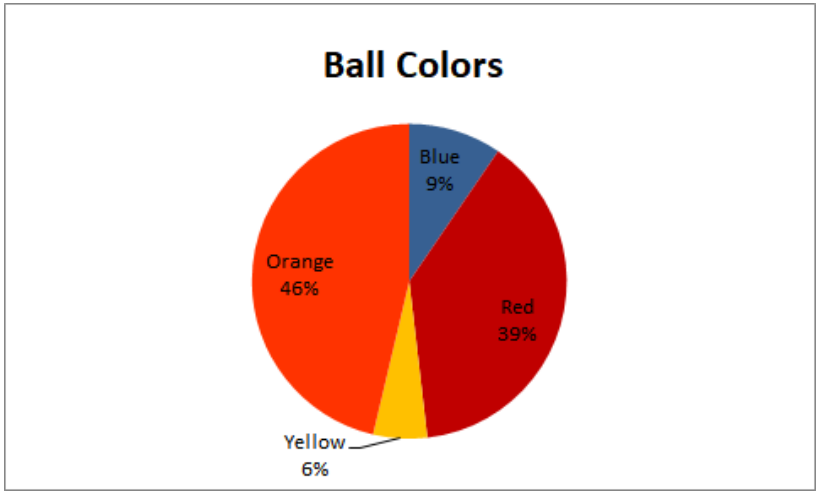
Basic Data Representation Mechanisms

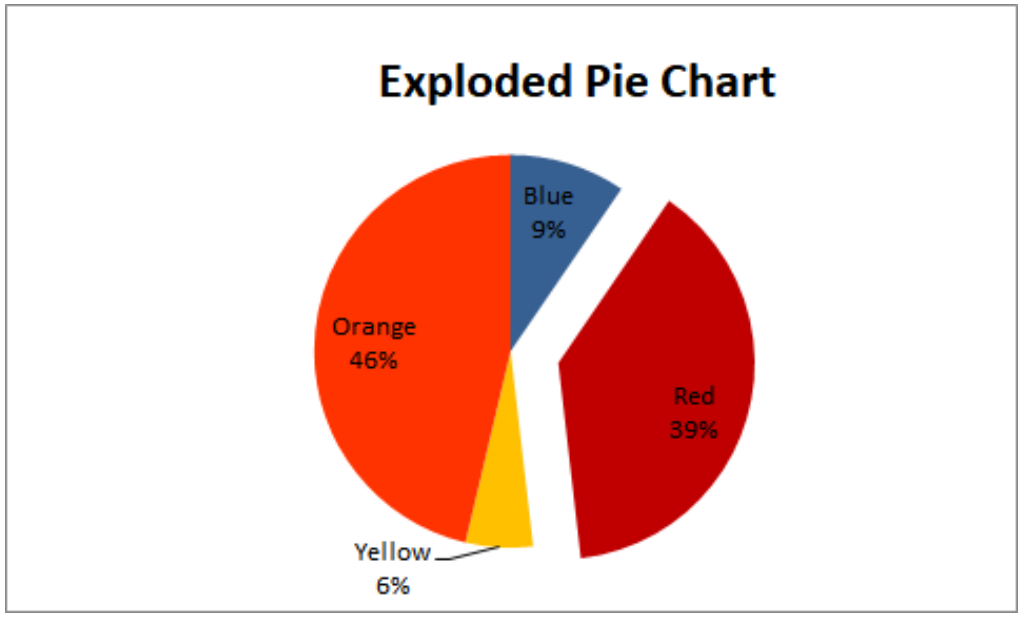
A pie chart - a circular representation divided into various sectors - illustrate numerical proportions

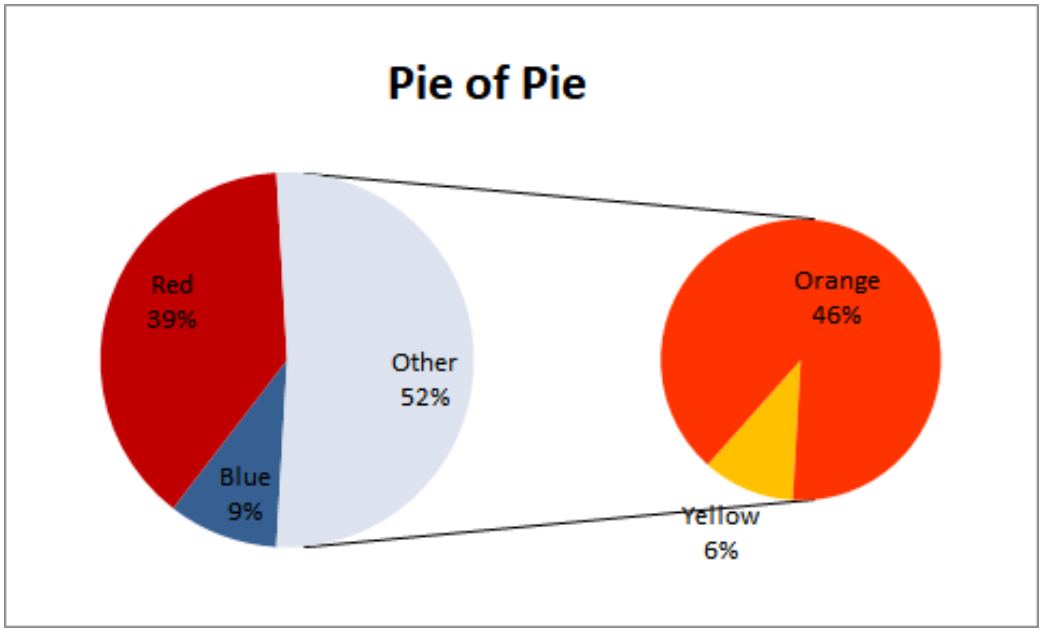
entire pie represents 100% of the dataset, each sector/pie slices represents a portion of the whole dataset.

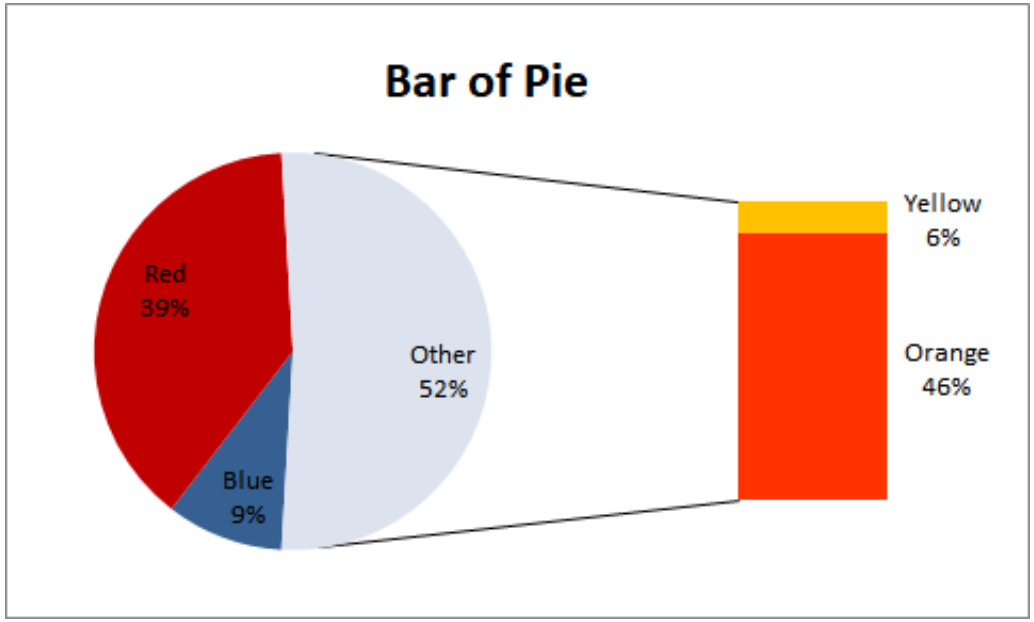
The portions that exist in the pie are measured as an angle of the total 360 degrees.

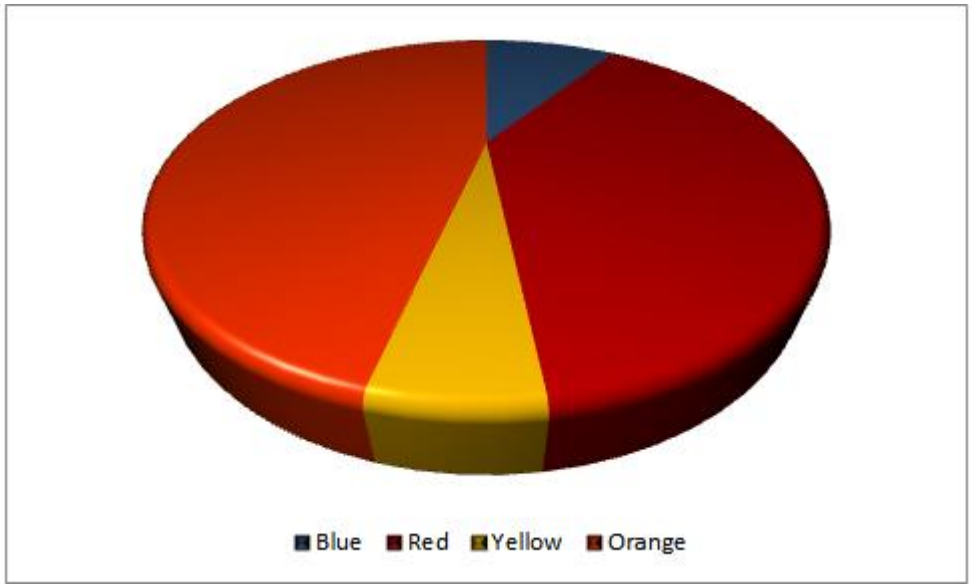


Colors	Frequency
Blue	112
Red	454
Yellow	65
Orange	544


Exploded Pie Chart
Blue
9%
Orange
46%
Red
39%
Yellow
6%


Pie of Pie
Red
39%
Other
52%
Blue
9%
Orange
46%
Yellow
6%


Bar of Pie
Red
39%
Blue
9%
Other
52%
Yellow
6%
Orange
46%


Blue
Red
Yellow
Orange

- Percentage of each value =(value/Total value)/*100
- Degree of each value =(Value/Total value)*360
- Total values = 112+454+65+544 = 1175

- Blue: (112/1175)100 = 9.53%
- Red: (454/1175)100 = 38.64%
- Yellow: (65/1175)100 = 5.53%
- Orange: (544/1175)100 = 46.30%

- Blue: (112/1175)360 = 34.3
- Red: (454/1175)360 = 139.1
- Yellow: (65/1175)360 = 19.9
- Orange: (544/1175)360 = 166.7

- **Advantages**
- A simple and easy-to-understand picture.
- It represents data visually as a fractional part of a whole, which can be an effective communication tool for the even uninformed audience.
- display relative proportions of multiple classes of data.
- require minimum addition explanations
- summarize a large data set into visual form.
- **Disadvantages**
- it does not easily reveal exact values. Values are expressed in terms of percentages or ratio therefore it is not easy to know the exact value represented
- the pie chart does not easily show changes over time
- pie charts fail to reveal key assumptions, causes, effects, or patterns
- pie charts can easily be manipulated to yield false impressions

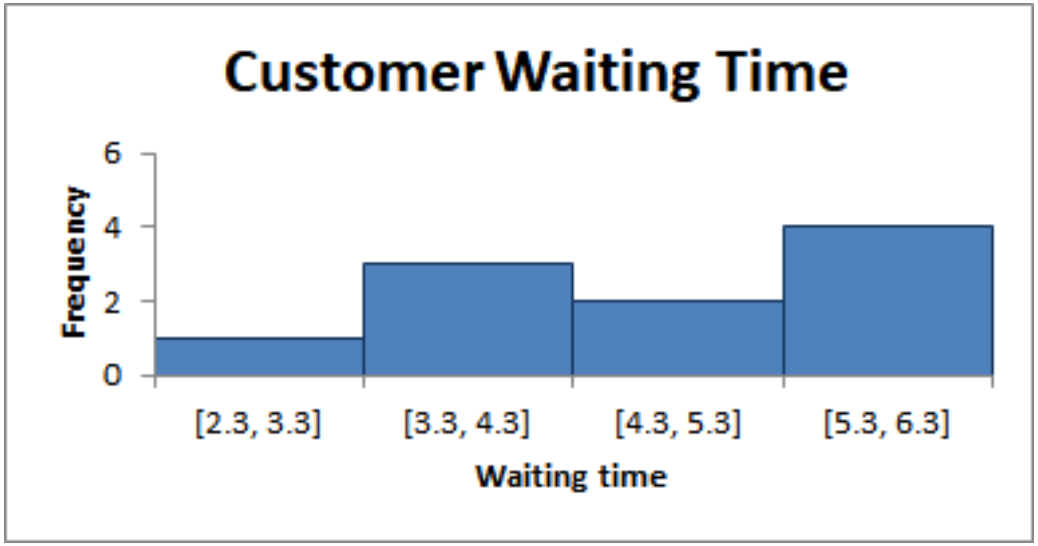
•A histogram - summarizes discrete or continuous data.

• visual interpretation of numerical data by showing the number of data points that fall within a specified range of values (called "bins").

•similar to a vertical bar graph.

•unlike a vertical bar graph, shows no gaps between the bars.

Customer Waiting Time (Minutes)
2.3
5.67
3.43
2.5
5.67
4.65
2.76
4.1
4.67
2.56



- **Components**
- **The title:** information included in the histogram.
- **X-axis:** scale of values which the measurements fall under.
- **Y-axis:** number of times that the values occurred within the intervals set by the X-axis.
- **The bars:** The height of the bar shows the number of times that the values occurred within the interval, while the width of the bar shows the interval that is covered.
- For a histogram with equal bins, the width should be the same across all bars.
- **Advantages**
- visual representation of data distribution
- can display a large amount of data and the frequency of the data values.
- Depicts median and distribution of the data and outliers or gaps in the data.

Normal Distribution

Bimodal

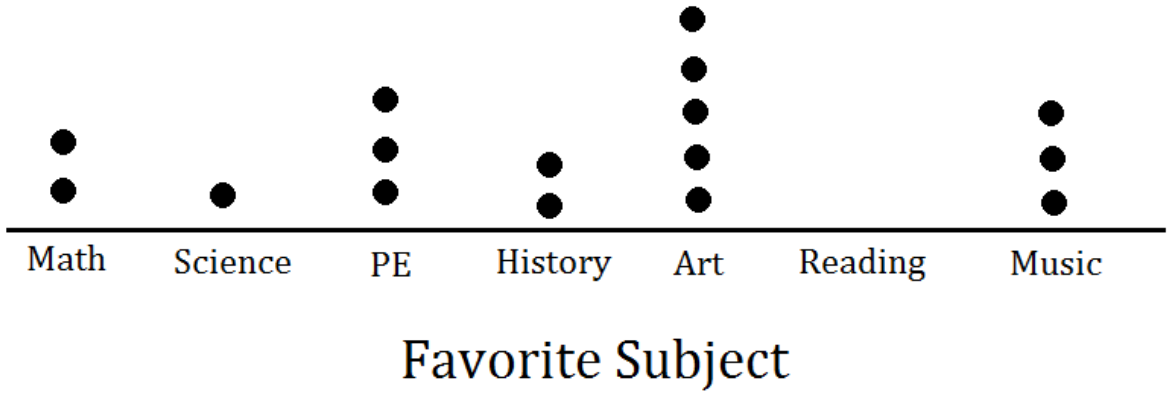
Right Skewed

Left Skewed

- +1 frequency of the data occurring in the dataset. and categories which are difficult to interpret in a tabular form.
- +2 helps to visualize the distribution of the data.
- -1 Cannot read exact values because data is grouped into categories.
- -2 More difficult to compare two data sets.
- -3 Use only with continuous data.

Dot Plot

A dot plot is a graphic display using dots and a simple scale to compare the frequency within categories or groups.



- suitable for small to moderate sized data sets
- conservation of numerical information
- larger data sets (around 20–30 or more data points) the related stemplot, box plot or histogram may be more efficient, as dot plots may become too cluttered

Stem Leaf Plots

7.6, 8.1, 9.2, 6.8, 5.9, 6.2, 6.1, 5.8, 7.3, 8.1, 8.8, 7.4, 7.7, 8.2

task completion times

stem	leaf
5	8 9
6	1 2 8
7	3 4 6 7
8	1 1 2 8
9	2

key: "9 | 2" means "9.2"

- Economics 9, 13, 14, 15, 16, 16, 17, 19, 20, 21, 21, 22, 25, 25, 26
- Libertarianism: 14, 16, 17, 18, 18, 20, 20, 24, 29

class sizes

Econ 101	stem	Pol 306
9	0	
3 4 5 6 6 7 9	1	4 6 7 8 8
0 1 1 2 5 5 6	2	0 0 4 9

key: "2 | 0" means "20"

- 104, 107, 112, 115, 115, 116, 123, 130, 134, 145, 147

Stems	Leafs
10	4 7
11	2 5 5 6
12	3
13	0 4
14	5 7

M

Decimal Between Stem and Leaf	Decimal in the Stem
12.3, 12.5, 13.0	1.23, 1.25, 1.30
Becomes	Becomes
12 | 3, 5 13 | 0	1.2 | 3, 5 1.3 | 0
Key: 12 | 3 = 12.3 units	Key: 1.2 | 3 = 1.23 units

www.LearnAlgebraFaster.com

a) How many students scored higher than 60 in section 1?

b) How many students scored higher than 60 in section 2?

c) What are the minimum and maximum scores in section 1?

d) What are the minimum and maximum scores in section 2?

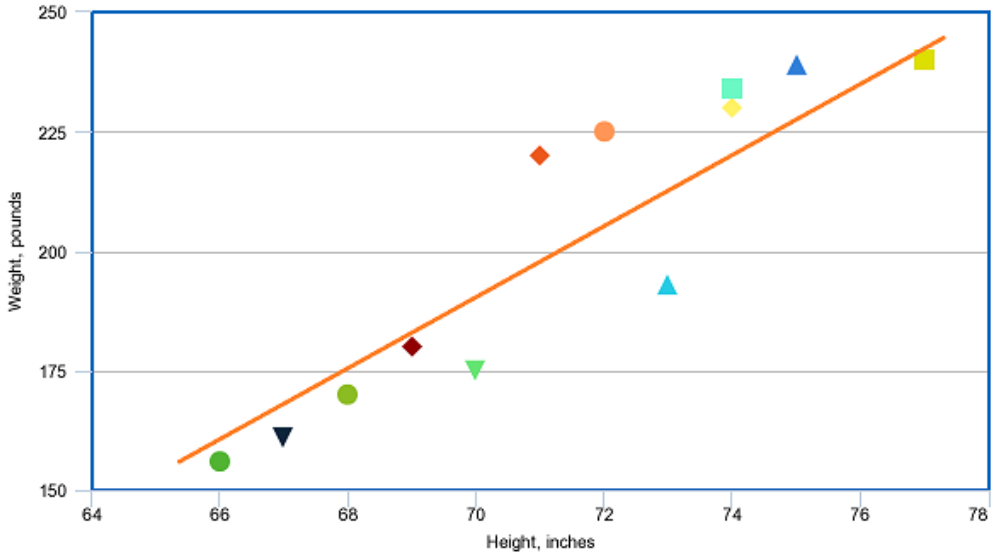
e) Without counting, which section has more students scoring 80 or more?

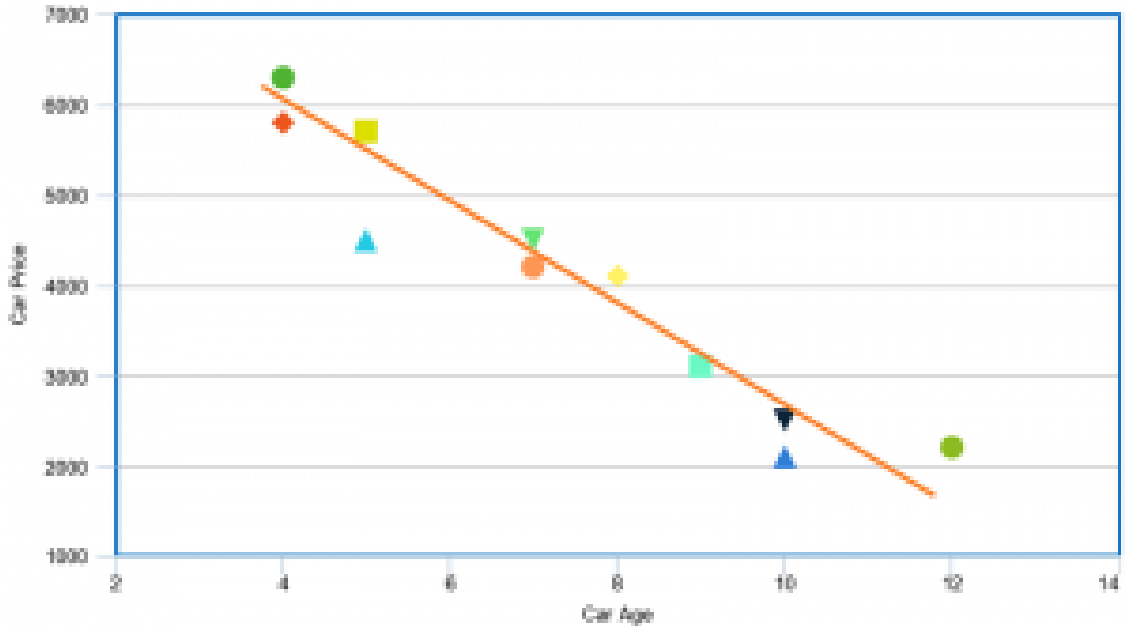
f) Without counting, which section has more students scoring 50 or less?

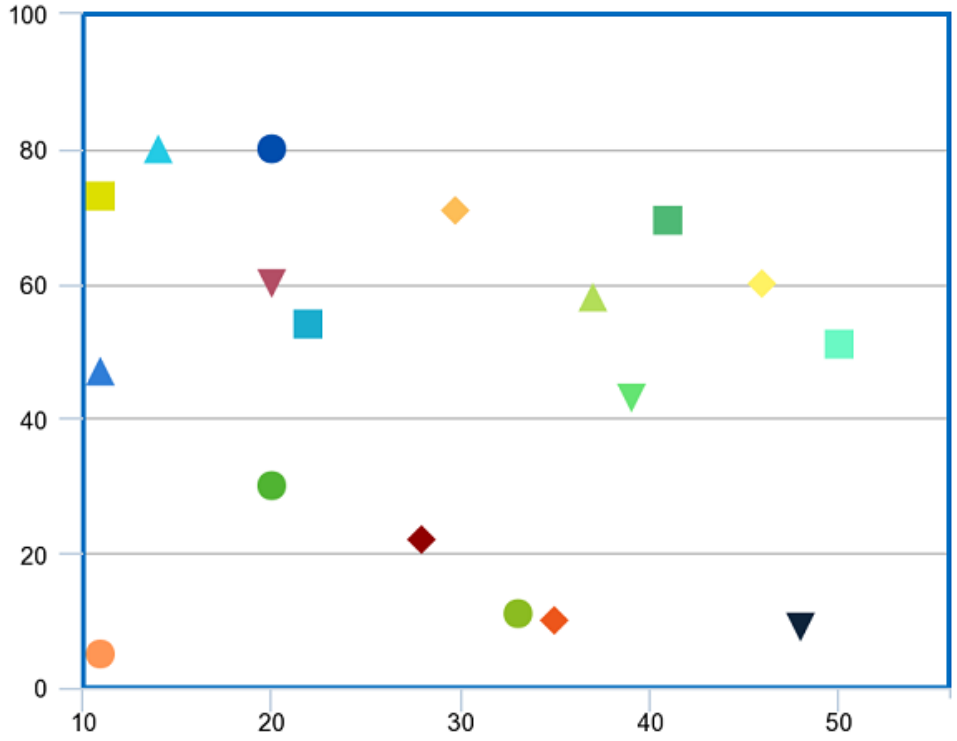
Section 1							Section 2						
				5	0	**4**	1	3	4	5			
			3	3	2	1	**5**	3	4	5	5	7	9
	8	6	5	4	3	1	**6**	1	2	2	5	6	6	7	9
9	7	6	3	1	0	0	**7**	0	3	4	6	8	9		
		7	4	3	2	1	**8**	1	6						
			5	3	2	0	**9**	0	1						

+1. It allows you to see the shape of the distribution.
+ 2. It gives the specific data values.
+ 3. It allows you to determine strong outliers.
-. Impractical for large data sets

- Scatter Plot - pairs of numerical data, with one variable on each axis, to look for a relationship between them.
- If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line.
- y variable tends to increase as the x variable increases, **positive correlation** .
- y variable tends to decrease as the x variable increases, **negative correlation**
- no clear relationship between the two variables **no correlation** between the two variables.




7000
6000
5000
4000
3000
2000
1000
Car Price
2
4
6
8
10
12
14
Car Age


100
80
60
40
20
0
10
20
30
40
50

- +'s:
- Show relationship and a trend in the data relationship.
- Show all data points, including minimum and maximum and outliers.
- Can highlight correlations.
- Retains the exact data values and sample size.
- Shows both positive and negative type of graphical correlation

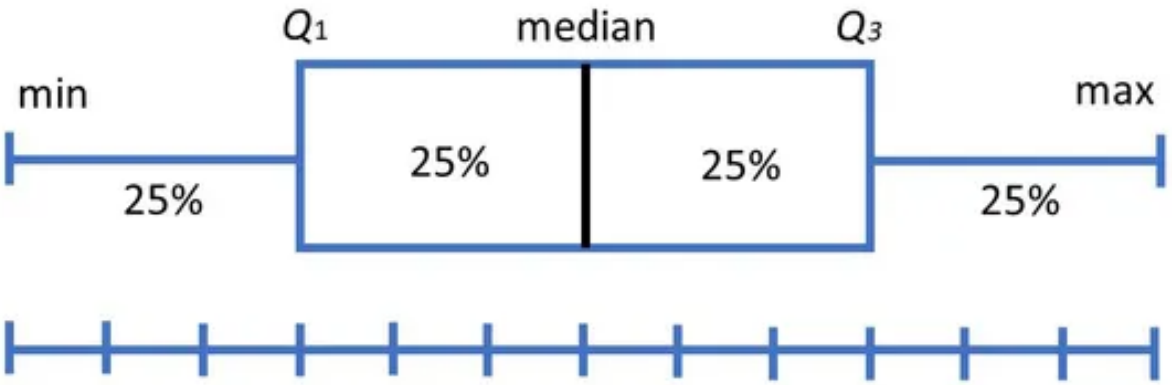
-'s:

- Flat best-fit line gives inconclusive results.
- Interpretation can be subjective.
- Correlation does not mean and not show causation.
- Interpretations becomes difficult with increased dimension
- Scatter plots are unable to give the exact extent of correlation.
- Scatter plot doesn't show the quantitative measure of the relationship between the two variables.

Box Plots / Box Whisker Plots

- A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum

Box plots divide the data into sections that each contain approximately 25% of the data in that set.

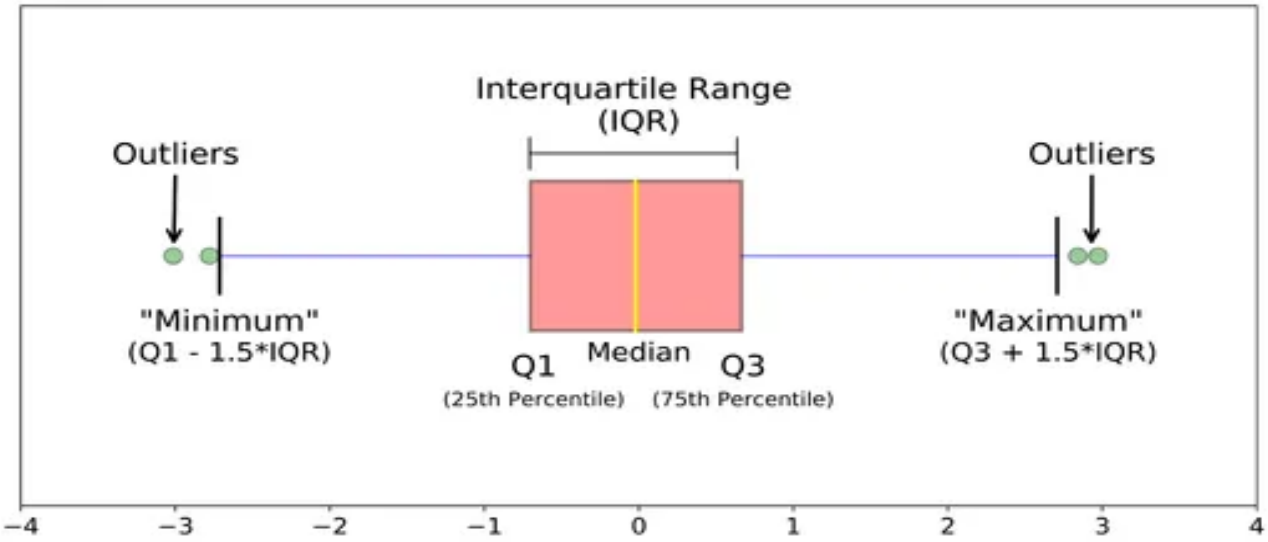


Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.

Box plots are useful as they show outliers within a data set.

An outlier is an observation that is numerically distant from the rest of the data.

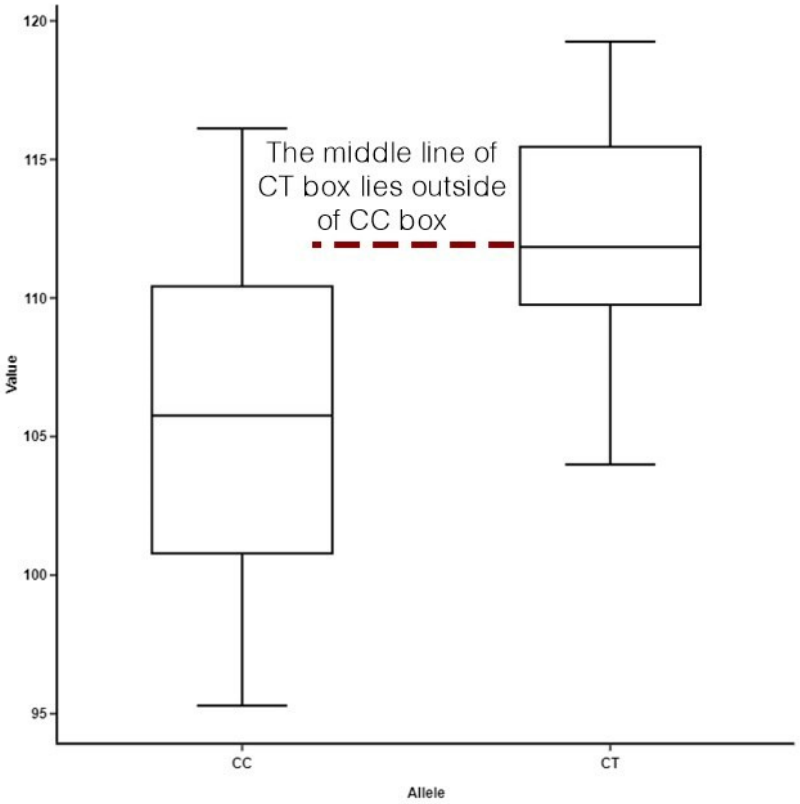
When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

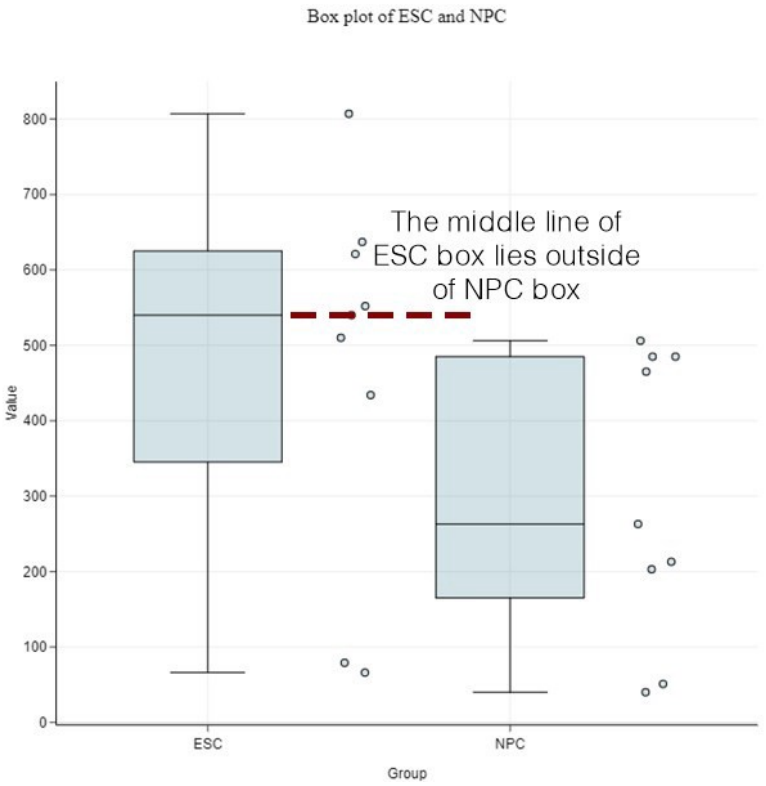


Source: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

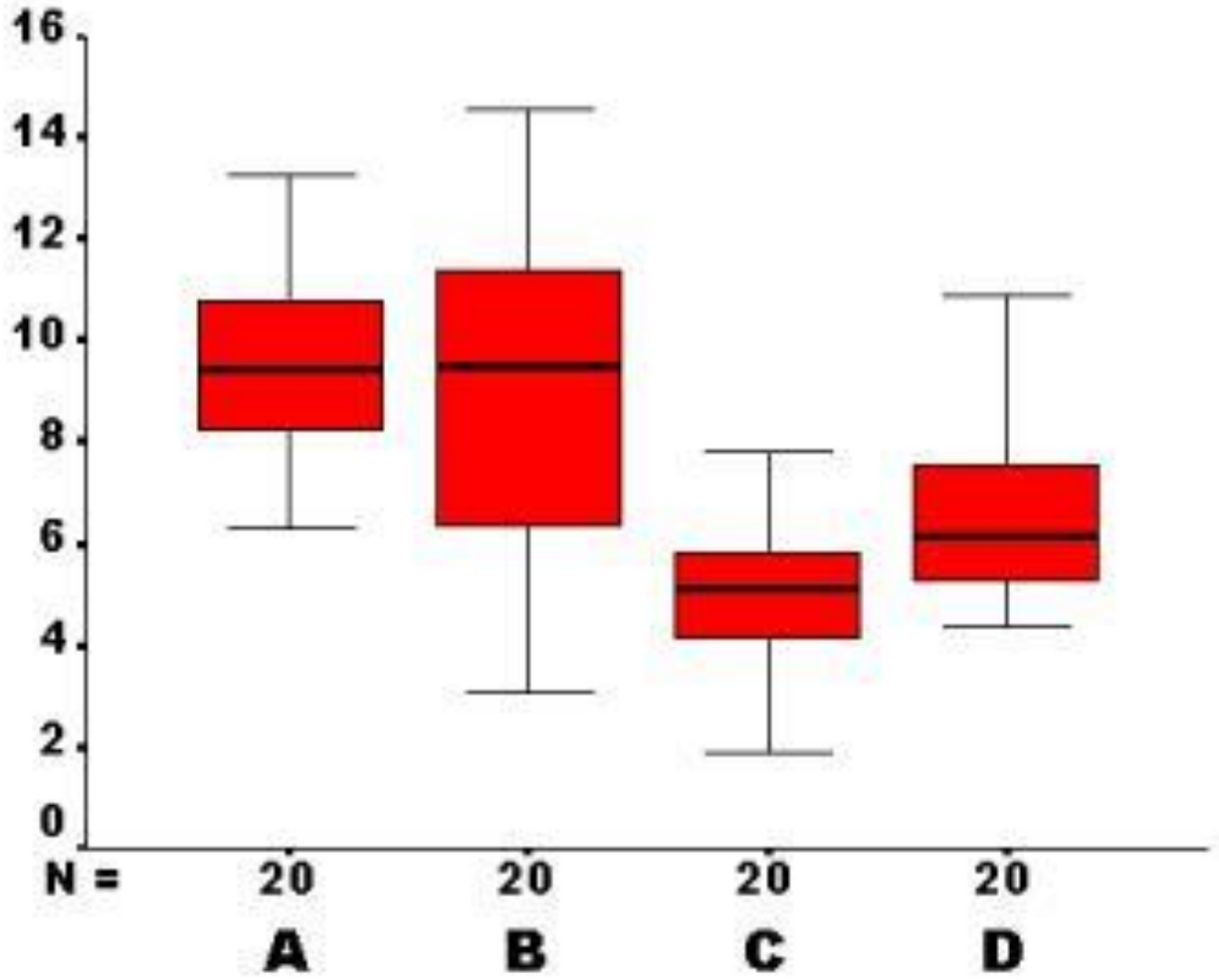
For example, outside 1.5 times the interquartile range above the upper quartile and below the lower quartile (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR).

- *Non-overlapping boxes, groups are different.*
- If they overlap, move on to the lines inside the boxes.
- *Boxes overlap but don't spread past both medians: groups are likely to be different.*



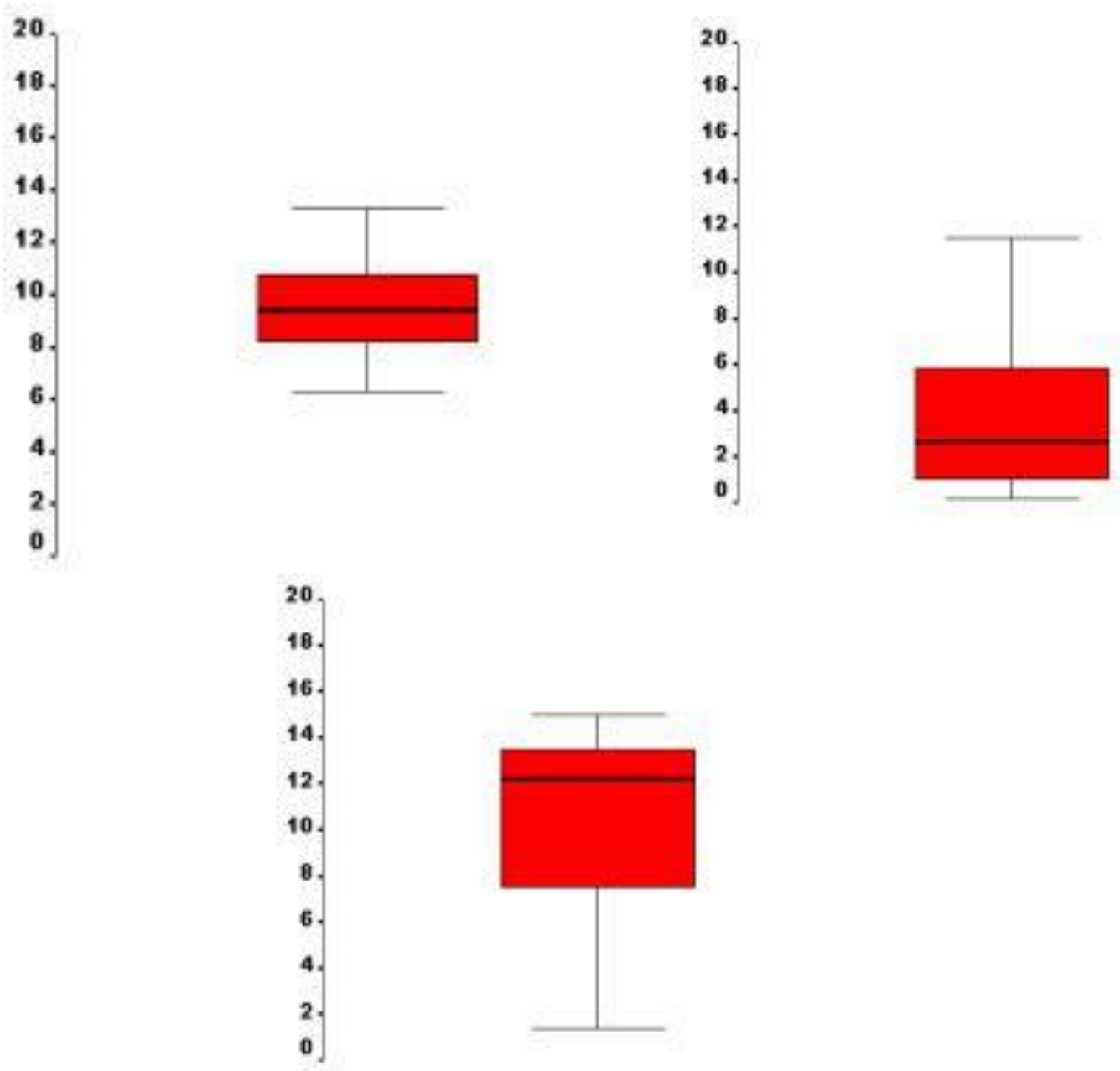


- If both median lines lie within the overlap between two boxes, we will have to take another step to reach a conclusion about their groups.
- Short boxes mean their data points consistently hover around the center values. Taller boxes imply more variable data. That's something to look for when comparing box plots, especially when the medians are similar.
- Larger ranges indicate wider distribution, that is, more scattered data.

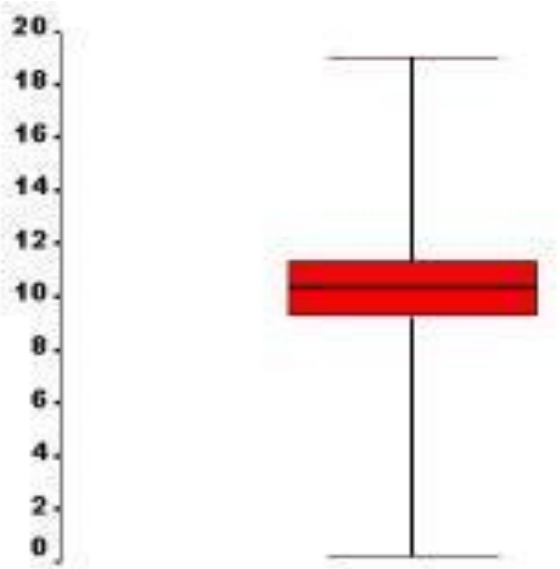


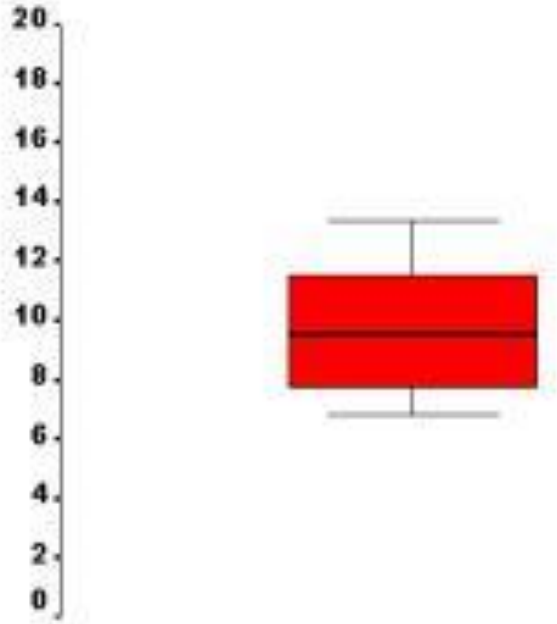
sample A and B appear to have similar centres, which exceed those of C and D. Sample B appears to have larger variability than the other three samples. Samples A, B and C are reasonably symmetric, but sample D is skewed to the right.

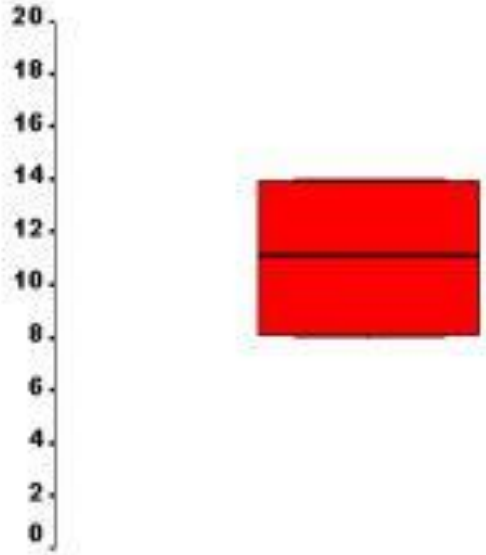
Boxplot as an Indicator of Symmetry



Boxplot as an Indicator of Tail Length







whiskers about the same length as the box, or maybe slightly longer.

length of the whiskers is shorter than the length of the box.

extremely short tails (actually a U-shaped population, with a dip in the middle rather than a hump). The whiskers are absent.

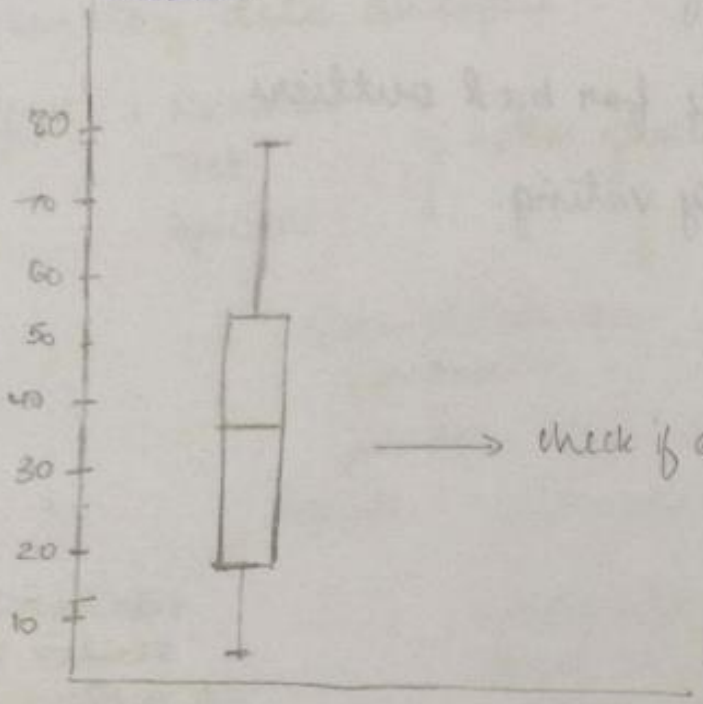
Consider 6, 15, 19, 21, 24, 38, 47, 55, 56, 65, 78

[6, 15, 19, 21, 24] median [47, 55, 56, 65, 78]

38 → median Q_2

[6, 15, 19, 21, 24] → Q_1 lower quartile median

[47, 55, 56, 65, 78] → Q_3 upper quartile median



→ check if evenly spread out / quartiles too close to whiskers

* Box plot is recommended for sets to compare related data and gain inferences

Outliers: $x < Q_1 - 1.5 * IQR \Rightarrow x$ is an outlier

$x > Q_3 + 1.5 * IQR$

Example: [1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 10]

Q_2 = avg (4, 5)

Q_2

for Q_1, consider 1 - 4 $\Rightarrow$ $Q_1 = 2$

Q_3, consider, 5 - 10 $\Rightarrow$ $Q_3 = 7$

Example:

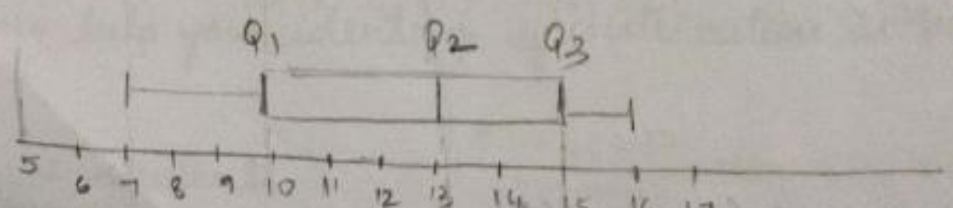
[2, 43, 49, 50, 51] [51] [53, 54, 60, 62, 63]

Q_2

$Q_1 - 1.5\,IQR = 32.5$ $\Rightarrow$ 2 is an outlier

$Q_3 + 1.5\,IQR = 76.5$

Example:



* You cannot reconstruct data given a box plot.

- Given the above box plot; comment on the following inferences;
- (i) All Data are Les than 17 - Yes
- (ii) Only 1 data point is 7 or 16 (cant say)
- (iii) Exactly 50% of data are > 13 – cant say
- (iv) 75% of data are >= 10 :
- 7 data points:

7 - - 13 - - 16 ; from the box Q1 and Q2 = 10,15

Hence the first blank in Lower Quartile shud be 10 and the last blank in the upper quartile shud be 15.

7 10 - 13 - 15 16 : concludes that 6 out of 7 points are greater than or equal to 10…not true in this case…

7 9 11 12 14 15 15 16 is the even scenario

10 13 15 ; 6/8 >=10 ; True in this case.