Pandas

- Pandas is a python library.
- Pandas is used to analyze data analysis.
- Pandas is used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Why Use Pandas?

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Relevant data is very important in data science.

Data Types

A data type is used by a programming language to understand how to store and manipulate data. The table below summarizes the different data types in Pandas.

Data type	Use
int	Integer number, eg: 10, 12
float	Floating point number, eg: 100.2, 3.1415
bool	True/False value
object	Test, non-numeric, or a combination of text and non-numeric values, eg: Apple
DateTime	Date and time values
category	A finite list of values

All statistical functions

Function	Description
count()	Returns the number of times an element/data has occurred (non-null)
sum()	Returns sum of all values
mean()	Returns the average of all values
median()	Returns the median of all values
mode()	Returns the mode
std()	Returns the standard deviation
min()	Returns the minimum of all values
max()	Returns the maximum of all values
abs()	Returns the absolute value

Input and Output

Often, you won't be creating data but will be having it in some form, and you would want to import it to run your analysis on it. Fortunately, Pandas allows you to do this. Not only does it help in importing data, but you can also save your data in your desired format using Pandas.

The below table shows the formats supported by Pandas, the function to read files using Pandas, and the function to write files.

Input type	Reader	Writer
CSV	read_csv	to_csv
JSON	read_json	to_json
HTML	read_html	to_html
Excel	read_excel	to_excel
SAS	read_sas	–
Python Pickle Format	read_pickle	to_pickle
SQL	read_sql	to_sql
Google Big Query	read_gbq	to_gbq

It supports two data structures:

- Series
- Dataframe

What is a Series?

A Pandas Series is like a column in a table.

It is a one-dimensional array holding data of any type.

DataFrames

Data sets in Pandas are usually multi-dimensional tables, called DataFrames.

Series is like a column, a DataFrame is the whole table.

Features of DataFrame

- Potentially columns are of different types
- Size – Mutable
- Labeled axes (rows and columns)
- Can Perform Arithmetic operations on rows and columns

Create DataFrame

A pandas DataFrame can be created using various inputs like –

- Lists
- dict
- Series
- Numpy ndarrays
- Another DataFrame

##Pandas use the `loc` attribute to return one or more specified row(s)