Mini Portfolio:
SQL for Data Science
[Retno Prabaningrum]

Intensive Data Science
by @myskill.id

Introduction

In a company, there are various departments. Now, the business department needs to know how many pieces kind of fruits were sold in August because of that the data department was asked to be able to present information according to the request from the business department.
Here's the data available:

	user_id [PK] integer	product_name character varying (255)	quantity integer	purchase_date date	store_city_id integer	price_per_kg integer
1	1	Mangga	13	2021-07-21	10	10000
2	2	Jeruk	17	2022-01-22	4	9000
3	3	Jeruk	18	2022-04-29	4	9000
4	4	Anggur	3	2022-03-30	9	13000
5	5	Pisang	13	2022-04-07	3	8000
6	6	Melon	11	2021-02-04	4	11000
7	7	Apel	6	2021-11-08	5	15000
8	8	Apel	6	2020-09-30	5	15000
9	9	Semangka	18	2021-10-23	4	12000
10	10	Nanas	5	2020-07-22	8	11000
11	11	Markisa	15	2021-03-31	3	11000
12	12	Pepaya	19	2021-06-07	5	8000
13	13	Nanas	7	2022-04-14	7	11000
14	14	Pepaya	2	2021-07-07	9	8000
15	15	Semangka	12	2021-08-14	8	12000
16	16	Pepaya	13	2021-12-07	3	8000
17	17	Apel	19	2021-01-14	2	15000
18	18	Kurma	3	2020-09-03	9	30000
19	19	Salak	15	2021-09-23	5	9000
20	20	Kurma	5	2021-07-17	4	30000
21	21	Apel	12	2021-03-24	3	15000

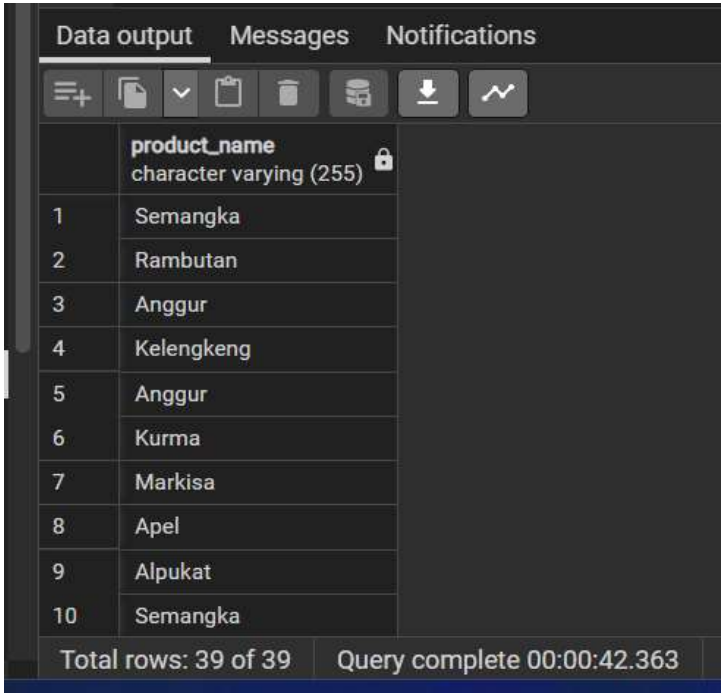
Total rows: 1000 of 1000 Query complete 00:00:09.524

So, how?

First, the data team did was look at all the necessary data. See all the columns to understanding whether the data in that column is complete in the query. The required query is:

```
SELECT product_name FROM
sales
WHERE purchase_date
BETWEEN '2021-08-01' AND
'2021-08-31'
```

And this is the result



Data output | Messages | Notifications

	product_name character varying (255)
1	Semangka
2	Rambutan
3	Anggur
4	Kelengkeng
5	Anggur
6	Kurma
7	Markisa
8	Apel
9	Alpukat
10	Semangka

Total rows: 39 of 39 | Query complete 00:00:42.363

Then..

The team was asked to display data on least of all sales of mangoes and apples at Store City 9 only. I fulfilled the request with this query:

```
SELECT * FROM sales
        WHERE
product_name        IN
('Mangga', 'Apel')
        AND
store_city_id=9
```

	user_id [PK] integer	product_name character varying (255)	quantity integer	purchase_date date	store_city_id integer	price_per_kg integer
1	191	Mangga	16	2020-07-26	9	10000
2	213	Mangga	2	2020-05-26	9	10000
3	224	Apel	18	2022-02-06	9	15000
4	235	Mangga	7	2021-01-03	9	10000
5	382	Mangga	10	2022-02-07	9	10000
6	388	Apel	9	2022-05-18	9	15000
7	430	Mangga	6	2021-01-25	9	10000
8	467	Apel	4	2020-09-04	9	15000
9	480	Apel	20	2021-02-11	9	15000
10	534	Apel	17	2022-01-17	9	15000
11	573	Apel	15	2022-03-07	9	15000
12	650	Apel	11	2022-05-18	9	15000
13	674	Mangga	5	2021-06-08	9	10000
14	692	Mangga	9	2021-09-07	9	10000
15	709	Mangga	11	2022-01-26	9	10000
16	725	Apel	13	2020-08-19	9	15000
17	729	Apel	17	2021-07-14	9	15000

Total rows: 19 of 19 Query complete 00:00:02.219

Suddenly there is a need that must be met. So the business department asks to display the salary for 3 years for employees starting with the letter 'ba'. The query needs is:

```
SELECT first_name,
        salary_2020,
        salary_2021,
        salary_2022
FROM employees
WHERE first_name LIKE 'Ba%'
```

Then the display appears as below:

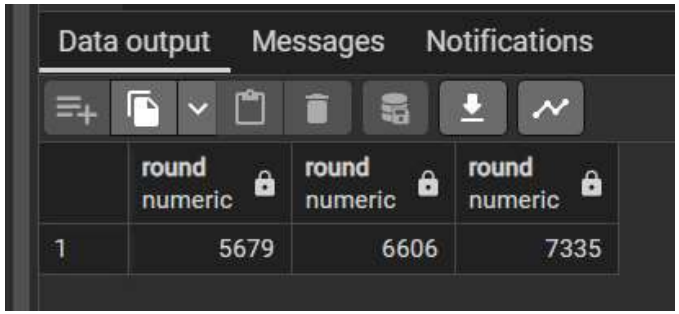
	first_name character varying (255)	salary_2020 integer	salary_2021 integer	salary_2022 integer
1	Barry	1019	7672	2138
2	Basia	7024	5610	3588
3	Baxter	3292	9713	1420
4	Baker	4460	1920	5395

Request From Financial Department..

Furthermore, there is a request from the finance department to display salary data on the average salary of employees per year. To make it easier, salary data must be rounded.

```
SELECT ROUND(AVG
(salary_2020)),
ROUND(AVG (salary_2021)),
ROUND(AVG (salary_2022))
FROM employees
```

Then the display appears as below:



Then, to answer the needs of the branch company, the finance department also asked the team to display records of the names of employees in Yogyakarta and West Java stores.

```
SELECT * FROM sales
WHERE store_city_id IN
(SELECT store_city_id
FROM region
WHERE region
IN('Yogyakarta','Jawa
Barat'))
```

The data of region is here:

	store_city_id [PK] integer	region character varying (255)
1	1	Sumatera
2	2	Yogyakarta
3	3	Jawa Timur
4	4	Sulawesi
5	5	Bali
6	6	Jawa Barat
7	7	Kalimantan
8	8	Jawa Tengah
9	9	Bandung
10	10	Jakarta

Then the display appears on the next slide:

Here’s the result:

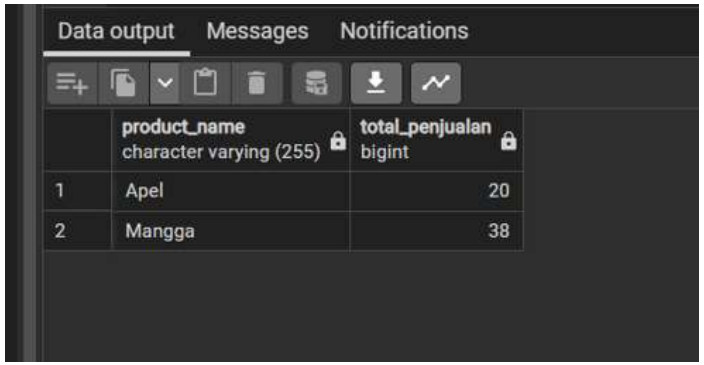
	employee_id [PK] integer	store_city_id integer	first_name character varying (255)	last_name character varying (255)	email character varying (255)	department character varying (255)	hire_date character varying (255)	salary_2020 integer	salary_2021 integer
1	4	6	Channing	Robert	robertchanning18@pro...	Asset Management	2019-11-17	7772	4146
2	5	2	Tanek	Ocean	tanek-ocean5079@yah...	Customer Service	2018-10-19	5145	11952
3	9	6	Orlando	MacKensie	orlando-mackensie@y...	Data Analyst	2019-03-03	6829	7614
4	12	6	Ima	Heather	heather-ima4730@hot...	Accounting	2019-03-07	5323	3414
5	13	2	Adam	Deborah	[null]	Sales and Marketing	2018-05-10	2146	3186
6	22	6	Xantha	Erin	erin-xantha280@icloud...	Tech Support	2019-11-04	4076	4435
7	28	2	Cruz	Leroy	c-leroy@aol.couk	Accounting	2019-09-22	5514	7930
8	30	2	Nolan	Kellie	n.kellie@google.com	Accounting	2020-02-14	8654	2187
9	32	6	Ryder	Simone	s-ryder@google.org	Human Resources	2019-07-13	2689	6764
10	34	6	Leigh	Carly	leigh-carly@icloud.com	Media Relations	2019-07-12	6873	6140
11	36	6	Darryl	Quyn	d_quyn@protonmail.edu	Advertising	2018-05-19	2022	4306
12	39	6	Ruby	Kiona	r_kiona@outlook.com	Sales and Marketing	2020-01-24	7539	9168
13	51	6	Gareth	Christian	christian_gareth@iclou...	Sales and Marketing	2019-05-23	3923	1308
14	53	6	Amelia	Kerry	kamelia22@yahoo.ca	Media Relations	2018-05-12	1882	4389
15	60	6	Dieter	Kylan	dkylan532@outlook.co...	Finances	2018-04-03	7972	2997
16	62	6	Palmer	Winter	winter_palmer4319@o...	Media Relations	2018-12-20	3269	1334

Back to Business Department Request

After a period of time, the business department asks for data that displays the record for the total quantity of Mangoes and Apples for 3 weeks after Idul Fitri 2022. The queries required are as follows:

```
SELECT product_name,
SUM(quantity) as total_penjualan
FROM sales
WHERE product_name IN ('Mangga',
'Apel')
AND purchase_date BETWEEN '2022-
05-01' AND '2022-05-22'
GROUP BY 1;
```

Then the display appears as below:



To filled the company needs, data is needed to display records employee. who work in Bali and Yogyakarta. This is done using the subquery method.

```
SELECT first_name, last_name, store_city_id
FROM employees
WHERE store_city_id IN(SELECT store_city_id
FROM region
WHERE region IN('Yogyakarta','Bali'))
```

Then the display appears on the next slide:

Data output Messages Notifications

	first_name character varying (255)	last_name character varying (255)	store_city_id integer
1	Willa	Grady	5
2	Tanek	Ocean	2
3	Ursa	Gary	5
4	Inez	Tashya	5
5	Adam	Deborah	2
6	Bruno	Honorato	5
7	Cameron	Jameson	5
8	Rylee	Jordan	5
9	Cruz	Leroy	2
10	Nolan	Kellie	2
11	Brian	Jada	5
12	Wayne	Dawn	5
13	Declan	Darius	5
14	Robert	Graham	5
15	Addison	Kristen	5
16	Mark	Nevada	5

Total rows: 116 of 116 Query complete 00:00:00.117

From the employees data, data is needed that displays the number of employees based on their salary category in store 9 in 2020.
The categories are LOWER, MIDDLE, and HIGH.
For the LOWER category in the range : < 4000 ;
MIDDLE : >= 4000 – 7000;
HIGHER : > 7000

Then the display appears as below:

	kategori_gaji text	jml_karyawan bigint
1	HIGH	1
2	LOWER	1
3	LOWER	1
4	HIGH	1
5	HIGH	1
6	HIGH	1
7	HIGH	1
8	MIDDLE	1

Also, there is data on car sales as well as shown below:

Data output Messages Notifications

	id_product [PK] integer	product_name character varying (255)	harga integer
1	1	Avanza	10000
2	2	Pajero	15000
3	3	Ford	30000
4	4	Pajero Sport	70000

From this data, I want to change the name of the Avanza car to Inova, while the Pajero to Tesla. The query is below:

```
SELECT penjualan,
       REPLACE(product_name,  'Pajero','Tesla')
AS modified_product,
           REPLACE(product_name,
'Avanza','Innova') AS modified_product
FROM penjualan
```

Then the display appears as below:

Data output Messages Notifications

	penjualan penjualan	modified_product text	modified_product text
1	(1,Avanza,10000)	Avanza	Innova
2	(2,Pajero,15000)	Tesla	Pajero
3	(3,Ford,30000)	Ford	Ford
4	(4,"Pajero Sport",70000)	Tesla Sport	Pajero Sport

From fruit sales data, a query is also carried out that returns/displays the average total income in the Sumatra and Kalimantan stores

```
WITH you AS(
        SELECT
        product_name,
        store_city_id,
SUM(quantity * price_per_kg) AS total_pendapatan
        FROM sales
        GROUP BY 1,2)
SELECT region, AVG(total_pendapatan)
FROM you a
JOIN region r
USING(store_city_id)
WHERE region IN ('Sumatera', 'Kalimantan')
GROUP BY 1
```

Then the display appears as below:

Data output | Messages | Notifications

	region character varying (255)	avg numeric
1	Kalimantan	1023800.000000000000
2	Sumatera	648285.714285714286

With the same request being made for the City of Yogyakarta and Sulawesi in 2021. Here’s the required query:

```
WITH table_a AS (
    SELECT product_name,
      store_city_id, purchase_date,
      SUM(quantity*price_per_kg) AS total
FROM sales
WHERE purchase_date BETWEEN '2021-01-01' AND '2021-12-31'
GROUP BY product_name, store_city_id, purchase_date
ORDER BY store_city_id
),
AVG_amount_per_region AS (SELECT store_city_id,
                          AVG(total) AS avg_total
FROM table_a
GROUP BY store_city_id)

SELECT * FROM AVG_amount_per_region
WHERE store_city_id IN (SELECT store_city_id FROM region)
```

Then the display appears as below:

	store_city_id integer	avg_total numeric
1	1	155218.750000000000
2	2	119416.666666666667
3	3	141150.943396226415
4	4	118803.921568627451
5	5	128029.411764705882
6	6	141043.478260869565
7	7	[illegible]

Total rows: 10 of 10 Query complete 00:00:00.813

Then, using the sales data and region_data also perform a query that displays the records total income from the sale of fruit outside Java and Java, categorize areas by island, For example, Bandung is part of the island of Java.

```
SELECT SUM(quantity * price_per_kg) AS Total_Pendapatan,
   CASE WHEN store_city_id = 2 THEN 'PULAU JAWA'
        WHEN store_city_id = 3 THEN 'PULAU JAWA'
        WHEN store_city_id = 6 THEN 'PULAU JAWA'
        WHEN store_city_id = 8 THEN 'PULAU JAWA'
        WHEN store_city_id = 9 THEN 'PULAU JAWA'
        WHEN store_city_id = 10 THEN 'PULAU JAWA'
        WHEN store_city_id = 1 THEN 'LUAR PULAU JAWA'
        WHEN store_city_id = 4 THEN 'LUAR PULAU JAWA'
        WHEN store_city_id = 5 THEN 'LUAR PULAU JAWA'
        WHEN store_city_id = 7 THEN 'LUAR PULAU JAWA'
        ELSE 'LUAR NEGRI'
   END AS region
FROM sales
GROUP BY store_city_id
```

Then the display appears as below:

Data output | Messages | Notifications

	total_pendapatan bigint	region text
1	15058000	PULAU JAWA
2	15357000	LUAR PULAU JAWA
3	14828000	PULAU JAWA
4	5979000	PULAU JAWA
5	9076000	LUAR PULAU JAWA
6	15752000	LUAR PULAU JAWA
7	14055000	PULAU JAWA

Total rows: 10 of 10 | Query complete 00:00:02.447

That’s all. Thank You!
Don’t forget to Follow me!

Instagram : @rprabaningrum
LinkedIn : Retno Prabaningrum
Github : Retno Prabaningrum

Intensive Bootcamp Data Science
by @myskill.id