Dominik Jung

The Modern Business Data Analyst

A Case Study Introduction into Business Data Analytics with CRISP-DM and R

Springer

The Modern Business Data Analyst

Dominik Jung

The Modern Business Data Analyst

A Case Study Introduction into Business Data Analytics with CRISP-DM and R

Springer

Dominik Jung
Software & Digital Business Group
TU Darmstadt, Darmstadt, Germany

ISBN 978-3-031-59906-4 ISBN 978-3-031-59907-1 (eBook)
https://doi.org/10.1007/978-3-031-59907-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Felix and Alexander

Preface

Solving Business Problems with Business Data Analytics

“Sorry, but I cannot tell fortunes!”, is my disappointing answer if people hear that I am an expert for *business data analytics*. They assume that my expertise in this field equates to crystal ball-like abilities or an uncanny knack for foreseeing. Some people ask me about my predictions on general topics like stock market trends, the next vote results, or specific things like the development of an obscure cryptocurrency or the future soccer results of their favorite hometown team.

Business data analytics isn't about divination or fortune telling. It's a methodical process, a toolset honed to derive insights from business data and make informed decisions. It's not a crystal ball. It's a sophisticated compass, guiding businesses through the maze of information to better understand the present and make calculated steps toward the future. Business data analytics isn't about magic - it's about leveraging information to drive strategic action.

Let me illustrate it with a short story of my life: During my time as a research assistant at the university, there was the 2018 World Cup in Russia. Because soccer is a big thing in Germany, the whole country and thus also our institute was in soccer fever. Some colleagues even took leave to watch certain games (and enjoy the celebration afterwards). Even during work, we often had a livestream running in our office to follow the current games. And so, it came about that my colleagues organized a betting game. Everyone had to participate, including me. The problem was that I'm probably one of the few Germans who isn't interested in soccer at all. So, I was faced with the challenge of predicting things I knew nothing about - classic everyday life for a data scientist.

The good thing in this case was that there was no problem to build up a database. Just a short Google search brought up a lot of freely available data and information. And besides the target was relatively clearly defined because there are only `3` meaningful outcomes in a soccer game in a world championship (`win`, `lose` or `draw`) and likely results (`1:0` or `1:1` and unlikely like `13:0`).

With that in mind, I built a model and bet exclusively on the most probable winner and one of the most likely results. I also occasionally initiated conversations about football to pretend that I was seriously into it. Although I still couldn't follow the soccer conversations with my colleagues, it took only a few match days until I got into the top tier with my method. And to my own surprise, I even made it to 1st place with my model in the end and won the betting game. And that without knowing the relevant players, backgrounds, or strategies.

A few years later I finished my PhD and left university to go to industry. And as it goes, there was again a soccer world championship coming up. As every championship, my new soccer-loving colleagues organized a betting game. Confident of victory, I took part and even bet a small sum of money. I dug out my old procedures and blindly bet on the results from the model (error 1). As I was otherwise very busy, I immediately bet on all available matches (error 2).

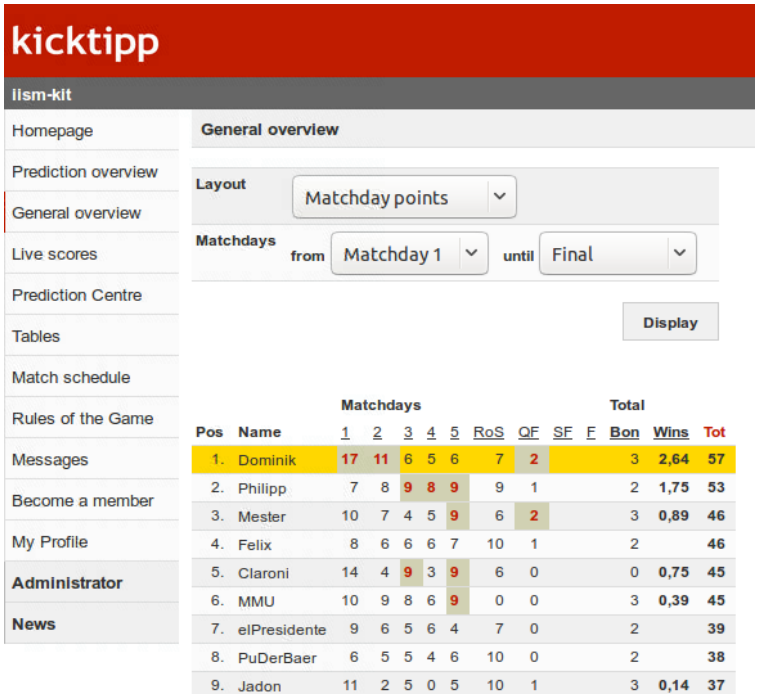

Figure 0-1 On my way to my glorious win against nearly 30 soccer-friendly colleagues.

When I looked at the table again a few weeks later (error 3), I was shocked! Almost all my results were wrong. I had almost always bet on the loser. When I looked at my model, it was clear that I had used it the wrong way (error 4). The betting game had progressed to the point where I had no chance of catching up, so I made one of the last places. And the cherry on top was that my boss forced me to bake a cake for the team as a punishment.

This anecdote illustrates very well what business data analytics is all about: It is about preparing a smart decision and then making it. You must not just blindly apply any code from your own archive or the internet (error 1). You have to crawl into the data and describe and understand the problem on the basis of the data. It is important to challenge the problem and its approaches regularly (error 2). Especially, at the beginning of a project a regular exchange with the stakeholders and the clear definition of a project plan and metrics are essential (error 3). These enable a clean evaluation and estimation of the quality of the data product (error 4).

As you can see, the reality of business data analytics is far from fortune-telling. It's not about gazing into a crystal ball but rather about deciphering complex data landscapes to extract meaningful insights. It's the art of transforming raw information into actionable strategies. And even if you proceed methodically and carefully, success is not guaranteed. Some days the models work out, and some days they don’t because the project is too ambitious. And sometimes the models predict the right thing but are poorly interpreted or configured by the users – so the project fails in the long term. But every time data and models can help us to make qualified decisions in domains, where we have no knowledge, if we use them right. And with the support of good data and well-defined models most people can even make better decisions than many experts in specific application areas (like my soccer colleagues).

Unfortunately, most analytics books primarily focus on predicting and building sophisticated data science models. They give you long pages with mathematical formulas but not a single line of code. But the story is not just about building models. It's about understanding the data and what it tells you about the problem and how to solve it with analytics. Analytics in industrial practice is about making smart decisions and supporting a decision-maker in difficult situations. However, this is not always as easy as it seems in my anecdote. There might be many pitfalls like "bad" data or data quality, difficult or unclear target variables or missing documentation. And many other challenges you can't even imagine yet. But do not worry, we will learn methods and workarounds to handle these issues!

In this book we will focus on prediction models not as an end in itself but in order to make smart business decisions. Business data analytics without considering the users of an analytics solution is near to worthless. A dashboard to manage KPIs with bad usability, will not be used by the management and hence, be worthless. Or sometimes an excellent model that is not understandable is worse than a somehow average model that has weaker performance but can be well explained and easily understood by the users.

Hence, this book is particularly about the intersection of business data analytics and data science to taggle the following questions:

- How to plan and setup a business data analytics project
- How to work with R as a programming language to conduct your project
- How to understand and work with your business data
- How to use analytics models to predict and understand your business
- How to build tools that a manager can use to make well-informed decisions, based on facts and not on feelings

R for Business Data Analytics

You probably ask yourself, why should I learn R to make business data analytics? Why can I not use Microsoft Excel, Python or C++? The answer is: I would not call R my favorite programming language. There are other languages like Python that I find better designed and more beginner friendly. I also teach other programming languages like Python at university and see many benefits in it. But R on the other hand was built by statisticians as environment for statistics and business data analytics, which makes it the perfect tool to wrangle and dive deep into business data. And will help you to generate and deploy your analytics models easily.

In the R community many free and useful packages exist to boost your business data analytics pipelines: For instance, making good-looking visualizations that have publication-quality is a difficult undertaking in most analytics tools like Python or Matlab. And I can tell you that R and its `ggplot2` package is probably the most advanced tool and best solution to make professional visualizations to support your analysis and communicate your results. From my time at university, I know that even if you started in another language like Python and plan now to publish your results or want to design posters or infographics for big journals and newspapers you will end up using R for this specific task.

Dashboards with `shiny`, or reports with `rmarkdown` are state-of-the-art solutions for your business data analytics workflow. The most relevant basics can be learned in 1-2 hours of work and you can produce first results for yourself in little time. R allows you to communicate your models and findings in an interactive webapp easily without changing or migrating to another analytics tool.

Every time I want to quickly dive deep into data and understand and model something statistical, I end up using R. With R you can produce higher quality outcomes with less effort compared to other programming languages like Java, Skala and in particular Python, if you want to make business data analytics. And that is what we plan to do. R is the tool that is designed to do business data analytics and statistics. And hence, that makes it my first choice to make business data analytics.

However, the goal of this book is not that you become a kind of religious R-fanatic. I know there is a healthy debate raging over the best language for data science, artificial intelligence, data mining, data engineering, etc. Many people believe Python, Go, Java, etc. is the better language for handling all these kinds of problems in every domain. We call these people cognitive miser. Or as Abraham Maslow said: “If all you have is a hammer, everything looks like a nail”. The goal of this book is to motivate you to go further in analytics and continuously learn new things. And this includes to learn further programming languages after reading this book.

How to Read this Book

Besides my job as data scientist, I am a longtime lecturer for data science and business data analytics at different universities and business schools in Germany. And I noticed that most students struggle that most concepts and algorithms of analytics and data science are described in academic language in thick books, sophisticated papers, or very abstract written blogs and websites with many errors. And hence, most people get wrongly the impression that they must get higher academic degrees in mathematics to figure this stuff out.

Thus, this book assumes that you have no academic degree or any experience with business data analytics or data science. Its goal is to illustrate and explain you the key concepts of business data analytics from scratch and to give you all the professional tools you need to start your journey as business data analytics professional. Business data analytics is art and science. There are many decisions where you have to rely on your intuition and cleverness. No scientific paper can teach you that. All you need is an open mind, willingness to puzzle and think mathematically, and a computer. And yes, this book!

We will cover a large number of different business data analytics techniques and best-practices. Rather than discussing the mathematical and theoretical foundations in detail, I will give you an intuitive understanding of the main concepts by illustrating you concrete practical examples. For that purpose, we focus on the application of the most relevant concepts and using mostly ready-to-go R frameworks and packages to get them working easily:

- `dstools` is a lightweighted package of different functions for professional data science development. It implements many useful algorithms and helper functions, which makes it a useful companion for data scientists and business data analysts.

- `tidyverse` is a popular collection of many data science algorithms and methods encapsulated in different packages (e.g., `dplyr`, `tidyr` or `ggplot2`). It's definitely the most relevant toolset for data science professionals and beginners. I recommend installing the `tidyverse`-collection instead of the many packages individually.

All these R packages are very powerful and provide an overwhelming number of functions and functionalities. The majority of packages we use in this book are part of the `tidyverse`-collection. The benefit is that these packages share a common way to do analytics and this will help you to get familiar with the topic.

Besides, we will use some other packages in this book that are not listed here. However, I made sure that these packages are compatible and follow the same principles. These specific packages are used for very specific tasks in this book (like web crawling) and therefore are not essential like the packages listed here.

You should also know, that it is not necessary that you know all R packages in detail or have already experiences in R or other programming languages. Nevertheless, coding experience will be helpful. If you are a Python programmer and you are familiar with some popular packages like `NumPy`, `Pandas`, and `Matplotlib` you will probably see many connections and similarities in the `dplyr`, `tidyr` and `ggplot` packages.

While you can read and understand the concepts in this book without writing one single line of code, I strongly recommend that you experiment and work with the different code examples in this book. For that purpose, I provide my code and further material for this book online at the website of the fictive *Junglivet Whisky Company*[1]:

www. junglivet.com

On this website I gathered all the material for this book and provide further information about the fictive company. You can use this material to deepen your background about the company, which can help you in the business cases of this book.

Furthermore, you find the lecture slides of the related university course of this book. You can use them to deepen your knowledge after reading this book or to look up some concepts, if you want another explanation about them or to learn more about their background.

Alternatively, you can also download the materials at my private git repository at:

https://github.com/dominikjung42/BusinessAnalyticsBook

If you are unsecure about git, or if you don't know how to use it yet and want to become familiar with it, the git tutorial documentation (*www.git-scm.com/docs/gittutorial*) is a great place to start. However, you don't need git skills in this book, but it will be definitely helpful in your future career as a business data analytics expert. After reading this book, you could continue your analytics journey by taking a deeper look at git and code versioning.

[1] *Similarities with people alive or other authorities are purely accidental and not intended.*

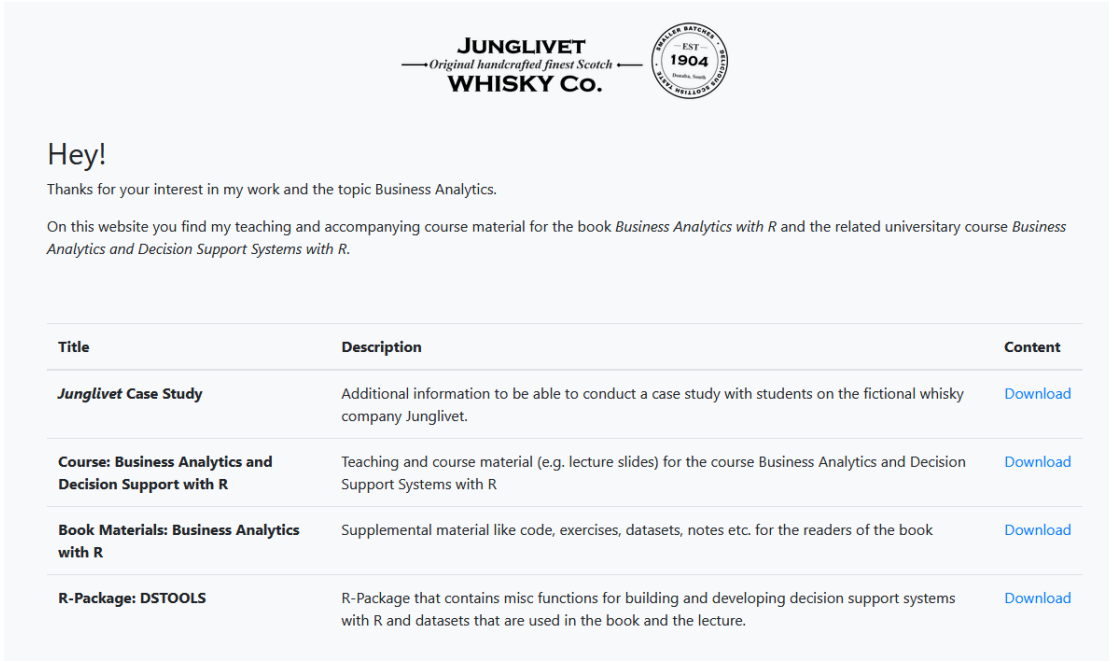

Figure 0-2 On the website of the Junglivet Whisky Company (www.junglivet.com) you can find further information related to the cases in this book.

Also, if you have never used R, in the chapter *Business Data Analytics Toolbox: R and RStudio* I will guide you through the installation process. You will learn the basic concepts of programming with R for business data analytics. The tutorials are for R beginners but also contain some information that might be also interesting for intermediate R programmers, so take at least a look at it before you decide to skip the chapter.

Throughout the book, we will learn more about business data analytics by solving case studies. You will be chief data scientist or chief business data analyst of the fictional *Junglivet Whisky Company*. By applying concepts and algorithms of this book you can test your skills and gradually leading the company along the path to success.

During this journey we will take a deeper look at all the different phases of analytics projects: *Business Understanding*, *Business Data Understanding*, *Business Data Preparation*, *Modeling*, *Evaluation* and *Deployment* (see the next chapter *How Business Data Analytics Experts Work* for more details). Sometimes we will focus on data processing and coding, sometimes we will look at the data itself and visualize it, and sometimes we will discuss concepts on a more general level.

In consequence, the book is organized in seven chapters that represent an introduction, the six different phases of typical business data analytics projects (whereby *Evaluation* and *Deployment* are combined in one chapter *Business Data Products*) and a last chapter to help you get started with real business cases.

As a first outlook, this book covers the following specific business data analytics topics and questions:

- What is business data analytics? And what are common business data analytics problems?
- What are the different phases and steps of business data analytics projects?
- Why business and data understanding are crucial for the success of your projects?
- How to handle, clean, and prepare your data for business data analytics?
- How to select and reduce your data to make it tidy?
- What are the most common analytics algorithms like regressions, k-nearest neighbors, decision trees, random forests, association analysis and ensemble methods?
- How to fit, optimized and train your analytics models?
- How to develop analytics solutions and data products for your users like the management or the domain experts?
- And finally, how to solve real business problems and case studies with business data analytics?

Conventions Used in this Book

The following typographical conventions are used in this book:

- *Italic,* indicates references like chapter names, URLs, email addresses and filenames.
- `Constant width`, indicates program code, as well as within paragraphs to refer to elements of the code such as variable or function names.

This symbol indicates other relevant things that are not in the text. If you see it, this box or element signifies a tip or important comment.

Code Examples

Further supplemental material (code examples, datasets, exercises, etc.) is available for download at: *www.junglivet.com*

Acknowledgments

I am also incredibly thankful to all reviewers and friends that helped me to make this book happen. They took themselves time to review my book in their free time besides their full-time jobs, family commitments and many other things they gave up to read my manuscript. In particular, I would like to thank Dr. Kevin Laubis for his feedback and for motivating me continuously for the project, Dr. Simon Behrendt for the comprehensive corrections and Dominik Raab for his numerous suggestions for improvement and the many entertaining discussions. Special thanks also to Prof. Dr. Peter Buxmann and Dr. Timo Sturm from University Darmstadt, with whom I was always able to discuss the application of machine learning in business data analytics and business informatics. I would also like to thank my former and actual colleagues at Karlsruhe Institute of Technology and Porsche AG, it was thanks to you that the idea for the project was born!

Above all, I want to thank my beloved wife, Trucy. She encouraged me to start and to continue working on this book. She is and was definitely one of my toughest critics, increasing the quality of this book.

Thank you all.

Contents

1 Introduction

1.1 Motivation

Congratulations to your new job! You are hired as business data analyst at the *Junglivet Whisky Company*. Through the book, we'll be learning more about business data analytics by building up analytics and reporting solutions for the different stakeholders at a traditional whisky distillery. However, the underlying problems are motivated by real problems which I faced during my analytics life-time. For educational purpose, I anonymized and adjusted the scenarios and details in this book.

We will start with simple straight-forward business problems and increase complexity step by step in each case study. Furthermore, we will implement basic algorithms and applications and even full decision-support systems from scratch. This will give you a strong understanding from business data analytics theory to application. And after working with this book, you will be ready to apply for a real business data analytics job in industry.

I would also like to point out again that you can download all the material we use such as code examples, datasets, notes, etc. from the course website at:

www.junglivet.com

So, you don't have to type all the coding examples by hand, just download the files and run them on your machine. However, I strongly recommend that you try out the exercises by yourself before looking at the solutions in this book or on the repository.

Additionally, I strongly recommend to install my R-package `dstools`. The name stands for "data science tools" and means that this package contains many functions that are useful for business data analytics and data science in general. Besides, it also contains many useful materials for this book. For instance, all the datasets of the case studies are included in the package so that you can load them easily with `data(<name of the dataset>)` in your R environment. You can solve all the exercises and tasks without the packages, but it will make your life much easier if you use it. And the good news: It is free to use!

If you decide to use it, you can install the newest version directly from GitHub with:

```
# Install the newest version from GitHub
install.packages("devtools")
devtools::install_github("dominikjung42/dstools")
```

Don't worry if you have problems to understand the R code in this book right now. In chapter 2, we will take you through an R crash course, including how to install and manage such packages in R. For the moment, it's enough if you try to get a general understanding of what we are doing in the code.

D. Jung, *The Modern Business Data Analyst*, https://doi.org/10.1007/978-3-031-59907-1_1

1.2 Whisky Quality Problems in the Junglivet Company

Are you ready to go for your first analytics jobs in the *Junglivet Whisky Company*? Let us start with an introductory example. The purpose is to give you a first impression of the business data analytics process and the central concepts and problems we will discuss in the book. You can recreate it directly on your computer if you already setup R. But it will not be necessary if you haven't, we will setup R and the integrated development environment (IDE) RStudio together in the next chapter. So, best would be if you just read the code examples and try to understand them. Then we setup R and RStudio together in the next chapter and you can try to solve the business case by your own after getting a better understanding of everything. If you continue with this book, we will also learn other and better techniques than used in this introduction to analyze analytics problems. So, feel free to come back to this example any time later in your reading.

In summary, for the moment it's enough if you read it and try to understand the code examples on a general level. The purpose is to show you an analytics workflow, to tease you for the upcoming chapters and not how to write code. I want to motivate you and illustrate the process of a business analyst. Please keep this in mind when reading the following scenario. Don't worry, there will be only some minimal code examples given to help you to illustrate a typical analytics project.

1.2.1 Welcome to the Junglivet Whisky Company

The company sits in a charming old building some five-minutes-walk from the center of the next village. Your office and the production hall seem to be next to each other. When you arrive at the building with the cab, a mid-aged lady waves to you cheerfully with both hands. "Cheers, welcome to the Junglivet factory!" says the lady as you leave the car. "Glad you are here darling! My name is Miss Gleck. And because there is a lot of work for you to do, it is best if you report to Mr. Gumble at the distillery right away. There are considerable problems. I will see to it that your suitcases are brought to your office in the meantime. Let's go! What are you waiting for?"

Somewhat confused, you leave the office and go next door to the production hall. A friendly older gentleman in a workman's outfit greets you from afar. He has a very strong handshake and introduces himself as Mr. Gumble. He is from the distillery and responsible for the production team. You make some small talk while Mr. Gumble is full of questions about you and your analytics background. And while you talk, you recognize that until now the company made most decisions based on feelings, intuition and as a consequence has now serious financial problems. Thus, Mr. Gumble and his colleagues are very excited to have you in the team, and hope that you can give new drive by bringing your analytics knowledge aboard.

For now, he needs your help with some problems in the production line. He received complains about the quality of the whisky and tries to find the reason for that. Mr. Gumble gives you an USB-stick with the production lane data of the last two weeks and asks you to analyze it (which is quite uncommon, because in reality people don't hand you the data you need and you have to search for it).

He says that the dataset should contain all the data that you need (which makes the situation even more suspicious – you never get "all" the data you need in the first try in your whole career

as business data analyst). As you plug in the USB-stick in the local computer you see that it consists of an excel sheet from the production lane containing 8 columns. An excerpt of the data is illustrated in Table 1-1.

Table 1-1. Excerpt from the dataset of the production lane (productionlog_sample.xlsx).

DAY	MONTH	MANUFACTURER	PRODUCT	SHIFT	COLOR	MALTING	TASTING
1	4	Leonard	Junglivet	Morning	0.27	Inhouse	895
1	4	Carlson	Junglivet Premium	Evening	0.27	Burns Best Ltd.	879
2	4	Leonard	Junglivet	Morning	0.28	Inhouse	938
…	…	…	…	…	…	…	…

Each row represents information of the shift and the final quality of the whisky production. For example, the first row shows all the information of the morning shift's Junglivet production (column SHIFT and PRODUCT) of the first of April (DAY and MONTH). The responsible brew master was Mr. Leonard (MANUFACTURER) and the malting was produced inhouse (MALTING). During production the color of the whisky was measured (COLOR) and after the production a quality assessment has been conducted. The final score in the quality check before delivery is 895 (TASTING), which is a quite good value (1000 is best says Mr. Gumble).

In summary the variables or features in the dataset are:

- DAY: The day of the production
- MONTH: The month of the production
- MANUFACTURER: Name of the leading brew master
- PRODUCT: Product type like Junglivet or Junglivet premium
- SHIFT: Indicates morning or evening shift
- COLOR: Color indicator on a scale between 0 and 1
- MALTING: Origin of the used malting products
- TASTING: Quality indicator based on a whisky tasting on a scale between 0 and 1000

If you want to learn more about the different features you can also download the additional material about whisky production on the *junglivet.com* website. But for the moment, let us continue helping Mr. Gumble!

1.2.2 Step 1: Business and Data Understanding

Unfortunately, you don't have your own laptop with your analytics toolbox with you, but you install Base R on the local computer of Mr. Gumble to take a quick look at the data. If you have R and RStudio installed and are familiar with it, you can follow the analysis by downloading the data from the book's website and open it in RStudio.

To start with your analysis, you have to load a couple of packages in your R environment (if you are not familiar with that, we will learn that later in chapter 2). While R has only some basic

functions, it can be easily extended by loading further packages. In this case you decide that you will work with an excel sheet and probably some simple data processing and visualization packages. For that purpose, you install the following packages on your new laptop:

```
install.packages("devtools")
devtools::install_github("dominikjung42/dstools")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("readxl")
```

And then load the freshly installed packages with the R-Bases `library()` function:

```
library("dstools")
library("dplyr")
library("ggplot2")
library("readxl")
```

Now that we have prepared everything for our analysis, we want to load the data into our R environment. For that purpose, you put the excel sheet from Mr. Gumble in the same folder as your R analytics project. If you are interested to follow this analysis you can download the file on the book's website, but if this is your first time reading this section it would be better if you just continue reading and come back to the chapter later.

After the dataset is in your local folder you load the production data directly form terminal with the following command in your R environment:

```
dataset = read_excel("productionlog_sample.xlsx")
```

Then you open the logfiles from the whisky production in your R environment. You want to check if your dataset is loaded correctly. You take a first visual look at the top lines of your dataset by calling the `head()` function:

```
> head(dataset)
DAY MONTH MANUFACTURER PRODUCT          SHIFT  COLOR MALTING          TASTING
1     4 Leonard    Junglivet          Morning 0.27  Inhouse              895
1     4 Carlson   Junglivet Premium Evening 0.27  Burns Best Ltd.      879
2     4 Leonard    Junglivet          Morning 0.28  Inhouse              938
2     4 Carlson    Junglivet          Evening 0.32  Inhouse              900
3     4 Leonard    Junglivet          Morning 0.32  Matro Ltd.           917
3     4 Carlson    Junglivet          Evening 0.28  Inhouse              900
```

If we have code examples that contain code and results / output of your code, I use the ">" symbol. The symbol means that this code is entered in the console of your RStudio IDE. Other lines in the code example without the symbol represent the R output.

As Mr. Gumble said, your dataset consists of `8` columns (good news – in most real cases your data description will not fit to the data you receive). As in the example, a single column represents further information about the whisky production shift. For instance, the third column

"MANUFACTURER" contains information about the responsible producer of the shift or the last column about the quality of the final product. These is quite important information and we want to use it to investigate the reason for the quality reduction in the last weeks.

All the datasets of this book are also available in the `dstools` package (online available at *https://github.com/dominikjung42/dstools*). Instead of loading them manually in your R environment you can just run `data("name of the dataset")` if you have the package installed. For instance, the following command loads the current dataset of this tasks from the `dstools` package:

```
my_data = data(" productionlog_sample")
```

Before you can start further investigating the problem, you decide to take a look at the consistency of the dataset of Mr. Gumble. This is crucial, because, as a business data analytics expert, you know that data is often dirty. It can contain missing values or the wrong values. Or the format of the dataset might be broken or contain some logical errors (like wrongly named columns or other curiosities).

Thus, a natural next step is to investigate the consistency of your dataset and the related columns by calling the table view:

```
View(dataset)
```

A table pops-up in your RStudio containing your dataset in a tabulated form. You can see the results in Figure 1-1.

R Code - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

dataset

Filter

	DAY	MONTH	MANUFACTURER	PRODUCT	SHIFT	COLOR	MALTING	TASTING
1	1	4	Leonard	Junglivet	Morning	0.27	Inhouse	895
2	1	4	Carlson	Junglivet Premium	Evening	0.27	Burns Best Ltd.	879
3	2	4	Leonard	Junglivet	Morning	0.28	Inhouse	938
4	2	4	Carlson	Junglivet	Evening	0.32	Inhouse	900
5	3	4	Leonard	Junglivet	Morning	0.32	Matro Ltd.	917
6	3	4	Carlson	Junglivet	Evening	0.28	Inhouse	900
7	4	4	Leonard	Junglivet	Morning	0.29	Inhouse	934
8	4	4	Gumble	Junglivet Premium	Evening	0.29	Matro Ltd.	951
9	5	4	Leonard	Junglivet	Morning	0.33	Matro Ltd.	852
10	5	4	Carlson	Junglivet	Evening	0.27	Inhouse	850

Figure 1-1 Table with your dataset representing the different whisky production shifts and their output.

The table view of RStudio allows you to investigate your data in an excel-like format. You can see the rows with different IDs on the left. Besides there are different columns like `DAY` or

MANUFACTURER. You can sort your dataset by a specific column by clicking the column name or the small rectangle beside the name. Furthermore, the RStudio view has a Filter option. You can find it in the menu, right on top of the column names on the left. If you click on the button, you can filter the dataset to find specific rows. The magnifying glass in the right corner allows you to search for specific words or numbers in your dataset.

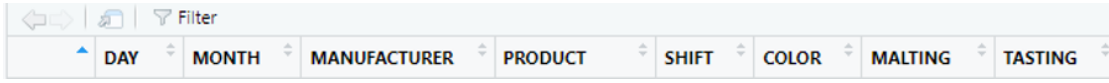

Figure 1-2 The filter function is very useful to look for specific values and check the quality of your dataset.

After eyeballing all the rows in the table, you see that row 11 has missing values. "Would have been too nice if the data quality would be perfect", you think to yourself. You decide to compute further descriptive statistics to check the quality of the dataset before you handle row 11.

A very common and useful function for this is the summary() function. You decide to run it by typing summary(dataset) in your R-console:

```
> summary(dataset)
      DAY            MONTH          MANUFACTURER       PRODUCT            SHIFT
 Min.   : 1.0   Min.   : 4     Length:21          Length:21          Length:21
 1st Qu.: 3.0   1st Qu.: 4     Class :character   Class :character   Class :character
 Median : 5.5   Median : 4     Mode  :character   Mode  :character   Mode  :character
 Mean   : 5.5   Mean   : 4
 3rd Qu.: 8.0   3rd Qu.: 4
 Max.   :10.0   Max.   : 4
 NA's   :1      NA's   :1

COLOR              MALTING            TASTING
Length:21          Length:21          Min.   : 822.0
Class :character   Class :character   1st Qu.: 875.0
Mode  :character                      Median : 925.5
                                      Mean   : 918.5
                                      3rd Qu.: 957.8
                                      Max.   : 999.0
                                      NA's   :1
```

As you can see R generated a short report of each column of your dataset. There is information if the column contains numeric or character values. And some more information like the different ranges, if the column is numeric.

Our missing row number 11 has also been detected by R (it is marked as NA). NA stands for "not available", which means in our case that there is one row that contains no values. Furthermore, the column or feature MONTH contains always the number 4. Features with only one value like MONTH are useless because they lack variability and provide no distinguishing information. Let us now have a look at these issues called data quality problems!

1.2.3 Step 2: Business Data Preparation

Therefore, you decide to start the next step of your analytics process, the business data preprocessing. Preprocessing means that you preprocess the data for the analysis and the modeling. This includes handling rows without values, rows with errors or remove variables from your dataset that are of no further interest.

You decide to remove row 11 and the column MONTH in the dataset. In a first step, you decide to start with the irrelevant column issue and remove it. Then you run names() to check if the column is removed and the dataset has the correct number of columns:

```
> dataset = subset(dataset, select = -c(MONTH))
> names(dataset)
[1] "DAY" "MANUFACTURER" "PRODUCT" "SHIFT" "COLOR" "MALTING" "TASTING"
```

Then you omit the row 11 with NA values to get a clean and usable dataset. You use the na.omit() command and to check if everything worked out correctly, you run nrow() to check the number of rows is reduced by one:

```
> dataset = na.omit(dataset)
> nrow(dataset)
20
```

For now, the dataset is fine and ready for the next step, so we can continue our analysis.

Data quality problems are very common in typical analytics projects. And this process is very relevant and will consume most of your time. We will handle further concepts and algorithms for data cleaning and other major techniques for data preprocessing more detailed in the chapter *Business Data Preparation*. This dataset is an example dataset for the introduction and is hence very simple and has no complicated data quality problems.

1.2.4 Step 3: Business Data Analytics

In a next step, you want to investigate why there might be quality problems with the whisky production lines of Mr. Gumble. You decide to plot some specific features in the dataset to help Mr. Gumble to find some possible drivers of the production problems.

For that purpose, you run different visualization commands from base R on selected features. The features MALTING, SHIFT and MANUFACTURER seem to be interesting. Sometimes the existing raw materials in whisky production are not enough and the company has to buy more raw materials from suppliers. It could be that these differ in quality and have a negative impact on the whisky. This is represented in the feature MALTING. On the other hand, it could also be that the leading brew master was careless and not all production processes were carried out with the necessary accuracy, which is represented in the feature MANUFACTURER. Or that the employees have difficulties with concentration in the evening shifts (or the morning shifts), which is represented in the feature SHIFT.

Hence, you decide to investigate the relationship between MANUFACTURER, SHIFT, MALTING and TASTING even further. To investigate relationships between non-numeric, class features like SHIFT or MANUFACTURER and the numeric quality (TASTING) you use boxplot(). A boxplot is a simple but informative visualization that is used very often in business data analytics. In addition, you can often quickly see the influence of different categorical features in it.

Let us now take a first look at the supplier and their possible influence on the quality. You can do it with:

```
boxplot(TASTING ~ MALTING, data = dataset)
```

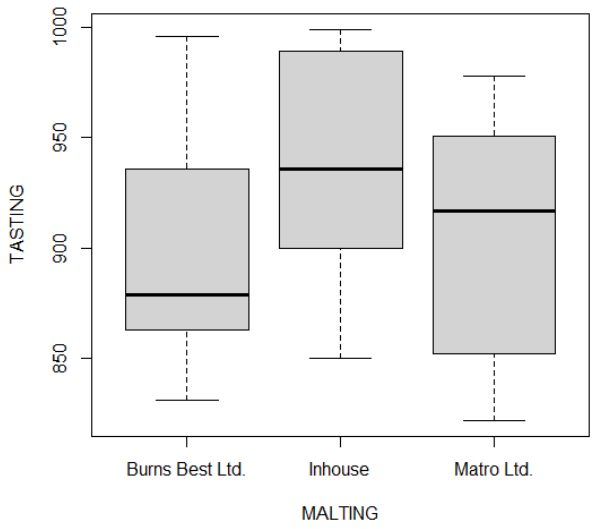

Figure 1-3. Boxplot visualization of the relationship between the classes "Burns Best Ltd.", "Inhouse", and "Matro Ltd.".

The result is a visualization with three boxes that you can see in Figure 1-3. The figure displays the distribution of the different supplier (represented by the column `MALTING`) of the *Junglivet Whisky Company* and their relationship on the quality in our whisky production dataset (represented by the column `TASTING`).

The box, which the name boxplot implies, shows in which range the middle 50% of all values are located. The bottom of the box is the point in the data where the bottom 25% have accumulated. The thick line in the middle of the box is the median. That means that up to this line 50% have accumulated. As we can see the Burns Best Ltd. Malting has the lowest median, indicating that it has wors quality. The top of the box marks the point where 75% of all values have accumulated.

We see that most whiskies with the supplier "Burns Best Ltd." have lower quality than if we use our own materials ("inhouse"). However, as a business data analyst you know that the plot gives only a short overview. It seems pretty clear that we have to further investigate the reason why "Burns Best Ltd." is associated with bad whisky. You might ask how can a business data analyst provide more evidence? A first step would be to compute further descriptive statistics to check your finding. Another step would be to run some statistical tests or models, which we will learn later in this book. For the moment, this seems to be enough for you because you know that Mr. Burns from Burns Best Ltd. has been condemned in the media for adulterated products in the last weeks.

To make a last quick check you decide to compute the mean whisky quality of your production with the bad supplier "`Burns Best Ltd.`". You can do that with:

```
> dataset_burns = subset(dataset, MALTING==c("Burns Best Ltd."))
> mean(dataset_burns$TASTING)
[1] 901
```

In the first line of code we make a new dataset named `dataset_burns`, which is a subset of our original dataset filtered by the supplier "`Burns Best Ldt.`". It contains only rows that have "Burns Best Ltd." as a supplier. The mean quality of whisky with this supplier is `901`.

To compare it with the whiskies that use inhouse materials you run:

```
> dataset_inhouse = subset(dataset, MALTING==c("Inhouse"))
> mean(dataset_inhouse$TASTING)
[1] 934.6
```

You were not surprised to see that the mean whisky quality is higher. You quickly close the laptop and decide to look at the other features `SHIFT` and `MANUFACTURER` later (which could be a very good exercise if you come back later to this chapter) and head for the office of Mr. Gumble. As you enter his office, you hear him discussing the whisky quality problems with Carl Leonard from the production team. You report your findings to Mr. Gumble and he is excited!

Mr. Gumble and Carl Leonard ask if you can also provide a possible indicator of the whisky quality that can be used in production. It would be very helpful if they could take a look at some easy measures like the whisky color during the production. And if the color has some deviations, they can stop the production instead of waiting of the final tasting.

What a day you think and go back to work. To investigate the relationship between numeric features like `COLOR` and quality you decide to use a quick scatterplot which can be made with `plot()`.

To investigate the relationship between `COLOR` and `TASTING` you run the following code, which puts `COLOR` on the x-axis and `TASTING` on the y-axis:

```
plot(x=dataset$COLOR, y=dataset$TASTING)
```

As you can see in the following Figure 1-4 there seems to be a relationship between the color of a whisky (`COLOR`) and the measured quality in the tasting (`TASTING`). The best color for *Junglivet* whisky is around `0.3` which indicates that color could be a possible indicator detecting quality issues already in the production.

After sharing your thoughts with Mr. Gumble (who is now completely enthusiastic about you), you realize, exhausted, that it is already really late. You go straight forward to your new appartement and find your bags and some papers with additional information about the company and a card with the weekly menu of the pub next door. You slip out of the building before anyone else can ask you for further help and go straight in the pub to get some food.

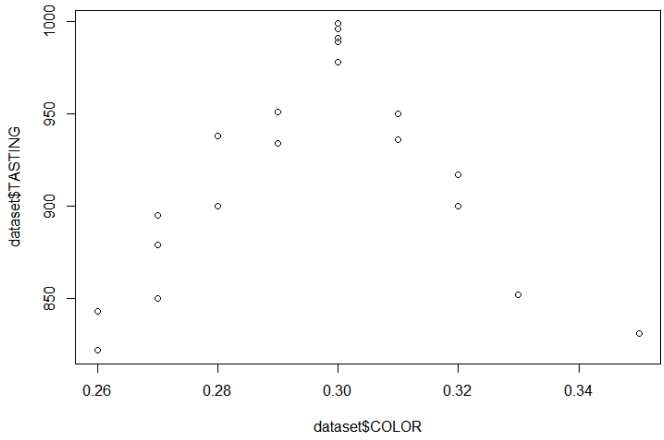

Figure 1-4. Basic plot of your dataset with the variables `TASTING` and `COLOR`.

"What an exciting first day, there is really a lot to do here" you think while taking a seat in the pub. You order some random stuff from the card and you doubt whether this was really the right decision to come here. While eating some horrible English food and drinking some glasses of 12-year-old *Junglivet* whisky to suppress the taste of the English food you start recapitulating your first day at the *Junglivet Whisky Company*.

You could help Mr. Gumble to solve his production problems, which will definitely increase future customer satisfaction. And your employees seem to be handsome and hard-working. However, if such simple problems have not been solved no wonder that the company got financial problems. After your fourth round of Junglivet whisky (the food doesn't seem so bad after all) you start to become optimistic. Based on some online reviews there is a huge number of customers loving the charm of the middle-class, family-controlled *Junglivet Whisky Company*. And besides all the problems in operational business, the whisky is delicious. Time to take some rest, there will be a lot of work for you to do!

The last chapter was an illustration of a small business data analytics case. We had a dataset, investigated and preprocessed it to derive further insights from it to support our business processes (whisky distillation). However, this was a quite simple business data analytics scenario. While, we detected the outlier in the analysis step, analytics experts normally check for outliers already in the data cleaning and do more advanced stuff in the analytics to detect patterns. In the following chapters of this book we will learn these sophisticated and automated techniques to detect relationships in your data instead of "fitting them by eye". Let's go!

1.3 The Business Data Analytics Mindset

Our first day at the *Junglivet Whisky Company* has illustrated that just running some R code will not make you a successful business data analyst. You must understand and clean your data. And if you present your first results to the management, there will always follow subsequent tasks

and analyses. From my experience, I can tell you that a business data analyst needs a very specific mindset to be successful in practice.

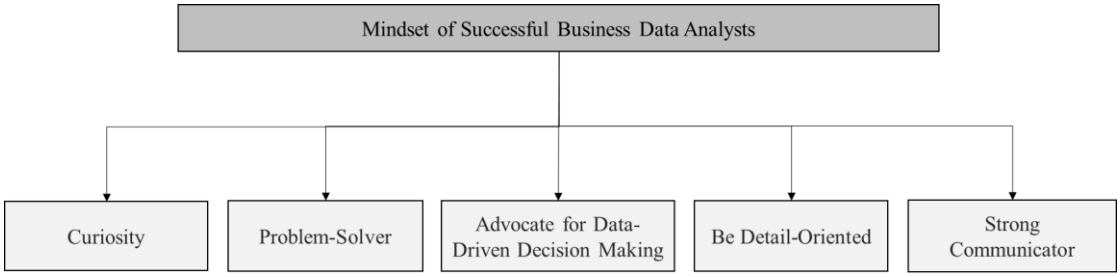

Figure 1-5. The key purpose of business data analytics is transforming data into actionable insights that guide business decisions.

First of all, you must see you as a kind of detective or investigator. You need an insatiable curiosity and inquisitiveness to explore business data in-depth. Ask yourself: How to handle and get your data for your project? What type of data is suited for your business case? Is the quality of the data sufficient to answer your questions? Is the dataset systematically biased or corrupted? How to handle missing values or outliers? How to use the data to support a business process? Rather than merely accepting numbers and figures at face value, you have to question the "why" and "how" behind the data patterns. This will lead you to meaningful connections and patterns that might go unnoticed by others.

Business data analysts also need a natural problem-solving mindset. They approach each data analysis challenge as an opportunity to provide solutions and make informed decisions. Armed with the methods and tools you will learn in this book, you have to break down complex problems into manageable parts, design statistical experiments, and conduct complex analyses to identify the drivers of a problem. There will be many challenges, especially if you work for no data-driven company but see them as chance to show the management potential for improvement and how business data analytics can contribute to the organization's success.

Third, business data analysts must see themselves as advocates of data-driven decision-making. They have to show that data is the most potent tool for informed decision-making. Rather than relying on gut feelings or intuition, you have to use and show how data can guide decisions. Build dashboards, setup reports and automate them, help others to use your findings. This data-first mindset is necessary to empower your colleagues and to influence key stakeholders in your organization with evidence-based insights.

In my humble opinion, in the world of data analysis, the devil is in the details. As a consequence, business data analysts have to understand the importance of accuracy in their work. They have to be meticulous and detail-oriented. Always pay close attention to business data quality, ensuring your business data is clean, consistent, and reliable. Have a keen eye for details allowing you to identify errors, outliers, or anomalies in the data early in your project, leading to more accurate and trustworthy results. And sometimes to better processes in your organization!

Fourth, business data analytics is useless if the insights and findings can and will not be used. If you want to provide the results of your business data analytics project like reports, visualizations, predictions, reports or dashboards to your users, your results should be designed always for non-experienced users. The insight or your analysis have to be communicated and provided in such a manner that your non-experienced users can use them effortless for decision-making. As a business data analyst, you must effectively communicate your findings to this non-technical audiences. Therefore, your results should represent the wording, working and cognitive capabilities of the end-users that are not firm with business data analytics.

In summary, the mindset of a business data analyst is the driving force behind turning data into actionable insights that propel businesses forward. Curiosity, problem-solving abilities, data-driven decision-making, attention to detail, and strong communication skills are key pillars of this transformative mindset of a data-driven organization. By embracing these characteristics, business data analysts can unlock the true power of data, revolutionizing how organizations make decisions, optimize processes, and gain a competitive advantage in the digital age.

1.4 How Business Data Analytics Experts Work

Since the beginning of business data analytics as a discipline, different process models have been proposed again and again as to what a successful analytics project should look like. There has been propositions from statistics-related organizations like the SEMMA framework from the SAS institute (SAS Institute, 2017). It organizes business data analytics as the process of sampling, exploring, modifying, modeling, and assessing. Other models come from data warehousing literature like the knowledge discovery process (Fayyad et al., 1996) that describe how unknown patterns can be identified in databases. Recent frameworks tackling new trends in software-development, like agile, are covered by Microsoft's TDSP - team data science process lifecycle (Microsoft Corporation, 2020), or focusing on specific subdomains like the DASC-PM – data science process model explicitly for data science in information systems (Schulz et al., 2020).

While analytics research is still proposing new guidelines and frameworks every year, there is one framework that has been established in business data analytics and data sciences since the beginning and most other frameworks build on it: The **cross-industry standard process for data-mining** (CRISP-DM). CRISP-DM has been developed by a consortium of over 200 analytics organizations in the nineties and published in 1999 to gather best practices across industries (Chapman et al., 1999). Since then it has become the most widely-used process standard for analytics and data science in general (Brown, 2015; Piatetsky, 2014).

In this book we will rely on the CRISP-DM methodology. If you are planning to work in an analytics team in industry your processes will be probably organized based on CRISP-DM. And if you understand the different phases and ideas behind them, you can easily translate them to the other frameworks. It's the most general and most widely used and will help you to organize your projects even if you work just for yourself.

The CRISP-DM framework divides analytic projects into six phases that are related to each other. The six phases are **business understanding**, **business data understanding**, **business data preparation**, **modeling**, **evaluation** and **deployment** and are illustrated in Figure 1-6.

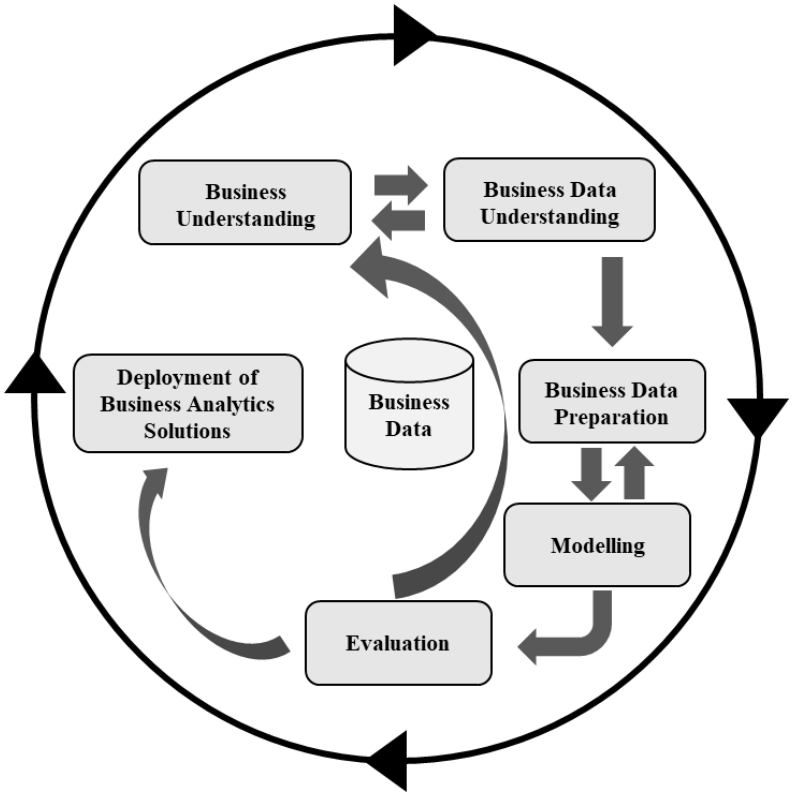

Figure 1-6. Business data analytics framework based on the cross-industry standard process for data-mining (CRISP-DM), adapted from the CRISP-DM 1.0 Guideline (Chapman et al., 1999).

The six phases build on each other but it is not required that you go through them step by step. In most real-live scenarios you will move back- and forth between them. In your analytics projects you will probably start with a business understanding workshop and setup the project management, then you will take a deeper look at the data basis. If you have a rough understanding of the data quality, you will go back to business understanding and update the project management timeline. Afterwards you can start with business data preparation. This illustrates that the phases depend on each other which is symbolized by the arrows in Figure 1-6 and explained more detail in Figure 1-8.

After you produced your first analytics results you can provide the results e.g., as a dashboard for your users. The new dashboard system can generate new insights and questions about the business problem, or your findings might influence your understanding of the initial problem, and hence the process starts again. This is symbolized by the outer circle.

The different steps or phases of a CRISP-DM based business data analytics process have different activities linked to them. The activities are common for every type of analytics projects and are necessary to avoid pitfalls and problems. It is important that you know them, and we will learn different methods how to actually do these activities in R (see Chapter 2). You will also find a checklist of the related activities for each phase at the end of each chapter. This way you can make sure you haven't forgotten anything!

Let us now look at the first phase. The main objective of the first phase **business understanding** or sometimes also called project understanding or domain understanding is to generate a general understanding of the objective of your business data analytics project in the team. Most clients will have a very abstract understanding of the potentials and limits of analytics. Or the risks or challenges of the projects have never been discussed. Some clients will probably have no technical understanding or unrealistic requirements. In this phase you have to taggle these pitfalls and identify the business problem you want to solve. Best-practice is to focus on the business problem with the highest impact (costs, revenue, etc.). Based on these specifications you organize your analytics project and try organizing and allocate manpower, time, hardware and budget over the whole project timeline. Your final project plan should be communicated as early as possible for every stakeholder. Find a shared consensus how to measure the success of the analytics project. This step is crucial for your project success and I will show you helpful methods in the next section.

The objective of the next phase is to overview the data sources related to the project. In particular, the data quality and availability are key success criteria. Do not forget that the best analytics system is useless if you cannot provide sufficient data for it. All activities of the **business data understanding** phase have the purpose to make sure that there is enough data and information about the data so that it can be used for analytics. Try to start your exploratory data analysis (EDA) with a small initial dataset and then go forward from there and investigate the other data sources. Check if the data quality is high enough to fulfill your requirements for the objective defined in the business understanding phase. If not, it is important to adjust the objective together with the stakeholder and discuss possibilities how to gather or generate the data that is needed. If that's not possible you have to cancel the project because there will be no usable solutions at the end.

If you got a rough understanding of the data, the data fields and data sources you have start the **business data preparation**. The data preparation step is the most time consuming one in every analytics project. Decide on which data you want to concentrate and how to transform the data to more general formats that can be used and shared easily in your team and between tools and systems. If you use raw business data you have to increase the data quality by removing corrupted data or replacing missing values. Do not forget that some specific data might require specific processing and anonymization steps due to national or international laws like the European general data protection regulation (Voigt & Von dem Bussche, 2017). Furthermore, redundant and unimportant data and features have to be removed or at least reduced as much as possible. Sometimes you have to generate new features or bootstrap your data because you have not enough of it. In addition, it is now also possible to carry out projects entirely on synthetic data.

Using your preprocessed data, you can finally start with **modeling**. In this phase, you apply business data analytics models and methods on your business data to generate insights to solve a specific analytics problem. Alternatively, you can use statistical methods to generate insights or simple models or decision rules that can also be used for decision support systems. Today exists a plethora of different methods and algorithms, the challenge is to find the right one for the right kind of problem.