|  |  |  |  |
| --- | --- | --- | --- |
| Attribute name | Attribute description | Datatype | Data values |
| ID | reservation id | object |  |
| n\_adults | number of adults | int64 |  |
| N\_less\_12 | number of children aged less than 12 | int64 |  |
| N\_more\_12 | number of children aged more than 12 | int64 |  |
| Weekend\_nights | number of weekend nights | int64 |  |
| Week\_nights | number of week nights | int64 |  |
| Board\_type |  | object | half board, nan, full board, breakfast, not selected |
| Booked\_tour | indicates whether a tour was included in the reservation | int64 |  |
| Room\_type |  | object | Room\_Type 1-7 |
| Lead\_time | number of days between the reservation date and the arrival date | float64 |  |
| Purchase\_type |  | object |  |
| Repeated | indicates whether the reservation is a repeat reservation | int64 |  |
| N\_p\_cancellation | number of previous reservations that were canceled by the customer prior to the current reservation | int64 |  |
| N\_p\_not\_cancellation | number of previous reservations not canceled by the customer prior to the current reservation | int64 |  |
| Price |  | float64 |  |
| N\_requests | number of special requests made by the guest | int64 |  |
| date | date of the reservation | object |  |
| Is\_canceled | target value, 0 – not canceled, 1 – canceled | int64 | 0 /1 |

**Data cleaning**

We took some cursory looks at the current data – looking generally at the columns and values in each column, and at the distribution of the values for each attribute of the numeric columns:

A graph with numbers and a bar graph

Description automatically generatedA graph with numbers and a bar graph

Description automatically generatedA graph with numbers and a bar

Description automatically generated

A graph with numbers and a bar

Description automatically generatedA graph of numbers and a number

Description automatically generatedA graph of a number of numbers

Description automatically generated with medium confidence

A graph with numbers and a number

Description automatically generatedA graph with numbers and a bar graph

Description automatically generatedA graph with numbers and a bar

Description automatically generated

A graph of a number of numbers

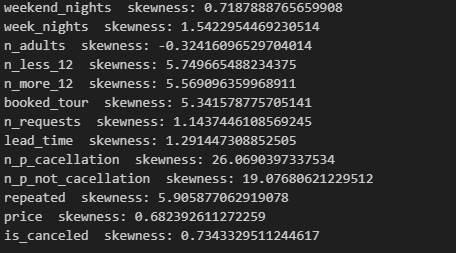
Description automatically generatedA graph with numbers and a bar

Description automatically generatedA graph with numbers and a bar graph

Description automatically generated

A graph with numbers and lines

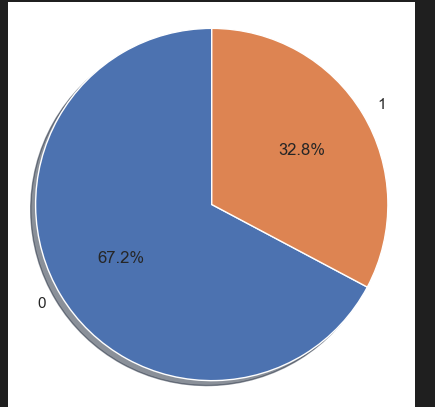
Description automatically generated

**Skewness  
  
**

As we can see from the graphs and from the skweness calculation it’s clear which one is more symmetrical, where some of the data is well distributed like “week\_nights” and which has long tail like “lead\_time”

Checking outliers in the price column looks like there is only one value 560.00, can’t see anything in the lower values.

Number of cancellations:



A graph of a number of bars

Description automatically generated with medium confidence

it looks like the lower the price the less cancelations we get!

**Value cleaning:**

We noticed that there was one value in the dates ’29-02-2018’ that was a different format than the other ones, so we corrected it. We then split the dates into three columns of day, month and year to make it more meaningful.

A graph of a number of months

Description automatically generated with medium confidenceA graph of different colored lines

Description automatically generatedA graph of a number of years

Description automatically generated

We took a quick look at the canceled vs not canceled for each of the date columns just to further explore the data.

**Filling missing values:**

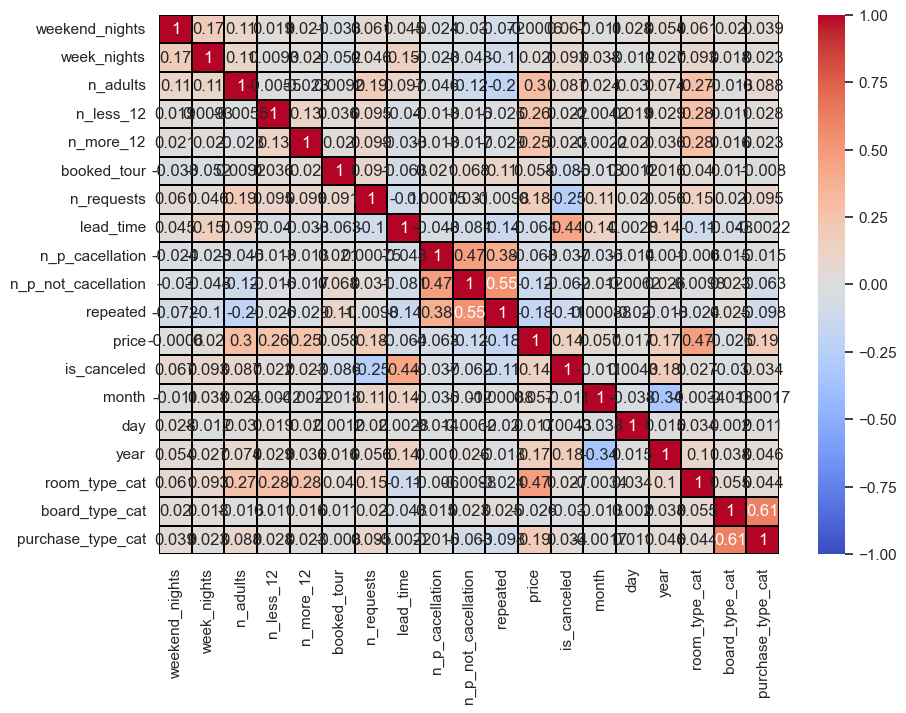
The next thing we did for the data cleaning was to look for missing values and to explore the best way to fill them. Below is the list of all the attributes and the number of values existing for each one.

A screenshot of a computer

Description automatically generated

As is apparent from the list, there were four attributes with missing values: board\_type, lead\_time, purchase\_type, and price.

The first step for us was to see if there was a correlation between one attribute to another that would help us fill in the values to a more precise degree. For this we converted our categorical data of board\_type, room\_type, purchase\_type into continuous values. We then performed a correlation function:



The only meaningful correlation that we found was between the price and the room\_type category – a correlation of 0.47.

**Price:**

Therefore, we filled the missing values of the price using the room\_type category.



These were the distributions of the price per room type.

For the other three categories with missing values, we decided the smartest way to fill them was by maintaining their current distributions:

**Board\_type:**

Before: After:

A graph of a number of people

Description automatically generated with medium confidenceA graph of a number of objects

Description automatically generated with medium confidence

**Lead\_time:**

Before: After:

A graph of a number of lead time

Description automatically generated A graph with numbers and a bar graph

Description automatically generated

**Purchase\_type:**

Before: After:

A graph of a number of blue bars

Description automatically generatedA graph of a number of blue squares

Description automatically generated

**Organizing data in different ways:**

Our next step was to create a binning system for the lead time – as it was a very spread out value. For this we used the min max method and split the lead time into five groups:

A graph with blue squares

Description automatically generated

We also created a column of the total children termed “Num\_Children” to see if that added any meaningful information for our model.

A graph of a number of children

Description automatically generated

From a cursory glance it looks like proportionally those without children our less likely to cancel, but there are a lot less with children so it was hard to really read anything into this.

We also tried other visuals to look at relationships between different attributes and their effect on cancellation – for example, we looked at the lead time:

A screen shot of a graph

Description automatically generated

And here is does appear that a higher lead time (booking further in advance) correlated to more cancellations – but it is difficult to say for certain because so many people book closer to their stay than farther away from their stay.

**Preparing data for models to integers:**

The next thing we did, was to take the categorical attributes and put them into numbers – so that our models could use them. For this purpose we relabeled the year, board\_type, purchase\_type, room\_type columns into “year\_relabel”, “board\_type\_relabel”, “purchase\_type\_relabel”, “room\_type\_relabel”.

**We looked for outliers:**

We found 273 rows of outliers specifically looking in the price attribute and seeing what was 3 standard deviations away from the mean. We chose this method – we could have also chosen the IQR method.

**Scaling/normalizing attributes:**

Scaled different attributes to further prepare for the classifier models - specifically the price attribute and lead time attribute.

We used the min-max for the price and z score normalization for the lead time.

A graph with numbers and lines

Description automatically generated A graph with numbers and a number

Description automatically generated

We then performed a PCA analyses to see if we could glean anything meaningful from that:

A graph with red and blue dots

Description automatically generated

**Final prep for model data:**

We then prepared the data by dropping the columns we did not want to use - 'ID', 'room\_type\_cat', 'board\_type\_cat', 'purchase\_type\_cat', 'room\_type','board\_type', 'purchase\_type', 'lead\_time', 'price', 'year', 'date'. This included data that was unnecessary like the ID column, and also data that we reconfigured for the purpose of the model and no longer needed the original columns.

And we also performed the cleaning on the test data.

**Our metric:**

The metric we chose to focus on is the precision metric. We want to be able to be as precise as possible that the people who canceled did indeed cancel (maximize the true positives out of the total things we categorized as positive). This way we will not have a scenario of double booking where we thought somebody canceled even though they did not, and this way we can better make sure to fill the spots where people canceled without worrying that we are double booking – we can make sure we lose the least amount of money this way by precisely filling in the rooms that are indeed now open.

**Classifiers:**

Decision tree classifier –

As is apparent from the data – our target data is not balanced – there are less cancellations than there are cancellations:

A graph with numbers and lines

Description automatically generated

First we calculated the effect of the depth on the decision tree precision (our metric of interest) before balancing the data:

A graph showing a line graph

Description automatically generated

We decided on a max depth of about 17 for our decision tree as this seemed pretty in line with the results from the above graph. We also performed a grid search to receive the best parameters for the decision tree – the min\_samples\_split and the min\_samples\_leaf.

We then looked at the classifiers using the ADASYN balancing method:

Gaussian Naïve Bayes classifier –