The Cost of Hotel Booking ~~Cancellation~~

ISSS602 AY2020-21 Assignment 1

Gabriella Pauline Djojosaputro

Introduction

Hotel Booking Cancellation

Problems

- Loss to the hotel revenue
- Not easy to find replacement booking
    - Need to lower price

Aims

Find patterns in cancellation

- Insights to drive efforts in reducing the possibility of cancellation

Data Source: 1 city hotel, 1 resort hotel in Portugal

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief, 22*(2019), 41-49. doi:10.1016/j.dib.2018.11.126

Data Preparation

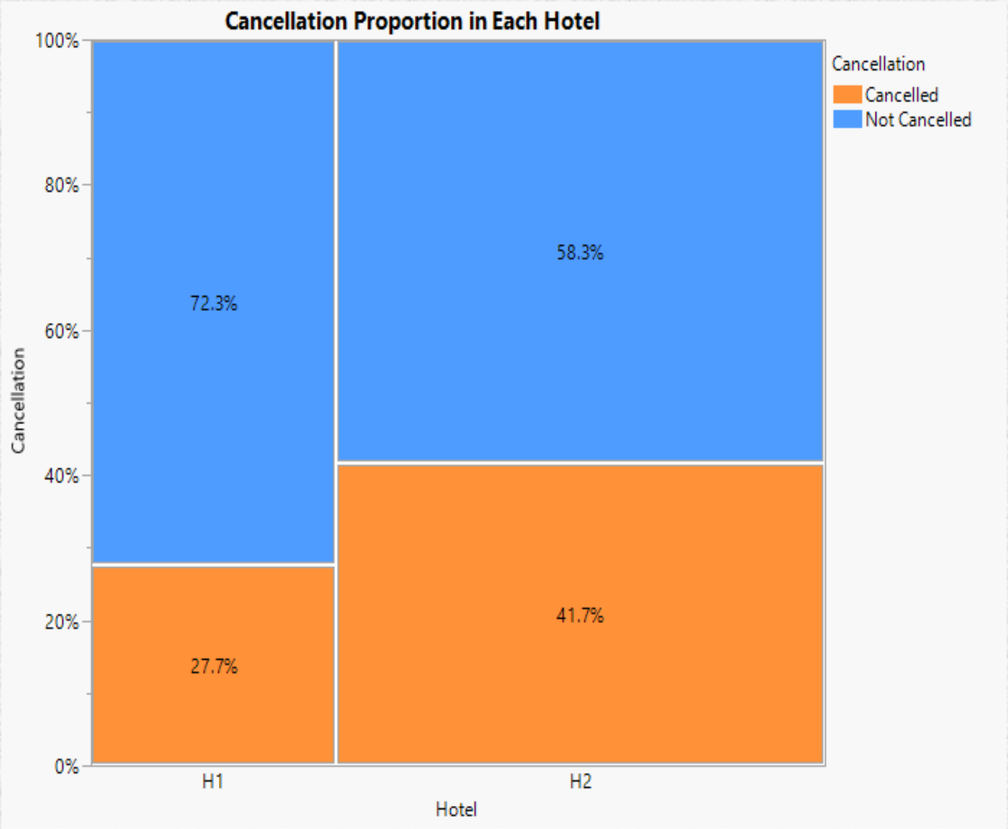
Issues

- Inconsistent way of storing date variables
- Inconsistent way of documenting missing values
- Non-missing values labelled as “NULL”
- Erroneous data: more than four adults in one booking
- Incorrect modelling type

Solution

- Recode values
- Hide and Exclude columns/records
- New formula column

Insight: H2 has different cancellation rate from H1



Cancellation Proportion

- H1: **27.73%**
- H2: **41.73%**

Chi-Square Test

- H_0: There is no difference in proportion of cancellation in H1 and H2.
- H_1: There is a difference in proportion of cancellation in H1 and H2.
- Confidence level: **95%**
- Assumption: No group has less than 5 observations
- P-value: **<.0001**

Contingency Table

Hotel	Count / Total % / Col % / Row %	Cancelled	Not Cancelled	Total
		Cancellation		
H1		11106	28938	40044
		9.30	24.24	33.54
		25.12	38.50	
		27.73	72.27	
H2		33102	46228	79330
		27.73	38.73	66.46
		74.88	61.50	
		41.73	58.27	
Total		44208	75166	119374
		37.03	62.97	

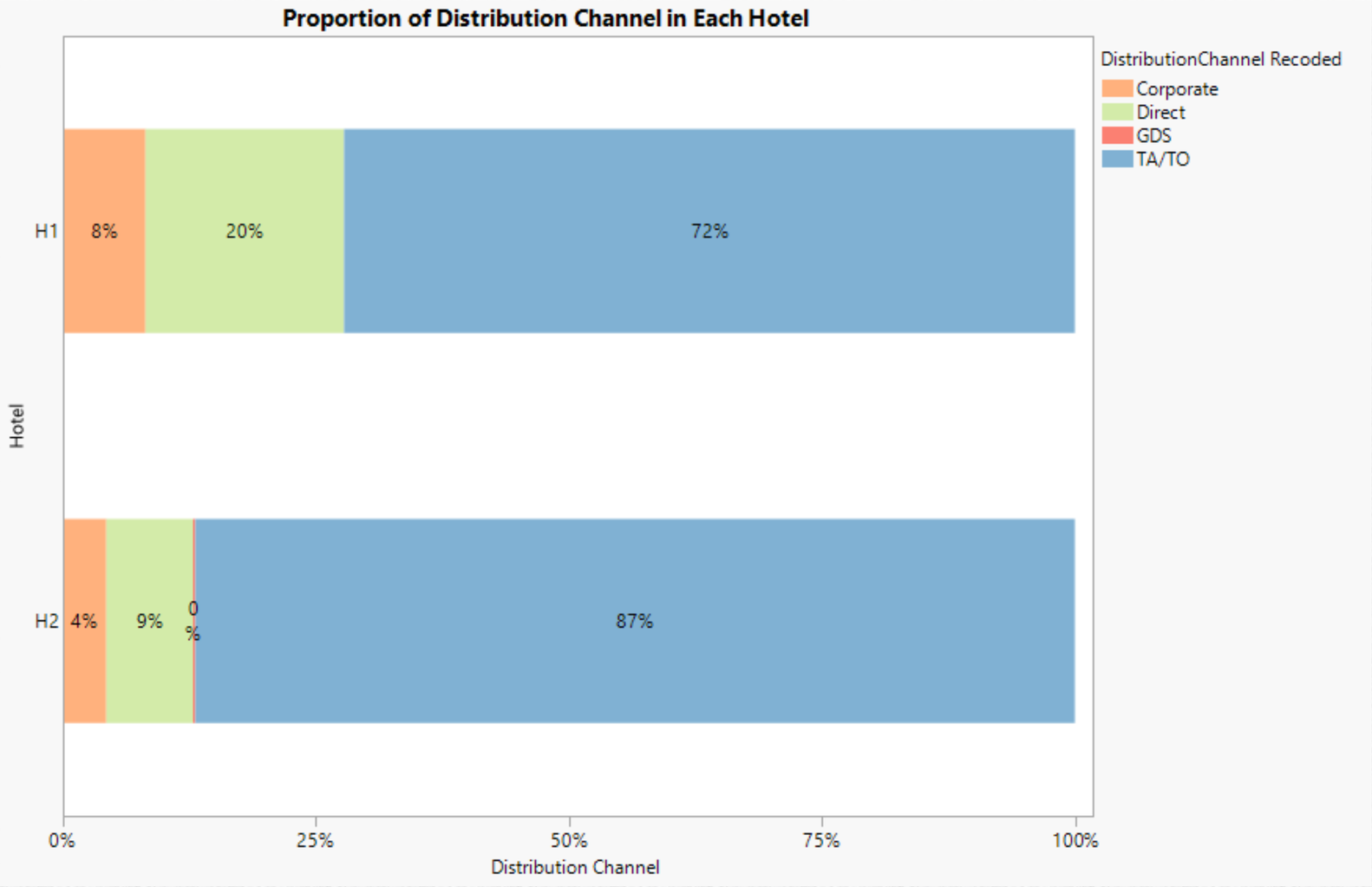
Tests

N	DF	-LogLike	RSquare (U)
119374	1	1143.6906	0.0145

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	2287.381	<.0001*
Pearson	2234.350	<.0001*

Analyze separately

Insight: Difference between H1 and H2



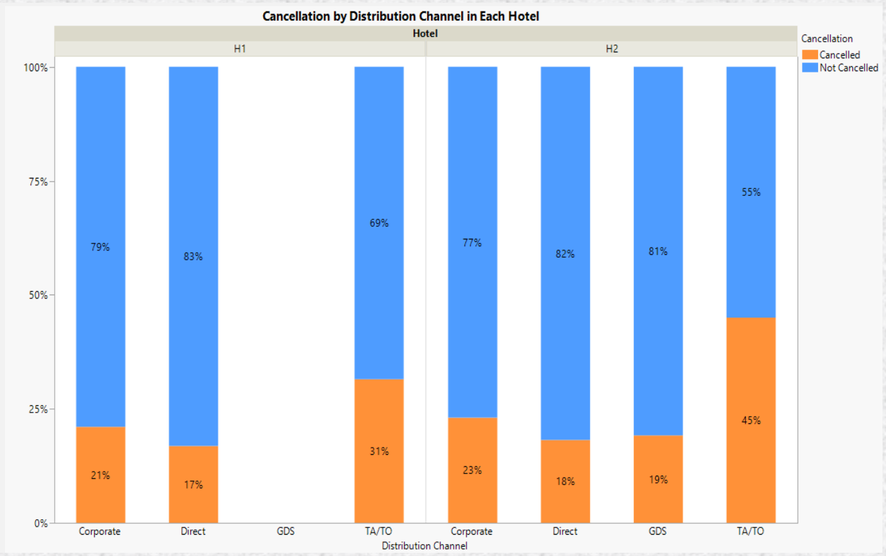
TA/TO Proportion

- H1: **72%**
- H2: **87%**

Other Observations

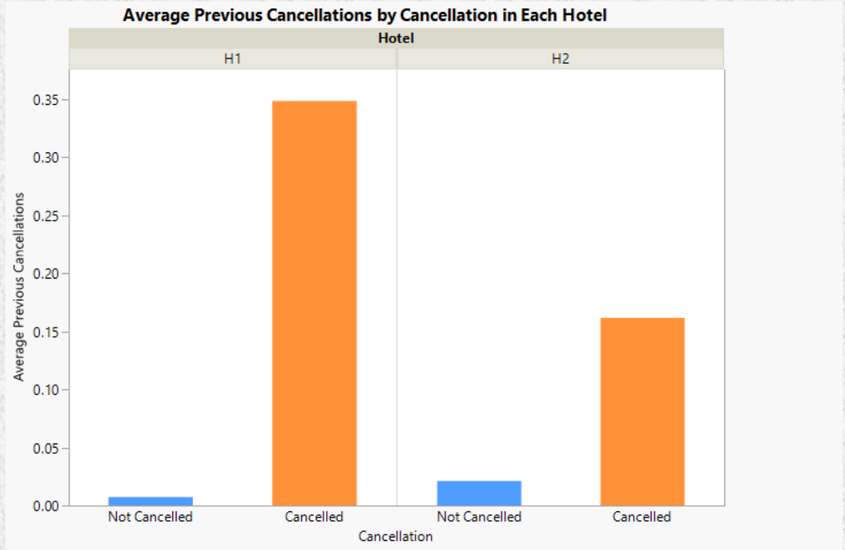
- H1 does not have GDS distribution channel
- Proportion of corporate and direct channel is twice higher in H1 than H2

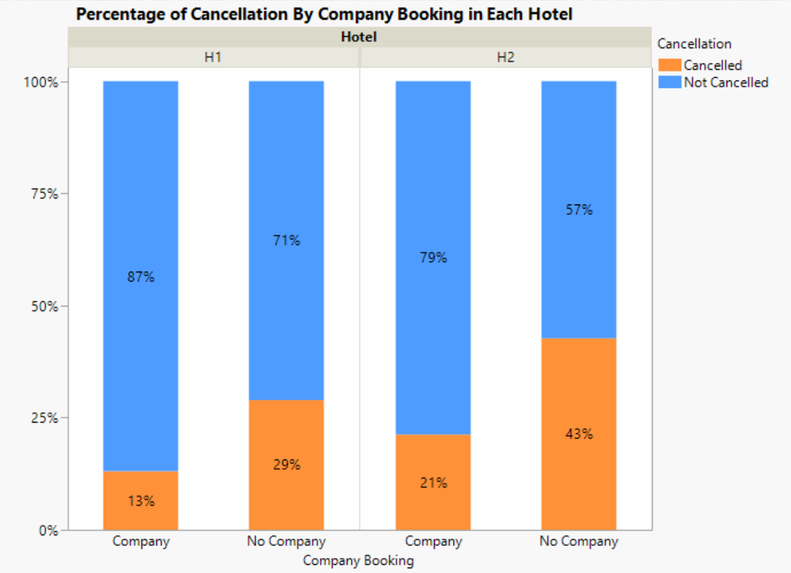
Insight: Common characteristics with higher cancellation proportion



1. Being made from TA/TO distribution channel rather than corporate, direct, or GDS channels.

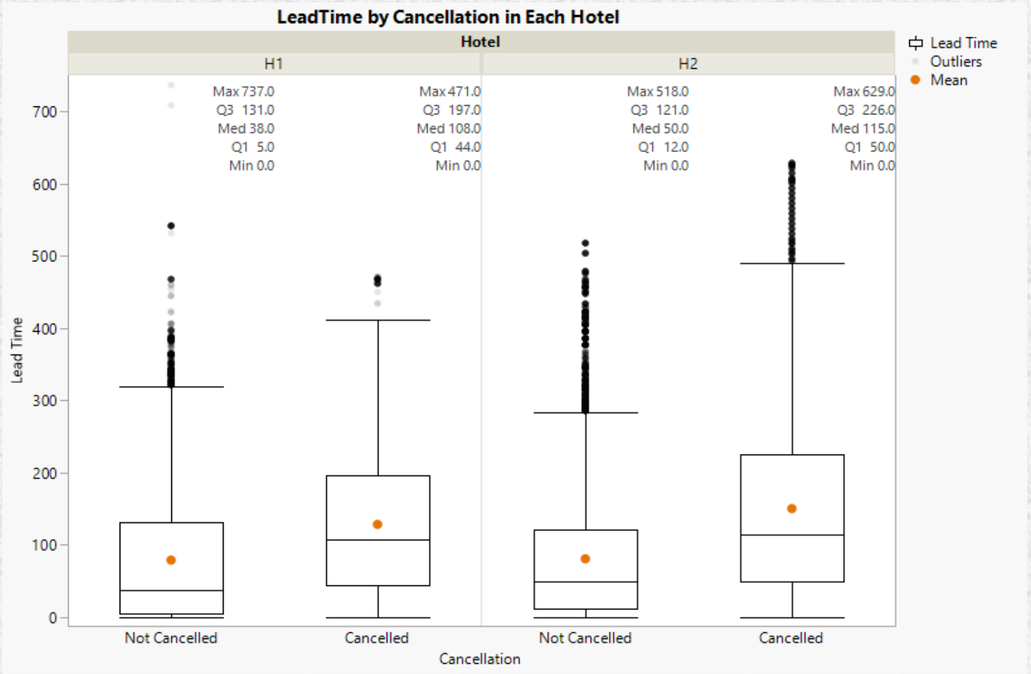
2. Having higher number of previous booking cancellations.

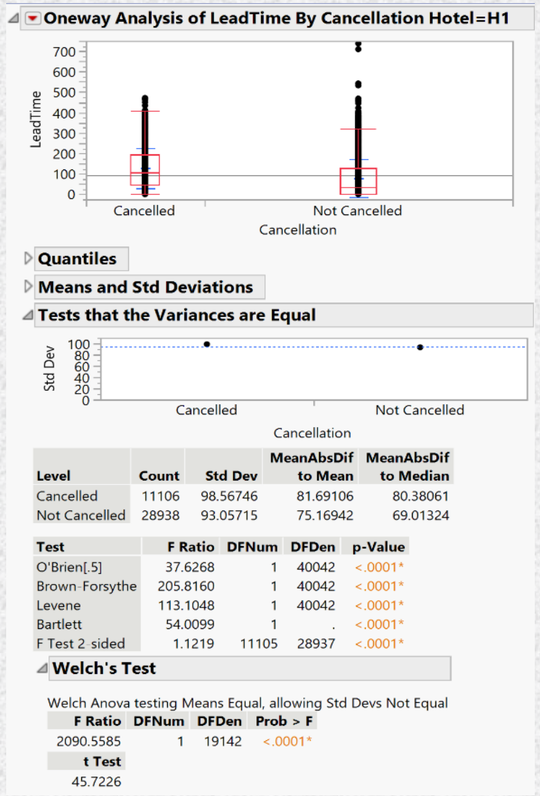




3. Not being made by or paid by a company.

Insight: Tested Hypothesis – Lead Time



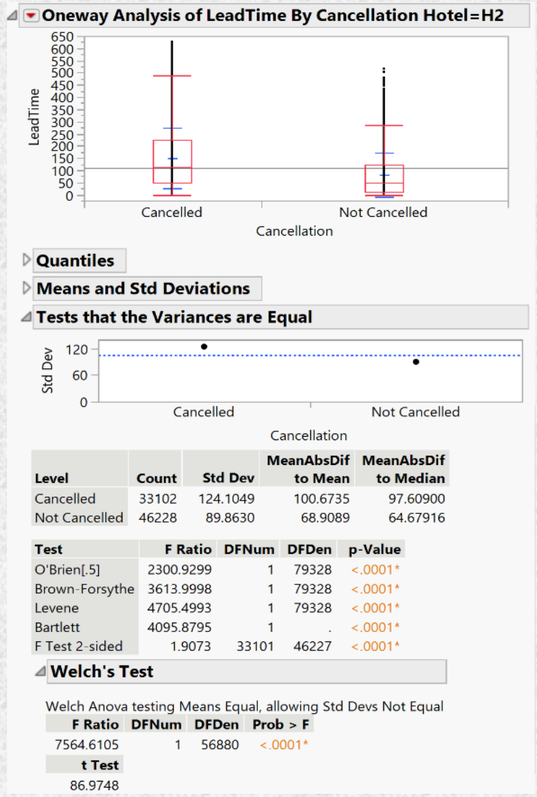


Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Cancelled	11106	98.56746	81.69106	80.38061
Not Cancelled	28938	93.05715	75.16942	69.01324

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	37.6268	1	40042	<.0001*
Brown-Forsythe	205.8160	1	40042	<.0001*
Levene	113.1048	1	40042	<.0001*
Bartlett	54.0099	1	.	<.0001*
F Test 2-sided	1.1219	11105	28937	<.0001*

F Ratio	DFNum	DFDen	Prob > F
2090.5585	1	19142	<.0001*

t Test
45.7226



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Cancelled	33102	124.1049	100.6735	97.60900
Not Cancelled	46228	89.8630	68.9089	64.67916

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	2300.9299	1	79328	<.0001*
Brown-Forsythe	3613.9998	1	79328	<.0001*
Levene	4705.4993	1	79328	<.0001*
Bartlett	4095.8795	1	.	<.0001*
F Test 2-sided	1.9073	33101	46227	<.0001*

F Ratio	DFNum	DFDen	Prob > F
7564.6105	1	56880	<.0001*

t Test
86.9748

Mean Lead Time

- H1:
  - Cancelled: **128.68 days**
  - Not Cancelled: **78.84 days**
- H2:
  - Cancelled: **150.28 days**
  - Not Cancelled: **80.70 days**

Welch's Test

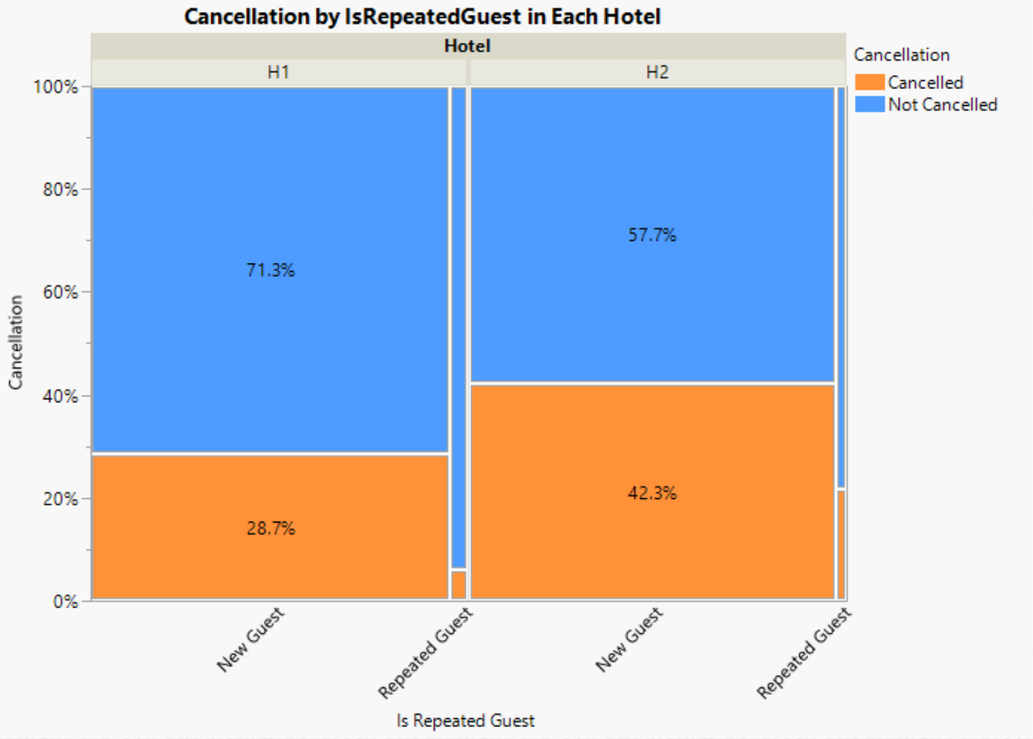
- H_0: There is no difference between the means of lead time for cancelled reservations and those that are not cancelled.
- H_1: There is a difference between the means of lead time for cancelled reservations and those that are not cancelled.

- Confidence level: **95%**
- Assumption: Not normal distribution and not equal variance
- P-value: **<.0001**

Insight: Tested Hypothesis – IsRepeatedGuest

Cancellation proportion

- H1:
  - New Guest: **28.73%**
  - Repeated Guest: **6.24%**
- H2:
  - New Guest: **42.25%**
  - Repeated Guest: **21.70%**



Contingency Analysis of Cancellation By IsRepeatedGuest Recoded Hotel=H1

Mosaic Plot

Contingency Table

IsRepeatedGuest Recoded	Count / Total % / Col % / Row %	Cancelled	Not Cancelled	Total
		Cancellation		
New Guest	Count	10995	27271	38266
	Total %	27.46	68.10	95.56
	Col %	99.00	94.24	
	Row %	28.73	71.27	
Repeated Guest	Count	111	1667	1778
	Total %	0.28	4.16	4.44
	Col %	1.00	5.76	
	Row %	6.24	93.76	
Total	Count	11106	28938	40044
	Total %	27.73	72.27	

Tests

N	DF	-LogLike	RSquare (U)
40044	1	277.94452	0.0118

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	555.889	<.0001*
Pearson	428.785	<.0001*

Contingency Analysis of Cancellation By IsRepeatedGuest Recoded Hotel=H2

Mosaic Plot

Contingency Table

IsRepeatedGuest Recoded	Count / Total % / Col % / Row %	Cancelled	Not Cancelled	Total
		Cancellation		
New Guest	Count	32661	44637	77298
	Total %	41.17	56.27	97.44
	Col %	98.67	96.56	
	Row %	42.25	57.75	
Repeated Guest	Count	441	1591	2032
	Total %	0.56	2.01	2.56
	Col %	1.33	3.44	
	Row %	21.70	78.30	
Total	Count	33102	46228	79330
	Total %	41.73	58.27	

Tests

N	DF	-LogLike	RSquare (U)
79330	1	186.05207	0.0035

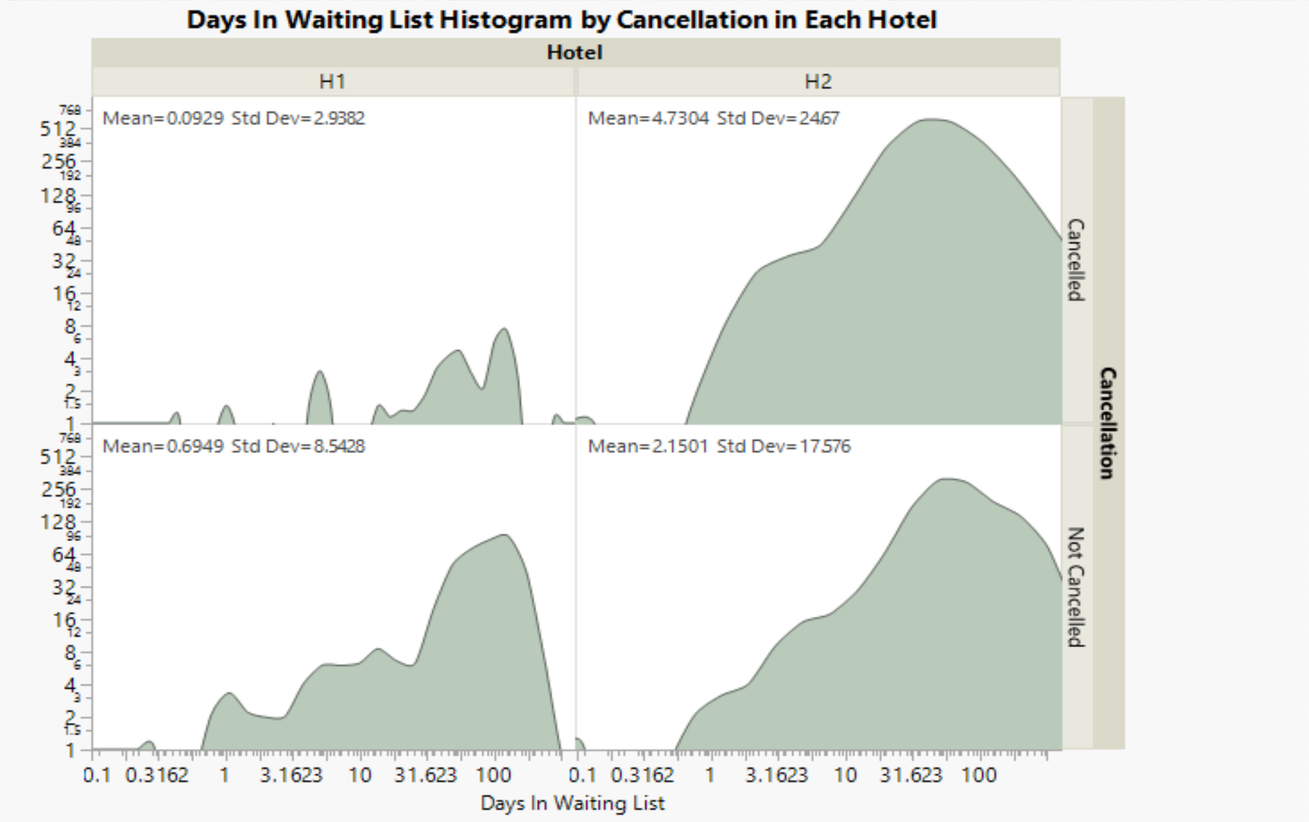
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	372.104	<.0001*
Pearson	343.890	<.0001*

Chi-Square Test

- H_0: There is no difference in the proportion of cancellation between new guest and repeated guest.
- H_1: There is a difference in the proportion of cancellation between new guest and repeated guest.

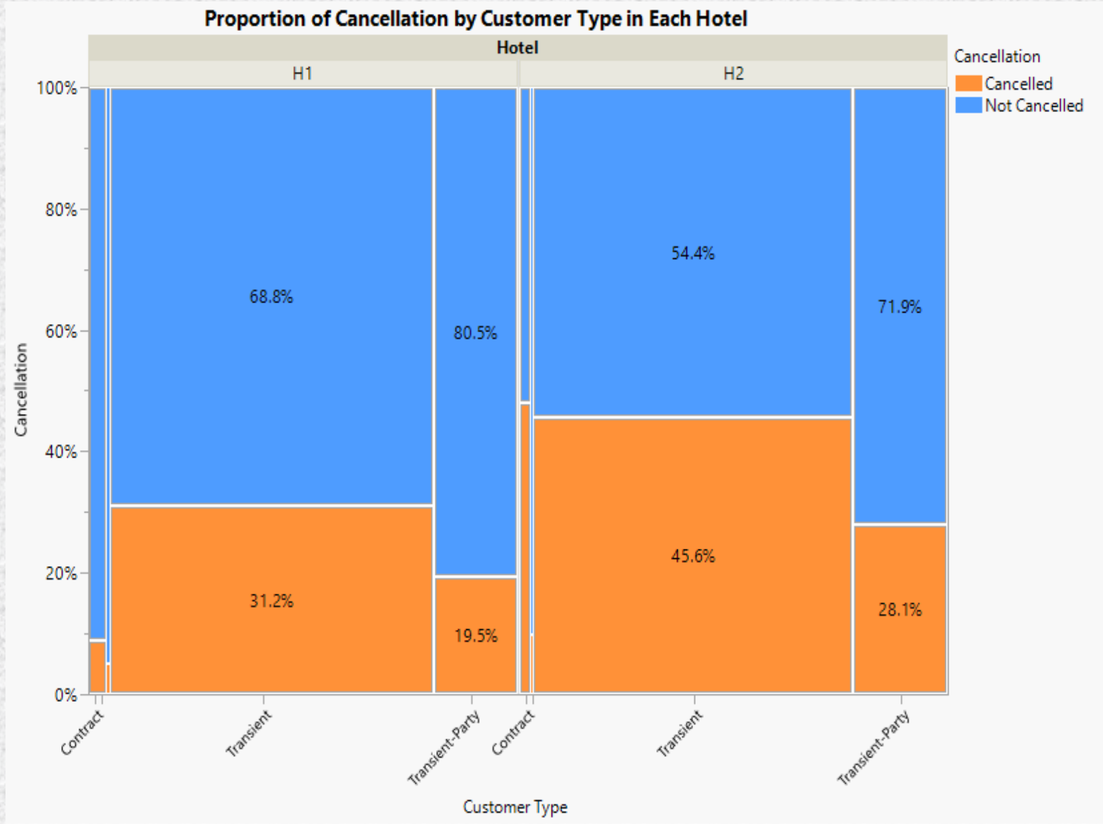
- Confidence level: **95%**
- Assumption: No group has less than 5 observations
- P-value: **<.0001**

Insight: Inconsistent characteristics with higher cancellation proportion

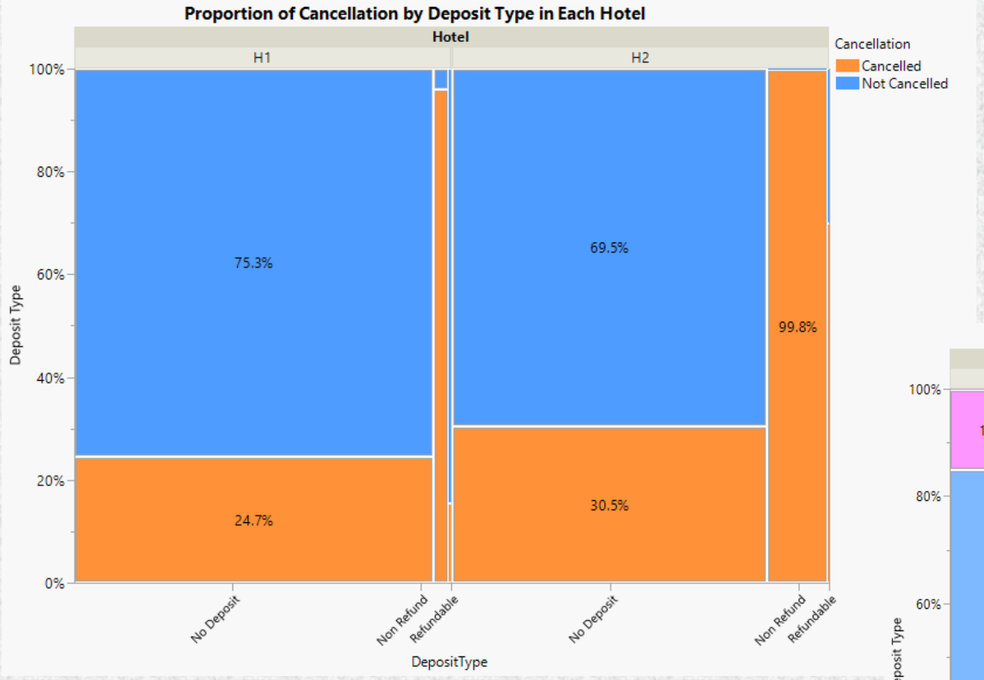


1. Higher average of days in waiting list
    - Lower cancellation in H1
    - Higher cancellation in H2

2. Customer type with higher cancellation proportion
    - Transient customers in H1
    - Contract customers in H2

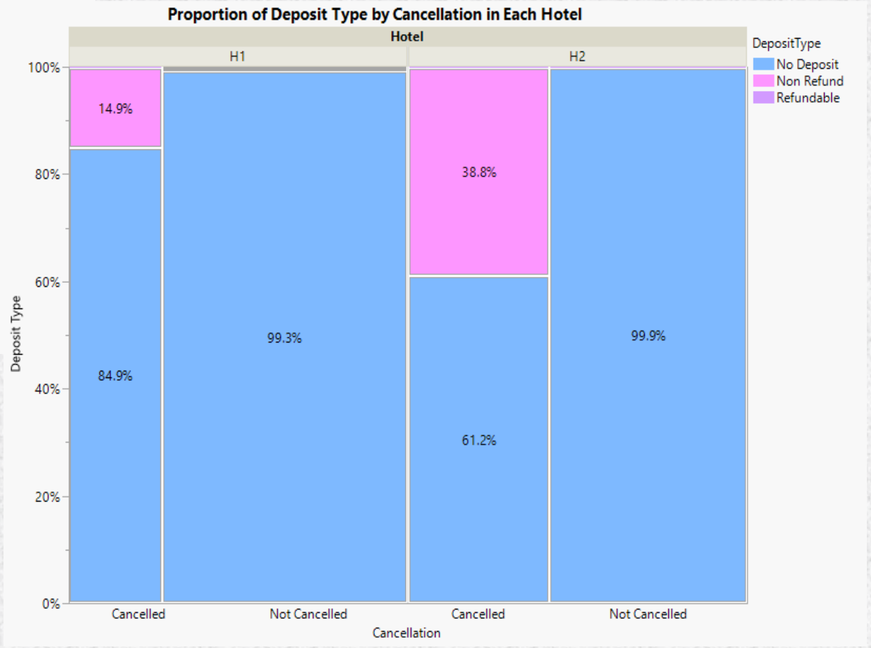


Insight: Needs further investigation – Deposit Type



Non-refundable deposit has the highest proportion of cancellation

- H1: **95.99%**
- H2: **99.81%**



Non-refundable deposit primarily exist in cancelled bookings

Possibility of logic error in categorization

- Current logic: Deposit type by checking whether there are any payments before arrival date
- Possible Loophole:
  - TA/TO booking paid in lump-sum periodically → all TA/TO bookings categorized as no deposit
  - Cancellation fee paid directly to hotel → counted as deposit

Managerial **Recommendations**

Characters to watch out for

→ HIGHER PROBABILITY of cancellation

1. Being made from TA/TO distribution channel rather than corporate, direct, or GDS channels.
2. Having higher number of previous booking cancellations.
3. Making the bookings further away from the arrival date, i.e. having a higher lead time.*
4. Being a new guest.*
5. Not being made by or paid by a company.

Data quality issues to improve

1. Consistency of data type, modelling type, and format
2. Consistency of missing value encoding
3. Data cleanliness
    - Erroneous values
    - Logic error in categorization