BASIC VOCABULARY REVIEW

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

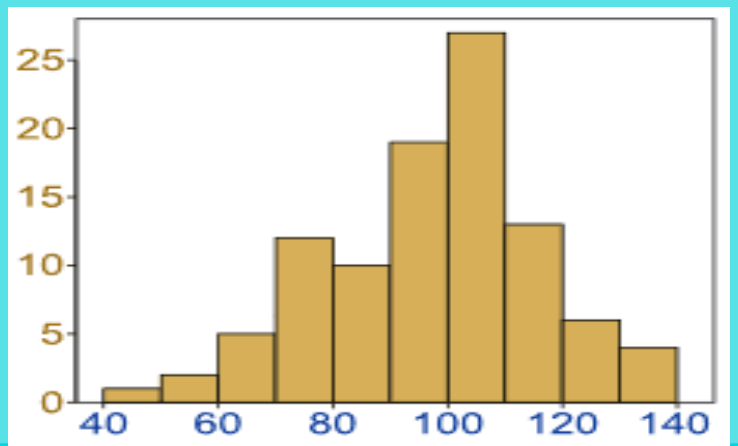
UNIT 6
TEST REVIEW

HISTOGRAM

A **frequency** plot that shows the number of times a data value or **range** of data values occurred in a data set.

Typically used for **larger data set**s and grouped into **equal intervals**.

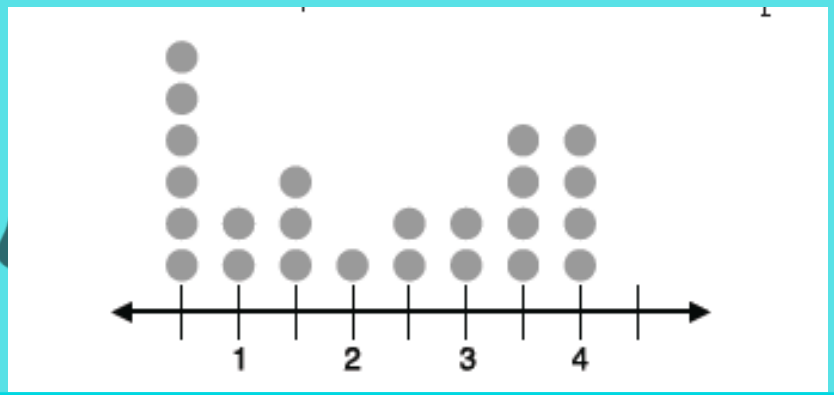
Example:



DOT PLOT

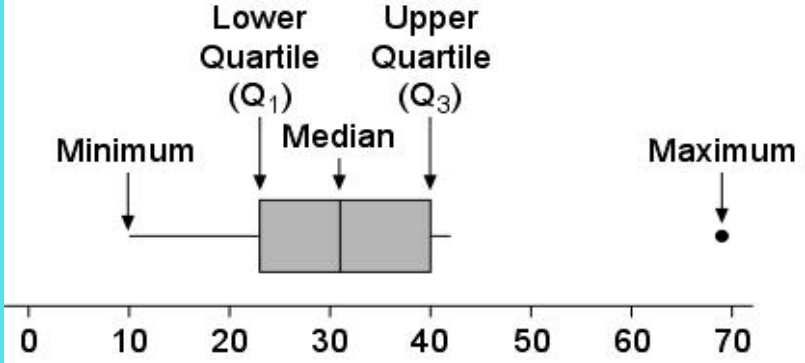
A **frequency** plot that shows the number of times a response occurred in a data set, where each **data value is represented by a dot**.

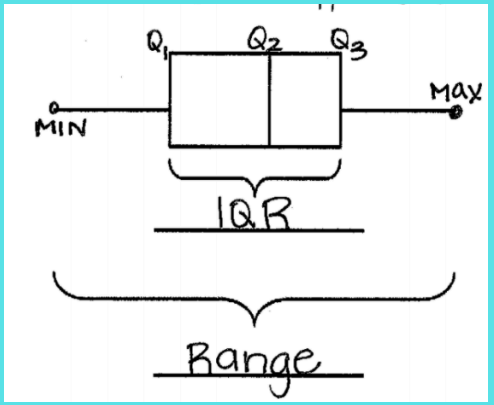
Example:

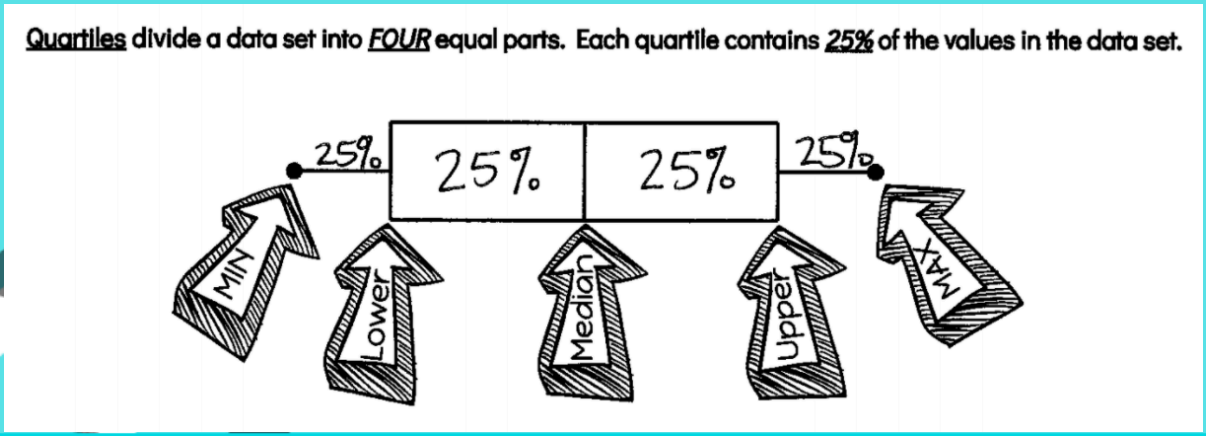


BOX PLOT

A plot showing the **5-number summary**: minimum, maximum, first quartile, median, and third quartile of a data set; the middle 50% of the data is indicated by a box







TO SUMMARIZE: DATA DISPLAYS

Definition	Advantages	Disadvantages
The **BOX PLOT** is a standardized way of displaying the distribution of data based on the 5 number summary of the data set.	• Shows 5 number summary • Shows outliers • Easy to compare 2 data sets • Handles large data sets	• Not visually appealing • Cannot tell exact values
A **HISTOGRAM** is a type of graph that shows the frequency distribution of data within equal intervals.	• Visually strong • Good for determining the shape of the data	• Cannot tell exact values • Difficult to compare 2 data sets
A **DOT PLOT** is a graphic display using dots and a simple scale to compare the frequency within categories or groups.	• Simple to make • Shows exact values	• Can be time consuming • Have to count to get exact total. (Fractions of units are hard to display.)

COMPARE DISTRIBUTIONS

When you compare two or more data sets, focus on 4 features:

1. **Center** - Graphically, the center of a distribution is the point where about half of the observations are on either side.
2. **Spread** - The **variability** of the data. If the observations cover a wide range, the spread is larger. If the observations are clustered around a single value, the spread is smaller.
3. **Shape** - The shape of a distribution is described by symmetry, skewness, number of peaks, etc.
4. **Unusual features** - Unusual features refer to gaps and outliers.

Our Example
{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

MEASURES OF CENTER & MEASURES OF SPREAD (VARIABILITY)

To Review

Data sets can be compared using measures of center and variability.

Measures of Center (central tendency)

1. Mean: use to describe the data set when an outlier is NOT present (symmetric data)
2. Median: use when outliers are present (skewed data)

★ The mean and median are both measures intended to be a single number that best represents an entire data set.

Measures of Variability

1. Range: max - min
2. Interquartile Range (IQR) = Q3-Q1. Used to describe the middle 50% of the data.
3. Mean Absolute Deviation (MAD): takes the average distance of the data points from the mean.

★ The range, IQR, and MAD are both measures intended to summarize the variability of the data using one number.

MEASURES OF CENTER: MEAN AND MEDIAN

Measures of Center (central tendency)

★ The mean and median are both measures intended to be a single number that best represents an entire data set.

Finding the Mean
1. Find the sum of the data values. 2. Divide the sum by the number of data points. This is the mean.

Finding the Median
1. First arrange the data from least to greatest. 2. Count the number of data points. If there is an even number of data points, the median is the average of the two middle-most values. If there is an odd number of data points, the median is the middle-most value.

MEAN

The average value of a data set, found by summing all values and dividing by the number of data points

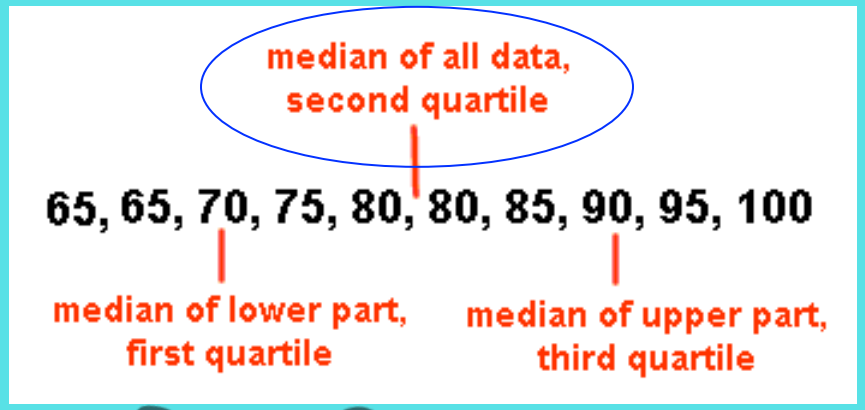
Example: **{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}**

MEDIAN

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

The middle-most value of a data set; 50% of the data is less than this value, and 50% is greater than it

Example:



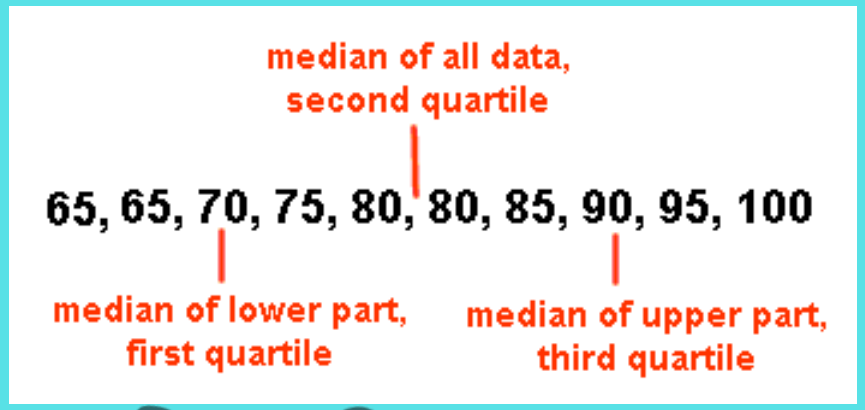
80 is the <u>MEDIAN</u>

MODE

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

- The value that appears the most often in a set of data.

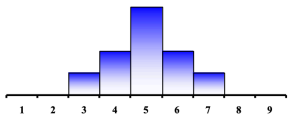
Example:



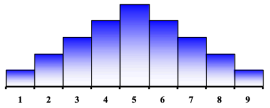
MEASURES OF SPREAD (VARIABILITY)

RANGE, IQR, MAD

vs. consistency



Less Spread



More Spread

Measures of Variability

★ The range, IQR, and MAD are both measures intended to summarize the variability of the data using one number.

Finding the Interquartile Range

1. Arrange the data from least to greatest.
2. Find the median of the data set. The median divides the data into two halves: the lower half and the upper half.
3. Find the middle-most value between the min. value and the median. This is the first quartile, Q_1.
4. Find the middle-most value between the median and the max value. This is the third quartile, Q_3.
5. Calculate the difference between the two quartiles, $Q_3 - Q_1$.

Finding the Mean Absolute Deviation (M.A.D.)

1. Find the mean.
2. Calculate the absolute value of the difference between each data value and the mean.
3. Determine the average of the differences found in step 2. This average is the mean absolute deviation.

RANGE

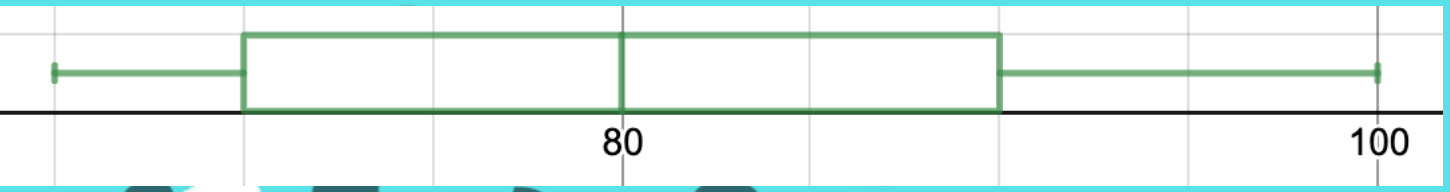
- The difference between the lowest and the highest value in a set of data.
- Maximum – minimum

2

$\text{stats}(L)$

Min	65
Q1	70
Median	80
Q3	90
Max	100

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}



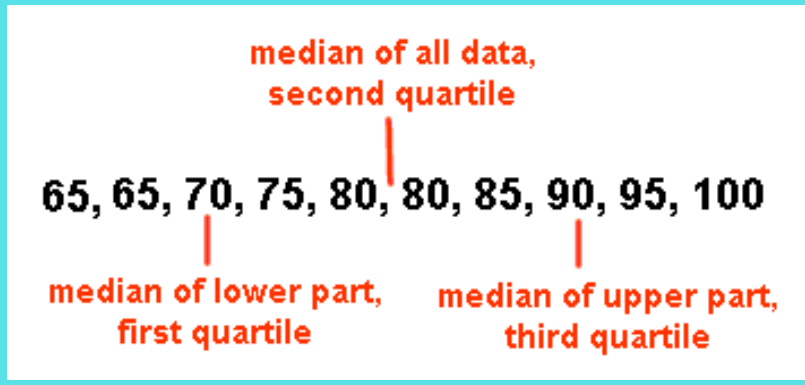
INTERQUARTILE RANGE (IQR)

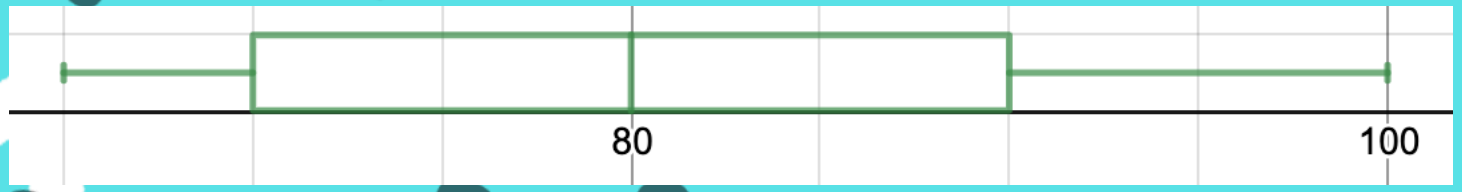
{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

The difference between the third and first quartiles; 50% of the data is contained within this range

SUBTRACT Q3 - Q1

IQR= THIRD QUARTILE - FIRST QUARTILE

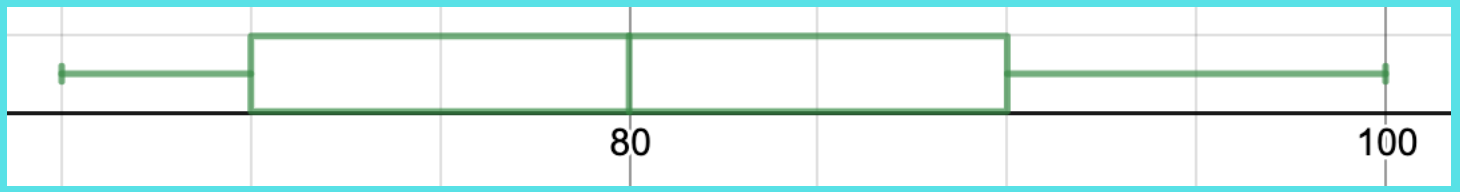


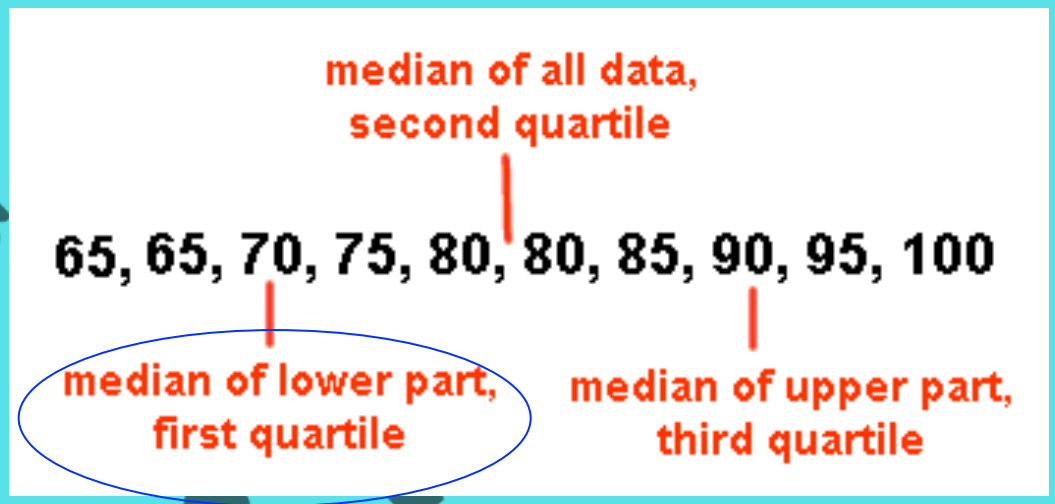


FIRST QUARTILE (LOWER)

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

The value that identifies the lower 25% of the data; the **median of the lower half** of the data set; written as Q_1 or $Q1$

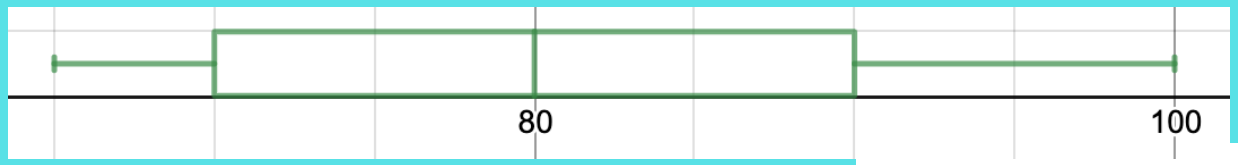


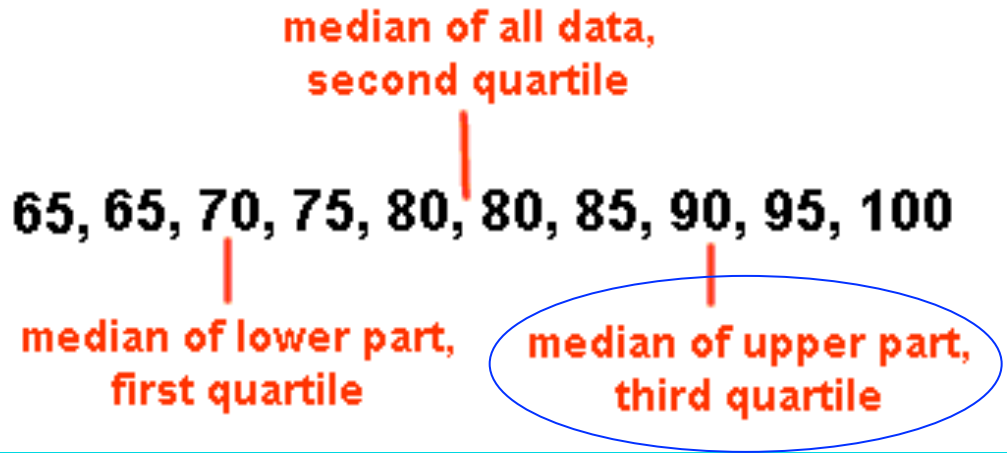


THIRD QUARTILE

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

Value that identifies the upper 25% of the data; the **median of the upper half** of the data set; 75% of all data is less than this value; written as Q_3 or $Q3$



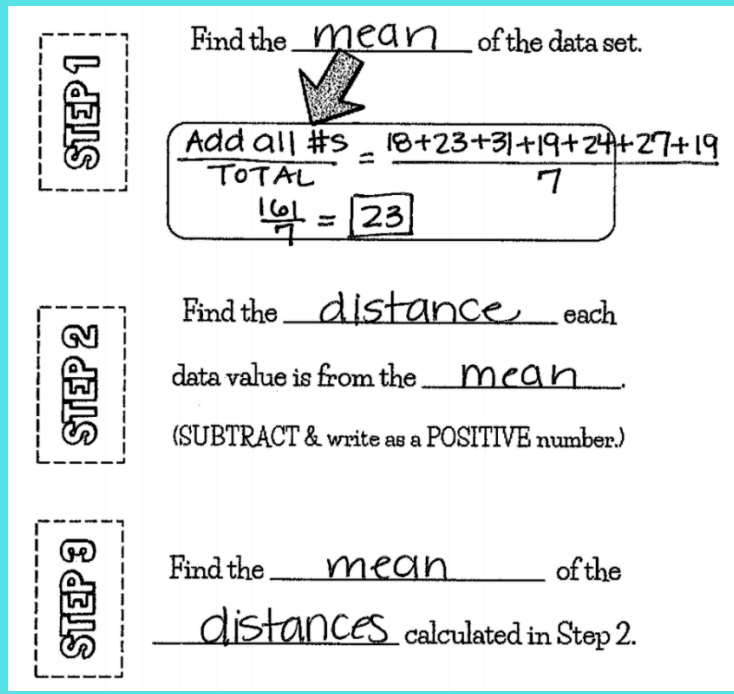


MEAN ABSOLUTE DEVIATION (M.A.D.)

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

The average distance between each data value and the mean. This is a way to describe variability (spread).

Different Example:



OUTLIER

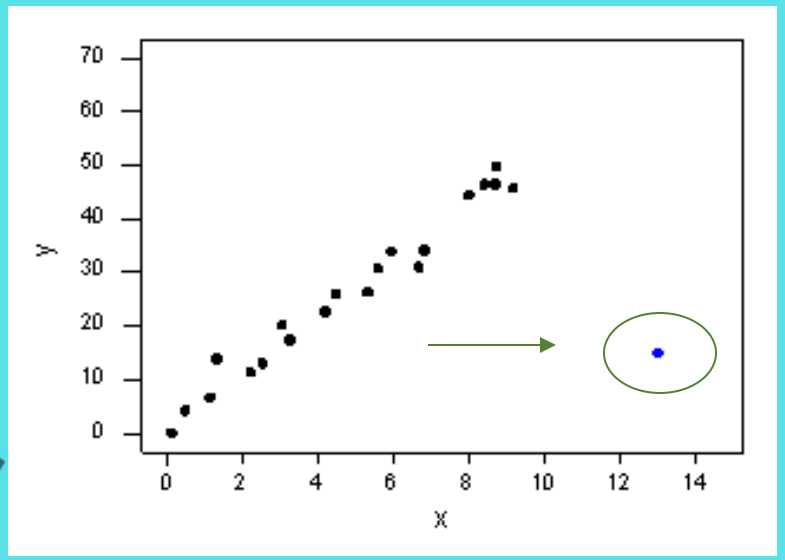
{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}

A data value that is much greater than or much less than the rest of the data in a data set; mathematically, any data less than Q1 + 1.5(IQR) or greater than Q3 + 1.5(IQR) is an outlier.

$Q_1 + 1.5(IQR)$

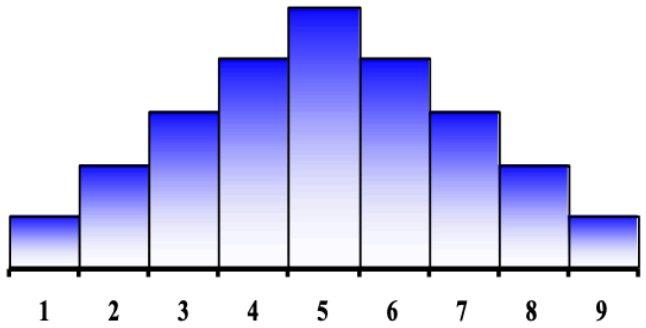
$Q_3 + 1.5(IQR)$

Example:



DESCRIBING: SHAPE

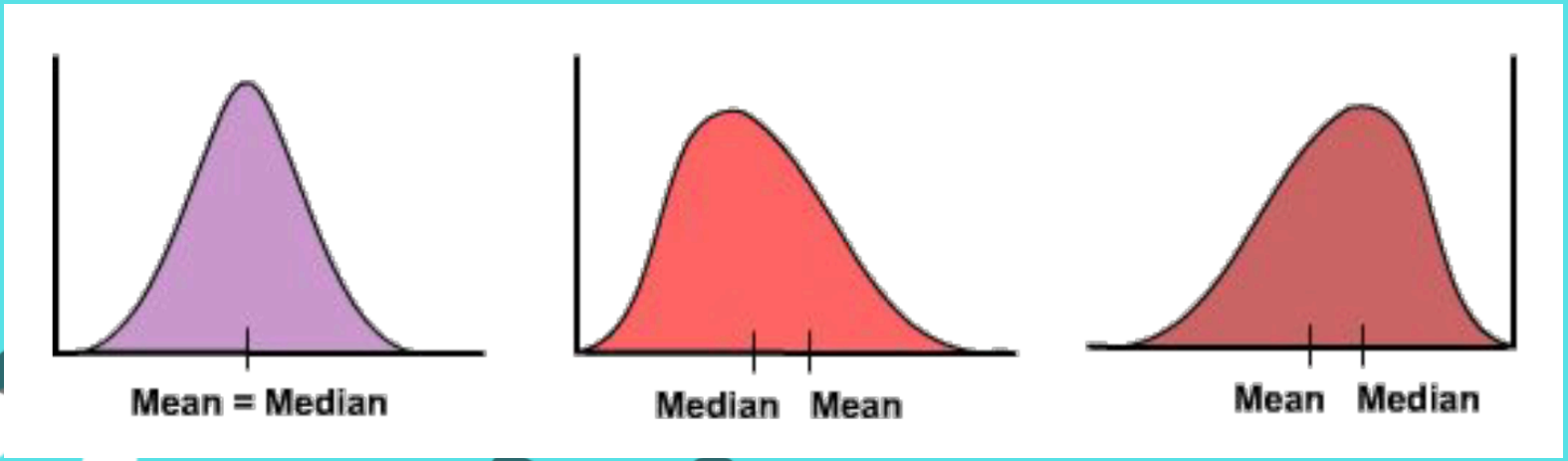
- The shape of a distribution is described by symmetry, number of peaks, direction of skew, or uniformity



Symmetric, Unimodal, Bell-shaped

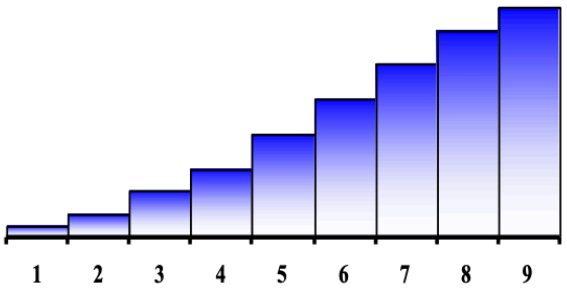
DESCRIBING SHAPE- SKEWNESS

- Skewness affects the mean the most. The mean is pulled in the same direction of the tail.
- That is why we use the median when describing the center of skewed data.
- We only use the mean to describe the center of symmetric data.

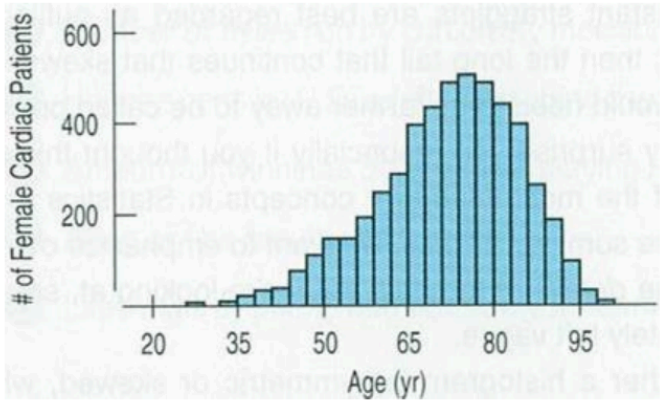


DESCRIBING SHAPE- SKEWED LEFT

- Data which is skewed LEFT will have a left "tail"

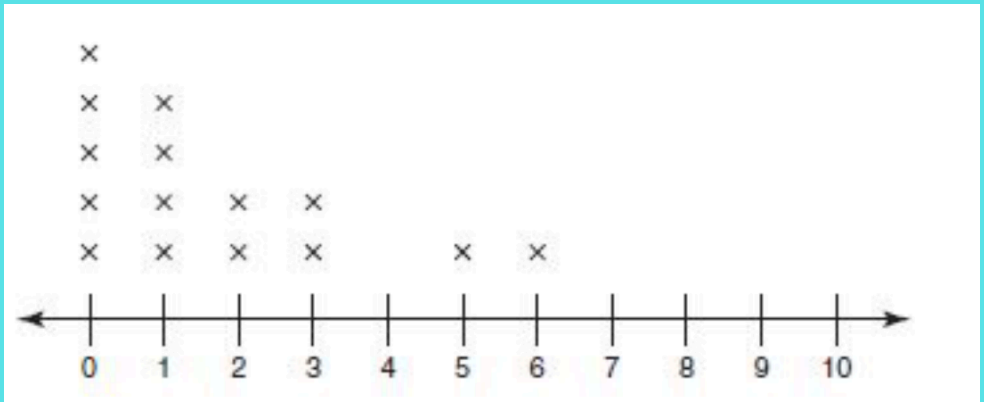


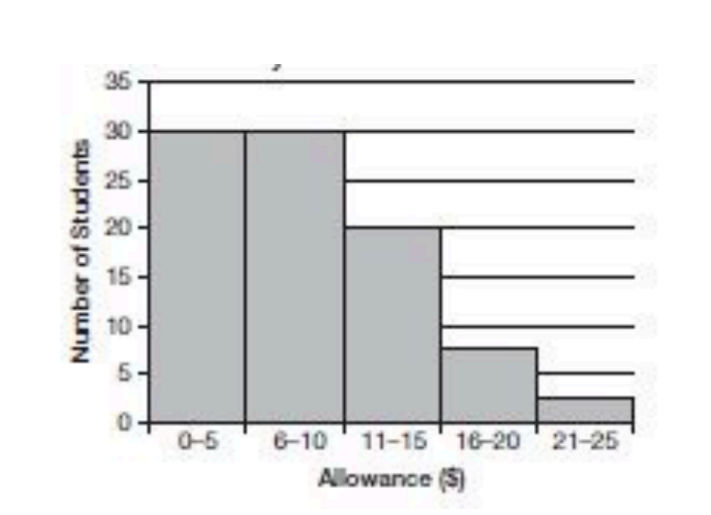
Skewed Left



DESCRIBING SHAPE- SKEWED RIGHT

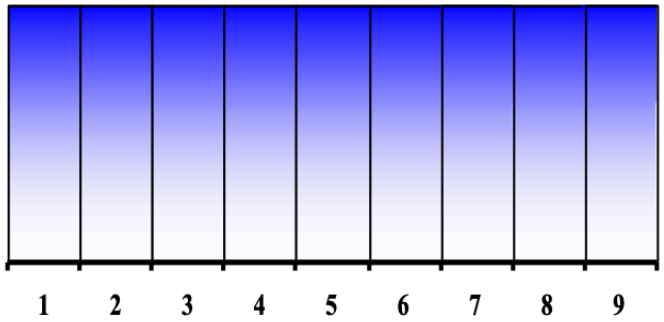
- Data which is skewed RIGHT will have a right "tail"



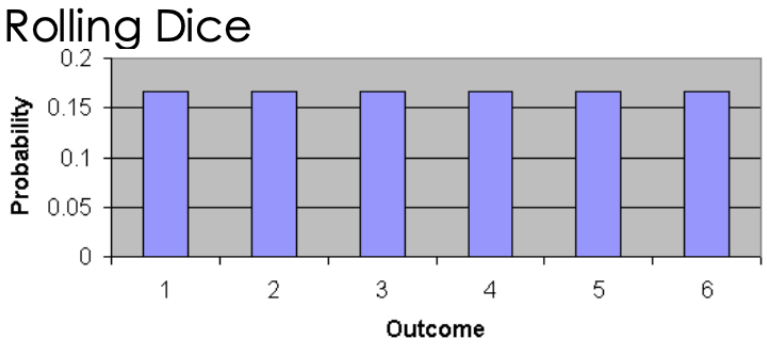


DESCRIBING: SHAPE

- Data which is skewed **UNIFORM** will all look the same. (like a football team where players all have the same UNIFORM on)

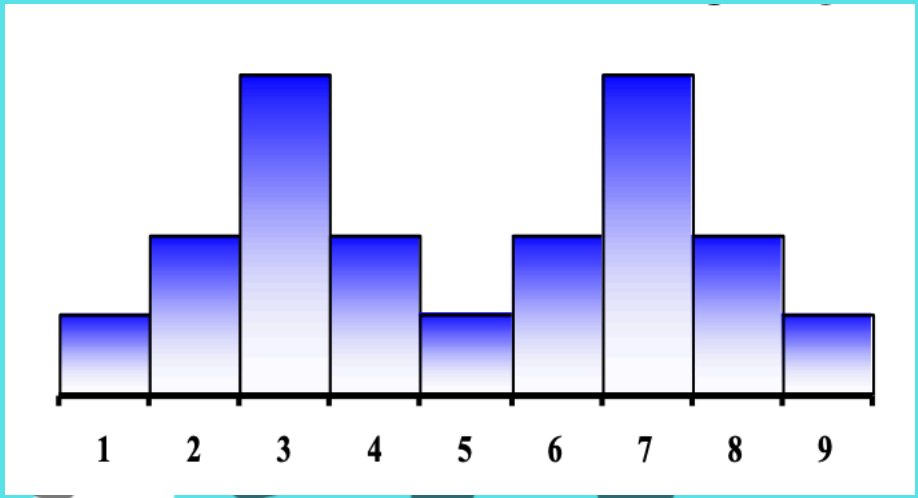


Uniform



DESCRIBING: SHAPE

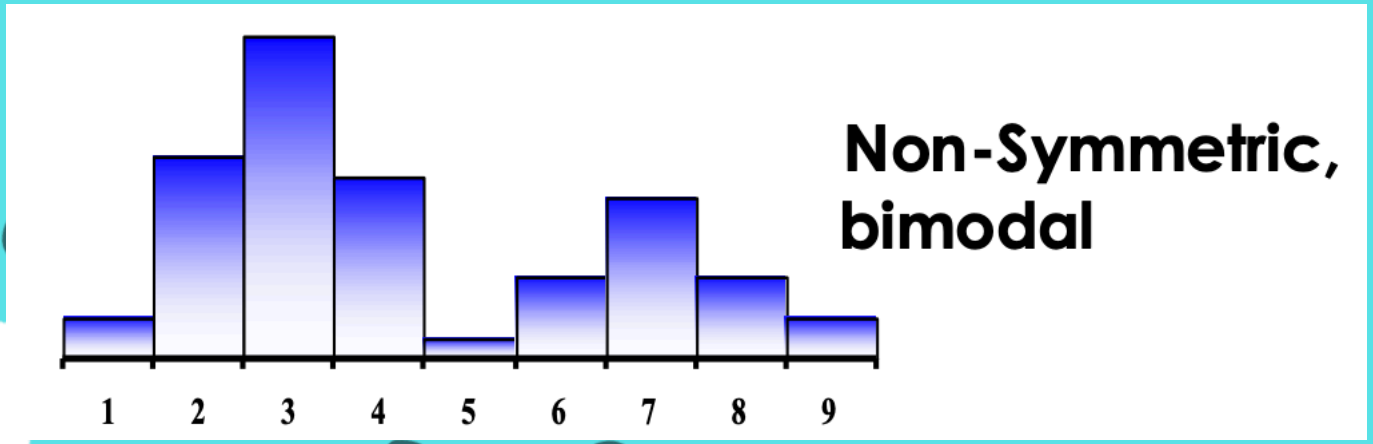
- Bimodal → "bi" means 2 → there are 2 mountains
- Symmetric - If I were to cut this in half, both sides would look the same (or close to the same)
  - A mirror image



Symmetric, Bimodal

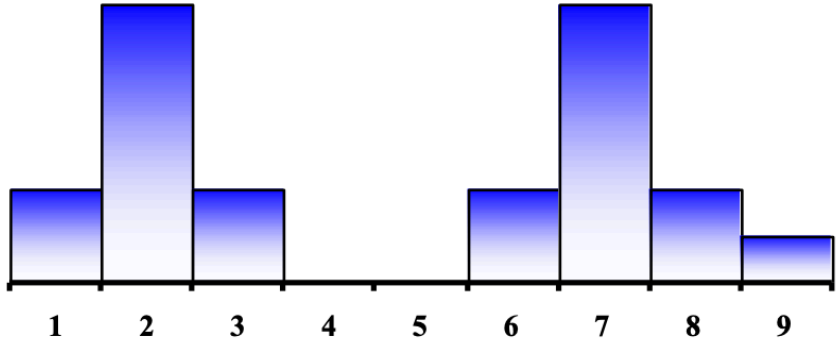
DESCRIBING: SHAPE

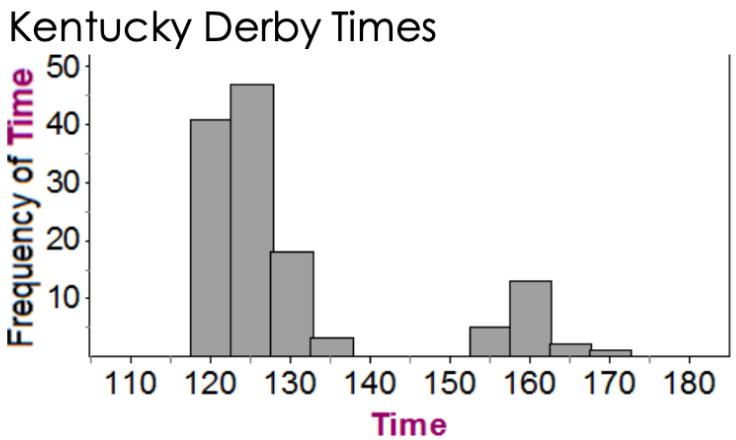
- Bimodal → "bi" means 2 → there are 2 mountains
- NOT Symmetric - If I were to cut this in half, each sides would look DIFFERENT (not a mirror image)



UNUSUAL FEATURES: GAPS

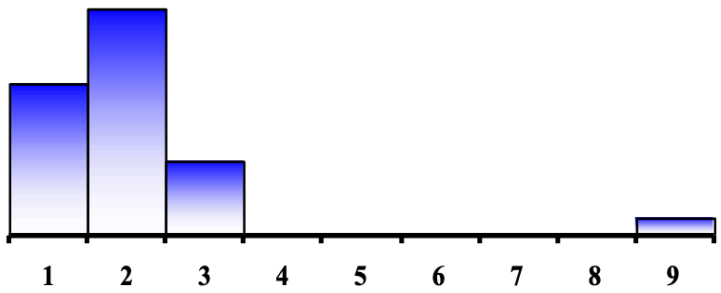
- Sometimes, statisticians refer to unusual features in a set of data.
- The two most common unusual features are gaps and outliers.

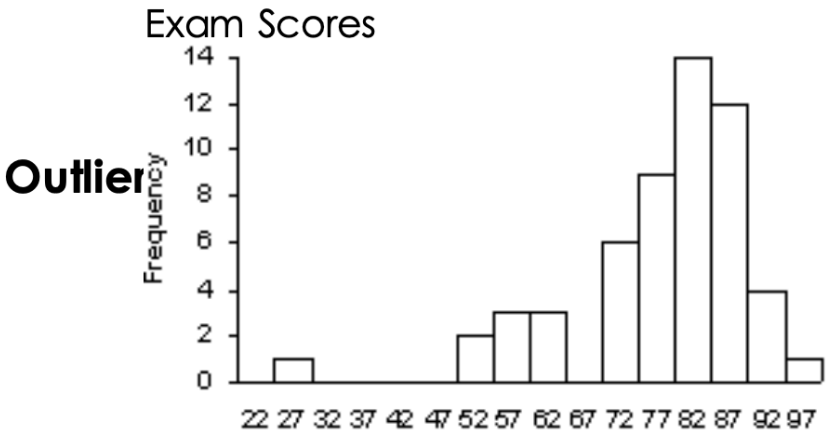




UNUSUAL FEATURES: OUTLIERS

- Sometimes, statisticians refer to unusual features in a set of data.
- The two most common unusual features are gaps and outliers.





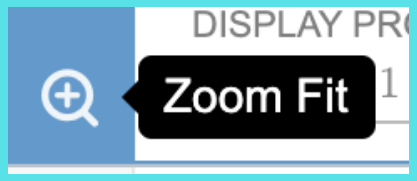
TO SUMMARIZE: MEASURES OF DATA

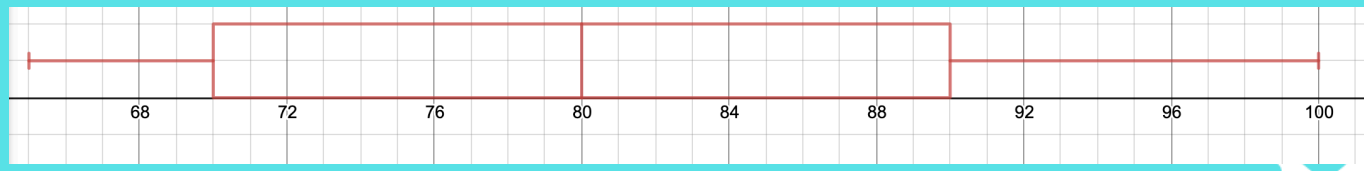
Comparing Measures of Center and Spread			
Center		**Spread (Range)**	
Mean	· Data is Symmetric · No Outliers	**More Spread**	· Data values are spread out · Greater MAD
Median	· Skewed Data · Outliers (Skewed left – mean < median) (Skewed right – mean > median)	**Less Spread**	· Data values are close together · Smaller MAD

vs. consistency

EXAMPLE IN DESMOS

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}





1
$L = [65,65,70,75,80,80,85,90,95,100]$

L = 10 element list

2
histogram(L)

Data Set, Bin Width = 1

BAR HEIGHTS

Count | Relative | Density

BIN ALIGNMENT

Center | Left

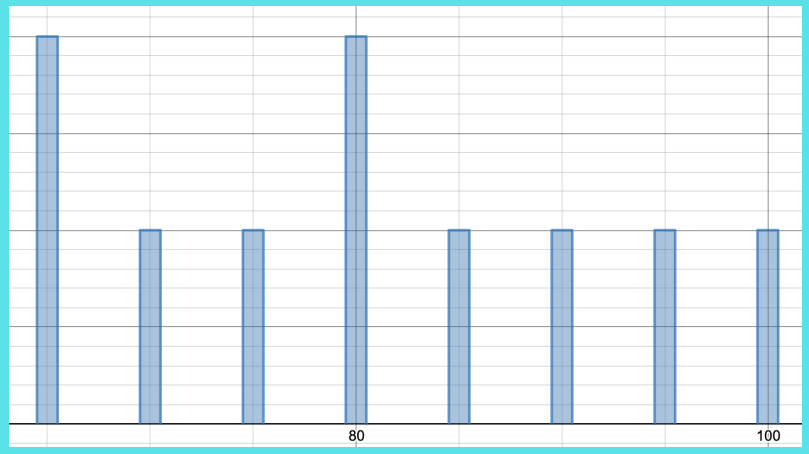
3
boxplot(L)

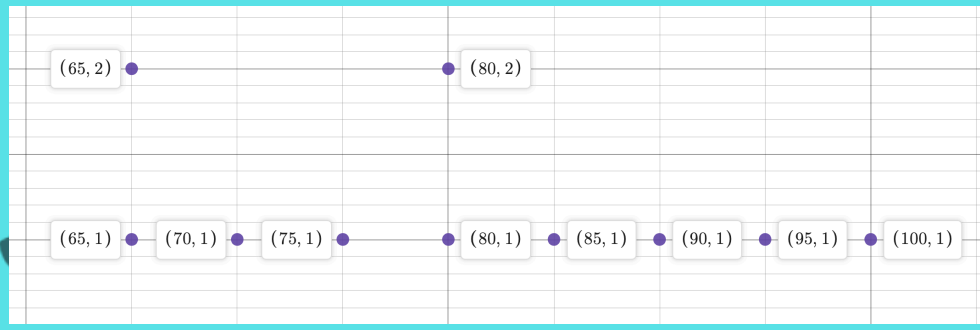
DISPLAY PROPERTIES

Offset: 1 Height: 1

4
dotplot(L)

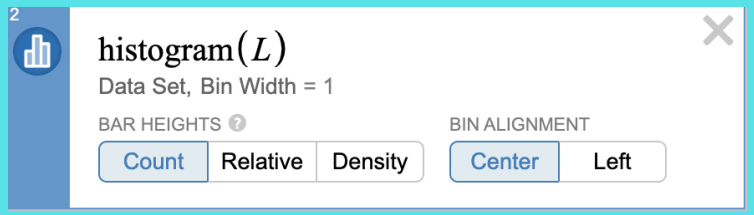
Data Set, Bin Width = 1

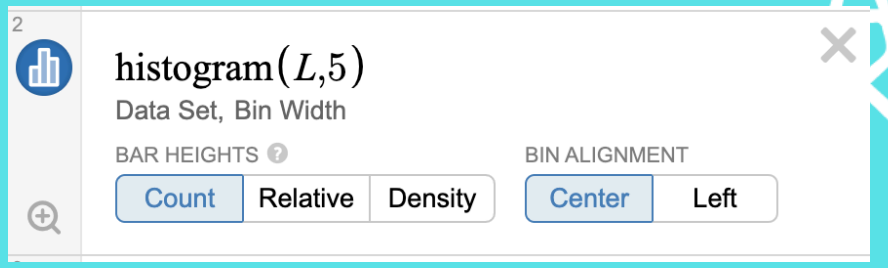


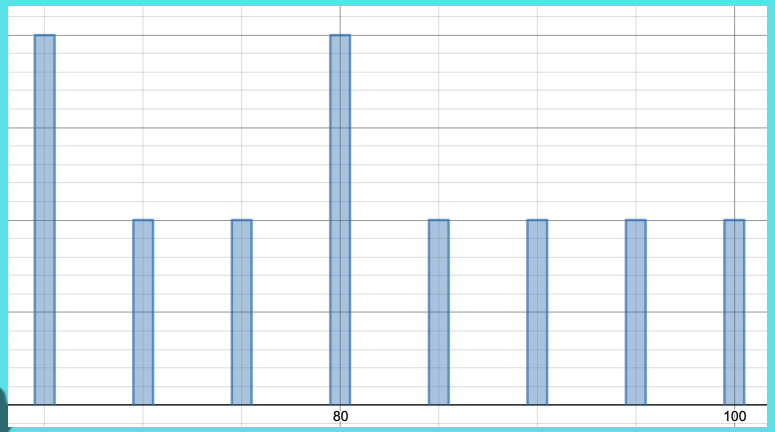


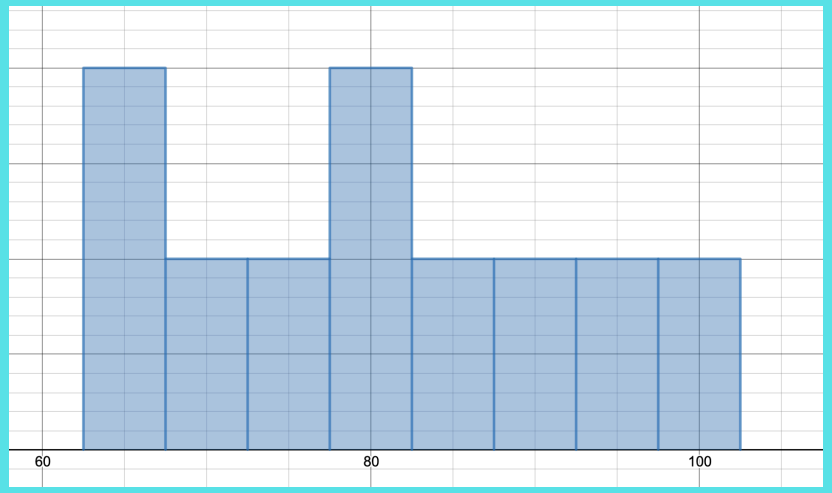
EXAMPLE IN DESMOS

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}









EXAMPLE IN DESMOS

{65, 65, 70 ,75, 80, 80, 85, 90, 95, 100}