MACHINE LEARNING – WORKSHEET

(CLUSTERING)

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is an application of clustering
a. Biological network analysis
b. Market trend prediction
c. Topic modeling
d. All of the above
Answer = d. All of the above

2. On which data type, we cannot perform cluster analysis?
a. Time series data
b. Text data
c. Multimedia data
d. None
Answer = d.None

3. Netflix's movie recommendation system uses-
a. Supervised learning
b. Unsupervised learning
c. Reinforcement learning
d. All of the above
Answer = c. Reinforcement Learning

4. The final output of Hierarchical clustering is-
a. The number of cluster centroids
b. The tree representing how close the data points are to each other
c. A map defining the similar data points into individual groups
d. All of the above
Answer= b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?
a. a distance metric
b. initial number of clusters
c. initial guess as to cluster centroids
d. None
Answer = d.None

6. Which is the following is wrong?
a. k-means clustering is a vector quantization method
b. k-means clustering tries to group n observations into k clusters
c. k-nearest neighbor is same as k-means
d. None

Answer = k-nearest neighbor is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. Single-link
2. Complete-link
3. Average-link

Options:
a. 1 and 2
b. 1 and 3
c. 2 and 3
d. 1, 2 and 3

Answer = d. 1,2& 3

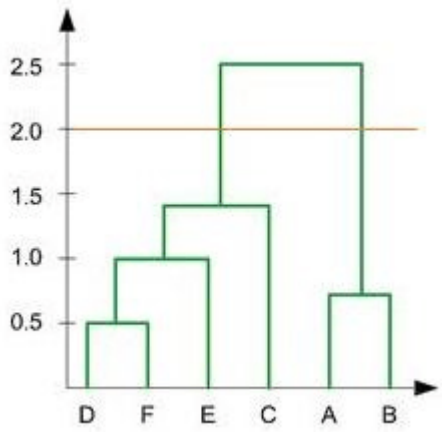
8. Which of the following are true?

1. Clustering analysis is negatively affected by multicollinearity of features
2. Clustering analysis is negatively affected by heteroscedasticity

Options:
a. 1 only
b. 2 only
c. 1 and 2
d. None of them

Answer = a. 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?



a. 2
b. 4
c. 3
d. 5

Answer = a.2

10. For which of the following tasks might clustering be a suitable approach?

a. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

b. Given a database of information about your users, automatically group them into different market segments.

c. Predicting whether stock price of a company will increase tomorrow.

d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

11. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

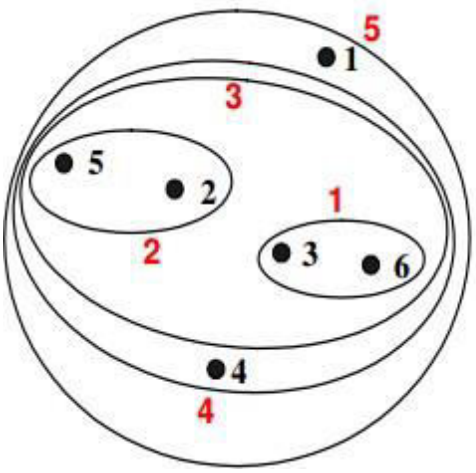
Table : X-Y coordinates of six points.

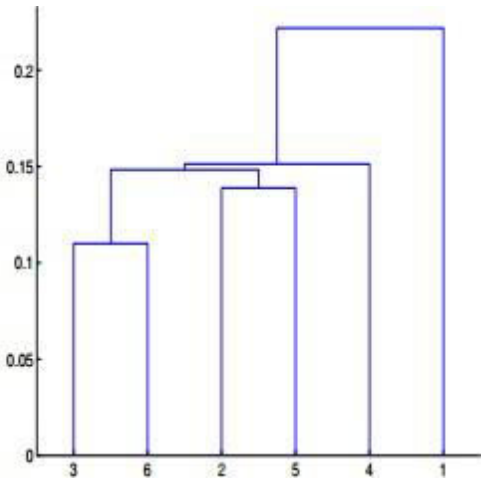
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

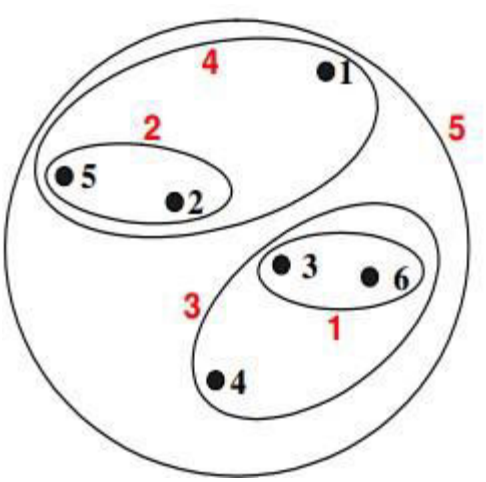
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

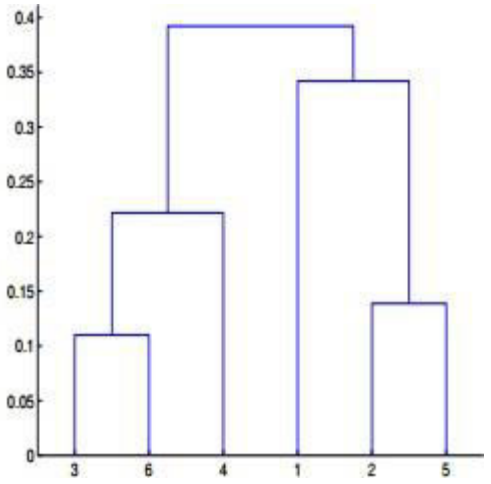
A.



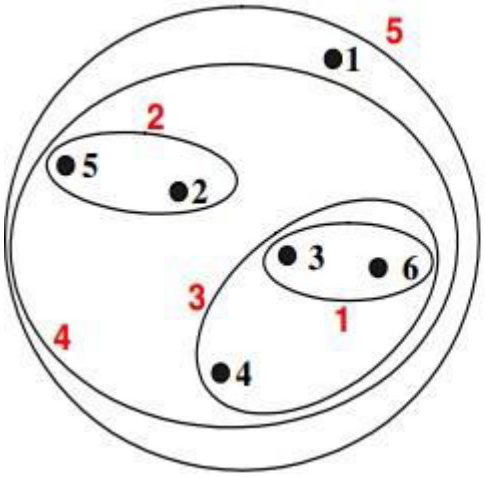


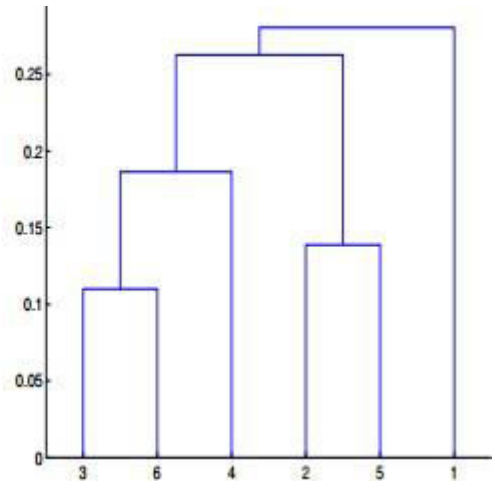
B



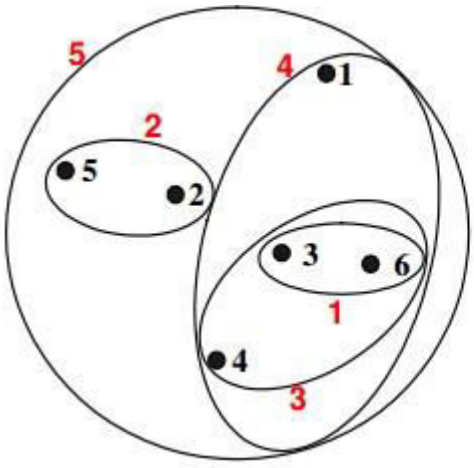


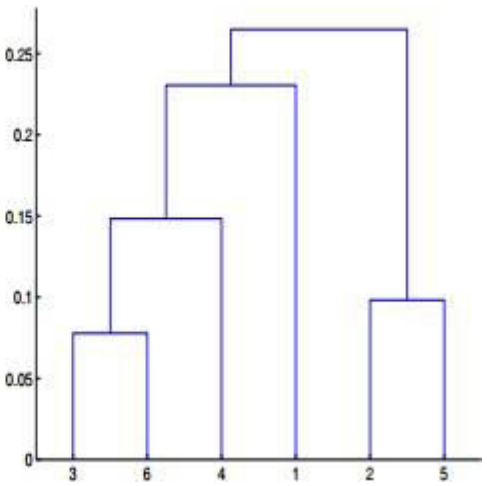
C.





D.





Answer = A

12. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

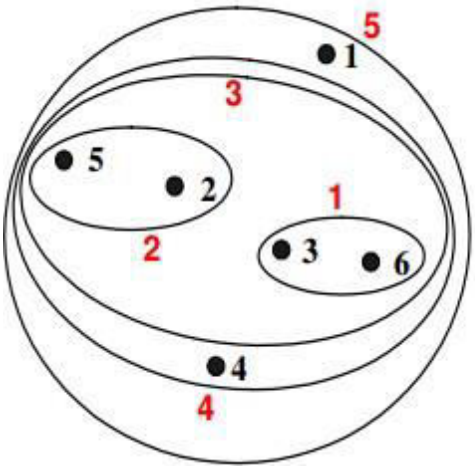
Table : X-Y coordinates of six points.

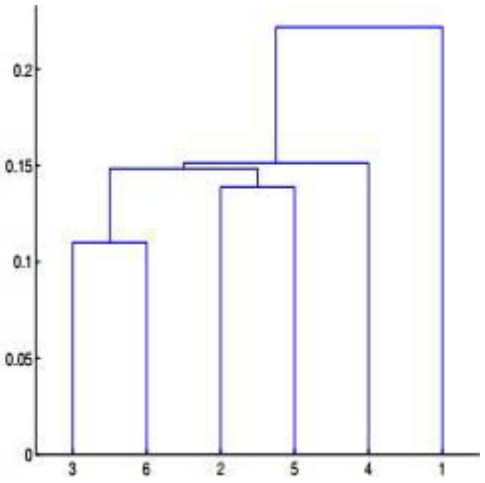
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

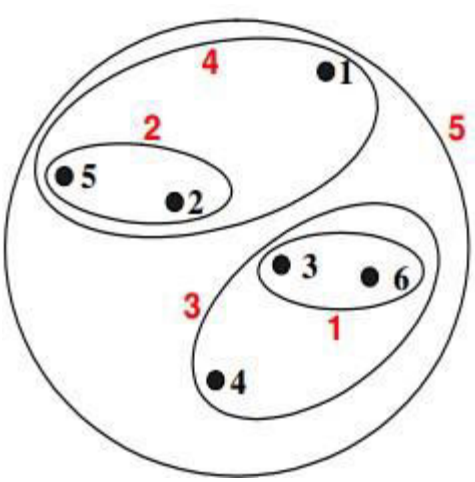
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering:

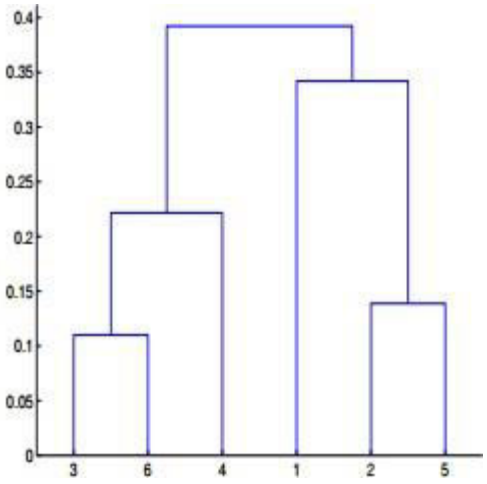
A



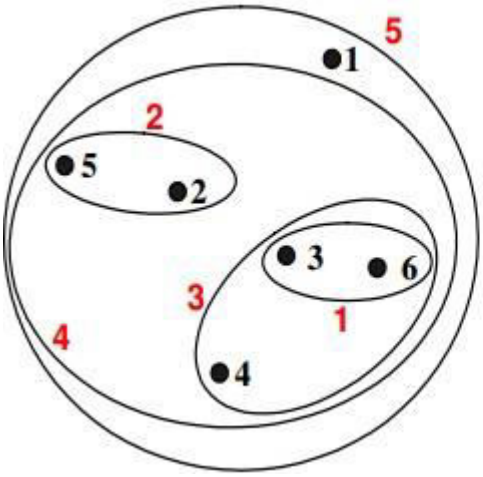


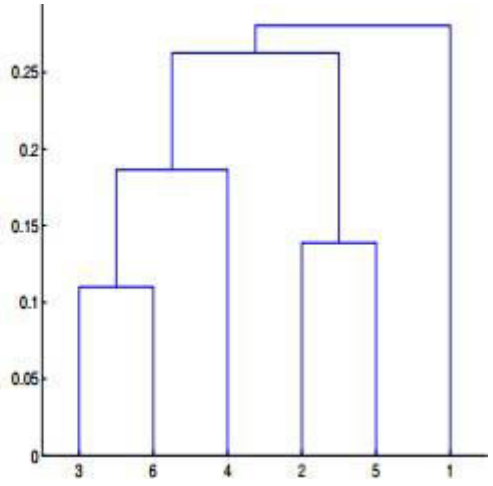
B.


4
1
2
5
5
2
3
6
3
1
4

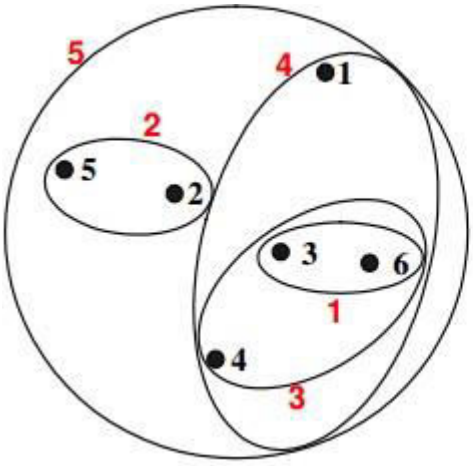

0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

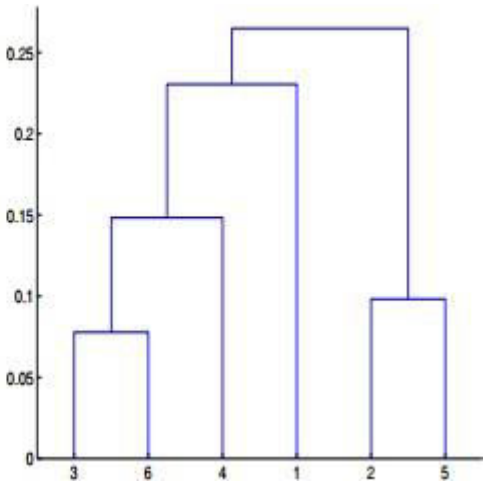
C.


5
1
2
5
2
3
6
3
1
4
4


0.25
0.2
0.15
0.1
0.05
0
3
6
4
2
5
1

D.


5
4
1
2
5
2
3
6
1
4
3


0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

Answer = B

13. What is the importance of clustering?

Answer = Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models.

14. How do you cluster a profile?

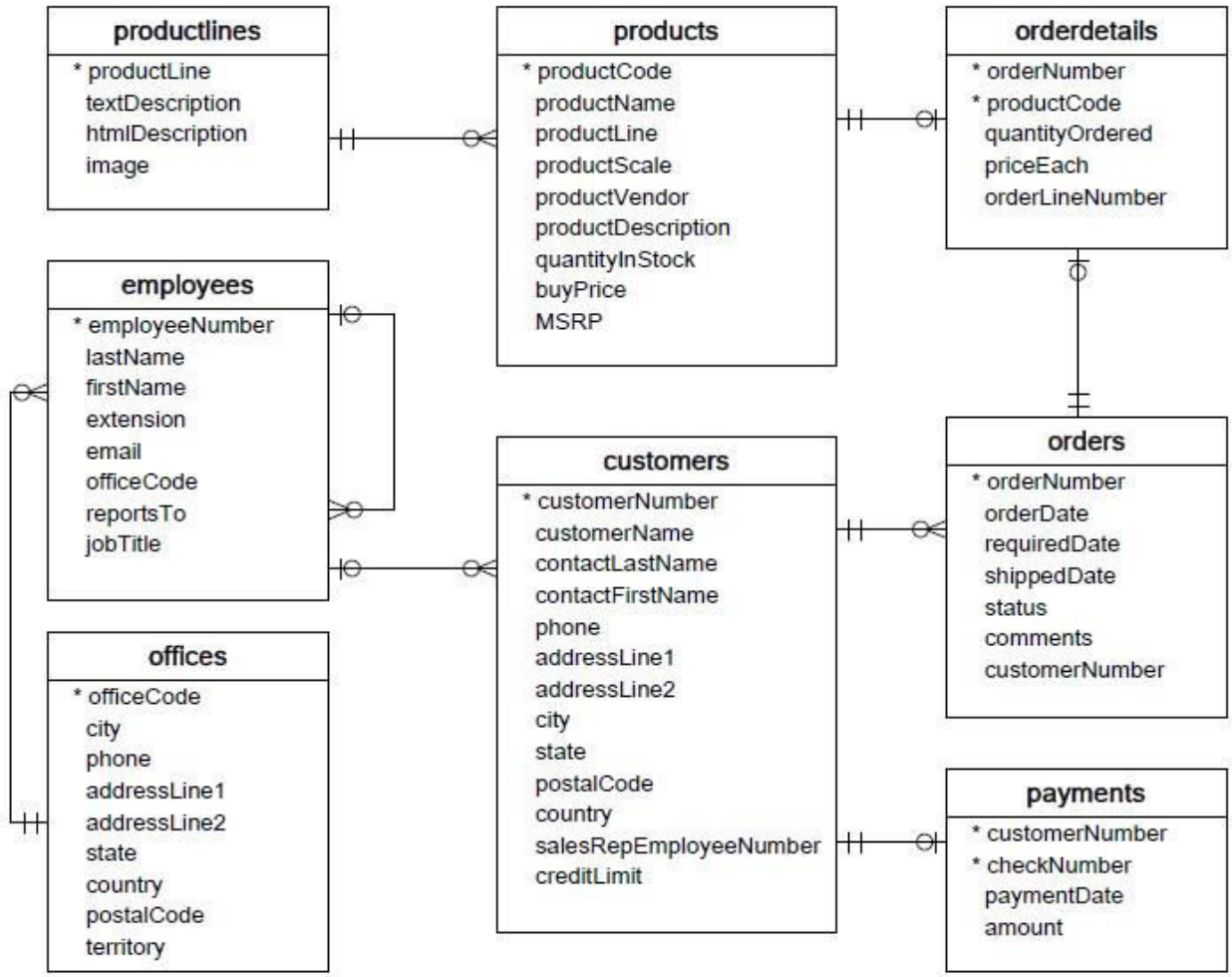
Answer =
Step 1: Confirm data is metric.
Step 2: Scale the data.
Step 3: Select Segmentation Variables.
Step 4: Define similarity measure.
Step 5: Visualize Pair-wise Distances.
Step 6: Method and Number of Segments.
Step 7: Profile and interpret the segments.
Step 8: Robustness Analysis.

15. How can I improve my clustering performance?

Answer = K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm. When the data has overlapping clusters, k-means can improve the results of the initialization technique

SQL – WORKSHEET 3

Refer the following ERD and answer all the questions in this worksheet. You have to write the queries using mysql for the required Operation.



- **Customers**: stores customer's data.
- **Products**: stores a list of scale model cars.
- **ProductLines**: stores a list of product line categories.
- **Orders**: stores sales orders placed by customers.
- **OrderDetails**: stores sales order line items for each sales order.
- **Payments**: stores payments made by customers based on their accounts.
- **Employees**: stores all employee information as well as the organization structure such as who reports to whom.
- **Offices**: stores sales office data.

1. Write SQL query to create table **Customers.**

Answer = Create Table Customers(customerNumber int NOTNULL
,customerName varchar
,contactLastname char,
contactFirstname char,
phone char(10),
address line 1 char,
address line 2 char,
city char,
postal code char(6),
country,
salesRepEmployeeNumber char(10),

credit limit char):

2. Write SQL query to create table **Orders.**
Answer = Create Table Orders(OrderNumber int NOT NULL,
Orderdate char(10) NOT NULL,
requiredDate char(10) NOT NULL,
shippedDate(10) NOT NULL,
status text NOT NULL,
comments text NOT NULL,
customerNumber char);

3. Write SQL query to show all the columns data from the **Orders** Table.
Answer = Select * from Orders;

4. Write SQL query to show all the comments from the **Orders** Table.
Answer = Select comments from Orders;

5. Write a SQL query to show orderDate and Total number of orders placed on that date, from **Orders** table.
Answer = Select orderDate, count(OrderNumber) where OrderDate='';

6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from **employees** table.
Answer = Select employeeNumber, lastName , firstNme from employees;

7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.
Answer = select orderNumber, customerName where customerName ='';

8. Write a SQL query to show name of all the customers in one column and salerepemployee name in another column.

9. Write a SQL query to show Date in one column and total payment amount of the payments made on that date from the **payments** table.
10. Write a SQL query to show all the products productName, MSRP, productDescription from the **products** table.
11. Write a SQL query to print the productName, productDescription of the most ordered product.
12. Write a SQL query to print the city name where maximum number of orders were placed.
13. Write a SQL query to get the name of the state having maximum number of customers.
14. Write a SQL query to print the employee number in one column and Full name of the employee in the second column for all the employees.
15. Write a SQL query to print the orderNumber, customer Name and total amount paid by the customer for that order (quantityOrdered × priceEach).

STATISTICS– WORKSHEET 3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
   a) Total Variation = Residual Variation – Regression Variation
   b) Total Variation = Residual Variation + Regression Variation
   c) Total Variation = Residual Variation * Regression Variation
   d) All of the mentioned

Answer = b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called_____outcomes.
   a) random
   b) direct
   c) binomial
   d) none of the mentioned

Answer = c) binomial

3. How many outcomes are possible with Bernoulli trial?
   a) 2
   b) 3
   c) 4
   d) None of the mentioned

Answer= a)2

4. If Ho is true and we reject it is called
   a) Type-I error
   b) Type-II error
   c) Standard error
   d) Sampling error

Answer =a) type-1 error

5. Level of significance is also called:
   a) Power of the test
   b) Size of the test
   c) Level of confidence
   d) Confidence coefficient

Answer = a) Power of the test

6. The chance of rejecting a true hypothesis decreases when sample size is:
   a) Decrease
   b) Increase
   c) Both of them
   d) None

Answer = c) both of them

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

Answer = b)hypothesis

8. What is the purpose of multiple testing in statistical inference?

a) Minimize errors
b) Minimize false positives
c) Minimize false negatives
d) All of the mentioned

Answer = d) all of the mentioned

9. Normalized data are centred at____and have units equal to standard deviations of the original data

a) 0
b) 5
c) 1
d) 10

Answer = a) 0

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

Answer= Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of “causes”

11. What is z-score?

Answer = A z score is simply defined as the number of standard deviation from the mean. The z-score can be calculated by subtracting mean by test value and dividing it by standard value

12. What is t-test?

Answer= The t-test is a test for the hypothesis of equal means, whereas the WMW test is less specific. If the underlying distributions of the variable in the two groups differ only in location, i.e. in means and medians, the WMW test is a test for the hypothesis of equal medians

13. What is percentile?

Answer= In statistics, a percentile is a score below which a given percentage of scores in its frequency distribution fall or a score at or below which a given percentage fall. The percentile and the percentile rank are related terms.

14. What is ANOVA?

Answer = An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you're testing groups to see if there's a difference between them.

15. How can ANOVA help?

Answer= ANOVA checks the impact of one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not. Another measure to compare the samples is called a t-test. When we have only two samples, t-test and ANOVA give the same results.