SQL FOR DATA ANALYSIS

انا دلوقتى هشتغل على داتا CSV_file حملته من على الانترنت و حطيته عندى على الجهاز عشان استخدمه و اعمل عليه Analysis و اطلع المعلومات اللى المدير طالبها منى على هيئه Dashboard او Report مثلا

Life cycle of Data Analysis :

1- Source Systems:

ده المكان اللى حملت منه الداتا من الانترنت و اللى هو هنا ال Web Page

2- Data Warehouse:

ده المكان اللى حطيت فيه الداتا على postgres بعد ما حملتها

3- **Queries and Analysis:**

ده بقى الجزء بتاع ال Data analyst

4- Presentation

ال dashboard او ال report النهائى

QUERIES AND ANALYSIS HAVE 5 STEPS:

1- Exploration:

دلوقتى انا معايا الداتا بس معرفش هى عباره عن ايه يعنى معرفش الداتا دى بتتكلم عن ايه ماهو اكيد اسم الفايل من برا مش هيعرفنى الداتا كلها من جوا فأنا لازم اعرف كل حاجه عن الداتا علشان اعرف هعمل analysis ازاى و هطلع منها ايه يفيد ال business حتى لو اضطريت اروح للى مسجل الداتا دى المهم انى اعرف كل column فى الدانتا دى يمثل ايه و ده اللى اسمه ال **Data Dictionary** يعنى اترجم من ال column name لل business name

2- Profiling:

بعد ما عرفت الداتا بتتكلم عن ايه و فهمت كل column بيمثل ايه بالظبط

لازم اعرف التوزيع بتاع الداتا دى ازاى علشان لما اعملها analysis ابقى فاهم هعمل ايه بالظبط لان مثلا الداتا لو distributed normally غير لما تبقى skewed اكيد كل ده هيفرق معايا حتى لما اجى اعمل normalization للارقام فى مرحله الML

يعنى من الاخر مثلا ايه ال columns اللى تهمنى اكتر طب و هتعامل معاها ازاى يعنى الحاجات اللى فيها nulls كتيره اوى دى؟ فلازم يبقى فى standard امشى بيه

3- Cleaning:

نضافه الداتا مهمه اوى يعنى مثلا اشيل ال special characters اللى فى الstrings زى !@#$%<>|+_

و ال null دى بقى هسال فيها ال business مثلا عاوز بداله 0’S ولا اشيلها ولا اسيبها زى ماهى ولا ايه

4- Shape:

شكل الداتا دى بقى بعد ما خلصت اول 3 مراحل و بقيت جاهزه لل analysis و مفيش حاجه توقفنا بطلع بقى شويه statistics منها زى مثلا Max/Min/SD/...etc

5- Analysis

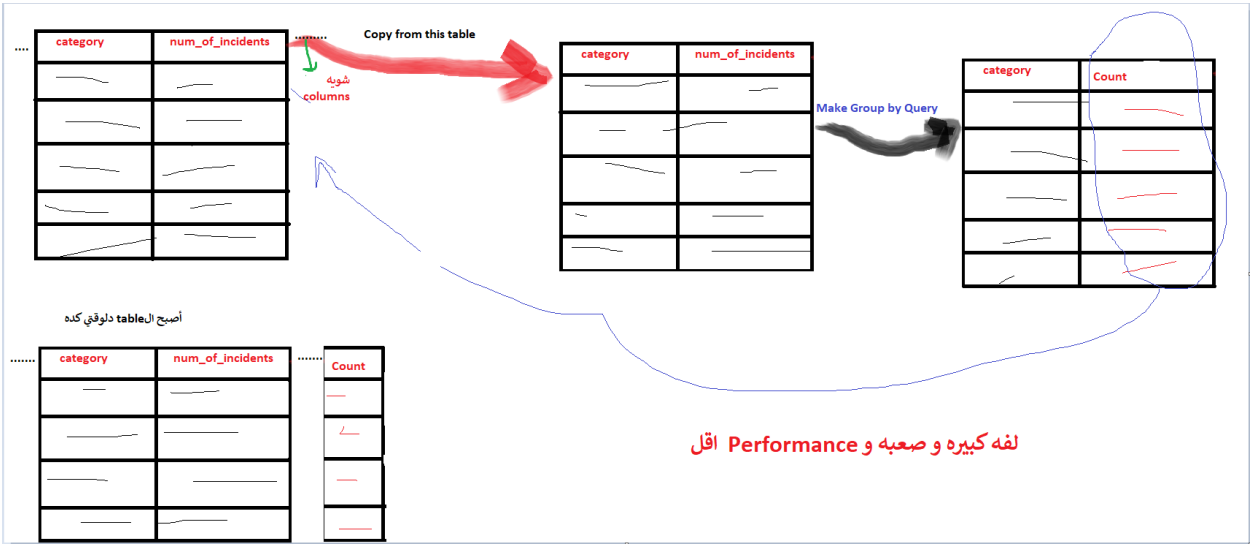
هنا بقى ببدأ اعمل analysis اللى هو اصلا انا مستنيه ..... **دورك جه**

Window Function:

انا دلوقتي لو معايا table كبير في fields كتيره منهم category و number of incidents و انا عاوز اعمل column جديد محطوط فيه ال count بتاع كل category ففي حلول:

1- انا لو عملته بالطريقه العاديه بتاع count و group by فكده هيطلع table واحد مكون من 2 columns بس اللي عاملهم select و هيطلع فعلا column جديد فيه ال count لكن مش ده اللي انا عاوزه ...
انا عاوز اطلع ال table بتاع ال database كللللللللللله و زياده عليه ال column بتاع الcount

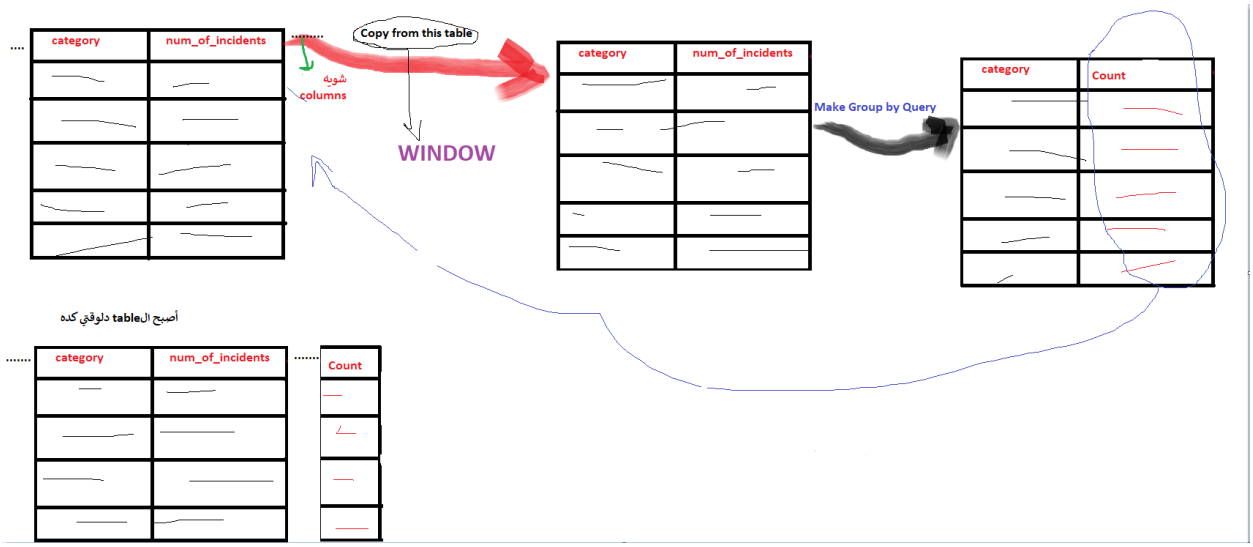
2- الطريقه التانيه و الصعبه



عندي الtable هاخد منه copy و اعمل عليه الquery هيطلعلي الcount اللي انا عاوزه هاخده احطه مع الtable الاصلي

3- في بقي طريقه احسن و انضف و اشيك ألا و هي ال **WINDOW FUNCTION**

ال window دي اللي هي الcopy فتتم ال query او ال(...../max/min/count) aggregation function
علي ال window مش علي الtable الاصلي و النتيجه تضاف directly للtable الاصلي



كل ده بيحصل ب window function اسمها Over()

Over()→ take copy from all table and do the aggregate function that you need without group by and the result will concatenated to the original table directly.

Over(Partition by + column name**)→** same Over() but the window will be copy of the specified column

-- take a window of the full table

هنا هيجيب عدد كل category مقارنه بكل ال fields اللي في الtable

```
Select category, pddistrict, descript, count(incidentnum) over() from police_incident_reports
```

Data Output | Messages | Graph Visualiser | Notifications

	category character varying (50)	pddistrict character varying (10)	descript character varying (100)	count bigint
1	ROBBERY	INGLESIDE	ROBBERY, BODILY FORCE	2129525
2	VEHICLE THEFT	PARK	STOLEN AUTOMOBILE	2129525
3	VEHICLE THEFT	SOUTHERN	STOLEN AUTOMOBILE	2129525
4	ARSON	INGLESIDE	ARSON	2129525
5	ASSAULT	SOUTHERN	BATTERY	2129525
6	ASSAULT	TARAVAL	BATTERY	2129525
7	ASSAULT	SOUTHERN	BATTERY	2129525
8	VEHICLE THEFT	TARAVAL	STOLEN AND RECOVERED VEHICLE	2129525

-- change the scope of the window to the category alone

انما هنا هيجيب عدد كل category موجوده في category column و هيبقي مش شايف باقي الcolumns

```
Select category, pddistrict, descript, count(incidentnum) over(partition by category)
from police_incident_reports
```

Data Output | Messages | Graph Visualiser | Notifications

	category character varying (50)	pddistrict character varying (10)	descript character varying (100)	count bigint
1	ARSON	INGLESIDE	ARSON OF AN INHABITED DWELLING	3875
2	ARSON	BAYVIEW	ARSON OF A VEHICLE	3875
3	ARSON	RICHMOND	ARSON OF A VEHICLE	3875
4	ARSON	BAYVIEW	ARSON OF A VEHICLE	3875
5	ARSON	CENTRAL	ARSON	3875
6	ARSON	INGLESIDE	ARSON OF A VEHICLE	3875
7	ARSON	MISSION	ARSON OF A VEHICLE	3875
8	ARSON	BAYVIEW	ARSON OF A VEHICLE	3875

بالنسبه لتأثير ال order by بقي هنا هيبقي عباره عن running counter كده بيزيد كل ما يشوف الcategory يعني لو عندي category اسمها Arson هتبقي 1 لما يشوفها تاني هتبقي 2 اللي بعدها 3 و هكذا لحد ما اول Arson تخلص و تيجي category تانيه مثلا اسمها Assault فهيصفر العداد و يبدا من الاول

-- running count of pd_id by incident category

```
SELECT pd_id, category, count(pd_id) over(partition by category order by pd_id) from
public.police_incident_reports;
```

Data Output | Messages | Notifications

	pd_id bigint	category character varying (50)	count bigint
1	3000333126029	ARSON	1
2	3001877226035	ARSON	2
3	3001915626030	ARSON	3
4	3003375826030	ARSON	4
5	3003850426031	ARSON	5
6	3004695526029	ARSON	6
7	3004853126031	ARSON	7
8	3005328926029	ARSON	8
9	3005778626031	ARSON	9
10	3005796826029	ARSON	10

Data Output | Messages | Notifications

	pd_id bigint	category character varying (50)	count bigint
3874	18033365026029	ARSON	3874
3875	18034361526030	ARSON	3875
3876	303280004134	ASSAULT	1
3877	705421604014	ASSAULT	2
3878	725956304134	ASSAULT	3
3879	1001501104134	ASSAULT	4
3880	1001501119057	ASSAULT	5
3881	1701797504134	ASSAULT	6
3882	2161221204013	ASSAULT	7
3883	3000018304154	ASSAULT	8

أكيد لاحظت صح؟!

Row_number()→ it's equal to count()

SELECT pd_id, category, **count(pd_id)** over(partition by category order by pd_id) from public.police_incident_reports;

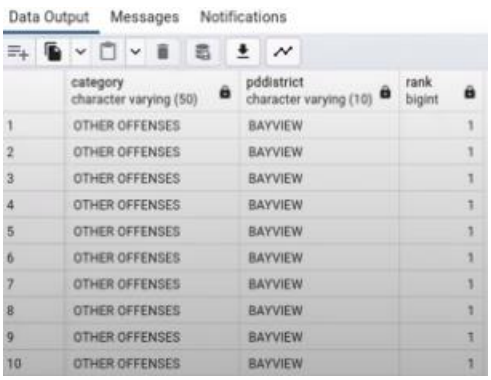
Equals to

select category, pd_id, date, **row_number()** over(partition by category) from police_incident_reports

Rank()→ give rank to the selected values

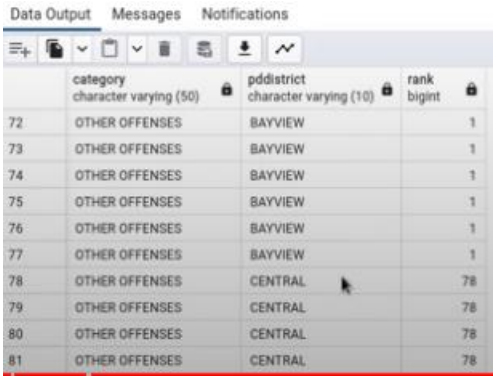
يعني بص كده

select category, pddistrict, **rank()** over(partition by "Incident Code" order by pddistrict) from police_incident_reports



Data Output | Messages | Notifications

	category character varying (50)	pddistrict character varying (10)	rank bigint
1	OTHER OFFENSES	BAYVIEW	1
2	OTHER OFFENSES	BAYVIEW	1
3	OTHER OFFENSES	BAYVIEW	1
4	OTHER OFFENSES	BAYVIEW	1
5	OTHER OFFENSES	BAYVIEW	1
6	OTHER OFFENSES	BAYVIEW	1
7	OTHER OFFENSES	BAYVIEW	1
8	OTHER OFFENSES	BAYVIEW	1
9	OTHER OFFENSES	BAYVIEW	1
10	OTHER OFFENSES	BAYVIEW	1



Data Output | Messages | Notifications

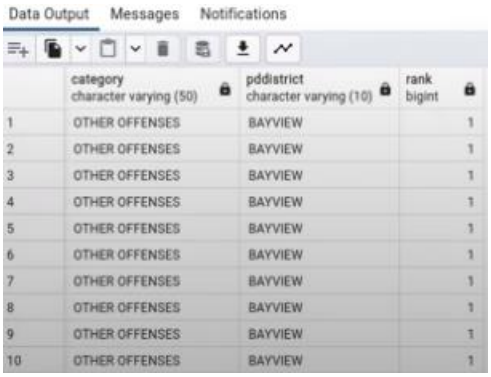
	category character varying (50)	pddistrict character varying (10)	rank bigint
72	OTHER OFFENSES	BAYVIEW	1
73	OTHER OFFENSES	BAYVIEW	1
74	OTHER OFFENSES	BAYVIEW	1
75	OTHER OFFENSES	BAYVIEW	1
76	OTHER OFFENSES	BAYVIEW	1
77	OTHER OFFENSES	BAYVIEW	1
78	OTHER OFFENSES	CENTRAL	78
79	OTHER OFFENSES	CENTRAL	78
80	OTHER OFFENSES	CENTRAL	78
81	OTHER OFFENSES	CENTRAL	78

الصوره اللي علي الشمال هنلاقي ان مفيش حاجه تميز ال rows عن بعضها فمفيش داعي ان ال value تتغير اصلا

الصوره اللي علي اليمين هنلاقي ان اول value تتغير اتغير معاها ال rank علطول بس هنا بقي بعد الواحد خد 78 علطول احتسابا ان في 77 row قبل كده و ده ال78

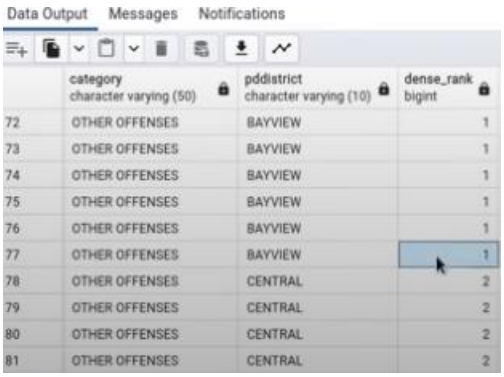
بس احنا اتعلمنا ان العد بيبقي .....-1-2-3-4 كده احنا هنروح ل Another Function بتعد العد بتاعنا

نفس فكره الاول و الاول مكرر بالظبط لما بيخلص بيجي التاني و التاني مكرر و هكذا ←Dense_Rank()



Data Output | Messages | Notifications

	category character varying (50)	pddistrict character varying (10)	rank bigint
1	OTHER OFFENSES	BAYVIEW	1
2	OTHER OFFENSES	BAYVIEW	1
3	OTHER OFFENSES	BAYVIEW	1
4	OTHER OFFENSES	BAYVIEW	1
5	OTHER OFFENSES	BAYVIEW	1
6	OTHER OFFENSES	BAYVIEW	1
7	OTHER OFFENSES	BAYVIEW	1
8	OTHER OFFENSES	BAYVIEW	1
9	OTHER OFFENSES	BAYVIEW	1
10	OTHER OFFENSES	BAYVIEW	1



Data Output | Messages | Notifications

	category character varying (50)	pddistrict character varying (10)	dense_rank bigint
72	OTHER OFFENSES	BAYVIEW	1
73	OTHER OFFENSES	BAYVIEW	1
74	OTHER OFFENSES	BAYVIEW	1
75	OTHER OFFENSES	BAYVIEW	1
76	OTHER OFFENSES	BAYVIEW	1
77	OTHER OFFENSES	BAYVIEW	1
78	OTHER OFFENSES	CENTRAL	2
79	OTHER OFFENSES	CENTRAL	2
80	OTHER OFFENSES	CENTRAL	2
81	OTHER OFFENSES	CENTRAL	2

الصوره اللي علي الشمال نفس ال () rank العاديه

الصوره اللي علي اليمين هنلاقي ان اول value تتغير اتغير معاها ال rank علطول بس هنا بقي بعد الواحد خد 2

Example:

```
select first_name,last_name, department_id, salary,
rank () over(partition by department_id order by salary desc) as rank ,
dense_rank () over(partition by department_id order by salary desc) as dense_rank
```

100 %

Results | Messages

	first_name	last_name	department_id	salary	rank	dense_rank
1	Shelley	Higgins	Accounting	12000.00	1	1
2	William	Gietz	Accounting	8300.00	2	2
3	Michael	Hartstein	Accounting Manager	13000.00	1	1
4	Pat	Fay	Accounting Manager	6000.00	2	2
5	Den	Raphaely	Administration Assistant	11000.00	1	1
6	Alexander	Khoo	Administration Assistant	3100.00	2	2
7	Shelli	Baida	Administration Assistant	2900.00	3	3
8	Sigal	Tobias	Administration Assistant	2800.00	4	4
9	Guy	Himuro	Administration Assistant	2600.00	5	5
10	Karen	Colmenares	Administration Assistant	2500.00	6	6
11	Steven	King	Executive	24000.00	1	1
12	Neena	Kochhar	Executive	17000.00	2	2
13	Lex	De Haan	Executive	17000.00	2	2
14	Nancy	Greenberg	Finance	12000.00	1	1
15	Daniel	Faviet	Finance	9000.00	2	2
16	John	Chen	Finance	8200.00	3	3
17	Jose Manuel	Urman	Finance	7800.00	4	4
18	Ismael	Sciarra	Finance	7700.00	5	5
19	Luis	Popp	Finance	6900.00	6	6
20	Alexander	Hunold	IT	9000.00	1	1
21	Bruce	Ernst	IT	6000.00	2	2
22	David	Austin	IT	4800.00	3	3
23	Valli	Pataballa	IT	4800.00	3	3

هنا انا بقوله اعمل window علي ال department_id يعني امسك كل department و هات ال salary بتاعه و رتبهم بناءا علي ال salary

Lag()→

بتقولي اللي قبله في ال rank كان ايه او كان كام

Lead()→

بتقولي اللي بعده في ال rank ايه او كام

Example:

```
select first_name,last_name, department_id, salary,
lead (salary,1) over(partition by department_id order by salary desc) as lead ,
lag (salary,1) over(partition by department_id order by salary desc) as lag
from employees ;
```

100 %

Results | Messages

	first_name	last_name	department_id	salary	lead	lag
1	Shelley	Higgins	Accounting	12000.00	8300.00	NULL
2	William	Gietz	Accounting	8300.00	NULL	12000.00
3	Michael	Hartstein	Accounting Manager	13000.00	6000.00	NULL
4	Pat	Fay	Accounting Manager	6000.00	NULL	13000.00
5	Den	Raphaely	Administration Assistant	11000.00	3100.00	NULL
6	Alexander	Khoo	Administration Assistant	3100.00	2900.00	11000.00
7	Shelli	Baida	Administration Assistant	2900.00	2800.00	3100.00
8	Sigal	Tobias	Administration Assistant	2800.00	2600.00	2900.00
9	Guy	Himuro	Administration Assistant	2600.00	2500.00	2800.00
10	Karen	Colmenares	Administration Assistant	2500.00	NULL	2600.00
11	Steven	King	Executive	24000.00	17000.00	NULL
12	Neena	Kochhar	Executive	17000.00	17000.00	24000.00
13	Lex	De Haan	Executive	17000.00	NULL	17000.00
14	Nancy	Greenberg	Finance	12000.00	9000.00	NULL
15	Daniel	Faviet	Finance	9000.00	8200.00	12000.00
16	John	Chen	Finance	8200.00	7800.00	9000.00
17	Jose Manuel	Urman	Finance	7800.00	7700.00	8200.00
18	Ismael	Sciarra	Finance	7700.00	6900.00	7800.00
19	Luis	Popp	Finance	6900.00	NULL	7700.00
20	Alexander	Hunold	IT	9000.00	6000.00	NULL
21	Bruce	Ernst	IT	6000.00	4800.00	9000.00
22	David	Austin	IT	4800.00	4800.00	6000.00
23	Valli	Pataballa	IT	4800.00	4200.00	4800.00

طب و ايه Null اللي موجوده دي؟

مش احنا عاملين الwindow علي ال department_id ف دي معناها ان في lead مفيش حد ال rank بتاعه اعلي من ال value اللي جنب null و ليكن اول null يعني مفيش في الAccounting حد السالري بتاعه اعلي من 8300

و معناها ان في lag مفيش حد ال rank بتاعه اقل من ال value اللي جنب null و ليكن اول null يعني مفيش في الAccounting حد السالري بتاعه اقل من 12000

Ntile()→

Its an abbreviation to percentage tile , it’s a statistical function

دي بتقسم الحاجه ل buckets او ranges

Example:

```
select first_name,last_name, department_id, salary,
ntile(4) over(partition by department_id order by salary desc) as quartile
from employees
order by
department_id, quartile;
```

100 %

Results | Messages

	first_name	last_name	department_id	salary	quartile
1	Shelley	Higgins	Accounting	12000.00	1
2	William	Gietz	Accounting	8300.00	2
3	Michael	Hartstein	Accounting Manager	13000.00	1
4	Pat	Fay	Accounting Manager	6000.00	2
5	Den	Raphaely	Administration Assistant	11000.00	1
6	Alexander	Khoo	Administration Assistant	3100.00	1
7	Shelli	Baida	Administration Assistant	2900.00	2
8	Sigal	Tobias	Administration Assistant	2800.00	2
9	Guy	Himuro	Administration Assistant	2600.00	3
10	Karen	Colmenares	Administration Assistant	2500.00	4
11	Steven	King	Executive	24000.00	1
12	Neena	Kochhar	Executive	17000.00	2
13	Lex	De Haan	Executive	17000.00	3
14	Nancy	Greenberg	Finance	12000.00	1
15	Daniel	Faviet	Finance	9000.00	1
16	John	Chen	Finance	8200.00	2
17	Jose Manuel	Urman	Finance	7800.00	2
18	Ismael	Sciarra	Finance	7700.00	3
19	Luis	Popp	Finance	6900.00	4
20	Alexander	Hunold	IT	9000.00	1
21	Bruce	Ernst	IT	6000.00	1
22	David	Austin	IT	4800.00	2
23	Valli	Pataballa	IT	4800.00	3

هنا ال window علي الdepartment_id يعني بيمسك كل department_id و يقسم الsalary اللي جواه ل 4 ranges