**Selection of Variables from Data [dplyr / select]**

pogRomcy danych (Data Masters) Season 1 / Episode 20

All rights reserved.  
The use shall be subject to prior consent of the author.

Press A to see the plain text instead of the slides.  
Press T to display the table of contents.

**What Is This Episode About?**

When working with the data on the car sales offers you could experience how uncomfortable the work with the data in many columns is, if you need only some of them. In the Episode 7, we showed how to select the columns, which met a specific criterion, using the operator [.

In our everyday work, in order to select the columns from the whole set of data, it is not only more convenient, but also quicker to use the dedicated function select() of the package dplyr.

In this episode, we will learn:

* How to select only the columns chosen by us from the set of data.
* How to select all columns except those chosen by us from the set of data.

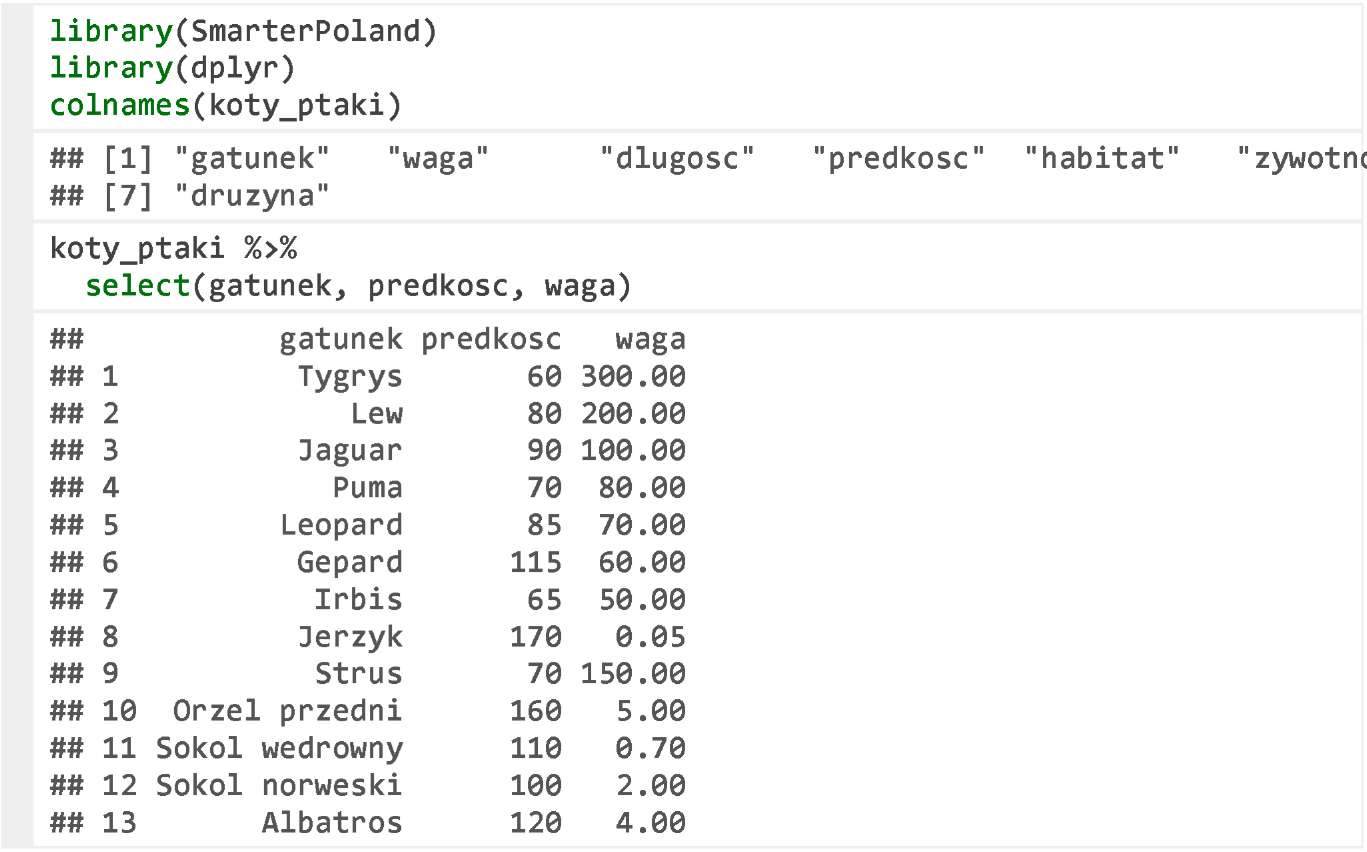
To illustrate these issues, we will use two sets of data. The first small set of data is koty\_ptaki and the second, considerably bigger one is auta2012, both available in the SmarterPoland package.

Positive Variable Selection

The data set koty\_ptaki includes few columns, and they all can be displayed on the console. Sometimes, however, the redundant variables only make it difficult to focus on what's important.

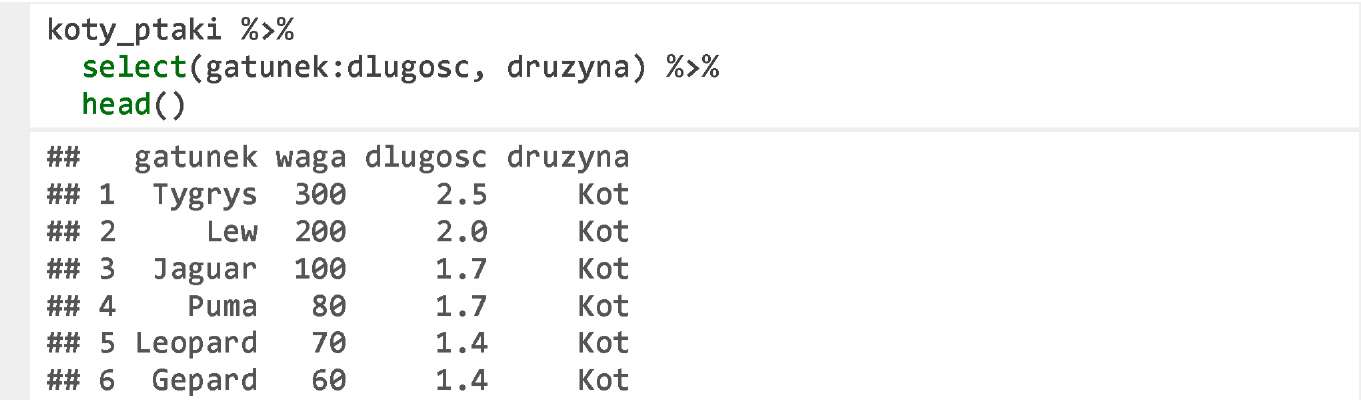
In order to change the column order or select only a subset of the chosen columns, we can use the function select ().

In the example below, we will display the column names in the data set koty\_ptaki. And then, we will choose three of these columns in the following order gatunek, predkosc, waga.

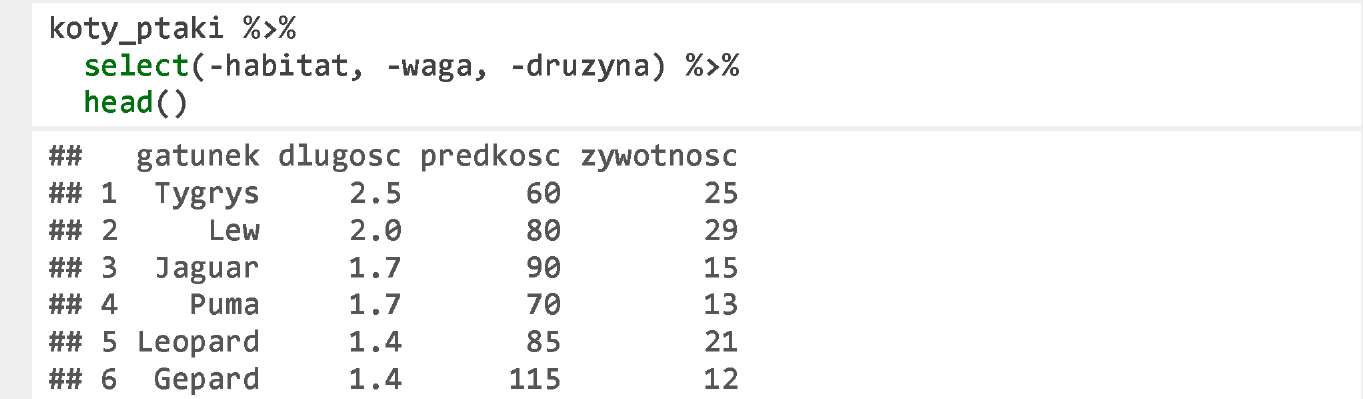


**Variable Selection Using Operators ‘:’ or ‘-’**

If there are many columns in a set of data, listing their names one after another can be inconvenient. In this case, it is worth using the operator :, which allows selecting the indicated range of the columns.



In the function select () you can use the operator - to select all columns from the data frame except those indicated.



**Cars**

We will practice the filtering operations on much bigger set of data, which cannot be viewed in whole on the screen. Namely, the set of data on the second-hand car offer prices.

The set of data auta2012 is available after downloading the package SmarterPoland. Please find below the list of the column names of this set of data. For more information on this set of data, see https://rawgithub.com/pbiecek/MOOC/master/0\_dane/0\_dane.html.

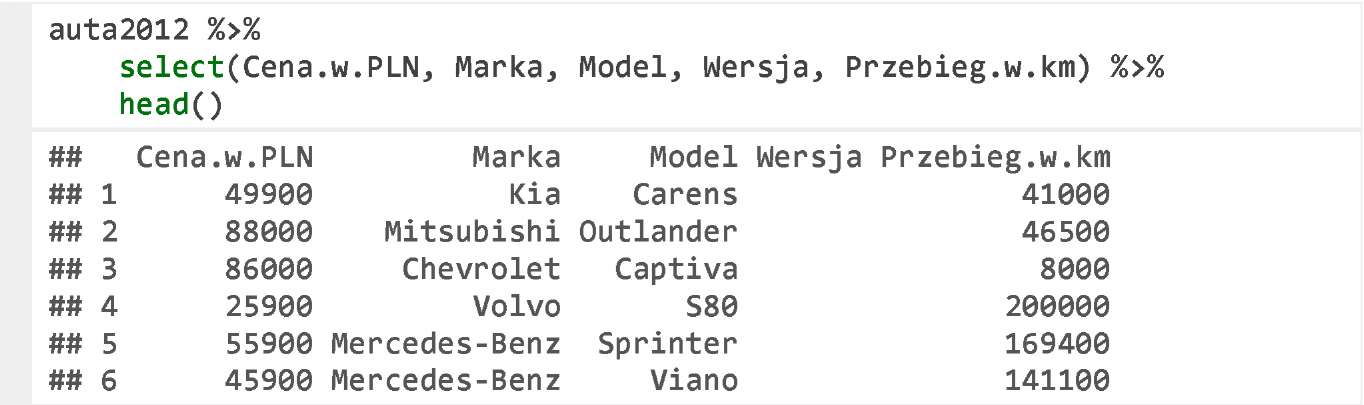
There are the following data in this set.



The function select () enables to easily select some useful variables from a large set of data. The first argument is the data frame, and the others are the variables, which we want to leave.

**Positive Variable Selection**

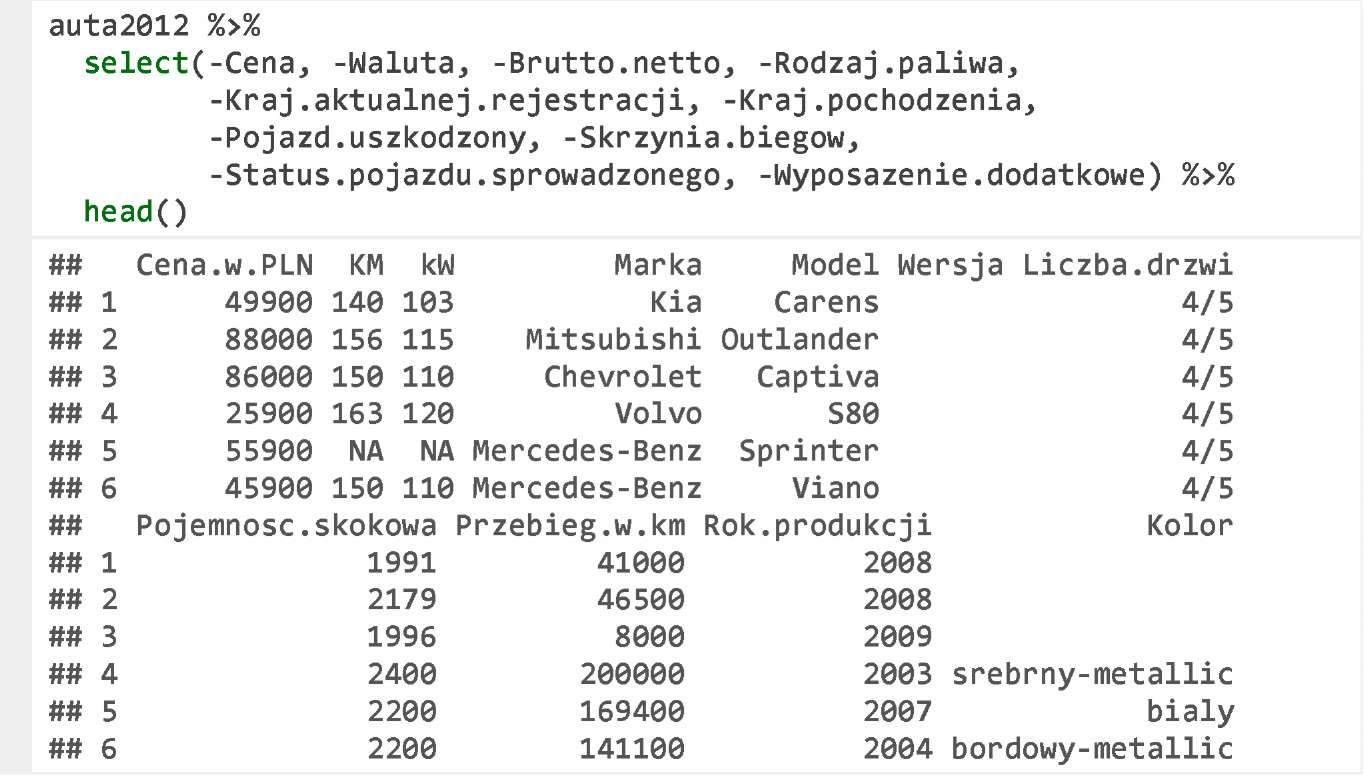
Let’s assume that we would like to select only price (Cena), make (Marka), model, version (Wersja), and mileage in kilometres (Przebieg.w.km). Just specify these variables as the arguments of the function select ().



**Negative Variable Selection**

If you want to remove some columns, you can use the operator -.

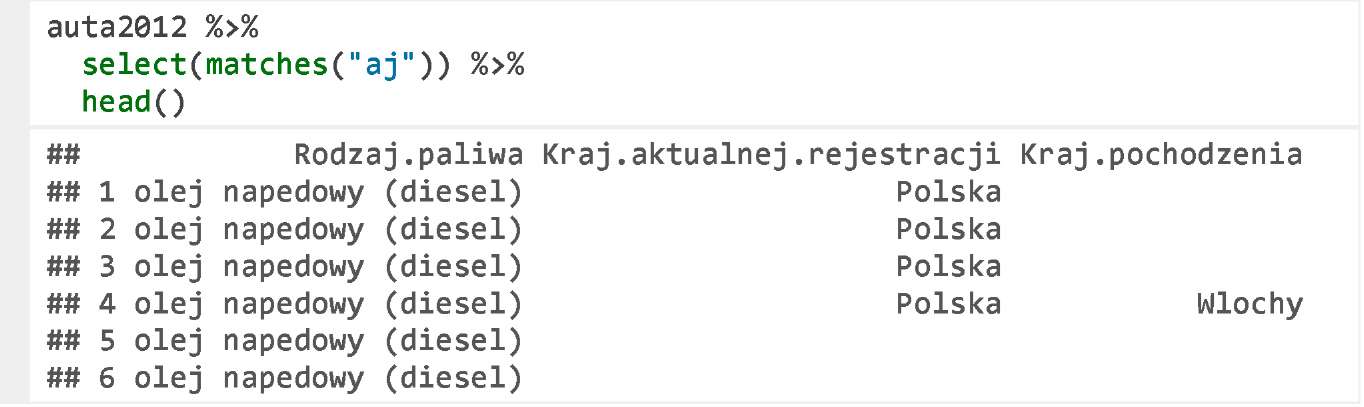
We remove nine indicated columns in the below example.



**Variable Selection with Pattern**

When selecting the columns, you can also indicate those, the names of which fit for a given pattern.

The below examples uses the function matches () to select all column names, which include the pattern aj in their name.



**Exercises**

* Select only Volvo cars, arrange them by cylinder capacity, and display only three columns: make (Marka), price (Cena.w.PLN ), and colour (Kolor).
* Select only those columns, the name of which ends with a.