data cleaning_missing value treatment

In [2]:
```
#Name : Vaibhav Laxman Karale
#Roll no. 58
#sub:E.T.1
#Section :3A
#Date:03/08/2024
```

In [17]:
```
 #Aim: to find out data cleaning ,missing value
```

In [18]:
```
import pandas as pd
```

In [19]:
```
import os
```

In [20]:
```
os.getcwd
```

Out[20]:
```
<function nt.getcwd()>
```

In [21]:
```
os.chdir("C:\\Users\\DELL\\OneDrive\\Desktop")
```

In [25]:
```
data=pd.read_csv("C:\\Users\\DELL\\OneDrive\\Desktop\\titanic.csv")
```

In [26]:

```
data.head(13)
```

Out[26]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	
5	897	0	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	
6	898	1	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	
7	899	0	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	
8	900	1	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	
9	901	0	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	
10	902	0	3	Ilieff, Mr. Ylio	male	NaN	0	0	349220	7.8958	
11	903	0	1	Jones, Mr. Charles Cresson	male	46.0	0	0	694	26.0000	
12	904	1	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23.0	1	0	21228	82.2667	

```
In [27]: data.tail(12)
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
406	1298	0	2	Ware, Mr. William Jeffery	male	23.0	1	0	28666	10.5
407	1299	0	1	Widener, Mr. George Dunton	male	50.0	1	1	113503	211.5
408	1300	1	3	Riordan, Miss. Johanna Hannah""	female	NaN	0	0	334915	7.7
409	1301	1	3	Peacock, Miss. Treasteall	female	3.0	1	1	SOTON/O.Q. 3101315	13.7
410	1302	1	3	Naughton, Miss. Hannah	female	NaN	0	0	365237	7.7
411	1303	1	1	Minahan, Mrs. William Edward (Lillian E Thorpe)	female	37.0	1	0	19928	90.0
412	1304	1	3	Henriksson, Miss. Jenny Lovisa	female	28.0	0	0	347086	7.7
413	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2
416	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0
417	1309	0	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3

```
In [28]: data.info
```

```
Out[28]: <bound method DataFrame.info of      PassengerId  Survived  Pclass  \
0            892         0       3
1            893         1       3
2            894         0       2
3            895         0       3
4            896         1       3
..           ...       ...     ...
413         1305         0       3
414         1306         1       1
415         1307         0       3
416         1308         0       3
417         1309         0       3

                                             Name     Sex   Age  SibSp  Parch  
\
0                                Kelly, Mr. James    male  34.5      0      0
1                Wilkes, Mrs. James (Ellen Needs)  female  47.0      1      0
2                       Myles, Mr. Thomas Francis    male  62.0      0      0
3                                Wirz, Mr. Albert    male  27.0      0      0
4    Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female  22.0      1      1
..                                            ...     ...   ...    ...    ...
413                            Spector, Mr. Woolf    male   NaN      0      0
414                  Oliva y Ocana, Dona. Fermina  female  39.0      0      0
415                  Saether, Mr. Simon Sivertsen    male  38.5      0      0
416                           Ware, Mr. Frederick    male   NaN      0      0
417                      Peter, Master. Michael J    male   NaN      1      1

                 Ticket      Fare Cabin  Embarked
0                330911    7.8292   NaN         Q
1                363272    7.0000   NaN         S
2                240276    9.6875   NaN         Q
3                315154    8.6625   NaN         S
4               3101298   12.2875   NaN         S
..                  ...       ...   ...       ...
413           A.5. 3236    8.0500   NaN         S
414            PC 17758  108.9000  C105         C
415  SOTON/O.Q. 3101262    7.2500   NaN         S
416              359309    8.0500   NaN         S
417                2668   22.3583   NaN         C

[418 rows x 12 columns]>
```

In [29]:

```
data.describe
```

Out[29]:

```
<bound method NDFrame.describe of      PassengerId  Survived  Pclass  \
0            892         0       3
1            893         1       3
2            894         0       2
3            895         0       3
4            896         1       3
..           ...       ...     ...
413         1305         0       3
414         1306         1       1
415         1307         0       3
416         1308         0       3
417         1309         0       3

                                             Name     Sex   Age  SibSp  Parch  
\
0                                Kelly, Mr. James    male  34.5      0      0
1                Wilkes, Mrs. James (Ellen Needs)  female  47.0      1      0
2                       Myles, Mr. Thomas Francis    male  62.0      0      0
3                                Wirz, Mr. Albert    male  27.0      0      0
4    Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female  22.0      1      1
..                                            ...     ...   ...    ...    ...
413                            Spector, Mr. Woolf    male   NaN      0      0
414                  Oliva y Ocana, Dona. Fermina  female  39.0      0      0
415                  Saether, Mr. Simon Sivertsen    male  38.5      0      0
416                           Ware, Mr. Frederick    male   NaN      0      0
417                      Peter, Master. Michael J    male   NaN      1      1

                 Ticket      Fare Cabin Embarked
0                330911    7.8292   NaN        Q
1                363272    7.0000   NaN        S
2                240276    9.6875   NaN        Q
3                315154    8.6625   NaN        S
4               3101298   12.2875   NaN        S
..                  ...       ...   ...      ...
413           A.5. 3236    8.0500   NaN        S
414            PC 17758  108.9000  C105        C
415  SOTON/O.Q. 3101262    7.2500   NaN        S
416              359309    8.0500   NaN        S
417                2668   22.3583   NaN        C

[418 rows x 12 columns]>
```

In [30]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Survived     418 non-null    int64  
 2   Pclass       418 non-null    int64  
 3   Name         418 non-null    object 
 4   Sex          418 non-null    object 
 5   Age          332 non-null    float64
 6   SibSp        418 non-null    int64  
 7   Parch        418 non-null    int64  
 8   Ticket       418 non-null    object 
 9   Fare         417 non-null    float64
 10  Cabin        91 non-null     object 
 11  Embarked     418 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
```

In [31]:
```
data.shape
```

Out[31]:
```
(418, 12)
```

In [32]:
```
data.size
```

Out[32]:
```
5016
```

In [33]:
```
data.ndim
```

Out[33]:
```
2
```

In [34]:
```
data.isna()
```

Out[34]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	En
0	False	False	False	False	False	False	False	False	False	False	True	
1	False	False	False	False	False	False	False	False	False	False	True	
2	False	False	False	False	False	False	False	False	False	False	True	
3	False	False	False	False	False	False	False	False	False	False	True	
4	False	False	False	False	False	False	False	False	False	False	True	
...	...	...	...	...	...	...	...	...	...	...	...	
413	False	False	False	False	False	True	False	False	False	False	True	
414	False	False	False	False	False	False	False	False	False	False	False	
415	False	False	False	False	False	False	False	False	False	False	True	
416	False	False	False	False	False	True	False	False	False	False	True	
417	False	False	False	False	False	True	False	False	False	False	True	

418 rows × 12 columns

In [35]:
```
data.isna().any()
```

Out[35]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare            True
Cabin           True
Embarked       False
dtype: bool
```

In [36]:

```
data.isna().sum()
```

Out[36]:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [38]:

```
data=data.dropna()
```

In [39]:
```
data
```

Out[39]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
12	904	1	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23.0	1	0	21228	82.266
14	906	1	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance...	female	47.0	1	0	W.E.P. 5734	61.175
24	916	1	1	Ryerson, Mrs. Arthur Larned (Emily Maria Borie)	female	48.0	1	3	PC 17608	262.375
26	918	1	1	Ostby, Miss. Helene Ragnhild	female	22.0	0	1	113509	61.979
28	920	0	1	Brady, Mr. John Bertram	male	41.0	0	0	113054	30.500
...	...	...	...	...	...	...	...	...	...	
404	1296	0	1	Frauenthal, Mr. Isaac Gerald	male	43.0	1	0	17765	27.720
405	1297	0	2	Nourney, Mr. Alfred (Baron von Drachstedt")"	male	20.0	0	0	SC/PARIS 2166	13.862
407	1299	0	1	Widener, Mr. George Dunton	male	50.0	1	1	113503	211.500
411	1303	1	1	Minahan, Mrs. William Edward (Lillian E Thorpe)	female	37.0	1	0	19928	90.000
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.900

87 rows × 12 columns

```
In [40]: data["Age"].fillna(30.272590)
```

```
Out[40]: 12     23.0
         14     47.0
         24     48.0
         26     22.0
         28     41.0
                ...
         404    43.0
         405    20.0
         407    50.0
         411    37.0
         414    39.0
         Name: Age, Length: 87, dtype: float64
```

```
In [41]: data.isna().sum()
```

```
Out[41]: PassengerId    0
         Survived       0
         Pclass         0
         Name           0
         Sex            0
         Age            0
         SibSp          0
         Parch          0
         Ticket         0
         Fare           0
         Cabin          0
         Embarked       0
         dtype: int64
```

```
In [ ]:
```