Data Analysis of 'takehome'

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [3]:
```
# uploading takehome_user_engagement.csv file
engage = pd.read_csv('takehome_user_engagement.csv',parse_dates = ['time_stamp'])
```

In [4]:
```
engage['time_stamp'] = pd.to_datetime(engage['time_stamp'])
```

In [5]:
```
# finding adopter_users from the file
engage['time_stamp'] = engage['time_stamp'].dt.floor('d').astype(np.int64)
engage = engage.sort_values(['user_id', 'time_stamp']).drop_duplicates()

a = engage.groupby('user_id')['time_stamp'].rolling(window=3)
b = pd.to_timedelta((a.max()- a.min())).dt.days
#print (b)
c = b[(b <= 7) & (b > 2)].index.get_level_values('user_id').tolist()
#print(c)

freq = {}
for i in c:
    if i in freq:
        freq[i] += 1
    else:
        freq[i] = 1
k1 = [k for k,v in freq.items()]
print(len(k1))
```

```
1504
```

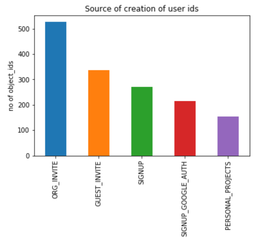
From takehome_user_engagement.csv file, Adopted users found to be 1504

In [10]:
```
# adopted_users dataframe
adopted_users = pd.DataFrame({'user_id':k1})

# uploading takehome_users.csv file
users = pd.read_csv("takehome_users.csv", parse_dates = ['creation_time'],encoding = 'latin-1')

# merging adopted_users dataframe with users dataframe
df = adopted_users.merge(users, how = 'inner' ,indicator=False, left_on = 'user_id',right_on = 'object_id')
df = df.drop('user_id',axis=1)
```

In [11]:
```
# count of users belong to which creation_source
df['creation_source'].value_counts()
```

Out[11]:
```
ORG_INVITE              526
GUEST_INVITE            337
SIGNUP                  271
SIGNUP_GOOGLE_AUTH      215
PERSONAL_PROJECTS       155
Name: creation_source, dtype: int64
```

In [12]:
```
df['creation_source'].value_counts().plot.bar()
plt.ylabel('no of object_ids')
plt.title('Source of creation of user ids')
plt.show()
```



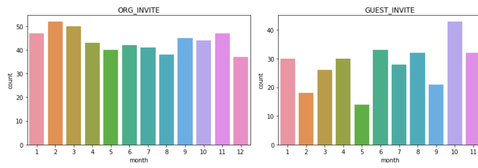
From the above graph, it is found that major source of adopted users is

ORG_INVITE with 526 users,

followed by GUEST_INVITE with 337 users.

Thus, ORG_INVITE and GUEST_INVITE can produce more future adopted users.

In [24]:
```
df['month'] = df['creation_time'].dt.month

import seaborn as sns
fig, ax = plt.subplots(1,2, figsize = (12,4))
sns.countplot(x = df[df['creation_source']=='ORG_INVITE']['month'], data = df , ax = ax[0])
ax[0].set_title('ORG_INVITE')
sns.countplot(x = df[df['creation_source']=='GUEST_INVITE']['month'], data = df , ax = ax[1])
ax[1].set_title('GUEST_INVITE')
plt.tight_layout()
```



From the above graphs we can conclude that from source "ORG_INVITE" majority of users created their account in the month of feb, followed by march.

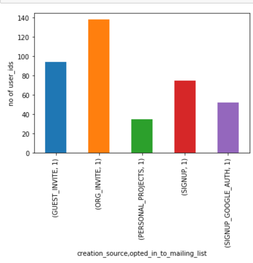
While incase of "Guest_INVITE" over 40 accounts were created in the month of october.

In [25]:
```
# count of opted_in_to_mailing_list
df['opted_in_to_mailing_list'].value_counts()
```

Out[25]:
```
0    1110
1     394
Name: opted_in_to_mailing_list, dtype: int64
```

394 users have opted into receiving marketing mails. From these users over 140 users belong to "ORG_INVITE"

In [27]:
```
df1 = df[df['opted_in_to_mailing_list']==1]
df1.groupby('creation_source')['opted_in_to_mailing_list'].value_counts().plot.bar()
plt.ylabel('no of user_ids')
plt.show()
```



In [31]:
```
# count of enabled for marketing drip
df['enabled_for_marketing_drip'].value_counts()
```

Out[31]:
```
0    1268
1     236
Name: enabled_for_marketing_drip, dtype: int64
```

236 users have enabled for marketing drip. From these users over 80 users belong to "ORG_INVITE"

In [28]:
```
df['enabled_for_marketing_drip'].value_counts()
df1 = df[df['enabled_for_marketing_drip']==1]
df1.groupby('creation_source')['enabled_for_marketing_drip'].value_counts().plot.bar()
plt.ylabel('no of user_ids')
plt.show()
```