**Hotel Cancellation Prediction**

Karim AbouDaoud
900212779
Department of Mathematics and Actuarial Science
The American University in Cairo
Cairo, Egypt
karimaboudaoud@aucegypt.edu

Youssef Nakhla
900201430
Department of Mathematics and Actuarial Science
The American University in Cairo
Cairo, Egypt
youssef_n9212@aucegypt.edu

## Description of The Dataset

After an extensive exploration across various data science platforms, including Kaggle, we have identified two datasets relating to our research focus. The first dataset centers on hotel reservations, encompassing comprehensive details across multiple categories, including dates, guest demographics, and hotel management intricacies. This dataset meticulously records booking dates, specifying the day, week, month, and year. It further highlights the guest count, categorizing individuals into adults, teenagers, and children, while also providing insights into the guest's nationality, recurrence status, and the type of room reserved.

Moreover, the dataset dissects into essential aspects of hotel management, disclosing details about the market segment responsible for concluding the guest transaction, whether through direct or corporate channels. Additionally, it documents the agent responsible for finalizing the deal, adding a layer to our understanding. The feature variables are of type categorical variables and numerical variables. The categorical variables are: “hotel”, “meal”, “country”, “market segment”, “distribution channel”, “is repeated guest”, “reserved room type”, “assigned room type”, “deposit type”, “customer type”, and “reservation status”. The numerical variables are: “lead time”, “arrival date year”, “arrival date month”, “arrival date week”, “arrival date day”, “stays in weekend nights”, “stays in week nights”, “adults”, “children”, “babies”, “previous cancellations”, “previous booking not canceled”, “booking changes”, “days in waiting list”, “adr”, “required car parking spaces”, and “total of special requests”. The label, which is the variable to be predicted, is the “is_canceled” variable, which identifies the status of each reservation as either canceled or not canceled.

The data set contains 32 variables and 119,000 observations. The table below has three columns: the first column states the variable name, the second column states the data type as Categorical or Numerical, and the third column gives a brief description and provides the unit of measurement of the variable.

| **Variable** | **Type** | **Description/Unit of Measurement** |
|---|---|---|
| Hotel | Categorical | H1: Resort Hotel & H2: City Hotel |
| Reservation Canceled | Binary | 0 refers to no, 1 refers to yes |
| Lead Time | Numerical | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| Arrival Year | Numerical | Year of arrival date |
| Arrival Month | Categorical | Month of arrival date |
| Week Number Arrival | Numerical | Week number of year for arrival date |
| Arrival Date Day of Month | Numerical | Day of arrival date |
| Stays in weekend nights | Numerical | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| Stays in week nights | Numerical | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| Adults | Numerical | Number of adults |
| Children | Numerical | Number of children |
| Babies | Numerical | Number of babies |

| | | |
|---|---|---|
| Meal | Categorical | Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC: no meal package; BB: Bed & Breakfast; HB; FB : Full board |
| Country | Categorical | Country of origin |
| Market Segment | Categorical | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| Distribution Channel | Categorical | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| Repeated Guest | Binary | 0 refers to no, 1 refers to yes |
| Previous Cancellations | Numerical | Number of previous bookings that were canceled by the customer prior to the current booking |
| Previous Bookings not Canceled | Numerical | Number of previous bookings not canceled by the customer prior to the current booking |
| Reserved Room Type | Categorical | Code of room type reserved |
| Assigned Room Type | Categorical | Code for the type of room assigned to the booking |

| | | |
|---|---|---|
| Bookings Changes | Numerical | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation |
| Deposit Type | Categorical | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay. |
| Agent | Numerical | ID of the travel agency that made the booking |
| Company | Numerical | ID of the company/entity that made the booking or responsible for paying the booking. |
| Days in Waiting List | Numerical | Number of days the booking was in the waiting list before it was confirmed to the customer |

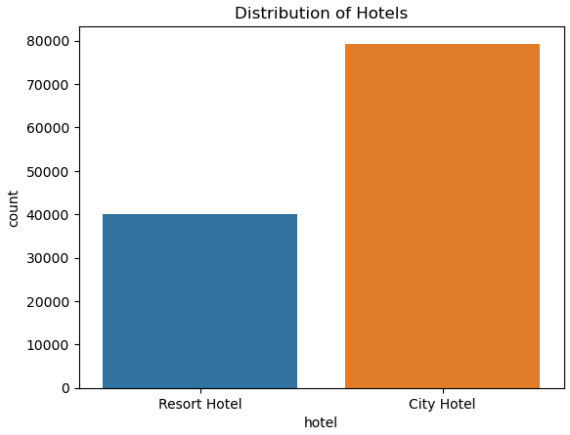
| | | |
|---|---|---|
| Customer Type | Categorical | Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking |
| ADR | Numerical | Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights |
| Required Car Parking Spaces | Numerical | Number of car parking spaces required by the customer |
| Total Special Requests | Numerical | Number of special requests made by the customer |
| Reservation Status | Categorical | Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why |
| Reservation Status Date | Numerical | Date at which the last status was set |

## Analysis of Feature Variables

The different features of the dataset will be analyzed in terms of their relationship with the label "is_cancelled", missing values, unique values for nominal features, semantic importance / relevance and statistical distribution.

## 1. Hotel

This feature is a categorical variable that states whether the booking occurred in either a Resort Hotel or City Hotel in Portugal. The feature does not seem to have any missing values and also it seems to be that the number of City Hotel reservations are much greater than that of the number of Resort Hotel reservations.




The question of which type of hotel has a higher cancellation rate would stand, hence after examination it was clear that City Hotels have higher

cancellation rates. This may be explained by the fact that they are larger in number within the dataset, or that City Hotels tend to attract more traffic of business travelers who may need to readjust their plans. As well as the fact that City Hotels may be affected by local events, conferences, etc. The cancellation rate for City Hotels is 41.72%, while the cancellation rate for Resort Hotels is 27.76%.

```
hotel
City Hotel      0.417270
Resort Hotel    0.277634
Name: is_canceled, dtype: float64
```

Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

```
Observed Contingency Table:
hotel        City Hotel  Resort Hotel
is_canceled
0                 46228         28938
1                 33102         11122

Chi-squared Test Statistics:
Chi2: 2224.924903923313
P-value: 0.0
Degrees of Freedom: 1

Expected Frequencies:
[[49944.8762878 25221.1237122]
 [29385.1237122 14838.8762878]]
```

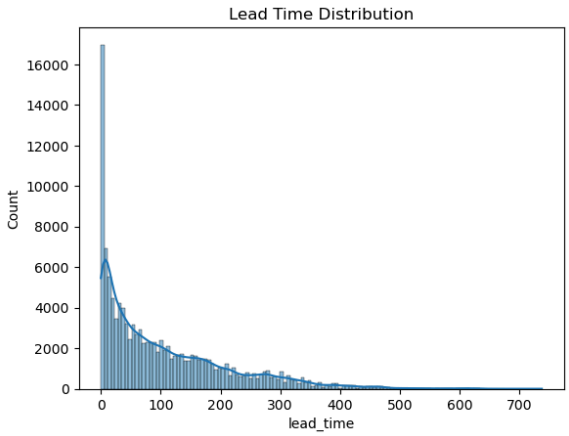
Since the p-value is 0 which is less than the significance level, we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and 'hotel' variables.

## 2. Lead Time

The feature lead time represents the number of days that elapsed between the entering date of the booking and the arrival date. It does not contain any missing values and seems to have the highest correlation with the label. The statistical summary for this feature is observed below.

```
count    119390.000000
mean        104.011416
std         106.863097
min           0.000000
25%          18.000000
50%          69.000000
75%         160.000000
max         737.000000
Name: lead_time, dtype: float64
```

Furthermore, the distribution of this variable was plotted and it appears to be skewed as seen below.




To further examine the relationship between the lead time and the cancellation rate, the lead time was temporarily changed to show it in the form of months to see whether the increase of lead time affected cancellation of reservations. The results showed that as lead time increases, the percentage of reservations that are canceled generally increases. Months with longer lead times (e.g., 10, 11, 14, 16, 17) tend to have higher percentages of cancellations. This could suggest that guests booking well in advance are more likely to cancel. There seems to be some variability in cancellation rates across different lead time months. For example, there is a dip in cancellations around month 5 and a peak around month 14. Yet, generally it is evident that as the lead time increases, it is more likely that the reservation is canceled.

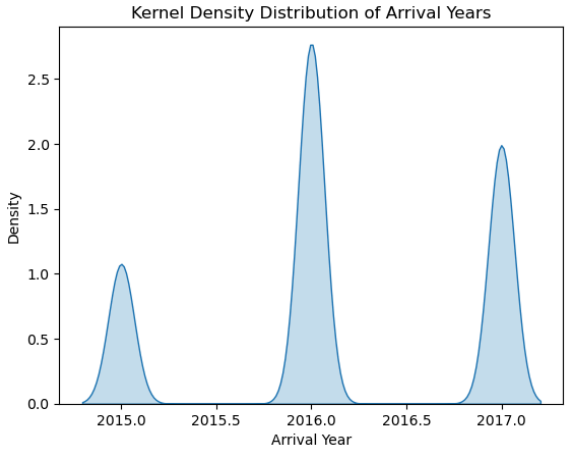
| is_canceled | 0 | 1 |
|---|---|---|
| lead_time_month | | |
| 0 | 81.754146 | 18.245854 |
| 1 | 63.659629 | 36.340371 |
| 2 | 60.262455 | 39.737545 |
| 3 | 55.961814 | 44.038186 |
| 4 | 56.403270 | 43.596730 |
| 5 | 53.797548 | 46.202452 |
| 6 | 55.274188 | 44.725812 |
| 7 | 53.078508 | 46.921492 |
| 8 | 44.854651 | 55.145349 |
| 9 | 36.176385 | 63.823615 |
| 10 | 30.651620 | 69.348380 |
| 11 | 29.441341 | 70.558659 |
| 12 | 42.126789 | 57.873211 |
| 13 | 37.500000 | 62.500000 |
| 14 | 27.255639 | 72.744361 |
| 15 | 35.159011 | 64.840989 |
| 16 | 17.073171 | 82.926829 |
| 17 | 18.852459 | 81.147541 |
| 18 | 25.274725 | 74.725275 |

## 4. Arrival Date Year

This feature indicates the year of the arrival date where the only years available in the dataset are 2015, 2016, and 2017; and their respective counts are highlighted below.

```
2016    56707
2017    40687
2015    21996
Name: arrival_date_year, dtype: int64
```

The variable presented no missing values and the statistical distribution plot shows that each year seems to approximately be normally distributed.




Finally, the year 2017 seems to have the highest cancellation rate with a rate of 38.69%, followed by the year 2015 with a rate of 37.02% and then the year 2016 with a rate of 35.86%.

```
arrival_date_year
2015    0.370158
2016    0.358633
2017    0.386979
Name: is_canceled, dtype: float64
```

## 5. Arrival Date Month

This feature indicates the month of the arrival date where they are labeled January, February, etc. It does not present any missing values and the respective counts of each month is shown below.

```
August       13877
July         12661
May          11791
October      11160
April        11089
June         10939
September    10508
March         9794
February      8068
November      6794
December      6780
January       5929
Name: arrival_date_month, dtype: int64
```

Based on these counts it is obvious that the season of Summer has the greatest amount of reservations which would make sense as the Portuguese hotel industry may be getting the most traffic during the summer months. Furthermore, the counts of each season are listed below.

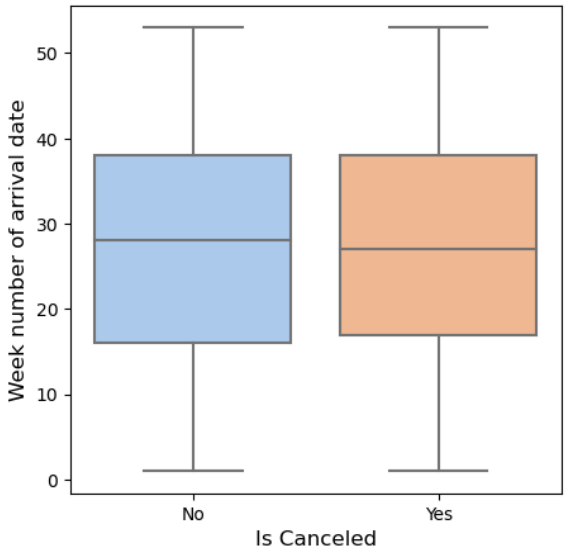
```
Summer    37477
Spring    32674
Autumn    28462
Winter    20777
Name: arrival_date_month, dtype: int64
```

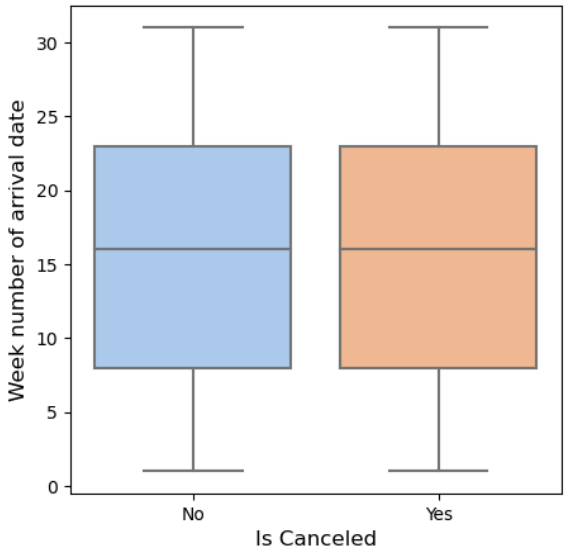
To further examine the relationship between the feature and the label, a chi-squared test was conducted where H0: there is no correlation and H1: a correlation exists between the variables. The test yielded a very small p-value which is much lower than the significance level and hence we reject the null hypothesis and conclude that there is a correlation between the feature and the label. Moreover, the highest cancellation percentages are observed in January, February, and December, while the lowest percentages are in November and August.

| arrival_date_month | is_canceled | percentage |
|---|---|---|
| April | 0 | 59.202814 |
| | 1 | 40.797186 |
| August | 0 | 62.246883 |
| | 1 | 37.753117 |
| December | 0 | 65.029499 |
| | 1 | 34.970501 |
| February | 0 | 66.584036 |
| | 1 | 33.415964 |
| January | 0 | 69.522685 |
| | 1 | 30.477315 |
| July | 0 | 62.546402 |
| | 1 | 37.453598 |
| June | 0 | 58.542828 |
| | 1 | 41.457172 |
| March | 0 | 67.847662 |
| | 1 | 32.152338 |
| May | 0 | 60.334153 |
| | 1 | 39.665847 |
| November | 0 | 68.766559 |
| | 1 | 31.233441 |
| October | 0 | 61.953405 |
| | 1 | 38.046595 |
| September | 0 | 60.829844 |
| | 1 | 39.170156 |

## 6. Week Number & Day of The Month

The two features are numerical variables that indicate the day and the week number of the guest's arrival date. Both features do not present any missing values, yet they have almost insignificant correlation coefficients with the label. Furthermore, box plots were plotted to observe the relationships between each variable and the label and the results show that these variables are not factors that determine whether a reservation is canceled or not. This can be observed on the box plots as there is no difference between their statistical distributions regardless of whether the booking was canceled or not.

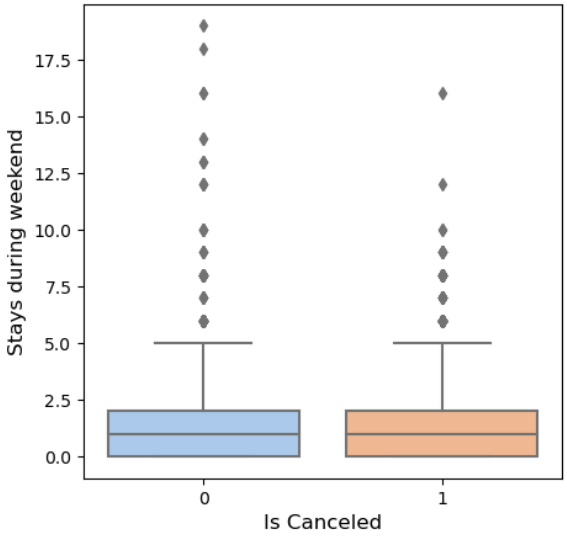


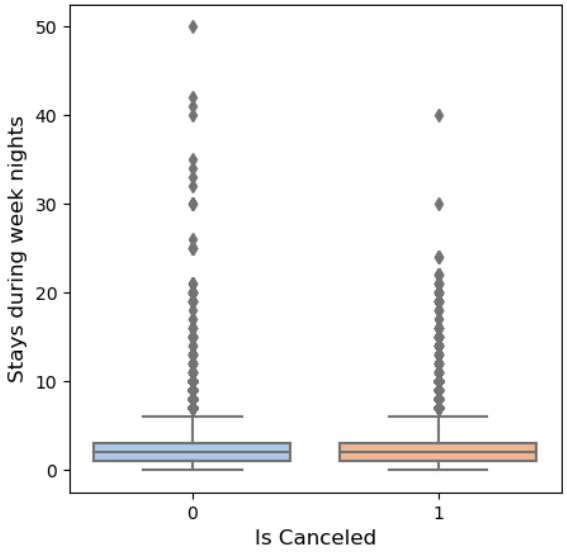




## 7. Stays in Weekends & Weeknights

The two features are numerical variables that indicate the number of weekend nights (Sunday and Saturday) and week nights (Monday to Friday) that is in the reservation. Both variables do not have any missing values and tend to have a very low correlation coefficient in relation to the label. The correlation of these two variables together is 0.498 which is a very high number and therefore may present a collinearity problem. An appropriate course of action would be to drop one of the variables

since the other is basically redundant. In addition, upon plotting both variables against the label, it can be seen that both variables are insignificant and do not contribute to the prediction of the reservation cancellation.

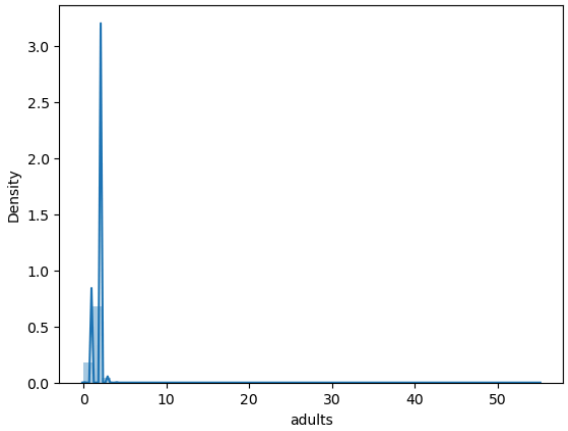



According to the above boxplots, the number of stay during weekends and weekdays has nothing to do with a booking cancellation. The two boxplots have a very similar distribution regardless of whether or not the booking was canceled.

## 8. Adults

This feature indicates the number of adults under the reservation and it does not seem to present any missing values. Most reservations seem to have 1-3 adults in record and also this feature is statistical appears to be be skewed.




To further examine the relationship between the number of adults and the label, the two features were put against each other only to show that the number of adults does not seem to impact the cancellation of the booking. Bookings with many adults are as almost likely to be canceled as reservations with only 1 adult.

| adults | is_canceled | count |
|---|---|---|
| 0 | 0 | 72.952854 |
| | 1 | 27.047146 |
| 1 | 0 | 71.016633 |
| | 1 | 28.983367 |
| 2 | 0 | 60.684657 |
| | 1 | 39.315343 |
| 3 | 0 | 65.317639 |
| | 1 | 34.682361 |
| 4 | 0 | 74.193548 |
| | 1 | 25.806452 |
| 5 | 1 | 100.000000 |
| 6 | 1 | 100.000000 |
| 10 | 1 | 100.000000 |
| 20 | 1 | 100.000000 |
| 26 | 1 | 100.000000 |
| 27 | 1 | 100.000000 |
| 40 | 1 | 100.000000 |
| 50 | 1 | 100.000000 |
| 55 | 1 | 100.000000 |

## 9. Children & Babies

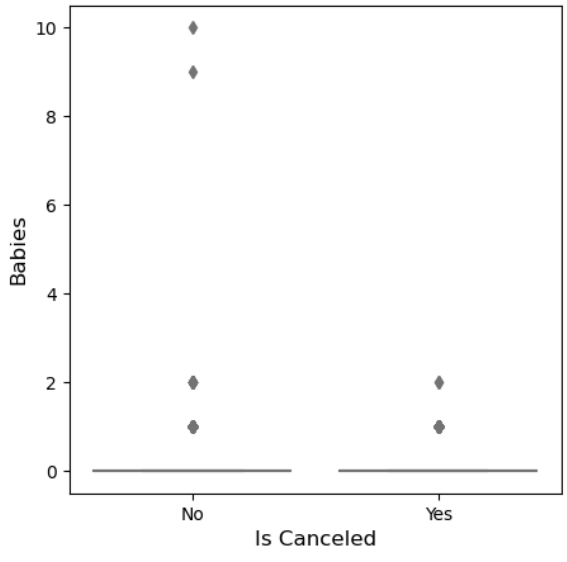
These two features represent the number of children and babies present in the reservation, where the dataset source does not appear to specify the age

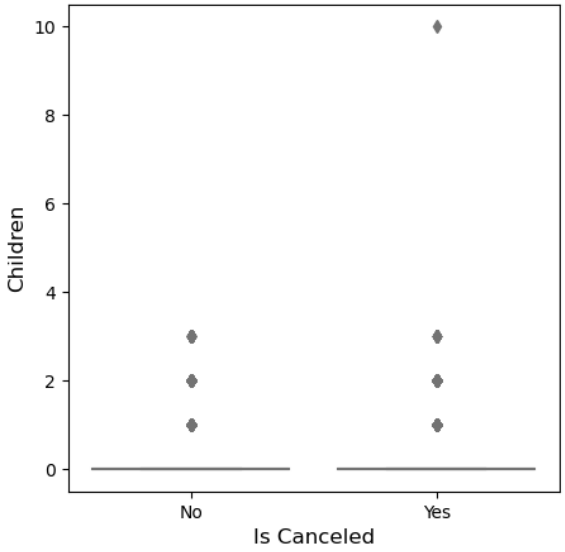
ranges corresponding to each variable. The children feature presents 4 missing values and the babies feature does not present any missing data. Most reservations did not seem to have any children or babies as it can be seen in the counts below.

```
0.0     110796
1.0       4861
2.0       3652
3.0         76
10.0         1
Name: children, dtype: int64
```

```
0     118473
1        900
2         15
10         1
9          1
Name: babies, dtype: int64
```

Furthermore, there does not appear to be a significant correlation between the number of children or babies on the cancellation of the reservation. According to the box plots below, the number of children and babies does not play a role in determining a cancellation. The two boxplots have a very similar distribution regardless of whether or not the booking was canceled.




## 10. Meal Type

The meal feature is a categorical variable differentiating between the type of meal booked and it does not present any missing values. The variable has four categories: HB: half board, BB: bed and breakfast, SC: self-catering and FB: full board. In addition, the category “Undefined” is the same as “SC” and it will be adjusted later in the data cleaning part of the report. The most common meal type was BB and the least common was FB which would make sense as it is the most expensive type of meal plan.

```
BB           92310
HB           14463
SC           10650
Undefined     1169
FB             798
Name: meal, dtype: int64
```

A chi-squared test was conducted to determine the existence of the a relationship between the meal type and the label where H0: there is no correlation and H1: a correlation exists between the variables. The low p-value indicates that the observed association between meal type and cancellation status is highly unlikely to be due to random chance and hence we reject the null hypothesis.

```
Observed Contingency Table:
is_canceled      0      1
meal
BB           57800  34510
FB             320    478
HB            9479   4984
SC            6684   3966
Undefined      883    286

Chi-squared Test Statistics:
Chi2: 304.23617668200444
P-value: 1.3212351959124216e-64
Degrees of Freedom: 4

Expected Frequencies:
[[58116.87293743 34193.12706257]
 [  502.40780635   295.59219365]
 [ 9105.66930229  5357.33069771]
 [ 6705.06658849  3944.93341151]
 [  735.98336544   433.01663456]]
```

The percentages provide insights into the cancellation rates for different meal types. Reservations with Full Board (FB) have a higher cancellation rate (59.90%), while reservations with Undefined meal types have the lowest cancellation rate (24.47%).

| meal | is_canceled | count |
|---|---|---|
| BB | 0 | 62.615101 |
| | 1 | 37.384899 |
| FB | 0 | 40.100251 |
| | 1 | 59.899749 |
| HB | 0 | 65.539653 |
| | 1 | 34.460347 |
| SC | 0 | 62.760563 |
| | 1 | 37.239437 |
| Undefined | 0 | 75.534645 |
| | 1 | 24.465355 |

## 11. Country

The country feature indicates the nationality of the guests presented in the ISO code format and it presented many (488) missing values that will be dealt with in the data cleaning part of the report. There exists 177 unique values in this feature and their respective counts can be seen below. Since the dataset presents hotels in Portugal, it is expected that the largest proportion of the guests are locals.

```
PRT    48590
GBR    12129
FRA    10415
ESP     8568
DEU     7287
       ...
DJI        1
BWA        1
HND        1
VGB        1
NAM        1
Name: country, Length: 177, dtype: int64
```

A feature with 177 unique values is nearly impossible and definitely very time consuming to analyze and hence it very difficult to make rigorous and accurate inferences. The data pre-processing section of the report discusses the approach that was implemented to deal with this problem.

## 12. Market Segment

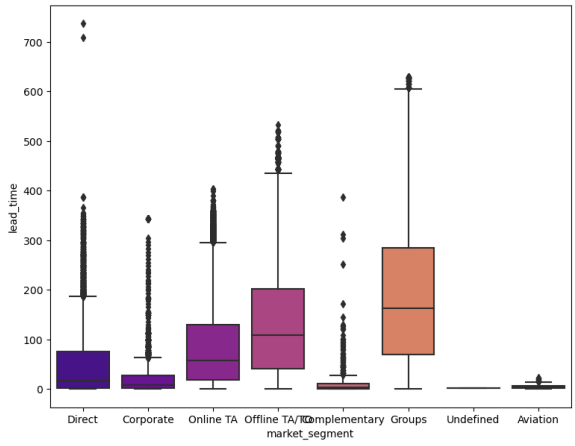
Market segment is a categorical feature that differentiates between different market segment designations as in the how each guest made the reservation with zero missing values. The variable has 7 categories: online TA (travel agent), offline TA, groups, direct, corporate, complementary and aviation. Most guests made reservations through Online TA as shown below.

```
Online TA        56477
Offline TA/TO    24219
Groups           19811
Direct           12606
Corporate         5295
Complementary      743
Aviation           237
Undefined            2
Name: market_segment, dtype: int64
```

A chi-squared test was conducted to determine the existence of the a relationship between the meal type and the label where H0: there is no correlation and H1: a correlation exists between the variables. The results show that there is strong evidence to reject the null hypothesis, indicating that there is a significant association between the market segment and the cancellation status since the p-value is less than the level of significance. Moreover, the highest cancellation rate (61%) was among the market segment category “groups” and the lowest cancellation rate (13%) was among the market segment “Complimentary”.

| market_segment | is_canceled | count |
|---|---|---|
| Aviation | 0 | 78.059072 |
| | 1 | 21.940928 |
| Complementary | 0 | 86.944818 |
| | 1 | 13.055182 |
| Corporate | 0 | 81.265345 |
| | 1 | 18.734655 |
| Direct | 0 | 84.658099 |
| | 1 | 15.341901 |
| Groups | 0 | 38.937964 |
| | 1 | 61.062036 |
| Offline TA/TO | 0 | 65.683967 |
| | 1 | 34.316033 |
| Online TA | 0 | 63.278857 |
| | 1 | 36.721143 |
| Undefined | 1 | 100.000000 |

The reasons behind these cancellation rates were explored and it was found that market segments with a respectively high cancellation rate also had a high lead time which reiterates the conclusion made before that the higher the lead time the more likely it is for a reservation to be cancelled.




## 13. Distribution Channel

This feature is a categorical variable that differentiates between different hotel booking channels and does not appear to have any missing values. The feature has 5 categories; TA/TO (travel agents/Tour operators), Direct, corporate and GDS (global distribution system), and Undefined. Their respective counts are shown below with TA/TO being the most frequent channels.

```
TA/TO        97870
Direct       14645
Corporate     6677
GDS            193
Undefined        5
Name: distribution_channel, dtype: int64
```

A chi-squared test was conducted to determine the existence of a relationship between the meal type and the label where H0: there is no correlation and H1: a correlation exists between the variables. The results of the test show that there is strong evidence to reject the null hypothesis, indicating that there is a significant association between the distribution channel and the cancellation status because the p-value is less than the level of significance.

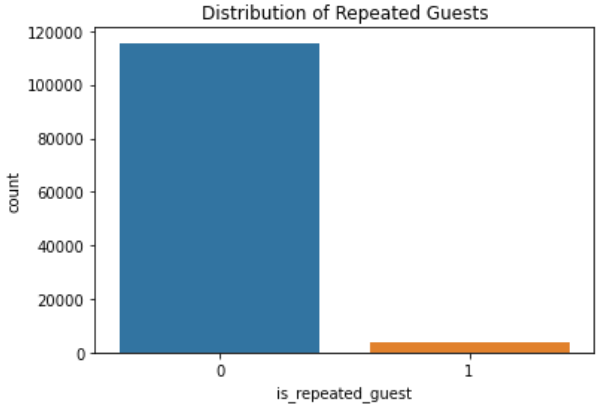
```
Observed Contingency Table:
is_canceled                0      1
distribution_channel
Corporate               5203   1474
Direct                 12088   2557
GDS                      156     37
TA/TO                  57718  40152
Undefined                  1      4

Chi-squared Test Statistics:
Chi2: 3745.794123751679
P-value: 0.0
Degrees of Freedom: 4

Expected Frequencies:
[[4.20373048e+03 2.47326952e+03]
 [9.22025354e+03 5.42474646e+03]
 [1.21509657e+02 7.14903426e+01]
 [6.16173584e+04 3.62526416e+04]
 [3.14791859e+00 1.85208141e+00]]
```

Furthermore, given that tour agents served as the primary booking channel, it is reasonable to observe that they also exhibited the highest cancellation rates, reaching nearly 41%. Notably, this cancellation rate is twice as significant as the cancellation rates observed for other distribution channels.

```
                                        count
distribution_channel is_canceled
Corporate            0             77.924217
                     1             22.075783
Direct               0             82.540116
                     1             17.459884
GDS                  0             80.829016
                     1             19.170984
TA/TO                0             58.974149
                     1             41.025851
Undefined            0             20.000000
                     1             80.000000
```

## 14. Repeated Guest Analysis

This feature is a binary variable with only two values; 0 represents a non repeated guest, and 1 represents a repeated guest. This variable does not contain any missing values. Furthermore, it was essential to examine the number of repeated guests and calculate the percentage of repeated and unrepeated guests. Therefore, I used a barplot to display the number of repeated and unrepeated guests.




As we can see, the number of unrepeated guests is way higher than the number of repeated guests. This makes us ask the question of whether this very high rate of unrepeated guests comes from City Hotels or Resort Hotels.

Therefore, we calculated the percentage of repeated guests for each type of hotel and we then subtracted it from 1 to obtain the number of unrepeated guests.

```
hotel
City Hotel      0.025615
Resort Hotel    0.044383
Name: is_repeated_guest, dtype: float64
```

For City Hotels, 1-0.025615 = 0.974

For Resort Hotels, 1-0.044383 = 0.956

Therefore, the percentage of unrepeated guests for city hotels is 97.4% and the percentage of unrepeated guests for resort hotels is 95.6%. This shows that there is no difference between City and Resort Hotels in terms of repeated guests. Therefore, we cannot conclude that there is a hotel type better than the other, but we can conclude that in general guests do not tend to return back to their previously booked hotels in Portugal.

Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

```
Observed Contingency Table:
is_repeated_guest      0     1
is_canceled
0                  71908  3258
1                  43672   552


Chi-squared Test Statistics:
Chi2: 857.4063180373694
P-value: 1.7841252215934033e-188
Degrees of Freedom: 1

Expected Frequencies:
[[72767.28603736  2398.71396264]
 [42812.71396264  1411.28603736]]
```

Since the p-value is 1.784e-188 which is less than the significance level (0.05), we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and 'is_repeated_guest' variables.

## 15. Previous Cancellations Analysis

This feature is a numerical variable that states the number of previous bookings that were canceled by the customer prior to the current booking. There are no missing values for this variable. The statistical summary for this feature is observed below.

```
count    119390.000000
mean          0.087118
std           0.844336
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max          26.000000
Name: previous_cancellations, dtype: float64
```

The statistical summary shows that the maximum number of cancellations was 26 times by one entity. This shows that this entity might be a company because it is not logical for a person to cancel 26 times and if so he would be then banned from registering again.
Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

```
Chi-squared Test Statistics:
Chi2: 9274.833707930893
P-value: 0.0
Degrees of Freedom: 14

Expected Frequencies:
[[7.10837792e+04 3.80961107e+03 7.30317112e+01 4.09229416e+
  1.95170952e+01 1.19620906e+01 1.38508418e+01 2.20354301e+
  7.55500461e+00 8.81417204e+00 1.19620906e+01 6.29583717e-
  3.02200184e+01 1.57395929e+01 1.63691766e+01]
 [4.18222208e+04 2.24138893e+03 4.29682888e+01 2.40770584e+
  1.14829048e+01 7.03790937e+00 8.14915822e+00 1.29645699e+
  4.44499539e+00 5.18582796e+00 7.03790937e+00 3.70416283e-
  1.77799816e+01 9.26040707e+00 9.63082335e+00]]
```

Since the p-value is 0 which is less than the significance level, we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and 'hotel' variables.

## 16. Previous Bookings not Canceled Analysis

This feature is a numerical variable that states the number of previous bookings that were not canceled by the customer prior to the current booking. There are no missing values for this variable. The statistical summary for this feature is observed below.

```
count    119390.000000
mean          0.137097
std           1.497437
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max          72.000000
Name: previous_bookings_not_canceled, dtype: float64
```

The summary shows that the maximum number of previous bookings not canceled is 72 which shows the consistency of this certain client.
In order to calculate whether there is a relationship between this variable and our main variable that we are testing the whole data on (is canceled variable), we calculated the correlation coefficient between both variables. Below is the output.

```
-0.05735772316594613
```

The correlation coefficient shows a weak negative correlation between both variables. The negative correlation coefficient makes a lot of sense because if a customer has not canceled the previous bookings, it is more likely that they will not cancel this current booking. If we look at the correlation coefficient below between previous cancellations and current booking cancellations, we can see that it is stronger and definitely a positive correlation because it makes a lot of sense that if a person cancels the previous bookings, it is more likely that they will cancel the current one.

```
0.1101328082228435
```

Afterwards, we calculated the percentage of previous cancellations and previous bookings not canceled.

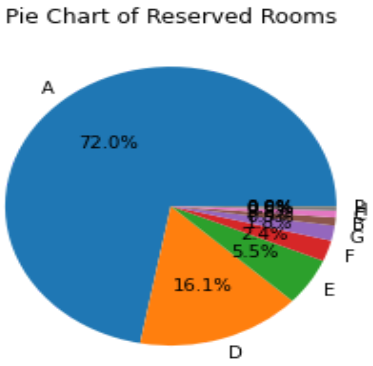
```
Previous Cancellations Percentage: 0.3885464529866637
Previous Bookings Not Cancelled Percentage: 0.6114535470133363
```

It is clear that the previous bookings not canceled are more than the canceled bookings which show that in general people do not cancel their bookings.

## 17. Reserved Room Type Analysis

This feature is a categorical variable that states the code of the room reserved. There are no missing values for this variable. The number of each room booked can be seen down below.

```
A    85994
D    19201
E     6535
F     2897
G     2094
B     1118
C      932
H      601
P       12
L        6
Name: reserved_room_type, dtype: int64
```




Above is a pie chart showing the highest percentages of assigned rooms and above the pie chart are the real values of each room. It is clear that room A is by far the most requested room by clients because 8,994 clients have requested it. The second highest is room D with only 19,201 requests. There are several reasons for why room A is the highest requested room in all hotels. In general, room A in hotels is identified as the room with the best view, or it may be the room that fits the amount of guests booking together or it could be the most recommended room on social media.

Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

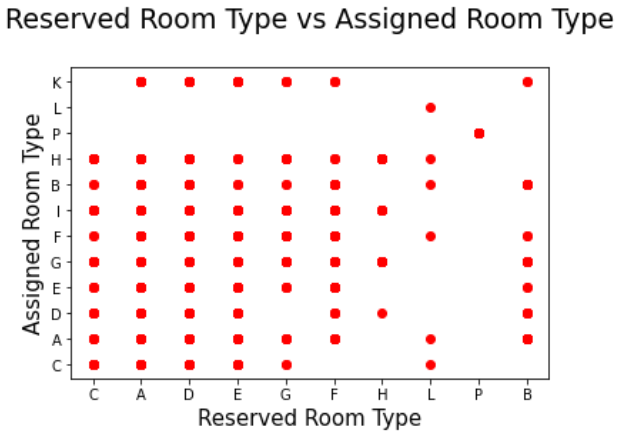
```
Observed Contingency Table:
is_canceled              0      1
reserved_room_type
A                    52364  33630
B                      750    368
C                      624    308
D                    13099   6102
E                     4621   1914
F                     2017    880
G                     1331    763
H                      356    245
L                        4      2
P                        0     12


Chi-squared Test Statistics:
Chi2: 647.8350973363271
P-value: 1.121956218424043e-133
Degrees of Freedom: 9

Expected Frequencies:
[[5.41404222e+04 7.03874596e+02 5.86772024e+02 1.20886370e+04
  4.11432959e+03 1.82390403e+03 1.31834830e+03 3.78379814e+02
  3.77750230e+00 7.55500461e+00]
 [3.18535778e+04 4.14125404e+02 3.45227976e+02 7.11236305e+03
  2.42067041e+03 1.07309597e+03 7.75651696e+02 2.22620186e+02
  2.22249770e+00 4.44499539e+00]]
```

Since the p-value is 1.219e-133 which is less than the significance level (0.05), we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and 'reserved_room_type' variables.

## 18. Assigned Room Type Analysis

This feature is a categorical variable that states the code of the room reserved. There are no missing values for this variable. This variable is similar to the previous one because it describes which rooms were assigned from the requested ones. Therefore, we will make a comparison between the assigned rooms and the requested ones. First, we need to understand that the assigned room types will also not have any correlation with our main variable which is_cancelled because the same one before it had no correlation.

To start with, we will draw a scatter plot between the assigned rooms and the requested ones to see the relationship between them.




Clearly there is no relationship between the room that the person reserves and the room that is assigned which means that it's of least probability that the person gets the room he/she wants.

## 19. Bookings Changes Analysis

This feature is a numerical variable that states the number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation. There are no missing values for this variable. The statistical summary for this feature is observed below.

```
: 0      101314
  1       12701
  2        3805
  3         927
  4         376
  5         118
  6          63
  7          31
  8          17
  9           8
  10          6
  13          5
  14          5
  15          3
  16          2
  17          2
  12          2
  11          2
  20          1
  21          1
  18          1
  Name: booking_changes, dtype: int64
```


It is by far that the most booking changes are 0 because it makes sense that most people do not change their bookings a lot. The highest number of booking changes are 21 which shows that it is definitely an entity or a company that has meetings or etc.

Furthermore, to test if there is a relationship between the booking changes and the cancellations we can compute the correlation coefficient to test our hypothesis.

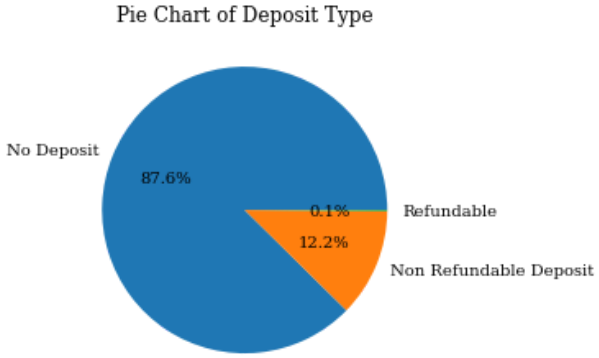
```
-0.14438099106132385
```

The correlation coefficient shows a weak negative correlation between both variables. The negative or weak correlation makes sense because if a person changes the booking, it has nothing to do with whether he/she cancels this current booking or not.

## 20. Deposit Type Analysis

This feature is a categorical variable that indicates if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay. We can find below displayed the values.

```
: No Deposit    104641
  Non Refund     14587
  Refundable       162
  Name: deposit_type, dtype: int64
```




Above is a pie chart showing the percentages of deposit types. It is clear that by far nearly everyone books without leaving a deposit because no deposit is 87.8% of the total. This could be for several reasons, maybe because most people tend not to leave a deposit because they want to change their mind, but the 12% of the people who paid a non refundable deposit are mostly people who want to make sure that the room is theirs. Also, the non refundable deposits could be in certain seasons such as holidays because there is always a very high demand on hotels then.

Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

```
Observed Contingency Table:
is_canceled          0      1
deposit_type
No Deposit       74947  29694
Non Refund          93  14494
Refundable         126     36


Chi-squared Test Statistics:
Chi2: 27677.32924132434
P-value: 0.0
Degrees of Freedom: 2

Expected Frequencies:
[[6.58802698e+04 9.18373768e+03 1.01992562e+02]
 [3.87607302e+04 5.40326232e+03 6.00074378e+01]]
```

Since the p-value is 0 which is less than the significance level, we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and deposit type variables.

## 21. Agent Analysis

This feature is a categorical variable that states the ID of the travel agency that made the booking. There are a lot of missing values for this variable. The most frequent occurrence for this feature is observed below.

```
9.0      31961
240.0    13922
1.0       7191
14.0      3640
7.0       3539
         ...
289.0        1
432.0        1
265.0        1
93.0         1
304.0        1
Name: agent, Length: 333, dtype: int64
```

It is clear that the agent with the highest bookings made is agent number 9 with 31,961. Yet since a large section of the feature is missing, it will be dropped.

## 22. Company Analysis

This feature is a categorical variable that states the ID of the company/entity that made the booking or responsible for paying the booking. Almost 94% of the observations of this feature is missing and hence the most appropriate course of action is to drop this feature.

## 23.Days in Waiting List Analysis

This feature is a numerical variable that states the number of days the booking was in the waiting list before it was confirmed to the customer. There are no missing values for this variable. The statistical summary for this feature is observed below.

```
count    119390.000000
mean          2.321149
std          17.594721
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max         391.000000
Name: days_in_waiting_list, dtype: float64
```

The summary shows that the maximum number of days the booking was in the waiting list is 391 days which shows that definitely this can be an outlier because it is way larger than the rest of the values.

```
0      115692
39        227
58        164
44        141
31        127
         ...
116         1
109         1
37          1
89          1
36          1
Name: days_in_waiting_list, Length: 128, dtype: int6
```

The statistics above clearly show that the most common days in the waiting list are 0 days. This shows that most of the hotels answer upon the requests very fast.

Furthermore, to test if there is a relationship between the days in the waiting list and the cancellations we can compute the correlation coefficient to test our hypothesis.

```
0.05418582411778037 6
```

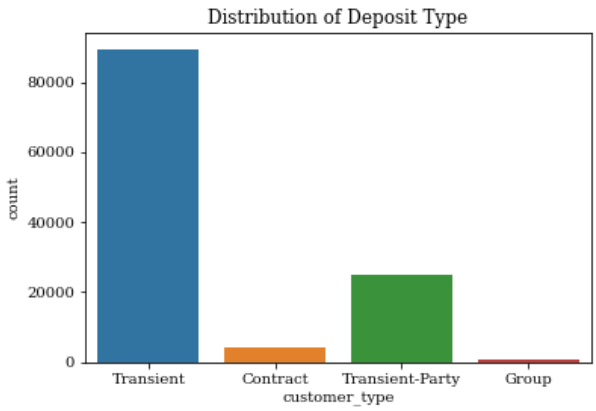
The correlation coefficient shows a weak positive correlation between both variables. The positive or weak correlation makes sense because the waiting list

days does not mean that the person will make the person cancel the current booking.

## 24.Customer Type Analysis

This feature is a categorical variable that states the type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking.

```
Transient          89613
Transient-Party    25124
Contract            4076
Group                577
Name: customer_type, dtype: int64
```




The statistics above clearly show that the most common customer type are the transient customers with 89613. This shows that most of the customers are not part of a group or contract and they are individuals.

Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

```
Observed Contingency Table:
is_canceled              0      1
customer_type
Contract              2814   1262
Group                  518     59
Transient            53099  36514
Transient-Party      18735   6389


Chi-squared Test Statistics:
Chi2: 647.8350973363271
P-value: 1.121956218424043e-133
Degrees of Freedom: 9

Expected Frequencies:
[[5.41404222e+04 3.18535778e+04]
 [7.03874596e+02 4.14125404e+02]
 [5.86772024e+02 3.45227976e+02]
 [1.20886370e+04 7.11236305e+03]
 [4.11432959e+03 2.42067041e+03]
 [1.82390403e+03 1.07309597e+03]
 [1.31834830e+03 7.75651696e+02]
 [3.78379814e+02 2.22620186e+02]
 [3.77750230e+00 2.22249770e+00]
 [7.55500461e+00 4.44499539e+00]]
```

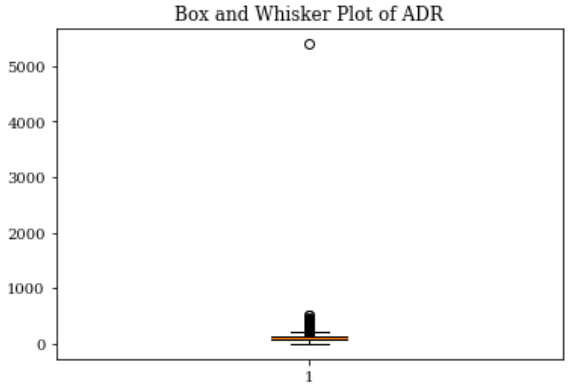
Since the p-value is 1.121e-133 which is less than the significance level (0.05), we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and 'customer_type_analysis' variables.

## 25.ADR Analysis

This feature is a numerical variable that states the Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights. The statistical summary for this feature is observed below.

```
count    119390.000000
mean        101.831122
std          50.535790
min          -6.380000
25%          69.290000
50%          94.575000
75%         126.000000
max        5400.000000
Name: adr, dtype: float64
```

The output above shows a lot of controversial information. This is because the output shows that the minimum values for the average daily rate is -6.38 and the maximum is 5,400. It could happen that the rate is negative, but it is not logical that we have a rate in thousands and the average rate as calculated above is 101.8. Therefore, the 5400 rate is definitely an outlier in the data because it does not make any sense.




The box plot has an irregular shape as it seems to be compressed and a single point seems to be an extreme outlier which is the 5400 data point that was mentioned above. As we can see the outlier has disrupted the data and is making us unable to see the true distribution of the data. This problem will be dealt with in the data cleaning and preprocessing section of the report.
Furthermore, to test if there is a relationship between the ADR and the cancellations we can compute the correlation coefficient to test our hypothesis.

```
0.047556597880385 8
```

The correlation coefficient calculated shows a very weak relationship between the variables which shows that there is no linear relationship between the average daily rate and our main variable (is_canceled).

## 26. Required Car Parking Spaces Analysis

This feature is a numerical variable that states the Number of car parking spaces required by the customer during his stay at each hotel.

```
count    119390.000000
mean          0.062518
std           0.245291
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           8.000000
Name: required_car_parking_spaces, dtype: float64
```

The summary shows that the minimum number of requested parking spaces is 0 and the maximum number of requested parking spaces is 8 reserved spaces.
Furthermore, to test if there is a relationship between the required car parking spaces and the cancellations we can compute the correlation coefficient to test our hypothesis.

```
-0.19549781749450643
```

The correlation coefficient calculated shows a very weak relationship between the variables which shows that there is no linear relationship between the required number of parking spaces and our main variable (is_canceled).

## 27. Total Special Requests Analysis

This feature is a numerical variable that states the number of special requests made by the customer. The statistical summary for this feature is observed below.

```
count    119390.000000
mean          0.571363
std           0.792798
min           0.000000
25%           0.000000
50%           0.000000
75%           1.000000
max           5.000000
Name: total_of_special_requests, dtype: float64
```

The summary shows that the minimum number of special requests is 0 and the maximum number of special requests is 5 special requests.
Furthermore, to test if there is a relationship between the total special requests and the cancellations we can compute the correlation coefficient to test our hypothesis.

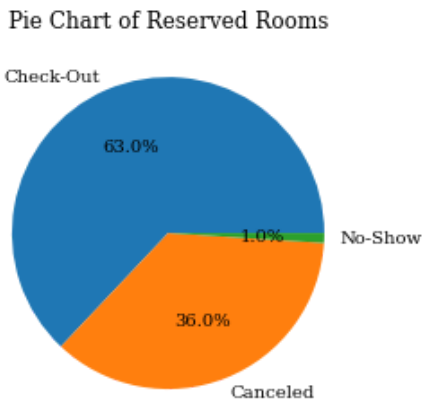
```
-0.2346577739690237
```

The correlation coefficient calculated shows a very weak relationship between the variables which shows that there is no linear relationship between the total special requests and our main variable (is_canceled).

## 28.Reservation Status Analysis

This feature is a categorical feature that describes the reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why.

```
Check-Out    75166
Canceled     43017
No-Show       1207
Name: reservation_status, dtype: int64
```




It is clear that the checkout percentage is the highest among the status and the no show is the least of all status because it is logical that most people show up for their reservations.

Furthermore, a chi-squared test was conducted to determine whether there is a significant association between two variables. The hypotheses were the following; H0: there is no correlation and H1: a correlation exists between the variables.

```
Observed Contingency Table:
is_canceled          0      1
customer_type
Contract          2814   1262
Group              518     59
Transient        53099  36514
Transient-Party  18735   6389


Chi-squared Test Statistics:
Chi2: 2222.50416048372
P-value: 0.0
Degrees of Freedom: 3

Expected Frequencies:
[[ 2566.18323143  1509.81676857]
 [  363.26980484   213.73019516]
 [56418.88565206 33194.11434794]
 [15817.66131167  9306.33868833]]
```

Since the p-value is 0 which is less than the significance level (0.05), we reject the null hypothesis and hence there is evidence to suggest that there is a correlation between the 'is_canceled' and 'reservation status' variables.

## 29. Reservation Status Date Analysis

This feature is a numerical variable which describes the date at which the last status of the booking was set. The statistical summary for this feature is observed below.

```
10/21/2015    1461
7/6/2015       805
11/25/2016     790
1/1/2015       763
1/18/2016      625
              ...
2/27/2015        1
4/25/2015        1
3/11/2015        1
6/14/2015        1
2/12/2015        1
Name: reservation_status_date, Length: 926, dtype: int64
```

As we can see above, the most repeated reservation status check date is 21/10/2015. We cannot say that there is a certain day that has the least reservation status check date because there are a lot of days that has only 1 check.

Furthermore, we cannot test if there is a relationship between the booking changes and the reservation status date.

## Data Cleaning & Preprocessing & Feature Engineering

After the data analysis performed in the first part of this report, it was clear that the data required extensive cleaning to prepare for the models to be implemented in later phases. This section will discuss every method implemented to clean and preprocess the data as well as any feature engineering to be performed.

### 1. Checking for NA's

For starters the data exhibited quite a few missing values during the analysis phase, so the first step was to find ways to fix the missing values. The four features that presented missing values were children (4), country (488), agent (16430), and company (112593).

```
children                         4
babies                           0
meal                             0
country                        488
market_segment                   0
distribution_channel             0
is_repeated_guest                0
previous_cancellations           0
previous_bookings_not_canceled   0
reserved_room_type               0
assigned_room_type               0
booking_changes                  0
deposit_type                     0
agent                        16340
company                     112593
```

Since the number of missing values in children were very few, these respective observations were dropped. Since the dataset is set in the Portuguese hotel industry, the most appropriate course of action to fill the missing NA values in the country feature would be fill in the missing values with the mode of the feature. In this case, the mode is the country Portugal ("PERT") and hence the missing values will be filled with it. Finally, the features agent and company present a great amount of missing values and hence they will be dropped.

### 2. Hotel

The feature hotel was one-hot encoded, which is a process of converting categorical variables into a binary matrix where each category or label is represented by a unique column. The feature presented to categorical variables. "City Hotel" and "Resort Hotel" which will be encoded to 1 and 0 respectively.

### 3. Lead Time

The numerical variable lead time was normalized using Min and Max scaling into a new variable "lead_time_normalized" and the original column was dropped.

### 4. New Feature Arrival Date

The categorical variable arrival date month was encoded such that each month is labeled to its corresponding number; for example January was encoded to 1, February to 2, etc. In addition, the variables arrival day of the month and week number presented a very weak correlation with the label and hence they will be dropped. Finally, to ease the readability of the data, the variables arrival year, month, and day of month will be combined into one variable "arrival_date" and the original columns will be dropped.

### 5. Weekends & Weekdays

Both of these variables will be combined into a new variable "total_stays" and the original columns will be dropped.

### 6. Adults

Since it is impossible to have a hotel reservation without an adult present, any observation with 0 adults was dropped.

### 7. New Feature Kids

The entirety of the hotel industry does not differentiate between babies and children and since the data source doesn't specify age ranges for both, the two variables will be combined into one new feature called "kids" and the original columns were dropped.

### 8. Meal Type

According to the data source, the meal type SC and Undefined are the same and hence any

observation will show the value “Undefined” was changed to “SC”. Moreover, the feature was one-hot encoded to form four new binary columns; BB, FB, HB, and SC and the original column was dropped.

## 9. Country

Since the dataset is based in portugal and the feature country presents 177 unique values making it nearly impossible to assess, the country feature will be generalized to two categories; Portugal and International (representing all other nationalities). Then it will be one-hot encoded and the original column will be dropped.

## 10. Market Segment

The feature market segment was one-hot encoded to form 7 new binary columns; Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Aviation and the original column will be dropped.

## 11. Distribution Channel

The feature distribution channel was one-hot encoded to form 5 new binary columns; Dist Direct, Dist Corporate, Dist TA/TO, SC, GDS and the original column will be dropped.

## 12. Reserved & Assigned Room Type

Both the reserved room type and the assigned room type variables are categorical variables with respectively 10 and 12 categories. The assigned room type variable was encoded using the following mapping: {'C': 0, 'A': 1, 'D': 2, 'E': 3, 'G': 4, 'F': 5, 'I': 6, 'B': 7, 'H': 8, 'P': 9, 'L': 10, 'K': 11}. Similarly, the reserved room type variable was encoded with the following mapping: {'C': 0, 'A': 1, 'D': 2, 'E': 3, 'G': 4, 'F': 5, 'H': 6, 'L': 7, 'P': 8, 'B': 9}.

## 13. Deposit Type

The feature deposit type was one-hot encoded to form 3 new binary columns; No Deposit, Refundable, Non Refund and the original column was dropped.

## 14. Customer Type

The feature customer type was one-hot encoded to form 4 new binary columns; Transient, Contract, Transient Party, Group and the original column was dropped.

## 15. ADR

During the analysis, it became clear that there in a single outlier with an adr value of 5400 which is assumed to be a typo and hence it was corrected to 540.

## 16. Reservation Status

The feature deposit type was one-hot encoded such that the values “Canceled” and “No show” were combined under one value “0” since both indicate that the reservation was canceled and the value “Check-out” was encoded to “1”. Yet, this created an exact copy of the label and hence this variable is redundant and will be dropped.

## 17. Reservation Date

The feature reservation date was converted into a format suitable for the application of machine learning models.

## Data Post Cleaning

The data post cleaning contains 43 variables and 118983 observations as can be seen below.

```
df.columns
```

```
Index(['hotel', 'is_canceled', 'adults', 'is_repeated_guest',
       'previous_cancellations', 'previous_bookings_not_canceled',
       'reserved_room_type', 'assigned_room_type', 'booking_changes',
       'days_in_waiting_list', 'adr', 'required_car_parking_spaces',
       'total_of_special_requests', 'reservation_status_date',
       'lead_time_normalized', 'arrival_date', 'total_stays', 'kids', 'BB',
       'FB', 'HB', 'SC', 'Portugal', 'International', 'Direct', 'Corporate',
       'Online TA', 'Offline TA/TO', 'Complementary', 'Groups', 'Aviation',
       'Dist Direct', 'Dist Corporate', 'Dist TA/TO', 'GDS', 'No Deposit',
       'Refundable', 'Non Refund', 'Transient', 'Contract', 'Transient-Party',
       'Group', 'total_guests'],
      dtype='object')
```

```
num_columns = df.shape[1]
num_columns
```

```
43
```

```
num_observations = len(df)
num_observations
```

```
118983
```