In [1]:
```
import numpy as np
import pandas as pd
import scipy as sp
import matplotlib
import sklearn as skl
```

In [2]:
```
df=pd.read_csv('Titanic.csv')
```

In [3]:
```
df_NOnaAll=df.dropna(how='all')
```

In [4]:
```
df_NOnaAny=df.dropna(how='any')
```

In [5]:
```
df_thresh=df.dropna(thresh=2)
```

In [6]:
```
df_fill_mean=df.fillna(df.mean())
```

In [7]:
```
df.head()
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050

In [8]:
```
df_AgeSet=df['Age'].fillna(df['Age'].median())
df_SibSpSet=df['Age'].fillna(df['Age'].median())
df_ParchSet=df['Age'].fillna(df['Age'].median())
```

In [9]:
```
df.describe()
```

Out[9]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.2042
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.6934
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91040
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.4542
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.0000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329

In [10]:
```
Categorical = df.dtypes[df.dtypes == "object"].index
print(Categorical)
df[Categorical].describe()
```

```
Index(['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'], dtype='object')
```

Out[10]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Carter, Mr. William Ernest	male	CA. 2343	B96 B98	S
freq	1	577	7	4	644

In [11]:
```
df.select_dtypes(include=['O']).columns.values
```

Out[11]:
```
array(['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'], dtype=object)
```

Many machine learning models, such as regression or Support Learning Models (SVMs), are based on algebraic expressions. Therefore, it is necessary to have numerical inputs to perform the analysis.

In [14]:
```
from sklearn import preprocessing as pp
le=pp.LabelEncoder()
le.fit(df[Categorical[0]])
Cat0=pd.DataFrame({Categorical[0]:le.transform(df[Categorical[0]])})
```

In [15]:
```
df[Categorical[0]]=Cat0[Categorical[0]].values
```

In [16]:
```
le.fit(df[Categorical[1]])
Cat1=pd.DataFrame({Categorical[1]:le.transform(df[Categorical[1]])})
df[Categorical[1]]=Cat1[Categorical[1]].values
```

In [17]:
```
df.head()
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cal
0	1	0	3	108	1	22.0	1	0	A/5 21171	7.2500	NaN
1	2	1	1	190	0	38.0	1	0	PC 17599	71.2833	C85
2	3	1	3	353	0	26.0	0	0	STON/O2. 3101282	7.9250	NaN
3	4	1	1	272	0	35.0	1	0	113803	53.1000	C12
4	5	0	3	15	1	35.0	0	0	373450	8.0500	NaN