DATA CLEANING,MISSING VALUE TREATMENT

In [2]:
```
#Name : Ashwini V Kayande
#Roll No : 60
#Section : 3A
#Date :03/08/2024
```

In [4]:
```
#Aim : TO Perform Data Processing, Data cleaning,Missing value treatment
```

In [6]:
```
import pandas as pd
```

In [8]:
```
import os
```

In [10]:
```
os.getcwd()
```

Out[10]:
```
'C:\\Users\\user'
```

In [14]:
```
os.chdir("C:\\Users\\user\\Desktop")
```

In [22]:
```
df=pd.read_csv("train.csv")
```

In [24]:
```
df
```

Out[24]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	T
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STC 31
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	3
...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	2
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	1
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	1
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	3

891 rows × 12 columns

In [26]:
```
df.head(40)
```

Out[26]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	P
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	
5	6	0	3	Moran, Mr. James	male	NaN	0	0	
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	
12	13	0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	
14	15	0	3	Vestrom, Miss. Hulda	female	14.0	0	0	

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	
				Amanda Adolfina					
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	
16	17	0	3	Rice, Master. Eugene	male	2.0	4	1	
17	18	1	2	Williams, Mr. Charles Eugene	male	NaN	0	0	
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1	0	
19	20	1	3	Masselmani, Mrs. Fatima	female	NaN	0	0	
20	21	0	2	Fynney, Mr. Joseph J	male	35.0	0	0	
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	
22	23	1	3	McGowan, Miss. Anna "Annie"	female	15.0	0	0	
23	24	1	1	Sloper, Mr. William Thompson	male	28.0	0	0	
24	25	0	3	Palsson, Miss. Torborg Danira	female	8.0	3	1	
25	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia...	female	38.0	1	5	
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NaN	0	0	
29	30	0	3	Todoroff, Mr. Lalio	male	NaN	0	0	
30	31	0	1	Uruchurtu, Don. Manuel	male	40.0	0	0	P

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	
				E					
31	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NaN	1	0	P
32	33	1	3	Glynn, Miss. Mary Agatha	female	NaN	0	0	
33	34	0	2	Wheadon, Mr. Edward H	male	66.0	0	0	
34	35	0	1	Meyer, Mr. Edgar Joseph	male	28.0	1	0	P
35	36	0	1	Holverson, Mr. Alexander Oskar	male	42.0	1	0	
36	37	1	3	Mamee, Mr. Hanna	male	NaN	0	0	
37	38	0	3	Cann, Mr. Ernest Charles	male	21.0	0	0	
38	39	0	3	Vander Planke, Miss. Augusta Maria	female	18.0	2	0	
39	40	1	3	Nicola-Yarred, Miss. Jamila	female	14.0	1	0	

In [28]:
```
df.tail(10)
```

Out[28]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	
881	882	0	3	Markun, Mr. Johann	male	33.0	0	0	
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A.
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SO
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [32]:
```
df.describe()
```

Out[32]:

	PassengerId	Survived	Pclass	Age	SibSp	Parc
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.00000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.38159
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.80605
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.00000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.00000
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.00000
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.00000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.00000

In [34]:
```
df.shape
```

Out[34]: (891, 12)

In [36]:
```
df.size
```

Out[36]: 10692

In [38]:
```
df.ndim
```

Out[38]: 2

In [40]:
```
df.isna()
```

Out[40]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False
887	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False
889	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False

891 rows × 12 columns

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

Out[42]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

In [44]:
```
df.isna().sum()
```

Out[44]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [54]:
```
df.columns = df.columns.str.strip()
```

In [59]:
```
df.isna().sum()
```

Out[59]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [61]:
```
df.any()
```

Out[61]:
```
PassengerId    True
Survived       True
Pclass         True
Name           True
Sex            True
Age            True
SibSp          True
Parch          True
Ticket         True
Fare           True
Cabin          True
Embarked       True
dtype: bool
```

In [63]:
```
df=df.dropna()
```

In [65]:
```
df.any()
```

Out[65]:
```
PassengerId    True
Survived       True
Pclass         True
Name           True
Sex            True
Age            True
SibSp          True
Parch          True
Ticket         True
Fare           True
Cabin          True
Embarked       True
dtype: bool
```

In [67]:
```
df.isna().sum()
```

Out[67]:
```
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 [ ]: