HOTEL BOOKING DEMAND ANALYSIS

Contributors

W.A.C Fernando (26545)
M.A Pasindu Malinda (26033)
R.T Dulmina (27999)
H.K.I Dhananjaya (27595)
W.G.M.M.A Mansoor (26775)

Contents

Business Scenario and Requirements

There are several factors that a customer and hotels in general would like to understand, when it comes to booking a hotel. Such as,

From the customer's side, deciding on the optimal length of stay in order to get the best daily rate.

From the hotel's side,

- Wanting to have a better understanding about what types of customers usually visit hotels, as in by analyzing the preferences of customers, what they most like and dislike, the hotels can make the necessary adjustments in a way that would satisfy those preferences.

- Next, by learning about those preferences, if the hotels can identify certain patterns amongst the customers, then the hotels can meet those accordingly and increase their sales and profits

- Also, if they can understand that at certain times the hotel would most likely be fully booked, then they can hire more staff to meet that big demand.

- By understanding what most of the guests like, the hotels can make personalized offers for various target groups. By doing so the hotels would be able to maintain customer loyalty and to also increase customer satisfaction.

Therefore, the dataset was chosen from Kaggle which was originally from the Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019. This dataset contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of the stay, number of adults, children/babies etc. (Nuno Antonio, 2019)

Link to the original dataset - hotel_bookings.xlsx

Data Warehouse Architecture

We created a star schema data warehouse here that includes one fact table (fact_bookings) and several other dimension tables (dim_customer, dim_hotel, dim_date) and this supports OLAP operations to analyze hotel booking trends, cancellations and customer behaviors.

ETL process

Firstly, the raw data was 1st imported into BigQuery as rawdata. Afterwards it was cleaned and transformed into structured tables using SQL. The ETL process did indeed include cleaning the data, normalizing the schema and creating the dimension and fact tables.

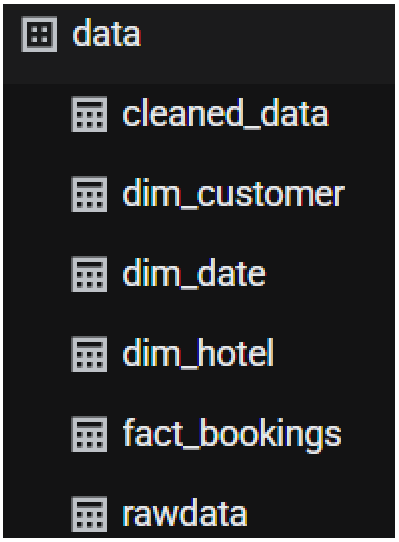
Issues in ETL process

- Deciding on the schema
- There were a lot of unnecessary columns in the raw data as well.

Link to the cleaned dataset - hotel_bookings_cleaned (1).xlsx

OLAP Cube Implementation in BigQuery

For this project an OLAP cube was successfully implemented on the hotel booking dataset and is hosted in Google BigQuery. The dataset consists of fact tables and dimension tables such as fact_bookings, dim_customer, dim_date, dim_hotel etc. acting as the basis of the OLAP cube as shown below.



data

- cleaned_data
- dim_customer
- dim_date
- dim_hotel
- fact_bookings
- rawdata

OLAP Architecture

Here the ROLAP (Relational OLAP) was incorporated, which means that the data is stored in relational tables and is queried by using SQL.

Reasons for choosing this architecture

- This seemed appropriate because BigQuery also operates on a distributed and scalable SQL engine and below are also some other reasons as to why we chose this.

- Compared to MOLAP (Multidimensional OLAP), since MOLAP requires pre-computed multidimensional cubes, ROLAP is way more flexible and is better suited for handling such a large dataset like this and since the fact tables and dimensional tables are all relational ROLAP seemed the better fit.

- Compared to HOLAP, even though this method uses both MOLAP and ROLAP, this introduces an additional level of complexity as we have to manage 2 storage formats here. Also, for this project, since it primarily requires only a few OLAP operations for the relational data, ROLAP seemed the better choice.

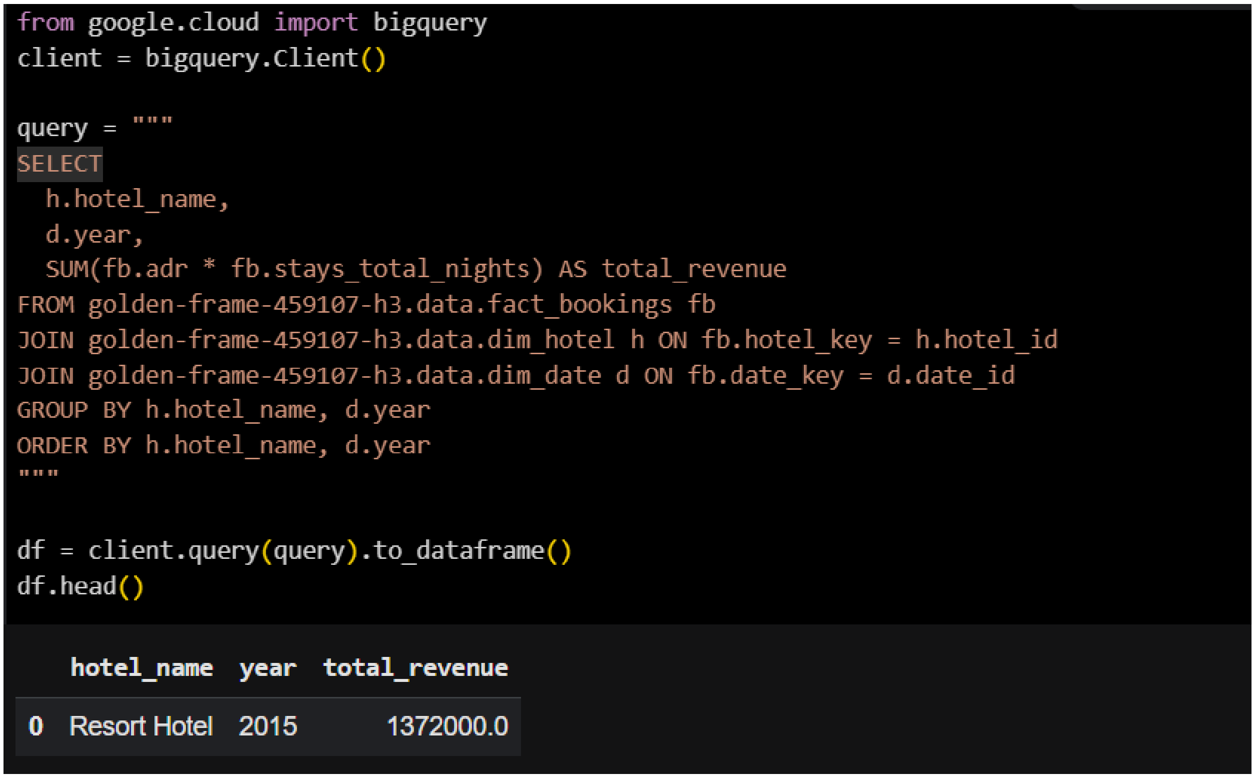
- Compared to DOLAP, it did not seem suitable for this project as its main focus involves basically downloading summarized data to a user's desktop for offline analysis. It is not suitable for large scale analysis such as this.

Construction of the Cube

The following dimensions were used to make the base of the cube.

- Hotel Name (dim_hotel)
- Year (dim_date)
- Market Segment (dim_customer)

For this, the primary measure was used as Total Revenue = adr*stays_total_nights (from fact_bookings), (adr=average daily rate).



```
from google.cloud import bigquery
client = bigquery.Client()

query = """
SELECT
  h.hotel_name,
  d.year,
  SUM(fb.adr * fb.stays_total_nights) AS total_revenue
FROM golden-frame-459107-h3.data.fact_bookings fb
JOIN golden-frame-459107-h3.data.dim_hotel h ON fb.hotel_key = h.hotel_id
JOIN golden-frame-459107-h3.data.dim_date d ON fb.date_key = d.date_id
GROUP BY h.hotel_name, d.year
ORDER BY h.hotel_name, d.year
"""

df = client.query(query).to_dataframe()
df.head()
```

	hotel_name	year	total_revenue
0	Resort Hotel	2015	1372000.0

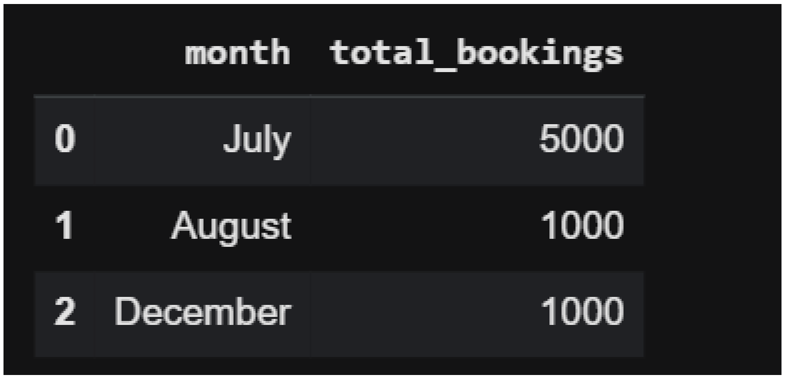
This is basically the base cuboid of the OLAP cube.

By using this cube, it allowed for the analysis of revenue generated per hotel per year and this was used as the foundation for the below-mentioned OLAP operations.

OLAP operations

Roll-Up

This was done to summarize total bookings from the day level to the month level in order to analyze any seasonal trends. Output is shown below.



	month	total_bookings
0	July	5000
1	August	1000
2	December	1000

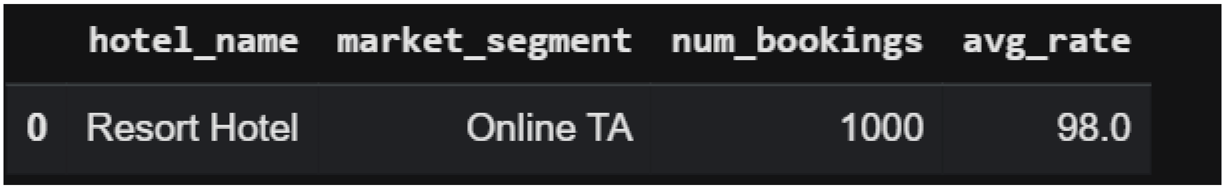
So here July clearly has the highest number of bookings for Resort Hotel, then August and December come after. Which means that July is indeed the peak season in terms of bookings for Hotel Resort.

This means the hotel should do certain things such as,

- Increasing their pricing, as in their room/suite rates a bit in July in order to make maximum level of profits and revenue.
- Schedule more staff and services as well in order to handle the big crowd that will be coming in July.
- Come up with new promotions and offers to focus on the off-peak months rather than focusing them on July since July already has a big, expected crowd.

Drill-Down

In order to gain better insights into customer behavior and overall pricing, this operation was done and by analyzing the bookings and breaking them down by hotel name and customer market segment. Output is shown below.



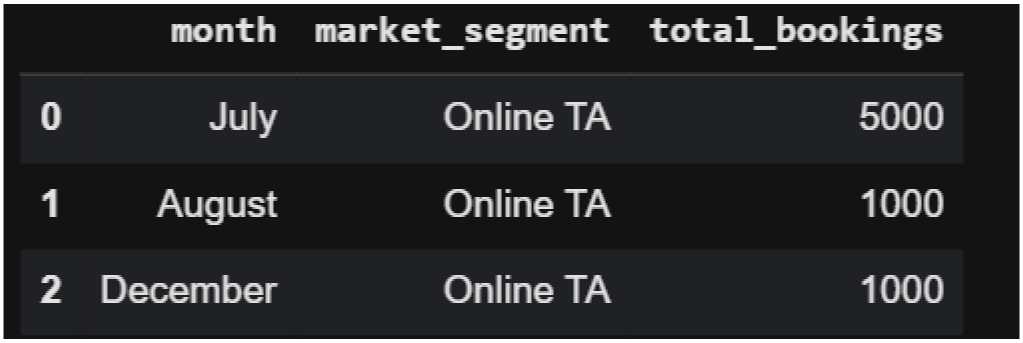
	hotel_name	market_segment	num_bookings	avg_rate
0	Resort Hotel	Online TA	1000	98.0

Here, it is clear which market segment is bringing in the most bookings and their average rate as well. In this case, its Online Travel Agencies (Online TA), they are a major segment for Resort Hotel.

Which means Hotel Resort should strengthen their partnerships with Online TAs and to also tailor more advertisements and promotions etc. towards Online TA users.

Slice

Isolated bookings regarding Resort Hotel to analyze and gain insights into monthly and segment-specific patterns. Output is shown below.



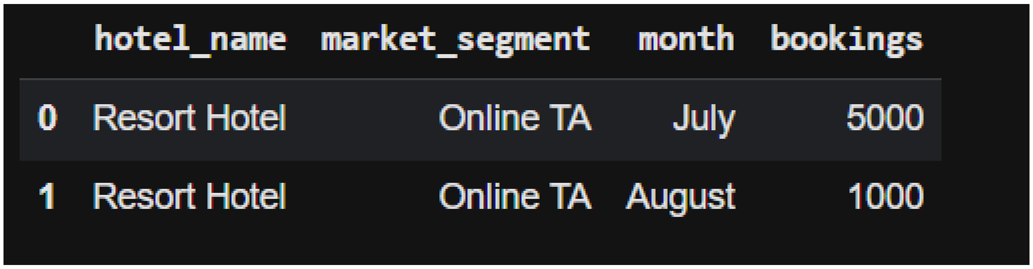
	month	market_segment	total_bookings
0	July	Online TA	5000
1	August	Online TA	1000
2	December	Online TA	1000

According to the above output, the majority of the July bookings come from Online TA, thus showcasing their dominance in peak periods.

Which means, the hotel should again offer promotions and packages targeting the OTA users, and also, they should understand the Online TA user behaviors better and make personalized offers and deals targeted for them.

Dice

Here to answer certain potential targeted questions, the data was filtered for ‘City Hotel’ limited to the ‘Corporate’ and ‘Group’ market segments in the month of August.



	hotel_name	market_segment	month	bookings
0	Resort Hotel	Online TA	July	5000
1	Resort Hotel	Online TA	August	1000

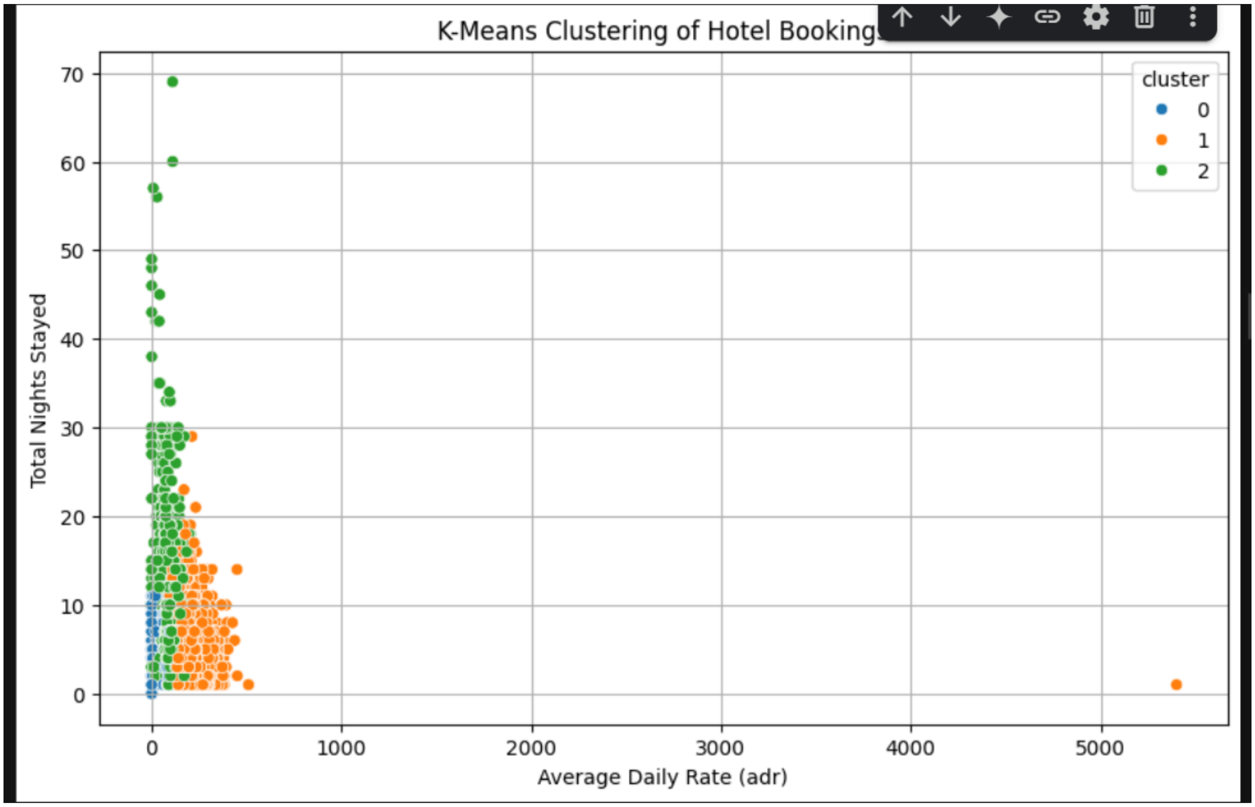
According to the above output it is clear that, in July the Resort hotel has received 5000 bookings from the online travel agents segment and in august the bookings has dropped to 1000 from that same segment. Which means that there is basically a major 80% drop in bookings over 1 month in the same market segment, which indicates possibly a seasonal peak in July, probably due to school vacations or holidays etc. and August may mark the end of that peak.

Therefore, based on this analysis hotels can make the necessary preparations to meet the demand in July by hiring more staff and having the relevant resources. They can also strategize to make promotions targeted for the offseason as well to maximize returns.

They can also change their prices accordingly. For instance, increase their prices in July to make the maximum profits etc.

Data mining

Kmeans implementation



According to the above output 3 clusters can be seen.

Cluster 0 (Blue)

Represents the guests who stayed for short periods due to being protective about their budgets. (low price adr)

Cluster (1) – orange color

Represents the people who were willing to stay from short to medium time periods. There is one outlier as well at an adr pass 5000 which could be an error of some kind. But it could also be interpreted as some kind of luxury booking as well.

Cluster (2) – green color

Represents the people who stayed for longer periods (up to 70 days) who were looking for affordable rates as well. And shows almost low to moderate level adr.

Based on these insights, hotels could create certain targeted packages/deals for the relevant clusters. As in,

Cluster 0 (since it represents the people who stayed for short time periods with low budget)

- Flash discounts could be offered to them or even minimal service rooms
- Also, there is a high chance that there are solo travelers and tourists who travel on weekends as well and by marketing the hotel to those kinds of people could attract more such people and help the hotels to increase more revenue.

Cluster 2 (since it represents people who stayed for longer periods with a low to moderate budget)

- There is a chance that some of those people could be certain business travelers, so marketing the hotels to such people then could attract more people.
- Also, by offering various monthly or even weekly rates, and by providing extended stay perks (ex- free use of gym or room service etc.) could also satisfy a lot of customers and could also attract more people.

Cluster 1 (since it represents the moderate to high spenders)

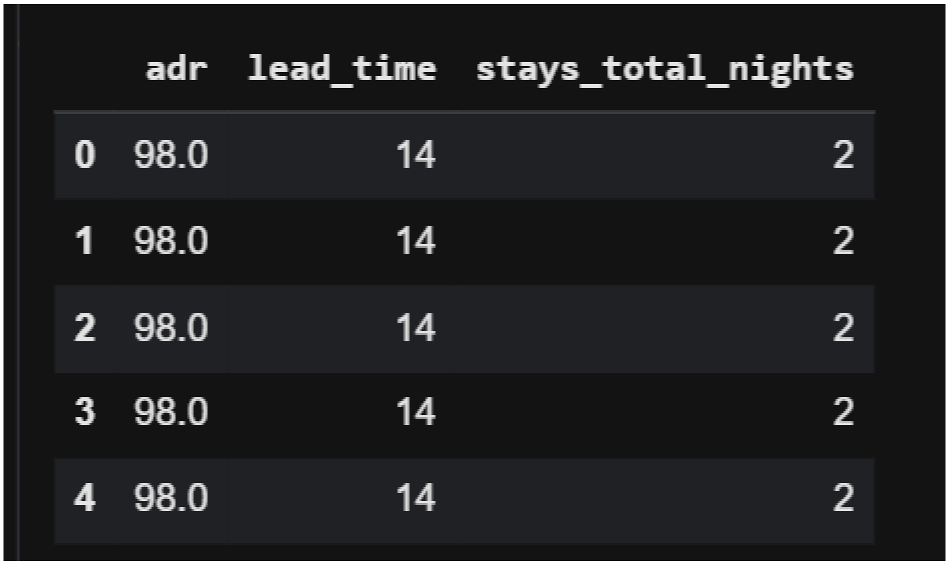
- Hotels should focus more overall comfort and dining and other premium level experiences to attract even more premium level customers.

By observing this cluster analysis, hotels can customize their marketing strategies according to each cluster and try to attract more people from each cluster in the future as well.

So, to summarize this kmeans clustering helped to identify the various behavior groups of people based on price and stay length. So, as mentioned earlier also, hotels can now use this analysis to make better data driven decisions in the future.

Applying DBSCAN

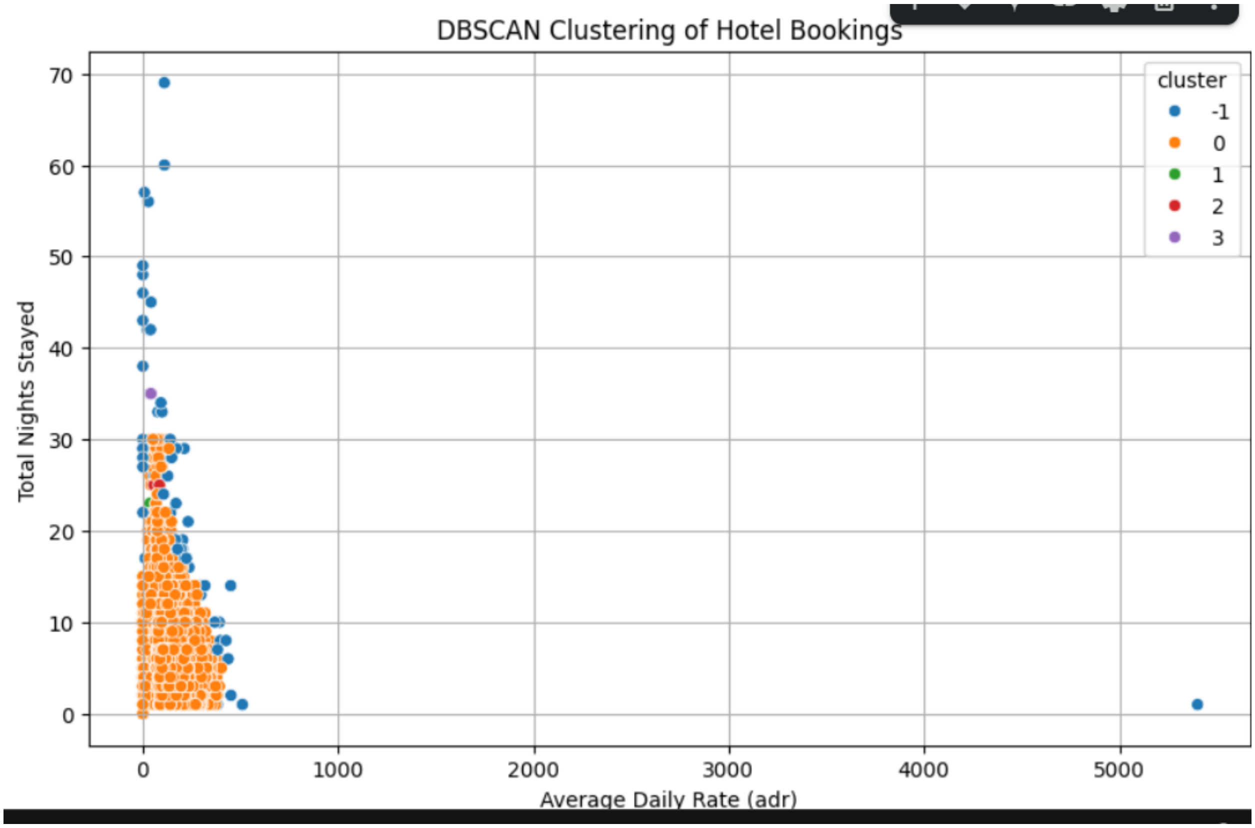
Before applying the DBSCAN, certain features had to be selected from the bookings dataset to better identify meaningful customer patterns. The output below shows the first few records that were retrieved.



	adr	lead_time	stays_total_nights
0	98.0	14	2
1	98.0	14	2
2	98.0	14	2
3	98.0	14	2
4	98.0	14	2

This basically shows a frequent booking pattern of sorts that involve 2-night stays at a fixed rate, which are booked about 2 weeks in advance. Then these patterns were used in DBSCAN to identify dense regions of similar behaviors. For instance, groups who are conscious about their budgets or people who stayed for longer periods etc.

DBSCAN

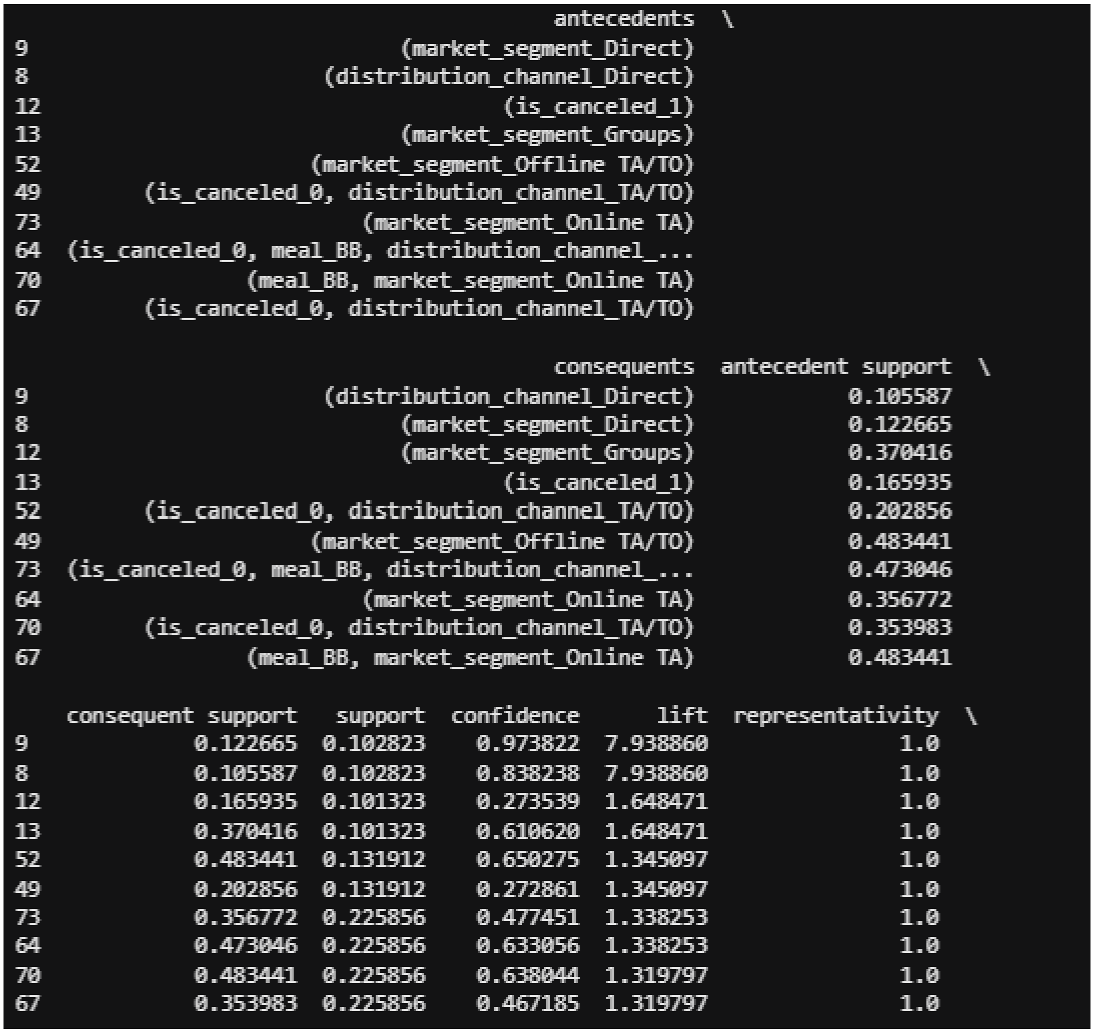


- In the above DBSCAN, each color represents a cluster, and the blue dots labeled -1 are noise that don't belong to any clusters.
- According to the above output there seems to be a dominant cluster in orange which basically represents the majority of the customers.
- Then there are several other smaller clusters as well such as green, red and purple that basically represent booking patterns.
- The blue represents unusual bookings, either very high ADRs or long stays or even odd combinations.
- So, by using the above analysis hotels can use it for customer segmentation.
- Since the orange cluster represents the majority of the customers with standard booking patterns, hotels can try to target them with various loyalty offers or promotions etc. to get them to visit again.
- Same goes for the smaller clusters as well where hotels can customize their promotions and deals according to the behaviors of those clusters.

- Regarding the outliers though, they could represent potential pricing mistakes made by the hotels or unusual bookings etc. and hotels should investigate them to prevent future errors.

Applying Apriori

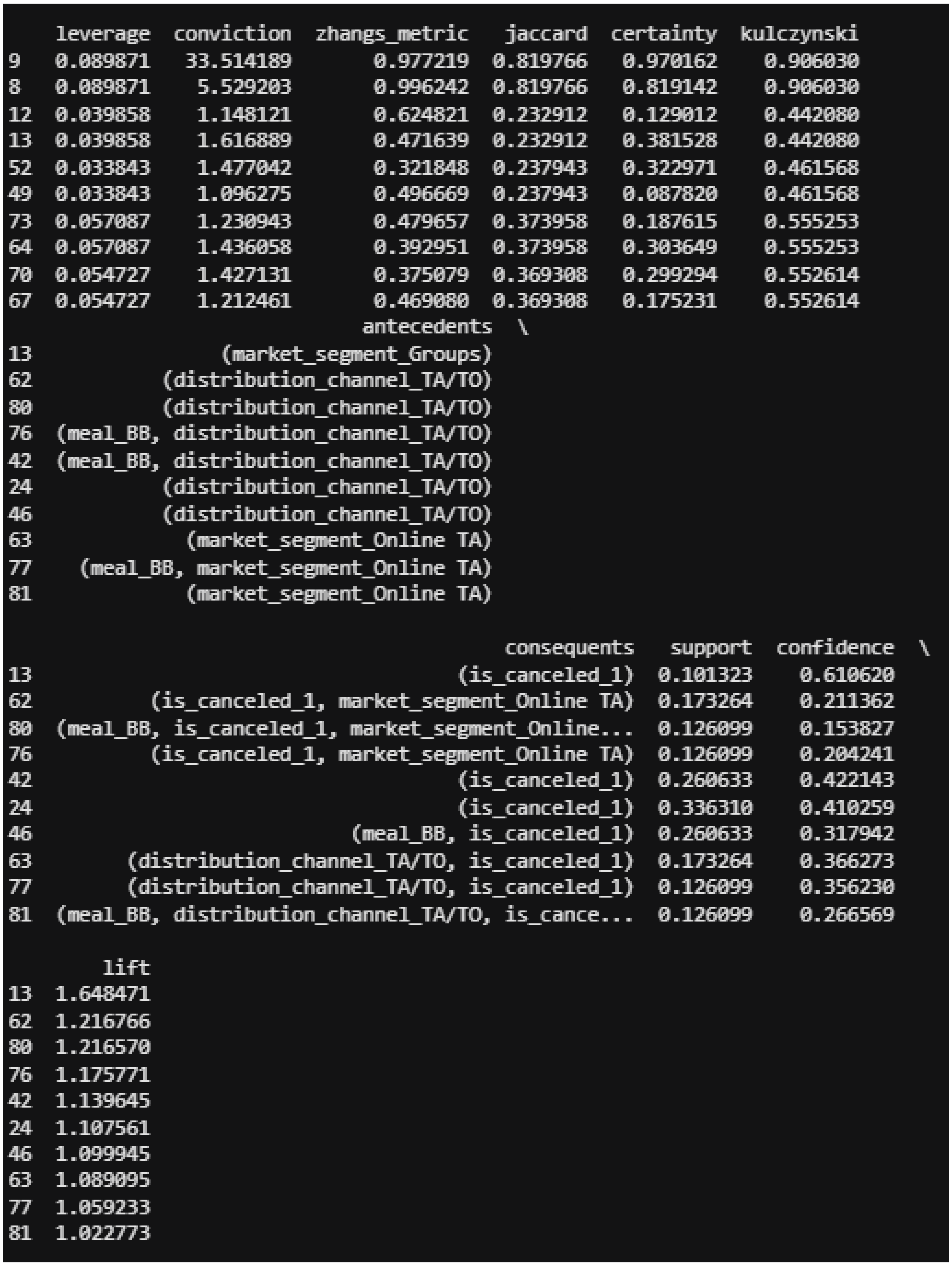
The Apriori algorithm was done to identify frequent itemsets and rules involving customer booking behavior. Then the focus was moved onto specifically onto rules that predict cancellations. The output is shown below.



```
                                          antecedents  \
9                             (market_segment_Direct)
8                       (distribution_channel_Direct)
12                                    (is_canceled_1)
13                            (market_segment_Groups)
52                     (market_segment_Offline TA/TO)
49        (is_canceled_0, distribution_channel_TA/TO)
73                         (market_segment_Online TA)
64  (is_canceled_0, meal_BB, distribution_channel_...
70                (meal_BB, market_segment_Online TA)
67        (is_canceled_0, distribution_channel_TA/TO)

                                          consequents  antecedent support  \
9                       (distribution_channel_Direct)            0.105587
8                             (market_segment_Direct)            0.122665
12                            (market_segment_Groups)            0.370416
13                                    (is_canceled_1)            0.165935
52        (is_canceled_0, distribution_channel_TA/TO)            0.202856
49                     (market_segment_Offline TA/TO)            0.483441
73  (is_canceled_0, meal_BB, distribution_channel_...            0.473046
64                         (market_segment_Online TA)            0.356772
70        (is_canceled_0, distribution_channel_TA/TO)            0.353983
67                (meal_BB, market_segment_Online TA)            0.483441

    consequent support   support  confidence      lift  representativity  \
9             0.122665  0.102823    0.973822  7.938860               1.0
8             0.105587  0.102823    0.838238  7.938860               1.0
12            0.165935  0.101323    0.273539  1.648471               1.0
13            0.370416  0.101323    0.610620  1.648471               1.0
52            0.483441  0.131912    0.650275  1.345097               1.0
49            0.202856  0.131912    0.272861  1.345097               1.0
73            0.356772  0.225856    0.477451  1.338253               1.0
64            0.473046  0.225856    0.633056  1.338253               1.0
70            0.483441  0.225856    0.638044  1.319797               1.0
67            0.353983  0.225856    0.467185  1.319797               1.0
```



```
    leverage  conviction  zhangs_metric   jaccard  certainty  kulczynski
9   0.089871   33.514189       0.977219  0.819766   0.970162    0.906030
8   0.089871    5.529203       0.996242  0.819766   0.819142    0.906030
12  0.039858    1.148121       0.624821  0.232912   0.129012    0.442080
13  0.039858    1.616889       0.471639  0.232912   0.381528    0.442080
52  0.033843    1.477042       0.321848  0.237943   0.322971    0.461568
49  0.033843    1.096275       0.496669  0.237943   0.087820    0.461568
73  0.057087    1.230943       0.479657  0.373958   0.187615    0.555253
64  0.057087    1.436058       0.392951  0.373958   0.303649    0.555253
70  0.054727    1.427131       0.375079  0.369308   0.299294    0.552614
67  0.054727    1.212461       0.469080  0.369308   0.175231    0.552614
                              antecedents  \
13                (market_segment_Groups)   
62           (distribution_channel_TA/TO)   
80           (distribution_channel_TA/TO)   
76  (meal_BB, distribution_channel_TA/TO)   
42  (meal_BB, distribution_channel_TA/TO)   
24           (distribution_channel_TA/TO)   
46           (distribution_channel_TA/TO)   
63             (market_segment_Online TA)   
77    (meal_BB, market_segment_Online TA)   
81             (market_segment_Online TA)   

                                          consequents   support  confidence  \
13                                    (is_canceled_1)  0.101323    0.610620   
62          (is_canceled_1, market_segment_Online TA)  0.173264    0.211362   
80  (meal_BB, is_canceled_1, market_segment_Online...  0.126099    0.153827   
76          (is_canceled_1, market_segment_Online TA)  0.126099    0.204241   
42                                    (is_canceled_1)  0.260633    0.422143   
24                                    (is_canceled_1)  0.336310    0.410259   
46                           (meal_BB, is_canceled_1)  0.260633    0.317942   
63        (distribution_channel_TA/TO, is_canceled_1)  0.173264    0.366273   
77        (distribution_channel_TA/TO, is_canceled_1)  0.126099    0.356230   
81  (meal_BB, distribution_channel_TA/TO, is_cance...  0.126099    0.266569   

        lift  
13  1.648471  
62  1.216766  
80  1.216570  
76  1.175771  
42  1.139645  
24  1.107561  
46  1.099945  
63  1.089095  
77  1.059233  
81  1.022773  
```

1st Rule

Antecedent: (market_segment_Groups)

Consequent: (is_canceled_1)

Here the support is 10.1%, which means that around 10% of all the bookings were group bookings and have ended up being canceled.

Confidence is 61%, which means around 61% of group bookings have also ended up canceled.

Lift is 1.65 which means group bookings are 1.65 times more likely to be canceled than average.

Therefore, this shows that group bookings in general are riskier when it comes to cancellations and hotels may want to implement more strict deposit or prepayment options/rules when it comes to people making group bookings.

2nd Rule

Antecedent (distribution_channel_TA/TO)

Consequent (is_canceled_1)

Here the confidence is 41% and the lift is 1.1-1.2 which means that bookings that were made through Travel Agencies or Tour Operators (TA/TO) show a higher cancellation risk. Therefore, hotels would need to re-evaluate their contacts with the relevant agencies and to also encourage more people to make direct bookings without using TA/Tos.

3rd Rule

Antecedent (meal_BB, market_segment_Online TA)

Consequent (is_canceled)

These types of bookings are potentially less likely to cancel based on the other parts of the table. Therefore, hotels might need to offer various promotions/deals that are targeted to this market segment in order to increase bookings.

To summarize by observing the above analysis, hotels can make better informed decisions and make the necessary changes accordingly.

Conclusion

In conclusion this project showcases how OLPA and other data warehousing and data mining concepts be applied using Google BigQuery in order to analyze hotel booking data. Initially a star schema was designed along with a fact table and the relevant dimension tables to support multidimensional analysis. By using various OLAP operations such as roll-up, roll-down, slicing and dicing and filtering we were able to extract key insights such as booking trends according to the hotels, seasonality and of course various market segments. Next, the ETL process involved cleaning the data and transforming the raw data into proper structured format ready for analytics. The insights that were gained from this project can help hotels to make better informed decisions when it comes to marketing, managing resources, and of course targeting customers. Despite some challenges this BigQuery environment did prove to be an effective tool for data analysis.

Project Link - https://console.cloud.google.com/bigquery?inv=1&invt=AbxKZg&project=golden-frame-459107-h3&ws=!1m5!1m4!4m3!1sgolden-frame-459107-h3!2sdata!3srawdata

Dashboard link - https://lookerstudio.google.com/s/iivgsb1EJmk