Assignment No: 9

Title of the Assignment: Data Visualization II

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of
age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age')
2. Write observations on the inference from the above statistics.

Objective:

Students should be able to perform the data Visualization operation using Python on any open-source dataset.

Prerequisite:

1. Basic of Python Programming
2. box plot Library, Concept of Data Visualization.

Contents for Theory:

1. Box plot Basics
2. Know your Data
3. Finding patterns of data.

Theory:

Data Visualisation plays a very important role in Data mining. Various data scientists spent their time exploring data through visualisation. To accelerate this process, we need to have a well-documentation of all the plots.

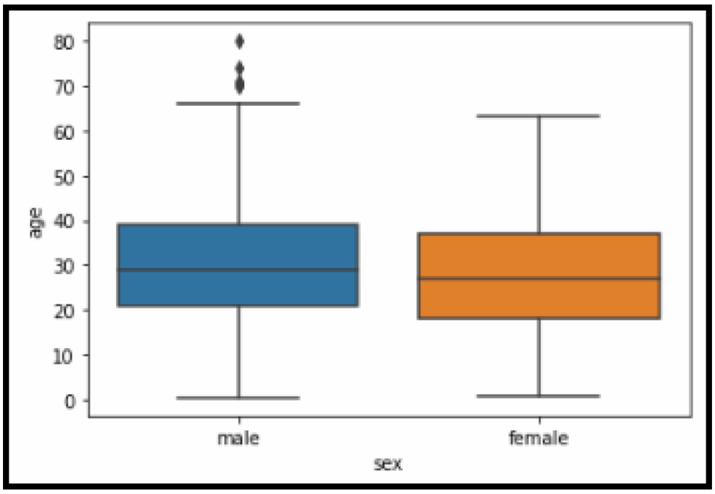
Even plenty of resources can't be transformed into valuable goods without planning and architecture.

The Box Plot

The box plot is used to display the distribution of the categorical data in the form of quartiles. The centre of the box shows the median value. The value from the lower whisker to the bottomof the box shows the first quartile. From the bottom of the box to the middle of the box lies the second quartile. From the middle of the box to the top of the box lies the third quartile and finally from the top of the box to the top whisker lies the last quartile.

Now let's plot a box plot that displays the distribution for the age with respect to each gender. You need to pass the categorical column as the first parameter (which is sex in our case) and the numeric column (age in our case) as the second parameter. Finally, the dataset is passed as the third parameter, take a look at the following script:

```
sns.boxplot(x='sex', y='age', data=dataset)
```

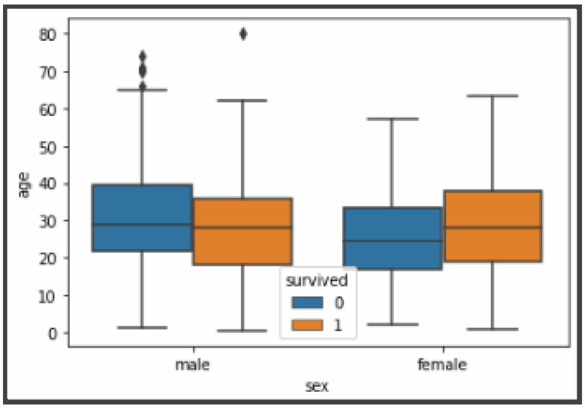


Let's try to understand the box plot for females. The first quartile starts at around 1 and ends at 20 which means that 25% of the passengers are aged between 1 and 20. The second quartile starts at around 20 and ends at around 28 which means that 25% of the passengers are aged between20 and 28. Similarly, the third quartile starts and ends between 28 and 38, hence 25% passengers are aged within this range and finally the fourth or last quartile starts at 38 and ends around 64.

If there are any outliers or the passengers that do not belong to any of the quartiles, they are called outliers and are represented by dots on the box plot.

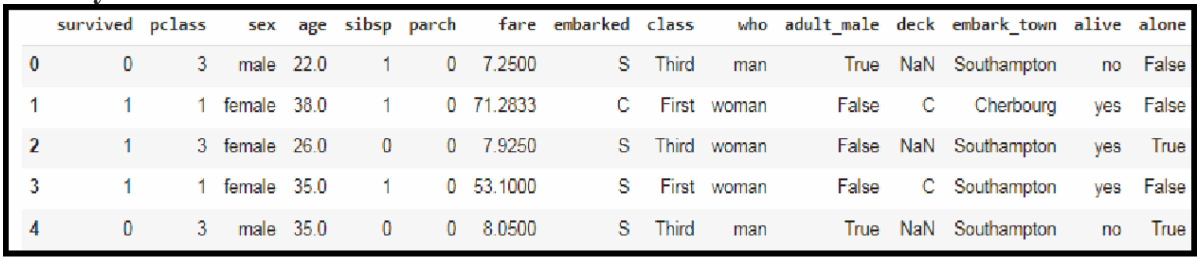
You can make your box plots more fancy by adding another layer of distribution. For instance, ifyou want to see the box plots of forage of passengers of both genders, along with the information about whether or not they survived, you can pass the survived as value to the hue parameter as shown below:

```
sns.boxplot(x='sex', y='age', data=dataset, hue="survived")
```



Now in addition to the information about the age of each gender, you can also see the distribution of the passengers who survived. For instance, you can see that among the male passengers, on average more younger people survived as compared to the older ones. Similarly, you can see that the variation among the age of female passengers who did not survive is much greater than the age of the surviving female passengers.

Know your data



	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

The dataset contains 891 rows and 15 columns and contains information about the passengers who boarded the unfortunate Titanic ship. The original task is to predict whether or not the passenger survived depending upon different features such as their age, ticket, cabin they boarded, the class of the ticket, etc.

Theory:

A box plot is a graphical representation of data that shows the distribution of a dataset. It shows the median, quartiles, and outliers of a dataset. The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The line inside the box represents the median. The whiskers represent the range of data, excluding outliers, and the dots or asterisks represent outliers.

Example:

Here is an example code to plot a box plot for the distribution of age with respect to each gender in the 'titanic' dataset using Python:

```
import seaborn as sns

import matplotlib.pyplot as plt

titanic = sns.load_dataset('titanic')

sns.boxplot(x='sex', y='age', hue='survived', data=titanic)

plt.show()
```

In this code, we first import the required libraries 'seaborn' and 'matplotlib.pyplot'. We then load the 'titanic' dataset using the 'sns.load_dataset()' function. Next, we use the 'sns.boxplot()' function to plot the box plot for the distribution of age with respect to each gender, along with information about whether they survived or not. We pass the 'sex' and 'age' columns of the dataset as the x and y parameters, respectively. We also pass the 'survived' column of the dataset as the hue parameter, which colors the box plot based on whether the passengers survived or not. Finally, we use the 'plt.show()' function to display the plot.

Conclusion-

We learned how to plot a box plot for the distribution of age with respect to each gender in the 'titanic' dataset, along with information about whether they survived or not. We also drew some observations from the statistics. Box plots are a useful tool to visualize the distribution of data and to draw inferences from it.

Assignment Questions

1. List out different types of plot to find patterns of data
2. Explain when you will use distribution plots and when you will use categorical plots.
3. Write the conclusion from the following swarm plot (consider titanic dataset)
4.

In [10]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [11]:
```
titanic = sns.load_dataset("titanic")
```

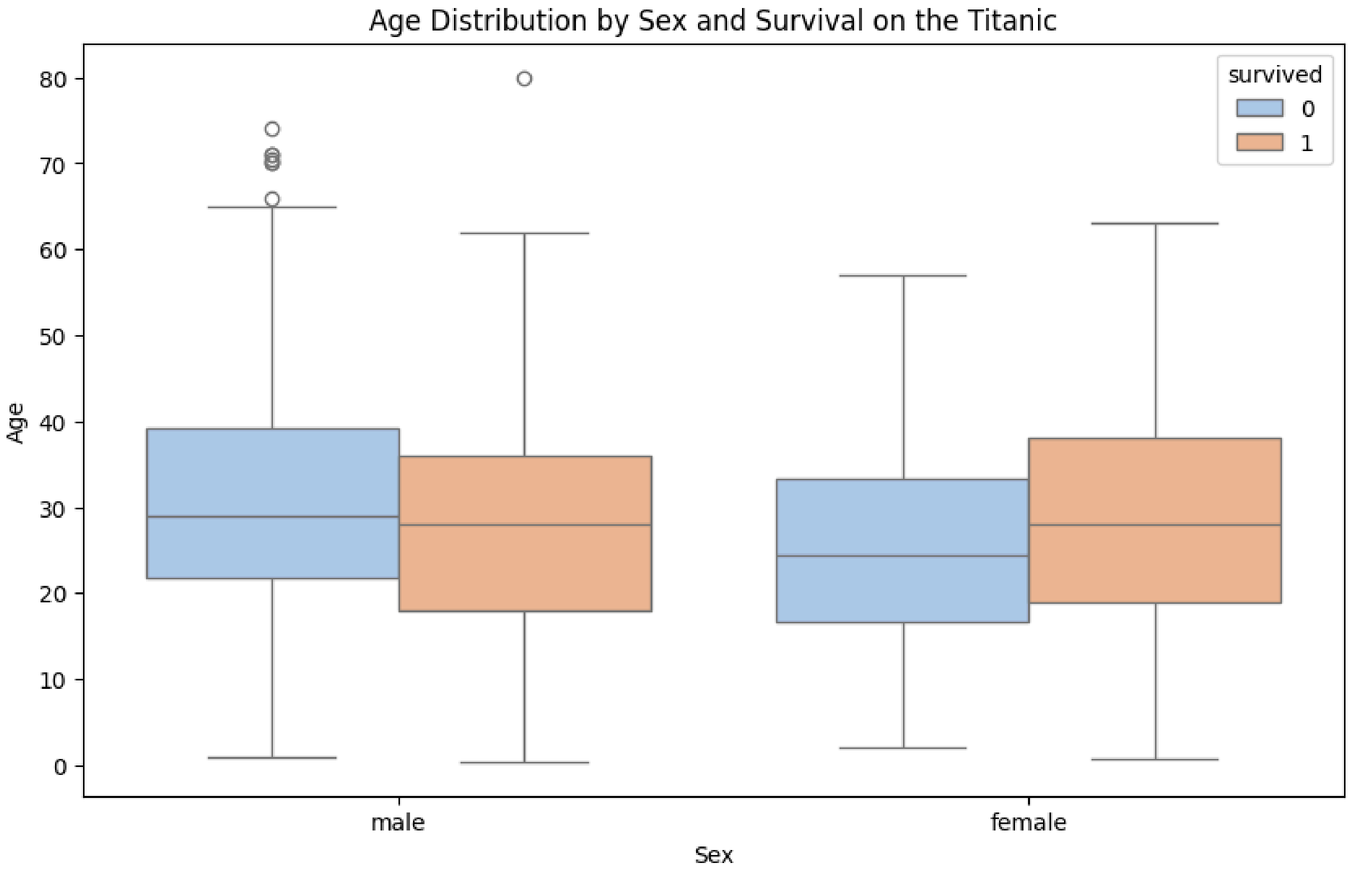
In [14]:
```
male=titanic[(titanic['sex']=='male')&(titanic['survived']==1)&(titanic['age']>20 )&(titanic[
```

In [19]:
```
male.shape[0]
```

Out[19]:
```
45
```

Box Plot

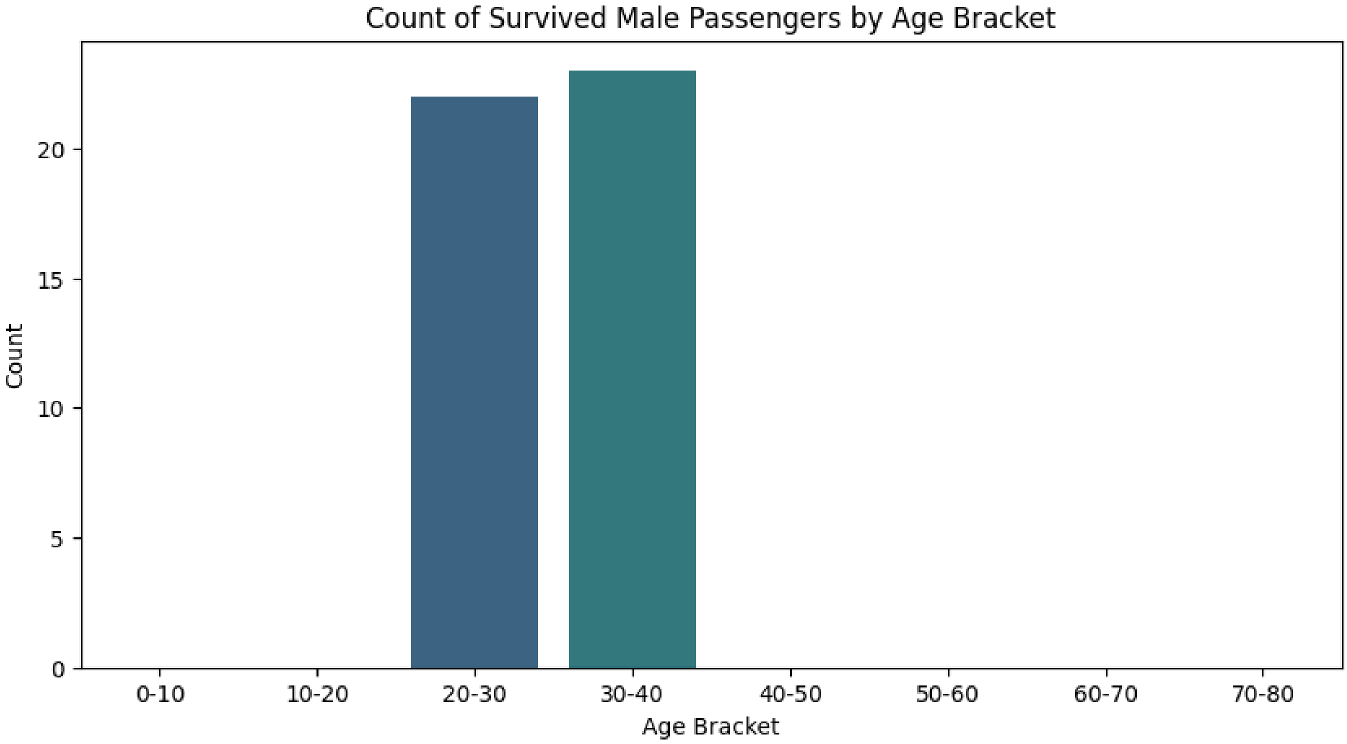
In [22]:
```
plt.figure(figsize=(10,6))
sns.boxplot(x='sex',y='age',hue='survived', data=titanic, palette="pastel")
plt.xlabel("Sex")
plt.ylabel("Age")
plt.title("Age Distribution by Sex and Survival on the Titanic")
plt.show()
```



In [30]:
```
import pandas as pd

male = male.dropna(subset=['age'])

bins = [0, 10, 20, 30, 40, 50, 60, 70, 80]
labels = ['0-10', '10-20', '20-30', '30-40', '40-50', '50-60', '60-70', '70-80']

# Create an 'age_bracket' column based on defined bins
titanic['age_bracket'] = pd.cut(male['age'], bins=bins, labels=labels, right=False)

# Plot a count plot for survived males per age bracket
plt.figure(figsize=(10, 5))
```

```
sns.countplot(x='age_bracket', data=male, palette="viridis")
plt.xlabel("Age Bracket")
plt.ylabel("Count")
plt.title("Count of Survived Male Passengers by Age Bracket")
plt.show()
```

```
C:\Users\91705\AppData\Local\Temp\ipykernel_17824\3013887974.py:13: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign
the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(x='age_bracket', data=male, palette="viridis")
```


Count of Survived Male Passengers by Age Bracket
Count
Age Bracket