PANDAS

What is Pandas?

- Pandas is a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool, built on top of the Python programming language.
- about pandas: https://pandas.pydata.org/about/index.html

Use of Pandas

- Data Cleaning and Preparation:
  Pandas provides tools for handling missing data, data alignment, transformation, and data normalization.
- Time Series Analysis:
  Pandas offers specialized functionality for working with time series data.
  Pandas integrate well with visualization libraries like Matplotlib and Seaborn which are used for Data Visualization.

Basic data structure in pandas

- Pandas provide two types of classes for handling data:
  ~ Series
  ~ DataFrame

Importing Pandas

There are different way to import pandas library:

1.Most common way is import statement with alias:

```
~ import pandas as pd
```

2.Importing all the function and class:

```
~ from pandas import *
```

3.Importing the specific function and class library:

```
~ from pandas import DataFrame, read_csv
```

Pandas Series

- A Pandas Series is like a column in a table.
- It is a one-dimenational labelled array capable of holding of any datatype.
- Integers, string, floating point numbers, python object etc.
- Each value in a pandas series is associated with the index.
- The default index value of it is from 0 to number – 1, or you can specify your own index values.

Note:

- 'Series' is a class provided by the panda's library.
- When you create a 'Series' in your code, means your creating an 'Series' object of that 'Series' class.
- Syntax: pd.Series(data, index, name)

Parameteres:

data:	It can be array, list, dictionary, csv file or excel file
index:	Default, also create custom index
name:	Series name

Creating a pandas series

- Pandas series can be created from the lists, dictionary and from other scalar values etc.
- Series can be created in different ways, here are some ways to create a series.

Create an empty series:

Code:

```
import pandas as pd
empty_series = pd.Series()
print(empty_series)
```

Output:

```
Series([ ], dtype: float64)
```

Create a series from lists:

We first creating a list after that we can create series of list.

```
# String
```

Code:

```
import pandas as pd
import numpy as np
countries = [ 'India' , 'Nepal' , 'Srilanka' , 'Bhutan' ]
print(pd.Series(countries))
```

Output:

```
0       India
1       Nepal
2    Srilanka
3      Bhutan
   dtype: object
```

```
# Custom index and name
```

Code:

```
marks = [ 58, 93, 89, 60 ]
subjects = [ 'C++' , 'Python', 'R', 'Java' ]
print(pd.Series(marks, index=subjects, name='student'))
```

Output:

```
C++       58
Python    93
R         89
Java      60
Name: student, dtype: int64
```

Create a series from dictionary:

We must first create a dictionary then we can pefrom series on dictionary.

Code:

```
import pandas as pd
marks = { 'maths' : 78,  'english' : 70, 'science' : 89 }
pd.Series(marks, name='student score')
```

Output:

```
maths    78
english  70
science  89
Name: student score, dtype: int64
```

Pandas Series Attribute

In Pandas, a Series object has several important attributes that is commonly used attributes of a Pandas Series include:

size attribute:

- Returns the number of elements in a Series, including any elements that might contain missing or NaN (Not-a-Number) values.

Code:

```
import pandas as pd
data = [ 10, 20, 30, None, 50 ]
print(pd.Series(data).size)
data1 = [ 10, 20, 30, 50 ]
print(pd.Series(data1).size)
```

Output:

```
5
4
```

dtype attribute:

- Returns the data type of the elements in the Series.
- Used to check the data type of the data within the Series.

Code:

```
import pandas as pd
data = [ 10, 20, 30, 50 ]
print(pd.Series(data).dtype)
```

Output:

```
int64
```

name attribute:

- Assign a name to a eries when creating it or later using the name attribute.
- The name is typically used in the context of DataFrames, where a Series can represent a column.

Code:

```
import pandas as pd
data = [ 10, 20, 30, 50 ]
print(pd.Series(data, name='my_data').name)
```

Output:

```
my_data
```

is_unique attribute:

- Returns a boolean value indicating whether all the values in the Series are unique (no duplicates) or not.

Code:

```
import pandas as pd
data = [ 10, 20, 30, 40, 50 ]
print(pd.Series(data).is_unique)
data1 = [ 10, 20, 30, 40, 50 ]
print(pd.Series(data1).is_unique)
```

Output:

```
True
False
```

values attribute:

- Returns the data in the Series as a NumPy array.

Code:

```
import pandas as pd
marks = { 'maths' : 78,  'english' : 70, 'science' : 89 }
print(pd.Series(marks).values)
```

Output:

```
[ 78 70 89 ]
```

Series using read_csv() function:

Parameters:

squeeze=True:

- Attribute specifies that the result should be squeezed into a series if it has only one column.
- Here we intentionally use this attribute to avoid the dataframe to understand the about series.

Code:

```
import pandas as pd
subs = pd.read_csv('subs.csv', squeeze=True )
print(subs)
print(type(subs))
```

Output:

```
0       48
1       57
        ...
363   144
364   172
Name: Subscribers gained, Length: 365, dtype: int64
<class 'pandas.core.series.Series'>
```

index_col:

- Used to specify which column in the CSV file should be used as the index for the resulting DataFrame/Series.
- The index is a way to uniquely identify each row in the DataFrame/Series.
- By default, if you don't specify the index_col parameter, Pandas will create a default integer index starting from 0.

Code:

```
import pandas as pd
vk = pd.read_csv('kohli_ipl.csv', index_col= 'match_no',
                 squeeze=True )
print(vk)
print(type(vk))
```

Output:

```
match_no
1           1
2          23
           ..
214       25
215        7
Name: runs, Length: 215, dtype: int64
<class 'pandas.core.series.Series'>
```

Code:

```
import pandas as pd
m = pd.read_csv('bollywood.csv', index_col= 'movie',
    squeeze=True )
print(m)
print(type(m))
```

Output:

```
movie
Uri: The Surgical Strike         Vicky Kaushal
Battalion 609                    Vicky Ahuja
                                     ...
Company (film)                    Ajay Devgn
Awara Paagal Deewana         Akshay Kumar
Name: lead, Length: 1500, dtype: object
<class 'pandas.core.series.Series'>
```

Note: Above dataset used to perform series methods

Series Methods

These methods are most used in pandas series.

head() method:

- Used to display the first few rows (default is 5) of a DataFrame or Series.
- If we provide negative number as parameter then return all rows.

Code:

```
import pandas as pd
print(vk.head()) # default is 5
print(vk.head(2))
```

Output:

```
match_no
1       1
2      23
3      13
4      12
5       1
Name: runs, dtype: int64
match_no
1       1
2      23
Name: runs, dtype: int64
```

tail() method:

- Used to display the last few rows (default is 5) of a DataFrame or Series.

Code:

```
import pandas as pd
print(vk.tail())
print(vk.tail(2))
```

Output:

```
match_no
211       0
212      20
213      73
214      25
215       7
Name: runs, dtype: int64
match_no
214      25
215       7
Name: runs, dtype: int64
```

sample() method:

- To randomly select a specified number of rows (default is 1) or elements from a dataframe or series.
- Useful to obtain a random sample from your data for data exploration, analysis, or testing.
- Ex. if you have biased in your datasets then we can use this method.

Code:

```
import pandas as pd
print(vk.sample()) # default 1
print(vk.sample(3)) # three random row from dataset
```

Output:

```
match_no
118     33
Name: runs, dtype: int64
match_no
202      5
150      8
110     82
Name: runs, dtype: int64
```

value_counts() method:

- To count the frequency of values that occur multiple times in a series.

Code:

```
import pandas as pd
print(m.value_counts())
```

Output:

```
Akshay Kumar          48
Amitabh Bachchan      45
                      ..
Akanksha Puri          1
Edwin Fernandes        1
Name: lead, Length: 566, dtype: int64
```

sort_values() method:

- Used to sort the values within a pandas series.
- By default, it sorts the values in ascending order, but you can specify the sorting order using the ascending parameter.
- **Syntax:** series.sort_values(ascending=True, inplace=False)

Parameters:

inplace:	when true then sort and replace sorted data with original data If false (default), it returns a new

	series with the sorted values while leaving the original series unchanged
ascending:	True (ascending) or False (descending)

Code:
```
import pandas as pd
print(vk.sort_values()) # default is ascending=True
```
Output:
```
match_no
135      0
8        0
        ..
126    109
128    113
Name: runs, Length: 215, dtype: int64
```

```
# with inplace=True for permanent changes in series
```
Code:
```
print(vk.sort_values(inplace=True)
print(vk)
```
Output:
```
match_no
128    113
126    109
       ...
8        0
135      0
Name: runs, Length: 215, dtype: int64
```

Method chaining:
- It is practice of applying multiple operations or methods to a Series in a single line of code.
- This approach is both efficient and readable, making it easier to perform complex data manipulations and transformations.

Code:
```
import pandas as pd
print(vk.sort_values(ascending=False).head(1).values[0])
```
Output:
```
113
```

sort_index() method:
- It similar in concept to the sort_values() method, but in this method it sorts the index (row labels) of the Series.
- Both methods allow you to control the sorting order, either ascending or descending, and both can be used with the inplace parameter to m odify the original Series.
- syntax: series.sort_index(ascending=True, inplace=False)

Code:
```
import pandas as pd
print(vk.sort_index()) # default is ascending=True
```
Output:
```
match_no
1        1
2       23
       ...
214     25
215      7
Name: runs, Length: 215, dtype: int64
```

```
# descending with inplace=True for permanent changes
```
```
print(vk.sort_index(ascending=False, inplace=True)
print(vk)
```
Output:
```
match_no
215      7
214     25
        ..
2       23
1        1
Name: runs, Length: 215, dtype: int64
```

Series Mathematical Methods:

Common statistical methods in Pandas Series for analyzing data:

count() method:
- Count the non-null elements in the Series.

Code:
```
print(vk.count())
print(subs.count())
```
Output:
```
215
365
```

sum() method:
- Used to calculate the sum of all the elements in a Series.

Code:
```
print(subs.count())
```
Output:
```
49510
```

product() method :
- Used to calculate the product of all elements in the Series.
- It multiplies all the values together and returns the result.

Code:
```
print(subs.count())
print(vk.subs)
```
Output:
```
215
```

mean() method:
- Calculates the mean (average) of the elements in a Series.

Code:
```
import pandas as pd
print('avg yt subs of channel:', subs.mean())
print('avg ipl runs of virat kohli:', vk.mean())
```
Output:
```
avg yt subs of channel: 135.64383561643837
avg ipl runs of virat kohli: 30.855813953488372
```

median() method:
- Calculates the median of the elements in a Series, which is the middle value when the data is sorted.
- It is a measure of central tendency.

Code:
```
import pandas as pd
print(subs.median())
print(vk.median())
```
Output:
```
123.0
24.0
```

mode() method:
- Returns the mode(s) of the elements in a Series, which is the most frequently occurring value(s).

Code:
```
import pandas as pd
print('Most frequent lead actor :', m.mode())
```
Output:
```
Most frequent lead actor : 0    Akshay Kumar
Name: lead, dtype: object
```

std() method:
- Computes the standard deviation of the elements in a Series, which measures the spread or dispersion of the data.

Code:
```
import pandas as pd
print(vk.std())
```
Output:
```
26.22980132830278
```

var() method:
- Calculates the variance of the elements in a Series, which is the average of the squared differences from the mean.

Code:
```
import pandas as pd
print(subs.var())
print(vk.var())
```
Output:
```
3928.1585127201565
688.0024777222343
```

min() method:

- Returns the minimum value in a Series or DataFrame.

Code:

```
import pandas as pd
print('minimum subs: ', subs.min())
print('minimum runs: ', vk.min())
```

Output:

```
minimum subs: 33
minimum runs: 0
```

max() method:

- The max() method returns the maximum value in a Series or DataFrame.

Code:

```
import pandas as pd
print('maximum subs: ', subs.max())
print('maximum runs: ', vk.max())
```

Output:

```
maximum subs: 396
maximum runs: 113
```

describe() method:

- It is a convenient function to generate descriptive statistics of a numeric Series.
- It provides a summary of various statistical measures, giving you insights into the data's distribution and central tendency.
- It's provides the following statistics:

count:	number of non-null elements in the Series.
mean:	mean (average) of the Series.
std:	standard deviation, which measures the spread of the data.
min:	minimum value in the Series.
25%:	25th percentile (lower quartile).
50%:	median (50th percentile).
75%:	75th percentile (upper quartile).
max:	maximum value in the Series.

Note: the describe() works on numeric data

Code:

```
import pandas as pd
print(subs.var())
print(vk.var())
```

Output:

```
count    215.000000
mean      30.855814
std       26.229801
min        0.000000
25%        9.000000
50%       24.000000
75%       48.000000
max      113.000000
Name: runs, dtype: float64

count   365.000000
mean    135.643836
std      62.675023
min      33.000000
25%      88.000000
50%     123.000000
75%     177.000000
max     396.000000
Name: Subscribers gained, dtype: float64
```

Some Important Series Methods/Functions:

These are some common methods and functions available for working with Pandas Series in Python:

astype() method:

- This method is used to cast the data type of the elements in a Series to the specified data type (e.g., int, float, str).
- Useful to reduce the memory space
- Syntax: series.astype(dtype)

Code:

```
import pandas as pd
import sys
print('Original size of dataset:',sys.getsizeof(vk))
vk_size =vk.astype('int32')
print('Reduce size of dataset:',sys.getsizeof(vk_size))
```

Output:

```
Original size of dataset: 3456
Reduce size of dataset: 2596
```

between() method:

- Checks if each element in the Series falls within the specified range.
- Returns a boolean Series.
- Syntax: series.between(left, right, inclusive=True)

Code:

```
import pandas as pd
print(vk.between(95,110)) # return boolean values
print(vk[vk.between(95,110)]) # printing values
```

Output:

```
match_no
1      False
2      False
       ...
214    False
215    False
Name: runs, Length: 215, dtype: bool
match_no
82      99
120    100
123    108
126    109
164    100
Name: runs, dtype: int64
```

clip() method:

- Clips values in the Series to be within the specified lower and upper bounds.
- Syntax: series.clip(lower, upper)

Code:

```
import pandas as pd
print(subs.clip(100,160))
```

Output:

```
0      100
1      100
       ...
363    144
364    160
Name: Subscribers gained, Length: 365, dtype: int64
```

drop_duplicates() method:

- Removes duplicate values from the Series.
- Syntax: series.drop_duplicates(keep='first', inplace=False)

Code:

```
import pandas as pd
temp = pd.Series([1,1,3,3,3,5,5])
# default first
print(temp.drop_duplicates())
# deleting first occurance
print(temp.drop_duplicates(keep='last'))
```

Output:

```
0    1
2    3
5    5
dtype: int64
1    1
4    3
6    5
dtype: int64
```

duplicated() function:

- Used to identify and mark duplicate values in a Series (column) of a DataFrame.
- It returns a Boolean Series.
- Syntax: Series.duplicated()

Code:

```
import pandas as pd
temp = pd.Series([1,1,3,3,3,5,5])
print(temp.duplicated()) # True means duplicate
print('Duplicate value count:',temp.duplicated().sum())
```

Output:

```
0    False
1     True
2    False
3     True
```

```
4     True
5    False
6     True
dtype: bool
Duplicate value count: 4
```

isnull() method:

- Returns a boolean Series indicating whether each element is NaN (missing data).
- Syntax: series.isnull()

Code:

```
import pandas as pd
import numpy as np
temp = pd.Series([1,3,np.nan,np.nan,5,np.nan,7,np.nan])
print(temp.isnull()) # return boolean
print('Missing values:',temp.isnull().sum())
```

Output:

```
0    False
1    False
2     True
3     True
4    False
5     True
6    False
7     True
dtype: bool
Missing values: 4
```

dropna() method:

- Removes missing (NaN) values from the Series.
- Syntax: series.dropna(axis=0, inplace=False)

Code:

```
import pandas as pd
import numpy as np
temp = pd.Series([1,3,np.nan,np.nan,5,np.nan,7,np.nan])
print(temp.dropna())
```

Output:

```
0  1.0
1  3.0
4  5.0
6  7.0
dtype: float64
```

fillna() method:

- Fills missing (NaN) values in the Series with the specified value.
- Syntax: series.fillna(value)

Code:

```
import pandas as pd
import numpy as np
temp = pd.Series([1,3, np.nan,5,np.nan,7,np.nan])
print(temp.fillna())
```

Output:

```
0  1.0
1  3.0
2  0.0
3  5.0
4  0.0
5  7.0
6  0.0
dtype: float64
```

isin() method:

- Checks if each element in the Series is in the provided list of values.
- Returns a boolean Series.
- Syntax: series.isin(values)

Code:

```
import pandas as pd
print(vk.isin([49,99])) # return Boolean
print(vk[vk.isin([49,99])]) # printing values
```

Output:

```
match_no
1      False
2      False
       ...
214    False
215    False
Name: runs, Length: 215, dtype: bool
match_no
82     99
86     49
Name: runs, dtype: int64
```

apply() method:

- Applies a given function to each element in the Series and returns a new Series with the results.
- Syntax: series.apply(func)

Code:

```
import pandas as pd
print(m.apply(lambda x:x.split()[0].upper()).head(5))
```

Output:

```
movie
Uri: The Surgical Strike                   VICKY
Battalion 609                              VICKY
The Accidental Prime Minister (film)       ANUPAM
Why Cheat India                            EMRAAN
Evening Shadows                            MONA
Name: lead, dtype: object
```

EXTRA SERIES METHOD THAT IS USED IN TASK:

to_numeric() method:

- Used to convert the values in a Series (or DataFrame column) to numeric data types.
- Useful when you have a Series containing strings or other non-numeric data, and you want to convert them to numeric types like integers or floating-point numbers.
- Syntax: pd.to_numeric(series, errors='coerce', downcast='integer')

Parameters:

errors:	'raise'	Raises an error if any value cannot be converted to a number.
	'coerce'	Replaces non-convertible values with NaN.
	'ignore'	Ignores non-convertible values

Code:

```
import pandas as pd
data = pd.Series(['1', '2', '3.14', 'hello', '5'])
numeric_data = pd.to_numeric(data, errors='coerce')
print(numeric_data)
```

Output:

```
0  1.00
1  2.00
2  3.14
3   NaN
4  5.00
dtype: float64
```

quantile() method:

- Return value at the given quantile.
- Syntax: pd.quantile(q=0.5, interpolation)

Parameters:

q:	float or array-like, default 0.5 (50% quantile)
interpolation:	{'linear', 'lower', 'higher', 'midpoint', 'nearest'}

Code:

```
import pandas as pd
data = pd.Series([1, 2, 3, 4, 5])
q = data.quantile()
print(q)
```

Output:

```
3.0
```

PANDAS DATAFRAMES

Pandas DataFrame

- A Pandas DataFrame is a two-dimensional data structure, like a two-dimensional array, or a table with rows and columns.
- Used for data manipulation, analysis, and cleaning, featuring labelled rows and columns, support for various data types, and flexibility for adding, removing, and transforming data.
- It is widely used in data science and analysis tasks for handling structured data efficiently.
- Syntax: pd.DataFrame(data, index, columns, dtype, copy)

Parameters:

data:	ndarray, Iterables, dict, or DataFrame
index:	Index or array-like
columns:	Index or array-like
dtype:	dtype, default None
copy:	bool or None, default None

Creating a DataFrame using lists

Code:
```
import pandas as pd
import numpy as np
student_data = [ [100,95,14], [107,87,16], [89,78,12] ]
print(pd.DataFrame(student_data,
columns= [ 'iq', 'marks', 'package' ] ))
```
Output:
```
    iq   marks   package
0  100     95        14
1  107     87        16
2   89     78        12
```

Creating a DataFrame using dictionary

Code:
```
import pandas as pd
import numpy as np
dictionary = {
                'iq' : [100,95,14],
                    'marks' : [107,87,16],
                'package' : [89,78,12]
              }
students = pd.DataFrame(dictionary)
print(students)
```
Output:
```
    iq   marks   package
0  100     95        14
1  107     87        16
2   89     78        12
```

Creating a DataFarame using read_csv() function

Code:
```
import pandas as pd
# movie datasets
movies = pd.read_csv('movies.csv')
# ipl matches dataset
ipl = pd.read_csv('ipl.csv')
```

Note: Above datasets use for performing dataframe attributes and methods.

Pandas DataFrame Attribute

Accessing a DataFrame through its attributes allows us to get the intrinsic properties of the DataFrame.

shape attribute:

- Used to display the total number of rows and columns of a particular data frame.
- Returns a tuple representing the dimensionality of the DataFrame.
- For example, if we have 3 rows and 2 columns in a DataFrame then the shape will be (3,2).

Code:
```
import pandas as pd
print('Shape of the DataFrame:',movies.shape)
print('Shape of the DataFrame:',ipl.shape)
```
Output:
```
Shape of the DataFrame: (1629, 18)
Shape of the DataFrame: (950, 20)
```

dtypes attribute:

- Return datatype of each column present in a dataframe.

Code:
```
import pandas as pd
print(ipl.dtypes)  # or print(movies.dtypes)
```
Output:
```
ID                int64
City             object
Date             object
Season           object
MatchNumber      object
Team1            object
    ...
Team2Players     object
Umpire1          object
Umpire2          object
dtype: object
```

index attribute:

- Display the row labels of a the dataframe object.

Code:
```
import pandas as pd
print(movies.index)
print(ipl.index)
print(students.indexs)
```
Output:
```
RangeIndex(start=0, stop=1629, step=1)
RangeIndex(start=0, stop=950, step=1)
RangeIndex(start=0, stop=3, step=1)
```

columns attribute:

Fetch the label values for columns present in a dataframe.

Code:
```
import pandas as pd
print(ipl.columns)
print(students.columns)
```
Output:
```
Index(['ID', 'City', 'Date', 'Season', 'MatchNumber', 'Team1', 'Team2',
       'Venue', 'TossWinner', 'TossDecision', 'SuperOver',
       'WinningTeam', 'WonBy', 'Margin', 'method',
       'Player_of_Match', 'Team1Players', 'Team2Players',
       'Umpire1', 'Umpire2'],  dtype='object')
Index(['iq', 'marks', 'package'], dtype='object')
```

values attribute:

- Represent the values/data of dataframe in numpy array from.

Code:
```
import pandas as pd
# print(movies.values)
# print(ipl.values)
print(students.values)
```
Output:
```
[[100 107  89]
 [ 95  87  78]
 [ 14  16  12]]
```

Pandas DataFrame Methods/Functions:

These methods are most used in Pandas DataFrame.

head() method:

- Used to display the first few rows (default is 5) of a DataFrame.

Code:
```
import pandas as pd
print(movies.head()) # default is 5
print(ipl.head(2))
```
Output:
```
No output taken
```

tail() method:

- Used to display the last few rows (default is 5) of a DataFrame.

Code:
```
import pandas as pd
print(movies.tail()) # default is 5
print(movies.tail(2))
```
Output:
```
No output taken
```

sample() method:

- To randomly select a specified number of rows or columns (default is 1) or elements from a dataframe.
- Useful to obtain a random sample from your data for data exploration, analysis, or testing.
- Ex. if we have biased in your datasets then we can use this method.

Code:

```
import pandas as pd
print(students.sample()) # default is 1
# specify the number for random sample
# print(ipl.sample(3))
```

Output:

```
    iq    marks   package
1  107      87        16
```

info() method:

- Used to get a concise summary of the dataframe.
- Prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.
- To get a quick overview of the dataset we use the info() method.

Code:

```
import pandas as pd
# print(ipl.info())
# print(movies.info())
print(students.info())
```

Output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3 entries, 0 to 2
Data columns (total 3 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   iq         3 non-null       int64
 1   marks      3 non-null       int64
 2   package    3 non-null       int64
dtypes: int64(3)
memory usage: 204.0 bytes
```

Pandas DataFrame Mathematical Methods

Common statistical methods in Pandas DataFrame

~ index(0) represents axis=0)

~ column(1) represents axis=1

sum() metnod:

- Return the sum of the values over the requested axis.
- Syntax: DataFrame.sum(axis, numeric_only ,skipna, **kwargs)

Parameters:

axis:	It represents index or column axis, '0' for index and '1' for the column. When the axis=0, method applied over the index axis and when the axis=1 method applied over the column axis.
skipna:	Bool (True or False). The default value is None.
numeric_only:	bool, default False Include only float, int, boolean columns. Not implemented for Series.
**kwargs:	Additional keyword arguments to be passed to the function.

Code:

```
import pandas as pd
print(students.sum()) # default (column-wise)
print(students.sum(axis=1)) # row-wise
```

Output:

```
iq       209
marks    210
package  179
dtype: int64
0  296
1  260
2   42
dtype: int64
```

max() method:

- Used to get the maximum of the values over the requested axis. It returns Series and if the level is specified, it returns the DataFrame.
- Syntax: DataFrame.max(axis, skipna, numeric_only **kwargs)

Code:

```
import pandas as pd
print(students.max())
```

Output:

```
iq   100
%    107
lpa   89
dtype: int64
```

min() method:

- Used to get the minimum of the values over the requested axis.
- It returns Series and if the level is specified, it returns the DataFrame.
- Syntax: DataFrame.min(axis, skipna, numeric_only, **kwargs)

Code:

```
import pandas as pd
print(students.min())
```

Output:

```
iq    0
%     0
lpa   0
dtype: int64
```

mean() method:

- Used to get the mean of the values over the requested axis.
- It returns Series and if the level is specified, it returns the DataFrame.
- Syntax: DataFrame.mean(axis, skipna, numeric_only, **kwargs)

Code:

```
import pandas as pd
print(students.mean())
```

Output:

```
iq    41.8
%     42.0
lpa   35.8
dtype: int64
```

median() method:

- Used to get the median of the values over the requested axis.
- It returns Series and if the level is specified, it returns the DataFrame.
- Syntax: DataFrame.median(axis, skipna, **kwargs)

Code:

```
import pandas as pd
print(students.mean())
```

Output:

```
iq    14.0
%     16.0
lpa   12.0
dtype: int64
```

mode() method:

- we can get each element mode along the specified axis.
- When this method applied to the DataFrame, it returns the DataFrame which consists of the modes of each column or row.
- Syntax: DataFrame.mode(axis=0, drpna=True)

Parameters:

dropna:	It represents the bool, and the default is True. It does not consider the null values.

Code:

```
import pandas as pd
print(students.mode())
```

Output:

```
   iq   %   lpa
0   0   0    0
```

std() method:

- Return sample standard deviation over requested axis. By default the standard deviations are normalized by N-1.
- It is a measure that is used to quantify the amount of variation or dispersion of a set of data values.
- Syntax: DataFrame.std(axis, skipna, level, ddof=1, numeric_only, **kwargs)

Parameters:

dropna:	It represents the bool, and the default is True. It does not consider the null values.
ddof:	int, default 1 Delta Degrees of Freedom. The divisor used in calculations is N - ddof, where N represents the number of elements.

Code:

```
import pandas as pd
print(students.std())
```

Output:

```
iq     51.197656
%      51.122402
lpa    43.990908
dtype: float64
```

var() function:

- Returns the unbiased variance over the specified axis.
- Syntax :DataFrame.var(axis, skipna, level, ddof=1)

Code:

```
import pandas as pd
print(students.var())
```

Output:

```
iq      2621.2
%       2613.5
lpa     1935.2
dtype: float64
```

sort_values() function: (series or dataframe)

- sort_values() is used to sort the values within a Pandas Series.
- By default, it sorts the values in ascending order, but you can specify the sorting order using the ascending parameter, where ascending=True sorts in ascending order, and ascending=False sorts in descending order
- syntax: series.sort_values(axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')

Parameters:

axis :	axis to direct sorting
ascending:	If True, sort values in ascending order, otherwise descending.
inplace:	If True, perform operation in-place.
kind:	choice of sorting algorithm.
na_position:	argument 'first' puts NaNs at the beginning, 'last' puts NaNs at the end.

Code:

```
import pandas as pd
movies = pd.read_csv('movies.csv')
movies.sort_values()
# Sorting based on column
movies.sort_values(by='title_x', ascending=False)
# Sorting with multiple columns movies.columns
movies.sort_values(by=[ 'year_of_release', 'title_x' ],
                   ascending=[False, False])
```

Output:

	Name	college	branch	cgpa	package
0	nitish	bit	eee	6.66	4.0
1	ankit	iit	it	8.25	5.0
2	rupesh	vit	cse	6.41	6.0
3	NaN	NaN	NaN	NaN	NaN
4	mrityunjay	NaN	me	5.60	6.0
5	NaN	vlsi	ce	9.00	7.0
6	rishabh	ssit	civ	7.40	8.0
7	NaN	NaN	cse	10.00	9.0
8	aditya	NaN	bio	7.40	NaN
9	NaN	git	NaN	NaN	NaN

code:

```
students = pd.DataFrame( {
        'name':[ 'nitish', 'ankit', 'rupesh', np.nan, 'mrityunjay',
                 np.nan, 'rishabh', np.nan, 'aditya', np.nan ],
        'college':[ 'bit', 'iit', 'vit', np.nan,np.nan, 'vlsi', 'ssit', np.nan,
                    np.nan, 'git' ],
        'branch':[ 'eee', 'it', 'cse', np.nan, 'me', 'ce', 'civ',' cse', 'bio',
                   np.nan ],
        'cgpa':[6.66, 8.25, 6.41, np.nan, 5.6, 9.0, 7.4,10, 7.4, np.nan ],
        'package':[4, 5, 6, np.nan, 6, 7, 8, 9, np.nan, np.nan ] } )
print(students)
# Sorting with NaN values
students.sort_values(by='name', na_position='first')
```

	Name	college	branch	cgpa	package
3	NaN	NaN	NaN	NaN	NaN
5	NaN	vlsi	ce	9.00	7.0
7	NaN	NaN	cse	10.00	9.0
9	NaN	git	NaN	NaN	NaN
8	aditya	NaN	bio	7.40	NaN
1	ankit	iit	it	8.25	5.0
4	mrityunjay	NaN	me	5.60	6.0
0	nitish	bit	eee	6.66	4.0
6	rishabh	ssit	civ	7.40	8.0
2	rupesh	vit	cse	6.41	6.0

rank() method: (series)

- Used to compute numerical data ranks (1 through n) along axis.
- After sorting (by default in ascending order), the position is used to determine the rank that is returned.
- If data contains equal values, then they are assigned with the average of the ranks of each value by default.
- Syntax: Series.rank(axis=0, method='average', numeric_only= NoDefault. no_default, na_option='keep', ascending=True, pct=False)

Parameters:

axis:	index to direct ranking
method:	{'average', 'min', 'max', 'first', 'dense'}
numeric_only:	Include only float, int, boolean data. Valid only for DataFrame or Panel objects
na_option :	{'keep', 'top', 'bottom'}
ascending :	False for ranks by high (1) to low (N)
pct :	Computes percentage rank of data

Code:

```
import pandas as pd
batsman = pd.read_csv('batsman_runs_ipl.csv')
# apply rank() function on specific column
batsman['ranks'] = batsman['batsman_run'].rank(ascending=False)
batsman.sort_values('ranks')
```

Output:

	batter	batsman_run	batting_rank
569	V Kohli	6634	1.0
462	S Dhawan	6244	2.0
130	DA Warner	5883	3.0
430	RG Sharma	5881	4.0
493	SK Raina	5536	5.0

sort_index(): (Series or DataFrame)

- Pandas Series.sort_index() function is used to sort the index labels of the given series or DataFrame.
- syntax: Series.sort_index(axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')

Code:

```
import pandas as pd
marks = { 'maths' : 78,  'english' : 70, 'science' : 89 }
marks_series = pd.Series(marks)
print(marks_series)
```

Output:

```
maths    78
english  70
science  89
dtype: int64
```

```
# alphabetically sorted index for string index
```

Code:

```
print(marks_series.sort_index())
# print(marks_series.sort_index(ascending=False))
```

Output:

```
english  70
maths    78
science  89
dtype: int64
```

set_index() : (dataframe)

- Used to re-assign a row label using the existing column of the DataFrame.
- It can assign one or multiple columns as a row index.
- Syntax: DataFrame.set_index(keys, drop=True, append=False, inplace=False, verify_integrity=False)

Parameters:

keys:	Column name or list of column name.
append:	Appends the column to existing index column if True.
inplace:	Makes the changes in the dataframe if True.
verify_integrity:	Checking new index column for duplicates if True.

Code:
```
import pandas as pd
# set batter column as index
batsman.set_index('batter').head()
# batsman.set_index('batter',inplace=True)
```
Output:

batter	batsman_run	batting_rank
A Ashish Reddy	280	166.5
A Badoni	161	226.0
A Chandila	4	535.0
A Chopra	53	329.0
A Choudhary	25	402.5

reset_index() : (series and dataframe)
- It is opposite to set_index()
- Convert the series into dataframe.
- Function reset the index to the default integer index beginning at 0, We can simply use the reset_index() function.
- Syntax: DataFrame.reset_index(inplace=False)

Code:
```
import pandas as pd
# reset the index column
batsman.reset_index().head()
# batsman.reset_index(inplace=True)
```
Output:

	batter	batsman_run	batting_rank
0	A Ashish Reddy	280	166.5
1	A Badoni	161	226.0
2	A Chandila	4	535.0
3	A Chopra	53	329.0
4	A Choudhary	25	402.5

Code:
```
# how to replace existing index without loosing
batsman.reset_index().set_index('batting_rank').head(3)
# series to dataframe using reset_index
marks_series.reset_index()
```
Output:

batting_rank	index	batter	batsman_run
166.5	0	A Ashish Reddy	280
226.0	1	A Badoni	161
535.0	2	A Chandila	4

Output:

	index	0
0	maths	67
1	english	57
2	science	89

rename() method: (dataframe)
- Used to rename any index, column, or row.
- Syntax: DataFrame.rename(mapper, index, columns, axis, copy, inplace, level)

Parameters:

mapper, index , columns:	Dictionary value, key refers to the old name and value refers to new name. Only one of these parameters can be used at once.
axis:	int or string value, 0/'row' for Rows and 1/'columns' for columns.
copy:	copies underlying data if True.
inplace:	makes changes in original Data Frame if True.
level:	specify level in dataframe if it have multiple level index.

Code:
```
import pandas as pd
# display specific columns to from movies dataset
movie_list = movies[['title_x','imdb_rating','imdb_votes']].head(3)
movie_list
```
Output:

	title_x	imdb_rating	imdb_votes
0	Uri: The Surgical Strike	8.4	35112
1	Battalion 609	4.1	73
2	The Accidental Prime Minister (film)	6.1	5549

Code:
```
# set the index
movie_list.set_index('title_x', inplace=True)
# rename the columns
movie_list.rename(columns={'imdb_votes' : 'votes',
                           'imdb_rating' : 'rating'}).head(3)
```
Output:

		rating	votes
title_x			
0	Uri: The Surgical Strike	8.4	35112
1	Battalion 609	4.1	73
2	The Accidental Prime Minister (film)	6.1	5549

Code:
```
# rename the index labels
movie_list.rename(index={ 'Uri: The Surgical Strike' : 'Uri' ,
                          'Battalion 609' : 'Battalion' }).head(3)
```
Output:

		rating	votes
title_x			
0	Uri	8.4	35112
1	Battalion	4.1	73
2	The Accidental Prime Minister (film)	6.1	5549

unique() method: (series)
- Return unique values based on a hash table.
- Multiple missing values consider as one NaN value.
- Syntax: Series.unique()

Code:
```
import pandas as pd
temp = pd.Series([1,1,2,2,3,3,4,np.nan,np.nan])
print(temp.unique())
# return unique ipl season form ipl dataset
print(ipl['Season'].unique())
# count the unique ipl season form ipl dataset
print(ipl['Season'].unique().shape[0])
```
Output:
```
array([ 1.,  2.,  3.,  4., nan])
['2022' '2021' '2020/21' '2019' '2018' '2017' '2016' '2015' '2014' '2
 013'
'2012' '2011' '2009/10' '2009' '2007/08']
15
```

nunique() method: (series and dataframe)
- The number of unique observations over the requested axis.
- It returns Series with a number of distinct observations.
- Syntax : DataFrame.nunique(axis=0, dropna=True)

Parameters:

axis:	It represents index or column axis, '0' for index and '1' for the column. When the axis=0, function applied over the index axis and when the axis=1 function applied over the column axis
dropna:	It represents the bool (True or False), and the default is True. It does not include NaN in the counts.

Code:
```
import pandas as pd
# Count unique values of the dataframe
temp = pd.DataFrame([1,1,2,2,2,3,3,4,np.nan,np.nan])
# count with default parameter
print('without dropna parameter :',temp.nunique()[0])
# count with dropna=False parameter
print('with dropna parameter :',temp.nunique(dropna=False)[0])
# return unique ipl season form ipl dataset
print('ipl season count :',ipl['Season'].nunique())
```
Output:
```
without dropna parameter: 4
with dropna parameter: 5
ipl season count: 15
```

Functions/Methods for handling missing values

isnull() method: (series and dataframe)

- Returns the DataFrame of the boolean values.
- If the resultant DataFrame consists of the True, it indicates that the element is a null value and if it is False, it indicates that the element is not a null value.
- Syntax: Series.isnull()

Code:

```
# apply on entire dataframe
students.isnull()
# return boolean values
students['name'].isnull()
# display nan values
students['name'][students['name'].isnull()]
```

Output:

```
3    NaN
5    NaN
7    NaN
9    NaN
Name: name, dtype: object
```

notnull() method: (series and dataframe)

- Used to detect the existing values.
- It returns a DataFrame consisting of bool values for each element in DataFrame that indicates whether an element is not a null value.
- While detecting the existing values, this method does not consider the characters such as empty strings '' or numpy.inf as null values.
- Syntax: DataFrame.notnull()

Code:

```
# apply on entire dataframe
students.notnull()
# detecting existing values
students['name'][students['name'].notnull()]
```

Output:

```
0     nitish
1      ankit
2     rupesh
4   mrityunjay
6     rishabh
8     aditya
Name: name, dtype: object
```

hasnans attribute: (series)

- Returns a boolean value.
- It returns True if the given Series object has missing values in it else it return False.
- Syntax: Series.hasnans

Code:

```
students['name'].hasnans
```

Output:

```
True
```

drona() method: (series or dataframe)

- It removes the missing values and returns the DataFrame with NA entries dropped from it or None if inplace=True.
- Syntax: DataFrame.dropna(axis=0, how='any', thresh, subset, inplace=False)

Parameters:

axis:	{0 or 'index', 1 or 'columns'}, default 0. Determine if rows or columns which contain missing values are removed. • 0, or 'index': Drop rows that contain missing values • 1, or 'columns': Drop columns that contain the missing value
how:	{'any', 'all'}, default 'any'. Determine if row or column is removed from DataFrame when we have at least one NA or all NA. • 'any': If any NA values are present, drop that row or column. • 'all': If all values are NA, drop that row or column.
thresh:	int, optional. Require that many non-NA values.
subset:	array-like, optional labels along another axis to consider, e.g. If you are dropping rows these would be a list of columns to include.
inplace:	bool, default False. If True, do operation inplace and return None.

Code:

```
# apply on sereis
students['name'].dropna()
# apply on dataframe & remove rows if any one columns have nan value
students.dropna()
# with how parameter
students.dropna(how='all')
# remove nan value from specific column
students.dropna(subset=[ 'name' ])
# remove nan values based on mutiple columns
students.dropna(subset=[ 'name','college' ])
```

Output:

	Name	college	branch	cgpa	package
0	nitish	bit	eee	6.66	4.0
1	ankit	iit	it	8.25	5.0
2	rupesh	vit	cse	6.41	6.0
3	NaN	NaN	NaN	NaN	NaN
4	mrityunjay	NaN	me	5.60	6.0
5	NaN	vlsi	ce	9.00	7.0
6	rishabh	ssit	civ	7.40	8.0
7	NaN	NaN	cse	10.00	9.0
8	aditya	NaN	bio	7.40	NaN
9	NaN	git	NaN	NaN	NaN

fillna() method: (series and dataframe)

- Fills NA/NaN values using the specified method.
- It returns the DataFrame object with missing values filled or None if inplace=True.
- Syntax: DataFrame.fillna(value, method, axis, inplace=False, limit, downcast)

Parameters:

value:	scalar, dict, Series, or DataFrame.Value to use to fill
method:	{'backfill', 'bfill', 'pad', 'ffill', None}, default None.
axis:	{0 or 'index', 1 or 'columns'}. Axis along which to fill missing values.

Note: Most of the time apply on specific columns

Code:

```
# Series
students['name'].fillna('unknown')
# replace students package nan values with their average package
students['package'].fillna(students['package'].mean())
```

Output:

```
0      nitish
1       ankit
2      rupesh
3     unknown
4     mrityunjay
5     unknown
6     rishabh
7     unknown
8      aditya
9     unknown
Name: name, dtype: object


0  4.000000
1  5.000000
2  6.000000
3  6.428571
4  6.000000
5  7.000000
6  8.000000
7  9.000000
8  6.428571
9  6.428571
Name: package, dtype: float64
```

drop_duplicates() method: (series and dataframe)

- It returns a DataFrame with duplicate rows removed.
- Considering certain columns is optional. Indexes, including time indexes, are ignored.
- Syntax : DataFrame.drop_duplicates(subset, keep='first', inplace=False, ignore_index=False)

Code:

```
import pandas as pd
import numpy as np
temp = pd.Series([ 1,1,1,2,3,3,4,4 ])
temp.drop_duplicates()
```

Output:

```
0  1
3  2
4  3
6  4
dtype: int64
```

```
# drop_duplicates() function with keep parameter
```

Code:

```
marks = pd.DataFrame( [ [100,80,10], [120,100,14],
                        [80,70,14], [80,70,14] ],
                      columns= [ 'iq', 'marks', 'package' ] )
print(marks)
print(marks.drop_duplicates(keep='last'))
```

Output:

```
    iq   marks   package
0  100      80        10
1  120     100        14
2   80      70        14
3   80      70        14

    iq   marks   package
0  100      80        10
1  120     100        14
3   80      70        14
```

drop() method: (series and dataframe)

- It drops specified labels from rows or columns.
- It removes rows or columns by specifying label names and correspon ding axis, or by specifying directly index or column names.
- When using a multi-index, labels on different levels can be removed b y specifying the level.
- It returns the DataFrame or None. DataFrame without the removed in dex or column labels or None if inplace=True.
- It raises KeyError exception if any of the labels are not found in the sel ected axis.
- Syntax: DataFrame.drop(labels,axis=0,index,columns,level , inplace=False,errors='raise')

Code:

```
temp = pd.Series([10,2,3,8,10])
print(temp)
# apply on sereis with index parameter
print(temp.drop(index=[0,3]))
# DataFrame with index & columns parameter
students.drop(index=[1,3],columns=['branch','cgpa']).head(4)
```

Output:

```
0    10
1     2
2     3
3     8
4    10
dtype: int64
1    2
2    3
4    10
dtype: int64
     name      college    package
0   nitish       bit         4.0
2  rupesh        vit         6.0
1  mrityunjay    NaN         6.0
1    NaN         vlsi        7.0
```

apply() method: (series and dataframe)

- Using this method we can apply different functions on rows and colu mns of the DataFrame.
- The objects passed to the method are Series objects whose index is e ither the DataFrame's index (axis=0) or the DataFrame's columns (axi s=1).
- Syntax: DataFrame.apply(func, axis=0, raw=False, result_type, args, **kwds)

Code:

```
points_df = pd.DataFrame({
        '1st point' : [(3,4),(-6,5),(0,0),(-10,1),(4,5)],
        '2nd point' : [(-3,4),(0,0),(2,2),(10,10),(1,1)] })
# print(points_df)
# creating function for apply() function
def euclidean(row):
    point_A = row['1st point']
    point_B = row['2nd point']
    return ((point_A[0] - point_B[0])**2 +
            (point_A[1] - point_B[1])**2)**0.5
# apply on DataFrame and asign new column
points_df['distance'] = points_df.apply(euclidean, axis=1)
print(points_df)
```

Output:

```
   1st point  2nd point    distance
0     (3, 4)     (-3, 4)   6.000000
1    (-6, 5)     (0, 0)    7.810250
2     (0, 0)     (2, 2)    2.828427
3   (-10, 1)   (10, 10)   21.931712
4     (4, 5)     (1, 1)    5.000000
```

nlargest() method: (series and dataframe)

- Returns a specified number of rows, starting at the top after sorting the DataFrame by the highest value for a specified column.
- This method is equivalent to df.sort_values(columns, ascending=False).head(n), but more performant.
- Syntax: DataFrame.nlargest(n, columns, keep='last')

Parameters:

n:	Required, a Number, specifying the number of rows to return
columns:	Optional, A String (column label), or a list of column labels, specifying the column(s) to order by
keep:	{'all','first''last'} Optional, default 'last', specifying what to do with d

Code:

```
import pandas as pd
df = pd.Series([10, 20, 65, 0, 30])
largest_values = df.nlargest(3 , keep='last')
# Get 3 largest values
largest_values
```

Output:

```
2  65
4  30
1  20
dtype: int64
```

nsmallest() method: (series and dataframe)

- Used to get n least values from a data frame or a series.
- This method is equivalent to df.sort_values(columns, ascending=True).head(n), but more performant.
- Syntax: DataFrame.nsmallest(n, columns, keep='last')

Code:

```
import pandas as pd
df = pd.Series([10, 20, 65, 0, 30])
smallest_values = df.nsmallest(3)
# Get 3 smallest values
smallest_values
```

Output:

```
3  0
0  10
1  20
dtype: int64
```

insert() method: (dataframe)

- Used to insert a column as a specific position in a pandas dataframe
- Syntax: df.insert(loc, column, value, allow_duplicates=False)

Parameters:

loc:	(int) The index where the new column is to be inserted. The index must be in the range, 0 <= loc <= len(columns).
column:	(str, num, or hashable object) The label (column name) for the inserted column.
value:	(scaler, series, or array-like) The column values.
allow_duplicates:	(bool) Optional argument. Determines whether you can have duplicate columns or not. It is False by default.

Code:

```
import pandas as pd
data = {
    'Name': ['Jim','Tobi'],
    'Age': [26, 28]
 }
df = pd.DataFrame(data)
df
```

Output:

	Name	Age
0	Jim	26
1	Tobi	28

Insert the New Column at specific position

Code:

```
import pandas as pd
df.insert(1, 'Department' , [ 'Sales' , 'Accounting' ])
df
```

Output:

	Name	Department	Age
0	Jim	Sales	26
1	Tobi	Accounting	28

copy() method: (series and dataframe)

- Create a copy of a dataframe.
- By default, the copy is a "deep copy" meaning that any changes made in the original DataFrame will NOT be reflected in the copy.
- Syntax: df.copy(deep=True)

  Parameters:

deep:	Optional. Default True. Specifies whether to make a deep or a shallow copy. • By default (deep=True, any changes made in the original DataFrame will NOT be reflected in the copy. • With the parameter deep=False, it is only the reference to the data (and index) that will be copied, and any changes made in the original will be reflected in the copy, and, any changes made in the copy will be reflected in the original.

Note:

- Use deep=True (default value) to create a deep copy.
- Use deep=False to create a shallow copy.

Code:

```
import pandas as pd
data = {
    'Name': ['Jim','Tobi'],
    'Age': [26, 28]
 }
df = pd.DataFrame(data)
df
```

Output:

	Name	Age
0	Jim	26
1	Tobi	28

create a deep copy

Code:

```
# deep copy is created by default
df1 = df.copy()
df1
```

Output:

	Name	Age
0	Jim	26
1	Tobi	28

Code:

```
# make changes to df1
df1.loc[0, 'Name'] = 'Goku'
# display df1
print(df1)
# display the original dataframe
print(df)
```

Output:

	Name	Age
0	Jim	26
1	Tobi	28

	Name	Age
0	Jim	26
1	Tobi	28

Create a shallow copy of a pandas dataframe

Code:

```
df2 = df.copy(deep=False)
df2
```

Output:

	Name	Age
0	Jim	26
1	Tobi	28

Code:

```
# make changes to df2
df2.loc[0, 'Name'] = 'Goku'
# display df2
print(df2)
# display the original dataframe
print(df)
```

Output:

	Name	Age
0	Jim	26
1	Tobi	28

	Name	Age
0	Jim	26
1	Tobi	28

GROUPBY OBJECT

What is Pandas Groupby?

- Pandas groupby splits all the records from your data set into different categories or groups so that you can analyze the data by these groups.
- When you use the groupby() function on any categorical column of DataFrame, it returns a GroupBy object, which you can use other methods on to group the data.
- Generally we two types of columns in datasets numerical and categorical
- Numerical Columns:
- Numerical columns contain data that consists of numbers. These numbers can be integers or floating-point numbers (decimals).
- Examples of numerical columns include columns like "Age," "Salary," "Temperature," "Number of Items Sold," and "Height.
- Categorical Columns:
- Categorical columns contain data that represents categories or discrete values. These values are often labels or strings.
- Categorical columns include columns like "Gender" (with values like "Male" and "Female"), "Product Category" (with values like "Electronics," "Clothing," and "Furniture"), and "Country" (with values like "USA," "Canada," and "UK").

Note: groupby() always apply on categorical columns

- Syntax: DataFrame.groupby(by, axis=0, level, as_index=True, sort=True, group_keys=True, squeeze=False, **kwargs)

Parameters:

by:	mapping, function, str, or iterables
axis:	int, default 0
level:	If the axis is a MultiIndex (hierarchical), group by a particular level or levels
as_index:	For aggregated output, return object with group labels as the index. Only relevant for DataFrame input. as_index=False is effectively “SQL-style” grouped output
sort:	Sort group keys. Get better performance by turning this off. Note this does not influence the order of observations within each group. groupby preserves the order of rows within each group.
group_keys:	When calling apply, add group keys to index to identify pieces
squeeze:	Reduce the dimensionality of the return type if possible, otherwise return a consistent type

```
# Importing the imdb movies datasets
Code:
    import pandas as pd
    movies = pd.read_csv('imdb-top-1000.csv')
    movies.head(20)
Output:
    Output not taken
```

```
# creating group by object
Code:
    import pandas as pd
    genres = movies.groupby('Genre')
Output:
                Runtime     IMDB_Rating
    Genre
    Action      129.046512  7.949419
    Adventure   134.111111  7.937500
    Animation    99.585366  7.930488
    Biography   136.022727  7.938636
    Comedy      112.129032  7.901290
```

```
# Applying builtIn aggregation function on groupby objects
Code:
    import pandas as pd
    genres.mean() # sum() min() mode() median() std() etc
    genres.mean()[['Runtime' , 'IMDB_Rating']].head()
Output:
                Runtime     IMDB_Rating
    Genre
    Action      129.046512  7.949419
    Adventure   134.111111  7.937500
    Animation    99.585366  7.930488
    Biography   136.022727  7.938636
    Comedy      112.129032  7.901290
```

```
# find the top 3 genres by total earning
Code:
    import pandas as pd
    movies.groupby('Genre').sum()['Gross'].sort_values(
    ascending=False).head(3)
Output:
    Genre
    Drama    3.540997e+10
    Action   3.263226e+10
    Comedy   1.566387e+10
    Name: Gross, dtype: float64
```

```
# efficienct way
Code:
    import pandas as pd
    movies.groupby('Genre')['Gross'].sum().sort_values(ascending=
    False).head(3)
Output:
    Genre
    Drama    3.540997e+10
    Action   3.263226e+10
    Comedy   1.566387e+10
    Name: Gross, dtype: float64
```

```
# find the genre with highest avgrage IMDB rating
Code:
    import pandas as pd
    movies.groupby('Genre')['IMDB_Rating'].mean().sort_values(asc
    ending=False).head(1)
Output:
    Genre
    Western   8.35
    Name: IMDB_Rating, dtype: float64
```

```
# find director with most popularity
Code:
    import pandas as pd
    movies.groupby('Director')['No_of_Votes'].sum().sort_values(asc
    ending=False).head(1)
Output:
    Director
    Christopher Nolan   11578345
    Name: No_of_Votes, dtype: int64
```

```
# find number of movies done by each actor
Code:
    import pandas as pd
    # movies['Star1'].value_counts()
    movies.groupby('Star1')['Series_Title'].count().sort_values
                    (ascending=False).head(3)
Output:
    Star1
    Tom Hanks        12
    Robert De Niro   11
    Clint Eastwood   10
    Name: Series_Title, dtype: int64
```

Groupby Attributes and Methods/function

len() method :

- To find the total number of groups created by the groupby operation.

```
# find total number of groups
Code:
    import pandas as pd
    len(movies.groupby('Genre')) # or movies['Genre'].nunique()
Output:
    14
```

size() method:

- To find the number of items in each group.

```
# find items in each group size & sort based on index
Code:
    import pandas as pd
```

```
movies.groupby('Genre').size().head()
```

Output:

```
Genre
Action       172
Adventure     72
Animation     82
Biography     88
Comedy       155
dtype: int64
```

```
# sort based on values
```

Code:

```
import pandas as pd
movies['Genre'].value_counts().head()
```

Output:

```
Drama      289
Action     172
Comedy     155
Crime      107
Biography   88
Name: Genre, dtype: int64
```

first() & last() method:

- To retrieve the first or last item within each group.

Code:

```
import pandas as pd
genres = movies.groupby('Genre')
# retrive the first item of each group using groupby
genres[ 'Series_Title', 'Runtime', 'IMDB_Rating' ].first().head()
# retrive the last item of each group using groupby
genres[ 'Series_Title', 'Runtime', 'IMDB_Rating' ].last().head()
```

Output:

Genre	Series_Title	Runtime	IMDB_Rating
Action	The Dark Knight	152	9.0
Adventure	Interstellar	169	8.6
Animation	Sen to Chihiro no kamikakushi	125	8.6
Biography	Schindler's List	195	8.9
Comedy	Gisaengchung	132	8.6

Genre	Series_Title	Runtime	IMDB_Rating
Action	Escape from Alcatraz	112	7.6
Adventure	Kelly's Heroes	144	7.6
Animation	The Jungle Book	78	7.6
Biography	Midnight Express	121	7.6
Comedy	Breakfast at Tiffany's	115	7.6

nth() method:

- To retrieve the nth item from each group within a DataFrame after performing a groupby operation.
- Syntax: DataFrameGroupBy.nth(n, dropna='all')

Code:

```
import pandas as pd
# nth()
genres['Series_Title'].nth(7).head()
```

Output:

```
Genre
Action                Star Wars
Adventure                Queen
Animation   Mononoke-hime
Biography             Amadeus
Comedy                  Amélie
Name: Series_Title, dtype: object
```

get_group() method:

- To retrieve a specific group by its name, which is useful for selective group access as opposed to filtering.

Code:

```
import pandas as pd
genres['Series_Title','Genre'].get_group('Family')
```

Output:

	Series_Title	Genre
688	E.T. the Extra-Terrestrial	Family
698	Willy Wonka & the Chocolate Factory	Family

groups attribute:

- To access the groups as a dictionary where keys are unique group labels and values are group indices.

```
# index position of movies based on genres
```

Code:

```
import pandas as pd
genres.groups
```

Output:

```
{'Action': [2, 5, 8, 10, 13, 14, 16, 29, 30, 31, 39, 42, 44, 55, 57, 59, 60, 6
8, 72, 106, 109, 129, 130, 134, 140, 142, 144, 152, 155, 160, ...],
 ...
 'Thriller': [700], 'Western': [12, 48, 115, 691]}
```

describe() method:

- To generate descriptive statistics for each group, providing information like mean, std deviation, min, max, and more.

```
# describe each column or specified columns
```

Code:

```
import pandas as pd
# genres.describe()
genres.describe()['Runtime'].head()
```

Output:

Genre	Count	mean	...	75%	max
Action	172.0	129.046512	...	143.25	321.0
Adventure	72.0	134.111111	...	149.00	228.0
Animation	82.0	99.585366	...	106.75	137.0
Biography	88.0	136.022727	...	146.25	209.0
Comedy	155.0	112.129032	...	124.50	188.0

sample() method:

- To obtain a random sample from each group.
- Syntax: pd.DataFrame.groupby().sample(n, replace=False)

Parameters:

n:	Sample size to return for each group.
replace:	default false and allow or disallow sampling of the same row more than once.

Code:

```
import pandas as pd
#genres.sample()
# genres.sample(2, replace=True)
genres.sample(2, replace=True)[['Series_Title', 'Released_Year',
                                'Genre']].head(4)
```

Output:

	Series_Title	Released_Year	Genre
900	Serbuan maut	2011	Action
223	Mad Max: Fury Road	2015	Action
675	Back to the Future Part II	1989	Adventure
406	The Princess Bride	1987	Adventure

nunique() method:

- To count the number of unique values within each group.

Code:

```
import pandas as pd
# genres.nunique()
genres.nunique()[['Series_Title','Released_Year']].head()
```

Output:

Genre	Series_Title	Released_Year
Action	172	61
Adventure	72	49
Animation	82	35
Biography	88	44
Comedy	155	72

agg() method :

- Apply multiple aggrigation function at same time.
- Syntax : DataFrameGroupBy.agg(arg, *args*, **kwargs)

```
# Passing dictionary
```

Code:

```
import pandas as pd
genres.agg(
    {
      'Runtime' : 'mean',
      'IMDB_Rating' : 'mean',
      'Gross' : 'sum',
    }
).head()
```

Output:

Genre	Runtime	IMDB_Rating	Gross

```
Genre
Action       129.046512   7.949419   3.263226e+10
Adventure    134.111111   7.937500   9.496922e+09
Animation     99.585366   7.930488   1.463147e+10
Biography    136.022727   7.938636   8.276358e+09
Comedy       112.129032   7.901290   1.566387e+10
```

Passing list

Code:

```
import pandas as pd
genres.agg([ 'min' , 'max' , 'mean' ]).[ 'Runtime'].head()
```

Output:

Genre	Runtime		
	min	max	mean
Action	45	321	129.046512
Adventure	88	228	134.111111
Animation	71	137	99.585366
Biography	93	209	136.022727
Comedy	68	188	112.129032

Adding both dictionary and list

Code:

```
import pandas as pd
genres.agg( { 'Runtime' : [ 'max' , 'min' , 'mean' ],
              'IMDB_Rating' : [ 'max' , 'min' ], }).head()
```

Output:

Genre	Runtime			IMDB_Rating	
	min	max	mean	min	max
Action	45	321	129.046512	9.0	7.6
Adventure	88	228	134.111111	8.6	7.6
Animation	71	137	99.585366	8.6	7.6
Biography	93	209	136.022727	8.9	7.6
Comedy	68	188	112.129032	8.6	7.6

Looping On Groups

- Groupby is a method in Pandas that allows you to group a DataFrame or Series by one or more columns.
- Once you have your data grouped, you can perform various operations on each group, such as aggregation, transformation, or filtering.

loop over groupby:

- Once you have your data grouped using the groupby method, you can loop over each group using a for loop.
- Syntax:

```
for group_name, group_data in df.groupby('column_name'):
    #perform some analysis or visualization
```

Find the highest rated movies of each genre

Code:

```
import pandas as pd
df = pd.DataFrame(columns=movies.columns)
for group, data in genres:
df = df.append(data[data['IMDB_Rating'] == data['IMDB_Rating'].
    max()])
df[['Series_Title', 'IMDB_Rating', 'Genre']].head()
```

Output:

	Series_Title	IMDB_Rating	Genre
2	The Dark Knight	9.0	Action
21	Interstellar	8.6	Adventure
23	Sen to Chihiro no kamikakushi	8.6	Animation
7	Schindler's List	8.9	Biography
19	Gisaengchung	8.6	Comedy

Split-Apply-Combine Strategy

By “group by” we are referring to a process involving one or more of the following steps:

- Splitting the data into groups based on some criteria.
- Applying a function to each group independently.
- Combining the results into a new DataFrame.

Find number of movies starting with A for each group

Code:

```
import pandas as pd
def alphabet(group):
    return group['Series_Title'].str.startswith('A').sum()
genres.apply(alphabet).head()
```

Output:

```
Genre
Action       10
Adventure     2
Animation     2
Biography     9
Comedy       14
dtype: int64
```

Find ranking of each movie in the group according to IMDB score

Code:

```
def ranking(group):
    # assign the new column for ranking
    group['genre_rank'] = group['IMDB_Rating'].rank
                          (ascending=False)
return group
genres['Series_Title', 'IMDB_Rating'].apply(ranking).head()
```

Output:

	Series_Title	IMDB_Rating	genre_rank
0	The Shawshank Redemption	9.3	1.0
1	The Godfather	9.2	1.0
2	The Dark Knight	9.0	1.0
3	The Godfather: Part II	9.0	2.5
4	12 Angry Men	9.0	2.5

Find normalized IMDB rating group wise

normalized formula: (X - Xmin) / (Xmax - Xmin)

Code:

```
import pandas as pd
def normalized(group):
    X = group['IMDB_Rating']
    Xmin = group['IMDB_Rating'].min()
    Xmax = group['IMDB_Rating'].max()
    group['Normalized_Rating'] = (X - Xmin) / (Xmax - Xmin)
    return group
genres['Series_Title', 'IMDB_Rating'].apply(normalized)
```

Output:

	Series_Title	IMDB_Rating	Normalized_Rating
0	The Shawshank Redemption	9.3	1.0
1	The Godfather	9.2	1.0
2	The Dark Knight	9.0	1.0
3	The Godfather: Part II	9.0	2.5
4	12 Angry Men	9.0	2.5

Groupby on multiple cols

Code:

```
import pandas as pd
duo = movies.groupby(['Director','Star1'])
# size
print(duo.size())
# get_group
duo['Series_Title', 'Director','Star1'].get_group(( 'Aamir Khan' ,
'Amole Gupte' ))
```

Output:

```
Director              Star1
Aamir Khan            Amole Gupte        1
Aaron Sorkin          Eddie Redmayne     1
Abdellatif Kechiche   Léa Seydoux        1
Abhishek Chaubey      Shahid Kapoor      1
Abhishek Kapoor       Amit Sadh          1
                                         ..
Zaza Urushadze        Lembit Ulfsak      1
Zoya Akhtar           Hrithik Roshan     1
                      Vijay Varma        1
Çagan Irmak           Çetin Tekindor     1
Ömer Faruk Sorak      Cem Yilmaz         1
Length: 898, dtype: int64
```

	Series_Title	Director	Star1
65	Taare Zameen Par	Aamir Khan	Amole Gupte

Find the most earning actor->director combo

Code:

```
import pandas as pd
duo['Gross'].sum().sort_values(ascending=False).head(1)
```

Output:

```
Director            Star1
```

```
Akira Kurosawa     Toshirô Mifune   2.999877e+09
Name: Gross, dtype: float64
```

```
# Find the best(in-terms of metascore(avg)) actor->genre combo
```
Code:
```
import pandas as pd
movies.groupby(['Star1','Genre'])['Metascore'].mean().reset_index()
.sort_values('Metascore',ascending=False).head()
```
Output:

	Star1	Genre	Metascore
230	Ellar Coltrane	Drama	100.0
329	Humphrey Bogart	Drama	100.0
360	James Stewart	Mystery	100.0
77	Bertil Guve	Drama	100.0
590	Orson Welles	Drama	100.0

```
# Agg on multiple groupby
```
Code:
```
import pandas as pd
duo.agg(['min','max','mean'])[['Runtime','IMDB_Rating']].head(2)
```
Output:

		Runtime		IMDB_Rating	
		min	max	min	max
Director	Star1				
Aamir Khan	Amole Gupte	165	165	8.4	8.4
Aaron Sorkin	Eddie Redmayne	129	129	7.8	7.8

```
# Importing the deliveries datasets
```
Code:
```
import pandas as pd
ipl = pd.read_csv('deliveries.csv')
ipl.info()
```
Output:
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179078 entries, 0 to 179077
Data columns (total 21 columns):
```

#	Column	Non-Null Count	Dtype
0	match_id	179078 non-null	int64
1	inning	179078 non-null	int64
2	batting_team	179078 non-null	object
3	bowling_team	179078 non-null	object
4	over	179078 non-null	int64
5	ball	179078 non-null	int64
6	batsman	179078 non-null	object
7	non_striker	179078 non-null	object
8	bowler	179078 non-null	object
9	is_super_over	179078 non-null	int64
10	wide_runs	179078 non-null	int64
11	bye_runs	179078 non-null	int64
12	legbye_runs	179078 non-null	int64
13	noball_runs	179078 non-null	int64
14	penalty_runs	179078 non-null	int64
15	batsman_runs	179078 non-null	int64
16	extra_runs	179078 non-null	int64
17	total_runs	179078 non-null	int64
18	player_dismissed	8834 non-null	object
19	dismissal_kind	8834 non-null	object
20	fielder	6448 non-null	object

```
dtypes: int64(13), object(8)
memory usage: 28.7+ MB
```

```
# Find the top 10 batsman in terms of runs
```
Code:
```
import pandas as pd
batsman = ipl.groupby('batsman')
batsman['batsman_runs'].sum().sort_values(ascending=False)
.head(10)
```
Output:
```
batsman
V Kohli          5434
SK Raina         5415
RG Sharma        4914
DA Warner        4741
S Dhawan         4632
CH Gayle         4560
MS Dhoni         4477
RV Uthappa       4446
AB de Villiers   4428
G Gambhir        4228
Name: batsman_runs, dtype: int64
```

```
# Find the batsman with max no of sixes
```
Code:
```
import pandas as pd
sixes = ipl[ipl['batsman_runs'] == 6]
sixes.groupby('batsman')['batsman'].count().sort_values
(ascending=False).head(1).index[0]
```
Output:
```
'CH Gayle'
```

```
# Find batsman with most number of 4's and 6's in last 5 overs
```
Code:
```
import pandas as pd
t_df = ipl[ipl['over'] > 15]
t_df[(t_df['batsman_runs'] == 4) | (t_df['batsman_runs'] == 6)]
t_df.groupby('batsman')['batsman'].count().sort_values
(ascending=False).head(1).index[0]
```
Output:
```
batsman
MS Dhoni         1548
Name: batsman, dtype: int64
```

```
# Find  V Kohli's record against all teams
```
Code:
```
import pandas as pd
t _df = ipl[ipl['batsman'] == 'V Kohli']
t_df.groupby('bowling_team')['batsman_runs'].sum().reset_index()
```
Output:

	bowling_team	batsman_runs
0	Chennai Super Kings	749
1	Deccan Chargers	306
2	Delhi Capitals	66
3	Delhi Daredevils	763
4	Gujarat Lions	283
5	Kings XI Punjab	636
6	Kochi Tuskers Kerala	50
7	Kolkata Knight Riders	675
8	Mumbai Indians	628
9	Pune Warriors	128
10	Rajasthan Royals	370
11	Rising Pune Supergiants	83
12	Rising Pune Supergiants	188
13	Sunrisers Hyderabad	509

```
# Create a function that can return the highest score of any batsman
```
Code:
```
import pandas as pd
def highestScore(batsman):
    t_df = ipl[ipl['batsman'] == batsman]
    return t_df.groupby('match_id')['batsman_runs'].sum()
            .sort_values (ascending=False).head(1).values[0]
highestScore('DA Warner')
```
Output:
```
batsman
MS Dhoni         1548
Name: batsman, dtype: int64
```

MERGING, JOINING AND CONCATENATION

```
# Datasets
Code:
    import pandas as pd
    courses = pd.read_csv('courses.csv')
    students = pd.read_csv('students.csv')
    nov = pd.read_csv('reg-month1.csv')
    dec = pd.read_csv('reg-month2.csv')
    matches = pd.read_csv('matches.csv')
    delivery = pd.read_csv('deliveries.csv')
```

pd.concat() method:

- Used to concatenate pandas objects such as DataFrames and Series.
- We can pass various parameters to change the behavior of the concatenation operation.
- Syntax: pd.concat(objs, axis, join, ignore_index, keys, levels, names, verify_integrity, sort, copy)

Parameters:

objs:	Series or DataFrame objects
axis:	axis to concatenate along; default = 0
join:	way to handle indexes on other axis; default = 'outer'
ignore_index:	if True, do not use the index values along the concatenation axis; default = False
keys:	sequence to add an identifier to the result indexes; default = None
levels:	specific levels (unique values) to use for constructing a MultiIndex; default = None
names:	names for the levels in the resulting hierarchical index; default = None
verify_integrity:	check whether the new concatenated axis contains duplicates; default = False
sort:	sort non-concatenation axis if it is not already aligned when join is 'outer'; default = False
copy:	if False, do not copy data unnecessarily; default = True

```
# Concat the columns vertically(default)
Code:
    import pandas as pd
    registered = pd.concat([nov,dec], ignore_index=True)
    registered
Output:
```

	student_id	course_id
0	23	1
1	15	5
2	18	6
3	23	4
4	16	9

```
Code:
    import pandas as pd
    d1 = {"Name": ["Pankaj", "Lisa"], "ID": [1, 2]}
    d2 = {"Name": "David", "ID": 3}
    df1 = pd.DataFrame(d1, index=[1, 2])
    df2 = pd.DataFrame(d2, index=[3])
    df3 = pd.concat([df1, df2])
    print(df3)
Output:
```

	Name	ID
1	Pankaj	1
2	Lisa	2
3	David	3

pd.append() method:

- Used to append rows of other data frames to the end of the given data frame, returning a new data frame object.
- Columns not in the original data frames are added as new columns and the new cells are populated with NaN value.
- Syntax: DataFrame.append(other, ignore_index=False, verify_integrity=False, sort)

Parameters:

other:	DataFrame or Series/dict-like object, or list of these The data to append.
ignore_index:	If True, do not use the index labels.
verify_integrity:	If True, raise ValueError on creating an index with duplicates.
sortPandas:	default False, Sort columns if the columns of self and other are not aligned.

Note:
Append method is deprecated and will be removed from pandas in a future version.

```
# Example
Code:
    import pandas as pd
    print(nov.append(dec, ignore_index=True).head())
Output:
```

	student_id	course_id
0	23	1
1	15	5
2	18	6
3	23	4
4	16	9

```
# Multiindex dataframe (keep original index as it is)
Code:
    import pandas as pd
    multi = pd.concat([nov, dec], keys=['Nov', 'Dec'])
    print(multi)
    # accessing each months
    # multi.loc['Nov']
    # multi.loc['Dec']
    # accessing the items
    print(multi.loc[('Nov',0)])
Output:
```

		student_id	course_id
Nov	0	23	1
	1	15	5
	2	18	6
	3	23	4
	4	16	9

```
student_id    23
course_id      1
Name: (Nov, 0), dtype: int64
```

```
# Concat dataframe horizontally
Code:
    import pandas as pd
    pd.concat([nov,dec], axis=1)
Output:
```

	student_id	course_id	student_id	course_id
0	23.0	1.0	3	5
1	15.0	5.0	16	7
2	18.0	6.0	12	10
3	23.0	4.0	12	1
4	16.0	9.0	14	9

merge() method:

- Used to merge two DataFrame objects with a database-style join operation.
- The joining is performed on columns or indexes.
- If the joining is done on columns, indexes are ignored.
- This function returns a new DataFrame and the source DataFrame objects are unchanged.
- Syntax : DataFrame.merge(self, right, how='inner', on, left_on, right_on, left_index=False, right_index=False, sort=False, suffixes=('_x', '_y'))

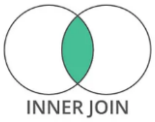
  OR

  Alternate syntax for merge: Ex. students.merge(regs)

Different Joins

Inner Join:

- An inner join returns only the rows where there is a match in both DataFrames' specified columns.
- It retains only the common elements from both DataFrames.
- Use the pd.merge() function with the how='inner' parameter or the .merge() method with the how='inner' argument



Datasets

Code:

```
import pandas as pd
print(students.head())
print (registered.head())
```

Output:

	student_id	name	partner
0	1	Kailash Harjo	23
1	2	Esha Butala	1
2	3	Parveen Bhalla	3
3	4	Marlo Dugal	14
4	5	Kusum Bahri	6

	student_id	course_id
0	23	1
1	15	5
2	18	6
3	23	4
4	16	9

inner join

Code:

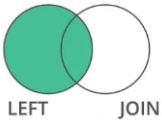
```
import pandas as pd
students.merge(registered, how='inner', on='student_id').head()
```

Output:

	student_id	name	partner	course_id
0	1	Kailash Harjo	23	1
1	1	Kailash Harjo	23	6
2	1	Kailash Harjo	23	10
3	1	Kailash Harjo	23	9
4	2	Esha Butala	1	5

left join:

- A left join returns all the rows from the left DataFrame and the matching rows from the right DataFrame.
- If there's no match in the right DataFrame, NaN values are filled in for columns from the right DataFrame.
- Use the pd.merge() function with the how='left' parameter or the .merge() method with the how='left' argument.



left join

Code:

```
import pandas as pd
print(courses.head())
# courses : left DataFrame
# registered : Right DataFrame
# join courses and registered dataset
courses.merge(registered, how='left', on='course_id').tail()
```

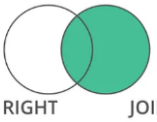
Output:

	course_id	course_name	price
0	1	python	2499
1	2	sql	3499
2	3	data analysis	4999
3	4	machine learning	9999
4	5	tableau	2499

	course_id	course_name	price	student_id
50	10	pyspark	2499	17.0
51	10	pyspark	2499	1.0
52	10	pyspark	2499	11.0
53	11	Numpy	699	NaN
54	12	C++	1299	NaN

right join:

- A right join is the opposite of a left join.
- It returns all the rows from the right DataFrame and the matching rows from the left DataFrame.
- If there's no match in the left DataFrame, NaN values are filled in for columns from the left DataFrame.
- Use the pd.merge() function with the how='right' parameter or the .merge() method with the how='right' argument.



right join

Code:

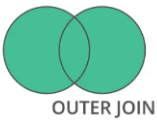
```
import pandas as pd
temp_df = pd.DataFrame({
    'student_id':[26, 27, 28],
    'name':['Akash','Vikas','Rahul'],
    'partner':[28,26,17]
})
students = pd.concat([students, temp_df], ignore_index=True)
students.merge(registered, how='right', on='student_id').tail(3)
```

Output:

	student_id	name	partner	course_id
50	42	NaN	NaN	9
51	50	NaN	NaN	8
52	38	NaN	NaN	1

outer join (full outer join):

- An outer join returns all the rows when there is a match in either the left or the right DataFrame.
- If there's no match in one of the DataFrames, NaN values are filled in for the corresponding columns.
- Use the pd.merge() function with the how='outer' parameter or the .merge() method with the how='outer' argument.



outer join

Code:

```
import pandas as pd
students.merge(registered, how='outer', on='student_id').tail(7)
```

Output:

	student_id	name	partner	course_id
56	25	Shashank D'Alia	2.0	10.0
57	26	Akash	28.0	NaN
58	27	Vikas	26.0	NaN
59	28	Rahul	17.0	NaN
60	42	NaN	NaN	9.0
61	50	NaN	NaN	8.0
62	38	NaN	NaN	1.0

1. Find total revenue generated

Code:

```
import pandas as pd
registered.merge(courses, how='inner', on='course_id')['price']
.sum()
```

Output:

```
154247
```

2. Find month by month revenue

Code:

```
import pandas as pd
t_df = pd.concat([nov,dec], keys=['Nov','Dec']).reset_index()
t_df.merge(courses, on='course_id').groupby('level_0')['price']
.sum()
```

Output:

```
level_0
Dec   65072
Nov   89175
Name: price, dtype: int64
```

3. Print the registration table

cols -> name -> course -> price

Code:

```
import pandas as pd
stu = registered.merge(students, on='student_id')
stu.merge(courses, on='course_id')[['name','course_name','price']]
.head()
```

Output:

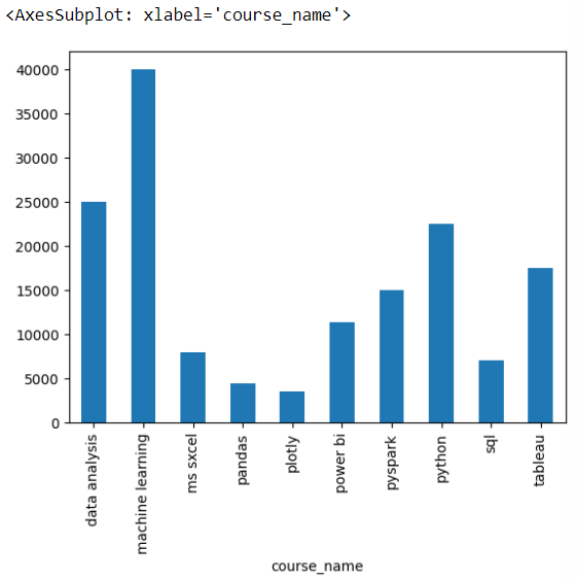
	name	course_name	price
0	Chhavi Lachman	python	2499
1	Preet Sha	python	2499
2	Fardeen Mahabir	python	2499

```
3  Kailash Harjo        python    2499
4  Seema Kota           python    2499
```

```
# 4.Find students who enrolled in both the monthsnov
```
Code:
```
import pandas as pd
common_student_id = np.intersect1d(nov['student_id'],
                                   dec['student_id'])
print(common_student_id)
students[students['student_id'].isin(common_student_id)]
```
Output:
```
array([ 1,  3,  7, 11, 16, 17, 18, 22, 23], dtype=int64)
```

	student_id	name	partner
0	1	Kailash Harjo	23
2	3	Parveen Bhalla	3
6	7	Tarun Thaker	9
10	11	David Mukhopadhyay	20
15	16	Elias Dodiya	25
16	17	Yasmin Palan	7
17	18	Fardeen Mahabir	13
21	22	Yash Sethi	21
22	23	Chhavi Lachman	18

```
# 5. Plot bar chart for revenue/course
```
Code:
```
import pandas as pd
registered.merge(courses,on='course_id').groupby('course_name')
['price'].sum().plot(kind='bar')
```
Output:




```
# 6. Find course that got no enrollment
# courses['course_id']
# regs['course_id']
```
Code:
```
import pandas as pd
course_list = np.setdiff1d(courses['course_id'],
                           registered['course_id'])
courses[courses['course_id'].isin(course_list)]
```
Output:

	course_id	course_name	price
10	11	Numpy	699
11	12	C++	1299

self join:

- A self-join in pandas is a way to combine rows from a single DataFrame by creating a relationship between columns within that DataFrame.
- This is useful when you have hierarchical or relational data stored within a single DataFrame.
- You can achieve a self-join by using the .merge() method or the .join() method.

```
# 9. Find top 3 students who did most number enrollments
```
Code:
```
import pandas as pd
registered.merge(students, on='student_id').groupby
(['student_id','name'])['name'].count().sort_values
(ascending=False).head(3)
```
Output:
```
student_id    name
23            Chhavi Lachman    6
7             Tarun Thaker      5
1             Kailash Harjo     4
Name: name, dtype: int64
```

```
#10. Find top 3 students who spent most amount of money on courses
```
Code:
```
import pandas as pd
registered.merge(students, on='student_id').merge(courses,
on='course_id').groupby(['student_id','name'])['price'].sum()
.head(3)
```
Output:
```
student_id  name
1           Kailash Harjo     7596
2           Esha Butala       2499
3           Parveen Bhalla    7498
Name: price, dtype: int64
```

```
# IPL Problems
# Find top 3 stadiums with highest sixes/match ratio
```
Code:
```
import pandas as pd
temp_df = delivery.merge(matches, left_on='match_id',
                         right_on='id')
sixes_df = temp_df[temp_df['batsman_runs'] == 6]
number_of_sixes = sixes_df.groupby('venue')['venue'].count()
number_of_matches = matches['venue'].value_counts()
(number_of_sixes/number_of_matches).sort_values
(ascending=False).head(3)
```
Output:
```
Holkar Cricket Stadium    17.600000
M Chinnaswamy Stadium     13.227273
Sharjah Cricket Stadium   12.666667
Name: venue, dtype: float64
```

```
# Find orange cap holder of all the seasons
```
Code:
```
import pandas as pd
temp_df = delivery.merge(matches, left_on='match_id',
                         right_on='id')
temp_df.groupby(['season','batsman'])['batsman_runs'].sum()
.reset_index() .sort_values('batsman_runs',ascending=False)
.drop_duplicates(subset=['season'], keep='first')
.sort_values('season')
```
Output:

	season	batsman	batsman_runs
115	2008	SE Marsh	616
229	2009	ML Hayden	572
446	2010	SR Tendulkar	618
502	2011	CH Gayle	608
684	2012	CH Gayle	733
910	2013	MEK Hussey	733
1088	2014	RV Uthappa	660
1148	2015	DA Warner	562
1383	2016	V Kohli	973
1422	2017	DA Warner	641

MULTIINDEX SERIES AND DATAFRAMES

MultiIndex Series

```
# can we have multiple index ? Let's try
index_val = [('cse',2019),('cse',2020),
             ('cse',2021),('cse',2022),
             ('ece',2019),('ece',2020),
             ('ece',2021),('ece',2022)]
a = pd.Series([1,2,3,4,5,6,7,8], index=index_val)
a
```

```
(cse, 2019)    1
(cse, 2020)    2
(cse, 2021)    3
(cse, 2022)    4
(ece, 2019)    5
(ece, 2020)    6
(ece, 2021)    7
(ece, 2022)    8
dtype: int64
```

Problem in MultiIndex Series

- What if we want to fetch values based on specific branch.
- Here branch and year is not independent then we cannot access specific branch, so it will raise KeyError.

```
# KeyError: 'cse'
a['cse']
```

MultiIndex Series (also known as Hierarchical Indexing)

- In pandas, a MultiIndex, also known as hierarchical indexing, allows you to have multiple index levels within single axis.
- This is particularly useful when you're working with higher-dimensional data that can be naturally represented as a hierarchical structure.
- MultiIndex can be applied to both rows (index) and columns, but we'll focus on row-based MultiIndexing here pd.MultiIndex is a class in the pandas library in Python that represents a multi-level or hierarchical index.
- It allows you to have multiple levels of indices on one axis, either for rows or columns in a DataFrame.
- syntax: pd.MultiIndex(levels=[level_values1, level_values2, ...], codes=[code_values1, code_values2, ...],names=[name1, name2, ...])

parameters:

levels:	A list of arrays containing the unique values for each level of the MultiIndex.
codes:	A list of arrays containing the integer codes that represent the labels for each level. The codes are indices into the corresponding levels.
names:	A list of names for each level. Names are optional and are used to provide meaningful labels to the levels.

Creating a MultiIndex (hierarchical index) object

- We can create a MultiIndex in several ways. One common way is to pass a list of arrays or tuples as the index when creating a DataFrame
- Also you can create a MultiIndex by using the pd.MultiIndex constructor or A MultiIndex can be created from a list of arrays (using MultiIndex.from_arrays()), an array of tuples (using MultiIndex.from_tuples()), a crossed set of iterables (using MultiIndex.from_product()), or a DataFrame (using MultiIndex.from_frame()).
- The Index constructor will attempt to return a MultiIndex when it is passed a list of tuples.

```
# pd.MultiIndex.from_tuples():
index_val = [('cse',2019),('cse',2020),('cse',2021),
             ('ece',2019),('ece',2020),('ece',2021)]
multiindex = pd.MultiIndex.from_tuples(index_val)
multiindex
```

```
MultiIndex([('cse', 2019),
            ('cse', 2020),
            ('cse', 2021),
            ('ece', 2019),
            ('ece', 2020),
            ('ece', 2021)],
           )
```

```
# pd.MultiIndex.from_product()
# it's like cartesion product
pd.MultiIndex.from_product([['cse','ece'],
                            [2019,2020,2021,2022]])
```

```
MultiIndex([('cse', 2019),
            ('cse', 2020),
            ('cse', 2021),
            ('cse', 2022),
            ('ece', 2019),
            ('ece', 2020),
            ('ece', 2021),
            ('ece', 2022)],
           )
```

levels attribute:

```
multiindex.levels
```

```
FrozenList([['cse', 'ece'], [2019, 2020, 2021]])
```

```
multiindex.levels[0]
```

```
Index(['cse', 'ece'], dtype='object')
```

```
multiindex.levels[1]
```

```
Int64Index([2019, 2020, 2021], dtype='int64')
```

Creating a series with MultiIndex objects

```
#Creating MultiIndex
multi_i = pd.MultiIndex.from_product([
    ['cse','ece'],
    [2019,2020,2021]])
# Creating a Series with the MultiIndex
ms = pd.Series([1,2,3,4,5,6], index=multi_i)
ms
```

```
cse  2019    1
     2020    2
     2021    3
ece  2019    4
     2020    5
     2021    6
dtype: int64
```

```
# how to fetch items from such a series
ms['cse']
ms['ece']
```

```
2019    4
2020    5
2021    6
dtype: int64
```

unstack() function :

- Convert the multiindex series into Dataframe

```
temp = ms.unstack()
temp
```

	2019	2020	2021
cse	1	2	3
ece	4	5	6

stack() function :

- Convert the dataframe into multiindex series

```
temp.stack()
```

```
cse  2019    1
     2020    2
     2021    3
ece  2019    4
     2020    5
     2021    6
dtype: int64
```

Main purpose of MultiIndexing Objects:

To representation of high dimensional data into lower dimensions like 1d(Series) and 2D(DataFrame)

MultiIndex DataFrame:

- A multi-index DataFrame in pandas refers to a DataFrame that has multiple levels of indexing for both rows and columns.
- It is a way of handling higher-dimensional data by creating a hierarchical index structure.

```
branch_df1 = pd.DataFrame(
    [
        [1,2],[3,4],
        [5,6],[7,8],
        [9,10],[11,12],
    ],
    index = multiindex,
    columns = ['avg_package','students'])
branch_df1
```

		avg_package	students
cse	2019	1	2
	2020	3	4
	2021	5	6
ece	2019	7	8
	2020	9	10
	2021	11	12

Are columns really different from index?

- In pandas, if you transpose a DataFrame using the T attribute or the transpose() method, the columns effectively become the index and vice versa.
- After transposing, what were originally columns become the index, and what were originally the index becomes the columns.

MultiIndex DataFrame based on columns

The MultiIndex has two levels: the first level consists of city names ('delhi' and 'mumbai'), and the second level consists of attributes ('avg_package' and 'students').

```
# multiindex in columns
branch_df2 = pd.DataFrame(
    [
        [1,2,0,0],
        [3,4,0,0],
        [5,6,0,0],
    ],
    index = [2019,2020,2021],
    columns = pd.MultiIndex.from_product([
        ['delhi','mumbai'],['avg_package','students']
    ])
)
branch_df2
```

	delhi		mumbai	
	avg_package	students	avg_package	students
2019	1	2	0	0
2020	3	4	0	0
2021	5	6	0	0

```
branch_df2['mumbai']
```

	avg_package	students
2019	0	0
2020	0	0
2021	0	0

```
branch_df2['delhi']['avg_package']
```

```
2019    1
2020    3
2021    5
Name: avg_package, dtype: int64
```

```
branch_df2.loc[2020]
```

```
delhi   avg_package    3
        students       4
mumbai  avg_package    0
        students       0
Name: 2020, dtype: int64
```

unstack() method:

- In simple words, unstack() method specified index labels becomes new columns and return new dataframe.
- It's used to pivot specified levels of the index labels into new columns, returning a new DataFrame.
- syntax: DataFrame.unstack(level=1, fill_value=None)

parameters:

level:	(default is 1) Specifies the level(s) of the index to unstack. If you have a MultiIndex, you can choose which level(s) you want to move to columns.
fill_value:	If there are missing values after unstacking, you can specify a value to replace those missing values. By default, missing values are filled with NaN.

```
branch_df1.unstack(level=1)
```

	avg_package			students		
	2019	2020	2021	2019	2020	2021
cse	1	3	5	2	4	6
ece	7	9	11	8	10	12

```
branch_df1.unstack(level=0)
```

	avg_package		students	
	cse	ece	cse	ece
2019	1	7	2	8
2020	3	9	4	10
2021	5	11	6	12

stack() function :

- In simple words, Convert the dataframe into multiindex series
- Used to reshape a DataFrame by moving or pivoting specified levels of columns to become inner-most levels of the index.
- This operation results in a new DataFrame or Series with a multi-level index.
- Here's a simpler explanation:
- Imagine you have a table (DataFrame) where some information is stored in both rows and columns.
- The stack() method allows you to take information from the columns and move it to the rows, creating a new structure.
- If your columns have only one level (like a regular DataFrame), using stack() will give you a Series.
- If your columns have multiple levels, you can choose which levels to move to the index, and the result will be a new DataFrame with a multi-level index.
- In essence, stack() helps you transform data by rearranging it from a wide format (with information in columns) to a long format (with information in rows).
- This can be useful for certain types of analyses or when you need the data in a different structure.

```
branch_df1.unstack()
```

	avg_package			students		
	2019	2020	2021	2019	2020	2021
cse	1	3	5	2	4	6
ece	7	9	11	8	10	12

```
# Here most inner columns becomes row
branch_df1.unstack().stack()
```

		avg_package	students
cse	2019	1	2
	2020	3	4
	2021	5	6
ece	2019	7	8
	2020	9	10
	2021	11	12

Working with MultiIndex dataframe

We can use pandas DataFrame methods, functions, and attributes on a MultiIndex DataFrame just like you would on a regular DataFrame.

```
# examples
branch_df3.head()
branch_df3.shape
branch_df3.info()
branch_df3.duplicated()
branch_df3.isnull()
```

Extracting rows and columns

To extract rows and columns from a MultiIndex DataFrame, you can use various methods, including .loc[], .iloc[], and other DataFrame indexing techniques. Here are some examples:

```
# single row
branch_df3.loc[('cse',2019)]
```

```
delhi   avg_package    1
        students       2
mumbai  avg_package    0
        students       0
Name: (cse, 2019), dtype: int64
```

```
# multiple row (similar to indexing)
branch_df3.loc[('cse',2019):('ece',2020):2]
```

		delhi		mumbai	
		avg_package	students	avg_package	students
cse	2019	1	2	0	0
	2021	5	6	0	0
ece	2020	9	10	0	0

```
branch_df3.iloc[0:5:3]
```

		delhi		mumbai	
		avg_package	students	avg_package	students
cse	2019	1	2	0	0
ece	2019	7	8	0	0

```
# Extracting column
branch_df3['delhi']['students']
```

```
cse  2019     2
     2020     4
     2021     6
ece  2019     8
     2020    10
     2021    12
Name: students, dtype: int64
```

```
# multiple columns
branch_df3.iloc[:,1:3]
```

		delhi	mumbai
		students	avg_package
cse	2019	2	0
	2020	4	0
	2021	6	0
ece	2019	8	0
	2020	10	0
	2021	12	0

```
# Extracting both
branch_df3.iloc[[0,4],[1,2]]
```

		delhi	mumbai
		students	avg_package
cse	2019	2	0
ece	2020	10	0

Sorting Index in MultiIndex

Sorting the index in a MultiIndex DataFrame can be done using the sort_index() method.

```
# default
branch_df3.sort_index()
```

		delhi		mumbai	
		avg_package	students	avg_package	students
cse	2019	1	2	0	0
	2020	3	4	0	0
	2021	5	6	0	0
ece	2019	7	8	0	0
	2020	9	10	0	0
	2021	11	12	0	0

```
# both -> descending sorting
branch_df3.sort_index(ascending=False)
```

		delhi		mumbai	
		avg_package	students	avg_package	students
ece	2021	11	12	0	0
	2020	9	10	0	0
	2019	7	8	0	0
cse	2021	5	6	0	0
	2020	3	4	0	0
	2019	1	2	0	0

```
# Sorting on level (0) and level (1)
branch_df3.sort_index(ascending=[False,True])
```

		delhi		mumbai	
		avg_package	students	avg_package	students
ece	2019	7	8	0	0
	2020	9	10	0	0
	2021	11	12	0	0
cse	2019	1	2	0	0
	2020	3	4	0	0
	2021	5	6	0	0

```
# sorting on specific level
branch_df3.sort_index(level=1, ascending=False)
```

		delhi		mumbai	
		avg_package	students	avg_package	students
ece	2021	11	12	0	0
cse	2021	5	6	0	0
ece	2020	9	10	0	0
cse	2020	3	4	0	0
ece	2019	7	8	0	0
cse	2019	1	2	0	0

```
branch_df1.unstack().stack().stack()
```

```
cse  2019  avg_package     1
           students        2
     2020  avg_package     3
           students        4
     2021  avg_package     5
           students        6
ece  2019  avg_package     7
           students        8
     2020  avg_package     9
           students       10
     2021  avg_package    11
           students       12
dtype: int64
```

```
# Example
branch_df2
```

	delhi		mumbai	
	avg_package	students	avg_package	students
2019	1	2	0	0
2020	3	4	0	0
2021	5	6	0	0

```
# Most-inner columns becomes row in stack() method
branch_df2.stack()
```

		delhi	mumbai
2019	avg_package	1	0
	students	2	0
2020	avg_package	3	0
	students	4	0
2021	avg_package	5	0
	students	6	0

```
# stacking one more level
branch_df2.stack().stack()
```

```
2019  avg_package  delhi     1
                   mumbai    0
      students     delhi     2
                   mumbai    0
2020  avg_package  delhi     3
                   mumbai    0
      students     delhi     4
                   mumbai    0
2021  avg_package  delhi     5
                   mumbai    0
      students     delhi     6
                   mumbai    0
dtype: int64
```

Transpose MultiIndex DataFrame

transpose() method:

- Used to transpose the rows and columns of a DataFrame.
- It switches the rows and columns, effectively converting the rows into columns and vice versa.
- Keep in mind that after transposing, the index becomes the columns, and the columns become the index.
- If your original DataFrame had column names, they would become the MultiIndex after transposing.

```
branch_df3
```

		delhi		mumbai	
		avg_package	students	avg_package	students
cse	2019	1	2	0	0
	2020	3	4	0	0
	2021	5	6	0	0
ece	2019	7	8	0	0
	2020	9	10	0	0
	2021	11	12	0	0

```
branch_df3.transpose()
```

		cse			ece		
		2019	2020	2021	2019	2020	2021
delhi	avg_package	1	3	5	7	9	11
	students	2	4	6	8	10	12
mumbai	avg_package	0	0	0	0	0	0
	students	0	0	0	0	0	0

swaplevel() method:

- The swaplevel() method in pandas is used to swap levels of a MultiIndex in a DataFrame.
- This can be particularly useful when you want to interchange the order of levels in a MultiIndex DataFrame.
- The method is applied to the DataFrame's index and can be useful for reorganizing or reshaping the data.
- Default is to swap the two innermost levels of the index.
- Syntax: DataFrame.swaplevel(i=-2, j=-1, axis=0)
- Swap levels i and j in a MultiIndex.
- Default is to swap the two innermost levels of the index.

Parameters:

i, j:	(int or str) Levels of the indices to be swapped. Can pass level name as string.
axis:	{0 or 'index', 1 or 'columns'}, default 0 The axis to swap levels on. 0 or 'index' for row-wise, 1 or 'columns' for column-wise

```
branch_df3.swaplevel()
```

		delhi		mumbai	
		avg_package	students	avg_package	students
2019	cse	1	2	0	0
2020	cse	3	4	0	0
2021	cse	5	6	0	0
2019	ece	7	8	0	0
2020	ece	9	10	0	0
2021	ece	11	12	0	0

```
# swaplevel with default and column
branch_df3.swaplevel(0,axis=1)
```

		avg_package	students	avg_package	students
		delhi	delhi	mumbai	mumbai
cse	2019	1	2	0	0
	2020	3	4	0	0
	2021	5	6	0	0
ece	2019	7	8	0	0
	2020	9	10	0	0
	2021	11	12	0	0

Long format Vs Wide format:

Long format: (tidy data):

- Long format is where, for each data point we have as many rows as the number of attributes and each row contains the value of a particular attribute for a given data point.
- A long format contains values that do repeat in the first column.
- Long format data common sources of obtain data is:
  - Surveys and Questionnaires
  - Time seires data
  - Sensor data
  - Clinical Trials and Medical Studies
- Long format:

Name	Attribute	Value
John	Height	160
John	Weight	67
wick	Height	182
wick	Weight	78

Wide format:

- Wide format is where we have a single row for every data point with multiple columns to hold the values of various attributes.
- A wide format contains values that do not repeat in the first column.
- Wide format data common sources of obtain data is:
  - Government Databases
  - CSV Files from Statistical Agencies
  - Machine Learning Datasets
- Wide format:

Name	Height	Weight
John	160	67
wick	182	78

What difference between them?

Both are similar, but he choice between long and wide formats in data storage and analysis is often dependent on the problem statement, the nature of the data, and the specific analysis or tasks you plan to perform.

Long format:

Name	Attribute	Value
John	Height	160
John	Weight	67
wick	Height	182
wick	Weight	78

Wide format:

Name	Height	Weight
John	160	67
wick	182	78

melt() function:

- In pandas, melt() function was used to transform the dataset from a Wide format into a Long format.
- Syntax: pd.melt(frame, id_vars, value_vars, var_name, value_name='value', col_level)

Parameters:

Frame:	DataFrame
value_vars:	[tuple, list, or ndarray, optional] Column(s) to unpivot. If not specified, uses all columns that are not set as id_vars.
id_vars:	[tuple, list, or ndarray, optional] Column(s) to use as identifier variables.
var_name:	[scalar] Name to use for the 'variable' column. If None it uses frame.columns.name or 'variable'.
value_name:	[scalar, default 'value'] Name to use for the 'value' column.
col_level:	[int or string, optional] If columns are a MultiIndex then use this level to melt.

Pivot Table

The pivot table takes simple column-wise data as input and group the entries into a two-dimenational table that provide a multidimensional summarization of the data.

```
import seaborn as sns
df = sns.load_dataset('tips')
df.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

```
# according to gender what is total average bill
df.groupby('sex')[['total_bill']].mean()
```

	total_bill
sex	
Male	20.744076
Female	18.056897

```
new_df = df.groupby(['sex','smoker'])[['total_bill']]
new_df.mean().unstack()
```

	total_bill	
smoker	Yes	No
sex		
Male	22.284500	19.791237
Female	17.977879	18.105185

pivot_table() function:

- Used to create pivot tables from a DataFrame.
- It provides a flexible and powerful way to reshape and summarize data.
- syntax: df.pivot_table(values=None, index=None, columns=None, aggfunc='mean', fill_value=None, margins=False, dropna=True, margins_name='All')

parameters:

values	This is the column that you want to aggregate. You can specify multiple columns if needed
index	The column whose unique values will become the index of the pivot table
columns	The column whose unique values will become the columns of the pivot table
aggfunc	The aggregation function to apply. It could be 'mean', 'sum', 'count', 'min', 'max', etc. You can also pass a dictionary to apply different aggregation functions to different columns
fill_value	A scalar value to replace missing values.
margins	If True, it adds all row/column margins (subtotals)
dropna	If True, it excludes NA/null values
margins_name	Name of the row/column that will contain the totals when margins is True

```
df.pivot_table(index='sex',
               columns='smoker',
               values='total_bill')
```

smoker	Yes	No
sex		
Male	22.284500	19.791237
Female	17.977879	18.105185

```
# aggfunc parameter
df.pivot_table(index='sex',
               columns='smoker',
               values='total_bill',
               aggfunc='sum')
```

smoker	Yes	No
sex		
Male	1337.07	1919.75
Female	593.27	977.68

```
# columns together
df.pivot_table(index='sex', columns='smoker')
```

	size		tip		total_bill	
smoker	Yes	No	Yes	No	Yes	No
sex						
Male	2.500000	2.711340	3.051167	3.113402	22.284500	19.791237
Female	2.242424	2.592593	2.931515	2.773519	17.977879	18.105185

```
# Multidimentinal
df.pivot_table(index=['sex','smoker'],
               columns=['day','time'],
               values='total_bill')
```

	day	Thur		Fri		Sat	Sun
	time	Lunch	Dinner	Lunch	Dinner	Dinner	Dinner
sex	smoker						
Male	Yes	19.171000	NaN	11.386667	25.892	21.837778	26.141333
	No	18.486500	NaN	NaN	17.475	19.929063	20.403256
Female	Yes	19.218571	NaN	13.260000	12.200	20.266667	16.540000
	No	15.899167	18.78	15.980000	22.750	19.003846	20.824286

```
# mergins parameter
df.pivot_table(index='sex',
               columns='smoker',
               values='total_bill',
               aggfunc='sum',
              margins=True)
```

smoker	Yes	No	All
sex			
Male	1337.07	1919.75	3256.82
Female	593.27	977.68	1570.95
All	1930.34	2897.43	4827.77

Plotting graphs using pivot table:

```
df = pd.read_csv('DATASETS/S21/expense_data.csv')
#df.head()
```

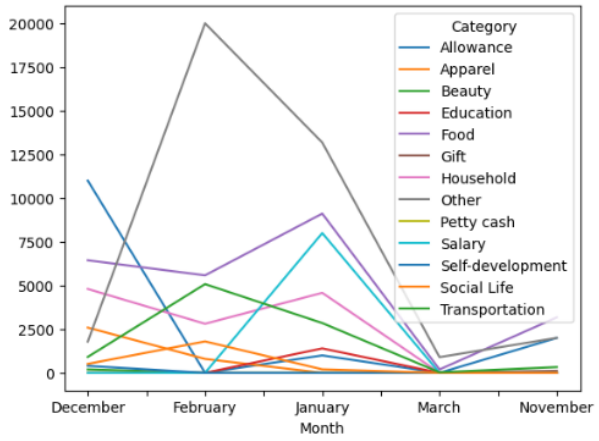
```
# Converting to date column type object to datatime64
df['Date'] = pd.to_datetime(df['Date'])
# add month column
df['Month'] = df['Date'].dt.month_name()
```

```
df.pivot_table(index='Month',columns='Category',
               values='INR',aggfunc='sum',fill_value=0)
```

Category	Allowance	Apparel	Beauty	Education	Food	Gift	Hous
Month							
December	11000	2590	196	0	6440.72	0	
February	0	798	0	0	5579.85	0	
January	1000	0	0	1400	9112.51	0	
March	0	0	0	0	195.00	0	
November	2000	0	0	0	3174.40	115	

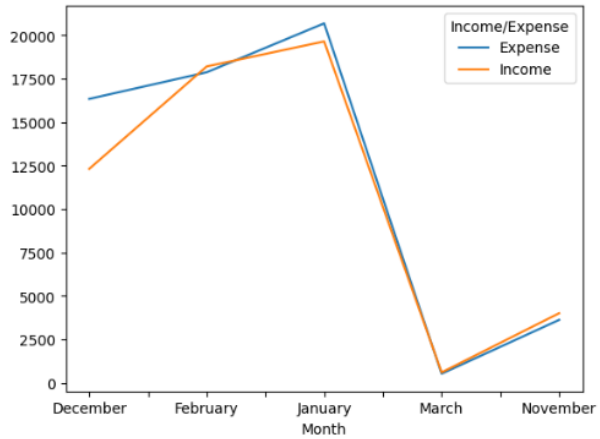
```
df.pivot_table(index='Month',columns='Category',
                values='INR',aggfunc='sum',fill_value=0
               ).plot()
```

```
<AxesSubplot: xlabel='Month'>
```




```
df.pivot_table(index='Month',columns='Income/Expense',
               values='INR',aggfunc='sum',fill_value=0
               ).plot()
```

```
<AxesSubplot: xlabel='Month'>
```



VECTORIZED STRING OPERATIONS

Where we can apply this topic?

Apply on textual dataset like movies description datasets, customer's reviews datasets, WhatsApp chat analysis review etc.

What are vectorized operations?

- **Here, a is a NumPy array with the elements [1, 2, 3, 4].**
- **The operation a * 4 is a vectorized operation. It multiplies each element of the array a by 4.**
- **The result is a new NumPy array where each element is the product of the corresponding element in the original array a and the scalar value is 4.**

```
import numpy as np
import pandas as pd
# Example of vectorized operation
a = np.array([1,2,3,4])
a * 4
```

```
array([ 4,  8, 12, 16])
```

Problems in vectorized operation in vanilla python

vanilla python:

- **It refers to the core, basic, or standard implementation of the Python programming language without any additional libraries or frameworks.**
- **It is the pure, unmodified form of Python, as defined by the Python Software Foundation.**
- **If we have None, missing values, and null values in datasets, in this scenario, Python functionality cannot handle.**
- **Vanilla Python may not be optimized for handling large datasets efficiently.**

```
# try to apply vectories operation on string
s = ['cat','mat',None,'rat']
[i.startswith('c') for i in s]
```

```
---------------------------------------------------------------
------------------
AttributeError                            Traceback (most
recent call last)
Cell In [96], line 3
      1 # try to apply vectories operation on string
      2 s = ['cat','mat',None,'rat']
----> 3 [i.startswith('c') for i in s]

Cell In [96], line 3, in <listcomp>(.0)
      1 # try to apply vectories operation on string
      2 s = ['cat','mat',None,'rat']
----> 3 [i.startswith('c') for i in s]

AttributeError: 'NoneType' object has no attribute 'start
swith'
```

How pandas solve this issue?

- **To solve this above issue is use .str accessor**
- **In pandas, the .str accessor is used to perform vectorized string operations on a Pandas Series containing strings.**
- **This accessor provides a collection of methods that allow you to manipulate strings efficiently without using explicit loops.**
- **The .str accessor simplifies and speeds up the process of working with string data in Pandas DataFrames, making it a powerful tool for data cleaning and manipulation.**

```
# apply vecorized string operation
s = pd.Series(['cat','mat',None,'rat'])
# string accessor
s.str.startswith('c')
```

```
0     True
1    False
2     None
3    False
dtype: object
```

Apply vectorized string operations in titanic dataset

```
df = pd.read_csv('DATASETS/S22/titanic.csv')
df.head(3)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0

```
# Choose 'Name' column for operations
df['Name']
```

```
0                              Braund, Mr. Owen Harris
1    Cumings, Mrs. John Bradley (Florence Briggs Th...
2                               Heikkinen, Miss. Laina
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                             Allen, Mr. William Henry
                           ...
886                              Montvila, Rev. Juozas
887                       Graham, Miss. Margaret Edith
888           Johnston, Miss. Catherine Helen "Carrie"
889                              Behr, Mr. Karl Howell
890                                Dooley, Mr. Patrick
Name: Name, Length: 891, dtype: object
```

Common .str methods/function:

Nearly all Python's built-in string methods are mirrored by a Pandas vectorized string method.

.str.lower() / .str.upper() / .str.capitalize() / .str.title() / .str.len()

```
df['Name'].str.lower().tail(3)
```

```
888    johnston, miss. catherine helen "carrie"
889                       behr, mr. karl howell
890                         dooley, mr. patrick
Name: Name, dtype: object
```

```
df['Name'].str.upper().tail(3)
```

```
888    JOHNSTON, MISS. CATHERINE HELEN "CARRIE"
889                       BEHR, MR. KARL HOWELL
890                         DOOLEY, MR. PATRICK
Name: Name, dtype: object
```

```
df['Name'].str.capitalize().tail(3)
```

```
888    Johnston, miss. catherine helen "carrie"
889                       Behr, mr. karl howell
890                         Dooley, mr. patrick
Name: Name, dtype: object
```

```
df['Name'].str.title().tail(3)
```

```
888    Johnston, Miss. Catherine Helen "Carrie"
889                       Behr, Mr. Karl Howell
890                         Dooley, Mr. Patrick
Name: Name, dtype: object
```

```
df['Name'].str.len().head()
```

```
0    23
1    51
2    22
3    44
4    24
Name: Name, dtype: int64
```

```
# find longest name
df['Name'][df['Name'].str.len() == 82].values[0]
```

```
'Penasco y Castellana, Mrs. Victor de Satode (Maria Josef
a Perez de Soto y Vallejo)'
```

.str.strip() method:

- **The .str.strip() method removes leading and trailing whitespace from each string in the Series.**
- **For the element with None and empty string, it becomes NaN after stripping.**
- **Useful in nlp projects.**

```
df['Name'].str.strip().tail()
```

```
886                     Montvila, Rev. Juozas
887              Graham, Miss. Margaret Edith
888  Johnston, Miss. Catherine Helen "Carrie"
889                     Behr, Mr. Karl Howell
890                       Dooley, Mr. Patrick
Name: Name, dtype: object
```

.str.split() method

- The .str.split() method splits each string in the Series into a list of substrings.
- If you want to split the string based on a specific delimiter, you can provide that delimiter as an argument to .str.split(). For example, to split based on a comma, you can use s.str.split(',').
- Syntax: .str.split(pat=None, n=-1, expand=False)

Parameters:

n:	Specifies the maximum number of splits to perform. By default, it is set to -1, which means there is no limit to the number of splits.
Expand:	Controls whether to expand the result into a DataFrame. If expand is set to True, the result will be a DataFrame with one column per split. If False (the default), the result is returned as a Series of lists.

.str.get() method:

- The .str.get() method in pandas is used to get the element at a specified position for each string in a Pandas Series of strings.
- It is a vectorized string method, meaning it operates on each element of the Series without requiring explicit loops.

```
df['Name'].head()
```

```
0                              Braund, Mr. Owen Harris
1    Cumings, Mrs. John Bradley (Florence Briggs Th...
2                               Heikkinen, Miss. Laina
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                             Allen, Mr. William Henry
Name: Name, dtype: object
```

```
# Create lastname columns
df['Lastname'] = df['Name'].str.split(',').str.get(0)
df.head(3)
```

Parch	Ticket	Fare	Cabin	Embarked	Lastname	Firstname	Title
0	A/5 21171	7.2500	NaN	S	Braund	Mr. Owen Harris	Mr.
0	PC 17599	71.2833	C85	C	Cumings	Mrs. John Bradley (Florence Briggs Thayer)	Mrs.
0	STON/O2. 3101282	7.9250	NaN	S	Heikkinen	Miss. Laina	Miss.

```
df['Firstname'] = df['Name'].str.split(',').str.get(1)
df.head(2)
```

Sp	Parch	Ticket	Fare	Cabin	Embarked	Lastname	Firstname	Title
1	0	A/5 21171	7.2500	NaN	S	Braund	Mr. Owen Harris	Mr.
1	0	PC 17599	71.2833	C85	C	Cumings	Mrs. John Bradley (Florence Briggs Thayer)	Mrs.

.str.replace() method:

```
# Count titles
df['Title'].value_counts().head(10)
```

```
Mr.        517
Miss.      185
Mrs.       125
Master.     40
Dr.          7
Rev.         6
Major.       2
Col.         2
Don.         1
Mme.         1
Name: Title, dtype: int64
```

```
# replace the title 'Ms' and 'Mlle' with 'Miss'
df['Title'] = df['Title'].str.replace('Ms','Miss')
df['Title'] = df['Title'].str.replace('Mlle','Miss')
```

```
df['Title'].value_counts().head(10)
```

```
Mr.        517
Miss.      185
Mrs.       125
Master.     40
Dr.          7
Rev.         6
Major.       2
Col.         2
Don.         1
Mme.         1
Name: Title, dtype: int64
```

Filtering

In pandas, vectorized string operations can be combined with boolean indexing to filter data based on string conditions efficiently.

.str.statswith() / .str.endswith() / .str.isalpha() / .str.isdigit

```
# Find Firstname starts with 'A'
bool_df = df['Firstname'].str.startswith('A')
df[bool_df]['Firstname'].head()
```

```
13       Anders Johan
22       Anna "Annie"
35    Alexander Oskar
38      Augusta Maria
61             Amelie
Name: Firstname, dtype: object
```

```
# Find Firstname ends with 'A'
bool_df = df['Firstname'].str.endswith('A')
df[bool_df]['Firstname'].head()
```

```
64     Albert A
303      Nora A
Name: Firstname, dtype: object
```

```
# Find the name it include digits ?
bool_df = df['Firstname'].str.isdigit()
df[bool_df] # digit not found
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticke

Advanced Filtering with regex and .str.contains() method

.str.contains() method:

The .str.contains() method in pandas is used to check if each element in a Pandas Series of strings contains a specified substring or matches a regular expression pattern.
This method returns a boolean mask indicating whether the condition is met for each element in the Series.

Syntax: .str.contains(pat, case=True, flags=0, na=nan, regex=True)

Parameters:

pat:	The substring or regular expression pattern to search for.
case:	If True, the matching is case-sensitive (default is True).
flags:	Additional flags for controlling the behavior of the regex (see the re module for options).
na:	The value to use for missing values (default is nan).
regex:	If True, treats the pat parameter as a regular expression (default is True).

```python
# Find the john name
bool_df = df['Firstname'].str.contains('john',case=False)
df[bool_df]['Firstname'].head()
```

```
1            John Bradley (Florence Briggs Thayer)
41     William John Robert (Dorothy Ann Wonnacott)
45                                    William John
98                          John T (Ada Julia Bone)
112                                    David John
Name: Firstname, dtype: object
```

```python
# Find Lastname with start and end char is vowel
pattern = '^[aeiouAEIOU].+[aeiouAEIOU]$'
bool_df = df['Lastname'].str.contains(pattern,case=False)
df[bool_df]['Lastname']
```

```
30          Uruchurtu
49     Arnold-Franchi
207          Albimona
210               Ali
353    Arnold-Franchi
493      Artagaveytia
518             Angle
784               Ali
840          Alhomaki
Name: Lastname, dtype: object
```

Slicing

- **In pandas, vectorized string operations allow you to perform slicing on each element of a Pandas Series of strings efficiently.**
- **You can use the .str accessor along with indexing or slicing to extract substrings based on position or conditions.**

```python
# basic slicing like pythons string
df['Name'].str[1:4].head()
```

```
0    rau
1    umi
2    eik
3    utr
4    lle
Name: Name, dtype: object
```

DATETIME

Timestamp Object

What is Timestamp?

- Timestamp refers as particular moments in time (e.g., Oct 24th. 2022 at 7:00pm)
- Pandas library provide different datatype to store that datatype

What is Timestamp object?

- Timestamp object is part of the pandas library and is particularly useful when working with time series data.
- Patient health metrics, stock price changes, weather records, economic indicators, servers, networks, sensors, and applications performance monitoring are examples of time-series data.
- We can also called vectorized datetime operations or vectorized datetime functions.

Creating Timestamp object

- Create a Timestamp object using the pd.Timestamp constructor.
- It can handle various input formats, including strings, datetime objects, or even numeric values representing timestamps.

Note: Always try to follow this date format: 'YYYY/MM/DD'

```
import numpy as np
import pandas as pd
# Creating a timestamp
pd.Timestamp('2023/10/18')
```

```
Timestamp('2023-10-18 00:00:00')
```

```
# diffrent variation in timestamp
pd.Timestamp('2023-10-18')
pd.Timestamp('2023, 10, 18')
```

```
Timestamp('2023-10-18 00:00:00')
```

```
# only year
pd.Timestamp('2023')
```

```
Timestamp('2023-01-01 00:00:00')
```

```
# use text
pd.Timestamp('18th oct 2023')
```

```
Timestamp('2023-10-18 00:00:00')
```

```
# Providing time as well
pd.Timestamp('2023, 10, 18, 8:12PM')
```

```
Timestamp('2023-10-18 20:12:00')
```

```
# creating timestamp using python datetime object
import datetime as dt
dt_object = dt.datetime(2023, 10, 18, 8, 18, 56)
t = pd.Timestamp(dt_object)
t
```

```
Timestamp('2023-10-18 08:18:56')
```

```
# fetching attributes like year,month,year etc
print(t.year)
print(t.month)
print(t.day)
print(t.hour)
print(t.minute)
print(t.second)
```

```
2023
10
18
8
18
56
```

(INTERVIEW QUESTION)

why separate objects to handle data and time when python already has datetime functionality?

- syntax wise datetime is very convenient
- But the performance takes a hit while working with huge data. List vs Numpy Array
- The weaknesses of Python's datetime format inspired the NumPy team to add a set of native time series data type to NumPy.
- The datetime64 dtype encodes dates as 64-bit integers, and thus allows arrays of dates to be represented very compactly.

```
import numpy as np
date = np.array('2015-07-04', dtype=np.datetime64)
date
```

```
array('2015-07-04', dtype='datetime64[D]')
```

```
# suporting vectorized operations
date + np.arange(6)
```

```
array(['2015-07-04', '2015-07-05', '2015-07-06', '2015-07-07',
       '2015-07-08', '2015-07-09'], dtype='datetime64[D]')
```

- Because of the uniform type in NumPy datetime64 arrays, this type of operation can be accomplished much more quickly than if we were working directly with Python's datetime objects, especially as arrays get large
- Pandas Timestamp object combines the ease-of-use of python datetime with the efficient storage and vectorized interface of numpy.datetime64
- From a group of these Timestamp objects, Pandas can construct a DatetimeIndex that can be used to index data in a Series or DataFrame

DatetimeIndex Objects

- In simple words, A collection of pandas timestamp
- Pandas DatetimeIndex makes it easier to work with Date and Time data in our DataFrame.
- DatetimeIndex() can contain metadata related to date and timestamp and is a great way to deal with DateTime related data and do the calculations on data and time.
- Syntax:
  pd.DatetimeIndex(data=None, freq=_NoDefault.no_default, tz=None, normalize=False, closed=None, ambiguous='raise', dayfirst=False, yearfirst=False, dtype=None, copy=False, name=None)
  Parameters:

data:	The data to be converted to datetime. It can be a list, array, or Series containing datetime-like objects, or a single datetime-like object.
freq:	The frequency of the datetime values if the data is not already a time series. It can be a string (e.g., 'D' for daily, 'H' for hourly) or a Timedelta object.
tz:	Timezone for the datetime values.
normalize:	If True, normalize the datetime values (set time to midnight).
closed:	Specify whether the interval is left-closed ('left'), right-closed ('right'), both closed ('both'), or neither closed ('neither').
ambiguous:	How to handle daylight savings time ambiguities. Default is 'raise', but it can also be set to 'infer' or 'NaT'.
dayfirst:	If True, parse dates with the day first (e.g., '10/12/2023' is October 12, 2023).
yearfirst:	If True, parse dates with the year first (e.g., '2023-10-12' is October 12, 2023).
dtype:	The dtype of the datetime values.
copy:	If True, ensure that the input data is copied. Default is False.
name:	Name to be assigned to the resulting DatetimeIndex.

```
# collection of timestaime object
d = pd.DatetimeIndex(['2023-10-18','2023-10-19'])
print(d)
```

```
DatetimeIndex(['2023-10-18', '2023-10-19'], dtype='datetim
e64[ns]', freq=None)
```

```
# checking type
print(type(d))
print(type(d[0]))
```

```
<class 'pandas.core.indexes.datetimes.DatetimeIndex'>
<class 'pandas._libs.tslibs.timestamps.Timestamp'>
```

```
# using python datetime object
dates = [dt.datetime(2023,10,18),dt.datetime(2023,10,19)]
pd.DatetimeIndex(dates)
```

```
DatetimeIndex(['2023-10-18', '2023-10-19'], dtype='datetim
e64[ns]', freq=None)
```

```
# using pd.Timestamp
dates = [pd.Timestamp(2023,10,18),pd.Timestamp(2023,10,19)]
dt_index = pd.DatetimeIndex(dates)
dt_index
```

```
DatetimeIndex(['2023-10-18', '2023-10-19'], dtype='datetim
e64[ns]', freq=None)
```

```
# using DatetimeIndex as Series
pd.Series([1,2], index=dt_index)
```

```
2023-10-18    1
2023-10-19    2
dtype: int64
```

date_range() function

- The pd.date_range() function in Pandas is used to generate a fixed-frequency DatetimeIndex.
- It's a convenient method for creating date sequences for time-based data, such as time series data.
- The function allows you to specify the start date, end date, and frequency of the date range.
- Syntax: pd.date_range(start=None, end=None, periods=None, freq=None, tz=None, normalize=False, name=None, closed=None, **kwargs)

Parameters:

start:	The start date of the range.
end:	The end date of the range.
periods:	The number of periods (int) to generate.
freq:	Frequency of the resulting date sequence. This can be a string representing a frequency alias (e.g., 'D' for day, 'H' for hour), a Timedelta object, or a custom frequency string.
tz:	Timezone for the datetime values.
normalize:	If True, normalize the datetime values (set time to midnight).
name:	Name to be assigned to the resulting DatetimeIndex.
closed:	Specify whether the interval is left-closed ('left'), right-closed ('right'), both closed ('both'), or neither closed ('neither').
**kwargs:	Additional keyword arguments that are passed to the underlying DatetimeIndex constructor.

```
# generate daily dates in given range
pd.date_range(start='2023/10/18',end='2023/10/21')
```

```
DatetimeIndex(['2023-10-18', '2023-10-19', '2023-10-20', '20
23-10-21'], dtype='datetime64[ns]', freq='D')
```

```
# alternative dates in given range
pd.date_range(start='2023/10/18',
              end='2023/10/25', freq='2D')
```

```
DatetimeIndex(['2023-10-18', '2023-10-20', '2023-10-22', '20
23-10-24'], dtype='datetime64[ns]', freq='2D')
```

```
# B --> business days
pd.date_range(start='2023/10/18',
              end='2023/10/25',freq='B')
```

```
DatetimeIndex(['2023-10-18', '2023-10-19', '2023-10-20', '20
23-10-23',
               '2023-10-24', '2023-10-25'],
              dtype='datetime64[ns]', freq='B')
```

```
# W --> One week per days
pd.date_range(start='2023/10/18',
              end='2023/10/28',freq='W')
```

```
DatetimeIndex(['2023-10-22'], dtype='datetime64[ns]', freq
='W-SUN')
```

```
# M -> Month end
pd.date_range(start='2023/10/18',
              end='2023/11/18',freq='M')
```

```
DatetimeIndex(['2023-10-31'], dtype='datetime64[ns]', freq
='M')
```

```
# MS -> Starting month
pd.date_range(start='2023/10/18',
              end='2023/11/18',freq='MS')
```

```
DatetimeIndex(['2023-11-01'], dtype='datetime64[ns]', freq
='MS')
```

```
# H -> Hourly data(factor)
pd.date_range(start='2023/10/18',
              end='2023/10/28',freq='12H')
```

```
DatetimeIndex(['2023-10-18 00:00:00', '2023-10-18 12:00:00',
               '2023-10-19 00:00:00', '2023-10-19 12:00:00',
               '2023-10-20 00:00:00', '2023-10-20 12:00:00',
               '2023-10-21 00:00:00', '2023-10-21 12:00:00',
               '2023-10-22 00:00:00', '2023-10-22 12:00:00',
               '2023-10-23 00:00:00', '2023-10-23 12:00:00',
               '2023-10-24 00:00:00', '2023-10-24 12:00:00',
               '2023-10-25 00:00:00', '2023-10-25 12:00:00',
               '2023-10-26 00:00:00', '2023-10-26 12:00:00',
               '2023-10-27 00:00:00', '2023-10-27 12:00:00',
               '2023-10-28 00:00:00'],
              dtype='datetime64[ns]', freq='12H')
```

```
# A -> Year end
pd.date_range(start='2023/10/18',
              end='2025/11/28',freq='A')
```

```
DatetimeIndex(['2023-12-31', '2024-12-31'], dtype='dateti
me64[ns]', freq='A-DEC')
```

```
# using periods(number of results)
pd.date_range(start='2023/10/18',periods=25,freq='M')
```

```
DatetimeIndex(['2023-10-31', '2023-11-30', '2023-12-31',
'2024-01-31',
               '2024-02-29', '2024-03-31', '2024-04-30',
'2024-05-31',
               '2024-06-30', '2024-07-31', '2024-08-31',
'2024-09-30',
               '2024-10-31', '2024-11-30', '2024-12-31',
'2025-01-31',
               '2025-02-28', '2025-03-31', '2025-04-30',
'2025-05-31',
               '2025-06-30', '2025-07-31', '2025-08-31',
'2025-09-30',
               '2025-10-31'],
              dtype='datetime64[ns]', freq='M')
```

to_datetme() function:

- The pd.to_datetime() function in Pandas is used to convert an object to a datetime.
- It can be used to convert a wide variety of input types, including strings, integers, floats, and other datetime-like objects, into Pandas datetime objects, such as Timestamp or DatetimeIndex.
- In simple words, Converts an existing objects to pandas timestamp/datetimeindex object
- Syntax: pd.to_datetime(arg, errors='raise', format=None, dayfirst=False, yearfirst=False, utc=None, format2=None, exact=True, unit=None, infer_datetime_format=False, origin='unix', cache=False)

Parameters:

arg:	The object to be converted to a datetime. It can be a single value or an iterables (e.g., list, array, or Series).
errors:	How to handle parsing errors. It can be set to 'raise' (default), 'coerce' (to force errors to NaT), or 'ignore' (to skip errors).
format:	A format string to specify the exact format of the input data.
dayfirst:	If True, parse dates with the day first (e.g., '10/12/2023' is October 12, 2023).
yearfirst:	If True, parse dates with the year first (e.g., '2023-10-12' is October 12, 2023).
utc:	If True, return UTC datetime objects.
unit:	The unit of the input data if the input is numeric (e.g., 's' for seconds or 'ns' for nanoseconds).
Infer_datetime_format:	If True, infer the datetime format of the input data (can improve parsing performance).
origin:	A reference date for numeric time data (default is 'unix' for Unix timestamps).
cache:	Whether to cache the datetime conversion results for performance improvement.

```
# simple example
s = pd.Series(['2015-07-04','2015-08-05','2015-09-06'])
# Series dtype is string
print(s)
```

```
0    2015-07-04
1    2015-08-05
2    2015-09-06
dtype: object
```

```
# convert into datatime object
d = pd.to_datetime(s)
print(d)
# access of year, month etc.
print(d.dt.year)
print(d.dt.month)
```

```
0   2015-07-04
1   2015-08-05
2   2015-09-06
dtype: datetime64[ns]
0    2015
1    2015
2    2015
dtype: int64
0    7
1    8
2    9
dtype: int64
```

```
# if date format is invalid and try to convert
s = pd.Series(['2015-07-04','2015-07-43','2015-17-06'])
pd.to_datetime(s)
```

```
ParserError: day is out of range for month: 2015-07-43
present at position 1
```

```
# Handling parsing errors using the errors parameter:
s = pd.Series(['2015-07-04','2015-07-43','2015-17-06'])
pd.to_datetime(s,errors='coerce')
```

```
0   2015-07-04
1          NaT
2          NaT
dtype: datetime64[ns]
```

```
df = pd.read_csv('DATASETS/S22/expense_data.csv')
df.head()
```

	Date	Account	Category	Subcategory	Note	INR	Incom
0	3/2/2022 10:11	CUB - online payment	Food	NaN	Brownie	50.0	
1	3/2/2022 10:11	CUB - online payment	Other	NaN	To lended people	300.0	
2	3/1/2022 19:50	CUB - online payment	Food	NaN	Dinner	78.0	
3	3/1/2022 18:56	CUB - online payment	Transportation	NaN	Metro	30.0	
4	3/1/2022 18:22	CUB - online payment	Food	NaN	Snacks	67.0	

```
df['Date'].info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 277 entries, 0 to 276
Series name: Date
Non-Null Count  Dtype
--------------  -----
277 non-null    object
dtypes: object(1)
memory usage: 2.3+ KB
```

```
# convert Date column dtype 'object' into 'datatime64'
df['Date'] = pd.to_datetime(df['Date'])
df['Date'].info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 277 entries, 0 to 276
Series name: Date
Non-Null Count  Dtype
--------------  -----
277 non-null    datetime64[ns]
dtypes: datetime64[ns](1)
memory usage: 2.3 KB
```

.dt accsessor:

- **In Pandas, the .dt accessor is used to access the datetime components of a Series or DataFrame.**
- **The .dt accessor provides a convenient way to work with date and time components like year, month, day, hour, minute, second, etc.**

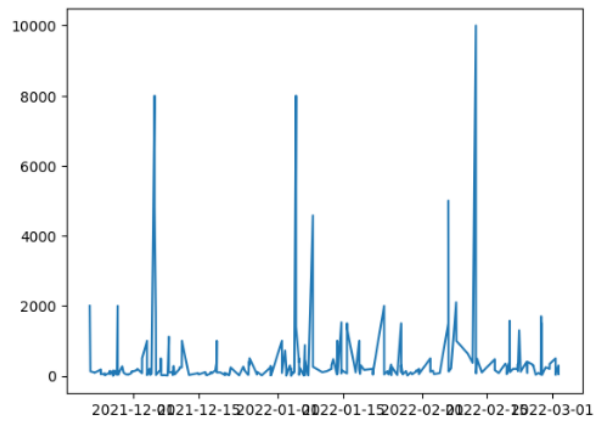
```
print(df['Date'].dt.year.head(3))
print(df['Date'].dt.month.head(3))
print(df['Date'].dt.day.head(3))
print(df['Date'].dt.day_name().head(3))
print(df['Date'].dt.month_name().head(3))
print(df['Date'].dt.is_month_end.head(3))
print(df['Date'].dt.is_quarter_end.head(3))
```

```
0    2022
1    2022
2    2022
Name: Date, dtype: int64
0    3
1    3
2    3
Name: Date, dtype: int64
0    2
1    2
2    1
Name: Date, dtype: int64
0    Wednesday
1    Wednesday
2      Tuesday
Name: Date, dtype: object
0    March
1    March
2    March
Name: Date, dtype: object
0    False
1    False
2    False
Name: Date, dtype: bool
0    False
1    False
2    False
Name: Date, dtype: bool
```

Plotting Graphs:

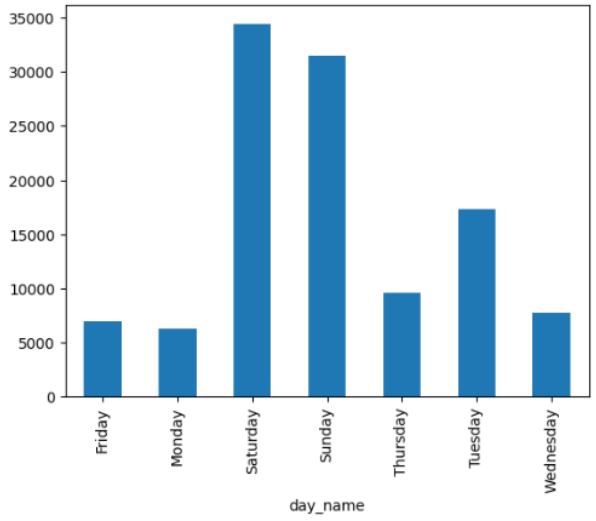
```
import matplotlib.pyplot as plt
plt.plot(df['Date'],df['INR'])
```

```
[<matplotlib.lines.Line2D at 0x25f4bc09600>]
```



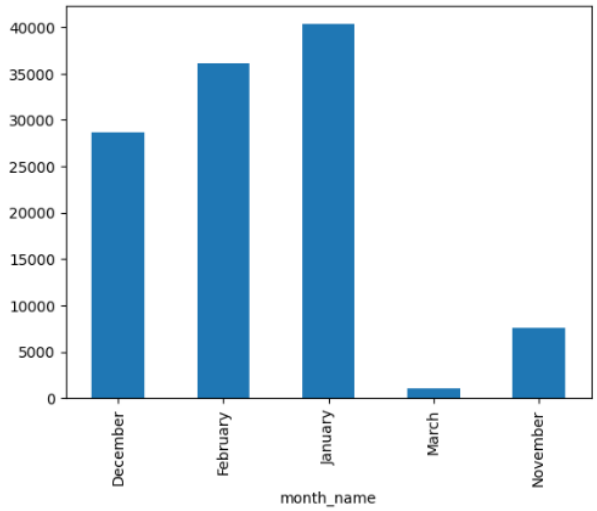

```
# daywise spending
df['day_name'] = df['Date'].dt.day_name()
df.groupby('day_name')['INR'].sum().plot(kind='bar')
```

```
<AxesSubplot: xlabel='day_name'>
```




```
# monthly spending
df['month_name'] = df['Date'].dt.month_name()
df.groupby('month_name')['INR'].sum().plot(kind='bar')
```

```
<AxesSubplot: xlabel='month_name'>
```



Timedelta Object (Pandas)

- Represents a duration, the difference between two dates or times.
- In pandas, a Timedelta object represents a duration, which is the difference between two dates, times, or a combination of both.
- It is useful to work with time-based data and perform various time-related operations.
- We can create a Timedelta object using the pd.Timedelta constructor and provide the duration as an argument, which can be expressed in various units such as days, hours, minutes, seconds, milliseconds, microseconds, and nanoseconds.

```
import pandas as pd
# Create Timedelta object using Timestamp object
td1 = pd.Timestamp('20th Oct 2023')
td2 = pd.Timestamp('20th Nov 2023')

# diff_td is Timedelta object
diff_td = td2 - td1
print(diff_td)
```

```
31 days 00:00:00
```

pd.Timedelta(value, unit):

- The pd.Timedelta() function in pandas is used to create a Timedelta object, which represents a duration or time difference.
- We can specify the duration as arguments to this function, using various units such as days, hours, minutes, seconds, milliseconds, microseconds, and nanoseconds.
- Syntax: pd.Timedelta(value, unit)

Parameters:

value:	The numerical value that represents the duration or time difference.
unit:	A string that specifies the unit of time. This can be one of the following: 'D' or 'days' for days 'H' or 'hours' for hours 'T' or 'minutes' for minutes 'S' or 'seconds' for seconds 'L' or 'milliseconds' for milliseconds 'U' or 'microseconds' for microseconds 'N' or 'nanoseconds' for nanoseconds

Note:

Timedelta is the pandas equivalent of python's datetime.timedelta and is interchangeable with it in most cases.

```
# standalone creation using pd.Timedelta()
pd.Timedelta(days=2,hours=10,minutes=35)
```

```
Timedelta('2 days 10:35:00')
```

```
# arithmetic operations
td = pd.Timedelta(days=10,hours=2)
new_date = pd.Timestamp('2023/08/20') + td
print(new_date)
```

```
2023-08-30 02:00:00
```

```
dates = pd.date_range(start='2023/10/18',
                      end='2023/10/25', freq='D')
td = pd.Timedelta(days=2,hours=10,minutes=35)
print(dates + td)
```

```
DatetimeIndex(['2023-10-20 10:35:00', '2023-10-21 1
0:35:00',
               '2023-10-22 10:35:00', '2023-10-23 1
0:35:00',
               '2023-10-24 10:35:00', '2023-10-25 1
0:35:00',
               '2023-10-26 10:35:00', '2023-10-27 1
0:35:00'],
              dtype='datetime64[ns]', freq='D')
```

```
# real life example (90's delivery dataset)
df = pd.read_csv('deliveries.csv')
df.head()
```

	order_date	delivery_date
0	5/24/98	2/5/99
1	4/22/92	3/6/98
2	2/10/91	8/26/92
3	7/21/92	11/20/97
4	9/2/93	6/10/98

```
# Calculating avg time period of delivery
df['order_date'] = pd.to_datetime(df['order_date'])
df['delivery_date'] = pd.to_datetime(df['delivery_date'])
# Timedelta
df['delivery_time'] = df['delivery_date'] - df['order_date']
df['delivery_time'].mean()
```

```
Timedelta('1217 days 22:53:53.532934128')
```