Importing Libraries And Dataset

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import LabelEncoder, OneHotEncoder, OrdinalEncoder, StandardScaler
from sklearn.impute import SimpleImputer

from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
```

```
df = pd.read_csv('/content/Titanic-Dataset.csv')
df.head()
```

	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

Next steps: Generate code with df New interactive sheet

Dataset Information

```
df.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
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
```

Replacing Null Values

```
print(df.isnull().values.any())
```

```
True
```

```
df.isnull().sum()
```

	0
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

```
df['Age'] = df['Age'].fillna(df['Age'].mean())
df['Cabin'] = df['Cabin'].fillna(df['Cabin'].mode()[0])
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

```
print(df['Age'].isnull().sum())
print(df['Cabin'].isnull().sum())
print(df['Embarked'].isnull().sum())
df.isnull().sum()
```

```
0
0
0
```

	0
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

```
df.duplicated().sum()
df.drop_duplicates(inplace=True)
```

Encoding

```
le = LabelEncoder()
df['Sex'] = le.fit_transform(df['Sex'])
df['Embarked'] = le.fit_transform(df['Embarked'])
df.head()
```

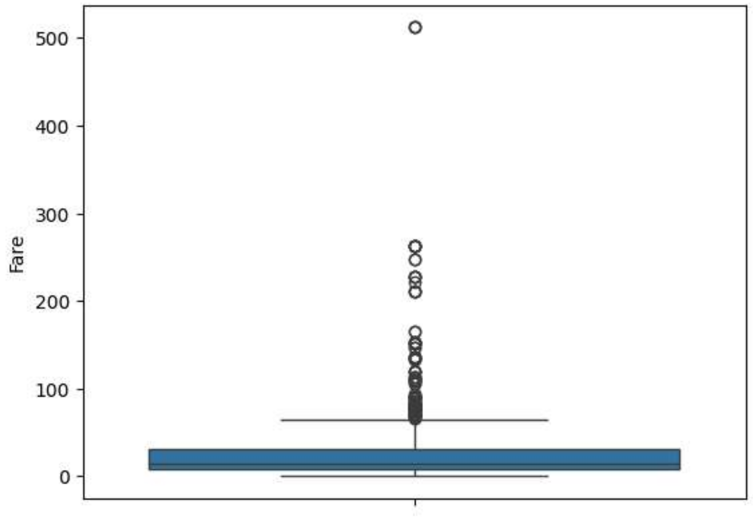
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.2500	B96 B98	2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.2833	C85	0
2	3	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.9250	B96 B98	2
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.1000	C123	2
4	5	0	3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.0500	B96 B98	2

Next steps: Generate code with df New interactive sheet

Removing Outliers

```
sns.boxplot(df['Fare'])
```

```
<Axes: ylabel='Fare'>
```



```
Q1 = df['Fare'].quantile(0.25)
Q3 = df['Fare'].quantile(0.75)
IQR = Q3 - Q1

lower = Q1 - 1.0 * IQR
upper = Q3 + 1.0 * IQR
dff = df[(df['Fare'] >= lower) & (df['Fare'] <= upper)]
```

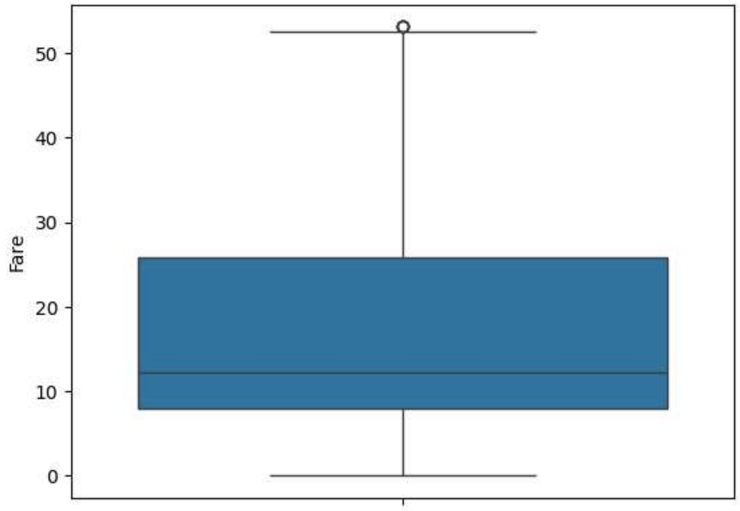
```
dff.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	1	22.000000	1	0	A/5 21171	7.2500	B96 B98	2
2	3	1	3	Heikkinen, Miss. Laina	0	26.000000	0	0	STON/O2. 3101282	7.9250	B96 B98	2
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.000000	1	0	113803	53.1000	C123	2
4	5	0	3	Allen, Mr. William Henry	1	35.000000	0	0	373450	8.0500	B96 B98	2
5	6	0	3	Moran, Mr. James	1	29.699118	0	0	330877	8.4583	B96 B98	1

Next steps: Generate code with dff New interactive sheet

```
sns.boxplot(dff['Fare'])
```

```
<Axes: ylabel='Fare'>
```



Normalization & Standardization

```
scaler = MinMaxScaler()
dff = dff.copy()
dff[['Age', 'Fare']] = scaler.fit_transform(dff[['Age', 'Fare']])
dff.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	1	0.271174	1	0	A/5 21171	0.136535	B96 B98	2
2	3	1	3	Heikkinen, Miss. Laina	0	0.321438	0	0	STON/O2. 3101282	0.149247	B96 B98	2
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	0.434531	1	0	113803	1.000000	C123	2
4	5	0	3	Allen, Mr. William Henry	1	0.434531	0	0	373450	0.151601	B96 B98	2
5	6	0	3	Moran, Mr. James	1	0.367921	0	0	330877	0.159290	B96 B98	1

Next steps: Generate code with dff New interactive sheet

```
scaler = StandardScaler()
dff[['Age', 'Fare']] = scaler.fit_transform(dff[['Age', 'Fare']])
dff.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	1	-0.542633	1	0	A/5 21171	-0.797316	B96 B98	2
2	3	1	3	Heikkinen, Miss. Laina	0	-0.228652	0	0	STON/O2. 3101282	-0.738993	B96 B98	2
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	0.477804	1	0	113803	3.164321	C123	2
4	5	0	3	Allen, Mr. William Henry	1	0.477804	0	0	373450	-0.728192	B96 B98	2
5	6	0	3	Moran, Mr. James	1	0.061711	0	0	330877	-0.692913	B96 B98	1

Next steps: Generate code with dff New interactive sheet