Hotel booking demand

Are we able to predict a booking cancellation?

Chiara De Luca Irene Caria Matteo Pernini

- ❑ Hotel booking demand dataset
- ❑ Cleaning data
- ❑ Exploratory data analysis
- ❑ Model data
- ❑ Conclusions

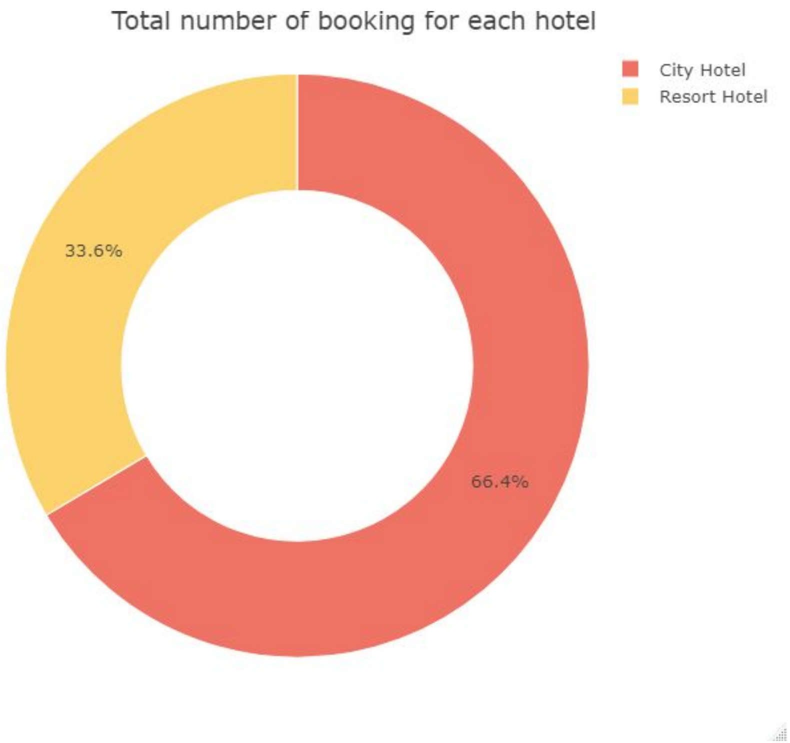
Dataset

- Observation -> hotel booking
- 32 variables describing 119390 observations
- Hotel : City and Resort
- Both located in Portugal
- Focus on variable *is_canceld*

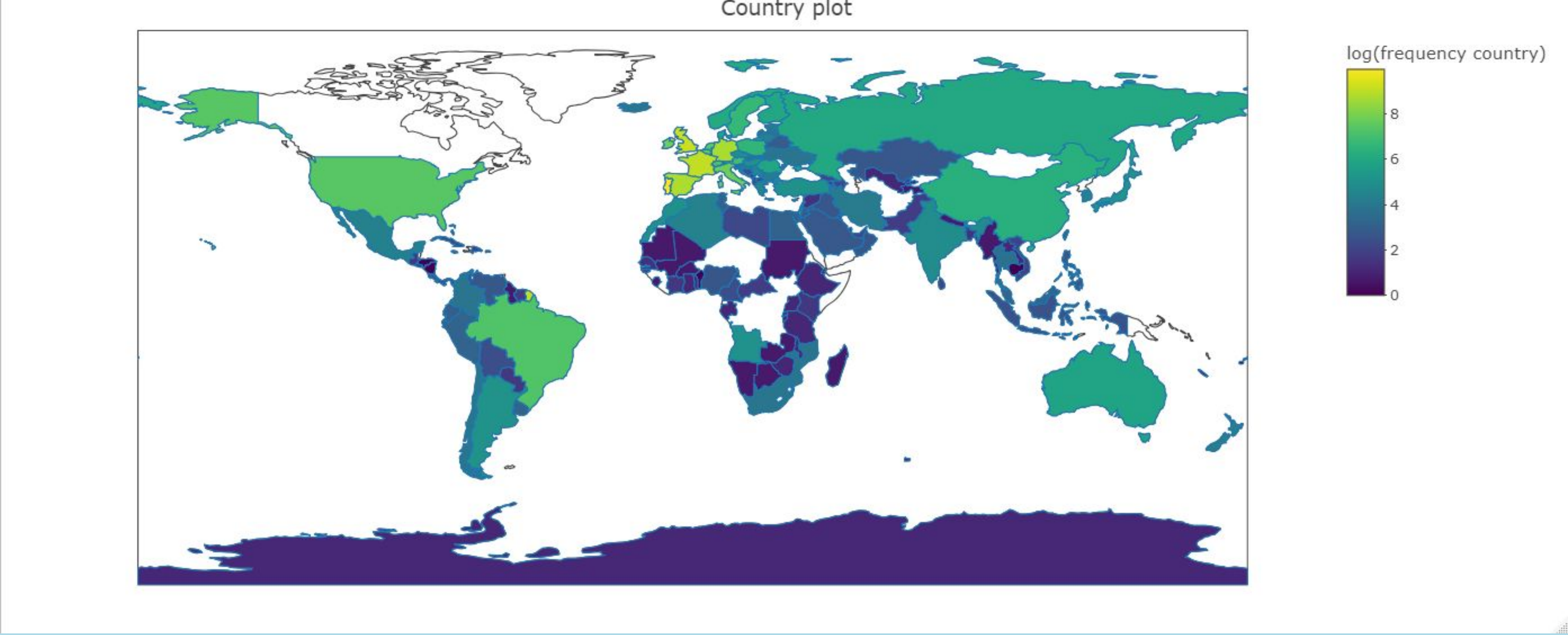
Cleaning data

- Convert characters into factors
- Add column *total_stays*
- Replacing missing values *children*
- Remove *company* and *agent*
- Most variables are categorical

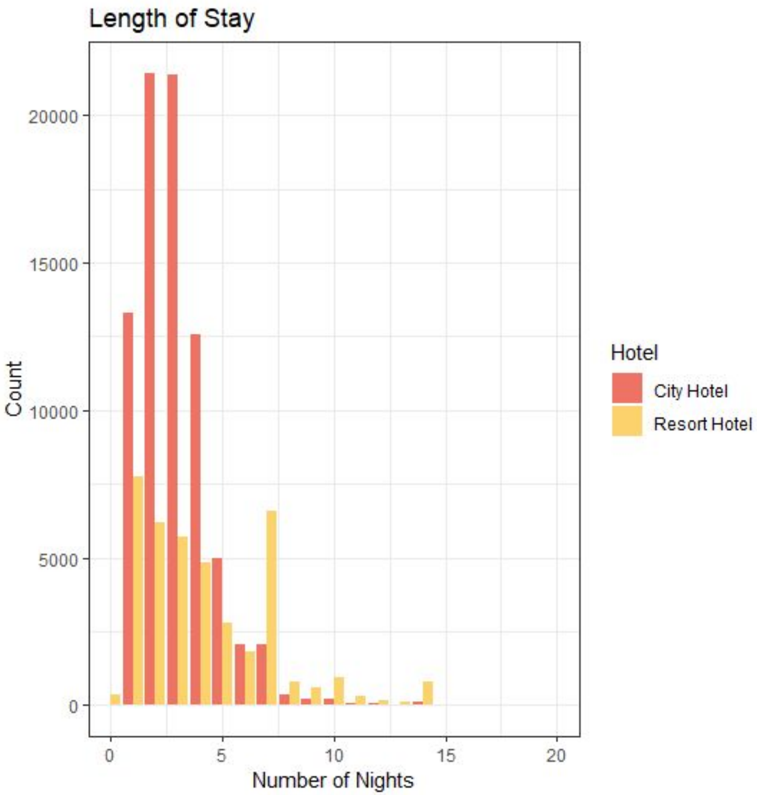
Exploratory Data Analysis



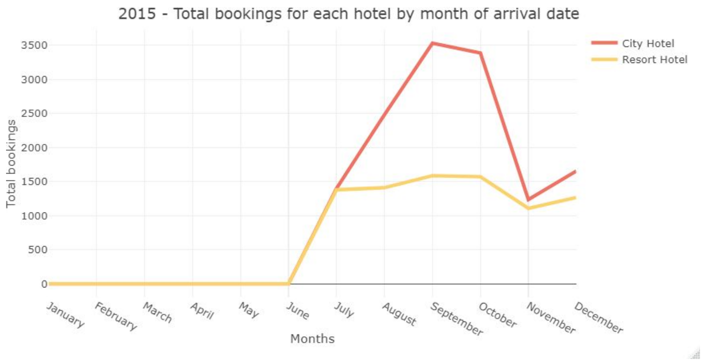
- ❑ City Hotel : 79330
- ❑ Resort hotel : 40060



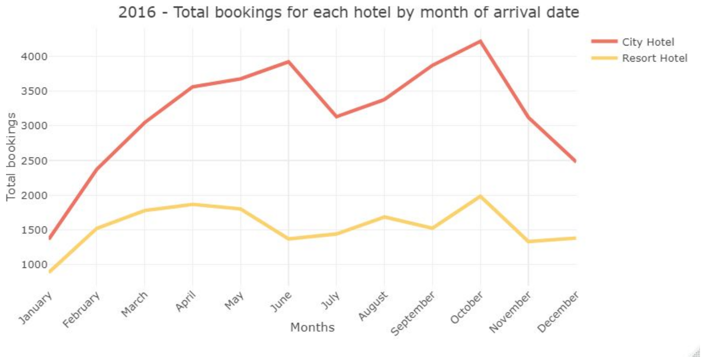
PRT : 48590 - GBR : 12129 - FRA : 10415



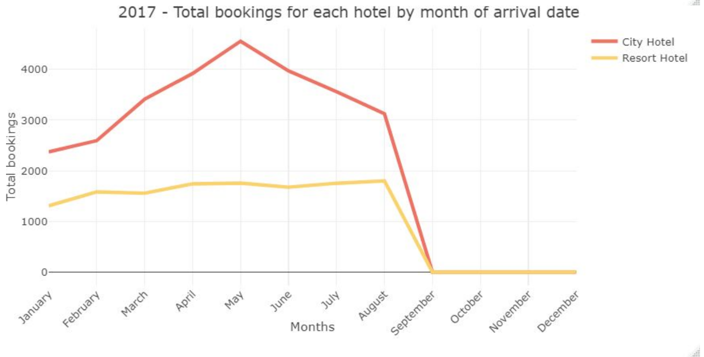
- City Hotel : 2-3 nights
- Resort hotel : 7 nights
- Long stay unusual for City Hotel



2015 - Total bookings for each hotel by month of arrival date



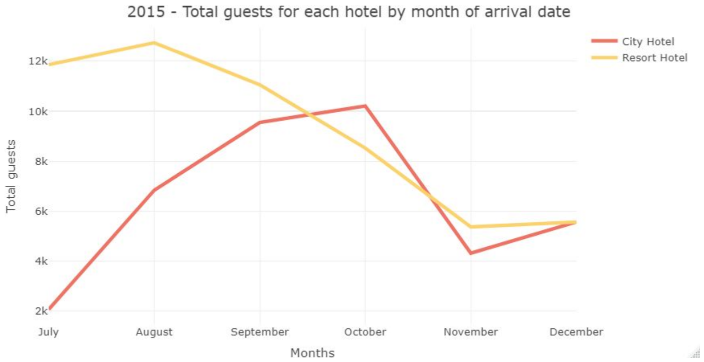
2016 - Total bookings for each hotel by month of arrival date

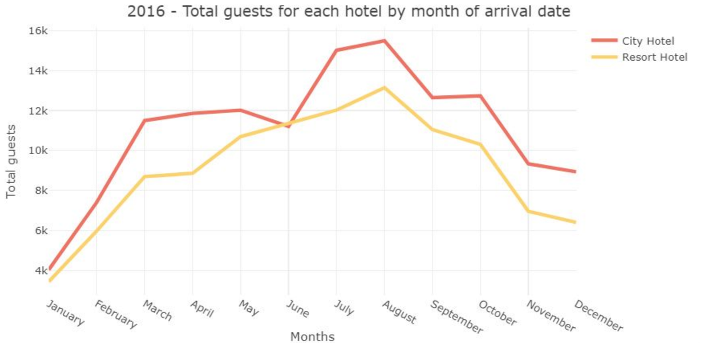


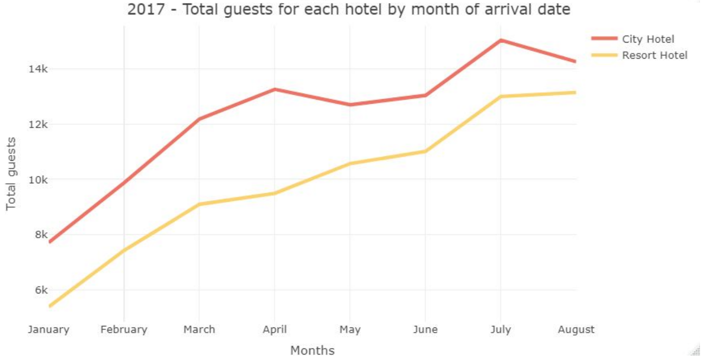
2017 - Total bookings for each hotel by month of arrival date

- Total bookings for each year

- Focus on summer period

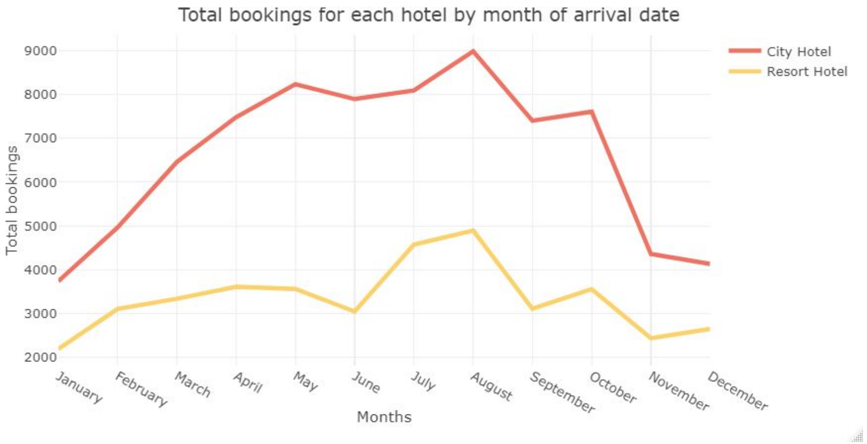


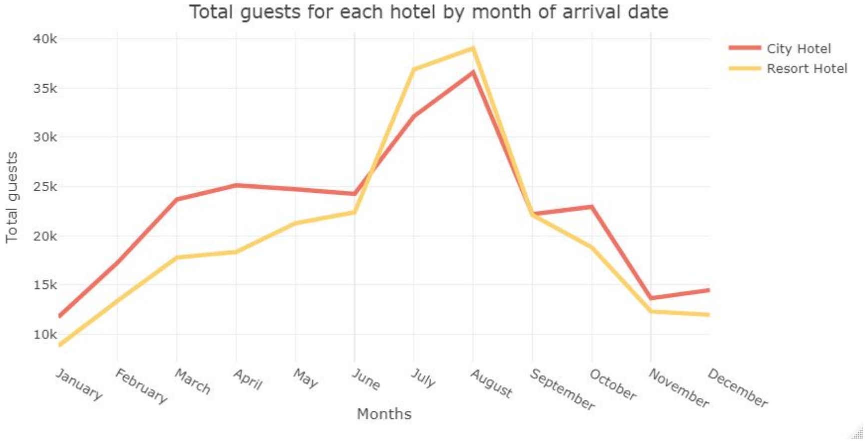




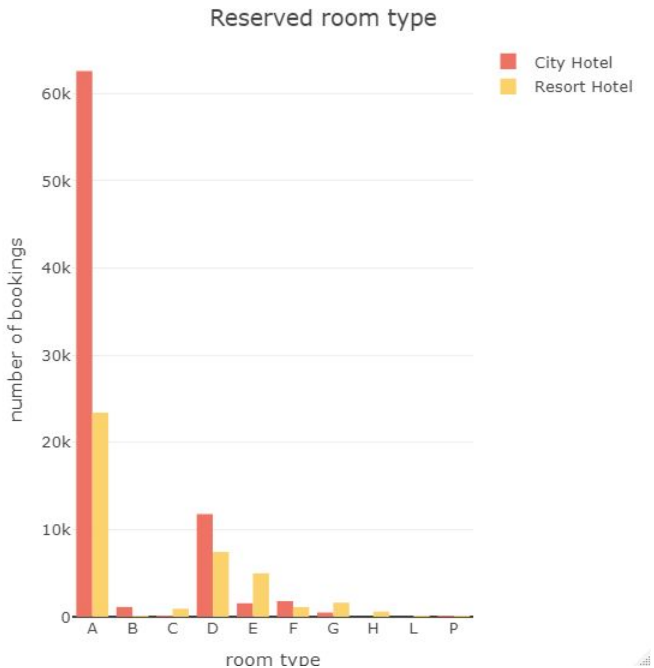
❑ Total guests for each year

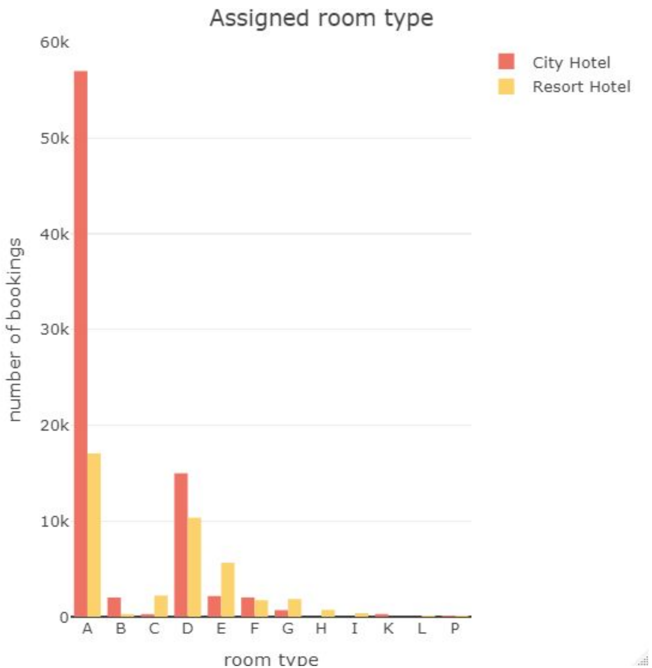
❑ Focus on summer period





- Total bookings vs total guests
- Focus on summer period



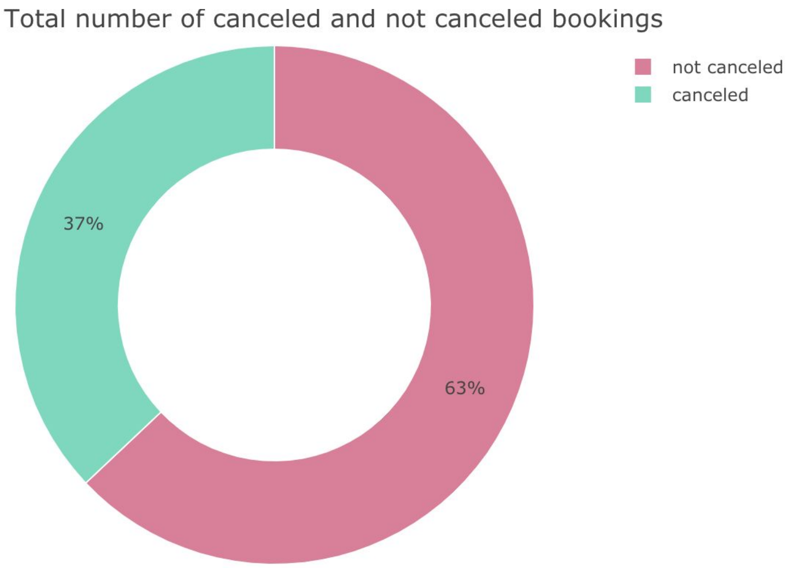


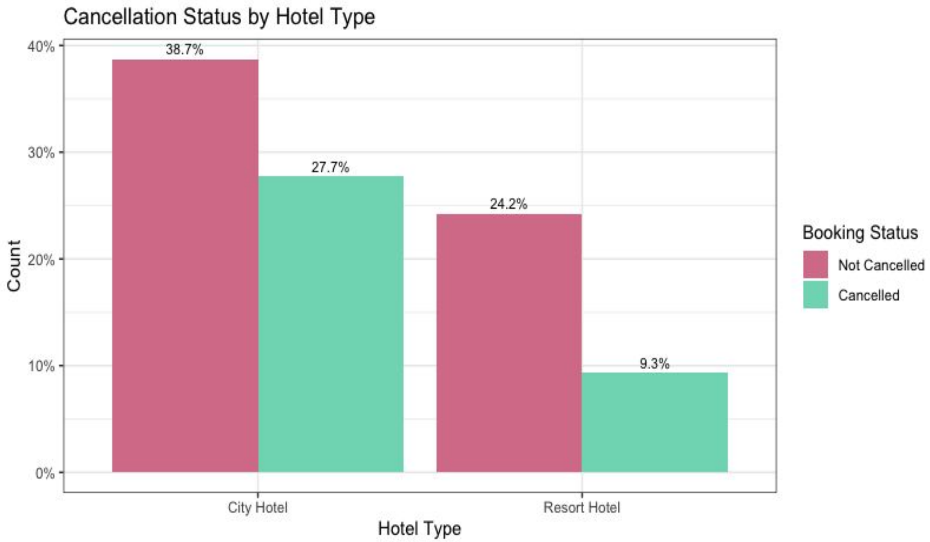
Reserved room type

```
                  A     B     C      D     E     F     G     H                 L     P
City Hotel    62595  1115    14  11768  1553  1791   484     0     -     -     0    10
Resort Hotel  23399     3   918   7433  4982  1106  1610   601     -     -     6     2
```

Assigned room type

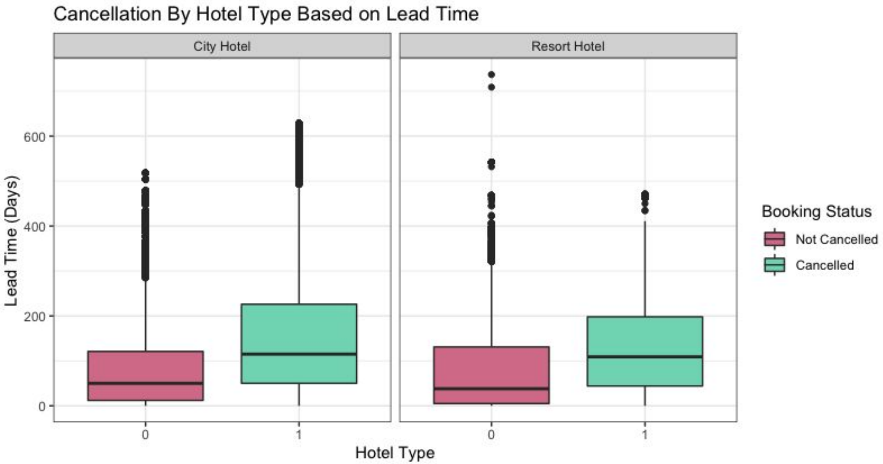
```
                  A     B     C      D     E     F     G     H     I     K     L     P
City Hotel    57007  2004   161  14983  2168  2018   700     0     0   279     0    10
Resort Hotel  17046   159  2214  10339  5638  1733  1853   712   363     0     1     2
```

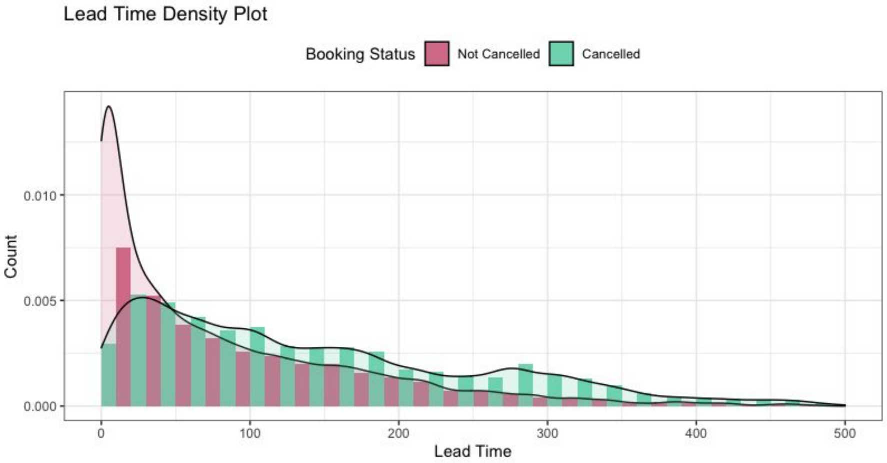




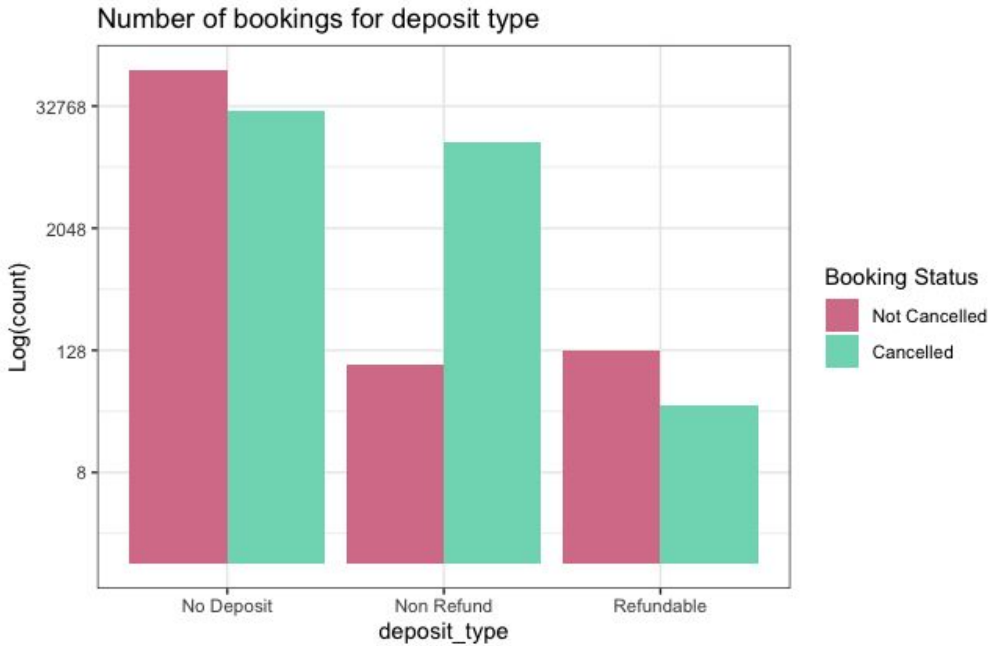
- ❑ is_canceled: binary variable
- ❑ is_canceled: unbalanced

- ❑ higher percentage of not canceled
- ❑ ratio canceled/not canceled lower for Resort





- lead_time: number of days that elapsed between the day of the booking and the arrival date or cancellation
- the median of cancelled is higher than for not cancelled
- the peak of the curves occur for a low value of lead_time
- from lead_time of about 40-50 days the “cancelled” curve is strictly above the other one



On average, for variable “Non Refund”:

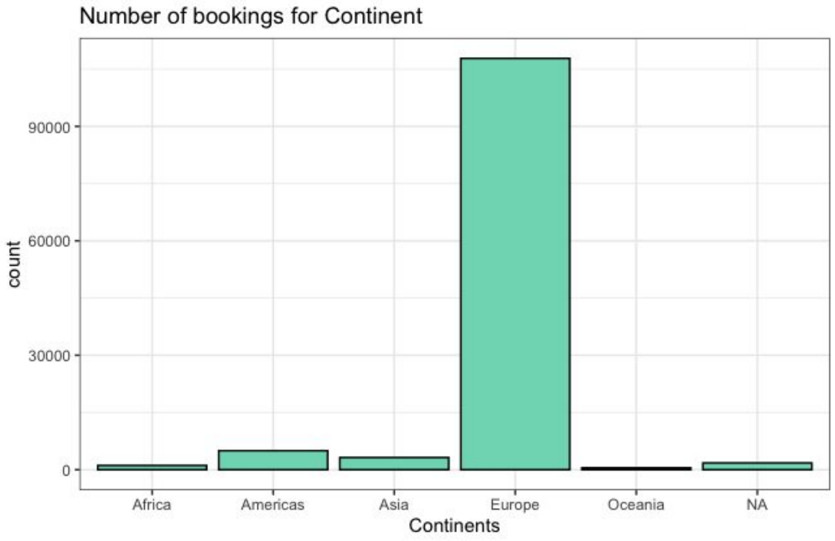
- ❏ lead time twice as long as No Deposit
- ❏ previous cancellation about 10 times higher than No Deposit
- ❏ especially adults
- ❏ very few special requests

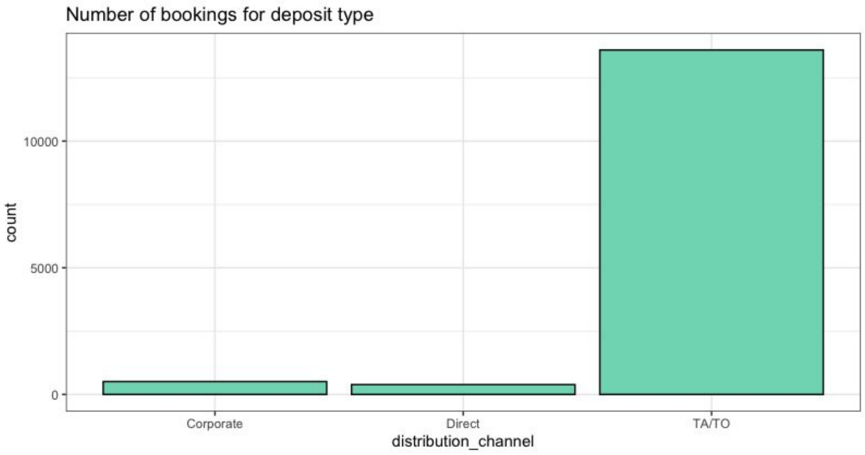
deposit_type	lead_time (avg)	previous_canc (avg)	adults (avg)	children (avg)	babies (avg)	special_request (avg)
No Deposit	88.75662	0.04203897	1.862597	0.1183952	0.00906	0.6514272
Non Refund	212.90889	0.004387468	1.811407	0.0006169	0.00000	0.0017824
Refundable	152.09877	0.024691358	1.907407	0.0308641	0.00000	0.1419753

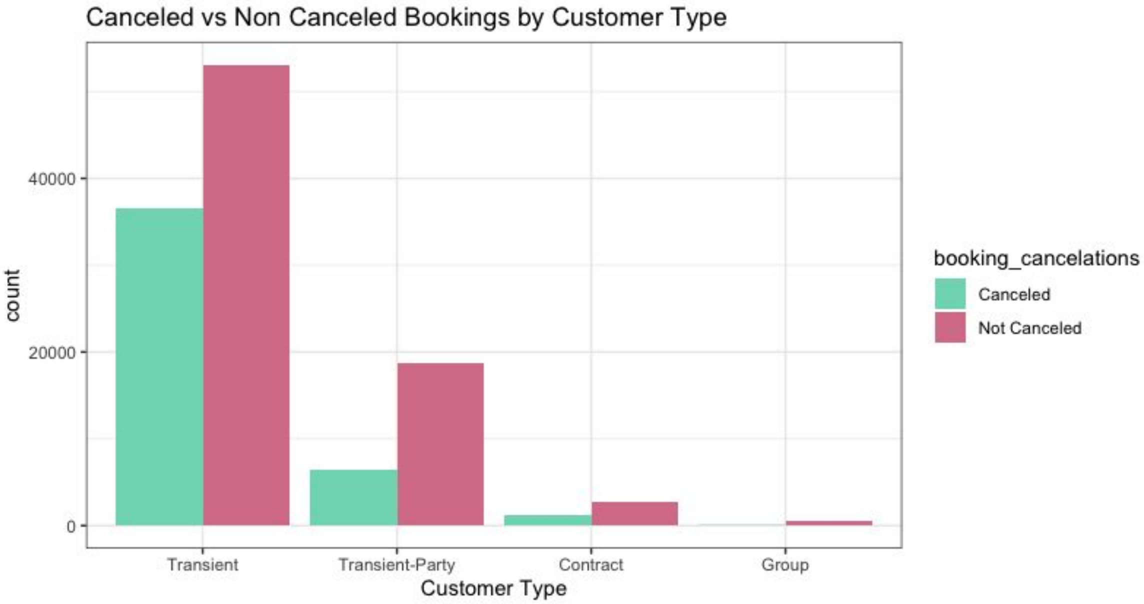
"As an example, if we look at the "Non refundable" canceled bookings in some Asiatic countries and from certain distribution channels, it is possible to understand why so many "Non refundable" bookings are canceled. These bookings are usually made through OTA using false or invalid credit card details. These bookings are issued as support for requests for visas to enter the country"

N. Antonio, A. De Almeida, L. Nunes
Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior, Cornell Hospitality Quarterly 60(4), May, 2019

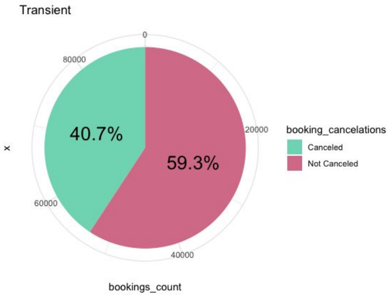
Data at our disposal are not sufficient to confirm what the authors of the paper argue



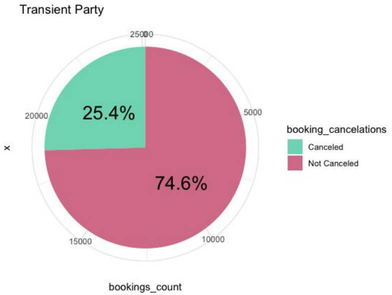




Canceled vs Non Canceled Bookings by Customer Type

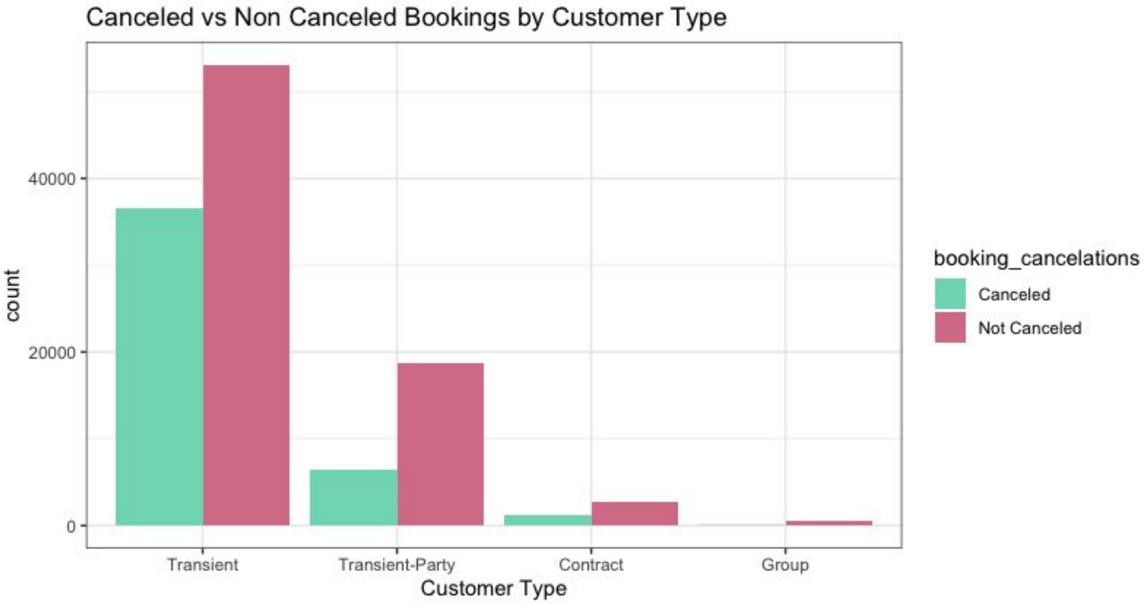


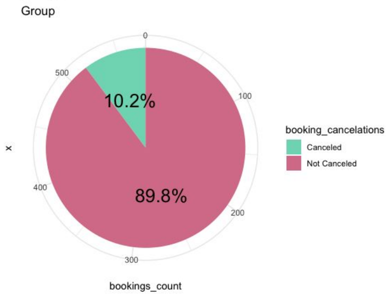
Transient

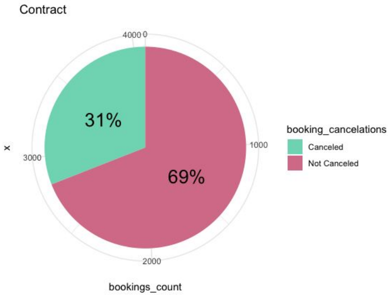


Transient Party

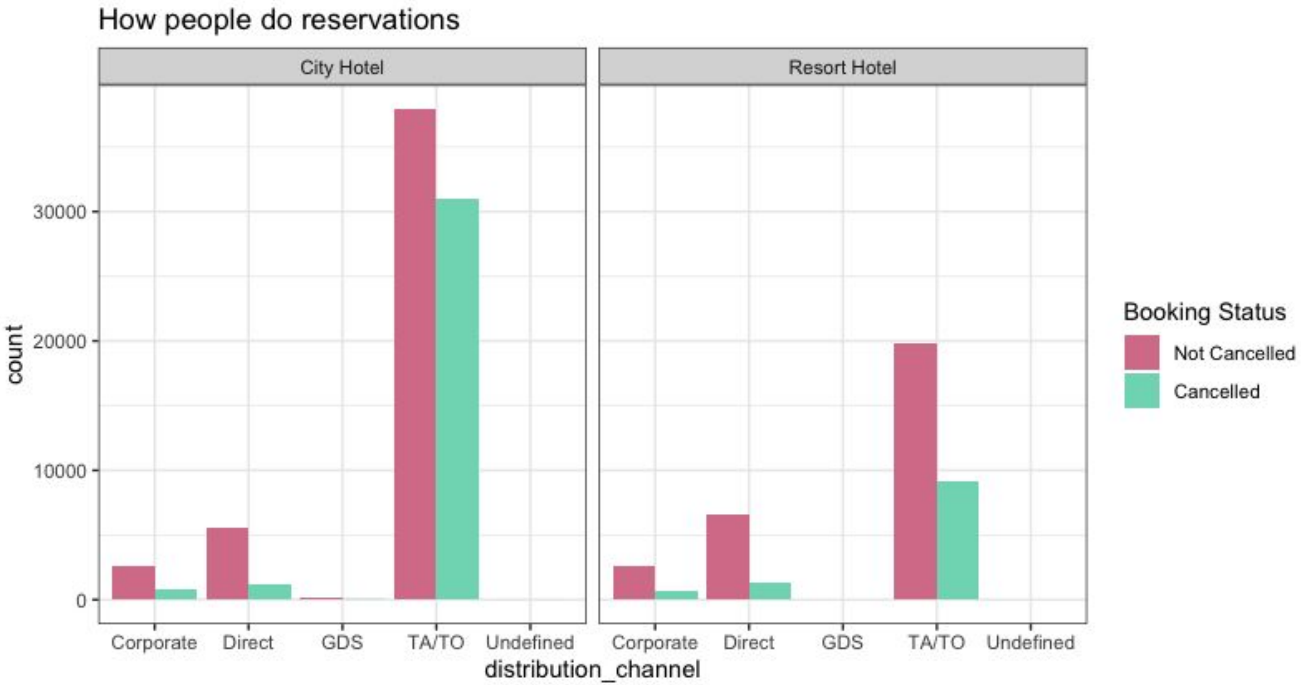
Is there a class of customers for whom the rate of cancellation is higher than the others?







Is there a class of customers where the rate of cancellation is higher than the others?



What is the main distribution channel for the two Hotels?

TA: Travel Agents

TO: Tour Operator

Association measures

1. Association between numerical variables :
   1.1. Correlation

2. Association between categorical variables :
   2.1. Chi-square
   2.2. Cramer’s V

3. Association between binary categorical variables:
   3.1. Relative risk
   3.2. Odds ratio
   3.3. Yule’s Q

Model Data

Implemented models:

- Logistic Regression with the relative selection model
- ridge regression
- lasso regression

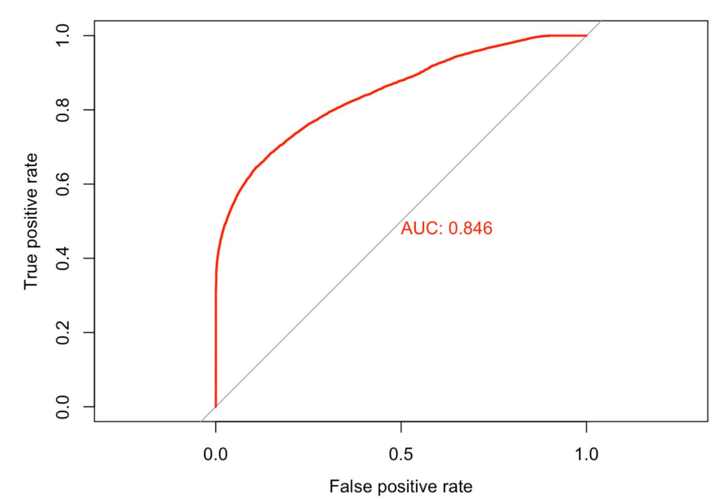
Some initial clarifications about:

- variables we've chosen
- statistical approach to the problem

LOGISTIC REGRESSION

MODEL

- Use of GLM function with family = “binomial”



EVALUATION

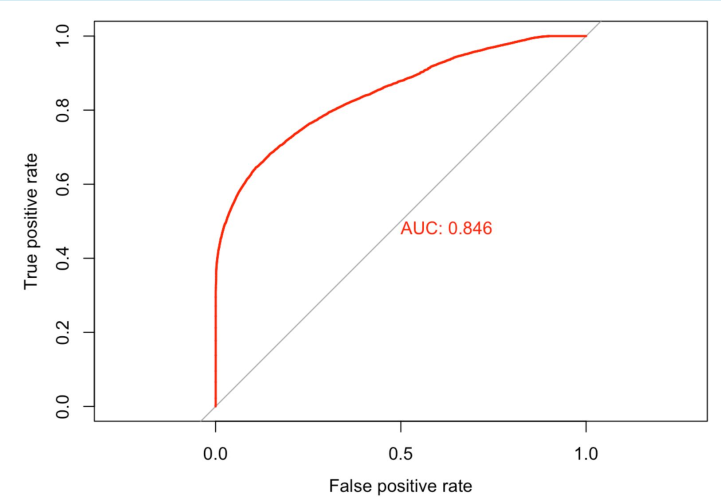
- Many significant variables as months, lead time, total stays…
- Some not-significant variables as parking spaces, babies, children…

AIC: 104011

MODEL SELECTION

STEPAIC FUNCTION

- What is it?
- How does it work?
- Why do we use it?



EVALUATION

- Model chosen: complete model without babies

BEST AIC: 104009

Insert table…

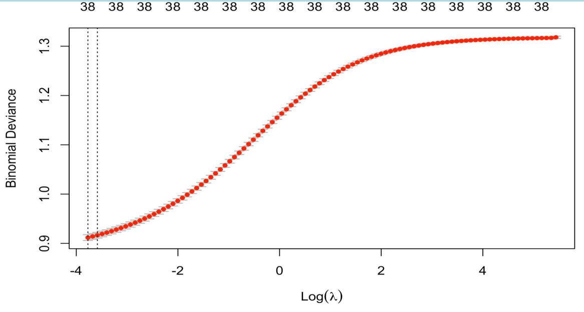
Ridge regression

Model

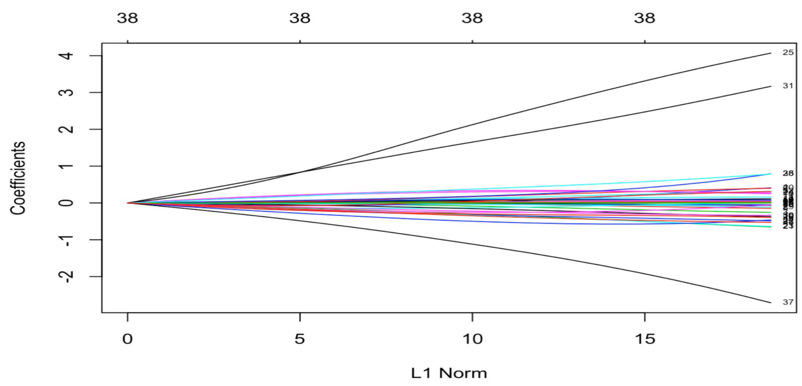
- How does it work?

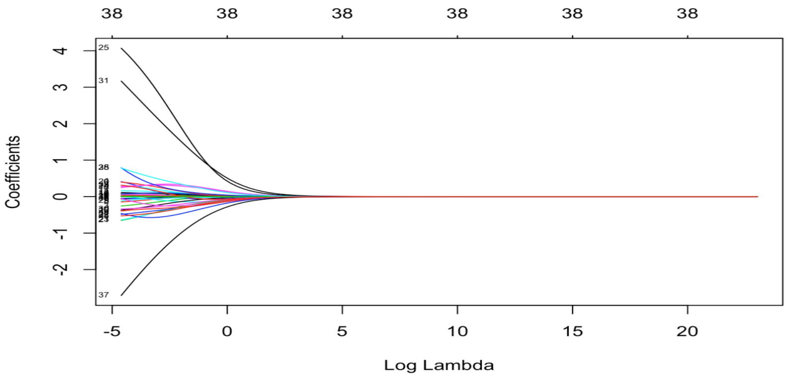
- Why we chose to use it

- Use of cross-validation to choose best lambda



Some useful plots





Evaluation

Confusion Matrix

	Not Canceled	Canceled
Not Canceled	71060	19375
Canceled	4106	24849

Accuracy

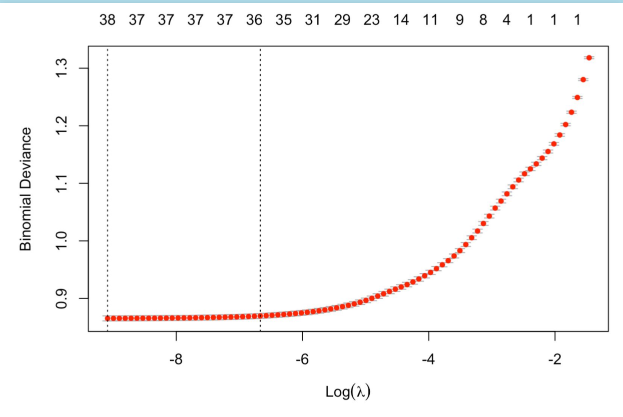
0.7719

```
Sensitivity: 0.7226
Specificity: 0.9808
```

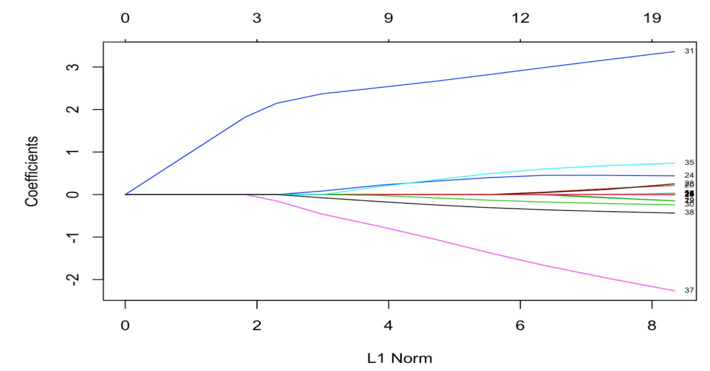
LASSO REGRESSION

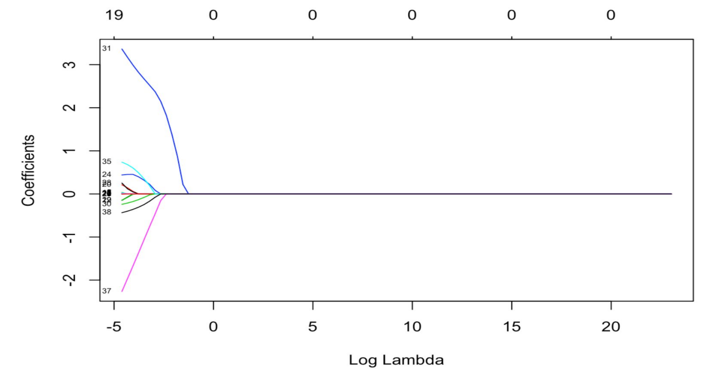
MODEL

- How does it work?
- Why we chose to use it
- Use of cross-validation to choose best lambda



SOME USEFUL PLOTS





Evaluation

Confusion Matrix

	Not Canceled	Canceled
Not Canceled	37445	151
Canceled	14373	7726

Accuracy

0.7567

```
Sensitivity: 0.7226
Specificity: 0.9808
```

Implemented models:

- logistic regression
- LDA
- QDA
- polynomial regression

Some initial clarifications about:

- variables we've chosen
- statistical approach to the problem

- *ADR*
- *lead time*
- *total of special requests*

Logistic regression

Evaluation

- All variables are very significant
- Model doesn’t work in a so good way

ACCURACY: 0.6968

Sensitivity: 0.8675
Specificity: 0.4067

LDA Evaluation

- Both model performances are very similar to the Logistic Regression performance
- Models work in a poor way

QDA Evaluation

ACCURACY: 0.6939

Sensitivity: 0.8724
Specificity: 0.3906

ACCURACY: 0.6885

Sensitivity: 0.8558
Specificity: 0.3960

POLYNOMIAL REGRESSION

MODEL

ADR

+

LEAD TIME

+

TOTAL OF SPECIAL REQUESTS

+

ADR^2

+

LEAD TIME^2

+

TOTAL OF SPECIAL REQUESTS^2

EVALUATION

- All variables are very significant
- Model doesn't work in a so good way, but in a better way respect to all the previous models
- Specificity is low, but higher than previous models

ACCURACY: 0.701

Sensitivity: 0.8452
Specificity: 0.456

Conclusions

Are we able to predict a booking cancellation? ✓