**Chapter 2: DATA PREPROCESSING**

**Topic – 1: Introduction**

**Name Meaning**

* Name **Pandas** was derived from **panel data** & **python data analysis**.

**Getting Version Info**



* Applicable on **all** libraries.

**Topic – 2: Pandas Series**

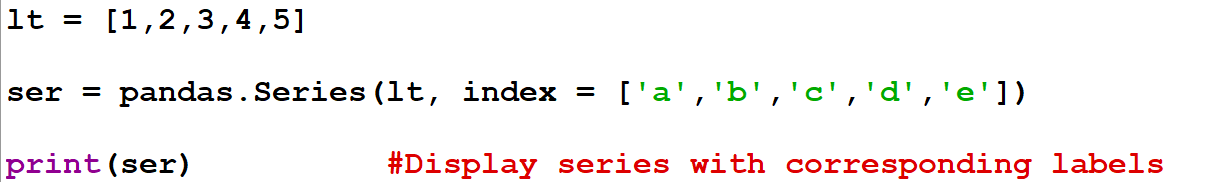
* **Series:** Series in **Pandas** means **column** of a table.
* **Label:** **Index** number of an element in series.
* We convert a given **non-primitive data structure** into a series.
* **Non-primitive data structure:** List, tuple, dictionary etc in **Python**.

**Default Labelling**



**Custom Labelling**

* We can also apply **custom labelling** using index parameter as **var2** shown below.



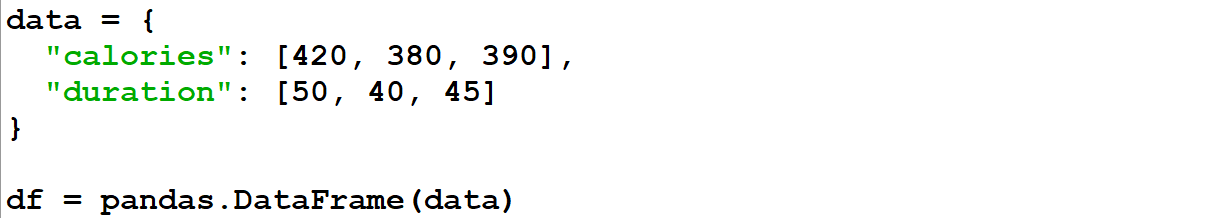
**Topic – 3: Direct Dictionary Labelling**

* For **dictionaries** converted into **series**, the **key’s** **values** are their **labels** by default.



**Topic – 4: Data Frames**

* Are **multi-dimensional** datasets.

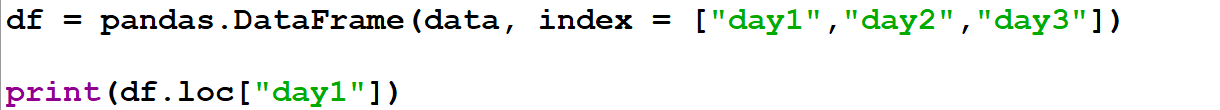


* **"calories"** and **"duration"** are **column** **name** & corresponding **lists** are **values** in those columns.

**Accessing Rows**



**Custom Row Indexing**



* Actually, after **custom indexing** or **labelling** the keys are **overwritten**.
* Thus, data **can’t** be accessed using **index** numbers anymore.

**Topic – 5: Reading CSV Files**

**Reading Limited Rows**

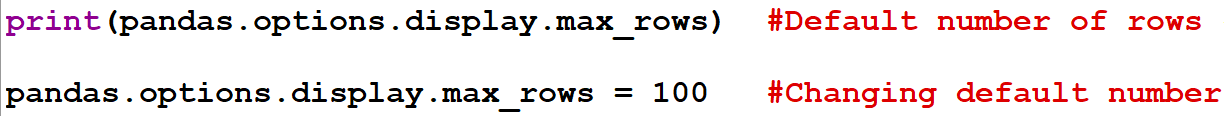


* Returns **5** **first** & **last** **rows** if they are more than a **default** number.

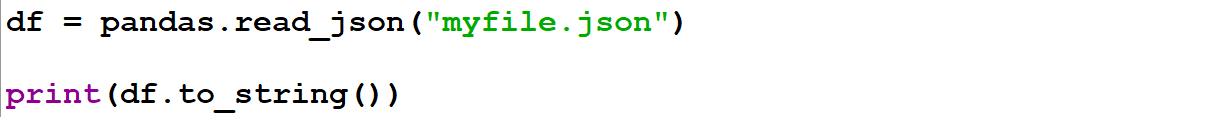
**Reading Entire Dataset**



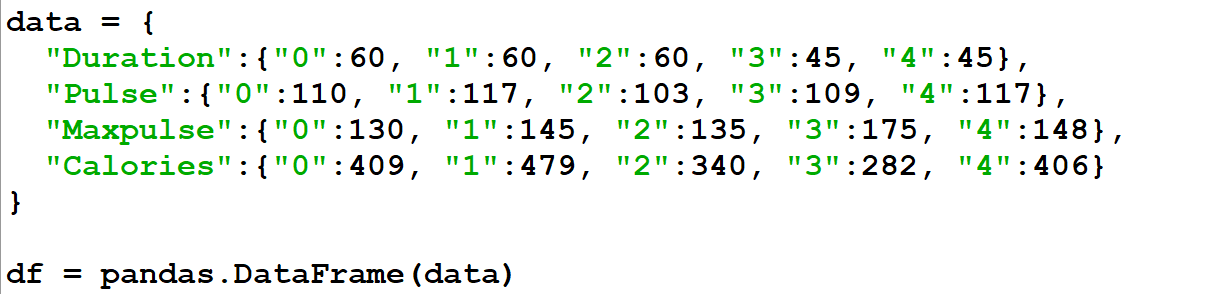
**Setting Rows**



**Topic – 6: Reading JSON Files**



**Custom JSON Dataset**



**Topic – 7: Analysing Data Frames**

* Are applicable to any form of data like **.csv**, **.json** etc.

**Returning Top Rows**

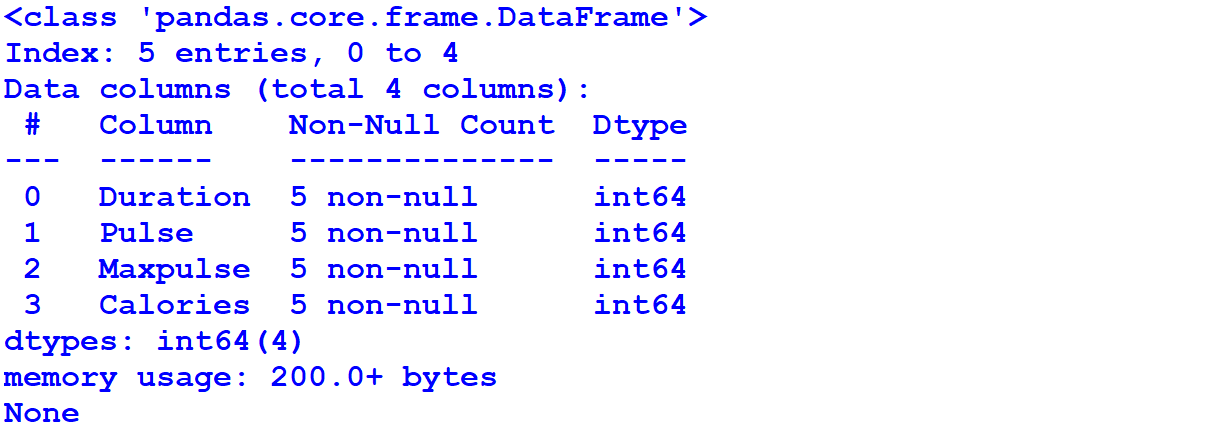


**Returning Bottom Rows**



**Displaying Frame Information**

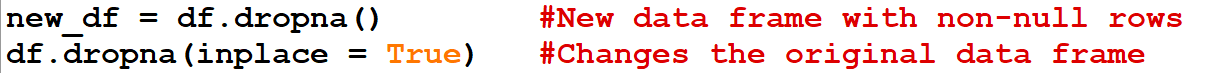




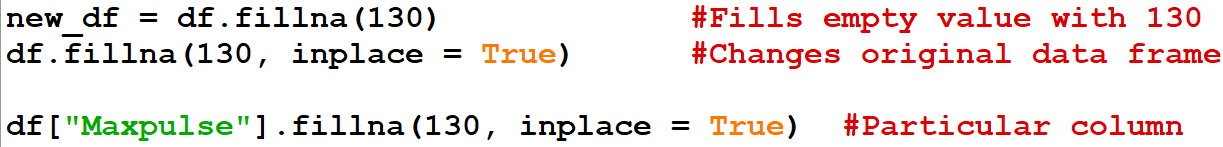
**Topic – 8: Cleaning Data**

* Means **removing** or **making changes** to a **"bad data"** in our dataset.
* **Bad data** can be **empty rows**, **data in wrong format**, **wrong data**, **duplicates** etc.

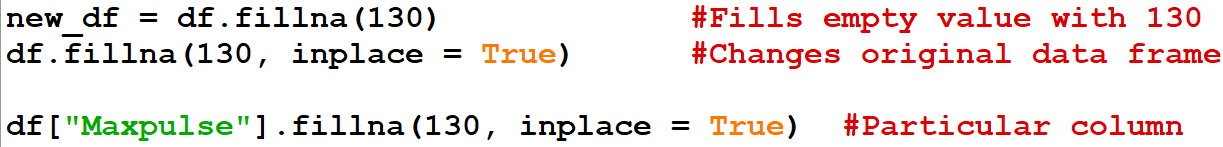
**Removing Empty Values**



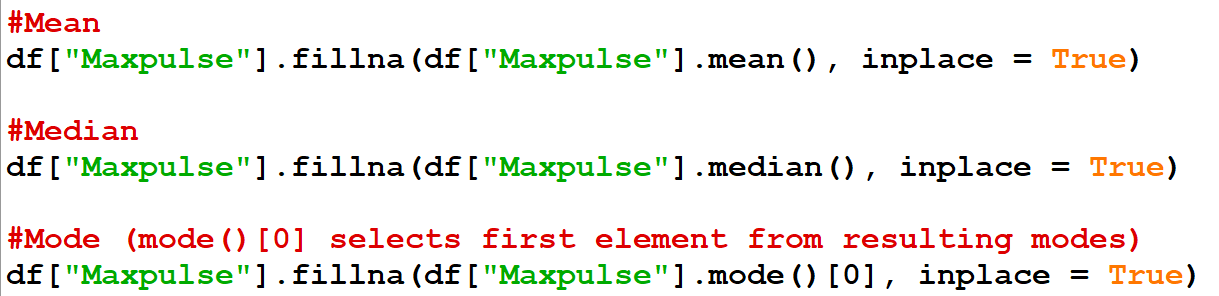
**Topic – 9: Filling Empty Cells**



**Fill Empty Values (Particular Column)**



**Filling Empty Values (With Mean, Median Or Mode)**

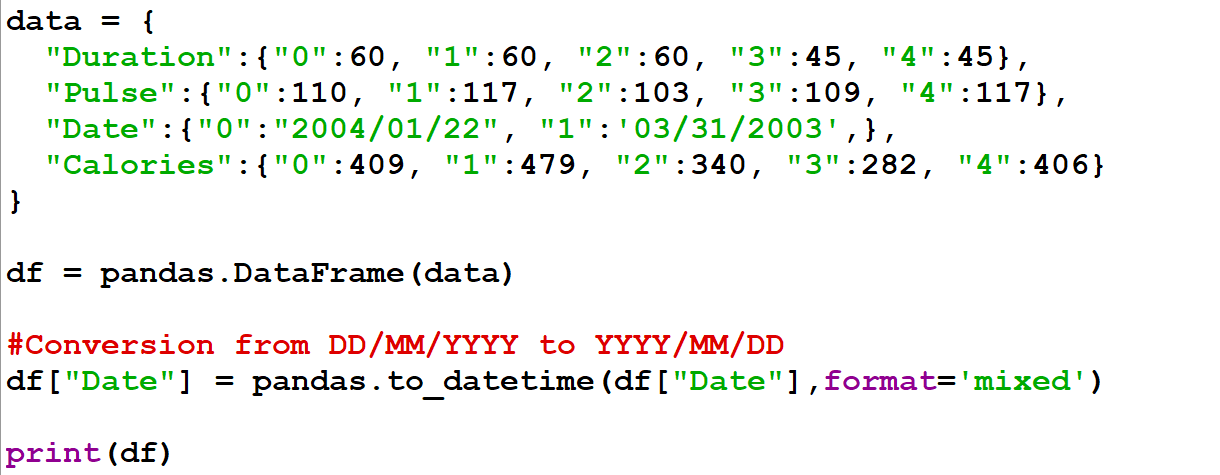


* Not writing **[0]** for **mode()** might result in **unexpected behaviour**, as a particular element from mode series is **not** selected.

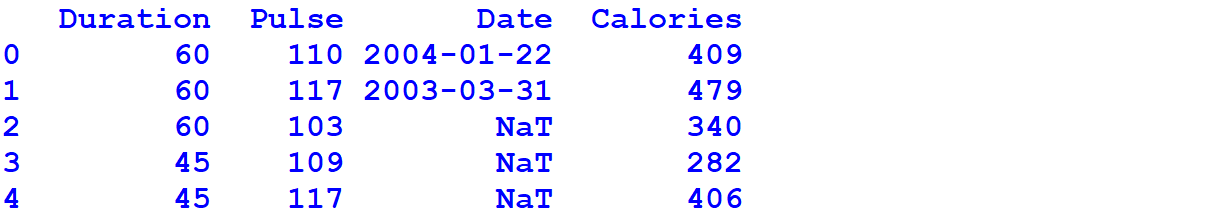
**Topic – 10: Cleaning Wrong Format Data**

**Format Conversion**

* **format='mixed'** tells that **date format** is **mixed** in data.



***\*String to date conversion failed\****

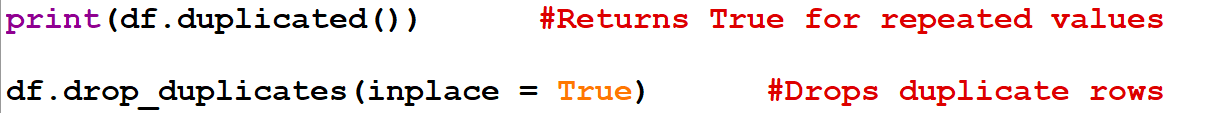


* **NaT** stands for **Not a Time**.



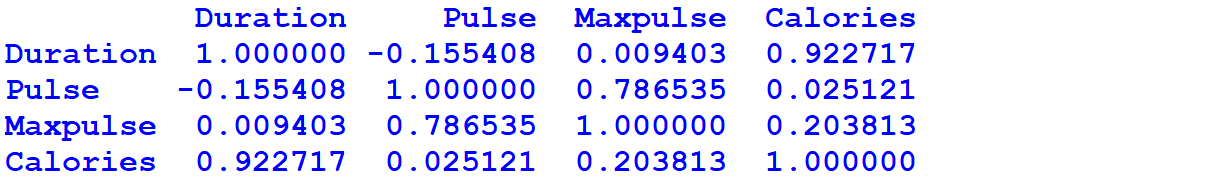
* Note that we have **manually** set our index as **"1"** by **labelling**, which is by default **1**.
* We can also set limits for values of an attribute using a **loop**.

**Topic – 11: Duplicate Rows**



**Topic – 12: Correlation**



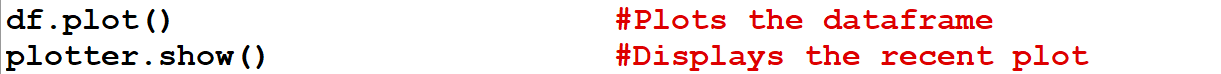


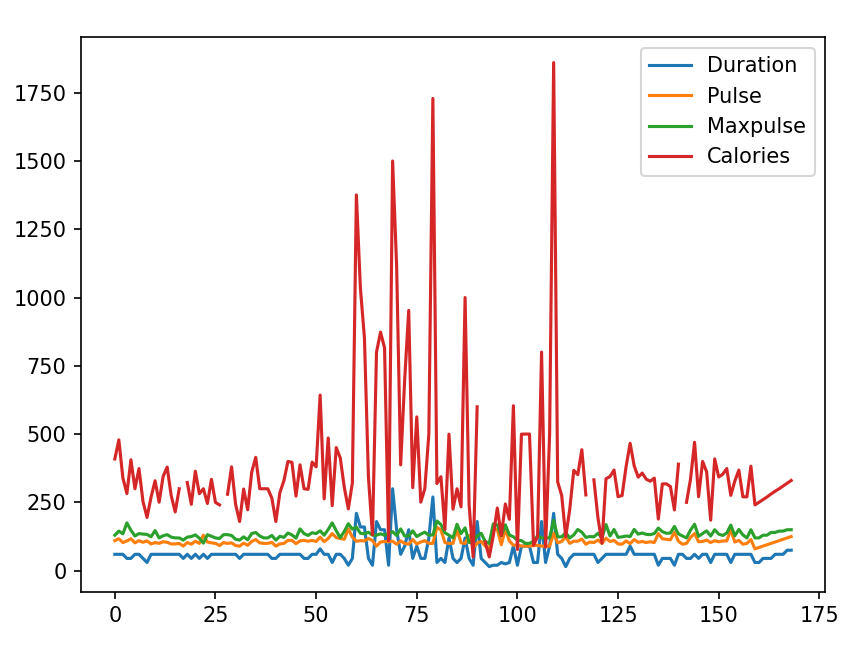
* Range is **[-1,1]** & it ignores the **non-numeric** columns.
* The more the value is away from **0**, the **perfect** the **correlation** is.
* Values **away** from **0** shows how two entities are **related**.
* Whereas values closer to **0** means that there is **no strong relation** between the figures.

**Topic – 13: Matplotlib With Pandas**



**Line Chart**



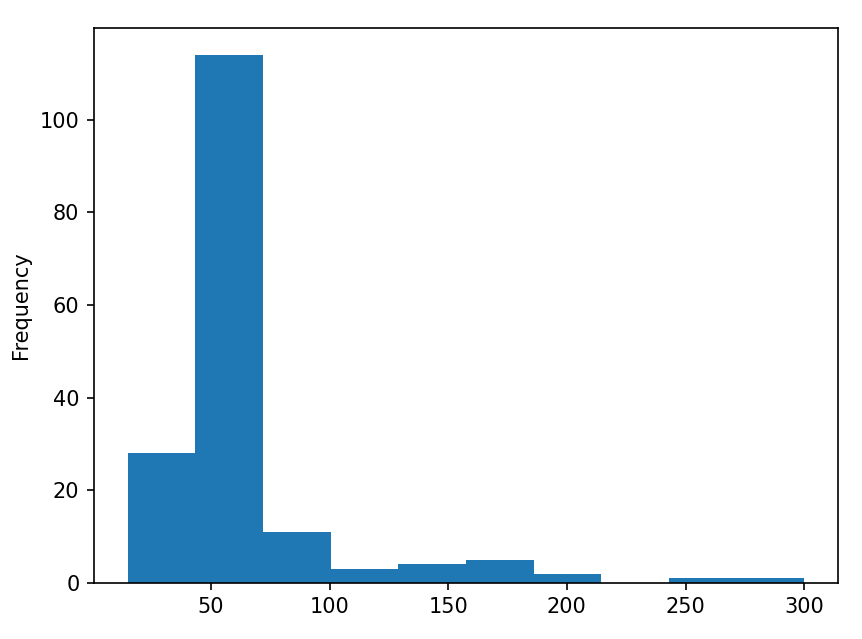


**Scattered Graph**



**Histogram**





* **X-axis** shows the **range** of **"Duration"** & **Y-axis** shows the **frequency** of **"Duration"**.
* For example, consider a data about **weight** of people.
* Now the **X-axis** will show different **ranges** (like **50-60 kg**).
* And **Y-axis** will show how many times the weights in **range** **50-60 kg** occur.

**Topic – 14: Miscellaneous**

**Row Iteration**

