8. Business analytics foundations including the R and SQL

What is business analytics (BA)? What problems does it solve and what tools does it use? What are R and SQL? How is BA related to R and SQL? Are there any examples of good practices where these software's are used to solve logistic business problems?

On these and similar questions, we will try to provide answers in the following chapter.

8.1 What is business analytics?

BA represents a holistic approach to data analysis and business decision-making. It is a data-driven environment with the goal of improving company business performance by providing a foundation for more informed decision-making. It is a systematic thinking process that applies qualitative, quantitative, and statistical computational tools and methods to analyse data, gain insights, inform, and support decision-making. Any particular analysis may use a variety of techniques including diagnostic, predictive, prescriptive, and optimisation models (Power et al., 2018). This Mikalef et al., (2019) provides a roadmap for both academic exploration and practical implementation, highlighting the transformative potential of analytics when properly integrated into organizational processes. Accordingly, the authors state that BA require organizations to radically redesign how such initiatives are approached, designed and refined, how resource planning and orchestration is executed and strategically aligned, as well as re-valuate their expected performance outcomes, their association with strategic objectives and, as a result, develop appropriate KPIs (Mikalef et al., 2019).

The main tasks of BA is to provide the knowledge pipeline to ensure a coherent link between raw data and business decisions. The overall goal is business effectiveness through 'verticalization,' usability, and integration with operational systems (Kohavi, Rothleder, & Simoudis, 2002). The BA has many application areas and related derivatives: Financial Analytics, Supply chain analytics, Crisis Analytics, Knowledge Analytics, Marketing Analytics, Customer Analytics, Service Analytics, Human Resource Analytics, Talent Analytics, Process Analytics, Risk Analytics (Holsapple, Lee-Post, & Pakath, 2014).

There are three types of Business Analytics platforms:

- **Descriptive** – They look at existing data and provide summary statistics and basic visualization.
- **Predictive** – They use existing data to estimate the most likely future scenarios.
- **Prescriptive** – They automatically process big data, business rules, market conditions, etc. These platforms utilize machine learning and artificial intelligence methods. The goal is fully automated decision-making on which actions a company should take considering the current situation to achieve the desired business objectives.

The interest in big data and business analytics has grown exponentially over the past decade (Mikalef et al., 2019). Modern-day BA is rooted in the ongoing advances of systems to support decision making. These advances include increasingly powerful mechanisms for acquiring, generating, assimilating, selecting, and emitting knowledge relevant to making decisions. Given its decision support heritage, business analytics necessarily partakes of and exploits these mechanisms. The knowledge that must be processed ranges from qualitative to quantitative and BA is concerned with operating on both knowledge types, as appropriate for the decision at hand (Kohavi, Rothleder, & Simoudis, 2002). The reason why should some organization apply BA is in the problems it solves. Problems which BA emphasise are the curtail problems for effective management of the company. Accordingly, there are several rationales for applying BA (Holsapple, Lee-Post, & Pakath, 2014):

- Achieving a competitive advantage
- Supporting of an organization's strategic and tactical goals
- Better organizational performance
- Better decision outcomes
- Better or more informed decision processes
- Knowledge production
- Obtaining value from data

Regardless, of the type of platform used in BA to solve and support the decision-making process in each company there are three key pillars of each BA solution (Figure 8.1). Generally, the BA occupies a place in the spectrum between Computer Science/Mathematics/Data Science

(on one hand) and Business and Management (on the other). Business analytics requires both technical and business knowledge. A major problem in designing a BA is that the boundaries are not distinct (Power et al., 2018). Accordingly, to perform the BA mathematical tools are needed to identify, extract and represent the insights in the right manner via tables, graphics, formulas, etc. Additionally, the programming tools serve as a support for this kind of activity and enable fast and error-less computations, compared to the traditional paper and pen approach. Last but not least, logistics expertise is needed in a given area or business problem to pinpoint the key influential factors and related ecosystem.

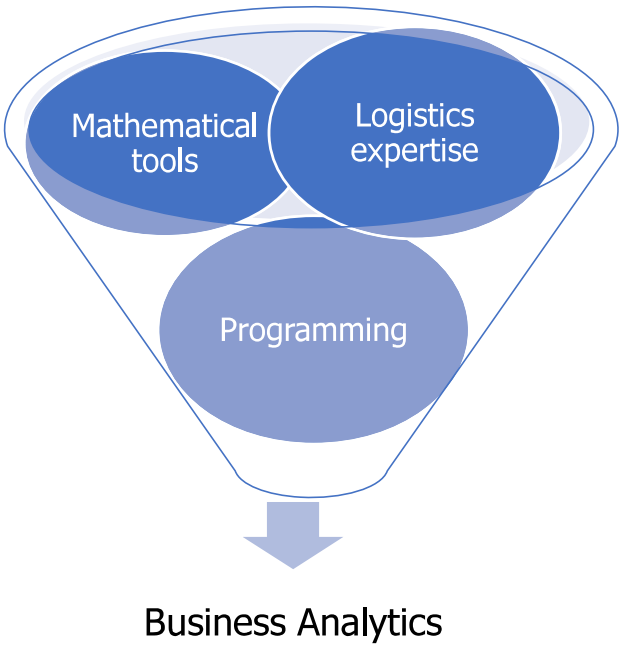

Figure 8.1 The key pillars of BA in the context of the supply chain and logistics.

The key consumer is the business user, whose job, possibly in merchandising, marketing, or sales, is not directly related to analytics per se, but who typically uses analytical tools to improve the results of some business process along one or more dimensions (such as profit and time to market). Business users do not want to deal with advanced statistical concepts; they want straightforward visualizations and task-relevant Outputs (Kohavi, Rothleder, & Simoudis, 2002).

8.2 What is R?

R is an integrated suite of software facilities for data manipulation, calculation and graphical

Display (R Core Team, 2019). Among other things it has:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- Graphical facilities for data analysis and display either directly at the computer or on hard-copy, and a well-developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed, most of the system supplied functions are themselves written in the S language.)

The main advantages of R are the fact that R is freeware and that there is a lot of help available online. It is quite similar to other programming packages such as MatLab (not freeware), but more user-friendly than programming languages such as C++ or Fortran (Torfs & Brauer, 2014). R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis (R Core Team, 2019).

Instal R and R Studio

To install R, go to cran.r-project.org and click the downalod the R for a specific operating system on your machine (usually Windows) (Figure 8.2).

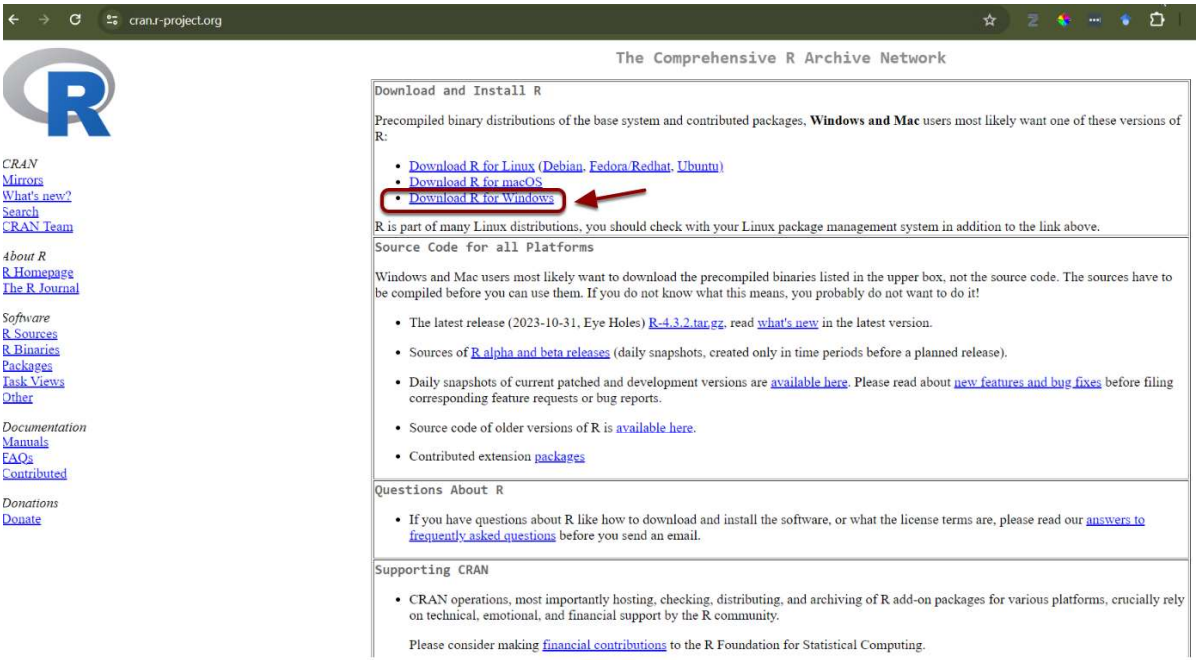

Figure 8.2 Download page for R software.

This will download the R software, and the installation procedure is the same as with any other software. When R software is installed, it will come without an advanced integrated development environment (IDE) to help and assist users in making different analyses. Although it is possible to do any kind of analysis with only R installed, it is preferable to pair it with some modern IDE, like RStudio, which is one of the most popular IDEs. The procedure on how to install the RStudio is similar to the core R software. Go to https://posit.co/download/rstudio-desktop/, search for RStudio Desktop open-source licence, download it, and install it. After installing the R and RStudio the user will have the following user interface screen (Figure 8.3).

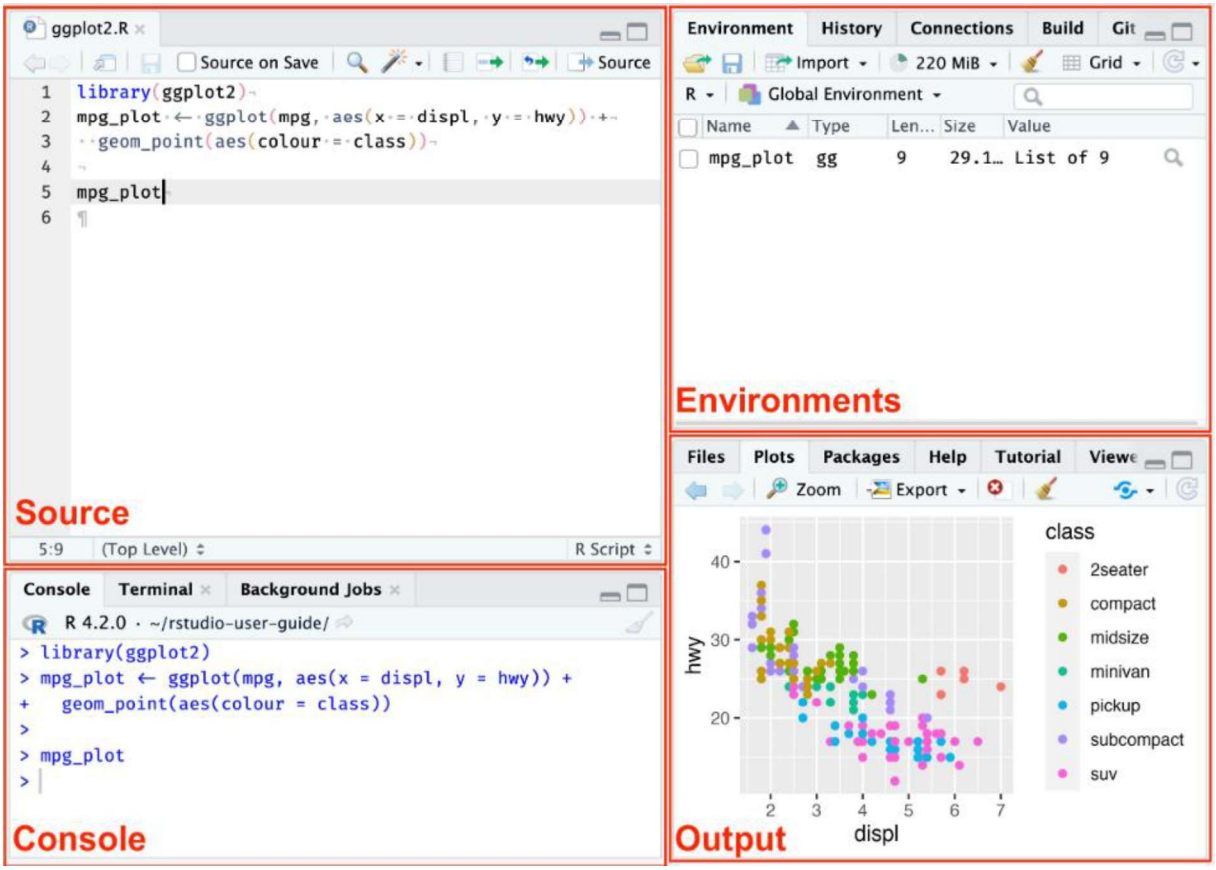

Figure 8.3 The user interface and the R & Rstudio (RStudio, 2024).

The RStudio user interface has 4 primary panes (RStudio, 2024):

- Source pane;
- Console pane;
- Environment pane, containing the Environment, History, Connections, Build, VCS, and Tutorial tabs;
- Output pane, containing the Files, Plots, Packages, Help, Viewer, and Presentation tabs.

Each of the panes and the subsection and options in it allow users to perform different operations, have control over some data analysis, or have a more structured and clear view of the data analytics process underway.

8.3 What is SQL and how is related to BA and R?

The Chinook Database is a sample database that is used for learning and demonstrating database management systems (DBMS) and SQL queries. It is designed as a digital media store using real data from an iTunes Library; fictitious names / addresses for customers and employees; and random data for the sales information's. The database contains a variety of tables that represent a music store's data, including information about artists, albums, tracks, customers, invoices, and more (Figure 8.4).

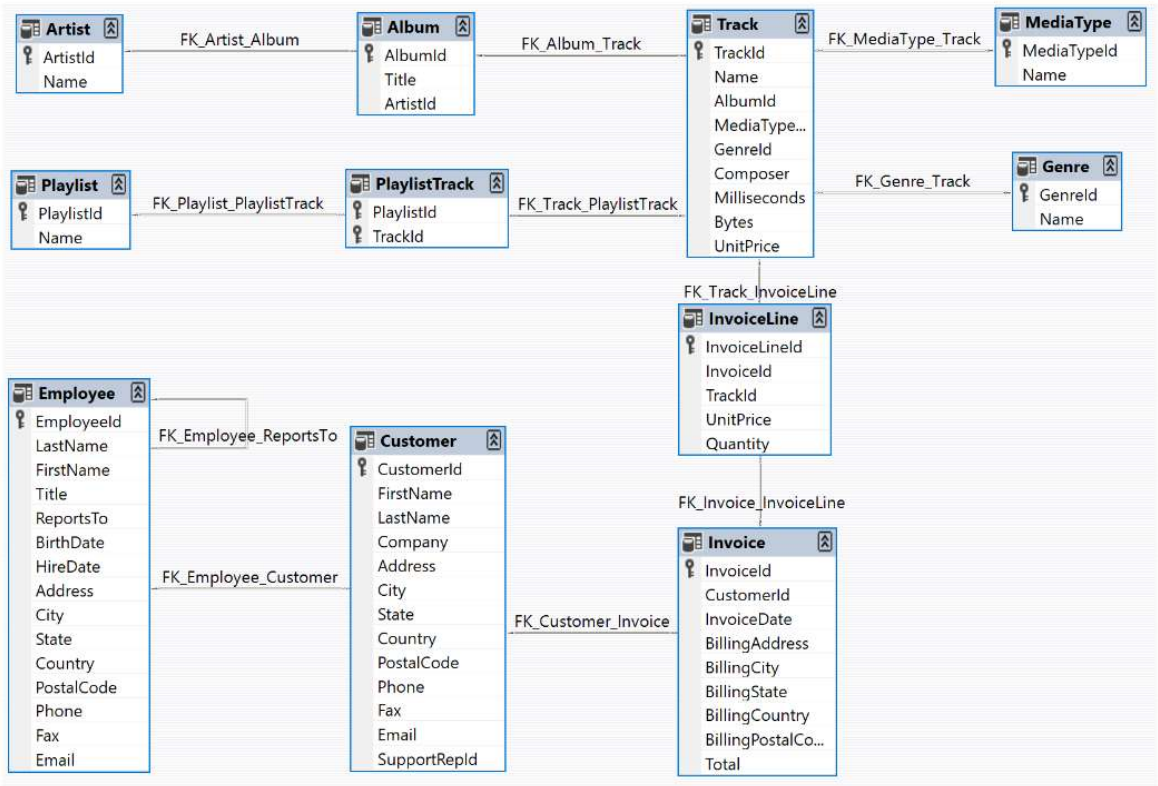

Figure 8.4 The data model of Chinook Database.

Figure 8.4 demonstrates the data model surrounding the Chinook database with the different data tables and their keys (unique identifiers) and joint tables like PlaylistTrack table. The tables convey different information's about the given digital store (Table 8.1).

Table 8.1 Information's contained in each of the tables of the Chinook Database.

Table Name	Description
Artist	Contains information about music artists.
Album	Contains information about music albums, each associated with an artist.

Track	Contains information about individual music tracks, including references to albums, media types, and genres.
Genre	Contains information about music genres.
MediaType	Contains information about different types of media (e.g., audio, video).
Customer	Contains information about customers, including contact details and support representative information.
Employee	Contains information about employees, including their roles, reports-to relationships, and contact details.
Invoice	Contains information about invoices, including customer details, billing information, and total amounts.
InvoiceLine	Contains detailed information about each item in an invoice, including references to tracks and quantities.
Playlist	Contains information about playlists.
PlaylistTrack	Links tracks to playlists, indicating which tracks are included in which playlists.

8.4 How are business analytics, SQL and R related?

The connection between BA, SQL and R is natural since all the business data should be stored in the SQL databases. This is still the idealistic goal since there is still poor data management in some fractions of the small and medium enterprises which still do not fully understand the power of the data. In large companies, this was recognized a long time ago and the data is properly structured in the databases (SQL or others, but usually in SQL). On the other hand, the analysis of the data can be performed in SQL, but for that purpose, it is better to use the statistically oriented software, where R comes in the focus, as one of the most popular statistical platforms to perform data analysis.

Accordingly, SQL and R can be seen as the perfect tool for collaboration when the problem at hand is from the BA area. There are several main reasons, and one of them is that the BA data is daily changing and updating according to the market reality and company activities: sales, employees, revenue, etc. The SQL databases are perfect for capturing those changes and updating existing data, while the R scripts are very good at automating tasks as well as designing new packages for analysing the given data. The reason for this is that the core R is more built around the concept of data analysing, than rather on general programming like Python for example.

Querying the SQL database with R

The R programming language and the SQL databases have a natural connection since the R is mainly built for statistical data analysis and the majority of the transactional data is found in databases. The "way" how the R operates to manage data manipulation from SQL databases is via DBI and RSQLite programming packages. The DBI package provides a standardized interface for interacting with various DBMS, allowing users to connect, query, and manage transactions consistently across different databases. The RSQLite package, which adheres to the DBI interface, specifically facilitates interaction with SQLite databases, enabling users to execute SQL queries, fetch data, and perform database operations directly from R. Together, these packages streamline the process of working with databases in R, offering a cohesive and efficient workflow.

To effectively demonstrate the performing the SQL operation from the R and generating desired insights from the data, regarding the problem on hand, we have provided several code snippets in Figures 8.5 and 8.6.

````
---
title: "BUSINES ANALYTICS FOUNDATINS INCLUDING THE R AND SQL"
format: html
editor: visual
---

# R & SQL

## Loading libraries

```{r setup, warning=FALSE, message=FALSE}
library(DBI)
library(RSQLite)
```

## Connect to the Chinook SQLite database

```{r}
con <- dbConnect(RSQLite::SQLite(), dbname = "Chinook_Sqlite.sqlite")
```

## List all tables in the database

```{r}
tables <- dbListTables(con)
print(tables)
```
````

Figure 8.5 The code snippet for establishing the connection between SQL & R and exploring the data tables contained in the SQL.

The first step in querying the SQL database via R is to establish the connection (Figure 8.5). The figure demonstrates the usage of DBI and RSQlite packages which enable establishing the connection via dbConect() function. The result of the connection and the data tables which are revealed via the aforementioned connection are then exported via dbListTables() functions which print a list of all the data found via the connection: Album, Artist, Customer, Employee, Genre, Invoice, InvoiceLine MediaType, Playlist, PlaylistTrack, Track.

After the connection is established, there are a number of possible analyses which can be performed, depending on a business goal and future usage of a given results. Here, due to the space restrictions, we will demonstrate only a fraction of possible data analysis, with a small code snippet and the set of code rules needed to extract the information from the SQL. The code performs database querying via R and displays the top-selling albums, their authors and the numbers sold (Figure 8.6). Table 8.2 represents the results of the data querying via code snippet in Figure 8.6.

```
## Choose a Album table from the database
```{r}
query_album <- "SELECT * FROM Album LIMIT 10"
data_album <- dbGetQuery(con, query_album)
print(data_album)
```

Query to get album details along with artist names
```{r}
query_album_artist <- "
SELECT Album.AlbumId, Album.Title AS AlbumTitle, Artist.Name AS ArtistName
FROM Album
JOIN Artist ON Album.ArtistId = Artist.ArtistId
LIMIT 10"

data_album_artist <- dbGetQuery(con, query_album_artist)
print(data_album_artist)
```

Query to get the top-selling albums along with artist names
```{r}
query_top_selling_albums <- "
SELECT
    Album.Title AS AlbumTitle,
    Artist.Name AS ArtistName,
    SUM(InvoiceLine.Quantity) AS TotalQuantitySold
FROM
    InvoiceLine
JOIN
    Track ON InvoiceLine.TrackId = Track.TrackId
JOIN
    Album ON Track.AlbumId = Album.AlbumId
JOIN
    Artist ON Album.ArtistId = Artist.ArtistId
GROUP BY
    Album.AlbumId, Album.Title, Artist.Name
ORDER BY
    TotalQuantitySold DESC
LIMIT 10"

# Execute the query
top_selling_albums <- dbGetQuery(con, query_top_selling_albums)
knitr::kable(top_selling_albums)
```
```

Figure 8.6 The code snippet for querying the SQL via R and determining the top 10 selling albums.

Table 8.2 The top 10 selling albums in the Chinook digital store.

Album title	Artist name	Quantity sold
Minha Historia	Chico Buarque	27
Greatest Hits	Lenny Kravitz	26
Unplugged	Eric Clapton	25
Acústico	Titãs	22
Greatest Kiss	Kiss	20
Prenda Minha	Caetano Veloso	19
Chronicle, Vol. 2	Creedence Clearwater Revival	19

Album title	Artist name	Quantity sold
My Generation - The Very Best Of The Who	The Who	19
International Superhits	Green Day	18
Chronicle, Vol. 1	Creedence Clearwater Revival	18

References Chapter 8

- Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. Decision Support Systems, 64, 130-141.
- Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging trends in business analytics. Communications of the ACM, 45(8), 45-48.
- Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data and business analytics: A research agenda for realizing business value. *Information & Management*. https://doi.org/10.1016/j.im.2019.103237
- Power, D. J., Heavin, C., McDermott, J., & Daly, M. (2018). Defining business analytics: An empirical approach. Journal of Decision Systems, 27(1), 40–53. https://doi.org/10.1080/2573234X.2018.1507605
- R Core Team. (2019). An introduction to R: Notes on R, a programming environment for data analysis and graphics. The R Foundation.
- RStudio. (2024). RStudio IDE cheatsheet: UI panes. Posit. https://docs.posit.co/ide/user/ide/guide/ui/ui-panes.html
- Torfs, P., & Brauer, C. (2014). A (very) short introduction to R. Hydrology and Quantitative Water Management Group, Wageningen University, The Netherlands, 1-12.