Assignment 6

Question 1:

What is the difference between supervised and unsupervised learning? Give one example (not a technique or algorithm) to demonstrate.

→ Supervised Learning: Supervised learning is a task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector/features) and the desired output value (also called the supervisory signal).

Example:

- Consider, we have a basket and filled it with various kinds of fruits like apple, banana, grape and cherry. Our task is to arrange them into groups.
- We already learn from your previous work about the physical characters of fruits
- So, arranging the same type of fruits at one place is easy.
- In data mining terminology, the earlier work/experience is called as training the data
- You already learn the things from your train data. This is because of response variable
- Response variable means just a decision variable
- Suppose you have taken a new fruit from the basket then you will see the size, color, and shape of that particular fruit.
- If size is Big, color is Red, the shape is rounded shape with a depression at the top, you will confirm the fruit name as apple and you will put in apple group.
- Likewise, for other fruits also.

Unsupervised Learning: The problem of an unsupervised learning task is trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution.

- Suppose you have a basket and it is filled with some different types of fruits and your task is to arrange them as groups.
- This time, you don't know anything about the fruits, honestly saying this is the first time you have seen them. You have no clue about those.
- So, how will you arrange them?
- What will you do first?
- You will take a fruit and you will arrange them by considering the physical character of that fruit.
- Suppose you have considered color.
    - Then you will arrange them on considering base condition as color.
    - Then the groups will be something like this.
        - RED COLOR GROUP: apples & cherry fruits.
        - GREEN COLOR GROUP: bananas & grapes.
- So now you will take another physical character such as size.
    - RED COLOR AND BIG SIZE: apple.
    - RED COLOR AND SMALL SIZE: cherry fruits.
    - GREEN COLOR AND BIG SIZE: bananas.
    - GREEN COLOR AND SMALL SIZE: grapes.

Question 2:

Explain how/why k-means is (almost) guaranteed to converge.

→

In k-means clustering, the first step is to select 'K' random cluster centers. The algorithm then moves the cluster centers around in space to minimize cost i.e. RSS. RSS is the objective function in K-means and the goal is to minimize it. Since 'N' is fixed, minimizing RSS is equivalent to minimizing the average squared distance, a measure of how well centroids represent their documents.

Steps,

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat {

   For every i, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set

   $$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

   }

For each iteration of the algorithm, we produce a new clustering with lower cost with respect the previous clustering, by adjusting the centroid 'μ' with the new mean values.

3. Cost is reduced iteratively by repeating two steps until a stopping criterion is met i.e. reassigning points to the cluster with the closest centroid; and re-computing each centroid based on the current members of its cluster.

4. Since cost is being reduced at each step of the algorithm, at certain point, assignment of points to each cluster does not change between iterations and centroids 'μ' do not change between iterations

5. In this case cost RSS is below the threshold value to stop the iterations and terminate it.

6. For small value of cost RSS, indicates that we are close to convergence.

Since there is only a finite set of possible clustering's, a monotonically decreasing algorithm will eventually arrive at a (local) minimum. Take care, however, to break ties consistently, e.g., by assigning a point to the cluster with the lowest index if there are several equidistant centroids. Otherwise, the algorithm can cycle forever in a loop of clustering that have the same cost.

Thus, it proves the convergence of K-means, there is unfortunately no guarantee that a global minimum in the objective function will be reached. This is a particular problem if a data set contains many outliers, points that are far from any other documents and therefore do not fit well into any cluster.

Question 3:

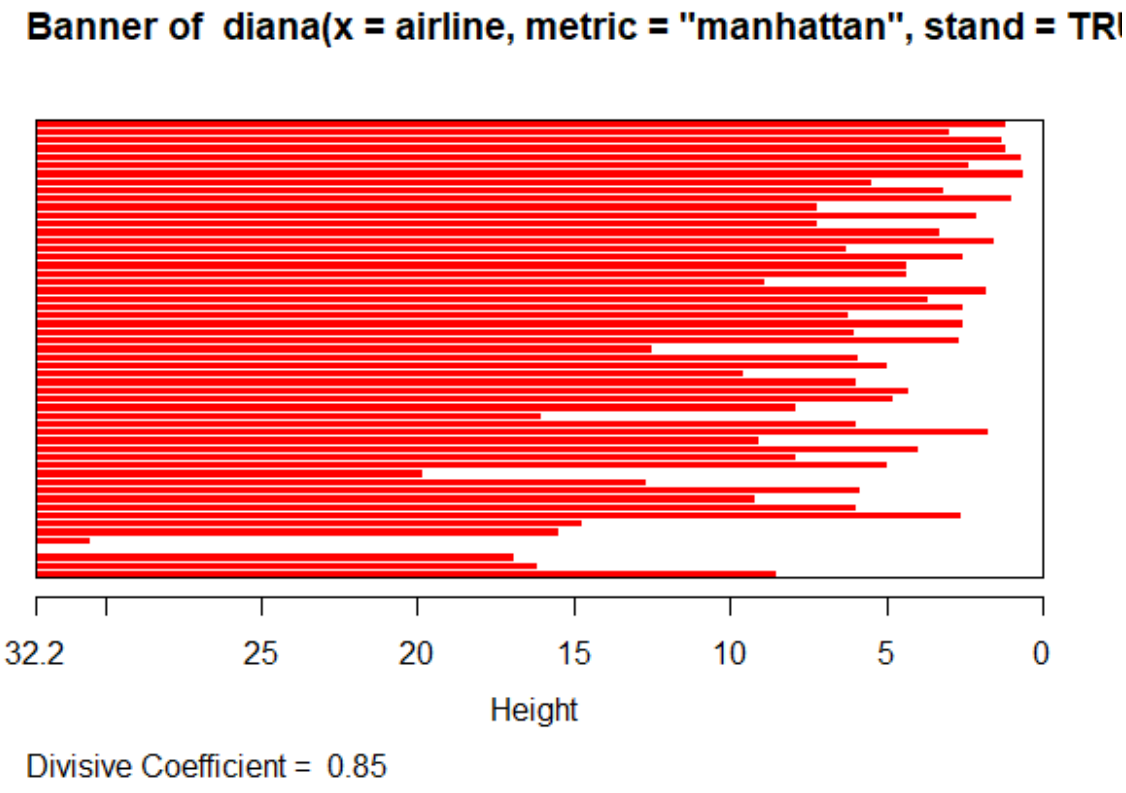
Airline Safety data set

Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?

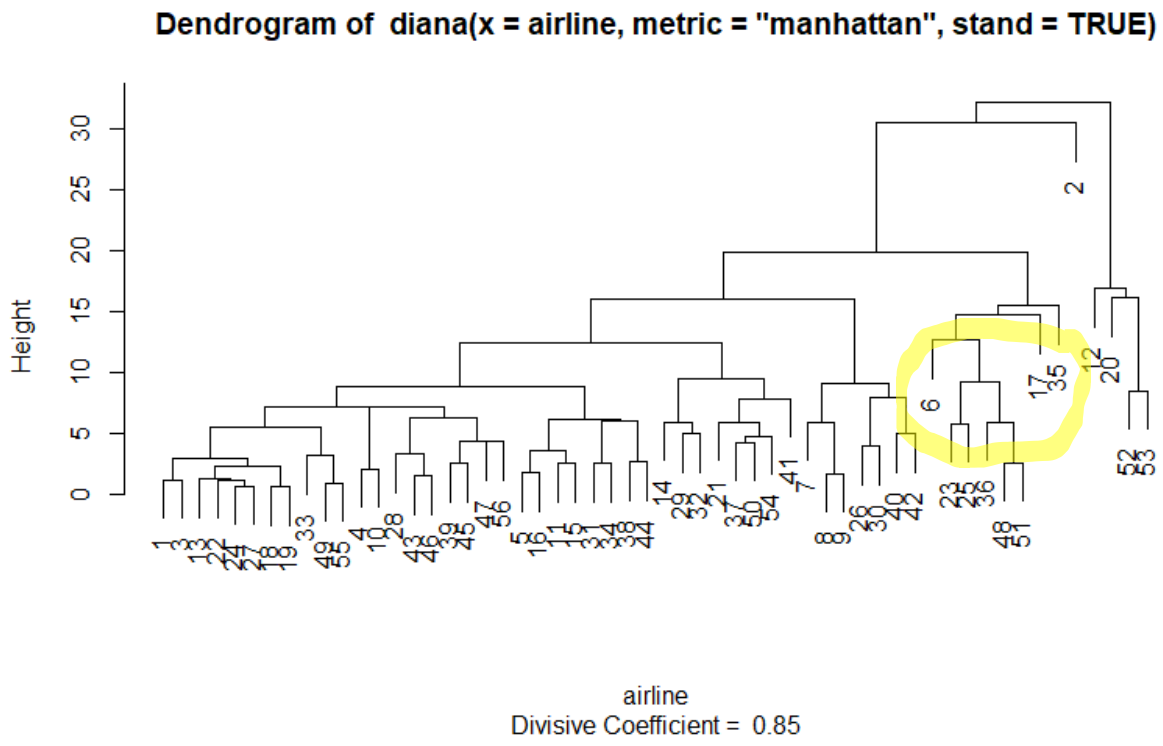
R Code

Divisive Method:

```
> airline <- read.csv(file.choose(), header = TRUE, sep = ",")
> View(airline)
> fit <- diana(airline, metric = "manhattan", stand = TRUE)
> plot(fit)
Hit <Return> to see next plot:
```



Hit <Return> to see next plot:



From the above Dendrogram, we can see that Malaysian airline (35), China Airline (17) and Air France (6) are clustered together.

	ï..airline	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fatalities_85_99	incidents_00_14	fatal_accidents_00_14	fatalities_00_14
35	Malaysia Airlines	1039171244	3	1	34	3	2	537
17	China Airlines	813216487	12	6	535	2	1	225
6	Air France	3004002661	14	4	79	6	2	337

From the data, these 3 airline shows high fatalities between 2000 to 2014. Also, their past fatalities are high for 1985 to 1999. Thus, we can say that travelers should avoid these airlines.

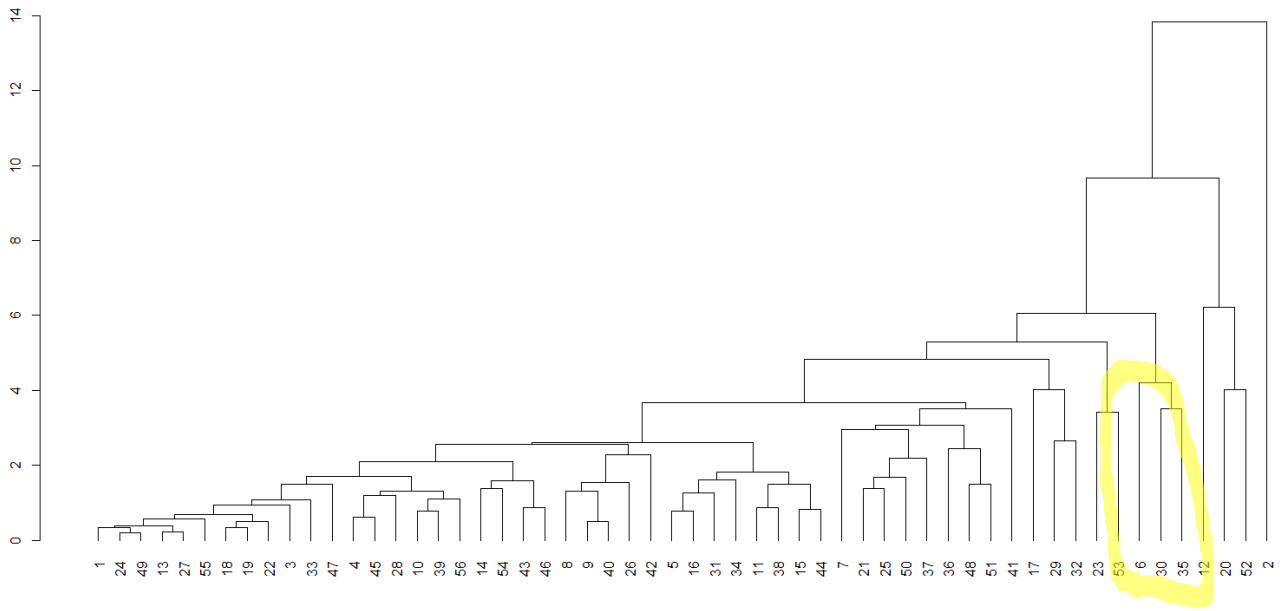
Agglomerative Approach:

```
> airline <- read.csv(file.choose(), header = TRUE, sep = ",")
```

```
> response <- airline[,1]
> predictor <- airline[,2:(dim(airline)[2])]
```

Method is ‘Average’

```
> library(cluster)
>
> clusters <- agnes(x=predictor, diss = FALSE, stand = TRUE, method = "average")
> DendCluster <- as.dendrogram(clusters)
> plot(DendCluster)
```



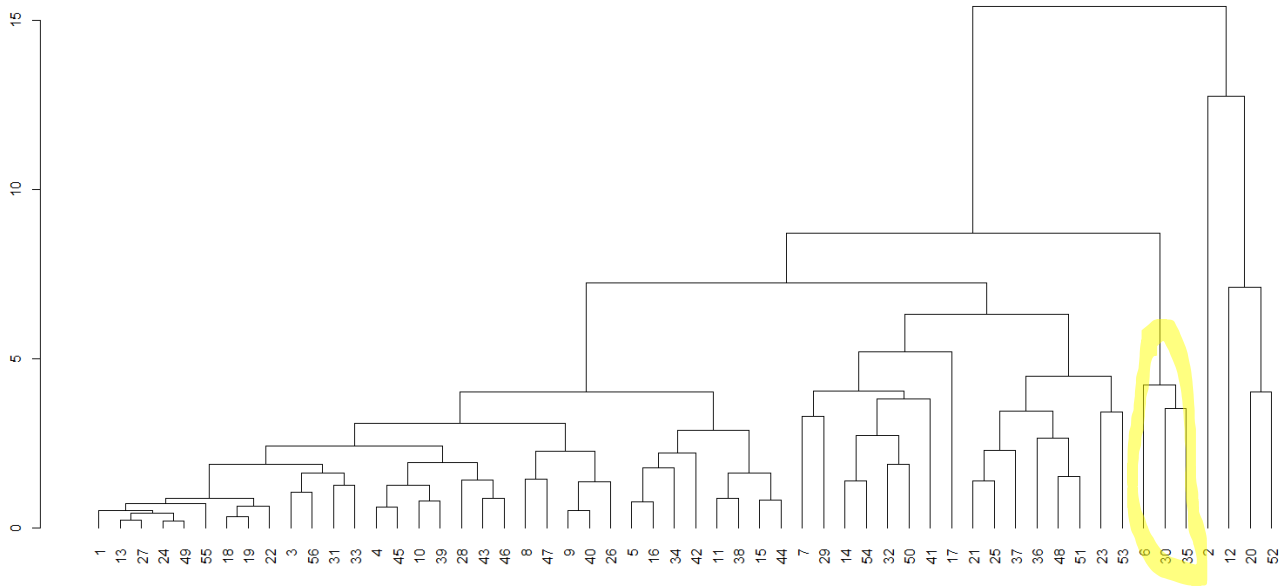
From the above Dendrogram, we can see that Malaysian airline (35), Kenya Airways (30) and Air France (6) are clustered together.

	ï..airline	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fatalities_85_99	incidents_00_14	fatal_accidents_00_14	fatalities_00_14
35	Malaysia Airlines	1039171244	3	1	34	3	2	537
30	Kenya Airways	277414794	2	0	0	2	2	283
6	Air France	3004002661	14	4	79	6	2	337

From the data, these 3 airline shows high fatalities between 2000 to 2014. Also, for Malaysia Airline and Air France, their past fatalities are high for 1985 to 1999. Thus, we can say that travelers should avoid these airlines.

Method is ‘Complete’

```
> clustersComplete <- agnes(x=predictor, diss = FALSE, stand = TRUE, method = "complete")
> DendClusterComplete <- as.dendrogram(clustersComplete)
> plot(DendClusterComplete)
```



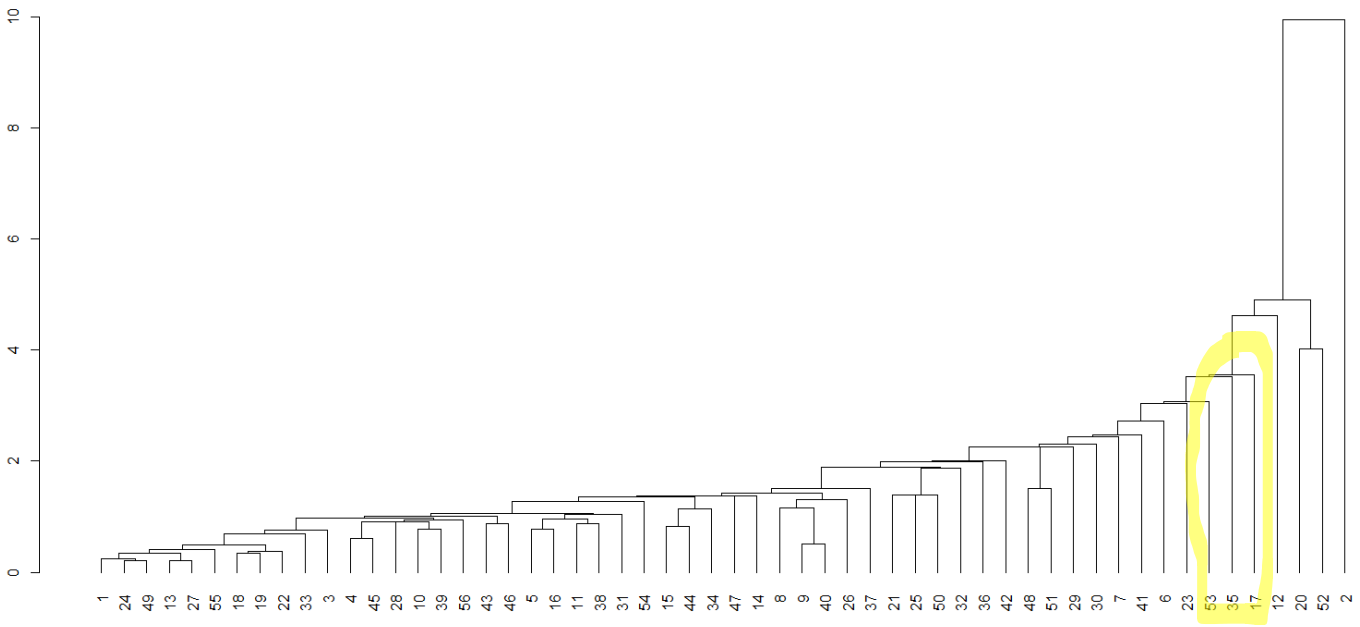
From the above Dendrogram, we can see that Malaysian airline (35), Kenya Airways (30) and Air France (6) are clustered together.

	ï..airline	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fatalities_85_99	incidents_00_14	fatal_accidents_00_14	fatalities_00_14
35	Malaysia Airlines	1039171244	3	1	34	3	2	537
30	Kenya Airways	277414794	2	0	0	2	2	283
6	Air France	3004002661	14	4	79	6	2	337

From the data, these 3 airline shows high fatalities between 2000 to 2014. Also, for Malaysia Airline and Air France, their past fatalities are high for 1985 to 1999. Thus, we can say that travelers should avoid these airlines.

Method is ‘Single’

```
> clustersSingle <- agnes(x=predictor, diss = FALSE, stand = TRUE, method = "single")
> DendClusterSingle <- as.dendrogram(clustersSingle)
> plot(DendClusterSingle)
```



From the above Dendrogram, we can see that Malaysian airline (35), US Airways / America West* (53) and China Airline (17) are closed to each other.

	ï..airline	avail_seat_km_per_week	incidents_85_99	fatal_accidents_85_99	fatalities_85_99	incidents_00_14	fatal_accidents_00_14	fatalities_00_14
35	Malaysia Airlines	1039171244	3	1	34	3	2	537
53	US Airways / America West*	2455687887	16	7	224	11	2	23
17	China Airlines	813216487	12	6	535	2	1	225

From the data, airlines (35) and (17) shows high fatalities between 2000 to 2014. Also, for US Airways and China Airline, their past fatalities are high for 1985 to 1999. Thus, we can say that travelers should avoid these airlines.