COHORT 7
WEEK 4

VISUALIZATION AND DATA EXPLORATION

OUTLINE

- Basics of Visualization
- Data Types and Visualization Type
- Software Plotting Libraries

Two types of visualization

Data exploration visualization:
figuring out what is true

Data presentation visualization:
convincing other people it is true

Visualization vs. Statistics

Visualization almost always presents a more informative (though less quantitative) view of your data than statistics (the noun, not the field)

	Data Set1			Data Set 2			Data Set 3			Data Set 4	
	X	y		X	y		X	y		X	y
	10	8.04		10	9.14		10	7.46		8	6.58
	8	6.95		8	8.14		8	6.77		8	5.76
	13	7.58		13	8.74		13	12.74		8	7.71
	9	8.81		9	8.77		9	7.11		8	8.84
	11	8.33		11	9.26		11	7.81		8	8.47
	14	9.96		14	8.1		14	8.84		8	7.04
	6	7.24		6	6.13		6	6.08		8	5.25
	4	4.26		4	3.1		4	5.39		19	12.5
	12	10.84		12	9.13		12	8.15		8	5.56
	7	4.82		7	7.26		7	6.42		8	7.91
	5	5.68		5	4.74		5	5.73		8	6.89
Average	**9**	**7.50**		**9**	**7.50**		**9**	**7.5**		**9**	**7.50**
Variance	**11**	**4.13**		**11**	**4.13**		**11**	**4.12**		**11**	**4.12**

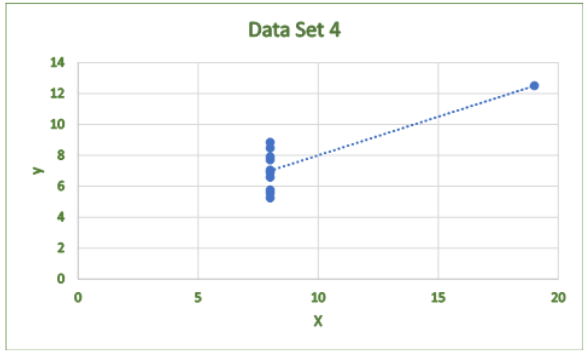
Now if you really look at these values, for each data sets the stats properties are either identical or similar.

- Each X's have average as 9.
- Each y's have average as 7.5.
- Each X's have variance as 11.
- Each y's have variance as 4.12 - 4.13.

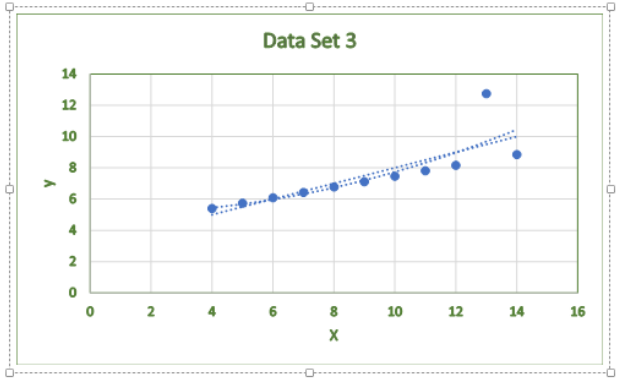
	Data Set1			Data Set 2			Data Set 3			Data Set 4	
	X	y		X	y		X	y		X	y
	10	8.04		10	9.14		10	7.46		8	6.58
	8	6.95		8	8.14		8	6.77		8	5.76
	13	7.58		13	8.74		13	12.74		8	7.71
	9	8.81		9	8.77		9	7.11		8	8.84
	11	8.33		11	9.26		11	7.81		8	8.47
	14	9.96		14	8.1		14	8.84		8	7.04
	6	7.24		6	6.13		6	6.08		8	5.25
	4	4.26		4	3.1		4	5.39		19	12.5
	12	10.84		12	9.13		12	8.15		8	5.56
	7	4.82		7	7.26		7	6.42		8	7.91
	5	5.68		5	4.74		5	5.73		8	6.89
Average	**9**	**7.50**		**9**	**7.50**		**9**	**7.5**		**9**	**7.50**
Variance	**11**	**4.13**		**11**	**4.13**		**11**	**4.12**		**11**	**4.12**

Statically they look identical.
But are these data same? While building model can same model be a good fit for all of these data set?

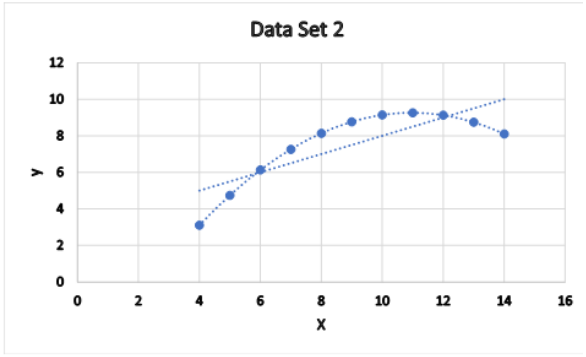
	Data Set 4	
	X	y
	8	6.58
	8	5.76
	8	7.71
	8	8.84
	8	8.47
	8	7.04
	8	5.25
	19	12.5
	8	5.56
	8	7.91
	8	6.89
Average	9	7.50
Variance	11	4.12



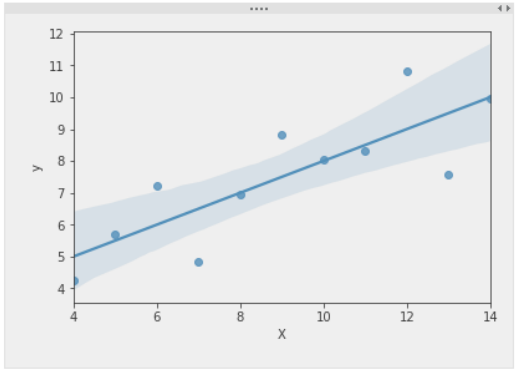
	Data Set 3	
	X	y
	10	7.46
	8	6.77
	13	12.74
	9	7.11
	11	7.81
	14	8.84
	6	6.08
	4	5.39
	12	8.15
	7	6.42
	5	5.73
Average	9	7.5
Variance	11	4.12



	Data Set 2	
	X	y
	10	9.14
	8	8.14
	13	8.74
	9	8.77
	11	9.26
	14	8.1
	6	6.13
	4	3.1
	12	9.13
	7	7.26
	5	4.74
Average	9	7.50
Variance	11	4.13



	Data Set1	
	X	y
	10	8.04
	8	6.95
	13	7.58
	9	8.81
	11	8.33
	14	9.96
	6	7.24
	4	4.26
	12	10.84
	7	4.82
	5	5.68
Average	9	7.501
Variance	11	4.127



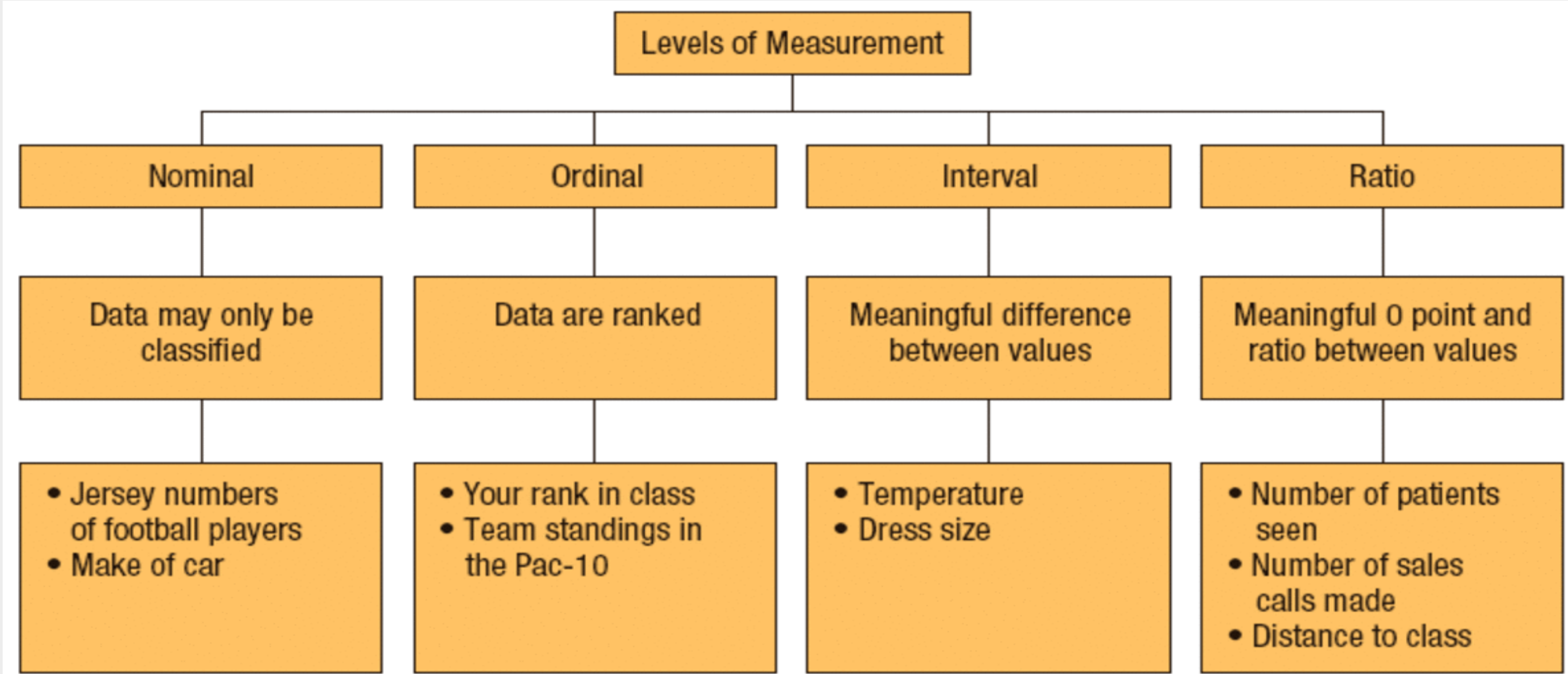
Statistically these are datasets look similar, but these are very different from each other. For first and third dataset may be a linear model can be a good fit but certainly not for the 2nd and 4th which is evident from the trend line.

So essentially statistics might not tell you the whole story . It's always better to understand the relationship by drawing graphs/charts and visualizing them.

(From wiki) This is called Anscombe's quartet . Established by the statistician **Francis Anscombe** to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Data types and visualization types

Data Types (Levels of Measurement)



What are the nominal, ordinal, interval, ratio scales really?

Nominal, ordinal, interval, and ratio scales can be defined as the 4 measurement scales used to capture and analyze data from surveys, questionnaires, data captured from devices, applications and similar research instruments.

Calculation	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Addition and subtraction	No	No	Yes	Yes
Mean, standard deviation	No	No	Yes	Yes
Multiplication and division	No	No	No	Yes
Ratios, coefficient of variation	No	No	No	Yes
Geometric mean	No	No	No	Yes

Calculation	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Addition and subtraction	No	No	Yes	Yes
Mean, standard deviation	No	No	Yes	Yes
Multiplication and division	No	No	No	Yes
Ratios, coefficient of variation	No	No	No	Yes
Geometric mean	No	No	No	Yes

Calculation	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Addition and subtraction	No	No	Yes	Yes
Mean, standard deviation	No	No	Yes	Yes
Multiplication and division	No	No	No	Yes
Ratios, coefficient of variation	No	No	No	Yes
Geometric mean	No	No	No	Yes

NOMINAL SCALES

Nominal scales (also known as a categorical variable scale) refer to variables, categories, or options that don't have a regular order or ranking that has universal application. For example, male and female are both categories but neither one can be ranked as number one or two in every situation.

In research, nominal data can be given a numerical value but those values don't hold true significance. If you use the assigned numerical value to calculate other figures like mean, median, etc. it would be meaningless.

Imagine using a nominal scale and giving male a value of 2, female a value of 4, and transgender a value of 6. If you were to calculate the mean, median, mode, etc. using those values it would have no real implication.

A simple way to think about nominal data is to consider them labels for the information you want to collect. Each label is exclusive, doesn't have any overlap, and lacks numerical significance on their own.

Example of nominal scales

- Samsung
- Apple
- Nokia
- Blackberry
- HTC

Example of nominal scales

What is your gender?

- Male
- Female
- Transgender
- Non-Binary
- Prefer not to say

What is your highest level of education?

- Some high school
- High school
- Trade school
- Associates degree
- Bachelor’s degree
- Master’s degree
- Ph.D. degree

ORDINAL SCALES

The ordinal scale is the second level of data measurement and encompasses the nominal scale. With an ordinal scale, the order of the values (ordinal variable) is important but the difference between values is inconsequential.

That’s because, due to the nature of the options presented on the scale, there’s often no way of knowing the degree of difference between them. Even when the difference between options is quantifiable, it doesn’t yield much insight when compared to the order of the values.

Example of ordinal scales

Income

Less than 25,000

25,000 – 50,000

50,000 – 100,000

Above 100,000

Response to a service

Very good

Good

Average

Poor

Very poor

Example of ordinal scales

Customer Experience

- Very good
- Good
- neutral
- Poor
- Very poor

How often do you go to GYM

- Very often
- Often
- Regularly
- Seldom
- Very seldom

Nominal scale vs ordinal scale

Nominal scales only take into consideration the label of the options while ignoring order. Ordinal scales take the label of the options into consideration as well as the order of those options. Both scales ignore the value of variables.

Because of this, ordinal scales have more applications than a nominal scale. Nominal scales can have as few as two options (dichotomous question) and can also work as a demographic question (what is your gender). Ordinal scales usually have more than two options to establish order.

Interval scale

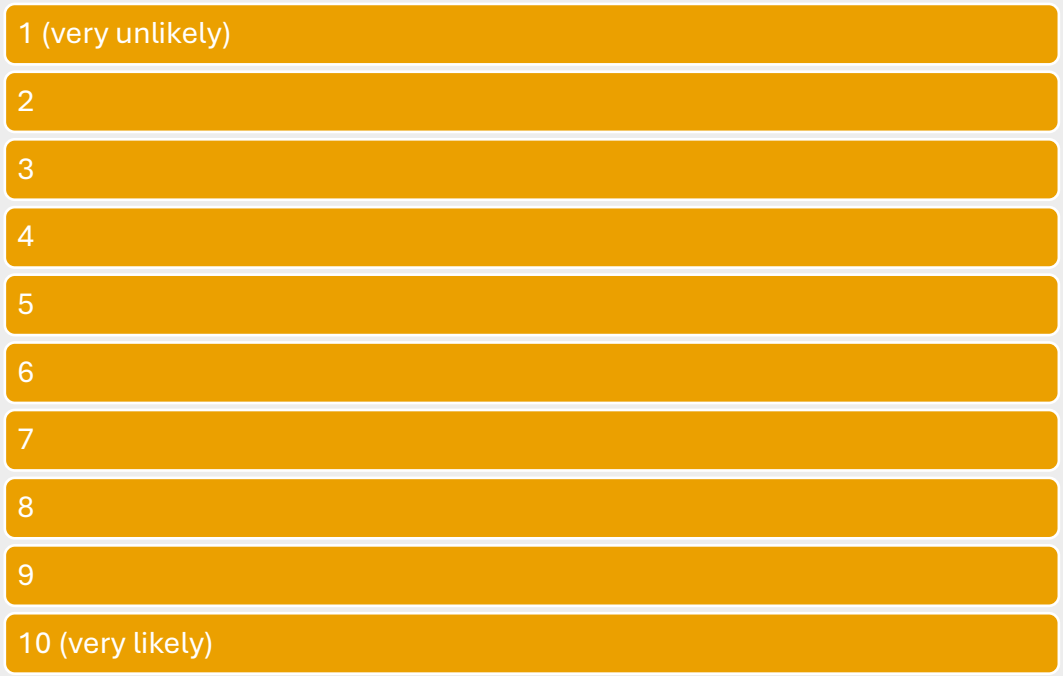
It’s a numerical scale in which the order is known and the difference between the values has meaning.

The interval scale is the third level of measurement and encompasses both nominal and ordinal scales.

This scale can also be referred to as an interval variable scale (**interval variable** is used to describe the meaningful nature of the difference between values).

Example of Interval scales

How likely are you to recommend us to a friend or colleague?



NOTE

The major challenge with interval data is that there's no true zero so deeper statistical analysis is impossible. This is where the ratio scale comes into play.

Ratio scale

Ratio scales are the cream of the crop when it comes to statistical analysis because they have everything you need.

A ratio scale has an order, a set value between units, and absolute zero. It’s an interval scale with a true zero.

Example of Ratio scales

What is your weight in KG

- Less than 70
- 70 – 120
- 121 – 150
- More than 150

What is your age?

- Less than 20
- 20 – 30
- 31 – 40
- 41 – 50
- More than 50

In Conclusion

You can only find mode with nominal scales, you can find median with ordinal scales, interval scales lend themselves to mean, mode, and median. Ratio scales can use all of that plus other methods such as geometric mean and coefficient of variation. Arguably, ratio data is the most versatile.

Interval scale vs ratio scale

The interval variable has order and the difference between the variables have meaning but the ratio between them doesn't have meaning. For example, if you increase the temperature from 10 to 20 degrees Celsius, it's not twice as hot. With a ratio variable scale, the difference between the variables has meaning and the ratio between them does as well. For example, if you increase height from 10 meters to 20 meters, it's twice as tall.

The second difference between the two scales is that the ratio scale has a true zero. That means if something is zero, it doesn't exist. If you weigh zero then your weight doesn't exist. Interval scale may have zero but it's not absolute. For example, the temperature can go into the negatives and zero is just another measurement on the scale.