soybean, nass.wheat: Yields and acres harvested in each state for the major agricultural crops in the United States, from approximately 1900 to 2011. Crops include: barley, corn, cotton, hay, rice, sorghum, soybeans, wheat.

year ..... year
state ..... state factor
acres ..... acres harvested
yield ..... average yield