# Pandas

Ayesha Sk

SCOPE, VIT Chennai

# Introduction-What is Pandas?

- Pandas is a Python library 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.

# Introduction-Why?

- 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.

# What can Pandas do?

- Is there a correlation between two or more columns?
- What is average value?
- Max value? Min value?
- Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called cleaning the data.

# Dataframe creation

- From own data through dictionary formats
  - df=pandas.dataframe(dictionary)
  - df=pandas.dataframe(dictionary,index=[...(as many experiences/rows)])
  - dictionary have format of keys and columns
- From existing csv or excel files
  - df=pandas.read_csv(path)
  - to collect data from other sources

# Databases can be created from

- own random data
- csv, xls, xlsxx
- json: structures can be very well mapped to dictionary
- tables stored in sqlite: sqlite is package supported by python in order to have relational database structure

# Pandas dataframe to save

- Data frame can be stored back as csv,excel file, sqlite tables, json data

# Dataframe operations

- **info** → to get description of data
- **shape** → row, columns
- **head** → to read first few records
- **tail** → to read last few records
- **drop_duplicates** → removing duplicates in a row
  - **inplace** =True→ Operations to be performed and stored back there itself
  - **keep**=(first/last/false) →
- **columns** → to point attributes of the dataset
- **rename** → renaming column names
- **isnull()** → check any value is NULL in dataset

# Common operations-Cont..,

- **dropna** $\rightarrow$ dropping NULL values
  - **axis** $=0/1\rightarrow$ rows drop/columns drop
- **describe** $\rightarrow$ to print complete statistical description of the dataset
- **mean** $\rightarrow$
- **median** $\rightarrow$
- **std** $\rightarrow$
- **var** $\rightarrow$
- **min** $\rightarrow$
- **max** $\rightarrow$
- **concat** $\rightarrow$ to concatenate two dataframes
  - **Ignore_index** $=$True$\rightarrow$ index mapping will be ignored
- **groupby** $\rightarrow$ grouping set of entitites based on some value
- **corr** $\rightarrow$ Correlation for feature prediction

# Dataframe-Slicing

- **loc** $\rightarrow$ by name
- **iloc** $\rightarrow$ by index

# Thank You