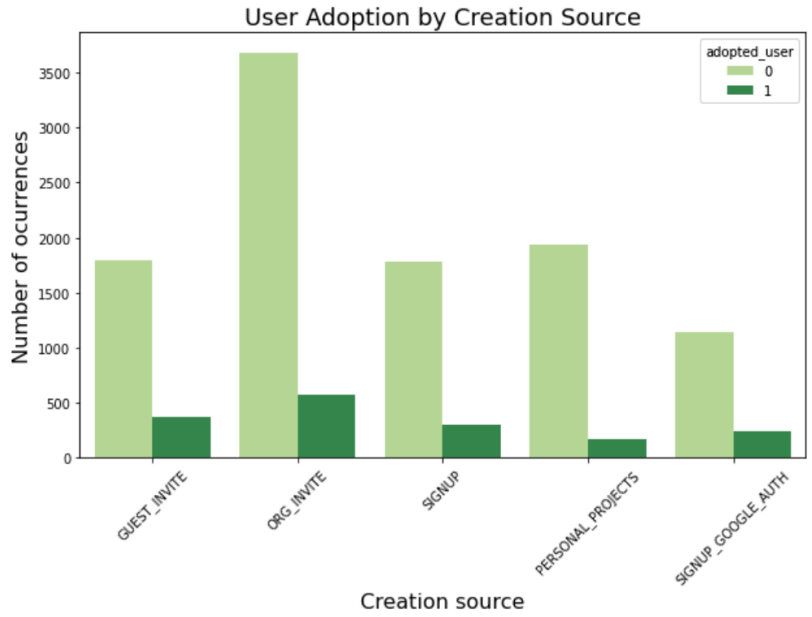
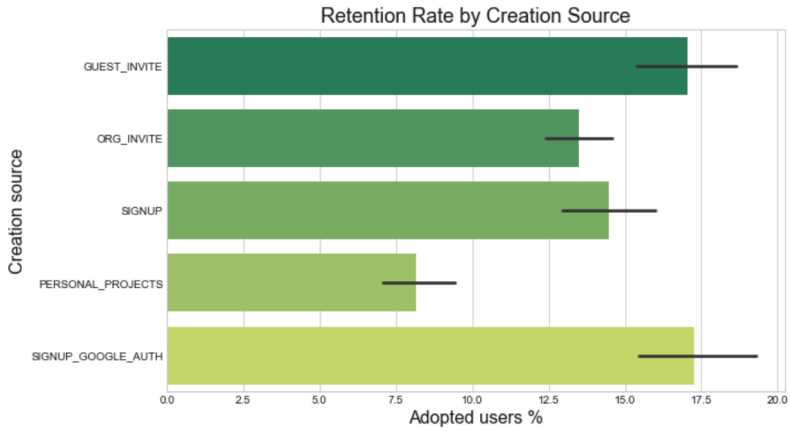
Overview

Exploratory data analysis: creation source seems to be a good predictor of retention rate. The features SIGNUP_GOOGLE_AUTH and GUEST_INVITE show a higher rate of adopted users compared to the other sources. The lowest retention rate is for users who sign up for personal projects. A chi-square confirms an association between creation_source and adopted_users with a p-value < 0.05.

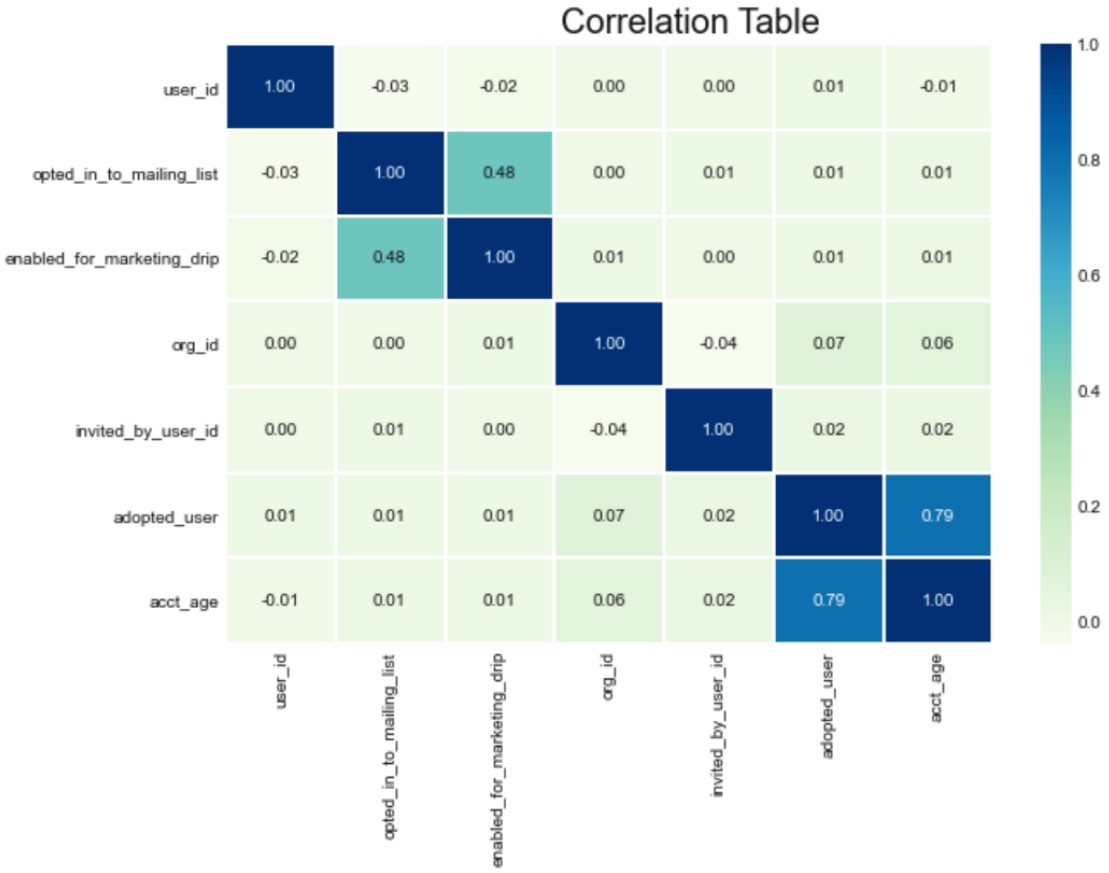




Predictive Modeling: Running a Random Forest Classifier with the original features returned a 0.72 test accuracy.

Feature Engineering: the first correlation heatmap showed a correlation between the target variable (adopted_user) and last_session_creation_time. Based on this

information, I created a feature called acct_age. This new feature has the age of each account and it turns out it has a high correlation 0.79 with adopted_user.



Predictive Modeling with new feature: This returned a test accuracy of 0.96 and the most important feature to predict user adoption turns out to be the engineered feature acct_age with the second one being org_id. Adding account age to the features significantly improved model performance.

```
Accuracy of test set was 0.9636666666666667
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      2586
           1       0.88      0.86      0.87       414

    accuracy                           0.96      3000
   macro avg       0.93      0.92      0.92      3000
weighted avg       0.96      0.96      0.96      3000
```

```
acct_age                      0.910104
org_id                        0.069466
PERSONAL_PROJECTS             0.008877
opted_in_to_mailing_list      0.002852
enabled_for_marketing_drip    0.002323
SIGNUP                        0.002305
ORG_INVITE                    0.002241
SIGNUP_GOOGLE_AUTH            0.001832
dtype: float64
```

Next steps: further explore org_id, perhaps apply KNN clustering algorithm to identify the natural grouping of users and create new features based on it. Apply other classification algorithms such as logistic regression, decision tree and gradient boosting.