**Problem 1 (10 points). Use the hw2\_p1.csv file for this problem(See R file).**

(1). Calculate the mean, standard deviation, minimum, and maximum of temp.

Mean of temp: 23.288

Standard deviation of temp: 7.600527

Minimum of temp: 4.6

Maximum of temp: 33.3

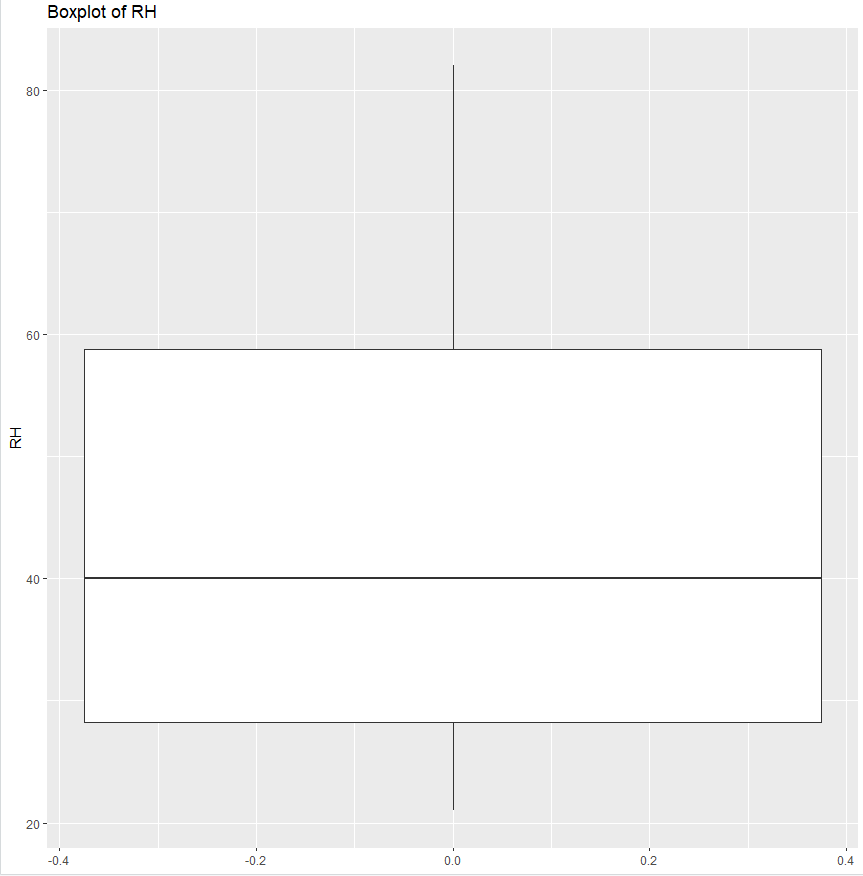
(2). Calculate the 25th, 50th, and 75th percentiles of temp using the percentiles formula that we discussed in the class.

25th percentile of temp: 18.35

50th percentile of temp: 25.2

75th percentile of temp: 29.05

(3). Plot the boxplot of RH.



(4). For this problem, consider only the wind attribute. Detect outliers using the IQR method that we discussed in the class and show all outlier values.

Outliers in wind: 9.4 9.4

**Problem 2 (10 points). Consider the following two objects with 7 binary attributes:**

(1). Calculate the distance between O1 and O2 assuming all attributes are symmetric attributes.

Matches: A2, A3, A6 = 3 matches

Total attributes = 7

SMC = 3/7

Distance = 1 – 3/7 = 4/7 = 0.5714

(2). Calculate the distance O1 and O2 assuming all attributes are asymmetric attributes with P being more important than N.

P – P Matches: A2 = 1 match

P – N Mismatches: A1, A4, A5, A7 = 4 mismatches

Jaccard = 1/ (1+4) = 0.2

Distance = 1 – 0.2 = 0.8

**Problem 3 (10 points) Consider the following dataset with two objects.**

Here, all attributes are ordinal attributes and ranks of their values are shown below (lowest rank on the left): A1: {1, 2, 3, 4, 5} A2: {first, second, third} A3: {bronze, silver, gold} A4: {small, medium, large} Calculate the distance between O1 and O2 using the method discussed in the class. Use the Euclidean distance measure.

From the dataset:

O1: 1, 2, 3, 2

02: 4, 3, 1, 3

Distance =

=

=

= 3.873

**Problem 4 (10 points). Consider the following dataset(See R file):**

(1). Calculate the cosine similarity between O1 and O2, cosine (O1, O2).

Cosine similarity between O1 and O2: 0.407040315179981

(2). Calculate the cosine similarity between O1 and O3, cosine (O1, O3).

Cosine similarity between O1 and O3: 0.757473586183857

(3). Is O1 closer to O2 or O3?

O1 is closer to O3 because the cosine O1 to O2 < cosine O1 to O3.

**Problem 5 (10 points). Consider the following dataset, which has attributes of mixed types.**

Calculate the distance between O7 and O8, d(O7, O8), and the distance between O7 and O9, d(O7, O9), using the method that we discussed in the class. Is O7 closer to O8 or closer to O9? You must do all calculations yourself.

A1:

O7 and O8: |52-36| = 16

O7 and O9: |52-12| = 4.

A2, A3:

mismatch = 1 for O7-O8

mismatch = 1 for O7-O9

A4, A5:

For O7 and O8: mismatch for A4 = 1, for A5 = 0 (Total = 1)

For O7 and O9: mismatch for A4 = 0, for A5 = 0 (Total = 0)

A6:

O7 and O8: 1

O7 and O9: 1

A7:

O7 and O8: 1/3

O7 and O9: 2/3

Distances:  
O7 and O8: 16 + 2 + 1 + 1 + 1/3 = 20.33

O7 and O9: 40 + 2 + 0 + 1 + 2/3 = 43.67

O7 is closer to O8 than to O9 because 20.33<43.67.