Hotel Booking Cancellation Analysis

Tan Yue Hang

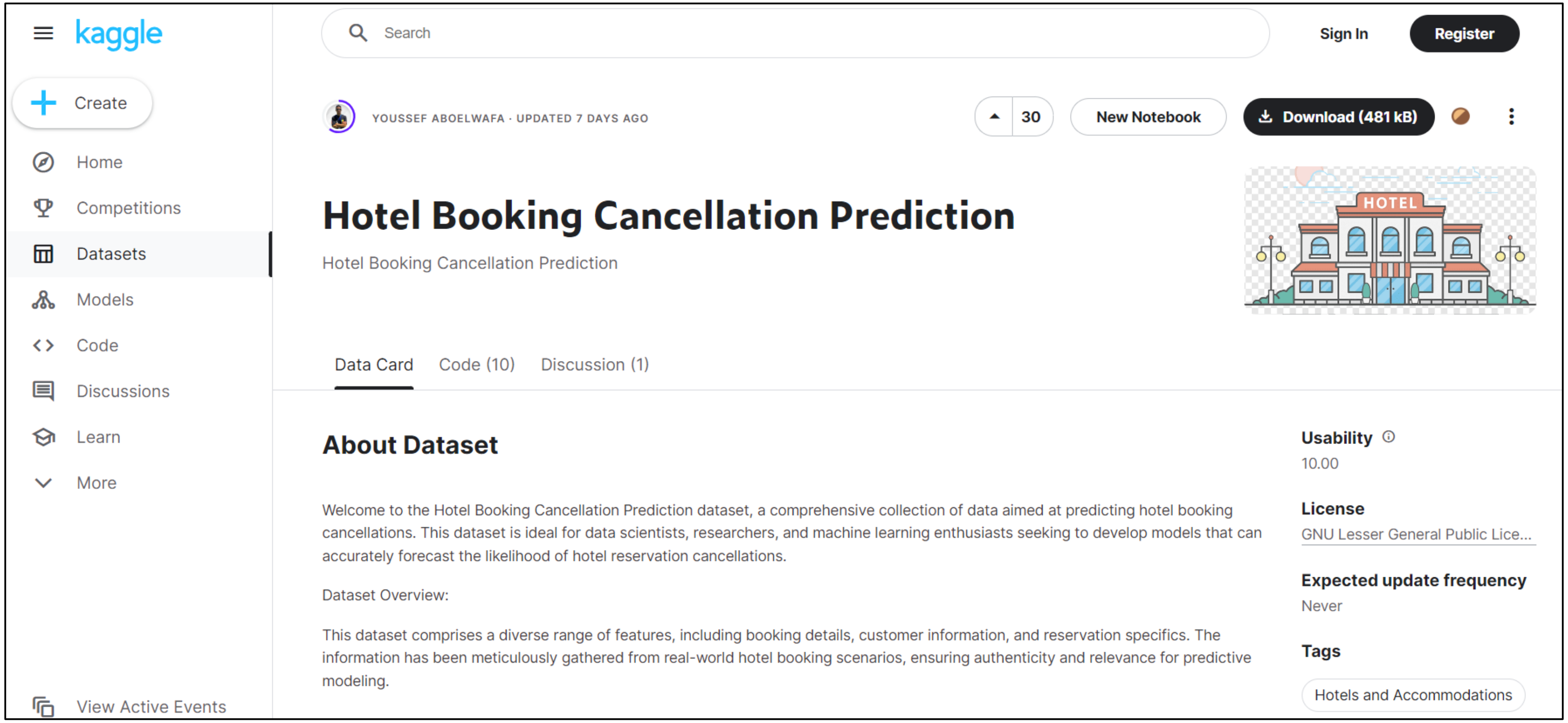
Data Analysis - Mini Project 1

Introduction

- This project will investigate the hotel booking cancellation data analysis and predictive modelling.
- The hotel business is a part of the hospitality industry that is closely tied with tourism and business traveling.
- Understanding the factors influencing cancellations is crucial. We will delve into the impact of seasonality (basic time series), booking lead time, room type and other factors on cancellation rates.

Source of Dataset

- The dataset was acquired from [Kaggle](), a fairly new dataset uploaded to Kaggle in Dec 2023.
- The dataset, as the author claimed, was gathered from real-world hotel booking scenarios, comprises of 17 data columns and >36k of observational rows, which makes it a good and moderately big dataset for data science and ML project.



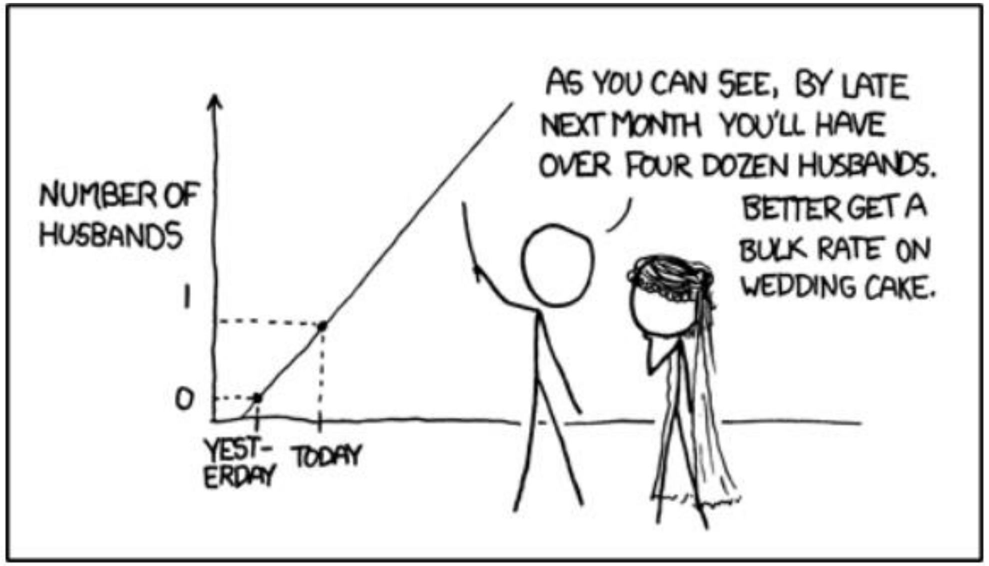
Presentation Contents

Section 1: Time Series EDA
Section 2: Features EDA and Data Cleaning
Section 3: ML Modelling

Section 1: Time Series EDA

<u>Notes:</u>

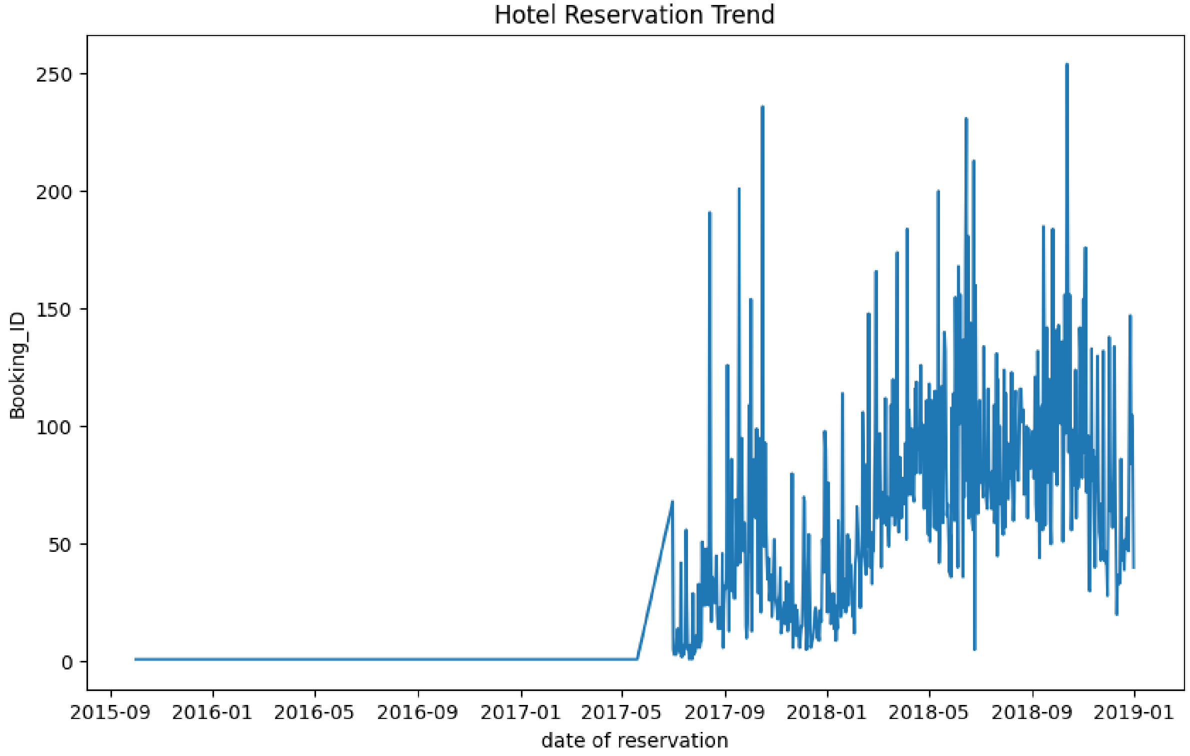
- The main purpose of this dataset is for predictive modelling analysis, and the information provided for time series analysis is kind of limited.
- Hence we can only do some basic time series EDA with this.

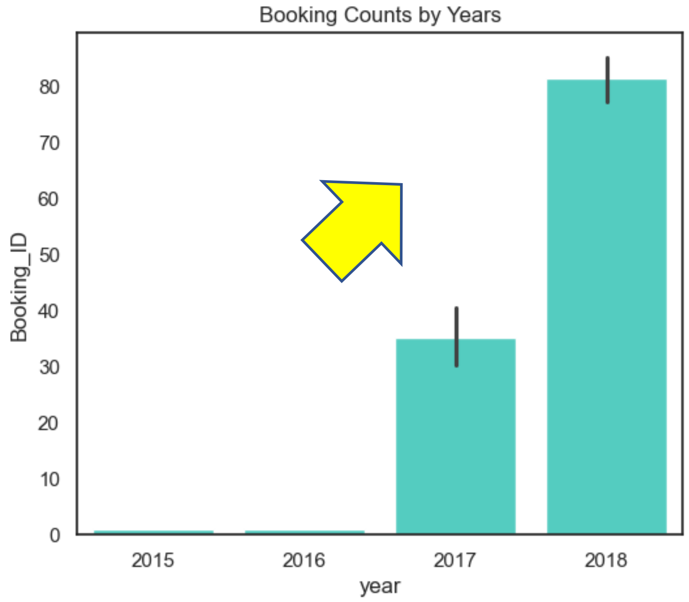


Cartoon source: https://xkcd.com/605/

Hotel Reservation Trend

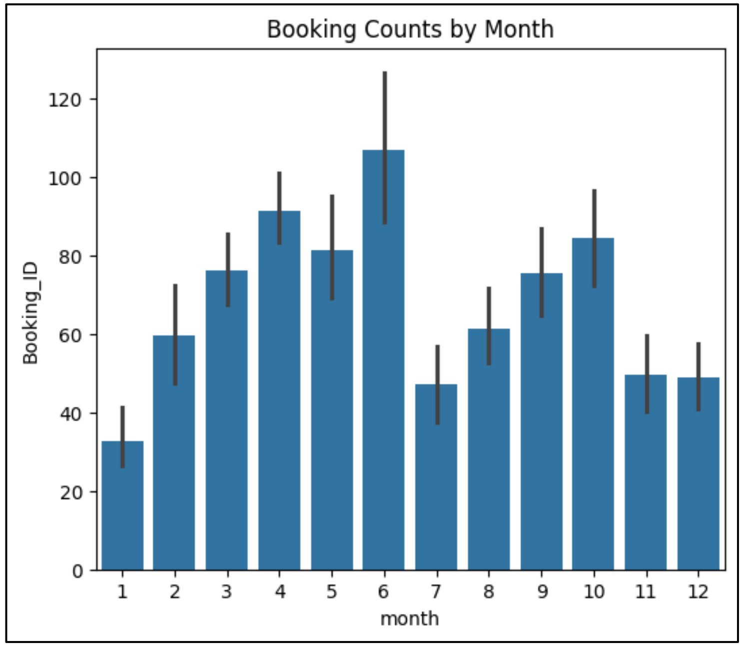
- ***What does the booking trend look like?***
- With "date of reservation" column, we can utilize this to create a simple time series chart.
- From the time series plot, it appears that while the hotel business was established in year 2015, the sales (booking) only started to pick up in mid of year 2017, and gradually increases after that with fluctuations.
- It's quite unlikely for hotel business to survive for almost two years with nearly zero booking, the hotel probably didn't have a good tracking system to record the earlier booking information.

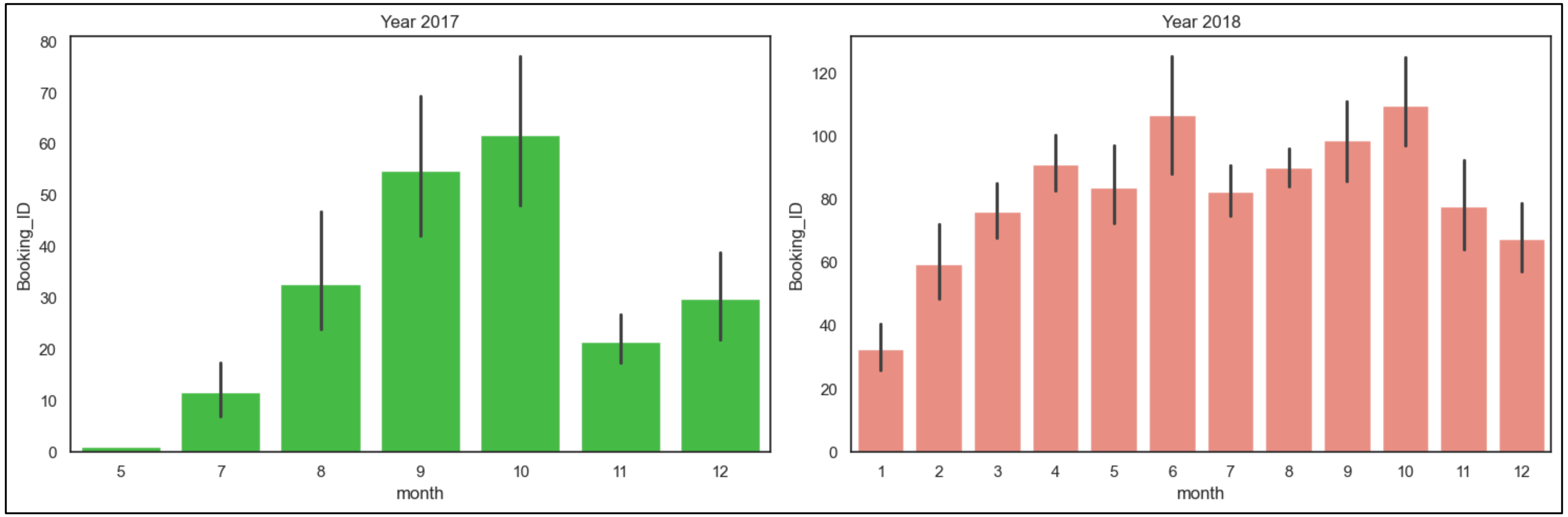




Reservation Trend by Months

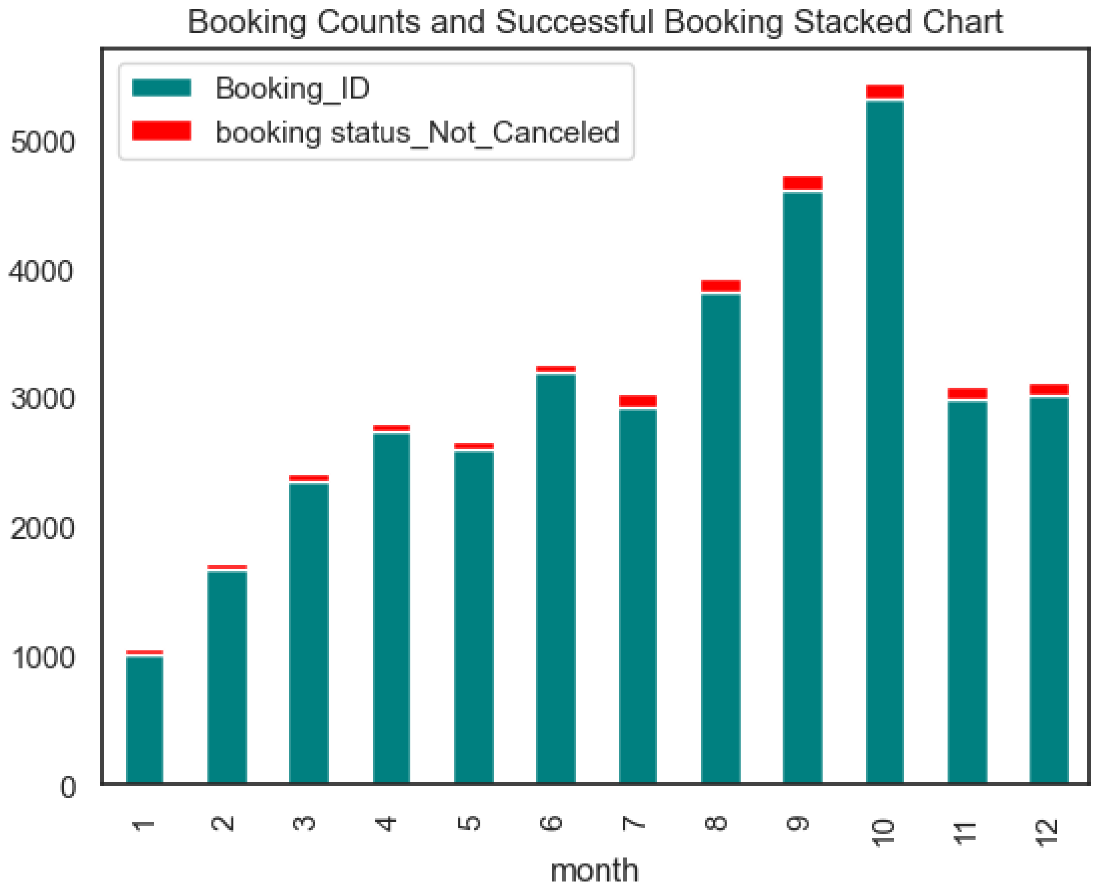
- ***Is there any pattern to the bookings made?***
- Hotel booking could be seasonal, and from the column plot by months there seems to be a fluctuating pattern to the booking counts over the months.
- Breaking down the booking counts further by year 2017 and year 2018, it's quite consistent that October usually observes higher booking counts after June.
    - This could be due to holidays seasons, weather patterns, events and etc, which are not provided by the author.
- We will be able to offer more insights on this only with more information provided.





Booking Counts vs. Cancellation Rates

- ***More bookings leads to more cancellation and vice versa?***
- From the figure, there doesn't seem to be significant correlation exists in between the booking counts and the cancellation rates.
- It also means that cancellation rates could be correlated with other factors instead.



Section 2: Features EDA and Data Cleaning

Dataset Information

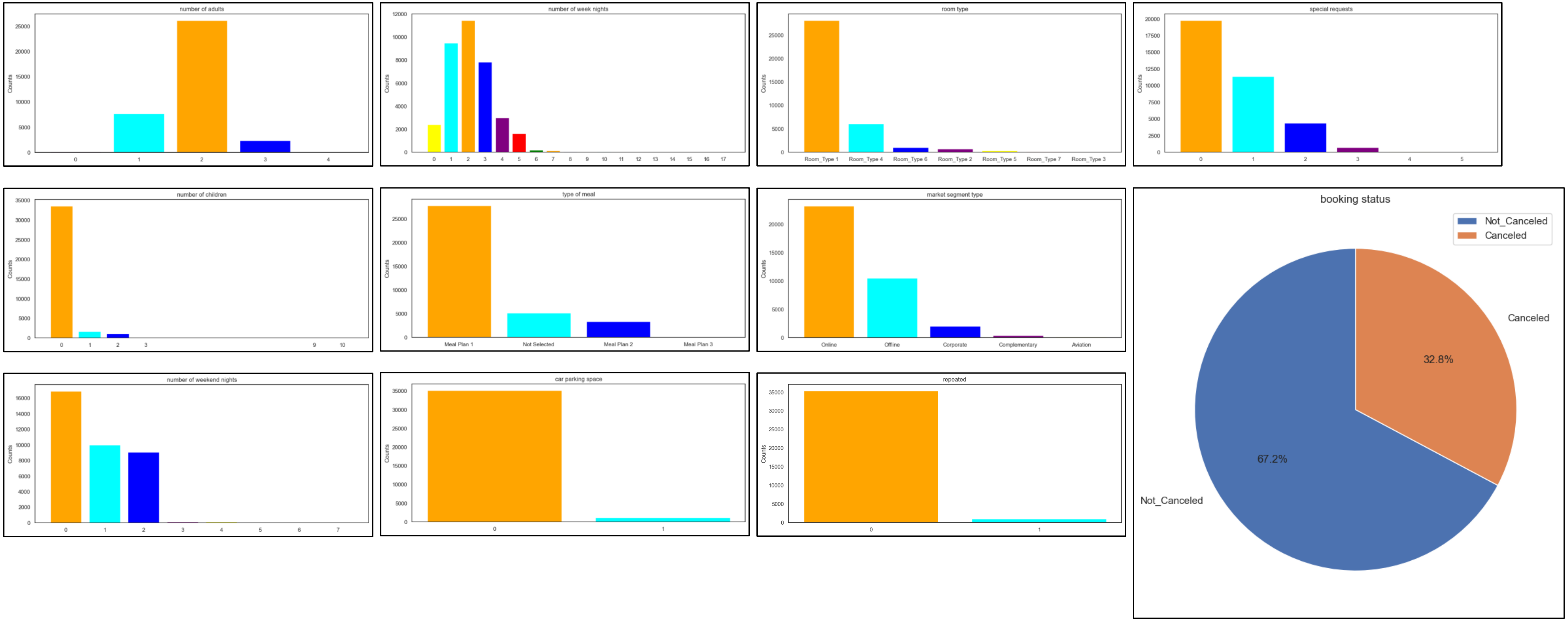
- The dataset comprises of 36,248 data rows and 17 features in total before data cleaning.
- The target (response) of the analysis would be “booking status”, which reflects whether a booking is cancelled.
- Most of the features are categorical, except for “lead time”, “average price”, and “date of reservation”.

```
<class 'pandas.core.frame.DataFrame'>
Index: 36248 entries, 0 to 36284
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Booking_ID                36248 non-null  object
 1   number of adults          36248 non-null  int64
 2   number of children        36248 non-null  int64
 3   number of weekend nights  36248 non-null  int64
 4   number of week nights     36248 non-null  int64
 5   type of meal              36248 non-null  object
 6   car parking space         36248 non-null  int64
 7   room type                 36248 non-null  object
 8   lead time                 36248 non-null  int64
 9   market segment type       36248 non-null  object
 10  repeated                  36248 non-null  int64
 11  P-C                       36248 non-null  int64
 12  P-not-C                   36248 non-null  int64
 13  average price             36248 non-null  float64
 14  special requests          36248 non-null  int64
 15  date of reservation       36248 non-null  datetime64[ns]
 16  booking status            36248 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(10), object(5)
memory usage: 5.0+ MB
```

	number of adults	number of children	number of weekend nights	number of week nights	car parking space	lead time	repeated	P-C	P-not-C	average price	special requests
count	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000	36285.000000
mean	1.844839	0.105360	0.810693	2.204602	0.030977	85.239851	0.025630	0.023343	0.153369	103.421636	0.619733
std	0.518813	0.402704	0.870590	1.410946	0.173258	85.938796	0.158032	0.368281	1.753931	35.086469	0.786262
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000	1.000000	0.000000	17.000000	0.000000	0.000000	0.000000	80.300000	0.000000
50%	2.000000	0.000000	1.000000	2.000000	0.000000	57.000000	0.000000	0.000000	0.000000	99.450000	0.000000
75%	2.000000	0.000000	2.000000	3.000000	0.000000	126.000000	0.000000	0.000000	0.000000	120.000000	1.000000
max	4.000000	10.000000	7.000000	17.000000	1.000000	443.000000	1.000000	13.000000	58.000000	540.000000	5.000000

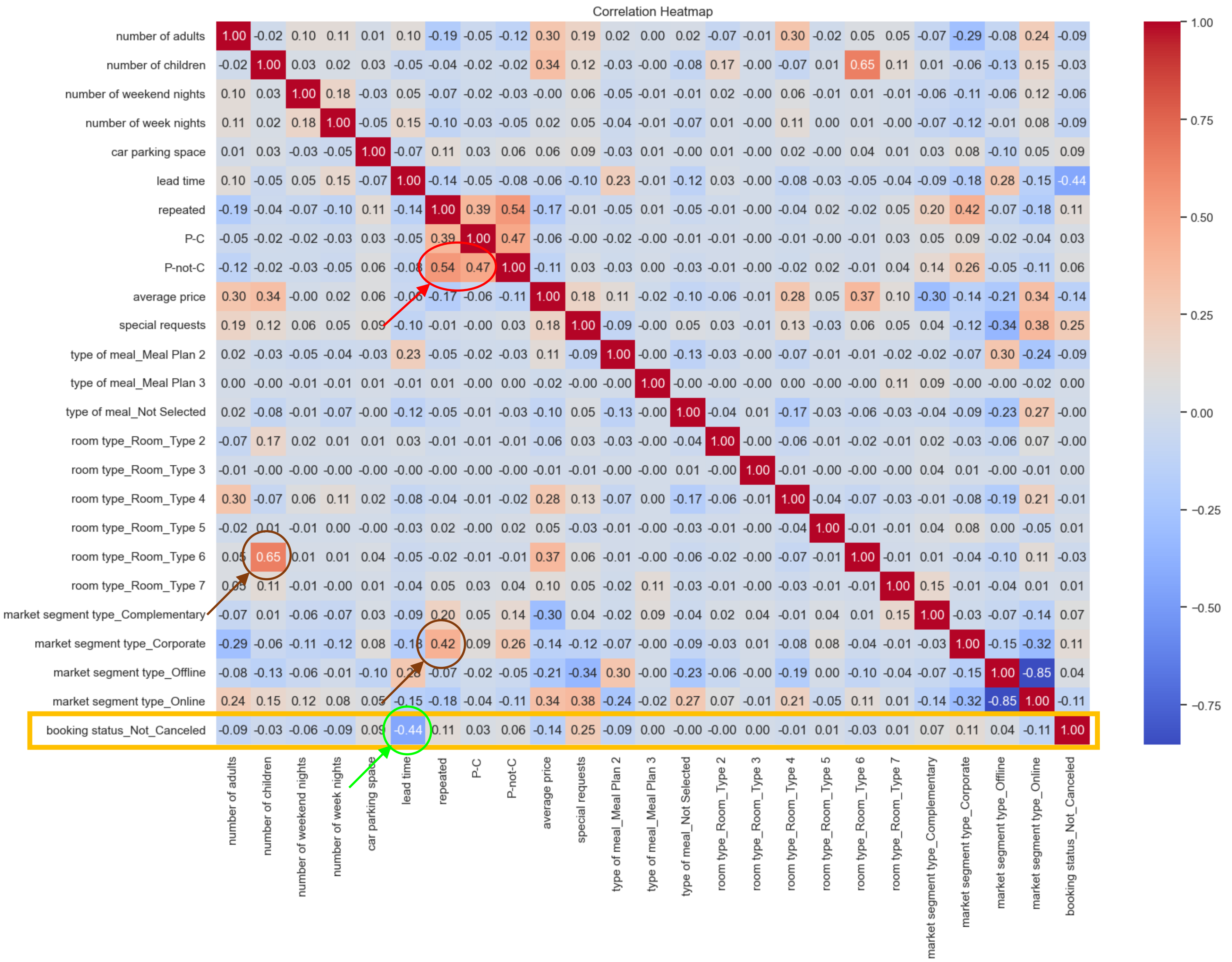
Features Exploration
(Categorical, w/o One-Hot Encoding)

- Create column charts to have a quick view of the categorical data distributions, as well as the classes of the target (booking status).
- It appears that the classes of booking status are slightly imbalanced, but still not too bad.



Correlation Plot
(with One-Hot Encoding)

- The correlation plot of all the features vs. the target is shown on the left.
- The features "P-not-C" and "P-C" will likely be dropped from the model later on, due to that:
  - They are not explained by the author what they are, hence no further insight can be derived from them.
  - From backward selection, it appears that the model performs slightly better by excluding one of these features.
- ***For some one-hot encoded features, why only part of the features seem to have significant correlation than the others?***



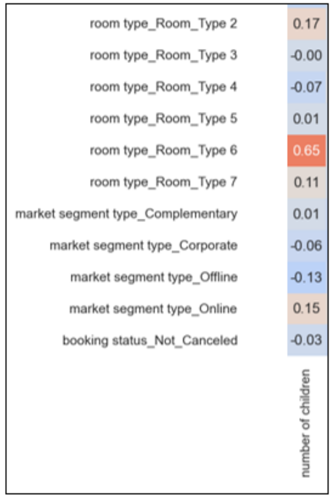
Room Type vs. Number of Children

Results Comparison of Pearson and ANOVA

	Pearson		ANOVA	
	R	Significance	P-value	Significance
Room Type 2	0.17	No	0.00	Yes
Room Type 3	-0.00	No	0.49	No
Room Type 4	-0.07	No	0.00	Yes
Room Type 5	0.01	No	0.26	No
Room Type 6	0.65	Yes	0.00	Yes
Room Type 7	0.11	No	0.00	Yes

number of children	0	1	2	3	9	10
room type						
Room_Type 1	26765	1312	34	1	1	0
Room_Type 2	482	25	179	5	1	0
Room_Type 3	7	0	0	0	0	0
Room_Type 4	5845	190	15	0	0	1
Room_Type 5	239	13	11	0	0	0
Room_Type 6	121	64	774	5	0	0
Room_Type 7	91	16	43	8	0	0

W/o One-Hot Encoding

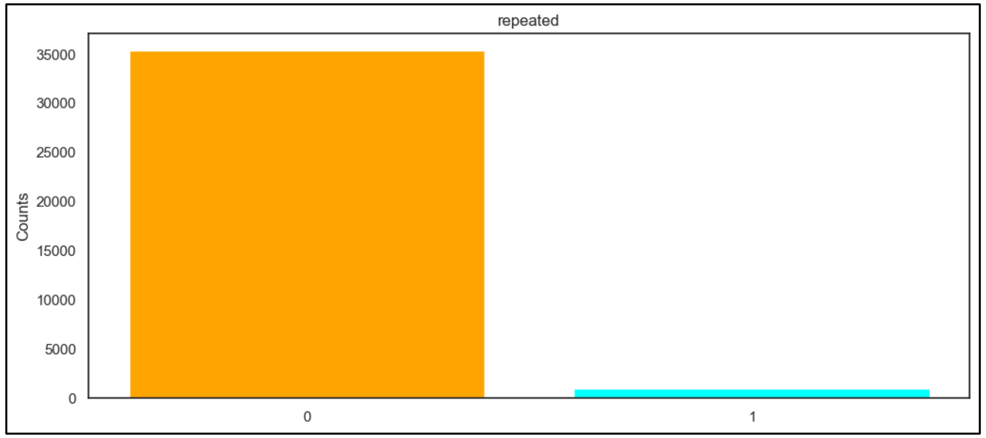


With **Pearson's r** and One-Hot Encoded

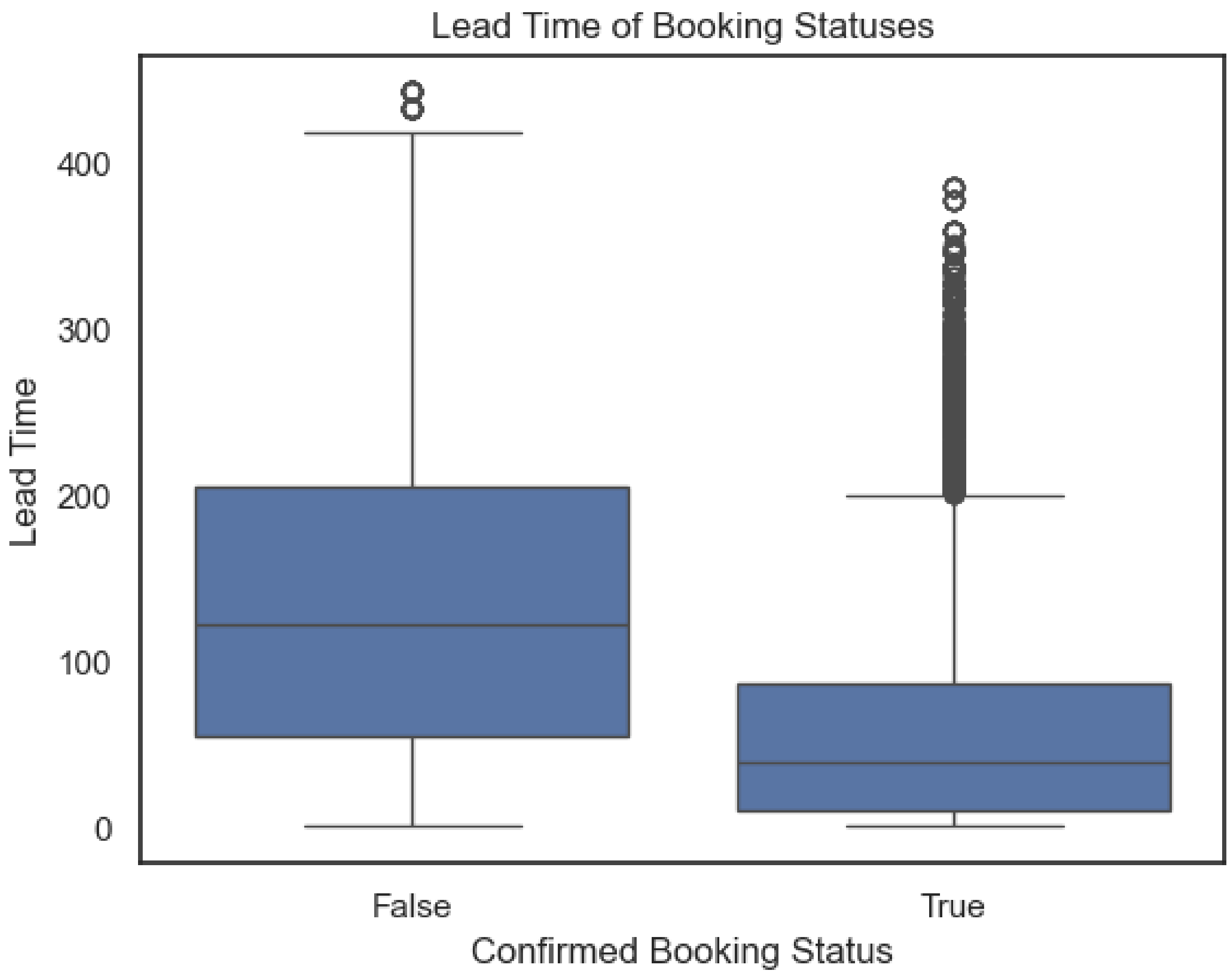
- It's worth noting that the heatmap constructed in the earlier slide was based on the default setting of **pd.corr()**, hence it applied Pearson's r by default to evaluate for all the features.
- However by looking at the cross-tabulation of "room type" vs. "number of children", it doesn't seem clear that how Pearson comes about all the r values with these two features, and since one of them is categorical data, <u>Pearson may not be the best method</u>.
- Try with **ANOVA** which is supposed to be more suitable for **categorical vs numerical data**, results summarized in the table above. Apparently both methods give **fairly different results**.
- Based on ANOVA results, it seems that there really is **strong correlation between "room type" and "number of children"** for the majority of the sub-columns.

Market Segment Type and Repeated Booking
(w/o One-Hot Encoding)

repeated	0	1
market segment type		
Aviation	109	16
Complementary	264	126
Corporate	1412	599
Offline	10431	90
Online	23106	95

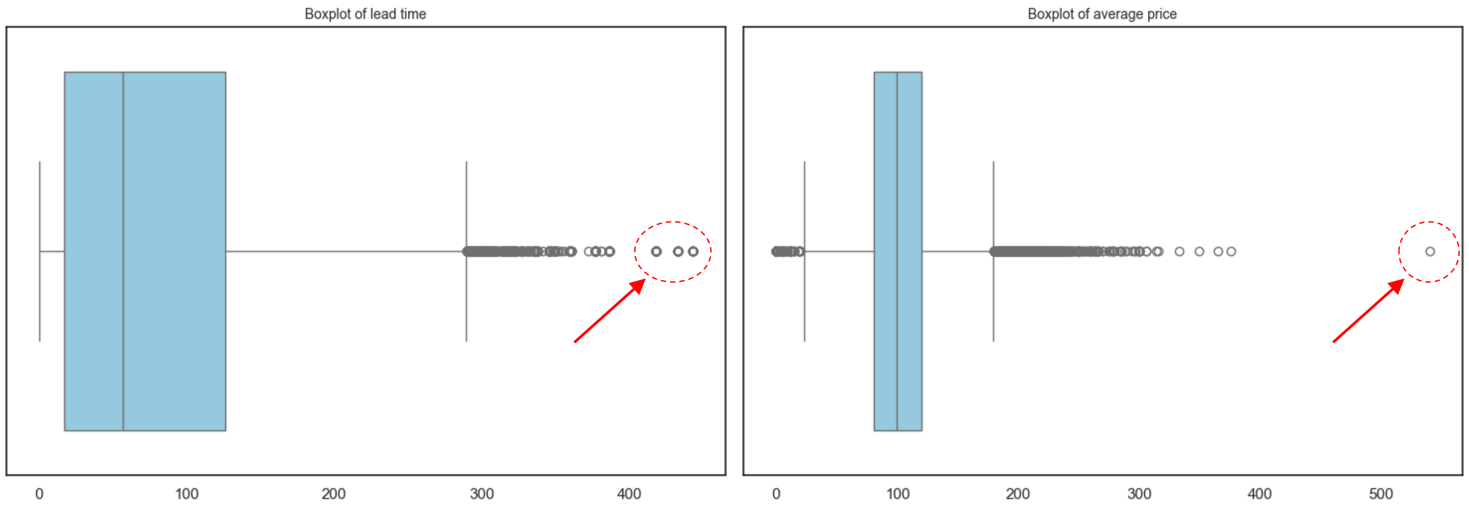


- From the cross tabulation, it’s true that repeated guests are seemingly less than first-time guests, regardless of market segment type.
- This aligns with the nature of many businesses, where the number of repeat customers is naturally less than the number of first-time customers.
- Besides that, based on **domain knowledge**, neither of these features could explain the other, so there shouldn’t be a multicollinearity issue between them, so keep them both in the modelling as a start.

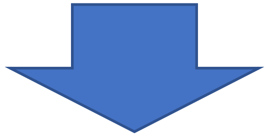


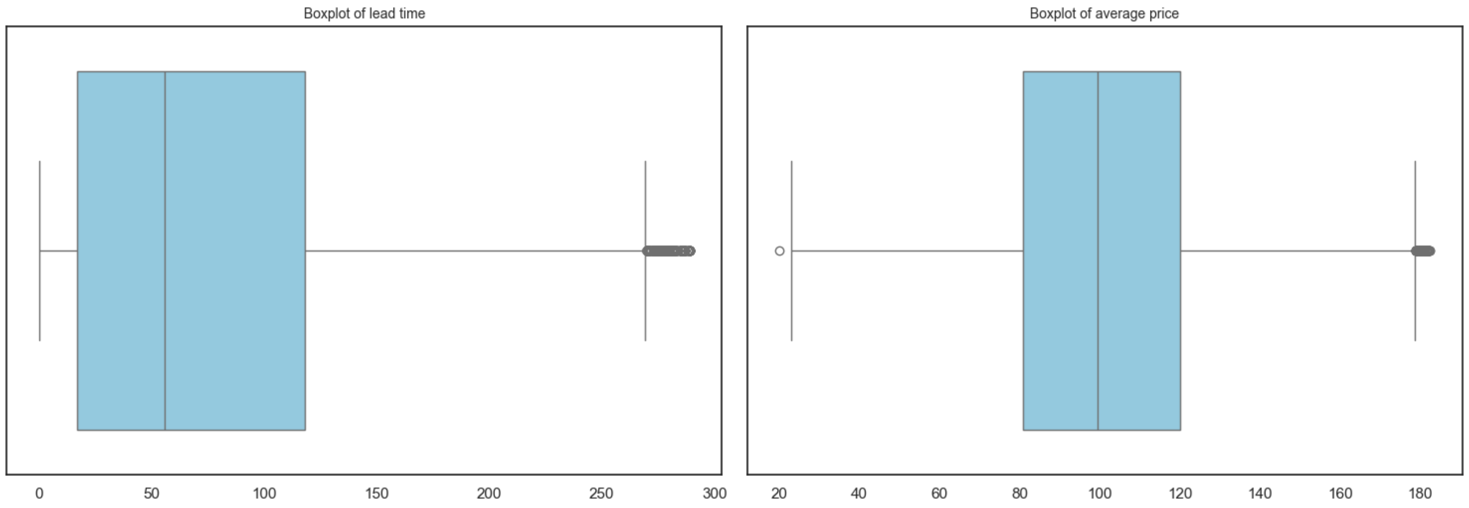
Booking Status and Lead Time

- ***How does the correlation of "lead time" and "booking status" look like, and the distribution?***
- Apparently there is visible correlation exists where confirmed booking status is generally associated with shorter lead time, but there are also more outliers come with it which could suggest that the datapoints in this groups have a lot of exceptions / unreliable datapoints for generalization.
- We will deal with this by removing some of the extreme outliers in the following slide.



Extreme outliers removal

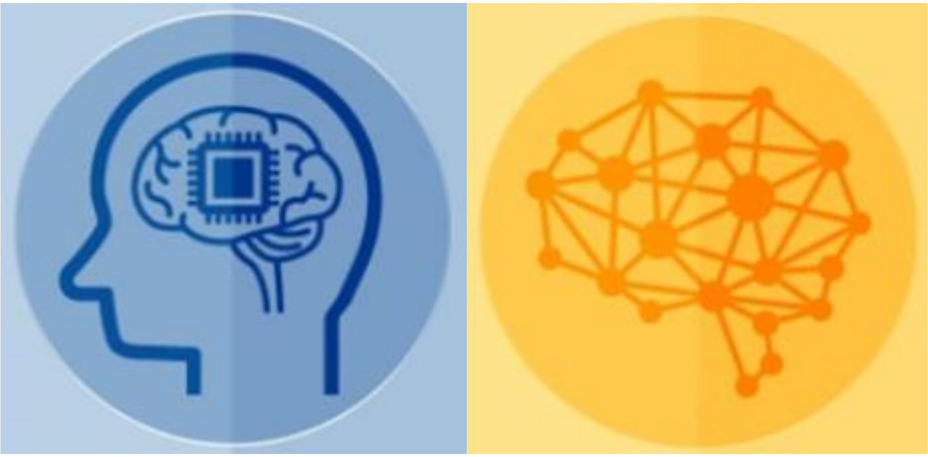




Features Exploration (Continuous)

- There are only two columns with continuous data, namely “lead time” and “average price”.
- From boxplots, there seem to be some extreme outliers found in both features. Extreme outliers are bad for data generalization.
- As part of data cleaning, remove the extreme outliers by **1.5*IQR** to improve the model performance later.
- With this operation, the size of the dataset has been reduced from **36,248** down to **33,312**, which is still acceptably large enough for data analysis and ML model building.

Section 3: ML Modelling



Cartoon source: https://www.futurefundamentals.com/

ML Model Selection
(with One-Hot Encoding)

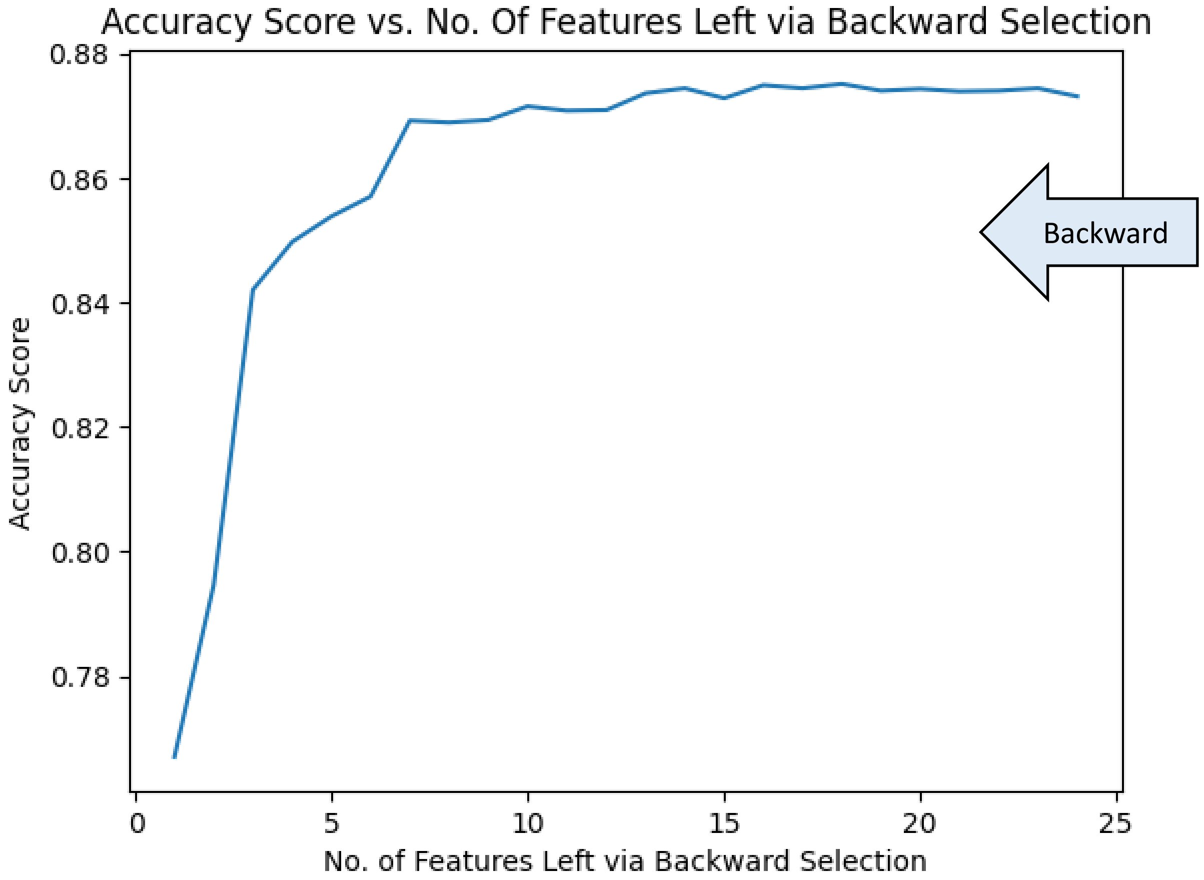
	Accuracy	Precision	Recall	F1	Elapsed Time (s)
LogisticRegression(max_iter=1000)	0.7977	0.8247	0.8956	0.8587	3.023
RandomForestClassifier()	0.8741	0.8921	0.9280	0.9098	41.355
GaussianNB()	0.3808	0.8896	0.1121	0.1985	0.405
SVC()	0.8187	0.8277	0.9292	0.8755	196.947
NuSVC()	0.7994	0.8073	0.9295	0.8641	284.351
LinearSVC()	0.7968	0.8220	0.8985	0.8585	13.229
HistGradientBoostingClassifier()	0.8657	0.8814	0.9302	0.9045	16.226
XGBClassifier()	0.8711	0.8881	0.9293	0.9082	12.653
SGDClassifier()	0.7823	0.8130	0.9071	0.8508	2.392

- All the models were fitted and evaluated with **cross validation of 5-fold** to ensure the assessment results are reliable and stable.
- Tried fitted with a few classification models, and **RandomForestRegressor** appears to perform the best in terms of accuracy score.
- The training time taken is also acceptable.

Features Backward Selection

(with One-Hot Encoding)

- **18** selected features (out of 24) appears to be the sweet spot to obtain the highest accuracy score with backward selection.




```
Best features: ['P-not-C', 'room type_Room_Type 5', 'room type_Room_Type 2', 'repeated', 'market segment type_Corporate', 'car parking space', 'number of children', 'type of meal_Meal Plan 2', 'type of meal_Not Selected', 'room type_Room_Type 4', 'market segment type_Offline', 'market segment type_Online', 'number of adults', 'number of weekend nights', 'number of week nights', 'special requests', 'average price', 'lead time’]

Accuracy score: 0.8789
Dataset: Training data
```

Appendices

Room Type and Number of Children

number of children	0	1	2	3	9	10
room type						
Room_Type 1	26765	1312	34	1	1	0
Room_Type 2	482	25	179	5	1	0
Room_Type 3	7	0	0	0	0	0
Room_Type 4	5845	190	15	0	0	1
Room_Type 5	239	13	11	0	0	0
Room_Type 6	121	64	774	5	0	0
Room_Type 7	91	16	43	8	0	0

W/o One-Hot Encoding

	number of children
room type_Room_Type 2	0.17
room type_Room_Type 3	-0.00
room type_Room_Type 4	-0.07
room type_Room_Type 5	0.01
room type_Room_Type 6	0.65
room type_Room_Type 7	0.11
market segment type_Complementary	0.01
market segment type_Corporate	-0.06
market segment type_Offline	-0.13
market segment type_Online	0.15
booking status_Not_Canceled	-0.03

With **Pearson's r** and One-Hot Encoded

	sum_sq	df	F	PR(>F)	room type
C(Q("room type_Room_Type 2"))	164.474278	1.0	1043.944362	8.163373e-226	room type_Room_Type 2
Residual	5710.586608	36246.0	NaN	NaN	room type_Room_Type 2
C(Q("room type_Room_Type 3"))	0.077635	1.0	0.478974	4.888920e-01	room type_Room_Type 3
Residual	5874.983251	36246.0	NaN	NaN	room type_Room_Type 3
C(Q("room type_Room_Type 4"))	32.890900	1.0	204.061775	3.622345e-46	room type_Room_Type 4
Residual	5842.169986	36246.0	NaN	NaN	room type_Room_Type 4
C(Q("room type_Room_Type 5"))	0.204411	1.0	1.261151	2.614408e-01	room type_Room_Type 5
Residual	5874.856475	36246.0	NaN	NaN	room type_Room_Type 5
C(Q("room type_Room_Type 6"))	2479.973792	1.0	26476.236862	0.000000e+00	room type_Room_Type 6
Residual	3395.087094	36246.0	NaN	NaN	room type_Room_Type 6
C(Q("room type_Room_Type 7"))	76.028211	1.0	475.203139	1.114145e-104	room type_Room_Type 7
Residual	5799.032675	36246.0	NaN	NaN	room type_Room_Type 7

With ANOVA and One-Hot Encoded

End of Presentation