Globant technical test

Jeison Mesa

16 abril, 2021

Introduction

In the workflow for data science professionals, it is important to achieve process automation. Automating processes allows to control human errors made by performing tasks manually. It is proposed to download files using web scraping. In this case, R is connected to the information available for each of the data sets in the web page. It is to recognise the structure of the information. It was possible to identify two types of data structure xls and csv files, for the set of csv files a totally different pattern was found to the xls format files, for this situation only the xls format files were keep. The next step is to have a logical file name. In this case, having files for different months, it is proposed to work with the following format "PCT_year_month" for all the files. This is because the structure presented on the website seemed rather "messy".

The additional step is done to identify the columns that are found for all the files. Resulting in the following variables:

Quality Control

The inferences that can be by a statistical model are largely influenced by the quality of the information, i.e. we must be able to guarantee that the information was verified and validated; for this purpose the following is generally checked:

1. **Problems associated with the scale of the variable**

Regardless of the type of software that we want to use, it is necessary that at the moment of loading the information those variables that by their typology are strictly numerical, the software can identify it because if we adjust a statistical model with a variable that is numerical and the software detects it as categorical we will reach erroneous conclusions (this is only an example of the many cases that can occur) or vice versa when giving numerical values to categorical variables the software detects them as numerical, in this case it is necessary to perform a transformation.