Chapter 1

Picturing Distributions with Graphs

Intended Learning Outcomes

- Define **individuals** and **variables** in the context of a dataset
- Distinguish **categorical** vs **quantitative**
- Construct **pie charts** and **bar graphs**
- Interpret pie charts and bar graphs
- Construct **histograms** and **stemplots**
- Interpret histograms and stemplots
  - Shape, centre, spread, outliers
- Construct and interpret **time plots**
  - Trends, seasonality, deviations

Datasets and observations

Dataset

A **dataset** is a structured collection of data containing information about a group of individuals and their variables.

	id	name	score
1	1	Alice	87
2	2	Bob	92
3	⋮	⋮	⋮

Datasets and observations

Dataset

A **dataset** is a structured collection of data containing information about a group of individuals and their variables.

	id	name	score
1	1	Alice	87
2	2	Bob	92
3	⋮	⋮	⋮

Observation

An **observation** is a single row in a dataset, containing all the variable values for one individual.

Individuals and Variables

Individuals

The objects described by a set of data.

Variable

Any characteristic of an individual that can take different values.

Example (`ex:ch1-icecream`): Canadian Ice Cream Survey

Consider the following dataset from a survey of Canadian adults about ice cream preferences:

	gender	**flavour**	**cones/mo**	**age**	**lactose intol.**
1	Female	Chocolate	3	22	No
2	Male	Vanilla	5	28	Yes
3	Female	Strawberry	2	19	No
4	Male	Mint	4	35	No
5	Female	Vanilla	6	41	Yes

Example (`ex:ch1-icecream`): Canadian Ice Cream Survey

Consider the following dataset from a survey of Canadian adults about ice cream preferences:

	gender	**flavour**	**cones/mo**	**age**	**lactose intol.**
1	Female	Chocolate	3	22	No
2	Male	Vanilla	5	28	Yes
3	Female	Strawberry	2	19	No
4	Male	Mint	4	35	No
5	Female	Vanilla	6	41	Yes

Individuals:

The Canadian adults surveyed

Example (`ex:ch1-icecream`): Canadian Ice Cream Survey

Consider the following dataset from a survey of Canadian adults about ice cream preferences:

	gender	flavour	cones/mo	age	lactose intol.
1	Female	Chocolate	3	22	No
2	Male	Vanilla	5	28	Yes
3	Female	Strawberry	2	19	No
4	Male	Mint	4	35	No
5	Female	Vanilla	6	41	Yes

Individuals:

The Canadian adults surveyed

Variables:

gender, flavour, cones/mo

age, lactose intol.

Types of Variables

Categorical

Places individuals into **groups** or **categories**.

Examples:

Gender, Favourite flavour of ice cream, marital status

Types of Variables

Categorical

Places individuals into **groups** or **categories**.

Examples:

Gender, Favourite flavour of ice cream, marital status

Quantitative

Takes **numerical values** where arithmetic makes sense.

Examples:

Age, Height, Number of courses

Example (`ex:ch1-dating`): Dating Survey Example

The following dataset was obtained through an online survey of Americans about how they met their partners:

meeting_method	age	gender	relationship_length	zipcode
Dating app	33	F	20	10001
Dating app	18	M	5	90210
Friends	37	M	21	60601
School/Work	26	F	5	77001
Social Media	20	F	4	98101
Other	43	F	11	85001

The individuals are American adults who responded to the survey

The variables are

- meeting_method (categorical)
- age (quantitative)
- gender (categorical)
- relationship_length (quantitative)
- zipcode (categorical)

Numbers $\neq$ Quantitative

⚠ **Caution:**

Just because a variable contains numbers does not make it quantitative

The following are categorical (in general)

- Postal codes
- Phone numbers
- Jersey numbers
- Student ID numbers

Numbers $\neq$ Quantitative

⚠ **Caution:**

Just because a variable contains numbers does not make it quantitative

The following are categorical (in general)

- Postal codes
- Phone numbers
- Jersey numbers
- Student ID numbers

Question to ask:

Do arithmetic operations make sense?

Exploratory Data Analysis

Exploratory Data Analysis (EDA)

Using graphs and numerical summaries to describe variables and relationships.

Exploratory Data Analysis

Exploratory Data Analysis (EDA)

Using graphs and numerical summaries to describe variables and relationships.

Distribution

The values a variable can take and how often it takes them.

Example (`ex:ch1-pizza-dist`) Distribution

A sample of 15 students was asked about their favourite pizza topping. The *distribution* of responses is shown below.

	favourite_topping
0	Veggie
1	Pepperoni
2	Mushroom
3	Veggie
⋮	⋮

Favourite Pizza Toppings

A sample of 15 students was asked about their favourite pizza topping. The distribution can be summarised in the following table:

Topping	Count	Percent
Pepperoni	6	40%
Mushroom	4	27%
Veggie	5	33%

Minutes to Campus: Raw Data

A survey of 58 DS 1000 students

Commute Time (minutes)

5.5	42.5	28.3	36.4	9.1	13.2	7.2	39.9	4.8	4.6	11.0
37.6	31.9	4.2	33.1	14.1	4.3	13.2	43.2	13.4	40.9	7.3
2.9	3.5	92.5	3.6	5.4	16.1	5.8	11.3	33.4	2.4	32.5
23.2	31.3	15.3	4.6	34.1	6.0	2.1	34.8	32.6	5.7	37.7
11.5	4.7	15.6	32.6	17.5	15.6	10.7	15.2	43.5	15.0	11.8
4.3	6.2	4.3								

Minutes to Campus: Raw Data

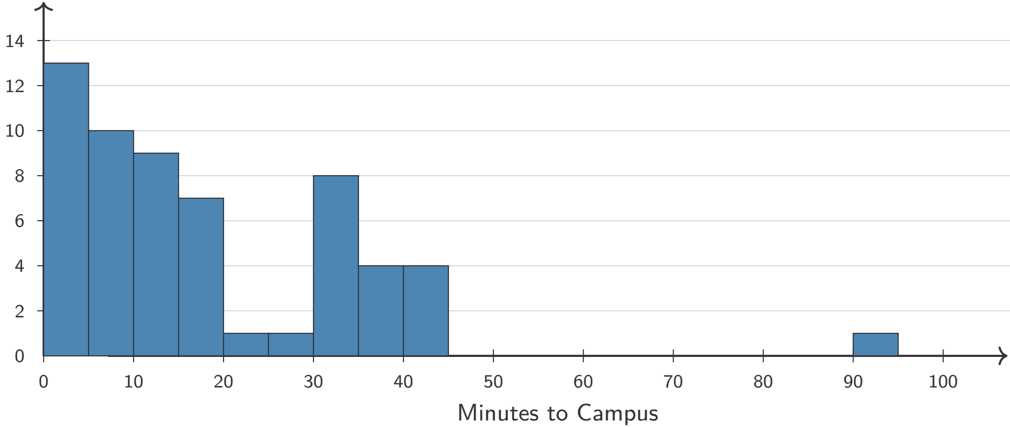
A survey of 58 DS 1000 students

Commute Time (minutes)										
5.5	42.5	28.3	36.4	9.1	13.2	7.2	39.9	4.8	4.6	11.0
37.6	31.9	4.2	33.1	14.1	4.3	13.2	43.2	13.4	40.9	7.3
2.9	3.5	92.5	3.6	5.4	16.1	5.8	11.3	33.4	2.4	32.5
23.2	31.3	15.3	4.6	34.1	6.0	2.1	34.8	32.6	5.7	37.7
11.5	4.7	15.6	32.6	17.5	15.6	10.7	15.2	43.5	15.0	11.8
4.3	6.2	4.3								

What do you see?

Minutes to Campus: Visualized

The same 59 students, now as a histogram



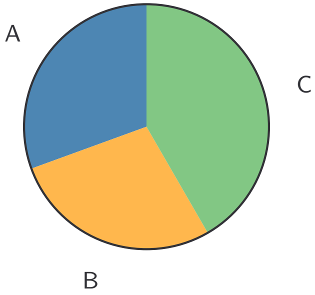
"A picture is worth a thousand numbers." - Aristotle, probably

Pie Charts

Pie Chart

The **pie chart** of a categorical variable is a circle which is

- divided into slices,
- where each slice's area (or angle) represents a category of the variable.

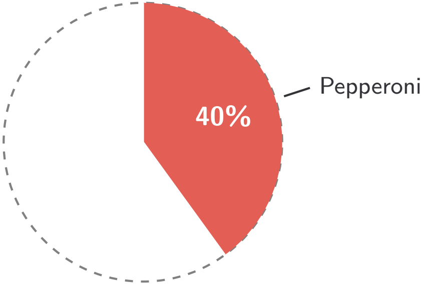


Example (`ex:ch1-pizza-pie`)

Topping	Count	Percent
Pepperoni	6	40%
Mushroom	4	27%
Veggie	5	33%

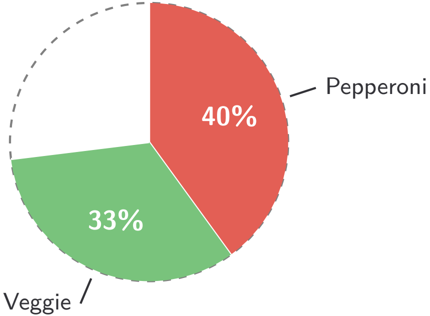
Example (`ex:ch1-pizza-pie`)

Topping	Count	Percent
Pepperoni	6	40%
Mushroom	4	27%
Veggie	5	33%



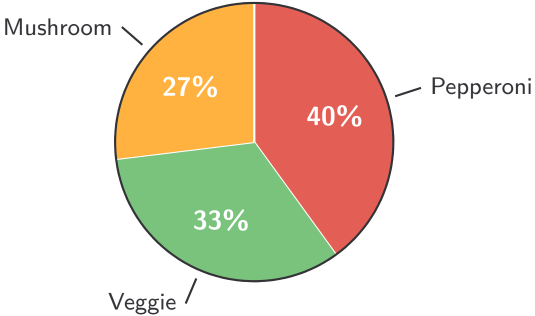
Example (`ex:ch1-pizza-pie`)

Topping	Count	Percent
Pepperoni	6	40%
Mushroom	4	27%
Veggie	5	33%



Example (`ex:ch1-pizza-pie`)

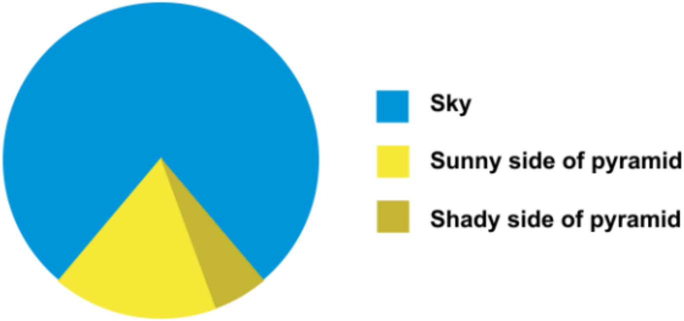
Topping	Count	Percent
Pepperoni	6	40%
Mushroom	4	27%
Veggie	5	33%



Main use of pie charts: Humour

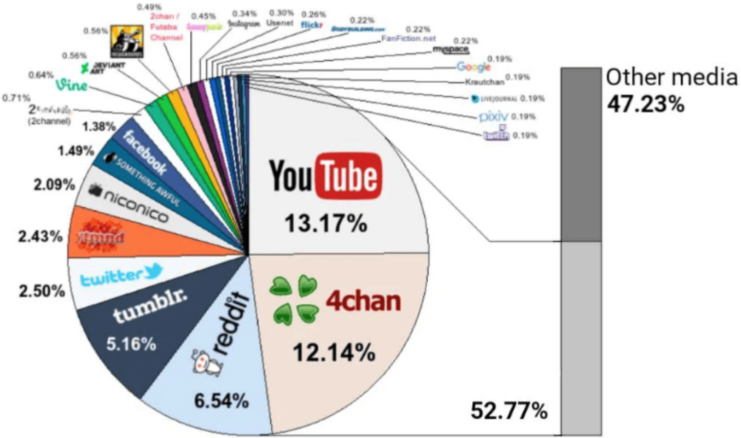
Hot take

the main benefit of pie charts is humour.



Pie I have eaten
Pie I have not yet eaten

Bad pie charts



Meme origins in 2010s

Example (`ex:ch1-transportation-modes`)

A survey of adults asked: “Which of the following modes of transportation do you use regularly?” The results are summarized below.

Mode	Percent of adults using (%)
Car	72
Public Transit	38
Bicycle	19
Walking	44
Rideshare	15
Other	6

Can we construct a pie chart for this distribution?

When are pie charts valid as visualizations?

⚠ **Caution:** For a pie chart to be a valid form of visualization of a categorical variable,

- Categories must be **mutually exclusive**
  no observation belongs to two or more categories

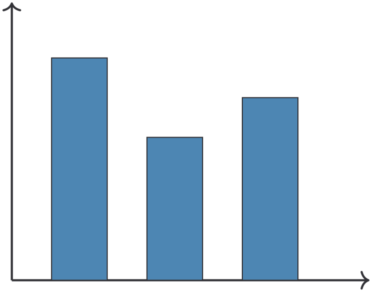
- Categories must be **exhaustive**
  every observation is in some category

Bar Graphs

Bar Graph

A **bar graph** displays the distribution of a categorical variable using rectangular bars where

- categories appear on the x-axis and
- the area or length of a bar represents the count or proportion of the corresponding category

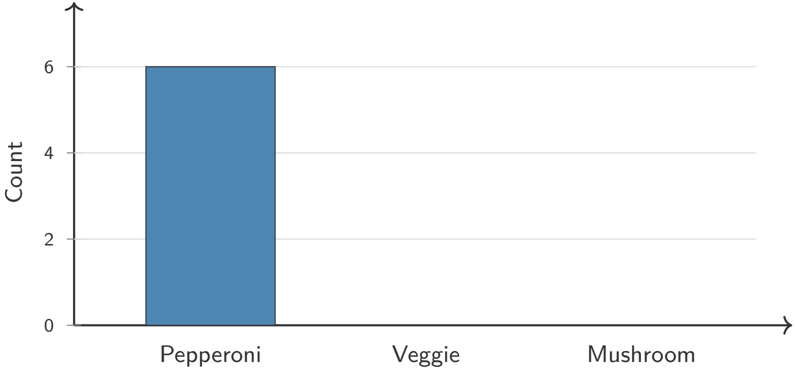


Example (`ex:ch1-pizza-bar`): Constructing a Bar Graph

Recall that the pizza toppings from our survey were: Pepperoni (6), Veggie (5), Mushroom (4).

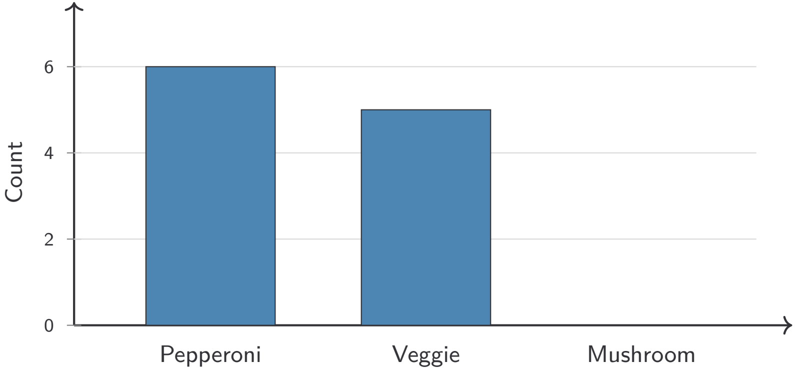
Example (`ex:ch1-pizza-bar`): Constructing a Bar Graph

Recall that the pizza toppings from our survey were: Pepperoni (6), Veggie (5), Mushroom (4).



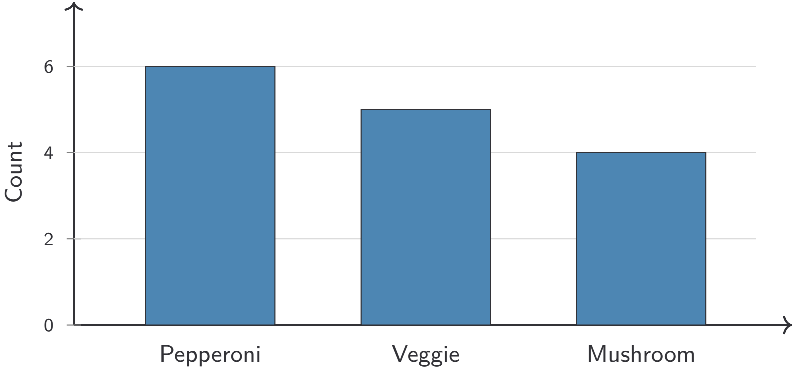
Example (`ex:ch1-pizza-bar`): Constructing a Bar Graph

Recall that the pizza toppings from our survey were: Pepperoni (6), Veggie (5), Mushroom (4).



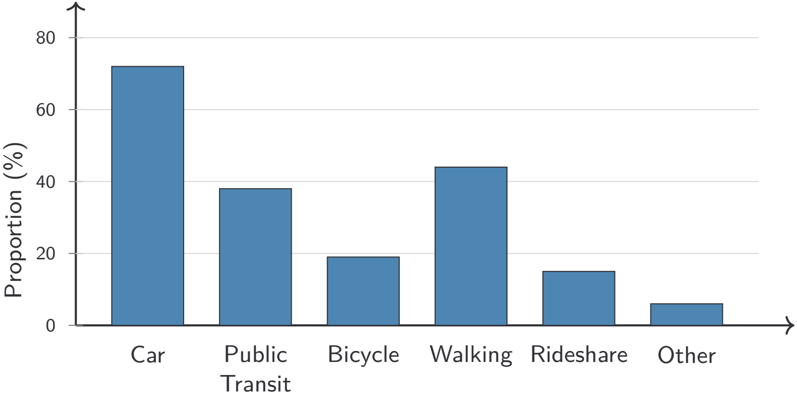
Example (ex:ch1-pizza-bar): Constructing a Bar Graph

Recall that the pizza toppings from our survey were: Pepperoni (6), Veggie (5), Mushroom (4).



Note that

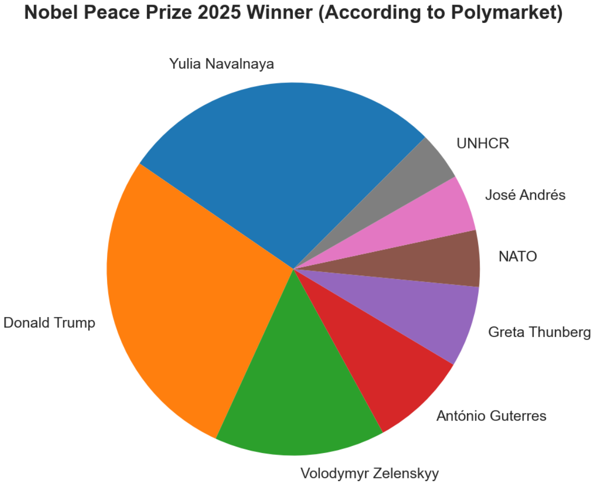
- Bars should have gaps between them
- Order of categories does not matter
- Works for any categorical variable



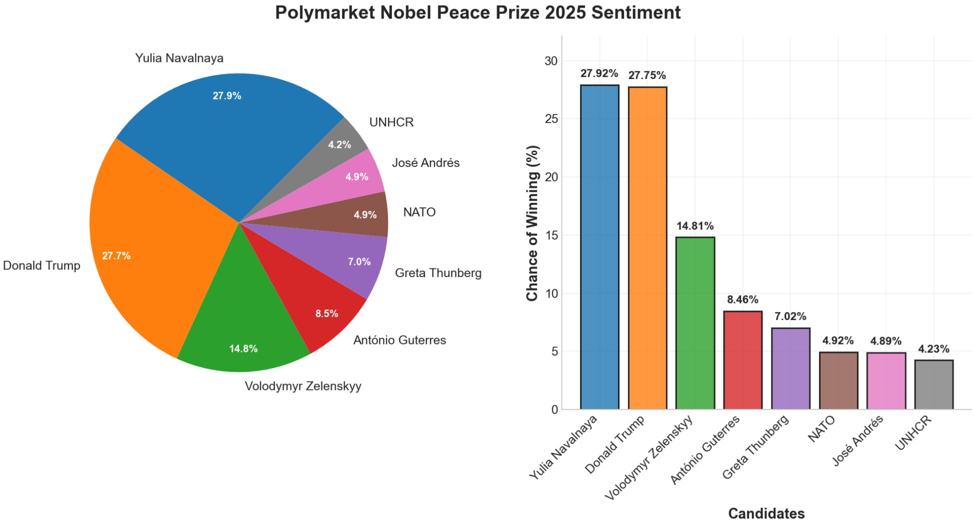
Pie Charts vs Bar Graphs

Example (`ex:ch1-pie-vs-bar`)

LIVE



Pie Charts vs Bar Graphs



Pie Charts vs Bar Graphs

Use pie charts when...

- showing parts of a whole

Pie Charts vs Bar Graphs

Use pie charts when...

- showing parts of a whole
- few categories (3–5)

Pie Charts vs Bar Graphs

Use pie charts when...

- showing parts of a whole
- few categories (3–5)
- the data pertains to pizza or other circles

Pie Charts vs Bar Graphs

Use pie charts when...

- showing parts of a whole
- few categories (3–5)
- the data pertains to pizza or other circles

Use bar graphs when...

Pie Charts vs Bar Graphs

Use pie charts when...

- showing parts of a whole
- few categories (3–5)
- the data pertains to pizza or other circles

Use bar graphs when...

- ...ever but especially when
- there are many categories

Pie Charts vs Bar Graphs

Use pie charts when...

- showing parts of a whole
- few categories (3–5)
- the data pertains to pizza or other circles

Use bar graphs when...

- ...ever but especially when
- there are many categories

🧠 Bar graphs versus pie charts

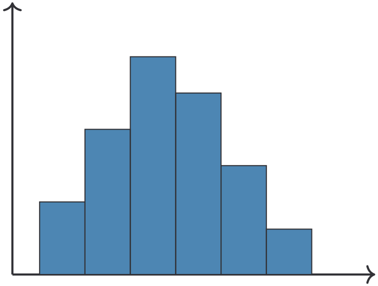
Bar graphs are almost always better as a visualization because humans are better at decoding lengths more accurately than angles.

Histograms

Histogram

A visualization for quantitative data where

- The x-axis is divided into bins,
- each covering a specific range of values of the variable.
- For each bin, a vertical bar is drawn whose height encodes the count or proportion of observations in that bin.



Building a Histogram

Start with Raw Data

A sample of 20 students were asked: *"How many hours of sleep did you get last night?"*

```
7, 5, 8, 6, 7, 9, 6, 7, 8, 4, 7, 6, 8, 5, 7, 10, 9, 6, 8, 7
```

Building a Histogram

Count Observations in Each Bin

Sorted data:

```
4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 9, 9, 10
```

Bin	Count
$[4, 5)$	
$[5, 6)$	
$[6, 7)$	
$[7, 8)$	
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

```
4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 9, 9, 10
```

Bin	**Count**
$[4, 5)$	1
$[5, 6)$	
$[6, 7)$	
$[7, 8)$	
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

```
4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 9, 9, 10
```

Bin	**Count**
$[4, 5)$	**1**
$[5, 6)$	2
$[6, 7)$	
$[7, 8)$	
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,

8, 8, 8, 8, 9, 9, 10

Bin	Count
$[4, 5)$	**1**
$[5, 6)$	**2**
$[6, 7)$	4
$[7, 8)$	
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

```
4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 9, 9, 10
```

Bin	Count
$[4, 5)$	**1**
$[5, 6)$	**2**
$[6, 7)$	**4**
$[7, 8)$	6
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,

8, 8, 8, 8, 9, 9, 10

Bin	Count
$[4, 5)$	**1**
$[5, 6)$	**2**
$[6, 7)$	**4**
$[7, 8)$	**6**
$[8, 9)$	4
$[9, 10)$	
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,
8, 8, 8, 8, 9, 9, 10

Bin	Count
$[4, 5)$	**1**
$[5, 6)$	**2**
$[6, 7)$	**4**
$[7, 8)$	**6**
$[8, 9)$	**4**
$[9, 10)$	2
$[10, 11)$	

Building a Histogram

Count Observations in Each Bin

Sorted data:

4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7,

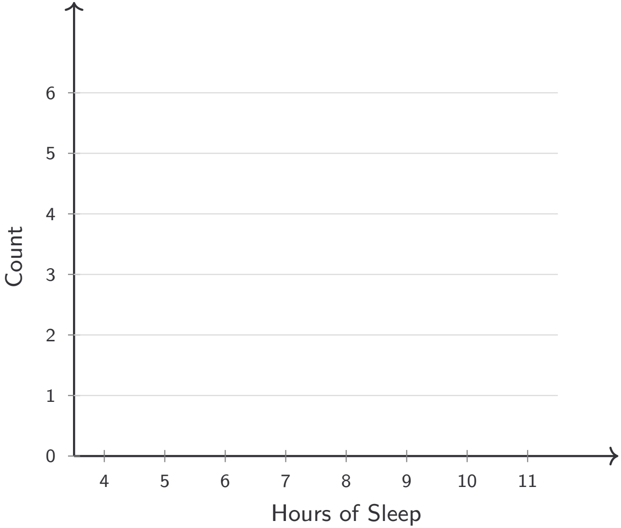
8, 8, 8, 8, 9, 9, 10

Bin	Count
$[4, 5)$	**1**
$[5, 6)$	**2**
$[6, 7)$	**4**
$[7, 8)$	**6**
$[8, 9)$	**4**
$[9, 10)$	**2**
$[10, 11)$	1

Building a Histogram

Draw bars with heights equal to counts

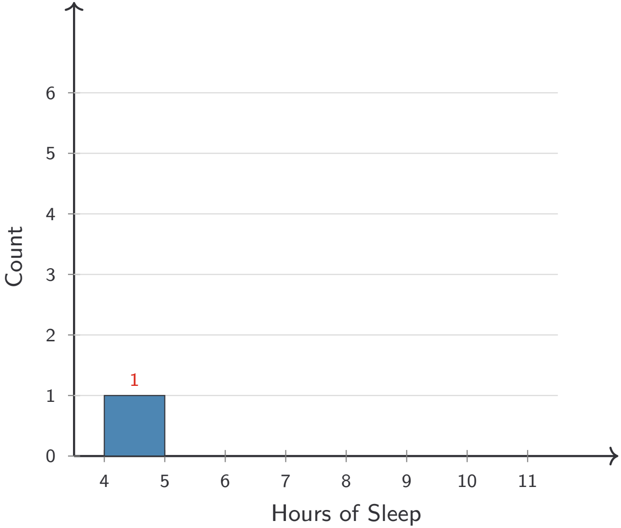
Bin	Count
$[4,5)$	
$[5,6)$	
$[6,7)$	
$[7,8)$	
$[8,9)$	
$[9,10)$	
$[10,11)$	



Building a Histogram

Draw bars with heights equal to counts

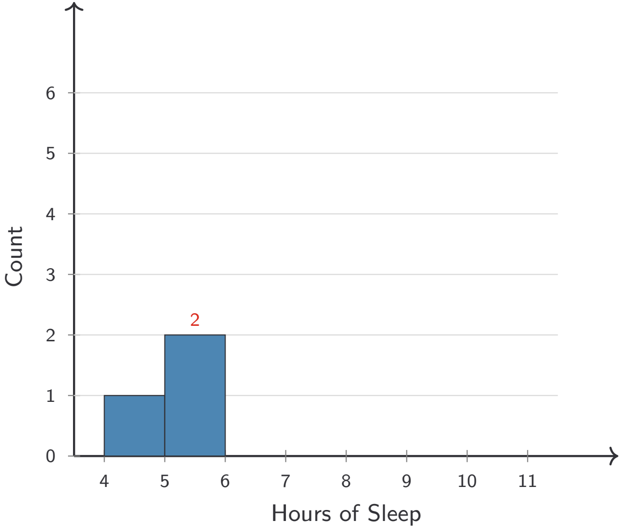
Bin	Count
$[4, 5)$	1
$[5, 6)$	
$[6, 7)$	
$[7, 8)$	
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	



Building a Histogram

Draw bars with heights equal to counts

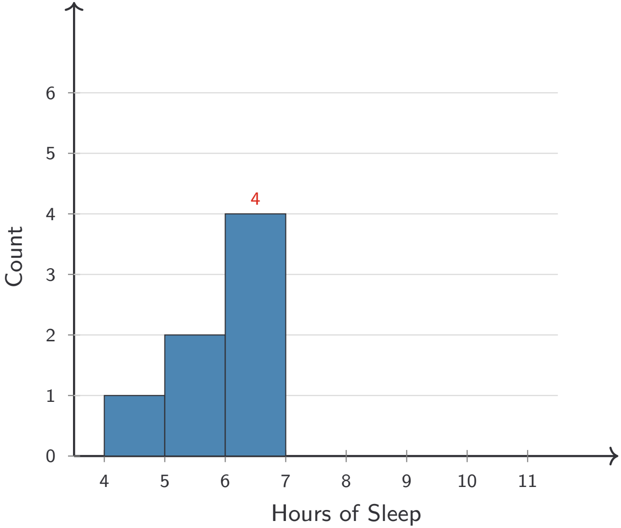
Bin	Count
$[4,5)$	1
$[5,6)$	2
$[6,7)$	
$[7,8)$	
$[8,9)$	
$[9,10)$	
$[10,11)$	



Building a Histogram

Draw bars with heights equal to counts

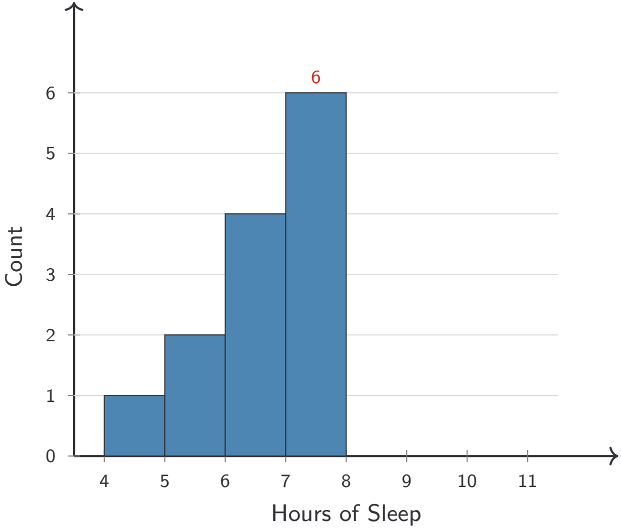
Bin	Count
$[4, 5)$	1
$[5, 6)$	2
$[6, 7)$	4
$[7, 8)$	
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	



Building a Histogram

Draw bars with heights equal to counts

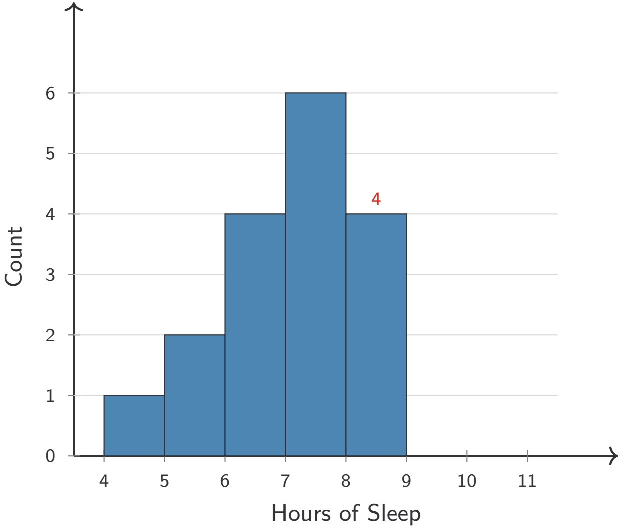
Bin	Count
$[4, 5)$	1
$[5, 6)$	2
$[6, 7)$	4
$[7, 8)$	6
$[8, 9)$	
$[9, 10)$	
$[10, 11)$	



Building a Histogram

Draw bars with heights equal to counts

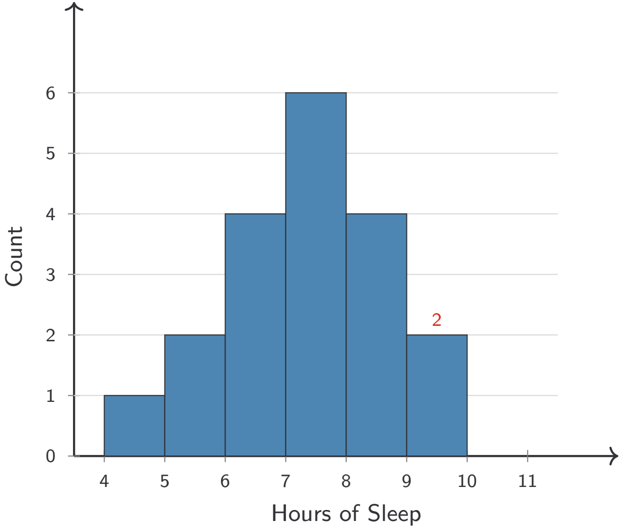
Bin	Count
$[4, 5)$	1
$[5, 6)$	2
$[6, 7)$	4
$[7, 8)$	6
$[8, 9)$	4
$[9, 10)$	
$[10, 11)$	



Building a Histogram

Draw bars with heights equal to counts

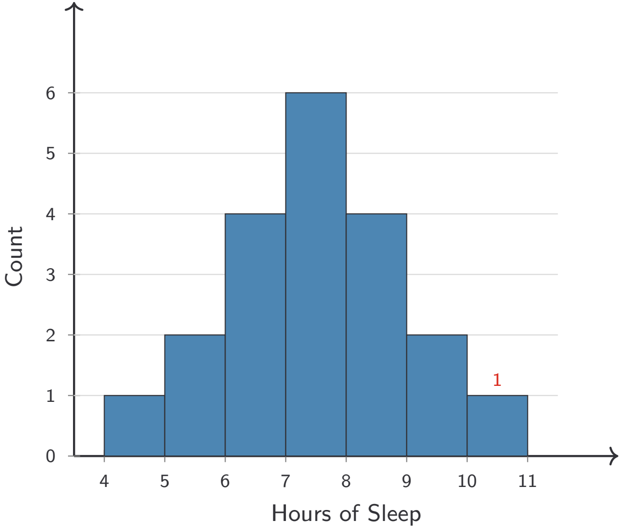
Bin	Count
$[4, 5)$	1
$[5, 6)$	2
$[6, 7)$	4
$[7, 8)$	6
$[8, 9)$	4
$[9, 10)$	2
$[10, 11)$	



Building a Histogram

Draw bars with heights equal to counts

Bin	Count
$[4, 5)$	1
$[5, 6)$	2
$[6, 7)$	4
$[7, 8)$	6
$[8, 9)$	4
$[9, 10)$	2
$[10, 11)$	1



How to Draw a Histogram

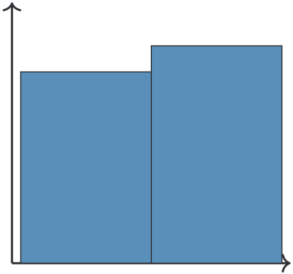
How to Draw a Histogram

1. Divide values into equal-width bins[a]
2. Count observations in each interval
3. Draw adjacent bars with heights representing counts

[a] The choice of bin width and location is arbitrary and can affect the appearance of the histogram

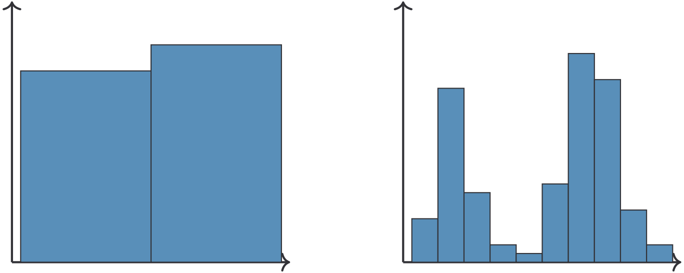
Choosing Bin Width: Old Faithful Geyser

Data: Duration of eruptions (in minutes)



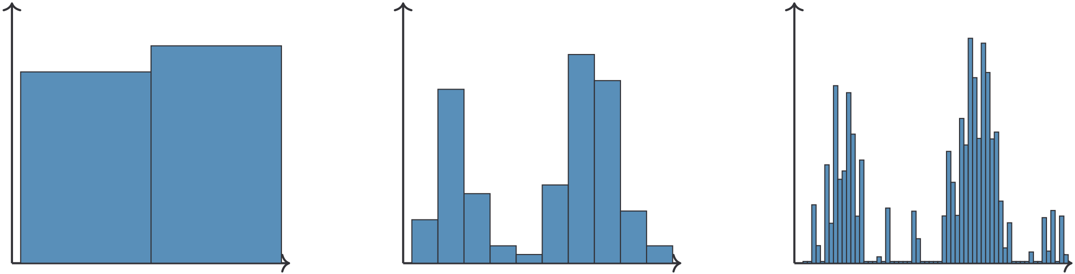
Choosing Bin Width: Old Faithful Geyser

Data: Duration of eruptions (in minutes)



Choosing Bin Width: Old Faithful Geyser

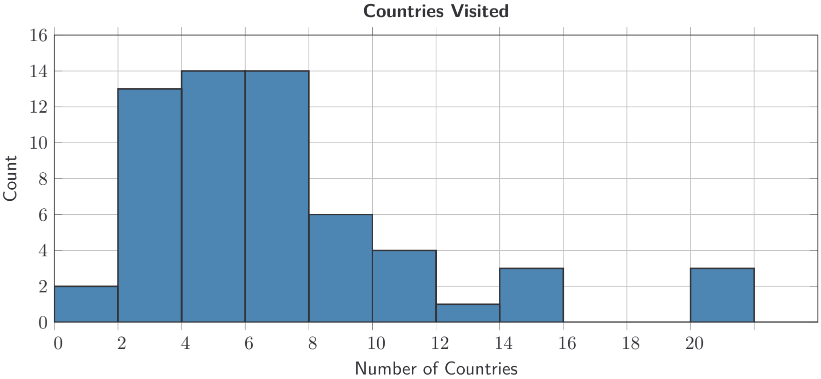
Data: Duration of eruptions (in minutes)



Example (`ex:ch1-countries-histogram`): Countries Visited Histogram

Countries Visited by DS 1000A Students

Find the number of students who visited fewer than 8 countries.



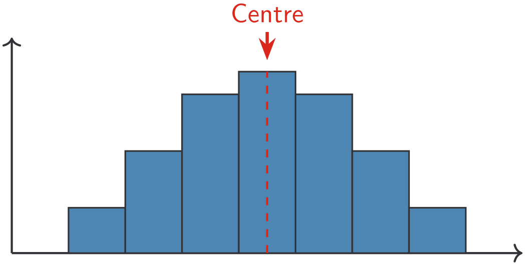
Describing Distributions

- Centre: What's a typical value?
- Spread: How much do the values vary?
- Outliers: Are there any unusual values?
- Shape: Symmetric? Skewed?

Centre

Centre

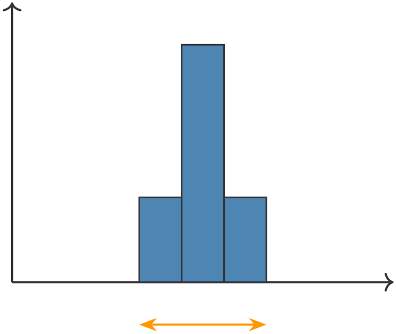
A **typical** or **representative** value for the distribution.



Spread

Spread

How much the data values can differ from one another.



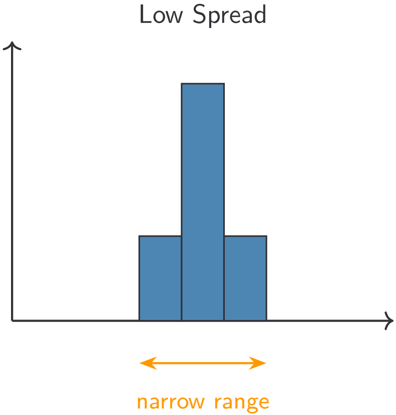
Low Spread

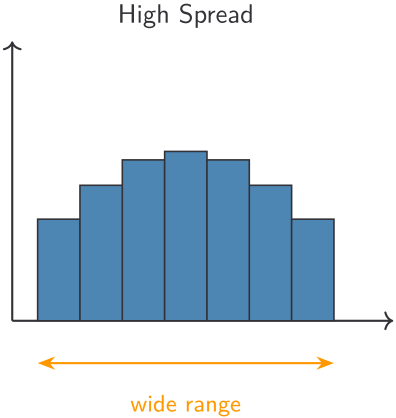
narrow range

Spread

Spread

How much the data values can differ from one another.


Low Spread


High Spread

Centre and spread

Center and spread

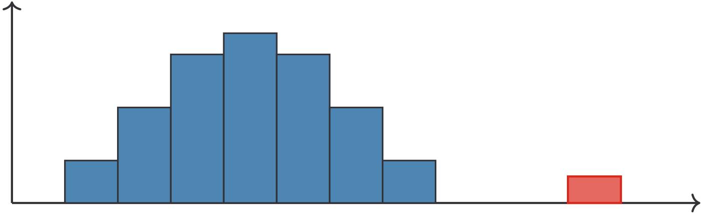
For now, we will keep the concepts of centre and spread abstract.

In Chapter 2, we will learn numerical summaries to quantify these features.

Outliers

Outlier

A value that lies far from the rest of the data.

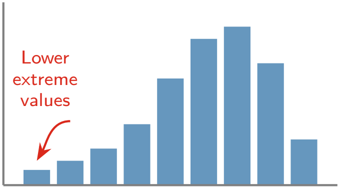


Tail of a Distribution

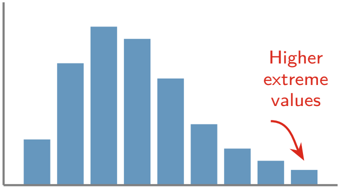
Tail of a Distribution

The portion of the distribution that extends away from the centre toward extreme values.

Longer Left Tail

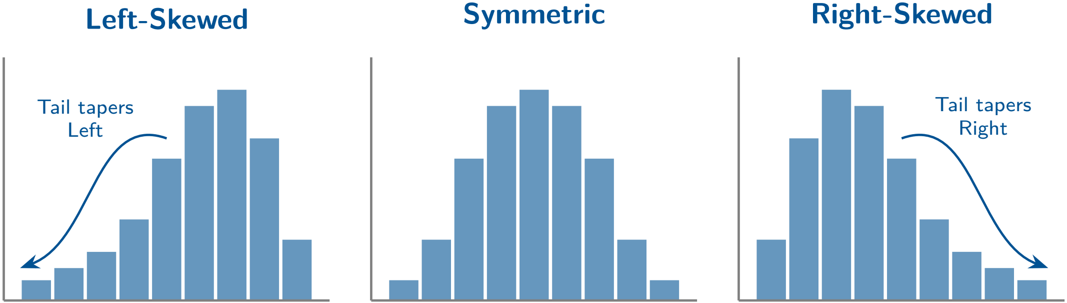


Longer Right Tail



The tail is the side that tapers off gradually.

Shape: Symmetric vs Skewed



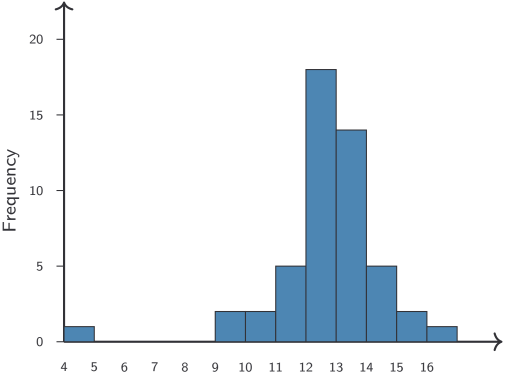
Left-Skewed

Symmetric

Right-Skewed

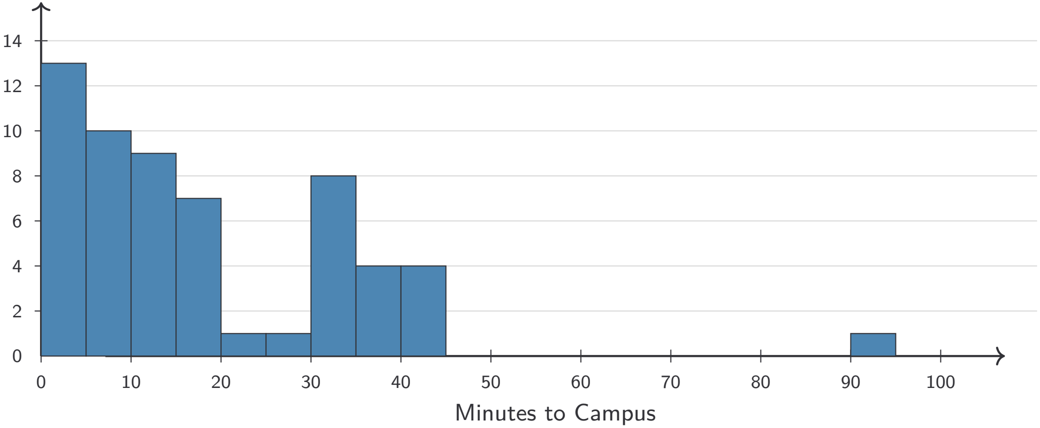
Shape of Distribution

What is the shape of the following distribution?



Shape of a Distribution

Describe the shape of the following histogram.



Stemplots (Stem-and-Leaf Plots)

Stemplot

Split each observation into a

- **stem** (all but the final digit) and a
- **leaf** (the trailing digit).

Data:

$13, \quad 22, \quad 22, \quad 24, \quad 43$

Stem	Leaf
1	3
2	2 2 4
3	
4	3

Stemplot, more digits

Data:

$$1364, \quad 1365, \quad 1366, \quad 1370, \quad 1371$$

Stemplot:

Stem	Leaf
136	4 5 6
137	0 1

Stemplot, with decimals

Data:

$$1364.03, \quad 1365.1, \quad 1366.00, \quad 1370.02, \quad 1371.9$$

Stemplot:

Stem	Leaf
1364	0 3
1365	1
1366	0
1370	0 2
1371	9

Example (`ex:ch1-test-scores`): Building a Stemplot

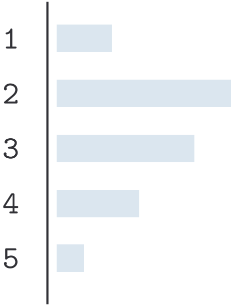
Construct the stemplot for the following test scores dataset:

$$67, \quad 72, \quad 74, \quad 78, \quad 81, \quad 83, \quad 85, \quad 87, \quad 88, \quad 91, \quad 95, \quad 98$$

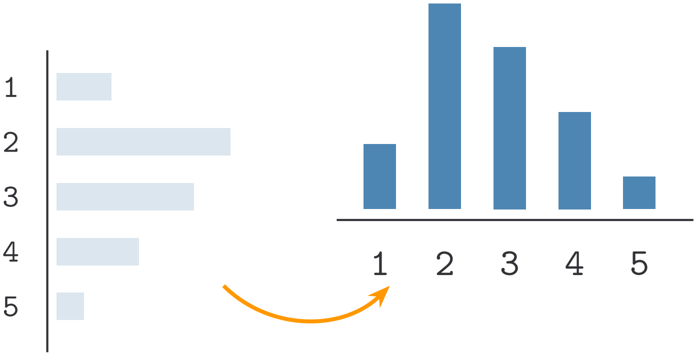
Stemplot:

Stem	Leaf

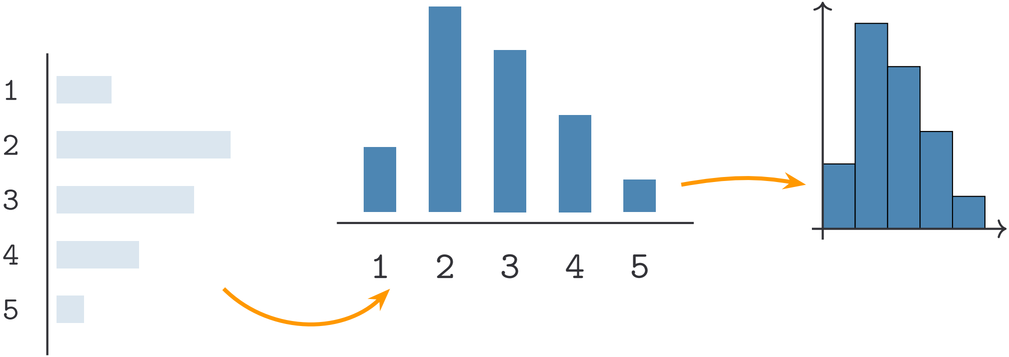
Determining Shape from Stemplots



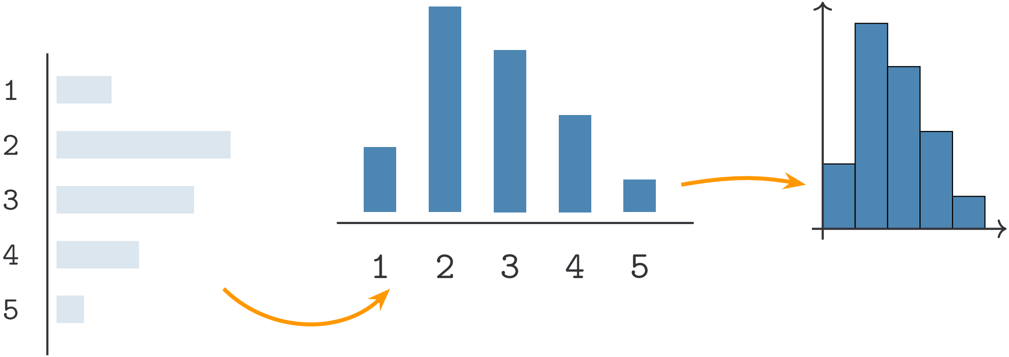
Determining Shape from Stemplots



Determining Shape from Stemplots



Determining Shape from Stemplots



Example (`ex:ch1-unemployment-stemplot`): Interpreting a Stemplot

The following stemplot shows the percentage unemployment rates for various countries in 2011.

Comment on the outliers and shape of the distribution.

4	2
5	
6	
7	
8	
9	7 9
10	0 8
11	1 5 5 6 6
12	0 1 2 2 2 3 4 4 4 4 5 7 8 8 8 9 9 9
13	0 1 2 3 3 3 3 3 4 4 4 8 9 9
14	0 2 6 6 6
15	2 3
16	8

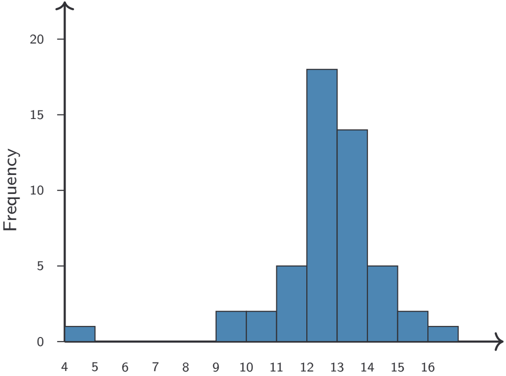
Stemplot

4	2
5	
6	
7	
8	
9	7 9
10	0 8
11	1 5 5 6 6
12	0 1 2 2 2 3 4 4 4 4 5 7 8 8 8 9 9 9
13	0 1 2 3 3 3 3 3 4 4 4 8 9 9
14	0 2 6 6 6
15	2 3
16	8

Stemplot

```
 4 | 2
 5 |
 6 |
 7 |
 8 |
 9 | 7 9
10 | 0 8
11 | 1 5 5 6 6
12 | 0 1 2 2 2 3 4 4 4 4 5 7 8 8 8 9 9 9
13 | 0 1 2 3 3 3 3 3 4 4 4 8 9 9
14 | 0 2 6 6 6
15 | 2 3
16 | 8
```

As Histogram



Using stemplots

Key Point: We can use stemplots when

- the datasets are small to moderate datasets (up to about 50 observations)
- it is important to see granularity (preserving actual data values) (unlike histograms)

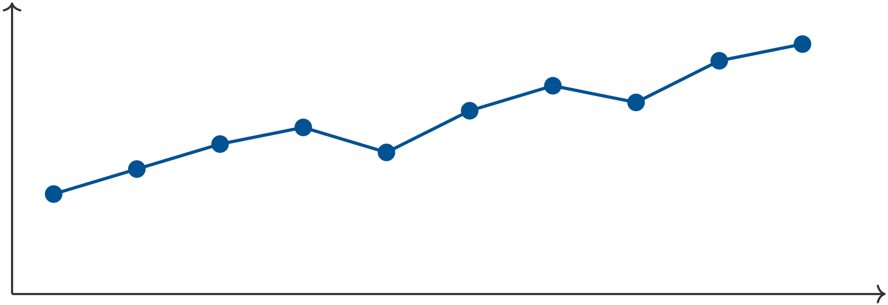
For larger datasets, histograms are usually preferred.

Time Plots

Time Plot (Time Series)

Shows how a variable **changes over time**.

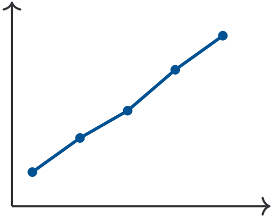
- Time on x-axis,
- quantitative variable on y-axis.



Time Plot Patterns

Features to look for in time plots:

Trend

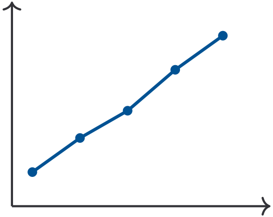


Long-term direction

Time Plot Patterns

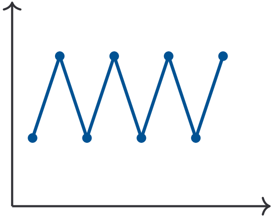
Features to look for in time plots:

Trend



Long-term
direction

Seasonality

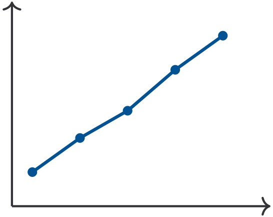


Cyclic
pattern

Time Plot Patterns

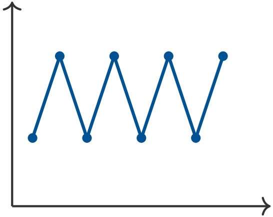
Features to look for in time plots:

Trend



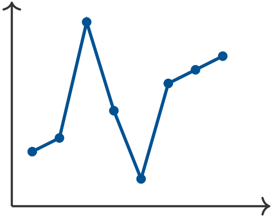
Long-term
direction

Seasonality



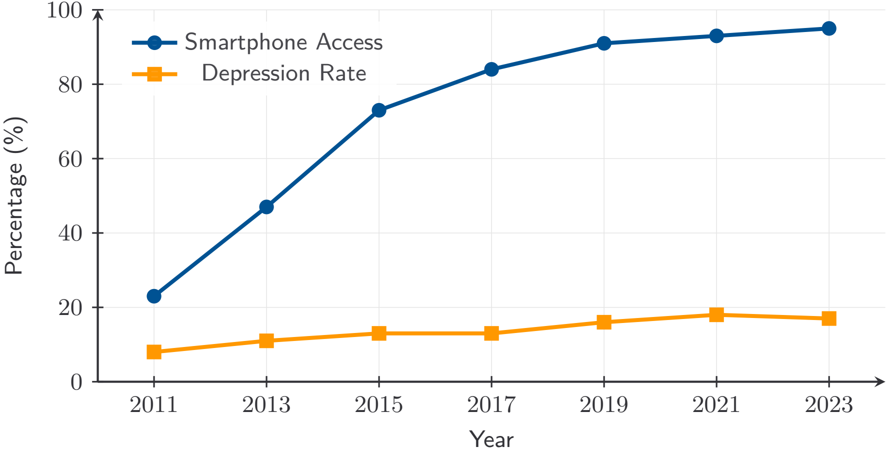
Cyclic
pattern

Deviations

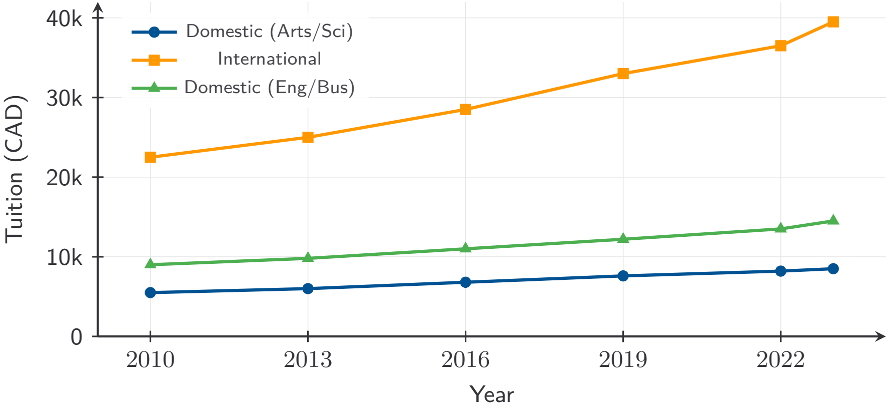


Unusual
behaviour

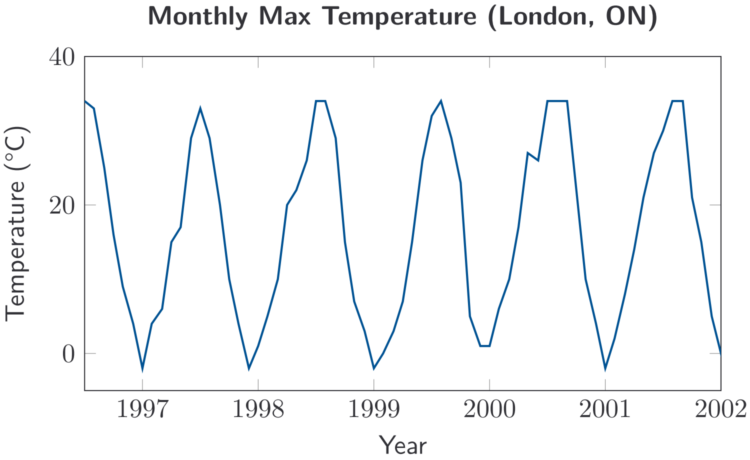
Example (`ex:ch1-smartphone-depression`): US Teen Smartphone Access & Depression



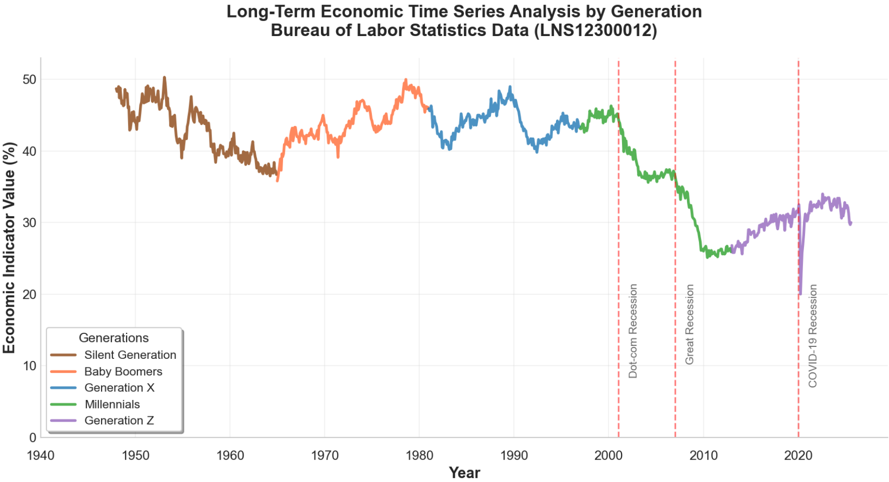
Example (`ex:ch1-tuition`): Ontario University Tuition Trends



Example (`ex:ch1-temperature`): Seasonal Temperature Patterns

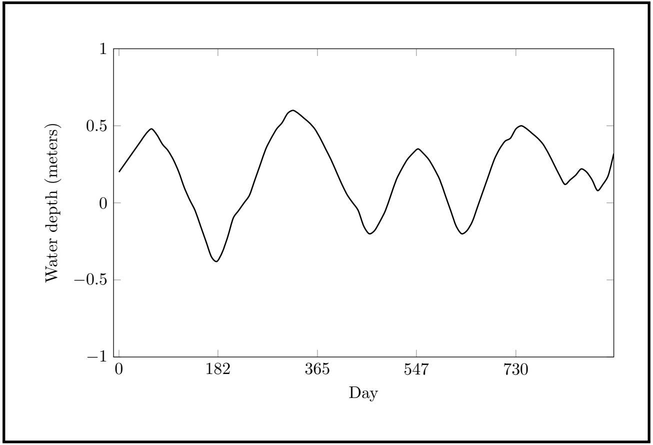


Example (ex:ch1-teen-employment): US Teen Employment Rate (1948–2022)

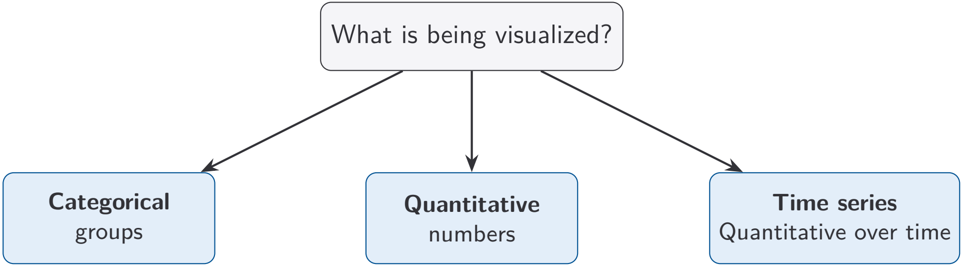


Exercise

Comment on each of the time series features we examined in the context of the following plot:



Choosing the Right Visualisation



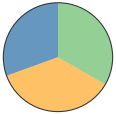
What is being visualized?

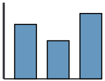
Categorical
groups

Quantitative
numbers

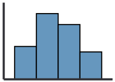
Time series
Quantitative over time

Pie Chart
Bar Graph





Histogram
Stemplot



```
2 | 3
3 | 1 5
4 | 0 2 2 4
```

Time Plot

Chapter 1 Summary

Concepts

- **Individuals** — objects described
- **Variables** — characteristics of individuals
- **Categorical** — groups or labels
- **Quantitative** — numbers
- **Distribution** — which values a variable takes, and how often

Interpreting Quantitative Data

- **Overall pattern**
    - Quantitative: centre, spread, shape
    - Quantitative + time: trend, seasonality
- **Deviations from patterns**
    - Quantitative: outliers
    - Quantitative + time: deviations from trend/seasonality