HOTEL CANCELLATION RATES – A STUDY

EVIDENCE FROM PORTUGAL

NGUYEN NHI THAI AN

Contents

1. Objective

This study examined aspects of customer relationship management (CRM) data in relation to hotel cancellation rates, using JMP Pro 15. Possible trends were obtained, from which business recommendations may be further suggested.

2. Data

2.1. Data Source & Preparation

The dataset is obtained from Antonio, Almeida & Nunes's "Hotel Booking Demand Datasets" (2019), in the form of two data tables H1 and H2, each representing bookings from a resort and city hotel, comprising 40,060 and 79,330 observations respectively. The bookings' arrival dates span 1/7/2015 to 31/8/2017 i.e. 26 months.

The data tables have the same structure, number of variables, and column orderings. Multiple File Import function was used to append H2 below H1 by rows (Figure 1), retaining the information of their filenames, from which the type of hotel (resort or city hotel) was recoded (Figure 2).

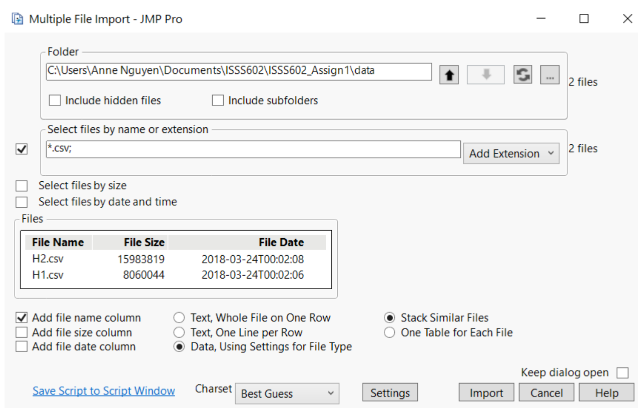


Figure 1. Multiple File Import to Combine H1 and H2 Data Tables

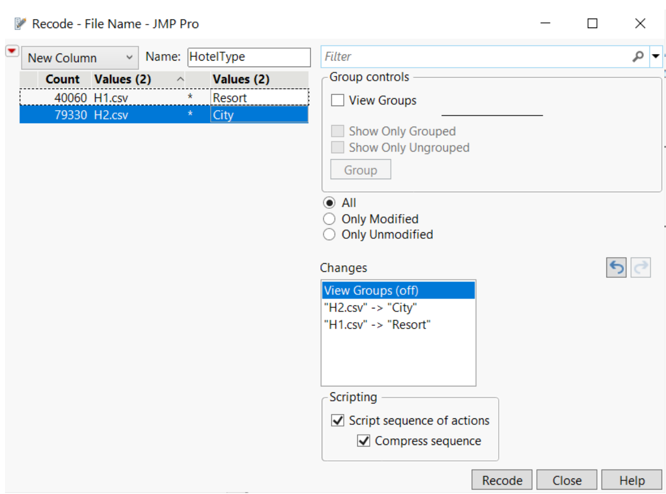


Figure 2. Recoding done to create 'HotelType' variable

2.2. Data Quality Issues & Mitigation

Multiple data quality issues were unearthed and largely classified into the following major categories, requiring additional mitigation measures.

2.2.1. Missing data

Despite assurance from documentation, inspection revealed there was missing data:

- There were 4 missing observations for 'Children' (Table 1).

Table 1. Summary statistics for 'Children' variable

Summary Statistics

1 Columns Clear Select Distribution

Columns	N	N Missing	Min	Max	Mean	Std Dev
Children	119386	4	0	10	0.1038899033	0.3985614448

- There were a number of 'Undefined' values for 'MarketSegment' (Table 2a) and 'DistributionChannel' (Table 2b), total 5 observations. Their meaning is ambiguous, and 4 of them are also the same 4 observations with missing 'Children' data.

Table 2a. Tabulated counts by 'MarketSegment'

	MarketSegment	N(MarketSegment)
1	Aviation	237
2	Complementary	743
3	Corporate	5295
4	Direct	12606
5	Groups	19811
6	Offline TA/TO	24219
7	Online TA	56477
8	Undefined	2

Table 2b. Tabulated counts by 'DistributionChannel'

	DistributionChannel	N(DistributionChannel)
1	Corporate	6677
2	Direct	14645
3	GDS	193
4	TA/TO	97870
5	Undefined	5

These were hidden and excluded from the dataset since the number is small.

2.2.2. Incorrect Modeling or Formatting

The dataset upon import without additional modifications suffered from several incorrect modeling or formatting issues:

- 'ReservationStatusDate' was not formatted as date. These required changing into numeric data type, continuous modeling type, and date format.
- Arrival dates were segregated into D/M/Y constituents. 'ArrivalDateYear', 'ArrivalDateMonthNumeric' (recoded from 'ArrivaldateMonth' as in Figure 3), 'ArrivalDateDayOfMonth' were aggregated into a new column called 'ArrivalDate' (Figure 4).

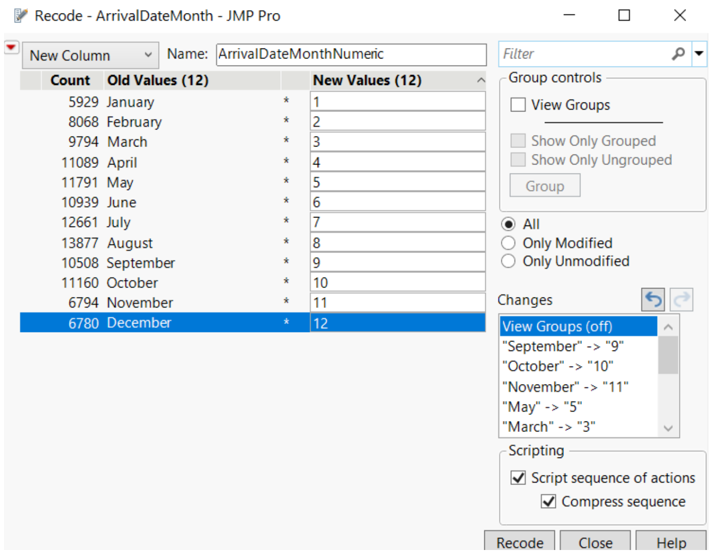


Figure 3. Recoding done for ‘ArrivalDateMonth’ to create ‘ArrivalDateMonthNumeric’

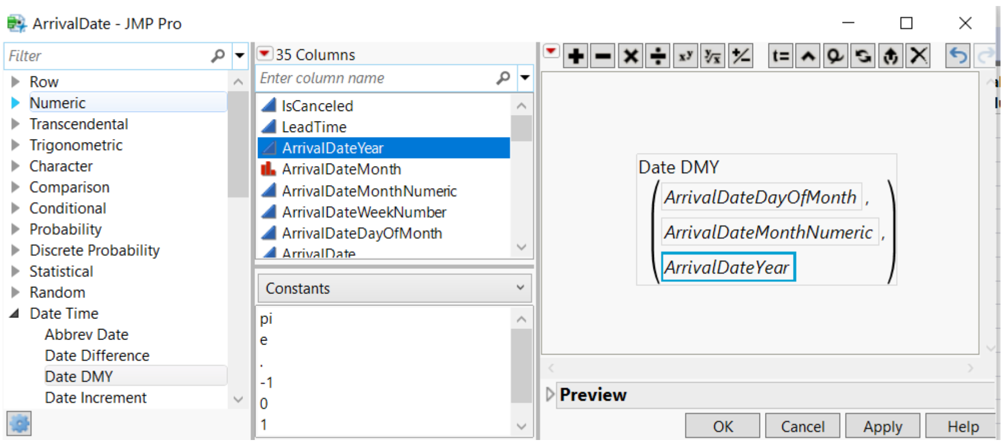


Figure 4. Formula used to create ‘ArrivalDate’

- ‘Children’ was in character data type instead of numeric.
- Variables such as ‘IsCanceled’ and ‘IsRepeatedGuest’ are encodings and should rightly be interpreted as categorical.

2.2.3. Presence of Long-Tailed Distributions and/or Extreme Values

The distribution for several variables included very low counts at very large values:

- There were 16 observations with ‘Adults’ > 4, 1 where ‘Children’ > 3, and 2 where ‘Babies’ > 2 (Figure 5). These should be excluded since they constituted an extremely small portion of the dataset.

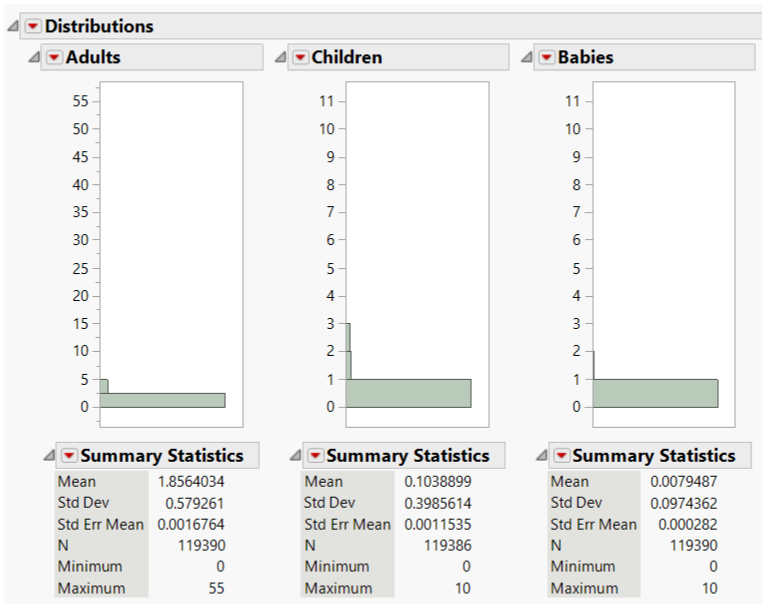


Figure 5. Distribution of ‘Adults’, ‘Children’, and ‘Babies’

- Other extreme outliers (e.g. ‘ADR’ = 5,400, ‘LeadTime’ > 700) were also excluded.
- There were numeric variables where potential outliers constituted a more significant portion of the dataset and the range was very large: ‘LeadTime’ with 2977 potential outliers – (~2.5%) or ‘StayInNightsTotal’ with 5219 potential outliers (4.4%), likewise for ‘PreviousCancellations’, ‘PreviousBookingsCanceled’, ‘DaysInWaitingList’ (Figure 6). These outliers evaluated by Tukey’s rule constituted more than 1% of the dataset, and were thus left as is. Since they are numeric, investigations into possible trends would still be relatively robust.

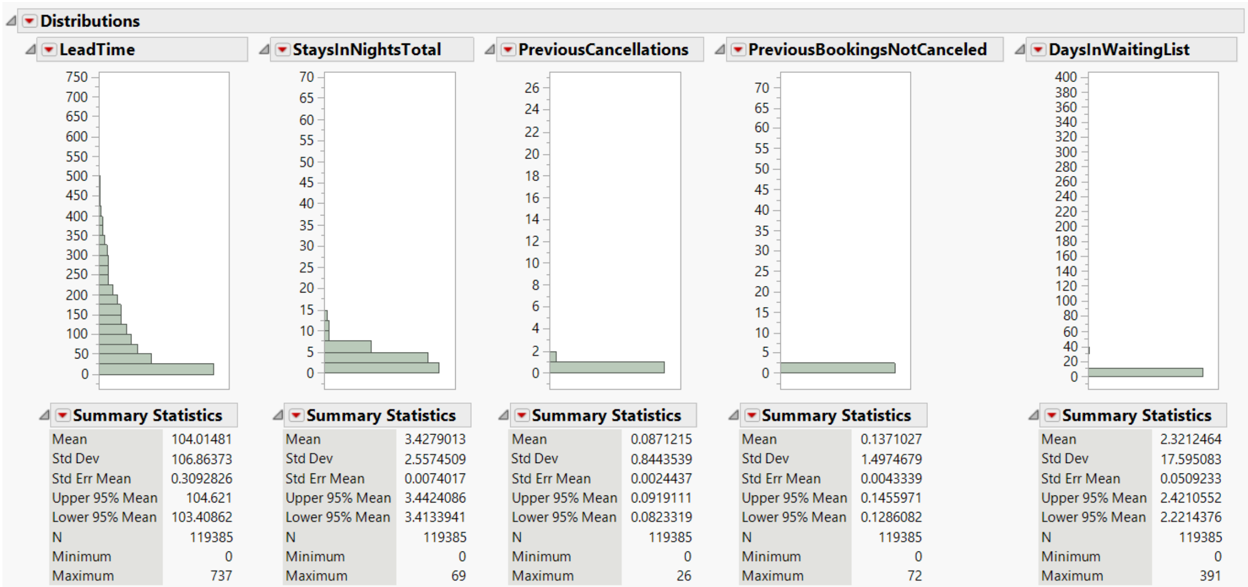


Figure 6. Distributions of 'LeadTime', 'StaysInNightsTotal', 'PreviousCancellations', 'PreviousBookingsNotCanceled', and 'DaysInWaitingList'

- Some categorical data had too many levels, which would make interpretation of potential relationships difficult and not particularly insightful (Table 3). 'Country' has 178 unique values. 'Agent' has 334 unique values, with 'NULL' having the 2nd highest frequency. 'Company's 'NULL; values which make up 94% of the dataset, and none of the remaining values exceeds 1%; this was recoded into 'IsCompany' for better clarity.

Table 3. Distribution of ‘Country’, ‘Agent’, and ‘Company’ categorical variables

Distributions

Country — Frequencies

Level	Count	Prob
PRT	48585	0.40696
GBR	12129	0.10160
FRA	10415	0.08724
ESP	8568	0.07177
DEU	7287	0.06104
ITA	3766	0.03155
IRL	3375	0.02827
BEL	2342	0.01962
BRA	2224	0.01863
NLD	2104	0.01762
USA	2097	0.01757
CHE	1730	0.01449
CN	1279	0.01071
AUT	1263	0.01058
SWE	1024	0.00858
CHN	999	0.00837
POL	919	0.00770
ISR	669	0.00560
RUS	632	0.00529
NOR	607	0.00508
ROU	500	0.00419
NULL	488	0.00409
FIN	447	0.00374
DNK	435	0.00364
AUS	426	0.00357

N Missing 0
178 Levels

Agent — Frequencies

Level	Count	Prob
9	31960	0.26771
NULL	16337	0.13684
240	13922	0.11661
1	7191	0.06023
14	3639	0.03048
7	3539	0.02964
6	3290	0.02756
250	2870	0.02404
241	1721	0.01442
28	1666	0.01395
8	1514	0.01268
3	1336	0.01119
37	1230	0.01030
19	1061	0.00889
40	1039	0.00870
314	927	0.00776
21	875	0.00733
229	786	0.00658
242	780	0.00653
83	696	0.00583
29	683	0.00572
171	607	0.00508
12	578	0.00484
85	554	0.00464
20	540	0.00452

N Missing 0
334 Levels

Company — Frequencies

Level	Count	Prob
NULL	112588	0.94307
6	1	0.00001
8	1	0.00001
9	37	0.00031
10	1	0.00001
11	1	0.00001
12	14	0.00012
14	9	0.00008
16	5	0.00004
18	1	0.00001
20	50	0.00042
22	6	0.00005
28	5	0.00004
29	2	0.00002
31	17	0.00014
32	1	0.00001
34	8	0.00007
35	1	0.00001
37	10	0.00008
38	51	0.00043
39	8	0.00007
40	927	0.00776
42	5	0.00004
43	29	0.00024
45	250	0.00209

N Missing 0
353 Levels

2.2.4. Seemingly “Erroneous” Data

There were 403 observations where ‘Adults’ is 0. Upon creating ‘StaysInNightsTotal’, there were 715 observations where ‘StaysInNightsTotal’ is 0. In total there were 1048 such observations. It is strange that bookings were made with 0 adults, or 0 nights. Further, a significant portion was made up of ‘ReservationStatus’ being ‘Check-out’ (Table 4), corresponding to the guests completing their stay i.e. this data is accurate and not caused by the hotel’s inability to update guests’ information until check-in. One possible explanation provided is that day guests may make 1-day booking without staying overnights. There was no explanation that could be garnered for observations where the number of ‘Adults’ is 0 however. For current analysis, these 1048 observations make up only 0.88% of the dataset, thus they would be excluded.

Table 4. Tabulated ‘ReservationStatus’ for bookings where either ‘StaysInNightsTotal’ or ‘Adults’ is 0

ReservationStatus	% of Total	N
Canceled	10.97%	115
Check-Out	87.40%	916
No-Show	1.62%	17

2.2.5. Data Redundancy

'ReservationStatus' is perfectly mapped to 'IsCanceled' (Table 5) while providing a lower level of data granularity. 'IsCanceled' is thus redundant, given that you can derive the same information from 'ReservationStatus', and one would expect 'IsCanceled' would be an intermediate variable.

Table 5. Tabulated data of 'ReservationStatus' vs. 'IsCanceled'

	IsCanceled	
ReservationStatus	**0**	**1**
Canceled	0	42882
Check-Out	74247	0
No-Show	0	1189

2.2.6. Incomplete Pricing Data

Data related to pricing would be highly relevant to the analysis of cancellation rates. While 'DepositType' and 'ADR' are included, there is no additional data included to shed light on the size of this deposit. Hypothetically, the greater the commitment to the booking placed, the less likely that booking would be cancelled, so the omission of this pertinent piece of data from the dataset is regrettable.

2.3. Additional Data Wrangling

Apart from additional variable creation outlined above, the following intermediate variables were created to assist in the analysis:

- 'DeltaReservationStatusArrival': created by taking the date difference between 'ArrivalDate' and 'ReservationStatusDate', with the interval name being 'Day' (Figure 7).

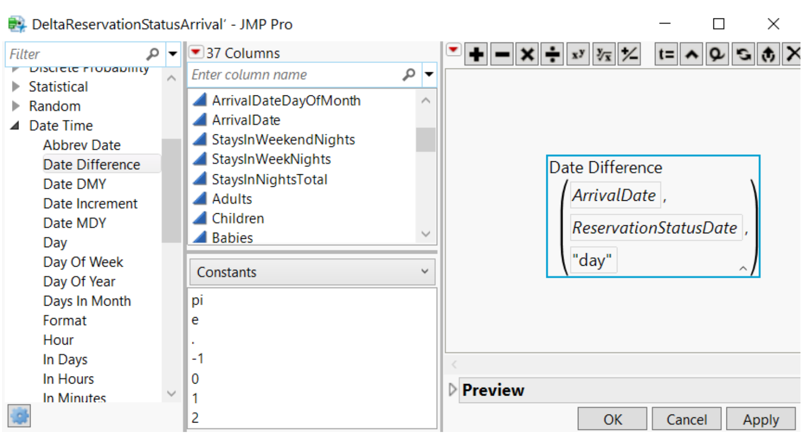


Figure 7. Formula used to calculate 'DeltaReservationStatusArrival'

- 'StaysInNightsTotal': created by summing up 'StaysInWeekendNights' and 'StaysInWeekNights' (Figure 8).

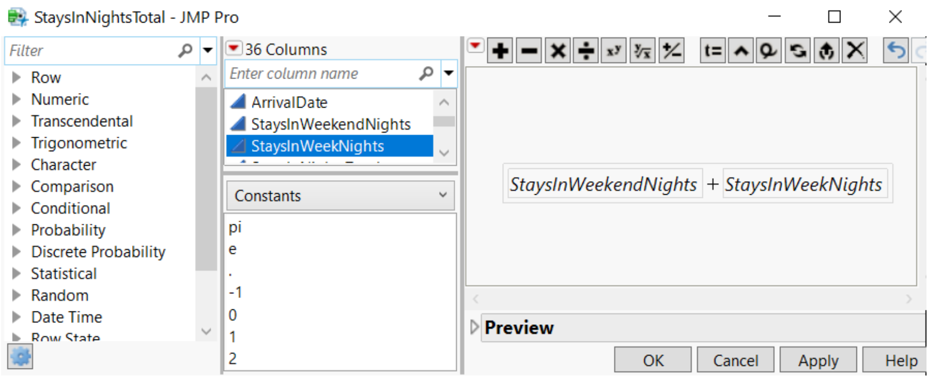


Figure 8. Formula used to calculate 'StaysInNightsTotal'

3. Data analysis

Interactive exploratory data analysis techniques were employed to explore possible trends in the data set. Ten significant trends are outlined below.

3.1. Key Insights

3.1.1. Seasonality by Hotel Type

To explore variations in booking volumes throughout the year, a bar chart of booking volume for each month scaled to percentage of total bookings for each hotel type subgroup was produced (Figure 9). Since the dataset spans 26 months, the observations from 1/7/2017 to 31/8/2017 were excluded to avoid inflating the contributions of July and August.

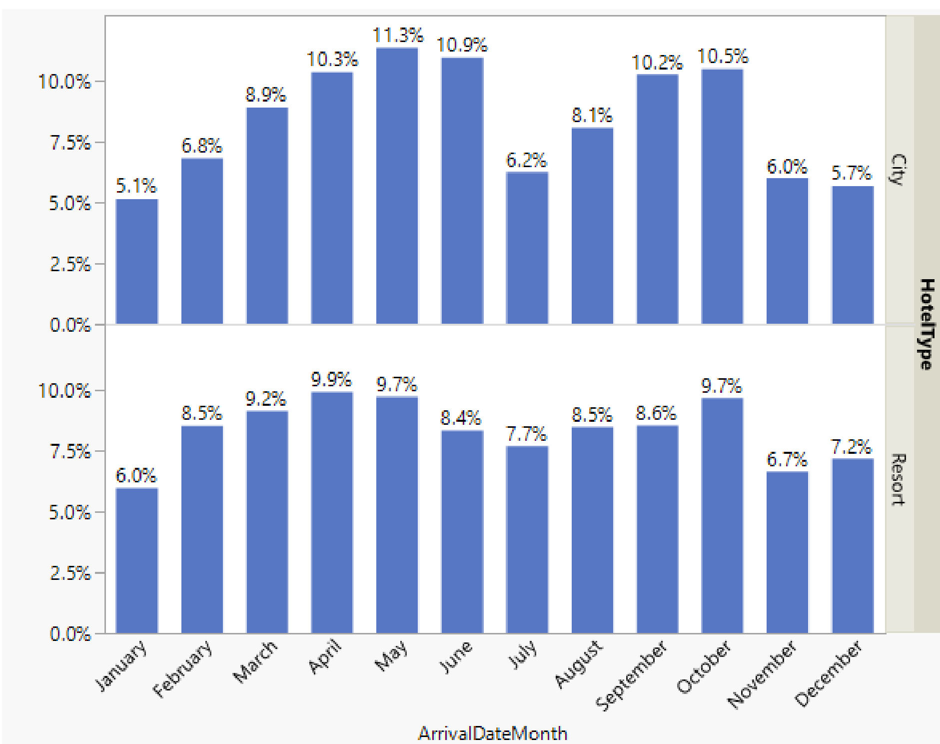


Figure 9. Standardized booking volume in arrival month, by hotel type

There was a definite seasonality effect to the volume of bookings throughout the year. Bookings were lowest around winter, picking up in spring around April or May, trending down slightly during the summer months and picking up again towards fall around October. This seasonality effect is also noted to be stronger for the city hotel compared to the resort.

3.1.2. Cancellation Rates by Hotel Type

To explore variations in cancellation rates across different hotel types, a mosaic plot of IsCanceled against HotelType was produced (Figure 10).

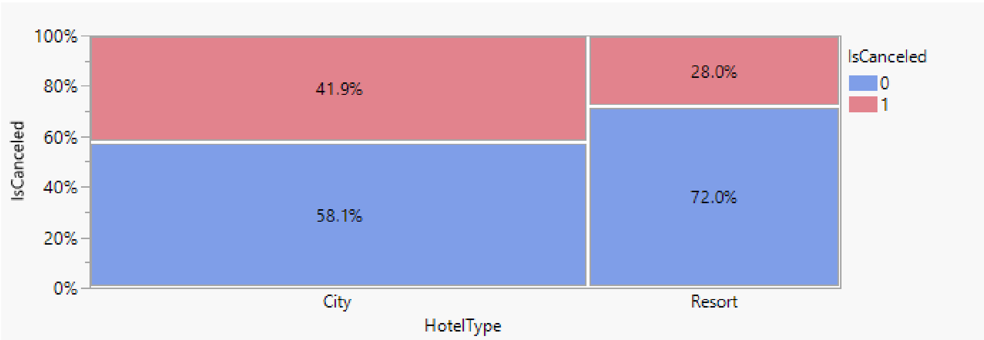


Figure 10. Mosaic plot of ‘IsCanceled’ against ‘HotelType’

The mosaic plot standardizes the cancellation rates for each subgroup of hotel type, and shows that the city hotel experienced a definitively greater rate of cancellation compared to the resort. This could be due to various reasons: the type of clientele that would typically go for a city hotel may be more ready to cancel, or competition in cities being more intensive may add more incentives for cancellation in search of better options.

3.1.3. Cancellation Rates by Market Segment

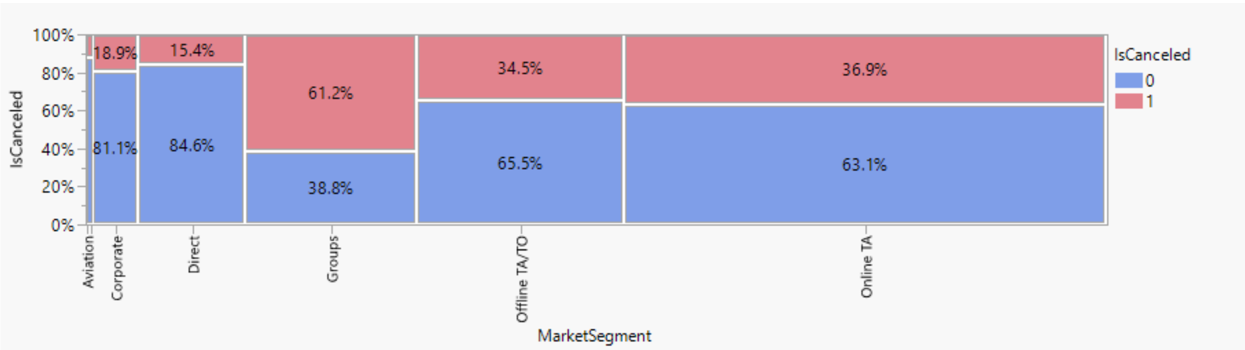


Figure 11. Mosaic plot of ‘IsCanceled’ against ‘MarketSegment’

An exploration into specific market segments was conducted via a similar mosaic plot but sliced by market segment instead (Figure 11). This shows that there are definite variations between the different market segments in cancellation rates: the most pronounced being the Groups market segment being particularly prone to cancellation.

3.1.4. Cancellation Rates by Paying Company

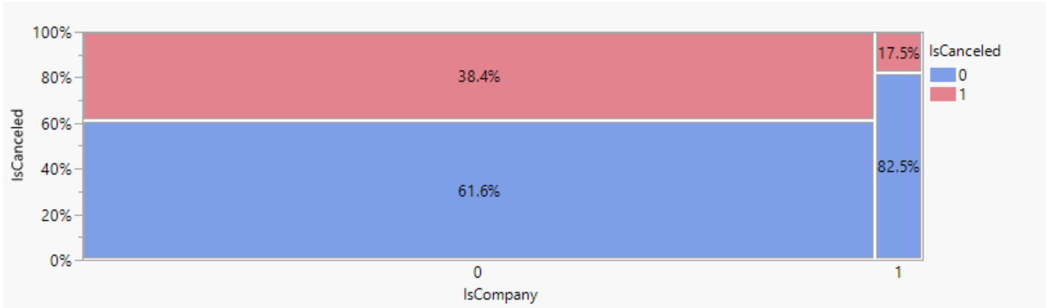


Figure 12. Mosaic plot of 'IsCanceled' against 'IsCompany'

Another aspect of clientele profile is to whether the payer is a company or not. From Figure 12, customers whose bookings were paid for by a company (presumably corporate customers) were more likely to cancel bookings, suggesting that they may be motivated by factors other than costs.

3.1.5. Cancellation Rates by Customer Loyalty

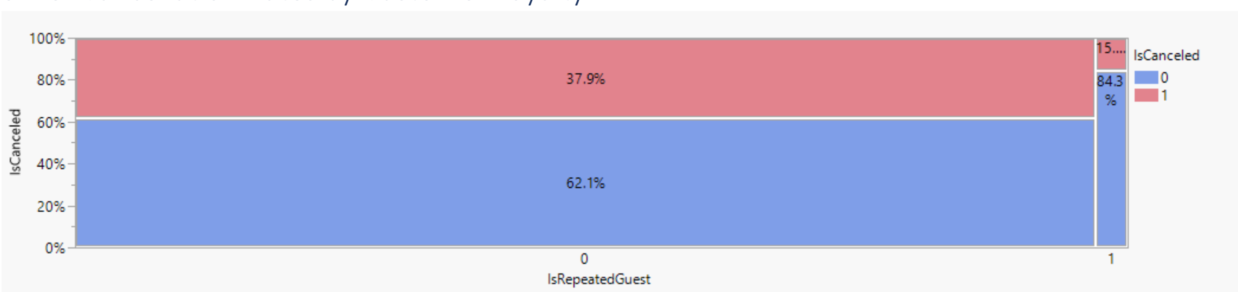


Figure 13. Mosaic plot of 'IsCanceled' against 'IsRepeatedGuest'

Customer loyal, represented by the 'IsRepeatedGuest' variable, factors into the cancellation decision, was explored by mosaic plot (Figure 13). It looks indeed possible that repeated guests cancel at a much lower rates than non-repeated guests.

3.1.6. How Early are the Cancellations

In this dataset, 'ReservationStatusDate' represents the last time a reservation was modified, so this date for cancelled bookings should be equivalent to cancellation date. A histogram was produced of the 'DeltaReservationStatusArrival', or time between cancellation and arrival, for bookings where 'ReservationStatus' is 'Canceled' (Figure 14).

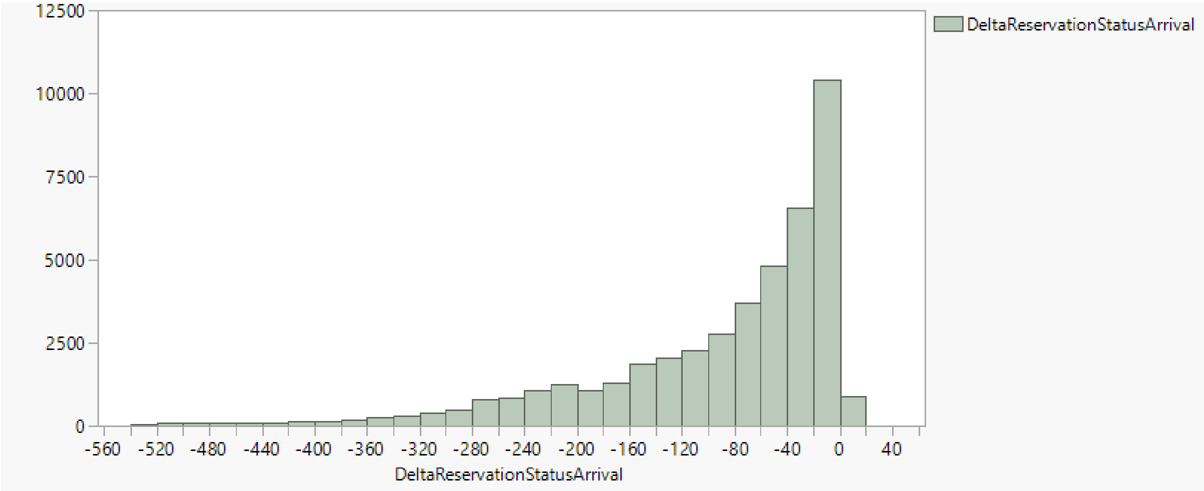


Figure 14. Histogram of 'DeltaReservationStatusArrival', for cancelled bookings only

This shows an exponential increase of cancellations as one approached the arrival date, with a drop of count at 0 for cancellations the-day-of, and 26% occurred within 3 weeks. If any intervention can be made to stop or confirm the bookings, these actions can most efficiently be carried out nearer to the arrival date.

3.1.7. Variations in Lead Time on Cancellation Status

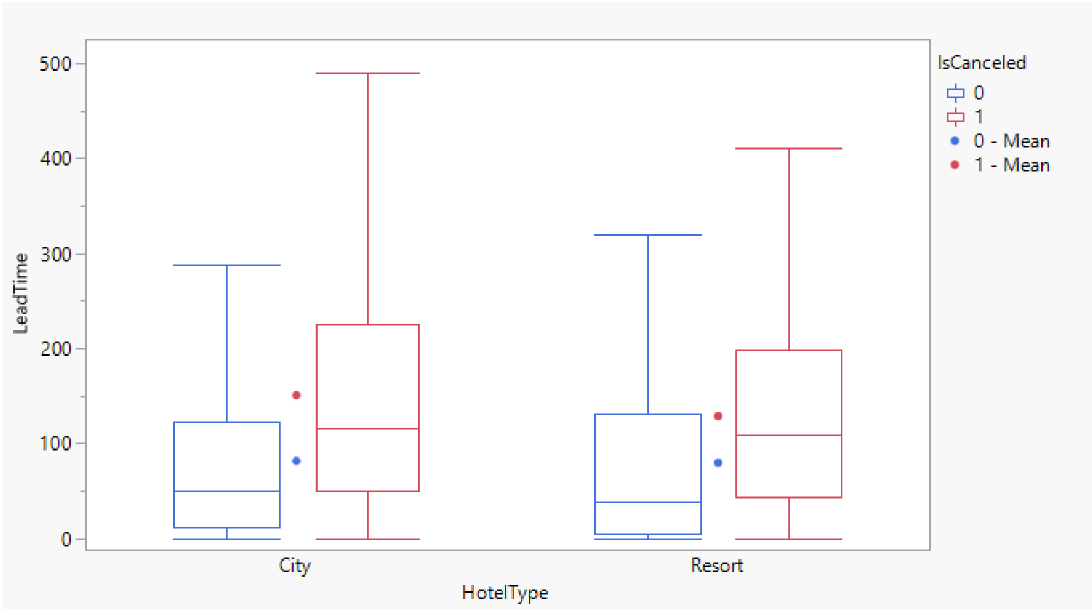


Figure 15. Box plot of 'LeadTime' vs 'IsCanceled' split by 'HotelType'

The variation in 'LeadTime' split by 'IsCanceled' categories is shown in Figure 15. It can be observed that the median lead time is significantly higher for cancelled bookings compared to those not cancelled, and this relationship holds for both hotel types.

3.1.8. Combined Seasonality and Hotel Type Effect on Cancellation Rates

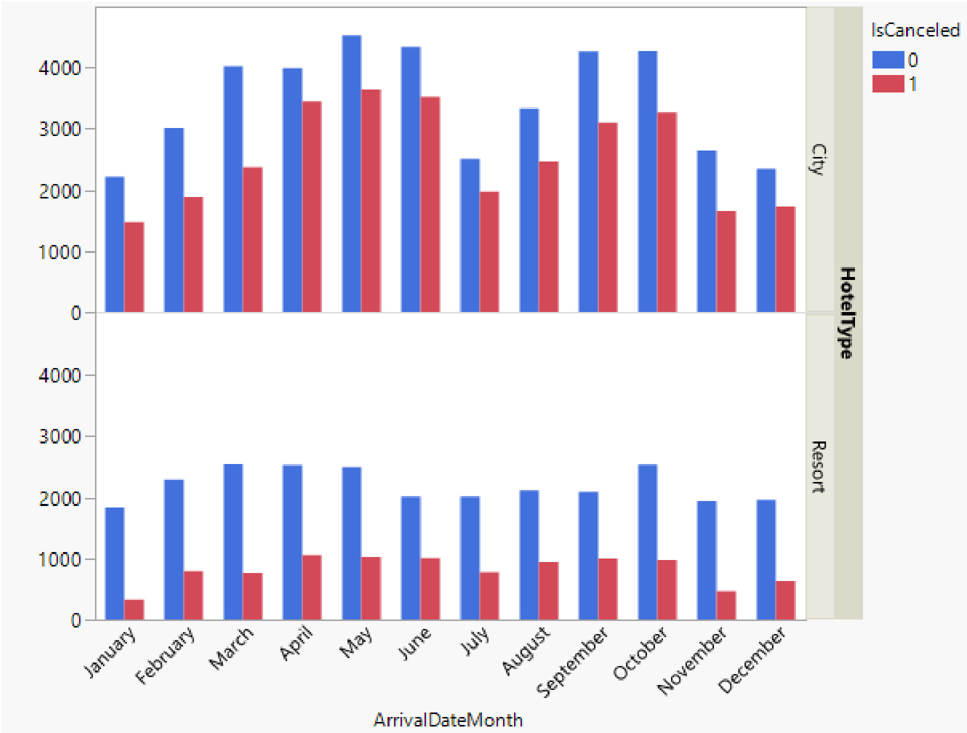


Figure 16. Bar chart showing seasonality and hotel type effect on cancellation counts

As suggested in Section 3.1.1. and 3.1.2., seasonality and hotel type have strong effect on cancellation rates. Figure 16 further sheds light on possible their possible interactions. For example, for the city hotel, cancellation rates are markedly lower in April and July; and for the resort, cancellation rates seem lowest in January and November.

3.1.9. Effect of Deposit Type on Cancellation Rates

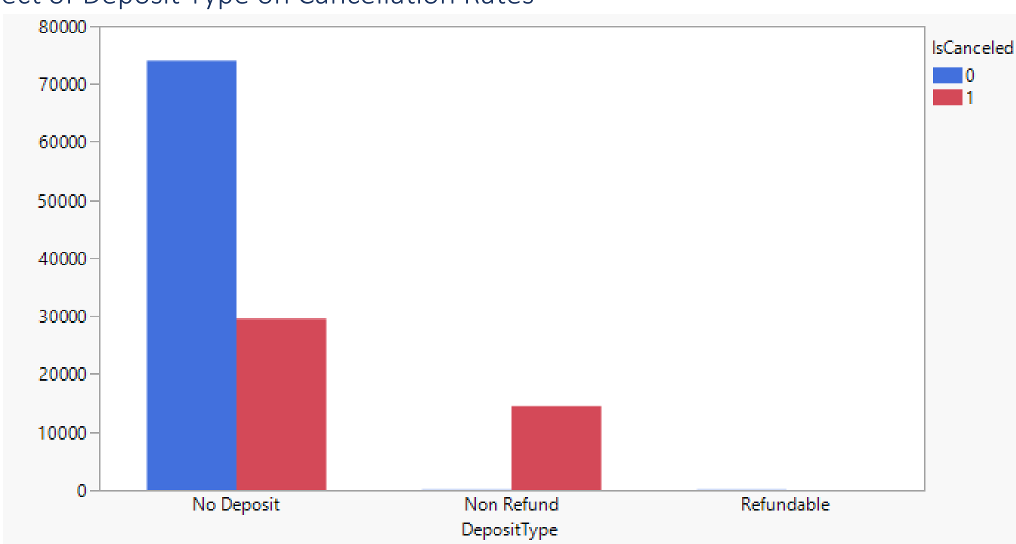


Figure 17. Bookings Count Split by 'DepositType' and 'IsCanceled'

Intuitively, the heavier the penalty, the lower the likelihood of cancellation. However, Figure 17 shows the opposite effect: for bookings that were fully non-refundable, there was disproportionally high cancellation, compared to the 'No Deposit' group's cancellations not exceeding non-cancelled bookings. This merits more investigation into why that is the case, whether a structural, or data integrity issue.

3.1.10. Effect of Hotel Type and Seasonality on Profitability

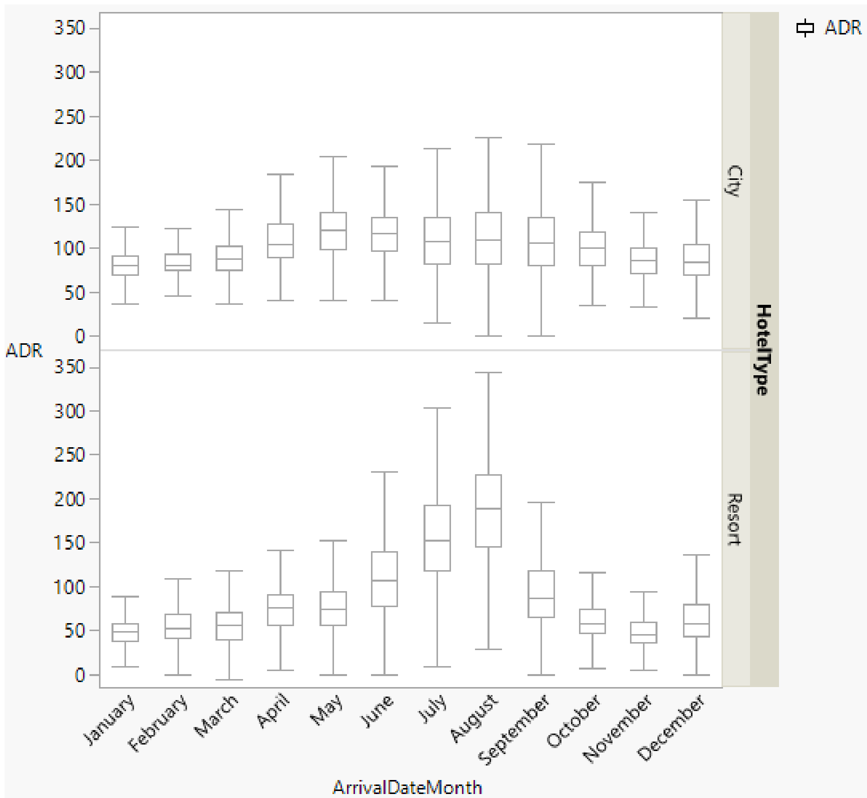


Figure 18. Box Plot of 'ADR' by 'ArrivalDateMonth', Split into 'HotelType'

'ADR' stands for 'average daily rate' and is a proxy for the hotel's profitability on a particular booking. As observed in the box plot (Figure 18), there is also seasonality effect in ADR for each type of hotel, but markedly different from that produced by bookings volume. For the city hotel, profitability somewhat tracked bookings volume's peak around April-May, then largely decreased for the remainder of the year. On the other hand, the resort had a very pronounced peak around August.

3.1.11. Variations in ADR by Hotel Type and Canceled Status

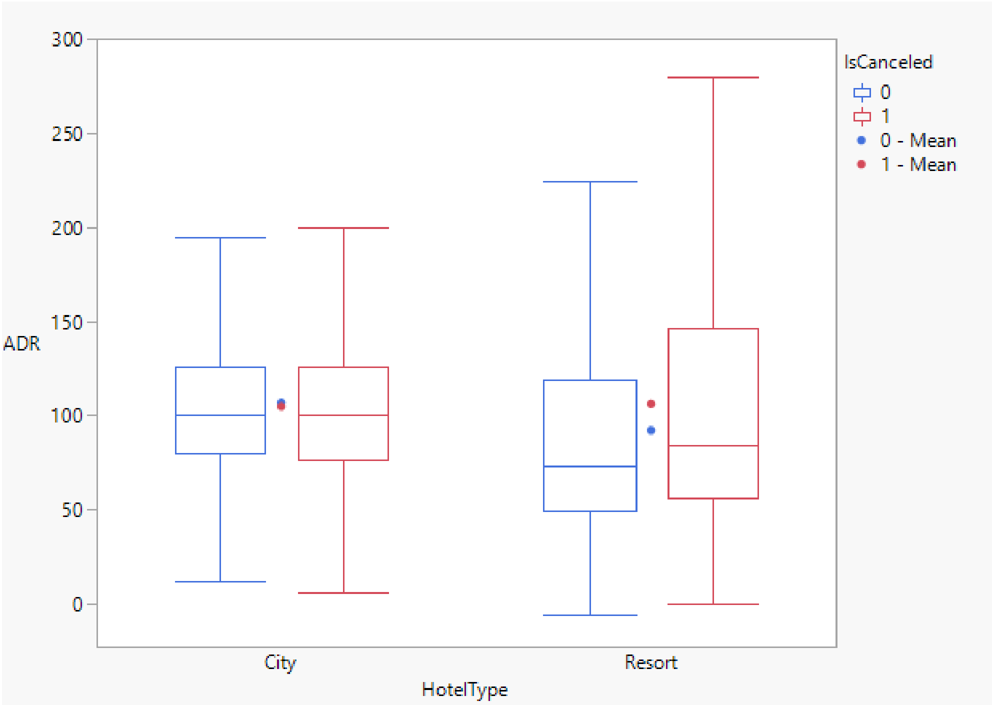


Figure 19. 'ADR' for 'IsCanceled' split by 'HotelType'

ADR is also explored for each hotel type vis-à-vis cancellation rates. There does not seem to be much difference between the profitability of cancelled bookings versus non-cancelled ones for the city hotel, shown in the box plot in Figure 19. On the other hand, for the resort, the plot suggests a significant difference in profitability: cancelled bookings tend to be those that net the resort more revenue as seen in the higher median & mean ADR compared to those that did not get cancelled.

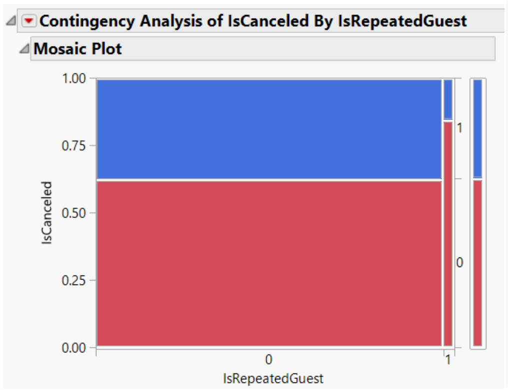
3.2. Business Hypotheses

From the several significant trends identified during the exploratory data analysis, three key hypotheses were formulated, requiring confirmatory data analysis, as outlined below.

3.2.1. Customer Loyalty Influences Cancellation Rates

With reference to Section 3.1.5., the following hypotheses were formulated, setting the confidence interval at 95% and significance level of 5%:

- H_0: there is no difference between the proportion of cancelled vs. non-cancelled bookings due to whether the person is a repeated guest.
- H_1: there is a difference between the proportion of cancelled vs. non-cancelled bookings due to whether the person is a repeated guest.



Contingency Table

IsRepeatedGuest	Count / Total % / Col % / Row %	IsCanceled 0	IsCanceled 1	Total
0	Count	71297	43521	114818
	Total %	60.26	36.78	97.04
	Col %	96.03	98.75	
	Row %	62.10	37.90	
1	Count	2949	549	3498
	Total %	2.49	0.46	2.96
	Col %	3.97	1.25	
	Row %	84.31	15.69	
Total	Count	74246	44070	118316
	Total %	62.75	37.25	

Tests

N	DF	-LogLike	RSquare (U)
118316	1	406.77335	0.0052

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	813.547	<.0001*
Pearson	716.375	<.0001*

Figure 20. Results of Chi-squared Test for 'IsCanceled' by 'IsRepeatedGuest

The contingency table showed that all categories' counts satisfied the assumptions that all expected frequencies are greater than 1 and at least 80% are greater than 5. The Chi-squared test returned a p-value which is less than the significance level, thus the null hypothesis was rejected. It was thus concluded that there is no statistical evidence to support the claim that there is no difference between the proportion of cancelled vs. non-cancelled bookings due to customer loyalty, with 95% confidence.

3.2.2. Cancelled Bookings Have Significantly Greater Lead Time

With reference to Section 3.1.7., the following hypothesis were formulated (without grouping by hotel types since the trend holds across this variable), setting the confidence interval at 95% and significance level of 5%:

- H_0: there is no difference in average lead time between the cancelled vs. non-cancelled bookings.
- H_1: there is a difference in average lead time between the cancelled vs. non-cancelled bookings.

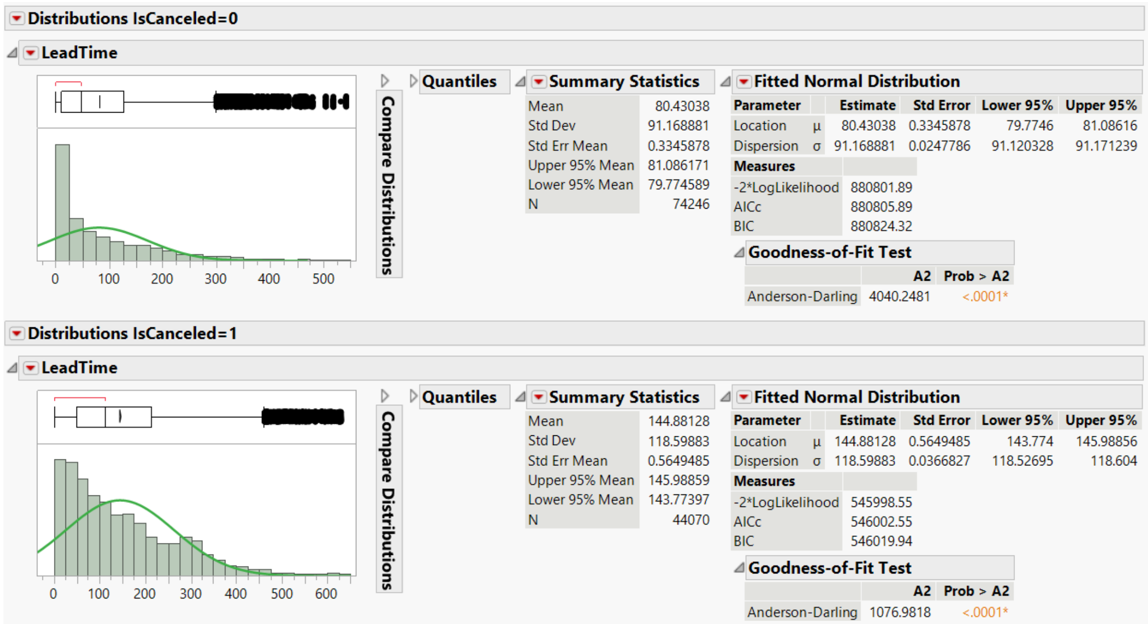


Figure 21. Anderson-Darling test results for 'LeadTime' for each 'IsCanceled' categories

Normality assumption for 'LeadTime' by 'IsCanceled' levels was evaluated via Anderson-Darling test (Figure 21). The p-values being less than the significance level confirmed that the distributions are not normal.

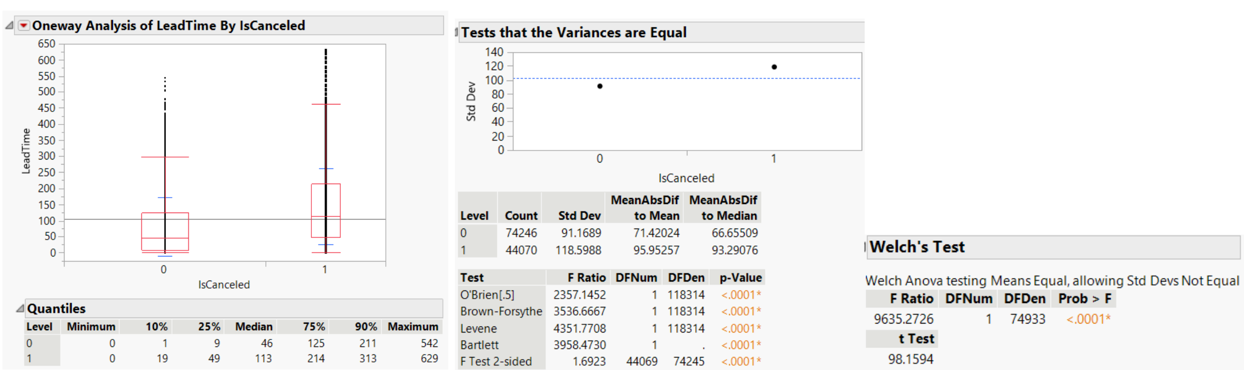


Figure 22. Unequal variance & Welch's test results for 'LeadTime' by 'IsCanceled' data

A test for unequal variance was also conducted, confirming that the groups' variances are unequal with p-value being less than significance level. Since nonparametric and unequal variance assumptions are satisfied, a Welch's test was carried out, returning the result of p-value less than 0.0001 (Figure 22). It is therefore concluded that there is no statistical evidence to support the claim that there is no difference between the average lead time between cancelled and non-cancelled bookings, with 95% confidence.

3.2.3. ADRs Differ Significantly for Cancelled Bookings, Regardless of Location

With reference to Section 3.1.11, the following hypotheses were formulated (for each individual hotel type being resort or city hotel), setting the confidence interval at 95% and significance level of 5%:

- H_0: there is no difference in mean ADR between the cancelled vs. non-cancelled bookings.
- H_1: there is a difference in mean ADR between the cancelled vs. non-cancelled bookings.

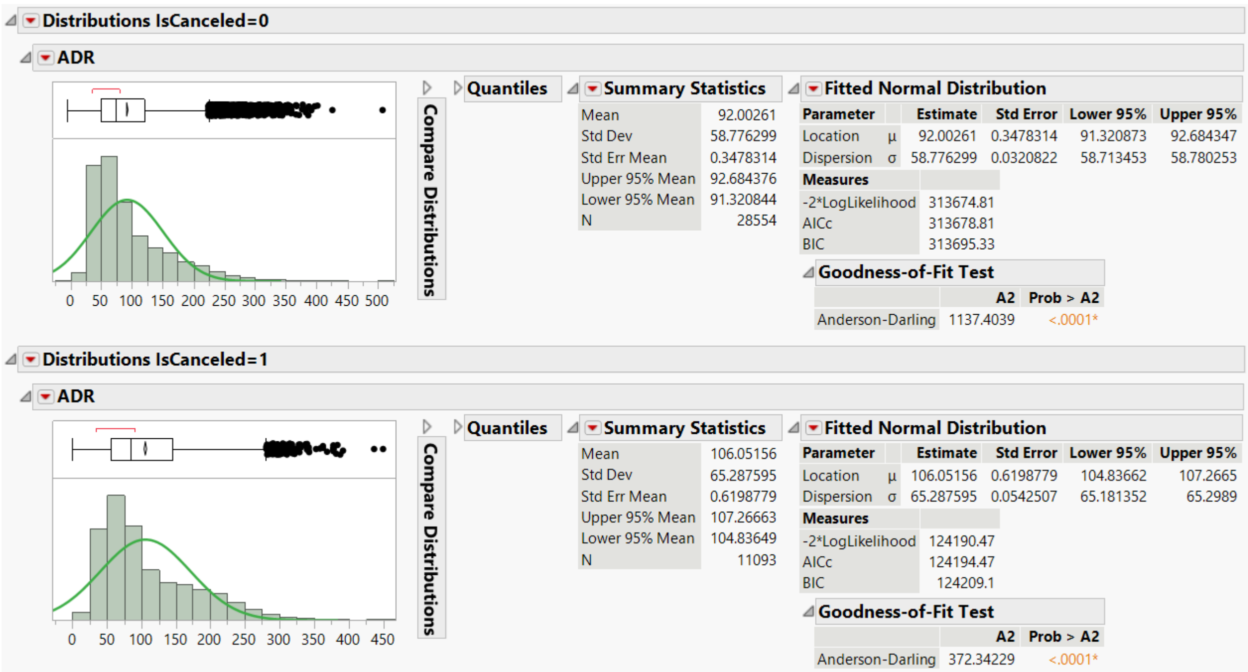


Figure 23. Anderson-Darling test results for ‘ADR for each ‘IsCanceled’ categories, for resorts

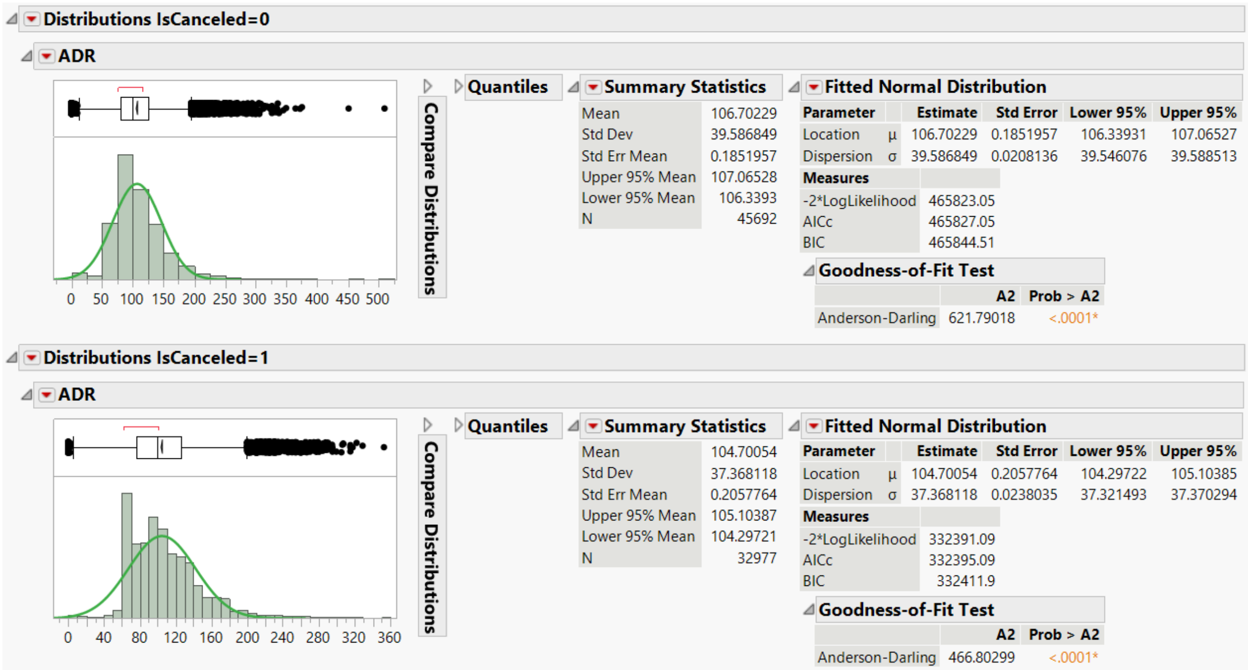


Figure 24. Anderson-Darling test results for 'ADR for each 'IsCanceled' categories, for <u>city hotels</u>

The Anderson-Darling test results for both resort and city hotels (Figure 23 & 24) returned a p-value less than significance level, meaning that the ADR is not normally distributed across all 'IsCanceled' and 'HotelType' categories.

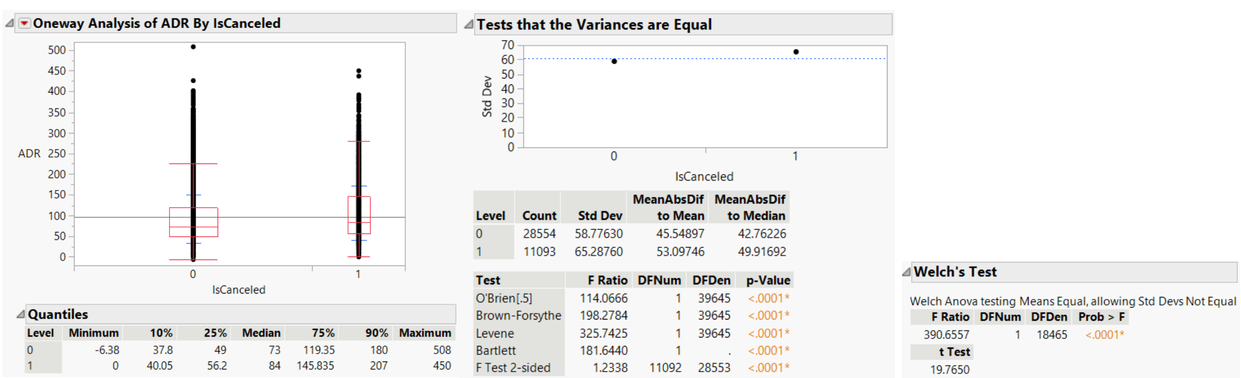


Figure 25. Unequal variance & Welch's test results for 'ADR by 'IsCanceled' data, for <u>resorts</u>

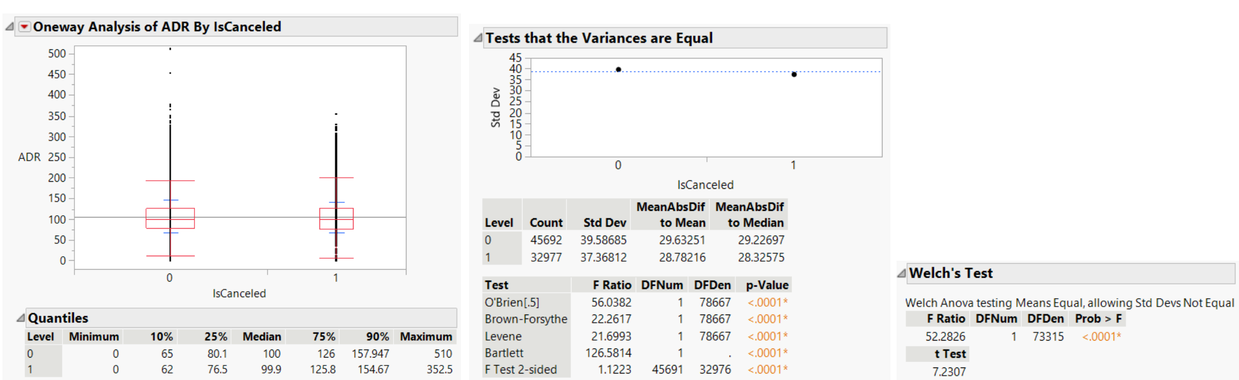


Figure 26. Unequal variance & Welch's test results for 'ADR by 'IsCanceled' data, for city hotels

Tests for unequal variance together with non-parametric Welch's tests were conducted for both resorts and city hotels. Previous visual analysis already revealed average ADR between cancelled and non-cancelled bookings differ for resorts, which was confirmed by Welch's test's p-value being less than significance level. Surprisingly, for city hotels, Welch's test's p-value was also less than significance level. It is therefore concluded mean ADR is statistically significantly lower for cancelled bookings in the city hotel, yet is higher for those in the resort, with 95% confidence.

4. Discussion

Relating the CRM data with the two hotels' bookings revealed invaluable information in potential prediction analysis for cancellation rates, room allocation, and pricing, owing to the data spanning a large, varied number of features pertinent to the hospitality business, and its large size.

It was revealed that there were definite seasonal trends in the hospitality business: whether by booking volume, ADR, or cancellation rates. Slicing the seasonal trends across different variables allow hotels to better manage their profits: prediction analysis for manpower management, pre-emptive actions in managing cancellation rates which differ from month to month, or setting price in response to peak seasons. While intuitively one expects business to pick up during "peak holiday season", this analysis gave definite answers as to which months constitute "holiday season". More importantly, it highlighted that the commonsensical definition of "peak season" differ depending on which variable is being looked at: peak ADR would differ from peak booking volume, and both differing depending on the hotel's location.

Other time-related aspects of managing cancellation rates also come into play. Specific probability distribution function can be built with the time between booking and cancellation, suggesting the possibility that hotels may efficiently intervene closer to arrival date to confirm cancellations, so rooms can be resold. Lead time also provides insight into cancellation risk calculation.

In addition, several key aspects pivotal to CRM were found to be intimately linked to cancellation rates: market segment, customer loyalty, or customer type. These help distinguish specific customer profiles and their respective distribution channels they normally avail themselves with; this in turns provides information to build pricing models with built-in discount calculated from risk of cancellations, allowing better price discrimination.

The dataset, while suffering from several drawbacks, showed that there is indeed great value in analyzing data aggregated from different sources.

Annex

Data Preparation Log

Item #	Issue	Action
1	Data comes in two separate tables for a resort and a city hotel.	Using Multiple File Imports, append H2 below H1 by rows, retaining source file's name. Recode 'File Name' into 'HotelType'.
2	'ArrivalDateMonth' is in character data type, not convenient for subsequent analysis.	Recode as ArrivalDateMonthNumeric & change modelling type to numeric & continuous.
3	The date of arrival variable is not in DateTime format and disaggregated into 3 other columns, unsuitable for analysis.	Create new formula column called 'ArrivalDate' by combining 'ArrivalDateYear', 'ArrivalDateMonth', 'ArrivalDateDayOfMonth', with numeric data type and continuous modelling type, formatted as yyyy-mm-dd.
4	'ReservationStatusDate' is in incorrect modelling type.	Change to numeric continuous, formatted as yyyy-mm-dd, consistent with ArrivalDate.
5	'Children' is in character data type.	Change to numeric continuous.
6	'IsCanceled', 'IsRepeatedGuest' should be categorical.	Change to numeric nominal.
7	'Adults', 'Children', and 'Babies' are unusually high for some observations.	Hide and exclude Adults > 4 (16 rows), Children > 3 (1 row), and Babies > 2 (2 rows), total 19 observations.
8	'Children' has 4 missing observations.	Hide & exclude from analysis, total 4 rows.
9	'MarketSegment' has 2 'Undefined' observations.	Hide & exclude from analysis, total 2 rows.
10	'DistributionChannel' has 5 'Undefined' observations.	Hide & exclude from analysis, total 5 rows.
11	Length of stay should be present for analysis.	Create 'StaysInNightsTotal' = 'StaysInWeekNights' + 'StaysInWeekendNights'
12	'Adult' has observations with 0 for value.	Hide & exclude from analysis, total 403 rows.
13	'StaysInNightsTotal' has observations with 0 for value.	Hide & exclude from analysis, total 715 rows.
14	'Meal' values being 'Undefined' or 'SC' mean the same thing – no meal package – per definition, yet encoded as two different values.	Recode 'Undefined' as 'SC' as well to prevent confusion that these observations are missing data and reduce the number of categories. New column is named 'MealRecoded'.
15	'Company' is made up majorly of NULL values (94%).	Create 'IsCompany' by encoding 'Company'.
16	'ADR' has 1 outlier with value 5400.	Hide & exclude from analysis, total 1 row.
17	'LeadTime' has 2 outliers with value above 700.	Hide & exclude from analysis, total 2 rows.

References

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in brief, 22*, 41-49.