The principles of cluster analysis

Table of contents

Learning outcomes:

LO1: Define cluster analysis, when it is applied, and its assumptions.

LO2: Discuss the application of cluster analysis to business.

1. Introduction

South Africa has 11 official languages, each with their own origins and dialects. Although each is different, some share similarities, for example, in the origin of words and sentence structure. If you wish to learn another language, but want to find the one that would be easiest to grasp, you might begin by grouping languages together based on certain shared characteristics. In this way, you would not only gain more insight into the languages, but also more clarity on which might be the easiest to learn, based on how similar it is to your native language.

Grouping based on similarities (or dissimilarities) is used in many ways. Biologists typically group plant or animal species into specific taxonomies based on shared characteristics, while medical practitioners group similar clinical symptoms to diagnose a patient and identify the most effective form of treatment. Local governments might group municipalities together based on the similar types of services they require, and social scientists might group together survey respondents based on similarities in education, income, or age as a means of understanding why a population responds in certain ways.

This lesson will examine a technique that allows you to group similar objects: cluster analysis. You will also be introduced to some examples of how cluster analysis can be applied in business and the value it can add in making business-related decisions.

2. Defining cluster analysis

Broadly speaking, cluster analysis refers to a group of multivariate analysis techniques that seek to group objects based on similar characteristics. Essentially, cluster analysis is used to group objects in such a way that the objects within each cluster are more similar to each other than to those in other clusters.

Cluster analysis can be likened to factor analysis in that both techniques attempt to assess structure and help to segment data. Like factor analysis, cluster analysis does not define a dependent variable and instead is an exploratory technique used to gain more insight into the data. Cluster analysis differs from factor analysis, however, in that it attempts to group **objects** (such as products or customers), while factor analysis attempts to group **variables** (for example, the socioeconomic status or education level of a group of respondents). Furthermore, cluster analysis groups objects based on a predefined distance metric, whereas factor analysis groups variables based on the degree of correlation between them (Hair et al., 2014).

To aid your understanding of what can be considered an “object”, consider a scenario where you are trying to measure the similarities of different cities based on population size. Each city in this example would be considered an object. The characteristics that define those cities (in this case, population size) would then be used to decide which cities are similar and which are significantly different.

Since cluster analysis groups objects based on their degree of similarity, it is helpful to define what constitutes similarity or dissimilarity:

- **Similarity:** This seeks to measure how closely objects resemble one another. If two objects have a high similarity value, they are considered very similar, whereas if they have a low similarity value, they are considered very dissimilar.

- **Dissimilarity:** This is the opposite of similarity. If two objects have a high dissimilarity value, they are considered very different, whereas if objects have a dissimilarity value of zero, they are considered identical.

(Hair et al., 2014)

These two concepts can be formally defined in most cases. Similarity values are measured between 0 and 1. Continuing with the previous example, if two cities, *a* and *b*, exhibit perfect similarity (i.e., they are the same city), the similarity value would be defined as:

$$s_{ab} = 1$$

Using this definition, you can deduce that the dissimilarity value would be defined as:

$$d_{ab} = 1 - s_{ab}$$

For any dissimilarity measure, the following is required:

$$d_{ab} \geq 0$$

$$If\ d_{ab} = 0\ then\ a = b$$

$$d_{ab} = d_{ba}$$

(Er, n.d.)

These equations simply show that dissimilarity can only be measured by positive values, with a minimum of 0. When the dissimilarity value is 0, this indicates that the objects are identical. Furthermore, if object *a* differs from object *b* by a certain number of units, we can deduce that object *b* differs from object *a* by the same number of units (Er, n.d.).

2.1 Business applications

Cluster analysis has multiple applications in different disciplines, including business. As with factor analysis, companies can use cluster analysis to make sense of large amounts of data. The results of this analysis can then be used to inform business decisions or to enhance business processes that benefit the company.

Some of the use cases for cluster analysis in business include:

- **Market or customer segmentation:** Companies can use cluster analysis to group similar customers together into groups known as customer personas. The characteristics used to create these groups can be derived from a variety of sources, including surveys, marketing and sales data, and performance measurements. This gives companies the benefit of creating more effective and targeted marketing campaigns through personalization (Optimove, 2019), allowing businesses to address the unique concerns, needs, and expectations of their customer base more effectively.

- **Strategic management research:** This research attempts to capture the organizational profile of a business based on leadership, environment, performance, and strategy (Ketchen & Shook, 1996). Through this multi-dimensional understanding of the way a business operates, strategy managers can provide insight on the changes necessary for business success based on configurations that have proven successful.

- **Fraud detection:** With the rise of credit card use, fraudulent transactions have increased substantially, and financial companies have sought ways to address this inconvenience. Cluster analysis has proven effective in allowing companies to group “good” transactions into a single cluster, based on similar transactional patterns and behaviors. In the event that a transaction does not conform to these patterns, the company is alerted to possible fraud and can take the necessary action to prevent it (Agarwal & Upadhyay, 2014). Interestingly, this approach has also been used in developing email spam filters, detecting cancerous (or fraudulent) cells and potentially fraudulent telecommunication.

- **Business recommendation algorithms:** Businesses can often include clustering-based methods into algorithms that improve user experience. This is most evident in Spotify and Netflix, both of which use recommendation algorithms based on the customer’s history and the histories of similar customers. This helps these companies to curate a list of recommendations best aligned to a specific group of customers with shared preferences (Huq, 2019; McFadden, 2019).

- **Document analysis, information retrieval, and organization:** Business processes can also be improved by employing clustering-based methods. In companies where large volumes of documents are used, it can be useful to implement an algorithm that studies the text of the documents and clusters documents with similar content together. In this way, document retrieval becomes quicker, more effective, and more precise (Bellot & El-Bèze, 2000).

2.2 Cluster analysis example

Now that you have learned what cluster analysis is and how it can be applied in business, engage with an example of how clustering analysis works on a practical level. Consider the grouping of a collection of trading cards featuring popular sports figures. Each trading card can be considered an object, and there are seven different trading cards with face values of 5, 10, 20, 50, 150, 175, and 400. What criteria do you think you could apply to group them together?

A good starting point may involve grouping the trading cards based on their face value. To do this, you would divide the cards into groups of 5, 10, 20, 50, 150, 175, and 400 face-value cards, respectively.

Now that you have clustered the trading cards by face value, what other criteria might you use to further group the cards? What about the color or the category of sport, such as water sports, catching sports, combat sports etc.?

Alternatively, instead of grouping by color or sport, you might want to group the cards by the inherent value or expense, with cards of closer values grouped together. You might group cards as follows:

Group 1: Cards of values 5, 10, 20

Group 2: Cards of value 50

Group 3: Cards of value 150 and 175

Group 4: Cards of value 400

In this final method, the trading cards have been clustered based on the distance between their inherent characteristics. Essentially, the final clusters represent distinct groups that have the least amount of intra-variability but maximum inter-variability. In other words, this means that the cards are similar within their groups but highly dissimilar between groups.

Pause and reflect:

Having gained an understanding of the concept of cluster analysis, can you think of ways this technique might be used within your business or career context?

3. Steps of cluster analysis

Before delving into how cluster analysis is performed, it is important to understand that clustering-based methods should be used with a conceptually-defined goal in mind. This goal can either be to reduce a large volume of data into a meaningful business output, or to test hypotheses about the data to help businesses make informed decisions (Hair et al., 2014).

For example, a sales department may provide a company with customer perceptions of a range of products they sell. In its raw format, however, this data is meaningless – unless it can be reduced. Cluster analysis can achieve this by segmenting customers into unique customer profiles based on similar characteristics. The company can then address the needs of specific groups of customers, especially those with poorer perceptions of the business.

On the other hand, a business might use cluster analysis to test the hypothesis that customer attitudes can help separate one type of consumer from another. This would allow the company to effectively market its product to different groups of customers.

The next section of the notes will expand on how cluster analysis works, including the use of distance metrics, clustering techniques, and deciding on the optimal number of clusters for your analysis.

3.1 Measuring similarity and dissimilarity

Consider the earlier trading cards example where similarity (and dissimilarity) were calculated based on a distance measure (in this case, the difference between the inherent values). This is visualized in Table 1, which calculates the distance between each type of trading card.

Table 1: Distances between different types of trading cards.

	5	**10**	**20**	**50**	**150**	**175**	**400**
5	0	5	15	45	145	170	395
10	5	0	10	40	140	165	390
20	15	10	0	30	130	155	380
50	45	40	30	0	100	125	350
150	145	140	130	100	0	25	250
175	170	165	155	125	25	0	225
400	395	390	380	350	250	225	0

Where the values equal 0, there is no dissimilarity and the objects are identical. All values above 0 indicate that some dissimilarity exists, the value of which is measured by the card's distance from one another (the difference in value, in this case). You will notice that the two cards with the greatest similarity are those with values 5 and 10. In contrast, the cards with the greatest dissimilarity are those with values 5 and 400.

These distances are the basic premise of using distance in cluster analysis. This course will address three distance metrics as a means of measuring dissimilarity between objects expressed as numerical variables:

1. **Euclidian distance:** This distance metric is one of the most commonly used in cluster analysis. In simple terms, it measures the distance between two objects as if a straight line were being drawn between them with a ruler.

2. **Manhattan or city-block distance:** This method places the data points on a 2D grid and measures the distance between the coordinates of each, as if you were walking along a city grid rather than in a direct, straight line.

3. **Correlation distance:** Correlation can be defined as either a positive or a negative relationship between two variables. It is, essentially, a measure of how two or more variables fluctuate together. Therefore, the correlation can often be indicative of how similar two variables are to each other. In the previous example, you might calculate the correlation between the two rows of the data matrix. It is worth noting that correlation distance is only a reliable measure when multiple variables are compared. This is because correlation values tend to be more unstable with fewer attributes. Recall that correlation is a linear or non-linear measure of relationship. However, in cluster analysis we are interested in measuring the similarity between "observations". Therefore, correlation-based distance considers two observations to be similar if their

features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. This is an unusual use of correlation, which is normally computed between variables as we did in Linear Regression Analysis. Here it is computed between the observation profiles for each pair of observations. This indicates that if we have 50 observations with 4 variables, we will have a 50x50 correlation matrix, not 4x4 as we would in regression analysis. Ultimately, Correlation-based distance focuses on the shapes of observation profiles rather than their magnitudes (James et al., 2013:396).

While these represent only a few of the metrics available, they are the core metrics that will be studied in this course. In cases where the data set includes both continuous and categorical variables, Gower distance can be used to measure dissimilarity between non-metric variables.

It is important to remember that cluster analysis is an exploratory technique, and that there is often no correct or incorrect strategy when selecting a clustering method or distance metric. Rather, the results of the clustering solution should be assessed in reference to the requirements of the business problem being addressed.

Explore further:

To gain a better understanding of other available distance metrics for performing cluster analysis, read this Analytics Vidhya article on the types of distances in machine learning.

3.2 Techniques for forming clusters

As discussed, cluster analysis is considered an exploratory technique, and it can involve the use of multiple different algorithms as a result. The use of a specific algorithm will depend on the type and quality of data set, as well as the specific goals of the analysis. The different clustering-based methods are illustrated in Figure 1.

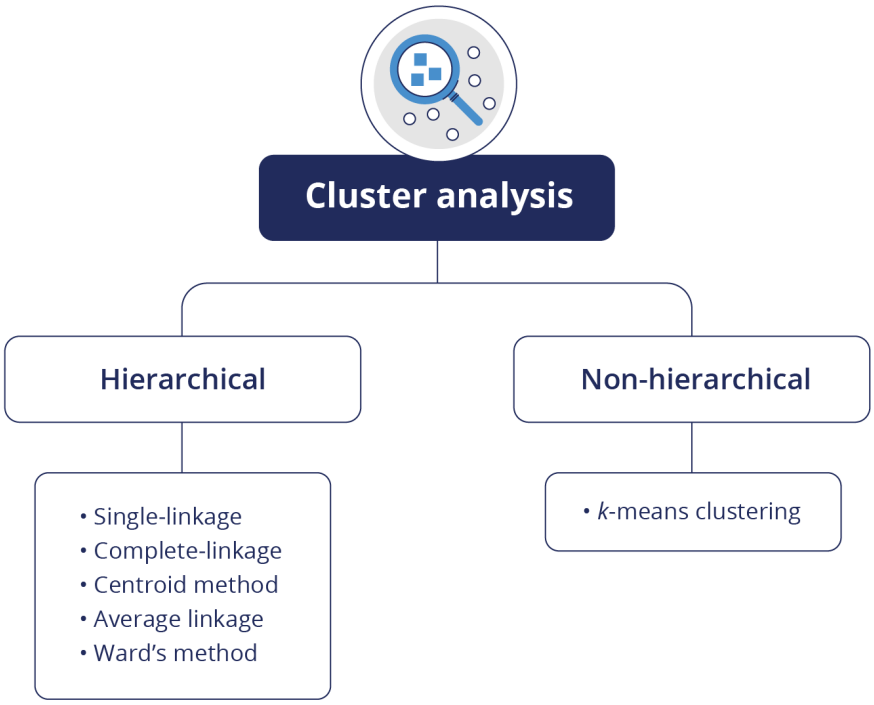

Figure 1: Clustering methods.

It is important to note that this is not an exhaustive list, nor are all these methods covered in this course. Cluster analysis is continually evolving and adapting, with newer methods being developed all the time. As you will notice in Figure 1, there are two major forms of cluster analysis: **hierarchical** and **non-hierarchical**. These are explained in further detail in the following sections, and you will see how each form works in practice in the IDE screencast in the next unit.

3.2.1 Hierarchical: Single-linkage, complete-linkage, and the centroid method

Hierarchical clustering can be divided into **agglomerative** and **divisive** methods. Agglomerative clustering considers each object as its own cluster and involves the sequential merging of two or more clusters to form a new cluster. Once these clusters merge, they cannot be separated and are considered a new, single cluster. Divisive clustering operates in the reverse, considering all objects to be in the same cluster and sequentially splitting this cluster into new, smaller clusters.

This course will focus on agglomerative methods, which include single-linkage, complete-linkage, and the centroid method:

- **Single-linkage:** With single-linkage clustering, each object is considered its own cluster. These clusters are sequentially merged based on the clusters closest to one another, until all objects belong to the same cluster. This is typically referred to as a bottom-up approach, and the merging of clusters is based on the dissimilarity value between two objects. With this method, dissimilarity is defined as the minimum of the dissimilarities between individual objects in each of the two clusters. In other words, the dissimilarity is taken as

the distance between the two closest objects within the two closest clusters, as illustrated in Figure 2.

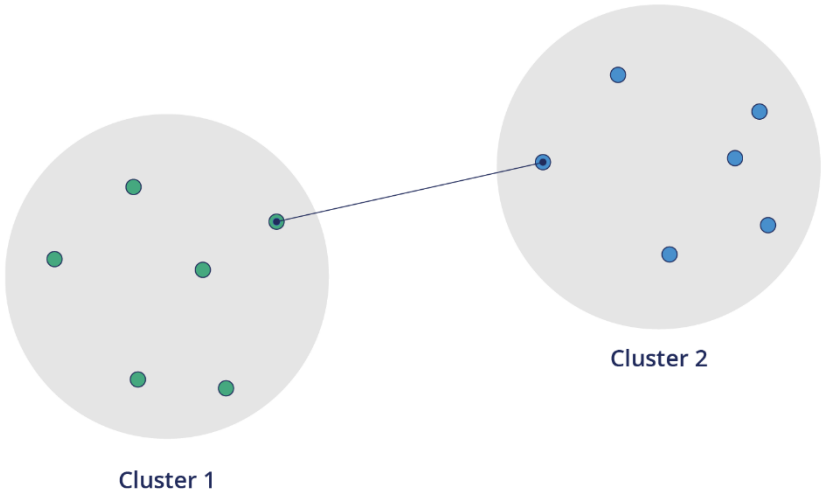

Figure 2: Single-linkage clustering.

- **Complete-linkage:** Complete-linkage clustering is similar to single-linkage in that each object is considered its own cluster, and these clusters are sequentially merged until all objects are housed within a single cluster. Once again, the two closest clusters are grouped together to form a single, new cluster. However, the difference lies in that the dissimilarity is defined as the maximum of the dissimilarities between individual objects in each of the two clusters. In other words, the dissimilarity is taken as the distance between the two farthest objects within the two closest clusters, as illustrated in Figure 3. This method is also referred to as farthest neighbor clustering.

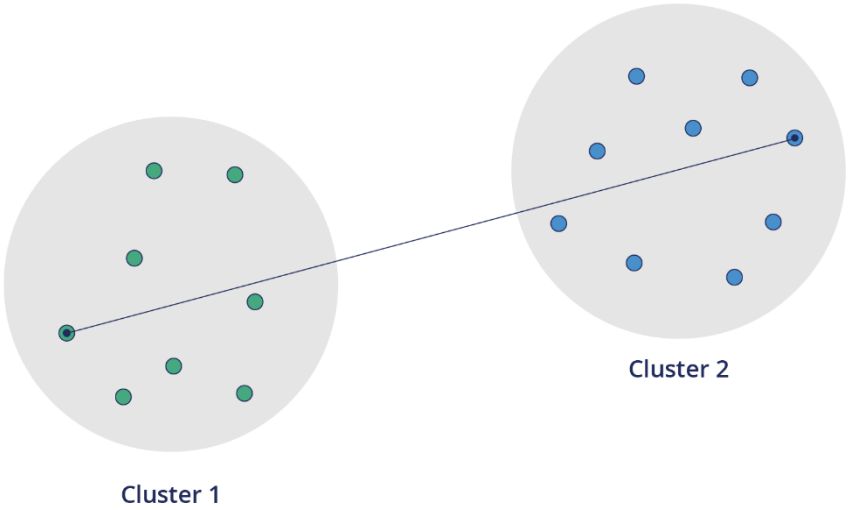

Figure 3: Complete-linkage clustering.

- **Centroid method:** Again, the centroid method considers each object its own cluster and sequentially merges two clusters together until all objects are

housed within the same cluster. However, in this method, the dissimilarity is defined as the dissimilarity between the centroids of the two clusters. A centroid is simply the center, or the average of each cluster, based on all the objects within that cluster. This method is illustrated in Figure 4.

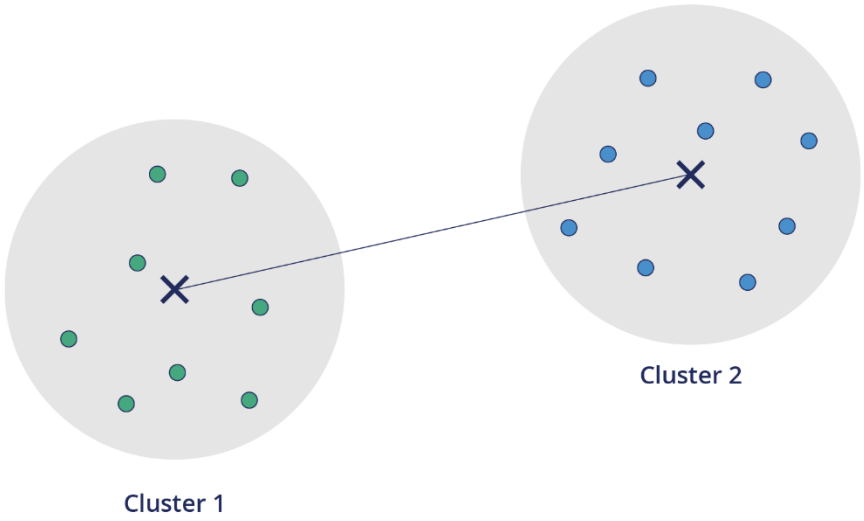

Figure 4: Centroid-based clustering.

Explore further:

To learn more about the concept of hierarchical clustering, read the Analytics Vidhya guide on hierarchical clustering and how to perform it in Python.

3.2.2 Non-hierarchical: *k*-means algorithm

Non-hierarchical clustering is typically an iterative algorithm that adjusts which clusters individual objects belong to during each step of the iteration. To achieve this, a data analyst will be required to specify the number of clusters based on the goals of the analysis.

The most commonly used non-hierarchical method is *k*-means clustering, which is an iterative and centroid-based algorithm. This method first requires you to define the number of clusters (*k*) desired for the analysis. The algorithm then randomly assigns data points as centroids within clusters, and any observations similar to those centroids will form part of a new cluster. After the first iteration, new centroids are randomly assigned, and the process is repeated until convergence is reached; in other words, the process continues until the cluster centroids no longer change.

3.3 Determining the number of clusters

Once a clustering technique has been used on a data set, the next step is to decide how many clusters are needed from the analysis. This decision is often not within the control of the data analyst, but rather specified by other departments within a business.

For example, the marketing department may already have developed targeted advertisements directed towards three different groups of customers (categorized by

age, location, or income bracket). In this instance, it would not make sense to group customers into more than three clusters.

In other cases, the number of clusters may not have been specified. Rather, the business uses this method as an exploratory technique to improve their understanding of their customer base. In this instance, the number of clusters should be determined subjectively, based on the appearance of the clustering tree and evidence of any noticeable clusters.

There is no right or wrong decision when selecting the number of clusters; this will depend on the analyst, the business goals, and how the data appears following the analysis. However, it is important to note that *k*-means clustering is the exception to this rule and relies on specifying the number of clusters at the start of the analysis.

3.4 Comparing different clustering techniques

As you have learned, when a researcher begins their analysis of a data set, they are faced with several clustering options. How do they decide which technique to use? How will they know whether the results are effective in describing the data accurately? Given that cluster analysis is an exploratory technique, there is no definitive answer to these questions. Once again, the technique used will depend on the goals of the analysis and the quality of the data.

The following should be considered when using **hierarchical clustering**:

- **Simplicity and efficiency:** Since the results of hierarchical clustering can be visualized with a clustering tree (or dendrogram), the interpretation of the results is often more intuitive and simpler to understand than non-hierarchical techniques. Furthermore, hierarchical clustering provides a time-efficient way to view multiple clustering options, including the use of different distance metrics.

- **Extensive similarity measure options:** The widespread usage of hierarchical techniques has resulted in the development or iteration of a multitude of similarity measures that can account for unique data sets, such as those with extensive multicollinearity. This means that there are often several more distance measure options available to the analyst when performing hierarchical clustering.

- **Sensitivity to outliers:** Outliers have the potential to mislead the analyst or skew the data. This is particularly true for complete-linkage clustering, which uses the maximum dissimilarities between individual objects within a cluster. As a result, the analyst will often be required to remove outliers from the data set. However, this can distort the results of the analysis if extreme care is not taken.

- **Large storage requirements:** While hierarchical clustering is time-efficient, large data sets present the problem of requiring large storage volumes.

(Hair et al., 2014)

In contrast, the following should be considered when using **non-hierarchical clustering**:

- **Lower sensitivity to clustering design:** Non-hierarchical techniques are inherently less sensitive to outliers, the similarity measures used, and whether extraneous variables are included in the analysis.

- **Ability to analyze large data sets:** Since non-hierarchical techniques only compare the distance between observations and the centroid, rather than the distances between all observations, this method can be used to analyze large data sets.

- **Lower efficiency with a greater number of clusters:** Non-hierarchical methods typically require the analyst to predefine the number of clusters desired in the analysis. However, the greater the number of clusters, the less efficient the analysis. This is because establishing each cluster solution involves its own analysis, whereas hierarchical clustering provides all possible cluster solutions from a single analysis.

(Hair et al., 2014)

To test your understanding of how cluster analysis works, the methods and distance metrics available, and the various business applications, answer the following quiz questions.

---

1. Select two options that best describe cluster analysis.

a. Cluster analysis is an exploratory technique, and the effectiveness of the clustering solution often relies on the business question being solved or addressed.

> Correct, well done. There is often no correct strategy for selecting a clustering method or distance metric. Rather, cluster analysis is an exploratory technique dependent on the business question being addressed. For example, if the marketing department requires customers to be separated into three unique clusters, you can select a method that best aligns with this requirement.

b. The appropriate number of clusters required in a clustering solution is always known.

> Incorrect. While it is true that the number of clusters can be specified by external sources (such as different departments within a company), it is often the case that an analyst is required to make an objective decision on the number of clusters, based on whether the clustering solution effectively separates the data.

c. Euclidean distance is a dissimilarity measure that places the data points on a 2D grid and measures the distance between two coordinates as if you were walking along a city grid rather than in a direct, straight line.

> Incorrect. This defines Manhattan (or city-block) distance. Euclidean distance measures the distance between two objects as if drawing a straight line between them with a ruler.

d. Cluster analysis attempts to group data observations together in such a way that intra-variability within clusters is minimized and inter-variability between clusters is maximized.

Correct, well done. Cluster analysis attempts to group together objects with a high degree of similarity within each grouping (minimized intra-variability within clusters), and also to separate the data to ensure that the groupings or clusters are highly dissimilar from each other (maximized inter-variability between clusters).

---

2. The product development department of a company is hoping to create novel, packaged product offerings, based on products that are often purchased together. To provide more insight into potential new product offerings, the business uses cluster analysis to group products based on shared characteristics and purchasing patterns. This is an appropriate use for cluster analysis.

True

Correct, well done. Cluster analysis could help the business to group products into well-defined clusters that share similar characteristics, such as product application, price, and customer purchasing patterns (for example, whether two or more products are frequently bought together). This would help the business to ascertain whether it can package certain products together and provide novel, unique, product offerings.

False

Incorrect. Cluster analysis could help the business to group products into well-defined clusters that share similar characteristics such as product application, price, and customer purchasing patterns (for example, whether two or more products are frequently bought together). This would help the business to ascertain whether it can package certain products together and provide novel, unique, product offerings.

The quiz **ends** and the following instructional content is displayed.

4. Assumptions of cluster analysis

Many statistical techniques need to satisfy assumptions that represent a population. Cluster analysis, on the other hand, assesses the structure of the data, which means that it usually does not need to satisfy the common assumptions of other techniques such as normality, linearity, or similar variance. However, there are three unique assumptions that should be considered with cluster analysis:

1. **Representativeness of the sample:** A data analyst rarely knows whether sampled data closely matches the observations for the entire population. Rather, it is assumed that the underlying structure of the data is representative of the population from which it was sampled. However, this assumption can be influenced by outliers, which may skew the data and indicate that the population

was undersampled. Therefore, it is imperative that the analyst confirms the representativeness of the sample, and that the results can be generalised accurately to the broader population. This can be achieved by ensuring that appropriate sampling methods are used during the data acquisition phase (Hair et al., 2014).

2. **Multicollinearity:** While other techniques can be negatively impacted by multicollinearity, you may recall that cluster analysis can use correlation as a distance metric to measure similarity. In other words, correlation between variables can help to separate observations into similar groups. However, variables that are **substantially** correlated can cause the observations to be inappropriately weighted, skewing the result output. Consequently, the analyst should either remove highly correlated variables or use distance metrics that can account for the existing multicollinearity, such as Mahalanobis distance (Hair et al., 2014).

3. **Data standardization:** Often, data sets will include variables measured on different scales. For example, consider weight measured in kilograms and height measured in centimeters. As cluster analysis relies on measuring the distance between observations, if one variable is measured on a different measurement scale, it will significantly influence the dissimilarity measures. As a result, it is assumed that the data is transformed in a way that allows comparisons to be made directly. This is usually achieved by subtracting the variable's mean value and dividing it by the standard deviation – otherwise termed the "Z-score".

Note:

Now that you have an understanding of the principles of cluster analysis, use the Unit 3 Notes to help you interpret the results when completing the IDE notebook.

5. Conclusion

The ability to segment a data set into well-defined clusters that share similar characteristics is an important technique to use in a business context. For example, cluster analysis can help group customers together to better understand their needs, or it can be applied to the products a company sells to identify novel product offering packages.

In this lesson, you were introduced to how cluster analysis operates, the different methods and distance metrics that can be used, and how it is an exploratory technique that analyses the underlying structure of the data. This knowledge makes it easier to understand how data segmentation can provide important business insights.

6. Bibliography

Agarwal, S. & Upadhyay, S. 2014. A fast fraud detection approach using clustering based method. *Journal of Basic and Applied Engineering Research*. 1(10):33-37.

Bellot, P. & El-Bèze, M. 2000. *A clustering method for information retrieval.* (Technical Report IR-0199). Avignon, France: Laboratoire Informatique d'Avignon.

Er, Ş. n.d. Applied multivariate data analysis: Part 1 [STA3022F Lecture notes]. Department of Statistical Sciences, University of Cape Town.

Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. 2014. *Multivariate data analysis*. Rev. 7th ed. Harlow, UK: Pearson Education Limited.

Huq, P. 2019. *Music to my ears: De-blackboxing Spotify's recommendation algorithm*. Available: https://blogs.commons.georgetown.edu/cctp-607-spring2019/2019/05/06/music-to-my-ears-de-blackboxing-spotifys-recommendation-algorithm/ [2019, December 11].

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An introduction to statistical learning* (Vol. 112,). New York: Springer.

Ketchen, D.J & Shook, C.L. 1996. The application of cluster analysis in strategic management research: an analysis and critique*. Strategic Management Journal.* 17(6):441-458.

McFadden, C. 2019. *How exactly does Netflix recommend movies to you?* Available: https://interestingengineering.com/how-exactly-does-netflix-recommend-movies-to-you [2019, December 11].

Optimove. 2019. *Customer segmentation via cluster analysis*. Available: https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis [2019, November 25].