

Split the data from Raw data to Clean data using Pandas

```
In [1]: import pandas as pd
import numpy as np
```

```
In [2]: data = pd.read_excel(r'C:\Users\Gopi Reddy\Downloads\Address1 (1).xlsx') # Read data
```

```
Out[2]:
```

	Address	Date	Time
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec

```
In [3]: print(type(data)) # prints the which type of data
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
In [4]: data.isnull() # Shows the any null values are present in the data
```

```
Out[4]:
```

	Address	Date	Time
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False

```
In [5]: data.shape # show the dimensions of the data
```

```
Out[5]: (5, 3)
```

```
In [6]: len(data) # Length of data
```

```
Out[6]: 5
```

```
In [7]: data.columns # shows the attributes of the data
```

```
Out[7]: Index(['Address', 'Date', 'Time'], dtype='object')
```

```
In [8]: data.dtypes
```

```
Out[8]: Address    object
        Date      object
        Time      object
        dtype: object
```

```
In [9]: data.info() # it will show the total information about the data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Address     5 non-null      object
1   Date        5 non-null      object
2   Time        5 non-null      object
dtypes: object(3)
memory usage: 252.0+ bytes
```

```
In [10]: data.info
```

```
Out[10]: <bound method DataFrame.info of                                     Address
        Date      Time
0   12thblock,Hyderabad,Telengana,500085  15th jan 2025  6hr 55min 7sec
1   29throad,Secundrabad,Telengana,500014  24th feb 1992  8hr 30min 4sec
2   15thblock,hyderabad,Telengana,500048  16th jan 2025  6hr 55min 7sec
3   6thblock,srnagar,Telengana,508242    25th feb 1992  8hr 30min 4sec
4   19thblock,jntu,Telengana,546128     17th jan 2025  6hr 55min 7sec>
```

```
In [11]: data.isnull().sum()
```

```
Out[11]: Address    0
        Date      0
        Time      0
        dtype: int64
```

```
In [12]: data
```

```
Out[12]:
```

	Address	Date	Time
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec

```
In [13]: data[:1] # slicing
```

```
Out[13]:
```

	Address	Date	Time
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec

.iloc[]:

It stands for "integer location" and is used to select data based on integer position.

You use it to select rows and columns by their index positions (i.e., integer-based indexing).

Example: `df.iloc[0, 1]` will select the element in the first row and second column of the DataFrame.

```
In [14]: data.iloc[:,1] # Gives total information about the Date column
```

```
Out[14]: 0    15th jan 2025
         1    24th feb 1992
         2    16th jan 2025
         3    25th feb 1992
         4    17th jan 2025
         Name: Date, dtype: object
```

```
In [15]: data.iloc[:,2]
```

```
Out[15]: 0    6hr 55min 7sec
         1    8hr 30min 4sec
         2    6hr 55min 7sec
         3    8hr 30min 4sec
         4    6hr 55min 7sec
         Name: Time, dtype: object
```

```
In [16]: data
```

```
Out[16]:
```

	Address	Date	Time
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec

```
In [17]: data[['hours','mins','sec']]=data['Time'].str.split(' ',expand=True) # here sp
         data.head()
```

Out[17]:

	Address	Date	Time	hours	mins	sec
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec	6hr	55min	7sec
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec	8hr	30min	4sec
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec	6hr	55min	7sec
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec	8hr	30min	4sec
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec	6hr	55min	7sec

```
In [18]: data[['day','month','year']]=data['Date'].str.split([' '],expand=True) # Here sp
data.head()
```

Out[18]:

	Address	Date	Time	hours	mins	sec	day	month
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec	6hr	55min	7sec	15th	jan
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec	8hr	30min	4sec	24th	feb
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec	6hr	55min	7sec	16th	jan
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec	8hr	30min	4sec	25th	feb
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec	6hr	55min	7sec	17th	jan



```
In [19]: data[['block','city','state','pincode']]=data['Address'].str.split(r'[, ]',exp
```

```
In [20]: data.head()
```

Out[20]:

	Address	Date	Time	hours	mins	sec	day	month
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec	6hr	55min	7sec	15th	jan
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec	8hr	30min	4sec	24th	feb
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec	6hr	55min	7sec	16th	jan
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec	8hr	30min	4sec	25th	feb
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec	6hr	55min	7sec	17th	jan

In [21]: data


Out[21]:

	Address	Date	Time	hours	mins	sec	day	month
0	12thblock,Hyderabad,Telengana,500085	15th jan 2025	6hr 55min 7sec	6hr	55min	7sec	15th	jan
1	29throad,Secundrabad,Telengana,500014	24th feb 1992	8hr 30min 4sec	8hr	30min	4sec	24th	feb
2	15thblock,hyderabad,Telengana,500048	16th jan 2025	6hr 55min 7sec	6hr	55min	7sec	16th	jan
3	6thblock,srnagar,Telengana,508242	25th feb 1992	8hr 30min 4sec	8hr	30min	4sec	25th	feb
4	19thblock,jntu,Telengana,546128	17th jan 2025	6hr 55min 7sec	6hr	55min	7sec	17th	jan

In [22]: new_data=data[['block','city','state','pincode','day','month','year','hours','mi
new_data

Out[22]:

	block	city	state	pincode	day	month	year	hours	mins	sec
0	12thblock	Hyderabad	Telengana	500085	15th	jan	2025	6hr	55min	7sec
1	29throad	Secundrabad	Telengana	500014	24th	feb	1992	8hr	30min	4sec
2	15thblock	hyderabad	Telengana	500048	16th	jan	2025	6hr	55min	7sec
3	6thblock	srnagar	Telengana	508242	25th	feb	1992	8hr	30min	4sec
4	19thblock	jntu	Telengana	546128	17th	jan	2025	6hr	55min	7sec



In [23]:

```
new_data.to_csv("clean_data.csv") # convert the new data to a csv file
```