Fundamentals of the R programming language

Project assignment

26.12.2019

Project assignment

As part of the "Fundamentals of Programming Language R" course, you need to perform an exploratory analysis of a chosen dataset. The task consists of the following:

Student must choose a dataset from a collection of provided datasets and perform an exploratory analysis. The result of the analysis must be a PDF report which will:

- **describe the process of dataset preparation**
- **display a series of interesting visualizations**
- **provide insights and conclusions related to the chosen dataset**

Analysis should be performed in *RStudio.*

DEADLINE FOR THE PROJECT TASK OF PROJECT IS February 1st, 2020

The project report should be submitted to the *Moodle* system, similar to the submission of laboratory exercises and homework assignments. To create a project task, use the knowledge from the lessons and available RStudio cheat sheets. The analysis does not have to contain predictive models since this material wasn't taught in the lectures yet, but, if desired, students who have sufficient knowledge to implement this type of analysis can also include it in the report.

Students can do the assignment **in pairs** in which case they submit a single report and need to choose two datasets.

The offered datasets are as follows:

boardgamesa	data related to boardgames
HighestMountains	the highest mountains in the world
HR_StanPremaSpoluStarost	census of the population of the cities of the Republic of Croatia by sex and age
HR_ZG_RI_OS_ST_Stan	census of the population of the towns of the larger cities of the Republic of Croatia by sex and age
HR_zaposleni_neto_placa	net wages in individual business sector in the Republic of Croatia
HR_ZupanijeTurizam	tourist information
IGN_game_reviews	information about the reviews of computer games of the IGN portal
IMDB_movie_dataset	information about films from IMDB
SettlersOfCatan	statistics related to the *Settlers of Catan* boardgame
Simpsons_episodes	information about brodcasts of the TV show *The Simpsons*
speed-dating-experiment	result data of a *speed dating* experiment
SongOfIceAndFireDatasets	various datasets relating to the **Song of Ice and Fire** series
SuperMarioMakerDatabase	game level information relate to the *Super Mario Maker* game on the *Nintendo Wii U* console
ufo-sightings	documented UFO sightings from USA

The main sources of data sets are **Kaggle** (www.kaggle.com) and **Central Bureau of Statistics** (www.dsz.hr). It is possible that additional datasets will be added - any dataset found in the project assignment folder is a valid choice for the project assignment, even if it isn't present in the above list.

Below is a brief instruction on how to organize the data analysis process that is best adhered to in order to make the task as efficient as possible.

Organization of the data analysis process

Since data analysis is often a demanding and complex process, it is recommended to prepare a specific organizational infrastructure in advance to make it easier to manage the process, allow for easy detection and correction of errors, and provide support for simple repetition and adjustment of the already processed steps.

There are many recommendations on how to best organize such an analysis, as well as adequate software support in the form of additional packages, interfaces and technologies. The more complex the analysis, the more important the need for a higher level of organization is, especially if the analysis is carried out in a multi-user environment where it is necessary to coordinate with other members of the project team.

For the purposes of the project task, a simpler version of the organization of the analysis process is recommended, which is adapted to less demanding analytical environments, and does not require any additional software support except for the creation of a specific folder hierarchy and using certain conventions during the analysis process. The procedure is as follows:

1. For the needs of analyzing a particular data set, it is recommended to create a separate folder that will be used strictly for this analysis. This can be done through RStudio GUI by creating a new "project", or simply by creating an appropriate folder on the disk and using it as a working directory during the project.
2. In the root project folder, it is recommended to create subfolders with the following names:

a) **R** - this folder stores R scripts to be used in the project; most often these are various function definitions (for example, for cleaning and data processing) that are expected to be called multiple times and which are not to be a part of the report. The files in this folder are called *.R* and can be load them using the `source` function.

b) **data** - this map contains datasets - whether it's input data, or data that has undergone certain steps of cleaning and preparation. It is recommended to store data in *CSV* form, with clearly given names (if we have more iterations, we use numbers that clearly indicate the process sequence - for example, `01_initial_dataset.csv`,`02_clean_dataset.csv` . . .).

c) **figures** - a folder for all the interesting visualizations we created during the analysis. It is recommended to store them in *PDF* or *PNG* format. Visualizations must have enough data to be easily interpreted (and reconstructed) later. For visualizations that will be included in the final report, we should also preserve the program code that created them. If we want, in this folder we can also create subfolders **expl** and **report** in order to separate the visualizations which resulted from the exploratory analysis and those we have chosen to include in the final report.

d) **Rmd** - in this folder we should put the R Markdown reports. With more complex processes of data analysis, it is recommended to number the reports (`01-loading.Rmd`,`02-transformation.Rmd`, etc.). It is recommended to carefully document the entire analysis process so that it could be carried out at any time again (over the same data with the same validation results, or as a template for new data).

3. In addition to the above mentioned folders and related files, it is recommended to create two text files in the project's main project folder:

    a) `log.txt` - after each project session, we briefly enter what has been done and what were the results.

    b) `TODO.txt` - a brief description of the planned tasks for the next project session.

This structure of files and folders will enable us easier implementation and monitoring of the analysis process. If it proves insufficient, it is recommended to explore resources on the Web that will provide additional information on somewhat more complex organization models or specific software support that automates certain elements of the organization's analysis process and enables high-quality process management.

NOTE: You do not need to submit your entire project folder, just the final report. The above steps are simply recommendations to be used for this and any future dataset analysis you will perform using R and RStudio.