MD - D5
Hotel Bookings

Yingxin Chen
Max Estradé
Louis Forster
David Morais
Amelie Tolstorukova

Index

1. Motivation for work

This data set contains booking information from various hotels between 2015 and 2016. This data set includes information such as the type of hotel (resort hotel or city hotel) when the booking was made, length of stay, number of adults, children, and/or babies, and the number of available parking spaces, among other things. You can find a detailed description of each parameter on Kaggle.

We decided to analyze this dataset because we, a group of young adults, are interested in travelling. This analysis lets us learn interesting information about what makes a good hotel experience.

2. Data Source Presentation

The dataset used in this study comes from the public repository Kaggle, specifically from the dataset “Hotel Booking Demand” available at:
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

It contains detailed booking information for city and resort hotels between 2015 and 2016, including variables such as hotel type, booking dates, length of stay, number of guests, type of meal plan, and customer type, among others. For this project, we used a subset of 5,000 records corresponding to bookings from July 2015 to April 2016 that was extracted from the original dataset to ensure manageable data volume while maintaining representativeness of the overall structure.

3. Metadata

Each row of our data matrix represents a booking at a particular hotel and contains information detailing the reservation. In the original dataset, each row had 32 variables, 15 of which were numerical, two binary, and 12 qualitative variables.

3.1. Metadata Table

Variable	Short variable name	Modalities	Short Mod Name	Meaning	Type	Measuring units	Missing code	Range	Role
hotel				Hotel type		H1 = Resort Hotel H2 = City Hotel			
is_canceled	canceled			Value indicating if the booking was canceled (1) or not (0)	Bool				
lead_time				Nº of days that elapsed between the entering date of the booking into the PMS and the arrival date	Num	Days		[0,626]	
arrival_date_year	year			Year of arrival date	Num	Year		[2015,2017]	
arrival_date_month	month			Month of arrival date	Quali				
arrival_date_week_	week			Week number for the	Num	Week number		[1,53]	

number				arrival date					
arrival_date_day_of_month	day			Day of arrival date	Num	Day number		[1,31]	
stays_in_weekend_nights	weekend_nights			Nº of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.	Num	Nights (days)		[0,8]	
stays_in_week_nights	week_nights			Nº of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.	Num	Nights (days)		[0,19]	
adults				Nº of adults	Num	People		[0,4]	
children				Nº of children	Num	People		[0,3]	
babies				Nº of babies	Num	People		[0,2]	
meal				Type of meal booked.	Quali		Undef		
		No meal package	SC						

		Bed & Breakfast	BB						
		Half board	HB	Breakfast and one other meal					
		Full board	FB	Breakfast, lunch, and dinner					
country				Country of origin.	Quali	Categories are represented in the ISO 3155–3:2013 format.			
market_segment	market_seg			Market segment designation.	Quali				
		Travel agents	TA						
		Tour operators	TO						
distribution_channel	channel			Booking distribution channel.	Quali				
		Travel	TA						

		agents							
		Tour operators	TO						
is_repeated_guest	reapeated			Value indicating if the booking name was from a repeated guest (1) or not (0)	Bool				
previous_cancellations	pre_cancel			Nº of previous bookings that the customer canceled before the current booking	Num			[0,26]	
previous_bookings_not_canceled	pre_bnc			Nº of previous bookings not cancelled by the customer before the current booking	Num			[0,63]	
reserved_room_type	rroom_type			Code of room type reserved.	Quali	Code is presented instead of designation for anonymity reasons.			
assigned_room_type	assroom_t			Code for the type of room	Quali	Sometimes the assigned			

	ype			assigned to the booking.		room type differs from the reserved room type due to hotel operational reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.			
booking_changes	changes			Nº of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation	Num			[0,15]	
deposit_type				Indication of whether the customer deposited to guarantee the booking.	Quali				
		No Deposit	ND	No deposit was made					

3.2. Final Scope

The final dataset used for the analysis consists of 4,867 hotel booking records, each representing a single reservation between July 2015 and April 2016. After preprocessing, only valid and complete records were retained.

Regarding the variables, we kept 31 variables for analysis after removing the column “company,” which contained more than 90% missing values. All other missing or erroneous entries were corrected or removed as part of the preprocessing step.

Therefore, the final scope includes numerical and categorical variables relevant to hotel booking behavior, ensuring that the dataset remains representative while being clean, consistent, and suitable for multivariate analysis such as PCA and clustering.

4. Data Mining Process

Data Visualization and Selection: After we obtained our dataset, we decided to reduce it to 5000 rows to fit the given scope. Then, we visualized every variable in the dataset.

Preprocessing: To prepare the data for mining, we had to clean and normalize our dataset.

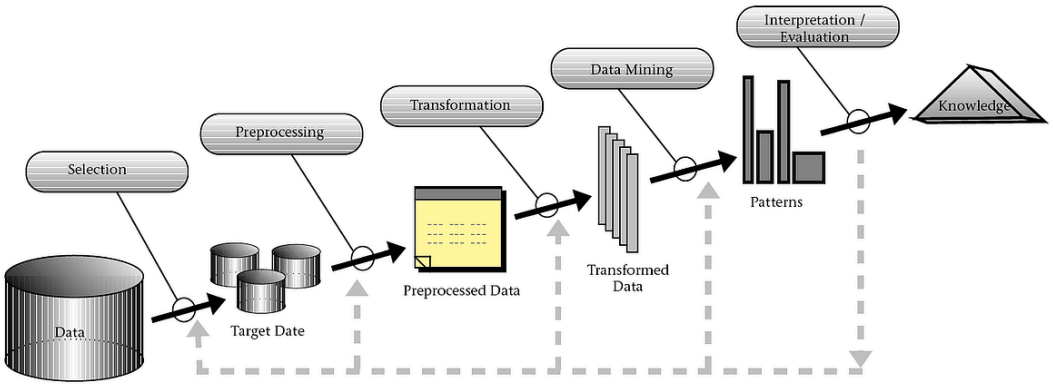
Data Mining: After preprocessing, we applied different data mining techniques to uncover hidden patterns and relationships among the variables.

- **Descriptive Statistical Analysis:** To explore the main characteristics of the dataset, both univariate and bivariate analyses were carried out.

- **Principal Component Analysis (PCA):** PCA was applied to the numerical variables to identify underlying structures and reduce dimensionality.

- **Hierarchical clustering:** The clustering gave us a dendrogram to determine the optimal number of clusters, grouping similar bookings according to their characteristics.

- **Profiling:** Once we had our clusters, we performed profiling to describe and compare them



5. Preprocessing

To ensure that our dataset was clean, accurate, and ready for further analysis, we ran many R scripts on it. These scripts were given to us in the laboratory sessions, and then they were adapted to fit our concrete dataset.

5.1. Visualization & Basic Descriptive Statistics

Initial visual exploration included histograms, boxplots, and barplots to understand distributions and identify anomalies. This phase provided insight into patterns such as average stay length, booking lead time, and common customer types.

5.2. Filtering

We removed irrelevant or redundant information to simplify subsequent analyses. For example, the variable “company” was excluded due to over 90% missing data, and records with incomplete key fields were filtered out.

5.3. Outlier Detection & Correction

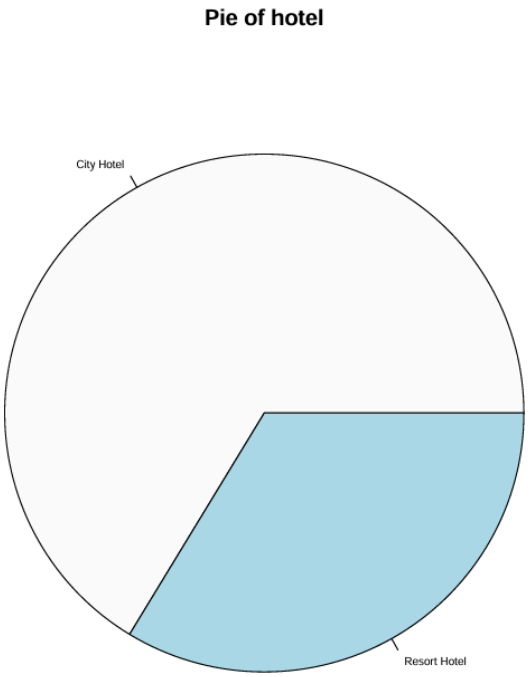
We identified potential outliers through boxplots and statistical summaries. We inspected extreme or erroneous values and corrected them when appropriate.

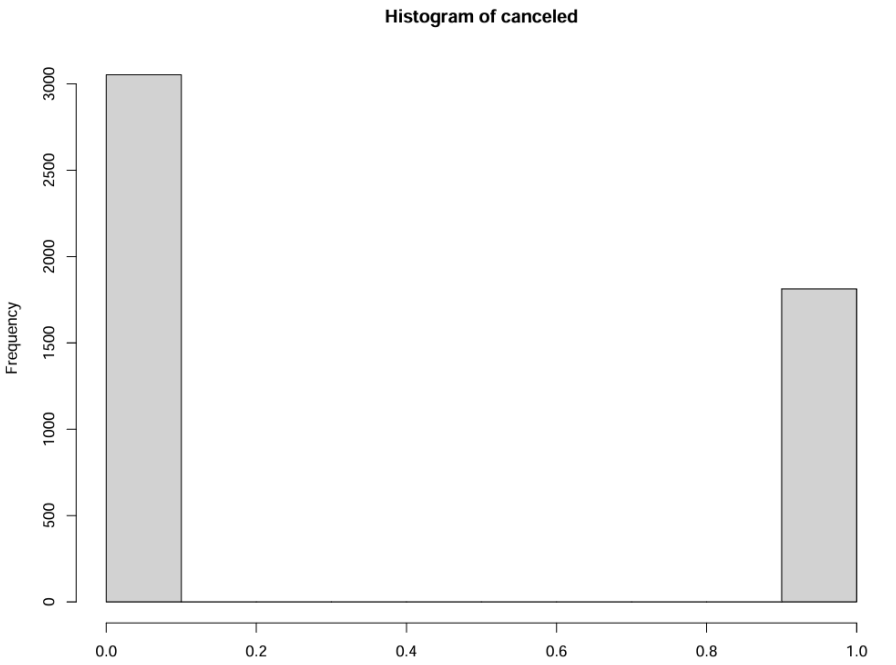
5.4. Missing Data Treatment

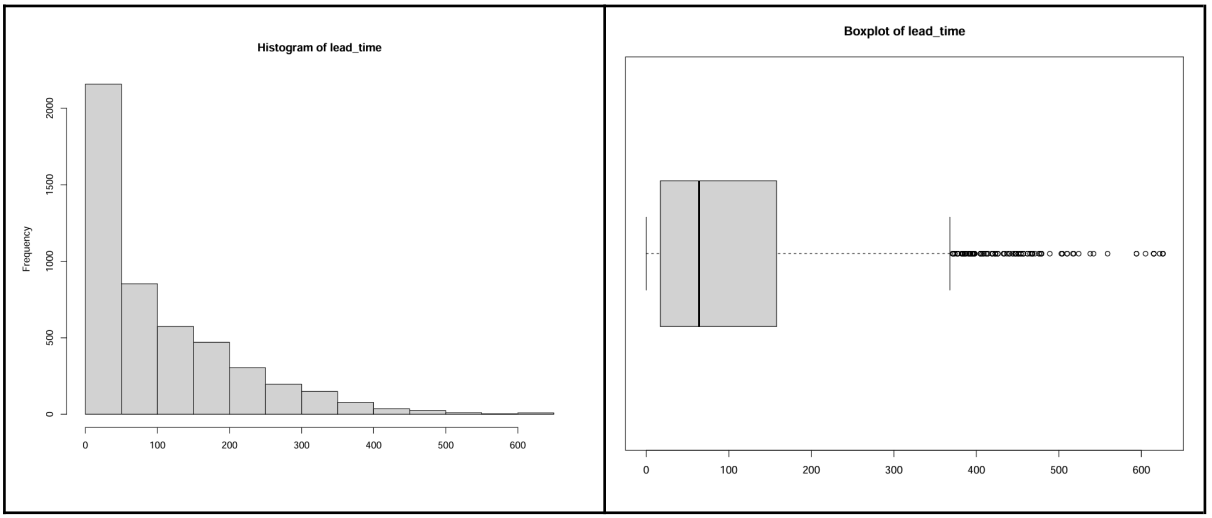
Missing values were handled using imputation methods. The *K-Nearest Neighbors (KNN, k=5)* algorithm was applied to fill missing entries in *“meal”*, *“country”, and “agent”*, ensuring minimal distortion of original distributions. We also removed the “*company”* column because, as we mentioned before, the amount of missing values was 90%.

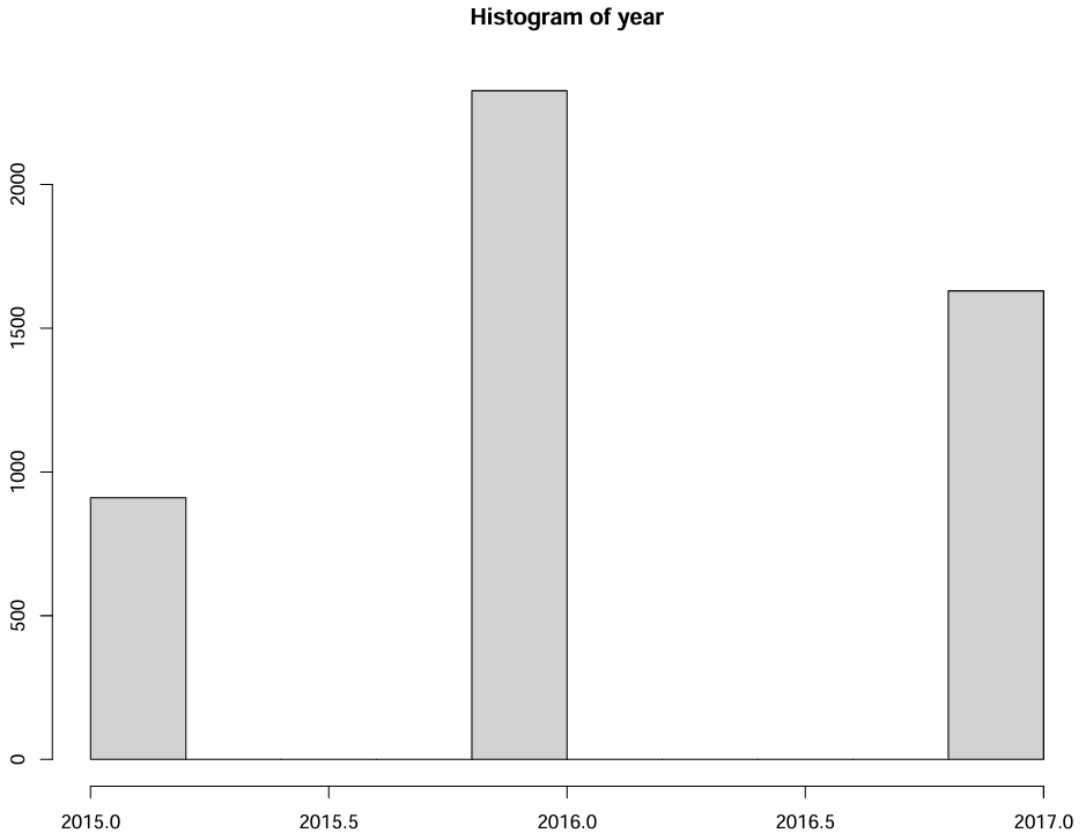
6. Statistical Descriptive Analysis

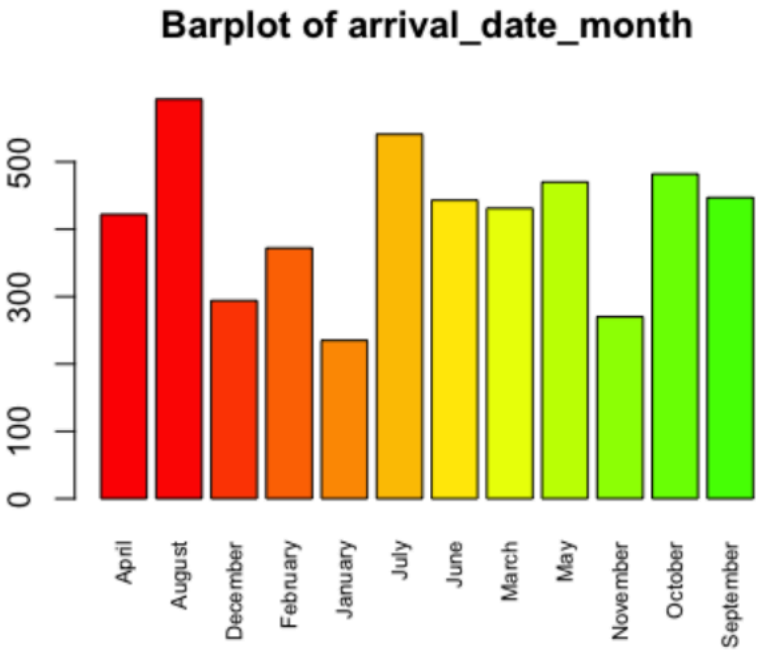
6.1. Univariate

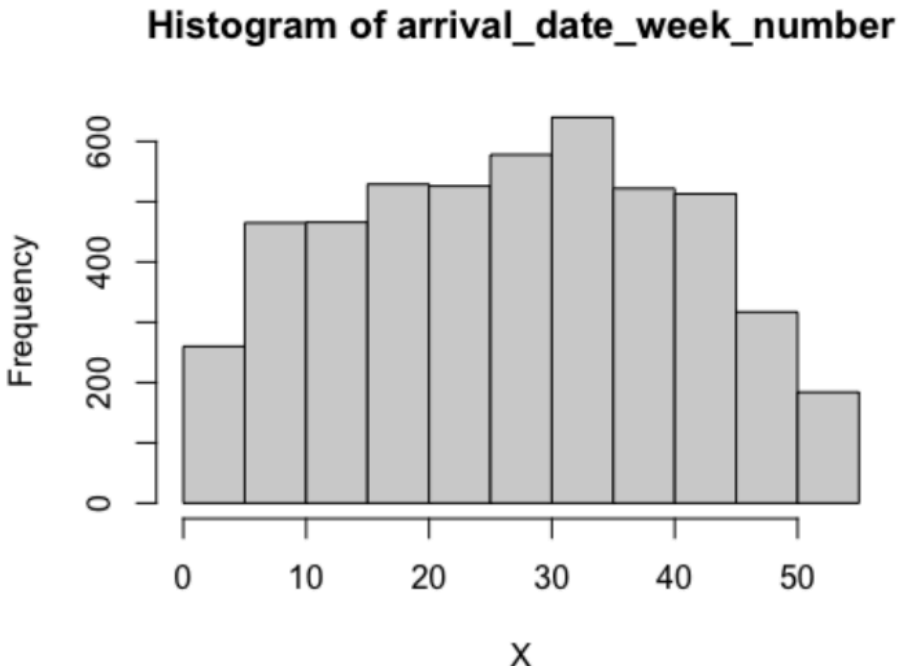


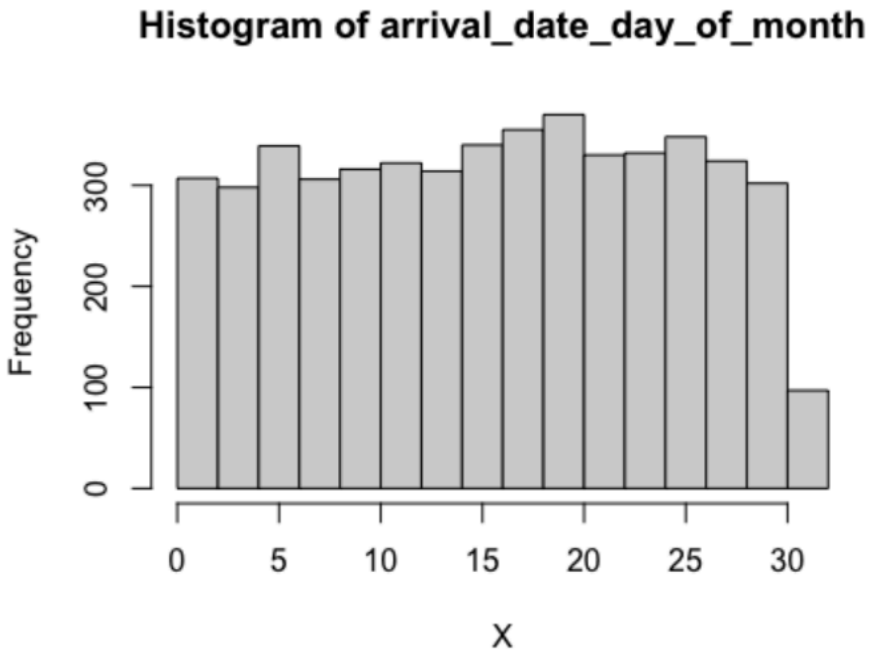

Histogram of canceled
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0

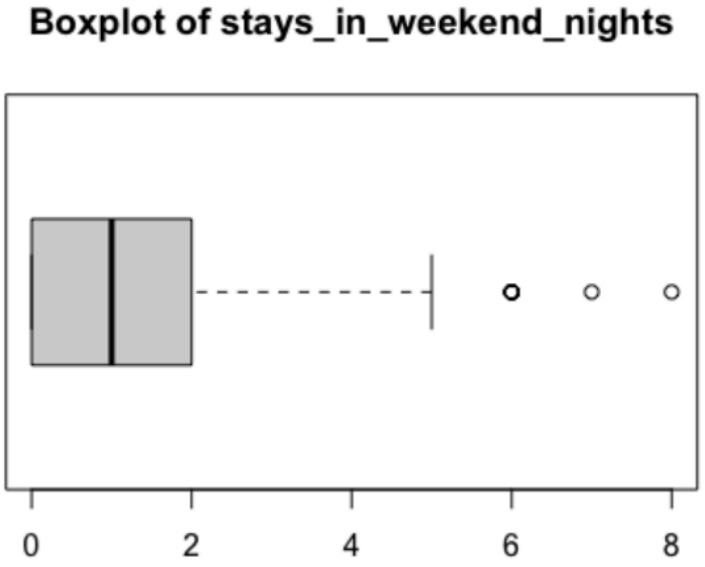

Histogram of lead_time
Frequency
0
500
1000
1500
2000
0
100
200
300
400
500
600
Boxplot of lead_time
0
100
200
300
400
500
600

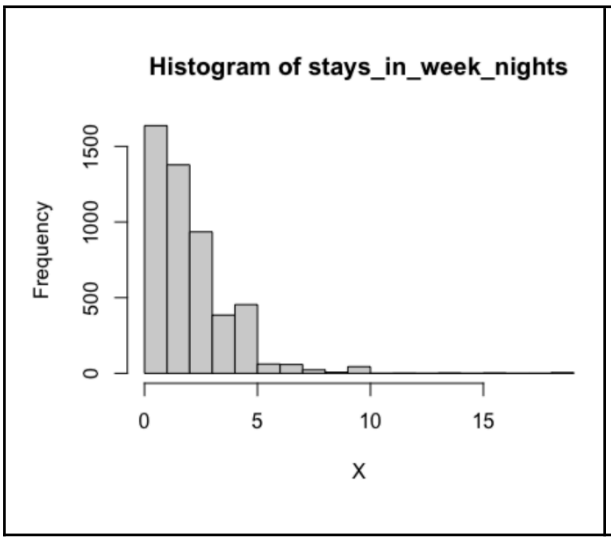

Histogram of year
0
500
1000
1500
2000
2015.0
2015.5
2016.0
2016.5
2017.0

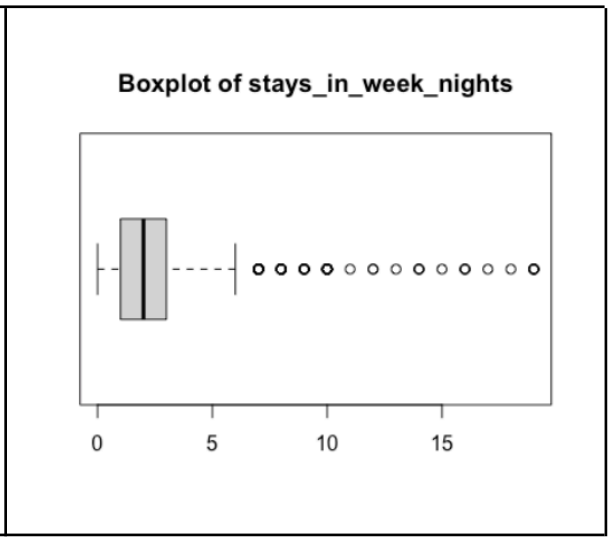

Barplot of arrival_date_month
0
100
300
500
April
August
December
February
January
July
June
March
May
November
October
September

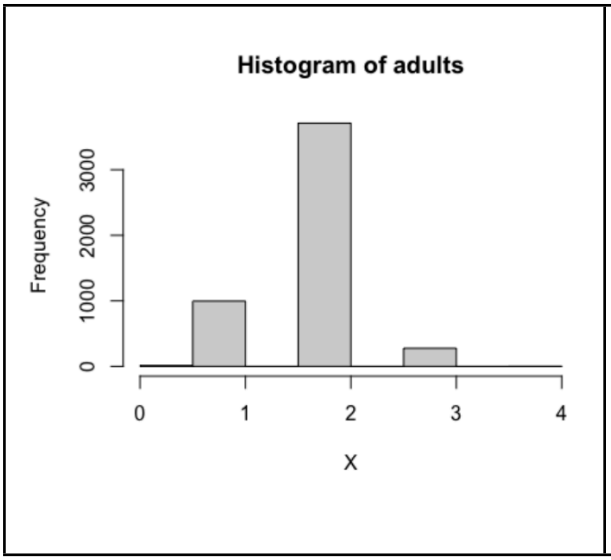

Histogram of arrival_date_week_number
Frequency
0
200
400
600
0
10
20
30
40
50
X

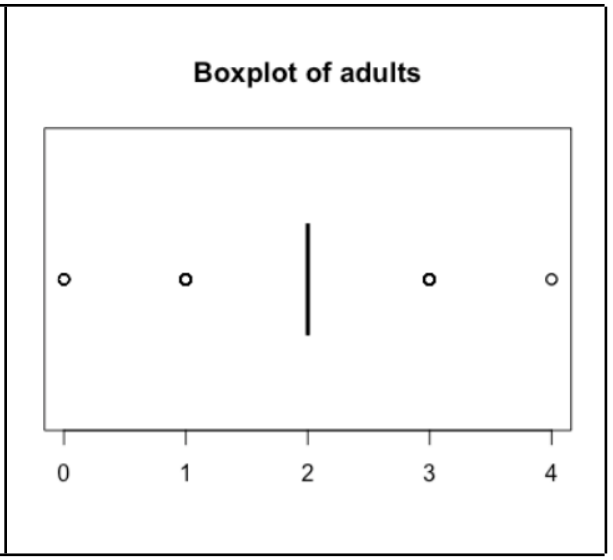

Histogram of arrival_date_day_of_month
Frequency
0
100
200
300
0
5
10
15
20
25
30
X

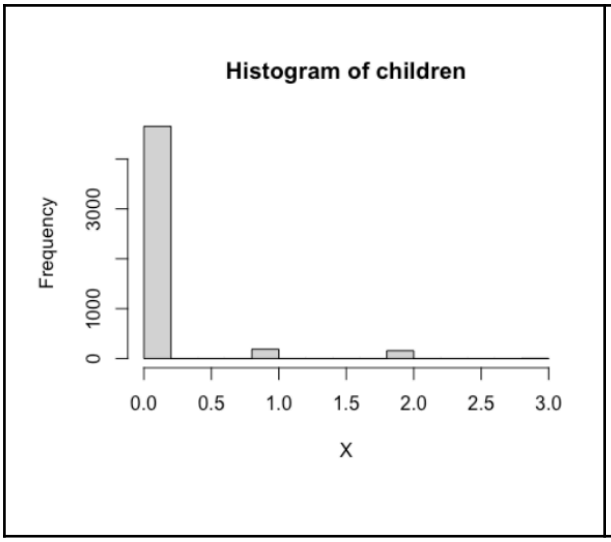

Boxplot of stays_in_weekend_nights
0
2
4
6
8

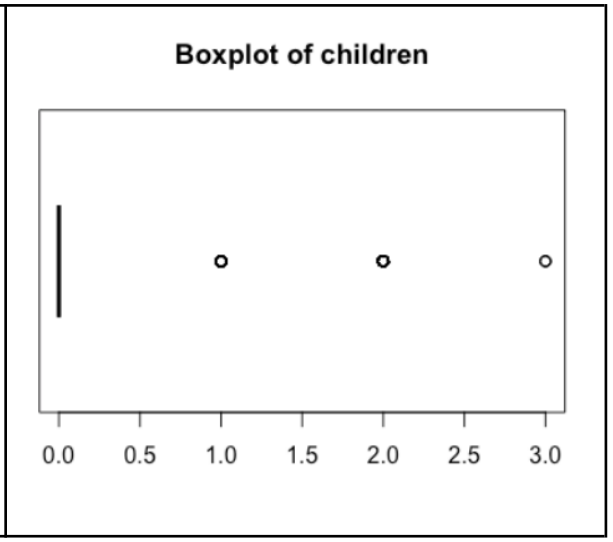

Histogram of stays_in_week_nights
Frequency
0
500
1000
1500
0
5
10
15
X

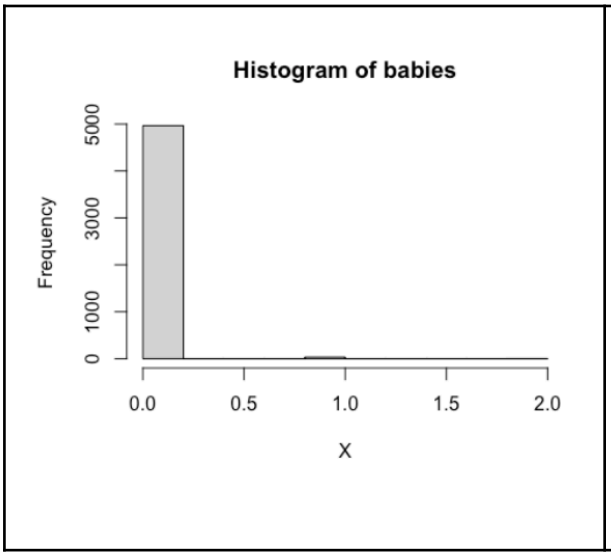

Boxplot of stays_in_week_nights
0
5
10
15

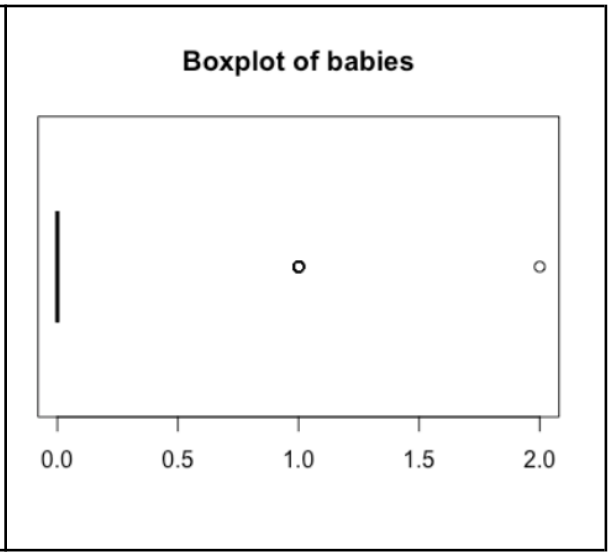

Histogram of adults
Frequency
0
1000
2000
3000
0
1
2
3
4
X


Boxplot of adults
0
1
2
3
4

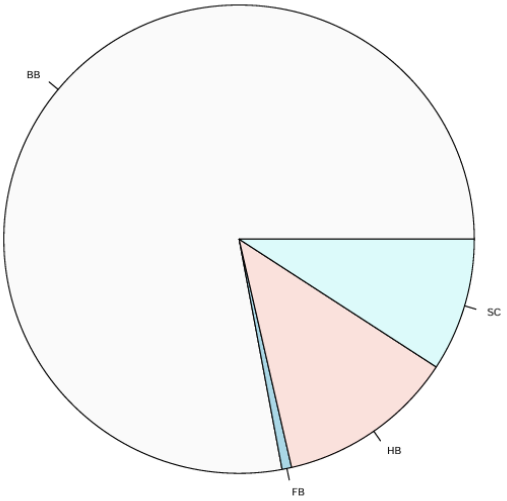

Histogram of children
Frequency
0
1000
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
X


Boxplot of children
0.0
0.5
1.0
1.5
2.0
2.5
3.0

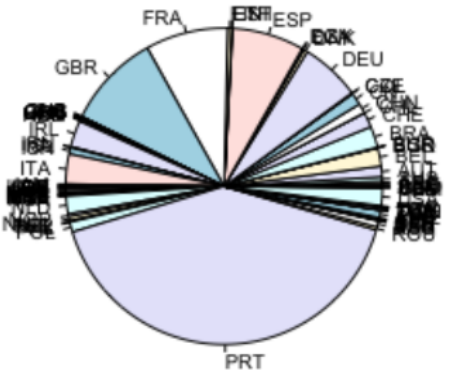

Histogram of babies
Frequency
0
1000
3000
5000
0.0
0.5
1.0
1.5
2.0
X


Boxplot of babies
0.0
0.5
1.0
1.5
2.0

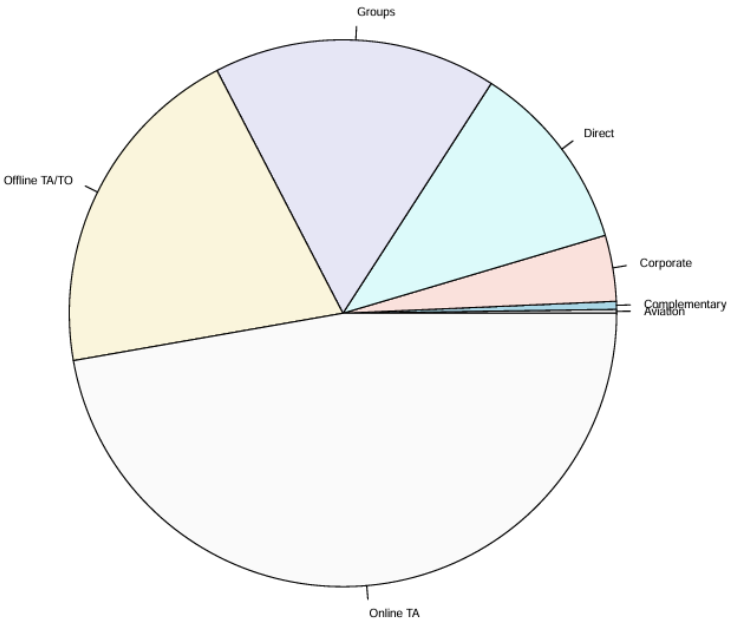
Pie of meal



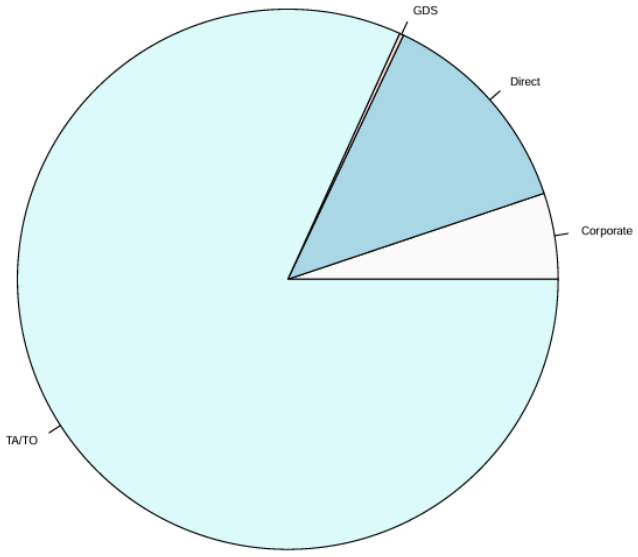
Pie of country

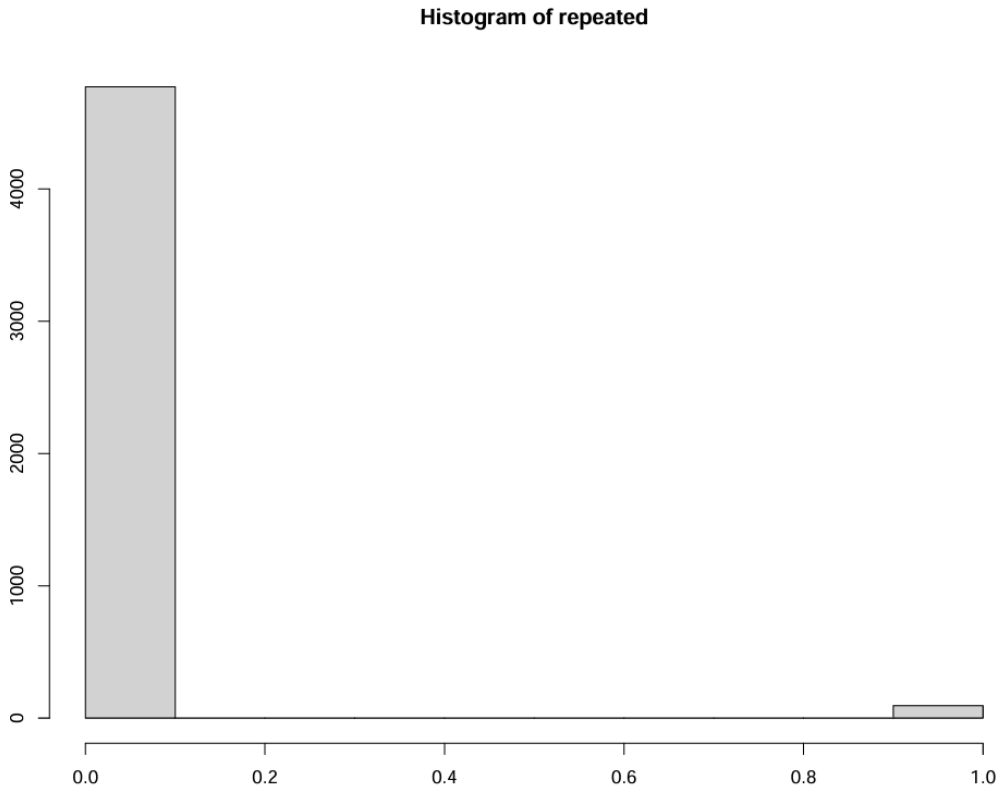


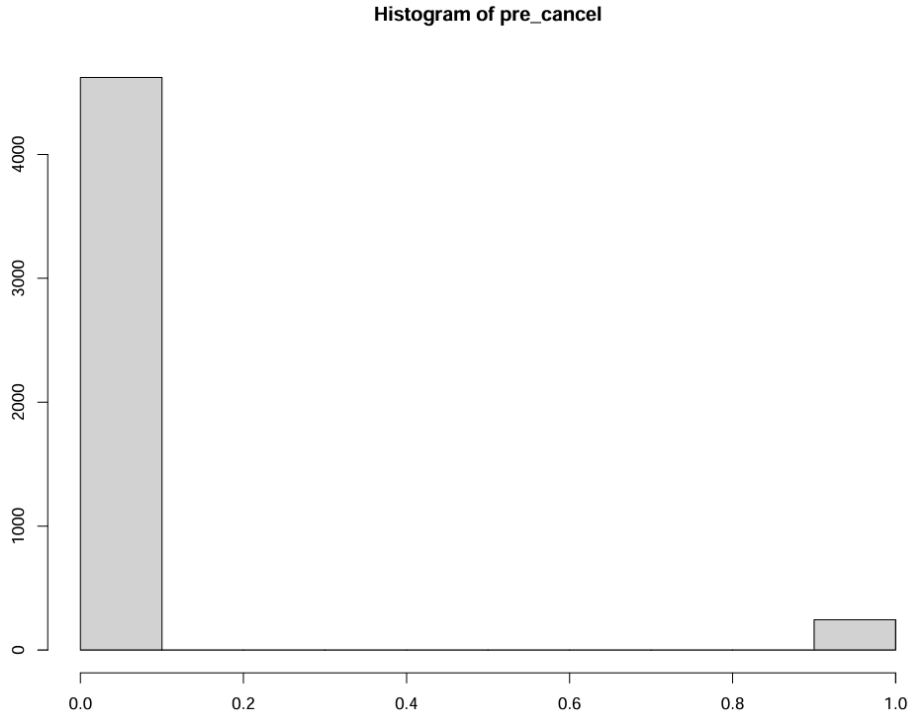
Pie of market_seg

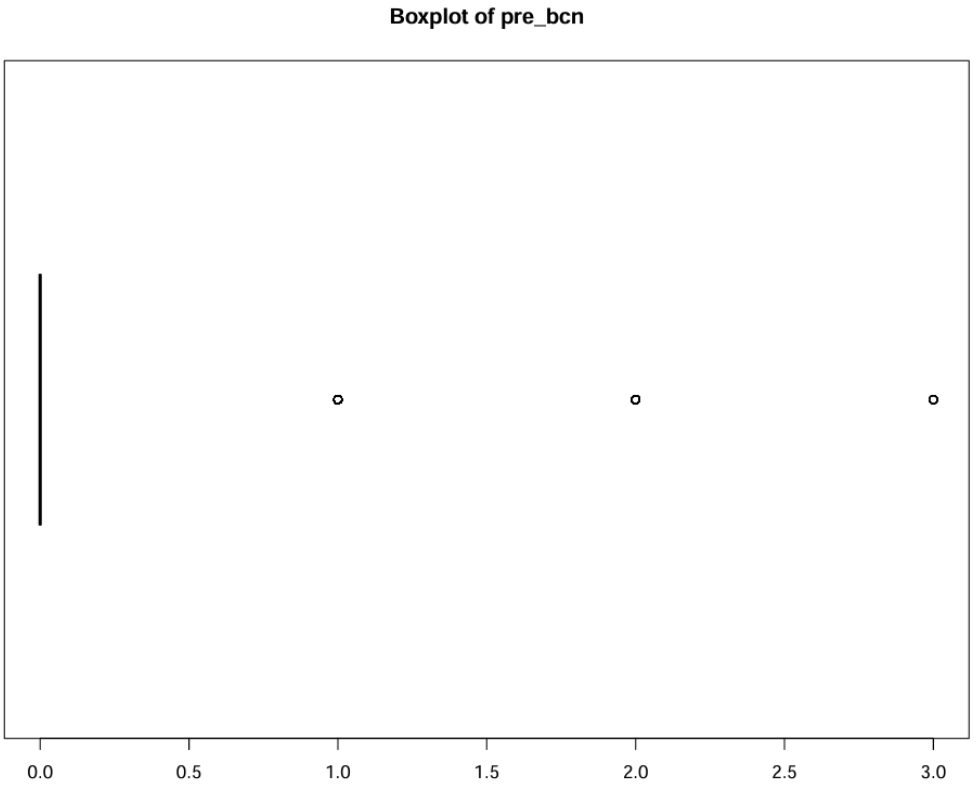


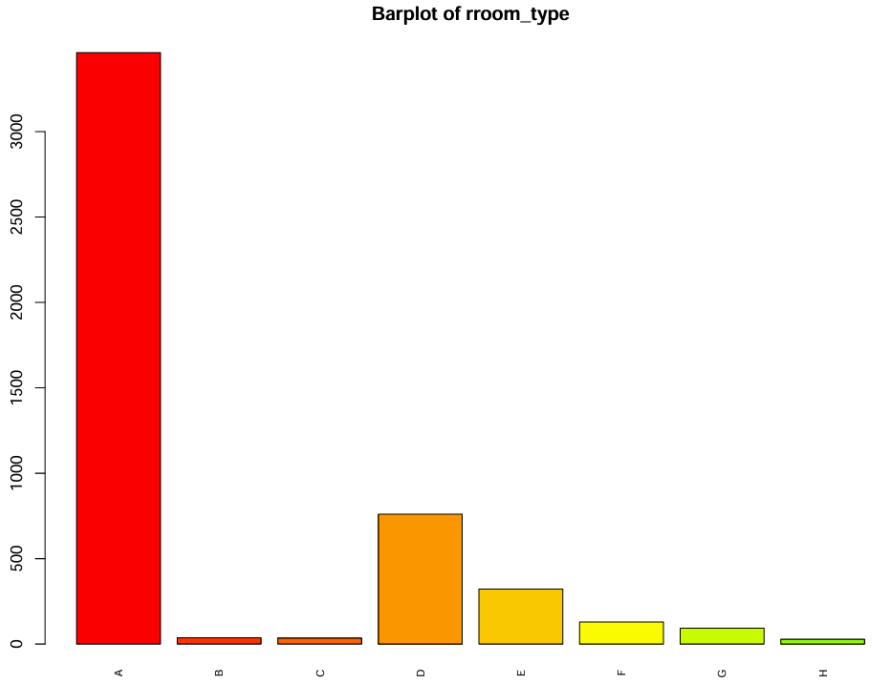
Pie of channel

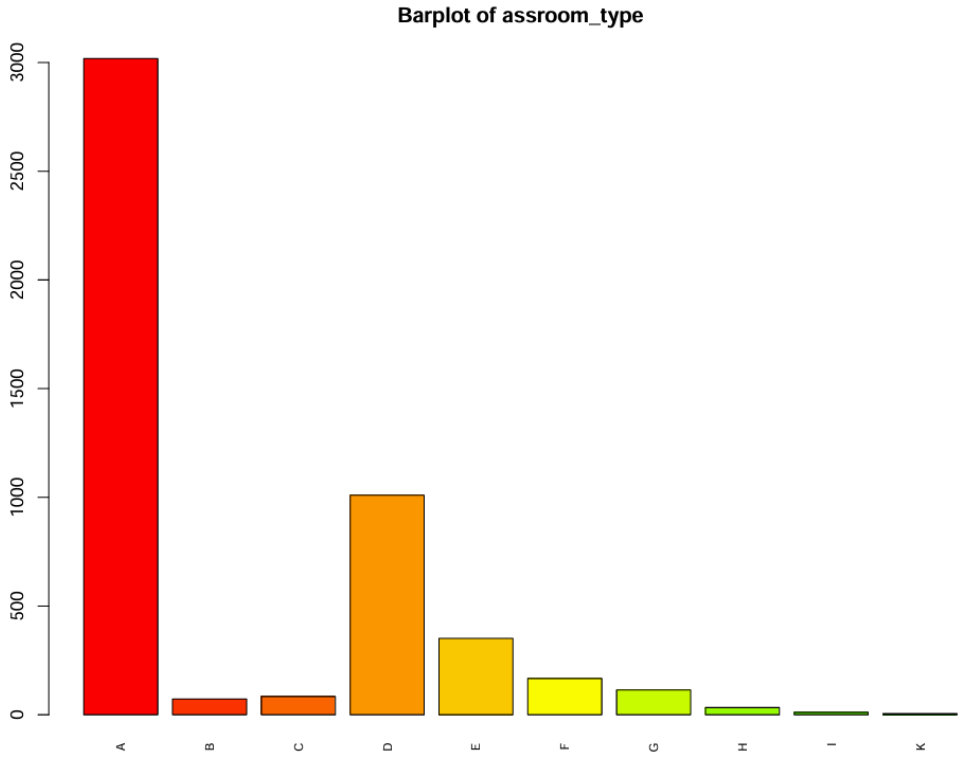


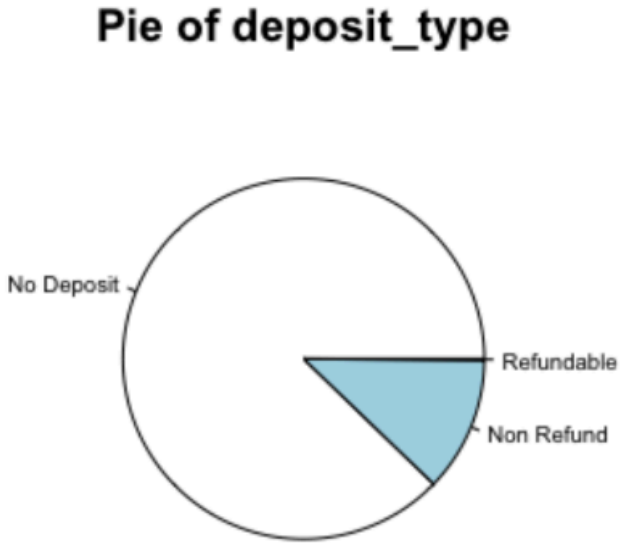

Histogram of repeated
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0

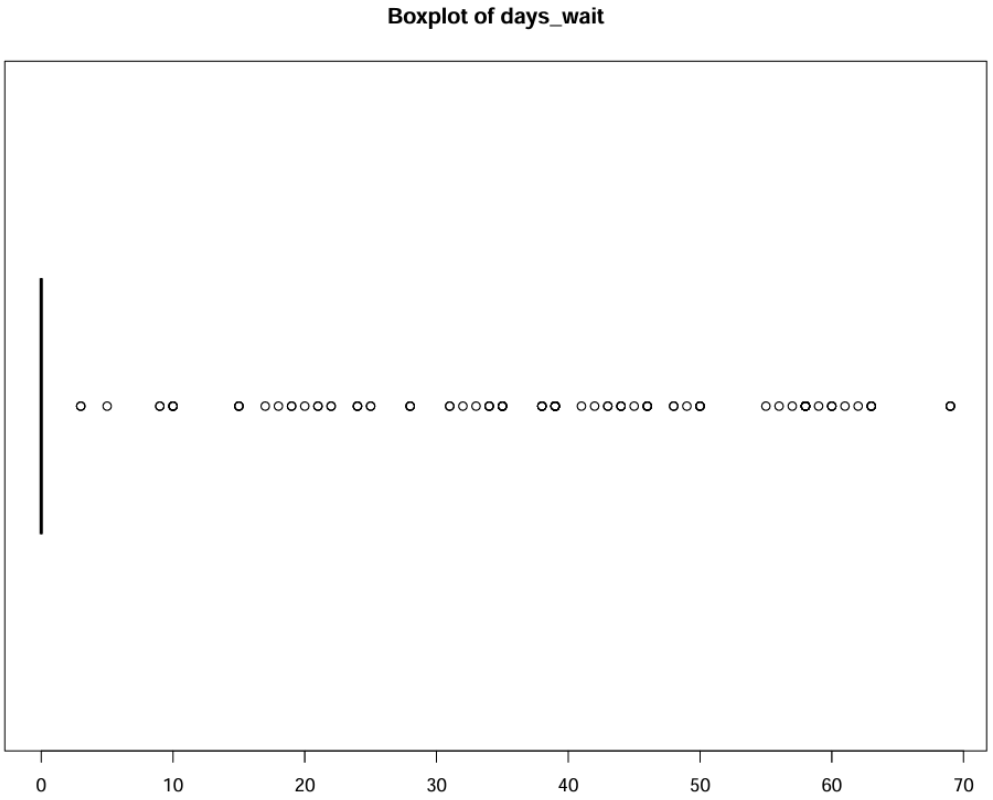

Histogram of pre_cancel
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0

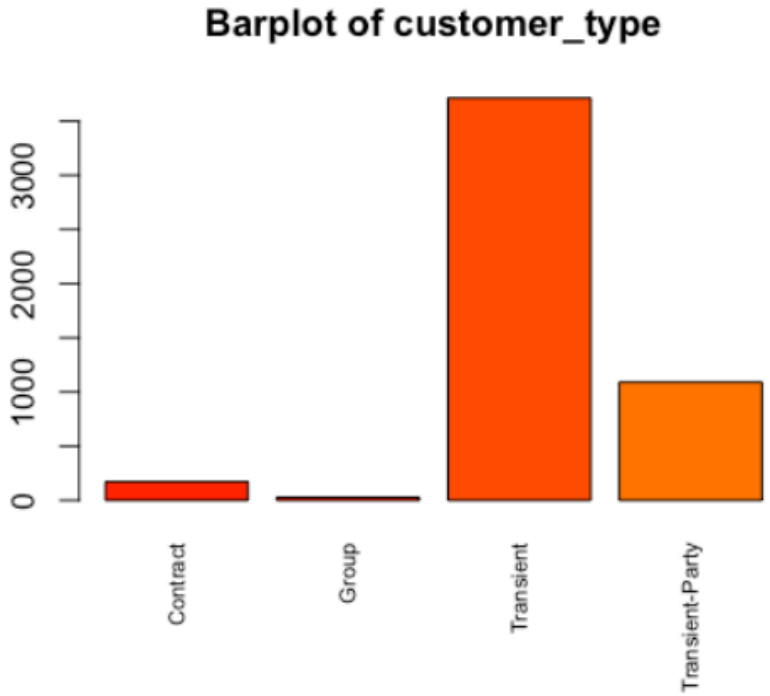

Boxplot of pre_bcn
0.0
0.5
1.0
1.5
2.0
2.5
3.0

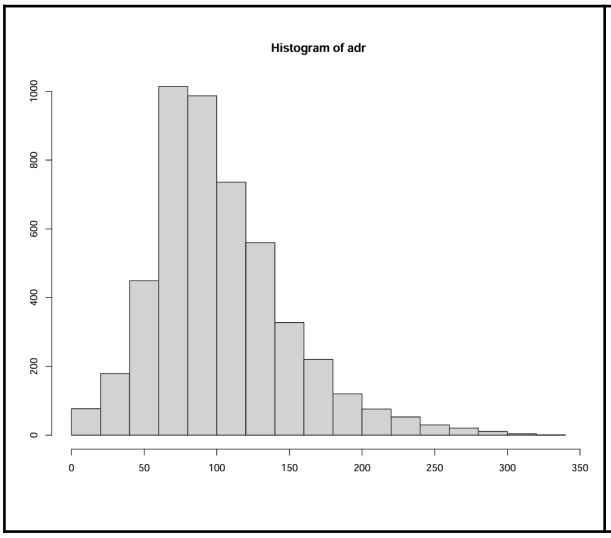

Barplot of rroom_type
0
500
1000
1500
2000
2500
3000
A
B
C
D
E
F
G
H

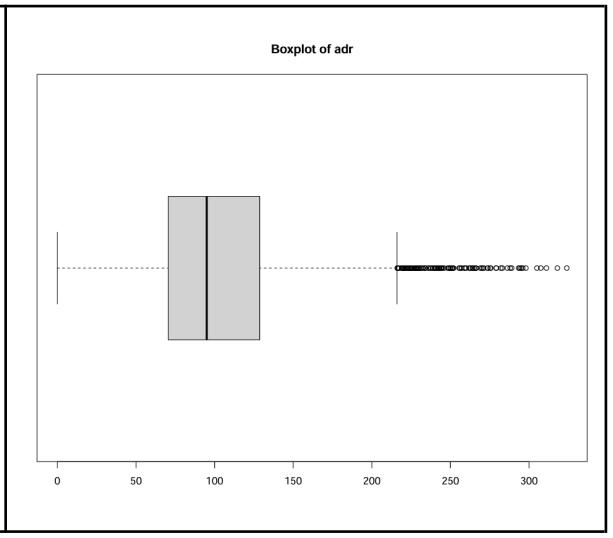

Barplot of assroom_type
3000
2500
2000
1500
1000
500
0
A
B
C
D
E
F
G
H
I
K

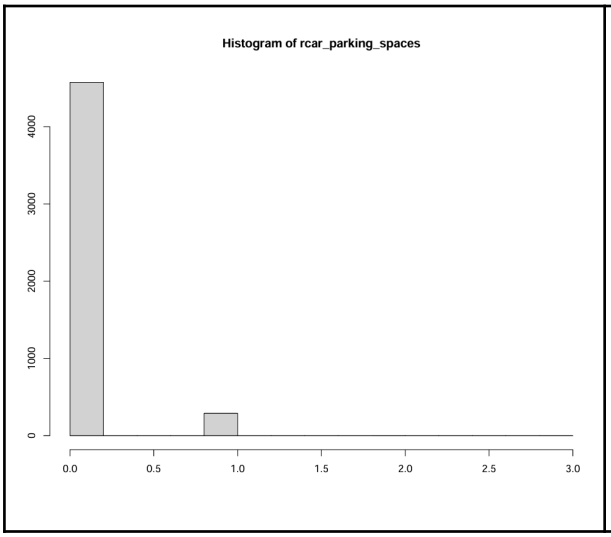

Pie of deposit_type
No Deposit
Refundable
Non Refund

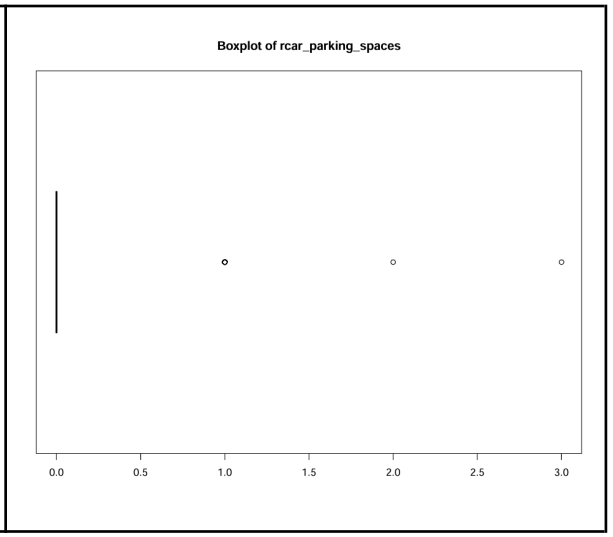

Boxplot of days_wait
0
10
20
30
40
50
60
70

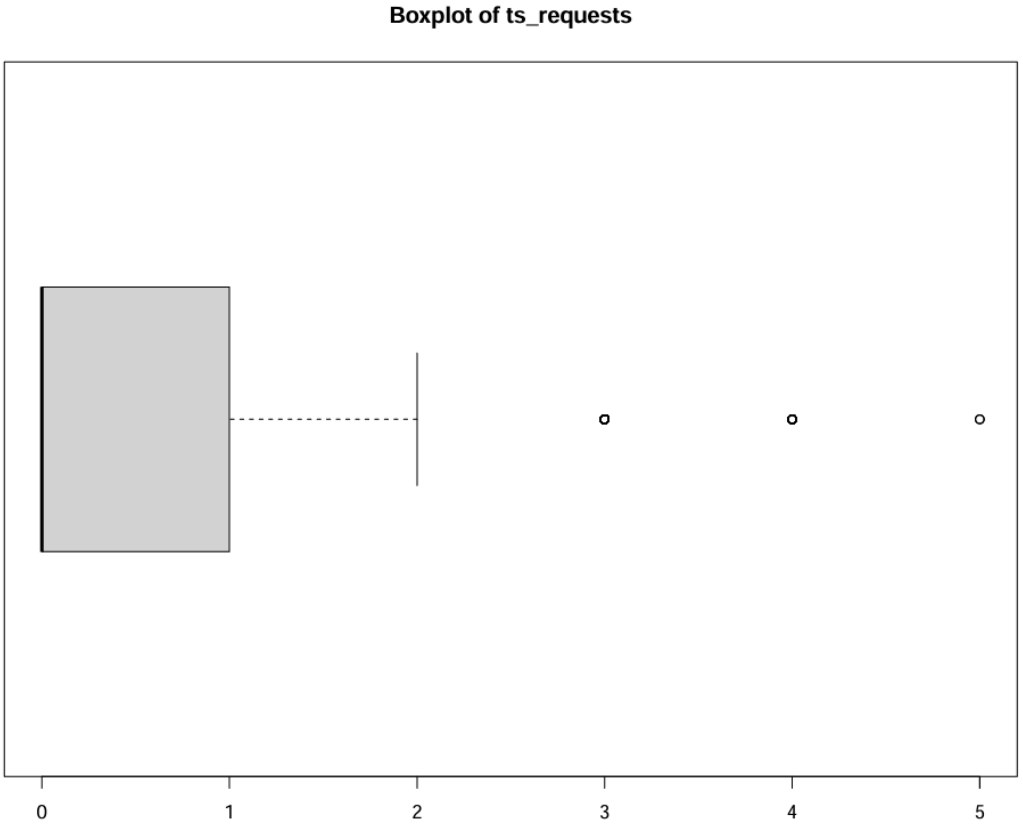

Barplot of customer_type
0
1000
2000
3000
Contract
Group
Transient
Transient-Party

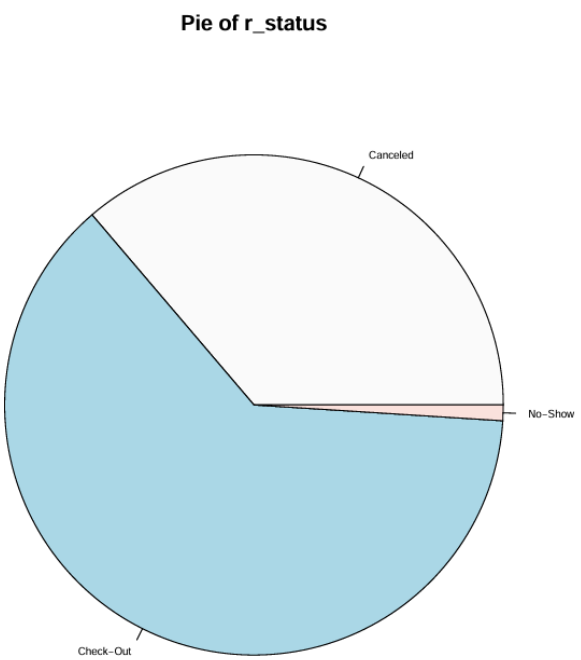

Histogram of adr
0
200
400
600
800
1000
0
50
100
150
200
250
300
350

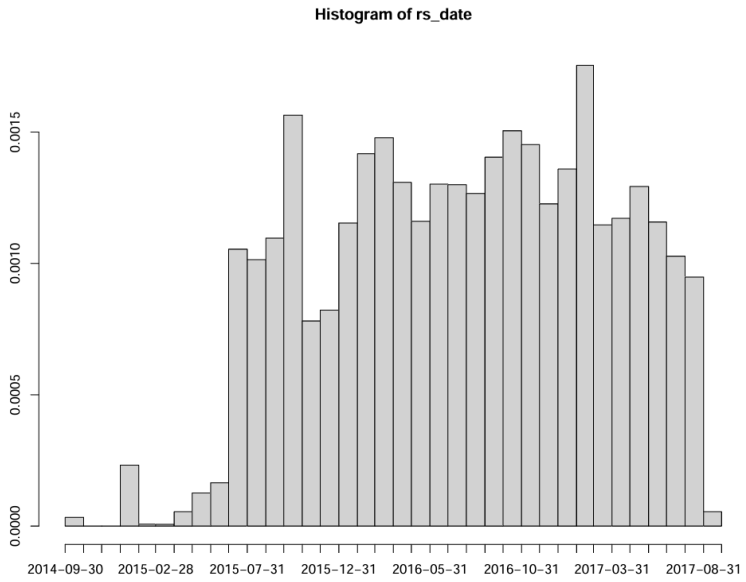

Boxplot of adr
0
50
100
150
200
250
300


Histogram of rcar_parking_spaces
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0

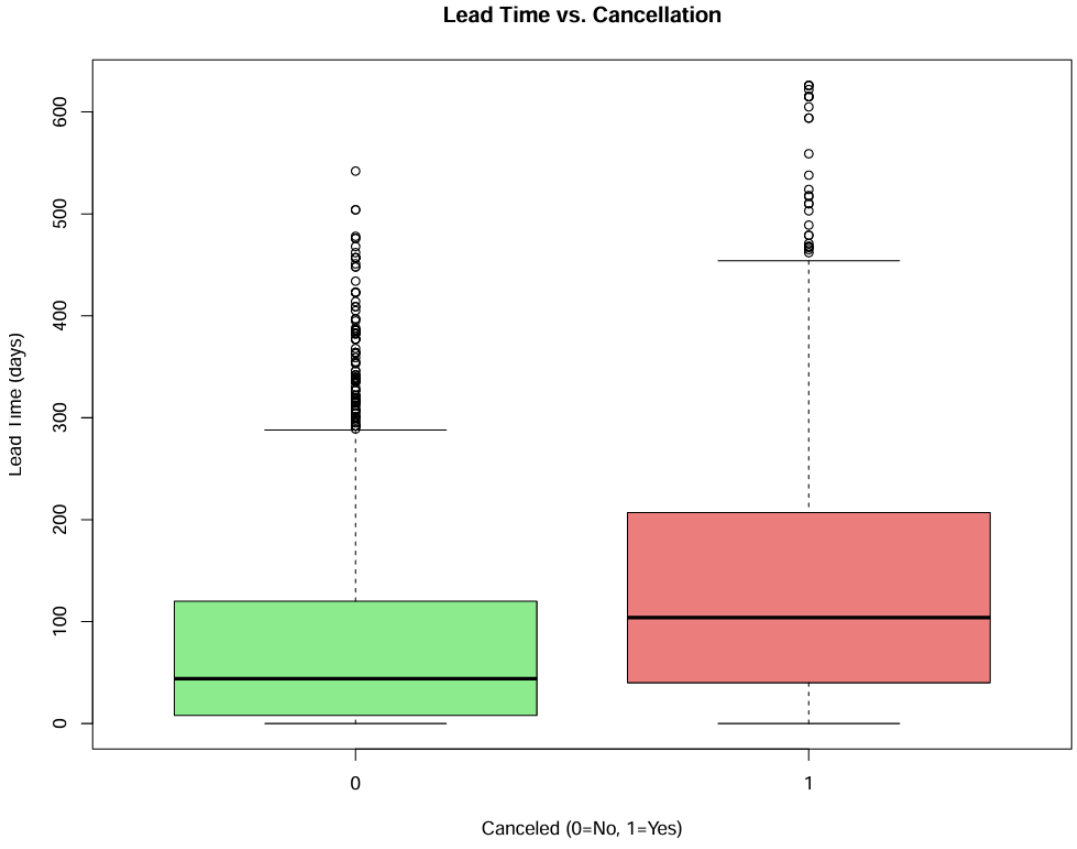

Boxplot of rcar_parking_spaces
0.0
0.5
1.0
1.5
2.0
2.5
3.0

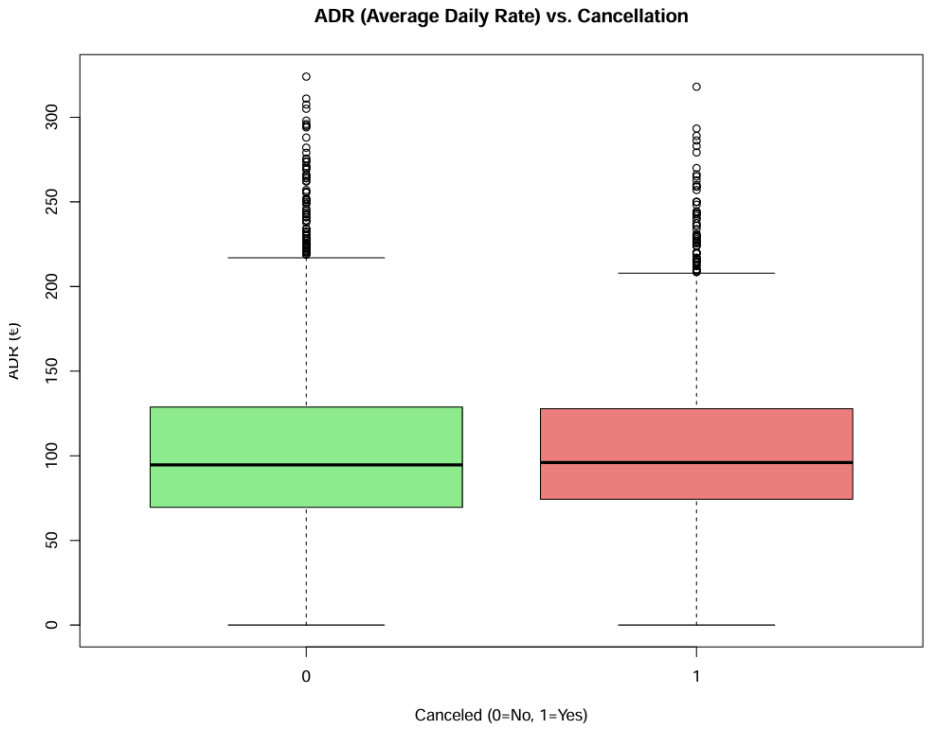

Boxplot of ts_requests
0
1
2
3
4
5

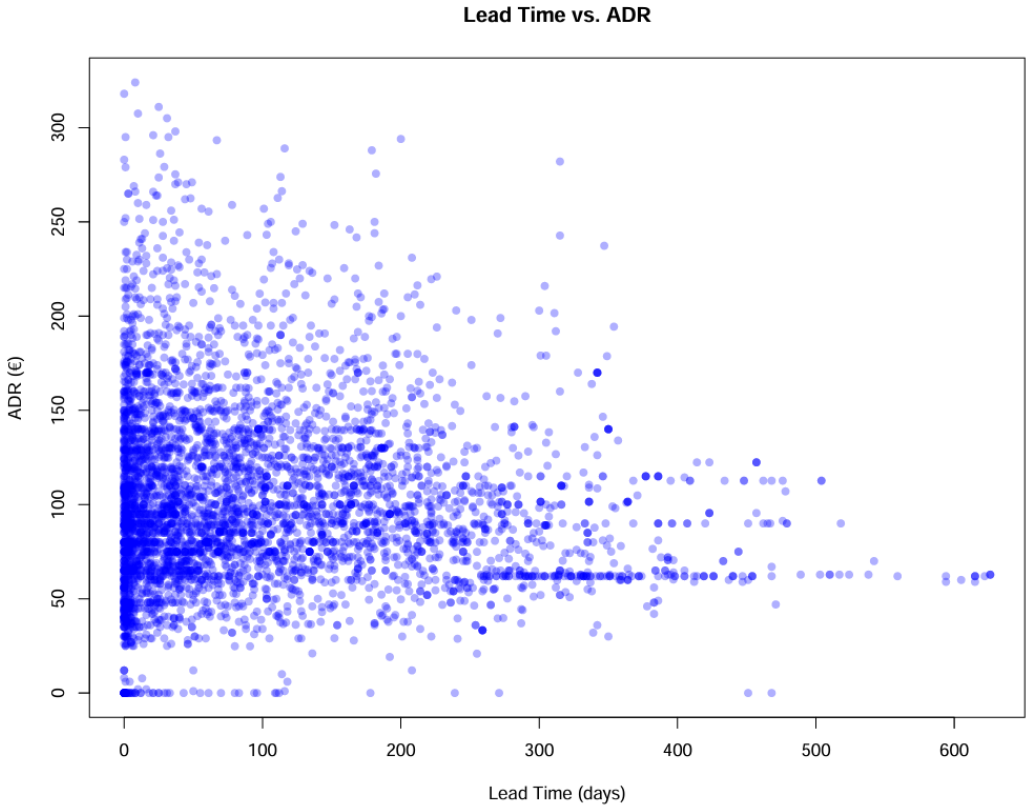

Pie of r_status
Canceled
No-Show
Check-Out

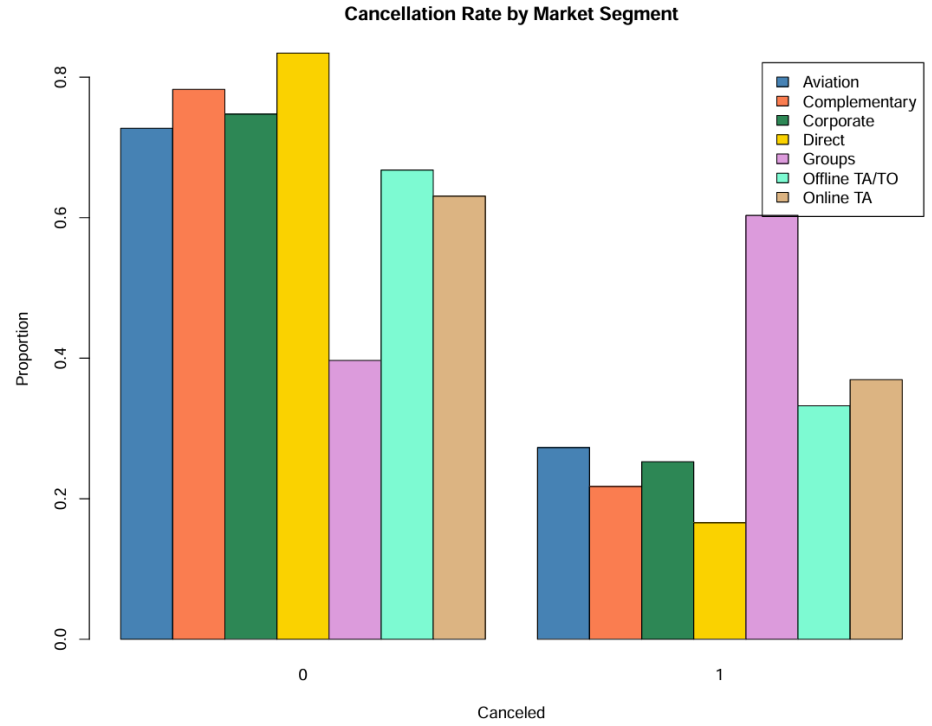


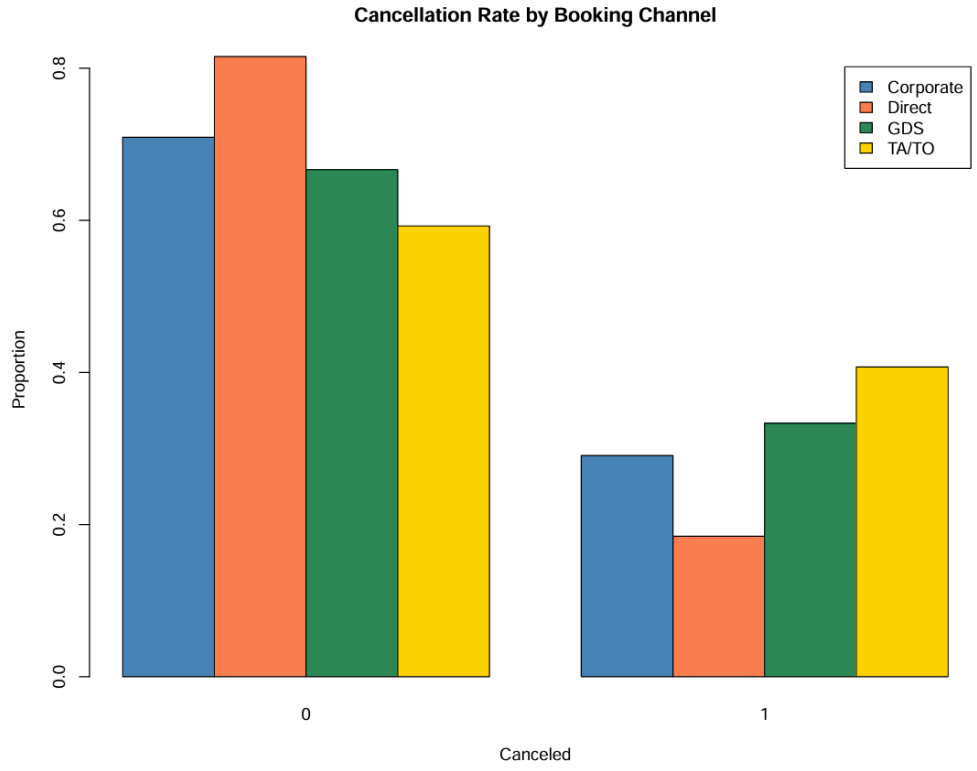
6.2. Bivariate

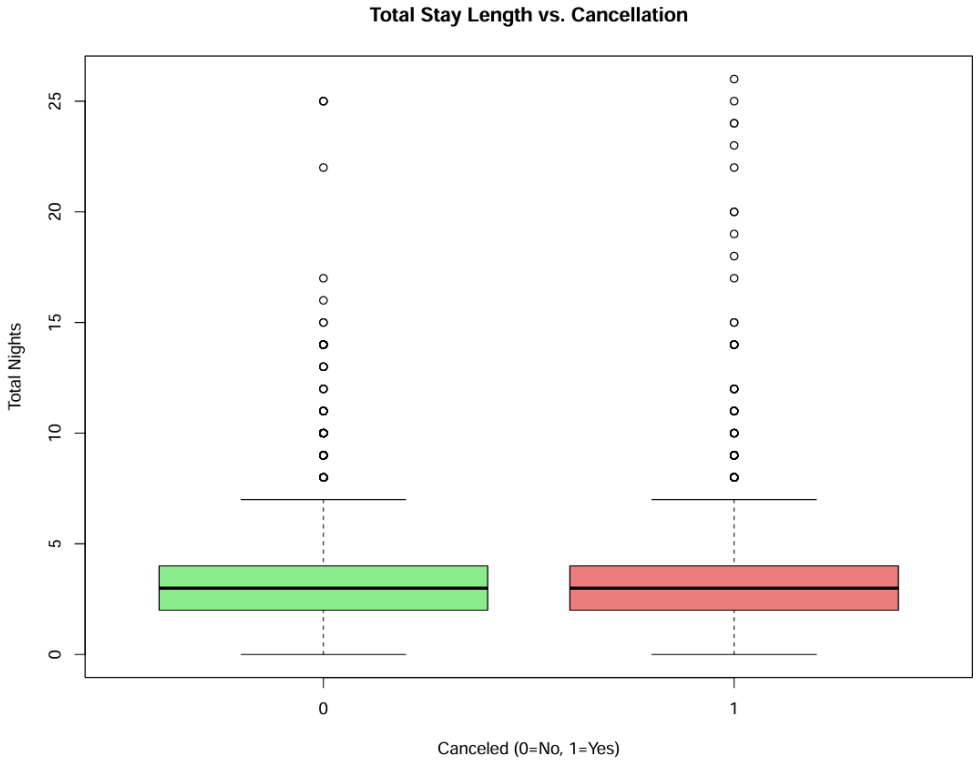


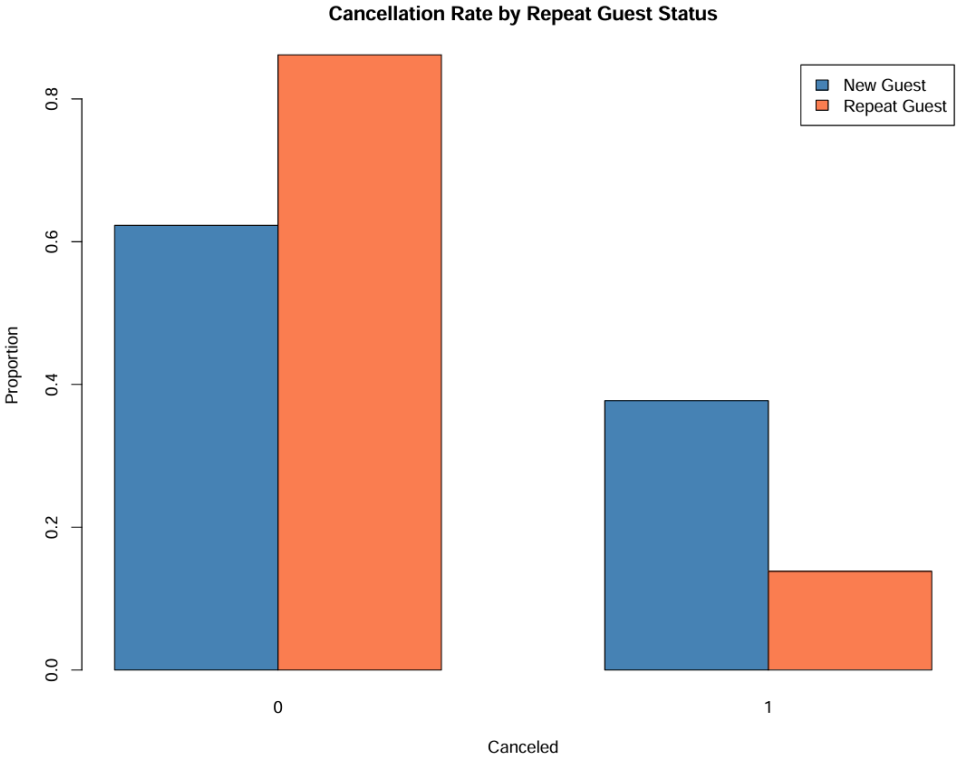

ADR (Average Daily Rate) vs. Cancellation
ADR (€)
0
50
100
150
200
250
300
0
1
Canceled (0=No, 1=Yes)

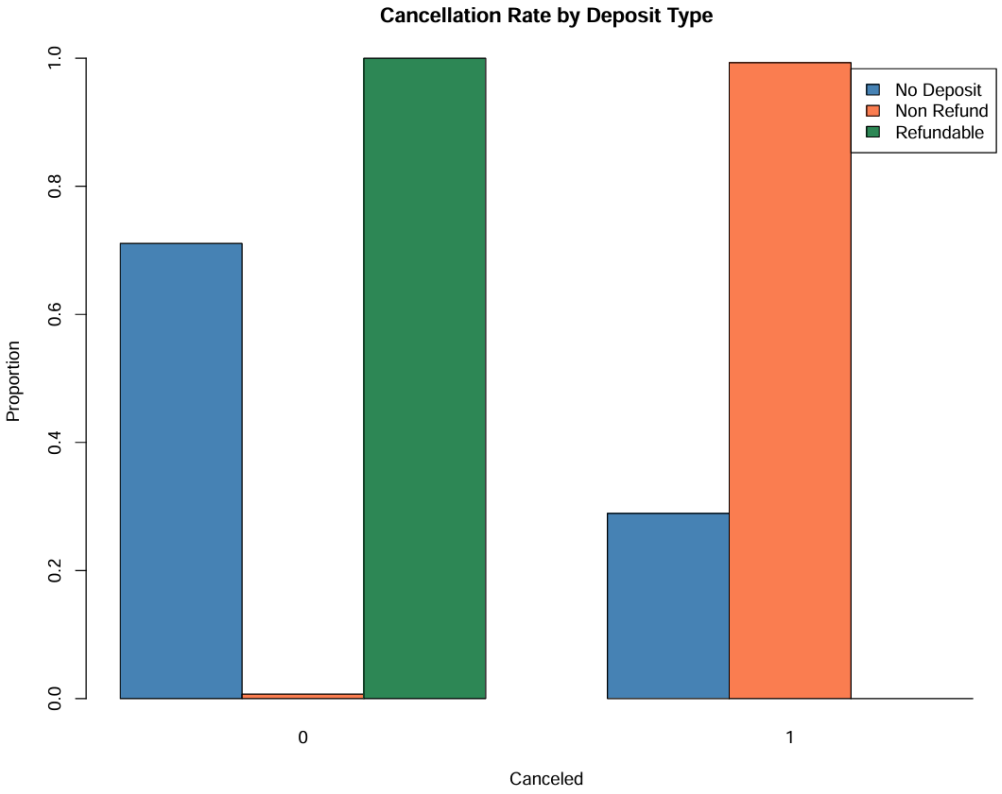

Lead Time vs. ADR
ADR (€)
0
50
100
150
200
250
300
0
100
200
300
400
500
600
Lead Time (days)

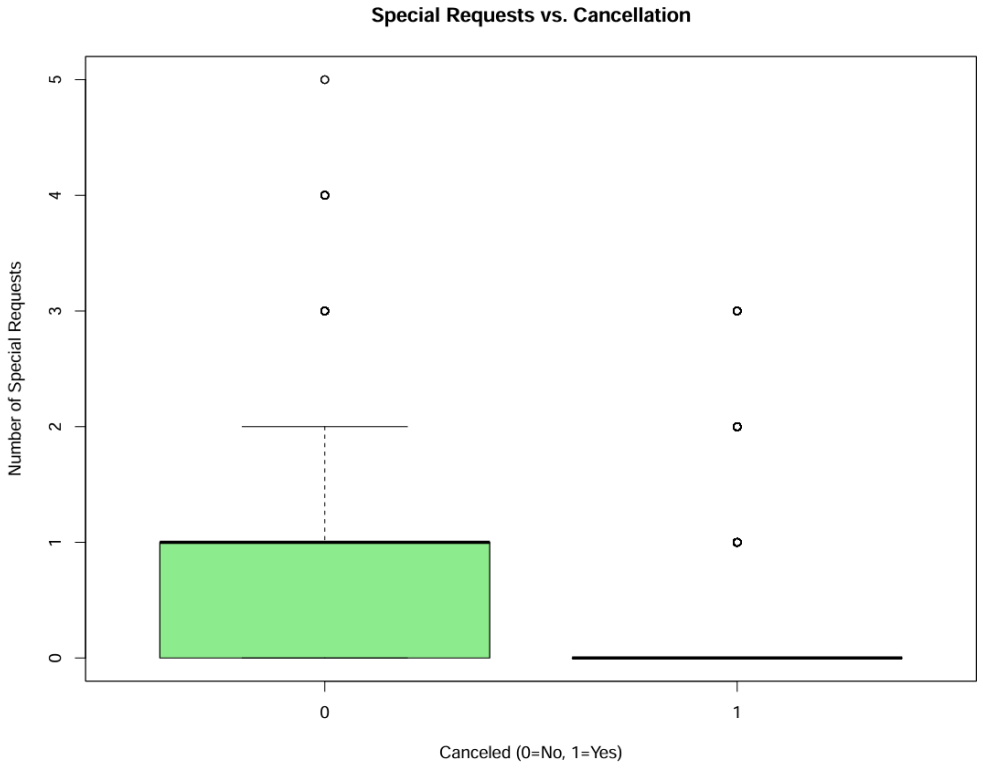

Cancellation Rate by Market Segment
Aviation
Complementary
Corporate
Direct
Groups
Offline TA/TO
Online TA
Proportion
0.0
0.2
0.4
0.6
0.8
0
1
Canceled

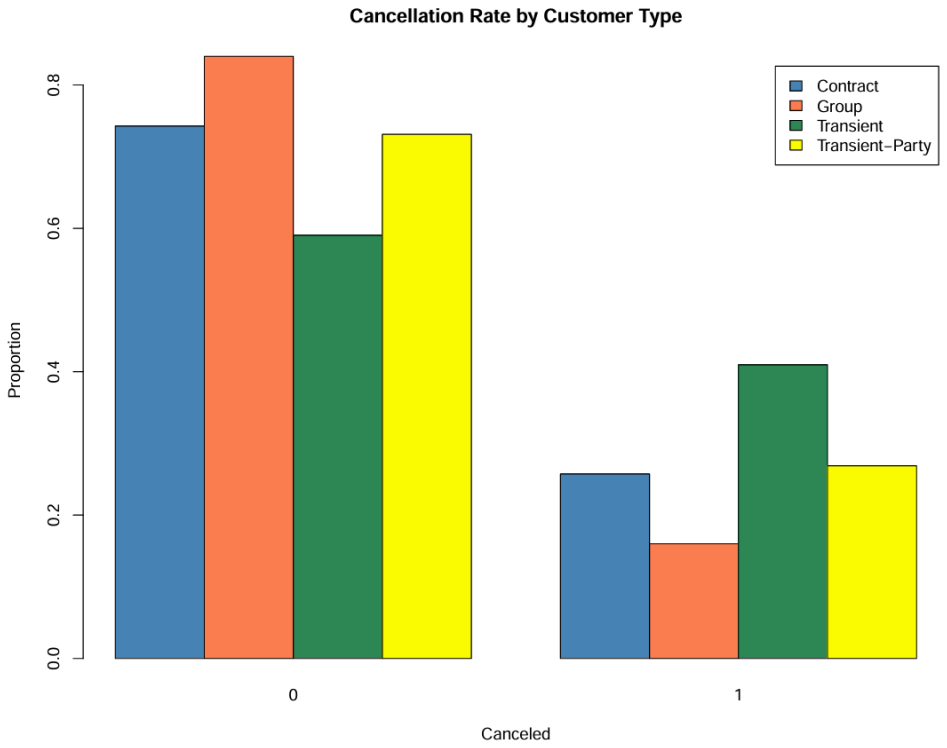

Cancellation Rate by Booking Channel
Corporate
Direct
GDS
TA/TO
Proportion
0.0
0.2
0.4
0.6
0.8
0
1
Canceled

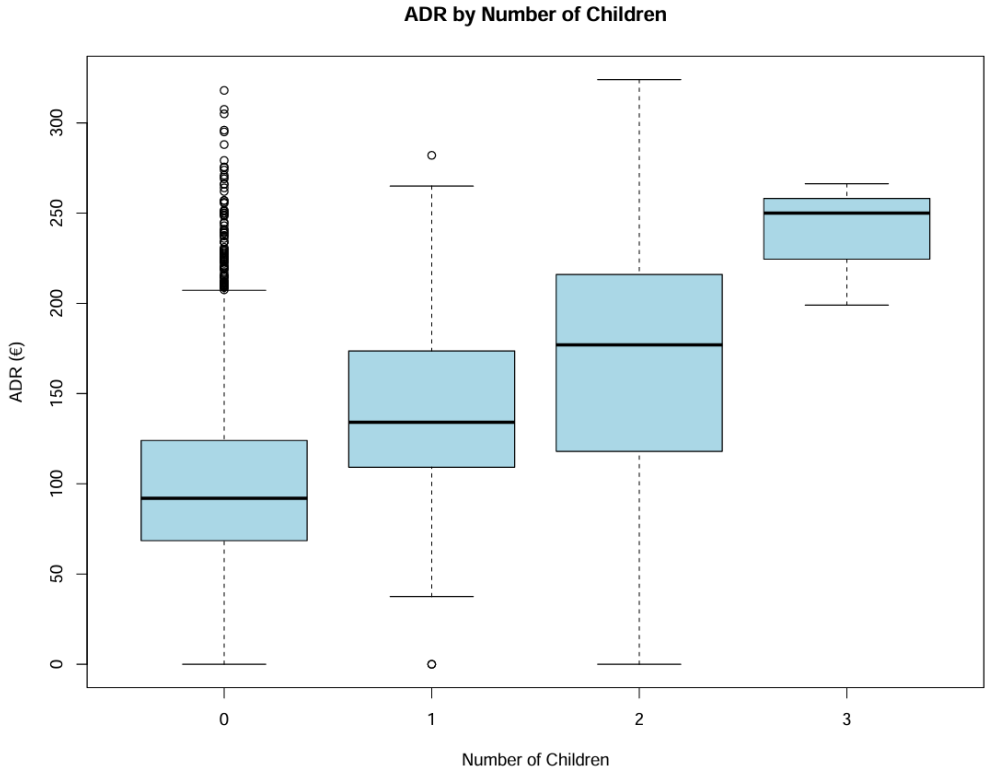

Total Stay Length vs. Cancellation
Total Nights
0
5
10
15
20
25
0
1
Canceled (0=No, 1=Yes)

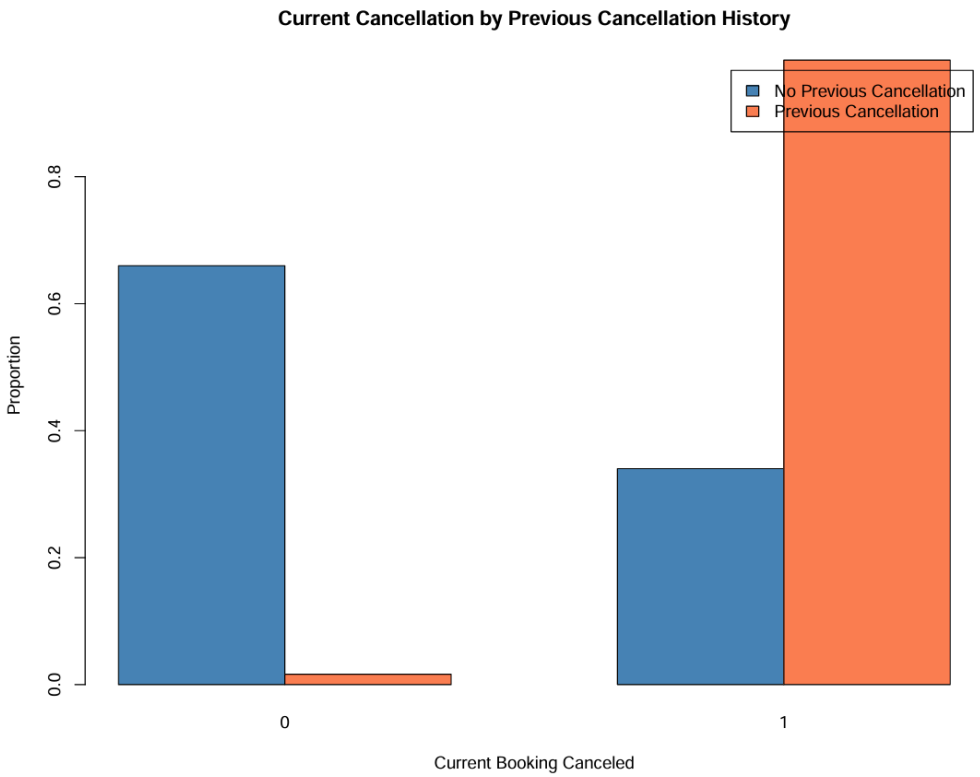

Cancellation Rate by Repeat Guest Status
New Guest
Repeat Guest
Proportion
0.0
0.2
0.4
0.6
0.8
0
1
Canceled


Cancellation Rate by Deposit Type
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
No Deposit
Non Refund
Refundable
0
1
Canceled


Special Requests vs. Cancellation
Number of Special Requests
0
1
2
3
4
5
0
1
Canceled (0=No, 1=Yes)


Cancellation Rate by Customer Type
Contract
Group
Transient
Transient-Party
Proportion
0.0
0.2
0.4
0.6
0.8
0
1
Canceled


ADR by Number of Children
ADR (€)
0
50
100
150
200
250
300
0
1
2
3
Number of Children



6.3. Analysis conclusions

The univariate analyses show that city hotels dominate the sample, with most reservations involving short stays of one to five nights and typically two adults. Most bookings are for bed-and-breakfast plans, and the distribution of average daily rates (ADR) is right-skewed, suggesting the presence of a few high-priced bookings. Seasonal trends are evident, with higher booking activity in summer months. Regarding cancellations, many reservations are not honored, reflecting variability across customer and booking characteristics. The bivariate analyses indicate that longer lead times and higher ADRs are associated with greater cancellation rates, while deposit type and customer segment strongly influence cancellation behavior. Online travel agencies show higher cancellation proportions than direct or corporate channels, and repeated guests tend to cancel less often. Overall, the plots highlight apparent behavioral and economic differences across booking types, providing insight into factors that may drive cancellation and revenue patterns.

7. PCA

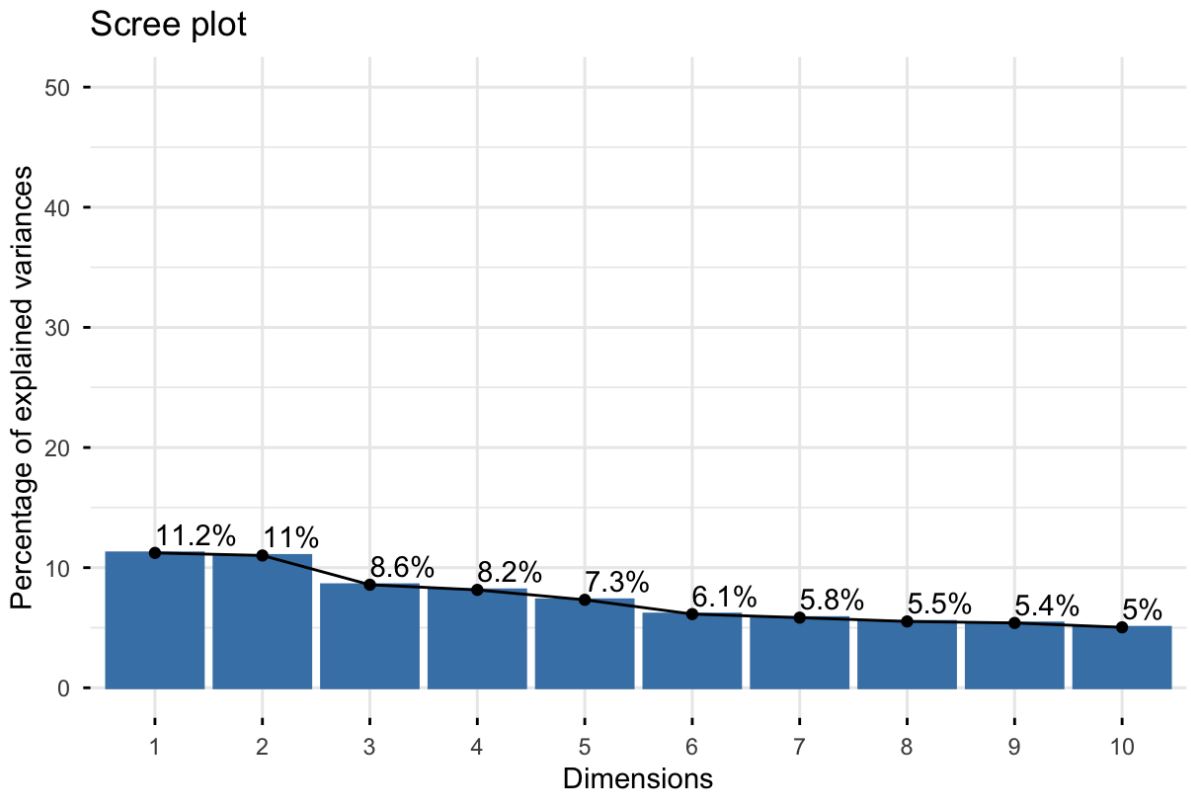
To see if some variables are not necessarily needed, we run a Principle Component Analysis. For that, we only select the numerical variables, scale everything, and then run PCA in R. After doing so we get the following information.

```
Importance of components:
                          PC1   PC2    PC3    PC4    PC5    PC6    PC7    PC8
Standard deviation     1.422 1.408 1.2427 1.2113 1.1483 1.0510 1.0254 0.9971
Proportion of Variance 0.112 0.110 0.0858 0.0815 0.0732 0.0614 0.0584 0.0552
Cumulative Proportion  0.112 0.222 0.3082 0.3898 0.4630 0.5244 0.5828 0.6380
                          PC9   PC10   PC11   PC12   PC13   PC14   PC15   PC16
Standard deviation     0.986 0.9519 0.9158 0.8865 0.8391 0.8018 0.7323 0.7084
Proportion of Variance 0.054 0.0503 0.0466 0.0437 0.0391 0.0357 0.0298 0.0279
Cumulative Proportion  0.692 0.7424 0.7890 0.8326 0.8717 0.9074 0.9372 0.9651
                          PC17   PC18
Standard deviation     0.5762 0.5439
Proportion of Variance 0.0184 0.0164
Cumulative Proportion  0.9836 1.0000
```

Now we can visualize our data in different plots and maps to get a better picture.

7.1 Scree plot

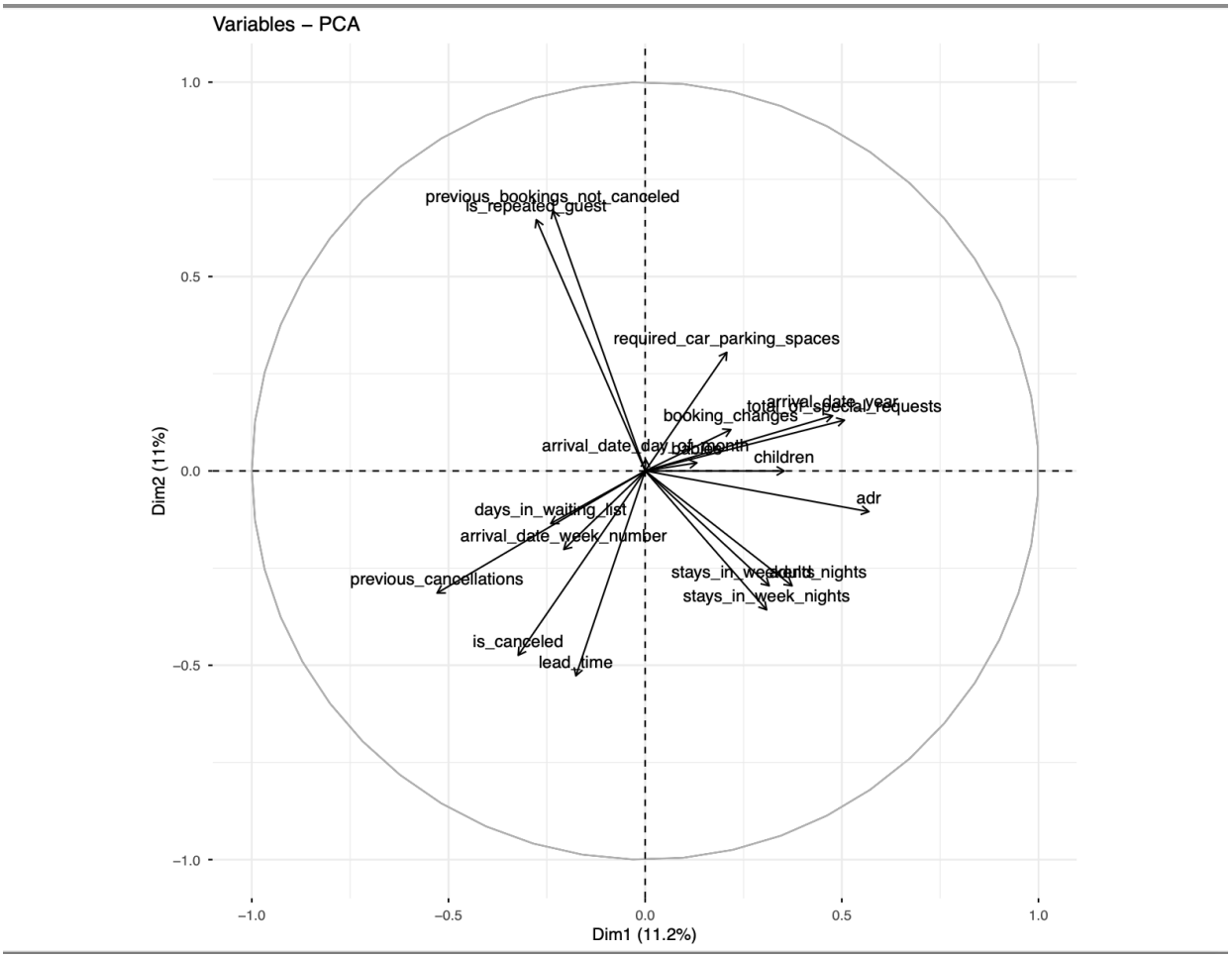
After running the PCA, we get the following scree plot.



The scree plot shows the proportion of variance explained by each component. Typically, you need to select the first components, before the curve flattens, in our case, that would be the first 3. But for a good result, we need at least 80%, and as we can see, our data does not fulfill that. Hence, we cannot remove anything.

7.2 Correlation circle

In this graph we can see how strongly the variables correlate with each other. A small angle means a strong correlation, whereas a 90º degrees angle means there is no correlation.

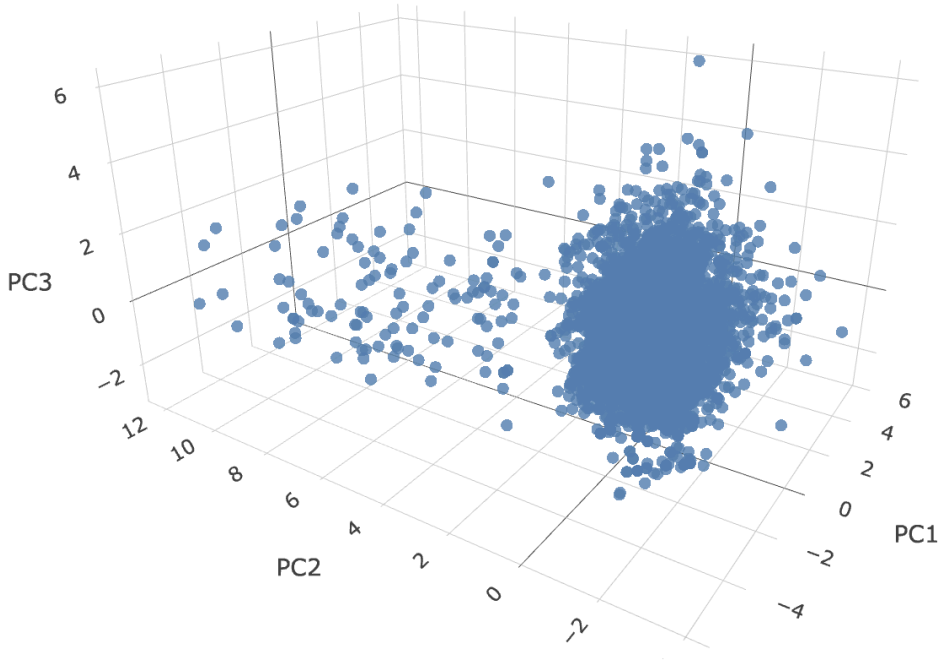


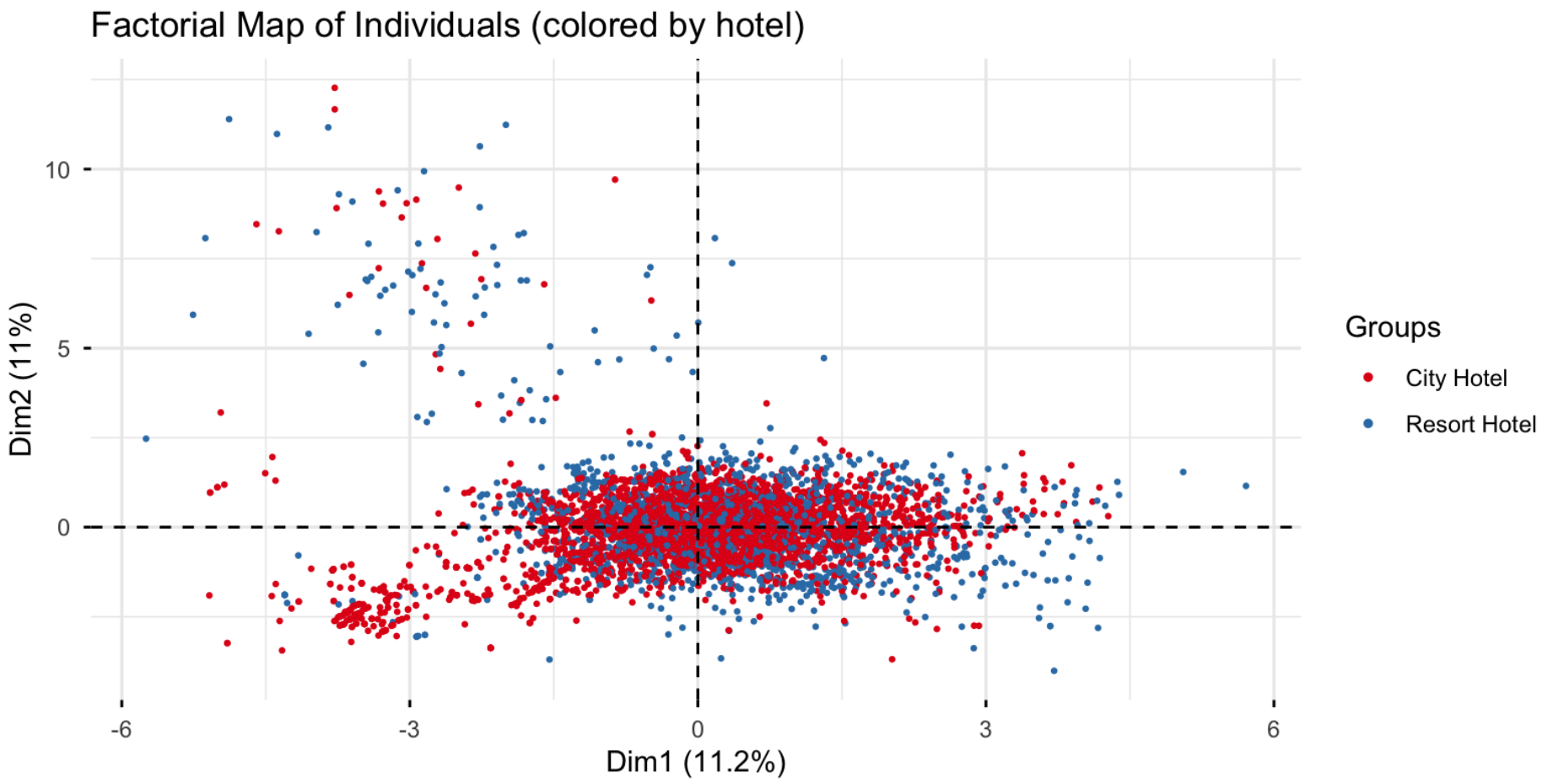
As we can see in our graph, there is no dominant axis, because our data has many different sources of variation. Each component captures a small slice of the overall variability. This confirms the complexity of our data, and that we should not reduce the dimensions.
The variables that are mostly correlated with Dim1 are adr, total_of_special_requests and children. With Dim2 however lead_time, is_cancelled and previous_cancelations.
In words, the Dim1 axis describes the size of bookings and prices, so the larger and more expensive the booking is, the larger Dim1, whereas Dim1 depends on the uncertainty of the guests.

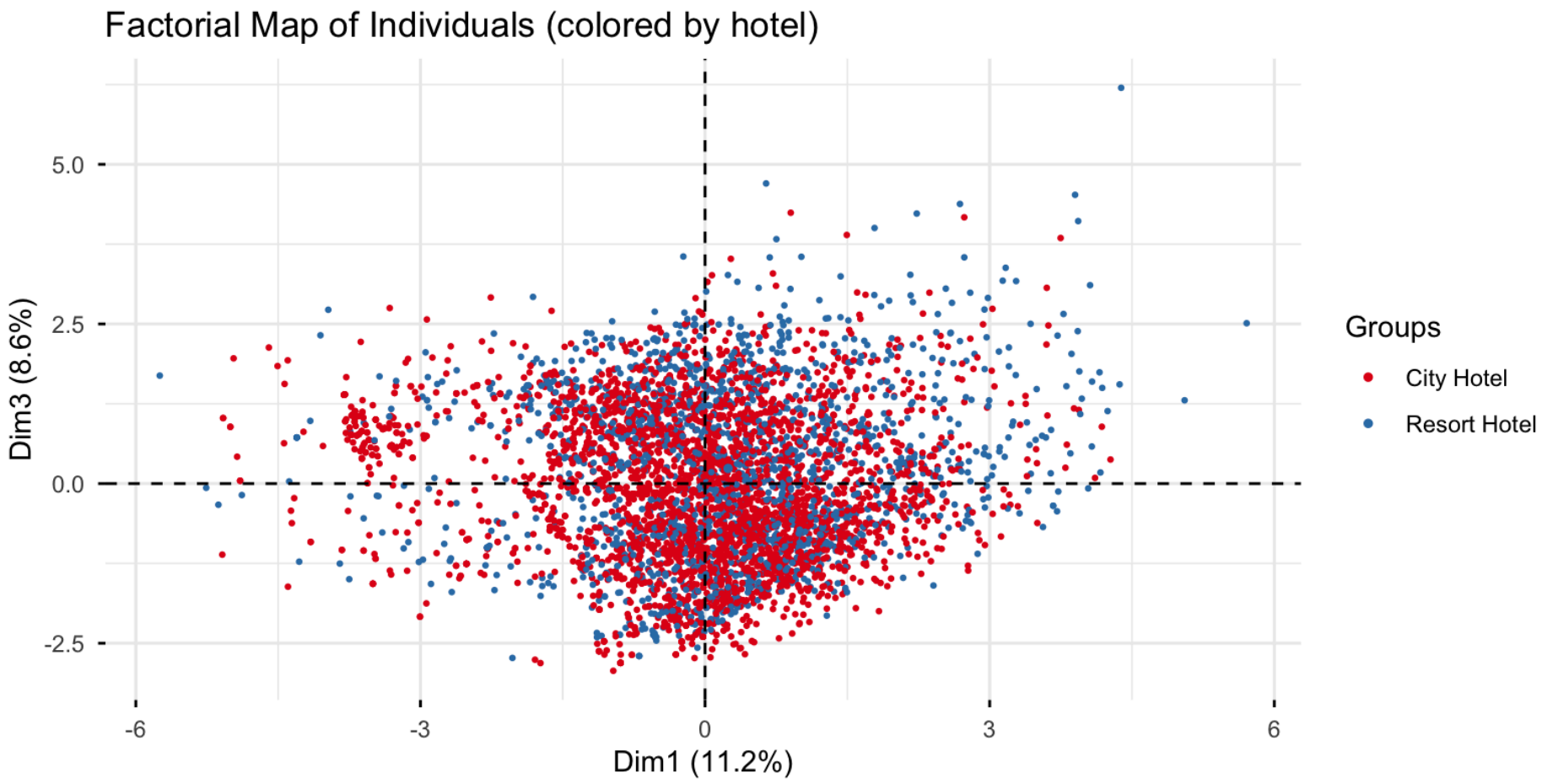
7.3 Factorial Maps

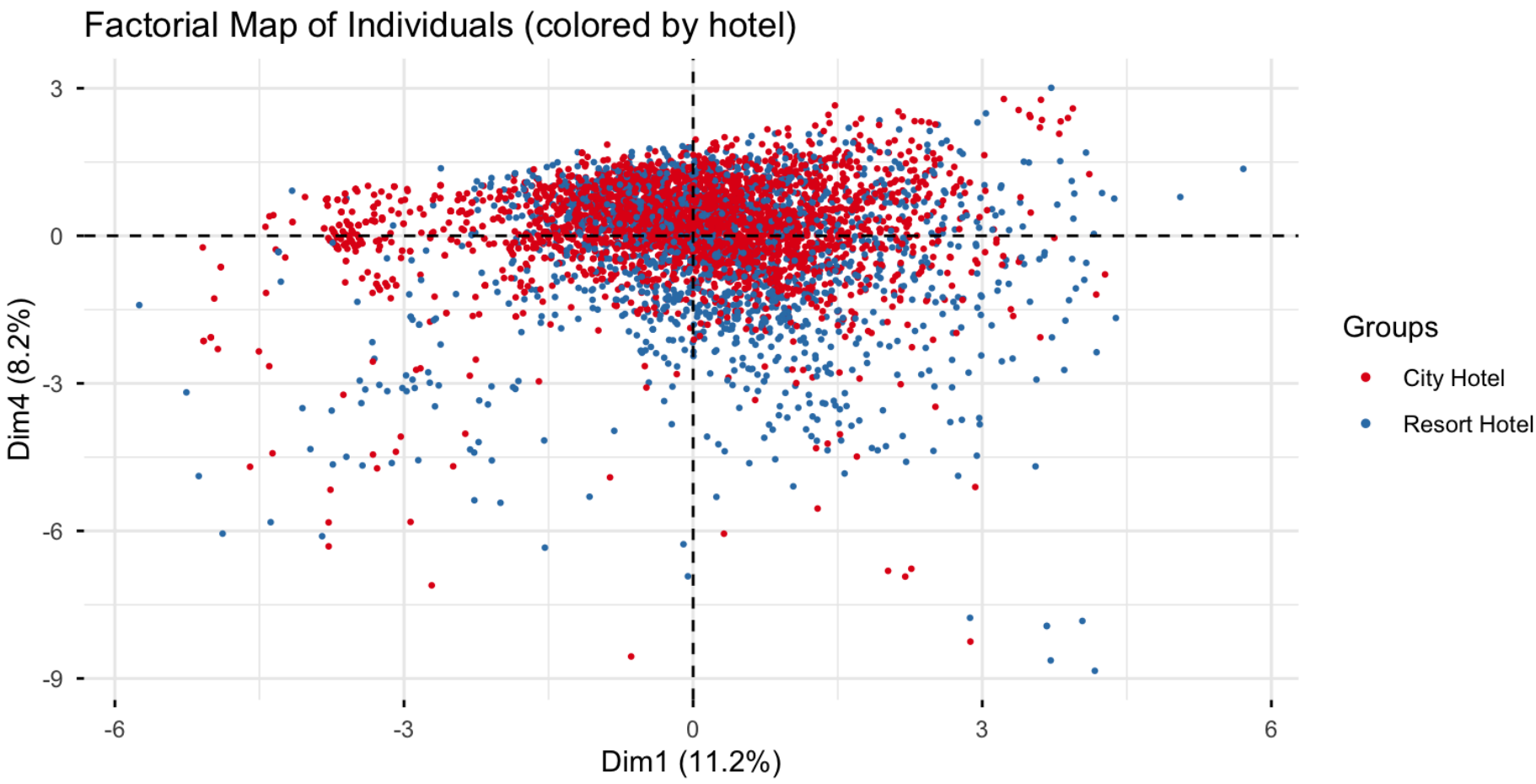
We represent the data on different factorial maps, to see how well the data is distributed. After doing so, we can see, that fno matter which dimensions we choose for the axes, the points are really close to the center, meaning that most of the variables have values close to the average of all variables, so they do not help in distinguishing patterns.

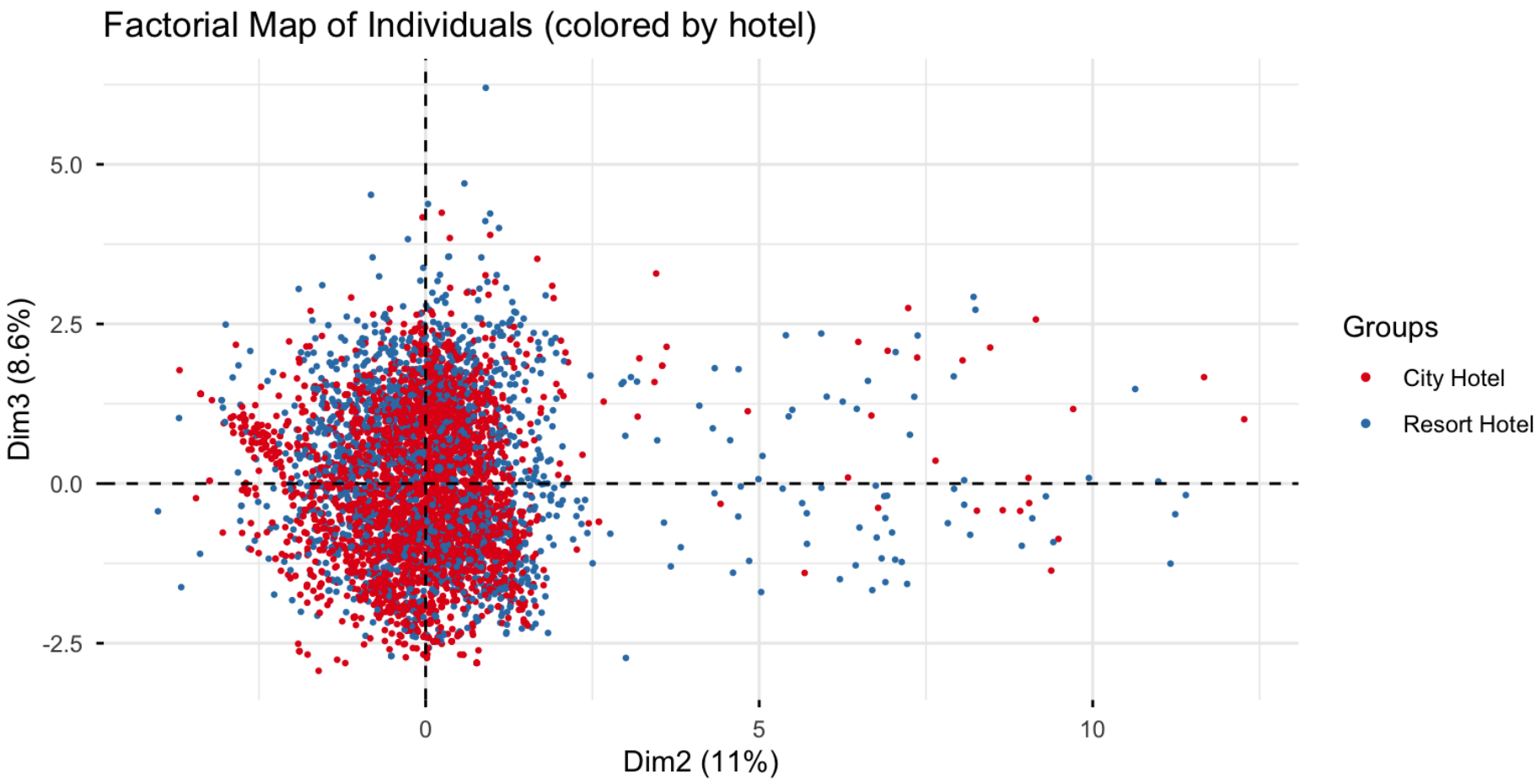
And even with a three dimensional visualization, the data is very centered.

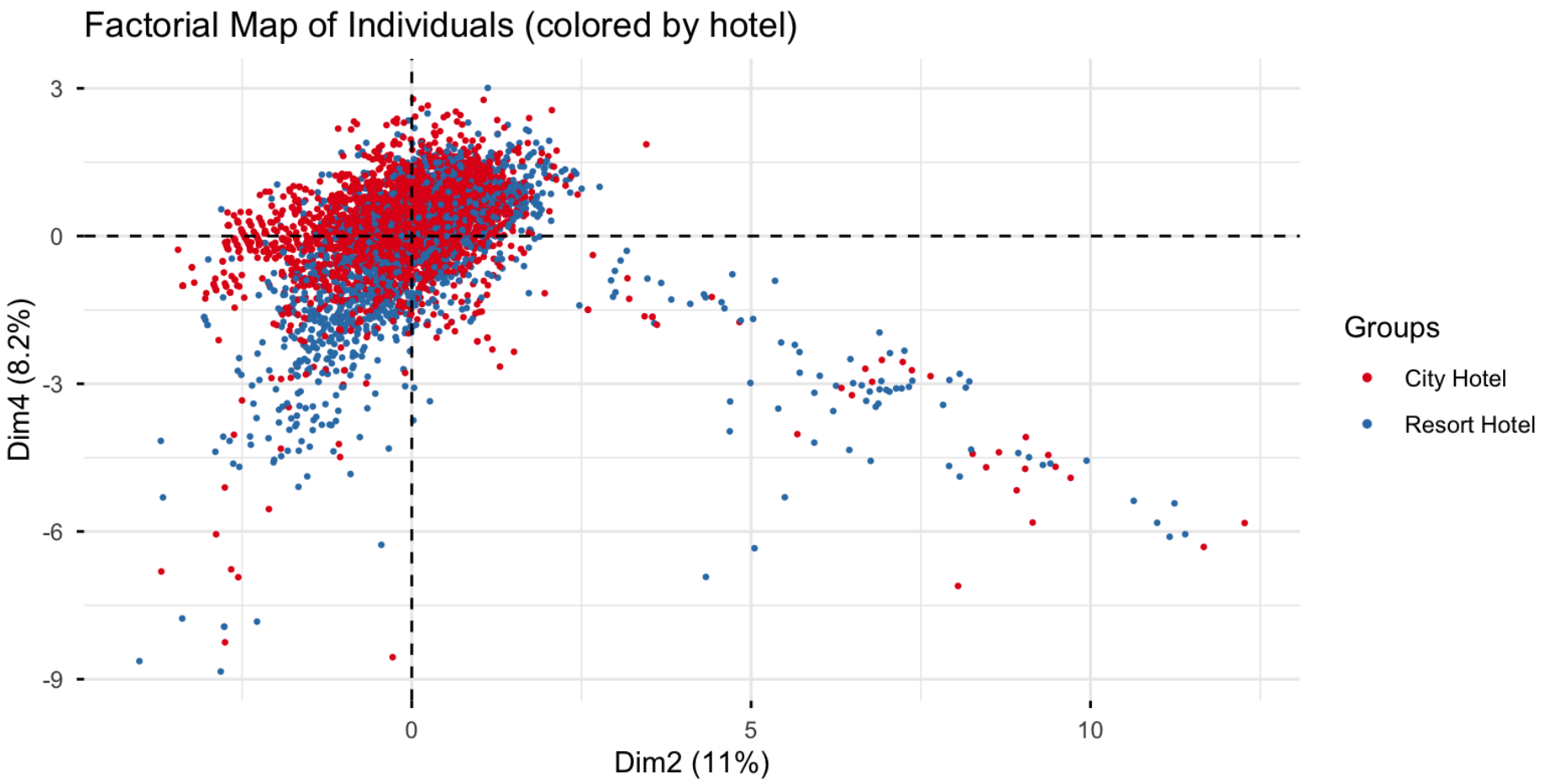


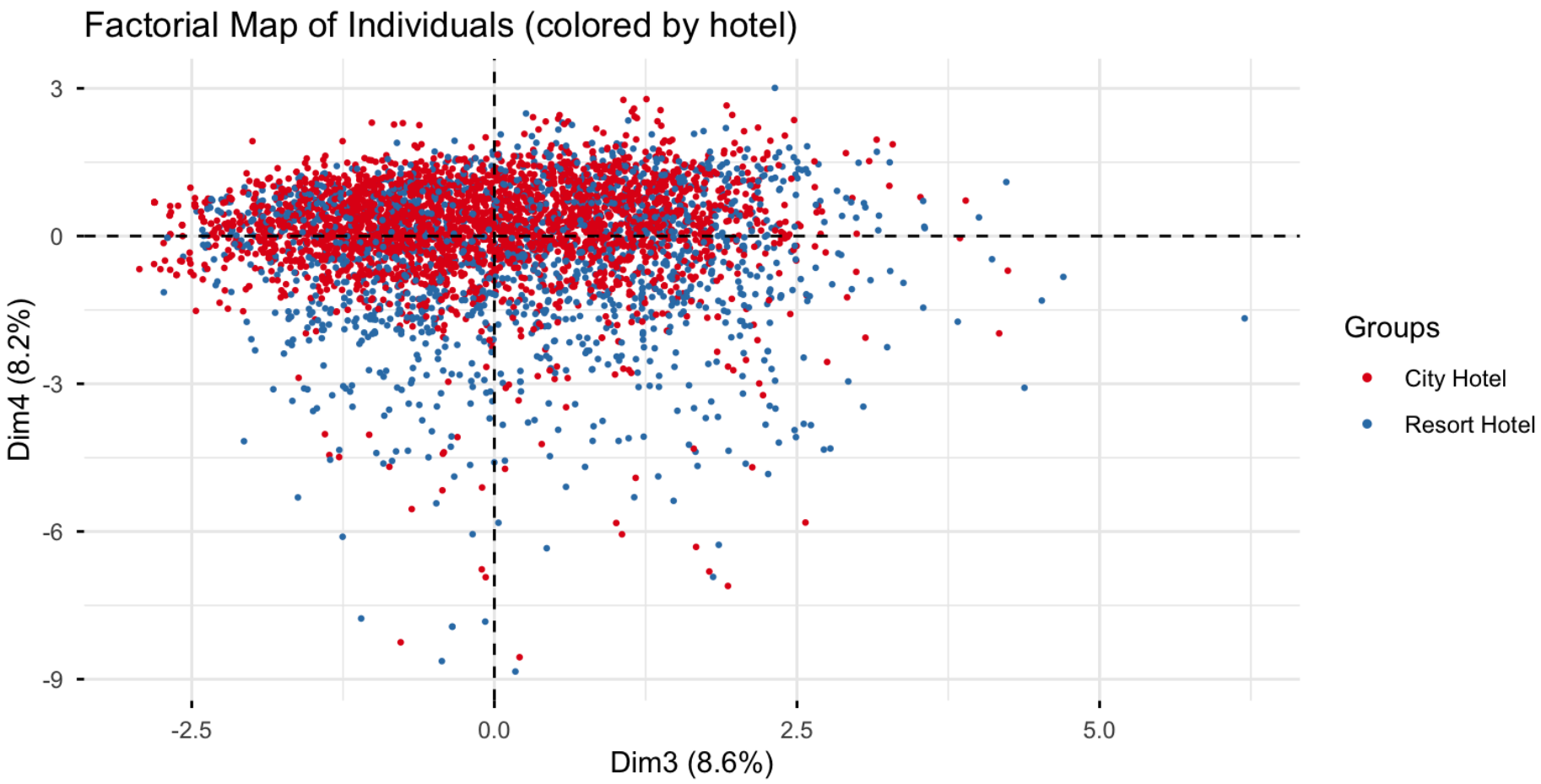

Factorial Map of Individuals (colored by hotel)
Dim2 (11%)
10
5
0
-6
-3
0
3
6
Dim1 (11.2%)
Groups
City Hotel
Resort Hotel

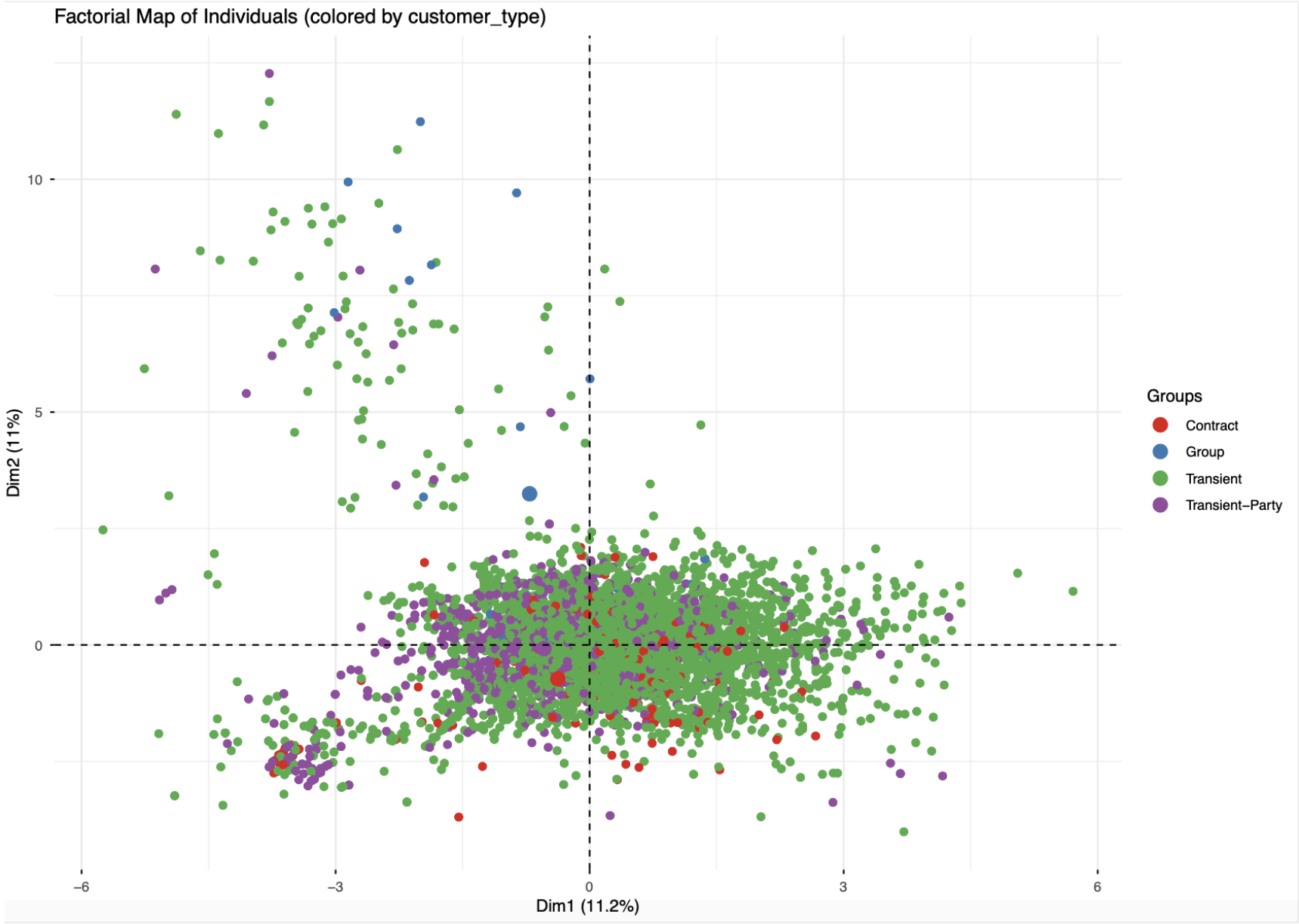

Factorial Map of Individuals (colored by hotel)
Dim3 (8.6%)
5.0
2.5
0.0
-2.5
-6
-3
0
3
6
Dim1 (11.2%)
Groups
City Hotel
Resort Hotel

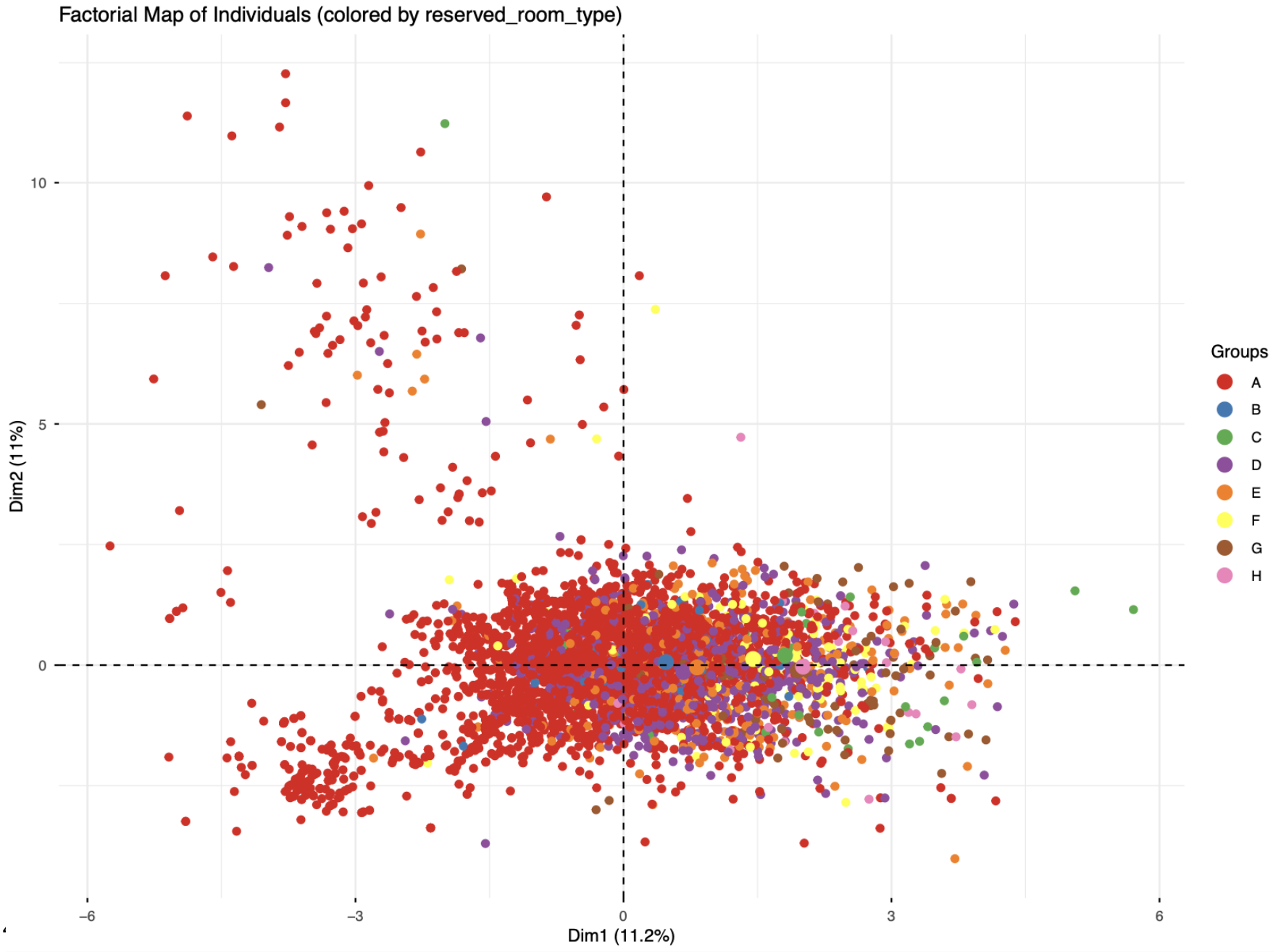

Factorial Map of Individuals (colored by hotel)
Dim4 (8.2%)
Dim1 (11.2%)
3
0
-3
-6
-9
-6
-3
0
3
6
Groups
City Hotel
Resort Hotel


Factorial Map of Individuals (colored by hotel)
Dim3 (8.6%)
5.0
2.5
0.0
-2.5
0
5
10
Dim2 (11%)
Groups
City Hotel
Resort Hotel


Factorial Map of Individuals (colored by hotel)
Dim4 (8.2%)
3
0
-3
-6
-9
0
5
10
Dim2 (11%)
Groups
City Hotel
Resort Hotel


Factorial Map of Individuals (colored by hotel)
3
0
-3
-6
-9
Dim4 (8.2%)
-2.5
0.0
2.5
5.0
Dim3 (8.6%)
Groups
City Hotel
Resort Hotel


Factorial Map of Individuals (colored by customer_type)
Dim2 (11%)
10
5
0
−6
−3
0
3
6
Dim1 (11.2%)
Groups
Contract
Group
Transient
Transient−Party


Factorial Map of Individuals (colored by reserved_room_type)
Dim2 (11%)
10
5
0
−6
−3
0
3
6
Dim1 (11.2%)
Groups
A
B
C
D
E
F
G
H

7.4 Conclusion

The PCA results show us that the dataset is very complex and multidimensional. The first two principal components together explain only around 22% of the total variance, meaning that the data cannot be represented by a small amount of linear combinations. The variable correlation plot shows that many variables are only weakly correlated with the first axes and point in different directions, also showing that the correlations among variables are not easily described. Similarly, the factorial map of individuals shows a dense cloud of points, even when using three dimensions, with no clear separation among groups, so that our data cannot be compromised.

Hence, PCA is not an appropriate method for dimensionality reduction in this dataset, because it fails to summarize the information efficiently. The data structure is too heterogeneous, with numerous variables contributing small, distinct portions of variance. That is why in the next section we will try to use clustering as a better compromise method.

8. Hierarchical Clustering

8.1 Precise description of the data used

For the clustering analysis, we used all numerical variables of the original dataset. These are lead_time,year,week, day, weekend_nights, week_nights,adults,children,babies, pre_cancel, pre_bcn,changes, days_wait,adr,rcar_parking_spaces, ts_requests.

8.2 Clustering method used, metrics, and aggregation criteria used

K-means

In our code, Ib represents the percentage of inertia, which describes the quality of the grouping. We observe that the inertia between clusters with k=5 is 99.92, a high percentage. With this result, we affirm our groups are well separated.

```
> Ib2

[1] 99.92975
```

Testing with k=8, we see that it only increased by 0.0206%; the difference is minimal. With 8 clusters, the groups can be too small or similar, which complicates their interpretation.

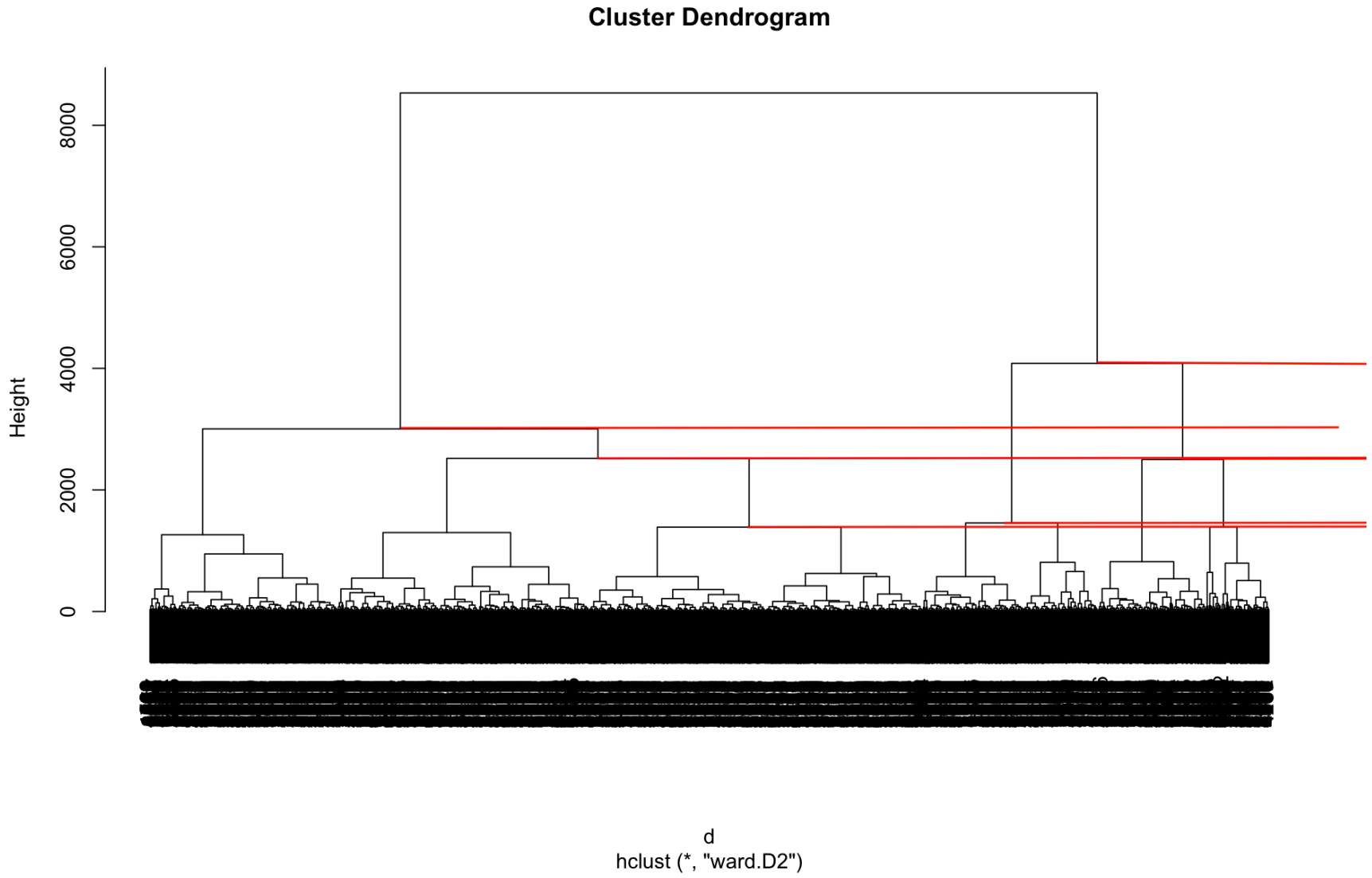
```
> Ib3
[1] 99.9503
```

Hierarchical clustering

Ward's method is used as the aggregation criteria in our script. h1 <- hclust(d,method="ward.D"). We chose it because it can form evenly sized, compact clusters, which is very suitable for customer segmentation.

Euclidean distance is used for calculating dissimilarity between numerical data points in the initial hierarchical clustering.

8.3 Resulting Dendrogram



Once we execute the hierarchical clustering on the whole dataset, we can look at the lines and find out the most significant gap (excluding the 2 clusters), which is the one that gives us a number of 6 classes.

8.4 Discuss how to get the final number of clusters

We performed hierarchical clustering using Ward's method, and the resulting dendrogram was examined to identify the most significant split. After the initial split into two clusters, the most significant vertical gap in the dendrogram was observed at the point that created six clusters.

8.5 Table with a description of the cluster size

Table showing how many members are in each cluster

```
> table(c2)
c2
   1    2    3    4    5    6
 765  567 2068  690  570  206

> table(c6)
c6
   1    2    3    4    5    6
 763  817 1091 1452  470  273
```

9. Profiling

9.1. Numerical variables traffic light method

	Numerical												
Mean	0.0193	0.3725	100.9058	0.9445	2.4897	0.1019	0.0503	0.0267	0.1886	0.7951	0.5748	0.0606	102.9327
	repeated	canceled	lead_time	wkend_nights	week_nights	children	pre_cancel	pre_bcn	changes	days_wait	ts_request	rcar_prkg_spc	adr
Cluster 1													
Cluster 2													
Cluster 3													
Cluster 4													
Cluster 5													
Cluster 6													

9.2. Categorical variables

channel: Clusters 1, 2, 3, and 4 heavily depend on the *Online TA* channel. Clusters 5 and 6 are distinguished by having a much more balanced channel distribution (Cluster 5) or favoring *Direct* reservations (Cluster 6).

deposit_type: Most reservations in clusters 1 to 5 are 'No Deposit'. Cluster 6 stands out as the only one where 'Non Refund' reservations are as common or more common than 'No Deposit' reservations, indicating a customer profile more committed to advance payment.

rroom_type: *Category A* is dominant in all clusters. Clusters 3, 4, and 6 greatly depend on this category. The difference between clusters can be seen in the mix of secondary categories *(B, C, D)*.

r_status: Clusters (1 and 2) are groups of reservations where cancellation is the norm. Clusters (3, 4, and 5) are defined as groups of reservations where check-out (completed reservation) is the norm.

market_seg: The *TA/TO* segment dominates overwhelmingly in all clusters. Cluster 4 is the largest in this segment. Cluster 6 is distinguished by its low total volume and the complete absence of the *Corporate* and *Direct* segments.

month: Clusters 1, 2, 4, and 5 are heavily concentrated in the high season (April to September). Clusters 3 and 6 show a more dispersed distribution throughout the year.

Once we figured out that our dendrogram had 6 clusters, we started the profiling section, where we examined each selected variable and obtained information about each cluster. We made profiling plots with every variable in our database. After collecting all the graphs and numerical values (ANOVA Test, Chi-square test, and p-values), we saw the variables that defined each cluster and how they did it.

To show how variables affect each cluster, we have created a traffic light matrix to represent how the numeric variables affect each cluster. We have also decided to show the most important plots of the categorical variables.

After executing the clustering and profiling process, it is possible to identify six marked customer clusters:

Cluster 1: Indecisive planners

This is a high-risk cluster. It possesses high planning and stay duration values, but the most principal characteristic for this cluster is the high cancellation rate. Their main channel for booking in peak seasons is the Online Travel Agency — Online TA. They have also shown very low commitment, as evidenced by the exclusive choice of "No Deposit" reservations. Besides that, these customers have a significant initial interest, with a low probability of completing their stay.

Cluster 2: High cancellation rate standard bookings

This cluster has a high cancellation rate, like Cluster 1, qualifying it as another low-quality cluster. Their bookings are regular by numerical values, and, as in Cluster 1, they are mostly made through Online TA during the high season without any financial commitment given as a warranty, No Deposit. This cluster can be generalized into transactional bookings, which run a highly probable risk of being canceled.

Cluster 3: Consistent economic clients (terminated basic bookings)

This cluster shows a high reliability rate despite the lower values. The most outstanding status is Check-Out, which describes a fantastic success rate. Their reservations are mostly in room category A (basic) and through Online TA, but unlike the risk clusters, these are distributed yearly. This segment ensures a continuing, although low-revenue, inflow.

Cluster 4: Tenacious and organized TA/TO segment clients.

Cluster 4 is the backbone of successful bookings since it always presents high levels of planning. It is the biggest group identified and is close to TA/TO, with the main channel being Online TA. Nearly all their bookings materialize in Check-Outs, with most in high season. They are organized customers whose results for agencies in occupancy rates are relevant.

Cluster 5: Stable and multi-channel clients

The peculiar characteristic of this cluster is its channel diversity, characterized by opting for a balanced mix rather than relying on Online TA exclusively. It has shown a blend of stable numerical behaviors, with a final status mainly Check-Out, with a low probability of cancellation. Their activity predominantly occurs in peak seasons. This profile represents a reliable customer acquired through different channels.

Cluster 6: Inbound direct clients devoted and prepaid

Cluster 6 is outstanding in the magnitude of its commitment. It's the only group that dominantly booked through the channel: Direct (non-agency). Not only that, but its pattern of deposit types has also been very different: it frequently uses the "Non-Refund" option, which in translation means pre-payment and means very high security regarding the reservation status. Their status generally is Check-Out, and their booking activity is more dispersed throughout the year. This profile is desirable regarding direct profitability and the certainty of completed stays.

10. Linear Regression

The primary objective of this section is to construct a robust linear regression model to predict the Average Daily Rate (adr) of a hotel booking. As response value we use adr which is a continuous variable representing the nightly revenue per room.

Two main models were explored to find the best balance between predictive power and parsimony:

The initial model included every numerical variable in the database. This gave us a high R^2 value but also, its higher complexity resulted in worse plots that were more difficult to read. We named this model “reg_full”.

Our final model used the variables: adults, children, meal, year, month, hotel, pre_bcn, pre_cancel, repeated, rroom_type, lead_time and country. This model gave us a similar R^2 value (both R^2 may be too small), and since it uses an inferior number of variables it creates much better plots. We named this model “reg3”.

Model	**R^2 (Explained Variability)**	**Adjusted R^2**	**Residual Standard Error (S)**
reg_full	0.6152	0.6031	29.76
reg3	0.5496	0.5469	31.8

These models establish a linear relationship between the ADR and the predictor variables. The majority of the estimated coefficients in the model were found to be statistically significant (p-values < 0.05), validating their contribution to ADR prediction.

The magnitude and sign of the estimated coefficients indicate how each regressor affects the ADR, holding all other variables constant:

Positive Effect:

- **Occupancy**: The adults and children variables have a direct positive impact, which is intuitive: a higher number of guests increases the base rate.
- **Room and Meal Type**: Higher categories in variables such as meal or certain rroom_type increase the ADR, reflecting the added value of these services.

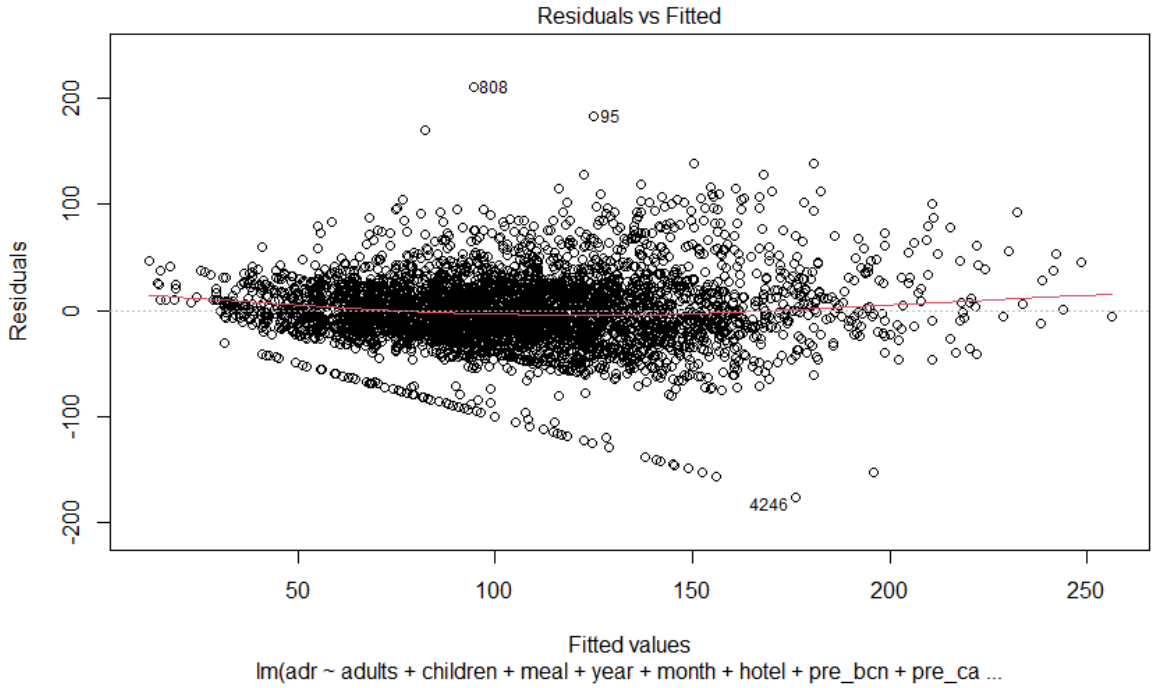
Negative Effect:

- **Booking Status (r_status)**: The coefficient for the "Cancelled" status (r_status = Cancelled) is often strongly negative.
- **Repeated bookings**: If the repeated (repeated guest) coefficient is negative, it suggests that returning customers may benefit from discounts or special rates.

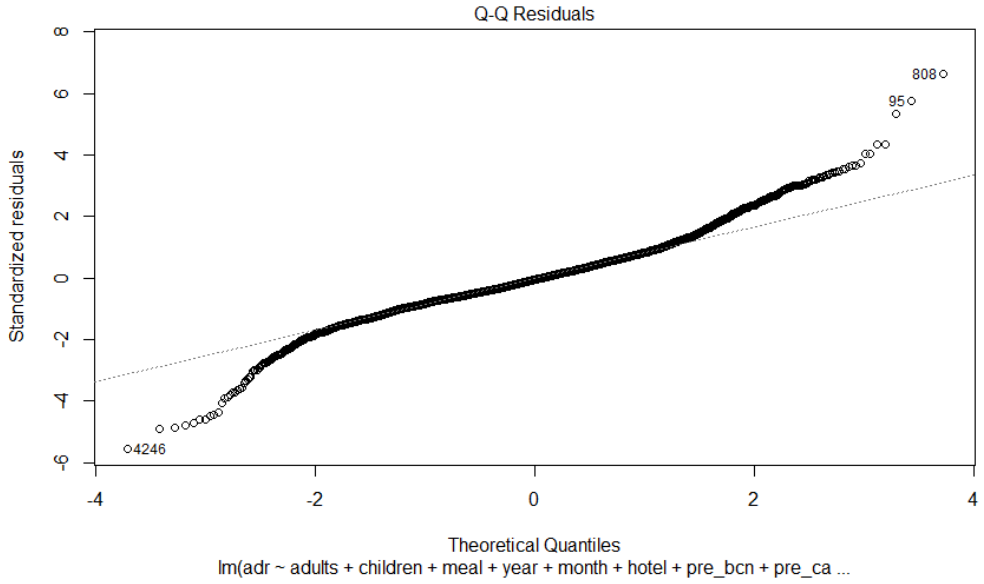
10.1. Graphical Residual Analysis

Assumption validation is performed by inspecting the residual plots generated by our final model:

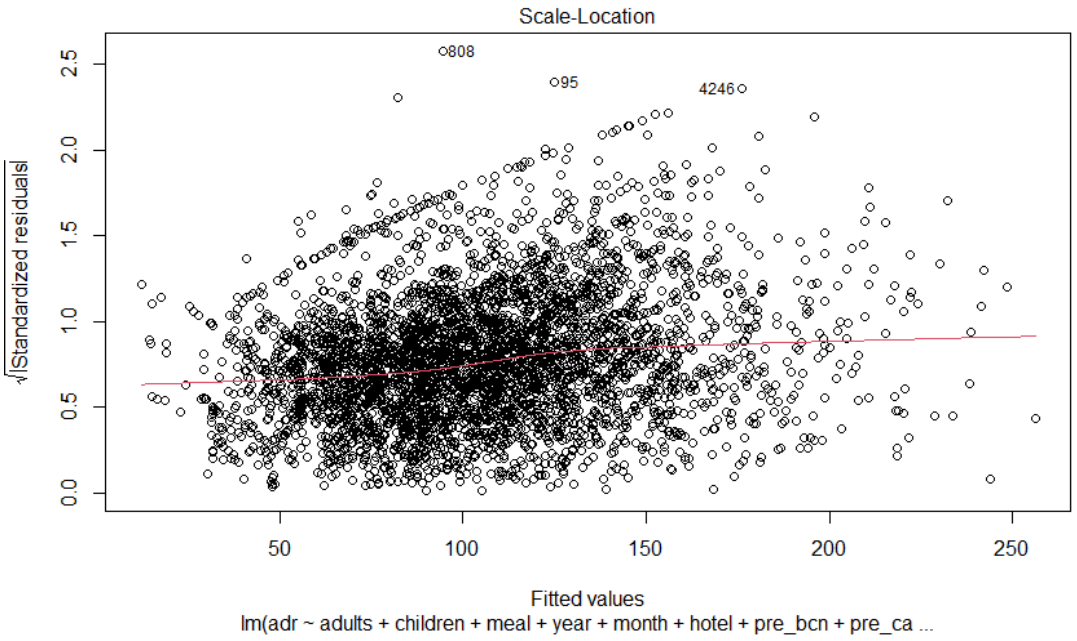
1. **Residuals vs. Fitted:** The expected shape of this plot is a random scatter of points around the horizontal zero line (heteroscedasticity). In our case we observe that the spread of the residuals widens as the fitted values (predicted ADR) increase. This indicates that our model predicts lower ADRs more accurately than higher ADRs. Additionally, a slight curve in the red trend line suggests that the relationship between the predictors and ADR may not be purely linear.



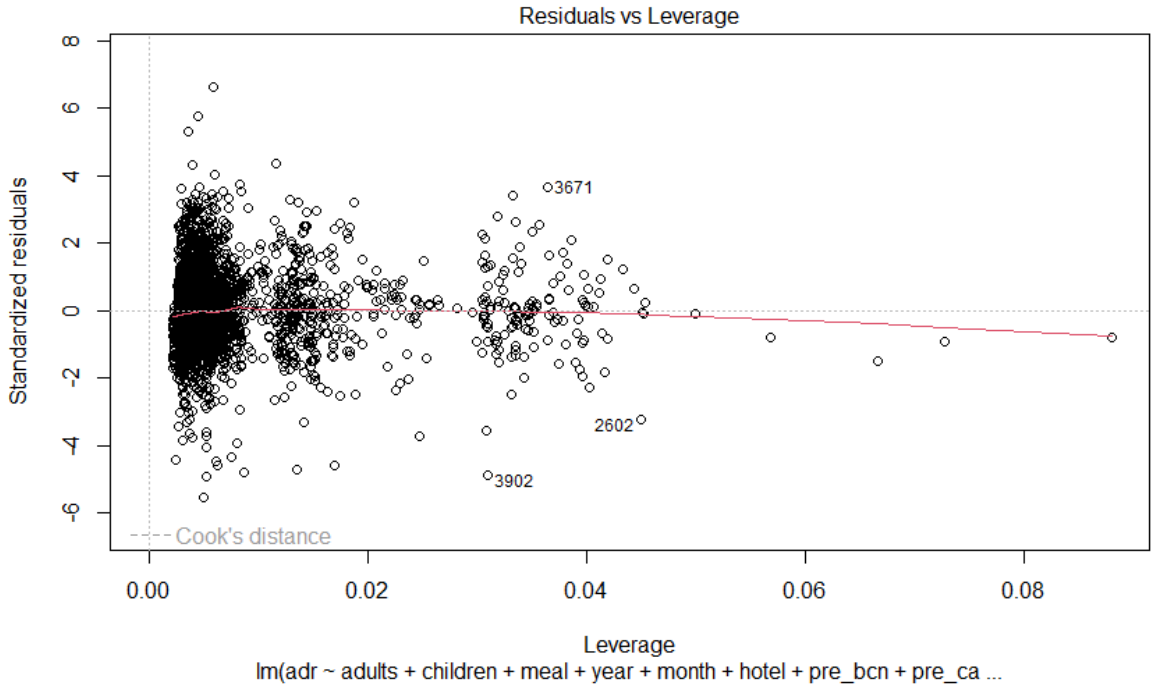
2. **Normal Q-Q Plot:** Assesses the normality of the error distribution. Points should closely follow the straight diagonal line. While the central points align reasonably well with the line, we observe significant deviations at both the upper and lower tails. These deviations indicate that the residuals are "heavy-tailed," meaning outliers occur more frequently than expected in a normal distribution



3. **Scale-Location:** Confirms homoscedasticity using the square root of the standardized residuals. The fitted line (red) should be flat and horizontal. The red line in our plot clearly slopes upward. This confirms the findings from the Residuals vs. Fitted plot: the variance of the errors is not constant but increases with the magnitude of the predicted value.



4. **Residuals vs. Leverage:** Identifies influential points that have a disproportionate impact on the model coefficients.



10.2. Conclusion

In conclusion, we selected the simplified model reg3 to predict the Average Daily Rate (ADR). While the full model offered marginally higher predictive power, reg3 was preferred for its interpretability and reduced complexity. The analysis of this model confirmed that factors such as the number of guests and premium services significantly increase revenue, while repeated bookings are associated with lower rates, likely reflecting loyalty discounts.

However, the graphical residual analysis strongly suggests that the underlying data does not follow a strictly linear relationship. The "Residuals vs. Fitted" plot displays a distinct curved pattern rather than a random scatter, indicating that a simple linear regression fails to capture the complexity of how ADR responds to changes in the predictor variables. We think that this linear regression is insufficient to explain this relationship.

11. Logistic Regression

The main objective of this section is to model and predict the probability that a hotel booking is canceled. As a qualitative response variable, we use *canceled*, a binary variable that indicates whether a reservation was canceled (1) or not (0).

An initial logistic regression model was constructed using several explanatory variables related to booking characteristics, customer profile, and price. After the variable selection process, the final model included the following predictors: **lead_time, week_nights, adults, deposit_type, customer_type and adr**. The final logistic model achieves a deviance reduction of 1415.29 and a McFadden Pseudo-R^2 of 0.22, indicating a good balance between explanatory power and parsimony.

The estimated coefficients describe how each regressor affects the log-odds of cancellation, holding all other variables constant. Several coefficients were found to be statistically significant (p-values < 0.05), confirming their relevance in explaining cancellation behavior.

FINAL MODEL
canceled= -2.926 + 0.003552 lead_time + 0.03520 week_nights + 0.2375 adults + 5.415 deposit_typeNon Refund -12.28 deposit_typeRefundable + 0.4925 customer_typeGroup + 1.082 customer_typeTransient + 0.5212 customer_typeTransient-Party + 0.002056 adr

This model describes the relationship between the selected explanatory variables and the probability that a reservation will end up being canceled.

The sign of the coefficients indicate how each regressor affects:

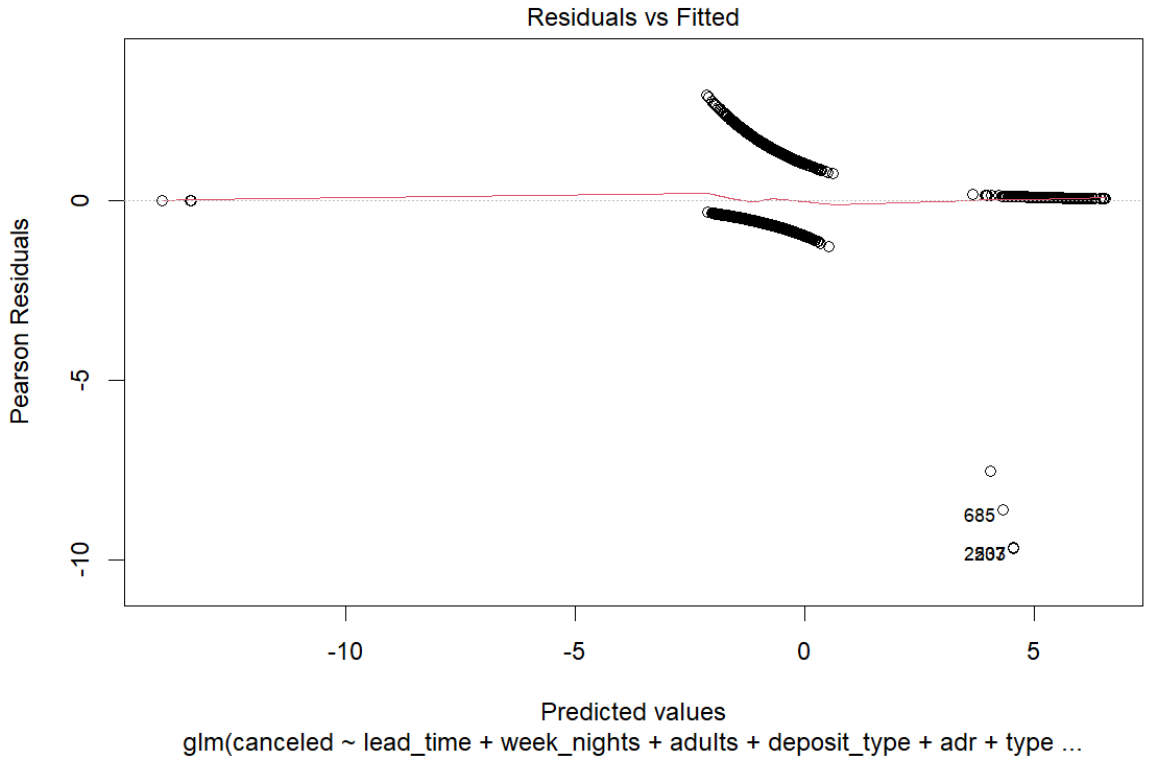
Positive effects on cancellation probability were observed for:

- Lead time: bookings made further in advance are more likely to be canceled.
- week_nights, adults:: Number of week nights and number of adults, which increase booking complexity.
- ADR (price): higher prices are associated with a higher probability of cancellation.
- deposit_type.Non Refund and customer_type.Transient, which show a higher tendency to cancel.

We see that deposit_typeRefundable has negative sign which would indicate that reservations with a refundable deposit have a lower probability of cancellation compared to the reference category. But as it is not statistically significant, nothing can be concluded with certainty.

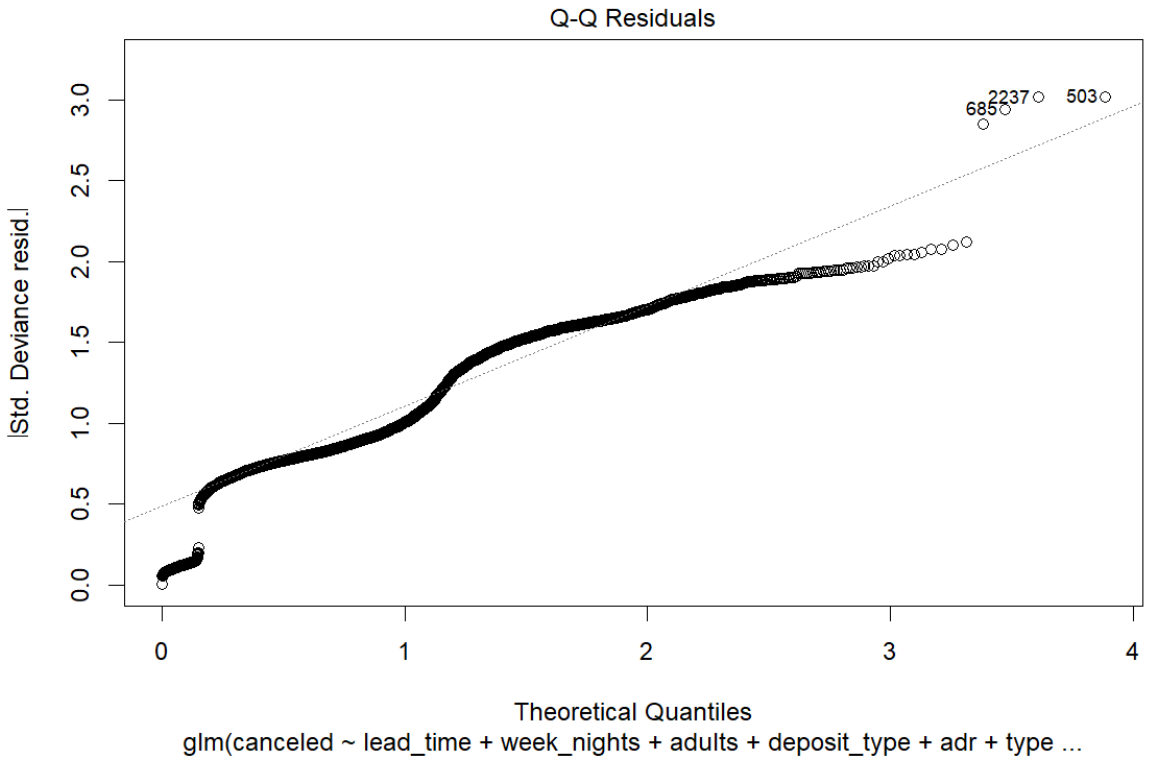
GRAPHICAL RESIDUAL ANALYSIS

Residuals vs. Fitted:



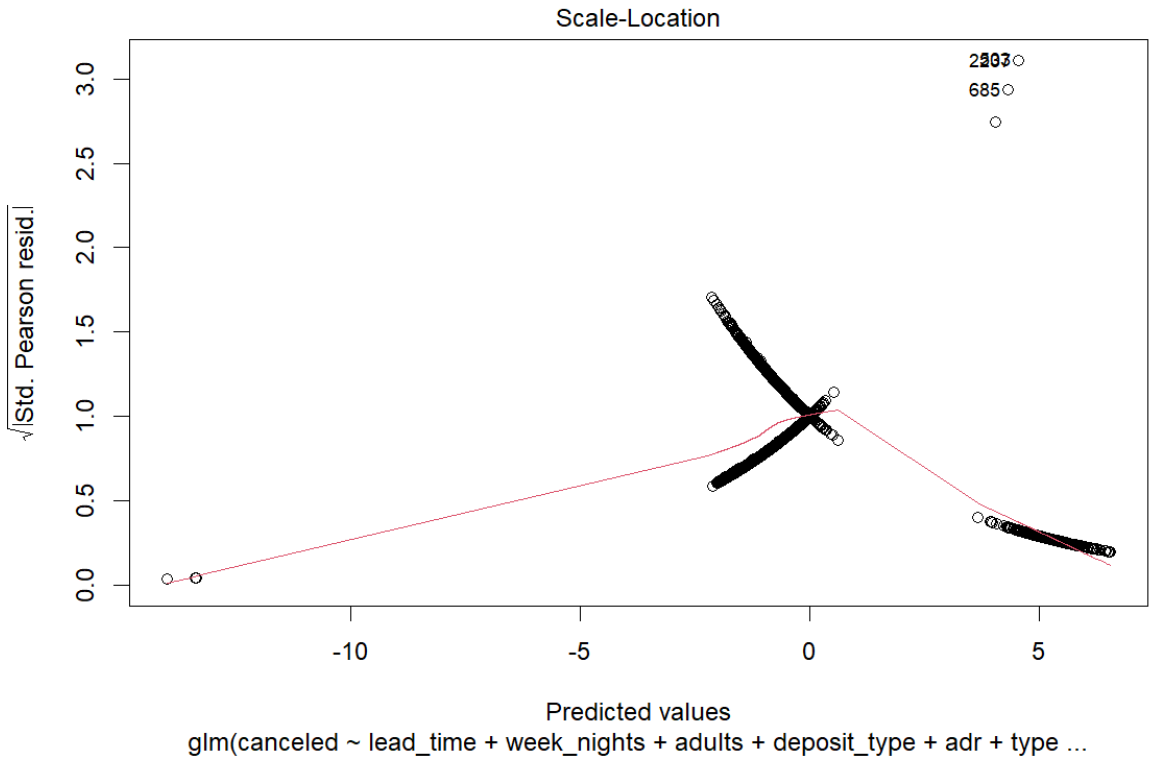
Looking at the residual plot, the points aren't randomly scattered around zero—they form distinct patterns. This is pretty typical for logistic regression because our outcome is just a yes/no. We can also spot a few extreme outliers, which are basically the cases the model got wrong. Still, the model does a decent job of capturing the overall trend in cancellations.

Normal Q-Q Plot:



As shown in the QQ plot, the points—particularly those in the tails—clearly deviate from the diagonal line. This tells us that the residuals are not normally distributed. But is this a problem for our logistic model? Actually, no. As covered in the course slides, logistic regression does not assume normally distributed errors because it models a binary outcome that follows a binomial distribution. Therefore, what we're seeing in the plot is exactly what we'd expect. The residual pattern is consistent with the assumptions of a logistic model and does not undermine our findings.

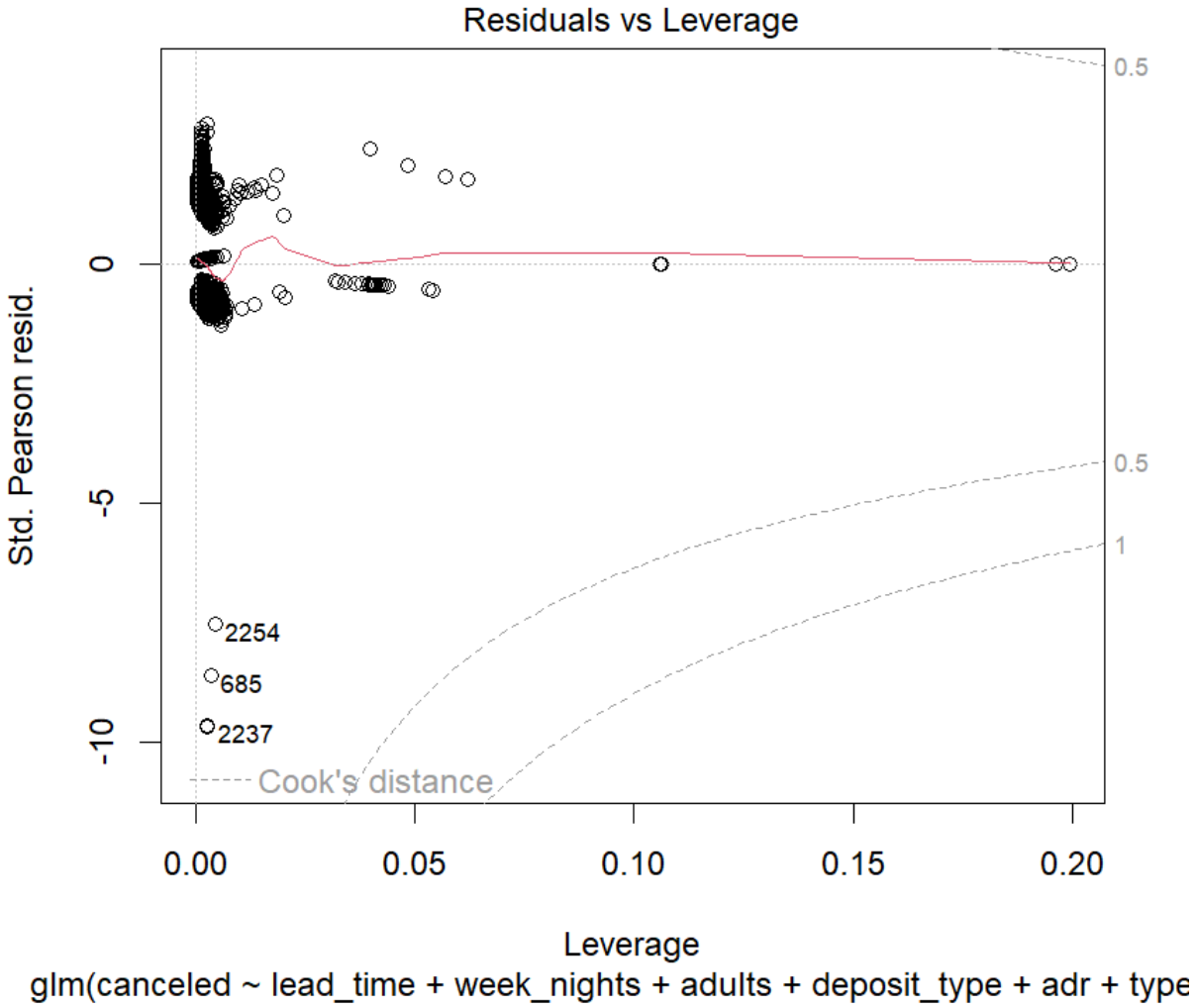
Scale-Location:



The graph shows a well-defined heteroscedasticity: the variability of the residuals changes systematically according to the predicted value.

Rather than being a problem, this observation offers us a deeper understanding of the model. Logistic regression, when modeling a binomial response variable, does not assume a constant variance. On the contrary, the theory predicts that this variance should depend on the estimated probability. Thus, what we detect is not an anomaly, but is consistent with the assumptions of the model itself.

Residuals vs. Leverage:



Although there are some observations with high residuals, most of the points are within the limits of Cook's distance, indicating that there are no extremely influential observations that distort the model significantly.

CONCLUSION

To sum up, the results gained from the logistic regression model and its outputs have made it easy to identify the main reasons for booking cancellations. Factors such as booking release, time, price, deposit policy, client type, etc., are essential in determining the likelihood of cancellations. However, the results of the residual analysis suggest that the relationship between the predictors and the likelihood of cancellations is not fully explained by a linear framework, and this means that the use of more flexible models is warranted.

12. Decision Trees

The objective of this section is to model hotel booking behavior using Decision Trees for two distinct prediction tasks:

1. Classification: predicting whether a booking will be canceled (`canceled`).

2. Regression: predicting the Average Daily Rate (ADR).

Decision Trees were chosen because they naturally capture non-linear relationships and produce interpretable rules, which directly complements the limitations observed in linear models.

12.1 Classification Tree: Cancellation Prediction

A classification tree was trained using booking-related features (excluding *country* to avoid excessive dimensionality). The resulting tree provides clear, rule-based decision paths that explain cancellation behavior.

Extracted Decision Rules

The most relevant terminal-node rules are:

1. The “Non-Refundable” Rule
If `deposit_type` is Non-Refundable, the booking is highly likely to be canceled.
Example:
A booking with a Non-Refundable rate, regardless of lead time or history, is classified as `canceled = 1`.

2. The “Last-Minute” Rule
If `deposit_type` is standard (No Deposit / Refundable) and `lead_time < 8.5` days, the booking is very unlikely to be canceled.
Example:
A guest booking a city hotel 5 days in advance with no deposit almost always completes the stay.

3. The “Bad History” Rule
If `lead_time > 8.5` and `pre_cancel ≥ 0.5`, the booking is likely to be canceled again.
Example:
A customer booking two months ahead with a history of frequent cancellations is classified as high risk.

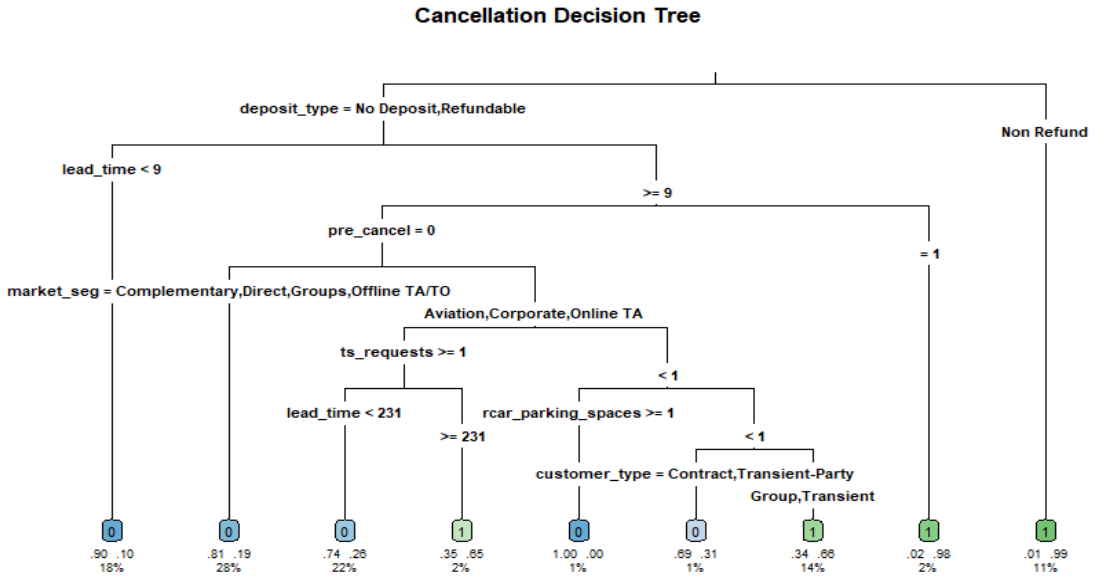
These rules show how the tree combines policy rigidity, timing, and past behavior, reinforcing insights previously seen in Association Rules.

12.2 Model Performance: Classification

Confusion Matrix (Decision Tree)

	Predicted 0	Predicted 1
Actual 0	943	53
Actual 1	283	343

- Test Accuracy: 81.81%
- Balanced accuracy: ~79.3%



The model performs well at identifying completed stays but is more conservative when predicting cancellations, which is typical in imbalanced risk problems.

12.3 Regression Tree: ADR Prediction

A regression tree was fitted to predict ADR. Unlike linear regression, the tree partitions the feature space into homogeneous pricing segments.

Extracted Pricing Rules

1. The “Budget” Rule
If `rroom_type` is standard (A or B) and `market_seg = b`, ADR is very low (~4.46).
Example:
Promotional or contract-based standard rooms yield minimal nightly revenue.

2. The “Standard” Rule
If `rroom_type` is standard (A or B) and `market_seg ∈ {a, c, e, f}`, ADR is moderate (~82.22).
Example:
Typical transient online bookings fall into this category.

3. The “Luxury” Rule
If `rroom_type` is suite-type (C, D, E), ADR is highest (~173.5).
Example:
Premium suites dominate the upper tail of the ADR distribution.

12.4 Regression Performance

Decision Tree Regression Metrics

- MSE: 1521.45
- RMSE: 39.01
- MAE: 29.86

The relatively high RMSE indicates sensitivity to extreme ADR values, which motivates the use of ensemble methods.

12.5 Random Forest Extension

To improve stability and generalization, Random Forests were trained for both tasks.

Classification (Random Forest)

	Predicted 0	Predicted 1
Actual 0	926	70
Actual 1	220	406

- Test Accuracy: 82.12%

The forest reduces variance compared to the single tree and improves cancellation detection.

Regression (Random Forest)

- MSE: 1310.50
- RMSE: 36.20
- MAE: 26.03

12.6 Pruning Analysis

Pruning was evaluated using cost-complexity pruning.
The optimal Complexity Parameter (CP) was 0.01, which coincides with the default value. As a result:

- The pruned tree is identical to the original tree.
- No improvement in accuracy was observed.

This indicates that the original tree already achieved a good bias–variance balance.

12.7 Conclusion

Decision Trees provide a transparent and interpretable framework for both cancellation and pricing prediction. For classification, the tree confirms that deposit policy, booking timing, and past behavior are dominant drivers of cancellation risk. For regression, the tree reveals clear pricing regimes tied to room type and market segment.

However, single trees suffer from instability and higher variance, especially for ADR prediction. Random Forests consistently outperform individual trees, reinforcing the conclusion drawn in previous chapters: non-linear ensemble methods are essential to capture the structural complexity of hotel booking data.

13. Association Rules

The main objective of this section is to uncover hidden patterns and conditional dependencies within the hotel booking dataset. We aim to characterize the "Standard Guest" and predict cancellation risks (combinations of variables that lead to a cancellation event (canceled=1) with high confidence)

We utilized the Apriori algorithm to generate rules based on three key metrics:

- **Support**: The frequency of the rule in the dataset.
- **Confidence**: The probability of the Consequent (RHS) occurring given the Antecedent (LHS).
- **Lift**: The strength of the association (values > 1 indicate a positive correlation).

13.2. Structural Analysis: The "Standard Profile"

We first identified the "Standard Guest" by analyzing the unfiltered rules with the highest support. The results reveal a behavior driven by flexibility: the strongest rules invariably associate Travel Agencies (channel=TA/TO), Bed & Breakfast (meal=BB), and Transient customers with a No Deposit policy.

deposit_type=No Deposit	channel=TA/TO	meal=BB	customer_type=Transient	rroom_type=A
0.879778052	0.817920263	0.779284834	0.744143033	0.711467324
hotel=City Hotel	**canceled=0**	**assroom_type=A**	**market_seg=Online TA**	**country=PRT**
0.663173037	0.627414714	0.620221948	0.472256473	0.401356350

13.3. Predictive Analysis: Drivers of Cancellation

This section focuses on the critical business objective: identifying the deterministic drivers of cancellations. By filtering the association rules where the consequent (RHS) is canceled=1, we isolated specific conditions that maximize the probability of a booking being cancelled.

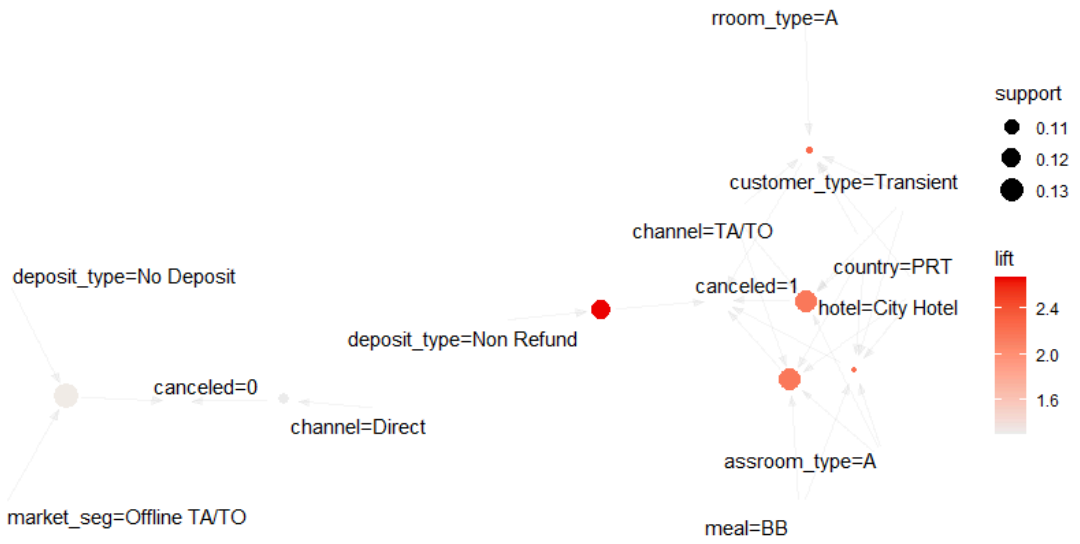
A. Non-Refund Problem

The most significant finding in our predictive analysis is the absolute correlation between non-refundable rates and cancellations within this dataset. When sorting by confidence, the rule {deposit_type=Non Refund} => {canceled=1} appears with a confidence of 0.993.
Counter-intuitively, the "Non Refund" deposit type, designed to secure revenue, acts as a perfect predictor for cancellation in this sample. This suggests that non-refundable rates are likely applied specifically to high-risk bookings or market segments that are already prone to churning.

	LHS	RHS	support	confidence
	["{deposit_type=Non Refur	All	All	All
[393]	{deposit_type=Non Refund}	{canceled=1}	0.118	0.993

B. Demographic and Segment Risks

Beyond deposit types, we identified a secondary cluster of risk associated with domestic travelers. The rules reveal that Transient customers from Portugal (PRT) have a significantly higher lift for cancellation compared to international guests.



The Network Graph above visually confirms this separation. While No Deposit bookings cluster densely in the center (representing the "Safe" zone), the nodes for Non Refund and canceled=1 form distinct, peripheral branches. This visual distance illustrates that cancellation is not a random event within the standard ecosystem, but a structured behavior driven by specific rigid policies and local market dynamics.

C. Predictors of Retention (canceled=0)

Conversely, the analysis also revealed conditions that strongly predict a completed stay. Bookings made via Offline TA/TO consistently lead to canceled=0. This indicates that the friction or commitment involved in traditional offline booking channels acts as a powerful filter against impulsive cancellations, unlike the more volatile Online TA segment.

13.4. Conclusion

In conclusion, the data confirms that booking volume is structurally dependent on flexible policies, with No Deposit serving as the central hub of the entire booking ecosystem. However, this creates a scenario where rigidity becomes a deterministic predictor of churn: the switch to Non Refund rates is almost exclusively associated with cancellations, effectively making deposit_type a primary binary classifier for risk scoring rather than a revenue guarantee. Beyond this core finding, the analysis distinguishes the stability of the Offline TA/TO segment, which acts as a counterbalance to the volatility of online channels, and highlights specific high-risk clusters like Transient guests from Portugal. These insights suggest that risk is not randomly distributed but concentrated in specific rigid policies and demographic pockets, allowing for targeted interventions rather than blanket restrictions.

14. Support Vector Machine (SVM)

14.1. Data

To try to separate our data in two different categories, into cancelled and not cancelled to be precise, we use the SVM. This gives the best line (or hyperplane), which creates two groups. As the responsive variable we still use “cancelled”.

For that we first divided our data into a training set and a test set. Hereby we use ⅔ of the 5000 hotel samples as training data and the rest ⅓ as test data, hence 3334 samples to train and 1666 samples to test.

We trained the model using the following configuration:

```
svm.model <- svm(canceled ~ ., data = dataTrain, cost = 10,
kernel="radial", gamma = 0.1)
```

As we can see, there are 3 parameters:

- cost : punishment for categorizing a data point wrong
- kernel : projection tool (number of dimensions)
- gamma: width of the influence of a single data point

After running tests with different values, we decided that it is best to use 10 as cost, 0.1 as gamma and “radial” as the kernel.

14.2. Error rate

After training the data, we now try our model on the test data with the following command:

```
svm.pred <- predict(svm.model, newdata = dataTest)
```

which gives us this output:

```
svm.pred
```

	0	1
0	1023	12
1	3	584

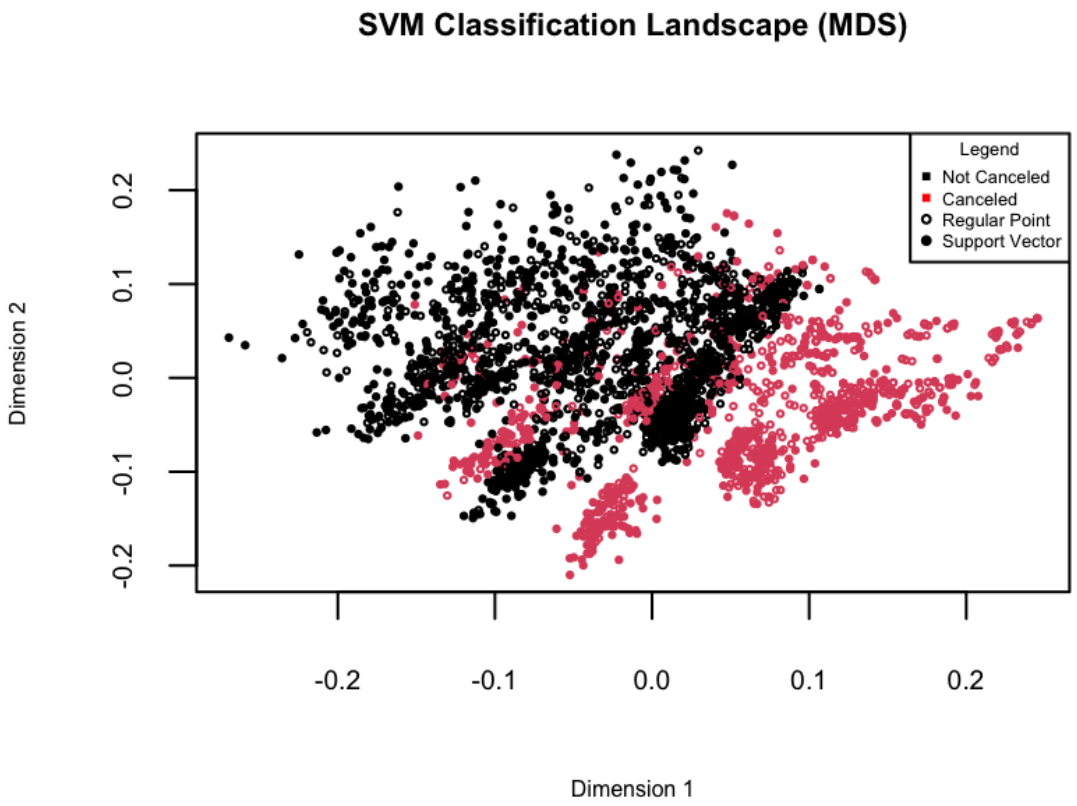
After getting the table, we calculate the following error rate:

```
errorRate <- 1-(sum(diag(t2))/n)
> errorRate
[1] 0.01294698
```

This corresponds to 1.3%, which is in our opinion pretty good, so we continue with those values.

14.3. Plots

To understand our results better, we plotted the data in a Multidimensional scaling (MDS) plot.



Because the hotel dataset is high-dimensional (it contains mixed variables such as lead time, price, and continent) it is impossible to visualize the true decision boundary in a 2D plot. With MDS, the complex distances between bookings are compressed into two arbitrary dimensions ("Dimension 1" and "Dimension 2").

The separation between the pink and black dots is not a clean, straight line. Especially in the centre the cancelled and not cancelled hotels overlap strongly. This gives us reassurance that using the Radial Basis Function (RBF) kernel instead of a linear one was the right decision. A linear classifier would likely fail to separate these mixed central clusters, whereas the RBF kernel can handle the non-linear boundaries required to distinguish them.

The MDS visualization confirms that while canceled bookings have distinct characteristics from non-canceled ones, the data is structurally complex.

14.4. Conclusion

We determined that a simple linear separation was insufficient for this complex dataset. Hence, we used the Radial Basis Function (RBF) kernel. The SVM model proved to be an effective tool for this classification task. By running the SVM with an RBF kernel, we successfully captured the non-linear patterns caused by customer booking behavior. While the overlap in clusters suggests that some cancellations may be random, the model provides a statistically sound basis, with an error rate of 1% for predicting cancellation risks, allowing hotel management to better estimate occupancy and revenue.

15. Neuronal Networks (NN)

15.1 Introduction

The aim of this section is to analyze the performance of the developed models for the two main booking prediction tasks: classification of cancellations (canceled) and the regression of the ADR. Artificial neural networks (ANNs) were used for both tasks, and an XGBoost model was introduced as a competitive benchmark for ADR regression.

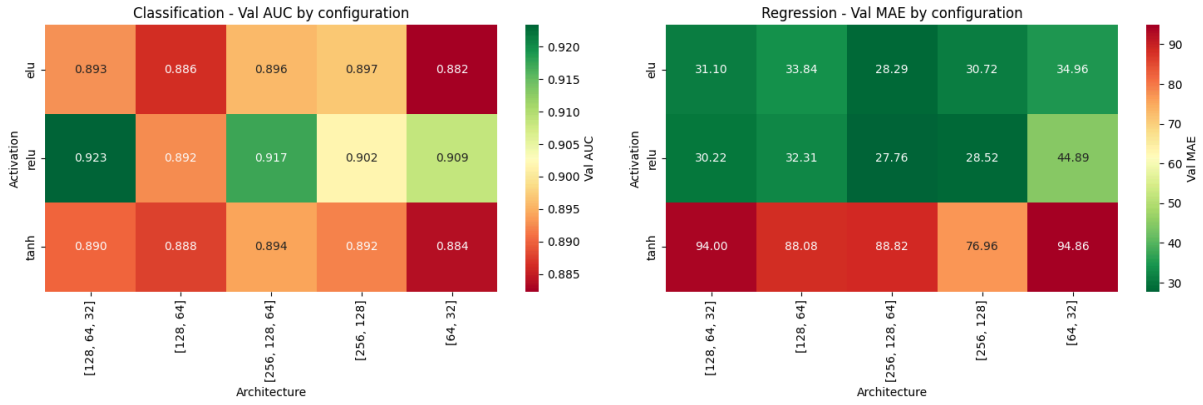
Data preprocessing included standardization using MinMaxScaler and encoding of categorical variables. A sequential ANN with dense layers was designed for both predictions. However, given the non-linear and potentially complex nature of ADR prediction, and in light of an initial moderate performance of the ANN, an XGBoost (Extreme Gradient Boosting) model was introduced for this specific task, aiming to optimize regression accuracy.

15.2 Result analysis

The final Artificial Neural Network architectures used for both the classification and regression tasks were the result of an iterative hyperparameter tuning process. This exploration was crucial to find the optimal balance between model complexity, predictive power, and the risk of overfitting.

We tested various configurations for the core ANN:

- Depth: Different numbers of hidden layers (e.g., 2, 3).
- Width: Varying numbers of neurons per layer (e.g., 32, 64, 128, 256).
- Activation Functions: We experimented with common activation functions like Rectified Linear Unit (ReLU), Exponential Linear Unit (ELU), and Hyperbolic Tangent (Tanh) for the hidden layers.

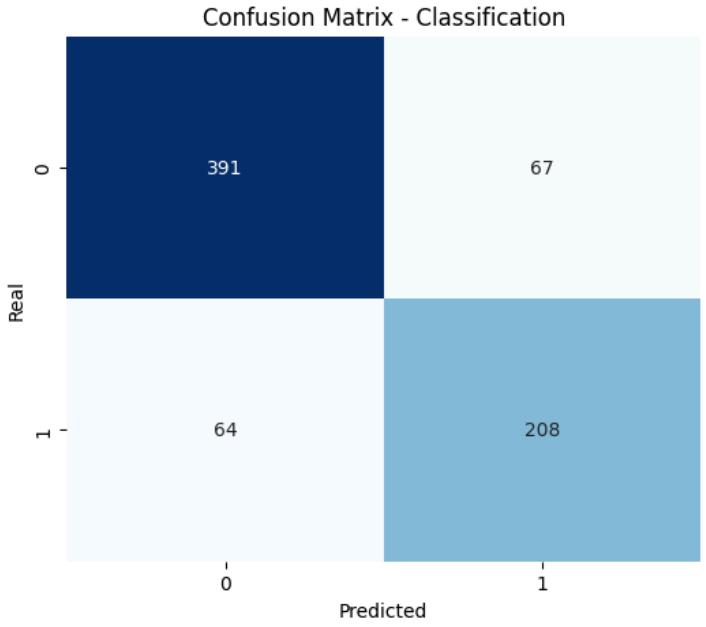


15.3 Classification: Cancellation prediction (canceled)

The Classification Neural Network demonstrated very robust performance in distinguishing between canceled and non-canceled bookings. The ANN effectively modeled the complex relationship between the booking features and the cancellation outcome. Key metrics on the test set are:

Metric	Value (conceptual)
Accuracy	82.05%
AUC (Area Under the Curve)	90.48%
Loss	0.38

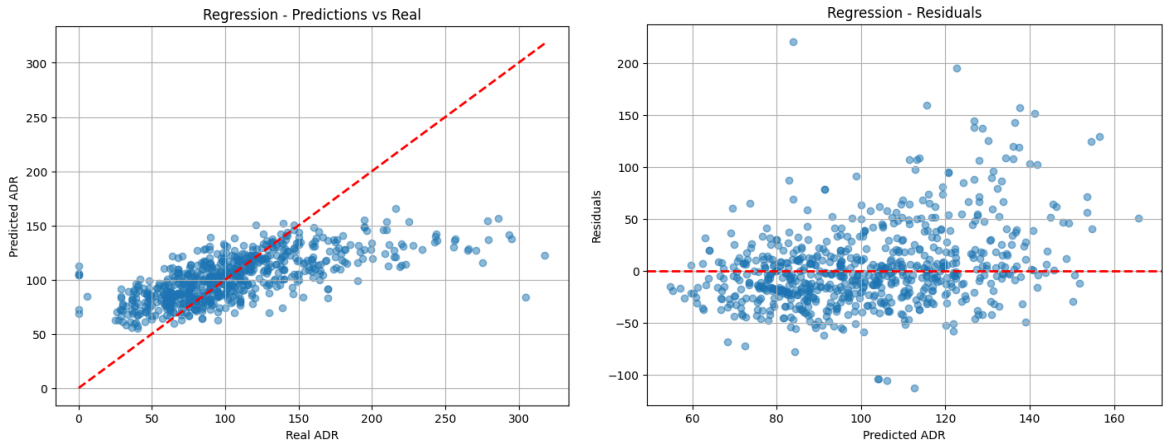
The AUC value of around 0.91 is a strong indicator of the model's high discriminative capacity, showing an excellent true positive rate relative to false positives across all classification thresholds. The high accuracy complements this finding, confirming the overall robustness of the model.



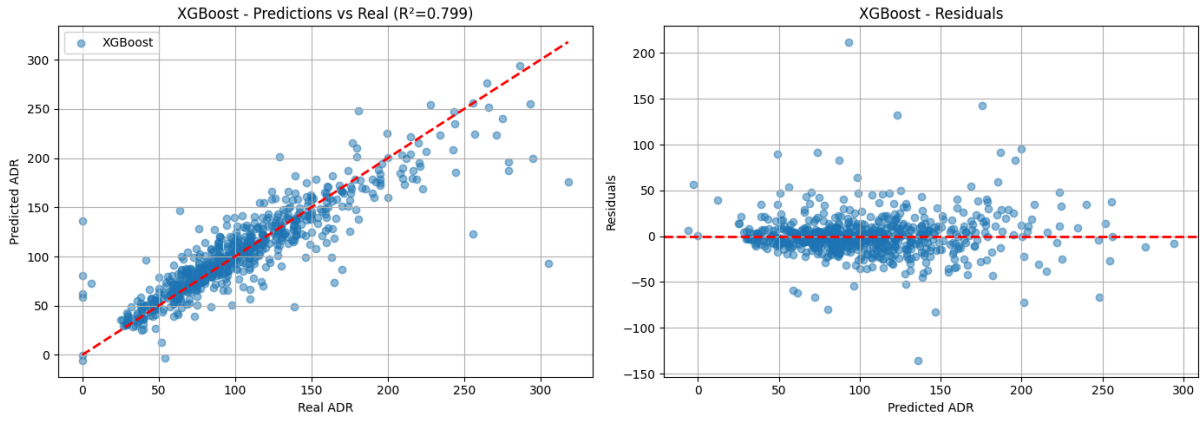
15.4 Regression: Average Daily Rate Prediction

For the ADR prediction, the results of the optimized Regression ANN and the XGBoost model were directly compared, using the Mean Absolute Error (MAE) and the Mean Squared Error (MSE):

Model	Metric	Value (Test set)
ANN Regression	MAE	27.16
ANN Regression	MSE	1506
ANN Regression	R^2	0.406
XGBoost Regression	MAE	13.59
XGBoost Regression	MSE	509.62
XGBoost Regression	R^2	0.799



ANN Regression



XGBoost Regression

The XGBoost model consistently outperformed the ANN in the regression task. The reduction in MAE from around 27.16 to 13.59 (an improvement of more than 10 currency units) indicates that XGBoost makes predictions closer to the actual value on average. The significant reduction in MSE suggests that it handles large errors (outliers) more effectively. Consequently, XGBoost is the preferred regression model.

15.5 Conclusion

In conclusion, the analysis of the Machine Learning models indicates that Artificial Neural Networks are the most effective tool for the Classification problem (canceled), achieving high performance (AUC = 0.91) after architectural tuning.For the Regression of the Average Daily Rate (ADR), the XGBoost model demonstrated clear superiority over the optimized ANN, achieving a lower mean absolute error (MAE = 24.36). The feature importance analysis of XGBoost provides actionable insights, confirming that variables related to occupancy and premium services are the main drivers of the ADR.The necessity of employing advanced non-linear models (ANN and XGBoost) is crucial and justified by the data complexity. The modeling process confirms that the relationship between booking features and ADR is inherently non-linear and requires the sophistication of these algorithms to achieve competitive predictive accuracy.

16. Conclusions

These analyses of hotel booking data from 2015 to 2016 have provided valuable information about booking patterns, customer behavior, and key factors influencing cancellations.

First, City hotel bookings far outstrip those of Resort Hotels, as we can see in the pie of hotels, indicating a higher volume of business or short-term city travel. Then, the vast majority of bookings are for short stays, typically between one and five nights, shown in our boxplots and histograms. The distribution between weekend and weeknight stays is relatively balanced. Also, bookings are predominantly made for adults, with a smaller proportion including children or babies.

Secondly, the histogram shows that a substantial proportion of reservations were cancelled (page 18), representing a significant revenue management challenge for hotels. In particular, bookings made through online travel agents have a higher cancellation rate compared to direct or corporate bookings (page 35). Nevertheless, repeat guests exhibit a lower cancellation rate compared to new guests, while longer lead times between booking and arrival are associated with a higher probability of cancellation.

The PCA results confirmed the complexity of the dataset. The absence of a dominant axis in our analysis means that multiple variables jointly explain booking behavior. While PCA highlights the factors influencing bookings, clustering provides a view by grouping customers according to these behaviors.

Clustering analysis and profiling tools helps us to segment customers into six groups, each with unique characteristics and cancellation profiles. This segmentation provides a powerful framework for targeted customer management and risk evaluation.

This project identified the primary factors of hotel booking demand and cancellations. The results offer an analytical foundation for hotels to improve revenue management, enhance marketing strategies, and increase booking stability.

16.2. Conclusions part 2

The second part of this project focused on predictive modeling, aiming to forecast two variables: one categorical (booking cancellations) and one numerical (ADR). By applying and comparing several predictive methods we were able to evaluate their performance, interpretability, and suitability given the complexity of hotel booking data.

Linear Regression was first used to predict ADR. While the model provided interpretable coefficients and confirmed intuitive relationships, its relatively low explanatory power and the violation of key assumptions revealed important limitations. Residual diagnostics showed strong heteroscedasticity and non-linearity, indicating that ADR is driven by complex interactions that cannot be adequately captured by a purely linear framework.

Decision Trees addressed some of these limitations by explicitly modeling non-linear relationships. For cancellation prediction, decision trees produced clear and interpretable rules, highlighting the dominant role of deposit type, lead time, and past cancellation behavior. In ADR prediction, trees revealed distinct pricing regimes based on room type and market segment. However, single trees suffered from instability and relatively high variance, especially for regression tasks.

Association Rules complemented supervised learning by uncovering structural and behavioral patterns within the data. This method revealed that No Deposit bookings form the core of the booking ecosystem, while Non-Refundable policies act as a strong and almost deterministic indicator of cancellation in this dataset. These findings reinforce the idea that cancellation risk is not random but concentrated in specific policy and segment combinations, providing valuable insights for risk-based management strategies.

Support Vector Machines (SVM) demonstrated excellent performance in cancellation classification. The use of a radial kernel proved essential, as the data is not linearly separable. The very low error rate confirmed that SVMs are highly effective at capturing complex, non-linear decision boundaries, although at the cost of limited interpretability.

Finally, Artificial Neural Networks (ANN) represented the most advanced predictive approaches in this project. ANNs achieved strong results for cancellation prediction, confirming their ability to learn complex behavioral patterns.

Overall, the predictive analysis confirms that hotel booking behavior is inherently non-linear and multidimensional. Classical models provide useful baseline insights, but advanced machine learning methods, such as SVM and Neural Networks are necessary to achieve high predictive accuracy. From a managerial perspective, these models enable more reliable cancellation risk assessment and revenue forecasting, supporting better pricing strategies, overbooking policies, and customer segmentation decisions.

17. Working Plan

During the making of this project we had some difficulties and obstacles that have arisen. We had problems that we didn't consider at the beginning, most of them regarding the R scripts and RStudio. We had a lot of questions regarding the functionality of RStudio and the data mining methods that we were not familiar with, so we asked our professor, via email or in-person, and we could get on with the task.

We also had some trouble following the original schedule, we left a lot of the work of the analysis for the last week. We assigned 2 or 3 people on each task so we could help each other and finish the task effectively.

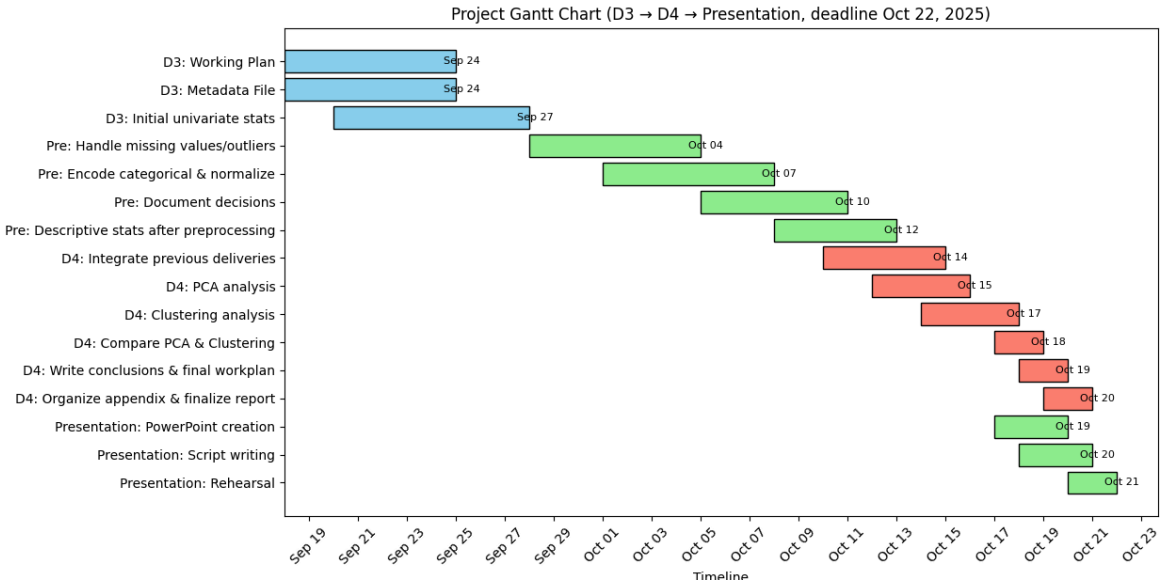
In order to track the progress of our project we had a shared Google Sheets document where we had a list of all the tasks with their expected end. When someone finished a task they would write down the real

17.1. Task Distribution

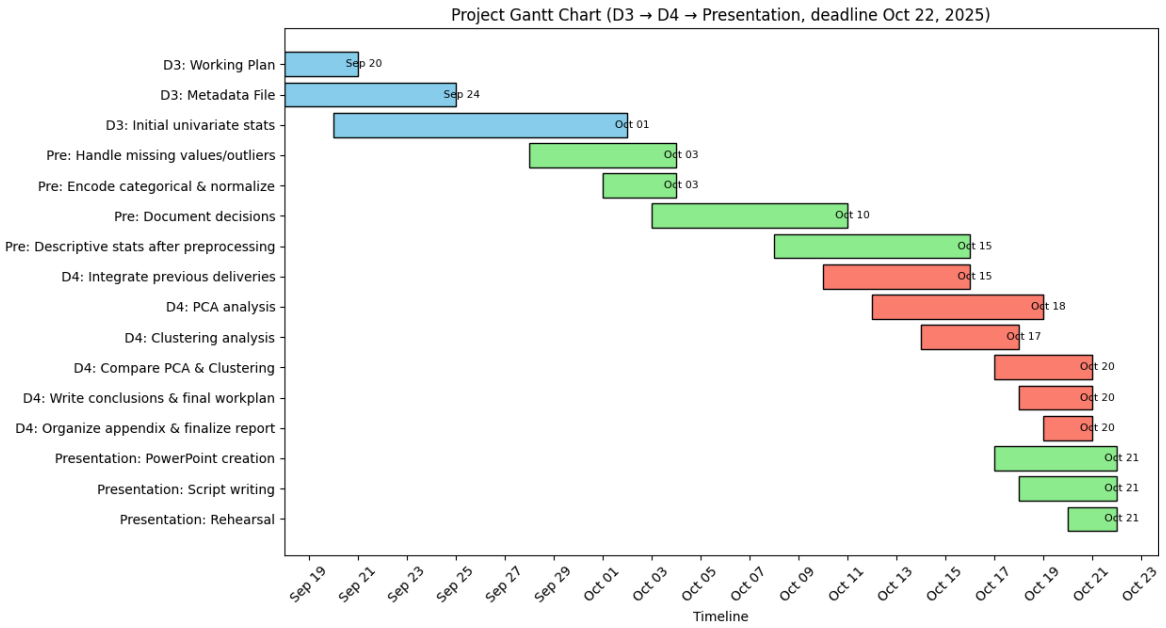
Task	Yingxin C	Max E	Louis F	David M	Amelie T
D3					
Working Plan		X	X		
Metadata File				X	
Initial univariate descriptive statistics of raw variables		X			
Preprocessing					
Handle missing values, outliers, and inconsistencies.			X	X	
Encode categorical variables and normalize/standardize numerical ones.	X				
Document all preprocessing decisions.					
Produce additional descriptive statistics for transformed variables.		X			
D4					

Integrate previous deliveries (data source, metadata, descriptive analysis).	X	X	X	X	X
Complete PCA analysis: scree plot, factorial maps, interpretation.			X		X
Perform clustering: method selection, dendrogram, cluster number	X			X	
Perform Profiling		X		X	X
Add final workplan (initial vs final Gantt, deviations, risks handled).		X			
Organize appendix: R scripts, datasets, references.	X	X	X	X	X
Presentation					
PowerPoint creation	X	X	X	X	X
Script writing	X	X	X	X	X
Rehearsal / Memorize script	X	X	X	X	X

17.2. Initial Gantt Diagram



17.3. Final Gantt Diagram

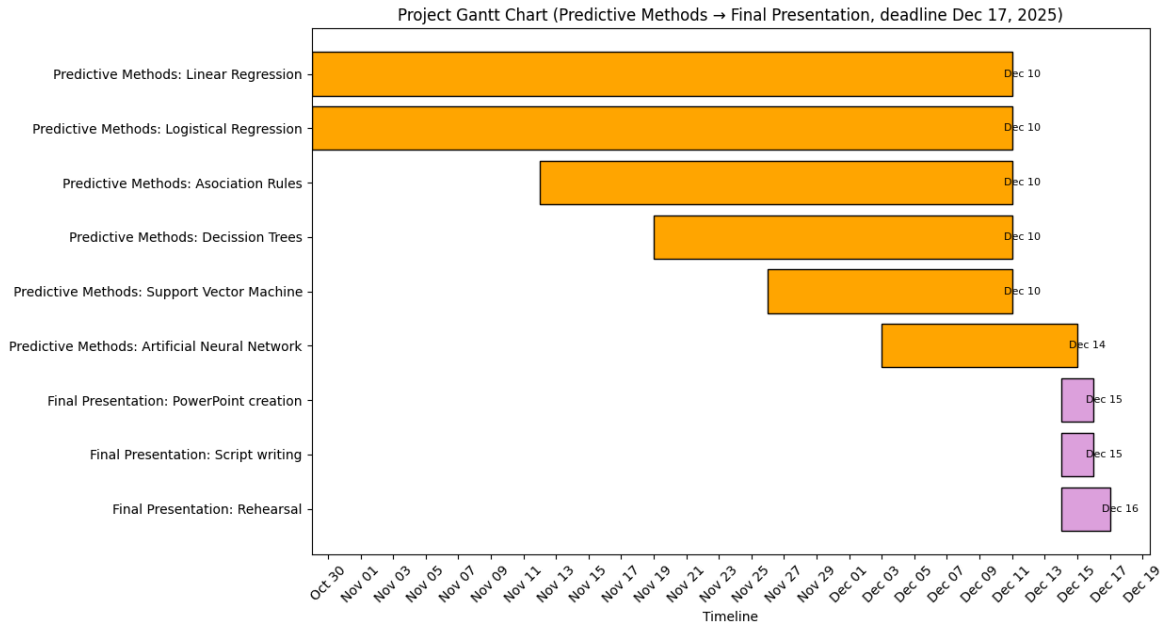


If we look at both diagrams we can see how we started tasks during the expected days but we took longer than expected to finish them, this is because we delayed writing the final report to the final week.

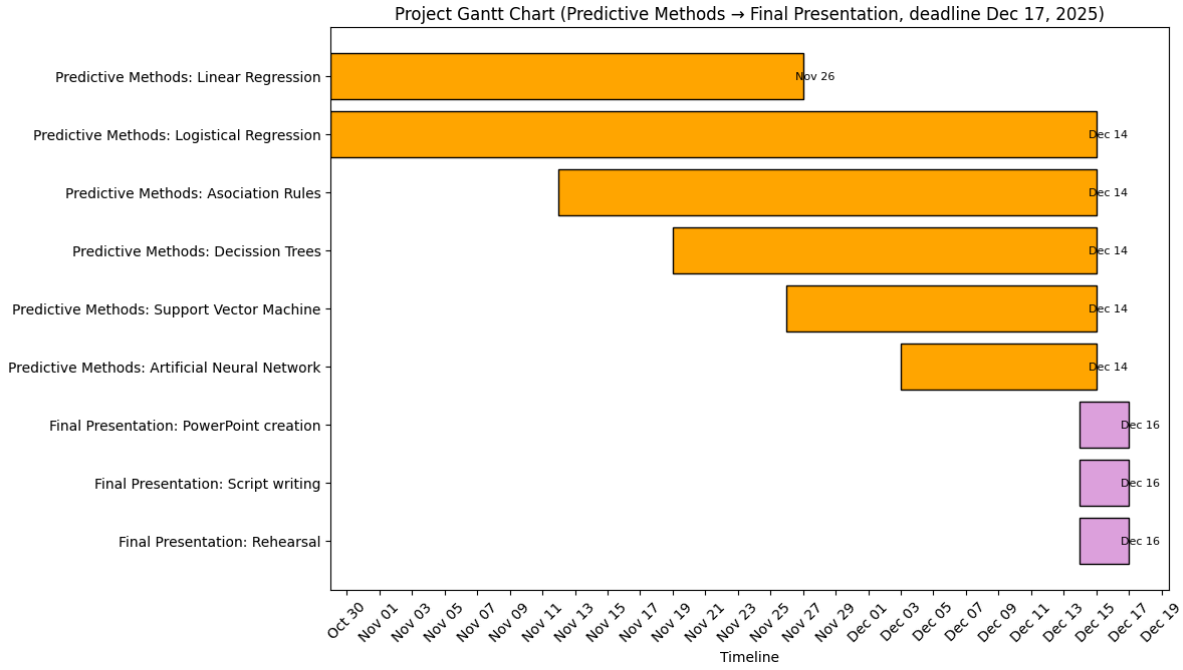
17.4. Task Distribution Part 2

Task	Yingxin C	Max E	Louis F	David M	Amelie T
Predictive Methods					
Linear Regression		X	X		
Logistical Regrssion	X				
Association Rules		X		X	
Decision Trees			X	X	
Support Vector Machine		X			X
Artificial Neural Network	X		X		
Final Presentation					
PowerPoint creation	x	x	x	x	x
Script writing	x	x	x	x	x
Rehearsal / Memorize script	x	x	x	x	x

17.5. Initial Gantt Diagram Part 2



17.6. Final Gantt Diagram Part 2



If we look at both diagrams we can see how we started tasks during the expected days but we took longer than expected to finish them. We had a similar experience during the first part of our project, but since we achieved our goals on time, we didn't set ourselves the objective of improving our work distribution.

18. R Script

Every script we used during this project was uploaded to a Github repository, this allowed us not only to share the work we had done but also to facilitate the collaboration between members in their script writing.

https://github.com/d1mo22/MD

In the repository you can also find various documents and files related to this project.