Question 2)

K-Means Clustering on a Multi-Class and Multi-Label Data Set

a) Determination of the optimal k-value

To determine the number of clusters to be used in the clustering process, I used a technique well-known as Elbow method.

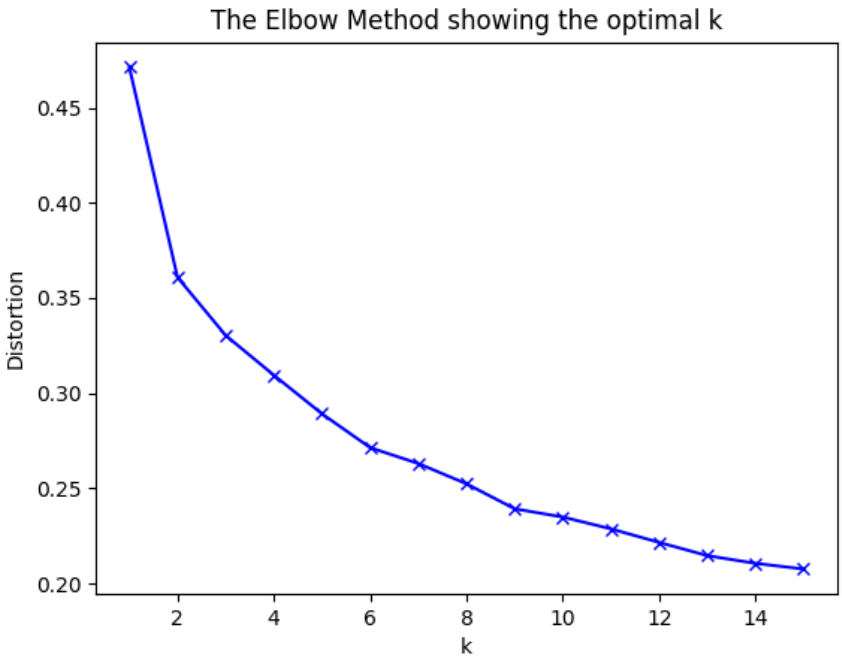
This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified.

After performing the elbow criterion test with 15 different values of k, I obtained the optimal value of **k to be 2.**

Here are the results obtained:

Distortions:
[0.6746431644135847, 0.5190084106643769, 0.4788931840662131, 0.4501682018094673, 0.4140918899382861, 0.3913976428644167, 0.3800351118686867, 0.36289188653734666, 0.34939280078645, 0.3417315625359848, 0.32726085355587753, 0.3232530771702386, 0.3121239227558629, 0.30649315447967757, 0.30424784594256005]

Elbow at k: 2



b. Performing k-means Clustering

After determining the value of k to be 2 for our problem, I proceeded by running the k-means clustering for the dataset.
Post-execution I was able to cluster the data into 2 different clusters, namely Cluster-0 and Cluster-1.

I then determined the label-family for each cluster by reading the true labels of the data-points in the corresponding cluster and then took a majority poll.

In a similar fashion I was able to perform this task of determining the label-genus and label-species for each cluster.

Here are the results I obtained:

```
#######   k-Means Clustering for Anuran Calls   #######

Fitting...

Centers of Centroids:

Cluster- 0 : [ 9.81912190e-01  3.68580302e-01  4.11499956e-01  3.38745161e- 01
  6.24981505e-02  1.66989514e-01  1.12450307e-01 -4.54103922e-02
 -3.26806081e-03  7.66570254e-02  3.29197462e-02 -9.09333134e-03
 -1.41984527e-02  3.32633353e-02  4.57955472e-02 -6.01663237e-03
 -1.87769057e-02  4.60654835e-03 -4.62719786e-04  1.21530071e-02
  1.62062920e-02 -1.24504200e-02]

Cluster- 1 : [ 0.99785469  0.27859975  0.21097633  0.55321929  0.19157641     0.02890766
-0.11521227  0.04465765  0.25965825  0.03534465 -0.26424321          0.09582125         0.31604205
-0.11173041 -0.24924994  0.090127    0.19610765        0.01090299     -0.09847145     -0.11862293
0.05841477  0.18755761]
```

Clustering for label-Family...

Cluster-0-label by Majority Polling: **Hylidae**
Cluster-1-label by Majority Polling: **Leptodactylidae**

Clustering for class-Genus...

Genus Cluster-0-label by Majority Polling: **Hypsiboas**
Genus Cluster-1-label by Majority Polling: **Adenomera**

Clustering for class-Species...

Species Cluster-0-label by Majority Polling: **HypsiboasCordobae**
Species Cluster-1-label by Majority Polling: **AdenomeraHylaedactylus**

Evidently, for *Cluster-0* **Hylidae**, **Hypsiboas** and **HypsiboasCordobae** are the classes for the labels family, genus and species respectively.

And **Leptodactylidae**, **Adenomera** and **AdenomeraHylaedactylus** correspond to the classes for labels family, genus and species respectively in *Cluster-1*.

c. Hamming Loss of the Majoirty Triplet

In the previous part of this question we determined a class for labels family, genus and species for every cluster.

These correspond to a majority triplet for that cluster.

Cluster 0:**Hylidae**, **Hypsiboas** and **HypsiboasCordobae**

Cluster 1: **Leptodactylidae**, **Adenomera** and **AdenomeraHylaedactylus**

I then read the true labels of the data-points and compared them to labels determined by the algorithm. I then determined the hamming loss for each label and also an average hamming loss for the dataset.

Out of curiosity I even calculated the Zero-One Loss for the dataset.

The following results were obtained:

```
#######  k-Means Clustering for Anuran Calls  #######

Fitting...

Clustering for label-Family...

        Cluster-0-label by Majority Polling: Hylidae
        Cluster-1-label by Majority Polling: Leptodactylidae
        Hamming Loss: 0.2334954829742877

Clustering for class-Genus...

        Genus Cluster-0-label by Majority Polling: Hypsiboas
        Genus Cluster-1-label by Majority Polling: Adenomera
        Hamming Loss: 0.2982626824183461

Clustering for class-Species...

        Species Cluster-0-label by Majority Polling: HypsiboasCordobae
        Species Cluster-1-label by Majority Polling: AdenomeraHylaedactylus
        Hamming Loss: 0.3638637943015983
```

Cluster-0 Majority Triplet: Hylidae Hypsiboas HypsiboasCordobae
Cluster-1 Majority Triplet: Leptodactylidae Adenomera AdenomeraHylaedactylus

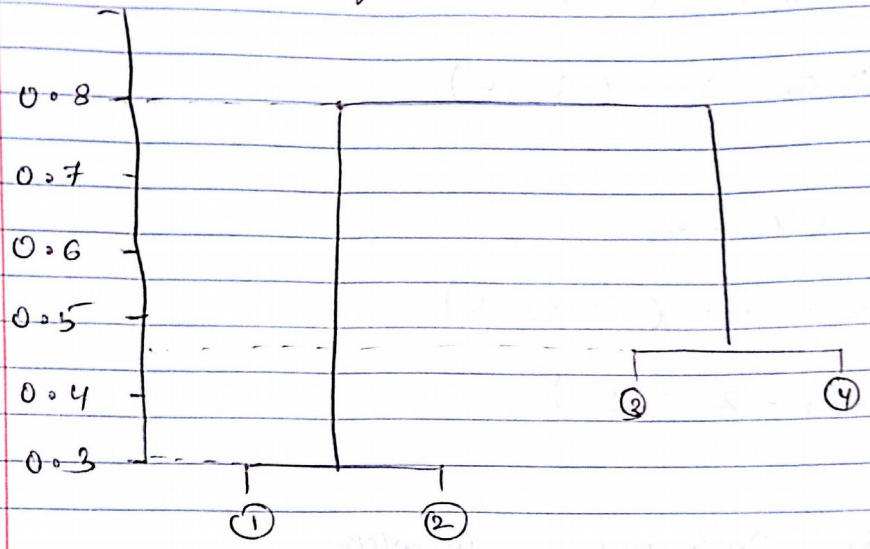
Average Hamming Loss: 0.2985406532314107
Average Zero-one Loss: 0.3638637943015983

The algorithm performs fairly, this is a tribute to the higher number of samples in the data-set.

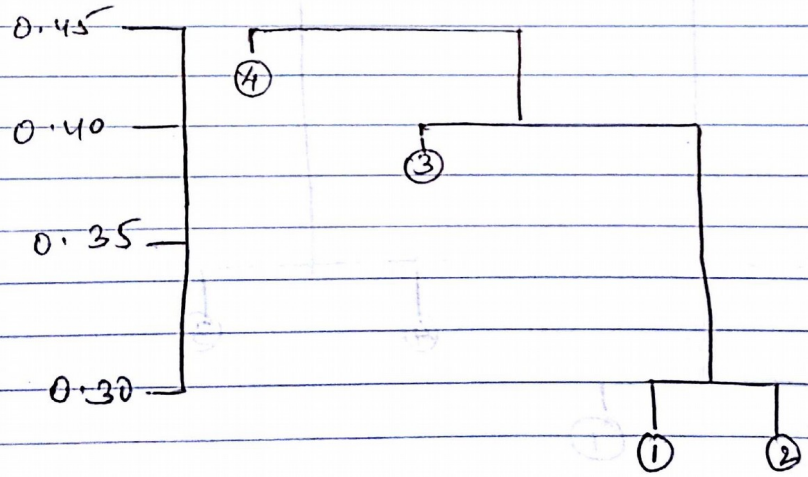
Question 3)

ISLR – 10.7.2.

(a). Cluster Dendogram: Complete Linkage.



(b) Cluster Dendogram: Single Linkage



c) We will have 2 clusters :

Cluster - 1 : (1, 2)

Cluster - 2 : (3, 4)

d) We will have :

Cluster - 1 : ((1, 2), 3)

Cluster - 2 : (4)

e) Cluster Dendogram : Complete.