ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

BUSINESS
ANALYTICS
Master of Science

**Department:** Department of Management Science and Technology

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

**Course:** Statistics for Business Analytics II

**Professor:** Dimitris Karlis

**Deliverable:** First Project (Logistic Regression)

**Student Name:** Evangelos Lakkas-Pyknis

**Registration Number**: f2822306

**E-mail:** eva.lakkaspyknis@aueb.gr

# Περιεχόμενα

# Introduction

This assignment aims to study the booking cancelations based on a given dataset of 2000 observations and 17 variables. Given that the response variable (Booking Status) is binary that takes values that belong in {“Canceled”, “Not-Canceled”} I used a logistic regression model to infer the behavior of the cancelations and model the probability $p_i$ of the booking getting canceled. This report consists mainly of three parts: Univariate and Bivariate Analysis of the data, Implementation of the model, Goodness of fit and Interpretation of the constructed model.

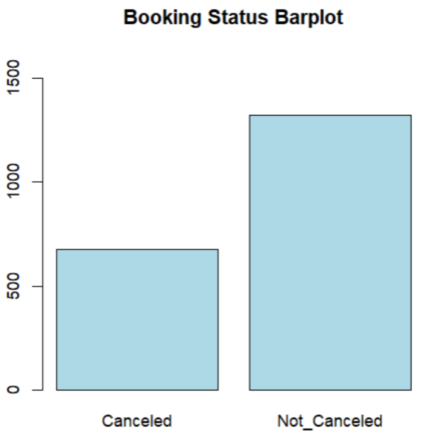
## Remarks about the data

Some brief comments on the processing of the dataset, there was an issue with the feature “Date of reservation” because many available values were not formatted as “year-month-day” but as numbers, I fixed this problem in R by setting all the dates in the same appropriate format. On top of that, I created a new column in the dataset that corresponds to the month of each date. I did this transformation because it is hard to infer the effect that each date has individually on a booking on the other hand it is reasonable to make a hypothesis that the months may play a role in the cancelations of bookings because generally hotel activities present seasonality (it is harder to find a booking in summer or December and easier on months like November or March), so this feature has a point to be examined in the context of my analysis. After the conversion of the dates two missing values were introduced in the dataset but given that the observations with NA values were only two out of 2000 (a very small part of the sample) I dropped the observations with these two values. Finally, I subtracted the feature “Booking ID” from the dataset because it is a unique identifier for each booking and by definition can’t help me to create a model for the inference of the cancelation’s behavior. So, the final dataset that I had available remained with 1998 observations and 17 features.

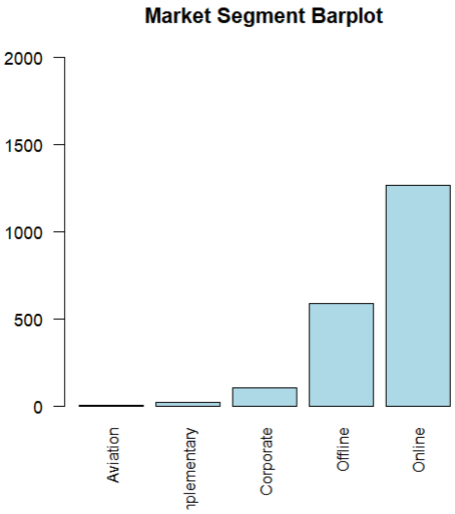
# 1) Univariate and Bivariate Analysis

## Univariate analysis

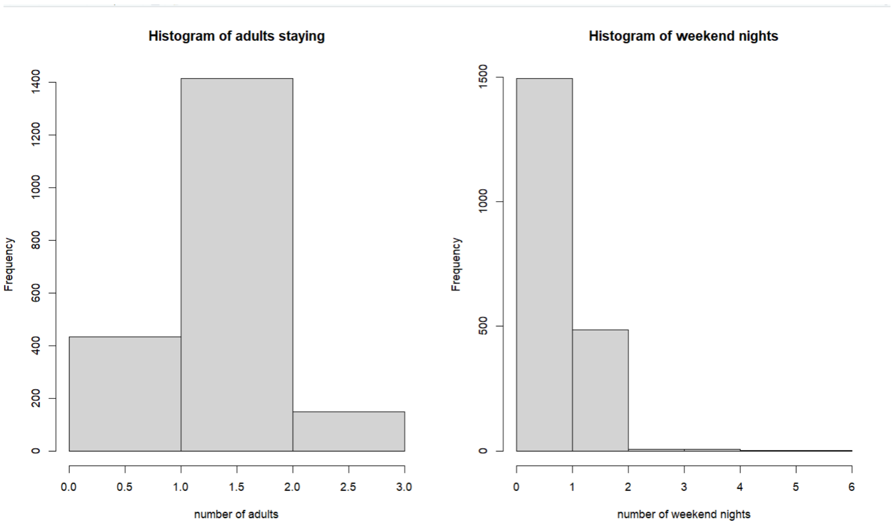
To gain an initial insight into the data I checked for information about my response variable and some other categorical variables that could play a role in the model. We notice that the bookings that were canceled were about half of the bookings that weren't canceled (678 bookings were canceled in total, corresponding approximately to 33% of the sample observations).



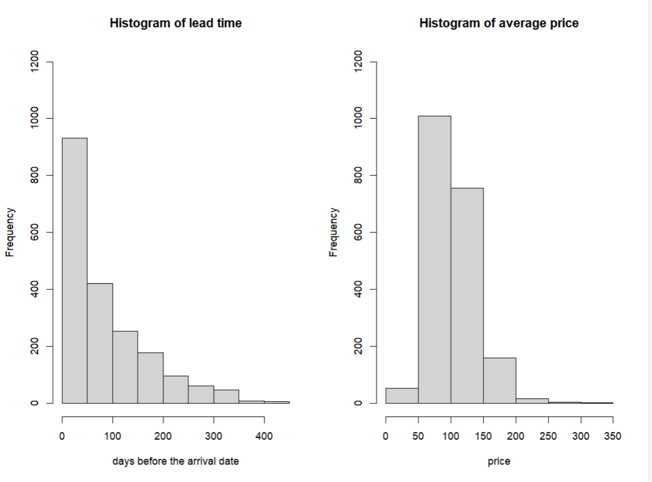
Market segment's values are concentrated on the online and offline levels and has some bookings that correspond to the corporate level, the other levels have very few observations. As for the parking space 1940 bookings didn't offer any parking and only 58 had this feature.



After that I got a description for the numeric variables of the dataset. As we can see from the histograms below most bookings host 2 adults with very few taking 3 adults, at the same time the visitors usually stay up to two weekend nights with the most common behavior being the stay for one weekend night.

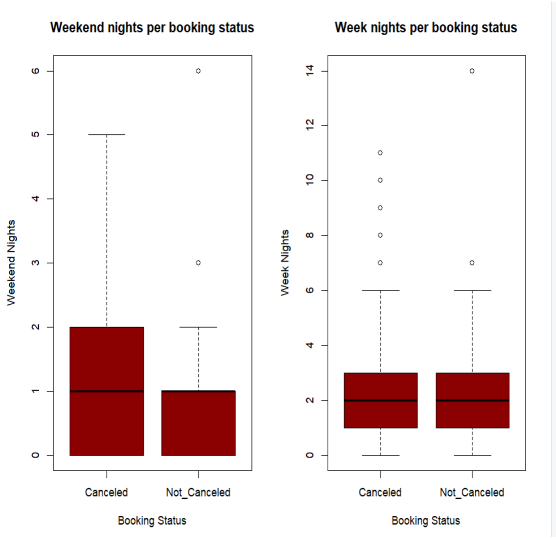


The dataset had also two continuous variables the average price of the booking and the lead time (how many days before the arrival day has the room been booked). We notice that neither of the two variables follow the normal distribution and have a right-skew. Most bookings have a price between 50 and 150 dollars and very few have a price higher than 200 dollars. In the same way most bookings have a lead time smaller than 100 days and the group with the lead time with the most bookings includes bookings from 0 to 50 days, after that while the lead time increases the bookings gradually decrease.

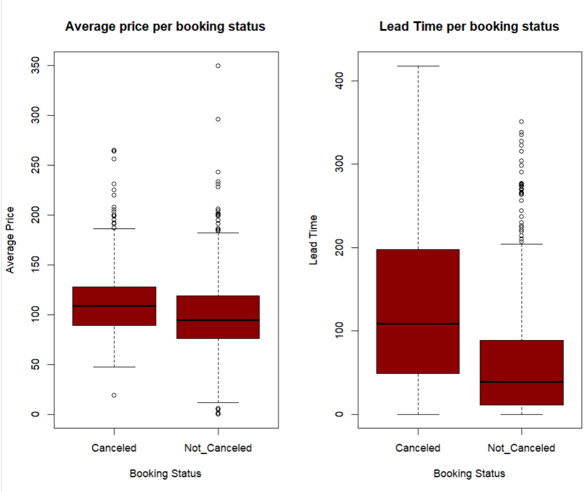


## Bivariate analysis

Now let’s see a comparison of different numeric variables for the canceled and not canceled bookings. Starting from two discrete variables of the dataset that reflect the duration of the staying of the visitors we notice that the number of the weekend nights are different between the canceled and not canceled bookings the median of the weekend may be approximately the same, but the canceled bookings generally include more weekend nights and have a greater dispersion than the not canceled ones. On the other hand, the weeknights don’t seem to have a significant difference between the two levels of the booking status the median is the same and so is the interquartile range, the canceled bookings may have some more outliers for the weeknights, but the general image remains the same. So these facts indicate that weekend nights could be an important variable for my model while weeknights are not so much.



As for the continuous variables, we notice that there is a difference for both average price and lead time for canceled and not canceled bookings. The median price is higher for the canceled bookings than the not canceled, also the cancelled bookings have a slightly less dispersed average price and less extreme values. The difference of lead time between the two levels of the booking status factor is obvious, the median of lead time is significantly higher for canceled bookings and in general these bookings have a lot dispersion for lead time this is showed by the size of the box that corresponds to the canceled bookings compared to this of the not canceled. These characteristics indicate that these variables may be useful for the creation of my model.



# 2) Model Implementation

## Fitting the model

First, I fitted the null model to see its ability for inference and goodness of fit. The null deviance was equal to 2559.8 and the ratio of the deviance to the degrees of freedom was equal to 1,28 (I want this ratio to be as close as possible to 1) so I started to see what variables I should select to build a better model.

In order to make that selection, I applied a LASSO method with a lambda value corresponding to the most regularized model within one standard error of the minimum,[1] and the coefficients that I got back are shown in the table below. The remaining variables are five (weekend nights, parking, lead time, average price, and special requests) which are statistically significant plus the intercept so my model will consist of these terms.

| Variable | Coefficient | Variable | Coefficient | Variable | Coefficient |
|---|---|---|---|---|---|
| **Intercept** | **2.57** | **Car Parking** | **0.39** | Date of reservation | . |
| Num. of adults | . | **Lead Time** | **-0.01** | Month of reservation | . |
| Num. of children | . | **Avg. Price** | **-0.13** | **Special Requests** | **0.71** |
| Weeknights | . | Repeated | . | Not P.C. | . |
| **Weekend nights** | **-0.05** | Market Segment | . | P.C. | . |
| Meal Type | . | Room Type | . | | |

So, the mathematical formulation of this model looks like this:

$$\log\left(\frac{p}{1-p}\right) = 3.58 - 0.22 \times weekend.nights - 0.01 \times lead.time + 1.82 \times parking - 0.02 \times avg.price + 1.05 \times sp.requests$$

Where:

p: the probability of the booking to get canceled.

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

parking: a dummy variable taking the value 0 if the booking doesn't include a parking space area and 1 if it does.

The estimates of the coefficients are also shown in the table below along with the indication for their statistical significance (all p-values are less than 0.05 which means that the coefficient cannot be assumed to be zero) and the standard error of the estimations.

| Coefficients | Estimates | Standard error | Pr(>\|z\|) |
|---|---|---|---|
| **Intercept** | 3.583 | 0.2323 | < 2e-16 |
| **Weekend.nights** | -0.224 | 0.0644 | 0.000506 |
| **Lead.time** | -0.012 | 0.0007 | < 2e-16 |
| **Parking** | 1.824 | 0.5505 | 0.000920 |
| **Avg. Price** | -0.020 | 0.0018 | < 2e-16 |
| **Sp. Requests** | 1.055 | 0.0881 | < 2e-16 |

[1] I preferred this lambda value over the minimum classification error because the second one might be sensitive to specific run

I tried to apply a stepwise forward method using the BIC criterion on this model, but the same variables remained.

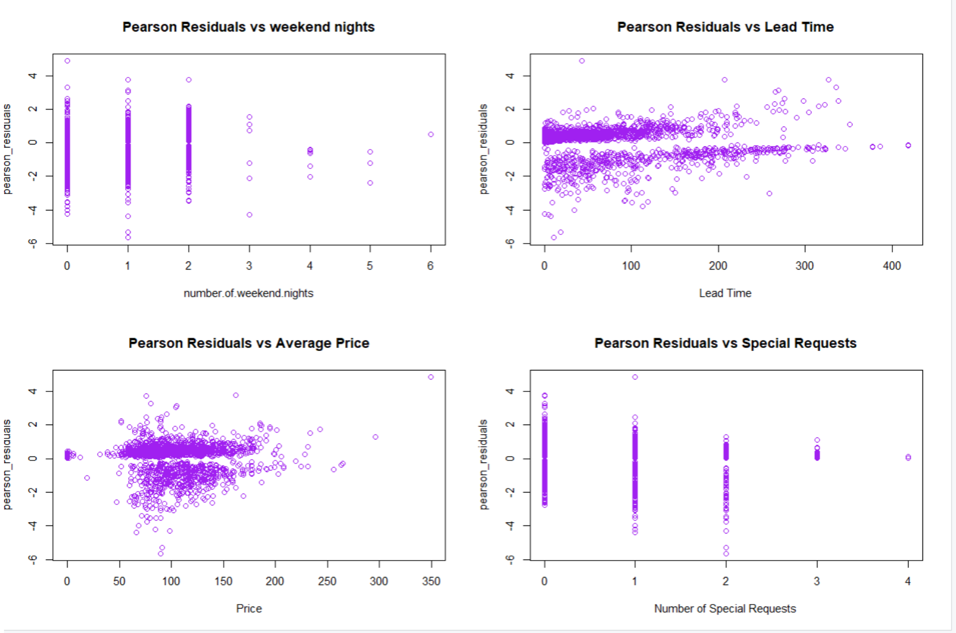
## Model’s goodness of fit and diagnostics

After I fitted the model, I examined if it was a good fit in my dataset. The residual deviance of the model is 1918.4 which is a significant improvement compared to the null model which had 2559.8 deviance. Also, if we divide the residual deviance with the degrees of freedom the result is pretty close to 1 (1918.4/1992≈0.96) which suggests a good fit in the data and an improvement compared to the null model which had null deviance divided by the degrees of freedom ≈ 1.28. The calculated adjusted pseudo-$R^2$ statistics of Cox-Snell and Nagelkerke (which are generally used for the goodness of fit for logistic regression) for the presented model are equal to 0.27 and 0.38 respectively, the values of these two statistics also imply a good fit. Finally, another way to see if the model has a good fit is to test if the deviance of the residuals of my model follows the $X^2$ distribution with 1992 degrees of freedom, I failed to reject the null hypothesis for this test at a=5% (p-value = 0.87), so I assume that the goodness of fit of my model is satisfactory.

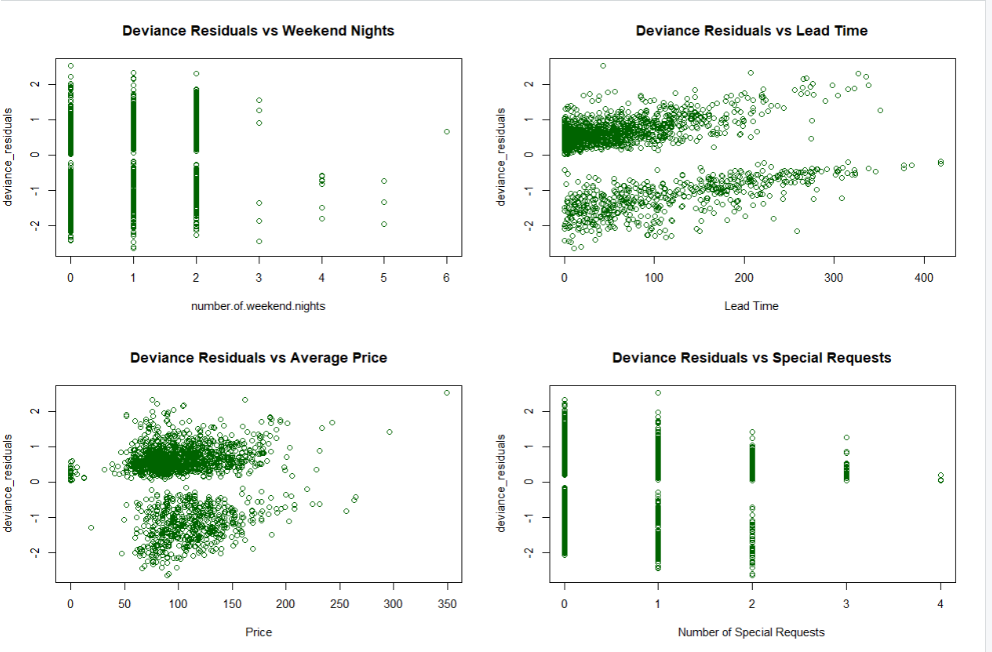
Then, I searched If the model could have a problem due to the possible presence of multicollinearity, however by looking at the results of the variance inflation factors (VIF) I confirmed that my model did not have such a problem because the VIF values for each variable that are shown in the table below are close to 1 (<8). So the predictors are not correlated with each other and the model can be used for inference.

| Variable | VIF |
|---|---|
| **weekend.nights** | 1.01 |
| **lead.time** | 1.07 |
| **car.parking.space** | 1.00 |
| **avg.price** | 1.14 |
| **special.requests** | 1.11 |

Finally, I plotted the Pearson and Deviance residuals of the model against the predictors to see if any unusual patterns would need further examination. The plots are shown below.



Both the Pearson and Deviance residuals don't seem to follow any patterns the residuals of the continuous variables shape 2 (slightly) distinct clouds of points as expected. The residuals of the discrete variables have an uneven distribution for the bigger values of the variables, but this seems reasonable because it is difficult for a booking to last more than 3 weekend nights and many special requests. So, everything seems fine.



## 3) Model Interpretation

As I referred before my model is given from the following formulation: $\log\left(\frac{p}{1-p}\right) = 3.58 - 0.22 \times weekend.nights - 0.01 \times lead.time + 1.82 \times parking - 0.02 \times avg.price + 1.05 \times sp.requests$

Each element of this model can be interpreted as explained below:

Intercept: the probability of the booking getting canceled is equal to $\frac{1}{1+e^{3.58}}$≈0.02 or 2% when all the other covariates are equal to 0 and the booking has no parking space available, this fact may not be very meaningful because the price cannot be assumed to be 0.

Weekend nights: The probability of the booking getting canceled has a negative relationship (this is shown by the negative sign of the coefficient) with the number of weekend nights which means that an increase in this variable will decrease the probability of the cancelation. For the increase by one weekend night, the odds ratio (i.e. the probability of the booking getting canceled divided by the probability of the booking being canceled) decreases by $e^{-0.22}$≈0.80, when all the other variables remain constant.

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

Car parking space: The probability of the booking cancelation has a positive relationship (this is shown from the positive sign of the coefficient) with the parking space which means that the existence of parking increases the probability of the cancelation. When the booking has available parking, and all the other covariates are constant the odds ratio becomes $e^{1.82}$≈6.17.

Average price: The probability of the booking cancelation has a negative relationship (this is shown by the negative sign of the coefficient) with the average price which means that an increase in this variable will decrease the probability of the cancelation. If the average price increases by one dollar the odds ratio (i.e. the probability of the booking getting canceled divided by the probability of the booking not getting canceled) decreases by $e^{-0.02}$≈0.98, when all the other variables remain constant.

Special requests: The probability of the booking cancelation has a positive relationship (this is shown by the positive sign of the coefficient) with the number of special requests which means that an increase in this variable will increase the probability of the cancelation. For every special request that a booking gets the odds ratio increases by $e^{1.05}$≈2.85, when all the other variables remain constant.