E-commerce Insights: Exploring Data in BigQuery with SQL

Overview:

In this project, I performed a comprehensive SQL-based analysis of TheLook's e-commerce data, a clothing retailer developed by the Google Looker team. The dataset, hosted on Google BigQuery, includes transactional records from August 2019 to September 2024.

The goal was to explore customer behavior, product performance, and sales trends by working with a subset of the available data, focusing on four key tables: Order_items, Orders, Products, and Users. This analysis provides actionable insights for product evaluation, customer segmentation, and business growth strategies.

1. Customers Responsible for the Most Revenue

This query identifies the top 10 customers contributing the most to revenue. The results highlight high-value customers, useful for targeting with special offers, loyalty programs, or retention strategies.

```sql
SELECT
  u.id,
  CONCAT(u.first_name," ",u.last_name) AS full_name,
  ROUND(SUM(oi.sale_price * o.num_of_item),2) AS total_spent
FROM bigquery-public-data.thelook_ecommerce.users AS u
JOIN bigquery-public-data.thelook_ecommerce.orders  AS o ON u.id = o.user_id
JOIN bigquery-public-data.thelook_ecommerce.order_items AS oi ON o.order_id = oi.order_id
GROUP BY u.id, full_name
ORDER BY total_spent DESC
LIMIT 10;
```

Row	id	full_name	total_spent
1	83144	Christopher Page	7286.0
2	93244	Anthony Villarreal	6220.0
3	60960	Jeffery Thompson	5775.02
4	50303	Benjamin Kidd	5595.92
5	65111	Patrick Warren	5359.89
6	18906	Edwin Edwards	5246.0
7	85252	Gabriel Johnson	5169.92
8	75467	Kathleen Lin	4946.45
9	96000	Andrew Blake	4825.36
10	59359	Alex Johnson	4748.91

- **Insight**: The top 10 customers are likely responsible for a significant portion of the business's revenue. Tailoring marketing campaigns or VIP programs for these customers could lead to increased loyalty and repeat purchases.

2. Average Order Value (AOV) Per Customer

The query calculates the average order value (AOV) per customer, providing an idea of how much, on average, each customer spends on a transaction.

```
SELECT
  o.user_id,
  CONCAT(u.first_name," ",u.last_name) AS full_name,
  AVG(SUM(items.sale_price * o.num_of_item)) OVER (PARTITION BY o.user_id) AS
avg_order_value
FROM bigquery-public-data.thelook_ecommerce.orders AS o
JOIN bigquery-public-data.thelook_ecommerce.users AS u ON o.user_id = u.id
JOIN bigquery-public-data.thelook_ecommerce.order_items AS items ON items.user_id=u.id
GROUP BY full_name,o.user_id
ORDER BY avg_order_value DESC
```

Row	user_id	full_name	avg_order_value
1	60960	Jeffery Thompson	20554.66008758...
2	83144	Christopher Page	18545.0
3	88792	Todd Sheppard	15212.50007629...
4	65111	Patrick Warren	15134.03999519...
5	97557	Kristen Gentry	13580.60004615...
6	27523	Lindsey Jimenez	13073.40000152...
7	81342	Heather Downs	12280.40000915...
8	9703	Selena Mosley	11826.61994743...
9	75467	Kathleen Lin	11437.84001159...
10	89348	Christina Young	11384.01000881...

- **Insight**: A higher AOV indicates that customers are purchasing more items or higher-priced items per transaction. This insight can be used to push strategies like cross-selling and upselling to maintain or increase the AOV.

❖ Customer By Gender

This query calculates the revenue and quantity of items sold based on the gender of the customer.

```
   SELECT
     o.gender,
     ROUND(SUM(oi.sale_price * o.num_of_item), 2) AS revenue,
     SUM(o.num_of_item) AS quantity
   FROM bigquery-public-data.thelook_ecommerce.order_items AS oi
   INNER JOIN bigquery-public-data.thelook_ecommerce.orders AS o ON oi.user_id =
   o.user_id
WHERE oi.status NOT IN ('Cancelled', 'Returned')
   GROUP BY gender
   ORDER BY revenue DESC;
```

Row	gender	revenue	quantity
1	M	14431624.65	230116
2	F	13007259.85	233365

Insight: Understanding which gender group brings in the most revenue and purchases the most items can help tailor marketing campaigns to either target underperforming segments or strengthen existing dominant segments.

4. Most Profitable Customer Segments (Age Group & Location)

This query analyzes the most profitable customer segments based on age group and location.

```
SELECT
  u.state,
  CASE WHEN u.age BETWEEN 18 AND 25 THEN '18-25'
       WHEN u.age BETWEEN 26 AND 35 THEN '26-35'
       WHEN u.age BETWEEN 36 AND 45 THEN '36-45'
  ELSE '46+'
  END AS age_group,
  ROUND(SUM(o.num_of_item * oi.sale_price),2) AS total_spent
FROM bigquery-public-data.thelook_ecommerce.order_items AS oi
JOIN bigquery-public-data.thelook_ecommerce.orders AS o ON oi.order_id = o.order_id
JOIN bigquery-public-data.thelook_ecommerce.users AS u ON o.user_id = u.id
GROUP BY u.state, age_group
ORDER BY total_spent DESC;
```

Row	state	age_group	total_spent
7	Zhejiang	46+	244187.29
8	Beijing	46+	222160.72
9	Jiangsu	46+	209819.59
10	Hebei	46+	208080.65
11	Guangdong	26-35	192371.82
12	Gyeonggi-do	46+	191442.28
13	Florida	46+	187588.81
14	Guangdong	36-45	175817.86
15	Henan	46+	168251.18

Insight: Identifying which age groups and states are contributing most to revenue helps in tailoring location-based marketing strategies and understanding demographic preferences. E.g., if the '26-35' age group is the most profitable, promotional strategies can be designed specifically for that cohort.

5. Return/Refund Rate & Products with the Highest Return Rate

This query calculates the return/refund rate and identifies which products have the highest return rates.

```
SELECT
    p.name,
    COUNT(o.returned_at) / COUNT(o.order_id) AS return_rate
FROM bigquery-public-data.thelook_ecommerce.orders AS o
JOIN bigquery-public-data.thelook_ecommerce.order_items AS oi ON o.order_id = oi.order_id
JOIN bigquery-public-data.thelook_ecommerce.products AS p ON oi.product_id = p.id
WHERE o.status = 'Returned'
GROUP BY p.name
ORDER BY return_rate DESC;
```

Row	name	return_rate
1	Wayfarer Style Sunglasses Dark Lens Black Frame	1.0
2	Blank Long Cuff Beanie Cap (Choose Many Colors ...	1.0
3	Pink Ribbon Breast Cancer Awareness Knee High Socks Great for Sports Teams Fundraising Relay for Life Walk Survivor (Style 26)	1.0
4	TopTie Mens Black & White Checkerboard Pre-Tied Satin Formal Bow Tie	1.0
5	Retractable Colorful Rhinestone Lanyards with Breakaway Feature ID Badge Holder & Key Chain	1.0

Insight: Products with high return rates may indicate quality issues or mismatched customer expectations. Businesses should investigate these products to either improve their quality or adjust how they are marketed and described online.

6. Average Time Between Purchases for Repeat Customers

This query calculates the average time between purchases for repeat customers.

```
WITH purchase_dates AS (
    SELECT
        o.user_id,
        CONCAT(u.first_name," ",u.last_name)AS full_name,
        o.created_at,
           LEAD(o.created_at) OVER (PARTITION BY user_id ORDER BY o.created_at) AS
next_purchase_date
    FROM bigquery-public-data.thelook_ecommerce.orders as o
    JOIN bigquery-public-data.thelook_ecommerce.users as u ON u.id = o.user_id
    WHERE status = 'Complete'
)

SELECT
    user_id,
    full_name,
    AVG(DATE_DIFF(next_purchase_date, created_at, DAY)) AS avg_time_between_purchases
FROM purchase_dates
WHERE next_purchase_date IS NOT NULL
GROUP BY user_id,full_name
```

Row	user_id	full_name	avg_time_between_p
1	88449	Monica Rodriguez	1875.0
2	81631	Alex Ruiz	1870.0
3	59376	Ashley Jones	1846.0
4	64023	Thomas Murray	1819.0
5	69651	Jamie Simmons	1798.0
6	40199	Tracie Moore	1780.0
7	17275	Jordan Davis	1729.0
8	14585	James Smith	1703.0
9	6871	Donald Meza	1670.0
10	73509	Tyler Jones	1644.0

Insight: Knowing the average time between purchases helps identify how frequently repeat customers return to buy. This can inform retention strategies, like the timing of follow-up emails or promotions to encourage shorter purchase cycles.

7. Order Value by Time Period (Month)

This query tracks order value across different time periods (months) to observe sales trends.

```
SELECT
  EXTRACT(MONTH FROM o.created_at) AS month_number,
  FORMAT_DATE('%B', o.created_at) AS order_month,
  ROUND(SUM(o.num_of_item * p.retail_price),2) AS total_revenue
FROM  bigquery-public-data.thelook_ecommerce.order_items AS oi
JOIN bigquery-public-data.thelook_ecommerce.orders AS o ON oi.order_id = o.order_id
JOIN bigquery-public-data.thelook_ecommerce.products AS p ON oi.product_id = p.id
WHERE o.status = 'Complete'
GROUP BY  month_number,order_month
ORDER BY month_number ASC;
```

Row	month_number	order_month	total_revenue
1	1	January	364362.32
2	2	February	364539.14
3	3	March	372916.91
4	4	April	401641.81
5	5	May	454650.77
6	6	June	477550.65
7	7	July	533669.47
8	8	August	646027.02
9	9	September	517626.17
10	10	October	306222.5
11	11	November	324131.71
12	12	December	324886.39

Insight: Seasonal trends can be identified through this analysis. For instance, if revenue spikes in November and December (due to holiday shopping), businesses can prepare for these peaks by stocking up inventory and increasing marketing efforts.

7. Best and Worst Performing Product Category

The query finds which product categories are selling the most and least.

```sql
SELECT
  p.category AS category,
  ROUND(SUM(sale_price * num_of_item), 2) AS revenue,
  SUM(num_of_item) AS quantity
FROM bigquery-public-data.thelook_ecommerce.order_items oi
JOIN bigquery-public-data.thelook_ecommerce.orders o ON oi.order_id = o.order_id
JOIN bigquery-public-data.thelook_ecommerce.products p ON oi.product_id = p.id
WHERE oi.status NOT IN ('Cancelled', 'Returned')
GROUP BY category
ORDER BY revenue DESC;
```

Row	category	revenue	quantity
1	Outerwear & Coats	2266744.64	15481
2	Jeans	2130430.06	21249
3	Sweaters	1451674.54	19205
4	Fashion Hoodies & Sweatshirts	1092197.55	19789
5	Swim	1090888.22	19154
6	Suits & Sport Coats	1076963.98	8644
7	Sleep & Lounge	961154.94	19499
8	Shorts	868007.52	19012
9	Tops & Tees	842358.37	20367
10	Dresses	776816.75	9170

Row	category	revenue	quantity
1	Clothing Sets	30477.52	372
2	Jumpsuits & Rompers	65974.02	1581
3	Socks & Hosiery	105922.96	6504
4	Leggings	144060.11	5426
5	Skirts	183671.41	3596
6	Suits	207333.5	1800
7	Socks	216757.01	10618
8	Plus	284885.69	7457
9	Pants & Capris	314489.31	5794

9. Best and Worst Performing Brands

The query finds which brands are selling the most and least.

```sql
SELECT
  p.brand AS brand,
  ROUND(SUM(sale_price * num_of_item), 2) AS revenue,
  SUM(num_of_item) AS quantity
FROM bigquery-public-data.thelook_ecommerce.order_items oi
JOIN bigquery-public-data.thelook_ecommerce.orders o ON oi.order_id = o.order_id
JOIN bigquery-public-data.thelook_ecommerce.products p ON oi.product_id = p.id
WHERE oi.status NOT IN ('Cancelled', 'Returned')
GROUP BY brand
ORDER BY revenue DESC;
```

Row	brand	revenue	quantity
1	Calvin Klein	290856.67	4670
2	True Religion	279206.78	1363
3	Diesel	269686.61	2129
4	7 For All Mankind	247791.24	1584
5	Carhartt	230800.4	3456
6	Tommy Hilfiger	181143.76	2431
7	Joe's Jeans	154835.55	1024
8	Volcom	154821.77	2645
9	Columbia	150391.18	2170
10	Quiksilver	141858.03	2463

Row	brand	revenue	quantity
1	marshal	0.06	3
2	Tabi Socks	5.99	1
3	Made in USA	6.37	13
4	Wholesale LOCS DG XLOOP CH...	7.98	2
5	California Costumes	9.09	1
6	e.g. smith	9.97	1
7	Wayfayrer	10.5	7
8	Wise Guy Productions	14.22	3
9	Skyblue	14.95	5
10	SilverHooks	14.99	1

Insight: Brands and Product Categories that perform the best can be featured in promotions or marketing campaigns. Conversely, the business might need to reevaluate or phase out underperforming brands or categories.

10. Performance by Marketing Channel

This query tracks the total number of customers acquired through different marketing channels.

```sql
SELECT
  u.traffic_source AS traffic_source,
  COUNT(DISTINCT oi.user_id) AS total_customers
FROM bigquery-public-data.thelook_ecommerce.order_items oi
JOIN bigquery-public-data.thelook_ecommerce.users u
ON oi.user_id = u.id
WHERE oi.status NOT IN ('Cancelled', 'Returned')
GROUP BY traffic_source
ORDER BY total_customers DESC;
```

Row	traffic_source	total_customers
1	Search	46183
2	Organic	9971
3	Facebook	4033
4	Email	3320
5	Display	2607

Insight: Knowing which marketing channels are driving the most customers helps allocate marketing budget more effectively. The business can focus more resources on high-performing channels (e.g., organic search, paid ads) and possibly reduce spending on underperforming ones.