cancelation cancelation cancelation !!

Introduction cancelation cancelation cancelation this is one of the biggest problems that faces the tourism industry which make this industry more risky than other investments , predicting the real demand is a real challenge for the managers in this field and helps them improve their profits, decrease the risk, and be always ready with enough facilities

This dataset contains two datasets with hotel demand data. One of the hotels (H1) is a resort hotel and the other is a city hotel (H2). Both datasets share the same structure, with 31 variables describing the 40,060 observations of H1 and 79,330 observations of H2. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Both hotels are located in Portugal: H1 at the resort region of Algarve and H2 at the city of Lisbon our main aimfrom this project is to discover the data ,knowing the distributions of our features and try to discover the realationships between them focusing the problem of high_cancelation_rate in order to find advices to decrease this rate or even predict it accordings the different features

the cancelation and its prediction is a real problem for the tourism industry and good understandig for this problem and the features that related with will be very useful to decrease the investments's risk of this important industry deep understanding of these relationships between cancelations and any features that related with will help managers to improve the confirmation rate of their reservations from the dataset we could find some possible relationships between the cancelations and some features even they are not strong or certain there is a possible relationship between the customers with previous cancelation and the future cancelations the cancelations increses when the adr is in its common mode specially when the total nights booked are more than 15 nights we couldn't find any clear and strong realationship between any of the features and the cancelation may be we have to include more factors in the dataset we can set a question in cancelation form asking customers to set a reason , that will be more helpful by deeper research we can build a stronger model in future to predict the actual confirmed reservations which will help the managers be ready for the real demand by deeper research we can advice the managers to improve choosing their agents and working together with them to improve the occupation of their hotels most customers are from Portugal , nice but the managers must give extra attention for the international advertising and international distribution channels

it seems that we have smoe outliers observations in the data specially in ADR feture so we have to check

we can note that:

> - we have two features that miss alot of data 'company' & 'agent'
> - some other features that have few unique values would be better to convert their type to **category**
> - we have a big number of **duplicates** in our data **but** whereas the variables don't include any unique identification column or names and this data is professionally gathered which dosen't allow this huge number of duplicates so we can freely assume that this observations are just repeated observations for different guests or groups
> - we can add a coloumn of the **total_nights** stayed which will be useful in our analysis
> - the data has some outliers observation that would be better to drop to improve the sense of the statistics and any predction-models we can generate from the data
> - its pretty clear that we have some **outliers** in the data and we have to handle to make sense to our plots and models

summary

> **two third of the reservation was for the city hote and one third fot the resort hotel**
>
> **the overall ratio of cancelation is 37% which tells that cancelations is a real problem**
>
> **the overall cancelation rate is 37% and this percentage is realy big and need to deep understand to get good predicted**
> **this percentage is differ between the two hotels , the city hotel has around 1.5 times worser rate city hotel is most booked with 66.5% from all bookings against 35.5% for resort hotel but also it has the biggest percentage of the total cancelation (74.5%) - the plots say that the distribution of the ADR is multimodal with multi peaks can also consider as right skewed distribution which is the favorite for the investorsin the long_term investment - while the distributions of the lead_time and the total_nights are straight and clearly positive skewed (the mean is right of the peak) it was clear that august and julie** are our top_season when we expect the double number of guests than other monthes like january and december but they are also the **top_season of the cancelation__ - as expected the ADR of the top_season_monthes is the greatest even with the high rate of cancelation** in augest an july but both hotels can expect to get their best rate in these monthes **travel agents & tour operators are have the biggest reservation ratio but also the biggest cancelation ratio while the direct bookings has the best cancelation ratio - online travel_agents are th bigest source for the reservations of our destination, while the direct_reservation has the least cancelation ratio - booking with no deposit is the most common type - almost all of nun_refund deposit type reservation were canceled that was totaly unexpected - the most common type of customers is the transient travelers that's may be why the most of bookings are from one to three nights the most pepole reserving the destination are native citizens - we saw from this graph that the pepole how had more than 6 previous cancelations will often cancel their future reseravtion thoese pepole may be not serious with their reservations from the graphs we saw that reservations with middle daily average rate have more chance to be canceled specially these reservations of more than 15 nights**

conclusions

> - **the cancelation and its prediction is a real problem for the tourism industry and good understandig for this problem and the features that related with will be very useful to decrease the investments's risk of this important industry**
> - **deep understanding of these relationships between cancelations and any features that related with will help managers to improve the confirmation rate of their reservations from the dataset we could find some possible relationships between the cancelations and some features even they are not strong or certain**
> - there is a possible relationship between the customers with previous cancelation and the future cancelations
> - the cancelations increses when the adr is in its common mode specially when the total nights booked are more than 15 nights
> - we couldn't find any clear and strong realationship between any of the features and the cancelation may be we have to include more factors in the dataset we can set a question in cancelation form asking customers to set a reason , that will be more helpful
> - by deeper research we can build a stronger model in future to predict the actual confirmed reservations which will help the managers be ready for the real demand
> - by deeper research we can advice the managers to improve choosing their agents and working together with them to improve the occupation of their hotels
> - most customers are from Portugal , nice but the managers must give extra attention for the international advertising and international distribution channels

resoures

> the site of sincedirect the dataset source
> pandas library doucumentation
> seaborn documentaion
> matplotlib documentation
> stack overflow