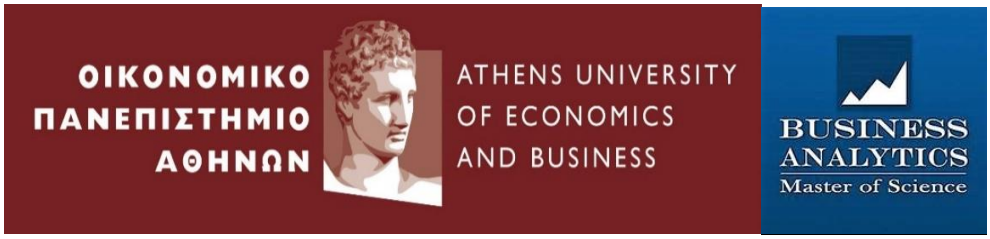


Department: Department of Management Science and Technology

M.Sc.: Business Analytics (Full-Time)

Course: Statistics for Business Analytics II

Professor: Dimitris Karlis

Deliverable: Second Project (Classification and Clustering)

Student Name: Evangelos Lakkas-Pyknis

Registration Number: f2822306

E-mail: eva.lakkaspyknis@aueb.gr

Introduction

This assignment aims to study a given dataset of 2000 observations and 16 variables, provide classification predictions for the cancelations (or not) of hotel bookings in the first part, and explore if clusters of bookings can be found based on common characteristics in the second part. So, this report consists mainly of two parts Classification and Clustering. In each part, I display all the tried-out methods along with an evaluation of their performance. Finally, I select the best method I found and give the best possible interpretation for it.

Remarks about the data

Some brief comments on the processing of the dataset, there was an issue with the feature "Date of reservation" because many available values were not formatted as "year-month-day" but as numbers, I fixed this problem as described in the report of the previous project. Also, I created a new column in the dataset corresponding to the month of each date[1]. I did this transformation because I assumed that month could give me some more useful information than individual dates. After the conversion of the dates two missing values were introduced in the dataset but given that the observations with NA values were only two out of 2000 (a very small part of the sample) I dropped the observations with these two values. I subtracted the feature "Booking ID" from the dataset because it is an identifier based on historical data so it will not be useful for my analysis. Finally, I did some cleaning on the average price column because some rows had unreasonably small price values, so I replaced these observations with a price equal to 30. So, the final dataset I had available had 1998 observations and 16 features. Also I used a 75-25 split for the training and testing datasets in the first part of the project.

[1] The criterion based on which I thought this assumption reasonable is explained more analytically in the introduction of my report of the first project.

Part I: Classification

1.1) Logistic Regression

The first method that I tried to predict if a booking will get canceled or not was logistic regression. I fitted the full model in order to perform a backward step-wise procedure using the Akaike Information Criterion (I chose AIC over BIC given that this index is generally preferred for prediction tasks). After doing this process I ended up with a logistic regression model that consists of the variables: number of weekend nights, car parking space, repeated, market segment type, average price, lead time, and special requests. I subtracted the repeated variable because it was statistically insignificant for the model (p-value= 0.97>0.05) and the accuracy had a minor change which didn't indicate a worse model. So, I preferred to simplify the model. The new model has an AIC score of 1356.6 against 1369.3 of the null model. Its mathematical formulation is shown below:

$$: \log\left(\frac{p}{1-p}\right) = 3.41 - 0.15 \times weekend.nights - 0.01 \times lead.time + 2.07 \times parking - 0.01 \times avg.price + 1.49 \times sp.requests + 12.70 \times ComplementaryType - 0.07 \times CorporateType + 0.81 \times OfflineType - 1.04 \times OnlineType$$

Where:

p: the probability of the booking to get canceled.

$log\left(\frac{p}{1-p}\right)$: the log odds ratio of the booking status variable (i.e. the logarithm of the probability of the booking to getting canceled divided by the probability of the booking not getting canceled).

Parking: a dummy variable taking the value 1 if the booking has parking space or the value 0 otherwise

Complementary Type, Corporate Type, Offline Type and Online Type are dummy variables which take the value 1 if the Market Segment variable correspond to their level or 0 otherwise. If all of them take value equal to zero then market segment type is Aviation.

The accuracy of the logistic regression model in the testing dataset is 78.57% which is good for a first try but I will try to find something better in terms of accuracy. The confusion matrix (i.e. the true values of the booking status against the predicted ones) from which accuracy metric is calculated is shown below:

	Predicted Values	
Booking status (true values)	Cancelled	Not Canceled
Cancelled	108	61
Not Canceled	44	286

The interpretation for each term of this model is given below:

Intercept: the probability of the booking getting canceled is equal to $\frac{1}{1+e^{3.39}}$≈0.03 or 3% when all the other covariates are equal to 0, the booking has no parking space available, and the market segment type is Aviation. This fact may not be very meaningful because the price cannot be assumed to be 0.

Weekend nights: The probability of the booking getting canceled has a negative relationship (this is shown by the negative sign of the coefficient) with the number of weekend nights which means that an increase in this variable will decrease the probability of the cancelation. For the increase by one weekend night, the odds ratio (i.e. the probability of the booking getting canceled divided by the probability of the booking being canceled) decreases by $e^{-0.15}$≈0.86, when all the other variables remain constant.

Lead time: The probability of the booking cancelation has a negative relationship (this is shown by the negative sign of the coefficient) with lead time which means that an increase in this variable will decrease the probability of the cancelation. If a booking gets reserved one day earlier (considering that the arrival date remains the same) the odds ratio decreases by $e^{-0.01}$≈0.99, when all the other variables remain constant.

Car parking space: The probability of the booking cancelation has a positive relationship (this is shown from the positive sign of the coefficient) with the parking space which means that the existence of parking increases the probability of the cancelation. When the booking has available parking, the market segment type is Aviation, and all the other covariates are constant the odds ratio becomes $e^{2.07}$≈7.92.

Price: The probability of the booking cancelation has a negative relationship (this is shown by the negative sign of the coefficient) with the average price which means that an increase in this variable will decrease the probability of the cancelation. If the average price increases by one dollar the odds ratio (i.e. the probability of the booking getting canceled divided by the probability of the booking not getting canceled) decreases by $e^{-0.01}$≈0.99, when all the other variables remain constant.

Special requests: The probability of the booking cancelation has a positive relationship (this is shown by the positive sign of the coefficient) with the number of special requests which means that an increase in this variable will increase the probability of the cancelation. For every special request that a booking gets the odds ratio increases by $e^{1.49}$≈4.43, when all the other variables remain constant.

Market Segment Type: The probability of the booking cancelation has a positive relationship with all the levels of the market segment type except the online type. This means that if a room gets booked online it is less likely to get cancelled.

1.2) High-Dimensional Discriminant Analysis

After that, I wanted to do a discriminant analysis based classification. I checked the assumption of the Linear Discriminant Analysis to see if homogeneity for the covariance matrices of the two classes can be assumed. This hypothesis is rejected (Box's M Test, p-value < 2.2×10^{-16} < 0.05) so LDA is not an appropriate method for this dataset. Also another issue is that my dataset contains several dimensions so I preferred the High-Dimensional Discriminant Analysis method and not Quadratic DA to avoid possible numeric

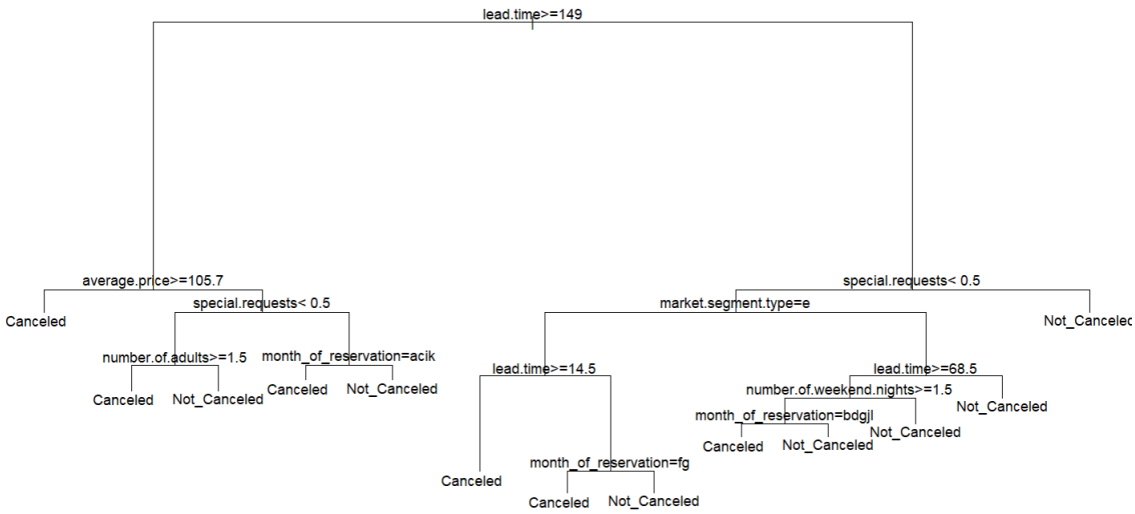
issues and select the most useful variables. After the training of the HDDA method in the training dataset I calculated its accuracy on the testing dataset which is equal to 66.13% which is poor performance if we compare it with the logistic regression model. The confusion matrix (i.e. the true values of the booking status against the predicted ones) from which accuracy metric is calculated is shown below:

	Predicted Values	
Booking status (true values)	Cancelled	Not Canceled
Cancelled	0	169
Not Canceled	0	330

Although the accuracy metric does not seem too bad, a thing that seems worrying on the results of this method is that there are 169 that were cancelled but the model did not classify them as such. So, the logistic model is better so far.

1.3) Decision Tree

The next method I tried was the decision tree, I tried this method to select variables more easily and based on the results to see If this method can give better predictive ability than the logistic regression and see if it is worth exploring (ensemble) methods based on decision trees (Random Forests). The original tree I created had the structure that is shown below. However, we can see that this classifier is overcomplex in its current form and this can lead to overfitting. Also, we observe that in many leaves particularly on the right part of the tree due to the overcomplexity the tree loses the easy explainability that it should have.

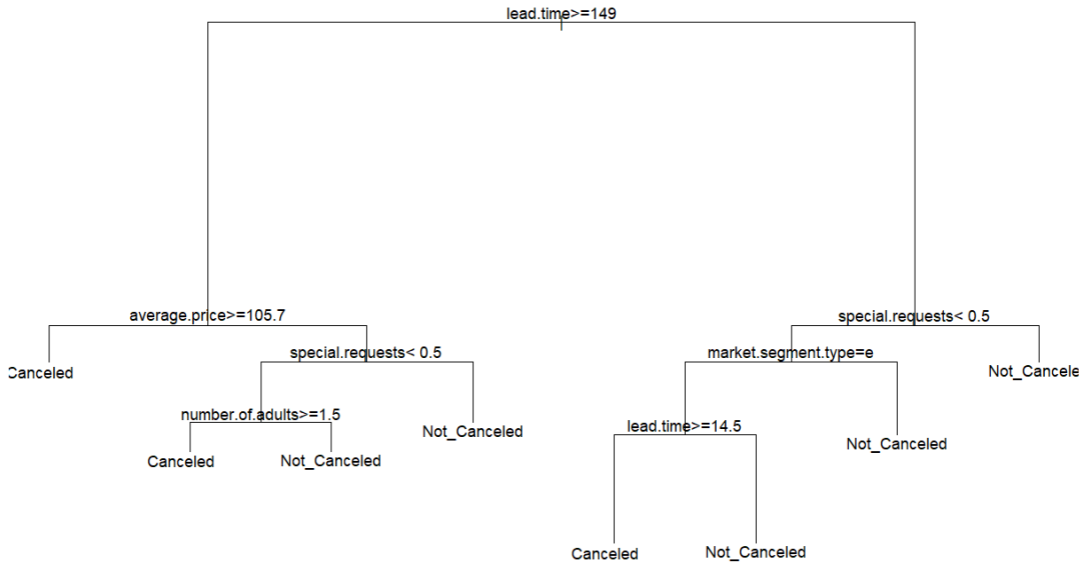


After checking the performance of the tree on the testing dataset I see that its accuracy which is 79.35% indexes a useful classifier the in-sample accuracy is 84.25% which is a respectable difference but based on this index overfitting does not seem to be a very severe problem but there might be a danger if we want to implement this classifier on a different dataset to make predictions in the future. So, I tried to prune the decision tree at the third split and when I calculated the accuracy in the testing dataset I saw that the pruned

tree that I have created is now marginally better in terms of accuracy because the accuracy is now 79.75% and the difference between the training and testing dataset has now decreased significantly as the in sample accuracy is 81.62%. The confusion matrix for the pruned decision tree is displayed below:

	Predicted Values	
Booking status (true values)	Cancelled	Not Canceled
Cancelled	111	58
Not Canceled	43	287

So given that the accuracy difference between the in sample and out of sample in the original tree was approximately 5% while in the pruned tree it is 2% so it is reasonable to say that the danger for overfitting has deteriorated. The final decision tree is shown below:



Also, after the pruning that I have implemented previously the decision tree has now become more explainable. Now we can distinguish the most important variables which are mainly lead time, special requests, and average price but also the number of adults and market segment type play a role. In this diagram, we can see more clearly not only which are the most important variables (we could do this also with the previous tree to an extent) but now we can interpret this model and say that if a booking has been booked 149 days or more before the arrival (we see this due to the presence of lead time in the top of the tree) and has average price larger than 105.7 it is most likely to get canceled. If a booking has lead time of less than 149 days, no special requests, and also a market segment different than online it is most likely not to get cancelled. Similarly, we can interpret the other leaves of the tree.

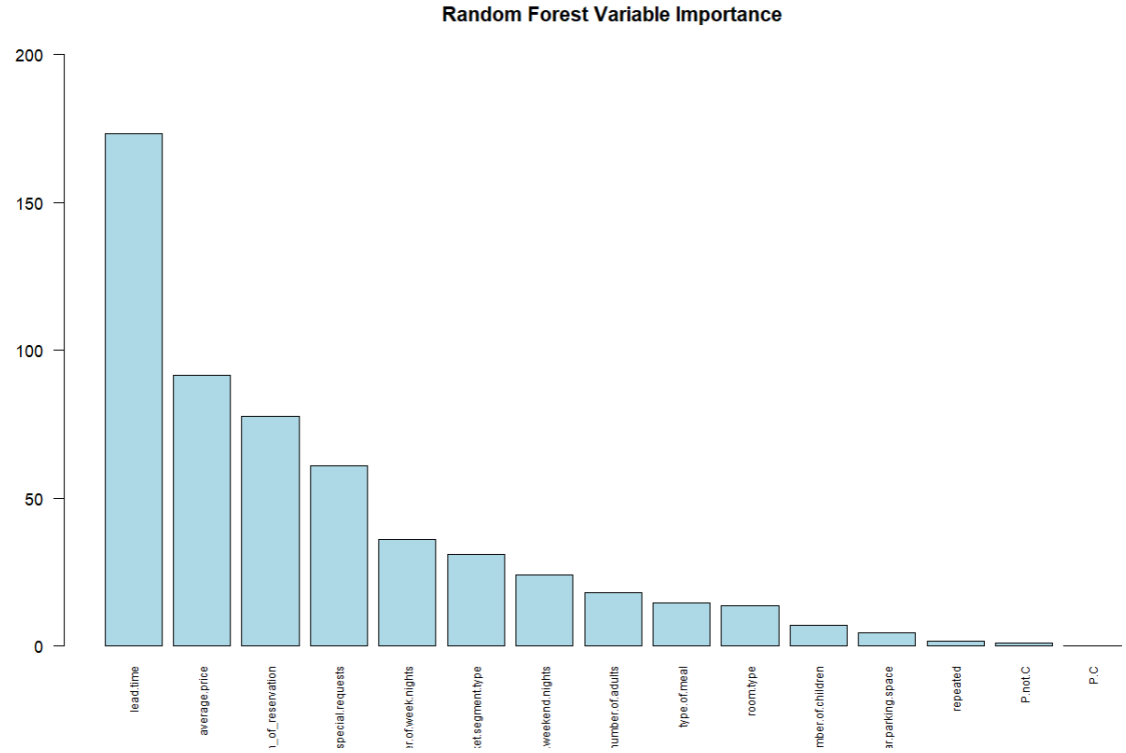
So far the decision tree is in terms of accuracy (and explainability) the best predictive model and it is slightly better than the logistic regression (accuracy= 79.75% > 78.57%). However, I will try one more method to because the accuracy is still relatively moderate due to the fact that no model has accuracy larger than 80%.

1.4) Random Forest

After seeing that decision trees are useful for this classification case I tried a Random Forest method (with 100 trees) in order to get better estimates for my data by not using an individual tree but an ensemble method. This method indeed gave me the best results in terms of accuracy which was 83.36% in the testing dataset[2]. After seeing that this method had the best performance in the testing dataset, I performed a 5-fold cross-validation procedure using the whole dataset this time (and not just the training) to examine if I would get equally good results in more than one testing datasets in case that the original testing data weren't representative. The mean accuracy from the 5-fold CV is 84.21%. So, Random Forest is the best method that I have found for this assignment below I show a table where there the accuracy scores for each tried-out method.

Method	Accuracy Score
Logistic Regression	78.57%
High-Dimensional Discriminant Analysis	66.13%
Decision Tree	79.75%
Random Forest (on testing dataset)	83.36%
Random Forest (5-fold CV evaluation)	84.21%

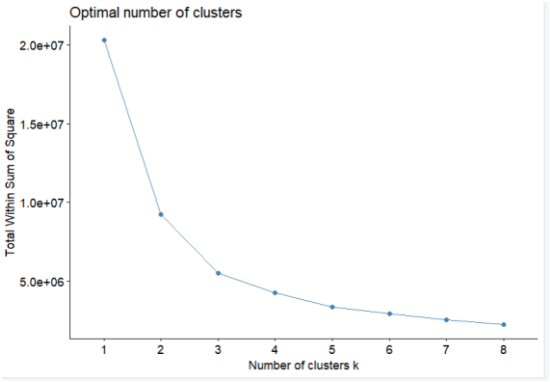
The downside of the random forest method is that it has not a physical interpretation. However, we can see from the bar plot below what the most important variables for the predictions of this method are.

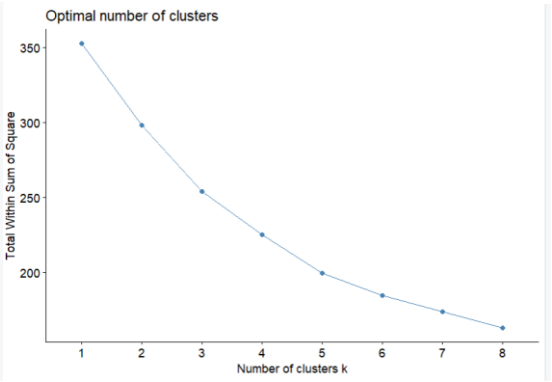


[2] Note that each time that I executed the random forest testing I was getting a slightly different accuracy however this accuracy was never less than 82% and larger than 84.5%. Here I am referring 83.36% indicatively because that was the last accuracy I calculated before writing the report.

Part II: Clustering

Now for the second part of the assignment before calculating the distances for the dataset and applying different clustering methods I created a second dataset where I scaled the data for the numeric variables by subtracting the minimum value (of each variable) for each observation and dividing by the range in order to see if the calculation of distances on these data can give a better clustering later on. Also note that given that the task which I want to solve now refers to unsupervised learning I will not split my dataset into training and testing.





After this transformation of the data, I examined what could be the optimal number of clusters with the elbow method on the left we see that the diagram suggests that the optimal number of clusters is 2 for the unscaled data due to sudden change in the slope of the curve at this point while on the right it is not so clear what is the optimal point for the scaled data it could be it appears that a good point for the choice of clusters could be between 3 and 4. The distance I used for the clustering is the Gauer distance due to the fact that my dataset has both categorical and numeric variables so I need a distance that can work with this type of data. I calculated the Gower distance both for scaled and unscaled data to see which of these two can give better clustering results.

2.1) Hierarchical Clustering

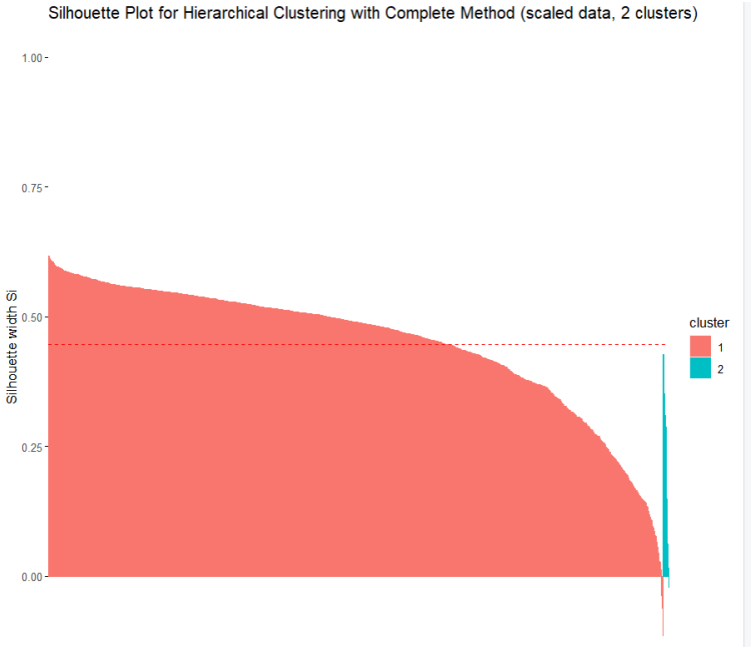
The first method I tried was Hierarchical clustering with Ward linkage. As I show below Ward method gave me a poor performance overall with the first try. The use of the distance that was calculated on the scaled data gave better clustering on the second cluster (the avg. silhouette value rises from 0.24 to 0.40) however this improvement does not happen in the first cluster when I compare results of this method from the scaled and unscaled. Also, there is an improvement in the second cluster (when comparing the unscaled data to the scaled) considering the number of negative silhouette values which on the table with silhouette values below which are shown to be significantly less.

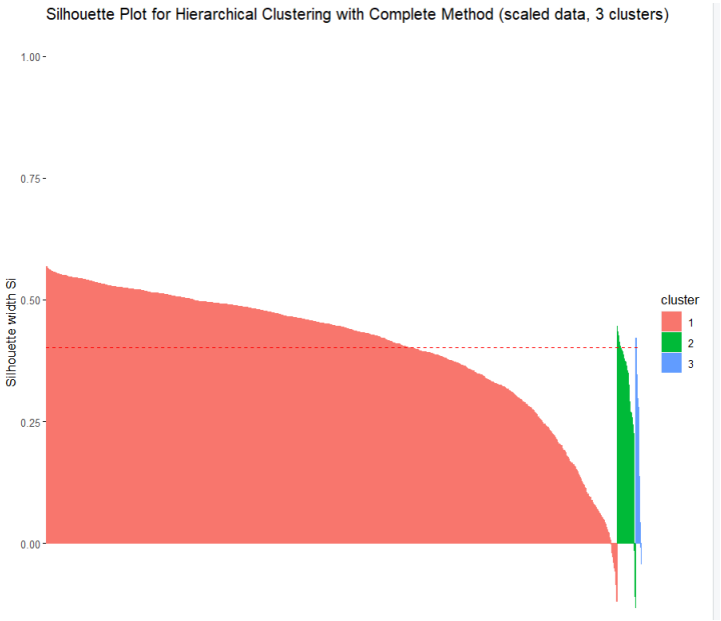
Method	Avg. Silhouette Values	Number of negative Silhouette Values	Observations per cluster
Ward (unscaled data)	0.23, 0.24	72	1247, 751
Ward (scaled data)	0.21, 0.40	23	1582, 416

After that I tried again the approach of Hierarchical Clustering but this time with complete linkage as we can see from the table below the complete linkage performs better than Ward on the unscaled data since the silhouette values are 0.35 and 0.42 for cluster 1 and 2 respectively. These silhouette values are also the best I have found so far which indicate for a moderate to good clustering. However, a thing that draws my attention is that the first cluster has 1954 observations and the second one only 44 which is not necessarily unreasonable but I will check if I can find another cluster method with equal or better results in terms of sil. values but with more balanced numbers of observations.

Method	Avg. Silhouette Values	Number of negative Silhouette Values	Observations per cluster
Complete (unscaled data)	0.35, 0.42	35	1954, 44
Complete (scaled data) 2 clusters	0.21, 0.45	8	1979, 19
Complete (scaled data) 3 clusters	0.41, 0.31, 0.20	29	1916, 63, 19

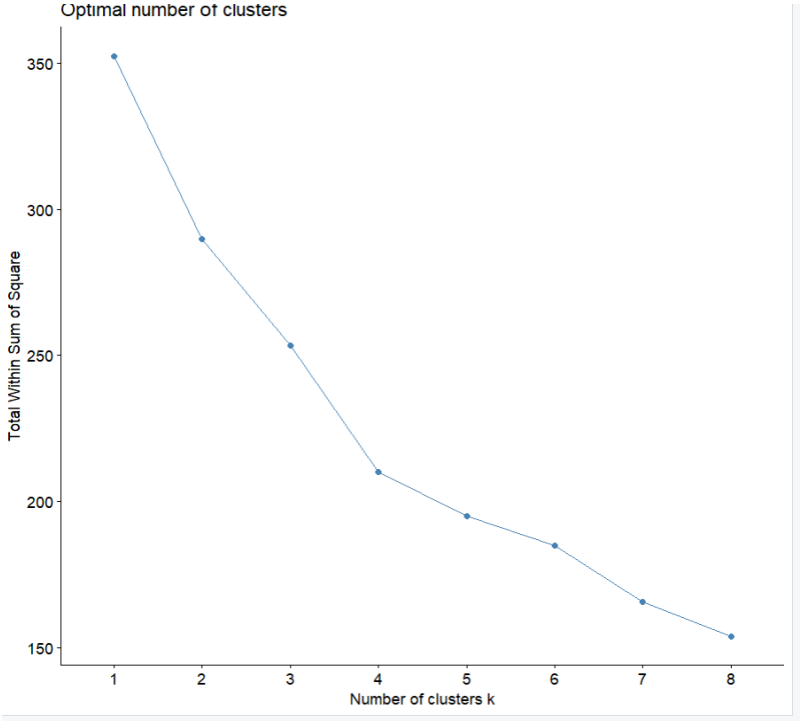
Also, considering the scaled data I tried 2 approaches one with 2 clusters and one with 3 clusters because the elbow diagram that I did in the beginning for the scaled data had not a clear angle in a specific point so I wanted to try different number of clusters to see which can be better. Some things that can be seen in the sil. plots below is that when I create a clustering with 3 groups the negative values rise significantly compared to the clustering with 2 groups that is indicating for placement of observations in wrong clusters. Also, we see that in the last cluster in both approaches the number of observations remains the same. So given these facts I prefer to keep the 2 clusters for the scaled data. Overall, for the complete linkage I found that it performs better on the unscaled dataset based on the silhouette values and also I notice that the sil. values have a relatively small gap between the two clusters. So, the best method I have found so far is the Hierarchical clustering with Complete linkage on the unscaled data.





2.2) K-Medoids Clustering

The last method of clustering that I tried was k-medoids on the scaled dataset but the results I got weren't as good as the hierarchical clustering's. First of all I implemented again the elbow method to see what could be the optimal number of clusters for this algorithm. The elbow diagram can be seen below

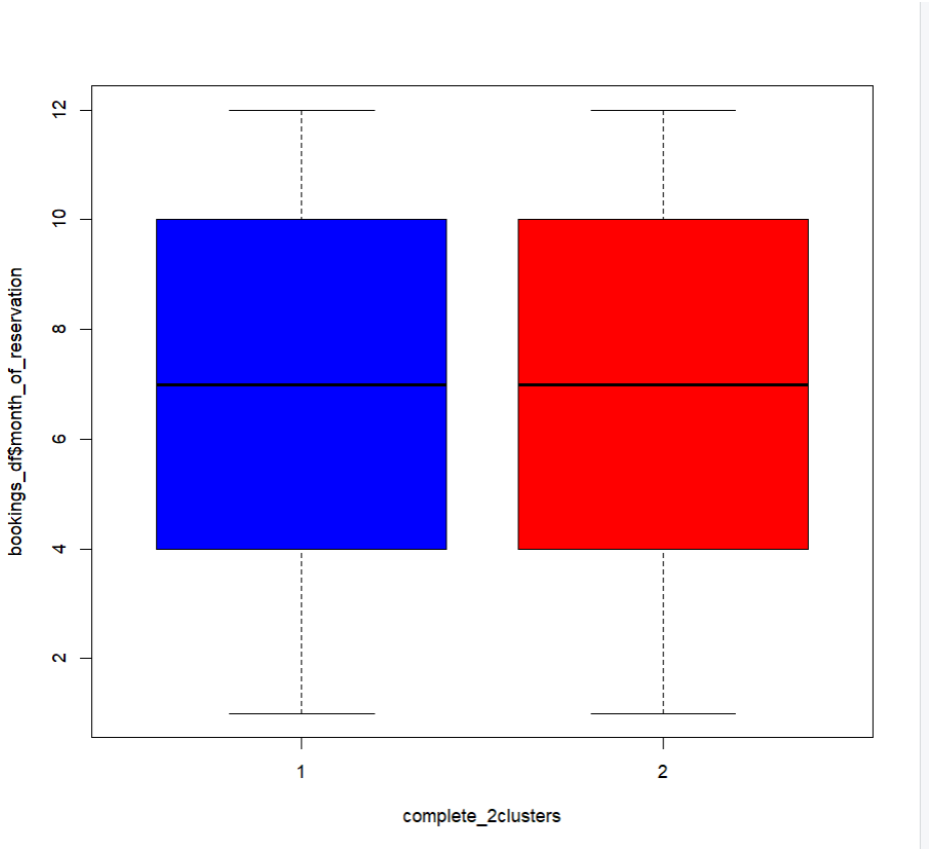


There are two sudden changes in the slope at point where clusters 2 and 4. I tried to cluster based on a k-medoids model with these parameters but the performance was worse when compared to Complet method. Especially in the case of the 4 clusters the negative sil. values were over 100.

Method	Avg. Silhouette Values	Number of negative Silhouette Values
K-medoids 2 clusters	0.17, 0.33	0
K-medoids 4 clusters	0.09, 0.36, 0.21,0.34	145

2.3) Variable selection

After finding the best method I could I plotted the most predictors against the clusters to find what variables hadn't difference between the 2 clusters to subtract them because they probably add noise. In the boxplot below we see that month does not add any information on the clustering so I will try to remove later ont.



Also I performed an ANOVA test and subtracted the variables number of children (p-value=0.06) as marginally insignificant and special requests (p-value = 0.66).

After removing these variables, I calculated again the Gower distance and both the Ward and Complete method improved significantly. Finally, I chose as the best model ward instead of complete method because it had more balanced clusters and less negative sil. values. Overall the final Ward linkage had

Method	Avg. Silhouette Values	Number of negative Silhouette Values	Observations per cluster
Ward (unscaled data)	0.38, 0.47	20	1272, 726