Antoine Arnoud - Data Challenge # 2

Yammer

Yammer is a social network for communicating with coworkers. Individuals share documents, updates, and ideas by posting them in groups, it's like Slack. Yammer is free to use indefinitely, but companies must pay license fees if they want access to administrative controls, including integration with user management systems like ActiveDirectory.

Investigating a Drop in User Engagement

The problem

You show up to work Tuesday morning, September 2, 2014. The head of the Product team walks over to your desk and asks you what you think about the latest activity on the user engagement dashboards (yes this really happens). You fire them up, and something immediately jumps out:

The above chart shows the number of engaged users each week. Yammer defines engagement as having made some type of server call by interacting with the product (shown in the data as events of type "engagement"). Any point in this chart can be interpreted as "the number of users who logged at least one engagement event during the week starting on that date."

Question

The head of product says "Can you look into this and get me a summary by this afternoon?" As she runs to a meeting.

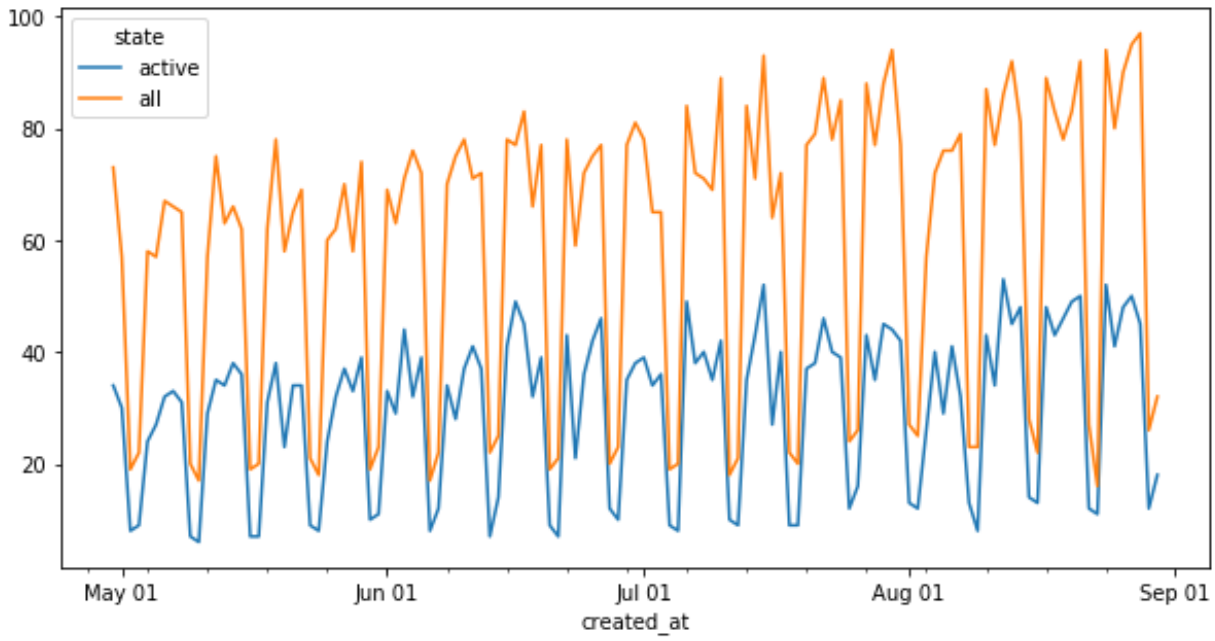
Is the server down?

I first check if new users suscribe. If not, it might mean that the server/app is down.

In [14]:
```
# Libraries
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
```

In [15]:
```
# Data
df_users = pd.read_csv('yammer_users.csv')
# Converting some columns to date format
df_users['created_at'] = pd.to_datetime(df_users['created_at']).dt.floor('1D')
df_users['activated_at'] = pd.to_datetime(df_users['activated_at']).dt.floor('1D')
```

In [16]:
```
# Counting the number of active/pending for each day
df_signup = df_users.groupby('created_at')['state'].value_counts().unstack()
# Add them to get the total nb of signups
df_signup['all'] = df_signup['active'] + df_signup['pending']
# PLot the active and all timeseries, keeping only recent data
df_signup[df_signup.index >= pd.to_datetime('2014-5-1')].plot(y=['active', 'all'], figsize=(10,5))
plt.gca().xaxis.set_major_locator(mdates.MonthLocator())
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%b %d'))
```



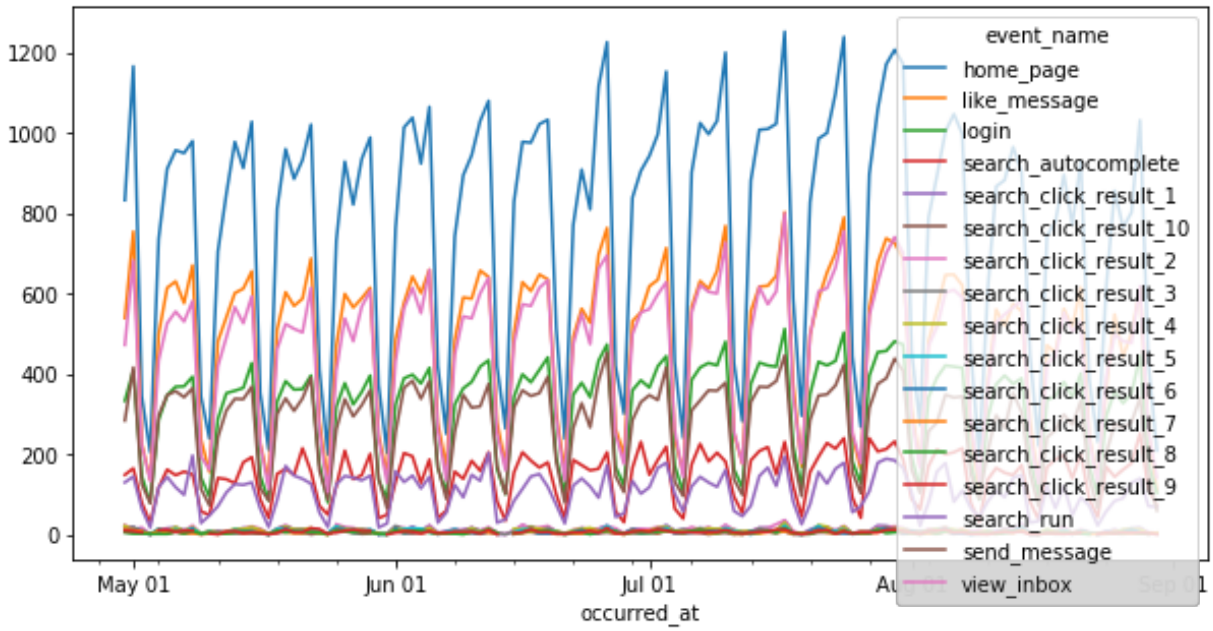
It seems that the server is not down.

Are users less engaged?

I check how much users are using the service.

In [17]:
```
# Loading the events
df_events = pd.read_csv('yammer_events.csv')
# Converting date
df_events['occurred_at'] = pd.to_datetime(df_events['occurred_at']).dt
.floor('1D')
# Keep only engagements
df_events = df_events[df_events.event_type == 'engagement'].drop('even
t_type', axis=1)
```

In [18]:
```
# Finding the nb of event names through time
df_event_name = df_events.groupby('occurred_at')['event_name'].value_c
ounts().unstack()
df_event_name.plot(figsize=(10,5))
plt.gca().xaxis.set_major_locator(mdates.MonthLocator())
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%b %d'))
```



It is not clear that engagement has decreased during the period.

Conclusion

It is not clear why the engagement drops in August. I did not have time but I would like to investigate the engagement by device category to see if this is due to a specific device (computer, phone or tablet).