**Raul Garay CAP 4770 Homework 1**

**2.2**

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the mean of the data? What is the median?

Mean: (809/27) = 29.96

Median: 25

(b) What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).

Bimodal: 25 & 35

(c) What is the midrange of the data?

Midrange: (13+70)/2 = 41.5

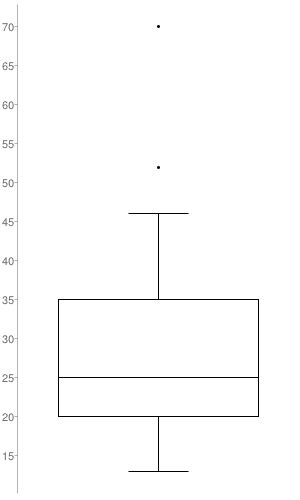
(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Q1: 20 Q3: 35

(e) Give the five-number summary of the data.

Min: 13 Q1: 20 Q2: 25 Q3: 35 Max: 70

(f) Show a boxplot of the data.



(g) How is a quantile–quantile plot different from a quantile plot?

A quantile plot only displays the distribution of one variable. However, a q-q plot displays that same distribution along with other chosen quantiles so that comparisons can be made.

**2.3**

Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

|  |  |
| --- | --- |
| **Age** | **Frequency** |
| 1-5 | 200 |
| 6-15 | 450 |
| 16-20 | 300 |
| 21-50 | 1500 |
| 51-80 | 700 |
| 81-110 | 44 |

Compute an approximate median value for the data.

= 11

N = 3194

= 950

= 1500

Width = 30

Median = 23.94

**2.6**

Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

(a) Compute the Euclidean distance between the two objects.

d = 6.708

(b) Compute the Manhattan distance between the two objects.

d = 11

(c) Compute the Minkowski distance between the two objects, using q = 3.

d = 6.15

(d) Compute the supremum distance between the two objects.

**2.8**

It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.

Suppose we have the following 2-D data set:

|  |  |  |
| --- | --- | --- |
|  | A1 | A2 |
| **X1** | 1.5 | 1.7 |
| **X2** | 2 | 1.9 |
| **X3** | 1.6 | 1.8 |
| **X4** | 1.2 | 1.5 |
| **X5** | 1.5 | 1.0 |

(a) Consider the data as 2-D data points. Given a new data point, x = (1.4, 1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

Euclidean Distance:

Manhattan Distance:

Supremum Distance:

Cosine Similarity:

3

(b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

|  |  |  |
| --- | --- | --- |
|  | A1 | A2 |
| **X1** | 0.375 | 0.778 |
| **X2** | 1 | 1 |
| **X3** | 0.5 | 0.889 |
| **X4** | 0 | 0.556 |
| **X5** | 0.375 | 0 |