

***ML Lab Assignment -2***

***Name-Vishal Dixit***

***Section- B (68)***

***University Roll no:- 201500792***

**Submitted to:-**

**Dr. Gaurav Kumar Sir**

**Link ->**

***Data Preprocessing Techniques:***

* Empty cells
* Data in wrong format
* Wrong data
* Duplicates

**Read Data:**

import pandas as pd  
  
df = pd.read\_csv('data.csv') #how to read data  
  
print(df.head(10)) #print first 10 rows of data

print(df.tail()) #print last 5 rows

print(df.head()) #print first 5 rows

print(df.(info)) #it gives information about dataset

**Remove Rows:**

import pandas as pd  
  
df = pd.read\_csv('data.csv')  
  
new\_df = df.dropna() #drop null values of dataset  
  
print(new\_df.to\_string())

**Replace Empty Values:**

import pandas as pd  
  
df = pd.read\_csv('data.csv')  
  
df.fillna(130, inplace = True) #fill null values with 130.

### **Replace Only For Specified Columns:**

import pandas as pd  
  
df = pd.read\_csv('data.csv')  
  
df["Calories"].fillna(130, inplace = True)

**Replace using Mean Median:**

import pandas as pd  
  
df = pd.read\_csv('data.csv')  
  
x = df["Calories"].mean()  
  
df["Calories"].fillna(x, inplace = True)

**Replace as Median:**

import pandas as pd  
  
df = pd.read\_csv('data.csv')  
  
x = df["Calories"].median()  
  
df["Calories"].fillna(x, inplace = True)

Cleaning Wrong Date Format:

import pandas as pd  
  
df = pd.read\_csv('data.csv')  
  
df['Date'] = pd.to\_datetime(df['Date']) #method for improve date format  
  
print(df.to\_string())

**Fixing Wrong Data:**

df.loc[7, 'Duration'] = 45 #Set "Duration" = 45 in row 7:

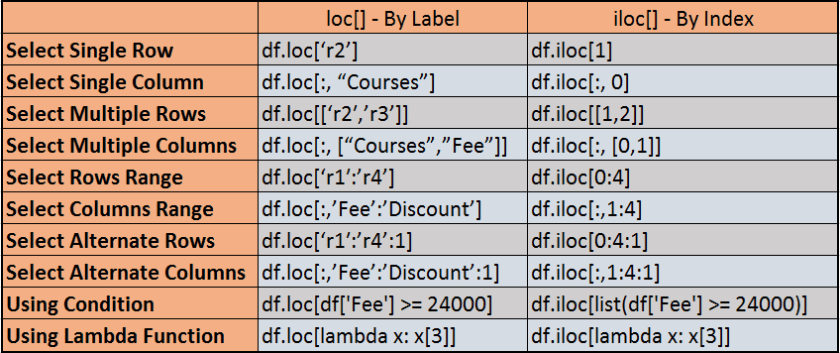
#If there is a condition:

for x in df.index:  
  if df.loc[x, "Duration"] > 120:  
    df.loc[x, "Duration"] = 120

**Delete Row when condition occurs:**

for x in df.index:  
  if df.loc[x, "Duration"] > 120:  
    df.drop(x, inplace = True)

**loc() or iloc() methods:**



**Removing Duplicates:**

data=pd.DataFrame({'k1':["one","two"]\*3+["two"],

                  "k2":[1,1,2,3,3,4,4]})

Print(data)

1)duplicated() = return true or false in Boolean as loop

0    False

1    False

2    False

3    False

4    False

5    False

6     True

dtype: bool

**2)drop\_duplicated() =** drop the duplicates

k1 k2

0 one 1

1 two 1

2 one 2

3 two 3

4 one 3

5 two 4