Clustering is the first data super-power a company should learn

Explaining the basic of clustering and why it could be valuable for you

Companies that use their data for effective decision making are set to grow their business much faster than their competitors. Whether your company stores its data using excel sheets or a full database system doesn't really matter. What matters is this:

Are you using your data to its full potential?

Now you might ask: "What Data Science technique can my company start with to get more insight out of my data?" Regardless of what kind of business you own, I would answer: **"Clustering."**

It is the technique behind Amazon's success with online sales. It is how Facebook and Google figure out what advertisements may apply to you when you visit their site. For many smaller companies, it has been a central tool for customer analysis and has helped them radically improve their marketing and sales strategies and sometimes even their product offering.

Most of all, it's an intuitive technique. Anyone with just a little data can get started with it.

In this article, you will be introduced to what clustering is. We will explore how and why successful companies like Netflix, bol.com and Amazon use clustering at the core of their business. More interestingly, we'll see how even a small restaurant can use it to make smarter decisions and grow their business. Lastly, you will learn that it is easy for even the smallest businesses to get started with clustering.

Interreg North Sea Region FBD
European Regional Development Fund
EUROPEAN UNION
FUTURES BY DESIGN
futuresbydesign.net

What is clustering?

Clustering methods create groups out of the data points in your data, based on their **similarity**. This process finds structure in data that doesn't have much structure by itself. That means that we could use it on almost any data that a small or big company has already lying around.

Initially, we may not know exactly what kinds of patterns we are looking for. That's okay. This is exactly what clustering will help us with.

Having created 'clusters', you can further investigate interesting patterns in your company's data. You can use the clustering results to ask valuable, pointed questions about the patterns and groups it has uncovered. This can have tremendous value for growing your business with actual data driven insights.

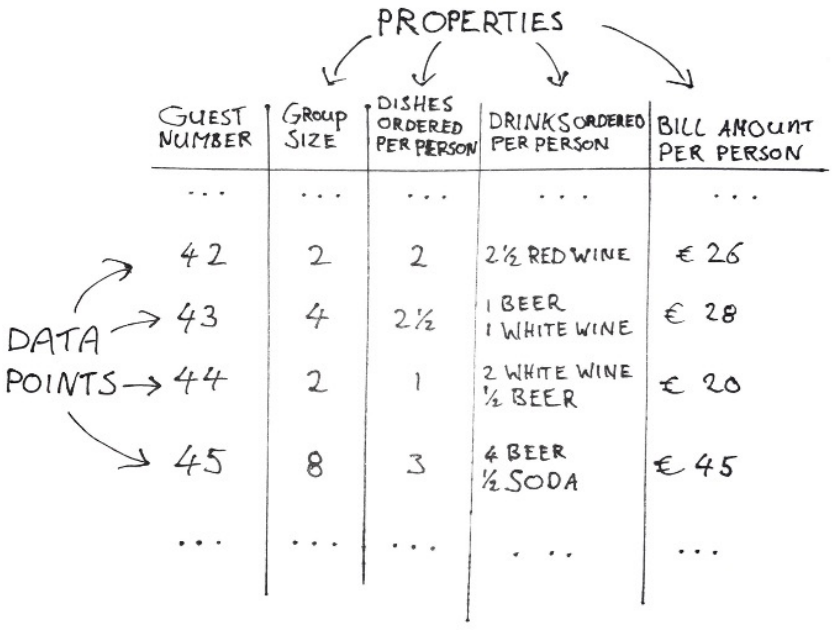
We'll get a better feel for how clustering would work in the next chapter.

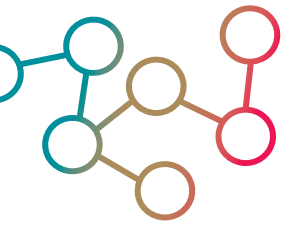
Clustering in a small restaurant

Let's look at a cozy little place called *"Restaurant La Estadística"*, as an illustrative example.

Like most restaurants, La Estadística uses an order system to inform the chef what to prepare. It also prints a receipt for the guests at the end of their visit.

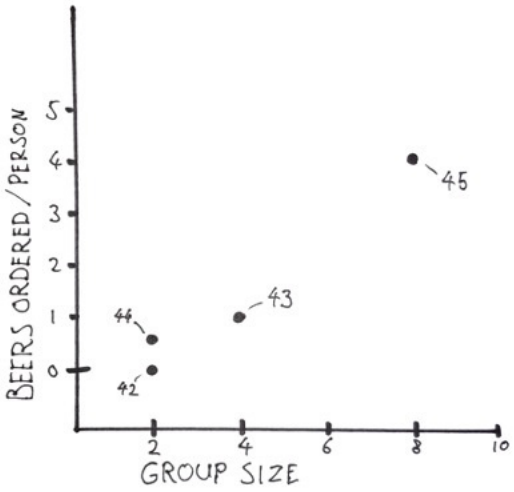
The system stores that data in a table. That table might look something like this:





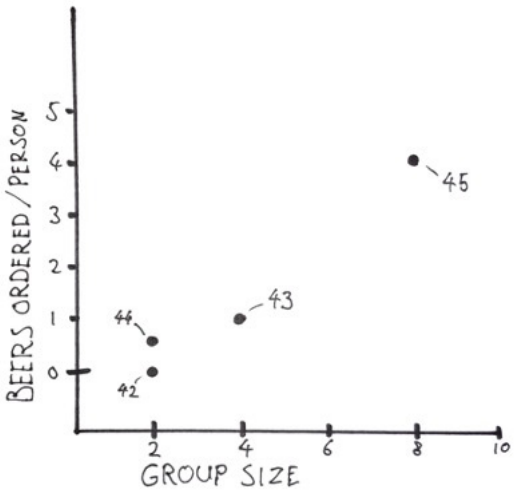
We see each group of guests represented as a row in the table. We can view each row as one **data point**. Every data point has **properties**. These are denoted in the columns.

We can manually create a clustering by drawing out all data points based on some of their properties in a figure. Let's pick *Group Size* and *Beers ordered*.

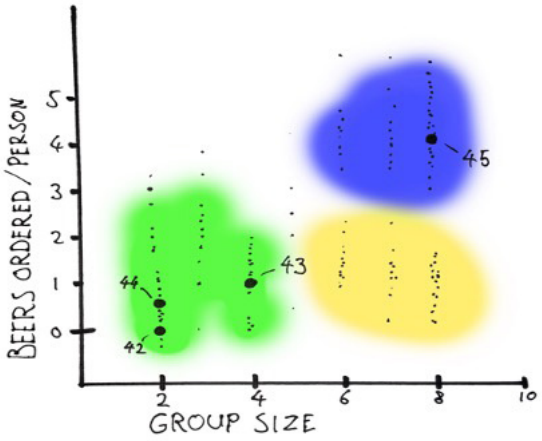


We can immediately see that the data point of 8 guests was quite different from the other three data points, if we look at how much beer they ordered per person.

Now we have only drawn four of the data points in the ordering system. The restaurant actually has many more. Let's draw the rest of the data points too.



You might already notice a pattern. Let's finish our manual clustering by adding a splash of color.



Can you see the general groups our clustering distinguishes?

The groups you see are called **clusters**. Clusters are groups of data points that are close together. This means that one cluster will contain guests that are very similar to each other, while guests that are less comparable will end up in different clusters.

A quick note on clustering algorithms

We were able to draw our data points on paper to obtain a simple clustering, but when we want to consider more than two properties in one clustering, we quickly run into trouble because it is hard to visualize more than 3 axes.

This is where clustering algorithms come in. You don't need to know exactly how they work, but it's useful to have a general idea of what they do.

An **algorithm** is a program that takes in some data points and produces an answer as a result. For example, the algorithm for summing two numbers takes in two numbers as input, say 1 and 2, and produces an answer; 3.

A clustering algorithm produces an answer to the problem:
Assign each data point in the input data to a cluster. Similar data points should end up in the same cluster, but data points that are very different must end up in different clusters.

A clustering algorithm will produce a far more exact answer than the manual clustering we made above. They can also take into account hundreds of properties at the same time, instead of just two. But the idea remains the same. We are still assigning similar data points to groups.

There exist a couple different flavours of clustering algorithms. Some handle data with many properties per data point better. Others handle weird shapes of clusters in the data better. But they all find patterns and a structure in data that, by itself, doesn't have any structure at all.

This structure is called a clustering.

Finding business value with clustering

Back to our example. With the clustering, the chef of La Estadística can easily filter on the blue cluster as seen above. He can quickly find that this cluster, on average, pays more per person than other clusters. He then notices that they often order more beer and more meat than other types of guests would.

The chef decides to introduce a group menu made specially for these guests. He includes some good meaty recommendations, combined with a discount on special beer when guests order at least three dishes. As a result, guests are much more satisfied with their food and the experience, and become returning customers. Moreover, the waiters spend less time collecting orders from big groups, because of the standardized menu.

The restaurant owner also starts to use the clustering to predict how much meat they should stock up on, based on the kinds of reservations they have. They know that bigger groups will very likely order a lot of meat, so the restaurant can prepare for those groups more efficiently.

Finally, the restaurant uses the cluster analysis on the demographics of their customers. Using their clustering, they can effectively analyse what types of customers have the most value for them. They can use that information to target the right people to show them Facebook advertisements.

Clustering applications everywhere

The illustration of *La Estadística* is just one of many examples of how clustering can help a business grow. To help you get an idea of the possibilities, here's a list of applications from all kinds of companies.

A webshop such as bol.com can use cluster analysis to group similar online visitors together and make effective recommendations for products they might like.

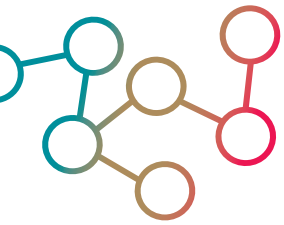
- Netflix uses cluster analysis to group both users and movies and show a useful variety of movies you might like, based on your user profile, genre preferences and preferences of comparable users.

- Airline companies use cluster analysis for customer support. It is possible to represent text in a numerical format, and with this, the airline can automatically group large numbers of questions and complaints together. These complaint categories can then be monitored and solved in a structured way. That is also why chatbots often give very generic answers.

- Housing companies use clustering to segment their customers, and predict outliers that might fail to pay their rent on time.

- Transport companies use cluster analysis on geographical data to discover the optimal central locations to build distribution centres for their goods.

- Law firms use clustering to group similar legal documents together and find cases that are similar to the one they are preparing.



Where to start?

You are probably already brewing on ideas on how to cluster your data. I'll give you some pointers to get started.

First, start thinking about what data you have right now. You need some data in a digital format. Think about the properties you already have for each data point.

Think about what patterns you expect to see if you would draw the data points in a figure. What properties might distinguish one group from another group?

Now think about what you could do with that information. Could you start making decisions that increase the value you can provide for your customers?

As a last note, like all Data Science techniques, don't expect clustering to hand you everything you need on a silver platter. Once you have a clustering, it is up to you to interpret it. You will need to investigate each cluster as to why it is so different from the rest and what makes that group interesting for you. That is the last essential step to get value out of the clustering and your data.

Conclusion

Companies that haven't started using their data to make decisions have more difficulty to fit their customers' needs very well. Clustering is a great first step to start using your data and make useful insights available to your company's team.

Don't overestimate the effort needed to get started with clustering. And don't underestimate the value that it can have for your business. Clustering is the first data superpower that a company should learn.

I expect that a couple of ideas might already have sprung up in your mind about what you could do with clustering. Write them down! You can then talk to a data scientist about your ideas and how to extract information out of your data. A first clustering prototype is quickly realized, and will show you what the potential is for your company.

Good luck on your data driven journey!