Homework Assignment 1

Name: **Asrar Syed**
Due by: **September 13, 2024**
Course Section: **CSC 4740-002**

Q1: Define an overall pipeline of data mining project step-by-step, using the reference from lecture slides. (10pt)
The KDD process is followed to extract an unknown, significant patterns or knowledge from a large data set.

Overall the first step would be to "understand" where your data is derived from and to analyze it to help build a initial hypothesis or expectations about patterns found in the data set. This could consist of acknowledging the source of the data and any relevant features such as missing data, the way the data is distributed or any outliers within the data set.

The next step of the process is "preprocessing" the data which can largely subdivided into several stages such as data cleaning, integration, reduction, and transformation. Data cleaning is the stage where we fill in the missing gaps in the data and remove any random errors, notably known as noisy data. Data integration is the stage where data from multiple sources are combined for analysis, and this ensures accuracy and speed efficiency of the data mining process. Data reduction is a stage at which we employ a technique to reduce collecting unnecessary data and to focus only on relevant data, done through dimensionality reduction for example. Data transformation is the state in which we overhaul the data into a suitable structured form to pick up on any patterns.

"Mining the Data" is the sequential process involving applying algorithms, techniques or tools to identify the patterns found in the data. This involves choosing the right data mining technique for the task then applying it to build the model. After the model is built, we need to test it against a test set to see how it preforms and then evaluate the models predictive ability compared to the actual outcomes in the test set.

"Post-processing" is another process that is subdividing into several stages such as verifying the results of the model, creating a visualization of what the data is doing, and making interpretations from the data.

The last step of the entire process is "Information" which focuses on discussing, concluding, and acting on the results of the data mining project. Conclusions are drawn based on the model's outputs, and actionable recommendations are provided to guide future actions or decisions.
Q2: Define the following data mining functionalities: Association and correlation analysis, classification, Regression, Clustering, and Outlier analysis. Give examples of what each data mining functionality can do, using a real-life example that you are familiar with. (10pt)

Association and Correlation Analysis:

Definition: This looks for patterns or relationships between different variables in your data. Association finds items that often occur together, while correlation measures how strongly two variables are related.

Example: Association analysis could reveal that “bread” and “butter” are frequently purchased together when someone goes to the grocery store. The store could then put those items in a visual distance of each other.

Classification:

Definition: Classification is about sorting data into categories based on certain features. You use a labeled dataset to train a model to recognize patterns and make predictions.

Example: Our college outlook email uses classification to identify incoming emails and then classifies them as “spam” or “not spam” depending on the sender, content or the subject line.

Regression:

Definition: Regression predicts a continuous outcome based on one or more input variables. It helps you understand how changes in the input variables affect the outcome.

Example: If you’re studying how the price of used cars changes with mileage and age, regression can help you estimate the price of a car based on these factors.

Clustering:

Definition: Clustering groups similar data points together without predefined labels. It helps find natural groupings in your data.

Example: A company might use clustering to group customers based on their purchasing behavior. This segmenting can be used as means to send targeted marketing advertisements.

Outlier Analysis:

Definition: Outlier analysis finds data points that are very different from the rest of the data. These anomalies could indicate something unusual or interesting.

Example: If a student suddenly has a much higher than average score on a test compared to their previous scores, outlier analysis can help identify if this is an error or something to investigate further.

Q3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order). (40pt)

5	11	11	12	13	15	16	16	19	20	20	21	22	22	25	25	25	25	25	26	26	26
27	28	29	31	33	34	34	34	35	35	35	35	35	35	36	36	36	37	40	45	46	52

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

Mean: 27.591

Median: 26.5

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

Mode: 35

Modality: Unimodal

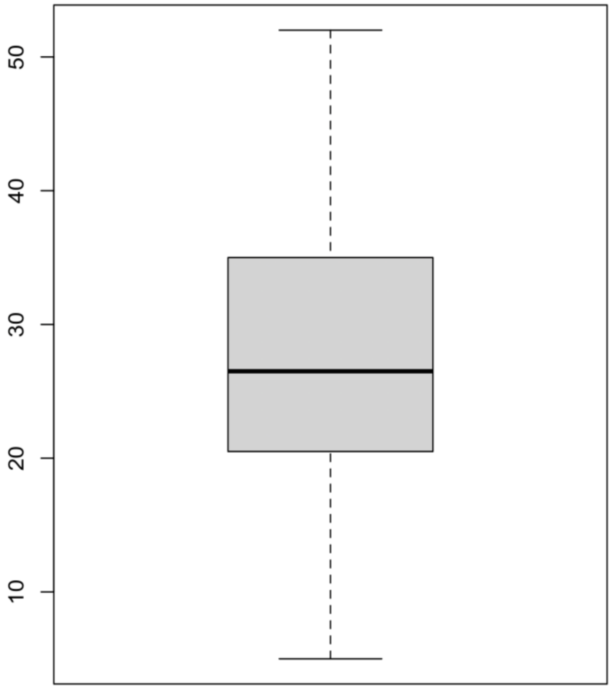
(c) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

First Quartile (Q1): 20.5
Third Quartile (Q3): 35

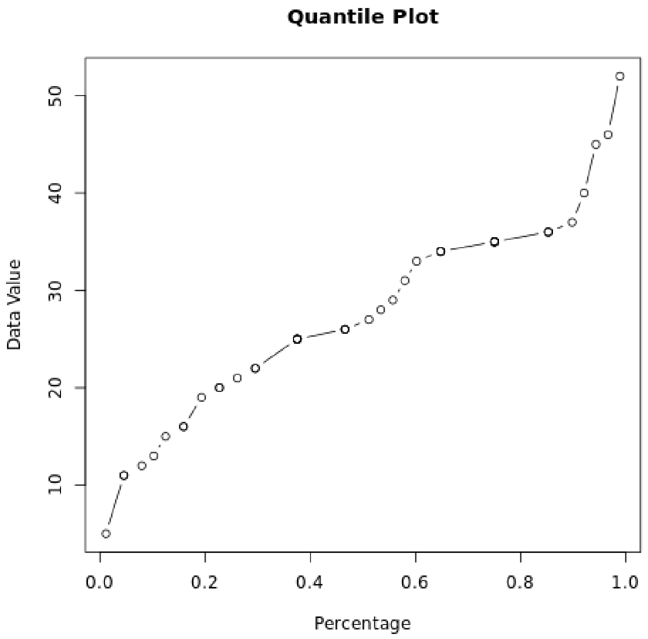
(d) Give the five-number summary of the data.

Five-Number Summary: 5.0 20.5 26.5 35.0 52.0

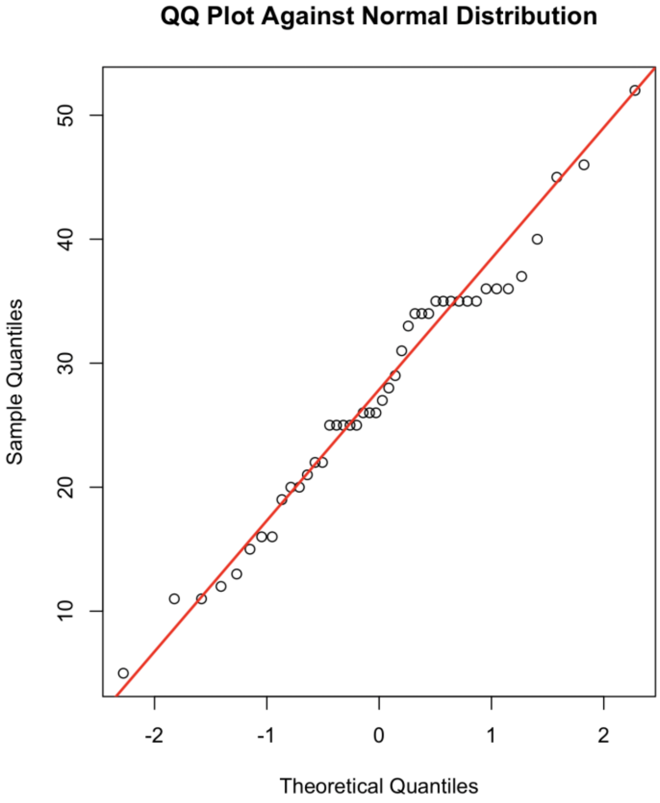
(e) Show a boxplot of the data.



(f) Show a quantile plot of the data



(g) Show a quantile-quantile plot against a normal distribution with sample mean and sample standard deviation?



Note: For (e), (f), (g) any programing language can be used to generate plot, even your hand drawing.
Q4: Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (10pt)

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

Euclidean Distance: 6.708204 or 6.71

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

Manhattan Distance: 11

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

Minkowski Distance: 6.153449 or 6.15

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

Supremum Distance: 6

Q5: 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 two-dimensional data set: (10 pt)

	A1	A2
X1	0.66162	0.74984
X2	0.72500	0.68875
X3	0.66436	0.74741
X4	0.62470	0.78087
X5	0.83205	0.55470

Consider the data as two-dimensional 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. (similar data points rank first)

	A1	A2	Euclidean Distance
X1	0.66162	0.74984	1.126045
X2	0.72500	0.68875	1.13402
X3	0.66436	0.74741	1.126089
X4	0.62470	0.78087	1.127858
X5	0.83205	0.55470	1.18963

	A1	A2	Manhattan Distance
X1	0.66162	0.74984	1.58854
X2	0.72500	0.68875	1.58625
X3	0.66436	0.74741	1.58823
X4	0.62470	0.78087	1.59443
X5	0.83205	0.55470	1.61325

	A1	A2	Supremum Distance
X1	0.66162	0.74984	0.85016
X2	0.72500	0.68875	0.91125
X3	0.66436	0.74741	0.85259
X4	0.62470	0.78087	0.81913
X5	0.83205	0.55470	1.0453

	A1	A2	Cosine Distance
X1	0.66162	0.74984	0.9999914
X2	0.72500	0.68875	0.9957523
X3	0.66436	0.74741	0.9999695
X4	0.62470	0.78087	0.9990284
X5	0.83205	0.55470	0.9653634

Q6: Required for graduate students, Bonus points for undergraduate students (10 pt) Compute Maximum Likelihood Estimation for the following problem;

If you get independent samples $x1$, $x2$, x3, x4 , from an $Exponential(\theta)$ distribution

Pdf (x, θ) = θexp(-θx) = $\theta e^{-\theta x}$; X> 0

What is the likelihood function?

L(x1,x2,x3,x4, θ) =

What is formular of MLE of θ;

Θ =

Here is data: (x1,x2,x3,x4) = (1.3,3.5,1.9, 2.2). Plug in the data values, what is the value of θ;

Q7: Please describe your idea(s) for data mining project. The idea(s) does not need to be final. You can still change, update, and revise it later. The following questions are meant to motivate your thinking, not for following exactly. What question(s) you want to answer? what type of data you are interested in? or which research field(s) you are interested in? or what do you expect your project can do using a certain type of data ? (10 pt)

One idea I am leaning towards at the current moment is – a prediction system on how bright headlights cause car crashes at night.

This will seek to provide analytical insight to a largely ignored topic, and will try to provide an answer to if brighter LED headlights cause fatal crashes and to what degree they contribute to. My project should provide a model that can predict fatality rates depending on the brightness of cars headlights.

I am also interested in researching some topics in healthcare, economics and entertainment.