ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Hotel Booking

Clustering Analysis

Statistical Machine Learning

Student Giorgia Treglia

Identification Number 6042123

Academic Year 2024/2025

Contents

Introduction

Nowadays, hotel bookings involve a large range of customer behaviours, from short weekend to business trips, and from early planners to last-minute reservations. Knowing the details of the characteristics behind each booking, especially for hotel managers, is very important, particularly when it comes to predicting whether a reservation might be cancelled or not. The importance behind this analysis is about the fact that cancellation can cause operational inefficiencies and lost revenue, and being able to anticipate them could improve booking organisation, pricing strategies, overbooking policies, and resource planning.

In this project the aim is to analyse a dataset of 2,000 hotel booking observations, with 17 variables representing information about the number of guests, the duration, the lead time, the type of room and meal, any special requests, and whether the reservation was eventually cancelled. The goal of this analysis is to identify significant and well-separated clusters based on booking characteristics using unsupervised machine learning techniques, which are part of a branch of machine learning that works with unlabelled data, using algorithms to find hidden patterns without any human intervention. In other words, the training model has only input parameter values and it has to discover the groups or patterns on its own. In this analysis, the booking cancellation status variable was excluded because it is a labelled outcome that could influence the clustering process by saying to the model what the correct answer should be. To perform the clustering, the methods used were K-Means Clustering, Model-Based Clustering (Mclust), and Hierarchical Clustering, which are going to be explained in the dedicated chapter. By comparing the results from each method, it was possible to evaluate the consistency and significance of the clusters, allowing to analysed also the correlation between cluster membership and cancellation probability. Additionally, the quality of each cluster was evaluated using internal validation techniques (silhouette scores), the Adjusted Rand Index and the NbClust tool, used to evaluate and select the most appropriate number of clusters based on objective statistical criteria. In conclusion, the analysis shows whether the clustering methods can provide useful signals for anticipating booking cancellations and show the characteristics that differentiate customer profiles.

Chapter 1

Descriptive and Exploratory Data Analysis

Before performing any clustering method, it is essential to understand the structure and the characteristics of the dataset through Descriptive and Exploratory Data Analysis (EDA). This chapter helps to summarize the main characteristics of the data, identify anomalies, and to give a general interpretation of each variable.

1.1 Dataset Overview and Summary Statistics

The first thing to do was to clean the dataset by removing the variables that are not appropriate for clustering. Specifically, the Booking ID was excluded, as it is simply an identifier which has no meaningful information about the booking itself, the date of reservation was also removed because, even if it represents the timing of the booking, it would require complex preprocessing to be meaningfully used, and since the focus of the study is on booking characteristics, the booking status (which indicates whether the booking was cancelled) was excluded. As said before, clustering is an unsupervised learning task, so the exclusion of the target variable is necessary to avoid inaccurate results. The exploratory analysis was conducted, and *Table 1* shows the summary statistics, like minimum, median, mean, and maximum values for each remaining variable.

VARIABLE	MIN	MEDIAN	MEAN	MAX
N. of adults	0	2	1.853	3
N. of children	0	0	0.109	3
N. of weekend nights	0	1	0.803	6
N. of weeknights	0	2	2.256	14
Car parking space	0	0	0.029	1
Lead time	0	56	83.3	418
Repeated guest	0	0	0.024	1
Prev. Cancellations	0	0	0.012	3
Prev. Not Cancellation	0	0	0.177	41
Average price	0	97.75	102.74	349.63
Special requests	0	0	0.63	4

The minimum of a variable is its smallest value observed, while the maximum is the largest. The mean, or also known as arithmetic average, gives an idea of the overall central tendency of every variable (the sum of all values divided by how many values there are for each). The median is the middle value that separates the lower half from the upper half of the data. As shown in the table, most bookings involve two adults, with very few including children (in fact the average number of children is close to zero) suggesting that most customers are couples or individual travellers. Regarding the duration, most people book for one or two weekend nights, though some stay for up to two weeks, while for lead time, which represents how far in advance the reservation was made, on average, bookings were made around 83 days before arrival, with some made more than a year before (up to 418 days). For the average price, while most bookings are centred around 100 euros, some go up to 349 euros, possibly for differences in room types, duration, or other characteristics. Other variables, like whether the guest requested a car parking space, special requests, or if they are a repeating customer, tend to be zero for most bookings, so the repeated guests and those with specific preferences are a minority in the data. The summary helped to understand the differences in the booking behaviour, especially the large range for lead time and average price, implies that there could be distinct types of customers, and clustering may help.

1.2 Correlation Analysis

To better understand how the characteristics of each booking are related to one another was helpful to calculate the correlation matrix using Pearson's correlation coefficient. The graph below, *Figure 1*, shows how strongly each variable is connected to the others, using values that range from -1 to 1. A value close to 1 means that as one variable increases, the other also tends to increase, a value near -1 means that as one goes up, the other goes down, and instead, values near 0 suggest little or no relationship between the variables. As shown above, most relationships between variables in the dataset are weak, meaning that booking characteristics, like the number of guests, number of nights, or price, tend to vary independently. For instance, the number of children had almost no connection with the number of weekend nights, or the number of adults showed a moderate positive correlation (0.26) with the room type, which is reasonable, as larger groups may book larger rooms. Additionally, a relatively low correlation (0.35) was found between the average price and the number of children, suggesting that booking with children should be typically more expensive, possibly because of the need of larger rooms. As previous mentioned some variables tend to vary independently, for instance, there is only a weak correlation (0.20) between the number of weekend

nights and the number of week nights, indicating that short and long stays are balanced across the multiple reservations. A few variables instead showed moderately stronger relationships, for instance, the room type was fairly correlated with the average price (0.43), which makes sense since better rooms tend to cost more. The number of previous cancellations is also somehow related to how many times a guest had previously cancelled or rebooked (0.44), which reflects their past behaviour, it has also a very strong correlation with the number of previous non-cancellations (0.71), which is expected due to the logic link. It is interesting to note that lead time and average price, two variables that may have been thought to be correlated, only exhibit a weak negative relationship, indicating that early bookings do not always translate into higher or lower prices. Overall, this analysis demonstrated that there is no significant redundancy in the data, and that each variable contributes to different information. Since each one may add to identify different types of booking, it is important to include all of them in the clustering step.

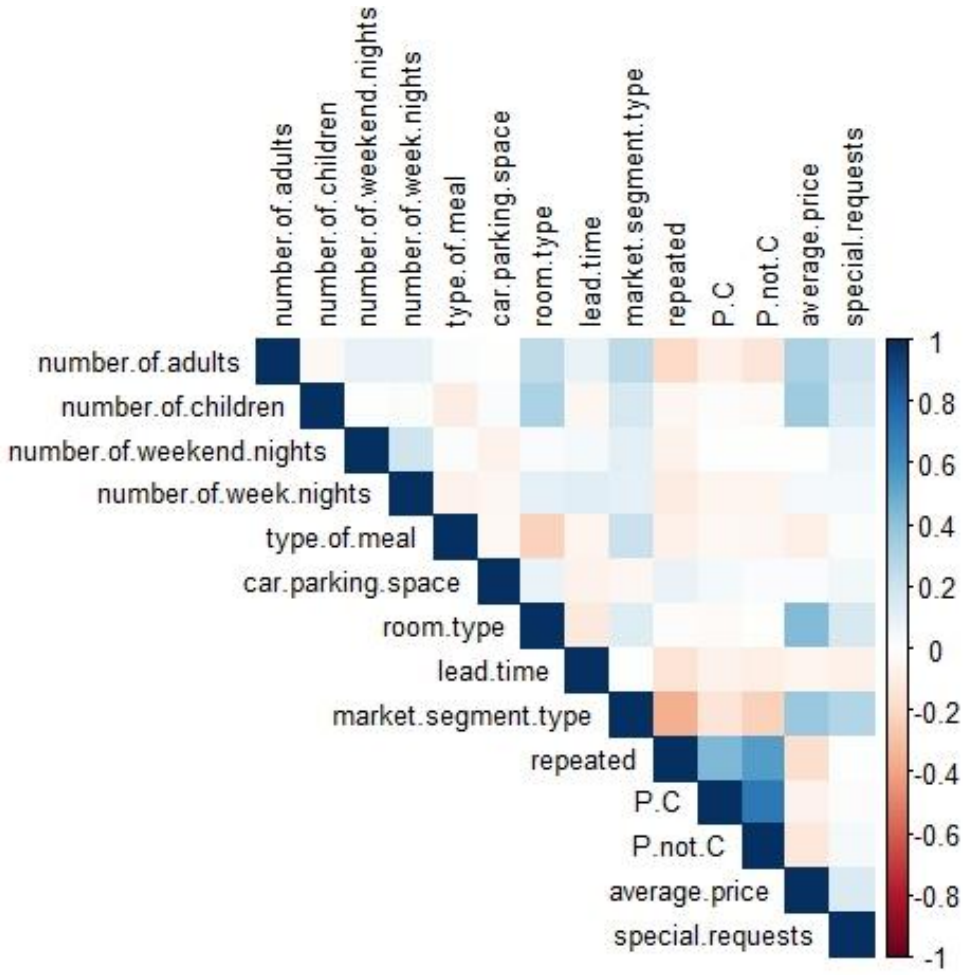
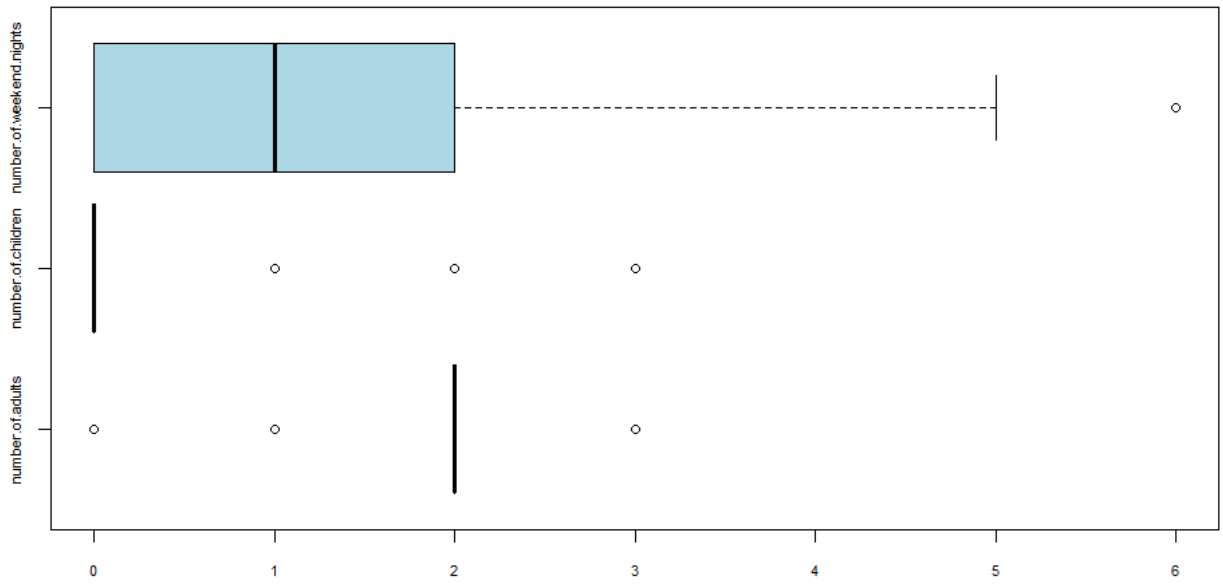
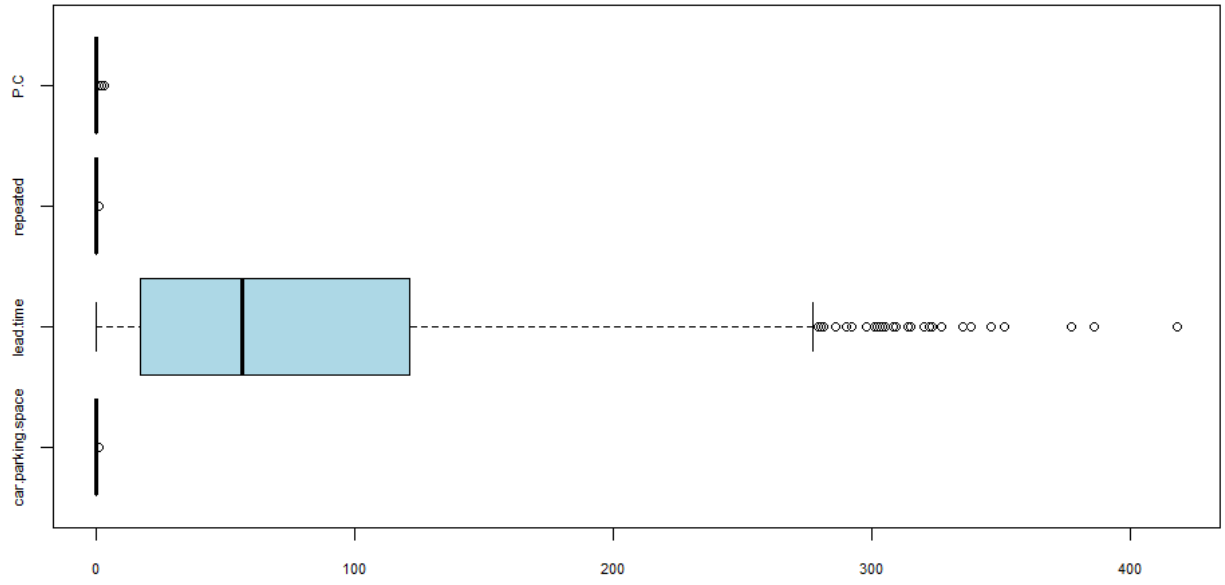


Figure 1 – Correlation Matrix

1.3 Boxplots

In order to understand how the values of different variables, like the number of nights a guest stays, the number of adults in a reservation, or how far in advance a booking is made, are distributed, is important to conduct the boxplot analysis, a plot that also help to understand the booking behaviour by identifying the overall distribution of the data and making it easier to identify unusual values, known as outliers (data points that may be very high or very low compared to the rest). The majority of the variables in the figure below (*Figure 2*) such as the number of adults, the number of children, nights, and car parking spaces, have compact distributions with very few extreme values. However, some other variables showed a lot more variation. In particular, lead time (which means how far in advance the booking was made), average price per night, and the number of previous bookings that were not cancelled show many outliers. These outliers represent bookings made super early, bookings that cost much more than average, or bookings from customers who had stayed at the hotel many times before. They might seem like mistakes or unusual data points, but that's not always the case. In this specific situation, these outliers likely reflect certain booking behaviours. As seen in the graph, some people plan their holidays extremely early, while others spend more money on fancier or longer stays. Although outliers often indicate possible data errors, it was better to keep them in the dataset because they might represent important information. For example, customers who pay more or book very early might behave differently than others when it comes to cancelling their reservations. Keeping these outliers gives us a full and accurate picture of booking behavior, which is important since the goal of this project is to identify patterns related to cancellations.





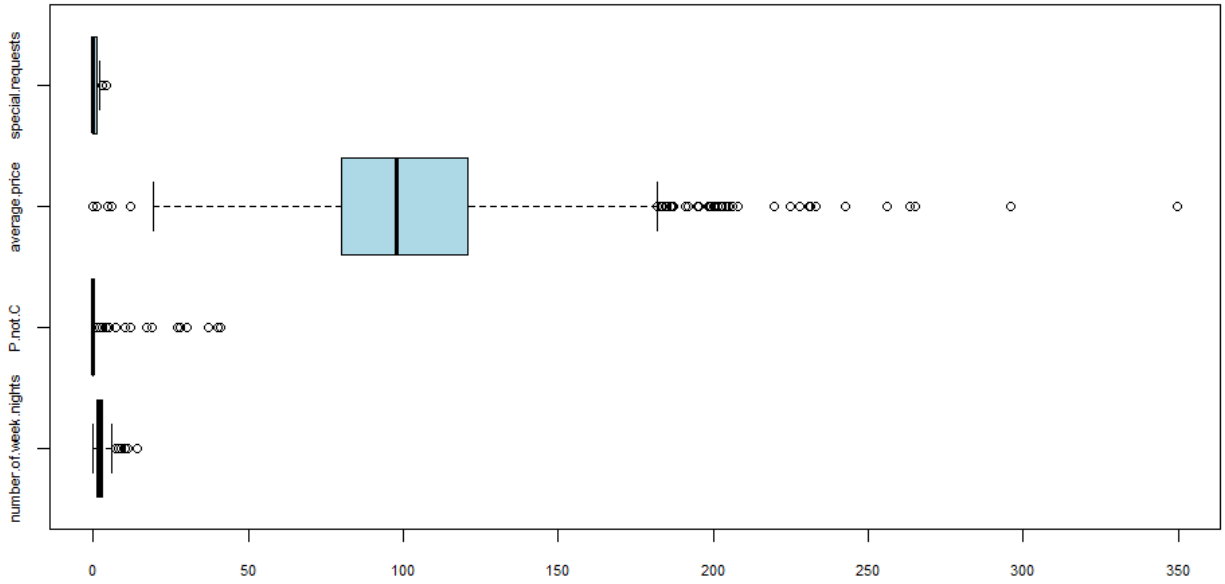


Figure 2 - Boxplots

Chapter 2

Clustering Methods

Grouping data points based on their similarity with each other is called Cluster Analysis, a method which aims to find natural groups within the data without any previous information. The fact of knowing if the reservations could be categorized in some way, like those that are more or less likely to be cancelled, is helpful. In this analysis the focus is on K-Means Clustering, which divides the data into a fixed number of groups by trying to make the observations in each group as similar as possible to each other. Hierarchical Clustering, which builds a tree structure of clusters starting by treating each data point as its own cluster and then gradually combine the closest clusters step by step. Lastly, Model-Based Clustering (MClust), which assumes that the data is generated from a mixture of probability distributions and uses statistical models to identify clusters.

2.1 K-Means Clustering

K-means clustering is a common method that groups similar observations based on their similarity where the first step is to decide how many groups is better to divide the data into. A useful method is the Elbow Method, a technique used to find the ideal number of clusters by looking at how the total variation within clusters (WSS) decreases as the number of clusters increase.

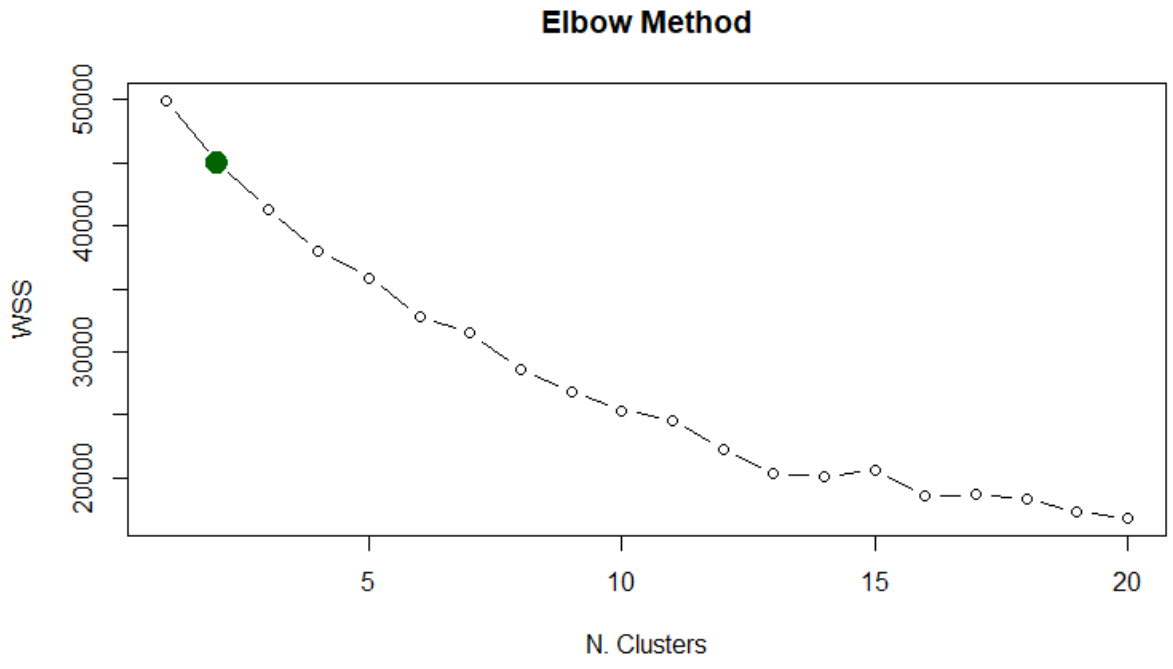


Figure 3- Elbow Method

The idea is to look for a point on the graph where the WSS stops decreasing as quickly, that point that looks like an elbow. In our case, the plot (shown below) shows the elbow at 2 clusters, the green dot, suggesting that dividing the data into 2 clusters is a reasonable choice, as adding more clusters brings only marginal improvement. After determining that 2 clusters were appropriate, the K-means algorithm, using the standardized version of the dataset (so that variables on different scales wouldn't dominate the clustering), divided the 2000 bookings into:

CLUSTER	SIZE
1	1,269
2	731

Each cluster has a center that represents the average value for each variable within that group. By examining these centers, is possible to understand the general characteristics of each cluster. For example, the Cluster 1 has lower average lead time (0.11) and higher average price (0.25), with more special requests (0.28). It also contains more people who did not cancel their booking. While the Cluster 2 tends to have higher lead times (meaning people book further in advance), lower average prices, and fewer special requests. It also has a slightly higher proportion of cancelled bookings. To better understand how these clusters relate to booking behavior, especially cancellation, we compared the clusters to the original booking.status variable:

CLUSTER	NOT CANCELLED	CANCELLED	PROPORTION N.C.
1	784	485	61.8%
2	538	193	73.6%

Interestingly, Cluster 2, which tends to have longer lead times and lower prices, had a higher proportion of not cancelled bookings (73.6%) compared to Cluster 1 (61.8%). This result indicates that certain booking characteristics, like earlier planning and lower prices, might be linked to a lower chance of cancellation, even if the booking seems simpler (fewer requests, lower costs). Clustering helped reveal these patterns by organizing the data in a way that highlights natural differences between types of customers.

2.2 Model-Based Clustering (Mclust)

After applying K-Means clustering, it seems useful to use a more flexible method called Model-Based Clustering, implemented with the Mclust package in R. Unlike K-Means, which partitions the data based on distance and requires the number of clusters to be chosen manually, Mclust automatically selects both the best number of clusters. It does this by testing various Gaussian mixture models and choosing the one that maximizes the Bayesian Information Criterion (BIC), a standard metric for balancing model complexity and fit. This method explored between 2 and 10 clusters, and it selected a solution with 2 clusters, the same results from the Elbow and Silhouette methods used in K-Means. The selected model type was VEI, which means it assumes clusters that have equal shape but different sizes, with a BIC value of 856,555, which indicates a strong fit to the data, and a log-likelihood of 428,570, suggesting the model captures meaningful structure.

CLUSTER	SIZE
1	1,554
2	446

To evaluate how well these clusters reflect actual booking outcomes, was good to compared them to the cancellation status of each booking. In Cluster 1, 552 out of 1,554 bookings were cancelled, with a cancellation rate of 35.5%, while Cluster 2 had 126 cancellations out of 446 bookings, corresponding to a lower rate of 28.3%. This difference suggests that Cluster 1 may group more uncertain bookings compared to Cluster 2. While the clusters are not completely distinct in terms of cancellation probability, there is difference, and this reinforces the idea that bookings can be grouped into types with different tendencies.

CLUSTER	NOT CANCELLED	CANCELLED	PROPORTION N.C.
1	1002	552	64.5%
2	320	126	71.7%

Instead, the plot below helped to visualize the structure in two dimensions where each point represents a booking, with different colours and shapes indicating which cluster it belongs to. Cluster 2 is more spreads out, especially along the first principal component (Dim1), possibly reflecting more variability in the booking characteristics it contains.

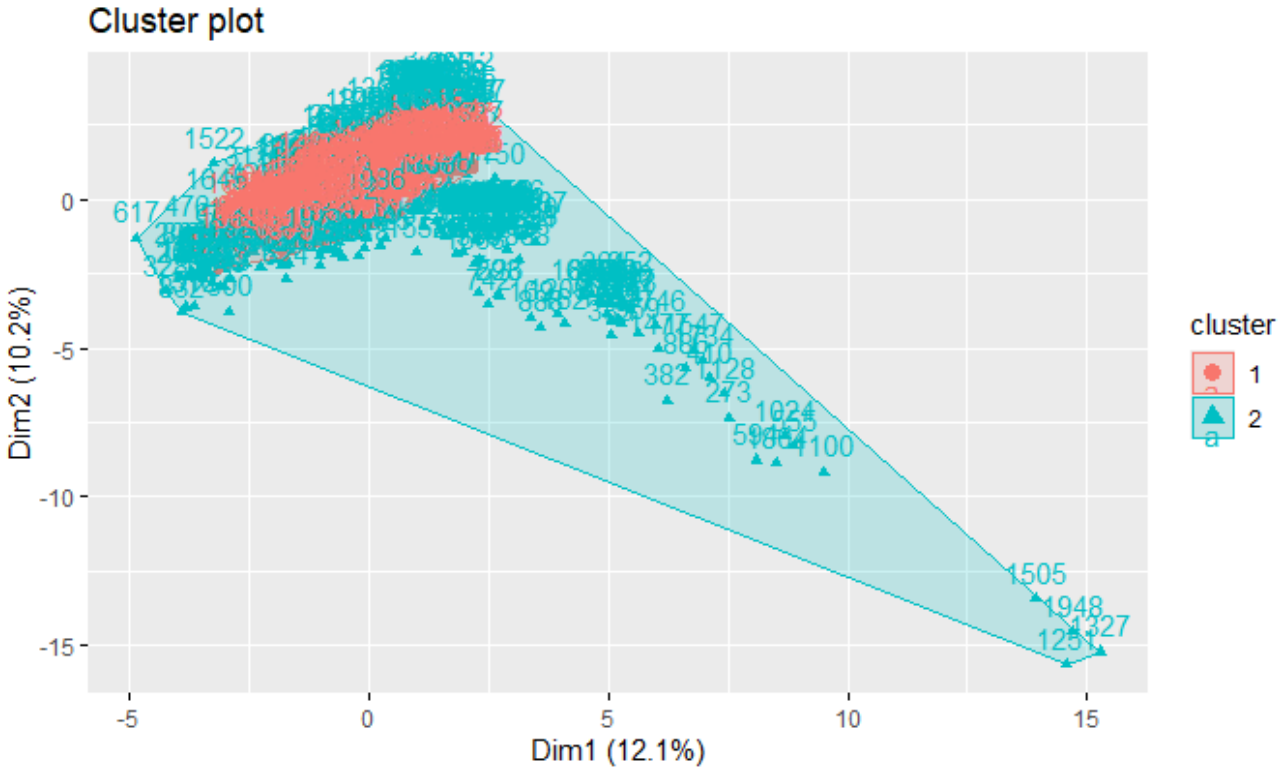


Figure 4 – Mclust Classification

2.3 Hierarchical Clustering Analysis (HCLUST)

To further investigate the structure of the data, a hierarchical clustering analysis was conducted using four different linkage methods: complete, single, average, and Ward. The single linkage groups clusters based on the shortest distance between any two points in different clusters, while the complete linkage uses the farthest one, the average linkage merges clusters by calculating the average distance between all points in the two clusters and the Ward linkage tries to minimize the total variation within clusters by merging the pair of clusters that have the smallest increase. Hierarchical clustering is a method that builds a tree structure (basically a dendrogram) to show how individual data or groups merge step by step, based on their similarity. The distance between the points is calculated using Euclidean distance after standardising the dataset. First, this method allows to display the height plots, *Figure 5*, which help identify how far apart clusters are at each merging step and where a significant jump in height occurs, which typically indicates the optimal number of clusters. All four methods show an increase near the end of the height axis, meaning a two-cluster solution, like the results from both the k-means and model-based clustering (Mclust).

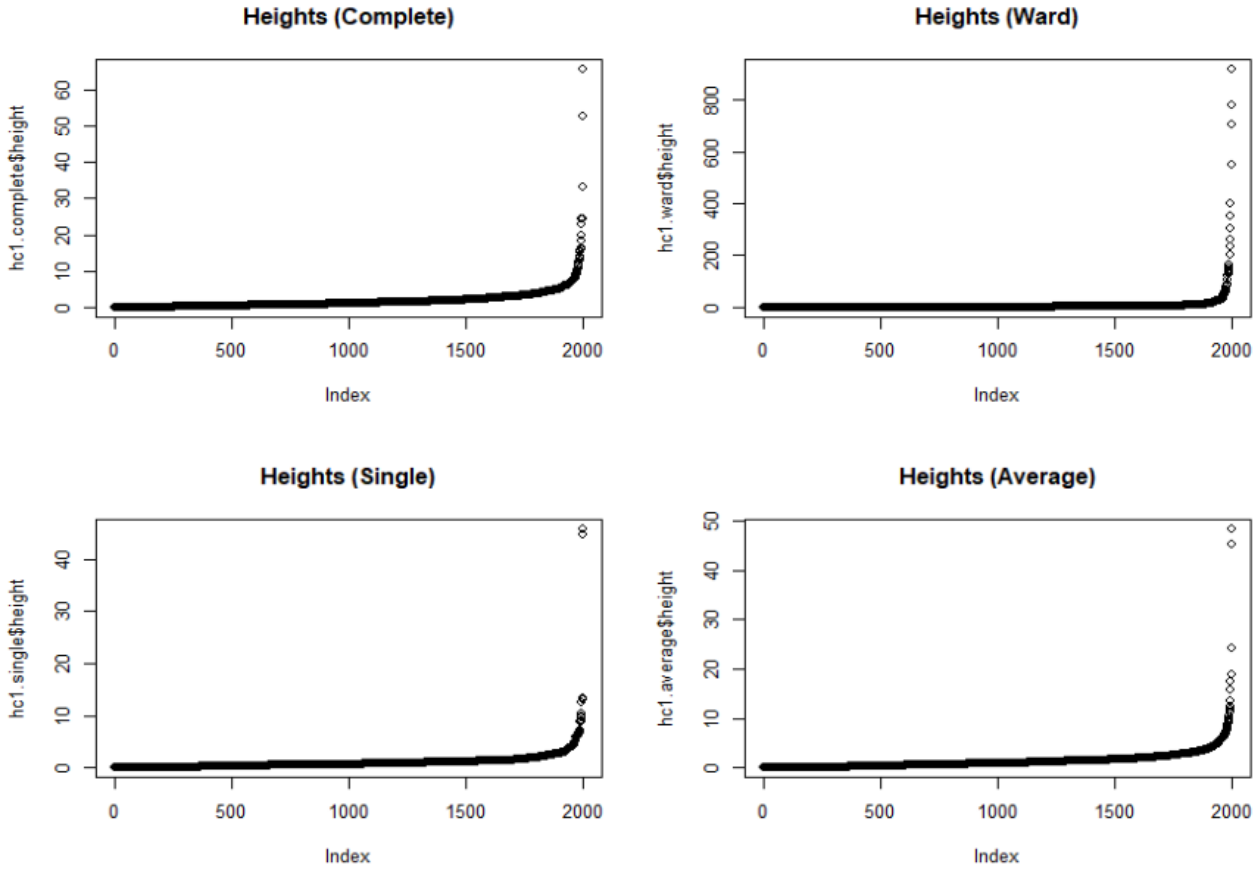
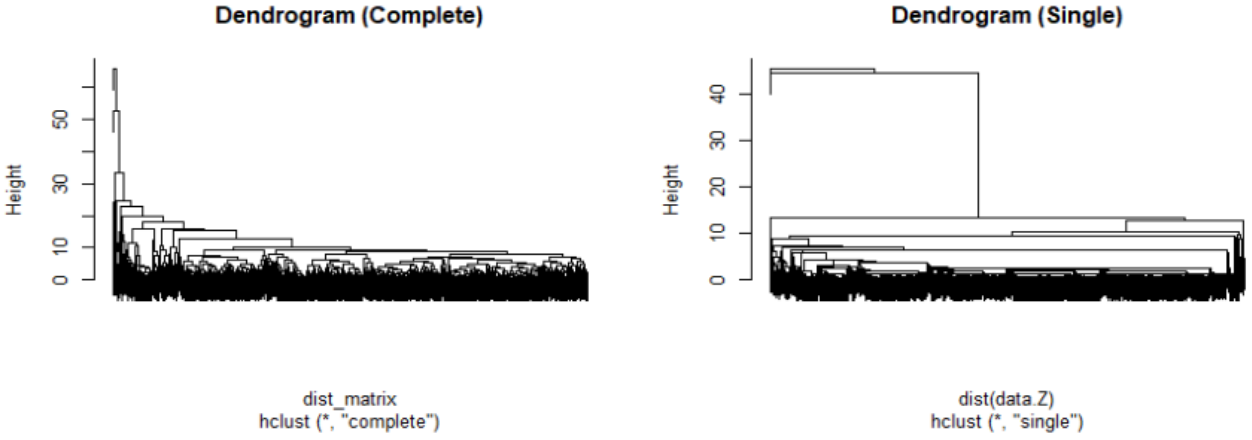


Figure 5 – Height Plot

After the height plot, dendrograms were created for all methods, but Ward's method provided the clearest and most balanced clustering structure. As show in *Figure 6*, by visually inspecting the dendrogram and drawing a horizontal line at the appropriate height (using rect.hclust), the data was cut into two main clusters. The first with 1,401 observations and the second of 599.



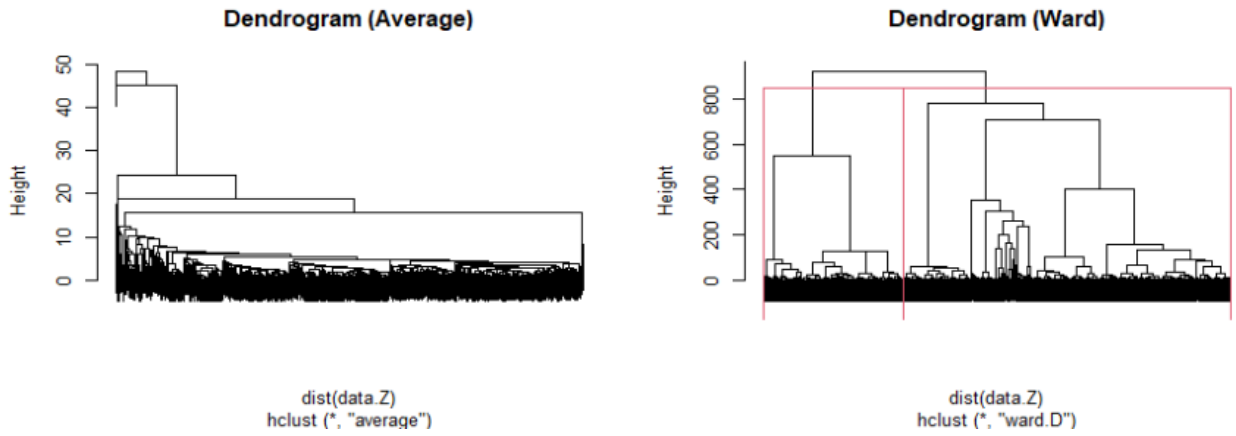


Figure 6 – Dendrogram

After the plots, the analysis was conducted. In cluster 1, about 65.7% of the bookings were not made, while 34.3% were made. Similarly, cluster 2 showed 67.1% not booked and 32.9% booked. These proportions suggest that while the clusters differ in size, both groups have a similar internal distribution of booking statuses, slightly favouring the no booking class. Overall, the hierarchical clustering, especially with Ward's method, confirmed previous findings and further validated the division of the dataset into two distinct clusters with reasonably interpretable patterns in terms of the booking behavior.

CLUSTER	NOT CANCELLED	CANCELLED	PROPORTION N.C.
1	920	481	65.6%
2	402	197	67.1%

Chapter 3

Comparison of Clustering Methods

After applying K-means, model-based clustering (Mclust), and hierarchical clustering (HCLUST), was useful to compare the results to understand how each method grouped the data and whether these groupings align. Although all three methods suggested two clusters as the optimal number (supported by the silhouette and elbow methods for K-means, the Bayesian Information Criterion for Mclust, and the dendrogram heights for HCLUST), the composition of clusters is different, as show in *Figure 7*, regarding the size of the different clusters.

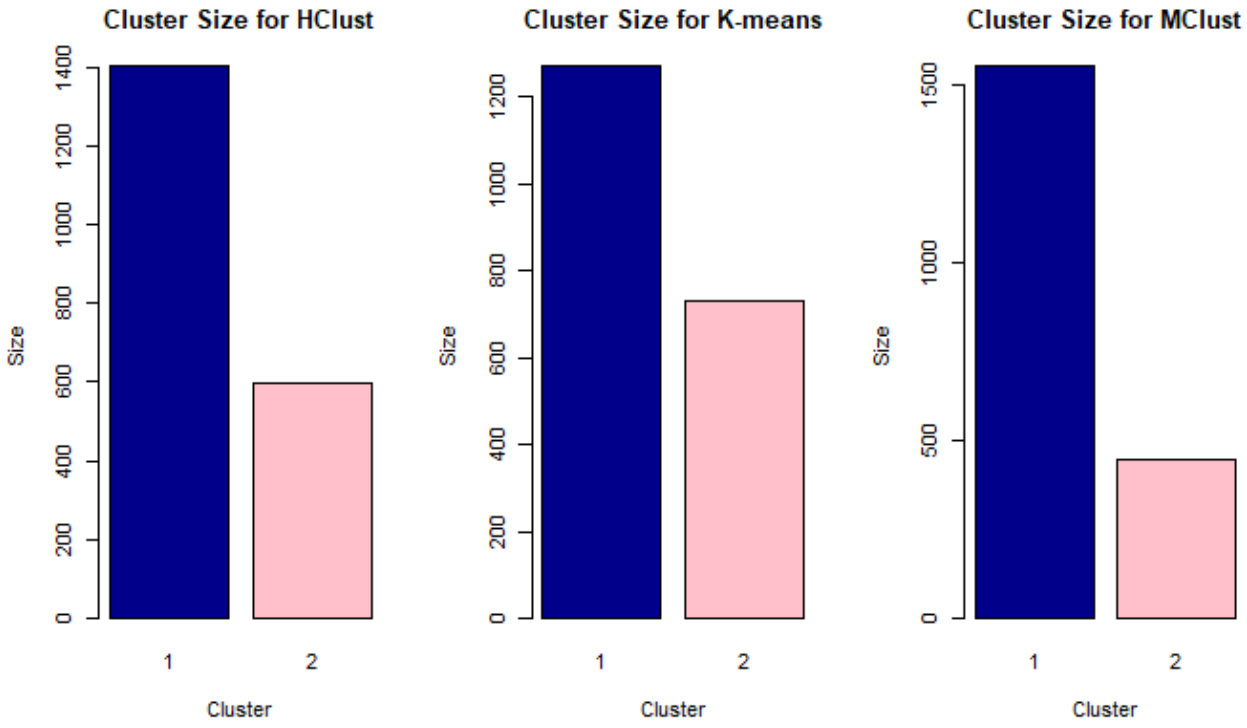


Figure 7 – Cluster Size

3.1 Cluster Characteristics and Evaluation

To see the similarity between clustering results, was important to calculate the Adjusted Rand Index (ARI), a metric that measures the agreement between two clustering solutions. An ARI of 1 indicates perfect agreement, while 0 means the clustering is no better than random. The highest similarity was observed between K-means and HClust (Ward), with an ARI of 0.62, suggesting

moderate agreement in how they classified the data. On the other hand, K-means and Mclust had a much lower ARI of 0.14, and Mclust and HClust showed minimal agreement with an ARI of 0.04. To better understand the differences between the clusters identified by K-means, Mclust, and Hierarchical Clustering, was important to analyse the average values of key customer variables within each cluster, specifically lead time, repeated guest (whether the customer had stayed before), average price, and number of special requests. This allows to interpret not only how the clustering algorithms split the data, but also what kind of customer behaviours they captured.

CLUSTER	LEAD TIME	REPEATED	A. PRICE	SPECIAL REQUESTS
1	73.92	0.0008	111.92	0.86
2	99.58	0.0643	86.79	0.23

CLUSTER	LEAD TIME	REPEATED	. PRICE	SPECIAL REQUESTS
1	85.05	0.0000	101.73	0.66
2	77.19	0.1076	106.23	0.52

CLUSTER	LEAD TIME	REPEATED	A. PRICE	SPECIAL REQUESTS
1	66.61	0.0343	106.19	0.79
2	122.33	0.0000	94.65	0.25

In the case of K-means, cluster 1 (the larger group) is made up of bookings with a shorter average lead time (around 74 days), a very low rate of repeated guests, and a higher average price and number of special requests. Cluster 2, on the other hand, included customers who booked earlier (around 100 days in advance), had a little higher rate of repeated stays, and paid lower prices with fewer special requests. For Mclust, the distinction was somewhat different. Cluster 1 showed an average lead time of about 85 days and no repeated guests at all, with a moderate price and request pattern. Cluster 2 had a shorter lead time (77 days), more frequent repeat guests, and little higher average prices and requests. HClust provided a clearer contrast: cluster 1 grouped customers who booked closer to their stay (67 days), had a higher rate of repeat bookings, and requested more services. Cluster 2, by contrast, included customers who booked much earlier (over 122 days on average), had no repeated stays, and fewer special requests. The last method seems to have combined the timing and customer loyalty behaviour, creating two defined profiles.

METHOD	AVERAGE SILHOUETTE
K-MEANS	0.184
MCLUST	0.313
HCLUST	0.105

To assess the quality of the clustering itself is appropriate the silhouette coefficient, a common metric that measures how well-separated the clusters are. Mclust achieved the highest average silhouette score (0.31), suggesting it produced the most compact and well-separated clusters. Additionally, to validate the choice of the number of clusters, the NbClust method was applied using different clustering algorithms: complete linkage, Ward's method, and K-means. NbClust uses a wide range of statistical indices to suggest the most appropriate number of clusters for the data, but because the dataset contained several categorical variables that complicated the analysis, these variables (room type, market segment type, and meal type) were excluded to allow the algorithms to focus on numerical characteristics. Across all three methods, most indices supported 2 clusters.

Conclusion

This clustering analysis aimed to identify meaningful cluster division based on behavioural and booking characteristics, using three different methods and NbClust as a validation tool. All these methods consistently pointed to two clusters as the optimal number. The elbow and silhouette methods supported two clusters in K-means; Mclust selected a two-component model using the BIC, the dendrogram structure of Hierarchical Clustering clearly revealed two main groups and most internal indices used by NbClust also concluded that two was the better choice. Despite the number of clusters, the way in which the observations were assigned differed. The silhouette score was 0.31 in Mclust, 0.18 in K-means, and 0.10 in Hierarchical Clustering, suggesting that Mclust has the clearest separation about the group of customers. About the results a consistent behavioural pattern emerged across all methods. One group included customers with shorter lead times, higher average prices, more special requests, and fewer cancellations. The other group had customers who booked earlier, paid lower prices, made fewer special requests, and had a higher likelihood of cancelling. However, not all methods grouped the same individuals in the same way. The Adjusted Rand Index for Mclust and the other two methods was 0.14 with K-means and 0.04 with HClust, indicating that Mclust captured a more flexible way to group. So, considering all the results, it should be the most effective clustering technique for this dataset, producing the clearest separation.