Data Analysis of Events at The Look E-commerce Platform

Data Source: BigQuery Public Database 'The Look E-commerce'

```sql
-- Number of sessions, number of purchases and conversion rate for each traffic source in 2024
WITH n AS (
SELECT traffic_source
,COUNT(DISTINCT session_id) as n_sessions
,SUM(CASE WHEN event_type = 'purchase' THEN 1 ELSE 0 END) as n_purchases
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01'
GROUP BY traffic_source
)

SELECT traffic_source
,n_sessions
,n_purchases
,ROUND(n_purchases/n_sessions*100, 2) as conv_rate
,RANK() OVER(ORDER BY n_sessions DESC) as n_sessions_rank
FROM n
ORDER BY conv_rate DESC;
```

Row	traffic_source	n_sessions	n_purchases	conv_rate	n_sessions_rank
1	Adwords	24465	14614	59.73	2
2	Facebook	8047	4767	59.24	4
3	Email	36535	21600	59.12	1
4	YouTube	8166	4823	59.06	3
5	Organic	4076	2374	58.24	5

According to information from the www.invespcro.com, “The latest survey and studies in 2023 show that the global average website conversion rate is 3.68%.” So, we have a fantastic conversion result :)
Adwords has the highest conversion rate, while the highest number of sessions occurred from the source of emails. Organic arrivals have the lowest figures.
It is recommended to estimate costs for each traffic source to optimize budgets.

```sql
-- Top 10 states with the largest number of purchases in 2024
SELECT state
,COUNT(*) as n_purchases
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE event_type = 'purchase' AND DATE(created_at) >= '2024-01-01'
GROUP BY state
ORDER BY n_purchases DESC
LIMIT 10;
```

Row	state	n_purchases
1	Guangdong	2659
2	England	1862
3	California	1808
4	Texas	1163
5	Shanghai	1138
6	São Paulo	1026
7	Beijing	1010
8	Zhejiang	1010
9	Hebei	922
10	Gyeonggi-do	881

It is recommended that for states where products are purchased the most, the following aspects be considered:
Assessment of demand and competitors in these states.
Diversification of marketing activities (such as attracting local opinion leaders).
Opportunities to improve delivery terms and costs for buyers.

```sql
-- Top 10 states with the largest number of purchases where the conversion is higher than the overall conversion of the site in 2024
WITH n AS (
SELECT state
,(SELECT ROUND(SUM(CASE WHEN event_type = 'purchase' THEN 1 ELSE 0 END) / COUNT(DISTINCT session_id) * 100, 2)
 FROM `bigquery-public-data.thelook_ecommerce.events`
 WHERE DATE(created_at) >= '2024-01-01') as site_conv
,COUNT(DISTINCT session_id) as n_sessions
,SUM(CASE WHEN event_type = 'purchase' THEN 1 ELSE 0 END) as n_purchases
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01'
GROUP BY state
)
```

```sql
SELECT state
,n_purchases
,ROUND(n_purchases/n_sessions*100, 2) as conv_rate
,site_conv
FROM n
WHERE n_purchases/n_sessions*100 > site_conv
ORDER BY n_purchases DESC
LIMIT 10;
```

Row	state	n_purchases	conv_rate	site_conv
1	Guangdong	2659	60.38	59.27
2	California	1808	59.81	59.27
3	Texas	1163	59.7	59.27
4	São Paulo	1026	59.62	59.27
5	Bahia	715	59.73	59.27
6	Fujian	646	59.93	59.27
7	Shandong	629	59.45	59.27
8	Hunan	598	61.59	59.27
9	Pará	596	59.48	59.27
10	Tokyo	566	61.72	59.27

It is also recommended to estimate the costs of attracting traffic from different states. These states, where the conversion rate exceeds the site's average, should be studied in more detail for potential expansion of the sales market.

```sql
-- Top 3 popular product categories for top 3 states with higher the overall conversion of the site
WITH r AS (SELECT e.state
,COUNT(*) as n_items
,p.category
,RANK() OVER(partition by e.state ORDER BY COUNT(*) DESC) as rank_by_n_items
FROM `bigquery-public-data.thelook_ecommerce.events` as e
JOIN `bigquery-public-data.thelook_ecommerce.order_items` as o
ON e.user_id = o.user_id
JOIN `bigquery-public-data.thelook_ecommerce.products` as p
ON o.product_id = p.id
WHERE DATE(e.created_at) >= '2024-01-01'
AND e.state IN ('Guangdong','California', 'Texas')
AND e.event_type = 'purchase'
GROUP BY e.state, p.category)
```

```sql
SELECT state
,category
,rank_by_n_items
FROM r
WHERE rank_by_n_items <= 3
ORDER BY state, r.n_items DESC;
```

Row	state	category	rank_by_n_items
1	California	Intimates	1
2	California	Jeans	2
3	California	Tops & Tees	3
4	Guangdong	Intimates	1
5	Guangdong	Shorts	2
6	Guangdong	Jeans	3
7	Texas	Intimates	1
8	Texas	Swim	2
9	Texas	Sleep & Lounge	3
10	Texas	Tops & Tees	3

In this example, we can see that the popularity of product categories varies by location. The preferences in California and Texas are somewhat similar, while they are completely different in Hong Kong.
Can we expand the range or add products for cross-selling to increase sales in prospective markets? A detailed study of preferences in different markets will enable better adjustment of advertising campaigns, assortment planning, and assortment expansion

```sql
-- What browser do website visitors use the most in 2024?
SELECT browser
,ROUND(COUNT(DISTINCT session_id) / (SELECT COUNT(DISTINCT session_id)
                            FROM `bigquery-public-data.thelook_ecommerce.events`
                            WHERE DATE(created_at) >= '2024-01-01') * 100, 2) as
percent_of_sessions
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01'
GROUP BY browser
ORDER BY percent_of_sessions DESC;
```

Row	browser	percent_of_sessions
1	Chrome	49.79
2	Firefox	20.18
3	Safari	19.97
4	Other	5.03
5	IE	5.03

Visitors primarily use three browsers: Chrome, Firefox, and Safari. Therefore, it's essential to prioritize monitoring the correct display of pages in these browsers when developing new pages or making changes to existing ones.
Additionally, it's recommended to collect information about the devices from which visits occur for future analysis. This will aid in evaluating the usability of both the mobile and desktop versions of the site, as well as provide insights into the target audience's demographics.

```
-- What is the cancellation event percentage breakdown by browser type in 2024?
SELECT browser
,ROUND(SUM(CASE WHEN event_type = 'cancel' THEN 1 ELSE 0 END)/COUNT(DISTINCT session_id)*100,
2) as cancel_percentage
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01'
GROUP BY browser
ORDER BY cancel_percentage DESC;
```

Row	browser	cancel_percentage
1	Firefox	10.39
2	IE	10.25
3	Chrome	10.24
4	Safari	10.23
5	Other	10.03

It appears that the cancellation rate is not significantly influenced by the type of browser.

```
-- What steps do visitors typically take before the Cancel event in 2024?
WITH cancel_event AS (
SELECT session_id
,sequence_number
,event_type
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01' AND
session_id in
     (SELECT session_id
     FROM `bigquery-public-data.thelook_ecommerce.events`
```

```
    WHERE event_type = 'cancel')
    )

SELECT event_type
,COUNT(*) as n_events
,sequence_number
FROM cancel_event
GROUP BY event_type, sequence_number
ORDER BY sequence_number;
```

Row	event_type	n_events	sequence_number
1	product	8214	1
2	cart	8214	2
3	cancel	8214	3

As you can see, the cancellation event follows a specific scenario: Product -> Cart -> Cancel. It's worth investigating why visitors cancel their session after reaching the cart. It's possible that visitors encounter an unclear interface in this part of the site or encounter information about additional costs (such as tax or delivery) that were not obvious before reaching this step.

```
----What are the user paths or typical navigation patterns on the site in 2024?
WITH user_paths AS (
SELECT STRING_AGG(event_type, ' -> ' ORDER BY sequence_number) AS path
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01' AND session_id NOT IN
                                        (SELECT DISTINCT session_id
                                        FROM `bigquery-public-data.thelook_ecommerce.events`
                                        WHERE DATE(created_at) BETWEEN '2023-12-01' AND
'2023-12-31')
GROUP BY session_id
)

SELECT path,
COUNT(*) AS n_sessions
FROM user_paths
GROUP BY path
ORDER BY n_sessions DESC;
```

Row	path	n_sessions
1	home -> department -> product -> cart -> purchas	22762
2	department -> product -> cart -> department -> product -> cart -> purchase	12936
3	department -> product -> cart	8227
4	department -> product	8223
5	product -> cart -> cancel	8214
6	product	8068
7	department -> product -> cart -> department -> product -> cart -> department -> product -> cart -> department -> product -> cart -> purchase	6618
8	department -> product -> cart -> department -> product -> cart -> department -> product -> cart -> purchase	4743

As we can see, there are several scenarios for making a purchase, and buyers take multiple steps to place an order. However, sessions that did not end with a purchase contain only 1 to 3 steps. It can be assumed that in these cases, the visitor was not interested in the product because it was either not relevant to them or not suitable in terms of price.
We can also see that there are some sessions where visitors abandoned the cart without pressing the cancel button.

```
-- What is the cart abandonment rate in 2024?
WITH user_paths AS (
SELECT session_id
,STRING_AGG(event_type, ' -> ' ORDER BY sequence_number) AS path
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01' AND session_id NOT IN
                                        (SELECT DISTINCT session_id
                                        FROM `bigquery-public-data.thelook_ecommerce.events`
                                        WHERE DATE(created_at) BETWEEN '2023-12-01' AND
'2023-12-31')
GROUP BY session_id
)

SELECT
ROUND((1- SUM(CASE WHEN path LIKE "%purchase%" THEN 1 ELSE 0 END) /
SUM(CASE WHEN path LIKE "%cart%" THEN 1 ELSE 0 END))*100, 2) as cart_aband_rate
FROM user_paths;
```

Row	cart_aband_rate
1	25.75

All cart additions are taken into account to calculate the cart abandonment rate. Achieving a rate of 25.75% is a positive result for this metric. (The average cart abandonment rate is 69.99%, according to the Baymard Institute).

```sql
-- What is the maximum, minimum and average duration of a session on the site?
WITH session_durations AS (
SELECT session_id
,TIMESTAMP_DIFF(MAX(created_at), MIN(created_at), SECOND)/ 60.0 AS session_duration_min
FROM `bigquery-public-data.thelook_ecommerce.events`
GROUP BY session_id
)

SELECT ROUND(MAX(session_duration_min)) as longest_sessioin_min
,ROUND(MIN(session_duration_min)) as shortest_session_min
,ROUND(AVG(session_duration_min)) as avg_session_min
FROM session_durations;
```

Row	longest_sessioin_mi	shortest_session_mi	avg_session_min
1	5788.0	0.0	410.0

We can observe that the dispersion of session duration data is indeed very large, indicating the presence of anomalies.This requires a more detailed study.
For now, let's focus on short sessions, specifically those that lasted less than a minute.

```sql
-- Which traffic source had the highest percentage of sessions lasting less than a minute in 2024?
WITH session_durations AS (
SELECT session_id
,traffic_source
,TIMESTAMP_DIFF(MAX(created_at), MIN(created_at), SECOND)/ 60.0 AS session_duration_min
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01'
GROUP BY session_id, traffic_source
),

all_traffic AS (
SELECT traffic_source
,COUNT(*) as count_all
FROM session_durations
GROUP BY traffic_source
),
```

```sql
min_traffic AS (
SELECT traffic_source
,COUNT(*) as count_min
FROM session_durations
WHERE session_duration_min = 0
GROUP BY traffic_source
)

SELECT a.traffic_source
,ROUND(count_min/count_all*100, 2) as min_duration_percent
FROM all_traffic a
JOIN min_traffic m
ON a.traffic_source = m.traffic_source
ORDER BY min_duration_percent DESC;
```

Row	traffic_source	min_duration_percent
1	Organic	11.64
2	YouTube	11.17
3	Facebook	10.96
4	Adwords	10.81
5	Email	10.81

The organic traffic source has the highest percentage of sessions lasting less than a minute. However, this source also has the lowest number of sessions and the lowest conversion rate. It can be assumed that organically, visitors encounter offers that are not relevant to them more often. Therefore, it is recommended to explore opportunities to improve SEO and organic traffic to the site.

```sql
-- Top 10 dates with the highest number of purchases on the site, along with the average number of orders per day in 2024
SELECT DATE(created_at) AS date
,COUNT(DISTINCT session_id) AS n_purchases
,ROUND(AVG(COUNT(DISTINCT session_id)) OVER ()) AS avg_daily_purchases
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE DATE(created_at) >= '2024-01-01'AND event_type = 'purchase'
GROUP BY date
ORDER BY n_purchases DESC
LIMIT 10;
```

Row	date	n_purchases	avg_daily_purchases
1	2024-05-08	1683	357.0
2	2024-05-07	1175	357.0
3	2024-05-10	1085	357.0
4	2024-05-09	996	357.0
5	2024-05-06	855	357.0
6	2024-05-05	672	357.0
7	2024-05-11	569	357.0
8	2024-05-04	520	357.0
9	2024-05-03	501	357.0
10	2024-05-01	497	357.0

It is recommended to investigate which advertising campaigns were active on these dates and to consider external factors such as holidays and sales events, which could have influenced the increase in the number of purchases.

```
-- What percentage of visitors made more than one purchase over the entire time period?
WITH repeat_purchases AS (
SELECT user_id
,COUNT(*) as n_purchases
FROM `bigquery-public-data.thelook_ecommerce.events`
WHERE user_id IS NOT NULL AND event_type = 'purchase'
GROUP BY user_id
ORDER BY n_purchases DESC)

SELECT ROUND(SUM(CASE WHEN n_purchases > 1 THEN 1 ELSE 0 END)/COUNT(*) * 100, 2) as
repeat_purchase_percent
FROM repeat_purchases;
```

Row	repeat_purchase_percent
1	56.24

This indicator is recommended for tracking customer loyalty.