PROJECT: ANALYZING AND FORMATTING POSTGRESQL SALES DATA

●

Cleaning a PostgreSQL Database

In this project, you will work with data from a hypothetical Super Store to challenge and enhance your SQL skills in data cleaning. This project will engage you in identifying top categories based on the highest profit margins and detecting missing values, utilizing your comprehensive knowledge of SQL concepts.

Data Dictionary:

`orders` :

Column	Definition	Data type	Comments
`row_id`	Unique Record ID	`INTEGER`	
`order_id`	Identifier for each order in table	`TEXT`	Connects to `order_id` in `returned_orders` table
`order_date`	Date when order was placed	`TEXT`	
`market`	Market order_id belongs to	`TEXT`	
`region`	Region Customer belongs to	`TEXT`	Connects to `region` in `people` table
`product_id`	Identifier of Product bought	`TEXT`	Connects to `product_id` in `products` table
`sales`	Total Sales Amount for the Line Item	`DOUBLE PRECISION`	
`quantity`	Total Quantity for the Line Item	`DOUBLE PRECISION`	
`discount`	Discount applied for the Line Item	`DOUBLE PRECISION`	
`profit`	Total Profit earned on the Line Item	`DOUBLE PRECISION`	

`returned_orders` :

Column	Definition	Data type
`returned`	Yes values for Order / Line Item Returned	`TEXT`
`order_id`	Identifier for each order in table	`TEXT`
`market`	Market order_id belongs to	`TEXT`

`people` :

Column	Definition	Data type
`person`	Name of Salesperson credited with Order	`TEXT`
`region`	Region Salesperson in operating in	`TEXT`

`products` :

Column	Definition	Data type
`product_id`	Unique Identifier for the Product	`TEXT`
`category`	Category Product belongs to	`TEXT`

Column	Definition	Data type
sub_category	Sub Category Product belongs to	TEXT
product_name	Detailed Name of the Product	TEXT

As you can see in the Data Dictionary above, date fields have been written to the `orders` table as `TEXT` and numeric fields like sales, profit, etc. have been written to the `orders` table as `Double Precision`. You will need to take care of these types in some of the queries. This project is an excellent opportunity to apply your SQL skills in a practical setting and gain valuable experience in data cleaning and analysis. Good luck, and happy querying!

Projects Data DataFrame as top_five_products_each_category

```
-- top_five_products_each_category
WITH cte AS (
    SELECT
    p.category,
    p.product_name,
    ROUND(SUM(o.sales) :: numeric,2) AS product_total_sales,
    ROUND(SUM(o.profit) :: numeric,2) AS product_total_profit

    FROM public.orders AS o
    LEFT JOIN public.products AS p
    USING(product_id)
    GROUP BY p.category,p.product_name
),
ranked AS (
    SELECT * ,
     ROW_NUMBER() OVER(PARTITION BY category ORDER BY product_total_sales DESC ) AS product_rank
    FROM cte
)
SELECT *
FROM ranked
WHERE product_rank BETWEEN 1 AND 5
ORDER BY category
```

	category	product_name	product_total_sa...	product_total_profit	prod...
0	Furniture	Hon Executive Leather Armchair, Adjustable	58193.48	5997.25	
1	Furniture	Office Star Executive Leather Armchair, Adju...	51449.8	4925.8	
2	Furniture	Harbour Creations Executive Leather Armch...	50121.52	10427.33	
3	Furniture	SAFCO Executive Leather Armchair, Black	41923.53	7154.28	
4	Furniture	Novimex Executive Leather Armchair, Adjust...	40585.13	5562.35	
5	Office Supplies	Eldon File Cart, Single Width	39873.23	5571.26	
6	Office Supplies	Hoover Stove, White	32842.6	-2180.63	
7	Office Supplies	Hoover Stove, Red	32644.13	11651.68	
8	Office Supplies	Rogers File Cart, Single Width	29558.82	2368.82	
9	Office Supplies	Smead Lockers, Industrial	28991.66	3630.44	
10	Technology	Apple Smart Phone, Full Size	86935.78	5921.58	
11	Technology	Cisco Smart Phone, Full Size	76441.53	17238.52	
12	Technology	Motorola Smart Phone, Full Size	73156.3	17027.11	
13	Technology	Nokia Smart Phone, Full Size	71904.56	9938.2	
14	Technology	Canon imageCLASS 2200 Advanced Copier	61599.82	25199.93	

Rows: 15 Expand

Projects Data DataFrame as i

```
-- impute_missing_values
WITH unit_prices AS (
    SELECT
        market,
        region,
        product_id,
        AVG(sales/(quantity*(1-discount))) AS unit_price
    FROM orders
    WHERE quantity IS NOT NULL
    GROUP BY market,region,product_id
)
SELECT
    product_id,
    discount,
    market,
    region,
    sales,
    quantity,
    sales/(unit_price *(1-discount)) AS calculated_quantity
FROM orders
INNER JOIN unit_prices
USING(market,region,product_id)
WHERE quantity IS NULL;
```

	product_id						calculated_quan...
0	TEC-STA-10003330	0	Africa	Africa	506.64		2
1	FUR-ADV-10000571	0	EMEA	EMEA	438.96		4
2	FUR-BO-10001337	0.15	US	West	308.499		3
3	TEC-STA-10004542	0	Africa	Africa	160.32		4
4	FUR-ADV-10004395	0	EMEA	EMEA	84.12		2

Rows: 5

Expand