Relax Data Science Challenge

The column active was created and populated with the code. For a given user_id, every date in user_engagement, is checked id there are atleast than 3 entries in the next 7 days. If yes, the that user is considered active which is marked as 1.

```
1 df_user['active'].isnull().sum()
```

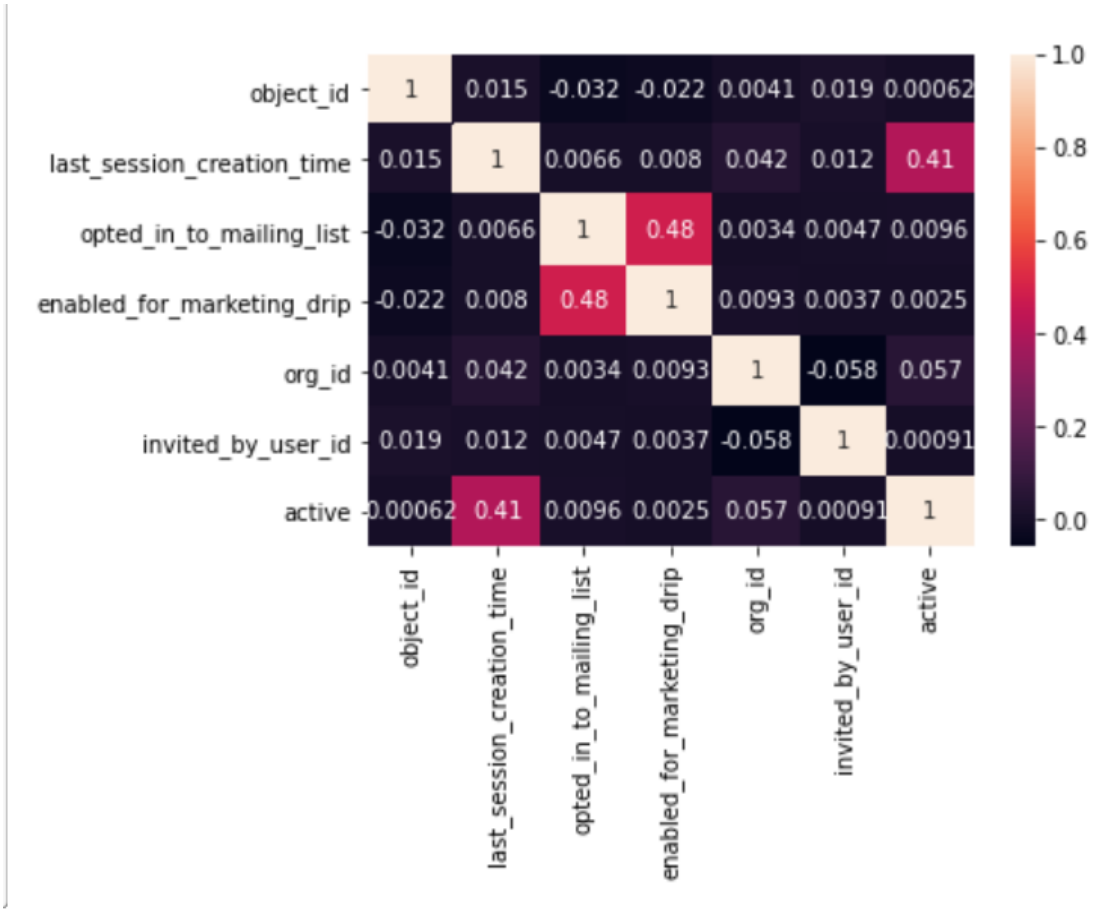
```
3177
```

There are 3177 users who just signed up and did not use the service

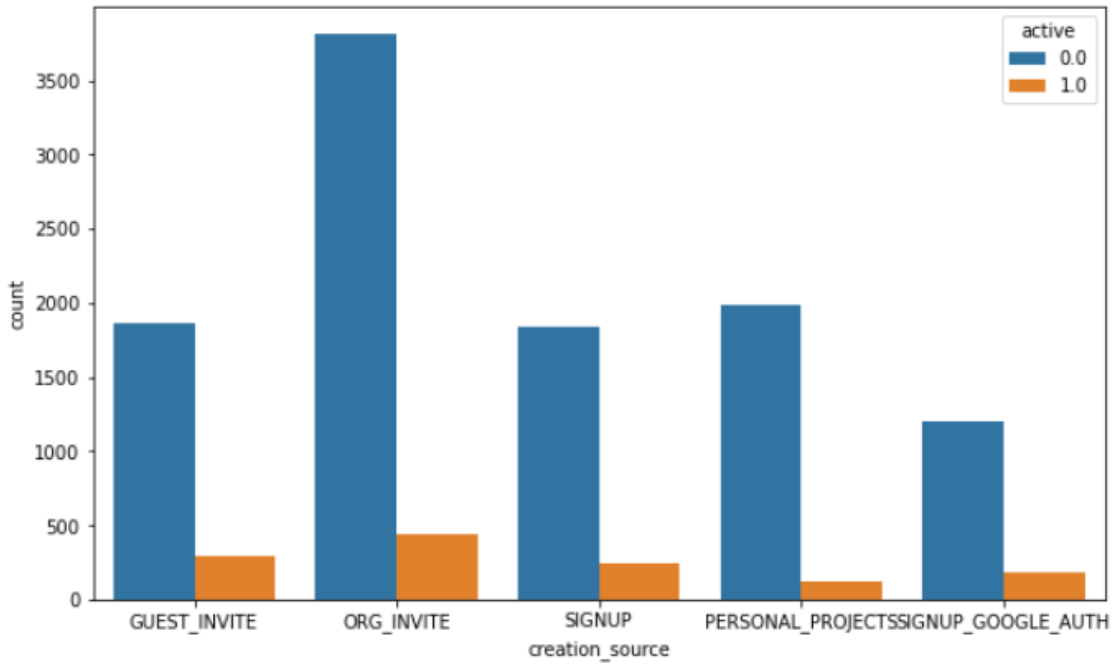
```
1 (df_user['active'].value_counts()/df_user.shape[0])*100
```

```
0.0    89.191667
1.0    10.808333
Name: active, dtype: float64
```

Highly skewed. Very small % of active users



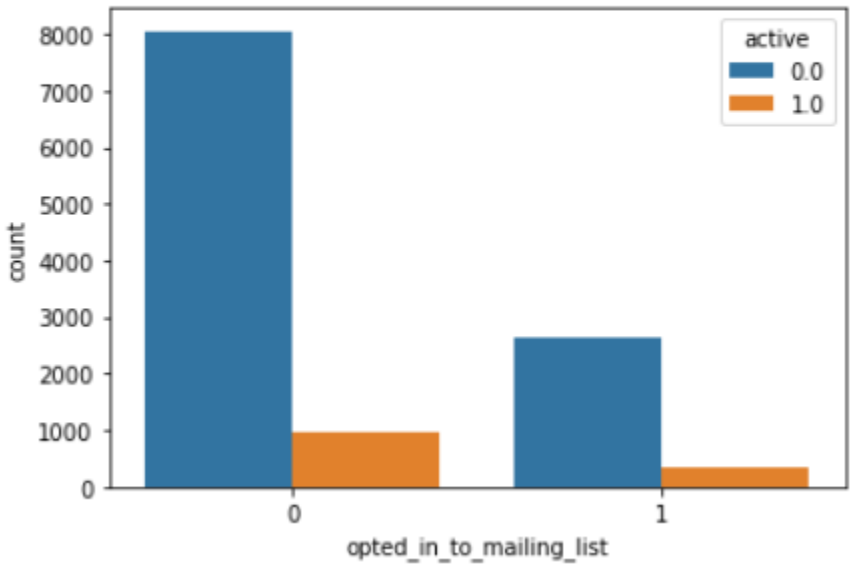
Not much correlation beween variables.



More users were invited by organizations as a full member. Useful for marketing.

```
1 sns.countplot('opted_in_to_mailing_list', hue='active', data=df_user)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1d2d66c4cd0>
```



Future Research:

Modelling can be done using SMOTE for oversampling. Feature importance will give better results.

.