|  |  |
| --- | --- |
| A picture of a winding road and trees  HOTEL BOOKING ANALYSIS  Using Excel,MySQl and Power BI | introduction  Hotel is important part of a journey as well as the key part aspect of economy, with the help of powerful tools like MS Excel, MySQL and Power BI, we could detect major insights and predictions, let us take a journey to discover various aspects and KPIs of what drives the Hotel and their bookings.  Zubair Baig  Data analyst |

# 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 23 specific problem statements related to hotel bookings.

The analysis begins with the collection and loading of data, followed by meticulous data cleaning to ensure accuracy and reliability. Through exploratory data analysis (EDA), we examine the distribution and characteristics of key variables, utilizing descriptive statistics, visualizations, and measures of skewness and kurtosis to assess data normality.

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.**
7. **Results and insights.**
8. **Conclusions.**
9. **Future Research and analysis.**
10. **References.**

# INTRODUCTION

The hotel industry is an important and competitive sector that relies heavily on data driven decision making, which helps business to enhance customer satisfaction, pricing strategies, and to improve operational efficiency. Understanding booking patterns and customer preferences is crucial for hotels to remain competitive and meet the evolving demands of their industry.

The primary aim of this project is to conduct a comprehensive analysis of hotel booking to discover patterns, trends, and insights that improve strategic decision-making in the hospitality industry. By addressing specific problem statements, we could provide actionable recommendations to improve marketing strategies, enhance customer experience, reduce cancellations and effectively lower the operating costs.

This project will focus on analyzing a dataset sourced from a MySQL database, which includes information on hotel bookings, customer demographics, booking channels and cancellation rates. The analysis will be conducted using MySQL query together with Microsoft Excel and visualized using Power BI.

The project will address 20 specific problem statements which includes identifying peak booking periods to understanding cancellation behaviors. However, the analysis is limited to the data provided and does not account for external factors such as market trends or competitor actions.

# Objectives

The primary objective of this project is to analyze the data and find answers for our problem statements by integrating multiple data analysis tools and techniques. The project aims to:

1. **Integration**:

* Utilize and combine various analysis tools to achieve the most accurate results and gain practical experience with co-working and understanding strengths of each tool.

1. **Query bookings:**

* Employ SQL commands to effectively query and extract the required data from a MySQL database that can address a specific problem statements related to hotel bookings.

1. **Visualize bookings using Power BI:**

* Leverage Power BI to create insightful visualizations that facilitate the understanding of booking patterns, customer preferences, and other key metrics.

1. **Capture important information:**

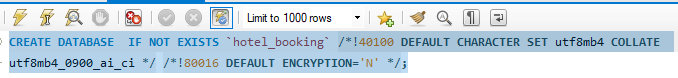
* Use Microsoft Excel, Microsoft Word and Microsoft Power Point to capture, document, and organize the entire analysis process, ensuring all steps and findings are systematically recorded.

1. **Conclusion**:

* Provide the actionable recommendations and comprehensive conclusions based on the insights derived from the data. And

# Data Loading

The Data for this project was provided in form of MySQL file, hence the was easily imported into the server just by creating a new database. The query for to create the database is provided below:

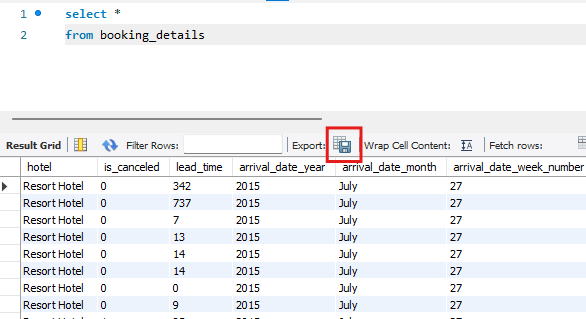


Once the data is imported to MySQL, we could now proceed to transform and process the data to multiple platforms.

And Upon observation the required no cleaning nor formatting or any manipulation required for analysis, hence we could 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, the windows will open file manager where we could save the file with a particular name and designated location. Moreover, we will also get option to convert .CSV files to. XLXS, in this project we have saved as CSV files, and below is the example query for one of the table.



SQL query to select all columns

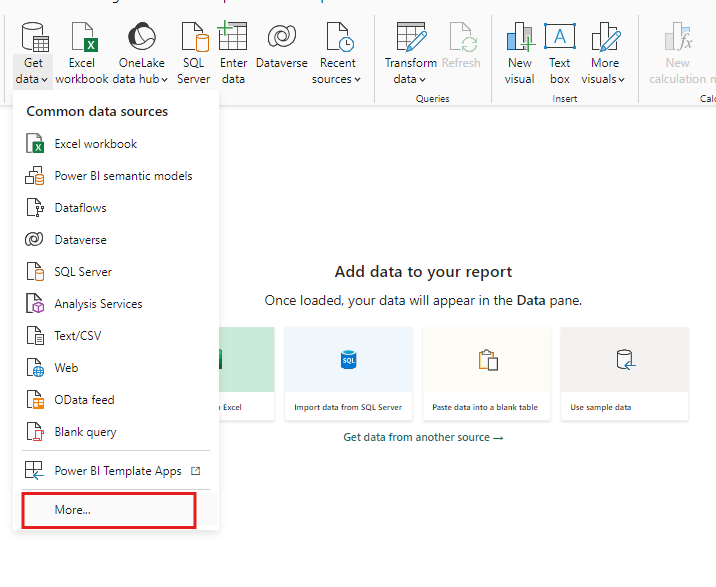
## Importing data to Power BI

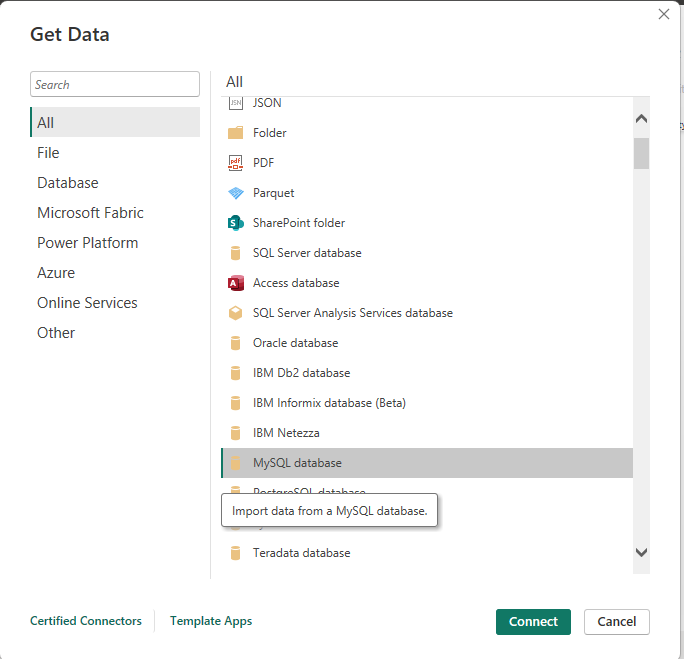
In this project we have utilized Microsoft Power BI desktop, a similar steps or approach are required for Power BI service or Premium platforms.

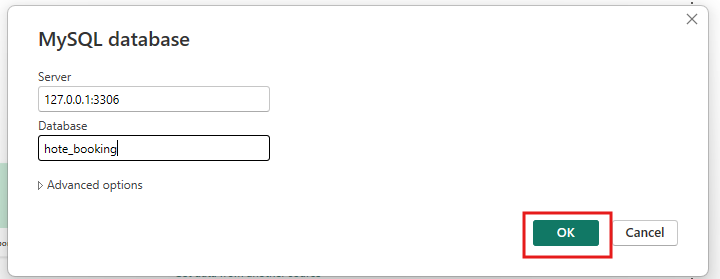
To import the data into power BI, we could either use CSV files we imported from MySQL, but that would require some manual work, in case if we have many tables which in our case is true, in order to import all the tables at once we used import from MySQL database option, which will directly load the database from MySQL to Power BI.

Below are the steps followed.

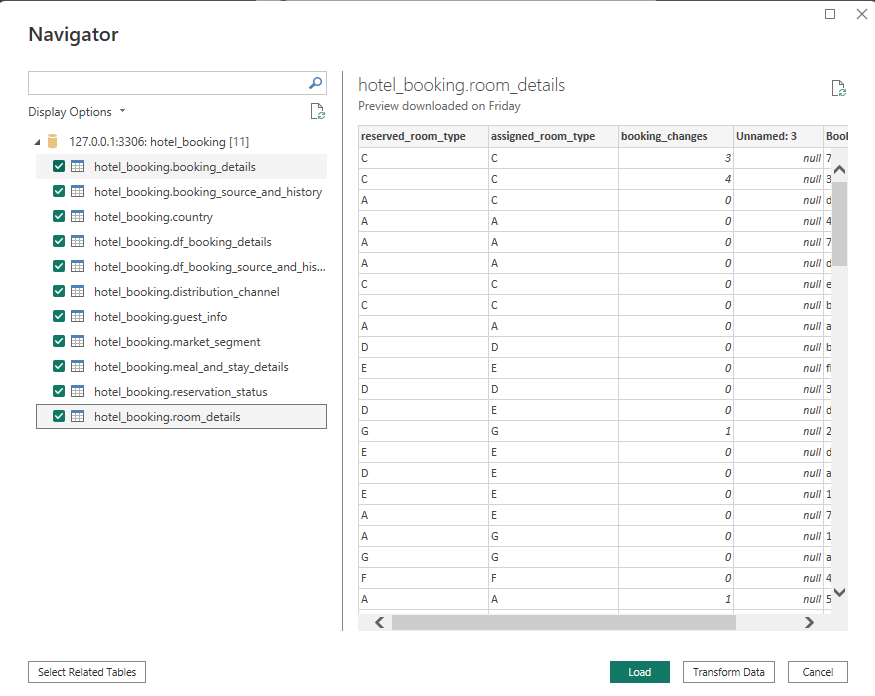
1. Select **Get data** icon, in top left corner and we will get the drop down of various platforms through which we import data, but in our case, we need to select **more.**



1. Scroll down to select **MYSQL database,** or we could even search in the search bar, and click **connect.**
2. Then we could enter our server and port number, and click **okay**.



1. We then get connected to our database and we could view all the files have loaded at once, we could even select the tables we and click **load** for power BI to import the tables.



# ER Diagram

## Overview

An ER(Entity-Relationship) diagram is a graphical tool used to model and represent the data structure of a database. It displays entities (Tables) and their relationships, including attributes, and how data elements interrelate within a system. The diagram helps in designing and understanding the logical schema of a database.

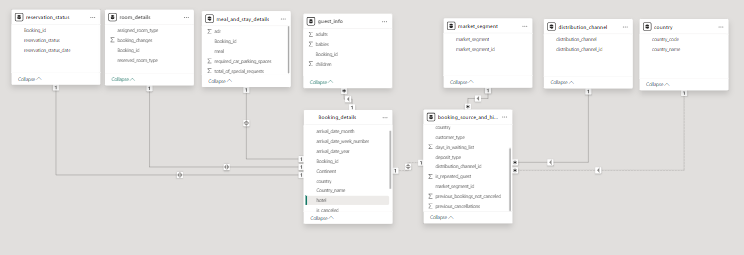
And in this project, we have conventionally utilized Fact, Query table concept to organize the ER-Diagram and to understand the relationship between the tables.

**Fact table:** A table that specifically designed to record all the events that occurred during an event. For instance, orders, whether records, sales.

**Query table:** A table which contains all the records of elements or categories which are involved or necessary to understand a transaction. Query table usually does not change it data, it remains a reference to fact table to get the required information related in the future, For instance, Location data, Product details, Customer profile.

A fact-query concept helps this project to understand the differentiation, connection, no. of relationship and the 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.



# Exploratory Data Analysis

Problem statement 1.

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

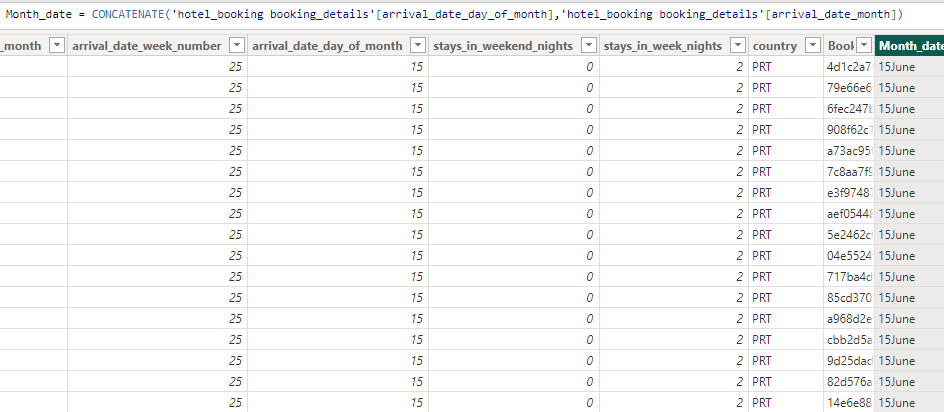
The distribution of the dates can be achieved in many ways; in this project we will be creating a distribution of

* **Bookings per Day in a year**: Distributing of days provide us information based on years, if we consider the complete date including year, we might not get complete information, but considering only the month and day, we get a combined distribution for a year and can infer meaningful insight about expected arrivals in year.

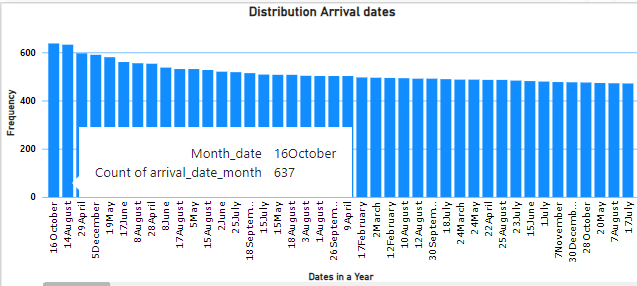
**Distribution of dates in a year**

In order to get the distribution of arrival dates considered in a year, we could add a new column, with the help of the DAX function mentioned below,

Month\_date = CONCATENATE('hotel\_booking booking\_details'[arrival\_date\_day\_of\_month],'hotel\_booking booking\_details'[arrival\_date\_month])



Adding custom column



Distribution of dates of a year

The above distribution illustrates ﻿ 16th October had the highest arrivals and was 615.73% higher than 11th December, which had the lowest (89).﻿﻿ Infact, 16th October accounted for 0.53% of arrivals.

**Statistical data for lead time.**

Statistical data help us give an overview of an numeric data, usually we could uitilize python function describe() to get the required result, but since this project is based on MySQL, we could achieve same results, using specific functions.

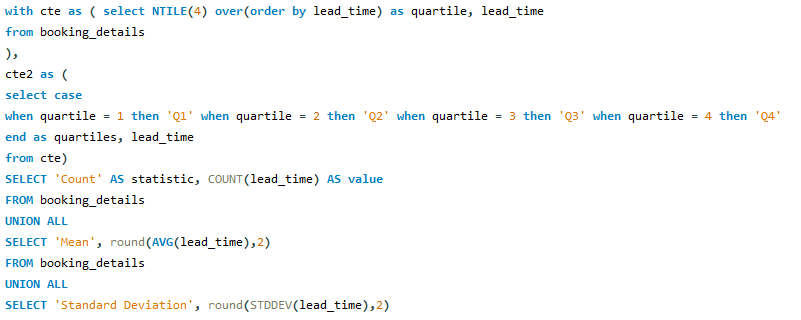
**Approach**

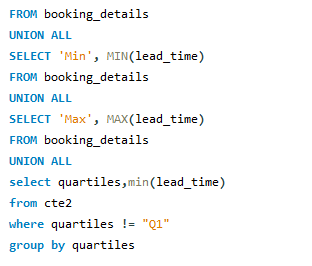
MySQl is not a programming language yet it is powerful, with the help of only inbuild functions, we could mimic the describe() function, in out approach we will be utilizing **union all** clause to align statistical aggregations in rows, each individual query will have aggregated repective functions to get the desired outcome.

However 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.

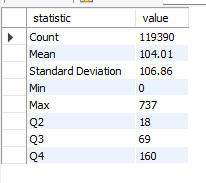
And for our quartiles we can divide the dataset into 4 parts and order it ascendingly, and if we retrieve the minimum of each division we can get the 1st, 2nd and 3rd quartiles repectively.

The query for the same is demonstrated below,



﻿

**Output**



The above outcome is the statistical information about the lead time data.

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

To analyze the peak booking months we can simply use power BI bar chart to aggregate the counts of booking per month, however to consider a month as peak booking month we might need a reference number.

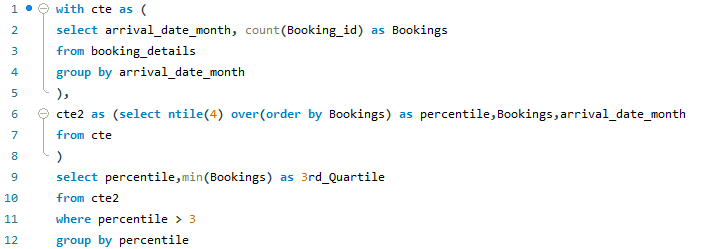
For our project we will be considering the months as peak months if the total bookings in a month is greater than 75% than other, in other words a peak month are all the months which are higher than 75th percentile of bookings.

But to calculate that we write a MySQL query that can retrieve us the 75th value,

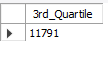
**Approach**

We can again use the same concept of Ntile() which was used in the previous problem statement, however this time we should also include count of booking in the **select** clause, also **where** clause to only retrieve 3rd quartile.

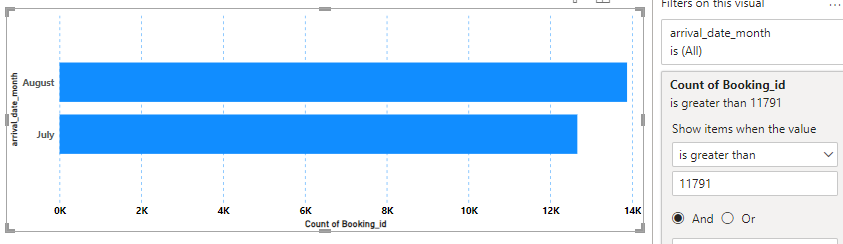
Below is the query,



**Outcome**

****

Now, we have found our reference number, now we could apply the filter and check for all the months, which has got booking above 11791



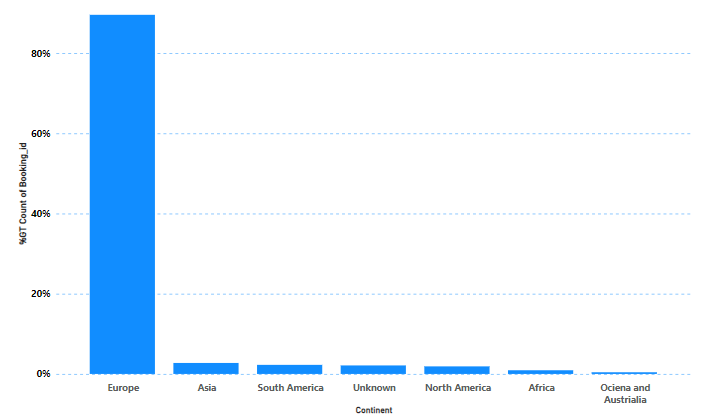
From the above chart it is we can confer that **July** and **August** are the months which received the higher bookings.

**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 it depends 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 various countries.

Upon observation we found that much of the bookings are made from Portugal and other European countries



As more than 80% of all bookings are from Europe and most of the countries within Europe has similar geographic configuration, we checked the events and occasions.

The first thing we could check for was summer vocations, and from a quick search we found out that the people in Europe get the summer occasion between June to August.

**Conclusion**

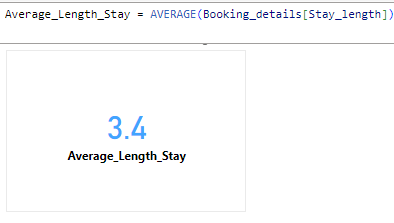
From all the research and data, we could conclude that the peak time for hotel booking from this particular is **July** and **August**, the primary reason for that is **Summer Vocations**.

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

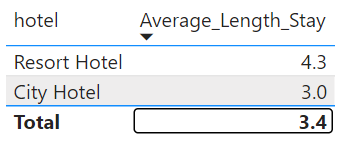
Average length of stays and variation of it by meal plans, can provide us very unique insight, it can infer the which meal plans customers prefer for long, medium and short stays, in order to arrive that we first need to calculate the length of stays for every booking which was checked out, and then find it’s average.

**Approach**

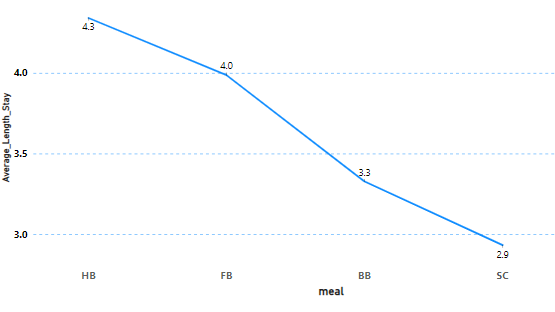
To get the length of stay, we added the columns from Booking\_details table, where we have the data of number of days spent in weekays and weekends. So can easily create a custom column and then we wrote a DAX function to calculate the average of the column,



Once we a measure of average stay length, we could get the average length of stay for each hotel type,

****

Now, 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, the dataset doesn’t provide the query table for meals plan names, however we get the information from data description,

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

**Conclusion**

We first saw that the overall length of stay as 3.4 days, which is a should be a decent number for any hotel, we also analyzed that resort hotel type has greator average length of stay of 4.3 compared to city hotel, which is justifiable as guests in cities might be preferable for short term stay compared to resorts.

The meal plan variation of length of stay is also interesting as guests with Half and Full board tend to stay longer comparted to people who only selects Bed and breakfast or self caterings.

This factors however may not primaliry affect the length of stay, as it be factor of many aspects, for example, hospitality, Economy of the country and many more, In the end we can conclude that by knowing plan of meal can be one of the aspects to infer that a guest might stay longer length, also hotel resorts can expect that guest who usually arrive in resorts can have plans for longer days camparted to city hotels.

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

A distribution is basically the amount of time the value or element have occurred, hence we could use the build in histogram chart in Excel to get the distribution, however due vast amount of data Excel failed to provide the correct distribution.

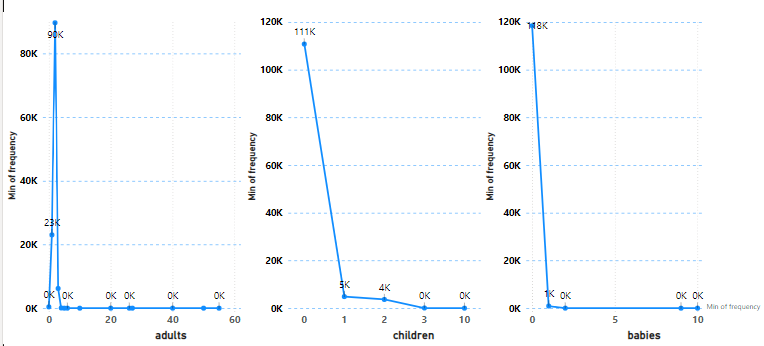
But here’s exactly MySQL and Power BI comes into picture, we used MySQL to get the frequency of occurance of number of adults,children and babies for each bookings and visualized it using power BI,

**Approach**

To write a SQL quert that returns the frequency of each occurance, we could simply group by clause to a particular column and use aggregate the count function on the same column to return the occurance of each value.

Once we have the data, we could then export it into CSV.

Visualize the data by using line chart with data labels turn on, and setting the aggregation of frequency column as minimum,



**Conclusion**

* Observations 1: In every distribution, there exist a single number of adults, children and babies who frequencies extensively higher than other numbers.

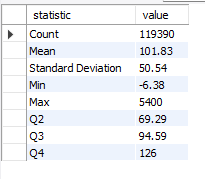
i.e Adults = 2, Children = 0, Babies = 0

This infers that hotels bookings are usually made by 2 adults, which could be couples or recently married.

* Observations 2: The data also affirms the fact that people would not prefer to take their babies to travel (As observed in problem set one).
* Anomaly 1: There 403 bookings made without any Adult. Consisting of children only
* Anomaly 2 :

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

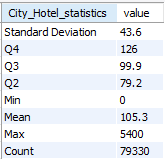
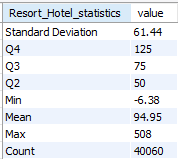
Let us first retrieve the summary statistics of Average daily rates (ADR),



**Differences between Resort and City Hotel bookings**

1. Statistical

Let us first start with statistical difference itself, to compare how ADR’s changes for Resort and city Hotels, we arrive at the below by creating a cte that contains the left join of booking details and meal and stay details and finally applying **where** clause to filter hotel type

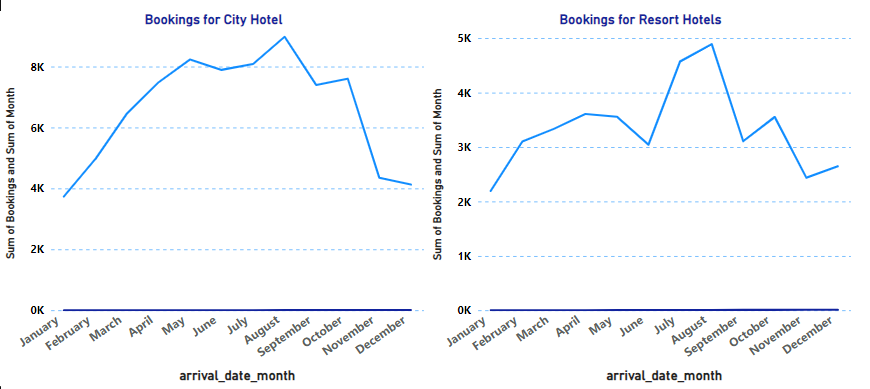
 

Some of the major differences are

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

1. **Bookings over a year**

Now we analyze the differences by observing the patterns, by visualizing count of booking per month.



The plot gives us the interpretation of SD of Resort being than City hotel, this can also implies that we can expect more an extensive booking increment in resort hotels during summer occasion compared to hotel which also tends to have higher orders in summer but not drastic increment.

That implies that the peak time for Resorts hotels is summer occasion compared to bookings in other months.

1. **Cancellation**

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



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**%

By this data we infer than we might expect more cancellations from Booking made in City Hotel type.

**Conclusion**

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

We saw that spending in City hotels can go upto 5400 averagely on daily base, whereas Resorts can experience immense boast in bookings during summer, where City bookings gets cancelled quite often but Resort bookings tends to be more serious.

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

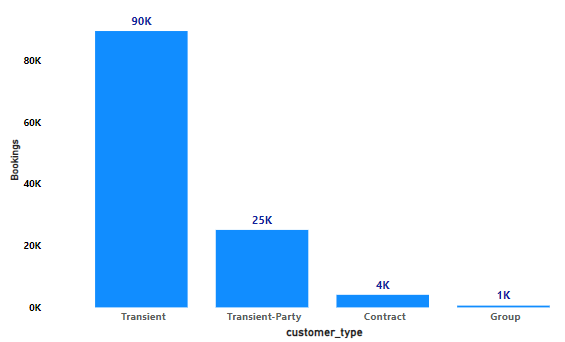
### Compare the total number of special requests made by different customer types (e.g., Transient, Group) and identify which customer type makes more requests.

To determine the number special requests made by different customer types, the best way is to make use of Power BI visualization,

**Approach**

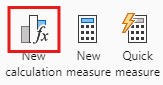
* Selecting chart type : We selected bar chart, which is most suitable for categorical visualization
* Select the fields and aggregation : We then select our fields, which in our case is Customer Type and Booking\_id, and change the aggregation to count for booking\_id, since the need the total nuber

We then get the visualization as,



Hence we get the required data and with number of bookings made for each customer type, however upon observation we could realise, that there is customer type which is **transient-party**, whose context is quite similar to transient, no specific information found about cutomer type transient-party in dataset details.

And in case we need to consider transient-party as transient itself, we could either change the values from transient-party to transient, or without changing any data, we could use the all new feature of Power BI, visual calculations

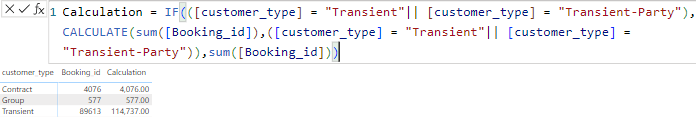


**Visual calculations**

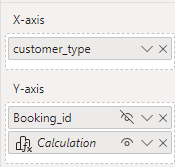
Visual calculation is the new feature in Power BI, that lets derive new columns or measures within your visual, this feature is explicitly helpful if we need any manipulation only for a particular Visual.

**Approach**

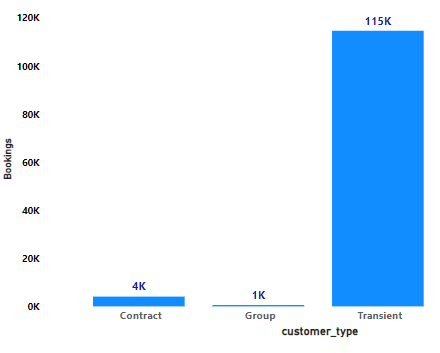
If we need to get the consider Transient party as Transient itself, then we can apply a filter to calculate the sum of Booking\_ids if customer\_type is transient or transient party,

****

Now Power BI allows us to hide fields with this new Visual calculation feature



We can hide our default aggregated values and display the calculated,

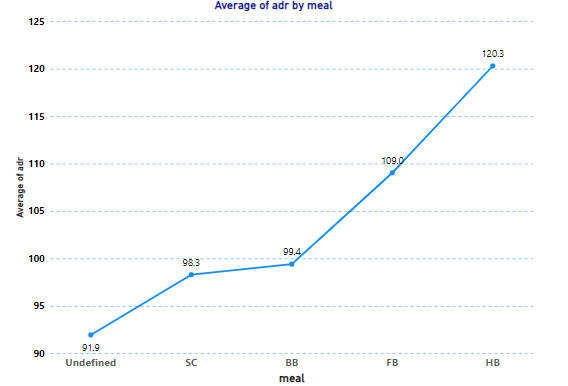


**Conclusion**

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

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

To Identify the ADR rakes by meal plan, we can again go with power BI visuals,



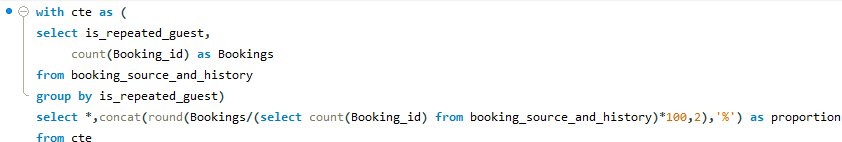
**Conclustion**

The average price for each meal plan does not vary drastically, however it does has 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**.
* However, the **BB** (Bed and Breakfast) plan tends to be the **best plan**, because the average daily rates is almost the same to self-catering or unknown plans.
* **HB** (Half Board) plan has the highest daily rates, as most of the HB plans does not include lunch and accommodations.

### Calculate the proportion of repeated guests and investigate their booking behavior. Identify any patterns or differences in preferences compared to firsttime guests.

To calculate the proportion of repeated guests we can write the MySQL query as

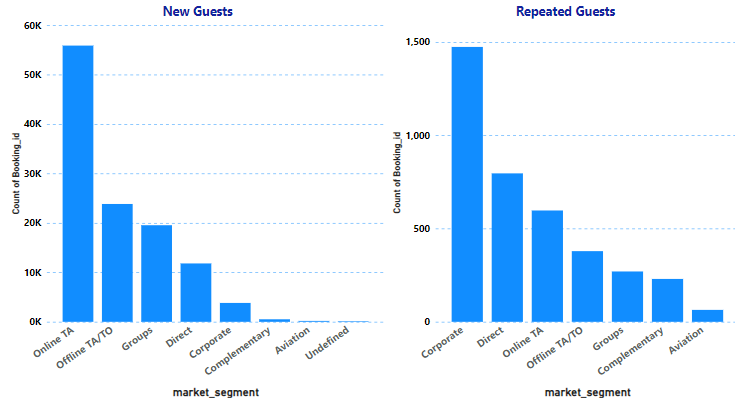


We get the output as



We can infer that most of the bookings are made by new guests, to understand the context behind it, we visualize the data using Power BI

The first thing we checked is the market segment, upon quick side by side comparison, we found that the reason of new guest being substantially high is because of nature of business and the online integration. Guest usually look for hotels online, and as the hotels from the data hotel are primarily booked for occasions and tourism. People usually prefer going to some other place.

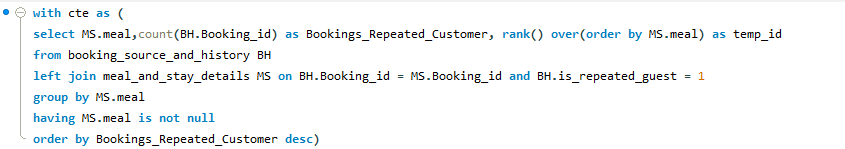


Other investigation is that **Corporate** customers are more likely to be a repeated guest.

Now let us differentiate some of the preferences between repeated and non-repeated guests.

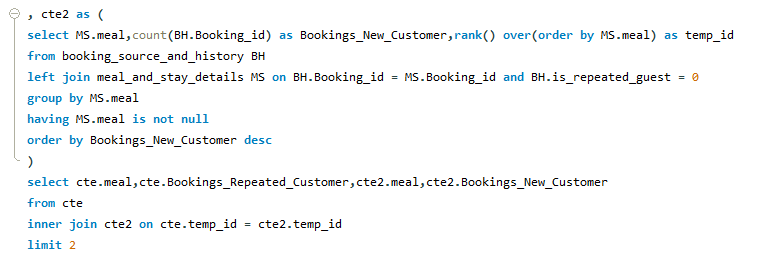
**Meal preferences**

Let us write a MySQL query to retrieve top 2 meal plan selected by repeated and non-repeated guests.

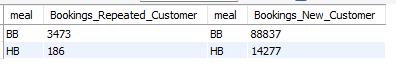


We are using cte and temp\_id, so that we can also return top 2 meal plans from non-repeated guests simultaneously.

The second part is exactly same as first except that it has a filter to retrieve non-repeated guests, below is continuation



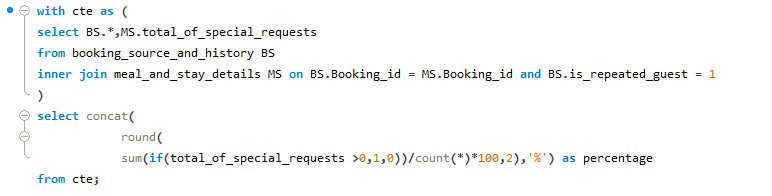
**Result**



We observed that both repeated and non-repeated customer has same preference over meal plan selections. Let us evaluate further different behavior

Percentage of special requests

MySQl query



For repeated guest

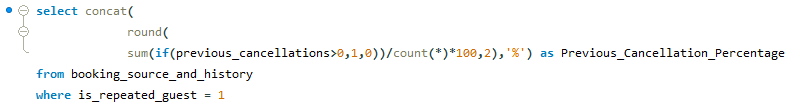


For Non- repeated guests



Previous cancellations

Studying previous cancellation can be very important aspect, the booking\_source\_and\_history table, provides us the sum of cancellations made by guest the current booking, now by differenciating the previous cancellation percentage we could examine the behavior, that whether a new guest or the existing who makes more cancellations.



For existing guest



For New guests



We see the guest who are repeated tends to make more cancellation than new guests.

**Conclusion**

We first calculated the amount of New guests and repeated guests and saw that new guests are **96.81** percent of total bookings, and the major reason for that is the being **online platforms**.

We then investigated certain preferences and behaviors, and observed that repeated and non-repeated both prefer **BB** and **HB** meal plans.

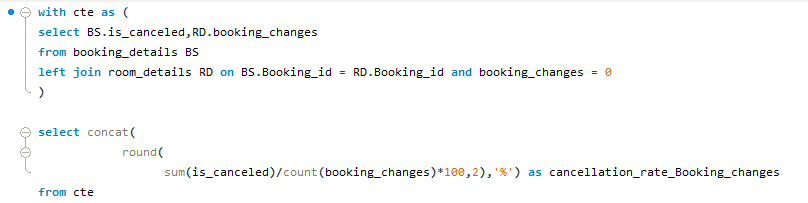
Then we proceeded to analyze special requests and saw that both groups have **similar** special requests.

Finally, the last examination was that the cancellation rate was higher for repeated guests and new guest, which generally would have being guests the other way.

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

To know the impact of booking changes on cancellation rates, we will have to calculate cancellation rates when there is no booking changes and cancellation rates when there is booking changes

We can achieve with the help of MySQL query



And for no booking changes we get the cancellation rate as



Now for cancellation rates when there any changes, we can just change the condition of join to greater than zero

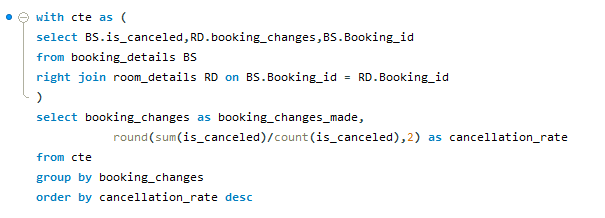
Cancellation rates when there is any changes



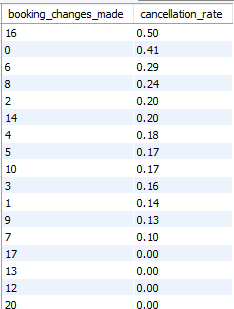
We can observe the impact here, and can infer that if there is any changes is bookings, it’s more likely that it will not be cancelled.

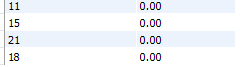
Calculation of cancellations rates for each booking changes number

We can write a MySQL query where in which we can aggregate by booking\_changes and calculated the total of cancelled /total bookings for each booking\_changes



Output





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

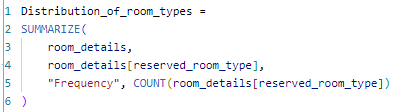
**Reserved types** are rooms which are chosen by customer or which are reserved at the time of booking, in our data we have rooms denoted from alphabetical order starting from A to Z.

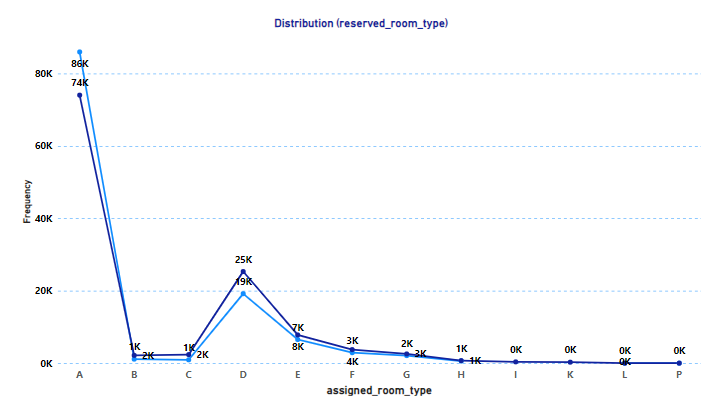
**Assigned types** are rooms which are assigned to the guest when the guests arrives at the hotel, assigned types can be different depending upon the availability or maintenance status, or guest requests.

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

Now let us create a table that 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.





Using line chart we could create a distribution, and we can observe that the distribution almost resembles the same, however major guests are assigned with D type room.

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

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, and some the important statistical measures for statistical data is,

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 =∑xi

​

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

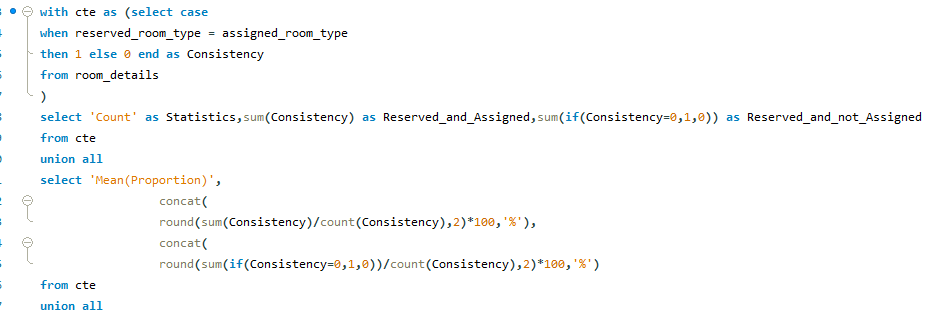
M=∑xi​​/n

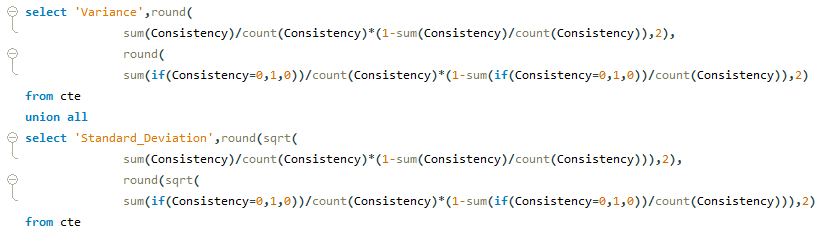
Variance - It gives the degree of variation in a set of values. For binary data, it measures how much the data points (0s and 1s) differ from the mean (the proportion of 1s).

σ 2 = p×(1−p)

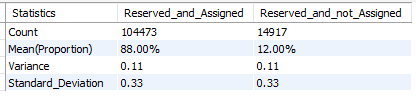
Standard Deviation – Standard Deviation is basically the measure of spread of data and is the square root of variance, the standard deviation, like the variance, is highest when the data is equally split between 1s and 0s (i.e. 𝑝 = 0.5). This is because there is maximum variability in the data. The standard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





Hence we get the output as



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 preferred majorly D type room.

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