

# **WORKSHEET SOLUTIONS- FLIPROBO ASSIGNMENT 8**

## **SET-1**

### **Worksheet-1:**

1. B) In hierarchical clustering you don't need to assign number of clusters in beginning.
2. A) max\_depth
3. A) SMOTE
4. D) 2 and 3
5. A) 3-1-2
6. D) Logistic Regression
7. C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node).
8. A) Ridge will lead to some of the coefficients to be very close to 0. & D) Lasso will cause some of the coefficients to become 0.
9. C) Use ridge regularization & D) use Lasso regularization
10. A) Overfitting & C) Underfitting
11. One-Hot-Encoding has the advantage that the result is binary rather than ordinal and that everything sits in a orthogonal vector space. The disadvantage is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality. Also, where for categorical variables where ordinal relationship exists, the one hot encoding is not enough. We have to use Label Encoder for ordinal data.
12. An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. Imbalanced classifications pose a challenge for predictive modeling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class.

Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.

1) Under-sampling- Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling.

2) Over-sampling- On the contrary, oversampling is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples. Rather than getting rid of abundant samples, new rare samples are generated by using e.g. repetition, bootstrapping or SMOTE (Synthetic Minority Over-Sampling Technique).

3) Cluster-Based Over Sampling- In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.

4) Modified synthetic minority oversampling technique (MSMOTE) for imbalanced data. It is a modified version of SMOTE. SMOTE does not consider the underlying distribution of the minority class and latent noises in the dataset. To improve the performance of SMOTE a modified method MSMOTE is used.

13. SMOTE: Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

ADASYN: Adaptive Synthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor.

The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data. The algorithm uses Euclidean distance for KNN Algorithm. The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to

compensate for the skewed distributions. The latter generates the same number of synthetic samples for each original minority sample.

14. Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. There are libraries that have been implemented, such as GridSearchCV of the sklearn library, in order to automate this process. Grid Search can be thought of as an exhaustive search for selecting a model. In Grid Search, the data scientist sets up a grid of hyperparameter values and for each combination, trains a model and scores on the testing data. In this approach, every combination of hyperparameter values is tried and when running it on larger dataset can be very inefficient.

For example, searching 20 different parameter values for each of 4 parameters will require 160,000 trials of cross-validation. This equates to 1,600,000 model fits and 1,600,000 predictions if 10-fold cross validation is used. While Scikit Learn offers the GridSearchCV function to simplify the process, it would be an extremely costly execution both in computing power and time.

15. There are three main errors (metrics) used to evaluate models, Mean absolute error, Mean Squared error and R2 score.

Mean Absolute Error (MAE): Lets take an example where we have some points. We have a line that fits those points. When we do a summation of the absolute value distance from the points to the line, we get Mean absolute error. The problem with this metric is that it is not differentiable.

Mean Squared Error (MSE): Mean Squared Error solves differentiability problem of the MAE. Consider the same diagram above. We have a line that fits those points. When we do a summation of the square of distances from the points to the line, we get Mean squared error.

R2 Score: R2 score answers the question that if this simple model has a larger error than the linear regression model. However, in terms of metrics the answer we need is how much larger. The R2 score answers this question. R2 score is  $1 - (\text{Error from Linear Regression Model} / \text{Simple average model})$ .

## Worksheet-2:

1. C) %
2. B) 0
3. C) 24
4. A) 2
5. D) 6
6. C) the finally block will be executed no matter if the try block raises an error or not.
7. A) It is used to raise an exception.
8. C) in defining a generator
9. A) \_abc B) 1abc  
C) abc2
10. A) yield B) raise
11. *# taking input from user*

```
num=int(input("Enter the num: " ))
```

```
factorial=1
```

```
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)
```

12. *# taking input from user*

```
number = int(input("Enter any number: "))
```

```
# prime number is always greater than 1
```

```
if number > 1:
```

```
    for i in range(2, number):
```

```
        if (number % i) == 0:
```

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

```
            break
```

```
    else:
```

```
        print(number, "is a prime number")
```

```
# if the entered number is less than or equal to 1
```

```
# then it is not prime number
```

```
else:
```

```

    print(number, "is not a prime number")
13.number1 = int(input("The given number is: "))
    x=number1
    rev=0

    while(number1 > 0):
        a = number1 % 10
        rev = rev * 10 + a
        number1 = number1 // 10
    if (x==rev):
        print("The number is a plaine")
    else:
        print("The number is not a plaine")
14.def pythagoras(opposite_side,adjacent_side,hypotenuse):
    if opposite_side == str("x"):
        return ("Opposite = " + str(((hypotenuse**2) -
(adjacent_side**2))**0.5))
    elif adjacent_side == str("x"):
        return ("Adjacent = " + str(((hypotenuse**2) -
(opposite_side**2))**0.5))
    elif hypotenuse == str("x"):
        return ("Hypotenuse = " + str(((opposite_side**2)
(adjacent_side**2))**0.5))
    else:
        return "You know the answer!"

print(pythagoras(3,4,'x'))
print(pythagoras(3,'x',5))
print(pythagoras('x',4,5))
print(pythagoras(3,4,5))
15.test_str = "programmer_muskan"
all_freq = {}

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

print ("Count of all characters in programmer_muskan is :\n " + str(all_))

```

## Worksheet-3:

1. a) The probability of rejecting  $H_0$  when  $H_1$  is true
2. b) null hypothesis
3. d) Type I error
4. b) the t distribution with  $n - 1$  degrees of freedom
5. a) accepting  $H_0$  when it is false
6. d) a two-tailed test
7. b) the probability of committing a Type I error
8. a) the probability of committing a Type II error
9. a)  $z > z_\alpha$
10. c) the level of significance
11. a) level of significance
12. d) All of the Above
13. 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. Examples of when you might want to test different groups:

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- Students from different colleges take the same exam. You want to see if one college outperforms the other.

What Does "One-Way" or "Two-Way Mean?"

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test.

- One-way has one independent variable (with 2 levels). For example: *brand of cereal*,
- Two-way has two independent variables (it can have multiple levels). For example: *brand of cereal, calories*.

What are "Groups" or "Levels"?

Groups or levels are different groups within the same independent variable. In the above example, your levels for "brand of cereal" might be Lucky Charms,

Raisin Bran, Cornflakes — a total of three levels. Your levels for “Calories” might be: sweetened, unsweetened — a total of two levels.

Let’s say you are studying if an alcoholic support group and individual counseling combined is the most effective treatment for lowering alcohol consumption. You might split the study participants into three groups or levels:

- Medication only,
- Medication and counseling,
- Counseling only.

Your dependent variable would be the number of alcoholic beverages consumed per day.

If your groups or levels have a hierarchical structure (each level has unique subgroups), then use a nested ANOVA for the analysis.

What Does “Replication” Mean?

It’s whether you are replicating (i.e. duplicating) your test(s) with multiple groups. With a two way ANOVA *with replication*, you have two groups and individuals within that group are doing more than one thing (i.e. two groups of students from two colleges taking two tests). If you only have one group taking two tests, you would use without replication.

Types of Tests.

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

- One-way ANOVA between groups: used when you want to test two groups to see if there’s a difference between them.
- Two way ANOVA without replication: used when you have one group and you’re double-testing that same group. For example, you’re testing one set of individuals before and after they take a medication to see if it works or not.
- Two way ANOVA with replication: Two groups, and the members of those groups are doing more than one thing. For example, two groups of patients from different hospitals trying two different therapies.

14. The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:

- An ANOVA can only be conducted if there is no relationship between the subjects in each sample. This means that subjects in the first group cannot also be in the second group (e.g. independent samples/between-groups).
- The different groups/levels must have equal sample sizes.
- An ANOVA can only be conducted if the dependent variable is normally distributed, so that the middle scores are most frequent and extreme scores are least frequent.
- Population variances must be equal (i.e. homoscedastic). Homogeneity of variance means that the deviation of scores (measured by the range or standard deviation for example) is similar between populations.

15. The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

- One-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.
- Two-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master's), and race finishing times in a marathon.

All ANOVAs are designed to test for differences among three or more groups. If you are only testing for a difference between two groups, use a t-test instead.



## **SET-2**

### **Worksheet-1:**

1. C) They are not optimal to use in case of outliers.
2. A) It is the most optimal classifier in a completely linearly separable data.
3. D) They can be used in case data is not completely linearly separable.
4. C) If the data is not linearly separable SVM technique cannot be used.
5. A) These functions gives value of the dot product of pairs of data-points in the desired higher. dimensional space without even explicitly converting the whole data in to higher dimensional space.
6. C) It is a model trained using supervised learning. It can be used for classification and regression.
7. D) All of the above
8. C) The data is noisy and contains overlapping points.
9. A) Misclassification would happen.
- 10.B) How accurately the SVM can predict outcomes for unseen data.

### **Worksheet-2:**

1. B) struct
2. C) 1\_no
3. A) in
4. A) Left to right
5. C) iv-iii-ii-i
6. C) 0.3333....
7. B) str
8. A) Division and multiplication have same precedence in python  
B) Python's operators' precedence is based on PEDMAS
9. A) abc = 1,000,000 C) a,b,c = 1000, 2000, 3000 D) a\_b\_c = 1,000,000
10. A) x\*\*4\*\*4
11. All of these are elements used in the Python language, but there is a fundamental difference between list, tuple, set, and dictionary in Python:

## What is a List?

A list is basically like a dynamically sized array that gets declared in other languages (Arraylist in the case of Java, vector in the case of C++). The lists don't always need to be homogeneous in nature. Thus, it is one of the most powerful tools used in the Python language.

## What is a Tuple?

The tuples refer to the collections of various objects of Python separated by commas between them. In some ways, the tuples are similar to the lists in terms of repetition, nested objects, and indexing. The difference is that a tuple, unlike a list, is immutable. The lists, on the other hand, are mutable.

## What is a Set?

The sets are an unordered collection of data types. These are mutable, iterable, and do not consist of any duplicate elements. The set class in Python represents the set's mathematical notion.

## What is a Dictionary?

In Python, the dictionary refers to a collection (unordered) of various data types. We use these for storing data values such as maps, and unlike other data types capable of holding only one value in the form of an element, a dictionary can hold the **key: value** pair. The key value in a dictionary is present to make it comparatively more optimized.

## List Vs Set Vs Dictionary Vs Tuple

Lists	Sets	Dictionaries	Tuples
List = [10, 12, 15]	Set = {1, 23, 34} Print(set) -> {1, 23, 24} Set = {1, 1} print(set) -> {1}	Dict = {"Ram": 26, "mary": 24}	Words = ("spam", "eggs") Or Words = "spam", "eggs"
Access: print(list[0])	Print(set). Set elements can't be indexed.	print(dict["ram"])	Print(words[0])
Can contains duplicate elements	Can't contain duplicate elements. Faster compared to Lists	Can't contain duplicate keys, but can contain duplicate values	Can contains duplicate elements. Faster compared to Lists
List[0] = 100	set.add(7)	Dict["Ram"] = 27	Words[0] = "care" -> TypeError
Mutable	Mutable	Mutable	Immutable - Values can't be changed once assigned
List = []	Set = set()	Dict = {}	Words = ()
Slicing can be done print(list[1:2]) -> [12]	Slicing: Not done.	Slicing: Not done	Slicing can also be done on tuples
<u>Usage:</u> Use lists if you have a collection of data that doesn't need random access. Use lists when you need a simple, iterable collection that is modified frequently.	<u>Usage:</u> - Membership testing and the elimination of duplicate entries. - when you need uniqueness for the elements.	<u>Usage:</u> - When you need a logical association b/w key:value pair. - when you need fast lookup for your data, based on a custom key. - when your data is being constantly modified.	<u>Usage:</u> Use tuples when your data cannot change. A tuple is used in combination with a dictionary, for example, a tuple might represent a key, because its immutable.

12. Strings are array in python, any character in python is string. Strings are immutable just like an array & tuples. Immutability is a clean and efficient solution to concurrent access. Having immutable variables means that no matter how many times the method is called with the same variable / value, the output will always be the same.

Immutable strings greatly simplify memory allocation when compared with C strings, you don't guess at a length and over-allocate hoping you over-allocated enough. The elements of the strings can be accessed by slicing and indexing, elements can be replaced using `replace( )` method, it generates a copy of the string with new replaced items.

Advantage:-

We also knowing that a string is immutable means we can allocate space for it at creation time, nad the storage requirements are fixed and unchanging. String in python are considered as "elemental" as numbers, No amount of activity will change the value 8 to anything else, and in python, no amount of activity will change the value 8 to anything else, and in Python no amount of activity will change the string "eight" to anything else.

Example:-

In [1]:

```
a = "I+Love+Python"           # defining a
new_a = a.replace("+"," ")     # replacing + with space(" ")
new_a                          # printing final output
```

13. It is inbuilt `ord( )` function in Python, given string of length one, return an integer representing the Unicode code point of the cahracter when the argument is a unicode object, or the value of the byte when the argument is an 8 bit string.

Example :- `ord("a")` return the integer 97, `ord("$")` returns 36. This is the inverse of `chr( )` for 8 bit strings and of `unichr( )`for unicode objects. If a unicode argument is given and python was built with UCS2 unicode, then the character's code point must be in the range[0...65535] inclusive.

Type( ) Function :- It is very simple syntax and can be used to find the type of any variable in python be it a collection type variable, a class object variable or a simple string or integer.

```
syntax:- type(variable name)
Example : - a=12      # defining variable a
type(a)      # printing type variable
```

14. We are going to solve the Quadratic Equation ( $ax^2 + bx + c = 0$ ),  
coefficient & constant values taken by the user input

Before we start, first we import some important library

```
import math
```

a, b, c are constant & x is a variable we have to calculate the value of x from  
finding the value of Discriminant

From discriminant ( $disc = b^2 - 4ac$ ) we will get 2 root values, in 3 conditions

$disc = 0$  = Roots are real and equal

$disc > 0$  = Roots are real & unequal

$disc < 0$  = Roots are imaginary

Roots calculation  $r1 = ((-b) + (\text{math.sqrt}(disc))) / 2a$  First root

$r2 = ((-b) - (\text{math.sqrt}(disc))) / 2a$  Second root

```
run = "yes"
```

```
while run=="yes":
```

```
    a = int(input("Enter the coefficient of x square : "))
```

```
    b = int(input("Enter the coefficient of x : "))
```

```
    c = int(input("Enter constant : "))
```

```
    disc = b*b - 4*a*c
```

```
    print("Discriminant = ", disc)
```

```
    if disc == 0:
```

```
        print("Roots are Real & Equal")
```

```
        r1 = (-b)/(2*a)  # in this disc is 0, so math.sqrt(disc) will be zero
```

```
        r2 = (-b)/(2*a)
```

```
        print("First Root :", r1)
```

```
        print("Second Root :", r2)
```

```
    elif disc > 0:
```

```
        print("Roots are Real & Unequal")
```

```
        r1 = ((-b) + (math.sqrt(disc)))/(2*a)
```

```
        r2 = ((-b) - (math.sqrt(disc)))/(2*a)
```

```
        print("First Root :", r1)
```

```
        print("Second Root :", r2)
```

```
    else:
```

```
        print("Roots are Imaginary")
```

```
        run = str(input("Type yes to return or Press any key to exit:")).casefold())
```

exit()

Output- Enter the coefficient of x square : 12

Enter the coefficient of x : 4

Enter constant : 7

Discriminant = -320

Roots are Imaginary

Type yes to return or Press any key to exit:yes

Enter the coefficient of x square : 1

Enter the coefficient of x : -5

Enter constant : 6

Discriminant = 1

Roots are Real & Unequal

First Root : 3.0

Second Root : 2.0

Enter the coefficient of x square : 1

Enter the coefficient of x : -2

Enter constant : 1

Discriminant = 0

Roots are Real & Equal

First Root : 1.0

Second Root : 1.0

After giving input values to quadratic equation, we can see above result showing all the 3 conditions with 2 roots in each case. We put yes for looping / to see again with new values, now it will continuously taking input values from user.

15. We are going to sum & average of "n" natural number, value of "n" will be taken by user without using any loop

```
n = int(input("Enter a number : "))
```

```
x = list(range(n))
```

```
add = sum(x)+n
```

```
# defining average
```

```
avg = add/n
```

```
# printing both sum(add) & average
```

```
print(add)
```

```
print(avg)
```

Output: Enter a number : 50

1275

25.5

## Worksheet-3:

1. C) Type I; Type II
2. B) We have made a correct decision
3. B) critical value
4. B) A Type I error was made.
5. C)  $\bar{x} = 17$ ,  $s = 7$
6. A) fail to reject  $H_0$
7. C) At  $\alpha = 0.05$ , reject the null hypothesis.
8. B) 0.041
9. C) 0.958
10. C) Left tail
11. A) Less than the significance level
12. B) 0.375
13. The  $t$ -distribution, also known as Student's  $t$ -distribution, is a way of describing data that follow a bell curve when plotted on a graph, with the greatest number of observations close to the mean and fewer observations in the tails.

It is a type of normal distribution used for smaller sample sizes, where the variance in the data is unknown.

In statistics, the  $t$ -distribution is most often used to:

- Find the critical values for a confidence interval when the data is approximately normally distributed.
- Find the corresponding  $p$ -value from a statistical test that uses the  $t$ -distribution ( $t$ -tests, regression analysis).

The  $t$ -distribution is a type of normal distribution that is used for smaller sample sizes. Normally-distributed data form a bell shape when plotted on a graph, with more observations near the mean and fewer observations in the tails.

The  $t$ -distribution is used when data are *approximately* normally distributed, which means the data follow a bell shape but the population variance is unknown. The variance in a  $t$ -distribution is estimated based on the degrees of freedom of the data set (total number of observations minus 1).

It is a more conservative form of the standard normal distribution, also known as the  $z$ -distribution. This means that it gives a lower probability to the center and a higher probability to the tails than the standard normal distribution.

Example: *t*-distribution vs *z*-distribution If you measure the average test score from a sample of only 20 students, you should use the *t*-distribution to estimate the confidence interval around the mean. If you use the *z*-distribution, your confidence interval will be artificially precise.

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.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD) of 1

It is useful to standardized the values (raw scores) of a normal distribution by converting them into *z*-scores because:

- (a) it allows researchers to calculate the probability of a score occurring within a standard normal distribution;
- (b) and enables us to compare two scores that are from different samples (which may have different means and standard deviations).

The formula for calculating a *z*-score is  $z = (x - \mu) / \sigma$ , where *x* is the raw score,  $\mu$  is the population mean, and  $\sigma$  is the population standard deviation.

As the formula shows, the *z*-score is simply the raw score minus the population mean, divided by the population standard deviation.

14. The *t*-distribution is a type of normal distribution that is used for smaller sample sizes. Normally-distributed data form a bell shape when plotted on a graph, with more observations near the mean and fewer observations in the tails.

The *t*-distribution is used when data are *approximately* normally distributed, which means the data follow a bell shape but the population variance is unknown. The variance in a *t*-distribution is estimated based on the degrees of freedom of the data set (total number of observations minus 1).

15. A *t*-score is the number of standard deviations from the mean in a *t*-distribution. You can typically look up a *t*-score in a *t*-table, or by using an online *t*-score calculator.

In statistics, *t*-scores are primarily used to find two things:

1. The upper and lower bounds of a confidence interval when the data are approximately normally distributed.
2. The  $p$ -value of the test statistic for  $t$ -tests and regression tests.



## SET-3

### Worksheet-1:

1. C) y intercept
2. A) True
3. B) the dependent variable
4. B) Linear Regression
5. C) the correlation coefficient squared
6. B) y increases as x increases
7. C) both linear and non-linear data
8. B) -1 to 1
9. B) RMSE D) MAE
10. A) Linear regression is a supervised learning algorithm. B) Linear regression supports multi-collinearity.
11. A) Ridge B) Lasso
12. A) Large amount of training samples with small number of features. D) The variables which are drawn independently, identically distributed
13. A) Linearity B) Homoscedasticity D) Normality
14. Linear regression is an algorithm used to predict, or visualize, a relationship between two different features/variables. In linear regression tasks, there are two kinds of variables being examined: the dependent variable and the independent variable. The independent variable is the variable that stands by itself, not impacted by the other variable. As the independent variable is adjusted, the levels of the dependent variable will fluctuate. The dependent variable is the variable that is being studied, and it is what the regression model solves for/attempts to predict. In linear regression tasks, every observation/instance is comprised of both the dependent variable value and the independent variable value.

The function of a regression model is to determine a linear function between the X and Y variables that best describes the relationship between the two variables. In linear regression, it's assumed that Y can be calculated from some combination of the input variables. The relationship between the input variables (X) and the target variables (Y) can be portrayed by drawing a line through the points in the graph. The line represents the function that best describes the relationship between X and Y (for example, for every time X increases by 3, Y increases by 2). The goal is to find an optimal "regression line", or the line/function that best fits the data.

Lines are typically represented by the equation:  $Y = m \cdot X + b$ .  $X$  refers to the independent variable while  $Y$  is the dependent variable. Meanwhile,  $m$  is the slope of the line, as defined by the “rise” over the “run”. Machine learning practitioners represent the famous slope-line equation a little differently, using this equation instead:

$$y(x) = w_0 + w_1 \cdot x$$

In the above equation,  $y$  is the target variable while “ $w$ ” is the model’s parameters and the input is “ $x$ ”. So the equation is read as: “The function that gives  $Y$ , depending on  $X$ , is equal to the parameters of the model multiplied by the features”. The parameters of the model are adjusted during training to get the best-fit regression line.

15. Regression analysis is a common statistical method used in finance and investing. Linear regression is one of the most common techniques of regression analysis. Multiple regression is a broader class of regressions that encompasses linear and nonlinear regressions with multiple explanatory variables.

Regression as a tool helps pool data together to help people and companies make informed decisions. There are different variables at play in regression, including a dependent variable—the main variable that you’re trying to understand—and an independent variable—factors that may have an impact on the dependent variable.

## Linear Regression

Also called simple regression, linear regression establishes the relationship between two variables. Linear regression is graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is 0.

In linear regression, every dependent value has a single corresponding independent variable that drives its value. For example, in the linear regression formula of  $y = 3x + 7$ , there is only one possible outcome of ‘ $y$ ’ if ‘ $x$ ’ is defined as 2.

If the relationship between two variables does not follow a straight line, nonlinear regression may be used instead. Linear and nonlinear regression are similar in that both track a particular response from a set of variables. As the relationship between the variables becomes more complex, nonlinear models have greater flexibility and capability of depicting the non-constant slope.

## Multiple Regression

For complex connections between data, the relationship might be explained by more than one variable. In this case, an analyst uses multiple regression which attempts to explain a dependent variable using more than one independent variable.

There are two main uses for multiple regression analysis. The first is to determine the dependent variable based on multiple independent variables. For example, you may be interested in determining what a crop yield will be based on temperature, rainfall, and other independent variables. The second is to determine how strong the relationship is between each variable. For example, you may be interested in knowing how a crop yield will change if rainfall increases or the temperature decreases.

Multiple regression assumes there is not a strong relationship between each independent variable. It also assumes there is a correlation between each independent variable and the single dependent variable. Each of these relationships is weighted to ensure more impactful independent variables drive the dependent value by adding a unique regression coefficient to each independent variable.

A company can not only use regression analysis to understand certain situations, like why customer service calls are dropping, but also to make forward-looking predictions, like sales figures in the future.

## Linear Regression vs. Multiple Regression Example

Consider an analyst who wishes to establish a relationship between the daily change in a company's stock prices and the daily change in trading volume. Using linear regression, the analyst can attempt to determine the relationship between the two variables:

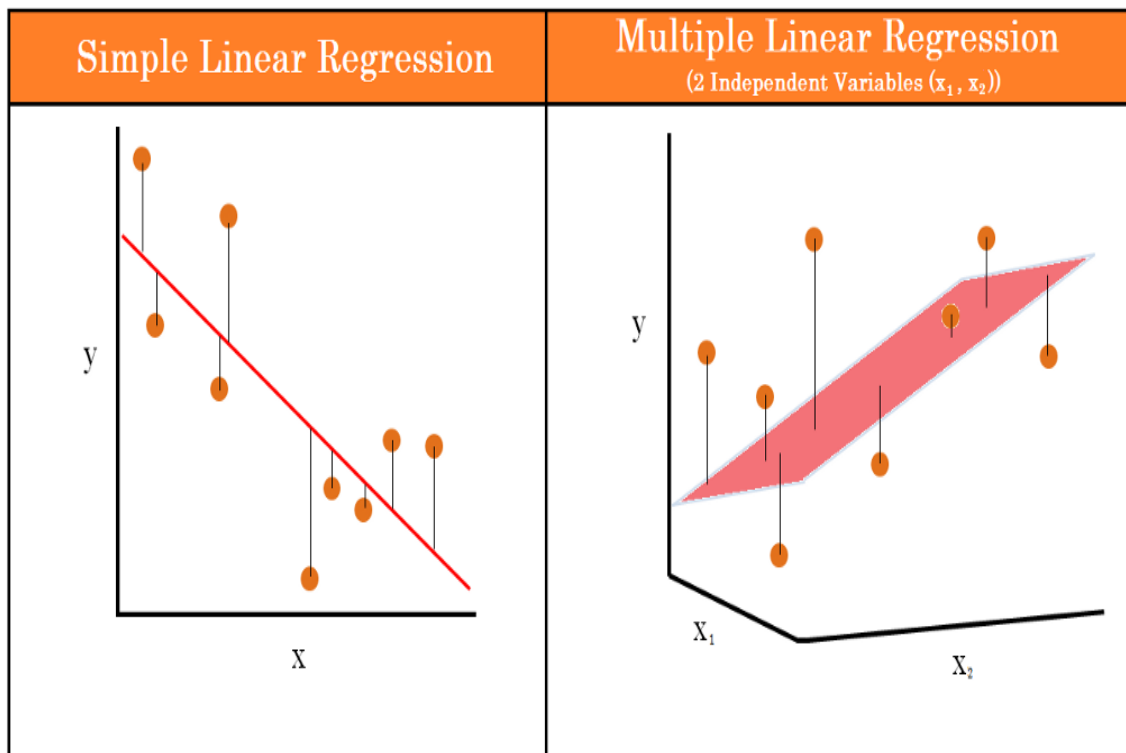
Daily Change in Stock Price = (Coefficient)(Daily Change in Trading Volume) + (y-intercept)

If the stock price increases \$0.10 before any trades occur and increases \$0.01 for every share sold, the linear regression outcome is:

Daily Change in Stock Price = (\$0.01)(Daily Change in Trading Volume) + \$0.10

However, the analyst realizes there are several other factors to consider including the company's P/E ratio, dividends, and prevailing inflation rate. The analyst can perform multiple regression to determine which—and how strongly—each of these variables impacts the stock price:

Daily Change in Stock Price = (Coefficient)(Daily Change in Trading Volume) + (Coefficient)(Company's P/E Ratio) + (Coefficient)(Dividend) + (Coefficient)(Inflation Rate)



## Worksheet-2:

1. D) int ('32')
2. C) 4
3. B) (a\*\*b)\*c
4. A) <class 'type'>
5. C) 65
6. D) Method
7. B) False
8. B) Sometimes
9. ALL A,B,C,D
10. B) You can pass keyword arguments in any order.  
C) You can call a function with positional and keyword arguments.  
D) Positional arguments must be before keyword arguments in a function call

11. num = int(input("Enter the Number :- "))

**for i in range(0,num):**

**for j in range(0,num-i-1):**

*print(end=" ") # if we remove space from bracket then all stars will start from left side following margin line*

**for j in range(0,i+1):**

*print("\*",end=" ")*

*print()*

Output: Enter the Number :- 8

```
*
* *
* * *
* * * *
* * * * *
* * * * * *
* * * * * * *
```

12. num = int(input("Enter the Number :- "))

**for i in range(num,0,-1):**

**for j in range(num,i,-1):**

*print(end=" ")*

**for k in range(0,i):**

*print("\*",end=" ")*

*print()*

**for i in range (1,num):**

**for j in range(0,num-i-1):**

*print(end=" ")*

```

    for k in range(0,i+1):
        print("*",end=" ")
    print()
Output: Enter the Number :- 8

```

```

* * * * *
* * * * *
* * * * *
* * * * *
* * * *
* * * *
* * *
* *
*
* *
* * *
* * * *
* * * * *
* * * * *
* * * * *
* * * * *

```

```

13.def fact(n):
    f=1
    if n==0:
        return 1
    for i in range(2,n+1):
        f=f*i
    return f
def comb(m,n):
    res=fact(m)//(fact(m-n)*fact(n))
    return res
num = int(input("Enter the Number :-"))
for i in range(0,num):
    for j in range(0,i+1):
        print(comb(i,j)," ",end=" ")
    print()

```

```

Output: Enter the Number :-8
1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1

```

1 7 21 35 35 21 7 1

```
14.num = int(input("Enter the Number :- "))
```

```
    for i in range (0,num):
        for s in range(0,num-i-1):
            print(end=" ")
        for j in range(0,i+1):
            print("* ",end="")
        print()
    for i in range(num-1,0,-1):
        for s in range(num,i,-1):
            print(end=" ")
        for j in range(0,i):
            print("* ",end="")
        print()
```

Output: Enter the Number :- 8

```
    *
  * *
* * *
* * * *
* * * * *
* * * * * *
* * * * * * *
* * * * * * *
* * * * * *
* * * * *
* * * *
* * *
* *
*
```

```
15.n = int(input("Enter the Number :- "))
```

```
    for i in range(n):
        print(' *(n-i-1),end=" ")
        for j in range(i+1):
            print(chr(65+j),end=" ")
        print()
    for i in range(n-1):
        print(' *i,end=" ")
        for j in range(n-i-1):
            print(chr(65+j),end=" ")
```

```
print()
```

Output: Enter the Number :- 8

A

A B

A B C

A B C D

A B C D E

A B C D E F

A B C D E F G

A B C D E F G H

A B C D E F G

A B C D E F

A B C D E

A B C D

A B C

A B

A



## Worksheet-3:

1. C) Neither
2. B) The underlying distribution
3. A) True
4. B) We are 95% confident that the result have not occurred by chance
5. C) If the region of rejection is located in one or two tails of the distribution
6. C) We accept the null hypothesis when it is not true
7. A) It is a sample proportion.
8. A) 0.13
9. C) 1.667
10. B) -2.50
11. C) There is a difference between the proportions of American men and American women who belong to sports clubs.
12. B) It is reasonable to say that more than 40% of Americans exercise regularly.
13. The test statistic is a number calculated from a statistical test of a hypothesis. It shows how closely your observed data match the distribution expected under the null hypothesis of that statistical test.

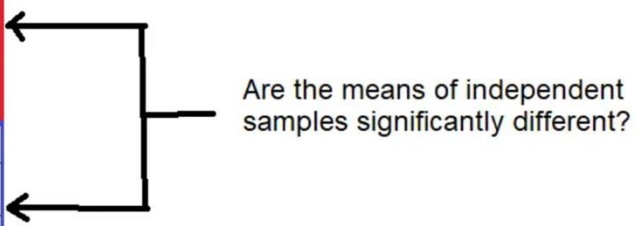
The test statistic is used to calculate the  $p$  value of your results, helping to decide whether to reject your null hypothesis.

A test statistic describes how closely the distribution of your data matches the distribution predicted under the null hypothesis of the statistical test you are using.

The distribution of data is how often each observation occurs, and can be described by its central tendency and variation around that central tendency. Different statistical tests predict different types of distributions, so it's important to choose the right statistical test for your hypothesis.

The independent samples T-test is defined as statistical hypothesis testing technique in which the samples from two independent groups are compared to determine if the means of the associated populations are significantly different. For example, let's say that we have two independent groups of male and female and we want to compare their income to determine whether their income is significantly different. Independent samples t-test is also called unpaired two-samples t-test because the test is performed with only two groups that are independent or unpaired or unrelated. The picture below shows the representation of two independent samples and the aspect of their means.

Id	Gender	Mathematics Marks
1	Male	98
2	Male	92
3	Male	89
4	Male	75
5	Female	83
6	Female	92
7	Female	85
8	Female	99



Are the means of independent samples significantly different?

The independent samples T-test is used when the two samples are independent and have normal distributions. The mandatory requirement is that you need to have two independent samples. The independent samples mean that *the two samples cannot be from the same group of people and they cannot be related in any way*. Another assumption for independent samples t-test is homogeneity of variance of the two groups. The two-sample T-test is used when the standard deviations of the populations to be compared are unknown and the sample size is small. The size of sample 30 or less is considered as small sample. That said, the size of the sample is not a strict condition for using T-test. However, two-sample T-test can also be used for pairwise comparisons when the “two” samples represent the same items tested in different scenarios.

The following are a few real-life examples where two-sample T-test for independent samples can be used:

- Comparing the average test scores of two classes from two different schools
- Comparing the average weights of two different groups of people
- Measuring the difference in height between men and women.
- Checking if work experience impacts job satisfaction.
- Examining if there is any change in productivity levels after introducing a new system at work.
- Evaluating if there is a difference in math scores between boys and girls.
- Checking if there is an improvement in productivity due to the implementation of new software.
- Assessing if there is a difference in blood pressure levels between patients taking medication and those not taking medications.

14. The mean difference (more correctly, 'difference in means') is a standard statistic that measures the absolute difference between the mean value in two groups in a clinical trial. It estimates the amount by which the experimental intervention changes the outcome on average compared with the control.

Formula

$$\text{Mean Difference} = \sum x_1/n - \sum x_2/n$$

Where –

- $x_1$  = Mean of group one
- $x_2$  = Mean of group two
- $n$  = Sample size

Example

Problem Statement:

There are 2 dance groups whose data is listed below. Find the mean difference between these dance groups.

Group 1	3	9	5	7
Group 2	5	3	4	4

Solution:

$$\sum x_1 = 3 + 9 + 5 + 7 = 24$$

$$\sum x_2 = 5 + 3 + 4 + 4 = 16$$

$$M_1 = \sum x_1/n = 24/4 = 6$$

$$M_2 = \sum x_2/n = 16/4 = 4$$

$$\text{Mean Difference} = 6 - 4 = 2$$

15. A two sample t-test is used to determine whether or not two population means are equal.

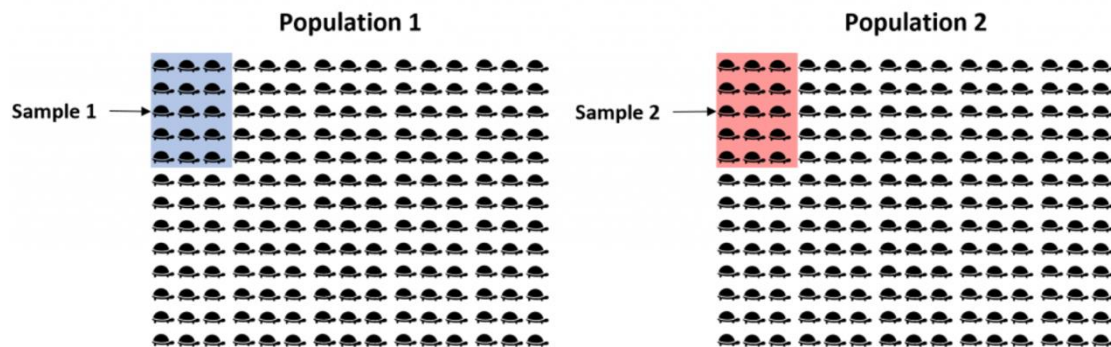
Here, we explain the following:

- The motivation for performing a two sample t-test.
- The formula to perform a two sample t-test.
- The assumptions that should be met to perform a two sample t-test.
- An example of how to perform a two sample t-test.

Two Sample t-test: Motivation

Suppose we want to know whether or not the mean weight between two different species of turtles is equal. Since there are thousands of turtles in each population, it would be too time-consuming and costly to go around and weigh each individual turtle.

Instead, we might take a simple random sample of 15 turtles from each population and use the mean weight in each sample to determine if the mean weight is equal between the two populations:



However, it's virtually guaranteed that the mean weight between the two samples will be at least a little different. The question is whether or not this difference is statistically significant. Fortunately, a two sample t-test allows us to answer this question.

### Two Sample t-test: Formula

A two-sample t-test always uses the following null hypothesis:

- $H_0: \mu_1 = \mu_2$  (the two population means are equal)

The alternative hypothesis can be either two-tailed, left-tailed, or right-tailed:

- $H_1$  (two-tailed):  $\mu_1 \neq \mu_2$  (the two population means are not equal)
- $H_1$  (left-tailed):  $\mu_1 < \mu_2$  (population 1 mean is less than population 2 mean)
- $H_1$  (right-tailed):  $\mu_1 > \mu_2$  (population 1 mean is greater than population 2 mean)

We use the following formula to calculate the test statistic t:

Test statistic:  $(\bar{x}_1 - \bar{x}_2) / s_p(\sqrt{1/n_1 + 1/n_2})$

where  $\bar{x}_1$  and  $\bar{x}_2$  are the sample means,  $n_1$  and  $n_2$  are the sample sizes, and where  $s_p$  is calculated as:

$$s_p = \sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2] / (n_1+n_2-2)}$$

where  $s_1^2$  and  $s_2^2$  are the sample variances.

If the p-value that corresponds to the test statistic  $t$  with  $(n_1+n_2-1)$  degrees of freedom is less than your chosen significance level (common choices are 0.10, 0.05, and 0.01) then you can reject the null hypothesis.

### Two Sample t-test: Assumptions

For the results of a two sample t-test to be valid, the following assumptions should be met:

- The observations in one sample should be independent of the observations in the other sample.
- The data should be approximately normally distributed.
- The two samples should have approximately the same variance. If this assumption is not met, you should instead perform Welch's t-test.
- The data in both samples was obtained using a random sampling method.