Q1.

1. Assuming both datasets to be imported as csv files, we first import the datasets with pd.csv() as two dataframes df and df2
2. Drop rows with device id as Nan in both datasets since id uniquely identifies the item, this is done through df.loc[df.index.dropna()] and df2.loc[df2.index.dropna()]
3. Merge the two datasets based on device id column: df = pd.merge(df, df2, how='inner', on='device_id')

Q2.

A.K-means clustering can be used here since this algorithm categorises data points into cluster and eventually find the number of optimal number of groups required.

B.
1. determine the number of clusters, start with 2 in this case for relatively small size of sample
2. determine the value of each cluster, here we randomly set cluster 1 to be 10 and cluster 2 to be 20
3. determine which cluster each data point belongs to, in this case B and D is closest to cluster 1 and A, C, E closest to cluster 2
4. calculate the center of each cluster as the mean of the data points in each cluster and change the value of each cluster to the mean, in this case, cluster 1 = (7+12)/2 = 9.5, and cluster 2 = (18+22+24)/3 = 21.3
5. repeat step 3 with the newly calculated centers of each cluster, there Is no need to repeat step 4 since there is no change to which cluster each data point belong to

Q3.
10-fold cross-validation means we partition the training samples randomly into 10 equal size partitions, then select one partition as validation set while the other 9 partitions as training set, to train the model, we repeat with each sub partition therefore we implement cross validation 10 times hence N1 = 10. And since we train the model with time with the remaining 9 partitions which equals 90 training samples, N2=90. Lastly, we test the model with the validation set of 10 training samples, N3 = 10

Q4.
Precision = TP/(TP+FP) = 8 / (2+8) = 0.8
Recall = TP/(TP+FN) = 8 / (8+12) = 0.4
F1-Score=2*Precision*Recall / (Precision + Recall) = 2*0.8*0.4/(0.8+0.4) = 0.53

Q5.

A. Basic authentication scheme can be used here where user authenticate with username and password, then a token can be issued to the user upon registration which is used in user's subsequent request and password is not passed around. One example is Json Web Token, which consist of 3 parts JSON data encoded and signed, used by user for any subsequent request, since its signed, validation from server can be done without database lookup

B. Api monitoring tool and rate limiting algorithm can be used, for example, return 429 Too Many Requests HTTP response code if requests are coming in too quickly

Q6.

```
201 Created
Location: api.coffeehouse.com/order?1234
Content-Type: application/xml

<order>
<drink>latte</drink>
</order>
```

Q7.

A. Use K nearest Neighbor, since we are given input data with features and known classifiers, we can then split the samples into training set to train model and validation set to test the model.
   Limits: slow at query
   Can create misleading result from irrelevant features
   Performance of training rely on greater size of samples

B. Identifier of people close to device, distance, duration of contact, GPS location changed to city and suburb, age

Q8.

A. Firstly, it cannot be content-based recommender system, which will recommend movie C to U1 as it most number of matched columns to U1's interest (release decade and director). Also movie B has the most number of reviews, an indication of its highest popularity so it could be collaborative. Hence R could be User-based or Item-based collaborative filtering

B. R will recommend movies to new user U2 if R has the profile or preference of U2 and if there are records of watched movies by U2 since R cannot recommend without any information from U2