**EDA on Hotel Booking Data**

**Deepak kumar Dubey**

Data science trainee,

AlmaBetter,Bangalore

**Abstract:**

The hotel industry is one of the most important components of the wider service industry, catering for customers who require overnight accommodation.

KPIs for the hotel industry are values or metrics that measure the performance of a particular area of hotel operations – or the property as a whole. They ensure clear visibility on the functionality and sustainability of your business within the hospitality landscape.

We had hotel booking data at our hands which had many important parameters in columns through which we were able to analyze some important KPI’s like average daily rate,average length of stay,meals ordered per booking,total bookings.

**1.Problem statement**

Our problem revolves around finding relations between different parameters which are arranged in different columns in our data frame.Lets look at what relations we will try to find out through analysis-

1)What months of the year get the most bookings?Does the average daily rate get affected by this??

2)Is there any relation between number of bookings happening and lead time?

3)How many bookings got canceled?

4)Is there any relation between the number of bookings and cancellations?

5)What type of hotel gets more bookings?

6)People who ought to stay for more nights opt for which type of hotel?

7)Does the size of the family affect which type of hotel the family opts for?

8)How many bookings required parking space and which type of hotel did they opt for?

9)What is the average size of families for all the bookings?

10)Through what type of distribution channels do customers get their bookings done?

11)How many customers had to wait for their bookings to get confirmed?Which type of hotel had more waiting time?

12)How many customers opted for a meal and which type of meal was most preferred?

13)What was the average number of nights stay when averaged over all bookings?

**2.Data summary-**

We have data stored in 32 columns out of which we will look at some relevant columns-

**1)Hotel -**Tell us about the type of hotel.We have two types of hotel in our data.

* City hotel
* Resort hotel

2)is\_canceled- Weather booking was canceled or not

* 0 not canceled
* 1 canceled

3)lead\_time -

* No of hours that elapsed between entering date of booking into property management system and arrival date

4)arrival date year

* 2015
* 2016
* 2017

5)arrival\_date\_month- january to december

6)arrival\_date\_week\_number

7)arrival\_date\_day\_of\_month

8)stays\_in\_weekend\_nights-

* No of weekend nights (Sat/Sun) the guest stayed or booked to stay at the hotel

9)stays\_in\_week\_nights

* No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel

10)Adults

11)children

12)Babies

13)meal

* Type of meal booked. Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)

14)country

15)market\_segment-(a group of people who share one or more common characteristics, lumped together for marketing purposes)

* TA: Travel agents
* TO: Tour operators

16)distribution\_channel-(A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer)

* TA/TO
* Corporate
* GDS
* Direct

17)is\_repeated\_guest (value indicating if the booking name was from repeated guest)

* 1: yes
* 0:no

18)previous\_cancellations-

* Number of previous bookings that were canceled by the customer prior to the current booking

9)previous\_bookings\_not\_canceled-

* Number of previous bookings not canceled by the customer prior to the current booking

20)deposit\_type-

* 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.

21)agent-

* ID of the travel agency that made the booking

22)customer\_type-

* 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 with at least other transient booking.

23)adr (average daily rate)-

* Average daily rate=(sum of all lodging transaction/total number of staying night)

24)required\_parking\_space

25)reservation\_status-

* 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

26)reservation\_status\_date-

* Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when the bookings were canceled or when the customer checked-out of the hotel.

**3.Null values:**

In our data ,only 3 columns had null values .And these columns were not involved in our analysis and didn’t create any problems so they were left as it is.

**4.Approach used:**

Which hotel type is more preferred amongst two types of hotel was determined by grouping the data based on two hotels and then use aggregate count function .Then after getting this data how customer bookings vary monthly was analyzed i.e which months of the year receive most bookings.Then ,weather avg monthly lead time has any relation with number of monthly bookings was analyzed by determining correlation coefficient.Then how many bookings out of total bookings got canceled was determined.Was there any correlation between monthly bookings and monthly cancellations was also found out through correlation coefficient.Then what was average stay per bookings was determined by using distribution plot and using mean aggregate function.Then customers who opted to stay for longer duration opted for which type of hotel was determined using group by and aggregate functions.

We divided all the bookings into 3 categories based on number of members who were visiting per bookings-First was bookings that had only one adult visiting,Second was which had two adults and third was bookings with more than two persons visiting.Then it was determined which among these 3 categories most bookings belonged to.Which hotel did families with children and babies preferred was determined by grouping the family dataframe and using aggregate mean function to know which type of hotel had more mean babies and children.

How many bookings required parking spaces was determined using the value counts method.Which type of hotel did those bookings prefer was also determined by grouping on hotel .What distribution channel do customers most prefer to get their bookings done was determined using the value counts method.How many bookings opted for meal was also determined using value counts and what type of meal was most preferred was also determined.

**6.Challenges faced:**

* Limiting the sns countplot so that only 10 most counts get plotted out of 177.

**5. Important python libraries used:**

* Pandas
* Matplotlib.pyplot
* Seaborn

**6.Conclusion:**

* May to September months were getting the most amount of bookings and average adr for these months was also high as compared to other months.
* Correlation coefficient came out around 0.82 which represents strong positive correlation.
* Around 30% of bookings were getting canceled.
* Correlation coefficient of 0.98 was found out which represents almost perfect correlation.
* City type hotels were getting twice the number of bookings than Resort type.
* People who ought to stay longer preferred Resort type hotels.
* Found out families having children and babies preferred Resort type hotels over City types a little more.
* Just around 6-7% of people required parking spaces and most of them opted for Resort type hotels.
* Couples(no babies and children) booked around 60% of the bookings followed by single adults.
* Most of the bookings are done through online traveling agents.
* Around 3 % people had to wait .Since most people opt for city type hotels ,city type has more avg waiting time.
* Almost 98 % of customers opted for meals and most opted for bed and breakfast time i.e BB type.
* 3.42 nights .Distribution plot was drawn for clear visualization.

**7.References:**

* W3schools
* Geek for geeks
* Analytics vidhya