Name: Shunyang Li

ZID: z5139935

Question 1

1. We need to drop the nan value in dataset1 and dataset2 by using `df.dropna()`, because in dataset 1 the Device ID is very import, so the value can not be nan
2. In dataset 1 we can convert *Quality Tested Date/Time* into datetime formate by using `pd.to_datetime()`, it could make the query more easier and it can avoid the error when execute query language.
3. In dataset 2, we also need to convert *Support Ticket Date/time* into datetime formate by using `pd.to_datetime()`. Before converting, we need to remove the chars *AM* or *PM*, otherwise it can not convert successfully.
4. The final step we can merge these two sets by using `pd.merge(ds1, ds2, on='Device ID')`

But we need to consider the storage problem, so, we need to remove the duplicate value after merge these two sets.

Question 2

A)

We can use **hierarchical clustering** algorithm, because it need to be divided into many groups, and we do not known how many goups we exactly need, that means we can not use k-means to solve.

B)

1. Calculate the distance between different points by using Euclidean distance.

   For example, the distance between A and B is $\sqrt{(18-7)^2} = 11$

	A	**B**	**C**	**D**	**E**
A	0	11	4	6	6
B	11	0	15	6	17
C	4	15	0	10	2
D	6	6	10	0	12
E	6	17	2	12	0

2. After get the table, we can find the distance between C and E is shortest. So, we can combine these two points as (C, E) and the calculate the distance bewteen (C, E) and A ($\frac{disC-A+disE-A}{2}$). And then **repeat** the rest points, unitl no isolation point.
3. And then we can get all the groups

C)

In part B), after merging all the points into groups, and then we can find the number of groups. There should have **two groups(ACE, BD).**

Question3

N1 = 10, N2 = 90, N3 = 10. According to the definition of cross validation, if the set is N, and the train set should be N-1, and the test set should 1. Because it is 10-Fold (10 fold, each fold have 10 data), each fold shoud have k-1 train set and 1 test set (where k = 10), so, the training set should be 90 and both validation set and the times of computing error should be 10.

Question 4

The formula of precision, recall and f1-score are:

$$precision = \frac{true\ true}{true\ true + false\ true}$$
$$recall = \frac{true\ true}{true\ true + false\ false}$$
$$F1 = \frac{2 * precision * recall}{precision + recall}$$

So, for precision we can get $\frac{8}{8+2}$ = 0.8, recall = 8/(8+11) = 8/19, and F1 = (2 * 8/10 * 8/19) / (8/10+8/19) = 0.55.

$$precision = \frac{8}{8+2} = \frac{8}{10}$$
$$recall = \frac{8}{8+11} = \frac{8}{19}$$
$$F1 = \frac{2 * \frac{8}{10} * \frac{8}{19}}{\frac{8}{10} + \frac{8}{19}} \approx 0.55$$

Question 5

A)

We can use **JWT token authentication**. Because for every different user it will generate different kind of tokens and different tokens can access different API. For example, in a shopping web, after the user login, it will return a jwt token, and by using this token the user can access shopping cart, but the user can not access to the backend management system. So, it can minimize the attack window. Generally for api keys, one api key can access all the api, so, api keys is not so safe. We should choose JWT token.

B)

We should use **API keys**. Because API keys can help to limit the behavior of user. For exmaple, an IT company sale an API with different price, the higher the price, the faster api, so, the company can use api keys to limit the ratin.

Question 6

```
Status code: 201 Created
Location: xxxx/order?xxxx
Content-Type: application/xml
<order xmlns=urn:coffeehouse>
  <drink>latte</drink>
  <order_id>xxxx</order_id>
  <link rel="payment" href="/payment/order?xxxx">
</order>
```

201 status code means created. And the api should return about how to access the order the the order id.

Question 7

A)

We can use **knn** to solve. Because the data has label, so we can not use unsupervised learning. As we known knn is non-parametric, that means knn does not make any assumptions about the data and the model structure will be built according to the data. knn algorithm based on similarity analysis, and people are infected with MOVID-19 and the place, the communication time is related, it is likely to be infected in a similar place. So knn is the best choice.

B)

For x: Distance, Duration of contact , GPS Location and Age.

For y: Infection Tested

Question 8

A)

R should be User-based collaborative filtering or Item-based collaborative filtering. If it uses Content-based recommender system the result should be C not B. Because the keywords are 2000s, D2 and Comedy. Content-based recommender system is based on the keywords.

B)

R cannot recommend. Because we donot known what kind of movies and the released year U2 likes, so the R cannot make recommend. However,