databricks

Booking_assignment_Detailed_Approch (Python)

Import Notebook

Technology Used: Azure Databricks(Apache Spark)

Langaue used : SQL/Python

1.High speed data querying, analysis, and transformation with large data sets.

2.Supports multiple languages and integrations with other popular products.

3.Can Schedule job and deal with realtime data.

4.Can be used to Create Visualization as well

Approch Used:

1.Schema Creation.

2.Data ingestion.

3.Data Cleaning & Transformation.

4.Creating Processed datset on which analysis can happen.

5.Doing Analysis.

```python
storage_account_name = "bookingassignment"
storage_account_key  = "s3qXVgHfZPAF3fP8MuK6d0O6FIDHjLieogBxlDA5IdMghTcfLNUN2oZEHTjvG6aaoZjrTorj7gyS+Eh3RY/yaA=="
spark.conf.set(
    f"fs.azure.account.key.{storage_account_name}.dfs.core.windows.net",
    f"{storage_account_key}")

container_name= "raw" # make sure to use the right container

from pyspark.sql.types import *
from pyspark.sql.functions import *
```

Data Ingestion & Schema Creation

1. Creating Schema for Both "Portfolio" and "Operational" table. So,we could have well defined datatype for all the columns of the table.
2. Doing data ingestion with well defined Schema

```python
#1 Schema Creation
Booking_schema = StructType([StructField("property_id", IntegerType(), False),
                                     StructField("Account_manager", StringType(), True),
                                  ])

operational_data_schema = StructType([StructField("property_id", IntegerType(), False),
                                     StructField("contact_id", IntegerType(), True),
                                     StructField("contact_date",StringType(), True),
                                     StructField("product", StringType(), True),
                                     StructField("product_status", StringType(), True),
                                     StructField("contact_channel", StringType(), True),
                                     StructField("office", StringType(), True),
                                     StructField("region", StringType(), True),
                                     StructField("account_manager", StringType(), True)])
```

```python
#2 Data ingestion from source with correct Schema
Portfolio = spark.read.csv(f"abfss://{container_name}@{storage_account_name}.dfs.core.windows.net/Data for RA
Assessment - Portfolio.csv",header = True ,schema = Booking_schema)
operational_data = spark.read.csv(f"abfss://{container_name}@{storage_account_name}.dfs.core.windows.net/Data for RA
Assessment - Operational Data.csv",header = True ,schema = operational_data_schema)
```

Data Cleaning and Data Preparation

1.Setting up formate for "Contact Date".

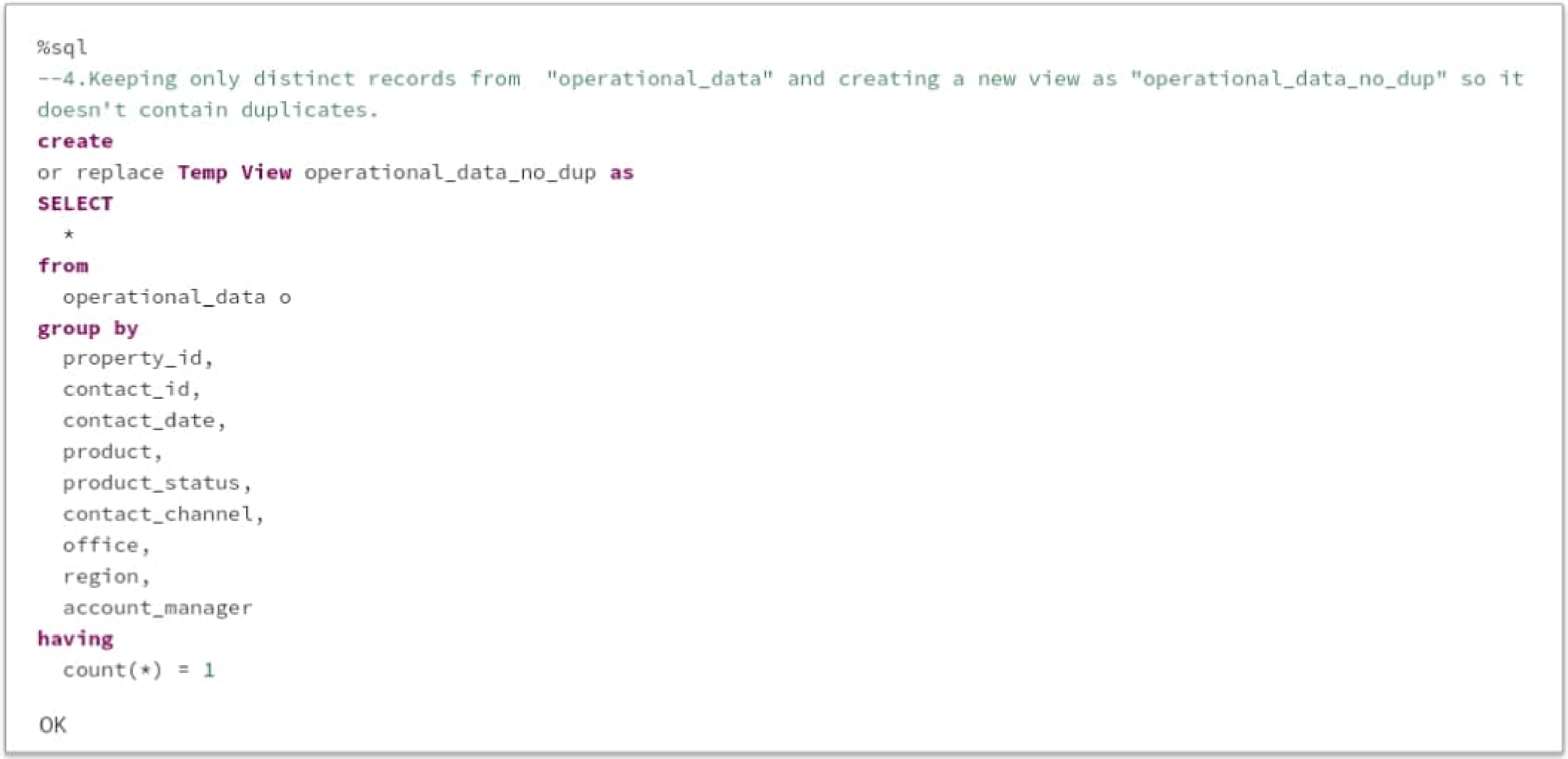
2.Value correction - Replacing "tōkyo" to "Tokyo" in the office column of operational_data.

3.Found Null value, Replace the null value of Region with "AMERICA" because office belong to Buenos Aires.

4.Keeping only distinct records from "operational_data" and creating a new view as "operational_data_no_dup" so it doesn't contain duplicates.

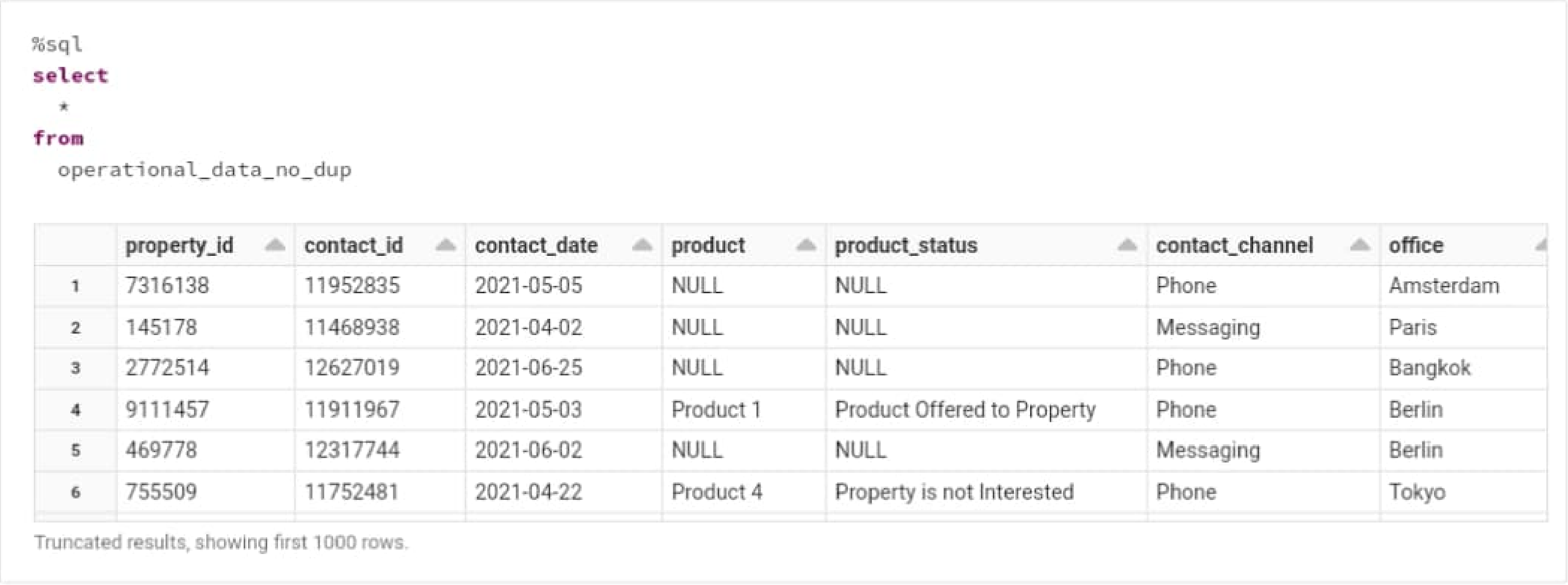
```
#1.Setting up formate for "Contact Date".
operational_data = operational_data.withColumn('contact_date',to_date(operational_data.contact_date, 'dd-MM-yyyy'))

#2. Value correction - Replacing "tōkyo" to "Tokyo" in the office column of operational_data.
operational_data=operational_data.withColumn('office', translate('office', 'ō', 'o'))

#3. Found Null value, Replace the null value of Region with "AMERICA" because office belong to Buenos Aires
operational_data = operational_data.na.fill({'region':'AMERICAS'})
```

```
Portfolio.createOrReplaceTempView("Portfolio")
operational_data.createOrReplaceTempView("operational_data")
```



```
%sql
--4.Keeping only distinct records from  "operational_data" and creating a new view as "operational_data_no_dup" so it
doesn't contain duplicates.
create
or replace Temp View operational_data_no_dup as
SELECT
  *
from
  operational_data o
group by
  property_id,
  contact_id,
  contact_date,
  product,
  product_status,
  contact_channel,
  office,
  region,
  account_manager
having
  count(*) = 1
```

OK



```
%sql
select
  *
from
  operational_data_no_dup
```

	property_id	contact_id	contact_date	product	product_status	contact_channel	office
1	7316138	11952835	2021-05-05	NULL	NULL	Phone	Amsterdam
2	145178	11468938	2021-04-02	NULL	NULL	Messaging	Paris
3	2772514	12627019	2021-06-25	NULL	NULL	Phone	Bangkok
4	9111457	11911967	2021-05-03	Product 1	Product Offered to Property	Phone	Berlin
5	469778	12317744	2021-06-02	NULL	NULL	Messaging	Berlin
6	755509	11752481	2021-04-22	Product 4	Property is not Interested	Phone	Tokyo

Truncated results, showing first 1000 rows.

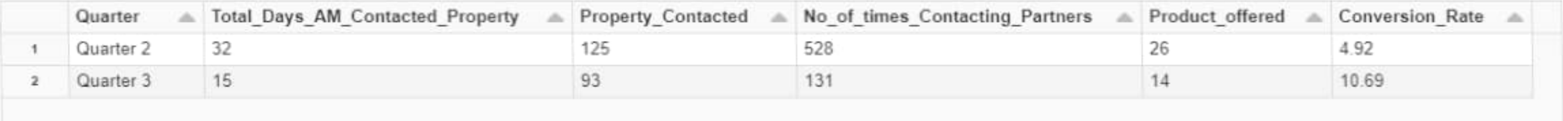
Question 1

“Dear Colleague, As an account manager I am responsible for contacting my portfolio (partners assigned to me) every quarter and offering products that can help these partners increase their performance.

I am an Account Manager (‘Account Manager 5’ in the dataset) in APAC and as a mid quarter check in, I would like to see how I have performed so far in this quarter and what are the focus areas for the rest of the quarter.”

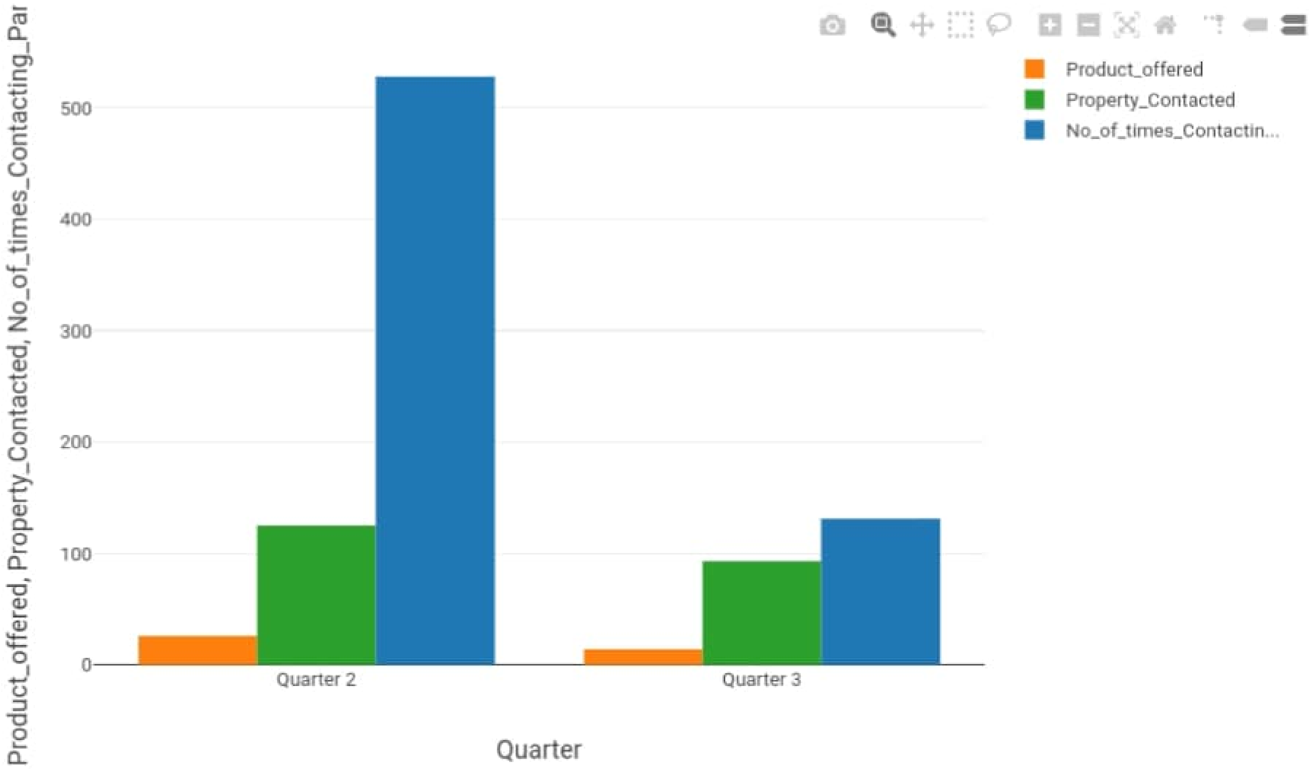
Solution 1.1 Mid Quarter-Q3 Analysis

```
%sql
-- Based On data available from:Q2 (April 1- june 30) , Q3 Ongoing (july-1 - Aug10)

select Quarter,
count(distinct f.Total_Days_AM_Contacted_Property) Total_Days_AM_Contacted_Property,
count(distinct f.Property_Contacted) Property_Contacted ,
count(f.No_of_times_Contacting_Partners) No_of_times_Contacting_Partners,
count(f.Product_offered) Product_offered,
round(count(f.Product_offered) / count(f.No_of_times_Contacting_Partners)*100,2) as Conversion_Rate


from (
SELECT contact_date ,
 case when contact_date  >='2021-01-04'  and contact_date<='2021-06-30' then  "Quarter 2"
       when contact_date  >='2021-07-01' and  contact_date  <= '2021-09-30' then "Quarter 3" end as Quarter,

       case when contact_date  >='2021-01-04'  and contact_date<='2021-06-30' then  contact_date
       when contact_date  >='2021-07-01' and  contact_date  <= '2021-09-30' then contact_date end as
Total_Days_AM_Contacted_Property,



  case when contact_date  >='2021-01-04'  and contact_date<='2021-06-30' then  p.property_id
    when contact_date  >='2021-07-01' and  contact_date  <= '2021-09-30' then  p.property_id end as Property_Contacted,

  case when contact_date  >='2021-01-04'  and contact_date<= '2021-06-30'  then o.contact_id
        when contact_date  >='2021-07-01'  and  contact_date  <= '2021-09-30'   then o.contact_id  end as
No_of_times_Contacting_Partners,


  case when contact_date  >='2021-01-04'  and contact_date<='2021-06-30' and product_status = 'Product Offered to
Property' then o.contact_id
         when contact_date  >='2021-07-01'  and  contact_date  <= '2021-09-30'  and product_status = 'Product Offered to
Property' then o.contact_id end as Product_offered


from
  portfolio p
inner join (
    select
      *
    from
      operational_data_no_dup
    where account_manager = 'Account Manager 5') o
on p.property_id = o.property_id) f
group by Quarter
order by Product_offered desc
```



	Quarter	Total_Days_AM_Contacted_Property	Property_Contacted	No_of_times_Contacting_Partners	Product_offered	Conversion_Rate
1	Quarter 2	32	125	528	26	4.92
2	Quarter 3	15	93	131	14	10.69

Show code



1.1(a): Performance Measure

It seems Account Manager 5 is taking a different Approach this Current quarter, he has Contacted less no of times(131) this quarter in comparison to Previous(528) but still manages to offer products to 14 Properties.

Conversion Rate(efficiency) is significantly higher(14/131)= 10.69 in Comparision to Previous Quarter 2 (26/528) = 4.92 . It seems he is giving more effort in convincing the Partners instead of Making more no of Blank calls.

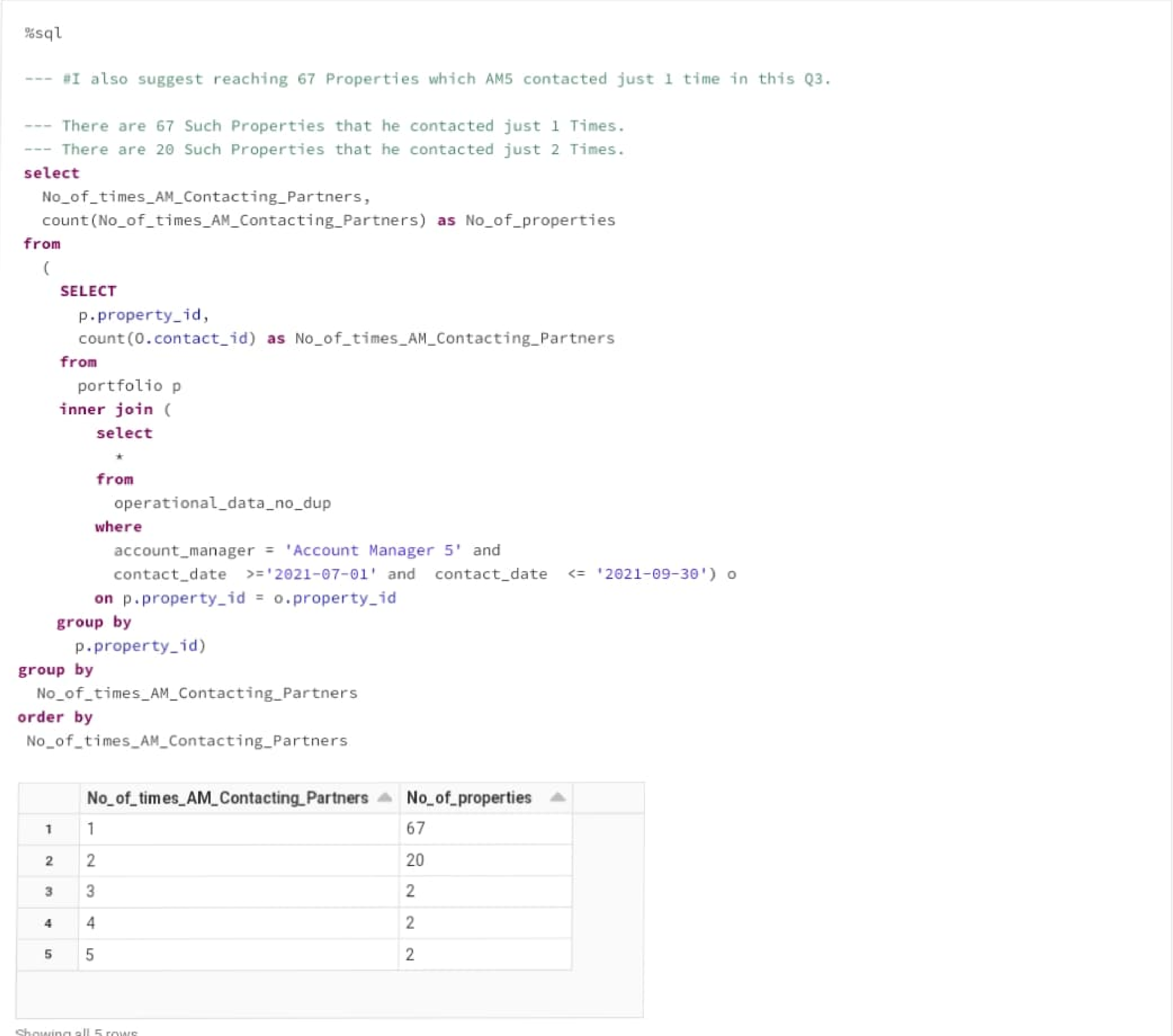
Overall Performance seems good,in fact improving and going in Right direction.

Conversion Rate = Product Offered/ No of contacts made to Partner

```
%sql
-- #I Highly recommend reaching out to those 7 Properties first which are in 'Property is Interested', 'Awaiting
Property Decision' Status.

select distinct property_id from  operational_data_no_dup
where account_manager = 'Account Manager 5' and  product_status in ('Property is Interested' ,'Awaiting Property
Decision')
```

	property_id ▲
1	2873228
2	10362730
3	7461934
4	492044
5	3604951
6	7043651
7	3209125

Showing all 7 rows

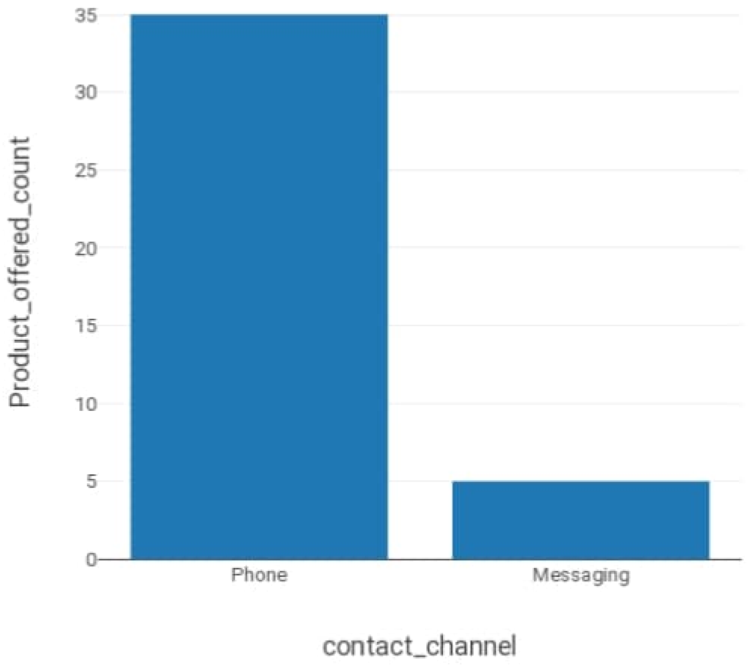


```
%sql

--- #I also suggest reaching 67 Properties which AM5 contacted just 1 time in this Q3.

--- There are 67 Such Properties that he contacted just 1 Times.
--- There are 20 Such Properties that he contacted just 2 Times.
select
  No_of_times_AM_Contacting_Partners,
  count(No_of_times_AM_Contacting_Partners) as No_of_properties
from
  (
    SELECT
      p.property_id,
      count(O.contact_id) as No_of_times_AM_Contacting_Partners
    from
      portfolio p
    inner join (
        select
          *
        from
          operational_data_no_dup
        where
          account_manager = 'Account Manager 5' and
          contact_date  >='2021-07-01' and  contact_date  <= '2021-09-30') o
        on p.property_id = o.property_id
    group by
      p.property_id)
group by
  No_of_times_AM_Contacting_Partners
order by
 No_of_times_AM_Contacting_Partners
```

	No_of_times_AM_Contacting_Partners ▲	No_of_properties ▲
1	1	67
2	2	20
3	3	2
4	4	2
5	5	2

Showing all 5 rows

```
%sql

-- # Overall from Q3 till date, AM5 can offer Product(40) more times incall(35) over Message(5). I highly recommend
using Phone over Message
SELECT
  contact_channel,
  count(product_status) as Product_Offered_Count
from
  portfolio p
  inner join (
    select
      *
    from
      operational_data_no_dup
    where
      account_manager = 'Account Manager 5'
  ) o on p.property_id = o.property_id
where
  product_status = 'Product Offered to Property'
group by
  contact_channel,
  product_status
```



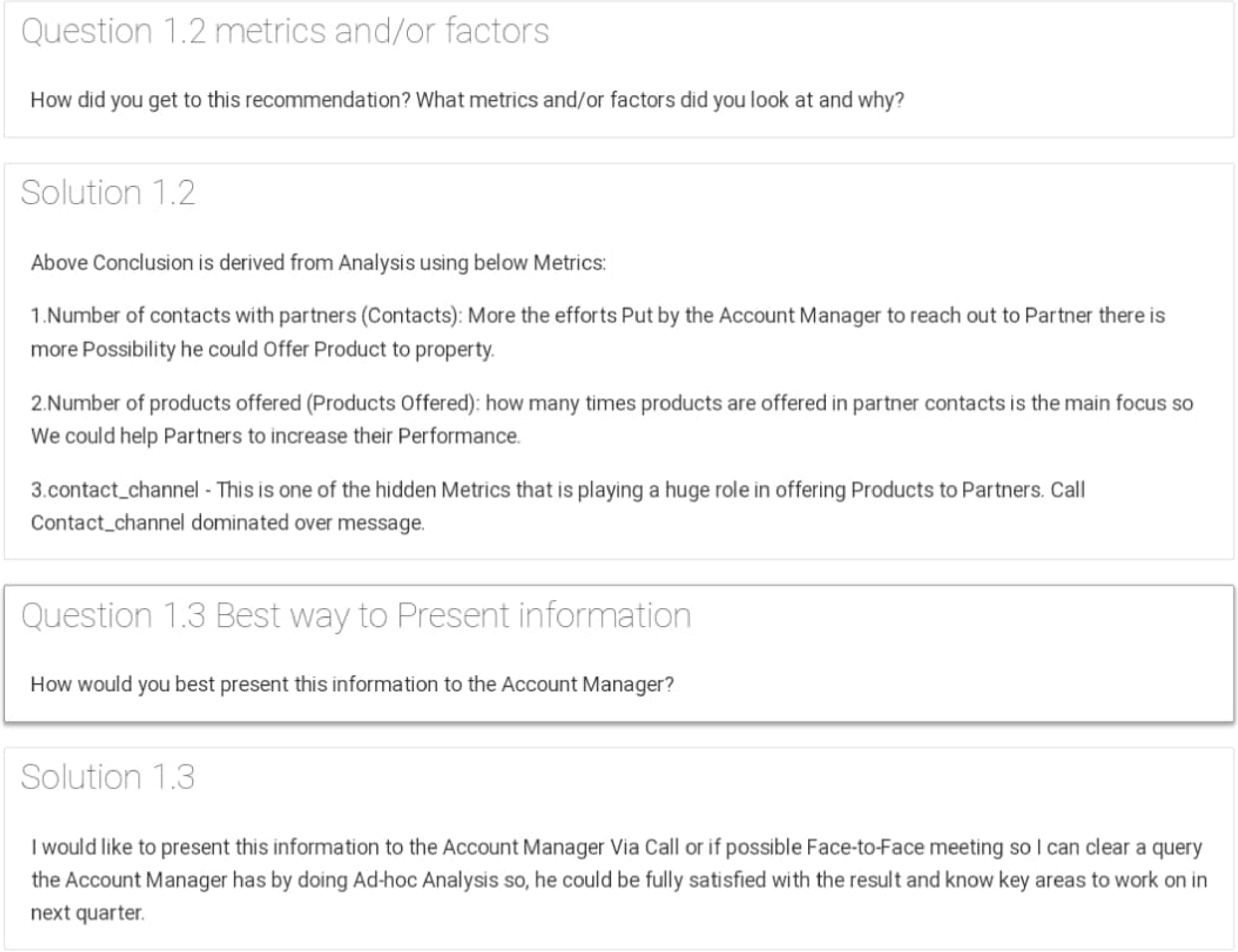
1.1(b) Recommendation

1.I Highly recommend reaching out to those 7 Properties first which are in 'Property is Interested', 'Awaiting Property Decision' Status.

2.AM5 approached 93 Properties so far out of 220 Properties in Q3. I recommend Contacting the remaining Properties as well i.e 127 Properties.

3.I also suggest reaching 67 Properties which AM5 contacted just 1 time in this Q3.

4.Overall from Q3 till date, AM5 can offer Product(40) more times incall(35) over Message(5). I highly recommend using Phone over Message



Question 1.2 metrics and/or factors

How did you get to this recommendation? What metrics and/or factors did you look at and why?

Solution 1.2

Above Conclusion is derived from Analysis using below Metrics:

1.Number of contacts with partners (Contacts): More the efforts Put by the Account Manager to reach out to Partner there is more Possibility he could Offer Product to property.

2.Number of products offered (Products Offered): how many times products are offered in partner contacts is the main focus so We could help Partners to increase their Performance.

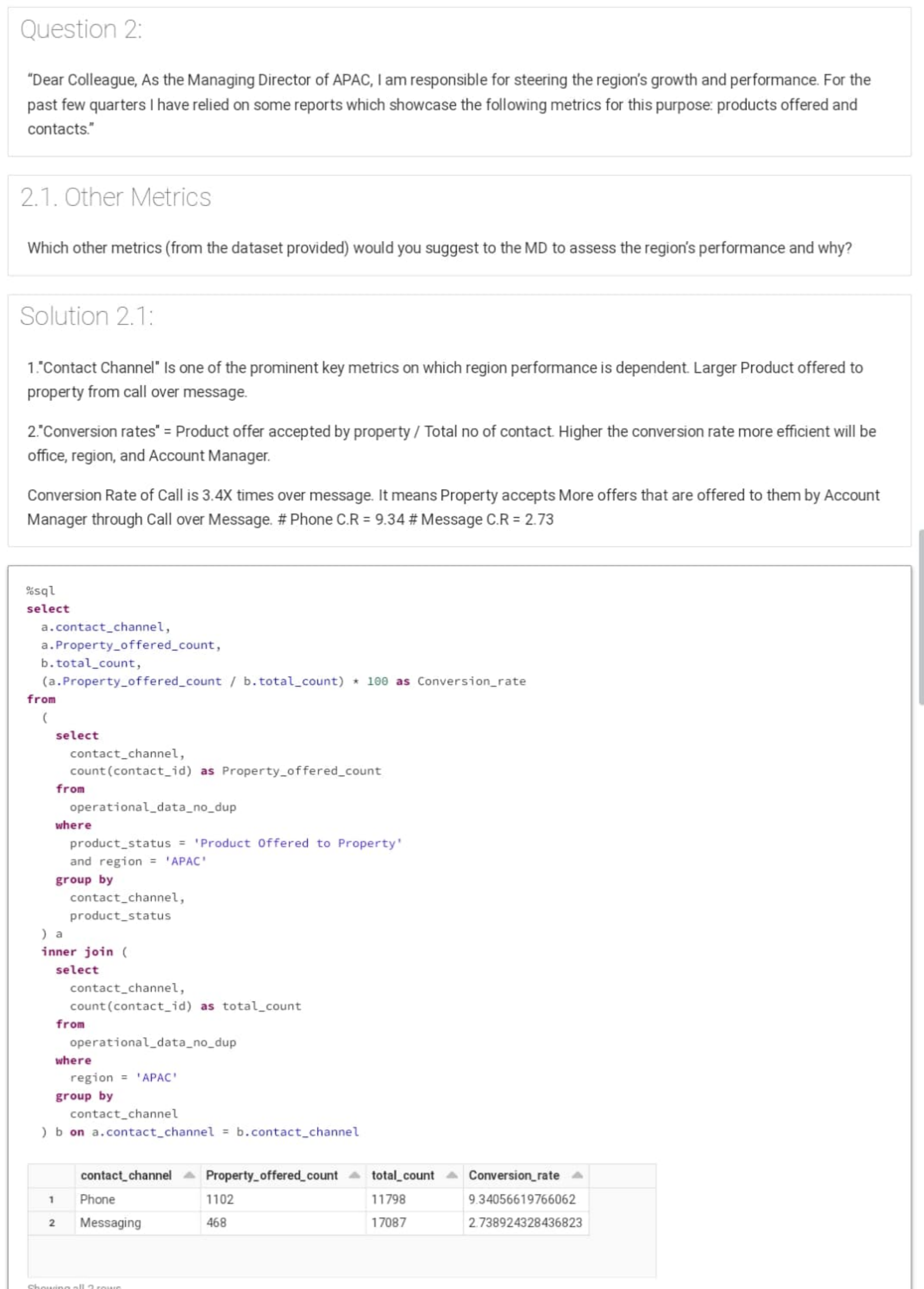
3.contact_channel - This is one of the hidden Metrics that is playing a huge role in offering Products to Partners. Call Contact_channel dominated over message.

Question 1.3 Best way to Present information

How would you best present this information to the Account Manager?

Solution 1.3

I would like to present this information to the Account Manager Via Call or if possible Face-to-Face meeting so I can clear a query the Account Manager has by doing Ad-hoc Analysis so, he could be fully satisfied with the result and know key areas to work on in next quarter.



Question 2:

"Dear Colleague, As the Managing Director of APAC, I am responsible for steering the region's growth and performance. For the past few quarters I have relied on some reports which showcase the following metrics for this purpose: products offered and contacts."

2.1. Other Metrics

Which other metrics (from the dataset provided) would you suggest to the MD to assess the region's performance and why?

Solution 2.1:

1."Contact Channel" Is one of the prominent key metrics on which region performance is dependent. Larger Product offered to property from call over message.

2."Conversion rates" = Product offer accepted by property / Total no of contact. Higher the conversion rate more efficient will be office, region, and Account Manager.

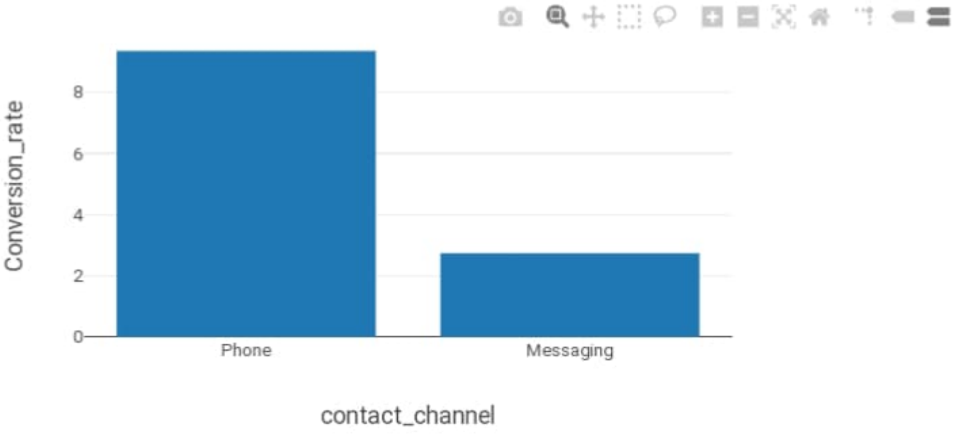
Conversion Rate of Call is 3.4X times over message. It means Property accepts More offers that are offered to them by Account Manager through Call over Message. # Phone C.R = 9.34 # Message C.R = 2.73

```
%sql
select
  a.contact_channel,
  a.Property_offered_count,
  b.total_count,
  (a.Property_offered_count / b.total_count) * 100 as Conversion_rate
from
  (
    select
      contact_channel,
      count(contact_id) as Property_offered_count
    from
      operational_data_no_dup
    where
      product_status = 'Product Offered to Property'
      and region = 'APAC'
    group by
      contact_channel,
      product_status
  ) a
  inner join (
    select
      contact_channel,
      count(contact_id) as total_count
    from
      operational_data_no_dup
    where
      region = 'APAC'
    group by
      contact_channel
  ) b on a.contact_channel = b.contact_channel
```

	contact_channel	Property_offered_count	total_count	Conversion_rate
1	Phone	1102	11798	9.34056619766062
2	Messaging	468	17087	2.738924328436823

Showing all 2 rows

Show code



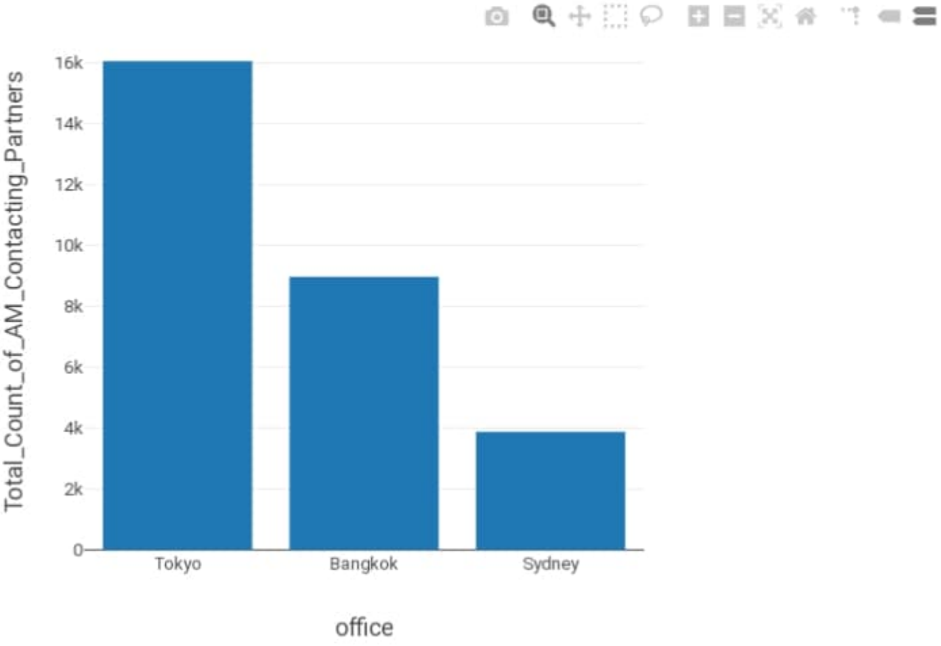
2.2 APAC region's performance

The MD needs to present his region's performance to the VP. How would your report look to accomplish his goal?

Solution 2.2:

1.No of Contacts with Partners: "Tokoyo" office in 'APAC' region has highest no in terms of reaching out to Partners. Tokoyo - 16044 Bangkok - 8965 Sydney - 3876

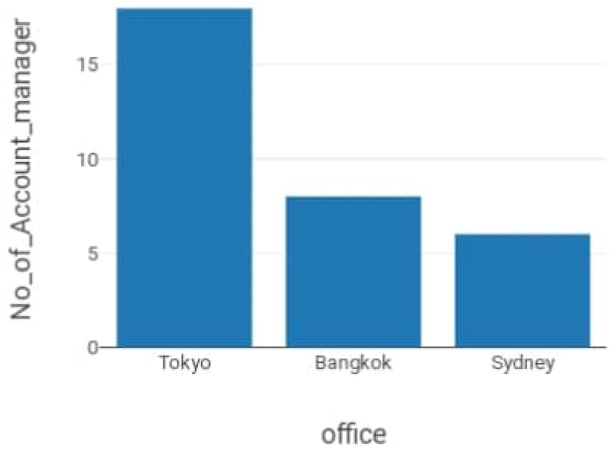
```
%sql
select
  office,
  count(contact_id) as Total_Count_of_AM_Contacting_Partners
from
  operational_data_no_dup
where
  region = 'APAC'
group by
  office
order by
  count(contact_id) desc
```

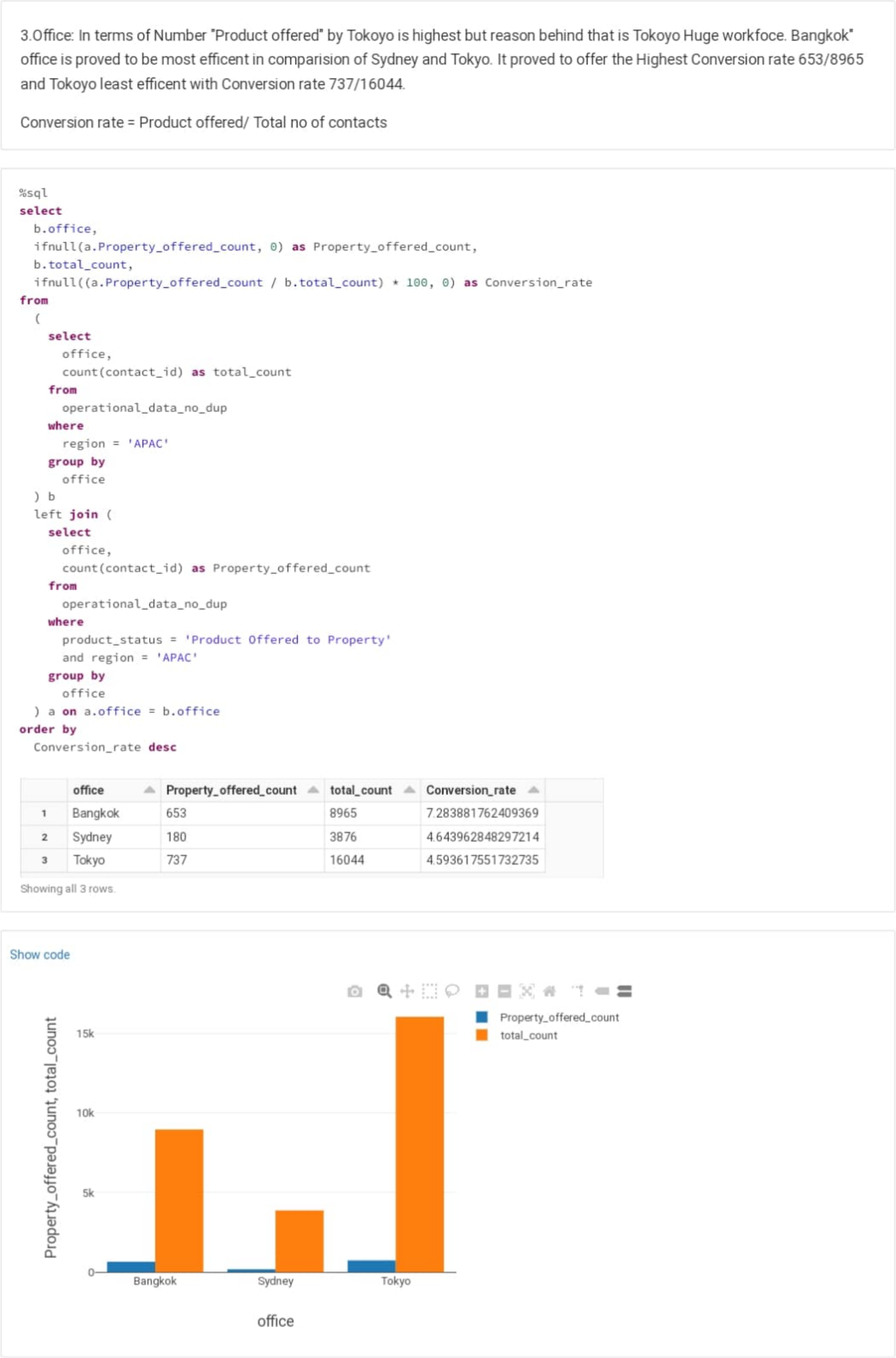


2.No of Account Manager in each office: "Tokoyo" office in 'APAC' region has highest no of employees. It clearly justify the reason why tokoyo office also have Highest No of Contacting Partners.

Tokoyo - 18 Bangkok - 8 Sydney - 6

```
%sql
select
  office,
  count(distinct account_manager) as No_of_Account_manager
from
  operational_data_no_dup
where
  region = 'APAC'
group by
  office
order by
  count(distinct account_manager) desc
```





3.Office: In terms of Number "Product offered" by Tokoyo is highest but reason behind that is Tokoyo Huge workfoce. Bangkok" office is proved to be most efficent in comparision of Sydney and Tokyo. It proved to offer the Highest Conversion rate 653/8965 and Tokoyo least efficent with Conversion rate 737/16044.

Conversion rate = Product offered/ Total no of contacts

```
%sql
select
  b.office,
  ifnull(a.Property_offered_count, 0) as Property_offered_count,
  b.total_count,
  ifnull((a.Property_offered_count / b.total_count) * 100, 0) as Conversion_rate
from
  (
    select
      office,
      count(contact_id) as total_count
    from
      operational_data_no_dup
    where
      region = 'APAC'
    group by
      office
  ) b
  left join (
    select
      office,
      count(contact_id) as Property_offered_count
    from
      operational_data_no_dup
    where
      product_status = 'Product Offered to Property'
      and region = 'APAC'
    group by
      office
  ) a on a.office = b.office
order by
  Conversion_rate desc
```

	office	Property_offered_count	total_count	Conversion_rate
1	Bangkok	653	8965	7.283881762409369
2	Sydney	180	3876	4.643962848297214
3	Tokyo	737	16044	4.593617551732735

Showing all 3 rows.

Show code

Question 3: Free Response

What additional insights can you derive from the provided dataset? Please feel free to look into and explore anything that seems interesting to you. Share your findings with the analytics team, which comprises both analysts and business stakeholders.

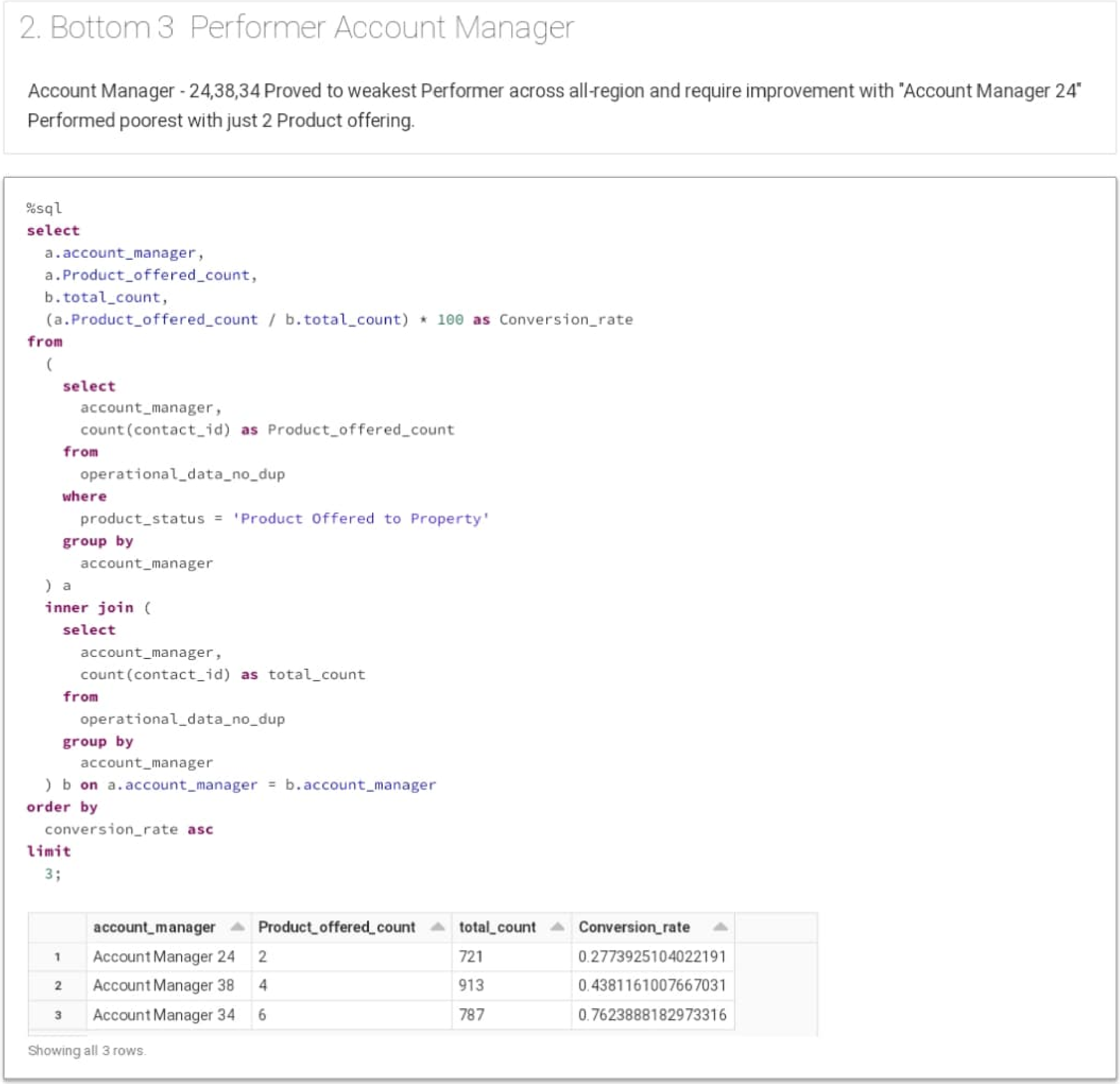
1.Top 3 Performer Account Manager

Account Manager - 36,23,73 Proved to top 3 Performer across all regions with "Account Manager 36" Performed exceptionally best with 535 Highest Product offering.

```
%sql
select
  a.account_manager,
  a.Product_offered_count,
  b.total_count,
  (a.Product_offered_count / b.total_count) * 100 as Conversion_rate
from
  (
    select
      account_manager,
      count(contact_id) as Product_offered_count
    from
      operational_data_no_dup
    where
      product_status = 'Product Offered to Property'
    group by
      account_manager
  ) a
  inner join (
    select
      account_manager,
      count(contact_id) as total_count
    from
      operational_data_no_dup
    group by
      account_manager
  ) b on a.account_manager = b.account_manager
order by
  conversion_rate desc
limit
  3;
```

	account_manager	Product_offered_count	total_count	Conversion_rate
1	Account Manager 36	535	797	67.1267252195734
2	Account Manager 23	170	516	32.94573643410852
3	Account Manager 73	259	886	29.23250564334085

Showing all 3 rows.



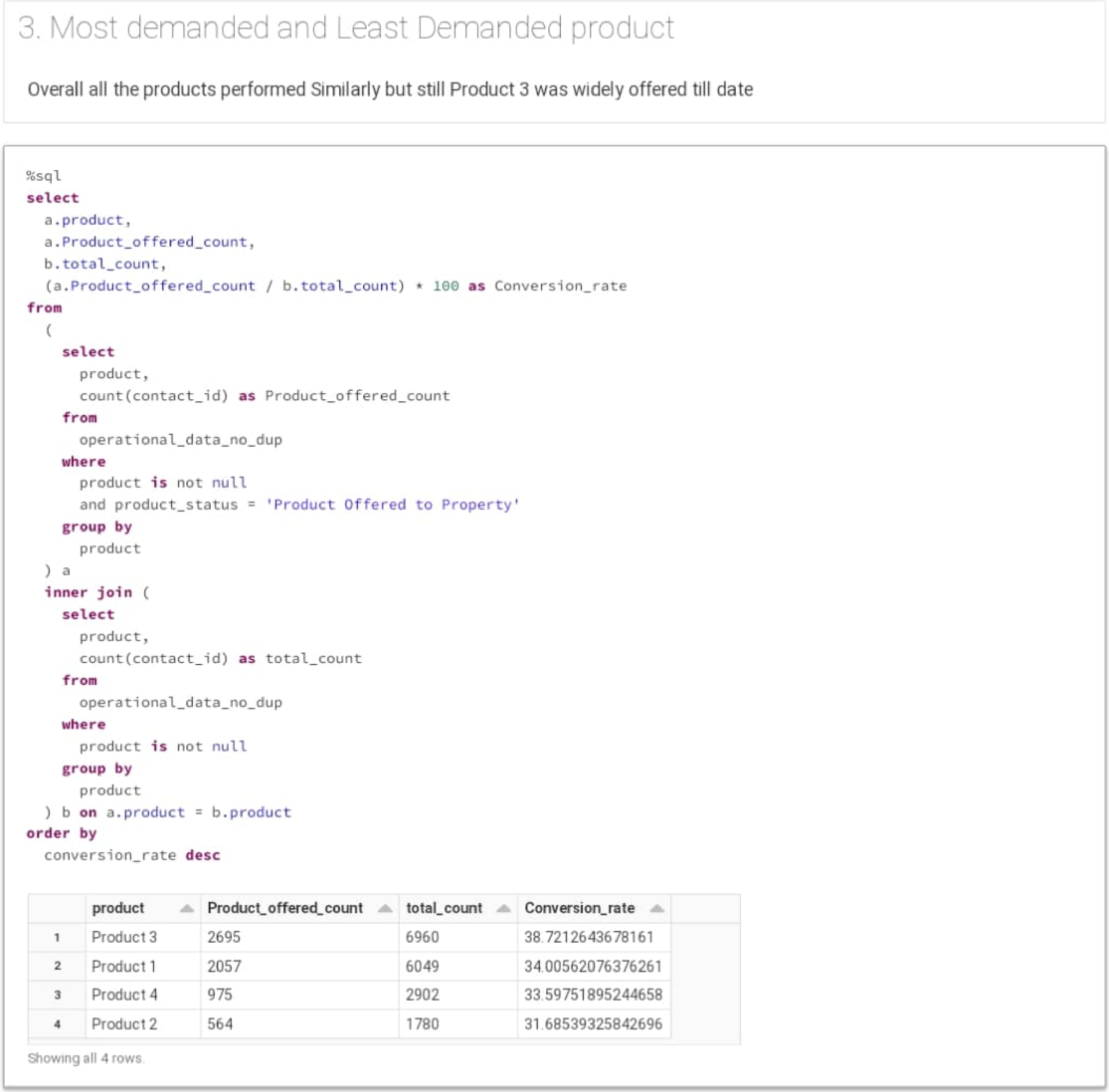
2. Bottom 3 Performer Account Manager

Account Manager - 24,38,34 Proved to weakest Performer across all-region and require improvement with "Account Manager 24" Performed poorest with just 2 Product offering.

```
%sql
select
  a.account_manager,
  a.Product_offered_count,
  b.total_count,
  (a.Product_offered_count / b.total_count) * 100 as Conversion_rate
from
  (
    select
      account_manager,
      count(contact_id) as Product_offered_count
    from
      operational_data_no_dup
    where
      product_status = 'Product Offered to Property'
    group by
      account_manager
  ) a
  inner join (
    select
      account_manager,
      count(contact_id) as total_count
    from
      operational_data_no_dup
    group by
      account_manager
  ) b on a.account_manager = b.account_manager
order by
  conversion_rate asc
limit
  3;
```

	account_manager	Product_offered_count	total_count	Conversion_rate
1	Account Manager 24	2	721	0.2773925104022191
2	Account Manager 38	4	913	0.4381161007667031
3	Account Manager 34	6	787	0.7623888182973316

Showing all 3 rows.



3. Most demanded and Least Demanded product

Overall all the products performed Similarly but still Product 3 was widely offered till date

```
%sql
select
  a.product,
  a.Product_offered_count,
  b.total_count,
  (a.Product_offered_count / b.total_count) * 100 as Conversion_rate
from
  (
    select
      product,
      count(contact_id) as Product_offered_count
    from
      operational_data_no_dup
    where
      product is not null
      and product_status = 'Product Offered to Property'
    group by
      product
  ) a
  inner join (
    select
      product,
      count(contact_id) as total_count
    from
      operational_data_no_dup
    where
      product is not null
    group by
      product
  ) b on a.product = b.product
order by
  conversion_rate desc
```

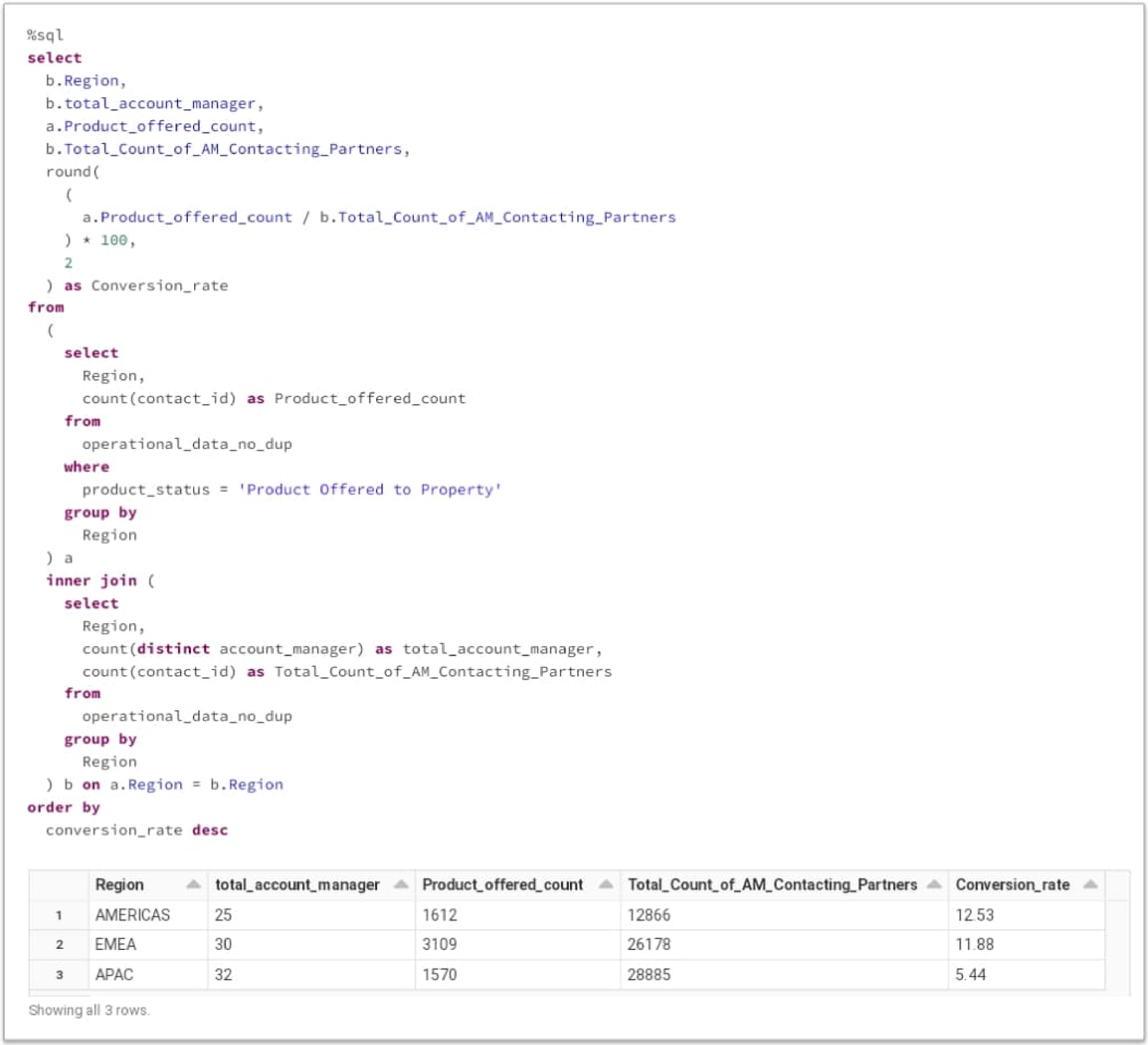
	product	Product_offered_count	total_count	Conversion_rate
1	Product 3	2695	6960	38.7212643678161
2	Product 1	2057	6049	34.00562076376261
3	Product 4	975	2902	33.59751895244658
4	Product 2	564	1780	31.68539325842696

Showing all 4 rows.

4. Region-wise Performance

1."AMERICAS" region Proved to be most efficient region who has highest Conversion Rate even with lesser number of Account Manager.

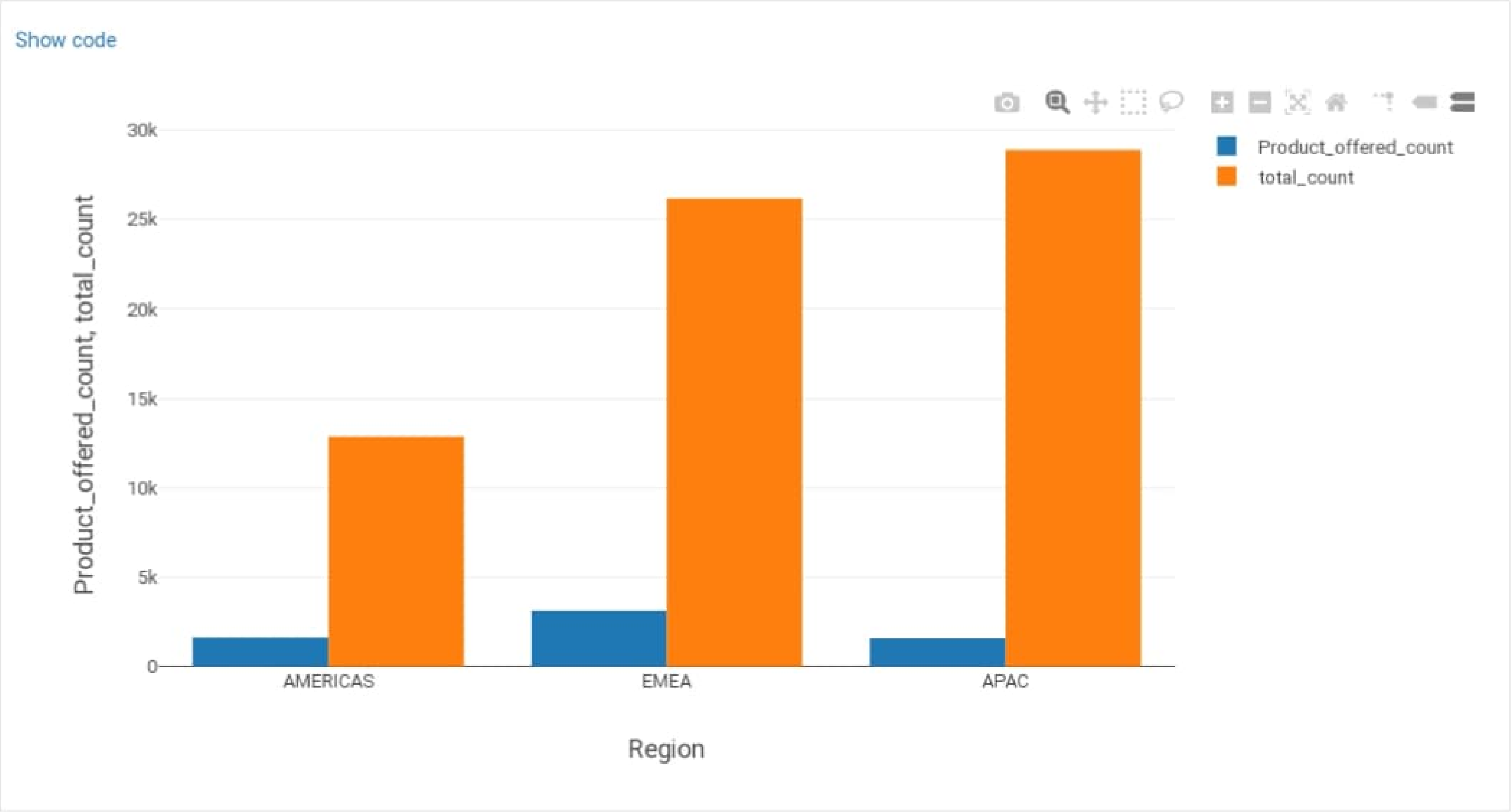
2."APAC" region proved to be weakest Performer even with most number of Account Manager 32, APAC has offered least product 1570 with Conversion Rate of just 5.44.



```sql
%sql
select
  b.Region,
  b.total_account_manager,
  a.Product_offered_count,
  b.Total_Count_of_AM_Contacting_Partners,
  round(
    (
      a.Product_offered_count / b.Total_Count_of_AM_Contacting_Partners
    ) * 100,
    2
  ) as Conversion_rate
from
  (
    select
      Region,
      count(contact_id) as Product_offered_count
    from
      operational_data_no_dup
    where
      product_status = 'Product Offered to Property'
    group by
      Region
  ) a
  inner join (
    select
      Region,
      count(distinct account_manager) as total_account_manager,
      count(contact_id) as Total_Count_of_AM_Contacting_Partners
    from
      operational_data_no_dup
    group by
      Region
  ) b on a.Region = b.Region
order by
  conversion_rate desc
```

	Region	total_account_manager	Product_offered_count	Total_Count_of_AM_Contacting_Partners	Conversion_rate
1	AMERICAS	25	1612	12866	12.53
2	EMEA	30	3109	26178	11.88
3	APAC	32	1570	28885	5.44

Showing all 3 rows.



Show code

30k
25k
20k
15k
10k
5k
0
Product_offered_count, total_count
AMERICAS
EMEA
APAC
Region
🟦 Product_offered_count
🟧 total_count