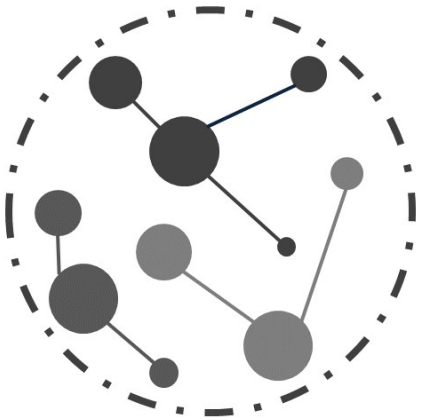
Clustering algorithms in the touristic sector:

Case study on travel agency data

Argyro Sioziou

A thesis presented for the degree of
Management Science and Technology



Management Science and Technology
Athens University of Economics and Business
Greece
20/02/2020

Clustering algorithms in the touristic sector

Case study on travel agency data

Argyro Sioziou

Abstract

Contents

Chapter 1

INTRODUCTION

1.1 Research Motivation

The fourth industrial revolution has come with many new technologies, and even though it has just begun, it is affecting people's lives and companies' behaviors in many aspects (Schwab, 2017). The latter are searching for new ways to address the former and engage them, while their expectations are continuously increasing. This leads companies of many industries, including tourism, to intend to deliver personalized experiences, using tools like marketing (So & Li, 2020). At the same time, the available data volume is increasing rapidly, giving the ability to enterprises worldwide to extract knowledge from them and create value (Cuzzocrea et al., 2017). This plethora of data provides an opportunity for marketers of the tourism sector, amongst others, to perform better segmentation, and get insights into their customers' needs using cluster analysis (D'Urso et al., 2019). The results of such techniques can also help businesses of the industry in their decision-making process and the creation of competitive advantage (D'Urso et al., 2016). However, this task is not that simple. Turistic data are to a large extent qualitative, which makes it hard for analysts to extract the maximum knowledge from them since clustering algorithms are in their majority used with quantitive data (Arimond & Elfessi, 2001). The process of eventually managing to extract the wanted information seems challenging. We could suppose though, that with the proper handling the results could prove rewarding for both the businesses and the customers. Based on this conjecture, we try to respond to this thesis how could a company of the tourism industry, using cluster analysis, take deal with its complex data and create value for itself and its customers.

1.2 Research Methology

To respond to the research objectives and questions, which were defined in the previous section, data of the back office system of a travel agency were used as a case study. The data were cleaned, reformed and metadata were extracted from them. The majority of the data were categorical. Next, two clustering algorithms were used to create meaningful groups. Before the analysis, a research on the algorithms was done in order to better understand their functionality. Afterward their results were analyzed and compared in order to conclude whether the results could be valueble and how.

1.3 Thesis Outline

The main goal of this thesis is to experiment with the use of clustering algorithms with categorical data and identify the value of its results in the formation of services, and the creation of marketing strategies in the tourism sector. Firstly, it defines machine learning and its categories, concentrating mainly on clustering and its applications in various sectors. Next, it analyses more in-depth two clustering methods, the kmeans, and agglomerative clustering. Further, in the last section of the research background, a thorough description of two clustering applications in the tourism sector are illustrated. Then, the case study description, available data, methodology, and results are presented. In conclusion, the results of the two algorithms are compared and their avail is discussed.

Chapter 2

BACKGROUND

2.1 Overview

2.1.1 Machine Learning

Machine learning is the field of study that focuses on training machines (e.g., computers) to identify patterns and derive logical conclusions (Bishop, 2006). It is technically an imitation of the human learning process and can be divided into different types based on the approach it uses to train the algorithm (Marsland, 2004, p. 5). The two main types are supervised and unsupervised learning (Dunham, 2002, p. 43):

- Supervised learning is used to classify instances to already known categories based on their characteristics. Algorithms that belong to this type are first trained on an already labeled with the possible categories dataset and, afterward, use their gained knowledge to classify new unlabeled datasets (Han et al., 2011, p. 24). Technically it is like a human trying to learn by first looking at all the correct answers and then, based on them, try to classify new ones (Marsland, 2004, p. 5). For example, a supervised learning problem would be images, each one containing one hand-written digit, and its label the number to which it corresponds. After training the algorithm with a labeled set, the algorithm would be capable of recognizing digits on images.

- Unsupervised learning, on the other side, is used to group data without knowing the labels beforehand and without any training proceeding. This type of learning allows the model to learn by itself. Usually, a person with knowledge in the sector is needed to interpret and extract the knowledge from the created groups (Han et al., 2011, p. 25). In this case, a human would only know whether or not his answer is correct, but nothing about the way to the right answer (Marsland, 2004, p. 5). The above example using unsupervised learning would be giving as input the images to the algorithm without labeling the digits. The expected output would be ten clusters, each one corresponding to one of the digits from zero to nine. A person inspecting the results would be in the position to recognize which is the digit shown on the images of each cluster.

Apart from those two main categories, two additional categories are worth mentioning, semi-supervised learning, and reinforcement learning.

- Semi-supervised learning falls between the above two types. It consists of a small amount of labeled data and a large amount of unlabeled data. The labeled data serve as initial training, and the results are using the unlabeled data to improve the process and, on many occasions, spot outliers (Han et al., 2011, p. 25). Following Marsland (2004)'s ratiocination, in real life, it would look like a person given a few correct answers as a starting point to solve more complex questions. Using the digits example, let us suppose that labeling all the pictures would be a very costly procedure, therefore someone would label a small amount of the images to use as the training set, and afterward apply the algorithm to the rest of the images to improve the model.

- Reinforcement learning is a more interactive method, as opposed to the previous ones. It uses unlabeled data and requires rewards or penalties based on its behavior. This type of machine learning takes decisions that lead to the maximization of the reward taken (e.g., a numerical value) (Sutton & Barto, 2015, pp. 2–3). Similarly to unsupervised learning, it is like giving feedback to a person, but not by telling it whether the answer is correct or not, but by grading its performance. Then that person using its sought knowledge would try to improve itself (Marsland, 2004, p. 5). Therefore, the digits paradigm would be similar to the one of unsupervised learning. However, instead of trying to discover hidden patterns, it would try to find the output with the maximized reward (or minimized penalty), making different decisions based on its previous experience.

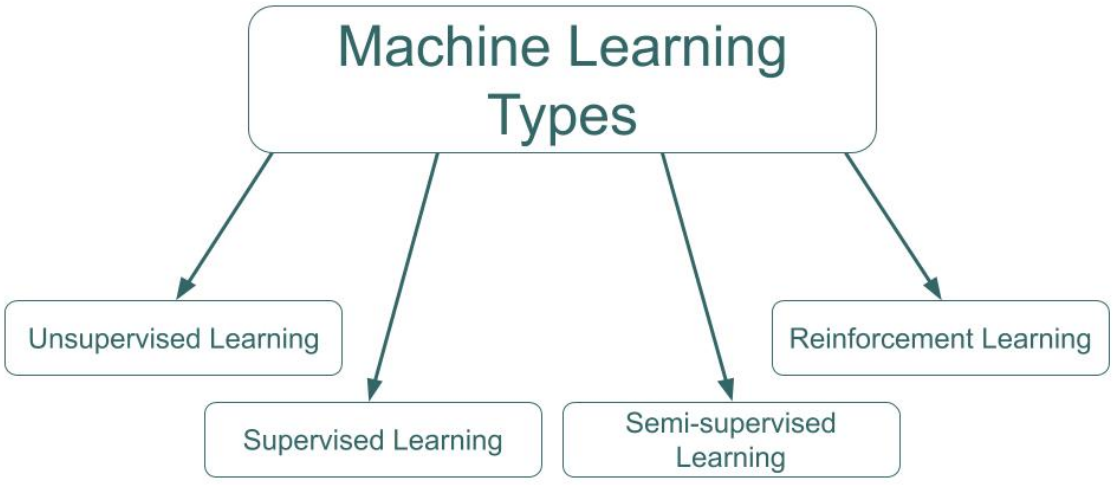


Figure 2.1.1: Types of machine learning.

2.1.2 Clustering

As indicated by the definitions given by Dunham (2002) and Tan et al. (2005), clustering or cluster analysis is a type of unsupervised machine learning. It groups instances to create coherent sets based on their similarities and dissimilarities. The aim is that instances that belong to the same sets are as much alike and as much different from the instances of the other sets as possible (Dunham, 2002; Tan et al., 2005).
As Kantardzic (2011) claims, humans are capable of detecting natural clusters from given data and compete with the clusters created by algorithms up until three-dimensional data. In real life, though, the dimensionality of the data examined is usually much higher than that, and therefore, if the cluster analysis were possible to be done by a human, it would not be as efficient (Kantardzic, 2011, p. 250). Since it is not humans that perform clustering, but algorithms, unexpected groups, and patterns can be discovered (Han et al., 2011, p. 444).
The usefulness of cluster analysis can be confirmed from its numerous applications in various fields of study. Some of them are mentioned by Everitt et al. (2011). Some of them briefly described are the following:

- Market research. A market researcher may want to use clustering to segment its audience and form its target groups. Furthermore, it can be useful to find the correlations between the financial measures of a company to its stock performance. Chakrapani (2015) describes an example in the market of sports cars. Whether a person will buy a sports car is not affected by demographics, but by its lifestyle, clustering can help to spot those lifestyles that lead to such a purchase (Everitt et al., 2011, p. 9).

- Astronomy. Astronomers need to discover distinct groups out of large data sets. For example, on Celeux and Govaert (1992)'s application, stars could be divided into groups differentiated by their volume and size based on their velocities (Everitt et al., 2011, p. 10).

- Psychiatry. Cluster analysis can group mental illnesses, detect patterns in people that commit suicide, investigate parasuicidal patients, and classify eating disorders. By further research on the results, new approaches can be proposed. Paykel and Rassaby (1978) found three different types of people that attempt suicide. They differ mainly on the cruelty of the methods they chose and their motivation behind the attempt (Everitt et al., 2011, p. 10).

- Bioinformatics and genetics. As Everitt et al. (2011) explain, the genetic code is complex, and its variations are responsible for the differences between organisms. In each cell, different genes are expressed, resulting in the production of several proteins. Combining different cell types makes an organism functional. In the case of alterations in the genetic code or malfunctions in those processes, diseases occur. Through clustering, it is possible to study the cause of a disease, investigate inheritance, discover the functionality of new gene sequences, and make research more efficient and affordable (Everitt et al., 2011, pp. 12–13). Ben-Dor et al. (1999) used clustering in their article to group genes that follow similar expression patterns in order to decrypt a gene's functionalities and regulative mechanisms (Ben-Dor et al., 1999).

Additional applications of clustering in the touristic sector will be analyzed more in-depth in section 2.4.

2.1.3 Types of Clustering

Clustering algorithms can be categorized based on two criterions. The first one is the relationships between the produced clusters of one iteration with the ones of a previous iteration. The second one is the number of clusters an instance can be a part of.
Based on the first criterion, Tan et al. (2005) describe the following types:

- Partitional clustering is a type of clustering which allows no overlapping between two or more sets of clusters, hence partitions the initial set to independent sets. This means that each instance can be part of exactly one cluster (Tan et al., 2005, p. 492). Therefore, it could be described as:

$$C_i \neq \emptyset,\ i = \{1, .., K\}$$
$$\bigcup_{i=1}^{K} C_i = X$$
$$C_i \cap Cj = \emptyset,\ i, j = \{1, ..., K\}\ and\ i \neq j$$

  where X is the initial set, K the number of partitions, and C_i, Cj the i^{th}, j^{th} clusters (Xu & Wunsch, 2005, p. 645).

- Hierarchical clustering produces clusters in each iteration, which are hierarchically connected to the clusters of the previous iteration. Based on whether the clusters are formed by combining clusters of the previous iteration (bottom-up) it is called agglomerative, or by decomposing clusters of the previous iteration (top-down) it is called divisive. This can occur by starting from one cluster, which contains all instances, and repeatedly partition the available clusters to even smaller ones (Tan et al., 2005, p. 492). Therefore, it could be described as:

$$C_i \in H_m, C_j \in H_l \atop m > l \Rightarrow \begin{cases} C_i \subset C_j \\ or \\ C_i \cap C_j = \emptyset \end{cases}, i \neq j\ and\ m, l = 1, ..., Q$$

where H_m, H_l partitions of the original set X, Q is the number of partitions, and C_i, C_j subsets of any of the Q partitions (Xu & Wunsch, 2005, p. 646).
Tan et al. (2005) mention three more types of clustering in their book, which are formed based on the second criterion and are the following:

- Exclusive clustering signifies that an instance can only be a part of only one cluster (Tan et al., 2005, p. 492). Let us define the binary variables x_i for each instance $x \in X$, where X are all the instances. If an instance x belongs to the i^{th} cluster x_i is set equal to 1, otherwise to 0. Then, if the number of clusters is n, we can describe each cluster by the function:

$$C_i = \{(x, x_i) | x \in X\}$$

where C_i is the i^{th} cluster, and which has the following restrictions:

$$\sum_{i=0}^{n} x_i = 1, x_i \in \{0, 1\}$$

- Overlapping or non-exclusive clustering signifies that an instance can be part of more than one cluster (Tan et al., 2005, p. 492). We can describe each cluster the same way we did in exclusive clustering, just by adjusting the restrictions:

$$C_i = \{(x, x_i)|x \in X\}$$
$$\sum_{i=0}^{n} x_i \geq 1, x_i \in \{0, 1\}$$

  where C_i is the i^{th} cluster, x is a specific instance, x_i is the binary variable that describes x in the i^{th} cluster, X are all the instances and n is the number of clusters.

- Fuzzy clustering signifies that every instance belongs to all the clusters (Tan et al., 2005, p. 492). As Kantardzic (2011) explains, the function that shows the membership of an element inside a cluster is called membership function (MF). The value of the MF that describes a given instance is a number between [0,1]. Each cluster can be described as follows:

$$C_i = \{(x, \mu_C[x])|x \in X\}$$

  where C_i is the i^{th} cluster, x is a specific instance, μ is the MF function, and X is the set that contains all the instances (Kantardzic, 2011, p. 416).

2.2 Kmeans

As Dunham (2002), Tan et al. (2005) explain kmeans is a prototype-based, iterative algorithm in which instances are assigned to a cluster in each iteration. The basic algorithm requires as input K points, called centroids, where K is the number of the desired clusters. To define the cluster to which one instance belongs to, the algorithm calculates its distance from all the centroids and assigns it to the closest one. Finally, using the produced clusters calculates the new centroids and repeats until it reaches the desired set. Each cluster's mean needs to be defined and recalculated in each iteration to redefine the new centroids (Dunham, 2002; Tan et al., 2005).
Basu et al. (2002) explain that kmeans is practically a minimization problem. The objective function (the function to be minimized) is the function that calculates the distances between one instance and its cluster's centroid. The smallest the sum of the distances of one cluster's centroid to its instances, the highest the homogeneity of the cluster (Basu et al., 2002, p. 2).
Disimalirites between data objects are their distance (Tan et al., 2005, p. 69), though there are many different methods of calculating the distance between two instances, the most common one used is the Euclidean distance (Xu & Wunsch, 2005, p. 648). Using the Euclidean distance to calculate the distances between the centroid of a

cluster, and the rest of the cluster's instances, for a one-dimensional dataset, the model is as follows:

$$d_{ij} = \sqrt{(c_i - x_j)^2}$$

where d_{ij} is the distance of the j^{th} element from the i^{th} cluster, c_i is the i^{th} centroid, and x_j is the j^{th} element (Tan et al., 2005, p. 69).
In the simple case where there is only one numerical value describing each instance the cluster mean can use the basic mean definition from statistics as follows:

$$m_i = \frac{1}{m} \sum_{j=1}^{m} (x_{ij})$$

where m_i is the mean of the i^{th} cluster, m is the number of instances that belong to the i^{th} cluster, and x_{ij} is the j^{th} instance of the i^{th} cluster (Dunham, 2002, p. 140).
As Dunham (2002) describes, the termination techniques of the algorithm may vary. It could be that the clusters of each iteration are the same or almost the same or that the algorithm terminates after a fixed number of iterations (Dunham, 2002, p. 140).

Algorithm 2.1: Kmeans

Input: $D = \{x_1, x_2, ..., x_n\}$
K initial centroids
Output: K cluster sets
1 **while** *termination condition is not reached* **do**
2 | Assign each element to the cluster of the closest centroid.
3 | Recompute the new centroids.
4 **end**

Algorithm 2.1 shows the kmeans' basic steps (Dunham, 2002; Tan et al., 2005).
The algorithm's time complexity is $O(tkn)$, where t is the number of iterations, n the number of elements, and k the number of clusters. (Dunham, 2002, p. 141). It is considered to produce good results and handles better storage and noise compared to some hierarchical agglomerative algorithms (Tan et al., 2005, p. 526). On the other side, it is not very scalable and time-efficient (Dunham, 2002, p. 141). Furthermore, outliers need to be handled exclusively in order for the clusters to be representative (Tan et al., 2005, p. 506).Finally, since kmeans bases on the euclidean measure, it works well with globular data, but not that well with other geometrical shapes nor with multi-dimensional (categorical) data (Xu & Wunsch, 2005, pp. 647, 649).

2.3 Agglomerative Clustering

As Han et al. (2011), Tan et al. (2005) explain agglomerative clustering is a hierarchical algorithm that starts by assigning each instance to its cluster. Then, on each iteration, it merges the two clusters that are the closest to each other. If no termination condition is set, on the final iteration, there is one big cluster that contains

all the initial instances. To decide which clusters to merge, a similarity measure is calculated in each iteration for its clusters and a linkage measure needs to be chosen (Han et al., 2011; Tan et al., 2005). Its results are usually presented using a binary tree on a dendrogram (Xu & Wunsch, 2005).
A simple alternative to Euclidean distance could be the Manhattan distance. Based on Shkaberina et al. (2020), it is the sum of the differences between two points coordinates. The model is as follows:

$$d_{xy} = \sum_{i=0}^{n} |x_i - y_i|$$

where d_{xy} the distance between two points x and y and i the i^{th} dimension out of n (Shkaberina et al., 2020). A distance matrix, called proximity matrix is usually created using the distance function to be given as input for the algorithm (Dunham, 2002; Tan et al., 2005).
Linkage measure is the way the similarity measure is used in order to choose the two clusters that are considered the closest (Kantardzic, 2011, p. 260). They fall into three main categories, single-linkage, average-linkage, and complete-linkage (Tan et al., 2005, p. 517). As Han et al. (2011) explain, single-linkage means that after the similarity measure for all the pairs of instances of each two clusters is calculated, the minimum distance for each pair of clusters is used to decide which two clusters are the closest. On the other side, complete linkage means that the max is chosen, and average linkage that the average distance is calculated. Therefore the different types of linkage can be described as follows:

$$Single\,linkage : dist_{min}(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} d(x^i, x^j)$$
$$Complete\,linkage : dist_{max}(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} d(x^i, x^j)$$
$$Average\,linkage : dist_{avg}(C_i, C_j) = \frac{1}{n_i nj} \sum_{x_i \in C_i, x_j \in C_j} d(x_i, x_j)$$

where dist is the function of the similarity measure, C_i, C_j are the i^{th} and j^{th} clusters and x^i, x^j are instances of the C_i^{th}, C_j^{th} clusters (Han et al., 2011, p. 460).

Algorithm 2.2: Agglomerative Clustering

Input: $D = \{x_1, x_2, ..., x_n\}$

Output: Cluster sets

```
1 Compute the proximity matrix.
2 while there is more than one clusters do
3 |   Merge the closest two clusters.
4 |   Recompute the proximity matrix.
5 end
```

Algorithm 2.2 shows agglomerative clustering's basic steps (Dunham, 2002; Tan et al., 2005).
The algorithm's time complexity is $O(n^2 lgn)$, where n is the number of proximities

assuming the matrix is symmetrical (Dunham, 2002; Tan et al., 2005). Because of that, it is not considered scalable and suitable for high-dimensional data (Kantardzic, 2011; Tan et al., 2005), though through the years new techniques have been developed to deal with this problem (Xu & Wunsch, 2005). Furthermore, a danger for low-quality results is lurking in case the splits are not wisely chosen, since they can not be undone, and instances can not move from one cluster to another (Han et al., 2011).

2.4 Applications in the touristic sector

As Dolnicar (2008) states, companies all around the globe use market segmentation to target their audiences more effectively and consume their resources more efficiently. Cluster analysis is a useful tool for the creation of those segments to create groups of people that seek similar benefits from their travel experiences. These groups not only help in the formation of marketing strategies but also to products that offer higher fulfillment to their consumers (Dolnicar, 2008, p. 17). This type of segmentation was introduced by Haley (1968) in 1968 and is called benefit segmentation. The need for the creation of a new way of grouping target groups arose from the fact that tourists' behavior is mainly determined by the benefits they seek to satisfy and not by descriptive factors (Haley, 1968, p. 31).

2.4.1 Benefit segmentation of potential wellbeing tourists

A sample application of benefit segmentation using cluster analysis describe Pesonen et al. (2011) in their article. Data about tourists were gathered in the region of Savonlinna, Finland, during the period with the highest footfall of the year through electronic questionnaires. The research aims to find the different segments of tourists based on the benefits they seek and then examine the interest of each segment on wellness holidays.
For the segment formulation, they used two algorithms, one hierarchical to find the best number of clusters, and kmeans to create the clusters. The solution they proposed contained four distinct clusters. From the four clusters, two of them seemed to have a higher preference for wellness services, as designated by Pesonen et al. (2011) 'Culturals' and 'Sightseers'. Both clusters portrayed people that seem to favor attractions and also have a tendency to go back to the same destination for their vacation. Additionally, the 'Culturals' show a preference for cultural activities, where 'Sightseers' show a preference for sightseeing activities. The illustrated interpretation of the results is that the previous approach, which promoted wellbeing products based on nature, might not be the most appropriate. The suggested alternative claims that combining wellbeing with history, culture, diverse experiences, and attractions could be more efficient (Pesonen et al., 2011, pp. 308–312).

2.4.2 A Clustering Method for Categorical Data in Tourism Market Segmentation Research

Another sample application that focuses more on the nature of the data to be analyzed describes Arimond and Elfessi (2001). Commonly, collected touristic data come from surveys, which, to a great extent, contain qualitative data. Furthermore, for the market segmentation to produce more representative results, many attributes should be used. Even so, most clustering methods used by marketers do not work well nor with multi-dimensional nor with categorical data (Arimond & Elfessi, 2001, p. 391).

Arimond and Elfessi (2001) used a Bed and Breakfast (B&B) survey in order to illustrate a more reliable way to deal with this type of data. First, they used multiple correspondence analysis (MCA) to produce some initial cluster groups, get a visualization of them, and understand them. Afterward, they performed cluster analysis, using the kmeans algorithm, in order to find the market segments. They concluded in four clusters, from which the three of them seemed to be worthy of further analysis. The other cluster did not contain enough data to lead to meaningful interpretation. The first of the three segments contained respondents who were seeking a romantic experience and were less likely to return to the B&B. Also, they were not very concerned about the cost and amenities provided. Tourists of the second cluster were looking for cozier, more peaceful, nature-related options and were more likely to go back to the same place if they liked it. They also preferred to take part in energetic activities and, similarly to the first cluster, do not care so much about price and amenities. The third cluster was technically a mixture of the two previous ones. The distinguishing factor was the eagerness of the people of the latter to socialize and the importance of price, value factors (Arimond & Elfessi, 2001, pp. 394–395).

Taking into consideration the results of the kmeans Arimond and Elfessi (2001) used the geographical characteristics of each cluster and suggested the usage of different marketing approaches on each of the three regions that the respondents came from (Wisconsin, Minnesota, Illinois). The B&B owners made some changes based on the formed clusters to improve their customer satisfaction. For example, they decided to offer a private breakfast to the first cluster, add more activities to please the second cluster, and provide information about social events to the people of the last cluster. Furthermore, they were planning on changing their marketing strategy based on the demographics provided by the research to better target their audience (Arimond & Elfessi, 2001, pp. 395–396).

Chapter 3

CASE STUDY: CLUSTER ANALYSIS ON TRAVEL AGENCY'S DATA

3.1 Data

3.1.1 Description

The data used for clustering were drawn from the back office system of a travel agency in Parga, Greece. The majority of the data are about hotel reservations. The rest are about excursions and transfers. The back office uses a relational database to store its data. After thoroughly examining the available tables, only a few of them were kept for the analysis, the ones that seemed to have the most analytical value. Their relationships are shown in the following simplified diagram.

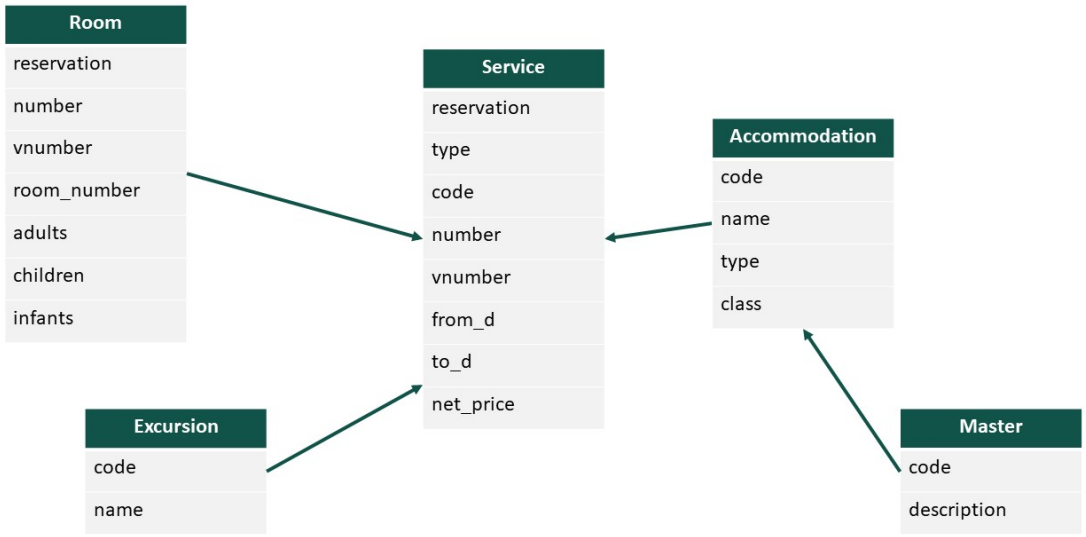


Figure 3.1.1: Relationships diagram.

One reservation may consist of many services (e.g., hotel reservation, excursion

activity). For each service booked, there is at least one corresponding row on the 'Service' table that contains its details in a total of sixty-seven columns. Eleven of them were used and are explained in table 3.1.

Column	Description
reservation	Service's reservation identifier.
type	Service type. Can be either hotel(HTL), excursion(EXC) or transfer(TRF).
code	Service code from either 'Excursion', 'Hotel', or 'Transfer' tables.
number	Service number. The service's identifier number per reservation.
vnumber	Service vnumber. A number starting from one for each service per reservation. Whenever the service is edited a new row is created with the updated details and vnumber incremented by one.
fromd	Service's starting date.
tod	Service's ending date.
net_price	The total price of the booked service.
status	Reservation status. Can be either canceled(CL) or confirmed(CF).
customer_id	Customer's identifier.

Table 3.1: Service table.

Whenever an accommodation service is booked, the 'Accommodation'(3.2) table, holds the information that describes it. Using the service code set on the 'Service' table the corresponding accommodation can be found.

Column	Description
code	Accommodation identifier.
name	Accommodation name.
type	Accommodation type identifier. Categorical value that shows the type of the accommodation.
class	Accommodation class identifier. Categorical value that shows the quality of the accommodation.

Table 3.2: Accommodation table.

The 'Master'(3.3) table contains several codes, and its descriptions, that can be found on many tables of the database, including the accommodation class and type codes.

Furthermore, whenever an accommodation is booked, at least one new row is entered in 'Room' table. These rows contain the information of the booked rooms of the accommodation, as described in table 3.4.

Similarly to an accommodation booking, when an excurision is booked its details can be found through the service code of 'Service' table in table 3.5.

Column	Description
code	A code.
description	Code's description.

Table 3.3: Master table.

Column	Description
reservation	Service's reservation identifier.
number	Service number. The service's identifier number per reservation.
vnumber	Service vnumber. A number starting from one for each service per reservation. Whenever the service is edited a new row is created with the updated details and vnumber incremented by one.
room_number	The number of rooms contained in this service booking with this specific composition.
adults	The number of adults in this room.
children	The number of children in this room.
infants	The number of infants in this room.

Table 3.4: Room table.

3.1.2 Processing and reforming

After cleaning the data and before performing the analysis, some additional preprocessing was done. This preprocessing contained the creation of new columns and tranformation of categorical data to numerical, ordinal whenever that was possible.Eventually, the tables were reformed to a flat structure. The additional data were used both for the analysis and the interpretation of the results.
First of all, the service booking dates were used in order to define the season status of the trip. The season status can either be low, medium or high, and is used to show whether, in a certain period, a small or a big number of reservations is expected. As defined from the travel agency, high season is during Christmas and Easter(in this case orthodox) holidays and from the 10th of July to the end of August. Medium season are considered September, June and from the 1st until the 9th of July. The rest of the year is considered low season. Furthermore, in order to keep each booked service only once, on its final state, only the maximum vnumber was kept for each service. Finally, for the accommodation services, to get a more representative value of the service, the price per person per night was calculated.
As shown in figure 'Relationships diagram.' on page 17, the 'Room' table contains information with regard to its composition. Instead of just using the number of adults and children a composition tag was created. The terms used by the travel agency to describe the composition are family, couple, single, crew and group. For interpration purposes these tags were added to the room data.
As already mentioned, information about the type and class can be drawn from the 'Master' table. The descriptions of the codes by themeselves provide many verbal information that can be summarized to create universal tags for the accommodations. Therefore, using those descriptions, five type tags were created, hotel, studio, apartment, house and villa, and three class tags, A(high class), B(medium class) or

Column	Description
code	Excursion identifier.
name	Excursion name.

Table 3.5: Excursion table.

C(low class). The class tags were represented by numerical values for the analysis. After transforming the initial tables to a flat structure, several plots were made with different combinations of columns and data formats using PCA(Principal Component Analysis). These plots helped to get a better idea of how the data are scattered and also test whether the data manipulation already done could result in meaningful clusters. The plot of the data that were later used for the analysis is shown in figure 3.1.2. By observing the figure, we would expect as an ideal output of the algorithms four clusters.

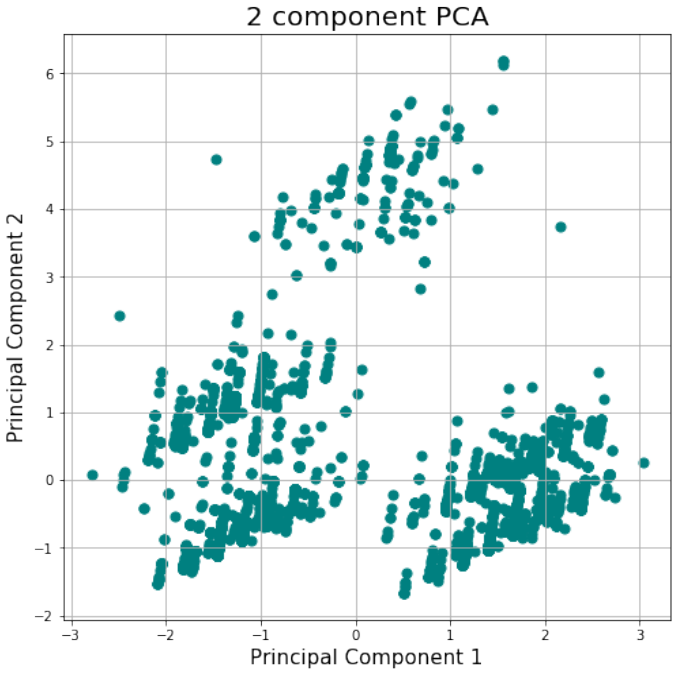
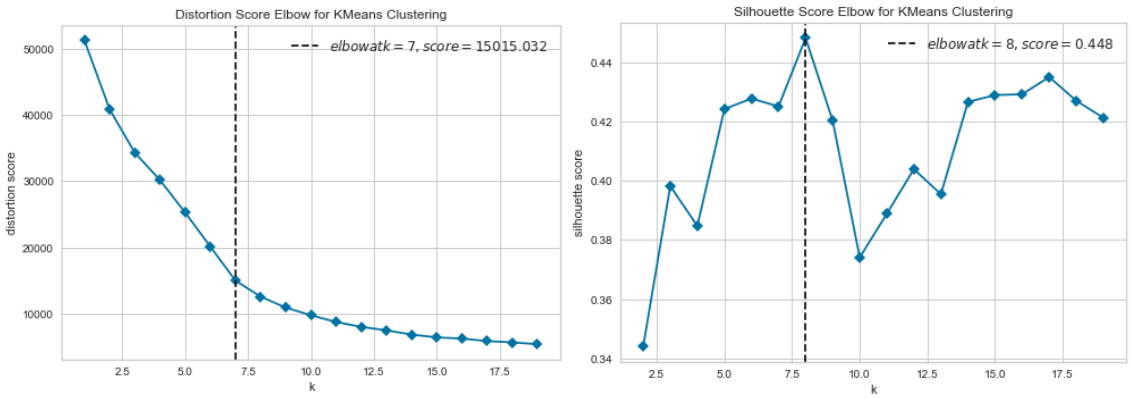


Figure 3.1.2: Data plot.

3.2 Kmeans

The first algorithm used was kmeans. The metric used to calculate the distances was Euclidean distance and the number of repetitions was 50.
Before running the algorithm, two methods were used to find the best number of clusters to be used for the value of k. The first one, the elbow method, returned that the ideal number of clusters is seven (3.2.1a). The second one used was the silhouette method. In this case, the ideal number of k was eight as shown in diagram 3.2.1b.
In order to conclude which would be the final kmeans clusters to be analyzed, kmeans



(a) Elbow method.

(b) Shilouette method.

Figure 3.2.1: Methods used to find the optimal number of clusters.

was run two times, using as k the numbers seven and eight respectively. Afterward, PCA was used on both results and seven was chosen as the best fit. From the seven produced clusters three of them were considered unworthy of further analysis because they were too small. So eventually, kmeans resulted in four clusters that can be observed in figure 3.2.2.

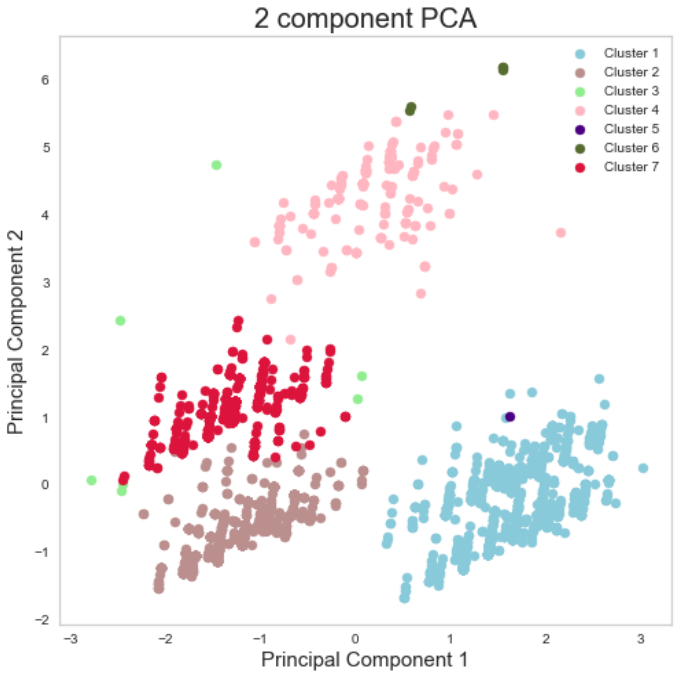
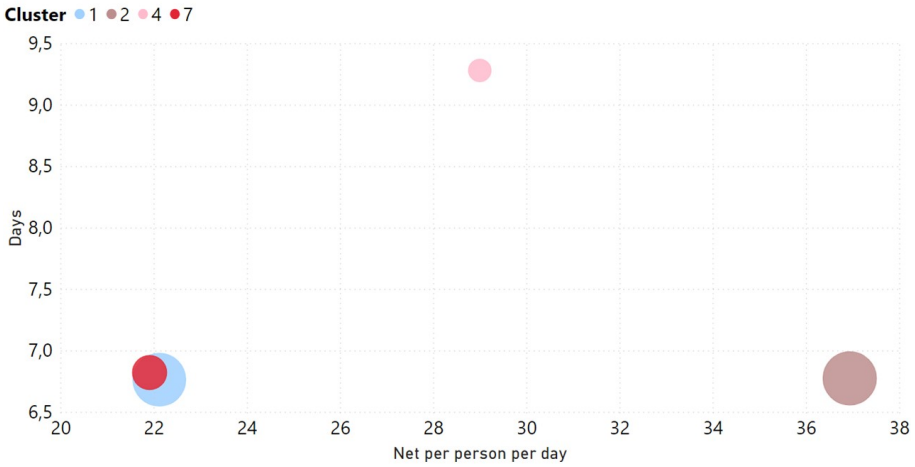


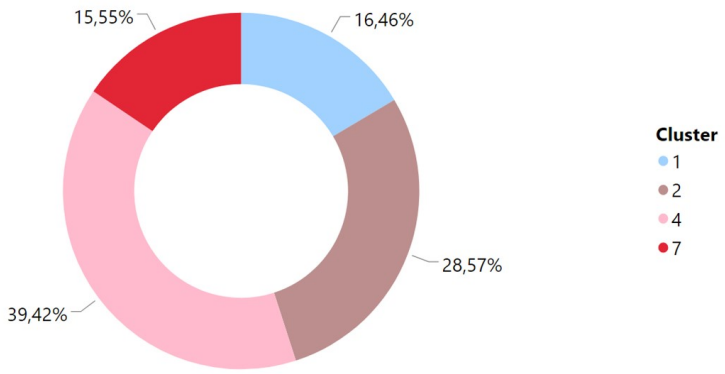
Figure 3.2.2: Kmeans algorithm.

Descriptive statistics were used to better understand the created groups and the reasons why they were formed the way they did. The formation of the clusters was affected to a great extent by the type of accommodation. The type of accommodation was the only categorical data used for the analysis that could not be converted to ordinal, therefore it is normal that the algorithm recognized those groups. In the following paragraphs, the clusters will be analyzed.

The first cluster(cluster 1 in diagrams) mainly contains low-cost, brief trips done by either families, crews or groups. These trips are more 'Value for money' oriented, since, even though they are very cheap, they contain relatively good quality accommodations (category B).
The second one(cluster 2 in diagrams) is also for brief trips. It differentiates on the category of the accommodation, the cost, and the composition. It contains accommodations of the highest quality, of the highest cost of all the clusters, and includes mainly singles and couples. This is considered a cluster that describes 'Luxury trips'.
The third cluster(cluster 4 in diagrams) differentiates mainly from the others on the trip's number of days, which appears to be very high. The reservations are mainly for studios of average quality and price. This is the 'Relaxing vacation' cluster.
Finally, the last cluster(cluster 7 in diagrams) is very similar to the first one, apart from one very important factor. The quality of the accommodation is the lowest of all the other clusters even though the price does not change that much, as it can be observed from diagram 3.2.3a. This is the 'Cheap trip' cluster.



(a) Average days and average net per person per day for each cluster.



(b) Total net for each cluster.

Figure 3.2.3: Descriptive statistics of the clusters.

As observed from diagram 3.2.3b, the most profitable for the agency clusters seem

to be clusters 'Luxury trips' and 'Relaxing vacation'. Based on those results the agency could adjust the provided product more to the customer needs. Therefore, in order to increase profits, the travel agency could include more trips of these types to their portfolios. These results can also help in the way the agency addresses its clients. It could change its marketing campaigns to target each group in a more personalized way and attract more clients. For example, when targeting a family it could present its package as a smart value for money option. Furthermore, it could concentrate more on promoting the two most profitable groups, especially cluster number 4, which seems to be one of the most profitable while counting a small number of reservations. In this way, it could effectively increase its revenue.

3.3 Agglomerative Clustering

The second method used to cluster the dataset was Agglomerative clustering. The metric used to compute the linkage between the clusters was the Manhattan distance, using average linkage. To observe the formation of clusters the dendogram 3.3.1 was created.

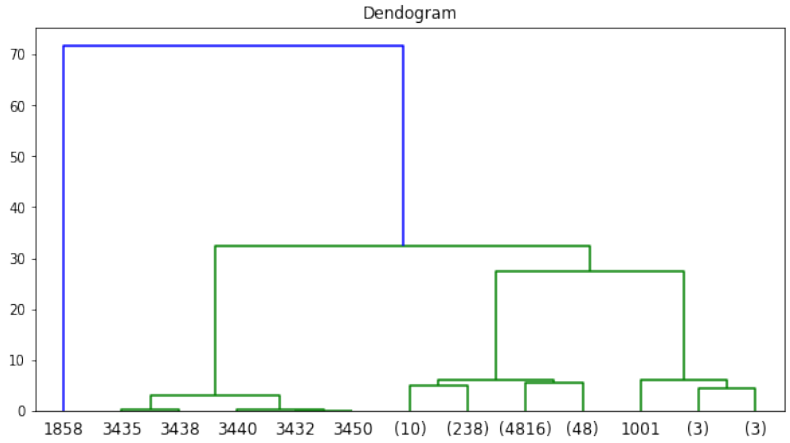


Figure 3.3.1: Dendrogram.

By observing it the expected clusters could be either four or ten. Though, PCA was also used for various values of n (between four and ten) and eventually the number of clusters was set to eight, with the result being the following.

Just as in the case of kmeans, some of the clusters formed did not provide meaningful results and were not considered worth analyzing. The ones that provide useful information are clusters one, two and four. Cluster 1 of agglomerative clustering is the 'Value for money' cluster from kmeans, and cluster 4 is the 'Relaxing Vacation' cluster. Cluster 2 seems to be a join of the other two kmeans clusters and will be named 'Low cost trip' cluster, since it contains the cheapest accommodations, of the lowest quality, and it is about trips that do not last many days (3.3.3b).
The importance of cluster 'Luxury vacation' on the total net is even more prominent now, as can be noticed from diagram 3.3.3a. On the other side, the 'Low cost trip' cluster, even though it contains the most reservations, has a very low impact on

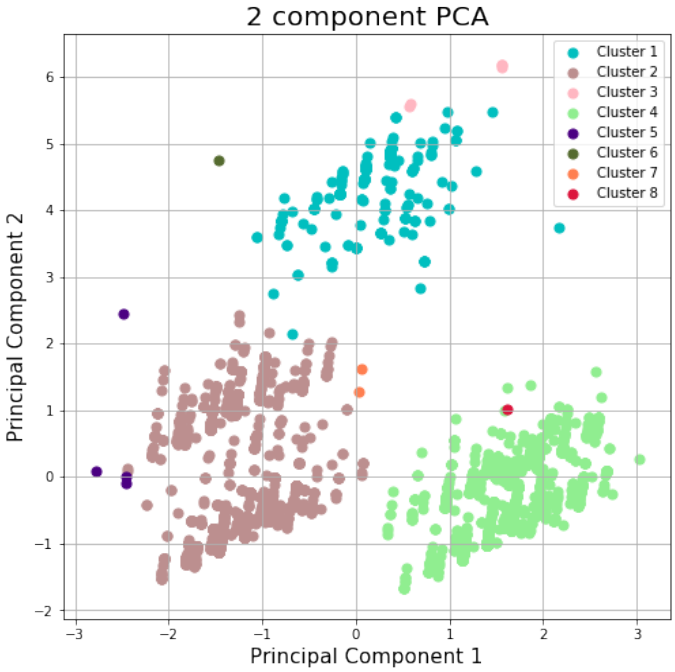
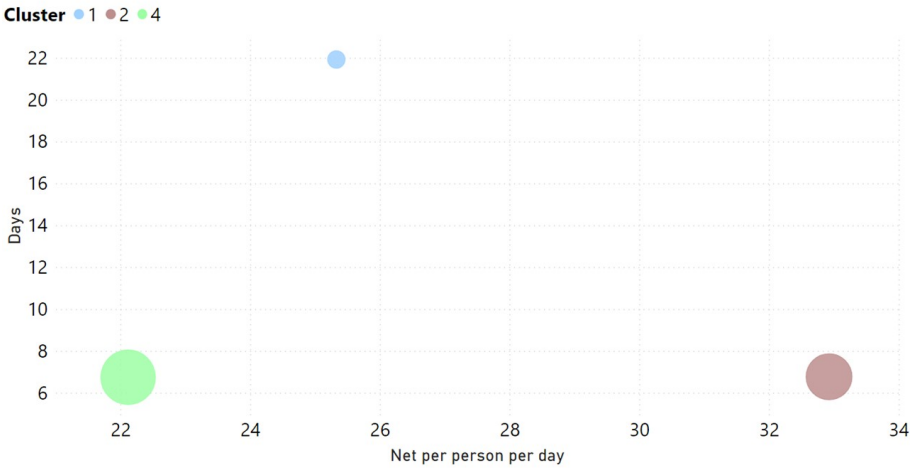
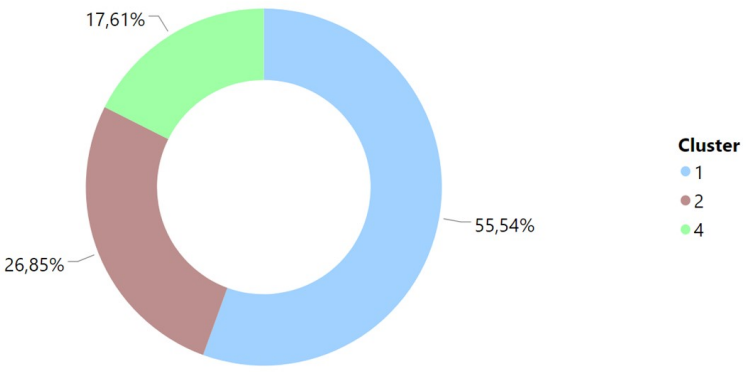


Figure 3.3.2: Agglomerative clustering.

the total revenue of the agency. Therefore, as proposed in the previous section the agency should concetrate more on the providance of the services of the clusters 'Luxury trip' and 'Relaxing vacation' and their promotion.



(a) Average days and average net per person per day for each cluster.



(b) Total net for each cluster.

Figure 3.3.3: Descriptive statistics of the clusters.

Chapter 4

CONCLUSIONS

4.1 Comparing Results

The task of dealing with categorical data is demanding and requires good data manipulation to produce satisfactory results. Through the usage of the two algorithms, we could say that both of them worked pretty well with them and provided useful information. Especially kmeans, even though it is not considered to work well with this type of data, produced the best results out of two, managing to recognize all the four groups.
Indeed the results could have been better if the categorical data did not affect them that much. Still, the travel agency can take advantage of the outcome to adjust both its product and marketing strategy in favor of the clusters 'Luxury trip' and 'Relaxing vacation' to become more competitive in the touristic market and increase its revenue.

4.2 Restrictions

Even though the dataset contained a sufficient volume of data, these were not evenly distributed throughout the different kinds of information. The data about the accommodations were severely outnumbering the excursion data. This resulted in the latter not playing any role in the cluster formation. The main reason why the excursion data were insufficient was due to the misuse of the back office system. Initially, a large amount of excursion data were available, but not properly inserted, at the time of the reservation, in the system and therefore was not possible to connect them to their corresponding accommodation reservation.
Furthermore, as mentioned, the clusters were to a large extent formed by the type of accommodation. If the type of accommodation was to be translated as the benefits that a client seeks on, for example, renting a studio instead of an apartment the results could offer a better picture of the factors that affect a traveler's decisions. This type of information was not currently available.

4.3 Future Work

The analysis did result in useful insights for the agency. Additional analysis though could be done if the current dataset was enriched by survey data given to the travel-

ers. These data could be about the benefits sought from their trips, and the factors that lead them to choose the services they chose. In this way, some categorical data could be converted to ordinal and the analysis would produce better results.
Furthermore, the same way this analysis was mainly focused on the accommodations since they could not be combined with all the equivalent excursions, a second analysis could be conducted based on the excursions. This analysis could produce excursions bundles to be promoted together. Finally, the agency's portfolio of excursions could be expanded based on the most profitable ones.

Bibliography

Arimond, G., & Elfessi, A. (2001). A clustering method for categorical data in tourism market segmentation research. *Journal of Travel Research*, *39*(4), https://doi.org/10.1177/004728750103900405, 391–397. https://doi.org/10.1177/004728750103900405

Basu, S., Banerjee, A., & Mooney, R. (2002). Semi-supervised clustering by seeding, In *In proceedings of 19th international conference on machine learning (icml-2002*.

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns [PMID: 10582567]. *Journal of Computational Biology*, *6*(3-4), 281–297. https://doi.org/10.1089/106652799318274

Bishop, C. M. (2006). *Pattern recognition and machine learning*.

Celeux, G., & Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*(3), 315–332. https://EconPapers.repec.org/RePEc:eee:csdana:v:14:y:1992:i:3:p:315-332

Chakrapani, C. (2015). *Machine learning: An algorithmic perspective*.

Cuzzocrea, A., Loia, V., & Tommasetti, A. (2017). Big-data-driven innovation for enterprises: Innovative big value paradigms for next-generation digital ecosystems, In *Wims '17*.

Dolnicar, S. (2008). A review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, *12*(1), 1–22. https://doi.org/10.1300/J073v12n01_01

Dunham, M. H. (2002). *Data mining: Introductory and advanced topics*.

D'Urso, P., Disegna, M., & Massari, R. (2019). Fuzzy clustering in travel and tourism analytics. https://doi.org/10.1007/978-3-030-06222-4_22

D'Urso, P., Disegna, M., Massari, R., & Osti, L. (2016). Fuzzy segmentation of postmodern tourists. *Tourism Management*, *55*, 297–308. https://doi.org/https://doi.org/10.1016/j.tourman.2016.03.018

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*.

Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. *Journal of Marketing*, *32*(3), https://doi.org/10.1177/002224296803200306, 30–35. https://doi.org/10.1177/002224296803200306

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms*.

Marsland, S. (2004). *Statistics in market research*.

Paykel, E. S., & Rassaby, E. (1978). Classification of suicide attempters by cluster analysis. *British Journal of Psychiatry*, *133*(1), 45–52. https://doi.org/10.1192/bjp.133.1.45

Pesonen, J., Laukkanen, T., & Komppula, R. (2011). Benefit segmentation of potential wellbeing tourists. *Journal of Vacation Marketing*, *17*(4), 303–314. https://doi.org/10.1177/1356766711423322

Schwab, K. (2017). *The fourth industrial revolution*. USA, Crown Publishing Group.

Shkaberina, G. S., Rozhnov, I. P., Popov, V. P., Kazakovtsev, L. A., & Lapunova, E. V. (2020). Detection of homogeneous production batches of semiconductor devices by greedy heuristic clustering algorithms with special distance metrics. *IOP Conference Series: Materials Science and Engineering*, *734*, 012104. https://doi.org/10.1088/1757-899x/734/1/012104

So, K. K. F., & Li, X. ((2020). Customer engagement in hospitality and tourism services. *Journal of Hospitality & Tourism Research*, *44*(2), 171–177. https://doi.org/10.1177/1096348019900010

Sutton, R. S., & Barto, A. G. (2015). *Reinforcement learning: An introduction*.

Tan, P.-N., Kumar, V., & Steinbach, M. (2005). *Introduction to data mining*.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, *16*, 645–678. https://doi.org/10.1109/TNN.2005.845141