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

Realtime Case Study - "Claimants" Dataset

In [2]:
```
# load the dataset
df= pd.read_csv("claimants.csv")
df.head()
```

Out[2]:

	CASENUM	CLMSEX	CLMINSUR	SEATBELT	CLMAGE	LOSS	ATTORNEY
0	5	0.0	1.0	0.0	50.0	34.940	0
1	3	1.0	0.0	0.0	18.0	0.891	1
2	66	0.0	1.0	0.0	5.0	0.330	1
3	70	0.0	1.0	1.0	31.0	0.037	0
4	96	0.0	1.0	0.0	30.0	0.038	1

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1340 entries, 0 to 1339
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   CASENUM   1340 non-null   int64  
 1   CLMSEX    1328 non-null   float64
 2   CLMINSUR  1299 non-null   float64
 3   SEATBELT  1292 non-null   float64
 4   CLMAGE    1151 non-null   float64
 5   LOSS      1340 non-null   float64
 6   ATTORNEY  1340 non-null   int64  
dtypes: float64(5), int64(2)
memory usage: 73.4 KB
```

In [4]:
```
# check for count of NA'sin each column
df.isnull().sum()
```

Out[4]:
```
CASENUM       0
CLMSEX       12
CLMINSUR     41
SEATBELT     48
CLMAGE      189
LOSS          0
ATTORNEY      0
dtype: int64
```

- There are 4 columns that have missing data

Replace 'Nan' values of CLMSEX, CLMINSUR, SEATBELT, CLMAGE

- CLMAGE --> Continous Variable --> Replace with either Mean or Median for contionus data
- CLMSEX,CLMINSUR,SEATBELT ---> discrete variable ---> Mode is used for discrete data

We have 2 ways to fill the missing values

1. fillna() using pandas

In [5]:
```
df["CLMAGE"].mean()
```

Out[5]:
```
28.414422241529106
```

In [6]:
```
df['CLMAGE'].fillna(28.414,inplace=True)   #df = df['CLMAGE'].fillna(28.414)
```

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
CASENUM      0
CLMSEX      12
CLMINSUR    41
SEATBELT    48
CLMAGE       0
LOSS         0
ATTORNEY     0
dtype: int64
```

2. SimpleImuter using sklearn

In [8]:
```
from sklearn.impute import SimpleImputer

mean_imputer = SimpleImputer(strategy='mean')
median_imputer = SimpleImputer(strategy='median')
mode_imputer = SimpleImputer(strategy='most_frequent')
```

In [9]:
```
df["CLMAGE"] = pd.DataFrame(mean_imputer.fit_transform(df[["CLMAGE"]]))
```

In [10]:
```
df["CLMAGE"].isnull().sum()
```

Out[10]:
```
0
```

Median Imputer

In [11]:
```
df = pd.read_csv("claimants.csv")
```

In [12]:
```
df.isnull().sum()
```

Out[12]:
```
CASENUM       0
CLMSEX       12
CLMINSUR     41
SEATBELT     48
CLMAGE      189
LOSS          0
ATTORNEY      0
dtype: int64
```

In [13]:
```
df["CLMAGE"].median()
```

Out[13]:
```
30.0
```

In [14]:
```
df["CLMAGE"] = pd.DataFrame(median_imputer.fit_transform(df[["CLMAGE"]]))
```

In [15]:
```
df["CLMAGE"].isnull().sum()  # all 189 records replaced by median=30
```

Out[15]:
```
0
```

Mode Imputer

In [16]:
```
# Mode Imputer
df["CLMSEX"] = pd.DataFrame(mode_imputer.fit_transform(df[["CLMSEX"]]))
df["CLMINSUR"] = pd.DataFrame(mode_imputer.fit_transform(df[["CLMINSUR"]]))
df["SEATBELT"] = pd.DataFrame(mode_imputer.fit_transform(df[["SEATBELT"]]))
```

In [17]:
```
df.isnull().sum()
```

Out[17]:
```
CASENUM     0
CLMSEX      0
CLMINSUR    0
SEATBELT    0
CLMAGE      0
LOSS        0
ATTORNEY    0
dtype: int64
```