Spanish Translation A/B Test

Goal

A/B tests play a huge role in website optimization. Analyzing A/B tests data is a very important data scientist responsibility. Especially, data scientists have to make sure that results are reliable, trustworthy, and conclusions can be drawn.

Furthermore, companies often run tens, if not hundreds, of A/B tests at the same time. Manually analyzing all of them would require lot of time and people. Therefore, it is common practice to look at the typical A/B test analysis steps and try to automate as much as possible. This frees up time for the data scientists to work on more high level topics.

In this challenge, you will have to analyze results from an A/B test. Also, you will be asked to design an algorithm to automate some steps.

Challenge Description

Company XYZ is a worldwide e-commerce site with localized versions of the site.

A data scientist at XYZ noticed that Spain-based users have a much higher conversion rate than any other Spanish-speaking country. She therefore went and talked to the international team in charge of Spain And LatAm to see if they had any ideas about why that was happening.

Spain and LatAm country manager suggested that one reason could be translation. All Spanish-speaking countries had the same translation of the site which was written by a Spaniard. They agreed to try a test where each country would have its one translation written by a local. That is, Argentinian users would see a translation written by an Argentinian, Mexican users by a Mexican and so on. Obviously, nothing would change for users from Spain.

After they run the test however, they are really surprised cause the test is negative. I.e., it appears that the non-localized translation was doing better!

You are asked to:

- Confirm that the test is actually negative. That is, it appears that the old version of the site with just one translation across Spain and LatAm performs better

- Explain why that might be happening. Are the localized translations really worse?

[Click here](#) to check out our site if interested in a feedback on your solutions, 1:1 mentorship, and more!

- If you identified what was wrong, design an algorithm that would return FALSE if the same problem is happening in the future and TRUE if everything is good and the results can be trusted.

Data

We have 2 table downloadable by clicking **here**.

The 2 tables are:

```
"test_table" - general information about the test results
```

Columns:

- **user_id** : the id of the user. Unique by user. Can be joined to user id in the other table. For each user, we just check whether conversion happens the first time they land on the site since the test started.
- **date** : when they came to the site for the first time since the test started
- **source** : marketing channel: *Ads, SEO, Direct . Direct* means everything except for ads and SEO. Such as directly typing site URL on the browser, downloading the app w/o coming from SEO or Ads, referral friend, etc.
- **device** : device used by the user. It can be mobile or web
- **browser_language** : in browser or app settings, the language chosen by the user. It can be *EN, ES, Other* (Other means any language except for English and Spanish)
- **ads_channel** : if marketing channel is ads, this is the site where the ad was displayed. It can be: *Google, Facebook, Bing, Yahoo ,Other*. If the user didn't come via an ad, this field is NA
- **browser** : user browser. It can be: *IE, Chrome, Android_App, FireFox, Iphone_App, Safari, Opera*
- **conversion** : whether the user converted (1) or not (0). This is our label. A test is considered successful if it increases the proportion of users who convert.
- **test** : users are randomly split into test (1) and control (0). Test users see the new translation and control the old one. For Spain-based users, this is obviously always 0 since there is no change there.

```
"user_table" - some information about the user
```

Columns:

- **user_id** : the id of the user. It can be joined to user id in the other table
- **sex** : user sex: Male or Female
- **age** : user age (self-reported)
- **country** : user country based on ip address

Example

> Let's check one user who went through the test

head(test_table, 1)

Column Name	Value	Description
user_id	315281	this is id of the user
date	2015-12-03	he came to the site on Dec, 3 for the first time since the test started
source	Direct	his marketing channel was direct. No SEO or Ads.
device	Web	he visited the site using "web" (i.e. laptop/desktop, but not mobile)
browser_language	ES	his browser language settings are Spanish
ads_channel	NA	didn't come via an ad, so this has to be NA
browser	IE	he used Internet Explorer!
conversion	1	he converted
test	0	he was in control. That is, he saw the old translation written by a Spaniard

```
    Let's now check the characteristics of that user
```

subset(user_table,user_id == 315281)

Column Name	Value	Description
user_id	315281	same id as in the example above
sex	M	he is a Male
age	32	he is 32 y/o
country	Spain	he is based in Spain. So, in his case, he could have not been in the test no matter what.