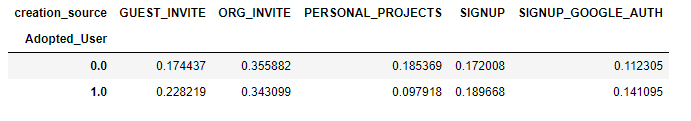
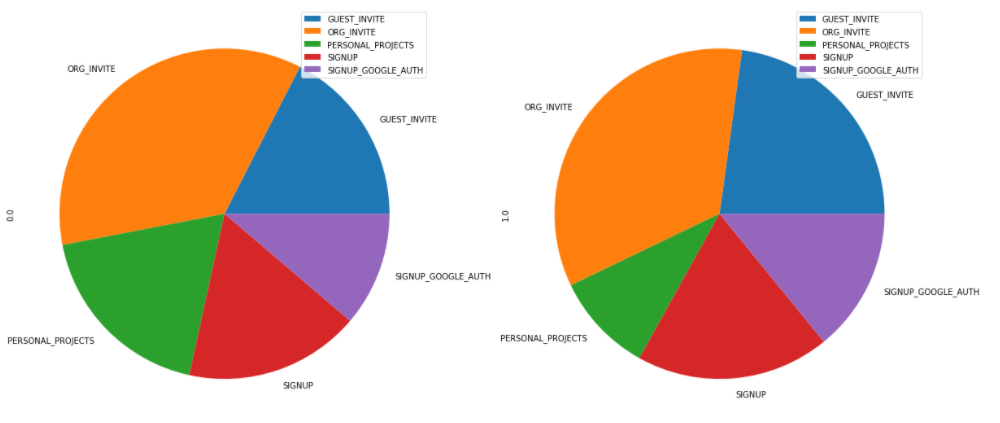
The following steps are documented in Relax\_Takehome.ipynb, in the same github repository.

Adopted users in the Relax user data are a minority class in an imbalanced dataset (1,297 Adopted out of 12,000 users). I self-joined the login dataset to find logins per user within 7 days of each other, then used pivot tabling to find unique users that had at least three logins within a single 7 day period.

I examined each of the features and decided to use chi square tests to establish if a given categorical variable was associated with Adopted Users, and then charted the frequencies for the Adopted v non-Adopted users. Looking at creation source, the chi square p value was sufficiently low (less than 0.05) to reject the null hypothesis that there was no difference between creation sources. Adopted Users were more likely from guest invites and less likely from personal projects. For the categories of marketing drip and opt in to mail list, neither yielded a low p value (0.82 and 0.31 respectively), and I did not reject the null hypothesis in either case. Below is a contingency table of creation source percentages with a pair of pie charts to highlight the differences. Guest invites had a higher percentage amongst Adopted Users, where Personal Projects had a lower percentage.





I ran a logistic regression, which was not successful due to the imbalanced nature of the dataset. While it achieved 89% accuracy overall, it achieved 0% precision in accuracy in predicting Adopted Users. I attempted to use the Imbalance Learn library but ran into a validation error. After looking up the error on Stack Overflow, I learned that I needed to update my version of SciKit Learn, and then found out it is not available in Anaconda. I would like to use the SMOTE methodology, randomly oversample the minority class and randomly under sample the majority class. This will have to be done after this library is available in Anaconda.