In preparation for applying machine learning algorithms to the predictor variables and target variable, I apply another step for determining relevance in my feature variables- statistical analysis. Up until this point I've used number and count thresholds, percentage cutoffs, and visualization filters. To complete my ML feature selection I will be implementing a statistical test on the remaining feature variables.

In the case of classification problems where input variables are categorical, we can use statistical tests to determine whether the output variable is dependent or independent of the input variables. If independent, then the input variable is a candidate for a feature that may be irrelevant to the problem and removed from the dataset.

For my statistical analysis I use the Pearson's chi-squared statistical hypothesis test for quantifying the independence of pairs of categorical variables. The chi-squared test assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable.

In python, the chi-squared test does this for a contingency table. During the visualization phase of this analysis, I wrote an algorithm that constructed a contingency table. This same algorithm will be used to construct the *observed frequencies* contingency tables to be used for the chi-squared test. First calculating the expected frequencies for the groups, then determining whether the division of the groups- the observed frequencies- matches the expected frequencies. The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the null hypothesis that the observed and expected frequencies are the same. The test statistic is then compared to the critical value computed at 99% probability. If the test statistic is greater than the critical value at, then we reject the null hypothesis and claim dependence.

The two functions that are used for the chi-squared test and to facilitate an interpretation of the results are chi2 and chi2_contingency from the scipy.stats module. My job then becomes to construct an algorithm that chains these two functions together and adds supporting lines of code to succinctly interpret the results.

I implement my chi-squared algorithm and print the corresponding results of independence/dependence. Of the 12 statistically tested feature variables, only one failed to reject the chi-squared null hypothesis of independence- 'ALT'. As a result, it is dropped from the dataframe.

The results of the chi-squared statistical analysis determined a statistically very likely non-independence correlation between 11 feature variables. These variables will be preprocessed before applying ML algorithms to them.