FORTH ASSIGMENT IN IN INTRENSHP

MACHINE LEARNING

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?
A) Least Square Error B) Maximum Likelihood
C) Logarithmic Loss D) Both A and B
ANS:- A) Least Square Error

2. Which of the following statement is true about outliers in linear regression?
A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers
C) Can’t say D) none of these
ANS:- A) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is ______?
A) Positive B) Negative
C) Zero D) Undefined
ANS:- B) Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable?
A) Regression B) Correlation
C) Both of them D) None of these
ANS:- C) Both of them

5. Which of the following is the reason for over fitting condition?
A) High bias and high variance B) Low bias and low variance
C) Low bias and high variance D) none of these
ANS:- C) Low bias and high variance
6. If output involves label then that model is called as:
A) Descriptive model B) Predictive modal
C) Reinforcement learning D) All of the above
ANS:- C) Reinforcement learning

7. Lasso and Ridge regression techniques belong to _________?
A) Cross validation B) Removing outliers
C) SMOTE D) Regularization
ANS:- D) Regularization

8. To overcome with imbalance dataset which technique can be used?
A) Cross validation B) Regularization
C) Kernel D)SMOTE
ANS:- D) SMOTE

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?
A) TPR and FPR B) Sensitivity and precision
C) Sensitivity and Specificity D) Recall and precision
ANS:- A) TPR and FPR

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
A) True B) False
Ans:- B) False

11. Pick the feature extraction from below:
A) Construction bag of words from a email B) Apply PCA to project high dimensional data
C) Removing stop words D) Forward selection
ANS:- B) Apply PCA to project high dimensional data

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?
A) We don’t have to choose the learning rate.
B) It becomes slow when number of features is very large.
C) We need to iterate.
D) It does not make use of dependent variable.
ANS:-A,B and C

13. Explain the term regularization?

:-Regularization concepts of Machine Learning important concepts of Machine Learning.
:-It is a technique to prevent the model from overfitting by adding extra information to it.
:-Regularization is that will try to apply some kind of pently
(PENTLY MEANS LAMBDA)
:-Regularization two part L1=LASSO
L2=RIDGE

14. Which particular algorithms are used for regularization?

:- Regularization two part L1=LASSO
L2=RIDGE
-LASSO(L1):-Omit Certain Attributes.
Y=a+b1x1+b2x2…bnxn
-RIDGE(L2):-reduce down the cofficent diffirences.
Alpha=.0001
Point:-1. Y=a+b1x1+b2x2…bnxn 2. Alpha=.0001
:-Point 1 and 2 is Elastic Net

Dipply explain L1 and L2

L1(LASSO).
f1,f2,f3← Label
:-L1 is completely vanish.
:-not contributing much to pridect.
:-f1 dosen't exist data set so it will is 0.
:-L1 is completely elimant is a physically but it's a given important zero.
:-Lasso will return best alpha and cofficients after perfoming 10 cross validations.
L2 (RIDGE).
f1, f2,f3← this is label.
:-suppose that if f1 feature is not contributing much to predict the level.
:-f1 is a very less importance to give.

15. Explain the term error present in linear regression equation?

:- An error term is a residual variable produced by a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables.
:- As a result of this incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis.

Formula:-

$Y=\alpha X+\beta\rho+\epsilon$**where:**$\alpha,\beta$=Constant parameters$X,\rho$=Independent variables$\epsilon$=Error term

PYTHON – WORKSHEET 1

1. Which of the following operators is used to calculate remainder in a division?
A)# B) & C) % D) $
ANS:- C) %

2. In python 2//3 is equal to?
A) 0.666 B) 0 C) 1 D) 0.67
ANS:-B)0

3.In python, 6<<2 is equal to?
A)36 B)10 C)24 D)45
ANS:-C)24

4. In python, 6&2 will give which of the following as output?
A) 2 B) True C) False D) 0
ANS:-A)2

5. In python, 6|2 will give which of the following as output?
A) 2 B)4 C) 0 D) 6
ANS:-D)6

6. What does the finally keyword denotes in python?
A) It is used to mark the end of the code
B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block.
C) the finally block will be executed no matter if the try block raises an error or not.
D) None of the above
ANS:- C) the finally block will be executed no matter if the try block raises an error or not.

7. What does raise keyword is used for in python?
A) It is used to raise an exception.
B) It is used to define lambda function
C) it's not a keyword in python.
D) None of the above
ANS:- A) It is used to raise an exception.

8. Which of the following is a common use case of yield keyword in python?
A) in defining an iterator
B) while defining a lambda function
C) in defining a generator
D) in for loop
ANS:- C) in defining a generator

9. Which of the following are the valid variable names?
A) _abc B) 1abc C) abc2 D) None of the above
ANS:-A,B AND C

10. Which of the following are the keywords in python?
A) yield B) raise C) look-in D) all of the above
ANS:-A and B

11. Write a python program to find the factorial of a number.

What is factorial?

:-Factorial is a non-negative integer. It is the product of all positive integers less than or equal to that number you ask for factorial. It is denoted by an exclamation sign (!).

EXAMPLE 1.

```
# Python program to find the factorial of a number provided by the user.

# change the value for a different result
num = 7

# To take input from the user
#num = int(input("Enter a number: "))

factorial = 1

# check if the number is negative, positive or zero
if num < 0:
   print("Sorry, factorial does not exist for negative numbers")
elif num == 0:
   print("The factorial of 0 is 1")
else:
   for i in range(1,num + 1):
       factorial = factorial*i
   print("The factorial of",num,"is",factorial)
RUN
The factorial of 7 is 5040.
```

12. Write a python program to find whether a number is prime or composite.

What is Prime number?

:-Any natural number that is divisible by 1 and itself called Prime Number in Python. Prime number is not divisible by any other numbers except one and itself.

Prime Numbers are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109 etc

What is composite number?

:-Any positive integer that can be formed by multiplying two smaller positive integers is called composite number. In other word, Composite number is a positive integer that has at least one divisor other than 1 and itself.

:-We can say that composite numbers are exactly the numbers that are not prime and not a unit.

Composite numbers are 4, 6, 8, 9, 10, 12, 14, 15, 16, 18, 20, 21, 22, 24, 25, 26, 27, 28, 30, 32, 33, 34, 35, 36, 38, 39, 40, 42, 44, 45, 46, 48, 49, 50, 51, 52, 54, 55, 56, 57, 58, 60, 62, 63, 64, 65, 66, 68, 69, 70, 72, 74, 75, 76, 77, 78, 80, 81, 82, 84, 85, 86, 87, 88, 90, 91, 92, 93, 94, 95, 96, 98, 99, 100, 102, 104, 105, 106, 108, 110, 111, 112, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 128, 129, 130, 132, 133, 134, 135, 136, 138, 140, 141, 142, 143, 144, 145, 146, 147, 148, 150 etc.

```
# Program to check if a number is prime or not
```

EXAMPLE 1

```
num = 29

# To take input from the user
#num = int(input("Enter a number: "))

# define a flag variable
flag = False

if num == 1:
    print(num, "is not a prime number")
elif num > 1:
    # check for factors
    for i in range(2, num):
        if (num % i) == 0:
            # if factor is found, set flag to True
            flag = True
            # break out of loop
            break

    # check if flag is True
    if flag:
        print(num, "is not a prime number")
    else:
        print(num, "is a prime number")
```

RUN

```
29 is a prime number >.
```

EXAMPLE 2

```
num = 407

# To take input from the user
#num = int(input("Enter a number: "))


if num == 1:
    print(num, "is not a prime number")
elif num > 1:
   # check for factors
   for i in range(2,num):
       if (num % i) == 0:
           print(num,"is not a prime number")
           print(i,"times",num//i,"is",num)
```

```
            break

    else:

        print(num,"is a prime number")



# if input number is less than

# or equal to 1, it is not prime

else:

    print(num,"is not a prime number")
```

RUN

```
407 is not a prime number
11 times 37 is 407
```

13. Write a python program to check whether a given string is palindrome or not.

:-Given a string, write a python function to check if it is palindrome or not. A string is said to be a palindrome if the reverse of the string is the same as the string. For example, “radar” is a palindrome, but “radix” is not a palindrome.

Examples:

```
Input : malayalam
Output : Yes

Input : geeks
Output : No
```

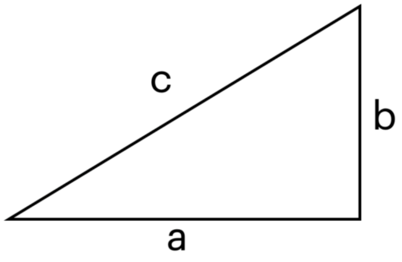
EXAMPLE:-

```
# function to check string is
# palindrome or not
def isPalindrome(s):
    # Using predefined function to
    # reverse to string print(s)
    rev = ''.join(reversed(s))

    # Checking if both string are
    # equal or not
    if (s == rev):
        return True
    return False

# main function
s = "malayalam"
ans = isPalindrome(s)

if (ans):
    print("Yes")
else:
    print("No")
```

Output

```
Yes.
```

14. Write a Python program to get the third side of right-angled triangle from two given sides.



$$c^2 = a^2 + b^2$$

:- The Pythagorean theorem states that given a right triangle, the **hypotenuse squared equals the sum of the sides squared**.

Here is an example.

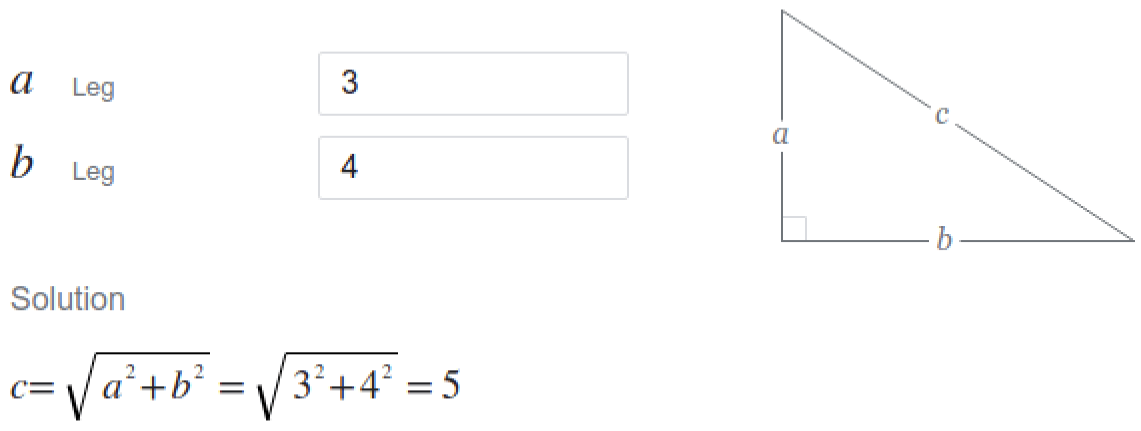
:-Given sides **a = 3** and **b = 4** in a right triangle, what is the length of the hypotenuse?

The solution:

:-Here is an example.

:-Given sides **a = 3** and **b = 4** in a right triangle, what is the length of the hypotenuse?

The solution:



a Leg 3

b Leg 4

Solution

$$c=\sqrt{a^2+b^2}=\sqrt{3^2+4^2}=5$$

Here is the code:

```
import math

a = 3

b = 4

c = math.sqrt(a ** 2 + b ** 2)
```

Output: **5.0**

```
a = float(input("Give side a: "))

b = float(input("Give side b: "))

c = math.sqrt(a ** 2 + b ** 2)

print(f"The length of the hypotenuse c is {c}")
```

Example run:

```
Give side a: 3

Give side b: 4

The length of the hypotenuse c is 5.0
```

15. Write a python program to print the frequency of each of the characters present in a given string.

:-Given a string, the task is to find the frequencies of all the characters in that string and return a dictionary with key as the character and its value as its frequency in the given string.

For example:-

```
Python3 code to demonstrate
# each occurrence frequency using
# naive method

# initializing string
test_str = "GeeksforGeeks"

# using naive method to get count
# of each element in string
all_freq = {}

for i in test_str:
    if i in all_freq:
        all_freq[i] += 1
```

```
  else:
    all_freq[i] = 1

# printing result
print("Count of all characters in GeeksforGeeks is :\n "
   + str(all_freq))
```

Output

```
Count of all characters in GeeksforGeeks is :

{'G': 2, 'e': 4, 'k': 2, 's': 2, 'f': 1, 'o': 1, 'r': 1}
```

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.

a) True b) False

ANS:- a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases? normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem b) Central Mean Theorem

c) Centroid Limit Theorem d) All of the mentioned

ANS:- a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data b) Modeling bounded count data
c) Modeling contingency tables d) All of the mentioned
ANS:- b) Modeling bounded count data

4.Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log-normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned
ANS:- d) All of the mentioned

5. ______ random variables are used to model rates.
a) Empirical b) Binomial
c) Poisson d) All of the mentioned
ANS:- c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True b) False
ANS:- b) False

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

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
a) 0 b) 5 c) 1 d) 10
ANS:- a) 0

9. Which of the following statement is incorrect with respect to outliers?
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned
ANS:- c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
:-NORMAL DISTRUBUTION= normal distribution also called the GAUSSIAN DISTRUTION.
:-normal distribution is commonly see continuous distribution.
:- mean,median and mode all line up such that the center of distribution.
:- classification -
Normal Distribution -

1.) Probability Distribution Function [P.D.F.]
2.) Cumulative Distribution Function [C.D.F.]

11. How do you handle missing data? What imputation techniques do you recommend?

:-What Is a Missing Value?

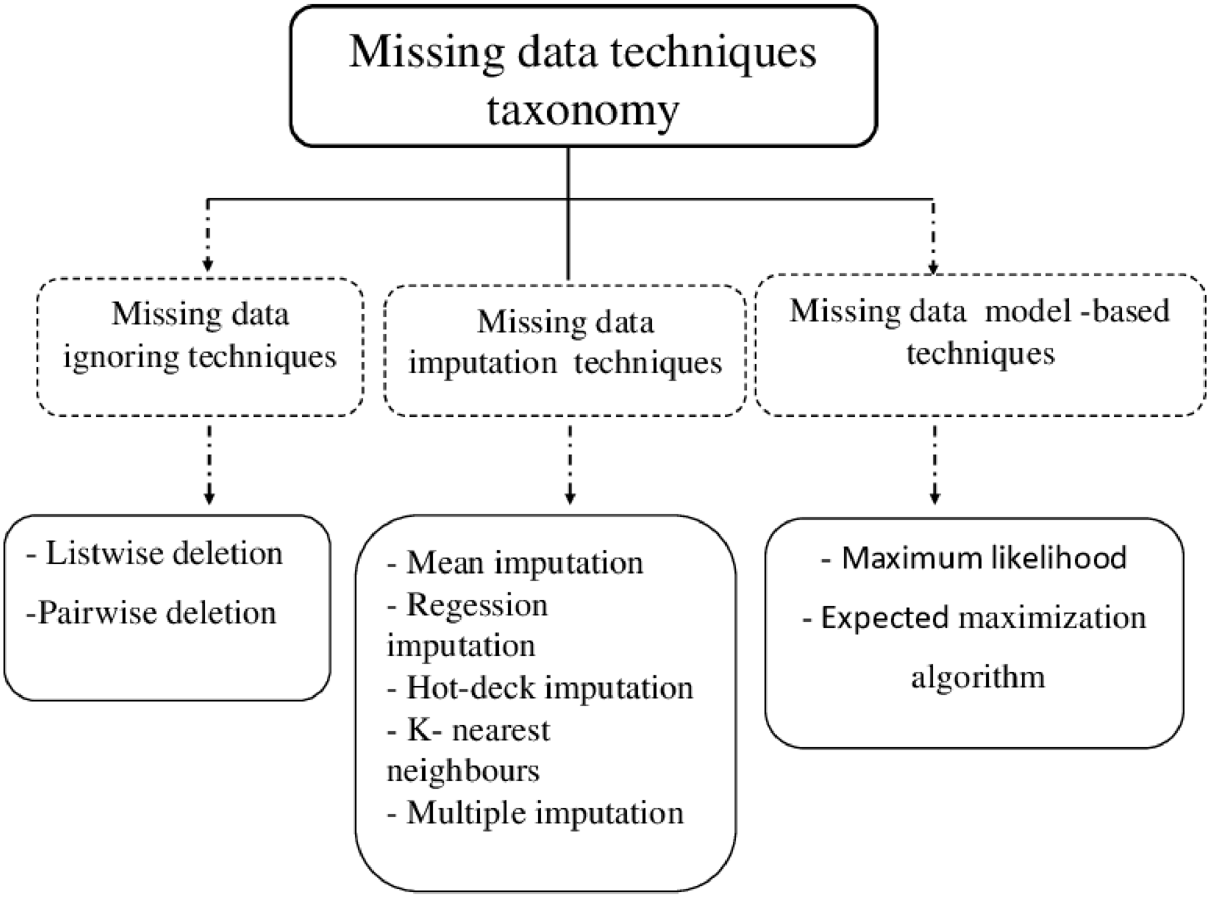
Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. Below is a sample of the missing data from the Titanic dataset. You can see the columns ‘Age’ and ‘Cabin’ have some missing values.

:-How Is a Missing Value Represented in a Dataset?

n the dataset, the blank shows the missing values.

:-In Pandas, usually, missing values are represented by **NaN**. It stands for **Not a Number**.

:-It’s simple structure


Missing data techniques taxonomy
Missing data ignoring techniques
Missing data imputation techniques
Missing data model -based techniques
- Listwise deletion
-Pairwise deletion
- Mean imputation
- Regession imputation
- Hot-deck imputation
- K- nearest neighbours
- Multiple imputation
- Maximum likelihood
- Expected maximization algorithm

12. What is A/B testing?

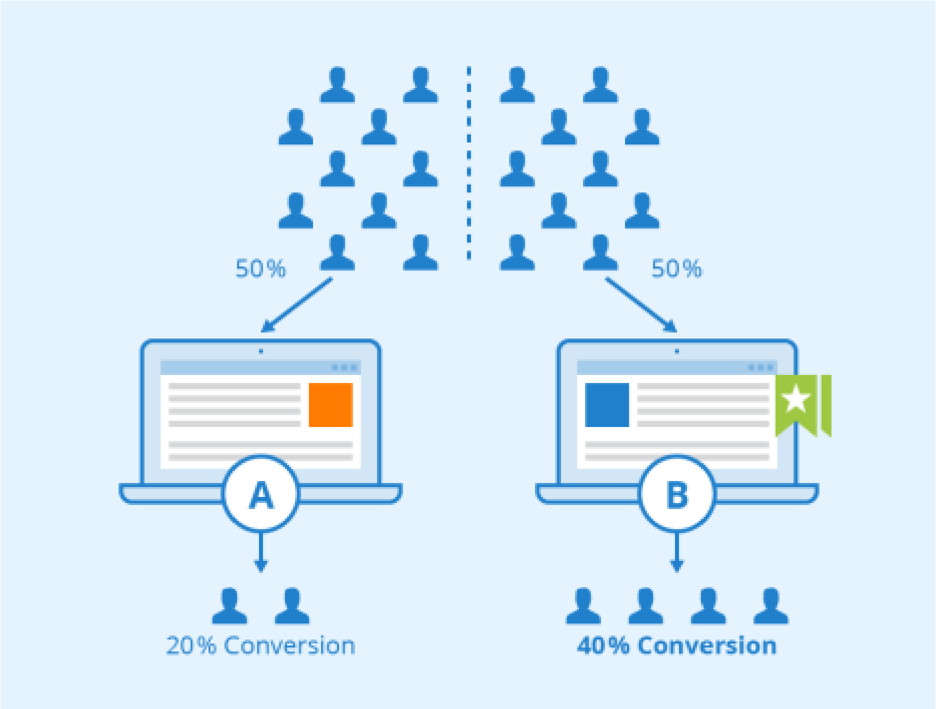
:- An A/B test, also called a controlled experiment or a randomized control trial, is a statistical

:-method of determining which of a set of variants is the best. A/B tests allow organizations

:-and policy-makers to make smarter, data-driven decisions that are less dependent on guess work.

:-In the simplest version of an A/B test, subjects are randomly assigned to either the controlgroup (group A) or the treatment group (group B).

:-Subjects in the treatment group receivethe treatment (such as a new medicine, a special offer, or a new web page design) while the control group proceeds as normal without the treatment. Data is then collected on theoutcomes and used to study the effects of the treatment.



:-This idea has been around for a long time. Historically, farmers have divided their fields into sections to test whether various treatments can improve their crop yield. Something like an A/B nutrition test even appears in the Old Testament!

13. Is mean imputation of missing data acceptable practice?

:-Imputation is a statistical procedure where you replace missing data with *some values*

- Unit imputation = single data point
- Item imputation = single feature valu

Missing Completely at Random (MCAR)

:-Missing Completely at Random, MCAR, means there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others.

:-The probability of missing data on a variable is unrelated to the value of it or to the values of any other variables in the data set.

Missing at Random (MAR)

:-Missing at Random, MAR, means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data. Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual's observed variables. So, for example, if men are more likely to tell you their weight than women, weight is MAR.

:-MAR is weaker than MCAR

$P(Y_{missing}|Y,X)=P(Y_{missing}|X)$

14. What is linear regression in statistics?

ASSUMPTION OF LINER REGRESSION.

:-the liner regression terms of cooficents and error them.

:-the mean of residual is zero.

:-the error terms are not correlected witch each other.

:-tThe independent variables are uncorrelated with each other.

:- The error terms have a constant variance.

:- the error terms are normally distributed

:- residual = error
:-mathematically residual is error is r=y-(mx+b)

:- line of regression Y mx+c

Simple liner regression (slr)

:-SLR is a method for predicting a quantitaive and to apply an email weep response using a single feature

Y=b0+b1x.

15. What are the various branches of statistics?

:-data collection, descriptive statistics and inferential statistics.

Data collection

:-Data collection is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data.

Descriptive statistics

:-Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).

Inferential statistics

:-Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking 'What is this data telling us, and what should we do?'