QUESTION 1

A. How is data mining different from statistical analysis?

Ans:

Data Mining: It is extracting unknown, implicit, and useful patterns from the data. It is used to build the models to do these tasks from the large datasets. It makes use of statistical techniques to do so.

Statistical Analysis: Although, both data mining and statistical analysis are used to extract the knowledge from the data, statistics is only about quantifying data. Data mining makes use of statistical analysis tools and techniques, which mainly involve math.

B. Differentiate between noise and outliers

Ans:

Noise: It is a modification of original values in the dataset. Noise is not considered as a part of dataset as it is an extraneous object.

Outlier: It is a data point which doesn't fit with the other data which is available. Unlike noise, Outliers are a part of dataset only.

C. List 5 unique data characteristics that makes traditional data mining techniques less effective.

Ans:

1. Large-scale data
2. High dimensional data
3. Heterogeneous data
4. Complex data
5. Distributed data

QUESTION 2

It is important to define or select similarity measures in data analysis. However, there is no commonly accepted similarity measure. Different similarity measures may come up with different results. And the similarities between data points may or may not change after data transformation. In this question, we want to explore these problems.

a) Consider the data as two-dimension data points. Give a new data point, p1 = (0.4,0.2) as a query, rank the data points based on the similarity with the query point using
(1) Euclidean distance (2) cosine similarity [round your similarity to three decimal places]

b) Transform each value in your data set and the query point using the sigmoid function:

$$y = \frac{1}{1 + e^x}$$

And re-rank the data points based on the similarity with the query point using
(1) Euclidean distance (2) cosine similarity [round your similarity to three decimal places]

Ans:

$$\text{Euclidean distance} = d(p,q) = \sqrt{\left(p_y - q_x\right)^2 + \left(p_y - q_y\right)^2}$$

$$\text{Cosine similarity} = \cos\theta = \frac{(p_x * q_x)+\left(p_y * q_y\right)}{\sqrt{p_x^2 + q_x^2}\sqrt{p_y^2 + q_y^2}}$$

a)

Data points	X1	X2	Euclidean distance	Rank by Euclidean Distance	Cosine similarity	Rank by Cosine Similarity
p1	0.3	0.8	0.608	2	0.733	4
p2	0.7	0.4	0.361	1	0.998	1
p3	1	0.1	0.608	2	0.934	3
p4	-0.1	-0.3	0.707	4	-0.707	5
p5	0.9	0.8	0.781	5	0.966	2
query point	0.4	0.2	-	-	-	-

b)

Transformed Data points	X1	X2	Euclidean distance	Rank by Euclidean Distance	Cosine similarity	Rank by Cosine Similarity
p1	0.574	0.69	0.142	3	0.991	5
p2	0.668	0.599	0.085	1	1	1
p3	0.731	0.525	0.134	2	0.993	4
p4	0.475	0.426	0.175	4	1	1
p5	0.711	0.69	0.179	5	1	1
query point	0.599	0.55	-	-	-	-

QUESTION 3

What is a metric space? What are the conditions a function must satisfy to be called a metric? Explain in full generality.

Ans:

Metric space is a set M with a function d (known as metric), in which each pair (x, y) ∈ M has a distance given by metric d(x, y). In the metric space, distances between all such pairs are defined by the metric.

For a function to be a metric, it must satisfy the following conditions.

Let d(p, q) be the distance metric between two data points p and q

1. **Positive definiteness:** $d(p,q) \geq 0,$ for all p and q and

   $d(p,q) = 0,$ for p = q
2. **Symmetry:** $d(p,q) = d(q,p),$ for all p and q
3. **Triangle Inequality:** $d(p,r) \leq d(p,q) + d(q,r),$ for all p, q, and r

QUESTION 4(A)

For each of the following, indicate whether the variable is binary/discrete/continuous, nominal/ordinal/interval/ratio.

Variable	Binary/Discrete/ Continuous	Nominal/Ordinal/ Interval/Ratio	Assumptions (If Any)
Brightness in terms of dark or light	Binary	Ordinal	–
Temperature in Fahrenheit	Continuous	Interval	There is no level of measurement for 0 temperature in Fahrenheit.
Barcodes	Discrete	Nominal	–
Student grades	Discrete	Ordinal	If course grades like A, B, C, D are considered, then they are ordinal since they represent student's performance and their order matters.
		Ratio	If the grades are considered as numbers like 80, 90, 100, etc., then they are ratio. There is a level of measurement for 0 value as well, thus the ratio.
Weight of an object	Continuous	Ratio	There is no level of measurement for 0 weight.