DATA CLEANING,MISSING VALUE TREATMENT

In [2]:
```
#Aim: To Perform Data Cleaning
#Exp no:6
#Name:Shrutika Vijay Ambekar
#Sec:3rd B
#Roll no:01
#Sub:ET-1
#Date:23/08/2024
```

In [4]:
```
import pandas as pd
```

In [6]:
```
import os
```

In [8]:
```
os.getcwd()
```

Out[8]:
```
'C:\\Users\\asus'
```

In [10]:
```
os.chdir("C:\\Users\\asus\\desktop")
```

In [12]:
```
df=pd.read_csv("titanic.csv")
```

In [14]:
```
df
```

Out[14]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1.0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1.0	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1.0	0.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1.0	0.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1.0	0.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1305	3.0	0.0	Zabour, Miss. Thamine	female	NaN	1.0	0.0	2665	14.4542	NaN	C	NaN	NaN	NaN
1306	3.0	0.0	Zakarian, Mr. Mapriededer	male	26.5000	0.0	0.0	2656	7.2250	NaN	C	NaN	304.0	NaN
1307	3.0	0.0	Zakarian, Mr. Ortin	male	27.0000	0.0	0.0	2670	7.2250	NaN	C	NaN	NaN	NaN
1308	3.0	0.0	Zimmerman, Mr. Leo	male	29.0000	0.0	0.0	315082	7.8750	NaN	S	NaN	NaN	NaN
1309	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

1310 rows × 14 columns

In [16]:
```
df.head(40)
```

Out[16]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1.0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1.0	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
			Allison, Miss.							C22				Montreal, PQ

2	1.0	0.0	Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C26	S	NaN	NaN	/ Chesterville, ON
3	1.0	0.0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1.0	2.0	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1.0	0.0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1.0	2.0	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
5	1.0	1.0	Anderson, Mr. Harry	male	48.0000	0.0	0.0	19952	26.5500	E12	S	3	NaN	New York, NY
6	1.0	1.0	Andrews, Miss. Kornelia Theodosia	female	63.0000	1.0	0.0	13502	77.9583	D7	S	10	NaN	Hudson, NY
7	1.0	0.0	Andrews, Mr. Thomas Jr	male	39.0000	0.0	0.0	112050	0.0000	A36	S	NaN	NaN	Belfast, NI
8	1.0	1.0	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.0000	2.0	0.0	11769	51.4792	C101	S	D	NaN	Bayside, Queens, NY
9	1.0	0.0	Artagaveytia, Mr. Ramon	male	71.0000	0.0	0.0	PC 17609	49.5042	NaN	C	NaN	22.0	Montevideo, Uruguay
10	1.0	0.0	Astor, Col. John Jacob	male	47.0000	1.0	0.0	PC 17757	227.5250	C62 C64	C	NaN	124.0	New York, NY
11	1.0	1.0	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18.0000	1.0	0.0	PC 17757	227.5250	C62 C64	C	4	NaN	New York, NY
12	1.0	1.0	Aubart, Mme. Leontine Pauline	female	24.0000	0.0	0.0	PC 17477	69.3000	B35	C	9	NaN	Paris, France
13	1.0	1.0	Barber, Miss. Ellen "Nellie"	female	26.0000	0.0	0.0	19877	78.8500	NaN	S	6	NaN	NaN
14	1.0	1.0	Barkworth, Mr. Algernon Henry Wilson	male	80.0000	0.0	0.0	27042	30.0000	A23	S	B	NaN	Hessle, Yorks
15	1.0	0.0	Baumann, Mr. John D	male	NaN	0.0	0.0	PC 17318	25.9250	NaN	S	NaN	NaN	New York, NY
16	1.0	0.0	Baxter, Mr. Quigg Edmond	male	24.0000	0.0	1.0	PC 17558	247.5208	B58 B60	C	NaN	NaN	Montreal, PQ
17	1.0	1.0	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50.0000	0.0	1.0	PC 17558	247.5208	B58 B60	C	6	NaN	Montreal, PQ
18	1.0	1.0	Bazzani, Miss. Albina	female	32.0000	0.0	0.0	11813	76.2917	D15	C	8	NaN	NaN
19	1.0	0.0	Beattie, Mr. Thomson	male	36.0000	0.0	0.0	13050	75.2417	C6	C	A	NaN	Winnipeg, MN
20	1.0	1.0	Beckwith, Mr. Richard Leonard	male	37.0000	1.0	1.0	11751	52.5542	D35	S	5	NaN	New York, NY
21	1.0	1.0	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0000	1.0	1.0	11751	52.5542	D35	S	5	NaN	New York, NY
22	1.0	1.0	Behr, Mr. Karl Howell	male	26.0000	0.0	0.0	111369	30.0000	C148	C	5	NaN	New York, NY
23	1.0	1.0	Bidois, Miss. Rosalie	female	42.0000	0.0	0.0	PC 17757	227.5250	NaN	C	4	NaN	NaN
24	1.0	1.0	Bird, Miss. Ellen	female	29.0000	0.0	0.0	PC 17483	221.7792	C97	S	8	NaN	NaN
25	1.0	0.0	Birnbaum, Mr. Jakob	male	25.0000	0.0	0.0	13905	26.0000	NaN	C	NaN	148.0	San Francisco, CA
26	1.0	1.0	Bishop, Mr. Dickinson H	male	25.0000	1.0	0.0	11967	91.0792	B49	C	7	NaN	Dowagiac, MI
			Bishop, Mrs											

27	1.0	1.0	Bishop, Mrs. Dickinson H (Helen Walton)	female	19.0000	1.0	0.0	11967	91.0792	B49	C	7	NaN	Dowagiac, MI
28	1.0	1.0	Bissette, Miss. Amelia	female	35.0000	0.0	0.0	PC 17760	135.6333	C99	S	8	NaN	NaN
29	1.0	1.0	Bjornstrom-Steffansson, Mr. Mauritz Hakan	male	28.0000	0.0	0.0	110564	26.5500	C52	S	D	NaN	Stockholm, Sweden / Washington, DC
30	1.0	0.0	Blackwell, Mr. Stephen Weart	male	45.0000	0.0	0.0	113784	35.5000	T	S	NaN	NaN	Trenton, NJ
31	1.0	1.0	Blank, Mr. Henry	male	40.0000	0.0	0.0	112277	31.0000	A31	C	7	NaN	Glen Ridge, NJ
32	1.0	1.0	Bonnell, Miss. Caroline	female	30.0000	0.0	0.0	36928	164.8667	C7	S	8	NaN	Youngstown, OH
33	1.0	1.0	Bonnell, Miss. Elizabeth	female	58.0000	0.0	0.0	113783	26.5500	C103	S	8	NaN	Birkdale, England Cleveland, Ohio
34	1.0	0.0	Borebank, Mr. John James	male	42.0000	0.0	0.0	110489	26.5500	D22	S	NaN	NaN	London / Winnipeg, MB
35	1.0	1.0	Bowen, Miss. Grace Scott	female	45.0000	0.0	0.0	PC 17608	262.3750	NaN	C	4	NaN	Cooperstown, NY
36	1.0	1.0	Bowerman, Miss. Elsie Edith	female	22.0000	0.0	1.0	113505	55.0000	E33	S	6	NaN	St Leonards-on-Sea, England Ohio
37	1.0	1.0	Bradley, Mr. George ("George Arthur Brayton")	male	NaN	0.0	0.0	111427	26.5500	NaN	S	9	NaN	Los Angeles, CA
38	1.0	0.0	Brady, Mr. John Bertram	male	41.0000	0.0	0.0	113054	30.5000	A21	S	NaN	NaN	Pomeroy, WA
39	1.0	0.0	Brandeis, Mr. Emil	male	48.0000	0.0	0.0	PC 17591	50.4958	B10	C	NaN	208.0	Omaha, NE

In [18]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1310 entries, 0 to 1309
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   pclass     1309 non-null   float64
 1   survived   1309 non-null   float64
 2   name       1309 non-null   object 
 3   sex        1309 non-null   object 
 4   age        1046 non-null   float64
 5   sibsp      1309 non-null   float64
 6   parch      1309 non-null   float64
 7   ticket     1309 non-null   object 
 8   fare       1308 non-null   float64
 9   cabin      295 non-null    object 
 10  embarked   1307 non-null   object 
 11  boat       486 non-null    object 
 12  body       121 non-null    float64
 13  home.dest  745 non-null    object 
dtypes: float64(7), object(7)
memory usage: 143.4+ KB
```

In [20]:
```
df.describe()
```

Out[20]:

	pclass	survived	age	sibsp	parch	fare	body
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	121.000000
mean	2.294882	0.381971	29.881135	0.498854	0.385027	33.295479	160.809917
std	0.837836	0.486055	14.413500	1.041658	0.865560	51.758668	97.696922
min	1.000000	0.000000	0.166700	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	21.000000	0.000000	0.000000	7.895800	72.000000
50%	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200	155.000000
75%	3.000000	1.000000	39.000000	1.000000	0.000000	31.275000	256.000000
max	3.000000	1.000000	80.000000	8.000000	9.000000	512.329200	328.000000

In [22]:
```
df.shape
```

Out[22]:
```
(1310, 14)
```

In [24]:
```
df.size
```

Out[24]:
```
18340
```

In [26]:
```
df.ndim
```

Out[26]:
```
2
```

In [28]:
```
df.isna()
```

Out[28]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	False	False	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False	True	False
2	False	False	False	False	False	False	False	False	False	False	False	True	True	False
3	False	False	False	False	False	False	False	False	False	False	False	True	False	False
4	False	False	False	False	False	False	False	False	False	False	False	True	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1305	False	False	False	False	True	False	False	False	False	True	False	True	True	True
1306	False	False	False	False	False	False	False	False	False	True	False	True	False	True
1307	False	False	False	False	False	False	False	False	False	True	False	True	True	True
1308	False	False	False	False	False	False	False	False	False	True	False	True	True	True
1309	True	True	True	True	True	True	True	True	True	True	True	True	True	True

1310 rows × 14 columns

In [30]:
```
df.isna().any()
```

Out[30]:
```
pclass       True
survived     True
name         True
sex          True
age          True
sibsp        True
parch        True
ticket       True
fare         True
cabin        True
embarked     True
boat         True
body         True
home.dest    True
dtype: bool
```

In [32]:
```
df.isna().sum()
```

```
Out[32]: pclass          1
         survived        1
         name            1
         sex             1
         age           264
         sibsp           1
         parch           1
         ticket          1
         fare            2
         cabin        1015
         embarked        3
         boat          824
         body         1189
         home.dest     565
         dtype: int64
```

```
In [34]: df["age"].fillna(29.699118)
```

```
Out[34]: 0       29.000000
         1        0.916700
         2        2.000000
         3       30.000000
         4       25.000000
                   ...
         1305    29.699118
         1306    26.500000
         1307    27.000000
         1308    29.000000
         1309    29.699118
         Name: age, Length: 1310, dtype: float64
```

```
In [36]: df.isna().sum()
```

```
Out[36]: pclass          1
         survived        1
         name            1
         sex             1
         age           264
         sibsp           1
         parch           1
         ticket          1
         fare            2
         cabin        1015
         embarked        3
         boat          824
         body         1189
         home.dest     565
         dtype: int64
```

```
In [38]: df.any()
```

```
Out[38]: pclass       True
         survived     True
         name         True
         sex          True
         age          True
         sibsp        True
         parch        True
         ticket       True
         fare         True
         cabin        True
         embarked     True
         boat         True
         body         True
         home.dest    True
         dtype: bool
```

```
In [40]: df=df.dropna()
```

```
In [42]: df.any()
```

Out[42]:

```
pclass       False
survived     False
name         False
sex          False
age          False
sibsp        False
parch        False
ticket       False
fare         False
cabin        False
embarked     False
boat         False
body         False
home.dest    False
dtype: bool
```

In [44]:

```
df.isna().sum()
```

Out[44]:

```
pclass       0
survived     0
name         0
sex          0
age          0
sibsp        0
parch        0
ticket       0
fare         0
cabin        0
embarked     0
boat         0
body         0
home.dest    0
dtype: int64
```

In [ ]:

```
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js