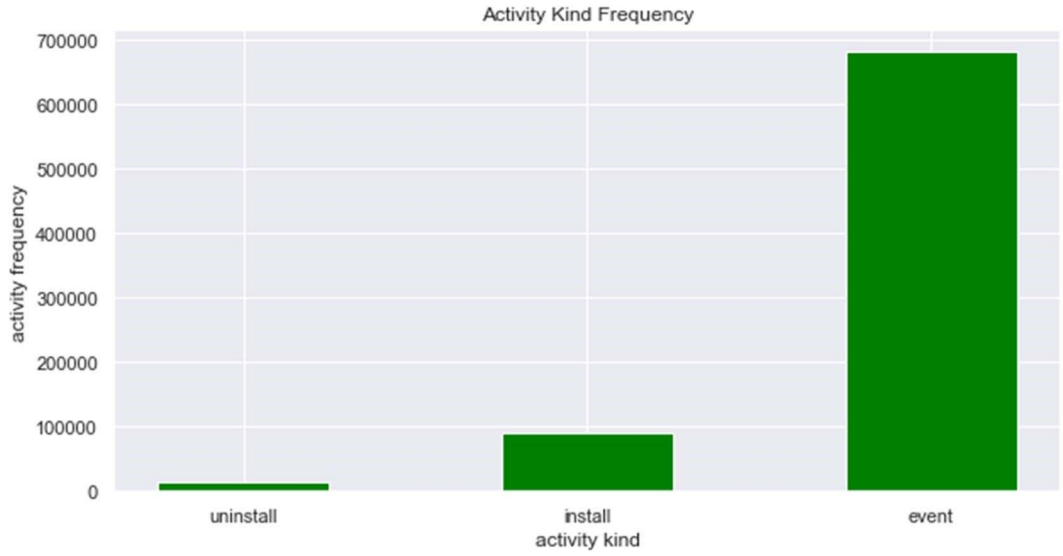
Cleaning data

In the initial step, we review the data to identify missing values and outliers. Outliers can be helpful in specific analyses, but not all should be retained. Additionally, we set aside any activities before installation, except those that may result from data issues. We then proceed to examine the remaining columns in the data discovery process.

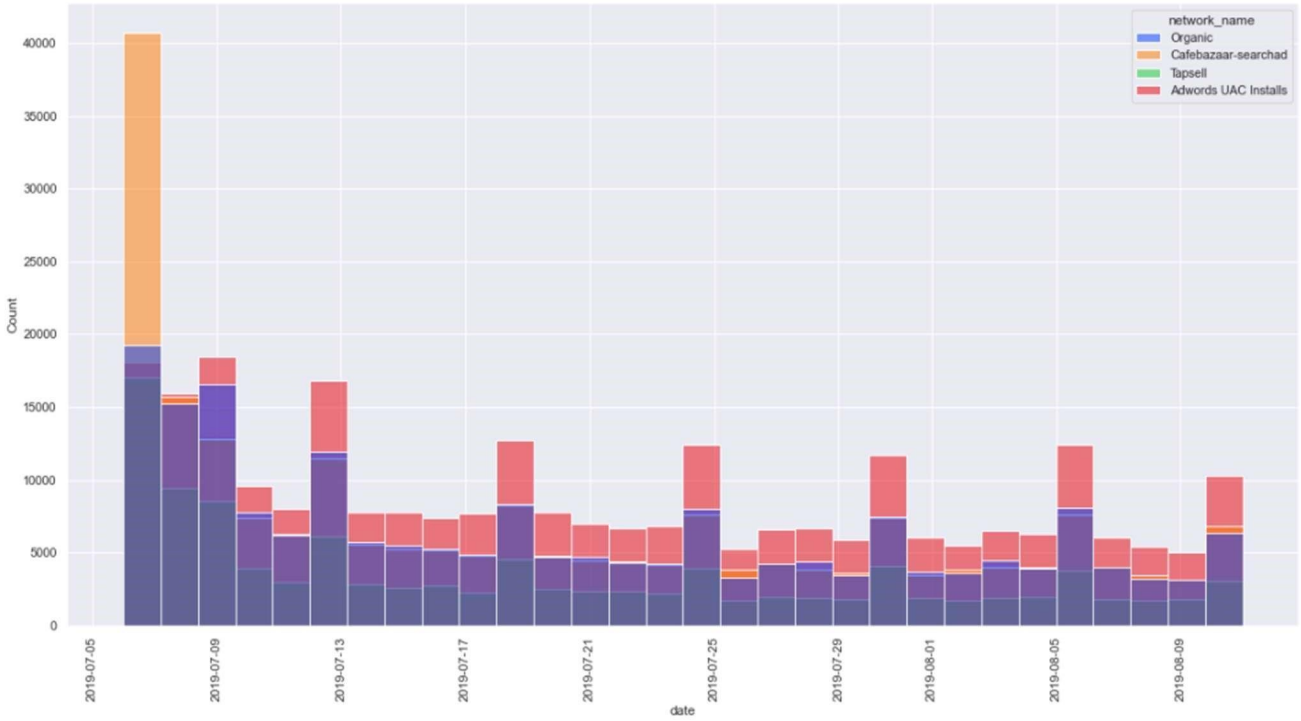
It's helpful to examine all the columns and their relationships to better understand the data. The first column we discussed was the type of activity column to compare the different activities in this month.

EDA

The bar chart shows that almost 100,000 activities were installed this month, in addition to approximately 680,000 events from users installed in previous months. However, this ratio is lower than expected, as some circumstances, such as searching in the app, viewing ads, and editing posts, may not be included in this data. It's important to note that these are just a few limited events we are examining.



Histogram diagram of event time



When examining the histogram, it is noticeable that Cafe Bazar is the primary source of events during the beginning of each month or period. This may be due to the varying behavior of users on weekdays versus weekends, resulting in a somewhat sinusoidal pattern. On average, the Adwords source has more events than other networks.

Activity diagram and type of activity done

In the innermost layer, we have the networks. In the next layer, each of them has a selection from each activity, and in the next stage, we said which event was the selection from these activities.

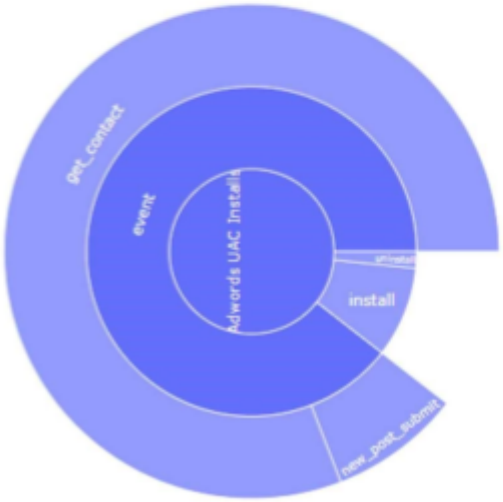
The number of organic and Cafebazaar networks seems almost the same, but significant differences can be seen in Tapsell and AdWords.

In the next step, we will look at these networks separately.

Network and shares

We can see that, as we saw in the bar plot, the share of the activities of organizing events and getting contact numbers in the events is the highest, so the difference with posting a new post is several times.

The share of calls in Tapsal seems less than the others, and the others are closer.



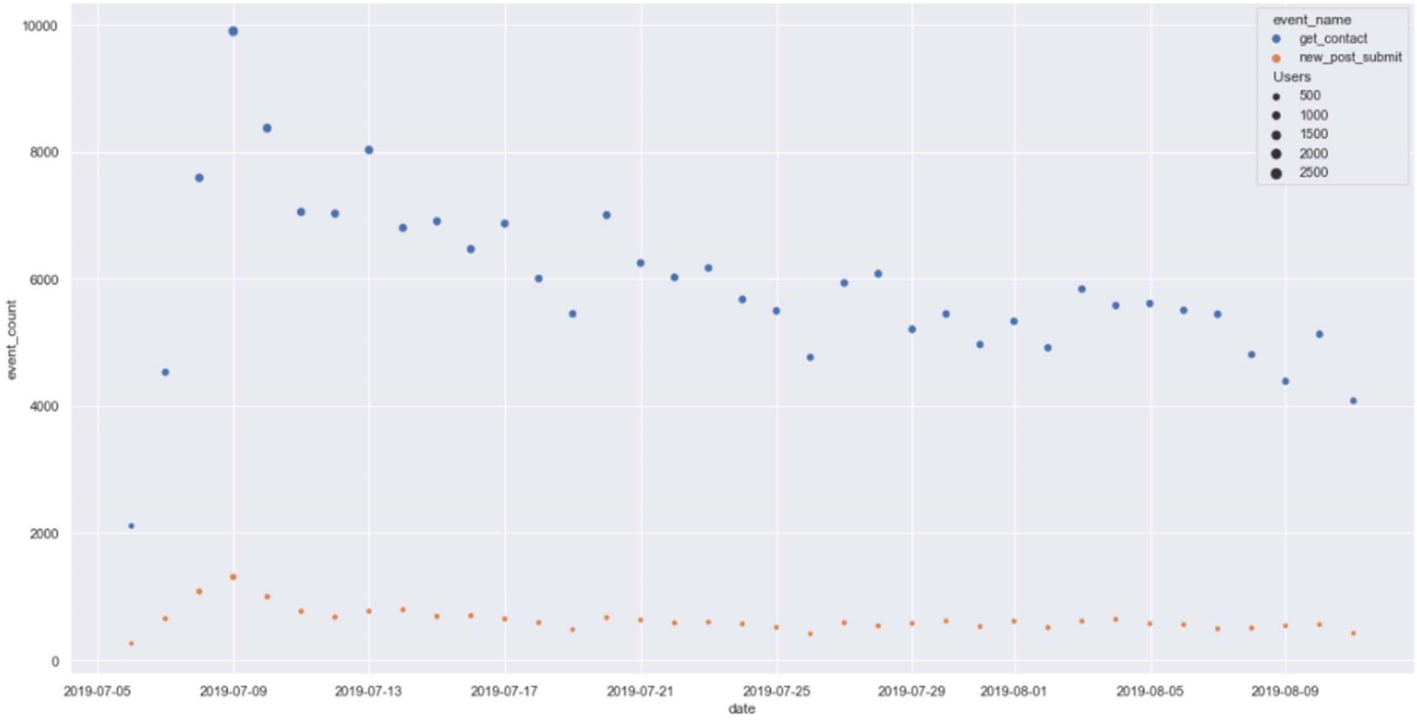
Now that I understand how the different events are, the number of them is almost the same. Can you give us more information about the events?
Scatter diagram of the number of events per user per day
The events of calling and posting are shown in two different colors because the type of event is different; just as it is clear in my graph, the number of receiving calls is multiple, because the behavior of the ad viewer and advertiser is different.

The size of the circles is related to the number of users, so the bigger it is, the more users perform this action, as we see at the beginning of the interval. The height shows the total number of moves.

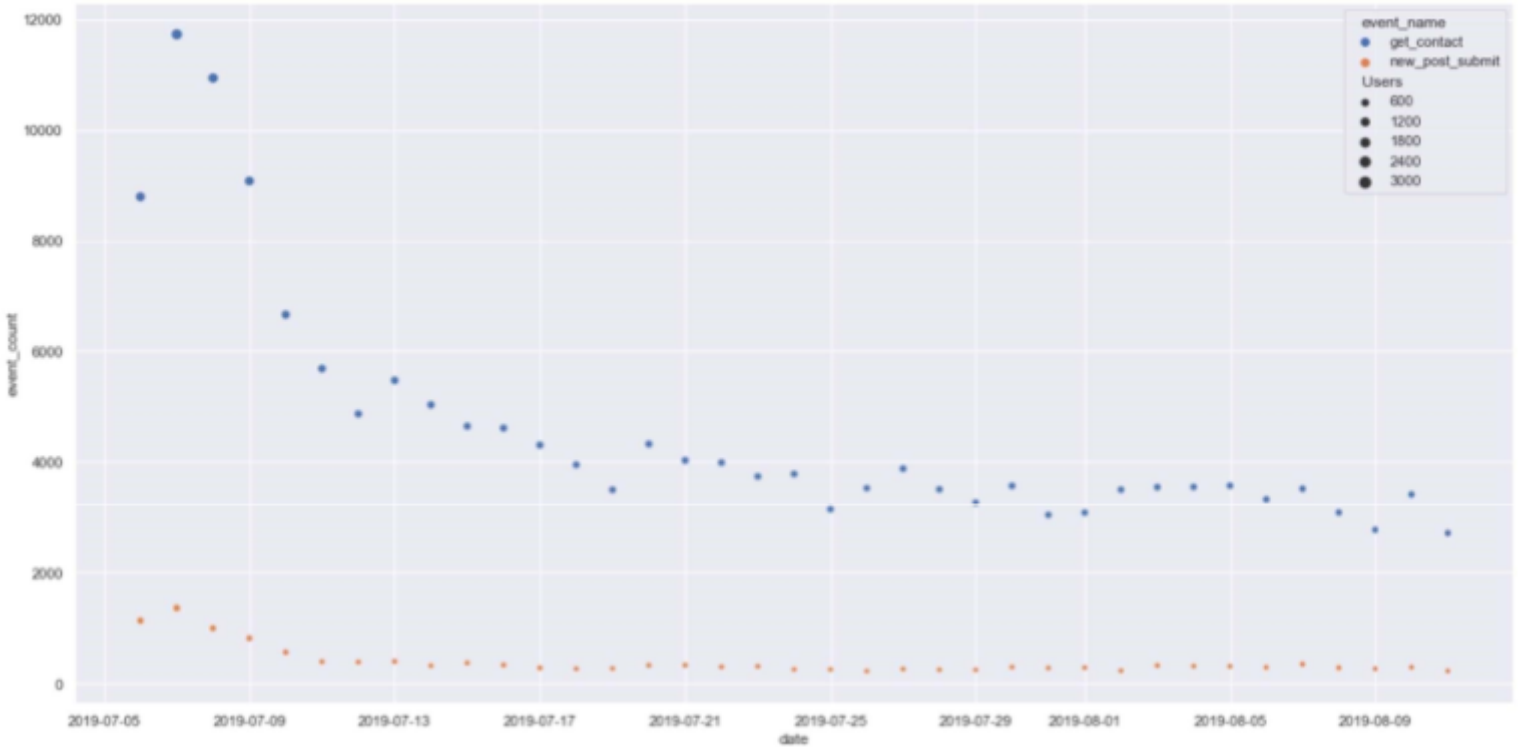
Now that we have checked the first chart, we should know that this chart was imported from the user's Organic source. At the beginning of the chart, both the actions of calling and posting are at their maximum, and we see a trend in it over time, which is related to the days of the week. Another thing is that we can see that the range of different networks is different; for example, for Organic, it has gone up to about 900, and the rest are different.

The calls started from 2000, and at 4000, there is a balance, and the sinuses increase and decrease except at the chart's beginning. Posting a new post below 2000 and even in a balanced manner, except for the first peak of the chart, is still below 1000.

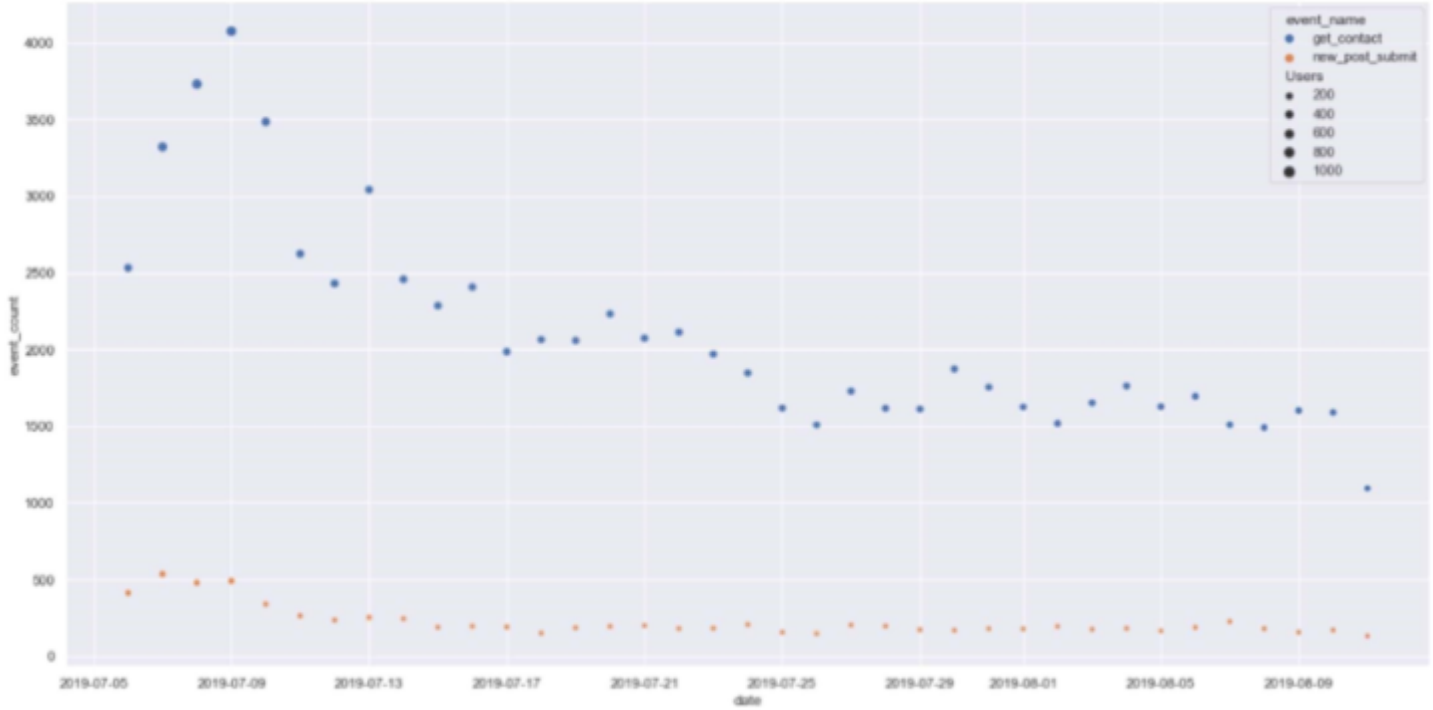
The graph below comes from the Awards source; as I mentioned, the interval is different. I went up to 10,000 to get a call, but the interval for posting a new post seems the same. And the number of my users has gone up to 2500.



In Cafe Bazar, this number reaches 12,000 and the number of users reaches 3,000.



In Tapsal, the number is the lowest, up to 4,000, and the maximum number of posts has reached almost 500.



As we have seen, the event data in the first interval behaves differently from the rest of the gap, which suggests that we need more data to confirm whether these may differ in each category.

Part 1:

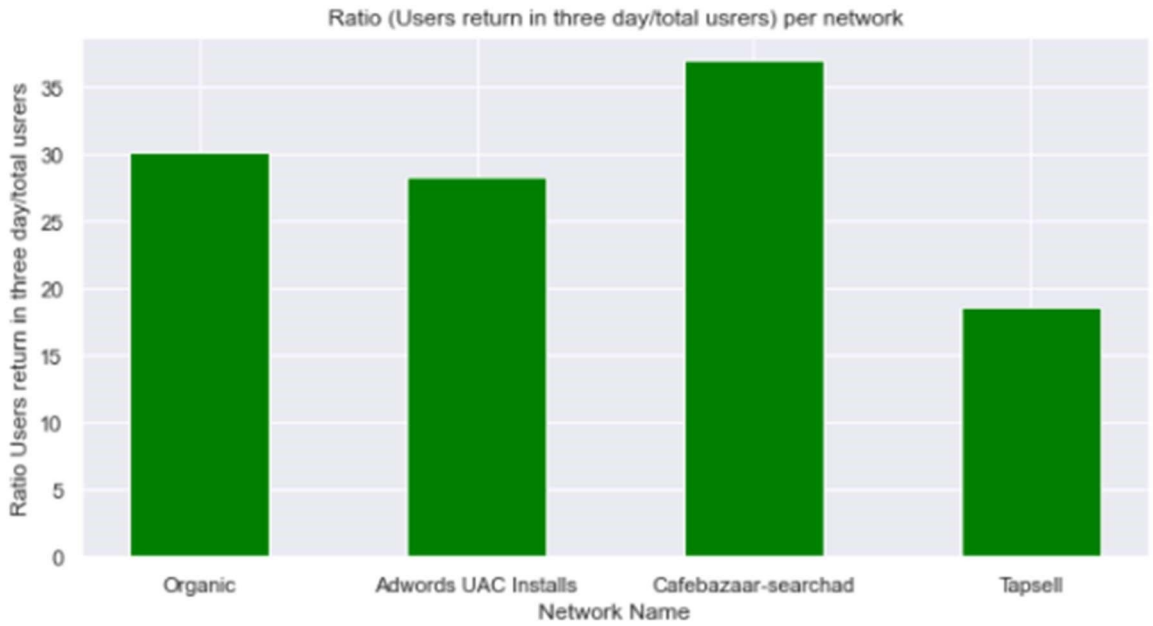
The percentage of users who finally performed an essential event within three days after installation.

If I consider this user as a particular segment of users, it can show that this type of user in Tapsel has a significant difference from the rest.

network_name	users_return_in_three_day	total_users	percentage_users_return_in_three_days_total
Organic	6866	22782	30.14
Adwords UAC Installs	7549	26662	28.31
Cafebazaar-searchad	9770	26451	36.94
Tapsell	3747	20209	18.54

(همینطور ي توي توتال متفاوتند)

که خب توي نمودار میله اي اینو بهتر میتونیم ببینیم:



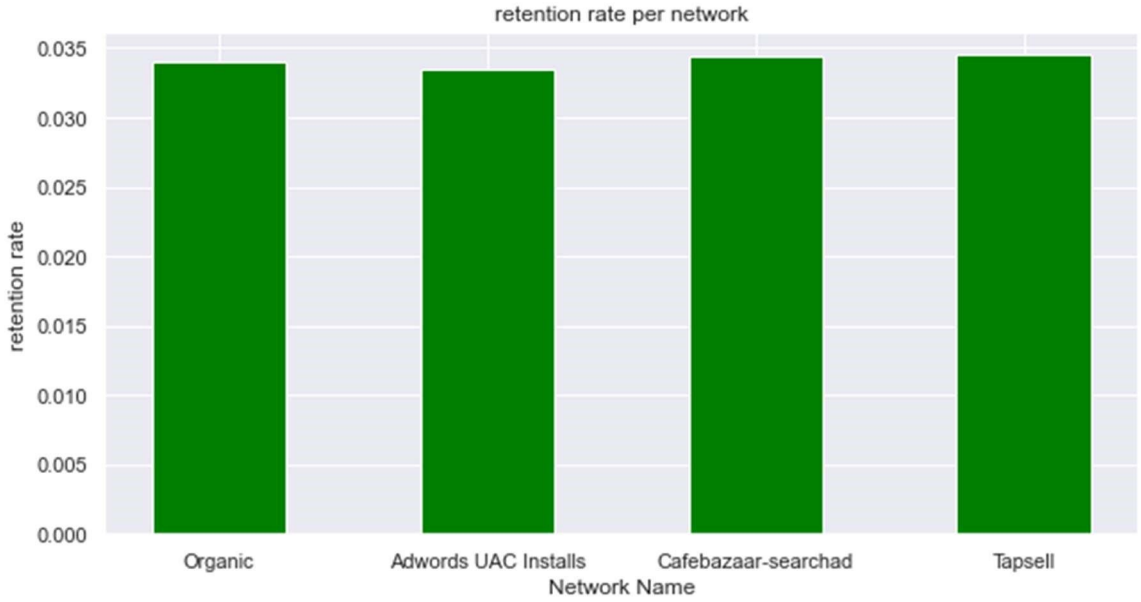
In Cafebazaar, the percentage of this type of user is more than others.

The return rate of these users: Users of any network in these 30 days can do an event comparable to the month. How many days of the month did each of them do the event? And the difference is not different for every network.

This shows that, on average, in each network, users act in one day out of 30 days. Users who had the previous condition, that is, in the first 3 days after installation, did an event.

network_name	retention_rate
Organic	0.034
Adwords UAC Installs	0.0335
Cafebazaar-searchad	0.0344
Tapsell	0.0345

In the bar graph, we can see that the networks are slightly different.



Part 2

In this question, we want to find fraud, which depends greatly on what we define as fraud. But here, we can talk about strange behavior or fraud with a series of graphs.

In this important question, which side are we on? Do we see the ad and take action, or the advertiser's point of view and post?

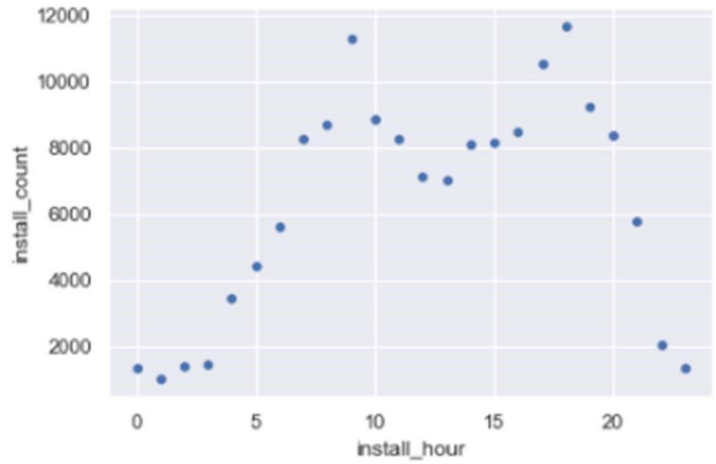
Maybe the trick is for the ad viewer to do a lot of get_contact actions in a short time, but this short time is different for the post, which means it will definitely take more time to post many posts.

Time distribution of installations throughout the day and night

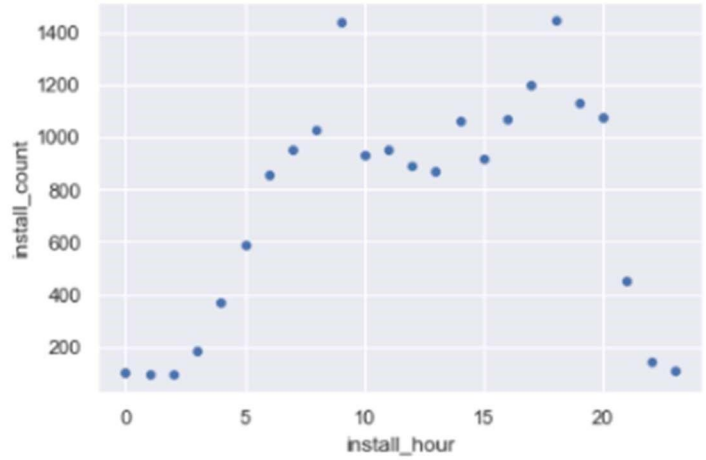
In the Organic network, we have two distribution charts in 24 hours of the day, the upper one for those who had the get_contact event, and the lower one for those who had the new_post_submit event.

which shows that in the hours from 11 p.m. to approximately 6 a.m., the number of installations of both modes is decreasing, it increases from 6 a.m. and reaches its peak at 9 a.m. and 4 p.m.

Farad's possibility to come to the organic network without any advertising is more than anything else.

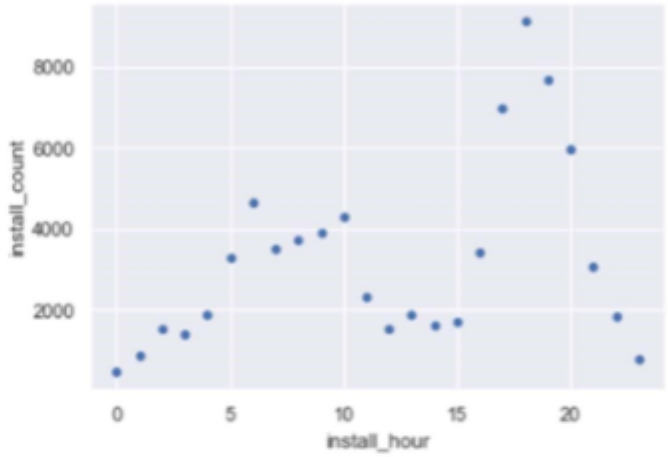



```
[n [24]: distribution_installed_per_hour(network_name='Organic',event_name=
```

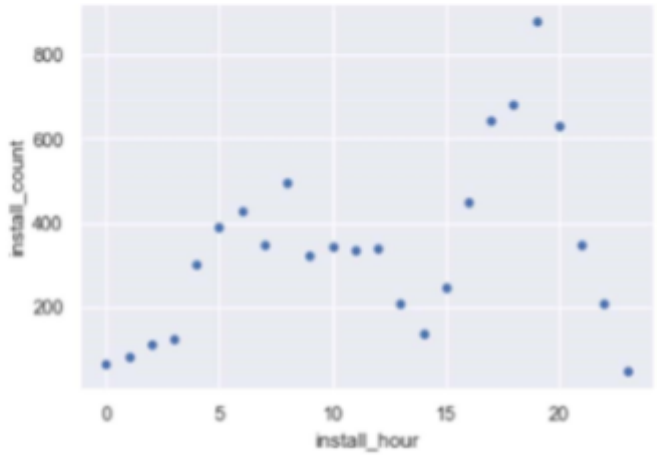


Examining the installation time in Tapsal: We can see that Tapsal has a completely different trend, which is increasing from 14:00 to 5:00 and reaches its peak at 15:00 to 20:00. Transfer to 800 means one-tenth of it.

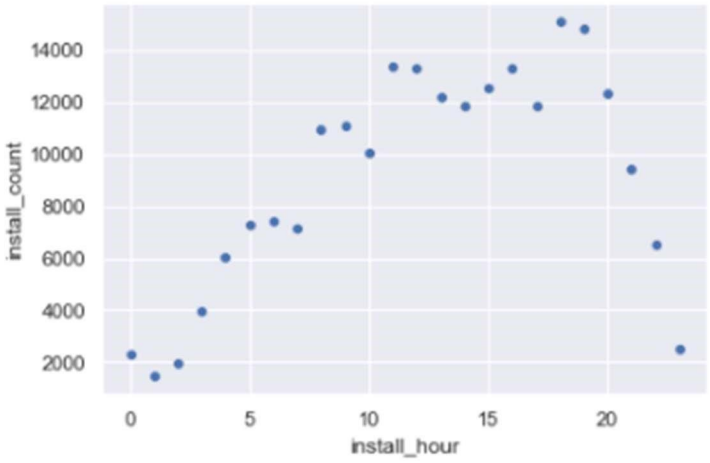
```
In [29]: distribution_installed_per_hour(network_name='Tapsell',event_name='get
```



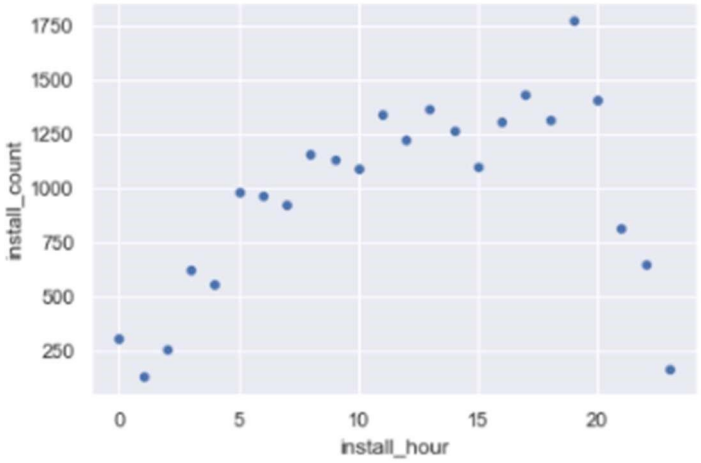

```
In [30]: distribution_installed_per_hour(network_name='Tapsell',event_name='new
```



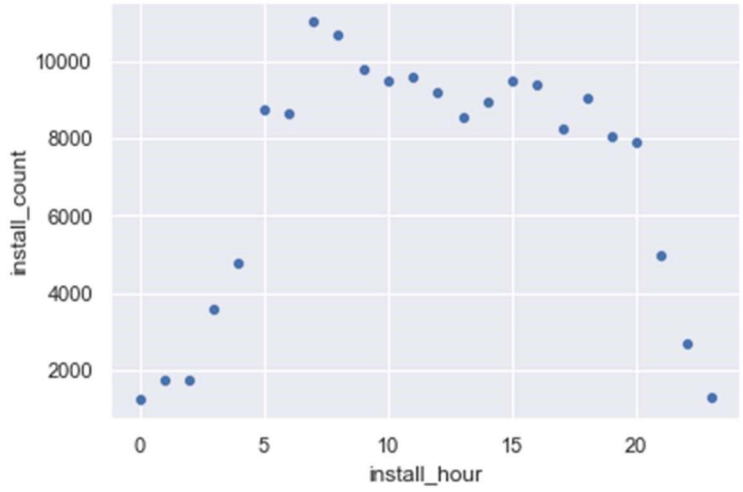
In adwords, the trend goes up to 10, then it reaches its peak in 15 to 20, and then it goes down. Again, the maximum number of ad installs by viewers has gone up to 14,000, but in that one, it has only reached 1,750.



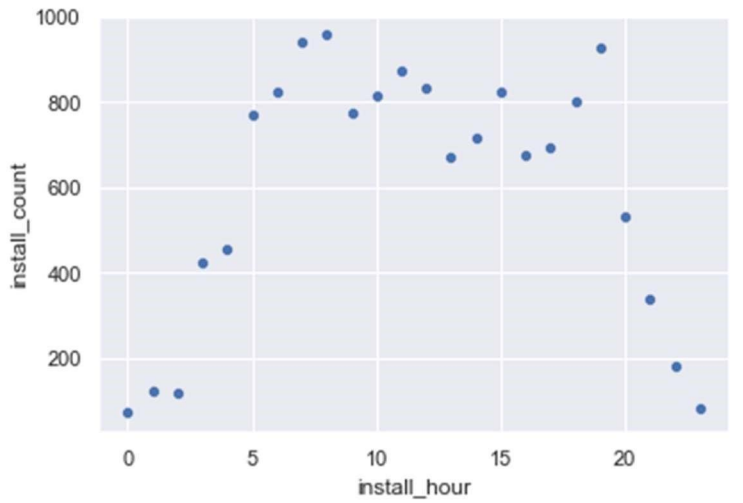
[26]:
```
distribution_installed_per_hour(network_name='Adwords UAC Insta
```



The ones that come from Cafe Bazar rise up to 7:8 and reach their peak between 5 and 10, after which it is almost constant until it decreases from 20 to 24, and the difference in the number of types of events is from 10,000 to 1,000 dice.



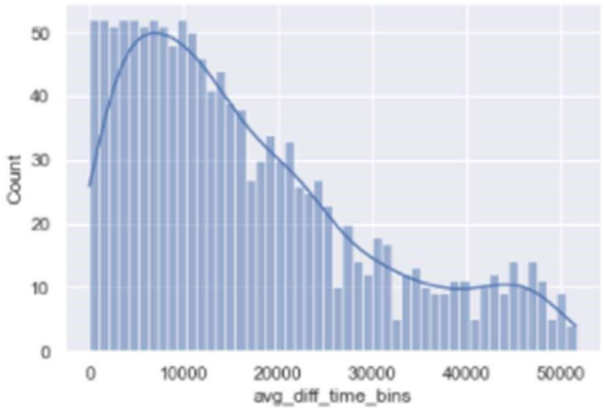

```
]: distribution_installed_per_hour(network_name='Cafebazaar-searc
```



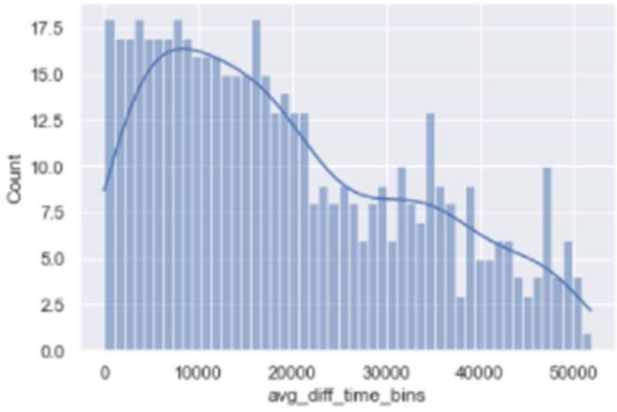
The distance between users' events in the app

For organic network users, the top chart is for the get_contac action and the bottom chart is for another.

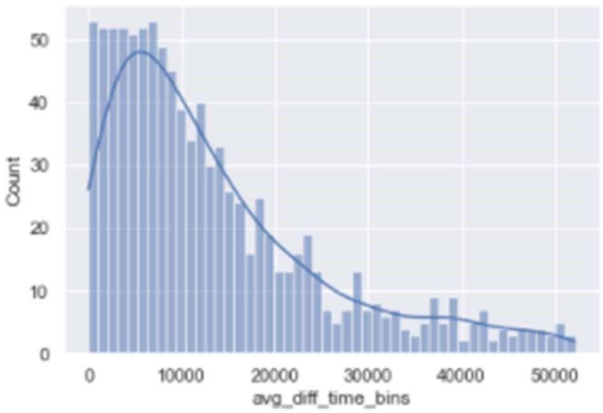
This graph shows the average between 2 user actions and the frequency of each of them. Between 0 and 1000 minutes, more users performed the action, and the greater the time difference, the lower the frequency. In the same way, for submission, with the difference that the frequency of the two graphs is different since this graph shows the average, we can consider farad users who have a significant difference with these graphs.



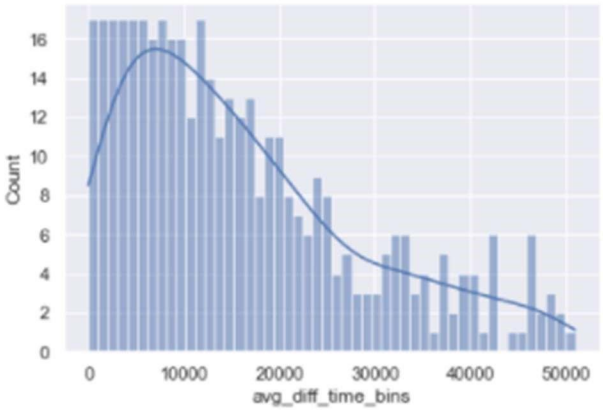
[33]:
```
diff_events_users_hist(
    bins=60,
    network_name='Organic',
    event_name='new_post_submit')
```



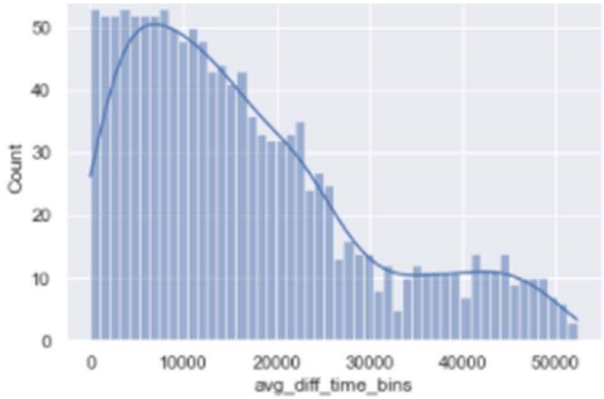
In tapsal, the situation is similar to the above, with the difference that we have a serious difference in abundance, which is logical because the number of tapsals is much less.



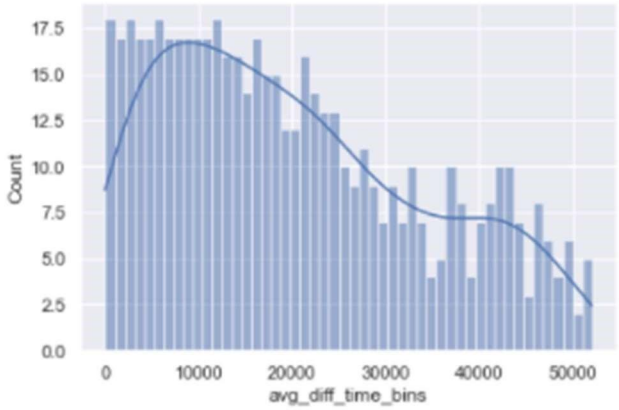

```
n [39]: diff_events_users_hist(
            bins=60,
            network_name='Tapsell',
            event_name='new_post_submit')
```



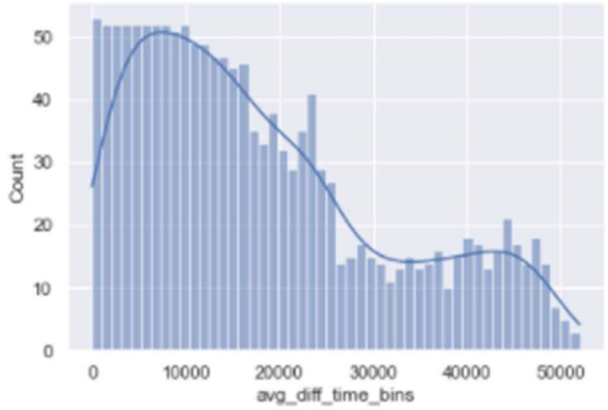
For Cafe Bazar, we differ only in the frequency, but we have the same process



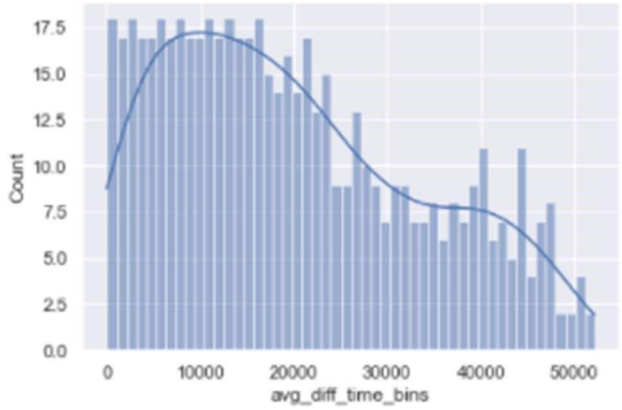
n [37]:
```
diff_events_users_hist(
    bins=60,
    network_name='Cafebazaar-searchad',
    event_name='new_post_submit')
```



It is the same for AdWords with the difference that the frequency is different and the process is easier.



[35]:
```
diff_events_users_hist(
    bins=60,
    network_name='Adwords UAC Installs',
    event_name='new_post_submit')
```



Part 4:

For question 4, because the installation cost is the same for all networkers, so I choose the one that, in addition to producing more value, also increases the ratio to the number of installations, which can be Adwords.

network_name	count_get_contact	count_new_post_submit	install_count	value_network
Organic	152098	17974	22783	145.65
Adwords UAC Installs	218762	23257	24092	187.34
Cafebazaar-searchad	168956	14401	23590	132.67
Tapsell	77525	8172	18917	84.18