AIG150- Week 1

Introduction To Data Wrangling Using Python

Reading Text:

Chap 01,Ch 02:Pandas for everyone

Chap 01-03:Data Wrangling with Python

Agenda

- Data Wrangling
- Python & Data Science
- Introduction to different data analysis tools in Python
- DataFrame
- Data Series

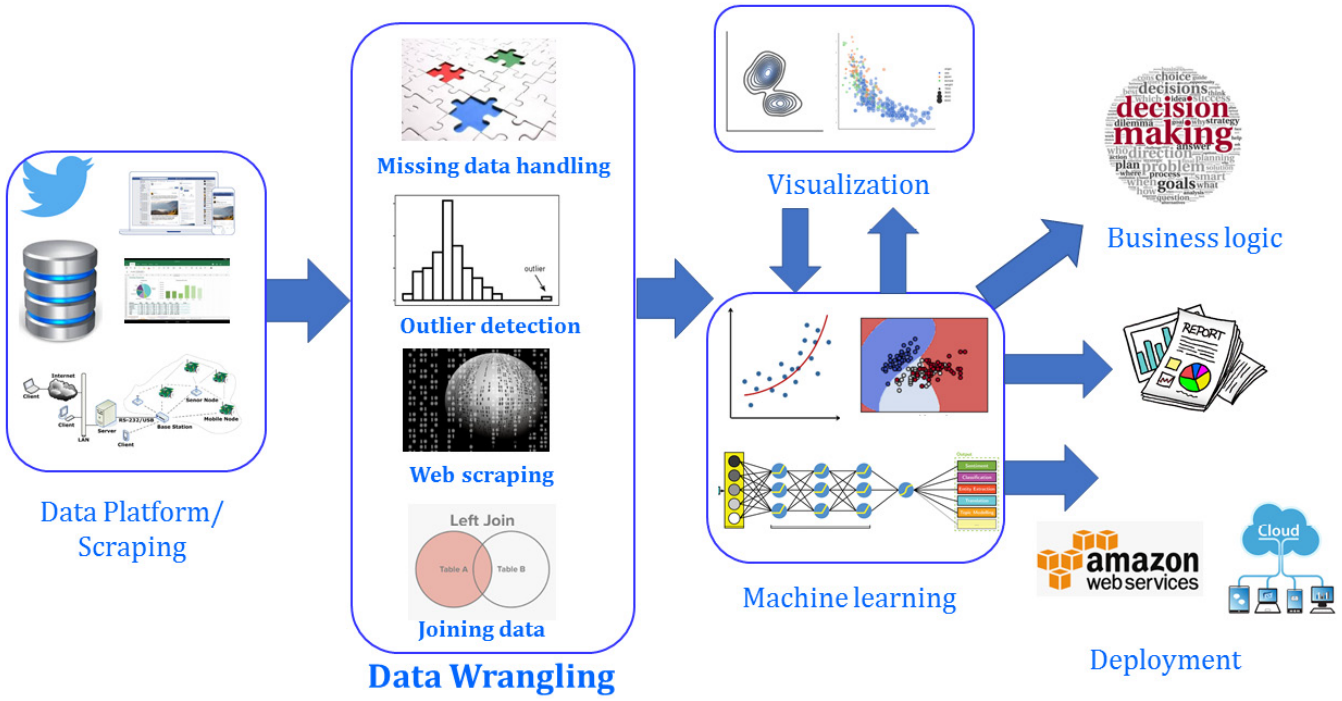
Data Wrangling

- The process that ensures that data is clean, accurate, formatted and ready to be use in data analysis
- Another word used for wrangling is ***"munging"***
- It is the first step of data analytical pipeline after selection of data sources
- Data is coming from different sources such as conventional databases, online feeds, smart devices, satellite imagery and many more

Data Wrangling --- The Process (1)

- Data Collection : Scraping raw data from multiple sources
- Data Transformation : Converting the data in one format to be used by the modelling process (look for missing values, data filtration, adding new data, data conversions…)
- Error Handling
- Outliers Detection
- Data Visualization

Data Wrangling --- The Process (2)



Ref: Chap 01, Data Wrangling With Python

Python & Data Science

- One of the top choices for data analysis
- Enrich in libraries such as *NumPy*, *SciPy*, *Matplotlib* and *pandas*
- Multiprocessing on large datasets --- reducing processing time
- Specialized IDEs

Core Functions of a Data Scientist

- Data Capture
  - ETL – Extract, Transform and Load
  - Python comes handy in extracting data from multiple sources --- CSV Files, DBMS, …
  - Data from multiple sources can then be transformed into a common format and loaded to analyze using Python
- Analysis
- Presentation

Fundamentals In Python -- Review

- List
- Sets
- Dictionaries
- Control Flow Tools
- Basic File Operation
- Import System

Introduction to NumPy, Pandas & Matplotlib

- Numpy offers comprehensive mathematical functions, random number generators, linear algebra routines and many more
- NumPy arrays are optimized data structures for numerical analysis
- Fast, powerful and reliable open-source data analysis tool build on top of basic Python programming language. https://pandas.pydata.org/
- Both NumPy and Pandas have numerous built-in statistical and visualization methods available for data analysis.
- Matplotlib is the most powerful and versatile visualization library in Python.

DataFrame

- The primary pandas data structure designed to work with relational or labeled data
- Two-dimensional, size-mutable, potentially heterogeneous tabular data structure also contains labeled axes (rows and columns)
- Performs arithmetic operations align on both row and column labels
- Can be thought of as a dictionary like container for Series objects
- For full details: pandas.DataFrame

Data Series

- One-dimensional labeled array capable of holding any data type (integers, strings, floating point numbers, Python objects, etc.)
- The axis labels are collectively referred to as the index
- Each column in DataFrame is a series
- Same as Python list