In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
df = pd.read_csv('Adops & Data Scientist Sample Data - Q1 Analytics.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	ts	user_id	country_id	site_id
0	2019-02-01 00:01:24	LC36FC	TL6	N0OTG
1	2019-02-01 00:10:19	LC39B6	TL6	N0OTG
2	2019-02-01 00:21:50	LC3500	TL6	N0OTG
3	2019-02-01 00:22:50	LC374F	TL6	N0OTG
4	2019-02-01 00:23:44	LCC1C3	TL6	QGO3G

In [4]:

```
df.shape
```

Out[4]:

```
(3553, 4)
```

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3553 entries, 0 to 3552
Data columns (total 4 columns):
ts            3553 non-null object
user_id       3553 non-null object
country_id    3553 non-null object
site_id       3553 non-null object
dtypes: object(4)
memory usage: 111.2+ KB
```

Q1: Consider only the rows with country_id = "BDV" (there are 844 such rows). For each site_id, we can compute the number of unique user_id's found in these 844 rows. Which site_id has the largest number of unique users? And what's the number?

In [6]:

```
df_BDV = df[df['country_id']=='BDV']
```

In [7]:

```
df_BDV['site_id'].unique()
```

Out[7]:

```
array(['N0OTG', '5NPAU', '3POLC'], dtype=object)
```

In [8]:

```
df_BDV = df_BDV[['user_id','site_id']]
```

In [10]:

```
df_BDV.groupby('site_id')['user_id'].nunique().sort_values(ascending = False)
```

Out[10]:

```
site_id
5NPAU    544
N0OTG     90
3POLC      2
Name: user_id, dtype: int64
```

Site '5NPAU' has maximum unique users, and number being 544

Q2 : Between 2019-02-03 00:00:00 and 2019-02-04 23:59:59, there are four users who visited a certain site more than 10 times. Find these four users & which sites they (each) visited more than 10 times. (Simply provides four triples in the form (user_id, site_id, number of visits) in the box below.)

In [11]:

```
df_ts = df[df['ts'] > '2019-02-03 00:00:00']
```

In [12]:

```
df_ts = df_ts[df_ts['ts'] < '2019-02-04 23:59:59']
```

In [14]:

```
df_ts.head()
```

Out[14]:

	ts	user_id	country_id	site_id
1049	2019-02-03 00:02:31	LC3C7E	TL6	3POLC
1050	2019-02-03 00:03:09	LC3C7E	TL6	3POLC
1051	2019-02-03 00:03:46	LC3C7E	TL6	3POLC
1052	2019-02-03 00:04:12	LC3C7E	TL6	3POLC
1053	2019-02-03 00:04:25	LC3C7E	TL6	3POLC

In [15]:

```
df_ts_top4Visits = df_ts.groupby(['user_id','site_id']).ts.count().sort_values(a
scending = False)[0:4]
```

In [16]:

```
df_ts_top4Visits
df_ts_top4Visits = pd.DataFrame(df_ts_top4Visits)
df_ts_top4Visits.rename(columns={'ts':'number_of_visits'})
```

Out[16]:

		number_of_visits
user_id	site_id	
LC3A59	N0OTG	26
LC06C3	N0OTG	25
LC3C9D	N0OTG	17
LC3C7E	3POLC	15

Q3 : For each site, compute the unique number of users whose last visit (found in the original data set) was to that site. For instance, user "LC3561"'s last visit is to "N0OTG" based on timestamp data. Based on this measure, what are top three sites? (hint: site "3POLC" is ranked at 5th with 28 users whose last visit in the data set was to 3POLC; simply provide three pairs in the form (site_id, number of users).)

In [17]:

```
df_visits = pd.read_csv('Adops & Data Scientist Sample Data - Q1 Analytics.csv')
```

In [18]:

```
df_visits['site_id'].nunique()  ## number of unique sites = 8
```

Out[18]:

```
8
```

In [19]:

```
df_visits['site_id'].unique()
```

Out[19]:

```
array(['N0OTG', 'QGO3G', 'GVOFK', '3POLC', '5NPAU', 'RT9Z6', 'JSUUP'
,
       'EUZ/Q'], dtype=object)
```

In [21]:

```
df_site_last = df_visits.groupby(['user_id']).agg({'ts':'max','site_id':'last'})
```

In [22]:

```
df_site_last.head(10)
```

Out[22]:

	ts	**site_id**
user_id		
LC00C3	2019-02-03 18:52:50	5NPAU
LC01C3	2019-02-04 11:35:10	5NPAU
LC05C3	2019-02-02 14:14:44	5NPAU
LC06C3	2019-02-07 01:16:12	N0OTG
LC07C3	2019-02-05 19:06:42	5NPAU
LC08C3	2019-02-05 16:11:30	5NPAU
LC0C32	2019-02-07 01:18:03	N0OTG
LC0C34	2019-02-06 21:01:55	5NPAU
LC0C35	2019-02-01 17:44:39	5NPAU
LC0C3B	2019-02-01 22:02:40	QGO3G

In [23]:

```python
df_visits[df_visits['user_id']=='LC0C3B']     #verify with some users that the ts and site aggregations computed above is correct
```

Out[23]:

	ts	user_id	country_id	site_id
505	2019-02-01 22:02:40	LC0C3B	TL6	QGO3G

In [24]:

```python
df_site_last.reset_index(inplace=True)
df_site_last
```

Out[24]:

	user_id	ts	site_id
0	LC00C3	2019-02-03 18:52:50	5NPAU
1	LC01C3	2019-02-04 11:35:10	5NPAU
2	LC05C3	2019-02-02 14:14:44	5NPAU
3	LC06C3	2019-02-07 01:16:12	N0OTG
4	LC07C3	2019-02-05 19:06:42	5NPAU
...	...	...	...
1911	LCFC3B	2019-02-05 04:53:03	N0OTG
1912	LCFC3D	2019-02-01 18:59:50	N0OTG
1913	LCFC3E	2019-02-01 20:49:13	5NPAU
1914	LCFEC3	2019-02-07 06:23:59	3POLC
1915	LCFFC3	2019-02-05 03:31:17	N0OTG

1916 rows × 3 columns

In [25]:

```
df_site_last.groupby('site_id').user_id.count().sort_values(ascending = False)
```

Out[25]:

```
site_id
5NPAU    992
N0OTG    561
QGO3G    289
GVOFK     42
3POLC     28
RT9Z6      2
JSUUP      1
EUZ/Q      1
Name: user_id, dtype: int64
```

In [26]:

```
df_site_last = df_site_last.groupby('site_id').user_id.count().sort_values(ascending = False)
df_top3 = pd.DataFrame(df_site_last[0:3])
df_top3.rename(columns={'user_id':'number_of_users'})
```

Out[26]:

	number_of_users
site_id	
5NPAU	992
N0OTG	561
QGO3G	289

Q4 : For each user, determine the first site he/she visited and the last site he/she visited based on the timestamp data. Compute the number of users whose first/last visits are to the same website. What is the number?

In [28]:

```
df = pd.read_csv('Adops & Data Scientist Sample Data - Q1 Analytics.csv')
```

In [29]:

```
df['user_id'].nunique()
```

Out[29]:

```
1916
```

Total Unique Users = 1916

In [30]:

```
df_user_visits = df[['ts','user_id','site_id']]
```

In [31]:

```
df_user_visits.head()
```

Out[31]:

	ts	user_id	site_id
0	2019-02-01 00:01:24	LC36FC	N0OTG
1	2019-02-01 00:10:19	LC39B6	N0OTG
2	2019-02-01 00:21:50	LC3500	N0OTG
3	2019-02-01 00:22:50	LC374F	N0OTG
4	2019-02-01 00:23:44	LCC1C3	QGO3G

In [32]:

```
#compute aggregations for the min/max timestamp (first and last visit for each user), and corresponding site names
df_site_first = df_user_visits.groupby(['user_id']).agg({'ts':['min','max'],'site_id':['first','last'] })
```

In [33]:

```
df_site_first.head(10)
```

Out[33]:

	ts		site_id	
	min	max	first	last
user_id				
LC00C3	2019-02-03 18:52:50	2019-02-03 18:52:50	5NPAU	5NPAU
LC01C3	2019-02-04 11:35:10	2019-02-04 11:35:10	5NPAU	5NPAU
LC05C3	2019-02-02 14:14:44	2019-02-02 14:14:44	5NPAU	5NPAU
LC06C3	2019-02-01 22:49:39	2019-02-07 01:16:12	N0OTG	N0OTG
LC07C3	2019-02-05 19:06:42	2019-02-05 19:06:42	5NPAU	5NPAU
LC08C3	2019-02-05 16:11:30	2019-02-05 16:11:30	5NPAU	5NPAU
LC0C32	2019-02-05 22:33:51	2019-02-07 01:18:03	QGO3G	N0OTG
LC0C34	2019-02-06 21:01:55	2019-02-06 21:01:55	5NPAU	5NPAU
LC0C35	2019-02-01 17:44:39	2019-02-01 17:44:39	5NPAU	5NPAU
LC0C3B	2019-02-01 22:02:40	2019-02-01 22:02:40	QGO3G	QGO3G

In [34]:

```
df[df['user_id']=='LC0C32'] #verify with some users that the ts and site aggregations computed above is correct by eyeballing
```

Out[34]:

	ts	user_id	country_id	site_id
2526	2019-02-05 22:33:51	LC0C32	TL6	QGO3G
3081	2019-02-07 01:18:03	LC0C32	TL6	N0OTG

In [35]:

```
df_site_first[df_site_first['site_id']['first'] == df_site_first['site_id']['last']]
```

Out[35]:

	ts		site_id	
	min	**max**	**first**	**last**
user_id				
LC00C3	2019-02-03 18:52:50	2019-02-03 18:52:50	5NPAU	5NPAU
LC01C3	2019-02-04 11:35:10	2019-02-04 11:35:10	5NPAU	5NPAU
LC05C3	2019-02-02 14:14:44	2019-02-02 14:14:44	5NPAU	5NPAU
LC06C3	2019-02-01 22:49:39	2019-02-07 01:16:12	N0OTG	N0OTG
LC07C3	2019-02-05 19:06:42	2019-02-05 19:06:42	5NPAU	5NPAU
...	...	...	...	...
LCFC38	2019-02-02 13:58:18	2019-02-02 13:58:18	5NPAU	5NPAU
LCFC3B	2019-02-05 04:53:03	2019-02-05 04:53:03	N0OTG	N0OTG
LCFC3D	2019-02-01 18:59:50	2019-02-01 18:59:50	N0OTG	N0OTG
LCFC3E	2019-02-01 20:49:08	2019-02-01 20:49:13	5NPAU	5NPAU
LCFFC3	2019-02-02 22:36:23	2019-02-05 03:31:17	N0OTG	N0OTG

1670 rows × 4 columns

Number of Users where first and last site is same : 1670

In [36]:

```
df_site_first[df_site_first['site_id']['first'] != df_site_first['site_id']['last']]
```

Out[36]:

	ts		site_id	
	min	max	first	last
user_id				
LC0C32	2019-02-05 22:33:51	2019-02-07 01:18:03	QGO3G	N0OTG
LC1C32	2019-02-04 13:24:34	2019-02-05 01:59:42	5NPAU	QGO3G
LC1C3C	2019-02-01 13:28:31	2019-02-04 18:49:03	N0OTG	5NPAU
LC1EC3	2019-02-04 19:14:01	2019-02-07 20:38:10	5NPAU	N0OTG
LC2C36	2019-02-05 14:54:53	2019-02-06 22:28:21	N0OTG	5NPAU
...	...	...	...	...
LCDC36	2019-02-01 19:48:37	2019-02-04 18:50:23	5NPAU	N0OTG
LCEDC3	2019-02-01 12:40:17	2019-02-06 00:42:54	GVOFK	5NPAU
LCF8C3	2019-02-02 20:11:26	2019-02-07 20:34:58	5NPAU	N0OTG
LCFC32	2019-02-04 21:02:40	2019-02-06 02:45:48	5NPAU	N0OTG
LCFEC3	2019-02-02 01:19:49	2019-02-07 06:23:59	N0OTG	3POLC

246 rows × 4 columns

Number of Users where first and last site is NOT same : 246

1670+246 = 1916 ##Numbers add up to unique users

In [37]:

```
df_site_first[df_site_first['site_id']['first'] == df_site_first['site_id']['last']].count()[0]
```

Out[37]:

```
1670
```

Ans for Q4 : 1670

In [ ]: