Read both the CSV files using “name = pd.read\_csv(‘file\_location’)” and print the Data. According to the 6 jars of Machine Learning, 1st is Data jar. Check for missing values, duplicates, outliers and encode in desired format for analysis. In the given assignment, creation\_time and last\_session\_creation\_time columns are in object format, whereas data entered is in Date-Time format. Convert the data to Date-Time format both in ‘takehome\_user\_ engagement’ and ‘takehome\_users’. In the ‘takehome\_users’ data, try to avoid or drop name and email columns, as considering them won’t have much significance on analysis. Check for duplicates and null or nan values in the dataset.

#Changing object to Date/Time format df1.creation\_time = pd.to\_datetime(df1['creation\_time']) df1.last\_session\_creation\_time=pd.to\_datetime(df1['last\_session\_creation\_time']) #Changing the column name df1['user\_id']=df1['object\_id'] #After creation of the new columns drop the exisiting ones df1.drop('object\_id', axis=1, inplace=True) #Drop data which aren't useful for analysis/do not consider the columns which are not useful df1.drop(['name','email'], axis=1, inplace=True) df1.head()

For the given task “adopted user”: adopted\_user\_dict={} user\_ids=list(set(takehome\_user\_engagement['user\_id'])) for i in range(len(user\_ids)): user\_id=user\_ids[i] reduced\_df=takehome\_user\_engagement[(takehome\_user\_engagement['user\_id']==user\_id)&(weeks.isin(weeks[weeks.duplicated()]))] week\_counts=reduced\_df.year\_week.value\_counts()[reduced\_df.year\_week.value\_counts()>2] three\_logins=reduced\_df[reduced\_df.year\_week.isin(list(week\_counts.index))] three\_logins=three\_logins[~three\_logins.duplicated()] adopted\_user\_dict[str(user\_id)]=len(three\_logins)

The above code is used to check if there are variables for the given task, i.e.,User logged in 3 or more times in a week. We are left with 1445 such users for analysis.

Check for correlation between the feature and the target variable for training the model. According to the data given and the task, there are only 1445 users from 12000 users. There is huge data loss, therefore the result obtained is biased or will not return correct value. Read both the CSV files using “name = pd.read\_csv(‘file\_location’)” and print the Data. According to the 6 jars of Machine Learning, 1st is Data jar. Check for missing values, duplicates, outliers and encode in desired format for analysis. In the given assignment, creation\_time and last\_session\_creation\_time columns are in object format, whereas data entered is in Date-Time format. Convert the data to Date-Time format both in ‘takehome\_user\_ engagement’ and ‘takehome\_users’. In the ‘takehome\_users’ data, try to avoid or drop name and email columns, as considering them won’t have much significance on analysis. Check for duplicates and null or nan values in the dataset.

#Changing object to Date/Time format df1.creation\_time = pd.to\_datetime(df1['creation\_time']) df1.last\_session\_creation\_time=pd.to\_datetime(df1['last\_session\_creation\_time']) #Changing the column name df1['user\_id']=df1['object\_id'] #After creation of the new columns drop the exisiting ones df1.drop('object\_id', axis=1, inplace=True) #Drop data which aren't useful for analysis/do not consider the columns which are not useful df1.drop(['name','email'], axis=1, inplace=True) df1.head()

For the given task “adopted user”: adopted\_user\_dict={} user\_ids=list(set(takehome\_user\_engagement['user\_id'])) for i in range(len(user\_ids)): user\_id=user\_ids[i] reduced\_df=takehome\_user\_engagement[(takehome\_user\_engagement['user\_id']==user\_id)&(weeks.isin(weeks[weeks.duplicated()]))] week\_counts=reduced\_df.year\_week.value\_counts()[reduced\_df.year\_week.value\_counts()>2] three\_logins=reduced\_df[reduced\_df.year\_week.isin(list(week\_counts.index))] three\_logins=three\_logins[~three\_logins.duplicated()] adopted\_user\_dict[str(user\_id)]=len(three\_logins)

The above code is used to check if there are variables for the given task, i.e.,User logged in 3 or more times in a week. We are left with 1445 such users for analysis.

Check for correlation between the feature and the target variable for training the model. According to the data given and the task, there are only 1445 users from 12000 users. There is huge data loss, therefore the result obtained is biased or will not return correct value. The most important features are org\_id, creation\_day, creation\_year and creatio\_month.