Data sets

Data tameRs

season 1 /episode 1

Press A to switch from slides to continuous text.

Press T to see the table of contents.

What is the topic of this episode?

You can process data in many ways. The course Data tameRs will introduce you to this variety. For this reason we will work on datasets of different size (from several to several hundred thousand rows), with various level of complexity (from two to two thousand columns) and prepared in different formats (text data, quantitative data and qualitative data).

All of these data sets were gathered in a single big package called PogromcyDanych to facilitate access to them. It contains all sorts of data sets such as data about offer prices of cars, results of opinion polls, information about changes in stock market indices, data about names given to newborn babies, ski jumping results or information about outcomes of breast cancer patients treatment.

In order to install this package you need to enter the following command in R program (you need to do this only once). It will download the data sets from the Internet and install them on your disc.

Once you install the package PogromcyDanych you need to activate it to be able to use the data sets. Use the following command to activate the package (after each activation of R).

Next slides will show you all the data sets included in that package.

Two language versions

Data tameRs course is available in two language versions –Polish and English. However, it was not necessary to create two language versions of data sets and thus double the data. For this reason your data loaded from the PogromcyDanych package will be in Polish.

If you want to switch to English use the command setLang() as in the example below. This function will translate the Polish names in the data sets (as well as names of columns and values) into English.

As a result of the function the following data sets will be copied into the namespaces:

|  |  |  |
| --- | --- | --- |
| **Polish name** | **English name** | **What type of data is this?** |
| koty\_ptaki | cats\_birds | Characteristics of 13 species |
| imiona\_warszawa | warsaw\_names | Names of babies born in Warsaw |
| mandatySejmik2014 | votes2014 | Results of the local elections |
| auta2012 | auta2012 | Offer prices of cars in 2012 |
| WIG | WIG | WIG Index in 2014 |
| pearson | pearson | Pearson’s data on height |
| galton | galton | Galton’s data on height |
| skiJumps2013 | skiJumps2013 | Results of ski jumping competitions season 2013/2014 |
| diagnoza | diagnosis | Data from the project Diagnoza Społeczna (Social Diagnosis) |
| diagnozaDict | diagnosisDict | Explanation of columns’ names |
| serialeIMDB | seriesIMDB | Data on TV series on the basis of the IMDB movie database |
| TCGA\_BRCA | TCGA\_BRCA | Data on breast cancer patients |

The setLang(‘’pol’’) command will translate the names back into Polish.

The fastest cats and birds in the world

The data set cats\_birds was created as a so-called toy-example. It is a small set (13 rows and 7 columns) which can be fully displayed on the computer screen. You can use it to practice basic operations on the data.

The data set juxtaposes characteristics of 13 chosen species of cats and birds presented in separate rows. The chosen species are the fastest members of both groups. Each row contains information on a different species.

The columns of this data set present respectively:

* name of the species,
* maximum weight of an individual animal expressed in kilograms and maximum body length in meters,
* maximum speed (in the horizontal flight in case of birds, not during the dive) expressed in kilometers per hour,
* habitat and length of life in years,
* the last column specifies whether the species is a big cat or a bird.

You can use the head() function to display the first 6 rows of this data set.

Names of infants born in Warsaw

The data set warsaw\_names was prepared to present methods of trend analysis and methods of text analysis.

The data set contains information on the number of babies with a certain name born in Warsaw in the subsequent months of the period between 2004 and 2014. This data was downloaded from the website http://gorny.edu.pl/imiona/index.php where you can also find information about the current number of births.

There are 84816 rows in this set and the data concerns different names, from very popular names to rather rare ones.

The columns of this data set present respectively:

* gender and name of the child,
* year and month in which the data was gathered,
* number of babies with a given name born in Warsaw in a certain year and month.

You can use the head() function to display the first 6 rows of this data set.

Use the levels() function to display a list of all the names included in this data set.

2014 local election results

The data set votes2014 was prepared to show ways in which spatial data can be analyzed and visualized. The data was downloaded from the website of the National Electoral Commission (http://wybory2014.pkw.gov.pl/pl/).

The data presents information about the number of seats in the regional parliaments in Poland won in 2014. The set consists of 16 rows and 9 columns. Each row presents results concerning a different voivodeship.

The subsequent columns present respectively:

* name of the voivodeship,
* number of seats won by PSL, PiS, PO, SLD and other political parties,
* what is the percentage of valid votes of the total number of votes cast by the people entitled to vote,
* longitude and latitude of the geographical centre of the voivodeship.

Below you can see 6 first rows of this data set.

Offer prices of second-hand cars in 2012

The data set auta2012 was designed with a view of creating opportunity to work on a big and interesting data set concerning transactions. Transaction data is usually rather extensive and it is processed in a very specific manner.

This data was extracted from the announcement service otomoto.pl in 2012. The set contains information about over 207 thousand car sales offers.

Each offer (each row) is described by 21 variables including:

* price, currency and information whether the given price is a net or a gross price,
* information on engine power expressed in horsepower and kW and its capacity,
* information about model, make and version of the car,
* information on the type of drive, year of manufacture and mileage (in km),
* information about color, number of doors, country of origin, country of registration and accessories.

Information about accessories is described by a very long line of text with values separated by comas.

Below you can see 6 first rows of this data set.

WIG, index of the Warsaw Stock Exchange

The data set WIG was prepared so that we could take a look at the analysis and visualization of data with longer time series. In this case the data concerns daily quotations on a stock exchange.

Information about the daily performance for WIG (Warsaw Stock Exchange Index) from the period December 2013-Nevember 2014 were downloaded from the webpage of the Warsaw Stock Exchange.

The subsequent columns present respectively:

* data of quotation,
* opening and closing share price, minimum and maximum share price,
* change in share price and the turnover in thousands of PLN.

There are 248 rows in the data set and each row corresponds to one day of quotation. The Warsaw stock Exchange is closed on Saturdays, Sundays and holidays and the number of quotation days is respectively smaller.

Below you can see 6 first rows of this data set.

Height of parents and children

The word regression comes from the pioneering studies carried out by Francis Galton and Karl Pearson concerning relationship between height of parents and their children.

The original data sets created by the scientists are available in the variables galton and pearson. Using these data sets we can describe relationship between the height of a son and his father (pearson) and relationship between the weighted average height of the parents (galton) and their son’s height.

Below you can see 6 first rows of each data set. There are 928 rows in the galton data set and 1078 rows in the pearson data set.

Results of ski jumping competitions

The data set skiJumps2013 contains results of ski jumping competitions season 2013/2014. It was created during construction of a diagram for a competition for the best data visualization during PAZUR conference.

The data set consists of 2130 rows. Each row describes one jump during the ski jumping competitions season 2013/2014.

Each jump is described by 16 columns. Their content is described in the data set skiJumps2013labels.

The skiJumps2013 data set contains the following information about each particular jump:

* information about competition –city, country, parameters of a ski jump,
* information about the ski jumper –name, surname, nationality, date of birth,
* information about the jumps themselves (two jumps if the jumper jumped twice or one jump if the jumper did not qualify for the second round) –speed, distance, points added to the classification.

Below you can see 6 first rows of this data set.

Social Diagnosis

Results of the panel research called Social Diagnosis (Diagnoza Społeczna) form a very interesting data set. The same people from a chosen group of households are polled every two years. The questions concern various issues and the answers of the participants allow for construction of a model of social changes taking place in Poland. You can find more information about this research, its results, as well as information about the data set on the project’s website http://diagnoza.com.

The data set in the form processable by R is available on the website https://github.com/pbiecek/Diagnoza. You can install it using a command install\_github("pbiecek/Diagnoza") after previous activation of the library(devtools) package.

The whole data set is large and small computers might have problems with it. For the purposes of this course I have prepared a subset of the data set from the Social Diagnosis research.

The subset is called diagnoza and it consists of 38461 rows. Each row presents answers of one person. The responses received in the pools are presented in 36 columns/variables. The names of the variables correspond to the questions asked in the poll http://diagnoza.com/pliki/kwestionariusze\_instrukcje/kwestionariusze\_2013.pdf. The data set diagnozaDict gives full versions of all the questions.

The variables describe among other things:

* names of the respondents,
* analytical balance resulting from the draw method,
* number of years of study, gender, education, height, weight, income,
* answers to chosen questions concerning the worldview.

The data set called diagnozaDict describes names of columns form the diagnoza data set. These are the first six columns.

Below you can see two chosen rows taken from that data set.

Assessment of episodes of TV series

The data set called serialeIMDB contains information about popularity of particular episodes of TV series.

I downloaded the data about assessment and number of votes cast for each episode of 200 most popular TV series from the website http:///www.imdb.com. The data set also includes names of the series and titles of particular episodes.

Each of 20122 rows describes one episode. The subsequent variables/columns give the following information:

* title of the series and title of an episode,
* number of the season and number of the episode,
* average assessment of a given episode,
* number of votes,
* identification number of the series in IMDB database.

Below you can see 6 first rows of this data set.

The prognosis of breast cancer

I downloaded a subset of clinical and genetic data concerning breast cancer patients (mainly female patients) from the database The Cancer Genome Atlas (TCGA) http://cancergenome.nih.gov/.

The data set TCGA\_BRCA contains data about 999 patients. Each row corresponds to one patient and gives five basic pieces of information:

* information on TP53 gene mutation,
* patient’s gender,
* information whether the patient lived for five years after the surgery,
* number of days between the surgery and death (if it occurred),
* information whether there was a relapse and a new tumor.

Below you can see 6 first rows of this data set.

Below you can see number of patients broken down by gender and type of TP53 gene mutation (Wild type means lack of mutation).