## **Data Analysis of Hotel Reservations Dataset**

+

# **Introduction** :

The online hotel reservation channels have dramatically changed booking possibilities and customers’ behavior. A significant number of hotel reservations are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with.

**About this file**

The file contains the different attributes of customers' reservation details. The detailed data dictionary is given below.

**Data Dictionary**

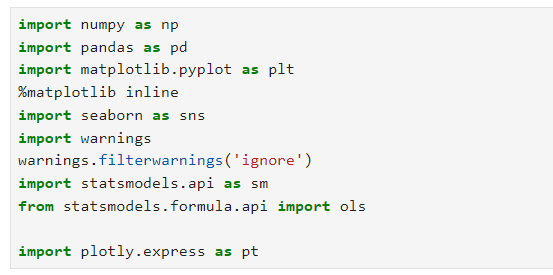
* Booking\_ID: unique identifier of each booking
* no\_of\_adults: Number of adults
* no\_of\_children: Number of Children
* no\_of\_weekend\_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* no\_of\_week\_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
* type\_of\_meal\_plan: Type of meal plan booked by the customer:
* required\_car\_parking\_space: Does the customer require a car parking space? (0 - No, 1- Yes)
* room\_type\_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
* lead\_time: Number of days between the date of booking and the arrival date
* arrival\_year: Year of arrival date
* arrival\_month: Month of arrival date
* arrival\_date: Date of the month
* market\_segment\_type: Market segment designation.
* repeated\_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
* no\_of\_previous\_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
* no\_of\_previous\_bookings\_not\_canceled: Number of previous bookings not canceled by the customer prior to the current booking
* avg\_price\_per\_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
* no\_of\_special\_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
* booking\_status: Flag indicating if the booking was canceled or not.

# Business question :

Can you predict if the customer is going to honor the reservation or cancel it ?

**Introduction to project:**

1. Importing libraries



1. Read file and Data Checks to Perform

**df = pd.read\_csv("Hotel Reservations.csv")**

* By using ”pd.read\_csv” we can read the file in python.
* After that we can see data types, head and tail of the data set.
* By using “df.duplicated().sum()” we can check is there any duplicated values in the data set.

1. Checking outliers and replace outliers with the median

df.describe().transpose()

· df.describe(): This method provides summary statistics for numeric columns in the DataFrame, including count, mean, standard deviation, min, 25th percentile, 50th percentile (median), 75th percentile, and max.

· .transpose(): This method flips the DataFrame so that rows become columns and vice versa, making it easier to read the statistics for each column.

From the data,

we found that “**avg\_price\_per\_room”**  column has a outlier.

\* The median is 99.45, suggesting that half of the room prices are below this value, which may indicate a skew in the data.

\* The standard deviation of 103.42 shows a moderate level of variability in room prices.

\* The variance of 35.09 confirms that there are some fluctuations in the pricing data.

\* The price range is substantial, from 0.0 to 540.0. This wide range suggests significant differences in room pricing, indicating diverse property types.

\* The minimum value of 0.0 may indicate a potential data entry error or special cases that could skew the analysis. It might be worth investigating further.

One common way to find outliers in a dataset is to use the inter-quartile range.

The inter-quartile range, often abbreviated IQR, is the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset. It measures the spread of the middle 50% of values.

The lower limit is calculated as:

Lower limit = Q1 – 1.5\*IQR

And the upper limited is calculated as:

Upper limit = Q3 + 1.5\*IQR

4) Data Analysis and Visualization :

**BOOKING STATUS**

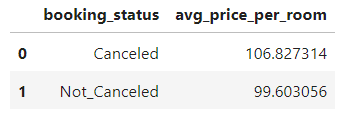
Based on “**df['booking\_status'].value\_counts()**” we counted the bookings status of that particular reservation. And we ploted the pie chart.



Cancellation rate is pretty high **11,885 of 36,275** bookings or **32.8% bookings is canceled**. Unfortunately, there is no data that can explain why the customers cancel their bookings. So we will looking that reason based on dataset using Exploratory Data Analysis method.

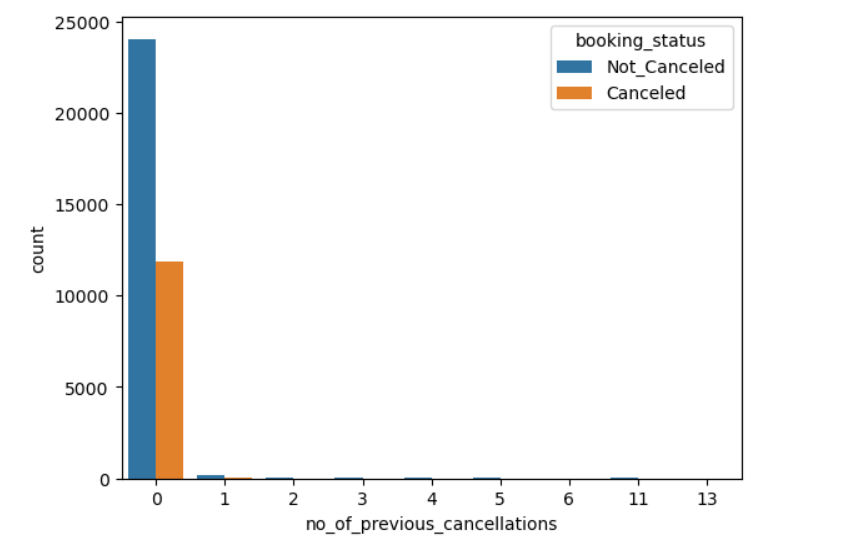
**AVERAGE PRICE PER ROOM**

**df.groupby('booking\_status')['avg\_price\_per\_room'].mean().reset\_index()** in pandas is used to calculate the average price per room, grouped by the booking status. This command is useful for analyzing how average room prices vary with different booking statuses, helping in decision-making and strategy formulation.



Canceled Average Price per Room is Higher than Not Canceled (106.827 > 99.6031).

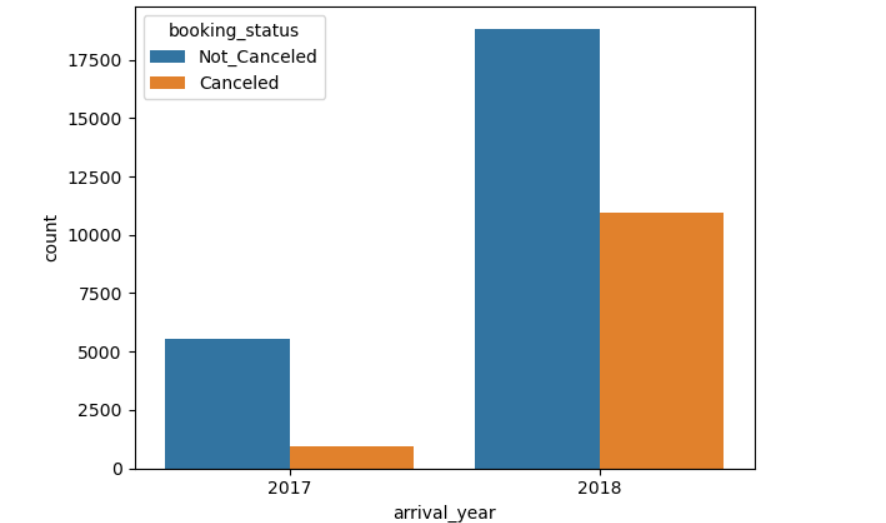
**PREVIOUS CANCELATION WITH RESPECT TO BOOKING STATUS**



This plot helps you understand how the number of previous cancellations relates to booking statuses, providing insights into customer behavior and potential risk factors in booking patterns.

* Higher lead\_time Leads to Higher Cancelation Rate.

**Count of Arrivals by Year and Booking Status**



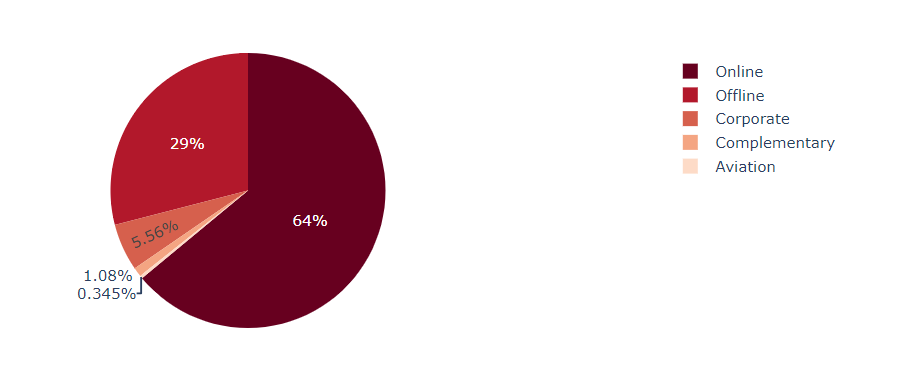
The number of bookings for each arrival year, split by booking status using different colors. In 2018 it registered more number of visiters.

**ROOM TYPE RESERVED WITH BOOKING STATUS**

This plot helps you understand how different room types are associated with various booking statuses, providing insights into customer preferences and potential booking trends.

* The cancellation rate is higher for the guests who reserved Room type 6 and type 1 compare to others.
* The cancellation rate is higher for the guests who don't have a special request compare to others.
* If Special Request is More than 2, booking\_status Will not be Canceled.

**MARKET SEGMENT TYPE**



**market\_segment\_type**

Online 23214

Offline 10528

Corporate 2017

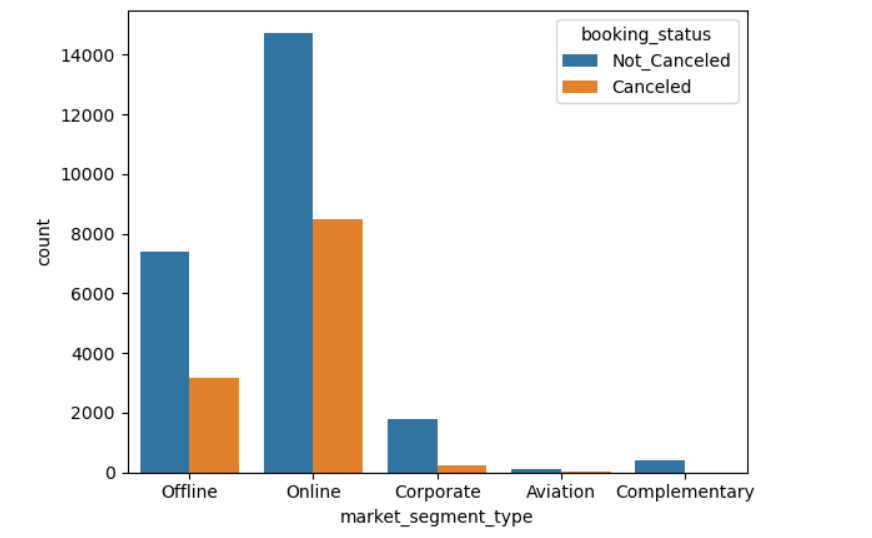
Complementary 391

Aviation 125

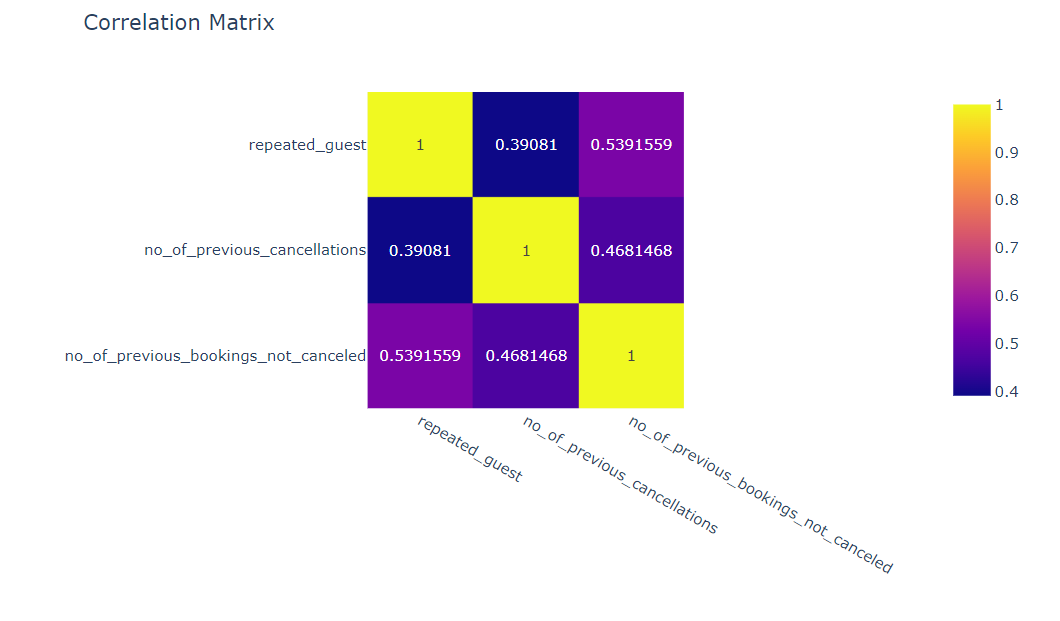
The pie chart will visually represent the proportions of each market segment, highlighting the dominance of the "Online" segment compared to the others.

**MARKET SEGMENT TYPE WITH BOOKING STATUS**

The command **df.groupby(["booking\_status", "market\_segment\_type"]).size()** is used to count the number of occurrences of each combination of booking\_status and market\_segment\_type in the DataFrame df. This command is useful for analyzing the relationships between booking statuses and market segments, helping identify trends or areas needing attention in your data analysis.



* High Cancellations in Online Segment: The Online market segment has the highest cancellations (8,475), suggesting potential issues with customer satisfaction or booking processes in this channel.
* Stable Corporate Bookings: The Corporate segment shows a healthy ratio of cancellations (220) to active bookings (1,797), indicating reliability among corporate clients.
* Offline bookings also show substantial counts, but with lower cancellations compared to Online (3,153 vs. 7,375), suggesting that customers may prefer direct engagement.
* Low Counts in Aviation: Both Canceled (37) and Not\_Canceled (88) for Aviation are low, indicating limited demand or niche market characteristics.
* Moderate Cancellations: The Complementary segment has a cancellation count of 391, which could reflect non-essential bookings.



#### repeated\_guest and no\_of\_previous\_cancellations:

Correlation: 0.39: This indicates a moderate positive correlation. As the number of previous cancellations increases, the likelihood of being a repeated guest also increases.

#### repeated\_guest and no\_of\_previous\_bookings\_not\_canceled:

Correlation: 0.54: This shows a stronger positive correlation, suggesting that repeated guests tend to have more previous bookings that were not canceled.

#### no\_of\_previous\_cancellations and no\_of\_previous\_bookings\_not\_canceled:

Correlation: 0.47: This indicates a moderate positive correlation. More previous cancellations are associated with fewer bookings that were not canceled.

#### Summary

Significant Relationships: The correlations indicate that being a repeated guest is associated with both a higher number of previous bookings not canceled and a moderate number of previous cancellations.

[ ]:

# Statistical Analysis :

Statistical analysis means investigating trends, patterns, and relationships using [quantitative data](https://www.scribbr.com/methodology/quantitative-research/). It is an important research tool used by scientists, governments, businesses, and other organizations.

1. Two sample Independent T test
2. Two way ANOVA
3. CHI - SQUARE TEST
4. **Two sample Independent T test**

A two-sample t-test always uses the following null hypothesis:

* H0: μ1 = μ2 (the two population means are equal)
* The alternative hypothesis can be either two-tailed, left-tailed, or right-tailed:
* H1 (two-tailed): μ1 ≠ μ2 (the two population means are not equal)
* H1 (left-tailed): μ1 < μ2 (population 1 mean is less than population 2 mean)
* H1 (right-tailed): μ1> μ2 (population 1 mean is greater than population 2 mean)

1. **Two way ANOVA**

ANOVA (Analysis of Variance) is a [statistical test](https://www.scribbr.com/statistics/statistical-tests/) used to analyze the difference between the means of more than two groups.

A two-way ANOVA test is a statistical technique that analyzes the effect of the independent variables on the expected outcome along with their relationship to the outcome itself.

1. **CHI - SQUARE TEST**

The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data. It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

Two way ANOVA,

1. The interaction among "no\_of\_adults", "no\_of\_children", "avg\_price\_per\_room"

p-value (PR(>F)): 0.402471, which is greater than 0.05. This suggests that the interaction between the number of adults and children does not significantly impact avg\_price\_per\_room.

1. The interaction among room type,average price per room and booking status

* The interaction between room type and booking status is significant, suggesting that the effect of room type on average price varies by booking status.
* Both room type and booking status have significant effects on avg\_price\_per\_room.
* There is also a significant interaction effect, meaning the relationship between room type and price is influenced by the booking status.

1. The interaction among "market\_segment\_type", "booking\_status", "avg\_price\_per\_room"

* The interaction between market segment type and booking status is significant, suggesting that the effect of market segment type on average price varies by booking status.
* Significant Effects: Both market segment type and booking status significantly affect avg\_price\_per\_room.
* Interaction Effect: There is a significant interaction effect, indicating that the impact of market segment type on price depends on the booking status.

Two sample Independent T test

Fail to reject the null hypothesis at alpha = 0.05

CHI - SQUARE TEST

Low p-value, we can conclude that there is a significant association between room type reserved and Booking status. In other words, the distribution of one variable depends on the levels of the other variable.