10th July 21



Hotel Cancellation Analysis



Project Notes 1 – Eda and business insights

Sivaramakrishanan s – july 2020 E

Table of Contents

[1) Introduction of the business problem 0](#_Toc76782022)

[Define The Problem Statement 0](#_Toc76782023)

[Need of the Study/Project 0](#_Toc76782024)

[Understanding Business/Social opportunity 1](#_Toc76782025)

[2)Data Report 1](#_Toc76782026)

[Understanding how data was collected 1](#_Toc76782027)

[Visual inspection of data (rows, columns, descriptive details) 1](#_Toc76782028)

[Descriptive Statistics of Numerical Variables 0](#_Toc76782029)

[Descriptive Statistics of Categorical Variables 0](#_Toc76782030)

[Understanding of attributes (Variable info, Renaming if required) 1](#_Toc76782031)

[Variable info 1](#_Toc76782032)

[Duplicate Value Check. 1](#_Toc76782033)

[3) Exploratory data analysis 0](#_Toc76782034)

[A) Univariant Analysis 0](#_Toc76782035)

[Target Variable 0](#_Toc76782036)

[Hotel Type 0](#_Toc76782037)

[Customer Type 1](#_Toc76782038)

[Meal Type 1](#_Toc76782039)

[Distribution Channel 0](#_Toc76782040)

[Market Segment 1](#_Toc76782041)

[Country 2](#_Toc76782042)

[Repeated Guest 3](#_Toc76782043)

[Reserved Room Type 3](#_Toc76782044)

[Assigned Room Type 4](#_Toc76782045)

[Deposit Type 4](#_Toc76782046)

[Total Stays – New Variable 5](#_Toc76782047)

[Total Guest – New Variable 6](#_Toc76782048)

[Waiting List Days 0](#_Toc76782049)

[Required Car Parking 0](#_Toc76782050)

[Required Special Request 0](#_Toc76782051)

[Previous Bookings Cancelled 0](#_Toc76782052)

[Lead Time – & Lead Time as Month - New Variable 1](#_Toc76782053)

[Arrival Month 2](#_Toc76782054)

[Arrival Day 2](#_Toc76782055)

[B) Bivariant Analysis 3](#_Toc76782056)

[Hotel Type and Cancellation 3](#_Toc76782057)

[Arrival Year and Cancellation 3](#_Toc76782058)

[Total Guest and Cancellation 4](#_Toc76782059)

[Meal Type and Cancellation 4](#_Toc76782060)

[Location and Cancellation 5](#_Toc76782061)

[Market Segment and Cancellation 5](#_Toc76782062)

[Distribution Channel and Cancellation 0](#_Toc76782063)

[Deposit Type and Cancellation 0](#_Toc76782064)

[Previous Cancellation and Cancellation 0](#_Toc76782065)

[Booking Via Company And Cancellation 0](#_Toc76782066)

[Agent Via Company and Cancellation 1](#_Toc76782067)

[Relationship Analysis 0](#_Toc76782068)

[Lead Time and Cancellation 1](#_Toc76782069)

[Total Number of Special request and Cancellation 0](#_Toc76782070)

[Total Number of Parking space request and Cancellation 0](#_Toc76782071)

[C) Removal Of Unwanted Variables. 1](#_Toc76782072)

[D) Missing Values 1](#_Toc76782073)

[E) Outlier Treatment Of Values 2](#_Toc76782074)

[F) Variable Transformations 3](#_Toc76782075)

[G) Addition of New Variables 4](#_Toc76782076)

[4) Business insights from EDA 5](#_Toc76782077)

[A) Is The Data Balanced? 5](#_Toc76782078)

[Identify the Metrics 5](#_Toc76782079)

[B) Any business insights using clustering  (if applicable) 5](#_Toc76782080)

[C) Any other business insights 5](#_Toc76782081)

[Business Insights 5](#_Toc76782082)

[Recommendations 0](#_Toc76782083)

Table of Tables

[Table 1 Removal of Unwanted Variables 1](#_Toc76780093)

[Table 2 Missing Values 1](#_Toc76780094)

[Table 3 Variable Transformation 3](#_Toc76780095)

[Table 4 New Variables 3](#_Toc76780096)

[Table 5 Business Insights 4](#_Toc76780097)

[Table 6 Recommendations 0](#_Toc76780098)

Table of Figures

[Figure 1 Cancellation Rate 0](#_Toc76782092)

[Figure 2 Rows And Columns 1](#_Toc76782093)

[Figure 3 Descriptive Statistics Numerical 0](#_Toc76782094)

[Figure 4 Descriptive Statistics Categorical 0](#_Toc76782095)

[Figure 5 Data Info 1](#_Toc76782096)

[Figure 6 Target Distribution 0](#_Toc76782097)

[Figure 7 Hotel Type 0](#_Toc76782098)

[Figure 8 Customer Type 1](#_Toc76782099)

[Figure 9 Meal Type 1](#_Toc76782100)

[Figure 10 Distribution Channel 0](#_Toc76782101)

[Figure 11 Market Segment 1](#_Toc76782102)

[Figure 12 Country 2](#_Toc76782103)

[Figure 13 Repeated Guest 3](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782104)

[Figure 14 Reversed Room Type 3](#_Toc76782105)

[Figure 15 Assigned Room Type 4](#_Toc76782106)

[Figure 16 Deposit Type 4](#_Toc76782107)

[Figure 17 Total Stays 5](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782108)

[Figure 18 Total Guest 6](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782109)

[Figure 19 Booking Changes 0](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782110)

[Figure 20 Booking Options Changes(NEW) 0](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782111)

[Figure 21 Lead Time 1](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782112)

[Figure 22 Arrival Month 2](#_Toc76782113)

[Figure 23 Arrival Week Day Name 2](#_Toc76782114)

[Figure 24 Hotel Type VS Cancellation 3](#_Toc76782115)

[Figure 25 Arrival Date VS Cancellation 3](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782116)

[Figure 26 Total Guest And Cancellation 4](#_Toc76782117)

[Figure 27 Meal Type And Cancellation 4](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782118)

[Figure 28 Location And Cancellation 5](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782119)

[Figure 29 Market Segment and Cancellation 5](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782120)

[Figure 30 Distribution Channel and Cancellation 0](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782121)

[Figure 31 Deposit Type and Cancellation 0](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782122)

[Figure 32 Previous Cancellation and Cancellation 0](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782123)

[Figure 33 Booking Via Company And Cancellation 0](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782124)

[Figure 34 Agent Via Company and Cancellation 1](file:////Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/Document/Sivaramakrishnan_S_Capstone_Project_Project_Notes_1.docx#_Toc76782125)

[Figure 36 Heat Map of Variables 0](#_Toc76782126)

[Figure 37 Lead Time and Cancellation 1](#_Toc76782127)

[Figure 38 Lead Time and Cancellation 0](#_Toc76782128)

[Figure 39 Total Number of Parking space request and Cancellation 0](#_Toc76782129)

[Figure 40 Outlier Treatment 2](#_Toc76782130)

**Tableau Public Link for visualizations –**

**https://public.tableau.com/app/profile/sivaramakrishnan3623/viz/Sivaramakrishnan\_Capstone\_Hotel\_Cancelation\_EDA\_1/SpecialRequestVSTarget**

# **1) Introduction of the business problem**

## Define The Problem Statement

The hotel industry has been transformed with a majority of bookings now made through Online Travel Agencies (OTA). These OTAs have transformed the cancellation policies from a footnote at the bottom of the page to the main selling point in their marketing campaigns [(source)](https://triptease.com/blog/the-real-cost-of-free-cancellations/). This results, the customers have become accustomed to free cancellation policies. Based on the [Fornova](https://www.fornova.com/) research conducted on Dec 201 on the 200K hotels for the cancellation rate across the industry the free cancellation policy hit 38% and 62% of no-refund policy on hotels where before the pandemic situation(COVID-19). The same survey ran on July 2020 the results are dramatically different as 58% of hotels now offering the free cancellation and 42% hotel are still refusing to offer the refund [(source)](https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/)

Below is the D-Edge Hospitality shows that the cancellation rate over 5 years shows that average change as 7.1 and average cancellation rate decreased from 41.3% to 39.6% [(source)](https://hospitalitytech.com/global-cancellation-rate-hotel-reservations-reaches-40-average)

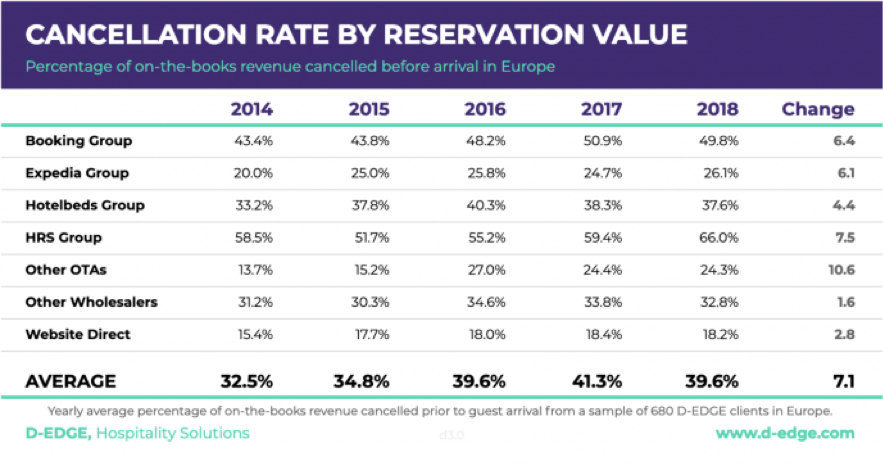


Figure 1 Cancellation Rate

## Need of the Study/Project

When hotels try to protect themselves by using services from OTA's "Risk Free Reservations", the burden then falls on OTAs. Indeed, this service requires the OTA to pay for the reservation if the booking is cancelled and they cannot find a new guest to occupy the room [(source)](https://triptease.com/blog/the-real-cost-of-free-cancellations/). One thing is clear, whether you are a hotel or an OTA, cancellations have an negative financial impact on your business.

In addition to the direct financial consequences of cancellations, they also cause operational problems (such as over or understaffing). Those problems may lead to decrease customer satisfaction and negative reviews. In a world where more and more customers check online reviews before picking a hotel, those reviews can have major impacts. Indeed, TripAdvisor’s reviews and scores influenced around $546 billions of travel spending during 2017 [(source)](https://www.stayntouch.com/blog/how-online-reviews-impact-hotel-revenue/). At a single hotel level, an increase in online reputation score has been linked to an increase in occupancy and revenue [(source)](https://vtechworks.lib.vt.edu/handle/10919/85353).

## Understanding Business/Social opportunity

Artificial intelligence is playing an increasingly important role in hospitality management [(source)](https://www.revfine.com/hospitality-management/), primarily because of its ability to carry out traditionally human functions at any time of the day. This potentially means that hotel owners can save significant money, eliminate human error and deliver superior service.

The AI is being utilised within the hotel industry away from pure customer service is in data analysis. In this capacity, the technology can be used to quickly sort through large amounts of data and draw important conclusions about customers, or potential customers.

An example of this has been seen with the Dorchester Collection hotel chain, which has made use of the Metis AI platform. By using this technology, the company has been able to sort through data collected via surveys, online reviews etc. and the AI has been able to then analyse this to draw conclusions about overall performance.

# **2)Data Report**

## Understanding how data was collected

* In terms of time, frequency and methodology

The dataset has collected with the details of the bookings scheduled to arrive between July, 1st 2017 and August, 31st 2020 from two types of hotels. The two type of hotel datasets were merged into one csv to proceed the analysis.

## Visual inspection of data (rows, columns, descriptive details)

Booking data from both types hotels share the same structure, with 25 variables describing the 40,060 observations of type-1 and 79,330 observations of type-2(total – **119390**). For a detailed list and description of those variables refer to the [data dictionary](https://file+.vscode-resource.vscode-webview.net/Users/s/Desktop/DS/Capstone%20Project/HotelCancellation/References/#Data-Dictionary)

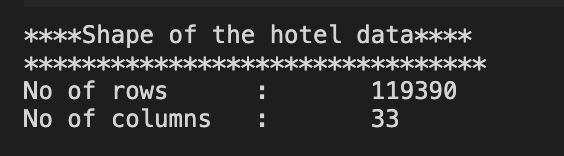


Figure 2 Rows And Columns

### Descriptive Statistics of Numerical Variables

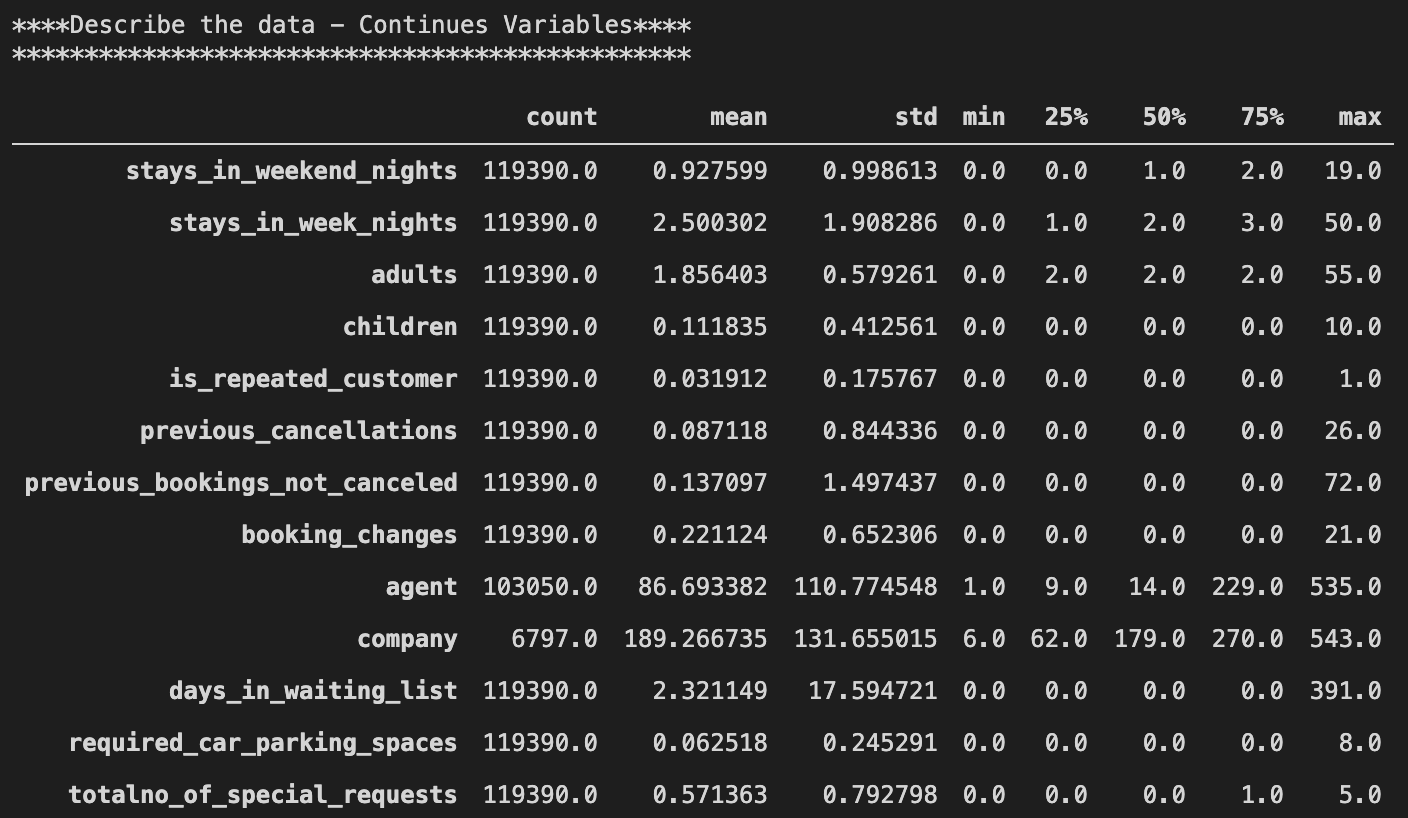


Figure 3 Descriptive Statistics Numerical

Based on the initial view the data provided are looks like a categorical we may need to convert to categorical type post detailed analysis.

### Descriptive Statistics of Categorical Variables

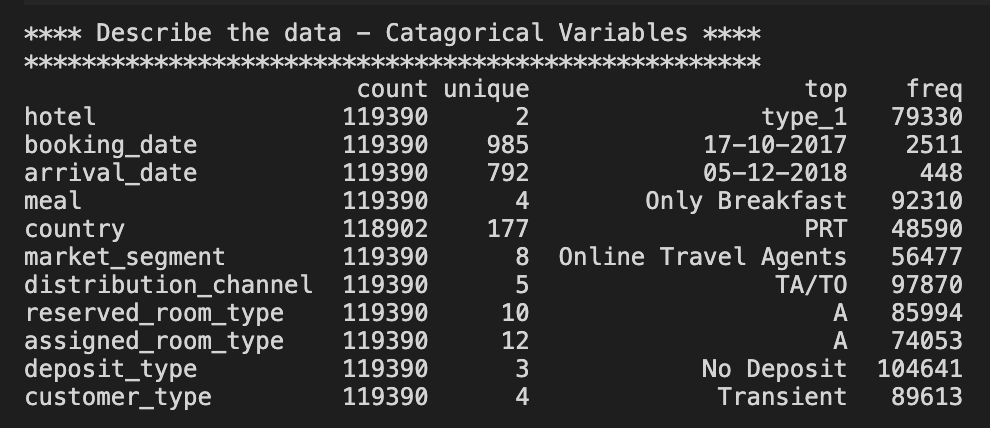


Figure 4 Descriptive Statistics Categorical

There are higher frequency for the No-Deposit which may lead to the highest cancellation since there is no loss for the customer due to the cancellation.

In the categorical variables like Agent or Company, “NULL” is presented as one of the categories. This should not be considered a missing value, but rather as “not applicable”. For example, if a booking “Agent” is defined as “NULL” it means that the booking did not came from a travel agent." As a result, "NULL" values for agent and company will be changed to No Agent and No Company for clarity purposes.

## Understanding of attributes (Variable info, Renaming if required)

### Variable info

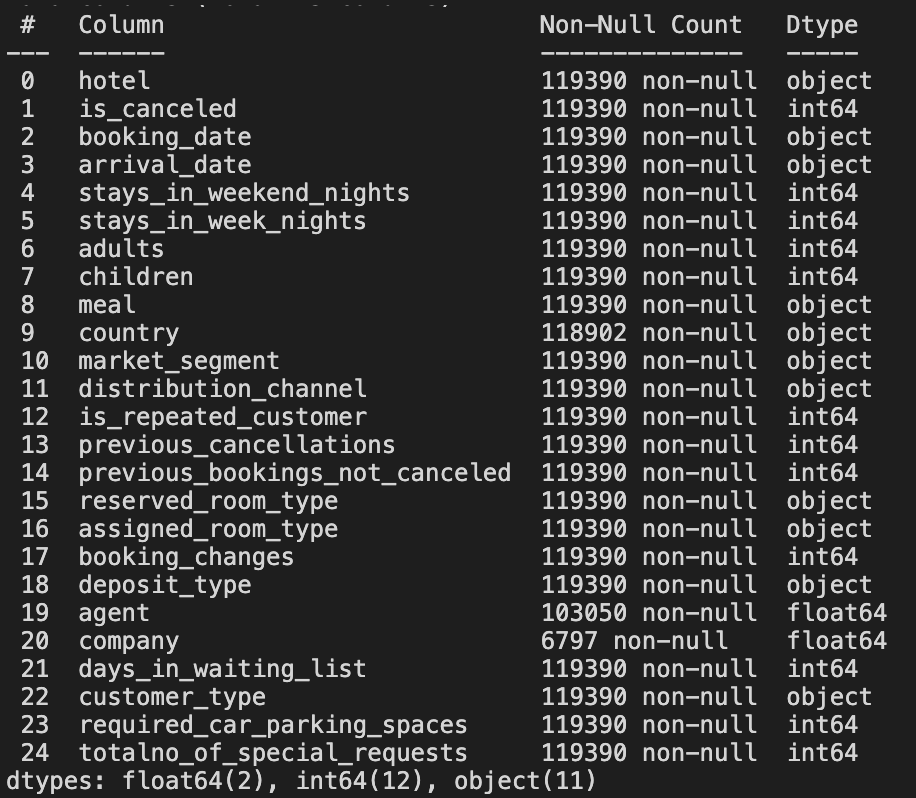


Figure 5 Data Info

* There are three features have missing values.
* Most of the features are in categorical/int64 type of variables.
* There are 2 date type features are considered as object which may need a need to convert to date for further analysis
* Looks like the variable names are more relevant and not required any renaming.

### Duplicate Value Check.

There are 33210 duplicate values in the dataset which is nearly 27.82% of total data this we can drop before the model building.

# **3) Exploratory data analysis**

## A) Univariant Analysis

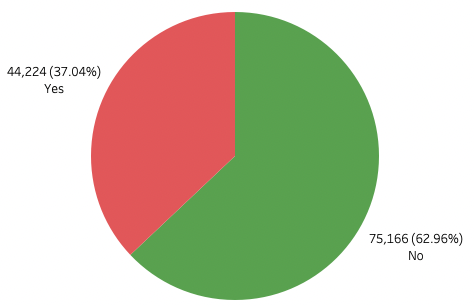
The Univariant analysis describes the distribution and spread for every continuous attribute, distribution of data in categories for categorical ones. The below analysis describes the univariant analysis for the features provided in the Hotel Cancellation data set.

### Target Variable

Let’s begin with the Target variable distribution on the Hotel Cancellation status. This data represented in the feature is\_canceled.

Cancelled : The count of Cancelled distribution is 44224 which is 37.04% of the total data.

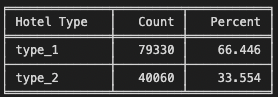
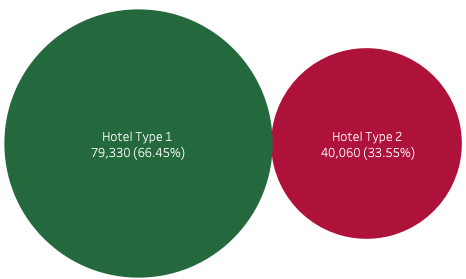
Not Cancelled : The count of Not Cancelled distribution is 75166 which is 62.96% of the total data.



* **The rate of cancellation is likely matching with the industry standard which is around 37% - 40%**
  + **Source: [Emerchantpay Link](https://www.emerchantpay.com/infographic-how-can-hotels-combat-rising-cancellation-rates/" \t "_blank)**
* **The problem that hospitality industries are facing that there are almost 4 cancellation in every 10 bookings**
* **The target data is almost balance, so later on for the machine learning process we won’t need to do an imbalance handing**

Figure 6 Target Distribution

### Hotel Type

The data collected from the two different types of hotels the hotel types are named as “type\_1” and “type\_2”

* There are more booking from the type1 hotel booking compared to the type2 hotel booking in this case we will see it later on how this affect cancellation

Figure 7 Hotel Type

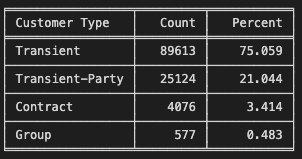
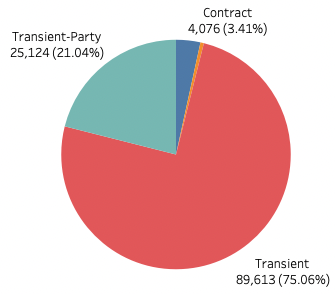
### Customer Type

The Customer Type is nothing but a type of stay by the customer. Below are the different type of customer types.

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 group or contract, and is not associated to other booking

Transient-party  – When the booking is transient, but is associated to at least other transient booking

* The customers in type Transient(Individual booking /Personal not related to company or anything) have majority of the booking and we will see how this customer type affecting the cancellation rate as well

Figure 8 Customer Type

### Meal Type

Type of meal booked. Categories are presented in standard hospitality meal packages are below :

* No meal
* Only Breakfast
* Breakfast-dinner
* Breakfast-lunch-dinner

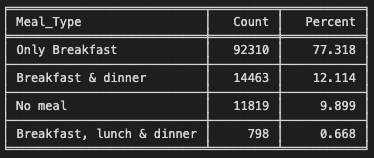
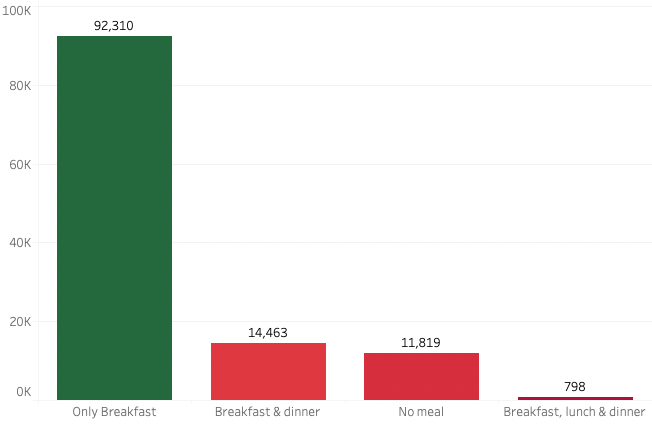
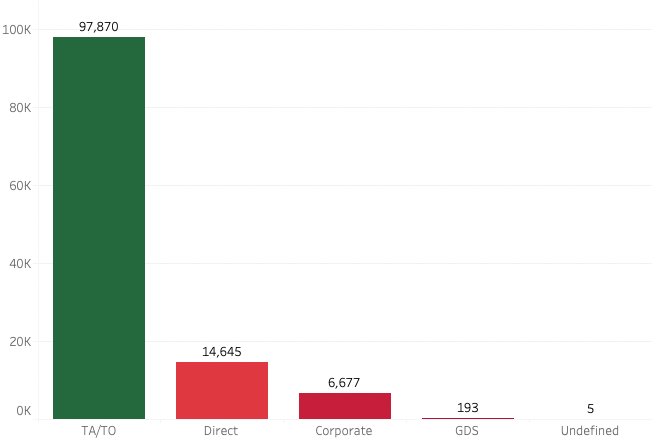
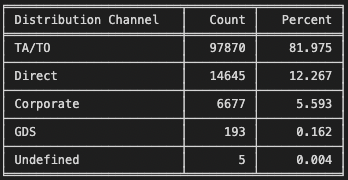


Figure 9 Meal Type

* Only Breakfast is most popular meal package compared to the others while Breakfast, lunch and dinner is the least popular meal package
* Since Only Breakfast is most popular so majority of the customers stays for the day

### Distribution Channel

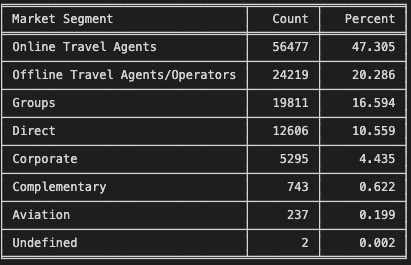
The distribution channel is a plan of an action for selling rooms profitability through varies channels. This would be the mix of Direct channels such as hotel website, in direct channels as

* OTA(Online Travel Agents)/TO(Tour Operators)
* Global Distribution Systems(GDS)
* Wholesalers(Corporate) bookings.

#### Figure 10 Distribution Channel

* The TA/TO (Travel Agents / Tour Operators) are the biggest booking channel compared to other channels
* The Undefined data needs to treated with most occurrence value as TA/TO

### Market Segment

The hotel market segmentation shall help to identify the purpose of the trip: either business or leisure. The price does not decide the market segmentation. A clear distinction must also be achieved between individual and group business.

Market segment categories are

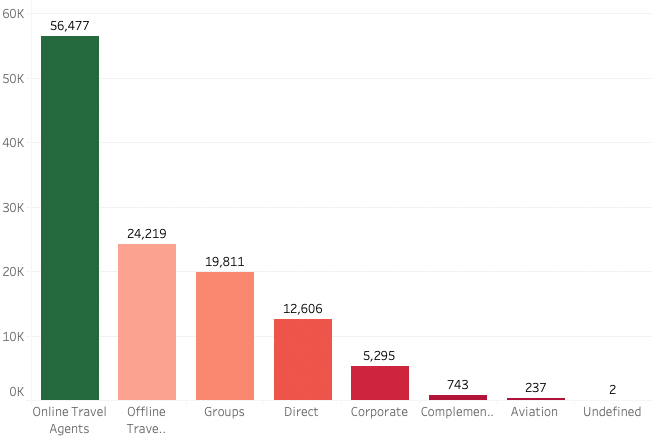
* Direct
* OTA(Online/Offline Travel Agents)/ TO(Tour Operators)
* Group Bookings
* Wholesale (Corporate) Bookings
* Complimentary
* Aviation
* The TA/TO (Travel Agents / Tour Operators) are the biggest booking channel compared to other channels
* The Undefined data needs to treated with most occurrence value as TA/TO

Figure 11 Market Segment

### Country

Country of origin. Categories are represented in the International Standards Organization (ISO) 3155–3:2013 format.

* Almost half of the booking is made from Portugal
* There are many unique values from all the countries where the booking comes from, since their highest booking from Portugal let’s assume both type of the hotels in the Portugal county and we try to group it into continent or we will group it into booking from Portugal and booking from outside Portugal we may assume that both hotel types are in Portugal itself as local booking have a highest count.
* So the new variable we create by splitting the booking into International Booking or Local Booking

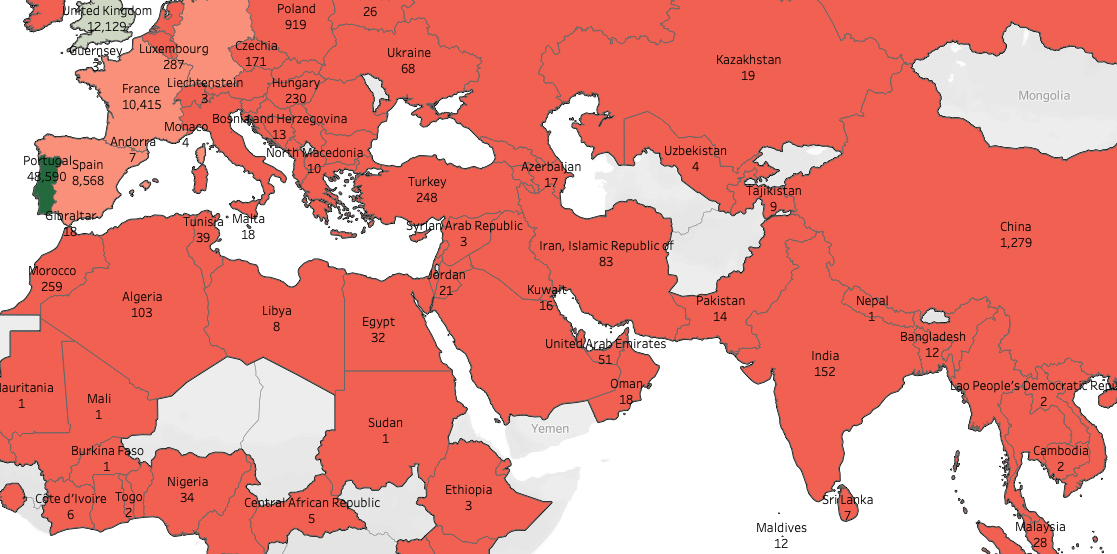
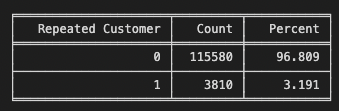


Figure 12 Country

After the continental split below is the distribution of the country.

* International with count of 70800 as 59.3% of total data.
* Local(Portugal) with count 48590 as 40.69% of total data.

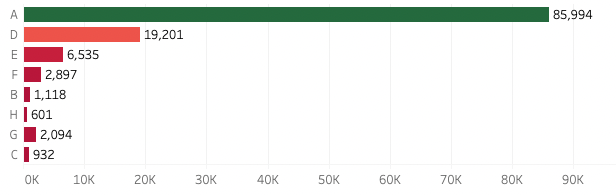
### Repeated Guest

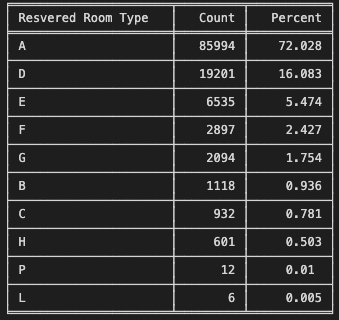
Value indicating if the booking name was from a repeated guest (1) or not (0). Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the Property Management System database it was assumed the booking was from a repeated guest.

* The Repeated customers are only 3.19% so there are very low rate of loyal customers
* The loyal customers are the most profitable then new customers, Below are some reasons
  + They are familiar with hotels offered services
  + Loyal guests usually spend more money at your hotel
    - * + The stay period for loyal guests is usually longer than that of new guests
        + Most probably they will brag to their friends about the hotel that amazing hotel that they keep going to every year - Source: [Sabee Link](https://www.sabeeapp.com/blog/loyal-customers-vs-new-customers" \t "_blank)

Figure 13 Repeated Guest

### Reserved Room Type

Code of room type reserved. Code is presented instead of designation for anonymity reasons.

Figure 14 Reversed Room Type

* The Reversed Room Type A is most popular room type since this has a highest number of booking
* The reason for this could be the cheapest room in the both hotel types, we may do later analysis to validate this assumption

### Assigned Room Type

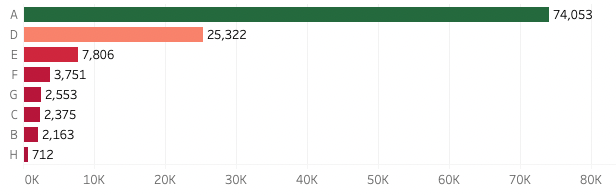
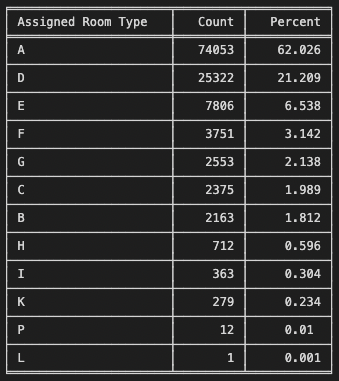
Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.

Figure 15 Assigned Room Type

* There are moderate difference from Reversed Room Type to Assigned Room Type
* The Room types(I & K) are not booked by customers but there are assignments in the dataset, these assignments may be due to the Reversed rooms are assigned to early arrived loyal customers visit on last minute

### Deposit Type

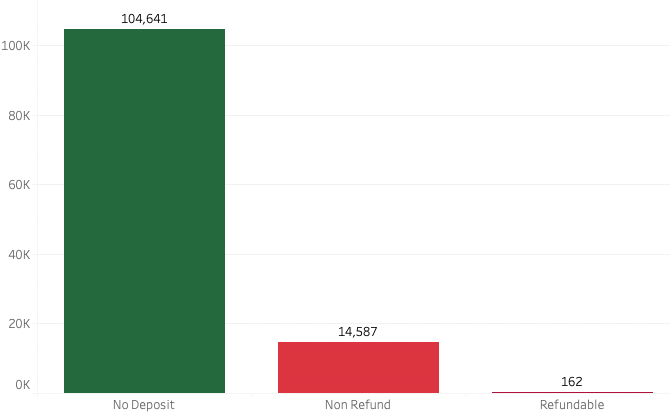
Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:

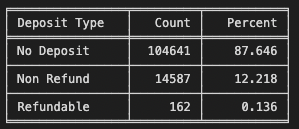
1) No Deposit (no deposit was made)

2) Non Refund (a deposit was made in the value of the total stay cost)

3) Refundable (a deposit was made with a value under the total cost of stay).

Value calculated based on the payments identified for the booking in the transaction (TR) table before the bookings arrival or cancellation date. In case no payments were found the value is “No Deposit”. If the payment was equal or exceeded the total cost of stay, the value is set as “Non Refund”. Otherwise the value is set as “Refundable”.



* The No Deposit type is the highest count this deposit type may impact highly affect the cancellation rate in both type of hotels

#### Figure 16 Deposit Type

### Total Stays – New Variable

The Total stays calculated based on the number of Stays in Weekend Night and number of Stays in Weekday Nights.

**Stays in Weekend Night**

Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel. Calculated by counting the number of weekend nights from the total number of nights.

**Stays in Weekday Nights.**

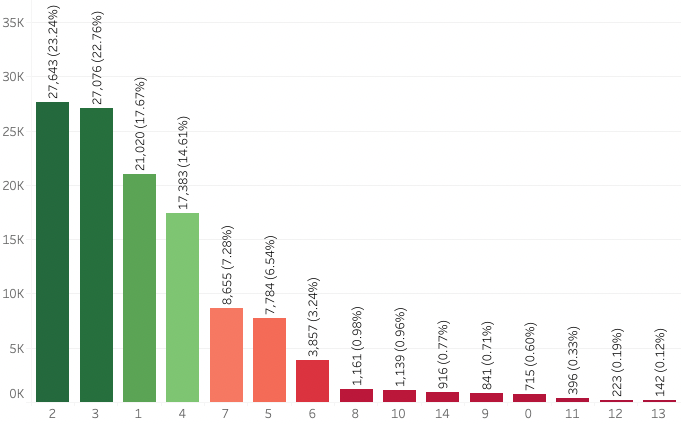
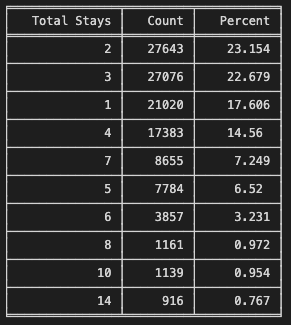
Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel. Calculated by counting the number of week nights from the total number of nights.

Figure 17 Total Stays

* Total Stays of 2,3,1 and 4 days are the most popular total stays
* There are customers to book hotel on the same day and check out (not an overnight stay) so we will not be dropping the rows that has 0 total stays (not an overnight) stay

### Total Guest – New Variable

We are creating the variable as total\_guest by combining the value of adults and children in the booking. The table shows the top 6 results of the count of Total Guest variable

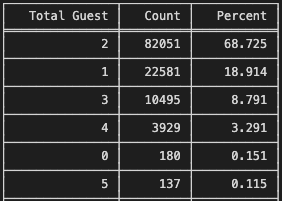
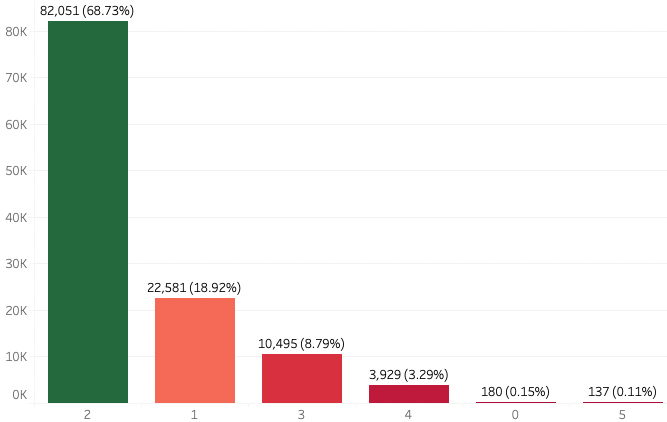


Figure 18 Total Guest

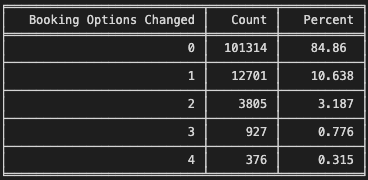
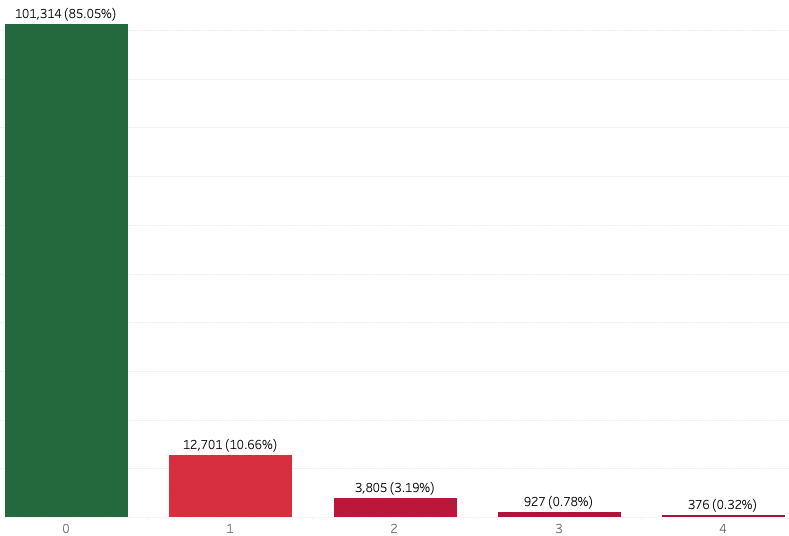
* Total Stays of 2,3,1 and 4 days are the most popular total stays
* There are customers to book hotel on the same day and check out (not an overnight stay) so we will not be dropping the rows that has 0 total stays (not an overnight) stay

**Adults – No of guest with age > 18**

**Children – No of guest with age < 18**

#### Booking Changes

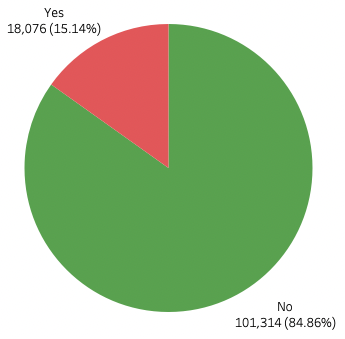
Number of changes/amendments made to the booking from the moment the booking was entered on the Property Management System until the moment of check-in or cancellation. Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal.



* Almost 85% of the customers are not changed their booking

Figure 19 Booking Changes

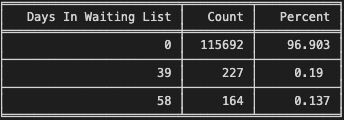
Considering the many different values in the booking changes we could make a two group as Booking Changed and Booking Not Changed



### Waiting List Days

Number of days the booking was in the waiting list before it was confirmed to the customer. Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the Property Management System.

Figure 20 Booking Options Changes(NEW)

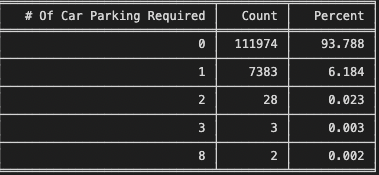


The table has a top 3 of more than 150 days of waiting list.

* Almost 97% customers are got the rooms without any waiting list
* The cause of the waiting list could be the reason the customer booking at the wrong time(Last minute travel) during days in high occupancy (Important Country festivals)
* This can be avoided book the rooms 40 days before Source: [USA Today](https://www.usatoday.com/story/travel/hotels/2014/08/14/hotel-booking-tips/14006883/)

### Required Car Parking

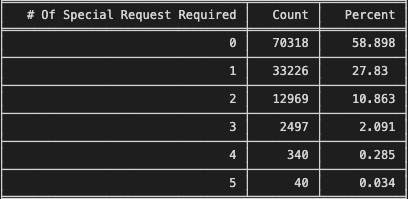
No of Car Parking space requested by customer during the booking of the Room.



* Over 94% customer not requested for the car parking
* There is 6% of customer required 1 car parking
* We need to do further analysis on the effect of the cancellation rate

### Required Special Request

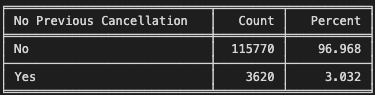
Number of special requests made by the customer (e.g. twin bed or high floor).



* Almost 59% of customers not requested for special request
* Over 28% customers are requested for 1 special request and 11% customers are requested for 2 special request

### Previous Bookings Cancelled

Number of previous bookings that were cancelled by the customer prior to the current booking. In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and cancelled.



#### Almost 95% of the booking never been cancelled before in this data set

#### We will group this into booking that's never been cancelled or have been cancelled before

### Lead Time – & Lead Time as Month - New Variable

Number of days that elapsed between the entering date of the booking into the Property Management System and the arrival date. Calculated by subtracting the entering date from the arrival date.

* Lead Time Days = (Date of Arrival - Date of Booking)
* There are many unique values in the lead time and we can group it as months to get more insights on the trend
* Lead Time Months = ((leadtime days) // 30) (// Returns Rounding off value)

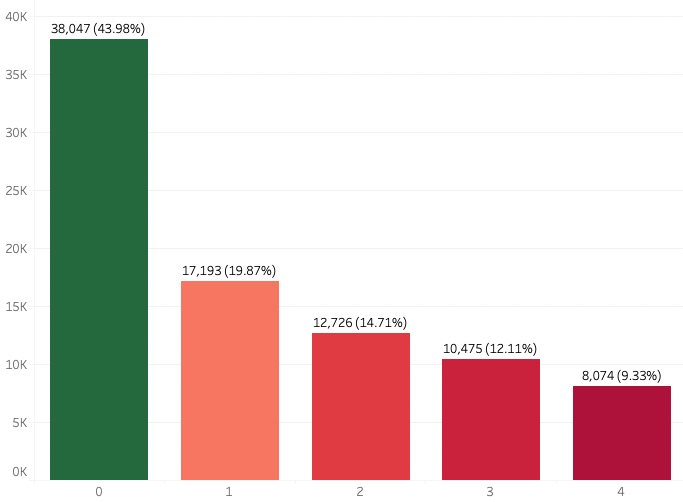
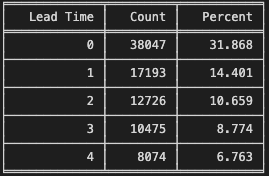
The table displayed the top 5 values of Lead time of Month

Figure 21 Lead Time

* The majority of leadtime booking and arrival on the same month and some bookings made a year or more ahead before the arrival
* We need to check this longer values how affect the cancellation rate

### Arrival Month

Month of arrival date with 12 categories: “January” to “December”.

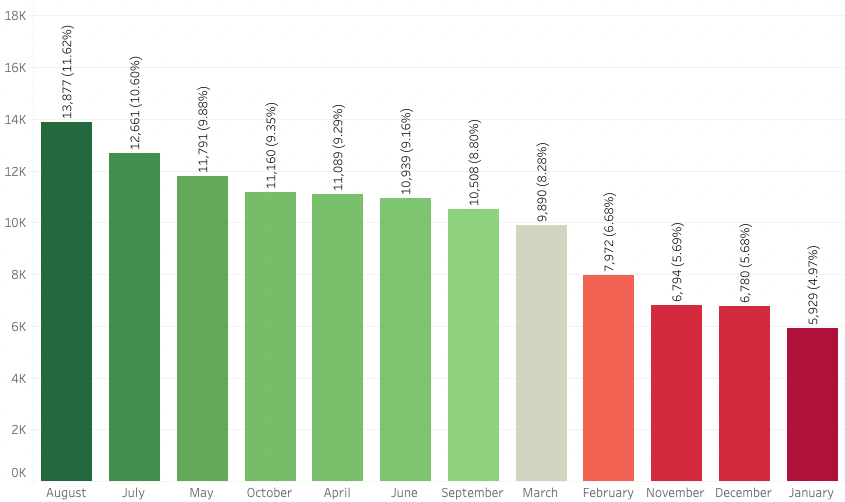


Figure 22 Arrival Month

* July and August (3 times of data) higher count since the data provided from July 2018 to August 2020
* The average shown in the above table is not reliable since there are difference in months over the year

### Arrival Day

Week day of arrival date with 7 categories: “Sunday” to “Saturday”.

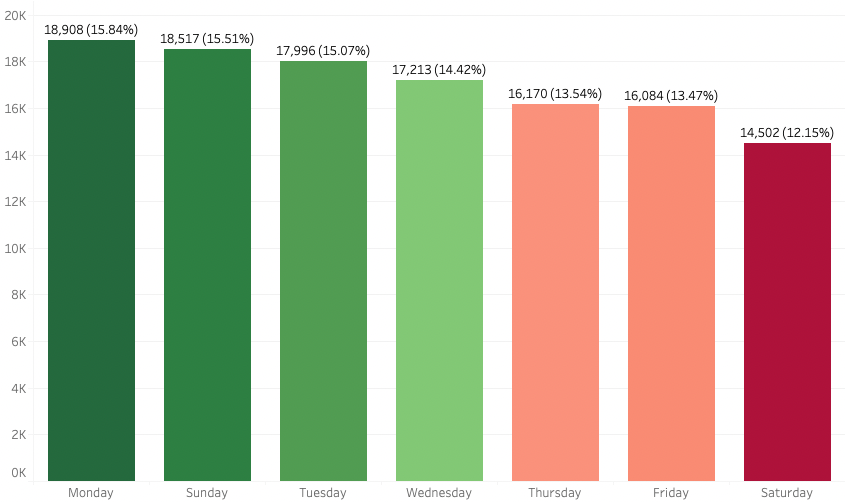


Figure 23 Arrival Week Day Name

* Booking Arrival on Monday and Sunday is higher than other days
* Also average difference in days of arrival is 2% to 3%

## B) Bivariant Analysis

### Hotel Type and Cancellation

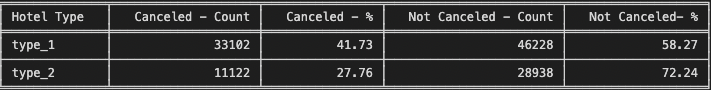
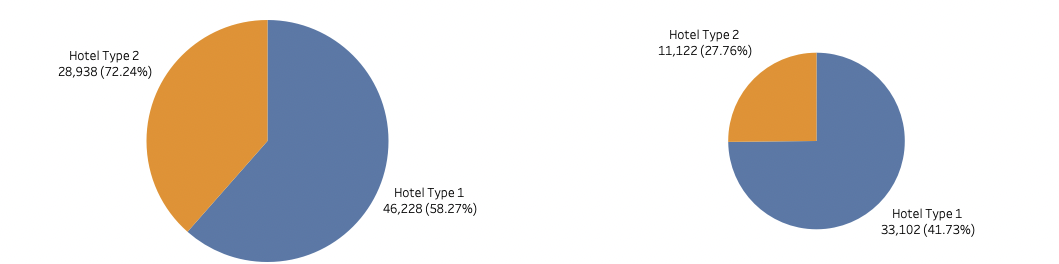
Here the Hotel type variable has been used to find the impact of the Target variable.

Figure 24 Hotel Type VS Cancellation

* The type1 hotels has a cancelling rate then type2. Based on the dataset the type1 more records so this may be cause of this
* One assumption can be made that increase number of booking will increase number of cancellation (Positively Correlated)

#### 

### Arrival Year and Cancellation

Figure 25 Arrival Date VS Cancellation

* The cancellation rate goes down about 1.2% from in 2019 however the cancellation rate goes up 2.84 % in 2020 to 38.7%
* The Years 2018 and 2020 only have the half of the year so this result may not represent the actual cancelation for both 2018 and 2020

### Total Guest and Cancellation

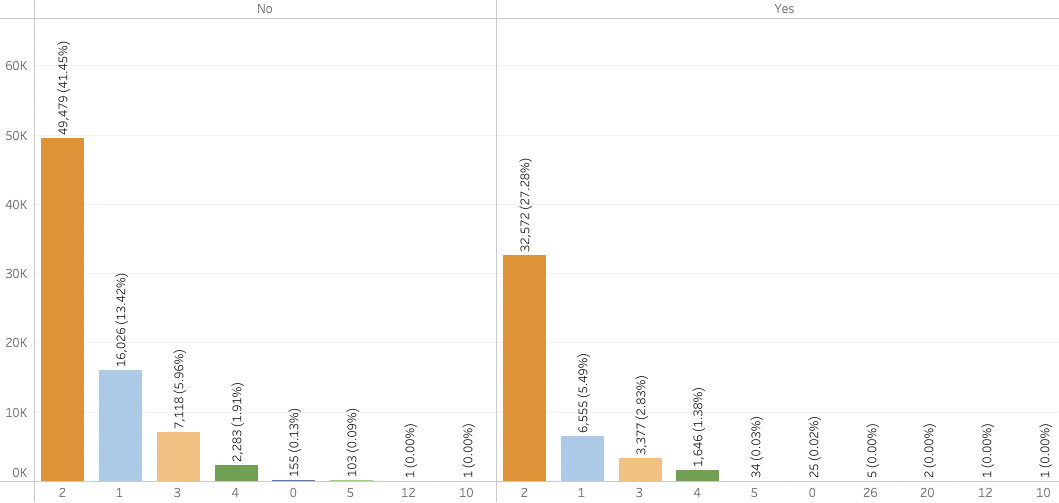
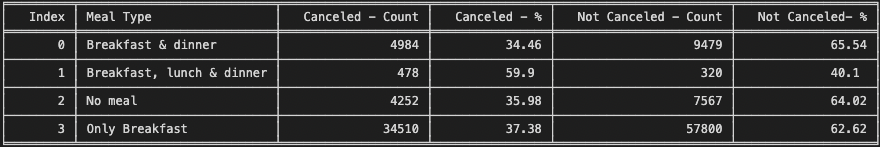


Figure 26 Total Guest And Cancellation

* Since we already reviewed this on the univariant the more of cancellations are with the stay of 2 persons
* The Total Guest with more than or equal to 20 then all bookings are cancelled
* The Total Guest with 50% cancelled for the count 10 and 12



### Meal Type and Cancellation

Figure 27 Meal Type And Cancellation

* Booking with Breakfast, Lunch & Dinner are more likely to cancel compared to booking with other meal package
* While other meal packages has cancellation around 34 - 37%

### Location and Cancellation

Figure 28 Location And Cancellation

* The booking made in Portugal are almost 2.5 X more likely to be cancelled compared to booking that's made outside Portugal

### Market Segment and Cancellation

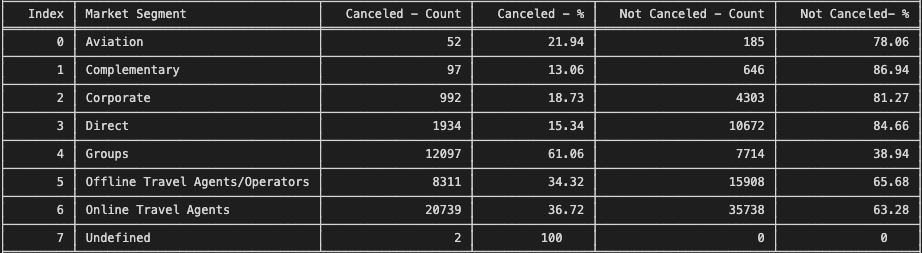
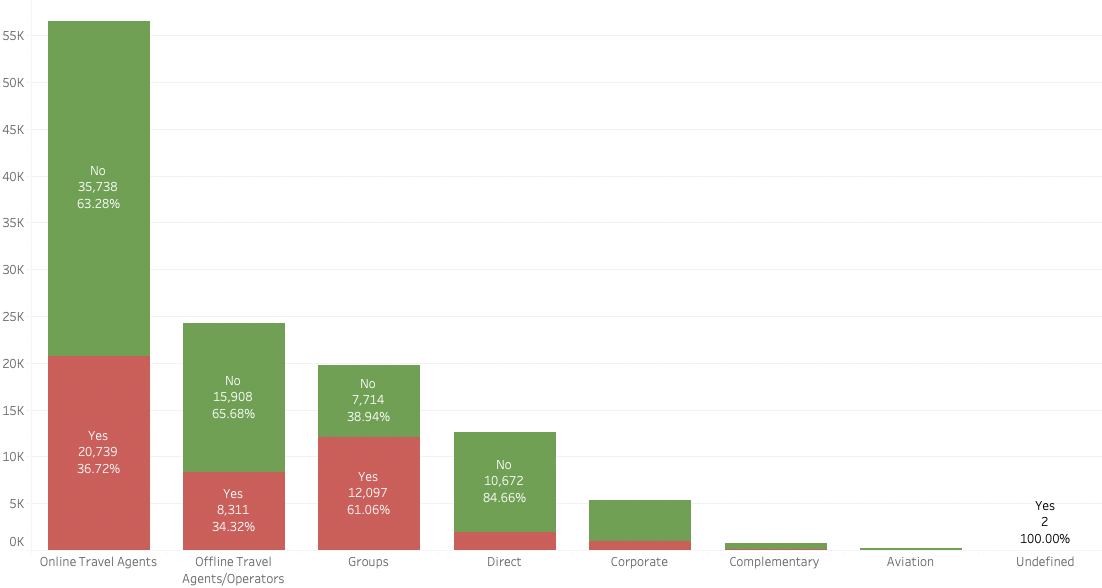


Figure 29 Market Segment and Cancellation

* The Groups market segment has a higher cancel rate(61.06) compared to its confirmed rate
* The Travel agent (online) and Offline has almost a similar number in term of cancellation rate
* The lowest cancellation rate is Direct Booking

### Distribution Channel and Cancellation

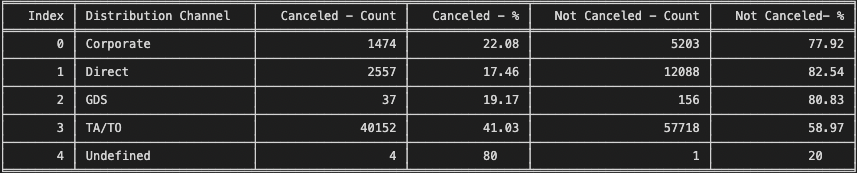
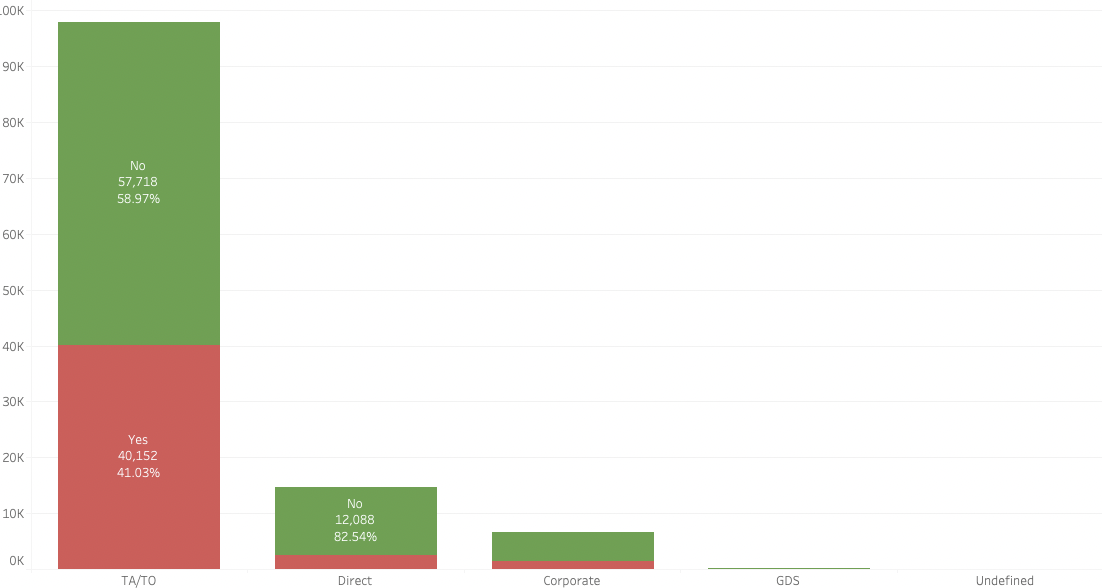
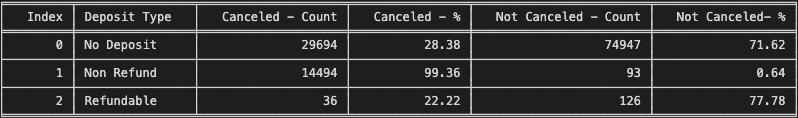


Figure 30 Distribution Channel and Cancellation

* TA/TO (Travel Agents/Tour Operators) have highest cancellation rate
* Direct has a similar to Market Segment has a lower cancellation rate

### Deposit Type and Cancellation

Figure 31 Deposit Type and Cancellation



* The Non Refundable Deposit has the highest cancellation then other Deposit Types
* The hotel has protection from losing out on revenue by implementing the Non Refund Deposit

### Previous Cancellation and Cancellation

Figure 32 Previous Cancellation and Cancellation

* Whoever cancelled the booking before have a high impact to the current cancellation.
* We can we can see the highest of the data is not cancelled and no previous cancellations so we assume there are more loyal customers in this scenario.

### Booking Via Company And Cancellation

Figure 33 Booking Via Company And Cancellation

* Surprisingly there is highest cancellation rate for the booking not through company.

### Agent Via Company and Cancellation

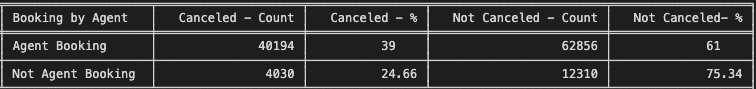


Figure 34 Agent Via Company and Cancellation

* The booking by agent has a highest impact to cancellation rate that the booking done by agents are cancelled.

### Relationship Analysis

##### Correlation Co-efficient between variables

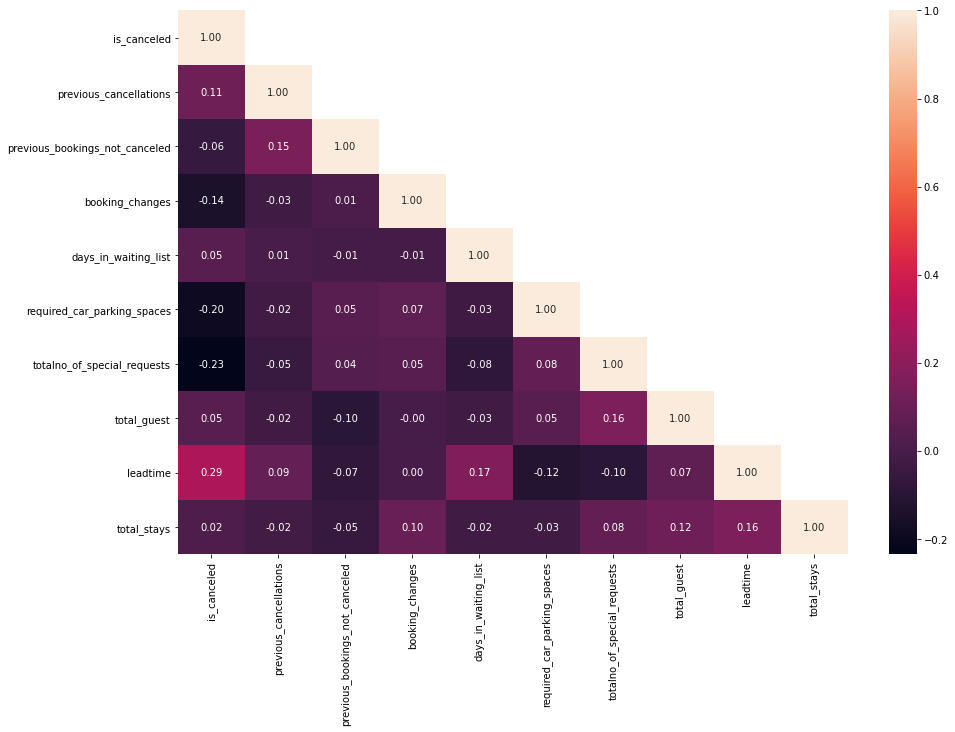
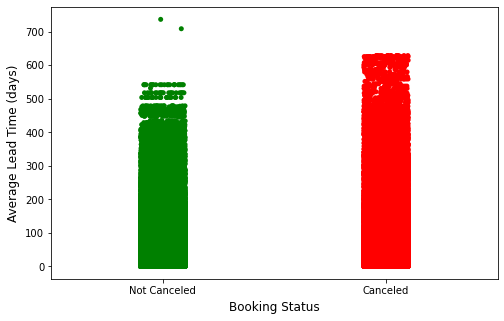


Figure 36 Heat Map of Variables

Relationship on Target Variable. - Top 3

* Lead time is the most highly correlated(0.29) feature with whether or not a booking is\_canceled. It makes sense that as the number of days between when the booking is made and the supposed arrival date increases, customers have more time to cancel the reservation and there is more time for an unforeseen circumstance derailing travel plans to arise.
* The total number of special requests is the second highest feature with the strongest correlation (-0.23)to our is\_canceled target. As the number of special requests made increases, the likelihood that a booking is canceled decreases. This suggests that engagement with the hotel prior to arrival and feeling like their needs are heard may make a customer less likely to cancel their reservation.
* The number of required car parking spaces is the third highest feature with the strongest correlation of (-0.20) to the is\_canceled target. As the number of parking spaces requests increases, the likelihood that a booking is canceled decreases. There is a potential reasons for this relationship are discussed later on.
* Relationship Between Predictors
* There is a modarate correlation(0.17) between days\_in\_waiting\_list and the lead\_time. Since both are related to no of days so can this could be the modarate correlation. We need further investigation on the multicoliniraty and decide on the feature seleaction.
* We also see more features have modorate correlation with lead\_time
  + lead\_time**VS**total\_stays**-**0.16
  + lead\_time**VS**required\_car\_parking**-**0.12
  + lead\_time**VS**required\_car\_parking**-**0.10

### Lead Time and Cancellation



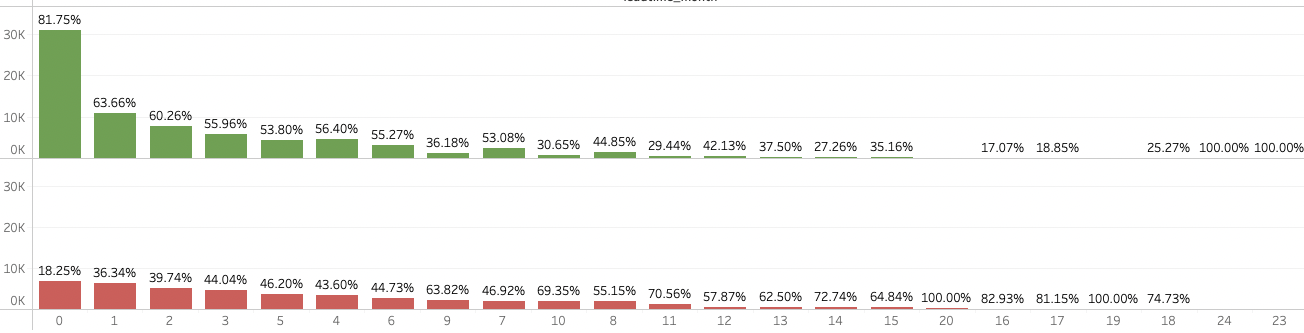


Figure 37 Lead Time and Cancellation

* Cancelled bookings have a longer lead time on average.
* There are 2 bookings not Cancelled with higher lead time these bookings could by the loyal customers
* Bookings that has more than 7 months lead time are more likely to be canceled compared to confirmed

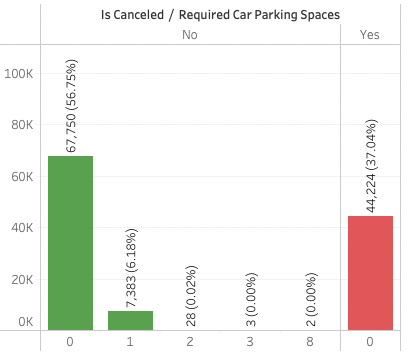
### Total Number of Special request and Cancellation

|  |  |
| --- | --- |
|  |  |

Figure 38 Lead Time and Cancellation

* Customers who cancel their bookings make on average fewer special requests. The potential reasons what were discussed above.

### Total Number of Parking space request and Cancellation



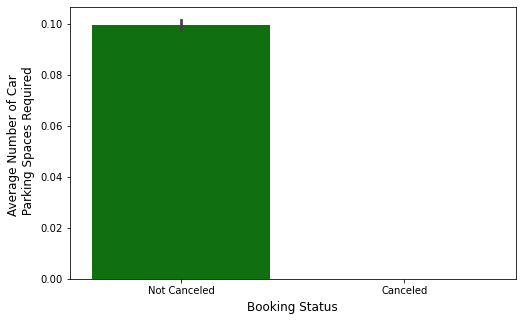


Figure 39 Total Number of Parking space request and Cancellation

* The customers are not cancelled their bookings tend to require more parking spaces.
* There are around **7383 (6.2 %)** that required car parking space(s) that require a parking space **there not a single booking that’s Cancelled (0 Cancellation)**

## C) Removal Of Unwanted Variables.

Since we have created and transformed the existing variables we can remove the variables which is not really required.

Table 1 Removal of Unwanted Variables

|  |  |  |
| --- | --- | --- |
| **Sno.** | **Variable** | **Reason** |
| 1 | booking\_date | This has been transferred to lead time and since this is a date variable which not required for classification algorithm |
| 2 | arrival\_date | This has been transferred to lead time and since this is a date variable which not required for classification algorithm |
| 3 | stays\_in\_weekend\_nights | Transformed to total stays |
| 4 | stays\_in\_week\_nights | Transformed to total stays |
| 5 | adults | Transformed to total\_guests |
| 6 | children | Transformed to total\_guests |
| 8 | agent | Transformed to agent\_by\_booking |
| 9 | company | Transferred to booking\_via\_company |
| 10 | booking\_changes | Transformed to booking\_changes\_new for gropping |

## D) Missing Values

In the categorical variables like Agent or Company, “NULL” is presented as one of the categories. This should not be considered a missing value, but rather as “not applicable”. For example, if a booking “Agent” is defined as “NULL” it means that the booking did not came from a travel agent." As a result, "NULL" values for agent and company will be changed to No Agent and No Company for clarity purposes.

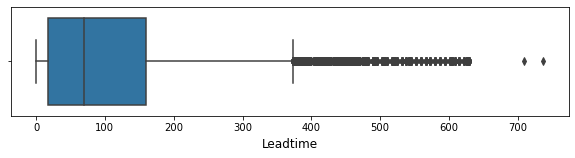
The missing value in the country can be updates as the maximum occurrence value since there is very minimal missing values.

Table 2 Missing Values

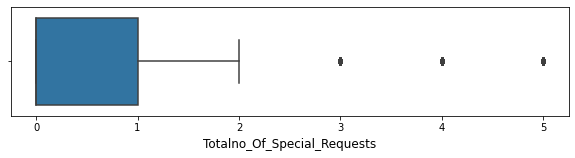
|  |  |  |  |
| --- | --- | --- | --- |
| SNo. | Feature Name | # of Missing value | Possible Action |
| 1 | Country | 488 (0.41%) | Can be filled with maximum occurrence |
| 2 | Agent | 16340 (13.69%) | This feature is related to ID information details which would not require for ML for further analysis we are creating the new variables as booking\_by\_agent by   * Agent\_Booking – the data has the values * Not Agent Booking – where the agent id exists |
| 3 | Company | 112593 (94.31%) | This feature is related to ID information details which would not require for ML for further analysis we are creating the new variables as booking\_via\_company by   * Booking Via Company – the data has the values * Booking Not Via Company – where the agent id exists |

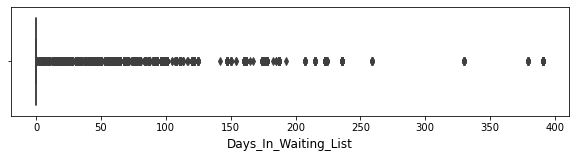
## E) Outlier Treatment Of Values

Outliers are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions. Unfortunately, all analysts will confront [outliers](https://statisticsbyjim.com/glossary/outliers/) and be forced to make decisions about what to do with them.









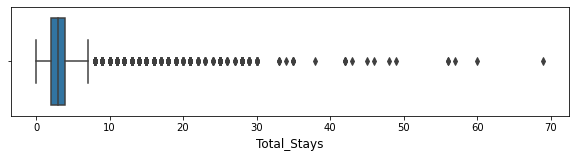
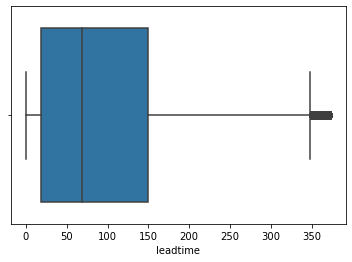




Figure 40 Outlier Treatment

* There are outliers present in all the above variables.
* The variables "Total\_Guest", "Total\_Stays", "Totalno\_Special\_Request" , "Required\_Car\_Parking\_Spaces" and “Days\_In\_Waiting\_List” are having less frequency so we can skip for the outlier treatment.
* The variable “Lead\_time” will be treated the outliers using IQR Value.



There is a much improvement in the outliers of the leadtime.

## F) Variable Transformations

Below are the variables transferred to new variables to get the meaning information from the data provided.

Table 3 Variable Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Sno. | | Variable | Reason |
| 1 | booking\_changes\_new | | Created new variable to split changes happened after booking or not. |
| 2 | booking\_via\_company | | Whether booking done via company |
| 3 | booking \_via\_agent | | Where the booking done by travel agent |
| 4 | Lead Time Month | | This feature created to find average month the lead time for each booking |

## G) Addition of New Variables

We have created the new feature to analyse more insights on the data and to identify how the new feature has impact to the target variable.

Table 4 New Variables

|  |  |  |
| --- | --- | --- |
| Sno. | Variable | Reason |
| 1 | a\_year | Year of Arrival date of the Booking |
| 2 | a\_month | Month of the Arrival date of the Booking(January, February…. December) |
| 3 | a\_day\_of\_week | Weekday of the Arrival date of the booking(Sunday, Monday….) |
| 4 | a\_weekno | Week of the Year for the Arrival date of the booking |
| 5 | lead\_time | Identify the difference of Booking Date and Arrival Date |
| 6 | Country\_name | Created new feature to more idea about the country code provided in the data this achieved as extracted the country details using iso3166 python package. |
| 7 | total\_guest | The customer classified as adult and children to identify the total guest we have added both the features. |
| 8 | total\_stays | The data provided as number of week day nights and week end night days so we have created this variable to calculate the total number of nights stayed by the customer |
| 9 | country\_new | We assume the all the local booking is from Portugal since nearly 40% of booking from Portugal so we split the data based on the Local(Portugal) and International |

# **4) Business insights from EDA**

## A) Is The Data Balanced?

This data is almost balance, so later on for the machine learning process we won’t need to do an imbalance handling.

### Identify the Metrics

Accuracy is one of the common evaluation metrics in classification problems, that is the total number of correct predictions divided by the total number of predictions made for a dataset. Accuracy is useful when the target class is *well balanced* but is not a good choice with unbalanced classes.

Since the data is balanced the we are going to use the metric as accuracy.

## B) Any business insights using clustering  (if applicable)

The features are pretty straight forward so we are not using clustering for to perform the EDA also it might not require before machine learning so as of now clustering is not applicable for this problem.

## C) Any other business insights

### Business Insights

The below table describes how the target variable has an impact to the each feature.

Table 5 Business Insights

|  |  |  |
| --- | --- | --- |
| SNo | Feature Name | Insights |
| 1 | Lead Time | We have grouped the lead time into monthly  (30 days month) lead time to make it more general to analyse compared to a specific number of days   * Booking that has 0 – 7 Months lead time have a higher confirmed booking rate **( > 50%)** to its cancelled rate. * Booking has more than 0-7 Months has higher cancelation rate **( > 50%)** to its confirmed rate.   + Booking that has **more than 7 months of lead time are more likely to be cancelled** than confirmed   + **Cancellation is positively correlated with lead time**(the higher the lead time the higher the cancellation rate) |
| 2 | Deposit Type | There are 3 kinds of deposit type in the data set are **NO Deposit, NO Refund, and Refundable**   * **No Refund Booking has the highest cancellation rate at 99.4%** * **No Deposit has cancellation rate of 28.3 %** * **While Refundable has cancellation rate around 22%**   From the hotels point of view there is nothing alarming since they don't lose revenue when no refund booking is cancelled, but it's always a good practice to question something is extraordinary, **why does non-refundable booking are most likely to be cancelled?** |
| 3 | Market Segment | * Based on the analysis the **corporate** , **Direct**, and **Aviation** has a cancellation rate around **18 - 22 %** of their booking * **Travel Agent (Online / Offline)** has a cancellation rate around **34 - 36 %** * Lastly **Group** has the highest cancellation rate around **61 %**   Based on this can be concluded the **group booking are the market segment that's most likely to be cancelled** compared to other market segment while **Direct has the lowest cancellation rate at 15%** (Outside Complimentary) |
| 4 | Location | The booking location in this dataset we originally have **178 countries**(including Portugal), it's not efficient and not effective to aggregating every country with Portugal in this one we split the **booking location into 2 Local (Booking that is from Portugal) and International (Booking Outside Portugal)**   * International Booking have **24% cancellation Rate** while Local Booking have **56% cancellation Rate** |
| 5 | Repeated Guests | * Whoever cancelled the booking before have a high impact(92%) to the current cancellation. * We can we can see the highest of the data is not cancelled and no previous cancellations so we assume there are more loyal customers in this scenario. * **Booking that's originally wasn't cancelled has 34% Cancellation rate** |
| 6 | Parking Space | This is one of the not common metrics to look at when it comes to predicting cancellation and analysing cancellation, however in this data set there are around **7383 (6.2 %)** that required car parking space(s).   * **7383 Bookings** that require a parking space **there not a single booking that’s Cancelled (0 Cancellation)** * **This conclude that booking that required a parking space will high likely to be confirmed** |
| 7 | Special Request | The number of special request(s) in a booking apparently affecting the cancellation rate of a booking from our analysis we see that booking that has no special request are more likely to cancelled compared to booking that has a special request   * **The cancellation rate of booking that has a special request is ranging from 5 - 22 % with booking with 5 special requests has the lowest cancellation rate** * **While Booking with no special request has cancellation rate of 48%** |

### Recommendations

Table 6 Recommendations

|  |  |  |
| --- | --- | --- |
| SNO | Head | Recommendations |
| 1 | Only Non Refundable Deposit For Group Booking | This analysis results that the group booking has the highest cancellation rate among all market segment, only allowing non-refundable deposit for group booking will help protect the hotel from losing revenue due to last minute cancellation and not able to find replacement. **Only Allowing Non Refundable Rates might result in fewer bookings for Group**, however it might protect the hotel from losing revenue |
| 2 | Setting Maximum Lead Time for Booking | The pattern of the booking that has more than **210 days of lead** time are more likely to be cancelled, setting up maximum lead time means it won’t be able to make booking that's too far in advance (**> 210 days**), and setting maximum advance reservation will help you to reduce cancellation |
| 3 | Increase Direct Booking Market Segment | The dataset has direct booking has the least cancellation rate **15%** (outside complimentary) compared to other market segment, with only being 10% of total booking market segment having more booking from direct market segment will likely to reduce the number of cancellation. |

**Few Strategies to increase Direct Booking**

1. Have a mobile-friendly hotel website

* Website should be accessible to any device
* Offer & Ensure Best Rate Guarantee
* Highlight the unique selling services

1. Optimize website to rank on Google

* Nowadays, your guests would always explore your hotel and more options on search engines like Google
* Need to perform search engine optimization (SEO) of your hotel website this will help to increase rank in the search engine so your hotel appears in first page of search.

1. Implement a live chatbot to attend guest inquiries.

* Its becomes easy for you to provide instant replies to your website visitors.
* When they will get their answer in a fraction of seconds, they will be able to make a decision instantly.

Source : [Ezeeabsolute](https://www.ezeeabsolute.com/blog/increase-direct-hotel-bookings/)