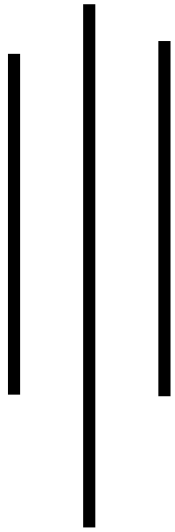
Project Proposal on

"Hotel Booking Strategies Research & Analysis"



TEAM 2:

Lina Duarte

Rubina Pradhan

Rachel Prathiba John

Vinay Karthick Jeyabalakrishnan

Prathiksha Bojamma Paleyada Thimmaiah

Date : 13/12/2021

Reviewed By: Professor Jin Fang

1. INTRODUCTION:

1.1 Research problem:

The hospitality industry is one of the fastest-growing sectors in the world which works on the fine balance of supply and demand. It is necessary for every business to understand the trends to maximize profits and efficiency. The data set used in this project consists of data points which directly or indirectly helps to understand consumer behavior and gives an advantage to the decision makers. Furthermore, as the expectations of customers in the industry constantly changes it requires accurate estimation of booking and cancellation ratio of the guests in order to maximize the revenue.

The main objective of this project is to predict the bookings and cancellation of the guest, analyze the customer segmentation and satiation as per the country and compare the trends seasonally. Understanding such issues avoid the industry from creating bottlenecks which impede the expected growth. The goal of the project is to better understand the industry issues that would give the hoteliers and hospitality managers a clear edge over the competition by contrasting dataset of city hotel and resort hotel to help in overcoming the limitation and determine if a different policy should be applied for each one.

1.2 Why is this research problem important?

i. Understanding the dependency of booking and cancellations among other variables in the dataset to maximize the efficiency, returns and inventory management. (Antonio, de Almeida, & Nunes, 2019)

ii. The research will help businesses to advertise to the right target segment at the right time of the year.

iii. The aviation industry faces a similar problem where cancelations impact the business hence advanced computation is done to predict the cancelation and respective additional bookings are done for maximum efficiency.

iv. This project will serve as a base for further investigation on the dependent variables that can be used to lay a foundation to prescriptive analytics and taking decisions at a real time.

v. The booking/cancelation behavior of the guests can be inferred and predicted to support operational strategies.

Benefit of understand hotel booking dynamics:

Hotel management industry comprises of target marketing, hotel administration, location, booking and cancellation rate, accounts and assisting guests. The primary goal is to run a hotel successfully while managing the other aspects of the business at the same time. Moreover, the hotel industry not only relates to the location i.e. resort hotels or city hotels, but it also includes estimation of popular booking month-wise, guest type, optimal length of stay and average daily rate. Our research project on hotel booking strategies can help understand those factors in-depth.

Intended Audience

This research project is helpful for HR Hotel Manager, Marketing Manager, and Hotel CEO.

1.3 How does it relate to the STAT 4600 class?

i. In our research, we used a lot of the statistical methods that were taught to us in class. Those methods include the numerical descriptive measures, probability, graphical visualizations, central limit theorem, etc.

ii. We were able to implement these statistical methods to perform our research and make inferences.

iii. We were able to find the trend of bookings and cancellations, thereby the attributes that affect them too.

iv. In a similar way, we were able to find the lead time distribution, standard deviation, etc.

2. METHODS:

2.1 Type/source/content of data used in the project

The hotel booking demand data originally comes from ScienceDirect, authors Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. The data was downloaded and cleaned by Thomas Mock and Antoine Bichat on February 11th, 2020. Our data set contains both qualitative data such as categorical data, for example, duration of booking, length of stay, the exact number of grown-ups, kids, as well as babies, deposit in addition to many other things.

The main dataset includes 119390 rows and 32 columns. The raw format of Hotel Booking data does contain null/ missing values.

Cleaning and Processing:

Data Cleaning: "the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted." [1]
Data processing: “data is collected and translated into usable information." [2]
Data Cleaning of Hotel Booking requires a number of unwanted columns to be drop down.

Reason for Data Cleaning: We have dropped reserved_room_type, agent, required_car_parking, day_is_waiting_list, total_of_special_requests, reservation_status and reservation_status_date as it do not have any direct or indirect impact on the research objective.

The selected variables are selected 24 columns after dropping down 6 columns are based on the dependent and independent variables on the research objective.

Index :

Index No.	Columns	Data Type
1	hotel	object
2	is_canceled	int64
3	lead_time	int64
4	arrival_date_year	int64
5	arrival_date_month	object
6	arrival_date_week_number	int64
7	arrival_date_day_of_month	int64
8	stays_in_weekend_nights	int64
9	stays_in_week_nights	int64
10	adults	int64
11	children	float64
12	babies	int64
13	country	object
14	market_segment	object
15	distribution_channel	object
16	is_repeated_guest	int64
17	previous_cancellations	int64
18	previous_bookings_not_canceled	int64
19	assigned_room_type	object
20	booking_changes	int64
21	deposit_type	object
22	company	float64
23	customer_type	object
24	adr	float64

Note: After the data cleaning and processing, the dataset contains 119390 rows and 24 columns as we have dropped the unwanted columns and NA values from the main dataset.

2.2 Analysis and modelling methods

As part of the research, analysis and modelling methods on the dataset will be performed in R and Python program to implement various descriptive statistical methods such as mean, median, quartiles, standard deviation, and more. Graphs will be constructed for these statistical models to provide a better visualization of the modelling tasks performed. We will also be making use of regression model, prediction model, interpreting the correlation, converting categorical variables to numerical in the analysis of the research project.

In this project we are comparing two main categories i.e., City Hotels and Resort Hotels. The booking and cancellation trends are our dependent variables. Keeping these as the main features of comparison, we are taking several factors that affect the booking and cancellation trends like guest type (individual/family), lead time, deposit type and more. We will be performing many comparisons based on multiple factors such as:

i. The number of bookings and cancellations that happened in one particular year for city hotels vs. resort hotels.

ii. Market segment of these bookings/cancellations for city hotels vs. resort hotels.

iii. Whether or not these bookings/cancellations have prior deposits.

iv. Whether the individual members of the family (adults, children and babies) have an impact on the bookings/cancellations.

v. Comparing the seasonal booking rate of the two different categories of hotel.

2.1.1 Understanding the dataset

Table 1.1. Table showing the total number of bookings and cancellations over 3 years in Resort and City Hotels

Hotel Type	Year	Number of Bookings	Number of Cancellations
Resort Hotel	2015	6,176	2,137
	2016	13,636	4,929
	2017	9,124	4,053
City Hotel	2015	7,677	6,003
	2016	22,732	15,406
	2017	15,816	11,694

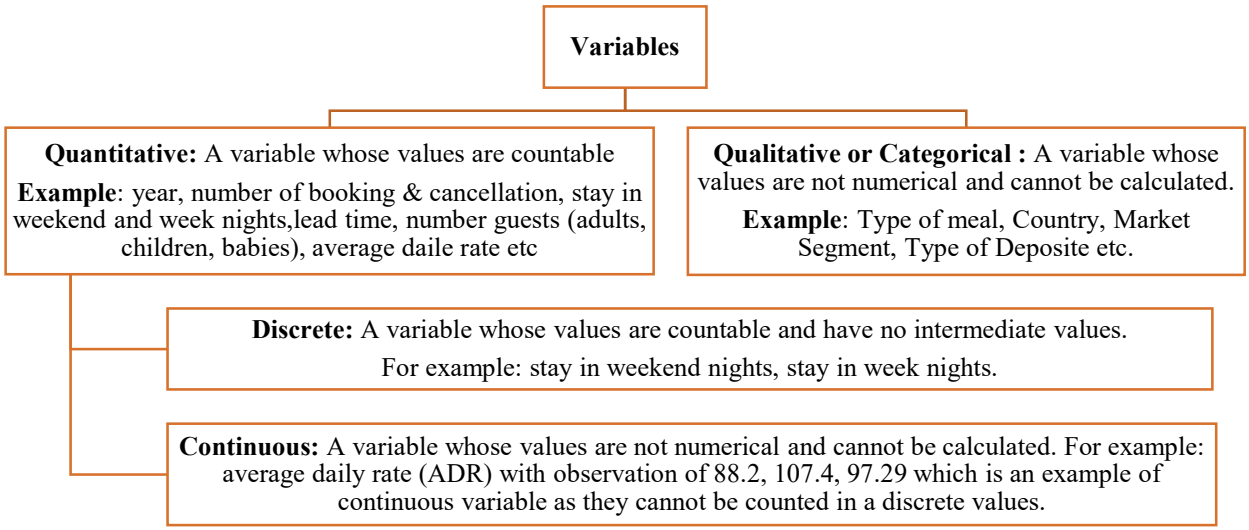
Variables

An element or member

An observation or measurement

Example 1.1. The element or member in your dataset

2.1.2 Types of Variables



2.1.3 Cross-section vs. Time-series Data

Cross-section data: Data collected on different elements at the same point in time or for the same period of time are called cross-section data. The observations for Resort hotel and city hotel in the month of July in the year 2015 is a cross-section data as they are different elements recorded in the same period of time. Example of cross section data shown below:

Hotel Type	Number of Bookings	Average Daily Rate (ADR)
Resort Hotel	1093	171.37
City Hotel	2234	125.33

Example 1.3. Table showing number of bookings and ADR of Resort and City Hotel in July 2015

Time-series data: Data collected on the same element for the same variable at different points in time or for different periods of time are called time-series data. As per our dataset, the observations of City hotel in different periods of time is an example of time-series data in the dataset. Example of Time-series data shown below:

Hotel Type	Year	Number of Bookings
Resort Hotel	2015	6,176
	2016	13,636
	2017	9,124
City Hotel	2015	7,677
	2016	22,732
	2017	15,816

Example 1.4. Number of bookings from year 2015 to 2017 for Resort and City Hotel

2.1.4 Population Vs. Sample

- **Population** is a study of all the elements or objects present in the dataset. The study of all the observations on both Resort hotel and city hotel is an example of a population.
- **Sample** is a study of only a portion of the population. The study of observations on Resort hotel alone, which is only a portion of the population, makes a sample in the dataset.

2.1.5 Census Vs. Sample Survey

- **Census** is a survey on all the members of the population. In the dataset, a survey on all the resort hotels in 2017 is an example of Census.

- **Sample survey** includes only a portion of the population. The survey on only few resort hotels in 2017 is an example of the sample survey.

2.1.6 Representative Sample

- A sample on City hotels alone can be a representative sample as the results obtained from this sample will be specific to City hotels alone and not Resort hotels. For example: The collection of guests who stayed during the weekend on 2015 in City Hotel.
- Any decisions made based on the results of this sample can be applied only to City hotels and not to Resort hotels. For example: The collection of guests who stayed for more than five weekends on 2015 in City Hotel.

2.1.7 Types of Sampling

- **Sampling with replacement**: Selecting a hotel and putting it back to the population is called sampling with replacement. Here, the population contains the same number of items after every selection. Hence, same item can be chosen any number of times.
- **Sampling without replacement**: a selection is made and not replaced in the population, the size of the population is reduced after each selection, and this is called a sampling without replacement.

2.1.8 Random Vs. Non-Random Samples

- **Random sample**: A mixture of resort and city hotels are selected from the population. This way there are chances for all the items to be selected.
- **Non-random sample**: Only resort hotels are selected from the population. Hence, the city hotels have no chance to be selected from the population

2.1.9 Sampling Errors vs Non- Sampling Error

- The number of booking cancellations is lesser in sample than the cancellations in the population, therefore this is called the **sampling error.**
- Data entered for the hotels in the sampling is incorrect, therefore it gives wrong results, this makes a **non-sampling error.**

2.1.10 Types of Errors

i) **Selection error:** A sample consisting of city and resort hotels to check their cancelation rate. We may miss hotels from few places as their details are not recorded anywhere online. Therefore, the sample results are different from the population result. This is called the selection error.

ii) **Non-response error**: The data set consist of hotels and its cancellation status along with some basic data about the customer like children and babies which could be used to calculate the cancellation rate. Some customers may not include these data which would cause a non-response error.

iii) **Response error:** The data set consist of hotels and its cancellation status along with some basic data about the customer like children and babies which could be used to calculate the cancellation rate as stated in question 13. Customers may not disclose the true details causing a response error in the data.

iv) **Voluntary response**: The dataset consists of a column where the customer has to mention the number of babies and this detail is specific to people with babies and not every customer would respond to that. This is called a voluntary response error.

2.1.11 Random Sampling Techniques:

i) Simple Random Sampling technique:

- Consider a selection of 10 hotels from a mixture of both resort and city hotels.
- Choose one hotel from the mixture

- Repeat the selection nine more times
- The 10 hotels selected from the mixture makes a simple random technique.

ii) Systematic random sampling technique :

- Arrange the 1,19,391 data as per the date.
- The size of sample should be 51, therefore the sample size 119,391/51 = 2,341
- Choose a hotel randomly from the first 2,341 hotels.
- If we choose 52nd hotel, then we select 2,341th hotel from the data.
- The Sample consist of hotels with position 52, 2,393, 4734 , 7075 and so on.

iii) Stratified random sampling technique :

- Divide the entire data on city and resort hotels into 3 sets based on the year 2015, 2016 and 2017.
- Form 3 groups based on the years, which is also known as strata
- Select a sample from each of these groups
- The collection of these three samples from the strata makes up a stratified random sample.

iv) Cluster sampling technique :

- Divide entire data set on hotel demand based on the months and create 12 clusters.
- The clusters should be similar to each other.
- Choose 7 clusters randomly from 40.
- Select hotels based on years randomly from these 7 clusters and analyze the cancellation rate on these selected hotels. This sampling technique is called cluster sampling

2.1.12 Design of Experiment:

This dataset is an observational study as we are using the data of resort and city hotels to predict the cancellation rate without putting these two types of hotels into any kind of experiment or

conditions. The dataset comprises of general booking and cancellations made on these hotels, using which we can predict the future trends.

2.2.1 The frequency distribution table for the qualitative variables in the dataset:

Qualitative Variables	Frequency
Resort Hotel	40060
City Hotel	79330
AGO	362
AUS	426
AUT	1263
BEL	2342
BGR	75
BRA	2224
CHE	1730
CHL	65
CHN	999
CN	1279
COL	71
CYP	51
CZE	171
DEU	7287
DNK	435
DZA	103
ESP	8568
FIN	447
FRA	10415
GBR	12129
GRC	128
HRV	100
HUN	230
IND	152
IRL	3375
IRN	83
ISL	57
ISR	669
ITA	3766
JPN	197
KOR	133
LTU	81
LUX	287
LVA	55
MAR	259
MEX	85
MOZ	67
NLD	2104
NOR	607
NULL	488
NZL	74
POL	919
PRT	48591
QAT	15
ROU	500
RUS	632
SRB	101
SVK	65
SVN	57
SWE	1024
THA	59
TUR	248
TWN	51
UKR	68
USA	2097
ZAF	80

Most number of bookings recorded in 2016 and most number of bookings in 33rd week of the year and 17th day of the month. Most bookings are done by 2 adults and they stay mostly for 2 days in the weekdays. ADR of the hotels are recorded to a mode of 62.

2.2.2 The relative frequency and percentage of the above mentioned qualitative variables:

Qualitative Variables	Frequency	Relative Frequency	Percentage
Resort Hotel	40060	0.056036201	5.603620112
City Hotel	79330	0.110967345	11.09673449
AGO	362	0.000506368	0.050636807
AUS	426	0.000595892	0.05958917
AUT	1263	0.001766693	0.176669301
BEL	2342	0.003276006	0.327600557
BGR	75	0.000104911	0.010491051
BRA	2224	0.003110946	0.311094636
CHE	1730	0.002419936	0.241993579
CHL	65	9.09224E-05	0.009092244
CHN	999	0.001397408	0.139740801
CN	1279	0.001789074	0.178907392
COL	71	9.93153E-05	0.009931528
CYP	51	7.13391E-05	0.007133915
CZE	171	0.000239196	0.023919597
DEU	7287	0.010193105	1.019310528
DNK	435	0.000608481	0.060848097
DZA	103	0.000144077	0.01440771
ESP	8568	0.011984977	1.198497681
FIN	447	0.000625267	0.062526665
FRA	10415	0.014568573	1.456857301
GBR	12129	0.016966128	1.696612789
GRC	128	0.000179047	0.017904727
HRV	100	0.000139881	0.013988068
HUN	230	0.000321726	0.032172557
IND	152	0.000212619	0.021261864
IRL	3375	0.004720973	0.472097301
IRN	83	0.000116101	0.011610097
ISL	57	7.9732E-05	0.007973199
ISR	669	0.000935802	0.093580176
ITA	3766	0.005267906	0.526790648
JPN	197	0.000275565	0.027556494
KOR	133	0.000186041	0.018604131
LTU	81	0.000113303	0.011330335
LUX	287	0.000401458	0.040145756
LVA	55	7.69344E-05	0.007693437
MAR	259	0.000362291	0.036229097

MEX	85	0.000118899	0.011889858
MOZ	67	9.37201E-05	0.009372006
NLD	2104	0.00294309	0.294308954
NOR	607	0.000849076	0.084907574
NULL	488	0.000682618	0.068261773
NZL	74	0.000103512	0.01035117
POL	919	0.001285503	0.128550347
PRT	48591	0.067969422	6.796942208
QAT	15	2.09821E-05	0.00209821
ROU	500	0.000699403	0.069940341
RUS	632	0.000884046	0.088404591
SRB	101	0.000141279	0.014127949
SVK	65	9.09224E-05	0.009092244
SVN	57	7.9732E-05	0.007973199
SWE	1024	0.001432378	0.143237818
THA	59	8.25296E-05	0.00825296
TUR	248	0.000346904	0.034690409
TWN	51	7.13391E-05	0.007133915
UKR	68	9.51189E-05	0.009511886
USA	2097	0.002933298	0.29332979
ZAF	80	0.000111905	0.011190455
Aviation	237	0.000331517	0.033151722
Complementary	743	0.001039313	0.103931347
Corporate	5295	0.007406682	0.74066821
Direct	12606	0.017633359	1.763335874
Groups	19811	0.027711762	2.771176187
Offline TA/TO	24219	0.033877702	3.387770232
Online TA	56477	0.079000413	7.900041265
Undefined	2	2.79761E-06	0.000279761
Corporate	6677	0.009339833	0.933983312
Direct	14645	0.020485526	2.048552585
GDS	193	0.00026997	0.026996972
TA/TO	97870	0.136901223	13.69012233
Contract	4076	0.005701537	0.570153659
Group	577	0.000807112	0.080711153
Transient	89613	0.125351275	12.53512754
Transient - Party	25124	0.035143622	3.514362249
Total	714895		

Figure 1.1 Pareto Chart

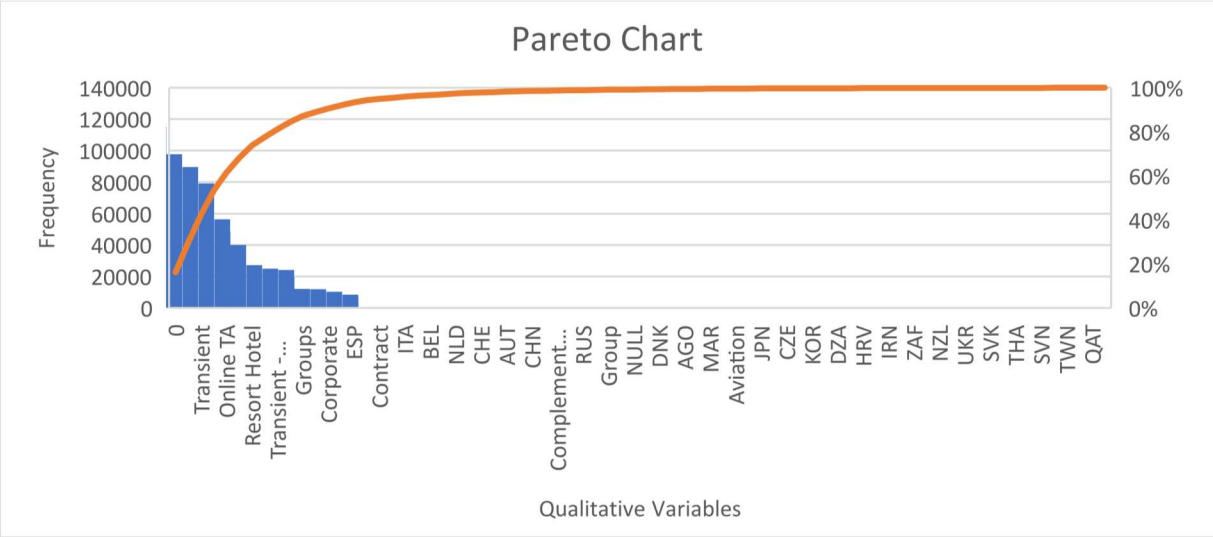
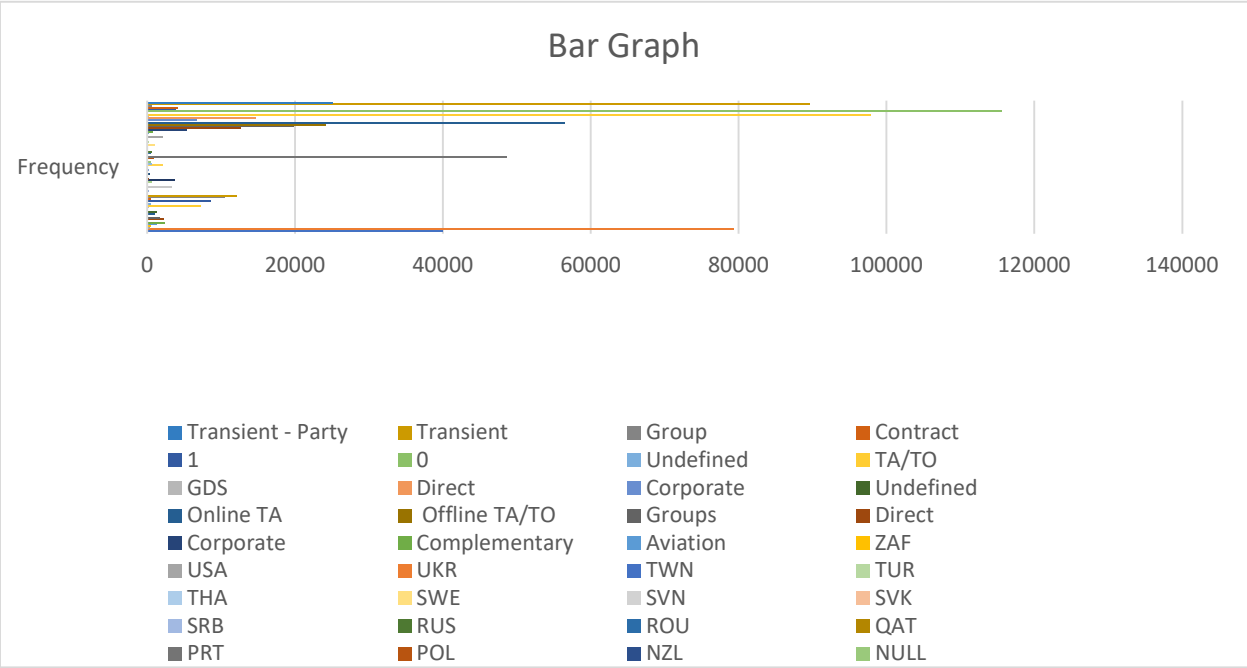


Figure 1.2 Bar Graph



2.2.2 The frequency distribution table:

Quantitative variables	Frequency
is_canceled	
0	75166
1	44224
lead_time	
0	6345
1	3460
2	2069
3	1816
4	1715
5	1565
6	1445
7	1331
8	1138
9	992
10	976
stays_in_weekend_nights	
0	51998
1	30626
2	33308
3	1259
4	1855
stays_in_week_nights	
0	7645
1	30310
2	33684
3	22258
4	9563
adults	
0	403
1	23027
2	89680
children	
0	110796
1	4861
2	3652
3	76
babies	
0	118473
1	900
2	15
9	1
booking_changes	
0	101314
1	12701
2	3805
3	927
Total	835379

Figure 1.3 Bar Graph of Lead Time

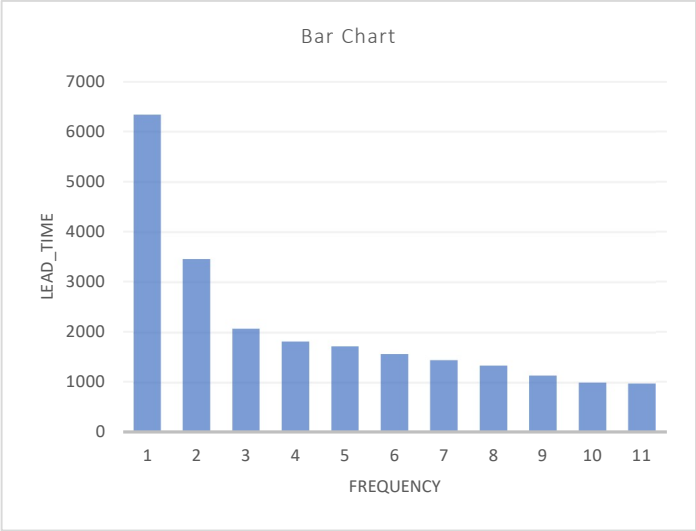


Figure 1.4 Frequency Polygon for stay in weekend night stay

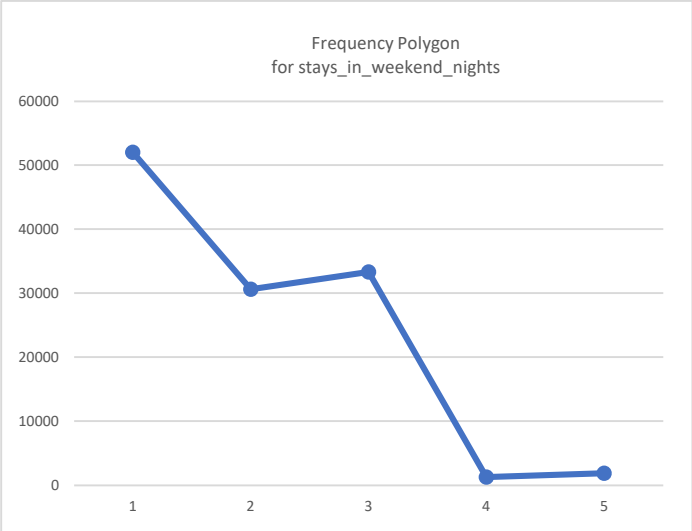


Figure 1.5 Frequency Polygon for stay in week night

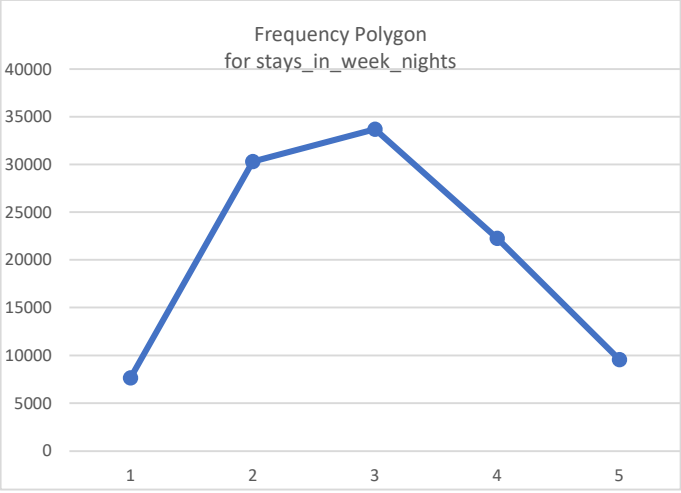


Figure 1.6 Frequency Polygon for number of children

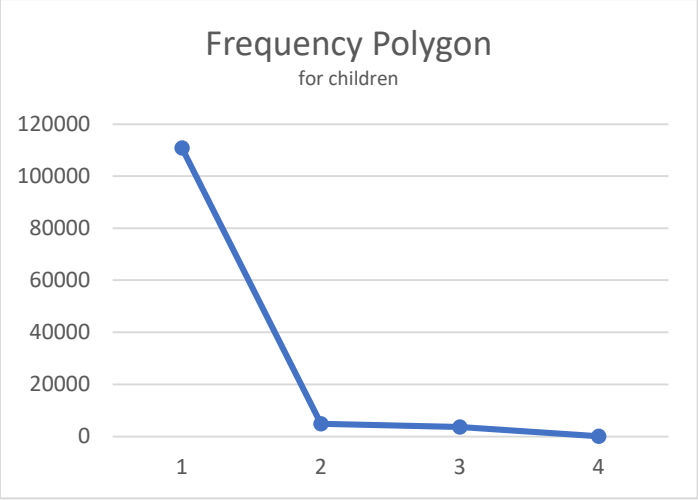
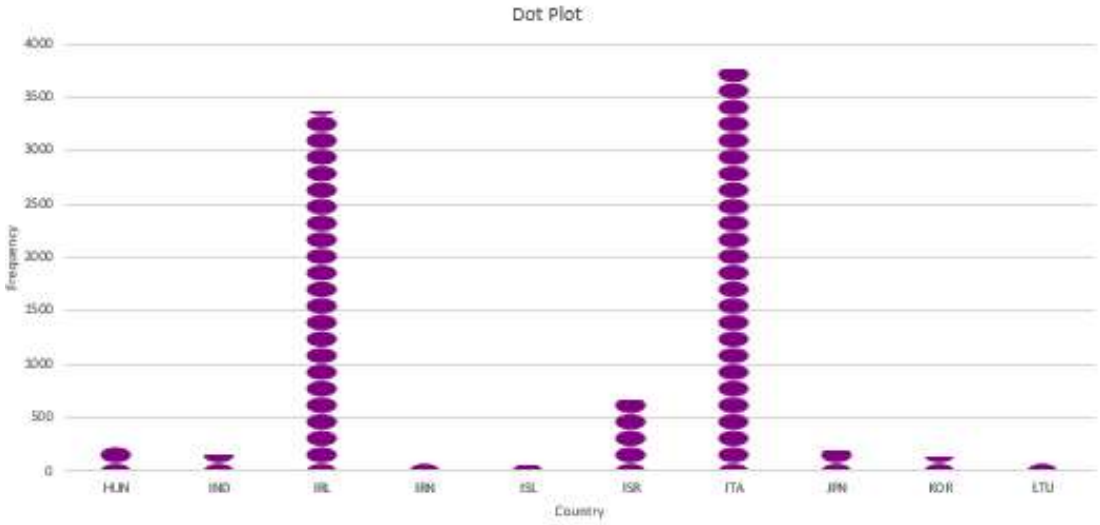


Figure 1.7 Stem and leaf of random observation

Countries	Frequency
QAT	15
TWN	51
SVN	57
THA	59
SVK	65
MOZ	67
NZL	74
BGR	75
LTU	81
IRN	83

1	5
2	
3	
4	
5	1 7 9
6	5 7
7	4 5
8	1 3

Figure 1.8 Dot Plot



2.2.3 The cumulative frequency distribution table with cumulative relative frequency and cumulative percentage for the quantitative variables in your dataset:

Quantitative variables	Frequency	Relative Frequency	cumulative frequency	cumulative relative frequency	cumulative percentage
is_canceled					
0	75166.00	0.09	75166.00	0.09	9.00
1	44224.00	0.05	119390.00	0.14	14.29
lead_time					
0	6345.00	0.01	125735.00	0.15	15.05
1	3460.00	0.00	129195.00	0.15	15.47
2	2069.00	0.00	131264.00	0.16	15.71
3	1816.00	0.00	133080.00	0.16	15.93
4	1715.00	0.00	134795.00	0.16	16.14
5	1565.00	0.00	136360.00	0.16	16.32
6	1445.00	0.00	137805.00	0.16	16.50
7	1331.00	0.00	139136.00	0.17	16.66
8	1138.00	0.00	140274.00	0.17	16.79
9	992.00	0.00	141266.00	0.17	16.91
10	976.00	0.00	142242.00	0.17	17.03
stays_in_weekend_nights					
0	51998.00	0.06	194240.00	0.23	23.25
1	30626.00	0.04	224866.00	0.27	26.92
2	33308.00	0.04	258174.00	0.31	30.91
3	1259.00	0.00	259433.00	0.31	31.06
4	1855.00	0.00	261288.00	0.31	31.28

2.3.1 Numerical Descriptive Measure

Description	Mean	Median	Mode	Standard Deviation	Variance	Co. eff. of variation	Max
Booking Cancellation	0.37	0.00	0.00	0.48	0.23	0.63	0.00
Lead Time	104.01	69.00	0.00	106.86	11419.63	109.79	737.00
Arrival Date	2016.16	2016.00	2016.00	0.71	0.50	0.00	0.00
Arrival Date Week	27.17	28.00	33.00	13.61	185.10	6.81	0.00
Arrival Day of the Month	15.80	16.00	17.00	8.78	77.10	4.88	0.00

Stays in weekend nights	0.93	1.00	0.00	1.00	1.00	1.08	0.00
Stays in weekday nights	2.50	2.00	2.00	1.91	3.64	1.46	0.00
Adults	1.86	2.00	2.00	0.58	0.34	0.18	0.00
Children	0.10	0.00	0.00	0.40	0.16	1.53	0.00
Babies	0.01	0.00	0.00	0.10	0.01	1.19	0.00
Repeated Guest	0.03	0.00	0.00	0.18	0.03	0.97	0.00
Booking Changes	0.22	0.00	0.00	0.65	0.43	1.92	0.00
ADR	101.80	94.58	62.00	50.54	2554.00	25.08	0.00

2.3.2 Discussing Mean Vs. Median:

The mean is the average where the sum of all the numbers is divided by the total number of numbers, whereas the median is the middle value in the list of given numbers numerically ordered from smallest to biggest.

One of the important variables of the dataset “Lead Time” which is the number of days in advance a booking is done, the Mean is 104 and median is 69. The mean might be affected due to outliers, but the median gives an estimation on how many days the bookings are made in advance, thus helping business owners to prepare for the operations.

2.3.3 Discussing Mode:

The mode is the value of the number which occurs most often in the list. In this dataset sue to the size of data the mode is 0 for cancellation, lead time, stays in weekend nights, children, babies, repeated guest, and booking changes. It has a positive impact too for variables like cancellation and booking changes which will help decision makers understand the consumer behaviour.

2.3.1 Numerical Descriptive Measures

20 Random variables from ADR
33
33
34
35
41
56
65
70
85
85
107
107
109
114
153
190
190
190
212
212
96

Calculating the approximate value of 35th percentile

The approximate value of 35th percentile = (k * n) / 100
= (35 * 20) / 100
= 7th item
= 65

Percentile rank of 65

Finding Percentile Rank of a Value

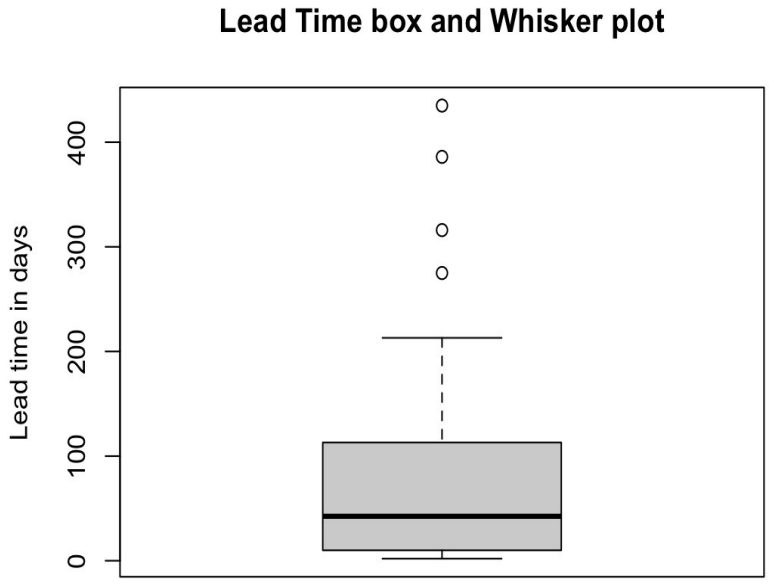
$$\text{Percentile rank of } x_i = \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100\%$$

Percentile Rank of 65 = (6 / 20) x 100% = 30%

Quartiles and IQR

i. (25th Percentile) = (n + 1) / 4 = 5th item = 41
ii. (Median) = 96
iii. (75th Percentile) = 180.75
iv. Interquartile range = Q3 – Q1 = 180.75 – 41 = 139.75

2.3.2 Box and Whisker Plot



2.4.1 Probability

Experiment, Outcome, and Sample Space

Experiment	Outcomes	Sample Spaces
Resort hotel or City Hotel	Adult (A), Babies (B), Children (C)	{A, AB, AC}
Booking and Cancellation	Booking(B), Cancel (C)	{B,C}
Meal type	BB - Bed & Breakfast; HB – Half board (breakfast + one meal); FB – Full board (breakfast+ lunch+ dinner) SC – no meal package	{BB, HB, FB, SC}

- **Simple Event :**

  Simple event includes one and only one of the (final) outcomes for an experiment. Picking two random booking from the dataset. Let the booking done by the new guest be (N) and the booking done by the repeated guest be (R). Each of the final four outcomes (NR, RN, NN, RR) for this experiment is a simple event. These four events can be denoted by E_1, E_2, E_3, and E_4, respectively.
  E_1 = {NR}, E_2 = {RN}, E_3 = {NN}, E_4 = {RR}.

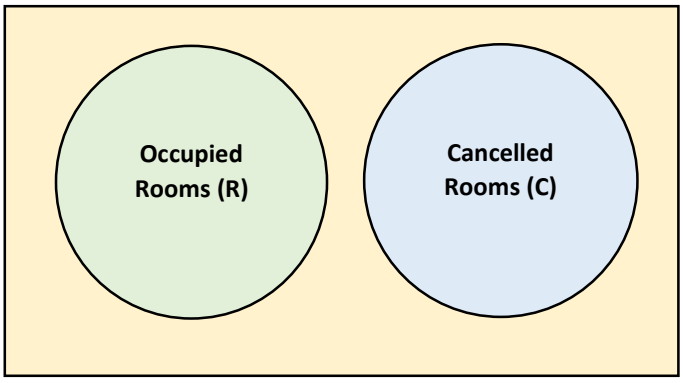
- **Compound Event :**

  Compound event is a collection of more than one outcome for an experiment. Picking two random booking from the dataset and observing whether the booking is done by at most one repeated guest. Let A be the event that at most one repeated guest is selected.
  A = at most one repeated guest has booked = {RN, NR, NN}
  Here, the event A contains more than one outcome, which is an example of compound event.

2.4.2 Mutually Exclusive Event :

As per the dataset, the booking and cancellation are mutually exclusive as it cannot occur together. The room occupied and cancellation cannot be occurred at once, hence it will always have zero intersection.
Let the probability of occupied room by (R) and cancelled room be (C).

Hence, P(R ∩ C)



2.4.3 Independent Event

Independent events are those events whose occurrence is not dependent on any other event. For example, in this dataset, if a guest books a resort hotel during the first trip and then again if the guest books the City hotel for the next trip, the event is independent event as the occurrence of one event does not affect the occurrence of the other event.

Let R be the guest booking in Resort hotel and C be the guest booking in City hotel.

P(R | C) = P (R) or P(C | R) = P(C)

2.4.4 Dependent Event

Two events are dependent if the outcome of the first event affects the outcome of the second event, so that the probability is changed. As per the dataset, we can observe that the cancellation is

dependent upon the booking hotel type. The probability of booking in a resort hotel is calculated, that is, the number of bookings in a resort hotel divided by the total of reservations. The probability of being cancelled and being a resort hotel is calculated dividing the total of rows cancel only for resort hotels by the total of reservations only in resort hotels.

P(R) = 11122/40060 = 0.277

The probability of cancelled hotel , P(C) = 44224/ 119390 = 0.37

Let C be the probability of being cancelled and R be the probability of booking in a Resort hotel.

= P(C | R) ≠ P(C)

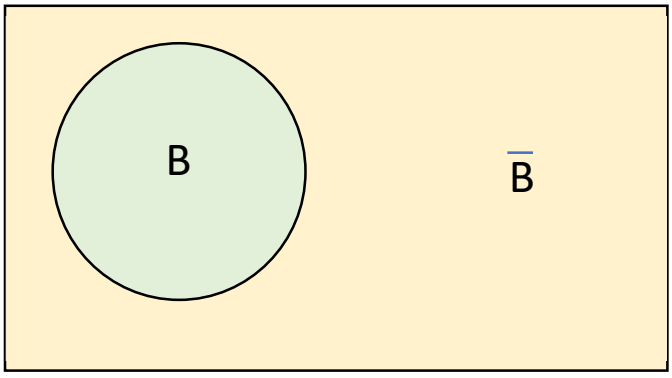
= 0.277 ≠ 0.37

As the probability of cancel a reservation given that the reservation is on a resort hotel is different than the probability of cancel a reservation no matter the hotel type, we can conclude these two events are dependent.

2.4.5 Complementary Event

The complement of event includes all the outcomes for an experiment that are not in A.

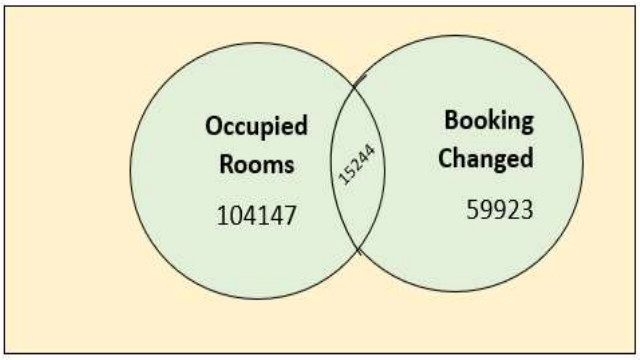
For example: The booking is cancelled, or it is not cancelled.



2.4.6 Intersection event

Let A and B be two events defined in a sample space. The intersection of A and B represents the collection of all outcomes that are common to both A and B and is denoted by A and B or A ∩ B or AB.

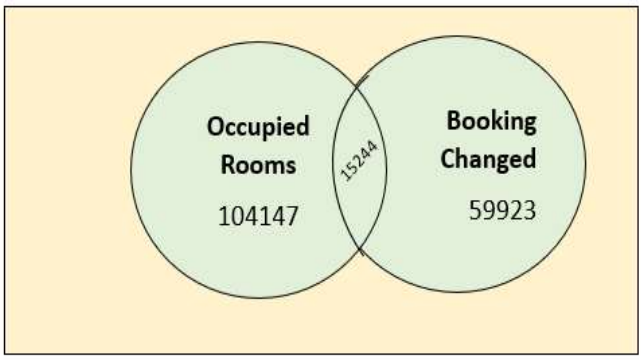
Bookings	Total
P(Occupied Rooms)	75,167
P(Bookings Changed)	119,391
P(Occupied rooms ∩ Bookings Changed)	15,244



2.4.7 Union Event

Let A and B be two events defined in a sample space. The union of events A and B is the collection of all outcomes that belong either to A or to B or to both A and B and is denoted by (A or B) or A ∪ B.

Bookings	Total
P(Occupied Rooms)	75,167
P(Bookings Changed)	119,391
P(Occupied rooms ∪ Bookings Changed)	179,314



$$A \cup B = P(A) + P(B) - P(A \cap B)$$
$$= 75167 + 119391 - 15244$$
$$= 179314$$

2.7.1 Sampling distributions: Standard deviation of a Sample mean

The mean of the lead time is 104 days. The standard deviation of the lead time is 106.831 days. The mean is the same no matter the sample size, so we have that the mean for each sample is 104 days.

If $\frac{n}{N} \leq 0.05$ $\frac{500}{119{,}390} = 0.0042 > 0.05$ The standard deviation for a sample of 500 is

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{107}{\sqrt{500}} = 4.78$

When $\frac{n}{N} > 0.05$ $\frac{10{,}000}{119{,}390} = 0.084 > 0.05$

The standard deviation for a sample of 10,000 is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$= \frac{107}{\sqrt{10{,}000}}\sqrt{\frac{119{,}390 - 10{,}000}{119{,}390 - 1}} = 1.07\sqrt{\frac{109{,}390}{119{,}389}} = 1.07\sqrt{0.916} = 1.024$$

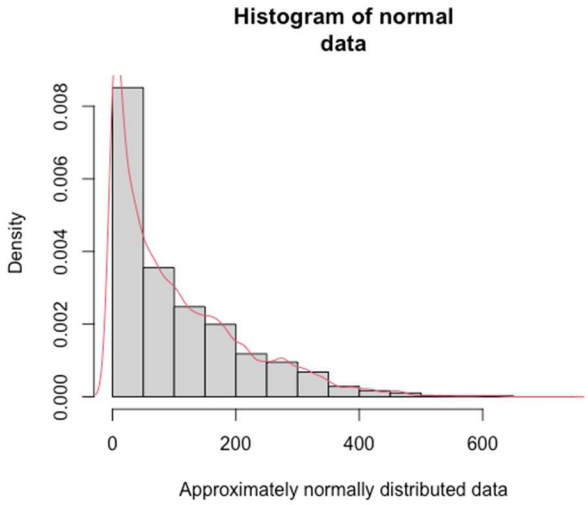
When $\frac{n}{N} > 0.05$ $\frac{50{,}000}{119{,}390} = 0.42 > 0.05$

The standard deviation for a sample of 50,000 is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$= \frac{107}{\sqrt{50{,}000}}\sqrt{\frac{119{,}390 - 50{,}000}{119{,}390 - 1}} = 0.478\sqrt{\frac{109{,}390}{119{,}389}} = 1.07 * 0.7623 = 0.816$$

2.7.2 Sampling distributions: Lead time distribution:



Assuming that the lead time of the hotel booking is approximately normally distributed skewed to the right as the density graph shows for lead time in the reservation data set with a mean of 104 days and a standard deviation of 106.831 days.

Find the probability that the mean of the lead time, of a random sample of 500 hotel booking will be between 95 and 100 days.

Even if the sample size is small as the population distribution is approximately normal, the sample too.

$\mu = 104\ days$

As $\frac{n}{N} \leq 0.05 \quad \frac{500}{119{,}390} = 0.0042 > 0.05$

The standard deviation of a 500 sample is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{107}{\sqrt{500}} = 4.78$$

$P(95 < \bar{x} < 100)$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{95 - 104}{4.78} = \frac{-9}{4.78} = -1.88$$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{100 - 104}{4.78} = \frac{-4}{4.78} = -0.84$$

Looking at the probability chart, the result is

(95<$\bar{x}$<100)=P(-1.88<z<-0.84)

P(-1.88<z)=0.0301

P(z< -0.84)=0.2005

P(95<$\bar{x}$<100)=0.2005-0.0301=0.1704

2.7.3 Sampling distributions: Central limit theorem for sample proportion

0	1	2	3	10
110796	4861	3652	76	1

According to the data set the reservations of hotel booking that include at least one children has a percentage of $\frac{number\ of\ rows\ with\ at\ least\ one\ children}{total\ rows} = \frac{8{,}590}{119{,}386} = 0.072$

Assuming that this is true for all the hotel booking population that includes at least one child. Let $\hat{p}$ be the proportion of hotel booking in a random sample of 1000 who include at least one child in the reservation. Find the mean and standard deviation of $\hat{p}$ and describe the shape of its sampling distribution.

Let p be the proportion of all booking that includes at least one child. P=0.072

$$q = 1 - p = 1 - 0.072 = 0.93$$

The mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p = 0.072$

The standard deviation of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.072 * 0.93}{1000}} = 0.0082$$

As np = 1000(0.072) = 72

As nq = 1000(0.93) = 930

Both are greater than 5, the central limit theorem is applied, so the sampling distribution of $\hat{p}$ is approximately normal with mean 0.072 and standard deviation of 0.0082

2.7.4 Sampling distributions: Application of the sampling distribution of $\hat{p}$

P(booking cancellation with lead time less than 30 days =

$$=\frac{booking\ cancelled\ with\ lead\ time\ less\ than\ 30\ days}{Total\ of\ booking\ with\ less\ than\ 30\ days\ lead\ time} = \frac{6{,}942}{31{,}105} = 0.223$$

22.3% of hotel booking with less than 30 days of lead time are cancelled. Supposing that is true for the current hotel booking population. Let $\hat{p}$ be the proportion in a random sample of 2000 hotel bookings with less than 30 days of lead time that will be cancelled. Find the probability that 20% to 22% of hotel booking in this sample will be cancelled

n = 2000, p = 0.223, and $q = 1 - p = 1 - 0.223 = 0.777$, where p is the proportion of a booking reservation with less than 30 lead time to be cancelled.

The mean of the sample proportion $\hat{p}$ is $\mu_{\hat{p}} = p = 0.223$

The standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.223 * 0.777}{2000}} = 0.0093$

As np = 2000(0.223) = 446

As nq = 2000(0.777) = 1554

As both are greater than 5, the central limit theorem could be applied to infers that the data is distributed approximately normal.

$$p(0.2 < \hat{p} < 0.22)$$

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.2 - 0.223}{0.0093} = -2.47$$

Finding in the z table, the probability is 0.0068

$$z = \frac{\bar{p} - p}{\sigma_{\bar{p}}} = \frac{0.22 - 0.223}{0.0093} = -3.22$$

Finding in the z table, the probability is 0.0006

$$p(0.2 < \hat{p} < 0.22) = 0.0068 - 0.0006 = 0.0062$$

2.7.5 Example to construct 90% confidence level

The owner of a city hotel kept careful records of the lead times. After first 36 records, he found a mean lead time of 105 and standard deviation of 42.

Construct a 90% confidence interval for the mean lead time.

Answer: Here, n> 30, hence we can use the normal distribution.

From the given information, n = 36, $\bar{x}$ = \$105, and $\sigma = \$42$.

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ = 42/ √36= \$ 7

We know that z = -1.65 and z = 1.65.

The 90% confidence interval for μ is $\bar{x} \pm z\sigma_{\bar{x}} = 105 \pm 1.65 * 7$= 93.45 to 116.55

Therefore, 90% confident that lead time is 93.45 to 116.55.

The owner of the resort hotels kept careful records of the lead times. After 49 records, he found a mean lead time for 120 and standard deviation of 52.

Construct a 95% confidence interval for the mean lead time

n = 49, x = 120, s = 52, Confidence level = 95% or .95

$s_{\bar{x}} = \frac{s}{\sqrt{n}}$ = 120/ √49= 120/7= 17.14

To find the value of t from Table V of Appendix B, we need to know the degrees of freedom and the area under the t distribution curve in each tail.

Degrees of freedom = n − 1 = 25 − 1 = 24

Area in each tail = 0.5 − (0.95⁄2) = 0.5 − 0.4750 = 0.025

Therefore z= 2.064 or – 2.064

The 90% confidence interval for μ is $\bar{x} \pm zs_{\bar{x}} = 120 \pm 2.064(17.14)$ = 84.62 to 155.38

2.7.6 Estimation of a Population Proportion: Large Samples

Using the dataset of hotel bookings, we took a sample of 1700 guests.

Eighty percent of the guests included in this sample said that having basic needs met in a hotel is very or extremely important in their vision of the ideal hotel stay to prevent cancellation.

What is the point estimate of the corresponding population proportion?

Find, with a 99% confidence level, the percentage of all guests who will say that having basic needs met is very or extremely important in their vision of ideal hotel stay to prevent cancellation. What is the margin of error of this estimate?

Let p be the proportion of all the guests who will say that having basic needs in a hotel stay met is very or extremely important I their vision of the ideal hotel stay to prevent cancellation, and let $\hat{p}$ be the corresponding sample proportion. From the given information,

n = 1700, $\hat{p}$ = 0.80, and $\hat{q} = 1 - \hat{p} = 1 - 0.80 = 0.20$

First, we calculate the value of the standard deviation of sample proportion as follows:

$S\hat{p} = \sqrt{\hat{p}\,\hat{q}/n} = \sqrt{(0.80)(0.20)/1700} = 0.0097015$

Note that np and ng are both greater than 5. Consequently, the sampling distribution of p is approximately normal and we will use the normal distribution to make a confidence interval about p.

(a) The point estimate of the proportion of all guests who will say that having basic needs met is very or extremely important in their vision of the ideal hotel stay to prevent cancellation is equal to .80; that is.

Point estimate of p= $\hat{p}$ = 0.80

(b) The confidence level is 99%, or .99. To find z for a 99% confidence level, first we find the area in each of the two tails of the normal distribution curve, which is (1 - .99)/2 =.0050.

Then, we look for .0050 and .0050 + .99 = .9950 areas in the normal distribution table to find the two values of z. These two z values are (approximately) -2.58 and 2.58. Thus, we will use 2=2.58 in the confidence interval formula. Substituting all the values in the confidence interval formula for p, we obtain $\hat{p}$ + z $S_{\hat{p}}$; = .80 ± 2.58 (0.0097015) = .80 ± 0.025 = .775 to .825 or 77.5% to 82.5%.

Thus, we can state with 99% confidence that .775 to .825 or 77.5% to 82.5% of all guests will say that having basic needs met is very or extremely important in their vision of the of the ideal hotel stay to prevent cancellation. The margin of error associated with this estimate of p is .025 or 2.5%, that is,

Margin of error = z $S_{\hat{p}}$ = 0.026 or 2.6%

2.7.7 Hypothesis Tests : Mean

Let μ be the mean lead time in days of a booking cancellation, and let $\bar{x}$ be the corresponding mean for the sample.

$$\mu = 104\ days$$

From the given information,

n = 300, $\bar{x}$ = 109 days , and s = 111.386 days

To calculate the p-value and to make the decision, we apply the following four steps.

$H_0 : \mu =$ 104 days The mean of the lead time is the same

$H_1 : \mu \neq$104 days The mean of the lead time is different

The t distribution must be used since it is a big sample and population σ is unknown

Area in each tail α/2 = .01/2=0.005

$$df = 299$$

From the distribution table, the rejection interval is 2.601 and – 2.601

Calculating the test statistic,

$$s_x = \frac{s}{\sqrt{n}} = \frac{111.386}{\sqrt{300}} = 6.4308$$

$$t = \frac{\hat{x} - \mu}{s_x} = \frac{109 - 104}{6.43087} = 0.777$$

The test statistic value 0.777 falls on the nonrejection region. So we cannot conclude that the mean differ and it must be different due to sampling error. The mean lead time of the sample is not significantly different from the population.

2.7.8 Hypothesis tests : Population proportion

According to all the data set, 37% of hotel booking ends up in cancellation. Supposing that is true for the bookings due to arrive between the 1st of July of 2015 and the 3rd of August 2017. A sample of 300 bookings shows that the proportion is 35%.

Find the p-value to test the hypothesis that the current percentage of hotel booking cancellations is different from 37%. What is your conclusion if the significance level is 5%?

$$n = 300; \hat{p} = 0.35; \propto = 0.05$$

$$p = 0.37; q = 0.63$$

$H_0 : p = 0.37$ The current percentage is 0.37

$H_1 : p \neq 0.37$ The current percentage differs from 0.37

$$np = 300 * 0.37 = 111$$

$$nq = 300 * 0.63 = 189$$

As np and nq are greater than 5, the sample size is large enough to use the normal distribution approximation.

α/2 = .05/2=0.025

The critical values are -1.96 and 1.96

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.37 * 0.63}{300}} = 0.0279$$

$$z = \frac{\hat{p} - p}{\sigma_p}$$

$$= \frac{0.35 - 0.37}{0.0279}$$

$$= -0.717$$

The test statistic -0.717 falls on the do not rejection error. We cannot conclude that the sample proportion is significantly different from the population proportion.

2.5 Logistic Regression

Given that 0 means on the data set that a booking was cancelled and 1 mean that the booking was not cancelled, if y in the logistic regression is close to 1 means the booking was not cancelled.

```
Call:
lm(formula = BookingDf$is_canceled ~ lead_time + babies + children +
    adults, data = BookingDf)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2025 -0.3305 -0.2387  0.5303  1.0663

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1939036  0.0045589  42.533  < 2e-16 ***
lead_time    0.0013113  0.0000126 104.092  < 2e-16 ***
babies      -0.1350975  0.0137112  -9.853  < 2e-16 ***
children     0.0192033  0.0033550   5.724 1.04e-08 ***
adults       0.0211030  0.0023233   9.083  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4613 on 119381 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.08756,   Adjusted R-squared:  0.08753
F-statistic:  2864 on 4 and 119381 DF,  p-value: < 2.2e-16
```

$$\hat{y} = 0.19 + 0.001x - 0.13x' + 0.02x'' + 0.21x'''$$

This equation means that the relationship between cancellation and lead time is positive but not very high. Which allow to conclude that the more lead time there are the closer to 1 y is going to be, so the more lead time, the more probability the booking is not going to be cancelled.

By other hand, babies and lead time have a negative relationship which means that including one baby, increase the probability of a booking to be cancelled.

3. RESULTS

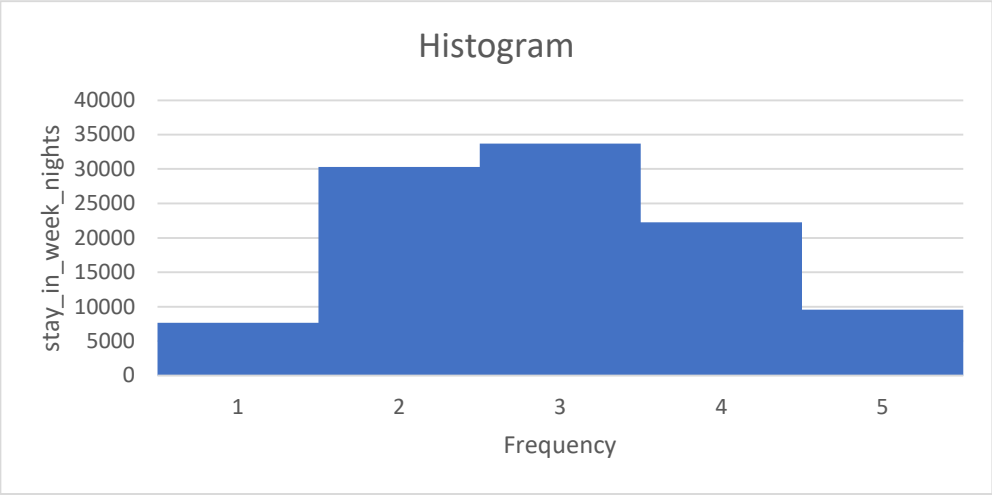
3.1 Presentation of Results

- With the Hotel Booking dataset the following has been concluded.
- The data has been cleaned and preprocessed to explore the exploratory data analysis for in-depth analysis.

i) Number of bookings that were confirmed and cancelled?
ii) Number of children guests had
iii) What is the booking ratio between Resort Hotel and City Hotel?
iv) What is the percentage of hotel booking trends in the year 2015, 2016 and 2017?
v) Which is monthly Hotel booking trends of City and Resort Hotel?
vi) What is the total number of nights stay in the City and Resort Hotel?

3.2 Data Visualization:

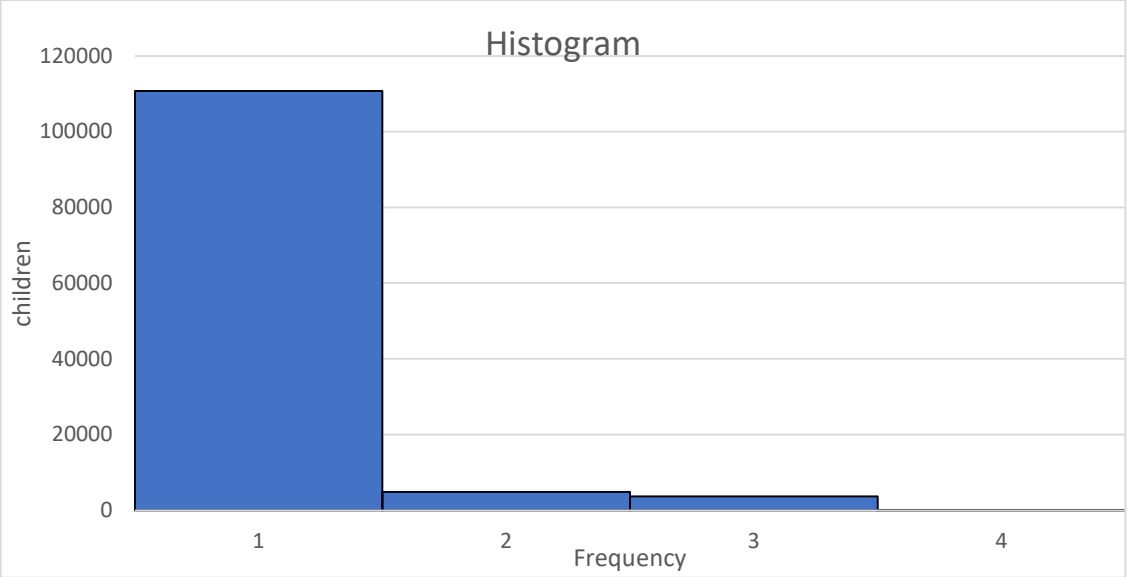
1. **What is the total number of nights stay in the City and Resort Hotel?**



The shape of the histogram for the quantitative variable-stays_in_week_nights suggests that it is normally distributed. We can infer that guests that are 3 in number are the ones that stay in week

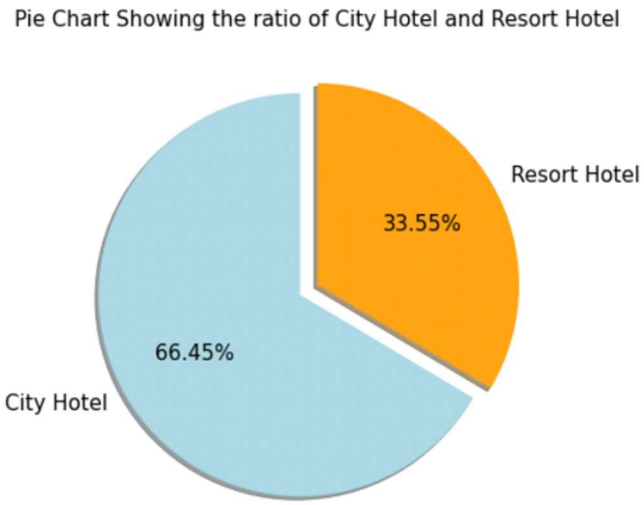
nights the most i.e., the number of entries of a set of 3 guests on week nights are approximately 34000.

2. **Total number of children guests had**



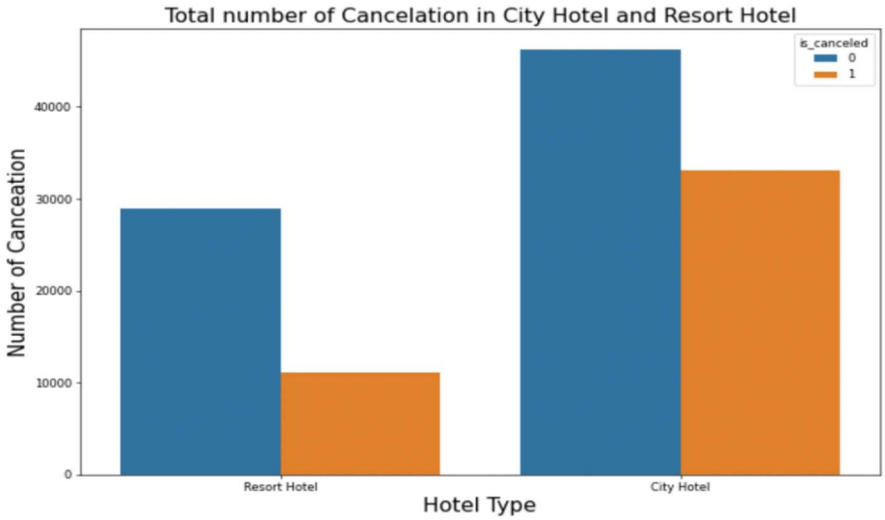
The shape of the histogram for the quantitative variable-children is right-skewed, hence it is positive. We can infer that guests with 1 child were highest in number i.e., approximately 11000.

3. What type of hotel is most preferred? City Hotel or Resort Hotel?



From this pie chart, we can infer that the number of city hotels at 66.45% are way more in number than the number of Resort hotels at 33.55%.

4. Total number of Cancelation in City Hotel and Resort Hotel

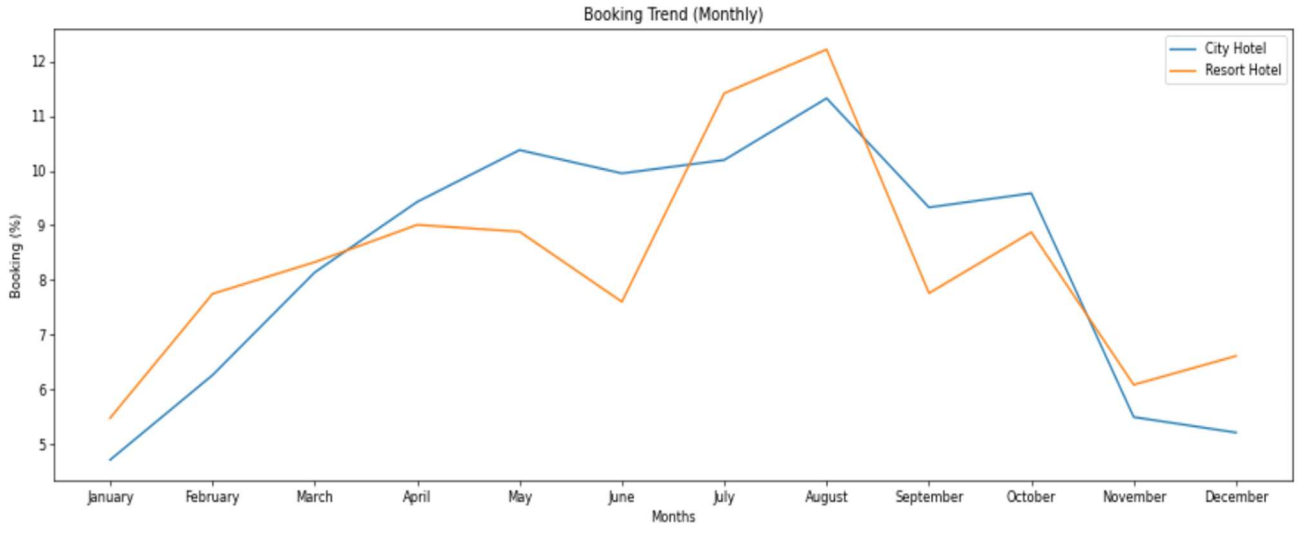


This graph depicts the number of cancellations for city and resort hotels. From this graph, we can see that the number of cancellations is lowest for resort hotel and the number of cancellations that didn't come through is highest for city hotel.

5. Number of total Bookings in the year 2015, 2016 and 2017

This graph depicts the number of bookings in city and resort hotels in the years 2015, 2016 and 2017. By this comparison, we can infer that in the year 2016, city hotel had the highest number of bookings while the least number of bookings was for resort hotel in the year 2015. On the whole, the year 2016 saw an increase in number of bookings for both types of hotels.

6. Monthly Trend for the bookings



This line-plot shows us the booking trend for city and resort hotel throughout the year. From this, we can notice that the bookings for both the hotels was the highest in the month of august i.e., towards the end of summer and the month that saw the least number of bookings was January.

4. Conclusion:

4.1 Key Findings

- The cancellation rate is higher is city hotels.
- More than 60% of the population booked the City Hotel.
- More than 30% of the population booked the Resort Hotel.
- Most booking were made during the month of July to August.
- Least bookings were made during the start and the end of the year.
- Higher booking was made for Resort hotel during the beginning and end of the year.
- During June and September, the booking has reduced.

4.2 The Business Insights/ Recommendations:

It is a well-known fact that data and revenue management in the hotel industry has a very important role. In a very similar manner, data science is also equally important in order to maximize the distribution as well as profit of a hotel.

Measuring the hotel bookings and cancellations would help us understand the problem, especially when deploying data analytics.

Regression analysis is needed to forecast the hotel's revenue and costs, group their guests on similar behaviors for understanding their opinions and thereby form effective marketing campaigns. Based on such analysis we can implement strict cancellations policies such as 24/48-hour notice, non-refundable rates, etc.

Appendix:

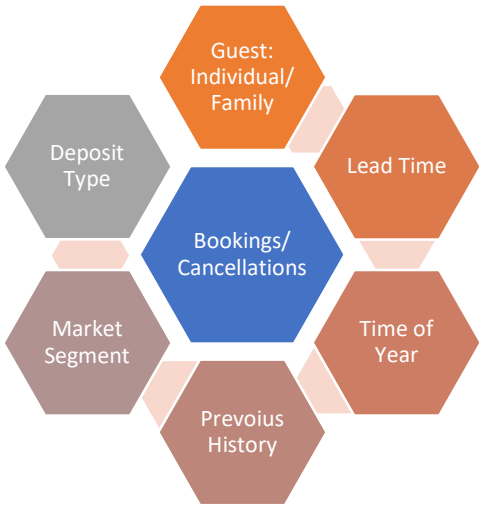


Figure 1.1 The booking cancellation is directly or indirectly dependent on these factors

Columns	Type	Description
Hotel	Categorical	Resort Hotel or City Hotel
is_canceled	Integer	canceled = 1 or not cancelled = 0
lead_time	Integer	Number of days between booking and check-in dates.
arrival_date_year	Date	Year of arrival date
arrival_date_month	Date	Month of arrival date
arrival_date_week_number	Date	Week number of year for arrival date
arrival_date_day_of_month	Date	Day of arrival date
stays_in_weekend_nights	Integer	Number of weekend nights (Saturday or Sunday)
stays_in_week_nights	Integer	Number of week nights (Monday to Friday)
adults	Integer	Number of adults
children	Integer	Number of children
babies	Integer	Number of babies
country	Categorical	Country of origin presented in ISO 3155–3:2013 format

market_segment	Categorical	"TA" = "Travel Agents" and "TO" = "Tour Operators"
distribution_channel	Categorical	"TA" = "Travel Agents" and "TO" = "Tour Operators"
is_repeated_guest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)
booking_changes	Integer	Number of changes made to the booking from the moment the booking was entered on the PMS
customer_type	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group–when the booking is associated to a group; Transient–when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party–when the booking is transient, but is associated to at least other transient booking
adr	Numerical	Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

Table 1.2. Description of the variables in the data set

Min	1st. Qu.	Median	Mean	3rd Qu.	Max
0	18	69	104	160	737

Table 1.3. Lead Time is 104 days

Min	1st. Qu.	Median	Mean	3rd Qu.	Max
00.000	2.000	2.000	1.856	2.000	55.000

Table 1.4. Guest type. Most bookings are done by 2 individuals. This gives the decision makers the insight of what kind of rooms are preferred by the guests.

Min	1st. Qu.	Median	Mean	3rd Qu.	Max
-6.38	69.29	94.58	101.83	126.00	5400.000

Table 1.5. Average daily rate. The average daily rate (ADR) is 101.83 which reflects the revenue earned for an occupied room per day. The operating performance of a hotel or other lodging business can be determined by using the ADR.

Using RStudio to import our Hotel dataset:

Summary information on Hotel Booking data frame

```
summary(hb.df)
```

Figure 1.2. Summary of hotel bookings

Finding the number of columns in the data set.

```
names(hb.df)
```

```
##  [1] "hotel"                      "is_canceled"
##  [3] "lead_time"                  "arrival_date_year"
##  [5] "arrival_date_month"         "arrival_date_week_number"
##  [7] "arrival_date_day_of_month"  "stays_in_weekend_nights"
##  [9] "stays_in_week_nights"       "adults"
## [11] "children"                   "babies"
## [13] "country"                    "market_segment"
## [15] "distribution_channel"       "is_repeated_guest"
## [17] "previous_cancellations"     "previous_bookings_not_canceled"
## [19] "reserved_room_type"         "assigned_room_type"
## [21] "booking_changes"            "deposit_type"
## [23] "customer_type"              "adr"
```

Figure 1.3. Variables in the data set

Number of bookings in City Hotel and Resort Hotel.

```
table(hb.df$hotel)
```

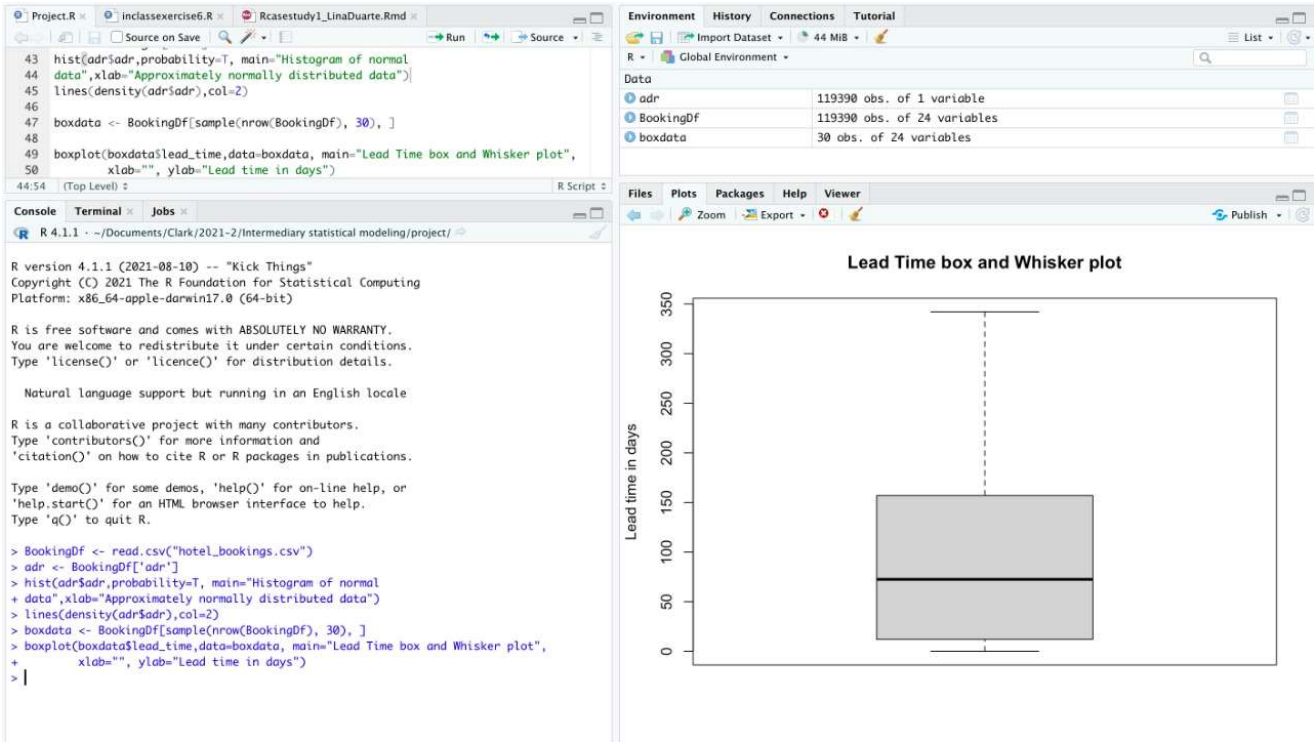
```
##
##   City Hotel Resort Hotel
##        79330        40060
```

Figure 1.4. The data of two categories of City Hotel vs. Resort Hotel. are compared based on other dependent variables.

```
library(dplyr)

#subset the data to count the frequency country
flow2 <- hb.df  %>%
  count (hb.df$country)

# providing the countries names
hb_seg2 <- flow2$`hb.df$country`[which(flow2$n > 50)]
cat(hb_seg2)


# providing the countries numbers
hb_se32 <- flow2$n[which(flow2$n > 50)]
cat(hb_seg32)

#combo
country_50 <-flow2  %>% filter(n >= 50)

# one code - n is the default of count unless changed ie count(dataname , name = " string" )
flow3 <- hb.df  %>%
  count (hb.df$country) %>%
  filter(n >= 50)
```

Figure 6. Country-wise Data Cleaning up the data where countries with at least 50 bookings are included for this research project. Filtering out the data where the booking value is greater than 50 using the R code. There are a total of 178 countries.

Box and Whisker Plot



Logistic Regression

```
regress<-lm(BookingDf$is_canceled ~ lead_time + babies + children + adults, data = BookingDf)
summary(regress)
```

Using Python Programming for dataset description:

Importing all the necessary libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as smf
import statsmodels.api as sm
import patsy
import matplotlib.ticker as ticker
%matplotlib notebook
```

Importing the main "Hotel Booking" Dataset:

```
data = "D:/Rubina/Clark Univeristy/Semester 1/STAT 4600-02 (Intermed Stats Model Analytics)/PROJECT/hotel_bookings.csv"
df = pd.read_csv(data)
```

Displaying the raw format of Hotel Booking Dataset:

```
df.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	a
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	
2	Resort Hotel	0	7	2015	July	27	
3	Resort Hotel	0	13	2015	July	27	
4	Resort Hotel	0	14	2015	July	27	

5 rows × 32 columns

Note: The above dataset contains 119390 rows and 32 columns which comprises of hotel type, booking and cancellation, lead time, arrival date year, stay in weekends and nights, number of adults, children and babies, deposit type, agent, company, customer type, adr, required car parking spaces, special guests and reservation details.

Locating null values in the dataset:

```
print(hotel_df.isnull())
```

```
        hotel  is_canceled  lead_time  arrival_date_year  arrival_date_month
\
0       False        False      False              False               False
1       False        False      False              False               False
2       False        False      False              False               False
3       False        False      False              False               False
4       False        False      False              False               False
...       ...          ...        ...                ...                 ...
119385  False        False      False              False               False
119386  False        False      False              False               False
119387  False        False      False              False               False
119388  False        False      False              False               False
119389  False        False      False              False               False

        arrival_date_week_number  arrival_date_day_of_month  \
0                          False                      False
1                          False                      False
2                          False                      False
3                          False                      False
4                          False                      False
...                          ...                        ...
119385                     False                      False
119386                     False                      False
119387                     False                      False
```

```
119388                    False                       False
119389                    False                       False

        stays_in_weekend_nights  stays_in_week_nights  adults  ...  \
0                         False                 False   False  ...
1                         False                 False   False  ...
2                         False                 False   False  ...
3                         False                 False   False  ...
4                         False                 False   False  ...
...                         ...                   ...     ...  ...
119385                    False                 False   False  ...
119386                    False                 False   False  ...
119387                    False                 False   False  ...
119388                    False                 False   False  ...
119389                    False                 False   False  ...

        deposit_type  agent  company  days_in_waiting_list  customer_type  \
0              False   True     True                 False          False
1              False   True     True                 False          False
2              False   True     True                 False          False
3              False  False     True                 False          False
4              False  False     True                 False          False
...              ...    ...      ...                   ...            ...
119385         False  False     True                 False          False
119386         False  False     True                 False          False
119387         False  False     True                 False          False
119388         False  False     True                 False          False
119389         False  False     True                 False          False

          adr  required_car_parking_spaces  total_of_special_requests  \
0       False                        False                      False
1       False                        False                      False
2       False                        False                      False
3       False                        False                      False
4       False                        False                      False
...       ...                          ...                        ...
119385  False                        False                      False
119386  False                        False                      False
119387  False                        False                      False
119388  False                        False                      False
119389  False                        False                      False

        reservation_status  reservation_status_date
0                    False                    False
1                    False                    False
2                    False                    False
3                    False                    False
4                    False                    False
...                    ...                      ...
119385               False                    False
119386               False                    False
119387               False                    False
119388               False                    False
```

Note: The main dataset includes 119390 rows and 32 columns. The raw format of Hotel Booking data does contain null/ missing values.

Removing the NA values and unwanted columns from the main dataset.

```
hotel_df=df.drop(['reserved_room_type','meal','agent','required_car_parking_spaces','days_in_waiting_list','total_of_special_requests','reservation_status','reservation_status_date'], axis=1)
hotel_df.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	a
0	Resort Hotel	0	342	2015	July	27	
1	Resort Hotel	0	737	2015	July	27	
2	Resort Hotel	0	7	2015	July	27	
3	Resort Hotel	0	13	2015	July	27	
4	Resort Hotel	0	14	2015	July	27	

5 rows × 24 columns

```
hotel_df.isnull().sum().sort_values(ascending=False)[:10]
```

```
company                         112593
country                            488
children                             4
market_segment                       0
customer_type                        0
deposit_type                         0
booking_changes                      0
assigned_room_type                   0
previous_bookings_not_canceled       0
previous_cancellations               0
dtype: int64
```

Modeling and Analysis

1. Which type of Hotel is most preferred? City Hotel or Resort Hotel?

```
plt.rcParams['figure.figsize']=8,8
labels=hotel_df['hotel'].value_counts().index.tolist()
sizes=hotel_df['hotel'].value_counts().tolist()
colors=['lightblue','orange']
explode=(0.1, 0)
plt.title("Pie Chart Showing the ratio of City Hotel and Resort Hotel", fontsize=18)
plt.pie(sizes,labels=labels,colors=colors,explode=explode,autopct='%1.2f%%',shadow= True, startangle=90, textprops={'fontsize':18})
plt.show()
```

2. Total number of Cancelation in City Hotel and Resort Hotel

```
hotel_df.groupby(['arrival_date_year'])['is_canceled'].mean()
```

```
arrival_date_year
2015    0.370158
2016    0.358633
2017    0.386979
Name: is_canceled, dtype: float64
```

From the above code, it can be indicated that the cancellation rate was highest in the year 2017 and lowest in the year 2016.

```
plt.figure(figsize=(12,8))

sns.countplot(x='hotel',hue='is_canceled', data=hotel_df)
plt.title("Total number of Cancelation in City Hotel and Resort Hotel", fontsize=18)
plt.xlabel("Hotel Type", fontsize=18)
plt.ylabel("Number of Canceation", fontsize=18)
plt.show()
```

3. Number of total Bookings in the year 2015, 2016 and 2017

```
plt.subplots(figsize=(12,8))
sns.countplot(x='arrival_date_year', hue='hotel',  data=hotel_df);
plt.title("Total number of Bookings in the year 2015, 2016, 2017", fontsize=18)
plt.xlabel("Year", fontsize=18)
plt.ylabel("Number of Bookings", fontsize=18)
plt.show()
```

4. Total number of Month-wise Bookings

```
## Order of months
new_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October',
             'November', 'December']


## Select only City Hotel
sorted_months = hotel_df.loc[df.hotel=='City Hotel' ,'arrival_date_month'].value_counts().reindex(new_order)
x1 = sorted_months.index
y1 = sorted_months/sorted_months.sum()*100

## Select only Resort Hotel
sorted_months = hotel_df.loc[hotel_df.hotel=='Resort Hotel' ,'arrival_date_month'].value_counts().reindex(new_order)
x2 = sorted_months.index
y2 = sorted_months/sorted_months.sum()*100

## Draw the line plot
fig, ax = plt.subplots(figsize=(18,6))
ax.set_xlabel('Months')
ax.set_ylabel('Booking (%)')
ax.set_title('Booking Trend (Monthly)')
sns.lineplot(x1, y1.values, label='City Hotel', sort=False)
sns.lineplot(x1, y2.values, label='Resort Hotel', sort=False)
plt.show()
```

References

[1] Data Mining Basics - What is Data Mining? (2021, March 29). Retrieved from https://www.sisense.com/glossary/data-mining-basics/

[2] What is Data Processing? Definition and Stages - Talend Cloud Integration. (n.d.). Retrieved from https://www.talend.com/resources/what-is-data-processing/

Antonio, N., Almeida, A. D., & Nunes, L. (2018, November 29). Hotel booking demand datasets. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352340918315191.

[3] Data in Brief. (n.d.). Retrieved from https://www.sciencedirect.com/journal/data-in-brief/vol/16.

(n.d.). Retrieved from https://web-p-ebscohost-com.goddard40.clarku.edu/ehost/detail/detail?vid=9&sid=10494d06-0d74-4ef5-b1ca-0303d6695640@redis&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ==#AN=128358351&db=aph.Mostipak, J. (2020, February 13). *Hotel Booking Demand*. Kaggle. Retrieved December 14, 2021, from https://www.kaggle.com/jessemostipak/hotel-booking-demand.