***Fact_app_event*. (Tasks are executed by Azure Data Studio)**

- We take a look to see the brands are mostly sold by query:

```
SELECT DISTINCT device
FROM Fact_app_event
```

	device
1	xiaomi
2	oppo
3	realme
4	iphone
5	samsung

- We try to see what is brand is favored the most, the first we begin with xiaomi:

```
SELECT COUNT (*)
FROM Fact_app_event
WHERE device = 'xiaomi';
```

Result is: 5955. Try the same with the rest of products and we have results:
oppo: 5941 / realme: 6007 / iphone: 5996 / Samsung: 6101

- If we want to save time, we can use this query:

```
SELECT device, COUNT(*) AS number_of_device
FROM Fact_app_event
GROUP BY device
ORDER BY number_of_device DESC
```

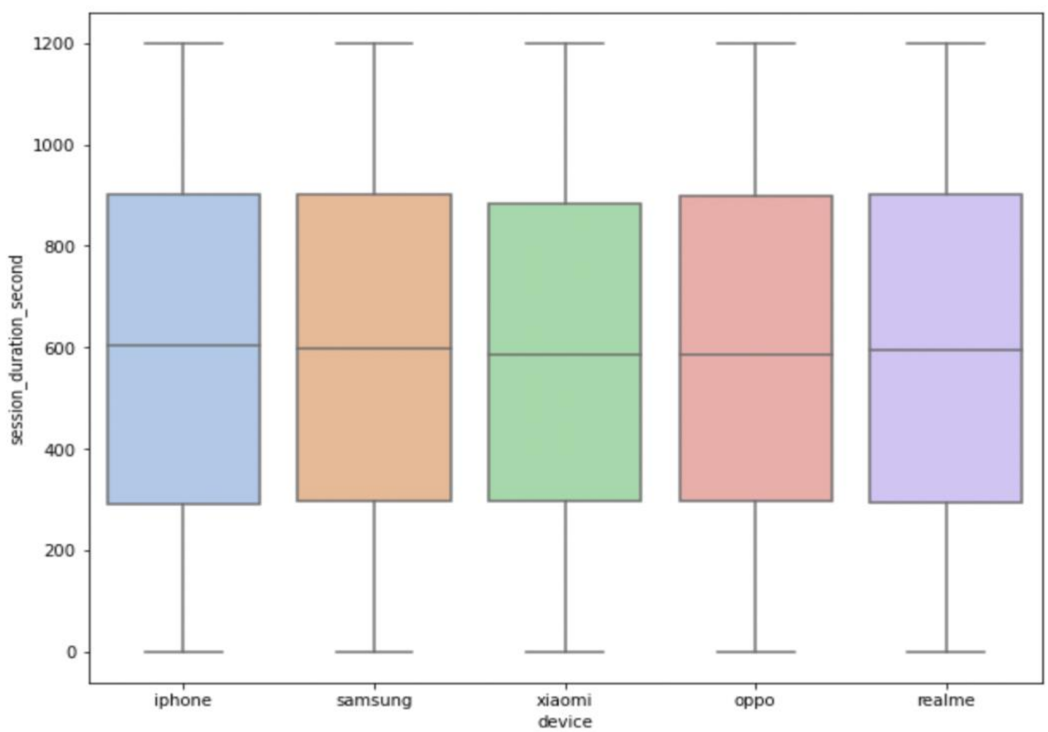
Results Messages

	device	number_of_device
1	samsung	6101
2	realme	6007
3	iphone	5996
4	xiaomi	5955
5	oppo	5941

The result show that top three products are Samsung, realme and iphone. While Xiaomi and Oppo lie at the bottom of table, we curious about it and try to survey the relationship between device and session_duration_second variable:

```
device = fact_app_event['device']
duration = fact_app_event['session_duration_second']
location = fact_app_event['location']
plt.figure(figsize=(10,8))
sns.boxplot(data=fact_app_event, x=device, y=duration, palette="pastel")
```

As the following boxplot, we can see the mean time of Xiaomi and Oppo's session duration is less than the rest.



- Next we want to see the number of clicks and number of impression in each province:

```
SELECT province,
SUM(click) AS num_of_click,
SUM(impression) AS num_of_impression
FROM Fact_app_event
GROUP BY province
```

	province	num_of_click	num_of_impression
1	dong_nai	27013	87746
2	ha_noi	27755	88680
3	binh_duong	28395	91520
4	ben_tre	27564	89083
5	hcm	27006	87031
6	long_an	26916	87275

Calculate the proportion between number of clicks and number of impressions for each province, we get the result:
Dong Nai: 0,30785 ; Ha Noi: 0,31298 ; Binh Duong: 0,31026 ; Ben Tre: 0,30942 ;
TP HCM: 0,31030 ; Long An: 0,30840

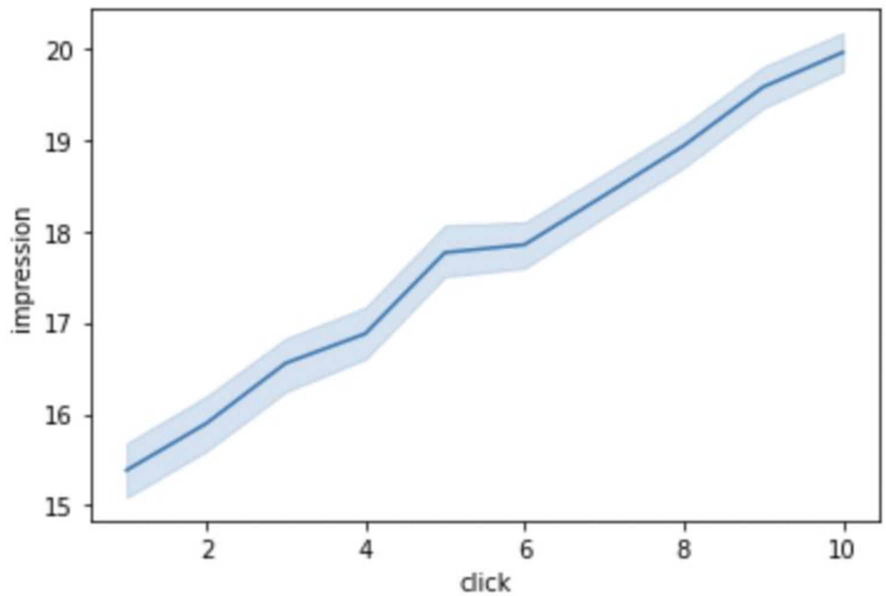
If we noticed, HCM city get the least of number of impression and number of clicks but it still gets very high proportion of people care about product and click. Beside Hanoi the province get the highest social interaction and impression, Binh Duong is quite a potential location to focus on. In addition, money spent advertising in Hanoi and HCM city should be maintained and pouring into other provinces such as: Dong Nai, Long An and Ben Tre in the future

- Then we are curious about Binh Duong so we want to see what is device brand is favored the most by people in Binh Duong. The result is Xiaomi so we can purchase it more for this province. Of course, we can apply the same to other provinces.

```
SELECT device, COUNT(*)
FROM Fact_app_event
WHERE province = 'binh_duong'
```

	device	(No column name)
1	xiaomi	1044
2	oppo	1023
3	realme	1007
4	samsung	1055
5	iphone	994

- We also see that the number of click increases in proportion to the number of impressions:



Plotted by seaborn library in python

Social_media **(Tasks are excuted in Jupyter Notebook integrated in Anaconda ecosystem by Python 3)**

- First, the necessary libraries are imported:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

- The excel file is converted into csv and imported:

```
social_media = pd.read_csv('social_media.csv')
social_media
```

	No.	Type of Post	Date	URL	Reach	Engagement
0	1	Celebrate National Holiday	2022-09-02	https://www.facebook.com/sobanhang.vn/posts/pf...	323	13
1	2	Minigame	2022-09-06	https://www.facebook.com/sobanhang.vn/posts/pf...	122	10
2	3	Go-to-market PRO	2022-09-05	https://www.facebook.com/sobanhang.vn/videos/s...	168	8
3	4	Teasing new feature	2022-09-03	https://www.facebook.com/sobanhang.vn/posts/pf...	414	15
4	5	Viral post	2022-09-07	https://www.facebook.com/sobanhang.vn/posts/pf...	700	52

- For the convenience when calling the columns in data frame, we assign them into variables. Because the range of Reach quite high comparing to Engagement, it is normalized and assigned into variable log_reach:

```
log_reach = np.log(social_media['Reach'])
reach = social_media['Reach']
engagement = social_media['Engagement']
type_of_post = social_media['Type of Post']
```

- The means of Reach and Engagement are calculated with results are 345.4 and 19.6, respectively.

```
mean_reach = social_media['Reach'].mean()
print(mean_reach)
mean_engage = social_media['Engagement'].mean()
print(mean_engage)
```

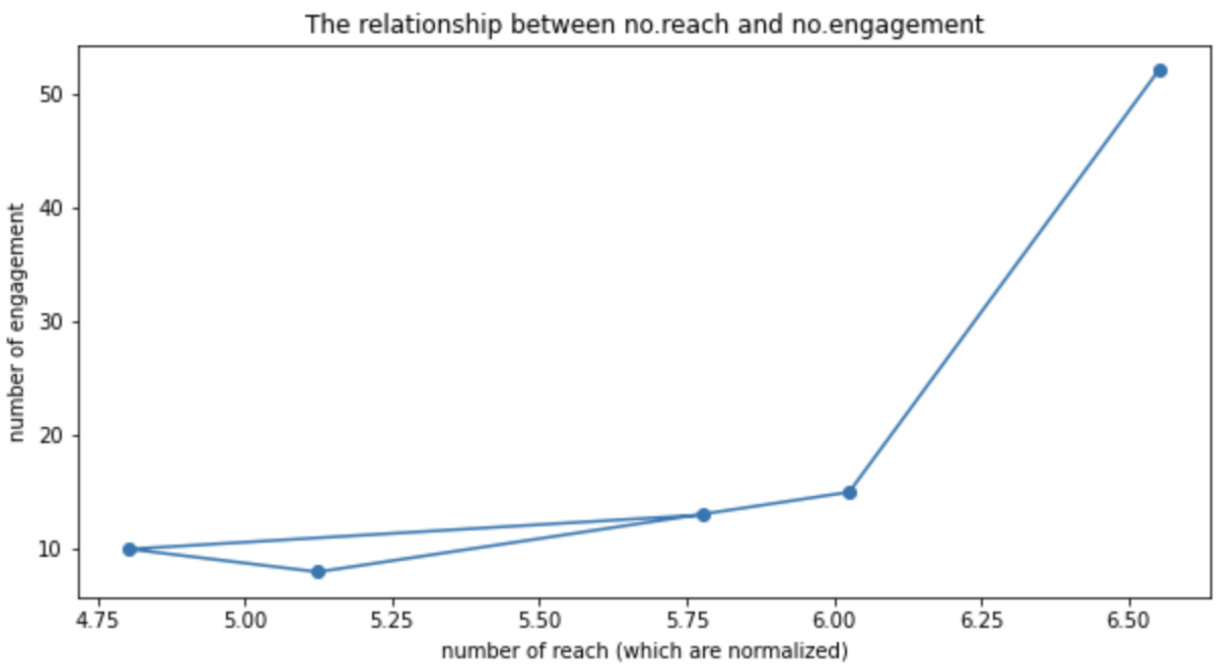
- The proportion between engagement and reach is calculated
  percent_engage = (mean_engage/mean_reach)*100

  The proportion is 5.67% => It seems to be the low probability that customers engage in advertisements.

- The relationship between number of Reach and number of Engagement is inspected:

```
plt.figure(figsize = (10,5))
plt.scatter(log_reach, engagement)
plt.plot(log_reach, engagement)
plt.title('The relationship between no.reach and no.engagement')
plt.xlabel('number of reach (which are normalized)')
plt.ylabel('number of engagement')
plt.show()
plt.clf()
```

The data show that the more reach the audience get, the more engagement will occur except in type of Go-to-market, the number of reaches is slightly higher than minigame, however the engagement is unexpectedly less than minigame's. It may prove that the advertising campaign have been effective.



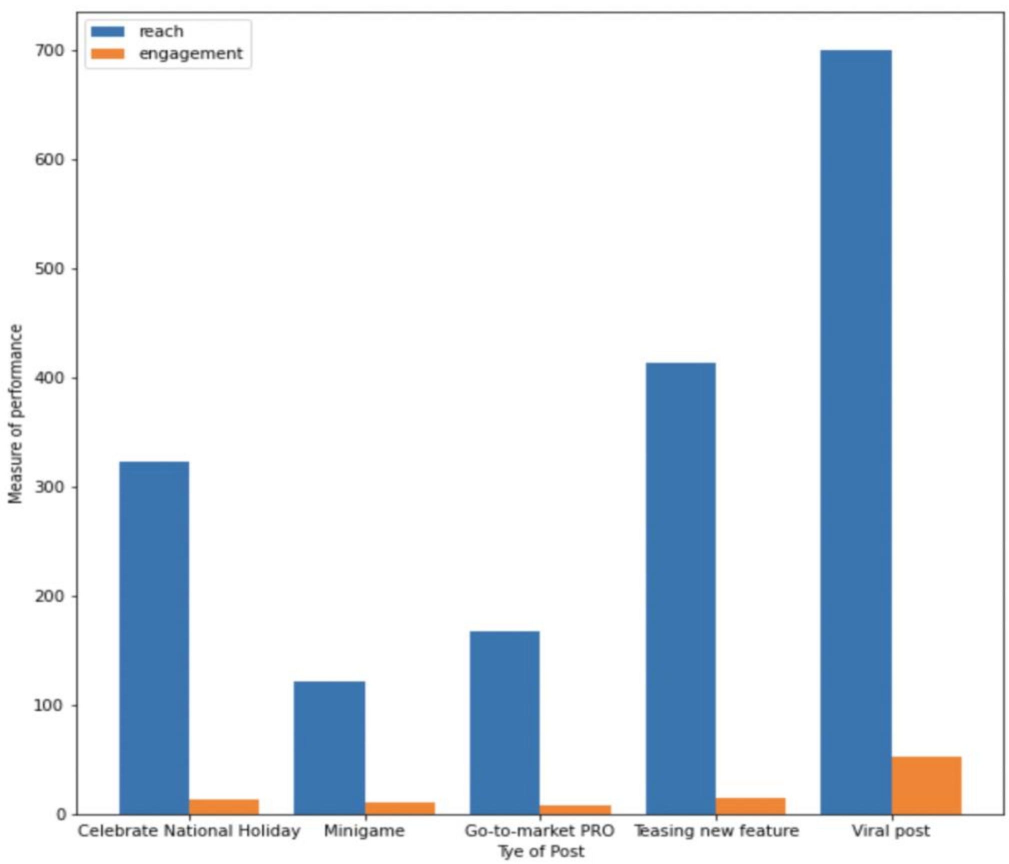
- We try to see it in bar chart whether it can provide a new insight:

```
plt.figure(figsize=(10,10))
n = 1
t = 2
d = 5
w = 0.8
x1 = [t*element + w*n for element in range(d)]
bar1 = plt.bar(x1, reach)

n = 2
x2 = [t*element + w*n for element in range(d)]
bar2= plt.bar(x2, engagement)

plt.xlabel('Tye of Post')
plt.ylabel('Measure of performance')
plt.legend((bar1,bar2),('reach', 'engagement'),loc='upper left')
ax=plt.subplot()
ax.set_xticks((np.array(x1)+np.array(x2))/2)
ax.set_xticklabels(type_of_post)
plt.show()
```

- Overall, you can see the type get the most reach and engagement is Viral post and the post get smallest number of reaches is minigame. However, it is not the type get the smallest number of engagements. Therefore, it is potentially deserved to be invested more in advertisement.



Dim_user_profile (Tasks are executed in Azure Data Studio)

- Let’s inspect what kinds of products are consumed the most.

```
SELECT main_category, COUNT(*) AS no_times
FROM dim_user_profile
GROUP BY main_category
ORDER BY no_times DESC
```

	main_category	no_times
1	thuc_pham	1693
2	my_pham	1690
3	f&b	1686
4	khac	1651
5	thoi_trang	1651
6	tap_hoa	1629

In main category: obviously, daily products such as food and drink invade the most. Besides, the needs of making beautiful like “my_pham” is also important.

- How about secondary category:

```
SELECT secondary_category, COUNT(*) AS no_times
FROM dim_user_profile
GROUP BY secondary_category
ORDER BY no_times DESC
```

secondary_category	no_times
thuc_pham	1746
thoi_trang	1678
tap_hoa	1675
f&b	1662
my_pham	1623
khac	1616

The leading is still “Thuc Pham”, the second one is fashion products

- How about smart phone brands:

```sql
SELECT device, COUNT(*) AS no_times
FROM dim_user_profile
GROUP BY device
ORDER BY no_times DESC
```

	device	no_times
1	samsung	2038
2	oppo	2023
3	xiaomi	2007
4	realme	1967
5	iphone	1965

It appears that Samsung get the bright view in the market city but not Iphone although it is not much different in quality. What causes the difference? We can guess by survey the relationship between device and other variables.

- Let's try with main_category:

```sql
SELECT main_category, COUNT(*) AS no_times
FROM dim_user_profile
WHERE device = 'samsung'
GROUP BY main_category
ORDER BY no_times DESC
```

	main_category	no_times
1	khac	361
2	tap_hoa	358
3	thoi_trang	351
4	thuc_pham	333
5	my_pham	327
6	f&b	308

```sql
SELECT main_category, COUNT(*) AS no_times
FROM dim_user_profile
WHERE device = 'iphone'
GROUP BY main_category
ORDER BY no_times DESC
```

	main_category	no_times
1	thuc_pham	370
2	f&b	351
3	tap_hoa	321
4	khac	319
5	my_pham	309
6	thoi_trang	295

We see customers concern to buy Samsung or Iphone have a tendency to view dairy products like foods and drinks first, then the beauty (we can see clearly in the Iphone table). While customers concern the smartphone brands which belonging lower average price class (oppo, realme) have tendency noticing more the beauty products than foods and drink.

```sql
SELECT main_category, COUNT(*) AS no_times
FROM dim_user_profile
WHERE device = 'oppo'
GROUP BY main_category
ORDER BY no_times DESC
```

	main_category	no_times
1	my_pham	372
2	f&b	354
3	thuc_pham	350
4	khac	332
5	thoi_trang	316
6	tap_hoa	299

```
SELECT main_category, COUNT(*) AS no_times
FROM dim_user_profile
WHERE device = 'realme'
GROUP BY main_category
ORDER BY no_times DESC
```

	main_category	no_times
1	my_pham	354
2	thoi_trang	348
3	f&b	338
4	tap_hoa	321
5	khac	306
6	thuc_pham	300

- Another special point is people often spent more money for high-class smart phone like Iphone at the beginning of year (January) but the hardly in the previous month (December)

```
SELECT device, COUNT(device) AS no_times
```

```sql
FROM dim_user_profile
WHERE registered_date LIKE '%/12/%'
GROUP BY device
ORDER BY no_times DESC
```

	device	no_times
1	oppo	262
2	samsung	262
3	xiaomi	255
4	realme	240
5	iphone	229

Iphone performance in December

```sql
SELECT device, COUNT(device) AS no_times
FROM dim_user_profile
WHERE registered_date LIKE '%/01/%'
GROUP BY device
ORDER BY no_times DESC
```

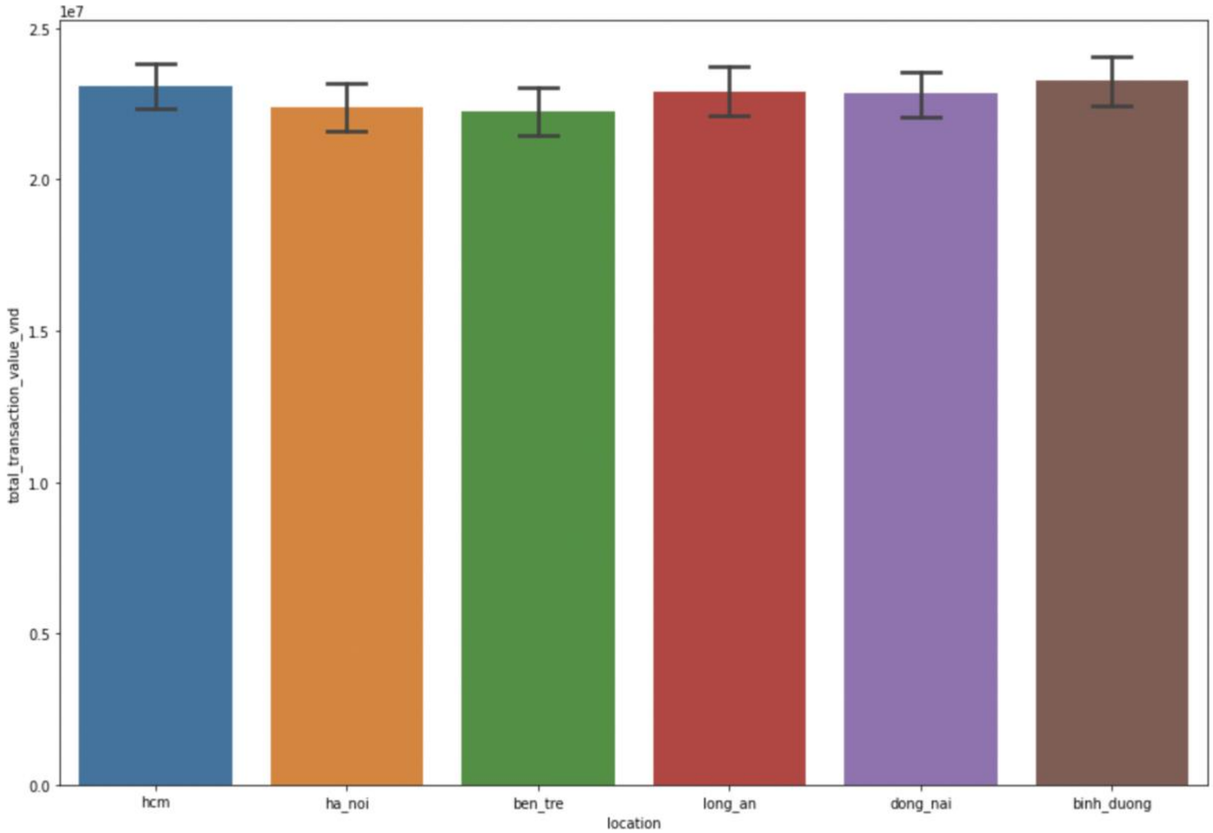
	device	no_times
1	iphone	267
2	realme	264
3	samsung	247
4	oppo	244
5	xiaomi	234

Iphone performance in January

- Especially, data's sale on electronic device does not appear from February, it incidentally at the same time after Tet holiday period in which the consumption volume is very high.

It can be predicted that people try to save money or compensate for New Year expenses.

- Finally, we survey what is province is the market get the most value:



From the bar chart above, we can see the East-South provinces (HCM, Binh Duong, Dong Nai) get the most total transaction value while the province Ben Tre in West region are at the lower position.