In [ ]:
```
project title
 Learn how to clean and prepare raw data for ML.
```

objective 1.Import the dataset and explore basic info (nu ls, data types). 2.Handle missing values using mean/median/imputation. 3.Convert categorical features into numerical using encoding. 4.Normalize/standardize the numerical features. 5.Visualize outliers using boxplots and remove them

loading dependencies...

In [8]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

loading the dataset...

In [23]:
```
data = pd.read_csv('train.csv')
```

In [25]:
```
data
```

Out[25]:

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [27]:
```
data.head()
```

Out[27]:

	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

statistical innformation..

In [29]:
```
data.describe()
```

Out[29]:

	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

datatypes information..

In [31]:
```
data.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
```

dealing with missing values .....

In [39]:
```
missing_values = data.isnull().sum()
missing_summary = missing_values[missing_values > 0].sort_values(ascending=False)
data_shape = data.shape
```

In [41]:
```
data_shape, missing_summary
```

Out[41]:
```
((891, 12),
 Cabin       687
 Age         177
 Embarked      2
 dtype: int64)
```

In [55]:
```
data['Age'].mean()
```

Out[55]:
```
29.36158249158249
```

In [ ]:
```
data['Age'].fillna(data['Age'].mean(), inplace=True)
```

In [69]:
```
data['Embarked'].mode()
```

Out[69]:
```
0    S
Name: Embarked, dtype: object
```

In [ ]:
```
data['Embarked'].fillna(data['Embarked'].mode()[0], inplace=True)
```

Converting categorical features into numerical using encoding.....

In [77]:
```
categorical_cols = data.select_dtypes(include=['object']).columns.tolist()
```

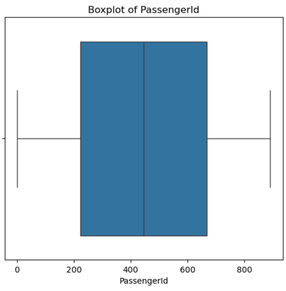
In [79]:
```
data_encoded = pd.get_dummies(data, columns=categorical_cols, drop_first=True)

data_encoded.shape, data_encoded.head()
```

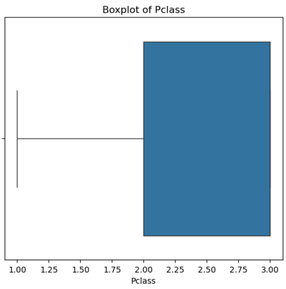
Out[79]:
```
((891, 1579),
    PassengerId  Survived  Pclass  SibSp  Parch     Fare  \
 0            1         0       3      1      0   7.2500   
 1            2         1       1      1      0  71.2833   
 2            3         1       3      0      0   7.9250   
 3            4         1       1      1      0  53.1000   
 4            5         0       3      0      0   8.0500   

    Name_Abbott, Mr. Rossmore Edward  Name_Abbott, Mrs. Stanton (Rosa Hunt)  \
 0                             False                                  False   
 1                             False                                  False   
 2                             False                                  False   
 3                             False                                  False   
 4                             False                                  False   

    Name_Abelson, Mr. Samuel  Name_Abelson, Mrs. Samuel (Hannah Wizosky)  ...  \
 0                     False                                      False  ...   
 1                     False                                      False  ...   
 2                     False                                      False  ...   
 3                     False                                      False  ...   
 4                     False                                      False  ...   

    Ticket_W./C. 14258  Ticket_W./C. 14263  Ticket_W./C. 6607  \
 0               False               False              False   
 1               False               False              False   
 2               False               False              False   
 3               False               False              False   
 4               False               False              False   

    Ticket_W./C. 6608  Ticket_W./C. 6609  Ticket_W.E.P. 5734  Ticket_W/C 14208  \
 0              False              False               False             False   
 1              False              False               False             False   
 2              False              False               False             False   
 3              False              False               False             False   
 4              False              False               False             False   

    Ticket_WE/P 5735  Embarked_Q  Embarked_S  
 0             False       False        True  
 1             False       False       False  
 2             False       False        True  
 3             False       False        True  
 4             False       False        True  

 [5 rows x 1579 columns])
```

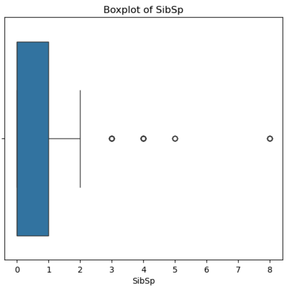
In [81]:
```
numerical_cols = data_encoded.select_dtypes(include=['float64', 'int64']).columns.tolist()
numerical_cols.remove('Survived')
```

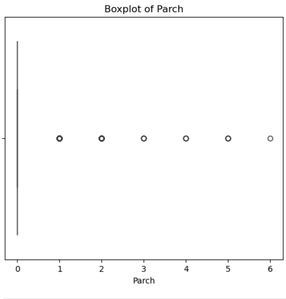
visualize outliers using boxplots and remove them.....

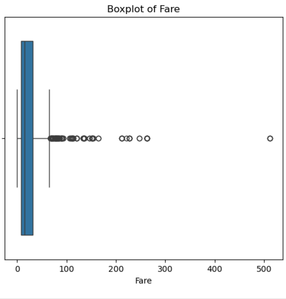
In [85]:
```
plt.figure(figsize=(15, 10))
for i, col in enumerate(numerical_cols[:6]):
    plt.subplot(2, 3, i + 1)
    sns.boxplot(x=data_encoded[col])
    plt.title(f'Boxplot of {col}')
plt.tight_layout()
plt.show()
```


Boxplot of PassengerId
0
200
400
600
800
PassengerId


Boxplot of Pclass
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
Pclass


Boxplot of SibSp
0
1
2
3
4
5
6
7
8
SibSp


Boxplot of Parch
0
1
2
3
4
5
6
Parch


Boxplot of Fare
0
100
200
300
400
500
Fare

In [ ]: