



BENGALURU

HOTEL BOOKINGS ANALYSIS

**USING MYSQL, EXCEL AND
POWER BI**

PREPARED BY: ZUBAIR BAIG

ABSTRACT

This project aims to conduct a comprehensive analysis of hotel booking data to uncover patterns, trends, and insights that can inform strategic decision-making in the hospitality industry. By leveraging data sourced from a MySQL database and analyzing it using Power BI, our primary objective is to address specific problem statements related to hotel bookings.

The analysis begins with the collection and loading of data, followed by exploratory data analysis (EDA), we examine the distribution and characteristics of key variables, utilizing descriptive statistics, visualizations.

The problem statements cover a range of issues, including identifying peak booking periods, understanding cancellation behaviors, and optimizing room pricing strategies. Solving these problems will provide actionable insights and trends within the data, which are presented through detailed visualizations.

Key findings are presented through insightful visualizations, highlighting actionable insights and trends within the data.

The results of this analysis provide valuable recommendations for improving booking strategies, enhancing customer satisfaction, and reducing cancellation rates. This project demonstrates the power of data-driven decision-making in the hospitality industry and lays the groundwork for future research and analysis.

Table of contents

1. Introduction.
2. Objectives.
3. Data Loading.
4. ER Diagram.
5. Problem Statements.
6. Exploratory Data Analysis.

INTRODUCTION

The hotel industry is an important and competitive sector that relies heavily on data driven decision making, helping business to enhance customer satisfaction, pricing strategies, and to improve operational efficiency.

We aim to analyze important aspects of hotel booking, using our 10 pre-defined problem statements.

To address that, we have used a dataset which contains real-world data of Booking details. This dataset contains 1,19,390 detailed record for each and every booking as well as the historical information, cancellation records and other crucial information. Included in 9 comprehensive table, some of the important tables:

- **Booking_Details** - It includes a unique booking identifier and information about the type of hotel (Resort Hotel or City Hotel). Additionally, it records the booking's cancellation status (0 for not canceled, 1 for canceled) as well as arrival date information.
- **Room_Details** - The Room_Details table provides information related to room reservations and changes made to them. It is associated with the Booking_Details table via the booking identifier.
- **Reservation_Status** - The Reservation_Status table records the status of reservations over time. This table captures the reservation's last status (e.g., Canceled, Check-Out) and the date on which this status was recorded.
- **Guest_Info** - This table records the number of adults, children, and babies accompanying the booking, offering an understanding of the composition of guests for each reservation.
- **Meal_And_Stay_Details** - This table includes the type of meal booked (e.g., Bed & Breakfast, Half Board), the Average Daily Rate (ADR) for the stay, the number of required car parking spaces, and the total count of special requests made by the guest.
- **Booking_Source_and_History** – This table is crucial for understanding the source of bookings and the historical behavior of guests. This table encompasses information such as the market segment (e.g., Online Travel Agents, Direct Booking), distribution channel (e.g., Online Travel Agents, Direct Booking), and whether the guest is a repeated visitor (0 for not repeated, 1 for repeated) and others.

Objectives

The primary objective of this project is to analyze the data and find key factors for data-driven decision making, the project also emphasizes the art of querying and visualization through continuous challenging and unique approach, providing concise insights, visualization and conclusion for each and every problem statement.

To achieve that, we will be analyzing the data-insights and providing conclusion to problem statement in a structured manner, divided into 6 **sections**, and those are:

1. Overview

- This section contains a brief explanation of problem statement and its significance, it provides background information and context necessary to understand the problem and provide necessary insights.

2. Objectives

- This section would outline the goals we need to aim to achieve the analysis, quote specific questions, as well as the discussion of best method to achieve our goal.

3. Data Preparations

- Data preparations would contain the steps taken to create calculative tables/cleaning necessary for get our specific goal.

4. Query/DAX approach

- Detailed explanation of the methods and techniques used for Querying data or visualizations.

5. Observations

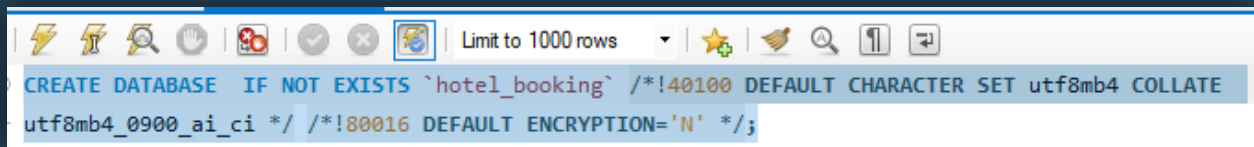
- The section would discuss the key findings or patterns from the analysis, it may also contain necessary charts and major insights.

6. Conclusion

- This section would consist a final verdict on the analysis and provide necessary explanation for the problem statement.

Data Loading

The Data for this project was provided in form of MySQL file. The data is present in form of a query which will be imported into a database once we run the query.



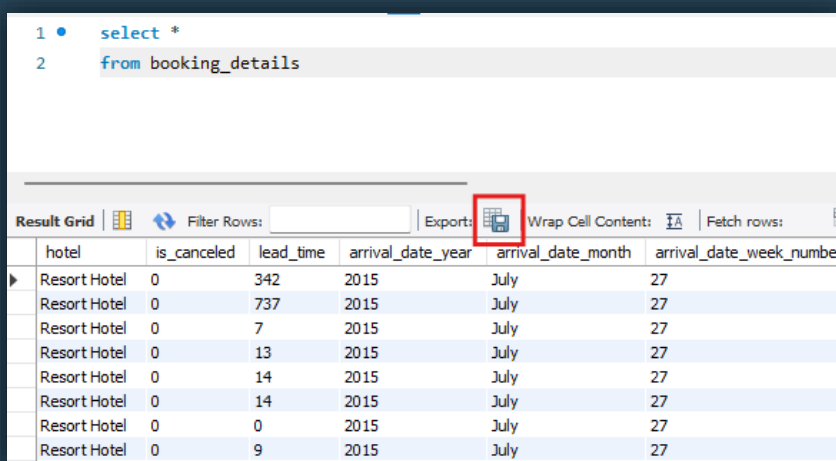
```
CREATE DATABASE IF NOT EXISTS `hotel_booking` /*!40100 DEFAULT CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci */ /*!80016 DEFAULT ENCRYPTION='N' */;
```

Once the data is imported to MySQL, we could now proceed to transform and process the data to multiple platforms.

And Upon observation the dataset clearly formatted and required minimal cleaning, hence we could proceed to import the data to other tools.

Exporting data as CSV file

To convert a data present in an SQL database we could simply query the enter table and click export option. And in Windows file saver we will also get option to save as .XLSX files, however in this project we have saved as CSV files, and below is the example to query booking_details table and export the same,



```
1 • select *
2   from booking_details
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number
▶	Resort Hotel	0	342	2015	July	27
	Resort Hotel	0	737	2015	July	27
	Resort Hotel	0	7	2015	July	27
	Resort Hotel	0	13	2015	July	27
	Resort Hotel	0	14	2015	July	27
	Resort Hotel	0	14	2015	July	27
	Resort Hotel	0	0	2015	July	27
	Resort Hotel	0	9	2015	July	27

SQL query to select all columns

Importing data to Power BI

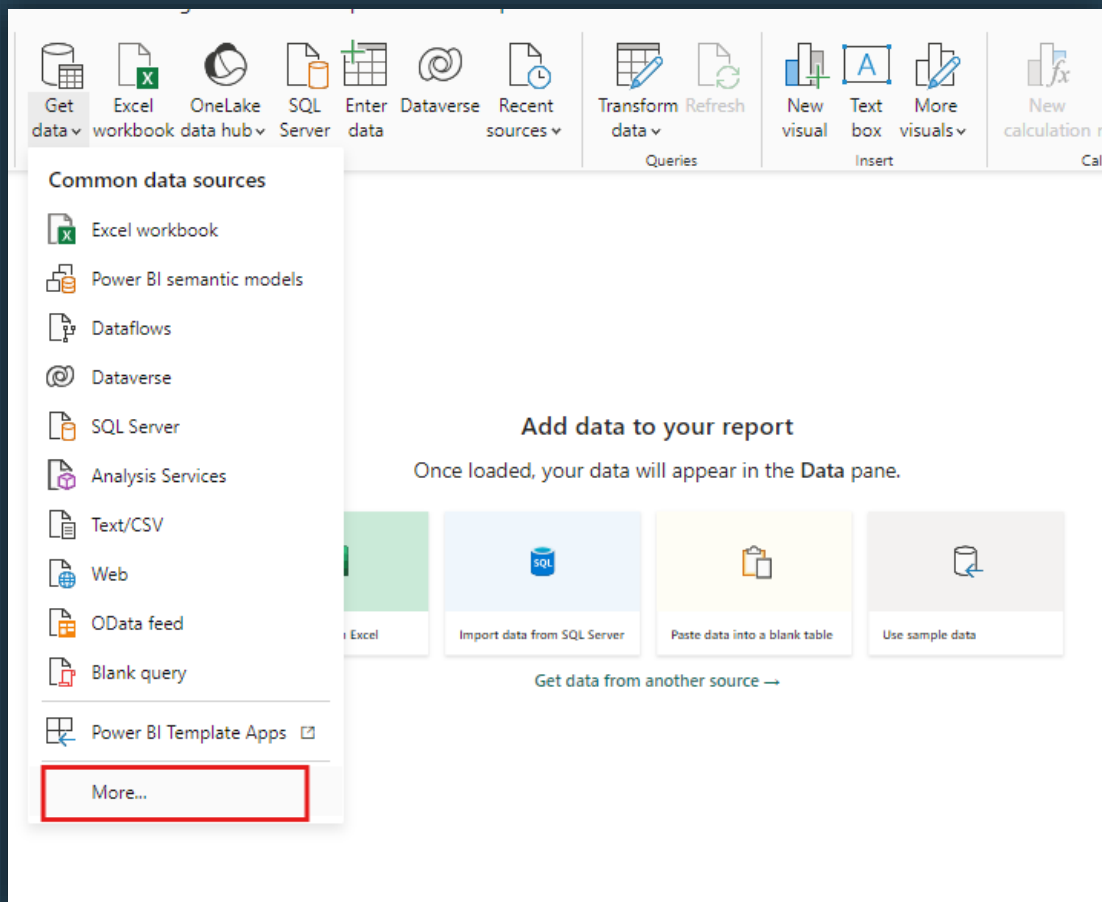
In this we will be working with Microsoft Power BI desktop, a similar steps or approach are required for Power BI service or Premium platforms.

To import the data into power BI, we could either use CSV files we imported from MySQL, however, that would require additional work, in case if we have many tables.

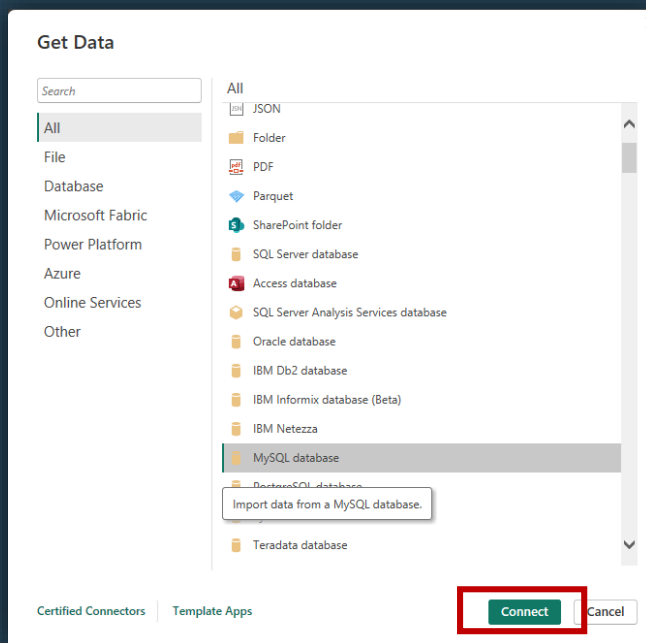
In order to import all the tables at once we used import from MySQL database feature, which will directly load the database from MySQL to Power BI.

Below are the steps followed,

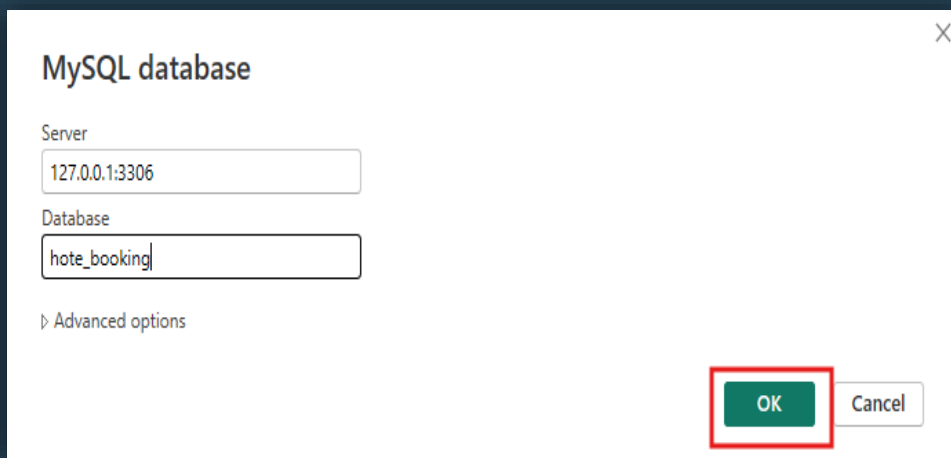
1. Select **Get data** icon, in top left corner and we will get the drop down of various platforms through which we import data, in our case, we would need to select **more**.



2. Scroll down to select **MYSQL database**, and click **connect**.



3. Then we could enter our server and port number, and click **okay**.



4. We then get connected to our database and power BI navigator lets user view all the tables that are present in a database, hence we could select the tables we need and click **load** for power BI to import the tables. This helps us to import all our tables at once.

Navigator

127.0.0.1:3306: hotel_booking [11]

- ☒ hotel_booking.booking_details
- ☒ hotel_booking.booking_source_and_history
- ☒ hotel_booking.country
- ☒ hotel_booking.df_booking_details
- ☒ hotel_booking.df_booking_source_and_his...
- ☒ hotel_booking.distribution_channel
- ☒ hotel_booking.guest_info
- ☒ hotel_booking.market_segment
- ☒ hotel_booking.meal_and_stay_details
- ☒ hotel_booking.reservation_status
- ☒ hotel_booking.room_details

hotel_booking.room_details
Preview downloaded on Friday

reserved_room_type	assigned_room_type	booking_changes	Unnamed: 3	Bool
C	C	3	null	7
C	C	4	null	3
A	C	0	null	d
A	A	0	null	4
A	A	0	null	7
A	A	0	null	d
C	C	0	null	e
C	C	0	null	b
A	A	0	null	a
D	D	0	null	b
E	E	0	null	fi
D	D	0	null	3
D	E	0	null	d
G	G	1	null	2
E	E	0	null	d
D	E	0	null	a
E	E	0	null	1
A	E	0	null	7
A	G	0	null	1
G	G	0	null	a
F	F	0	null	4
A	A	1	null	5

Select Related Tables

Load Transform Data Cancel

ER Diagram

Overview

An ER(Entity-Relationship) diagram is a graphical tool used to model and represent the data structure of a database.

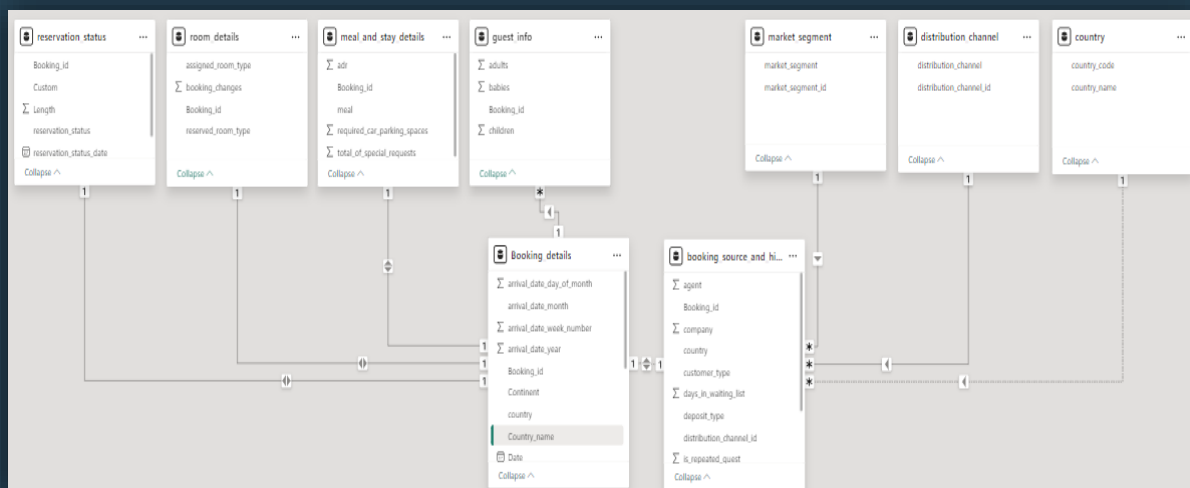
And in this project, we have conventionally utilized Fact, Query table concept to organize the ER-Diagram and to understand the relationship between the tables.

1. **Fact table:** A table that specifically designed to record all the events that occurred during an event. This tables contains an unique ID for each and every record, which is crucial to understand attributes of events and to relate it other tables, Few examples include, orders_table, whether_records, sales_records.
2. **Query table:** A table which contains all the records of elements or categories which are involved or necessary to understand an event, it remains a reference to fact table to get the required information, For instance, Location data, Product details, Customer profile.

A fact-query concept helps to understand attributes, connections and type of relationships.

And after differentiating the tables the facts tables are placed in the **center**, and all the query tables are placed at the **top**. This method helps to understand the relationship in simpler way.

Data Model



Exploratory Data Analysis

1. Understand the distribution of arrival dates, including the most common arrival days and summary statistics for lead times.

Overview:

A **Distribution** is nothing but the plot of a reading/object vs it's occurrence/frequency, hence by plotting the distribution of arrival dates we could realize the dates which on which arrivals are usually high.

Lead times is the time taken for booking to convert it from lead into confirmed bookings, statistical summary includes measures of central tendencies, Variation, Standard Deviation and others, by calculating statistical reading we could derive insightful data about lead times.

Objectives:

- Our objective is to create a distribution that will demonstrate the occurrence of a date in a complete year cycle. This will make a distribution aggregating bookings for each year cycle.
- Find the top 10% dates which has the highest arrivals.
- The second goal for this problem statement is to calculate summary statistics for lead time, using MySQL.

Data Preparation:

Since the arrival date is captured in three separate columns each containing year, month and month day, we could add a new column, with the help of the **CONCATENATE** function as illustrated below,

```
Month_date = CONCATENATE('hotel_booking booking_details'[arrival_date_day_of_month], 'hotel_booking booking_details'[arrival_date_month])
```

month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	country	Bool	Month_date
	25	15	0	2	PRT	4d1c2a	15June
	25	15	0	2	PRT	79e66e	15June
	25	15	0	2	PRT	6fec247	15June
	25	15	0	2	PRT	908f62c	15June
	25	15	0	2	PRT	a73ac9	15June
	25	15	0	2	PRT	7c8aa7	15June
	25	15	0	2	PRT	e3f9748	15June
	25	15	0	2	PRT	aef0544	15June
	25	15	0	2	PRT	5e2462	15June

Approach

Creating distribution to find most common dates:

To achieve the distribution we would need to aggregate the count of our custom Month_date column by the same Month_date column

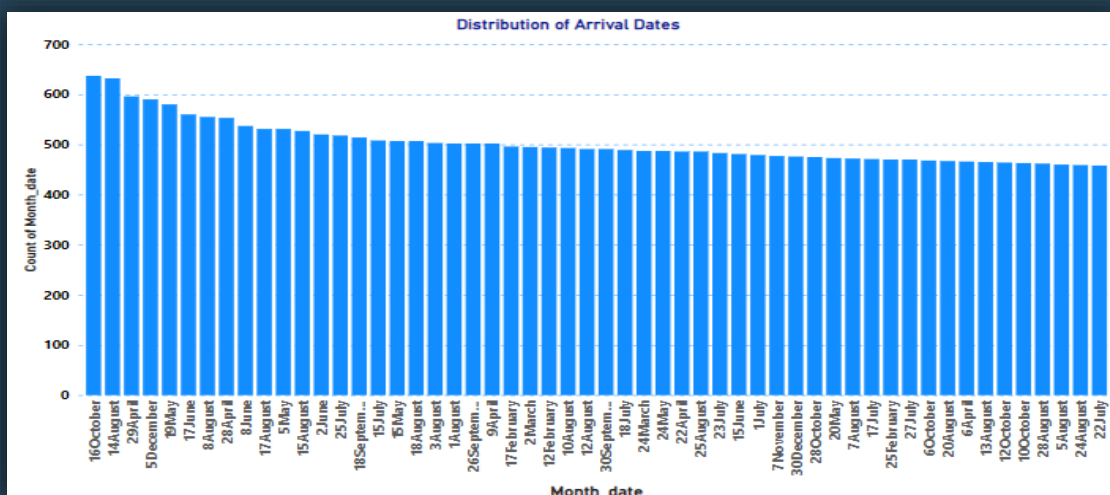
X-axis

Month_date

Y-axis

Count of Month_date

Outcome



Distribution of dates of a year

To summarize lead time: MySQL is not a programming language yet it is powerful, with the help of only in build functions, we could mimic the statistical function as describe() function in python.

In our approach we will be utilizing **union all** clause to align statistical aggregations in rows, each individual query will have aggregated respective functions to get the desired outcome.

Now there is not in-build function to calculate quartiles directly, but we could take help of **ntile()** function, ntile clause is a window function which divides all the records into desired divisions.

In our case quartiles, hence the dataset is divided into 4 parts and ordered in ascending, and if we retrieve the minimum of each division we can get the 1st, 2nd and 3rd quartiles respectively.

Below the MySQL Query for the same,

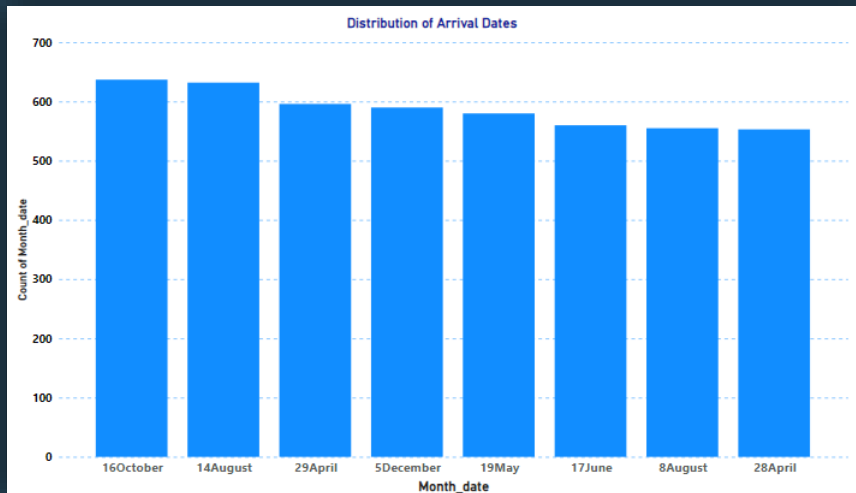
```
with cte as ( select NTILE(4) over(order by lead_time) as quartile, lead_time
from booking_details
),
cte2 as (
select case
when quartile = 1 then 'Q1' when quartile = 2 then 'Q2' when quartile = 3 then 'Q3' when quartile = 4 then 'Q4'
end as quartiles, lead_time
from cte)
SELECT 'Count' AS statistic, COUNT(lead_time) AS value
FROM booking_details
UNION ALL
SELECT 'Mean', round(AVG(lead_time),2)
FROM booking_details
UNION ALL
SELECT 'Standard Deviation', round(STDDEV(lead_time),2)
FROM booking_details
UNION ALL
SELECT 'Min', MIN(lead_time)
FROM booking_details
UNION ALL
SELECT 'Max', MAX(lead_time)
FROM booking_details
UNION ALL
select quartiles,min(lead_time)
from cte2
where quartiles != "Q1"
group by quartiles
```

Output

	statistic	value
►	Count	119390
	Mean	104.01
	Standard Deviation	106.86
	Min	0
	Max	737
	Q2	18
	Q3	69
	Q4	160

Observations:

From the distribution we can clearly say that October 16th has the most arrivals, however we see dates in August and July are also most common, now by applying filter to only show dates than has top 10% occurrence, we get



Hence we have 16th October, 14th Aug, 29th April, 5th December, 19th May, 8th August and 28th April as dates which has most occurred or most common.

Observations for statistical observation of Lead time.

- The average value as 104.01.
- Much of the Data fall below 4th Quartile, it has value closer to mean.
- The last Quartile has values ranging from 160 to Max Value 737, hence the density in last quadrant is low.

Conclusion

From all the Analysis and observations, we can infer that common dates tend to occur over a period of year and depending upon the observations we could expect more bookings on particular dates, although there are also chances of random events which might affect the arrivals.

2. Identify peak booking months and analyze reasons for peak in bookings, including holidays or events.

Overview:

Peak booking months are great to understanding for a Hotel that on which months they can expect more bookings, and also understanding the reason behind could also help Hotels manage their requirements.

Objectives:

- Analyze which months has higher bookings compared to rest.
- Find reason for the higher bookings.

Data preparation

In this problem statement we has to differentiate bookings by continents, however the dataset didn't consisted the continent information, however we prepared a table which consists country code along with respected continent referenced from [worldpopulationreview](#).

Continent	country
Africa	AGO
Africa	BDI
Africa	BEN
Africa	BFA
Africa	BWA

Query Approach

The objective of the query is to check out for 75th percentile, since we need a threshold number above which we will consider a month as peak bookings month.

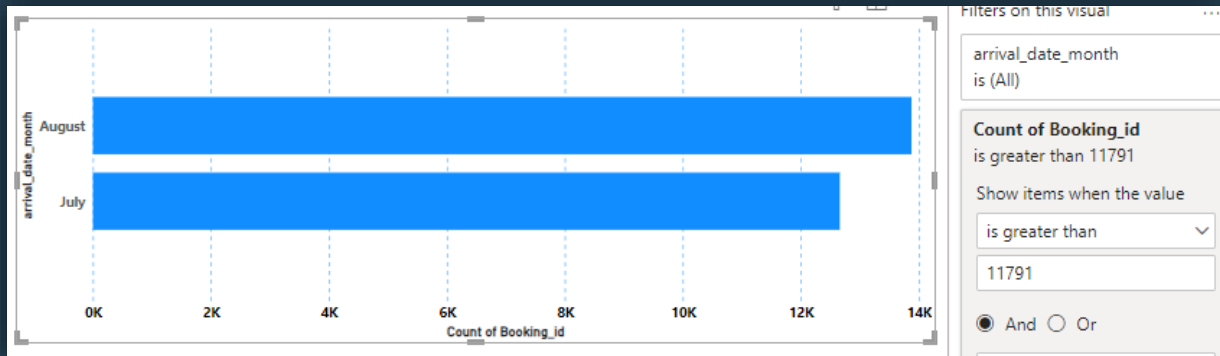
To achieve that, we aggregated count of booking made in each month and then used the ntile() function to divide the records into 4 parts, finally we applied filter to retrieve the 3rd quartile which is our 75th percentile value.

```
1 • with cte as (  
2   select arrival_date_month, count(Booking_id) as Bookings  
3   from booking_details  
4   group by arrival_date_month  
5   ),  
6   cte2 as (select ntile(4) over(order by Bookings) as percentile, Bookings, arrival_date_month  
7   from cte  
8   )  
9   select percentile, min(Bookings) as 3rd_Quartile  
10  from cte2  
11  where percentile > 3  
12  group by percentile
```


Outcome

	3rd_Quartile
▶	11791

Now, we have found our threshold value, now we could apply the filter and check for all the months, which has got booking above 11,791



Observations:

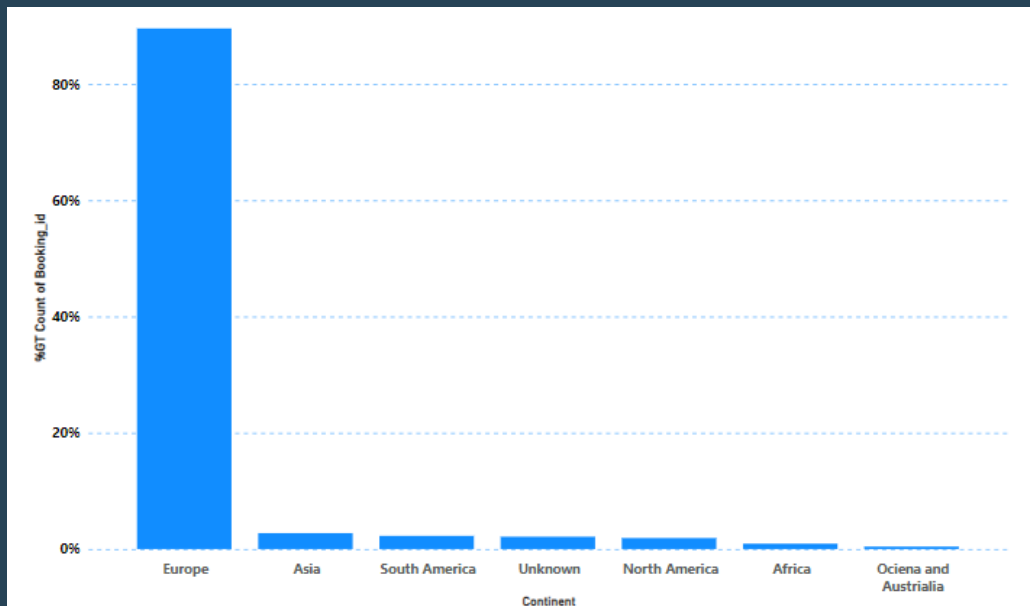
From the above chart we can confer that **July** and **August** are the months which have the peak bookings.

Analyzing the reason for higher bookings

In the hotel booking industry, occasions, events and festivals are the three biggest reasons why people might travel and spent time outside than usual.

We can search for events and occasions in July and August, however events and occasions depends specifically on the geographical and cultural aspects of a location in which the guest has made booking from, since our data consist of information from hotel from numerous countries.

We created a distribution of bookings from each continent



As more than 80% of all bookings are from Europe and most of the countries within Europe has similar geographic configuration, we could now look for events and occasions in particularly in Europe

The first thing we could check for was summer vacations, and from a detailed search from [Wikipedia Article](#) , we found out that the people in Europe usually the summer occasion between June to August. And is the primary reason for peak bookings in the respective months

Conclusion

From all the research and data, we could conclude that the peak time for hotel booking from this particular is **July** and **August**, the primary reason for that is **Summer Vacations**.

3. Understand the distribution of reserved and assigned room types. Calculate summary statistics for the consistency between reserved and assigned room types.

Overview:

Reserved types are rooms which are chosen by customer or which are reserved at the time of booking, in our data we have rooms denoted from alphabetical order starting from A to Z.

Assigned types are rooms which are assigned to the guest when the guests arrives at the hotel, assigned types can be different depending upon the availability or maintenance status, or guest requests.

The distribution of reserved and assigned type can help us understand amount of time the room type changes as well as any underlying patterns.

Objectives:

- The goal is creating a distribution for each room type and compare in the same chart.
- To create a binary summary statistic for frequency of times reserved and assigned types was same.

Data Preparation:

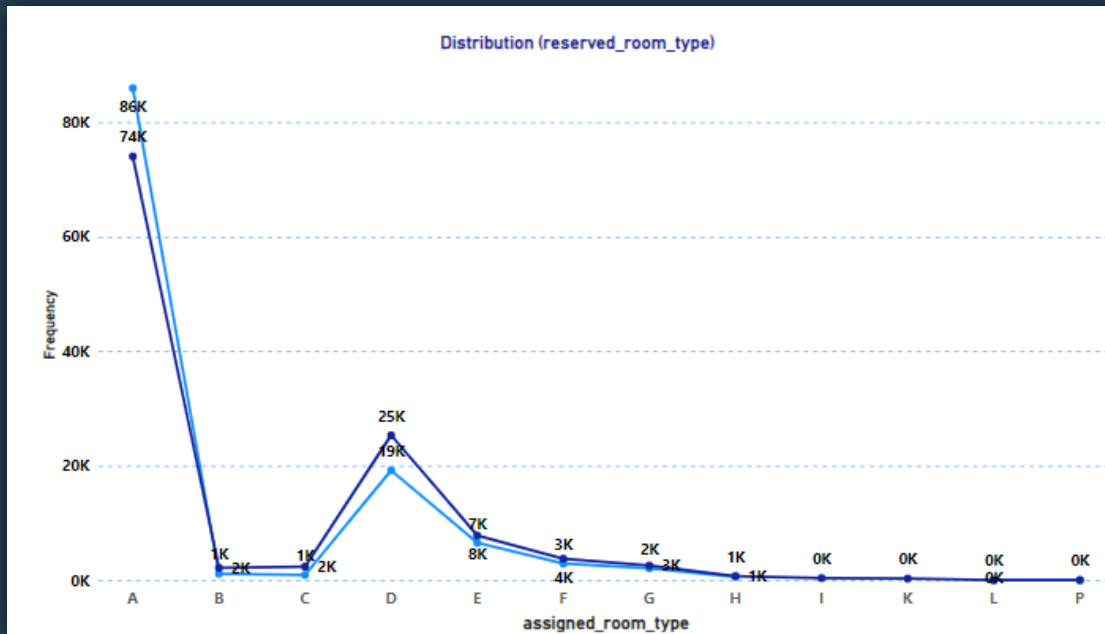
We need to create a separate table which contains the frequency of each room assigned for reserved and assigned types.

Using the Dax function **Summarize** we aggregated room_types and applied count function to get the frequency.

```
1 Distribution_of_room_types =  
2 SUMMARIZE(  
3     room_details,  
4     room_details[reserved_room_type],  
5     "Frequency", COUNT(room_details[reserved_room_type])  
6 )
```

reserved_room_type ▾	Frequency ▾	assigned_room_type ▾	Frequency ▾
C	932	C	2375
A	85994	A	74053
D	19201	D	25322
E	6535	E	7806
G	2094	G	2553
F	2897	F	3751
H	601	I	363
L	6	B	2163
P	12	H	712
B	1118	P	12
		L	1
		K	279

Now we could create a line chart in power using the above tables.



Observations:

we can observe that the distribution almost resembles the same, however major assigned guests are preferably changed to D type room.

Let us further evaluate difference by creating a summary statistic table.

Approach

Now we need to create a new column that will have the data for consistency of two types, if reserved and assigned are same we could update 1 else 0.

Once we create column, we get a binary information of consistency, but statistical data for binary data can be different let us define some of the measure we will be calculating.

Count – Count of 1's, this gives us information of how much from total bookings the reserved and assigned room types where same, or wise versa.

$$N = \sum x_i$$

Mean – This gives the average or proportion of time bookings had reserved and assigned where equivalent or not equivalent.

$$M = \sum x_i / n$$

Variance - It gives the degree of variation in a set of values. For binary data, it measures how much the data points (0s and 1s) differ from the mean (the proportion of 1s).

$$\sigma^2 = p \times (1-p)$$

Standard Deviation – Standard Deviation is basically the measure of spread of data and is the square root of variance, the standard deviation, like the variance, is highest when the data is equally split between 1s and 0s (i.e. $p = 0.5$). This is because there is maximum variability in the data. The standard deviation decreases as the data becomes more skewed towards either all 1s or all 0s.

Now using the above definitions and formulas we can retrieve the required statistical data by using with the help of MySQL

```
with cte as (select case
  when reserved_room_type = assigned_room_type
  then 1 else 0 end as Consistency
  from room_details
)
select 'Count' as Statistics, sum(Consistency) as Reserved_and_Assigned, sum(if(Consistency=0,1,0)) as Reserved_and_not_Assigned
from cte
union all
select 'Mean(Proportion)',
      concat(
        round(sum(Consistency)/count(Consistency),2)*100,'%'),
      concat(
        round(sum(if(Consistency=0,1,0))/count(Consistency),2)*100,'%')
from cte
union all
```

```
select 'Variance', round(
  sum(Consistency)/count(Consistency)*(1-sum(Consistency)/count(Consistency)),2),
  round(
    sum(if(Consistency=0,1,0))/count(Consistency)*(1-sum(if(Consistency=0,1,0))/count(Consistency)),2)
from cte
union all
select 'Standard_Deviation', round(sqrt(
  sum(Consistency)/count(Consistency)*(1-sum(Consistency)/count(Consistency))),2),
  round(sqrt(
    sum(if(Consistency=0,1,0))/count(Consistency)*(1-sum(if(Consistency=0,1,0))/count(Consistency))),2)
from cte
```

Hence we get the output as

Statistics	Reserved_and_Assigned	Reserved_and_not_Assigned
Count	104473	14917
Mean(Proportion)	88.00%	12.00%
Variance	0.11	0.11
Standard_Deviation	0.33	0.33

Conclusion

From all the above analyzation we can infer that 88% of the guests are assigned with room types that were reserved, and the deviation is more towards D room type, where guest or some internals have preference have majorly chosen type D room.

Statistical Analyzation shows deviation of data is 33% towards the reserved and assigned to be the same.

4. Calculate the average length of stays for different hotel types and explore variations by meal plans.

Overview:

Average length of stays and variation of it by meal plans, can provide us very unique insight, it can infer which meal plans customers prefer for long, medium and short stays, and other important insights.

Objective:

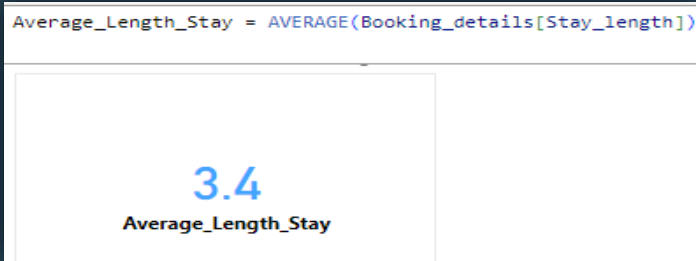
- Calculate the length of days a guest has stayed, for every booking.
- Plot a chart using power BI to detect variation of average length
- Find out whether specific meal plans have any significant or preference on length of stays.

Data Preparation

To get the length of stay, in our we have details of number of days spend in weekdays and number of days spend in weekends for each booking in Booking_details table, hence we created a custom column by adding both the columns.

Query Approach

Now once we have the column that has length of stays for each booking we could use **Average** function in Power BI to calculate the Average length of stay.

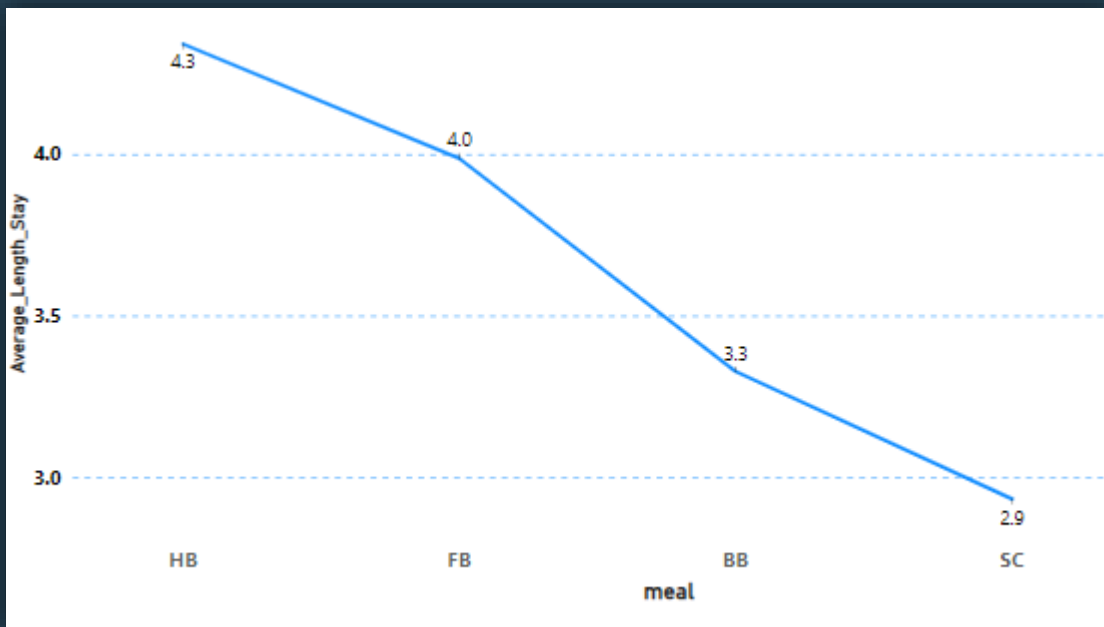


Once we have a measure of average stay length, we could get the average length of stay for each hotel type, which is our first finding of our problem statement.

hotel	Average_Length_Stay
Resort Hotel	4.3
City Hotel	3.0
Total	3.4

Observations

Let us explore the variation of average length of stay, across different meal plans



Utilizing the same measure, we get the variation of average length stay as per meal plans, in the chart above the codes in x-axis represents meal plans, which could conventionally mean,

BB: Bed and breakfast, **FB:** Full board, **HB:** Half Board, **SC:** Self catering.

Here are some insights,

- ❖ Calculated the average length of stay as 3.4 days, which should be a nominal number for any hotel.
- ❖ Analyzed that resort hotel type has a greater average length of stay of 4.3 days compared to city hotels.
 - Justifiable as guests in cities might prefer short-term stays compared to resorts.
- ❖ Meal plan variation in length of stay is also interesting:
 - Guests with Half and Full board tend to stay longer.
 - Guests who select Bed and Breakfast or self-caterings tend to have shorter stays.

Conclusion

In the end we can conclude the plan of meal can be one of the aspects to guesstimate that a guest might stay longer length or shorter length, also hotel resorts can expect that guest who usually arrive in resorts can have plans for longer days compared to city hotels.

5. Understand the distribution of the number of adults, children, and babies and identify any outliers.

Overview

The distribution of the number of adults, children and babies plays an important role to understand the majority of guest's type, for example, if we have adults equal it is most probable that the guests are a couple, similarly depending upon the numbers we could make Data-Driven decision to focus more particular type of guests.

Objective

- Our main goal is to create distributions for number of adults, children and babies in each booking.
- Analyze the underlying factors and detect outliers using the statistics.

Data Preparations

To create a distribution chart, we need to retrieve an aggregated table which counts each occurrence of a particular number.

Thereby we have to create 3 tables for occurrence of 1. Adults, 2. Children, 3. Babies.

Hence we could write a MySQL query as which, has aggregate function has **count** and **group by** the same columns

```
select adults, count(adults) as Frequency
from guest_info
group by adults
```

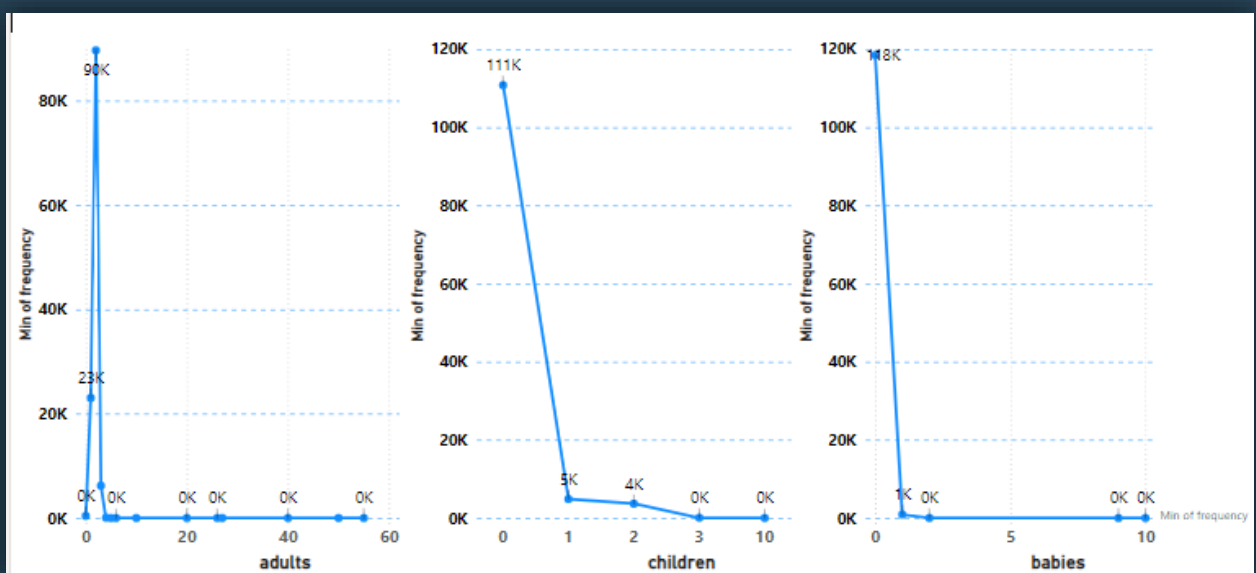
Once we have the data, we could then export it into CSV, and import it to Power BI, Visualize the data by using line chart with data labels turn on, and setting the aggregation of frequency column as minimum.

adults	frequency
2	89680
1	23027
3	6202
4	62
40	1
26	5
50	1
27	2
55	1
0	403
20	2
6	1
5	2
10	1

babies	frequency
0	118473
1	900
2	15
10	1
9	1

children	frequency
0	110796
1	4861
2	3652
10	1
3	76

Observations



Observation 1: In every distribution, there exist a single number of adults, children and babies whose occurrence is extensively higher than others. i.e Adults number occurred most is 2 with occurrence of 90k, Children number 0 = 111k, Babies number 0 = 118k.

Observation 2: The most common guest type is Adult = 2, Children = 0, Babies = 0, with total of 81557 bookings, which is almost 68% booking of all bookings.

Observation 3: The number of bookings made with Adult more than 5 is only 14.

Anomaly 1: There are 403 bookings made without any Adult. Consisting of children only.

Anomaly 2: There are 3 bookings made with zero Adults but with at least 1 babies.

adults	children	babies	Booking_id
0	2	1	81e5fc02
0	2	1	afc8dbb6
0	2	1	01c5ef38

Conclusion

With all our observations and analyzation we could conclude that most of bookings are probably made by couples or guests involving 2 adults, the observation also suggests that as the bookings where there are zero babies are extensively higher, it confers the fact that guests may not prefer include babies in their journeys or occasions.

We could also conclude that the number of time we a higher numbered group occupying the hotel is quite extensively low, and the hotel might have to work on attracting group events and team based guests.

6. Calculate summary statistics for ADR and explore differences between Resort Hotel and City Hotel bookings.

The analyzation for this particular problem statement will be divided into part 1 and part 2.

Part 1 - ADR

Overview

ADR – Average Daily Rate, which is nothing but Average spending of guests per day, and calculating summary statistics provides us important insights and interpretation, for example Density of guests in 4th Quartile, by understanding the number of guests in 4th Quartile helps take important data-driven decision making in pricing.

Approach

The similar MySQL query concept which is used in previous problem statement can be utilized to get required output.

	statistic	value
►	Count	119390
	Mean	101.83
	Standard Deviation	50.54
	Min	-6.38
	Max	5400
	Q2	69.29
	Q3	94.59
	Q4	126

Observations:

Observation1: We see that from large data availability of 119390 records, we can infer that dataset can provide reliable insights.

Observation2: The difference between **Max** value and the 4th quadrant is significantly high, which means that 75% of the ADR's lies below 4th quadrant that is 126.

Observation3: The mean value is 101.83 which gives us a reference number to guestimate the expected spending of a guests.

Part 2 - Resort Hotel and City Hotel bookings.

Overview:

Studying the guest's behavior can be vital in Hotel business in order to craft the plans and theme accordingly, hence studying the difference between Resorts and City Hotels, we could get important initial classification of behaviors based on Resort and City bookings.

Objectives:

As an Analyst, we would need to explore and present the differences in numbers, hence let us define our goals as per which we will analyzing the difference of attributes between Resorts and City Hotels.

1. Statistical.
2. Booking over a year.
3. And Cancellations.

Differences between Resort and City Hotel bookings

1. Statistical

Now let's compare how ADR's changes for Resort and city Hotels, ADR spending would provide us the key difference between the spending behavior of guests for Resorts and City.

Approach

We arrive at the statistical summary by creating a cte that contains the booking details and left joined with meal and stay details and applying filter **where** clause to consider only a particular hotel type

City_Hotel_statistics	value
Standard Deviation	43.6
Q4	126
Q3	99.9
Q2	79.2
Min	0
Mean	105.3
Max	5400
Count	79330

Resort_Hotel_statistics	value
Standard Deviation	61.44
Q4	125
Q3	75
Q2	50
Min	-6.38
Mean	94.95
Max	508
Count	40060

Observations

- **Observations1** – Bookings from City hotels are almost **twice** than Resort.
- **Observations2** – We observed that the average daily rates in City hotels tends to be **higher** compared to Resort, with Max spending reaching up to 900% more than max spending of resort.
- **Observation3** – We can infer that rates in Resorts can vary more than City rate.

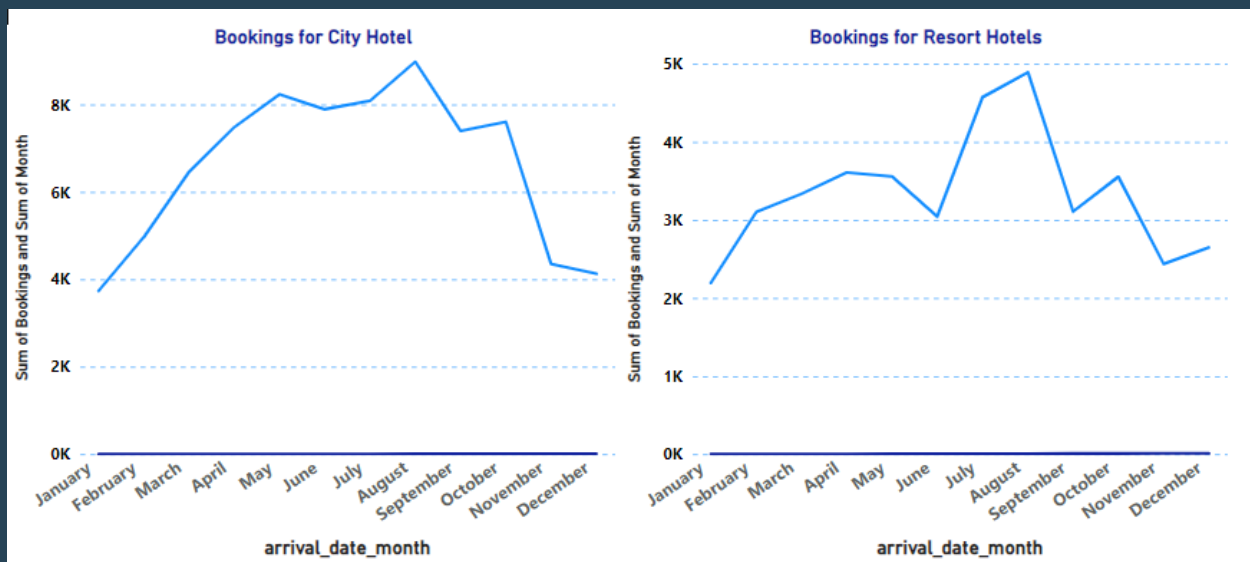
2. Bookings over a year

Analyzing the amount of booking pattern each month, for both the types can provide seasonal difference and guest behaviors.

Approach

To analyze the differences we will be using a visualization, that will not only make it easier to detect difference in bookings, but also the discrepancies

Now to plot the required chart we would just drop the arrival_month in x-axis and drop count of booking in y-axis and sort it as x-axis in ascending order,



Observations

Observation1: The plot interprets that of bookings in City hotel start rising steadily in first 3 months, then gradually increasing from **April** up until mid of **November**, whereas bookings in Resort type peaks up only in Summer seasons.

Observation2: In the month of **October** we could expect a higher bookings compared to previous months, in both type of Hotels

Cancellation

The amount of cancellation across two types of hotel can also be one of the important aspects to understanding the difference in booking behavior.

Approach

By applying this right filters in Power BI, we could get the count of bookings that were cancelled or not cancelled as per the hotel type,

Below is the details

Total cancelled (City Hotels)	Total cancelled (Resort Hotels)
33.10K	11.12K
Total not cancelled (City Hotels)	Total not cancelled (Resort Hotels)
46.23K	28.94K

Observations

The above data shows that cancellations for City hotels are **42%** of total bookings which is much higher than Resort hotels which is **27.7%**

Conclusion

Upon analyzing the some of the aspects of Resort and City hotels, like statistical, booking over time and cancellation

We saw that averagely spending in City hotels can go up to 5400 on daily bases, Resorts can experience immense boost in bookings during summer, where City bookings gets cancelled quite often but Resort bookings tends to be more stable.

In the end, we can conclude that Resorts and City hotels both experience different customer behavior and business patterns.

7. Compare the total number of special requests made by different customer types (e.g., Transient, Group) and identify which customer type makes more requests.

Overview:

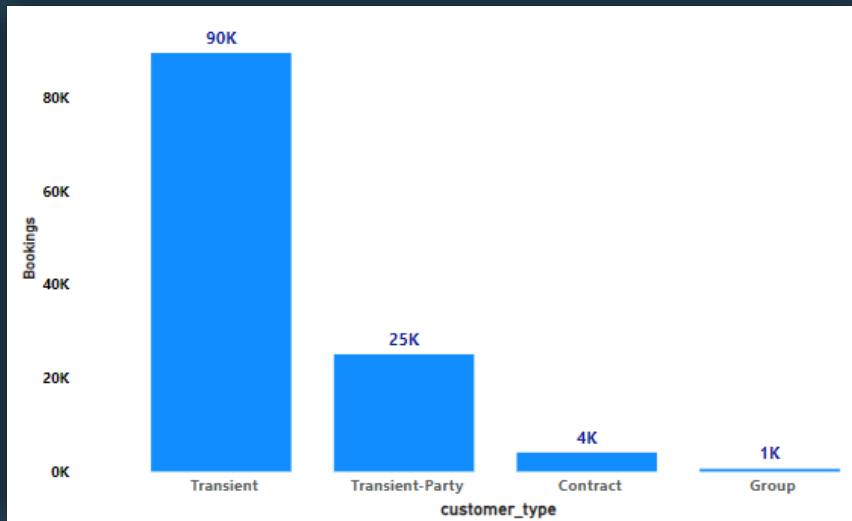
Special requests can be demanding and difficult to figuring out what type of customer might request more special request could allow Hotels to focus more on one type of guests and work to figure the key improvement aspects.

Objectives:

- Plot the bar graph, between Customer type vs total number of special requests.
- Utilize the same chart and determine the customer type with most requests.

Approach

We can use Power BI visualization and add Customer_type column in X-axis and Booking_id's in Y-axis with aggregation as count.



Conclusion

The number of bookings made by each customer type is calculated, from data we found that **Transient** customer have made 95% of all bookings.

8. Analyze Average Daily Rates (ADR) by meal plan type to identify variations in pricing.

Overview:

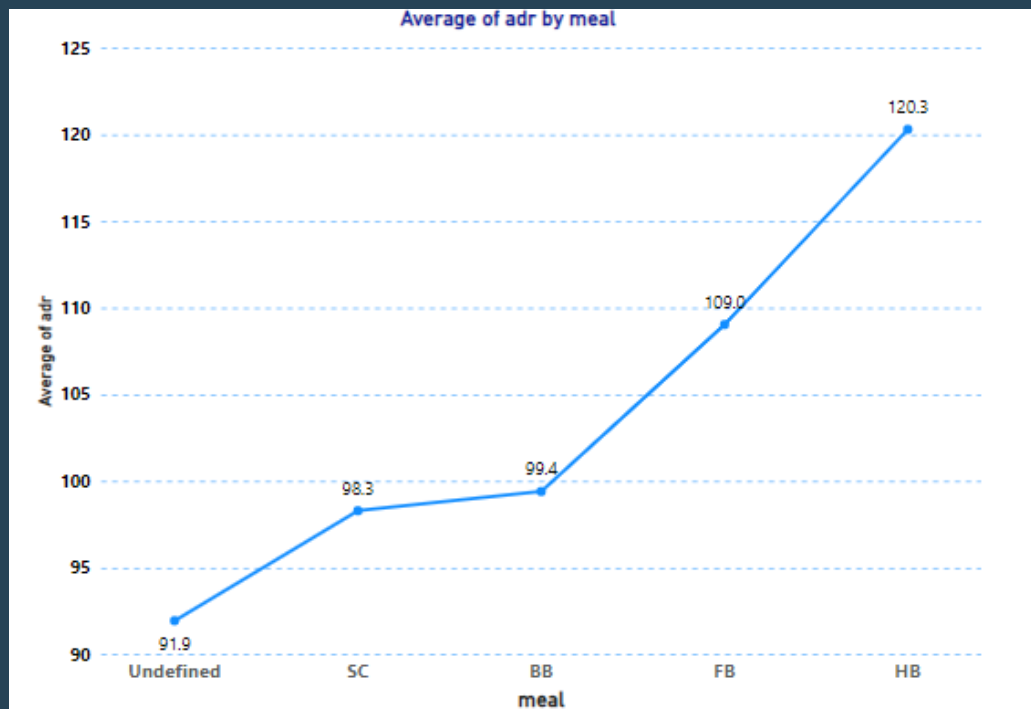
The hospitality industry often categorizes room rates by various factors, one of which is the type of meal plan included in the booking. The Average Daily Rate (ADR) is a crucial metric in the industry, representing the average rental income per paid occupied room per day.

Objectives:

- To plot a line chart between Meal plan type and Average daily rates.
- Analyze the underlying reason for any variations.

Approach

To Identify the ADR rates by meal plan, we can again go with power BI visuals, drop Meal plan column in X-Axis and ADR column in Y-Axis. Selecting chart type as line chart.



Observations

Observations1: While SC, BB, FB and HB can offers comprehensive plan for meals, but there are still customers who opt none.

Observation2: The **BB** plan (i.e Bed and Breakfast) have a nominal average spending of 99.4 dollars.

Conclusions

The average price for each meal plan does not deviates drastically, however it does have a range of 28.4 dollars. The following can be some of the observations.

- When the guest take care of Meals by **himself** i.e SC or unknown meal plan the ADR is **lower**.
- **HB** (Half Board) plan has the highest daily rates, as most of the HB plans does not include lunch and accommodations.
- However, the **BB** (Bed and Breakfast) plan tends to be the **budget friendly** or plan opted for quick stay, because the average daily rates is almost the same to self-catering or unknown plans.

9. Calculate the proportion of repeated guests and investigate their booking behavior. Identify any patterns or differences in preferences compared to first -time guests.

Overview

This problem statement focuses on calculating the proportion of repeated guests and analyzing their booking behavior in comparison to first-time guests. By identifying patterns and differences, hotels can mend their services and promotional efforts to better meet the needs of both groups, thereby achieving loyalty and optimizing revenue.

Objectives

- Write a MySQL query to retrieve the proportion of repeated and new guests.
- Investigate comprehensively by analyzing behavior across Market segment, Meals preferences, Special requests and number of cancellations.

Proportion of Repeated and new guests

With the help of concept of **CTE** we can get the table is grouped by `is_repeated` columns which contains the binary information about guest history. Then to calculate proportion of it, we could select count of bookings for repeated and non-repeated, and with the help of **subquery**, divide it by total bookings.

```
with cte as (  
  select is_repeated_guest,  
         count(Booking_id) as Bookings  
  from booking_source_and_history  
  group by is_repeated_guest)  
select *, concat(round(Bookings/(select count(Booking_id) from booking_source_and_history)*100,2), '%') as proportion  
from cte
```

Outcome

is_repeated_guest	Bookings	proportion
0	115580	96.81%
1	3810	3.19%

Observations

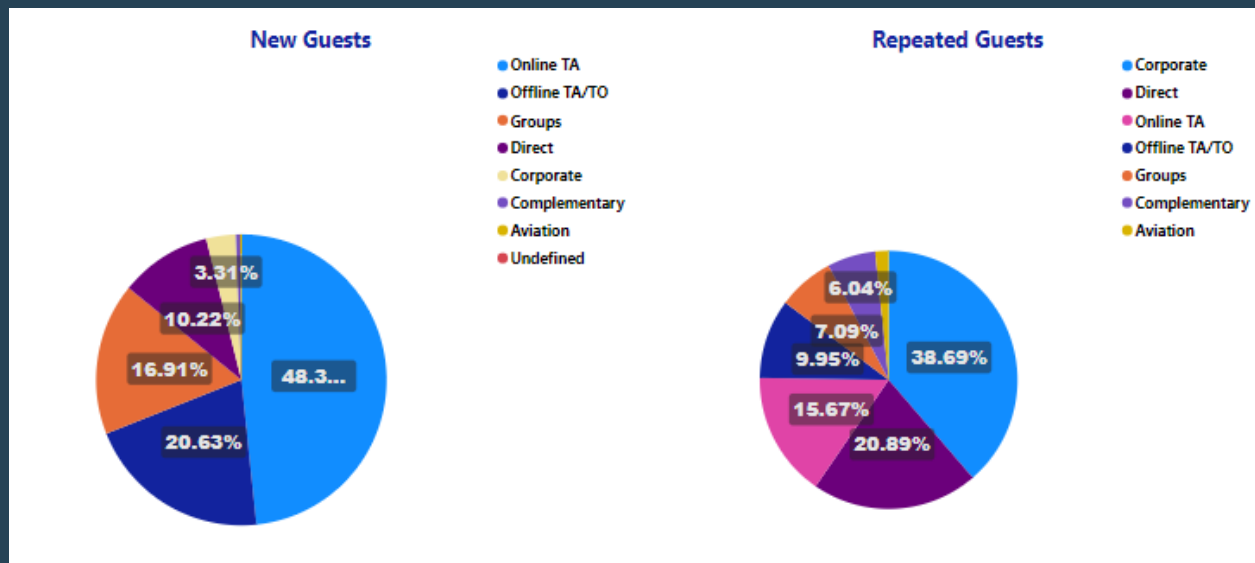
- 96.81% of the bookings are made by new guests.

To understand the context behind it, let us investigate the underlying factors.

Market Segment

Market segment could help us examine, which market segment provide us the most repeated and non-repeated guests.

Using a Pie-chart we can understand the source of guests, and its proportion.



Observations

- Online TA is the biggest reason for extensive bookings from new guests.
- On the other hand, guests from corporate segment have extensively has major repeated bookings

Meal preferences

Meal preferences of new guest and repeated guest can be helpful information in order tailor targeted promotion and Personalize communications.

Approach

A MySQL query to retrieve top 2 meal plan selected by repeated and non-repeated guests.

```
with cte as (  
  select MS.meal, count(BH.Booking_id) as Bookings_Repeated_Customer, rank() over(order by MS.meal) as temp_id  
  from booking_source_and_history BH  
  left join meal_and_stay_details MS on BH.Booking_id = MS.Booking_id and BH.is_repeated_guest = 1  
  group by MS.meal  
  having MS.meal is not null  
  order by Bookings_Repeated_Customer desc)
```

We are using cte and temp_id, so that we can also return top 2 meal plans from non-repeated guests simultaneously

The second part is exactly same as first except that we have changed the filter to retrieve non-repeated guests, below is continuation"

```

, cte2 as (
select MS.meal, count(BH.Booking_id) as Bookings_New_Customer, rank() over(order by MS.meal) as temp_id
from booking_source_and_history BH
left join meal_and_stay_details MS on BH.Booking_id = MS.Booking_id and BH.is_repeated_guest = 0
group by MS.meal
having MS.meal is not null
order by Bookings_New_Customer desc
)
select cte.meal, cte.Bookings_Repeated_Customer, cte2.meal, cte2.Bookings_New_Customer
from cte
inner join cte2 on cte.temp_id = cte2.temp_id
limit 2

```

Outcome

meal	Bookings_Repeated_Customer	meal	Bookings_New_Customer
BB	3473	BB	88837
HB	186	HB	14277

Observations

- The major preferred plans are BB and HB.
- Repeated and non-repeated customer has same preference over meal plan selections.

Percentage of special requests

Analyzing Special request as per repeated and new guests could help a Hotel to customized guest services, Improved Revenue Management and enhance operational efficiency.

Approach

The query is designed to retrieve the proportion of special requests It does not take into picture the amount of requests but only takes into picture if the guest has made atleast 1 request or not.

It first gets the data by joining all bookings from booking_details and meal_and_stay_details table, with an additional condition to only include the desired guest type.

Next the query uses certain formatting functions like concat and round to manipulate the outcome, and a ratio is taken using a conditional clause to get the count of guest who atleast made one requests vs the total bookings of desired guest type.

```

with cte as (
  select BS.*,MS.total_of_special_requests
  from booking_source_and_history BS
  inner join meal_and_stay_details MS on BS.Booking_id = MS.Booking_id and BS.is_repeated_guest = 1
)
select concat(
  round(
    sum(if(total_of_special_requests >0,1,0))/count(*)*100,2), '%' ) as percentage
from cte;

```

* For repeated guest

* For Non- repeated guests

percentage
41.08%

percentage
41.81%

Observations

- The amount of special requests made by both kind of customers are near to 41%.

Previous cancellations:

Studying previous cancellation can be very important aspect, we could examine the behavior by the guests and help hotel to Enhanced its Booking policies, improve guest retention and optimize revenue management.

Objectives:

- Calculate cancellation percentage for new guests and repeated guests.

Approach

In order to find the cancellation percentage, we would need to divide number of bookings which was cancelled by total number of booking multiplied with 100, the MySQL query execute the same calculations, since the cancellation information is stored in form of binary data we could use sum function along with conditional clause to get the number of cancellations, which is divided by total count of the table, which have been filtered by desired guest type.

```

select concat(
    round(
        sum(if(previous_cancellations>0,1,0))/count(*)*100,2), '%') as Previous_Cancellation_Percentage
from booking_source_and_history
where is_repeated_guest = 1

```

For existing guest

Previous_Cancellation_Percentage
24.33%

For New guests

Previous_Cancellation_Percentage
4.81%

Observations

- Guests who are repeated tends to make more cancellation than new guests.

Conclusion

Our analysis revealed several key insights about new and repeated guests:

Guest Composition: New guests constitute 96.81% of total bookings, primarily due to the hotel's strong presence on online platforms. The nature of the business, focused on occasions and tourism, attracts guests who often prefer exploring new locations.

Meal Plan Preferences: Both new and repeated guests show a preference for Bed and Breakfast (BB) and Half Board (HB) meal plans, indicating these are popular and well-received options.

Special Requests: Both groups have similar special requests, suggesting that the hotel meets common guest needs effectively regardless of their booking history.

Cancellation Rates: Surprisingly, repeated guests have a higher cancellation rate than new guests. This unexpected trend may require further investigation to understand underlying causes and address potential issues.

In summary, the high proportion of new guests is driven by online bookings and the hotel's appeal for special occasions and tourism. The preference for BB and HB meal plans and similar special requests across both groups indicate consistent guest expectations. However, the higher

cancellation rate among repeated guests highlights an area for potential improvement in guest retention strategies.

10. Analyze the impact of booking changes on cancellation rates. Calculate cancellation rates for bookings with different numbers of changes.

Overview

The flexibility to modify bookings is an important service that can influence guest satisfaction and retention. However, frequent booking changes may also correlate with higher cancellations, which can affect a hotel's revenue management and operational planning.

The problem statement aims to analyze the impact of booking changes on cancellation rates. By understanding this relationship, hotels can optimize their booking policies and improve overall operational efficiency.

Objectives

- Calculation of cancellations rate when there is booking changes vs when there are no booking changes.
- Retrieving a table which contains cancellation rates for each booking changes frequency.
- Deriving the Co-Relationship between booking changes and cancellation rate.
- Realizing impact of booking changes on cancellation rate.

Approach

Cancellations rates

To understand impact of cancellations, we need to calculate cancellation rates with and without booking changes.

In our dataset cancellation information is stored as 1 or 0, we could get the number of booking cancelled with the help of sum and divide it with total bookings, but before that we need to filter out the data as per the booking_changes per our required conditions.

Below is an example of MySQL to retrieve cancellation rate only for those bookings which has zero changes

```

with cte as (
  select BS.is_canceled, RD.booking_changes
  from booking_details BS
  left join room_details RD on BS.Booking_id = RD.Booking_id and booking_changes = 0
)

select concat(
  round(
    sum(is_canceled)/count(booking_changes)*100,2), '%') as cancellation_rate_Booking_changes
from cte

```

And for no booking changes we get the cancellation rate as

cancellation_rate_no_Booking_changes
43.65%

Cancellation rates when there is any changes

cancellation_rate_Booking_changes
24.47%

Cancellation rates for every booking changes frequency

This number will provide us an understanding of relationship cancellation and booking changes for instance we could analyze if there exist a threshold number of booking changes, after which the booking gets cancelled or any specific changing frequency that result in cancellation of bookings

```

with cte as (
  select BS.is_canceled, RD.booking_changes, BS.Booking_id
  from booking_details BS
  right join room_details RD on BS.Booking_id = RD.Booking_id
)

select booking_changes as booking_changes_made,
       round(sum(is_canceled)/count(is_canceled),2) as cancellation_rate
from cte
group by booking_changes
order by cancellation_rate desc

```

Outcome

booking_changes_made	cancellation_rate
16	0.50
0	0.41
6	0.29
8	0.24
2	0.20
14	0.20
4	0.18
5	0.17
10	0.17
3	0.16
1	0.14
9	0.13
7	0.10
17	0.00
13	0.00
12	0.00
20	0.00

Co-relation between Booking changes and cancellation rates

Now by exporting the data to as .CSV and with the help of **CORREL** function of calculated the co-relation

Booking_changes	Cancellation_rate
16	0.5
0	0.41
6	0.29
8	0.24
2	0.2
14	0.2
5	0.17
10	0.17
3	0.16
1	0.14
9	0.13
7	0.1
17	0
13	0
12	0
20	0
=correl(A2:A17,B2:B17)	

We get **Correation** as -0.332052664

Observations

- The cancellation rate for bookings with at least 1 change is lower than bookings with no booking changes.
- The data indicates a non-linear relationship between the number of booking changes and the cancellation rate. For instance, with 0 booking changes, the cancellation rate is 0.41, which is relatively high compared to other values.
- Higher booking changes does not affect the cancellation rates.

Conclusion

It is observed that bookings with at least one booking changes exhibit a lower cancellation rate compared to those with no changes. This suggests that allowing modifications to bookings may increase customer commitment and reduce the likelihood of cancellations.

Infect, the relationship between the number of booking changes and the cancellation rate is non-linear.