Forecast Analysis and SQL Queries

Christiana Mensah

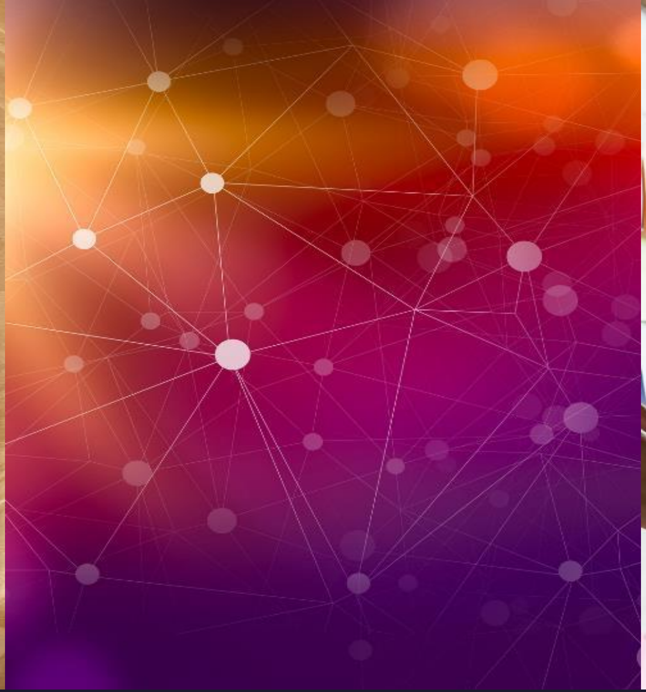
Agenda

1.0. Forecasting Case

- Thought process & results

2.0. SQL Quiz

3.0. Site Analytics Quiz



1.0. Forecast Problem Statement

- Part 1: Visualize this data and include any insights about the business that you inferred based on sales trends. (E.g., When do sales spike? What do you think that suggests about the nature of the business?)

- Part 2: Given that this data ends in March 2020, leverage a data driven method to forecast sales through December 2021. Explain your methodology and share any assumptions that you make.

- Part 3: In addition to sales, what additional information or context would you have liked in order to better forecast sales? How would this information have helped inform your forecast and / or your understanding of the business?

Forecasting Case

Thought Process Involves:

- Understanding the problem statement and tools to needed to solve it.
- Examining the data – The type of dataset, graphing the data and drawing inference, deciding on the type of analysis and prediction model needed.
- Checking results of analysis and answering the problem statement and concluding insights.

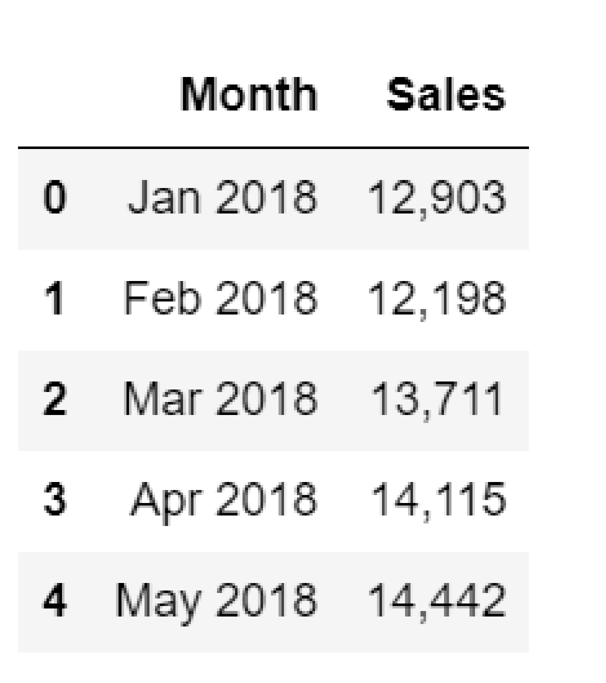
Applying the Process:

1. The question was to analyze data, draw inference and predict future events
2. Data contains two columns, a date and sales data.
3. Having a date column, shows we can apply time series analysis.
4. Prophet model chosen usually can be used for prediction. A model developed by Facebook and designed for automatic forecasting of univariate time series data.
5. Python programming language which is normally suited for machine learning modeling and widely used by data scientist was used for the analysis in a Jupyter notebook.

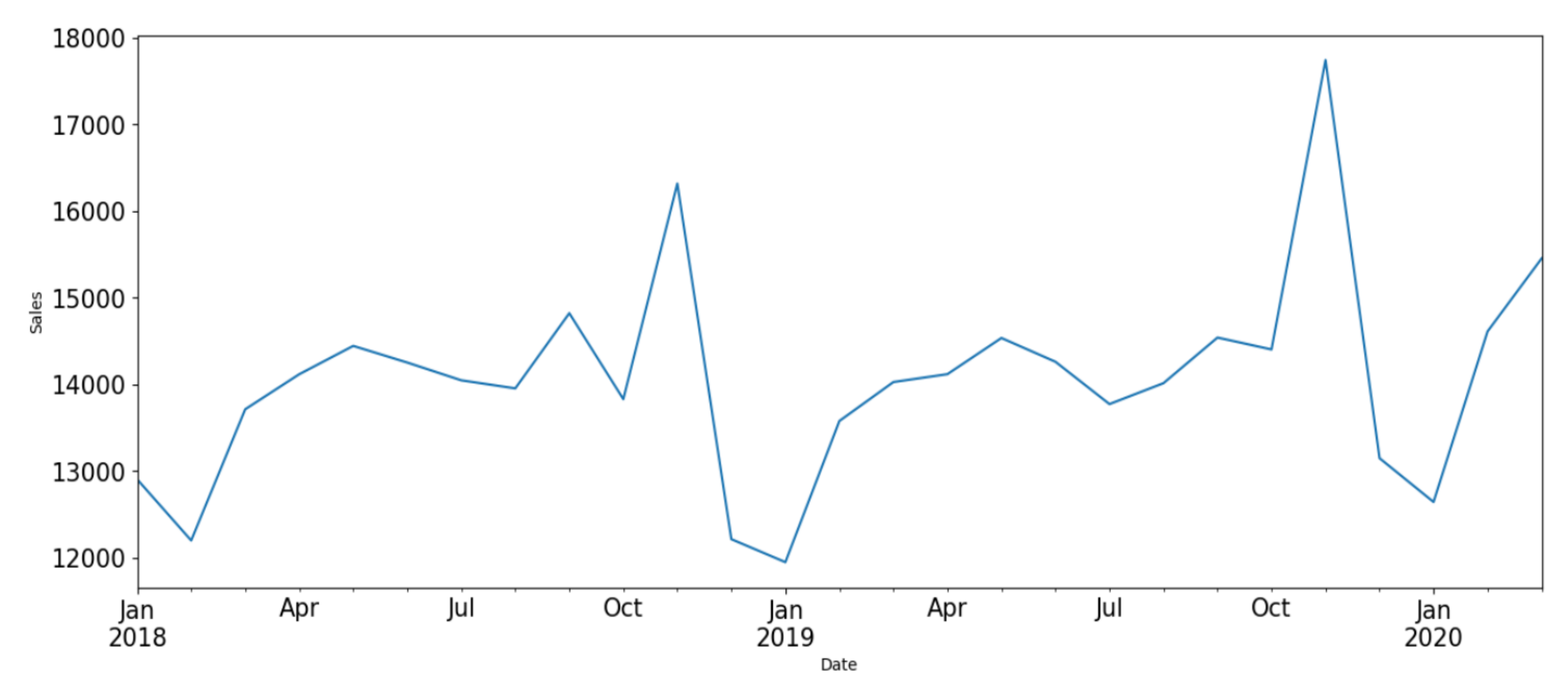
** The complete data analysis done within Jupyter notebook can be found here.

Chart

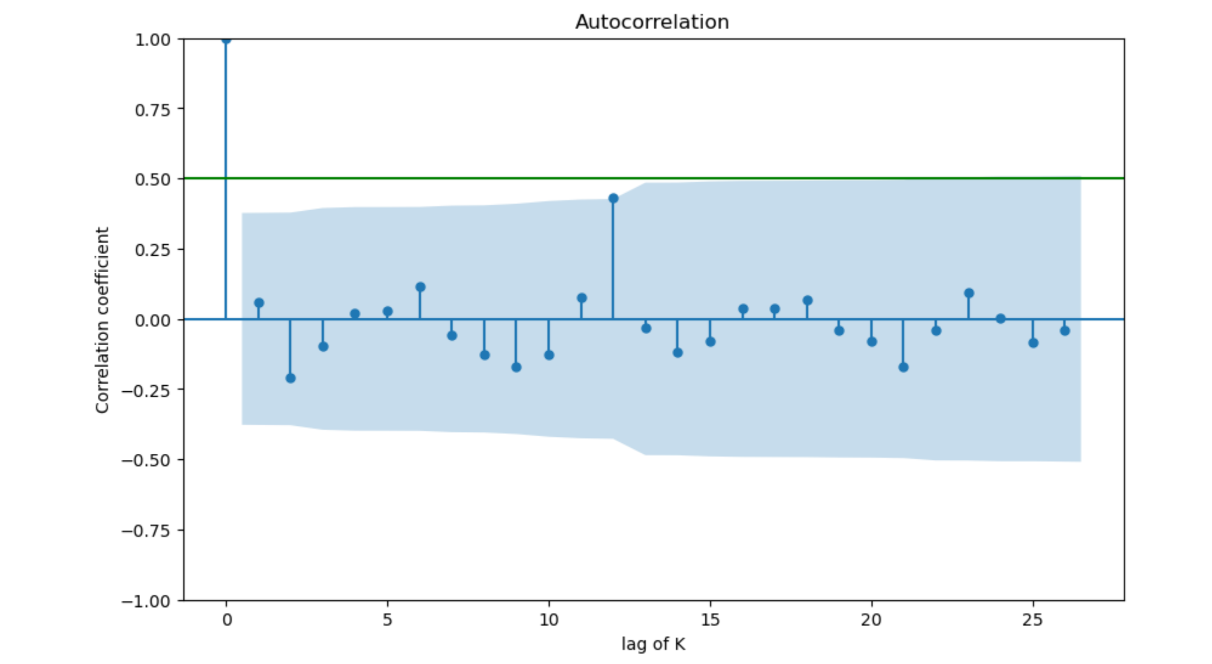
- A Simple graphing of the data showed the data spans a little over two years and consistent peaks and deeps around the same time of each year showing a trend in the data.



	Month	Sales
0	Jan 2018	12,903
1	Feb 2018	12,198
2	Mar 2018	13,711
3	Apr 2018	14,115
4	May 2018	14,442



Autocorrelation

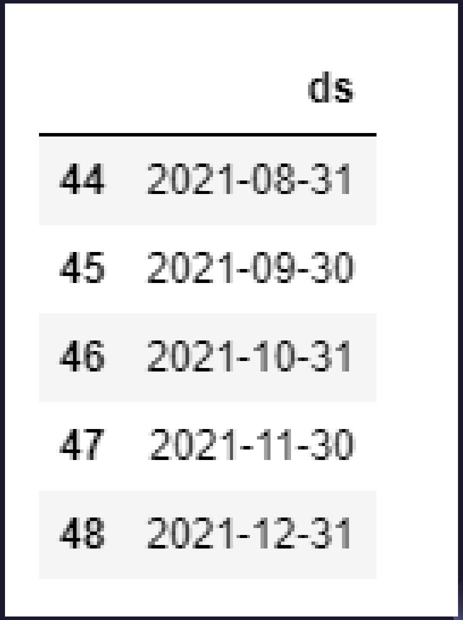


- In an Autocorrelation graph, there is a trend when the autocorrelation tends to be high at small lags like 1 or 2 and when seasonality exists, the autocorrelation goes up periodically at larger lags like we see at lag 12.
- In this graph we see a trend but because there is only one spike which is also below the 0.50 line shows little significance, there is a likelihood that seasonality does not exist.
- We also see a deep in from January to February and a slow rise in sales up until August and then there is a fluctuation and a sharp rise from October to November, then a sharp drop in December and finally slow progression in January.
- The fact that this pattern repeats in the following year suggests a trend.
- The sharp rise in **October** sales may also suggest that this store sells **Halloween products or costumes** and right after we get into the thanksgiving and Christmas season where those products are out of season the sales drop significantly.

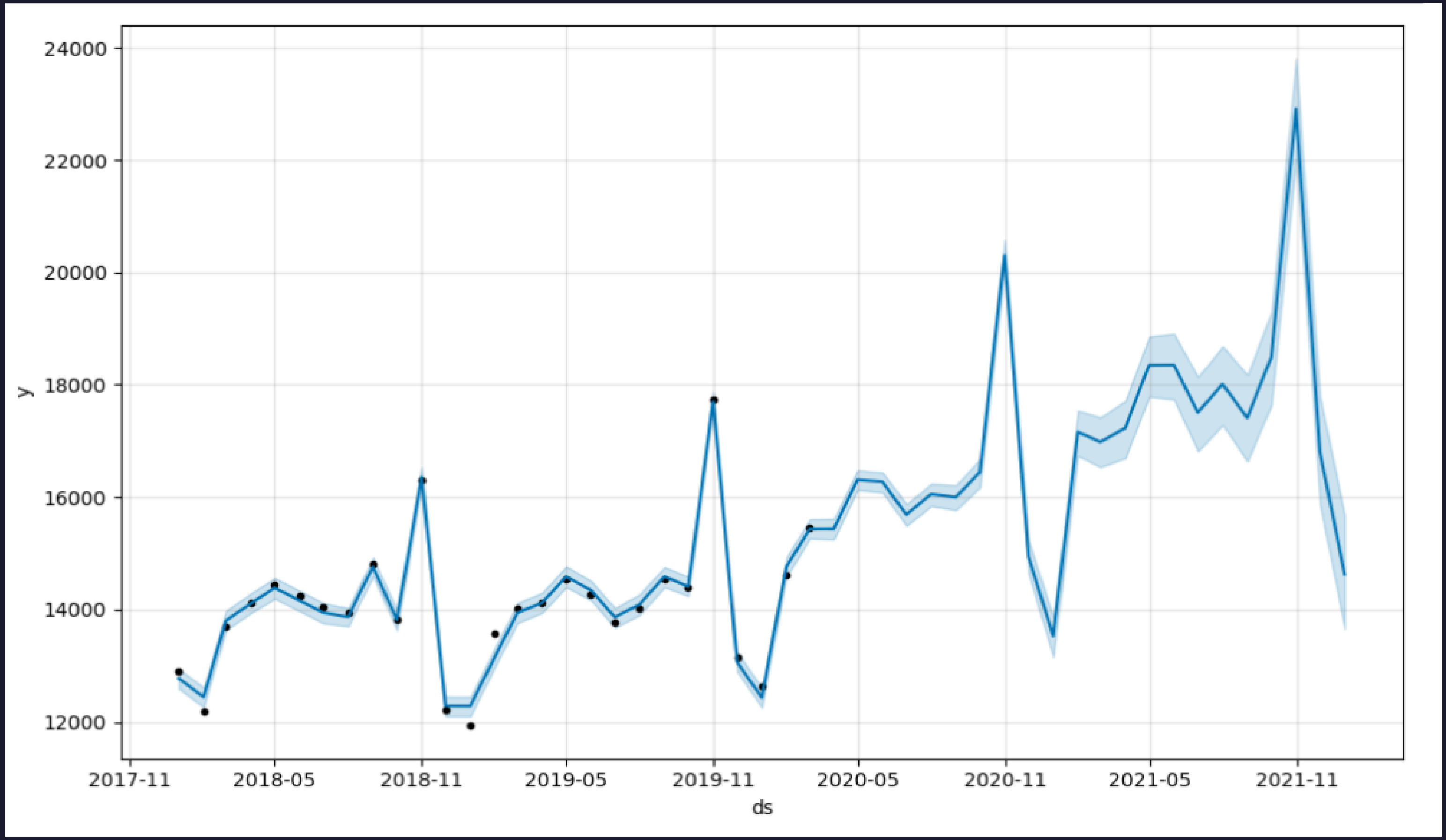
Prophet Forecast Model

The Prophet model was used for prediction automatic forecasting because we have a univariate time series data.

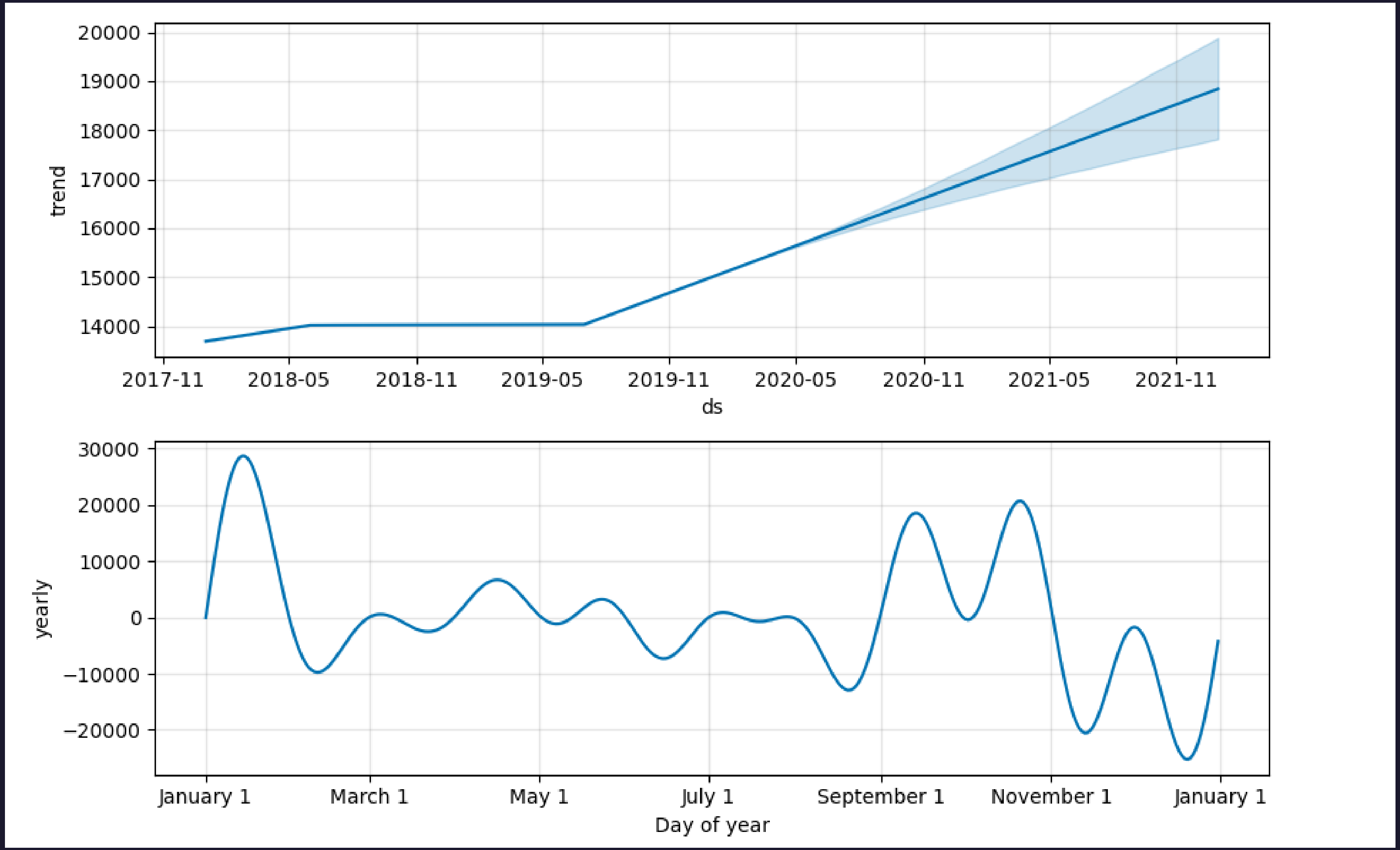
The chart shows a 22-period forecast to the end of December 2021 as requested.



	ds
44	2021-08-31
45	2021-09-30
46	2021-10-31
47	2021-11-30
48	2021-12-31



Forecast Trends



Summary

From the graphs above it shows there is a trend, and the store is most likely a costume store that sells Halloween items.

Forecast Assumptions

The following are assumptions typical of time series data:

- It was assumed that the data is stationary and that the mean and variance are constant over time.
- It can also be assumed, that the data is seasonal but there wasn't enough data to truly prove that fact because the little there was wasn't below the significant level.
- Having more data would have been helpful to ascertain seasonality.

2.0. SQL Quiz

2.1. Which product category has the highest percent of reseller customers

Solution

```
Select Product.category, (count(reseller_flag)/count(Customer.customer_id)* 100) as "Premium Customer Percent" from Customer
join  Order on  Customer.customer_id = Order.customer_id
join Product on Order.product_id =  Product.product_id
where reseseller_flag like 'Y’  group by category
order by "Premium Customer Percent" Desc  Limit by 1
```

Explanation:

Given the following information, I realize a reseller flag shows repeat customer and assumed that there is either a ‘Y’ for yes or ‘N’ for No value in the column.

Based on this I decided to calculate the premium customer percentage and then display the count of it by category.

Customer: customer_id, name, reseller_flag	(name here is customer name)
Order: order_id, order_item_id, customer_id, date, product_id, revenue	
Product: product_id, name, category	(name here is product name)

2.0. SQL Quiz

2.2. How many new customers did we acquire in 2018. (If a customer made their first purchase, then we call them as new customer (for that order only))

Solution

```
select  count(Customer.customer_id)
from (select customer_id from Order where date BETWEEN '$2018-01-01' AND '$2018-12-31') AS result
join Customer on result.customer_id = Customer.customer_id
having count(customer_id) >1
```

Explanation: Given the following information, I queried for orders in 2018 and placed the result as a subquery as a limited table to extract the count of customers that had made a exist more than once as those are considered repeat customers.

Customer: customer_id, name, reseller_flag	(name here is customer name)
Order: order_id, order_item_id, customer_id, date, product_id, revenue	
Product: product_id, name, category	(name here is product name)

3.0. Site Analytics Quiz

An electronics ecommerce store has 2 datasets capturing customer data and their online visits.

3.1. Questions

A) How many customers do we have in each region?

B) Which region has the highest share of Mobile customers?

C) How will you check : Does any customer belong to multiple regions?

Solutions:

A) *Select Distinct region, count(customer_id) from Customer group by region order by count(customer_id) Desc*

B) *Select Distinct region, max(customer_ id) from Customer where device like 'Mobile' group by region order by region Desc Limit 1*

C) *Select region, customer_Id from Customer where customer_id IN (select customer_id from Customer group by customer_id having count(distinct region)>1)*

Explanations:

A) Based on the table information given, I grouped the data by region and count the number of customers. Also ordered the region in descending order so the region with highest customer number of customer count is at the top.

B) Since we want to know about mobile devices specifically the region with the highest count, so I grouped by region, for just mobile devices and ordered by region which makes the top region the highest mobile customers.

C) I did a subquery where am selecting the region and customers from the subquery that has each customers appearing in more than one region. The resulting display will have the only customers by ID that exist in multiple regions.

3.0. Site Analytics Quiz

An electronics ecommerce store has 2 datasets capturing customer data and their online visits

3.2. Questions

A) How many visits did we receive from North America in January 2021?

B) Do customers spend more avg. time in visits on Mobile or Desktop?

C) Are BUY visits more valuable than BROWSE visits? If yes, is it more valuable on Mobile or Desktop?

D) Which product has the highest "average value per customer"?

Solutions:

A) *Select region, count(visit_id) from Customer, Visit Where (Customer.customer_id = Visit.customer_id) AND region like 'North America’ AND visit_date BETWEEN '$2021-01-01' AND '$2021-12-31') group by region*

B) *Select device, avg(visit_duration) from Customer, Visit Where Customer.customer_id = Visit.customer_id group by device order by avg(visit_duration) desc*

C) *Select device, visit_type, avg(visit_value) from Customer, Visit Where Customer.customer_id = Visit.customer_id group by visit_type, device order by avg(visit_value), device Desc*

D) *Select product, sum(visit_value)/count(customer_id) AS "Average value per customer" from Visit group by product order by "Average value per customer" Limit 1*

Explanations:

A) This was achieved by grouping customers by region and then filtering just by North American region and counting the number of times they visited just in 2021.

B) Grouping by device I found the average customer visit duration and ordering that result shows the device with the highest count at the top, assuming Mobile and Desktop are the only devices present.

C) We find this by calculating the average visit value grouping the result by visit type and device, then order by the average see the different in value by device.

D) Find the top product with the calculated average value per customer value.

3.0. Site Analytics Quiz

An electronics ecommerce store has 2 datasets capturing customer data and their online visits.

3.3. Questions

A) [BONUS]What is avg. time in days between visits in North America?

B) How can this metric be useful for analysis? Hint : For each customer, for each visit in North America, find the difference between that visit and the previous visit. Average this quantity over all visits"

Solutions:

A) *Select Customer.customer_id, avg(visit_duration)/365 AS "Average Days" from Customer,Visit Where Customer.customer_id = Visit.customer_id AND region like 'North America' group by Customer.customer_id*

B) *It shows the average frequency of a customer visiting a site and can help with logistics, overhead expenses in terms or when someone should be available to schedule shipping and if shipping could be done in bulk based on the average frequency of customer visits.*

Explanations:

A) Since we are looking for time in days, I divided the average visit duration by 365 days in a year, grouping result by individual customer and filtering the results to include just North American region.

B) Understanding the number of days traffic flows or how often it does on a site, helps to plan for marketing, logistics and other important actions such as IT maintenance or site refresh.

Thank You

Christiana Mensah

Github Link