Athens University of Economics and Business

Master of Science (MSc) in Business Analytics

Statistics for Business Analytics II - FT

-Project II -

Instructors: Dimitris Karlis

Student Name: Eleni Ralli

Student ID: f2822312

30 March, 2024

Contents

1. Introduction

This study focuses on analyzing cancellation behaviors for room bookings in a hotel, aiming to explain the factors influencing the decision to cancel rather than predicting it. We utilize a random sample of bookings that includes a variety of variables such as the number of adults and children, room type, price, special requests, and others. This approach offers a deeper understanding of the dynamics shaping cancellation decisions.

For this study, we utilize a dataset comprising 2000 observations of individual room bookings in some hotel, encompassing 17 diverse variables (including nominal, ordinal, continuous, and discrete types).

Also, we create 3 more variables that will be useful for our study the number of total quests (the sum of total number of adults and total number of children), the number of total nights (the sum of total number of weekend nights and total number of week nights) and the reservation month.

Table 1: Data Table

Variable Number	Name	Type of Variable	Description
1.	Booking_ID	Nominal	Unique identifier for each booking
2.	Number of Adults	Discrete	Number of adults included in the booking
3.	Number of children	Discrete	Number of children included in the booking
4.	number of weekend nights	Discrete	Number of weekend nights included in the booking
5.	number of week nights	Discrete	Number of week nights included in the booking
6.	type of meal	Nominal	Type of meal included in the booking
7.	car parking space	Nominal	Indicates whether a car parking space was requested or included in the booking
8.	room type	Nominal	Type of room booked
9.	lead time	Discrete	Number of days between the booking date and the arrival date
10.	market segment type	Nominal	Type of market segment associated with the booking
11.	repeated	Nominal	Indicates whether the booking is a repeat booking
12.	P-C	Discrete	Number of previous bookings that were canceled by the customer prior to the current booking
13.	P-not-C	Discrete	Number of previous bookings not canceled by the customer prior to the current booking
14.	average price	Continuous	Average price associated with the booking
15.	special requests	Discrete	Number of special requests made by the guest
16.	date of reservation	Nominal	Date of the reservation
17.	booking status	Nominal	Status of the booking (canceled or not canceled)
18.	number of total quests	Discrete	The sum of total number of adults and total number of children
19.	number of total nights	Discrete	The sum of total number of weekend nights and total number of week nights
20.	reservation month	Nominal	The month of the reservation

2. Descriptive analysis and exploratory data analysis

The analysis will be performed using the R statistical package. The dataset initially contained 'Not Available' (NA) values in two observations. Due to this, these observations would provide incomplete information for our analysis and could cause issues in conducting the analysis using R package. So, they were removed.

We are examining certain variables that we find interesting on an individual basis to understand the values they hold and to perform some descriptive measures. These measures help us in better understanding each variable. For the quantitative variables like number of total quests, number of total nights, lead time, number of week nights, average price, and special requests, we are looking at their mean, standard deviation, median, minimum and maximum value, skewness, and kurtosis (see Table 2).

From this numeric variable that we examine no one is close to a normal distribution. Number of total quests, number of total nights, lead time, number of week nights, and special requests are discrete variables so this is one more reason to doesn't be close to a normal distribution. All these numeric variables that we choose to examine show a tendency to have more values on the left side of the distribution (right skew- in a perfect normal distribution, skewness is 0). Also, number of total nights and number of week nights have higher kurtosis values in contrast to a normal distribution (in a perfect normal distribution kurtosis is 3) and variables number of total quests, lead time average price, and special requests have lower kurtosis in contrast to a normal distribution (fewer extreme outliers than a normal distribution) (see Figure 1).

We will examine these hypotheses by conducting normality tests and creating the corresponding QQplots (see Appendix - Figure 2). The tests reveal that none of the numeric variables follows a normal distribution (Shapiro-Wilk and Kolmogorov-Smirnov p-value < 2.2e-16), a finding that is also graphically confirmed by observing both the probability density diagrams (see Figure 1) and the QQplots (see Appendix - Figure 2).

Numeric Variables	Mean	Median	Sd	Minimum	Maximum	Skewness	Kurtosis
Number of total quests	2	2	1	1	5	0.6	4.2
Number of total nights	3	3	2	0	20	2.2	14.4
lead time	83	56	83	0	418	1.3	4.2
number of week nights	2	2	1	0	14	1.5	9.7
average price	103	98	35	0	350	0.6	6.0
special requests	1	0	1	0	4	1.1	3.5

Table 2: Descriptive Statistics (rounded) Table for some Quantitative Variables

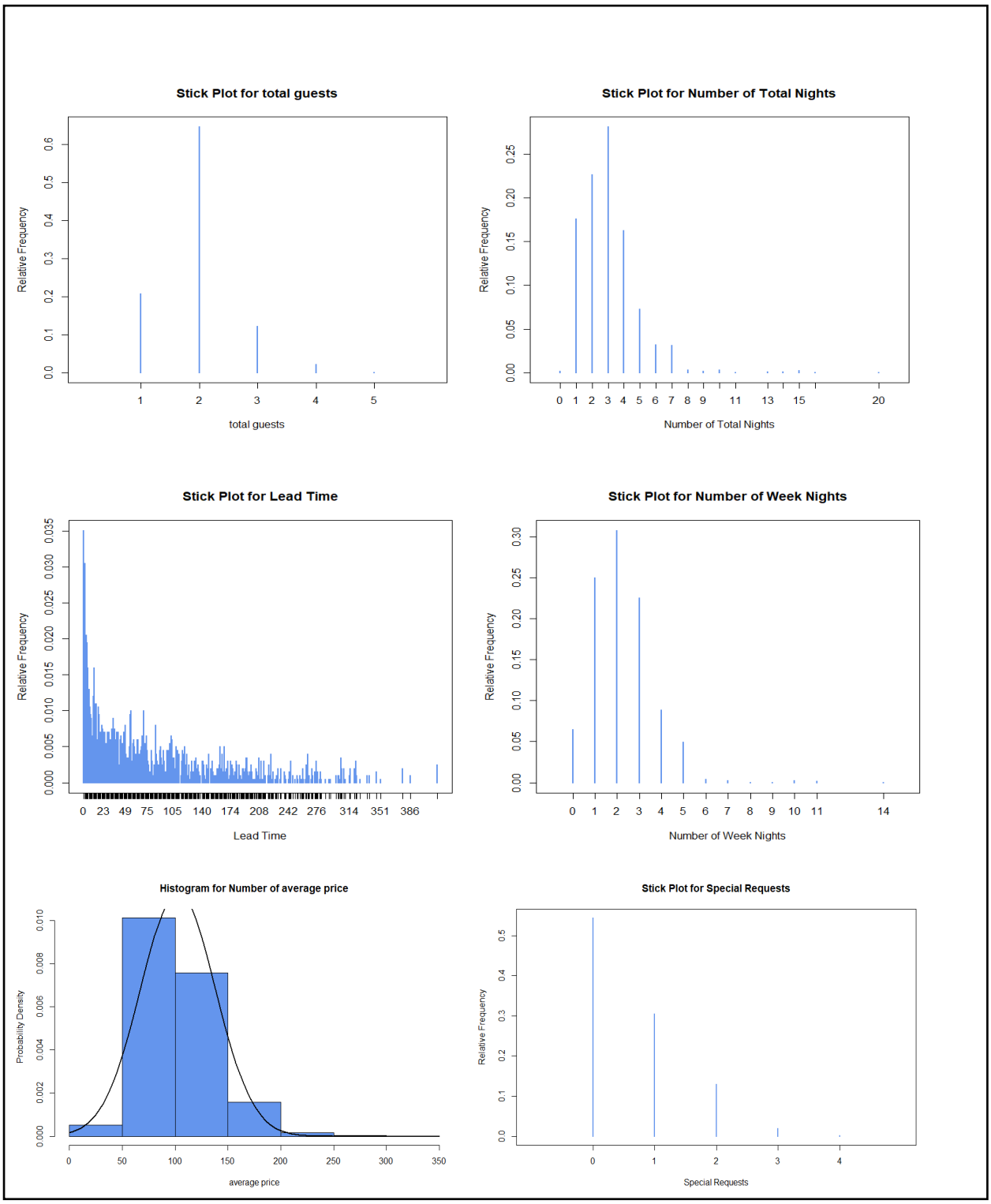

Figure 1: Probability Density Diagrams/ Relative Frequency Diagrams of some Quantitative Variables

We will examine 3 categorical variables from the dataset. The variables market segment type, reservation month and booking status and we see the frequency and percentage distribution of them. These variables show a distribution skewed towards higher levels (see Figure 3 and Table 3).

Categorical Variable	Level: Aviation	Level: Complementary	Level: Corporate	Level: Offline	Level: Online
market segment type	7 0.4%	23 1.2%	110 5.5%	590 29.5%	1270 63.4%

Categorical Variable	Level 01	Level 02	Level 03	Level 04	Level 05	Level 06	Level 07	Level 08	Level 09	Level 10	Level 11	Level 12
reservation month	101 5.1%	137 6.8%	139 7.0%	167 8.4%	161 8.0%	187 9.4%	146 7.3%	183 9.2%	245 12.2%	230 11.5%	133 6.6%	171 8.5%

Categorical Variable	Level: Canceled	Level: Not_Canceled
Booking status	678 33.9%	1322 66.1%

Table 3: Frequency table and percentages(rounded) of the levels of the categorical variable market segment type and reservation month

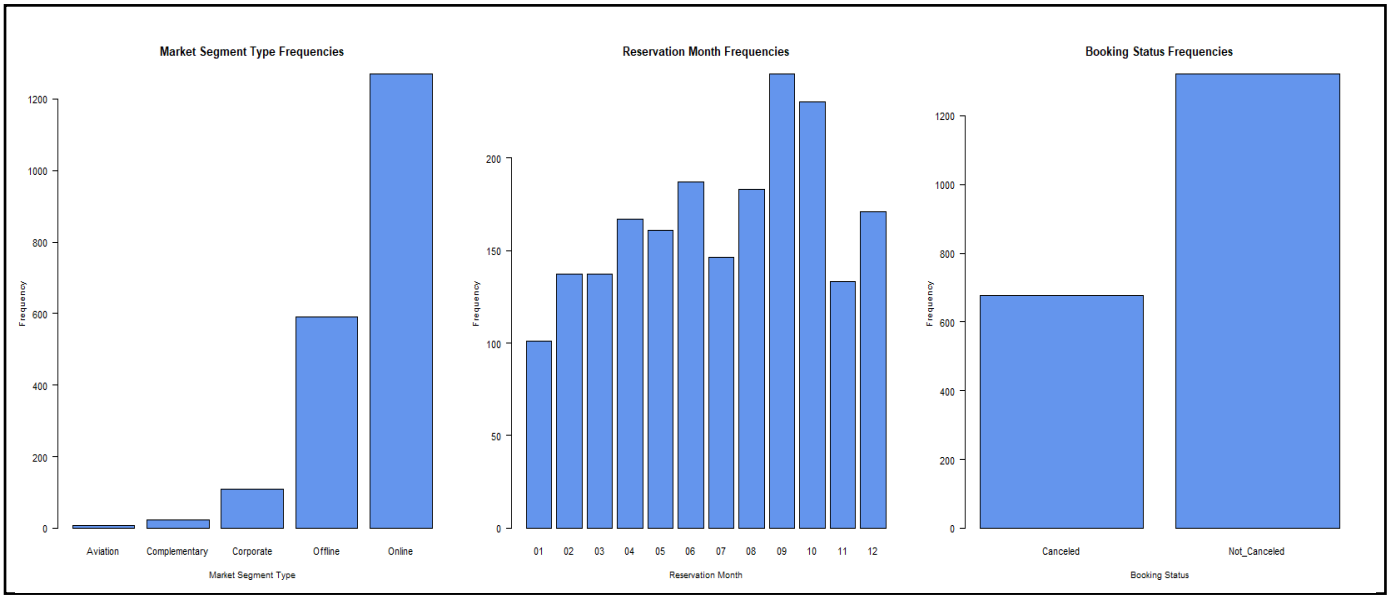

Figure 3: Bar charts of the categorical variable market segment type and reservation month

3. Pairwise comparisons

An important thing is to examine the relationship of the variable (booking status) for which we want to draw conclusions about how it is affected by the other variables in the dataset. Initially, we will examine the relationship of some numeric variables in pairs to get a first picture of how they relate to each other and how each relates to the booking status.

Pairwise comparison between numeric variables

To explore the relationship between certain numerical variables of interest, we apply the Spearman correlation matrix. This reveals that some of the numerical variables we are examining do not have a monotonic relationship with each other (Pearson linear correlation coefficient lower than 0.2 in each pair - see Figure 4). Also, as we have seen earlier, the distributions of these variables do not follow a normal distribution, so we check whether they have a monotonic relationship using a non-parametric test. For the variables average price and special requests (Spearman correlation test p-value < 2.483e-15 < 0.05), the number of weeknights with lead time (Spearman correlation test p-value < 2.2e-16 < 0.05), and the number of weeknights with special requests (Spearman correlation test p-value < 0.04597 < 0.05), there is an indication of a monotonic relationship between them. For average price with number of weeknights (Spearman correlation test p-value = 0.1175 > 0.05) or lead time (Spearman correlation test p-value = 0.3095 > 0.05), as well as lead time with special requests (Spearman correlation test p-value = 0.07097 > 0.05), there is no indication of a monotonic relationship between them.

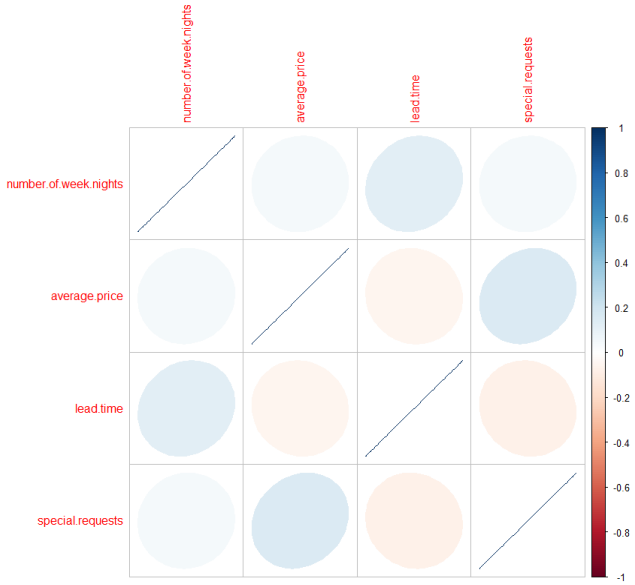

Figure 4: Spearman Correlation Matrix of some Hotel Booking Variables: Number of Week Nights, Average Price, Lead Time, and Special Requests

Pairwise comparison between categorical and numeric variables

We will examine the relationship between the categorical variable booking status and some numerical variables that appear to be of interest. For the relationship of booking status with average price, the hypothesis of normality is rejected (S-W p-value and K-S p-value < 2.2e-16 < 0.05), and the hypothesis of equality of means among groups is also rejected (Welch's t-test p-value < 2.671e-14< 0.05) These differences indicate that the average price has a significant impact on the booking status. (see Appendix -Figure 5). For the relationship of booking status with lead time, the hypothesis of normality is rejected (S-W p-value and K-S p-value < 2.2e-16 < 0.05), and the hypothesis of equality of medians among groups is also rejected (Wilcoxon test p-value < 2.2e-16 < 0.05) These differences indicate that the lead time has a significant impact on the booking status. (see Appendix -Figure 6). For the relationship of booking status with special requests, the hypothesis of normality is rejected (S-W p-value and K-S p-value < 2.2e-16 < 0.05), and the hypothesis of equality of medians among groups is also rejected (Wilcoxon test p-value < 2.2e-16 < 0.05) These differences indicate that the special requests have a significant impact on the booking status.

Pairwise comparison between categorical variables

We will examine the relationship between the categorical variable booking status and some categorical variables that seem interesting to us. For the relationship of booking status with market segment type I have indications that there is a dependency between the market segment type and booking status (Pearson's Chi-squared simulate p-value = 0.0005< 0.05) (see Appendix -Figure 7). For the relationship of booking status with car parking space I have indications that there is dependency between booking status and car parking space (Pearson's Chi-squared p-value = 1.039e-05<0.05) (see Appendix -Figure 8).

4. Predictive model

4.1 Description of selected forecasting methods and Reasoning for choice

Our aim is at creating a predictive model to classify whether a booking will be cancelled or not. predicting We did three predictive modeling techniques- machine learning methods:

- Random Forest
- Support Vector Machine (SVM)
- Logistic Regression.

Random Forest is a good choice for predicting if a booking will be canceled because it's an ensemble method- this means that it builds multiple decision trees (each one giving its own guess) and merges them together to get more accurate and stable predictions. Also, doesn't have so the risk of overfitting that a single decision tree might have.
Also, when the data is complicated (for example non-linear relationships between variables that we have in our case) can still perform good.
It uses parts of the data to train and other parts to test, which helps it not to memorize the data but to really understand it so we don't overfitting problem (the problem that a model learn the data too much and start to learn the random noise and the random variances). This can do it because, each decision tree in the forest is constructed using a different bootstrap sample from the data, with some of the original data left out of the bootstrap sample and used as an "out-of-bag" sample to estimate error, strength, and correlation (internal validation mechanism for the model).
When it builds each tree, Random Forest uses different parts of the data each time (a sample m from all the variables), making each tree unique. This makes the whole model strong and not too focused on just one part of the data. It doesn't need to pruning the trees to stop them from over-growing and overfitting.
The model also shows which things matter most when predicting cancellations. This helps to understand what makes people cancel their bookings.

Support Vector Machine (SVM) is a powerful method that can do and classification and regression tasks. It works well when we have lots of features in our data.
SVM finds the best line or a plane or a curve that separates the categories of the booking status canceled or not. This line (or a plane or a curve) is created in a way that keeps it as far as possible from the closest points of each category (giving both sides equal space). It is an optimization process, which seeks to maximize the margin between the nearest data points of any class (support vectors), and the decision boundary.
Also, it allows for some points to be on the wrong side of the line with a level of confidence in the classification, ensuring that the model is not only accurate but also generalizable (predict unseen data). This method do not perform automatic feature selection so we will do it with a penalized method (lasso is a type of regression that adds a penalty to the model for having too many variables).

Logistic Regression predicts whether something will happen or not, like if a booking will be canceled. This method looks at the relationship between our data features and whether or not a booking canceled. We can say that it adds up all

the reasons why a booking might be canceled, giving more weight to the stronger reasons, and then calculates the odds of cancellation. The transparency of this method is important when we need to explain the results to others, like to understand why losing bookings. The bad thing of this method is that assumes each variable affects the cancellation independently of any other variable. Also, if the relationship between the variables and cancellations is very complex, might not capture this as well as some other methods.

4.2 General description of our process - Training and parameter tuning process for each model and Optimization methods used

First, we split our data into two sets: a training set and a testing set. We use 80% of our data for training and hold 20% for testing. This split is for evaluating the model's performance later (to see if we have overfitting problem as we mentioned before). Also, we ensure that the proportion of booking statuses (canceled or not canceled) is maintained across both training and testing datasets (to help maintaining the dataset original distribution).

For the Random Forest: It use a 10-fold cross-validation. So, the training data is divided into ten parts, training the model ten times, each time using a different part as the test set to measure model performance. We do this to optimize the 'mtry' parameter, which represents the number of variables randomly sampled as candidates at each split. A grid of potential 'mtry' values was created, and the model was trained across these values to find the optimal. The maximum limit is set to 10. This means that for each tree in the forest, a different number of variables will be tested to find the best splitting point at each node. We choose to select the number 10 to increase the diversity among trees (lower mtry values mean that each tree in the forest is likely use different subsets of variables for their decisions, this is good for accuracy because reducing the correlation between individual trees- generally we want the mtry to be smaller that the total number of variables).

The Lasso method was used for variable selection for the SVM in Logistic Regression. This method uses a penalty term to shrink some of the coefficients to zero, effectively selecting a simpler model that does not include those variables. The optimal value of the penalty term was determined using cross-validation. The function that we use returns the value of lambda that minimizes the cross-validation error, which is then used to select the most important variables. After the variable selection process, a logistic regression model was built using the selected variables and the Akaike Information Criterion (AIC) is used to further variable selection for the model (stepwise model selection). This function starts with the initial model and tries to improve it by adding or dropping variables, one at time. So, we will keep the explanatory variables: number of weekend nights, car parking space, lead time, market segment type, average price, special requests.

For the SVM: The explanatory variable was selected and we input them to the model. We use a 10-fold cross-validation, repeated three times, to train the model. So, the training data is divided into ten parts, we use nine for training, and keep one for test (each subset was one time test). We repeat this process three times with different random splits so to reduce the variability in the model's performance estimation (more reliable). To optimizing the model, we tune two

hyperparameters: sigma and C. The sigma parameter affects how flexible the model is when drawing the plane (or decision boundary) that separates different classes of data. Controls how the model balances between fitting the training data closely and being able to generalize well to new, unseen data. A lower sigma value means the model tries to closely follow the training data, which might capture more details but can lead to overfitting. A higher sigma value makes the model less strict in following the training data, making it more capable of generalizing to new data. The C parameter is like deciding how much we let some observations of the wrong group be on the wrong side of the plane without care too much. With high C, the model trying really hard to get every observation on the correct side, even if it means that our plane has to zigzag a lot (very big variance). This is great for the observations we already have, but if we add new observations, we will see the overfitting problem. With low C, we are more okay with a few wrong observations on the wrong side and we have a smoother plane (work better for unseen observations). So, we want the optimal combination of sigma and C values and for this we make a grid search across three different positions for each hyperparameter from this two. For sigma, we try 0.001, 0.01, and 0.1. For C, we try 0.1, 1, and 10. So, we will test all their combinations to see which setting makes our model perform best. After training our model with these settings, we choose the combination of sigma and C that gave the best results according to accuracy (all these steps were done by the library caret).

Logistic Regression: We use a 10-fold cross-validation, repeated three times, to train the model. So, the training data is divided into ten parts, we use nine for training, and keep one for test (each subset was one time test). We repeat this process three times with different random splits so to reduce the variability in the model's performance estimation (more reliable). Here we don't have hyperparameters to tune but we use the binomial family (we have a binary outcome) and generalized linear model.

4.3 Comparison and selection of the best model (Evaluation metrics)

For the Random Forest Model Evaluation:

Confusion Matrix	Reference		
Prediction		Canceled	Not Canceled
	Canceled	92	15
	Not Canceled	43	249

The confusion matrix shows that out of the total predictions, 92 were correctly predicted as "Canceled" (true positives) and 249 as Not Canceled (true negatives). There were 43 false negatives, meaning that these bookings were incorrectly predicted as Not Canceled when they were actually Canceled. Also, there were 15 false positives, where bookings were incorrectly identified as Canceled but were actually Not Canceled.

The model achieved an accuracy of approximately 85.46% (significantly higher than the No Information Rate of 66.17%- better than a naive baseline that always predicts the most frequent class). This tells us the proportion of the total number of predictions that were correct for the test data. However, accuracy alone doesn't provide all the information that we want to evaluate our model, especially in cases where the dataset might be imbalanced between the two classes. The 95% confidence interval for accuracy (0.8162, 0.8877) suggests that we can be 95% confident that the true accuracy of the model is within this range.

Metric	Value
Accuracy	0.8546
95% CI	(0.8162, 0.8877)
No Information Rate	0.6617
Sensitivity	0.6815
Specificity	0.9432
Positive Predictive Value	0.8598
F1 Score	0.7603306

The F1 score is calculated (F1=2×((Precision+Recall) / (Precision×Recall))). Recall -Sensitivity is 0.6815 and is the proportion of actual positives (cancellations) that are correctly identified by the model. Precision -Positive Predictive value is the proportion of positive identifications (predicted as cancellations) that were actually correct (0.85). The F1 score is 0.76. This score means that the model is good at identify relevant instances, but can be better to reduce the number of false negatives. (F1 Score closer to 1 show a very good model to precision)

For the Support Vector Machine (SVM) Model Evaluation:

Confusion Matrix	Reference		
Prediction		Not Canceled	Canceled
	Not Canceled	238	45
	Canceled	27	90

The confusion matrix shows 238 true negatives and 90 true positives, show the model correctly identify 238 non-cancellations and 90 cancellations. There are 45 false positives, where bookings were incorrectly predicted as cancellations, and 27 false negatives, where cancellations were incorrectly predicted as non-cancellations.

The model achieved an accuracy of approximately 82% (significantly higher than the No Information Rate of 66.25%- better than a naive baseline that always predicts the most frequent class). This tells us the proportion of the total number of predictions that were correct for the test data. However, accuracy alone doesn't provide all the information that we want to evaluate our model, especially in cases where the dataset might be imbalanced between the two classes. The 95% confidence interval for accuracy (0.7788, 0.8564) suggests that we can be 95% confident that the true accuracy of the model is within this range.

Metric	Value
Accuracy	0.82
95% CI	(0.7788, 0.8564)
No Information Rate	0.6625
Sensitivity	0.8981
Specificity	0.6667
Pos Pred Value	0.8410
F1 score	0.869

The F1 score is calculated (F1=2×((Precision+Recall) / (Precision×Recall))). Recall -Sensitivity is 0.8981 and is the proportion of actual positives (cancellations) that are correctly identified by the model. Precision -Positive Predictive value is the proportion of positive identifications (predicted as cancellations) that were actually correct (0.8410). The F1 score is 0.869 (F1 Score closer to 1 show a very good model to precision). Here, we can see that the SVM is good at both accurately identifying true cancellations (precision) and making sure it identifies most of the actual cancellations (recall), without favoring one at the expense of the other too much. This balance is crucial because focusing too much on precision could mean missing out on identifying a lot of actual cancellations, while focusing too much on recall could mean wrongly labeling many non-cancellations as cancellations.

For the Logistic Regression Model Evaluation:

Confusion Matrix	Reference		
Prediction		Not Canceled	Canceled
	Not Canceled	230	53
	Canceled	35	82

The model predicted 230 bookings as not canceled (true negatives) and 82 bookings as canceled (true positives) correctly. It false predicted 53 bookings as canceled when they were not (false positives), and 35 bookings as not canceled when they were (false negatives).

The model achieved an accuracy of approximately 78 % (not significantly higher than the No Information Rate of 66.25% - better than a naive baseline that always predicts the most frequent class). This tells us the proportion of the total number of predictions that were correct for the test data. However, accuracy alone doesn't provide all the information that we want to evaluate our model, especially in cases where the dataset might be imbalanced between the two classes. The 95% confidence interval for accuracy (0.7362, 0.8196) suggests that we can be 95% confident that the true accuracy of the model is within this range.

Metric	Value
Accuracy	0.78
95% CI	(0.7362, 0.8196)
No Information Rate	0.6625
Sensitivity	0.8679
Specificity	0.6074
Pos Pred Value	0.8127
F1 score	0.8394

The F1 score is calculated (F1=2×((Precision+Recall) / (Precision×Recall))). Recall -Sensitivity is 0.867 and is the proportion of actual positives (cancellations) that are correctly identified by the model. Precision -Positive Predictive value is the proportion of positive identifications (predicted as cancellations) that were actually correct (0.8127). The F1 score is 0.83. This score means a good balance between precision (the model's ability to correctly identify cancellations out of all predicted cancellations) and recall (the model's ability to find all actual cancellations) (F1 Score closer to 1 show a very good model to precision).

4.4 The choice of the final model

The SVM model is the best choice for predicting cancellations based on the metrics accuracy and F1 score. The SVM model exhibits the highest F1 Score of 0.869 among the three models so not only accurately find cancellations (with a high precision of 0.8410) but also effectively recognizes a high proportion of actual cancellations (with a high sensitivity of 0.8981). Also, the SVM model accuracy is 82% (explained in detail in the previous paragraph), while the Random Forest model has a slightly higher accuracy of 85.46%, the bigger F1 Score of the SVM model show a better balance in performance across different types of bookings, making it more reliable. So, while each model has its strengths, the SVM has ability to balance precision and recall , with a high level of accuracy, makes it the most effective model for this prediction task.

5. Clustering

5.1 Distance measure and Linkage method selection

Firstly, we will select the distance measure and the linkage method to cluster our data. We choose the Gower distance because we are working with mixed data types (numeric and categorical variables) and we calculate the distance matrix. This distance measure allows us to effectively quantify the similarity between data points, considering the nature of our diverse dataset. For the linkage method, we choose the Ward linkage approach. The reason behind this selection is its efficiency in minimizing the total within-cluster variance, essentially ensuring that the clusters formed are compact and have minimal internal variance. This method is particularly effective in generating well-defined, cohesive groups. So, for the two clustering models that we will do we will use this distance measure and linkage method.

Agglomerative hierarchical clustering is a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged together step by step as one moves up the hierarchy. This method builds a tree of clusters called a dendrogram, which provides a visual representation of the each observation merging into larger clusters. K-medoids clustering is a partition technique similar to k-means, but instead of using the mean of objects in a cluster as center of the cluster, it uses actual objects, known as medoids. It's more robust to noise and outliers compared to k-means because medoids are less influenced by extreme values. This method partitions the data into k distinct non-overlapping (hard method of clustering) clusters based on minimizing the sum of dissimilarities between objects and their corresponding medoid.

5.2 Method Selection and Approach Description and choice of number of clusters

We will do clustering using agglomerative hierarchical clustering with Ward's method linkage and gower distance and K-medoids based on Gower distance.

We start be using 17 variables number of adults, number of children, number of weekend and weeknights, type of meal, car parking space, room type, lead time, market segment type, repeat customer flag, two numerical variables P.C and P.not.C, average price, special requests, reservation month, total guests, and total nights.

We start by doing hierarchical clustering because a lot of times the hierarchical clustering used to see from the dendrogram in how many clusters the data are split (see Appendix – Figure 9). From the dendrogram we can see 2 main clusters, these two clusters are depicted and in the silhouette plot. We can see there that the average silhouette width for all data points is 0.2, which falls into the positive but close to 0 range, so there is some structure to the clusters but not be very strong, and the separation between clusters isn't particularly distinct. The cluster labeled 1 contains 1,401 observations and has a low average silhouette width of 0.14, suggesting that the points within this cluster are not very close to each other, implying a weak structure. The cluster labeled 2 contains 599 observations with a higher average silhouette width of 0.31 (see Appendix-Figure 10). (the silhouette width ranges take values from -1 to 1, where values

near 1 imply that the observation is well matched to its own cluster and poorly matched to neighboring clusters, values near 0 suggest overlapping clusters, and values near -1 indicate that points might have been assigned to the wrong cluster- each line is an observation). In conclusion, the silhouette plot show that the current hierarchical clustering might not be optimal as the clusters are not well-separated and individual observations are not close to the centroid of their respective clusters, especially in cluster 1.

We did at first hierarchical clustering to take a first picture of how many clusters probably exist from the dendrogram and then we will continue with K-medoits because a lot of times gives better results.

With K-medoits clustering in the silhouette plot we can see that in cluster 1 we have 1,268 observations with an average silhouette width of approximately 0.18, which show a weak group (the observations in this cluster are not very close to each other, and there's some uncertainty in the cluster assignments. In cluster 2 we have 732 observations with a better average silhouette width of approximately 0.27. The overall average silhouette width for the clustering is 0.21, which is relatively low, so the separation between the two clusters is not very good (see Appendix-Figure 11). Also, we conducted K-medoids with multiple starting points, but the results we took were the same.

5.2 Determining the Number of Clusters

Initially, 2 clusters were considered optimal from the first picture that we take from the dendrogram (see Appendix-Figure 9). For K=2 and with K-medoits, the silhouette plot show one cluster with a relatively higher average silhouette width (0.266), suggesting better cohesion and separation for this cluster compared to the K=3. Despite K=3 showing a higher average silhouette score for one of the clusters, the overall structure wasn't as clear as with K=2.We did the elbow method (visualize through within-cluster sum of squares and silhouette) and we see a sharp decline from 1 to 2 clusters, after which the curve flattens, showing that increasing the number of clusters beyond 2 does not lead to significantly better within-cluster cohesion . However the “elpbow” is created for K=3 clusters (see Appendix -Figure 12). The silhouette method showed that the average silhouette width for K=2 was slightly lower than for K=3 (see Appendix -Figure 13). However, considering both the silhouette scores and the elbow method, we saw that K=2 provided a more distinct and cohesive clustering solution.

Also, the Adjusted Rand Index (ARI) is a measure used to evaluate the similarity between two clusterings (an ARI value close to 1 indicates a perfect match between two clustering solutions, while a value close to 0 indicates random clustering, and negative values suggest less than chance agreement).We compared the clustering results for K-medoids with K=2 and K=3 resulted in an ARI of approximately 0.628. This suggests a moderate agreement between the two clustering solutions. They are not identical, showing some differences in how the clusters are composed, but the ARI value also indicates that they are not entirely dissimilar.

After that we split into two subsets to check the consistency of K-medoids clustering regardless of how the data was divided. The aim was to assess the stability of the clusters: whether similar clusters emerge from different subsets of the data and also if something exist and in the 2 subsets then probably is true. For both subsets, the K-medoids algorithm was

applied with K=2, and the silhouette plots were generated to evaluate the clustering quality. The results showed that subset 1 had clusters with average silhouette widths of 0.28 and 0.18, while subset 2 showed widths of 0.19 and 0.25, respectively. Although the silhouette scores are not high, showing moderate separation, the cluster frequencies were consistent across both subsets. The aggregate statistics calculated for both subsets also indicated similarities in the clustering characteristics. This consistency in results across different subsets suggests that the K-medoids clustering is stable and the groupings are not highly sensitive to the specific subset of data used, implying that the clusters are a genuine reflection of the structure in the data (see Appendix -Figure 14).

After the variable selection, the choice of 2 clusters was more clear (see Appendix – Figure 15).

5.3 Clustering Variables Selection

We aim to retain only those variables that have meaningful information for clustering and exclude any that may introduce noise. So we did ANOVA test on the numeric variables in the dataset to see which variables significantly contribute to the difference between clusters. The test revealed that the variable number of week nights, average price , P C showed no significant difference between clusters, as their p-values were higher, showing that they may not be important for the clustering process. Additionally, chi-squared and Fisher's exact tests were applied to the categorical variables to assess their relationship with the clusters. It was found that type of meal, car parking space, room type, market segment type, repeated customer, and reservation month were all significantly associated with the clusters, thus they should be retained in the model.Based on these results, the less significant variables were removed from the dataset, and after the remove of the variables we did again in the dataset ANOVA and chi-squared and Fisher's exact tests with the remain variables and we remove and the P not C .Then for categorical variables, the association between each categorical variable and the clusters was measured using tables and the Adjusted Rand Index (ARI). The ARI compares the similarity of the cluster assignments given by the k-medoids algorithm to the actual labels provided by each categorical variable(a higher ARI indicates a stronger agreement).From the frequency tables and ARIs , we show that variables like market.segment.type have a big impact on clustering with an ARI 0.77, so a strong correspondence between this variable and the 2 clusters . However, other variables like type.of.meal, car.parking.space, and room.type show little to no correspondence with the cluster assignments and we remove them . In the end we remain with 7 variables number.of.adults ,number.of.children ,number.of.weekend.nights, lead.time ,market.segment.type ,special.requests ,total.nights .

5.4 Assessing Cluster Quality

The clustering quality of our analysis was done using silhouette scores. The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters, with a higher score showing a better fit (as we explained before with more details).For your k-medoids clustering we have an average silhouette width of 0.64, so we have a good separation between the two clusters. The first cluster, with 1268 units, has a slightly higher silhouette width of 0.65, show that the observations within this cluster are more cohesive and well-matched. The second cluster, containing 732

observations, also shows a strong silhouette width of 0.61, so this show that the observations within are appropriately grouped. These scores are above average, showing that the clusters are well-separated and that observations are generally closer to the centroids of their own clusters than to those of other clusters. The consistency of the silhouette scores before and after the hierarchical clustering confirms the reliability of our clustering results (see Appendix–Figure 15,Figure 16).

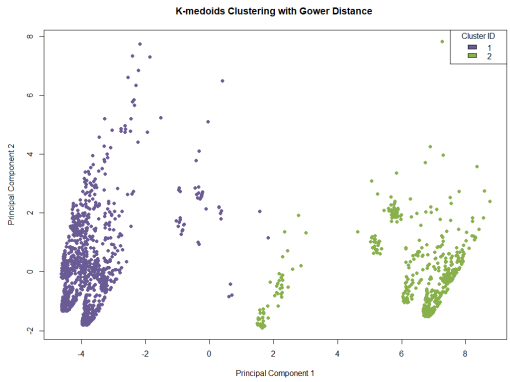

Figure 16 : K medoids clustering

5.5 Interpreting Clusters

In the end we remain with 7 variables number.of.adults ,number.of.children ,number.of.weekend.nights, lead.time ,market.segment.type ,special.requests ,total.nights that have meaningful information for our clustering .

For the first cluster, the average number of adults is slightly higher (1.92) compared to the second cluster (1.73), showing a potential trend of larger adult groups within the first cluster's bookings.Both clusters show a preference for booking with two adults as their median. The number of children is notably low across both clusters, with the first cluster having a slightly higher mean of children (0.16) compared to the mean in the second cluster (0.02).For weekend night stays, the first cluster appears to slightly favor longer weekend stays with an average of 0.86 nights compared to 0.71 in the second cluster. This preference is also seems in the median numbers, where the first cluster tends to book at least one weekend night, while the second cluster has a median of zero, showing no strong preference for weekend stays. Lead time, shows a significant difference between the clusters. The second cluster plans their stays well in advance, with an average lead time of 104.32 days, compared to the first cluster's 71.16 days. Market segment type mean values indicate that the first cluster might be slightly more inclined towards a specific market segment (mean of 4.91) compared to the more diverse market segment interests in the second cluster (mean of 3.91). The special requests metric reveals a stark contrast between the two clusters; the first cluster has a higher tendency to make special requests (average of 0.90) versus the second cluster (average of 0.17).The total nights stayed show a relatively close average between the clusters, with the first cluster booking an average of 3.17 nights and the second cluster booking slightly fewer nights on average (2.86).

We make some boxplots you've to visualize the distributions of the number of adults, children, weekend nights, and special requests for two distinct clusters. These visualizations help to see central tendencies, variability, and outliers within each cluster. Cluster 1 tends to have more adults on average compared to Cluster 2, as we show by the boxplot and descriptive statistics. Special requests are higher in Cluster 1, both in terms of the average number and the presence of

outliers. The number of children is relatively low in both clusters, but Cluster 1 has a few more observations with a higher number of children. Weekend nights booked show a wide range in Cluster 1, showing more variability in the length of stay for weekends within this cluster.

We make some histograms and we show that for Cluster 1, the histogram of total nights shows a strong preference for shorter stays, with the majority of stays being fewer than 5 nights. Cluster 2, shows a more evenly distributed range of stays, extending up to 14 nights, but still the max is at shorter stays. For the lead time for bookings, Cluster 1 displays a skewed distribution with a high frequency of short lead times, peaking near the time of booking. This show that customers in Cluster 1 tend to make their reservations closer to their stay date. Cluster 2 lead time histogram also leans towards shorter lead times but has a more gradual decline and a wider spread, showing that customers in this cluster plan their stays with more varied notice periods (see Appendix – Figure 16).

8. Appendix

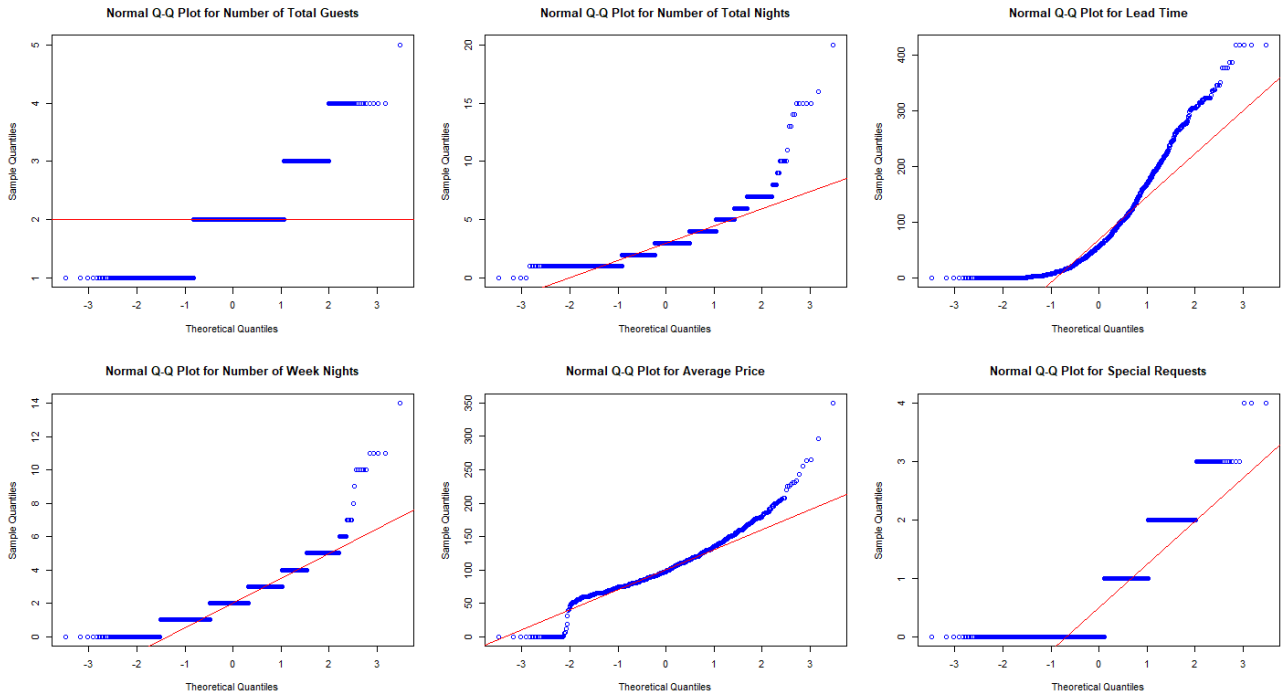

Figure 2: QQplots for number of total quests, number of total nights, lead time, number of week nights, average price, and special requests variables

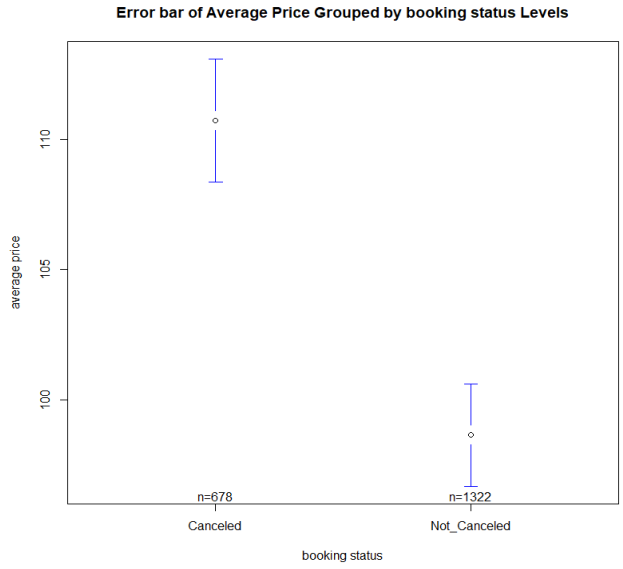

Figure 5: Error bar of Average Price Grouped by booking status Levels

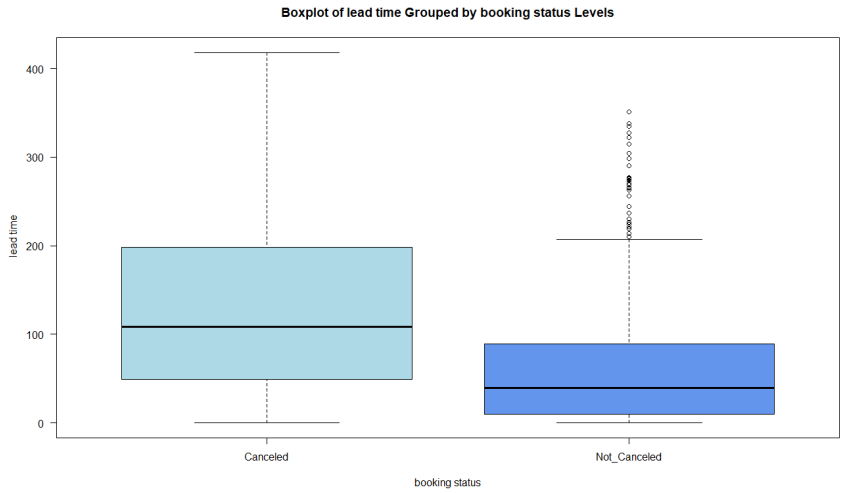

Figure 6: Boxplot of lead time Grouped by booking status Levels

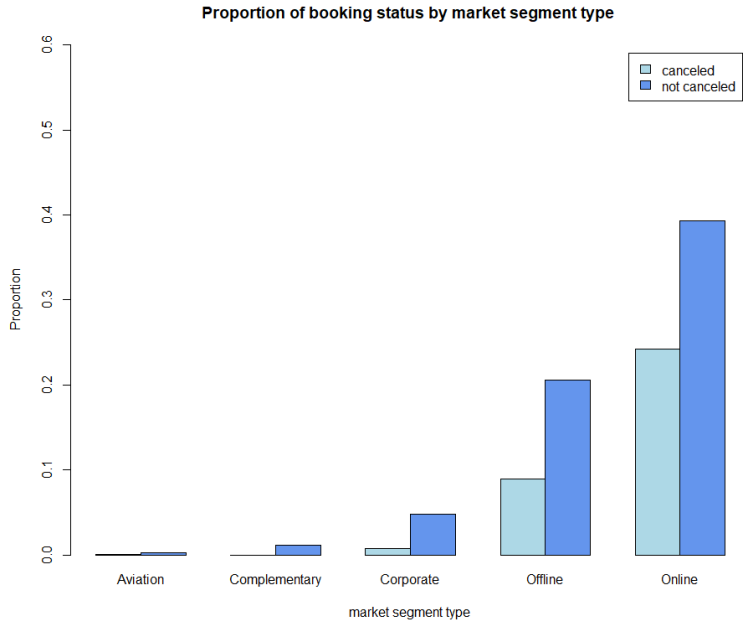

Figure 7: Barplot for the Proportion of booking status by market segment type

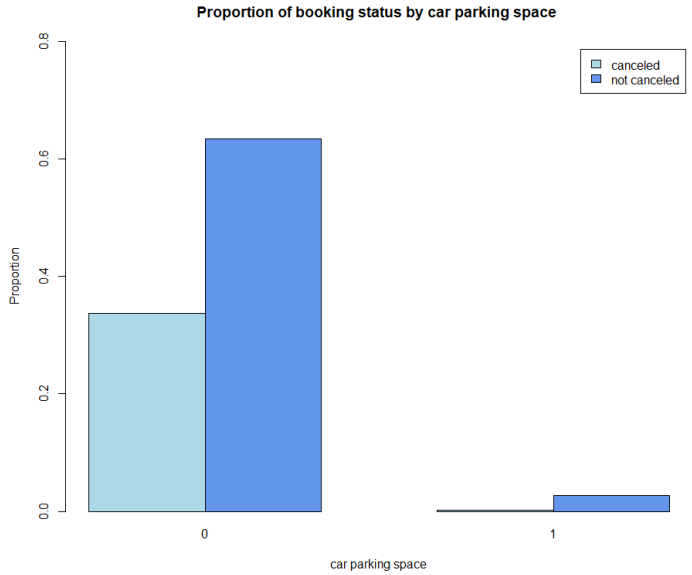

Figure 8: Barplot for the Proportion of booking status by car parking space

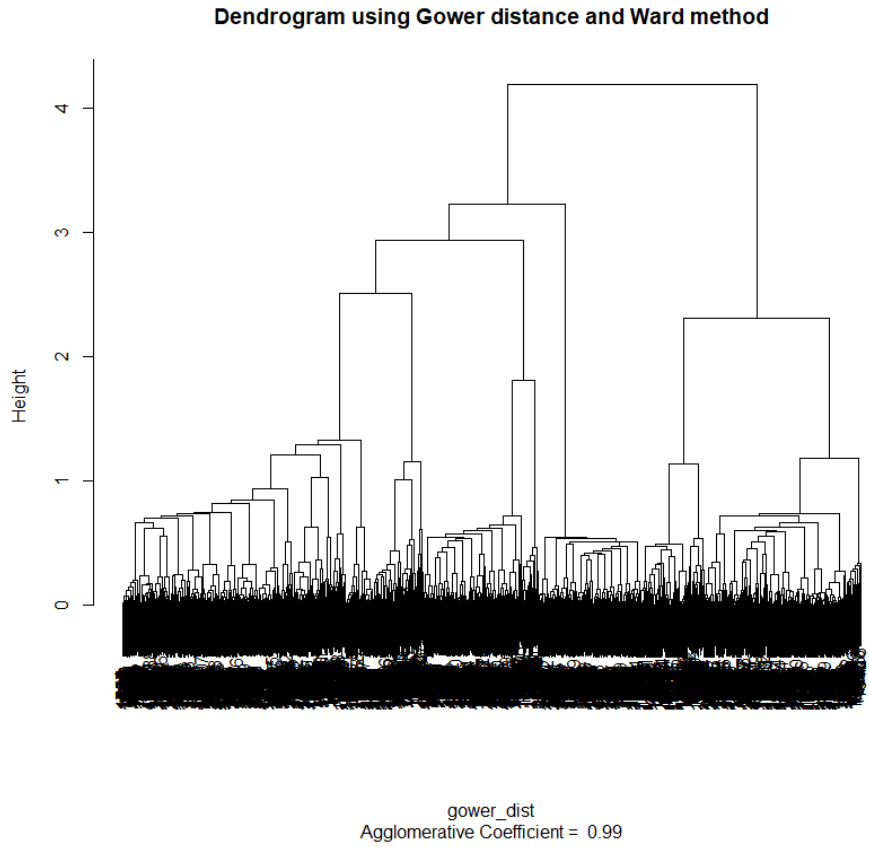

Figure 9: Dendrogram from the hierarchical clustering.

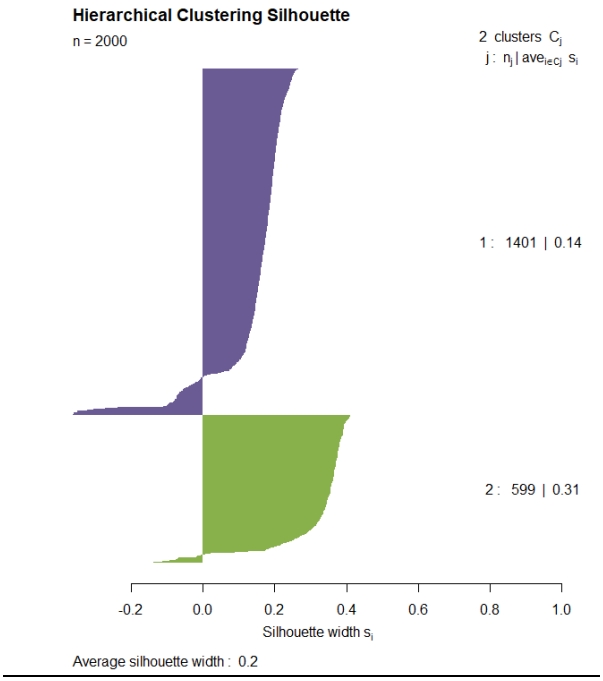

Figure 10: Hierarchical clustering Silhouette

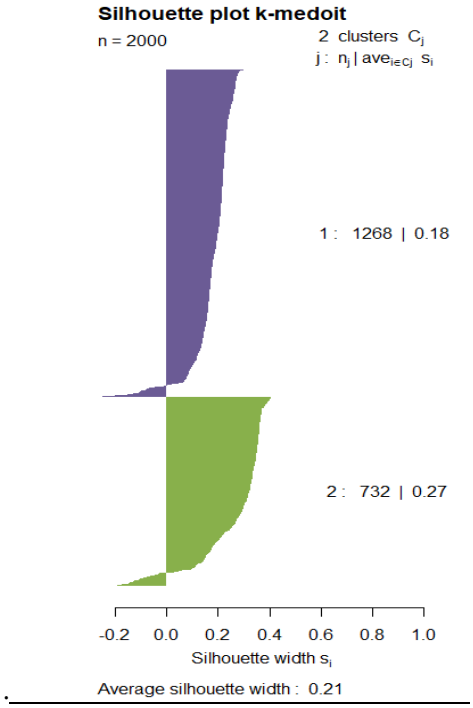

Figure 11: K-medoits clustering Silhouette

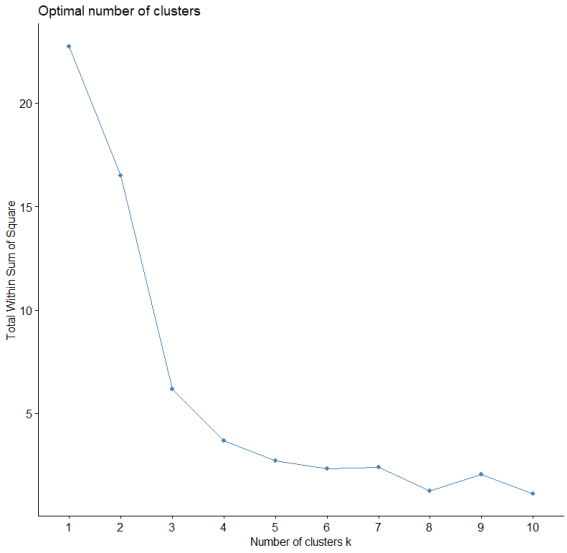

Figure 12: elbow method

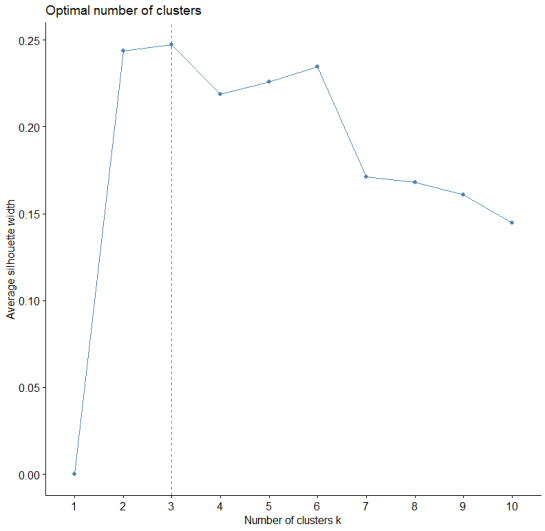

Figure 13: Average width Silhouette scores for the number of cluster

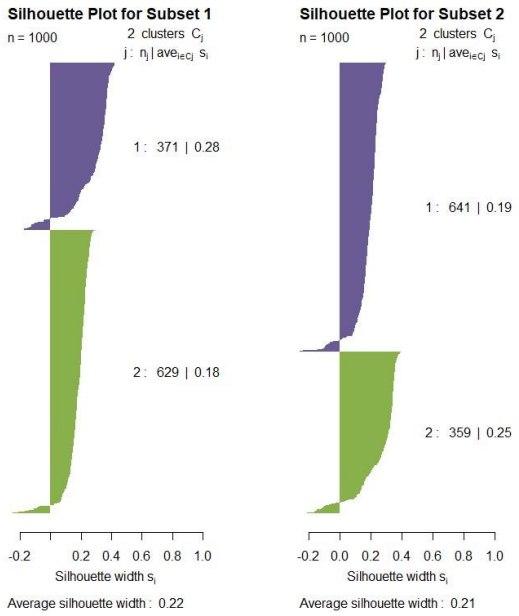

Figure 14: Average width Silhouette scores for the two subsets

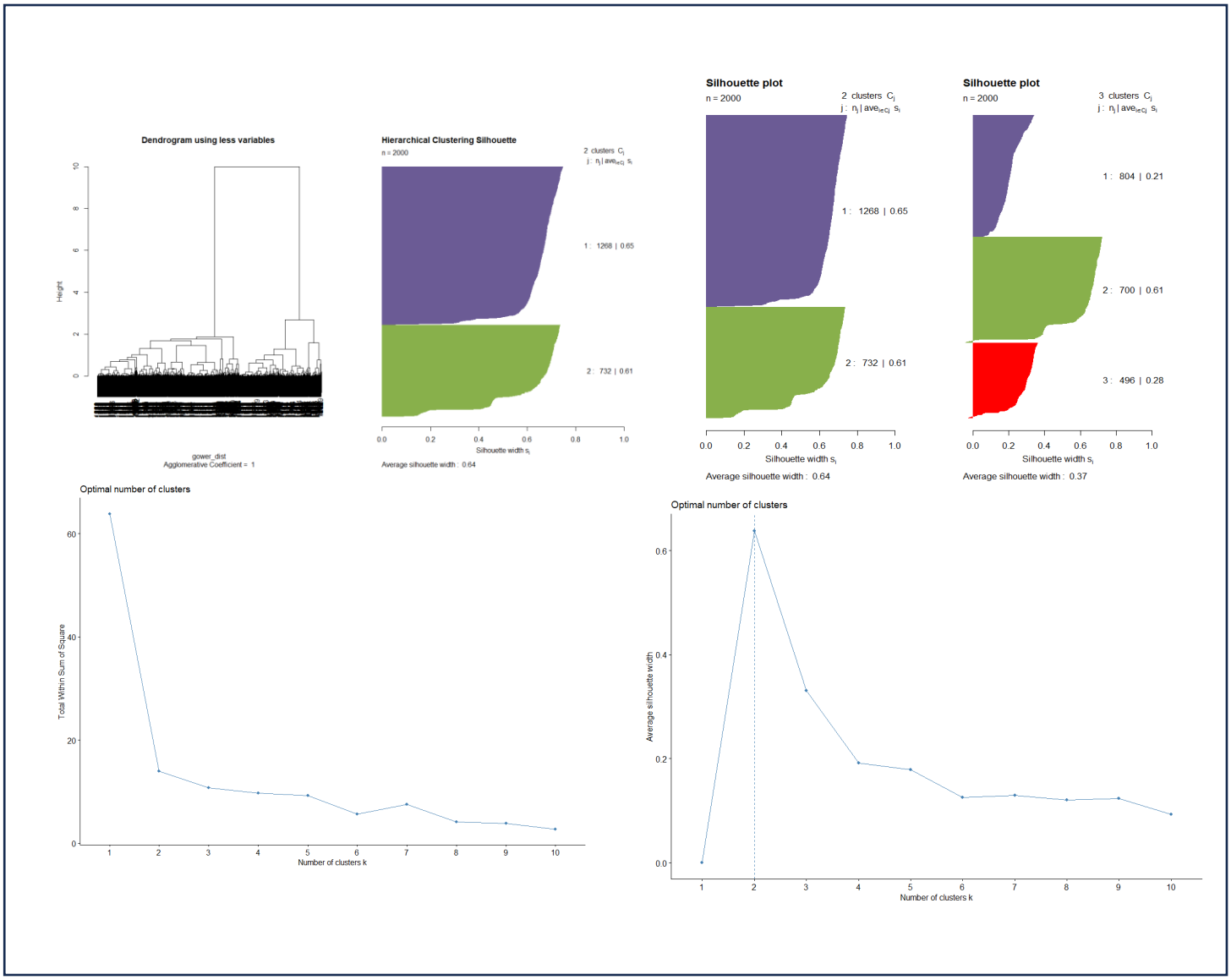

Figure 15: dendogram and Average width Silhouette scores after the removal of the variables and elbom method plots

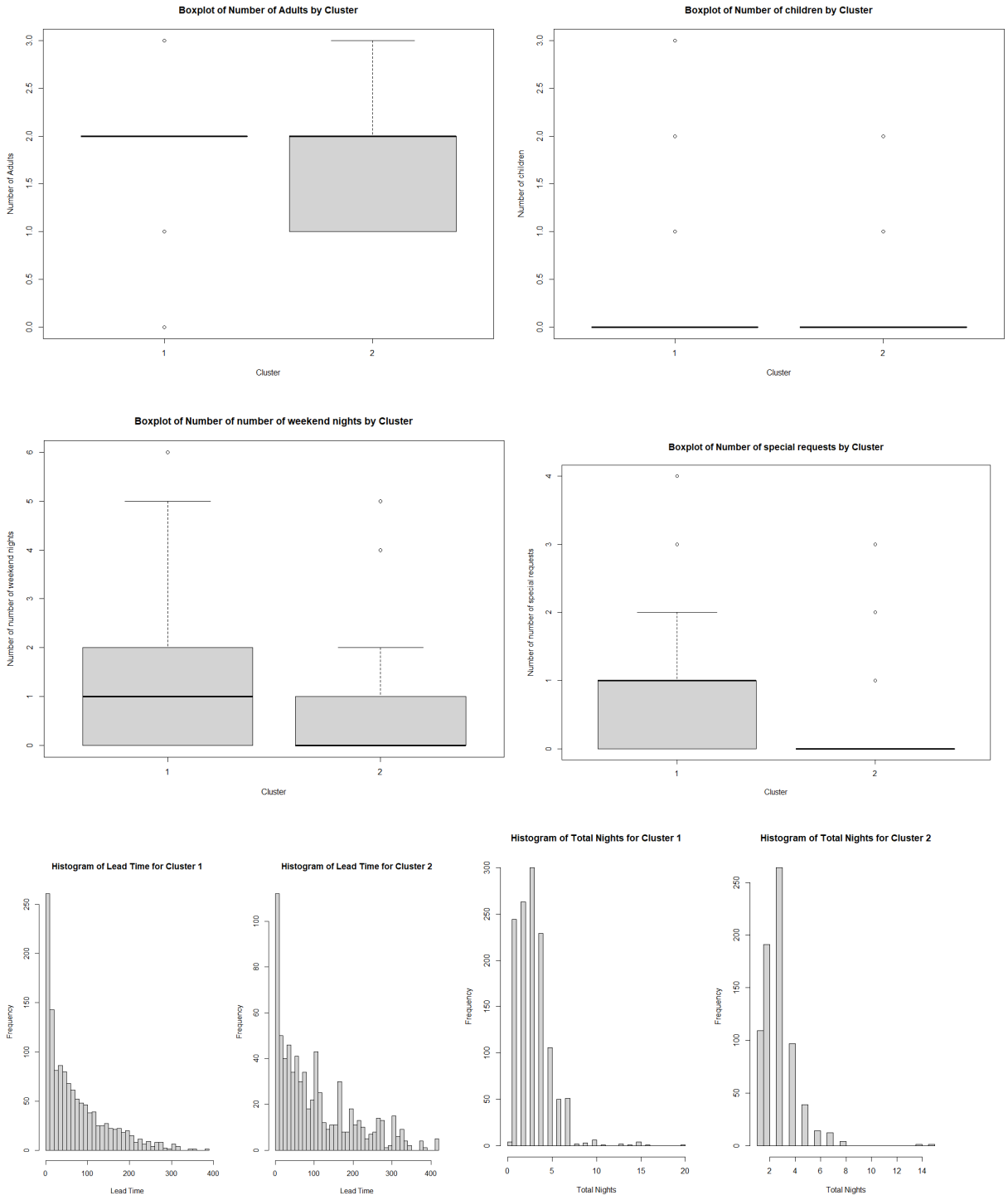

Figure 16: boxplots and histograms for clustering interpretation .