COURSE NOTES:

DESCRIPTIVE STATISTICS

FIRST WE NEED TO KNOW THE VARIABLES ...

TYPES OF DATA

CATEGORICAL

NUMERICAL

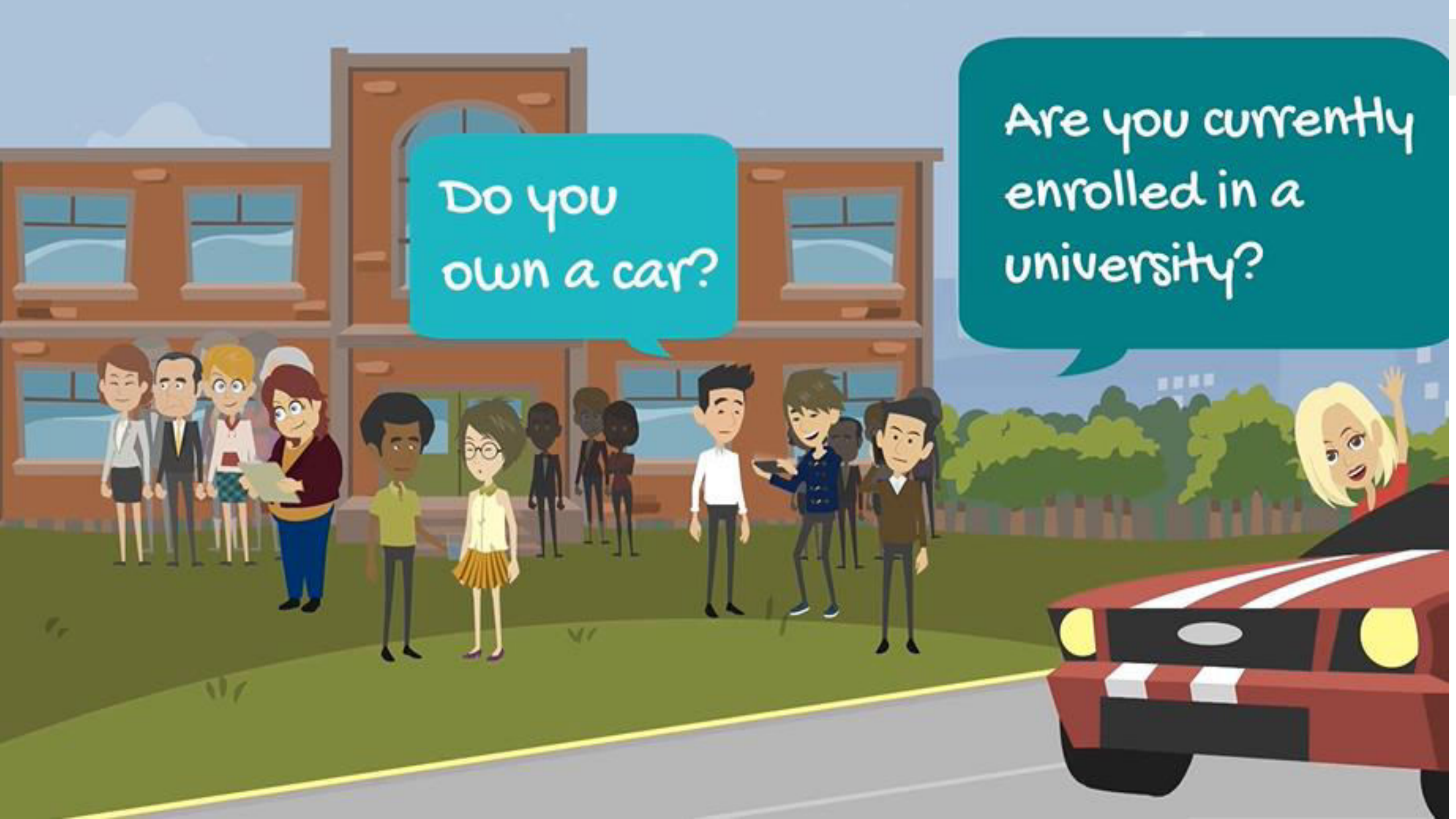
MEASUREMENT LEVELS

CATEGORICAL

Categories or groups

CAR BRANDS

YES AND NO QUESTIONS

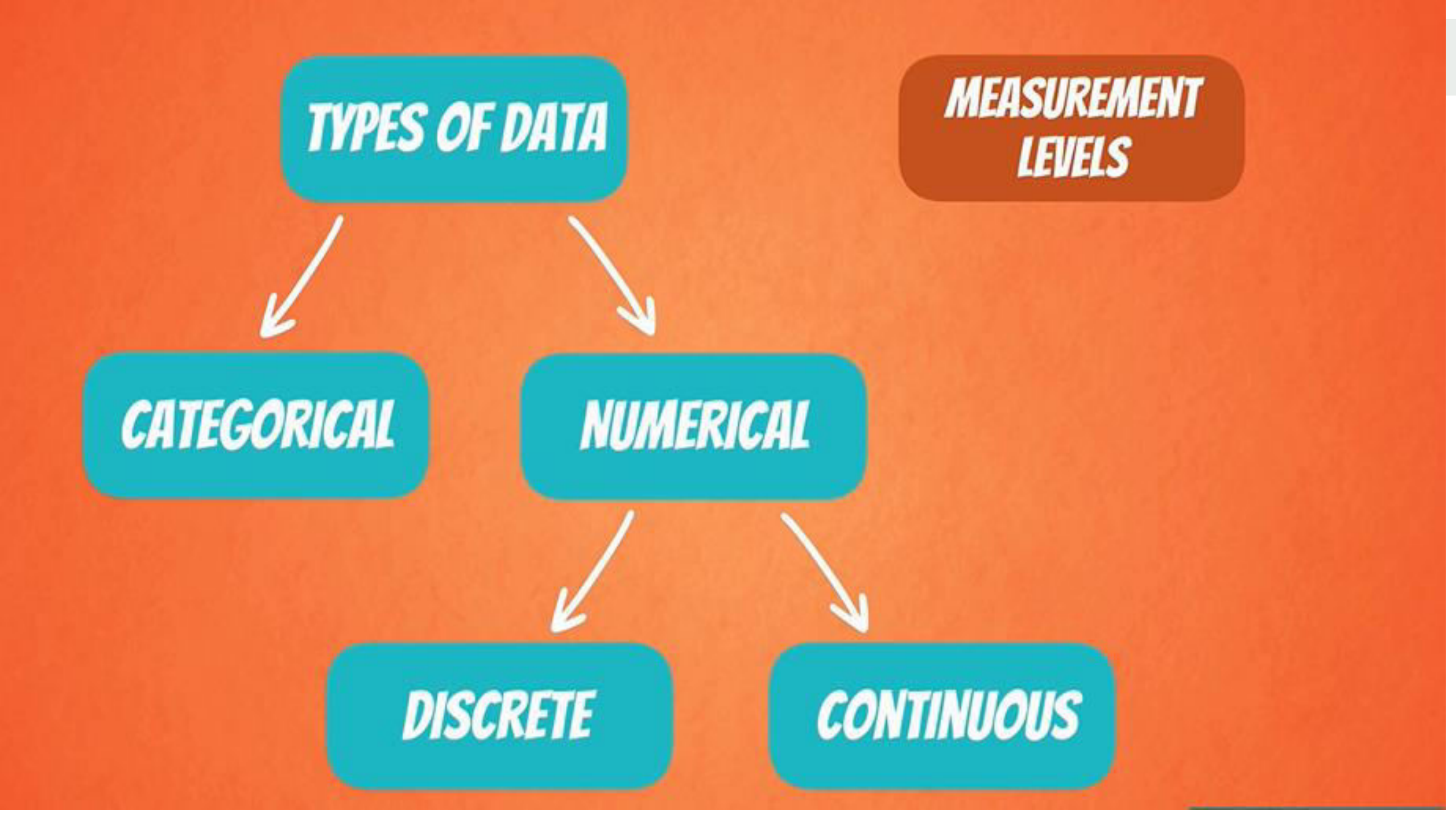

Do you own a car?
Are you currently enrolled in a university?

TYPES OF DATA

CATEGORICAL

NUMERICAL

MEASUREMENT LEVELS


TYPES OF DATA
MEASUREMENT LEVELS
CATEGORICAL
NUMERICAL
DISCRETE
CONTINUOUS

DISCRETE

of children

S.A.T score

F-

A+

1000; 1560; 1570; 2400

CONTINUOUS

68.0389 kg

CONTINUOUS

Weight

DISCRETE

of children

EXAMPLES OF DISCRETE

A, B, C,
D, E, F
or
0 to 100%

#1,000

Grades

number of objects

Money

EXAMPLES OF CONTINUOUS

Height

Area

Distance

Time

TIME ON A CLOCK IS DISCRETE

TIME IN GENERAL IS CONTINUOUS

A variable represents the weight of a person. What type of data does it represent?

1. Categorical, discrete
2. Categorical, continuous
3. Numerical, discrete
4. Numerical, continuous

Answer is

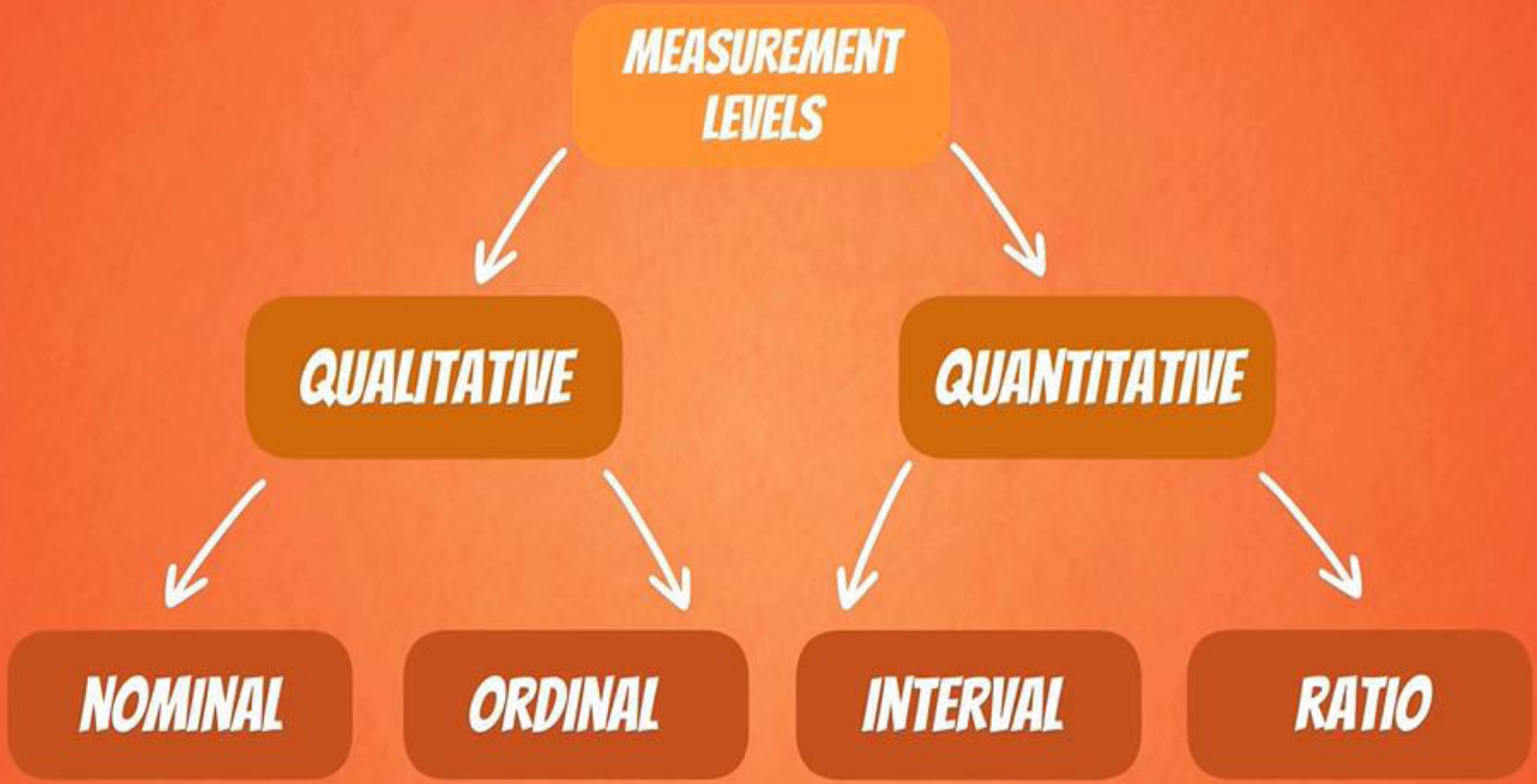
numerical, continous

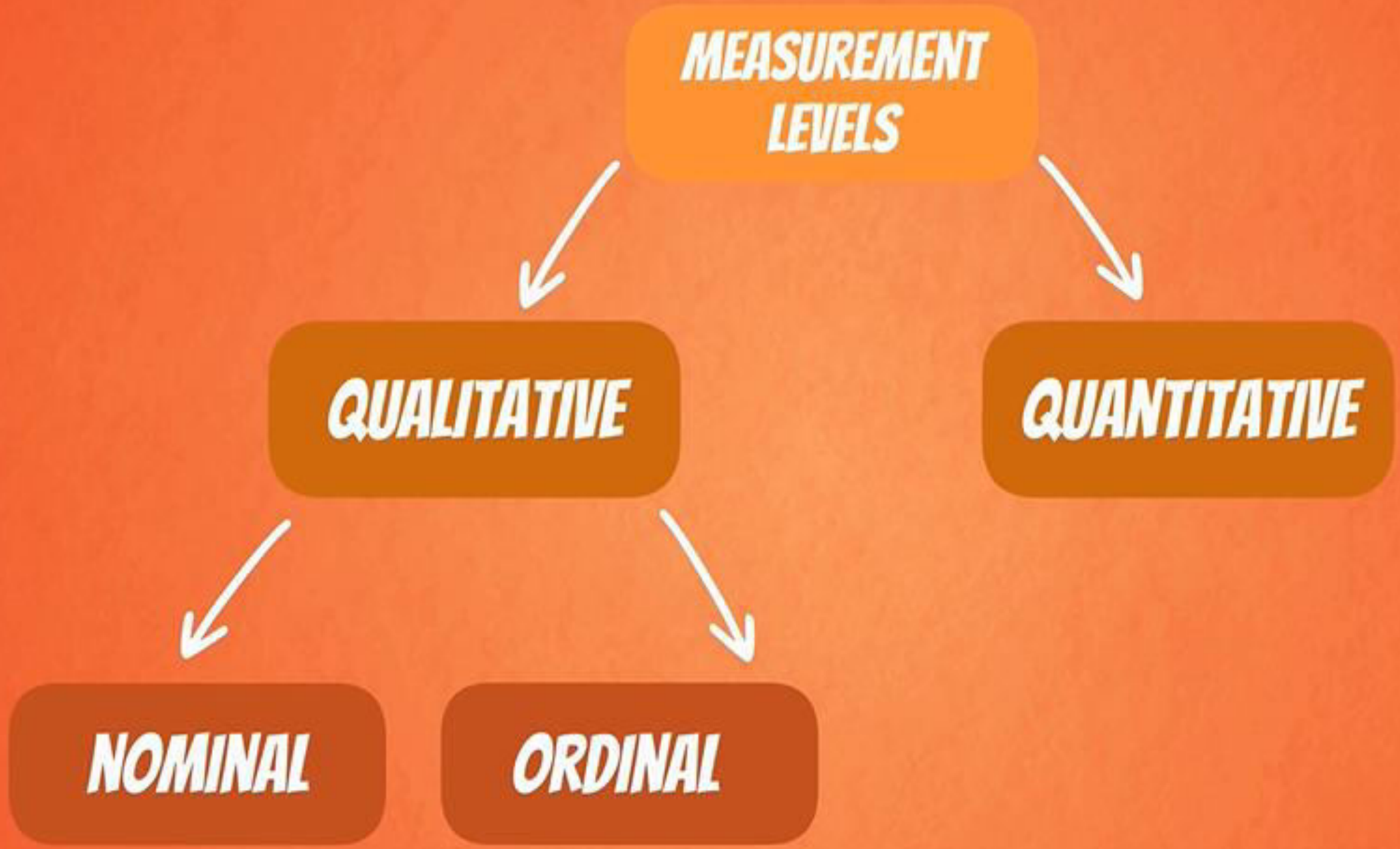
A variable represents the gender of a person.
What type of data does it represent ?

1. Categorical
2. numerical, discrete
3. Numerical, continuous

Answer is

Categorical


MEASUREMENT LEVELS
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL
INTERVAL
RATIO


MEASUREMENT LEVELS
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL

NOMINAL

1. Which of the following courses would you like to register for? (Choose any 3)

- Web development
- Product design
- Graphics design
- Content writing
- Animation

II. What toppings would you like on your pizza?

- Pepperoni
- Sausage
- Spinach
- Sardines
- Extra Cheese

ORDINAL

Disgusting Unappetizing Neutral Tasty Delicious

Ordinal Data Collection Tool

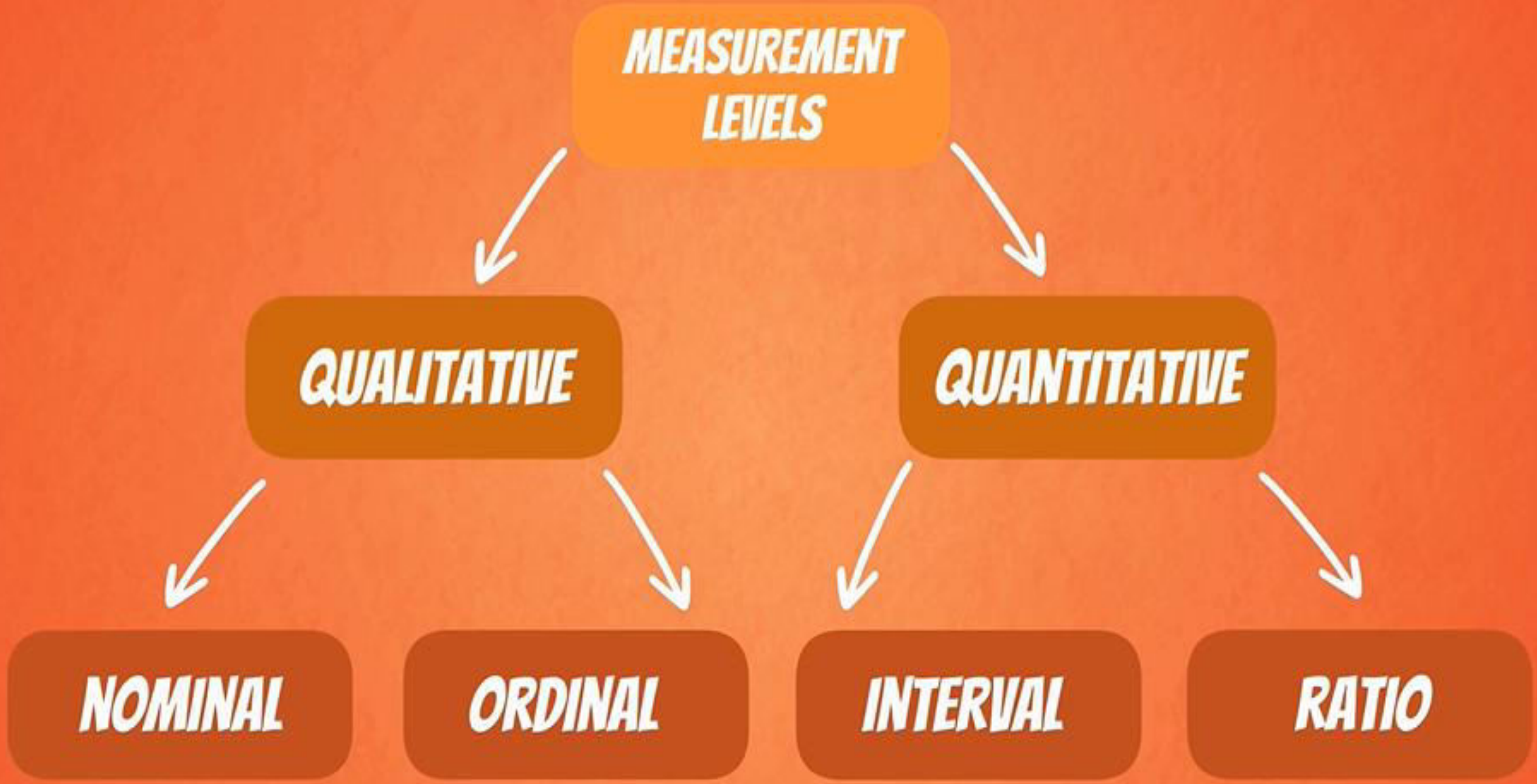
Which of the following best describes your current level of financial happiness?

- ○ Very happy
- ○ Happy
- ◉ Neutral
- ○ Unhappy
- ○ Very unhappy

Ordinal Data Collection Tool

How can you rate your knowledge of Excel?

- ○ Advanced
- ◉ Intermediate
- ○ Basic
- ○ Novice
- ○ Zero


MEASUREMENT LEVELS
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL
INTERVAL
RATIO

RATIO

HAS A TRUE 0

INTERVAL

DOESN'T HAVE A TRUE 0

RATIO

You have 3 times as many as I do.

RATIO OF 6/2 IS 3

RATIO

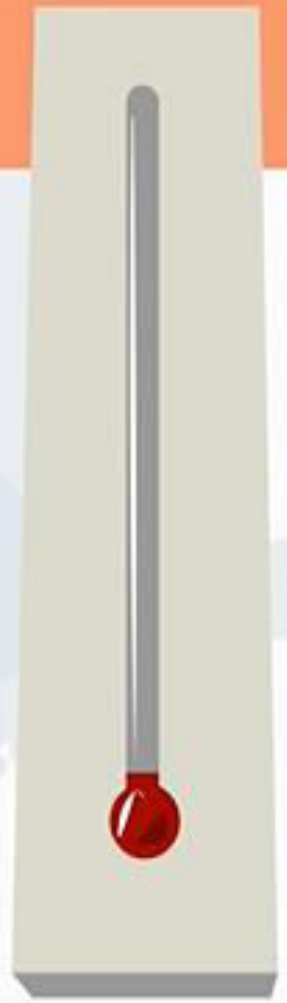
NUMBER OF OBJECTS

DISTANCE

TIME

INTERVAL

DOESN'T HAVE A TRUE 0

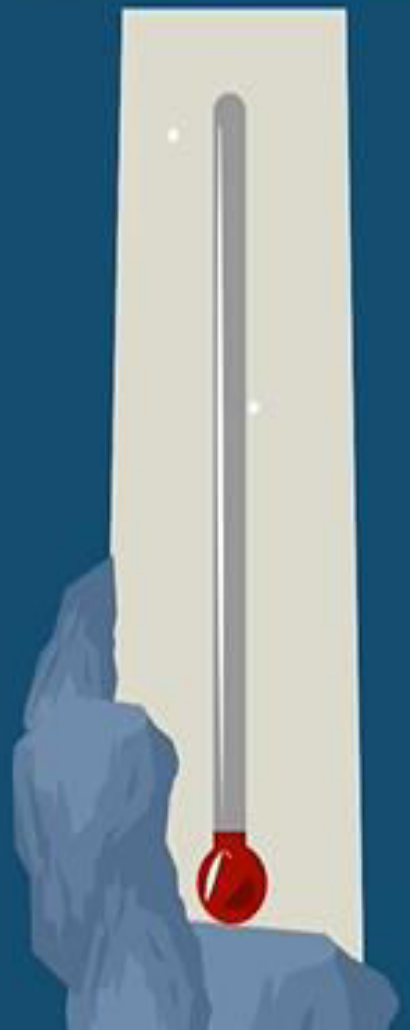



INTERVAL
TODAY
seems colder in C°
in terms of F°... not really

INTERVAL

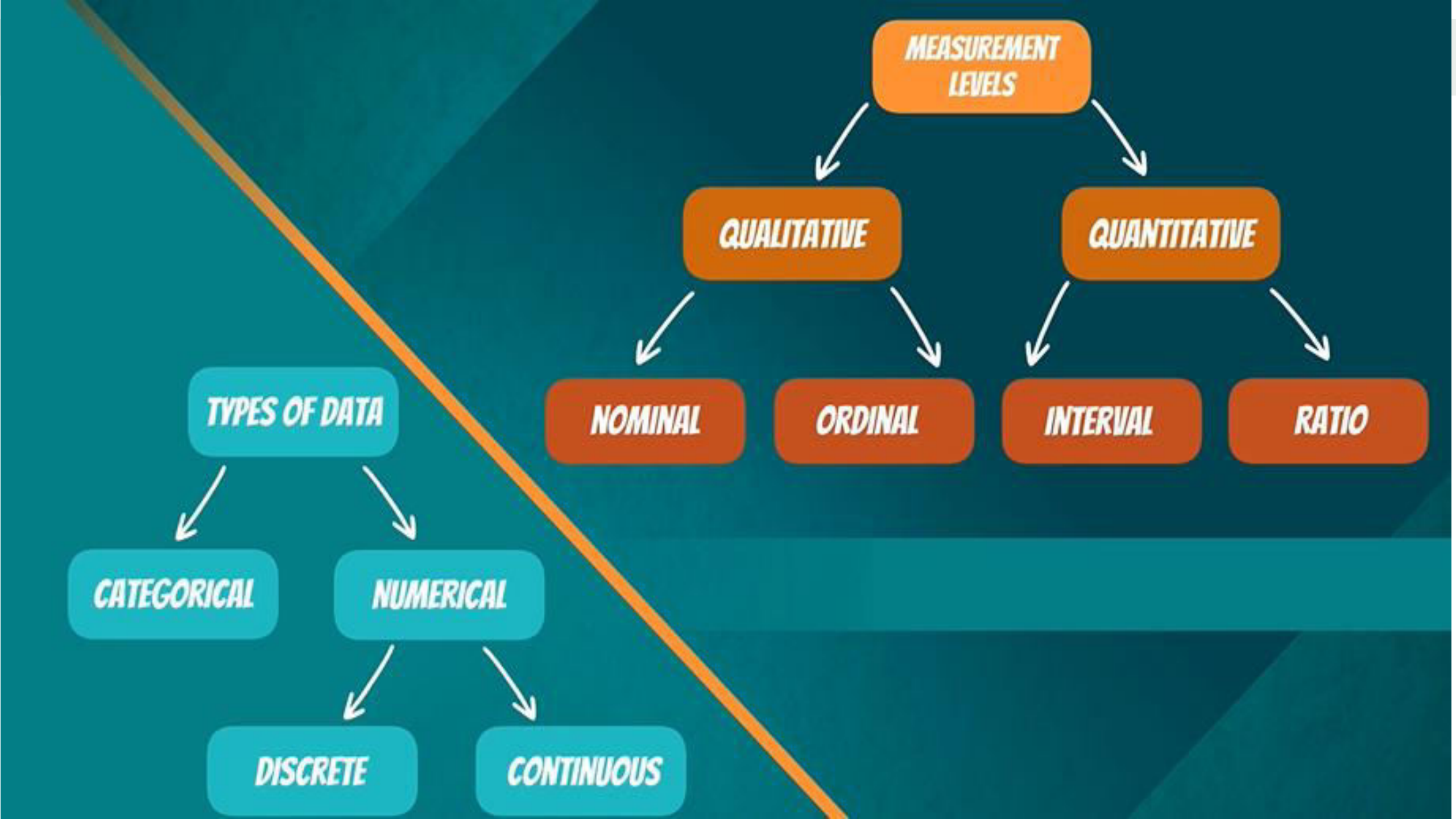
0°C AND 0°F ARE NOT TRUE ZEROS

INTERVAL



$0^{\circ}C$ AND $0^{\circ}F$ ARE NOT TRUE ZEROS

$0^{\circ}K = -273.15^{\circ}C = -459.67^{\circ}F$


MEASUREMENT LEVELS
QUALITATIVE
QUANTITATIVE
NOMINAL
ORDINAL
INTERVAL
RATIO
TYPES OF DATA
CATEGORICAL
NUMERICAL
DISCRETE
CONTINUOUS

Question 1:

A variable represents the gender of a person. What is the level of measurement?

- ○ nominal
- ○ ordinal
- ○ interval
- ○ ratio

Question 2:

A variable represents the weight of a person. What is the level of measurement?

- nominal
- ordinal
- interval
- ratio

VISUALIZING DATA IS THE MOST INTUITIVE WAY TO INTERPRET IT

TYPES OF DATA

MEASUREMENT LEVELS

CATEGORICAL

NUMERICAL

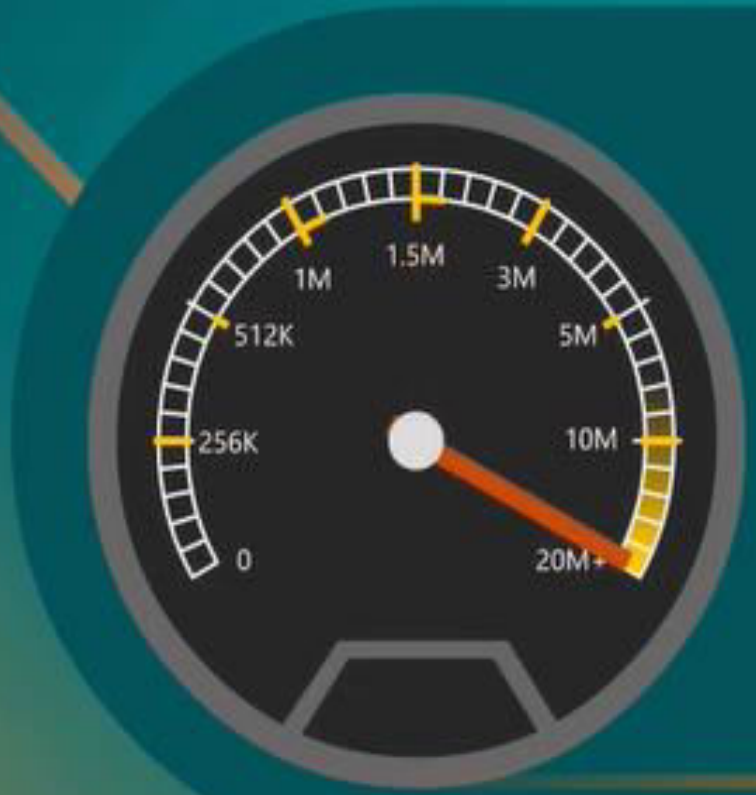
REPRESENTATION OF CATEGORICAL VARIABLES

Frequency distribution tables

Bar charts

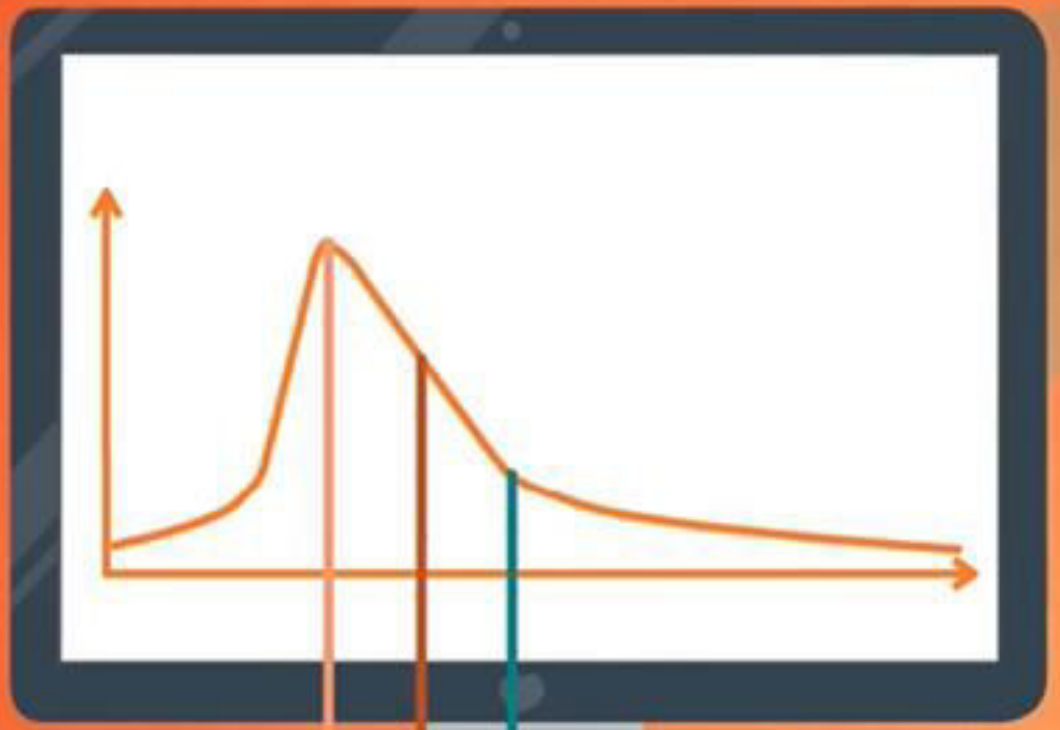
Pie charts

Pareto diagrams



MEASURES OF CENTRAL TENDENCY

MEASURES OF CENTRAL TENDENCY



Mean

Median

Mode

MEASURES OF CENTRAL TENDENCY

a.k.a. simple average

MEAN

Population μ

Sample $\bar{x}$

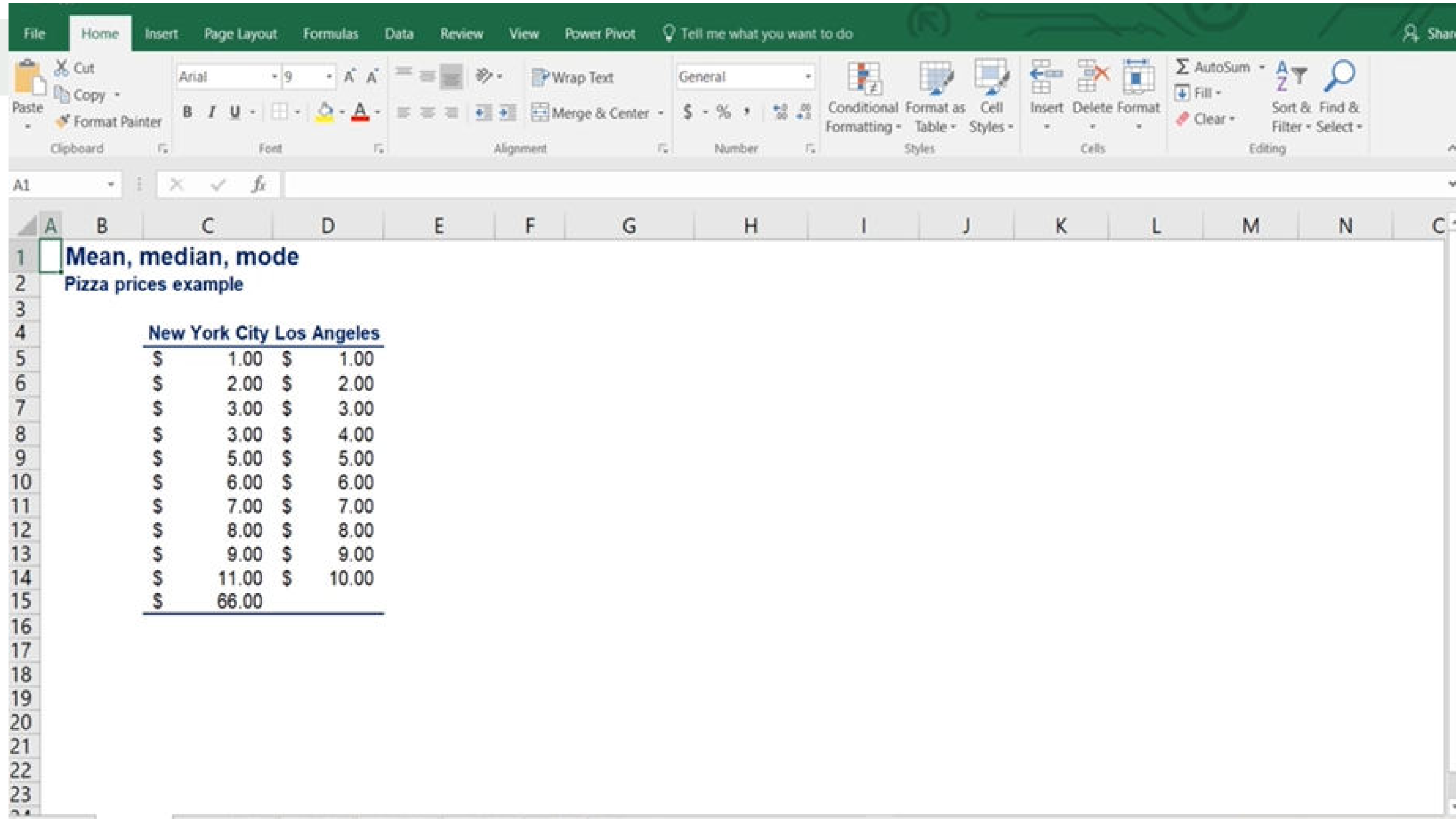
HOW DO WE FIND THE MEAN

$$\frac{\sum_{i=1}^{N} x_i}{N}$$

By adding up all the components and then dividing by the number of components

or

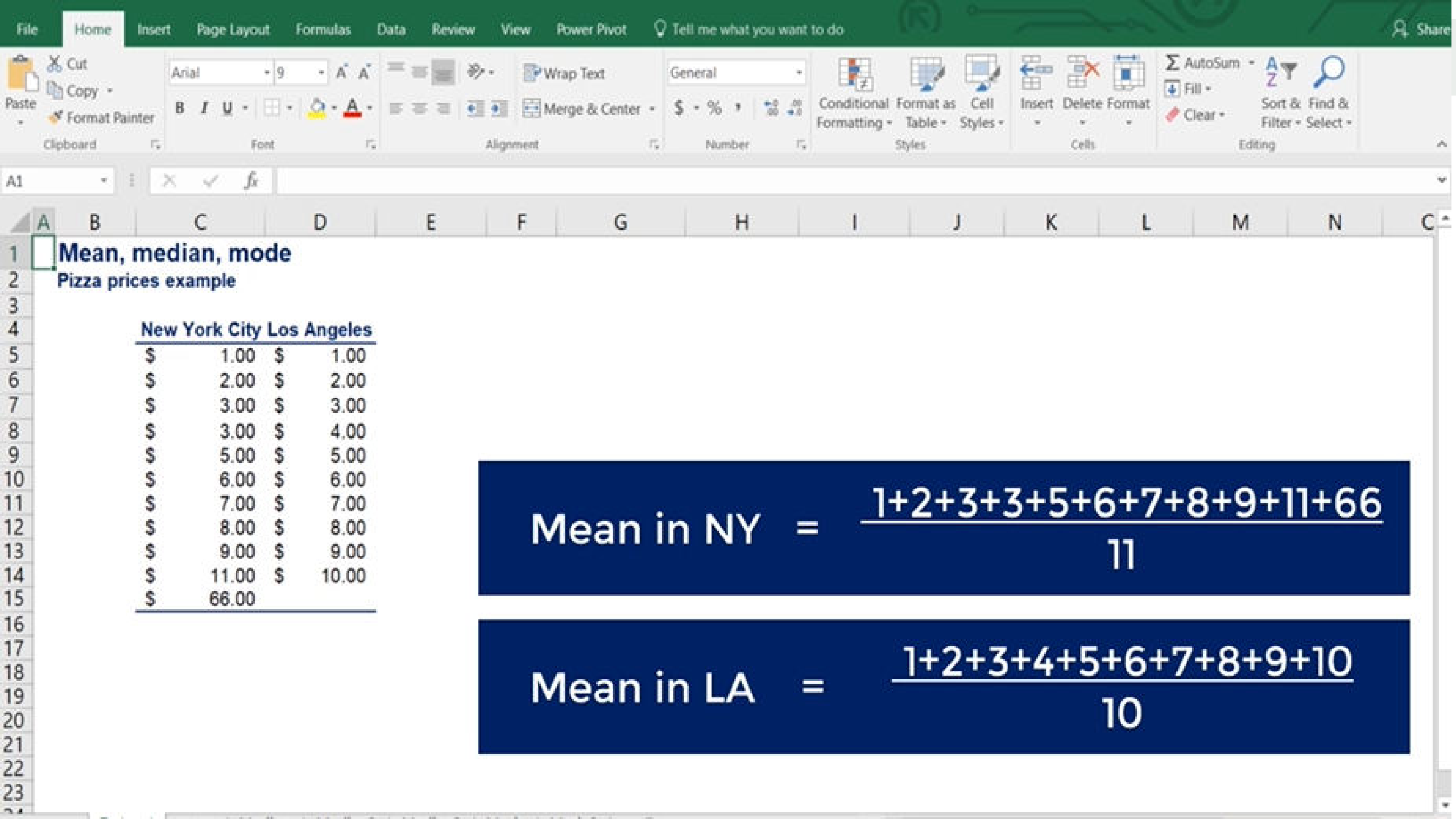
$$\frac{x_1 + x_2 + x_3 + \ldots + x_{N-1} + x_N}{N}$$

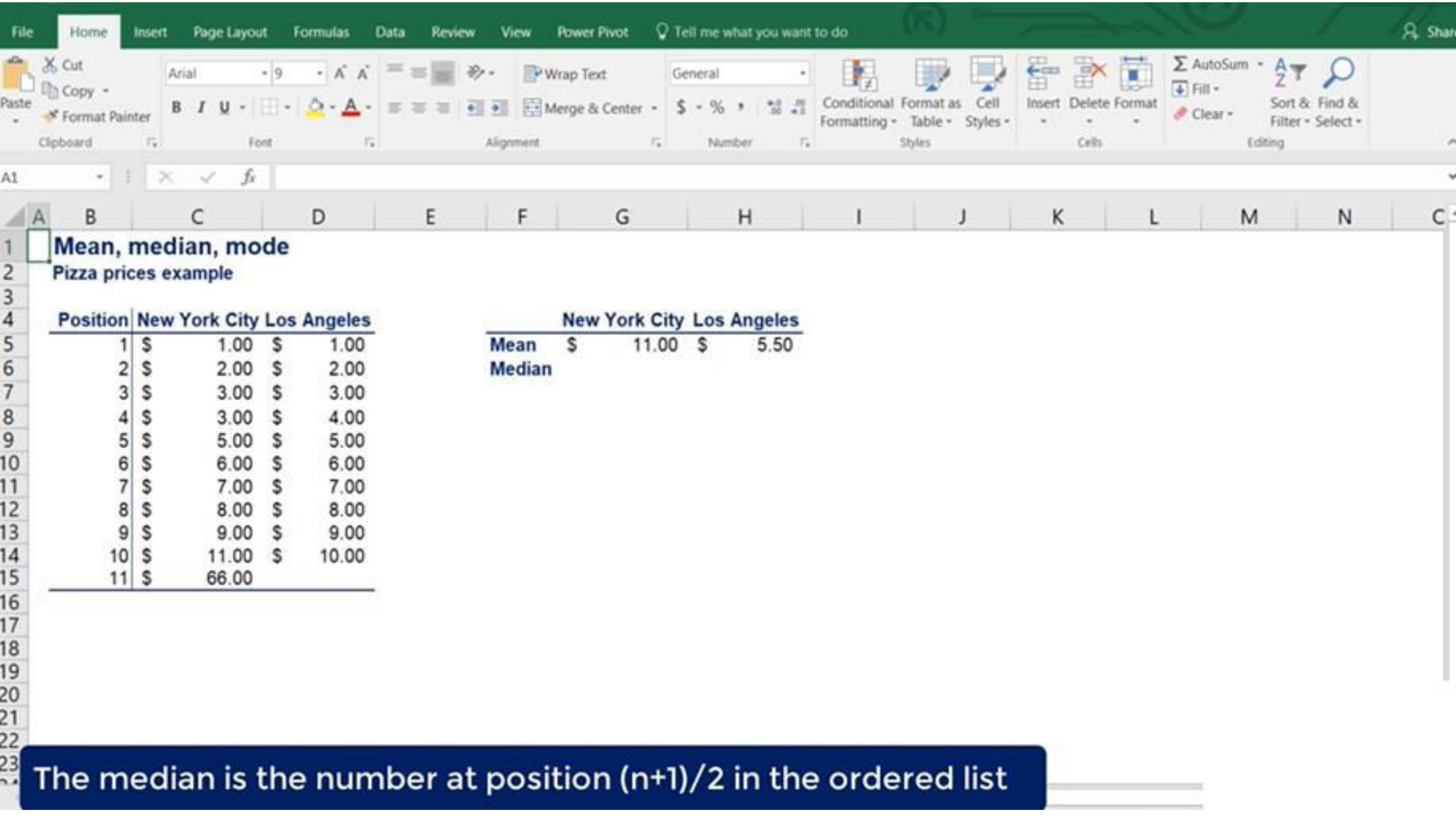


Mean, median, mode

Pizza prices example

New York City		Los Angeles	
$	1.00	$	1.00
$	2.00	$	2.00
$	3.00	$	3.00
$	3.00	$	4.00
$	5.00	$	5.00
$	6.00	$	6.00
$	7.00	$	7.00
$	8.00	$	8.00
$	9.00	$	9.00
$	11.00	$	10.00
$	66.00		


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Mean, median, mode
Pizza prices example
New York City Los Angeles
$ 1.00 $ 1.00
$ 2.00 $ 2.00
$ 3.00 $ 3.00
$ 3.00 $ 4.00
$ 5.00 $ 5.00
$ 6.00 $ 6.00
$ 7.00 $ 7.00
$ 8.00 $ 8.00
$ 9.00 $ 9.00
$ 11.00 $ 10.00
$ 66.00
Mean in NY = 1+2+3+3+5+6+7+8+9+11+66 / 11
Mean in LA = 1+2+3+4+5+6+7+8+9+10 / 10



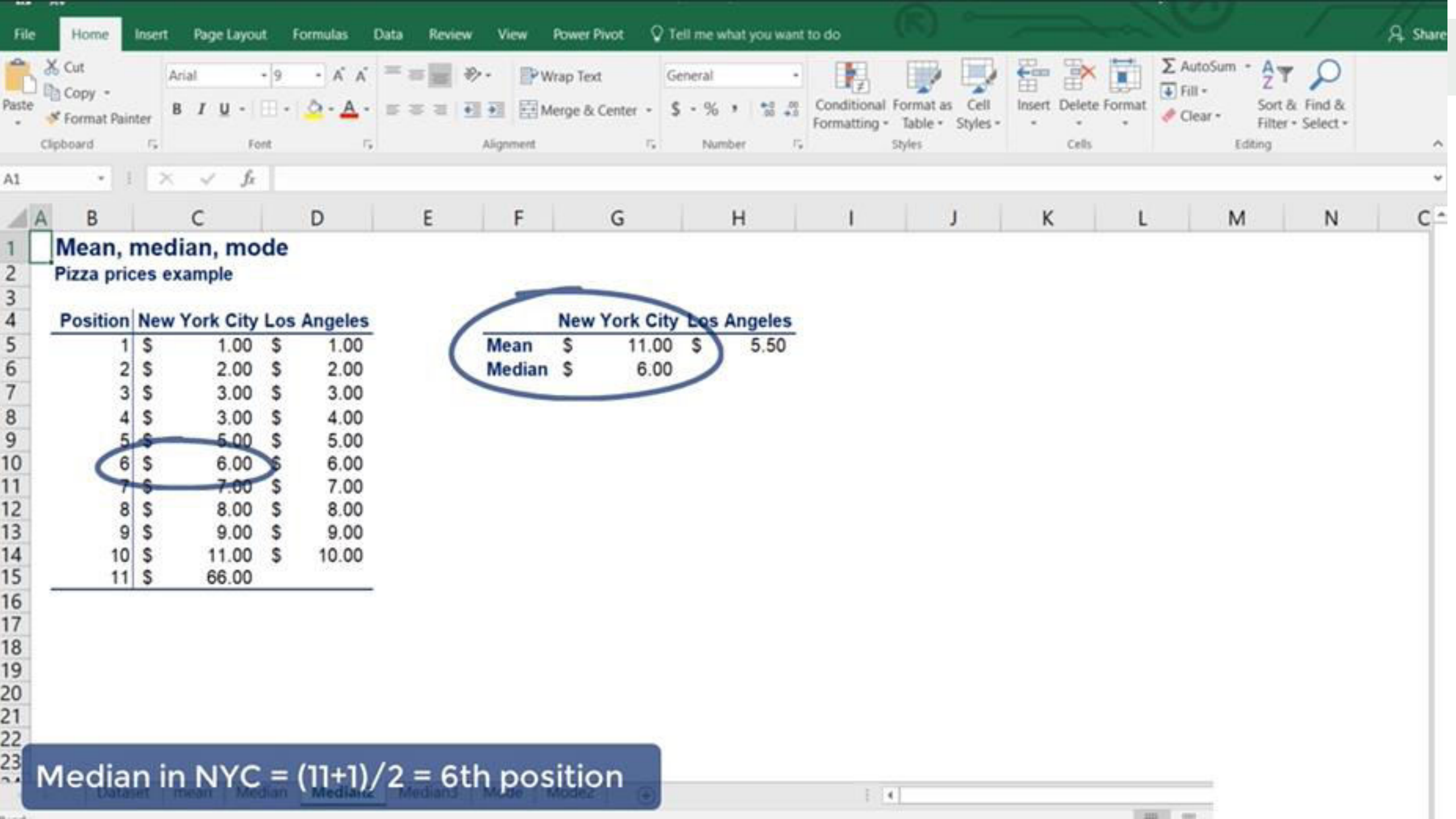
Mean, median, mode

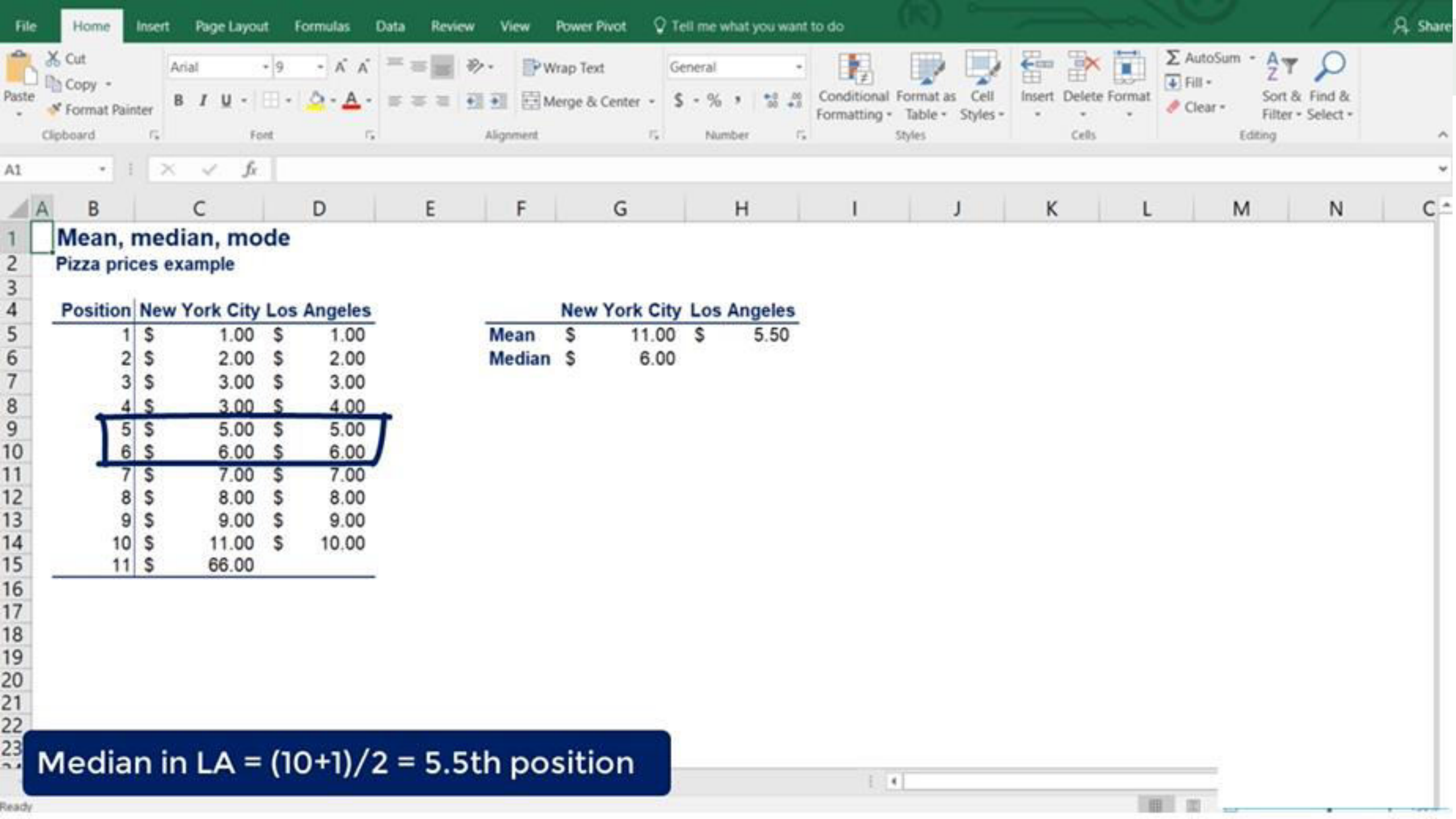
Pizza prices example

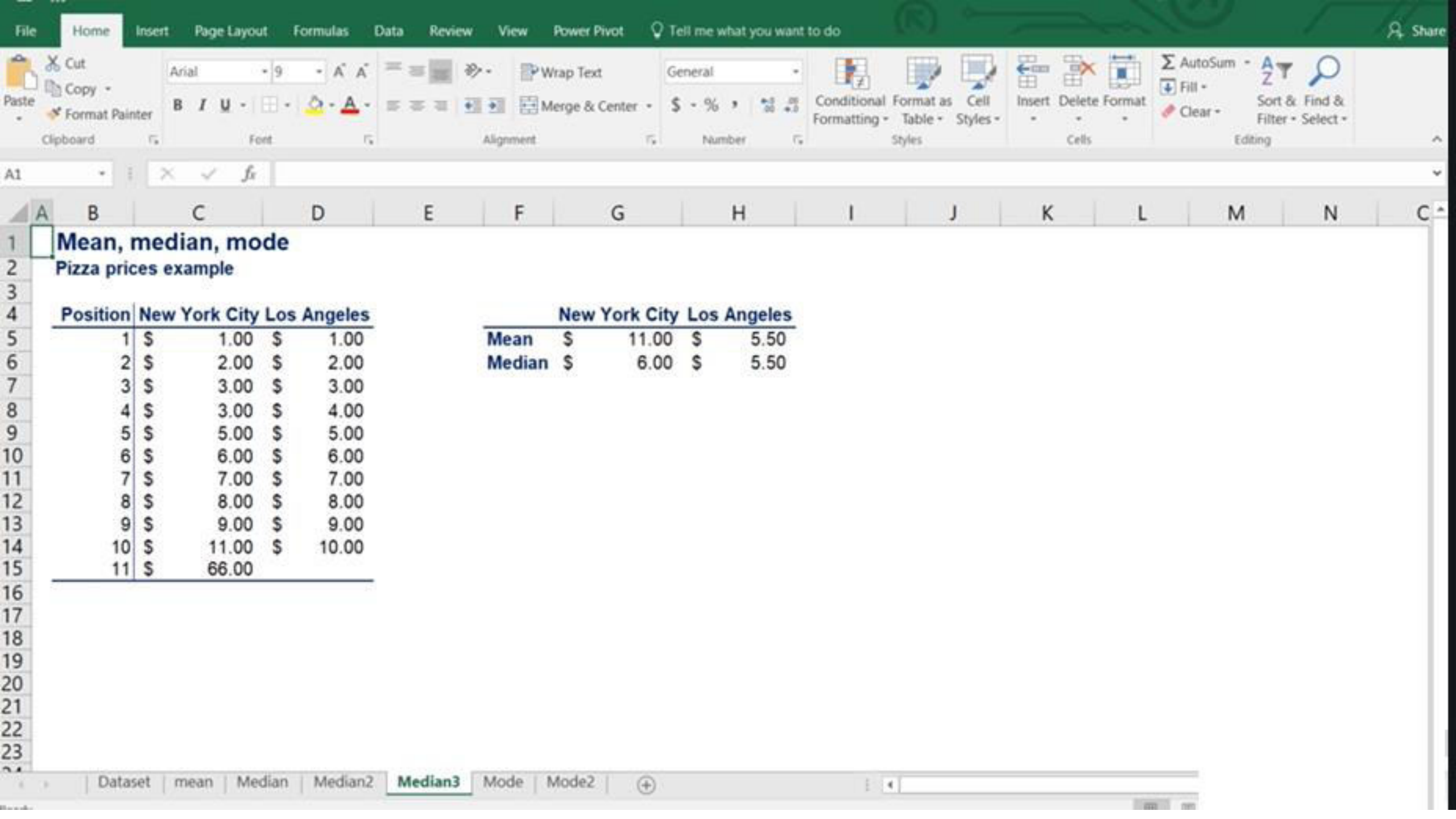
Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median		

The median is the number at position (n+1)/2 in the ordered list


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Mean, median, mode
Pizza prices example
Position | New York City | Los Angeles
1 | $ 1.00 | $ 1.00
2 | $ 2.00 | $ 2.00
3 | $ 3.00 | $ 3.00
4 | $ 3.00 | $ 4.00
5 | $ 5.00 | $ 5.00
6 | $ 6.00 | $ 6.00
7 | $ 7.00 | $ 7.00
8 | $ 8.00 | $ 8.00
9 | $ 9.00 | $ 9.00
10 | $ 11.00 | $ 10.00
11 | $ 66.00
New York City | Los Angeles
Mean | $ 11.00 | $ 5.50
Median | $ 6.00
Median in NYC = (11+1)/2 = 6th position


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Mean, median, mode
Pizza prices example
Position
New York City
Los Angeles
1 $ 1.00 $ 1.00
2 $ 2.00 $ 2.00
3 $ 3.00 $ 3.00
4 $ 4.00 $ 4.00
5 $ 5.00 $ 5.00
6 $ 6.00 $ 6.00
7 $ 7.00 $ 7.00
8 $ 8.00 $ 8.00
9 $ 9.00 $ 9.00
10 $ 11.00 $ 10.00
11 $ 66.00
Mean $ 11.00 $ 5.50
Median $ 6.00
Median in LA = (10+1)/2 = 5.5th position
Ready



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste Clipboard

Arial 9 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting Format as Table Cell Styles Styles

Insert Delete Format Cells

AutoSum Fill Clear Sort & Filter Find & Select Editing

A1

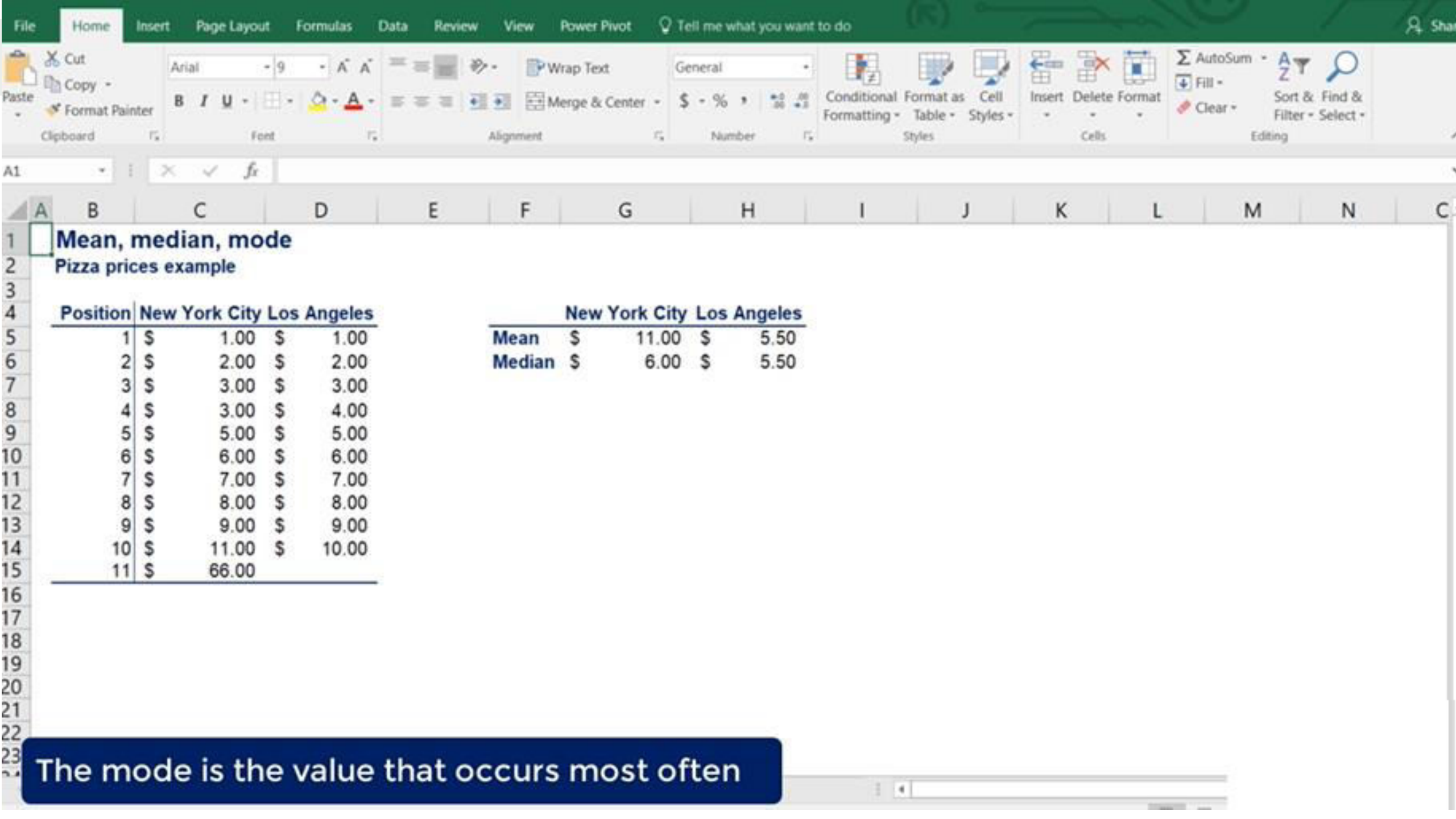
Mean, median, mode

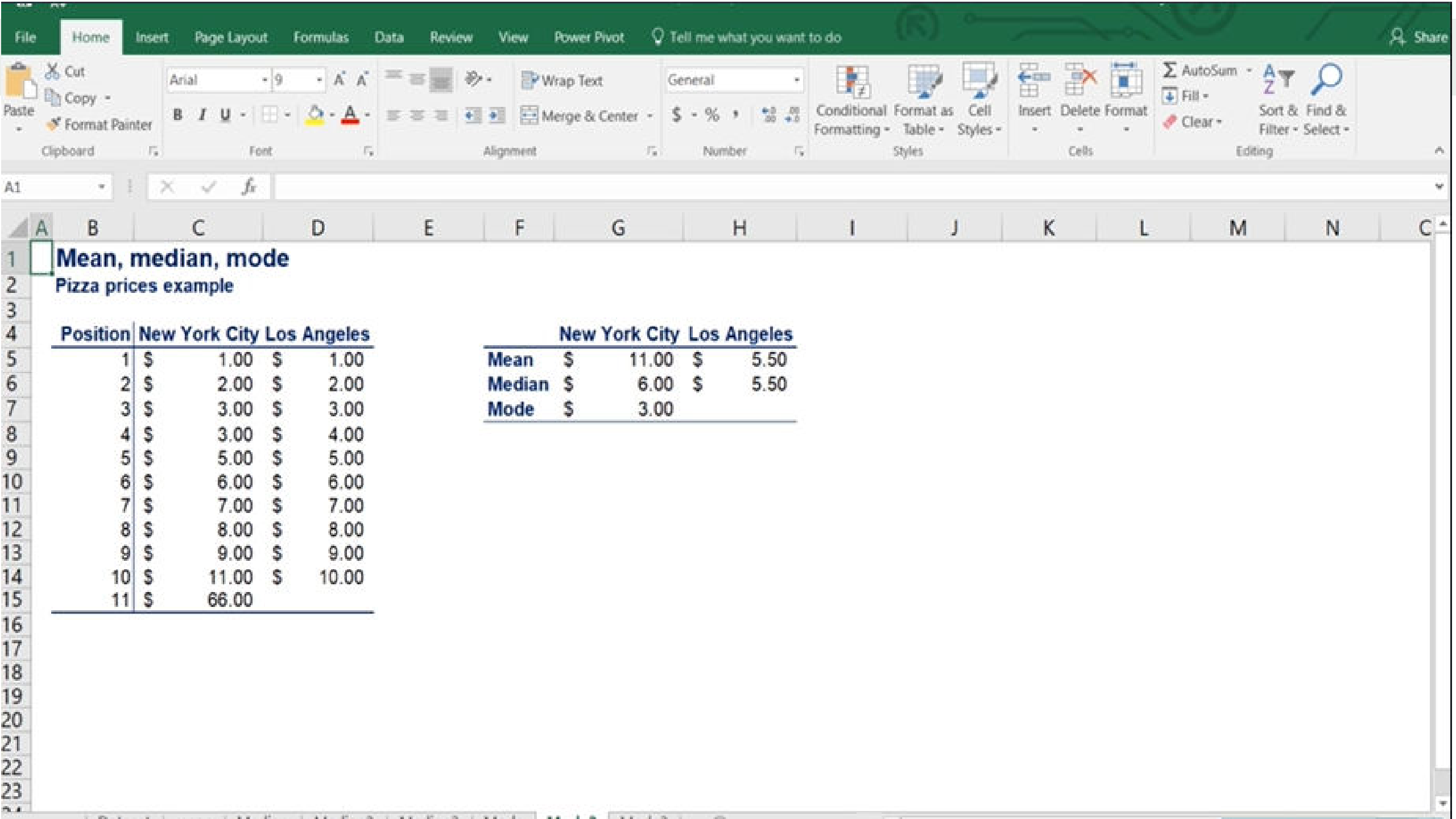
Pizza prices example

Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median	$ 6.00	$ 5.50

Dataset mean Median Median2 Median3 Mode Mode2


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Mean, median, mode
Pizza prices example
Position | New York City | Los Angeles
1 $ 1.00 $ 1.00
2 $ 2.00 $ 2.00
3 $ 3.00 $ 3.00
4 $ 3.00 $ 4.00
5 $ 5.00 $ 5.00
6 $ 6.00 $ 6.00
7 $ 7.00 $ 7.00
8 $ 8.00 $ 8.00
9 $ 9.00 $ 9.00
10 $ 11.00 $ 10.00
11 $ 66.00
New York City Los Angeles
Mean $ 11.00 $ 5.50
Median $ 6.00 $ 5.50
The mode is the value that occurs most often

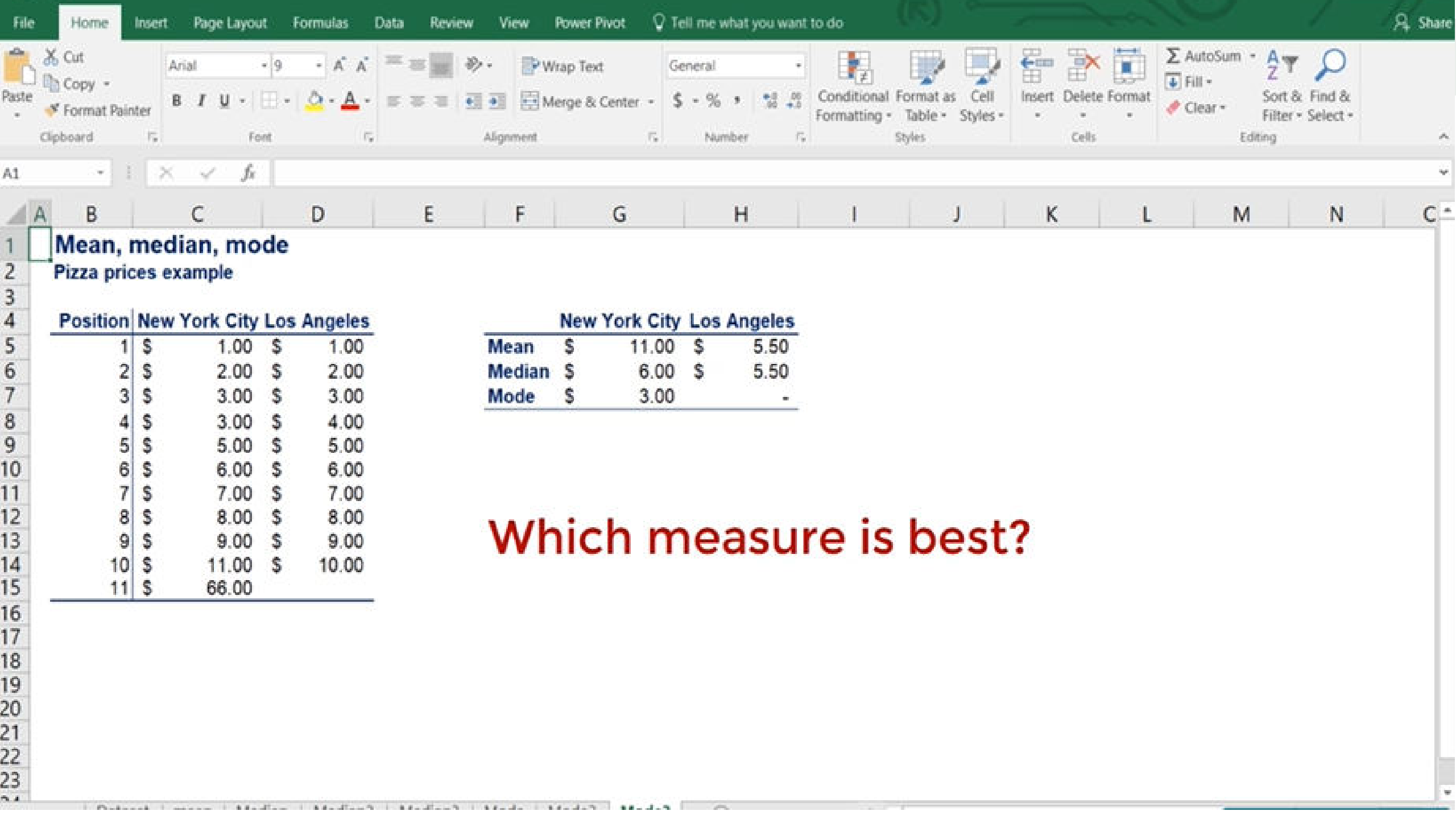


Mean, median, mode

Pizza prices example

Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median	$ 6.00	$ 5.50
Mode	$ 3.00	



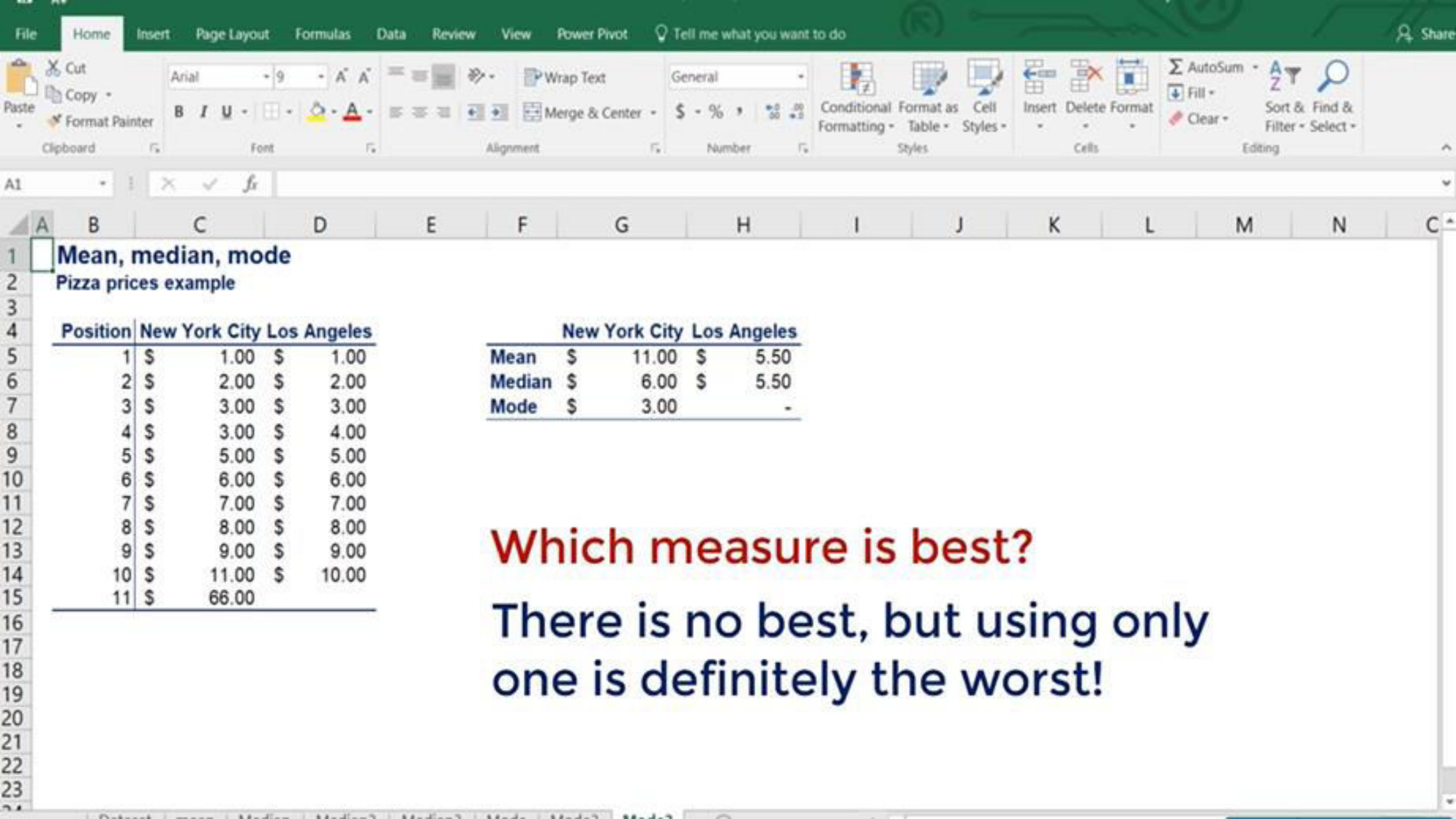
Mean, median, mode

Pizza prices example

Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50
Median	$ 6.00	$ 5.50
Mode	$ 3.00	-

Which measure is best?

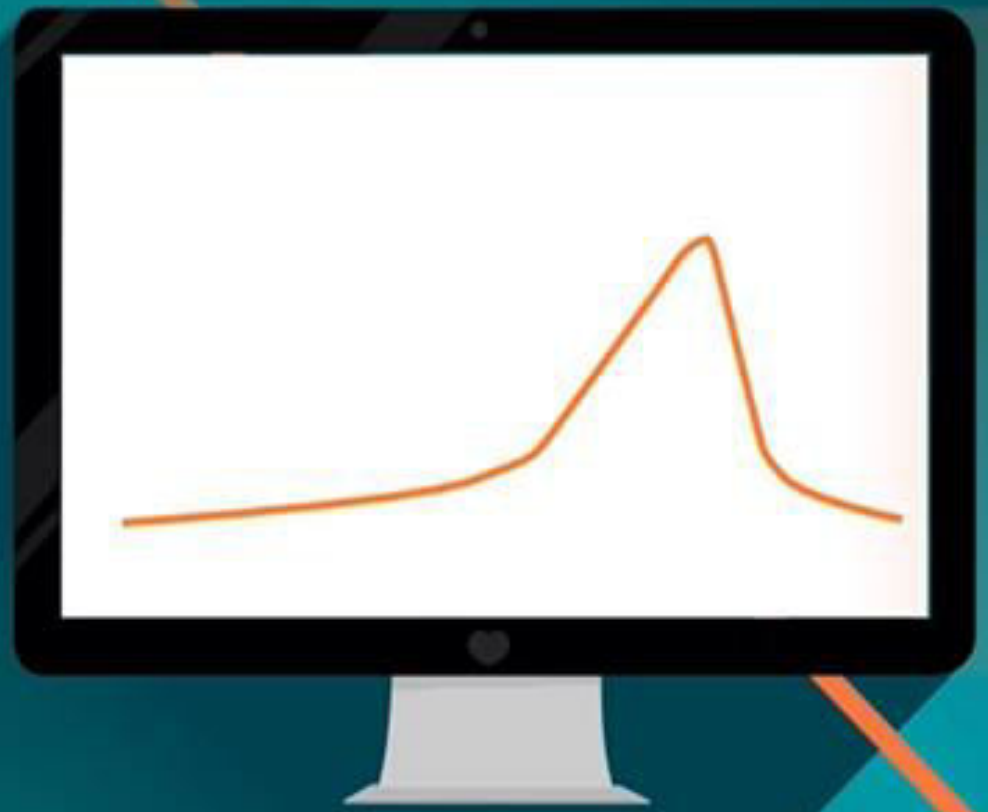

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Mean, median, mode
Pizza prices example
Position
New York City
Los Angeles
1 $ 1.00 $ 1.00
2 $ 2.00 $ 2.00
3 $ 3.00 $ 3.00
4 $ 3.00 $ 4.00
5 $ 5.00 $ 5.00
6 $ 6.00 $ 6.00
7 $ 7.00 $ 7.00
8 $ 8.00 $ 8.00
9 $ 9.00 $ 9.00
10 $ 11.00 $ 10.00
11 $ 66.00
New York City
Los Angeles
Mean $ 11.00 $ 5.50
Median $ 6.00 $ 5.50
Mode $ 3.00 -
Which measure is best?
There is no best, but using only one is definitely the worst!

MEASURES OF CENTRAL TENDENCY

MEASURES OF ASYMMETRY

SAMPLE SKEWNESS FORMULA

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}^3}$$



Almost always, you will use software that performs the calculation for you

Skewness indicates whether the data is concentrated on one side

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste — Clipboard

Arial 9 — Font

Wrap Text Merge & Center — Alignment

General — Number

Conditional Formatting Format as Table Cell Styles — Styles

Insert Delete Format — Cells

AutoSum Fill Clear Sort & Filter Find & Select — Editing

A1

Skewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
2
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

mean > median

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

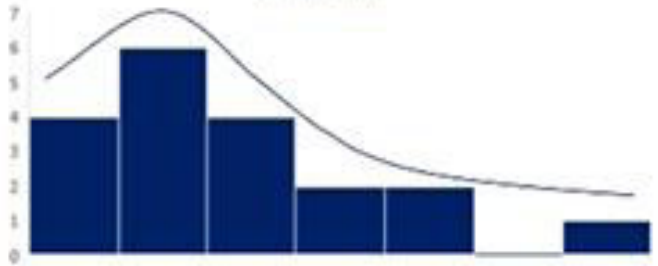
Mean	Median	Mode
4.00	4.00	4.00

Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
7
7
7

Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00

Positive skew



Parents | Frequency | Mean-median-mode | Graphs Part1 | Graphs Part2 | Graphs Part3

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do

Cut Copy Format Painter Paste — Clipboard

Arial 9 — Font

Wrap Text Merge & Center — Alignment

General — Number

Conditional Formatting Format as Table Cell Styles — Styles

Insert Delete Format — Cells

AutoSum Fill Clear Sort & Filter Find & Select — Editing

A1

Skewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
2
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

mean > median

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00

Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
7
7
7

Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste Clipboard

Arial 9 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting Format as Table Cell Styles Styles

Insert Delete Format Cells

AutoSum Fill Clear Sort & Filter Find & Select Editing

A1

Skewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

Zero (no skew)

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00

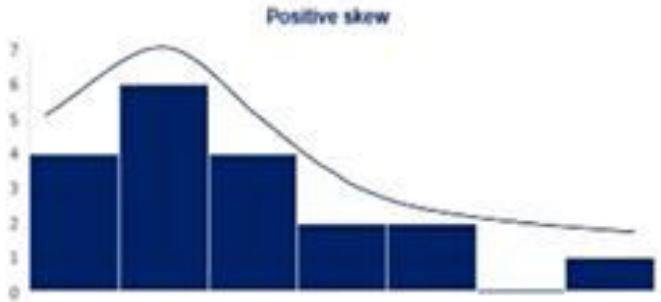
mean = median = mode

Negative (left)

Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
7
7
7

Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste Clipboard

Arial 9 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting Format as Table Cell Styles Styles

Insert Delete Format Cells

AutoSum Fill Clear Sort & Filter Find & Select Editing

A1

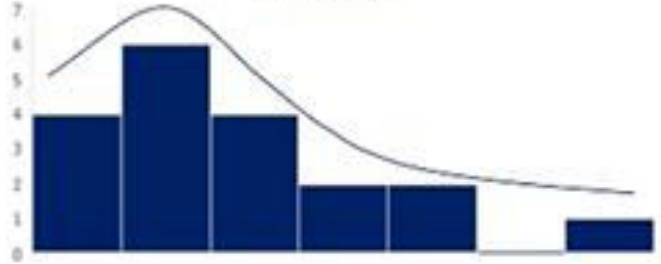
Skewness

Positive (right)

Dataset 1
1
1
1
1
2
2
2
2
2
3
3
3
4
4
5
5
7

Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00



Zero (no skew)

Dataset 2
1
1
2
2
3
3
3
4
4
4
4
4
5
5
5
6
6
7
7

Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00

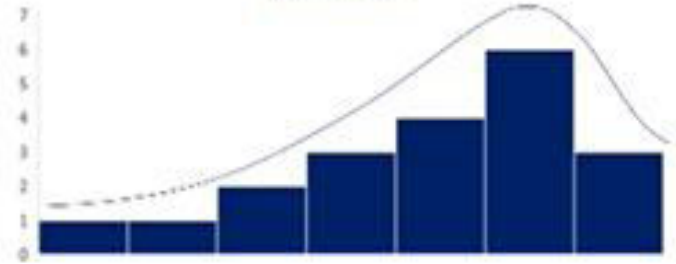
Negative (left)

Dataset 3
1
2
3
3
4
4
4
5
5
5
6
6
6
6
6
7
7
7

Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

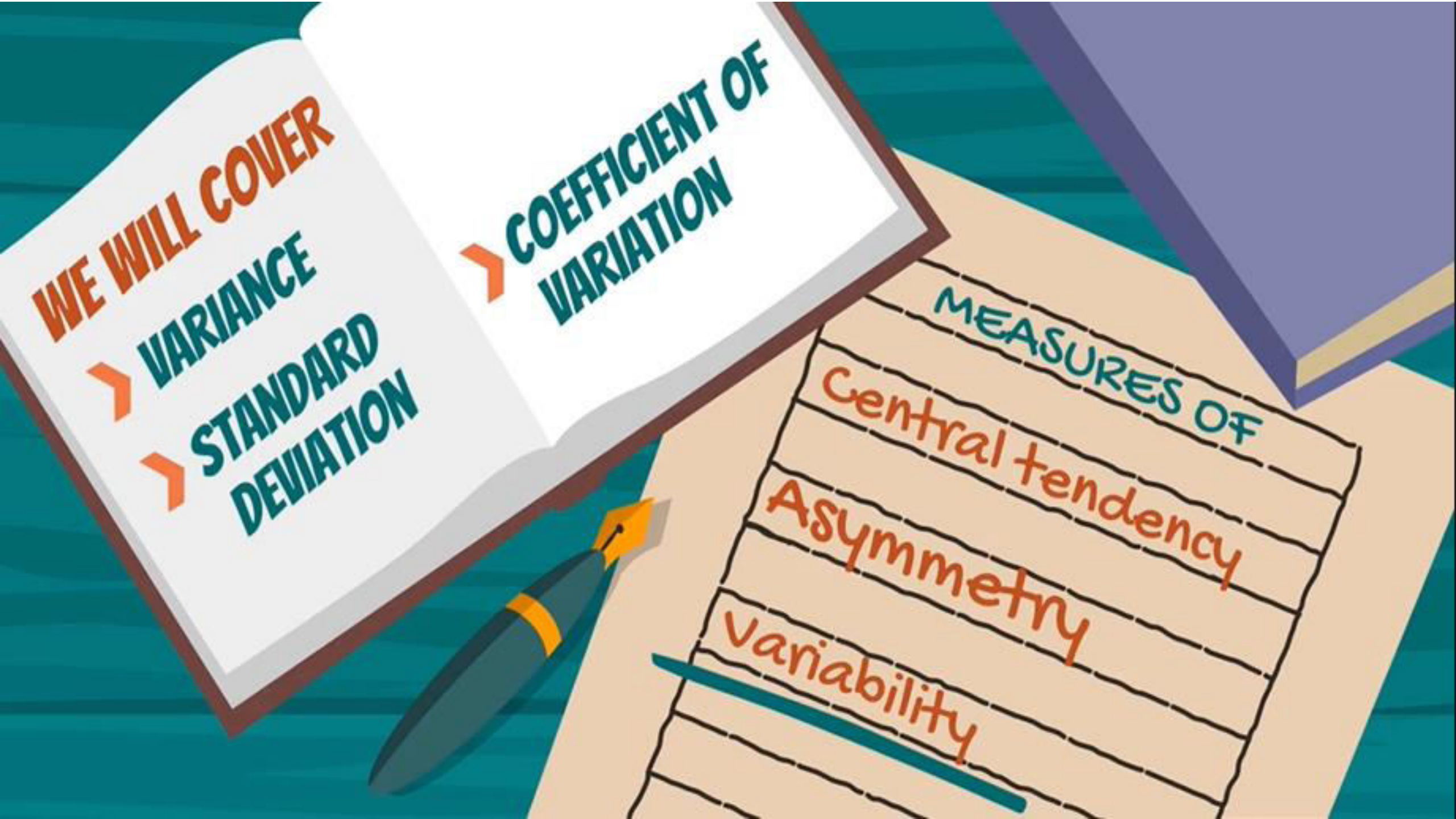
Mean	Median	Mode
4.90	5.00	6.00

mean < median



Measuring how data is spread out : calculating variance


MEASURES OF
Central tendency
Asymmetry
Variability


WE WILL COVER
VARIANCE
STANDARD DEVIATION
COEFFICIENT OF VARIATION
MEASURES OF
Central tendency
Asymmetry
Variability

DIFFERENT FORMULAS FOR

SAMPLE DATA

POPULATION DATA

WHEN YOU TAKE SAMPLE DATA

SAMPLE DATA

A sample statistic is an approximation of the population parameter

WHEN YOU TAKE SAMPLE DATA

SAMPLE DATA

10 different samples give

10 different measures

WHEN YOU TAKE SAMPLE DATA

FORMULA

population formula

sample formula

Mean

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

sample formula

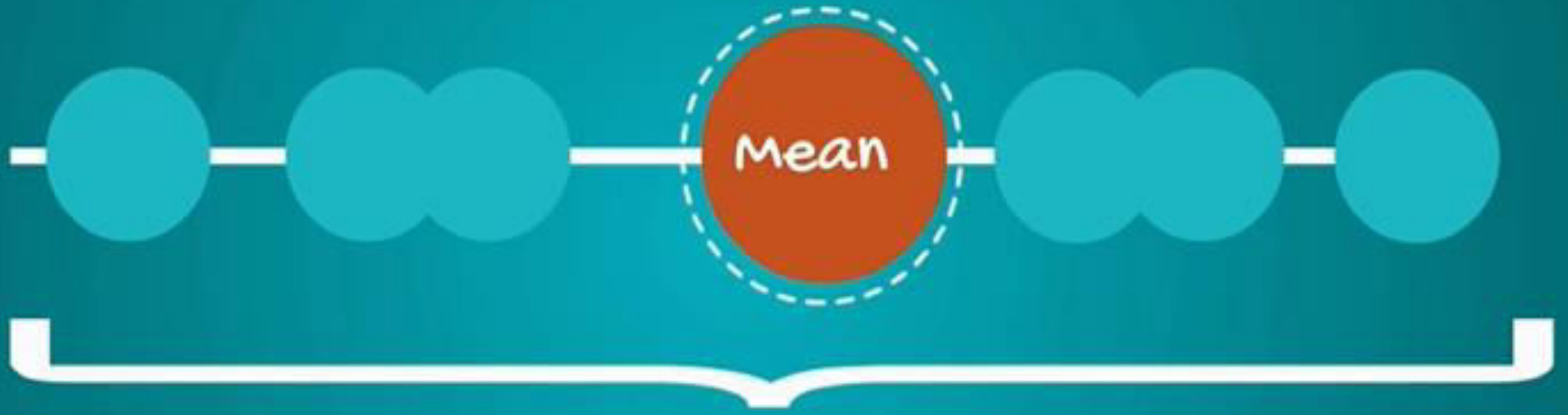
n is the size of the sample

$$\frac{\sum_{i=1}^{N} x_i}{N}$$

population formula

N is the size of the population

VARIANCE



Variance measures the dispersion of a set of data points around their mean

VARIANCE

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

population variance

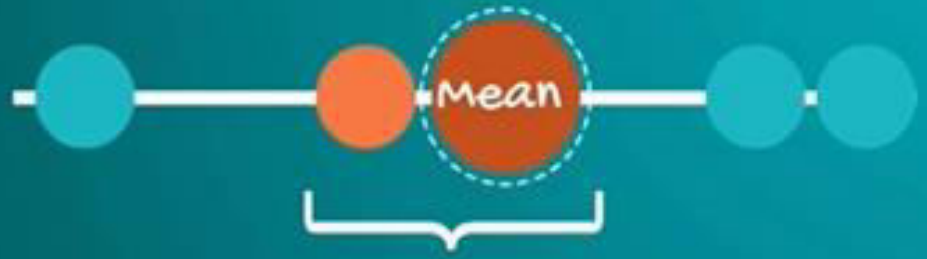
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

sample variance

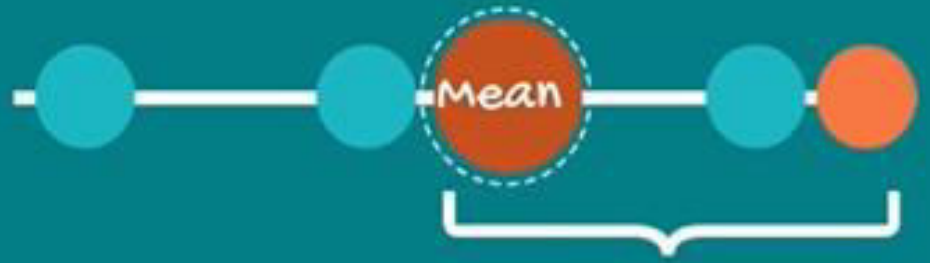
VARIANCE

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

population variance

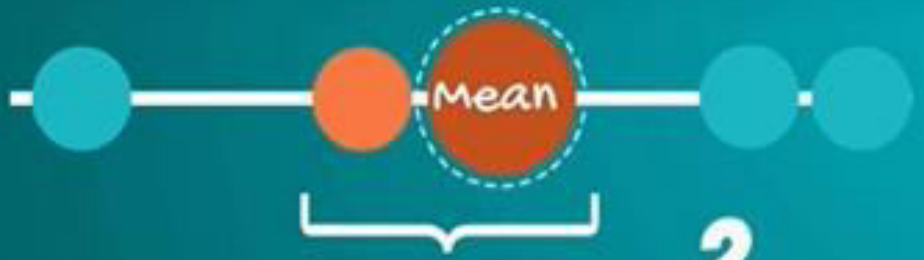


lower result



higher result

- Dispersion is non-negative.
- Non-negative values don't cancel out
- Amplifies the effect of large differences



Variance

Population
1
2
3
4
5

Mean 3.00

We start by calculating the mean. $\text{Mean} = \frac{1+2+3+4+5}{5}$

A1

Variance

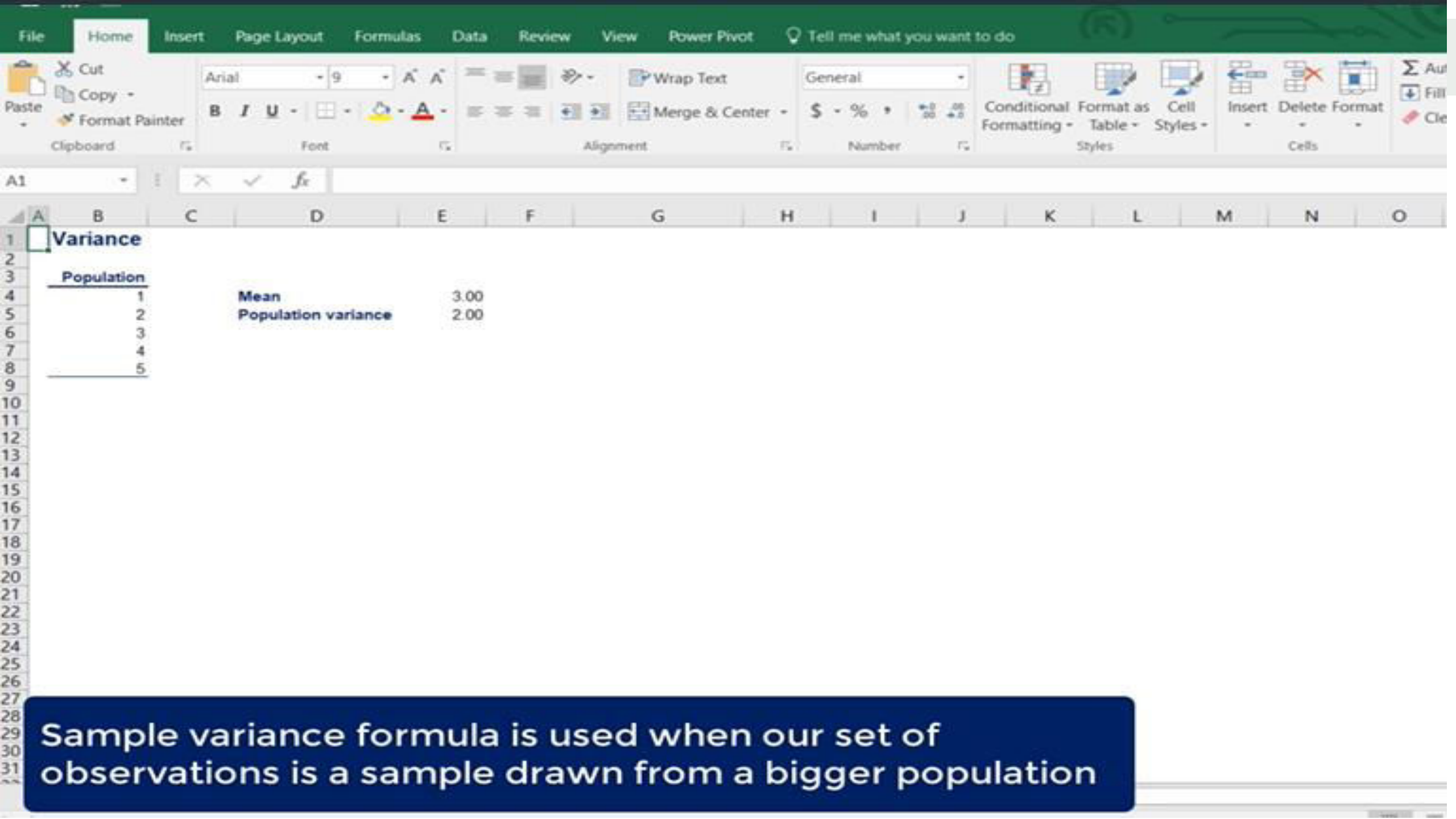
Population
1
2
3
4
5

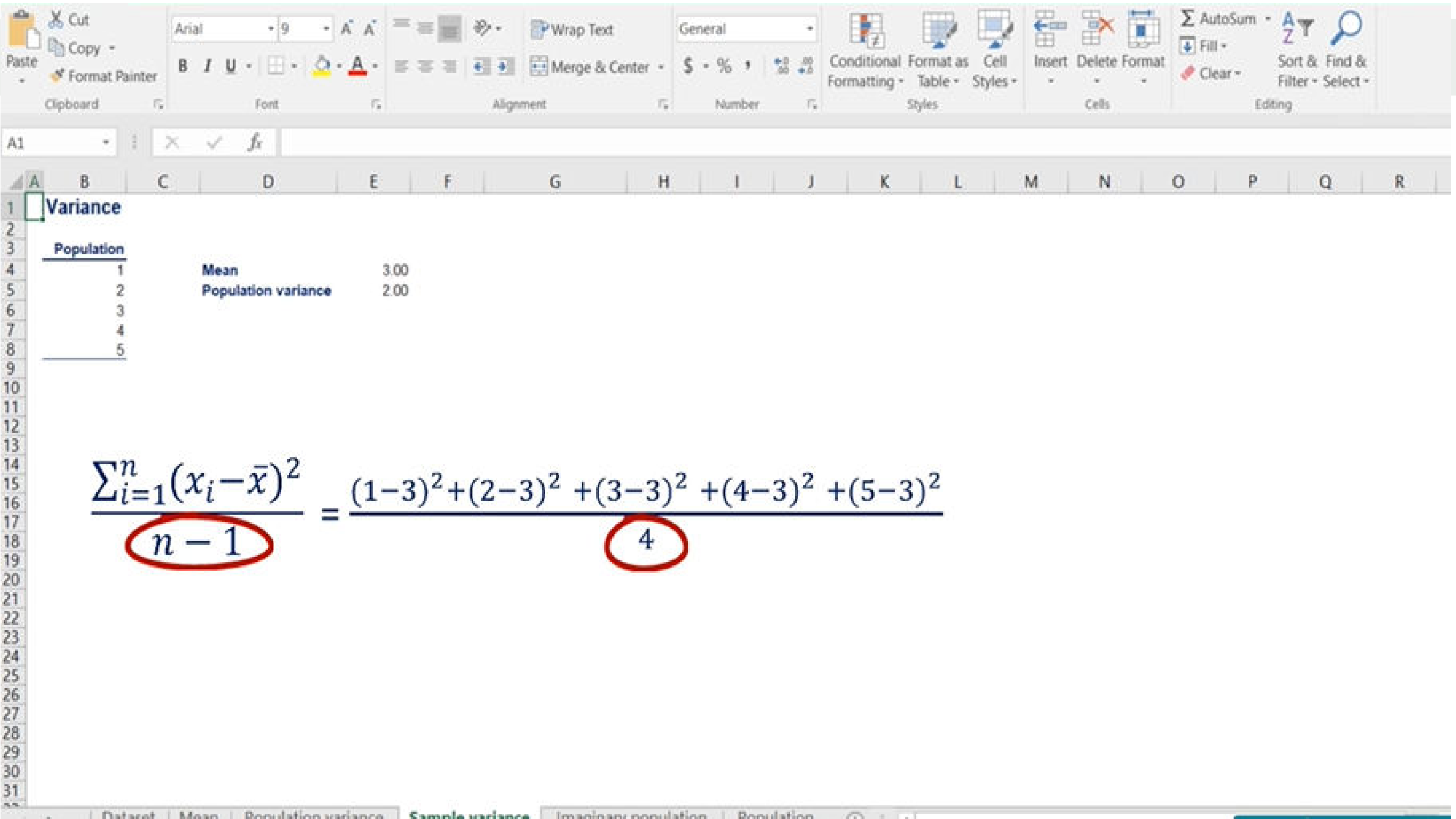
Mean	3.00
Population variance	2.00

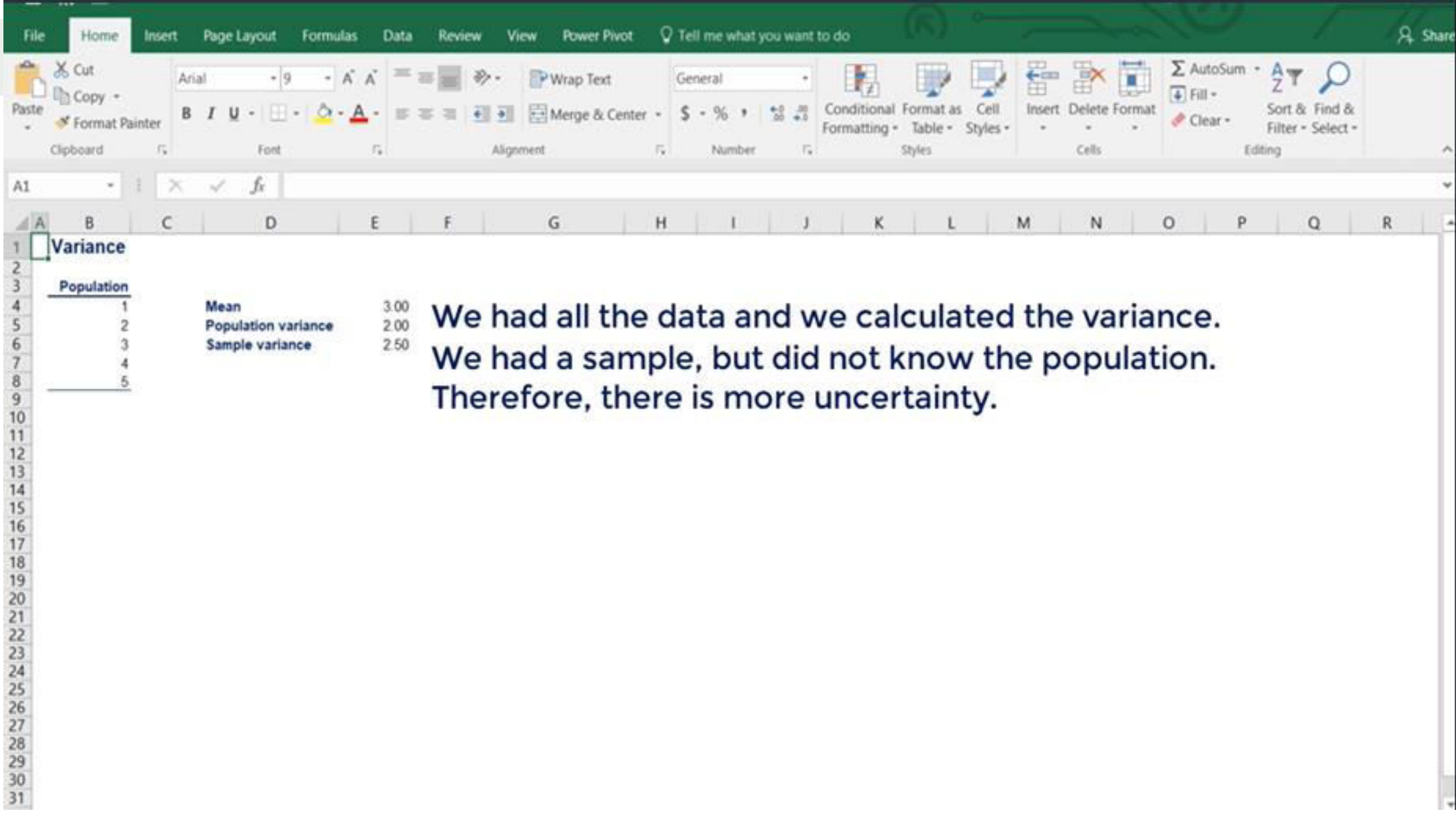
$$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}$$

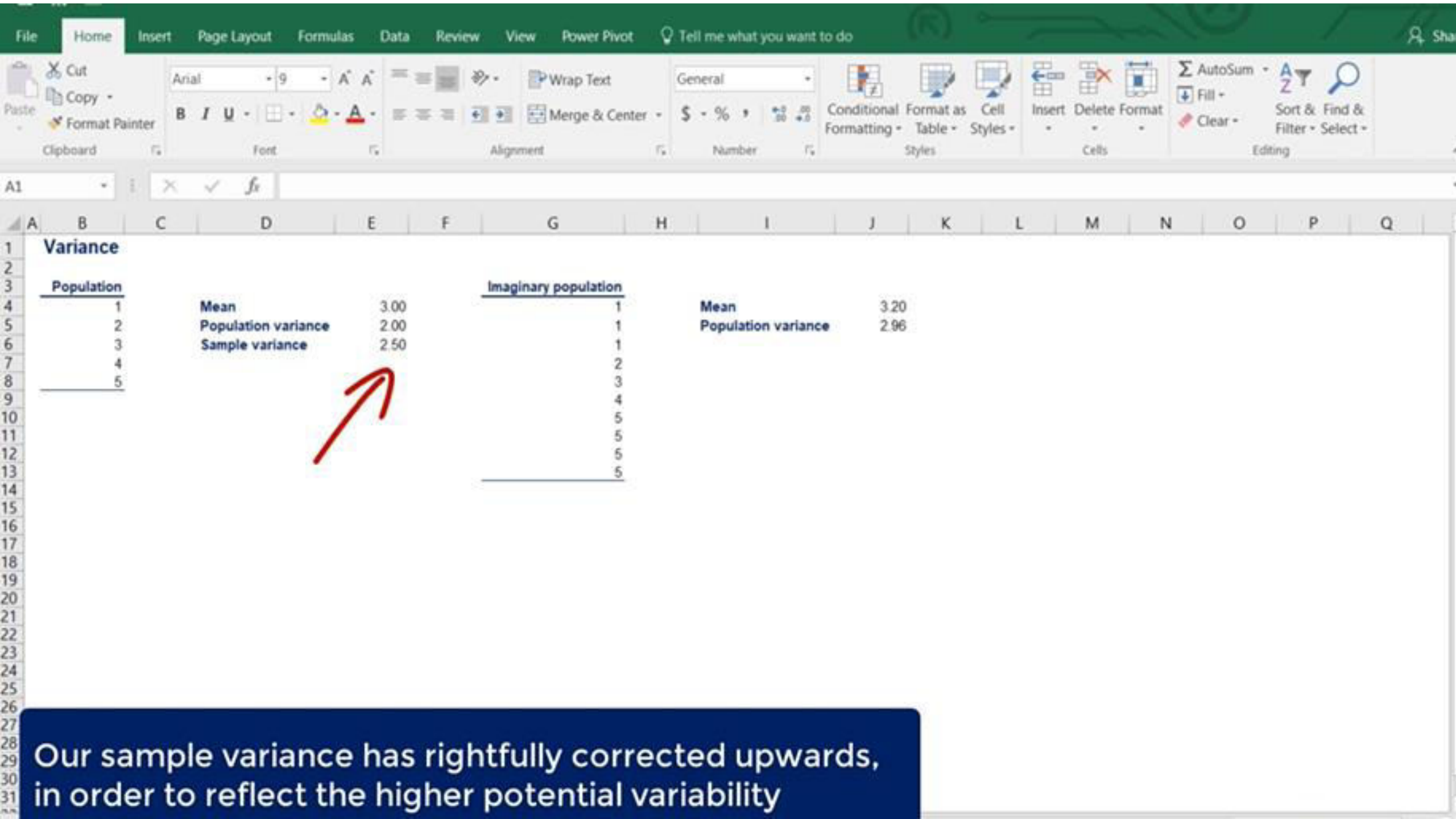
Population variance formula

Population size


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Paste
Cut
Copy
Format Painter
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
A1
Variance
Population
1
2
3
4
5
Mean
3.00
Population variance
2.00
Sample variance formula is used when our set of observations is a sample drawn from a bigger population


Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Variance
Population
1
2
3
4
5
Mean
3.00
Population variance
2.00
$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{4}$
Dataset
Mean
Population variance
Sample variance
Imaginary population
Population


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Variance
Population
1
2
3
4
5
Mean 3.00
Population variance 2.00
Sample variance 2.50
We had all the data and we calculated the variance.
We had a sample, but did not know the population.
Therefore, there is more uncertainty.


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Variance
Population
1
2
3
4
5
Mean 3.00
Population variance 2.00
Sample variance 2.50
Imaginary population
1
1
1
2
3
4
5
5
5
5
Mean 3.20
Population variance 2.96
Our sample variance has rightfully corrected upwards,
in order to reflect the higher potential variability


WE WILL COVER
VARIANCE
STANDARD DEVIATION
COEFFICIENT OF VARIATION
VARIANCE
large
units
2


WE WILL COVER
VARIANCE
STANDARD DEVIATION
COEFFICIENT OF VARIATION
VARIANCE
large 2 units

WHEN YOU TAKE SAMPLE DATA

FORMULA

population formula

sample formula

STANDARD DEVIATION FORMULAS

sample standard deviation

population standard deviation

STANDARD DEVIATION FORMULAS

$$\sigma = \sqrt{\sigma^2}$$

population standard deviation

sample standard deviation

$$S = \sqrt{S^2}$$

COEFFICIENT OF VARIATION (CV)

/relative standard deviation/

$$\frac{\text{standard deviation}}{\text{mean}}$$

COEFFICIENT OF VARIATION (CV)

$$c_v = \frac{\sigma}{\mu}$$

Population formula

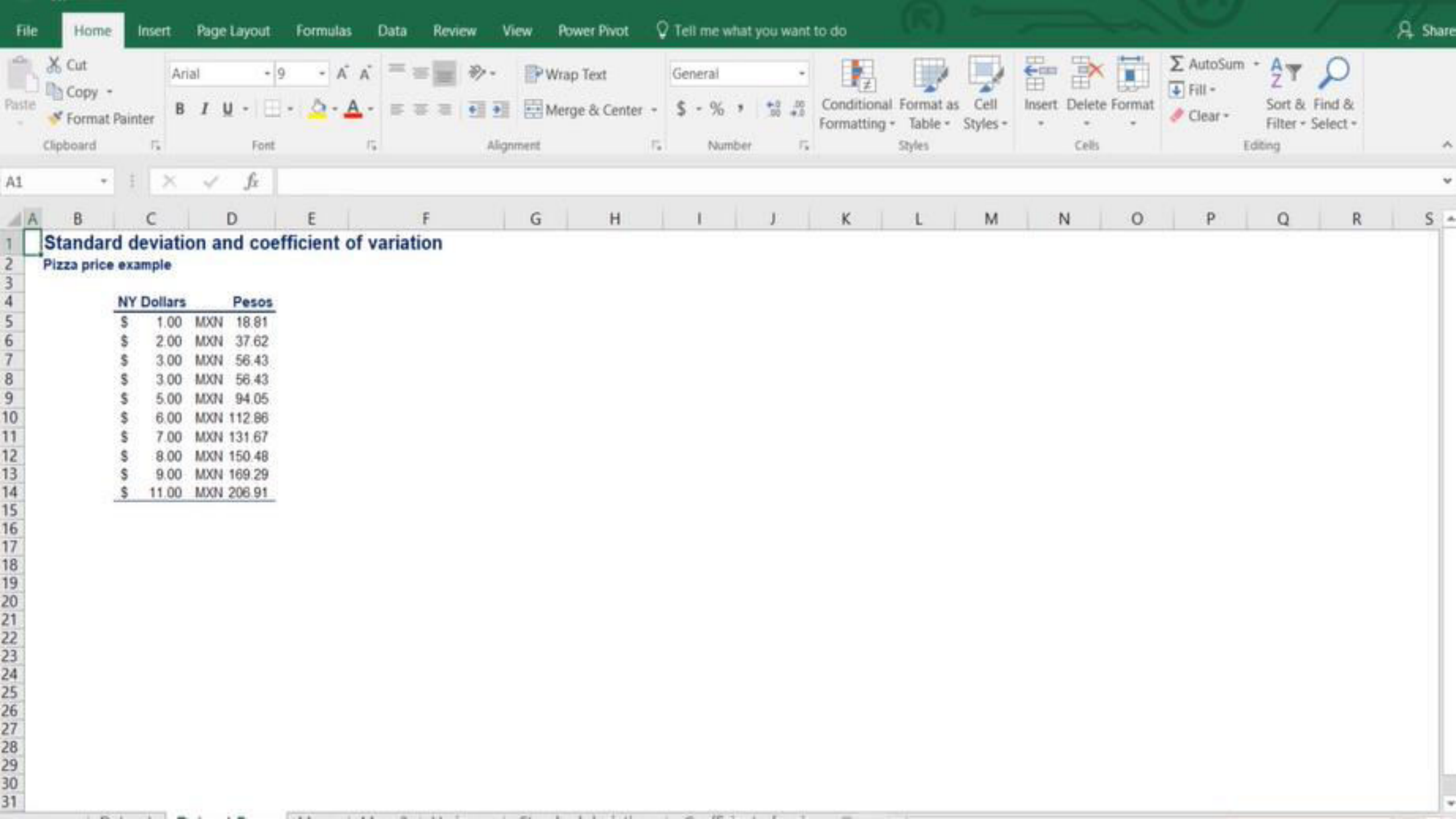
Sample formula

$$\hat{c}_v = \frac{s}{\bar{x}}$$

Standard deviation is the most common measure of variability for a SINGLE DATASET

Standard deviation is the most common measure of variability for a SINGLE DATASET

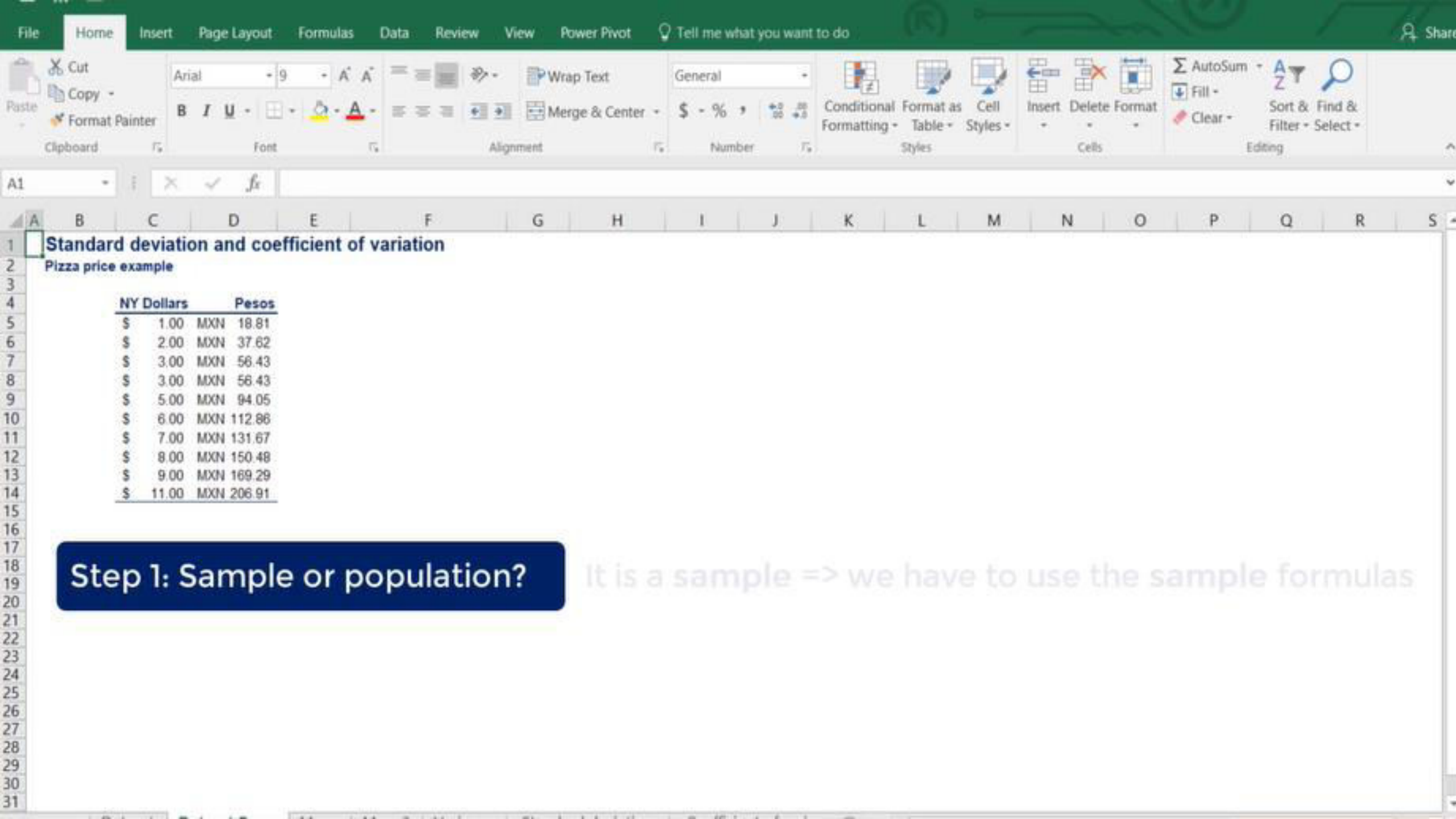
Comparing TWO OR MORE datasets

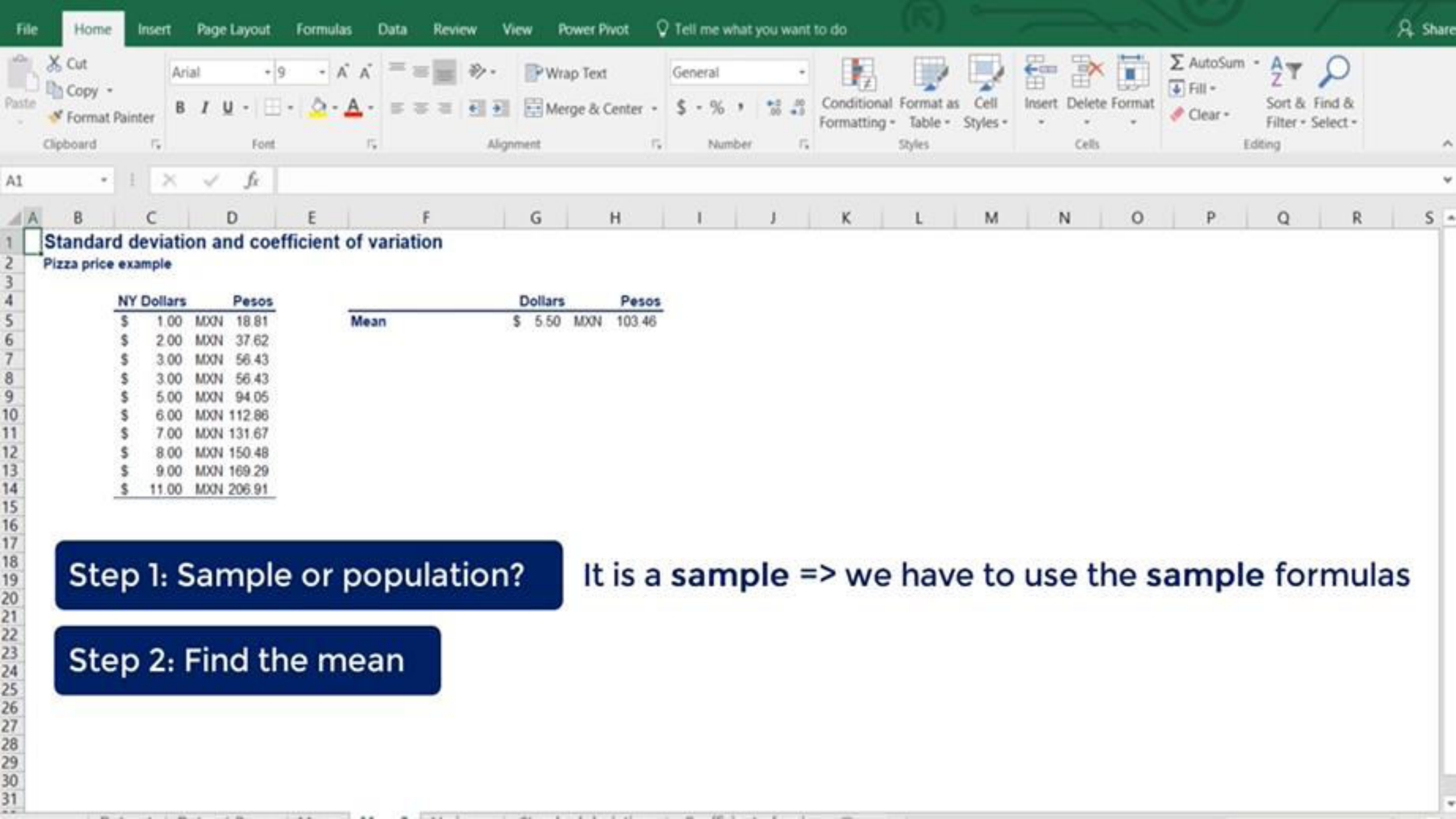


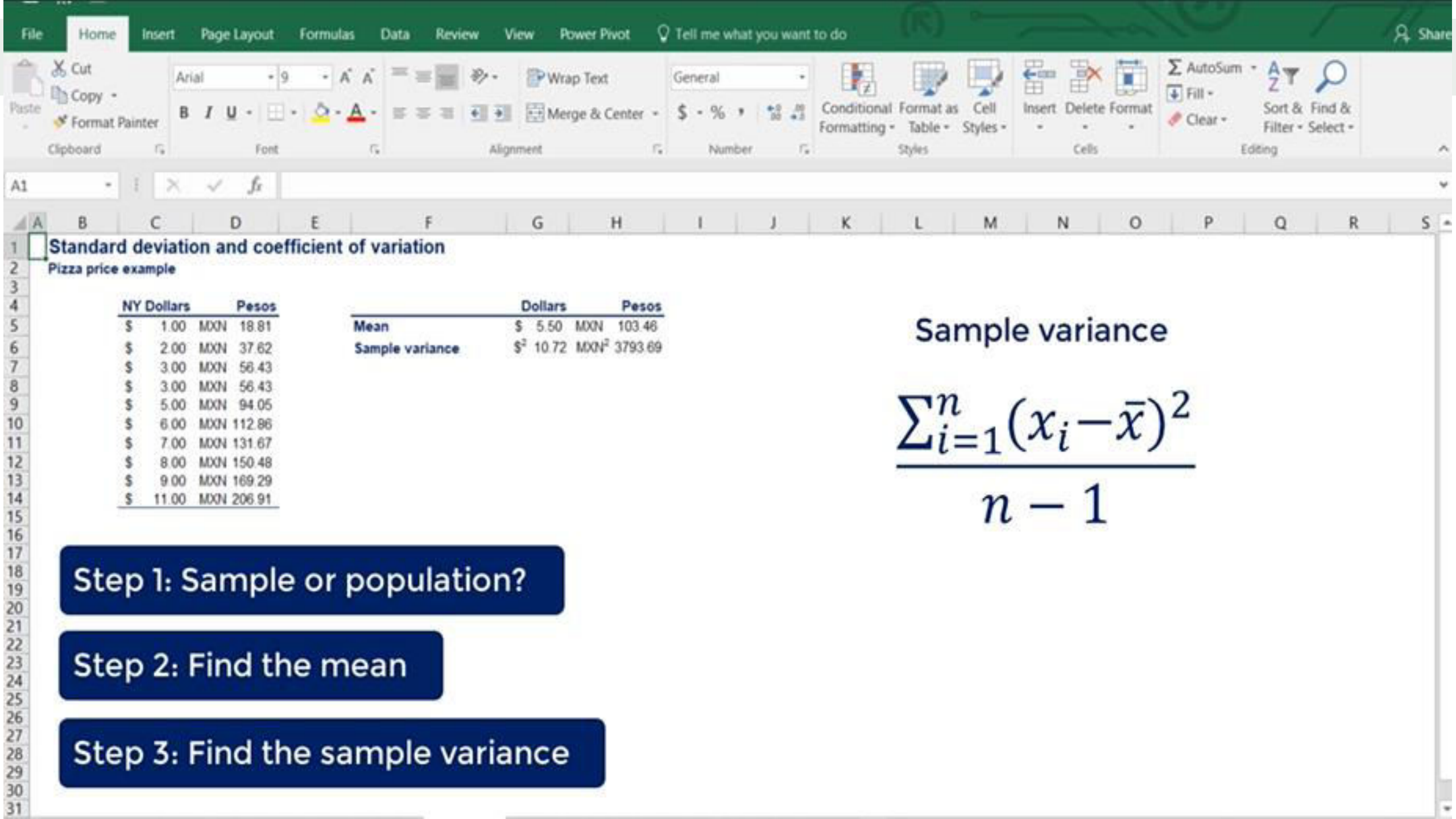
Standard deviation and coefficient of variation

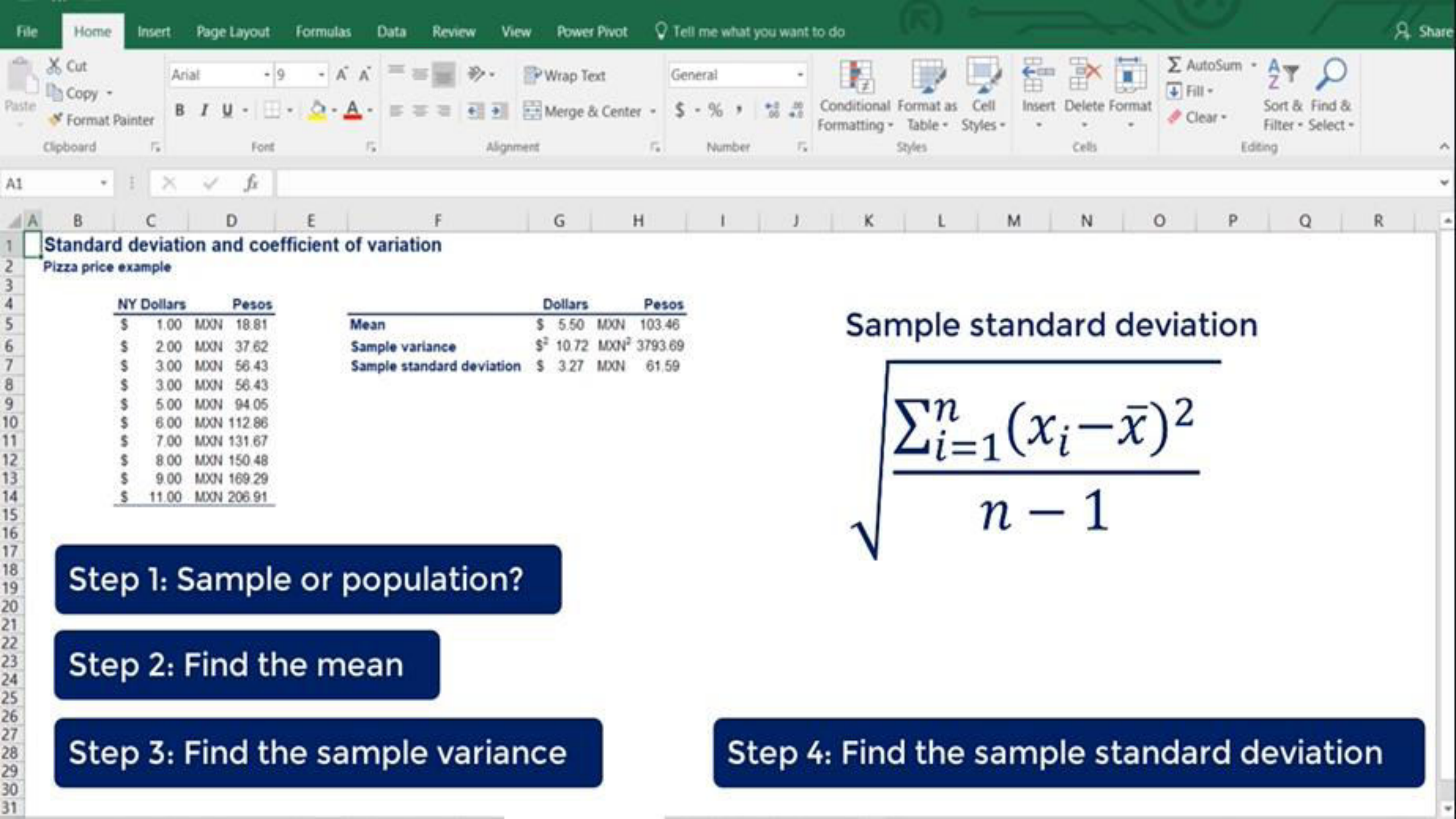
Pizza price example

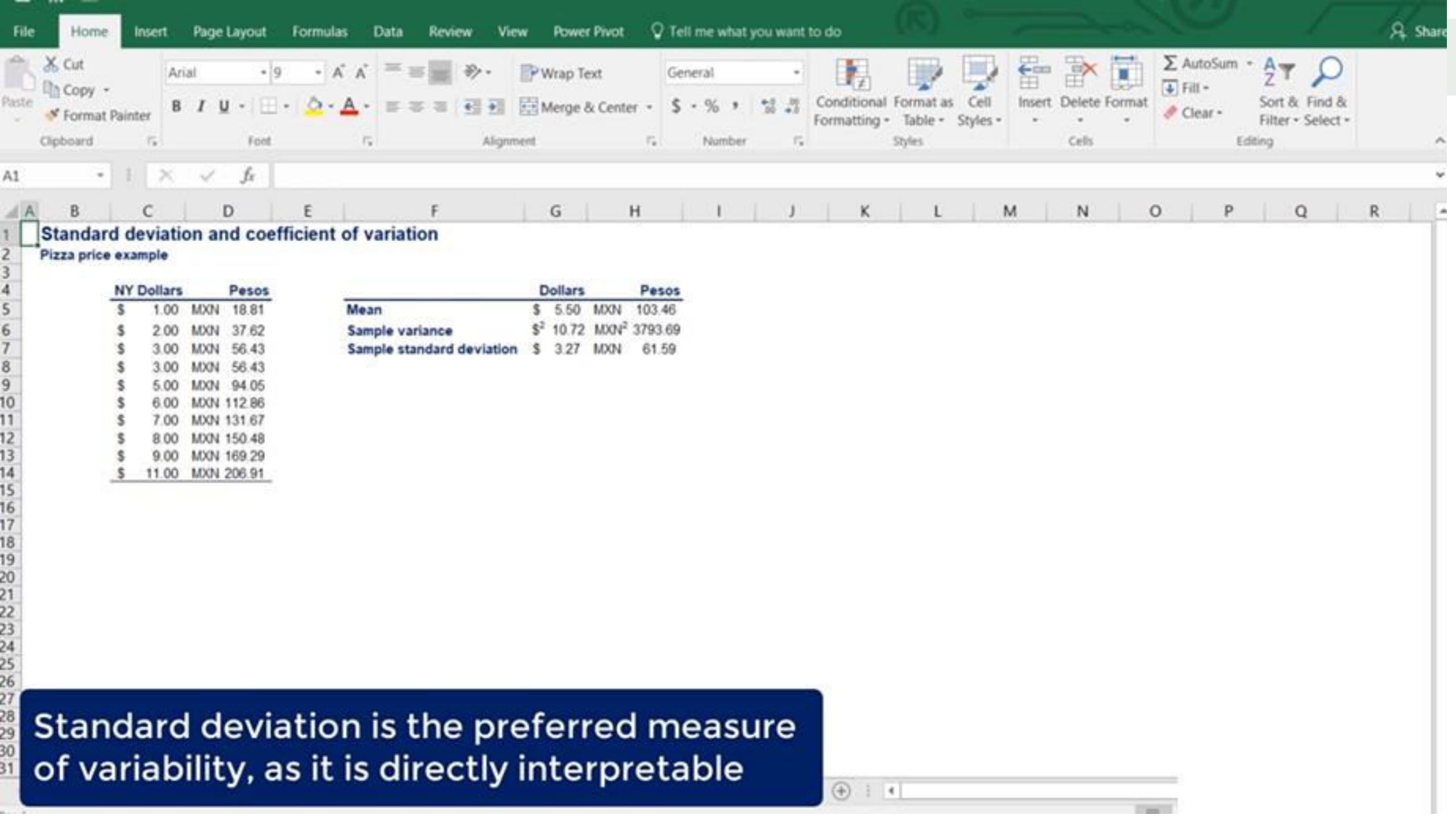
NY Dollars			Pesos
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

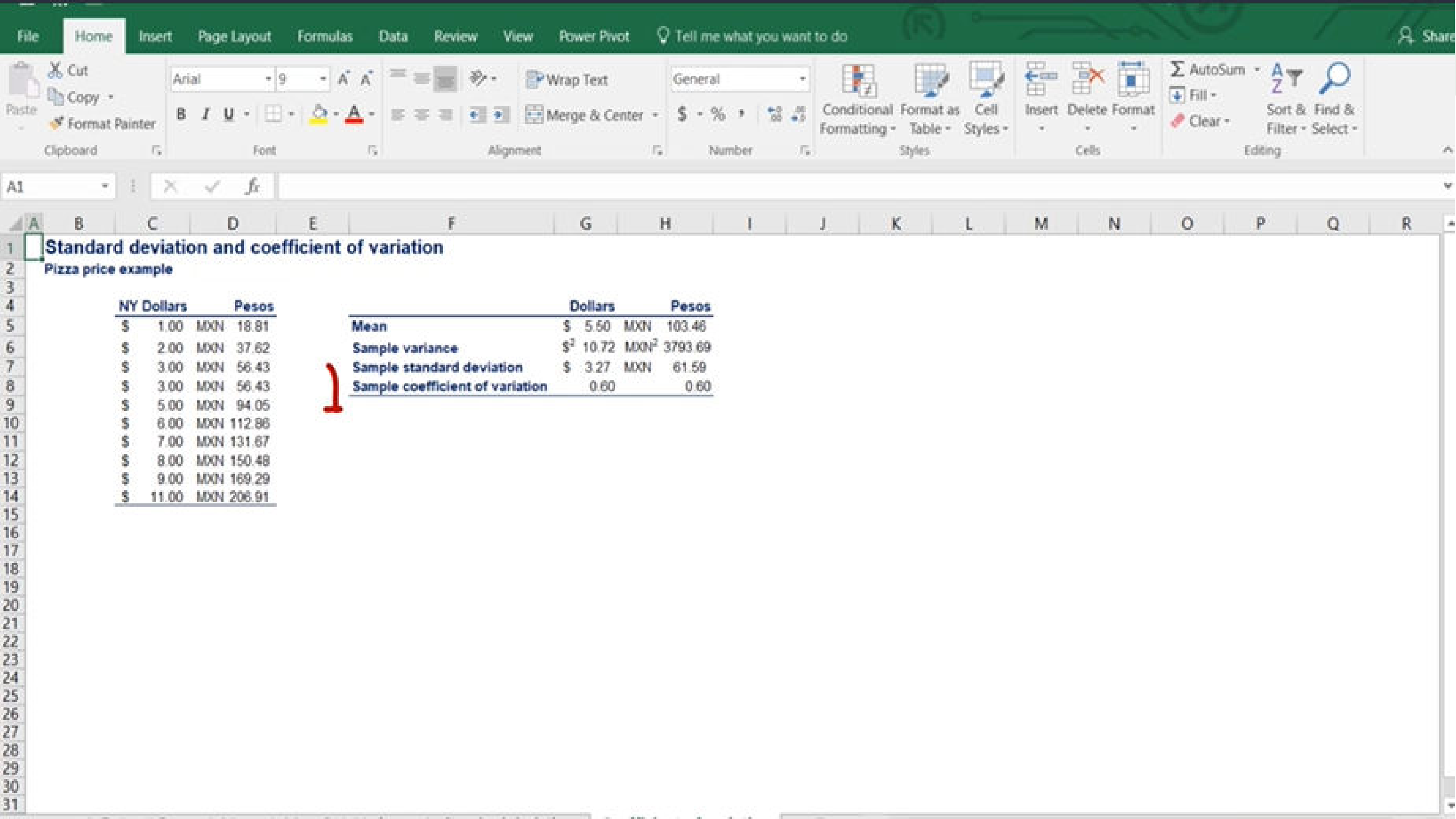

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Standard deviation and coefficient of variation
Pizza price example
NY Dollars
Pesos
$ 1.00 MXN 18.81
$ 2.00 MXN 37.62
$ 3.00 MXN 56.43
$ 3.00 MXN 56.43
$ 5.00 MXN 94.05
$ 6.00 MXN 112.86
$ 7.00 MXN 131.67
$ 8.00 MXN 150.48
$ 9.00 MXN 169.29
$ 11.00 MXN 206.91
Step 1: Sample or population?
It is a sample => we have to use the sample formulas


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Standard deviation and coefficient of variation
Pizza price example
NY Dollars
Pesos
$ 1.00 MXN 18.81
$ 2.00 MXN 37.62
$ 3.00 MXN 56.43
$ 3.00 MXN 56.43
$ 5.00 MXN 94.05
$ 6.00 MXN 112.86
$ 7.00 MXN 131.67
$ 8.00 MXN 150.48
$ 9.00 MXN 169.29
$ 11.00 MXN 206.91
Dollars
Pesos
Mean
$ 5.50 MXN 103.46
Step 1: Sample or population?
It is a sample => we have to use the sample formulas
Step 2: Find the mean


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Standard deviation and coefficient of variation
Pizza price example
NY Dollars
Pesos
$ 1.00 MXN 18.81
$ 2.00 MXN 37.62
$ 3.00 MXN 56.43
$ 3.00 MXN 56.43
$ 5.00 MXN 94.05
$ 6.00 MXN 112.86
$ 7.00 MXN 131.67
$ 8.00 MXN 150.48
$ 9.00 MXN 169.29
$ 11.00 MXN 206.91
Dollars
Pesos
Mean
$ 5.50 MXN 103.46
Sample variance
$² 10.72 MXN² 3793.69
Sample variance
$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
Step 1: Sample or population?
Step 2: Find the mean
Step 3: Find the sample variance


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Standard deviation and coefficient of variation
Pizza price example
NY Dollars
Pesos
$ 1.00 MXN 18.81
$ 2.00 MXN 37.62
$ 3.00 MXN 56.43
$ 3.00 MXN 56.43
$ 5.00 MXN 94.05
$ 6.00 MXN 112.86
$ 7.00 MXN 131.67
$ 8.00 MXN 150.48
$ 9.00 MXN 169.29
$ 11.00 MXN 206.91
Dollars
Pesos
Mean $ 5.50 MXN 103.46
Sample variance $² 10.72 MXN² 3793.69
Sample standard deviation $ 3.27 MXN 61.59
Sample standard deviation
$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$
Step 1: Sample or population?
Step 2: Find the mean
Step 3: Find the sample variance
Step 4: Find the sample standard deviation


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Arial
9
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Standard deviation and coefficient of variation
Pizza price example
NY Dollars Pesos
$ 1.00 MXN 18.81
$ 2.00 MXN 37.62
$ 3.00 MXN 56.43
$ 3.00 MXN 56.43
$ 5.00 MXN 94.05
$ 6.00 MXN 112.86
$ 7.00 MXN 131.67
$ 8.00 MXN 150.48
$ 9.00 MXN 169.29
$ 11.00 MXN 206.91
Dollars Pesos
Mean $ 5.50 MXN 103.46
Sample variance $² 10.72 MXN² 3793.69
Sample standard deviation $ 3.27 MXN 61.59
Standard deviation is the preferred measure of variability, as it is directly interpretable

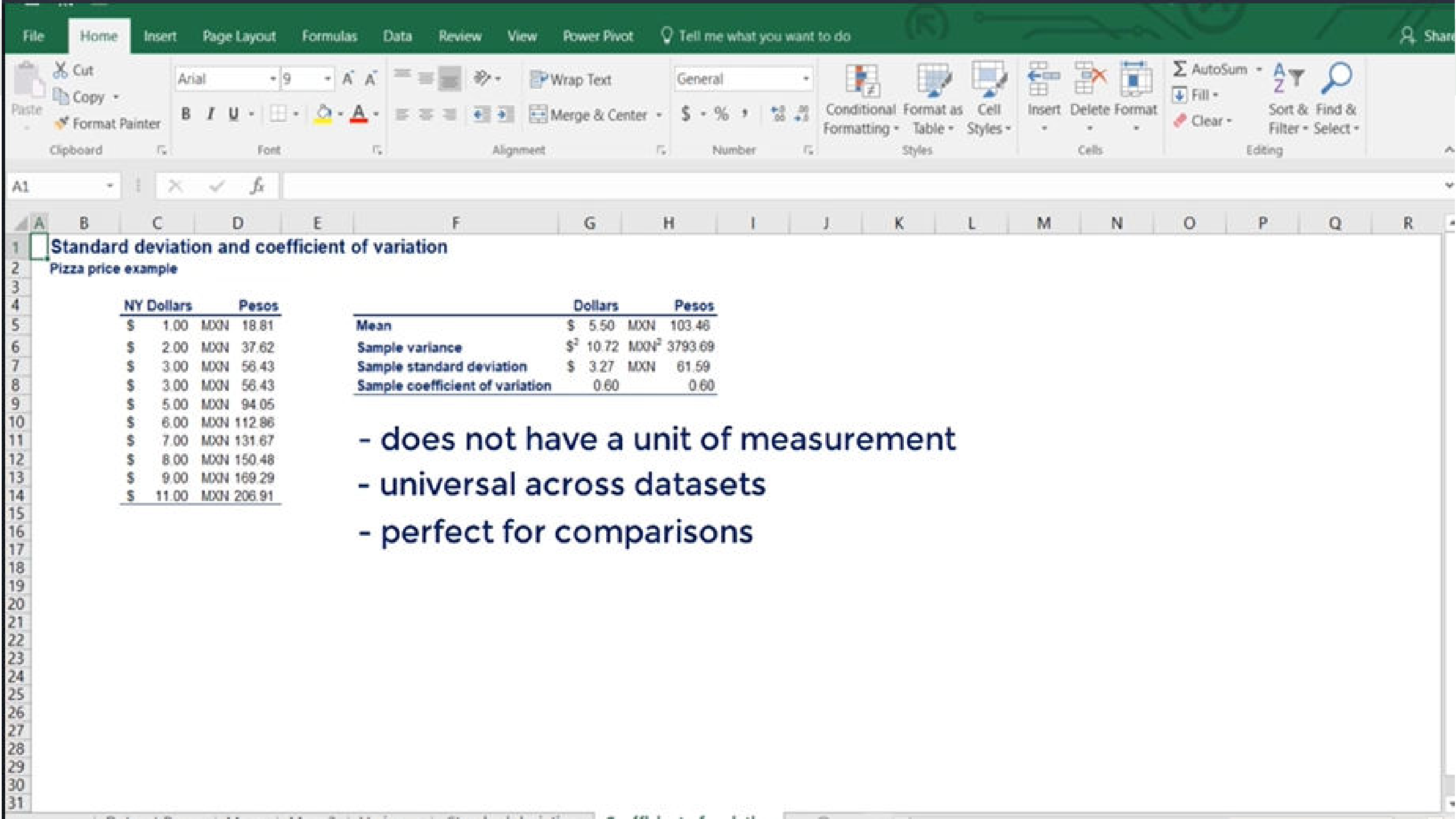


Standard deviation and coefficient of variation

Pizza price example

NY Dollars			Pesos
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

	Dollars		Pesos
Mean	$ 5.50	MXN	103.46
Sample variance	$^2 10.72	MXN^2	3793.69
Sample standard deviation	$ 3.27	MXN	61.59
Sample coefficient of variation	0.60		0.60



Standard deviation and coefficient of variation

Pizza price example

NY Dollars			Pesos
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

	Dollars		Pesos
Mean	$ 5.50	MXN	103.46
Sample variance	$^2 10.72	MXN^2	3793.69
Sample standard deviation	$ 3.27	MXN	61.59
Sample coefficient of variation	0.60		0.60

- does not have a unit of measurement
- universal across datasets
- perfect for comparisons

RECAP

σ^2 Variance

σ Standard deviation

c_v Coefficient of variation

Question 1:

Johnny wanted to know how many days per week people in his class exercise, so he asked 3 of his friends. The answers he got were 1, 3 and 5. Based on this, which of the following is true about the results Johnny found?

Hint: Do the numbers [1,3,5] represent a population or a sample?

- ◯ $\mu = 3, \sigma = 2$
- ◯ $\bar{x} = 3, \sigma = -2$
- ◯ $\bar{x} = 3, s = 2$
- ◯ $\bar{x} = 3, s = 4$


Univariate measures
Central tendency
Asymmetry
Variability


in the next two lessons...
MEASURES OF RELATIONSHIP BETWEEN VARIABLES
hivariate measures
Central tendency
Asymmetry
Variability


in the next two lessons...
MEASURES OF RELATIONSHIP BETWEEN VARIABLES
OUR FOCUS:
Covariance
Linear correlation coefficient

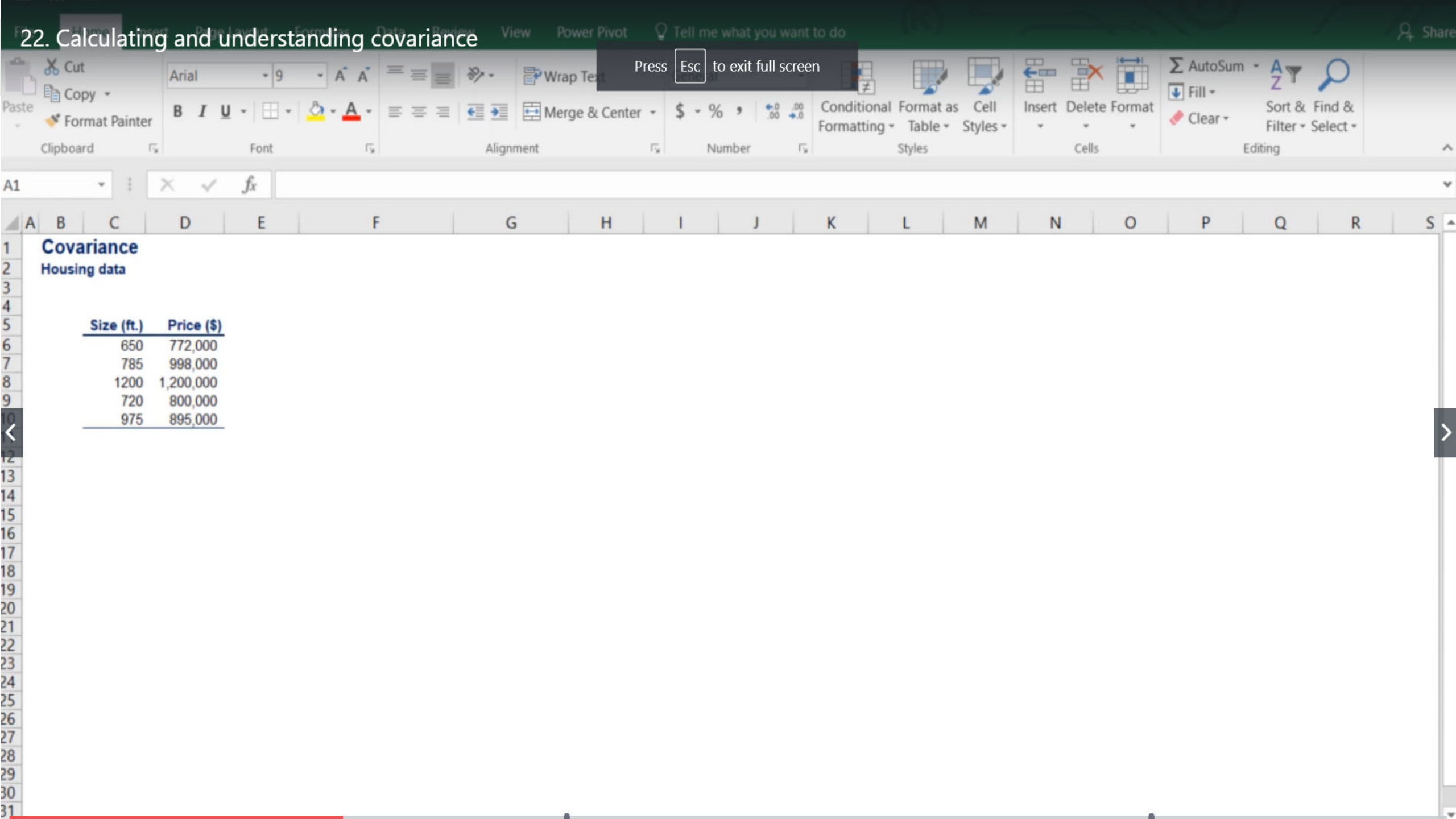
REAL ESTATE

what determines house prices?

REAL ESTATE

Their size!

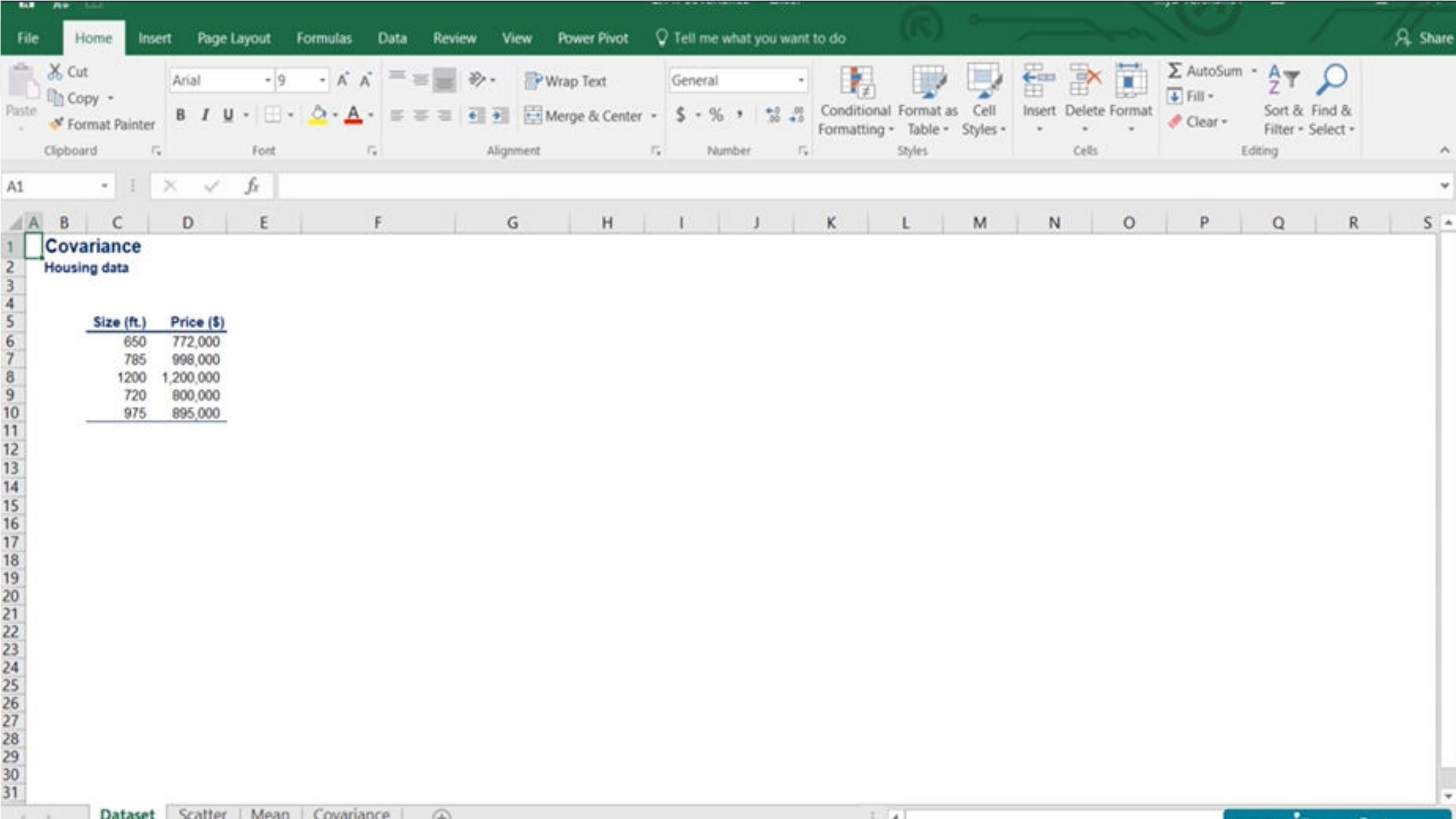
22. Calculating and understanding covariance



Covariance

Housing data

Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
Cut
Copy
Format Painter
Paste
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
A1
Covariance
Housing data
Size (ft.)
Price ($)
650
772,000
785
998,000
1200
1,200,000
720
800,000
975
895,000
Dataset
Scatter
Mean
Covariance

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Cut Copy Format Painter Paste — Clipboard

Arial 9 — Font

Wrap Text Merge & Center — Alignment

General — Number

Conditional Formatting Format as Table Cell Styles — Styles

Insert Delete Format — Cells

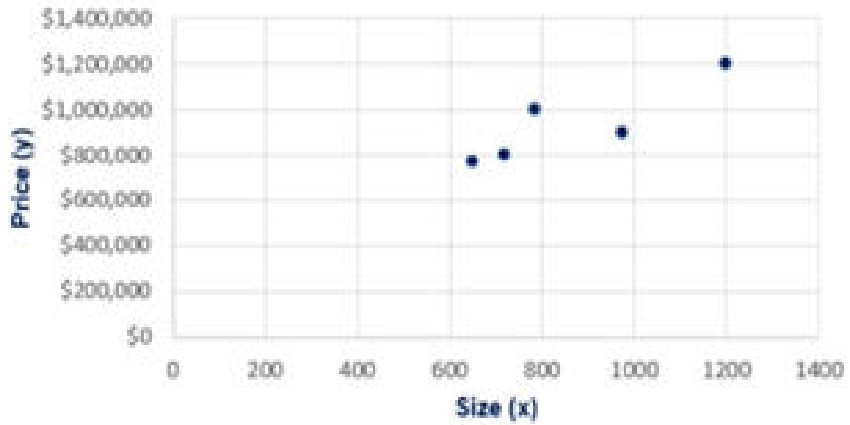
AutoSum Fill Clear Sort & Filter Find & Select — Editing

A1

Covariance

Housing data

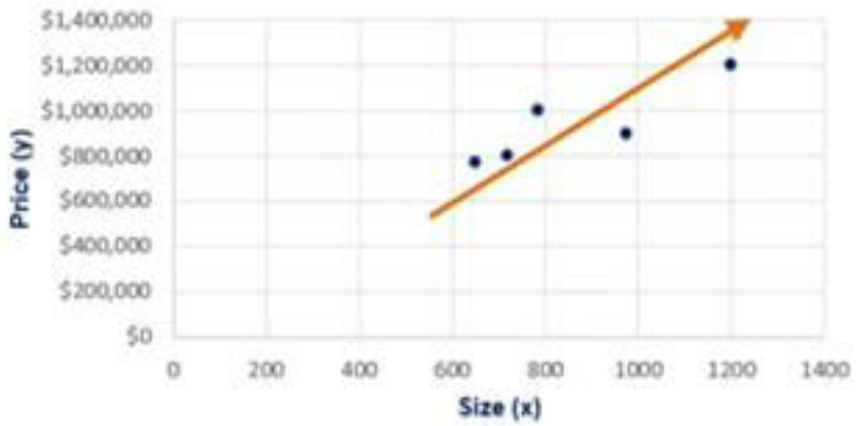
Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000



Covariance

Housing data

Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000

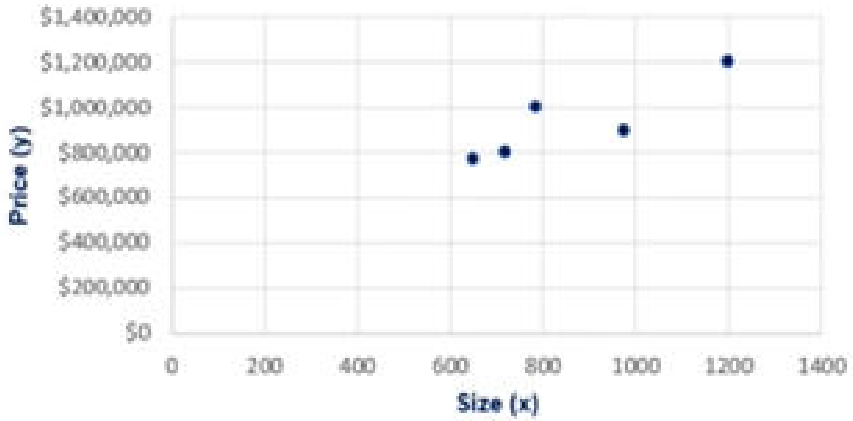


The two variables are correlated and the main statistic to measure this correlation is called covariance

Covariance

Housing data

Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000



Covariance may be:

> 0
= 0
< 0

Covariance

Housing data

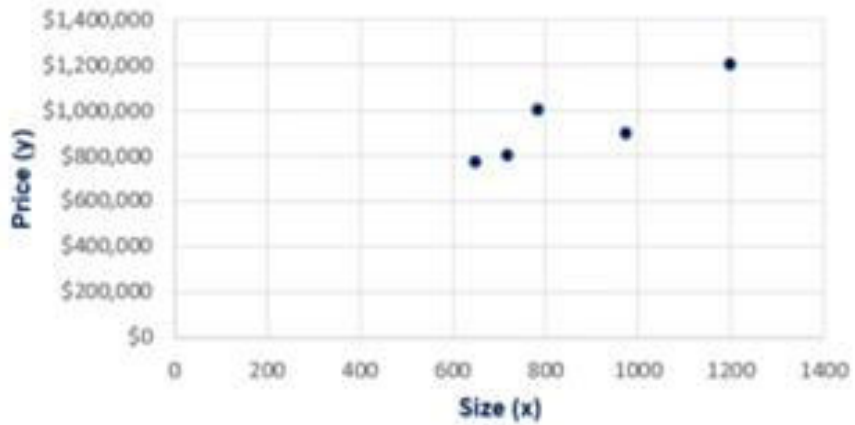
Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000

Sample formula

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population formula

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$



Covariance

Housing data

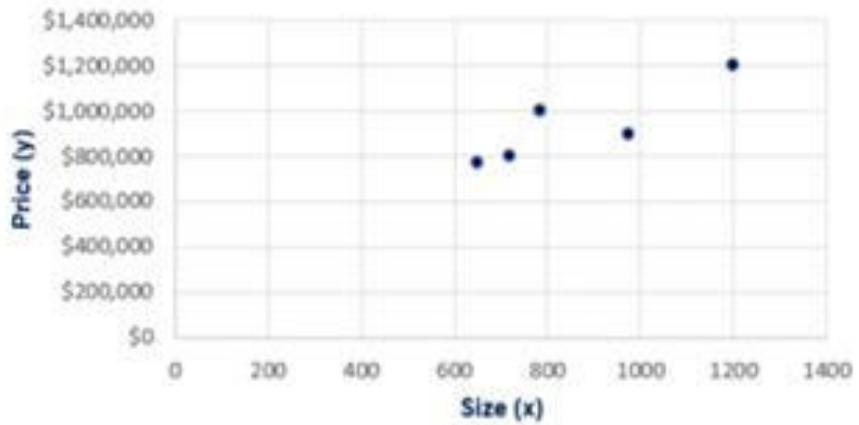
Size (ft.)	Price ($)
650	772,000
785	998,000
1200	1,200,000
720	800,000
975	895,000

Sample formula

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Population formula

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$



G6

Covariance

Housing data

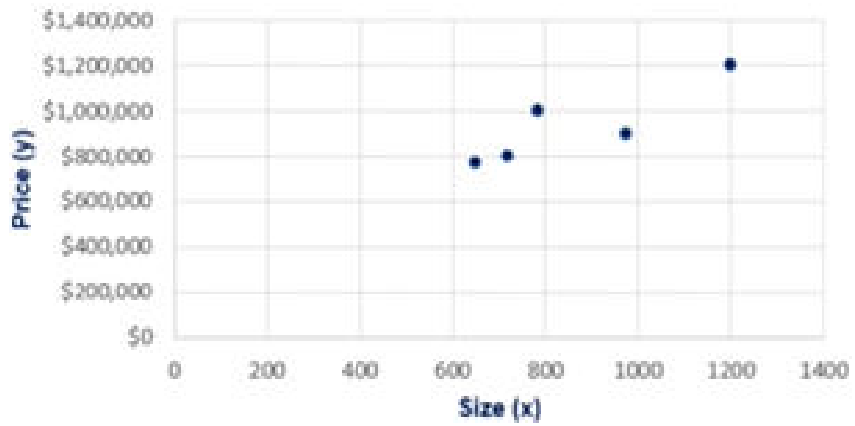
	x	y
	Size (ft.)	Price ($)
	650	772,000
	785	998,000
	1200	1,200,000
	720	800,000
	975	895,000
Mean	866	933,000

$(x-\bar{x})*(y-\bar{y})$

Sum

Sample size

Cov. Sample



Sample formula

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})*(y_i-\bar{y})}{n-1}$$

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Insert Function | AutoSum | Recently Used | Financial | Logical | Text | Date & Time | Lookup & Reference | Math & Trig | More Functions — Function Library

Name Manager | Define Name | Use in Formula | Create from Selection — Defined Names

Trace Precedents | Trace Dependents | Remove Arrows | Show Formulas | Error Checking | Evaluate Formula — Formula Auditing

Watch Window

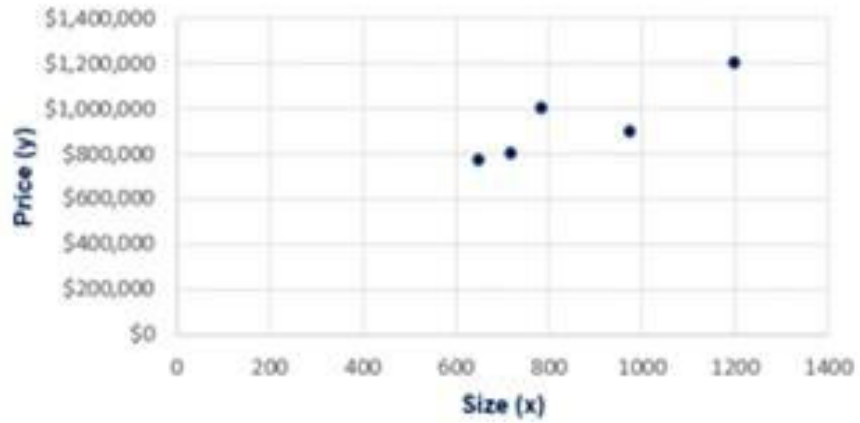
Calculation Options | Calculate Now | Calculate Sheet — Calculation

G14

Covariance

Housing data

	x	y		
	Size (ft.)	Price ($)		$(x-\bar{x})*(y-\bar{y})$
	650	772,000		34,776,000
	785	998,000		- 5,265,000
	1200	1,200,000		89,178,000
	720	800,000		19,418,000
	975	895,000		- 4,142,000
Mean	866	933,000	Sum	133,965,000
			Sample size	5
			Cov. Sample	33,491,250



File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Insert Function | AutoSum | Recently Used | Financial | Logical | Text | Date & Time | Lookup & Reference | Math & Trig | More Functions — Function Library

Name Manager | Define Name | Use in Formula | Create from Selection — Defined Names

Trace Precedents | Trace Dependents | Remove Arrows | Show Formulas | Error Checking | Evaluate Formula — Formula Auditing

Watch Window

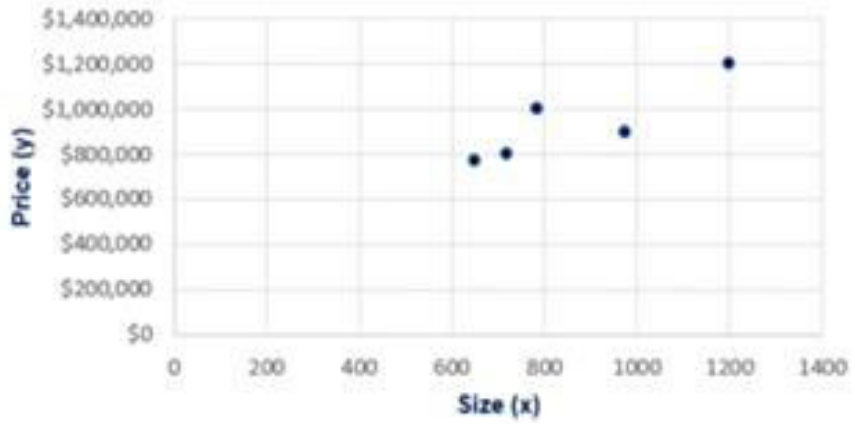
Calculation Options | Calculate Now | Calculate Sheet — Calculation

G14

Covariance

Housing data

	x	y		
	Size (ft.)	Price ($)		$(x-\bar{x})*(y-\bar{y})$
	650	772,000		34,776,000
	785	998,000		- 5,265,000
	1200	1,200,000		89,178,000
	720	800,000		19,418,000
	975	895,000		- 4,142,000
Mean	866	933,000	Sum	133,965,000
			Sample size	5
			Cov. Sample	33,491,250



Covariance gives a sense of direction

> 0, the two variables move together
< 0, the two variables move in opposite directions
= 0, the two variables are independent

NEXT LESSON: CORRELATION COEFFICIENT

File Home Insert Page Layout Formulas Data Review View Power Pivot Tell me what you want to do Share

Paste Cut Copy Format Painter | Arial 9 | Wrap Text Merge & Center | General | Conditional Formatting Format as Table Cell Styles | Insert Delete Format | AutoSum Fill Clear Sort & Filter Find & Select

Clipboard Font Alignment Number Styles Cells Editing

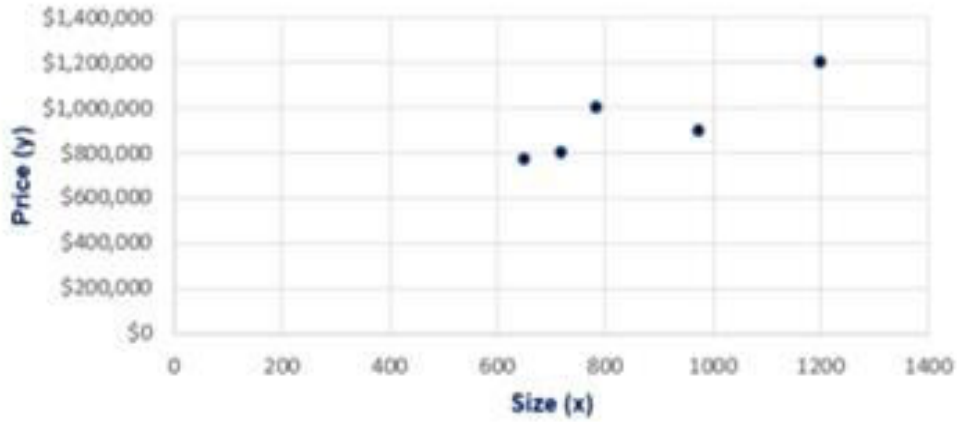
A1

Correlation coefficient

Housing data

	Size (ft.)	Price ($)		(x-x̄)*(y-ȳ)
	650	772,000		34,776,000
	785	998,000		- 5,265,000
	1200	1,200,000		89,178,000
	720	800,000		19,418,000
	975	895,000		- 4,142,000
Mean	866	933,000	**Sum**	133,965,000
Standard dev.	222	173,615	**Sample size**	5
			Cov. Sample	33,491,250

$$\frac{Cov(x,y)}{Stdev(x) * Stdev(y)}$$

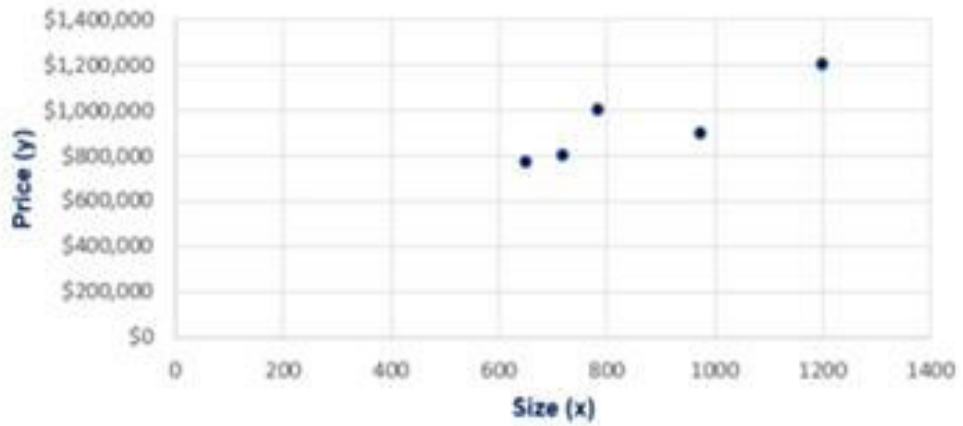


Correlation coefficient

Housing data

	Size (ft.)	Price ($)	(x-x̄)*(y-ȳ)
	650	772,000	34,776,000
	785	998,000	- 5,265,000
	1200	1,200,000	89,178,000
	720	800,000	19,418,000
	975	895,000	- 4,142,000
Mean	866	933,000	
Standard dev.	222	173,615	

Sum	133,965,000
Sample size	5
Cov. Sample	33,491,250



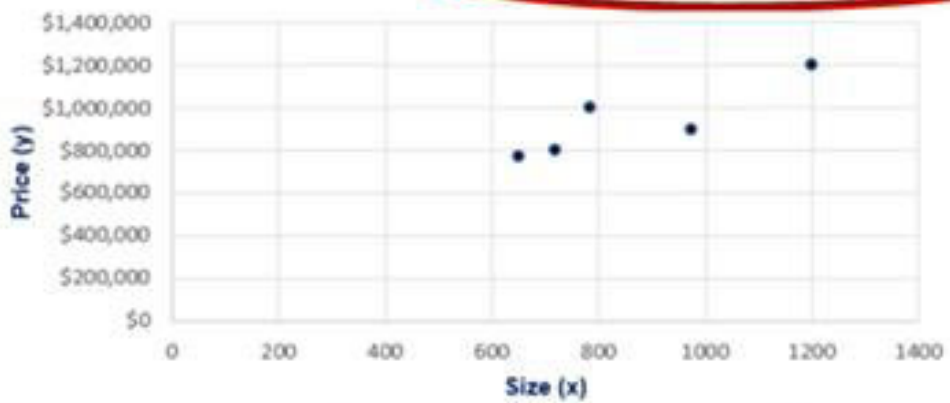
$$\frac{Cov(x,y)}{Stdev(x) * Stdev(y)}$$

$$\frac{S_{xy}}{S_x S_y} \qquad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlation coefficient

Housing data

	Size (ft.)	Price ($)	$(x-\bar{x})*(y-\bar{y})$
	650	772,000	34,776,000
	785	998,000	- 5,265,000
	1200	1,200,000	89,178,000
	720	800,000	19,418,000
	975	895,000	- 4,142,000
Mean	866	933,000	**Sum** 133,965,000
Standard dev.	222	173,615	**Sample size** 5
			Cov. Sample [illegible]
			Correlation coeff. 0.87

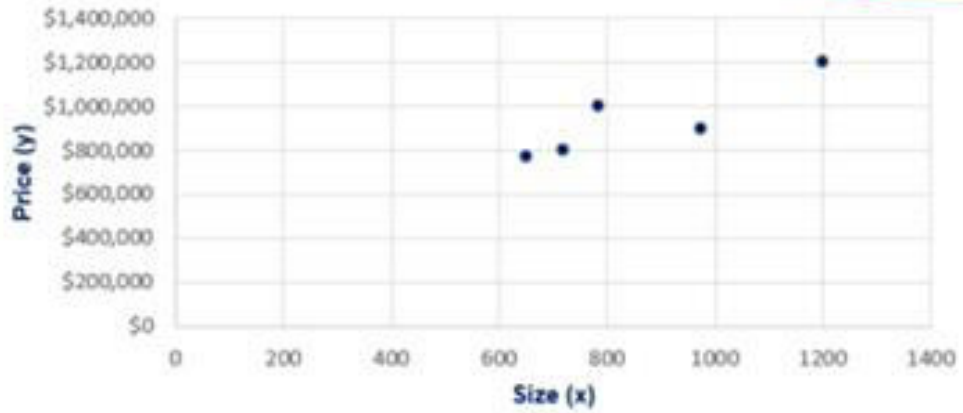


$-1 \leq$ correlation coefficient ≤ 1

Correlation coefficient

Housing data

	Size (ft.)	Price ($)	$(x-\bar{x})*(y-\bar{y})$
	650	772,000	34,776,000
	785	998,000	- 5,265,000
	1200	1,200,000	89,178,000
	720	800,000	19,418,000
	975	895,000	- 4,142,000
Mean	866	933,000	**Sum** 133,965,000
Standard dev.	222	173,615	**Sample size**
			Cov. Sample 33,491,250
			Correlation coeff. 0.87



There is a STRONG relationship between the two variables

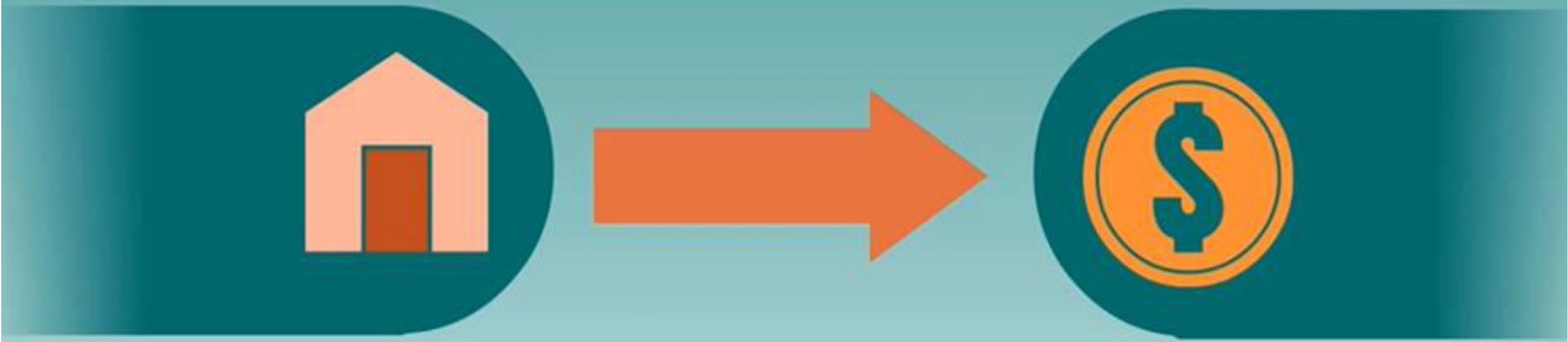
PERFECT POSITIVE CORRELATION

Correlaton coeff.= 1

the entire variability of one variable is explained by the other

RELATIONSHIP DIRECTION

Size determines price



CORRELATION OF 0

Absolutely independent variables

CORRELATION OF 0

They have nothing in common!

Coffee in Brazil

Houses in London

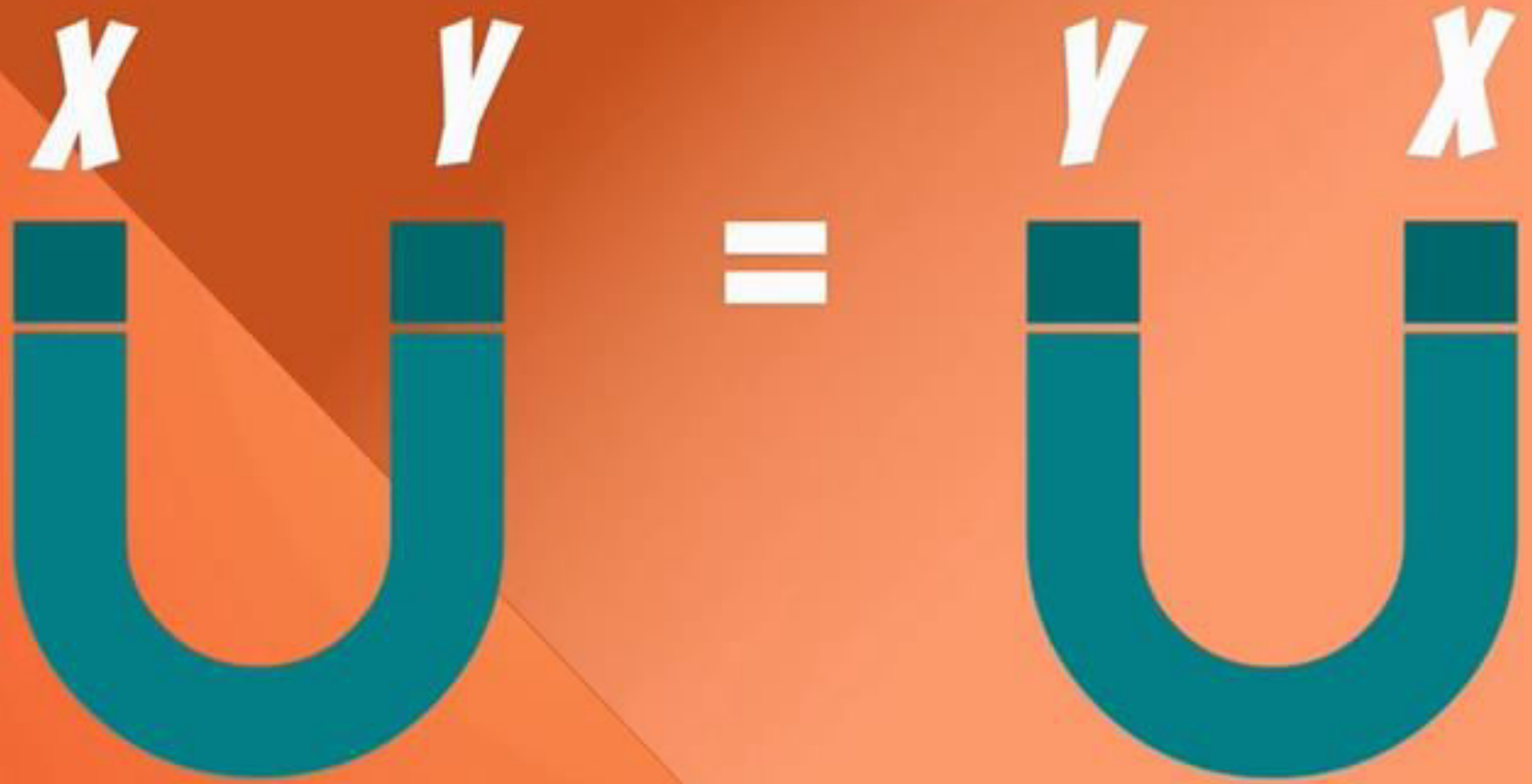
NEGATIVE CORRELATION

Perfect negative correlation of - 1

Imperfect negative correlation: (- 1,0)

NEGATIVE CORRELATION

CORRELATION



CORRELATION

$$\frac{Cov(x, y)}{Stdev(x) * Stdev(y)} = \frac{Cov(y, x)}{Stdev(y) * Stdev(x)}$$

Symmetrical with respect to both variables

CAUSALITY

Important to understand the direction of causal relationships

CAUSALITY

Important to understand the direction of causal relationships

CORRELATION DOES NOT IMPLY CAUSATION

Question 1:

Which of these is true about correlation and causality?

- ◯ Causality is a symmetric relation
- ◯ Correlation in an asymmetric relation (x affects y is different from y affects x)
- ◯ Causality is an asymmetric relation. (x causes y is different from y causes x)
- ◯ None of the above