Using the Libre Excel Tools for Data Science

How to Let Forth Data Science with LET4DataSci

Author(s): Jacob J. Walker



**0**

***μ***

**±**

**Σ**

**1**

Last Updated: 2/7/2014



This work is licensed under a [Creative Commons Attribution-ShareAlike 3.0 Unported License](http://creativecommons.org/licenses/by-sa/3.0/).

The Wizard Girl logo is based upon material from the [Scratch Project](http://Scratch.MIT.edu), used as part of the above license.



Most of the Libre Excel Tools for Data Science are released under the [LGPL 3](http://www.gnu.org/copyleft/lesser.html)

# Contents

[Contents 2](#_Toc379432376)

[Introduction 4](#_Toc379432377)

[Using Excel for Data Science 4](#_Toc379432378)

[Using LET4DataSci 5](#_Toc379432379)

[LET4ETL Tools – Extracting, Transforming, and Loading Data 5](#_Toc379432380)

[Introduction to ETables 5](#_Toc379432381)

[Tools for Extracting Data from Different Data Sources 6](#_Toc379432382)

[Tools for Transforming Data 6](#_Toc379432383)

[Tools for Loading Data Elsewhere 6](#_Toc379432384)

[Lookup, Referencing, and Querying tools 6](#_Toc379432385)

[LET4Text Tools 7](#_Toc379432386)

[Tools for Parsing N-Grams from Text 7](#_Toc379432387)

[LET4Stats Tools 7](#_Toc379432388)

[Tool to Generate a Survey Size 7](#_Toc379432389)

[Tools for Probabilistic Logic 7](#_Toc379432390)

[LET4Privacy Tools 7](#_Toc379432391)

[Tools to Obfuscate People’s Names 7](#_Toc379432392)

[LET4Integration Tools 8](#_Toc379432393)

[UDFs to use VBA Functions & Methods in Formulas 8](#_Toc379432394)

[Contributing to the LET4DataSci Project 8](#_Toc379432395)

[Why You Should Contribute 8](#_Toc379432396)

[Design Goals of the Project 8](#_Toc379432397)

[Works Cited 9](#_Toc379432398)

[Appendices 10](#_Toc379432399)

[Appendix A – Recommended Learning Resources 10](#_Toc379432400)

[Resources to Learn More about Data Science with Excel 10](#_Toc379432401)

[Resources to Learn More about Data Science in General 10](#_Toc379432402)

[Resources to Learn More about Excel and VBA in General 10](#_Toc379432403)

[Other Resources worth Investigating 10](#_Toc379432404)

# Introduction

Libre Excel Tools for Data Science (LET4DataSsci) is a set of open source tools for Excel to help with various data science and analysis tasks. LET4DataSci includes libraries of user defined functions (UDFs), classes, and sub-routines that can be used in your formulas, macros, and VBA code. LET4DataSci also includes component “templates” that can be copied and pasted into your own spreadsheets. At this point, the number of tools LET4DataSci offers is limited to those I developed to handle some of the data science tasks I have faced so far. But, with this as a start, I hope others will join the project, and that the number of tools will expand over time to become a fuller set.

## Using Excel for Data Science

Surveys of attendees at several Strata data science conferences suggest that Excel ranks in the top 5 most commonly used tools for data science, with over 1/3 of data scientists using it for analysis [Magoulas and King 2014]. While Excel cannot handle really big data, it can handle “mid-size” data, now handling 1,048,576 rows. And even for projects that may ultimately need to be done with Hadoop, Excel can be used for prototyping [Foreman 2013].

Further, for those dipping their toes into data science, especially for someone who isn’t a programmer, Excel may be easier to start with. Also, Excel is very flexible and has a lot of analysis tasks it can do quickly and easily. It is also fairly extensible, which is what the LET4DATASCI does: adds tools to help with data science tasks.

Excel does have some drawback for data science work. First, as mentioned 1 million rows may seem like a lot, but can become limiting quickly. Next, Excel can be a bit slow on certain tasks. Excel is clearly proprietary and runs well only on Windows (although with emulation such as WINE, it can run on other systems like Linux). And last, Excel can have security challenges, because there is a lot of potentially private data in a single file. (To help with some of the security challenges, LET4DATASCI includes some privacy tools that can obfuscate data, to remove personally identifiable information (PII)).

But, most problematically, Excel is missing some key features. That is where LET4DataSci comes in. While it cannot solve all of Excel’s shortcomings (such as not having a built-in Median function in PivotTables), it does provide ways of working around these problems. It also provides other tools that Excel could not be expected to have built-in, but which none the less, are valuable to data science work, such as an obfuscation method for people’s names (although, granted, this tool is currently U.S. centric).

Anyone who uses Excel as their only tool for analysis should be careful of the “[Law of the Instrument](http://en.wikipedia.org/wiki/Law_of_the_instrument)”, which Abraham Maslow warned about when he wrote, "I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail." [Maslow 1966:15] Although, in many ways Excel is more of a Swiss Army knife® than a hammer, but it is still good for Excel users to expand their repertoire. As such, LET4DataSci also has integration tools to connect with other applications, programming languages, databases, web services, etc.

## Why Excel and Not LibreOffice or OpenOffice Calc?

It may seem strange to some folks that a set of open source tools are being released for a highly proprietary program like Microsoft Excel, especially when there is LibreOffice (and its sister Apache OpenOffice). There are several reasons, all fairly straightforward:

1. LibreOffice and OpenOffice do not yet have as rich set of tools available as Microsoft Excel. Specifically, while Excel 2007 introduced having tables within a spreadsheet, Calc has not caught up, and from the few comments I can find in forums, does not plan to do so.
2. While Excel is clearly proprietary, it is also relatively cheap (compared to many statistical packages), especially with an academic discount. So the barrier to entry is relatively low.
3. There is an interesting paradox about Excel. Microsoft is heavily proprietary, but the community of Excel nerds that has grown over time are generally extremely non-proprietary in sharing their knowledge and VBA code on forums and websites.
4. To be quite honest, I have devoted a lot of time to learning Excel in my life, and while I’m trying to expand my data science toolset to other platforms, with a focus on helping with FLOSS (Free/Libre/Open Source Software) projects, I don’t want my proprietary skills to go to waste.   
   I suspect I am not alone in being in this situation.

So, while I truly hope much of what is being created in LET4DataSci is ported to LibreOffice & OpenOffice, for now it is an open source project for Excel.

Also, some may wonder why the word “Libre” is used, if there is no connection to LibreOffice.   
The answer is even simpler: Libre means freedom, and that is the point of LET4DataSci.

# Using LET4DataSci

LET4DataSci has been mostly modularized so that if you only want one type of tool, you only need to put that one module into your spreadsheet. The licensing of this project is also modularized, so that if you are working with proprietary data, it is not required to become open source or open content. But, with that being said, I encourage you to use data science to help humanity, not just to make a profit, and one of the best ways you can do this is to release your work under an open license, like Creative Commons.

While this documentation attempts to explain how to use the various tools in LET4DataSci, it is currently geared towards someone who already is fairly proficient in Excel. In the future, I hope that this documentation will expand to become both a tutorial on conducting data science tasks with Excel, as well as a reference of the LET4DataSci tools. But how soon that can happen depends a lot upon who else becomes involved in this project, as there are currently a lot of tools that are planned to be in the project, but only a few currently implemented. For now, there is a Recommended Learning Resources list in the appendices that can help the beginner get a handle on prerequisite skills.

So without further ado, let’s start looking at the various parts of LET4DataSci, starting with LET4ETL.

## LET4ETL Tools – Extracting, Transforming, and Loading Data

The LET4ETL tools are the foundation of the LET4DataSci project, as generally the first step of data science projects is to *extract* data from somewhere, often needing to *transform* it along the way (aka data wrangling or munging), and if you are using Excel as semi-ad hoc middleware, then in the end you may wish to *load* the data someplace else. To do these tasks, there is also a need to do tasks that involves figuring out where stuff is, so the LET4ETL Tools also include a robust set of lookup, referencing, and querying tools.

While, it would be nice to be able to do ETL and lookup tasks well with only built-in functions, objects, and subroutines; unfortunately, Excel’s built-in functions and objects have limitations that often stops the data scientist from doing what they need to do. For example, Web Queries using QueryTables objects cannot easily be used to extract data from web pages that require form entries (such as password protected web sites). Normally someone would need to write code from scratch to get around this problem, and each solution would be non-standardized and difficult to reuse. That is why LET4ETL has created new objects, functions, and subroutines that are reusable for different types of projects. The core of these is a new class of objects called *LETables*.

### Introduction to LETables

A Libre Excel Table (LETable, which I like to pronounce as Lé Table) is an expansion of a standard Excel table (ListObject) that has added functionality, which can help with the following:

* Web scraping (better than a QueryTable)
* Running SQL Queries on them (by integrating with Access)
* Summarizing data (in some cases better than a PivotTable)
* Mail merging with Microsoft Word (better than a standard Excel table)
* “Mirroring” a PivotTable (or normal table)

LETables add much of the functionality that Excel *should* have already, but that it either doesn’t have, or functionality that is only available in newer versions of Excel. (Although to be clear, Excel 2007 or newer is generally required for most of LET4DataSci.) When using the LET4DataSci tools, any time a table is created, it will create a LETable by default.

### Tools for Extracting Data from Different Data Sources

#### Web Scraping with LETables

One of the frustrations that many of us who work with Excel have, is that many of the built-in features of Excel have some critical features missing that stop them from being used in many important situations. This has clearly been the case with Web Queries, which use the QueryTable object. A Web Queries is missing the key component that there is no ability to have transactions with a website, and this causes major problems if a web page you want to scrape requires a password to be entered first.

This is where the LETable object comes in, because Excel thankfully has VBA, we have been able to create an object that has most of the features of a QueryTable, and also store transactional data.

#### “Mirroring” a PivotTable or standard Excel Table with a LETable

### Tools for Transforming Data

In data science tasks, there is often a need for data wrangling (also called data munging), which is generally transforming the data from one form, to being in a form that can more easily be analyzed.

#### Summarizing Data with LETables (similar to a PivotTable)

### Tools for Loading Data Elsewhere

#### Using LETables with Microsoft Word’s Mail Merge

### Lookup, Referencing, and Querying tools

While technically these types of functions are not ETL, they are required for a lot of ETL work, and many of the other LET4ETL tools use them, so it was decided to have these tools within the LET4ETL module.

#### Using SQL Queries on LETables

## LET4Text Tools

### Tools for Parsing N-Grams from Text

## LET4Stats Tools

### Tool to Generate a Survey Size

### Tools for Probabilistic Logic

## LET4Privacy Tools

For ethical and legal reasons, when an organization or researchers wish to release data to the public, it is critical that the data does not contain personally identifiable information (PII).

1. The legalities of database release are still murky, and it is wise to research these thoroughly and possibly check with a legal expert. This tool is released without warranty, and it is the responsibility of users of this tool to determine whether they are acting legally or not.

2. There are two competing ethical issues that drive the legalities and ethic:

\* The right of the general public to have full disclosure of government data, and the ability of that data to improve humanity

\* The right of individuals in the data to have privacy

3. It is nearly impossible to fully make any data non-personally identifiable without obfuscating so much information that the data is meaningless for research purposes. But, generally the data can and should be obfuscated sufficiently that it would take a lot of additional research on the part of the would-be privacy thief to use it any meaningful way.

4. Any obfuscation of personally identifiable information reduces the validity of research done with that data. For instance, if all names in a database are changed to random names using a random name generator, then no valid research can be conducted upon the name portion of the database. For instance, surnames are a weak proxy for paternal country of origin, and a full randomization would eliminate any ability to analyze this proxy in any meaningful way. (Note: Surnames are a very unreliable proxy for African Americans whose ancestry included American slaves, due to many slaves taking the surname of their former masters.)

### Tools to Obfuscate People’s Names

But, often the data does not make as much sense if there is not at least a placeholder for personally identifiable information, such as a name. The following Excel spreadsheet can help overcome this problem, by using public data from the U.S. Census Bureau and the Social Security Administration, random names can be generated.

## LET4Integration Tools

By necessity the use of any integration tool requires having other software. And sometimes it is necessary to have additional middleware (generally in the form of 3rd party add-ons) for Excel to handle this integration. But, the goal of LET4DataSci is to require the least amount of 3rd party middleware possible, and where 3rd party software is required, to use open source software as much as possible.

### Python Integration

### R Integration

### Java Integration

## UDFs to use VBA Functions & Methods in Formulas

One of the frustrating things about Excel, is that there are many functions and methods on objects that can be used in VBA that are not available in Excel as a function to use in a spreadsheet. While these are not necessarily specific to data science, they often can be useful for data science tasks, along with other types of tasks.

# Contributing to the LET4DataSci Project

## Why You Should Contribute

Contributing to the LET4DataSci project is a “win-win” all the way around.

## Design Goals of the Project

## How to Contribute to the Project

# Works Cited

Foreman, J.W. 2013. *Data Smart: Using Data Science to Transform Information into Insight*. Wiley.

Magoulas, R. and King, J. 2014. *2013 Data Science Salary Survey: Tools, Trends, What Pays (and What Doesn’t) for Data Professionals*. O’Reilly.

Maslow, A.H. 1966. *The Psychology of Science: A Reconnaissance*. Maurice Bassett.

# Appendices

## Appendix A – Recommended Learning Resources

The following recommended learning resources represents a curriculum that an autodidact could use to start from being a novice to becoming proficient in using Excel for data science, assuming you have subject matter expertise within a discipline already (or you know how to gain such expertise).

For anyone who is just starting with data science, it is helpful to start with learning some of the very basic concepts and skills that will be needed to do any substantive work in the future. One commonly cited definition of “data science” comes from Drew Conway’s [Data Science Venn Diagram](https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html), which has been more formalized in the [Data Science Venn Diagram v2.0](http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html) by Steven Geringer (now adapted by me):

### Introduction to Excel

#### Head First Excel: A learner's guide to spreadsheets

Available at: [Safari Books Online](http://my.safaribooksonline.com/9780596807719), [Google Books](http://books.google.com/books?id=uWbUKkpboH0C), Amazon, etc.

### Introduction to VBA Programming in Excel

#### Excel VBA Programming For Dummies

Available at: [Safari Books Online](http://my.safaribooksonline.com/9781118490389), Google Books, Amazon, etc.

#### Excel Programming with VBA Starter

Available at: [Safari Books Online](http://my.safaribooksonline.com/book/programming/vba/9781849688444), Google Books, Amazon, etc.

#### Excel® 2013 VBA and Macros

Available at: [Safari Books Online](http://proquestcombo.safaribooksonline.com/9780133259483), Google Books, Amazon, etc.

#### Programming Excel with VBA and .NET

Available at: [Safari Books Online](http://my.safaribooksonline.com/book/office-and-productivity-applications/0596007663), [Google Books](http://books.google.com/books?id=9YyH5L2y5ggC), Amazon, etc.

### Introduction to Probability and Statistics

Assuming a person has the skills and knowledge of traditional algebra, they should be able to start to learn about probability and statistics, which has become important to every scientific discipline, and is generally the most used form of mathematics in data science.

#### Probability and Statistics, from the Khan Academy

<http://www.khanacademy.org/math/probability>

#### The Manga Guide to Statistics

Available at: [Safari Books Online](http://my.safaribooksonline.com/book/statistics/9781593271893), [Google Books](http://books.google.com/books?isbn=1439587965), [Amazon](http://www.amazon.com/Manga-Guide-Statistics-Shin-Takahashi/dp/1593271891), etc.

#### Statistical Analysis with Excel for Dummies

Available at: [Safari Books Online](http://my.safaribooksonline.com/book/statistics/9781118464328), Google Books, Amazon, etc.

### Introduction to Linear Algebra

#### The Manga Guide to Linear Algebra

Available at: Safari Books Online, Google Books, Amazon, etc.

### Introduction to Data Science (With or Without Excel)

#### Head First Data Analysis: A Learner's Guide to Big Numbers, Statistics, and Good Decisions

Available at: [Safari Books Online](http://my.safaribooksonline.com/9780596806224), [Google Books](http://books.google.com/books?isbn=0596153937), Amazon, etc.

#### Data Smart: Using Data Science to Transform Information into Insight

Available at: [Safari Books Online](http://my.safaribooksonline.com/book/databases/9781118661468), Google Books, Amazon, etc.

### Conducting Data Science with Excel

#### Statistical Analysis, Decision Analytics, Predictive Analytics: Microsoft® Excel

Predictive Analytics: Microsoft® Excel available at: [Safari Books Online](http://my.safaribooksonline.com/book/office-and-productivity-applications/9780132967266), [Google Books](http://books.google.com/books?isbn=0132967251), Amazon, etc.

Statistical Analysis: Microsoft® Excel 2010 available at: [Safari Books Online](http://my.safaribooksonline.com/book/office-and-productivity-applications/9780132681872), Google Books, Amazon, etc.

Decision Analytics: Microsoft® Excel® available at: [Safari Books Online](http://my.safaribooksonline.com/book/databases-and-reporting-tools/9780133490589), Google Books, Amazon, etc.

These three books represent the trilogy of how to use Excel for foundational data science tasks.

#### Microsoft Excel 2010: Data Analysis and Business Modeling

Available at: Safari Books Online, Google Books, Amazon, etc.

### Working with Databases

#### Data Analysis Using SQL and Excel

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Access Data Analysis Cookbook

Available at: Safari Books Online, Google Books, Amazon, etc.

### Expanding to Other Programming Languages: Python & R

#### Head First Programming: A Learner's Guide to Programming Using the Python Language

Available at: [Safari Books Online](http://my.safaribooksonline.com/9780596806682), Google Books, Amazon, etc.

### Advanced Data Science Topics

#### Mining Text Data

Available at: [Google Books](http://books.google.com/books?id=vFHOx8wfSU0C), [Amazon](http://www.amazon.com/Mining-Text-Data-Charu-Aggarwal-ebook/dp/B00BW1AR72), etc.

#### Excel Data Analysis: Your visual blueprint for creating and analyzing data, charts, and PivotTables

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Data Analysis Microsoft with Access 2010: From Simple Queries to Business Intelligence

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Data Mining Applications for Empowering Knowledge Societies

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Social and Political Implications of Data Mining: Knowledge Management in E-Government

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Data Mining For Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Visual Intelligence: Microsoft Tools and Techniques for Visualizing Data

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Business Intelligence for Dummies

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Even You Can Learn Statistics: A Guide for Everyone Who Has Ever Been Afraid of Statistics

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Statistics Hacks

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Think Stats

Available at: Safari Books Online, Google Books, Amazon, etc.

### Other Resources worth Investigating

#### Gamification by Design

Available at: Safari Books Online, Google Books, Amazon, etc.

#### Who Is Fourier? A Mathematical Adventure

Available at: Safari Books Online, Google Books, Amazon, etc.

## Appendix B – Details about Licensing