Q 1.

Below are number of outliers present in 37 attributes.

﻿0 0

1 12

2 33

3 2

4 16

5 12

6 39

7 24

8 11

9 15

10 1

11 24

12 39

13 22

14 17

15 13

16 10

17 2

18 1

19 13

20 26

21 41

22 14

23 6

24 16

25 25

26 34

27 15

28 30

29 16

30 8

31 8

32 44

33 27

34 28

35 26

36 26

37 39

38 41

As the data is continuous and with outliers, we replaced the missing value by the mean of each attribute.

For normalizing the values, used ﻿sklearn.preprocessing ﻿ normalize which normalizes dataset and translates each feature individually such that it is in the given range on the training set that is between 0 and 1.

Q 2.

For all possible values of k, we will use k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster center.

sum of squarted distances or value of inertia becomes zero as ‘k’ increases.

We get the following graph of sum of squared distances vs k

A close up of a map

Description automatically generated

As the name of the method suggests, above graph looks like an arm with elbow at k=4 as after 4, change in the value of inertia is not significant.

Q 5.

**Modified k-means on given data:**

﻿{2: 1, 3: 53, 0: 225, 1: 248}

We can verify that there are number of clusters which are 4 has same data which means our prediction of k=4 is true

**k-means on reduced dimensions by PCA:**

After applying PCA on the dataset, we choose number of principle components as 5 from the below graph as they explain the majority of variance from our data.

A screenshot of a cell phone

Description automatically generated

As the name of the method suggests, above graph looks like an arm with elbow at k=4 as after 4, change in the value of inertia is not significant.

A close up of a map

Description automatically generated

﻿{3: 1, 0: 55, 2: 227, 1: 244}

We can verify that there are number of clusters which are 4 has same data which means our prediction of k=4 is true.

We can conclude from above that there is no difference as number of clusters remain the same on given dataset and on the reduced data by PCA.