EXTERNSHIP
ASSIGNMENT 2

Done By:

Lalitha Sri. B

20BCE7127.

1. Download the dataset:
2. Load the dataset:

```
import pandas as pd
titanic = pd.read_csv(r'C:\Users\lalit\Downloads\titanic.csv')
print(titanic)
```

```
     survived  pclass     sex   age  sibsp  parch     fare embarked   class  \
0           0       3    male  22.0      1      0   7.2500        S   Third
1           1       1  female  38.0      1      0  71.2833        C   First
2           1       3  female  26.0      0      0   7.9250        S   Third
3           1       1  female  35.0      1      0  53.1000        S   First
4           0       3    male  35.0      0      0   8.0500        S   Third
..        ...     ...     ...   ...    ...    ...      ...      ...     ...
886         0       2    male  27.0      0      0  13.0000        S  Second
887         1       1  female  19.0      0      0  30.0000        S   First
888         0       3  female   NaN      1      2  23.4500        S   Third
889         1       1    male  26.0      0      0  30.0000        C   First
890         0       3    male  32.0      0      0   7.7500        Q   Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True
..   ...         ...  ...          ...   ...    ...
886  man        True  NaN  Southampton    no   True
887  woman      False    B  Southampton   yes   True
888  woman      False  NaN  Southampton    no  False
889  man        True    C    Cherbourg   yes   True
890  man        True  NaN   Queenstown    no   True

[891 rows x 15 columns]
```

```
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

3. Perform Below Visualizations.

Univariate Analysis

Bi - Variate Analysis

Multi - Variate Analysis

```
titanic.head()
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

```
titanic.shape
```

```
(891, 15)
```

```
titanic.dtypes
```

```
survived        int64
pclass          int64
sex            object
age           float64
sibsp           int64
parch           int64
fare          float64
embarked       object
class          object
who            object
adult_male       bool
deck           object
embark_town    object
alive          object
alone            bool
dtype: object
```

4. Perform descriptive statistics on the dataset.

```
titanic['survived'].replace({0:'Not survived', 1:'survived'}, inplace=True)
```

```
titanic.head()
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	Not survived	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	survived	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	survived	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	survived	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	Not survived	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

```
category = pd.crosstab(titanic['pclass'],titanic['survived'], margins=True)
print(category)
```

```
survived  Not survived  survived  All
pclass
1                   80       136  216
2                   97        87  184
3                  372       119  491
All                549       342  891
```

5. Handle the Missing values.

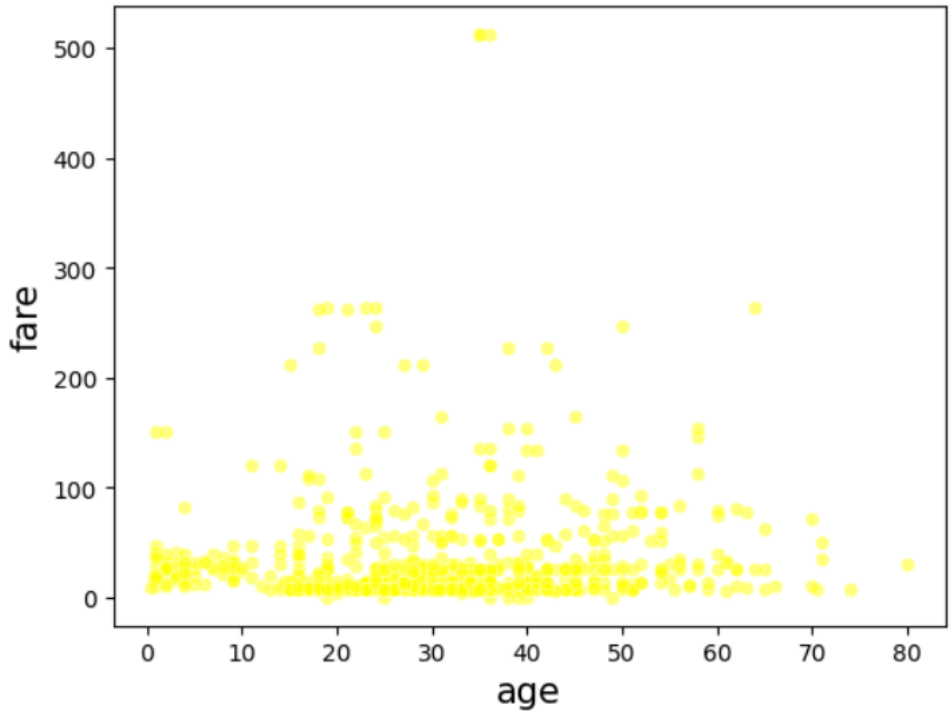
```
titanic.fillna(titanic.mean(numeric_only=True).round(1), inplace=True)
print(titanic)
```

```
      survived  pclass     sex   age  sibsp  parch     fare embarked  \
0   Not survived       3    male  22.0      1      0   7.2500        S
1       survived       1  female  38.0      1      0  71.2833        C
2       survived       3  female  26.0      0      0   7.9250        S
3       survived       1  female  35.0      1      0  53.1000        S
4   Not survived       3    male  35.0      0      0   8.0500        S
..          ...     ...     ...   ...    ...    ...      ...      ...
886 Not survived       2    male  27.0      0      0  13.0000        S
887     survived       1  female  19.0      0      0  30.0000        S
888 Not survived       3  female  29.7      1      2  23.4500        S
889     survived       1    male  26.0      0      0  30.0000        C
890 Not survived       3    male  32.0      0      0   7.7500        Q

     class    who  adult_male deck  embark_town alive  alone
0    Third    man        True  NaN  Southampton    no  False
1    First  woman       False    C    Cherbourg   yes  False
2    Third  woman       False  NaN  Southampton   yes   True
3    First  woman       False    C  Southampton   yes  False
4    Third    man        True  NaN  Southampton    no   True
..     ...    ...         ...  ...          ...   ...    ...
886  Second   man        True  NaN  Southampton    no   True
887  First  woman       False    B  Southampton   yes   True
888  Third  woman       False  NaN  Southampton    no  False
889  First    man        True    C    Cherbourg   yes   True
890  Third    man        True  NaN   Queenstown    no   True

[891 rows x 15 columns]
```

6. Find the outliers and replace the outliers

```
sns.scatterplot(x=titanic['age'], y=titanic['fare'], alpha=0.5, color='yellow')
plt.xlabel('age', fontsize=15)
plt.ylabel('fare', fontsize=15)
```

```
Text(0, 0.5, 'fare')
```



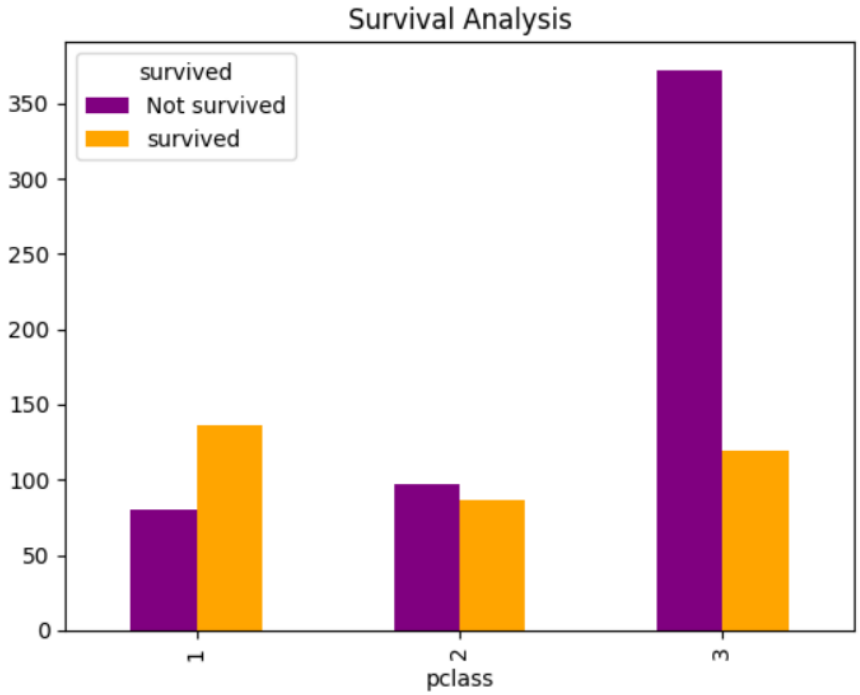
7. Check for Categorical columns and perform encoding.

```
category = pd.crosstab(titanic['pclass'],titanic['survived'], margins=True)
print(category)
```

```
survived  Not survived  survived  All
pclass
1                 80       136  216
2                 97        87  184
3                372       119  491
All              549       342  891
```

```
category.iloc[:-1,:-1].plot(kind='bar',stacked=False, color=['purple','orange'], grid=False, title='Survival Analysis')
```

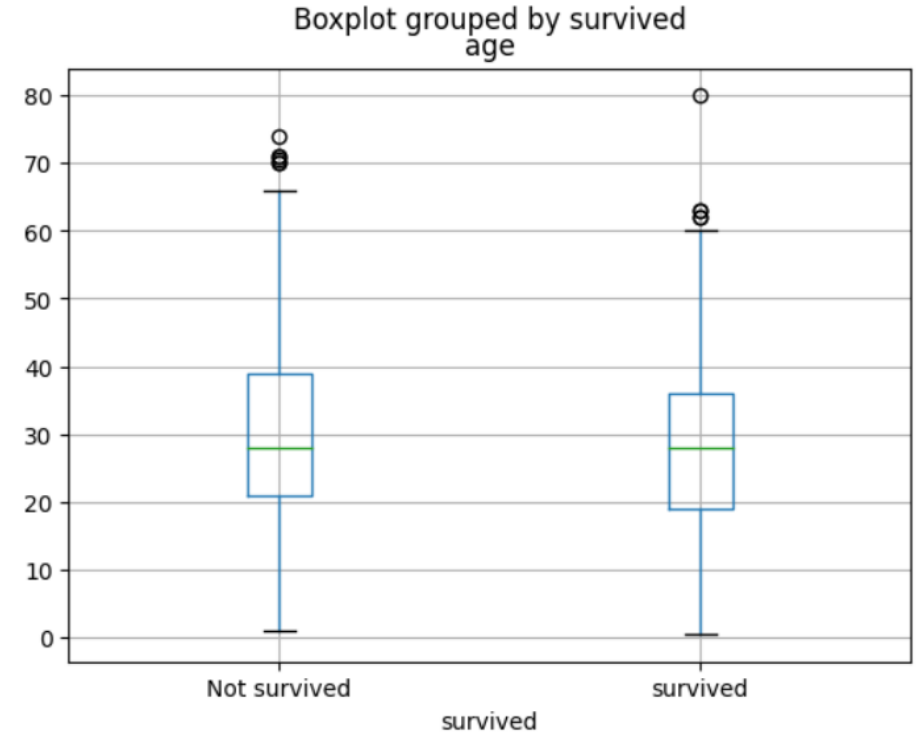
```
<Axes: title={'center': 'Survival Analysis'}, xlabel='pclass'>
```



8. Split the data into dependent and independent variables.

```
titanic.boxplot(column='age', by='survived')
```

```
<Axes: title={'center': 'age'}, xlabel='survived'>
```



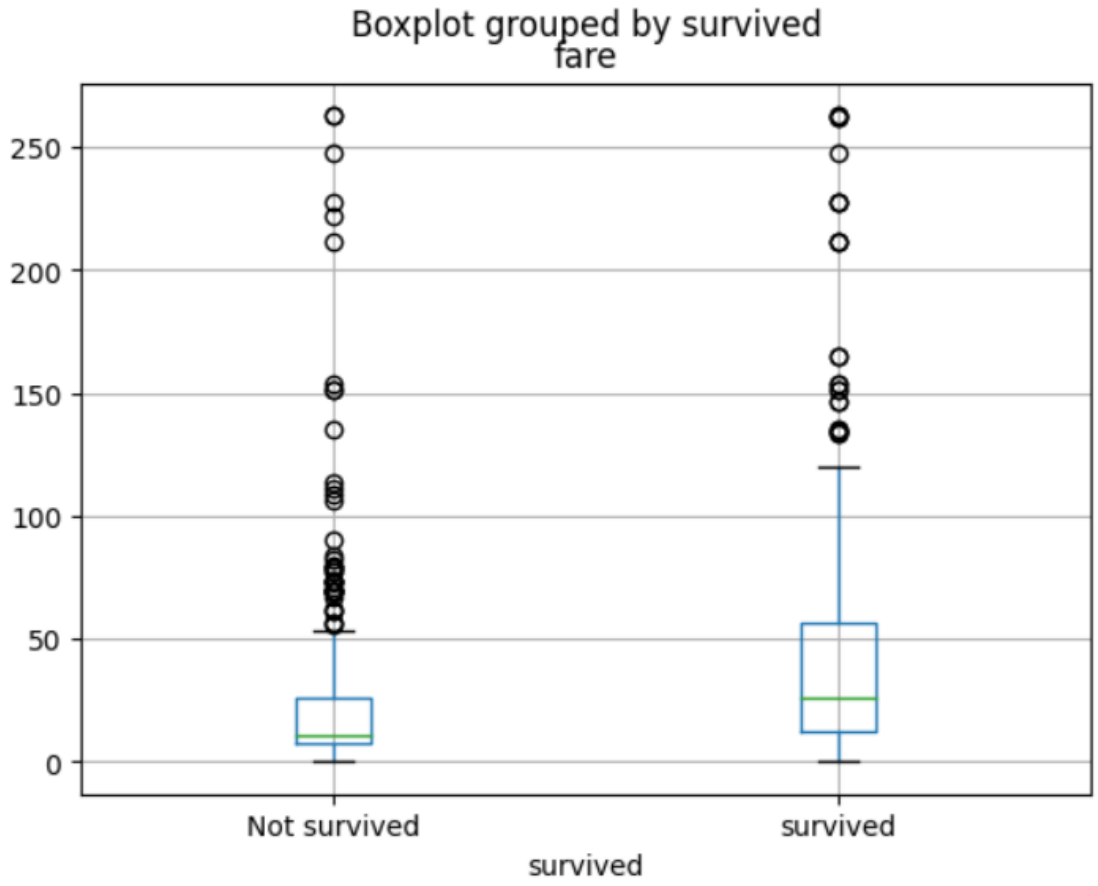
9. Scale the independent variables

```
fare_filter_df = titanic[titanic['fare']<=300]
fare_filter_df.head()
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	Not survived	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	survived	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	survived	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	survived	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	Not survived	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

```
fare_filter_df.boxplot(column='fare', by='survived')
```

```
<Axes: title={'center': 'fare'}, xlabel='survived'>
```



10. Split the data into training and testing

INPUT:

```
import pandas as pd
from sklearn.model_selection import train_test_split
titanic = pd.read_csv(r'C:\Users\lalit\Downloads\titanic.csv')

features = titanic.drop('survived', axis=1)
labels = titanic['survived']
print(features)
print(labels)

X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.4, random_state=42)

X_val, X_test, y_val, y_test = train_test_split(X_test, y_test, test_size=0.5, random_state=42)

for dataset in [y_train, y_val, y_test]:
    print(round(len(dataset) / len(labels), 2))
```

OUTPUT:

```
     pclass    sex   age  sibsp  parch     fare embarked  class    who  \
0         3   male  22.0      1      0   7.2500        S  Third    man
1         1 female  38.0      1      0  71.2833        C  First  woman
2         3 female  26.0      0      0   7.9250        S  Third  woman
3         1 female  35.0      1      0  53.1000        S  First  woman
4         3   male  35.0      0      0   8.0500        S  Third    man
..      ...    ...   ...    ...    ...      ...      ...    ...    ...
886       2   male  27.0      0      0  13.0000        S Second    man
887       1 female  19.0      0      0  30.0000        S  First  woman
888       3 female   NaN      1      2  23.4500        S  Third  woman
889       1   male  26.0      0      0  30.0000        C  First    man
890       3   male  32.0      0      0   7.7500        Q  Third    man

     adult_male deck  embark_town alive  alone
0          True  NaN  Southampton    no  False
1         False    C    Cherbourg   yes  False
2         False  NaN  Southampton   yes   True
3         False    C  Southampton   yes  False
4          True  NaN  Southampton    no   True
..          ...  ...          ...   ...    ...
886        True  NaN  Southampton    no   True
887       False    B  Southampton   yes   True
888       False  NaN  Southampton    no  False
889        True    C    Cherbourg   yes   True
890        True  NaN   Queenstown    no   True


[891 rows x 14 columns]
0    0
1    1
2    1
3    1
4    0
    ..
886  0
887  1
888  0
889  1
890  0
Name: survived, Length: 891, dtype: int64
0.6
0.2
0.2
```