DATA MINING

Group 6 - Hotelbook

Gerard Álvarez, Tomás Calaf, Natalia Dai,
Adrià Espinoza and Mario Martín

23/10/2023

List of contents

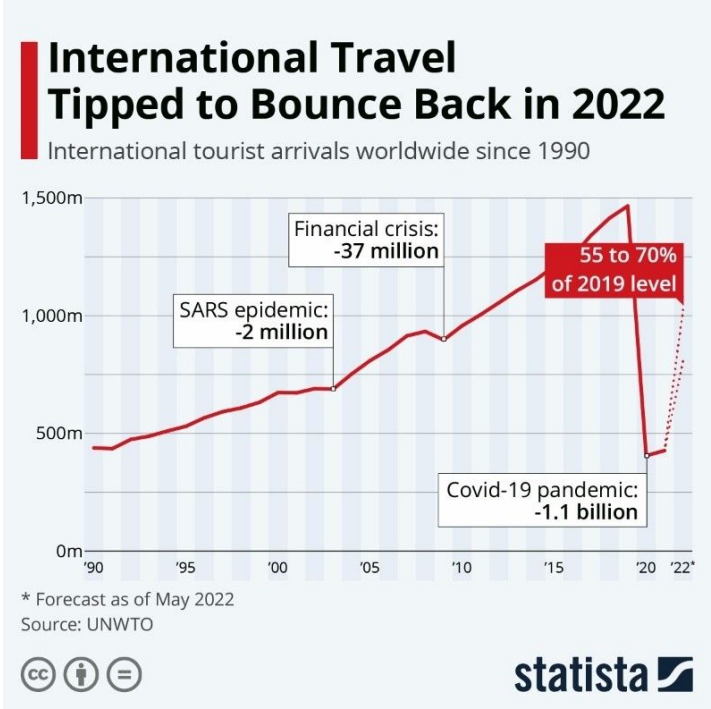
- Introduction
- Data mining process
- Descriptive analysis of variables
- Univariate descriptive analysis
- Preprocessing
- PCA
- Clustering
- Profiling
- PCA vs Clustering
- Conclusions
- Original vs Final scheduling

Introduction

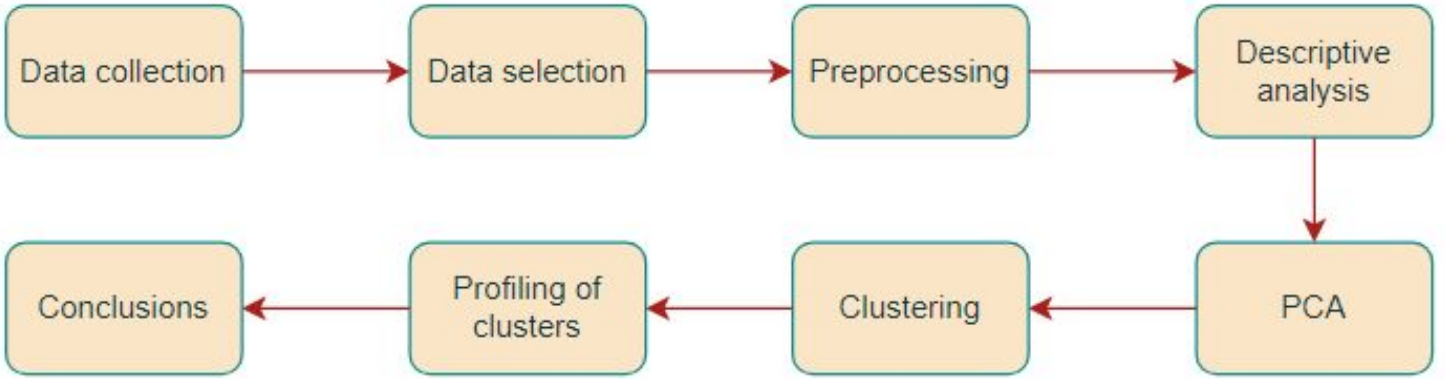
- Detect the trends among the tourists when they book an hotel
- Categorize all bookings into groups

kaggle

https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?select=hotel_bookings.csv



Data mining process

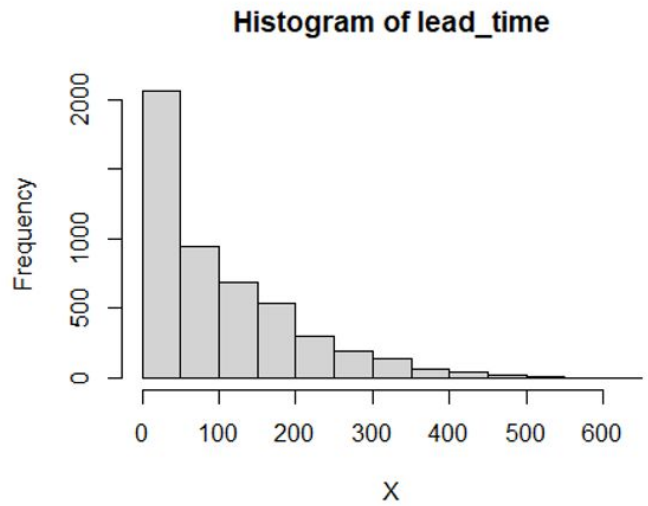


Descriptive analysis of variables

- hotel
- is_canceled
- lead_time
- arrival_date_month
- arrival_date_week_number
- arrival_date_day_of_month
- stays_in_weekend_nights; stays_in_week_nights
- adults; children; babies
- meal
- country
- days_in_waiting_list
- customer_type
- adr
- reservation_status

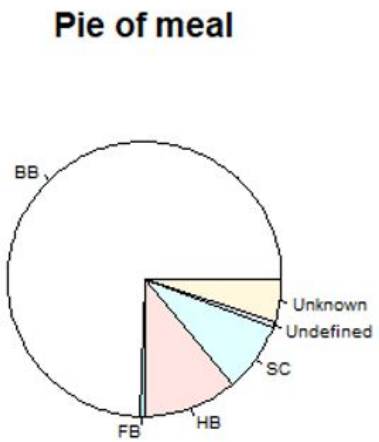
Univariate descriptive analysis

➔ Numerical



Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc
lead_time	0.00	18.00	72.63	101.30	155.00	629.00	101.47	1.00

➔ Categorical

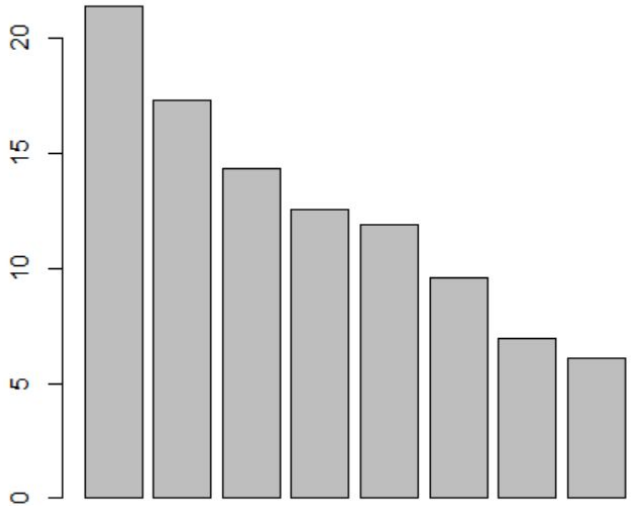


Preprocessing

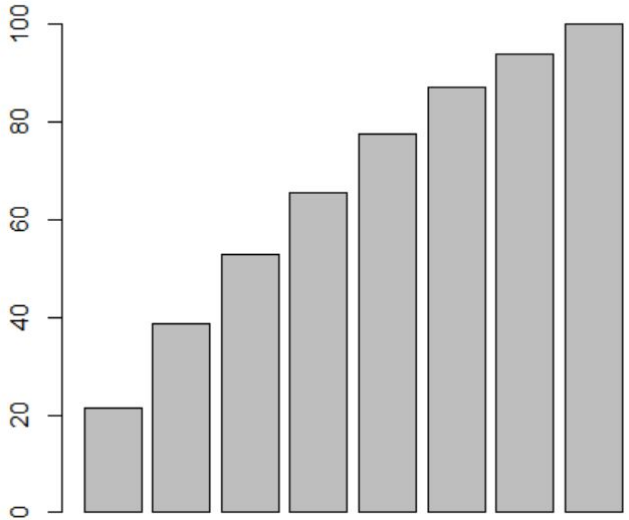
- ➔ Determining working matrix, filtering
- ➔ Creation of the missing values
- ➔ Identification and treatment of missing data
- ➔ Identification and treatment of outliers/errors

PCA

- 80% of information → 6 Dimensions
- PC2 & PC3 → 45º angle

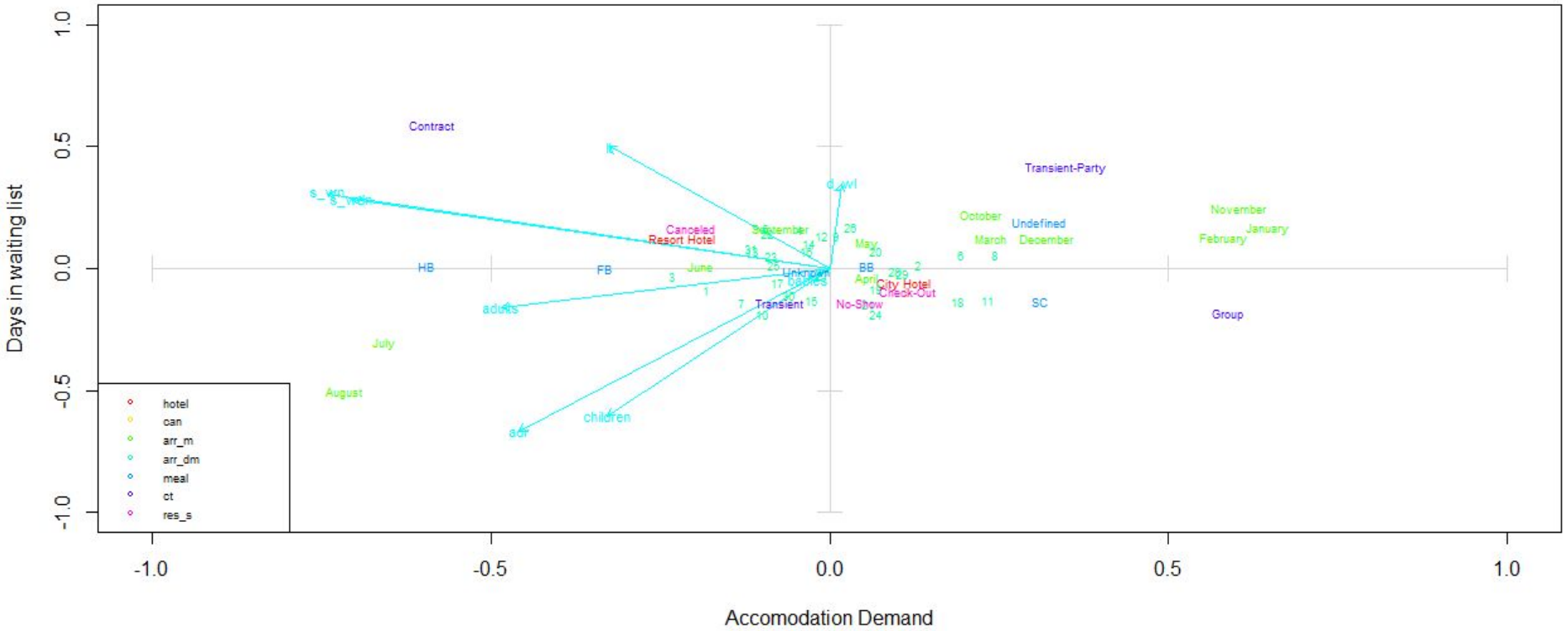


PCx information

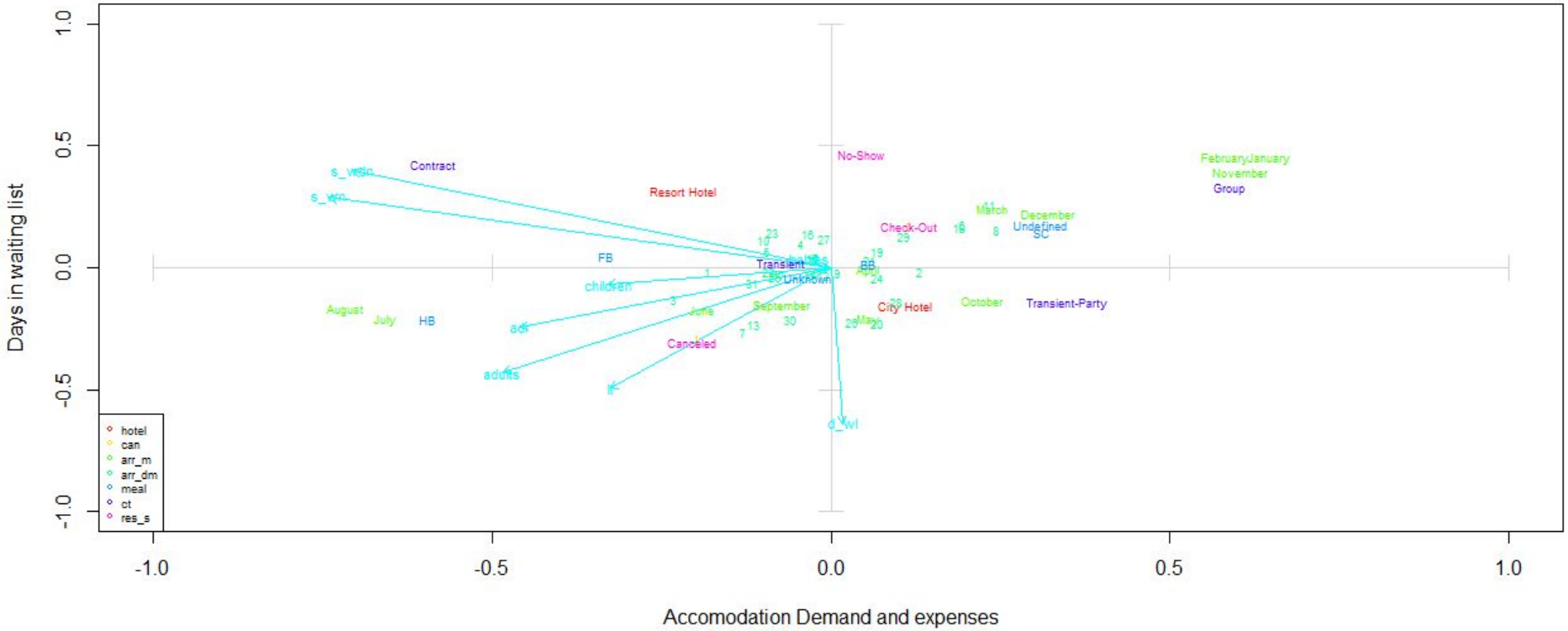


Total information

PCA - PC1 & PC2


Days in waiting list
Accomodation Demand
1.0
0.5
0.0
-0.5
-1.0
Contract
lt
Transient-Party
Canceled
Resort Hotel
HB
FB
adults
children
July
August
June
Transient
No-Show
City Hotel
Check-Out
BB
October
Undefined
March
December
November
January
February
SC
Group
hotel
can
arr_m
arr_dm
meal
ct
res_s

PCA - PC1 & PC3

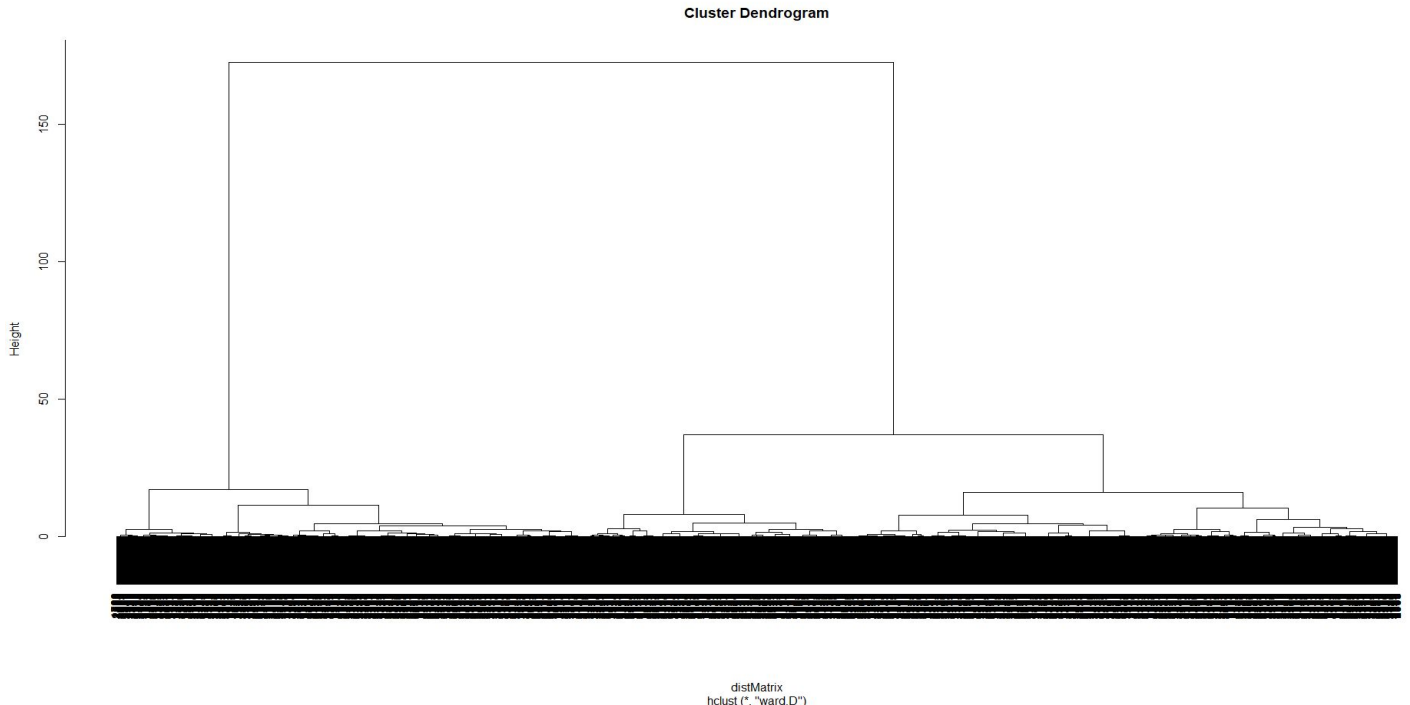

Days in waiting list
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
Accomodation Demand and expenses
hotel
can
arr_m
arr_dm
meal
ct
res_s
Contract
Resort Hotel
No-Show
FebruaryJanuary
November
Group
March
December
Undefined
Check-Out
FB
Transient
BB
children
Unknown
August
July
HB
June
September
City Hotel
October
Transient-Party
adr
adults
Canceled
d_wl

PCA - Conclusions

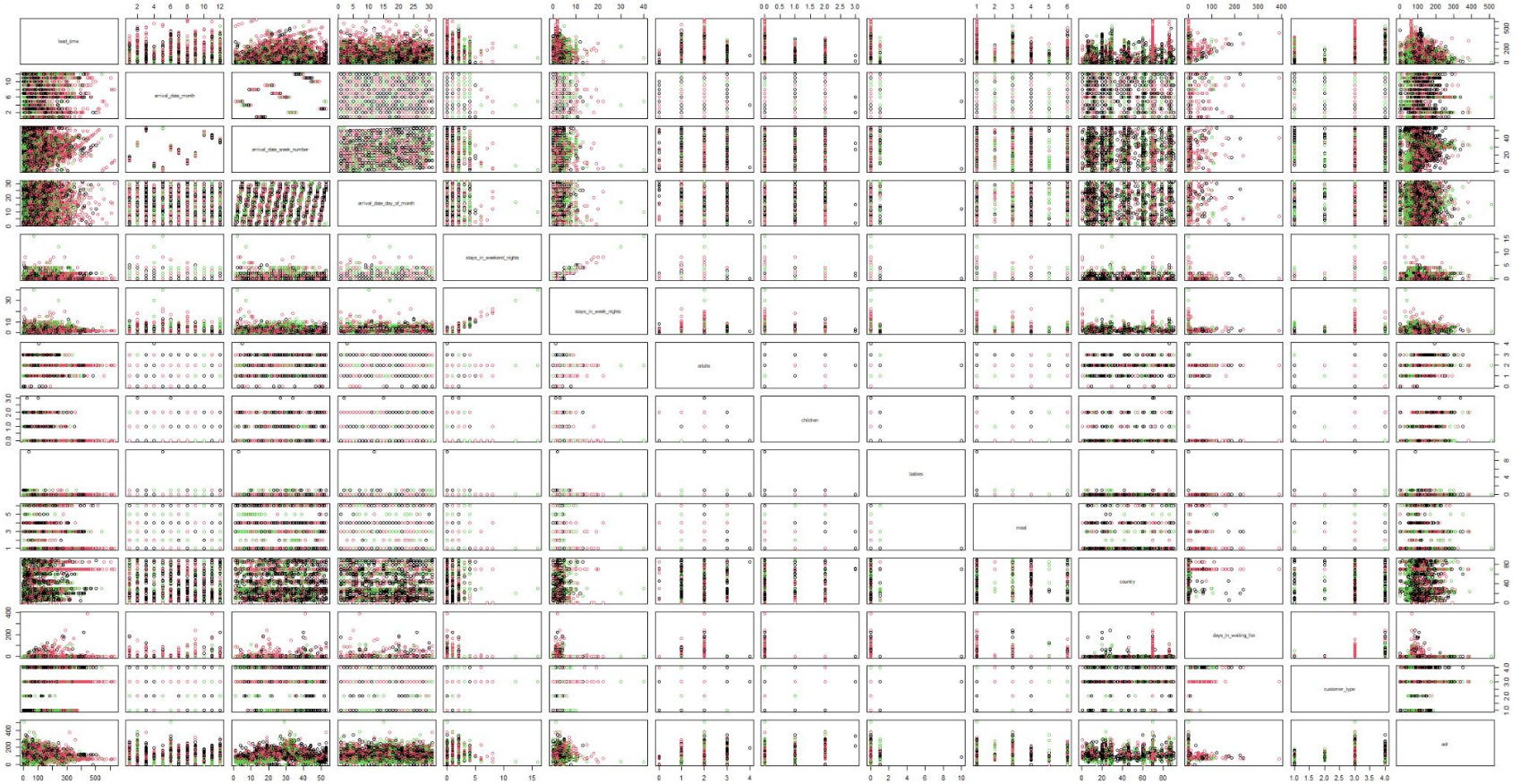
- ➔ Summer months
- ➔ The number of weekday nights and weekend nights and type of customer
- ➔ Contract > transient > transient-party > group

Clustering - Process

- First, we executed the script to obtain the dendrogram.
- We cut the dendrogram and obtained nc = 3.
- Overall class graphic for all the variables.

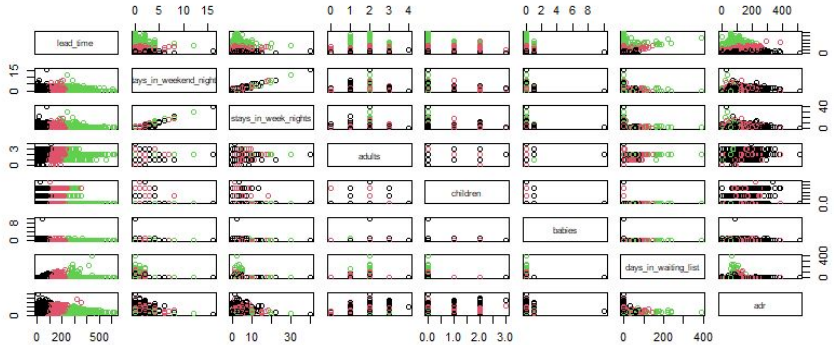


Clustering - Interpretation

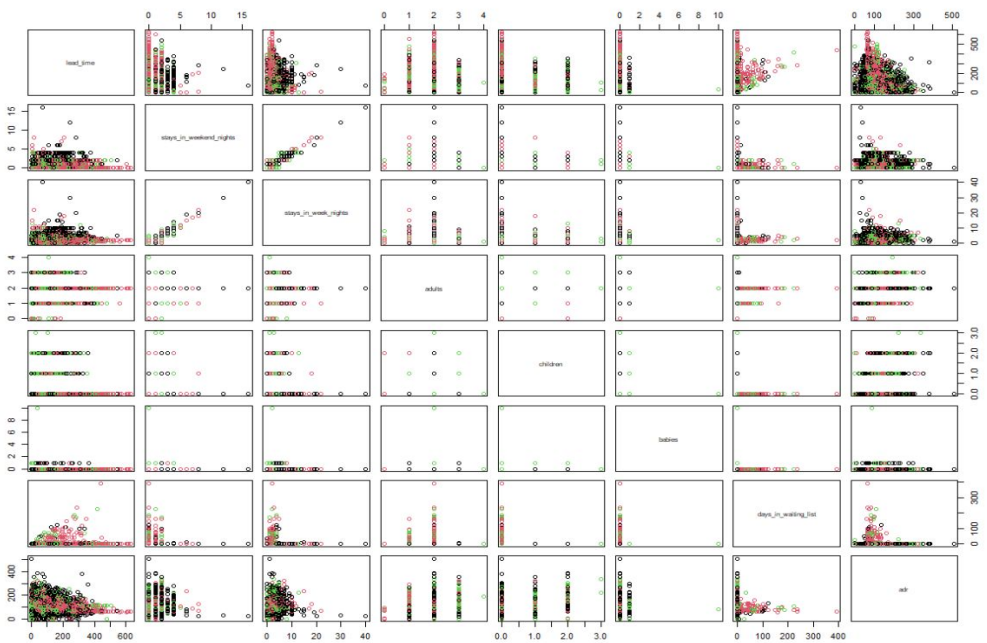


Pairplots using Gower distance

Clustering - Interpretation

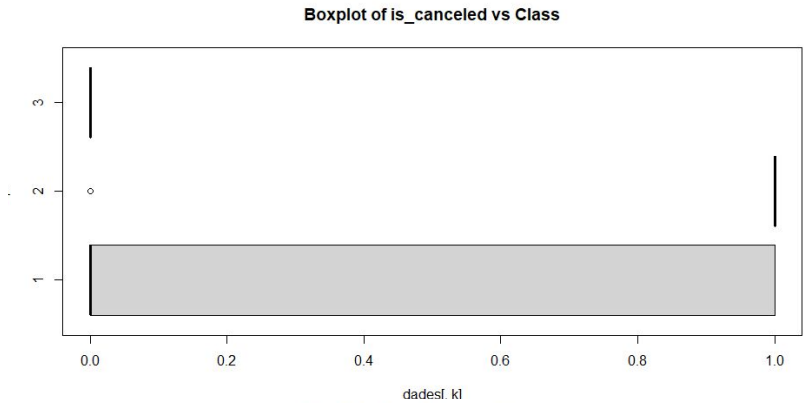


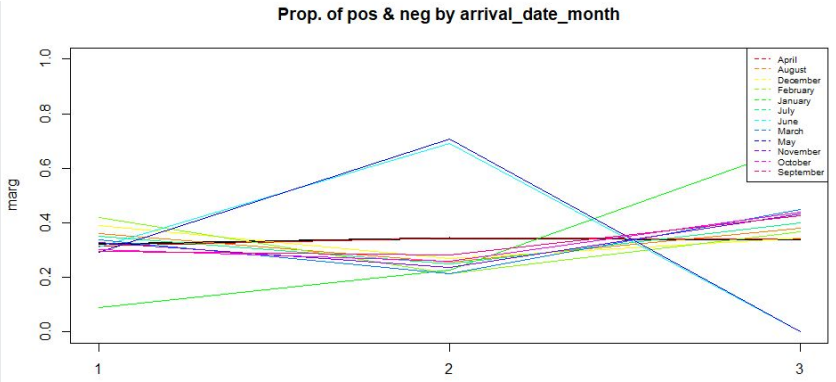
Pairplots using Euclidean distance

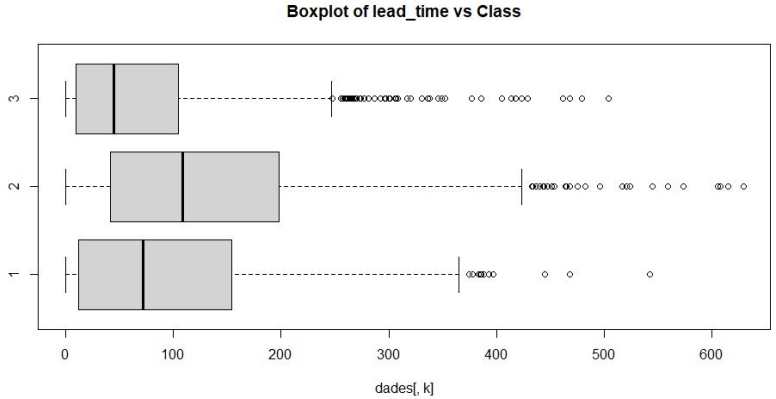


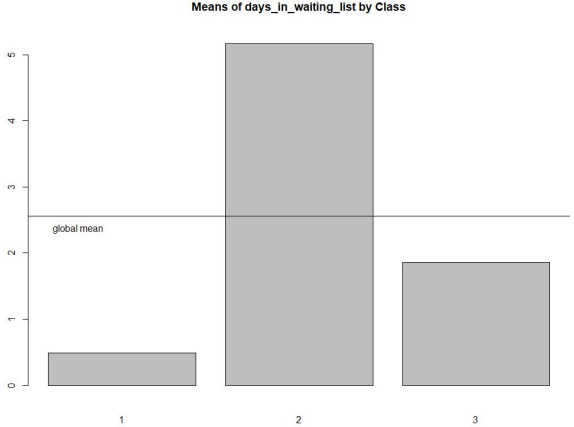
Pairplots using Gower distance
(only numerical values)

Profiling graphs









Final Class Profiling

- Class 1 and 2 have no noticeable separation.
- Class 2:
  - high “days_in_waiting_list” and summer values “arrival_date_of_month”
  - medium high “lead_time” and likely to be cancelled(“is_cancelled”)
- Class 2, with the majority of summer reservations, is the busiest time of the year.
- No further conclusions can be seen with class profiling.

Conclusions: PCA vs Clustering

- PCA: Travel patterns, customer types and seasonal reservations.
- Clustering: Prices fluctuations over the time.
- No clear results, no good data
- PCA better than clustering

PCA

Original vs Final scheduling

	September						October																					
Tasks	25	26	27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Assignment Grid																												
Gantt Chart																												
Risk Plan																												
Metadata file																												
Description of variables																												
Preprocessing Steps																												
Justification of preprocessing decisions																												
Descriptive of modified variables																												
Bivariate description																												
PCA																												
Hierarchical Clustering																												
Profiling of clusters																												
Conclusions																												
Presentation																												

Initial Gantt diagram

	September						October																					
Tasks	25	26	27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Assignment Grid																												
Gantt Chart																												
Risk Plan																												
Metadata file																												
Description of variables																												
Preprocessing Steps																												
Justification of preprocessing decisions																												
Descriptive of modified variables																												
Bivariate description																												
PCA																												
Hierarchical Clustering																												
Profiling of clusters																												
Conclusions																												
Presentation																												

Final Gantt diagram

Thank you

Time to answer your questions