FREQUENCY TABLE

A **frequency table** is a table with two columns.
One column lists the *categories*, and another for the *frequencies* with which the items in the categories occur (how many items fit into each category).

Example

A local movie theater chain wants to understand which genres of movies are most popular among their customers to guide future movie screenings. They collect data on ticket sales for different genres over the past three months. The data is summarized in the following frequency table:

Example

A local movie theater chain wants to understand which genres of movies are most popular among their customers to guide future movie screenings. They collect data on ticket sales for different genres over the past three months. The data is summarized in the following frequency table:

Genre	Frequency (Tickets Sold)
Action	102
Comedy	85
Drama	76
Horror	58
Sci-Fi	64
Romance	43

Example

A hospital is analyzing the ages of patients admitted for routine check-ups in the past month to better understand the age demographics of their visitors. The ages of 30 patients are recorded below:

24 26 29 31 33 36
36 38 38 42 45 47
48 50 52 54 56 59
60 61 63 65 65 67
70 73 75 77 79 82

Complete the frequency distribution for the data:

Example

A hospital is analyzing the ages of patients admitted for routine check-ups in the past month to better understand the age demographics of their visitors.

24 26 29 31 33 36
36 38 38 42 45 47
48 50 52 54 56 59
60 61 63 65 65 67
70 73 75 77 79 82

Age Range	Frequency
20-29	
30-39	
40-49	
50-59	
60-69	
70-79	
80-89	

BAR GRAPH

A **bar graph** is a graph that displays a bar for each category with the length of each bar indicating the frequency of that category.

Example

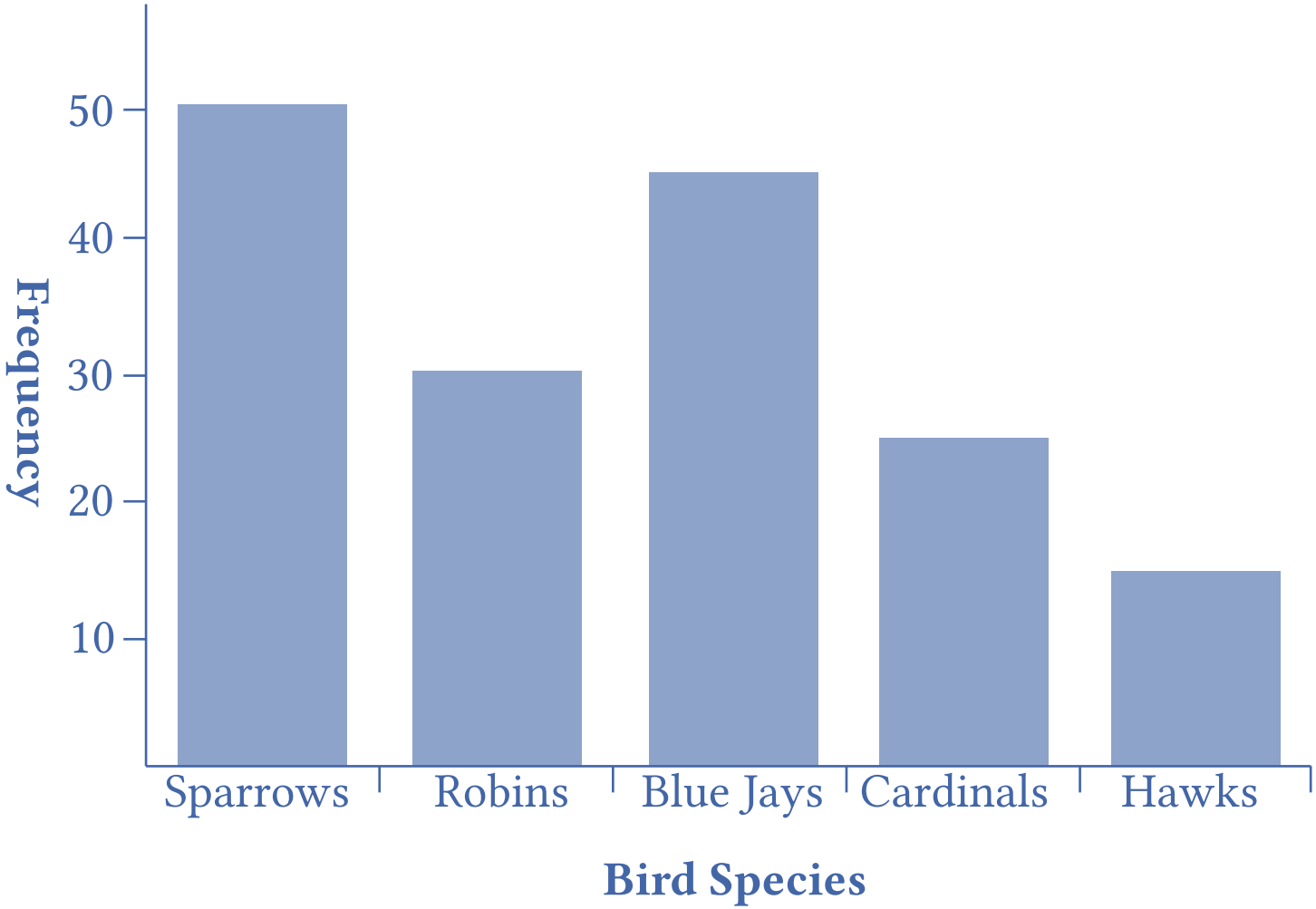
A local wildlife reserve is studying the behavior of different bird species visiting their park. Over the course of a month, park rangers record the number of sightings for five common bird species. They want to use this data to decide which areas of the park to focus conservation efforts on, based on which species are most commonly seen. The data collected is shown below:

Bird Species	Frequency
Sparrows	50
Robins	30
Blue Jays	45
Cardinals	25
Hawks	15

Example

A local wildlife reserve is studying the behavior of different bird species visiting their park. Over the course of a month, park rangers record the number of sightings for five common bird species. They want to use this data to decide which areas of the park to focus conservation efforts on, based on which species are most commonly seen. The data collected is shown below:

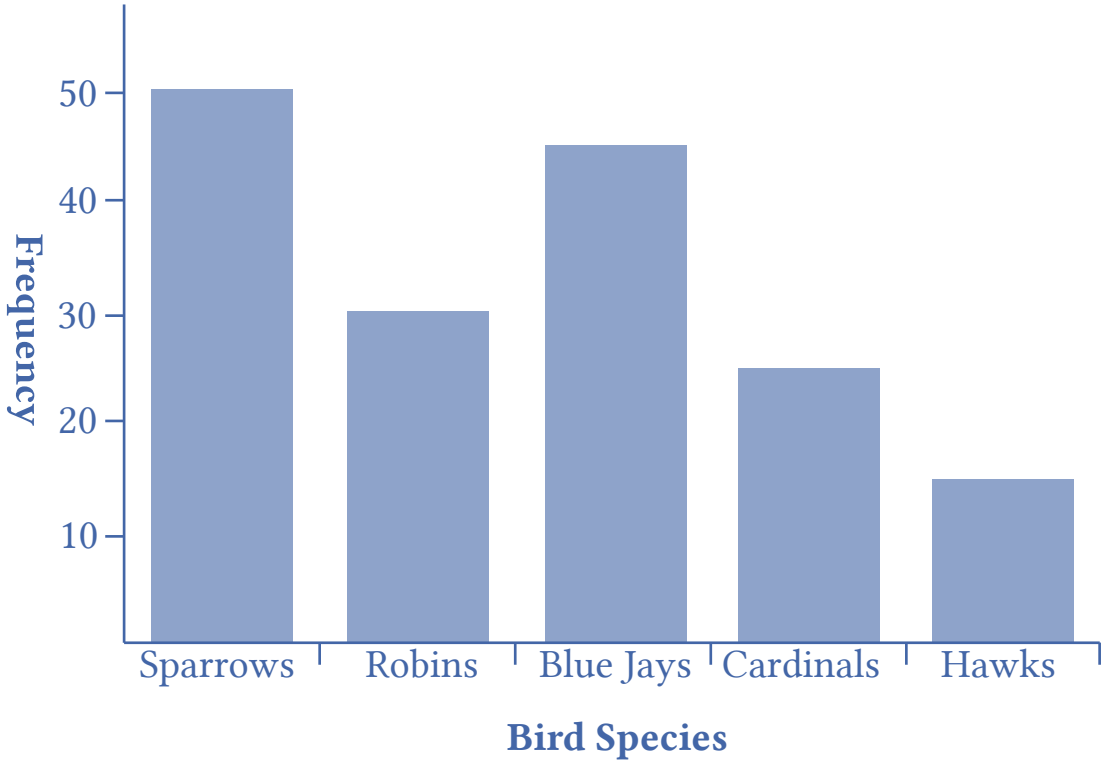
Bird Species	Frequency
Sparrows	50
Robins	30
Blue Jays	45
Cardinals	25
Hawks	15

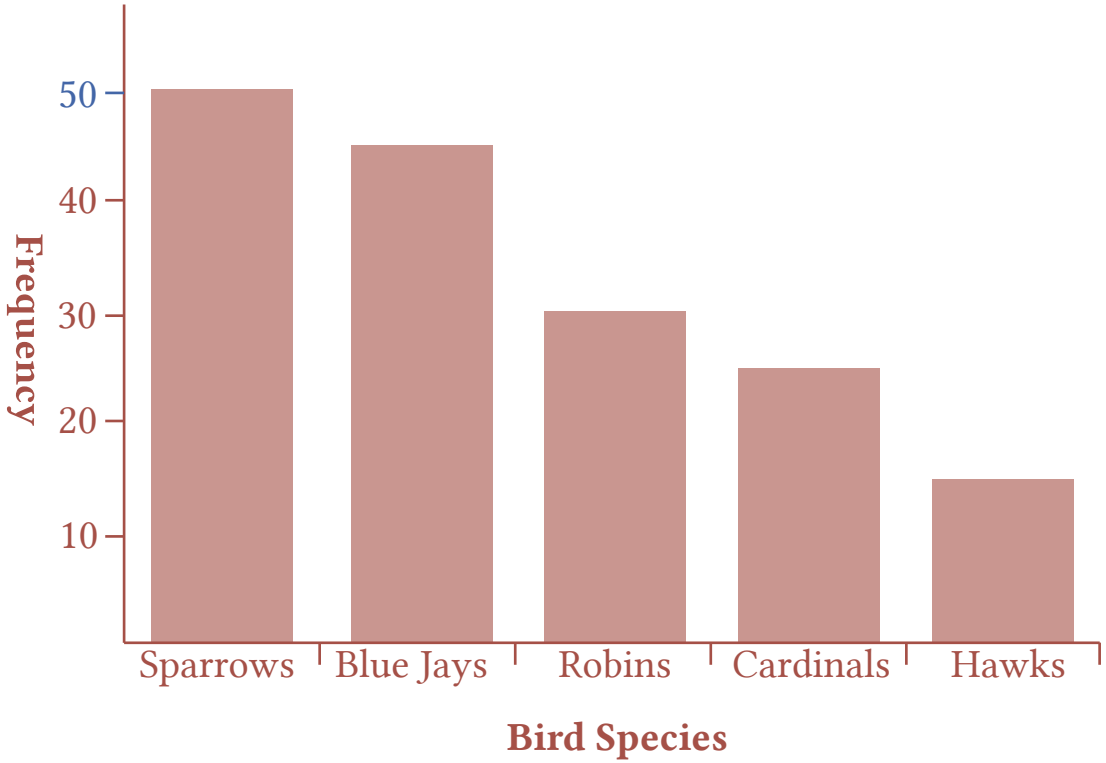


PARETO CHART

A **Pareto chart** is a bar graph ordered from *highest* to *lowest* frequency

Pareto chart



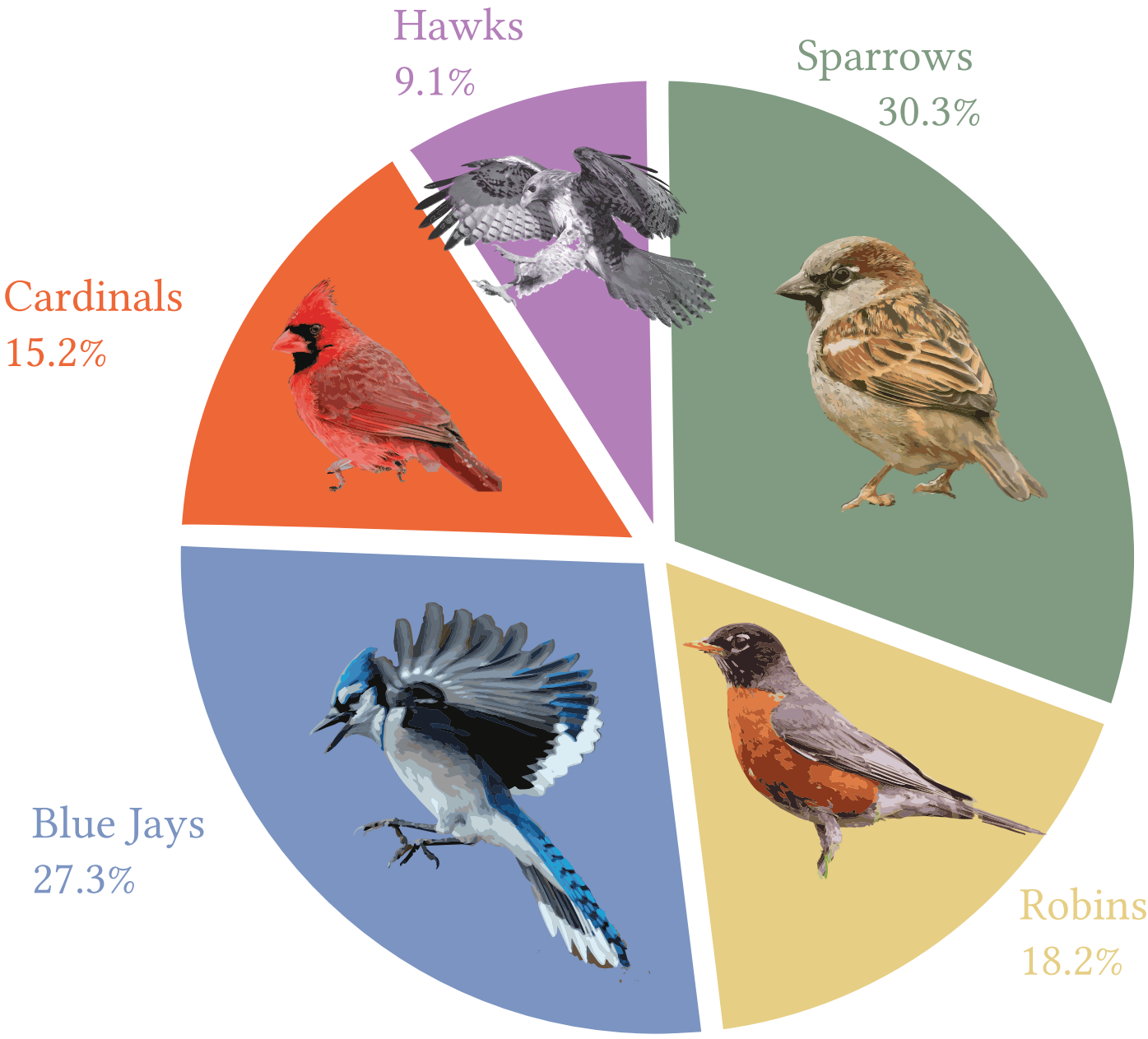


PIE CHART

A **pie chart** is a circle with wedges cut of varying sizes marked out like slices of pie or pizza.

The relative sizes of the wedges correspond to the relative frequencies of the categories.

Bird Species	Frequency
Sparrows	50
Robins	30
Blue Jays	45
Cardinals	25
Hawks	15



Question

Create a pie chart for the following frequency table:

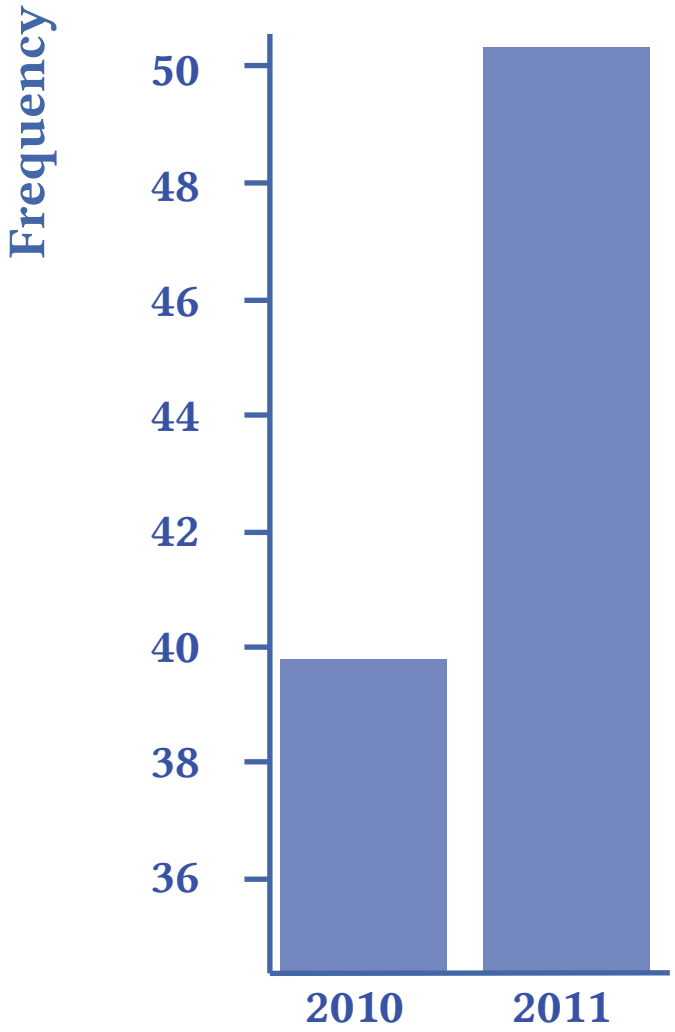
Workshop Type	Frequency
Painting	3
Sculpture	1
Music	2
Dance	4

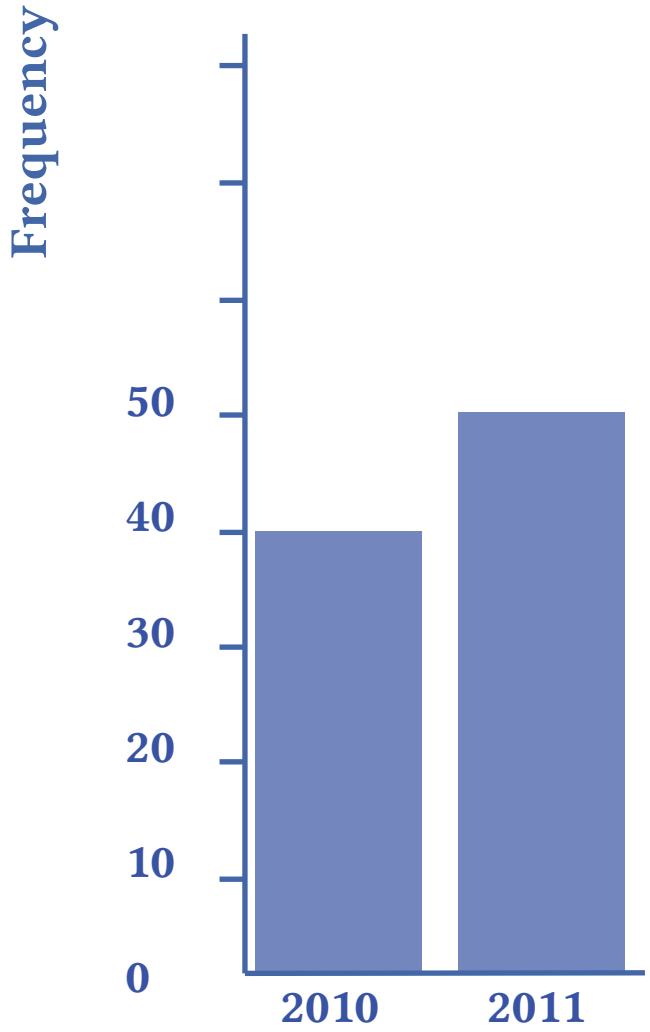
PICTOGRAM

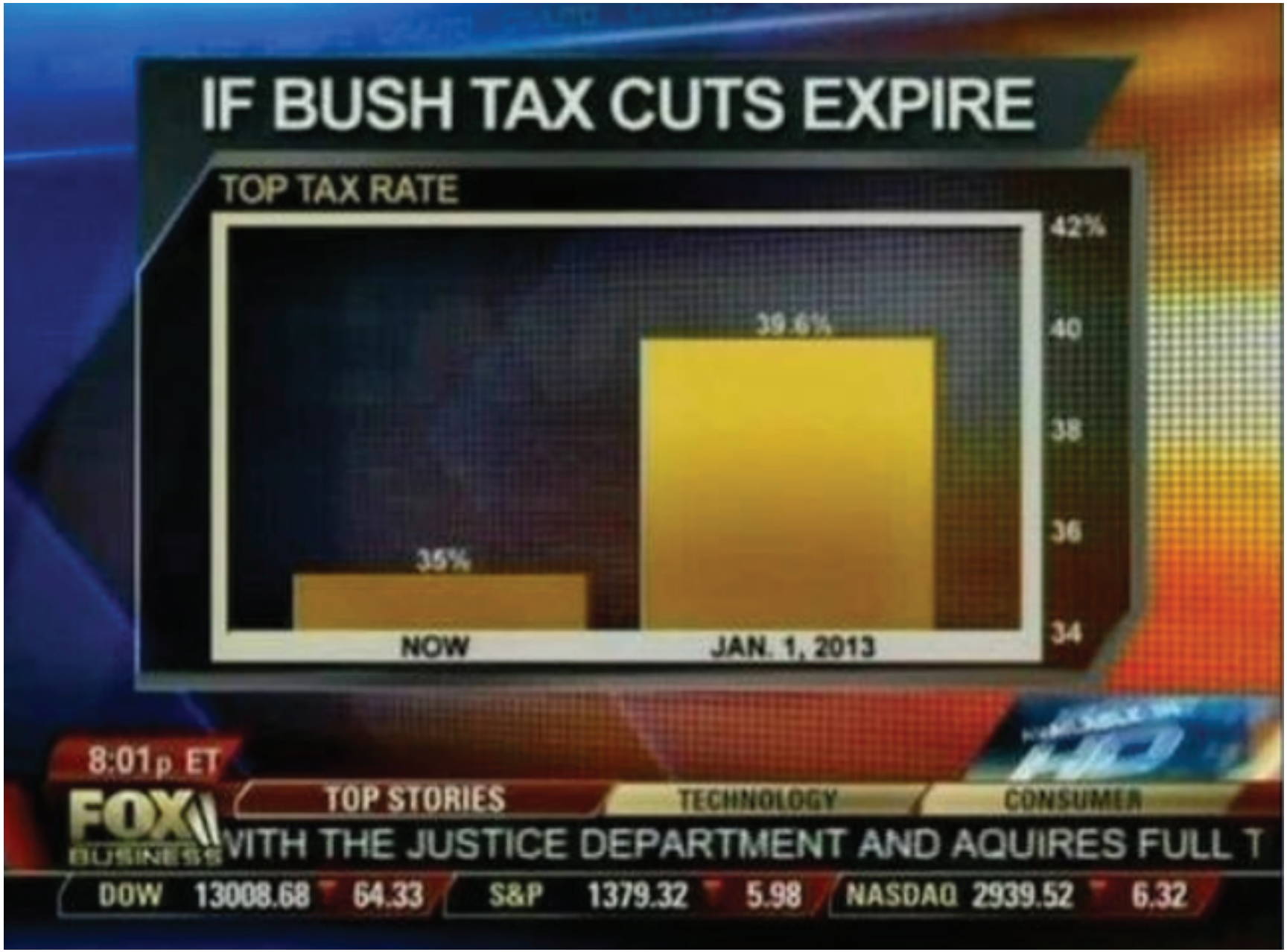
A **pictogram** is a statistical graphic in which the size of the picture is intended to represent the frequencies or size of the values being represented.

Beware:
It is easy to be misled with charts

Can you compare the following bar charts? What is misleading here?

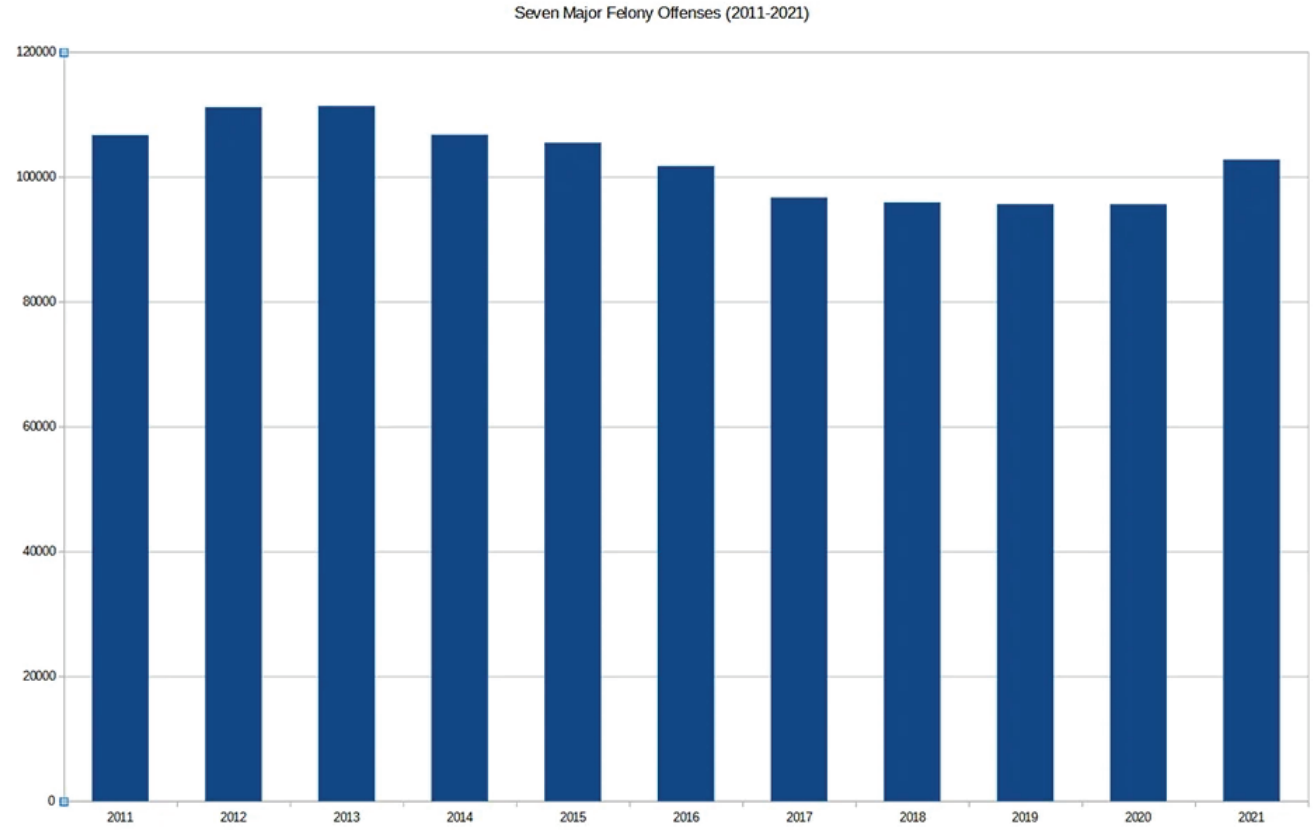




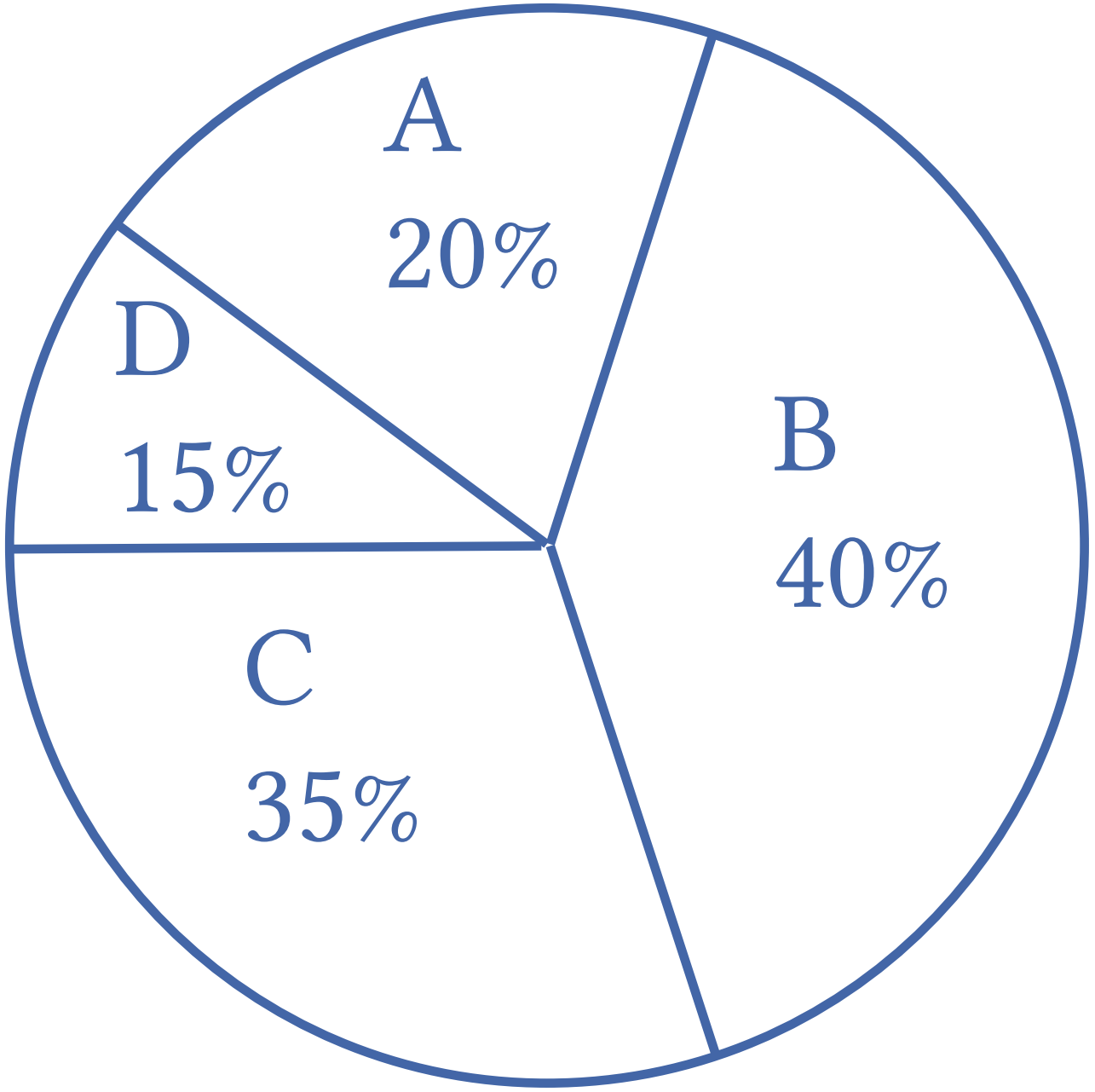

IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
TOP STORIES
TECHNOLOGY
CONSUMER
FOX BUSINESS
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32


EARLY APPROVAL RATINGS
SOURCE: CNN/ORC POLL | CNN/USA TODAY/GALLUP | GALLUP
SAMPLING ERROR: +/-3% PTS
80%
70%
60%
50%
40%
EISENHOWER '53
KENNEDY '61
NIXON '69
CARTER '77
REAGAN '81
BUSH '89
CLINTON '93
BUSH '01
OBAMA '09
TRUMP NOW


NYC
Major and Violent Crime Has Increased
Seven Major Felony Offenses (2017 - 2021)
104,000
102,000
100,000
98,000
96,000
96,658
95,883
95,606
95,593
102,741
7.5% more offenses than in 2020
2017
2018
2019
2020
2021
Calendar Year
Source: NYPD
MORE STREET PATROLS,


Seven Major Felony Offenses (2011-2021)
120000
100000
80000
60000
40000
20000
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

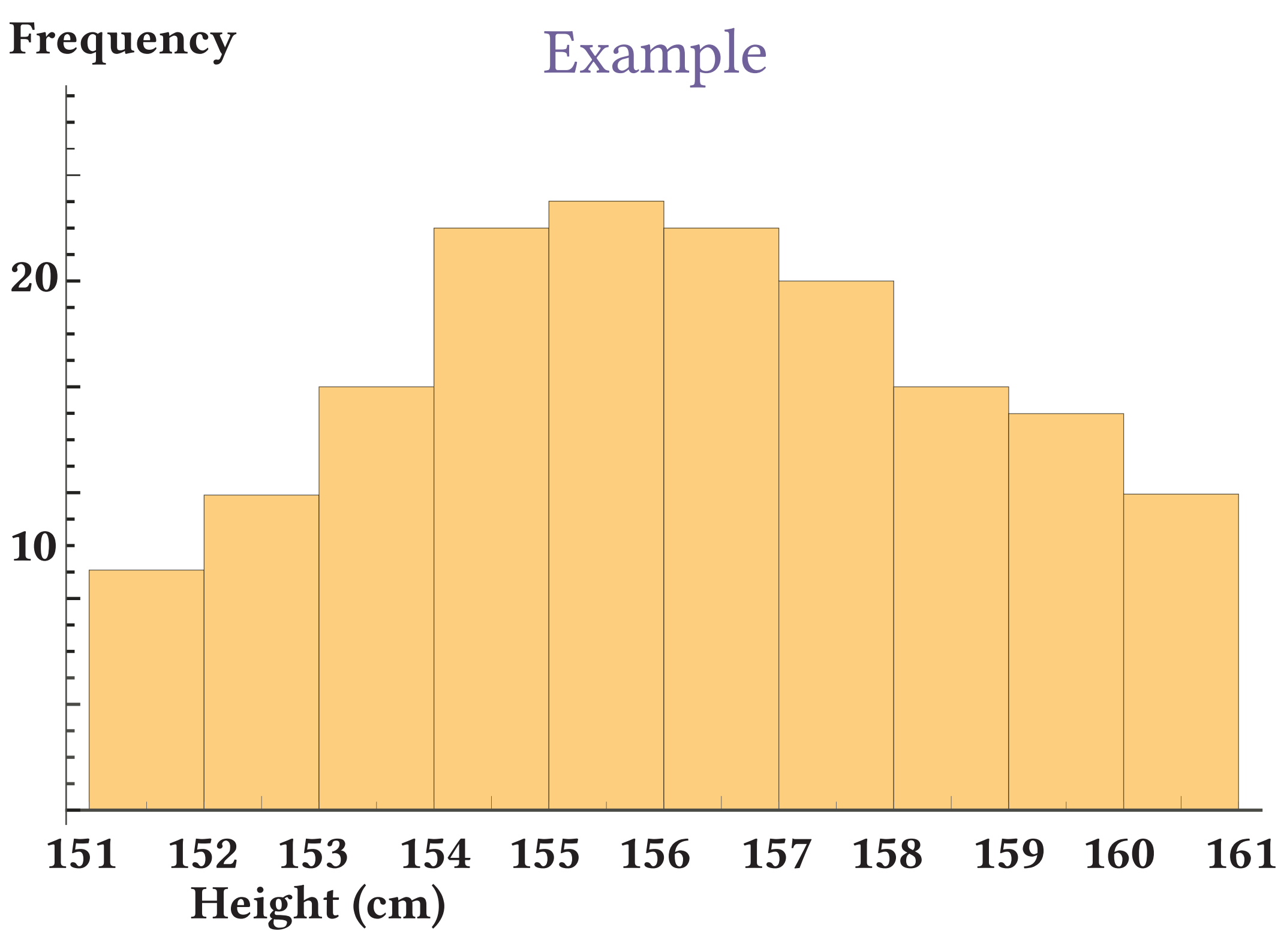
What's wrong with this pie chart?



HISTOGRAM

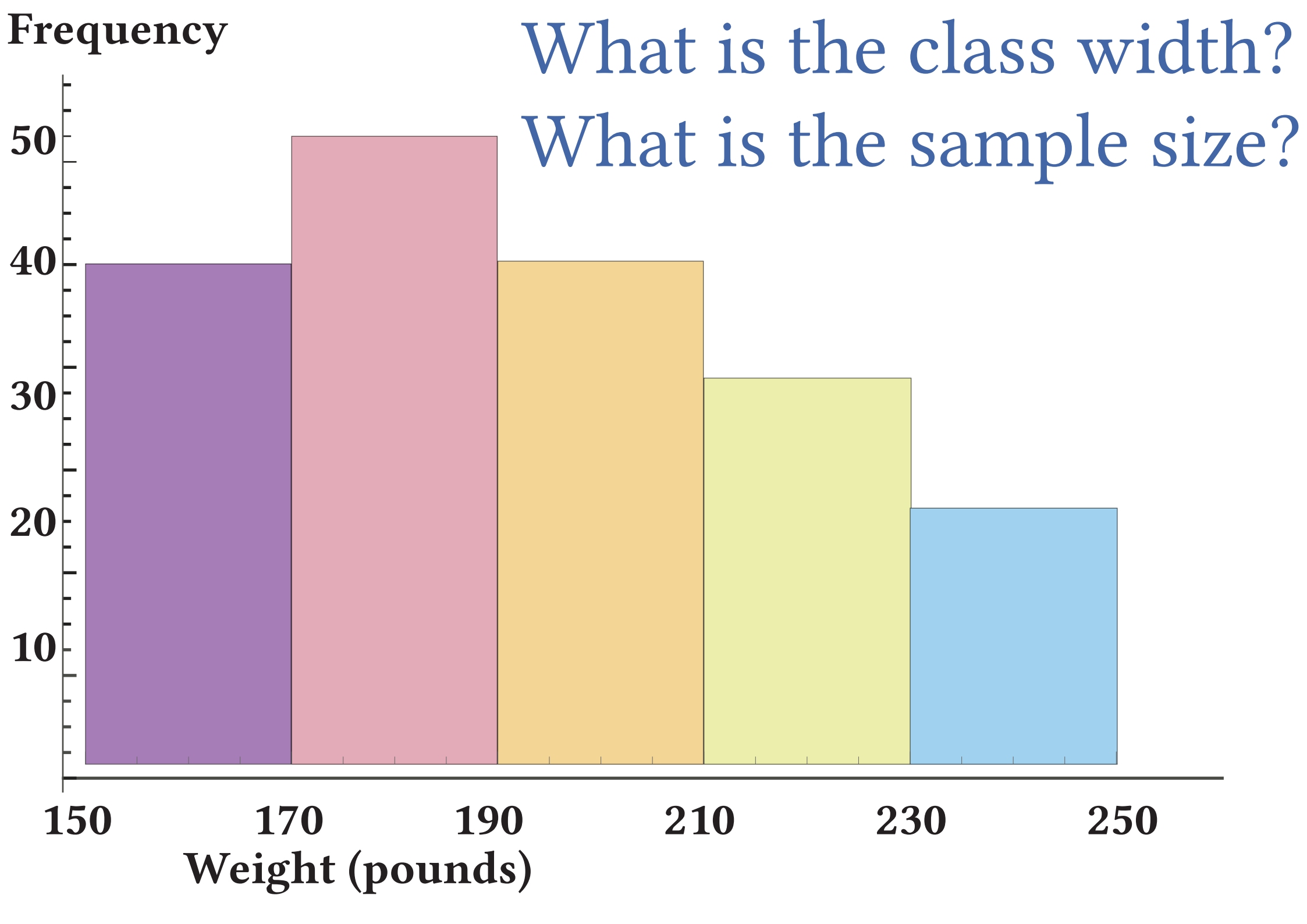
A **histogram** is like a bar graph, but where the horizontal axis is a number line.

Example


Frequency
20
10
151
152
153
154
155
156
157
158
159
160
161
Height (cm)

CLASS INTERVALS

Class intervals are groupings of the data. In general, we define class intervals so that each interval is equal in size.
For example, if the first class contains values from 120-129, the second class should include values from 130-139.
We have somewhere between 5 and 20 classes, typically, depending upon the number of data we're working with.



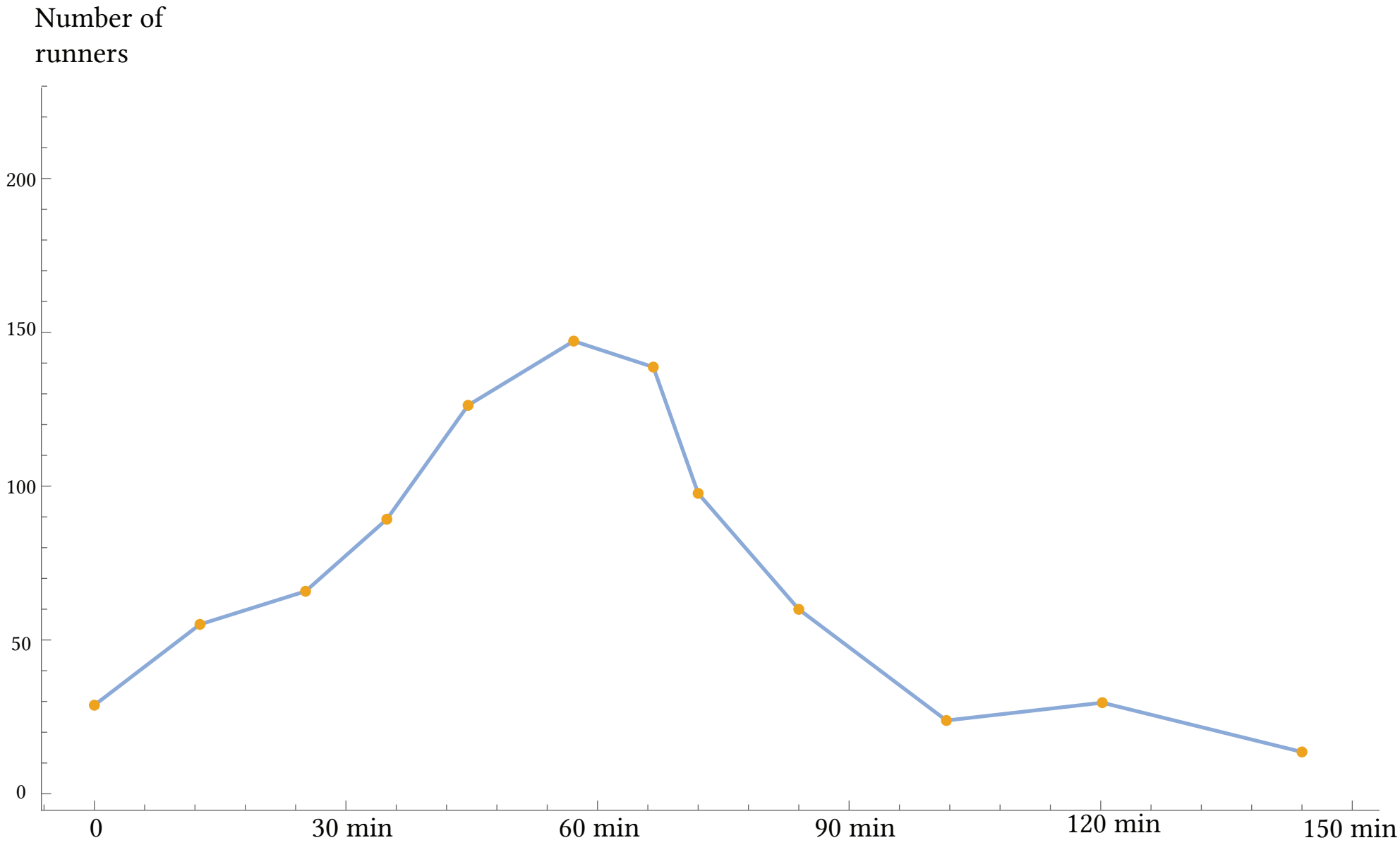
What is the class width?

What is the sample size?

FREQUENCY POLYGON

An alternative representation is a **frequency polygon**. A frequency polygon starts out like a histogram, but instead of drawing a bar, a point is placed in the midpoint of each interval at height equal to the frequency. Typically the points are connected with straight lines to emphasize the distribution of the data.

Example


Number of runners
200
150
100
50
0
0
30 min
60 min
90 min
120 min
150 min

Question

A tourism agency conducted a survey to analyze the ages of participants in a group travel program. The ages of 30 participants are listed below:

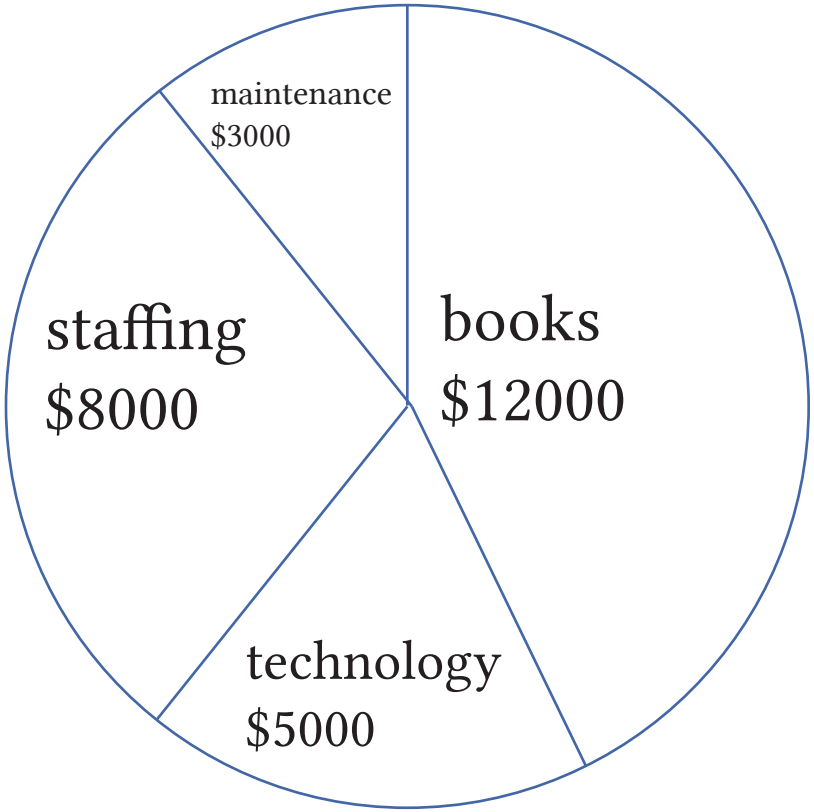
18, 21, 22, 25, 27, 28, 29, 31, 32, 34, 35, 37, 39, 41, 43, 45, 46, 48, 50, 52, 53, 54, 56, 58, 60, 62, 64, 66, 68

Complete the frequency distribution for the data.

Age	Frequency
20-29	
30-39	
40-49	
50-59	
60-69	
70-79	

Question

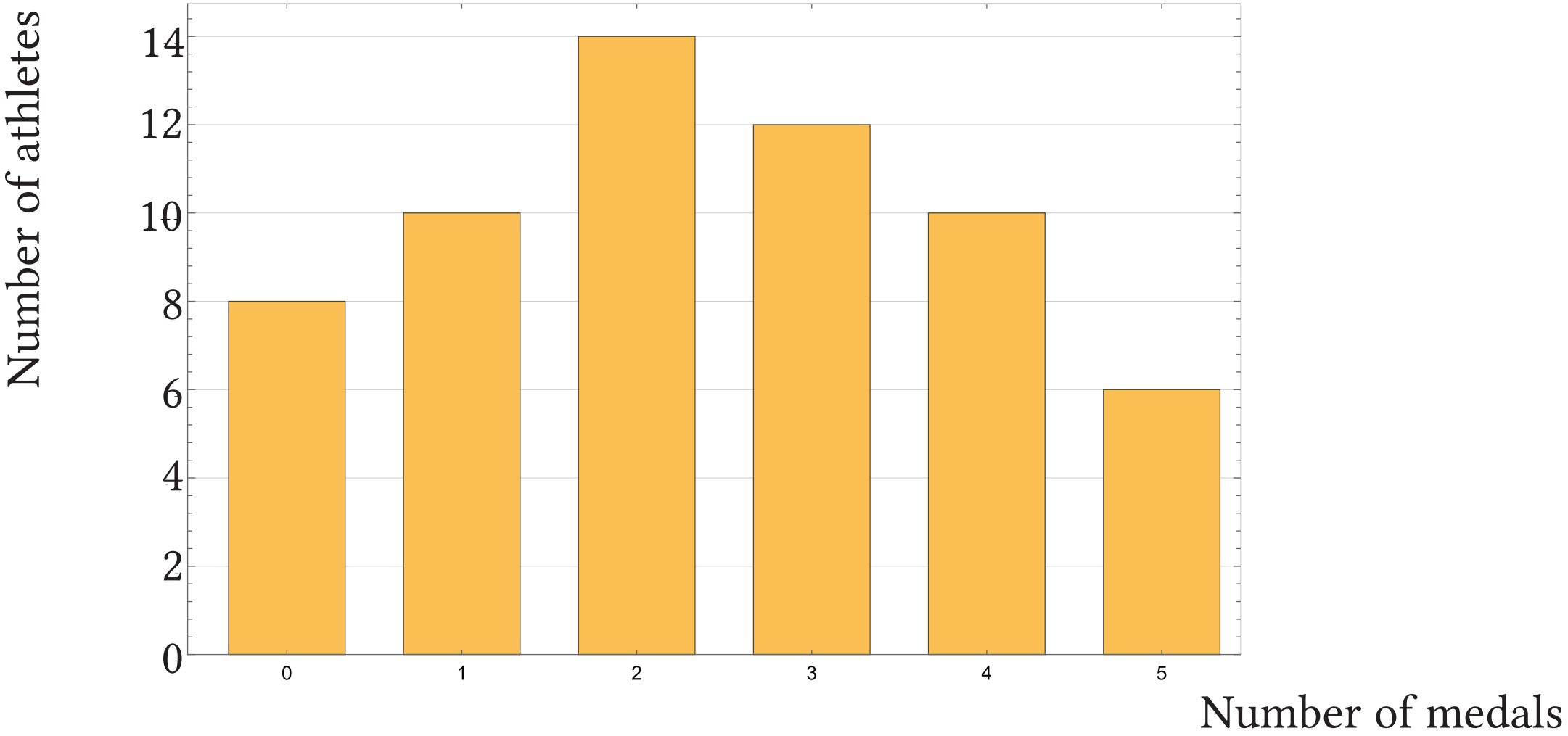
A local library tracks its annual spending in different areas: Books, Technology, Staffing, and Maintenance. For this year, the spending in each category is as follows:



Calculate the percentage of the total budget that was spent on Books.

Question

The data in the figure below represents the number of medals won by 150 athletes in a sports competition.



How many athletes won exactly 2 medals?

MEAN

The **mean** of a set of data is the sum of the data values divided by the number of values.

Example

During the track season, Marci ran four races, with times of 12.8 seconds, 13.4 seconds, 12.6 seconds, and 13.1 seconds. What is the mean of her race times?

Example

During the track season, Marci ran four races, with times of 12.8 seconds, 13.4 seconds, 12.6 seconds, and 13.1 seconds. What is the mean of her race times?

$$\frac{12.8+13.4+12.6+13.1}{4} = 12.975$$

Question

Marci baked four batches of cookies, with the following quantities: 24, 30, 27, and 33 cookies. What is the mean number of cookies per batch?

Question

Marci baked four batches of cookies, with the following quantities: 24, 30, 27, and 33 cookies. What is the mean number of cookies per batch?

$$\frac{24+30+27+33}{4} = 28.5$$

Question

A survey was conducted in a town where 80 households were asked about their monthly grocery expenses, rounded to the nearest $10. The results are summarized in the table below.

Grocery Expenses (dollars)	Frequency
150	5
160	10
170	14
180	18
190	16
200	9
210	5
220	3

What is the mean average monthly grocery expense for households in this town?

MEDIAN

The **median** of a set of data is the value in the middle when the data is in order.

To find the median, begin by listing the data in order from smallest to largest, or largest to smallest.

If the number of data values, N, is odd, then the median is the middle data value. This value can be found by rounding N/2 up to the next whole number.

If the number of data values is even, there is no one middle value, so we find the mean of the two middle values (values N/2 and N/2 + 1)

Example

Steve has the following list of his daily step counts for the last 30 days and wants to calculate the median:

4,500 4,600 4,800 5,000 5,200 5,500 5,700 5,800 6,000 6,200 6,500 6,500 6,800 7,000 7,100 7,200 7,400 7,600 7,800 8,000 8,200 8,300 8,500 8,600 8,800 9,000 9,200 9,400 9,600 10,000

What is the median step count value?

Example

Steve has the following list of his daily step counts for the last 30 days and wants to calculate the median:

4,500 4,600 4,800 5,000 5,200 5,500 5,700 5,800 6,000 6,200 6,500 6,500 6,800 7,000 7,100 7,200 7,400 7,600 7,800 8,000 8,200 8,300 8,500 8,600 8,800 9,000 9,200 9,400 9,600 10,000

What is the median step count value?

As N is **even**, the median will be the average of the 15th and 16th values in ascending order. Thus the median is

$$\frac{7{,}100 + 7{,}200}{2} = 7{,}150$$

Question

A group of employees tracked the number of emails they sent in a day, listed in increasing order:

5, 13, 7, 8, 17, 10, 11, 12, 19, 14, 15, 16, 9, 18, 6

What is the median number of emails sent?"

Example

A city conducted a survey on the number of hours people spent volunteering in a month. The results are summarized in the table below.

Hours Volunteered	Frequency
5	8
10	12
15	15
20	18
25	14
30	10
35	7
40	6

What is the mean number of hours spent volunteering?

MODE

The **mode** is the element of the data set that occurs most frequently.

It is possible for a data set to have more than one mode if several categories have the same frequency, or no modes if each every category occurs only once.

Example

In a survey of favorite fruits among a group of friends, the following data was collected:

Fruit	Frequency
Apple	4
Banana	6
Cherry	2
Date	3
Elderberry	5

Which fruit is the mode?

Example

In a survey of favorite fruits among a group of friends, the following data was collected:

Fruit	Frequency
Apple	4
Banana	6
Cherry	2
Date	3
Elderberry	5

Which fruit is the mode?
Bananas occur with a maximum freqency of 6, thus bananas are the mode.

Question

In a survey of customer satisfaction with a new service, the following data was collected on the number of issues reported:

Number of Issues	Frequency
0	12
1	15
2	8
3	5
4	3

Find:

The average number of issues reported

The median number of issues reported

The mode of the number of issues reported

RANGE

The **range** is the difference between the *maximum value* and the *minimum value* of the data set.

Example

In a study of the number of hours worked per week by employees in a department, the following data was collected:

Hours Worked per Week: 38, 42, 45, 37, 50, 41, 39, 48

What is the range of this data?

Example

In a study of the number of hours worked per week by employees in a department, the following data was collected:

Hours Worked per Week: 38, 42, 45, 37, 50, 41, 39, 48

What is the range of this data?

Range = maximum value - minimum value

= 50 - 37

= 13

STANDARD DEVIATION

The **standard deviation** is a measure of variation based on measuring how far each data value deviates, or is different, from the mean. A few important characteristics:

Standard deviation is always positive. Standard deviation will be zero if all the data values are equal, and will get larger as the data spreads out.
Standard deviation has the same units as the original data.
Standard deviation, like the mean, can be highly influenced by outliers.

STANDARD DEVIATION

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

TO COMPUTE THE STANDARD DEVIATION

1. Calculate the mean
2. Find the deviation of each data from the mean. In other words, subtract the mean from the data value.

2. Square each deviation.

3. Add the squared deviations.

4. Divide by n, the number of data values, if the data represents a whole population; divide by n – 1 if the data is from a sample.

5. Compute the square root of the result.

Example

The temperatures recorded in five different cities for a particular day were 72°F, 68°F, 75°F, 70°F, and 80°F. Find the standard deviation of the temperatures.

Example

The temperatures recorded in five different cities for a particular day were 72°F, 68°F, 75°F, 70°F, and 80°F. Find the standard deviation of the temperatures.

$$\text{mean} = \mu = \frac{72+68+75+70+80}{5} = \frac{365}{5} = 73$$

difference from the mean:	Squaring the difference:	Adding the values:
$72 - 73 = -1$	$(72 - 73)^2 = (-1)^2 = 1$	$1 + 25 + 4 + 9 + 49 = 88$
$68 - 73 = -5$	$(68 - 73)^2 = (-5)^2 = 25$	
$75 - 73 = 2$	$(75 - 73)^2 = (2)^2 = 4$	$88/5 = 17.6$ Dividing by the number of values
$70 - 73 = -3$	$(70 - 73)^2 = (-3)^2 = 9$	$\sqrt{17.6} = 4.19523539268$ taking square root
$80 - 73 = 7$	$(80 - 73)^2 = (7)^2 = 49$	≈ 4.20

Question

The number of miles run each day for a week by a runner were: 4, 6, 5, 7, 8, 6, and 5. Calculate the standard deviation of the number of miles run.

Question

The number of items sold by a small shop over seven days were: 20, 25, 22, 30, 27, 24, and 26. Calculate the standard deviation of the number of items sold.

Question

The number of attendees at five different workshops were: 45, 50, 55, 60, and 52. Calculate the standard deviation of the number of attendees.

Question

The number of calls received by a customer service center each hour over an 8-hour shift were: 32, 27, 35, 30, 28, 31, 29, and 34. Calculate the standard deviation of the number of calls received.

QUARTILES

Quartiles are values that divide the data in quarters.

The first quartile (Q1) is the value so that *25% of the data values are below it*; the third quartile (Q3) is the value so that *75% of the data values are below it*. You may have guessed that the second quartile is the same as the median, since the median is the value so that 50% of the data values are below it.

This divides the data into quarters; 25% of the data is between the minimum and Q1, 25% is between Q1 and the median, 25% is between the median and Q3, and 25% is between Q3 and the maximum value.

FIVE NUMBER SUMMARY

The five number summary takes this form:

Minimum, Q1, Median, Q3, Maximum

TO FIND THE FIRST QUARTILE, Q1

Begin by ordering the data from smallest to largest
Compute the locator: **L = 0.25n**
If L is a decimal value:
Round up to L+
Use the data value in the L+th position
If L is a whole number:
Find the mean of the data values in the Lth and L+1th positions.

TO FIND THE THIRD QUARTILE, Q3

Use the same procedure as for Q1, but with locator: L = 0.75n

Example

Suppose a group of 10 athletes have their running speeds (in meters per second) recorded, and their speeds sorted from slowest to fastest are:

5.8, 6.2, 6.4, 6.6, 6.8, 7.0, 7.2, 7.4, 7.6, 7.8

What are the first and third quartiles of their running speeds?

Example

Suppose a group of 10 athletes have their running speeds (in meters per second) recorded, and their speeds sorted from slowest to fastest are:

5.8, 6.2, 6.4, 6.6, 6.8, 7.0, 7.2, 7.4, 7.6, 7.8

What are the first and third quartiles of their running speeds?

Q1: 0.25 x 10 = 2.5
Rounding up, we get 3, the third position. So Q1 = 6.4 m/s

Q3: 0.75x10 = 7.5
Rounding up, we get 8, the eight position. So Q3 = 7.4 m/s

Example

Suppose a group of 10 athletes have their running speeds (in meters per second) recorded, and their speeds sorted from slowest to fastest are:

5.8, 6.2, 6.4, 6.6, 6.8, 7.0, 7.2, 7.4, 7.6, 7.8

What is the 5 number summary?

Example

Suppose a group of 10 athletes have their running speeds (in meters per second) recorded, and their speeds sorted from slowest to fastest are:

5.8, 6.2, 6.4, 6.6, 6.8, 7.0, 7.2, 7.4, 7.6, 7.8

What is the 5 number summary?

Median: mean of 5th and 6th position

$$\text{Median} = \frac{6.8 + 7.0}{2} = 6.9$$

Min = 5.8 , Max = 7.8

Five number summary: 5.8, 6.4, 6.9, 7.4, 7.8

Question

The monthly rent paid by 24 individuals was recorded, and the amounts sorted from lowest to highest are:

$800, $850, $850, $900, $950, $950, $975, $1000, $1025, $1050, $1075, $1100, $1125, $1150, $1200, $1250, $1300, $1300, $1350, $1400, $1450, $1500, $1550, $1600

Find the 5-number summary of this data.

Example

Using the data on weekly working hours from a group of employees, create the five-number summary.

Weekly Working Hours (hours)	Frequency
30	4
35	6
40	10
45	15
50	18
55	12
60	8
65	5

Example

Weekly Working Hours (hours)	Frequency
30	4
35	6
40	10
45	15
50	18
55	12
60	8
65	5

The total amount of employees is the total sum of the frequencies:

4 + 6 + 10 + 15 + 12 + 8 + 5 = 78

Min = 30, Max = 65

Q1: 78/4 = 19.5, rounding up we get 20. 40 hours is in the 20th position and so Q1 = 40.

Median: 78/2 = 39. The mean of the 39th and 40th position is 50.

Q3: 3x78/4 = 58.5, rounding up we get 59. 55 hours is in the 59th position and so Q3 = 59.

5 number summary: 30, 40, 50, 59, 65

Question

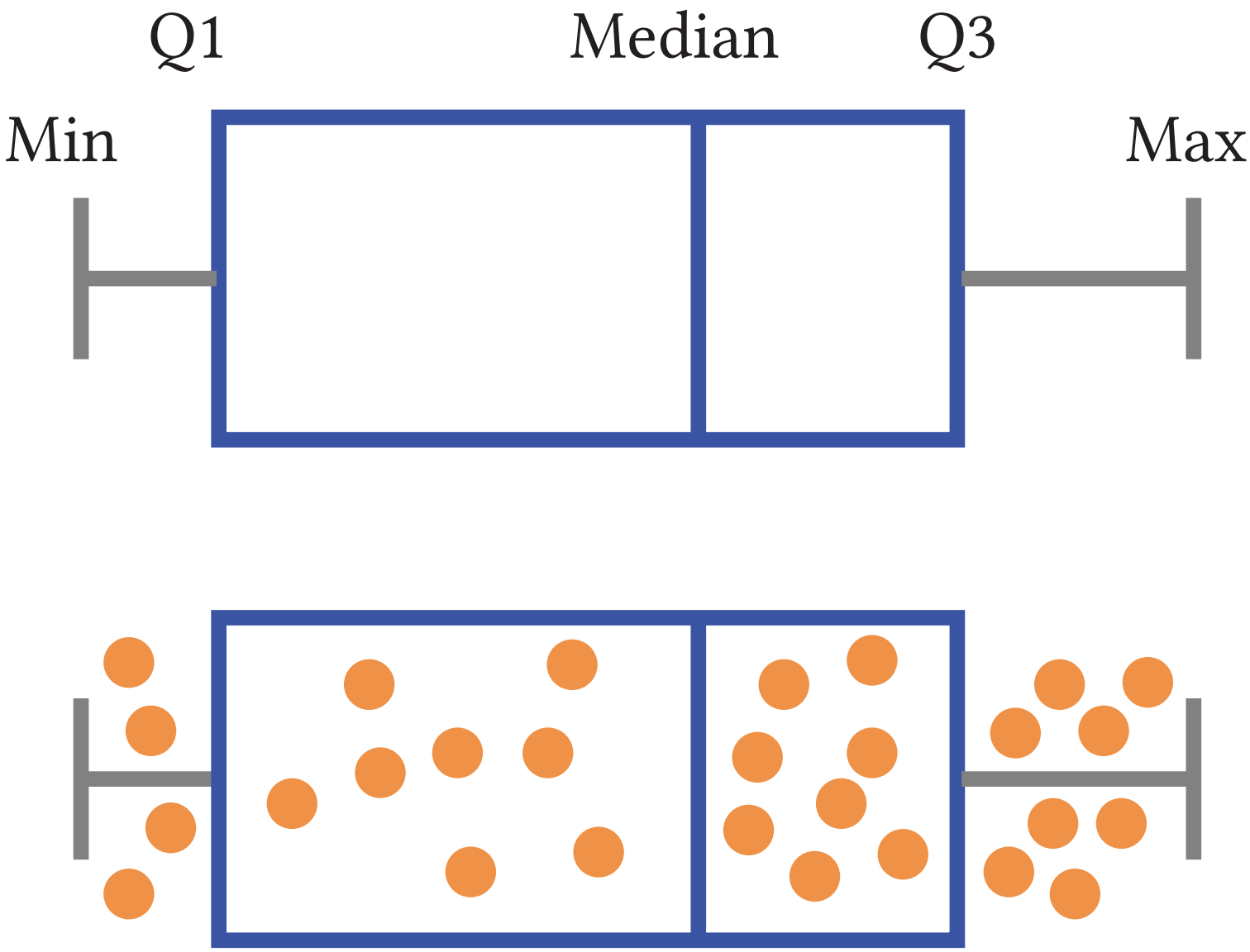
The daily caloric intake (in kilocalories) of 60 individuals was recorded, and their intake data is grouped as follows:

Calories (kcal)	Frequency
1800	5
2000	8
2200	12
2400	15
2600	10
2800	6
3000	4

Using this data, calculate the five-number summary.

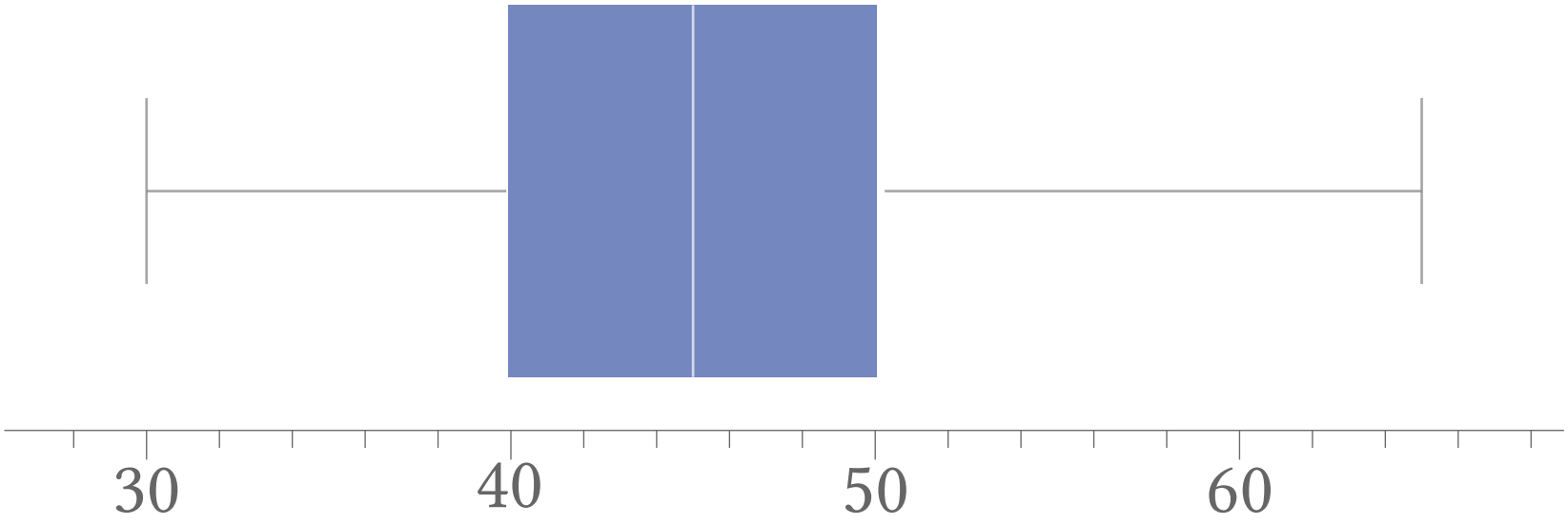
BOX PLOT

A **box plot** is a graphical representation of a five-number summary.


Q1
Median
Q3
Min
Max

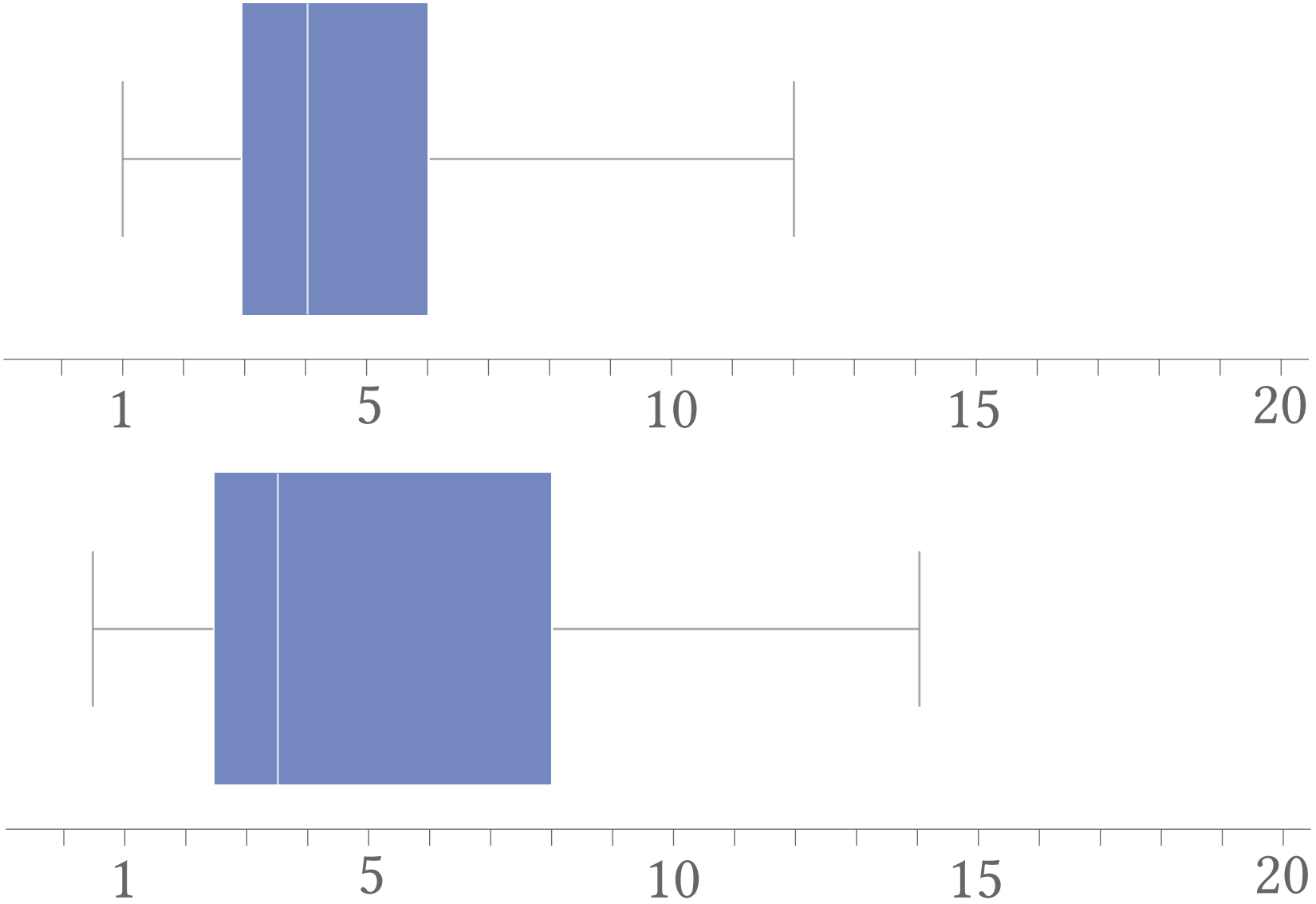
Example

five-number summary:
30, 40, 45, 50, 65



Example

The box plot of delivery times in days for two retailers is shown below:



Which retailer would you choose for a faster delivery?