

# Analysis of Hotel Bookings & Cancellation patterns

## Data Intensive Computing 587 PHASE 1

Manikanta Kalyan Gokavarapu - mgokavar - 50465129

Rakesh Kumar Gavara - rgavara - 50483851

## 1. Problem Statement

The aim of this project is to detect patterns and tendencies in customer actions that can be utilized to improve revenue management, enhance resource allocation, optimize customer service, and shape marketing and promotional strategies. Through an analysis of historical booking and cancellation data, the goal is to develop precise forecasts about future demand and make necessary operational adjustments to meet customer requirements and maximize profitability.

### 1.1 Background

Hotel bookings is one of the most dynamic flows in the market i.e. it is quite difficult to understand trends in bookings. This might be due to a number of factors such as climate factors, region, population, hotel reviews and so on. Some of the customers tend to cancel their bookings, reschedule as well due to various reasons such as change of travel plans, last minute pop up events and so on. It would either result in overbooking or underbooking, understaffing or overstaffing. Cancellations and dynamicity often incur loss to hotels.

### 1.2 Significance

Understanding hotel booking and cancellation patterns can be significant for several reasons:

- **Revenue management:** Hoteliers can use this data to optimize their revenue management strategies by identifying the times when demand is highest and adjusting their rates accordingly. They can also offer discounts or promotions during the times when demand is lower to attract more customers.
- **Staff planning:** Hotels can use the data on booking patterns to forecast how many staff members they need at any given time. This can help them schedule employees more efficiently and avoid over- or under-staffing, which can impact customer service.

- **Resource allocation:** Knowing the booking patterns can help hotels allocate resources such as room inventory, housekeeping, and food and beverage services more efficiently. For example, if a hotel sees a surge in bookings for a particular weekend, they can ensure that they have enough staff and supplies on hand to meet the increased demand.
- **Marketing and promotions:** Hotels can use booking and cancellation data to identify trends and preferences among their customers. This can help them tailor their marketing and promotional efforts to target specific demographics or customer segments.
- **Forecasting:** By analyzing past booking and cancellation patterns, hotels can make more accurate predictions about future demand. This can help them plan ahead and make informed decisions about pricing, inventory, staffing, and marketing.

**Summary:** If we gain proper intelligence from the data we can very well come up with strategies to improve the hotel profits. In order to do so we can use Exploratory data analysis techniques such as Visualisation and Statistics to understand the data. Some of the focus areas involved here are understanding at what time of the year the bookings are high, which sector of the people do the regular bookings, why there are more cancellations and so on. Once we gain intelligence from the data we can develop prediction models to take strategic decisions to meet the demands and improve sales.

## 2. Data Sources.

### 2.1 Data Source References

Originally the datasource is from the website 'Sciencedirect' - Data in Brief , Volume 22, February 2019, Pages 41-49 authored by Nuno Antonio, Ana de Almedia and Luis Nunes.

**Link :** <https://www.sciencedirect.com/science/article/pii/S2352340918315191>

Later this dataset was modified by jthomasmock in tidy tuesday on 11-02-2020. For our analysis we have used the raw data that is available in the following link in the github.

**Link:**

<https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-11/hotels.csv>

## 2.2 Dataset Description

The Dataset contains 32 columns and 119000 columns.

The columns are

1. Hotel: The name of the hotel or resort.
2. Is\_canceled: A binary column that indicates whether the reservation was canceled (1) or not (0).
3. Lead\_time: The number of days between the booking date and the arrival date.
4. Arrival\_date\_year: The year of the arrival date.
5. Arrival\_date\_month: The month of the arrival date.
6. Arrival\_date\_week\_number: The week number of the arrival date.
7. Arrival\_date\_day\_of\_month: The day of the month of the arrival date.
8. Stays\_in\_weekend\_nights: The number of weekend nights (Saturday or Sunday) the guest stayed.
9. Stays\_in\_week\_nights: The number of weekday nights (Monday to Friday) the guest stayed.
10. Adults: The number of adults in the reservation.
11. Children: The number of children in the reservation.
12. Babies: The number of babies in the reservation.
13. Meal: The type of meal booked. (e.g. Breakfast, Half board, Full board, etc.)
14. Country: The country of origin of the guest.
15. Market\_segment: The market segment the booking was made from (e.g. Direct, Online Travel Agent, Corporate, etc.).
16. Distribution\_channel: The distribution channel through which the booking was made (e.g. Direct, Travel Agent, Online Travel Agent, etc.).
17. Is\_repeated\_guest: A binary column that indicates whether the guest has stayed at the hotel before (1) or not (0).
18. Previous\_cancellations: The number of previous cancellations the guest has made.

19. Previous\_bookings\_not\_canceled: The number of previous bookings the guest has made that were not canceled.
20. Reserved\_room\_type: The type of room reserved by the guest.
21. Assigned\_room\_type: The type of room assigned to the guest upon arrival.
22. Booking\_changes: The number of changes made to the reservation.
23. Deposit\_type: The type of deposit made for the reservation.
24. Agent: The ID of the travel agent through which the booking was made.
25. Company: The ID of the company or entity that made the booking.
26. Days\_in\_waiting\_list: The number of days the booking was on the waiting list before it was confirmed.
27. Customer\_type: The type of customer that made the booking (e.g. Transient, Group, Contract, etc.).
28. ADR: The average daily rate, which is the total revenue divided by the number of days stayed.
29. Required\_car\_parking\_spaces: The number of parking spaces required by the guest.
30. Total\_of\_special\_requests: The total number of special requests made by the guest (e.g. extra towels, late check-out, etc.).
31. Reservation\_status - This feature implies one of the following : Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in.
32. Reservation\_status\_date - Date at which last reservation status is done.  
To determine when the reservation was canceled or the client checked out of the hotel, we can use this variable in association with the ReservationStatus.

### 3. Data Cleaning/Processing.

#### 3.1 Checking and Handling Null/Missing/NAN values

- In this step we find the missing/null values in each column and we apply a solution to the column based on the null value percentages.
- We have found the no.of null values and null value percentages for all the columns and following features have null values and their null value percentages are given in the Figure [1].

children	4	0.003350
babies	0	0.000000
meal	0	0.000000
country	488	0.408744
agent	16340	13.686238
company	112593	94.306893

Figure [1].

- As the columns 'agent' and 'company' have a large number of missing/null values we drop 'agent' and 'company' columns.
- As the columns 'country' and 'children' have less number of missing/null values we just remove the missing value rows instead of whole columns.

#### 3.2 Checking and Handling Duplicate values

- In this step we find the duplicated rows in the dataset and retain only one unique row in the duplicated rows and remove all other repeated rows.
- In the dataset we can see that there are in total **31984** duplicated rows as shown in the figure [2].

[31984 rows x 30 columns]

Figure [2].

- So we have dropped the duplicate rows and retained only the one unique row in the dropped duplicate rows.

### 3.3 Data Type Conversions.

- In this step we find the datatypes of all the features and convert the datatype of a feature to appropriate format.
- In the Dataset we could notice that only the reservation\_status\_date is in the wrong format. It is in object format , but needs to be in data and time format. You could find the same in figure [3].

```
29 reservation_status_date      86914 non-null object
```

**Figure [3].**

- We have converted it to data and time format as show in the figure [4]

```
28 reservation_status_date      86753 non-null datetime64[ns]
```

**Figure [4]**

### 3.4 Encoding Categorical Data

- In this step we encode the categorical features and convert them to numerical values. This is done because it is easy to draw relationships and patterns for the numerical data. To achieve this we can use one-hot encoding or label-encoding, I have used label-encoding in this.
- There were many categorical variables in the dataset but we have encoded only for the needed features.
- The first encoded feature is 'hotel' where we have 2 unique values i.e, City, Resort I have encoded them to binary format using label encoding . So, encodings are City Hotel - 0, Resort Hotel - 1. The unique values of the hotel column after applying label encoding is given in the figure [5].

```
[1 0]
```

**Figure [5]**

- The second encoded feature is Arrival\_date\_month. We have converted the month variable to values using a dictionary mapping. After encoding the Arrival\_data\_month feature's previous and after transformation unique values are given in the figure [6]

```
['July' 'August' 'September' 'October' 'November' 'December' 'January'  
 'February' 'March' 'April' 'May' 'June']
```

---

```
[ 7  8  9 10 11 12  1  2  3  4  5  6]
```

**Figure [6]**

- The third encoded feature is market\_segment. I have encoded the feature using label encoding and the encoding is done as 0 -> Aviation, 1 -> Complementary, 2 -> Corporate, 3 -> Direct, 4 -> Groups , 5 -> Offline TA/TO, 6 -> Online TA.
- The unique values of market\_segment feature before and after encoding is given in the figure [7]

```
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'  
'Aviation']
```

---

```
[3 2 6 5 1 4 0]
```

**Figure [7]**

- The fourth encoded feature is meal. I have encoded the feature using label encoding and the encoded values are as follows: 0-> BB, 1 -> FB, 2 -> HB, 3 -> SC, 4 -> Undefined.
- The unique values of the meal feature before and after encoding is given in the figure[8].

```
['BB' 'FB' 'HB' 'SC' 'Undefined']
```

```
[0 1 2 3 4]
```

**Figure [8]**

### 3.5 Remove Unwanted Columns

- In this step we usually remove the unwanted/unnecessary columns in our dataset for our analysis.
- As 'marketing segment' and 'distribution channel' columns implying the same data we drop the 'distribution\_channel' column.
- After dropping distribution\_channel Initial dataset's 30 columns are reduced to 29 columns as shown in the figure[9] below.

---

```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 86914 entries, 0 to 119389  
Data columns (total 29 columns):
```

**Figure [9]**

### 3.6 Handling Inconsistencies in the data.

- In this step we need to Identify and handle inconsistencies and errors in the data for example by correcting typos, or resolving conflicts in the data.
- So, in the dataset - 'adult', 'babies', 'children', all the columns at once cannot be zero because for every booking there must be at least a single occupant.
- I have filtered the rows with all three attributes equal to zero and removed the filtered rows from the dataset
- The before and after filtering the inconsistent data is given in the below figure [10]. Total 161 rows were removed.

(86914, 29)

(86753, 29)

**Figure [10]**

### 3.7 Data Integration

- In this step we need to Integrate data by matching and merging data based on a common identifier or key variable.
- In our dataset babies and children nearly imply the same thing and we cannot get more insights by having two separate columns. So, we have combined 'babies' and 'children' columns and created a new column 'Kids'. With the new column we can get better insights.
- In our dataset we don't have the total occupants for a booking. So, we have created a new column 'Guests' by combining 'adults' and 'Kids'. This new column will give various insights like we can tell whether 'Guests' column influences the cancellation of a booking or not. The new columns are given in the figure11.

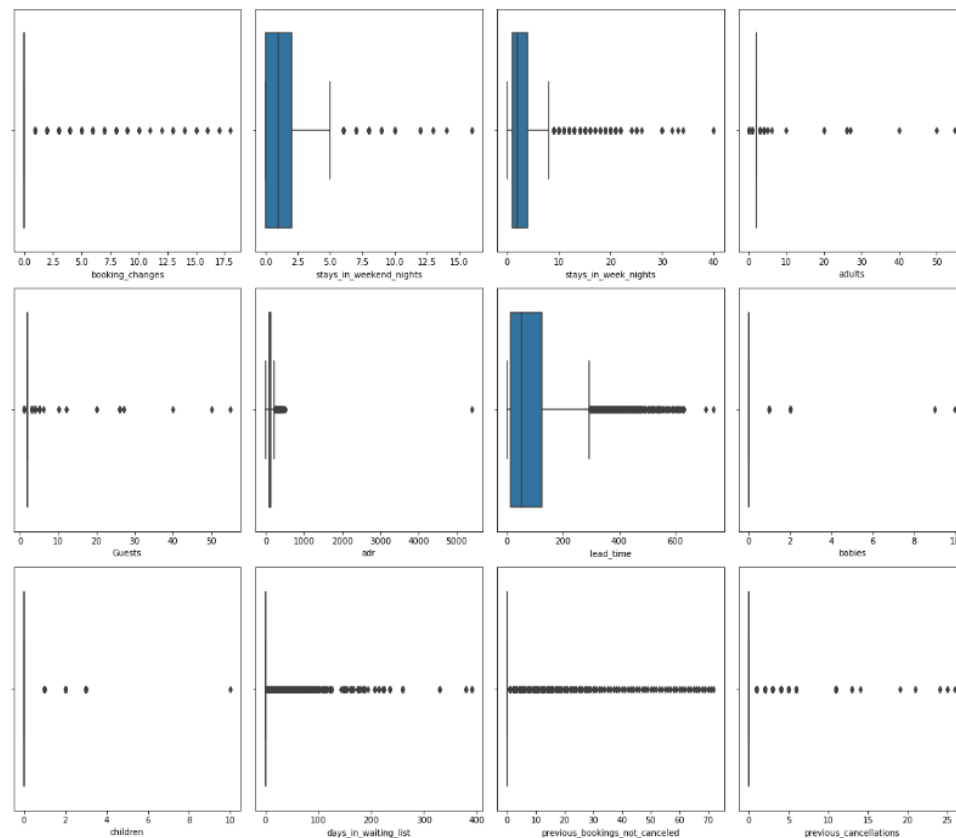
Kids	Guests
0.0	2.0
0.0	2.0
0.0	1.0
0.0	1.0
0.0	2.0

**Figure [11]**



### 3.8 Handling Outliers in the data.

- In this step we identify the outliers in the data and then we remove them or replace them with appropriate values or transform the data.
- To find the outliers, first I have found the continuous and categorical variables in the dataset by considering the unique values in each feature. If there are more than 13 unique values in the feature we consider the feature as 'Continuous feature' and if there are less than or equal to 13 unique features we consider it as 'Categorical feature'. There were in total 13 Continuous features and 18 Categorical features.
- Then I have plotted the box plots of Continuous variables and printed the Continuous Variable min, max and mean etc values for understanding the outliers.
- From the box plots in Figure [12] and data description there were many outliers in the continuous features.



- For the continuous features 'adults', 'booking\_changes', 'lead\_time', 'previous\_booking\_not\_canceled', 'PrevCancel', 'days\_in\_waiting\_list', 'stays\_in\_weekend\_nights', 'stays\_in\_week\_nights' all these features have outliers, so we have treated those outliers by imputing values based on the max, median and box plot ranges.
- For the categorical features there are some extreme values for some features like babies, children, kids, required\_car\_parking\_spaces so I have imputed those extreme values/outliers with zeros.
- The extreme values/outliers are identified using the below data from categorical features max values as shown in Figure [13]

	Kids	hotel	is_repeated_guest	children
count	86753.000000	86753.000000	86753.000000	86753.000000
mean	0.150346	0.386154	0.038731	0.139511
std	0.473062	0.486869	0.192953	0.457232
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	0.000000	1.000000	0.000000	0.000000
max	10.000000	1.000000	1.000000	10.000000

required_car_parking_spaces	arrival_date_year	babies
86753.000000	86753.000000	86753.000000
0.083548	2016.211900	0.010835
0.280557	0.685937	0.113614
0.000000	2015.000000	0.000000
0.000000	2016.000000	0.000000
0.000000	2016.000000	0.000000
0.000000	2017.000000	0.000000
8.000000	2017.000000	10.000000

Figure [13]

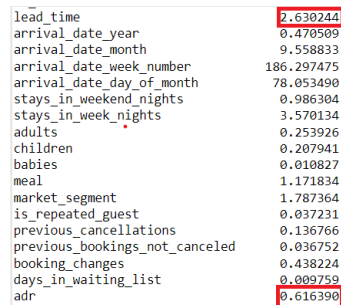
### 3.9 Normalizing the data

- In this step we normalize the data to decrease the variance, scale and increase uniformity in the data.
- To Check whether Normalization required or not for my data I have found the variance of all the features and two features 'lead\_time' and 'adr' have very high variance as shown in the Figure [14]

lead_time	7370.794126
arrival_date_year	0.470509
arrival_date_month	9.558833
arrival_date_week_number	186.297475
arrival_date_day_of_month	78.053490
stays_in_weekend_nights	0.986304
stays_in_week_nights	3.570134
adults	0.253926
children	0.207941
babies	0.010827
meal	1.171834
market_segment	1.787364
is_repeated_guest	0.037231
previous_cancellations	0.136766
previous_bookings_not_canceled	0.036752
booking_changes	0.438224
days_in_waiting_list	0.000750
adr	3007.377580

Figure [14]

- From the above it is evident that lead\_time and adr have very high amounts of variance so to remove this we use log transformation on these columns, this is done because when we apply regression models, these extreme values can have a disproportionate influence on the results. so I am applying the log transformation to reduce the effect/magnitude of the extreme values on the model.
- After applying the log transformation on the high variance features the variance got decreased as shown in the figure 15.



lead_time	2.630244
arrival_date_year	0.470509
arrival_date_month	9.558833
arrival_date_week_number	186.297475
arrival_date_day_of_month	78.053490
stays_in_weekend_nights	0.986304
stays_in_week_nights	3.570134
adults	0.253926
children	0.207941
babies	0.010827
meal	1.171834
market_segment	1.787364
is_repeated_guest	0.037231
previous_cancellations	0.136766
previous_bookings_not_canceled	0.036752
booking_changes	0.438224
days_in_waiting_list	0.009759
adr	0.616390

Figure [15]

### 3.10 Renaming the Columns.

- In this step we rename the columns in a dataset, so it will increase the readability of the data and make it easier to work with.
- In the dataset the column names are not consistent and are not meaningful, so I have renamed the needed columns as below to make them consistent and meaningful.

'arrival\_date\_year': 'arrival\_year', 'arrival\_date\_month': 'arrival\_month', 'arrival\_date\_week\_number': 'arrival\_week\_number', 'stays\_in\_weekend\_nights': 'weekend\_night\_stays', 'stays\_in\_week\_nights': 'week\_night\_stays', 'days\_in\_waiting\_list': 'waiting\_days', 'Kids': 'kids', 'Guests': 'guests', 'PrevCancel': 'prev\_cancel', 'total\_of\_special\_requests': 'special\_requests', 'previous\_bookings\_not\_canceled': 'previous\_uncanceled\_bookings'

- Final Columns after renaming is given in the below figure 16.

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_year', 'arrival_month',
      'arrival_week_number', 'arrival_date_day_of_month',
      'weekend_night_stays', 'week_night_stays', 'adults', 'children',
      'babies', 'meal', 'country', 'market_segment', 'is_repeated_guest',
      'previous_cancellations', 'previous_uncanceled_bookings',
      'reserved_room_type', 'assigned_room_type', 'booking_changes',
      'deposit_type', 'waiting_days', 'customer_type', 'adr',
      'required_car_parking_spaces', 'special_requests', 'reservation_status',
      'reservation_status_date', 'kids', 'guests', 'prev_cancel'],
      dtype='object')
```

Figure [16].

## 4. Exploratory Data Analysis

### 4.1 Bar plot Analysis

- Bar plots represent the visualization of the data between categorical values(x axis) and the data values (y axis). Comparison is done between the discrete values in this analysis.
- In this technique we came up with two plots - 'arrival\_month' vs 'canceled\_bookings' & 'arrival\_month' vs 'repeated\_guests' as below:

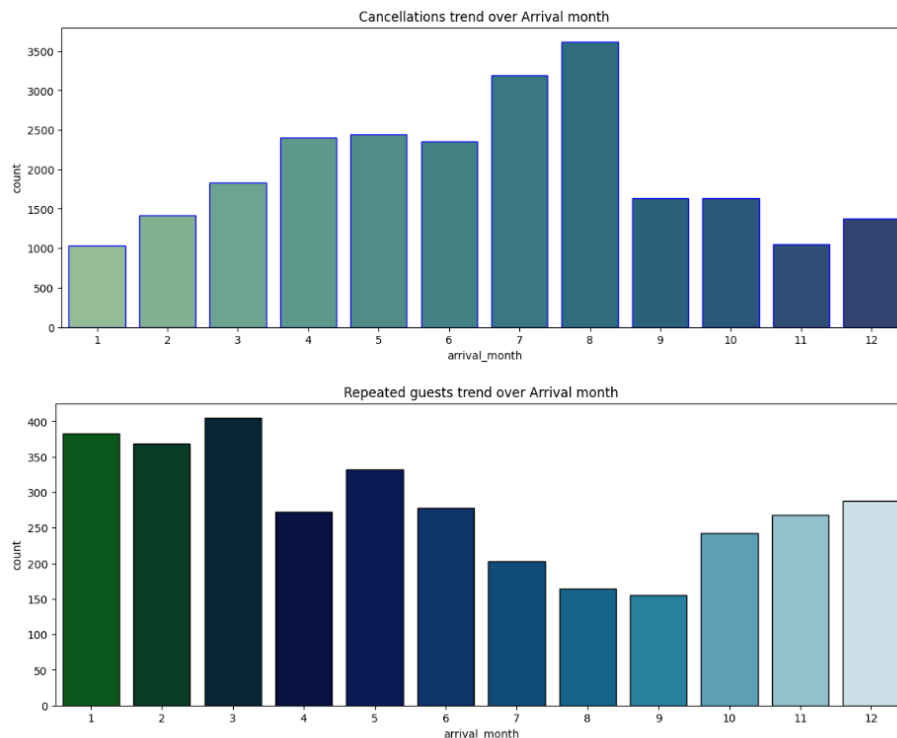


Figure [17]

- Observation1: From the figure 17, first plot we observe that most of the cancellations occur in the month of August.
- Observation2: In the figure 17 second plot we observe that most of the repeated customers have done the bookings in the month of March.
- Observation3: In the figure 17 first plot we observe that November month records minimum cancellations.
- Observation4: In the figure 17 second plot we observe that September month records minimum repeated guests.

## 4.2 Pie Chart Analysis

- Pie charts are circular charts used to represent a single series of data, where the area of the chart reflects the total percentage of the data. The chart is divided into wedges, each representing a percentage of the data.
- In this technique we have quantified the percentage of bookings in various categorized features such as reserved\_room\_type, assigned\_room\_type and customer\_type as below:

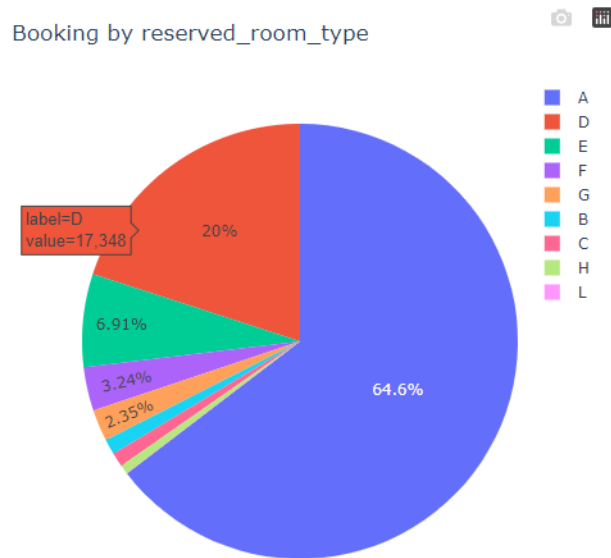


Figure [18]

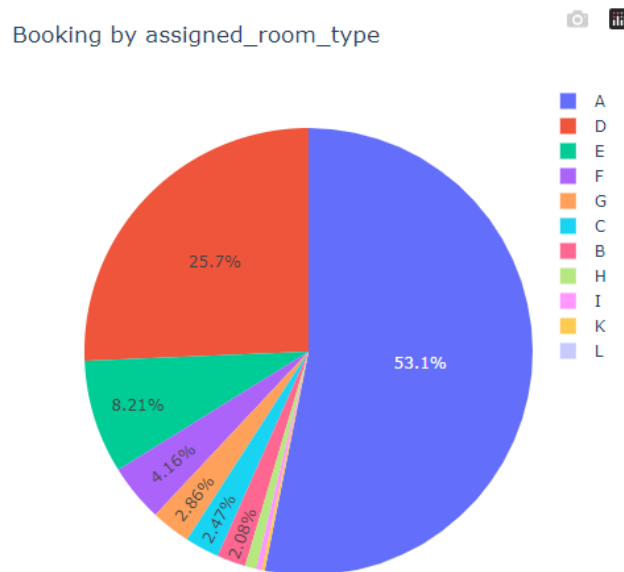
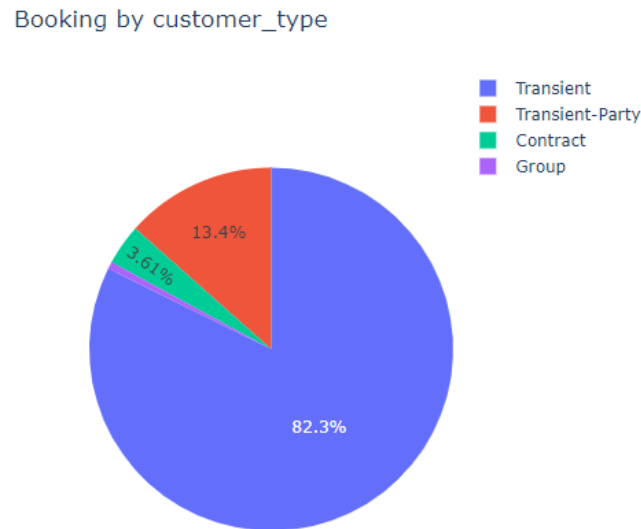


Figure [19]



**Figure [20]**

- Observation1: from the figures 18,19 & 20 reserved\_room\_type of type 'A', assigned\_room\_type of type 'B' and customer\_type of type 'Transient' fall in the majority portion of their respective distributions.
- Observation2: from the figure 18,19 & 20 we reserved\_room\_type of type 'L', assigned\_room\_type of type 'L' and customer\_type of type 'Group' fall in the majority portion of their respective distributions.

#### **4.3 Correlation Matrix Analysis between target variable & features**

- A correlation matrix is a tabular representation that contains correlation coefficients indicating the strength and direction of the relationship between different variables. Each entry in the table denotes the correlation between two variables and falls within the range of -1 to 1. This matrix is useful for summarizing data, serving as a diagnostic tool for advanced analyses, and serving as an input for more sophisticated analysis techniques.
- Our target variable is 'is\_canceled'. As part of this step we came with the heat map i.e.correlation matrix to understand the features which are highly, poorly correlated to the target variable 'is\_canceled' as below:

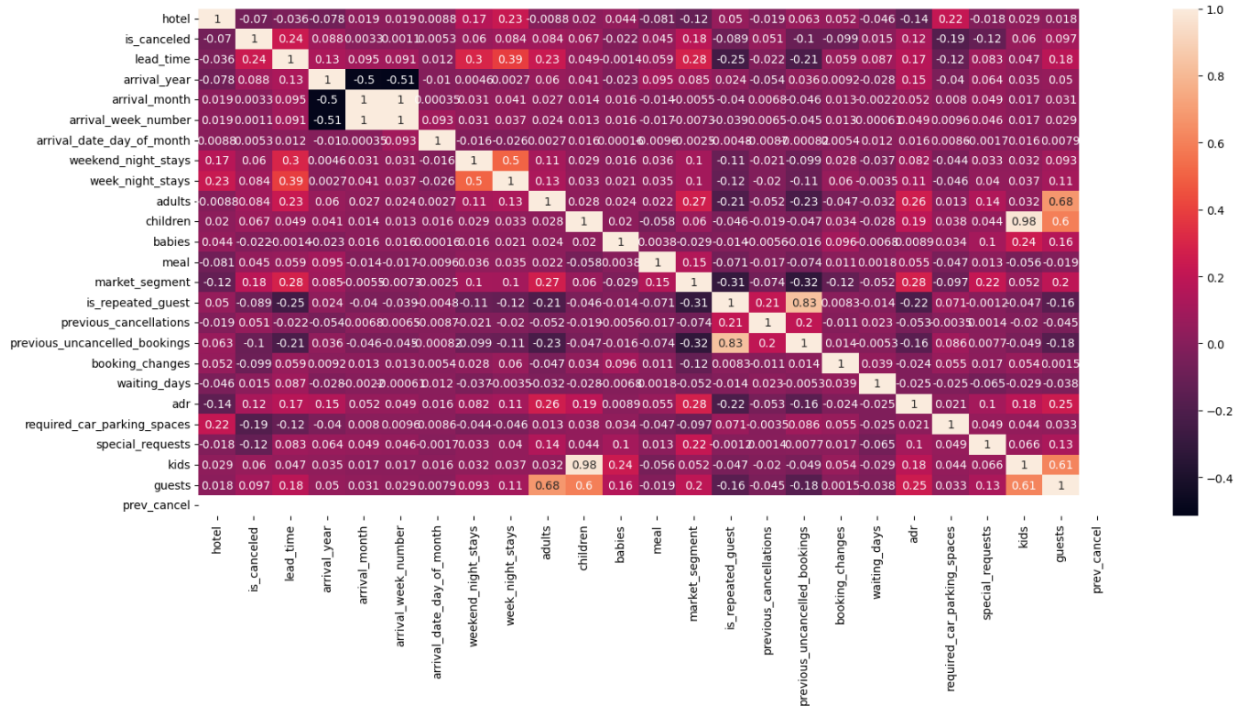
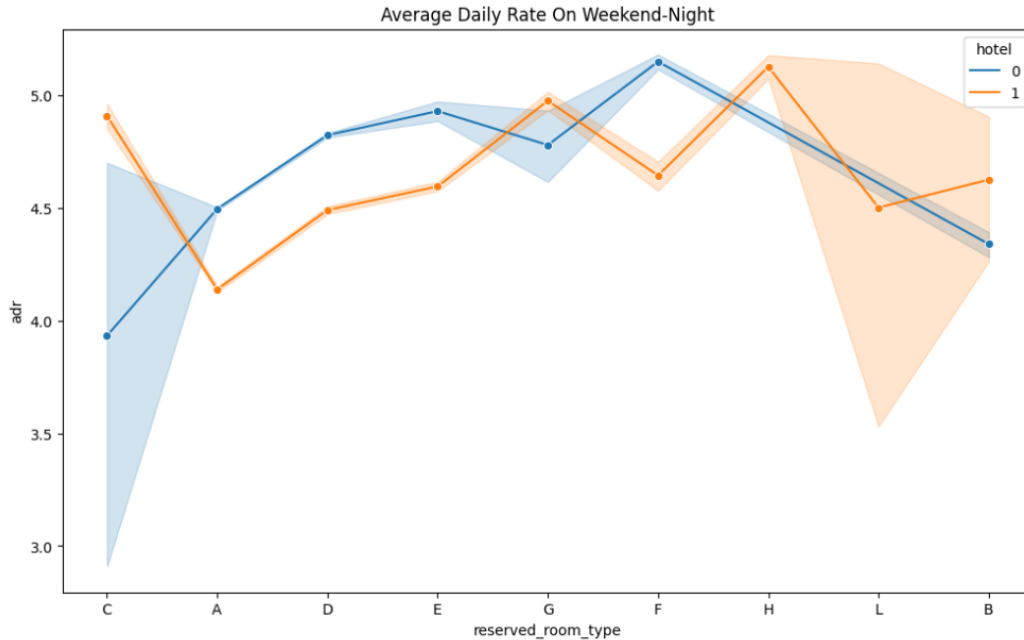


Figure [21]

- Observation 1: from the figure 21 highly correlated feature is 'lead\_time'.
- Observation 2: Poorly correlated feature is 'required\_car\_parking\_spaces'.

#### 4.4 Line Plot Analysis

- A line plot is a graphical representation that depicts data points either as discrete markers or connected line segments to demonstrate the continuous variation in the data concerning time or any other relevant variable.
- As part of this step we tried to plot Average daily rate against reserved room type for City Hotel and Resort Hotel.
- This analysis will shed info variance between the average daily rates category wise as below:



**Figure [22]**

- Observation 1 : From the figure 22 the category 'L' the variation in the daily rate is minimum and for the category 'C' the variation in the daily rate is maximum.
- Observation 2: From the figure 22, category 'C' of hotel 0 records minimum average daily rate.
- Observation 3: From the figure 22, category 'A' of hotel 1 records minimum average daily rate

#### 4.5 Scatter Plot Analysis

- Scatter plots depict the correlation between variables through the use of dots. These dots are used to represent the relationship between the variables. The scatter() method in the matplotlib library is utilized to generate a scatter plot. The application of scatter plots is extensive as they illustrate the connection between variables and the impact of changes in one on the other.
- From the Correlation matrix top 6 features which are highly correlated with the target variable 'is\_Canceled' are - 'lead\_time', 'market\_segment', 'adr', 'guests', 'arrival\_year', 'adults' as below:



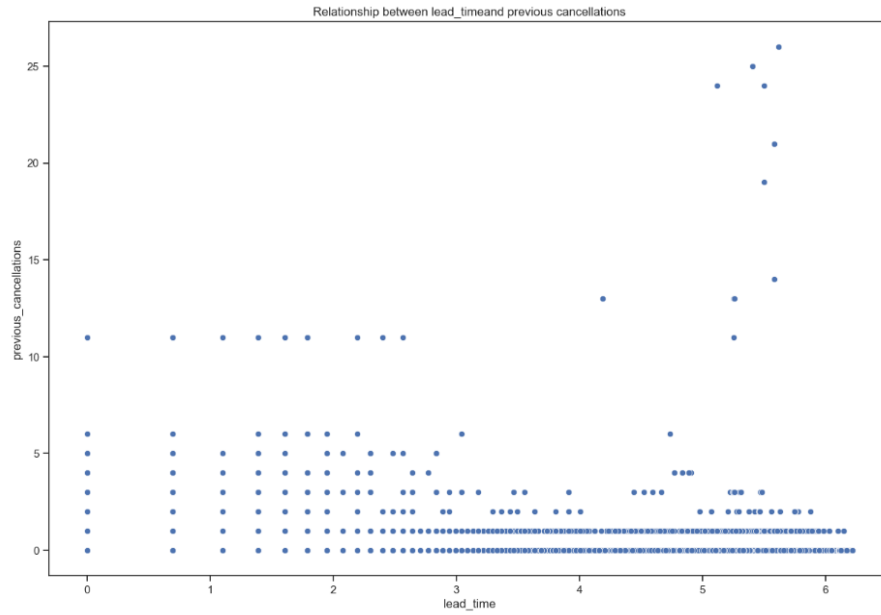


Figure [23]

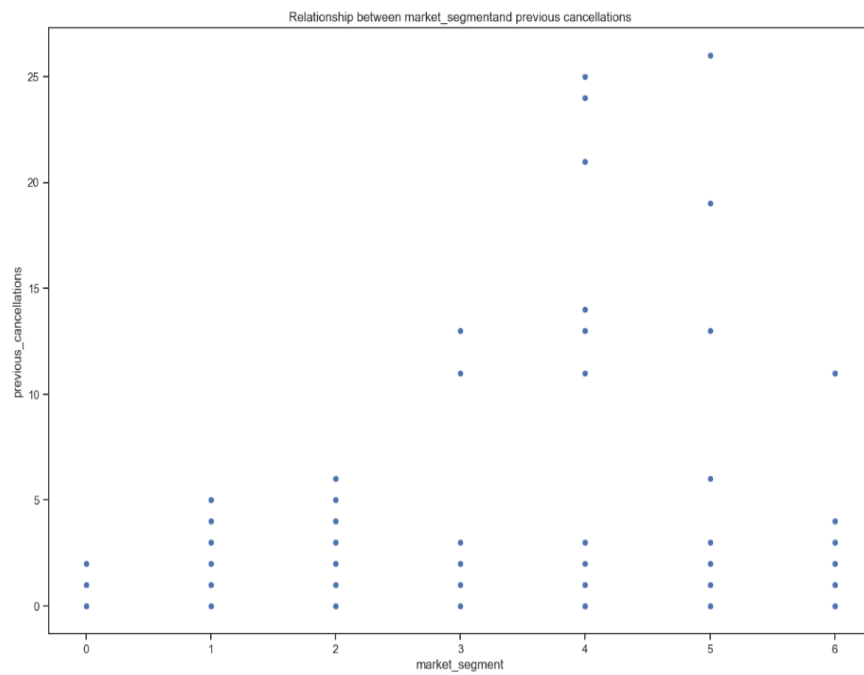


Figure [24]

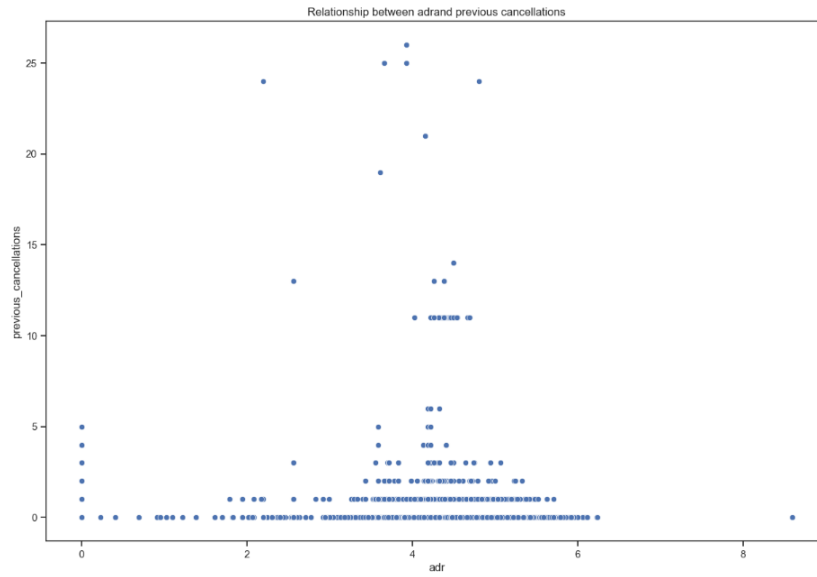


Figure [25]

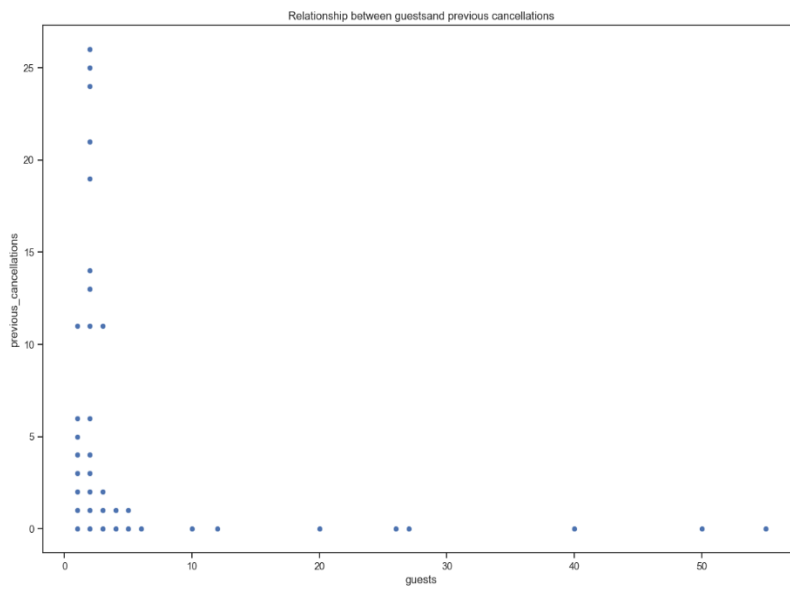
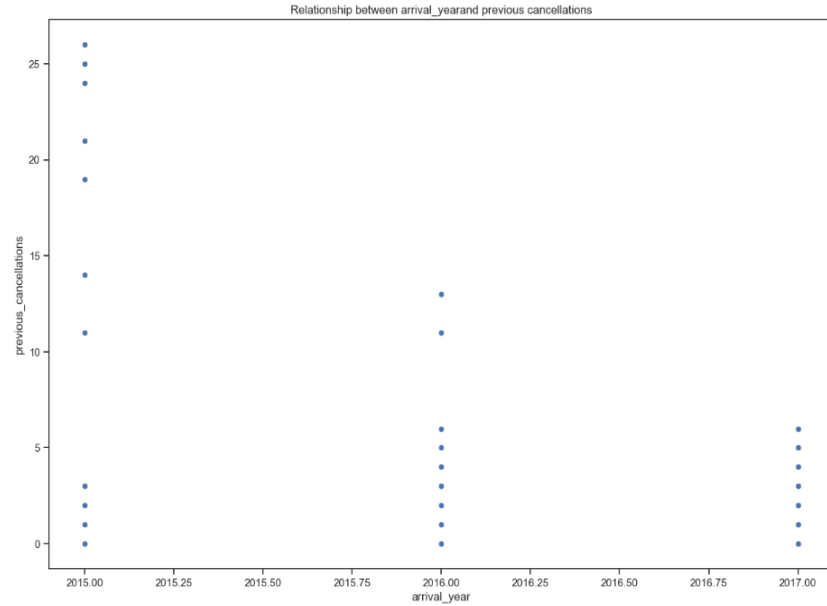
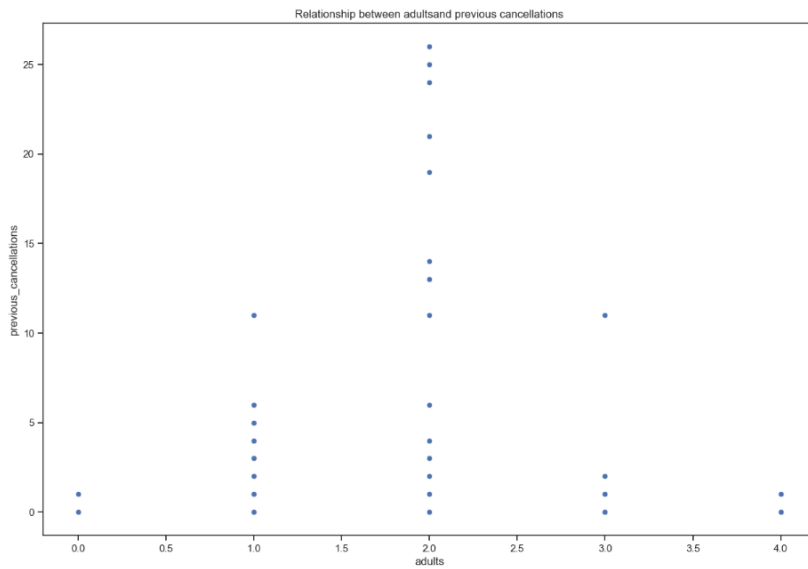


Figure [26]



**Figure [27]**



**Figure [28]**

- Observation1: data points are uniformly distributed for lead\_time vs is\_cancelled plot.
- Observation2: data are distributed around median/mean for average\_daily\_rate vs is\_cancelled plot.
- Observation3: Cancellations for the guests count ranges from 1 to 10 are more frequent compared to bookings where the number of guests is greater than 10.

## 4.6 Box Plot Analysis

- This type of chart is commonly used to display numerical data. This type of chart uses quartiles to represent the data, which is a statistical measure that divides the data into four equal parts. The quartiles are represented as boxes on the chart, with the top and bottom of the box representing the upper and lower quartiles, respectively.

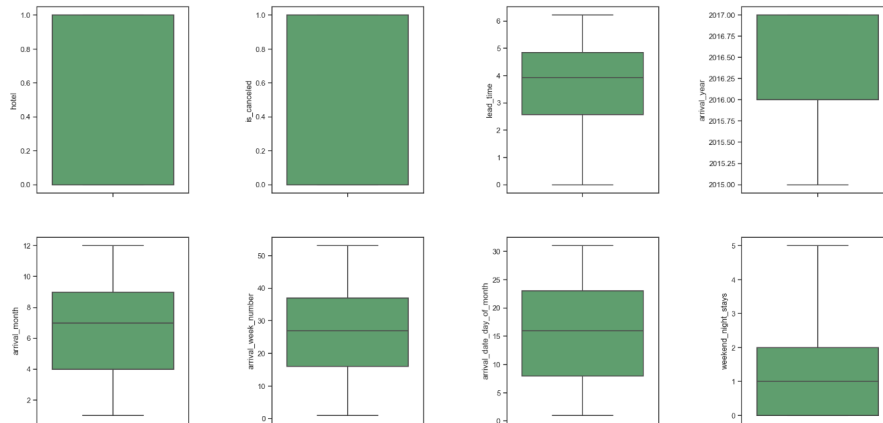


Figure [29]

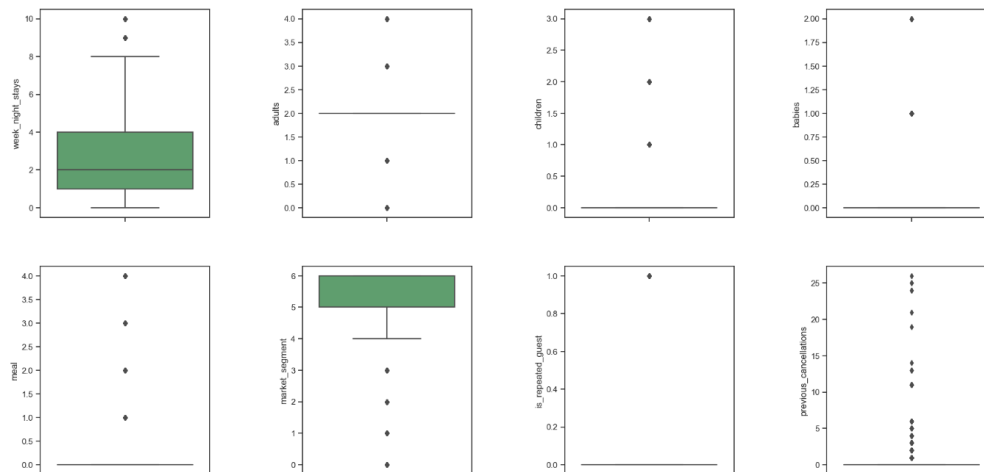


Figure [30]

- Observation1: In figure 30, for the last plot for 'previous cancellations' we can observe all outliers being displayed.
- Observation2: In figure 29 for the 'arrival\_year' plot we see quartile shifted primarily for recent years.

- Observation3: In the figure 29 for the 'weekend\_night\_days' plot we see quartiles shifted primarily for days < 2.
- Observation4: In the figure for plots 'arrival\_month', 'arrival\_week\_number' and 'arrival\_date\_day\_of\_the\_month' the interquartile range primarily fall close to the mean values.

## 4.7 Cat plot analysis

- Catplot is a function in the Seaborn data visualization library that creates a categorical plot by allowing you to plot categorical data in a variety of ways. It is a high-level interface for creating many kinds of statistical graphics that show the relationship between a numerical variable and one or more categorical variables.
- As part of this step we have cat plotted cancellations against 'Adults', 'children' and 'babies'. With this analysis we are trying to understand how different sections of humans play a role in booking cancellations as below:

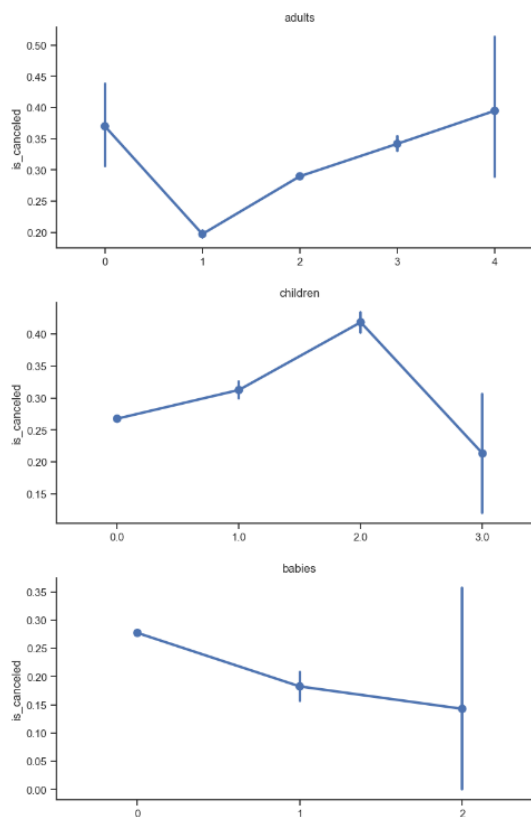


Figure [31]

- Observation1: Cancellations are inversely proportional to the number of children.
- Observation2: Cancellations are directly proportional to the number of adults.
- Observation3: For children the cancellations increase till it reaches children 2, crashes after that.

#### 4.8 Pair Plot.

- Pair plot analysis is a useful tool for visualizing the relationships between different variables in a dataset.
- In our dataset, pair plot helped to identify correlations between different features and the is\_canceled target variable, which typically implies the demand for hotel rooms.
- The pair plot is given in the below figure [32]



Figure [32]

- Observation1: If we see the arrival year and the hotel features there were three year were the hotels had bookings which are 2015, 2016, 2017
- Observation2:The is\_canceled feature has an increasing positive nearly linear relationship with features like lead\_time, arrival year, month and week, weekend and week night stays, market segment and adr, guests, previous cancellations, waiting days.
- Observation3:The is\_canceled feature has a decreasing or inverse relationship with features like repeated\_guests this is because if he is a repeated guest ( more value in this feature ) gives you less chance of cancellation.
- Observation4:The inverse relationship with features like parking spaces, special requests, previous\_booking not canceled this might be because if the booking was made with request for more parking spaces or special requests then there is a less chance of cancellation and also if the bookings were not previously not canceled there will be less chance of cancellations.

#### 4.9 Violin Plot Analysis.

- Generally a violin diagram is primarily used to display the size, spread, and density of the data distribution. It is especially helpful when contrasting the distributions of various groups or classifications because it makes it simple to identify variations in central tendency, variation, and skewness and violin plot also helps in revealing patterns and trends in data.
- I have plotted a violin plot across deposit\_type, hotel, cancellations. The violin plot is given in the below figure [33].

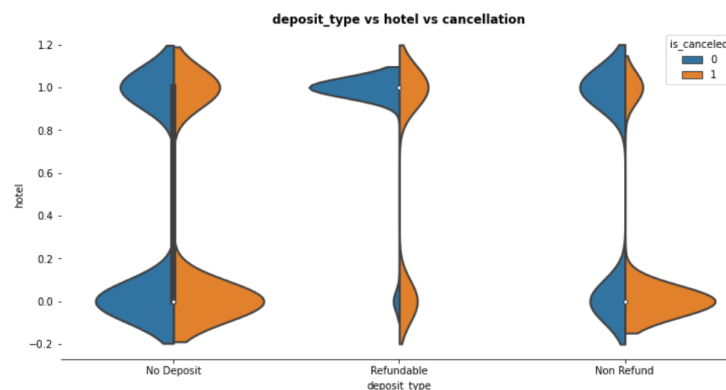
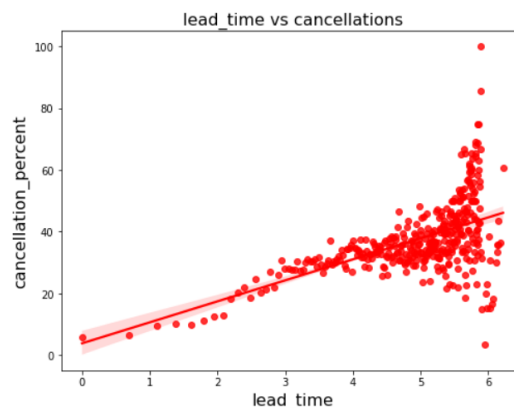


Figure [33]

- Observation1: Even though the deposit type is non refundable there were more cancellations in city hotels but there were less cancellations in resort hotels. This might be due to resort hotels having more deposit value.
- Observation2: When the deposit\_type is refundable there is an equal distribution of cancellations in both resort and city hotels.
- Observation3: When there is no deposit there were more cancellations in city hotels than resort hotels.
- Observation4: Out of all the deposit types when there is a refundable deposit there were very less honourings in city hotels. This is because even though the customer cancels the booking they will get their deposit back.

#### 4.10 Regressor Plot Analysis.

- A regressor plot enables you to see how two variables are related and evaluate how well the model matches the data. We can learn more about the nature and strength of the connection between the variables by examining the correlation coefficient, error bands, data points, trend line, and data points.
- After plotting the correlation matrix we could see that lead\_time and is\_cancellation have a good correlation of 0.24. So, to understand the relation between lead\_time and Cancellations I have plotted a regressor plot.
- In this plotting I have grouped lead\_time and only kept 10 bookings per graph. The Regressor plot is given in the below figure 34



**Figure [34]**



- Observation1: From the regressor plot we could observe that as the lead\_time increases there are more percent of cancellations. For the lead\_time 0 to 2 there are very less percent of cancellations.
- Observation2: This implies that customers who booked the hotels just/few days before the arrival have honored their booking. But customers who have booked the hotels many days before have made more percent of cancellations.

## 5. References:

- Data intensive Computing Lecture Slides
- [https://practice.geeksforgeeks.org/courses/data-science-live?utm\\_source=GfG&utm\\_medium=gfg\\_submenu&utm\\_campaign=DS\\_Submenu/](https://practice.geeksforgeeks.org/courses/data-science-live?utm_source=GfG&utm_medium=gfg_submenu&utm_campaign=DS_Submenu/)
- <https://www.analyticsvidhya.com/>
- <https://realpython.com/pandas-plot-python/#analyze-categorical-data>