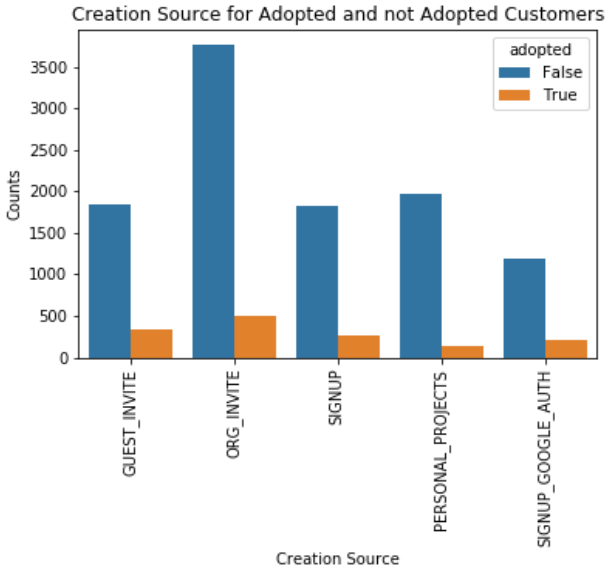
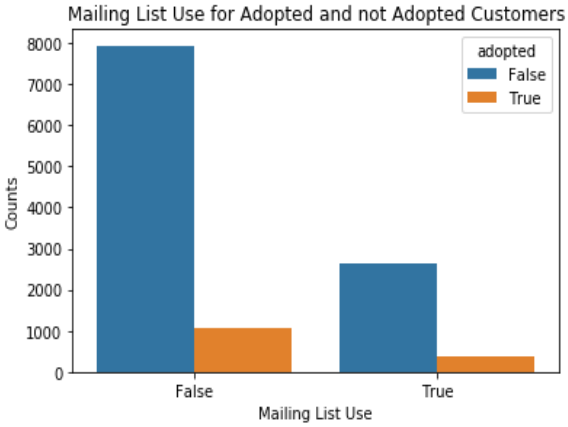
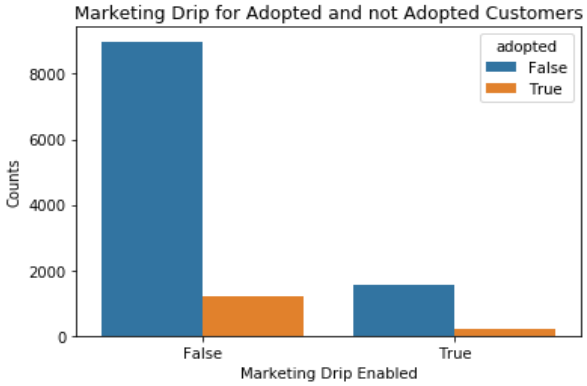
Report: Relax Data Science Challenge
Maia Paddock

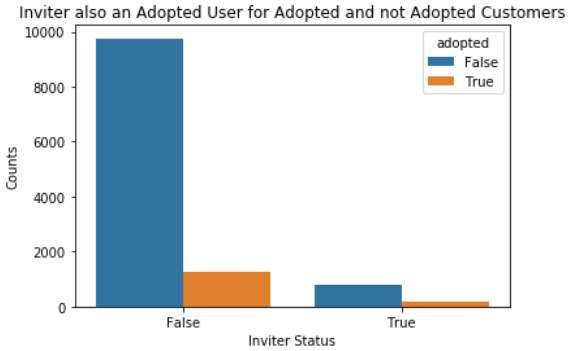
From my investigation of the data, after grouping the users and counting which ones had logged on at least three times in a seven day period at least one time, I found that 1439 users fell under this category. These I defined as adopted users. Approximately 12% of the users in the database were adopted users.

To look for the most important factors to predict future user adoption, I first plotted all of the categorical data into bar charts, with each category split between adopted users and non adopted users. These bar charts showed that users whose account was created through an organization invite were less likely than others to be adopted users. The most successful were guest invites, signups, and signups with google authentication. From the bar charts, both the mailing list and the marketing drip seemed to have very little correlation with whether or not a user was retained. Finally, I created a class to see if the inviting user being an adopted user or not would influence the change of the user being an adopted user. From the bar chart, there does not appear to be a correlation.









To examine the non-categorical data, I attempted to create scatter plots. However, the data was too dense and I was unable to find any quick correlations. After examining the bar charts, I quickly put together a random forest classifier to look at the feature importance in predicting an adopted user. The random forest classifier was not trained under optimal hyperparameters as the performance did not need to be optimized for this purpose. Plotting the feature importance found that the last email list and marketing drip were the least useful features, which is in line with what was seen on the bar charts. The time of the last session was by far the most important predictor, followed by the original creation time. The IDs of the organizations the users belong to was the next most important factor, showing that some organizations are more active than others. The creation source was not a very important factor despite the differences seen on the bar chart.