CHAPTER 1

Picturing Distributions with Graphs

Goals

- Define data, individuals, and variables
- Distinguish between categorical and quantitative variables
- Create and interpret pie charts and bar graphs for categorical data
- Create and interpret histograms and stemplots for quantitative data
- Describe distributions using shape, centre, spread, and outliers
- Recognise trends, seasonality, and deviations in time plots

1.1. Individuals and Variables

Definition 1.1 *Individuals and Variables*
Individuals are _____.

A **variable** is any _____. A variable can _____.

Example 1.2 *Canadian Ice Cream Survey*
Consider the following dataset from a survey of Canadian adults about ice cream preferences:

	gender	favourite flavour	cones per month	age	lactose intolerant
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
6	Male	Chocolate	1	25	No

Identify the individuals and variables in this dataset.

Individuals: _____

Variables: _____

Example 1.3 *Netflix Viewing Patterns*
A streaming platform surveyed 200 households about their Netflix viewing habits. The following is

a sample of 6 households from the dataset:

household	primary genre	hours/week	members	subscription	active months
1	Drama	12.5	3	Premium	48
2	Comedy	8.3	2	Standard	36
3	Action	15.2	4	Premium	60
4	Drama	6.1	1	Basic	12
5	Documentary	4.7	2	Standard	24
6	Comedy	10.8	3	Premium	54

Identify the individuals, variables, and classify each variable as categorical or quantitative (after reviewing the definition below).

1.2. Types of Variables

Definition 1.4 *Variable types*
A **categorical variable** is a variable that can take on one of a _____ _____. It assigns each individual to a particular group or category based on some property.

A **quantitative variable** is a variable that takes _____ _____. It describes a measurable quantity rather than a category or group.

Example 1.5 *Dating Survey*
The following dataset was obtained through an online survey of Americans about how they met their partners:

meeting_method	age	gender	relationship_length (months)	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

List the individuals in the dataset and the variables, classifying each variable as categorical or

quantitative.

Common Mistake: Just because a variable contains numbers does **not** make it quantitative. Postal codes, phone numbers, and jersey numbers are categorical because arithmetic operations on them are meaningless.

1.3. Exploratory Data Analysis

Definition 1.6 *Distribution*
The **distribution** of a variable tells us what values it takes and how often it takes these values.

Exploratory data analysis is the process of using graphs and numerical summaries to describe the variables in a dataset or the relationships among them. This will be the aim of this chapter and chapter 2.

Example 1.7 *Favourite Pizza Toppings*
A sample of 15 students was asked about their favourite pizza topping. The results of the survey are shown below.

	favourite_topping
0	Veggie
1	Pepperoni
2	Mushroom
3	Veggie
4	Pepperoni
5	Mushroom
6	Pepperoni
7	Veggie
8	Mushroom
9	Pepperoni
10	Veggie
11	Pepperoni
12	Mushroom
13	Veggie
14	Pepperoni

The distribution can be summarised in the following table:

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

🧠 The formal description of a distribution is more technical than we present here. In fact, it can refer to one of several equivalent (but different) concepts. But for our purposes, we will think of it as follows (for now):

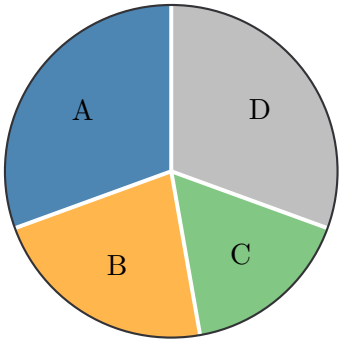
Distribution for a categorical variable can be described by listing the categories and giving the count or percent of individuals in each category.

For quantitative variables, we will learn how to visualise distributions.

1.4. Categorical Variables: Pie Charts and Bar Graphs

1.4.1 Pie Charts

Definition 1.8 *Pie Chart*
The **pie chart** of a categorical variable is a circle divided into _____, where each _____ represents a _____ of the variable.

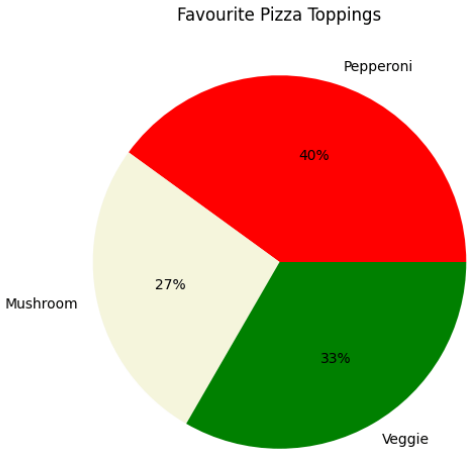


Example 1.9 *Favourite Pizza Toppings*
A sample of 15 students was asked about their favourite pizza topping. The *distribution* of responses is shown below.

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

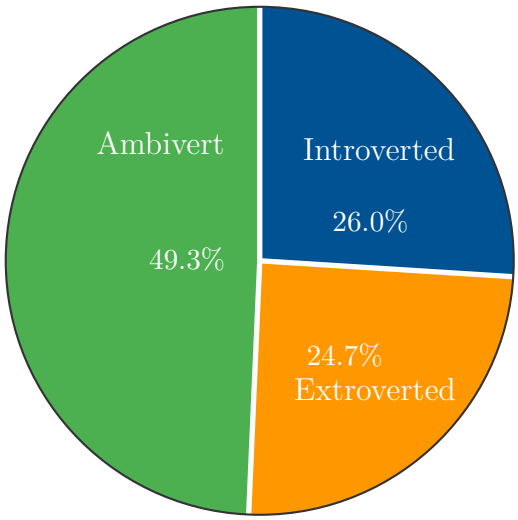
Basic pie chart

```python
import matplotlib.pyplot as plt

# Step 1: Define our data
# List of pizza topping names
toppings = ['Pepperoni', 'Mushroom', 'Veggie']

# Number of people who like each topping
counts = [6, 4, 5]

# Colours for each slice (red, orange, green)
colours = ['red', 'beige', 'green']

# Step 2: Create a figure for our chart
# figsize makes the chart 6 inches by 6 inches
fig, ax = plt.subplots(figsize=(6, 6))

# Step 3: Create the pie chart
# autopct shows percentages on each slice
ax.pie(counts, labels=toppings, autopct='%1.0f%%',
 ↪  colors=colours)

# Step 4: Add a title
ax.set_title('Favourite Pizza Toppings')

# Step 5: Save and display the chart
plt.savefig('pizza_pie.png')
plt.show()
```



Interpretation: Pepperoni is the most popular topping (40%), followed closely by Veggie (33%) and Mushroom (27%).

Example 1.10 *Student Introversion Distribution*
The following pie chart shows the distribution of introversion types from the Fall 2025 DS 1000A class survey.



Creating the Introversion Pie Chart

```
import pandas as pd
import matplotlib.pyplot as plt

# Read the survey data
df = pd.read_csv('DS1000_survey.csv')

# Count the frequency of each introversion
 ↪  type
intro_counts =
 ↪  df['introversion_type'].value_counts()

# Calculate percentages
intro_pct = (intro_counts /
 ↪  intro_counts.sum()) * 100

# Define colours matching the Swiss modern
 ↪  palette
colours = ['#4CAF50', '#FF9800', '#005293']

# Create the pie chart
fig, ax = plt.subplots(figsize=(8, 6))
ax.pie(intro_pct, labels=intro_pct.index,
  autopct='%1.1f%%', colors=colours,
   ↪  startangle=90)

ax.set_title('Student Introversion Type
 ↪  Distribution')
plt.tight_layout()
plt.savefig('introversion_pie.png', dpi=150)
plt.show()
```

Interpretation: Nearly half of the respondents identify as ambivert, with introverts and extroverts making up roughly equal portions of the remaining students.

Remark
The pie chart requires that the categories be _____ (no overlap) and _____ (all individuals must fit into one of the categories).

If these conditions are not met, a pie chart should **not** be used.

Example 1.11 *Transportation Modes*
A survey of adults asked: "Which modes of transportation do you use regularly? (Select all that apply)"

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

Can this data be represented with a pie chart? Explain why or why not.

1.4.2 Bar Graphs

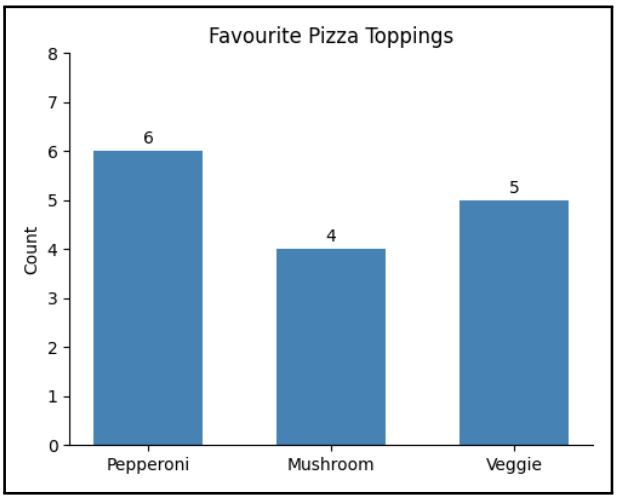
Another way to visualise categorical data is with a bar graph.

Definition 1.12 *Bar Graph*
A **bar graph** or **bar chart** is a graphical display in which _____ are shown along one axis, and a _____ is shown along the other axis representing a _____.

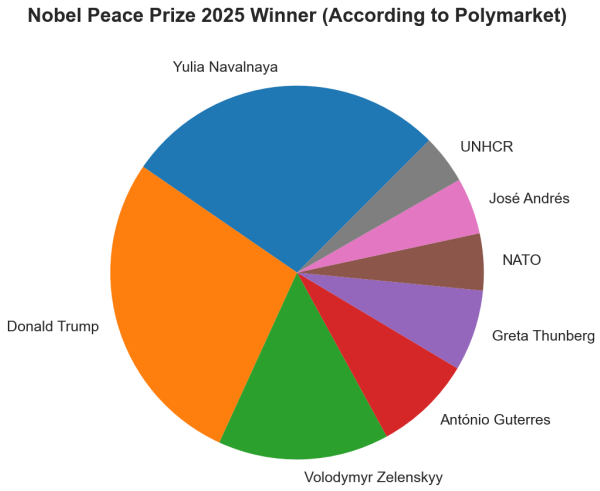
Example 1.13 *Favourite Pizza Toppings (Bar Graph)*
Using the same data as the pie chart example, we can visualise the pizza preferences using a bar graph.

Basic bar graph

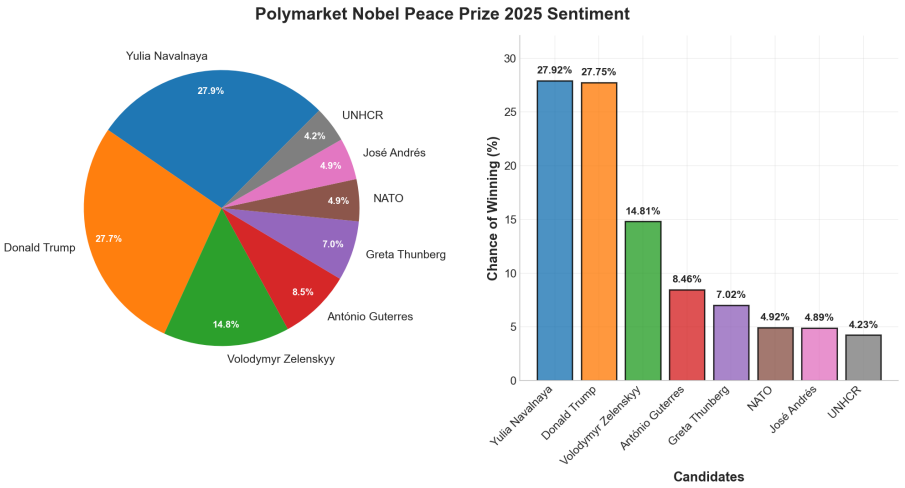
```python
import matplotlib.pyplot as plt

# Data
toppings = ['Pepperoni', 'Mushroom',
 ↪  'Veggie']
counts = [6, 4, 5]

# Create figure
fig, ax = plt.subplots(figsize=(5, 4))

# Create bars
ax.bar(toppings, counts, color='steelblue',
 width=0.6)

# Add count labels on top of each bar
for i, count in enumerate(counts):
    ax.text(i, count + 0.15, str(count),
     ↪  ha='center')

# Labels and title
ax.set_ylabel('Count')
ax.set_title('Favourite Pizza Toppings')

plt.tight_layout()
plt.savefig('pizza_bar.png')
plt.show()
```



Example 1.14 *Comparing Pie Charts and Bar Graphs*
Consider the following visualization of Polymarket betting on 2025 Peace Prize winners.



The pie chart makes it difficult to compare the public sentiment in favour of the largest leading candidates: Donald Trump and Yulia Navalnaya. Why? The slices are similar in size, and our brains are not very good at comparing angles.



Adding percentages can help, but the bar graph on the right makes it much easier to compare the categories directly.

Key point:

- Humans are better at comparing lengths (bars) than angles (slices)
- Pie charts are limited to situations where the categories represent parts of a whole (i.e., the data must sum to 100%)
- Pie charts are weak at showing small differences between categories
- Bar graphs can display many more categories clearly

1.5. Quantitative Variables

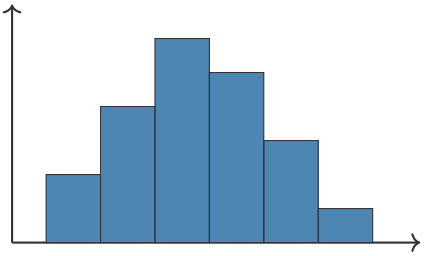
We will now turn our attention to **quantitative variables** and examine two visualizations commonly used to display their distributions: histograms and stemplots. Later, we will see how to visualize a quantitative variable over time using time plots.

1.5.1 Histograms

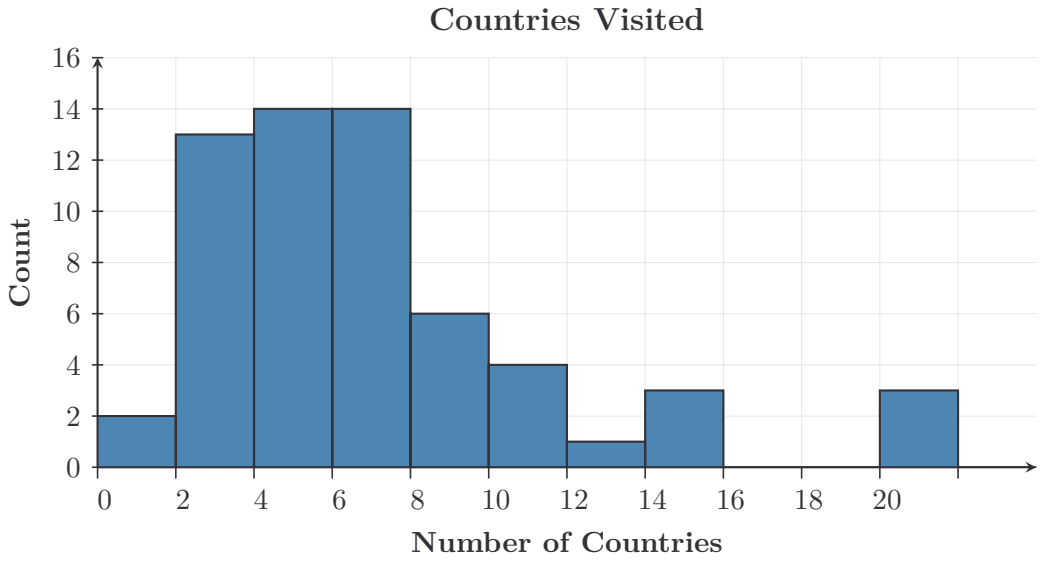
Definition 1.15 *Histogram*
A **histogram** is a graphical display of a quantitative variable, which consists of adjacent bars whose heights represent the counts (or percents) of values falling in each interval, called a bin.

How to Draw a Histogram

1. Divide values into equal-width intervals (bins)
2. Count observations in each interval
3. Draw adjacent bars with heights representing counts



Example 1.16 *Countries Visited by DS 1000 Students*
The following histogram shows the distribution of the number of countries visited by students in the DS 1000 class.



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

1.5.2 Interpreting Histograms

Visualisations make it easy to understand the **overall pattern** of a distribution as well as **deviations** from that pattern.

Overall pattern:

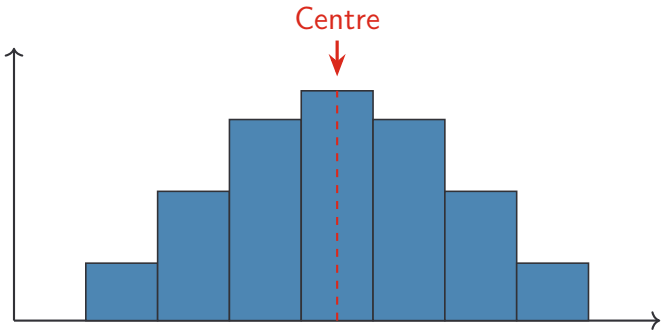
- Centre (location): _____
- Spread: _____
- Shape: _____

Deviations:

- Outliers: _____

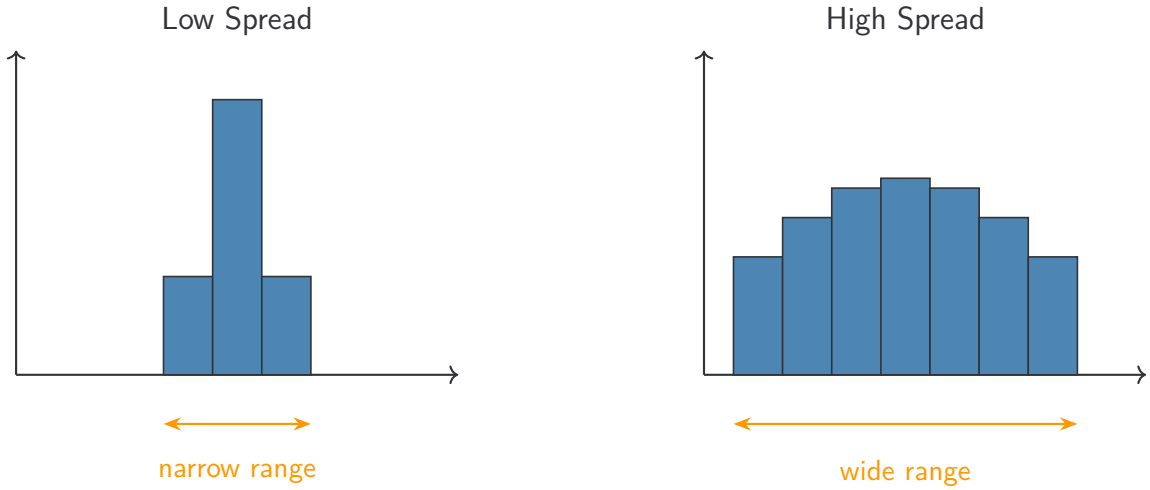
Centre

Definition 1.17 *Centre*
The **centre** of a distribution is a value that is representative of the entire dataset.



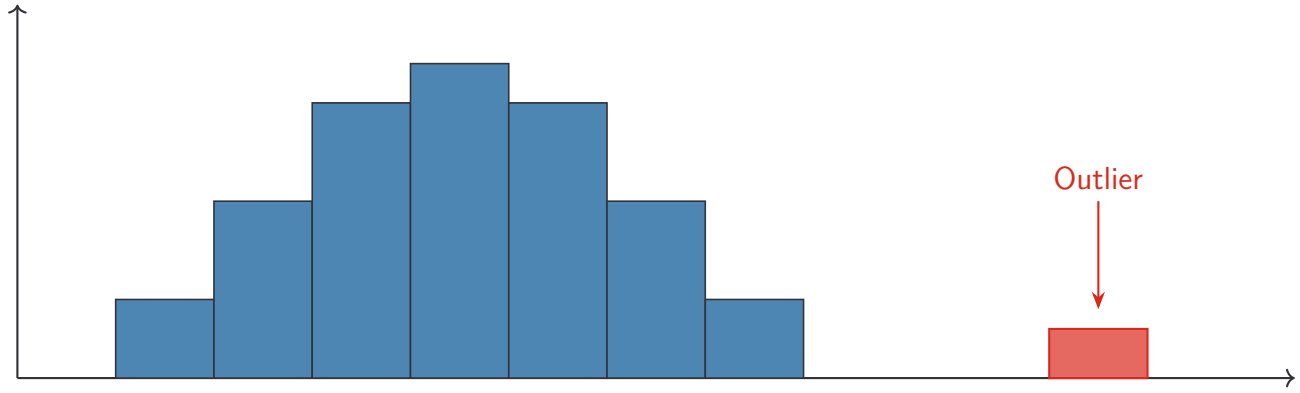
Spread

Definition 1.18 *Spread*
The **spread** (or variability) of a distribution describes how much the data values differ from one another. Distributions with greater spread have values that are more widely scattered.



Outliers

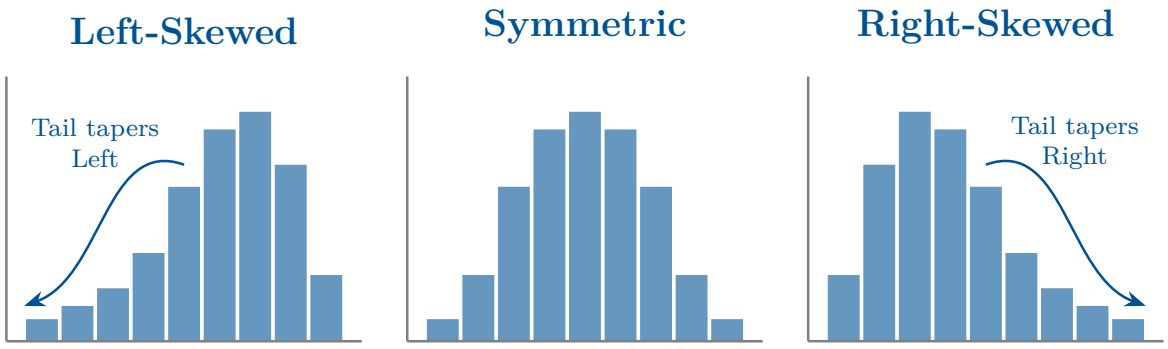
Definition 1.19 *Outlier*
An **outlier** is an observation that lies an abnormal distance from other values in a dataset.



Shape

Definition 1.20 *Tail of a Distribution*
The **tail** of a distribution refers to the portion of the distribution that extends away from the centre toward extreme values. In a histogram, the tail is the side that tapers off gradually.

- A distribution with a longer tail on the right has higher extreme values
- A distribution with a longer tail on the left has lower extreme values



Definition 1.21 *Shape*
A distribution is _____ if the left and right sides of the graph are approximately mirror images of each other.

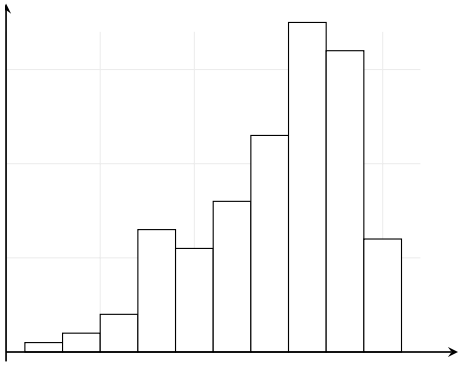
A distribution is _____ if the right side of the graph (larger values) extends much farther out than the left side.

A distribution is _____ if the left side of the graph (smaller values) extends much farther out than the right side.

Intuition: Think of skewness as which direction the "tail" points. Right-skewed distributions have a long tail to the right (toward higher values). Left-skewed distributions have a long tail to the left (toward lower values).

Example 1.22

Describe the shape of the following histogram.

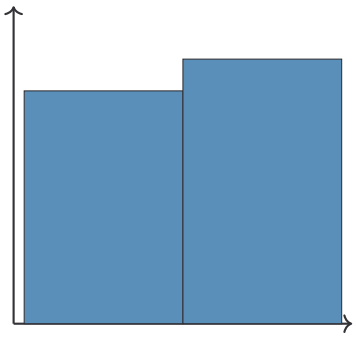


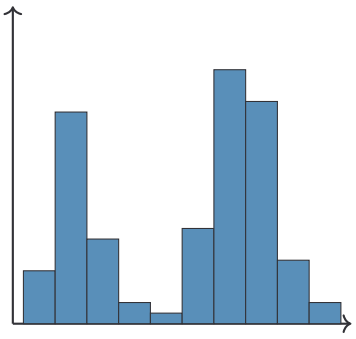
Effect of Bin Width

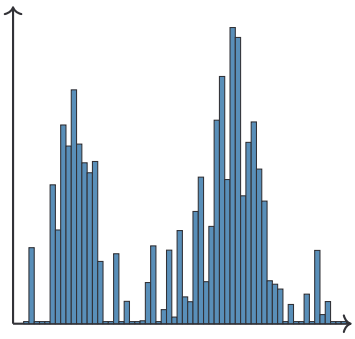
Notice how the shape changes when the bin width and count change.

Choosing the "right" number of bins or bin size influences what the histogram reveals about the data.

Too few bins can obscure important details (oversmoothing), while too many bins can introduce noise and make it hard to see the overall pattern.







1.5.3 Stemplots

Definition 1.23 *Stemplot*
A **stemplot** (or stem-and-leaf plot) is a simple way to display the distribution of a quantitative variable while preserving the actual data values. Each observation is split into a **stem** (leading digits) and a **leaf** (trailing digit).

How to Create a Stemplot

1. Separate each observation into a stem (all digits except the last) and a leaf (the last digit)
2. Write stems in a vertical column from smallest to largest
3. Write each leaf to the right of its stem, in order from smallest to largest

Example 1.24 *Test Scores*
The following are test scores (out of 100) for 15 students:

67 72 74 78 81 82 83 85 87 88 91 93 95 97 98

Create a stemplot for this data.

Example 1.25 *Unemployment Rates Stemplot*
The following stemplot shows the percentage unemployment rates for various countries in 2011.

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

Example 1.26 *Split Stems*
When there are many observations, we can **split stems** to show more detail. Each stem appears twice: once for leaves 0–4 and once for leaves 5–9.

For the data: 51 52 53 55 56 57 58 59 61 62 63 64 65 66 67 68

Stem	Leaves
5L	1 2 3
5H	5 6 7 8 9
6L	1 2 3 4
6H	5 6 7 8

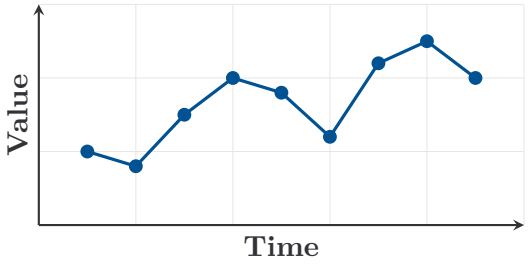
Key: 5|7 means 57.

> **Key point:** Stemplots are useful for:
>
> - Small to moderate datasets (up to about 50 observations)
> - Preserving actual data values (unlike histograms)
> - Quickly identifying the shape, centre, and spread
>
> For larger datasets, histograms are usually preferred.

1.6. Time Plots

For some variables, observations are recorded over time. Examples include daily temperatures, monthly sales figures or population counts over years. To visualize how these variables change over time, we use **time series plots**.

Definition 1.27 *Time Series Plot*
A _____ or a _____ is a graphical display of a quantitative variable against time.

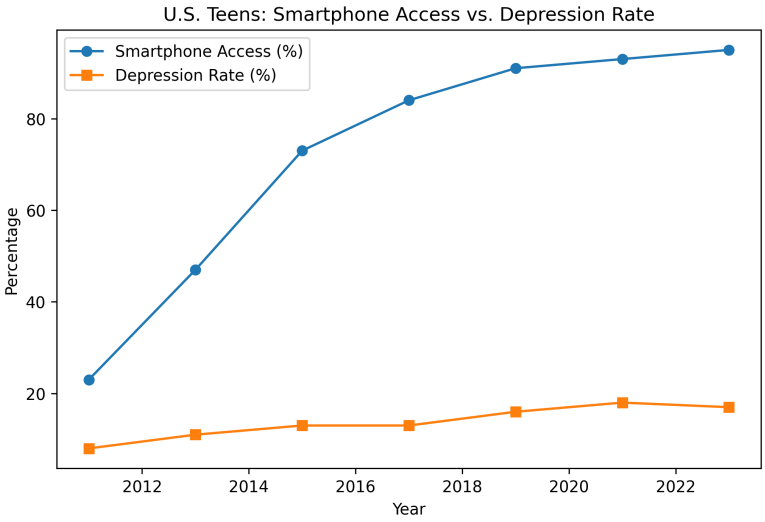


Example 1.28

The following time series plot depicts Pew Research Center's estimates of the proportion of US teens who own a smart phone from 2011 to 2023 and SAMHSA's estimates of US teens who experienced a major depressive episode by year.

Time series in Python

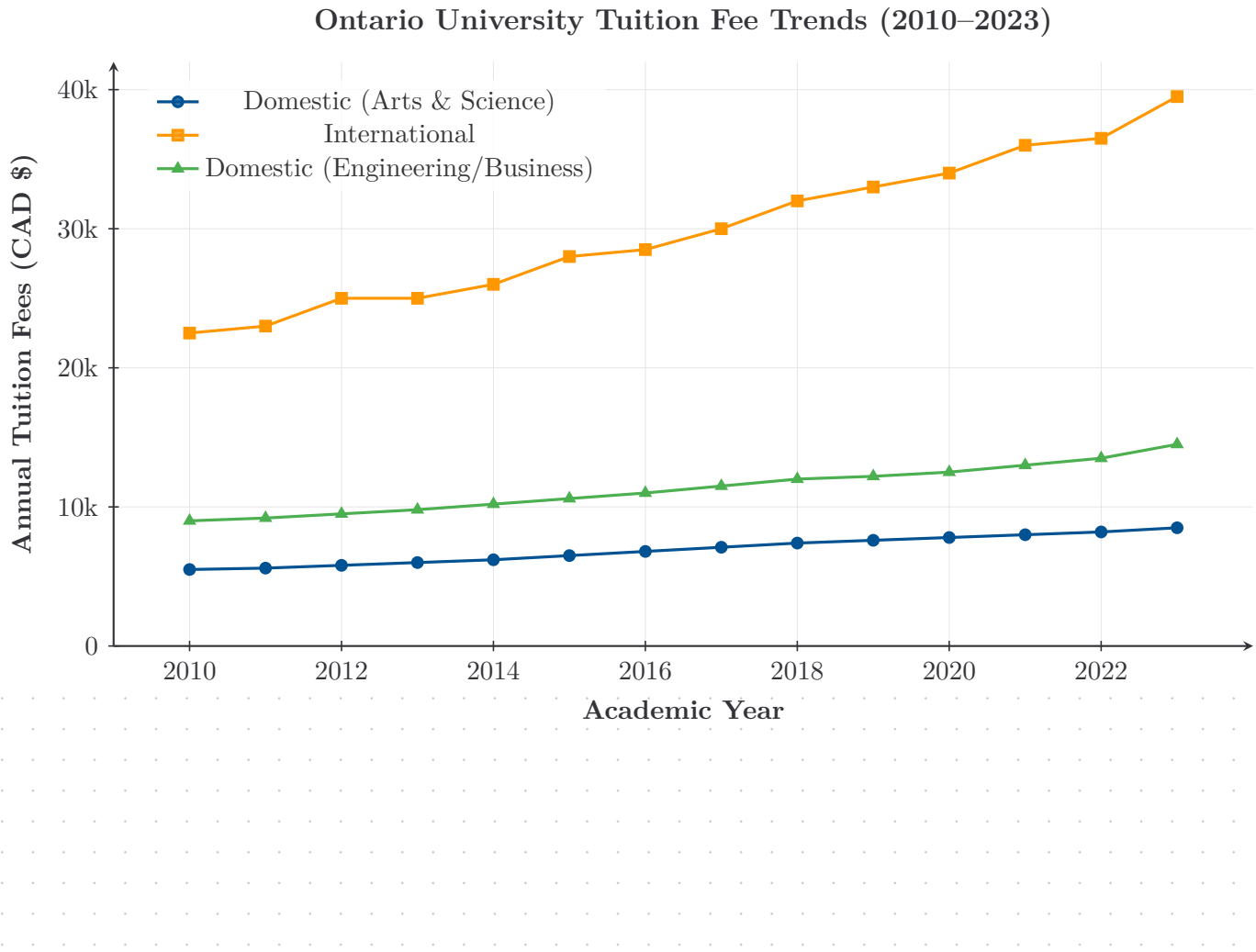
```
import matplotlib.pyplot as plt

# Step 1: Write up / read in the data
years = [2011, 2013, 2015, 2017, 2019, 2021,
 ↪  2023]
smartphone = [23, 47, 73, 84, 91, 93, 95]
depression = [8, 11, 13, 13, 16, 18, 17]

# Step 2: Create the plot
plt.figure(figsize=(8, 5))

# Step 3: Plot both lines
plt.plot(years, smartphone, marker='o',
  label='Smartphone Access (%)')
plt.plot(years, depression, marker='s',
  label='Depression Rate (%)')

# Step 4: Add labels and title
plt.xlabel('Year')
plt.ylabel('Percentage')
plt.title('U.S. Teens: Smartphone Access vs.
 ↪  Depression Rate')
plt.legend()

# Step 5: Save and display
plt.savefig('time_series_simple.png',
  dpi=300, bbox_inches='tight')
plt.show()
```



Note that time series plots always have time on the $x-$axis and the quantitative variable of interest on the $y-$axis.

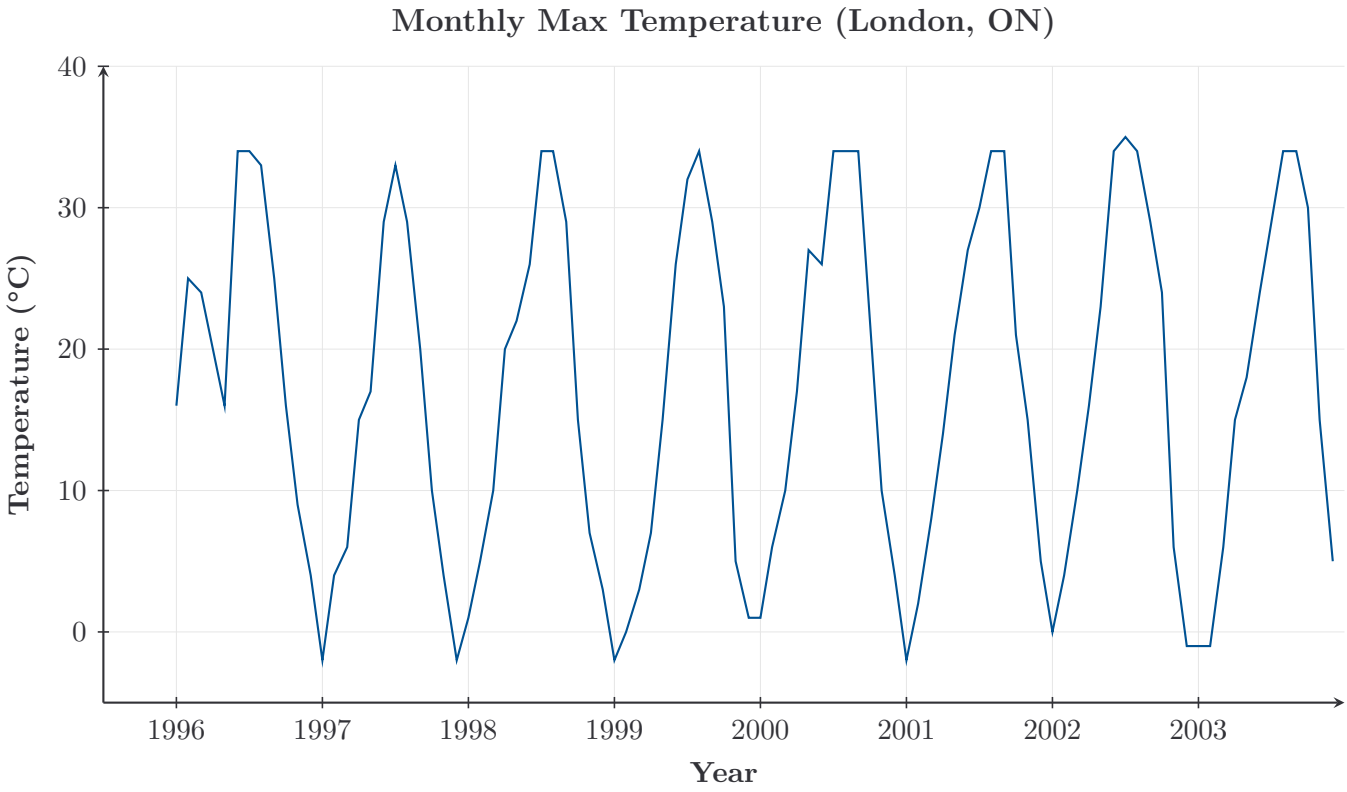
A **trend** in a time series is a pattern indicating a tendency for the variable to either rise or fall over time.

Example 1.29 *Ontario University Tuition Trends*
The following time plot shows how tuition fees have changed at Ontario universities from 2010 to 2023. What does the plot reveal about changes in tuition fees over time for domestic and international students? Compare the magnitude of these changes.



The **seasonality** of a time series refers to regular, repeating patterns or cycles that occur at regular intervals over time.

Example 1.30 *Seasonal Temperature Patterns*
Monthly maximum temperatures show clear seasonal patterns.



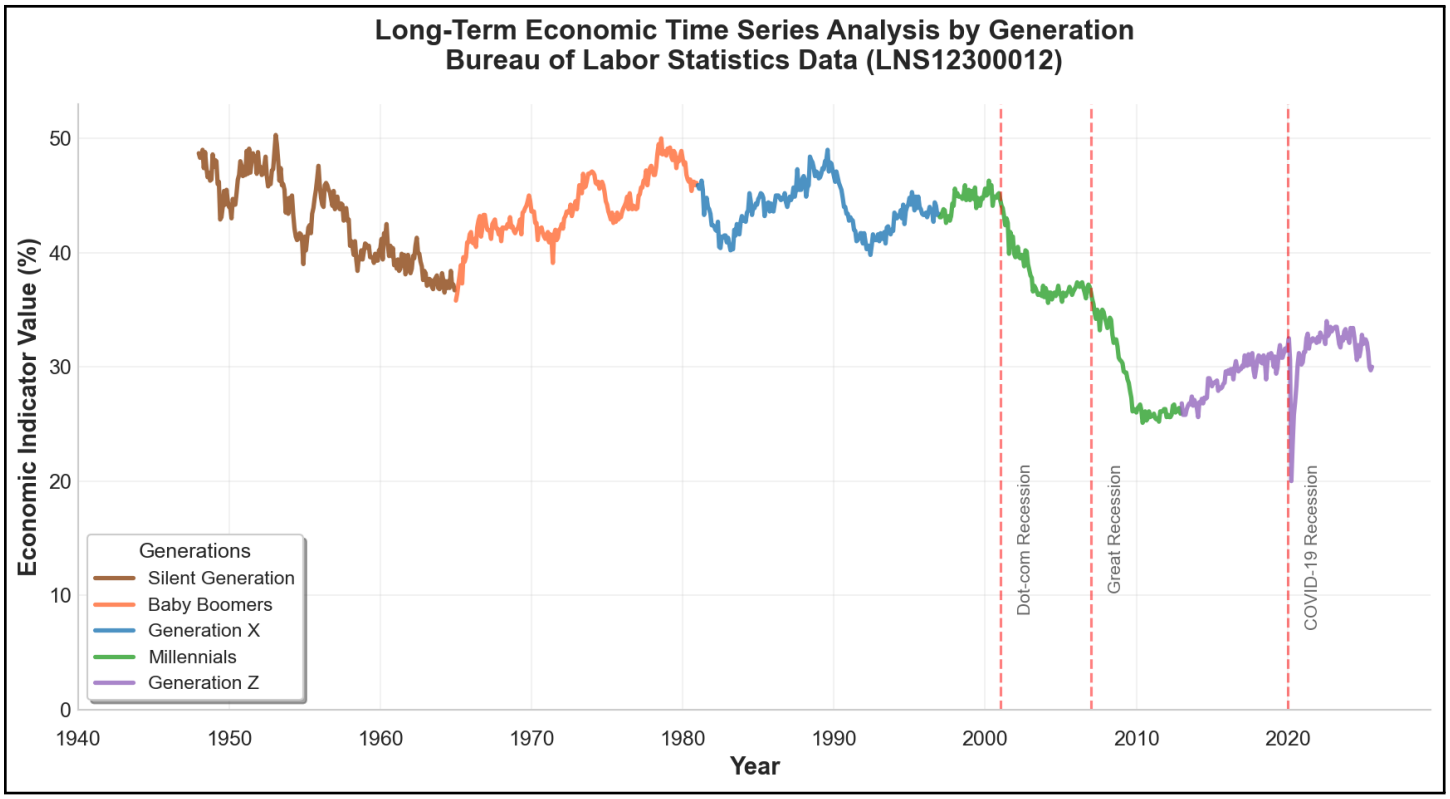
Interpretation: The temperature data shows a clear **seasonal pattern**: temperatures peak in July/August around 34°C and reach their lowest in January around 0°C or below. This cycle repeats every year with consistency.

Much like in histograms, we can also spot deviations from expected patterns in time plots, such as unusual spikes or drops.

Example 1.31

The following time plot shows the employment rate of teenagers (ages 15–19) in the United States from 1948 to 2022.

The red dashed lines indicate points in time where major economic events occurred.



Interpretation: Clear sharp deviations from the overall pattern at each of the marked economic events. For example, the 2008 financial crisis caused a significant drop in teen employment rates, which only partially recovered in subsequent years. The COVID-19 pandemic in 2020 led to an even more pronounced decline, with rates plummeting to historic lows before beginning to recover in 2021 and 2022.

Key point: When examining time plots, look for:

- **Trends:** Long-term upward or downward movement
- **Seasonality:** Regular cycles that repeat (e.g. weekly, monthly, yearly)
- **Deviations:** Sudden changes that deviate from the pattern

1.7. Chapter Summary

Chapter 1: Picturing Distributions with Graphs

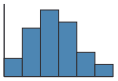
Key Concepts

- **Individuals:** Objects described by data
- **Variables:** Characteristics that vary across individuals
- **Categorical:** Groups/categories
- **Quantitative:** Numerical measurements
- **Distribution:** Pattern of values and their frequencies

Describing Distributions

- **Shape:** Symmetric, left-skewed, or right-skewed
- **Centre:** Typical or representative value
- **Spread:** How much values vary
- **Outliers:** Values that deviate from the pattern

Visualisation Tools Summary

Categorical Data (Qualitative groups)		Quantitative Data (Numerical values)		
Pie Chart Parts of a whole	**Bar Graph** Compare counts	 **Histogram** Shape & Spread	1 | 2 2 | 3 5 3 | 1 8 **Stemplot** Exact values	**Time Plot** Trends over time

1.8. Additional exercises

Exercise 1.32 *Identifying Variables*
A university conducts a survey of first-year students. For each variable below, state whether it is categorical or quantitative:

(a) Student ID number

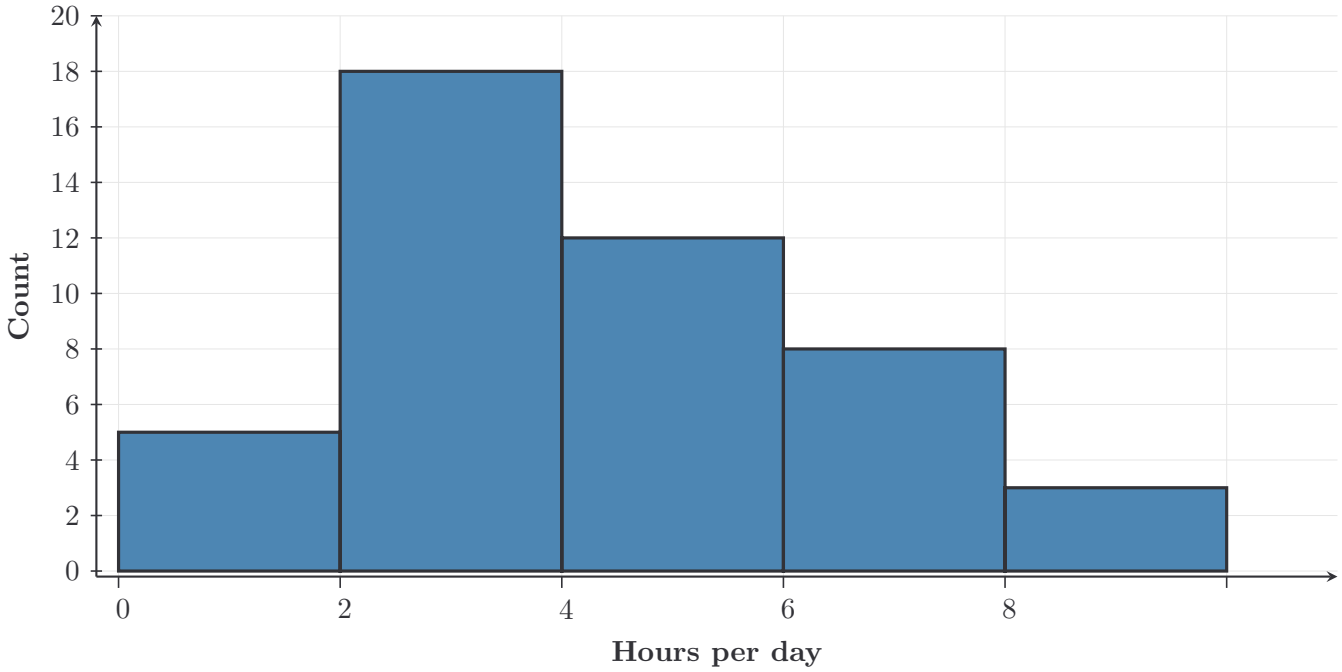
(b) Number of courses enrolled

(c) Faculty (e.g., Arts, Science, Engineering)

(d) Distance from home to campus (in km)

(e) Favourite streaming service

Exercise 1.33 *Interpreting a Histogram*
The histogram below shows the distribution of hours spent on social media per day for a sample of teenagers.



(a) Describe the shape of the distribution.

(b) How many teenagers were surveyed?

(c) What percentage of teenagers spend fewer than 4 hours per day on social media?

Exercise 1.34 *Pie Chart Appropriateness*
A survey asks respondents: "Which of these activities do you enjoy? (Select all that apply): Reading, Gaming, Sports, Music, Cooking."

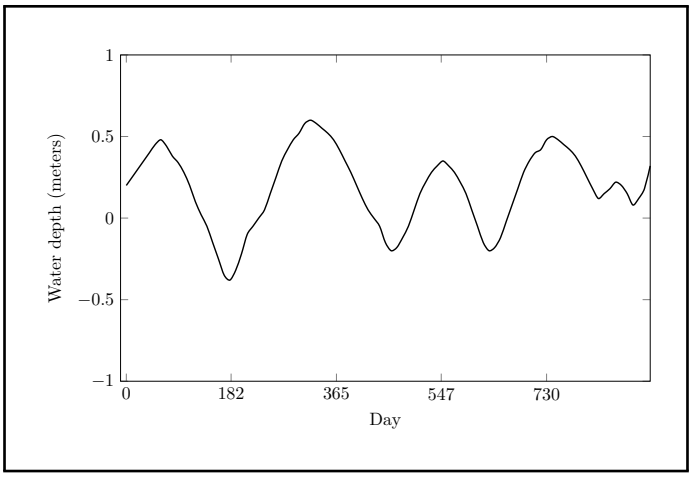
Can the results be displayed in a pie chart? Explain why or why not.

Exercise 1.35 *Creating a Stemplot*
Create a stemplot for the following dataset representing the ages of participants in a fitness class:

23 25 27 31 33 35 36 38 41 42 44 45 47 52 55

Then describe the shape of the distribution and comment on the presence of any outliers.

Exercise 1.36 *Time series interpretation*
Comment on each of the time series features we examined in the context of the following plot:



Additional practice: Odd problems in the text, 1.13 to 1.45 (except 1.29)