In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
pd. pandas.set_option('display.max_columns', None)
```

In [8]:

```
dataset=pd.read_csv('Titanic_Dataset.csv')
print(dataset.shape)
dataset.head()
```

```
(891, 12)
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [10]:

```
## Let us capture all the nan values
## First lets handle Categorical features which are missing
features_nan=[feature for feature in dataset.columns if dataset[feature].isnull().sum()>1 and dataset[featur

for feature in features_nan:
    print("{}: {}% missing values".format(feature,np.round(dataset[feature].isnull().mean(),4)))
```

```
Cabin: 0.771% missing values
Embarked: 0.0022% missing values
```

In [11]:

```
## Replace missing value with a new label
def replace_cat_feature(dataset,features_nan):
    data=dataset.copy()
    data[features_nan]=data[features_nan].fillna('Missing')
    return data

dataset=replace_cat_feature(dataset,features_nan)

dataset[features_nan].isnull().sum()
```

Out[11]:

```
Cabin       0
Embarked    0
dtype: int64
```

In [ ]:

```
dataset.head()
```

In [12]:

```
numerical variables the contains missing values
ature for feature in dataset.columns if dataset[feature].isnull().sum()>1 and dataset[feature].dtypes!='O']

umerical nan variables and percentage of missing values

al_with_nan:
sing value".format(feature,np.around(dataset[feature].isnull().mean(),4)))
```

```
Age: 0.1987% missing value
```

In [ ]:

In [13]:

```
## Replacing the numerical Missing Values

for feature in numerical_with_nan:
    ## We will replace by using median since there are outliers
    median_value=dataset[feature].median()

    ## create a new feature to capture nan values
    dataset[feature+'nan']=np.where(dataset[feature].isnull(),1,0)
    dataset[feature].fillna(median_value,inplace=True)

dataset[numerical_with_nan].isnull().sum()
```

Out[13]:

```
Age    0
dtype: int64
```

In [14]:

```
dataset.head(50)
```

17	18	1	2	Williams, Mr. Charles Eugene	male	28.0	0	0	244373	13.0000	Missing	S	1
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1	0	345763	18.0000	Missing	S	0
19	20	1	3	Masselmani, Mrs. Fatima	female	28.0	0	0	2649	7.2250	Missing	C	1
20	21	0	2	Fynney, Mr. Joseph J	male	35.0	0	0	239865	26.0000	Missing	S	0
21	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.0000	D56	S	0

In [ ]: