**Data science Course**

**Day 1 Live class 4th Nov 23**

**General information session**

Data science Foundation- (Excel and Basic stats)- 8 sessions

Programing foundation- 4 sessions

SQL (Programing language)- (MySQL- Database)- 6-8 sessions

Python programing- 7-8 sessions

Data visualization- 4 sessions

Statistical Analysis- 3-4 sessions

Machine Learning- 15-20 session

**Specializations-**

NLP with ML-: Pre- processing of data of text data

Deep learning

Deep NLP

Advanced Excel

R & R Shiny

Tableu

Power BI

It can only be mastered by practicing them regularly, from day 1 once one concept is covered.

**Addition language**- (200+ Programming Language)

Manually in Calc, Excel, Python, MySQL (DB), C++, C#, Ruby on Rails, Java script, Java, Matlab,

* ML is a brain of AI
* Java can also be used for ML, not just Python

First job you can apply after this Course

* Data Analyst
* Product Analyst (ML)
* Business Intelligence Analyst
* Analyst- (Service based Org., Consultant type job roles, handle clients)
* Data Engineers- Gateway to ML Engineers jobs
* Data Scientist- Need all skills- Python, sql, ML, cleaning, maths, pre-processing

Specialized Jobs-

* ML Engineer
* NLP Engineer
* SQL Developer
* Database Administrator – Create and maintain the data base
* Tableu Developer
* Power BI Developer

Portfolio should contain uncommon projects which makes you different

* For Data analyst:- the projects you are good at like 2-3 projects on SQL, 3-4 Python projects. Eg. Accident analyst, Crime analysts
* Data scientist- 5 ML projects, 2-3 NLP, 3-4 DL

Recommendations:-

Reference Books to follow for course:-

Python:-

* Learning Python- O rielly publications
* Core Python programming- Dr R Nageshwara Rao

SQL books:-

* Learning SQL- Orielly publ.

Statistics:-

* Think Stats- Orielly Publications
* How to lie with statistics

Machine Learning:-

Hands on Machine Learning with Scikit-learn, Keras and tensorflow

Online resources:-**For Learning**

* Towards Data Science – search Towards Data Science statistics, Python or regression
* Medium – same search as above – **Its paid**

**For errors**

* Stack overflow for solution
* Gist from GitHub
* Stack Exchange

Resources (ML)

* Kaggle- Ready made data sets, Frame your own data and get the solution- Make your profile for learning and participate in completions
* Mention your Kaggle platform in resume

**For Practice:- IMP**

* Codechef.com

Hackerrank

Hacker earth

Leet code

**Challenge IMP**

Solve 1000 problems total in Python in all above sites

500 in SQL

**Day 2 Live class 5th Nov 23**

Watch same recorded videos multiple times and practice all the concepts.

Maths and Statistics is a core foundation of ML and data science, so keep learning it.

Basic Excel- Explained all tabs by tutor

Sum- Add the values- Can add colan or coma in formula

Subtraction- Subtracting the values- no formuala, Manually select cell with coma

Multiplication- multiply values – Use product formula and put coma between cell . Even use range with colan :

Division- Divide the values- no formula, Manually select cell with coma

* To select cells you can select colan(:), like B3:E3 so B3, C3,D3 and E3 selected, like B2:D5, all between selected to sum
* To save the storage in excel (& to solve error), instead of same reference, you can us as a fix (flash fill) reference in one cell and use in formula. Use Fn and F4 in laptop to fix the cell.

**Day 3 Live class 11th Nov 23**

BOD MAS-Brackets, Off/Order, Division, Multiplication, Addition, Subtraction

For scholarship we will use If formula, Apply if then set a true and false condition. =If(B2>75, “Yes”, “No”)

**19th Nov 2023**

How to download data from github- Go through preparation video preloaded

<https://github.com/datatrained123/DS1023--> with HW

=Countif formula for full column of dept. and get sum count

=Countifs for multiple columns of dept and get sum count =countifs(F:F, “male”, “country”)

=average sum is mean of data which is E symbol formula

=averageif function to select one column then another to find average age or salary hike, we cannot use uniq function inside avgif, .. Mention aveg range at the end

=averageifs is for a multiple conditions.. average range will be at the start

Watch pre recorded before always joining session

**25th Nov classes**

**Percentiles:-** Fraction of whole- Comparison b/w competitors or a position hold by competitor .. It validates you scored better than 80% and less than 20% if your percentage is 80%. We can say it’s a rank. If salary percentile is 30th than you are higher than 30% employees and less than 70%

It is he organizer who is declaring the course and not you.

Type of percentile:-

1. Inclusive (Inc) and (Min and max value are included )

2. Exclusive percentile (Exc) (Min and Max value only are not included)

Percentile can be used every where like Excel, Python, Numpy, Panda, Python and many more. It can only define in continuous format

Scaled= X- Xmin/ Xmax- Xmin

Value of k would always be from 0 to 100 which is 0/100, So formula percentile inc(array which is column, k which is 0/100) and same in exc. In K 0 should be selected in formula and not write

**Quartile**- quart means 4th

100 can be divided in to 4 which is 0 to 25, 25 to 50, 50 to 75 and 75 to 100. And we will use formula on 25th, 50th, 75th and 100th column. We will get same value as percentile using quartile inc or exc

Formula= =quantile inc and ec (array of salary, k select a cell with 1 to 100 number)

Formula- =small(array, k) and large(array, k select cell), as k increases or decreases you get next salary higher/lower salary amount

**Mode-** mode is value which maximum times (occurrence) I get in column. Mode will only hold one value.

Mode of data- It will be a smaller value in data while getting mode

Mean, Median and mode give you 3 different info. Combined together all 3 called measure of central tendency.

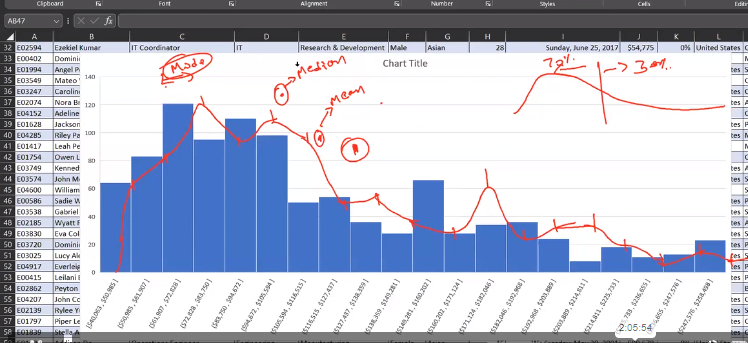
Mean- avg value of data

Median- Middle value which is present at center of data.

Mode- Value which is occurring max number of times.

This is the first step of analysis.

KDE- kernel Density Estimation which identifies the density in each bins of histogram chart (Middle dots in each bins)



Right tailed distribution- Right side of lower arrow

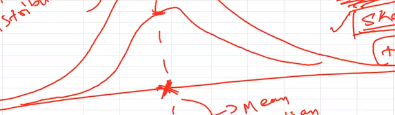
**For Machine learning algorithms** – pre process is to convert into **normal distribution** which is mandate

Moments:- 1. Mean

2. Variance- It will give you figure about distribution of data or the spread of the data.

3. Skewness – It tells you if data is normally distributed or not. Like density in graph is 70% left and 30% on right. It tells **how data is biased**

When we preprate data for ML algorithms, for **normal distribution skewness** should be always 0, if it’s not 0 then it has be pre processed again by **transformation** of data. So, skewness become 0 and normal distribution can be done of data.

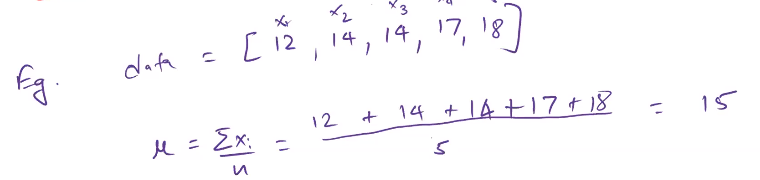
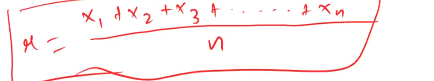
4. Kurtosis- It tells you how high and low the peak of data is. 

5. Co relation- eg. Total number of yrs of experience = High salary = Skill set . It can be identified by co relation

**26th Nov class**

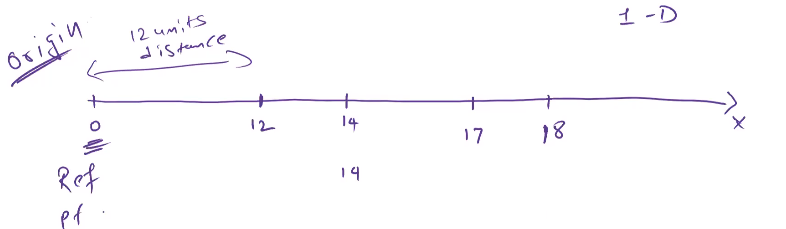
**Moments:--** Measure of statistics

* In Stats, measure anything,
* Quantification in technical term (Represent something as quantity (as a numbers)is quantification )

**1st moment- Mean/Average ,** Formula is μ = Xi/n, i= 1 to n like X1+X2+X3/n total number of qty which is 3 

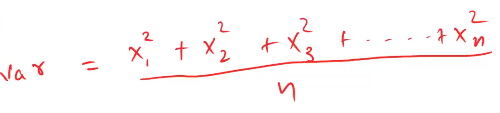
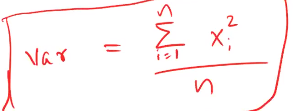
Origin of data is 0. Reference point is always 0

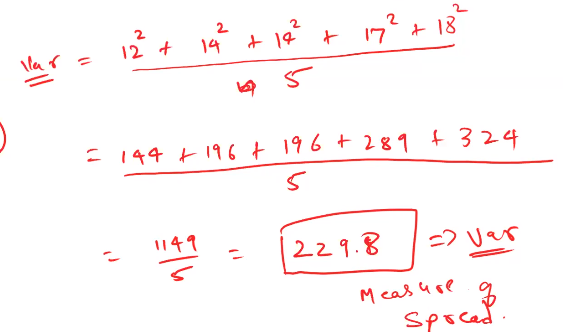
Measure of distance we calculate.



**2nd moment- Variance** (Imp term) -Measure of spread. In ML algorithms spread of data should be v low. During analyzing variance should be v high. Higher the variance better the analysis.

Var= square of Xi /n , i = 1 to n . It is Raw/Crude formula





Eg. Price patterns of Rubik’s cube. You can analyze the difference in features of Cars/Properties.

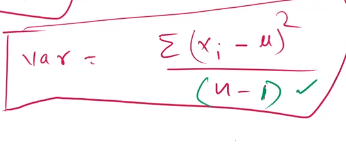
Features=> Columns=> Information. We pass this features in ML model so model will be trained.

Data should follow some set of pre requisite to process ML algorithm. And one **variance of data** is one of the pre requisite. And Variance should be low. Skewness should be 0, so no biasedness in model.

While analyzing the data the variance should be high and while preparing the data for ML algorithms the variance should be low.

Value will increase if mean of data will increase. Mean is derived from element of dataset.

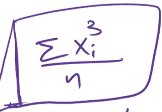
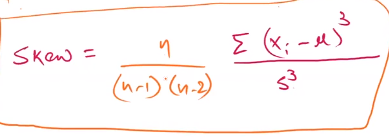
var= Xi square2/n , so Var= (Xi-myu)2 /n-1. You will not apply this formula though, as this will be readymade in all tools.



n-1 is sampling adjustment**. You can study article on towards data science to understand working maths behind it. Give enough time and logic**

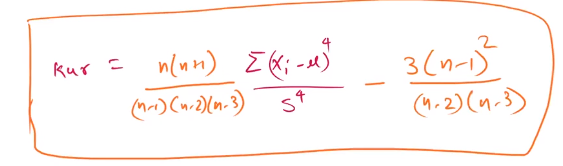
**3rd Moment- Skewness** – Measure of biasedness/Tiltedness

Formua- whole cube to Xi/n

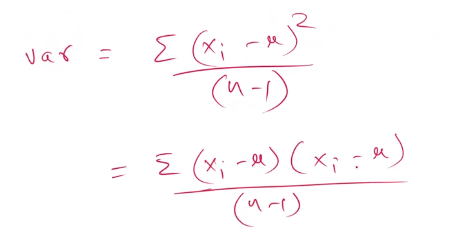
We will also have derived formulas. This is just a raw formula. No one will ask you this formula in interview, they only want to know what insights you have from this data. **Its all game of insights you can get from data.**

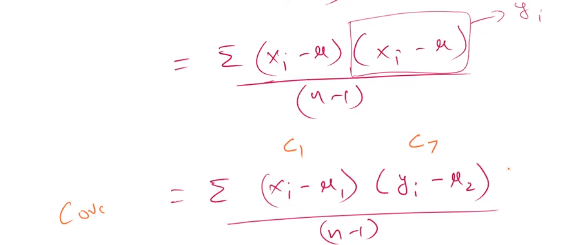
**4th Moment- Kurtosis**- Peakedness. Peak (Highest point) of the data.



This is used in all analytic tools.’ “**Think stats” book you can go through as recommended earlier.**

**Co-variance-**

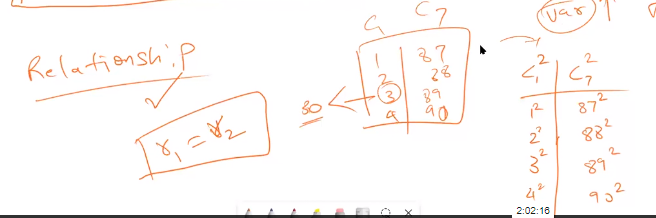


C1 and C2 are columns.

This is not generalized so can not be dependent.

**Corelation** (r) is a generalized form of a covariance.





This correlation will help you to analyze the relationship between features.

Eg. How to measure salary of an employee. r(s, exp)= r1, r(s, age)= r2 , r(s, skill)= r3.

r(s, mob. No.)= 0 – No contribution in decision making.

These are foundation of ML. You will come across these terms everywhere.

**Towards data science website you can go through.**  Search in google- moments/Variance towards data science.

**2nd Dec. class**

**Descriptive statistics**

1. Mean- Average value
2. Variance- Measure of spread 1. Population- Complete dataset, 2. Sample- Subset of a data. Use Formula- =VarS or P.

-Standard deviation- It’s a square root of variance. Formula- =stdev(select salary colum)

1. Skewness- Measure of biasness. Formula =skew(sal column). Skewness can never come to 0. So bring down to as close to 0
2. Kurtosis- Measure of Peakedness . It is hardly used in data. Formula = kur(sal column). Value can come positive or negative

**Inferential statistics**

**Covariance**- Common spread shared y two columns – Subject to change with values.

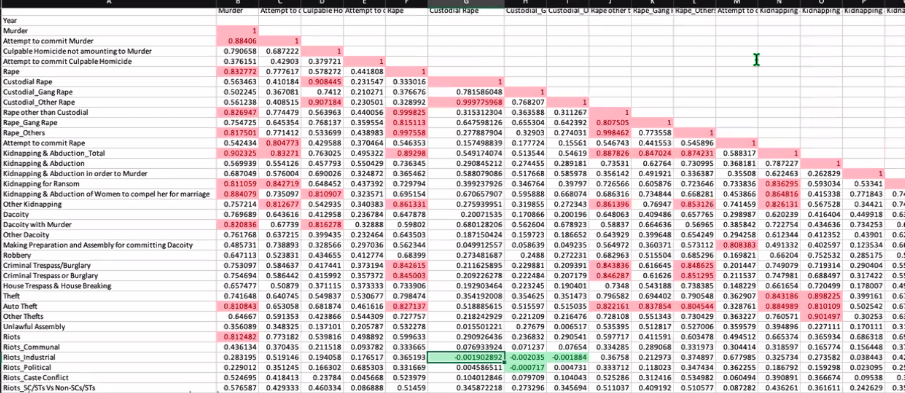
For salary hike,, Formula. =covariance.s(salary colum, hike colum).

Ppl will high salary will get high hike.

**Correlation**- Relation between two features- Not to subject to change. It will be (-1, 1)

Formula =corr(sal, hike), 0.9256 value comes and it is highly positive.

Correlation matrix



Highlight with conditional formatting. Enter 0.8 and you will all highlighted.

**Pivot table-**

Eg. Make a dashboard requirement in accidents

1. 2021-2022
2. Total number of ac involved
3. Severity of the accidents of one type vs others
4. Casualties based on type of vehicles
5. Casualties on road surface
6. Casualties on road type
7. Rural vs urbans
8. Monthly analysis of 2021-22
9. Maximum Casualties by vehicle type

Go to insert, Pivot table, then select as per needs. To combine values select filed, items and click in calculated field and + it to merge to one name

Github link- <https://github.com/datatrained123/DS1023>

**3rd Dec. class**

<https://www.kaggle.com/datasets/jhalls/accident-data>

Select crime data- go to data- data analysis (from Analysis tool), then you get a plotted data.

Go to home- data validation- and color it after putting values

Principal components analysis (PLC)- where co variance is used.

Vlook up and Hlook up

<https://github.com/datatrained123/DS1023> for HW

**Programming Foundation**

**9th Dec**

Computer Is an intelligent Machine.

Machine is only following the instructions which is called AI and it can not do more than that

Eg. Zomato-

Binary language- Computer scientist created voltage for a communication But It was difficult, so they came up with high voltage as 1 and low voltage as 0. So that was Binary language came in to picture.

**Programing languages** are predefined keywords which holds some meaning in order to execute a task, perform any instruction given to a computer.

We use programing languages to talk with computer or to give instructions to a computer. Those instructions are then converted in to binary form and then give the output. Binary is a number system.

Codes:- Specific keyword to give instructions to a computer.

**Data**- Collection of Information which then is stored somewhere permanently.

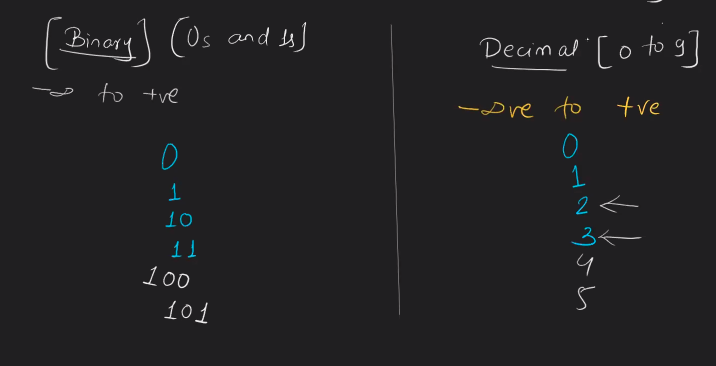
**Variable** is a name of storage location in computer’s memory. eg. x = 10, y = “Aisha”

Two types of memory

1. RAM- Random access memory- It is permanent memory which when we install game.
2. ROM- Read Only memory- It is temporary memory which works while we play game

**Numbers** are from negative infinite to positive infinite. So numbers are not countable

Very decimal numbers can be converted into Binary form



**Data type** is defining the type of data storing in particular variable. 5 types of data

1. Int
2. String
3. Char
4. Float
5. Boolean

**10th Dec- to reapt**

Value of string will be always in double inverted comma.

Int a- 10 which is a number

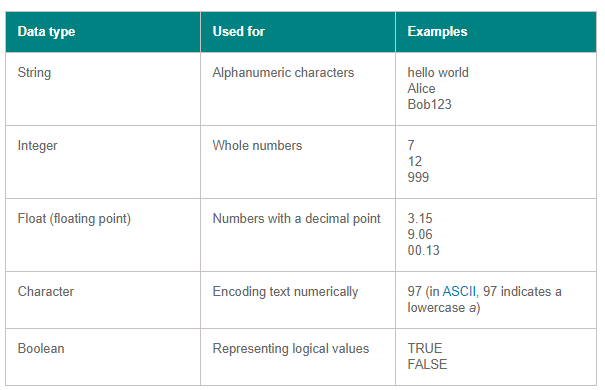
String a= “10” which is a symbol. String can store multiple numbers.

Bullean data type will be either true or false

Char data type will be in single inverted coma and can be stored only 2 numbers.

String can store combination of data type and can contain numeric number in form of character only.

For Python this data type is not needed. It is just for knowledge



**Operators—**are symbols

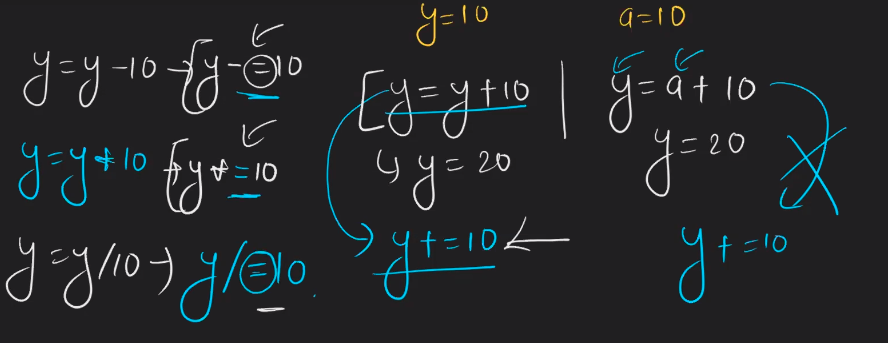
x + y – x and y is operands, + (addition) is operation.

1. Arithmetic operator- + , - , \* , / , % (Modulo) which is a remainder/Division.

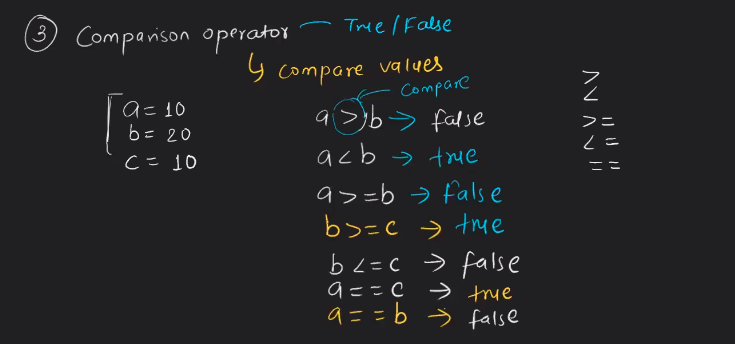
Modulo is shesh in gujrati which will be 0 or number

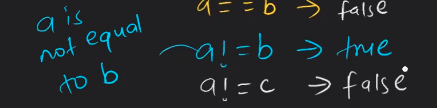
1. Assignment Operator- It is to assign values. x = 5 , x is assignment.

Z = x+y, + is arithmetic, z= 10+20 which becomes z = 30 where = is assignment.

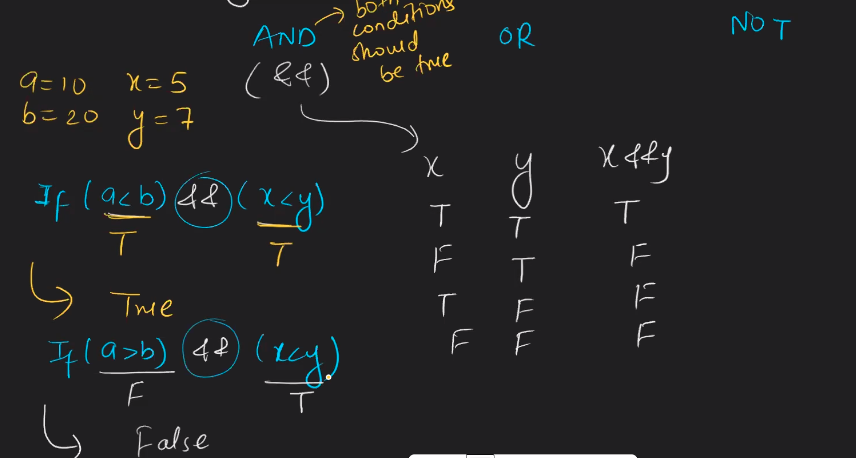


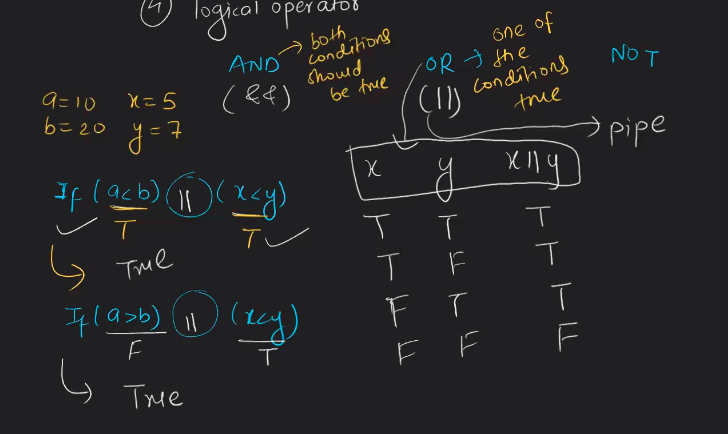
1. Comparison operator-In True or False. It is to compare values.





1. Logical operator- Both conditions should be true.

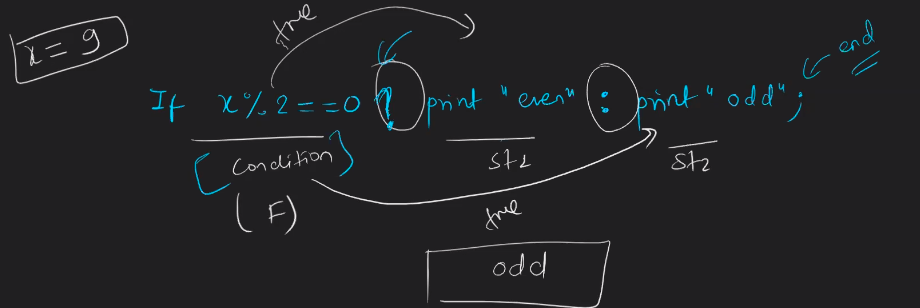




&& operator needs both conditions equal otherwise it will be false.

Not is going to reverse the value.

1. Conditional operator- Help us to define a condition.



**16th Dec**

**Algorithms**

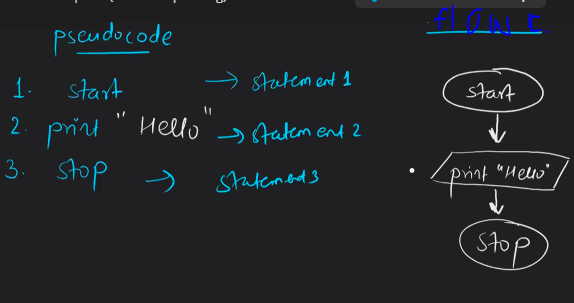
Programing foundation is just for theory. Practical will be done in Python.

We create lot of algorithms in daily life.

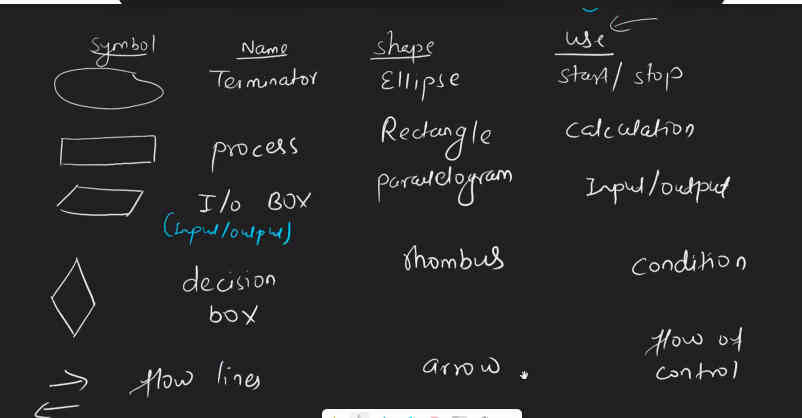
**Algorithms** is step by step instruction to complete a task. It can be write in 2 ways.

1. **Pseudo code**- Textual representation. Turn on stove, put on vessel, pour water, add sugar-tea, turn off stove, serve it (We can not skip any)

First step will be always start.

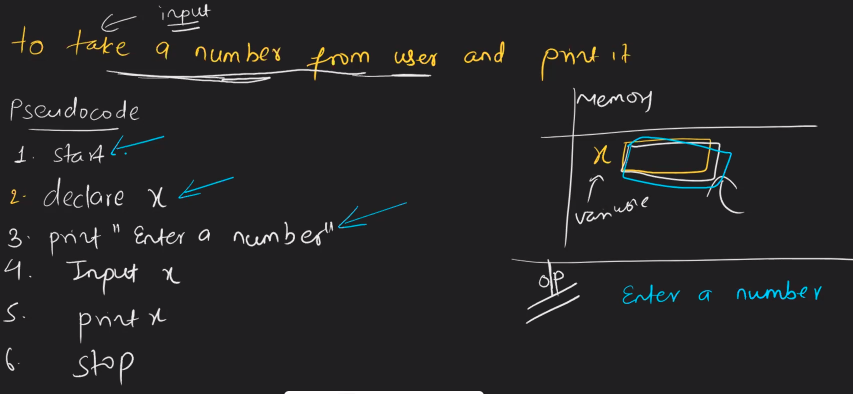


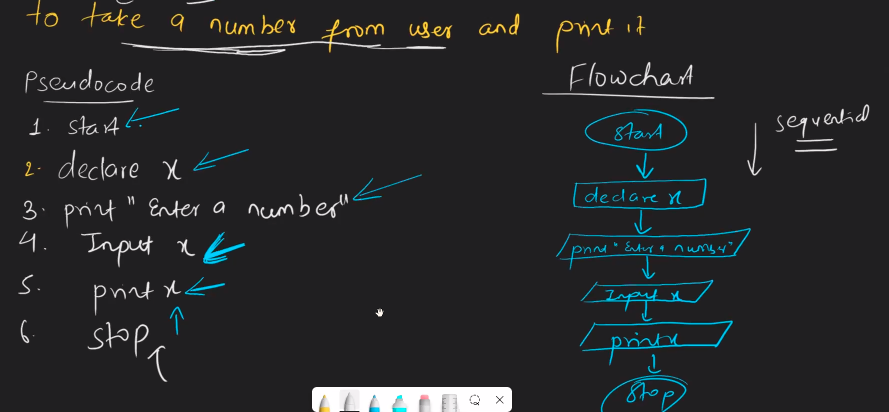
1. **Flow chat**- Graphical representation. It will be in vertical tables, step by step. Refer below table.



* To take a number from user and print it.

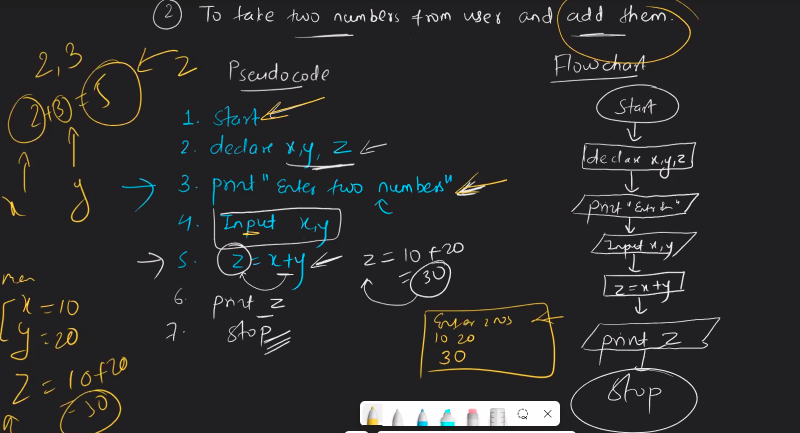
Variable was to store value. Variable is



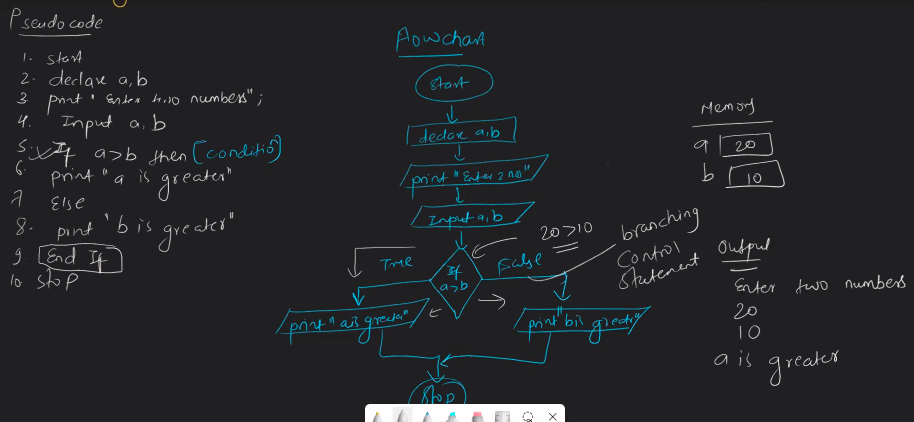


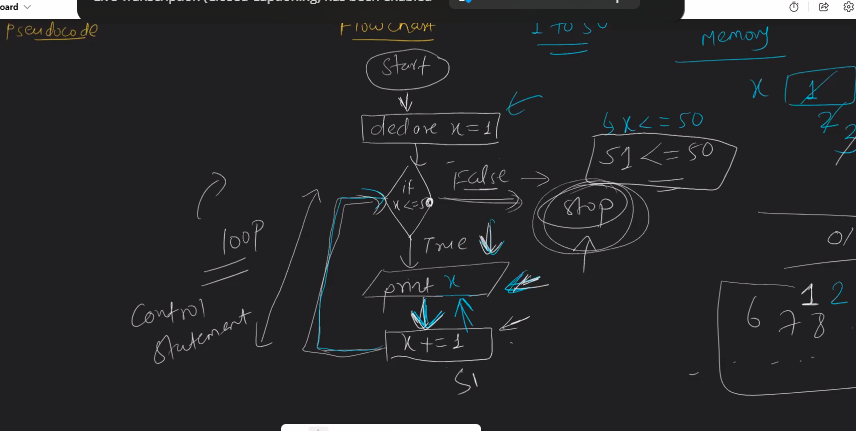
If you write x, number will display as a variable, If you write “x” then x will display

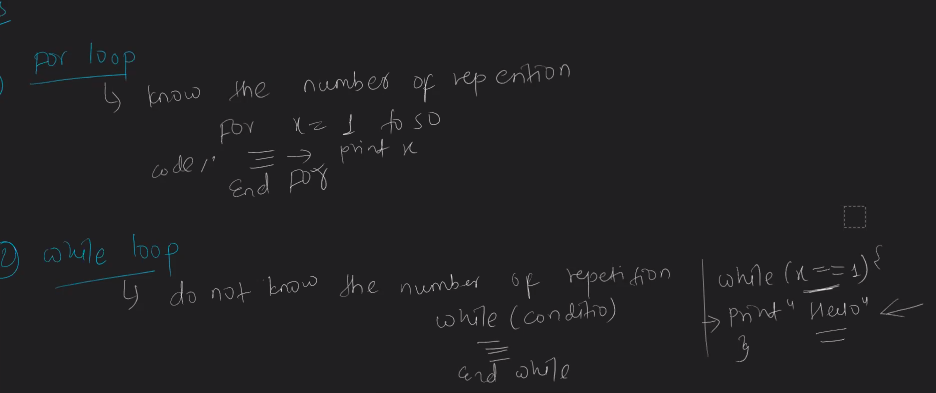
1. To take 2 numbers from user and add them.

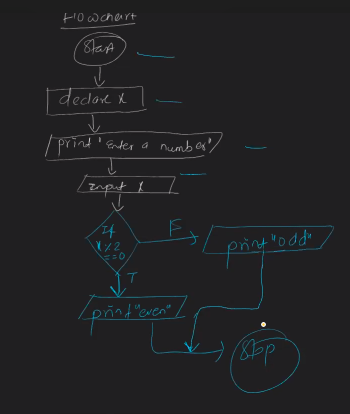


1. To take 2 numbers from user and find out which one is greater number.

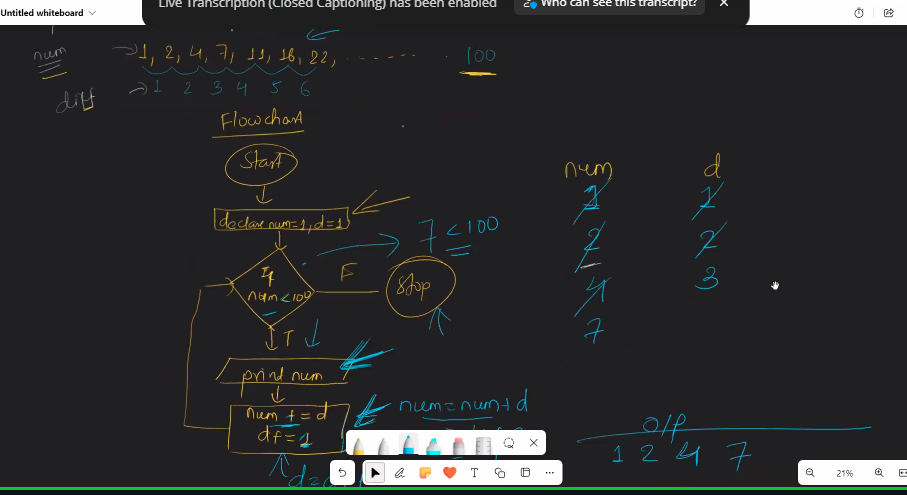






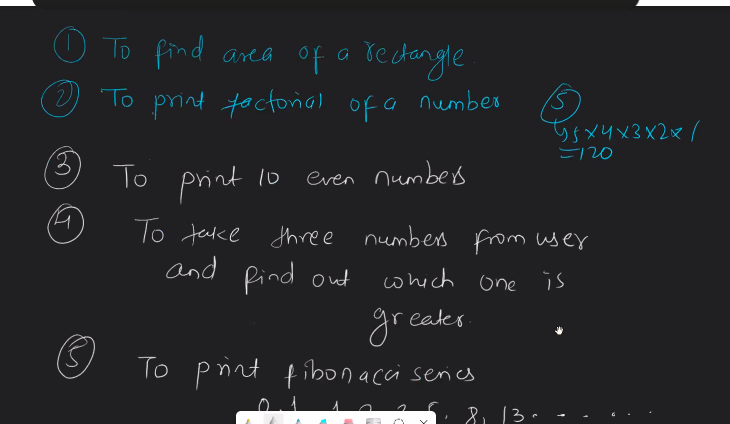


Loop example



You can always raise a ticket for one on one sessions.

Write home work in algorithms with code.



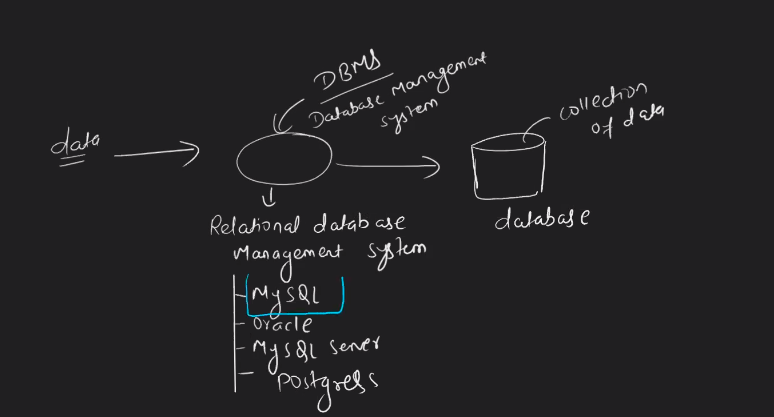
Submit this assignment in ask query. Subject name- Programing foundation

Use note book for this assignment. Also, upload these questions while raising query.

**17th Dec- Re watch**

**SQL- Structured Query Language – A language to deal with relational database.**

Database- Collection of data stored in such a way that it can be accessed easily. For that they need software application, which is DBMS (Database Management System)

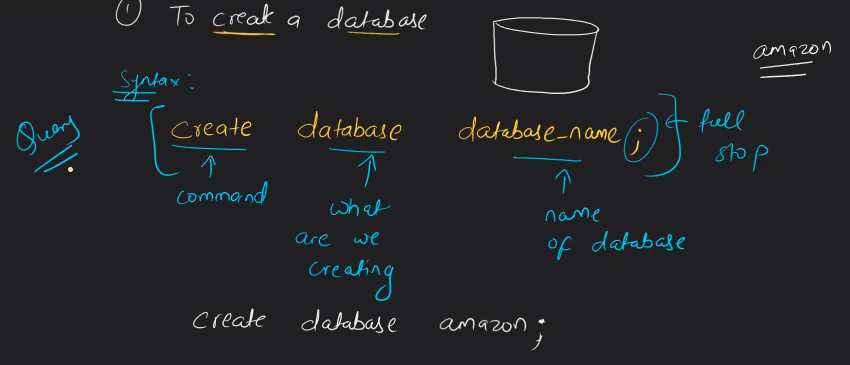


SQL is a language and excel is software application. Excel is for small things but for large skills we use data base (My SQl)

Data base is a collection of data. My SQL is a software to write a command.

CRUD operations- Create, Reqd, Update, Delete

**Syntax**- Rules to write commands.



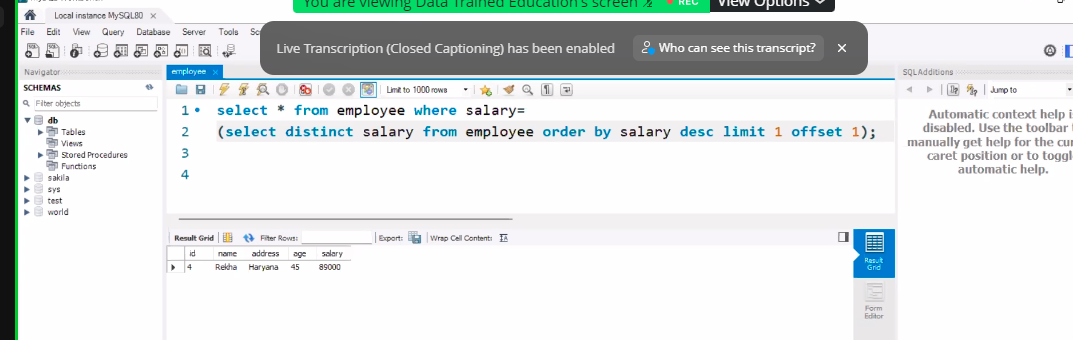
Every time we create new data base, it should be new name.

30th Dec assignment link

<https://docs.google.com/document/d/1kM1ziEfddNXYY8ciiK8wCBkTz6c9acXQ0sh5XtbNTc4/edit?pli=1>

Notes in rough book

31stDec



Assignment

<https://docs.google.com/document/d/1ODgGqXJUgDIo81ddx4vcrC3tN6d4eFf4lbcH96h8u1A/edit>

<https://docs.google.com/document/d/1kM1ziEfddNXYY8ciiK8wCBkTz6c9acXQ0sh5XtbNTc4/edit>

submit it in ask doubt question

**10th Feb**

**Pandas**

[GitHub - training-ml/Files](https://github.com/training-ml/Files)