INSIGHTS :

1. Activity_data is a lot larger than user_data
2. It is user activity data for 7 days
3. A lot of device_type values were Nan, so, we had 3 options -
    a. Delete entire column
    b. Replace Nan with mean
    c. Replace Nan with median

    We replaced it with the medium as it seemed to be the best approach out of the three.
4. user_activity_type_id - 16 heavily dominates the user_activity_type_id list, so, the entire Dataset is skewed in its favour.
5. After 16, 12 and 11 are the major affecting values.
6. Heatmap doesn't give us any important insights.
7. If a user has registered for the service/platform, then there is a high chance that they will Come back to explore it further. This is clearly shown in scatter plot (IN [51] : may change serial no. with time).
8. After registering for the service/platform, the user spends considerable amount of time On it/comes back as it can be seen in our plots on (IN [58] : may change serial no. with time).
9. The most common choice of device type for our users is device_type 1, followed by 4th.

MODELS:

1. Logistic Regression Model:
   Accuracy : 94.46%
   Status : Good. Can be improved.

2. Decision Tree Classifier:
   Accuracy : 100%
   Status : Possible case of overfitting.

3. Decision Tree Regressor:
   Accuracy : 100%
   Status : Possible case of overfitting.

4. KMeans clustering:
   Status : no insights as of now, need more logical parameters as input.

WHAT MORE CAN BE DONE:

1. At the end moment, I realised that the source value has been eradicated from the dataset, so, will need to work on that again.
2. Do better clustering and collaborative filtering

3. SQL query can not be run on my laptop or any free online ide due to the large size of activity_data. Will need to do that again tomorrow and showcase the results.