SQL DATA ANALYSIS PROJECT

In this project. We will use SQL to analyse our data. The data is from the Northwind Database. Let's begin.

In [2]:
```
%load_ext sql
```

```
The sql extension is already loaded. To reload it, use:
  %reload_ext sql
```

In [4]:
```
%sql postgresql://postgres:Jjake247!@localhost:5432/northwind
```

Out[4]:
```
'Connected: postgres@northwind'
```

We have connected our notebook to the northwind database. We will now begin our analysis by checking the tables in our database. For our analysis we will be using the postgresql relational database management system.

In [6]:
```
%%sql
SELECT table_name
FROM information_schema.tables
WHERE table_schema = 'public';
```

```
 * postgresql://postgres:***@localhost:5432/northwind
14 rows affected.
```

Out[6]:

table_name
customer_demographics
customer_customer_demo
customers
employees
categories
products
suppliers
orders
shippers
region
territories
employee_territories
order_details
us_states

We will seek to use SQL to answer various questions regarding our data.

First, we'll begin by checking the number of customers we have in our customers table in the database.

In [ ]:
```
%%sql
select count(*)
from customers;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
1 rows affected.
```

Out[ ]:

count
91

We have seen we have 91 customers in the customers table.

In [11]:
```
%%sql
select *
from customers
limit 5;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
5 rows affected.
```

Out[11]:

customer_id	company_name	contact_name	contact_title	address	city	region	postal_c
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin	None	1
ANATR	Ana Trujillo Emparedados y helados	Ana Trujillo	Owner	Avda. de la Constitución 2222	México D.F.	None	0
ANTON	Antonio Moreno Taquería	Antonio Moreno	Owner	Mataderos 2312	México D.F.	None	0
AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London	None	WA1
BERGS	Berglunds snabbköp	Christina Berglund	Order Administrator	Berguvsvägen 8	Luleå	None	S-95

From the query above we see that the customers table has various fields including the company they work for, their address, city and their country. We can group our data by country so that we can see the countries were most of the customers are coming from.

In [15]:
```
%%sql
select country, count(*)
from customers
group by country
order by count DESC
limit 5;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
5 rows affected.
```

Out[15]:

country	count
USA	13
France	11
Germany	11
Brazil	9
UK	7

From the line of code above we see the countries that make up the top 5. USA is leading with the most customers and France and Germany are tied for second.

Next, we'll calaculate the subtotals for each of the orders placed by the customers. To do this we'll use the order_details table.

In [30]:
```
%%sql
select *
from order_details
limit 5;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
5 rows affected.
```

Out[30]:

order_id	product_id	unit_price	quantity	discount
10248	11	14.0	12	0.0
10248	42	9.8	10	0.0
10248	72	34.8	5	0.0
10249	14	18.6	9	0.0
10249	51	42.4	40	0.0

In [37]:
```
%%sql
select  order_id,
        sum(unit_price * quantity * (1-discount)) as subtotal
from order_details
group by order_id
order by subtotal DESC
limit 10;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
10 rows affected.
```

Out[37]:

order_id	subtotal
10865	16387.49998714775
10981	15810.0
11030	12615.050067901611
10889	11380.0
10417	11188.400139808655
10817	10952.84492739618
10897	10835.240051269531
10479	10495.60012435913
10540	10191.699981689453
10691	10164.800018310547

From the query above we get the 10 most expensive orders.

In [39]:
```
%%sql
select *
from orders
limit 5;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
5 rows affected.
```

Out[39]:

order_id	customer_id	employee_id	order_date	required_date	shipped_date	ship_via	freight	s
10248	VINET	5	1996-07-04	1996-08-01	1996-07-16	3	32.38	
10249	TOMSP	6	1996-07-05	1996-08-16	1996-07-10	1	11.61	Sp
10250	HANAR	4	1996-07-08	1996-08-05	1996-07-12	2	65.83	
10251	VICTE	3	1996-07-08	1996-08-05	1996-07-15	1	41.34	
10252	SUPRD	4	1996-07-09	1996-08-06	1996-07-11	2	51.3	

In order to get the contact name of the customers who placed the various orders, we'll need to join three tables. The customers tables, the orders table and the order_details table.

In [47]:
```
%%sql
select  c.contact_name,
        od.order_id,
        o.order_date,
        sum(od.unit_price * od.quantity * (1-od.discount)) as order_total
from customers as c
join orders as o
on c.customer_id = o.customer_id
join order_details as od
on o.order_id = od.order_id
```

```
group by c.contact_name, od.order_id, o.order_date
limit 10;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
10 rows affected.
```

Out[47]:

contact_name	order_id	order_date	order_total
Michael Holz	10419	1997-01-20	2097.6000273466107
José Pedro Freyre	10872	1998-02-05	2058.4599686691163
Pascale Cartrain	10463	1997-03-04	713.299994468689
Pirkko Koskitalo	11025	1998-04-15	269.99999955296516
Annette Roulet	10832	1998-01-14	475.1100083673
Victoria Ashworth	10484	1997-03-24	386.2000026702881
Mary Saveley	10546	1997-05-23	2811.999969482422
Paolo Accorti	10710	1997-10-20	93.49999904632568
Helen Bennett	10318	1996-10-01	240.39999389648438
Philip Cramer	10542	1997-05-20	469.1099952833355

From the query above we have used joins to find the customer name, the order date and the total of their order.

We can also find the customers who placed their orders first. To do this we will use the 'order by' clause on the order_date column.

In [50]:

```
%%sql
select  c.contact_name,
        od.order_id,
        o.order_date,
        sum(od.unit_price * od.quantity * (1-od.discount)) as order_total
from customers as c
join orders as o
on c.customer_id = o.customer_id
join order_details as od
on o.order_id = od.order_id
group by c.contact_name, od.order_id, o.order_date
order by o.order_date asc
limit 10;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
10 rows affected.
```

Out[50]:

contact_name	order_id	order_date	order_total
Paul Henriot	10248	1996-07-04	439.99999809265137
Karin Josephs	10249	1996-07-05	1863.4000644683838
Mary Saveley	10251	1996-07-08	654.0599855789542
Mario Pontes	10250	1996-07-08	1552.600023412704
Pascale Cartrain	10252	1996-07-09	3597.9001445159315
Mario Pontes	10253	1996-07-10	1444.7999839782715
Yang Wang	10254	1996-07-11	556.62000967741
Michael Holz	10255	1996-07-12	2490.4999780654907
Paula Parente	10256	1996-07-15	517.8000068664551
Carlos Hernández	10257	1996-07-16	1119.899953842163

In [59]:
```
%%sql
select  extract(year from shipped_date) as shipped_year,
        count(order_id)
from orders
group by shipped_year
order by count DESC;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
4 rows affected.
```

Out[59]:

shipped_year	count
1997	398
1998	268
1996	143
None	21

From the query above we see the year that was busiest in terms of shipment was 1997.

We can also calculate the total amount of sales for each product. To do this we will need to utilise joins and also to utilise nested subqueries.

In [68]:
```
%%sql
select *
from products
limit 5;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
5 rows affected.
```

Out[68]:

product_id	product_name	supplier_id	category_id	quantity_per_unit	unit_price	units_in_stock	u
1	Chai	8	1	10 boxes x 30 bags	18.0	39	
2	Chang	1	1	24 - 12 oz bottles	19.0	17	
3	Aniseed Syrup	1	2	12 - 550 ml bottles	10.0	13	
4	Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22.0	53	
5	Chef Anton's Gumbo Mix	2	2	36 boxes	21.35	0	

In [71]:
```
%%sql
select *
from suppliers
order by company_name
limit 3
```

```
 * postgresql://postgres:***@localhost:5432/northwind
3 rows affected.
```

Out[71]:

supplier_id	company_name	contact_name	contact_title	address	city	region	postal_code
18	Aux joyeux ecclésiastiques	Guylène Nodier	Sales Manager	203, Rue des Francs-Bourgeois	Paris	None	75004
16	Bigfoot Breweries	Cheryl Saylor	Regional Account Rep.	3400 - 8th Avenue Suite 210	Bend	OR	97101
5	Cooperativa de Quesos 'Las Cabras'	Antonio del Valle Saavedra	Export Administrator	Calle del Rosal 4	Oviedo	Asturias	33007

In [83]:
```
%%sql
select  p.product_name,
        s.company_name,
        sum(od.unit_price * od.quantity * (1-od.discount)) as total_sales
from products p
join suppliers s
on p.supplier_id = s.supplier_id
join order_details od
on p.product_id = od.product_id
group by p.product_name, s.company_name
order by total_sales desc
limit 5;
```

```
 * postgresql://postgres:***@localhost:5432/northwind
5 rows affected.
```

Out[83]:

product_name	company_name	total_sales
Côte de Blaye	Aux joyeux ecclésiastiques	141396.7356273254
Thüringer Rostbratwurst	Plutzer Lebensmittelgroßmärkte AG	80368.6724385033
Raclette Courdavault	Gai pâturage	71155.69990943
Tarte au sucre	Forêts d'érables	47234.969978504174
Camembert Pierrot	Gai pâturage	46825.48029542655

From the query above we see that Cote de Blaye is the best-selling product.

SUMMARY

In this project, we used SQL to answer various questions regarding the database. Some of the questions we answered include:

1. How many customers are there?
2. Which countries have the most customers?
3. Which customers placed the most orders?
4. Which customers placed the most expensive orders?
5. Who placed their orders earliest?
6. Which is the best selling product?