Data analysis project

Here we present the different stages of a data analysis project.

1 Data analysis project

A data analysis project can be divided into five main stages:

1. Defining objectives
2. Data
3. Developing and validating models
4. Implementation
5. Performance monitoring and improvement

When planning a project, it is important to bear in mind that each stage has a different level of importance, but also that each stage takes a different amount of time to complete. Pyle (1999) provides an estimate of the time required for each stage, as well as their importance in the success of the project (given as a percentage of the total, see Table 1).

Table 1: Breakdown of a data analysis project.

Stage	Time	Importance
Understanding the problem	10%	15%
Exploring the solution	9%	14%
Implementing the solution	1%	51%
Prepare the data	60%	15%
Analyze the data	15%	3%
Model the data	5%	2%

Two important facts stand out: just because a step is very important does not mean it will take a long time. Implementing the solution is very important (otherwise there will be no result), but it will not usually take very long to do (possibly just a few lines of code). Conversely, data preparation is a stage of medium importance (although this is debatable), but it takes up most of the project time. For example, it is necessary to manage missing data, outliers, possible accents for French data, etc.

2 Setting objectives

Do we want to: visualize data? explore and make assumptions? test? group? understand? predict?

How do we do this in practice? We ask questions! First, we need to clarify the terms. Who will use the model and how? What is the target population?

Why is it important to have clear objectives when undertaking a data analysis project? It helps guide data collection and formatting. It helps define an appropriate model (e.g., classification vs. prediction). It allows you to analyze the results in light of the objective and thus allow others to judge their relevance. It is important to define the objectives before looking at the data so as not to be biased by it.

Example

The National Bank of Canada would like to launch a new savings product and is giving you access to its customer database.
Bad objective: Analyze the customer database.
Better objective: Can you predict which customers will buy the new savings product?

Example

The Montreal Canadiens hockey team wants to learn more about its opponents in order to develop new game tactics.
Bad objective: Analyze the opponents' data.
Better objective: Can you characterize the opponents' playing style in order to identify weaknesses?

Example

Pharmascience wants to know if its new drug is effective.
Bad objective: Analyze the drug data.
Better objective: Can you determine a testing protocol (statistics) to determine if the drug is effective?

3 Data

Data is at the heart of the matter. To be useful, data must be available and of good quality. Once the objectives have been defined, preliminary processing and basic exploration of the data is carried out, followed by more developed models.

3.1 Where can data be found?

The simple answer is: on the Internet! Here is a non-exhaustive list of websites that collect data sets:

- Google datasets;
- Kaggle;
- UC Irvine Machine Learning Repository;
- Time Series Machine Learning website;
- Physionet Database.

You can also check out the official websites of data sources that can be found for most countries around the world:

- Canada: StatCan;
- France: data.gouv.fr;
- USA: data.gov;
- England: data.gouv.uk;
- etc.

For data on more specific topics, you can look at the different branches of government. For example, the Canadian Centre for Mapping and Earth Observation provides geospatial data for Canada (here).

When working for a company, you generally have access to internal data sources, e.g. databases on production, customers and employees, lists of transactions and potential customers, information on website visits.

3.2 Quality

There is a popular saying in computer science that also applies to data analysis: "Garbage in, garbage out." Basically, this means that no matter how sophisticated the model is, if the input data is poor quality, biased, incomplete, etc., then the output results will be poor quality. This ensures a certain level of credibility, reproducibility, and usability for our conclusions.

But what do we mean by data quality? To ensure data quality, we can ask ourselves the following questions:

- Is the data representative of the target population?
- Is the data correct and relevant?
- Is there any missing or redundant data?

3.3 Building the database

Once our data has been retrieved, we need to load it into memory so that we can then perform analyses. In Python, the `pandas` library can read most of the file formats we will be dealing with. In R, different libraries need to be used to load different types of data (see Table 2).

Table 2: Different libraries for different file formats.

Format	Extension	Library
Text	.txt; .csv	`readr`
Excel	.xlsx	`readxl`
SAS	.sas7bdat	`haven`
SPSS	.sav; .zsav	`haven`
JSON	.json	`jsonlite`

Over the past decade, the concept of "tidy data" has emerged (see Wickham (2014)). Each "tidy" dataset follows three principles:

1. Each variable is a column in the table.
2. Each observation is a row in the table.
3. Each cell in the table corresponds to a single measurement.

This allows for a unified approach to data analysis. In general, we will always try to put our dataset in "tidy" format. The `tidyr` package in R and the `pandas` library in Python allow you to format data in "tidy" format.

3.4 Exploration and preliminary processing

Once the data has been loaded and formatted in "tidy" format, we perform an initial exploration of the data before moving on to the model development stage itself. Although this stage is **very important**, it is not the focus of this course. Here are a few tips for this initial exploration:

- Data cleaning: remove duplicates, standardize modalities, check the format of special values, etc.
- Data exploration: rare modalities, too many modalities, asymmetry, class imbalance, extreme or outlier values, highly correlated variables, missing values.

4 Model development and validation

This course focuses on model development and validation. For now, we can identify four main components for this:

1. a (mathematical) **space** in which to work;
2. a **distance** for comparing observations;
3. a **model** (or algorithm);
4. a **function** for measuring the quality of the model.

We will look at each of these components in detail in the following sections.

5 Implementation

Once the model has been chosen and validated, we may want to **put it into production**. Putting a model into production means making it available to as many people as possible. Generally, this involves automating data collection and cleaning, then "feeding" the created model and producing analysis reports of the results. This part is called **data engineering**. A **data engineer** is therefore responsible for setting up the data processing pipeline, from data collection to model output.

6 Performance monitoring and improvement

Finally, once the model is in production, its performance must be monitored. When new data arrives, the model that was considered may no longer be very suitable (e.g., the assumptions made are no longer correct). To prevent this from happening, the model is **monitored** by regularly checking its performance. You may also want to improve your model, for example because you have more accurate assumptions or better quality data.

Pyle, Dorian. 1999. *Data Preparation for Data Mining*. Morgan Kaufmann.
Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (September): 1–23. https://doi.org/10.18637/jss.v059.i10.