STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation

D. Uma
Department of Computer Science and Engineering
umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation
Boxplot

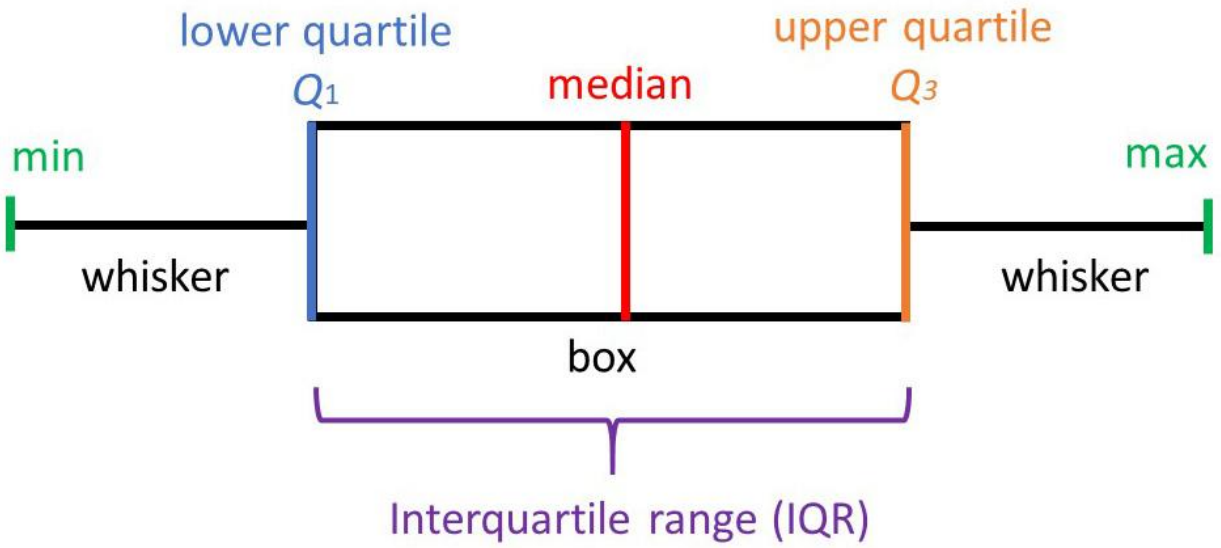
D. Uma
Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Boxplot

A **box and whisker plot** is a way of summarizing a set of data measured on an **interval scale**.

It shows the distribution of a set of data along a number line, dividing the data into four parts using the median and quartiles.



Source: Simply Psycl

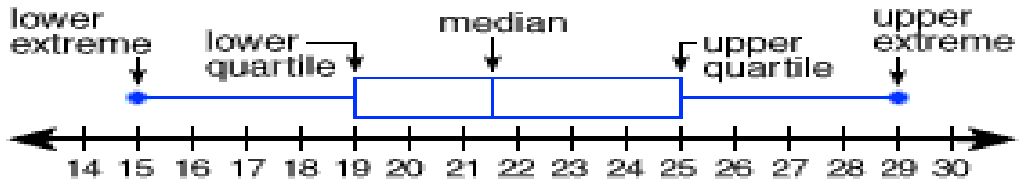
STATISTICS FOR DATA SCIENCE

Why Boxplot?

A **box and whisker plot** is a way of summarizing a set of data measured on an **interval scale**.

It is a graph that presents information from a **five-number summary**.

It is often used in explanatory data analysis.



This type of graph is used to show the shape of the distribution, its central value, and its variability.

Box and whisker plots are ideal for comparing distributions because the centre, spread and overall range are immediately apparent.

Source: www.150.statcan.gc.ca

STATISTICS FOR DATA SCIENCE

Why Boxplot?

It **does not show a distribution** in as much detail as a stem and leaf plot or **histogram** does, but is especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.

Box and whisker plots are also very useful when large numbers of observations are involved and when two or more data sets are being compared.

Shows how the values in the data are spread out.

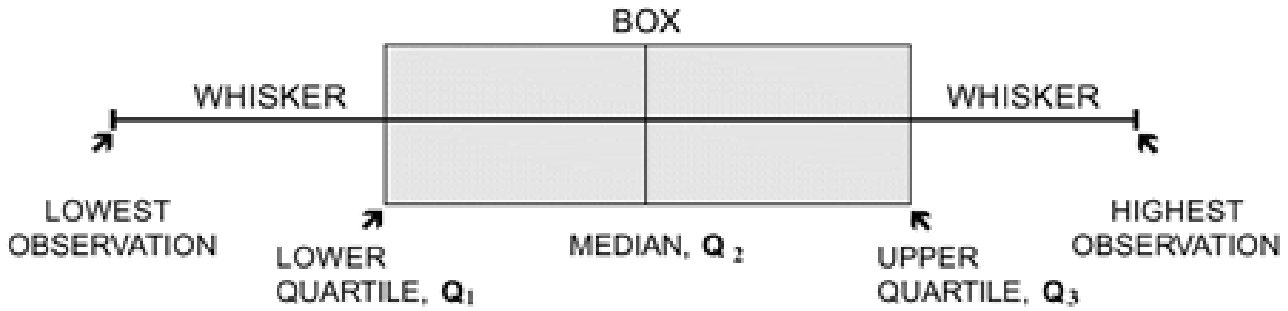
Source: www.150.statcan.gc.ca

Boxplot

In a box and whisker plot:

- the **ends of the box** are the **upper and lower quartiles**, so the box spans the **interquartile range**.
- the **median** is marked by a **vertical line** inside the box.
- the **whiskers** are the **two lines** outside the box that **extend to the highest and lowest observations.**

Figure 1. Box and whisker plot



Source: www.150.statcan.gc.ca

STATISTICS FOR DATA SCIENCE

Five Number Summary - Example

10	11	12	25	25	27	31	33
34	34	35	36	43	50	59	

Arrange the data in order n= 15

Position of the 1st Quartile = 0.25(n+1)=4 ; **First Quartile** = **25**

Position of the Median = $\frac{n+1}{2}$ or 0.50(n+1)= 8 ; **Median** = **33**

Position of the 3rd Quartile = 0.75(n+1)=12 ; **Third Quartile** =**36**

Minimum = **10** **Maximum** = **59**

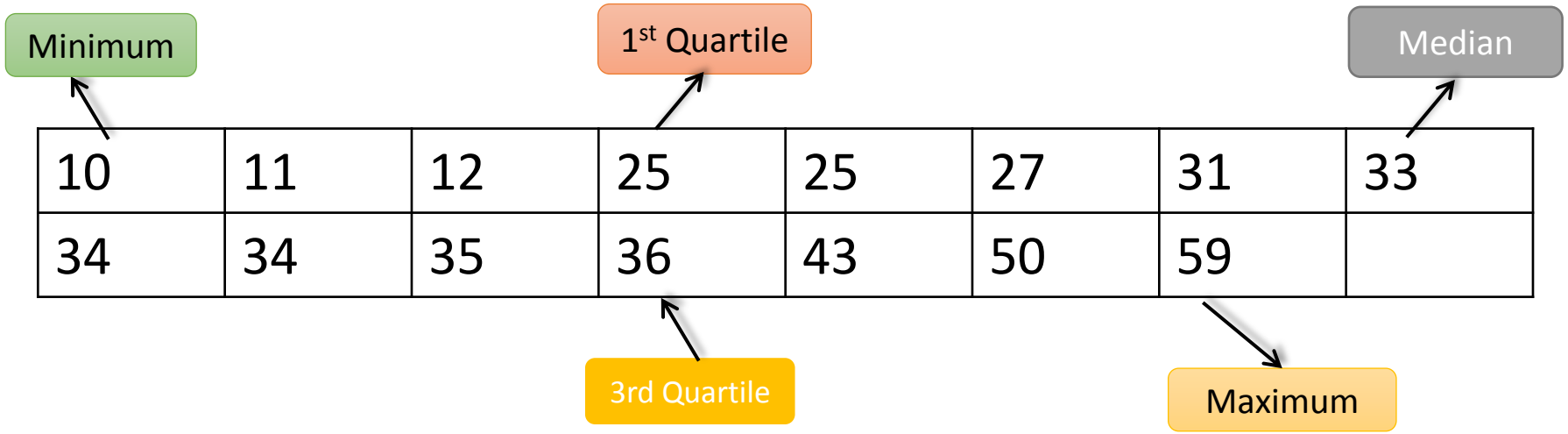
Steps to Construct Boxplot

Step 1: Order the data from smallest to largest.
Step 2: Find the median.
Step 3: Find the quartiles.
Step 4: Complete the five-number summary by finding the min and the max.
Step 5: Making a boxplot.

- **Step 1:** Scale and label an axis that fits the five-number summary.
- **Step 2:** Make solid dots against Q1, Q2 and Q3 values above the number line.
- **Step 3:** Draw vertical lines from number line to those 3 points and complete the box .
- **Step 4:** Draw two horizontal lines from outside box(either side) to till the minimum and maximum value respectively. These lines are called whiskers.

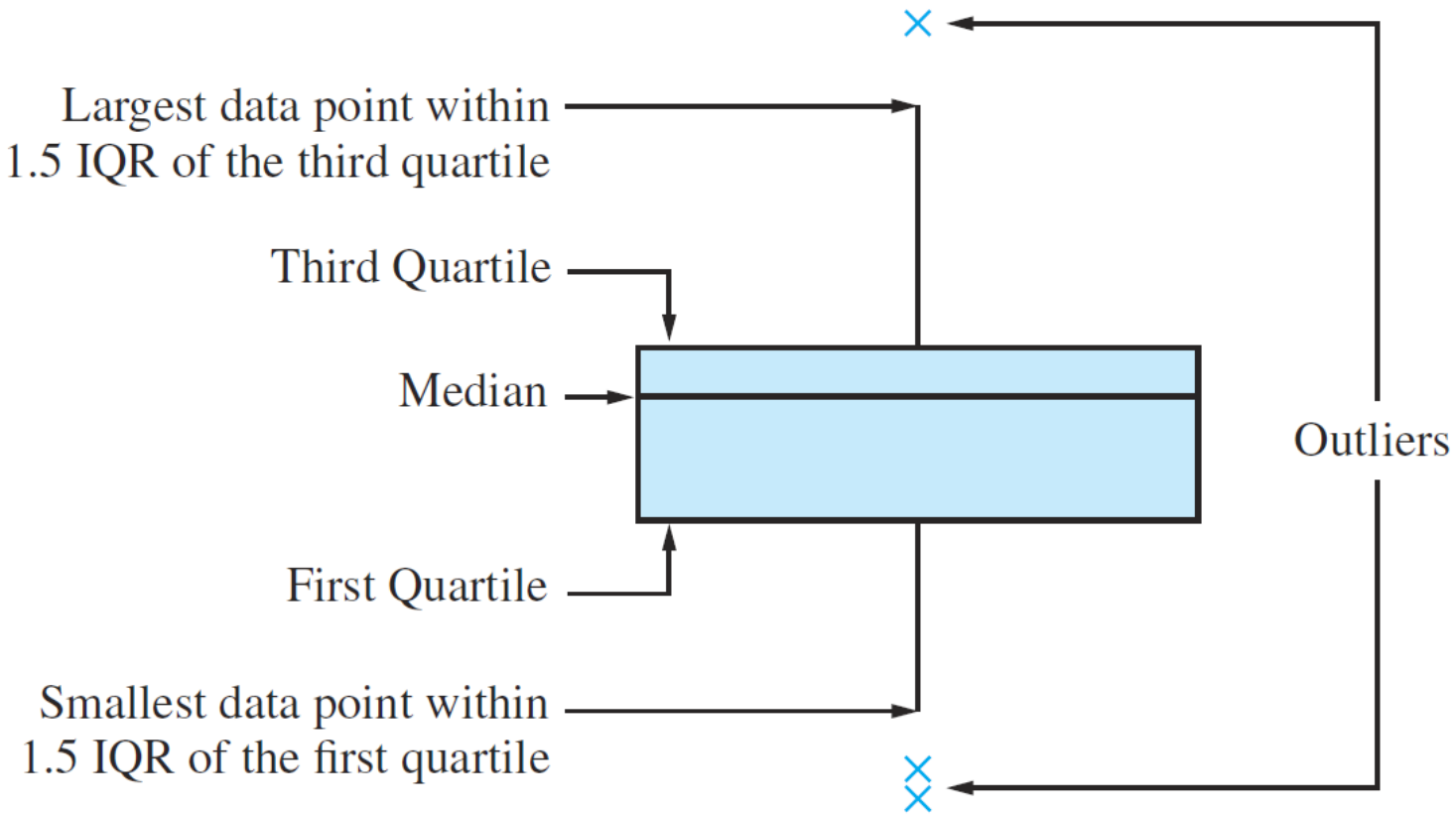
STATISTICS FOR DATA SCIENCE

Five Number Summary - Example



STATISTICS FOR DATA SCIENCE

Detailed Anatomy of Boxplot



STATISTICS FOR DATA SCIENCE

Boxplot:Example

Like Angela, Carl works at a computer store. He also recorded the number of sales he made each month. In the past 12 months, he sold the following numbers of computers:

51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13.

Give a five-number summary of Carl's and Angela's sales.
Make two box and whisker plots, one for Angela's sales and one for Carl's.
Briefly describe the comparisons between their sales.

Source: www.150.statcan.gc.ca

STATISTICS FOR DATA SCIENCE

Boxplot:Example

Given: 51, 17, 25, 39, 7, 49, 62, 41, 20, 6, 43, 13

First, put the data in ascending order.
Then find the median.

6, 7, 13, 17, 20, 25, 39, 41, 43, 49, 51, 62.

Median = (6th + 7^{th})÷ 2 = 6.5th value
= (sixth + seventh observations) ÷ 2
= (25 + 39) ÷ 2
= **32**

Source: www.150.statcan.gc.ca

Boxplot:Example

There are six numbers below the median, namely:
6, 7, 13, 17, 20, 25.
Q1 = the median of these six items
= (6 + 1) ÷ 2= 3.5th value
= (third + fourth observations) ÷ 2
= (13 + 17) ÷ 2
= **15**

Here are six numbers above the median, namely:
39, 41, 43, 49, 51, 62.
Q_3 = the median of these six items
= (6 + 1) ÷ 2= 3.5th value
= (third + fourth observations) ÷ 2
= **46**

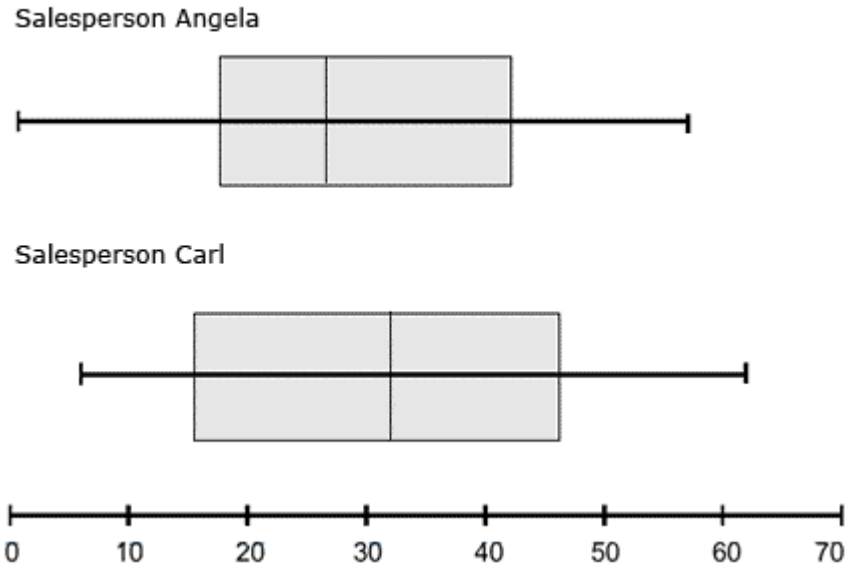
Source: www.150.statcan.gc.ca

Boxplot:Example

The five-number summary for Carl's sales is 6, 15, 32, 46, 62.

Using the same calculations, we can determine that the five-number summary for Angela is 1, 17, 26, 42, 57.

Figure 2. Carl's and Angela's box and whisker plots



Source: www.150.statcan.gc.ca

Boxplot:Example

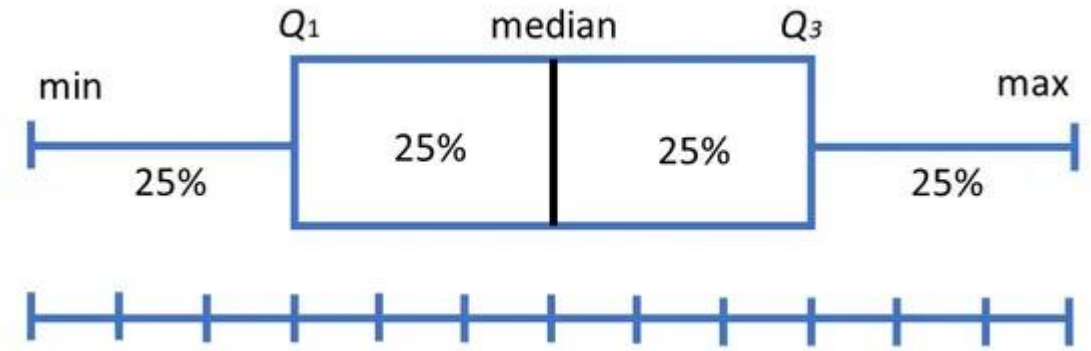
The five-number summary for Carl's sales is 6, 15, 32, 46, 62.

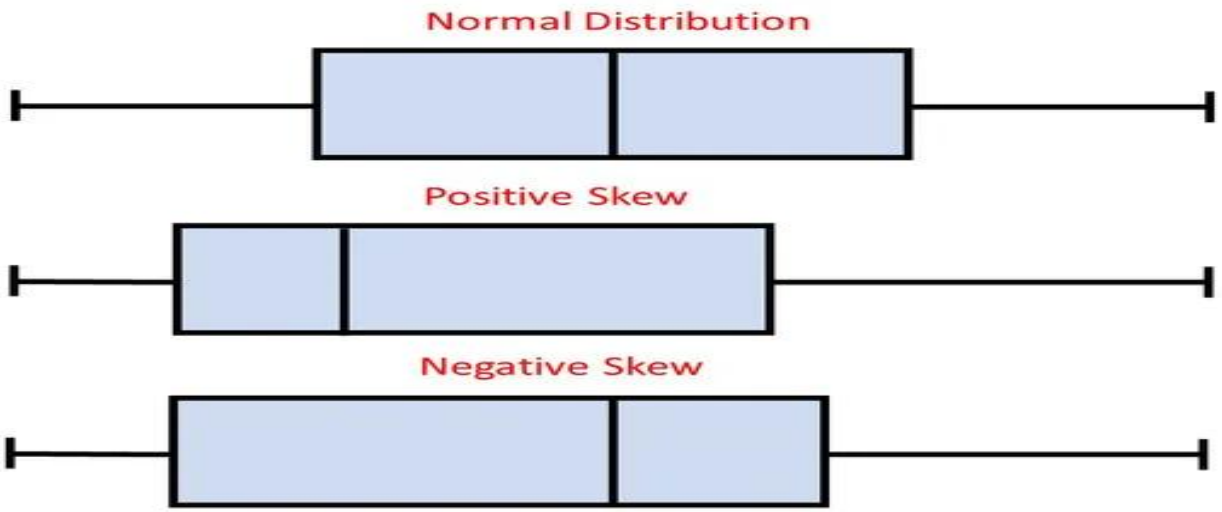
Using the same calculations, we can determine that the five-number summary for Angela is 1, 17, 26, 42, 57.

Source: www.150.statcan.gc.ca

STATISTICS FOR DATA SCIENCE

Boxplot: Symmetry Vs Non-Symmetry





Source: https://www.simplypsychology.org/

STATISTICS FOR DATA SCIENCE

Boxplot:Example

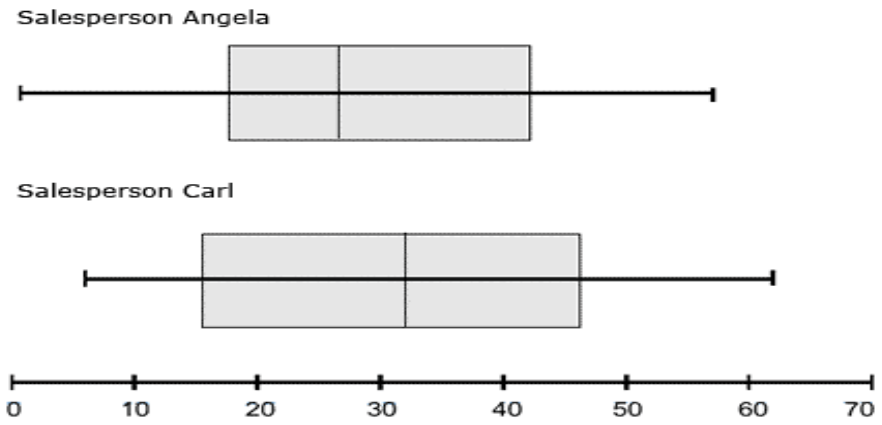
Summary:

Carl's highest and lowest sales are both higher than Angela's corresponding sales, and Carl's median sales figure is higher than Angela's.

Also, Carl's interquartile range is larger than Angela's.

These results suggest that Carl consistently sells more computers than Angela does.

Figure 2. Carl's and Angela's box and whisker plots



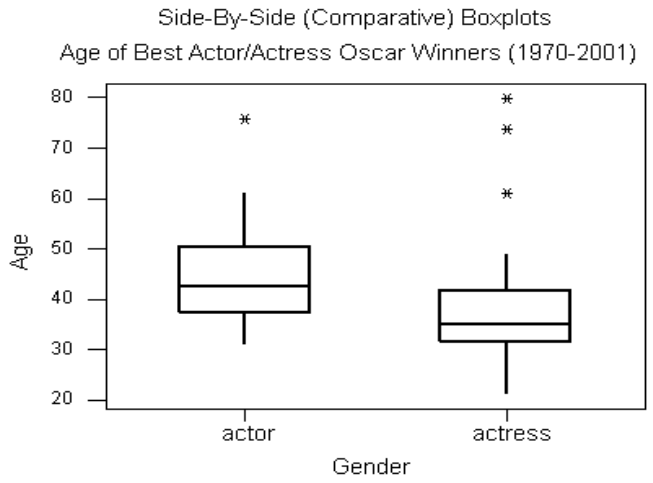
Source: www.150.statcan.gc.ca

Comparative Boxplot

We can compare the boxplots of the two (or more) samples side-by-side.

This will allow us to compare how the medians differ between samples, as well as the first and third quartile.

It also tells us about the difference in spread between the two samples.



Example – Outlier Calculation

Outliers are points that are unusually large or small.

A data value is considered to be an outlier if,

DataValue $<$ $Q_1 - 1.5\ (IQR)$

DataValue $>$ $Q_3 + 1.5\ (IQR)$

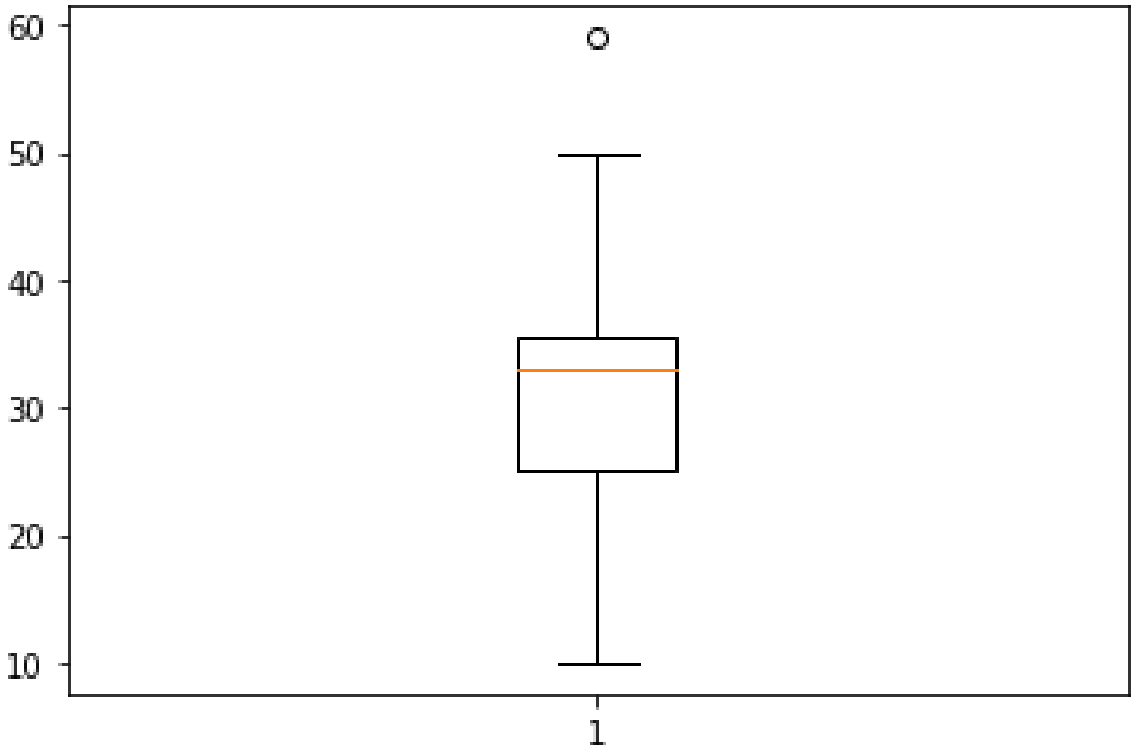
Let's calculate the IQR = $Q_3 - Q_1$ = 36 – 25 = 11

Let's substitute now,

$Q_1 - 1.5\ (11) = 8.5$

$Q_3 + 1.5\ (11) = 52.5$

Resulting Boxplot



Summary

Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines.

Draw vertical lines to complete the box.

Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is no more than 1.5 IQR below the first quartile.

Extend vertical lines (whiskers) from the quartile lines to these points.

Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Plot each outlier individually.

Summary

Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines.

Draw vertical lines to complete the box.

Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is no more than 1.5 IQR below the first quartile.

Extend vertical lines (whiskers) from the quartile lines to these points.

Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Plot each outlier individually.

THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335