RELAX TAKE HOME CHALLENGE

1. Problem Statement
    - Identify the Adopted User it has 2 conditions: First is have log in 3 separate days, and Second is have to log in 7 day period.
    - Which feature are lead to adopted user
2. I have two data tables
    - First is user table have information of the users and how they are join
    - Second is useable summary table is tell the day of each user is login
3. Predefined statement and processing data
    - I know that the second table has information for me to identify the Adopted used, because it shows what day the user is login. First I just extract the Year, Month and day, and Week number of each row of date time. Then I just group by Year,Month, Day and then count the total visit for each day. Then I used describe() method to see the statistics of our data, and it showed each user only log in 1 per a day.

	user_id	Year	Month	Day	total_visit
count	207917.000000	207917.000000	207917.000000	207917.000000	207917.0
mean	5913.314197	2013.377468	5.930112	15.799117	1.0
std	3394.941674	0.590782	3.529426	8.807596	0.0
min	1.000000	2012.000000	1.000000	1.000000	1.0
25%	3087.000000	2013.000000	3.000000	8.000000	1.0
50%	5682.000000	2013.000000	5.000000	16.000000	1.0
75%	8944.000000	2014.000000	9.000000	23.000000	1.0
max	12000.000000	2014.000000	12.000000	31.000000	1.0

    - Now I know each user just login one per day, so now I will just count how many times they are login in a week(or 7-day period). But we know the are period of 2 year, and 52 week for each year, so table will have a lot duplicate for user, but our goal is just find the user have login at least 3 day per week at least one time, so I use groupby the user again and use max() method to get the maximum the total visit of each user. Then, I will assign 1 to the Adopted user if the user has a total login in each week greater than or equal 3 and 0 if not . Table below, row total user visit, column is number time of each week. You can say that we have 6821 users who only visit 1 time per week.

	1	2	7	3	5	6	4	8
total_visit	6821	558	480	282	235	223	219	5

    -
    - Now I have identified the Adopted user, so we need to see which features predict the Adopted user. I know a useful feature in the user table column are: user_id,

creation_scoure, opt in to mail list, enable for marketing drip, origin ID. I don't need name, email and creation time, last session log in, and invited user Id.

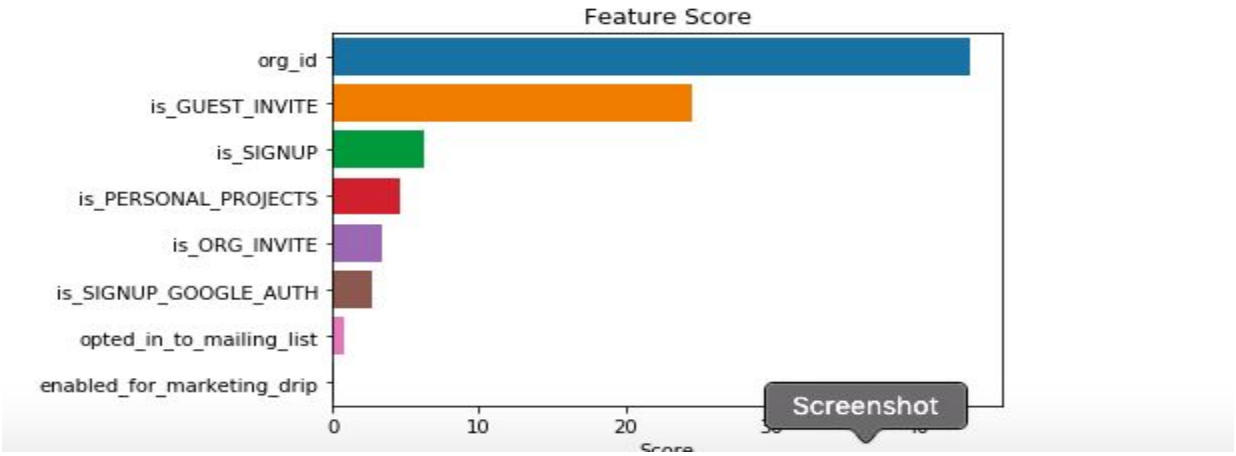
- But creation_scoure is category type, so I need to convert it to a dummy numerical variable, create a new column for each category in the column, then assign 1 to column if category matches with column name.
- After I convert to numerical data type, then I merge the data frame I did above for Adopted user with this user information to created the new data with all feature I have, then use SelectKBest model in the scikit learn to see which feature have a best score to predict the Adopted User.

```
skb = SelectKBest(f_classif, k='all')
skb.fit(X,y)
```

```
SelectKBest(k='all', score_func=<function f_classif at 0x1208f29d8>)
```

```
New_data = skb.transform(X)
```

```
feature_df = pd.DataFrame(skb.scores_, index = X.columns)
feature_df.columns = ['Score']
feature_df.sort_values(by= 'Score', ascending = False, inplace = True)
_ = sns.barplot(x= 'Score', y= feature_df.index, data = feature_df)
_ = plt.title('Feature Score')
```



4. Conclusion
    a. The table users have 12000 users sign up, but only 8823 users are actual log in to the product.
    b. In 8823 users are login, but 6821 users only login 1 time per week.
    c. After identifying the Adopted User, it shows only 1444 users meet the condition of the Adoption user.
    d. Use the SelectKBest method and use parameter f_classif, it returns the best 2 features is org_id, and Guest_invite.
    e. You can use processing and coding in this Jupyter Notebook.