# **Description**

To begin this exploratory analysis, first import libraries and define functions to plot data using matplotlib. Depending on the data, not all graphs will be made. In this analysis, we will focus only on the data set out below.

* Training Features: These are the input variables your model will use to predict the likelihood that people will receive H1N1 flu and seasonal influenza vaccines. There are a total of 35 feature columns, each of which is an answer to a survey question. These questions cover several different issues, such as whether people observe safe behavioral practices, their views on diseases and vaccines, and their demographic characteristics.
* Training Labels: These are labels corresponding to observations in educational characteristics. There are two target variables: h1n1\_vaccine and season\_vaccine. Both are binary variables; It indicates that 1 person received the relevant flu vaccine and 0 people did not receive the relevant flu vaccine. Note that this is known as a "multi-label" modeling task.
* Test Features: These are the observation features that you will use to generate delivery predictions after training a model. We don't give you labels for these samples - it's up to you to create them.

First, we need to remove null values ​​from the data set provided and also verify that no missing values ​​are left in the data. Now let's double check that the lines between attributes and tags match. We don't want to have the wrong tags. If the two arrays (row indexes of the two data frames) do not match, Numpy's assert\_array\_equal error will return an error. The claim works successfully, so there is no problem. If the two index sequences were not the same, there would be an error. Let's start by taking a look at the distribution of the two target variables.

Apparently, about half of the people have received the seasonal flu vaccine, but only about 20% of people have received the H1N1 flu vaccine. In terms of class balance, we say that the target for seasonal flu vaccine has balanced classes, but the H1N1 flu vaccine target has moderately unbalanced classes. The phi coefficient of these two variables is 0.37, indicating a moderately positive correlation. We can also see this in cross tabulation. Most people who got the H1N1 flu vaccine also got the seasonal flu vaccine. While a small portion of those who had seasonal vaccines had the H1N1 vaccine, a higher proportion of them had H1N1 vaccine than those who did not have the seasonal vaccine. Now we learn how properties relate to target variables. We'll start by trying to visualize whether there is a simple bivariate correlation. If a feature is associated with the target, we expect different vaccination patterns as you change the attribute's values.

It is difficult to jump right to the final visualization. Instead, we can choose a feature and a goal and try to reach more features and a prototype before implementing both goals. We'll use h1n1\_concern, the level of concern the person shows about the H1N1 flu, and h1n1\_vaccine as the target variable.

In the graph, each pair of blue (no vaccine) and orange (vaccinated) bars add up to 1.0. We can clearly see that even if most people do not have the H1N1 vaccine, they are more likely to have higher levels of anxiety. Graphs have a stronger correlation with the seasonal vaccine, but much less h1n1\_vaccine so far. We see a particularly strong correlation with age\_group with seasonal vaccine, but not with h1n1\_vaccine. It seems that in seasonal flu, people act appropriately by the fact that people are more affected and the risk of flu-related complications is higher with age.