Homepage Experiment Data

April 26, 2019

In [9]:
```
import pandas as pd

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

Out[9]:

	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

0.0.1 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 [10]:
```
# total number of actions
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8188 entries, 0 to 8187
Data columns (total 4 columns):
timestamp    8188 non-null object
id           8188 non-null int64
group        8188 non-null object
action       8188 non-null object
dtypes: int64(1), object(3)
memory usage: 256.0+ KB
```

In [13]:
```
# number of unique users
df['id'].nunique()
```

Out[13]: 6328

In [19]:
```
# size of control group and experiment group
df.query("group == 'control'").id.nunique(), df.query("group == 'experiment'").id.nuniq
```

Out[19]: (3332, 2996)

0.0.2 2. How long was the experiment run for?

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

```
In [20]: # duration of this experiment
         df.timestamp.min(), df.timestamp.max()

Out[20]: ('2016-09-24 17:42:27.839496', '2017-01-18 10:24:08.629327')
```

0.0.3 3. What action types are recorded in this dataset?

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

```
In [21]: # action types in this experiment
         df.action.unique()

Out[21]: array(['view', 'click'], dtype=object)
```

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

Answer Getting the proportion of the users who click is more effective than getting the number of users who click when comparing groups of different sizes.ie..more total clicks could occur in one version, even if there is a greater percentage of clicks in the other version (simpson's paradox)

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

Answer The No.of unique visitors who click at least once / The No.of unique visitors who view the page

0.0.6 6. What are the null and alternative hypotheses?

Use CTR_{old} and CTR_{new} in your hypotheses.
H0: CTRnew CTRold less than equal to 0
H1: CTRnew CTRold > 0

```
In [ ]:
```