In [1]:

```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
print(os.getcwd())
```

```
C:\Users\Acer\DS application Lab
```

In [2]:

```
df = pd.read_csv("C:/Users/Acer/Downloads/Titanic-Dataset.csv")
```

In [3]:

```
df.head()
```

Out[3]:

	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 [4]:

```
df.duplicated()
```

Out[4]:

```
0      False
1      False
2      False
3      False
4      False
       ...
886    False
887    False
888    False
889    False
890    False
Length: 891, dtype: bool
```

In [5]:

```
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
```

In [6]:

```
cat_col = [col for col in df.columns if df[col].dtype == 'object']
print('Categorical columns :',cat_col)
```

```
Categorical columns : ['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked']
```

In [7]:

```
num_col =[col for col in df.columns if df[col].dtype !='object']
print('Numberical columns :',num_col)
```

```
Numberical columns : ['PassengerId', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']
```

In [8]:

```
df1 = df.drop(columns = ['Name','Ticket'])
df1.shape
```

Out[8]:

```
(891, 10)
```

In [48]:

```
df.isnull()
```

Out[48]:

	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	False	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	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

891 rows × 12 columns

In [50]:

```
round((df1.isnull().sum()/df1.shape[0])*100,2)
```

Out[50]:

```
PassengerId     0.00
Survived        0.00
Pclass          0.00
Sex             0.00
Age            19.87
SibSp           0.00
Parch           0.00
Fare            0.00
Cabin          77.10
Embarked        0.22
dtype: float64
```

In [52]:

```
df2 = df1.drop(columns='Cabin')
df2.dropna(subset=['Embarked'], axis=0, inplace=True)
df2.shape
```

Out[52]:

```
(889, 9)
```

In [54]:

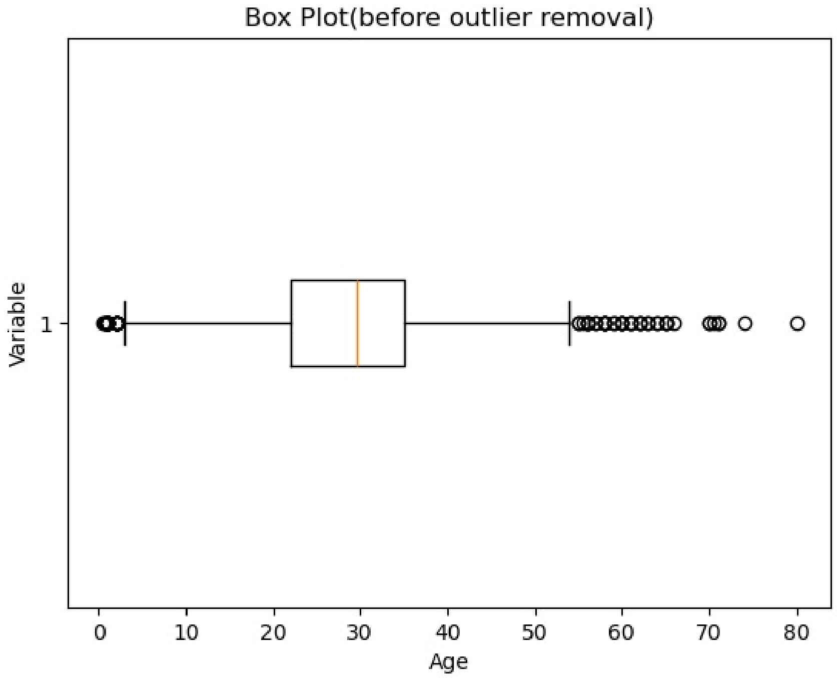
```
df3= df2.fillna(df2.Age.mean())
df3.isnull().sum()
```

Out[54]:

```
PassengerId    0
Survived       0
Pclass         0
Sex            0
Age            0
SibSp          0
Parch          0
Fare           0
Embarked       0
dtype: int64
```

In [56]:

```
plt.boxplot(df3['Age'], vert=False)
plt.ylabel('Variable')
plt.xlabel('Age')
plt.title('Box Plot(before outlier removal)')
plt.show()
```



In [58]:

```
mean = df3['Age'].mean()
std = df3['Age'].std()

lower_bound = mean - std*2
upper_bound = mean + std*2

print('Lower Bound :',lower_bound)
print('Upper Bound :',upper_bound)

df4=df3[(df3['Age']>=lower_bound)&(df3['Age']<=upper_bound)]
df4.shape
```

```
Lower Bound : 3.705400107925648
Upper Bound : 55.578785285332785
```

Out[58]:

```
(821, 9)
```

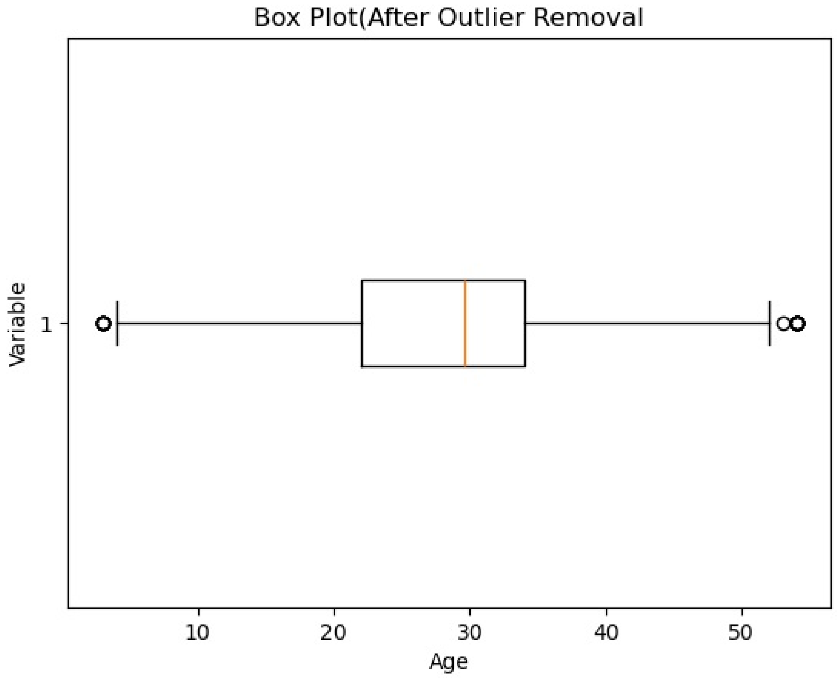
In [60]:

```
Q1 = df3['Age'].quantile(0.25)
Q3 = df3['Age'].quantile(0.75)
IQR = Q3 - Q1

lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR

# Get index labels (not just positional index)
upper_array = np.where(df3['Age'] >= upper)[0]
lower_array = np.where(df3['Age'] <= lower)[0]

print("Upper outlier indices:",upper_array)
print("Lower outlier indices:",lower_array)
outlier_idx = df3.index[np.concatenate([upper_array,lower_array])]
df3.drop(index=outlier_idx, inplace=True)

print("New Shape:", df3.shape)
```

```
Upper outlier indices: [ 11  15  33  54  93  95 115 151 169 173 194 231 251 267 274 279 325 365
 437 455 466 482 486 491 492 544 554 569 586 624 625 629 646 658 671 683
 693 744 771 849 877]
Lower outlier indices: [  7  16  77 118 163 171 182 204 296 304 339 380 385 468 478 529 641 643
 754 787 802 823 826 829]
New Shape: (824, 9)
```

In [62]:

```
plt.boxplot(df3['Age'], vert=False)
plt.ylabel('Variable')
plt.xlabel('Age')
plt.title('Box Plot(After Outlier Removal')
plt.show()
```



In [64]:

```
X = df3[['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked']]
Y = df3['Survived']

print("Features shape:", X.shape)
print("Target shape:", Y.shape)
```

```
Features shape: (824, 7)
Target shape: (824,)
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js