In [1]:
```
import pandas as pd

df = pd.read_csv('homepage_actions.csv')
df.head()
```

Out[1]:

	timestamp	id	group	action
0	2016-09-24 17:42:27.839496	804196	experiment	view
1	2016-09-24 19:19:03.542569	434745	experiment	view
2	2016-09-24 19:36:00.944135	507599	experiment	view
3	2016-09-24 19:59:02.646620	671993	control	view
4	2016-09-24 20:26:14.466886	536734	experiment	view

1. Match the following characteristics of this dataset:

- total number of actions
- number of unique users
- sizes of the control and experiment groups (i.e., the number of unique users in each group)

In [2]:
```
df.shape
```

Out[2]:
```
(8188, 4)
```

In [3]:
```
# total number of action

df.nunique()# the output shows that there are 2 different types of actio
n
```

Out[3]:
```
timestamp    8188
id           6328
group           2
action          2
dtype: int64
```

Using the output from the 'df.shape' we can say that there are 8188 total number of actions which are categorically represented by 2 values.

In [4]:
```
# number of unique users
df.id.nunique()
```

Out[4]:
```
6328
```

In [5]:
```
# size of control group and experiment group
df.groupby('group').nunique()
```

Out[5]:

group	timestamp	id	group	action
control	4264	3332	1	2
experiment	3924	2996	1	2

2. How long was the experiment run for?

Hint: the records in this dataset are ordered by timestamp in increasing order

In [8]:
```
# duration of this experiment

df.timestamp.max(), df.timestamp.min()
```

Out[8]:
```
('2017-01-18 10:24:08.629327', '2016-09-24 17:42:27.839496')
```

We can observe the latest and oldest value of timestamp, we can calculate that this period would be somewhere around 4 months.

3. What action types are recorded in this dataset?

(i.e., What are the unique values in the action column?)

In [11]:
```
# action types in this experiment
df.action.value_counts()
```

Out[11]:
```
view     6328
click    1860
Name: action, dtype: int64
```

4. Why would we use click through rate instead of number of clicks to compare the performances of control and experiment pages?

Answer: More clicks can occur in a particular version of the page even though when there is a greater percentage of clicks in other version. Hence to balance out such possibilities, we use the click through rate instead of number of clicks to compare the performances of control and experimental pages.

5. Define the click through rate (CTR) for this experiment.

```
Answer : For this experiment, the fraction of The number of clicks on the 'v
iew courses' button over the total views of the home-page button.
```

6. What are the null and alternative hypotheses?

Use CTR_{old} and CTR_{new} in your hypotheses.

$$H_0 : CTR_{new} =< CTR_{old}$$
$$H_1 : CTR_{new} > CTR_{old}$$

In [ ]: