SQL PROJECT

. . .

Zuwa Ojefua August, 2023

Transforming and Analyzing Data with SQL

The main goal of the project was Exploratory Data Analysis.

Database Creation

Data Cleaning

Data Analysis

Database Creation

Query:

```
CREATE TABLE public.sales report
  product sku character varying,
  total ordered integer,
  name character varying,
  stock level integer,
  restocking lead time integer,
  sentiment score numeric,
  sentiment magnitude numeric,
  ratio numeric
ALTER TABLE IF EXISTS
public.sales report
  OWNER to postgres;
```

Query:

```
CREATE TABLE public.sales_by_sku
(
    product_sku character varying,
    total_ordered integer
);
ALTER TABLE IF EXISTS
public.sales_by_sku
    OWNER to postgres;
```

```
> 1...3 Sequences
> == all_sessions
  > == analytics
  > == products

▼ Ii Columns (2)
       product_sku
       total_ordered
   > > Constraints
   > 🚠 Indexes
   > 🔓 RLS Policies
   > IIII Rules
   > 🖈 Triggers
 product_sku
       total_ordered
       name
       stock_level
       restocking_lead_time
       sentiment_score
       sentiment_magnitude
       ratio
     ▶ ■ Constraints
```

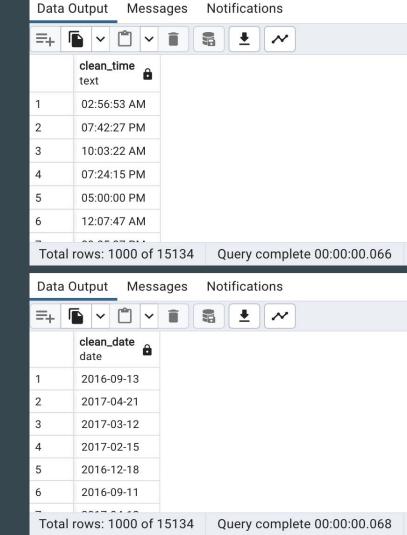
SELECT

TO_CHAR(TO_TIMESTAMP(time), 'HH:MI:SS AM') clean_time FROM all_sessions;

SELECT

DATE(date) AS clean_date

FROM all_sessions;



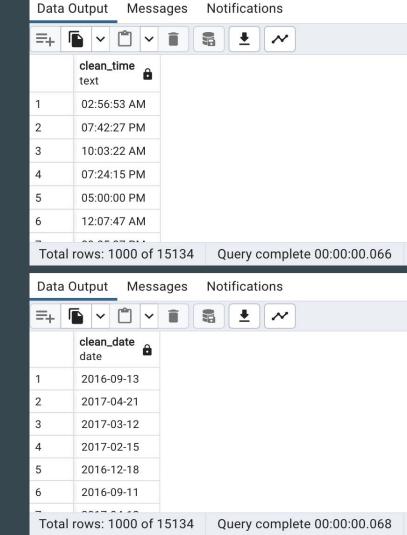
SELECT

TO_CHAR(TO_TIMESTAMP(time), 'HH:MI:SS AM') clean_time FROM all_sessions;

SELECT

DATE(date) AS clean_date

FROM all_sessions;

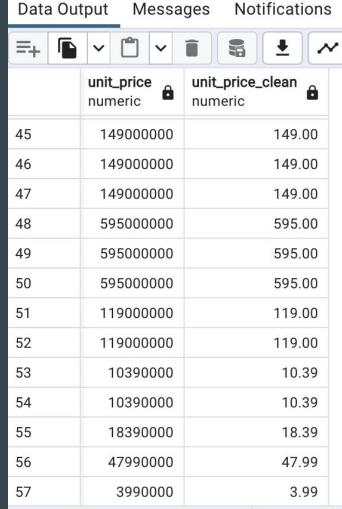


```
SELECT
              country,
              CASE WHEN country IN ('(not set)', 'not available in demo dataset')
                             THEN NULL
                             ELSE country
              END AS clean_country,
              city,
              CASE WHEN city IN ('(not set)', 'not available in demo dataset')
                             THEN (CASE WHEN country NOT IN ('(not set)', 'not available in demo dataset')
                                                  THEN country
                                     END)
                      ELSE COALESCE(city, country)
              END AS clean_city
FROM all_sessions
```

	country character varying	clean_country character varying	city character varying	clean_city character varying
1	Taiwan	Taiwan	(not set)	Taiwan
2	United States	United States	not available in demo dataset	United States
3	United States	United States	not available in demo dataset	United States
4	United States	United States	not available in demo dataset	United States
5	United States	United States	London	London
6	(not set)	[null]	(not set)	[null]
7	El Salvador	El Salvador	not available in demo dataset	El Salvador
8	United States	United States	not available in demo dataset	United States
9	United Kingdom	United Kingdom	not available in demo dataset	United Kingdom
10	Australia	Australia	Sydney	Sydney
11	United States	United States	Philadelphia	Philadelphia
12	Serbia	Serbia	not available in demo dataset	Serbia
13	Canada	Canada	not available in demo dataset	Canada

Total rows: 1000 of 15134 Query complete 00:00:00.345

unit_price **SELECT** numeric unit price, ROUND((unit price / 1000000), 2) unit price clean 45 149000000 FROM analytics 46 WHERE unit_price IS NOT NULL 149000000 47 149000000 48 595000000 49 595000000 50 595000000 51 119000000



Total rows: 1000 of 4301122 Query complete 00:00:03.596

Data Analysis: Revenue across regions

```
WITH clean cte AS
       SELECT
              CASE WHEN country IN ('(not set)', 'not available in demo dataset')
                            THEN NULL
                            ELSE country
              END AS country,
              CASE WHEN city IN ('(not set)', 'not available in demo dataset')
                            THEN (CASE WHEN country NOT IN ('(not set)', 'not available in demo dataset')
                                                  THEN country
                                    END)
                     ELSE COALESCE(city, country)
              END AS city,
              ROUND((total transaction revenue / 1000000), 2) total transaction revenue
       FROM all sessions
       WHERE total transaction revenue IS NOT NULL
SELECT
       country,
       city,
       MAX(total transaction revenue) max total revenue
FROM clean cte
GROUP BY country, city
ORDER BY max total revenue DESC
```

Data Analysis: Revenue across regions

Data Output Messages Notifications				
	country character varying	city character varying	max_total_revenue numeric	
1	United States	United States	1015.48	
2	United States	Atlanta	742.48	
3	United States	Sunnyvale	649.24	
4	Israel	Tel Aviv-Yafo	602.00	
5	United States	Los Angeles	363.00	
6	Australia	Sydney	358.00	
7	United States	Seattle	358.00	
8	United States	Chicago	306.00	
9	United States	Palo Alto	305.00	
10	United States	San Francisco	301.00	
11	United States	Nashville	157.00	
12	United States	Mountain View	156.00	
13	United States	San Jose	154.00	
Total rows: 21 of 21 Query complete 00:00:00.496				

Data Analysis: Top-selling product across regions

```
WITH all sessions cte AS
       SELECT
               CASE WHEN country IN ('(not set)', 'not available in demo dataset')
                       THEN NULL
                       ELSE country
               END AS country,
               CASE WHEN city IN ('(not set)', 'not available in demo dataset')
                               THEN (CASE WHEN country NOT IN ('(not set)', 'not available in demo dataset')
                                                      THEN country
                                       END)
                       ELSE COALESCE(city, country)
               END AS city,
               v2 product name product name,
               COUNT(v2 product name) OVER (PARTITION BY v2 product name) AS product count
       FROM all sessions
       ORDER BY product count DESC, country
SELECT
       country,
       city,
       product name,
       MAX(product count) max sold
FROM all sessions cte
GROUP BY country, city, product name, product count
ORDER BY max sold DESC, country
```

Data Analysis: Top-selling product across regions

Data (Output Messages	Notifications		
=+ 1				
	country character varying	city character varying	product_name character varying	max_sold bigint
77	United States	Sunnyvale	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
78	United States	Austin	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
79	United States	San Francisco	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
80	United States	Los Angeles	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
81	United States	Palo Alto	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
82	Uruguay	Montevideo	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
83	Vietnam	Hanoi	Google Men's 100% Cotton Short Sleeve Hero Tee Whi	295
84	Albania	Albania	22 oz YouTube Bottle Infuser	245
85	Argentina	Argentina	22 oz YouTube Bottle Infuser	245
86	Australia	Australia	22 oz YouTube Bottle Infuser	245
87	Australia	Sydney	22 oz YouTube Bottle Infuser	245
88	Australia	Melbourne	22 oz YouTube Bottle Infuser	245
89	Bangladesh	Bangladesh	22 oz YouTube Bottle Infuser	245
Total rows: 1000 of 7611 Query complete		Query complete 00:	00:00.149	

QA Process

The data was inspected and observed to be fully anonymized as it did not contain any personal user_id information, so there were no data privacy risks associated with the data set.

The data was observed to have many structural issues ranging from column names not stored in acceptable case; duplicate and null entries in different formats; and inconsistent units for time, and financial data.

Overall, the data was observed to contain an unusually high proportion of null entries and inconsistencies so the integrity of the data would not be expected to yield high confidence results after data analysis.

QA Process

UNION ALL

SELECT 'products' as table_name,

COUNT(*) AS count_rows,

COUNT(DISTINCT(product_sku)) AS count_distinct_id

FROM products

UNION ALL

SELECT 'sales_by_sku' as table_name,
 COUNT(*) AS count_rows,
 COUNT(DISTINCT(product_sku)) AS count_distinct_id
FROM sales_by_sku

UNION ALL

UNION ALL

SELECT 'sales_report' as table_name,

COUNT(*) AS count_rows,

COUNT(DISTINCT(product_sku)) AS count_distinct_id

FROM sales_report

Data	Data Output Messages Notifications					
=+						
	table_name text	count_rows bigint	count_distinct_id bigint			
1	all_sessions	15134	14223			
2	products	1092	1092			
3	sales_by_sku	462	462			
4	analytics	4301122	120018			
5	sales_report	454	454			
Tota	l rows: 5 of 5	Query complete 00:00:04.247				

Future work

Review the normalization of the tables.

Find answers to more questions to gain deeper insights into the dataset.

Thank you.