Data Analysis Interview Challenge
Relax Challenge – Eleanor Hoyt

Problem: Use the users and engagement datasets to explore what features can be used to best predict user adoption.

Approach: This is a classification problem where we are interested in whether users become "adopted". Adopted is defined as "a user who has logged into the product on three separate occasions in at least one seven-day period". This approach included two phases:

Phase One: Generated data describing user adoption based on the engagement dataset. This included counting the number of logins over the previous seven days for each time a user engaged with the program. This adopted feature was then considered per user where a user was indicated as adopted (1) if they had a max engagement of 3 or more times over a 7 day period, otherwise the user was considered not adopted (0).

Phase Two: Join the adopted feature with the users dataset. Perform data wrangling and cleaning to prepare dataset for modeling. A simple logistic regression was used to predict the dependent variable "adopted".

General data observations:
There were some missing values in the users dataset including users who never logged in and users who were not referred by another user. These values were filled with 0 in both instances.

There are a number of users who signed up but never logged in. This resulted in an imbalanced dataset where there were many more not-adopted users compared to adopted.

Model performance: The logistic regression model trained for this dataset had an accuracy of 0.86. However, other important metrics including recall, precision, and F1 score were not high. This may be related to the imbalanced dataset or potentially an issue encountered during data wrangling and cleaning. This should be addressed first before any results are implemented into decision-making.

Important features: Based on the logistic regression model trained, a user's organization ID and users who signed up using Google Authentication had the largest positive influence on adoption. Conversely, the year users signed up and users who were invited to join another user's workspace had the largest negative influence on adoption. These observations may allow the company to invest in enhanced marketing to target particular organization IDs or those who signed up as part of a personal project to increase the adoption metric.

Future work: There is a lot of room for improvement on this model. With additional time and resources, a more in-depth exploratory data analysis phase may allow for additional interpretation of the dataset as well as engineering of additional features from existing

variables. For example, the email address and datetime variables have potential for many interested derived features.

Additionally, further model selection and tuning will result in a better-performing model that can more reliably produce insights on features that influence user adoption. Exploring additional model types such as decision trees or random forests may also result in better performance.