# WORKSHEET SET 3

# STATISTICS

**1. Which of the following is the correct formula for total variation?**

b) Total Variation = Residual Variation + Regression Variation

**2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.**

c) binomial

**3. How many outcomes are possible with Bernoulli trial?**

a) 2

**4. If Ho is true and we reject it is called**

a) Type-I error

**5. Level of significance is also called:**

b) Size 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

**7. Which of the following testing is concerned with making decisions using data?**

b) Hypothesis

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

d) All of the mentioned

**9. Normalized data are centered at and have units equal to standard deviations of the original data**

a) 0

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

### 10. What Is Bayes' Theorem?

**Ans –** Bayes' Theorem is one of the most important concepts in machine learning. Bayes' theorem is named after the statistician and philosopher Mr. Thomas Bayes. Bayes' theorem is used to calculate the conditional probability machine learning technology. Byes' theorem gives the precise and accurate result.

Naïve Bayes Classification is simplified version of Bayes' Theorem.

### 11. What is z-score?

**Ans –** Z-score is a statistical term for calculating the mean of a set of data. When z=0 then it is identical to the mean value and the distribution is normal. Z-score may be a positive z-score or negative z-score. If the value of z-score is above the mean, z-score should be positive z-score and if the value of z-score is below the mean, then it should be negative z-score.

Z-score is also used to remove the outliers from dataset. Z-score is also called a Standard Score. Formula for calculating the z-score is;

**Z = (x-mue)/sigma**

Where, x be the data point

Mue be the mean of population

Sigma is the standard deviation of population

By using z-score technique we can also calculate the percentile and probabilities.

### 12. What is t-test?

**Ans –** t-test is statistical test which is used to compare the means of two groups. T-test is also known as student's t-test. t-test performed with the help of hypothesis test. In t-test, both the test, null hypothesis and alternate hypothesis are used. The null hypothesis shows the difference of means between the two groups is zero but in the case of alternate hypothesis, it shows different result than zero.

t-test follows the normal distribution.t-test mainly divided into 3 types;

1.one sample t-test

2.two sample t-test

3.paired t-test

## 13. What is percentile?

**Ans –** Percentile is an important technique which is used in statistic. Percentile is different concept than the percentage. Percentile ranges from 1 to 99. There is no $0^{th}$ or $100^{th}$ percentile in statistic. Percentile is nothing but the comparison of values, means it indicates the percentage of scores that a given value is greater or less.

Ex. If a student x scores $70^{th}$ percentile marks in exam, then should say that, x scores better than 70% student of the student in the class and we also say that 30% student scores better than x in the class.

There are several percentiles which known as quartile like, $25^{th}$ percentile called as first quartile or Q1. $50^{th}$ percentile called as second quartile or Q2. $50^{th}$ percentile is also equal to the median it splits the distribution in half. $75^{th}$ percentile called as a third quartile or Q3.

## 14. What is ANOVA?

**Ans –** ANOVA stands for the Analysis of Variance. ANOVA is an important test in statistic. ANOVA is used to find the difference of means between the two or more groups or ANOVA is used to check if the means of two or more groups are different from each other. In ANOVA if the value of p (significance value) is less than 0.05 then we have to reject null hypothesis and accept alternate hypothesis.

If value of p is greater than 0.05 the we have to accept null hypothesis and reject alternate hypothesis. In ANOVA Null hypothesis is valid only when all the sample means are equal and Alternate hypothesis is valid only when the sample means are different from each other. ANOVA is divided between two types One way ANOVA and Two-way ANOVA.

## 15. How can ANOVA help

**Ans –** ANOVA is used to determine the difference of means between two or more groups at a time. Only ANOVA is applied if we have to find the difference of means between two groups. It is very simple and easy test to apply.

# 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?**

d. All of the above

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

d. None

**3. Netflix's movie recommendation system uses?**

c. Reinforcement learning and Unsupervised learning

**4. The final output of Hierarchical clustering is?**

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?**

d. None

**6. Which is the following is wrong?**

c. 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?**

i. Single-link

ii. Complete-link

iii. Average-link

Ans - d. 1, 2 and 3

**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

Ans - 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?**

Ans - a. 2

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

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

**11. Given, six points with the following attributes Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:**

Ans - a.

**12. Given, six points with the following attributes: Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.**

Ans - b.

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

**13. What is the importance of clustering?**

**Ans –** Clustering is the unsupervised learning in machine learning. In unsupervised learning we get the references from the datasets consisting of input data without labelled. Unlabeled data are present in clustering and which type of data have to used is totally depends on the user. Clustering is nothing but the group of datasets which has common characteristics. Different datasets are present in cluster which has a similar characteristic.

Clustering is used many different fields like marketing, biology, insurance etc. Clustering has a two main types;

1.Hard Clustering

2.Soft Clustering

## 14. How can I improve my clustering performance

**Ans –** Clustering is the unsupervised machine learning technique which divided the dataset into different groups according to their attributes.

# SQL

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

**Ans -** customer` (`customerNumber`, `customerName`, `contactLastName`, `contactFirstName`, `phone`, `addressLine1`, `addressLine2`, `city`, `state`, `postalCode`, `country`, `salesRepEmployeeNumber`, `creditLimit`) VALUES (NULL, 'Allenj', 'Den', 'Allenj', '908478390', '7th cross road', 'near market', 'Bangaluru', 'KA', '578900', 'India', '10', '10000'), (NULL, 'july', 'ken', 'july', '983467209', '3rd cross road', 'behind temple', 'Bangaluru', 'KA', '567009', 'India', '20', '20000'), (NULL, 'Nick', 'Linen', 'Nick', '894578120', '420,K.N.Apartment', 'near temple', 'Mumbai', 'MH', '401209', 'India', '30', '30000'), (NULL, 'Jenny', 'Lopez', 'Jenny', '984576800', 'P.D. Apartments', 'Beside royal mart', 'Pune', 'MH', '560052', 'India', '40', '40000'), (NULL, 'Daniel', 'Vittori', 'Daniel', '875643908', '2nd cross road', 'above rana medical', 'Nagpur', 'MH', '404001', 'India', '50', '50000'), (NULL, 'Laura', 'Breint', 'Laura', '875568900', 'M.G. road', 'ville parle', 'Ooty', 'TN', '570009', 'India', '60', '60000'), (NULL, 'John', 'D\'souza', 'John', '770983762', '1st cross road', 'above bakery', 'Kunnor', 'KL', '880009', 'India', '70', '70000')

**2. Write SQL query to create table Orders.**

**Ans** - INSERT INTO `orders` (`orderNumber`, `orderDate`, `requiredDate`, `shippedDate`, `status`, `comment`, `customerNumber`) VALUES ('11', '2022-08-18', '2022-08-23', '2022-08-20', 'free delivery', 'very nice', '1'), (NULL, '2022-08-21', '2022-08-26', '2022-08-24', 'order is dispatched', 'great', '2'), (NULL, '2022-08-27', '2022-08-31', '2022-08-29', 'delivery charges apply', 'so costly', '3'), (NULL, '2022-09-01', '2022-09-06', '2022-09-04', 'currently not available', 'oops!', '4'), (NULL, '2022-09-04', '2022-09-08', '2022-09-06', 'shipped', 'excellent', '5'), (NULL, '2022-09-10', '2022-09-15', '2022-09-12', 'delivered by 10', 'ok', '6'), (NULL, '2022-09-19', '2022-09-24', '2022-09-22', 'out of stock', 'oops!', '7')

**3. Write SQL query to show all the columns data from the OrdersTable.**

**Ans -** - INSERT INTO `orders` (`orderNumber`, `orderDate`, `requiredDate`, `shippedDate`, `status`, `comment`, `customerNumber`) VALUES ('11', '2022-08-18', '2022-08-23', '2022-08-20', 'free delivery', 'very nice', '1'), (NULL, '2022-08-21', '2022-08-26', '2022-08-24', 'order is dispatched', 'great', '2'), (NULL, '2022-08-27', '2022-08-31', '2022-08-29', 'delivery charges apply', 'so costly', '3'), (NULL, '2022-09-01', '2022-09-06', '2022-09-04', 'currently not available', 'oops!', '4'), (NULL, '2022-09-04', '2022-09-08', '2022-09-06', 'shipped', 'excellent', '5'), (NULL, '2022-09-10', '2022-09-15', '2022-09-12', 'delivered by 10', 'ok', '6'), (NULL, '2022-09-19', '2022-09-24', '2022-09-22', 'out of stock', 'oops!', '7')

**4. Write SQL query to show all the comments from the OrdersTable.**

**Ans** - INSERT INTO `orders`(`comment`) VALUES ('very nice', 'great', 'so costly', 'oops!', 'excellent', 'ok', 'oops!')

**5. Write a SQL query to show orderDate and Total number of orders placed on that date, from Orderstable.**

**Ans -**

**6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from employees' table.**

**Ans** - INSERT INTO `employees` (`employeeNumber`, `lastName`, `firstName`, `extension`, `email`, `officeCode`, `reportsTo`, `jobTitle`) VALUES ('10', 'Benny', 'Den', 'skfujunj65', 'den7@email.com', 'BD123', 'Manager', 'Engineer'), (11, 'Dayal', 'Jimmy', 'judfhuun', 'jimmy@email.com', 'JIM456', 'Manager', 'Engineer'), (12, 'Manuel', 'Evan', 'djgnurgjnk', 'emanuel@email.com', 'EM890', 'Team leader', 'Sales person'), (13, 'Stocks', 'Ben', 'jfurmvji', 'ben@email.com', 'SB1112', 'Manager', 'Sales person'), (14, 'Soudi', 'Jen', 'judbvuen', 'jsoudi@email.com', 'JS1314', 'Team leader', 'Engineer'), (15, 'Warner', 'Ellie', 'cirugnnkk', 'ellie@email.com', 'EL1516', 'Manager', 'Sales person'), (16, 'Lurroso', 'Nancy', 'nsflusr', 'nancy@email.com', 'NA1718', 'Manager', 'Engineer')

**7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.**

**Ans -**

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

**Ans -** customer` (`customerName`, `salesRepEmployeeNumber`) VALUES ('Allenj', '10'), ('july', '20',), ('Nick', '30'), ('Jenny', '40'), ('Daniel', '50'), ('Laura', '60'), ('John', '70')

**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.**

**Ans** - INSERT INTO `payments` (`payment Date`, `amount`) VALUES ('2022-08-21', '10000'), ('2022-08-21', '20000'), ('2022-08-25', '20000'), ('2022-08-31', '30000'), ('2022-08-21', '40000'), ('2022-08-29', '10000'), ('2022-08-28', '50000')

**10. Write a SQL query to show all the products productName, MSRP, productDescription from the products table.**

**Ans -** create table`products` (`productCode` VARCHAR(30) NOT NULL AUTO_INCREMENT , `productName` VARCHAR(80) NOT NULL , `productLine` VARCHAR(40) NOT NULL , `productScale` VARCHAR(50) NOT NULL , `productVendor` VARCHAR(50) NOT NULL , `productDescription` VARCHAR(80) NOT NULL , `quantityinStock` INT(70) NOT NULL , `buyPrice` INT(40) NOT NULL , `MSRP` INT(50) NOT NULL

**11. Write a SQL query to print the productName, productDescription of the most ordered product.**

**Ans -**

**12. Write a SQL query to print the city name where maximum number of orders were placed.**

**Ans –** SELECT city, COUNT (DISTINCT,city_name)

MAX (orderNumber)

FROM order table

GROUP_BY city

**13. Write a SQL query to get the name of the state having maximum number of customers.**

**Ans -**

**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.**

**Ans** INSERT INFO `employees` (`employeeNumber`, `lastName`, `firstName`) VALUES ('10', 'Benny', 'Den'), (11, 'Dayal', 'Jimmy'), (12, 'Manuel', 'Evan'), (13, 'Stocks', 'Ben'), (14, 'Soudi', 'Jen'), (15, 'Warner', 'Ellie'), (16, 'Lurroso', 'Nancy')