VIDYASAGAR UNIVERSITY

PREDICTING THE HOTEL BOOKING CANCELATION BEHAVIOUR OF THE EUROPIAN HOTEL

Subject:- Statistics

Registration No: - 1160497 of 2020-2021

Roll No: - 1126116-200140

Session: - 2022-2023

PREDICTING THE HOTEL BOOKING CANCELATION BEHAVIOUR OF TWO EUROPIAN HOTEL

CONTENTS

INTRODUCTION

For this project we will be analyzing the hotel booking data. This Dataset contains the booking information of two European hotel **City Hotel** and **Resort Hotel.** The dataset contain 118566 observation represent the hotel booking between the hotel booking between 1st July , 2015 to 31st august ,2017 The data includes the booking that effectively arrived and the booking that were canceled.

Hotel industry is a very volatile industry and the booking depends on various factor given in the dataset.

PROBLEM SEGMENT

- The main objective behind this project to explore and analyze data to discover the important factors in which the cancelation depends.
- And we want to fit a model on various factor to predict the guest booking behavior.

SOURCE OF DATA

- The data is originally from the article Hotel booking Demand Dataset written by Nuno Antonio, Ana Almeida, and Luis Nunes for the data in Brief, Volume 22, February, 2019

Data source

- Available on Kaggle

DATA POINT AND DESCRIPTION

- Hotel : Resort or city Hotel
- Is_canceled : value indicating if the booking was canceled (1) or not(0).
- Lead_time : the number of days that elapsed between entering the date of the booking and the arrival date.
- Market_segment : market segment designation .
- Distribution_chanel : booking distribution channel.
- Previous_booking : (0) represent the guest is new, the previous booking that not canceled by the customer (1) and previous booking that canceled by the customer (2)
- Reaserved_room_changes : if the guest getting the same reaserved room or other rooms.
- Deposit_type : No deposit (0) , non refundable (1) , refundable (2).
- Days_in_wating_list : the no of days the booking was in the waiting list before it was confirmed to the customer .
- Customer_type : type of customer. Contract, group, transient, transient-party.
- ADR : Average daily rates define by dividing the sum of all lodging transaction by the total no of staying nights.
- Required car parking space: the no of required car parking space by the customer.

DATA

hotel	is_canceled	lead_time	Booking_night	m.arket_segment	repetation_type	Reaserved room_changes	days_in_w
City Hotel	0	3	5	Aviation	0	0	
City Hotel	0	1	1	Aviation	1	0	
City Hotel	0	194	194	Online TA	0	0	
City Hotel	0	194	194	Online TA	0	0	
City Hotel	0	74	74	Online TA	0	0	
City Hotel	0	84	84	Online TA	0	0	
City Hotel	0	74	74	Online TA	0	0	
City Hotel	0	76	76	Online TA	0	0	
City Hotel	0	83	83	Online TA	0	0	
City Hotel	0	28	28	Online TA	0	6	
City Hotel	0	6	6	Online TA	0	0	
City Hotel	0	5	5	Online TA	0	0	
City Hotel	0	23	23	Direct	0	0	
Resort Hotel	1	110	111	Online TA	0	0	

Among 118566 observation first 15 observation are presented in the observation

GRAFICAL ANALYSIS ON VARIOUS FACTOR

- Booking cancelation vs lead time

➢ **In the billow table we arranged the observation of cancelation with respect to lead time**

lead_time	resort_0	reasort_1	Proportion_success_ reasort	city_0	city_1	Proportion success city
0_100	19580	5216	0.789643	31753	14750	0.682816
101_200	5352	3207	0.625307	9300	8795	0.513954
201_300	2675	1960	0.577131	3259	4839	0.402445
301_400	900	669	0.573614	1257	3649	0.256217
401_500	393	913	0.300919	264	969	0.214112
above 500	68	412	0.141667	0	64	0

➢ **X-square test**

H_0 : lead time and booking cancelation are independent.

H_1 : lead time and booking cancelation are dependent.

REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 3335.8, df = 5, p-value < 2.2e-16

We reject the null hypothesis.

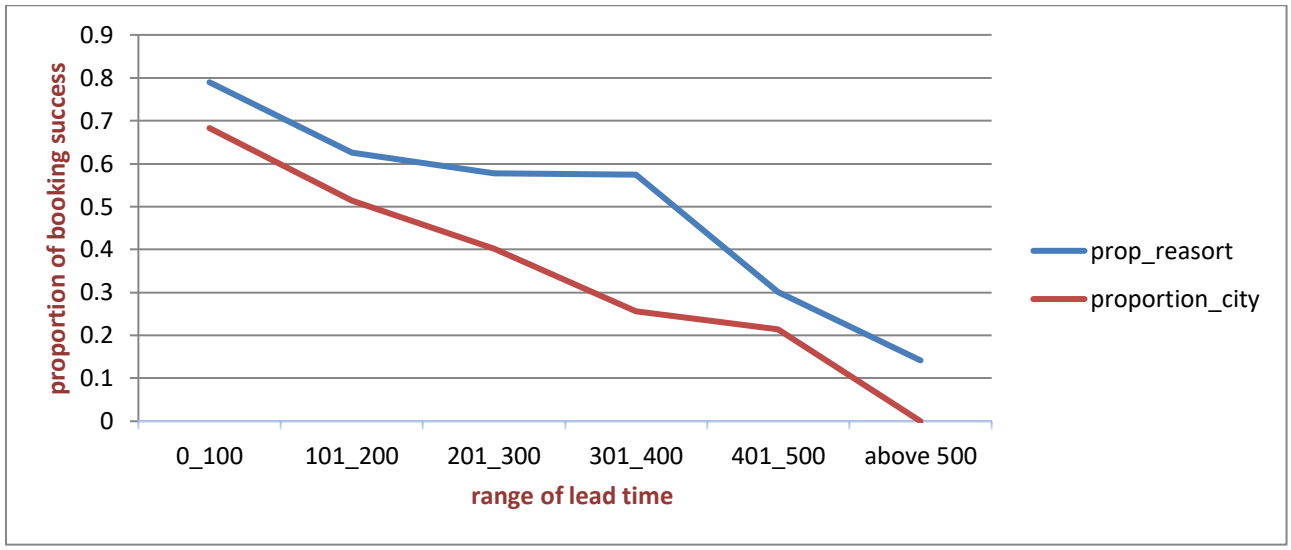
City hotel:

Pearson's Chi-squared test

X-squared = 6270.9, df = 5, p-value < 2.2e-16

We reject the null hypothesis.

- **Graphical representation**



From the upper graph we can analysis that if lead time increases then the booking success rate decreases. We can also see that city hotel booking success rate is comparatively low with respect to resort hotel.

Booking cancelation vs booking nights

- **In the billow table we arranged the observation of cancelation with respect to no of booking nights.**

night_spend	city_0	city_1	proportion_success_city	reasort_0	reasort_1	proportion_success_reasort
1_10	45691	32776	0.582296	27370	10632	0.7202253
11_20	153	240	0.389313	1089	406	0.7284281
above_20	7	50	0.122807	97	72	0.5739645

- **X-square test**

H_0 : booking nights and booking cancelation are independent.

H_1 : booking nights and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 109.02, df = 2, p-value < 2.2e-16

We reject the null hypothesis.

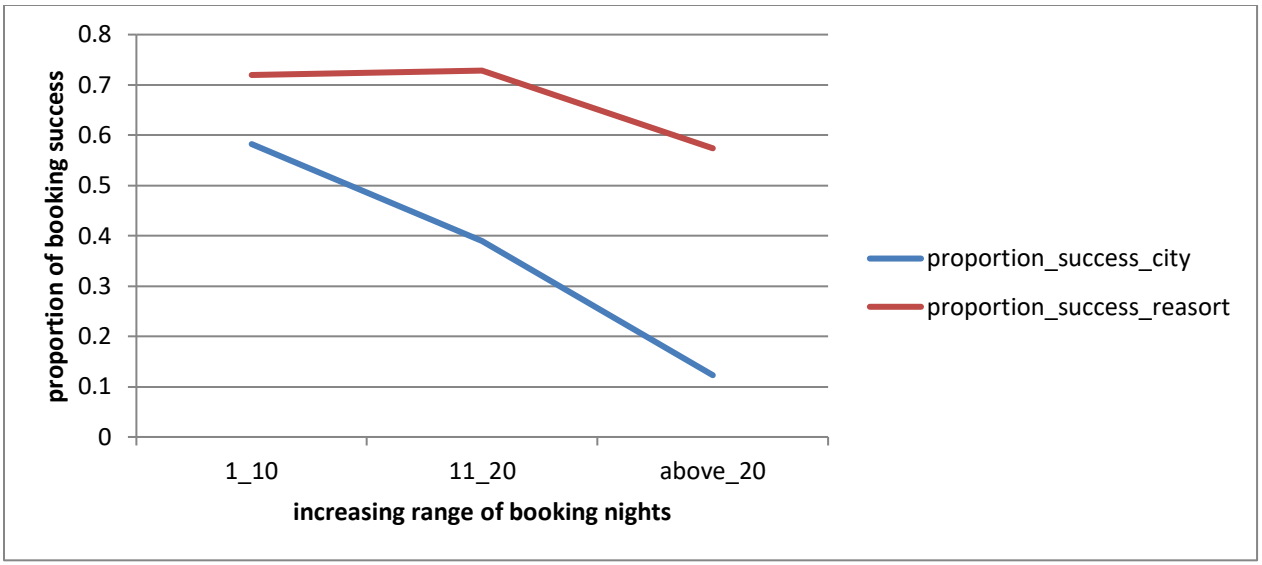
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 18.409, df = 2, p-value = 0.0001006

We reject the null hypothesis.

- **Graphical representation**



From the upper graph we can analysis that if booking night increases then the booking success rate decreases. We can also see that city hotel booking success rate is comparatively high than resort hotel with respect to no of booking nights.

Booking cancelation vs market segment

- **In the billow table we arranged the observation of cancelation with respect to no of market segment**

market _segment	city_0	city_1	proportion_success_city	reasort_0	reasort_1	proportion_success_reasort
Aviation	180	51	0.779221			
Complementary	460	54	0.894942	168	33	0.835821
Corporate	2322	639	0.784195	1920	350	0.845815
Direct	4963	1053	0.824967	5556	877	0.863672
Groups	4335	9620	0.310641	3331	2473	0.573915
Offline TA/TO	9477	7166	0.569429	6271	1135	0.846746
Online TA	24096	14481	0.624621	11310	6242	0.644371
Undefined	0	2	0			

- **X-square test**

H_0 : market segment and booking cancelation are independent.

H_1 : market segment and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X.-squared = 6721.6, df = 7, p-value < 2.2e-16

We null hypothesis reject the

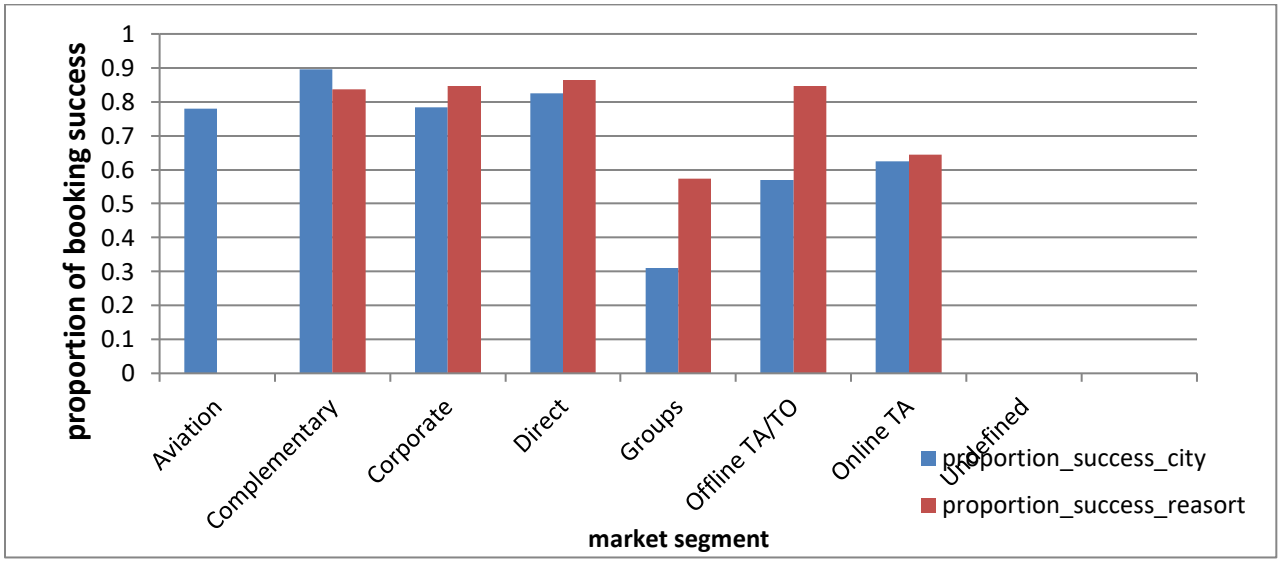
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 2540.1 , df = 5, p-value < 2.2e-16

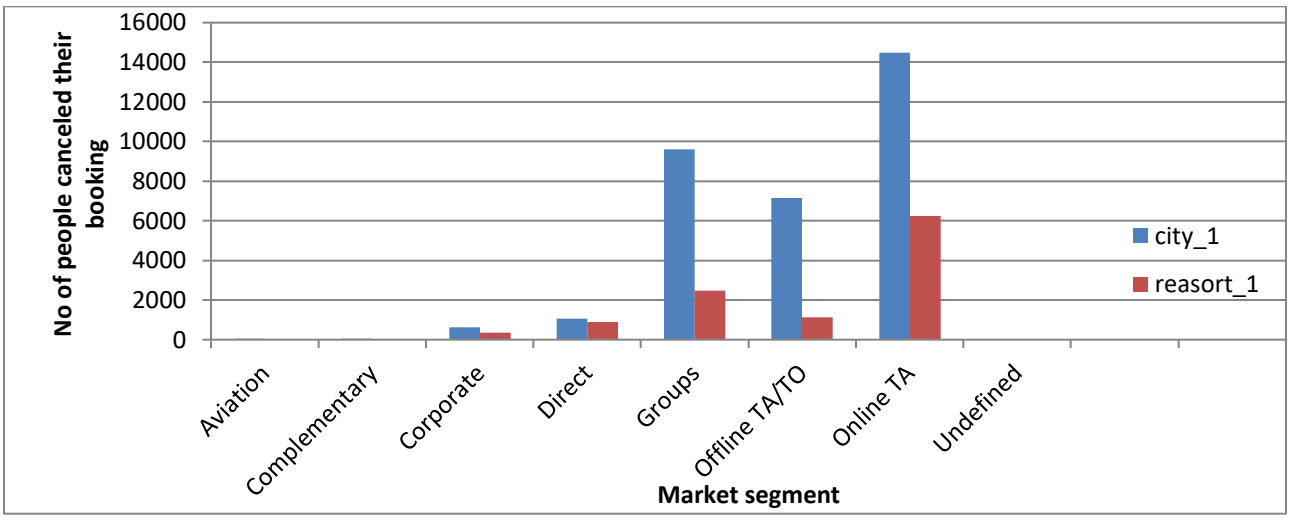
We reject the null hypothesis.

- **Graphical representation**



From the upper graph we can analysis that most of the booking success rate in city hotel in those market segment (aviation, complementary, corporate and direct) and for the resort hotel most of the booking success rate in those market segment (complementary, corporate, direct and offline travel agent)

if lead time increases then the booking success rate decreases. We can also see that city hotel booking success rate is comparatively low with respect to resort hotel.



And by this graph most of the cancelation is done in through online offline and groups segment. And for reasort hotel online and groups and offline segment also.

- Booking cancelation vs previous booking behavior
  - **In the billow table we arranged the observation of cancelation with respect to previous booking behavior**

PREVIOUS_TYPE	C_0	C_1	proportion_cancel_city	R_0	R_1	proportion_cancel_reasort	Percentage of guest
not repeted	44281	27982	0.387224	26420	10153	0.277609	92%
preveous not canceled	1187	68	0.054183	1966	33	0.016508	3%
preveous canceled	365	5016	0.932169	170	924	0.844607	5%

➢ X-square test

H_0 : previous booking behavior and booking cancelation are independent.

H_1 : previous booking behavior and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 6806.4, df = 2, p-value < 2.2e-16

We reject the null hypothesis.

REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 2418.9, df = 2, p-value < 2.2e-16

We reject the null hypothesis.

➢ Graphical representation

In both hotel approximately 92% guest are not repeated and other 8% are repeated among which 3% are previous booking is not canceled and least 5% previous booking is canceled.

And by the upper graph we can analysis that in city hotel approximately 40% new guest are canceled their booking and in resort hotel approximately 37%.

And those who are previous booking is not canceled their cancelation rate very low in both hotel approximately less than 5%.

- Booking cancelation vs booking room changes
- **In the billow table we arranged the observation of cancelation with respect to booking room and assigned room are same or not**

ROOM	C_0	C_1	proportion_cancelation_city	percent of changing	R_0	R_1	proportion_cancelation_reasort
same	39279	32654	0.45395	91.17%	21448	10729	0.333437
change	6554	412	0.535764	9%	7108	381	0.416305

- **X-square test**

H_0 : room changes and booking cancelation are independent.

H_1 : room changes and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 4066.214, df = 1, p-value < 2.2e-16

We reject the null hypothesis

REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 2425.1, df = 1, p-value < 2.2e-16

We reject the null hypothesis

- **Graphical representation**

In both city hotel and resort hotel approximately 91% and 81% time they got the same hotel which they are booked and in city hotel 45% guest canceled their booking although they are getting their same booked room and resort hotel canceled their booking although they are getting their same booked room. And if the room changes the cancelation of booking is comparatively go high.

- Booking cancelation vs Deposit type
- **In the billow table we arranged the observation of cancelation with respect to deposit type**

Row Labels	city_0	city_1	proportion_cancelation_city	percent of changing	reasort_0	reasort_1	proportion_cancelation_reasort	percent of changing
No Deposit	45803	20208	0.306131	83.67%	28367	9438	0.24965	95.31%
Non Refund	24	12844	0.998135	16.31%	69	1650	0.95986	4.33%

Refundable	6	14		0.7	0.03%	120	22	0.15493	0.36%

➢ X-square test

H_0 : deposit type and booking cancelation are independent.

H_1 : deposit type and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 21188, df = 2, p-value < 2.2e-16

We reject the null hypothesis

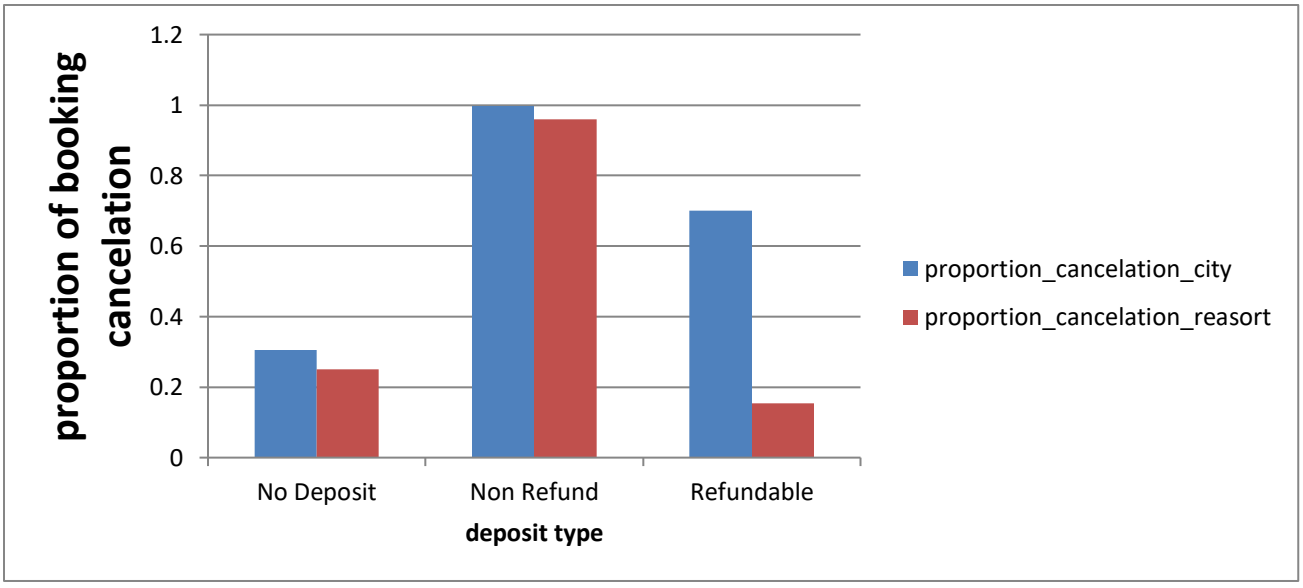
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 4124.1, df = 2, p-value < 2.2e-16

We reject the null hypothesis

➢ Graphical representation



In both city and resort hotel approximately 84% and 93% booking are done with no deposit and 16% and 4% booking are not refundable , other less than 1% booking are refundable

from the upper graph we can analysis that in city hotel approximately 30% guest are canceled their booking who does not give any deposit while booking and in resort hotel approximately 25%.

And those whose booking are non refundable in city hotel approximately 99% guest are canceled their booking and in resort hotel approximately 96% are canceled. And among refundable bookings 70% in city hotel and 15% in resort hotel cancel their booking.

Booking cancelation vs days in waiting list

- **In the given table we arranged the observation of cancelation with respect to waiting days**

wating_days	city_0	city_1	proportion_cancelation_city	**percent of booking**	reasort_0	reasort_1	proportion_cancelation_reasort	**percent of booking**
on date	44736	30724	0.407156	95.64%	28320	11093	0.281455	99.36%
1_30	115	442	0.793537	0.71%	29	5	0.147059	0.09%
31_60	424	1060	0.714286	1.88%	37	5	0.119048	0.11%
60_90	228	331	0.592129	0.71%	52	1	0.018868	0.13%
above 90	330	511	0.60761	1.07%	117	6	0.04878	0.31%

- **X-square test**

H_0 : waiting days and booking cancelation are independent.

H_1 : waiting days and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 1087.6, df = 4, p-value < 2.2e-16

We reject the null hypothesis

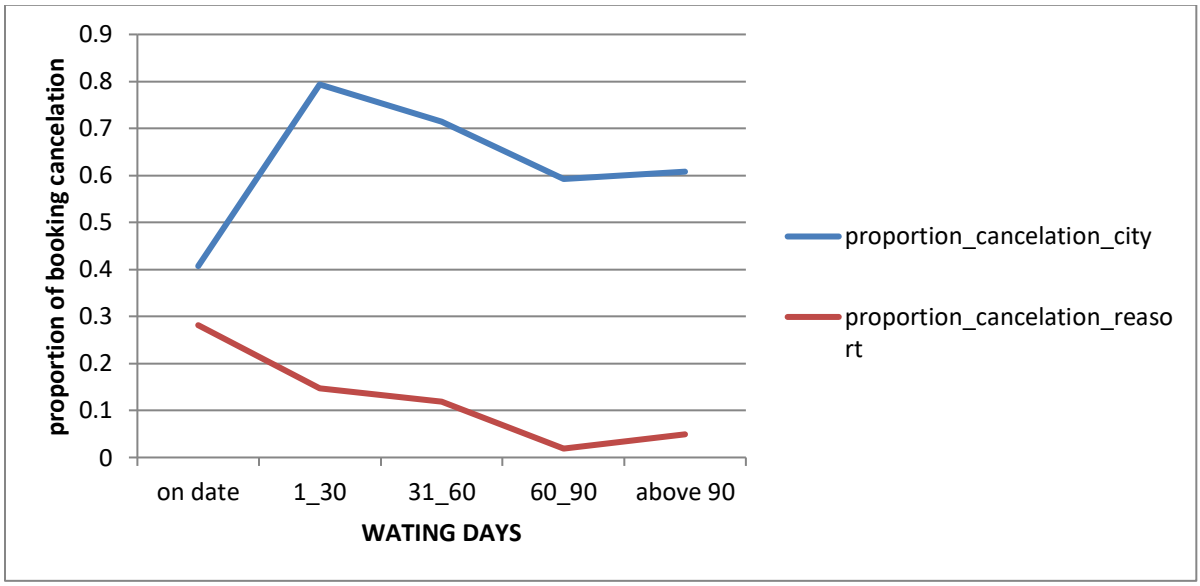
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 59.711, df = 4, p-value = 3.336e-12

We reject the null hypothesis

- **Graphical representation**



From the upper table we can say that in city hotel about 95% booking are completed on date and for resort hotel more than 99% booking are completed are on date

In city hotel the cancelation rate increases with respect to waiting days increases but other hand on resort hotel cancelation rate decreases with respect to increasing of waiting days.

- Booking cancelation vs customer type
  - **In the billow table we arranged the observation of cancelation with respect to customer type**

CUSTOMER_TYPE	C_0	C_1	proportion_cancelation_city	percent of booking	R_0	R_1	proportion_cancelation_reasort	percent of booking
Contract	1184	1105	0.482744	2.90%	1610	157	0.088851	4.45%
Group	261	29	0.1	0.37%	250	29	0.103943	0.70%
Transient	31973	27066	0.458443	74.83%	20489	9407	0.314657	75.37%
Transient-Party	12415	4866	0.281581	21.90%	6207	1517	0.196401	19.47%

  - **X-square test**

H_0 : customer type and booking cancelation are independent.

H_1 : customer type and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 1877.1, df = 3, p-value < 2.2e-16

We reject the null hypothesis

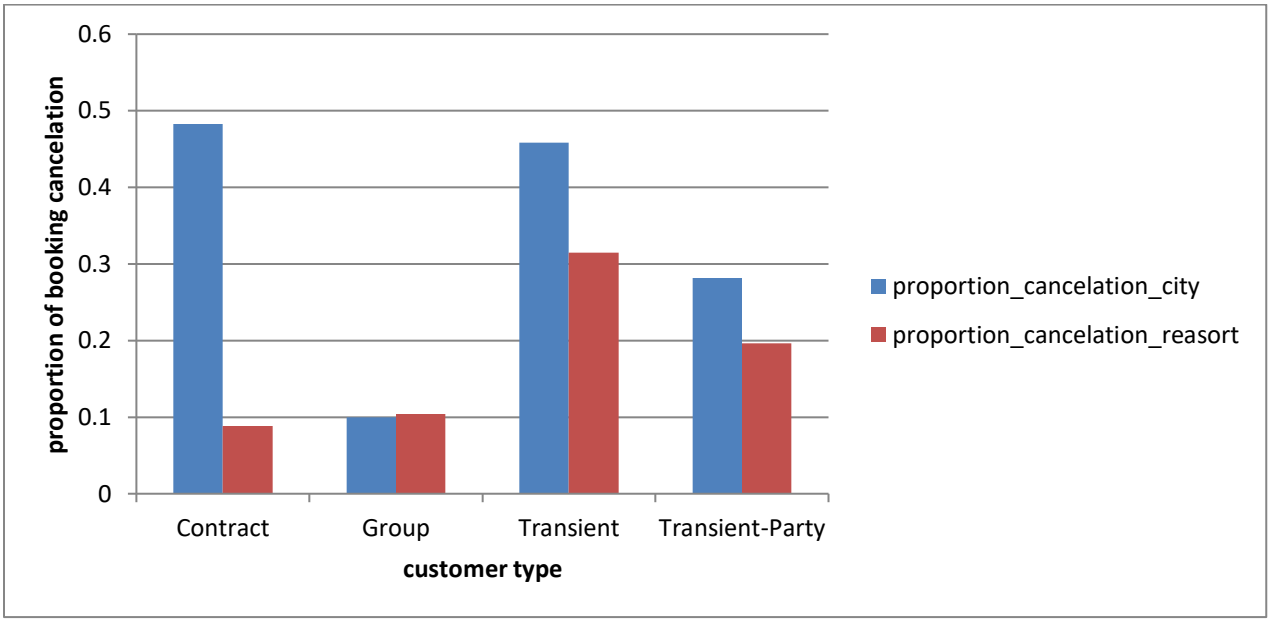
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 808.88, df = 3, p-value < 2.2e-16

We reject the null hypothesis

- **Graphical representation**



Maximum customer in both hotel transient and transient-party type. In city hotel we can see higher cancelation rate in contract and transient type party and in resort hotel intransient and transient party type customer.

- Booking cancelation vs ADR
- **In the billow table we arranged the observation of cancelation with respect to ADR**

price Labels	c_0	c_1	proportion_cancelation_city	**percent of booking**	r_0	r_1	proportion_cancelation_reasort	**percent of booking**
0_50	30304	24384	0.445875	69.31%	22480	9209	0.290606	79.89%
51_100	10967	6726	0.38015	22.42%	4502	1436	0.241832	14.97%
above 100	4562	1956	0.300092	8.26%	1574	465	0.228053	5.14%

- **X-square test**

H_0 : lead time and booking cancelation are independent.

H_1 : lead time and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 650.47, df = 2, p-value < 2.2e-16

We reject the null hypothesis

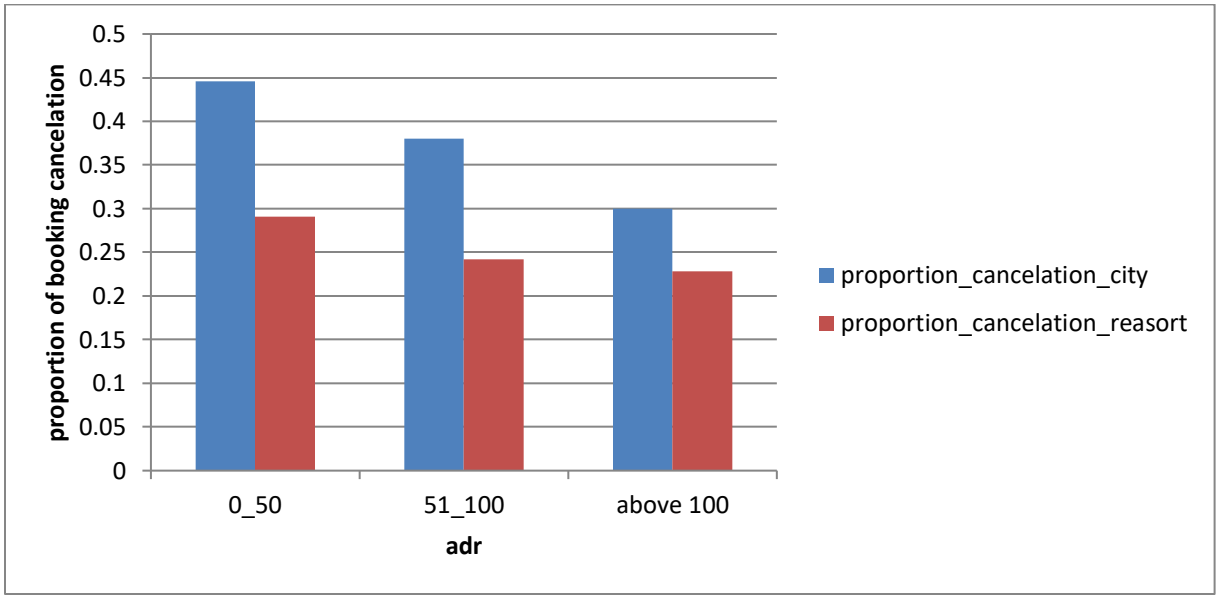
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 87.863, df = 2, p-value < 2.2e-16

We reject the null hypothesis

- **Graphical representation**



Maximum no of guest booked hotel in the price segment 0 to 50 dollar

And the above diagram in the increases of price the cancelation rate decreases.

Booking cancelation vs no of car parking required

- **In the billow table we arranged the observation of cancelation with respect to no of car parking required**

car_parking	C_0	C_1	proportion_cancelation_city	percent of booking	R_0	R_1	proportion_cancelation_resort	percent of booking
0	43915	33066	0.429535	97.57%	23081	11110	0.324939	86.20%
1	1913	0	0	2.42%	5447	0	0	13.73%
2	3	0	0	0.00%	25	0	0	0.06%
3	2	0	0	0.00%	1	0	0	0.00%
8	0	0	0	0.00%	2	0	0	0.01%

- **X-square test**

H_0 : **no of car parking required** and booking cancelation are independent.

H_1 : **no of car parking required** and booking cancelation are dependent.

City hotel:

Pearson's Chi-squared test

X-squared = 12530, df = 3, p-value < 2.2e-16

We reject the null hypothesis

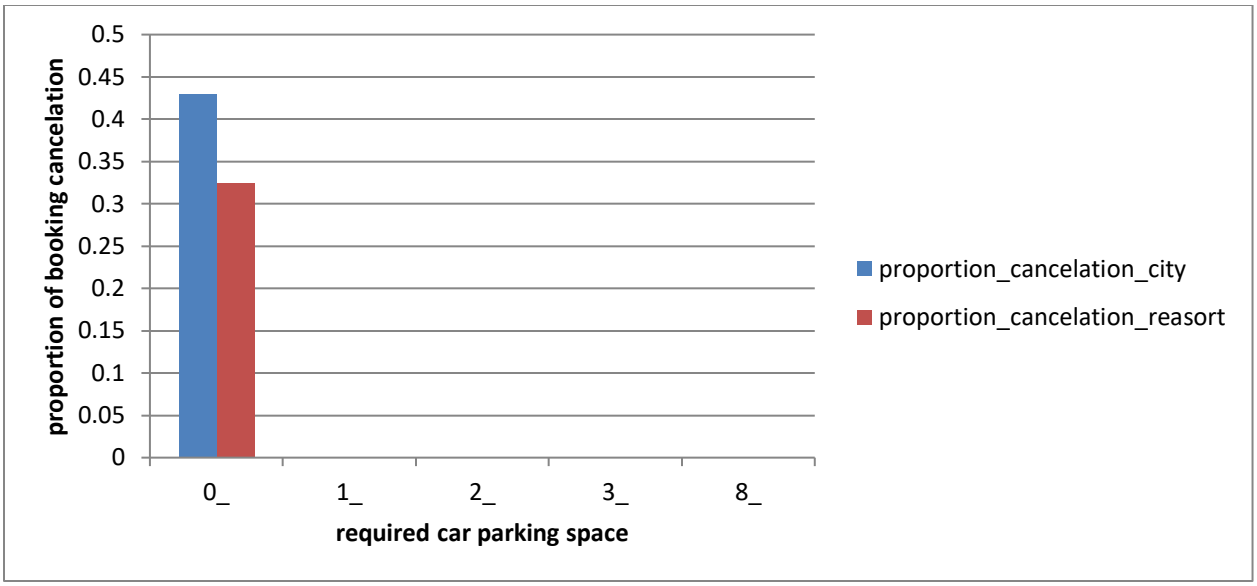
REASORT HOTEL:

Pearson's Chi-squared test

X-squared = 87.863, df = 2, p-value < 2.2e-16

We reject the null hypothesis

- **Graphical representation**



From the upper dataset and diagram we can analyze that the guest who have single car they are looking for choices otherwise not.

Fitting the logistic regression

PROBABLITY CONCLUSSION

p/(1-p) = known as 'Odds'

ln[p/(1-p)]= u + b1 X1 +b2X2 + b3X3 + ……

where, p = probability of accepting

Let,

ln[p/(1-p)] = Y = u + b1 X1 +b2X2 + b3X3 + ……

or, p/(1-p) = e^y

or, p = e^y/(1+e^y)

In this way we can obtain that a guest will cancel or not cancel his booking .

Let, H0 : bi =0 ag H1 : bi ≠ 0

- Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.908049	0.061208	-47.511	< 2e-16	***
lead_time	0.193034	0.009485	20.351	< 2e-16	***
night_spend	0.209675	0.048803	4.296	1.74e-05	***
market_segment	0.389321	0.008217	47.378	< 2e-16	***
Previous booking	1.120974	0.025442	44.061	< 2e-16	***
booking_changes	-0.518188	0.011892	-43.573	< 2e-16	***
deposit_type	4.547851	0.065410	69.529	< 2e-16	***
days_in_waiting_list	-0.053308	0.021493	-2.480	0.0131	*
customer_type	0.015201	0.014963	1.016	0.3097	
ADR	-0.011136	0.013728	-0.811	0.4173	
Required_car_parking	-18.834243	46.732995	-0.403	0.6869	

- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- (Dispersion parameter for binomial family taken to be 1)
- Null deviance: 124336 on 94154 degrees of freedom
- Residual deviance: 87771 on 94144 degrees of freedom
- AIC: 87793

Number of Fisher Scoring iterations: 16

- So by logistic regression we accept null hypothesis for customer_type, ADR , Required_car_parking_spaces at 5% level of significance

Conclusion

- Booking cancelation is increases instead of lead time increases that behave in both hotels.
- If the no of booking days increases then booking cancelation also increases. This cancelation rate is very high specifically for city hotel.
- Groups ,online, offline TA cancelation rate is high for both hotel and most of the guest comes from those segment . specifically city hotel management should be aware on this.
- In both hotel who are previously canceled the booking usually they are avoid to prefer this hotel again
- In both hotel if the booked room changes according to the customer preference then there is very low cancelation rate.
- Wonderfully for both hotel non-refundable booking cancelation rate is too high .And for city hotel refundable booking cancelation also too high.
- In city hotel increasing the days of waiting the cancelation rate is also increases but for resort hotel the cancelation rate decreases.
- For city hotel contractual and transient guest have higher cancelation rate. But for resort hotel transient guest comparatively high cancelation rate than others.

- Booking price increases then the cancelation rate is also increases.

The guest who has more than one car they usually doesn’t cancel their booking

THANK YOU