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*?

(b) 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?

(d) Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data?

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

(f) Show a *boxplot* of the data.

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

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.

2.4 Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of *age* and *%fat*.

(b) Draw the boxplots for *age* and *%fat*.

(c) Draw a *scatter plot* and a *q-q plot* based on these two variables.

2.5 Briefly outline how to compute the dissimilarity between objects described by the following:

(a) Nominal attributes

(b) Asymmetric binary attributes

(c) Numeric attributes
(d) Term-frequency vectors

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.
(b) Compute the *Manhattan distance* between the two objects.
(c) Compute the *Minkowski distance* between the two objects, using $q = 3$.
(d) Compute the *supremum distance* between the two objects.

2.7 The *median* is one of the most important holistic measures in data analysis. Propose several methods for median approximation. Analyze their respective complexity under different parameter settings and decide to what extent the real value can be approximated. Moreover, suggest a heuristic strategy to balance between accuracy and complexity and then apply it to all methods you have given.

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:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	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.
(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.

2.7 Bibliographic Notes

Methods for descriptive data summarization have been studied in the statistics literature long before the onset of computers. Good summaries of statistical descriptive data mining methods include Freedman, Pisani, and Purves [FPP07] and Devore [Dev95]. For

CPSC 8650 - Assignment 1

Parampreet Singh

2.2 Given -

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) Mean of the data :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$= \frac{(13 + 15 + 16 + \cdots 70)}{27}$$

$$= 29.962 \simeq 29.96$$

Median → 27 observations i.e odd

$$\text{median} = \left(\frac{27+1}{2}\right)^{th} \text{term}$$

$$\text{median} = 25$$

b) Mode → 25 - frequency 4
30 - frequency 4

∴ mode is both 25 & 30

c) Mid range of data:

$$= \frac{\text{smallest} + \text{largest}}{2}$$

$$= \frac{13 + 70}{2}$$

$$= 41.5$$

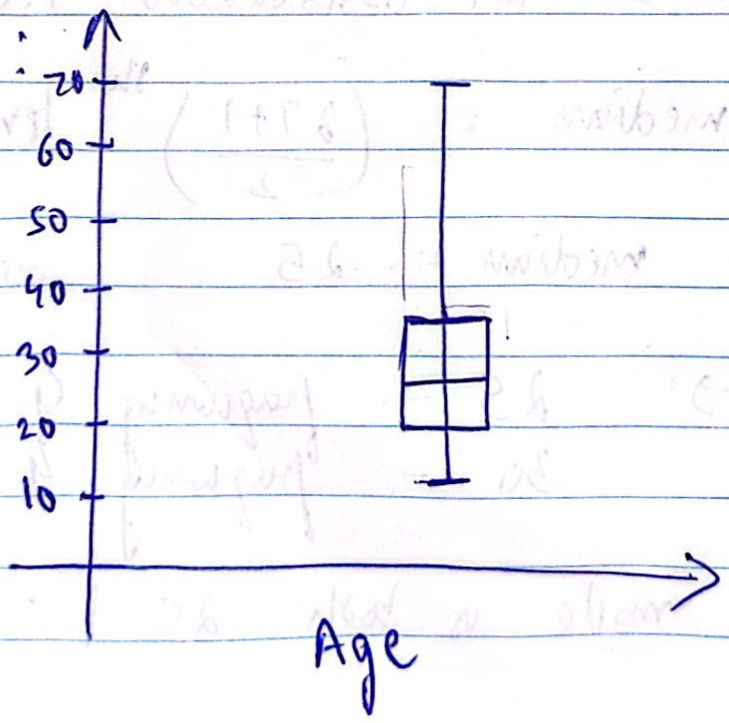
d) Q_1 (25th percentile) = 20

Q_3 (75th percentile) = 35

e) 5 number summary of data:

minimum	Q_1	median	Q_3	Maximum
13	20	25	35	70

f) Boxplot:



g) Quantile - Quantile is a graphical representation graph which compares data from one set to another set, it is a univariant distribution against the corresponding quantiles of another univariant distribution, both axis display the range of values measured for their corresponding distributions

Quartile plot is a graphical representation method which displays all the data in the increased order which plots quartile information, measured for the independent variable are plotted against their corresponding quartile plot

Ans 2.3 :-

Given data grouped into intervals

age	frequency	cummulative frequency
1-5	200	200
6-15	450	650
16-20	300	950
21-50	1500	2450
51-80	700	3150
81-110	44	3194

$$\text{Median} = l + \frac{\frac{N}{2} - (\Sigma \text{freq}) \text{ of } l}{\text{freq}_{\text{median}}} \times \text{width}$$

$N \rightarrow$ Sum of freq

$l \rightarrow$ lower boundary of median class $= 21$

$(\Sigma \text{freq.}) \text{ of } l = 950$

$\text{freq}_{\text{median}} = 1500$

width $= 30$

$$\text{Median} = 21 + \frac{\left(\frac{3194}{2} - 950\right) \times 30}{1500}$$

$$= 21 + 12.94$$

$$\boxed{\text{Median} = 33.94}$$

Ans 2.4

(a) Mean (sample mean) for age $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$$= \frac{1}{18}(23 + 23 + \cdots + 61)$$

$$= \frac{836}{18} = 46.44$$

Mean for % fat $= \frac{1}{18}(9.5 + 26.5 + \cdots + 35.7)$

$= \frac{518.1}{18}$

$= 28.783$

≈ 28.78

Median for age $= \frac{50+52}{2} = 51$

Median for % fat $= \frac{30.2+31.2}{2} = 30.7$

standard deviation (σ) for sample age

$\sigma = \sqrt{\text{Variance}}$

$= \sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2}$

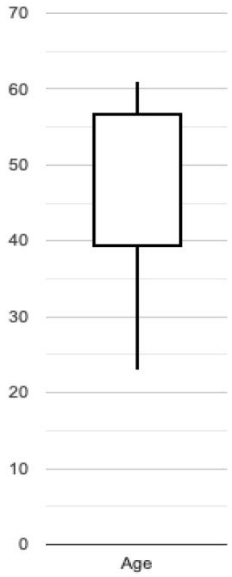
$= \sqrt{\frac{1}{17}[(23-46.44)^2 + \cdots + (61-46.44)^2}$

$= \sqrt{\frac{2970.4447}{17}} = 13.2186 \approx 13.2$

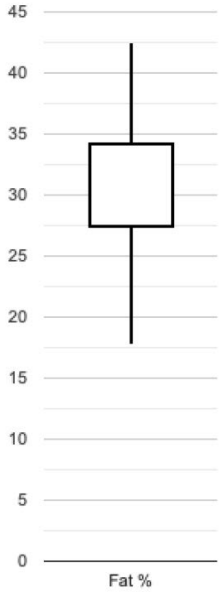
σ fat % $= \sqrt{\frac{1}{17}(9.5-28.78)^2 + \cdots \cdot (35.7-28.78)^2}$

$= \sqrt{\frac{1455.945}{17}} = 9.2544 \approx 9.2$

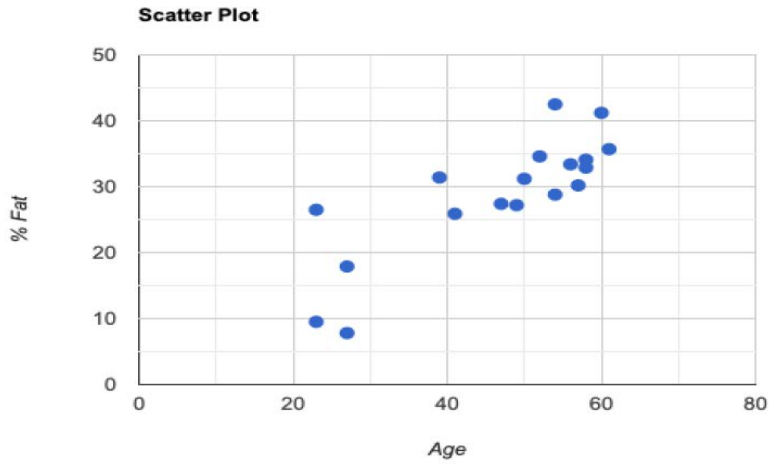
Box Plot for age



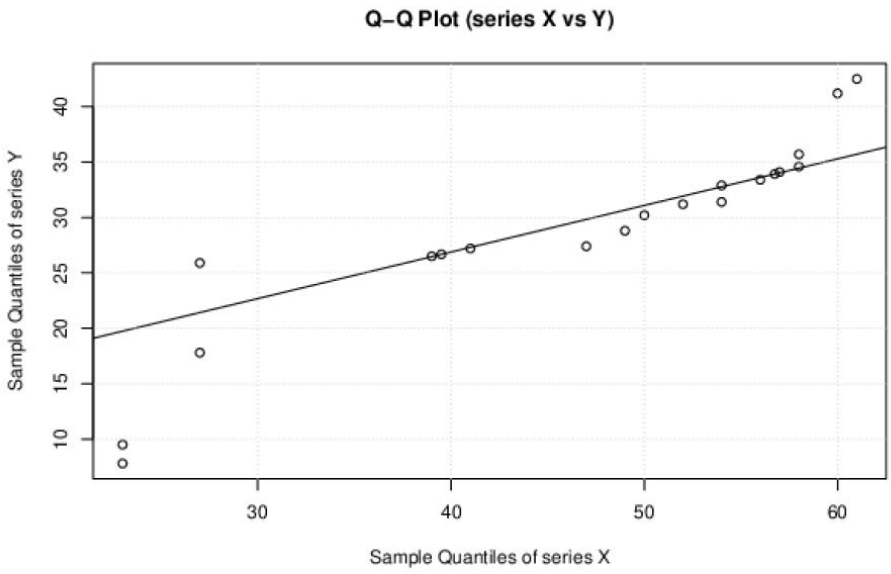
Box Plot for %fat



Scatter Plot



Q-Q Plot



2.5 (a) Nominal attribute:

We can compute dissimilarity b/w objects for the nominal attribute using simple matching method

$$d(i,j) = \frac{p-m}{p}$$ where m = no. of matches, p = total no. of variables

(b) Assymetric binary attribute:

Generally binary attribute values are 0 & 1 both has same meaning and a binary attribute has 2 types to measure the distance they are assymetric and symmetric

Contigency table for binary data:

		object 2 (i)		
		1	0	sum
object 2 (i)	1	q	r	q+r
	0	s	t	s+t

distance measure for asymetric binary variables are

$$d(i,j) = \frac{r+s}{q+r+s}$$

c) Numeric attribute -

we can find the dissimilarity b/w objects using the numeric attribute with minkowski distance with p dimensional data objects and further we can find using L-1 norm (Manhattan), L-2 norm (Euclidean), supremum.

Minkowski distance

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \cdots + |x_{ip}-x_{jp}|^h}$$

where $i = x_{i1}, x_{i2} \cdots x_{ip}$

$j = x_{j1}, x_{j2}, \cdots x_{jp}$

are 2-D data objects, h is the order (the distance so defined is also called L-h norm)

L-1 norm (Manhattan Distance)

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \cdots + |x_{ip} - x_{jp}|$$

L-2 norm (Eucledian Distance)

$$d(i,j) = \sqrt{|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

If when $h \to \infty$,
then it is called Supremum Distance
(also Lmax norm or L_∞ norm)
This is the maximum difference b/w
any component attribute of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{1/h}$$

$$= \max_{f}^{p} |x_{if} - x_{jf}|$$

(d) Term Frequency vectors:
It is also called cosine similarity
where it counts the frequency of the words
particularly used in documents.

It will use cosine measure to calculate the similarity b/w the documents. It is also known as term frequency vector

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \, |d_2|}$$

2.b Given 2 objects represented by the tuples (22, 1, 42, 10) & (20, 0, 36, 8)

(a) Euclidean distance

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

$$= \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}$$

$$= \sqrt{4+1+36+4}$$

$$= \sqrt{45}$$

$$= 6.708203$$

$$\approx 6.71$$

(b) Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

$= |22-20| + |1-0| + |42-36| + |10-8|$

$= 11$

c) Minkowski Distance for $q=3$

$$d(i,j) = \sqrt[3]{|x_{i1}-x_{j1}|^3 + |x_{i2}-x_{j2}|^3 + |x_{i3}-x_{j3}| + |x_{i4}-x_{j4}|}$$

$$= \sqrt[3]{(22-20)^3 + |1-0|^3 + |42-36|^3 + |10-8|^3}$$

$$= \sqrt[3]{233}$$

$= 6.1534 \approx 6.15$

d) Supremum distance

$$d(i,j) = \max_f |x_{if} - x_{jf}|$$

$= |42-36|$

$= 6$

Ans 2.7 So median $= L_1 + \left(\frac{N/2 - (\sum freq)l}{freq_{median}}\right) width$

where L_1 is the lower boundary of the median interval, N is the number of the frequency of the entire dataset and $(\sum freq)_l$ is the sum of the frequency

above the median interval, $freq_{median}$ is the frequency of the median interval and width is the width of the interval. The most straight forward way is to divide all data into k equal lengths intervals and the error incurred will be decreased as k becomes larger & larger, however time used will also increase.

So the error made and time used is a good optimality measure, and there are some other approaches to approximate median, one of them would be hierarchically divide the whole dataset into intervals, firstly divide into k regions, then find the region in which median resides then secondly divide this particular region into k sub regions find the sub region in which median resides and the same procedure goes on. we will repeatedly do this, until the width of the sub region reaches the pre defined sub threshold then median approximation formula as above mentioned is applied. This would be faster than globally partitioning all data into

shorter intervals which is expensive.

Ans 2.8

(a) New data point $x = (1.4, 1.6)$

Euclidean distance

$d(1, x) = \sqrt{(1.5-1.4)^2 + (1.7-1.6)^2} = 0.1414$

$d(2, x) = \sqrt{(2-1.4)^2 + (1.9-1.6)^2} = 0.6708$

$d(3, x) = \sqrt{(1.6-1.4)^2 + (1.8-1.6)^2} = 0.2828$

$d(4, x) = \sqrt{(1.2-1.4)^2 + (1.5-1.6)^2} = 0.2236$

$d(5, x) = \sqrt{(1.5-1.4)^2 + (1.0-1.6)^2} = 0.6083$

Manhattan Distance

$d(1, x) = |1.5-1.4| + |1.7-1.6| = 0.2$

$d(2, x) = |2-1.4| + |1.9-1.6| = 0.9$

$d(3, x) = |1.6-1.4| + |1.8-1.6| = 0.4$

$d(4, x) = |1.2-1.4| + |1.5-1.6| = 0.3$

$d(5, x) = |1.5-1.4| + |1.0-1.6| = 0.7$

Supremum dist

$d(1, x) = 0.1$

$d(2, x) = 0.6$

$d(3, x) = 0.2$

$d(4, x) = 0.2$

$$d(s, n) = 0.6$$

Using Similarity

$$\cos(d_1, n) = \frac{(1.5 \times 1.4) + (1.7 \times 1.6)}{\sqrt{(1.5)^2 + (1.7)^2} \times \sqrt{(1.4)^2 + (1.6)^2}} = 0.9999$$

$$\cos(d_2, n) = \frac{(2 \times 1.4) + (1.9 \times 1.6)}{\sqrt{(2)^2 + (1.9)^2} + \sqrt{(1.4)^2 + (1.6)^2}} = 0.99875$$

$$\cos(d_3, n) = \frac{(1.6 \times 1.4) + (1.8 \times 1.6)}{\sqrt{(1.6)^2 + (1.8)^2} + \sqrt{(1.4)^2 + (1.6)^2}} = 0.99997$$

$$\cos(d_4, n) = \frac{(1.2 \times 1.4) + (1 \times 1.6)}{\sqrt{(1.2)^2 + 1^2} + \sqrt{(1.4)^2 + (1.6)^2}} = 0.96536$$

when comparing values of the above datapoints, the ranking is below

Euclidean Distance : n_1, n_4, n_3, n_5, n_2
Manhattan Distance : n_1, n_4, n_3, n_5, n_2
Supremum Distance : n_1, n_4, n_3, n_5, n_2
Cosine Similarity : n_1, n_3, n_4, n_2, n_5

b) $x = (1.4, 1.6)$

Normalized value of x is $\frac{x}{\text{norm}(x)}$

Here normalized $(x) = \sqrt{1.4^2 + 1.6^2}$

Normalized x is $\left(\frac{1.4}{2.13}, \frac{1.6}{2.13}\right) = (0.6573, 0.7512)$

x_1 value $(1.5, 1.7)$ normalized $x_1 = \sqrt{1.5^2 + 1.7^2}$

normalized $x_1 = \left(\frac{1.5}{2.27}, \frac{1.7}{2.27}\right) = 0.6608$ $x_1 = 2.27$

x_2 $(2, 1.9)$ normalized $(x_2) = \sqrt{2^2 + 1.9^2} = 2.76$

normalized $x_2 = \left(\frac{2}{2.76}, \frac{1.9}{2.76}\right)$

$= (0.7246, 0.6884)$

Similarly for

	A_1	A_2
x_3	0.6639	0.7464
x_4	0.6250	0.7813
x_5	0.8333	0.5555

So euclidean distance normalized x_1 to normalized datapoint

$x_1 = 0.004$ $\quad$ $x_4 = 0.04$

$x_2 = 0.09$ $\quad$ $x_5 = 0.96$

$x_3 = 0.007$

So, based on these values

Ranks

↳ x_1, x_3, x_4, x_2, x_5