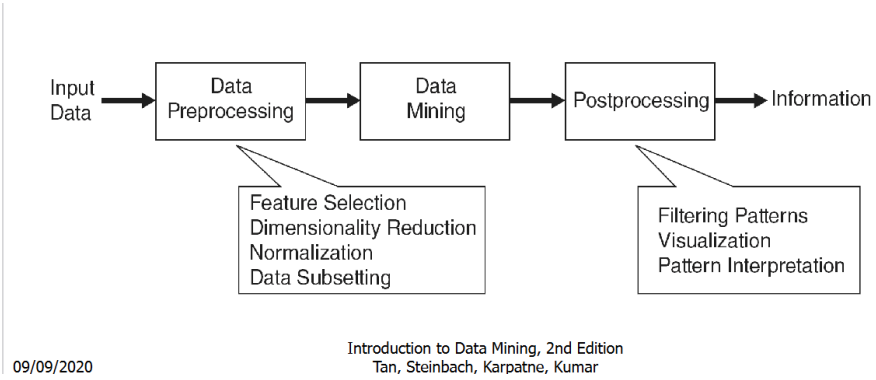
CSC 6740 Data Mining Assignment 1

Name: **Salil Dinesh Apte**
PantherID: **002821755**
MS CS Graduate student

1. Overall pipeline of a Data Mining project
    a. Data mining is non-trivial extraction of implicit, previously unknown and potentially useful information from data related to a business problem
    b. There are various stages involved -
        i. **Data Preprocessing**: This takes in the raw input data and then feature selection, dimensionality reduction, normalization and data subsetting takes place where the data gets cleaned and relevant fields are extracted for further steps. The actual data for the next step gets selected
        ii. **Data Mining**: In this phase, primarily machine learning algorithms get applied to the data to build predictive or descriptive models. This involves selecting the appropriate algorithm, training the model and tuning hyperparameters
        iii. **Data Post Processing**: Here filtering patterns and visualization takes place to interpret whether there patterns present within the data and their correlation. This is effective in evaluating efficiency in solving the business problem. It may involve domain experts to ensure the insights are actionable
        iv. **Deployment and Information**: Implementing the model in a production environment where it can be used to make real-time predictions or decisions, where finally meaningful information has been extracted from the data



2. **Data Mining functionalities**
    a. **Association and correlation analysis**
        i. Identifies relationships between variables in large datasets, often in the form of "if-then" rules

ii. Example: In a supermarket, if a customer buys milk, they are likely to also buy eggs. This is often used in market basket analysis to find product pairings and aisle structures

b. **Classification**
   i. Assigns items to predefined categories based on input data
   ii. Example: Spam detection in email systems, where emails are classified as "spam" or "not spam" based on their content.

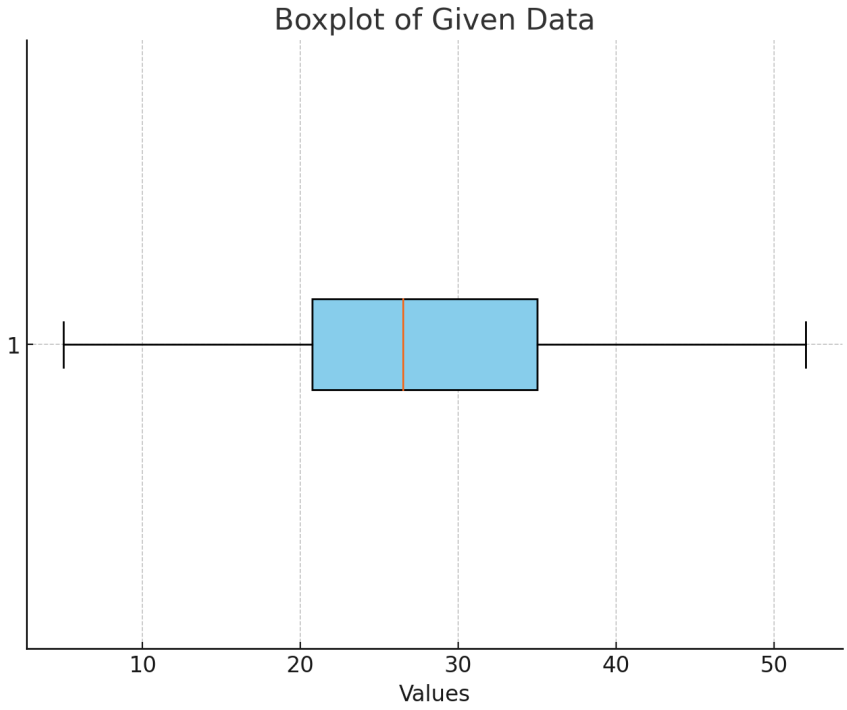
c. **Regression**
   i. Predicts a continuous value based on input variables
   ii. Example: Predicting house prices based on factors like size, location, and number of bedrooms, number of tenants, floor

d. **Clustering**
   i. Groups similar data points into clusters without predefined labels
   ii. Customer segmentation in marketing, where customers are grouped into clusters based on purchasing behavior and online in-app activity parameters

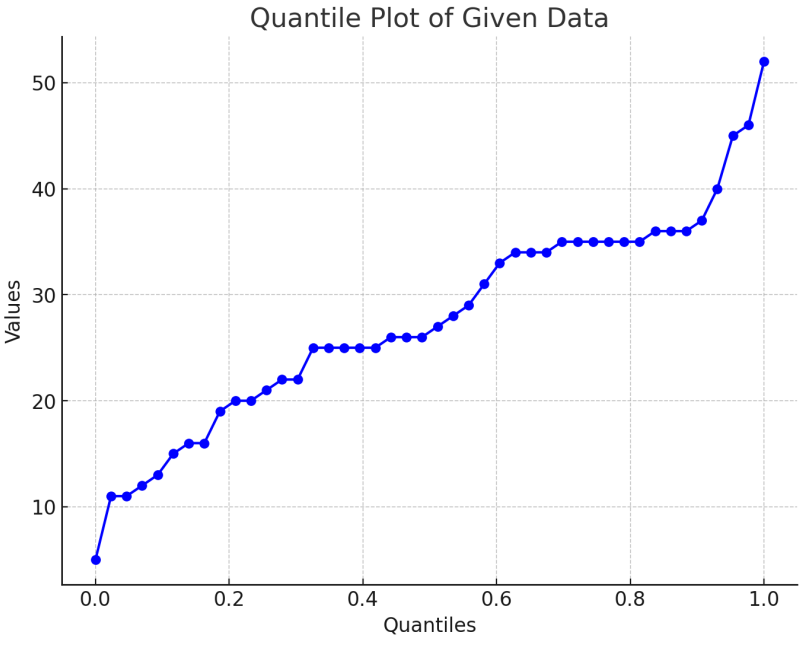
e. **Outlier analysis**
   i. Identifies data points that deviate significantly from other observations
   ii. Example: Detecting fraudulent credit card transactions by identifying transactions that are unusual compared to typical spending patterns or a peculiar amount

3. Analysis of data
   a. Mean : **27.59**, Median : (26 + 27) / 2 = **26.5** since it is an even set of data
   b. Mode: **35** (element with highest frequency being 6) and data is **multimodal**
      i. **bimodal** would be the simplest classification based on the highest frequency values 35 and 26
   c. Quartiles
      i. 1st quartile :
         1. Q1 is the value at the **25th percentile**.
         2. Since there are 44 data points, the position is (44+1)/4 = 11.25. So Q1 is the 11th value
         3. **Q1 = 20**.
      ii. 3rd quartile : Average of values at positions 33 and 34 = (35 + 35)/2 = **35**
   d. **Five number summary**
      i. **Minimum**: The smallest value is **5**
      ii. **Q1** (1st quartile): Value at 25th percentile as described above is **20**
      iii. **Q2** (Median): Value at 50th percentile, which is 22nd and 23rd values in the sorted dataset (average of 26 and 26)
         1. Median = **26**
      iv. **Q3** : Average of values at positions 33 and 34 = (35 + 35)/2 = **35**
      v. **Maximum**: The largest value is **52**

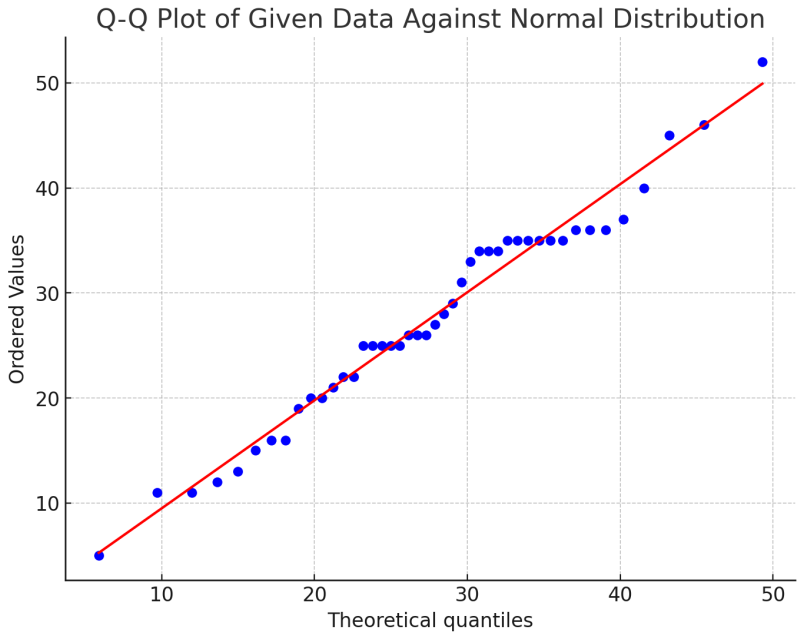
vi. **Boxplot**



vii. **Quantile** plot



viii. **Q-Q Plot** against **Normal Distribution**



4. Two object tuples
    a. **Euclidean distance**
        i. D = sqrt((22−20)^ + (1−0)^2 + (42−36)^2 + (10−8)^2)
        ii. D = sqrt(4 + 1 + 36 + 4)
        iii. D = sqrt(45)
        iv. Euclidean Distance = **6.71**
    b. **Manhattan distance**
        i. |22−20|+|1−0|+|42−36|+|10−8|
        ii. MD = 2 + 1 + 6 + 2
        iii. Manhattan Distance = **11**
    c. **Minkowski distance**
        i. Formula: D = (|x1−x2|^h+|y1−y2|^h+|z1−z2|^h+|w1−w2|^h)^(1/h)
        ii. Here h = 3
        iii. D = (|22−20|^3 + |1−0|^3 + |42−36|^3 + |10−8|^3)^(⅓)
        iv. Minkowski distance = 233 ^ (⅓)
        v. Minkowski distance = **6.13**
    d. **Supermum distance**
        i. D = max(|22−20|,|1−0|,|42−36|,|10−8|)
        ii. D = max(2,1,6,2)
        iii. Supermum distance = **6**
5. Solution
    a. **Data points** from the table
        i. x1 = (0.66162, 0.74984)
        ii. x2 = (0.72500, 0.68875)
        iii. x3 = (0.66436, 0.74741)
        iv. x4 = (0.62470,0.78087)

v. x5 = (0.83205,0.55470)

b. **Query point**
   i. xQ = (1.4, 1.6)
c. **Euclidean distance**
   i. d(x1,x2) = sqrt((x1 − x2)^2 + (y1 − y2)^2)
   ii. Distances: [1.1260, 1.1340, 1.1261, 1.1279, 1.1896]
   iii. Ranking (most similar first): **x1, x3, x4, x2, x5**
d. **Manhattan distance**
   i. d(x1,x2) = |x1 − x2| + |y1 − y2|
   ii. Distances: [1.5885, 1.5863, 1.5882, 1.5944, 1.6133]
   iii. Ranking (most similar first): **x2, x3, x1, x4, x5**
e. **Supremum distance**
   i. d(x1,x2) = max(|x1 − x2|, |y1 − y2|)
   ii. Distances: [0.8502, 0.9113, 0.8526, 0.8191, 1.0453]
   iii. Ranking (most similar first): **x4, x1, x3, x2, x5**
f. **Cosine similarity**
   i. Formula: CS = (x1.x2 + y1.y2) / (sqrt(x1^2 + y1^2) + sqrt(x2^2 + y2^2))
   ii. Similarities: [0.99999, 0.99575, 0.99997, 0.99903, 0.96536]
   iii. Ranking (most similar first): **x1, x3, x4, x2, x5**
g. **Summary of rankings**
   i. x1 is generally the closest in most metrics
   ii. x3 also ranks highly across distances
   iii. x5 is the least similar to the query point

6. **Maximum Likelihood Estimation**
   a. f(x;θ)=θe^(−θx) for x > 0
   b. The likelihood function L(x1,x2,x3,x4;θ) is the joint probability of observing the data x1, x2, x3, x4, given the parameter θ
   c. For independent data points, the likelihood function is the product of the individual PDFs for each xi:
      i. L(x1, x2, x3, x4; θ) = ∏ f(xi;θ) where i ranges from 1 to 4
      ii. L(x1, x2, x3, x4; θ) = ∏ θ e^(−θx) where i ranges from 1 to 4
      iii. L(x1, x2, x3, x4; θ) = θ^(4) e^(−θ(x1 + x2 + x3 + x4))
      iv. Thus, the likelihood function for the four independent samples is:
      v. **L(x1, x2, x3, x4; θ) = θ^(4)e^(−θΣxi) where i ranges from 1 to 4**
   d. MLE of θ
      i. Likelihood Function**:** The likelihood function for n independent observations is:
         1. L(x1, x2, x3, x4; θ) = θ^(4)e^(−θΣxi) where i ranges from 1 … n
      ii. To make the math easier, we take the natural logarithm (log-likelihood):
         1. logL(θ) = nlogθ - θΣxi where i ranges from 1 … n
      iii. Differentiate the log-likelihood function with respect to θ and set it to zero
         1. (d/dθ) logL(θ) = n/θ - Σxi = 0 where i ranges from 1 … n
      iv. Solve the above equation to get the MLE for θ
         1. Θ = n / (Σxi) where i ranges from 1 … n

2. Therefore **MLE for θ = n / (sum of the observations)**

v. For given data
   1. n = 4 because there are 4 data points
   2. Sum of observations = 1.3 + 3.5 + 1.9 + 2.2 = 8.9
   3. Θ = n / (Σxi) = 4/8.9 = 0.4494
   4. The MLE for **θ = 0.4494**

7. Data Mining Project (Human / AI comment classifier)

a. For my Data Mining project, I am exploring the idea of detecting and classifying AI-generated comments on social media, specifically identifying whether a given comment is written by a human or generated by models like ChatGPT and Gemini.

b. **Questions to answer**
   i. Can we develop a machine learning model to accurately classify whether a social media comment was generated by an AI?
   ii. What linguistic features differentiate
   iii. AI-generated comments from human-written ones?
   iv. How do AI-generated comments impact online communication, and can these differences be measured?

c. **Data Of Interest**
   i. I am interested in working with a dataset of social media comments, which includes both human-written and AI-generated text
   ii. The dataset will contain features like the length of the text, sentence structure, and specific word usage, which are essential for identifying patterns in AI-generated content

d. **Research Field**
   i. This project lies at the intersection of natural language processing (NLP) and social media analytics.
   ii. It can contribute to fields such as AI ethics, content moderation, and digital communication studies.

e. **Expected outcomes**
   i. By data mining and analyzing social media data, I expect to build a model that can:
      1. **Accurately classify** AI-generated text and comments, providing insights into the increasing role of AI in communication.
      2. Develop methods to **detect and differentiate** AI-generated content, which can be useful for content moderation and fact-checking.
      3. Help in developing **corrective measures** if a high amount of AI generated content is detected by a user based on frequency and keeping the community informed.