DIGITAL CULTURE
ITMO UNIVERSITY

Introduction to Data Analysis

High School of Digital Culture
ITMO University
dc@itmo.ru

Contents

1 Basic Concepts of Data Analysis

1.1 What is Data Analysis?

We live in an era where technology keeps advancing and the volume of data is growing inexorably. Some studies claim that data will grow to 400 zettabytes by 2025. The key is that all the collected data contains important information. The ability to find and analyze the required information is useful, both at work and in everyday life.

Data analysis is used in many branches of science, industry, and services. Data analysis helps companies to make the right business decisions and grow. To paraphrase the famous Rothschild's saying *those who own information rule the world*, we say: *those who know how the information is structured and have the tools for its processing rule the world.* So, when had it started?

Since the beginning of time, people have tried to describe the world they were living in. Some used poetry, prose, or art to express their thoughts, while others described the world around us using numbers.

To apply numerical descriptions in practice, it's essential to understand the rules and logic that lie behind them. Numerical descriptions can be good or not good at all (recall, for example, the well-known expression 'the average temperature in a hospital'). In less obvious cases, it's hard to distinguish a good description from bad. We hope that this course will help you to understand numerical descriptions of the real world and to assess them critically, as well as to create them yourself and apply them correctly. Some descriptions will be rather obvious and simple to understand, while others will require the use of mathematical concepts of statistics, probability theory, and the like. However, once you're armed with these tools for description and analysis, you will be rewarded with the opportunity to acquire nontrivial knowledge about the world around us and the patterns behind it.

Data analysis is a science that combines methods for collecting, structuring, presenting, generalizing, analyzing, and interpreting data (in other words, it allows making conclusions based on a studied phenomenon related to the data).

1.2 Basic Concepts

Today, we are going to look at possible objects of study in data analysis, find out what objects are described with different attributes (variables), define population and sample, as well as to review the steps in solving a data-analysis problem.

1.2.1 Variables

A variable (also called an attribute) is a common characteristic of all the studied objects or a property the instances of which may differ for different objects. The instances of an attribute are called values, alternatives, or gradations.

The ability to think in attributes and identify variables for a research purpose is one of the most important qualities of a good analyst. Look at the variables and their possible values. The following values correspond to the variable profession: analyst, programmer, cook, manager, teacher, doctor. The variable size takes the values 5 meters, 7 meters, 100 kilometers.

Figure 1: Examples of variables

1.2.2 Value Distribution of Variables

Data analysis studies values that a variable takes given the described objects and phenomena. That's why we want to find the rate of occurrence of various values that each variable takes. This rate is called the distribution of a variable. It's expressed as an absolute value or a percentage. The distribution depends on the dataset for which it's calculated. For example, Figure 2 shows the distribution of the variable "animals" in two zoos. Pie charts show both absolute values of the variable distribution and percentages. The distribution of animals in the zoos clearly differs.

1.2.3 Population and Sample

Analysts can rarely analyze the properties of the entire set of the studied objects. Assume the task is to carry out a poll in a large city. It's not feasible to get information about the opinion of each citizen. The realistic approach is to collect the opinions of 5,000 citizens and use the findings to draw conclusions on public opinion in the city. It leads us to the concepts of population and sample.

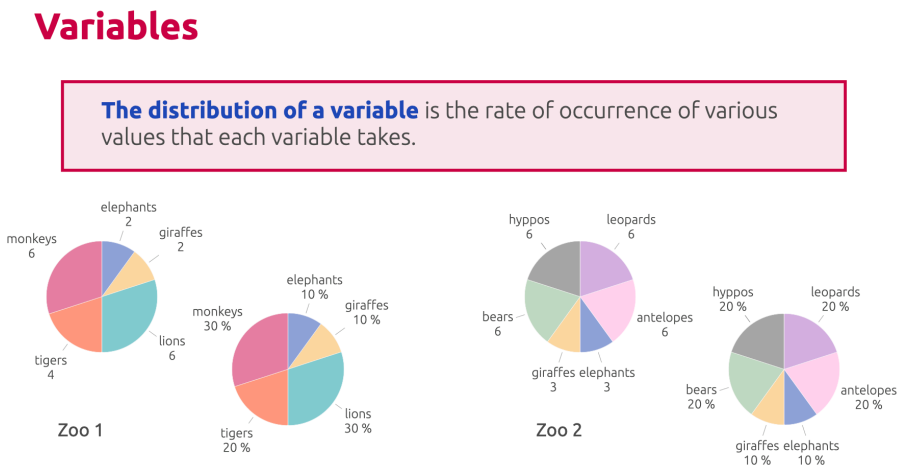


Figure 2: An example of value distribution of variables

- A **population** is the entire group of objects that the researcher studies.
- A **sample** is often a small portion of the whole population. Samples are selected in a particular way to study the properties of the population.

When researchers use the term sample, they are referring to a representative sample. It's a sample the variable values in which are distributed approximately just like in the population. That's why researchers use special methods to build representative samples for opinion polls.

1.2.4 Hypothesis

Another important concept in data analysis is a hypothesis. A **hypothesis** is an assumption about the values of the variables in the population (it can be made after a sample analysis).

The figure shows examples of hypotheses. The assumption that electricity consumption depends on the season and day of the week is a hypothesis. Another hypothesis is that the evening rush hour is the period from 6 to 7 p.m. on weekdays.

A weather forecast is also a hypothesis that requires analyzing the available data on current weather and weather conditions.

1.3 Steps of Data Analysis Process

The steps of the data analysis process depend on the features of studied objects and phenomena. However, some steps are the same for any dataset analysis. They include, for example, the process of describing studied objects, identifying variables, collecting and preprocessing data, and so on. Let's look closely at each step.

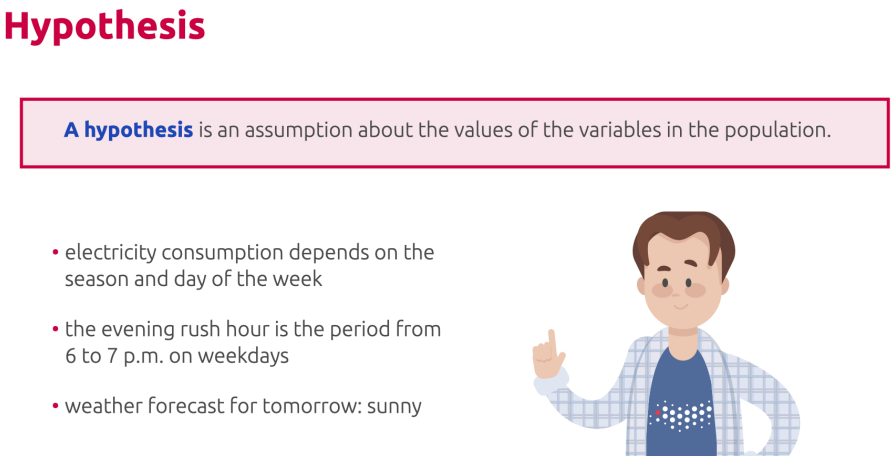


Figure 3: Examples of hypotheses

1.3.1 Describing Studied Objects

Describing studied objects. This step aims to prepare a theoretic description of studied objects or phenomena. It often requires professional services. A well-written description will contribute to the next steps of analysis.

1.3.2 Identifying Variables and Formulating Hypotheses

The next step is to identify variables, formulate hypotheses, and operationalize the concepts. We transition from abstract concepts of the domain to variables to measure them both quantitatively and qualitatively. At this point, we also formulate the hypotheses in terms of variables.

1.3.3 Collecting and Preparing Data

The next step is to collect and prepare data for hypothesis testing. Ways of data collection include opinion polls, measurements, external sources, and so on. At this point, we collect the entire studied population (it's technically possible at times) or build a representative sample after a selective inquiry.

This step also assumes data preprocessing before the data is loaded into a storage. Before loading into the storage, the data is also carefully preprocessed, storage structures are created, missing values are added, credibility is checked, and so on.

Starting from this step, analysts should have a tool for loading the data into the storage and preparing it for further analysis.

1.3.4 Exploratory Data Analysis

The next step consists of exploratory data analysis. Exploratory data analysis is an approach of analyzing the main properties of data, finding patterns,

distributions, anomalies in them, and, if possible, building initial models. Analysts don't have a strong opinion on what is "explored" in this step.

The term exploration was coined by mathematician John Tukey who formulated *the goals of exploratory data analysis*:

- Immerse in the data
- Reveal basic structures
- Select the most important variables
- Detect outliers and anomalies
- Test basic hypotheses

So, this step often reduces to visualizing the prepared dataset, describing it with descriptive statistics, and finding outliers in the data. This step may reveal the need to clean and converse the data even further. Sometimes, even these simplest methods of analysis produce impressive results that confirm the hypotheses.

1.3.5 Data Cleaning: Removing Noise and Anomalies

The next common step is to remove noise and anomalies from data. In practice, data for analysis is rarely of good quality. Exploratory data analysis detects noise and various anomalies in the data. Noise and anomalies may distort the overall picture of the patterns in the analyzed data. This step aims to remove noise and anomalies.

1.3.6 Data Conversion

Data conversion. Exploratory data analysis may show that some variables have too distinct values. In this case, they are not only difficult to analyze but also won't be visualized accurately. However, the simplest mathematical transformations easily solve this problem. Moreover, some analysis algorithms require not only the commensurability of variables in scale but also the so-called normalization of the variables, so that variable values fall within a given range of possible values (for example, from 0 to 1). We will discuss the methods developed for this transformation later in this course.

1.3.7 Building Models

Building models. This step is for testing the hypotheses and building mathematical models that describe the behavior of the variables and the relationships between them.

1.3.8 Interpretation

The final step is interpretation. Interpretation is a process of transforming data into meaningful information. Interpretation depends on many factors. For example, it matters who interprets the data, what information the interpreter has had, from what point of view the interpreter examines the provided data, and so on.

Interpretation of the created models is made by one person or a group of people. The process can be informal, creative, or formal (metric-based). The developed models represent formalized expert knowledge that can be, and should be, reproduced.

2 Measurements and Scales

As has been noted, data analysis represents real-world objects as variables. Variables have corresponding values. To match a value with a particular variable, one takes a measurement according to a rule.

In practice, it's necessary to measure or classify various values that characterize the properties of objects, phenomena, and processes. Some properties are only qualitative while others are quantitative. Variable measuring is closely connected to the concept of a measurement scale.

A measurement scale is a set of numerical or symbolic values, reflecting the permissible variations in the values of the measured object.

There are four main types of measurement scales classified according to the logical structure of the properties:

- Nominal
- Ordinal
- Interval
- Ratio

2.1 Nominal Scale

A nominal scale that consists of names, titles, or categories classifies objects and phenomena on a certain basis. Examples of a nominal scale include animal species, marital status, occupation, and academic degree.

A variable is dichotomous if it has exactly two possible values, for example, yes/no, 0/1, true/false, or know/do not know. It's a special case of a nominal scale. A good example of nominal data is names of supermarket sections. Here are a few examples of dichotomous data: the variable "goods availability" taking

two possible values (available or unavailable), the variable "statement" that turns into true or false, and the variable "married" taking two possible values (yes or no).

Nominal data aggregation. Since the only characteristic of nominal data is its belonging to a particular class of objects, nominal data is not related to the concept of zero, a unit of measurement, or the ability to compare the objects by their order. And yet you can calculate the number of different nominal values and a percent of the whole, but you cannot calculate the average. For example, we can find out how many alumni, such as programmers, mathematicians, physicists, chemists, geologists, and philologists, a university has, but we cannot obtain the average of all the professionals. Calculations done on these numbers are useless as they have no quantitative significance.

2.2 Ordinal Scale

An ordinal scale uses numbers (or an ordered set of text characteristics) to depict relative qualitative positions of the objects. The ordinal scale allows us to classify objects and compare the qualitative characteristics of different objects.

A good example of an ordinal scale is the 5-star hotel rating. According to it, a four-star hotel is better than that with three stars, but it's not clear to what extent.

Ordinal data aggregation. The well-known example in the figure shows the Beaufort scale that estimates wind force at sea. It has 13 categories from 0 to 12. Each category corresponds to points that characterize the force of sea wind. For example, category-9 winds (strong gales) are stronger than those of category 0 (calm winds).

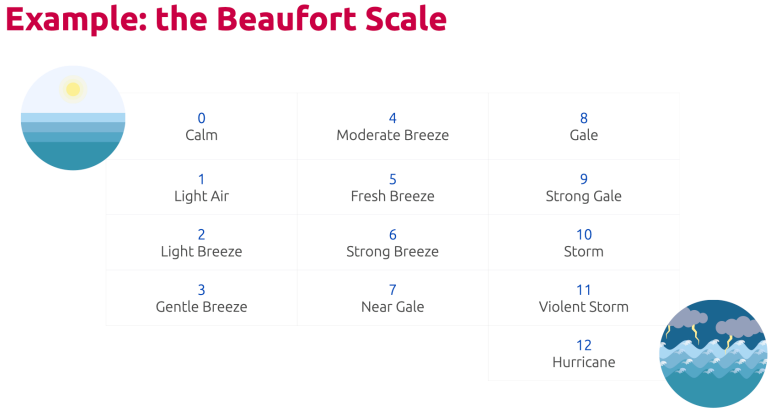


Figure 4: An example of ordinal data

It's possible to calculate the number of ordinal values and the percent of the whole, but different views exist as to whether it is possible to calculate an average for ordinal data. On the one side, it's impossible to determine the average of the variable "wind force at sea". Even with a formal numerical value, it won't have

an actual meaning. On the other side, some studies show that the difference in values between consecutive categories is approximately the same.

For example, if possible answers 1, 2, 3, 4, and 5 in a poll correspond to the responses strongly disagree, disagree, neither agree nor disagree, agree, strongly agree, we assume that the difference between the answers on this scale is approximately the same.

Sociologists use it to calculate the average for such answers and reasonably interpret them. Such aggregation heavily depends on the characteristics of the studied objects.

The use of ordinal data in such calculations is considered a bad practice in some fields while being totally acceptable in others.

The figure illustrates the attempt to calculate aggregated values for the wind force variables and the poll results. The quantity can be calculated both in the first and second cases, but the average can only be found for an opinion poll.

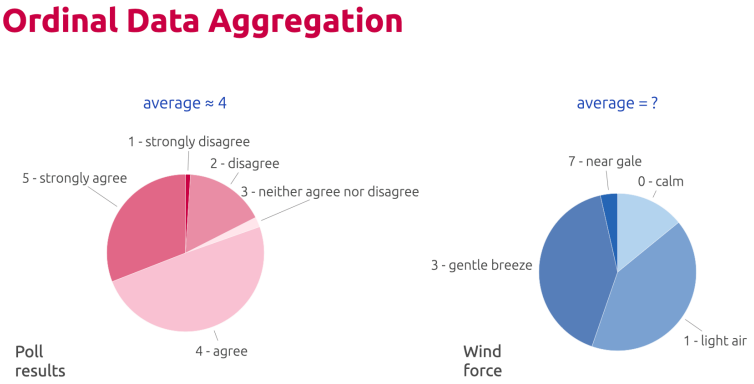


Figure 5: An example of calculating the aggregated values

2.3 Interval Scale

An interval scale consists of equal intervals. It has a unit of measurement and an arbitrary starting point also called an arbitrary zero point.

The interval scale has all the properties of the nominal and ordinal scales and also allows specifying the quantitative value of the measured feature.

It allows finding the difference between two values in the given units of measurement. The downside is the absence of an absolute zero as a reference point.

One example of an interval scale is a chronology with a starting point at the creation of the world (that is also interpreted differently), for example, the date of Jesus' birth or the foundation of Rome, and so on.

Celsius and Fahrenheit temperature scales are also examples of the interval scales.

When the air temperature is 30 degrees Celsius, we say it's 10 degrees higher than the temperature of 20 degrees, but we cannot say that it is one and a half

times warmer because there is no absolute zero as a starting point. Unit intervals can in turn be divided into equal intervals. Here's another example. A time scale is divided into years, months, days, hours, minutes, seconds, and so on. In theory, the interval scale is divisible into an infinite number of parts, which allows adjusting the scale depending on the research objectives.

Interval data aggregation. Arithmetic operations, such as addition, subtraction, multiplication, and so on, can be performed on the values on the interval scale, and all of them are formally possible. However, such operations are performed if it's possible to interpret the results. For example, one can calculate the average temperature in June in a particular location, but finding the average patients' temperature in a hospital is a waste of time.

2.4 Ratio Scale

A ratio scale has all the properties of the interval scale and also an absolute zero as a starting point.

The absolute zero on the ratio scale represents the absence of the property being measured, for example, the speed of a car, the number of apples in a fridge. In these cases, zero means that a measured thing is absent, which distinguishes it from an arbitrary zero in the interval data.

Aggregation of ratio scales. All arithmetic operations, including division, are meaningful for the data on this scale. For example, a product price is measured on the ratio scale in rubles or another currency. You can compare the product prices and conclude that one is three times more expensive than another. The price of 0 rubles means that the product is free.

3 Types of Data

When transforming data into useful information, it's important to know what data you're going to work with. Data may be simple factoids (the results of a preliminary analysis) or raw transactions that haven't been processed yet.

Depending on the level of pre-aggregation, the following types of input data are distinguished:

- Factoids
- Series
- Tables
- Transactions

Let's consider each of these types.

3.1 Factoids

A factoid is an aggregated part of the general information. The factoid is calculated based on the input (raw) data and focuses on a particular detail. The very word factoid seems to distrust the data aggregated somehow. Nevertheless, a significant part of the data provided for analysis comes in the form of factoids.

The sentence "95% of tourists in St. Petersburg visit the Hermitage" is a factoid. It focuses on the Hermitage visitors. This factoid is the result of a preliminary analysis of the information about the tourists visiting St. Petersburg museums. This factoid says nothing about tourists visiting the Russian Museum or the Peter and Paul Fortress. Perhaps, the conclusion was based on the number of tickets bought by museum visitors. However, tickets do not contain information about the visitor's status (a city resident or a traveler). The source may also be the results of a travel survey. We can only hope that the data was obtained and aggregated based on a quite representative sample. Anyway, the data accuracy is questionable.

3.2 Series

A series is data in which one type of information (dependent variable) is mapped to another type of information (independent variable). The information corresponding to the dependent variable can be of aggregated nature, which makes it a factoid.

For example, the table illustrates the percentage of trips by different types of public transport. In the example, the independent variable is a type of public

Transport	Number of Trips
Bus	26%
Trolley bus	6%
Tram	6%
Subway	62%

transport, and the dependent variable is the percent of trips. The percentages in the example are factoids. Hopefully, the percentages were calculated based on the information about the actual fares for each type of public transport.

When time is an independent variable, a series is called **a time series**. As an example, we can consider the St. Petersburg theater attendance statistics.

The source of the data in the table is the website of the Government of St. Petersburg. The total number of theater visits in the example depends on the year. Thus, the year is an independent variable, and the number of theater visits is a dependent variable that stands for the number of visits in a given year (for example, 1,702,200 people visited St. Petersburg theaters in 2011). Probably,

2010	2011	2012	2013	2014	2015	2016
1,492,600	1,702,200	1,823,300	1,866,300	2,117,200	2,310,680	2,267,129

the data was aggregated based on the theater reports on sold tickets, making it trustworthy.

3.3 Tables

Tabular data is of particular interest since many government services publish aggregated statistics about their activities in a tabular form.

Tabular data contains several units of dependent information and one unit of independent information. Information corresponding to dependent variables can be of aggregated nature (presented in the form of factoids).

The table shows an extended public transport example. Here, the type of public transport is an independent variable, and the number of trips and the average duration of the trip are dependent. The number of trips and their average

Transport	Number of trips	Average trip
Bus	26%	31 min
Trolley bus	6%	22 min
Tram	6%	21 min
Subway	62%	28 min

time for each type of transport are presented in the form of factoids. However, the question is how reliable these factoids are. The number of trips can be easily calculated based on the tickets sold, but the average duration is a more complicated indicator. Public transport information systems don't track the entry and exit of passengers explicitly, which means that the obtained number has been calculated based on probabilistic algorithms or observations. Nevertheless, even this information expands our knowledge about public transport. The problem is how to connect this information and obtain practical results. Aggregation of these dependent values seems incongruous because they are expressed in different units of measurement.

Let's consider another example of tabular data where time is an independent variable. St. Petersburg statistics show the number of theater visitors and students studying at universities. Both dependent variables (theater visitors and

Year	2010	2011	2012	2013	2014	2015	2016
Theater	1,492,600	1,702,200	1,823,300	1,866,300	2,117,200	2,310,680	2,267,129
University	807,000	749,500	707,800	654,500	593,000	555,600	553,700

university students) are expressed in the same unit of measurement (people). However, data aggregation still doesn't make sense. The data is unrelated. In

this case, aggregation between dependent variables or any comparison is meaningless. How to connect the number of theater visits in 2015 (2,310,680 people) and 555,600 university students? It's not possible to establish the relationships between data items despite that the year and place are the same.

The picture changes when we consider the tabular data statistics showing the number of students at universities and academies. It's possible to aggregate the

Year	2010	2011	2012	2013	2014	2015	2016
Universities	807,000	749,500	707,800	654,500	593,000	555,600	553,700
Academies	109,100	109,000	111,900	104,700	105,100	101,200	101,300
$\sum$	916,100	858,500	819,700	759,200	698,100	656,800	655,000

data. But a more formal justification is preferable. In fact, university students and academy students are two categories of the student variable. Thus, we can sum up the values expressed in the same units of measurement and related to the categories of one variable.

3.4 Transactions

Transaction (raw) records are data describing events. They are often created in the form of rows or tables. However, transaction records don't aggregate data by any parameter. Data is temporary, and it isn't stored. And yet analysts are particularly interested in it.

Pre-aggregation destroys the original history of the data but saves storage space. For example, the original transactions in St. Petersburg public transport contained about 60 million transactions (2 weeks of observation). However, when trips are aggregated by type of transport in terms of quantity and average duration, a lot of useful information is lost before the analysis. For example, the input data on trips provides important insights about the traffic flows in a city, including the speed of public transport vehicles in different areas of the city and the most popular routes. It also helps to assess the quality of transport infrastructure and so on.

The same goes for students. Student data may also give valuable insights. For example, we can find out what educational institutions or jobs students prefer. Keeping the original transactions has always demanded, and still demands, computing resources in terms of storage and processing power. That's why data collection often uses aggregation methods that significantly reduce the amount of analyzed data and minimize the time of subsequent processing.

The recent progress in the efficient storage technologies, parallel computing, and cloud technologies has made the storage and instant aggregation of data arrays a reality.

An analyst choosing between factoids and raw transactions should prefer the

latter because raw transactions will allow analyzing the data in different ways and use any suitable models.

4 Data Sources

Many questions arise when obtaining the data for analysis. How is it collected? Is data collected in the same way? Who is collecting the data? Are there any open data sources? Let's find the answers to all these questions.

The main data sources are:

- Opinion polls
- Observations
- Documents
- Results of direct measurements
- Social media and external sources

Let's consider each of these types.

4.1 Opinion Polls

Many researchers think of opinion polls as being the simplest and easiest method of collecting primary sociological data. There is no doubt that the efficiency, simplicity, and cost-effectiveness of this method make it very popular compared to other research methods. However, this simplicity is often apparent. The problem lies in obtaining quality data, which requires appropriate conditions and compliance with certain requirements. So, what are the conditions?

Firstly, it's the need to correctly construct the questionnaires for a poll. Secondly, there should be means for filling out a questionnaire and analyzing answers. Thirdly, it's about creating an environment of trust and support to carry out the poll. Sociologists are responsible for the first and third conditions, but they can also engage analysts in this step. It's necessary to make sure that the answers to the questions are non-contradictory and include all possible options. Such categories of answers are called comprehensive and mutually exclusive.

For example, the answers to the question: "What is your age group?" with possible options of ordinal answers may be intervals 20-30, 30-40, 40-50, 50-60. These options are ambiguous because a 30-year-old person will doubt which of the two possible options to choose. And a person aged 65 will have no choice at all.

In this case, the correct options are intervals that cover all possible answers and do not overlap. For example, under 20, 21-30, 31-40, 41-50, and 51 years old or older.

An analyst would prefer to see the input data in an electronic form rather than on paper. Fortunately, many tools are freely available for the creation of any questionnaire in an electronic form. One of the best examples is Google Forms (https://docs.google.com/forms/) allowing to create such questionnaires. You can create a questionnaire, share the link, and wait for responses. Here's an

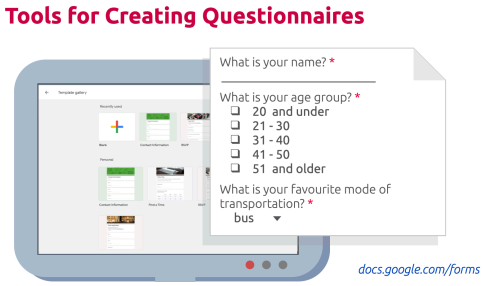


Figure 6: Google Forms

example of the questionnaire created with Google Form. Such services also help to collect and visualize the statistics of responses.

4.2 Observations

Observations are usually outsourced. They may be used to understand the behavior. Observations can come in handy when the objects of interest cannot fill out a questionnaire.

One of the main drawbacks of this method is that the data obtained by observing has low reliability due to human error (subjective judgments).

Every point of view is subjective to some extent and may affect the perception of an event. To minimize human error, several observers may be engaged, so that the collected data represents different points of view. High accuracy requires a large number of observers, which is rare in practice.

Despite the drawbacks, observations are the only way to collect data in some fields (for example, about animals, plants, and so on).

Oddly enough, this data collection method is still used even in those areas of human activity where the level of process automation is already high enough. For example, this method is widely used to study traffic flows in large cities although most of them are connected to digital information systems capable of providing any information on the movement of passengers.

4.3 Documents

Well-structured documents, or at least semi-structured, are an excellent source of data. However, documents are mostly ill-structured, which makes them

difficult to analyze formally.

A particular challenge is to analyze medical documents of past times. They were written by hand, and most of them are illegible because of handwriting or uncommon notations. As a result, the medical data collected over decades is lost because there's no way to formalize it.

4.4 Results of Direct Measurements

The best data for analysts is the results of direct measurements. Unlike the data from other sources, they are less prone to subjective interpretation. These are raw transactions with no aggregation. Such data is usually collected by various sensors that track studied objects (such as airplanes, vehicles, devices, people, and so on). These devices collect impressive amounts of disorganized data. Thanks to the progress in the efficient storage technologies, parallel computing, and cloud technologies, it's possible to keep such data arrays.

Here's an example of a service that uses direct measurement results. It's called Flightradar24 (https://www.flightradar24.com). Its measurements are used by many other helpful online services.

However, it's only a small portion of knowledge that we can extract from the data. By analyzing the operational data from sensors of an aircraft on a flight, we can track the flight trajectory, predict with a given accuracy what kind of maintenance is needed after landing, and the like. Since data collection devices record transactions in different systems, interact with an environment, or even fail, the data requires preprocessing before the analysis (syncing the identification of objects in different systems, removing noise, handling missing values, and so on). However, preprocessing doesn't reduce the importance of data. The amount of data only grows, and there's no doubt that the data contains valuable knowledge that we can extract with a meaningful data analysis.

Note that this kind of data is rarely public, and to obtain it, one has to deal with the owners of data storages.

4.5 Social Media

Social media are also a type of data source. Most social media sites provide an application programming interface (API) to access open data. The data from such sites is an excellent source for analyzing and predicting social activity. The article on the developers' website (https://developers.facebook.com/docs/graph-api/overview/) describes how to get data into and out of the Facebook platform.

4.6 External Data Sources

Sometimes there's no need to collect data because it's available on public resources that support a popular initiative for open access and free content. Different governments and companies have implemented data availability policies to

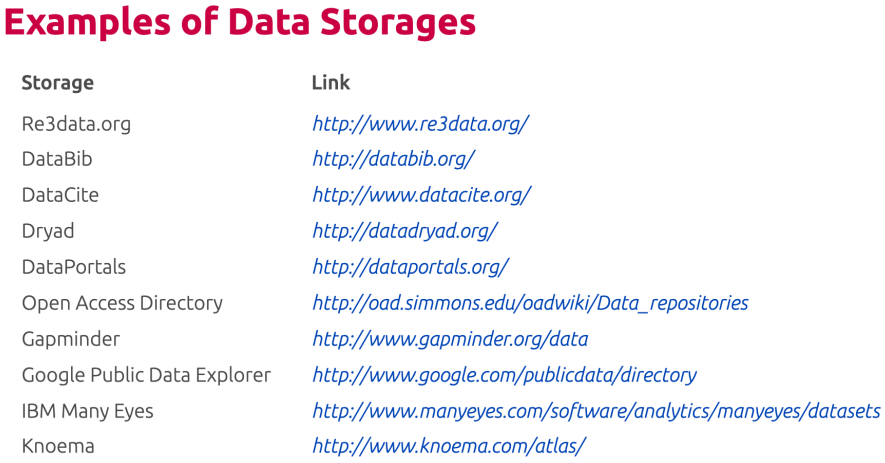
be more open and transparent, as well as to encourage the development of new products and services. Open data sources include search engines, data storages, government databases, and institutional repositories. Let's consider examples of such systems.

4.6.1 Examples of Search Engines

The most popular search engines are Google and Bing. A correctly formulated query may return a lot of useful information, but also partially useful information that contains the desired data.

4.6.2 Examples of Data Storages

The next source is open data storages. They don't have news feeds, books, or magazines. It's just a variety of data storages and nothing more. Here're the examples of such data storages.



Examples of Data Storages

Storage	Link
Re3data.org	http://www.re3data.org/
DataBib	http://databib.org/
DataCite	http://www.datacite.org/
Dryad	http://datadryad.org/
DataPortals	http://dataportals.org/
Open Access Directory	http://oad.simmons.edu/oadwiki/Data_repositories
Gapminder	http://www.gapminder.org/data
Google Public Data Explorer	http://www.google.com/publicdata/directory
IBM Many Eyes	http://www.manyeyes.com/software/analytics/manyeyes/datasets
Knoema	http://www.knoema.com/atlas/

Figure 7: Examples of data storages

4.6.3 Examples of Government Databases

The Open-Data Portal of the Russian Federation is an example of a government database. It contains useful and well-structured information.

4.6.4 Examples of Research Databases

The next class of data sources is databases of research institutions. One of such databases is AcademicTorrents (http://academictorrents.com). AcademicTorrents provides numerous datasets with good academic descriptions and also offers many amazing courses.

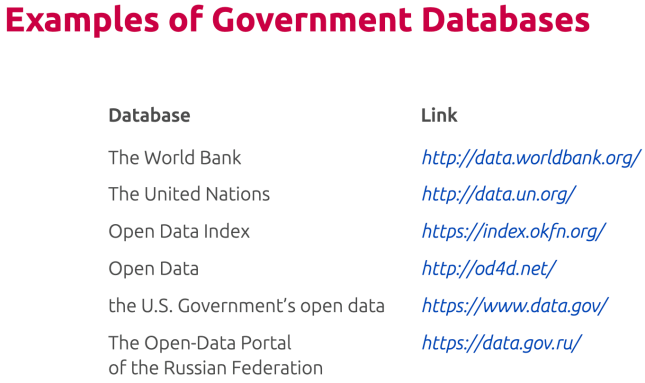


Figure 8: Examples of government databases

Making references to open data sources is a must. While the format of references may depend on a publisher, a reference should always include the author or owner of the data source, the name of the original source and the links to it.

5 Data Preprocessing

Data preprocessing is one of the steps that precede data analysis. Some preprocessing operations are often carried out when the input data is raw transactions collected from several sources.

These operations include loading data into storages, data splitting, bringing data to the same units of measurement, converting to a unified vocabulary, merging data from multiple sources, combining data from different sources, adding missing numerical values, and data cleaning. The challenge is that there are no universal techniques for carrying out these operations. Each dataset is unique, and some training techniques can be used only once. However, the successful completion of these operations may significantly affect the results of the analysis. So, let's consider the basic operations of data preprocessing.

5.1 Loading Data into Storages

It's the first step of data preprocessing. When data is loaded into a single storage offering the tools for data handling, we can merge the data from multiple sources and perform primary data processing. Storage systems usually offer special utilities for downloading data from external sources. However, even this step may bring many surprises, such as unreadable characters, unexpected types of data, and so on. Because of that, the following loading recommendations are suggested:

- Remove all unreadable characters from the input files.
- Load the input data as text fields (to deal with data types later on).

- When the amount of data is large, load to the storage server directly.

5.2 Data Splitting

A simple example of a problem that analysts often solve is separating first and last names, as well as addresses. In cases like this, the input data is first and last names located in one cell, and the task is to separate them. This task requires knowledge of the peculiarities of both names and surnames. Or, the input data is separate cells for first and last names, but some of the entries contain first and last names altogether or their order is different from the majority of entries. Please look at an example. How to split the data? There are many possible solutions,

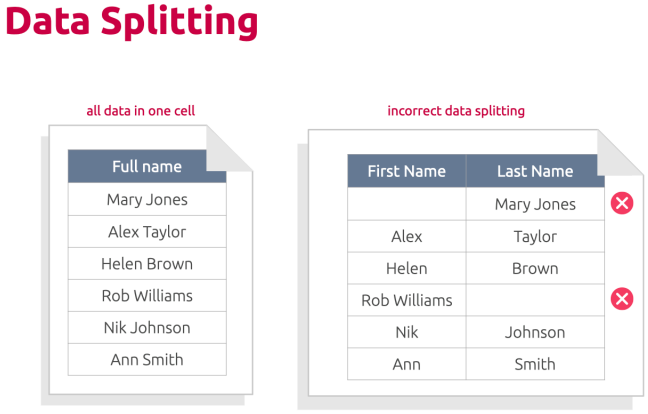


Figure 9: An example of data splitting

but the simplest and most effective is to create a directory of common names and use it to analyze the input data. Non-standard names should be analyzed separately. But their number will be limited.

5.3 Converting Data to the Same Units

This step focuses on another important aspect of data preprocessing to ensure that all the values related to the same variable are represented in the same units. Take medical data from different countries for example. Some countries use pounds as a measure of weight while others use kilograms (the measurement scales differ). We choose one unit of measurement and convert all the values to the same unit. Otherwise, we cannot compare or aggregate the values. You can see an example of the conversion in the figure.

5.4 Converting to a Unified Vocabulary

This step usually accompanies the input of nominal data. When the nominal data is input as a free text and not selected from the list, the problems will surely come. For example, some people may type “the English language” in a text field, but some will simply write “English” or even “eng” with no capital letters. Despite that all these entries refer to the same thing, they will affect data processing or

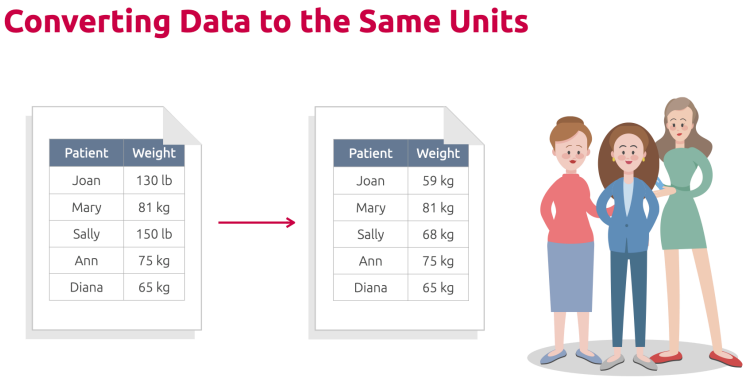


Figure 10: An example of data conversion

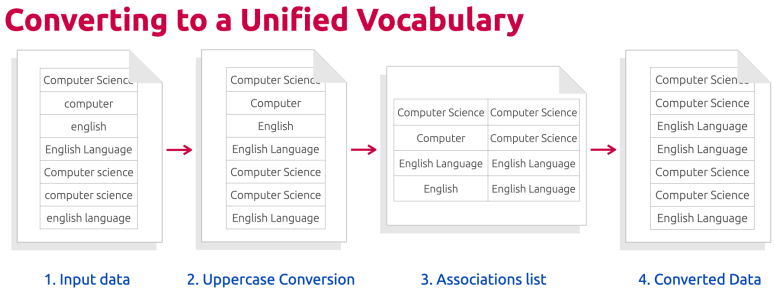


Figure 11: An example of converting to a unified vocabulary

even lead to undesirable results. To avoid this, we should convert data to a unified vocabulary. How to do this? One way is to convert the values to one case (lower or upper), make a list of values that differ to create associations, and replace the inconsistent values.

5.5 Merging Data from Multiple Sources

Imagine you are making a list of house residents. The residents of each unit have provided their lists of occupants, and your task is to combine all the lists into one, or to put it differently, to merge different entities. What problems can arise during this operation? In set theory, the union is an operation with elements of the same structure. So, to merge data, first make sure that the provided data has the required structure. Suppose that for this task, you need to know the name of the responsible resident and the apartment number. In this case, there are only 2 household units, and the data is as follows: How can we merge the data? We should at least convert the data to the same structure. First, we match the fields to see which fields in the first table correspond to the fields in the second table (even if they are called differently). After matching, we rename the fields for consistency. The next question is what to do with the fields that are only found in one place (in our example, it's the total number of residents). There are two options. The first one is to delete the fields that are not found in all the sources or do the opposite and expand the definition of the source structure to include

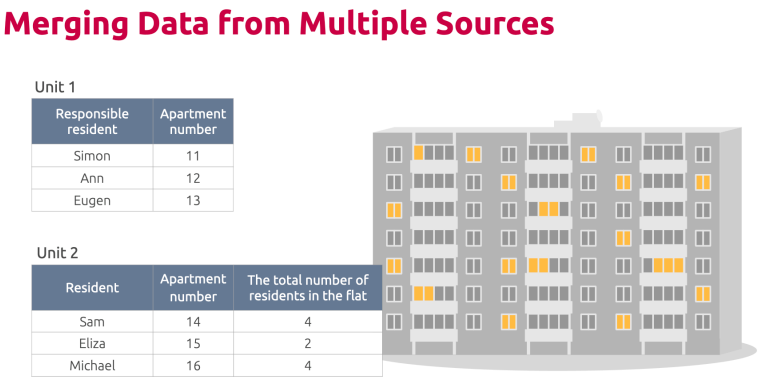


Figure 12: Household unit data

fields found in at least one other source (to use the data in another task). And now we can merge the data. Examples are shown in the figures. However, data

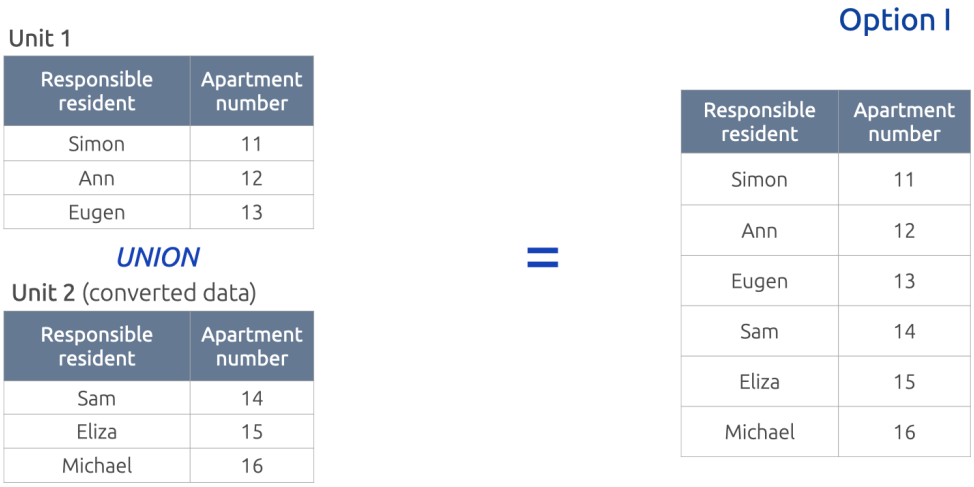


Figure 13: The first option shows minimal structuring of data followed by merging

sources being merged may contain data in different units of measurement or have no unified vocabulary. In this case, the first step is to convert the data to the same scales, units of measurement, and a unified vocabulary, and the second is to perform data merging as described earlier.

5.6 Combining Data from Multiple Sources

Different sources of real-world data complement each other and enrich our understanding of the object. Combining data from all possible sources is doable but it comes with a cost.

The first problem is matching the fields. As when merging data from multiple sources, it's necessary to investigate the matching fields and convert names consistently.

The second problem is to convert data from different sources to the same scales, units of measurement, and unified vocabulary.

Unit 1 (converted data)

Responsible resident	Apartment number	The total number of residents in the flat
Simon	11	
Ann	12	
Eugen	13	

UNION

Unit 2

Responsible resident	Apartment number	The total number of residents in the flat
Sam	14	4
Eliza	15	2
Michael	16	4

=

Option II

Responsible resident	Apartment number	The total number of residents in the flat
Simon	11	
Ann	12	
Eugen	13	
Sam	14	4
Eliza	15	2
Michael	16	4

Figure 14: The first one shows the data conversion to the extended structure followed by merging

The third problem is to identify data associated with one object (for example, identify data on a particular buyer in different supermarkets). Sources of information often use different systems of object identification (for some, each identifier is a number on the identity document, for others, it's a mobile phone number, and so on). We aim to find shared attributes that are unique for the object in each system (for example, a mobile phone number, e-mail, and so on) even if these attributes are not identifiers in their original systems.

Finally, data sources can be structures of various formats (tables, JSON, XML, and so on). However, this task is quite easy to solve, as today's storage management systems often allow querying data with varied structures. Otherwise, an additional step is added to convert data to one format.

After all these steps, we can combine data from multiple sources.

Data from a sports club

First name	Last name	Date of birth	E-mail	Mobile phone number	Passport	Type of sports
Nicolas	Brown	08.02.1998	nbrown@gmail.com	8(922)468-2929	4004 271492	swimming
Sam	Wilson	03.01.1999	samwilson@yahoo.com	8(931)852-9582	3003 262899	volleyball

Data from a supermarket

First name	Last name	E-mail	Mobile phone number	Occupation
Nik	Brown	nbrown@gmail.com	8(922)468-2929	student
Samuel	Wilson	samwilson@yahoo.com	8(931)852-9582	official

Data Combination Results

First name	Last name	Date of birth	E-mail	Phone number	Passport	Type of sports	Occupation
Nicolas	Brown	08.02.1998	nbrown@gmail.com	8(922)468-2929	4004 271492	swimming	student
Samuel	Wilson	03.01.1999	samwilson@yahoo.com	8(931)852-9582	3003 262899	volleyball	official

Figure 15: Combining data on the sports club and supermarket

5.7 Adding Missing Numeric Values

Empty or incomplete numeric fields are one of the problems of data processing. If the data is missing because it hasn't been collected, there might be a chance to do it and fill in the missing values. However, it's also possible that

the source is no longer available. For example, you have sensor readings, and no other data is available. There are two approaches to such cases:

- Label the fields with special values (and exclude the fields from the analysis).
- Approximate missing values based on existing data.

5.7.1 Approximation of Missing Values

Missing numerical values are usually approximated by calculating the mean for the entire dataset. Some types of data, such as time series, require a different approach. They are calculated based on the mean, the so-called sliding window consisting of the nearest neighbors. The window width (number of neighbors) taken to find the mean depends on the task, but it is usually limited to 4-6 neighbors (on the left and right). In some cases, we apply non-standard algorithms

Approximation of Missing Values

Stop name	Arrival time	Stop no.
Primorskaya Station	16:00	1
Nalichnaya str.	?	2
Malyj ave.	?	3
Gavanskaya str.	?	4
Shkipersky str.	?	5
Sredniy ave.	?	6
Tram Park	?	7
The 12th line	16:35	8

→

Stop name	Arrival time	Stop no.
Primorskaya Station	16:00	1
Nalichnaya str.	16:05	2
Malyj ave.	16:10	3
Gavanskaya str.	16:15	4
Shkipersky str.	16:20	5
Sredniy ave.	16:25	6
Tram Park	16:30	7
The 12th line	16:35	8

Figure 16: An example of missing values corresponding to tram arrivals

that are strongly tied to the analyzed domain. For example, a tram had arrived at the destination, but the time of the stops made was lost. The task is to restore the data on time based on the historical data on time recorded before and after the loss.

Without additional data, we can calculate the period between the filled-in fields and evenly divide it between stops. However, if there is historical information about when the tram made the stops earlier, we can find the time intervals between those stops and distribute the time interval among stops in question proportionally to the historical intervals.

5.8 Data Cleaning

Data cleaning usually includes the following steps:

- Removing duplicates

- Checking ranges
- Comparing with references or regular expressions

5.8.1 Removing Duplicates

Duplicates may appear in the input data due to technical issues. They are typical for raw transactions. Duplicates may cause incorrect data aggregation. For example, an illustrated automated row counter will return an incorrect number of vehicle types.

Duplicates are easy to find and remove with the data processing tools that are provided by storages.

5.8.2 Regular Expression Check

Some types of data should have a particular format. For example, it applies to email addresses or phone numbers. To check the validity, data is compared with the template. The template is given as a so-called regular expression. Different

Attribute	Template
Email	%@%.%
Phone number	+7(DDD) DDD-DD-DD

tools offer various ways to create such templates, but this feature is available almost everywhere.

The question is rather what to do with the data that failed the validation. We have at least three options. The first one is trying to obtain valid data, the second is to fill in the values with explicitly specified symbols with an undefined value, and the third is to exclude them from the analysis.

5.8.3 Range Check

Another step in data validation is a range check. At first glance, the range check is a very simple operation used in numeric variables to see what values are higher or lower than the values acceptable for this variable in the dataset. The point is how to find this range. For the best results, the range is usually set by the domain rules. For example, the range is points from 0 to 100. So, the value of 787 in the input data is an obvious error.

In this case, when loading the data into the storage, we can check the loaded values and exclude those that are obviously incorrect. It's worse when no distinct ranges exist, and some values look suspicious. We inherently understand that because they are very different from others. But how to formalize the concept "very different"? Such values are termed outliers. Statistics offer formal methods

for outlier detection, and we will certainly consider them but not now because, to define outliers, we need other statistical concepts.

Next time, we will consider some tools for data processing that will allow us to apply the acquired knowledge in practice.