New York Times Web Analytics

September 21, 2014

1 Introduction

We are presented with the web log of the **New York Times** Website containing information about the number of advertisements the users are exposed to (**Impressions**), and the number of advertisements those users click through to (**Clicks**).

The goal is to understand how different factors affect a user clicking through an advertisement. These factors include the number of **impressions**, **gender**, **age** and whether the user is **signed in**.

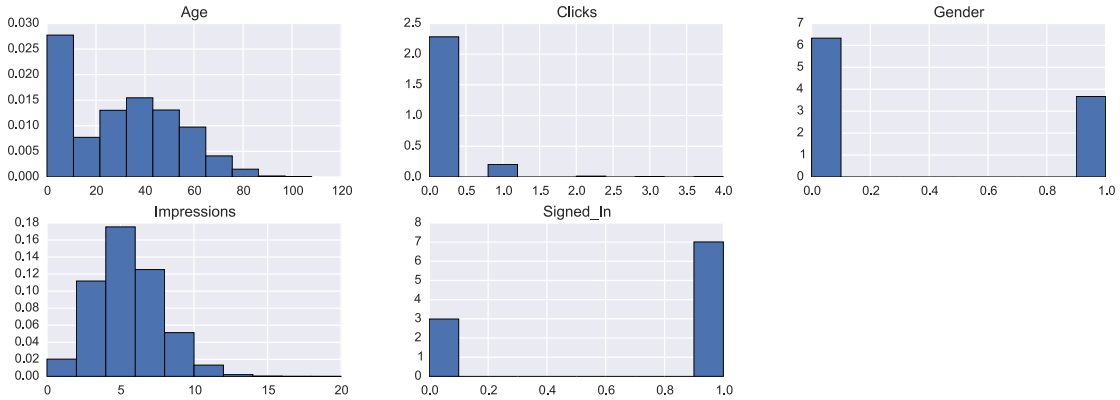
To start with, here is how the data looks like

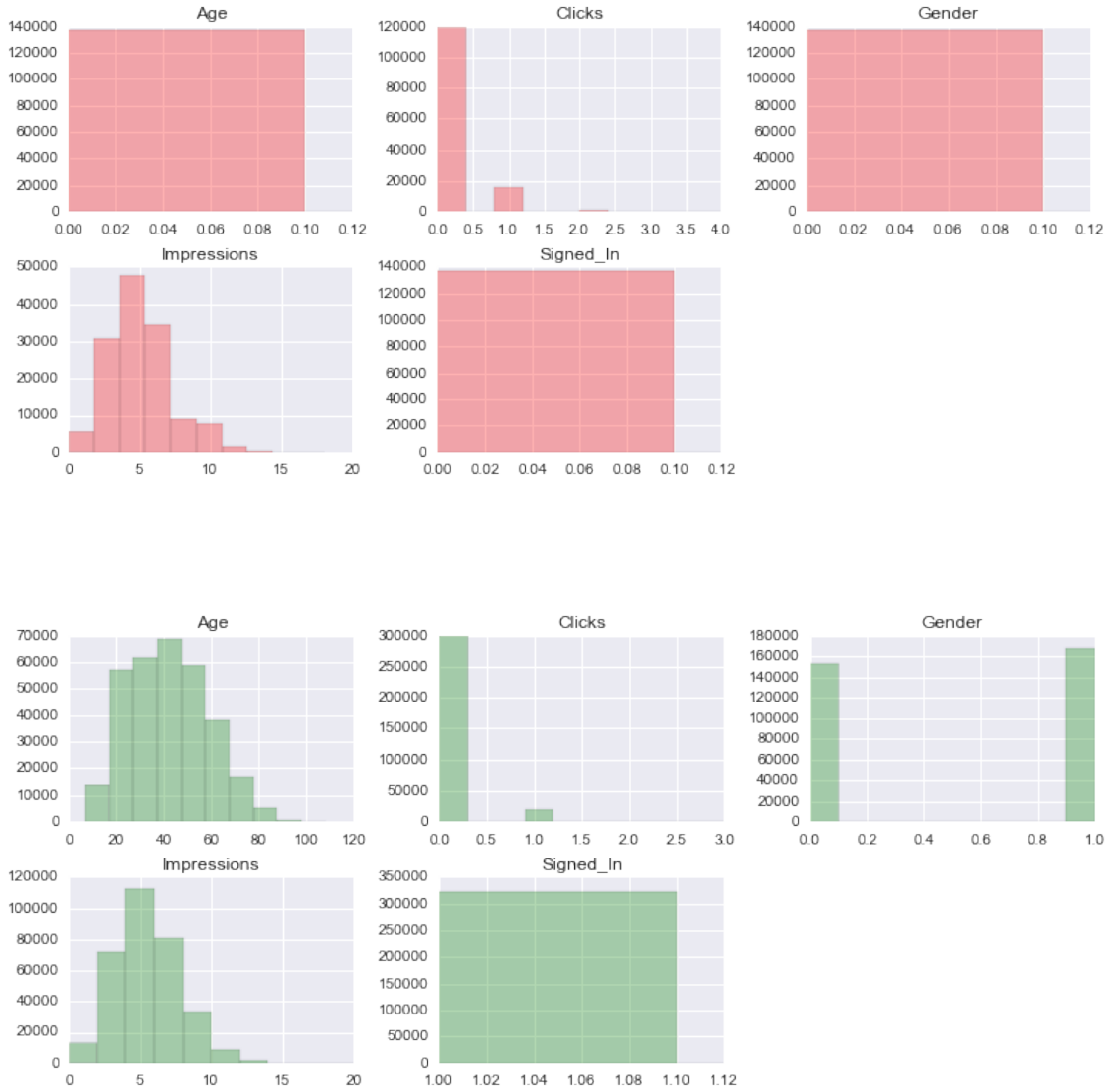
```
Out[80]:    Age  Gender  Impressions  Clicks  Signed_In
         0   36       0            3       0          1
         1   73       1            3       0          1
         2   30       0            3       0          1
         3   49       1            3       0          1
         4   47       1           11       0          1

         [5 rows x 5 columns]
```

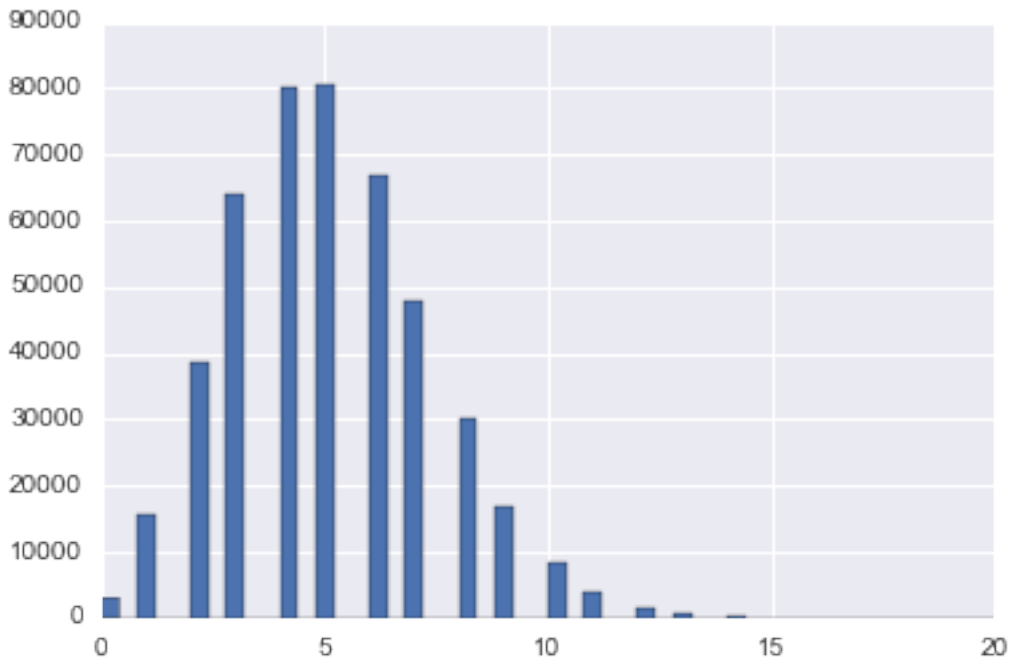
2 Analysis

Let's get an idea how the distribution of the data looks like ...



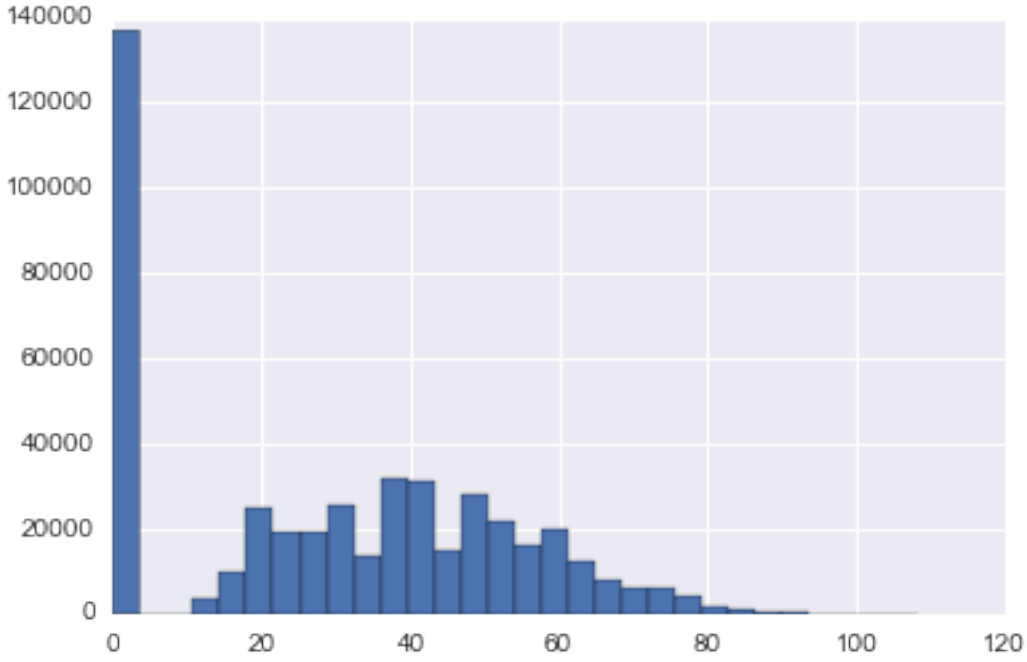


Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x10c1426d0>



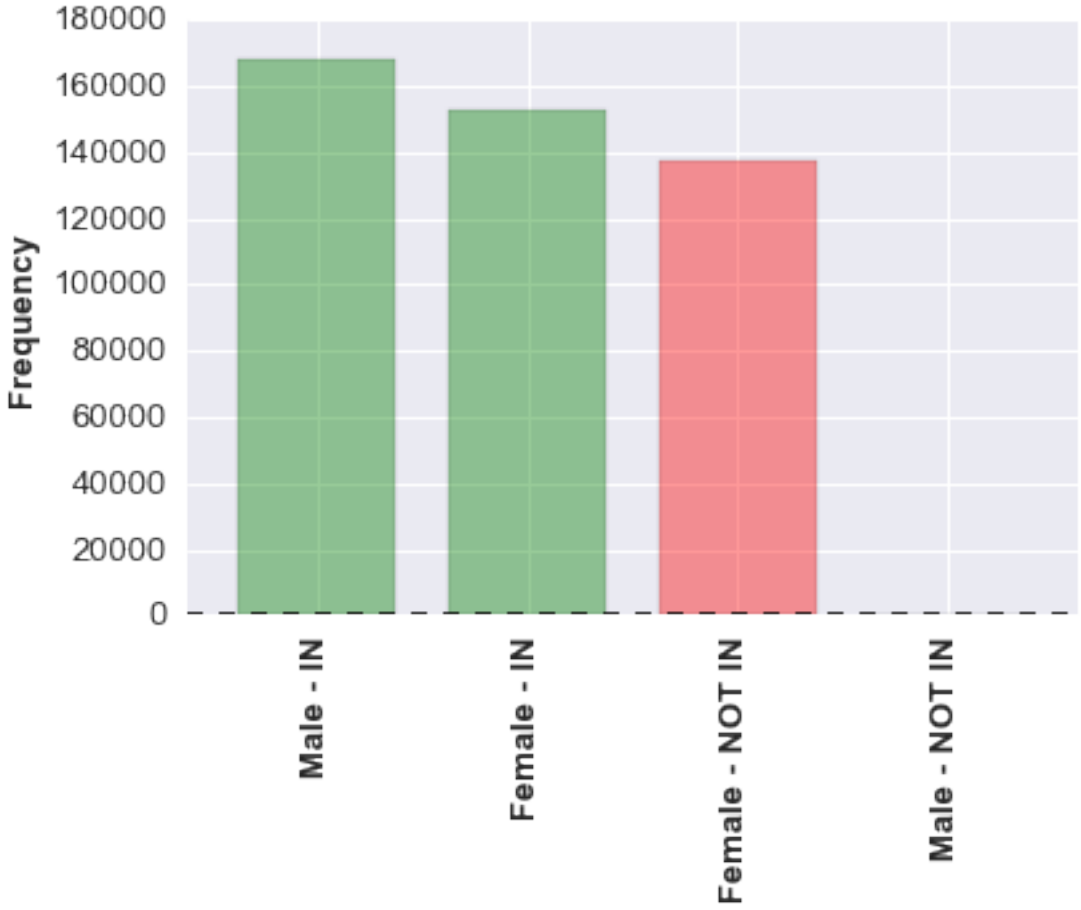
Impressions looks like a Gamma distribution

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x112580950>
```



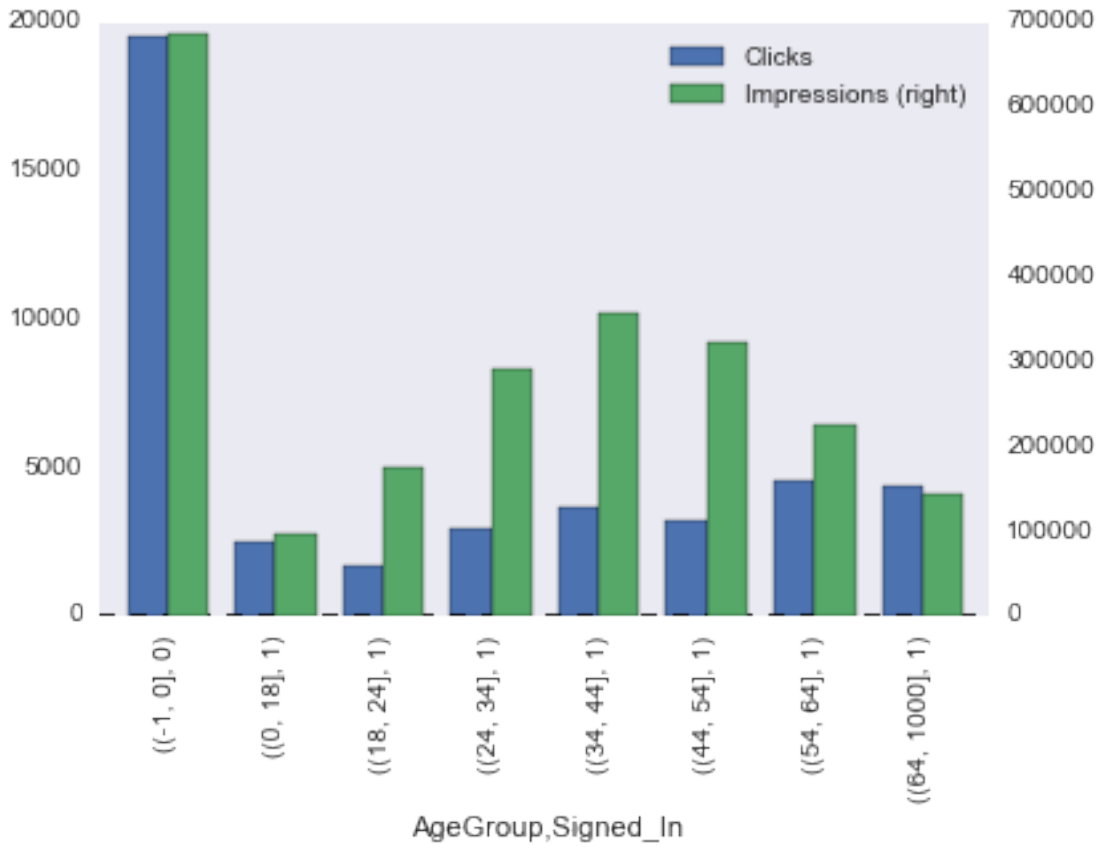
Obviously we have a bunch of people with **age 0**

Out[7]: <matplotlib.text.Text at 0x11261d0d0>



All males signed ?? Females on the other hand doesn't always sign in, but there are more users who are female. Are the female bots???

Out[11]: array(['(34, 44]', '(64, 1000]', '(24, 34]', '(44, 54]', '(-1, 0]',
'(0, 18]', '(18, 24]', '(54, 64]'], dtype=object)



Out[16]:

	GP1	GP1_mean	GP2	GP2_mean	GPonetwo_diff	p_val
7	(64, 1000]	0.029803	(44, 54]	0.009958	0.019845	1.430923e-295
0	(34, 44]	0.010286	(64, 1000]	0.029803	-0.019516	5.245541e-288
6	(64, 1000]	0.029803	(24, 34]	0.010146	0.019656	7.860398e-285
9	(64, 1000]	0.029803	(18, 24]	0.009720	0.020082	2.458627e-272
15	(44, 54]	0.009958	(0, 18]	0.026621	-0.016663	1.300520e-151
17	(44, 54]	0.009958	(54, 64]	0.020307	-0.010349	2.525271e-151
3	(34, 44]	0.010286	(0, 18]	0.026621	-0.016334	1.498876e-146
12	(24, 34]	0.010146	(0, 18]	0.026621	-0.016474	2.458928e-146
18	(0, 18]	0.026621	(18, 24]	0.009720	0.016900	2.346743e-144
5	(34, 44]	0.010286	(54, 64]	0.020307	-0.010020	7.523228e-144
14	(24, 34]	0.010146	(54, 64]	0.020307	-0.010160	5.668132e-141
20	(18, 24]	0.009720	(54, 64]	0.020307	-0.010586	1.007813e-130
10	(64, 1000]	0.029803	(54, 64]	0.020307	0.009496	9.214903e-56
19	(0, 18]	0.026621	(54, 64]	0.020307	0.006314	5.323948e-20
8	(64, 1000]	0.029803	(0, 18]	0.026621	0.003182	4.361969e-05

[15 rows x 6 columns]

Out[20]:

	GP1	GP1_mean	GP2	GP2_mean	GPonetwo_diff	p_val
1	(34, 44]	0.010286	(24, 34]	0.010146	0.000140	6.246618e-01
11	(24, 34]	0.010146	(44, 54]	0.009958	0.000189	5.146885e-01
16	(44, 54]	0.009958	(18, 24]	0.009720	0.000237	4.779019e-01
2	(34, 44]	0.010286	(44, 54]	0.009958	0.000329	2.339282e-01

```
13    (24, 34]  0.010146    (18, 24]  0.009720       0.000426   2.136576e-01
4     (34, 44]  0.010286    (18, 24]  0.009720       0.000566   8.747009e-02
8   (64, 1000]  0.029803     (0, 18]  0.026621       0.003182   4.361969e-05
19     (0, 18]  0.026621    (54, 64]  0.020307       0.006314   5.323948e-20
10  (64, 1000]  0.029803    (54, 64]  0.020307       0.009496   9.214903e-56
20    (18, 24]  0.009720    (54, 64]  0.020307      -0.010586  1.007813e-130
14    (24, 34]  0.010146    (54, 64]  0.020307      -0.010160  5.668132e-141
5     (34, 44]  0.010286    (54, 64]  0.020307      -0.010020  7.523228e-144
18     (0, 18]  0.026621    (18, 24]  0.009720       0.016900  2.346743e-144
12    (24, 34]  0.010146     (0, 18]  0.026621      -0.016474  2.458928e-146
3     (34, 44]  0.010286     (0, 18]  0.026621      -0.016334  1.498876e-146
17    (44, 54]  0.009958    (54, 64]  0.020307      -0.010349  2.525271e-151
15    (44, 54]  0.009958     (0, 18]  0.026621      -0.016663  1.300520e-151
9   (64, 1000]  0.029803    (18, 24]  0.009720       0.020082  2.458627e-272
6   (64, 1000]  0.029803    (24, 34]  0.010146       0.019656  7.860398e-285
0     (34, 44]  0.010286  (64, 1000]  0.029803      -0.019516  5.245541e-288
7   (64, 1000]  0.029803    (44, 54]  0.009958       0.019845  1.430923e-295

[21 rows x 6 columns]
```

Out[79]: u'/Users/JeffreyTang/Desktop/Zipfian/ab-testing'