Data processing in R vs Traditional Programming Languages

While C++ and other traditional programming languages are excellent for software development, they offer poor solutions for data exploration. The lack of a GUI and a verbose syntax, limit the number of people that could perform data analysis and other statistical computing.

The development of R has somewhat mitigated this issue. While R maintains a strict syntax like many programming languages, all complex operations and implementation are abstracted, and users are able to focus on data exploration/processing. Aside from the simple graphical interface and low barrier to entry, R abstracts a lot of the internal code required to make all functions work and focusses on providing developers with a set of tools to manipulate their data of choice. Common statistical measures used in data exploration include:

- mean (the average of the values in a data set),
- median (the middle value/s of a data set after sorting), and
- range (the minimum and maximum values in a data set).

These values, among others, allow data scientists to recognize patterns in data and points out outliers. These figures help them make predictions as well as determine why the outlier are present.

In the context of machine learning, data sets provide us with a set of useful statistics when determining predictors. They are covariance and correlation.

- Covariance refers to the relationship between two variables in an observation. The greater the number, the greater the reliance on values has on the other. As an example, height and weight typically coincide with one another.
- Correlation measures the rate at which two random variables move in sequence. It tracks the changes in variables over time and determines whether their paths are similar or not.