



8 Essential Tools for Data Scientists

- Posted by [Joydeep Bhattacharya](#) on November 10, 2016 at 6:54am

There is growing demand of data scientists in every organization. For growth of any business enterprise there is need to evaluate data in order to streamline the strategies and to keep a record. The skills of a data scientist boils down to the tools that they are able to use and are aware of. In this article we will talk of both the programming data science tools as well as ones which do not require much of coding to get started with.

Here is list of 8 essential data science tools which you need to be aware of:

1- The R Project for Statistical Computing

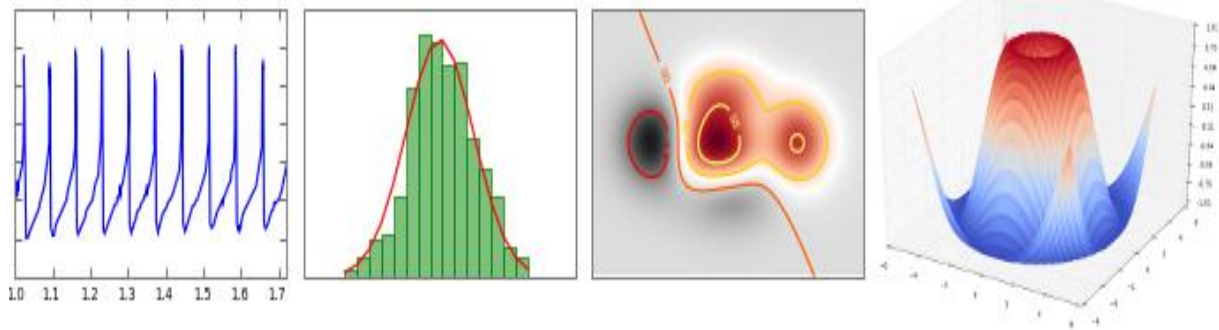
R is a perfect alternative to statistical packages such as SPSS, SAS, and Stata. It is compatible with Windows, Macintosh, UNIX, and Linux platforms and offers extensible, open source language and computing environment. The R environment provides with software facilities from data manipulation, calculation to graphical display.

A user can define new functions and manipulate R objects with the help of C code. As of now there are eight packages which a user can use to implement statistical techniques. In any case a wide range of modern statistics can be implemented with the help of CRAN family of Internet websites.

There are no license restrictions and anyone can offer code enhancements or provide with bug report.

2- Matplotlib

matplotlib is a python 2D plotting library with which statistical data representation is easy to intercept. It produces quality representation of data in the form of histograms, graph, bar charts, power spectra etc.



matplotlib can be used in python scripts, python, ipython shell and web application servers. It provides an interface similar to MATLAB for simple plotting. For advanced plotting features, user has full control over font properties, line styles with object oriented interface

and various set of functions which people familiar with MATLAB will find easy to get started with. Keep a track of latest developments through the source code or in what's new page section. You can mail the professionals about some bug, patches on github tracker.

3- RapidMiner

RapidMiner's open source data science platform provides assistance of rich data science to make calculative decisions. An all in one data science platform, which has pre-defined data preparation tools and machine learning algorithms to support all of data science needs. The visual data science workflow designer provided by RapidMiner has built-in templates and repeatable workflows.

For a data scientist, there are plenty of impressive features which are highlighted below:

- There are more than 60 file types and formats for structured and unstructured data.
- A user can access, load and extract valuable information from unstructured data. It supports plain texts, HTML, PDF, RTF, and many more.
- Graphical representation of data with more than 30 visualization options.
- The visual workflow designer comes with validation techniques such as preprocessing models, cross validation and split validation.

It is currently being used in array of industries such as automotive, banking, insurance, retail etc.

4- Logical Blue

"Data Science is not Rocket Science" is what this company believes in. It provides with user friendly software platform which lets businesses use data to automate the decision making and improve revenue generation. Irrespective of what business goals you have for predictive analysis the professionals staff is well skilled to provide assistance related to machine learning and statistics.

It uses latest statistics techniques and computational intelligence to provide with near perfect predictive models in the process makes it all easy for data scientists to contribute in the growth of the business. The data report is arranged in such as form that the data scientists can separate the less useful data from the most productive one.

The logical blue API allows for easy integration into tools and workflows such as underwriting process. The models which you build are hosted in cloud and can be used straight away without any complex process for deployment.

5- Bar Code Generator

Bar code scanning apps have great potential to find pertinent data from huge data cache. Especially for industries which belong to logistics or shipping. With an assigned value to a barcode with respective to products, you can find specific products without any trouble of manually searching through database. For data scientists who need to manage and keep track of inventory bar code generator is perfect remedy to find and locate product with quick scanning.

Gospaces provides with simple process to get started. Just assign a value you want the products to be classified in the database. No lengthy processes or wait, just generate and download unique bar code with a single click. Add the generated bar code to the products and it can be scanned with the assigned value.

6- Paxata

Paxata focuses more on data cleaning and preparation and not on machine learning or statistical modeling part. The application is easy to use and its visual guidance makes it easy for the users to bring together data, find and fix any missing or corrupt data to be resolved accordingly. The data can be shared and re-used with other teams. It is apt for people with limited programming knowledge to handle data science.

Here are the processes offered by Pixata:

- The Add Data tool obtains data from wide range of sources.
- Any gaps in the data can be identified during data exploration.
- User can cluster data in groups or make pivots on data.
- Multiple data sets can be easily combined into single AnswerSet with the help of SmartFusion technology solely offered by Paxata. With just a single it automatically finds out the best combination possible.

If you are looking for extensive data cleaning then Paxata is your apt companion.

7- Trifacta

Trifacta offers with free stand alone software and a licensed professional version. For data cleaning, the software takes in data as input and provides with summary with several statistics by column. For every column a recommendation is provided for transformations with a single click. There are many pre defined combinations which can be performed on the data for transformation.

For data preparation Trifacta uses:

- Discovering: the data is analyzed to get a quick sense of what you have.
- Structure: Assign proper shape and variable types to the data in order to solve anomalies if any.
- Cleaning: With functions such as imputation, text standardization, etc the data model is readied for the user.
- Enriching: the quality of analytics is improved by either adding data from other sources or by performing feature engineering on existing data.
- Validating: A final sense check is done to validate the data.
- Publishing: The data is exported for further use.

8- Narrative Science

What separates Narrative Science from its competitors is its ability to provide with automated reports based on the data. The advanced natural language processing tool works like a story telling tool to create reports on the lines of consulting reports.

The advanced patented NLG platform referred to as Quill is driven by Artificial Intelligence which creates data driven communications automatically. Quill automates the data driven report writing workflows and eliminates manual involvement.

Data scientists can generate personalized reports which are targeted to specific audience. It uses past data and also incorporate specific statistics to generate data reports. Narrative science is servicing diverse industry clienteles which belong to financial, insurance, government and e-commerce domains.

The above highlighted tools can be helpful to get hold of pertinent points in a data. These tools can help data scientists to get done with their work without much difficulty and reap productive results.

<https://www.datasciencecentral.com/profiles/blogs/8-essential-tools-for-data-scientists>