Introduction to Data Science

Dr. Irfan Yousuf
Department of Computer Science (New Campus)
UET, Lahore

(Week 7; February 26 – March 01, 2024)

Outline

- Multivariate Data Analysis

Multivariate Data Analysis

- Multivariate analysis is used to explore more than two variables at once.

- Correlating more than two variables at a time.

Multivariate Data Analysis

- **Dependence techniques**
  - Dependence methods are used when one or some of the variables are dependent on others. Dependence looks at cause and effect.

- **Interdependence techniques**
  - Interdependence methods are used to understand the structural makeup and underlying patterns within a dataset.

  - Interdependence methods seek to give meaning to a set of variables or to group them together in meaningful ways.

Multivariate Data Analysis

- **Factor analysis** is an interdependence technique which seeks to reduce the number of variables in a dataset.
- Factor analysis works by detecting sets of variables which correlate highly with each other.
- Example 1: Let's imagine you have a dataset containing data pertaining to a person's income, education level, and occupation. You might find a high degree of correlation among each of these variables, and thus reduce them to the single factor "socioeconomic status."
- Example 2: You might also have data on how happy they were with customer service, how much they like a certain product, and how likely they are to recommend the product to a friend. Each of these variables could be grouped into the single factor "customer satisfaction"

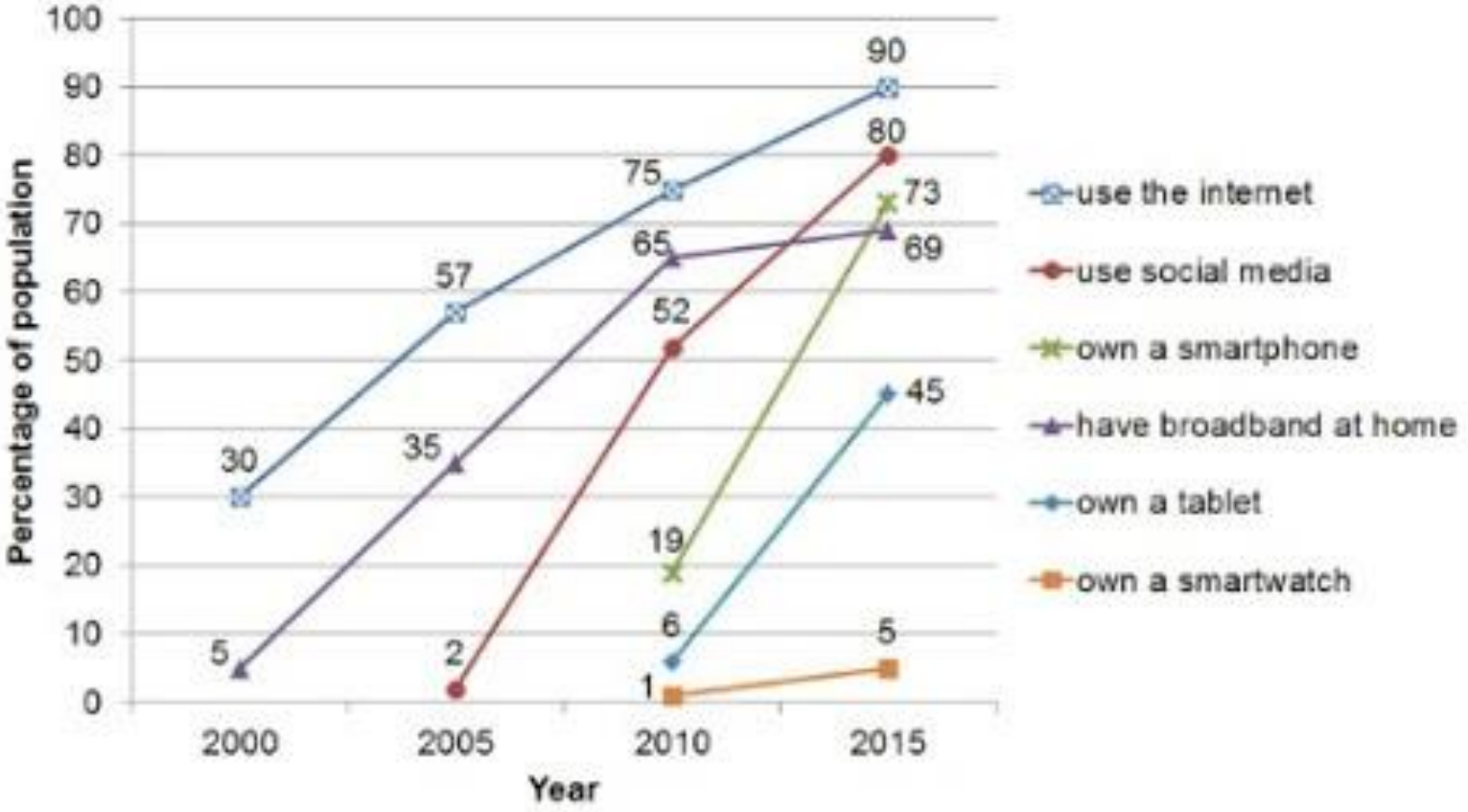
Multivariate Data Analysis

- Another interdependence technique, **cluster analysis** is used to group similar items within a dataset into clusters.
- When grouping data into clusters, the aim is for the variables in one cluster to be more similar to each other than they are to variables in other clusters.
- Example 1: A prime example of cluster analysis is audience segmentation. If you were working in marketing, you might use cluster analysis to define different customer groups which could benefit from more targeted campaigns.
- Example 2: As a healthcare analyst, you might use cluster analysis to explore whether certain lifestyle factors or geographical locations are associated with higher or lower cases of certain illnesses.

Describing a Plot

Example Plot

The graph shows information about technology usage in the UK over time. Summarise the information by selecting and reporting the main features. Make comparisons where relevant.



Example Plot

The graph shows the rate at which British people adopted new technology over a 15-year period from 2000 to 2015. The figures are given as percentages of the population.

Overall, there was widespread adoption of new technology during these years. Nearly nine out of ten people in the UK were online by 2015. The figures for having broadband in the home, ownership of a smartphone and use of social media platforms were all high that year too, at around 70 to 80 per cent, and nearly half the population owned a tablet. The only exception to this is smartwatch ownership, which remained comparatively low at 5 per cent.

If we look at the trends over time, we can see that the uptake of new technology increased dramatically in this period. For example, internet usage tripled and social media usage grew strikingly by 78 percentage points. Smartphones and tablets appeared in 2010 and, similarly, these followed a steep upward trajectory. However, for some products, the graph shows that growth slowed down noticeably after an initial surge. Social media usage, for instance, was near zero in 2005 and shot up to 52 per cent in 2010, before climbing more slowly to 80 percent in 2015. Also, broadband subscriptions rose steadily by 30 percentage points every five years to 2010, but by a modest 4 percentage points after then. In contrast, the newer technologies such as tablets showed no sign of levelling off.

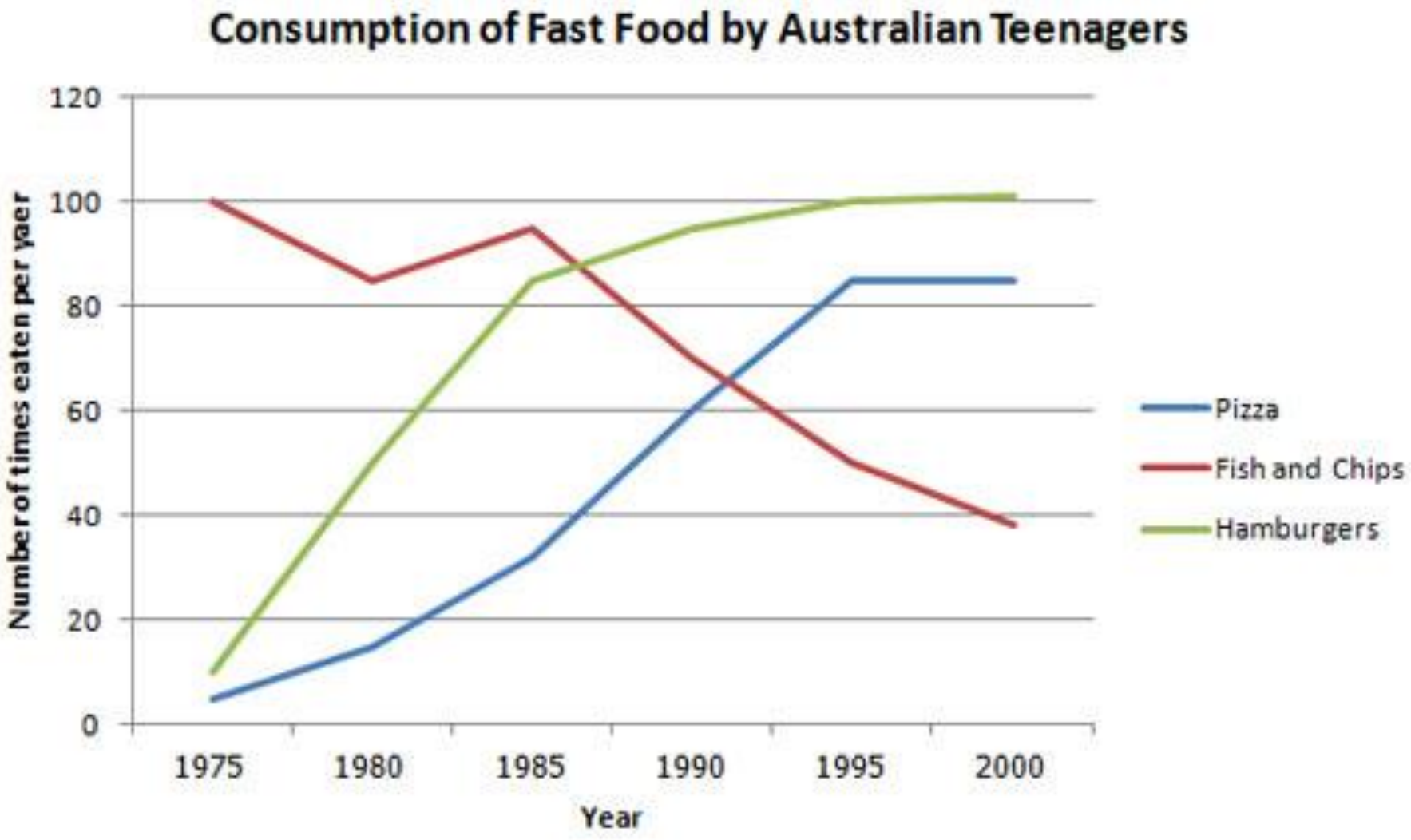
Ownership of all the technologies was increasing; it will be interesting to see when it peaks.

How to Describe a Graph

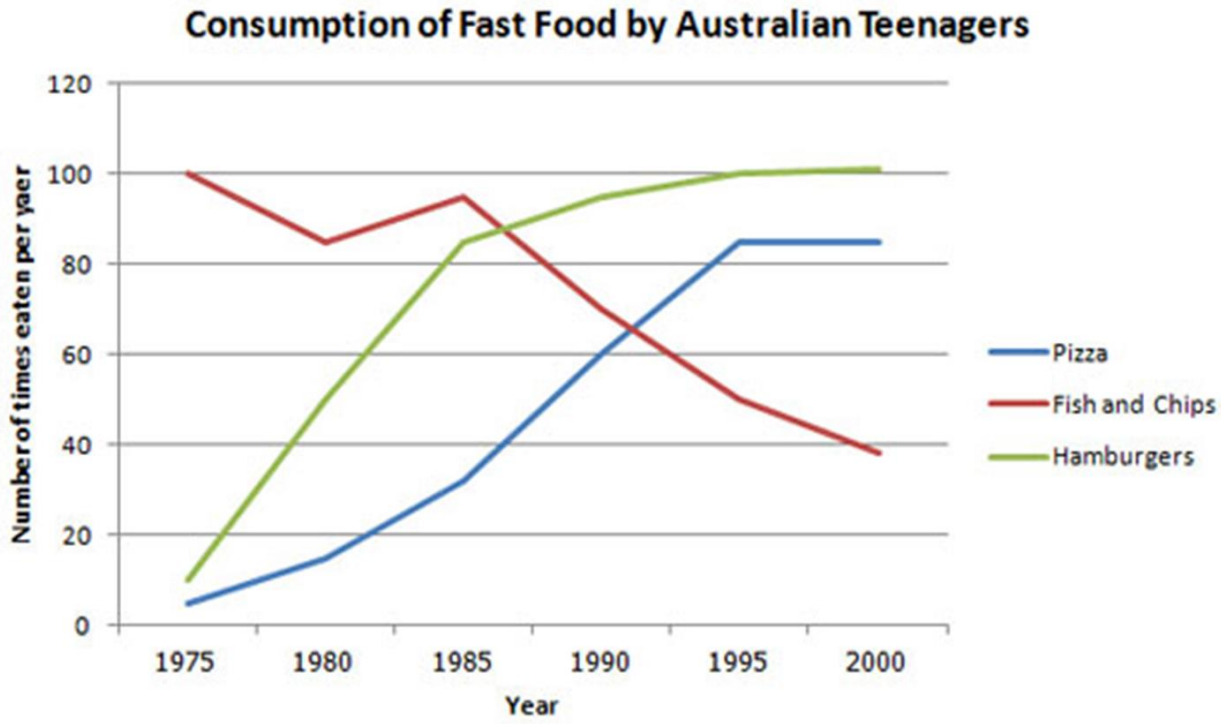
- Introduction
- Body
- Conclusion / Summary

- Introduction
- Overview / Summary
- Details

How to Describe a Graph

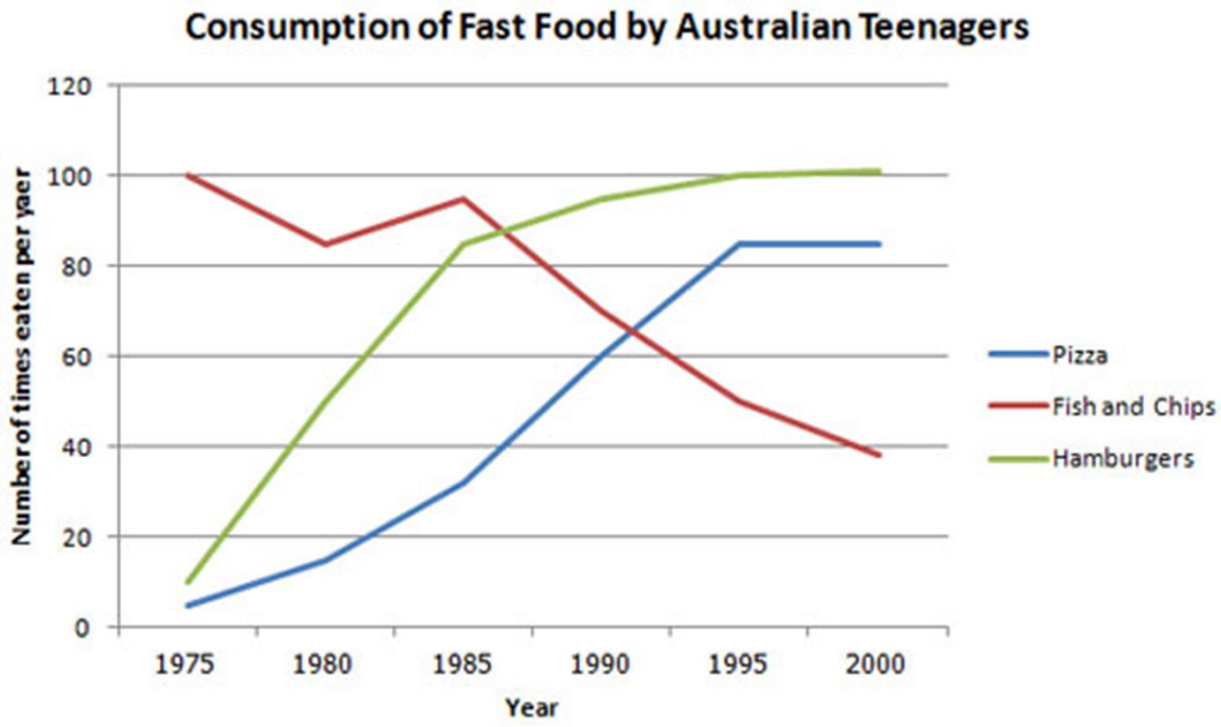


How to Describe a Graph



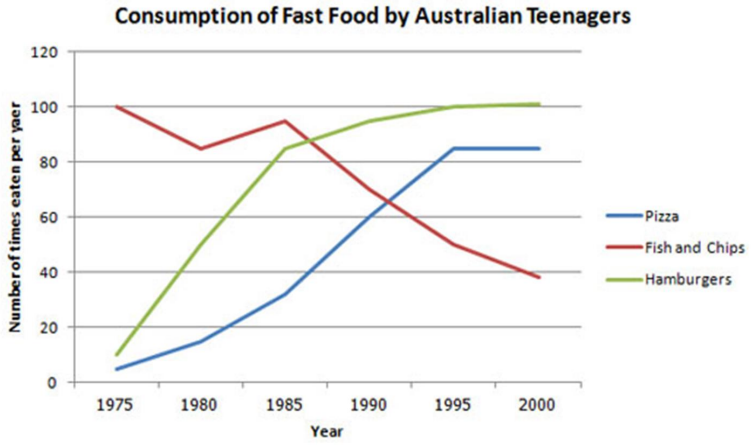
- The line graph compares the fast food consumption of teenagers in Australia between 1975 and 2000, a period of 25 years.

How to Describe a Graph



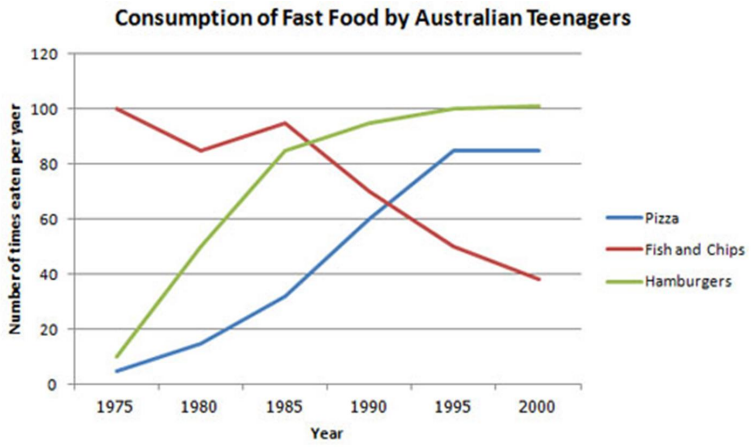
- Overall, the consumption of fish and chips declined over the period, whereas the amount of pizza and hamburgers that were eaten increased.

How to Describe a Graph



- In 1975, the most popular fast food with Australian teenagers was fish and chips, being eaten 100 times a year. This was far higher than Pizza and hamburgers, which were consumed approximately 5 times a year. However, apart from a brief rise again from 1980 to 1985, the consumption of fish and chips gradually declined over the 25 year timescale to finish at just under 40.

How to Describe a Graph



- In sharp contrast to this, teenagers ate the other two fast foods at much higher levels. Pizza consumption increased gradually until it overtook the consumption of fish and chips in 1990. It then levelled off from 1995 to 2000. The biggest rise was seen in hamburgers as the occasions they were eaten increased sharply throughout the 1970's and 1980's, exceeding that of fish and chips in 1985. It finished at the same level that fish and chips began, with consumption at 100 times a year.

How to Describe a Graph

The line graph compares the fast food consumption of teenagers in Australia between 1975 and 2000, a period of 25 years. Overall, the consumption of fish and chips declined over the period, whereas the amount of pizza and hamburgers that were eaten increased.

In 1975, the most popular fast food with Australian teenagers was fish and chips, being eaten 100 times a year. This was far higher than Pizza and hamburgers, which were consumed approximately 5 times a year. However, apart from a brief rise again from 1980 to 1985, the consumption of fish and chips gradually declined over the 25 year timescale to finish at just under 40.

In sharp contrast to this, teenagers ate the other two fast foods at much higher levels. Pizza consumption increased gradually until it overtook the consumption of fish and chips in 1990. It then levelled off from 1995 to 2000. The biggest rise was seen in hamburgers as the occasions they were eaten increased sharply throughout the 1970's and 1980's, exceeding that of fish and chips in 1985. It finished at the same level that fish and chips began, with consumption at 100 times a year.

Line Graph

- A line chart graphically displays data that changes continuously over time.
- Each line graph consists of points that connect data to show a trend (continuous change).
- Line graphs have an x-axis and a y-axis. In the most cases, time is distributed on the horizontal axis.

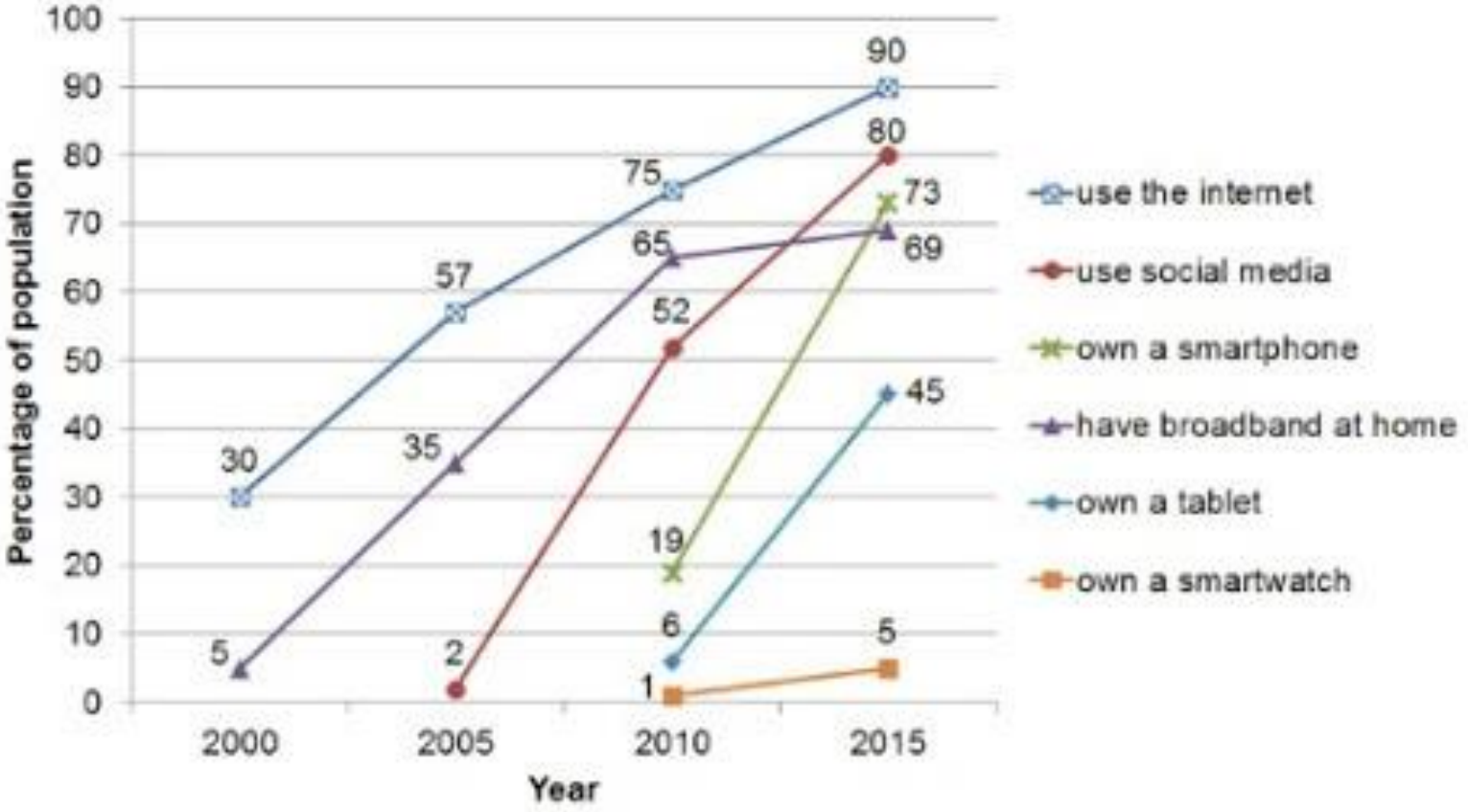
- When you want to show trends. For example, how house prices have increased over time.
- When comparing two or more different variables, situations, and information over a given period of time.

Line Graph

- These lines show movement over time affected by the increase or decrease in the key factors.
- To express the movement of the line, you should use appropriate verbs, adjectives, and adverbs depending on the kind of action you need to show.
- **Verbs:** rise, increase, grow, go up to, climb, boom, peak, fall, decline, decrease, drop, dip, go down, reduce, level up, remain stable.
- **Adjectives:** sharp, rapid, huge, dramatic, substantial, considerable, significant, slight, small, minimal, massive.
- **Adverbs:** dramatically, rapidly, hugely, massive, sharply, steeply, considerably, substantially, significantly, slightly, minimally, markedly.

Example Plot

The graph shows information about technology usage in the UK over time. Summarise the information by selecting and reporting the main features. Make comparisons where relevant.



Example Plot

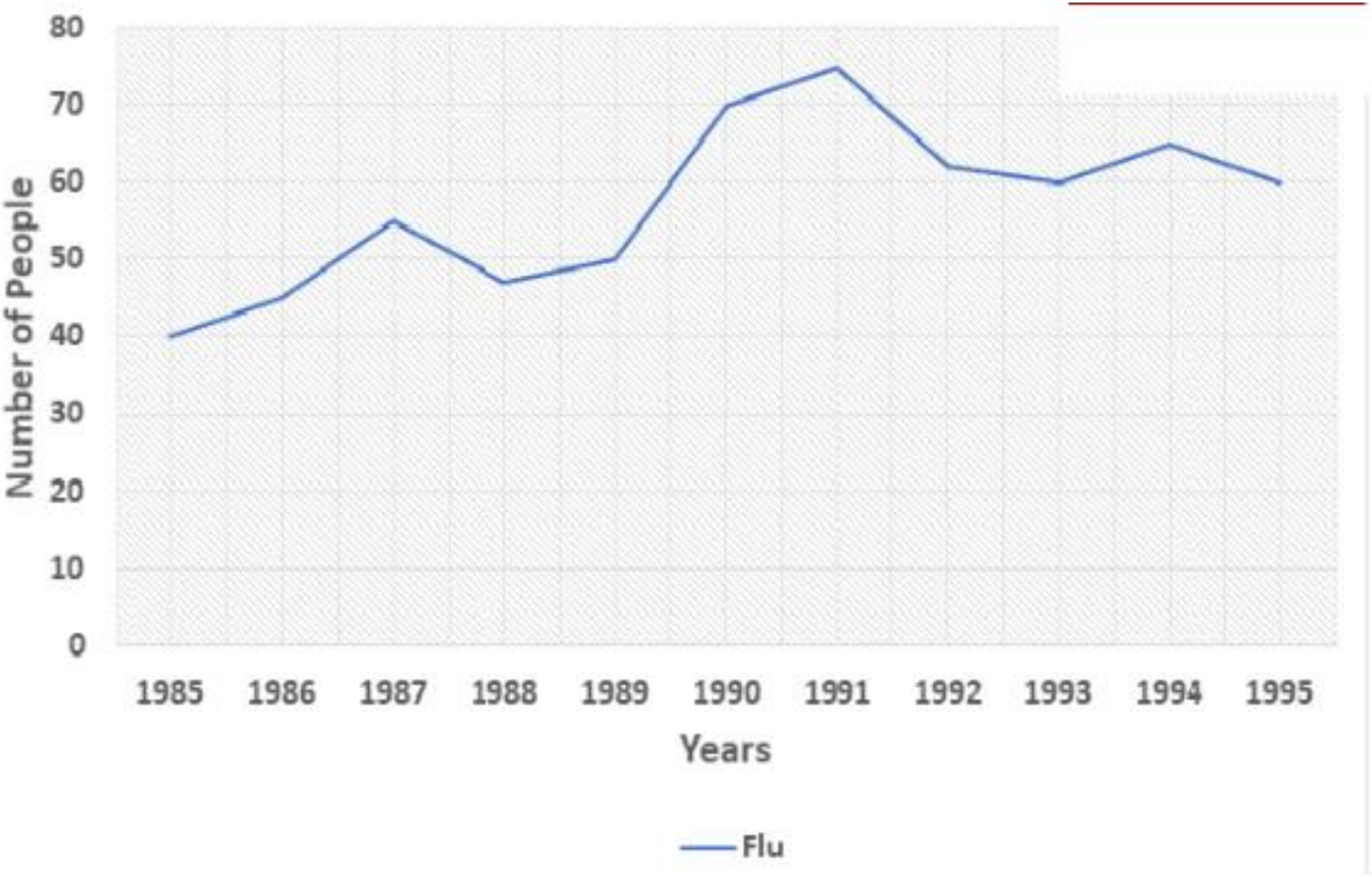
The graph shows the rate at which British people adopted new technology over a 15-year period from 2000 to 2015. The figures are given as percentages of the population.

Overall, there was widespread adoption of new technology during these years. Nearly nine out of ten people in the UK were online by 2015. The figures for having broadband in the home, ownership of a smartphone and use of social media platforms were all high that year too, at around 70 to 80 per cent, and nearly half the population owned a tablet. The only exception to this is smartwatch ownership, which remained comparatively low at 5 per cent.

If we look at the trends over time, we can see that the uptake of new technology increased dramatically in this period. For example, internet usage tripled and social media usage grew strikingly by 78 percentage points. Smartphones and tablets appeared in 2010 and, similarly, these followed a steep upward trajectory. However, for some products, the graph shows that growth slowed down noticeably after an initial surge. Social media usage, for instance, was near zero in 2005 and shot up to 52 per cent in 2010, before climbing more slowly to 80 percent in 2015. Also, broadband subscriptions rose steadily by 30 percentage points every five years to 2010, but by a modest 4 percentage points after then. In contrast, the newer technologies such as tablets showed no sign of levelling off.

Ownership of all the technologies was increasing; it will be interesting to see when it peaks.

The graph below shows the reported number of cases of influenza in people over 65 in a certain village in the UK from 1985 to 1995.



Sample Answer 1

The line chart illustrates the number of cases of influenza among people above 65 years in the countryside in the United Kingdom between 1985 to 1995. Units measured in years and number of people.

Looking from the overall perspective, during the decade, the graph fluctuated. In the beginning year, it had the least amount of cases and after 6 years it was at its peak.

In 1985, 40 people were infected and then it gradually increased till 1987 with nearly 55 persons. For the next 2 years, 5 cases were decreased compared to 1987. For the next couple of years, it was rapidly raised to one-quarter of humans and it was the highest number of cases registered in a decade.

From 1991 to 1993, cases dropped by approximately 15. For the last 2 years, there were ups and downs with 10 cases in 1994 and 1995 respectively and ending with 60 cases in the year 1995.

Sample Answer 2

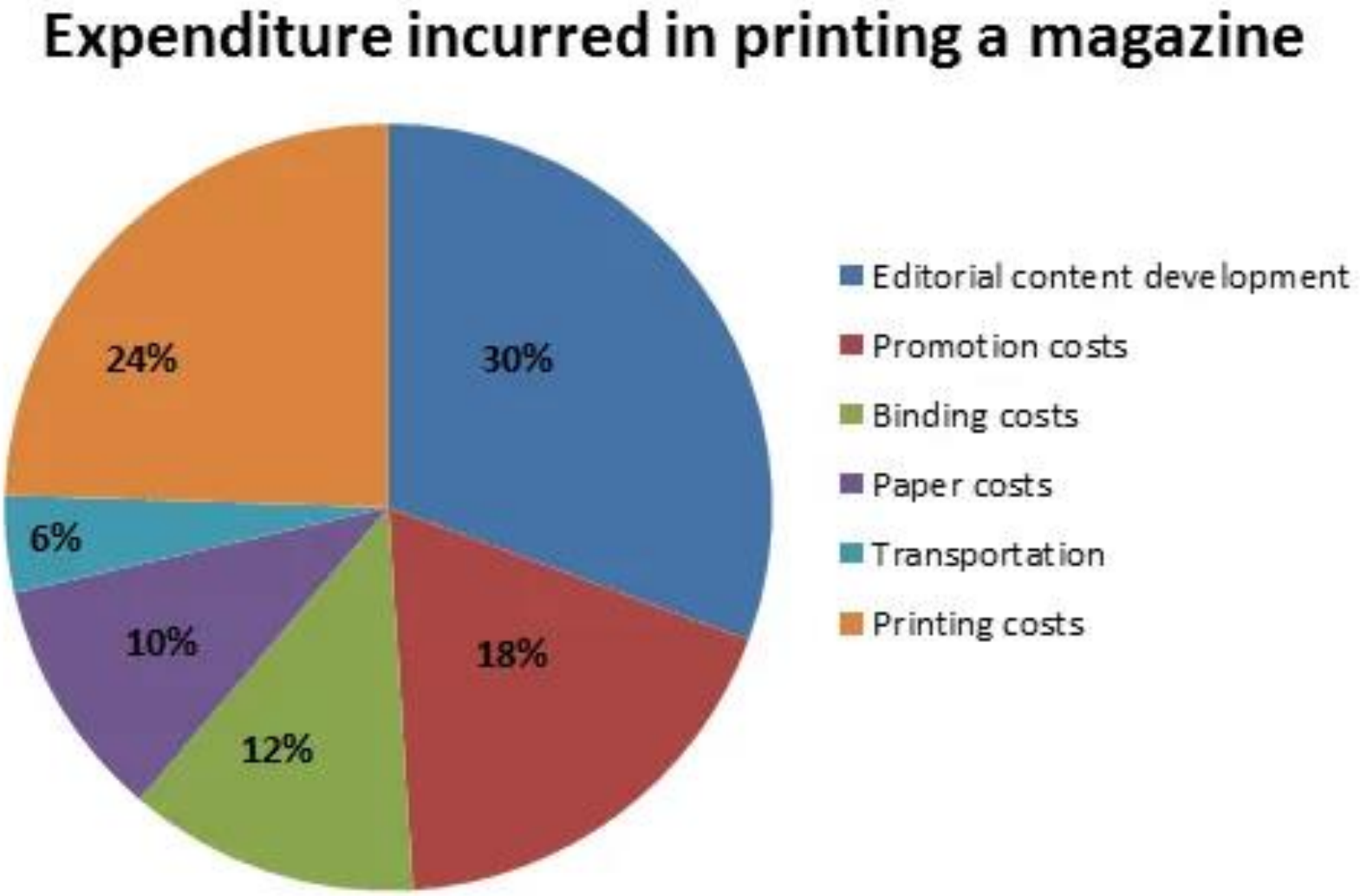
The graph illustrates the number of people over 65 who were infected by influenza, in a particular village. Th data described the survey period start from 1985 to till the end of 1995.

Overall, the graph shows how, within a decade many people are suffered by flue in every years.

In the beginning of the year 1985, total 40 people have infected by the flu. The disease was steeply increased to 55 in 1987. In 1988 the virous spread rate slightly dope to 48 but not last long till 1989, the flu hit to 50 people. Since than, the virous spread rate sharply increase to 70. The report from the graph, in the 1991, 75 people have been infected by influenza which is the highest cause rate in this decade.(1985 to 1995)

The effected rate was gradually drop to 62 in 1992. The following year of 1993 to end of 1995, the number of people are slightly up and down. The flu rate was end up as over 65 people are suffered in this ten years.

Pie Chart


Expenditure incurred in printing a magazine
30%
18%
12%
10%
6%
24%
Editorial content development
Promotion costs
Binding costs
Paper costs
Transportation
Printing costs

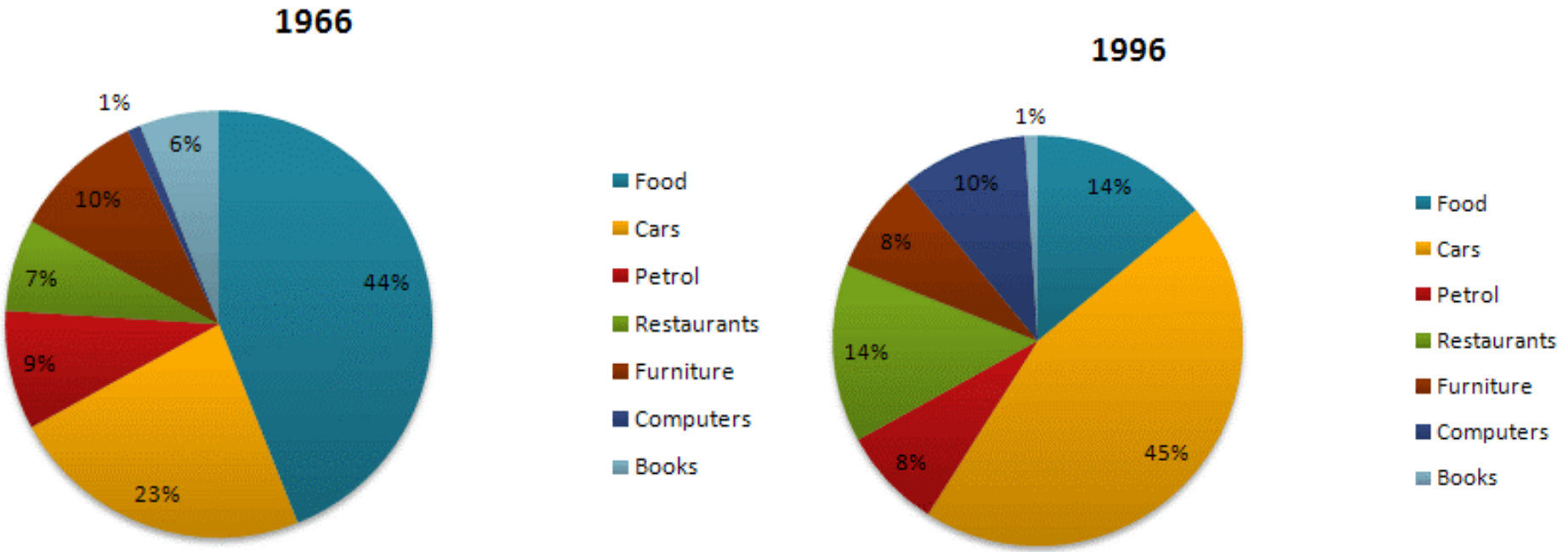
The given pie chart shows the different expenses involved in printing a magazine. Six types of expenses are shown in the diagram and they are: editorial content development, printing, promotion, paper, binding and transportation. While editorial content development is the biggest expense, transportation is the smallest.

When we study the graph, it is not hard to see that the costs associated with content development and printing are the biggest expenses incurred in publishing a magazine. While 30 percent of the total cost goes towards developing editorial content, printing accounts for 24 percent of the total expenditure. Together, they account for more than half of the total expenses a publisher has to incur to bring out a magazine. Promotion costs, too, are significant. At 18 percent, the cost involved in promoting the magazine is the third biggest.

Paper costs are surprisingly low. Only 10 percent of the total expenditure goes towards buying the paper. At 12 percent, binding costs aren't quite significant. Of all the expenses incurred in printing a magazine, the costs involved in transporting it is the smallest.

The given pie charts compare the expenses in 7 different categories in 1966 and 1996 by American Citizens.

Write a report to describe the information below.



The pie charts compare the expenditure of US residents in two different years in seven categories namely food, cars, petrol, restaurants, furniture, computers and books.

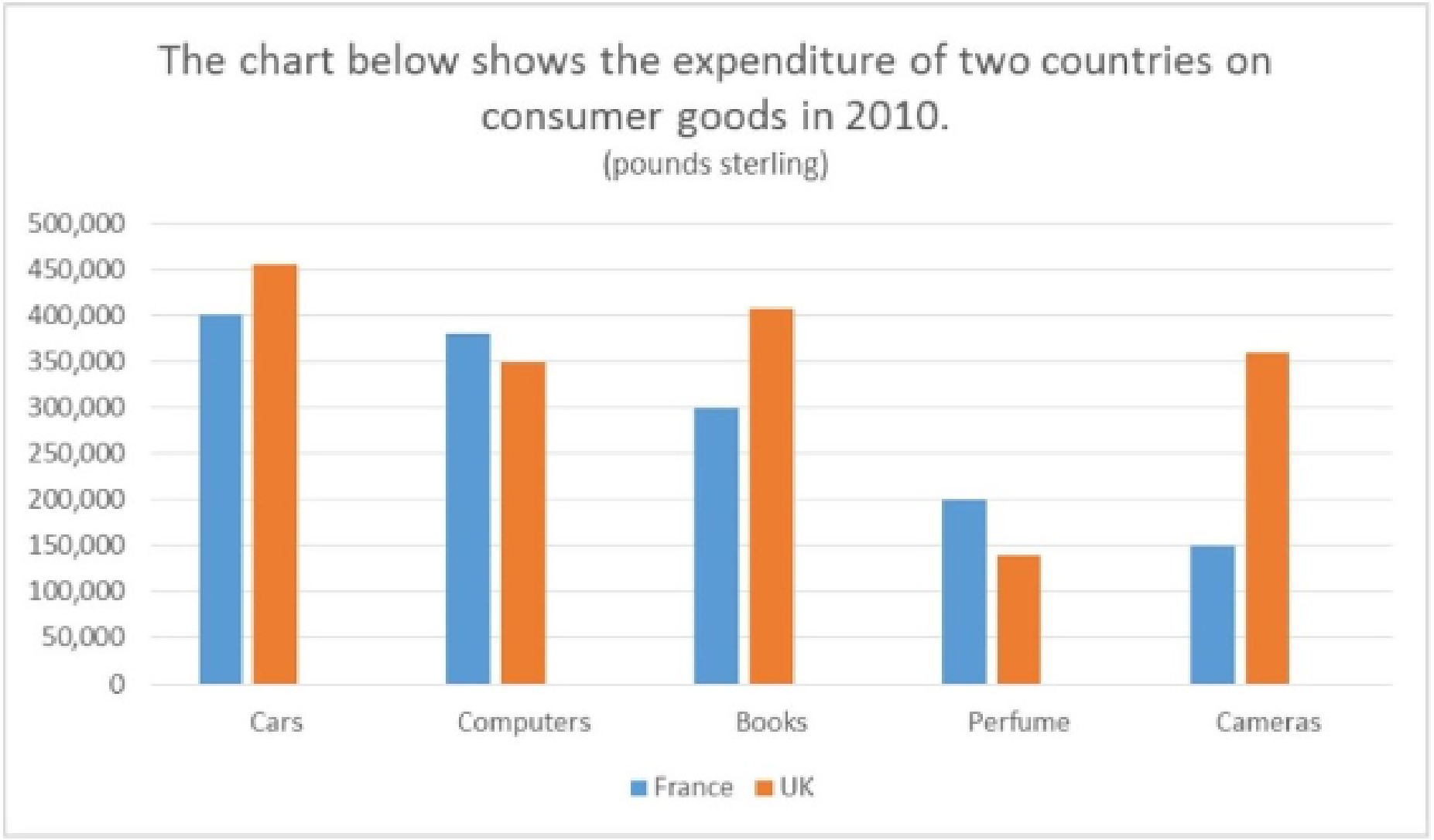
It is clear that the largest proportion of American citizens' spending went on foods and cars. On the other hand, computers and books have the lowest percentage in the chart in 1966 and 1996 respectively. In 1966, 23% of American citizens' expenditure went on cars. The percentage rose to nearly double at 45% in 1996.The proportion of spending on food fell from 44% in 1966 to only 14% in 1996.

Expenditure on computers stood at only 1% in 1966 but reached 10% in 1996. The percentage of American citizens spending on restaurants had doubled from 7% in 1966 to 14% in 1996. Spending on books was highest in 1966, at 6%. By contrast, there was no significant change in the proportions of petrol and furniture over a period as a whole.

Bar Plot

- A Bar Graph (or Bar Chart) represents categorical data with comparison. A Bar Graph can be horizontal or vertical while plotting. In general, you find rectangular bars with lengths or heights.

Bar Plot


The chart below shows the expenditure of two countries on consumer goods in 2010.
(pounds sterling)
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Cars
Computers
Books
Perfume
Cameras
France
UK

Sample Answer 1

The chart illustrates the amount of money spent on five consumer goods (cars, computers, books, perfume and cameras) in France and the UK in 2010. Units are measured in pounds sterling.

Overall, the UK spent more money on consumer goods than France in the period given. Both the British and the French spent most of their money on cars whereas the least amount of money was spent on perfume in the UK compared to cameras in France. Furthermore, the most significant difference in expenditure between the two countries was on cameras.

In terms of cars, people in the UK spent about £450,000 on this as opposed to the French at £400,000. Similarly, the British expenditure was higher on books than the French (around £400,000 and £300,000 respectively). In the UK, expenditure on cameras (just over £350,000) was over double that of France, which was only £150,000.

On the other hand, the amount of money paid out on the remaining goods was higher in France. Above £350,000 was spent by the French on computers which was slightly more than the British who spent exactly £350,000. Neither of the countries spent much on perfume which accounted for £200,000 of expenditure in France but under £150,000 in the UK.

Sample Answer 2

The given bar demonstrates the difference between the expenditure of France an the United Kingdom on purchasing goods in 2010.

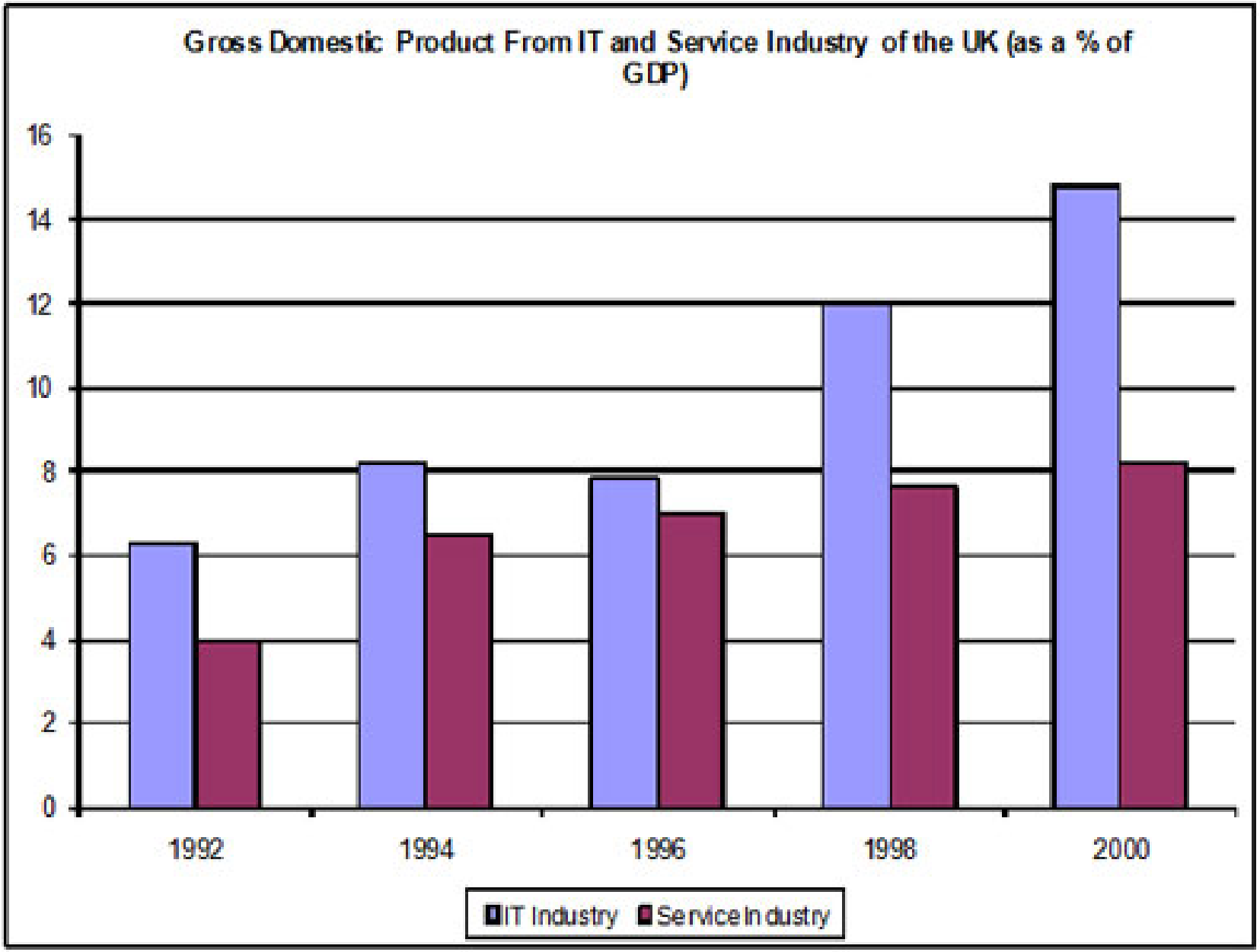
Overall, the French ranked first in types of goods, while the British had a share in three.

The highest figure for the UK and France were reported in expenditure of cars and computers, respectively.

The was a significant differnce between the amount of money spent on cameras in the UK and France , the former standing first with over 350000 poud sterling , while the latter was 150000. Consumers in the UK had also a highest share in expenditure of cars and books , while the former was 450000 and the latter was 400000.

The disparity between the expenditure of compuers was almost the same in two countries. France standing first with nearly 400000 and the share of the UK was low by a narrow margin 350000.

The lowest share for France and the UK in buying perfume, while the former spent 200000 and the latter spent less than 150000.


Gross Domestic Product From IT and Service Industry of the UK (as a % of GDP)
16
14
12
10
8
6
4
2
0
1992
1994
1996
1998
2000
IT Industry
Service Industry

The bar chart illustrates the gross domestic product generated from the IT and Service Industry in the UK from 1992 to 2000. It is measured in percentages. Overall, it can be seen that both increased as a percentage of GDP, but IT remained at a higher rate throughout this time.

At the beginning of the period, in 1992, the Service Industry accounted for 4 per cent of GDP, whereas IT exceeded this, at just over 6 per cent. Over the next four years, the levels became more similar, with both components standing between 6 and just over 8 per cent. IT was still higher overall, though it dropped slightly from 1994 to 1996.

However, over the following four years, the patterns of the two components were noticeably different. The percentage of GDP from IT increased quite sharply to 12 in 1998 and then nearly 15 in 2000, while the Service Industry stayed nearly the same, increasing to only 8 per cent.

At the end of the period, the percentage of GDP from IT was almost twice that of the Service Industry.

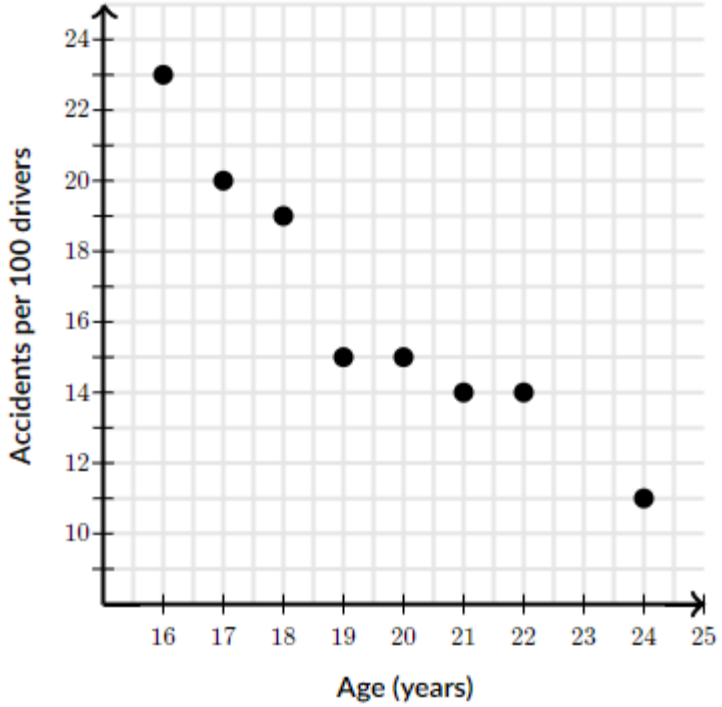
Scatter Plot

- When trying to find out whether there is a relationship between 2 variables.
- To predict the behavior of dependent variable based on the measure of the independent variable.
- When having paired numerical data.
- When you just want to visualize the correlation between 2 large datasets without regard to time.

Scatter Plot

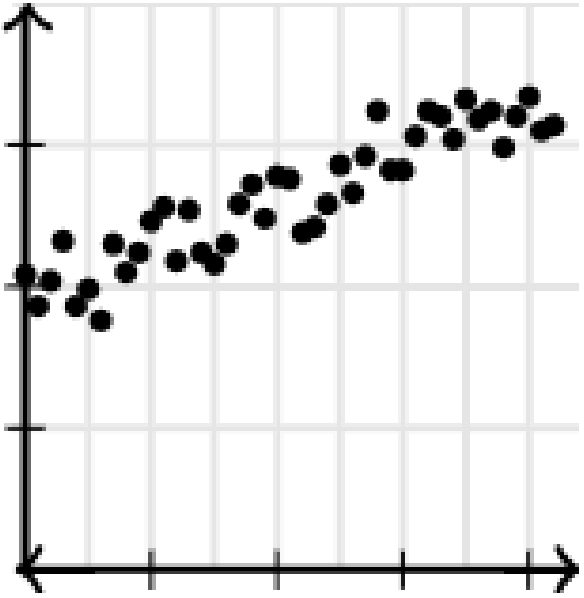
- A quick description of the association in a scatterplot should always include a description of the form, direction, and strength of the association, along with the presence of any outliers.
- **Form:** Is the association linear or nonlinear?
- **Direction:** Is the association positive or negative?
- **Strength:** Does the association appear to be strong, moderately strong, or weak?
- **Outliers:** Do there appear to be any data points that are unusually far away from the general pattern?

Scatter Plot



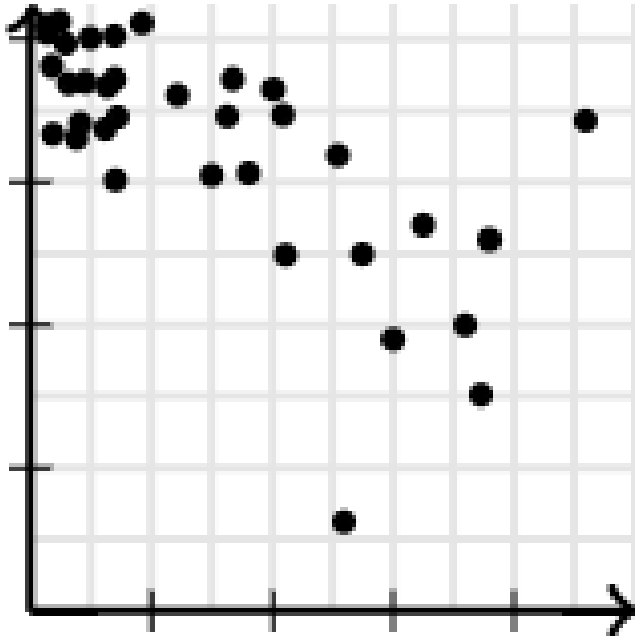
This scatterplot shows a strong, negative, linear association between age of drivers and number of accidents. There don't appear to be any outliers in the data

Scatter Plot



There is a strong, positive, linear association between the two variables

Scatter Plot



There is a moderately strong, negative, linear association between the two variables with a few potential outliers.

Summary

Multivariate Data Analysis
Describing a Plot