Background

One of the main reasons for the recent increase of R's popularity is the fast, efficient and intuitive way to work with datasets. Looking at the workflow of R data scientists, almost 50% of the work consists of data exploration tasks (featured at r-bloggers, http://www.wsj.com/articles/what-data-scientists-do-all-day-at-work-1457921541).

While the number of packages for various statistical tests increased over the years, R developers often follow their habits and use the very same statistical analysis techniques for every dataset. Here we propose a new package called **discovr** to increase the number of considered statistical techniques and to reduce the trial and error approach during the data exploration and analysis.

The package reflects the thoughts and input of multiple science departments of the Radboud University Nijmegen, on how to speed up the daily analysis workflow.

Related work

While yet no package exists that analyzes the viability of a variety of different statistical methods, there are packages that help with the model selection for specific techniques. While the expansion of existing packages was not an option, as it would completely shift the purpose of those packages, we will integrate a variety of existing packages. An example of those packages is glmulti, which screens and selects the best-suited glm models (different distributions and scoring).

Details of your coding project

Package Features and syntax:

```
discovr(x, data = dataframe, options = c(xyz))
```

As input data, discovr accepts data.frames and data.tables.

The first parameter (here x), represents one or multiple (with c(x,y)) explanatory variables. If no parameter is given here, it will perform the analysis the dataset independently of any explanatory variable, such as correlations and what the data distribution is (normal distribution etc.).

Output: The discovr package will have an interactive D3 heatmap (http://rpubs.com/jcheng/mtcars-heatmap) integrated, which will indicate a quality score for different statistical methods. It will thereby also indicate why a specific statistical method was rated as not viable:

As example, let's have a closer look at linear fittings: The 6 most often tested indicators for linear fits are:

- normal distributed data
- heteroscedasticity or homoscedasticity
- correlations between the dependend variables
- outliers and their effect (cooks distance)
- zero-inflation
- significance of the estimated fitting parameters

The all these tests, will then accompany for the quality score, and indicate whether a linear fitting is viable or not.
Additionally to the d3 heatmap, all those test scores and underlying tests can also be accessed via the command line, such as discovrVariable$LF (for just the linear fitting).
Timeplan: During the 3 Months, the first version of the **discovr** package will created and the CRAN documentation written. This first version will include a test for the 10 most used statistical techniques to analyse datasets. Those techniques will span frequental statistics (different correlation analysis, nls, glm) aswell as bayes statistics (random forrest).

- 1st month: Explore statistical backgrounds of the integrated statistical methods and working out a user friendly and efficient package input syntax.
- 2nd month: Integration of the statistical method tests.
- 3rd month: Integrating the interactive d3 heatmap and the package output and starting to write the CRAN package documentation

.

Expected impact

The discovr package will be of use for at least two R user groups:

The first group is the group of scientists and data analysts, starting to grasp R's statistical toolset. This group will be guided to an array of analysis techniques specificly usefull for their kind of dataset. Furthermore let them explore less common statistical analysis techniques.

The second user group consists of experianced R users that know almost all statistical techniques and their requirements and scoring parameters. Here, discovr offers the chance to check a variety of statistical methods for a given dataset in a matter of seconds, streamlining the exploratory/first analysis process. Moreover, it will help to check rather rarely viable statistical methods, which are otherwise, due to the additional time investment, often left out.

Mentors

Prof. Dr. R. Brock (email: Roland.Brock@radboudumc.nl). Prof. Brock is the head of the Department of Biochemistry at the RIMLS Nijmegen with years of experience in scientific analysis.

Drs. S. Schmidt (mail: Samuel.Schmidt@radboudumc.nl). Drs. Schmidt is an expert in the scientific data analysis in R.

Experience

- I created various shiny packages, which can upload and analyze confocal microscopy datasets, with various input parameters. The type of dataset and the specific analysis methods were determined based on numerical values of these datasets.
- I worked (big-data) observations in the area of ecology, and learned the various statistical methods applied in scientific publications.
- Finished all data camp courses yet available and learned many techniques presented by r-bloggers.