Intro to the course ‘Data Crunchers’

Data Crunchers

season 1 /episode 1

All rights reserved.

Use of the content permitted only with the consent of the author.

Press A to switch from slides to continuous text.

Press T to see the table of contents.

What kind of course is this?

Data Crunchers is a set of two courses on processing, visualizing and modeling data.

The target audience of the courses are people who work with data, those who would like to work with data and also those who just want to know what work with data is like. Both courses are at the elementary level. The only thing that is taken for granted is a general familiarity with tables with data from spreadsheets such as e.g. Excel and basic computer skills.

Some prior knowledge of programming would be very useful, although it is not necessary. All new concepts will be explained on the examples connected with analysis of interesting data.

The first course will show you how to work with R –a free and very powerful tool for data analysis. You will learn how to write programs, create functions and loops, load, process and save data.

The second course will teach you how to identify good and bad diagrams. Then you will be instructed how to design proper charts using R. The next part of the second course is devoted to statistical models and methods. It will show you how to estimate relationships between two or more variables.

Both courses end with a set of 20 exercises which you try to do on your own. Those who complete 13 or more tasks will be granted the title of ‘Data Cruncher’ and a certificate of course completion signed by the University of Warsaw.

You need to be registered to access the course materials. Registration is free of charge and it is open to everybody interested. Once you are registered, you have access to the forum, educational materials and exercises. The educational materials were divided into short units devoted to various issues. At the end of each unit you will find a short task checking your understanding. Materials also contain a key to the exercises.

Participants of both courses can communicate through a special forum. If some part of material is exceptionally difficult or problematic, or if you discover some interesting solutions during your work, you can share them on the forum. The support team reads the posts on the forum and answers all questions, even the most basic ones. Other participants of the course may suggest something or help you in your problems as well.

The first season is about R

The first season is devoted to R. Its first four episodes present reasons for learning to use that program; they explain how to install R, a useful editor and additional packages necessary for work with R. Later on in the course you will learn basics of working with R –you will be taught how to write comments, where to enter them and where to look for the results.

The next three episodes discuss the problem of loading and saving tables with data and basic operations performed on such tables. They will show you how to load a text or Excel file into R as well as how to save the results and select only certain columns or rows from the results.

Episodes 8, 9 and 10 will help you to automate calculations. They explain the workings of loops and conditional commands. They also show how to create your own functions.

Episodes from 11 to 15 will introduce you to functions and operations typical for the most popular types??? You will learn uses of number, text and data vectors. You will discover what are logical and qualitative variables and what operations can be performed on them.

Episodes from 16 to 24 present a wide range of functions allowing for free conversion of data. When you master these functions you will be able to perform 90% of the typical operations associated with data processing. You will know how to filter, convert, aggregate and group data and how to change its structure.

This is an elementary course but it is concluded with an episode showing where you may look for further information on R. If you grow fond of R, you will know how to further explore its possibilities.

The second season is about visualization and modeling

While the first course shows how to process data, the second one tells why it is a worthwhile occupation. This part presents issues connected with data exploration, visualization and statistical modeling.

The second season starts with three episodes presenting the milestones in the history of data visualization, difficulties connected with data presentation and problems resulting from the manner in which our mind interprets images. Then I will show you examples of honest as well as dishonest diagrams.

Episodes 4, 5, 6, and 7 will introduce you into the world of ggplot2 package. You will learn about the simplest and the most advanced commands allowing for creation of any type of diagram in R. If you become interested in designing diagrams in R, you will be able to develop your skills in data visualization thanks to the materials suggested in the episode no 8.

The subsequent episodes present basic statistical tools for data exploration necessary in statistical data analysis. You will learn from them how to compare two means in a statistical manner. You will also learn how to check whether two phenomena occur independently of each other or are connected by some relationships. What is more, you will be shown how to create regression models in order to construct for example a forecast concerning some characteristic. This unit will be also concluded with a list of suggested sources in which you will be able to find more information on statistical modeling.

Both courses include examples of very interesting yet unusual uses of data analysis. They show in what ways data analysis is used in the analysis of survey data, robotics, linguistic engineering and meteorology.

You should not be afraid of a steep learning curve

A steep learning curve is characteristic of data analysis –especially so in case of the analysis using R. Many new things seem to be difficult at the beginning. However, you will have a lot of materials at your disposal. The forum is continuously monitored and all the questions are answered so that every participant of the course has necessary assistance in his first encounter with R. We assure you that your effort put in learning R will not be wasted. You will have better understanding of data analysis in general. You will process the data more efficiently with time. Due to R’s elasticity and expression you will meet virtually no barriers in analyzing the most diverse types of data.

When you work with the slides you can press T at any time in order to see the tale of contents or A to switch from slides to continuous text.