

In this project, we examine A/B testing data for ShoeFly.com to determine which version of an ad resulted in more clicks across email, Google, Facebook, and Twitter. The csv file of all the data as well as the full Excel spreadsheet with accompanying pivot tables are included in the repository. In this document, I walk through the steps of the project with screenshots of each pivot table created for the different parts of the analysis.

1. After importing the data and cleaning up the column names and data (removing separators, etc.), we see that the data has the following columns:
 1. **user_id** (an ID number identifying a unique user) COLUMN A
 2. **utm_source** (where the user saw the ad: Google, Facebook, Twitter, or email) COLUMN B
 3. **day** (which day of the week the user saw the ad) COLUMN C
 4. **ad_click_timestamp** (the time at which the user clicked on the ad; this column is empty if a user didn't click on the ad) COLUMN D
 5. **experimental_group** (did the user see the A or B version of the ad) COLUMN E

1	user_id	utm_source	day	ad_click_timestamp	experimental_group
2	008b7c6c-7272-471e-b90e-930d548bd8d7	google	6 - Saturday	7:18	A
3	009abb94-5e14-4b6c-bb1c-4f4df7aa7557	facebook	7 - Sunday		B
4	00f5d532-ed58-4570-b6d2-768df5f41aed	twitter	2 - Tuesday		A
5	011adc64-0f44-4fd9-a0bb-f1506d2ad439	google	2 - Tuesday		B
6	012137e6-7ae7-4649-af68-205b4702169c	facebook	7 - Sunday		B
7	013b0072-7b72-40e7-b698-98b4d0c9967f	facebook	1 - Monday		A

2. To aid in the analysis, we created a new column “**is_click**” (COLUMN F) that takes on binary values (TRUE or FALSE). If the **ad_click_timestamp** column is empty, the value of this column is FALSE. Otherwise, it is TRUE. The formula for this column is:

✓ *f_x* =IF(ISBLANK(D2), FALSE, TRUE)

3. Now we can begin our analysis. First, we want to know how many people clicked on the ad (regardless of version) for each source (Google, Facebook, Twitter, email). To do this,

we create a pivot table where the **rows** are the sources, and the **values** field is a count of rows with a nonnull entry in the **ad_click_timestamp** column:

3	Row Labels	Count of ad_click_timestamp
4	email	80
5	facebook	180
6	google	239
7	twitter	66
8	Grand Total	565

4. We can create a more detailed pivot table where the **rows** are the sources, the **columns** are the possible values from is_click (TRUE or FALSE) and **values** are a count of user_id. We'll also create a **Percent Clicked** column for each source, whose value equals $\text{TRUE}/(\text{TRUE} + \text{FALSE})$ or the number of people who clicked on the ad from that source divided by the total number of people who saw the ad on that source, whether they clicked it or not. From here, we can also see that of the 1654 users who saw an ad, about 34% of these people clicked on the ad:

2	Count of user_id	Column Labels			
3	Row Labels	<input type="checkbox"/> FALSE	<input type="checkbox"/> TRUE	Grand Total	Percent Clicked
4	email	175	80	255	0.31372549
5	facebook	324	180	504	0.357142857
6	google	441	239	680	0.351470588
7	twitter	149	66	215	0.306976744
8	Grand Total	1089	565	1654	0.341596131

We see that Facebook and Google generated the highest percentage of clicks (~35%), while email and Twitter generated only ~31%.

- Now let's get more information on the results for each version of the ad. First, we'll verify that the same number of people saw each ad:

22	Row Labels	Count of user_id
23	A	827
24	B	827
25	Grand Total	1654

- Now let's create a pivot table that totals the aggregate number of clicks generated by the A and B versions of the ad across all platforms. The **rows** will be A and B, the **columns** TRUE and FALSE and the **values** a count of user_id. We'll also create a Percent Clicked column to calculate the percentage of people who clicked on each ad:

30	Count of user_id	Column Labels			
31	Row Labels	FALSE	TRUE	Grand Total	Percent Clicked
32	A	517	310	827	0.374848851
33	B	572	255	827	0.30834341
34	Grand Total	1089	565	1654	0.341596131

From this alone, we see that across all platforms, ad A resulted in a 37% click rate while ad B only generated a 31% click rate.

- Next, let's evaluate the performance of each ad based on day of the week. To do this, we'll first apply the **Filter** function to our spreadsheet based on the "experimental_group" column to create separate spreadsheets for users who saw ad A and for those who saw ad B. Then for each group we'll create a separate pivot table where the **rows** are the days of the week, the **columns** are the "is_click" values, the **values** are the count of user_ids (for TRUE, FALSE as well as a sum of the total number of users who saw the ad on a particular day). Then, we'll create a Percent Clicked column:

37	Experimental Group A				
38	Count of user_id	Column Labels			
39	Row Labels	<input type="checkbox"/> FALSE	TRUE	Grand Total	Percent Clicked
40	1 - Monday	70	43	113	0.380530973
41	2 - Tuesday	76	43	119	0.361344538
42	3 - Wednesday	86	38	124	0.306451613
43	4 - Thursday	69	47	116	0.405172414
44	5 - Friday	77	51	128	0.3984375
45	6 - Saturday	73	45	118	0.381355932
46	7 - Sunday	66	43	109	0.394495413
47	Grand Total	517	310	827	0.374848851
48					
49					
50	Experimental Group B				
51	Count of user_id	Column Labels			
52	Row Labels	<input type="checkbox"/> FALSE	TRUE	Grand Total	Percent Clicked
53	1 - Monday	81	32	113	0.283185841
54	2 - Tuesday	74	45	119	0.378151261
55	3 - Wednesday	89	35	124	0.282258065
56	4 - Thursday	87	29	116	0.25
57	5 - Friday	90	38	128	0.296875
58	6 - Saturday	76	42	118	0.355932203
59	7 - Sunday	75	34	109	0.311926606
60	Grand Total	572	255	827	0.30834341

With the exception of Tuesday, ad A performed better on every day of the week. Tuesday also saw the smallest difference in performance between the two ads (a ~1% difference) while the other days saw a noticeably larger performance disparity (as much as a 15% difference on Thursday).

5	Day of the Week	Percent Clicked A	Percent Clicked B	Difference in Performance
6	1 - Monday	0.380530973	0.283185841	0.097345133
7	2 - Tuesday	0.361344538	0.378151261	0.016806723
8	3 - Wednesday	0.306451613	0.282258065	0.024193548
9	4 - Thursday	0.405172414	0.25	0.155172414
0	5 - Friday	0.3984375	0.296875	0.1015625
1	6 - Saturday	0.381355932	0.355932203	0.025423729
2	7 - Sunday	0.394495413	0.311926606	0.082568807
3	Average	0.37539834	0.308332711	0.071867551

Across all days, the average difference across each day for the two ads is 7%.