DATA CLEANING,MISSING VALUE TREATMENT

In [1]:
```
#Name : Ankita Gulde
#Roll no. : 44
#section : 3A
```

In [2]:
```
#Aim : TO Perform Data Processing, Data cleaning,Missing value treatment
```

In [3]:
```
import pandas as pd
```

In [5]:
```
import os
```

In [6]:
```
os.getcwd()
```

Out[6]:

In [7]:
```
os.chdir("C:\\Users\\HP\\Desktop")
```

In [9]:
```
df=pd.read_csv("titanic.csv")
```

In [10]:
```
df
```

Out[10]:

	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.			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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
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

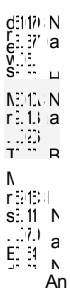
1310 rows × 14 columns

In [11]:
```
df.head(40)
```

Out[11]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1.0	1.0	Walton			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				Montreal, PQ / Chesterville, ON
2	1.0	0.0	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
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.
7	1.0	0. 0
8	1.0	1.
9	1.0	0.
10	1.0	0. 0
11	1.0	1.
12	1.0	1.
13	1.0	1.
14	1.0	1.
15	1.0	0.
16	1.0	0.
17	1.0	1.
18	1.0	1.
19	1.0	0.
20	1.0	1. 0
21	1.0	1.
22	1.0	1.
23	1.0	1.
24	1.0	1. 0
25	1.0	0.
26	1.0	1.
27	1.0	1.
28	1.0	1.
29	1.0	1. 0
30	1.0	0.
31	1.0	1.
32	1.0	1.
33	1.0	1.0



								Korneli							
								Andrews,							Jr
								Appleton,							
								Da							
	male	71.0000	0.0	0.0						PC 17609	49.5042	NaN	C	NaN	22.0	Montevideo, Uruguay
	male	47.0000	1.0	0.0											
	female	18.0000	1.0	0.0											
								Artagaveytia, Mr. Ramon							
								Astor, Col. John Jacob		PC 17757	227.5250	C62 C64	C	NaN	124.0	New York, NY
								Astor, Mrs. John Jacob		PC 17757	227.5250	C62 C64	C	4	NaN	New York, NY
								(Madeleine Talmadge Force)								
								Aubart, Mme. Leontine Pauline								
								Barber, Miss. Ellen "Nellie"								
								Barkworth, Mr. Algernon Henry Wilson								
								Baumann, Mr. John D								
								Baxter, Mr. Quigg Edmond								
								Baxter, Mrs. James (Helene DeLaudeniere	female	50.0000	0.0	1.0	PC 17558	247.5208	B58 B60	C	6	NaN	Montreal, PQ
	Bazzani, Miss. Albina	female	32.0000	0.0	0.0	11813	76.2917	D15	C	8	NaN		NaN				
	Beattie, Mr. Thomson	male	36.0000	0.0	0.0	13050	75.2417	C6	C	A	NaN		Winnipeg, MN				
	Beckwith, Mr. Richard Leonard	male	37.0000	1.0	1.0	11751	52.5542	D35	S	5	NaN		New York, NY				
	Beckwith, Mrs. Richard Leonard (Sallie		47.0000	1.0	1.0	11751	52.5542	D35	S	5	NaN		New York, NY				
	Behr, Mr. Karl Howell	male	26.0000	0.0	0.0	111369	30.0000	C148	C	5	NaN		New York, NY				
								Chaput)								

Bidois, Miss. Rosalie	female	42.0000	0.0	0.0	PC 17757	227.5250	NaN	C	4	NaN	NaN	NY
Bird, Miss. Ellen female		29.000 0	0.0	0.0	PC 1748 3	221.77 92	C97	S	8	NaN	NaN	
									Na N	148. 0	San Francis co, CA	
p, Mr. Dickinson male		25.000 0	1.0	0.0	1196 7	91.0792	B49	C	7	NaN	Dowagiac, MI	
, Mrs. Dickinson female		19.000 0	1.0	0.0	1196 7	91.0792	B49	C	7	NaN	Dowagiac, MI	
Birnbaum, Mr. Jakob	male	25.0000	0.0	0.0	13905	26.0000	NaN	C				

Bisho H

Bishop H (Helen Walton)

Bissette, Miss. Amelia female	35. 00 00	0.0	0.0	PC 17 76 0	13 5.6 33 3	C9 9	S	8	aN	NaN
Bjorn strom-Steffan sson, Mr. male Maurit	28. 00 00	0.0	0.0	11 05 64	26. 55 00	C5 2	S	D	NaN	Stock holm, Swed en / Wash
Weart								Na N	NaN	Trenton, NJ
Blank, Mr. male	40. 00 00	0.0	0.0	11 22 77	31. 00 00	A3 1	C	7	NaN	Glen Ridge , NJ
onnell, Miss. Caroline female	30. 00 00	0.0	0.0	36 92 8	16 4.8 66	C7	S	8	NaN	Youngsto

Blackwell, Mr. Stephen	male	45.0000	0.0	0.0	113784	35.5000	T	S

B

OH

Birkdale,

34	1.0	0.0
35	1.0	1.0
36	1.0	1.0
37 1.0	1.0	

Bonnell, Miss. Elizabeth	female	58.0000	0.0	0.0	113783	26.5500	C103	S	8	NaN	England Cleveland, Ohio
James male		42.000 0	0.0	0.0	11048 9	26.550 0	D22	S	NaN	NaN	London / Winnipeg, MB

Borebank, Mr. John										
Bowen, Miss. Grace Scott	female	45.0000	0.0	0.0	PC 17608	262.3750	NaN	C	4	NaN Cooperstown, NY
werman, Miss. Elsie Edith	female	22.0000	0.0	1.0	113505	55.0000	E33	S	6	NaN St Leonards-on-Sea, England Ohio
Bradley, Mr. George ("George Arthur Brayton")	male	NaN	0.0	0.0	111427	26.5500	NaN	S	9	NaN Los Angeles, CA
Bo										

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 [12]:
```
df.tail(10)
```

Out[12]:

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1300	3.0	1.0	Yasbeck, Mrs. Antoni (Selini Alexander)	female	15.0	1.0	0.0	2659	14.4542	NaN	C	NaN	NaN	NaN
1301	3.0	0.0	Youseff, Mr. Gerious	male	45.5	0.0	0.0	2628	7.2250	NaN	C	NaN	312.0	NaN
1302	3.0	0.0	Yousif, Mr. Wazli	male	NaN	0.0	0.0	2647	7.2250	NaN	C	NaN	NaN	NaN
1303	3.0	0.0	Yousseff, Mr. Gerious	male	NaN	0.0	0.0	2627	14.4583	NaN	C	NaN	NaN	NaN
1304	3.0	0.0	Zabour, Miss. Hileni	female	14.5	1.0	0.0	2665	14.4542	NaN	C	NaN	328.0	NaN
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	26.5	0.0	0.0	2656	7.2250	NaN	C	NaN	304.0	NaN
1307	3.0	0.0	Zakarian, Mr. Ortin	male	27.0	0.0	0.0	2670	7.2250	NaN	C	NaN	NaN	NaN
1308	3.0	0.0	Zimmerman, Mr. Leo	male	29.0	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

In [13]:
```
df.info()
```

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

In [14]:
```
df.describe()
```

	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 [18]:
```
df.isna()
```

Out[18]:
```
2
```

Out[18]:

	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 [19]:
```
df.isna().any()
```

Out[19]:

```
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
```

In [20]:

Out[20]:

```
pclass          1
survived        1
name            1
sex             1
age           264
sibsp           1
parch           1
ticket          1
fare            2
cabin           1
embarked        3
boat          824
body          118
              9
home.dest     565
dtype:
int64
```

In [21]:
```
df["age"].fillna(29.699118)
```

Out[21]:

```
0       29.000000
Name: age, Length: 1310, dtype: float64
```

```
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
```

In [22]:
```
df.isna().sum()
```

Out[22]:
```
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 [23]:
```
df.any()
```

Out[23]:
```
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
```

In [24]:
```
df=df.dropna()
```

In [25]:
```
df.any()
```

Out[25]:
```
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
```

In [26]:
```
df.isna().sum()
```

Out[26]:
```
pclass      0.0
survived    0.0
name        0.0
sex         0.0
age         0.0
sibsp       0.0
dtype: float64
parch       0.0
ticket      0.0
fare        0.0
cabin       0.0
embarked    0.0
boat        0.0
body        0.0
home.dest   0.0
```

In [ ]:
```

```