

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

▼ Tables (5)

> all_sessions

> analytics

> products

▼ sales_by_sku

▼ Columns (2)

product_sku

total_ordered

> Constraints

> Indexes

> RLS Policies

> Rules

> Triggers

▼ sales_report

▼ Columns (8)

product_sku

total_ordered

name

stock_level

restocking_lead_time

sentiment_score

sentiment_magnitude

ratio

> Constraints

Data Cleaning

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

```
SELECT
    DATE(date) AS clean_date
FROM all_sessions;
```

	clean_time	
1	02:56:53 AM	
2	07:42:27 PM	
3	10:03:22 AM	
4	07:24:15 PM	
5	05:00:00 PM	
6	12:07:47 AM	
7	02:25:07 PM	

Total rows: 1000 of 15134 Query complete 00:00:00.066

	clean_date	
1	2016-09-13	
2	2017-04-21	
3	2017-03-12	
4	2017-02-15	
5	2016-12-18	
6	2016-09-11	
7	2017-04-10	
Total rows: 1000 of 15134		Query complete 00:00:00.068

Data Cleaning

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

```
SELECT
    DATE(date) AS clean_date
FROM all_sessions;
```

	clean_time	
1	02:56:53 AM	
2	07:42:27 PM	
3	10:03:22 AM	
4	07:24:15 PM	
5	05:00:00 PM	
6	12:07:47 AM	
7	02:25:07 PM	

Total rows: 1000 of 15134 Query complete 00:00:00.066





	clean_date
1	2016-09-13
2	2017-04-21
3	2017-03-12
4	2017-02-15
5	2016-12-18
6	2016-09-11
7	2017-04-10

Total rows: 1000 of 15134 Query complete 00:00:00.068

Data Cleaning

```
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
```

Data Cleaning

	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

Data Cleaning

```
SELECT
    unit_price,
    ROUND((unit_price / 1000000), 2) unit_price_clean
FROM analytics
WHERE unit_price IS NOT NULL
```

Data Output

Messages

Notifications

	<div>unit_price</div> <div>numeric</div> <div></div>	<div>unit_price_clean</div> <div>numeric</div> <div></div>	
45	149000000	149.00	
46	149000000	149.00	
47	149000000	149.00	
48	595000000	595.00	
49	595000000	595.00	
50	595000000	595.00	
51	119000000	119.00	
52	119000000	119.00	
53	10390000	10.39	
54	10390000	10.39	
55	18390000	18.39	
56	47990000	47.99	
57	3990000	3.99	
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
<div><div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div></div><div><div></div><div></div><div></div></div><div><div></div><div></div><div></div></div></div>		<div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div>
	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				
<div><div><div><div></div></div><div><div></div></div><div><div></div></div><div><div></div></div><div><div></div></div><div><div></div></div><div><div></div></div><div><div></div></div><div><div></div></div></div></div>				
	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 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

```
SELECT 'all_sessions' as table_name,  
      COUNT(*) AS count_rows,  
      COUNT(DISTINCT(full_visitor_id)) AS count_distinct_id  
FROM all_sessions
```

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

```
SELECT 'analytics' as table_name,  
      COUNT(*) AS count_rows,  
      COUNT(DISTINCT(full_visitor_id)) AS count_distinct_id  
FROM analytics
```

UNION ALL

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

Data Output

Messages

Notifications

table_name

count_rows

count_distinct_id

text

bigint

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

Total 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.