Preliminary Operations

Jordi Tarroch Mejón

***Abstract*—Hotels have multiple ways to increase profits using data science but is not so clear how. Part of solving that issue starts by extracting and transforming the data offered by hotel's Property Management Systems to finally load it to each of the models created. In the process of extraction is also important to explore the data to understand how later this should be transformed and used for further research. The results obtained in this paper will be of vital importance to later apply Clustering and Prediction algorithms.**

***Index Terms*—matrix correlation, time series, data wrangling, exploratory data analysis, encoding, sampling, boxplot**

CONTENTS

I Introduction 1

II Data Wrangling 1
- II-A Common . . . 1
  - II-A1 Target Variable Definition . . 1
  - II-A2 Encoding of Categorical Variables . . . 1
  - II-A3 Imputation . . . 2
  - II-A4 Target Variable . . . 2
  - II-A5 Predictive Variables . . . 2
- II-B Specific to Clustering and Dimension Reduction . . . 2
- II-C Specific to Prediction . . . 2
  - II-C1 Logistic Regression for Reservations Cancellation Prediction . . . 2
  - II-C2 Time Series Analysis for Number of Reservations Prediction . . . 2

III Exploratory Data Analysis 3
- III-A Target Variable . . . 3
- III-B Relationships and Distributions . . . 3
- III-C Specific to Prediction . . . 3
  - III-C1 Matrix Correlation . . . 3
  - III-C2 Linear Relationship of Numeric Predictive Variables with IsCanceled . . . 3
  - III-C3 Contingency table of categorical variables . . . 3
  - III-C4 Time Series EDA . . . 3

IV Conclusion 3

V References 3

VI Annexes 4

Jordi Tarroch Mejón is with Data Science, CUNEF, e-mail: jordi.tarroch@cunef.edu

I. INTRODUCTION

In the Data Wrangling part: multiple extraction and transformation processes has been done in order to apply clustering, dimension reduction and prediction algorithms in further research.

In the Exploratory Data Analysis: multiple visualizations has been done to get a good understanding of the studied data.

II. DATA WRANGLING

A. Common

1) Target Variable Definition: IsCanceled column, column that represents the canceled bookings are defined as the following ReservationStatus values:

- A booking is canceled when IsCanceled equals to 1 and ReservationStatus takes either the values 'No-Show' or 'Canceled'.
- A booking is not canceled when IsCanceled equals to ' and ReservationStatus takes either the value 'Checkout'.

So ReservationStatus does not add any other value than an explanation of what a booking canceled means. Reason enough to erase the column ReservationStatus as will be explained further.

2) Encoding of Categorical Variables: In order to perform computations faster an encoding for the categorical variables has been done. Variables that will be used in all clustering, dimension reduction and prediction. Except for the exploratory data analysis.

a) Ordinal encoding: ArrivalDateMonth is a categorical variable associated to a certain order, reason why an ordinal encoding has been done.

Agent and Company are inherently ordinal encoded although a further treatment is done later for these variables.

b) Nominal encoding: The following variables have been encoded with a method called onehot encoding, which is a vector representation where all the elements of the vector are 0 except one, which has 1 as its value:

- ReservedRoomType
- AssignedRoomType
- Meal
- Country
- MarketSegment
- DistributionChannel
- DepositType
- CustomerType

c) Null to numeric: These two variables have some special features:

- Agent
- Company

First of all, they are categorical variables which due to its values are already ordinal encoded except the value 'NULL', value that will be encoded to 0.

Also, the fact that these are the only columns with 'NULL' values, it may give some clues that could be useful for the algorithms used. So two columns has been added as:

- AgentYes
- CompanyYes

Both of them take the value '1' if the booking has been done by an Agent or a Company and '0' if it hasn't been done by any of them.

3) Imputation:

a) Pre-Encoded: The previous encoded variables with its original values has been erased as its information has been already processed accordingly.

b) Common sense:

- ReservationStatus is erased for the reasons explained before, its information is already included in IsCanceled.
- ReservationStatusDate does not add any value to any of the goals of this study.
- ArrivalDateYear has been erased because a year is not repeatable and thus its predicting power is not applicable in the future.

4) Target Variable: Finally, the study starts with a good understanding of what the Target Variable IsCanceled means for the purpose of both, clustering and prediction. Variable that is separated from the rest.

5) Predictive Variables: Finally, the study starts with a data set for the predictive variables properly encoded and selected that will be used in clustering and prediction.

B. Specific to Clustering and Dimension Reduction

Due to the complex computations needed for clustering a sampling of the main data set is done to compute the different algorithms faster.

The data sets used for clustering contain:

- A sample of the Predictive Variables with numeric and categorical variables to calculate the dissimilarity matrix with Gower distances for most of the clustering techniques used.
- A sample of the Predictive Variables with only **numeric data** to calculate the dissimilarity matrix with Euclidean distances for the CLARA clustering technique.

TABLE I
PREDICTIVE NUMERIC VARIABLES

- ADR	- Children	- ArrivalDateOfMonth
- Adults	- DaysInWaitingList	
- TotalOfSpecialRequests	- StaysInWeekNights	- ArrivalDateWeekNumber
- Babies	- Leadtime	- PreviousCancellations
- BookingChanges	- PreviousBookingsNotCanceled	- RequiredCarParkingSpaces
- StaysInWeekendNights		

- A sample of the IsCanceled column to perform the contingency table and calculate the cancellation ratio of each group of the different clusters created by each clustering technique. It must have a very similar proportion of 'canceled bookings' and 'not canceled bookings' to the original data set given by the hotel.
- A table of couples of categorical variables to perform a correspondence analysis to optimize hotel room efficiency by studying the relationship between 'Room Type' - 'Country' and 'Room Type' - 'Distribution Channel'.

C. Specific to Prediction

1) Logistic Regression for Reservations Cancellation Prediction: After the common transformations, the prediction study starts with a data set that contains categorical and numerical data properly encoded.

a) Train - Test for GLM:: In order to train the model for the **Logistic Regression**, two sets of data extracted from the complete data set are built for the purpose of the logistic regression with a fixed seed to make it repeatable:

- Train set: a random uniform distribution sample made of the 80% of the original data set. This is used to fit the parameters of the model.
- Test set: the rest of the data left out of the training set. This is used to provide an unbiased evaluation of the final predictions given by the model.

If hyperparameters tuning was needed a validation set would be created, providing a similar evaluation as the test set but while tuning the hyperparameters. But at this point is not required.

Proportions of 'canceled bookings' and 'not canceled bookings' must be done to ensure a real representation of the data studied.

b) Train - Test for GLMNET:: In order to train the model for the **Regularization process**, train and test sets are needed but with target and predictive variables separated to fit the R function with its parameters.

- Hyperparemeters tuned for the Elastic Net:
  - Alpha equal to 1 to fit with a Lasso Model.
  - Lambda that produces the highest accuracy.

2) Time Series Analysis for Number of Reservations Prediction : Variables used and created to end up with the Time Series are:

- IsCanceled: target variable which helps to differentiate between the number of reservations that were canceled and not canceled.
- Dates: these are created from the different data variables existing in the encoded data set.
  - **Reservations made for an exact Arrival Date,** the steps followed to create this time series are:
    - Year, month and day are extracted and converted to an Arrival Date at the hotel.
    - Grouping and counting the reservations made by Arrival Date a data set is obtained.
    - Previous data set is transformed to Weekly data using 'xts package'.
  - **Reservations made an exact Reservation Date,** the steps followed to create this time series are:

  - ∗ Taking the previous data set before it was converted to 'xts' and subtracting LeadTime (number of days between Reservation date and Arrival date).
  - ∗ Previous data set is transformed to Weekly data using 'xts package'.
- **Cancel and Non Canceled Reservations made for Arrival Date and made an exact Reservation Date :** the goal of this time series is to test prediction algorithms splitting the time series information. These 4 time series has been obtained by grouping and canceling them based on canceled information.

III. Exploratory Data Analysis

A. Target Variable

A little study of the proportions of 'canceled bookings' and 'not canceled bookings' must be done to ensure that in the following steps the study keeps having a real representation of the original data set.

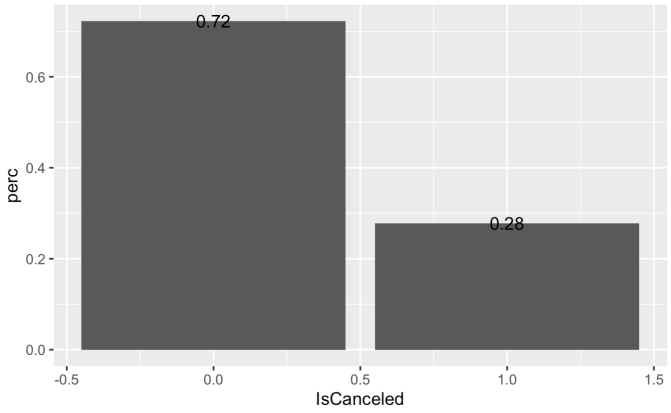

Fig. 1. IsCanceled original data set percentage

B. Relationships and Distributions

Numeric variables are studied using:

- A frequency table to see the distributions of its values.
- A frequency table distinguishing the bookings that has been canceled from the ones that hasn't been canceled, to see in the distribution if some values are more related to booking cancellations or not.

All plots can be seen here following this format:
Figure 2
Categorical variables are studied using:

- A frequency table distinguishing the bookings that have been canceled from the ones that haven't been canceled, to see in the distribution if some values are more related to booking cancellations or not.

All plots can be seen here following this format:
Figure 22

C. Specific to Prediction

1) Matrix Correlation: A matrix correlation has been done to see possible linear correlations between the variables studied (numeric and ordinal encoded).
Figure 20

2) Linear Relationship of Numeric Predictive Variables with IsCanceled: A list of the most correlated variables with IsCanceled is done to start seeing interesting information about the relationship with booking cancellations.
Figure 21

3) Contingency table of categorical variables: A visual representation based on the percentage that results from the contingency table between categorical variables and IsCanceled target variable has been done to see their relationship.
All plots can be seen here following this format:
Figure 22

4) Time Series EDA:

- Time Series: number of reservations in a specific date over time.

All plots are here:
Figure 33

- Seasonality: time series over 4 periods to observe any seasonal component.

Figure 36

- Pre-ARIMA, EDA before performing ARIMA models. It includes:
  - Time Series

Figure 38

-
  - ACF (Auto-Correlation Function): it gives the autocorrelation of any series with its lagged values. Non-Stationarity is found when the ACF are slowly decreasing.
  - PACF (Partial Auto-Correlation Function): it gives the correlation of the residuals (remaining after removing the effects already explained by the earlier lag/s).

Figure 39

IV. Conclusion

Extraction and Transformation of the data is a key process to later run models smoothly and give value to the work done. Depending on the clustering, dimension reduction or prediction algorithm used data must be transformed and extracted properly to reach the desired results.

The Exploratory Data Analysis gives a good understanding about the data that is being studied before any modeling algorithm is done and starts giving some clues about what role each variable plays.

V. References

- https://www.r-bloggers.com/clustering-mixed-data-types-in-r/
- https://datascience.stackexchange.com/questions/22/k-means-clustering-for-mixed-numeric-and-categorical-data
- https://chrisalbon.com/machine_learning/preprocessing_structured_data/encoding_ordinal_categorical_features/

VI. Annexes

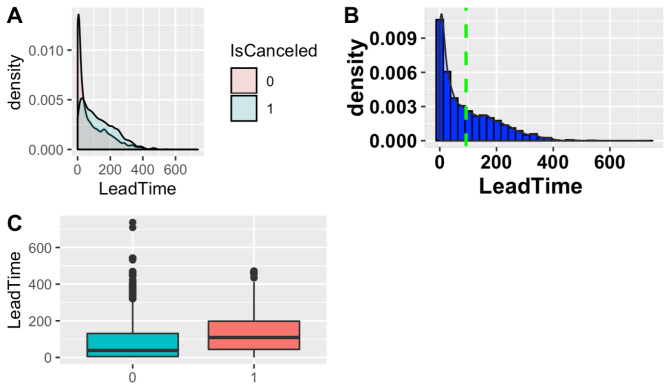

Fig. 2. EDA LeadTime

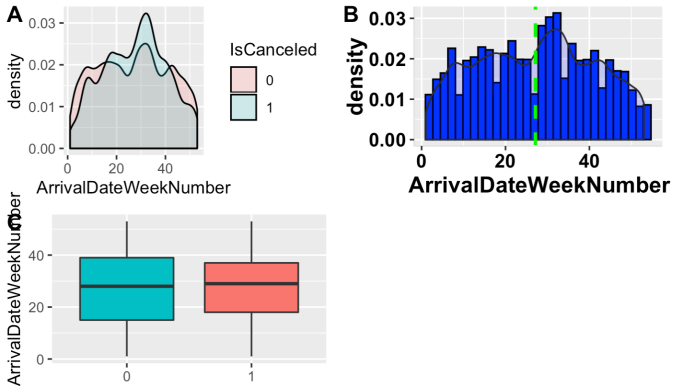

Fig. 3. EDA ArrivalDateWeekNumber

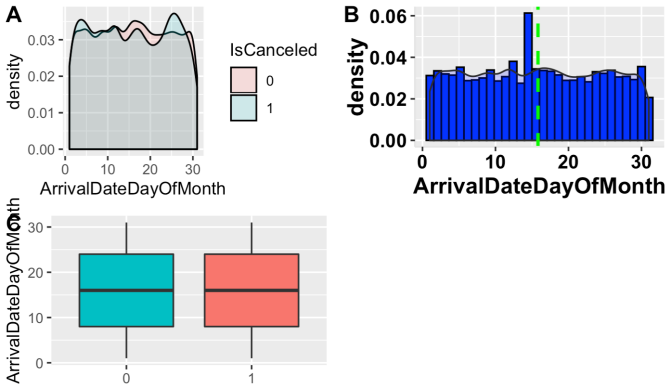

Fig. 4. EDA ArrivalDateDayOfMonth

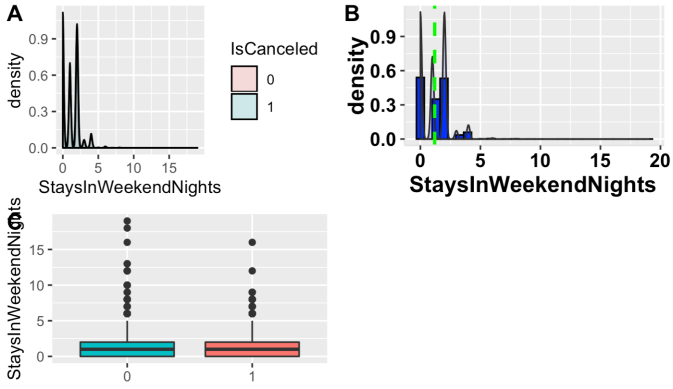

Fig. 5. EDA StaysInWeekendNights

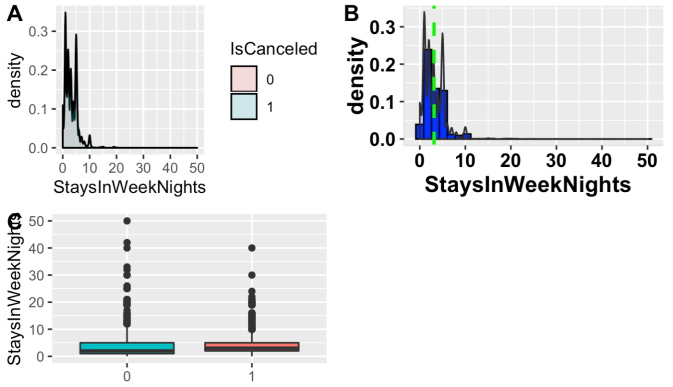

Fig. 6. EDA StaysInweekNights

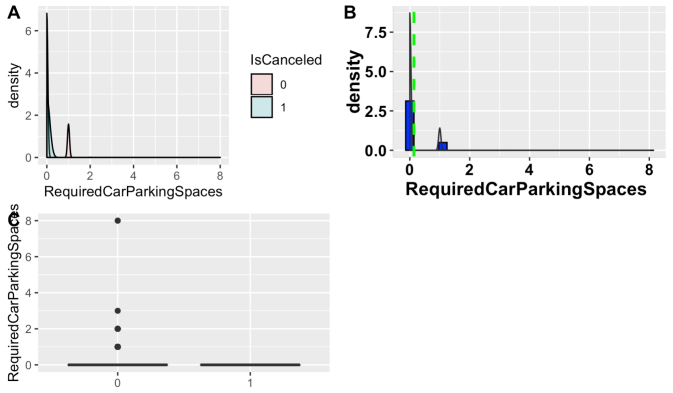

Fig. 7. EDA RequiredCarParkingSpaces

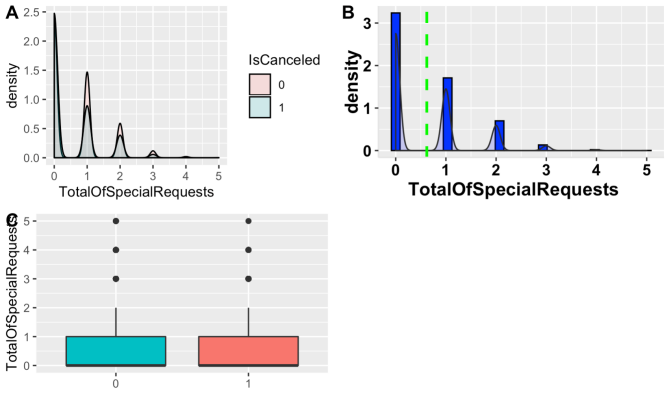

Fig. 8. EDA TotalOfSpecialRequests

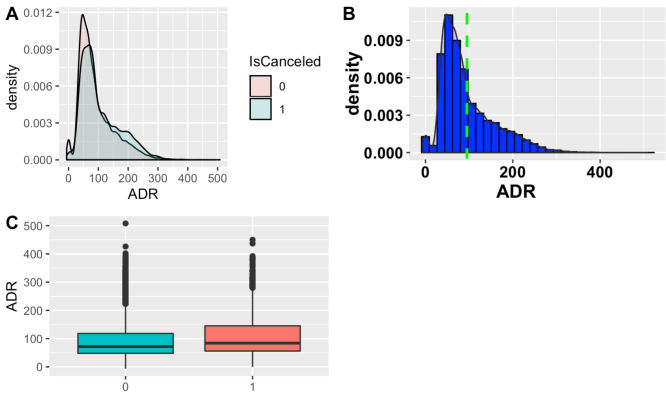

Fig. 9. EDA ADR

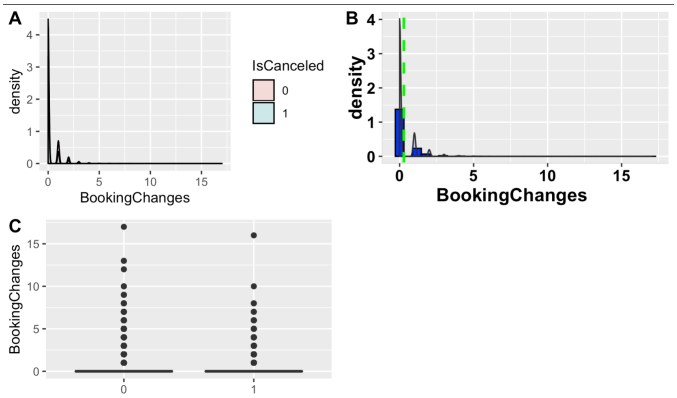

Fig. 10. EDA BookingChanges

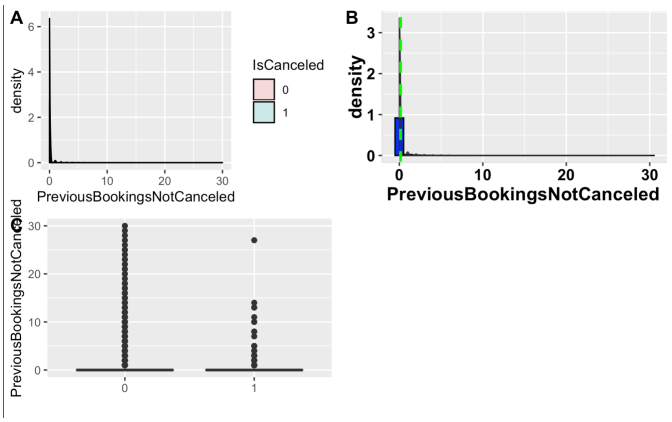

Fig. 11. EDA PreviousBookingsNotCanceled

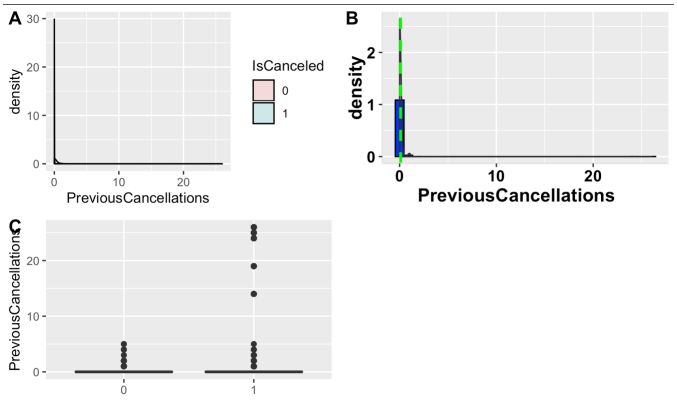

Fig. 12. EDA PreviousCancellations

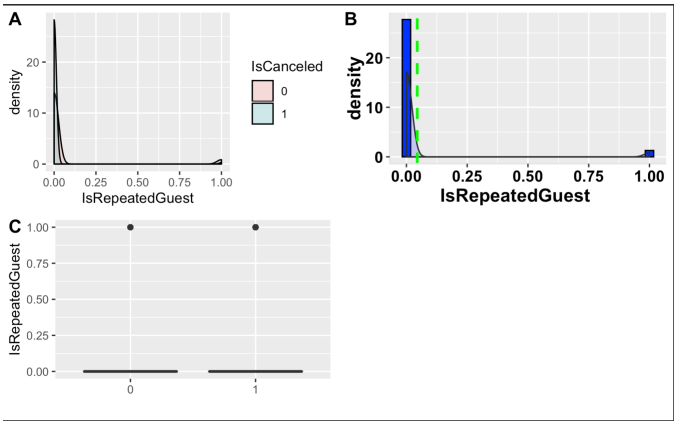

Fig. 13. EDA IsRepeatedGuest

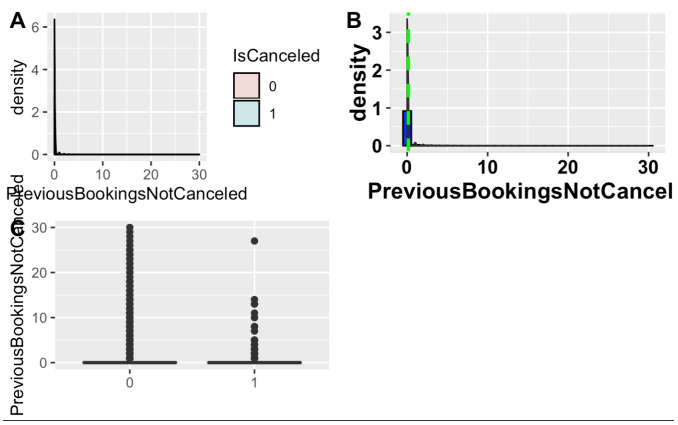

Fig. 14. EDA PreviousBookingsNotCanceled

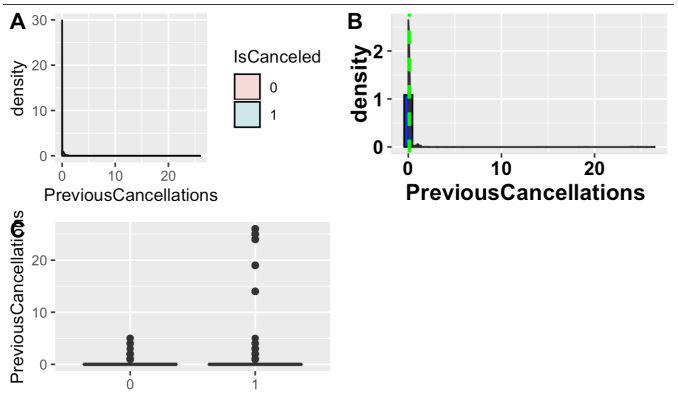

Fig. 15. EDA PreviousCancellations

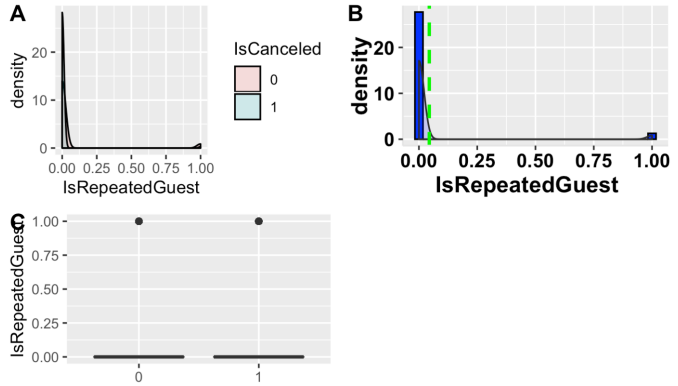

Fig. 16. EDA IsRepeatedGuest

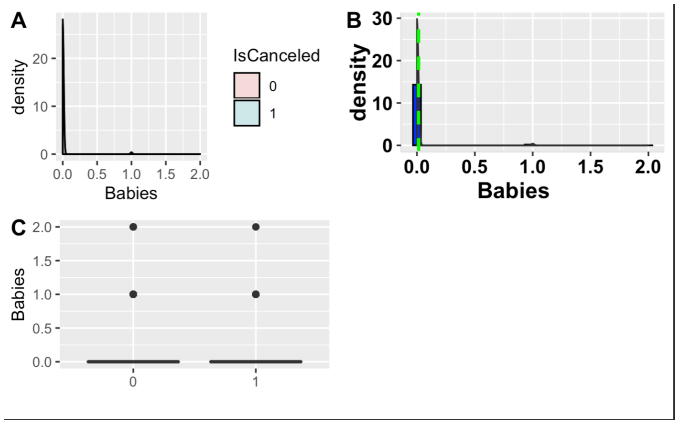

Fig. 17. EDA Babies

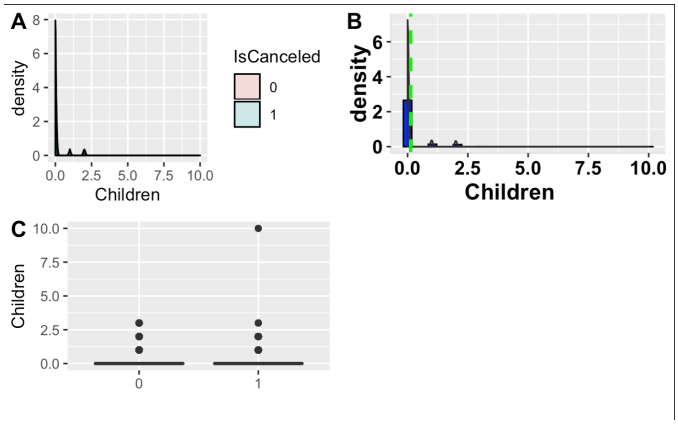

Fig. 18. EDA Children

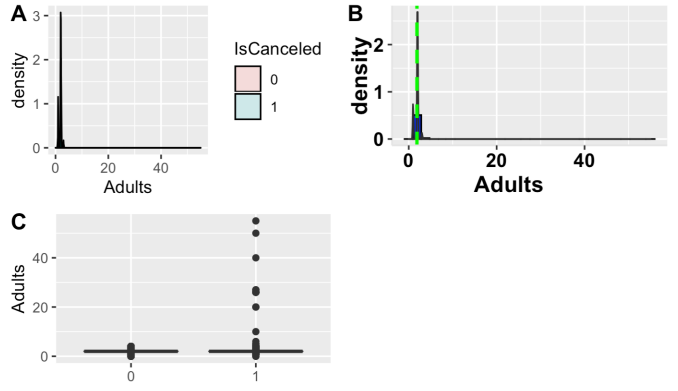

Fig. 19. EDA Adults

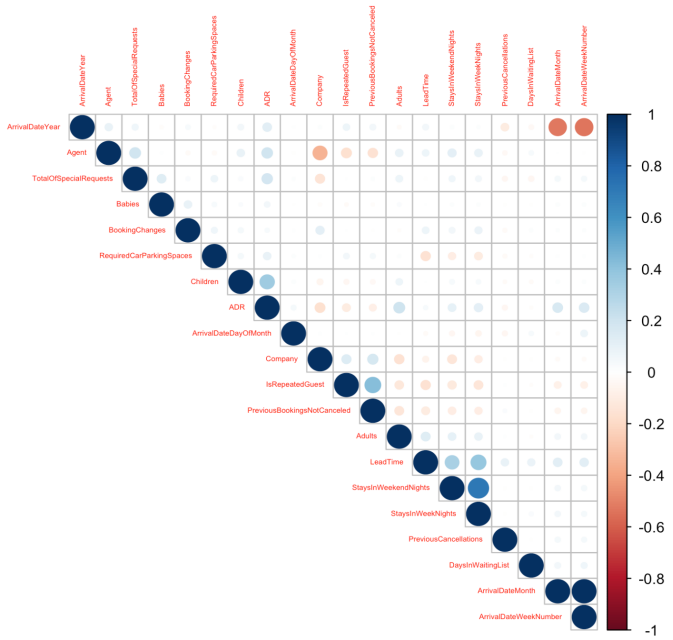

Fig. 20. Matrix Correlation

variable <fctr>	correlated <dbl>	coefficient_limit <dbl>
IsCanceled	1.0000000	0.1
LeadTime	0.2294438	0.1
IsRepeatedGuest	-0.1035628	0.1
PreviousCancellations	0.1141725	0.1
BookingChanges	-0.1148349	0.1
Agent	0.1120368	0.1
ADR	0.1093167	0.1
RequiredCarParkingSpaces	-0.2438634	0.1
TotalOfSpecialRequests	-0.1012946	0.1

Fig. 21. Predictive variables with High Correlation with IsCanceled

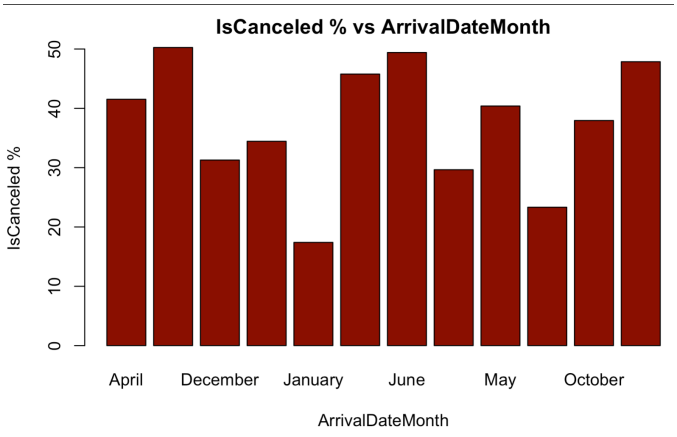

Fig. 22. EDA ArrivalDateMonth

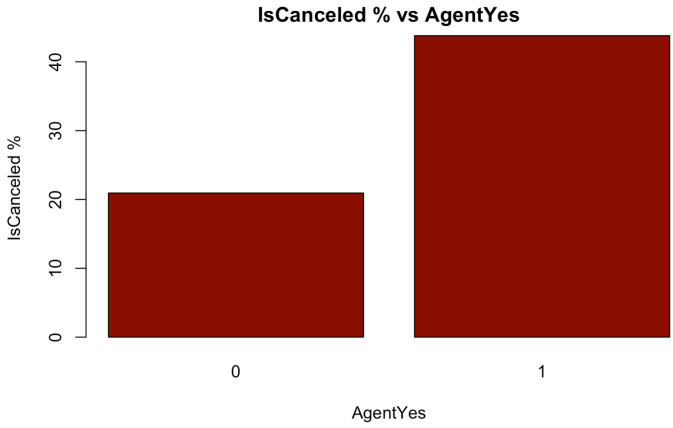

Fig. 23. IsCanceled% vs AgentYes

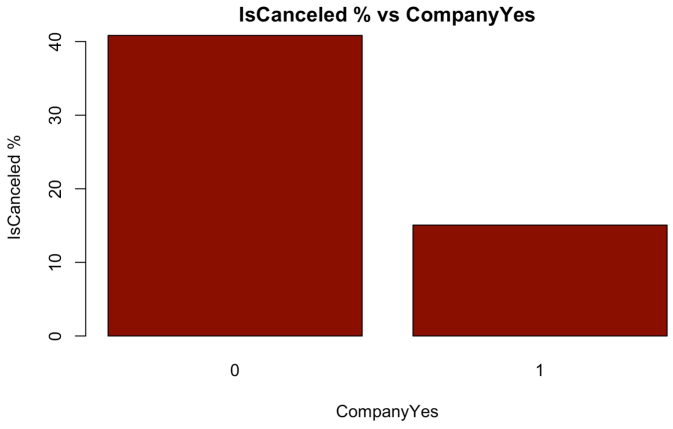

Fig. 24. IsCanceled% vs CompanyYes

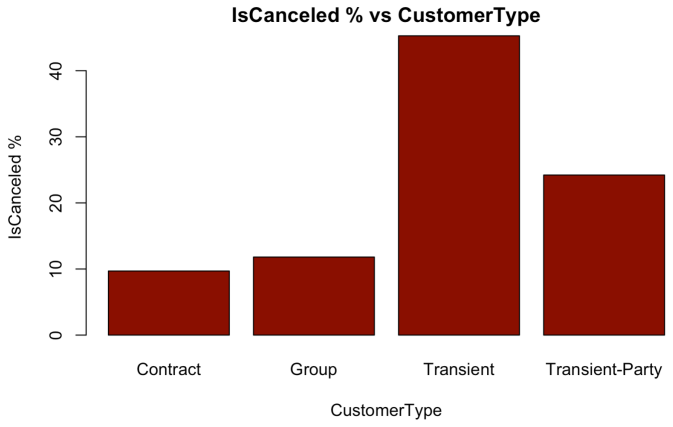

Fig. 25. IsCanceled% vs CustomerType

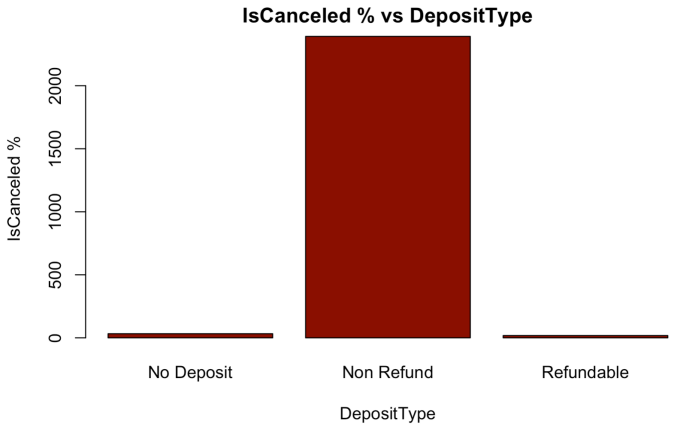

Fig. 26. IsCanceled% vs DepositType

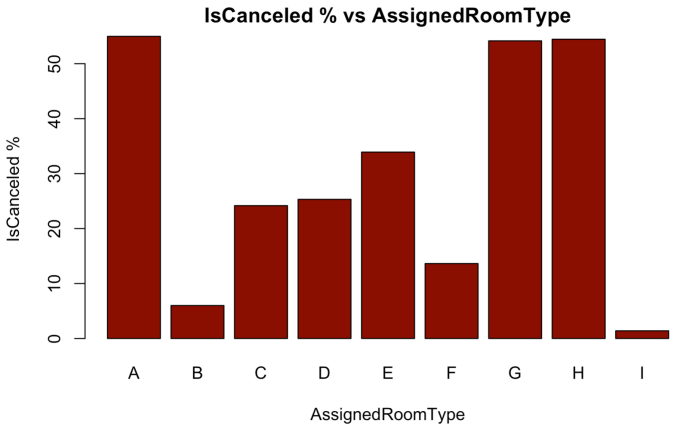

Fig. 27. IsCanceled% vs AssignedRoomType

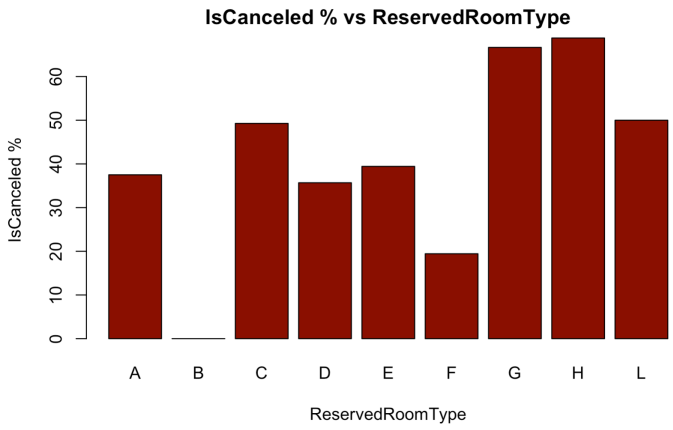

Fig. 28. IsCanceled% vs ReservedRoomType

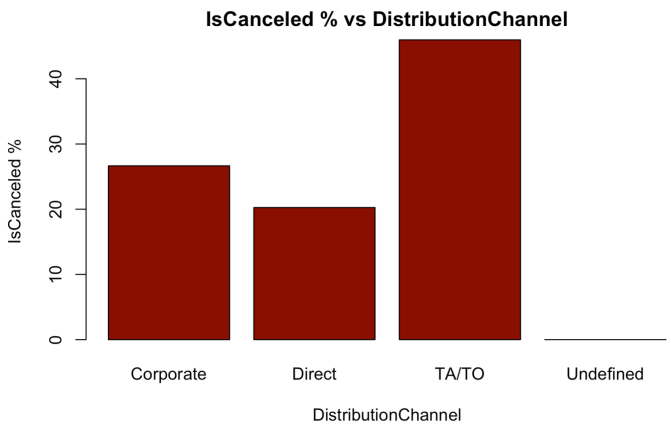

Fig. 29. IsCanceled% vs DistributionChannel

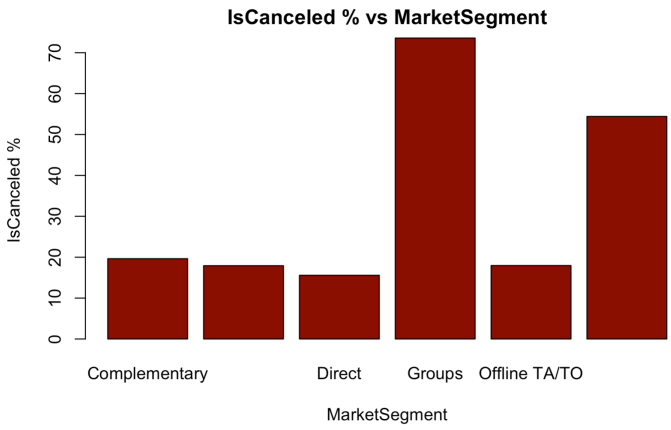

Fig. 30. IsCanceled% vs MarketSegment

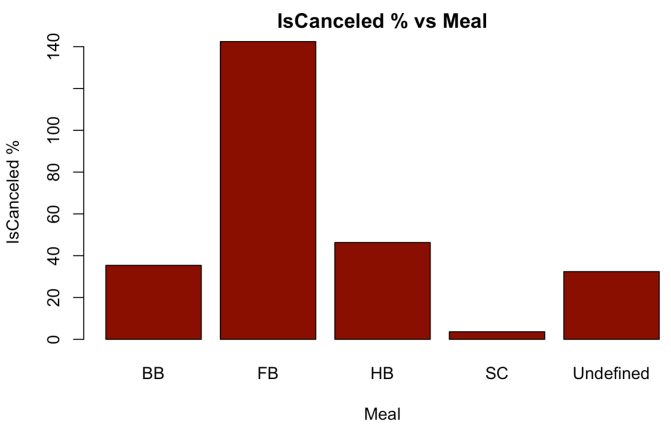

Fig. 31. IsCanceled% vs Meal

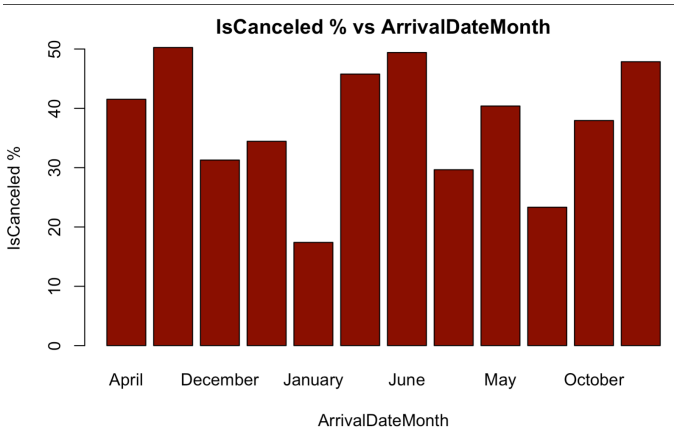

Fig. 32. IsCanceled% vs ArrivalDateMonth

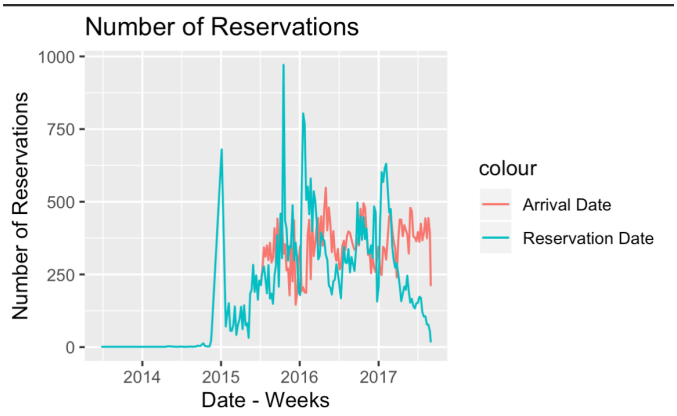

Fig. 33. Arrival Date and Reservation Date

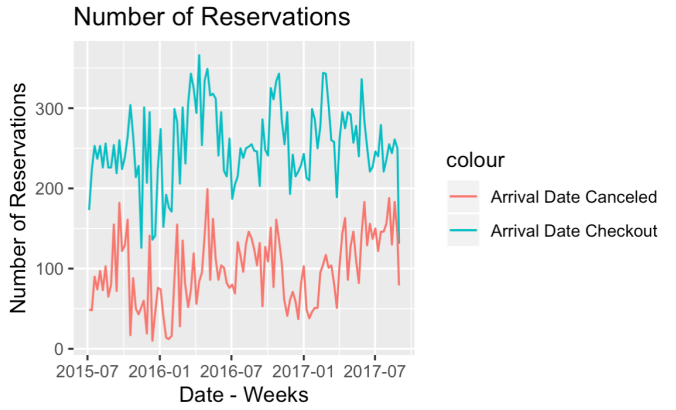

Fig. 34. Reservation Date Check

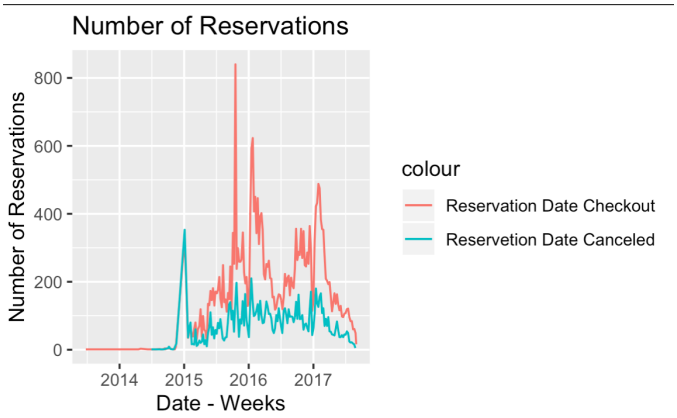

Fig. 35. Reservation Date Checkout and Canceled

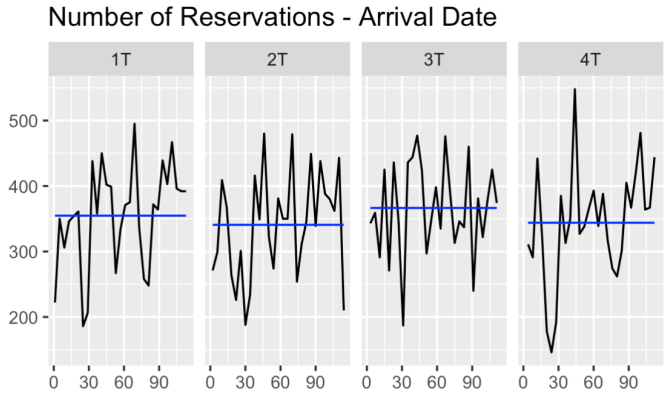

Fig. 36. Arrival Date sesasonality

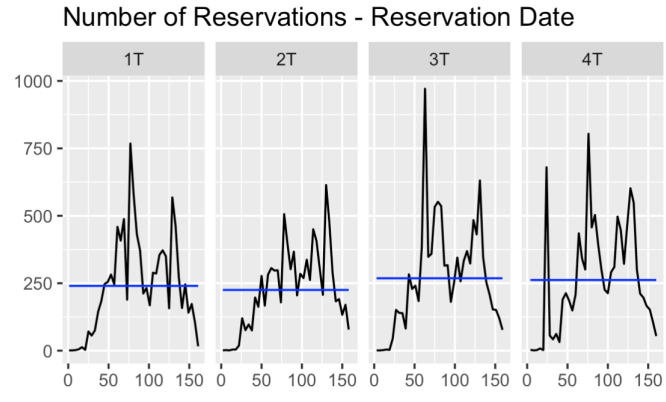

Fig. 37. Reservation Date seasonality

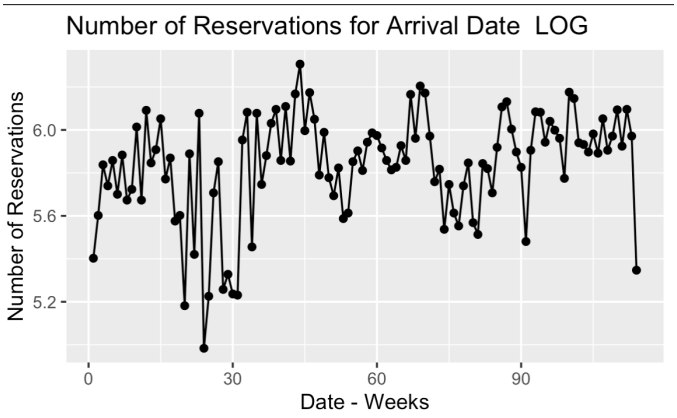

Fig. 38. Arrival Date Log Scale

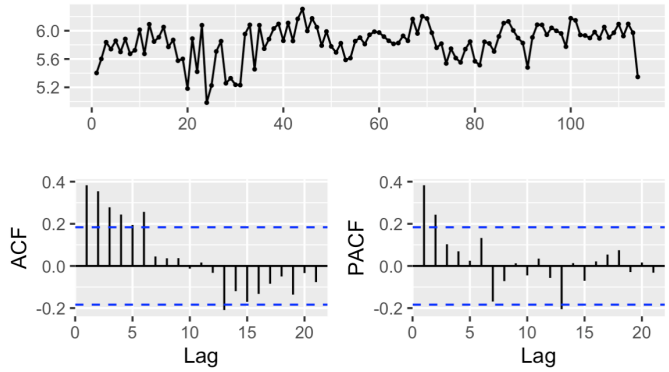

Fig. 39. Arrival Date Log Scale 1 lag

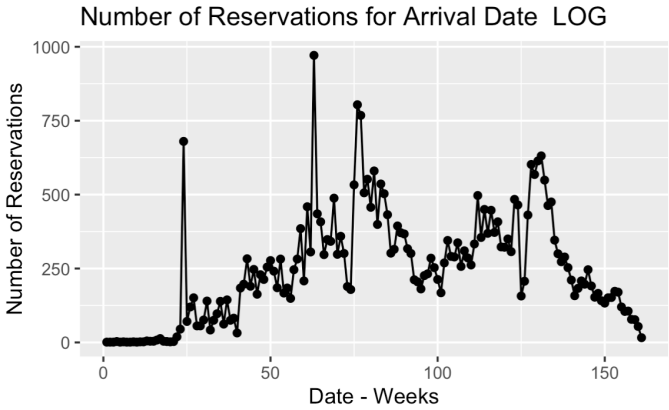

Fig. 40. Reservation Date Log Scale

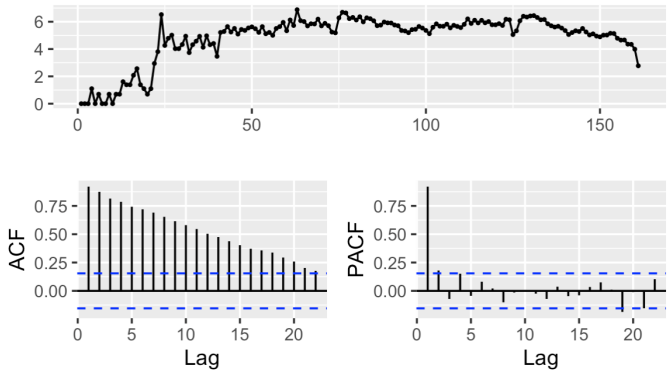

Fig. 41. Reservation Date Log Scale 1 lag