Dealing With Missing Values

Importing Libiraries ¶

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

Load Dataset

In [2]:
```
df=pd.read_csv("claimants.csv")
df
```

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
...	...	...	...	...	...	...	...
1335	34100	0.0	1.0	0.0	NaN	0.576	1
1336	34110	1.0	1.0	0.0	46.0	3.705	0
1337	34113	1.0	1.0	0.0	39.0	0.099	1
1338	34145	1.0	0.0	0.0	8.0	3.177	0
1339	34153	1.0	1.0	0.0	30.0	0.688	1

1340 rows × 7 columns

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
```

Checking for Null valus

In [4]:
```
df.isnull().sum()
```

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


We can replace Null values normally or using sklearn libiraries

fillna

In [5]:
```
df["CLMSEX"].isnull()
```

Out[5]:
```
0       False
1       False
2       False
3       False
4       False
        ...
1335    False
1336    False
1337    False
1338    False
1339    False
Name: CLMSEX, Length: 1340, dtype: bool
```

In [6]:
```
df["CLMSEX"].isnull().sum()
```

Out[6]: 12

In [7]:
```
df[df["CLMSEX"].isnull()]
```

Out[7]:

	CASENUM	CLMSEX	CLMINSUR	SEATBELT	CLMAGE	LOSS	ATTORNEY
132	3087	NaN	NaN	NaN	NaN	3.040	0
465	901	NaN	NaN	NaN	NaN	1.069	0
491	1580	NaN	1.0	0.0	14.0	5.500	0
522	3570	NaN	1.0	0.0	NaN	7.857	0
902	13179	NaN	1.0	0.0	50.0	0.800	1
920	13658	NaN	1.0	0.0	17.0	1.039	0
964	14810	NaN	1.0	0.0	8.0	1.365	0
1009	15805	NaN	1.0	0.0	0.0	6.300	0
1166	19857	NaN	NaN	NaN	NaN	0.000	1
1175	30306	NaN	1.0	0.0	NaN	3.675	0
1223	30751	NaN	1.0	0.0	34.0	0.150	1
1246	31008	NaN	1.0	0.0	45.0	0.358	1

In [8]:
```
x=df["CLMSEX"].mean()
x
```

Out[8]: 0.5587349397590361

In [9]:
```
df["CLMSEX"].fillna(x,inplace=True)
```

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

Out[10]:
```
CASENUM       0
CLMSEX        0
CLMINSUR     41
SEATBELT     48
CLMAGE      189
LOSS          0
ATTORNEY      0
dtype: int64
```

Above we can observe there is no Null value in CLMSEX Feature

Every Time i can't Do do like this my hand is pain to write code so write simple program

In [11]:
```
df.columns
```

Out[11]:
```
Index(['CASENUM', 'CLMSEX', 'CLMINSUR', 'SEATBELT', 'CLMAGE', 'LOSS',
       'ATTORNEY'],
      dtype='object')
```

In [12]:
```
for i in df.columns:
    x=df[df[i].isnull()]
    if len(x)==0:
        pass
    else:
        meen=df[i].mean()
        df[i].fillna(meen,inplace=True)
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
CASENUM     0
CLMSEX      0
CLMINSUR    0
SEATBELT    0
CLMAGE      0
LOSS        0
ATTORNEY    0
dtype: int64
```

Congratulations We fill all null values

SimpleImputer

In [14]:
```
from sklearn.impute  import SimpleImputer
```

Load Dataset

In [15]:
```
df=pd.read_csv("claimants.csv")
df.head()
```

Out[15]:

	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

Initalizing

In [16]:
```
mean_imputer=SimpleImputer(strategy="mean")
```

In [17]:
```
mode_imputer=SimpleImputer(strategy="most_frequent")
```

In [18]:
```
median_imputer=SimpleImputer(strategy="median")
```

In [19]:
```
df.isnull().sum()
```

Out[19]:
```
CASENUM       0
CLMSEX       12
CLMINSUR     41
SEATBELT     48
CLMAGE      189
LOSS          0
ATTORNEY      0
dtype: int64
```

In [20]:
```
df["CLMSEX"]=pd.DataFrame(mean_imputer.fit_transform(df[["CLMSEX"]]))
```


In [21]:
```
df["CLMSEX"].isnull().sum()
```

Out[21]:
```
0
```

No Null Values

Again Boring let's write a program

In [22]:
```
for i in df.columns:
    x=df[df[i].isnull()]
    if len(x)==0:
        pass
    else:
        df[i]=pd.DataFrame(mean_imputer.fit_transform(df[[i]]))
```

In [23]:
```
df.isnull().sum()
```

Out[23]:
```
CASENUM     0
CLMSEX      0
CLMINSUR    0
SEATBELT    0
CLMAGE      0
LOSS        0
ATTORNEY    0
dtype: int64
```

Congratulations We fill all null values