DATA SCIENCE

BY TECHOLAS

WHAT IS DATA??

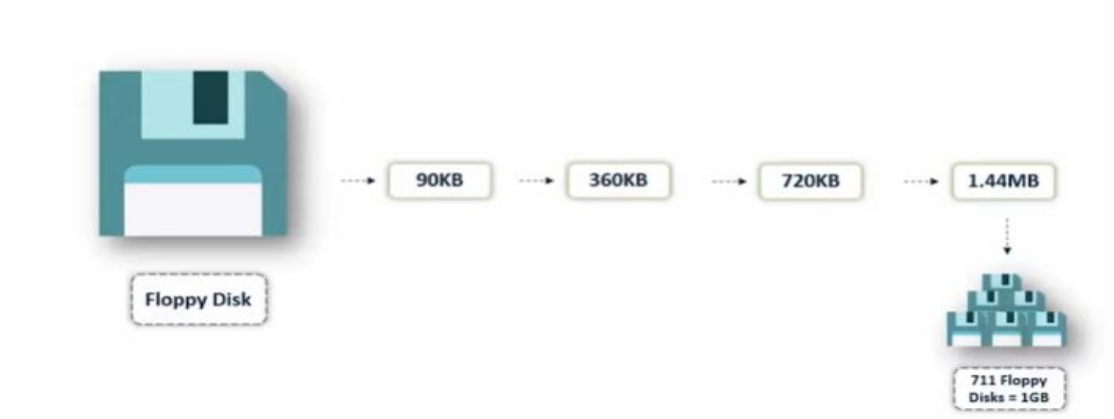
Data is just a collection of facts..

Eg: 100, ’ADWAITH’ , (a+b)^2=a^2+2ab+b^2

TECHOLAS
TECHNOLOGY DEMYSTIFIED

DATA BEFORE SOME YEARS AGO..

- Consider data ,some 10 or 15 years before ,It was all about structured data and its size was in kb’s and mb’s
- And Storing and processing this data was very easy by using traditional systems.



TECHOLAS
TECHNOLOGY DEMYSTIFIED

DATA NOW..

- In this Decade data is not that much simple
- Its size its complexity everything matters
- Now data comes in pb’s and eb’s
- And its type is not structured thats called unstructured data
- So Data processing is not so simple by our traditional processing system..

DATA.. DATA.. DATA...

DATA IS NOW EVERYWHERE..

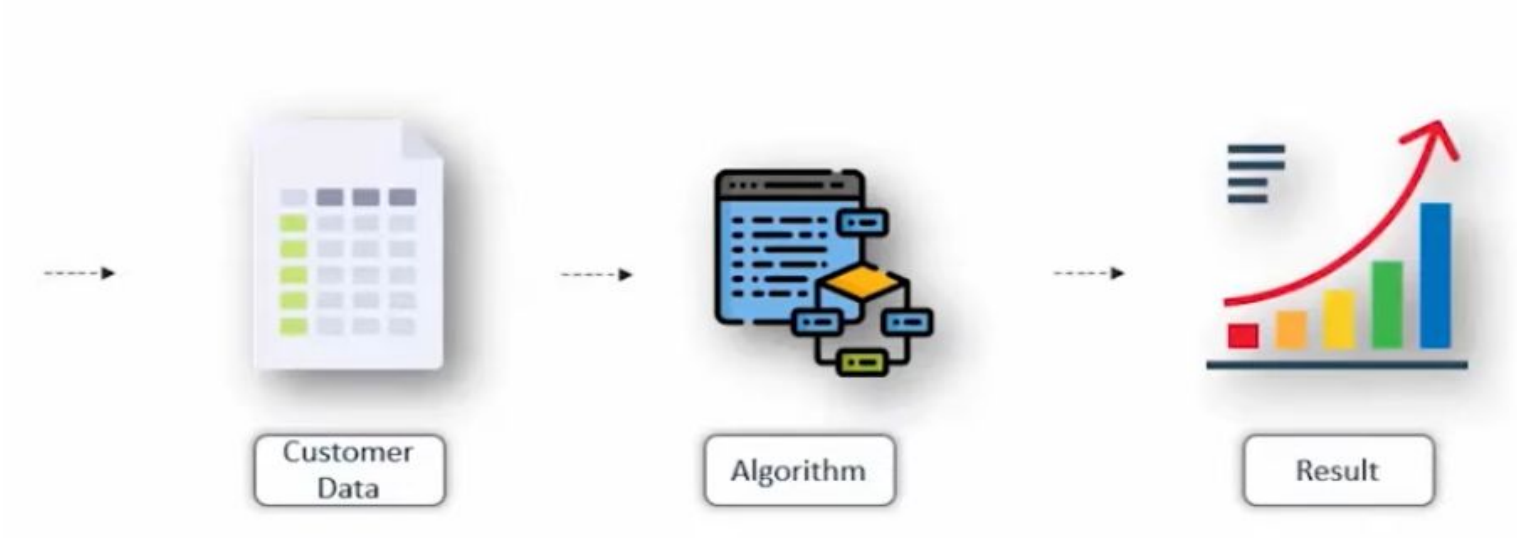
- SOCIAL MEDIA APPLICATIONS
- GOVERNMENT
- SMART CAR AND OTHER SMART DEVICES
- IOT DEVICES
- HOSPITAL AND HEALTH CARE..
- FINANCE SECTORS
- MANUFACTURING SECTORS
- ....

NEED OF DATA SCIENCE??

- We are getting huge amount of data from everywhere
- What to do with this data??
- Before that what is data science??

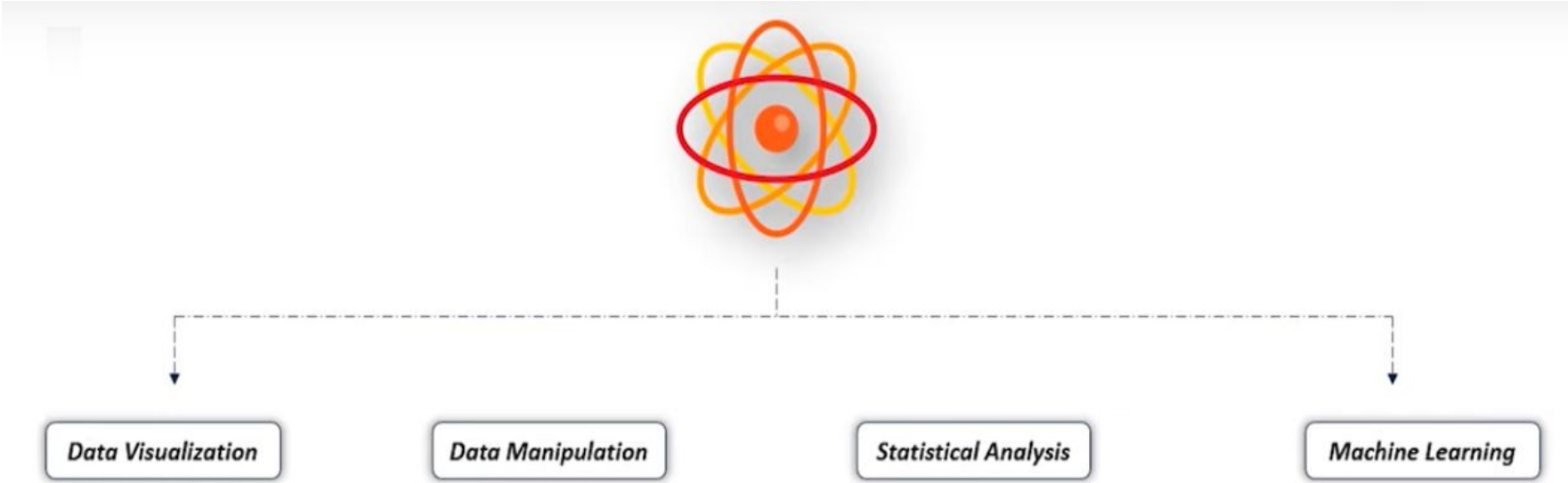
WHAT IS DATA SCIENCE??

- Its an act of Apply some science (or skills) on Data to make data talk to us.



CHOLAS
TECHNOLOGY DEMYSTIFIED

What are those skills??



TECHOLAS
TECHNOLOGY DEMYSTIFIED

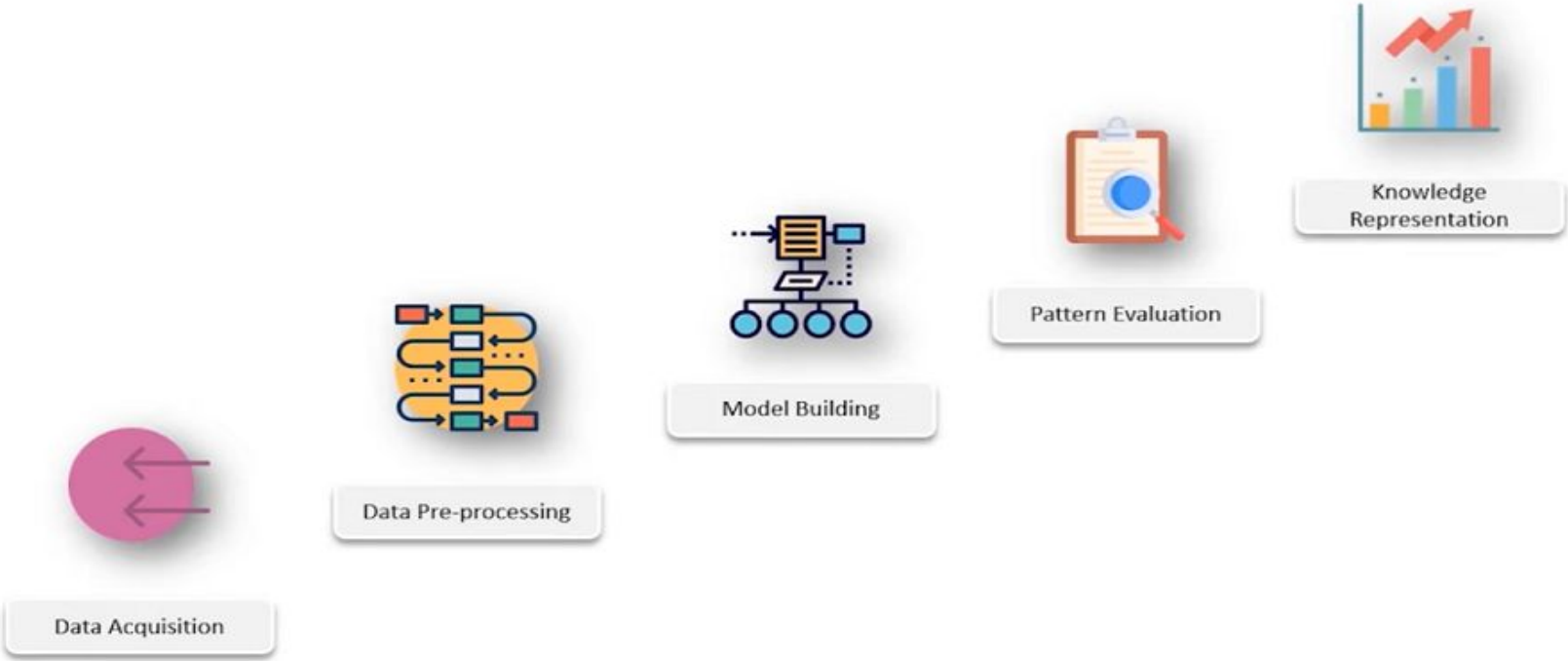
So ...Where Data Science ... ??

Consider An Example..

- John is MD of a Holiday hotel.. And he is not getting customers even he is losing customers
- Then he contact his friend Thejas, a Data Scientist
- He first collects data ,mine the data
- Creates some algorithms to analyse the data
- He reached some reasons and solutions from that
- He visualizes that and shown to John
- John does the remedies and he got his customers back

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Life Cycle Of Data Science..



DATA ACQUISITION..

- As you know data comes from multiple sources ..with multiple formats
- We integrate all this data and store in one place that is called Data warehouse
- And from this integrated data we will select particular section to start Data Science That is called TARGET

Data Pre Processing..

- The Target Data cannot be used for data science
- It's time to apply data pre processing..
- The row data will be transformed to different forms by using techniques like
  - NORMALIZATION
  - SUMMARIZATION
  - AGGREGATION
  - TRANSFORMATION

DATA MODELLING..

- Then we will apply some scientific algorithms to create the data models or find the interesting insights in the data
- The algorithms are like
  - LINEAR REGRESSION
  - K-MEANS
  - RANDOM FOREST

Data Validation..

- In this step we are checking the created model or pattern is correct or not
- For checking these things different validation techniques are available
- By checking these patterns if its not validated the pattern or model will be discarded

Knowledge Representation

- If the data Model is validated then it's time to represent the knowledge by using graphs or diagrams

Application of Data Science..

1. CHAT BOTS
   siri,cortana
2. SELF DRIVING CARS
3. SENTIMENT ANALYSIS
   Pre Election result predicting
4. FACE DETECTION
   Image tagging in facebook

5...

TECHOLAS
TECHNOLOGY DEMYSTIFIED

NUMPY...

Installation..

We are using jupyter notebook to code numpy

Installing Jupyter Notebook

1. **Installing using PIP**
   pip3 install jupyter
   pip3 install numpy
2. **Installing through anaconda**

   For new users, we **highly recommend** installing Anaconda. Anaconda conveniently installs Python, the Jupyter Notebook, and other commonly used packages for scientific computing and data science.

   Download Anaconda https://www.anaconda.com/distribution/

To Open: jupyter notebook

What is Numpy..?

- Numpy(Numerical Python) is a linear algebra library of python..
- Used to perform mathematical and logical operations on arrays
- It provides methods to perform operations on multidimensional arrays and matrices

Numpy Array..

- Numpy arrays are basically python list with some additional feature.
- Numpy arrays support 1D,2D,3D .....nD arrays

How to create Numpy array

- IMPORT NUMPY MODULE

  ```
  import numpy as np
  ```

- CREATE NUMPY 1D ARRAY

  ```
  a=np.array([2,5,6,4])   #python list
  b=np.array((1,2,34))    #python tuple
  print(a)
  print(b)
  ```

- Numpy array is 0 index based

  ```
  print(a[0])
  print(a[0:])
  print(a[:2])
  print(a[1:3])
  ```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

2D array..

```
b=np.array([[1,2,3,4],[5,6,3,6]])
print(b)
print(b[0])
print(b[0:])
print(b[0:2])

print(b[0:2][0])
```

output:[1 2 3 4]

```
print(b[1][1])
```

output:6

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Numpy array continues.. Initialization

- Zeros function
  a=np.zeros((2,2))
- arange() function
  arange(start,end,difference)
  a=np.arange(2,50,3)
- full() function
  full((rows,columns),default_value)
  d=np.full((2,3),10)
- linspace() function
  linspace(start,end,number_of_elements)
  a=np.linspace(2,8,4)

TECHOLAS
TECHNOLOGY DEMYSTIFIED

random

- random(number_of elements) function
  out= np.random.random(5)
- randint(start_range,end_range,number_of_elements)
  out_arr = np.random.randint(2, 10,5)
  out_arr = np.random.randint(2, 10)
  out_arr = np.random.randint(2, 10,(2,2))
- shape
  Prints the structure of the array
  print(out_arr.shape)

continues..

- Size of the array
  abc=np.arange(20)
  print(abc.size) -->size of the array
  print(abc.itemsize) -->size of the individual item in the array
- ndim-Dimension of the array
  print(abc.ndim)
- Dtype- finding the element type of the array
  Array in numpy are not heterogeneous but homogeneous
  print(abc.dtype)
- Nbytes → total number of bytes using
  print(a.nbytes)

Astype- to convert the type of the array

```
arr=np.array([1,2,3,4

print(arr.dtype)

arr=arr.astype('int8')

print(arr.dtype)

])
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Add or remove elements

- Append
  arr=np.array([1,2,3,4])
  np.append(arr,6)
- Insert
  np.insert(abc,1,2,axis=1) → 2nd arg: index at which insert, 3rd arg: content to insert
- Delete
  np.delete(abc,1,axis=1)

Copy & Sort

- COPY

  abc=np.array([1,6,3,2])

  ab=np.copy(abc) Or

  ab=abc.copy()
- SORT

  np.sort(abc,axis=1)

  abc.sort()

TECHOLAS

TECHNOLOGY DEMYSTIFIED

Element operation

```
a=np.array([4,5,6,7])
a*2
a-2
a+2
a/2
a  ** 2

b=np.array([2,6,4,3])
a+b
a-b
a*b
a/b
```

continues..

```
arr=np.array([[1,2,3,4,5],[6,7,8,9,10],[2,5,7,9,1]])
arr1=np.array([[1,2,3,4,5],[6,7,8,9,10],[2,5,7,9,1]])
```

- subtract()

  ```
  np.subtract(5,4)
  np.subtract(arr,arr1)
  ```

- divide()

  ```
  np.divide(2,4)
  np.divide(arr,arr1)
  ```

- multiply()

  ```
  np.multiply(2,4)
  np.multiply(arr,arr1)
  ```

- sqrt()

  ```
  np.sqrt(2)
  np.sqrt(arr)
  ```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

continue..

- sin()
  np.sin(0)
- cos()
  np.cos(0)
- tan()
  np.tan(0)
- log()
  np.log(10)
  np.log10(2)
- exp()
  np.exp(5) → e=2.71

TECHOLAS
TECHNOLOGY DEMYSTIFIED

- std() -->standard deviation

  np.std(arr)
- +,-, * .%

  arr=np.array([[1,2,3,4,5],[6,7,8,9,10]])

  arr1=np.array([[1,2,3,4,5],[6,7,8,9,10]])

  arr+arr1

  arr-arr2

  arr*arr2

  arr/arr2

  arr%arr1

continues..

- Vstack() and hstack()

  np.vstack([arr,arr1]) → vertically concatenate 2 arrays

  np.concatenate([arr,arr1],axis=0)

  np.hstack([arr,arr1]) → horizontally concatenate 2 arrays

  np.concatenate([arr,arr1],axis=1)

- ravel()

  np.ravel(arr) → Make a single dimension array

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Numpy slicing

```
arr=np.array([[1,2,3,4,5],[6,7,8,9,10],[2,5,7,9,1]])

print(arr[0,1])

print(arr[0:])

print(arr[0:2])

print(arr[0:2,3])
```

Linear Algebra

- Matrix multiplication:

```
a=np.ones((2,3))
b=np.full((3,2),5)
print(a)
print(b)
out=np.matmul(a,b)
```

Statistics..

```
abc=np.array([[1,2,3],[2,5,7],[3,6,9]])
print(np.max(abc))
print(np.min(abc))
print(np.max(abc,axis=1))
print(np.max(abc,axis=0))
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

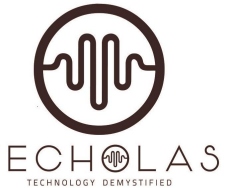
continues..

- sum()

```
arr=np.array([2,8])
arr2=np.array([4,6])
np.sum([5,10])
np.sum((2,5,6))
np.sum(arr)
np.sum([arr,arr2])
np.sum([arr,arr2],axis=1) #axis=0 for column wise sum, =1 for row wise sum
arr.sum(axis=0)
```

Load Data From File

```
data=np.genfromtxt('abc.txt',delimiter=',')
new_data=data.astype('int32')
```



Boolean Masking and Advanced Indexing

- **Boolean Masking**

  ```
  data=np.genfromtxt('abc.txt',delimiter=',')
  data>5
  data>=5
  data==5
  np.all(data>5,axis=1)
  np.any(data>5,axis=1)
  ```

- Indexing

  ```
  data[data>5]
  data[((data>5) & (data<10))]
  data[data%2==0]
  ```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

continues..

- Indexing with list

```
abc=np.array([2., 4., 6., 8., 4., 6., 8., 2., 6., 4., 2., 4., 2., 4.])
print(abc[[0,4,6,8]])
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Index It

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Array comparisons..

- Element wise Comparisons

  ```
  a=np.array([1,2,3])
  b=np.array([2,4,3])
  np.equal(a,b)
  ```

- Comparing full array

  ```
  np.array_equal(a,b)
  ```

Array BroadCasting..

- When doing aggregate functions with array if both array is not in the same shape the lower shape array will expand to accomplish the function

```
aa=np.array([[1,2,3],[4,5,6],[7,8,9]])
bb=np.array([[2,3,4]])
np.sum((aa,bb))
np.subtract(aa,bb)
```

Split array..

```
aaa=np.array([[1,2,3],[4,5,6],[7,8,9]])
[[1 2 3]
 [4 5 6]
 [7 8 9]]
```

AS PASSING INDEX

```
np.split(aaa,3,axis=0)
np.split(aaa,3,axis=1)
```

AS PASSING LIST

```
np.split(aaa,[1,3],axis=0)
np.split(aaa,[1,3],axis=1)
```

Split continues..

Passed Index=[a,b]

axis=0
Arr[:a,:]
Arr[a:b,:]
Arr[b:,:]

axis=1

Arr[:,:a]
Arr[:,a:b]
arr[:,a:]

TECHOLAS
TECHNOLOGY DEMYSTIFIED

exercises..

- Write a NumPy program to create a 3x3 matrix with values ranging from 2 to 10
- Write a NumPy program to get the unique elements of an array
- Write a NumPy program to find the union of two arrays. Union will return the unique, sorted array of values that are in either of the two input arrays.
- Write a NumPy program to create a 2-dimensional array of size 2 x 3 (composed of 4-byte integer elements), also print the shape, type and data type of the array.
- Write a NumPy program to create a 2D array whose diagonal equals [4, 5, 6, 8] and 0's elsewhere.
- Write a NumPy program to create a null vector of size 6 (1D) and update sixth value to 9
- Write a NumPy program to create a array with values ranging from 12 to 39
- Write a NumPy program to reverse a 1D array (first element becomes last).

AS
TECHNOLOGY DEMYSTIFIED

Exercise continues...

- Write a NumPy program to an array converted to a float type
- Write a NumPy program to create a 8x8 matrix and fill it with a checkerboard pattern
- Write a NumPy program to convert the values of Centigrade degrees into Fahrenheit degrees. Centigrade values are stored into a NumPy array
- Write a NumPy program to find the number of elements of an array, length of one array element in bytes and total bytes consumed by the elements.
- Write a NumPy program to test whether each element of a 1-D array is also present in a second array.
- Write a NumPy program to find common values between two arrays

Exercises continues...

- Write a NumPy program to remove specific elements in a numpy array.
  Original array:
  [10 20 30 40 50 60 70 80 90 100]
  Delete first, fourth and fifth elements:
  [20 30 60 70 80 90 100]
- Write a NumPy program to replace the negative values in a numpy array with 0.
  [-1 -4 0 2 3 4 5 -6]
- Write a NumPy program to count the occurrence of a specified item in a given NumPy array.
  [10 20 20 20 20 0 20 30 30 30 0 0 20 20 0] find the occurence of 20

TECHOLAS
TECHNOLOGY DEMYSTIFIED

PANDAS...

Intro..

- Pandas is a python module that makes data science easy and effective.
- Pandas is an open source library in Python. It provides ready to use high-performance data structures and data analysis tools.
- Pandas module runs on top of Numpy and it is popularly used for data science and data analytics.
- NumPy is a low-level data structure that supports multi-dimensional arrays and a wide range of mathematical array operations. Pandas has a higher-level interface. It also provides streamlined alignment of tabular data and powerful time series functionality.

Intro continues..

- DataFrame and Series are the key data structures in Pandas.
- Pandas provides a rich feature-set on the DataFrame. For example, data alignment, data statistics, slicing, grouping, merging, concatenating data, etc.
- Pandas Series is a one-dimensional labelled array capable of holding data of any type.

Pandas Series

Creating a pandas Series

```
import numpy as np
data = np.array([1,2,3,4,5])
pd.Series(data)
```

Data Frame

- Data frame is a main object in Pandas
- It is used to represent data with rows and columns as a 2-D data structure
- Tabular or Excel spreadsheet like data

Different ways to create a dataframe

1. Through reading csv file
2. Through reading excel file
3. Through python Dictionary
4. Through python tuple list
5. Through python dictionary list

Create DataFrame using python dictionary..

- Using method **DataFrame()**

```
import pandas as pd
emp={ 'id':[101,102,103,104,105],
'Name':['deepak','shyam','arun','manu','jeena'],
'Age':[25,22,24,20,27],
'salary':[25000,30000,150000,10000,30000] }

df=pd.DataFrame(emp)
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Create Dataframe using python tuple list

- Using **DataFrame()** method

```
Import pandas as pd
emp_data=[(101,'deepak',25,25000),
(102,'shyam',22,30000),
(103,'arun',24,150000),
(104,'manu',20,10000),(105,'jeena',27,30000)]

df=pd.DataFrame(emp_data,columns=['id','name','age','salary'])
```

- Each and every element of the tuple list is actually a row in your dataframe
- You should mention all column names as a list as a second argument

Create Dataframe using a list of dictionaries..

- Using **DataFrame()** method

```
Import pandas as pd
emp_data_list=[{'id':101,'name':'deepak','age':25,'salary':25000},
{'id':102,'name':'shyam','age':22,'salary':30000},
{'id':103,'name':'arun','age':24,'salary':150000},
{'id':104,'name':'manu','age':20,'salary':10000},
{'id':105,'name':'jeena','age':27,'salary':30000}]

df=pd.DataFrame(emp_data_list)
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Create dataframe using CSV

- We use **read_csv()**

```
import pandas as pd
df=pd.read_csv(r'C:\Users\mithu\Desktop\employee.csv')
print(df)
```

- read_csv() method accepts the path of the csv file ,if that exist in the home folder or current working directory ,then you can directly mention your filename or else you should mention the complete absolute path
- os.getcwd() → to find the current working directory

```
import os
print(os.getcwd())
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Skip the unnecessary header or add yours..

- While you are reading a csv file,If you have some unnecessary Header in your file you can skip it using the following command
  data=pd.read_csv(r'C:\Users\mithu\Desktop\stock_data.csv',skiprows=1) → tells to skip the given number from the top of file
  OR
  data=pd.read_csv(r'C:\Users\mithu\Desktop\stock_data.csv',header=1) → shows in which row your header locates
- If you don't have header in your file and you want to add some column header
  data=pd.read_csv(r'C:\Users\mithu\Desktop\stock_data.csv',header=None,names=['tickers','eps','revenue','price','people'])

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Read csv continues..,

- Read a limited number of rows from the top of the file
  data=pd.read_csv(r'C:\Users\mithu\Desktop\py-master\pandas\4_read_write_to_excel\stock_data.csv',nrows=3)
- Replace all unwanted value with NaN
  data=pd.read_csv(r'C:\Users\mithu\Desktop\stock_data.csv',na_values=['n.a.','not available'])
- If you want to replace your column value with NaN by specific to column
  data=pd.read_csv(r'C:\Users\mithu\Desktop\stock_data.csv',
  na_values={'eps':['n.a.','not available'],
  'revenue':['n.a.','not available',-1],'people':['n.a.','not available']})

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Write csv ..

- If you want to write your data frame to a CSV file
  data.to_csv(r'C:\Users\mithu\Desktop\abcde.csv',index=False)
- Write only specific columns
  data.to_csv(r'C:\Users\mithu\Desktop\abcde.csv',index=False,columns=['price', 'people'])
- Write Without Header
  data.to_csv(r'C:\Users\mithu\Desktop\abcde.csv',index=False,header=False)
- To append
  data.to_csv(r'C:\Users\mithu\Desktop\abcde.csv',mode='a',index=False,header=False)

Read txt file

- We can use the read_csv method to read a text file,with mentioning the delimiter
- Import pandas as pd
  data=pd.read_csv('abc.txt',delimiter='/t')

Create dataframe using excel file

- **read_excel()**

```
import pandas as pd
df= pd.read_excel(r'C:\Users\mithu\Desktop\employee.xlsx','Sheet1')
```

- read_excel() method contains 2 arguments first one the file path and the second one the sheet name in that excel file that we are looking to access.

Read Excel...

- data=pd.read_excel(r'C:\Users\mithu\Desktop\stock_data.xlsx','Sheet1')
- data=pd.read_excel(r'C:\Users\mithu\Desktop\stock_data.xlsx','Sheet1',na_values=['n.a.','not available'])
- data=pd.read_excel(r'C:\Users\mithu\Desktop\stock_data.xlsx','Sheet1',index=False)
- data=pd.read_excel(r'C:\Users\mithu\Desktop\pandas\4_read_write_to_excel\stock_data.xlsx','Sheet1',header=None, names=['a','b','c','d','e'])

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Using converters..

Using converters to convert data using conditions

- STEP 1:create the converter method

```
def convert_people(co):
if co=='larry page':
    return 'Mithun'
else:
    return co
```

- STEP2: use the converter

```
data=pd.read_excel(r'C:\Users\mithu\Desktop\py-master\pandas\4_read_write_to_excel\stock_data.xlsx','Sheet1', converters={'people':convert_people})
```

Write EXCEL ...

- data.to_excel(r'C:\Users\mithu\Desktop\new.xlsx')

Write two data frames in a single excel file

Using write ExcelWriter() method

```
with pd.ExcelWriter(r'C:\Users\mithu\Desktop\new2.xlsx') as wr:
    data1.to_excel(wr,sheet_name='Sheet1')
    data2.to_excel(wr,sheet_name='Sheet2')
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Read SQL

- By connecting pandas with mysql you can read sql tables .queries and you can insert values
- STEP 1:For connecting Sql with pandas you need to install 2 modules
  pip3 install PyMysql
  pip3 install sqlalchemy
- STEP 2: Next need to create a connection by using create_engine() method of **sqlalchemy** module
  conn=sqlalchemy.create_engine("mysql+pymysql://root:12345@localhost:3306/employee")
- STEP 3: read sql table by using pandas read_sql_table() method
  data=pd.read_sql_table('dept',conn)
  data=pd.read_sql_table('dept',conn,columns=['dloc','dname'])

read_sql_query()

- You can read the sql query by using read_sql_query()

```
query='select * from dept'
data=pd.read_sql_query(query,conn)
```

Write to sql table..

- We can write dataframe directly to mysql table

```
data=pd.read_csv('weather_data.csv')
data.to_sql(name='weather_table',con=conn,index=False)
data.to_sql(name='weather_table',con=conn,index=False,if_exists='append')
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Basics..

- **Shape**
  df.shape → provides the shape of the Dataframe
- **head**()
  df.head() → provides the first 5 rows of the DataFrame
  df.head(3) → provides first 3
- **tail**()
  df.tail() → provides the last 5 rows of the DataFrame
  df.tail(2) → Provides the last 2 rows of the dataframe

Basics continues..

- **Columns**
  df.columns
- **To print a specific column data**
  df.name or df['name'] → where name is the column name in df dataframe
- **To get some columns**
  Df[['id','name','age']]
- **Index**
  df.index → to view the index assigned to the data frame
  df.set_index('id') → To change the index to some other column value
  df.reset_index() → Reset index to previous stage

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Basics continues..

- **Adding new column to a dataframe**

  data[‘new_column_name’] = column_values

- **sort_values()**
  data.sort_values('Name',ascending=True)

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Basics continues..

- **loc**

  .loc() is a label based data selecting method which means that we have to pass the name of the row or column which we want to select. It can accept boolean data.

- **iloc**

  iloc() is a indexed based selecting method which means that we have to pass integer index in the method to select specific row/column.

Basics continues..

- **dtypes**
  It returns a Series with the data type of each column
  data = pd.read_csv('weather_data.csv')
  data.dtypes

- **unique**
  To return the unique elements of each column of a dataframe
  data.city.unique()

- **value_counts**
  returns counts of unique rows
  data.city.value_counts()

Iterate over pandas dataframes..

- To print all column names in the dataframe

```
for i in data:
 print(i)
```

- By using iteritems()

```
for column,values in data.iteritems():
    print (column)
    print(values)
```

- By using iterrows()

```
for index,row in data.iterrows():
     print('row',index)
     print(row)
```

Slicing and indexing

- df[2:5] → Provides all rows from 2 to 4
- df[:] → Provides all rows
- df[['name','age']][2:5] → will provide all rows from 2 to 4 with columns name and age

Basic Operations..

- **max()**
  df['age'].max() → Provides the maximum age in the age column
- **min()**
  df['age'].min()
- **mean()**
  df.salary.mean() → provides the average salary of the dataframe
- **std()**
  df.age.std() → **provides the standard deviation**
- **describe()**
  df.describe() → **provides all the status**

Conditional Selecting..

- To select rows from the Dataframe with conditions
  df[df.age>23] → Will provide all rows where age>23
- Find the name and salary of the person who is having highest age??

Date as the index

- data=pd.read_csv(r'C:\Users\mithu\Desktop\datas\weather_data.csv')
- data.set_index('day')
- type(data['day'][0])
- data=pd.read_csv(r'C:\Users\mithu\Desktop\datas\weather_data.csv',parse_dates=['day'])

TECHOLAS
TECHNOLOGY DEMYSTIFIED

date_range() ..

date_range() function is used to create a range of data from a start date and end date

drng=pd.date_range(start='1/1/2017',end='1/22/2017',freq='D') → freq: it means the frequency of generating date range it have various options

https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html#timeseries-offset-aliases

TECHOLAS
TECHNOLOGY DEMYSTIFIED

to_datetime()

- To_datetime function is very useful in pandas to convert any String date to type datetime

```
dates = ['2017-12-05', 'Jan 5, 2017', '01/05/2017', '2017.01.05',
'2017/01/05','20170105','Jan 5 2017','5 jan 2017','6 january 2019']

pd.to_datetime(dates)
pd.to_datetime('01/2/15',yearfirst=True)
pd.to_datetime('05/2/2015',dayfirst=True)
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Replace missing values in a dataframe..

- isna()-> check whether null values are present
  ```
  data.isna()
  data.isna().sum()
  data.isnull().sum()
  ```
- fillna()
  ```
  new_df=data.fillna(0)
  new_df=data.fillna({'temperature':0,'windspeed':0,'event':'unknown'})
  ```
- fillna() -> forward fill
  `new_df=data.fillna(method='ffill')` —>It fill with previous row value
- fillna() ->backward fill
  `new_df=data.fillna(method='bfill')` —>It fill with next row data
  ```
  new_df=data.fillna(method='ffill',limit=1)
  ```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Linear Interpolation

Linear interpolation is often used to approximate a value of some function *f* using two known values of that function at other points

interpolate()

new_df=data.interpolate()

TECHOLAS
TECHNOLOGY DEMYSTIFIED

DROP rows having NaN values

- dropna()
  new_df=data.dropna() —> drop row having any NaN value
  new_df=data.dropna(how='all') —> drop row having all column value NaN
- dropna() with threshold value
  new_df=data.dropna(thresh=2) —> maintain all rows having at least 2 non NaN values

replace()..

- replace() is a function to replace some special values to another value in data frame
- It can use in different ways

replace() continues..

- data=pd.read_csv(r'C:\Users\mithu\Desktop\datas\weather_data_replace.csv')
  data.replace(-99999,np.NaN) —> replace the value with NaN
- data.replace([-99999,-88888],np.NaN)—> Replace multiple values with NaN
- Replace values with specific to columns
  data.replace({'temperature':-88888,'windspeed':-99999,'event':'0'},np.NaN)
  data.replace({'temperature':[-88888,-99999],'windspeed':[-99999,-88888],'event':'0'},np.NaN)
- Map data with replace()

```
data.replace({-99999:np.NaN,
              -88888:np.NaN,
                  '0':'sunny'})
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

replace() continues..

- Replace value with regex
  data.replace({'temperature':'[A-Za-z]','windspeed':'[A-Za-z]'},'',regex=True)
- Replace list of values with another list of values
  data.replace(['Rain','Sunny','Snow'],[101,102,103])

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Group By..

- Group by is the operation in pandas that create groups of rows by a column or set of columns
- Group by creates sub dataframes,that means it's actually split the main dataframe into sub dataframes based on a single column value or combination of multiple column value
- groupby() is the method used to do this ,We can pass a single column name or a list of column names as the argument of this method

```
data=pd.read_csv('weather_by_cities.csv')
grp_result=data.groupby('city')
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Pictorial representation..

day	city	temperature	windspeed	event
1/1/2017	new york	32	6	Rain
1/2/2017	new york	36	7	Sunny
1/3/2017	new york	28	12	Snow
1/4/2017	new york	33	7	Sunny
1/1/2017	mumbai	90	5	Sunny
1/2/2017	mumbai	85	12	Fog
1/3/2017	mumbai	87	15	Fog
1/4/2017	mumbai	92	5	Rain
1/1/2017	paris	45	20	Sunny
1/2/2017	paris	50	13	Cloudy
1/3/2017	paris	54	8	Cloudy
1/4/2017	paris	42	10	Cloudy

df.groupby('city') →

DataFrameGroupBy

new york ->

day	city	temperature	windspeed	event
1/1/2017	new york	32	6	Rain
1/2/2017	new york	36	7	Sunny
1/3/2017	new york	28	12	Snow
1/4/2017	new york	33	7	Sunny

mumbai ->

day	city	temperature	windspeed	event
1/1/2017	mumbai	90	5	Sunny
1/2/2017	mumbai	85	12	Fog
1/3/2017	mumbai	87	15	Fog
1/4/2017	mumbai	92	5	Rain

paris ->

day	city	temperature	windspeed	event
1/1/2017	paris	45	20	Sunny
1/2/2017	paris	50	13	Cloudy
1/3/2017	paris	54	8	Cloudy
1/4/2017	paris	42	10	Cloudy

Group by continues..

- How to fetch the group that creates??
  grp_result.get_group('paris') —> will provide the group (data frame) by specific to ‘paris’ city
- Fetch all groups.. By using for loop..

```
for city ,city_df in grp_result:
    print(city)
    print(city_df)
```

Group operations..

- Aggregate functions on group objects
  grp_result.max()
  grp_result.mean()
- Analytics on specified group columns
  grp_result.get_group('mumbai').max()

Group by multiple columns..

- We can pass a list of columns

```
grp=data.groupby(['city','event'])
```

- While fetching you should pass a tuple to get the group

```
grp.get_group(('new york','Sunny'))
```

- Fetch all using for loop

```
for city ,city_df in grp:
    print(city)
    print(city_df)
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

concat..

- **concat()** function is used to concatenate two or more DataFrames
- Concat can do both vertically(axis=0) as well as horizontal (axis=1)

```
tcs={'id':[101,102,103,104],
'name':['Mithun','Dipin','jose','Rahul'],
 'Age':[25,26,27,29]}

wipro={'id':[101,102,103,104],
'Name':['dev','sreenath','sreerag','shafeel'],
'Age':[24,24,27,26]}

tcs_emp=pd.DataFrame(tcs)
wipro_emp=pd.DataFrame(wipro)
pd.concat([tcs_emp,wipro_emp])
```

concat..

- Ignore index —> TO get common index after concatenation
  pd.concat([tcs_emp,wipro_emp],ignore_index=True)
- keys—> To provide separate keys for each concatenate data frames
  data=pd.concat([tcs_emp,wipro_emp],keys=['TCS','WIPRO'])
- Concatenating vertically
  data=pd.concat([tcs_emp,wipro_emp],axis=1)
- Concatenating a dataframe with a series
  salary=pd.Series([65000,75000,15000,25000],name='salary')
  pd.concat([tcs_emp,salary],axis=1)

merge...

- Merge is a process to merge two tables horizontally

```
tcs={'id':[101,102,103,106],
     'name':['Mithun','Dipin','jose','maneesh'],
     'age':[25,26,27,29]}
wipro={'id':[101,102,103,105],
     'name':['dev','sreenath','sreerag','varun'],
     'salary':[25000,34000,37000,20000]}
tcs_emp=pd.DataFrame(tcs)
wipro_emp=pd.DataFrame(wipro)
pd.merge(tcs_emp,wipro_emp,on='id')
pd.merge(tcs_emp,wipro_emp,on='id',suffixes=['_left','_right'])
```

merge..

- pd.merge(tcs_emp,wipro_emp,on='id',how='outer')
- pd.merge(tcs_emp,wipro_emp,on='id',how='outer',indicator=True)
- pd.merge(tcs_emp,wipro_emp,on='id',how='left')
- pd.merge(tcs_emp,wipro_emp,on='id',how=right)

Pivot and pivot tables...

- Pivot allows to reshape the data frame

```
datas=pd.read_csv('weather.csv')
datas.pivot(index='date',columns='city')
datas.pivot(index='date',columns='city',values='temperature')
```

- Pivot Table is used to summarize or aggregate data with in a dataframe

```
datas=pd.read_csv('weather2.csv')
datas.pivot_table(index='city',columns='date')
datas.pivot_table(index='city',columns='date',aggfunc='sum')
datas.pivot_table(index='city',columns='date',aggfunc='sum',margins=True)
```

melt..

- Melt allows to transform or reshape the data frame

```
datas=pd.read_csv('weather.csv')
pd.melt(datas,id_vars=['day'])
pd.melt(datas,id_vars=['day'],value_name='temperature',var_name='city')
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Cross tab or contingency table..

Contingency table: a table showing the distribution of one variable in rows and another in columns, used to study the correlation between the two variables.

```
data=pd.read_excel('survey.xls')
pd.crosstab(data.Nationality,data.Handedness)
pd.crosstab(data.Nationality,data.Handedness,margins=True)
pd.crosstab(data.Nationality,[data.Sex,data.Handedness],margins=True)
pd.crosstab(data.Nationality,data.Sex,values=data.Age,aggfunc=np.average)
```

map..

- Map is a python built in function
- It is used to apply some functions on each element of a sequence or collection item
- It can be used in the pandas dataframe as well

```
def square(x):
    return x*x
abc=[1,2,3,4,5]
out=list(map(square,abc))
out
```

Map on dataframe ⟶ applymap()

- We have a built in function applymap() to use this map concept in pandas

```
import pandas as pd
data=pd.read_csv('weather.csv')
data[['chennai','chicago']].applymap(square)
```

- Apply with lambda function

```
data[['chennai','chicago']].applymap(lambda x: x*x)
```

map() on pandas series

```
import pandas as pd

data=pd.read_csv('weather.csv')

data['chennai'].map(lambda x: x*x)
```

MATPLOTLIB...

Matplotlib Intro..

- It is the python visualization library
- Provides wide variety of graphs
- Very Strong visualization tools
- Provides a module named pyplot
- It supports very simple functions to visualize
- Easy integration with pandas and numpy

Installation and import..

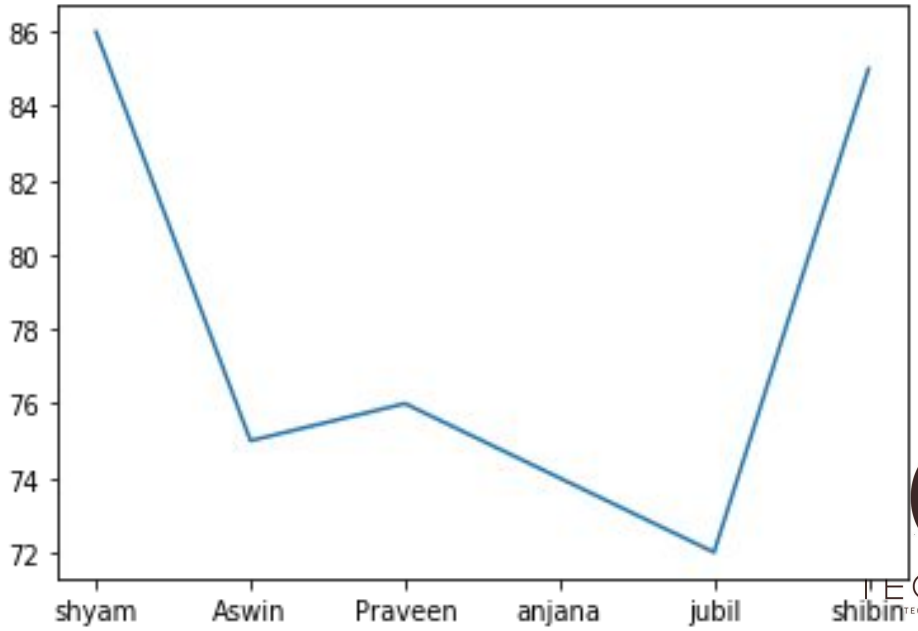
- pip3 install matplotlib

Open the jupyter notebook

- from matplotlib import pyplot

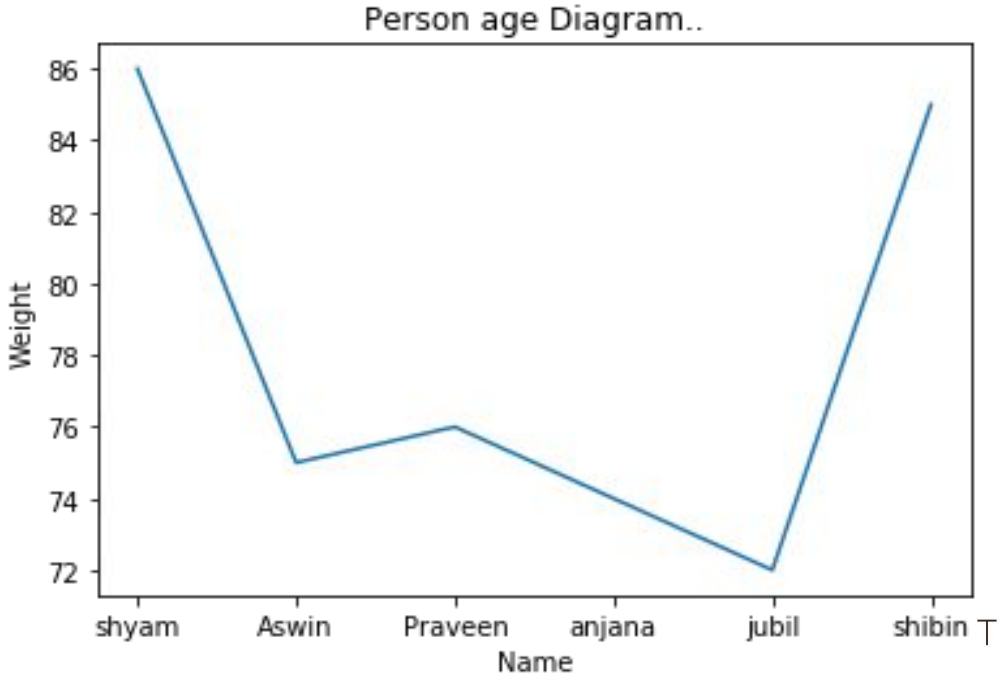
Plot your first..

- from matplotlib import pyplot as plt
- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value)
- plt.show()



Add title and Label

- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.show()

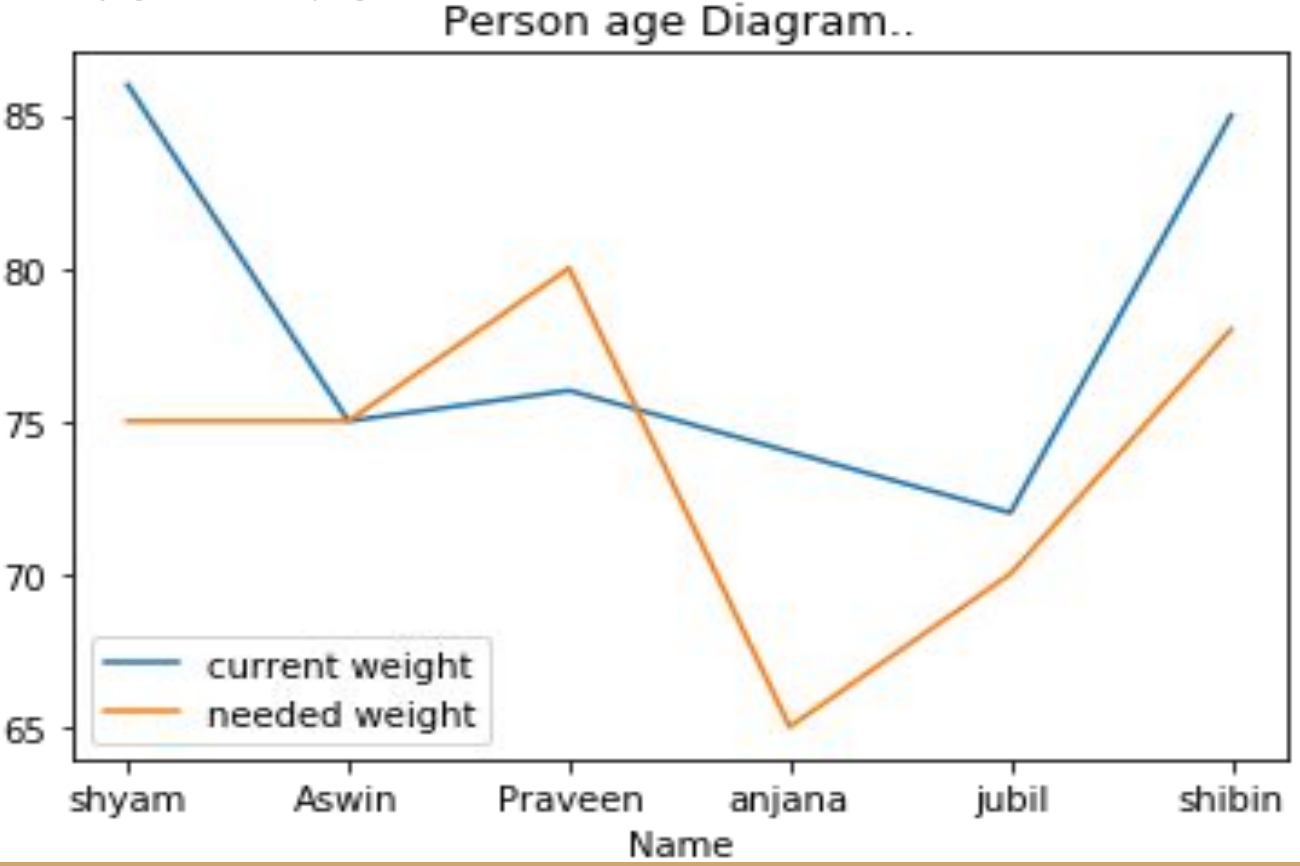


TECHOLAS
TECHNOLOGY DEMYSTIFIED

Multiple plot and adding legends

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value)
- y_value_2=[75,75,80,65,70,78]
- plt.plot(x_val,y_value_2)
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.legend(['current weight','needed weight'])
- plt.show()

continues..



TECHOLAS
TECHNOLOGY DEMYSTIFIED

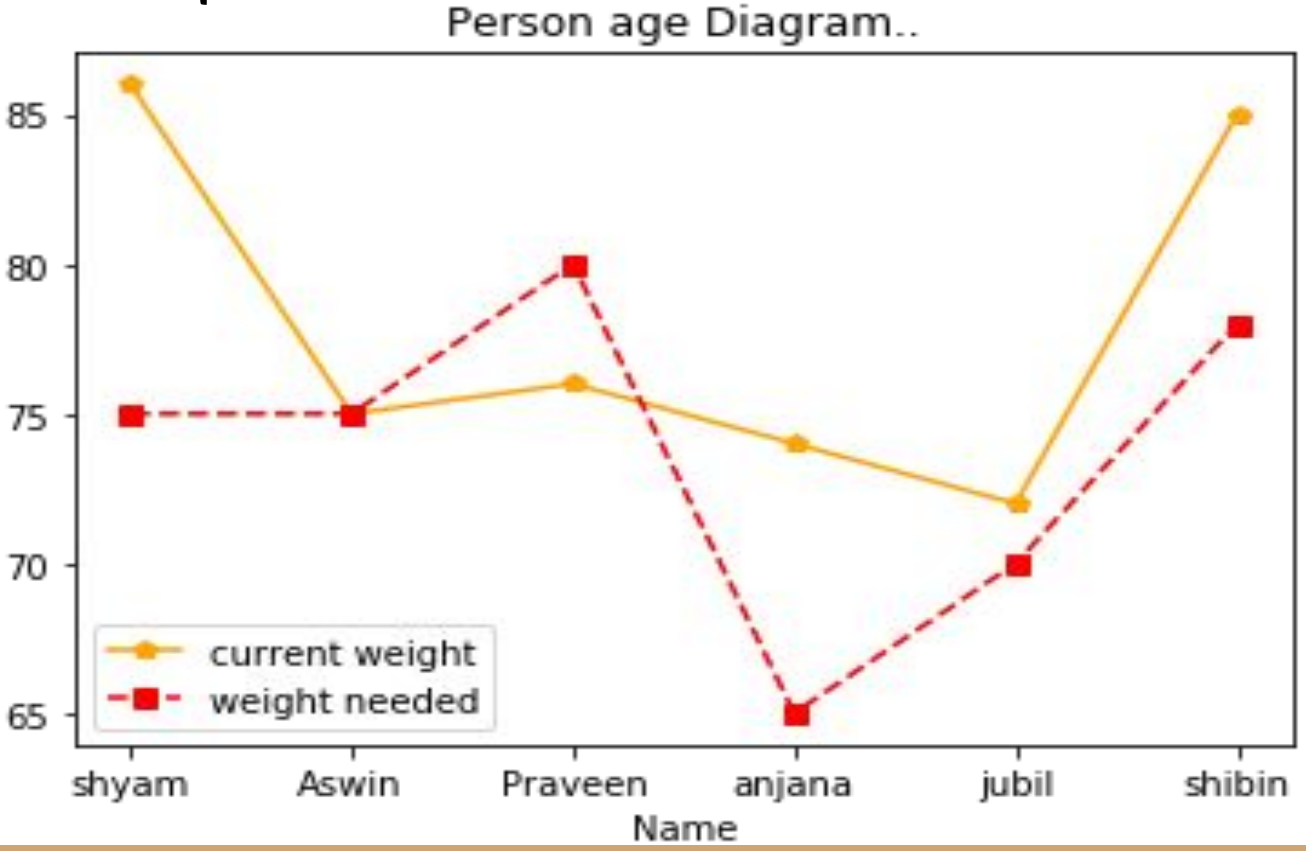
Same can do in this way as well..

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value,label='current weight')
- y_value_2=[75,75,80,65,70,78]
- plt.plot(x_val,y_value_2,label='weight needed')
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.legend()
- plt.show()

color,linestyle,marker

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value,label='current weight',c='orange',linestyle='-',marker='p')
- y_value_2=[75,75,80,65,70,78]
- plt.plot(x_val,y_value_2,label='weight needed',color='red',linestyle='--',marker='s')
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.legend()
- plt.show()

output..



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Same using format String..

- Format String

  fmt = '[marker][line][color]'

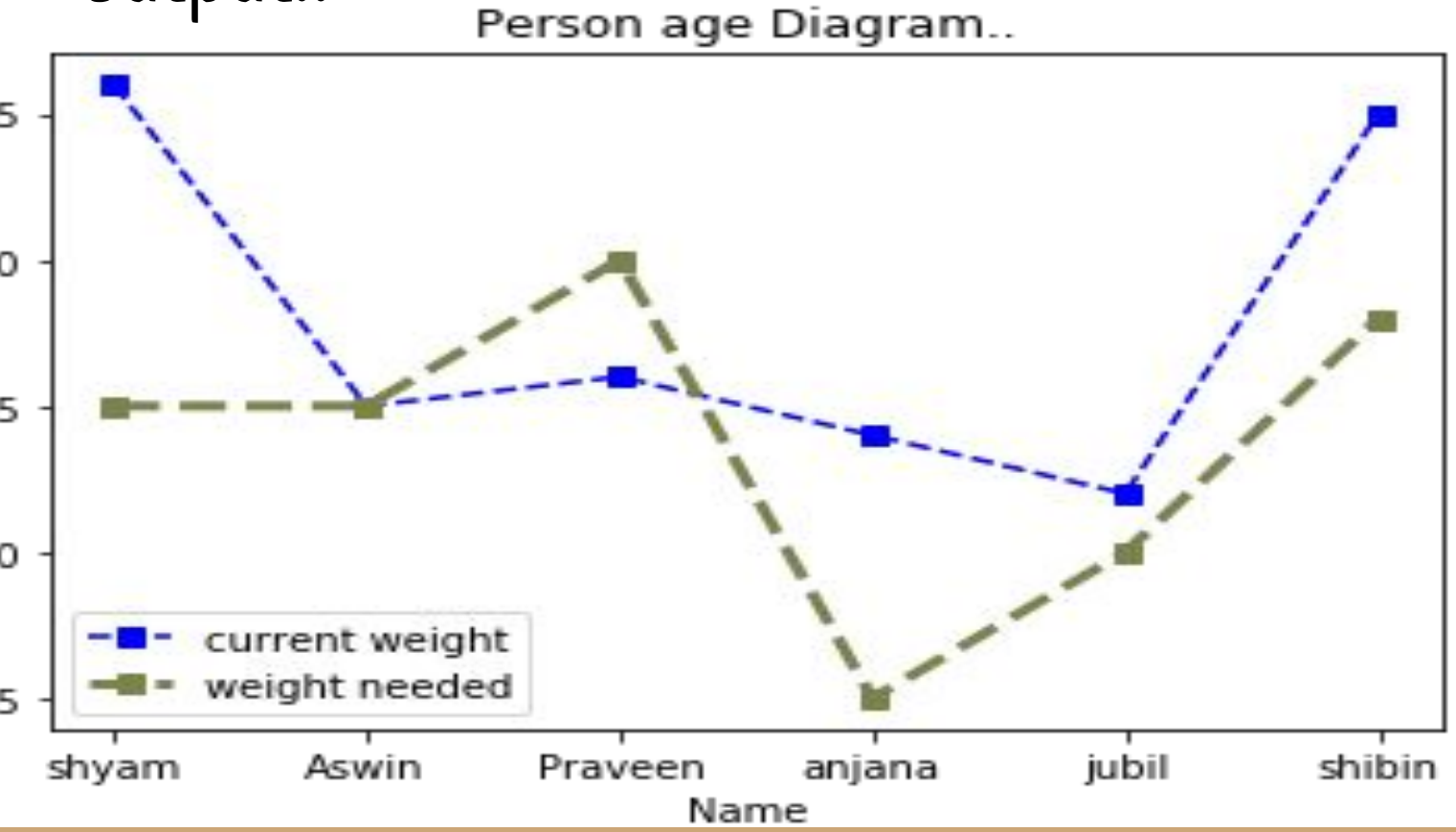
  fmt=’s--b’

  - x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
  - y_value=[86,75,76,74,72,85]
  - plt.plot(x_val,y_value,'s--b',label='current weight',)
  - y_value_2=[75,75,80,65,70,78]
  - plt.plot(x_val,y_value_2,label='weight needed',color='red',linestyle='--',marker='s')
  - plt.title('Person age Diagram..')
  - plt.xlabel('Name')
  - plt.ylabel('Weight')
  - plt.legend()
  - plt.show()

Color code and Increase line width..

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value,'s--b',label='current weight',)
- y_value_2=[75,75,80,65,70,78]
- plt.plot(x_val,y_value_2,label='weight needed',color='#7a834c',linewidth=3,linestyle='--',marker='s')
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.legend()
- plt.show()

output..



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Adding Grid..

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value,'s--b',label='current weight',)
- y_value_2=[75,75,80,65,70,78]
- plt.plot(x_val,y_value_2,label='weight needed',color='#7a834c',linewidth=3,linestyle='--',marker='s')
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.grid(True)
- plt.legend()
- plt.show()

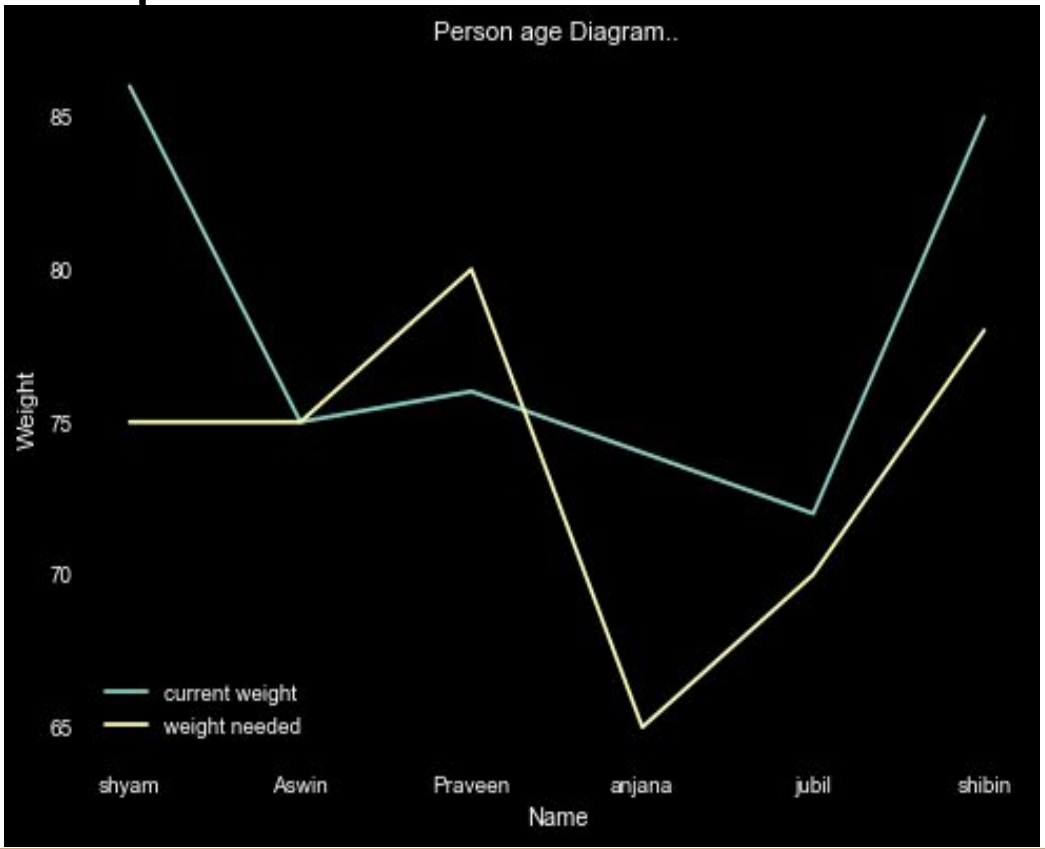
Styles in pyplot..

- There are plenty of built in styles that can apply on out plot provided by the matplotlib
- To get all available styles
  plt.style.available
- Use the available styles..
  plt.style.use('dark_background')

Apply the style..

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- y_value=[86,75,76,74,72,85]
- plt.plot(x_val,y_value,label='current weight',)
- y_value_2=[75,75,80,65,70,78]
- plt.plot(x_val,y_value_2,label='weight needed')
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.grid(False)
- plt.legend()
- plt.style.use('dark_background')
- plt.show()

output..



TECHOLAS
TECHNOLOGY DEMYSTIFIED

To save plot as image

- plt.savefig('weight_graph.png')

Bar chart..

- Instead of plot() we need to use bar() to get a barchart

```
x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
```

  - y_value=[86,75,76,74,72,85]
  - plt.bar(x_val,y_value,label='current weight',)
  - plt.title('Person age Diagram..')
  - plt.xlabel('Name')
  - plt.ylabel('Weight')
  - plt.grid(True)
  - plt.legend()

```
plt.style.use('dark_background')
plt.show()
```

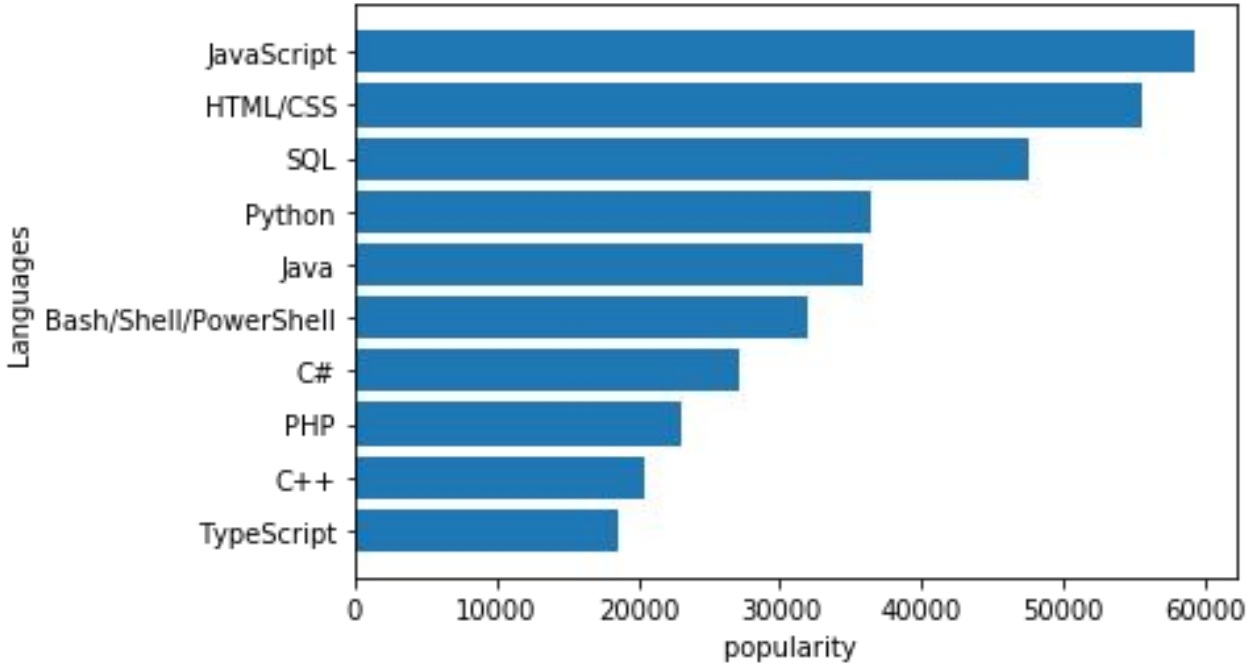
TECHOLAS
TECHNOLOGY DEMYSTIFIED

Multiple plotting in barchart..

- x_val=['shyam','Ashwin','Praveen','anjana','jubil','shibin']
- x_range=np.arange(len(x_val))
- width=.4
- y_value=[86,75,76,74,72,85]
- plt.bar(x_range-width,y_value,width=width,label='current weight',)
- y_value_2=[75,75,80,65,70,78]
- plt.bar(x_range,y_value_2,width=width,label='weight needed')
- #plt.bar(x_range+width,y_value_2,label='weight needed')
- plt.title('Person age Diagram..')
- plt.xlabel('Name')
- plt.ylabel('Weight')
- plt.xticks(ticks=x_range,labels=x_val)
- plt.grid(True)
- plt.legend()
- plt.style.use('dark_background')
- plt.show()

Horizontal barchart

- We use barh() method to plot a horizontal barchart

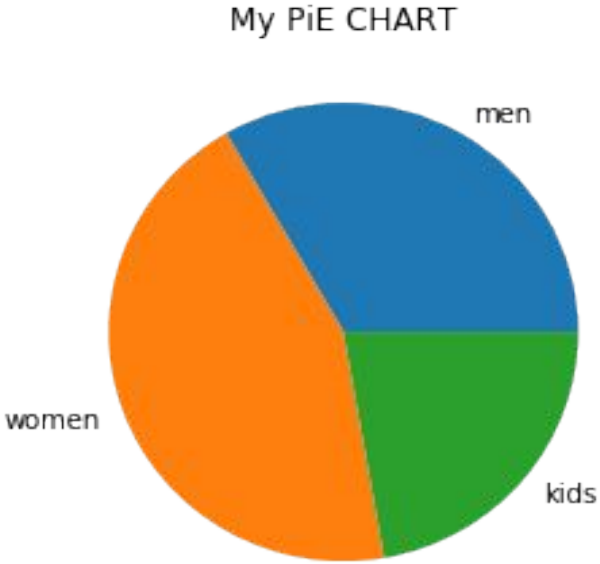


TECHOLAS
TECHNOLOGY DEMYSTIFIED

Pie Chart..

- It uses a method pie()

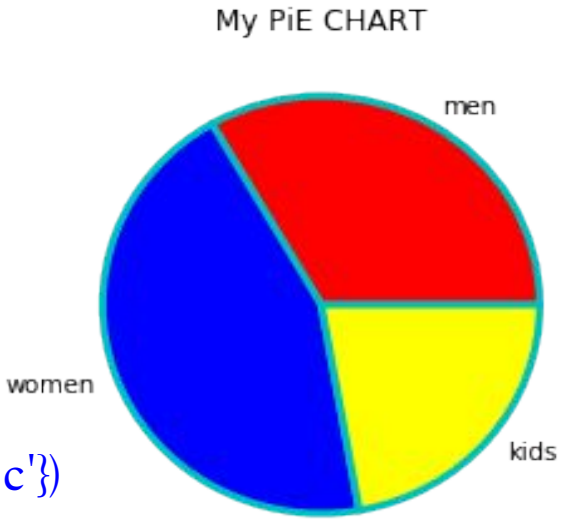
```
from matplotlib import pyplot as plt
slices=[30,40,20]
labels=['men','women','kids']
plt.title('My PiE CHART')
plt.pie(slices,labels=labels)
plt.show()
```



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Custom colors.. And wedge properties

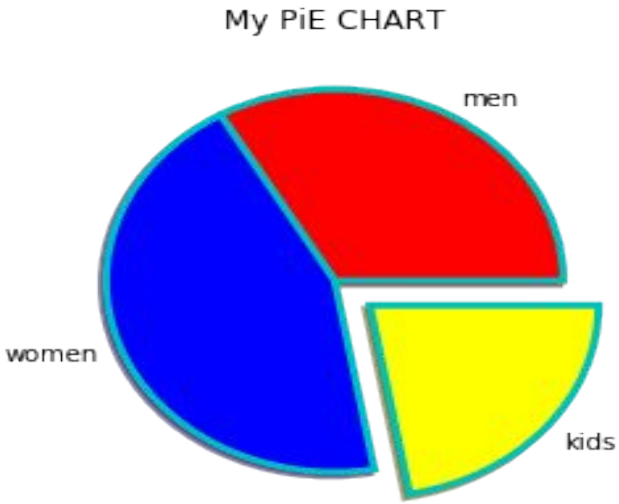
- from matplotlib import pyplot as plt
- slices=[30,40,20]
- labels=['men','women','kids']
- colors=['red','blue','yellow']
- plt.title('My PiE CHART')
- plt.pie(slices,labels=labels,
  colors=colors,
  wedgeprops={'linewidth':3,'edgecolor':'c'})



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Explodes.. And shadow

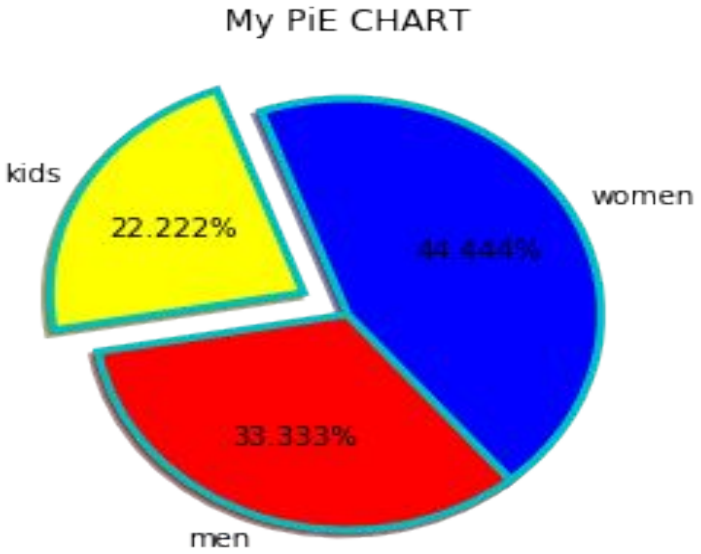
- from matplotlib import pyplot as plt
- slices=[30,40,20]
- labels=['men','women','kids']
- colors=['red','blue','yellow']
- plt.title('My PiE CHART')
- explodes=[0,0,.2]
- plt.pie(slices,labels=labels,colors=colors,
  wedgeprops={'linewidth':3,'edgecolor':'c'},
  explode=explodes,shadow
- =True)



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Startangle and percentage value..

- from matplotlib import pyplot as plt
- slices=[30,40,20]
- labels=['men','women','kids']
- colors=['red','blue','yellow']
- plt.title('My PiE CHART')
- explodes=[0,0,.2]
- plt.pie(slices,labels=labels,colors=colors,
  startangle=190,
  autopct='%1.3f%%',
  wedgeprops={'linewidth':3,'edgecolor':'c'},
  explode=explodes,shadow=True)



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Scatter plots..

- Scatter plots are great useful when you want to show the relationship between two set of datas and to find the correlation between them

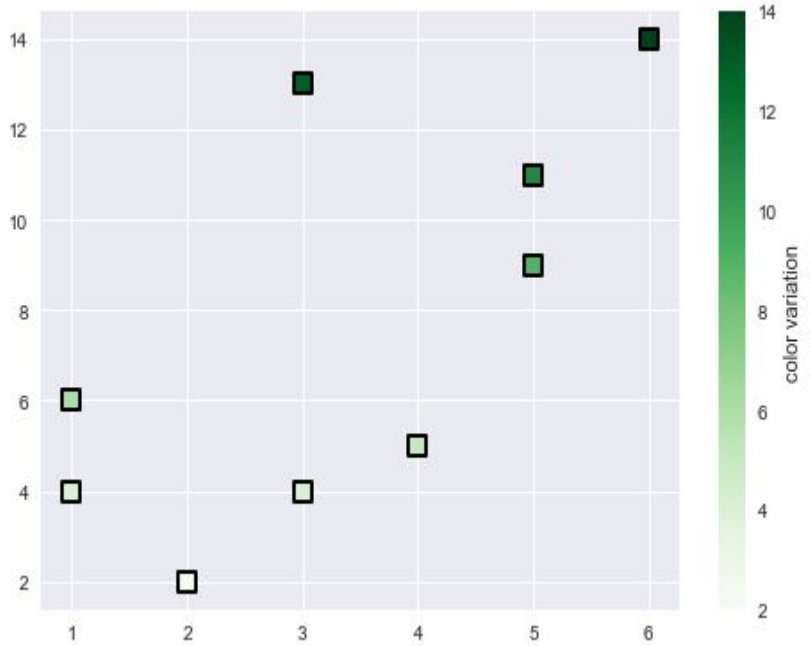
```
a=np.random.randint(1,10,size=9)
b=np.random.randint(2,15,size=9)
plt.scatter(a,b)
```

Scalar plot..

1. a=np.random.randint(1,10,size=9)
2. b=np.random.randint(2,15,size=9)
3. plt.style.use('seaborn')
4. plt.scatter(a,b,s=100,color='red',marker='s',edgecolors='black',linewidths=2)

continue..

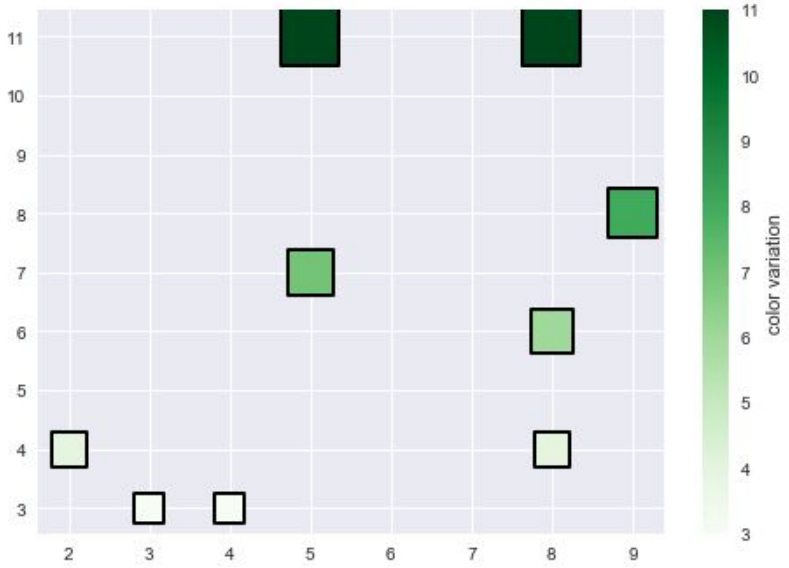
- a=np.random.randint(1,10,size=9)
- b=np.random.randint(2,15,size=9)
- colors=b
- plt.style.use('seaborn')
- plt.scatter(a,b,s=100,marker='s',
- c=colors,cmap='Greens',
- edgecolors='black',linewidths=2)
- cbar=plt.colorbar()
- cbar.set_label('color variation')



TECHOLAS
TECHNOLOGY DEMYSTIFIED

continues...

- a=np.random.randint(1,10,size=9)
- b=np.random.randint(2,15,size=9)
- colors=b
- sizes=b*100
- plt.style.use('seaborn')
- plt.scatter(a,b,s=sizes,marker='s',
- c=colors,cmap='Greens',
- edgecolors='black',linewidths=2)
- cbar=plt.colorbar()
- cbar.set_label('color variation')



TECHOLAS
TECHNOLOGY DEMYSTIFIED

STATISTICS FOR DATA SCIENCE

Statistics

The discipline concerned with the collection, organization, analysis, interpretation and presentation of data.

Mainly there are 2 types of statistics: **descriptive** and **inferential** statistics.

- Descriptive statistics deals with the summarization of the data.
- Inferential statistics deals with the interpretation of the data i.e., the analysis is used to describe the meaning of the data.

What is Data??

Data are measurements or observations that are collected as a source of information.

There are a variety of different types of data, and different ways to represent data.

Eg:

The value of sales of a particular product, or the number of times India has won a cricket match, are all examples of data.

Data..

Data unit: **A data unit is one entity in the set of all data (population)being studied, about which data are collected.** A data unit is also referred to as a unit record or record.

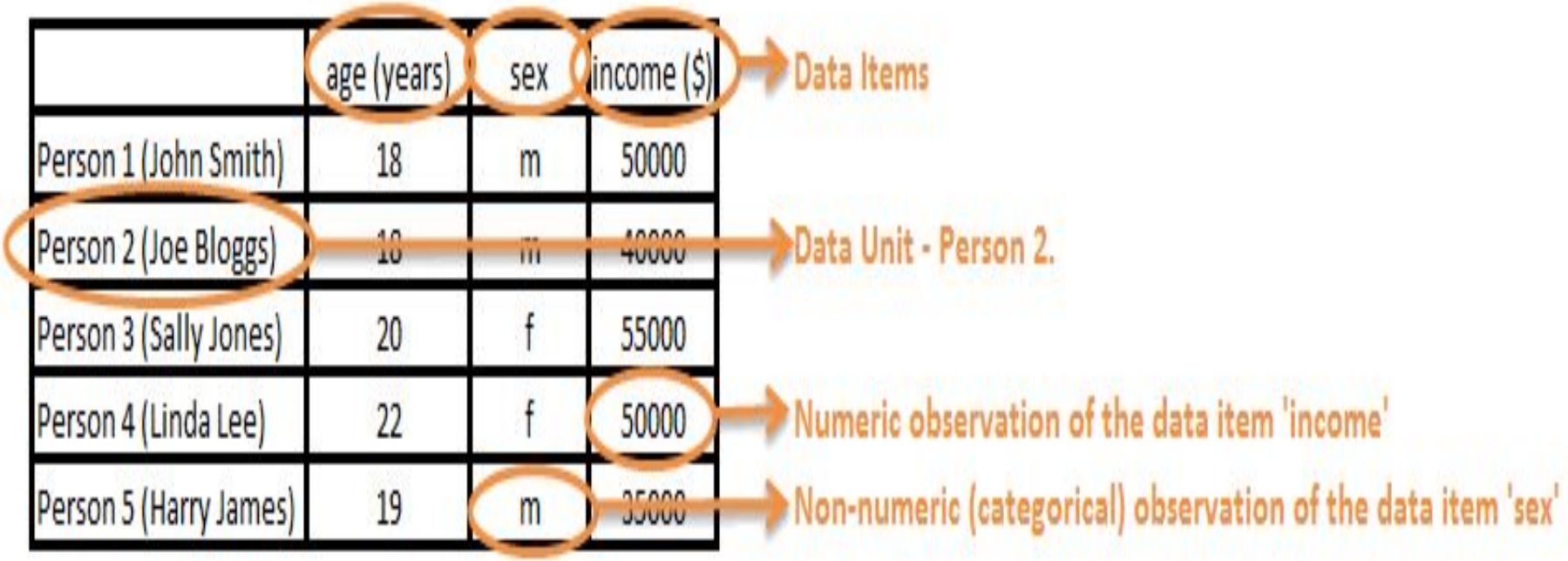
Data Item: A data item is a characteristic (or attribute) of a data unit which is measured or counted, such as height, country of birth, or income. A data item is also referred to as a **variable** because the characteristic may vary between data units, and may vary over time.

Observation:**An observation is an occurrence of a specific data item that is recorded about a data unit**.

DATASET : A dataset is a complete collection of all observations.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

....



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Quantitative and qualitative data..

Quantitative data are measures of values or counts and are expressed as numbers.

Quantitative data are data about **numeric variables** (e.g. how many; how much; or how often).

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

Qualitative data are data about **categorical variables** (e.g. what type).

Quantitative = Quantity

Qualitative = Quality

TECHOLAS
TECHNOLOGY DEMYSTIFIED

VARIABLE..

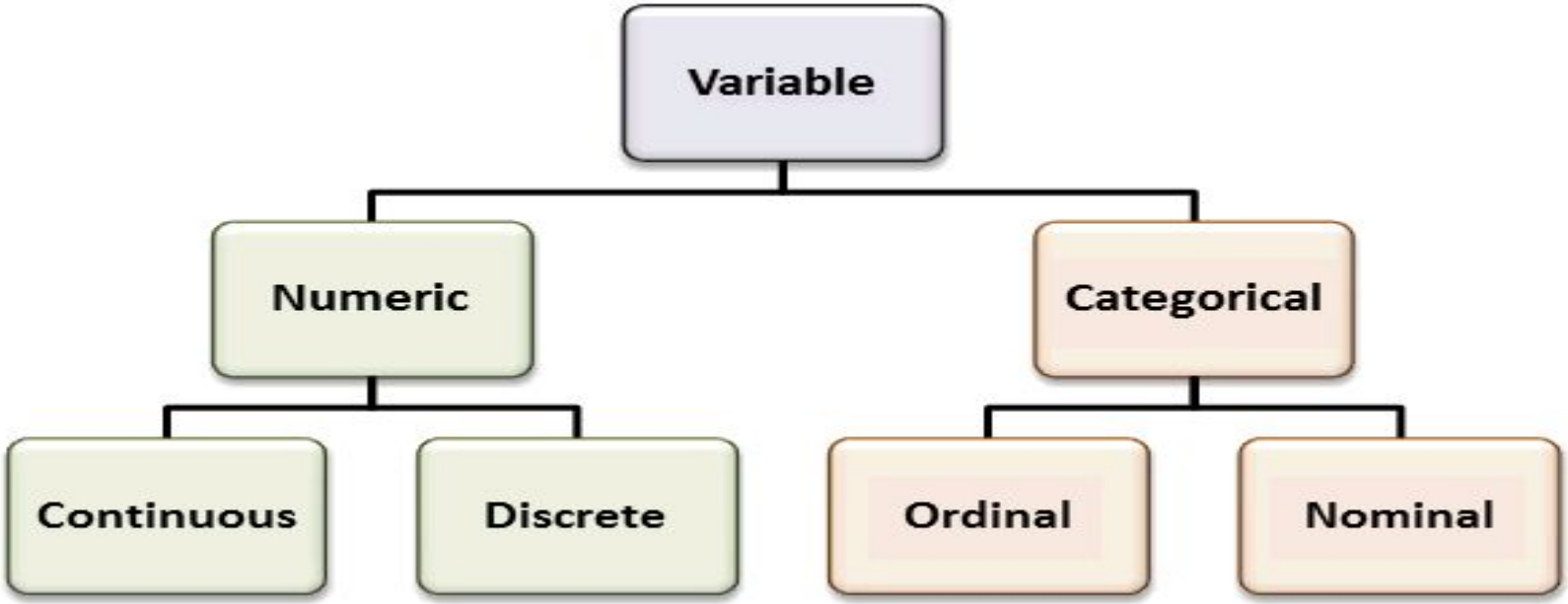
A variable is any characteristics, number, or quantity that can be measured or counted.

A variable may also be called a **data item**.

Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.
It is called a variable because the value may vary between data units

TECHOLAS
TECHNOLOGY DEMYSTIFIED

TYPES oF Variables..



TECHOLAS
TECHNOLOGY DEMYSTIFIED

Numeric Variables..

Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore numeric variables are quantitative variables

- **A continuous variable is a numeric variable. Observations can take any value between a certain set of real numbers.** The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include height, time, age, and temperature.
- **A discrete variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values.** A discrete variable cannot take the value of a fraction between one value and the next closest value. Examples of discrete variables include the number of registered cars, number of business locations, and number of children in a family, all of of which measured as whole units (i.e. 1, 2, 3 cars).

Categorical Variable..

Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'. Categorical variables are qualitative variables and tend to be represented by a non-numeric value.

- **An ordinal variable is a categorical variable. Observations can take a value that can be logically ordered or ranked.** The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category. Examples of ordinal categorical variables include academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra large) and attitudes (i.e. strongly agree, agree, disagree, strongly disagree).
- **A nominal variable is a categorical variable. Observations can take a value that is not able to be organised in a logical sequence.** Examples of nominal categorical variables include sex, business type, eye colour, religion and brand.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Population and samples..

POPULATION: It is set of all values of observations..

Sample: It is a subset of population

- population=np.random.randint(140,190,100)
- sample1=np.random.choice(population,20)
- sample2=np.random.choice(population,25)

SAMPLING

Sampling is the process of selecting the sample from the population

Sampling categorized into 2

1.Probability sampling
2.Non probability sampling

PROBABILITY SAMPLING

Choosing the sample from the large population by using the theory of probability.

1. RANDOM SAMPLING
2. SYSTEMATIC SAMPLING
3. STRATIFIED SAMPLING

RANDOM SAMPLING

In this sampling all the elements have same probability of being selected to form a sample.ie it is choosing randomly

Eg: from a large set of students choosing 10 students randomly.

Systematic sampling.

In systematic sampling every nth element is chosen from different groups of population to form a sample

Eg: if considering indian teenagers as population taking 10 elements from all states is systematic sampling

STRATIFIED SAMPLING..

In this a sufficient number will be selected from each stratum of the population

STRATUM: subset of population that having at least one common behavior

Eg: if considering indian teenagers as population the girls is one stratum and boys are another one and selecting 10 elements from each and creates a sample is stratified sampling..

TYPES OF STATISTICS

1. **DESCRIPTIVE STATISTICS:**descriptive statistics uses data to provide descriptions about population either through numerical calculations or through graph or tables
   It concentrates main characteristics of data
   It provides a graphical summary about data
2. **INFERENTIAL STATISTICS:**It makes inferences and predictions about a population based on the sample.Inferential statistics generalize a huge dataset and apply probability to draw conclusions

TECHOLAS
TECHNOLOGY DEMYSTIFIED

TOOLS FOR DESCRIPTIVE STATISTIC ANALYSIS

1. CENTRAL TENDENCY
2. MEASURE OF SPREAD
3. SKEWNESS
4. KURTOSIS

TECHOLAS
TECHNOLOGY DEMYSTIFIED

MEASURES OF CENTRAL TENDENCY

Central Tendency provides the idea of distribution of data around the central value..

There are 3 kinds of Central Tendencies

1. MEAN
2. MEDIAN
3. MODE

TECHOLAS
TECHNOLOGY DEMYSTIFIED

MEAN..

Mean is the sum of the value of each observation in a dataset divided by number of observations

We use numpy module to find the mean

```
import numpy as np

list=np.random.randint(3,10,20)

list.mean()

np.mean(list)
```

Median..

Median is the middle value of the distribution When the values are arranged in ascending or descending order..

We use median() function in numpy to find the median value

```
list=np.random.randint(3,10,20)
```

```
np.median(list)
```

MODE..

Mode is the most commonly occuring value in the distribution

In numpy there is no direct function to find mode .So we will use statistics module to use it

```
import numpy as np

import statistics as st

list=np.random.randint(3,10,20)

st.mode(list)
```

MEASURE OF SPREAD..

Measures of spread describe how similar or varied the set of observed values are for a particular variable (data item).

Measures of spread include the

- **Range**
- **quartiles and the interquartile range**
- **variance**
- **standard deviation**.

Summarising the dataset can help us understand the data, especially when the dataset is large

Measures of spread summarise the data in a way that shows how scattered the values are and how much they differ from the mean value.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Measure of spread continues..

4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

Consider the above 2 dataset both having the same mean value that is 6

if we look at the spread of the values in the dataset , we can see that second Dataset is more dispersed than first Dataset . Used together, the measures of central tendency and measures of spread help us to better understand the data

TECHOLAS
TECHNOLOGY DEMYSTIFIED

range...

The range is the difference between the smallest value and the largest value in a dataset.

Dataset A
4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Dataset B
1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

DATASET A: The range is 4, the difference between the highest value (8) and the lowest value (4).

DATASET B: The range is 10, the difference between the highest value (11) and the lowest value (1)

range..

```
data=np.array([4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8])
range=data.max()-data.min()
print(range) → 4
```

Quartiles..

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point *between* the quarters.

Quartiles						
25% of values	**Q1**	25% of values	**Q2**	25% of values	**Q3**	25% of values

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Quartile..

DATASET A: [4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8]

Dataset A														
4	5	5	**Q1**	5	6	6	**Q2**	6	6	7	**Q3**	7	7	8

- **As the quartile point falls between two values, the mean (average) of those values is the quartile value:**
- **Q1 = (5+5) / 2 = 5**
- **Q2 = (6+6) / 2 = 6**
- **Q3 = (7+7) / 2 = 7**

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Interquartile range..

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles, and describes the middle 50% of values when ordered from lowest to highest.

The IQR is often seen as a better measure of spread than the range

- The IQR for Dataset A is = **2**
- **IQR = Q3 - Q1**
- **= 7 - 5**
- **= 2**

TECHOLAS
TECHNOLOGY DEMYSTIFIED

QUARTILE .. in notebook

- data=np.array([4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8])
- # first quartile
- q1=np.percentile(data,25)
- # second quartile
- q2=np.percentile(data,50)
- # third quartile
- q3=np.percentile(data,75)

Find the interQuartile range =3rd quartile- 1st quartile

- #interquartile range
- IQR=q3-q1
- print(IQR)

Variance and standard deviation

The variance and the standard deviation are measures of the spread of the data around the mean. They summarise how close each observed data value is to the mean value.

- In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.
- The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Variance equations..

The population **Variance** σ^2 (pronounced *sigma squared*) of a discrete set of numbers is expressed by the following formula:

- Where: X_i represents the *ith* unit, starting from the first observation to the last
- μ represents the population mean
- N represents the number of units in the population

$$\sigma^2 = \frac{\sum_{i-1}^{N}(X_i - \mu)^2}{N}$$

The **Variance** of a sample s^2 (pronounced *s squared*) is expressed by a slightly different formula:

$$s^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1}$$

- x_i represents the *ith* unit, starting from the first observation to the last
- $\bar{x}$ represents the sample mean
- n represents the number of units in the sample

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Standard deviation ..

The **standard deviation** is the square root of the variance. The standard deviation for a population is represented by σ, and the standard deviation for a sample is represented by *s.*

finding the variance and std

Dataset A

- Calculate the population mean (μ) of Dataset A.
  (4 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 8) / 12
  mean (μ) = 6
- Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset
  = -2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2
- Square each individual deviation value
  = 4, 1, 1, 1, 0, 0, 0, 0, 1,1,1, 4
  Calculate the mean of the squared deviation values
  =(4 + 1 +1 +1 + 0 + 0 + 0 + 0 +1 +1 +1 + 4) / 12
  Variance σ^2= 1.17
- Calculate the square root of the variance
  Standard deviation σ = 1.08

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Finding in notebook..

Variance

```
data=np.array([4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8])
np.var(data)
sample=np.random.choice(data,5)
np.var(sample)
```

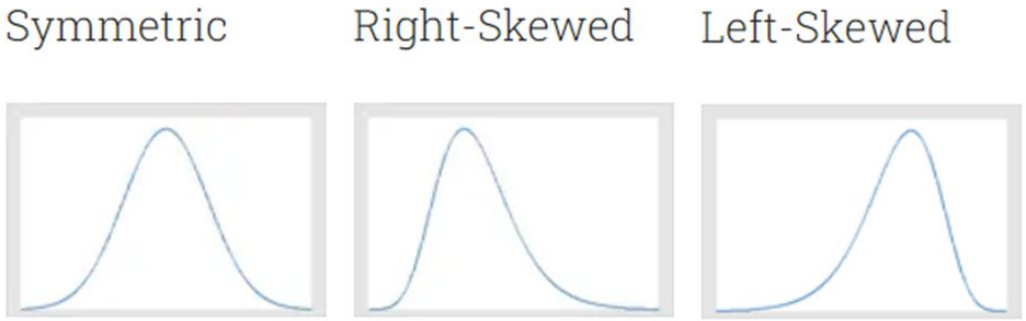
Standard Deviation

```
np.std(data)
np.std(sample)
```

TECHOLAS
TECHNOLOGY DEMYSTIFIED

SKEWNESS

- It defines the asymmetry of a distribution



- Right skewed distributions are also said to be positively skewed. This occurs because probabilities taper off more slowly for higher values. Thus, the extreme values are found far from the peak on the high end more frequently than on the low.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

- Left skewed distributions are also said to be negatively skewed. This condition occurs because probabilities taper off more slowly for lower values. Thus, the extreme values are found far from the peak on the low side more frequently than the high side.
- The long tail is the place where we find majority of the exceptional values.

KURTOSIS

- Kurtosis is the measure of skewness of a distribution. Kurtosis measures extreme values in either tail.
- Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution.
- Distributions with low kurtosis exhibit tail data that are generally less extreme than the tails of the normal distribution.

TOOLS FOR INFERENTIAL STATISTIC ANALYSIS

- Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn.
- The goal of inferential statistics is to draw conclusions from a sample and generalize them to a population.
- Thus we need to have confidence that our sample accurately reflects the population.
- Different sampling techniques can be used to select the appropriate group to represent the population aptly.
- Common inferential statistical tools are:
    1. HYPOTHESIS TESTING
    2. CONFIDENCE INTERVAL
    3. REGRESSION ANALYSIS

TECHOLAS
TECHNOLOGY DEMYSTIFIED

HYPOTHESIS TESTING

- **Hypothesis testing:** A hypothesis is an educated guess about something based on limited evidence as a starting point for further investigation. It should be testable by experiment or observation.
- A random population sampling is done to test 2 different hypothesis: the **null hypothesis** and the **alternate hypothesis**.
- The null hypothesis denoted as H_0 is a hypothesis of equality between the population parameters. The alternate hypothesis is the opposite of a null hypothesis.
- Thus they are mutually exclusive and only one of them can be true and one of them will always be true.

CONT...

Eg: We want to test that a coin has exactly a 50% chance of landing on heads. Then,

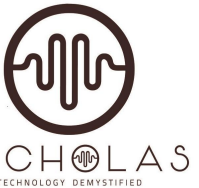
H_0: P = 0.5

H_a: P≠ 0.5

A random sample of 100 coin flips is taken, and the null hypothesis is then tested. If it is found that the 100 coin flips were distributed as 40 heads and 60 tails, then the coin does not have a 50% chance of landing on heads and would reject the null hypothesis and accept the alternative hypothesis.

CONFIDENCE INTERVAL

- Confidence interval(CI) is a range of values that is likely to contain the value of an unknown population parameter. Different random samples drawn from the same population are likely to produce slightly different intervals.
- The **confidence level** is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval.
- A confidence interval is the mean of our estimate plus and minus the variation in that estimate.
- If we construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.



CONT...

Eg: A survey of 100 Brits and 100 Americans about their television-watching habits, found out that both groups watch an average of 35 hours of television per week.

However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.

REGRESSION ANALYSIS

Regression analysis is used to understand the relationship between two variables (X and Y) in a data set as a way to estimate the unknown variable to make future projections on events and goals. This helps estimating the value of a random variable based on the values of some known variables.

PROBABILITY

- A number that reflects the chance or likelihood that a particular event will occur.
- Probabilities can be expressed as proportions that range from 0 to 1, and they can also be expressed as percentages ranging from 0% to 100%.
- Probability 0 : Impossible event
- Probability 1 : Sure event

Probability is calculated as :

P(event) = Number of favorable outcomes/ total no of outcomes

CONT...

Study of obesity in children 5-10 years of age who are seeking medical care

	Age (years)						
	5	**6**	**7**	**8**	**9**	**10**	**Total**
Boys	432	379	501	410	420	418	2,560
Girls	408	513	412	436	461	500	2,730
Totals	840	892	913	846	881	918	5,290

If we select a child at random, each child has the same probability of being chosen and is equal to 1/5290 = 0.0002. Such outcomes are called **equally likely** outcomes.

CONT...

Q. What is the probability of selecting a girl?

Q. What is the probability of selecting a 7 year-old?

Q. What is the probability of selecting a boy who is 10 years of age?

Q. What is the probability of selecting a child (boy or girl) who is at least 8 years of age?

All these are cases of unconditional probability. Everyone in the entire population has equal chance of being selected and no other factor can affect its selection.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONDITIONAL PROBABILITY

- Conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred.
- Denoted as P(B|A) and read as *probability of B given A.*

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

- If 2 events are independent, the conditional probability of event B given event A is *P(B).*
- Bayes' theorem for finding conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A).P(B|A)}{P(B)}$$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Eg1: In a card game, suppose a player needs to draw two cards of the same suit in order to win. Of the 52 cards, there are 13 cards in each suit.

Suppose first the player draws a heart. Now the player wishes to draw a second heart.

Since one heart has already been chosen, there are now 12 hearts remaining in a deck of 51 cards.

So the conditional probability

$P(\text{Draw second heart} \mid \text{First card a heart}) = 12/51$.

Eg2: Consider rolling a fair die. Let A be the event that the outcome is an odd number, i.e., A={1,3,5}. Also let B be the event that the outcome is less than or equal to 3, i.e., B={1,2,3}.

Then $P(A \mid B) = 2/3$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Considering the previous example,

Q. What is the probability of selecting a 9 year old from the sub-population of girls?

RANDOM VARIABLES

- A random variable is a numerical description of the outcome of a statistical experiment.

- A random variable taking a finite number or infinite sequence of values is said to be **discrete**.

Eg: Number of automobiles sold in a dealership

- A random variable taking any value in an interval is said to be **continuous**.

Eg: Weight of a person in kilograms.

PROBABILITY DISTRIBUTIONS

They are statistical functions that describe the likelihood of obtaining possible values that a random variable can take.

General properties of a probability distribution

i. Probability of an event will be non negative.

ii. The probability for a particular value or range of values must be between 0 and 1.

iii. The sum of all probabilities for all possible values must equal 1.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Probability distribution of discrete random variables

The probability distribution of a discrete random variable is the list of all possible values of the variable and their probabilities which sum to 1.

The cumulative probability distribution function is the probability that the random variable is less than or equal to a particular value.

Consider the rolling of a die.

Sample space = {1,2,3,4,5,6}

This contains 6 mutually exclusive outcomes. Let x be the random variable that denotes each point in the sample space.

CONT...

Then **P(X=x)** where x = {1,2,3,4,5,6}, is the probability that the random variable takes a value from the sample space and is called the probability mass function (PMF).

CDF is given as **P(X ≤x)** where x = {1,2,3,4,5,6}

PDF and CDF of a rolling die

Outcome	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6
Cumulative Probability	1/6	2/6	3/6	4/6	5/6	1

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Probability distribution of continuous random variables

The probability distribution for a continuous random variable is defined by the probability density function.

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{dx}$$

Cumulative distribution function is defined in the same manner as discrete random variable.

$$P(X \leq b) = \int_{-\infty}^b f(x)\,\mathrm{dx}$$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Types of probability distributions

Bernoulli distribution

- A bernoulli distribution has only 2 outcomes; success(1) or failure(0).
- A random variable X with bernoulli distribution can take only 2 values: *p* (success) and *q* (failure) where
- *q = 1-p*
- The values of p and q need not be the same(i.e., need not be equally likely)
- The probability mass function is given by

$$p(x) = \begin{cases} p & ; x = 1 \\ 1 - p & ; x = 0 \end{cases}$$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

- Expected value (mean) of X is **$E(X)=(1*p)+(0*(1-p))=p$**
- Variance of X is **$V(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)$**
- All events that have exactly 2 outcomes can be expected to have a bernoulli distribution.

CONT...

Uniform distribution (Rectangular distribution)

- All the possible outcomes of a uniform distribution are equally likely.
- Probability of getting a number from 1 to 6 when rolling a die is a uniform distribution.
- Its probability mass function is

$$f(x) = \frac{1}{b-a} \quad for \ -\infty < a \leq x \leq b < \infty$$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Mean

$$E(X) = \frac{a+b}{2}$$

Variance

$$V(X) = \frac{(b-a)^2}{12}$$

- A standard uniform density has parameters a=0 and b=1. So the PDF is given by

$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Geometric distribution

- It is a type of discrete probability distribution that represents the probability of the number of successive failures before a success is obtained in a Bernoulli trial.
- In other words, in a geometric distribution, a Bernoulli trial is repeated until a success is obtained and then stopped.
- It is based on 3 important assumptions:
  - The trials being conducted are independent
  - There can only be 2 outcomes; success or failure.
  - The success probability is same for each trial.
- A geometric distribution can have an indefinite number of trials until the first success is obtained.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

- Example:

A die repeatedly rolled until a 3 is obtained.

Here p=1/6 and the random variable X can take any value from 1 to 6 until the first success is obtained.

- PMF: $P(X = x) = (1 - p)^{(x - 1)}\ p$ where $0 < p \leq 1$
- CDF: $P(X \leq x) = 1 - (1 - p)^x$
- Mean, $E(X) = 1/p$
- Variance, $V(X) = (1 - p)/p^2$

CONT...

Binomial distribution

- A distribution with only 2 possible outcomes: success or failure where the probability of success or failure is same for all trials is a binomial distribution.
- The outcomes need not be equally likely.
- Consider a fight between you and wrestling champion; probability of you winning can be 0.2 and probability of loosing will be 1-0.2=0.8.
- The parameters of a binomial distribution are n and p; n is the total number of trials and p is the probability of success in each trial.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

- Properties of a binomial distribution:
  - Each trial is independent
  - Only 2 possible outcomes are there; success or a failure
  - A total number of n identical trials are conducted.
  - The probability of success and failure is same for all trials.
- A binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

- Mean E(x) = np
- Variance V(x) = npq

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Poisson distribution

- When events occur at random points of time and space and the no of occurrences of the event is the only thing that matters, then it follows a poisson distribution.
- Eg: No of customers arriving in a restaurant, no of emergency calls recorded in a hospital in a day, etc.
- A distribution is called poisson when the following are valid:
  - Any successful event should not influence the outcome of another successful event.
  - The probability of success over a short interval must equal the probability of success over a longer interval.
  - The probability of success in an interval approaches zero as the interval becomes smaller.

CONT...

- PMF of poisson distribution

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}; \mathrm{x} = 0,1,2 \ldots.$$

  where λ is the rate at which an event occurs and X is the number of events in a time interval t.

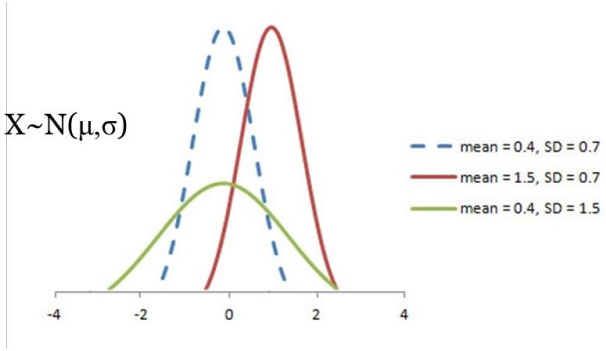
- Mean E(x) = λ
- Variance V(x) = λ

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Normal distribution

- Any distribution with the following features can be a normal distribution:
  - The mean , median and mode of the distribution coincide
  - A bell-shaped distribution curve symmetrical about the line x = μ.
  - Total area under the curve =1
  - Exactly half of the values are to the left of the center and the other half to the right.



TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

- PDF of normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)} for \ -\infty < x < \infty$$

- Mean E(x) = μ
- Variance V(x) = σ^2
- When mean is 0 and standard deviation is 1, the distribution is called a standard normal distribution. Then the PDF will be

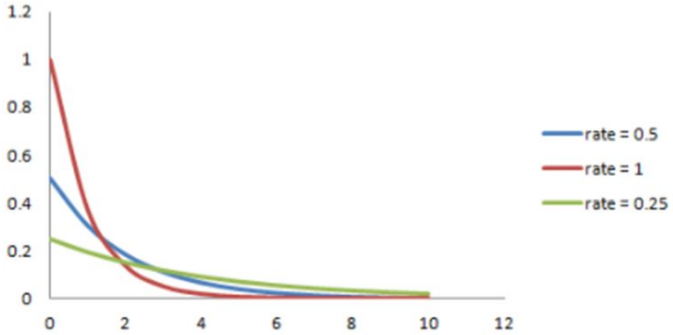
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{(\frac{-x^2}{2})} for \ -\infty < x < \infty$$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Exponential distribution

- It is the probability distribution of the time between the events in a poisson process.
- The amount of time until the event occurs means during the waiting period, not a single event has happened. This is Poisson (X=0)
- Eg: Length of time between metro arrivals, interval of time between the calls in a call centre etc.



TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

- PDF of an exponential distribution:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x \end{cases}$$

where $\lambda > 0$ is called the rate

- Mean $E(x) = 1/\lambda$
- Variance $V(x) = 1/\lambda^2$

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Covariance and Correlation

$$Cov(x,y) = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

$$Correlation = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$$

Covariance	Correlation
Measure of how 2 random variables change together	Measure that represents how strongly 2 random variables are related to each other
Measure of correlation	Scaled form of covariance
Indicates the direction of linear relationship between variables	Measures the strength and direction of linear relationship between 2 variables
They can vary between -∞ to +∞	They vary between -1 and +1
Zero for independent variables	Zero for independent variables

Central Limit Theorem (CLT)

- *It states that the distribution of a sample variable approximates a normal distribution (bell curve) as the sample size becomes larger, regardless of the population's actual distribution shape.*
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- The average of the sample means and standard deviations will equal the population mean and standard deviation which helps in accurately predicting the characteristics of populations.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

Describing Frequencies

- Frequency is the number of times a particular value for a variable has occurred.
- Measurement of frequencies can be done in different ways.
- **Absolute frequency**: It is the number of times a particular value for a variable has occurred. It is the simplest way to represent frequency.
- **Relative frequency**: It is the number of times a particular value for a variable has been observed to occur in relation to the total number of values for that variable. It is calculated by dividing the absolute frequency by the total number of values for the variable.

CONT...

Methods of expressing relative frequency

- ★ ***Ratio*** : It compares the frequency of one value for a variable with another value for the variable.

  Eg: In a total of 20 coin tosses where there are 12 heads and 8 tails, the ratio of heads to tails is 12:8.

  Similarly, the ratio of tails to heads is 8:12.

- ★ ***Rate:*** It is a measurement of one value for a variable in relation to another measured quantity.

  Eg: In a total of 20 coin tosses where there are 12 heads and 8 tails, the rate is 12 heads per 20 coin tosses. Similarly, the rate is 8 tails per 20 coin tosses.

CONT...

- ***Proportion:*** It describes the share of one value for a variable in relation to a whole. It is calculated by dividing the number of times a particular value for a variable has been observed, by the total number of values in the population.

  Eg: in a total of 20 coin tosses where there are 12 heads and 8 tails, the proportion of heads is 0.6 (12 divided by 20).

  Similarly, the proportion of tails is 0.4 (8 divided by 20).

- ***Percentage***: It expresses a value for a variable in relation to a whole population as a fraction of one hundred.

  Eg: In a total of 20 coin tosses where there are 12 heads and 8 tails, the percentage of heads is 60% (12 divided by 20, multiplied by 100). Alternatively, the percentage of tails is 40% (8 divided by 20, multiplied by 100).

MEASURE OF ERRORS

Absolute (Standard error)

It is the absolute difference between the measured value and true value of a quantity.

Random error

An error is considered random if the value of what is being measured sometimes goes up or sometimes goes down. Eg. Blood pressure of a healthy person may go up and down each time it is measured.

TECHOLAS
TECHNOLOGY DEMYSTIFIED

CONT...

Constant error

It cause measurements to deviate constantly from the true value. Eg: measuring device errors.

Relative error

It is defined as the ratio of the absolute error of the measurement to the actual measurement.

Percentage error

It is relative error expressed in percentage.

TECHOLAS
TECHNOLOGY DEMYSTIFIED