MACHINE LEARNING

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

2. On which data type, we cannot perform cluster analysis?

a. Time series data

b. Text data

c. Multimedia data

d. None

Answer: a

3. Netflix's movie recommendation system uses-

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning and Unsupervised learning

d. All of the above

Answer: a

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: d

5. Which of the step is not required for for K-means clustering?

a. A distance metric

b. Initial number of clusters

c. Initial guess as to cluster centroids

d. None

Answer: d

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 neighbour is same as k-means

d. None

Answer: d

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

i. Single-link

ii. Complete-link

iii. Average-link

Options:

a. 1 and 2

b. 1 and 3

c. 2 and 3

d. 1, 2 and 3

Answer: d

8. Which of the following are true?

i. Clustering analysis is negatively affected by multicollinearity of features

ii. Clustering analysis is negatively affected by heteroscedasticity

Options:

a. 1 only

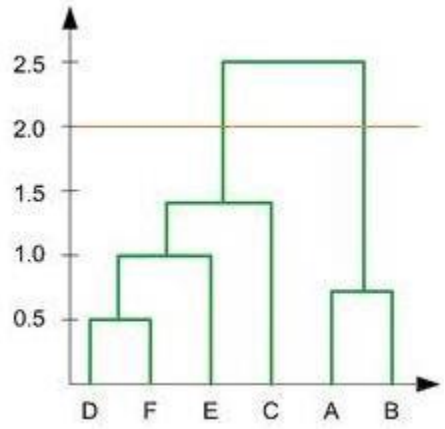
b. 2 only

c. 1 and 2

d. None of them

Answer: a

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

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.

Answer: b

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

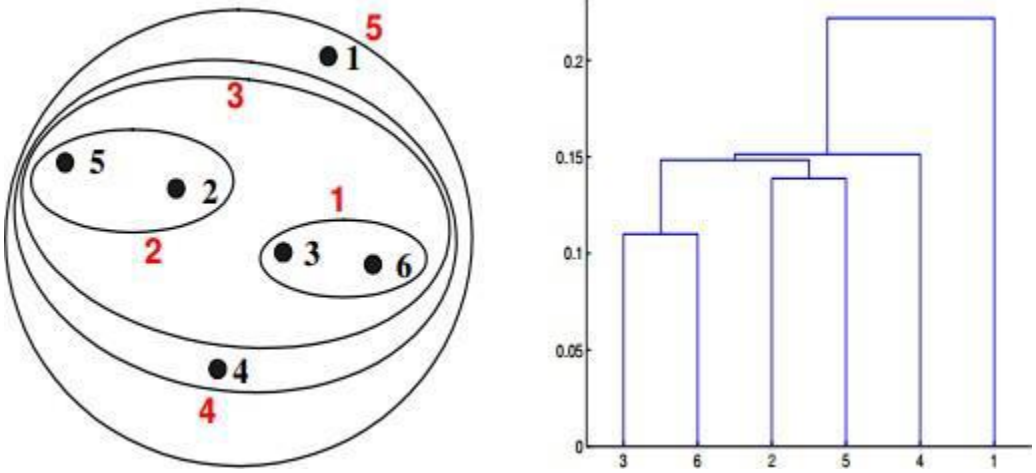
Table : X-Y coordinates of six points.

	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

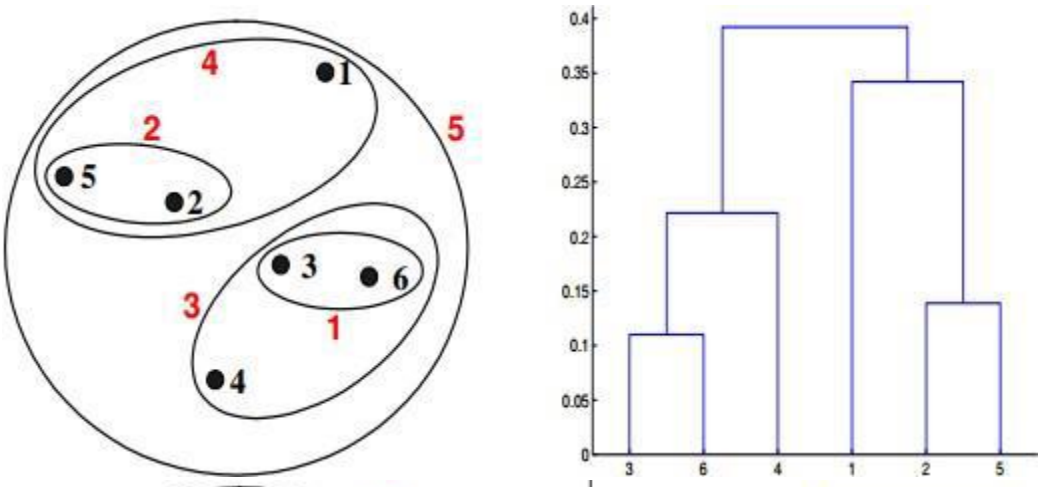
Table : Distance Matrix for Six Points

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

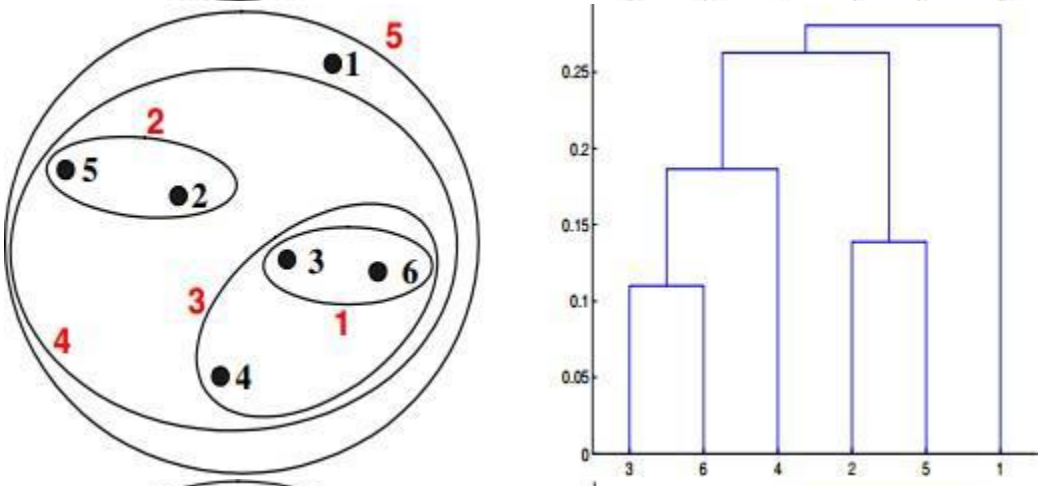
a.



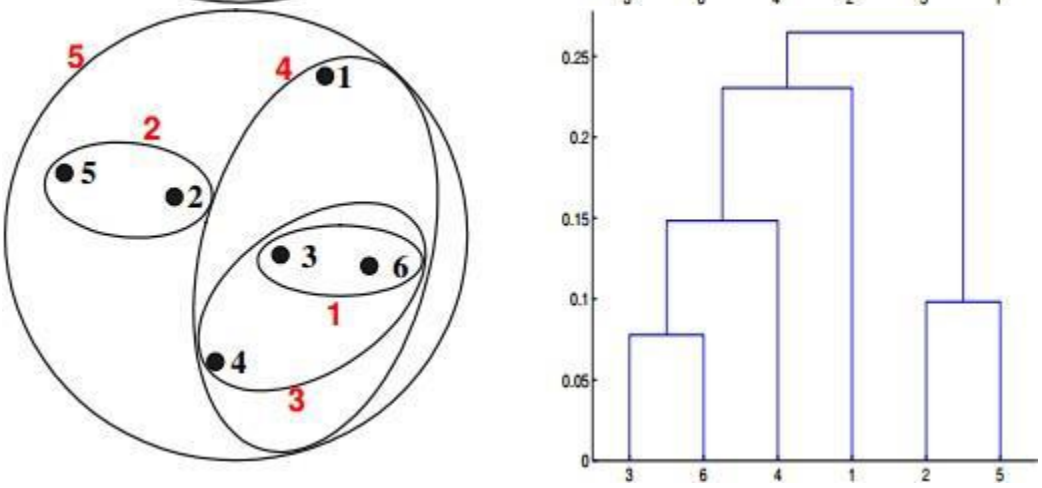
b.



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

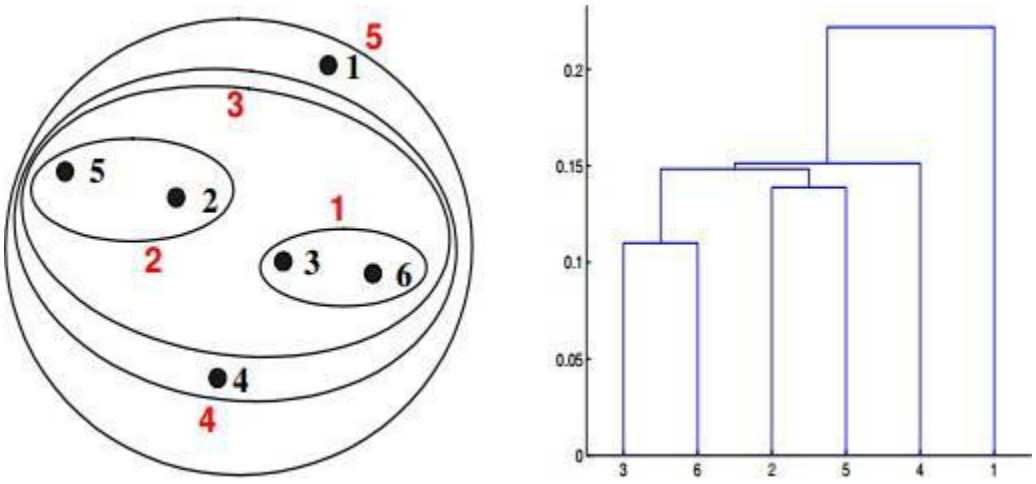
Table : X-Y coordinates of six points.

	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

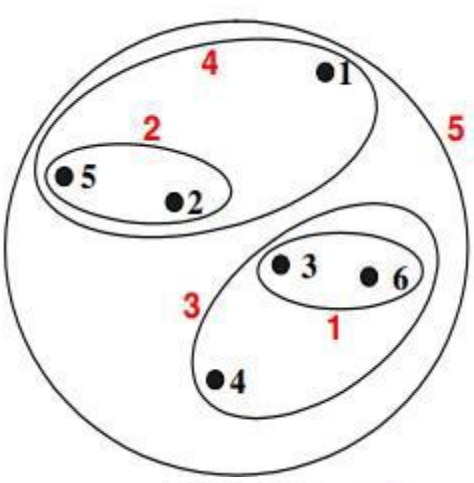
Table : Distance Matrix for Six Points

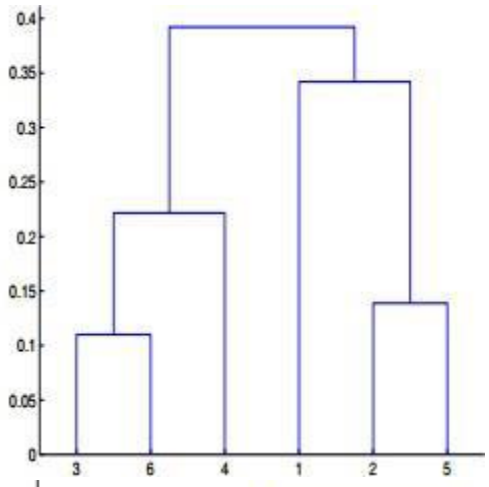
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

a.

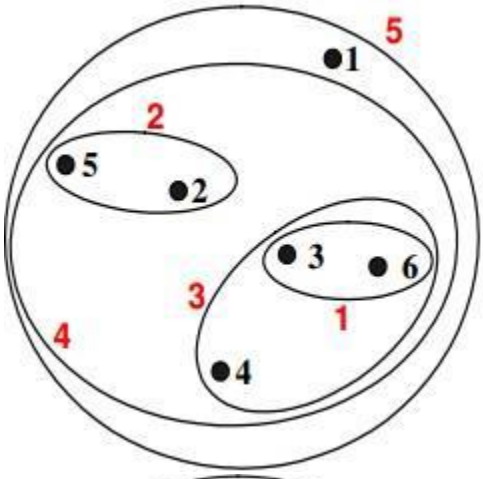


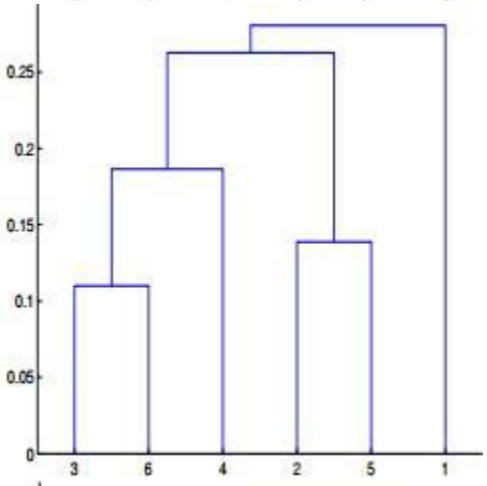
b.



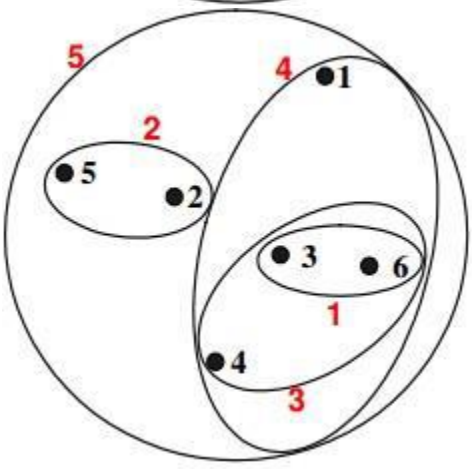


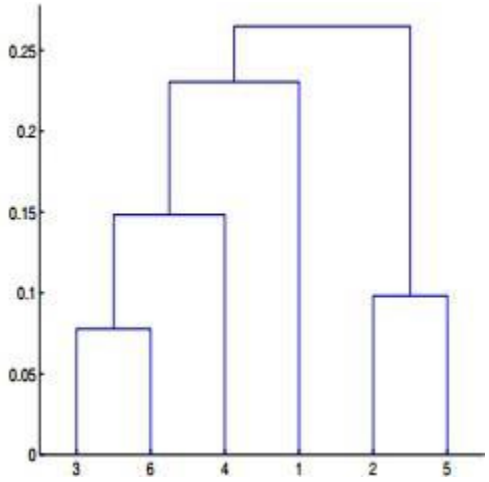
c.





d.





Answer: b

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

Answer: The process of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. The process of clustering is important in data analysis and data mining applications. There are various types of clustering methods; they are

Hierarchical methods
Partitioning methods
Density-based
Model-based clustering
Grid-based model
Clustering Intelligence Servers provides the following benefits:

- Simplified management:
- Increased resource availability:
- Greater scalability:
- Strategic resource usage:
- Increased performance:

14. How can I improve my clustering performance?

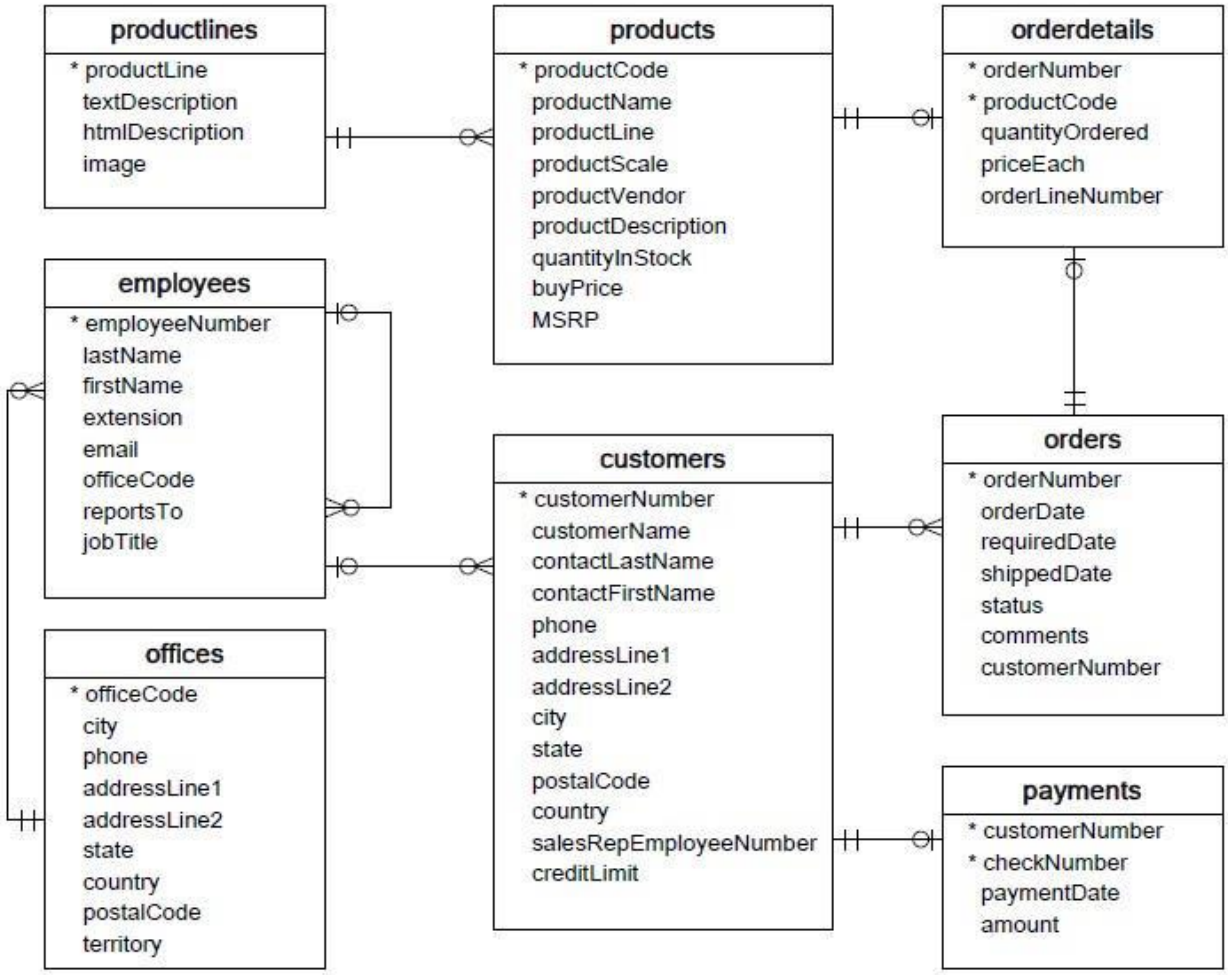
Answer: An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means.

PCA is a classical multivariate data analysis method that is useful in linear feature extraction. Without class labels it can compress the most information in the original data space into a few new features, i.e., principal components. Handling high dimensional data using clustering techniques obviously a difficult task in terms of higher number of variables involved. In order to improve the efficiency, the noisy and outlier data may be removed and minimize the execution time and we have to reduce the no. of variables in the original data set. The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables. It is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set.

K-means is a prototype-based, simple partitioned clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice. The algorithm consist of two separate phases: the first phase is to select k centers randomly, where the value of k is fixed in advance. The next phase is to assign each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, recalculating the average of the clusters. This iterative process continues repeatedly until the criterion function becomes minimum

WORKSHEET 3 SQL

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 Customers

2. Write SQL query to create table Orders.

Answer: CREATE Orders

3. Write SQL query to show all the columns data from the Orders Table.

Answer: SELECT * FROM Customers.

4. Write SQL query to show all the comments from the Orders Table.

Answer: SELECT * FROM Orders.
5. Write a SQL query to show order Date and Total number of orders placed on that date, from Orders table.
Answer: SELECT order Date and Total number FROM Orders
6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from employees table.
Answer: SELECT employeNumber, lastName, firstName FROM employees
7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.
Answer: SELECT orderNumber, customerName FROM customers, orderdetails SORT BY orderNumber.
8. Write a SQL query to show name of all the customers in one column and saleRepemployeenumber in another column.
Answer: SELECT customerName, saleRepemployeenumber FROM customers
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.
Answer: SELECT orderDate, checkNumberpaymentDate amount FROM orders, payment.
10. Write a SQL query to show all the products productName, MSRP, productDescription from the products table.
Answer: SELECT productName, MSRP, productDescription amount FROM products
11. Write a SQL query to print the productName, productDescription of the most ordered product.
Answer: SELECT porductCode, productName, productDescription FROM orderdetails, products SORT BY orderNumber
12. Write a SQL query to print the city name where maximum number of orders were placed.
Answer: SELECT porductCode, productName, productDescription FROM orderdetails, products SORT BY orderNumber
13. Write a SQL query to get the name of the state having maximum number of customers.
Answer: SELECT city FROM customers,orders WHERE max(orderNumber)
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.
Answer: SELECT employee number,lastName, firstName FROM employee
15. Write a SQL query to print the orderNumber, customer Name and total amount paid by the customer for that order (quantityOrdered × priceEach).
Answer: SELECT orderNumber, customer Name, amount FROM orders, payment GROUPBY 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: a

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: a

3. How many outcomes are possible with Bernoulli trial?

a) 2

b) 3

c) 4

d) None of the mentioned

Answer: a

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

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: c

6. The chance of rejecting a true hypothesis decreases when sample size is:

a) Decrease

b) Increase

c) Both of them

d) None

Answer: b

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

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

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

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

10. What Is Bayes' Theorem?

Answer: Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

The statement of Bayes Theorem is as follows: Let E1,E2,E3,...,EnE1,E2,E3,...,En be a set of events associated with a sample space S, where all events E1,E2,E3,...,EnE1,E2,E3,...,En have non-zero probability of occurrence and they form a partition of S. Let A be any event which occurs with E1orE2orE3...orEnE1orE2orE3...orEn, then according to Bayes Theorem, $P(Ei|A)=P(Ei)P(A|Ei)\sum nk=1P(Ek)P(A|Ek),i=1,2,3,...,nP(Ei|A)=P(Ei)P(A|Ei)\sum k=1nP(Ek)P(A|Ek),i=1,2,3,...,n$

- Here Eii ∩ Ejj = ϕ, where i ≠ j. (i.e) They are mutually exhaustive events
- The union of all the events of the partition, should give the sample space.
- $0 \leq P(Eii) \leq 1$

11. What is z-score?

Answer: A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean. A z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.
A z-score can be placed on a normal distribution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean μ and also the population standard deviation σ.
Z-scores are a way to compare results to a "normal" population. Results from tests or surveys have thousands of possible results and units; those results can often seem meaningless. For example, knowing that someone's weight is 150 pounds might be good information, but if you want to compare it to the "average" person's weight, looking at a vast table of data can be overwhelming (especially if some weights are recorded in kilograms). A z-score can tell you where that person's weight is compared to the average population's mean weight.

12. What is t-test?

Answer: A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another. A t-test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.

13. What is percentile?

Answer: A percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular score surpassed. For example, if you score 75 points on a test, and are ranked in the 85th percentile, it means that the score 75 is higher than 85% of the scores.

14. What is ANOVA?

Answer: Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study. The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.12 ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers."3 It was employed in experimental psychology and later expanded to subjects that were more complex.

15. How can ANOVA help?

Answer: There are many types of ANOVA
1. One-way ANOVA
2. Two-way ANOVA
3. MANOVA

The different types of ANOVA are:

1. ANOVA is helpful for testing three or more variables
2. ANOVA gives most significant and influential factors
3. ANOVA also gives the most contributing factors.