Homework 7

Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. Only **one person** in the group should submit to Canvas.
Due: check on Canvas.

1. What would the 1-Nearest Neighbor approach predict is the class when the feature set is {Soymilk, Apple, Eggs} and the following data is used as the training data (same table as the classwork)? Use the Jaccard similarity to calculate neighbors. Show your calculations. (Note: do NOT write any code for this problem. The answers are to be computed by hand.)

	Feature set	*Class*
A	{Apple, Banana, Soymilk, Yogurt}	Vegetarian
B	{Apple, Peanuts, Yogurt}	Vegetarian
C	{Tomatoes, Potatoes, Yogurt}	Non-Vegetarian
D	{Apple, Tomatoes, Potatoes}	Non-Vegetarian

J(A) = (Intersection: {Apple, Soymilk}) / (Union: {Apple, Banana, Soymilk, Yogurt, Eggs}) = 2/5

J(B) = (Intersection: {Apple}) / (Union: {Apple, Soymilk, Peanuts, Yogurt, Eggs}) = 1/5

J(C) = (Intersection: {}) / (Union: {Apple, Soymilk, Tomatoes, Potatoes, Yogurt, Eggs}) = 0/6

J(D) = (Intersection: {Apple}) / (Union: {Apple, Tomatoes, Soymilk, Potatoes, Eggs}) = 1/5

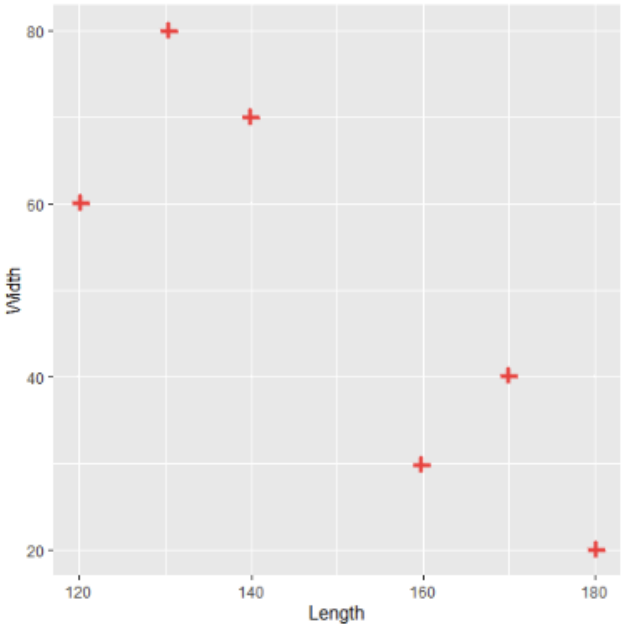
It will be Vegetarian since J(A) has the most significant Jaccard value.

2. Consider the following dataset. (Note: do NOT write any code for this problem. The answers are to be computed by hand and marked on the graph. You can visually guess some of the answers.)

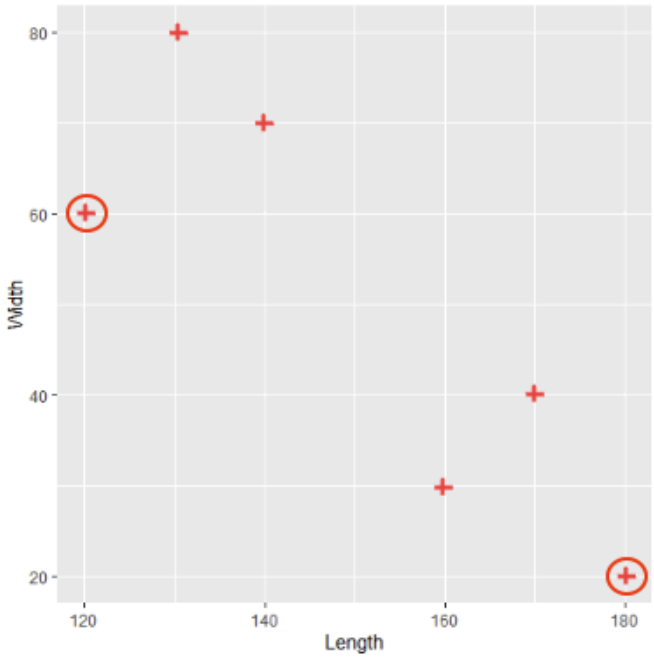
Length	120	140	130	170	160	180

Width	60	70	80	40	30	20

a) Mark the data points on the graph below (*use '+' to indicate each point*).



b) Let k=2. Let one of the two initial centers be (Length=120, Width=60). Select the second center using the **Farthest Distance Heuristic**. Indicate the two centers on the graph (*circle the centers*).



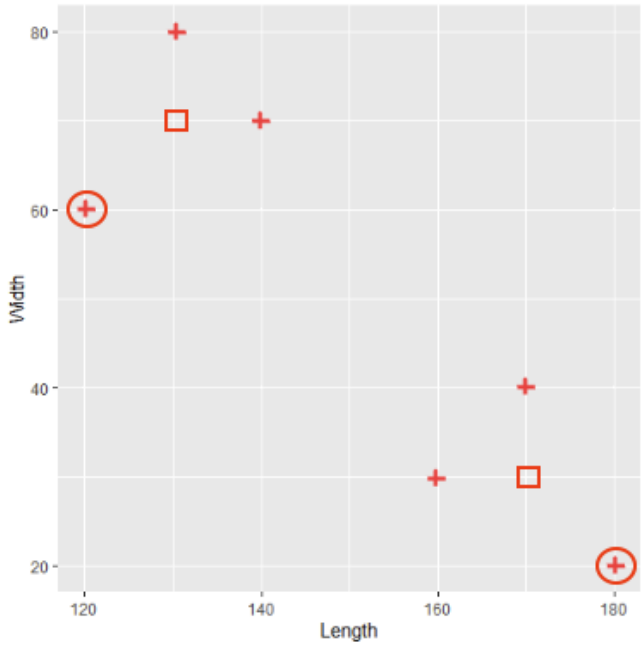
c) Recompute the centers after the first iteration of the k-means algorithm.

New Center 1:
(120 + 140 + 130) / 3 = 130
(60 + 70 + 80) / 3 = 70

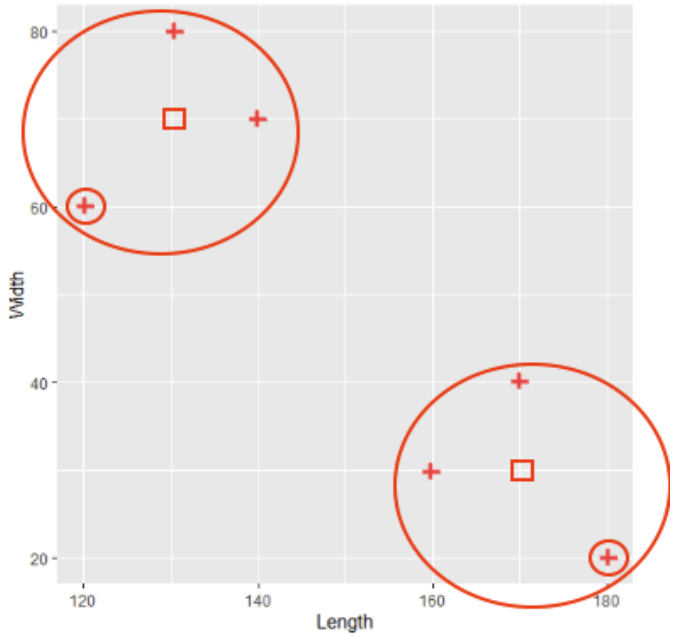
New Center 2:
(170 + 160 + 180) / 3 = 170
(40 + 30 + 20) / 3 = 30

The new centers after the first iteration are as follows:
New Center 1: (130, 70)
New Center 2: (170, 30)

Indicate the two new centers on the graph (*mark new centers with squares*).



d) What are the two clusters after this first iteration? *Draw two ovals, each containing all the points in one cluster in the graph above.*



e) Will the k-means algorithm terminate after this first iteration or will it continue? Answer in 1-2 sentences.

The k-means algorithm will continue after the first iteration because the centers have changed. The algorithm will iterate until the centers no longer change significantly between iterations or a predefined stopping criterion is met.

f) If a new point (Length=140, Width=60) is given, to which cluster will it belong?

It will belong to the cluster with the current center (130, 70) because it is the closest

3. Consider the file `breast-cancer-wisconsin.csv` (in the Datasets module on Canvas) which contains "Features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass."[1] The goal is to cluster the data based on the features to distinguish Benign and Malignant cases.

a. Read the data from the file into an object called "mydata". Column 1 ("Code") is the anonymized subject code and will not be used here. Columns 2-10 are the 9 features. Column 11 is the diagnosis: [B]enign or [M]alignant.
   i. How many total cases are there in the data?: 683
   ii. How many [B]enign cases are there in the data?: 444
   iii. How many [M]alignant cases are there in the data?: 239

b. Run k-means clustering using **all the rows** and **only the following features**: ClumpThickness, CellSize, and Nuclei. Use nstart=10.
   i. What should be the value of k? k = 2

[1]Original dataset from Breast Cancer Wisconsin (Diagnostic) Data Set
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

ii. Give R code:

```
result <- kmeans(mydata[, c(2, 3, 7)], centers = 2, nstart = 10)
```

c. Evaluation: Compare the resulting clusters with the known diagnosis.

i. What is the contingency table of your clustering? (Hint: use R's table() function. You can arbitrarily assign cluster 1/2 to Benign/Malignant)

	Cluster 1	Cluster 2
Benign	7 (FP)	437 (TN)
Malignant	219 (TP)	20 (FN)

ii. Give R code:

```
table(mydata$Class, result$cluster)
```

4. Using the contingency table that you obtained from the previous problem (3.c), calculate the following metrics (consider Malignant as the Positive class):

1. Accuracy: (TP + TN)/All = 219 + 437 / 683 = 0.96
2. Error: 1 – accuracy = 1 - 0.96 = 0.04
3. Precision: TP/(TP + FP) = 219 / 219 + 7 = 0.969
4. Recall: TP/(TP + FN) = 219 / 219 + 20 = 0.916
5. F-score: 2(P*R)/(P+R) = 2(0.969*0.916)/(0.969+0.916) = 0.942

Consider a "silly" classifier for this problem that makes every prediction as Malignant. Calculate the metrics for this "silly" classifier.

1. Accuracy: (TP + TN)/All = (219 + 0) / 683 = 0.32
2. Error: 1 - accuracy = 1 - 0.32 = 0.68
3. Precision: TP/(TP + FP) = 219 / (219 + 7) = 0.97
4. Recall: TP/(TP + FN) = 219 / (219 + 20) = 0.92
5. F-score: 2(P*R)/(P+R) = 2 * (0.97 * 0.92) / (0.97 + 0.92) = 0.946