**Experiment No. 9**

**Aim: 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.**

Introduction:

In this article we will look at [Seaborn](https://seaborn.pydata.org/) which is another extremely useful library for data visualization in Python. The Seaborn library is built on top of Matplotlib and offers many advanced data visualization capabilities.

Though, the Seaborn library can be used to draw a variety of charts such as matrix plots, grid plots, regression plots etc., in this article we will see how the Seaborn library can be used to draw distributional and categorial plots. In the [second part](https://stackabuse.com/seaborn-library-for-data-visualization-in-python-part-2/) of the series, we will see how to draw regression plots, matrix plots, and grid plots.

### Downloading the Seaborn Library

The seaborn library can be downloaded in a couple of ways. If you are using pip installer for Python libraries, you can execute the following command to download the library:

pip install seaborn

Alternatively, if you are using the Anaconda distribution of Python, you can use execute the following command to download the seaborn library:

conda install seaborn

Theory:

### The Dataset

**Titanic Dataset –**  
It is one of the most popular datasets used for understanding machine learning basics. It contains information of all the passengers aboard the RMS Titanic, which unfortunately was shipwrecked. This dataset can be used to predict whether a given passenger survived or not.

The dataset that we are going to use to draw our plots will be the Titanic dataset, which is downloaded by default with the Seaborn library. Now we have to use the load\_dataset function and pass it the name of the dataset.

Let's see what the Titanic dataset looks like. Execute the following script:

import pandas as pd

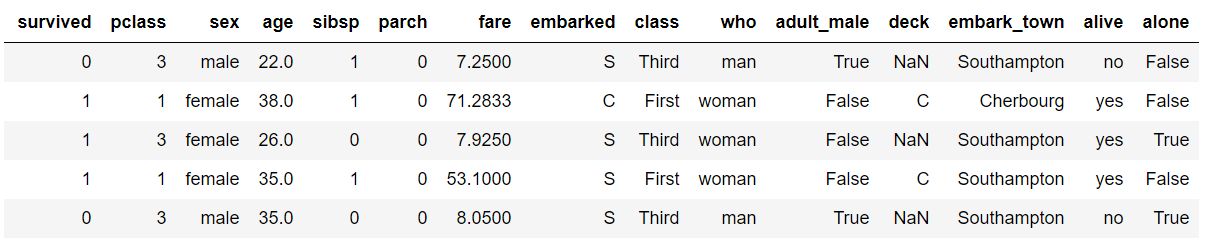
import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

dataset = sns.load\_dataset('titanic')

dataset.head()

.

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. We will use the Seaborn library to see if we can find any patterns in the data.

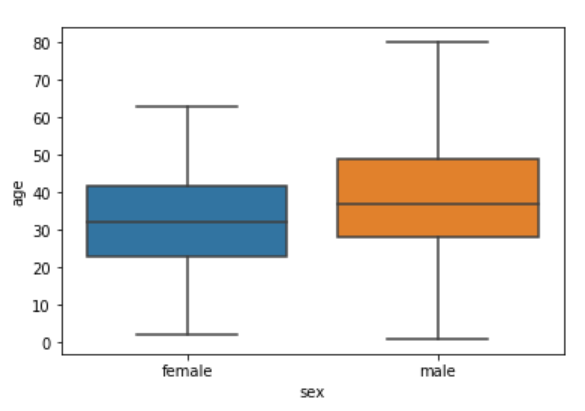
#### The Box Plot:

The box plot is used to display the distribution of the categorical data in the form of quartiles. The center of the box shows the median value. The value from the lower whisker to the bottom of 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 plot a box plot that displays the distribution for the age with respect to each gender. Here we 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)

**Output:**

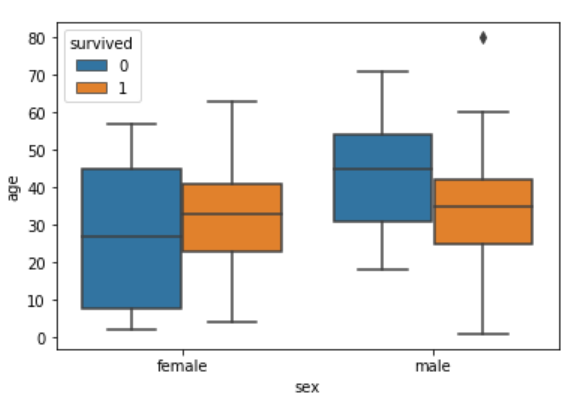


Now try to understand the box plot for female. The first quartile starts at around 5 and ends at 22 which means that 25% of the passengers are aged between 5 and 25. The second quartile starts at around 23 and ends at around 32 which means that 25% of the passengers are aged between 23 and 32. Similarly, the third quartile starts and ends between 34 and 42, hence 25% passengers are aged within this range and finally the fourth or last quartile starts at 43 and ends around 65. 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.

We can make our box plots more fancy by adding another layer of distribution. For instance, if we 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")

**Output:**



Now in addition to the information about the age of each gender, we can also see the distribution of the passengers who survived. For instance, we can see that among the male passengers, on average more younger people survived as compared to the older ones. Similarly, we 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.

### Conclusion: [Seaborn](https://seaborn.pydata.org/) is an advanced data visualization library built on top of [Matplotlib library](https://matplotlib.org/). In this practical we looked at how we can draw box plot using Seaborn library.