WORKSHEET SET 3: SOLUTIONS

STATISTICS WORKSHEET-3

Q1. (B) Total Variation = Residual Variation + Regression Variation

Q2. (C) Binomial

Q3. (A) 2

Q4. (A) Type-I error

Q5. (B) Size of the test

Q6. (B) Increase

Q7. (B) Hypothesis

Q8. (D) All of the mentioned

Q9. (A) 0

Q10. Bayes' Theorem, named after mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. It states that, the conditional probability of an event which is 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. In other words, it describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

Q11. Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. It measures the distance between a data point and the mean using standard deviations.

Q12. T-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing. It is used to determine if there is a significant difference between the means of two groups and how they are related. It is used to determine whether a process has an effect on the population of interest, or whether two groups are different from one another.

Q13. A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. It is a score below which a given percentage of scores in its frequency distribution falls.

Q14. ANOVA stands for Analysis of Variance. It was developed by Ronald Fisher in 1918. ANOVA is a collection of statistical models and their associated estimation procedures which are used to analyze the differences between means. It is a powerful & common statistical procedure used in social sciences.

Q15. ANOVA is helpful for testing three or more variables. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. ANOVA is used to test differences among multiple means without increasing the Type I error rate. As the number of groups increases, the number pair comparisons increases substantially and calculations become overwhelming very quickly. If we test enough pairs, we begin to make observations that are less significant, until we find p values that are insignificant. ANOVA puts all the data into one F number and gives us one P to test the null hypothesis.

SQL WORKSHEET-3

Q1. CREATE TABLE customers (customerNumber INT UNSIGNED AUTO_INCREMENT PRIMARY KEY, customerName VARCHAR (20) NOT NULL, contactLastName VARCHAR (20), contactFirstName VARCHAR (20), phone INT(10), addressLine1 VARCHAR (25), addressLine2 VARCHAR (25), city VARCHAR (15), state VARCHAR (15), postalCode INT (8), country VARCHAR (15), FOREIGN KEY(salesRepEmployeeNumber) REFERENCES employees(employeeNumber));

Q2. CREATE TABLE orders (orderNumber INT UNSIGNED AUTO_INCREMENT PRIMARY KEY, orderDate DATE NOT NULL, requiredDate DATE, shippedDate DATE, status VARCHAR (20), comments VARCHAR (30), FOREIGN KEY (customerNumber) REFERENCES customers (customerNumber));

Q3. SELECT * FROM orders;

Q4. SELECT comments FROM orders;

Q5. SELECT orderDate, count (orderNumber) FROM orders GROUP BY orderNumber;

Q6. SELECT employeeNumber, lastName, firstName FROM employees;

Q7. SELECT orderNumber, customerName FROM orders customers WHERE orders.customerName = customers.customerName order by orderNumber;

Q8. SELECT customerName, concat (lastName, “”, firstName) FROM customers JOIN employees on customers.salesRepEmployeeNumber=employees.employeeNumber;

Q9. SELECT paymentDate, sum (amount) FROM payments GROUP BY paymentDate;

Q10. SELECT productName, MSRP, productDescription FROM products;

Q11. SELECT productName, productDescription FROM products WHERE productName = MAX (productName);

Q12. SELECT customers.city, count (orders.orderNumber) as c FROM orders INNER JOIN customers on orders.customersNumber=customers.customernumber GROUP BY customers.city ORDER BY c desc LIMIT 1;

Q13. SELECT state FROM from customers GROUP BY state ORDER BY count (state) desc limit 1;

Q14. SELECT employeeNumber, concat (firstName, lastName) as ‘fullName’ FROM employees;

Q15. SELECT orderdetails.orderNumber, customerName, (priceEach*quantityOrdered) as total_amount FROM orders JOIN orderdetails on orders.orderNumber=orderdetails.orderNumber JOIN customers on customers.customerNumber=orders.customerNumber GROUP BY orderdetails.orderNumber ORDER BY orderdetails.orderNumber

MACHINE LEARNING WORKSHEET-2

Q1. (D) All of the above.

Q2. (D) None.

Q3. (C) Reinforcement learning and Unsupervised learning

Q4. (B) The tree representing how close the data points are to each other.

Q5. (D) None.

Q6. (C) k-nearest neighbor is same as k-means.

Q7. (D) 1, 2 and 3.

Q8. (A) 1 only.

Q9. (A) 2

Q10. (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.

Q11. (A).

Q12. (B).

Q13. Clustering methods help in grouping the data points into clusters, using the different techniques are used to find the appropriate result for the particular problem, these clustering techniques help in grouping the data points into similar categories, and each of these categories is further divided into subcategories to assist the exploration of the queries output. Few clustering methods include: Hierarchical method, Density-based method, Model-based method, Grid-based method & Partitioning.

Q14. K-means is a very simple and ubiquitous clustering algorithm. But quite often it does not work on our problem because the initialization is bad. There is an improved initialization method, k-means++, which can help to alleviate this problem, which uses a different initialization. Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step. Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.