Hotel Bookings

Friday, July 5, 2025

By:

- Ahmad Alqaisi
- Samer Magdy
- Toka Abdelghafar
- Nouran magdy
- Osama almomani

Outlines

- Getting Familiar with Data
- Data Preparation and Transformation
- Descriptive statistical data analysis
- Market Segmentation Analysis
- Bookings Status Over Time Analysis
- Average Price and Utility Analysis
- Conclusion

Getting Familiar with Data

This shows the structure of the dataset. It has a *shape* of (36285 rows, 17 columns) with-out any missing values.

11 of them are quantitative variables
(including binary ones) and rest are qualitative variables
(including the identifier and date columns).

```
RangeIndex: 36285 entries, 0 to 36284
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Booking_ID                36285 non-null  object
 1   number of adults          36285 non-null  int64
 2   number of children        36285 non-null  int64
 3   number of weekend nights  36285 non-null  int64
 4   number of week nights     36285 non-null  int64
 5   type of meal              36285 non-null  object
 6   car parking space         36285 non-null  int64
 7   room type                 36285 non-null  object
 8   lead time                 36285 non-null  int64
 9   market segment type       36285 non-null  object
 10  repeated                  36285 non-null  int64
 11  P-C                       36285 non-null  int64
 12  P-not-C                   36285 non-null  int64
 13  average price             36285 non-null  float64
 14  special requests          36285 non-null  int64
 15  date of reservation       36285 non-null  object
 16  booking status            36285 non-null  object
dtypes: float64(1), int64(10), object(6)
memory usage: 4.7+ MB
```

Data Preparation and Transformation

In this phase we cast the binary integer variable into categorical variables as well as checking for any duplicated row or missing values.

Then, we converted the *date of registration* attribute into standard pandas datetime object for consistency, since it has multiple formats, and easier use. And any wrong entered date (e.g. 2018/2/29) is discarded.

Warning

For probability of cancellation and not cancellation (**P-C and P-not-C**) values are inconsistent (e.g. the summation of the corresponding values should be equal to 100% which is not hold) and for this purpose they will be discard from this analysis.

```
df.dropna(inplace=True)
df.isna().sum().sum(), df.duplicated().sum()
```

```
(0, 0)
```

Descriptive statistical data analysis

Well, that reveals a lot!
First of all, the **number of adults** are ranging from 0 to 4 adults with a majority of 2 adults in each booking observation.

For the **number of childrens**, the maximum number of childrens were 10 childrens and almost of bookings were limited to adults (no childrens).

Weekend stays typically range from 0 to 2 nights, with some extending up to 7 nights and the same for **number of week nights**, it stays range from 0 to 3 with some exceptions going up to 17 nights.

	count	mean	std	min	25%	50%	75%	max
number of adults	36248.0	1.845178	0.518671	0.0	2.0	2.00	2.0	4.0
number of children	36248.0	0.105302	0.402597	0.0	0.0	0.00	0.0	10.0
number of weekend nights	36248.0	0.810445	0.870938	0.0	0.0	1.00	2.0	7.0
number of week nights	36248.0	2.204508	1.410825	0.0	1.0	2.00	3.0	17.0
lead time	36248.0	85.282360	85.961536	0.0	17.0	57.00	126.0	443.0
P-C	36248.0	0.023339	0.368432	0.0	0.0	0.00	0.0	13.0
P-not-C	36248.0	0.152919	1.753126	0.0	0.0	0.00	0.0	58.0
average price	36248.0	103.435350	35.081308	0.0	80.3	99.45	120.0	540.0
special requests	36248.0	0.620034	0.786429	0.0	0.0	0.00	1.0	5.0

Cont.

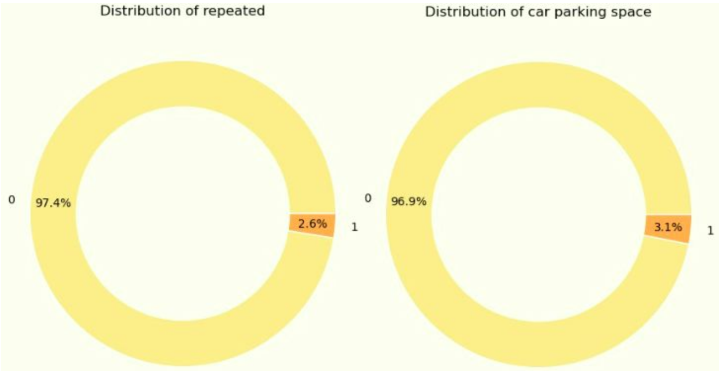
Most bookings, more than half, do not include **special requests**, with some having up to 5 requests.

The **average prices** are normally distributed and they have almost symmetric belled shape, with prices mean of 103.

The median **lead time** is approximately 57 days with a high variability as indicated by a standard deviation of 85.94 days. The values are ranging from 0, indicating some bookings are made for immediate arrivals and 443 days, indicating some bookings made in advance.

Cont.

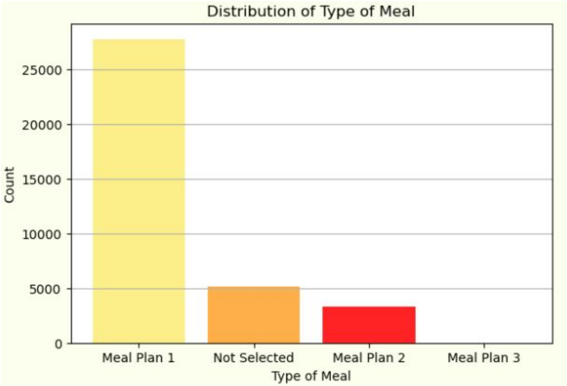
For both, **car parking space** and **repeated**, a very small proportion of 3% were found with a space for car parking and they are not repeated bookings.



repeated		car parking space	
0	0.974454	0	0.969074
1	0.025546	1	0.030926

Cont.

Also, there are 3 primarily **meal plans**, plan 1, 2, and 3.
There is an option for non selected plan.
The majority of 77% bookings were with a plan 1. 14% of them did not select any meal plan.
The last 9% were found with plan 2 and a very small percentage, 0.01, of bookings were booked with plan meal 3.

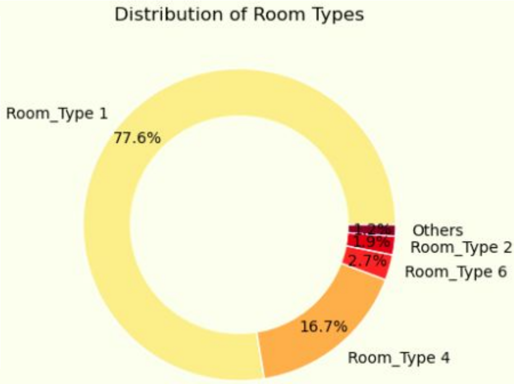


type of meal	
Meal Plan 1	0.767187
Not Selected	0.141553
Meal Plan 2	0.091122
Meal Plan 3	0.000138

Cont.

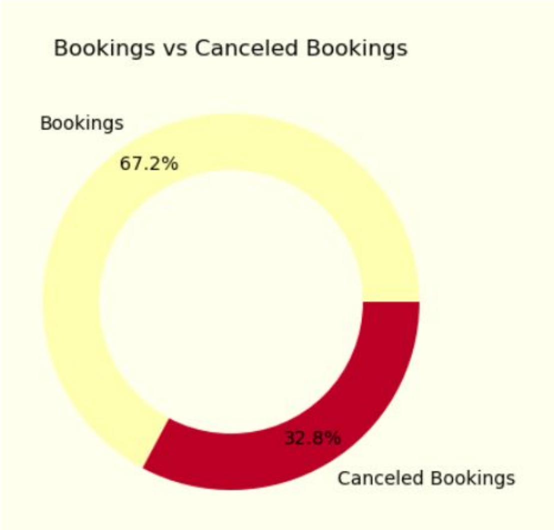
For **type of rooms** there are 7 types. The most popular room type booked was rooms of type 1, with a percentage of 77% and 17% for rooms of type 4. The last 6% was distributed among the room types.

room type	
Room_Type 1	0.775574
Room_Type 4	0.166933
Room_Type 6	0.026595
Room_Type 2	0.019091
Room_Type 5	0.007256
Room_Type 7	0.004359
Room_Type 3	0.000193



cont.

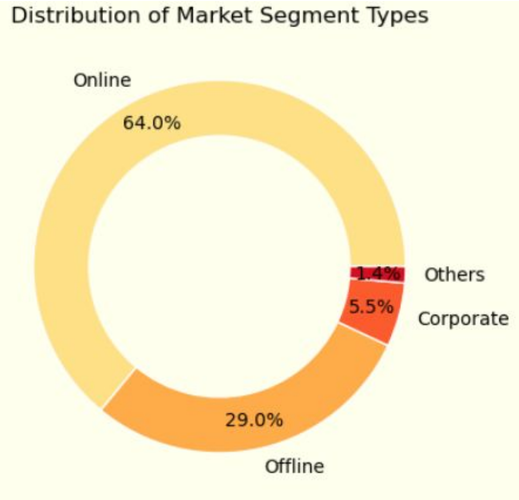
Well about 33% of booking are cancelled in fact it is not a small ratio!




```
booking_status
Not_Canceled     0.672203
Canceled         0.327797
```

cont.

Well, the majority of bookings, 64%, were made through online channels, making it the most popular choice, likely due to their accessibility. A significant portion of bookings, 29%, were made offline. Corporate bookings account for 5.56% of the total. This segment consists of business travelers whose stays are often arranged through company accounts or corporate travel agencies. The last proportion segment, about 1.5%, were for complementary and aviation bookings.

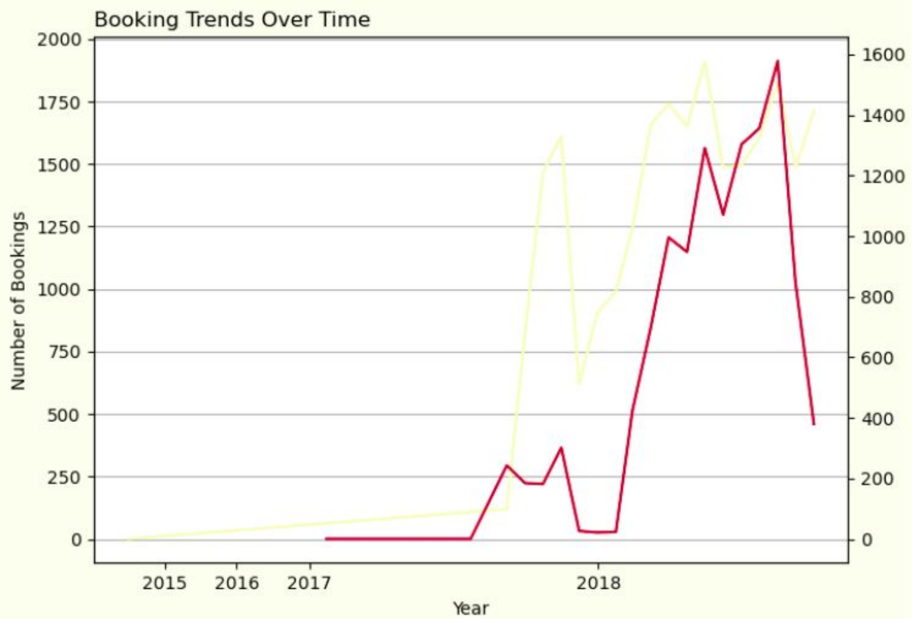


market segment type	
Online	0.640063
Offline	0.290250
Corporate	0.055479
Complementary	0.010759
Aviation	0.003448

Bookings Status Over Time Analysis

Well, about 33% of booking are canceled, in fact it is not a small ratio! so, let's dive deeply in our analysis.

As shown here, there was a relative increasing in the number of booking, from 2015 until mid-2017. Then, a significant increasing in the number of booking until the end of November 2017 which is with sharp decline in the number of bookings until the beginning of 2018. From 2015 there was a relatively stable number of canceled booking until 2018, which get also increased as the number of bookings were increased in general.

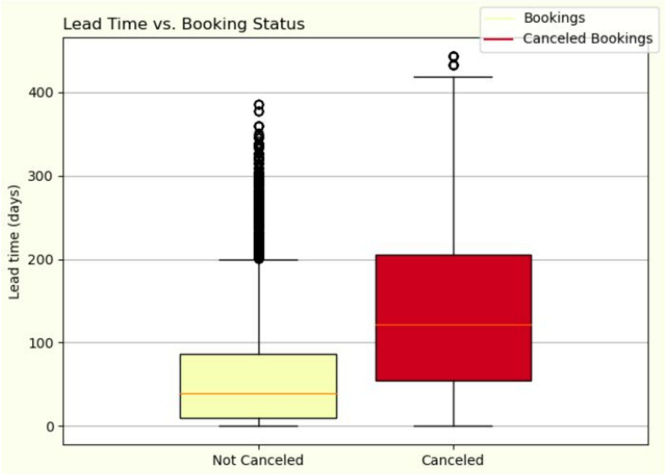


Cont.

In fact, there is a strong positive relationship about 0.79 between the number of canceled bookings and average prices, as the average prices increases the number of booking cancelation is also increases.

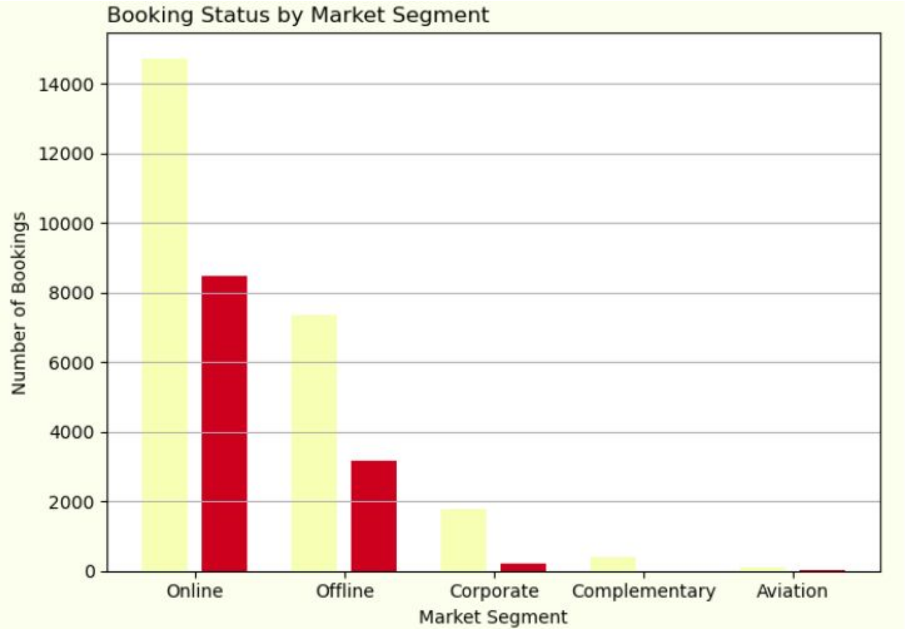
Cont.

In the boxplot above, it is clearly shown that the impact of the duration time on the likelihood of booking to be canceled, since the boxplot that representing the canceled bookings has a bigger range and the data points are spread in that range.

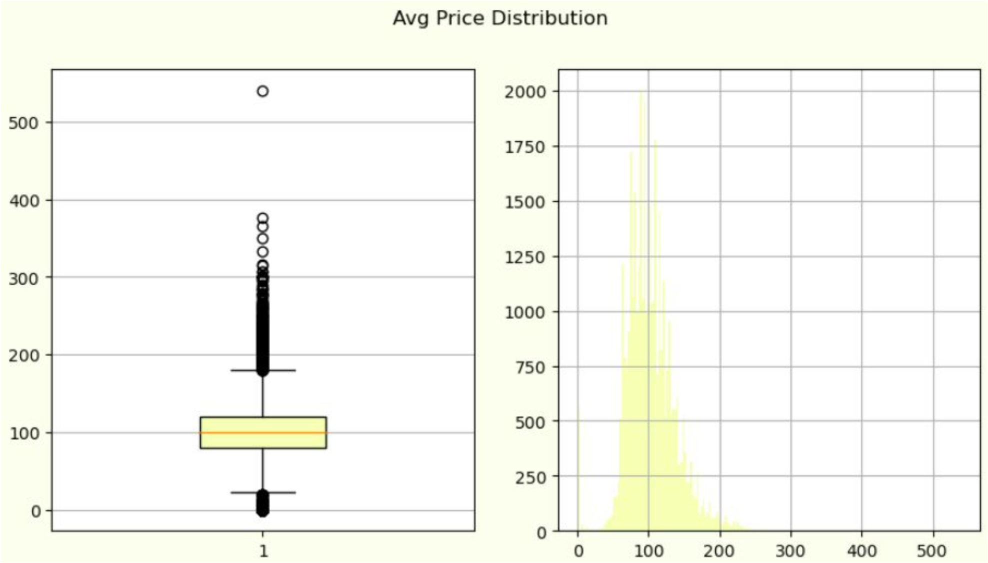


Cont.

Finally, the majority of bookings are done and canceled online platforms, due to their accessibility and its usability than offline or traditional methods.



Average Price Analysis

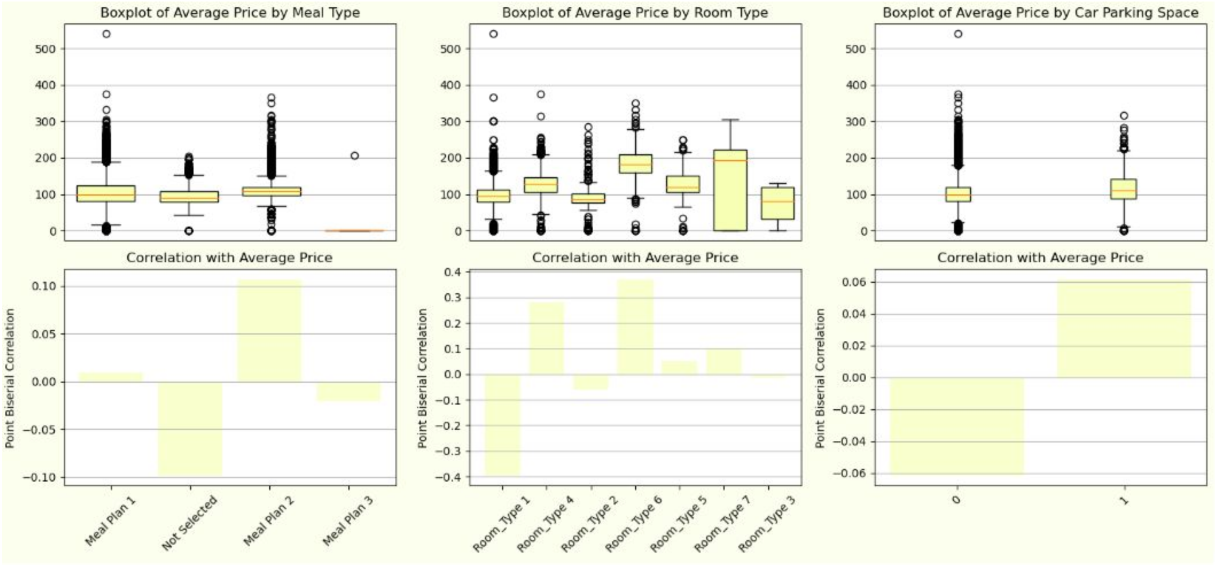


There are multiple factors that can affect the booking price, including length of stay, type of meal, the special requests and more.

The average prices are normally distributed from 30 to 180 with relatively small outliers with prices up to 500.

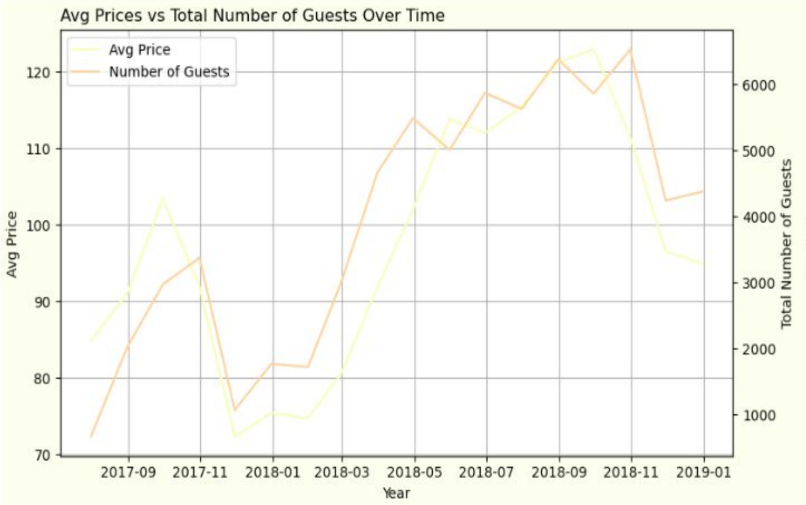
Utility Analysis

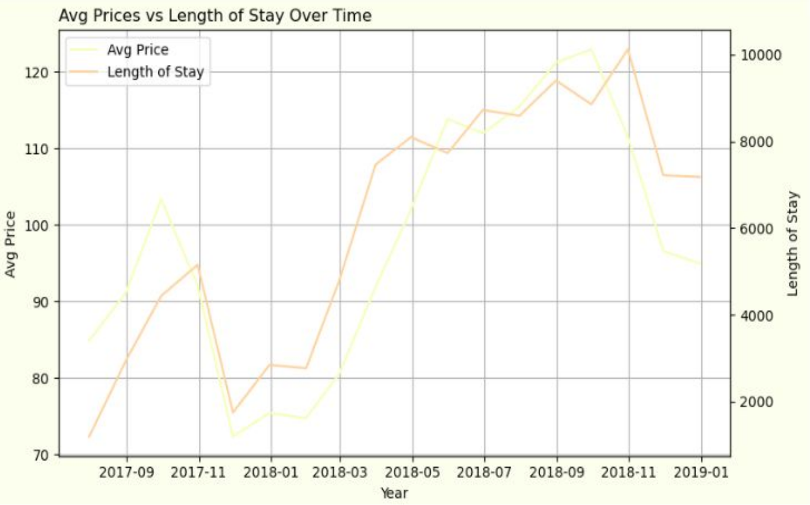
The variability in data may due to request some special utilization such as requesting for space car parking, highly cost meal plan or any something else which will explained in the following chart.



Cont.

As we seen before, there was a disruption in average booking prices from mid-2017 to the end of 2018. So, let's turn our focus in this interval:





Cont.

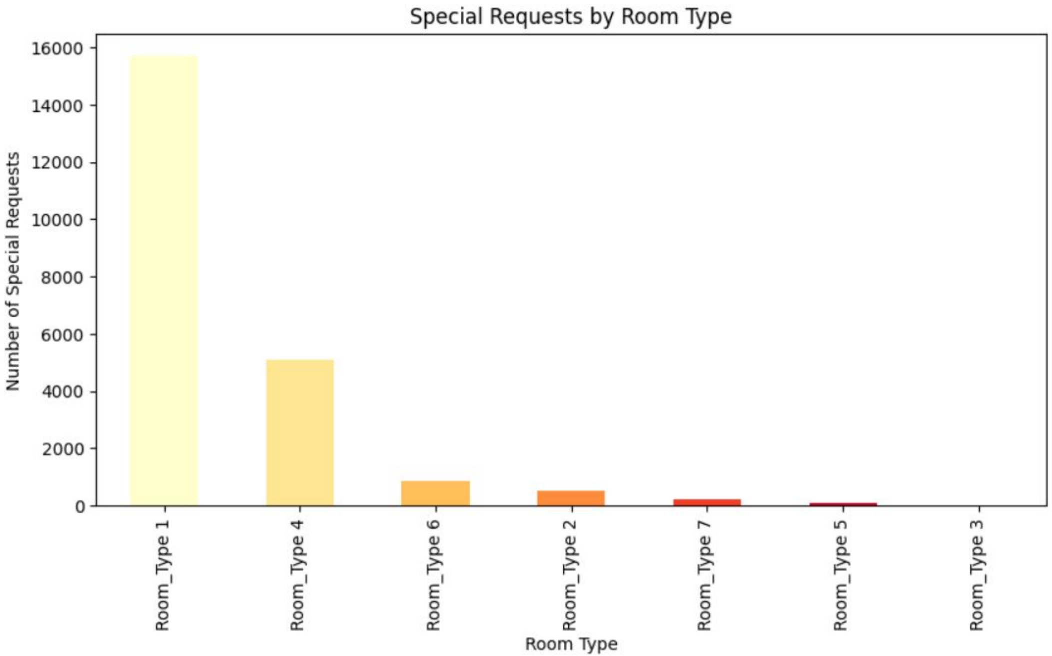
Well, maybe these results are not very exciting since it is obvious that the more people in the booking, the more staying time, the more the cost!

But from this chart we can ensure that the primary factor on the price is both, the number of bookers and stay time:

Correlation coefficient between average prices and Total Number of Bookers between mid of 2017 and 2018: **0.87**

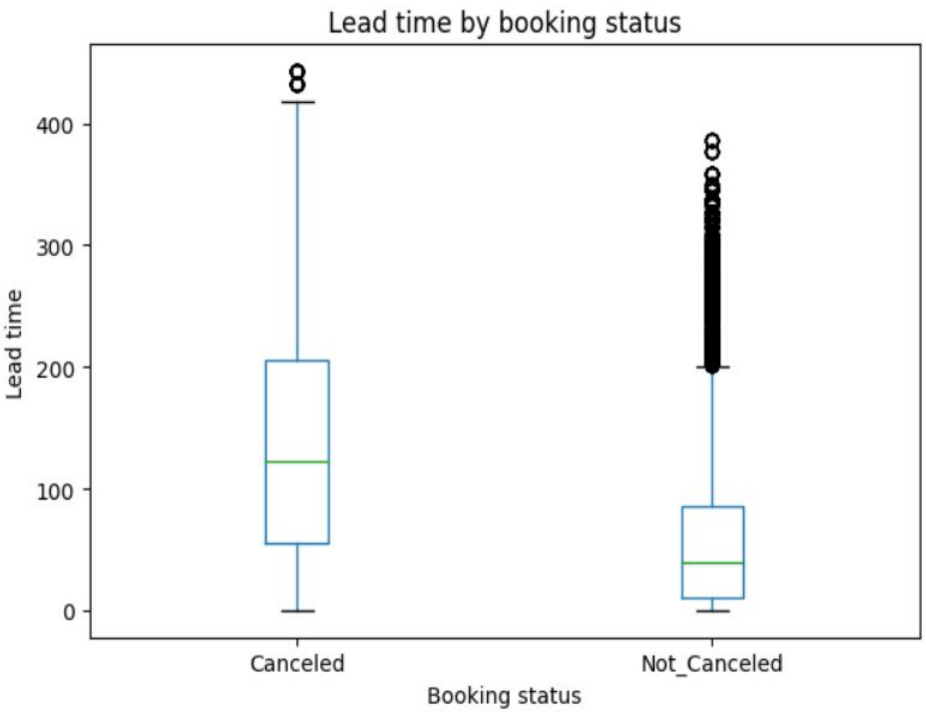
Correlation coefficient between average prices and Staying Period between mid of 2017 and 2018: **0.85**

Special Requests By Room Types



The bar chart analyzes special requests by room type. Room type 1 has the most requests, followed by room types 4 and 2. This suggests guests in room type 1 frequently request additional services or amenities. By understanding these preferences, the hotel can tailor its offerings to better serve guests in different rooms.

Lead Time By Booking Status



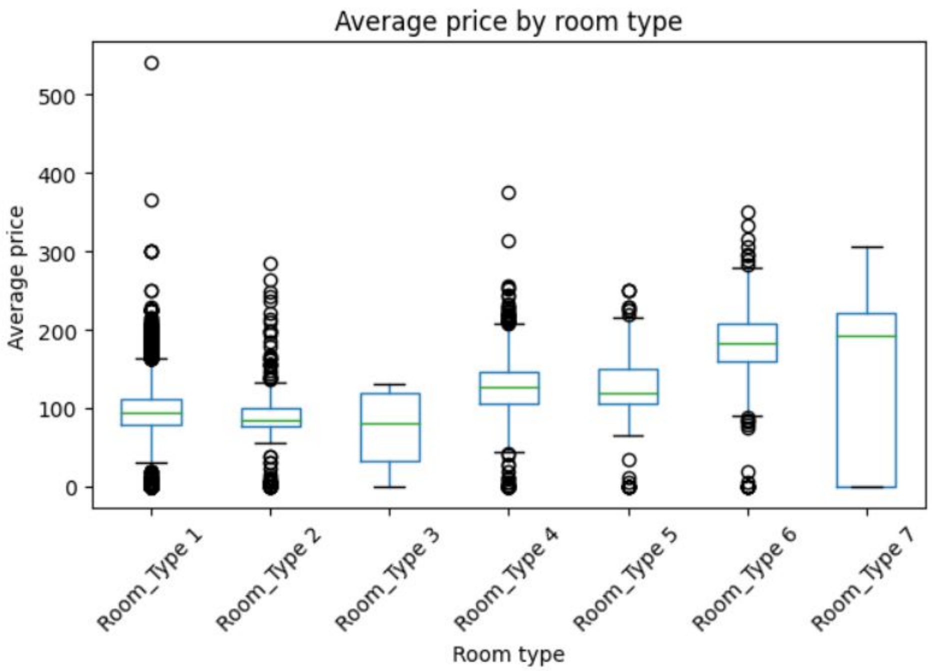
Lead Time by Status: Canceled bookings tend to have longer lead times (median around 100 days) compared to non-canceled bookings (median around 50 days).

Spread of Lead Time: Canceled bookings show a wider range of lead times, with some exceeding 300 days. Non-canceled bookings have a tighter spread.

Lead Time and Cancellation: The data suggests a possible link between longer lead times and higher cancellation rates. Customers might change plans over extended booking windows.

Actionable Insights: This information can help hotels understand booking patterns and anticipate cancellations, potentially leading to improved policies.

Average Price By Room Type



Median Price Differences: Room types have distinct median prices. For example, Room_Type 4 is pricier than Rooms 1 and 2 on average.

Price Range within Types: The spread of prices within a type varies. Room_Type 4, for instance, shows a wider range, suggesting more price flexibility.

Potential Pricing Factors: These variations likely reflect room size, amenities, location, or demand influencing room prices.

Pricing Exceptions: Outliers exist, potentially due to special offers, discounts, or specific booking situations.

This analysis helps hotels understand their pricing strategy, identify potential improvements, and ensure room rates are competitive and aligned with customer expectations.

Booking Status By Market Segment

Insights:

The proportion of canceled bookings is higher than not canceled bookings for all market segments except for the Aviation market segment.