PROJECT II

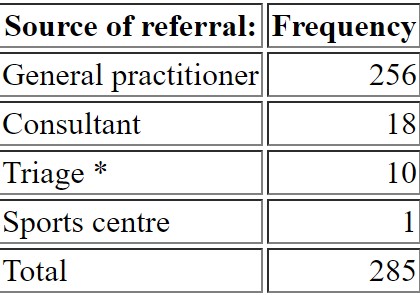
**Name: Bushra Naeemi**

**ID: 29064**

ITC 360: Data Mining

# PROJECT

1. Suppose that the values for a given set of source of referral of patients in a physiotherapy trial patients recruited to a randomized controlled trial of physiotherapy. The sources of referral and corresponding frequencies are as follows.



Compute the median of the above dataset,

***Answer of Question 1:***

Median from frequency table formula: ((n +1) **÷ 2) th value in the dataset**

* (285 + 1) **/ 2 = 286 / 2 = 143**

1. Suppose that the data for analysis includes the attribute height. The height values for the data tuples are (in random order): 77 75 77 77 74 73 74 73 72 71 73 61 62 63 61 62 62 63 64 65 64 66 75 63 64 67 75 66 68 68 69 70 70 72 71 73 72 75 76.
   1. What is the mean of the data? What is the median?
   2. What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data?
   3. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
   4. Give the five-number summary of the data.
   5. Show a bloxplot of the data

***Answer of Question 2:***

First Step: We should sort the data in assenting order:

61, 61, 62, 62, 62, 63, 63, 63, 64, 64, 64, 65, 66, 66, 67, 68, 68, 69, 70, 70, 71, 71, 72, 72, 72, 73, 73, 73, 73, 74, 74, 75, 75, 75, 75, 76, 77, 77, 77

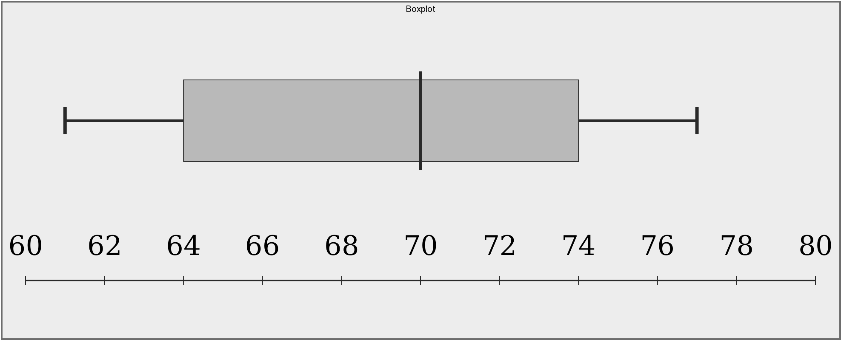
1. **Mean** = 69.307 and **Median** = 70
2. **Mode** = 73, 75.

It contains two modes, so it is Bimodal data

1. The midrange of the data is: M = (max + min) / 2;

M = (77+61) / 2 = 69

1. Q1 = 64 and Q3 = 74
2. Minimum = 61, Q1 = 64, Median = 70, Q3 = 74, Maximum = 77



1. Given two objects represented by the tuple (45, 5, 20, 10) and (40, 7,25, 8):
2. Compute the Euclidean distance between the two objects.
3. Compute the Minkowski distance between the two objects, using p = 2.
4. Do you observe any differences between the distances’ measures. Explain why we

use Euclidean distance. On top of that, explain in which scenario Minkowski is preferred.

**Answer of Question 3:**

P = (45, 5, 20, 10)

q = (40, 7,25, 8)

1. d(p, q) =
2. As we can see from the outputs of the equations, there is no difference between them. We use Euclidean distance in order to find the distance between two points in a plane.

In this question we are calculated the distance between object P and Q.

Euclidian distance is an extension of Minkowski distance where p = 2.

Minkowski formula is used for calculation of the distance between two points when it is needed in different ways. For instance, if p = 1, we have the Manhattan Distance which calculates the distance between two points in a grid or with p = 2, in this case we are calculating the distance in a plane.

1. Determine the cosine similarity relationships of the following documents

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Document*** | ***team*** | ***coach*** | ***hockey*** | ***baseball*** | ***soccer*** | ***penalty*** | ***score*** | ***win*** | ***loss*** | ***season*** |
| *Document1* | 5 | 0 | 3 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |
| *Document2* | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| *Document3* | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| *Document4* | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

**Answer for Question 4:**

d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

d3 = (0, 7, 0, 2, 1, 0, 0, 3, 0, 0) d4 = (0, 1, 0, 0, 1, 2, 2, 0, 3, 0)

Document 1:

d1 \* d2 = 5 X 3 + 0 X 0 + 3 X 2 + 0 X 0 + 2 X 1 + 0 X 1 + 0 X 0 + 2 X 1 + 0 X 0 + 0 X 1 = 25

d1 \* d3 = 0 + 0 + 0 + 0 + 2 + 0 + 0 + 6 + 0 + 0 = 8

d1 \* d4 = 0 + 0 + 0 + 0 + 2 + 0 + 0 + 0 + 0 + 0 = 2

* ||d1|| = 6.481

Document 2:

d2 \* d1 = 25

d2 \* d3 = 4

d2 \* d4 = 3

* ||d2|| = 4.12

Document 3:

d3 \* d1 = 8

d3 \* d2 = 4

d3 \* d4 = 8

* ||d3|| = 7.937

Document 4:

d4 \* d1 = 2

d4 \* d2 = 3

d4 \* d3 = 8

* ||d4|| = 4.358

Cosine Similarities:

Cos (d1, d2) = 25 / (6.481 X 4.12) = 0.935 => Document 1 and 2 are almost identical

Cos (d1, d3) = 0.155 => Document 1 and 3 are mostly dissimilar

Cos (d1, d4) = 0.07 => Document 1 and 4 are not similar at all

Cos (d2, d3) = 0.122 => Document 2 and 3 are mostly dissimilar

Cos (d2, d4) = 0.166 => Document 2 and 4 are mostly dissimilar

Cos (d3, d4) = 0.231 => Document 3 and 4 are fairly similar

1. Use the CHANNING HOUSE dataset and perform the following analysis (a) Show the first 10 rows of the dataset.

ITC 360 Data Mining Page 3 of 4

Project: Data Mining

* 1. Show how many rows and columns in the dataset
  2. Show if there is any missing values in the dataset
  3. Visualize the number of male vs female who died in the Channing house.
  4. Visualize the number of male vs female who left the Channing house before July 1, 1975.
  5. Compute the rate of female who died in the Channing house.
  6. Compute the rate of male who died in the Channing house.
  7. Normalize the dataset using minmax normalization method.
  8. Apply PCA on the dataset.
  9. Submit your jupyter notebook.

**Answer for Question 5:**

* Done in jupyter.

1. Given the following sets:
   1. = {Sam, Kyesha, Derek, Lorrie, Robin, Raúl, Shirley, Nathan, Chris, Dana}*,*
   2. = {Sam, Lorrie, Raúl, Derek}
   3. = {Robin, Derek, Kyesha}

* 1. Construct a contingency table.
  2. Find the Jaccard coefficient between A and B, between A & C, and between B & C (c) Explain how similar each to one another.

**Answer for Question 6:**

1. **Table:**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Sets* / Values | *Sam* | *Kyesha* | *Derek* | *Lorrie* | *Robin* | *Raúl* | *Shirley* | *Nathan* | *Chris* | *Dana* | *Total* |
| *A* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 10 |
| *B* | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 4 |
| *C* | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 |
| Total | 2 | 2 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 17 |

**b)**

Na - number of elements in set А

Nb - number of elements in set B

Nc - number of elements in intersecting set

T (A, B) = 4 / 10 + 4 – 4 = 4 / 10 = 0.4

T (A, C) = 0.3

T (B, C) = 0.166

**c)** The sets of (A, B) and (A, C) are closely similar to each other. But the set (B, C) is not that much similar because it has only one interesting set so it.

ITC 360 Data Mining Page 4 of 4