Course notes: Descriptive statistics

1.60

POPULATION

Collection of all items of interest

N

parameters

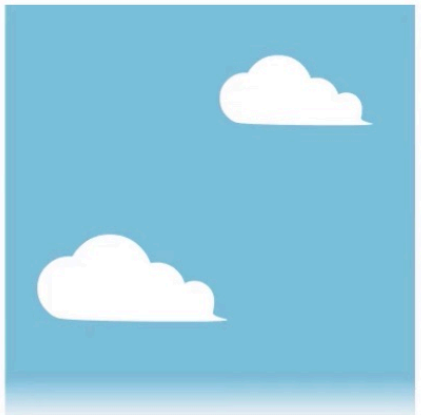
SAMPLE

A subset of the population

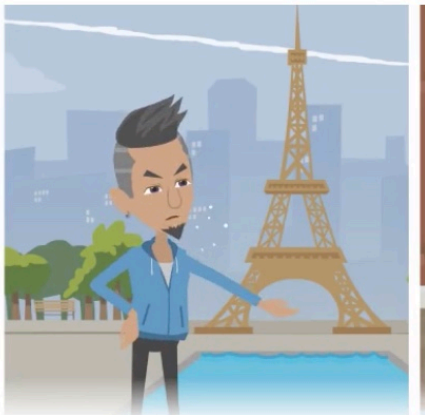
n

statistics

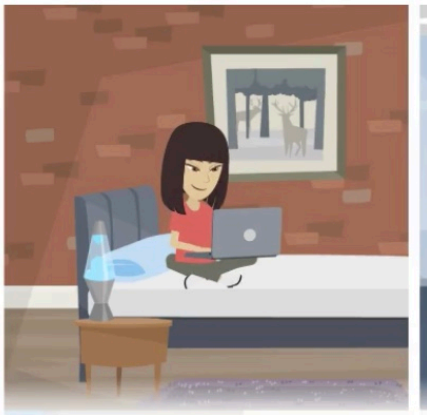
HOME



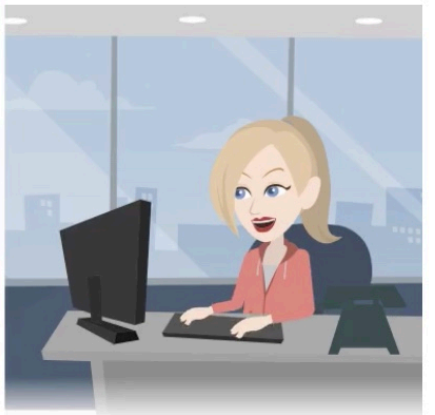
ON EXCHANGE



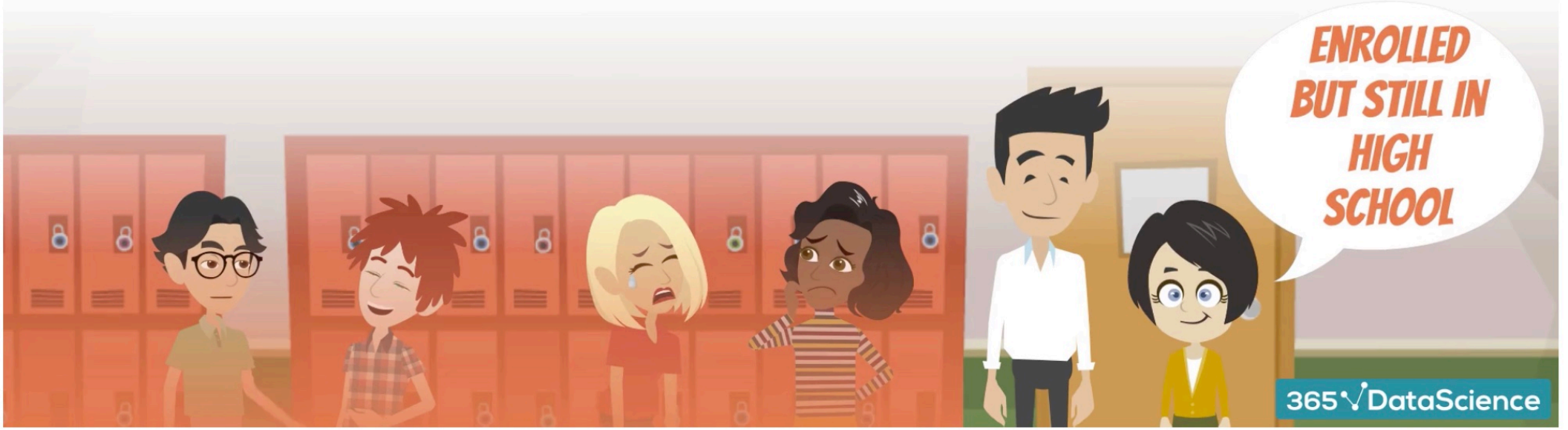
ABROAD



DISTANCE



PART TIME



POPULATIONS ARE HARD TO DEFINE AND HARD TO OBSERVE IN REAL LIFE

SAMPLE

Less time consuming

Less costly (cheaper)

Fullscreen

SAMPLE

Canteen

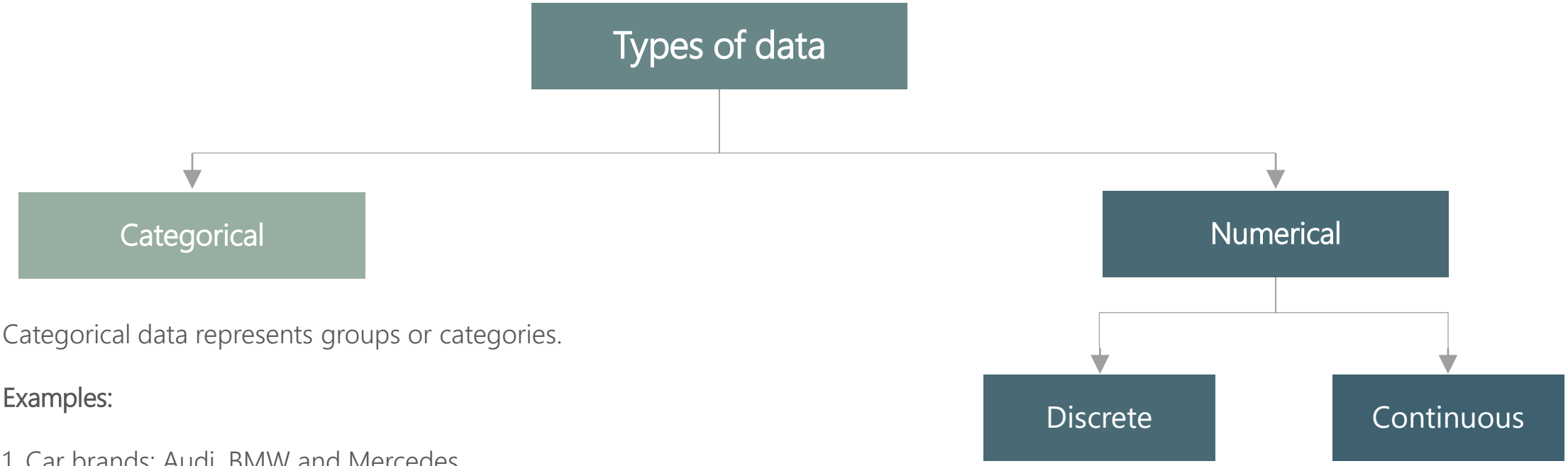
RANDOMNESS

A random sample is collected when each member of the sample is chosen from the population strictly by chance.

REPRESENTATIVENESS

A representative sample is a subset of the population that accurately reflects the members of the entire population.

Types of data



Categorical data represents groups or categories.

Examples:

1. Car brands: Audi, BMW and Mercedes.

2. Answers to yes/no questions: yes and no

Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.

Examples:
Discrete: # children you want to have, SAT score
Continuous: weight, height

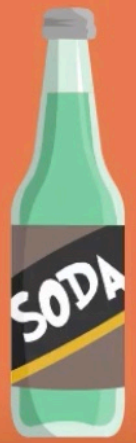
Discrete: You can imagine all possible outcomes

1.50

EXAMPLES OF DISCRETE

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

Grades



#1,000

number of objects

Money

1.50

EXAMPLES OF CONTINUOUS

Height

Area

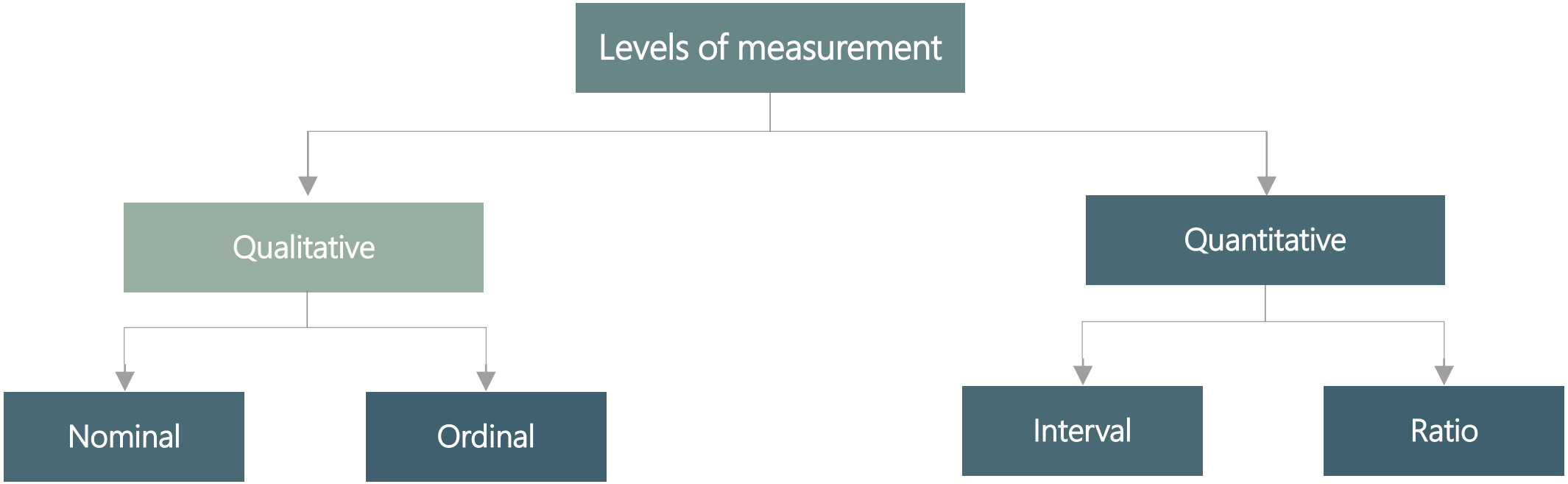
Distance

Time

TIME ON A CLOCK IS DISCRETE

TIME IN GENERAL IS CONTINUOUS

Levels of measurement



There are two qualitative levels: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that **can** be ordered.

Examples:
Nominal: four seasons (winter, spring, summer, autumn)
Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

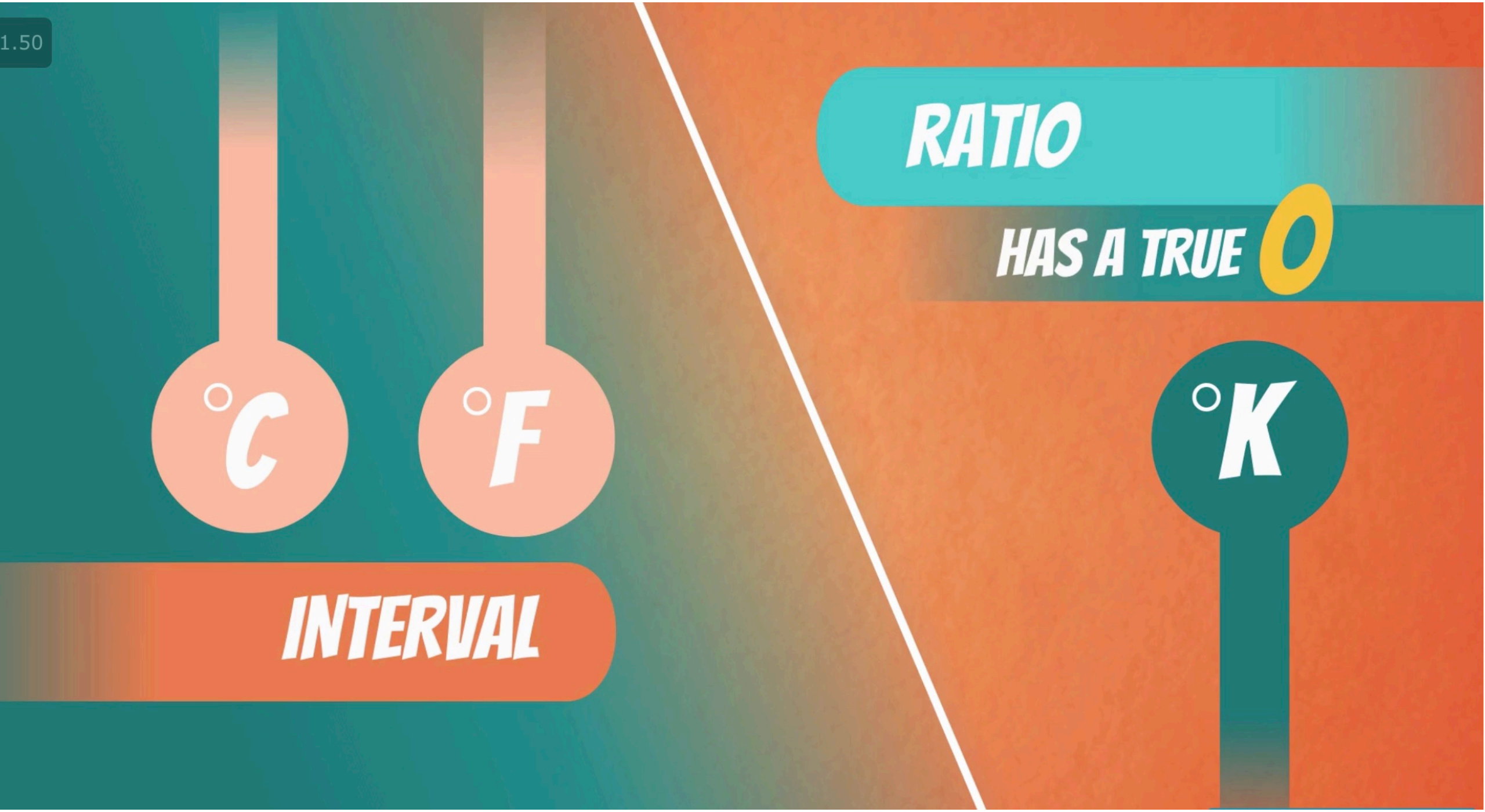
There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios **have a true zero,** while intervals don't.

Examples:
Interval: degrees Celsius and Fahrenheit
Ratio: degrees Kelvin, length

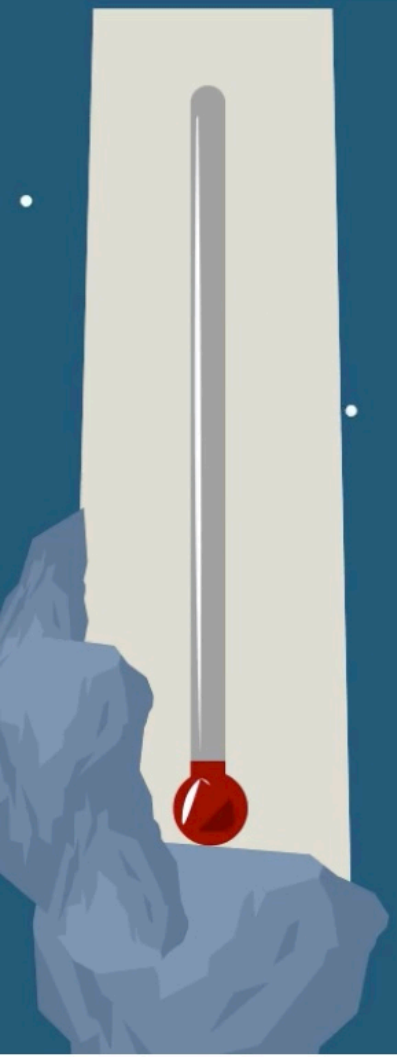
RATIO

You have 3 times as many as I do.

RATIO OF 6/2 IS 3


1.50
°C
°F
INTERVAL
RATIO
HAS A TRUE 0
°K

INTERVAL



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

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

Graphs and tables that represent categorical variables

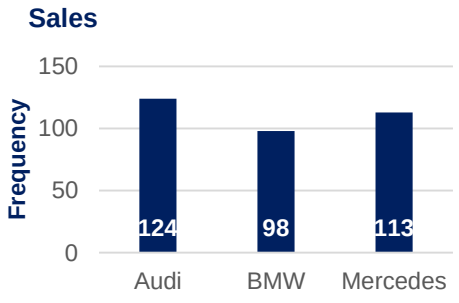
All graphs are very easy to create and read, once you have identified the type of data you are dealing with.

Frequency distribution tables

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

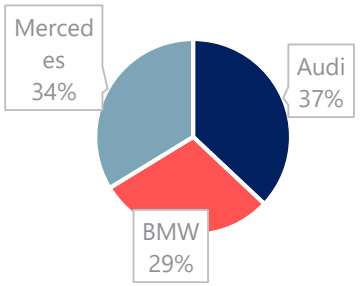
Frequency distribution tables show the category and its corresponding absolute frequency.

Bar charts



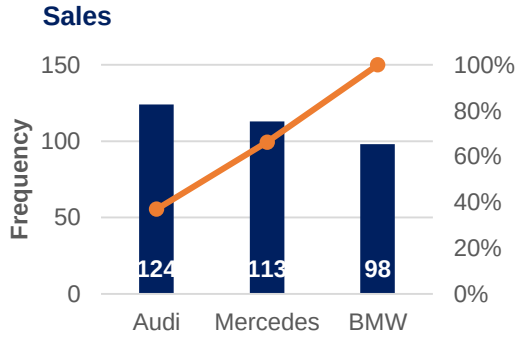
Bar charts are very common. Each bar represents a category. On the y-axis we have the absolute frequency.

Pie charts



Pie charts are used when we want to see the share of an item as a part of the total. Market share is almost always represented with a pie chart.

Pareto diagrams



The Pareto diagram is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency.

Graphs and tables that represent categorical variables. Excel formulas

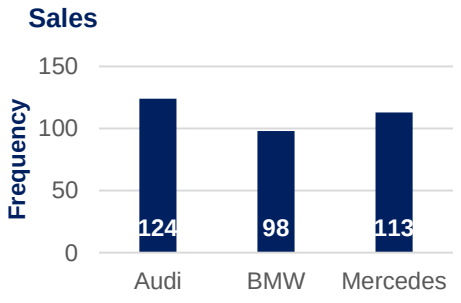
Frequency distribution tables

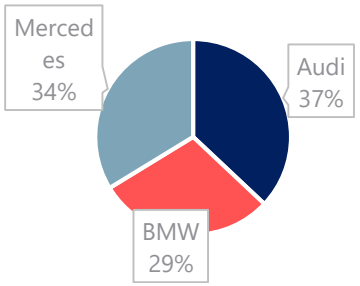
Bar charts

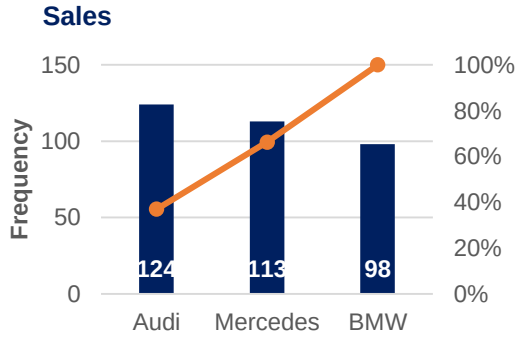
Pie charts

Pareto diagrams

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335







In Excel, we can either hard code the frequencies or count them with a count function. This will come up later on.
Total formula: =SUM()

Bar charts are also called clustered column charts in Excel. Choose your data, Insert -> Charts -> Clustered column or Bar chart.

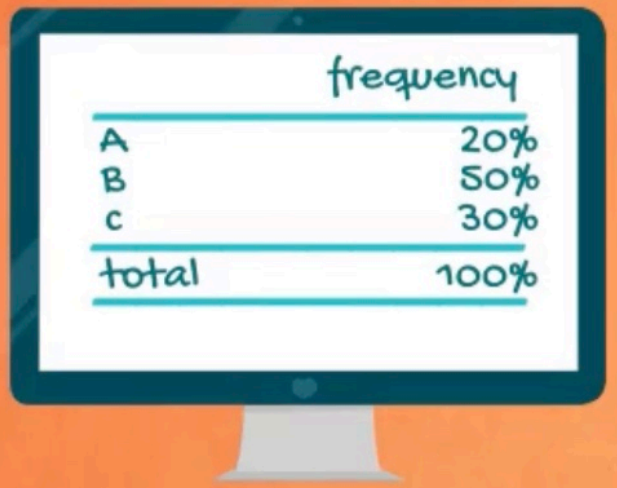
Pie charts are created in the following way:
Choose your data, Insert -> Charts -> Pie chart

Next slide.

1.60

REPRESENTATION OF CATEGORICAL VARIABLES

Frequency distribution tables



Bar charts

Pie charts

Pareto diagrams

1.60

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

Percentage Number

Conditional Formatting Format as Table Cell Styles Styles

Insert Delete Format Cells

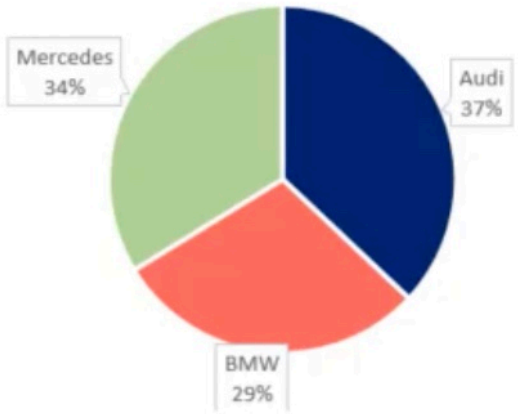
AutoSum Fill Clear Sort & Filter Find & Select Editing

SUM =C5/$C$8

Graphs and tables for categorical variables

German car shop

	Frequency	Relative frequency
Audi	124	=C5/$C$8
BMW	98	29%
Mercedes	113	34%
Total	335	100%

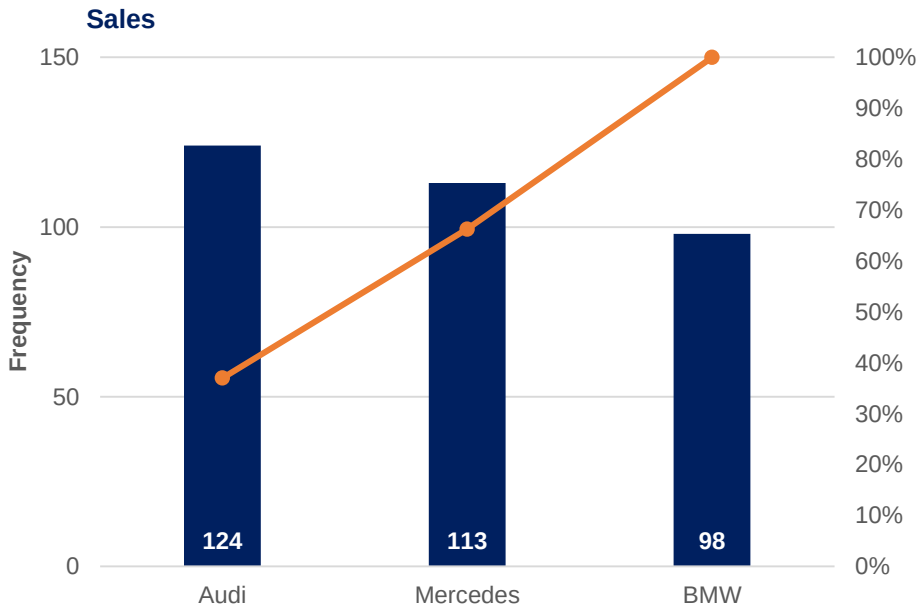


Relative frequency is the percentage of the total frequency for each category

1:57

Frequency distribution table | bar chart | Pie chart | Pareto diagram | Pareto diagram Part2

Pareto diagrams in Excel



A Pareto diagram is a special type of bar chart, where categories are shown in descending order of freqency

Frequency is the number of occurrences of each item.

Cumulative Frequency is the sum of the relative frequency (curve on the graph)

Creating Pareto diagrams in Excel:

1. Order the data in your frequency distribution table in descending order.
2. Create a bar chart.
3. Add a column in your frequency distribution table that measures the cumulative frequency.
4. Select the plot area of the chart in Excel and **Right click**.
5. Choose **Select series.**
6. Click **Add**
7. Series name doesn't matter. You can put 'Line'
8. For **Series values** choose the cells that refer to the cumulative frequency.
9. Click **OK.** *You should see two side-by-side bars.*
10. Select the plot area of the chart and **Right click.**
11. Choose **Change Chart Type.**
12. Select **Combo.**
13. Choose the type of representation from the dropdown list. Your initial categories should be **'Clustered Column'.** Change the second series, that you called 'Line', to **'Line'.**
14. Done.

1.60

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

Paste Arial 9 Wrap Text General AutoSum Fill Clear

Conditional Formatting Format as Table Cell Styles Insert Delete Format Sort & Filter Find & Select

Clipboard Font Alignment Number Styles Cells Editing

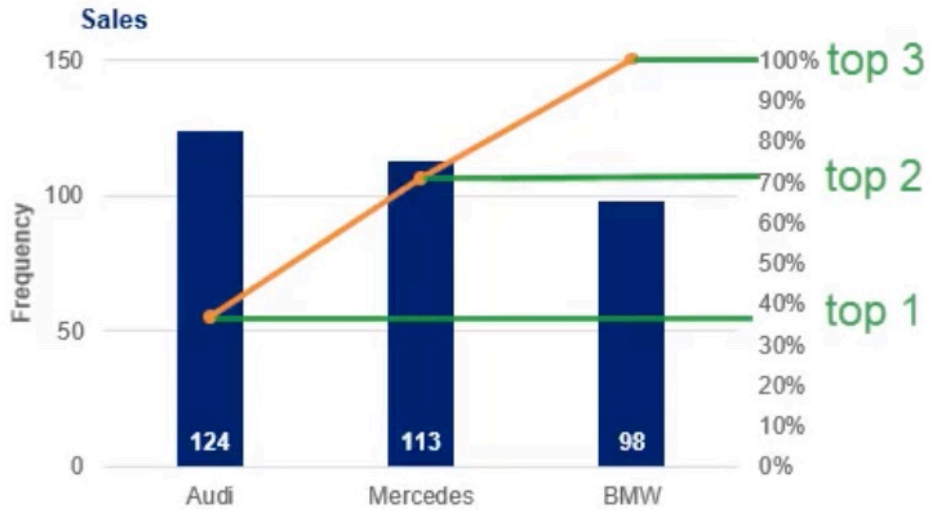
U30

Graphs and tables for categorical variables

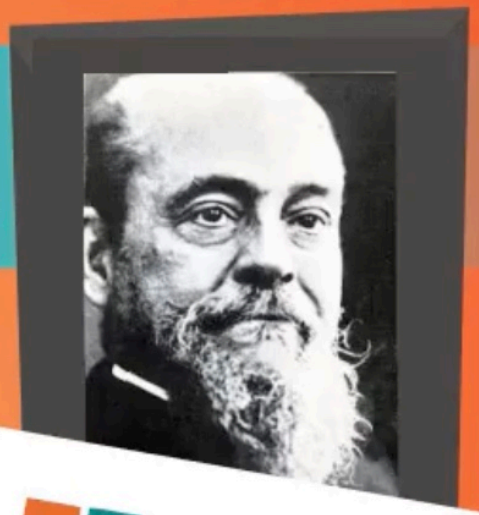
German car shop

	Frequency	Relative frequency
Audi	124	37%
BMW	98	29%
Mercedes	113	34%
Total	335	100%

Ordered	Frequency	Relative frequency	Cumulative frequency
Audi	124	37%	37%
Mercedes	113	34%	71%
BMW	98	29%	100%



PARETO PRINCIPLE



80% of the effect come form 20% of the causes

80-20 RULE

MICROSOFT

fixing 20% of its software bugs, they manage to solve 80% of the problems

PARETO DIAGRAM

It shows how subtotals change with each additional category and provide us with a better understanding of our data.

Numerical variables. Frequency distribution table and histogram

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

Frequency distribution tables for numerical variables are different than the ones for categorical. Usually, they are divided into intervals of equal (or unequal) length. The tables show the interval, the absolute frequency and sometimes it is useful to also include the relative (and cumulative) frequencies.

The interval width is calculated using the following formula:

$$Interal\ width = \frac{Largest\ number\ - smallest\ number}{Number\ of\ desired\ intervals}$$

En general, stastician prefer 5 to 20 itervals

Creating the frequency distribution table in Excel:

1. Decide on the number of intervals you would like to use.
2. Find the interval width (using a the formula above).
3. Start your 1st interval at the lowest value in your dataset.
4. Finish your 1st interval at the lowest value + the interval width. (= start_interval_cell + interval_width_cell)
5. Start your 2nd interval where the 1st stops (that's a formula as well - just make the starting cell of interval 2 = the ending of interval 1)
6. Continue in this way until you have created the desired number of intervals.
7. Count the absolute frequencies using the following COUNTIF formula:
   =COUNTIF(dataset_range,">="&interval start) -COUNTIF(dataset_range,">"&interval end).
8. In order to calculate the relative frequencies, use the following formula: = absolute_frequency_cell / number_of_observations
9. In order to calculate the cumulative frequencies:
   i. The first cumulative frequency is equal to the relative frequency
   ii. Each consequitive cumulative frequency = previous cumulative frequency + the respective relative frequency

Note that all formulas could be found in the lesson Excel files and the solutions of the exercises provided with each lesson.

1.50

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

Paste Arial 9 Wrap Text 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

Z18

Graphs and tables for numerical variables. Frequency distribution table

Dataset	Frequency
1	1
9	1
22	1
24	1
32	1
41	1
44	1
48	1
57	1
66	1
70	1
73	1
75	1
76	1
79	1
82	1
87	1
89	1
95	1
100	1
Total	20

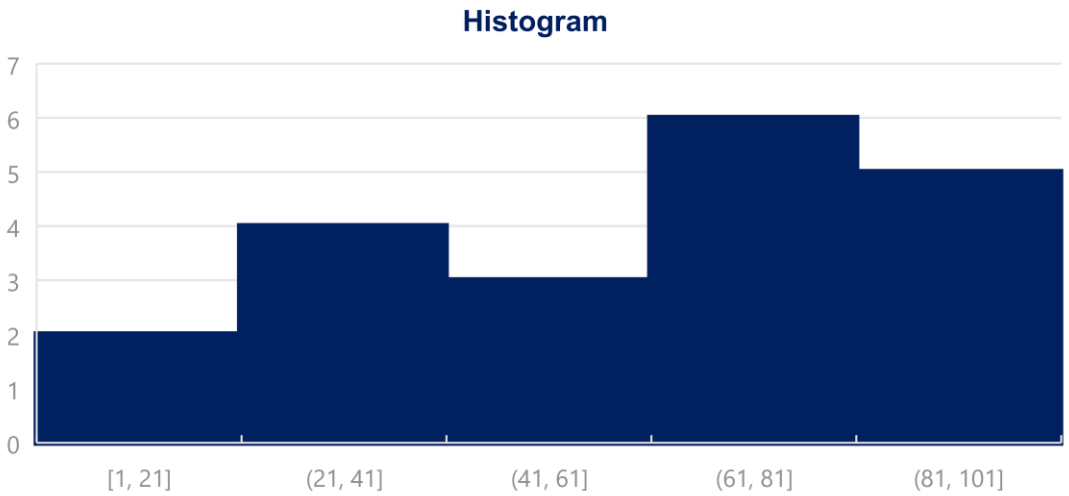
Frequency distribution table

Interval width 20

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

$$\text{relative frequency} = \frac{\text{Frequency}}{\text{Total frequency}} = \frac{2}{20} = 0.10$$

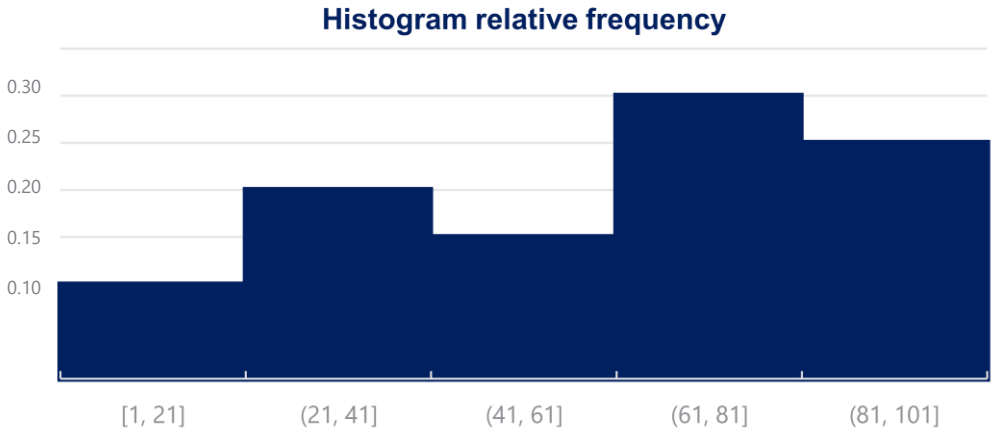
Numerical variables. Frequency distribution table and histogram



Histograms are the one of the most common ways to represent numerical data. Each bar has width equal to the width of the interval. The bars are touching as there is continuation between intervals: where one ends -> the other begins.

Creating a histogram in Excel:

1. Choose your data
2. Insert -> Charts -> Histogram
3. To change the number of bins (intervals):
    1. Select the x-axis
    2. Click Chart Tools -> Format -> Axis options
    3. You can select the bin width (interval width), number of bins, etc.



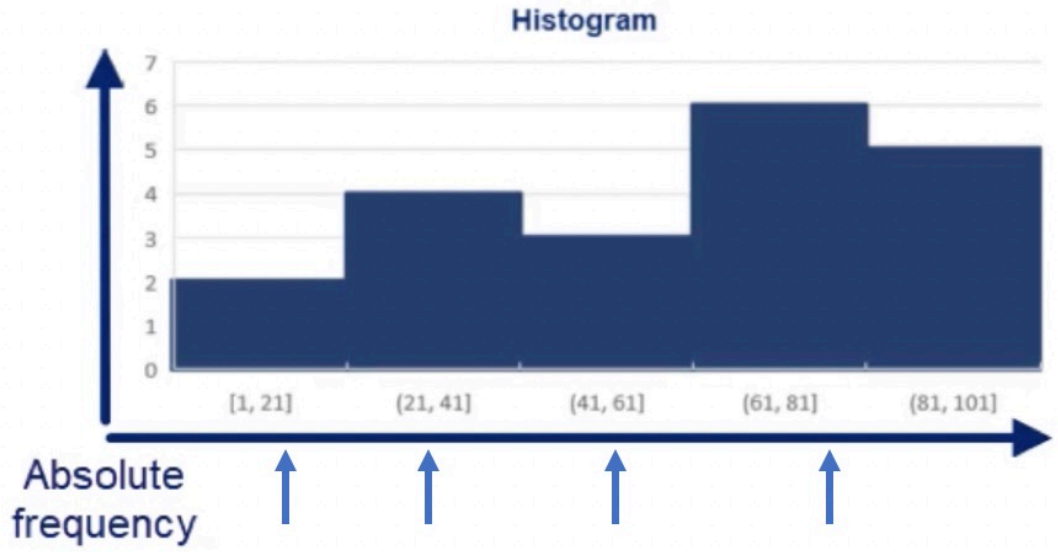
1.50

The histogram

Data set
1
9
22
24
32
33
42
44
57
66
70
73
75
76
79
82
87
89
95
100

Frequency distribution table

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25



Horizontal axis is numerical, too.

Bars do not have gaps as they represent intervals.
While the bars in in the bar charts represent categories

1.60

ONLY ONE VARIABLE

Variable: Units sold

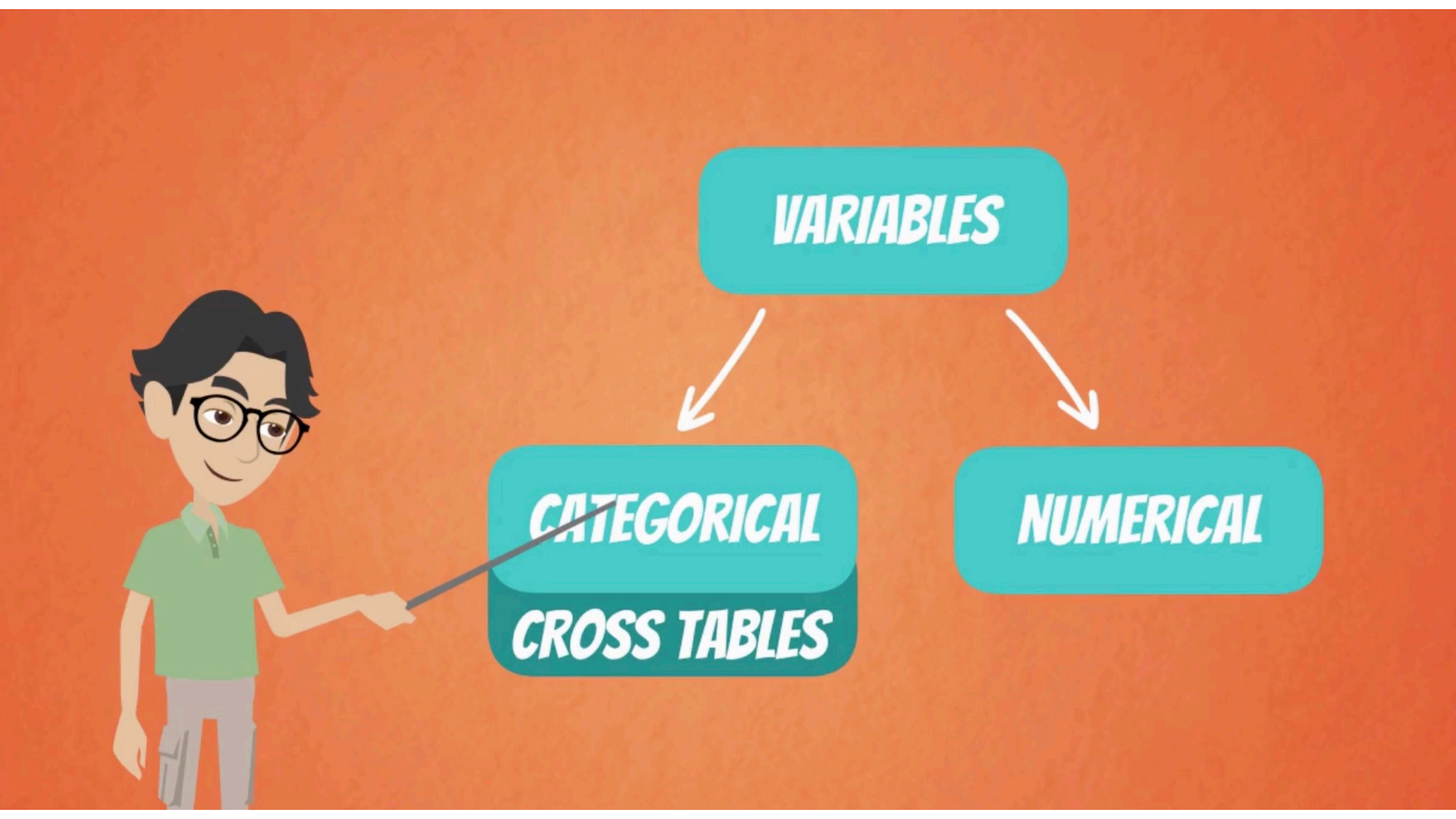
HOW DO WE REPRESENT RELATONSHIPS BETWEEN

TWO VARIABLES?

CROSS TABLES

investment	investitor		total
stocks	96	185	281
bonds	181	3	184
real estate	88	152	240
total	365	340	705

SCATTER PLOTS


VARIABLES
CATEGORICAL
CROSS TABLES
NUMERICAL

1.50

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

Paste Arial 9 Wrap Text 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

Z18

Graphs and tables for numerical variables. Frequency distribution table

Dataset	Frequency
1	1
9	1
22	1
24	1
32	1
41	1
44	1
48	1
57	1
66	1
70	1
73	1
75	1
76	1
79	1
82	1
87	1
89	1
95	1
100	1
Total	20

Frequency distribution table

Interval width 20

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

$$\text{relative frequency} = \frac{\text{Frequency}}{\text{Total frequency}} = \frac{2}{20} = 0.10$$

1.60

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

Calibri 11 — 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

Graphs and tables for relationships between variables

Cross table

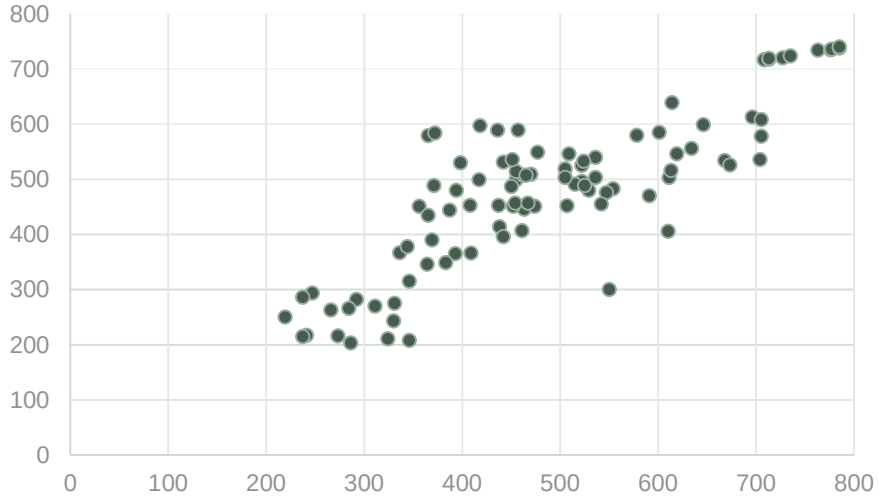
Type of investment \ Investor	Investor A	Investor B	Investor C	Total	
Stocks	96	185	39	320	Total investment in stocks
Bonds	181	3	29	213	Total investment in bonds
Real Estate	88	152	142	382	Total investment in real estate
Total	365	340	210	915	

↑ ↑ ↑

Holdings of each investor

Graphs and tables for relationships between variables. Scatter plots

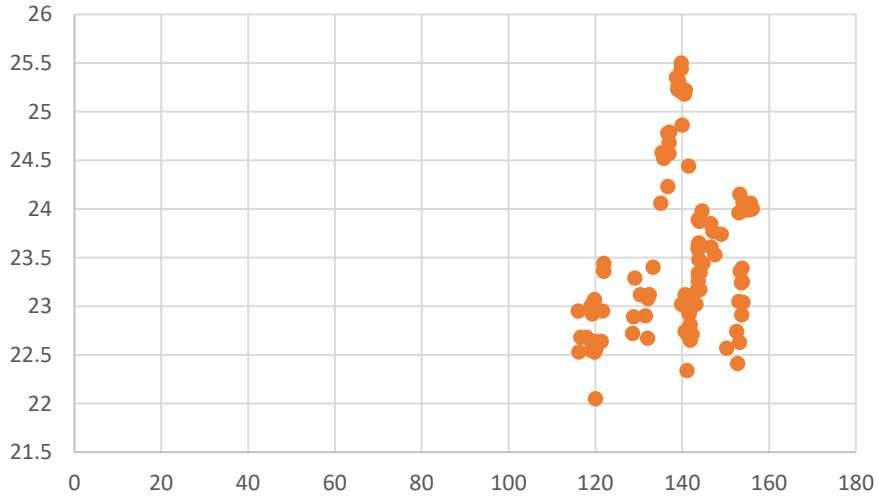
Representing two numerical variables



When we want to represent two numerical variables on the same graph, we usually use a scatter plot. Scatter plots are useful especially later on, when we talk about regression analysis, as they help us detect patterns (linearity, homoscedasticity).
Scatter plots usually represent lots and lots of data. Typically, we are not interested in single observations, but rather in the structure of the dataset.

Creating a scatter plot in Excel:

1. Choose the two datasets you want to plot.
2. Insert -> Charts -> Scatter



A scatter plot that looks in the following way (down) represents data that **doesn't have a pattern.** Completely vertical 'forms' show no association.

Conversely, the plot above shows a linear pattern, meaning that the observations move together.

Graphs and tables for relationships between variables. Cross tables

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	0.10	0.20	0.04	0.35
Bonds	0.20	0.00	0.03	0.23
Real Estate	0.10	0.17	0.16	0.42
Total	0.40	0.37	0.23	1.00

Cross tables (or contingency tables) are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals. Sometimes, these tables are constructed with the *relative frequencies* as shown in the table below.

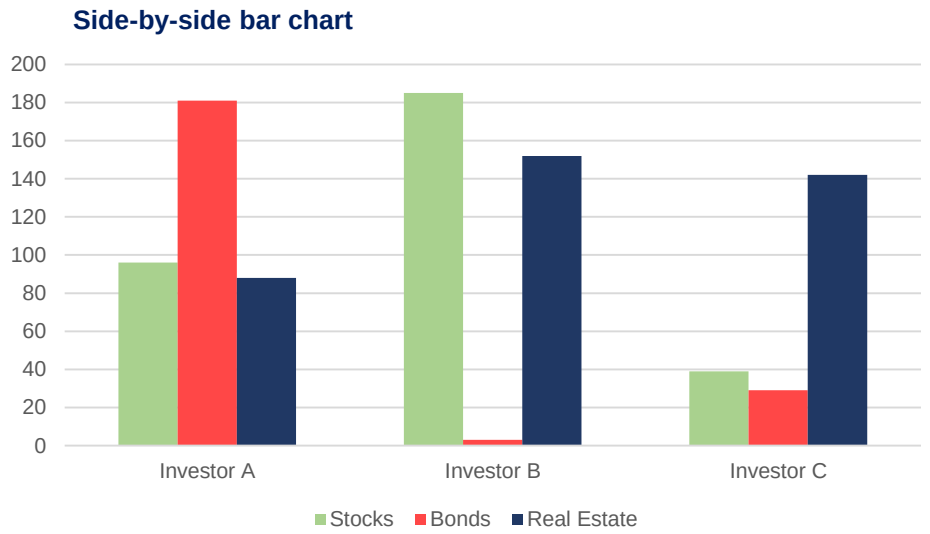
A common way to represent the data from a cross table is by using a side-by-side bar chart.

Creating a side-by-side chart in Excel:

1. Choose your data
2. Insert -> Charts -> Clustered Column

Selecting more than one series (groups of data) will automatically prompt Excel to create a side-by-side bar (column) chart.

The side by side bar chart is a variation of the bar chart



1.60

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

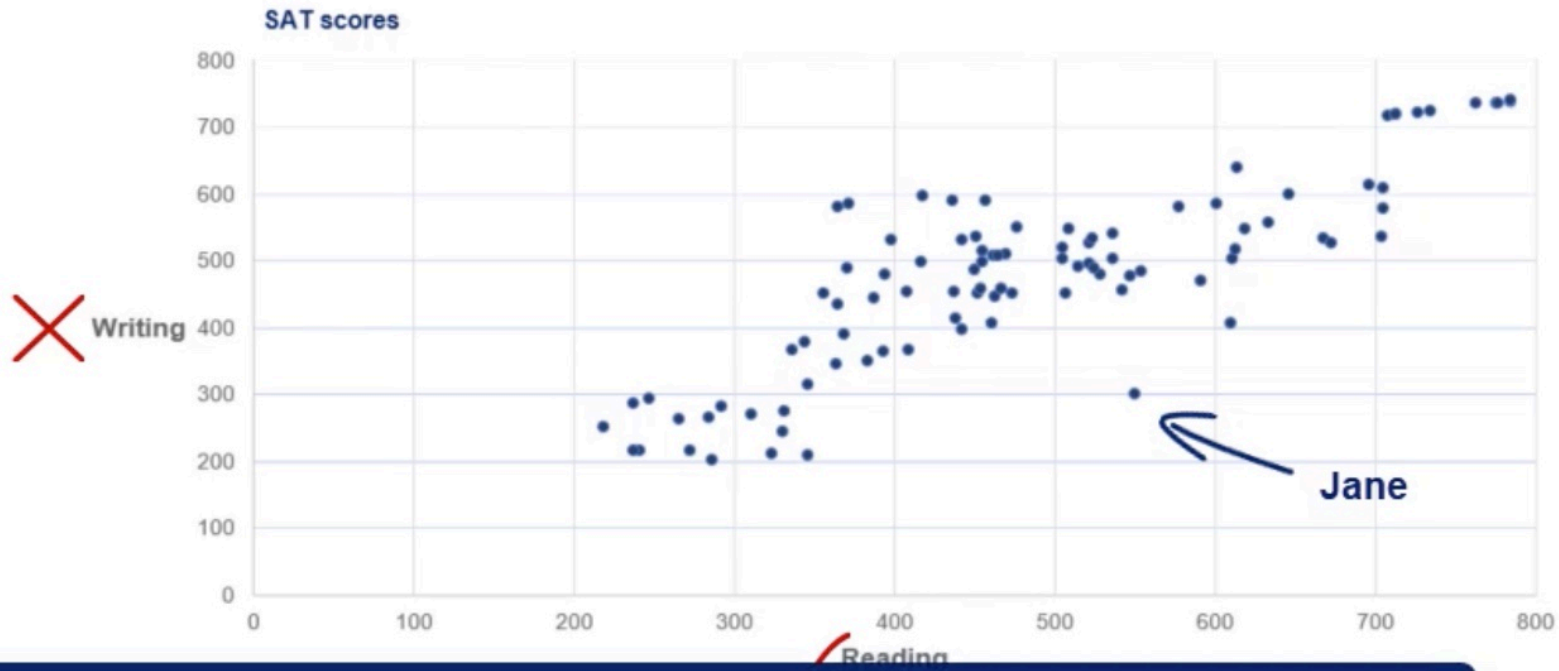
AutoSum Fill Clear Sort & Filter Find & Select Editing

A1

Graphs and tables for relationships between variables

Scatter plot

Student ID	Reading	Writing
1	273	216
2	292	282
3	219	250
4	241	217
5	284	266
6	247	294
7	237	215
8	286	203
9	237	286
10	266	263
11	311	270
12	324	211
13	330	243
14	331	275
15	336	367
16	344	378
17	346	315
18	346	208
19	356	451
20	364	346
21	365	435
22	365	579
23	369	390
24	436	589
25	393	365
26	394	480
27	417	499



Outliers are data points that go against the logic of the whole dataset

365 DataScience

Mean, median, mode

Mean

The mean is the most widely spread measure of central tendency. It is the simple average of the dataset.

Note: easily affected by outliers

The formula to calculate the mean is:

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

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

In Excel, the mean is calculated by:

=AVERAGE()

Median

The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.

In an ordered dataset, the median is the number at position $\frac{n+1}{2}$.

If this position is not a whole number, it, the median is the simple average of the two numbers at positions closest to the calculated value.

In Excel, the median is calculated by:

=MEDIAN()

Mode

The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes.

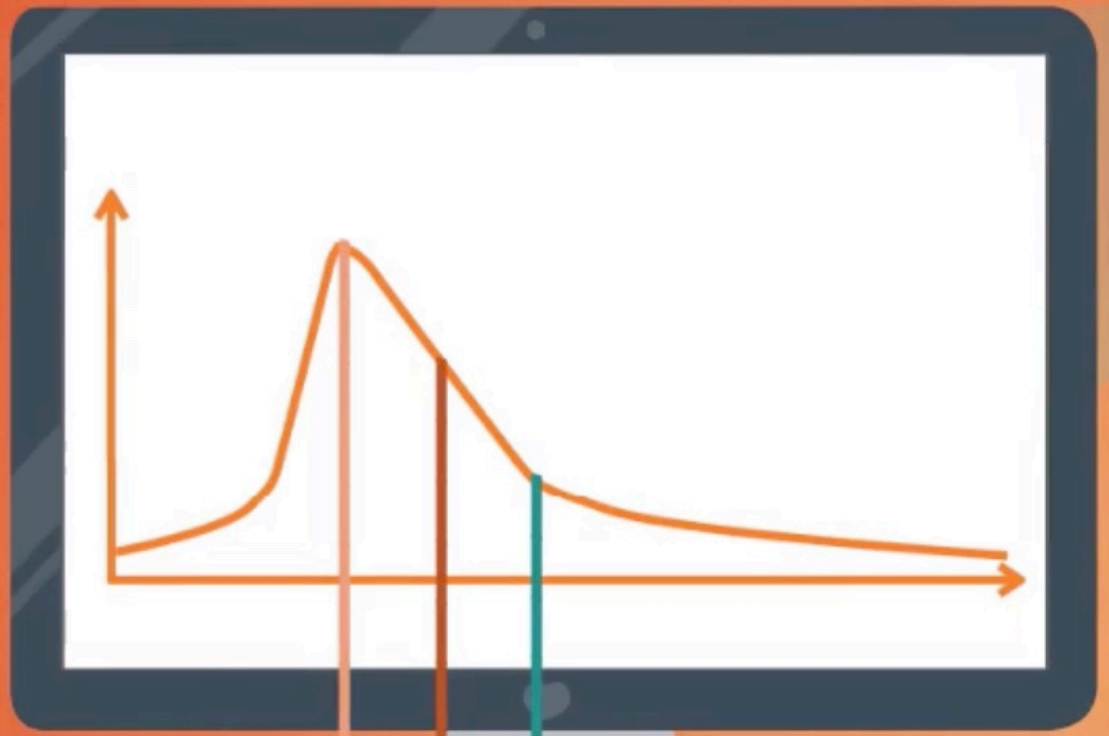
The mode is calculated simply by finding the value with the highest frequency.

In Excel, the mode is calculated by:

=MODE.SNGL() -> returns one mode

=MODE.MULT() -> returns an array with the modes. It is used when we have more than 1 mode.

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

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

.60

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

1.50

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	

	New York City	Los Angeles
Mean	$ 11.00	$ 5.50

The mean is not enough to make definite conclusions!

1:50

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	-

Each price appears only once... we say that there is NO mode

1:50

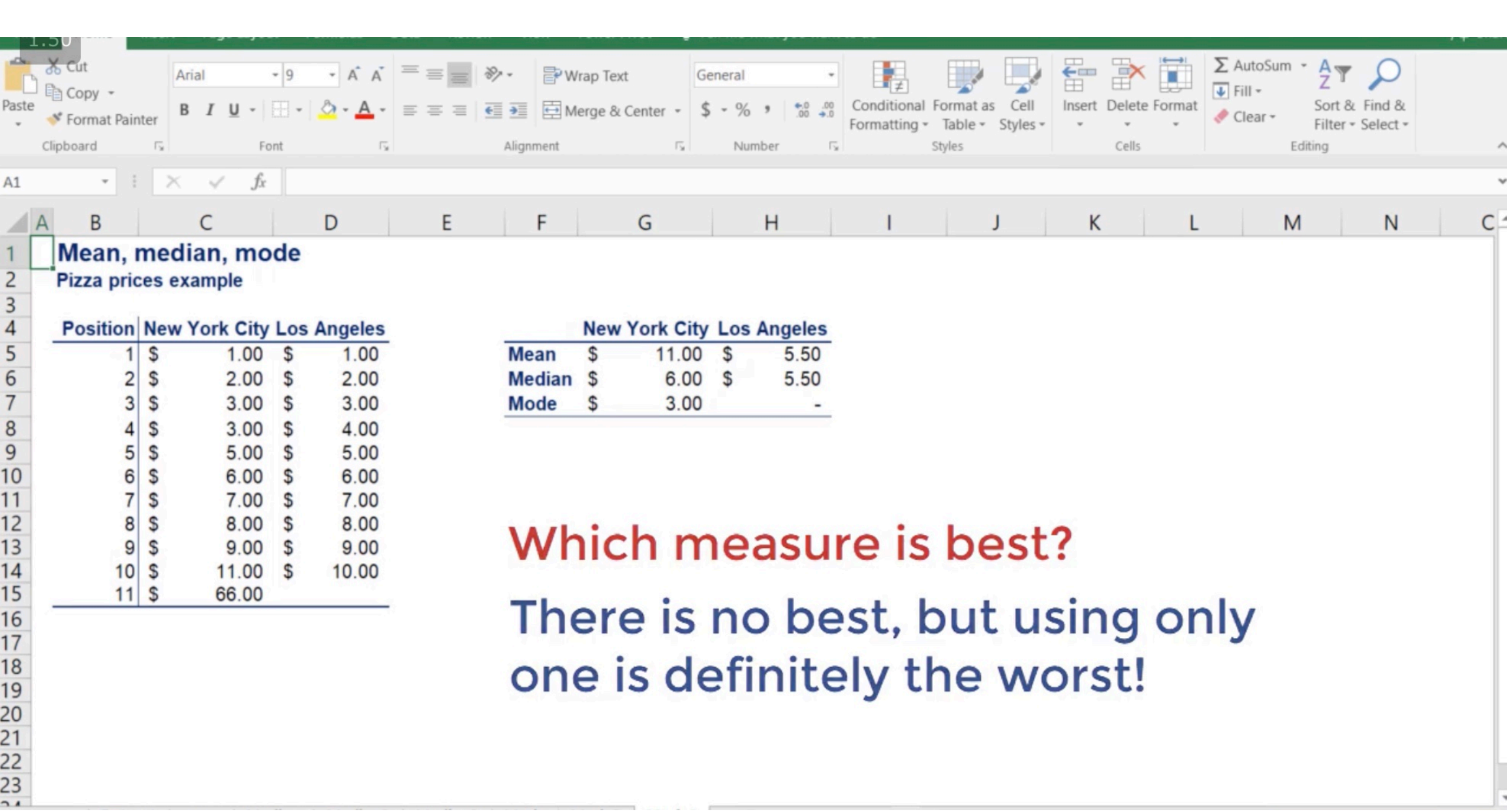
Mean, median, mode

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	-

Each price appears only once... we say that there is NO mode


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
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

1.50

Mean, median, mode

Pizza prices example

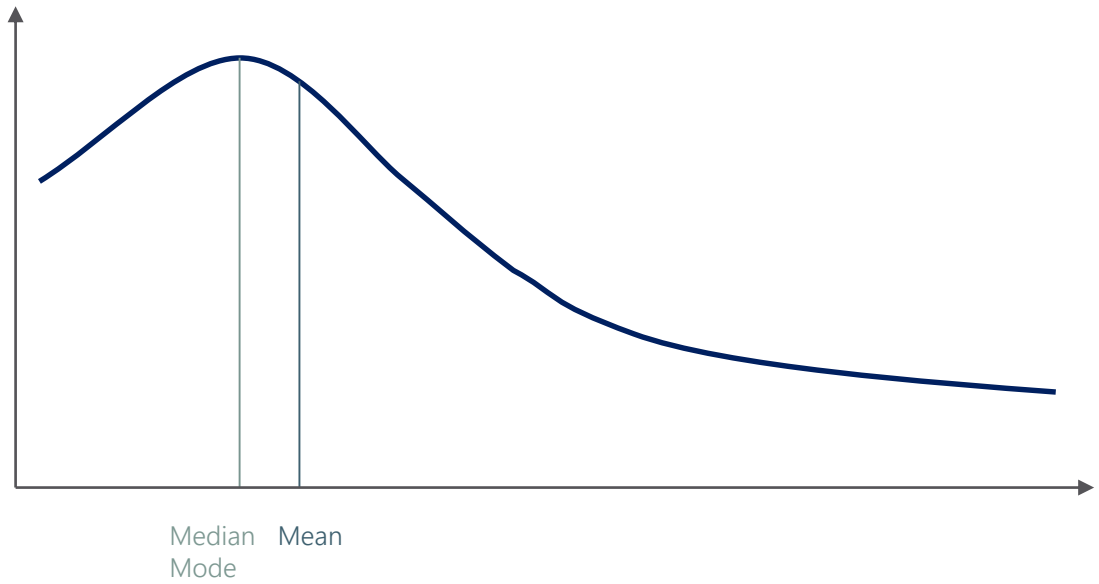
Position	New York City	Los Angeles
1	$ 1.00	$ 1.00
2	$ 2.00	$ 2.00
3	$ 3.00	$ 3.00
4	$ 3.00	$ 4.00
5	$ 5.00	$ 5.00
6	$ 6.00	$ 6.00
7	$ 7.00	$ 7.00
8	$ 8.00	$ 8.00
9	$ 9.00	$ 9.00
10	$ 11.00	$ 10.00
11	$ 66.00	

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

($5 + $6)/2 = $5.5

Median in LA = (10+1)/2 = 5.5th position

Skewness



Skewness is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side.

Right (positive) skewness looks like the one in the graph. It means that the **outliers** are to the right (long tail to the right).

Left (negative) skewness means that the outliers are to the left.

Usually, you will use software to calculate skewness.

Calculating skewness in Excel:

=SKEW()

Skewness shows to which side is the longer tail,
not where the data is concentrated.

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

1.60

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

mean > median

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

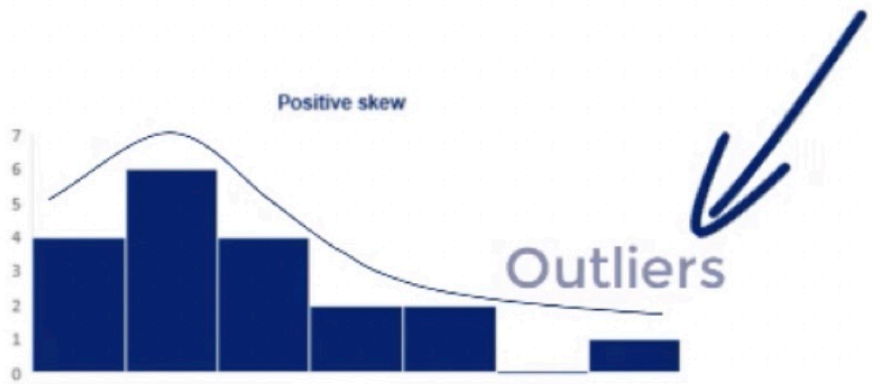
Negative (left)

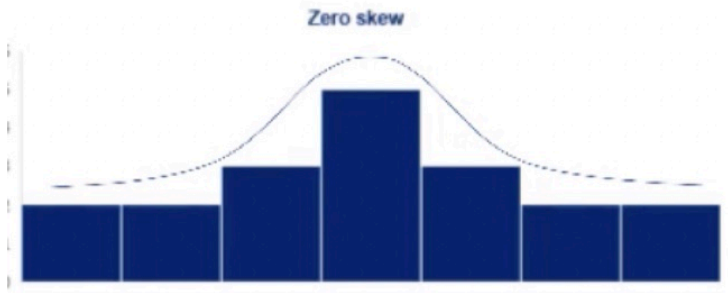
Dataset 3
1
2
3
3
4
4
4
5
5
5
5
6
6
6
6
6
7
7
7

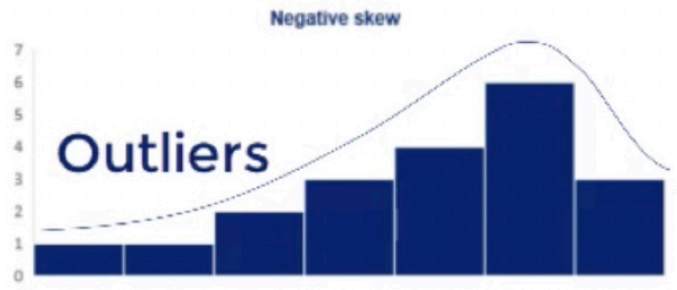
Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

Mean	Median	Mode
4.90	5.00	6.00

mean < median







1:50

Mean, median, mode

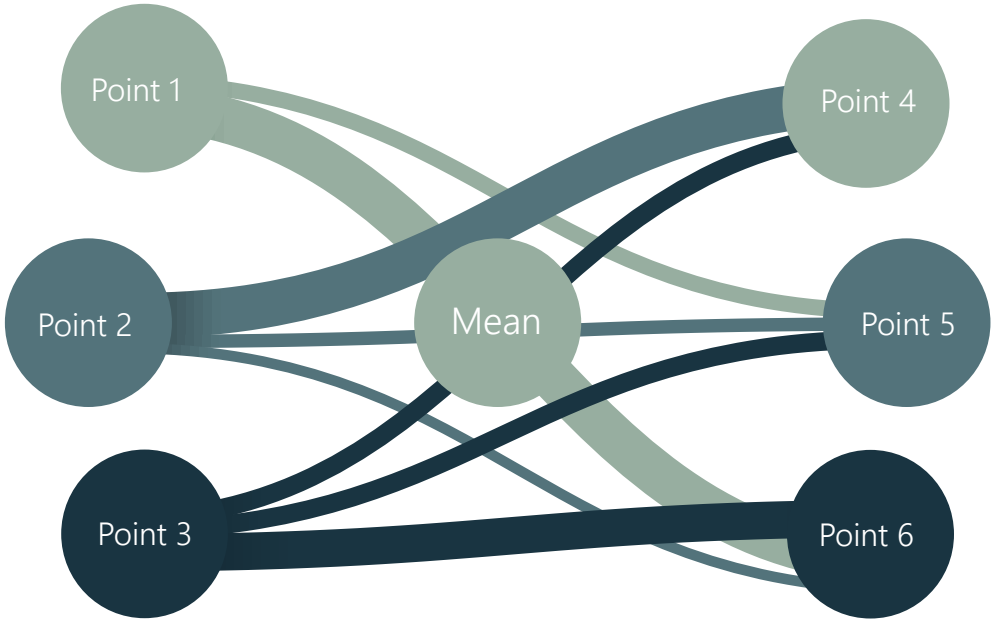
Pizza prices example

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	-

Each price appears only once... we say that there is NO mode

Variance and standard deviation



Variance and standard deviation measure the dispersion of a set of data points around its mean value.

There are different formulas for population and sample variance & standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas. More on the mathematics behind it.

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

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

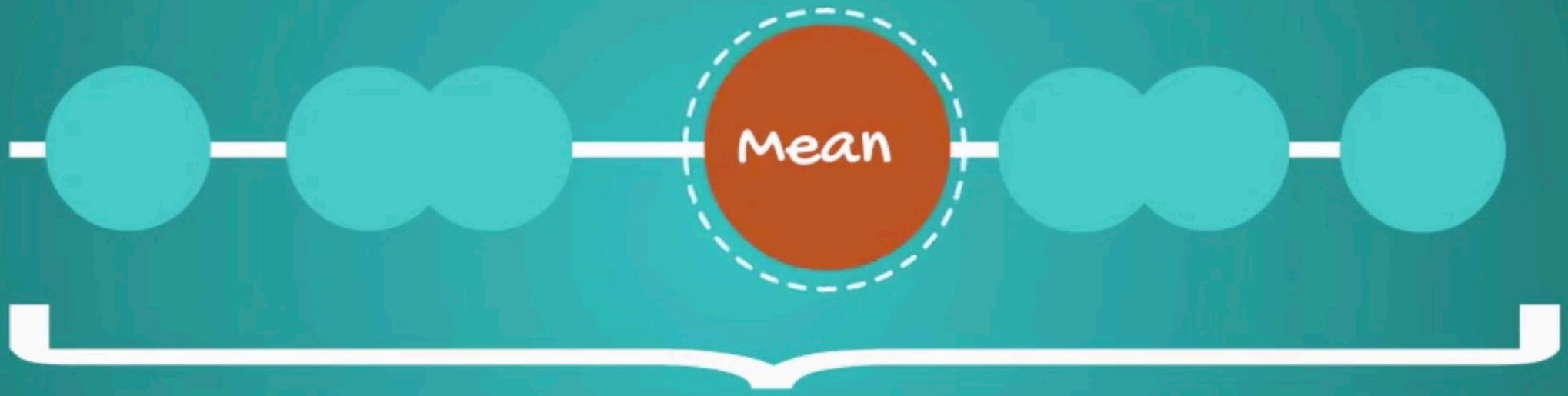
Sample standard deviation formula: $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

Population standard deviation formula: $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

Calculating variance in Excel:

Sample variance: =VAR.S()
Population variance: =VAR.P()
Sample standard deviation: = STDEV.S()
Population standard deviation: =STDEV.P()

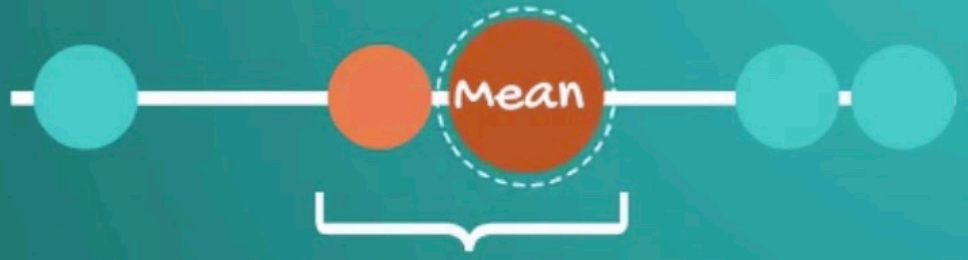
VARIANCE



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

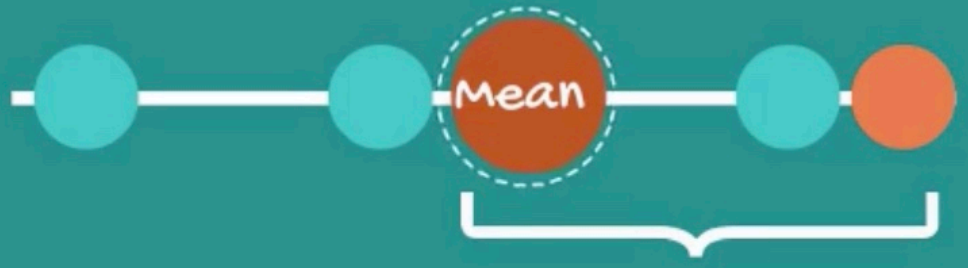
VARIANCE

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



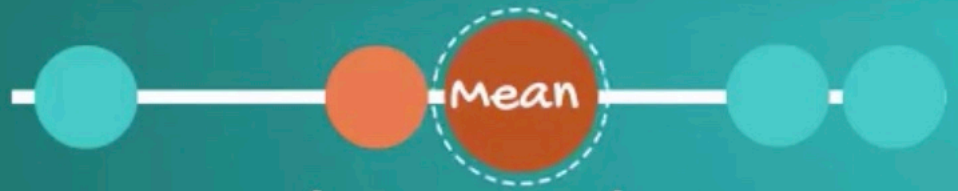
lower result

population variance



higher result

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



lower result ^{2}

10 SQUARED IS 100

higher result ^{2}

100 SQUARED IS 10,000!

1.60

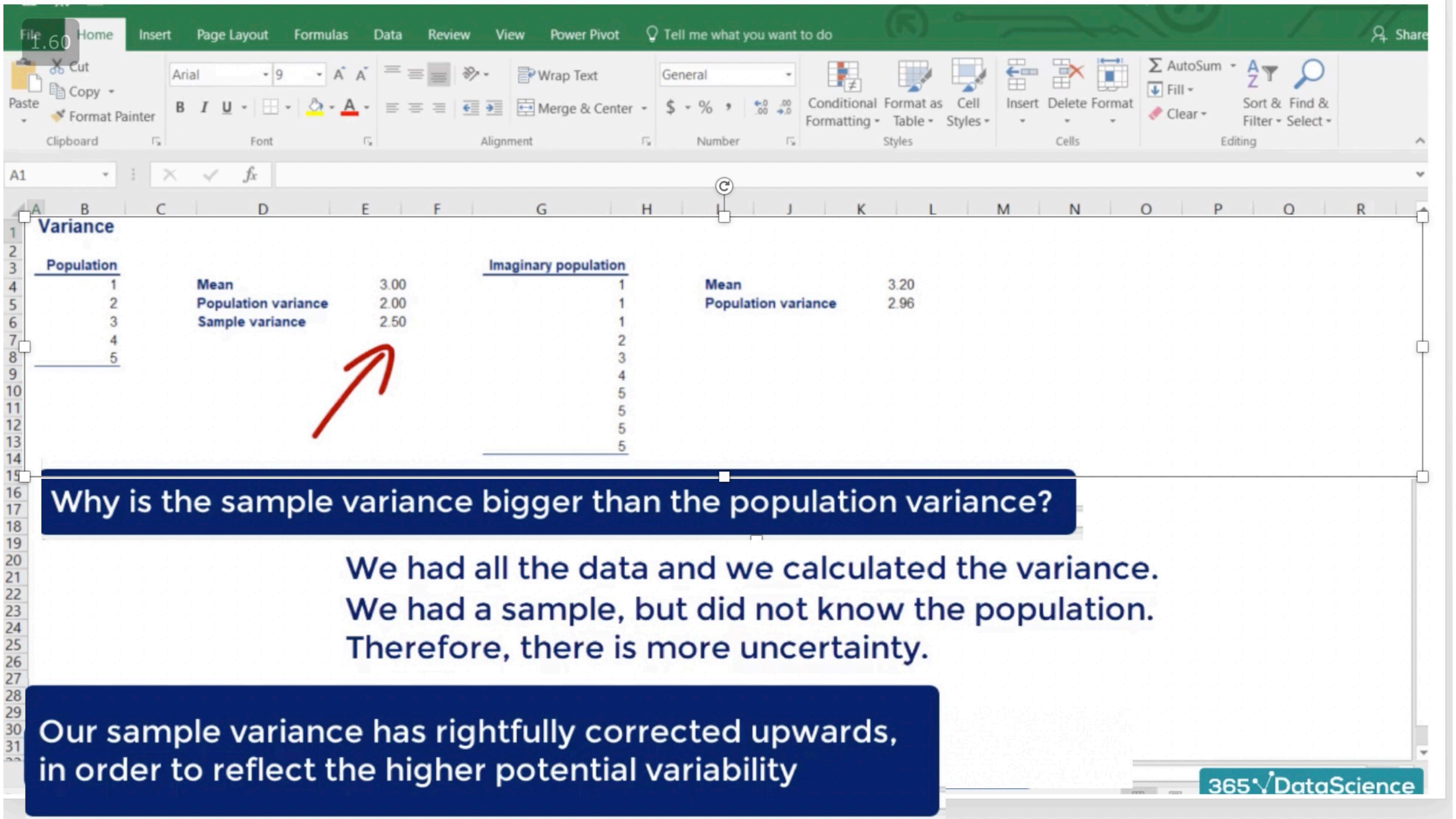
VARIANCE

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

population variance

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

sample variance


1.60
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Power Pivot
Tell me what you want to do
Share
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
AutoSum
Fill
Clear
Sort & Filter
Find & Select
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
Why is the sample variance bigger than the population variance?
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.
Our sample variance has rightfully corrected upwards,
in order to reflect the higher potential variability
365 DataScience

STANDARD DEVIATION FORMULAS

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

population standard deviation

sample standard deviation

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

1.50

COEFFICIENT OF VARIATION (CV)

/relative standard deviation/

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

COEFFICIENT OF VARIATION (CV)

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

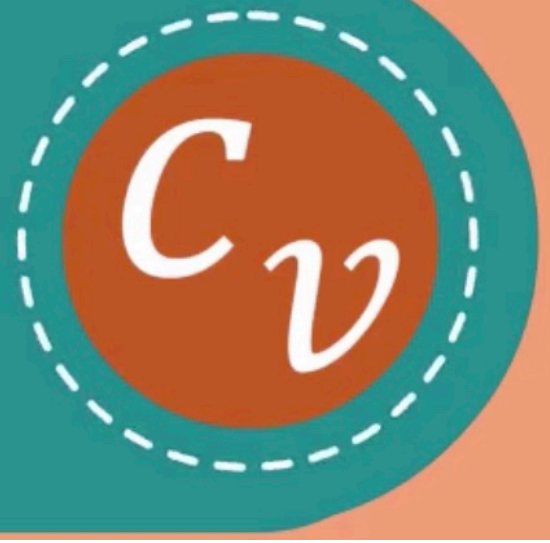
Population formula

Sample formula

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

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

Comparing TWO OR MORE datasets

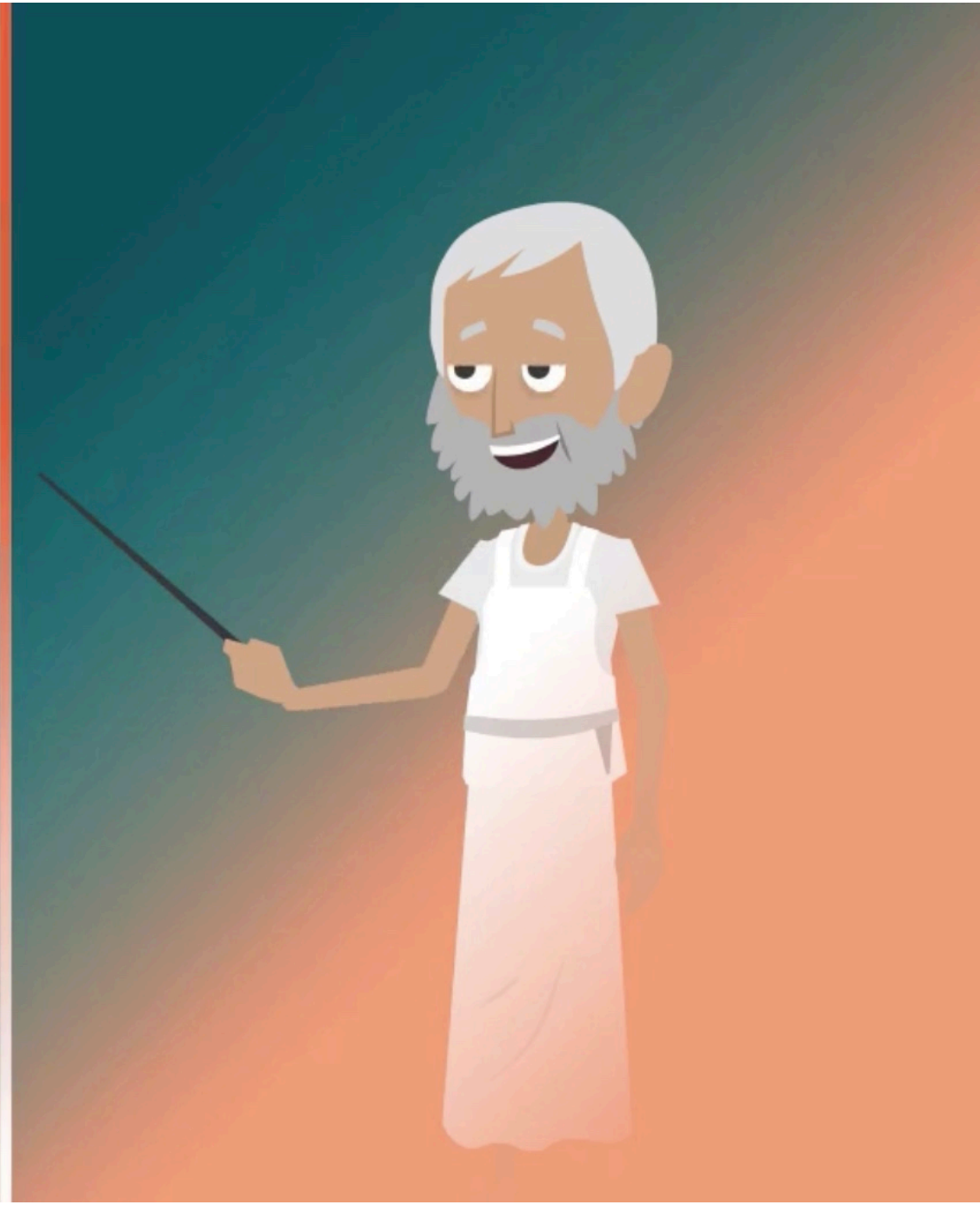


1.50

Tell me, I'll forget

Show me, I'll remember

Involve me, I'll understand



1.50

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	MXN^2 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

1.50

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	MXN^2	3793.69
Sample standard deviation	$ 3.27	MXN	61.59
Sample coefficient of variation	0.60		0.60

Same pizza, same restaurants, but different standard deviations...

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

- **Both datasets have same variability**

$$CV = \frac{s}{\bar{x}}$$

Dataset Pesos | Mean | Mean2 | Variance | Standard deviation | Coefficient of variation

Ready

365 DataScience

1.50

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

($5 + $6)/2 = $5.5

Median in LA = (10+1)/2 = 5.5th position

1.50

REAL ESTATE

what determines house prices?

1.50

REAL ESTATE

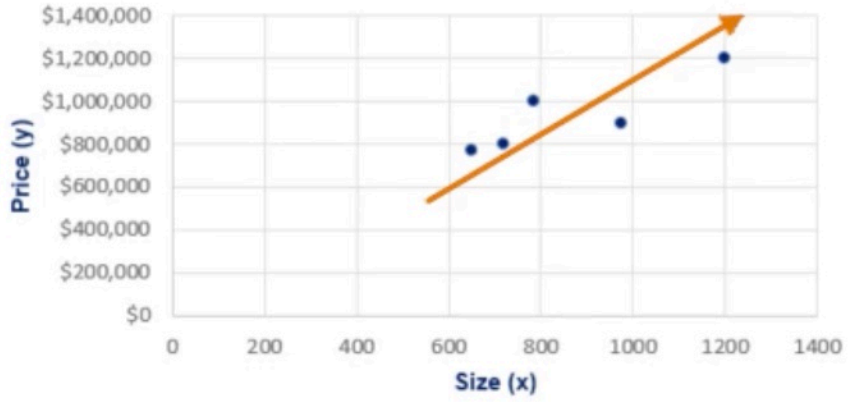
Their size!

1.50

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

Sample formula

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

Population formula

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

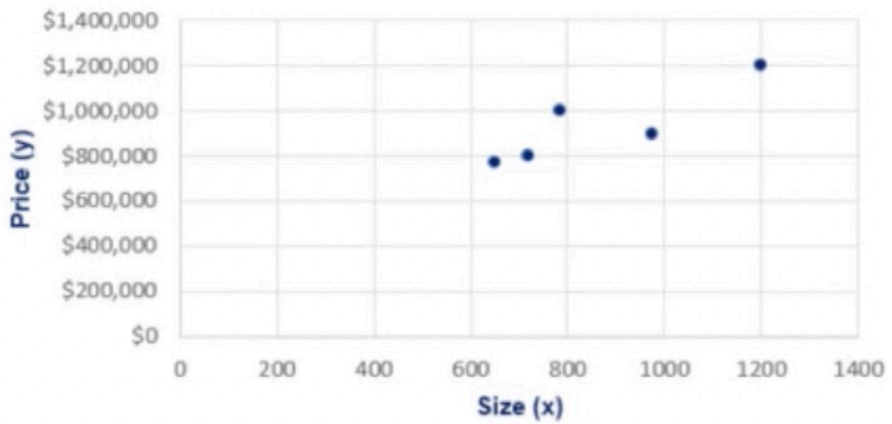
Covariance may be:

> 0
= 0
< 0

Covariance

Housing data

	x	y		
	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
			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

But the value itself does not have a comparative meaning.

- 5
- 50
- 0.0023456
- 33,491,250

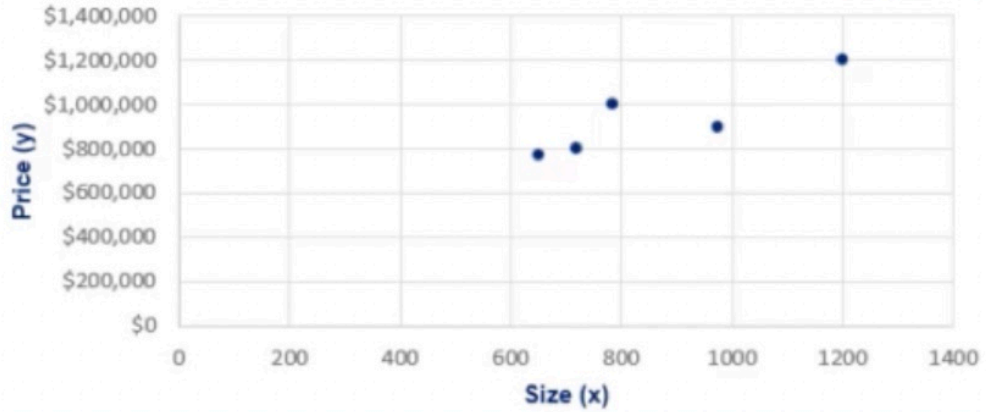
We manipulated the strange covariance value in order to get something intuitive

$-1 \leq \text{correlation coefficient} \leq 1$

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
			Cov. Sample 33,491,250
			Correlation coeff. 0.87



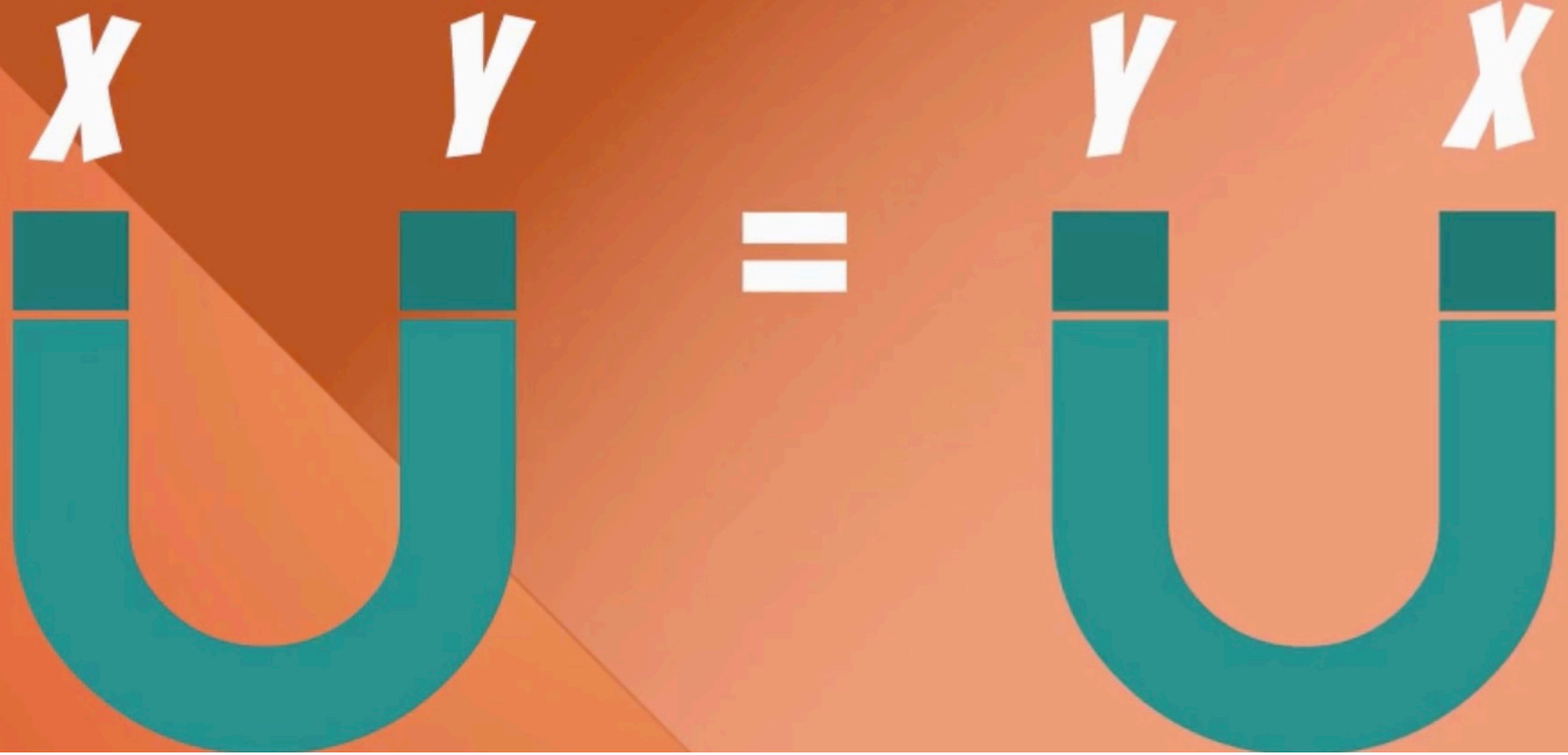
$$\frac{Cov(x,y)}{Stdev(x) * Stdev(y)}$$

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

A common practice is to disregard correlations below 0.2
The data is dispersed and there isn't obvious trend.
And y does not depend on x

There is a STRONG relationship between the two variables

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

CORRELATION DOES NOT IMPLY CAUSATION

PERFECT POSITIVE CORRELATION

X Y

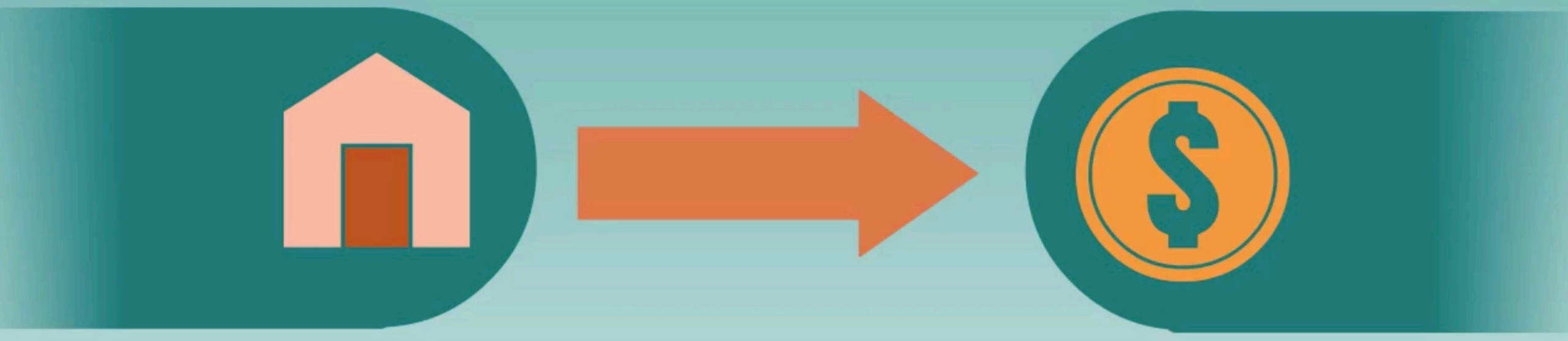
Correlaton coeff.= 1

the entire variability of one variable is explained by the other

1.60

RELATIONSHIP DIRECTION

Size determines price



CORRELATION OF 0

Absolutely independent variables

Coffee in Brazil

Houses in London

NEGATIVE CORRELATION

Perfect negative correlation of - 1

Imperfect negative correlation: (- 1,0)

NEGATIVE CORRELATION

Covariance and correlation

Covariance

Covariance is a measure of the joint variability of two variables.

- A positive covariance means that the two variables move together.
- A covariance of 0 means that the two variables are independent.
- A negative covariance means that the two variables move in opposite directions.

Covariance can take on values from -∞ to +∞ . This is a problem as it is very hard to put such numbers into perspective.

Sample covariance formula: $$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})*(y_i-\bar{y})}{n-1}$$

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

In Excel, the covariance is calculated by:

Sample covariance: =COVARIANCE.S()
Population covariance: =COVARIANCE.P()

Correlation

Correlation is a measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result.

- A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other.
- A correlation of 0 means that the variables are independent.
- A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.

Sample correlation formula: $$r = \frac{s_{xy}}{s_x s_y}$$

Population correlation formula: $$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

In Excel, correlation is calculated by:

=CORREL()

DATABASE OF A REAL ESTATE COMPANY OPERATING IN CALIFORNIA

THE COMPANY IS LAUNCHING A MARKETING CAMPAIGN
BUT IT WANTS TO TARGET ITS AUDIENCE PROPERLY

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

A1

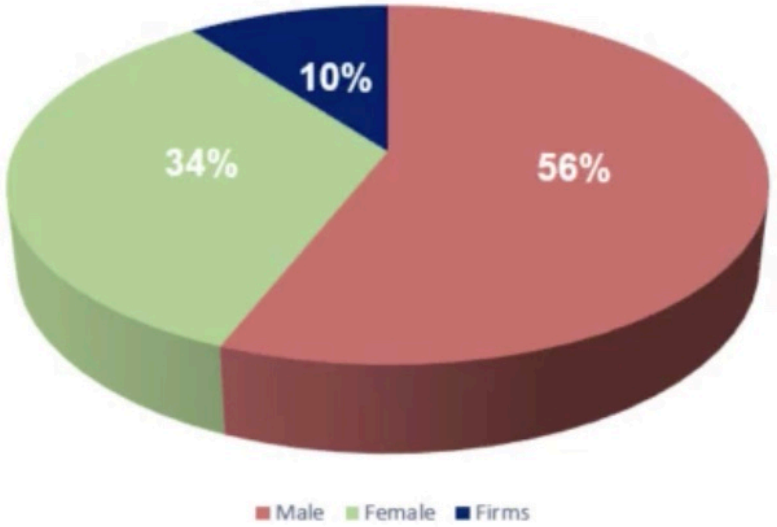
365 DataScience RE California Database

Gender

Currently summarizing

Frequency distribution table

	Frequency	Relative frequency
Male	93	56%
Female	56	34%
Firms	17	10%
Total	166	100%



1. **Males** are more likely to sign contracts and are **potentially** a better audience for our ads (unclear).

1.50

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

A1

Fullscreen

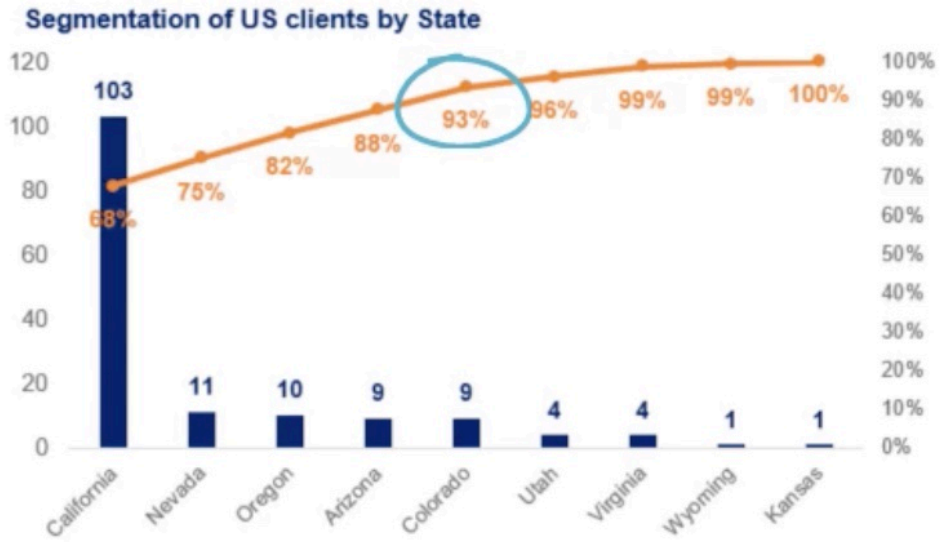
365 DataScience RE California Databa

Location

Currently summarizing

Frequency distribution table

	Frequency	Relative frequency	Cumulative frequency	Cumulative US only
California	103	53%	53%	68%
Nevada	11	6%	58%	75%
Oregon	10	5%	64%	82%
Arizona	9	5%	68%	88%
Colorado	9	5%	73%	93%
Utah	4	2%	75%	96%
Virginia	4	2%	77%	99%
Wyoming	1	1%	77%	99%
Kansas	1	1%	78%	100%
None	43	22%	100%	
Total	195	100%		



1. **Males** are more likely to sign contracts and are **potentially** a better audience for our ads (unclear).

2. 68% of our sales came from **California**, with **Nevada**, **Oregon**, **Arizona** and **Colorado** forming 93% of the US customer base.

1.50

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

A1

Fullscreen

365 DataScience RE California Database

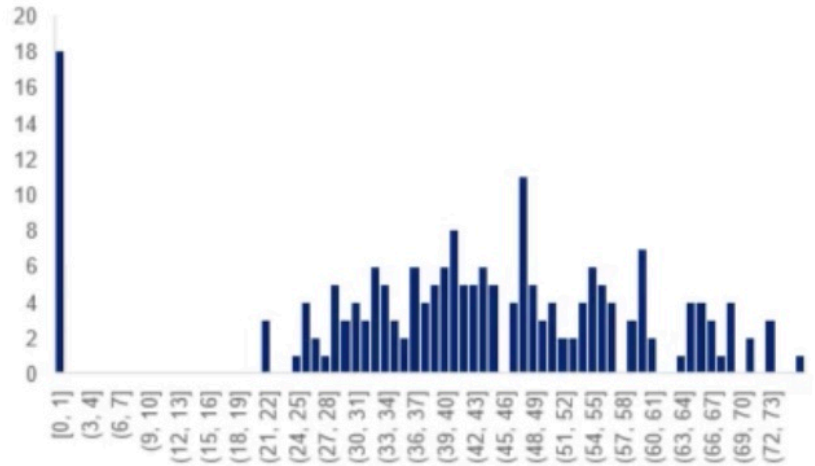
Age

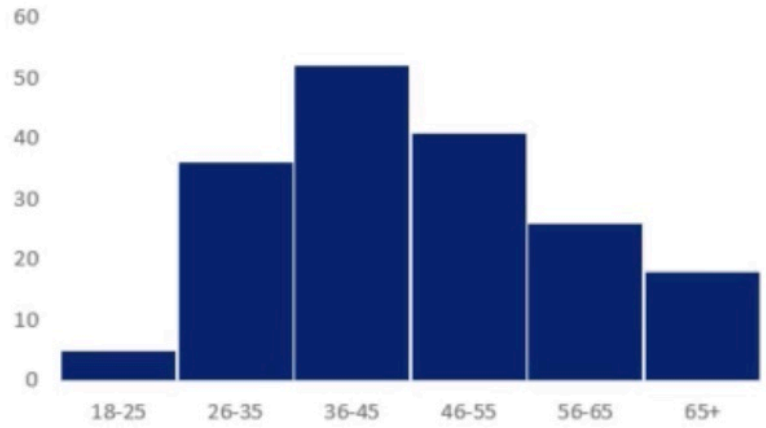
Currently summarizing

Frequency distribution table

	Frequency	Relative frequency
18-25	5	3%
26-35	36	20%
36-45	52	29%
46-55	41	23%
56-65	26	15%
65+	18	10%
Total	178	100%

Mean	46.15
Median	45.00
Mode	48.00
Skew	0.24
Variance	164.91
St. dev.	12.84





1. **Males** are more likely to sign contracts and are **potentially** a better audience for our ads (unclear).

2. 68% of our sales came from **California**, with **Nevada**, **Oregon**, **Arizona** and **Colorado** forming 93% of the US customer base.

3. 71% of the sales were made with customers aged **between 26 and 55 years old**, with a mean of **46 years** and a standard deviation of **13 years**. Younger people buy more property than older people.

1.50

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

A1

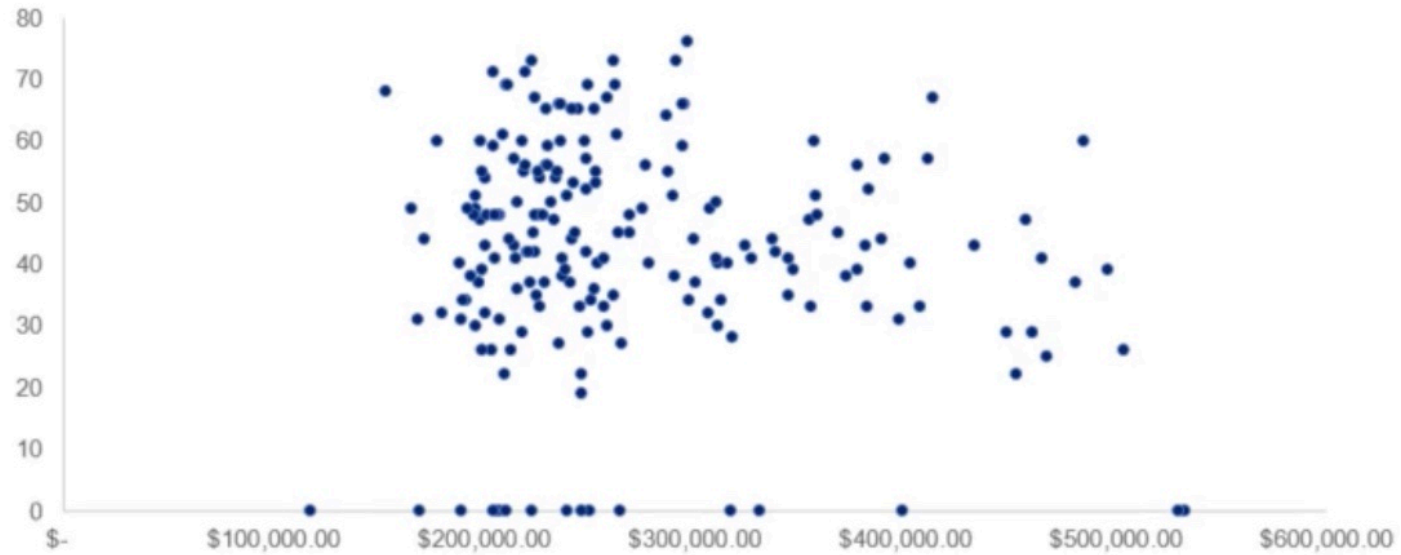
365 DataScience RE California Database

Relationship between age and price

Currently summarizing

Age and price

Covariance	-176361.87
Correlation coefficient	-0.17



1. **Males** are more likely to sign contracts and are **potentially** a better audience for our ads (unclear).

2. 68% of our sales came from **California**, with **Nevada**, **Oregon**, **Arizona** and **Colorado** forming 93% of the US customer base.

3. 71% of the sales were made with customers aged **between 26 and 55 years old**, with a mean of **46 years** and a standard deviation of **13 years**. Younger people buy more property than older people.

4. There is **no relationship** between the age of a given customer and the price they are willing to pay.