Introduction to Data Analysis and Key Concepts

Contents

In this lecture, we will consider the main data analysis concepts.

3.1. What is Data? An informal Definition

We will start by giving an informal definition of data. It is not by any means a complete definition, but it will suffice as a starting point to get an intuitive understanding of what data is:

> Data is a **set of values collected** with respect to some **variables** which describe a given **phenomenon**.

In the definition above, we mentioned some key concepts, which are discussed in the following sections.

3.2. Observations, Populations, Samples

We will see three fundamental concepts: observations, populations, samples.

3.2.1. Observations

When we deal with data, we actually deal with *multiple instances* of values associated to the same event or phenomenon. Examples:

Example	Comment
We want to study how the heights of students change with the years	We need sets of recorded heights, not just one height
We want to study how a given drug affects the recovering from a disease	We need to record sets of values about drug assumption and recover, just one case will not be enough
We need to create a system which can generate text from a prompt	We need to record several examples of prompt-text generation, just one would not be enough to study how such generation should be made

We will call observations, the units by which we measure data. These could be persons, cars, animals, plants, etc. We often indicate an observation as x. Please consider these as "abstract" entities, not necessarily numerical observations. E.g., "let's consider a person x".

3.2.2. Population

When we study a given phenomenon, we will be interested in a set of observations, which is called a "population". For instance:

- if we want to study the distribution of heights of people in the world, we will

need to look at the population of all people in the world.
- if we want to study the age of people attending a computer science course in Italy, then we need to look at the population of all students of computer science courses in Italy.

Note that a population can sometimes be a theoretical concept and identify sets of elements which are not even finite. E.g., *"all movies which will ever be filmed"*.

We can denote a population with the symbol Ω. Al our observations will be $\omega \in \Omega$.

3.2.3. Sample

In practice, working with population can be very hard, as it is not always possible to obtain observations from those large sets. Intuitively, in practice, working on a large enough set of observations from a population could be good enough. We refer to a subset of a population as "a sample": $\{\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(n)}\} \subseteq \Omega$.

3.2.4. Example

We want to study how the height of people in the world changed in the years. Here the **population** Ω is the set of people ever existed on earth, a **sample** $\{\omega^{(1)}, \ldots, \omega^{(n)}\} \subseteq \Omega$ is a subset of people for which we have some data (e.g., say we measured height of many people in most countries since the 16th century). An observation is a person $\omega^{(i)}$.

3.3. Variables

We have identified our problem, defined a suitable population and identified a sample of observations. While observations are abstract concepts such as "a person" or "a student", we usually want to capture specific features of such observations, such as "the person's age" or "the student's height". We collect these features by means of **statistical variables**.

Statistical variables are concept similar to that of mathematical of programming

variables, in that they can be seen as sort of “containers for the data”.

We may also be interested in different features of an observation. For instance, for each person in a population, we may want to record their age, gender, and height. We can introduce a variable to capture each of these features. For instance, given observation ω, we may obtain “height = 180cm”, “weight=80Kg”, “gender=male”.

Formally, we’ll define a variable X as:

$$X : \Omega \to S$$

$$\omega \mapsto x$$

Where S is the set of possible values for variable X. The definition above specifies that a variable maps an abstract observation ω to some (possibly more concrete) value $x \in S$.

3.3.1. Example

Given the population of all people currently living in the world Ω, we define a variable H to collect the heights of the observed people ω:

$$H : \Omega \to \mathbb{R}$$

$$\omega \mapsto h$$

Given an observation $\omega^{(1)}$, we may obtain $H(\omega) = 180$. We often say that H assumes the value 180 and write: $H = 180$.

3.3.2. Quantitative and Qualitative

Variables can be of two main kinds:

- **Qualitative**: represent properties that can be observed and cannot generally be measured with a numerical value (e.g., ‘color’, ‘sex’);

- **Quantitative**: represent properties that can be quantified with a number (e.g., ‘height’, ‘weight’, ‘age’).

3.3.3. Discrete and Continuous

Variables can also be discrete or continues:

- **Discrete variables** can assume a finite number (or a countable infinite number) of possible values
- **Continuous variable** assume a continuous, infinite number of values, which can be generally denoted with real numbers

3.3.4. Scalar and Multi-Dimensional

Variables can be:

- **scalar** or uni-dimensional: they assume real numbers (e.g., $X = 1$)
- **multi-dimensional**: they assume vector or matrix values, (e.g., $X = \binom{1}{3}$, or $X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$).

3.3.5. Examples

Let’s see some examples:

3.3.5.1. Discrete Scalar Variables

We want to assess if a coin is fair or not.

- We consider as population all possible tosses of that coin.
- An observation will be a specific tossing.
- A discrete scalar variable X may record the outcome of a given tossing. The set of possible values will be $S = \{head,\ tail\}$ (discrete values). The

variable is scalar as it will contain a single value.
- If we toss a coin, we may get $X = tail$.

3.3.5.2. Continuous Scalar Variables

We want to study the heights in centimeters of students in this class.

- Our population is the set of all students in this class.
- We can use a continuous scalar variable X to record the heights of the students. In this case, we can choose $S = \mathbb{R}_+$.
- If we pick a student, we may get $X = 175$.

3.3.5.3. Continuous Multi-Dimensional Variables

We want to study the positions of all cars in the world.

- Our population is the set of all cars in the world.
- We could use the variable X to denote the latitude and longitude coordinates of a car in the world. The set of possible values may be $S = \mathbb{R}^2$.
- Once we pick a car, we may have $X = \binom{37}{15}$.

3.3.6. Scales

The values of the variables can sometimes be ordered. For instance, we can say that a student is taller than another one. In other cases, an order is not possible. For instance, it may not be meaningful to order students by color of the eyes or gender. We can classify whether and how a variable can be sorted using scales. Variables can be characterized with respect to the following scales:

- **Nominal scale**: nominal variables cannot be ordered. Examples are the gender of a person and the color of the eyes.
- **Ordinal scale**: ordinal variables can be ordered, but the difference between two ordinal variables is usually not meaningful. For instance, we may have the

following scale to classify the level of expertise of a basketball player: *novice, amateur, intermediate, expert*. While this scale allows to meaningful sort players, the difference between two levels may not be meaningful.

- **Continuous scale**: continuous variables are generally expressed with real numbers and can be ordered. For instance, the *height of people* and the *amount of money in the bank account* are examples of continuous variables.

3.4. Data Collection

Data collection is the crucial first step in the data analysis process. It involves gathering observations that will serve as the foundation for all subsequent analysis. The methods of data collection can vary widely depending on the goals of the analysis, the type of data required, and available resources.

3.4.1. Survey

Surveys are a common method of data collection used to gather information from individuals or groups. They involve asking specific questions to respondents, often in a structured format. Surveys can be conducted through various means, such as paper questionnaires, online forms, telephone interviews, or in-person interviews. Surveys are particularly useful when seeking opinions, attitudes, preferences, or demographic information from a target population.

Key characteristics of surveys:

- Structured and standardized questions.
- Can be administered in various formats.
- Allow for quantifiable responses.

3.4.2. Experiments

Another way to collect data is through experiments. These allow to collect data in a **controlled** way and they are usually employed to establish cause-and-effect

relationships between variables. In an experiment, researchers manipulate one or more independent variables to observe their effects on dependent variables. Experiments are characterized by their ability to provide evidence of causation, which is a significant advantage in many scientific and practical applications.

The most common form of experiments for data collection are **randomized controlled experiments** (or randomized controlled trials), which have the following characteristics:

- Controlled manipulation of variables.
- Random assignment of participants (in randomized experiments).
- Replicability to test hypotheses.

3.4.2.1. Randomized Controlled Experiments - Example

We want to test whether a given drug is effective in treating a disease. We will consider a suitable population (e.g., all people affected by the disease), obtain a suitable sample (a random set of people, diverse by age, gender and health conditions) and set up an experiment in which half of the people take the drug, while the remaining half take a placebo. The assignment are **randomized** to avoid any bias in the data. We will discuss this better when we'll talk of causal inference.

3.4.3. Observational Data

Observational data collection involves the passive recording of information as it naturally occurs, without any interference or manipulation by the researcher. This method is commonly used when experiments are not feasible, ethical, or practical. For instance, if we want to affect the effects of smoke on young people, it would be unethical to ask people to smoke. Observational studies can be valuable for exploring patterns, relationships, and behaviors in real-world settings.

Key characteristics of observational data:

- No direct manipulation of variables.

- Captures data as it naturally occurs.
- Useful for studying complex, uncontrolled environments.

3.5. Data Sets, Design Matrix, and Features

A set of data related to a specific phenomenon is called a “dataset”. Datasets are usually stored in tables in which **columns represent the different variables** and **rows represent the different observations**. If you have ever had a look at a spreadsheet, you probably already saw an example of a dataset!

Let’s consider the following example of a dataset of marks of 5 subjects obtained by three students:

	Maths	Geography	English	Physics	Chemistry
x001	8	9	30	8	10
x038	9	7	27	6	
x002	6	-1	18	5	6
x012	7	7	25	4	10
x042	10	10	30	10	10

Each row is an **observation** related to a given student, while each column represents a different **variable** (the marks obtained in the different courses).

Before moving on, think for a moment what you could do with a dataset like this (maybe imagine a larger one):

- You could take the average of all marks obtained by a student (average by rows) to get a ranking of the students. This could be useful to understand which students may need help.
- You could compute the average of the votes obtained in each course (average by column) to identify the subjects which are “more difficult” for the students

than others.

- You could group the courses into humanity-based and science-based to identify which students excel in each field.

The examples above are all (very simple) examples of data analysis. As you can see, even with a simple dataset like this and no knowledge of complex notions of data analysis, we can already do a lot of analysis.

The "table" structure containing a dataset is often called a **data matrix** or a **design matrix**. In this format, each column of the matrix represents a variable, which is often also referred to as a **feature**, while each row is a different observation.

Let's have a look at a real, but simple, dataset we'll use often in the future, the `Titanic` dataset, which like like this:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0
5	0	3	Allen, Mr. William Henry	male	35.0	0	0
...	...	...	...	...	...	...	...

887	0	2	Montvila, Rev. Juozas	male	27.0	0	0
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0
891	0	3	Dooley, Mr. Patrick	male	32.0	0	0

891 rows × 11 columns

The dataset shows different observations related to the passengers of the Titanic. We can see each row of this matrix as a multi-dimensional variable (a vector) summarizing the main properties of the observations. Indeed, in the example above, each passenger is a different observation, while the different column are the different **features**.

3.6. The Data Analysis Workflow

Data Analysis can be defined following this definition adapted from wikipedia (https://en.wikipedia.org/wiki/Data_analysis):

> Data analysis is the **process** of inspecting, cleaning, transforming, and modeling data with the goal of **understanding** something about a given phenomenon, supporting **decision-making**, and **making predictions on unseen data**.

While this definition is not very formal, it is a good starting point to get an understanding of what data analysis is. Let's dive into the main aspects of this definition in the following sections.

3.6.1. Data Analysis as a Process

In the definition above, data analysis is regarded to as the process of inspecting, cleaning, transforming and modeling data. Indeed, it is very important to understand that **data analysis is a process**. In this sense, data analysis is not just a single algoritm or a single technique you can apply to the data. It is instead a collection of techniques, statistical and machine learning tools can be applied to achieve a given goal. Four main set of techniques that we can usually apply when processing data are:

- **Inspecting**: the process of looking at the data to assess some of its main properties, such as the number of observations (rows), variables (columns), what are the typical values of variables (e.g., in the previous examples of students and courses we should expects marks out of 10), etc.
- **Cleaning**: is the process of "fixing" some aspects of the data which are likely to be incorrect. For instance we could remove rows with missing or out of range values (students with marks over 10 or with negative marks).
- **Transforming**: transforming the data from a format to another one. For instance, we may add a new "mean" column to each student in the example above, or, after realizing that English marks are out of 30 while the others are out of 10, transforming that column by dividing all numbers by 3 (so that they are in the same range as other variables).
- **Modeling**: choosing and tuning a statistical model which can be used to summarize, explain or make predictions about the data. We will discuss more models later, but for the moment it is important to understand that a model tries to abstract away and represent such aspects of the data in a way that, in some sense, the model becomes a good explanation of the data itself. For instance, if we could create a model which can allow us to reconstruct the correct value of English from the values of the other courses, we have found some kind of mathematical explanation of how marks in other course affect

marks in English and we have shown that marks in English are not independent from the other marks.

3.6.2. Goals of Data Analysis

We can simplify (as in the spirit of this introductory material) by saying that data analysis processes can aim to address three main types of goals:

- Understanding something about a given phenomenon;
- Supporting decision-making;
- Making predictions on unseen data.

We'll see some examples of each of these goals in the following sections.

3.6.3. Main Types of Data Analysis

While we have discussed of broad goals of data analysis above, we can divide data analysis processes into five main types (different categorization may be possible, but we will stick to this one in this course). Each category describes a *mindset* and a *set of tools* that can be used to perform data analysis processes towards the accomplishment of one or more of the goals above. It should be clear that, during a data analysis process, one may mix the different kinds of data analysis or perform them in a sequential fashion. As we will see, the different approaches also rely on similar statistical techniques and one approach does not exclude the others.

Popular types of data analysis are:

- Descriptive and analysis;
- Exploratory analysis;
- Inferential analysis;
- Predictive analysis.

We will analyze each of these approaches during the course.

3.6.4. Data Analysis Workflow

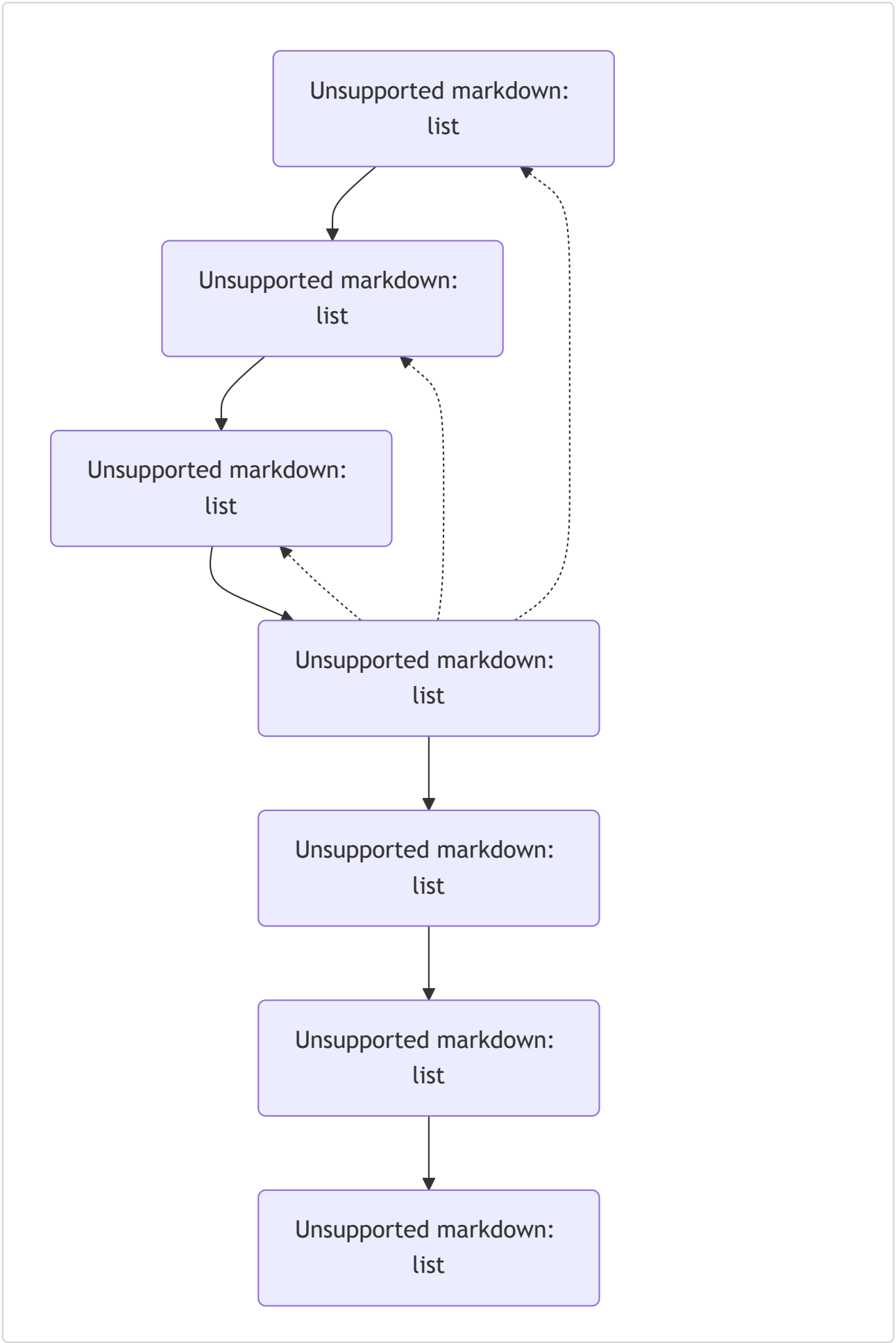
The typical workflow of data analysis is composed of the following stages:

1. Define your data analysis question
2. Collect the data needed to answer the data analysis question
3. Clean and format the data
4. Explore and describe the data
5. Choose suitable models for the analysis
6. Fit, fine-tune, evaluate, and compare the models for the considered data analysis
7. Review the data analysis when new data is available

3.6.4.1. Non-Linear Workflow

It should be noted that the process is not exactly linear. A good data analysis will iterate over the different steps and possibly jump back to any of those steps to revise it. For instance, after fitting the models on the data in step 5, one may note that some other models could give better results, and jump back to step 4 to refine the choice of the models, to then return to step 5.

The figure below shows the described workflow. Solid arrows illustrate the main flow, while dashed arrows show possible alternative paths after performing data exploration (step 4). Similar dashed arrows pointing to any past node may apply to the other nodes as well. In the example below, after exploratory data analysis (step 4), we may note that the data is not clean (maybe we find some outliers) or not adequately formatted, so we jump back to step 3. Similarly, we may find that we need more data, and jump back to step 2, or we may want to revisit and refine our data analysis question (step 1).


Unsupported markdown: list
Unsupported markdown: list
Unsupported markdown: list
Unsupported markdown: list
Unsupported markdown: list
Unsupported markdown: list
Unsupported markdown: list

3.6.4.2. Example

A data analyst working for an e-commerce company is tasked with analyzing customer reviews to improve product quality and customer satisfaction. The analyst starts by defining the question that the data analysis needs to answer (**step 1 - define your data analysis question**): "What are the common themes and issues in customer reviews, and how can we address them to improve product quality and satisfaction?" To answer this question, the analyst collects a diverse dataset of customer reviews from various sources (**step 2 - collect the data needed**), including the company's website, social media, and third-party review platforms. This dataset includes text reviews, ratings, and timestamps.

Once the data is collected, the analyst proceeds with data cleaning and formatting (**step 3 - clean and format the data**) since the initial dataset is messy with spelling errors, duplicate reviews, and inconsistent formatting. This involves data preprocessing, such as removing punctuation and converting text to lowercase. After cleaning the data, the analyst explores and describes it (**step 4 - explore and describe the data**) using techniques like word clouds, frequency distributions, and sentiment analysis to uncover common words, sentiments, and trends within the reviews.

To extract more meaningful insights from the text data, the analyst decides to apply natural language processing (NLP) techniques (**step 5 - choose suitable models for analysis**). Specifically, Latent Dirichlet Allocation (LDA) for topic modeling is chosen to identify recurring themes within the reviews. However, upon implementing LDA and reviewing the results, the analyst realizes that some topics are unclear and overlapping (**step 6 - fit, fine-tune, evaluate, and compare the models**).

Acknowledging the need to revisit the analysis, the analyst decides to backtrack to the data exploration step (**step 4 - explore and describe the data**) to gain a deeper understanding of the customer feedback. During this reassessment, it becomes evident that customers frequently mention product quality issues when discussing customer service experiences, leading to topic overlap. Armed with this insight, the analyst decides to categorize reviews into "product-related" and "service-related" (**step 5 - choose suitable models for analysis**) before applying

topic modeling.

With the revised approach, the analyst proceeds to categorize reviews and then applies LDA separately to the two categories (**step 6 - Fit, fine-tune, evaluate, and compare the models for the considered data analysis**). This results in more interpretable topics, such as “defective products,” “timely shipping,” and “responsive customer support.” The analyst then analyzes these topics to generate actionable insights for product improvement and customer service enhancements.

To ensure continuous improvement, the analyst commits to ongoing review and monitoring (**step 7 - Review the data analysis when new data is available**) of customer reviews, periodically updating the analysis to identify new emerging issues or trends. This iterative approach to data analysis allows for refining strategies and addressing evolving customer concerns over time.

3.7. References

- Chapter 1 of *Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.*

Previous
2. Introduzione a Python

Next
4. Probability for Data Manipulation