<https://chartio.com/learn/charts/what-is-a-scatter-plot/>

<https://calcworkshop.com/functions-statistics/line-best-fit/>

<https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

http://web.mnstate.edu/peil/MDEV102/U4/S36

**In previous sections we discussed various charts. This section will cover plots (Some time also referred as charts only)**

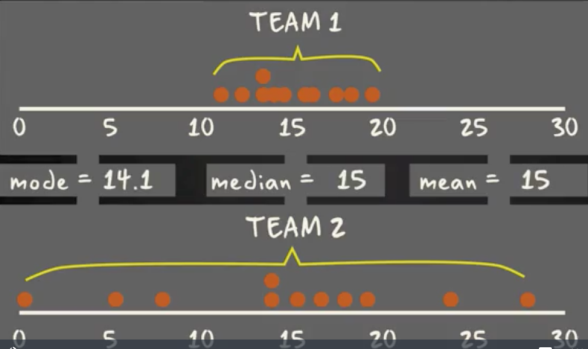
1. **Scatter plot (Scatter Chart)**
2. **Bubble Plot (Chart)**
3. **Box Plot**
4. **Time line Plot chart**

# C. Box Plot

<https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

# What is a Boxplot (Box and whisker Plot)?

For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode).



There are times when mean, median, and mode aren’t enough to describe a dataset. In the figure above two Teams with same mode, median and mean are having different dispersion of the data.

A box-plot is a graph that gives you a good indication of how the values in the data are spread out.

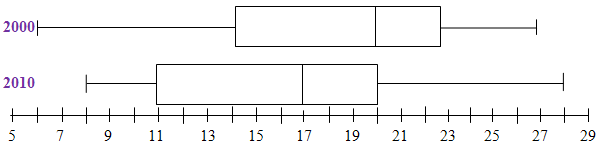
It takes less space compared to histograms and barchart , which is useful when **comparing distributions between many groups or datasets.**

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

**Significance of Box Plot**

Interpret the information given in the following box-and-whisker plot.

The results from a pre-test for students for the year 2000 and the year 2010 are illustrated in the box plot. What do these results tell us about how students performed on the 29 question pre-test for the two years



* If we compare only the lowest and highest scores between the two years, we might conclude that
  + the students in 2010 did better than the students in 2010.
  + since the lowest score of 8 in 2010 is greater in value than the lowest score of 6 in 2000.
  + And the highest score of 28 in 2010 is greater in value than the highest score of 27 in 2000.
* But the box portion of the illustration gives us more detailed information.
  + The middle bar in each box shows us that the median score of 20 in 2000 is greater in value than the median score of 17 in 2010.
  + Further, we note that the box and whiskers divide the illustration into four pieces. Each of these four pieces represents the same portion of students.
  + So, the upper half of the students in 2000 scored in the same score range as the upper one-fourth of the students in 2010, see the illustration at a score of 20.
* By considering the upper one-fourth, upper half, and upper three-fourths instead of just the lowest and highest scores, we would conclude that the students as a whole did much better in 2000 than in 2010. We would conclude that as a whole the students in 2010 are less prepared than the students in 2000.
* The Box Plot (box-and-whisker plots) follows **five-number summary**, which consists of the [**minimum**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S36.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif, [**first quartile**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S36.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif, [**median**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S36.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif, [**third quartile**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S36.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif, and [**maximum**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S36.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif.

# Median (For reference)

In previous sessions, we worked problems involving the [**mean**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S362.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif and [**median**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S362.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif. For this session, we primarily use the median. Here is a brief review of terms used with the median:

* A **data set** is any finite set of real numbers.
* A data set is in **increasing order** if the numbers in the data set are arranged from the least value to greatest value with the least value on the left and the greatest value on the right.
* The **median** of a data set is the number that, when the set is put into increasing order, divides the data into two equal parts.
* If a data set has an **odd** number of data points, then the **median is the** **middle data value** (when the data is in increasing order).
* If a data set has an **even** number of data points, then the **median is the** **mean of the two middle data values** (when the data is in increasing order).

**Example 1:** Find the median of the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.

First, we put the values in the data set into increasing order: 3, 5, 7, 8, 12, 13, 14, 18, 21. Notice that the number of data values is 9, which is odd (there are nine numbers in this data set). Then the middle data value is the 5th value, counting from either the left or the right. Therefore the median is 12.  
(The values 3, 5, 7, and 8 are to the left of 12, and 13, 14, 18, and 21 are to the right of 12.)

**Example 2:** Find the median of the data set {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}.

Note that here we consider the two 14's to be distinct elements and not representing the same item; consider this like you obtained a score of 14 on two different quizzes.  
First, we put the values into increasing order: 3, 5, 7, 8, 12, 14, 14, 15, 18, 21. Notice that there are 10 values, which is even. Then the middle data values are the 5th value from the left and the 5th value from the right. That is, 12 and 14 (the leftmost of the two 14s). Therefore the median is the mean of the two middle values: The median is http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-equation.gifhttp://web.mnstate.edu/peil/MDEV102/U4/S36/lessonimages/equation_image2.gif

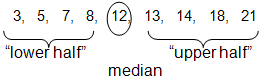
# First Quartile and Third Quartile

**Definitions:**

* The **lower half** of a data set is the set of all values that are to the left of the median value when the data has been put into increasing order.
* The **upper half** of a data set is the set of all values that are to the right of the median value when the data has been put into increasing order.
* The **first quartile,** denoted by ***Q*1** , is the median of the *lower half* of the data set. This means that about 25% of the numbers in the data set lie below *Q*1 and about 75% lie above *Q*1 .
* The **third quartile,** denoted by ***Q*3** , is the median of the *upper half* of the data set. This means that about 75% of the numbers in the data set lie below *Q*3 and about 25% lie above *Q*3 .

**Example 1:** Find the first and third quartiles of the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.

First, we write data in increasing order: 3, 5, 7, 8, 12, 13, 14, 18, 21.



As on the previous page, the median is 12.

Therefore, the lower half of the data is: {3, 5, 7, 8}.

The first quartile, *Q*1, is the median of {3, 5, 7, 8}.

Since there is an even number of values, we need the mean of the middle two values to find the first quartile:

http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-equation.gifhttp://web.mnstate.edu/peil/MDEV102/U4/S36/lessonimages/equation_image4.gif .

Similarly, the upper half of the data is: {13, 14, 18, 21}, so

http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-equation.gifhttp://web.mnstate.edu/peil/MDEV102/U4/S36/lessonimages/equation_image5.gif .

**Example 2:** Find the first and third quartiles of the set {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}.

Note that here we consider the two 14's to be distinct elements and not representing the same item; consider this like you obtained a score of 14 on two different quizzes.

First, we write the data in increasing order: 3, 5, 7, 8, 12, 14, 14, 15, 18, 21.

As before, the median is 13 (it is the mean of 12 and 14 — the pair of middle entries).

Therefore, the lower half of the data is: {3, 5, 7, 8, 12}.

Notice that 12 is included in the lower half since it is below the median value.

Then ***Q*1 = 7** (there are five values in the lower half, so the middle value is the median). Similarly, the upper half of the data is: {14, 14, 15, 18, 21}, so ***Q*3 = 15**.

# Five-Number Summary

**Definitions:**

* The **minimum** **value** of a data set is the least value in the set.
* The **maximum** **value** of a data set is the greatest value in the set.
* The **range** of a data set is the distance between the maximum and minimum value. To compute the range of a data set, we subtract the minimum from the maximum:  
  *range* = *maximum* – *minimum.*
* The **interquartile range** of a data set is the distance between the two quartiles.  
  *Interquartile range* = *Q*3 – *Q*1*.*

**Example 1:** Find the range and interquartile range of the set {3, 7, 8, 5, 12, 14, 21, 13, 18}.

First, we write the data in increasing order: 3, 5, 7, 8, 12, 13, 14, 18, 21.

*range* = *max* – *min* = 21 – 3 = 18.

Recall from the previous page that *Q*1 = 6 and *Q*3 = 16.

Therefore, the *interquartile range = Q*3 – *Q*1 = 16 – 6 = 10.

The range is 18 and the interquartile range is 10.

**Example 2:** Find the range and interquartile range of the set {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}.

First, we write the data in increasing order: 3, 5, 7, 8, 12, 14, 14, 15, 18, 21.

*range* = *max* – *min* = 21 – 3 = 18.

Recall from the previous page that *Q*1 = 7 and *Q*3 = 15.

Therefore, the *interquartile range = Q*3 – *Q*1 = 15 – 7 = 8.

The range is 18 and the interquartile range is 8.

## Self Check Problem

The following dollar amounts were the hourly collections from a Salvation Army kettle at a local store one day in December: $19, $26, $25, $37, $32, $28, $22, $23, $29, $34, $39, and $31. Determine the range and interquartile range for the amount collected.

[**Solution**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S364.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif

The five-number summary from the previous page is

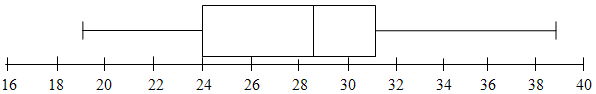
Minimum - 19,

*Q*1- 24,

Median - 28.5,

*Q*2 - 33,

Maximum - 39.



**Definition:** The **five-number summary** of a data set consists of the five numbers determined by computing the **minimum**, ***Q*1** *,* **median***,****Q*3** *,* and **maximum** of the data set.

**Example 1:** Find the five-number summary for the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.

From our Example 1's on the previous pages, we see that the five-number summary is:  
Minimum: 3         *Q*1 : 6            Median: 12            *Q*3 : 16           Maximum: 21

**Example 2:** Find the five-number summary for the data set {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}.

From our Example 2's on the previous pages, we see that the five-number summary is:

Minimum: 3         *Q*1 : 7            Median: 13            *Q*3 : 15           Maximum: 21

## Self Check Problem

The following dollar amounts were the hourly collections from a Salvation Army kettle at a local store one day in December: $19, $26, $25, $37, $32, $28, $22, $23, $29, $34, $39, and $31. Find the five-number summary for the amount collected.

Solution

The results follow from the Self-Check Problems on the previous pages.

Minimum $19,

First Quartile $24,

Median $28.50,

Third Quartile $33,

Maximum $39.

# Box-and-Whisker Plot

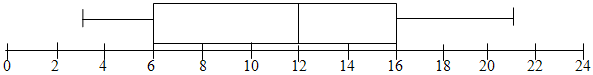
**Definition:** A **box-and-whisker plot** or **boxplot** is a diagram based on the [**five-number summary**](http://web.mnstate.edu/peil/MDEV102/U4/S36/S365.html)http://web.mnstate.edu/peil/MDEV102/U4/S36/ada-annotation.gif of a data set.

* To construct this diagram, we first draw an **equal interval** scale on which to make our box plot. **Do not** just draw a boxplot shape and label points with the numbers from the 5-number summary. The boxplot is a visual representation of the distribution of the data. Greater distances in the diagram should correspond to greater distances between numeric values.
* Using the equal interval scale, we draw a rectangular box with one end at *Q*1 and the other end at *Q*3. And then we draw a vertical segment at the median value. Finally, we draw two horizontal segments on each side of the box, one down to the minimum value and one up to the maximum value, (these segments are called the "**whiskers**").

**Example 1:** Draw a box-and-whisker plot for the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}.

From our Example 1 on the previous page, we had the five-number summary:

Minimum: 3, *Q*1 : 6, Median: 12, *Q*3 : 16, and Maximum: 21.



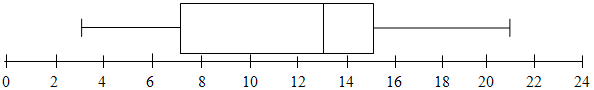
Notice that in any box-and-whisker plot,

* the left-side whisker represents where we find approximately the lowest 25% of the data
* and the right-side whisker represents where we find approximately the highest 25% of the data
* The box part represents the interquartile range and represents approximately the middle 50% of all the data.
* The data is divided into four regions, which each represent approximately 25% of the data. This gives us a nice visual representation of how the data is spread out across the range.

**Example 2:**Draw a box-and-whisker plot for the data set {3, 7, 8, 5, 12, 14, 21, 15, 18, 14}.

From our Example 2 on the previous page, we had the five-number summary:

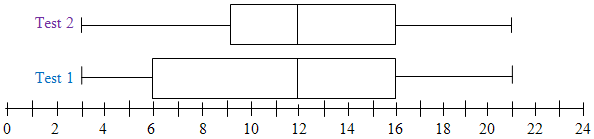
Minimum: 3, *Q*1: 7, Median: 13, *Q*3: 15, and Maximum: 21.



When we relate two data sets based on the same scale, we may examine box-and-whisker plots to get an idea of how the two data sets compare.

**Example 3:** Suppose that the box-and-whisker plots below represent quiz scores out of 25 points for Quiz 1 and Quiz 2 for the same class.

What do these box-and-whisker plots show about how the class did on test #2 compared to test #1?



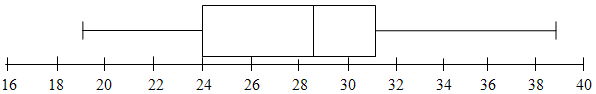
These box-and-whisker plots show that the lowest score, highest score, and *Q*3 are all the same for both exams, so performance on the two exams were quite similar. However, the movement *Q*1 up from a score of 6 to a score of 9 indicates that there was an overall improvement. On the first test, approximately 75% of the students scored at or above a score of 6. On the second test, the same number of students (75%) scored at or above a score of 9.

## Self Check Problem

The following dollar amounts were the hourly collections from a Salvation Army kettle at a local store one day in December: $19, $26, $25, $37, $32, $28, $22, $23, $29, $34, $39, and $31. Construct the box-and-whisker plot for the amount collected.

The five-number summary from the previous page is

Minimum - 19, *Q*1- 24, Median - 28.5, *Q*2 - 33, Maximum - 39.



### ****Characteristics of box plots****

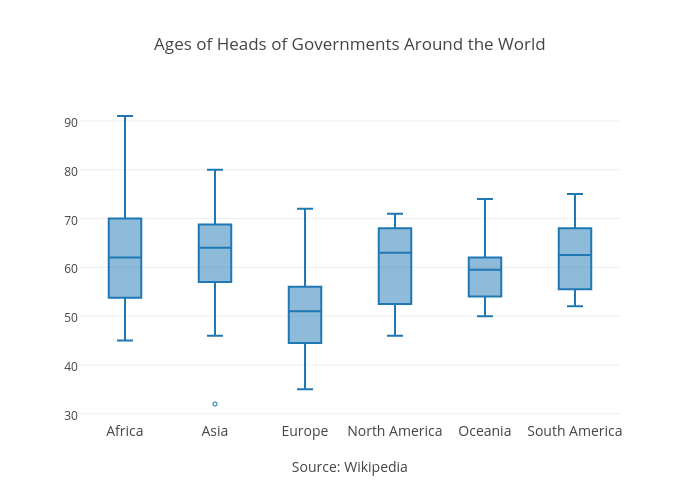
1. **Handles Large Data Easily**Due to the five-number data summary, a box plot can handle and present a summary of a large amount of data. A box plot consists of the median, which is the midpoint of the range of data; the upper and lower quartiles, which represent the numbers above and below the highest 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 too unmanageable for other graphs, such as line plots or stem and leaf plots.
2. **Exact Values Not Retained**The box plot does not keep the exact values and details of the distribution results, which is an issue with handling such large amounts of data in this graph type. A box plot shows only a simple summary of the distribution of results so that you can quickly view it and compare it with other data. Use a box plot in combination with another statistical graph method, like a histogram, for a more thorough, more detailed analysis of the data.
3. **A clear summary**  
   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.
4. **Displays 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. By extending the lesser and greater data values to a max of 1.5 times the inter-quartile range, the box plot delivers outliers or obscure results. Any results of data that fall outside of the minimum and maximum values known as outliers are easy to determine on a box plot graph.

**Application of BOX Plot**

* + If you are given the data for the heads of the governments and you are asked to compare among that of different continents

If you use histogram you need to draw separate histograms for each continent and then compare the graph

Instead the BOX plot representation would be as below from which we can easily compare median og age from every continent and percentage of people below or above the median can also be easily compared.



Using Africa as an example, the median age of government leaders is 62. The two quartiles (that form the box) are 53.75 and 70, which means that half of the data points are found within this range. The “whiskers” show the minimum and maximum of 45 and 91.

* + A box plot is ideal for comparing the distribution of a series of datasets, like this data for each continent.
  + It is often used to track different trials of an experiment that is run many times. If the trials are exactly the same, a box plot will show the consistency of results. If they vary by a parameter that is being tested, a box plot could reveal trends or patterns.
  + Histogram will demonstrate poor performance when there are few samples or when the boxes are the wrong sizes.
  + box-plots provide some information that the histogram does not explicitly, that is, median, 25th and 75th percentile, min/max that is not an outlier and explicitly separates the points that are considered outliers.

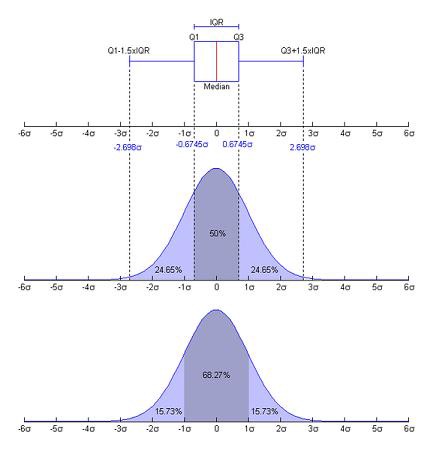
<https://dimensionless.in/what-is-a-box-plot/>

**Managing Outliers in BOX Plot**

**Deciding Minimum and Maximum considering Outliers**

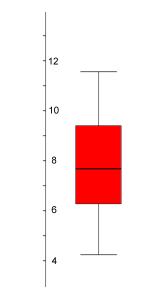
1. **Highest Value (Maximum)**This point in the box plot represents the highest value in the data distribution over which the box plot is built which is not an outlier. This point does not correspond to the Maximum value in your dataset. Suppose you have some data like 65,76,87,100,105,100000. Here the Maximun value is 100000 but it is most likely to be an outlier and hence the box plot will not mark this as the maximum value. The most feasible option will be 105 as the maximum value of the box plot.
2. **Lowest Value (Minimum)**This point in the box plot represents the lowest value in the data distribution over which the box plot is built and is not an outlier (smallest value in the Interquartile range of the distribution). This point does not correspond to the smallest value in your dataset. Suppose you have some data like 0.005,65,76,87,100,105. Here the smallest value is 0.005 but it is most likely to be an outlier and hence the box plot will not mark this as the minimum value. The most feasible option will be 65 as the minimum value of the box plot.

**In this case Boxplot statistic would be based on the updated Maximum and Minimum values**



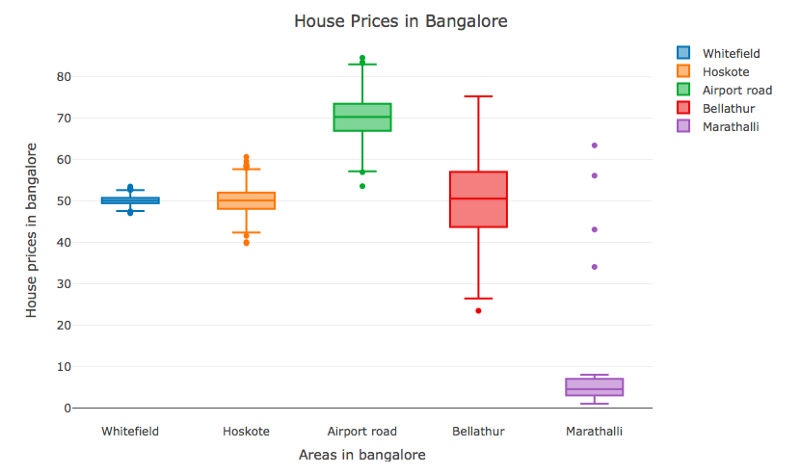
### ****Understanding different box plots****

Although boxplots can be drawn in any orientation, most statistical packages seem to produce them vertically by default, as shown below, rather than horizontally. The length of the box becomes its height. The width across the page signifies nothing.



Boxplots of several samples are lined up alongside one another. The box length gives an indication of the sample variability and the line across the box shows where the sample is centred. The position of the box in its whiskers and the position of the line in the box also tells us whether the sample is symmetric or skewed, either to the right or left. For a symmetric distribution, long whiskers, relative to the box length, can betray a [heavy tailed](http://web.pdx.edu/~stipakb/download/PA551/boxplot.html#TailLength) population and short whiskers, a short tailed population. So, provided the number of points in the sample is not too small, the boxplot also gives us some idea of the "shape" of the sample, and by implication, the shape of the population from which it was drawn. This is all important when considering appropriate analyses of the data.

Data on different house prices in 5 different areas of Bangalore is plotted. Let us understand the distribution of this data and try to find some insights out of it.



* **The Box plot as an Indicator of Centrality**We will try to gather our first insight by observing the centrality of the box plots. Centerline represents the median value for the house price in different areas. Houses on airport road have the highest median value of the house which makes it a comparatively expensive place to live in whereas houses in Marathali have the least median value which allows us to conclude that houses here are relatively cheapest to live.

|  |  |  |
| --- | --- | --- |
| http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b17.jpg |  | http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b18.jpg |
|  | | |
| **The boxplot of a sample of 20 points from a population centered on 7.** |  | **The boxplot of a sample of 20 points from a population centered on 12.** |

* **The Box plot as an indicator of the spread**The spread of a box plot talks about the variance present in the data. More the spread, more the variance. If you look closely at the first two box plots, both Whitefield and Hoskote areas have the same median house price value so it seems like both places fall into the same budget category. But if we look more closely, we can observe that width of Hoskote box plot is more than Whitefield box plot. Hoskote area has more variance in house price as compared to Whitefield i.e. Hoskote offers more variety of budget in houses as compared to Whitefield. If we look at the overall graph, we find that Bellathur area has the most spread in its box plot. This clearly states that this area has the widest variety in the budget of the houses.

|  |  |  |
| --- | --- | --- |
| http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b20.jpg |  | http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b19.jpg |
|  | | |
| **The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 1.** |  | **The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 3.** |

* **The Box plot as an indicator of symmetry**Symmetry around the median talks about skewness present in the data. If the median line is towards the lower half of the box plot, then it is right skewed (positive skew) and if the median line is towards the upper portion of the box plot then it is left-skewed (negative skew). If we look at the box plot representing Marathalli, we can observe that median is towards the lower half of the box plot and hence it is right skewed (positive skew) which means that most of the houses are on the cheaper side in Marathalli and only a few are expensive.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b1.jpg |  | http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b11.jpg |  | http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b9.jpg |
|  | | | | |
| **The boxplot of a sample of 20 points from a symmetric population. The line is close to the centre of the box and the whisker lengths are the same.** |  | **The boxplot of a sample of 20 points from a population which is skewed to the right. The top whisker is much longer than the bottom whisker and the line is gravitating towards the bottom of the box.** |  | **The boxplot of a sample of 20 points from a population which is skewed to the left. The bottom whisker is much longer than the top whisker and the line is rising to the top of the box.** |

* **The Box plot as an indicator of tail length**Tail length talks about the kurtosis present in data. There are three cases here. Either your data will be normally distributed or it will have more data in its tail as compared to a normal distribution(platykurtic) or it will have fewer data in tails as compared to a normal distribution(leptokuritc). A long tail shows that the distribution is platykurtic and shorter tail gives the idea of distribution being leptokurtic. In above example, Marathalli has the shortest tail as compared to other box plots which may mean that in Marathalli most of the house prices lie in the interquartile range (q3-q1).
* The tails are the extremities of the sample or population, rather than the centre. Lack of symmetry entails one tail being longer than the other. Populations are usually referred to as being heavy-tailed or light-tailed, or the Greek equivalent, leptokurtic (slender arched) or platykurtic (flat arched). The ideal level of kurtosis, neither too heavy or too light, is represented by the Normal population - the bell shaped curve. The box-plot of a sample from a Normal population should exhibit whiskers about the same length as the box, or perhaps marginally longer. The symmetric example [above](http://web.pdx.edu/~stipakb/download/PA551/boxplot.html#Symmetry) is from a Normal population.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b00.jpg |  | http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b5.jpg |  | http://web.pdx.edu/~stipakb/download/PA551/boxplot_files/b8.jpg |
|  | | | | |
| **The boxplot of a sample of 20 points from a population with long tails. The length of the whiskers far exceeds the length of the box. (A well proportioned tail would give rise to whiskers about the same length as the box, or maybe slightly longer.)** |  | **The boxplot of a sample of 20 points from a population with short tails. The length of the whiskers is shorter than the length of the box.** |  | **The boxplot of a sample of 20 points from a population with extremely short tails (actually a U-shaped population, with a dip in the middle rather than a hump). The whiskers are absent.** |