Raw to Clean Data Task

In [21]:
```
# importing libraries

import pandas as pd
```

In [22]:
```
# Reading the excel file

Task = pd.read_excel(r"C:\Users\ratho\.ipynb_checkpoints\DATA\TASK.xlsx")
Task
```

Out[22]:

	Unnamed: 0	Unnamed: 1	ADDR	DT	NAME	time
0	NaN	NaN	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21
1	NaN	NaN	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21

In [23]:
```
# Dropping the Unnamed columns

Task = Task.drop(columns = ['Unnamed: 0','Unnamed: 1'])
Task
```

Out[23]:

	ADDR	DT	NAME	time
0	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21
1	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21

In [24]:
```
Task.shape                # checking the shape of data
```

Out[24]:
```
(2, 4)
```

In [25]:
```
Task.isnull()             # check if there is any null value or not
```

Out[25]:

	ADDR	DT	NAME	time
0	False	False	False	False
1	False	False	False	False

In [26]:
```
Task.info()               # geting info
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2 entries, 0 to 1
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype         
---  ------  --------------  -----         
 0   ADDR    2 non-null      object        
 1   DT      2 non-null      datetime64[ns]
 2   NAME    2 non-null      object        
 3   time    2 non-null      object        
dtypes: datetime64[ns](1), object(3)
memory usage: 192.0+ bytes
```

In [27]:
```
# spliting ADDR column

Task[['ROAD','STREET','AREA','CITY']] = Task['ADDR'].str.split(',', expand = True)
Task


# meaning of expand here is :

#              When you use `.str.split(' ', expand=True)`,
#              pandas breaks the text wherever there is a space (' ')
#              and puts each part into a separate column in the DataFrame.
```

Out[27]:

	ADDR	DT	NAME	time	ROAD	STREET	AREA	CITY
0	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21	45 rd	kenith street	btm	bangalore 500038
1	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21	45 rd	street	hitech	Hd 500038

In [28]:
```
# Spliting City column

Task['CITY']=Task['CITY'].str.strip()
Task[['PLACE','ZIP']]= Task['CITY'].str.split(' ', expand = True)
Task


# Remove extra spaces from the CITY column using .str.strip().
# Split the CITY column into PLACE and ZIP based on spaces.
# str.strip(): Removes leading/trailing spaces from strings.
```

Out[28]:

	ADDR	DT	NAME	time	ROAD	STREET	AREA	CITY	PLACE	ZIP
0	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21	45 rd	kenith street	btm	bangalore 500038	bangalore	500038
1	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21	45 rd	street	hitech	Hd 500038	Hd	500038

In [29]:
```
# spliting Task column

Task[['HOUR','MIN','SEC']]= Task['time'].str.split(':', expand = True)
Task
```

Out[29]:

	ADDR	DT	NAME	time	ROAD	STREET	AREA	CITY	PLACE	ZIP	HOUR	MI
0	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21	45 rd	kenith street	btm	bangalore 500038	bangalore	500038	8	3
1	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21	45 rd	street	hitech	Hd 500038	Hd	500038	5	3

In [30]:
```
Task.isnull().sum()        # checking total null vlaues in any column of the data
```

Out[30]:
```
ADDR      0
DT        0
NAME      0
time      0
ROAD      0
STREET    0
AREA      0
CITY      0
PLACE     0
ZIP       0
HOUR      0
MIN       0
SEC       0
dtype: int64
```

In [31]:
```
Task['DT'] = Task['DT'].astype(str)         # typecasting datatype of Dt into string
```

In [32]:
```
Task.dtypes                                 # checking all the data types of columns a
```

Out[32]:
```
ADDR      object
DT        object
NAME      object
time      object
ROAD      object
STREET    object
AREA      object
CITY      object
PLACE     object
ZIP       object
HOUR      object
MIN       object
SEC       object
dtype: object
```

In [33]:
```
Task                    # calling Task
```

Out[33]:

	ADDR	DT	NAME	time	ROAD	STREET	AREA	CITY	PLACE	ZIP	HOUR	MI
0	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21	45 rd	kenith street	btm	bangalore 500038	bangalore	500038	8	3
1	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21	45 rd	street	hitech	Hd 500038	Hd	500038	5	3

In [34]:
```
# after typecasting dtype ; spliting the DT column

Task[['YEAR', 'MONTH', 'DAY']] = Task['DT'].astype(str).str.split('-', expand=True)
Task
```

Out[34]:

	ADDR	DT	NAME	time	ROAD	STREET	AREA	CITY	PLACE	ZIP	HOUR	MI
0	45 rd, kenith street, btm, bangalore 500038	2022-04-03	abc	8:30:21	45 rd	kenith street	btm	bangalore 500038	bangalore	500038	8	3
1	45 rd, street, hitech, Hd 500038	2022-04-02	dec	5:30:21	45 rd	street	hitech	Hd 500038	Hd	500038	5	3

In [35]:
```
# drop the unwanted columns now

Task = Task.drop(columns =['ADDR','DT','time','CITY'])
Task
```

Out[35]:

	NAME	ROAD	STREET	AREA	PLACE	ZIP	HOUR	MIN	SEC	YEAR	MONTH	DAY
0	abc	45 rd	kenith street	btm	bangalore	500038	8	30	21	2022	04	03
1	dec	45 rd	street	hitech	Hd	500038	5	30	21	2022	04	02

In [36]:
```
# saving the file

Task.to_excel('Task_DONE.xlsx',index = False)
```