



Cohort Analysis

-TheLook eCommerce-

By Cessa Mutiara Aziz

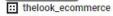
01 **OVERVIEW**



Overview of The Dataset

TheLook is a fictious eCommerce clothing site developed by the Looker team. The dataset contains synthetic database and provides contents for the purposes of product discovery, testing, and evaluation.





events

inventory_items

order_items

orders

products

users



02 Assignment Goals

Goals

Create monthly retention cohorts (based upon the date that a user purchased a product) and then how many of them (%) coming back for the following month in 2019 - 2022 and gain the insight.



Create monthly retention cohorts (based upon the date that a user purchased a product) and then how many of them (%) coming back for the following month in 2019 - 2022 and gain the insight

Field name	Type	Mode
cohort_month	DATE	NULLABLE
cohort_size	INTEGER	NULLABLE
num_month	INTEGER	NULLABLE
total_users	INTEGER	NULLABLE
percentage	FLOAT	NULLABLE

Row	cohort_month	cohort_size	num_month	total_users	percentage
1	2019-01-01	26	0	26	1.0
2	2019-01-01	26	1	2	0.076923076923076927
3	2019-01-01	26	6	1	0.038461538461538464
4	2019-01-01	26	9	1	0.038461538461538464
5	2019-01-01	26	10	1	0.038461538461538464
6	2019-01-01	26	12	1	0.038461538461538464
7	2019-01-01	26	14	2	0.076923076923076927
8	2019-01-01	26	17	1	0.038461538461538464
9	2019-01-01	26	19	1	0.038461538461538464
10	2019-01-01	26	22	1	0.038461538461538464
11	2019-01-01	26	24	1	0.038461538461538464
12	2019-01-01	26	26	1	0.038461538461538464
13	2019-01-01	26	29	1	0.038461538461538464
14	2019-01-01	26	39	2	0.076923076923076927
15	2019-02-01	85	0	85	1.0

SQL Syntax

```
WITH cohort_item AS
  (SELECT
    user_id,
    MIN(DATE(DATE_TRUNC(created_at, MONTH))) AS cohort_month
  FROM bigquery-public-data.thelook_ecommerce.order_items
  GROUP BY user_id),
--- Find the size of each cohort by counting the number of ids that purchased for the first time in a month - cohort_size
cohort_size AS
  (SELECT
    cohort_month,
    COUNT(1) as num_users
  FROM cohort_item
  GROUP BY cohort_month),
-- 2. Measure Activity After Cohort Month
--- Find what months there's been activity after their cohort month.
user activities AS
(SELECT
    DATE_DIFF ((DATE_TRUNC(oi.created_at, MONTH))), cohort_item.cohort_month, MONTH) AS num_month,
    oi.user id AS user id
    FROM bigguery-public-data, thelook ecommerce, order items AS oi
    LEFT JOIN cohort item
    ON cohort item.user id = oi.user id
    WHERE EXTRACT(YEAR FROM cohort_item.cohort_month) IN (2019,2020,2021,2022)
    GROUP BY 2.1).
---count how many users were retained in each month after their cohort month -- retention table
retention table AS
(SELECT
  c.cohort month.
  a.num month.
  COUNT(1) as num users
FROM user_activities AS a
LEFT JOIN cohort_item AS c
ON a.user_id = c.user_id
GROUP BY 1, 2)
--3. FINAL
----divide the number of remaining users by the cohort size
---- (cohort_month, size, month_number, percentage)
SELECT
  r.cohort_month,
  s.num_users AS cohort_size,
  r.num_month,
 r.num_users AS total_users,
r.num_users/s.num_users AS percentage
FROM retention_table AS r
LEFT JOIN cohort_size AS s
ON r.cohort_month = s.cohort_month
WHERE r.cohort_month IS NOT NULL
ORDER BY 1,3;
```

Result Overview

The query results a monthly retention cohorts based upon the first date of a user's purchase and the following months in 2019 - 2022.

But for the further analysis we'll analyze the time frame of the latest 1 year (2021-2022) to measure our user engagement for the next marketing strategy.



Cohort Analysis

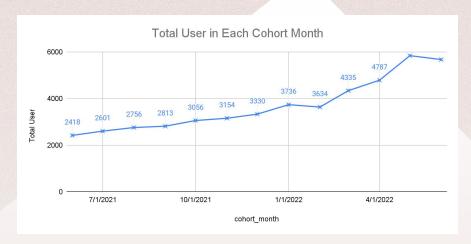
AVERAGE of pe nu	ım_month													9
cohort_month													12 Gra	and Total
2021-06-01	100.00%	6.74%	4.43%	4.18%	4.63%	4.47%	5.17%	4.51%	3.72%	5.42%	4.63%	4.59%	2.44%	11.92%
2021-07-01	100.00%	7.15%	4.61%	4.81%	5.54%	5.61%	5.61%	4.46%	5.57%	4.58%	5.38%	2.42%		12.98%
2021-08-01	100.00%	6.46%	4.68%	5.48%	6.17%	5.62%	5.30%	5.88%	4.75%	5.62%	2.87%			13.89%
2021-09-01	100.00%	7.18%	5.76%	5.37%	5.51%	5.08%	5.55%	5.37%	5.76%	3.09%				14.87%
2021-10-01	100.00%	8.21%	7.04%	6.15%	5.66%	6.41%	6.45%	6.18%	3.01%					16.57%
2021-11-01	100.00%	8.08%	6.56%	6.79%	7.83%	6.28%	7.13%	3.30%						18.25%
2021-12-01	100.00%	10.78%	6.94%	8.35%	7.48%	7.63%	4.38%							20.79%
2022-01-01	100.00%	9.07%	9.21%	8.75%	8.43%	4.58%								23.34%
2022-02-01	100.00%	12.99%	9.74%	9.19%	5.01%									27.39%
2022-03-01	100.00%	13.70%	11.76%	5.63%										32.77%
2022-04-01	100.00%	16.88%	8.17%											41.68%
2022-05-01	100.00%	14.88%												57.44%
2022-06-01	100.00%													100.00%
Grand Total	100.00%	10.18%	7.17%	6.47%	6.25%	5.71%	5.66%	4.95%	4.56%	4.68%	4.29%	3.51%	2.44%	19.79%
SUM of total_usen														
cohort_month	0	1											12 (Grand Total
2021-06-01	2418	163	107	101	112	108	125	109	90	131	112	111	59	374
2021-07-01	2601	186	120	125	144	146	146	116	145	119	140	63		405
2021-08-01	2756	178	129	151	170	155	146	162	131	155	79			421
2021-09-01	2813	202	162	151	155	143	156	151	162	87				418
2021-10-01	3056	251	215	188	173	196	197	189	92					455
2021-11-01	3154	255	207	214	247	198	225	104						460
	0000	359	231	278	249	254	146							484
2021-12-01	3330	338												523
	3330 3736	339	344	327	315	171								020
2021-12-01				327 334	315 182	171								
2021-12-01 2022-01-01	3736	339	344			171								497
2021-12-01 2022-01-01 2022-02-01	3736 3634	339 472	344 354	334		171								497 568
2021-12-01 2022-01-01 2022-02-01 2022-03-01	3736 3634 4335	339 472 594	344 354 510	334		171								497 568 598 670
2021-12-01 2022-01-01 2022-02-01 2022-03-01 2022-04-01	3736 3634 4335 4787	339 472 594 808	344 354 510	334		171								497 568 598

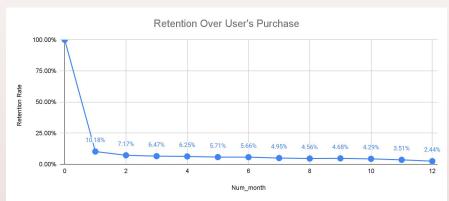
Cohort Analysis

The total user in each cohort month shows an upward trend in a year (number of user increasing over time).

The retention curve shows the retention of the cohorts over time, out of all new users during the time range (48131 users), 10.18% are retained on 1 month after, 5.66% on 6 months after, and 2.44% on 12 months after.

Overall, in the time frame of **6** months, the user who come back to purchase in the marketplace is decreasing around **50%**.





Hypothesis

1. Decreasing activity in Active users in the same cohort within a year;

2. Decreasing activity in non-active users in the same cohort within a year;

3. The total user who completed their orders has decreased within a year.

Hypothesis 1: Decreasing activity in Active users in the same cohort within a year

Field name	Туре	Mode
cohort_month	DATE	NULLABLE
cohort_size	INTEGER	NULLABLE
num_month	INTEGER	NULLABLE
total_users	INTEGER	NULLABLE
percentage	FLOAT	NULLABLE

SCH	HEMA DET	TAILS F	PREVIEW					
Row	cohort_month	cohort_size	num_month	total_users	percentage			
1	2022-01-01	269	1	50	0.18587360594795538			
2	2022-01-01	269	2	53	0.19702602230483271			
3	2022-01-01	269	3	49	0.18215613382899629			
4	2022-01-01	269	4	43	0.15985130111524162			
5	2022-01-01	269	5	29	0.10780669144981413			
6	2022-01-01	269	0	269	1.0			
7	2022-02-01	272	1	59	0.21691176470588236			
8	2022-02-01	272	2	53	0.19485294117647059			
9	2022-02-01	272	3	59	0.21691176470588236			
10	2022-02-01	272	4	33	0.12132352941176471			
11	2022-02-01	272	0	272	1.0			
12	2019-01-01	26	0	26	1.0			
13	2019-01-01	26	6	1	0.038461538461538464			
14	2019-01-01	26	9	1	0.038461538461538464			
15	2019-01-01	26	10	1	0.038461538461538464			

Active users

Cohort month = registered date

Hypothesis 1 - SQL Syntax For the query process & more info, click here

```
#Addition for Question 7 - Measure the activities of Active Users (cohort month = registered date) throughout the period
WITH cohort_items_temp AS
  (SELECT
    user id.
    MIN(DATE(DATE_TRUNC(created_at, MONTH))) AS cohort_month
  FROM bigquery-public-data.thelook_ecommerce.order_items
  GROUP BY user_id),
cohort_item AS
(SELECT cohort_items_temp.*,
from cohort items temp
left join bigquery-public-data.thelook_ecommerce.users AS user
ON cohort_items_temp.user_id = user.id
WHERE cohort_month = date(date_trunc(user.created_at,MONTH))),
--- Find the size of each cohort by counting the number of unique ids that purchased for the first time in a month - cohort_size
cohort_size AS
  (SELECT
    cohort_month,
    COUNT(1) as num_users
  FROM cohort_item
  GROUP BY cohort_month),
-- 2. Measure Activity After Cohort Month
--- Find what months there's been activity after their cohort month.
user_activities AS
(SELECT
    DATE_DIFF ((DATE(DATE_TRUNC(oi.created_at, MONTH))),cohort_item.cohort_month,MONTH) AS num_month,
    FROM bigquery-public-data.thelook_ecommerce.order_items AS oi
    LEFT JOIN cohort_item
    ON cohort item.user id = oi.user id
    WHERE EXTRACT(YEAR FROM cohort_item.cohort_month) IN (2019,2020,2021,2022)
    GROUP BY 1, 2),
---count how many users were retained in each month after their cohort month -- retention table
retention_table AS
(SELECT
 c.cohort month.
 a.num_month,
 COUNT(1) as num_users
FROM user_activities AS a
LEFT JOIN cohort item AS c
ON a.user_id = c.user_id
GROUP BY 1, 2)
----divide the number of remaining users by the cohort size
---- (cohort_month, size, month_number, percentage)
SELECT
 r.cohort_month,
 s.num users AS cohort size.
 r.num_month,
 r.num_users AS total_users,
r.num_users/s.num_users AS percentage
FROM retention table AS r
LEFT JOIN cohort_size AS s
ON r.cohort_month = s.cohort_month
WHERE r.cohort month IS NOT NULL
ORDER BY 1,3;
```

Hypothesis 1:

Decreasing activity in Active users in the same cohort within a year

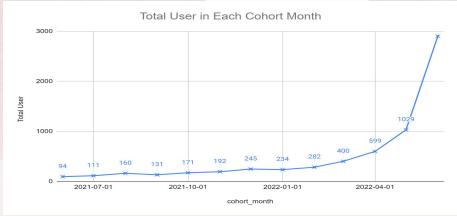
	1029	393												14
2022-04-01	599	236 393	133											9
2022-03-01	400	116	98	48										6
2022-02-01	282	76	75	64	34									
2022-01-01	234	37	55	42	48	25								4
2021-12-01	245	50	47	39	38	31	21							
2021-11-01	192	33	19	21	26	24	30	15						
2021-10-01	171	30	22	19	20	23	15	21	16					
2021-09-01	131	19	17	9	13	6	14	10	13	10				
2021-08-01	160	20	5	16	12	17	13	15	10	11	5			
2021-07-01	111	13	15	4	13	9	12	8	18	11	10	2		
2021-06-01	94	12	6	6	7	8	5	4	4	11	7	9	4	
ohort_month													12 Gra	and Total
SUM of total_usen	um_month													
Frand Total	100.00%	21.33%	15.89%	11.75%	12.21%	10.14%	9.70%	8.09%	9.20%	9.03%	6.53%	5.69%	4.26%	24.
2022-06-01	100.00%													100.0
2022-05-01	100.00%	38.19%												69.1
2022-04-01	100.00%	39.40%	22.20%											53.8
2022-03-01	100.00%	29.00%	24.50%	12.00%										41.
2022-02-01	100.00%	26.95%	26.60%	22.70%	12.06%									37.
2022-01-01	100.00%	15.81%	23.50%	17.95%	20.51%	10.68%								31.
2021-12-01	100.00%	20.41%	19.18%	15.92%	15.51%	12.65%	8.57%							27.
2021-11-01	100.00%	17.19%	9.90%	10.94%	13.54%	12.50%	15.63%	7.81%						23.
2021-10-01	100.00%	17.54%	12.87%	11.11%	11.70%	13.45%	8.77%	12.28%	9.36%					21.
2021-09-01	100.00%	14.50%	12.98%	6.87%	9.92%	4.58%	10.69%	7.63%	9.92%	7.63%				18.4
2021-08-01	100.00%	12.50%	3.13%	10.00%	7.50%	10.63%	8.13%	9.38%	6.25%	6.88%	3.13%			16.
2021-07-01	100.00%	11.71%	13.51%	3.60%	11.71%	8.11%	10.81%	7.21%	16.22%	9.91%	9.01%	1.80%		16.
2021-06-01	100.00%	12.77%	6.38%	6.38%	7.45%	8.51%	5.32%	4.26%	4.26%	11.70%	7.45%	9.57%	4.26%	14.4

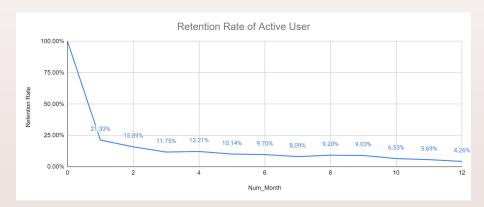
Hypothesis 1: Decreasing activity in Active users in the same cohort within a year

The total of active users in each cohort month shows an **upward exponential trend** in a year because of the high rise up in the 12 months after.

The retention curve shows the retention of the cohorts over time, out of all new active users during the time range (6549 users), 21.33% are retained on 1 month after, 9.70% on 6 months after, and 4.26% on 12 months after.

In summary, in the time frame of **6 months**, the active user who come back to purchase in the marketplace is **decreasing around 50%** (the same as our main cohort analysis)





Hypothesis 2: Decreasing activity in non-active users in the same cohort within a year

cohort_month DATE NULLABLE cohort_size INTEGER NULLABLE num_month INTEGER NULLABLE total_users INTEGER NULLABLE percentage FLOAT NULLABLE	Field name	Type	Mode
num_month INTEGER NULLABLE total_users INTEGER NULLABLE	cohort_month	DATE	NULLABLE
total_users INTEGER NULLABLE	cohort_size	INTEGER	NULLABLE
	num_month	INTEGER	NULLABLE
percentage FLOAT NULLABLE	total_users	INTEGER	NULLABLE
	percentage	FLOAT	NULLABLE

a	question7_a	addition_no	nactiveuser	90	QUERY	* SHARE	COPY	■ SNAPSHOT	TOTAL	≜ EXPORT	
SCH	HEMA D	ETAILS	PREVIEW								
Row	cohort_month	cohort_size	num_month	total_users	percen	tage					
	2019-02-01	58	11	5	0.0862	06896551724144					
2	2019-02-01	58	31	4	0.0689	65517241379309					
	2019-02-01	58	0	58	1.0						
ı	2019-02-01	58	1	3	0.0517	24137931034482					
i	2019-02-01	58	5	3	0.0517	24137931034482					
	2019-02-01	58	7	3	0.0517	24137931034482					
	2019-02-01	58	18	3	0.0517	24137931034482					
3	2019-02-01	58	26	3	0.0517	24137931034482					
)	2019-02-01	58	29	3	0.0517	24137931034482					
0	2019-02-01	58	2	1	0.0172	41379310344827					
1	2019-02-01	58	6	1	0.0172	41379310344827					
2	2019-02-01	58	8	1	0.0172	41379310344827					
3	2019-02-01	58	10	1	0.0172	41379310344827					
4	2019-02-01	58	15	1	0.0172	41379310344827					
5	2019-02-01	58	17	1	0.0172	41379310344827					

Non-Active users

Cohort month > registered date

Hypothesis 2 - SQL Syntax For the query process & more info, click here

```
#Addition for Question 7 - Measure the activities of Non-Active Users (cohort month > registered date) throughout the period
WITH cohort_items_temp AS
  (SELECT
    user id.
    MIN(DATE(DATE_TRUNC(created_at, MONTH))) AS cohort_month
  FROM bigquery-public-data.thelook_ecommerce.order_items
  GROUP BY user_id),
cohort_item AS
(SELECT cohort_items_temp.*,
from cohort items temp
left join bigquery-public-data.thelook_ecommerce.users AS user
ON cohort_items_temp.user_id = user.id
WHERE cohort_month > date(date_trunc(user.created_at,MONTH))),
--- Find the size of each cohort by counting the number of unique ids that purchased for the first time in a month - cohort_size
cohort_size AS
  (SELECT
    cohort_month,
    COUNT(1) as num_users
  FROM cohort_item
  GROUP BY cohort_month),
-- 2. Measure Activity After Cohort Month
--- Find what months there's been activity after their cohort month.
user_activities AS
(SELECT
    DATE_DIFF ((DATE(DATE_TRUNC(oi.created_at, MONTH))),cohort_item.cohort_month,MONTH) AS num_month,
    FROM bigquery-public-data.thelook_ecommerce.order_items AS oi
    LEFT JOIN cohort_item
    ON cohort item.user id = oi.user id
    WHERE EXTRACT(YEAR FROM cohort_item.cohort_month) IN (2019,2020,2021,2022)
    GROUP BY 1, 2),
---count how many users were retained in each month after their cohort month -- retention table
retention_table AS
(SELECT
 c.cohort month.
 a.num_month,
 COUNT(1) as num_users
FROM user_activities AS a
LEFT JOIN cohort item AS c
ON a.user_id = c.user_id
GROUP BY 1, 2)
----divide the number of remaining users by the cohort size
---- (cohort_month, size, month_number, percentage)
SELECT
 r.cohort_month,
 s.num users AS cohort size.
 r.num_month,
 r.num_users AS total_users,
r.num_users/s.num_users AS percentage
FROM retention table AS r
LEFT JOIN cohort_size AS s
ON r.cohort_month = s.cohort_month
WHERE r.cohort month IS NOT NULL
ORDER BY 1,3;
```

Hypothesis 2:
Decreasing activity in non-active users in the same cohort within a year

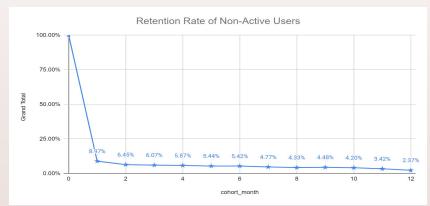
AVERAGE of pe nu	m_month													
cohort_month													12 Gra	and Total
2021-06-01	100.00%	6.50%	4.35%	4.09%	4.52%	4.30%	5.16%	4.52%	3.70%	5.16%	4.52%	4.39%	2.37%	11.81
2021-07-01	100.00%	6.95%	4.22%	4.86%	5.26%	5.50%	5.38%	4.34%	5.10%	4.34%	5.22%	2.45%		12.80
2021-08-01	100.00%	6.09%	4.78%	5.20%	6.09%	5.32%	5.12%	5.66%	4.66%	5.55%	2.85%			13.769
2021-09-01	100.00%	6.82%	5.41%	5.29%	5.29%	5.11%	5.29%	5.26%	5.56%	2.87%				14.699
2021-10-01	100.00%	7.66%	6.69%	5.86%	5.30%	6.00%	6.31%	5.82%	2.63%					16.25
2021-11-01	100.00%	7.49%	6.35%	6.52%	7.46%	5.87%	6.58%	3.00%						17.919
2021-12-01	100.00%	10.02%	5.96%	7.75%	6.84%	7.23%	4.05%							20.269
2022-01-01	100.00%	8.62%	8.25%	8.14%	7.62%	4.17%								22.809
2022-02-01	100.00%	11.81%	8.32%	8.05%	4.42%									26.529
2022-03-01	100.00%	12.15%	10.47%	4.98%										31.909
2022-04-01	100.00%	13.66%	6.16%											39.949
2022-05-01	100.00%	9.90%												54.959
2022-06-01	100.00%													100.009
Grand Total	100.00%	8.97%	6.45%	6.07%	5.87%	5.44%	5.42%	4.77%	4.33%	4.48%	4.20%	3.42%	2.37%	19.389
SUM of total_use nu														
cohort_month													12 Gra	and Total
2021-06-01	2324	151	101	95	105	100	120	105	86	120	105	102	55	356
2021-07-01	2490	173	105	121	131	137	134	108	127	108	130	61		382
2021-08-01	2596	158	124	135	158	138	133	147	121	144	74			392
2021-09-01	2682	183	145	142	142	137	142	141	149	77				394
2021-10-01	2885	221	193	169	153	173	182	168	76					422
2021-11-01	2962	222	188	193	221	174	195	89						424
2021-12-01	3085	309	184	239	211	223	125							437
2022-01-01	3502	302	289	285	267	146								479
2022-02-01	3352	396	279	270	148									444
2022-03-01	3935	478	412	196										502
2022-04-01	4188	572	258											501
2022-05-01	4810	476												528
2022-06-01	2771													277
Grand Total	41582	3641	2278	1845	1536	1228	1031	758	559	449	309	163	55	5543

Hypothesis 2: Decreasing activity in Non-Active users in the same cohort within a year

The total of non-active users in each cohort month shows an upward trend until 11 months after and a large drop occurs in the 12 months after.

The retention curve shows the retention of the cohorts over time, out of all new active users during the time range (41582 users), 8.97% are retained on 1 month after, 5.42% on 6 months after, and 2.37% on 12 months after.





Hypothesis 3: The total user who completed their orders has decreased within a year

Field name	Туре	Mode
ourchased_month	DATE	NULLABLE
total_user_purchased	INTEGER	NULLABLE
total_user_completed_orders	INTEGER	NULLABLE
total_user_cancelled_orders	INTEGER	NULLABLE
total_user_returned_orders	INTEGER	NULLABLE

	destion/_addi	ion_purchasedst	atus Q QUERY	+2 SHARE COPY	■ SNAPSHOT ■ DELETE ■ EXPORT ■ EXPOR
SCH	EMA DETAIL	S PREVIEW			
Row	purchased_month	total_user_purchased	total_user_completed_orders	total_user_cancelled_orders	total_user_returned_orders
	2019-01-01	26	7	8	1
2	2019-02-01	87	22	12	6
3	2019-03-01	157	42	23	15
4	2019-04-01	228	59	40	24
5	2019-05-01	349	79	56	41
5	2019-06-01	387	92	49	43
7	2019-07-01	497	136	77	55
В	2019-08-01	553	132	73	63
9	2019-09-01	658	178	113	53
10	2019-10-01	695	173	99	70
11	2019-11-01	819	202	132	80
12	2019-12-01	924	242	137	100
13	2020-01-01	1054	267	155	125
14	2020-02-01	1080	287	163	99
15	2020-03-01	1378	347	212	147

Hypothesis 3 - SQL Syntax

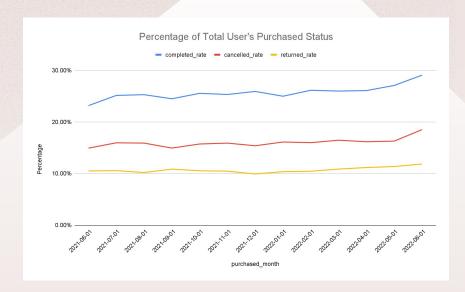
For the query process & more info, click here

```
#Addition for Question 7 - user's purchased status (completed orders, cancelled orders, and returned orders)
throughout the period
WITH obs AS
(SELECT
 DATE(DATE_TRUNC(created_at, MONTH)) AS purchased_month,
 COUNT (DISTINCT user_id) AS total_user_purchased,
 COUNT (DISTINCT CASE WHEN status = 'Complete'then user_id END) AS total_user_completed_orders,
 COUNT (DISTINCT CASE WHEN status = 'Cancelled'then user_id END) AS total_user_cancelled_orders,
 COUNT (DISTINCT CASE WHEN status ='Returned'then user_id END) AS total_user_returned_orders
FROM bigguery-public-data.thelook_ecommerce.order_items
WHERE EXTRACT (YEAR FROM created_at) IN (2019,2020,2021,2022)
GROUP BY 1
SELECT *
FROM obs
```

Hypothesis 3: The total user who completed their orders has decreased within a year

The percentage of total user who completed their orders within a year shows stationary trend and tend to slightly showing an uptrend in 2022.

So, the total user who completed their order hasn't decreased in a year.



Insight & Recommendation

As the overall of total user in each cohort month is increasing over time within the year, we can conclude that they have a good onboarding experience.

However, as the retention curve indicates that users aren't getting quickly engaged to the marketplace, resulting in drop-offs.

TheLook eCommerce needs to improve the user's engagement as quickly as possible, so it can boost the retention.

