

# 1. Introduction

Welcome to the series of E-learning modules on box plot. In this module, we are going to cover the definition, utility, construction and analysis of box plot as a graphical tool of representation of data.

By the end of this session, you will be able to:

- Explain box plot
- Explain the uses of box plot
- Explain the construction of box plot for the given data
- Explain the analysis and interpretation of the data

In descriptive statistics, it is assumed that data points are clustered around some central value. Hence, the data are generally summarized around these values. A box plot also known as a box-whisker plot is a good way for visual representation of the summarized values of a data set.

Let us first understand the meaning of box plot.

A box plot displays the range and distribution of a data along a number line. It is a convenient way of graphically depicting groups of numerical data through their five-number summaries. Boxplots can be drawn either horizontally or vertically.

Now, we will find out when we should use a box plot.

- Box plots are particularly useful for comparing distributions of the results from several experimental conditions.
- A box plot helps in indicating which observations might be considered outliers.
- A box plot tells the spacing between the different parts of the box and helps indicate the degree of dispersion (spread) and skewness in the data.

After understanding the use of box plot, let us understand the parts of a box plot.

Title: The title briefly describes the information that is contained in the box plot.

**Figure 1**



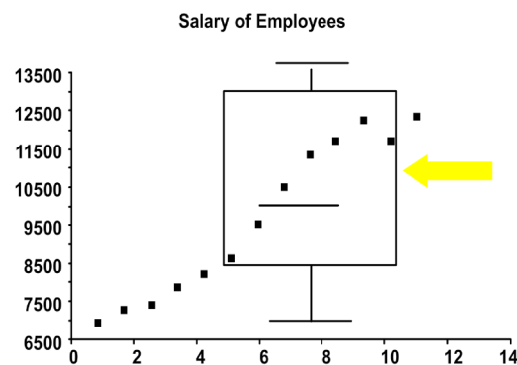
Number line: A number line is a line representing the range and distribution of the data.

**Figure 2**



Box: One or more boxes are plotted in a graph each representing a cluster of data. A box is formed by connecting the lower quartile, median and the upper quartile.

**Figure 3**



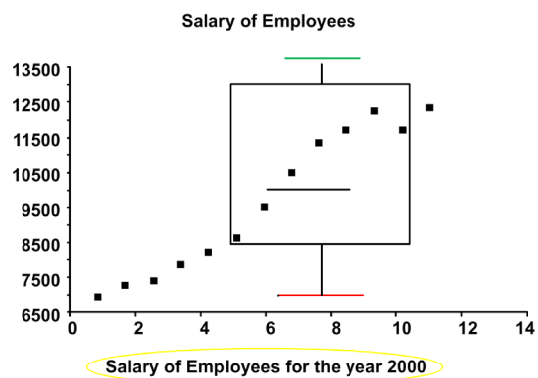
Whiskers: Whiskers on either side of the boxes represent the minimum and maximum samples. They are lines joining the box and the extreme values of the data.

**Figure 4**



Legend: The legend provides additional information about the documents like where the data came from and how the measurements were gathered.

**Figure 5**



## 2. Understanding of Box Plot

Let us understand how a box plot can be constructed for a given data.

A Box-and-Whisker plot is made of:

One or more boxes each represent a cluster of data and whiskers on either side of the boxes representing the minimum and maximum samples, which are drawn on a number line to depict the data.

In a box-and-whisker plot, the concentration of data is divided into two quartiles based on medians. The concentration is then visually depicted using boxes and line connectors.

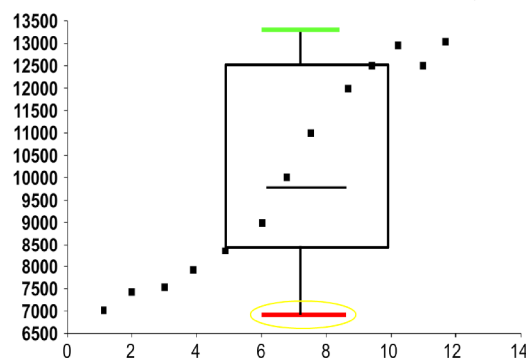
Box plots are depicted using the five-member summaries of numerical data. These summaries are:

- The Smallest Observation
- Lower Quartile
- The Median
- Upper Quartile
- The Largest Observation

Now, let us know each summary.

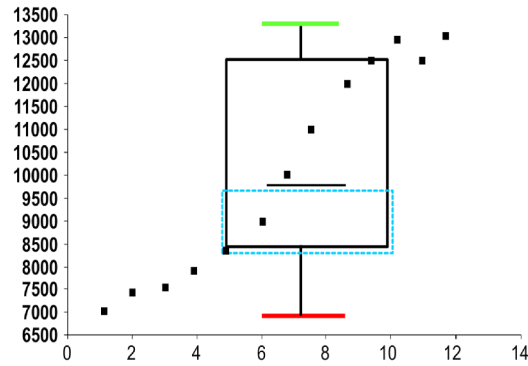
The Smallest Observation: This is the smallest data point in the data set and called as the minimum value of the data set. It is the first value in an ordered set of data.

**Figure 6**



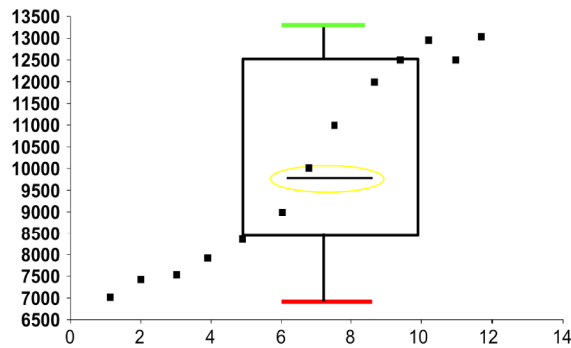
Lower Quartile ( $Q_1$ ): This value divides the distribution in such a way that one-fourth of the total items fall below it and three-fourth fall above it. If there is no middle value in the set of data, use the average of the two middle values of the data.

**Figure 7**



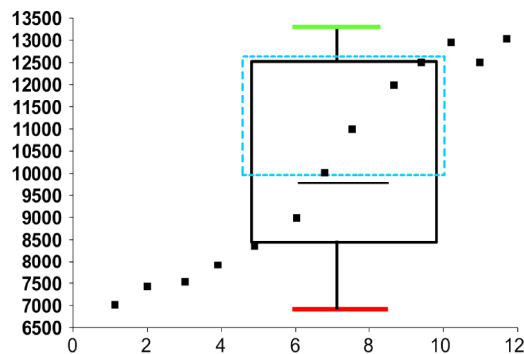
The Median ( $Q_2$ ): This is a value of the distribution, which splits the series into two equal parts. One part comprising the all the values greater than the median and the other part comprise of values less than the median. If there is no middle value in the set of data, use the average of the two middle values of the data.

**Figure 8**



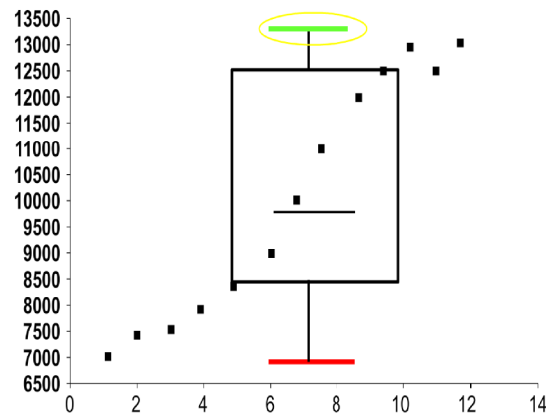
Upper Quartile ( $Q_3$ ): this value divides the distribution in such a way that three fourth of the total items fall below it and one fourth of the item falls above it. If there is no middle value in the set of data, use the average of the two middle values of the data.

**Figure 9**



The Largest Observation: this is the largest data point in the distribution and is called the maximum value of the data set. It is the last value in an ordered set of data.

**Figure 10**



Now, let us discuss the steps involved in the construction of a box plot.

Step 1: Ordering the data: This is the first step in constructing the box plot, wherein the data is arranged from the least to the greatest value of the distribution.

For example in a given data set 15, 21, 47, 55, 78, 65, 56, 6, 19, 38, 24

It is ordered as 6, 15, 19, 21, 24, 38, 47, 55, 56, 65, 78

Step 2: Finding the median that will split the data into two equal halves.

Example : For the same data set taken in step 1 we will get the median as 38, with data from 6 to 24 in the lower half and 47 to 78 as the upper half.

Step 3: Finding the lower quartile that is the median of the lower half of the data set.

Example: In continuation with the same data, we will get 19 as the median for the lower half of the data set.

Step 4: Finding the upper quartile that is the median of the upper half of the data set.

Example: using the same data set we will get 56 as the median for the upper half of the data set.

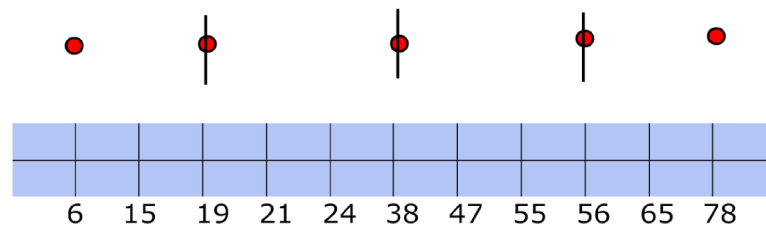
Step 5: Identifying the extreme values of the data that is the maximum and minimum value of the data set.

Example: In the same data set, we can identify 6 and 78 as the extreme values 6 being the minimum value and 78 being the maximum value.

Step 6: Plotting the graph:

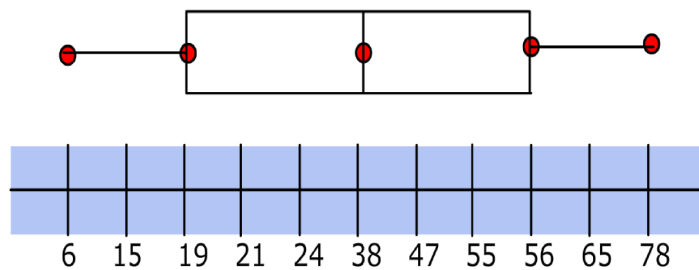
- In this step, first we plot the five values above a number line.  
Here we have plotted the five values that is 6, 19, 38, 56 and 78
- Then we draw vertical lines through the lower quartile, median and the upper quartile.  
The lower quartile is 19, median 38 and upper quartile is 56.

**Figure 11**



- Next, we draw a box by connecting the vertical lines drawn through the lower quartile, median and the upper quartile.
- Next we draw the whiskers from the extremes to the box  
The extreme values are 6 and 78

**Figure 12**



# 3. Interpretation of Box Plot

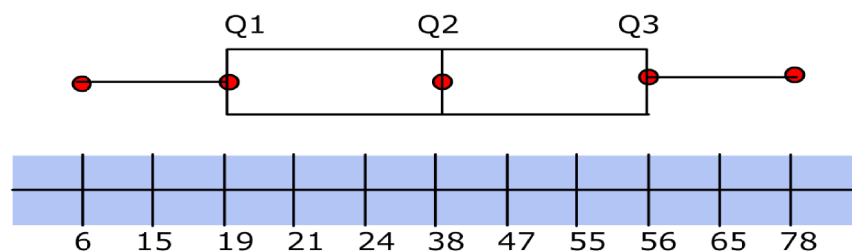
Let us learn how to interpret a box plot.

While interpreting the box plot we need to keep few points in mind.

- First, we need to know that by finding the middle values of the data set we have divided the data into four equal groups called quartiles.

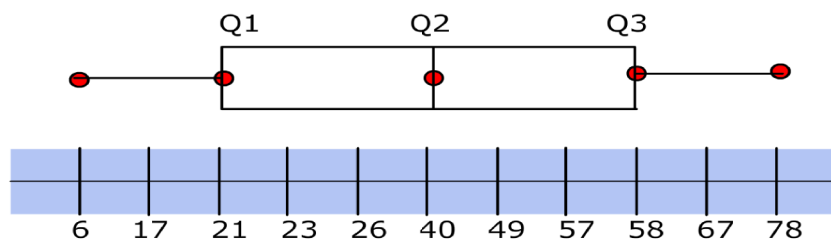
In this example, the first group is 6 to 19, second is 19 to 38, third is 38 to 56 and last group is 56 to 78.

**Figure 13**



- We use the medians for comparisons of the data. The data with higher median is considered to be a better data.

**Figure 14**



Along with this data, let us consider another set of data whose Q1 is 21, Q2 is 40 and Q3 is 58. Since this set of data has higher median, it is considered to be a better data.

We will be able to compare the lower and the higher values for any two sets of data.

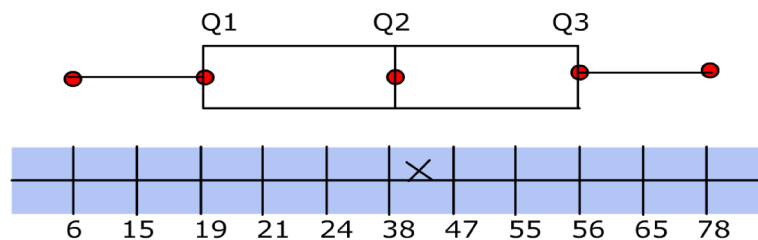
The box-and-whisker plot may include a cross or an "X" marking the mean value of the data, in addition to the line inside the box that marks the median. The difference between the "X" and the median line can then be used as a measure of "skew".

In the first set of data, the mean value is 38.54 and the median is 38. Therefore, the skew is



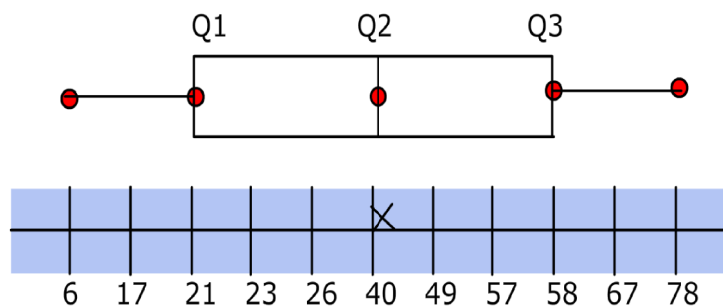
measured 38.54 minus 38 that is equal to 0.54.

**Figure 15**



Similarly, for the second data, the mean value is 40.18 and the median is 40. Therefore, the skew is measured as 0.18.

**Figure 16**



The IQR is the length of the box in your box-and-whisker plot. An outlier is any value that lies on more than one and a half times the length of the box from either end of the box. That is, if a data point is below  $Q1 - 1.5 \times IQR$  or above  $Q3 + 1.5 \times IQR$ , it is viewed as being too far from the central values to be reasonable.

#### Example 1

Let us consider an illustration to understand the concept of outliers. Find the outliers, if any, for the following data set:

6, 15, 19, 21, 24, 38, 47, 55, 56, 65, 78

Solution:

First, we have to find the IQR.

$Q1 = 19$

$Q2 = 38$

$Q3 = 56$

Then,  $IQR = Q3 - Q1$

$$\text{IQR} = 56 - 19 = 37$$

The outliers are

$$Q1 - (1.5 \times \text{IQR}) = 19 - (1.5 \times 37) = -36.5$$

$$Q3 + (1.5 \times \text{IQR}) = 56 + (1.5 \times 37) = 111.5$$

Therefore, any value falling below -36.5 and more than 111.5 are not found in the given data.

## 4. Examples of Box Plot

### Example 2

Let us consider an illustration to understand the concept of outliers. Find the outliers, if any, for the following data set:

10.2, 14.1, 14.4, 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

Solution:

To find out if there are any outliers,

First, we have to find the IQR.

There are fifteen data points, so the median will be at position  $15 \text{ plus } 1 \text{ divided by } 2$  is equal to 8. Then  $Q2$  is equal to 14.6.

There are seven data points on either side of the median, so  $Q1$  is the fourth value in the list and  $Q3$  is the twelfth:  $Q1 = 14.4$  and  $Q3 = 14.9$ .

Then  $IQR = 14.9 \text{ minus } 14.4 = 0.5$ .

Outliers will be any points below  $Q1 \text{ minus } 1.5 \times IQR = 14.4 \text{ minus } 0.75 = 13.65$  or above  $Q3 \text{ plus } 1.5 \times IQR = 14.9 \text{ plus } 0.75 = 15.65$ .

Then the outliers are at 10.2, 15.9, and 16.4.

Let us do a real-life exercise.

Let us look at the following data, depicting basic salaries of same class (grade / role) of employees of a company.

Now, let us assume that the company is planning to do a salary revision exercise. Typical to all companies have a range of salaries that are being offered to the employees. In this case, the range varies from Rs.7000 to Rs.12,900.

The company wants to know how the employees are distributed across the range i.e., whether they are evenly distributed with varying salaries or more concentrated in one range or towards lower or upper sides.

**Figure 17**

Emp. No.	Salary
1	12200
2	8800
3	9800
4	12900
5	8400
6	7000
7	8000
8	7400
9	11800
10	10900
11	12200
12	12800
13	7500

Step 1: Ordering of the data

Since the salary data is in focus, first step is to sort the set of data by salaries. Here is the sorted list in ascending order of salaries.

**Figure 18**

Emp. No.	Salary
6	7000
8	7400
13	7500
7	8000
5	8400
2	8800
4	9800
10	10900
9	11800
11	12200
12	12800
1	12200
4	12900

Step 2: Find out the medians

There are 13 values in this ordered list. The 7<sup>th</sup> will be the median i.e. 9800. Let us call this Q2.

**Figure 19**

Emp. No.	Salary	
6	7000	
8	7400	
13	7500	
7	8000	
5	8400	
2	8800	
4	9800	Q2
10	10900	
9	11800	
11	12200	
12	12800	
1	12200	
4	12900	

The Median divides the data into two halves.

**Figure 20**

Emp. No.	Salary
6	7000
8	7400
13	7500
7	8000
5	8400
2	8800
4	9800
10	10900
9	11800
11	12200
12	12800
1	12200
4	12900

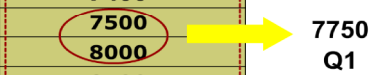
Step 3: Find the lower quartile

The Median for this half will be sum of midpoints  $(7500+8000)$  divided by  $2 = 7750$ .

Let us call this as Q1.

**Figure 21**

Emp. No.	Salary
6	7000
8	7400
13	7500
7	8000
5	8400
2	8800
4	9800
10	10900
9	11800
11	12200
12	12800
1	12200
4	12900




Step 4: Finding the upper quartile

The Median for this half will be sum of midpoints  $(12200 + 12800)$  divided by  $2 = 12500$ .

Let us call this Q3

**Figure 22**

Emp. No.	Salary
6	7000
8	7400
13	7500
7	8000
5	8400
2	8800
4	9800
10	10900
9	11800
11	12200
12	12800
1	12200
4	12900



Step 5: Finding the extreme values

The minimum value for this data will be 7000.

The maximum value for this data will be 12900.

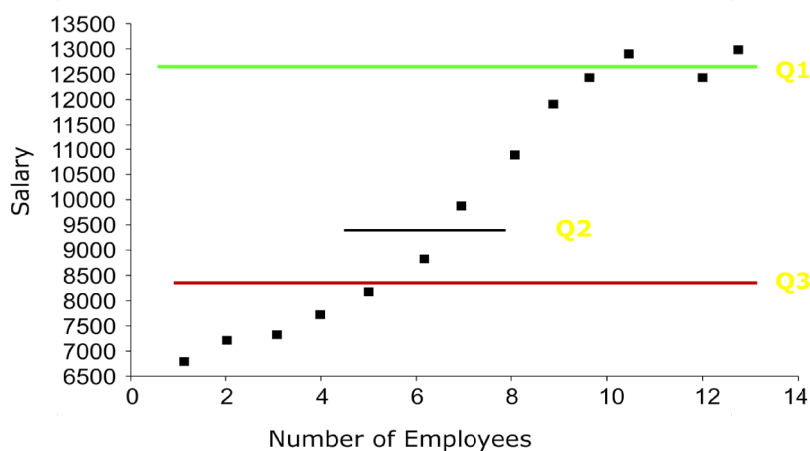
**Figure 23**

Emp. No.	Salary
6	7000
8	7400
13	7500
7	8000
5	8400
2	8800
4	9800
10	10900
9	11800
11	12200
12	12800
1	12200
4	12900

Step 6: Plotting the graph:

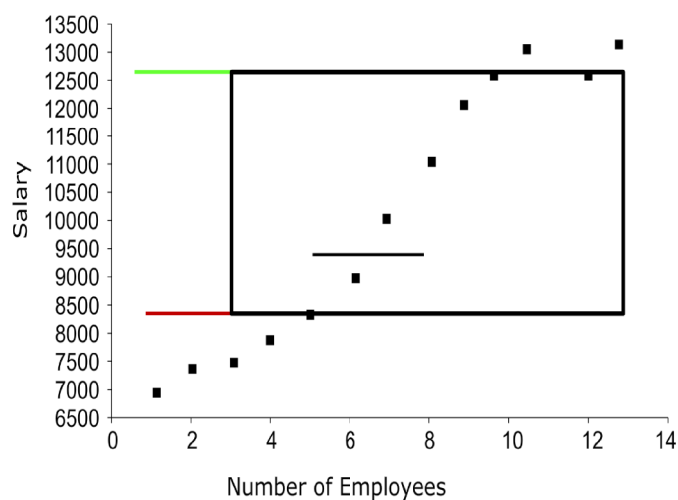
- In this step, first we plot the five values 7750, 9800 and 12,500. The graph will now look like this:

**Figure 24**



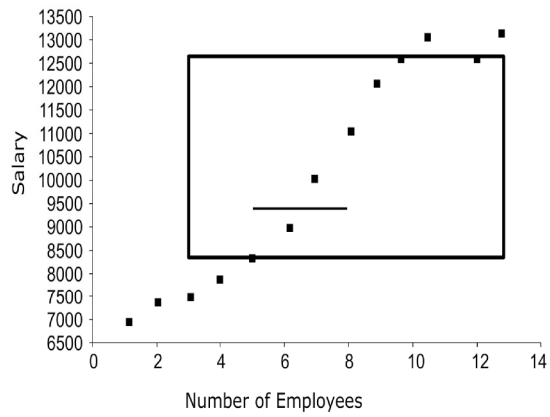
Then we draw vertical lines through the lower quartile 7750, median 9800 and the upper quartile 12,500.

**Figure 25**



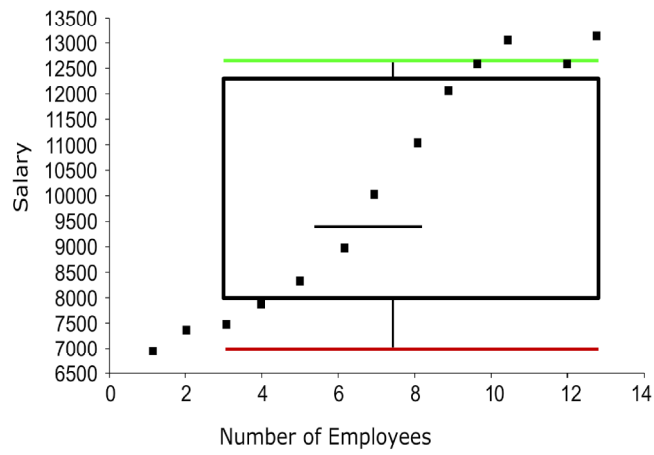
Next, we draw a box by connecting the vertical lines drawn through the lower quartile, median and the upper quartile.

**Figure 26**



Next we draw the whiskers form the extremes 7000 and 12900 to the box

**Figure 27**



# 5. Interpreting Box-and-Whisker Charts

## Interpreting Box-and-Whisker Charts

In this example, what does the box chart tell us?

The Plot tells us several things.

There is a larger concentration of data points in the upper quartile. This means that more employees are concentrated on the upper salary range.

There is a larger concentration of data points outside the box. This means quite a few employees are farther from the mid-point salary level.

Even distribution of salaries can be achieved by bringing up the salaries in the lowest quartile. This will also be the most economical approach for the company.

## Considerations for making Box Plots

We have used a plotter chart in this example. However this is not mandatory, you can use any other type of chart. In fact it is not mandatory to show the data points within the box at all. All we need is to show the 5 summaries i.e. sample minimum, lower quartile, median, upper quartile and sample maximum.

In the example discussed earlier, we had prepared a vertical Box plot. However, boxes can be plotted vertically or horizontally as per your convenience. There is no colour code prescribed for Box plots. You can choose what works for you.

If required, multiple boxes (each representing a unique data set related to same matter) can be represented in same graph, each with its own whiskers. See this example.

## Key usages of Box plots:

Box-and-whisker plots are used for a variety of purposes in our daily lives.

Some examples are:

1. Policymaking: To determine the most deserving segment to benefit from policies.
2. Financial decision-making: To determine the most economical actions to achieve objectives.
3. Understanding of imbalances: For example: Social & Economic conditions across regions.

Let us discuss the advantages and disadvantages of Box plot.

The advantages are:

1. Handles Large Data Easily: Due to the five number data summary, a box plot is able to handle and present a summary of a large amount of data. A box plot consists of the median, which is the midpoint of the range of data; the upper and lower quartiles, which represent the numbers above and below the highest and lower quarters of the data; and the minimum and maximum data values. Organizing data in a box plot by using five key concepts is an efficient way of dealing with large data that is too unmanageable for other graphs, such as line plots or stem and leaf plots.
2. Summarizing: A box plot is a highly visually effective way of viewing a clear summary of



one or more sets of data. It is particularly useful for quickly summarizing and comparing different sets of results from different experiments. At a glance, a box plot allows a graphical display of the distribution of results and provides indications of symmetry within the data.

3. Outliers: A box plot is one of very few statistical graph methods that show outliers. There might be one outlier or multiple outliers within a set of data, which occurs both below and above the minimum and maximum data values. An outlier is an obscure result that can be detected by extending the minimum and maximum data values to a maximum of 1.5 times the inter-quartile range. Any results of data that fall outside of the minimum and maximum values are considered outliers, which are easy to determine on a box plot graph.

The following are the disadvantages of box plot:

1. It is not as visually appealing as other graphs.
2. Exact Values are not retained. The issue with handling such large amounts of data in a box plot is that the exact values and details of the distribution of results are not retained. A box plot shows only a simple summary of the distribution of results, so that it can be quickly viewed and compared with other data. For a thorough, more detailed analysis of data a box plot should be used in combination with another statistical graph method, such as a histogram.

Here's a summary of our learning in this session:

- Understanding of what a box plot is and its usage
- How to segregate and prepare the data for constructing the graph
- A detail understanding of representing the summarized values in the data
- In depth knowledge in understanding its application