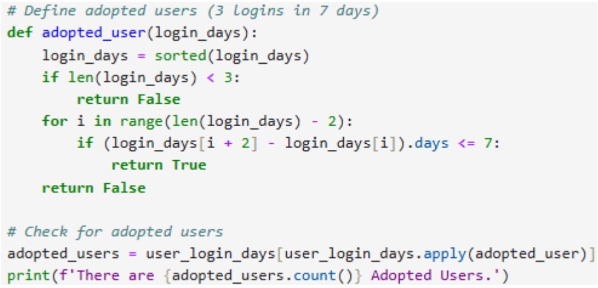
Relax Data Science Challenge Findings

For this take-home challenge, I was given two different datasets in which I had to identify “adopted users.” Adopted users are defined as a user who has logged into the product on three separate days within a seven-day period. The goal was to predict future user adoption based on various factors.

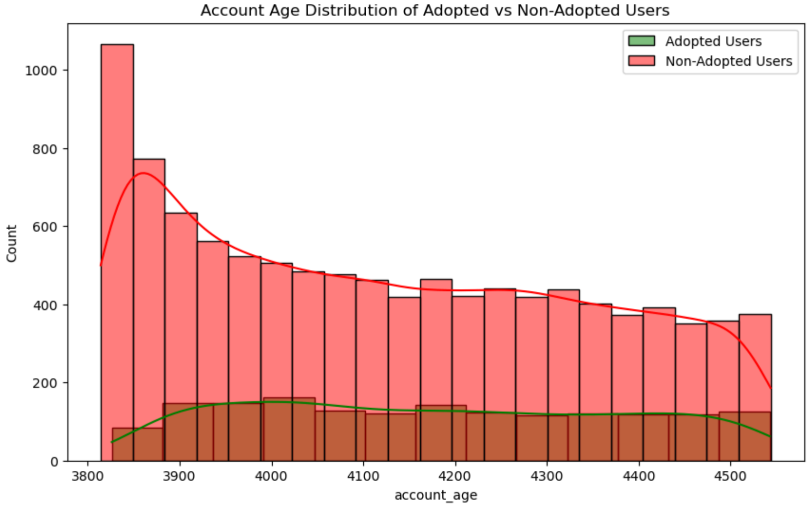
The initial step involved loading the datasets. The ‘takehome_users.csv’ could not be read, therefore having to use the encoding ‘latin1’. Timestamp columns were converted to datetime format for further analysis. My next step was to identify how many users were considered “adopted users.” To do so, I used the following code:



```python
# Define adopted users (3 logins in 7 days)
def adopted_user(login_days):
    login_days = sorted(login_days)
    if len(login_days) < 3:
        return False
    for i in range(len(login_days) - 2):
        if (login_days[i + 2] - login_days[i]).days <= 7:
            return True
    return False

# Check for adopted users
adopted_users = user_login_days[user_login_days.apply(adopted_user)]
print(f'There are {adopted_users.count()} Adopted Users.')
```

There are 1,656 out of 12,000 users that are considered as “adopted users.” I merged both datasets and filled in any missing values. To understand the distribution of adopted and non-adopted users, I visualized the data using a histogram. The plot revealed that the number of non-adopted users peaked and then declined, while the number of adopted users increased, plateaued, and subsequently declined.



To predict future adopted users, I tested out two machine learning models: Logistic Regression and Random Forest. The Logistic Regression model had a slightly better recall for the “True” class, but the Random Forest had a slightly better precision for the “True” class. Both models did, however, have login_count as the highest indicator of determining “adopted users”.