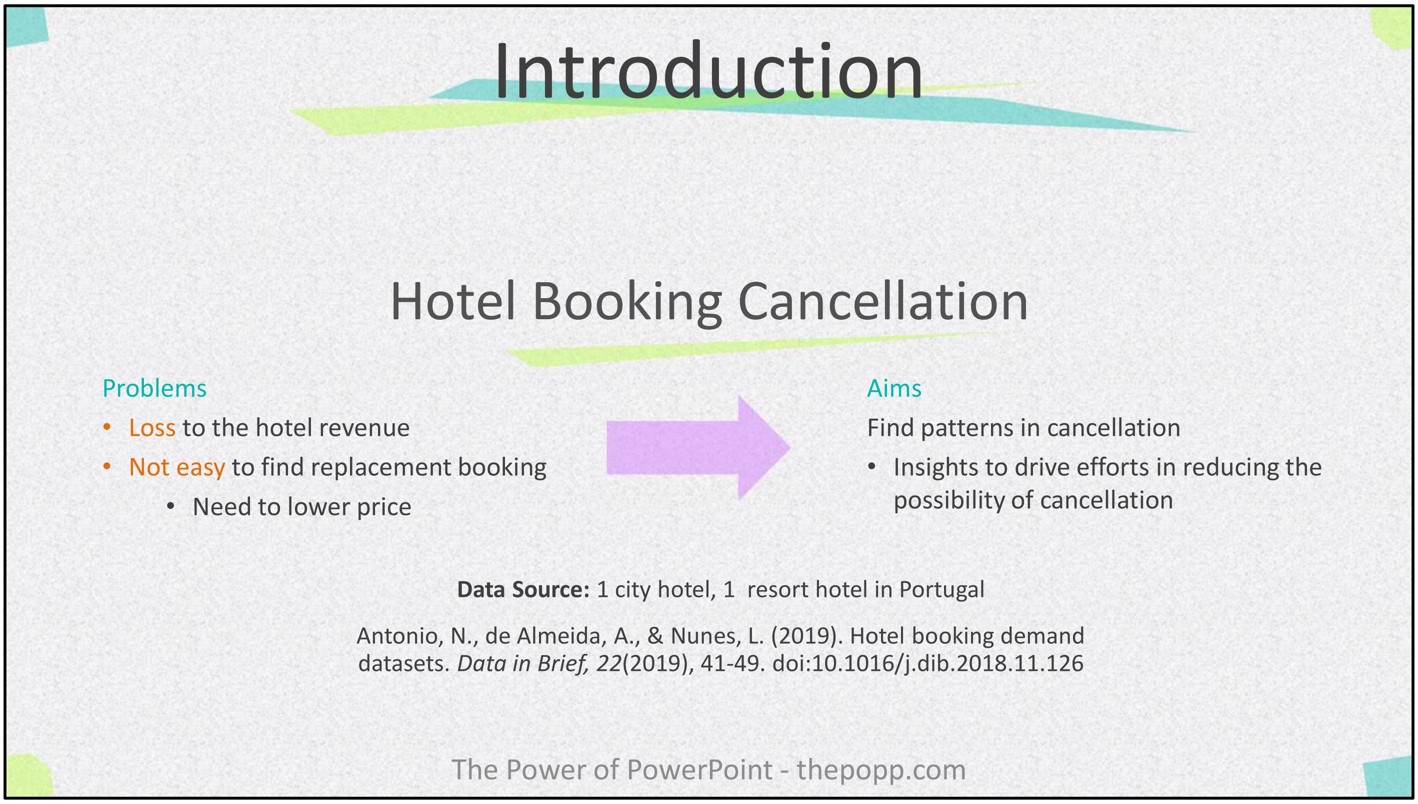
The Cost of Hotel Booking ~~Cancellation~~

ISSS602 AY2020-21 Assignment 1

Gabriella Pauline Djojosaputro

This report about “The Cost of Hotel Booking Cancellation” is done as the first assignment for ISSS602 Data Analytics Lab course in AY2020-21 Term 1.



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

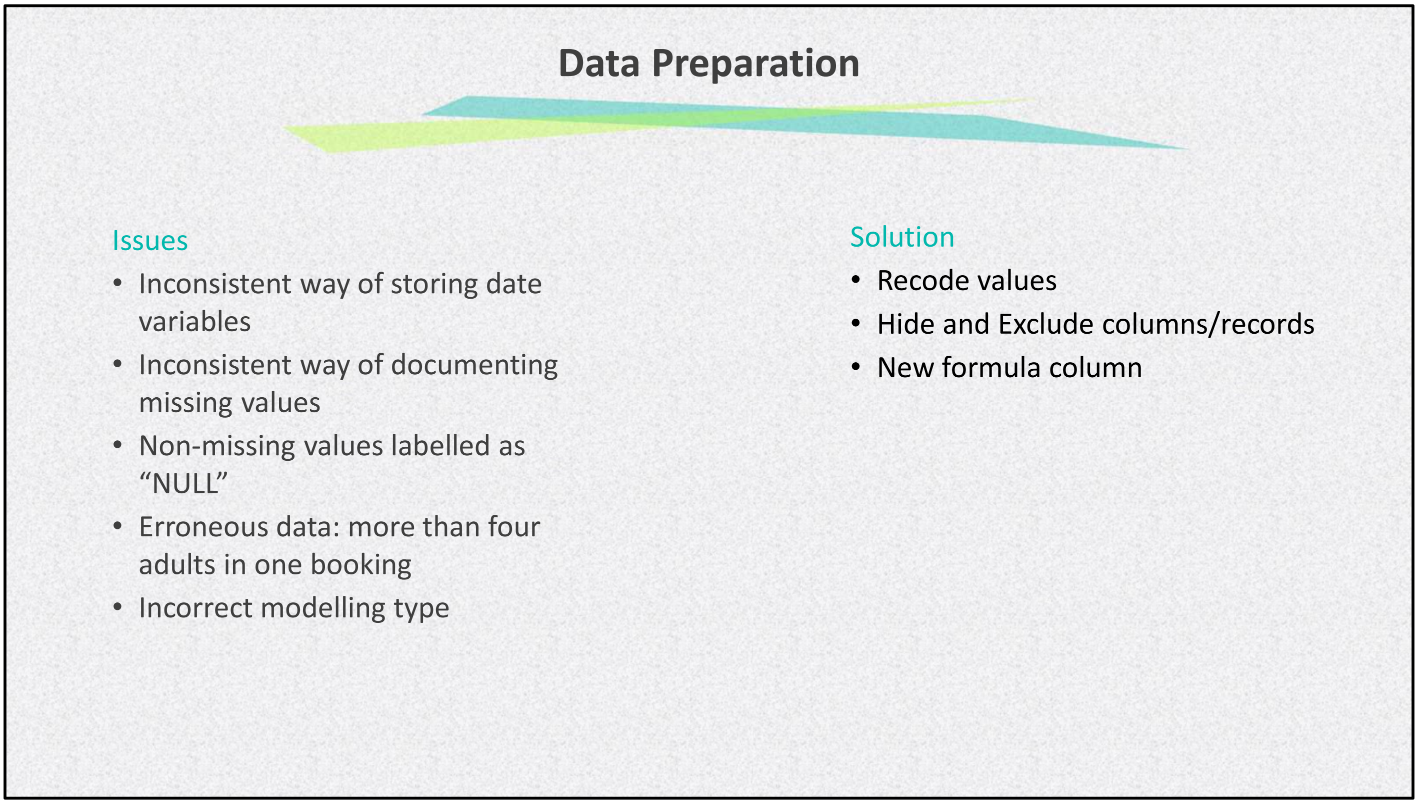
Data Source: 1 city hotel, 1 resort hotel in Portugal

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

Hotel booking cancellation is one of the challenges that affect tourism and hospitality industry. Once the reservation is cancelled, it would not be easy to find a replacement booking, especially if the cancellation is performed last-minute. Even if replacement booking can be found, usually the hotel will need to lower the price. This would lead to direct loss of the booking revenue, as well as the downstream revenue from restaurants and other facilities that may be used by the customers during the stay.

This report aims to find patterns in the cancellation so that hotel managements can use the new insights to drive efforts in reducing the possibility of cancellation.

The data source used in provided by Antonio et al. (2019) in Data in Brief, with the title Hotel booking demand datasets. The datasets contain reservations from 2 hotels in Portugal—a city hotel and a resort hotel.



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

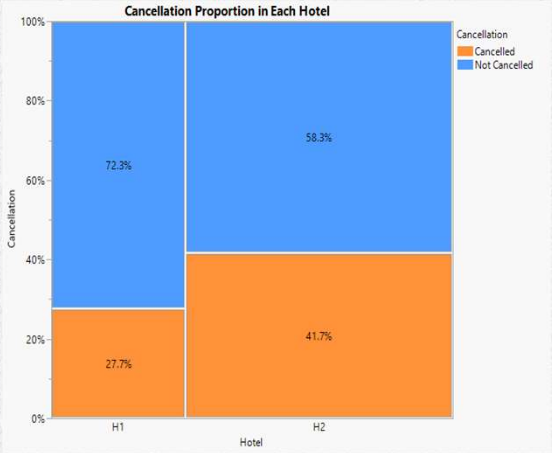
- Recode values
- Hide and Exclude columns/records
- New formula column

In preparing the data, several issues as discovered and resolved before going further into analysis.
The issues are:

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

If these issues are not resolved, the analysis can be inaccurate or hindered. In general, the steps taken to resolve them are either to recode the values, hide and exclude columns or records, create new formula columns, or some combination of them.

Insight: H2 has different cancellation rate from H1



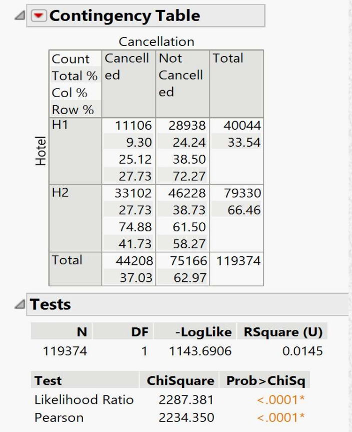
Cancellation Proportion

- 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

Analyze separately



Contingency Table

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

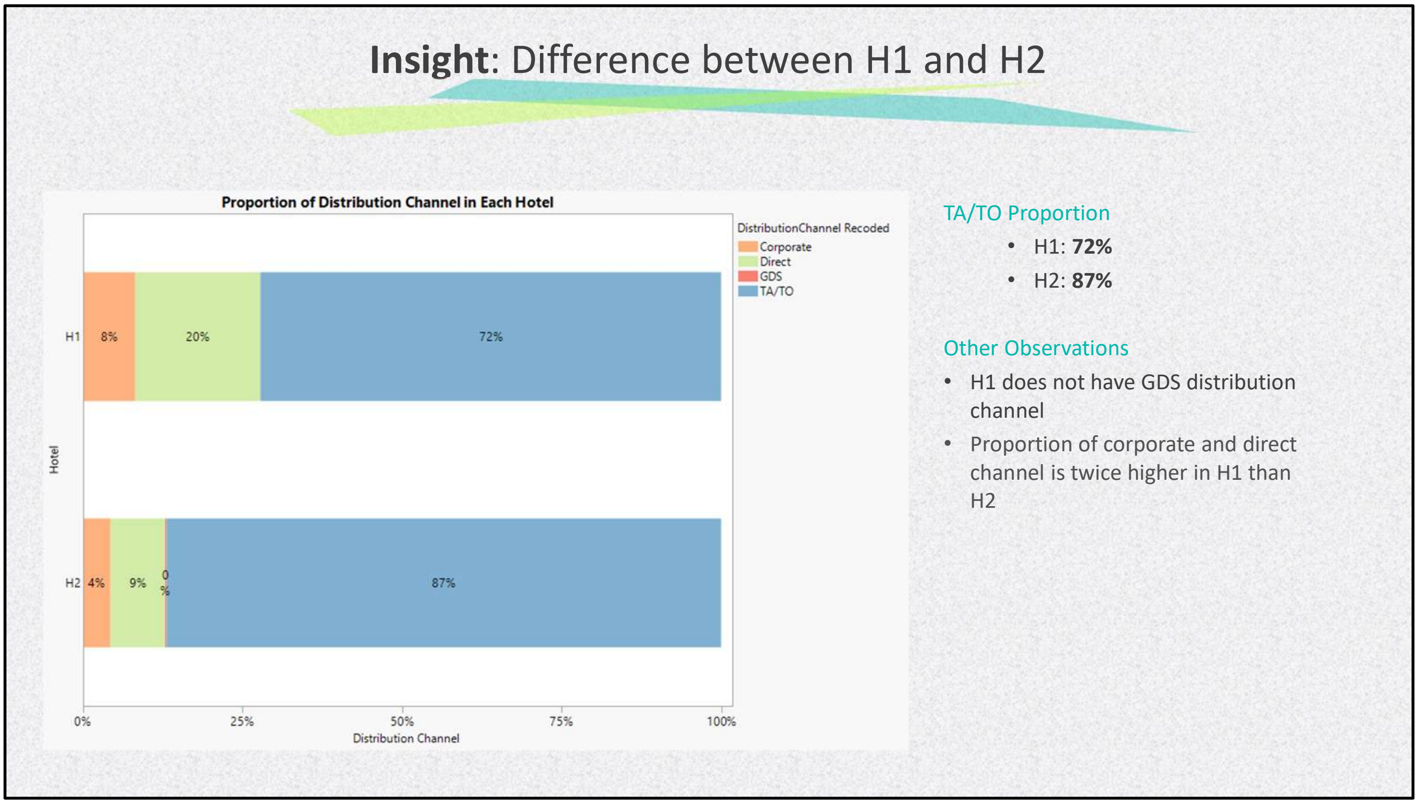
Tests

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

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

Cancellation rate in H1 is first compared to H2 to determine whether the two hotels are similar and can be analyzed together as a group. Otherwise, the analysis must be done separately on each hotel.

H1 and H2 are shown to have different cancellation proportion, with H2 having higher rate (41.73%). Chi-square is used to test the hypothesis, after the assumption that no group has less than 5 observation is verified. The p-value is less than the set confidence level (95%), so we reject the null hypothesis that “There is no difference between the proportion of hotel booking cancellations in H1 and H2”.



Insight: Difference between H1 and H2

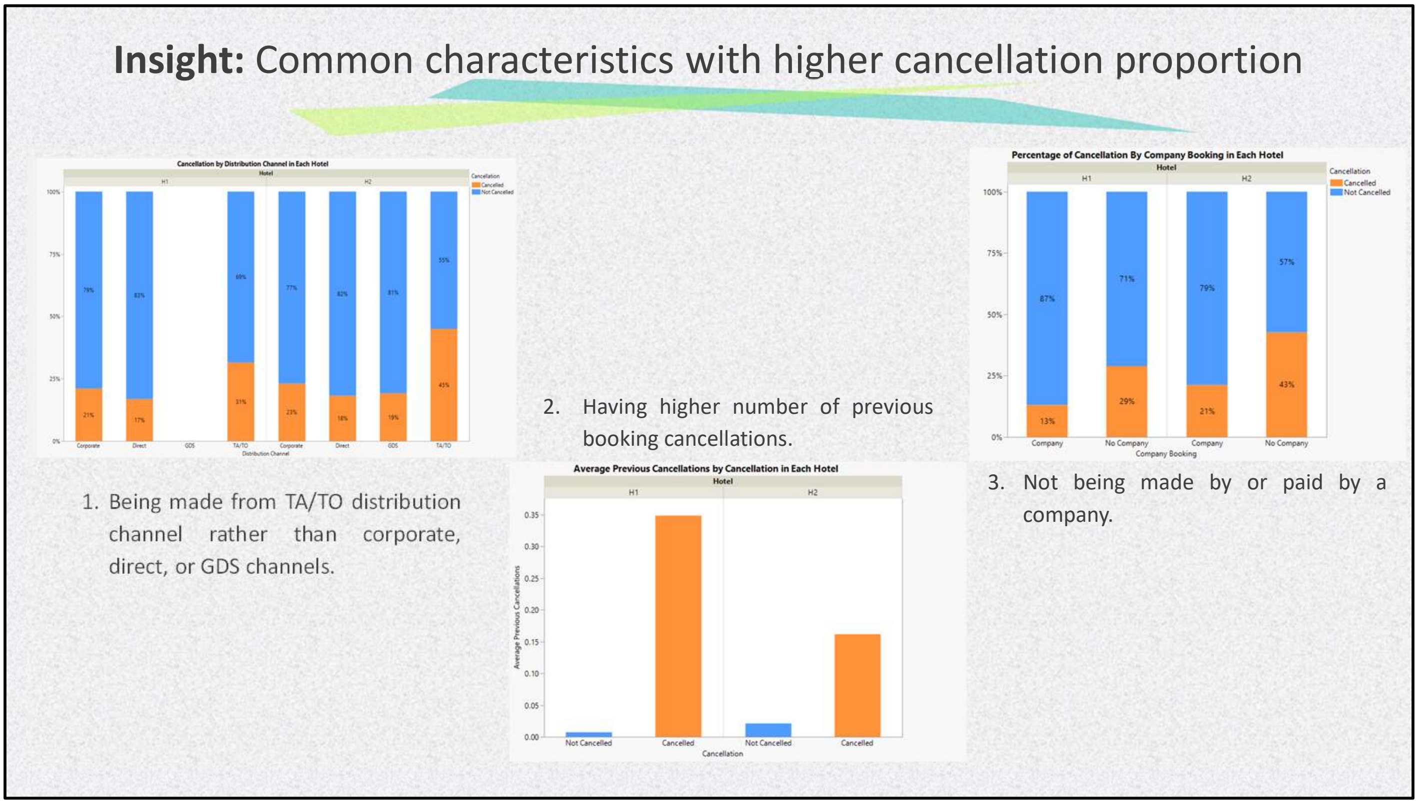
TA/TO Proportion

- 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

Some of the things that differentiate H1 and H2 is that they have different distribution channel for their bookings. H2 have higher proportion of bookings from TA/TO (87%) than H1 (72%). Other observations made on this matter is that the proportion of corporate and direct channel is twice higher in H1 than H2, and H1 does not have GDS distribution channel.

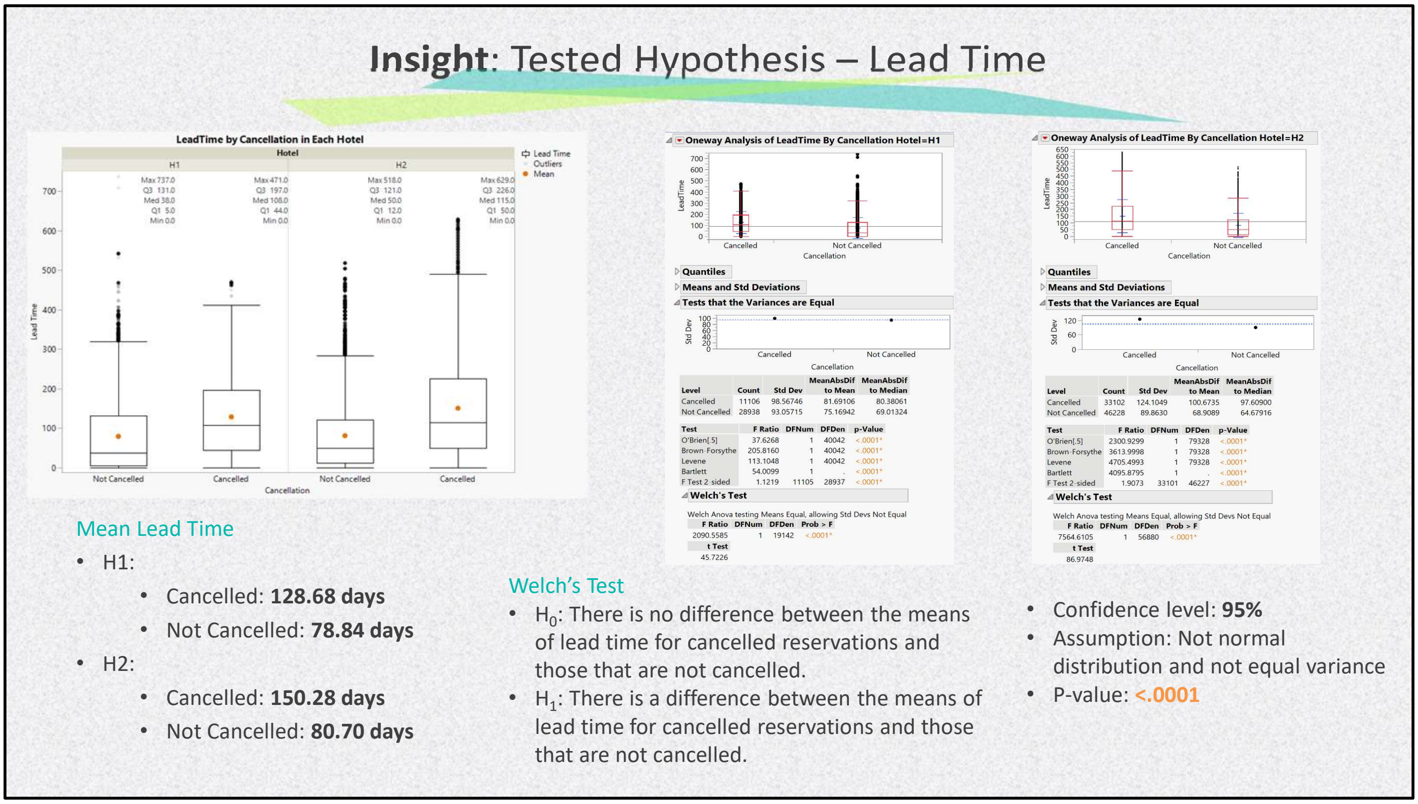


Insight: Common characteristics with higher cancellation proportion

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.

Some characteristics of guest booking consistently have higher cancellation proportion in both H1 and H2, despite being a different type of hotel in a different location in Portugal. These observed characteristics that are not yet tested are:

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

Hotel=H1

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*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

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

t Test
45.7226

Hotel=H2

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*

Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

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

t Test
86.9748

Mean Lead Time

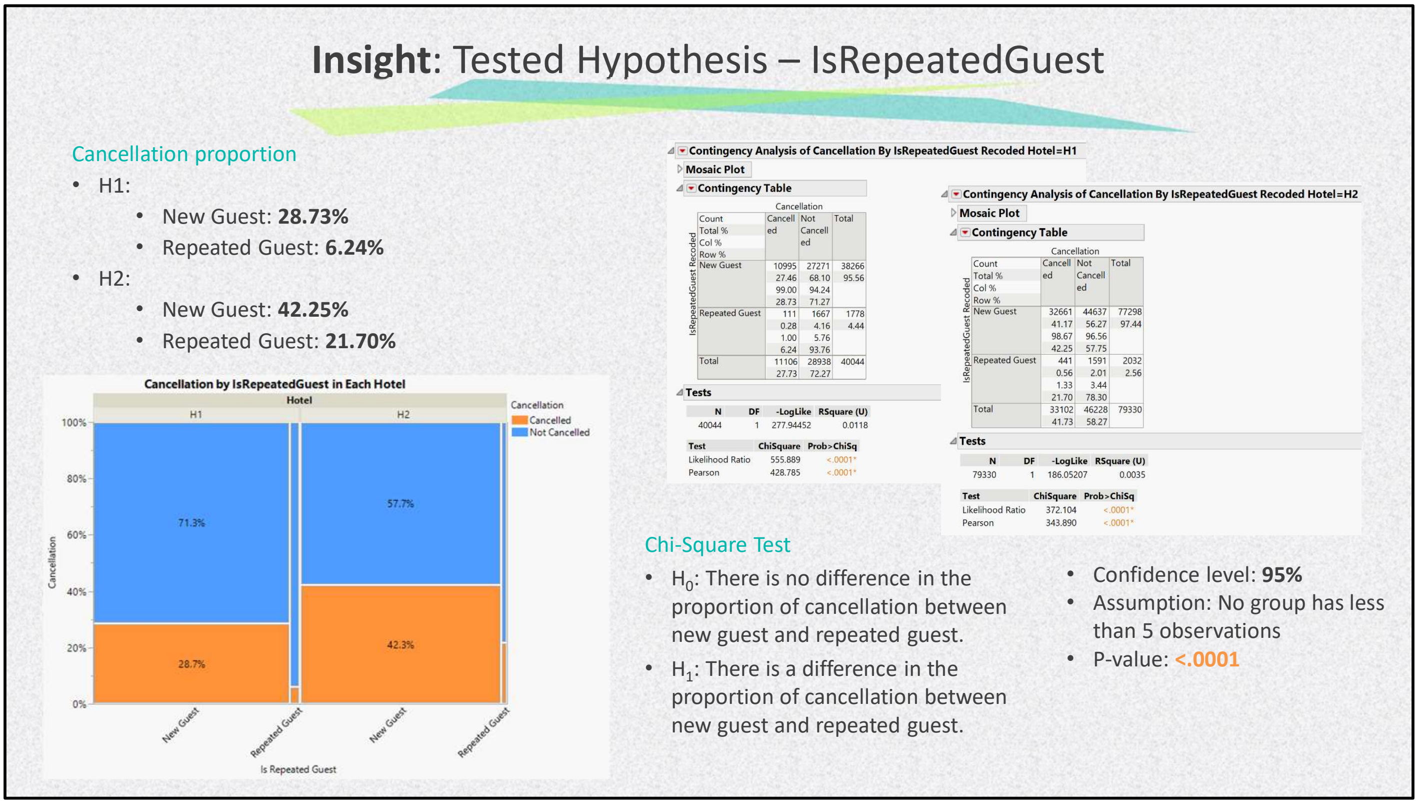
- H1:
  - Cancelled: **128.68 days**
  - Not Cancelled: **78.84 days**
- H2:
  - Cancelled: **150.28 days**
  - Not Cancelled: **80.70 days**

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**

Some observations have been tested using statistical hypothesis testing. The first one is the difference in lead time between the cancelled and non-cancelled bookings.
Cancelled bookings are observed to have a higher mean of lead time. The data does not have normal distribution and equal variance, so Welch's Test is used. The p-value of the Welch tests are less than the critical value of 0.05. This means in both hotels, there is not enough evidence to show that "There is no difference between the means of lead time for cancelled reservations and those that are not cancelled."



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

Cancellation

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

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

Cancellation

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

Tests

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

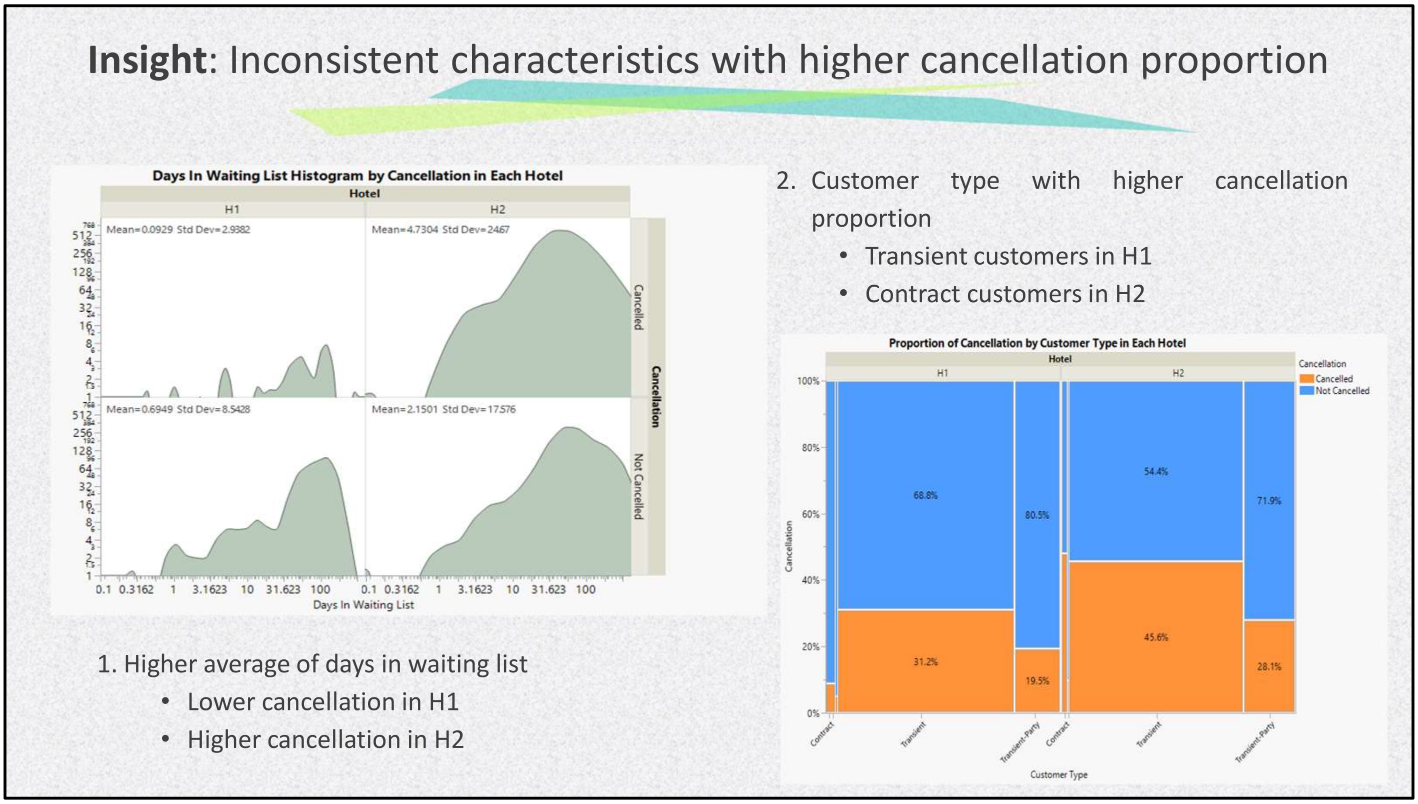
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	372.104	<.0001*
Pearson	343.890	<.0001*

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

Another insight that was tested is whether there is any difference in cancellation for new guests and repeated guests. New guests have higher cancellation rate compared to repeated guests. To test this insight, Chi-Squares's assumption that there is no group with less than 5 observations is verified, and the test obtains a result of 0.0001. For both H1 and H2, the p-value of Chi-Square test is lower than the critical value of 0.05, so we reject the null hypothesis that "There is no difference in the proportion of cancellation between new guest and repeated guest."



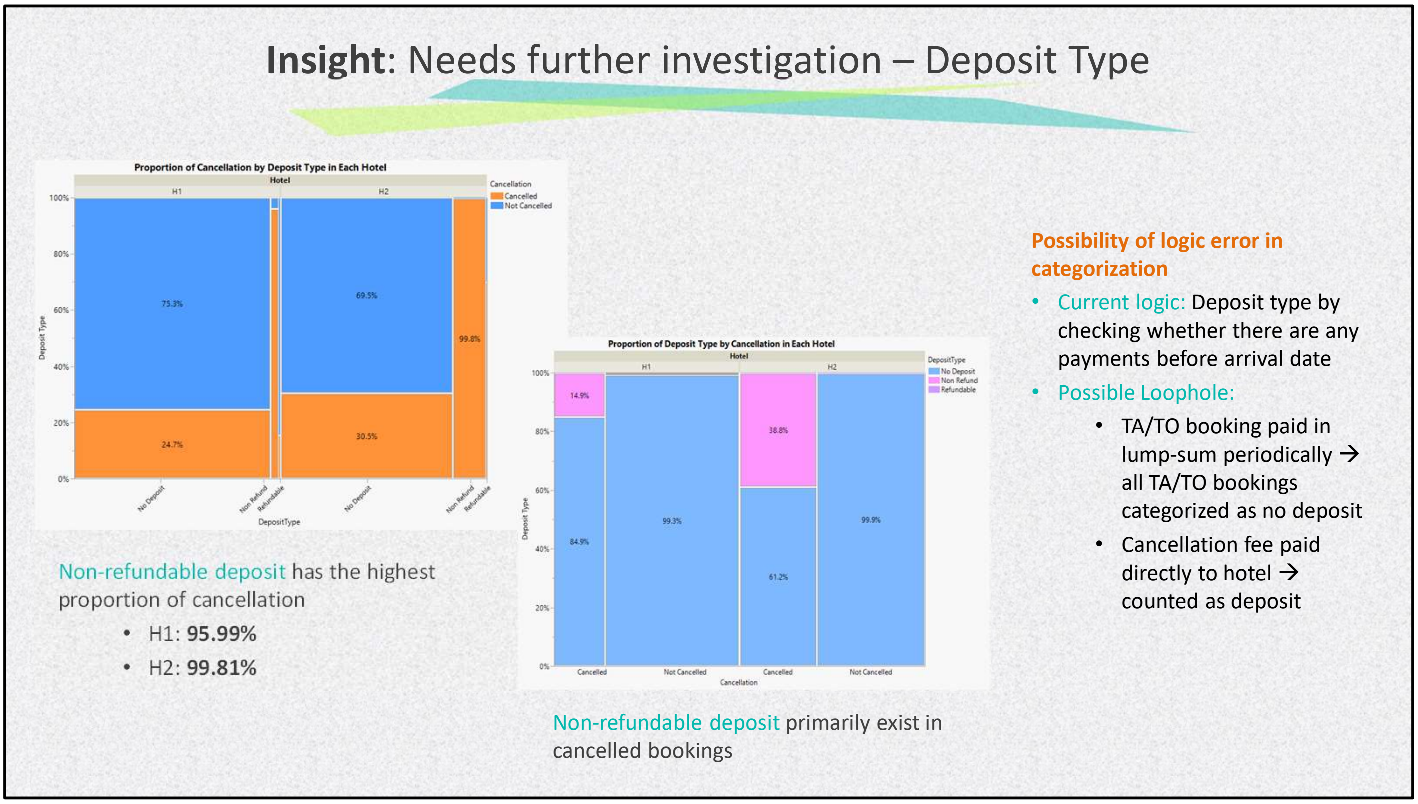
Insight: Inconsistent characteristics with higher cancellation proportion

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

Some insights are found to be inconsistent between H1 and H2.

1. Cancelled bookings have higher average of days in waiting list than non-cancelled bookings in H2. However, in H1, non-cancelled bookings have higher average of days in waiting list.
2. The type of customers with the highest proportion of cancellation is different for H1 (Transient customers) and H2 (Contract customers)



Insight: Needs further investigation – Deposit Type

Non-refundable deposit has the highest proportion of cancellation

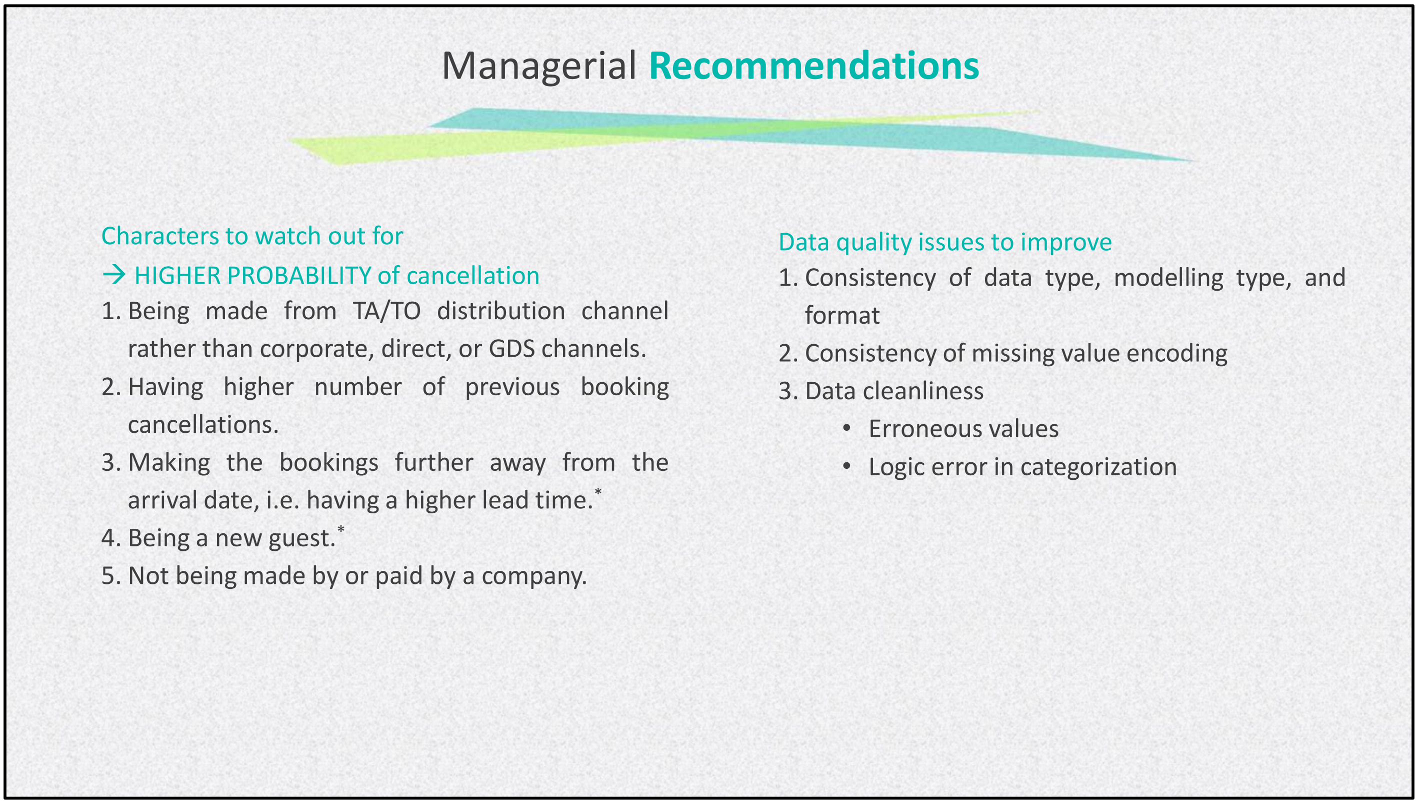
- H1: **95.99%**
- H2: **99.81%**

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

There is a suspicious insight regarding the deposit type that needs further investigation. Non-refundable deposit has more than 90% on cancellation, which does not seem to make sense. There is a possibility that this is due to logic error in the categorization of deposit type, which is currently done by determining whether there are any payments made before the arrival date. A possible loophole in this logic is if the payments for booking via TA/TO are only paid in lump sum periodically, while the cancellation fee will be directly paid to the hotel, then TA/TO bookings will always be no deposit type and only categorized as non-refundable when they are cancelled.



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

The recommendations regarding the hotel booking are to look out for certain characteristics that have 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 also need to be improved, which are:

1. Consistency of data type, modelling type, and format
2. Consistency of missing value encoding
3. Data cleanliness, to eliminate erroneous values and logic error in categorization