**CP421 Assignment 1**

Nathaniel Carr

Note: In some cases, I have neglected to write out long summations, since it would be too wide for the page and make it hard to see what I'm actually doing. I've instead used a sigma summation notation. An example of this is found in 1.a, where I could have typed all 27 data points with a + between them, but it would exceed the page width and formatting it on separate lines would be confusing.

1.a)

Let A be the array of ages in sorted order.

The mean is:

The median is the 13th item (mid-point): 25.

1.b)

This data is bimodal, since it has 2 numbers that appear at the highest frequency: 25 and 35.

1.c)

The midrange of this data is:

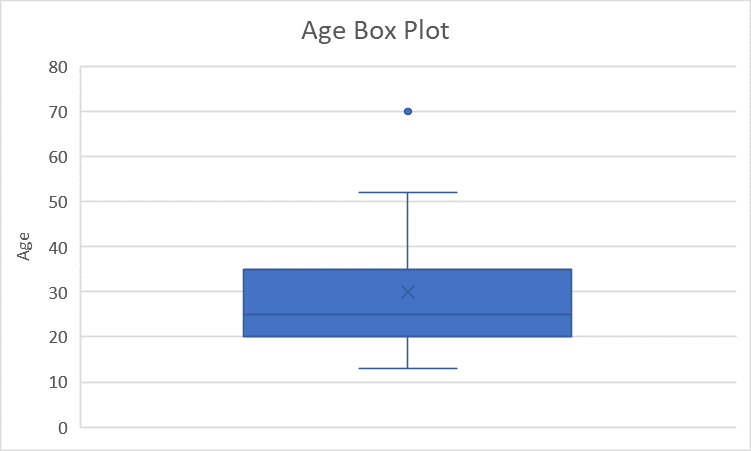
1.d)

Since the median is found at the 14th item, we look at the median of the 13 items to the left and 13 items to the right of the 14th item to find Q1 and Q3, respectively. Thus,

1.e)

1.f)

Note that 70 is an outlier because:



2.a)

Using Euclidean distance:

Since , the ranking of this data by similarity with the query using Euclidean distance (in ascending order of similarity): x1, x4, x3, x5, x2.

Using cosine similarity:

Ranking by similarity with the query using Euclidean distance: x1, x3, x4, x2, x5.

2.b)

For all xi:

Consequently, for all Aj of xi:

=

The normalized data set is:

|  |  |  |
| --- | --- | --- |
|  | A1 | A2 |
| x | 0.6585 | 0.7526 |
| x1 | 0.6616 | 0.7498 |
| x2 | 0.7250 | 0.6887 |
| x3 | 0.6644 | 0.7474 |
| x4 | 0.6247 | 0.7809 |
| x5 | 0.8321 | 0.5547 |

Repeating the Euclidean distance calculations:

3.a)  
The number of baskets some item, x, appears in can be calculated by .

The highest value of x for which is greater than or equal to the frequency threshold is 20, so every number below 21 is frequent and all others are infrequent.

The following set of items are frequent:

3.b)

5 and 7 are found together in the following baskets: 35, 70.

2, 5, and 7 are found together in the following baskets: 70.

The confidence of {5, 7} → 2 is 0.5.

2, 3, and 4 are found together in the following baskets: 12, 24, 36, 48, 60, 72, 84, 96.

2, 3, 4, and 5 are found together in the following baskets: 60.

The confidence of {2, 3, 4} → 5 is 0.125.