In [7]:
```
import pandas as pd
```

In [8]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

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

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

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
413	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	NaN	S
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C105	C
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	NaN	S
416	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	NaN	S
417	1309	0	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	NaN	C

418 rows × 12 columns

In [11]:
```
df.isnull()
```

Out[11]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	True	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	True	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
413	False	False	False	False	False	True	False	False	False	False	True	False
414	False	False	False	False	False	False	False	False	False	False	False	False
415	False	False	False	False	False	False	False	False	False	False	True	False
416	False	False	False	False	False	True	False	False	False	False	True	False
417	False	False	False	False	False	True	False	False	False	False	True	False

418 rows × 12 columns

In [12]:
```
df.shape
```

Out[12]:
```
(418, 12)
```

Scanning all variables for missing values and inconsistencies

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

In [14]:
```
df.isnull().sum().sum()
```

Out[14]:
```
414
```

In [15]:
```
df4=df.drop('Cabin',axis=1)
df4
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.0500	S
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	S
416	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.0500	S
417	1309	0	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.3583	C

418 rows × 11 columns

In [16]:
```
df4.isnull().sum()
```

Out[16]:
```
PassengerId     0
Survived        0
Pclass          0
Name            0
Sex             0
Age            86
SibSp           0
Parch           0
Ticket          0
Fare            1
Embarked        0
dtype: int64
```

In [17]:
```
df4=df4.fillna(method='pad')
df4
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	0	3	Spector, Mr. Woolf	male	28.0	0	0	A.5. 3236	8.0500	S
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	S
416	1308	0	3	Ware, Mr. Frederick	male	38.5	0	0	359309	8.0500	S
417	1309	0	3	Peter, Master. Michael J	male	38.5	1	1	2668	22.3583	C

418 rows × 11 columns

In [18]:
```
df4.isnull().sum()
```

Out[18]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

In [19]:
```
df4
```

Out[19]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	S
...	...	...	...	...	...	...	...	...	...	...	...
413	1305	0	3	Spector, Mr. Woolf	male	28.0	0	0	A.5. 3236	8.0500	S
414	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.9000	C
415	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.2500	S
416	1308	0	3	Ware, Mr. Frederick	male	38.5	0	0	359309	8.0500	S
417	1309	0	3	Peter, Master. Michael J	male	38.5	1	1	2668	22.3583	C

418 rows × 11 columns

Using IQR to deal with outliers

In [20]:
```
import sklearn
from sklearn.datasets import load_boston
```

In [21]:
```
import seaborn as sns
```

In [22]:
```
df4.shape
```

Out[22]:
```
(418, 11)
```

In [23]:
```
Q1 = np.percentile(df4['Fare'],25,method = 'midpoint')
Q3 = np.percentile(df4['Fare'],75,method = 'midpoint')
IQR = Q3 - Q1
```

In [24]:
```
df4.shape
```

Out[24]: (418, 11)

In [25]:
```
upper = np.where(df4['Fare'] >= (Q3+1.5*IQR))

lower = np.where(df4['Fare'] <= (Q1-1.5*IQR))
```

In [26]:
```
df4.drop(upper[0], inplace = True)
df4.drop(lower[0], inplace = True)
df4.shape
```

Out[26]: (363, 11)

In [28]:
```
df5=df4
```

In [62]:
```
df5=df5.drop('Name',axis=1)
df5=df5.drop('Sex',axis=1)
df5=df5.drop('Embarked',axis=1)
df5
```

Out[62]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Ticket	Fare
0	892	0	3	34.5	0	0	330911	7.8292
1	893	1	3	47.0	1	0	363272	7.0000
2	894	0	2	62.0	0	0	240276	9.6875
3	895	0	3	27.0	0	0	315154	8.6625
4	896	1	3	22.0	1	1	3101298	12.2875
...	...	...	...	...	...	...	...	...
412	1304	1	3	28.0	0	0	347086	7.7750
413	1305	0	3	28.0	0	0	A.5. 3236	8.0500
415	1307	0	3	38.5	0	0	SOTON/O.Q. 3101262	7.2500
416	1308	0	3	38.5	0	0	359309	8.0500
417	1309	0	3	38.5	1	1	2668	22.3583

363 rows × 8 columns

Data transformations - Changing the scale of for better understanding of the variable

In [63]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df5.drop('Ticket',axis=1),
                                                    df5['Ticket'],
                                                    test_size = 0.3,
                                                    random_state = 0)

X_train.shape, X_test.shape
```

Out[63]:
```
((254, 7), (109, 7))
```

In [64]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [65]:
```
scaler=MinMaxScaler()
```

In [68]:
```
X_train_scaled=scaler.fit_transform(X_train)
X_test_scaled=scaler.fit_transform(X_test)

print(X_train_scaled)
```

```
[[0.27577938 0.         1.         ... 0.2        0.         0.22237231]
 [0.53717026 1.         0.         ... 0.         0.         0.42224308]
 [0.5323741  1.         0.5        ... 0.         0.         0.16153846]
 ...
 [0.3117506  0.         1.         ... 0.         0.         0.12384615]
 [0.11990408 0.         0.         ... 0.2        0.         0.92307692]
 [0.46282974 0.         0.5        ... 0.         0.         0.19      ]]
```

In [69]:
```
print(X_test_scaled)
```

```
[[0.84558824 1.         0.         0.71047129 0.         0.16666667
  0.97474308]
 [0.83823529 1.         1.         0.59787679 0.         0.
  0.11121846]
 [0.10294118 0.         1.         0.71047129 0.         0.
  0.11923077]
 [0.05637255 0.         0.         0.64613157 0.         0.
  0.46923077]
 [0.59558824 1.         0.5        0.45311243 0.25       0.
  0.4       ]
 [0.95343137 0.         1.         0.45311243 0.75       0.16666667
  0.33884615]
 [0.23284314 1.         0.         0.67830143 0.25       0.
  0.85294923]
 [0.32843137 0.         1.         0.63004665 0.25       1.
  0.72153846]
 [0.29901961 1.         1.         0.34051793 0.5        0.
  0.35769231]
 [0.00245098 1.         1.         0.46919736 0.         0.
```

In [ ]: