

**TASK**

**Exploratory Data Analysis on the Hotel Bookings Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

‘hotel\_bookings.csv’ data set reflects hotel bookings with many features that are considered before the bookings are processed e.g. hotel type, number of children/ babies/ adults, reservation date, lead time(when the booking was done away from the reservation date), the reserved room type, meal option, etc.

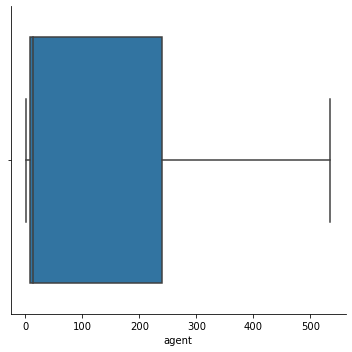
**DATA CLEANING**

All the columns of this data set seem to be of value for our EDA except for 'arrival\_date\_week\_number' column since it represents the week number of the year. It doesn’t seem like it will be useful, so we drop it.

We also drop any duplicate rows that may be present in our data set, and reset the index.

MISSING DATA

We look for columns with missing data i.e. 'children', 'country', 'agent', 'company' and we find that ‘children’ has roughly 0.005% missing values,’country’ has 0.616%, ‘agent’ has 20.625% and ‘company’ has roughly 94%.

We decide to drop ‘company’ since most of its data is missing, we drop the entries missing children and country, and we leave ‘agent’ as it is because first of all, it contains categorical data,it has a mean of 95.692181, max of 535, min of 1, median of 14 and a mode of 9 (this distribution is shown on the diagram on the right hand side).

Its standard deviation is 113.3011,

which is very high. This tells us

that our data is very dispersed away

from each other. So we decide to not

do anything about its missing values. Dropping such entries is not a good

idea considering that more than 20%

is missing.

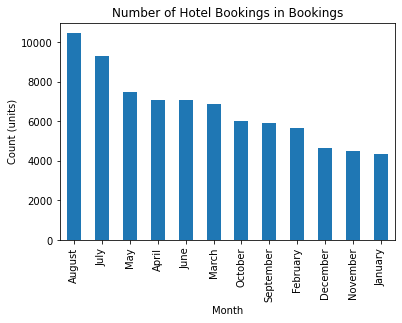
DATA STORIES AND VISUALIZATIONS



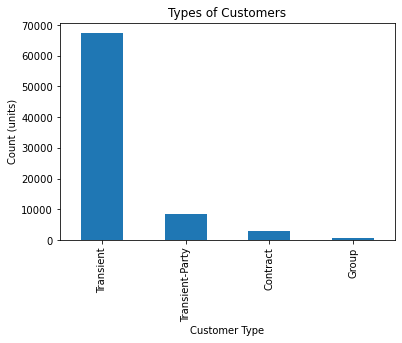
We plot a graph to see the types of hotels included in this data set and how the bookings are distributed between these hotel types. We then find city hotel with the highest bookings of 47 582, along with resort hotel with the bookings of 31 643. (figure on the left)



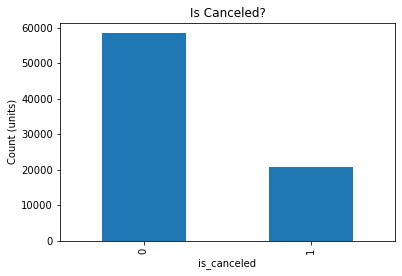
We check for the types of deposits that occurred during these hotel bookings and we find that around 80 000 had no deposits, less than 300 were non refundable deposits and only 86 were refundable. (see diagram on the right.)



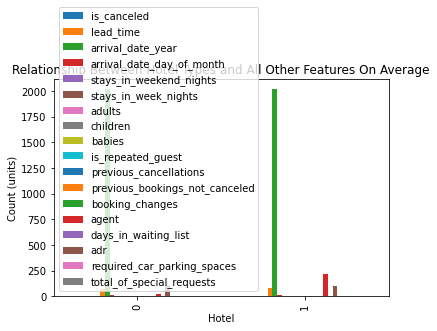
We also check to see how these hotel bookings are distributed with the year and we find that we get the the highest number of bookings in August(more than 10 000), and the lowest in January (less than 5000). (see diagram on the left).



We check for the types of customers and how the bookings distribute among them, and we find that the transient customers do the most bookings of above 60 500, and group customers hardly book. (see on the right).

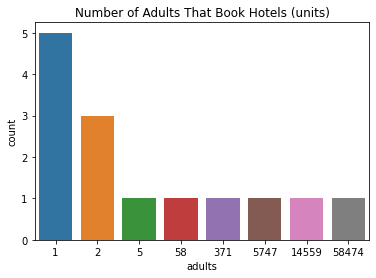
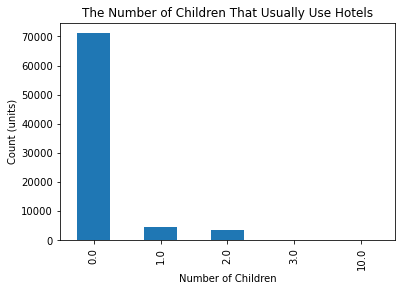
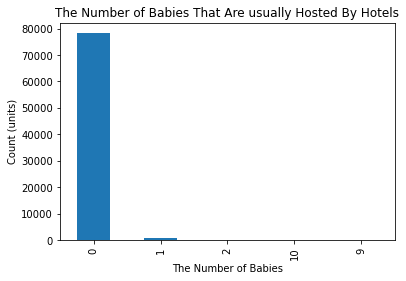


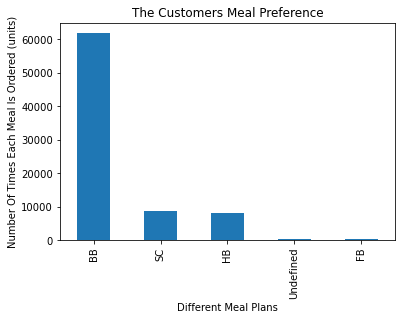
We check to see how likely it is for the bookings to be canceled. We deduce that more than 60 500 bookings were not canceled while close to 20 000 were canceled. We can now check to see what could be the factors of these cancellations.



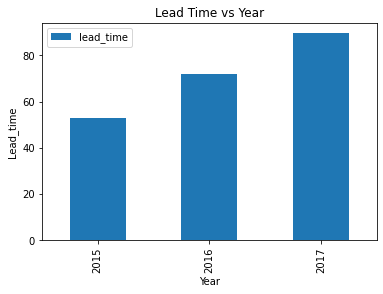
We check to see which hotel type performs better. We see that City hotels have the highest bookings, very little cancellations, most agencies and companies use these hotels more than the resorts.

We check for the number of adults, children, and babies who come to hotels (get bookings) and we find that most of the bookings are done for only 1 adult, children and babies hardly spend nights in hotels.





We also check which meals are mostly ordered in the hotels. We deduce that most customers love BB meal and FB is the least ordered.

****

Lead time is the number of days that elapsed between the entering date of the booking into the PMS and the arrival date

So on average, hotel bookings were done earlier in 2017 than in any other year. Lots of factors could influence such decision e.g. more people are able to afford hotel outings etc.

​

# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : LUNGILE MNGONYAMA**

