Module 1

Data Preprocessing: From Raw Data to Ready to Analyse

Dr. Anil Dolgun

Last updated: 09 April, 2018

Overview

Summary

Module 1 will set the background for the entire course. Data preprocessing will be defined and the importance of data preprocessing inside the data analysis workflow will be explored. The module will define 5 major tasks for data preprocessing and will provide a quick overview of these tasks. Then, in the following modules of this course, we will unwrap each data preprocessing task by providing details of operations related to that task. I will also introduce you to R in this module, discuss the benefits it provides, and start to get you comfortable with R by giving R and RStudio basics. Recommended reading and useful online resources to get started with R will be given at the end of this module.

Learning Objectives

The learning objectives of this module are as follows:

- Define data preprocessing
- Identify steps for data analysis, understand the place of data preprocessing inside the data analysis workflow
- Understand the reasons why data preprocessing is important
- Identify major tasks in data preprocessing
- Understand main benefits of using R statistical programming language in data preprocessing
- Learn how to install R and RStudio, know the overview of the RStudio interface.
- Know how to install and load packages
- Learn mathematical/logical operations and basic programming in R
- Know how to get further help for R statistical programming language

Data

"Data! Data! Data! I can't make bricks without clay!" - Sir Arthur Conan Doyle.

Data. Our world has turned out to be increasingly dependent upon this resource. Sir Conan Doyle's famous detective, Sherlock Holmes, wouldn't shape any theories or draw any conclusions unless he had adequate data. Data is the fundamental building piece of all that we do in analytics such as the analyses we perform, the reports we build, and the decisions we made.

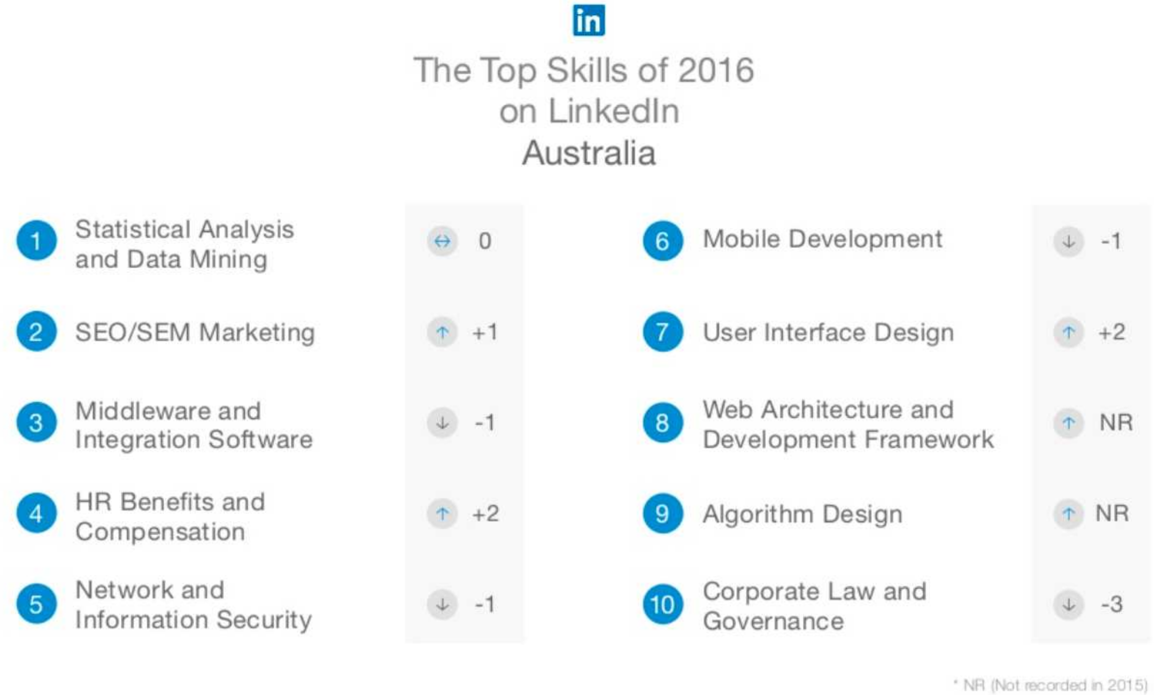
"In God we trust. All others must bring data." – W. Edwards Deming.

This quote by W. Edwards Deming (https://en.wikipedia.org/wiki/W._Edwards_Deming) (statistician, professor, author, lecturer, and consultant), emphasises the significance of data-driven decisions. Obviously, everyone has better bring (also understand and interpret) data to back up their claims.

The rise of the Data Analyst

Today's organisations have access to more data than ever before, but more data isn't better data unless you know what to do with it. Organisations are struggling to find people who can turn their data into insights and value, which in turn has created a high demand across the world for data analysts. During an interview in 2009, Google's Chief Economist Dr. Hal R. Varian stated, ***"The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades."*** Read the full article here (https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/#2a2b655c52ad)

Moreover, LinkedIn recently reported that the “data analysis” is one of the hottest skill categories over the past two years for recruiters, and it was the only category that consistently ranked in the top 4 across all of the countries they analysed. Read the full article here (https://blog.linkedin.com/2016/10/20/top-skills-2016-week-of-learning-linkedin). Below are the top ten skills of 2016 reported on LinkedIn for Australia.



in

The Top Skills of 2016 on LinkedIn Australia

1	Statistical Analysis and Data Mining	↔ 0	6	Mobile Development	↓ -1
2	SEO/SEM Marketing	↑ +1	7	User Interface Design	↑ +2
3	Middleware and Integration Software	↓ -1	8	Web Architecture and Development Framework	↑ NR
4	HR Benefits and Compensation	↑ +2	9	Algorithm Design	↑ NR
5	Network and Information Security	↓ -1	10	Corporate Law and Governance	↓ -3

* NR (Not recorded in 2015)

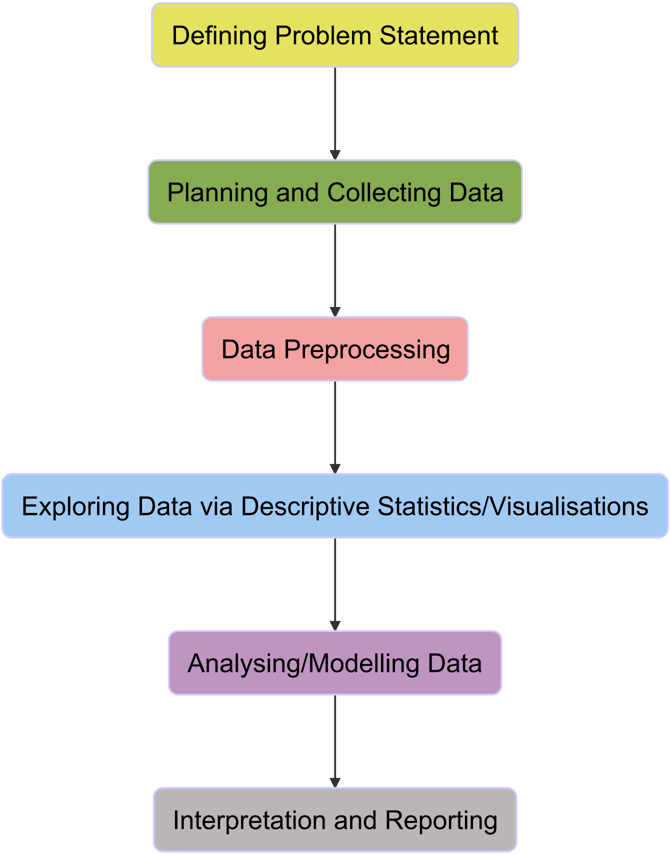
According to a recent IBM report (Read the full article here (https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/IML14576USEN.PDF)):

- 59% of all data science and analytics job demand is in Finance and Insurance, Professional Services, and IT.
- By 2020, the number of data science and analytics job listings is projected to grow by nearly 364,000 listings to approximately 2,720,000.
- 39% of data scientists and advanced analyst positions require a Master’s or Ph.D. in the related area.

Consequently, it is safe to say that skilled data analysts are now crucial to all industries. There is a need for becoming fluent in the data analysis process and staying up-to-date by continually learning and adding new knowledge in this field. I am assuming that’s the reason why you are here!

Data Analysis Steps

The statistical approach to data analysis is much broader than just analysing data. Data analysis process starts with defining problem statement, continues with planning and collecting data, preprocessing data, exploring data using descriptive statistics and data visualisations, analysing/modelling data, and finalises with interpreting and reporting findings. This process is depicted in the following illustration.


Defining Problem Statement
Planning and Collecting Data
Data Preprocessing
Exploring Data via Descriptive Statistics/Visualisations
Analysing/Modelling Data
Interpretation and Reporting

- ***Defining Problem Statement***: This is the first step of data analysis. In this step, the problem statement is identified by the organisations/researchers. The data analyst should thoroughly understand the problem and the domain of the problem.
- ***Planning and Collecting Data***: In this step, the appropriate tools for data collection related to the problem statement will be identified. This step may include designing a survey for data collection, scraping data from web or accessing an Excel/a database file.
- ***Data Preprocessing***: The objective of this step is to make the data ready for the further statistical analysis. This step is considered to be one of the important phases in data analysis. The accuracy of the statistical analysis depends on the quality of the data gained in this step. A number of operations such as importing data, reshaping data from long to wide format, filtering data, cleaning data, identifying outliers, and transforming variables can be applied to the data to make ready for the statistical analysis.
- ***Exploring Data via Descriptive Statistics/Visualisations***: The objective of this step is to understand the main characteristics of the data. Exploratory analyses are generally done using descriptive statistics (i.e. mean, median, standard deviation, frequencies, percentages etc.) and visualisation tools (i.e. scatter plots, box-plots, histograms, interactive data visualisations etc.). Exploratory analysis will show you the things that you didn’t expect, or raise new questions about the data.
- ***Analysing/Modelling Data***: The statistical analysis/modelling step can include a broad range of techniques like statistical hypothesis testing, statistical modelling, and machine learning algorithms. Generally, the type of the variables in the data set and the purpose of the investigation will determine the appropriate analysis technique.
- ***Interpretation and Reporting***: The last step of the data analysis is the reporting and the interpretation of the results. This step is also critical as if you cannot understand and communicate your results to others, it doesn’t matter how well you conducted your

analysis.

What is Data Preprocessing?

Most statistical theory concentrates on data modelling, prediction, and statistical inference while it is usually assumed that data are in the correct state for the data analysis. However, in practice, a data analyst spends most of his/her time (usually 50%-80% of an analyst time) on making ready the data before doing any statistical operation (Dasu and Johnson (2003)). Despite the amount of time it takes, there has been surprisingly very little emphasis on how to preprocess data well (Wickham and others (2014)). Real-world data are commonly incomplete, noisy, inconsistent, and don't have all the correct labels and codes that are required for the analysis. ***Data Preprocessing, which is also commonly referred to as data wrangling, data manipulation, data cleaning, etc., is a process and the collection of operations needed to prepare all forms of untidy data (incomplete, noisy and inconsistent data) for statistical analysis***.

Why Data Preprocessing is important

Data preprocessing may significantly influence the statistical conclusions based on the data. "**Garbage in, garbage out (GIGO)**" is a famous saying that is used to emphasise "***the quality of the statistical analyses (output) always depends on the quality of the data (input)***". By preprocessing data, we can minimise the garbage that gets into our analysis so that we can minimise the amount of garbage that our analyses/models result.

Why do you learn Data Preprocessing?

The road to becoming an expert in data analysis can be challenging and in fact, obtaining expertise in the broad range of data analysis is a career-long process. In this course, you will take a step closer to fluency in the early stages; namely in the data preprocessing step, as you need to be able to import, manage, manipulate and transform your data before performing any kind of data analysis.

Major Tasks in Data Preprocessing

We will define 5 major tasks for data preprocessing framework, namely : ***Get***, ***Understand***, ***Tidy & Manipulate***, ***Scan*** and ***Transform***.

A typical data preprocessing process usually (but not necessarily) follows the following order of tasks given below:

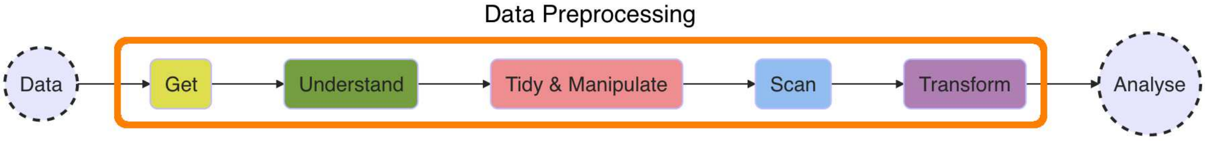


Fig1. Major tasks in data preprocessing

Get: A data set can be stored in a computer or can be online in different file formats. We need to get the data set into R by importing it from other data sources (i.e., .txt, .xls, .csv files and databases) or scraping from web. R provides many useful commands to import (and also export) data sets in different file formats.

Understand: We cannot perform any type of data preprocessing without understanding what we have in hand. In this step, we will check the data volume (i.e., the dimensions of the data) and structure, understand the variables/attributes in the data set, and understand the meaning of each level/value for the variables.

Tidy & Manipulate: In this step, we will apply several important tasks to tidy up messy data sets. We will follow Hadley Wickham's "Tidy Data" principles (Wickham and others (2014)):

1. Each variable should form a column.
2. Each observation should form a row.
3. Each type of observational unit should form a table.

We may also need to manipulate, i.e. filter, arrange, select, subset/split data, or generate new variables from the data set.

Scan: This step will include checking for plausibility of values, cleaning data for obvious errors, identifying and handling outliers, and dealing with missing values.

Transform: Some statistical analysis methods are sensitive to the scale of the variables and it may be necessary to apply transformations before using them. In this step we will introduce well-known data transformations, data scaling, centering, standardising and normalising methods.

There are also other steps related with preprocessing special types of data including dates, time and characters/strings. The last module of this course will introduce the special operations used for date, time and text preprocessing.

Why you learn this subject using R?

In this course, you won't learn anything about Excel, SPSS, SQL, SAS, Python, Julia, or any other statistical package/programming language useful for data preprocessing. This isn't because I think that these tools are bad or redundant. They're not. In practice, most data analytics teams use a mixture of these tools and programming languages. I strongly believe that R is a great place to start your data analysis journey as it is a comprehensive language for data analysis. You can use R effectively in almost each step of data analysis, from data collection to reporting. You can collect, preprocess, visualise and analyse your data using R functions, report and publish your findings using RMarkdown.

Since any typical data preprocessing actions like missing value imputation or outlier handling obviously influence the results of statistical analyses, data preprocessing should be viewed as a statistical operation and should be performed in a reproducible way. The R software provides us a good environment for reproducible data preprocessing as all actions can be scripted and therefore reproduced. Moreover, through the Master of Analytics and Master of Statistics & Operations Research programs, you will learn many data analysis subjects (i.e., Introduction to Statistics, Data Visualisation, Machine Learning, Analysis of Categorical Data, Time Series Analysis, Forecasting and Applied Bayesian Analysis courses) using R environment. So it is better to use the same tool for data preprocessing as well.

Introduction to R and RStudio IDE

There are many reasons why R is a good solution for the problems that are covered in this course. According to Munzert et al. (2014), the most important points are:

- R is freely and easily accessible. You can download, install, and use it wherever and whenever you want.
- For a software environment with a primarily statistical focus, R has a large community that continues to flourish. R is used by various disciplines, such as social scientists, medical scientists, psychologists, biologists, geographers, linguists, and also in business.
- R is open source. This means that you can easily retrace how functions work and modify them with little effort.
- R is reasonably fast in ordinary tasks.
- R is powerful in creating data visualizations. Although this not an obvious plus for data collection, you would not want to miss R's graphics facilities in your daily workflow.
- Work in R is mainly command line based. This might sound like a disadvantage to R rookies, but it is the only way to allow for the production of reproducible results compared to point-and-click programs.
- R is not picky about operating systems. It can generally be run under Windows, Mac OS, and Linux.
- Finally, R is the entire package from start to finish.

The following videos will cover the installation of R, RStudio and R packages, and introduce you the RStudio's basic features.

Installing R

This video will introduce you how to download and install R for different operating systems.

Installing RStudio

This video will introduce you how to download and install RStudio for different operating systems.

00:52

R Studio Overview

This video will give you a quick overview of the RStudio IDE.

02:34

Installing Packages

This video will introduce you how to install packages in R.

R Programming Basics

We will also use Dr. James Baglin’s (https://www.linkedin.com/in/james-baglin-ab3764114/) R Bootcamp notes to introduce you to the basics of R and RStudio environment. Please complete the following two modules in order to get started with learning R. Do not assume you will be an R guru after completing all of them. These are just beginnings. R has a slow learning curve and you will get heaps of practice during this course.

- R Bootcamp Overview (https://astral-theory-157510.appspot.com/secured/RBootcamp_Overview.html)
- R Bootcamp Course 1 - Getting Started (https://astral-theory-157510.appspot.com/secured/RBootcamp_Course_01.html)

After finishing R Bootcamp Overview (https://astral-theory-157510.appspot.com/secured/RBootcamp_Overview.html) and R Bootcamp Course 1 - Getting Started (https://astral-theory-157510.appspot.com/secured/RBootcamp_Course_01.html) modules, you will be able to:

- Install R and RStudio
- Access RStudio from MyDesktop
- Have an overview of the RStudio interface
- Install and load packages
- Learn mathematical/logical operations
- Learn basic programming in R

Additional Resources and Further Help in R

There are lots of amazing things that you can do with R and RStudio. Here are my favourite links to help you learn more.

- R for Data Science (http://r4ds.had.co.nz/): The online (and free!) book “R for Data Science” by Garrett Grolemund and Hadley Wickham, published by O’Reilly, January 2017. This is also one of the recommended textbook for our course.

- R Programming for Data Science (https://bookdown.org/rdpeng/rprogdatascience/): The online (and free!) book “R Programming for Data Science” by Roger D. Peng.

- R Cheatsheets (https://www.rstudio.com/resources/cheatsheets/): RStudio cheat sheets make it easy to learn about and use some of the favourite packages.
- Rseek (http://rseek.org): A custom search engine for R related resources created and maintained by Sasha Goodman.
- R-bloggers (https://www.r-bloggers.com): R news and tutorials contributed by (750) R bloggers.
- swirl (http://swirlstats.com/): Learn R, in R. Swirl teaches you R programming and data science interactively, at your own pace, and right in the R console!
- DataCamp (https://www.datacamp.com/): Free hands-on courses on R programming.
- Quick-R web page (http://www.statmethods.net): A comprehensive web site including R tutorials on basic and advanced statistics.
- R Help pages: R has a very powerful but complicated help file achieve. If you don't know how to use a function, you can search it by typing `help()` (e.g., `help(mean)`) in the console and this will display the help files (e.g., associated with `mean()` function).
- RStudio Community (https://community.rstudio.com/): This is a warm and welcoming place to ask any questions you might have about the R and RStudio.
- Google and Stack Exchange (https://stackexchange.com/): Sometimes what you really need to see are some examples. You can ask your question to Google or Stack Exchange. Just add "**with R**" at the end of any search. There is a huge community there willing to help when you have issues.

References

Dasu, Tamraparni, and Theodore Johnson. 2003. *Exploratory Data Mining and Data Cleaning*. Vol. 479. John Wiley & Sons.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons.

Wickham, Hadley, and others. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). Foundation for Open Access Statistics: 1–23.