Course: **Data Mining** 08-Jan-2023 **(Fall 2022)**

Resource Person: **Dr. Muhammad Faheem**

ASSIGNMENT–1 (Introduction and Getting to know Data)

Total Points: 30
Submission Due: Saturday Jan 14, 2023
(Google Classroom Course Page)

Instructions: Please Read Carefully!

- **This is an individual assignment. Everyone is expected to complete the given assignment on their own, without seeking any help from any website or any other individual.** There will be strict penalties for any work found copied from any source and the university policy on plagiarism will be strictly enforced.
- Assignment is to be submitted via Google Classroom.
    - You should already have created your account on Google Classroom as per my earlier email. If not, then follow the link in that email to create your account.
    - Submit your assignment on or before due date. **No late submissions will be possible**.
    - The viva of this assignment will be conducted.

Question 1: [5]

How is a data warehouse different from a relational database? Explain your justification with a set of examples.

Differences between a data warehouse and a database: A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing. For more details, please refer to the section "Differences between operational database systems and data warehouses."

Similarities between a data warehouse and a database: Both are repositories of information, storing huge amounts of persistent data.

Question 2: [5]

Compare and discuss the role of different statistical description techniques for data preprocessing step. Discuss each measure of central tendency and dispersion of data and elaborate which one should be preferred in a specific situation.

The basic statistical descriptions include measures of central tendency (Section 2.2.1), which measure the location of the middle or center of a data distribution. Intuitively speaking, given an attribute, where do most of its values fall? In particular, we discuss the mean, median, mode, and midrange.

In addition to assessing the central tendency of our data set, we also would like to have an idea of the dispersion of the data. That is, how are the data spread out? The most common data dispersion measures are the range, quartiles, and interquartile range; the five-number summary and boxplots; and the variance and standard deviation of the data These measures are useful for identifying outliers and are described in Section 2.2.2.

Finally, we can use many graphic displays of basic statistical descriptions to visually inspect our data (Section 2.2.3). Most statistical or graphical data presentation software packages include bar charts, pie charts, and line graphs. Other popular displays of data summaries and distributions include quantile plots, quantile–quantile plots, histograms, and scatter plots.

Refer to chapter 2 for details discussion.

Question 3: [10]

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

Work Experience	**Frequency**
1-2	100
3-4	350
5-6	600
7-8	400

9-10	300
11-12	400
13-14	15

Compute an approximate median value for the data.

Presented during the class and refer to chapter 2.

Question 4: **[10]**

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	27	27	39	41	47	49	50
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a) Draw the boxplots for age and %fat.
b) Draw a scatter plot and q-q plot based on these two variables.

Calculate the mean, median and standard deviation of age and %fat.

For the variable age the mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable %fat the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

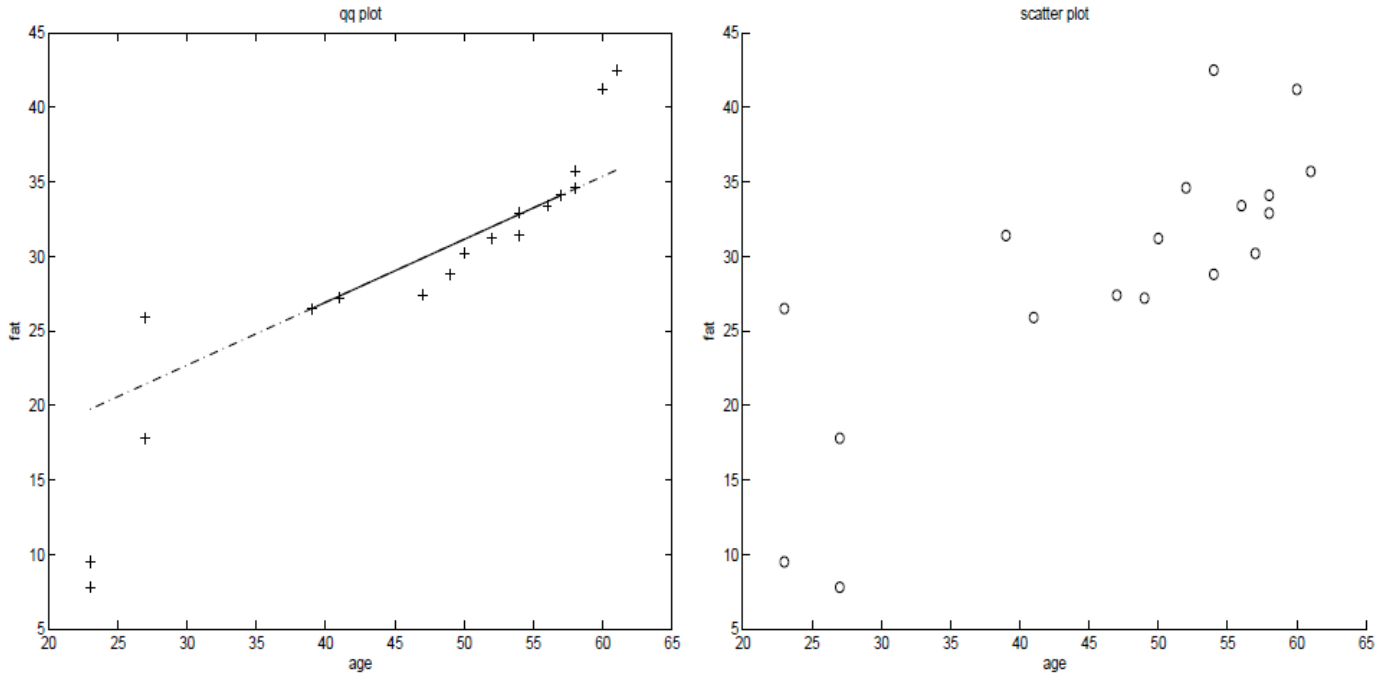

Figure 2.3: A *q-q plot* and a *scatter plot* of the variables *age* and *%fat* in Exercise 2.4.

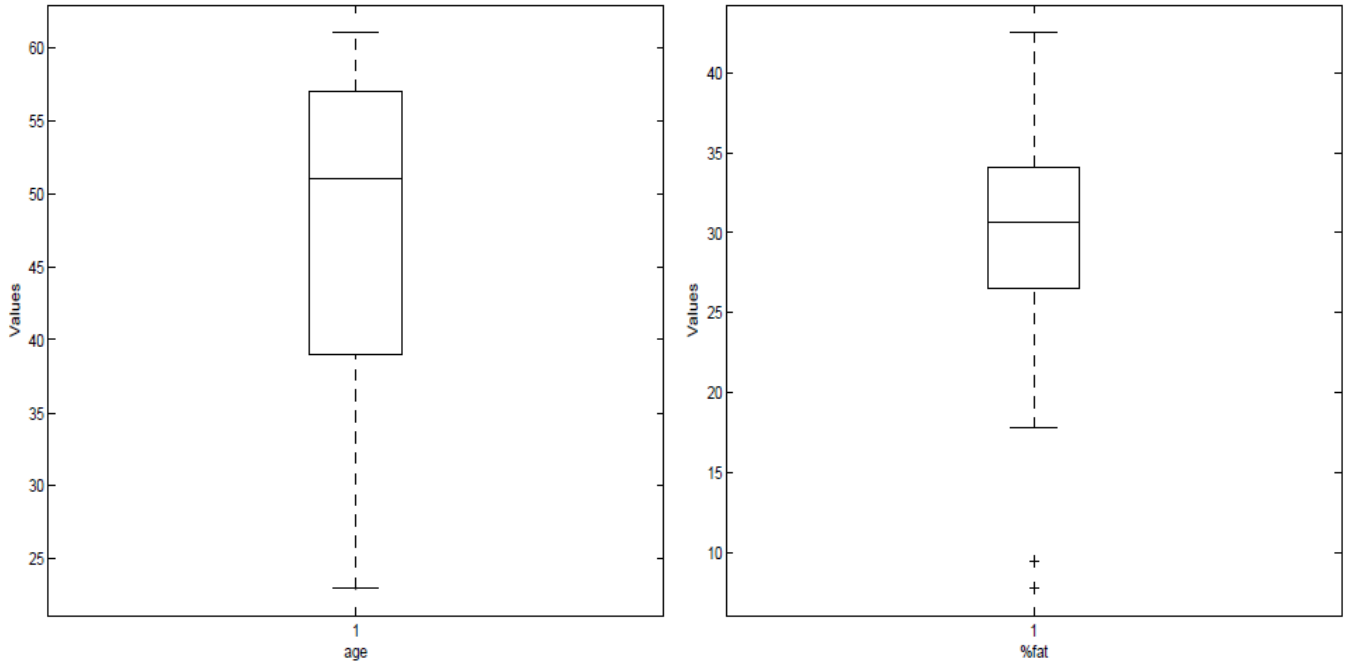

Figure 2.2: A boxplot of the variables *age* and *%fat* in Exercise 2.4.