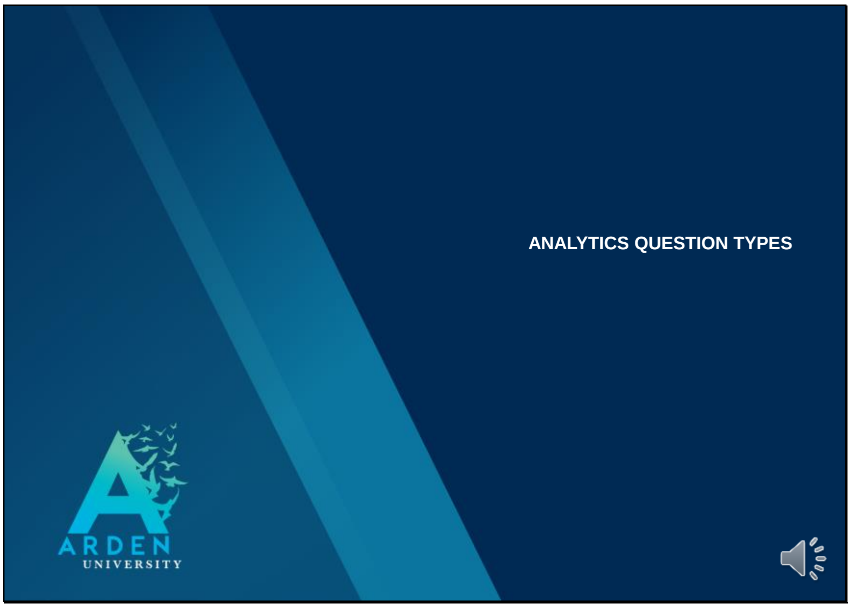
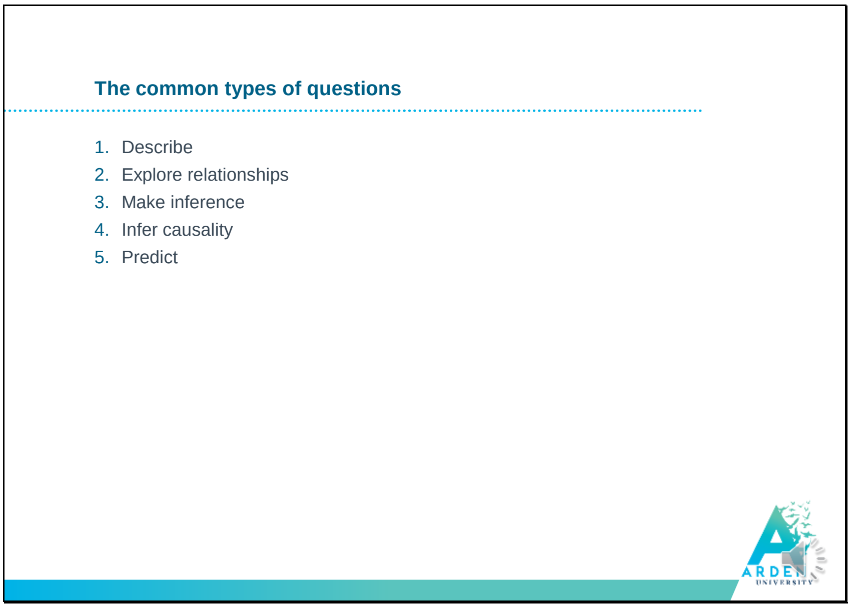
Slide 1



ANALYTICS QUESTION TYPES

Here we will mention the most common types of analytics questions. After defining our business question, we should identify its type.

Slide 2

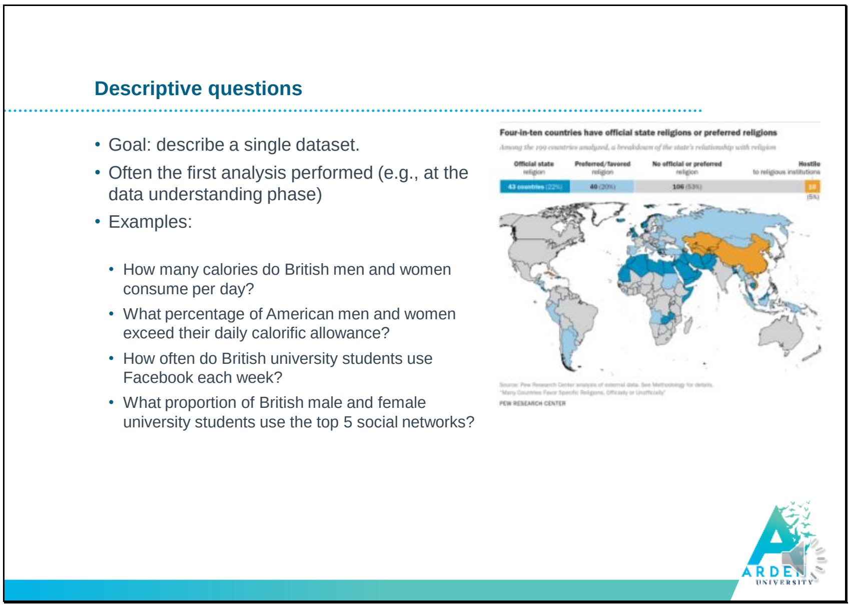


The common types of questions

1. Describe
2. Explore relationships
3. Make inference
4. Infer causality
5. Predict

We will briefly describe five main types of questions: descriptive, relationship-based, inferential, causal and predictive.

Slide 3



Descriptive questions

- Goal: describe a single dataset.
- Often the first analysis performed (e.g., at the data understanding phase)
- Examples:
  - How many calories do British men and women consume per day?
  - What percentage of American men and women exceed their daily calorific allowance?
  - How often do British university students use Facebook each week?
  - What proportion of British male and female university students use the top 5 social networks?

Four-in-ten countries have official state religions or preferred religions

Among the 199 countries analyzed, a breakdown of the state's relationship with religion

Official state religion
Preferred/favored religion
No official or preferred religion
Hostile to religious institutions
43 countries (22%)
40 (20%)
106 (53%)
10
(5%)

Source: Pew Research Center analysis of external data. See Methodology for details.
"Many Countries Favor Specific Religions, Officially or Unofficially"

PEW RESEARCH CENTER

A descriptive data analysis seeks to summarize the measurements in a single dataset.

This is usually the first analysis carried out since it requires almost no knowledge or assumption about the data.

Note that its goal is to merely describe. Thus, descriptive analysis does not include further interpretations.

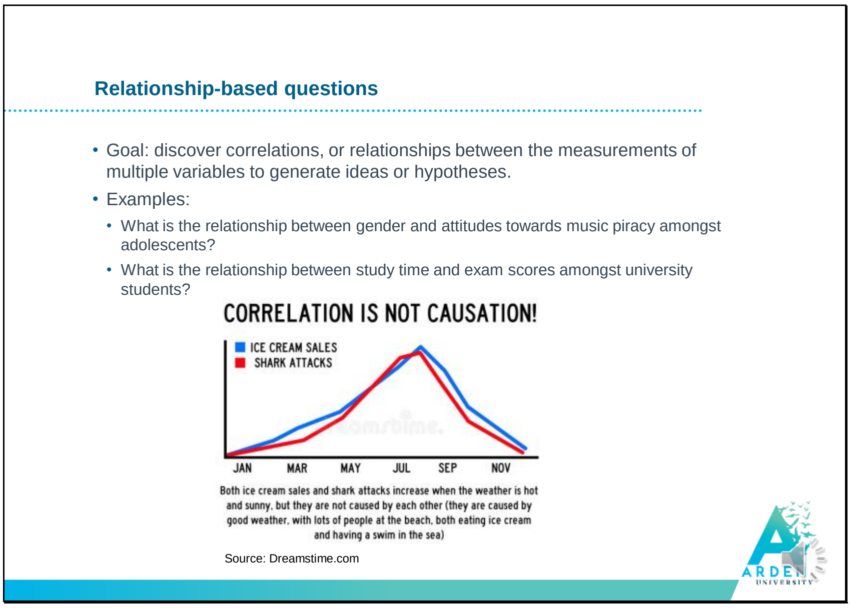
Examples are:

How many calories do British men and women consume per day?
What percentage of American men and women exceeds their daily calorific allowance?

On the right you can see an example of a descriptive analysis carried out by the Pew research centre, which describes which religions are favoured by which countries.

This example might go for further analysis then and compare countries or ask what affects the extent to which a religion is preferred.

Slide 4



Relationship-based questions

- Goal: discover correlations, or relationships between the measurements of multiple variables to generate ideas or hypotheses.
- Examples:
  - What is the relationship between gender and attitudes towards music piracy amongst adolescents?
  - What is the relationship between study time and exam scores amongst university students?

CORRELATION IS NOT CAUSATION!

Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

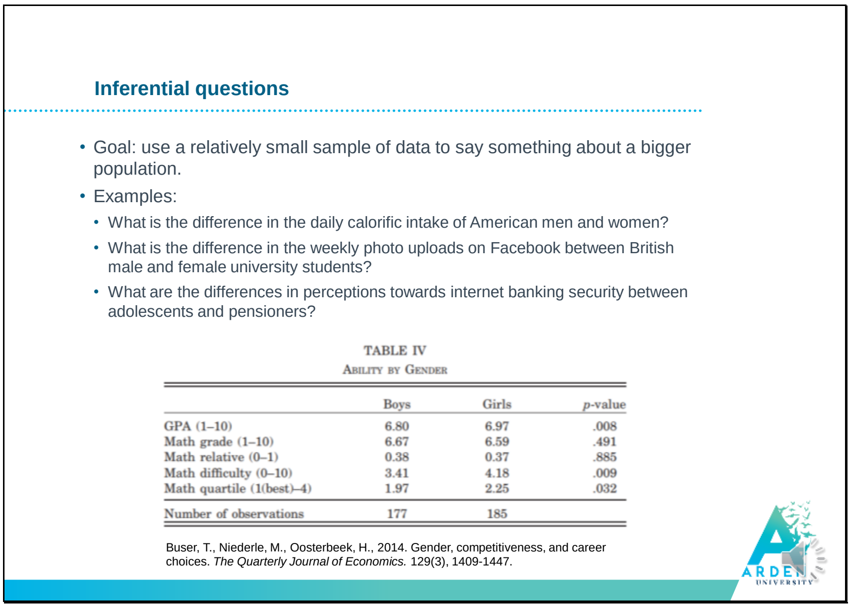
Source: Dreamstime.com

Relationship-based analysis usually builds on descriptive analysis to discover relationships between two variables.

For example: What is the relationship between study time and exam scores amongst university students?

Bear in mind that merely because two variables behave similarly (hence correlate), this does not imply causation and should not be used to predict.
In the example at the bottom of the slide, this correlation between ice cream sales and shark attacks does not imply that an increase in ice cream sales will result (or predict) an increase in shark attacks. It might, but the correlation itself does not imply that. More reasonable is that there is another, third variable (which might be the weather in this case), which affects both variables independently.

Slide 5



Inferential questions

- Goal: use a relatively small sample of data to say something about a bigger population.
- Examples:
  - What is the difference in the daily calorific intake of American men and women?
  - What is the difference in the weekly photo uploads on Facebook between British male and female university students?
  - What are the differences in perceptions towards internet banking security between adolescents and pensioners?

TABLE IV
ABILITY BY GENDER

	Boys	Girls	*p*-value
GPA (1–10)	6.80	6.97	.008
Math grade (1–10)	6.67	6.59	.491
Math relative (0–1)	0.38	0.37	.885
Math difficulty (0–10)	3.41	4.18	.009
Math quartile (1(best)–4)	1.97	2.25	.032
Number of observations	177	185	

Buser, T., Niederle, M., Oosterbeek, H., 2014. Gender, competitiveness, and career choices. *The Quarterly Journal of Economics.* 129(3), 1409-1447.

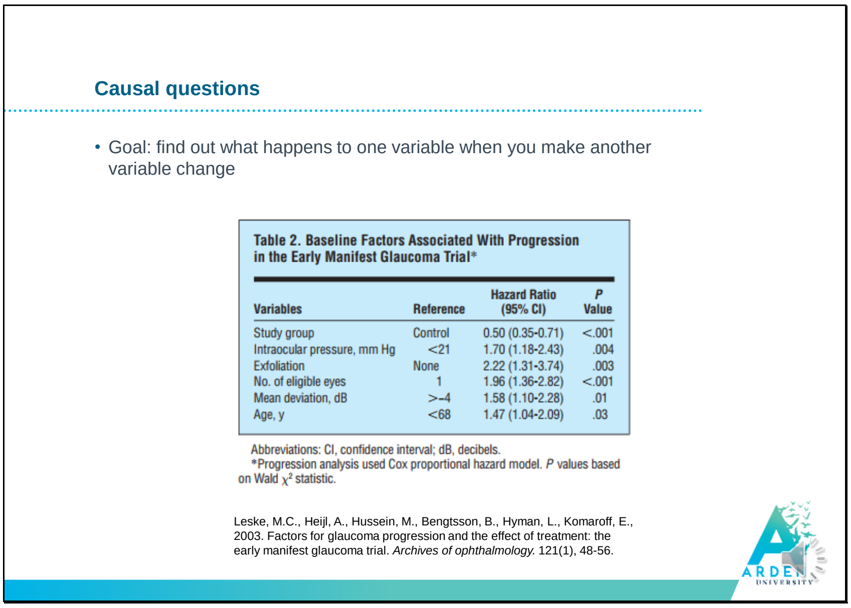
In Lesson 7 we will discuss inferential statistics. In essence its goal is to analyse a small sample and be able to generalize the results to the whole population.

We use this kind of question when we don't have access to the whole population we are studying.

Thus, inferential statistics will help us estimate to what extent we can generalize the results of an analysis of a sample to the whole population.
Pay attention to the size of the sample. If the sample is not large enough, we will need to settle with descriptive analysis, and an inference will not be possible.

The figure below shows differences between boys and girls in grades and estimation of maths difficulty. The right-hand column shows the p-value, which applies to the extent to which these results can be generalized to the whole population, rather than only to the sampled girls and boys.

Slide 6



Causal questions

- Goal: find out what happens to one variable when you make another variable change

Table 2. Baseline Factors Associated With Progression in the Early Manifest Glaucoma Trial*

Variables	Reference	Hazard Ratio (95% CI)	*P* Value
Study group	Control	0.50 (0.35-0.71)	<.001
Intraocular pressure, mm Hg	<21	1.70 (1.18-2.43)	.004
Exfoliation	None	2.22 (1.31-3.74)	.003
No. of eligible eyes	1	1.96 (1.36-2.82)	<.001
Mean deviation, dB	>−4	1.58 (1.10-2.28)	.01
Age, y	<68	1.47 (1.04-2.09)	.03

Abbreviations: CI, confidence interval; dB, decibels.

*Progression analysis used Cox proportional hazard model. *P* values based on Wald χ^2 statistic.

Leske, M.C., Heijl, A., Hussein, M., Bengtsson, B., Hyman, L., Komaroff, E., 2003. Factors for glaucoma progression and the effect of treatment: the early manifest glaucoma trial. *Archives of ophthalmology.* 121(1), 48-56.

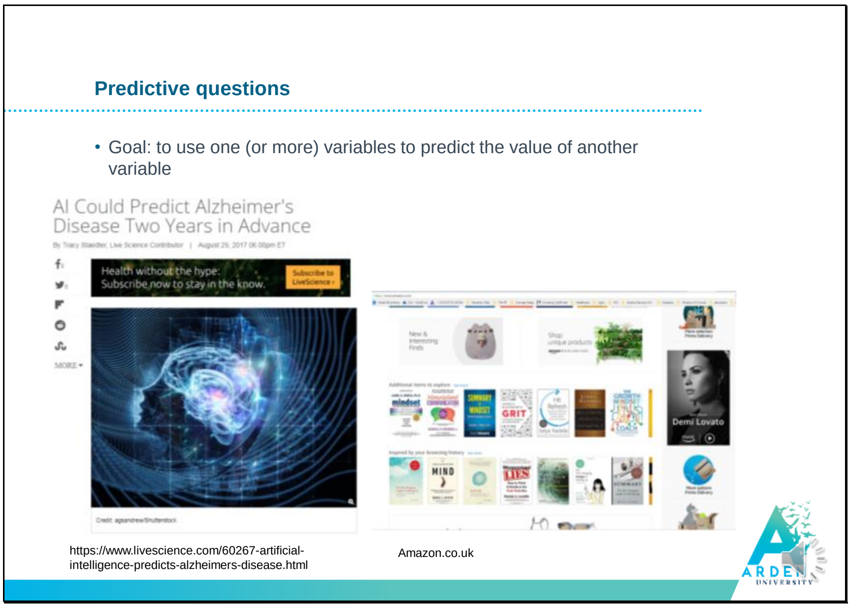
In causal questions (causal – from the word “cause”), it is not enough to show that variables X and Y correlate; there is also a need to establish causality: that is – X causes Y.

Establishing causality will typically involve an experimental design (which we will discuss in Lesson 7).

Usually, randomized clinical trials use causal questions to establish the effect of a treatment on a patient.

In the example below you can see the results of a clinical trial, showing that the study group, which was treated with the tested drug, showed half the risk of progressing in Early Manifest Glaucoma, relatively to the control group.

Slide 7



Predictive questions

- Goal: to use one (or more) variables to predict the value of another variable

AI Could Predict Alzheimer's Disease Two Years in Advance

By Tracy Staedter, Live Science Contributor | August 29, 2017 06:00pm ET

Health without the hype: Subscribe now to stay in the know.

Credit: agsandrew/Shutterstock

https://www.livescience.com/60267-artificial-intelligence-predicts-alzheimers-disease.html

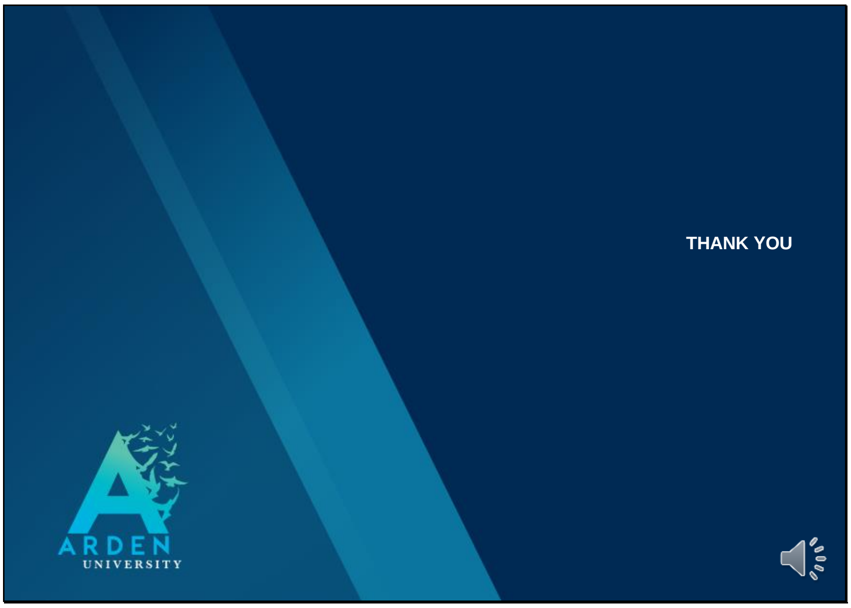
Amazon.co.uk

If X predicts Y, it does not mean that X causes Y.

Big e-commerce companies such as Amazon use predictive analytics to "recommend" products, where each recommendation area uses a slightly different predictive algorithm based on different data.

In the left figure, an artificial-intelligence-driven algorithm can recognize the early signs of dementia in brain scans, and may accurately predict who will develop Alzheimer’s disease up to two years in advance

Slide 8



THANK YOU

ARDEN
UNIVERSITY

Thank you.