Data Structure Guide

Introduction

Often we hear that (big) data is the oil of the 21st century. Which may give you the feeling you should start blindly collecting all data that you could possibly collect. However, imagine yourself a company that has collected a lot of customer data over the last 10 years which they want to analyse. A first step for this company might be to have a conversation with a Data Analyst. The company might tell the data analyst they have already done the data collecting part, so it should be easy to get a lot of interesting insights from the data, right? However, when the Data Analyst looks at the available data, there might be a big chance that the Data Analyst will say : "I will need a couple of months before we get to the desired insights, because your data is not yet ready for the analysis".

This guide will help you understand why it could take longer than expected to create meaningful insights by analysing your company data and what you as an SME can do to prepare your data in the right way to smoothen the data analysis process. If you understand the concept of clean and structured data, you can start to look at your own data and see where it needs improvement. After cleaning and structuring your data, your data will be ready for analysis, and you can start using the tools provided by Futures by Design to start your own data exploration!

How will this guide help me preparing my data?

The following steps are explained is this guide followed by some examples and questions that can support you in the process of cleaning your own data:

1. Data collection: What data should you collect? (if you haven't yet)
2. Collecting the right data: How to collect the data right?
   a. Structured data
   b. Clean data
   c. Relational Database
3. Preparing data for further analysis

Data collection: What data should you collect?

This section is especially relevant for the SME's that have not yet collected any data. First of all, from the perspective of future analysis, the more data you can collect at little to zero effort and resources, the better. However, when you are just starting you should keep in mind what the reason is to collect certain variables and what the additional costs are. It rarely makes sense to record the temperature of the day while registering sales orders for example, but if you are interested in analysing the effect of the weather on your sales, it might be interesting[1]. Therefore, you do not need to spend time and resources on collecting irrelevant data, that you

[1] It is good to note that in this example there is no need for you to look at the thermometer every day and note down the temperature. Most weather data is publicly available.

are not going to use anyway. Answering the following questions can help in determining which data you need to collect for your initial analyses:

1. What is my the question I want to answer using analytics?
2. What data do I need to analyse to answer this question?
3. Which relevant data can I collect?
4. Is that enough to answer my initial question?

Example

For example, you have a company that sells laptops to customers and your goal is to target your customers better with personalized deals. Filling out the questions above could look somewhat like this:

1. *What is my the question I want to answer using analytics?*
   How could I target our customers with more personalized deals based on their laptop preferences?

2. *What are the sub questions to answer this question and which data do I need to make my analyses?*
   - What different type of customers do I have?
     The relevant data to see whether there are different type of customer groups, this often requires some domain knowledge:
       - Country
       - Age
       - Gender
       - City size
       - Marital status
       - Income
       - Family size
       - Last purchase (date)
       - Amount of purchases
       - Total amount of money spent
       - …..
   - What do my customers mostly buy?
     The relevant data to see whether there a different products bought:
       - Product name
       - Brand
       - Memory
       - Screen size
       - Camera
       - Bluetooth
       - Amount of USB ports
       - ……
   - Can I see a pattern between the different customers buying different products?
       - Requires all the information above

3. *Which relevant data can I now collect?*
   - *Customer data* – Name, address, phone number, e-mail, country, age

- *Product data* – Product name, product description, different specifications of a laptop, e.g. screen width, brand, camera, etc.
- *Sales data* – Date, Product, Customer, Quantity, price

4. Is that enough to answer my initial question?
   Data analysis is an iterative process, meaning that we can start with the available data. Once it turns out we don't have enough information to come to interesting insights, it is always possible to collect more data. However, start with as much of the possible and relevant data sources as possible.

Of course you can have many different Key Performance Indicators (KPI's), requiring different information provision. Therefore, it is not a bad idea to collect a lot of information. You should, however, at least have the data necessary to answer your research question(s). These questions help you to identify what data and variables you need to answer those questions.

Once we have identified the different variables, it is important to store this data in the right way. There are many options on how to store your data, but Excel is a good place to start if you are just starting to collect your data. If you have a lot of data already, or you are more experienced, you can look into Databases such as MySQL, PostgreSQL or MongoDB.

Collecting the right data - Structured data

Structured data is usually found in Excel datasets or relational databases (we will cover this later). It means that everything is stored in columns or fields. Every row denotes a new instance (person, item, company, or other piece of information). The reason for having structured data is that it makes it more simple to search for pieces of information for you as a person, but also for the computer to eventually perform faster and better analyses.

Unstructured data is essentially everything else. Unstructured data might have an internal structure but is not structured via pre-defined data models or schema. It may be textual or non-textual, and human- or machine-generated.

Examples of unstructured data:

- Text files: Word processing, spreadsheets, presentations, email, logs.
- Email: Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- Social Media: Data from Facebook, Twitter, LinkedIn.
- Website: YouTube, Instagram, photo sharing sites.
- Mobile data: Text messages, locations.
- Communications: Chat, IM, phone recordings, collaboration software.
- Media: MP3, digital photos, audio and video files.
- Business applications: MS Office documents, productivity applications.

Example of unstructured data and structured data

Unstructured
Today we had a meeting with Martin Jones, CEO of a possible supplier, Lenovo. They are willing to give us a 10% discount on every laptop that we buy if we buy it in batches of minimum 100.

Structured	
Date:	21-07-2020
Type of meeting:	Supplier intake
Contact person:	Martin Jones
Company:	Lenovo
Discount:	10%
Minimum batch size:	100

Questions

1. What types of unstructured data do you have?

   ____

   ____

2. If so, how could you transform this unstructured data into structured data?

   ____

   ____

   ____

   ____

Collecting the right data - Clean data

Data cleaning means the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according to the necessity. Data cleaning is a large part of the work, when you want to create value from your data. It takes up most of a data scientist's time.

Data is essential for analytics and machine learning, but when it comes to the real world data, it is likely that data may contain incomplete, inconsistent or missing values. Therefore, the importance of data cleaning cannot be overstated. No matter the algorithm you use — if your data is bad, you will get bad results. Professional data scientists know this and have revealed that data cleaning takes up to 70% of the time spent on a data science project.

Missing values

This is perhaps the most common trait of unclean data. These values usually take the form of NaN or None in a table. This is, for example, the case when you open your data in Excel, and there are empty cells. Missing values can also lead models that predict NaN values, which we would not want. In order to handle missing data, it is important to identify the cause of the missing data. This will guide us in deciding on the best way to handle them.

There are several causes of missing data values. Some values could be missing because they do not exist, others could be missing because of improper collection of data or poor data entry. For example, if someone is single and a question applies to married people, then the question will contain a missing value. In cases like this, it would be wrong to fill in a value for that question. So how can you handle these missing values ? It depends on the type of data you are collecting:

Types of data

- **Numerical**: You can either fill in a 0 or the average for missing values.
- **Text** : Most of the time text is not relevant for analysis, especially when the column is the description of an item.
- **Category** : If there is a row that doesn't fall in the available categories, you can create a new category "No category". Which can be used to fill up the empty values.
- **Dates** : If a date is missing, while this is essential for your analysis you can best delete the whole row if you cannot recall the date. Deleting rows or columns is the least preferred but a good alternative if you can't fill in the missing values.

Inconsistent data

This mostly occurs in text or categorical data. For example, when you have to write down the names of different countries, employee #1 might write down "netherlands", employee #2 might write down "The Netherlands" and employee #3 might write down "Holland". While they all mean the same thing, it is written different and a computer will not see them as equal. If we were to do an analysis based on the demographics of your customer, a result would give that there is one person living in "netherlands", "The Netherlands", and "Holland" rather than 3 persons living in The Netherlands.

So how can you handle this inconsistent data? Figure out for yourself how you are collecting the data now. Is it possible for your employees to fill in certain values? If so, are they limited by a dropdown menu (consistent data) or is it an open text box allowing for all different types of values (inconsistent data). If it is the latter, check your data for any inconsistencies and try to prevent having open fields but rather choose dropdown menu's or write clear instructions on how to enter the data.

Example

Name	Address	Phone	E-mail
Tyrone Lowe	89 St James Boulevard, EX39 0YB, Horns Cross	+44 77 6428 2654	t.lowe@gmail.com
Spears, Loren	17 Ponteland Rd, Mortimer WE, TD5 7XT	070-09974620	lorenspears@outlook.com
Maria Randall	96 Preston Rd, Mortimer West End	077 1879 1573	m.randall@icloud.com

Table 1. Structured, unclean data

First name	Last name	Street	Zipcode	City	Phone	E-mail
Tyrone	Lowe	89 St James Boulevard	EX39 0YB	Horns Cross	+44 77 6428 2654	t.lowe@gmail.com
Loren	Spears	17 Ponteland Rd	TD5 7XT	Mortimer West End	+44 70 0997 4620	lorenspears@outlook.com
Maria	Randall	96 Preston Rd	RG7 5DS	Mortimer West End	+44 77 1879 1573	m.randall@icloud.com

Table 2. Structured, clean data

Can you see the differences? Let me help you:

- First name and last name separated in columns to prevent messing with the order
- Address split up in street, zipcode, city to prevent messing with the order
- City written down consistently (Mortimer WE/ Mortimer West End)
- Phone written down consistently with the land code

Questions

1. How clean is the data in your systems?

____

2. Can you write down any data sources that are collected in a clean, consistent way?

   ____

3. Can you write down any data sources thar are not (yet) collected in a clean, consistent way?

   ____

4. What steps could you undertake to increase the data quality of these data sources?

   ____

   ____

   ____

Collecting the right data - Relational database

In the above examples, we mentioned different data sources such as Customer data and Product data. It is common for organisations to have multiple tables that store different types of data. To get a more holistic/rich view on the business question you are trying to answer it might be necessary to combine different data sources within your company. A **relational database** is a way of storing and organizing your data into tables, and links them, based on defined relationships. These relationships enable you to easily retrieve and combine data from one or more tables. Therefore a database can be seen as a collection of tables. If you understand the logic behind these relational databases, it might help you structure your own data in the correct way.

- A **database** contains one or more tables of information.
- The rows in a table are called **records**.
- The columns in a table are called **fields** or **attributes**.
- A database that contains two or more related tables is called a **relational** database.

In this guide we will not elaborate too much on creating a fully functioning relational database, but rather give you some basic knowledge and rules. These rules should be kept in mind when collecting or organizing your data. We will explain these different steps with an example and end with an hands-on example in Excel.

Step 1. Start with a clean and structured table with information

Imagine that you are responsible for keeping track of all the books being checked out of a library. You fill out the single table below every time when someone borrows a book to track all the critical information:

First name	Last Name	Address	Phone	Book title	Author	Year	Due Date
Bob	Smith	123 Main St.	555-1212	Don Quixote	Miguel Cervantes	1605	14-07-2020
Alicia	Petersohn	136 Oak St.	555-1234	Three Men in a Boat	Jerome K. Jerome	1889	16-07-2020
Bob	Smith	123 Main St.	555-1212	Things fall Apart	Chinua Achebe	1958	15-08-2020
Bob	Smith	123 Main St.	555-1212	Anna Karenina	Leo Tolstoy	1873	15-08-2020
Zayn	Murray	248 Pine Dr.	555-1248	Heidi	Johanna Spyri	1880	17-08-2020
Bob	Smith	123 Main St.	555-1212	The Old Man and the Sea	Earnest Hemingway	1952	10-09-2020

Table 3. Original table

This table meets the basic need to keep track of who has checked out which book, but does have some serious flaws in terms of efficiency, space required, and maintenance time.

Problems with this table:

- Re-enter existing information
  When Bob checks out more books over time, you will have to re-enter all of his contact information for every book. This is a waste of time, since you've entered his contact information already, but also increases opportunity for error.

- Update all rows when change occurs
  When an update is necessary (e.g. Bob's phone number changes), each of Bob's records must be located and corrected. If one of Bob's records has a different phone number from the rest, is it a correction, a record overlooked during the last update, or a data-entry mistake?

- A lot of duplicate information
  For each book the author and the year it is published in, is entered. This is also duplicate information and again increases the opportunity for error.

These problems can be decreased by **normalizing** our data – in other words, dividing the information into multiple tables with the goal of having "a place for everything, and everything in its place." Each piece of information should appear just once, simplifying data maintenance and decreasing the storage space required.

Reasons for normalization:

1. Minimize duplicate data
2. Minimize or avoid data modification issues
3. Simplify extracting important data

Step 2. Create unique rows, prevent duplicate information

In order to prevent the collection of duplicate data and leaving less room for errors, there are some basic guidelines you can follow:

Basic guidelines

- Every table has a primary key
  The primary key is one or more columns whose values are unique in this table, and so can be used to identify different rows, e.g. order number, employee number, invoice number. The primary key is the column that will help you connect different tables of information correctly.

- Tables cannot contain sub-columns
  You cannot list multiple cities in one column and separate them with a semi-colon. For example, the value "Chicago" is desired; whereas "Chicago; Los Angeles; New York" is not.

- A table should not contain repeating groups of columns.
  Examples are Customer1Name, Customer2Name, and Customer3Name.

- If a column contains duplicate information, it should be listed in a new table.
  For example, if the address of a person is repeated multiple times in a column it is sensitive for errors and should be registered in a separate table (Customer data).

Example

According to the rules above, we can separate the data in two tables. Instead of repeating everything we know about a client whenever he checks out a book, we will instead give each library client an ID, and repeat *only the ID* whenever we want to associate that person with a record in another table. This helps to show which records in the Clients table correspond to which records in the Checkout table – in other words, who checked out which book.

Clients table

Client ID	First name	Last name	Address	Phone
1	Bob	Smith	123 Main St.	555-1212
2	Alicia	Petersohn	136 Oak St.	555-1234
3	Zayn	Murray	248 Pine Dr.	555-1248

Checkout table

Checkout ID	Client ID	Book title	Author	Year	Due date
1	1	Don Quixote	Miguel Cervantes	1605	14-07-2020
2	2	Three Men in a Boat	Jerome K. Jerome	1889	16-07-2020
3	1	Things fall Apart	Chinua Achebe	1958	15-08-2020
4	1	Anna Karenina	Leo Tolstoy	1873	15-08-2020
5	3	Heidi	Johanna Spyri	1880	17-08-2020
6	1	The Old Man and the Sea	Earnest Hemingway	1952	10-09-2020

Table 4. Clients table & Checkout table separated

Now the two tables can be related by using the Client ID. In Clients table, the Client ID field is the primary key and so its values must remain unique. For example, the value "2" can appear only on one record - Alicia's - and Alicia can have only one Client ID - "2."

Client ID cannot be the primary key of the Checkout table, because as we see Client ID "1" appears more than once and is thus not unique. This makes sense, because we want the clients to be able to rent multiple books. If Client ID were the primary key for the Checkout table, each person would only be permitted to check out one book, and afterwards be forbidden to check out any more books, ever.

We can't make Book Title the primary key, or we'd have a similar problem – each book could only be checked out once, and afterwards no one would be permitted to check it out ever again. We can't make Due Date the primary key, or else only one book could be due each day. Since none of the existing fields works as a primary key, a new field is added to identify each record named Checkout ID.

Step 3 – Primary key related columns

The primary key is used to uniquely identify each row in a table. When we talk about columns that depend on the primary key, we mean that in order to find a particular value, such as what color is Kris' hair, you would first have to know the primary key, such as an EmployeeID, to look up the answer. When all the columns relate to the primary key, they naturally share a common purpose, such as describing an employee.

Once you identify a table's purpose, look at each of the table's columns and ask yourself, **"Does this column describe what the primary key identifies?"**

- If you answer "yes," then the column is dependent on the primary key and belongs in the table.
- If you answer "no," then the column should be moved different table.

Example

When looking at our Checkout table in Table 2 above, we see that there is still some additional information (author, year) being collected for a book while this is not the purpose of the table. Therefore, we create a new table in Table 3 below called Book. Now we have the optimal database structure for the given dataset which satisfies al the guidelines. This structure allows for fast analyses and minimalizes the probability for errors.

Clients table

Client ID	First name	Last name	Address	Phone
1	Bob	Smith	123 Main St.	555-1212
2	Alicia	Petersohn	136 Oak St.	555-1234
3	Zayn	Murray	248 Pine Dr.	555-1248

Books table

Book ID	Book title	Author	Year
1	Don Quixote	Miguel Cervantes	1605
2	Three Men in a Boat	Johanna Spyri	1880
3	Things fall Apart	Chinua Achebe	1958
4	Anna Karenina	Leo Tolstoy	1873
5	Heidi	Johanna Spyri	1880
6	The Old Man and the Sea	Earnest Hemingway	1952

Checkout table

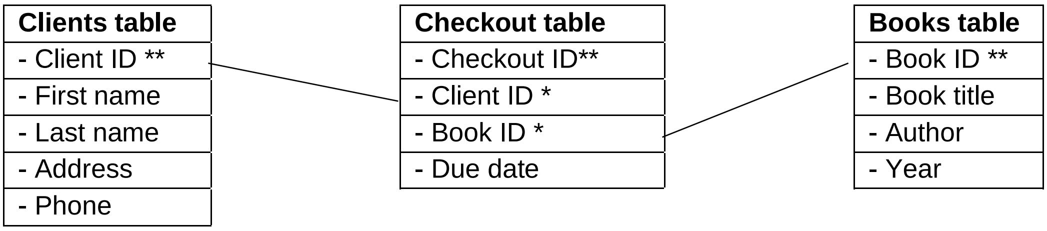
Checkout ID	Client ID	Book ID	Due date
1	1	1	14-07-2020
2	2	2	16-07-2020
3	1	3	15-08-2020
4	1	4	15-08-2020
5	3	5	17-08-2020
6	1	6	10-09-2020

Table 5. Optimal database structure

Step 4. How to still get a clear overview?

Diagram

Imagine there are many different tables in your database, you might want to get a clear overview of all these tables and how they are related. We can create a diagram that illustrates these relations and shows the **primary** (**) and **foreign** (*) keys. Foreign keys are primary keys from another table, such as Client ID in the Checkout table. This is especially handy if you are collecting a lot of columns in many different tables.



Clients table
- Client ID **
- First name
- Last name
- Address
- Phone

Checkout table
- Checkout ID**
- Client ID *
- Book ID *
- Due date

Books table
- Book ID **
- Book title
- Author
- Year

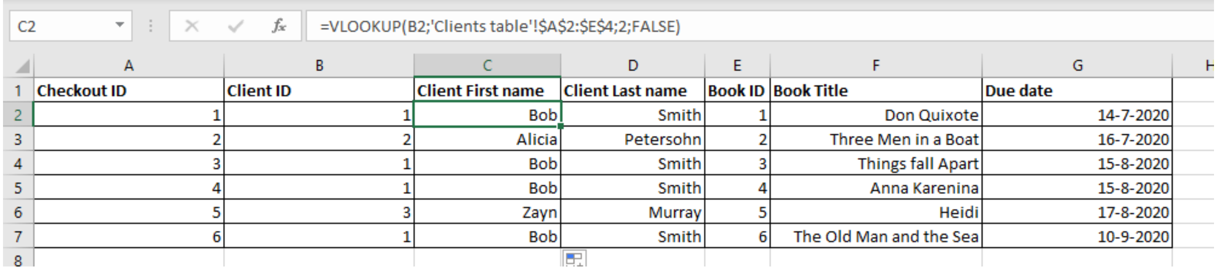
Figure 1. Relational Database Diagram

Overview table

Even though the table structure in Table 5 is the best version of how to structure your database, it is unclear to see in one blink which client has lend which book in the Checkout table. Therefore, it is not necessarily a bad thing to have one table that contains all the relevant information. However, instead of manually filling in this information, it is possible to link it to the Books or Clients table in such a way that it is automatically updated in all present and past fields when something changes in any the underlying tables.

Excel Example

Imagine the Clients table, Books table and Checkout table are all different sheets in Excel and you would like to have a readable overview of the checkout table.



C2 =VLOOKUP(B2;'Clients table'!$A$2:$E$4;2;FALSE)

Checkout ID	Client ID	Client First name	Client Last name	Book ID	Book Title	Due date
1	1	Bob	Smith	1	Don Quixote	14-7-2020
2	2	Alicia	Petersohn	2	Three Men in a Boat	16-7-2020
3	1	Bob	Smith	3	Things fall Apart	15-8-2020
4	1	Bob	Smith	4	Anna Karenina	15-8-2020
5	3	Zayn	Murray	5	Heidi	17-8-2020
6	1	Bob	Smith	6	The Old Man and the Sea	10-9-2020

Figure 2. Overview table (in Excel)

As you can see, there is a formula used to display the name of the client and the book title. This is a very practical formula which will be explained below. By using this formula, we can make changes in the 'Books table' and the 'Clients table' which will be automatically updated in this overview list. However, if you have a lot of data this can be a very time consuming formula. Therefore you can best use this if it doesn't affect your core tasks in Excel. If it does affect your tasks and you have an enormous amount of data, it might be an idea to look further into database management systems such as MySQl, PostgreSQL or MongoDB.

- Formula used in this case:
  =VLOOKUP(B2;'Clients table'!$A$2:$E$4;2;FALSE)
- Syntax
  =VLOOKUP (value, table, col_index, [range_lookup])
  - value - The value to look for in the first column (primary key) of a table.
  - table - The table from which to retrieve a value.
  - col_index - The column in the table from which to retrieve a value. In the example above column A is 1, Column B is 2, etc.
  - range_lookup - [optional] TRUE = approximate match (default). FALSE = exact match.

Preparing data for analysis

When your data is clean and structured, we have created some tools for you to explore your data with. However, since machines can't handle text very well, we have to make some small adaptions to get the most interesting insights of your data. This is a slight summary of the steps above, with some added steps.

Make sure that your are using a copy of the original dataset so you don't make any changes in your original data.

Step 1 – Check for missing values

Open your data in Excel and check if every cell contains information in your data. Is there a column that contains many missing values?

- Try filling the empty cells with the right information
- Is the information unknown, but the column relevant? You can fill with 0 or the average.
  - Does one of the columns contain a lot of empty values, but you want to know whether the information is present or not? For example, you want to know whether a client has a website yes or no, you can add a column "Website available?". If there is a website you can fill in the value 1, if not you fill in 0. By doing so, you handle your missing data perfectly and still get more information out of the available data.
- Is the information unknown, but the column not relevant? Such as description of a product. You can delete the column.

Step 2 – Check for inconsistencies

Is the collected data consistent?

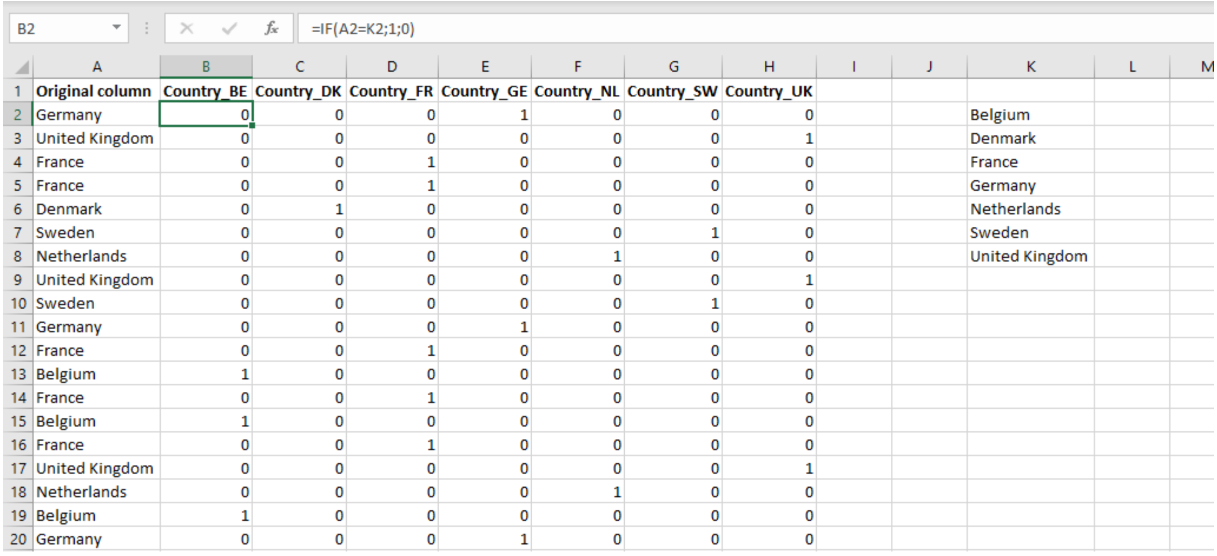
- Are phone numbers written the same way? (+31 61042455 vs. 06 11042455)
- Is categorical data collected consistently? E.g. you can't have Mortimer WE and Mortimer West End, this should be the same.
  - You can check the unique values in Excel by using the formula
  =SORT(UNIQUE(column)). This can help to find your inconsistencies and clean up your data.

Step 3 – Create as much numerical values as possible

As said before, computers have a hard time with handling text to analyse patterns. For example, if you want to predict the price of a product, the total description of the product is not very useful. What we, however, can ask is whether there is a description or not. This question immediately translates to a true or false answer, which can be written in numbers as 0 (false) and 1 (true).

For categorical values, such as countries, brands, or gender, the column often contains text values such as Netherlands, Apple, or Female. Again this is hard to understand for a computer. However, it is also possible to transform the data to numerical values such as above. This time we ask ourselves the question “Do they live in the Netherlands?”, “Is the brand Apple?”, “Is the gender female?”. The answer to these questions is again false (0) or true (1). An example of how to translate these countries to 0 or 1 in Excel is elaborated below.

Excel example



B2 =IF(A2=K2;1;0)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Original column	Country_BE	Country_DK	Country_FR	Country_GE	Country_NL	Country_SW	Country_UK					
2	Germany	0	0	0	1	0	0	0			Belgium		
3	United Kingdom	0	0	0	0	0	0	1			Denmark		
4	France	0	0	1	0	0	0	0			France		
5	France	0	0	1	0	0	0	0			Germany		
6	Denmark	0	1	0	0	0	0	0			Netherlands		
7	Sweden	0	0	0	0	0	1	0			Sweden		
8	Netherlands	0	0	0	0	1	0	0			United Kingdom		
9	United Kingdom	0	0	0	0	0	0	1					
10	Sweden	0	0	0	0	0	1	0					
11	Germany	0	0	0	1	0	0	0					
12	France	0	0	1	0	0	0	0					
13	Belgium	1	0	0	0	0	0	0					
14	France	0	0	1	0	0	0	0					
15	Belgium	1	0	0	0	0	0	0					
16	France	0	0	1	0	0	0	0					
17	United Kingdom	0	0	0	0	0	0	1					
18	Netherlands	0	0	0	0	1	0	0					
19	Belgium	1	0	0	0	0	0	0					
20	Germany	0	0	0	1	0	0	0					

On the left in Column A we see the original column. Each row denotes a client and the values in column A tell us where the customer is from.

On the right in Column K we see a list of unique values of the country column (A), which is the original column.

- This unique list is retrieved by the formula = SORT(UNIQUE(A2:A20))

The columns B to H indicate whether the client from this row is from one of those countries. Thus column B asks the question “Is this customer from Belgium or not?”

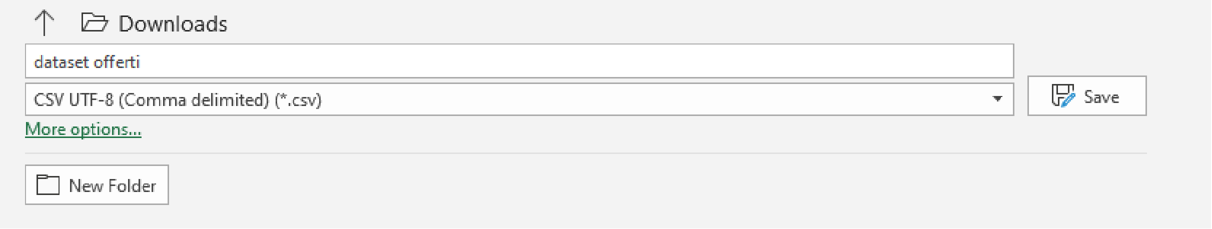
- The used formula is =IF(A2=K2;1;0), which translates to: If the value in A2 (the original country column, in this case Germany) is equal to the value in K2 (which is Belgium), write down 1, if they are not equal write down 0.
- Can you figure out the formula that would be in cell F10?

____

Step 4. Save as CSV file

Once most of the data is numerical and all cells have been filled with consistent data. The file can be saved as a .csv file and is ready for exploration!

- In Excel: File > Save As >



Downloads

dataset offerti

CSV UTF-8 (Comma delimited) (*.csv)