**Assignment 9**

**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.

**What is Data Visualization?**

Data Visualization represents the text or numerical data in a visual format, which makes it easy to grasp the information the data express. We, humans, remember the pictures more easily than readable text, so Python provides us various libraries for data visualization like matplotlib, seaborn, plotly, etc. We will use Matplotlib and seaborn for performing various techniques to explore data using various plots.

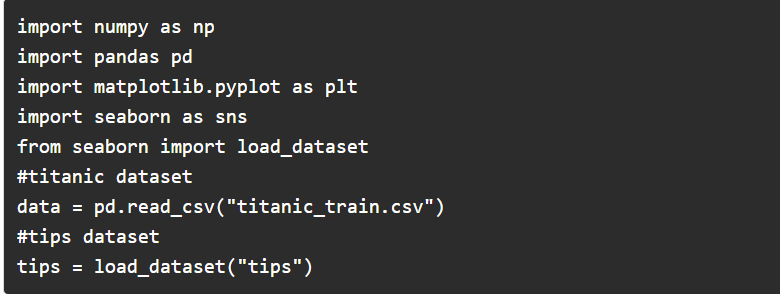
**Exploratory Data Analysis**

Creating Hypotheses, testing various business assumptions while dealing with any Machine learning problem statement is very important and this is what EDA helps to accomplish. There are various tootle and techniques to understand your data, And the basic need is you should have the knowledge of Numpy for mathematical operations and Pandas for data manipulation.

We will use a very popular Titanic dataset with which everyone is familiar with and you can download it from [here](https://www.kaggle.com/c/titanic/data).

Now lets us start exploring data and study data visualization plots with Boxplot.

let’s get started by importing libraries and loading Data

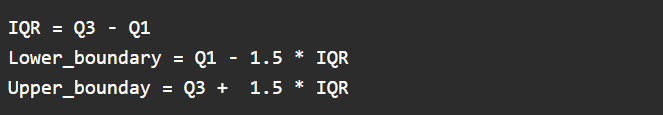


Analysing Numerical data is important because understanding the distribution of variables helps to further process the data. Most of the time you will find much inconsistency with numerical data so do explore numerical variables.

**Boxplot**

Boxplot is a very interesting plot that basically plots a 5 number summary. to get 5 number summary some terms we need to describe.

* **Median** – Middle value in series after sorting
* **Percentile** – Gives any number which is number of values present before this percentile like for example 50 under 25th percentile so it explains total of 50 values are there below 25th percentile
* **Minimum and Maximum** – These are not minimum and maximum values, rather they describe the lower and upper boundary of standard deviation which is calculated using Interquartile range (IQR).



Here Q1 and Q3 is 1st quantile (25th percentile) and 3rd Quantile (75th percentile)

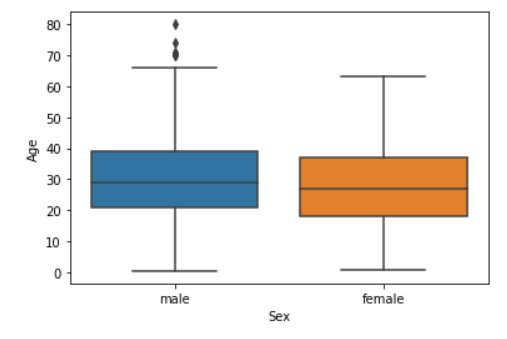
**Bivariate/ Multivariate Analysis**

Bivariate Analysis is used when we have to explore the relationship between 2 different variables and we have to do this because, in the end, our main task is to explore the relationship between variables to build a powerful model. And when we analyze more than 2 variables together then it is known as Multivariate Analysis.

First, let’s explore the boxplots when if one variable is numerical and one is categorical then there are various plots that we can use for Bivariate and Multivariate analysis.

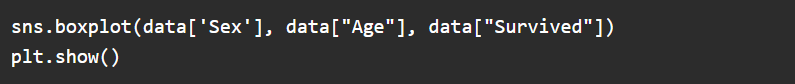
We can draw a separate boxplot for both the variable, let us explore gender with age using a boxplot.

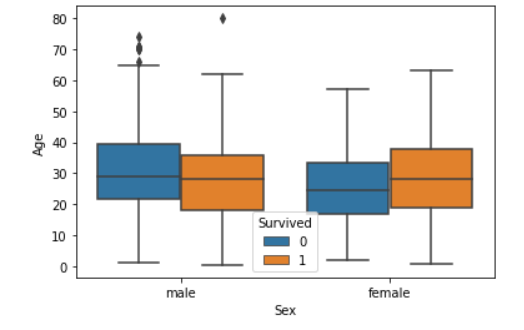




**Multivariate analysis with boxplot**

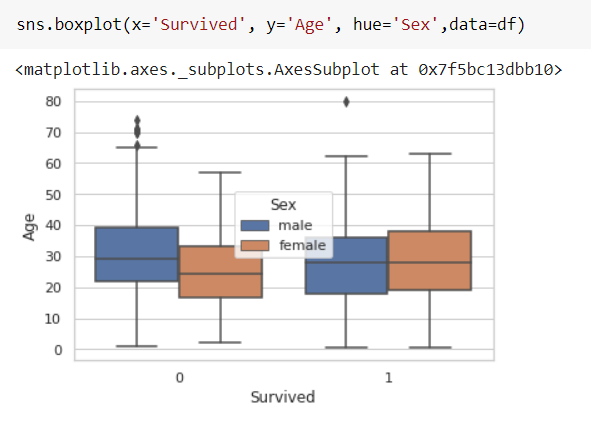
Along with age and gender let’s see who has survived and who has not.

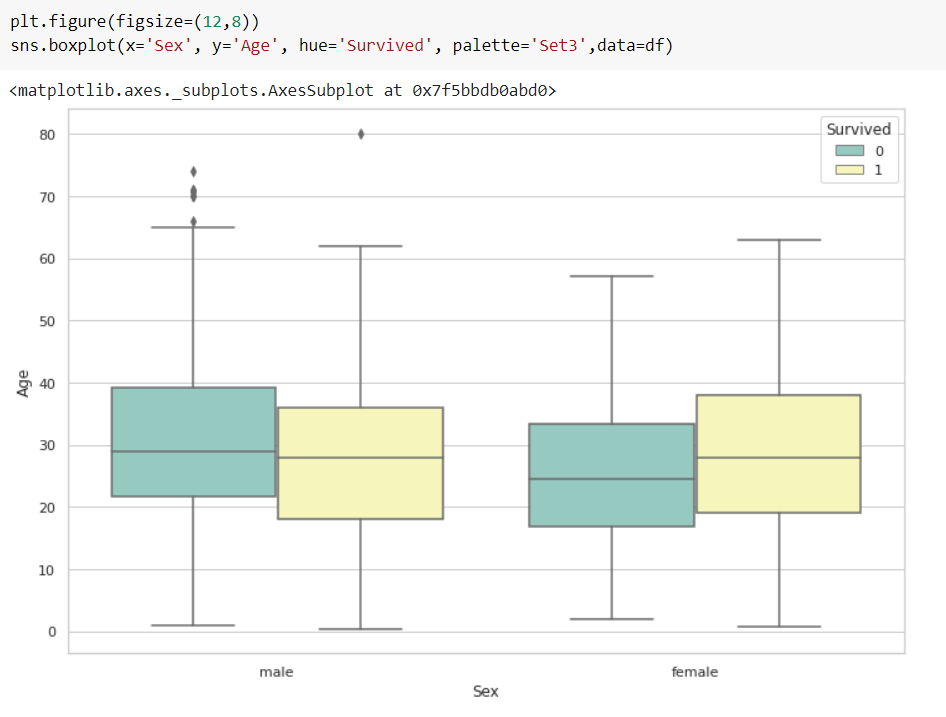


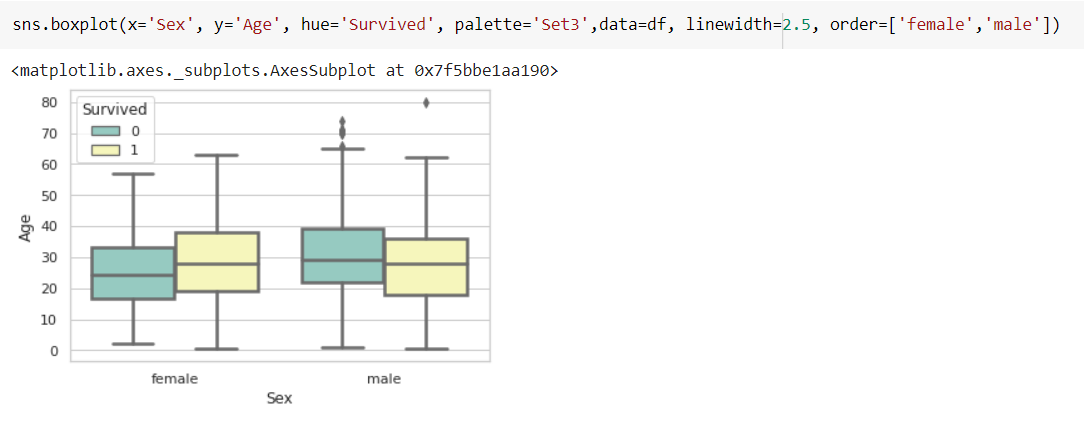


**Code Snipplet:**

****

****

****

****