Annals of "Dunarea de Jos" University of Galati
Fascicle I. Economics and Applied Informatics
Years XXIII – nº3/2017
ISSN-L 1584-0409 ISSN-Online 2344-441X
www.eia.feaa.ugal.ro

Data Science and Some Instruments

Corina SBUGHEA*

ARTICLE INFO

Article history:
Accepted November 2017
Available online December 2017

JEL Classification
C80, C88

Keywords:
Data, Information, Knowledge, Coding skills

ABSTRACT

This paper is addressed to beginners, who want to form an overview on the field of Data Science, on the skills needed to access available IT tools, for obtaining meaningful and valuable analyzes in developing new strategies.

© 2017 EAI. All rights reserved.

1. Introduction

In the last decades, the development of the IT domain has facilitated the storage of large amounts of data. Complex economic organisations have exponentially stored data regarding their current activity, their clients or the products and services provided. Accordingly, this has led to the need to extract information and knowledge from these amounts of data. In addition to the demand for specialists in managing large databases, another demand has emerged on the labour market, that for specialists in the processing and analysis of this data, who are able to make the most accurate forecasts. The know-how required from these specialists has exceeded that required from statisticians, in the classical sense of the word. They must have analytical, mathematical, statistical abilities and, above all, they must have computer science background and coding skills.

2. Theoretical framework

The IT domain specifies the difference between data and information. The term *data* comes from the Latin *datum* (a given fact) and it refers to facts or measures, which can be observed or measured. If these messages are contextualised or subjectively interpreted, then they become items of *information*.

In the domain-specific literature, there are more definitions given to the phrase *Data science*. Broadly speaking, *data science* includes all the automated means, technologies and methods used for collecting, preparing, analysing and visualising large amounts of data, with the ultimate goal of gaining new knowledge.

James Nicholas Gray, an American computer scientist, asserts that Data science represents the fourth paradigm of science, alongside the empirical, the theoretical and the computational ones. This new paradigm is data driven. Gray considers that science is constantly changing due to information technology evolution and to data abundance. IBM estimates that 90% of the data available in the world today has been created in the last two years [12]. This storage trend will evolve exponentially. Nowadays, we collect data in various forms, from a wide variety of sources such as mobile devices, specialised tools, the Internet, and together with the development of technology, the phenomenon will continue to increase.

In accordance with this trend, many researchers believe that, in a growing number of traditional subject matters, new subdomains will emerge, which will deal with their "computational" or "quantitative" side. Within an organisation, the ultimate goal of data science is to constantly support and improve the decision-making strategies and processes connected to data (Florian V., Neagu G., 2016). All this involves going through different stages, from collecting data to their manipulation, exploration, modelling, validation, visualisation and transmission to decision-making factors, with the ultimate goal of drawing correct, and especially most efficient, strategic conclusions.

*Dunarea de Jos University of Galati, Romania. E-mail address sbughea@yahoo.com

Although data science seems to be connected more to domains such as databases or computer science, computer skills are not the only one required. The researcher Jeffrey Stanton, with the help of a case study, identified a number of features that arise from the roles played by specialists in shaping and implementing mechanisms related to: data architecture, data acquisition and data archiving. He groups the identified features into three categories: communication skills, data analysis skills and ethical reasoning skills (Stanton J., 2012):

- knowledge of the application domain and of the context in which the data will be used;
- a characteristic, which is considered vital by the author, is the ability to communicate with the end user; consequently, the data science specialist must be able to translate both the computational terms and the terms specific to the application domain;
- creating an overview of the analysed system, i.e. the ability to imagine data flows and circuits between relevant subsystems;
- knowing how to represent data, i.e. understanding how data and metadata is stored and correlated, fact which reflects a particular way of structuring them;
- the ability to analyse and process data, which refers to operations such as transformation, synthesis and drawing conclusions;
- the ability to communicate the results of the analysis in a meaningful manner to the user, because numerical data is most often than not difficult to grasp by non-specialists;
- the ability to identify the limits imposed by the collected data because real systems may imply a certain degree of uncertainty but also the possibility of increased data accuracy;
- to all the above considerations, the ethical component of the process may be added, because one must also take into account data privacy; as a result, specialists must identify and communicate limits in order to prevent the misuse of data.

3. Popular instruments

The most common instruments in the field are R and Python, but their utility is not identical. R is, indeed, a very useful instrument for data analysis, whereas Python is a dynamic multi-paradigm programming language, which can both develop object-oriented applications and allows for imperative, functional or procedural programming. Since it is a high-level programming language, Python can develop complex applications, much easier and faster than C or C++, although the basic implementation of Python is written in C and was designed to be portable.

But these are not the only instruments specialists can use. SAS is also a useful software tool, which allows the processing of different types of data, with multiple functionalities, being comparable to R in terms of difficulty. A drawback of SAS is that it is not open source, this is why it is used only by big companies that can afford to buy it. A less useful tool but, nevertheless widely used, is SPSS (Statistical Package for the Social Sciences).

Author Michael Brzustowicz proposes and develops an insight into the mathematical functionalities of the Java programing language (Brzustowicz , 2017), considering it can be a robust, convenient and accurate tool in big data processing. The approach is addressed to the learners of the language, who want to acquire data science skills, too.

Many authors consider that Scala is the new star of the domain. This is also a multi-paradigm programming language that supports both object-oriented and functional programming. Scala provides powerful functional libraries for interacting with databases and building scalable frameworks (Bugnion, 2016). This is considered to be more a language for engineers, a substitute for Java, unlike Python, which is a scientific programming language.

There are also some data science specialists who prefer low level coding, like C++, due to the speed and control advantages brought by the machine code resemblance.

The choice of the right instruments is arbitrary, depending on the proposed objectives and on the specialist's level of training. The research regarding the dynamic of the number of people, who are already studying a certain language or just planning to start learning it, based on the data of the first quarter from 2017, show that the most progressing languages are Python (+2.7%), JavaScript (+0.5%), and R (+0.5%) [11].

3.1. R language

First developed as a statistical software for didactic purposes by Ross Ihaka and Robert Gentleman, project R gathered an impressive community of users and volunteer developers (R Development Core Team). They are constantly updating the programming environment, the newest methods and techniques usually being implemented in R first. R programming language is considered to be a dialect of the S language, designed by AT&T Bell Laboratories; nevertheless, there are significant differences between the two, aspects which are highlighted by the creators themselves in the article ”R: a language for data analysis and graphics”, written in 1996.

R is an interpreted language, which means that the speed of execution is slower than in the case of a compiler, and any syntax errors are only noticed during program execution. The language is addressed to users who have mathematical and statistical knowledge because it has a relatively high complexity level.

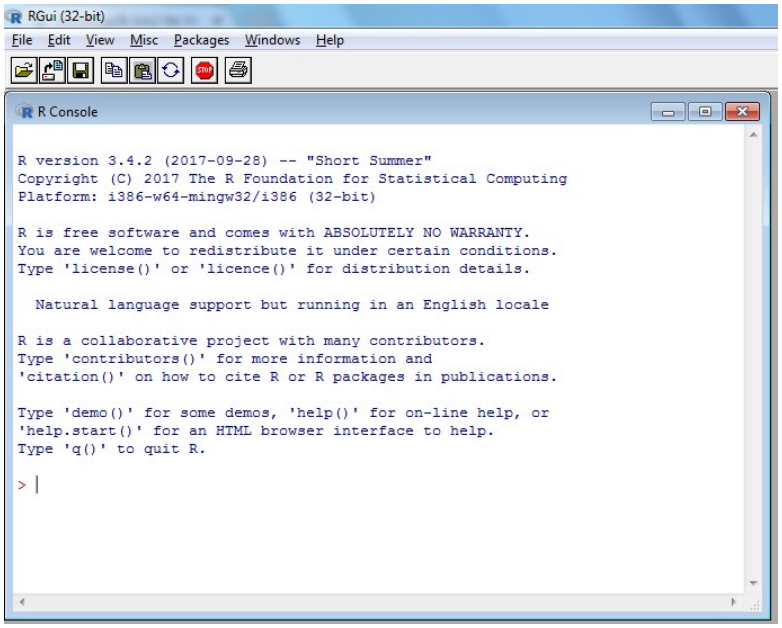

Fig.nr. 1. R window

Nevertheless, there have also been developed a number of simple graphical user interfaces (GUI) for R, which minimise the effort of the user, such as RStudio or RCommander.

RStudio represents an integrated development environment and it is probably the most used R graphical interface at present. It includes a menu bar (File, Edit, Code, View, Plots, Session, Built, Debug, Tools, Help) and a toolbar and the window is divided into four quadrants, two of which contain more tabs: Editor, Console, Environment/ History/Build, Files/Plots/Packages/Help/Viewer.

The creation of a new script can be done from the File menu or with the help of the key combination Ctrl+Shift+N. Here, one can write and execute commands with the help of the Run button or of the Ctrl+R accelerator. The results will be displayed in the Console quadrant and all the objects created during a work session (data series/functions) will be available in the Environment quadrant. The data series can be viewed in spreadsheets, too, with the help of the buttons available on their right.

The tag History contains the list of the commands executed up to the current moment. These commands can be partially or totally saved in a script. In the fourth quadrant, the Files tag contains files and folders, and the graphs can be found under the tab Plots. If there are more graphs, the navigation between graphs is made with the help of the arrows. In order to use them outside the work environment, you can use the Export option, in JPG, PNG, PDF formats or you can copy them in the Clipboard memory so that you can insert them in other files.

The facilities offered by the R environment can be extended with a series of add-ons, grouped under the tag Packages. These are the default packages when installing RStudio, but others can also be added.

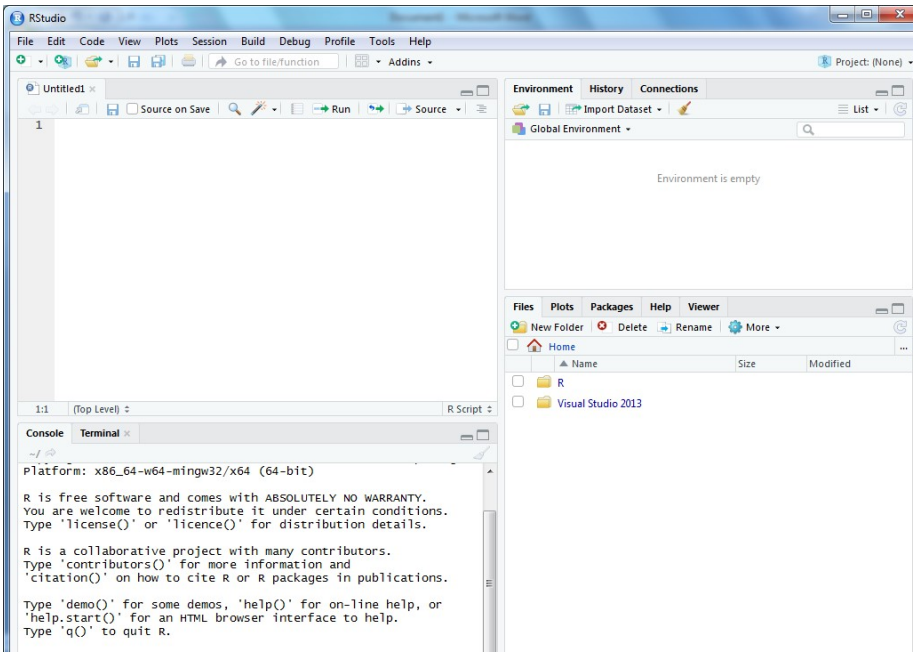

Fig.nr. 2. RStudio interface

R Commander is a graphical user interface, implemented as an R package by John Fox, who promises a point-and-click type of interaction. Nevertheless, Rcmdr offers access only to a part of the multiple facilities provided by R language and can be extended through plug-in packages [10]. R Commander has a menu bar (File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, Help), a toolbar, an area for typing commands (Script Window), an area for results (Output Window) and an area for displaying messages and warnings (Message Window).

The options from the Data menu allow for data access and manipulation, the submenus subordinated to the Statistics menu offer access to basic statistical analyses as well as non-parametric tests, and the Models menu contains options that refer to trust intervals, hypothesis testing, probability distributions etc. The buttons from the toolbar refer to the current data set and allow editing, viewing and application operations of an available model. The data set can be manually typed or can be imported from another file. The exploration of data from a graphical point of view is limited to the options from the menu in Rcmdr, but advanced users can extend the investigation by using the command line. R Commander may be installed from Rstudio, too, with the Install option, under the tag Packages and will be loaded when you tick it.

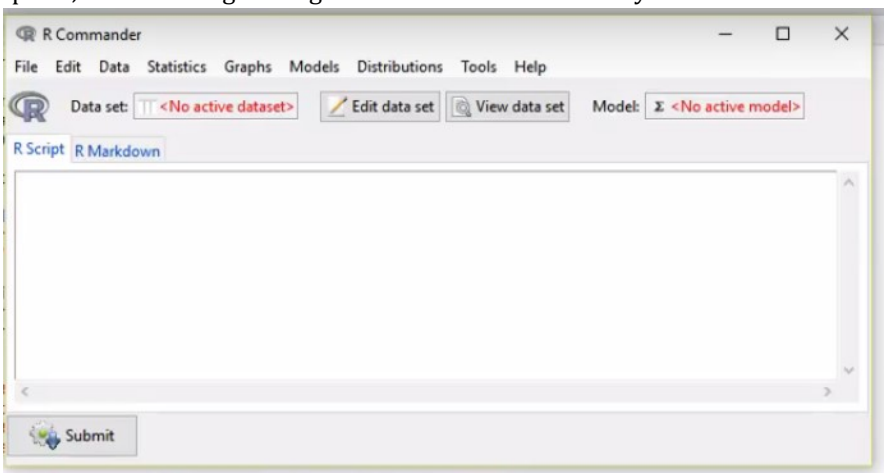

Fig.nr. 3. R Commander interface

Graham Williams considers that R is all-encompassing and the best statistical analysis package available at present, with high portability, being also compatible with other instruments from the domain. It includes all the standard statistical tests, analysis models, as well as remarkable graphic capabilities and it offers a comprehensive language for data management and manipulation (Williams, 2011). Nevertheless, there are also some disadvantages, such as the difficulty to be learnt by non-specialists or the memory management problems, because R can very quickly occupy all the available memory.

3.2. Python

Python programming environment combines the data visualisation and analysis capacity, offered by R programming language with the development opportunity offered by Java [13]. This is why it is considered to be a complex and yet flexible instrument, which may support all the stages involved in designing an application.

Its flexibility also derives from the fact that it is a multi-paradigm language, which supports object-oriented, imperative, functional programming, and procedural styles. Since it is an interpreter and not a compiler, it has the advantage of being portable on different computing systems of the resulted programs. The basic implementation of Python, CPython is written in C and functions on Linux, Unix, Windows, Mac OS X.

Python functionalities range from working with files to working with processes, threads, sockets, serialization etc. It also includes a variety of libraries for developing graphical interfaces such as PyQT , PyGTK, wxWidgets. For numerical analysis and graphs there are the libraries NumPy, SciPy și Matplotlib.

Due to its accessibility and flexibility, Python is the most used language in order to learn programming, as shown by a study, which was performed by Philip Guo in 2014, for the top 39 universities from the USA [14].

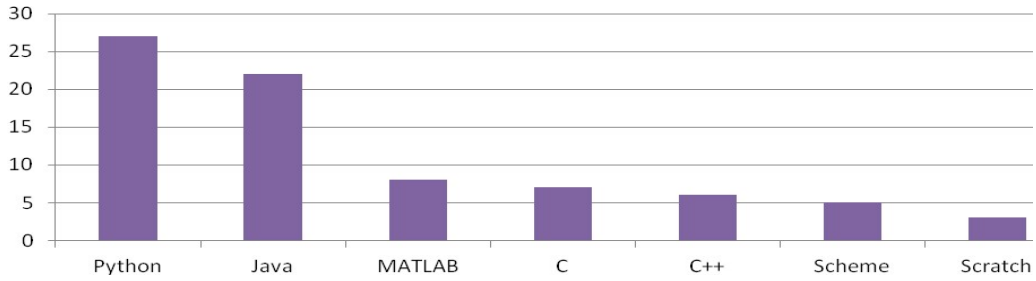

Fig.nr. 4. Languages for learning, data source: www.pgbovine.net

The syntax of the Python programming environment uses expressions and keywords, which are well-known to programmers, but what differentiates it from other languages is the signification of the indentation, because routines and code blocks are delimited by the indentation method.

The indentation, without using special characters, such as braces in C, allows an easier viewing of the code. Thus, the instructions that are at the same indentation level will form a block of code, the character ';' being optional, but it must be preceded by the symbol ':' . The unindented lines generally correspond, in the syntax of a Python program, to global variables, to definitions on classes, procedures, functions etc. This characteristic of the language makes us think about the Java-specific feature, where classes are delimited in different files.

Python has a robust set of libraries, especially machine learning libraries (scikit-learn, Theano, TensorFlow), making it the first choice of developers, who must include statistical analyses and models in their projects, or for data scientists, who must integrate their results in web applications or other development environments [13].

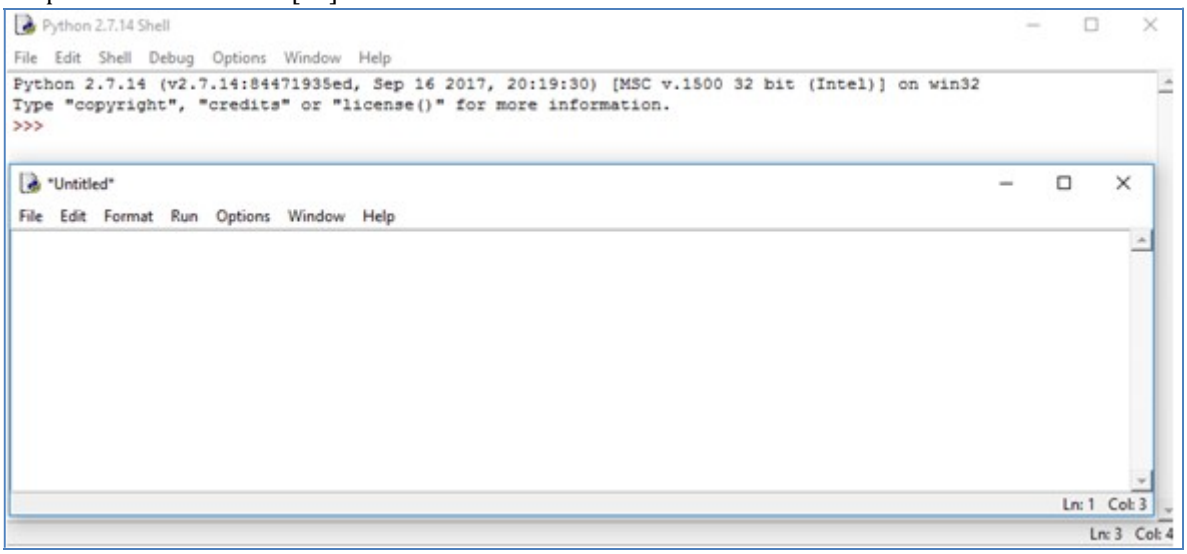

Fig.nr. 5. Python window

According to the survey performed by Stack Overflow in 2016, Python is the most popular language for data science, the data presented there regarding developers being illustrated in the following graph:

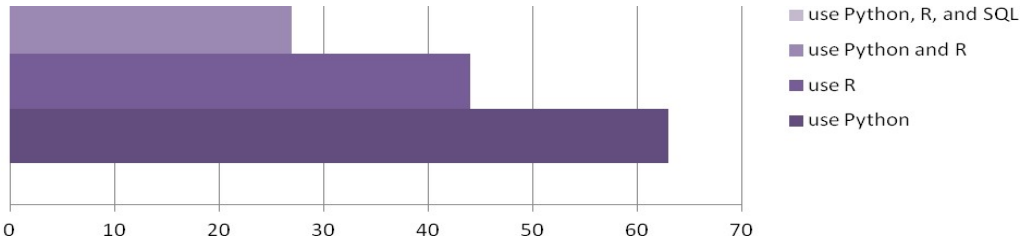

Fig.nr. 6. Languages popularity for data science,
data source: https://insights.stackoverflow.com/survey/2016#technology

Every year, CodeEval draws up a ranking of the "Most Popular Coding Languages". In 2016, they took into account 26 programming languages, and the study revealed the fact that, for the fifth consecutive year, Python occupied the first position in the ranking, although the total percentage decreased as compared to the previous year.

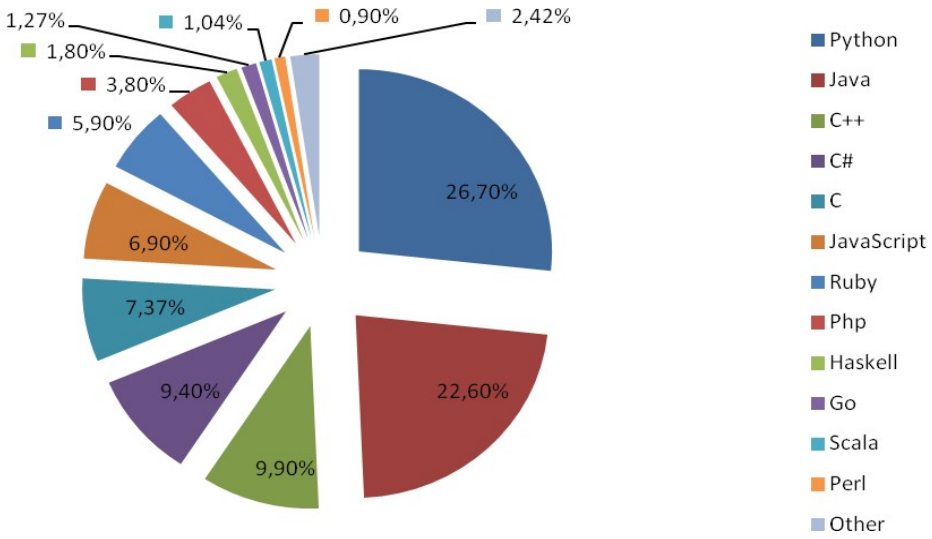

Fig.nr. 7. The use of development environments in 2016,
data source: http://blog.codeeval.com/codeevalblog/2016/2/2/most-popular-coding-languages-of-2016

A possible drawback of Python is that it doesn't offer compatibility with the previous versions; accordingly, programs written for one version can generate interpretation errors on another version. Guido van Rossum, the author of Python, states that Python 3 includes changes that makes it, on purpose, inconsistent with the 2.x versions.

5. Conclusions

Today's society faces an unprecedented problem, namely the abundance of information. Under these circumstances, it has become imperative to conduct research on the quality of the available information. This is the reason why data organisation, classification and synthesizing techniques are constantly developing. Starting from these needs, important communities of scientists and developers have been formed to help implement these techniques in more or less accessible tools for non-specialist users. The present article has attempted to review both the advantages and drawbacks of two important programming environments, R and Python. The choice between the two depends on the user's training and skills. It seems that R is the choice of researchers with a background in statistics and data analysis and less in programming, whereas

Python is used by the programmers who need to include data analysis and research methods in their complex projects.

References

1. *Brzustowicz M. (2017), Data Science with Java, Practical Methods for Scientists and Engineers, O'Reilly Media*
2. *Bugnion P. (2016), Scala for Data Science - Leverage the power of Scala with different tools to build scalable, robust data science applications, Packt Publishing*
3. *Florian V., Neagu G. (2016), Abordări și soluții specifice în managementul, guvernanța și analiza datelor de mari dimensiuni (Big Data) , Revista Română de Informatică și Automatică, vol. 26, https://rria.ici.ro/wp-content/uploads/2016/03/03-art-1.-Neagu_Florian.pdf*
4. *Granville, V. (2014), Developing Analytic Talent: Becoming a Data Scientist, Publisher: John Wiley & Sons, Inc.*
5. *Ihaka R. & Gentleman R.(1996), R: a language for data analysis and graphics, Journal of Computational and Graphical Statistics 5: 299–314*
6. *Lutz M. (2013) Learning Python, 5th Edition,Powerful Object-Oriented Programming, Publisher: O'Reilly Media*
7. *R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL:http://www.R-project.org.*
8. *Stanton J. (2012), AN INTRODUCTION TO Data Science, , Syracuse University (With A Contribution By Robert W. De Graaf) , file:///D:/FACULTATE/articole%202017/DataScienceBookV3.pdf*
9. *Williams G. (2011), Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery, Springer*
10. *http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf*
11. *https://www.cleveroad.com/blog/research-of-most-popular-programming-languages-for-2017*
12. *https://datascience.berkeley.edu/about/what-is-data-science/*
13. *https://www.upwork.com/hiring/data/15-python-libraries-data-science/*
14. *https://cacm.acm.org/blogs/blog-cacm/176450-python-is-now-the-most-popular-introductory-teaching-language-at-top-u-s-universities/*
15. *http://www.kdnuggets.com/2016/05/10-must-have-skills-data-scientist.html*
16. *https://docs.python.org/2/contents.html*
17. *https://learnpythonthehardway.org/book/*
18. *http://www.e-learn.ro/tutorial/python/introducere-in-python-partea-i/149/1/367.htm*
19. *https://www.upwork.com/hiring/data/15-python-libraries-data-science/*
20. *https://www.infoworld.com/article/2951779/application-development/in-data-science-the-r-language-is-swallowing-python.html*