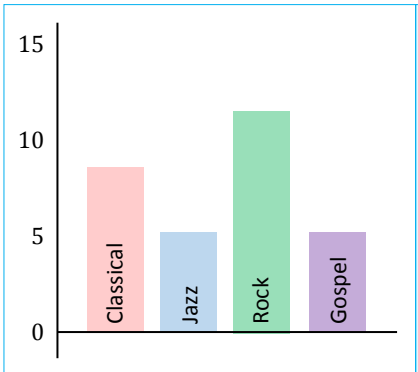
***Histograms***

Consider the following data:

30 people were asked which music genre they listen to most:

| Classical | Jazz | Rock | Gospel |
|---|---|---|---|
| 8 | 5 | 12 | 5 |

We can represent this data as a bar chart:




Bar charts are appropriate for representing ***discrete*** categories, as is the case here.

The ***height*** of each bar represents the size (frequency) of each category.

The bars are drawn with spaces between.

Notice that in a bar chart, only the $y$ axis has a numerical scale.

***Histograms***

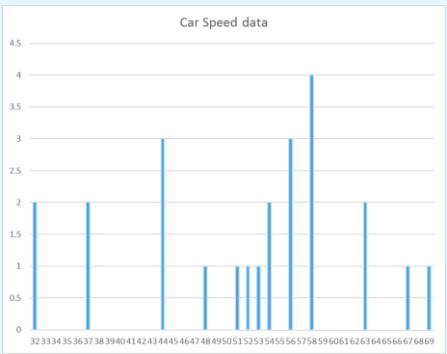
Here again is the car speed data:

$32, 32, 37, 37, 44, 44,\quad 44, 48, 51, 52, 53, 54,$

$54, 56, 56, 56, 58, 58,\quad 58, 58, 63, 63, 67, 69.$

We could treat each speed as a separate Category and draw a bar chart:

But doing so is already cumbersome even for this small numerical data set.




Notice again, only the $y$ axis has a continuous scale; the $x$ axis represents ***discrete*** categories.

For data such as this, a ***histogram*** is more appropriate.

A ***histogram*** is a bar chart used to represent ***continuous data***.

There are no gaps between the bars representing the categories.

In a histogram ***both*** axes have a continuous scale.

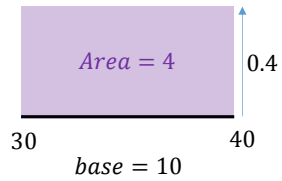
For this reason, the ***area*** of each bar represents frequency.

***Histograms***

To plot a histogram of the car speed data, we will use the grouped data table shown below.

Remember, the ***area*** of each bar will represent the ***frequency***. $Area = base \times height$

So, for the interval $30 \longrightarrow 40$:

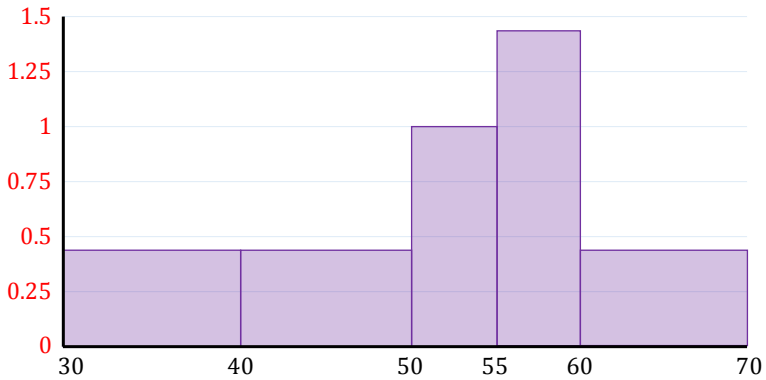



so $height = \dfrac{Area}{base} = \dfrac{4}{10} = 0.4$

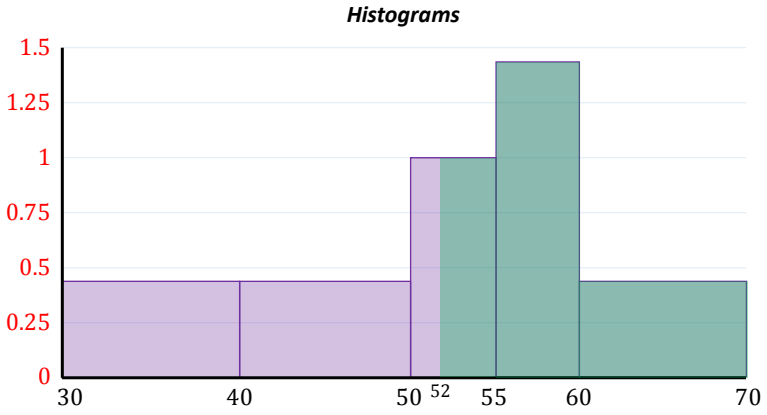
The height of the bar is known as ***frequency density***

| | $f$ (frequency) $= bar\ area$ | Interval base width | Bar height |
|---|---|---|---|
| $30 \rightarrow 40$ | 4 | 10 | $4 \div 10 = 0.4$ |
| $40 \rightarrow 50$ | 4 | 10 | $4 \div 10 = 0.4$ |
| $50 \rightarrow 55$ | 5 | 5 | $5 \div 5 = 1$ |
| $55 \rightarrow 60$ | 7 | 5 | $7 \div 5 = 1.4$ |
| $60 \rightarrow 70$ | 4 | 10 | $4 \div 10 = 0.4$ |

***Histograms***




| | $f$ (frequency) $= bar\ area$ | Interval base width | Bar height |
|---|---|---|---|
| $30 \rightarrow 40$ | 4 | 10 | $4 \div 10 = 0.4$ |
| $40 \rightarrow 50$ | 4 | 10 | $4 \div 10 = 0.4$ |
| $50 \rightarrow 55$ | 5 | 5 | $5 \div 5 = 1$ |
| $55 \rightarrow 60$ | 7 | 5 | $7 \div 5 = 1.4$ |
| $60 \rightarrow 70$ | 4 | 10 | $4 \div 10 = 0.4$ |




***Example*** Use the histogram to estimate the number of car with a speed over $52mph$.

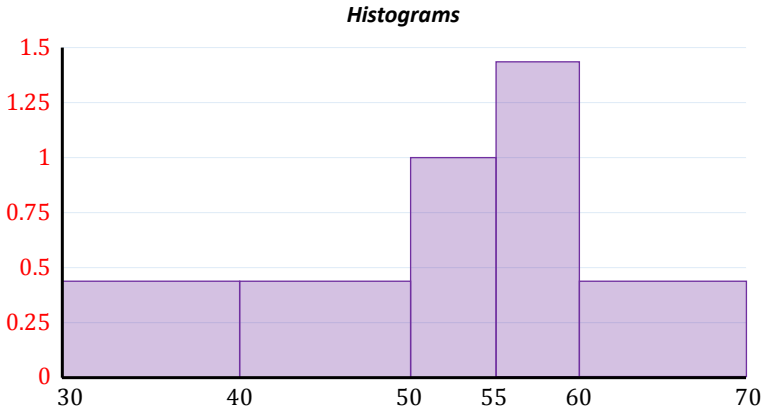
Remember, in a histogram, the ***area*** represents frequency.

So we need to work out the size of the green area shown above.

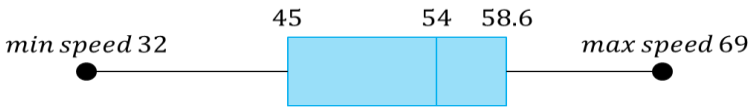
We do this using the simple $area = base \times height$ formula for rectangles.

Note that the green rectangle between 50 and 55 is 3/5 of the purple rectangle beneath it.

$$\frac{3}{5} \times (5 \times 1) \; + \; (5 \times 1.4) \; + \; (10 \times 0.4) \; = 14 \text{ cars with a speed} > 52.$$

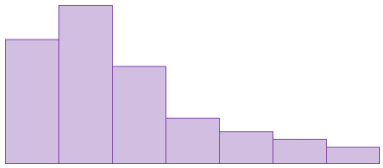



Compare the histogram above with the box-whisker plot for the same data:

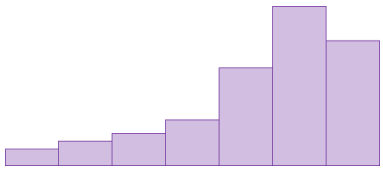



The data is negatively skewed; notice how this can be seen also in the histogram.

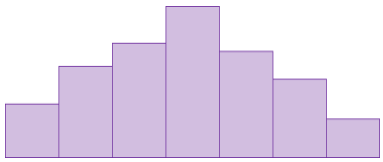
***Histograms***



***Positively skewed data***



***Negatively skewed data***



***Symmetric data***