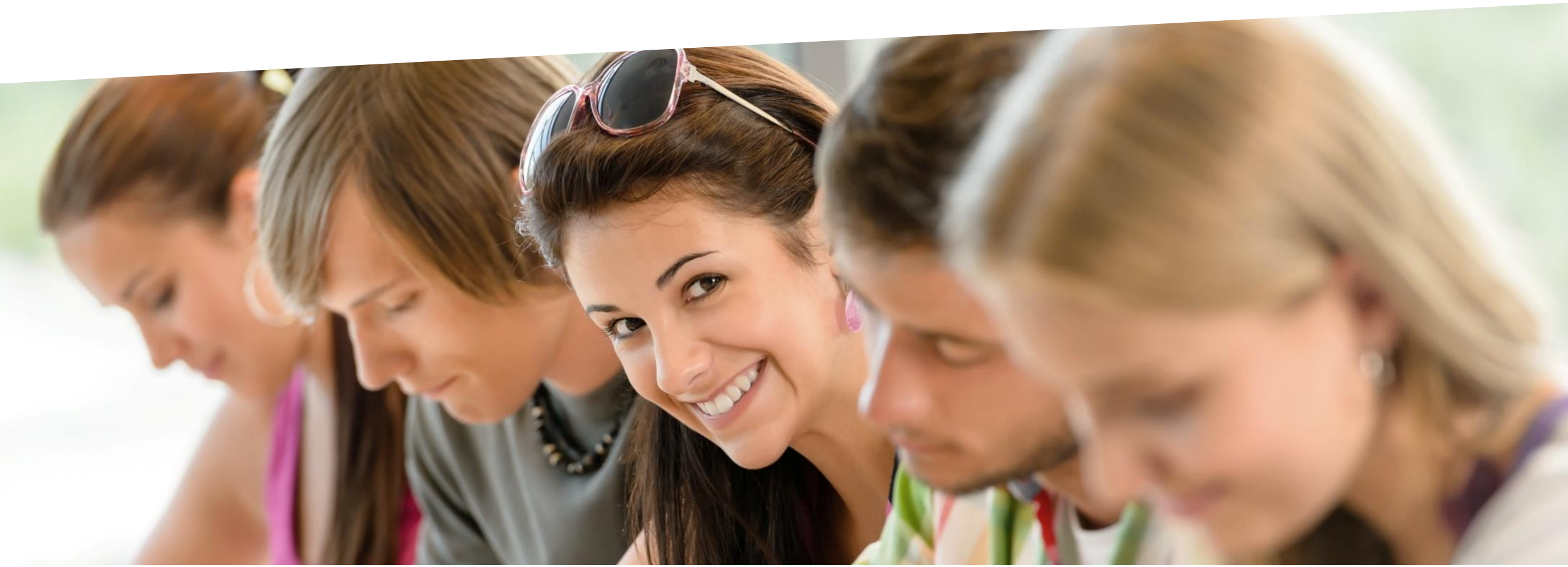
Fachhochschule Graubünden
University of Applied Sciences

Empirical Social Research

Prof. Dr. Thorsten Merkle / Dr. Ramon Leemann



graubünden Education and Research

Agenda

- Learning objectives
- The nature of quantitative data
- Quantitative data analysis
- Q&A

FH GR

Learning objectives

Students are able to …

- ✓ … characterise and distinguish between different types of variables.
- ✓ … name the purposes of and distinguish between exploratory and descriptive data analysis.
- ✓ … create and interpret simple tabulation tables.
- ✓ … apply different ways of presenting data.
- ✓ … apply measures of central tendency and dispersion.

FH GR

The nature of quantitative data

Types of variables

Interval/ratio

– regular distances between all categories in range

Ordinal

– categories can be ranked, but unequal distances between them

Nominal

– qualitatively different categories - cannot be ranked

Dichotomous

– only two categories (e.g. gender)

FH GR

The nature of quantitative data

Types of variables

Nominal variables are organized into non-numeric categories that cannot be ranked or compared quantitatively. This type of data is often referred to as qualitative.

- Appropriate mathematical operation: counting the number of cases per category.
- Categories must be both mutually exclusive and exhaustive.
  - mutually exclusive: each case can only fit into one category
  - exhaustive: there is a category for each possible case.
- Example: favorite breakfast food
  - This is a nominal variable because people can be grouped by their favorite type of food, but we cannot rank one type of food as "more than" another, or "lower than" another.
- Mutual exclusivity: when asked this way, the measurement of favorite breakfast food loses its mutual exclusivity: *Which is your favorite breakfast food?*

FH GR

The nature of quantitative data

Types of variables

Ordinal variables are organized into rankable categories.

- Appropriate mathematical operations: counting and ranking.
- The categories still must be mutually exclusive and exhaustive, but **also** have a logical order that allows them to be ranked.
- Not only can individuals be categorized, as with a nominal variable, they can also be ranked.

FH GR

The nature of quantitative data

Types of variables

3. **Interval** variables have an exact interval between categories, allowing for a direct comparison between categories, such that the difference between any two sequential data points is exactly the same as the difference between any other two sequential data points.
    - Appropriate mathematical operations: counting, ordering, and addition, subtraction, multiplication and division of the interval between values (but not the values themselves).
    - Example: time of the day: 10:00 am, 10:20 am, noon, 4:00 pm, 8:00 pm, etc.
        - In this example, we can say that 10:20 is exactly 20 minutes later than 10:00, but we can't say that 8:00 is "twice as late" as 4:00, and it doesn't make sense to add noon + 4:00.
        - With interval variables, we can perform mathematical operations such as addition and multiplication on the intervals between values (e.g., it takes 20 minutes to drive to the store, 30 minutes to shop, 10 minutes to check out, and 20 minutes to drive home, for a total of 80 minutes), but we can't perform these calculations on the values themselves.

FH GR

The nature of quantitative data

Types of variables

4. **Ratio** variables have all of the characteristics of nominal, ordinal and interval variables, but also have a meaningful zero point.
    - Appropriate mathematical operations: counting, ordering, and addition, subtraction, multiplication and division of the interval between values as well as the values themselves.
    - For most of the course, we will work with ratio rather than interval variables, as they ar much more common, and we'll find many more practical uses for them.
    - Example: weight in pounds (or kilos)
        - When talking about weight, we can:
            - classify people into mutually exclusive and exhaustive groups (as with nominal variables)
            - rank order people (as with ordinal variables)
            - calculate the exact interval between people's weights (as with interval variables)
            - but we can **also** add, subtract and compare values in a meaningful way.

FH GR

Data input

A simple data matrix

	Id	Variable 1	Variable 2	Variable 3	Variable 4
Case 1	1	27	1	2	1
Case 2	2	19	2	1	2
Case 3	3	24	2	3	1

FH GR

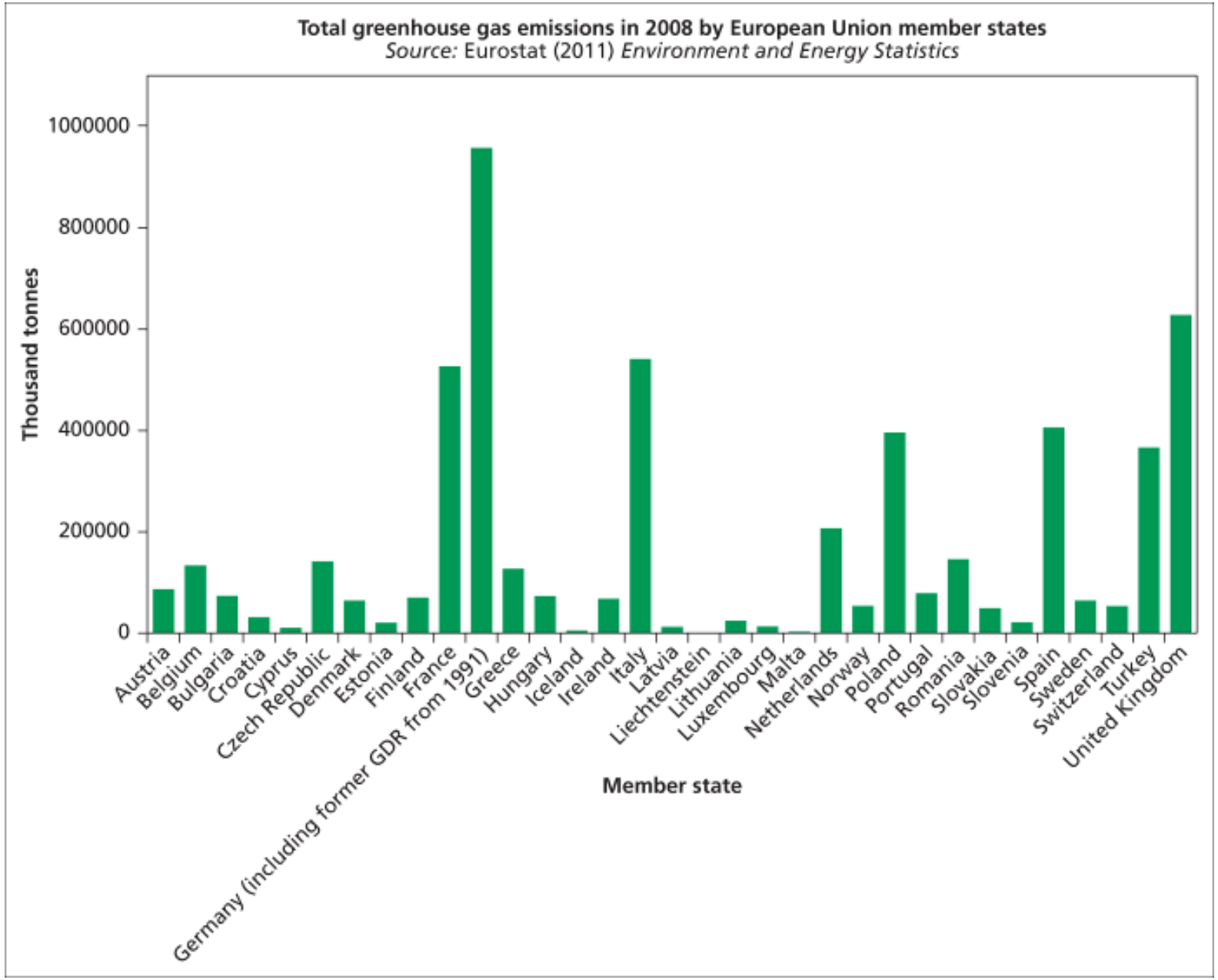
Exploring data

Exploratory data analysis can include:

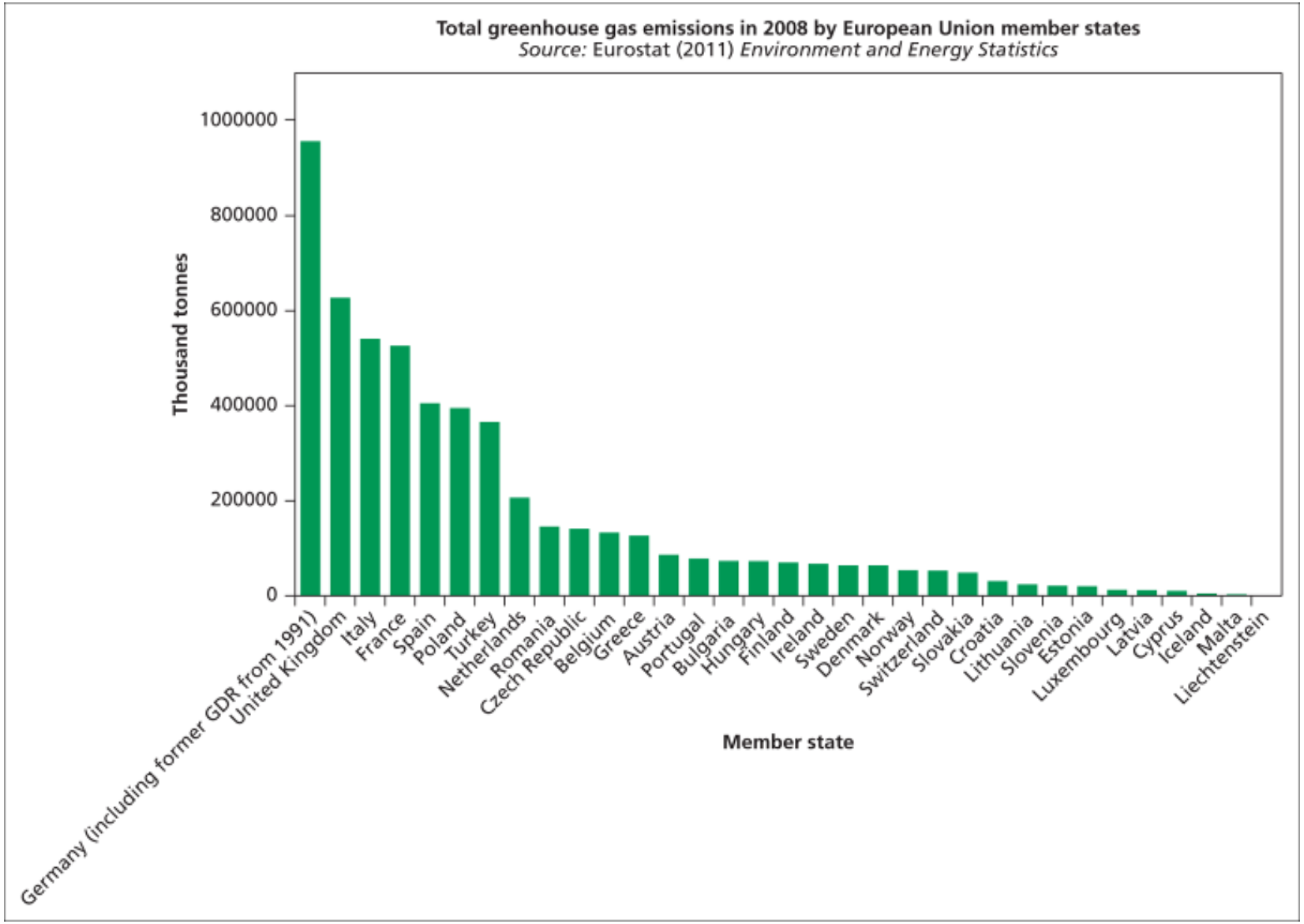
- Specific values
- Highest and lowest values
- Trends over time
- Proportions
- Distributions

FH GR

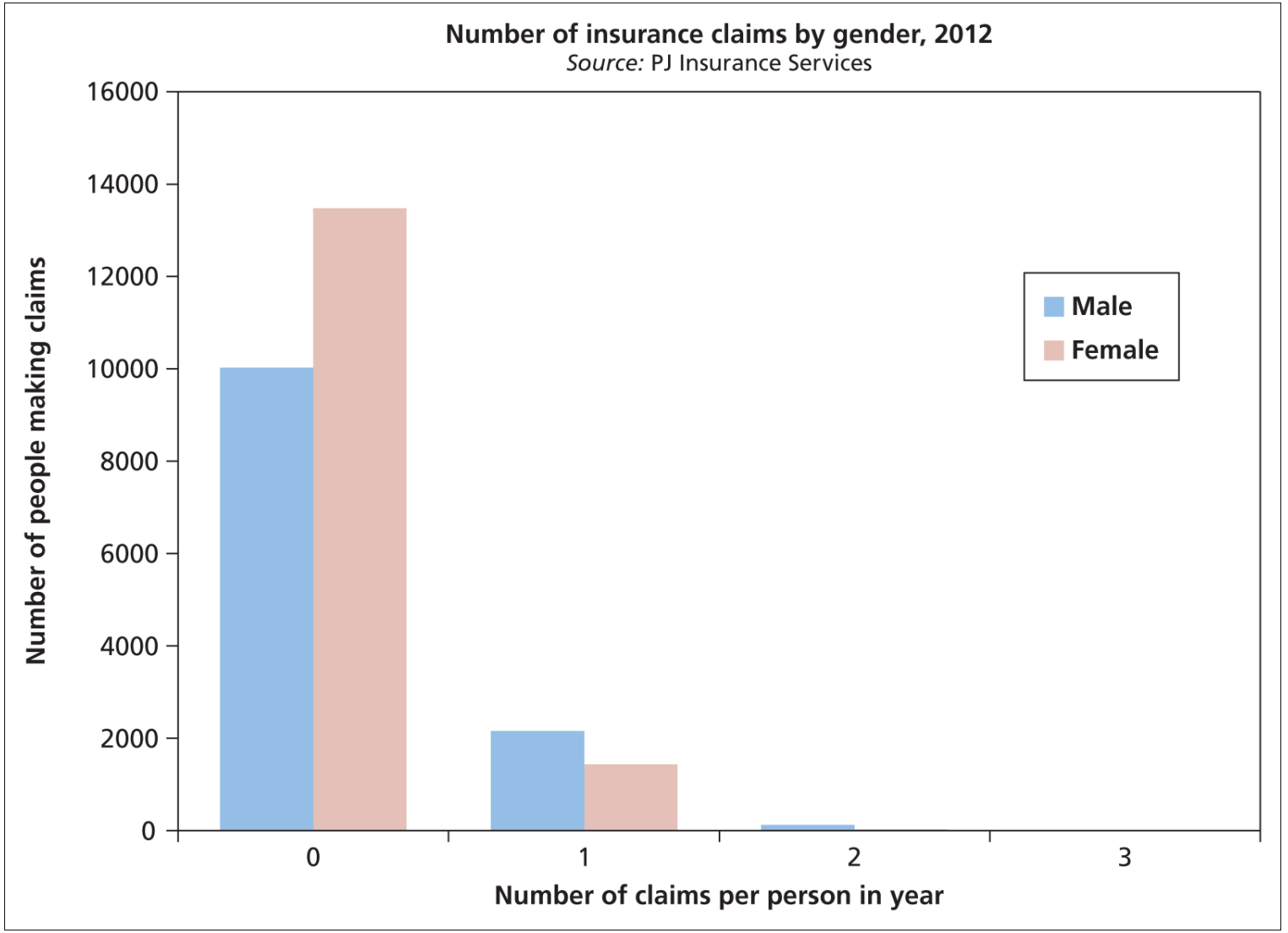
A bar chart


Total greenhouse gas emissions in 2008 by European Union member states
Source: Eurostat (2011) Environment and Energy Statistics
Thousand tonnes
1000000
800000
600000
400000
200000
0
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Republic
Denmark
Estonia
Finland
France
Germany (including former GDR from 1991)
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Netherlands
Norway
Poland
Portugal
Romania
Slovakia
Slovenia
Spain
Sweden
Switzerland
Turkey
United Kingdom
Member state

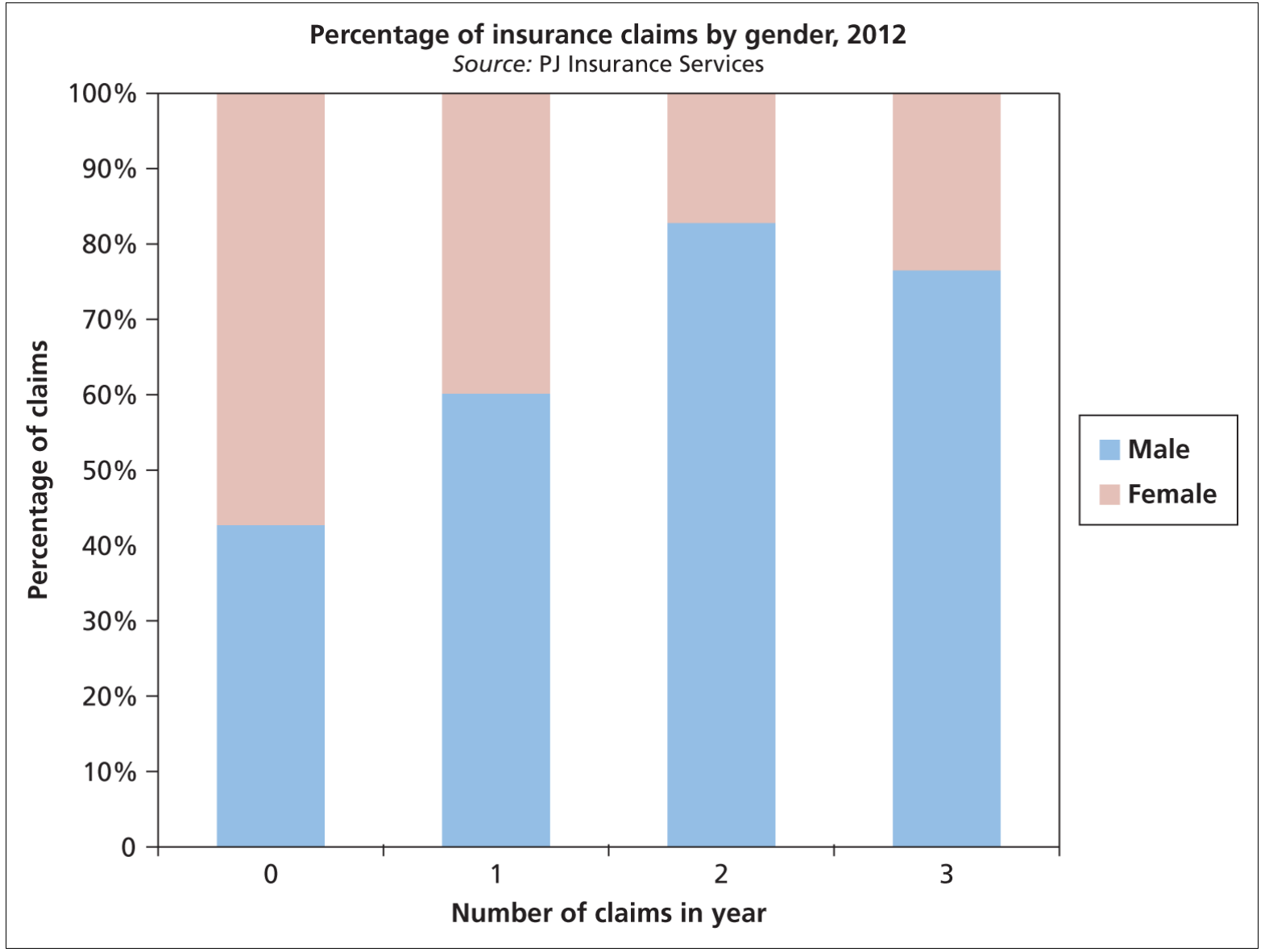
A bar chart (data re-ordered)



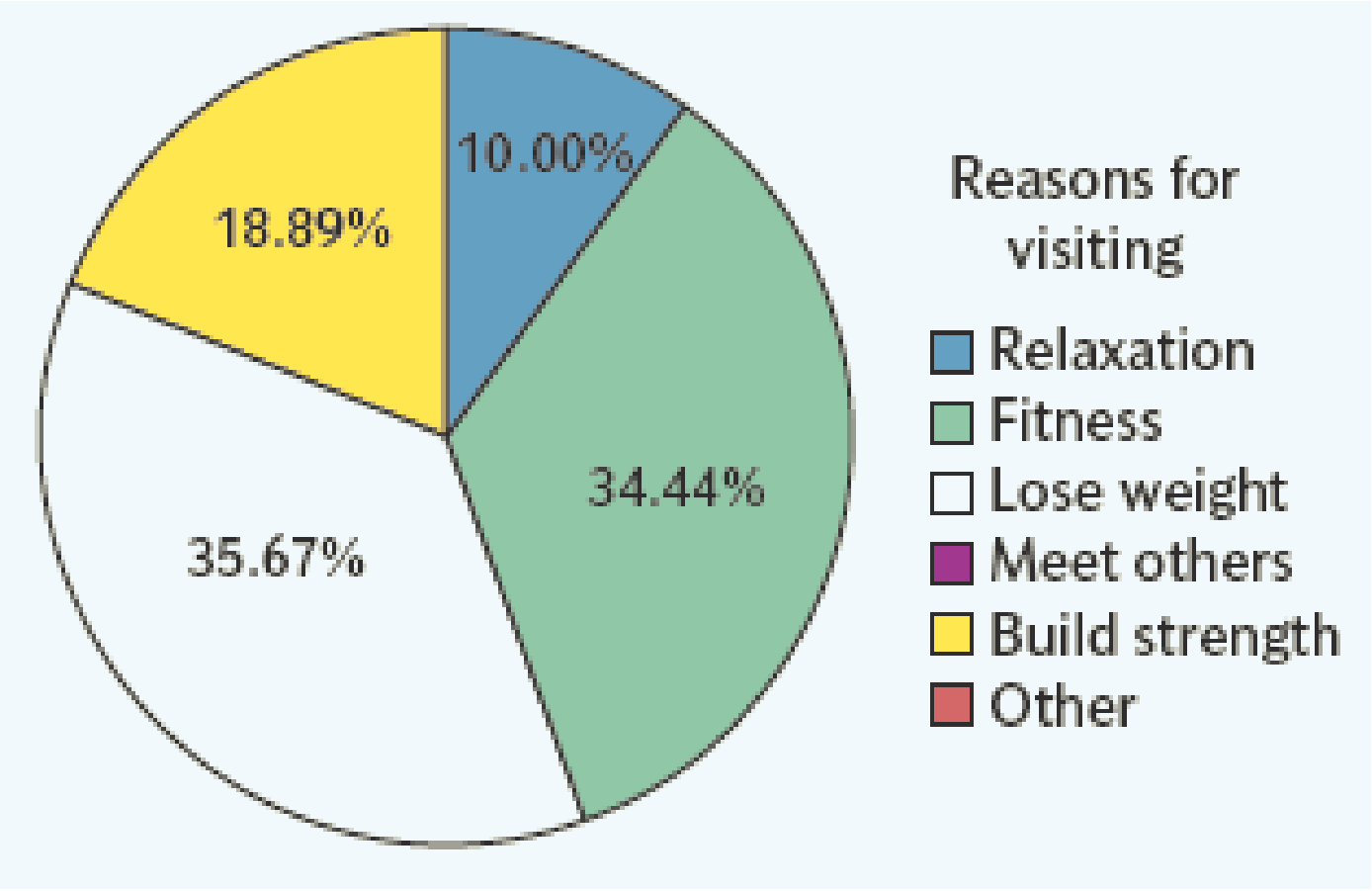
A multiple bar chart



A percentage component bar chart



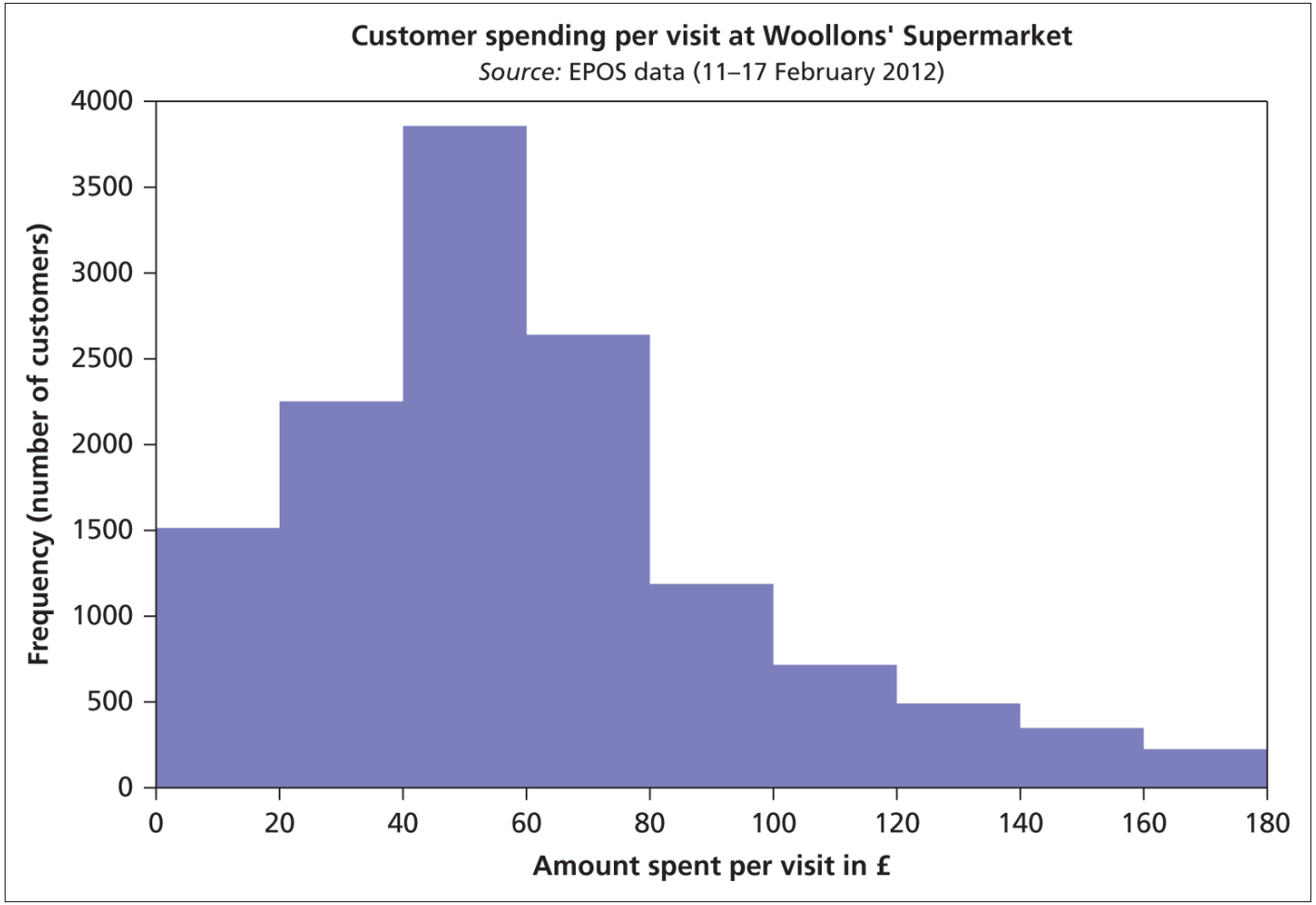
A pie chart


10.00%
18.89%
34.44%
35.67%
Reasons for visiting
Relaxation
Fitness
Lose weight
Meet others
Build strength
Other

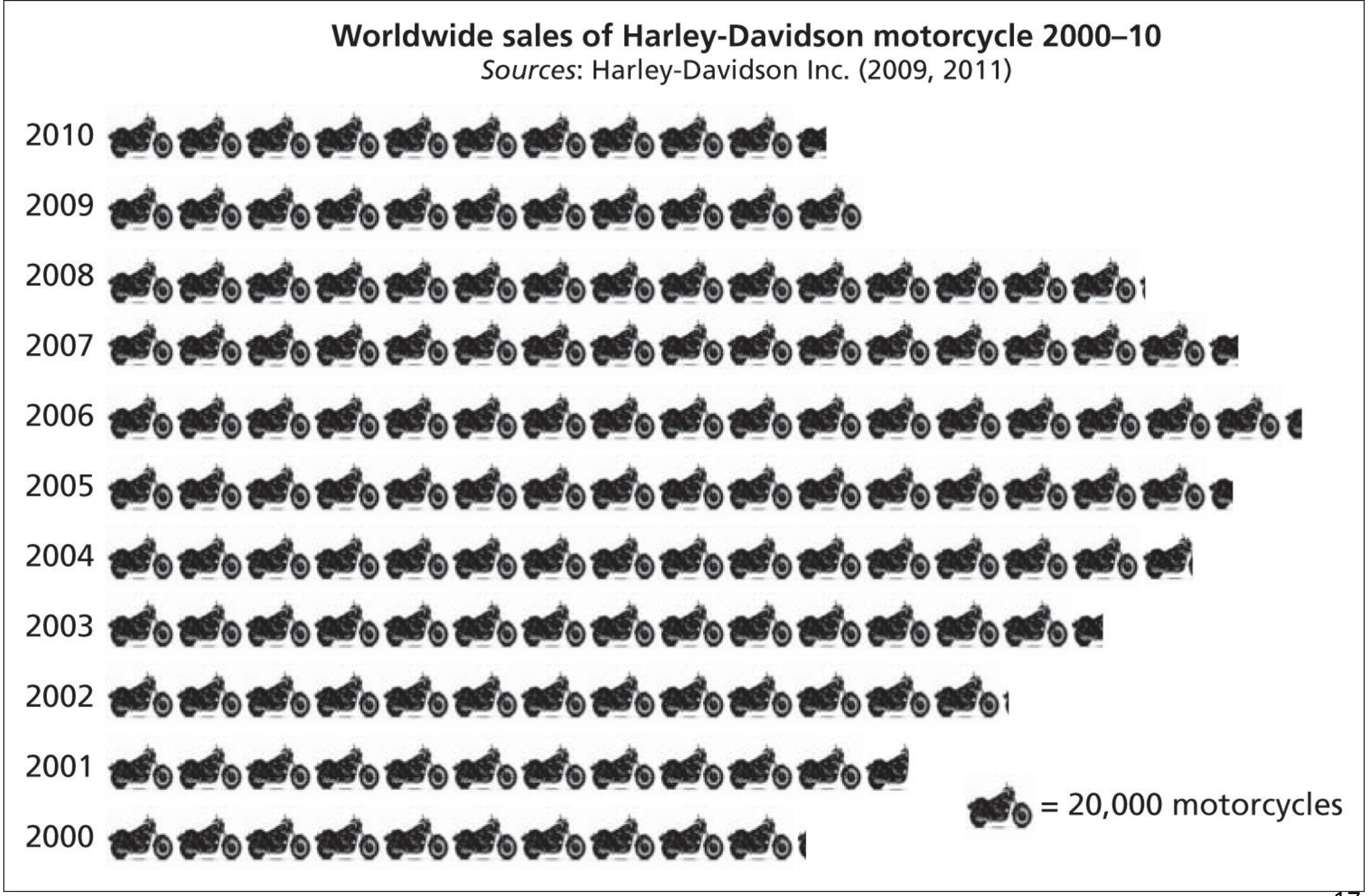
A histogram

Histogram

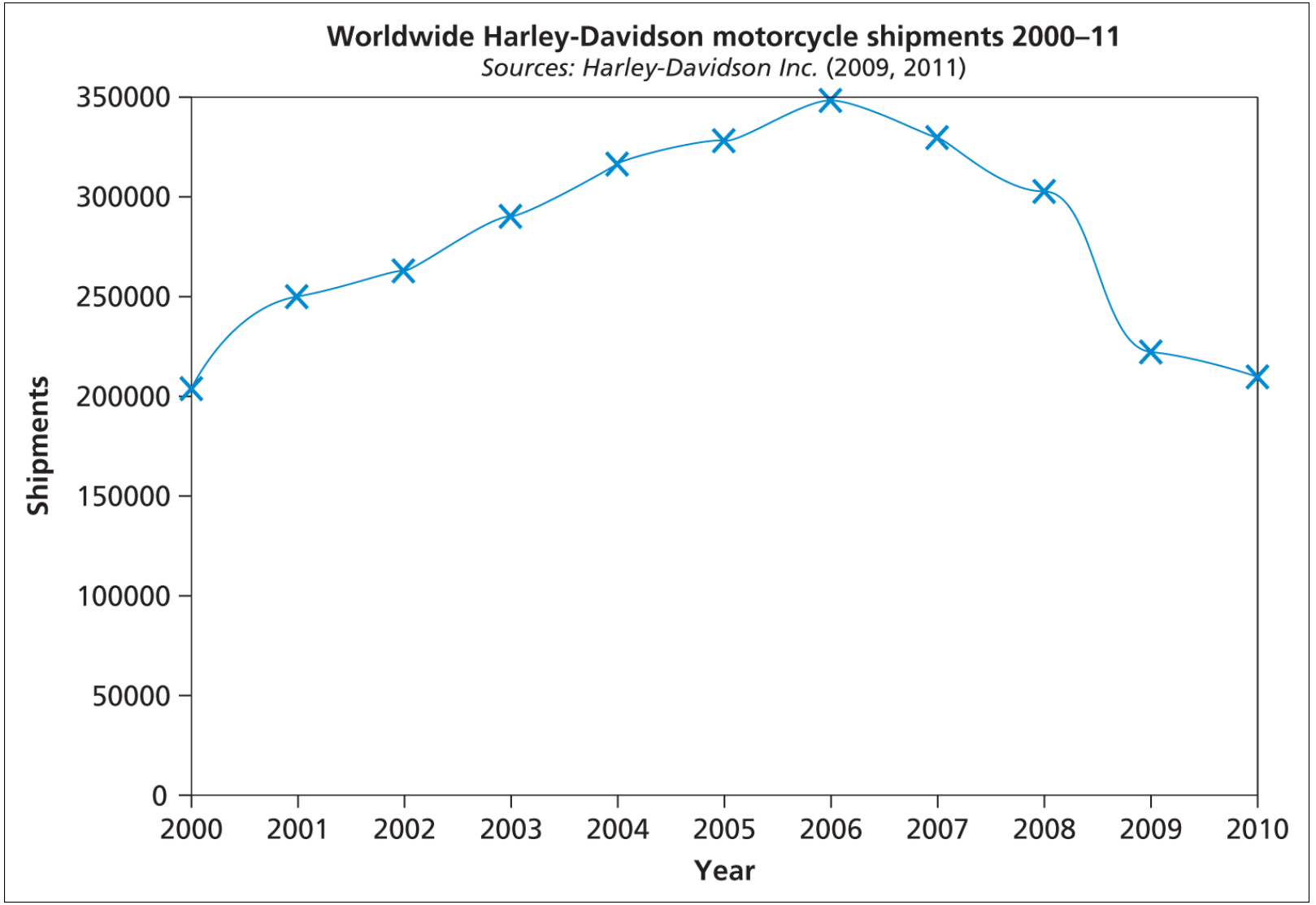
- A graphical way of showing a frequency distribution in which the height of a bar corresponds to the observed frequency of the category.



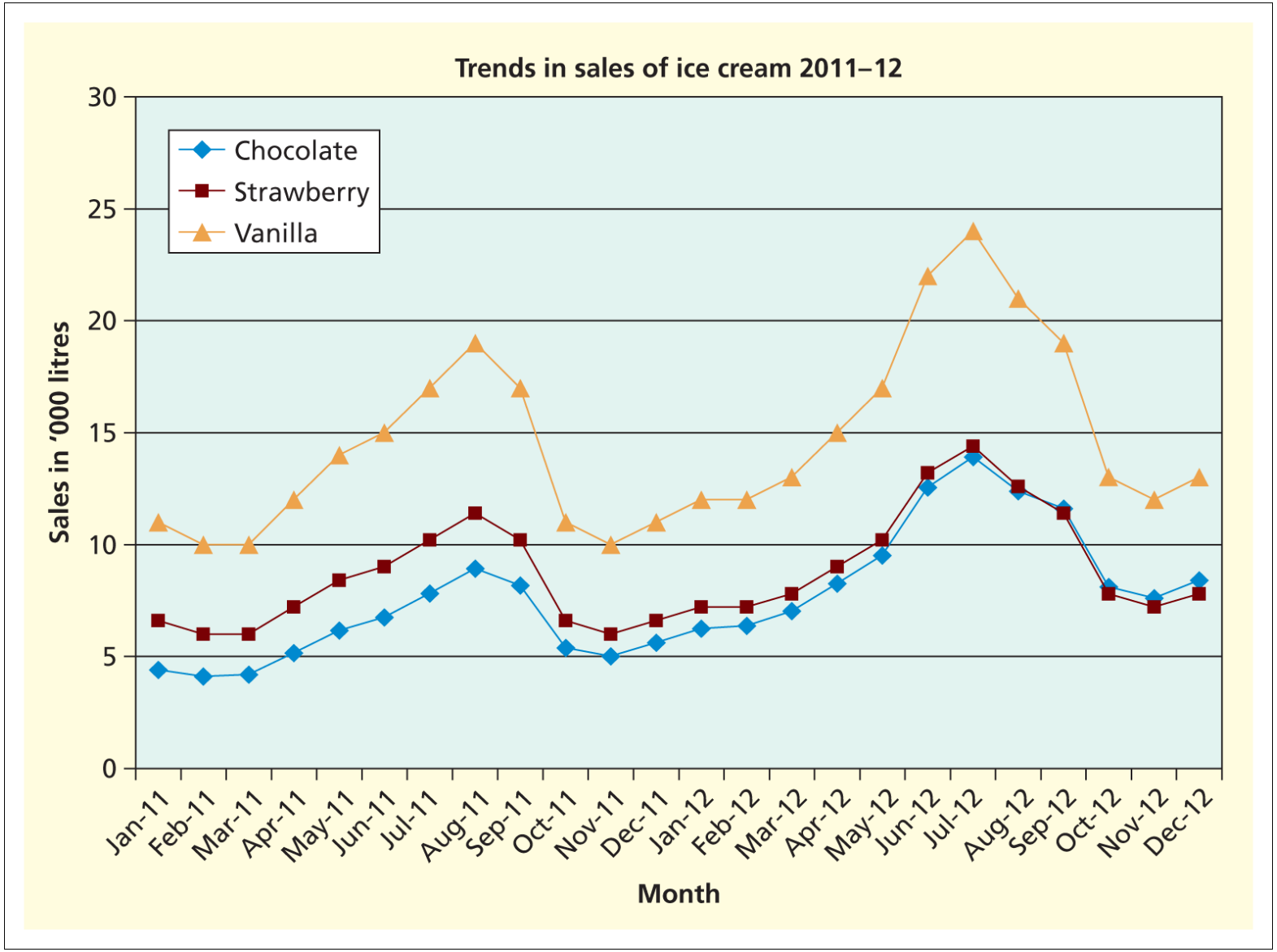
A pictogram


Worldwide sales of Harley-Davidson motorcycle 2000–10
Sources: Harley-Davidson Inc. (2009, 2011)
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
= 20,000 motorcycles

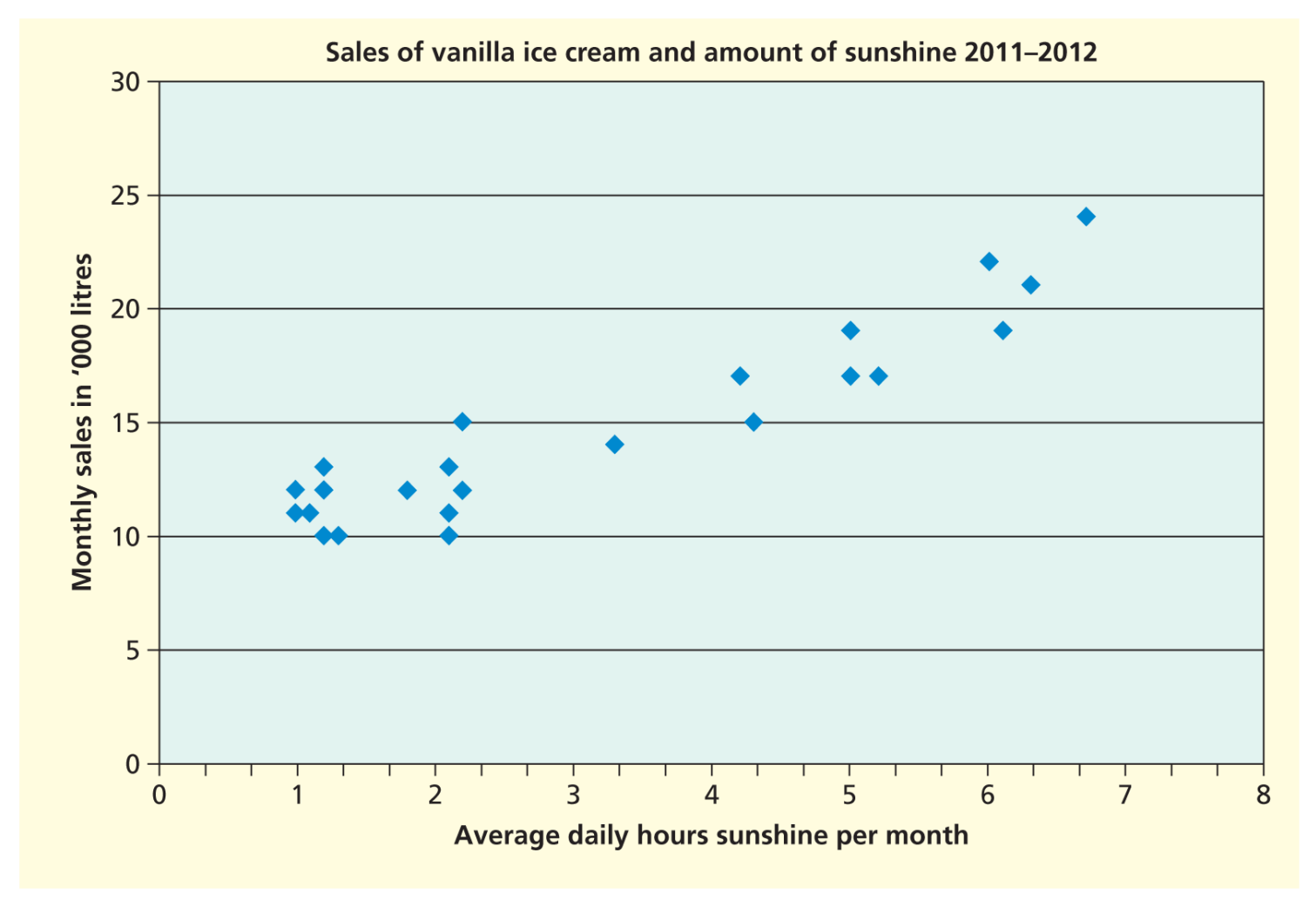
A line graph


Worldwide Harley-Davidson motorcycle shipments 2000–11
Sources: Harley-Davidson Inc. (2009, 2011)
Shipments
350000
300000
250000
200000
150000
100000
50000
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year

Trendlines


Trends in sales of ice cream 2011–12
Chocolate
Strawberry
Vanilla
Sales in '000 litres
30
25
20
15
10
5
0
Jan-11
Feb-11
Mar-11
Apr-11
May-11
Jun-11
Jul-11
Aug-11
Sep-11
Oct-11
Nov-11
Dec-11
Jan-12
Feb-12
Mar-12
Apr-12
May-12
Jun-12
Jul-12
Aug-12
Sep-12
Oct-12
Nov-12
Dec-12
Month

A scatter graph



The nature of descriptive data analysis

Descriptive analysis

- The elementary transformation of raw data in a way that describes the basic characteristics such as central tendency, distribution, and variability.

Describing data using statistics

Statistics to describe a variable focus on two aspects:

- The central tendency
- The dispersion

Measures of central tendency

Mode

- most frequently occurring value

Median

- middle value or mid-point within entire range of values
- not distorted by outliers

Mean

- sum all values in distribution, then divide by total number of values

FH GR

Measures of central tendency

1. **Mode**: the value that occurs most often
   - only measure of central tendency that can be used with nominal variables.
   - also appropriate for ordinal, interval, and ratio variables.
   - Note: some sets of numbers have no mode, and others have multiple modes.
   - Example: favorite breakfast food of 20 of your classmates

FH GR

Measures of central tendency

2. **Median**: the value at the exact center of a distribution, so that 50% of cases are below and 50% are above (when the scores are arranged in rank order).
    - The median is appropriate for ordinal, interval, and ratio variables, but not nominal variables (because nominal data cannot be arranged in rank order).
    - Example: how 15 of your classmates rate the last movie they saw

FH GR

Measures of central tendency

3. **Mean**: mathematical average (sum of all values divided by the number of values).
   - The mean is only appropriate with interval and ratio level data, and is the only measure of central tendency that incorporates all of the scores in a dataset.

$$\bar{x} = \frac{\sum x_i}{N}$$

   - In words: Σ = the sum of all of the scores (of all of the Xs), N = the number of scores

Measures of dispersion

Standard deviation

- Extent to which values differ from the mean.

Inter quartile range

- Difference within the middle 50 percent of all values.

Measures of dispersion

Dispersion means the amount of variation in a sample.

Measures of dispersion compare levels of variation in different samples to see if there is more variability in a variable in one sample than in another.

The range is the difference between the minimum and maximum values in a sample

The standard deviation is the average amount of variation around the mean, reducing the impact of extreme values (outliers)

FH GR

Measures of dispersion

Standard deviation

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

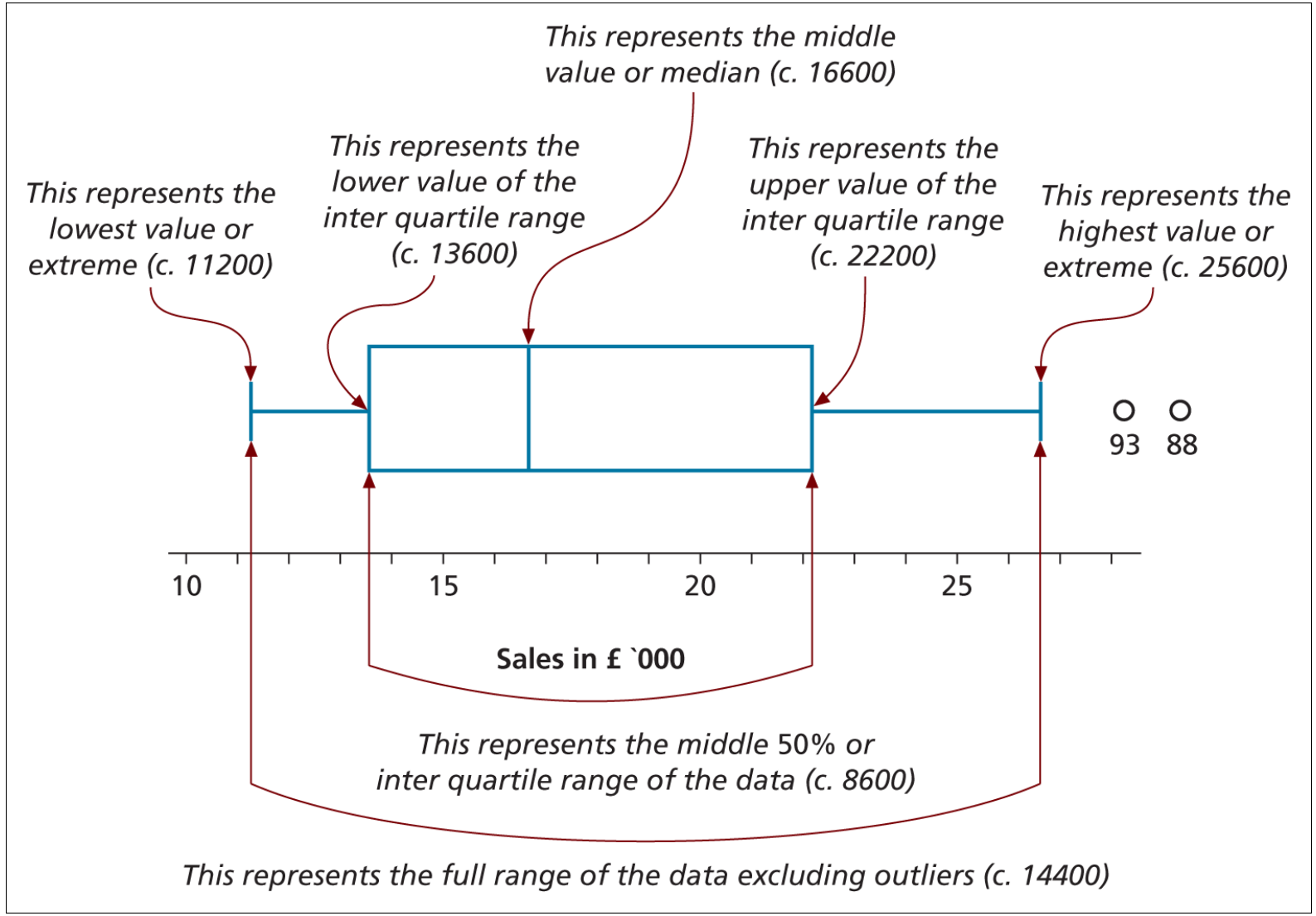
In words: the square root of the average squared deviation. (Remember, deviation is the difference between each score and the mean.)

Six steps for calculating the standard deviation

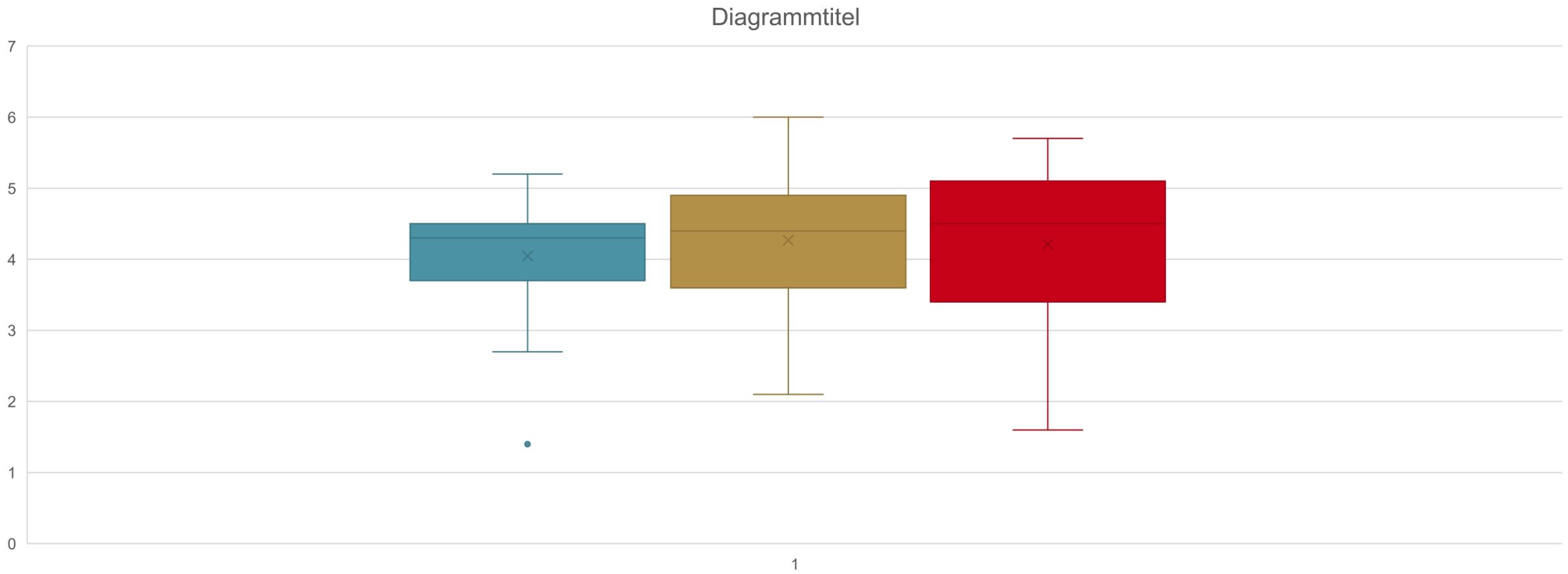
1. find the mean of the dataset
2. calculate the deviation for each score
3. square each deviation
4. sum the squared deviations
5. divide that sum by N
6. take the square root of the result

FH GR

Inter quartile range / box plot



Box plots / (Excel: box and whisker)



Wrap up

Students are able to …

- ✓ … characterise and distinguish between different types of variables.
- ✓ … name the purposes of and distinguish between exploratory and descriptive data analysis.
- ✓ … create and interpret simple tabulation tables.
- ✓ … apply different ways of presenting data.
- ✓ … apply measures of central tendency and dispersion.

FH GR

University of Applied Sciences of the Grisons
Pulvermühlestrasse 57
7000 Chur
T +41 81 286 24 24
info@fhgr.ch

Thank you very much for your attention.

Fachhochschule Graubünden
Scola auta spezialisada dal Grischun
Scuola universitaria professionale dei Grigioni
University of Applied Sciences of the Grisons