Statistics Essentials for Data Science

Introduction to Data Visualization

Learning Objectives

By the end of this lesson, you will be able to:

- Explain the commonly used visualization and interpret them
- Understand how to select the appropriate visualization
- List the dos and don'ts of data preprocessing
- Provide insights into data using a visualization

Business Scenario

ABC is an organization that runs a social media platform. The organization wants to present the platform data to its senior management and clients. The presentation is supposed to provide insights into the users and their likes.

The organization decides to present the data with charts that show the required data and convey the meaning.

To do so, it must explore different data visualization tools and techniques and create the presentation using these.

Data Visualization

Discussion

Discussion

Duration: 15 minutes

Does visualizing data help in any way?

- Why do we need to visualize the data?
- What are the tools to visualize the data?

Data Visualization

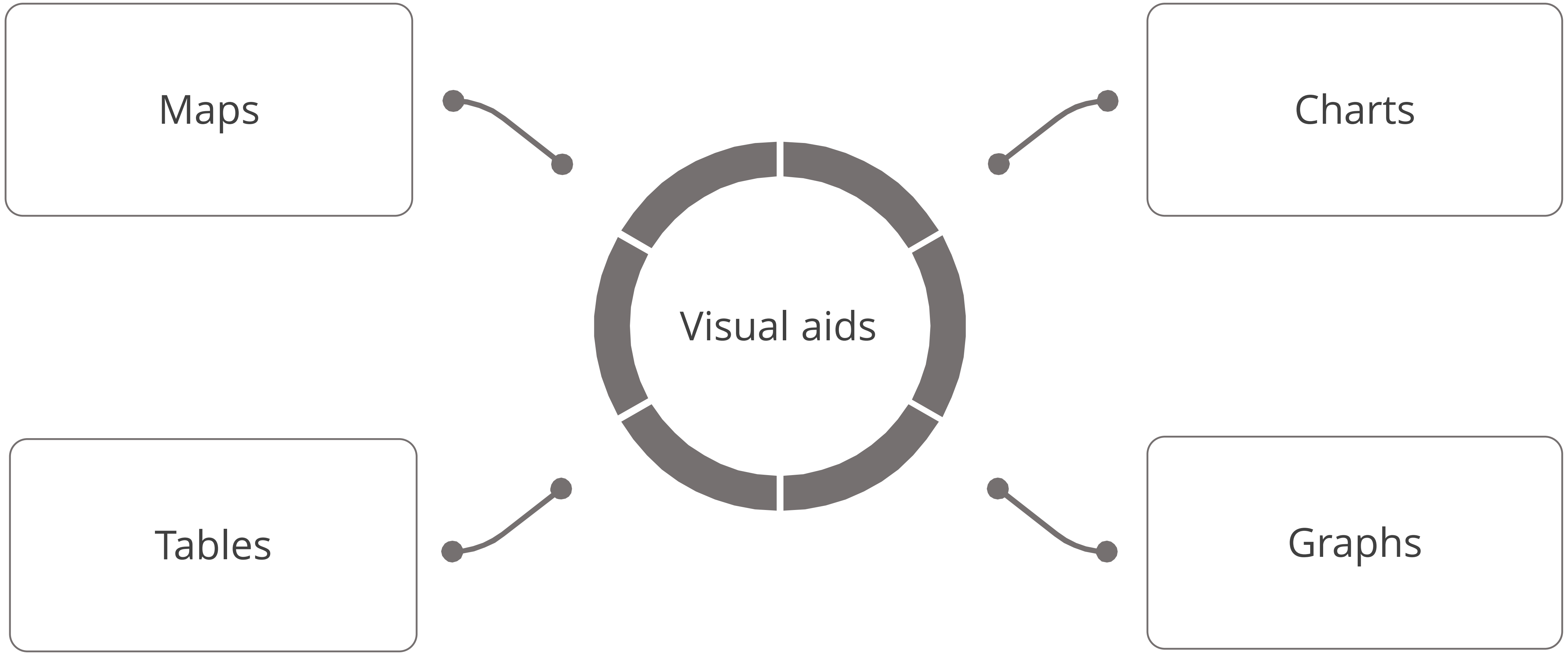
The graphical representation of information and data is known as data visualization.

These visualizations help users draw insights.

Pictorial or visual aids facilitate better communication and assimilation of information.

Data Visualization

It is the pictorial representation of data using:



Data Visualization Tools

They provide an easy way to understand datasets in terms of data patterns, such as:

Trends

Presence of outliers

These tools are helpful when information is included in executive reports.

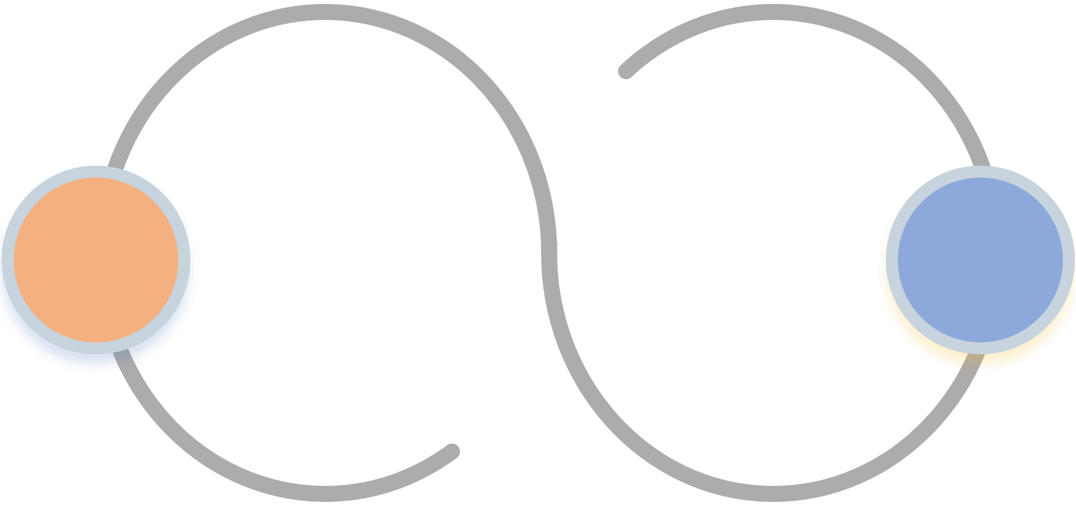
Data Visualization Tools

These tools help reinforce descriptive information and enhance assimilation by readers.

They are also useful in executive presentations.

Data Visualization Tools

These tools are useful in communicating data in media reports.



Newspapers

Magazines

Information technology advancements have also increased the volume and variety of data.

Discussion

Duration: 15 minutes

Does visualizing data help in any way?

- Why do we need to visualize the data?

Answer: Pictorial or visual aids facilitate better communication and assimilation of information.

- What are the tools to visualize the data?

Answer: Charts, graphs, tables, and maps are a few tools to visualize the data. They capture and readily draw the attention of viewers.

Basic Charts

Discussion

Discussion

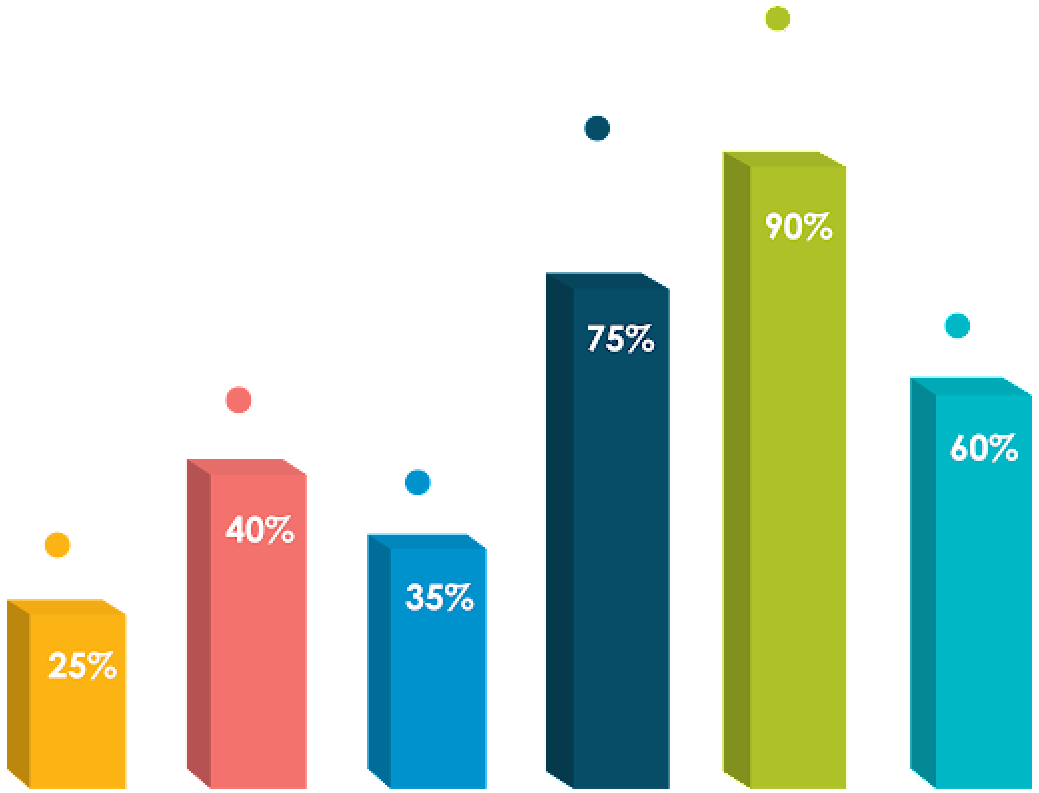
Duration: 15 minutes

How should data be visualized?

- Why is it sometimes ineffective to convey data in a tabular format?
- What are the types of visualization?

Bar Chart

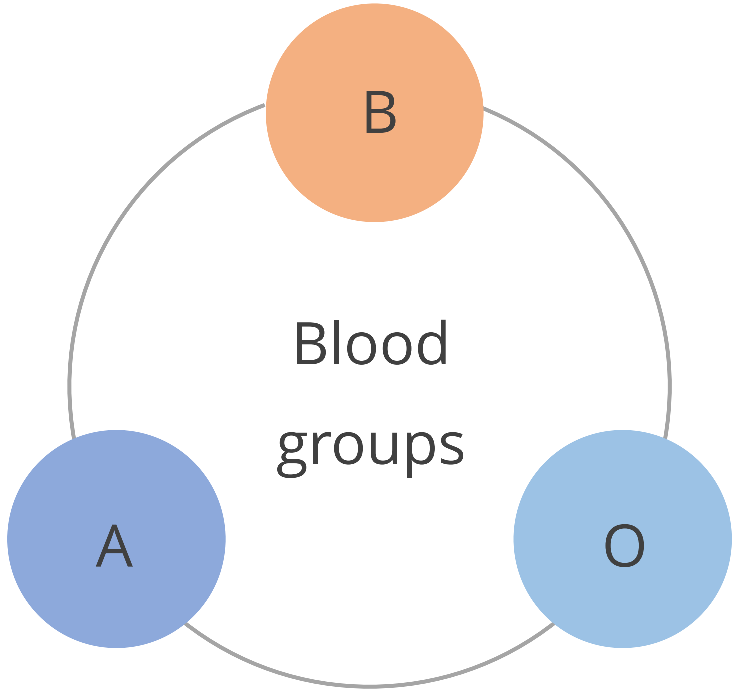
It is a graphical representation of data attributes using rectangular bars with heights or lengths proportional to the values they represent.



For example, in the image above, the heights are proportional to the frequencies of occurrences.

Bar Chart

Example: A blood bank supplies three groups of blood.



Blood is stored in plastic bags.

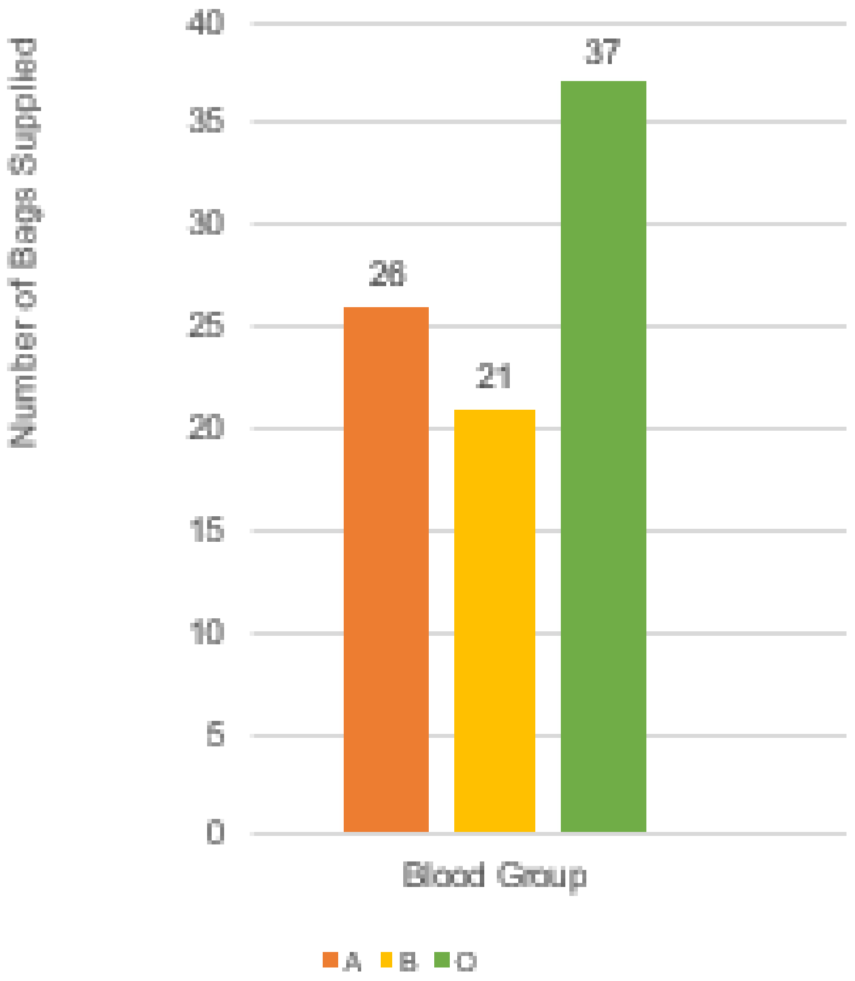
Bar Chart

The number of plastic bags supplied for each group is shown in the table below:

Blood group	No. of bags supplied
A	26
B	21
O	37
	Total: 84

Bar Chart

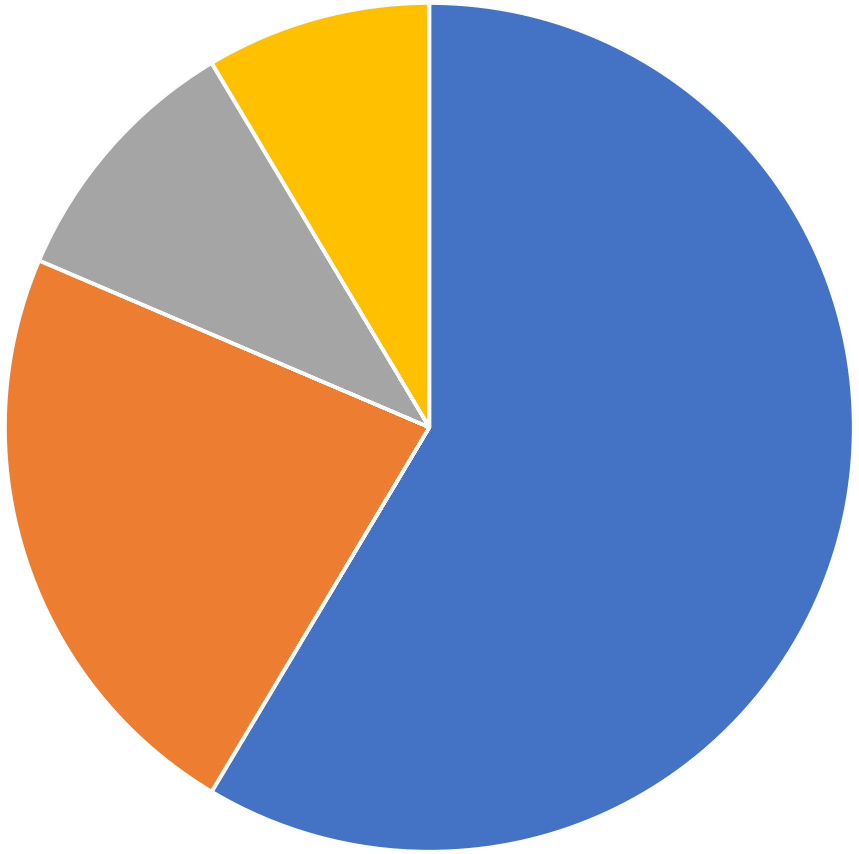
Shown below is the bar chart representing the data from the table.



The x-axis denotes the blood group, and the y-axis denotes the number of plastic bags.

Pie Chart

It consists of a circle with several sectors, one for each attribute.

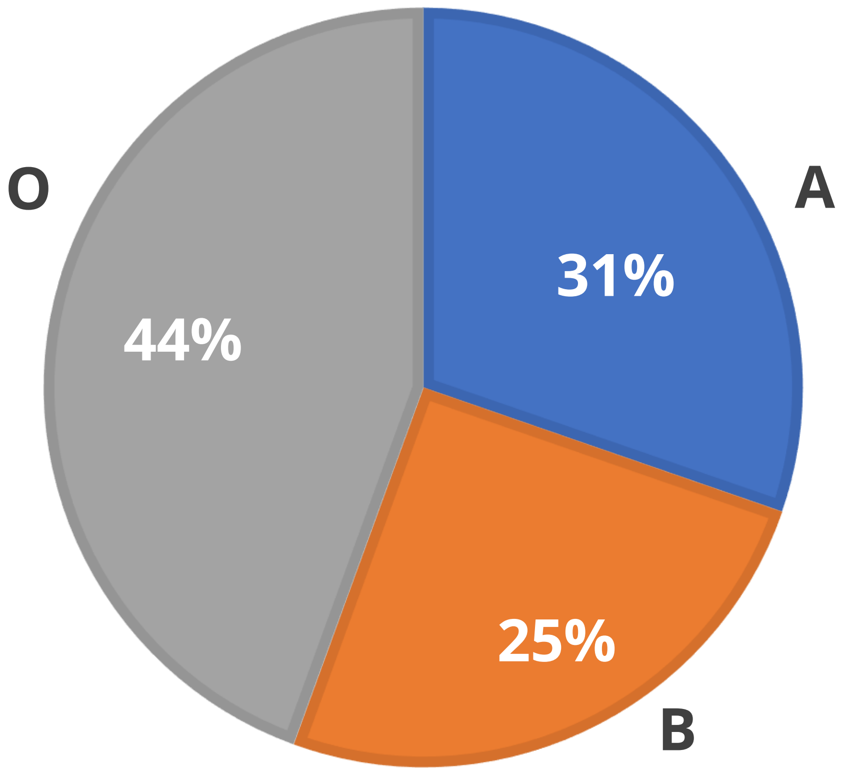


The chart's fraction is the sum of each category in an attribute.

Pie Chart

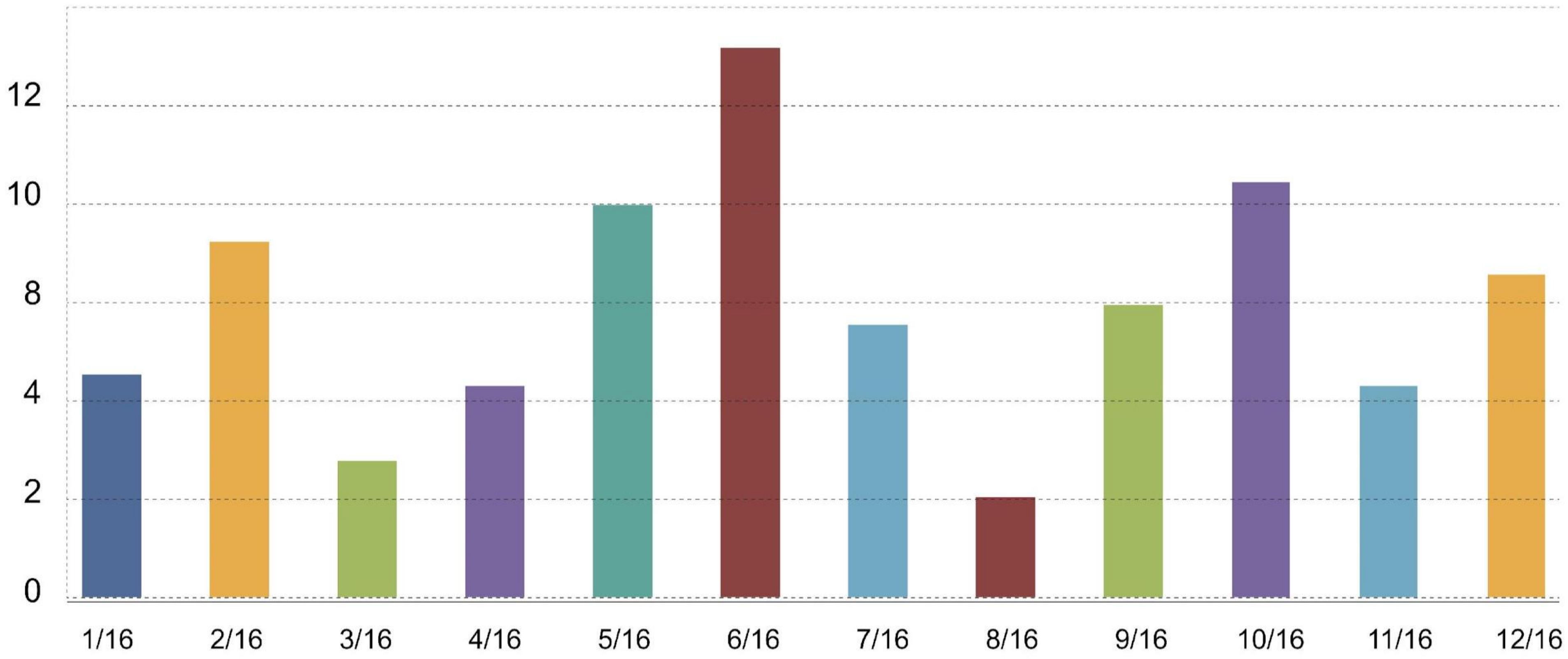
Example: The pie chart for the blood group data is shown below.

Blood group	No. of bags supplied
A	26
B	21
O	37
	Total: 84



Histogram

These are created using frequency distributions.



They visually display data points organized into specified ranges determined by the user.

Histogram

A histogram for a variable frequency distribution consists of several rectangles.

Width of rectangles ∝ Width of class intervals

Height of rectangles = Frequencies

Successive rectangles are adjacent to one another.

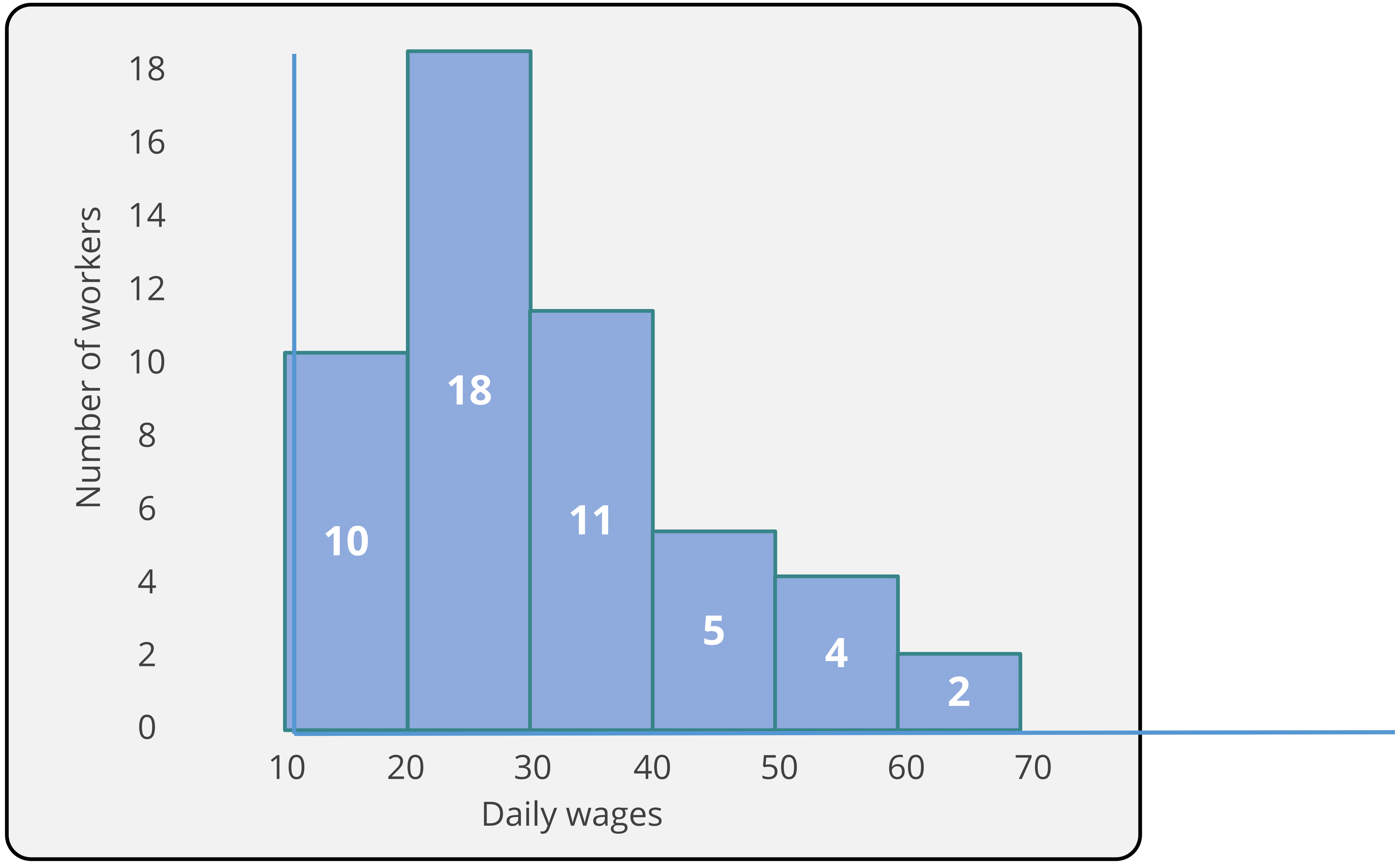
Histogram

Example: Consider a frequency distribution table that depicts the occupancy rate for daily wages and the number of workers

Class intervals	Frequencies
10–20	10
20–30	18
30–40	11
40–50	5
50–60	4
60–70	2
Total	50

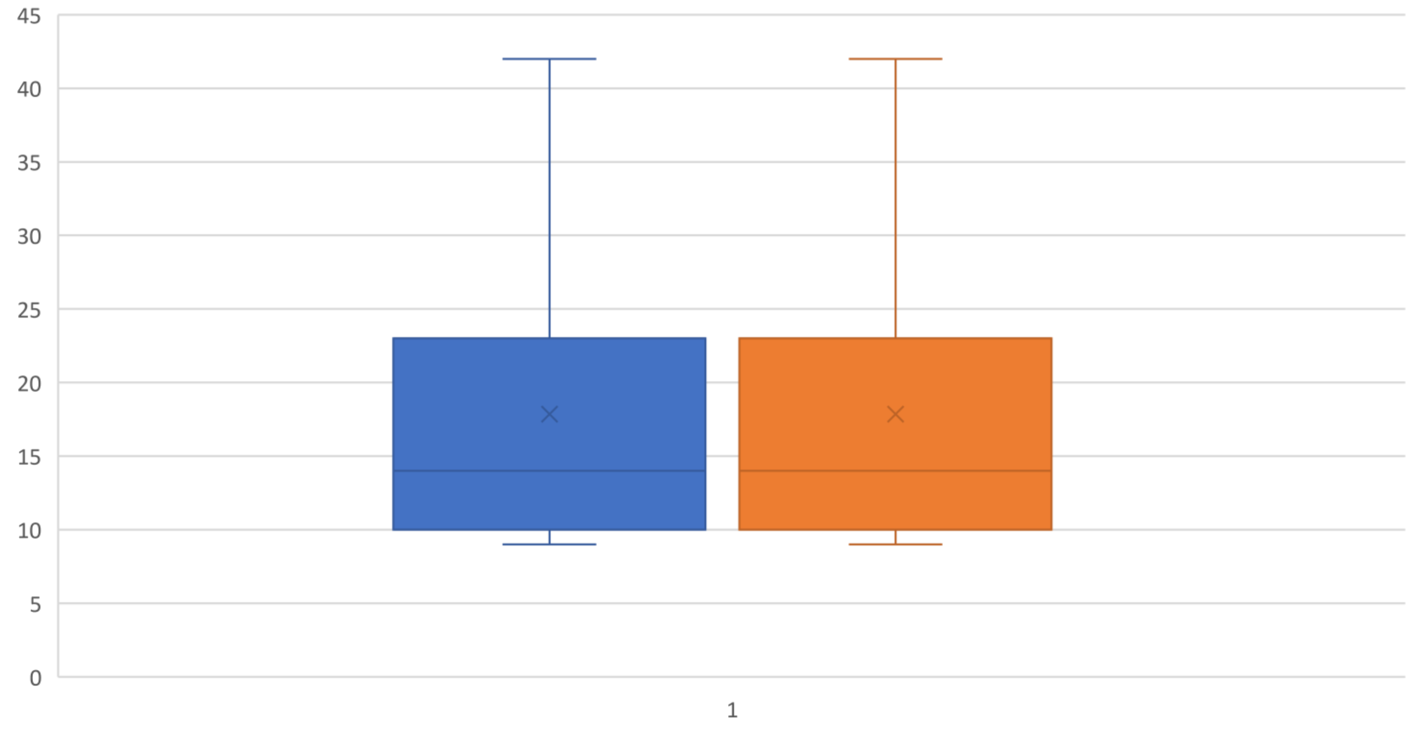
Histogram

Shown below is the histogram that represents the data from the table.



Box Plot Chart

A box plot, also known as a box chart or box and whisker chart, is a graphical representation that showcases groups of numerical data based on their quartiles.



They summarize data spread using five important values namely minimum, maximum, first quartile, third quartile, and median.

Box Plot Chart: Example

Determine the maximum, minimum, median, first, second, and third quartiles for the following dataset: 23, 42, 12, 10, 15, 14, and 9.

Given: 23, 42, 12, 10, 15, 14, 9.

Step 1: Arrange the given dataset in ascending order

- Resulting data = 9, 10, 12, 14, 15, 23, 42

Step 2: Compute the first and third quartiles

- First Quartile = 10 (Middle value of 9, 10, 12 is 10)
- Third Quartile = 23 (Middle value of 15, 23, 42 is 23)

Discussion

Duration: 15 minutes

- Why is it sometimes ineffective to convey data in a tabular format?

Answer: Data when represented visually can be easily understood and therefore can save time. On the other hand, tabular data can be confusing, hard to understand, and time-consuming. Hence tabular data is sometimes ineffective in conveying the information.

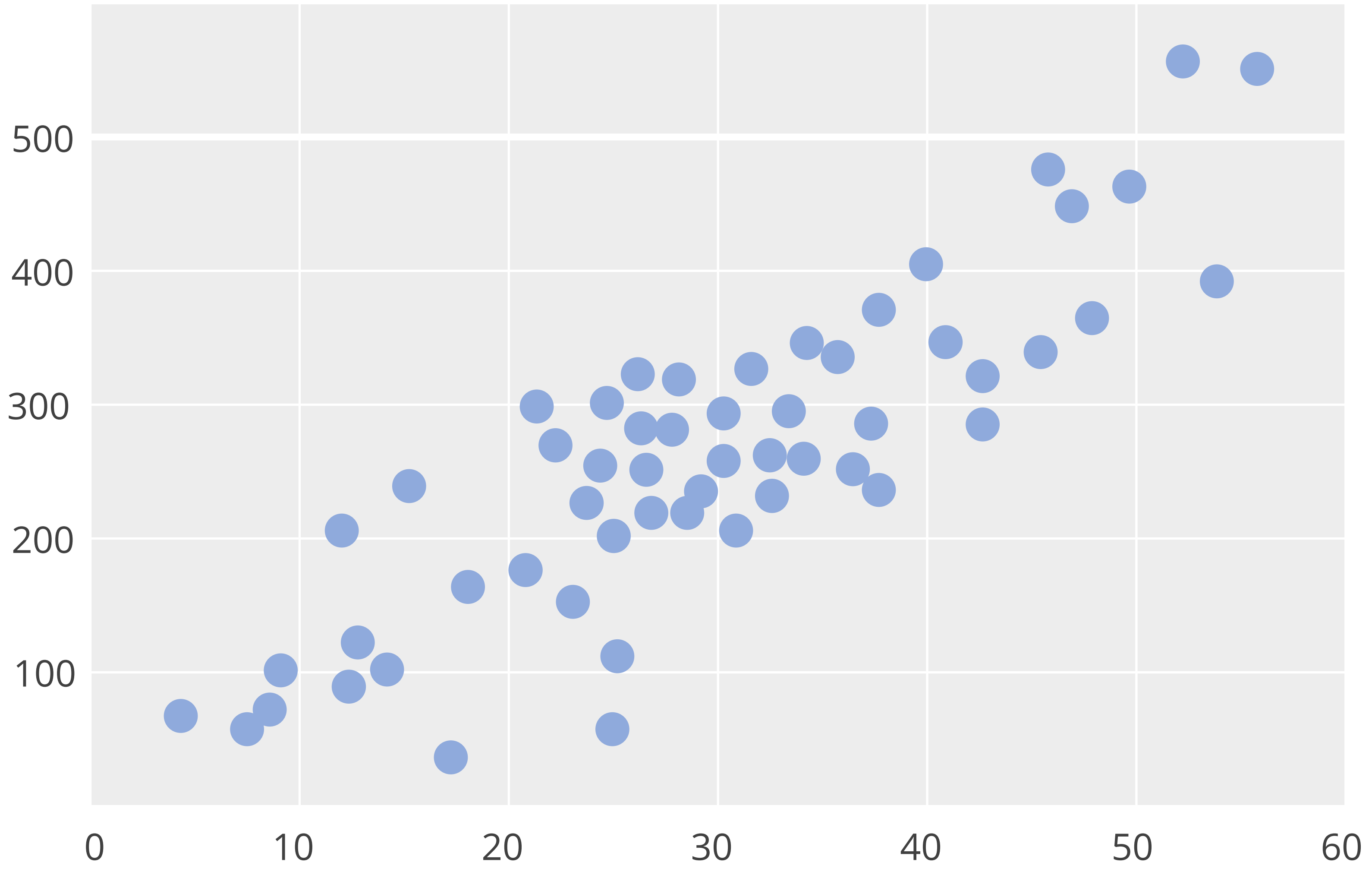
- What are the types of visualization?

Answer: The data can be visualized using an extensive array of techniques, including bar charts, pie charts, histograms, and box plots, among others.

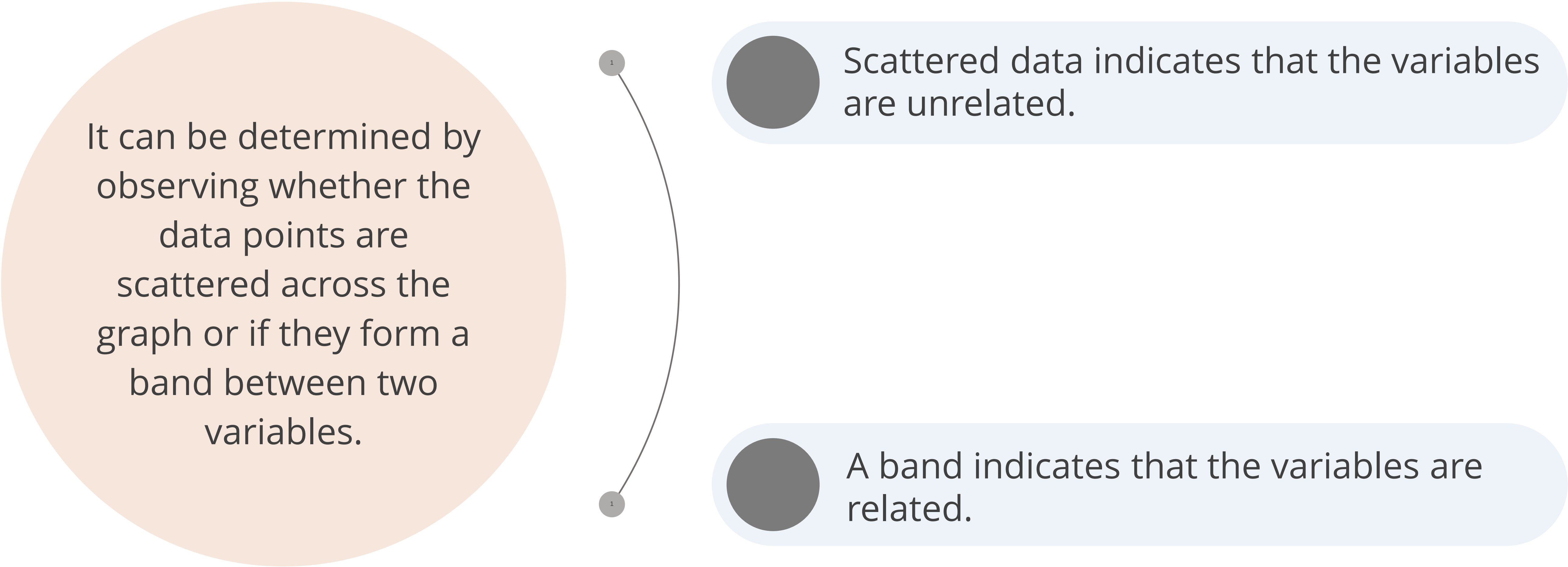
Advanced Charts

Scatter Plot

These graphs display the relationship between two variables in a dataset.



Scatter Plot



It can be determined by observing whether the data points are scattered across the graph or if they form a band between two variables.

Scattered data indicates that the variables are unrelated.

A band indicates that the variables are related.

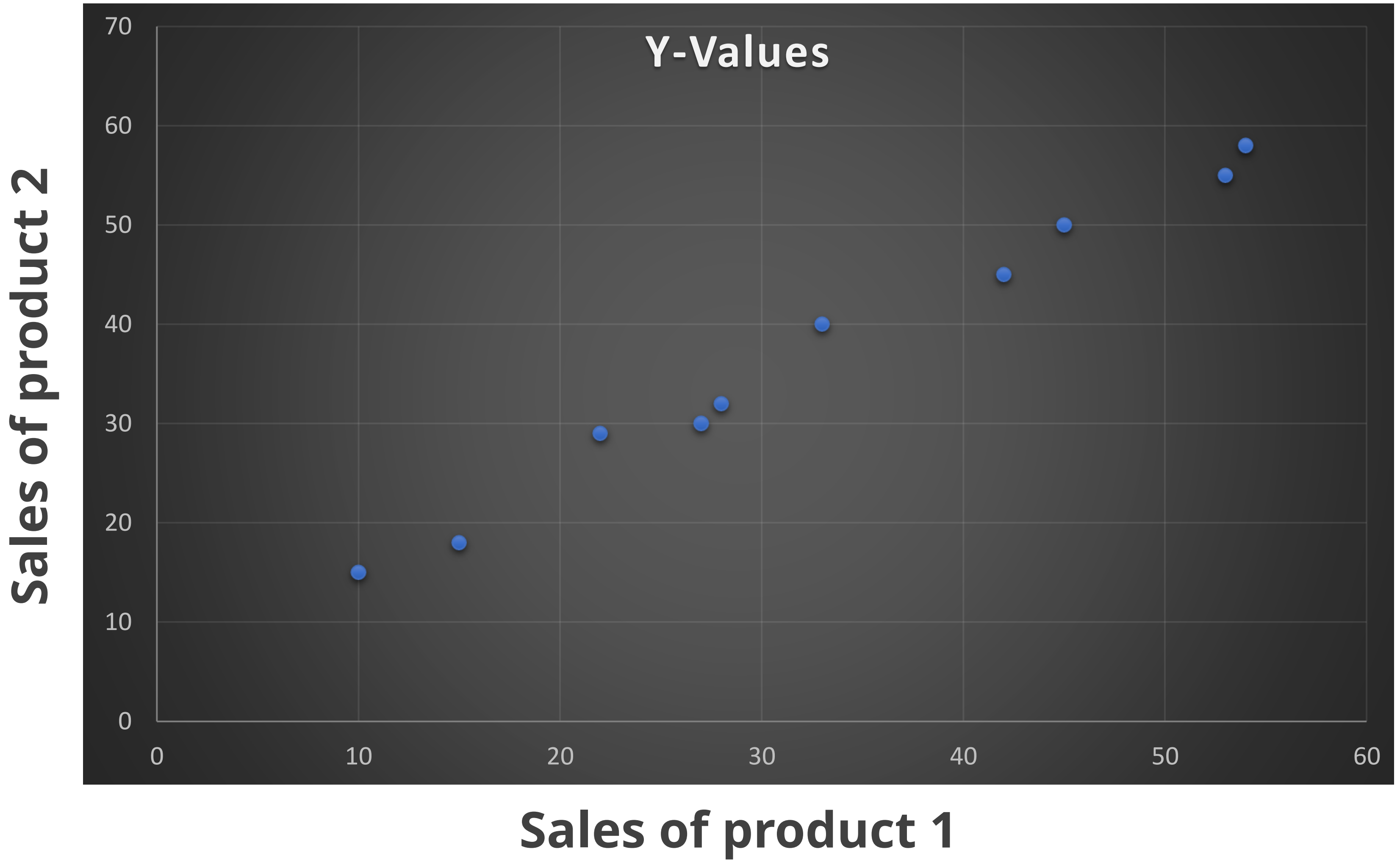
Scatter Plot

The following data represents the sales of two products at a retail outlet over a 10-day period

Product 1	10	15	21	27	28	33	41	44	51	52
Product 2	15	19	27	30	35	39	46	60	58	59

Scatter Plot

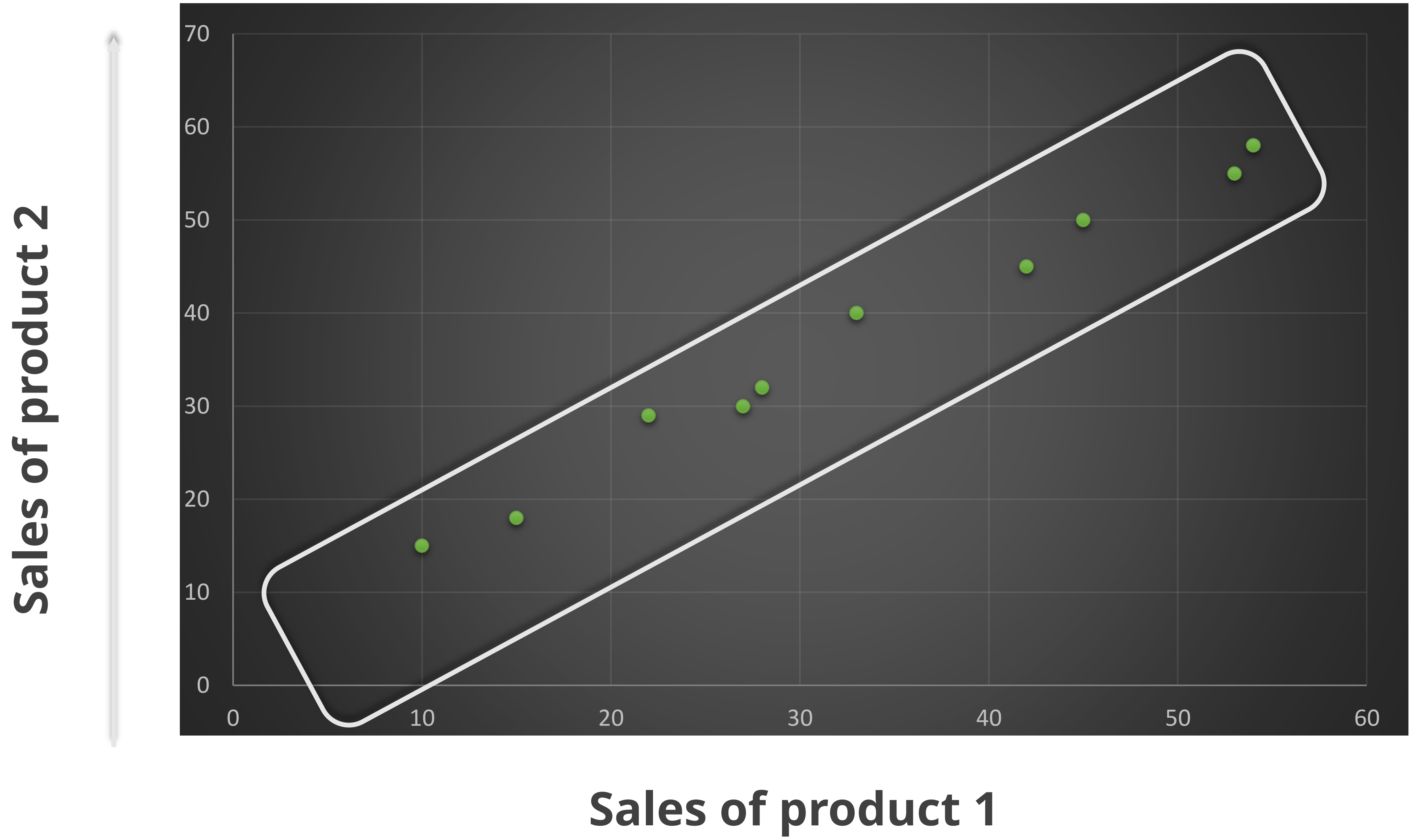
Shown below is the scatter plot chart depicting the data given in the table.



Values of sales for the first product are shown on the X-axis and sales for the second product are shown on the Y-axis.

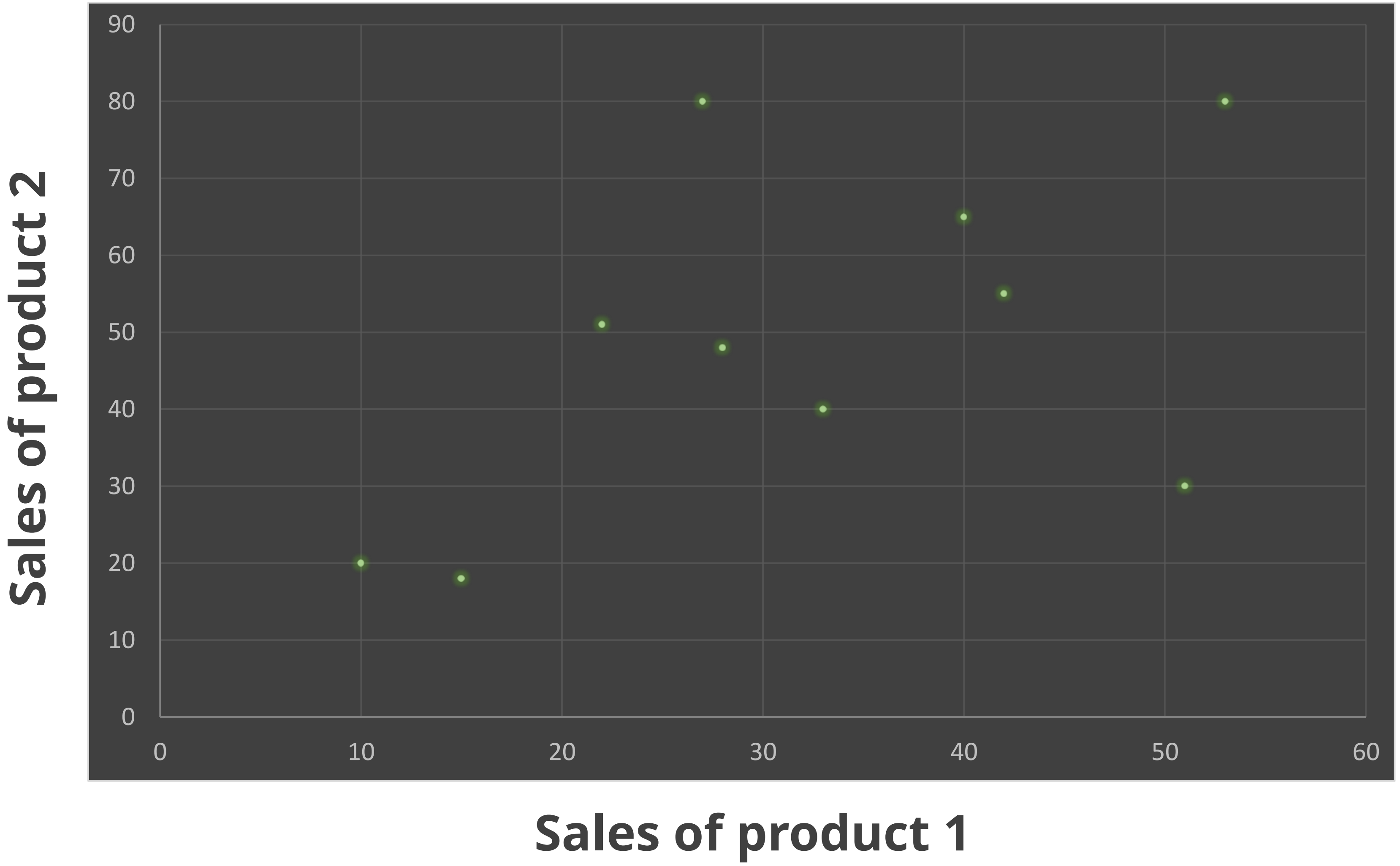
Scatter Plot

A narrow band indicates a relationship between the two variables, such as the sales of two products on different days.



Scatter Plot

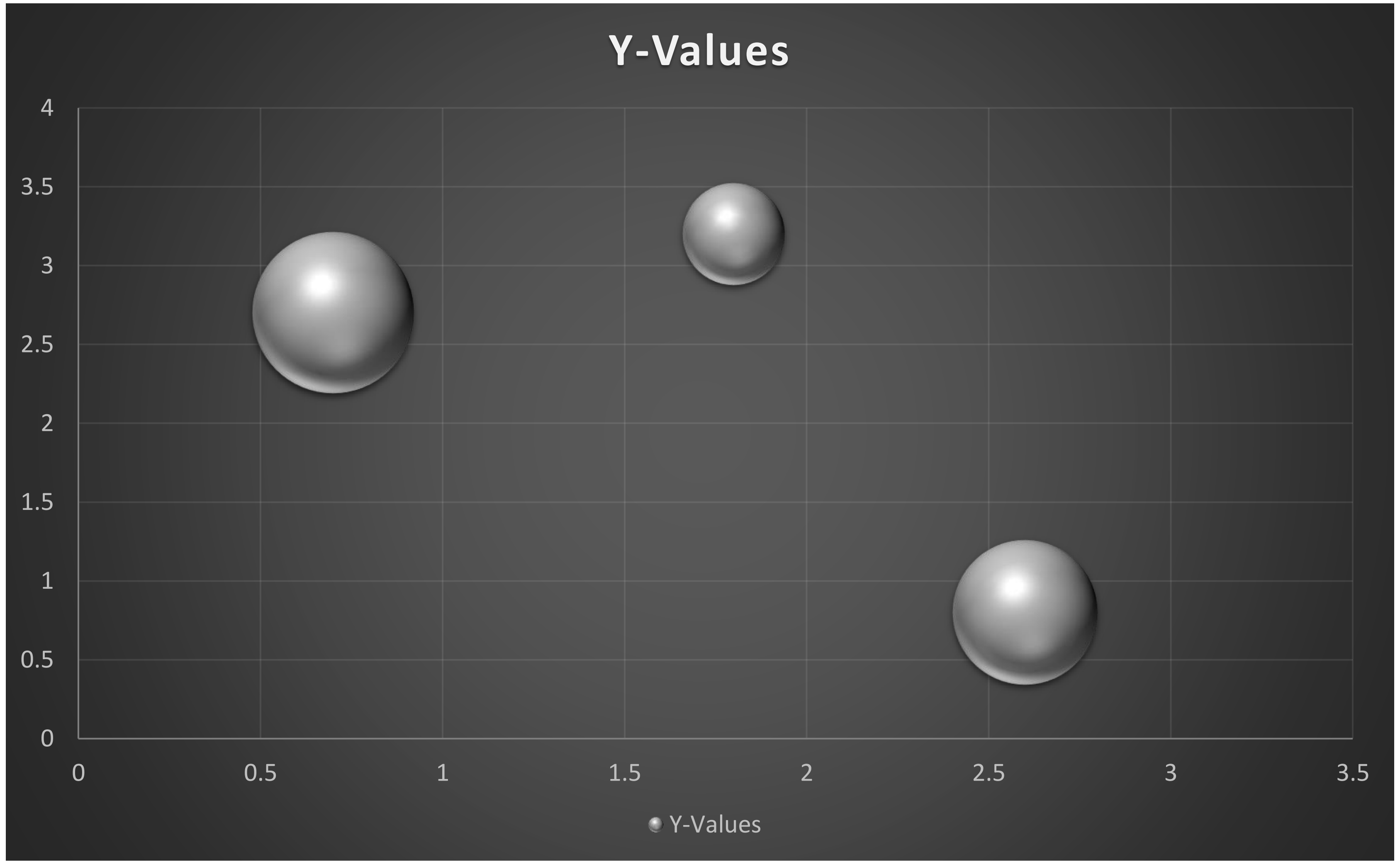
Plotted values are scattered across the chart and values do not fall in the band.



In such cases, there is no relationship between the variables.

Bubble Plot

This is an extension of the scatter plot used to identify relationships between three numerical variables.



Bubble Plot

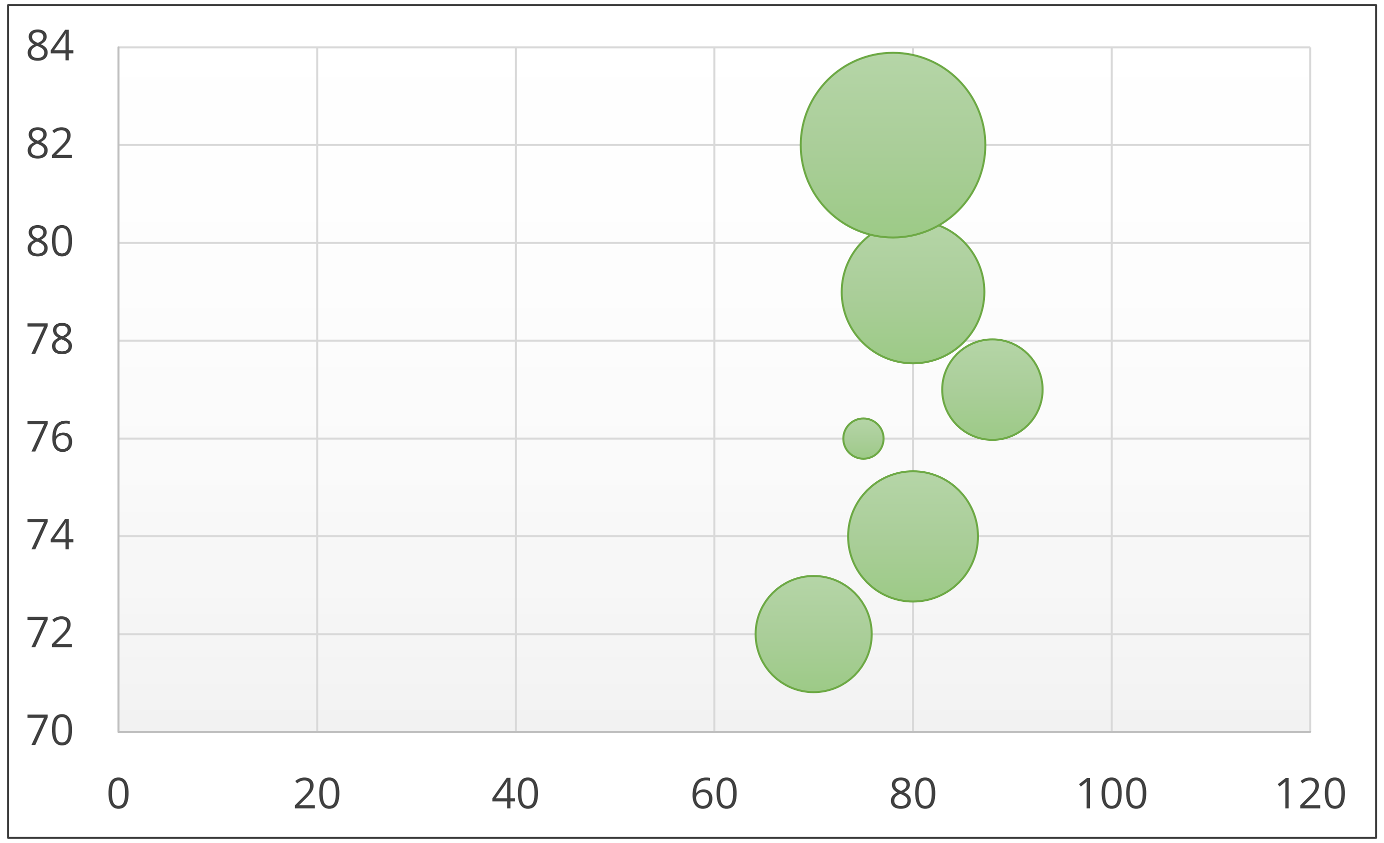
It is a data visualization that uses bubbles to represent data points, with the size of the bubble indicating a third dimension of data.

Example: Plot a bubble chart using the three variables in the given data

Variable 1	78	80	88	78	70	75
Variable 2	82	79	77	74	72	76
Variable 3	87	79	77	80	78	74

Bubble Plot

The size of the bubbles should be proportional to the third variable's value.



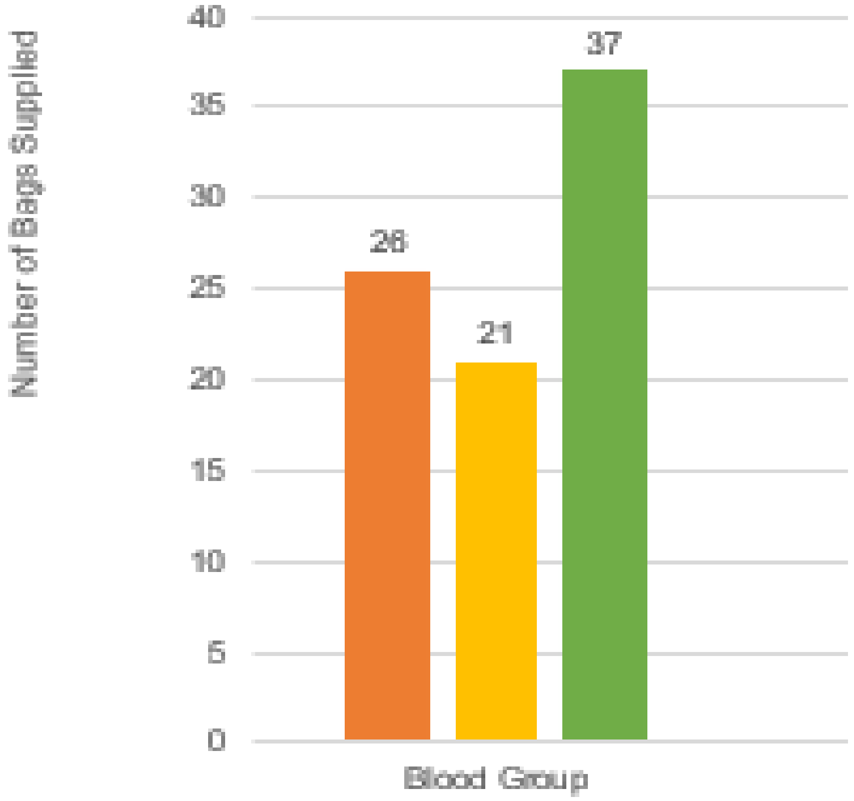
The first variable varies over a broader range, and hence the Y-axis is spread over a wider range.

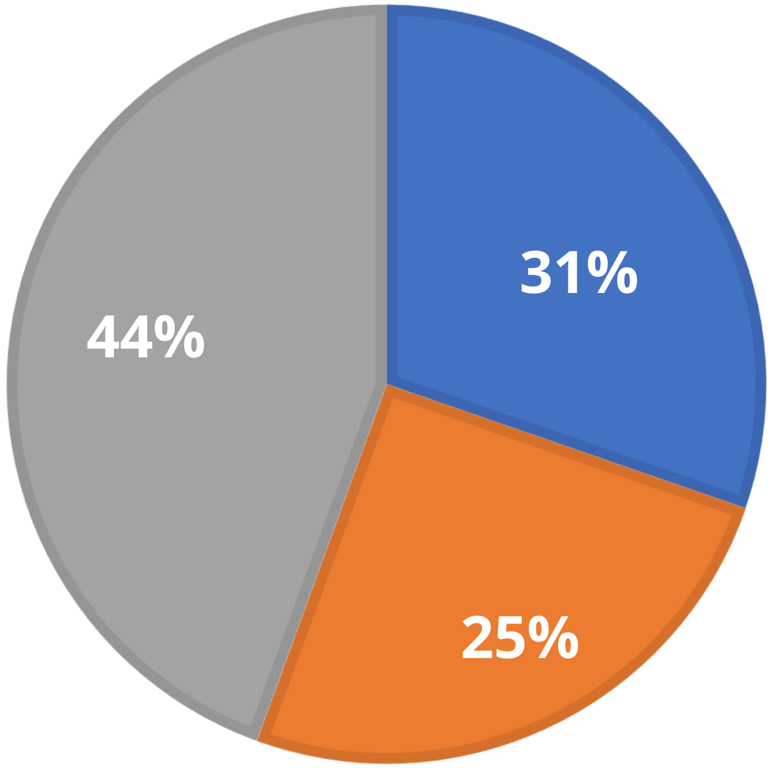
The reduced spread of the second variable results in a narrower range on the X-axis.

Interpretation of the Charts

Identifying Attributes in Charts

Values that occur more frequently in the attributes can be easily identified.

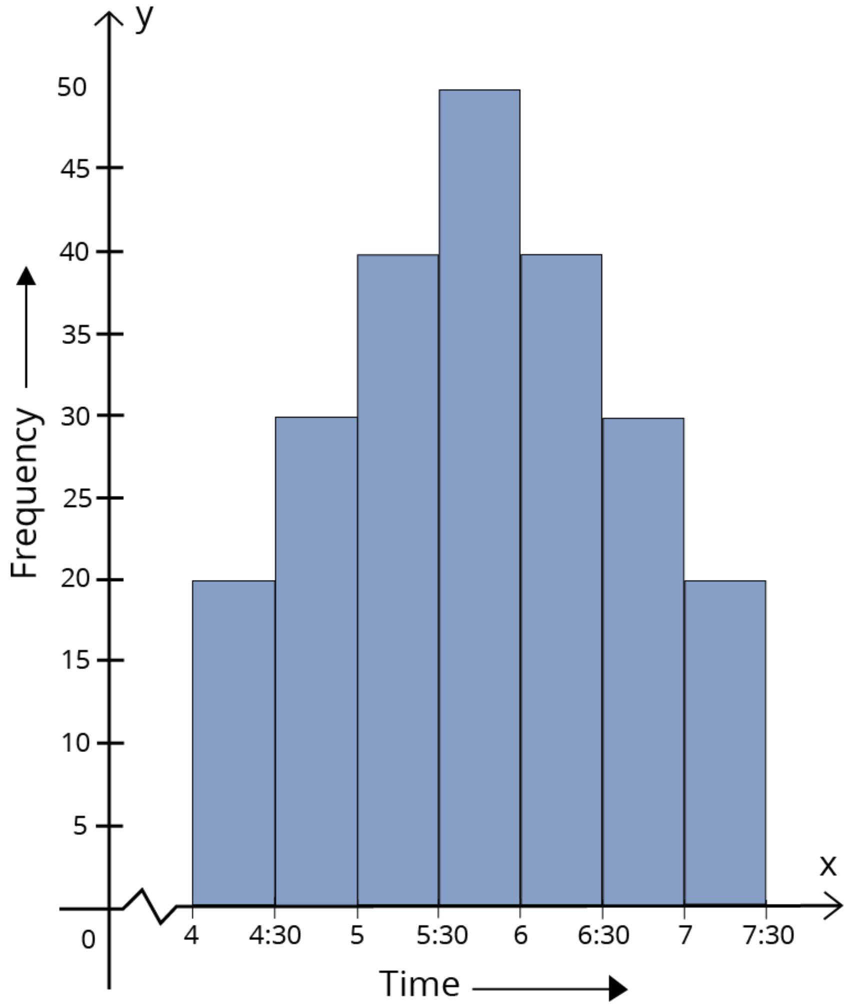




The heights of bars in a bar chart and the areas of sectors in pie charts are proportional to the frequencies of the data they represent.

Presence or Absence of a Pattern

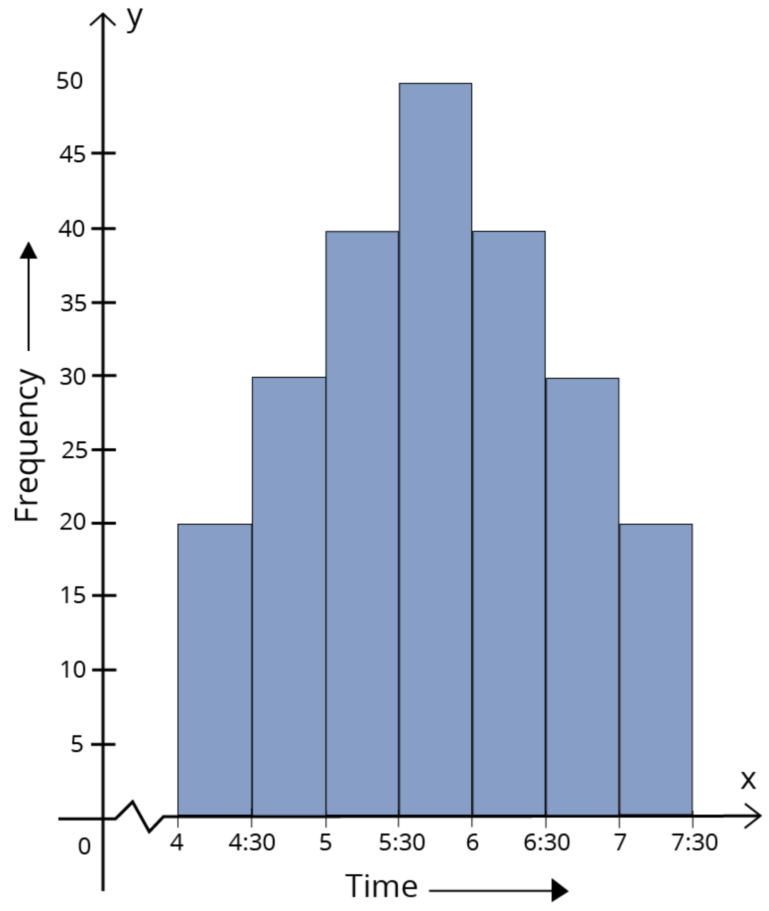
Here, the pattern is obtained as the frequency initially increases, reaches a peak, and then gradually decreases.



- Sometimes, multiple peaks occur, indicating the absence of a normal distribution pattern.
- Different histograms for each season help in identifying the pattern.

Degree of Symmetry

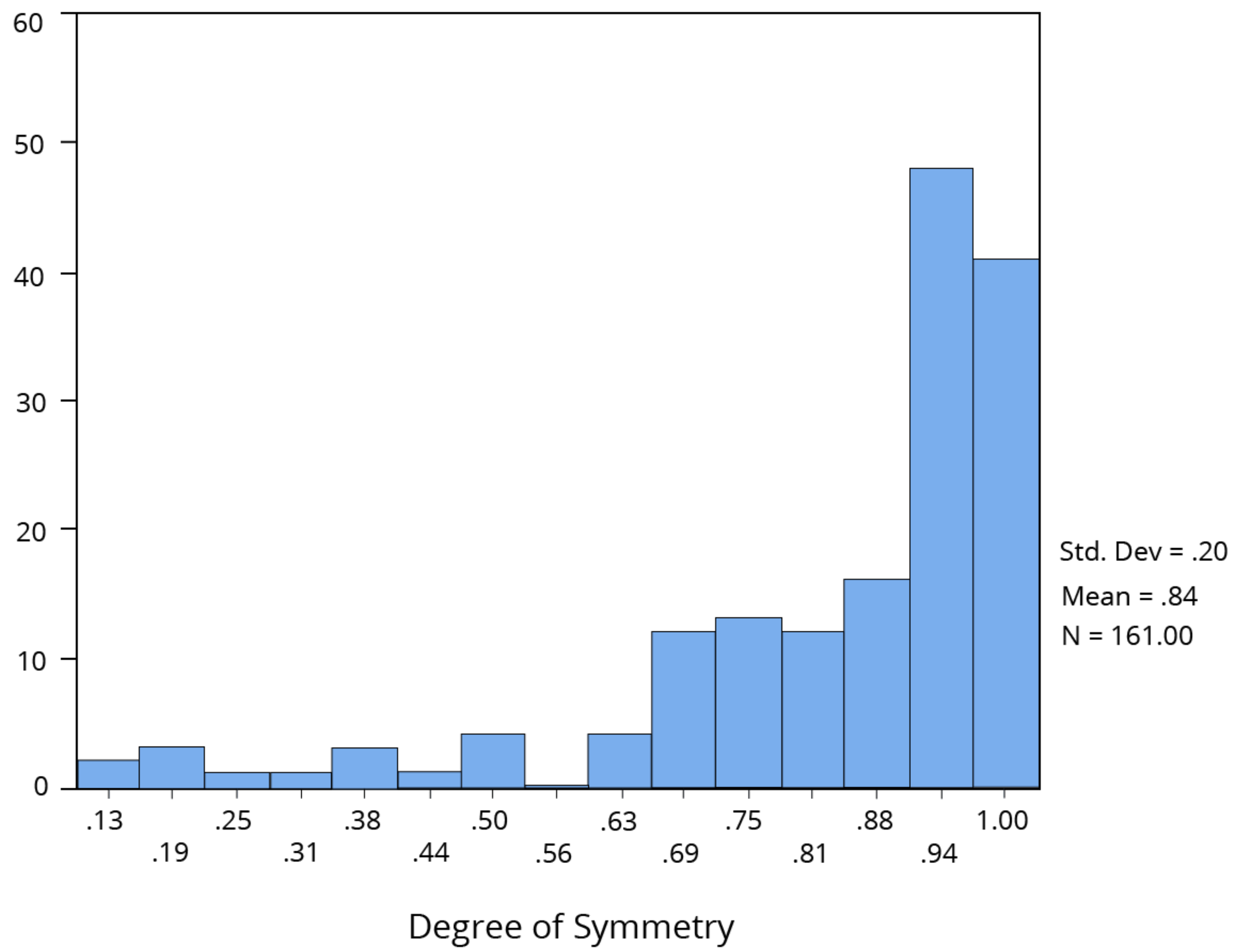
When a line is drawn through the center of a symmetry histogram, its two halves are identical.



The mean, median, and mode values are identical, and all fall within the center of the distribution for a symmetric histogram.

Degree of Symmetry

Data is skewed to the left.

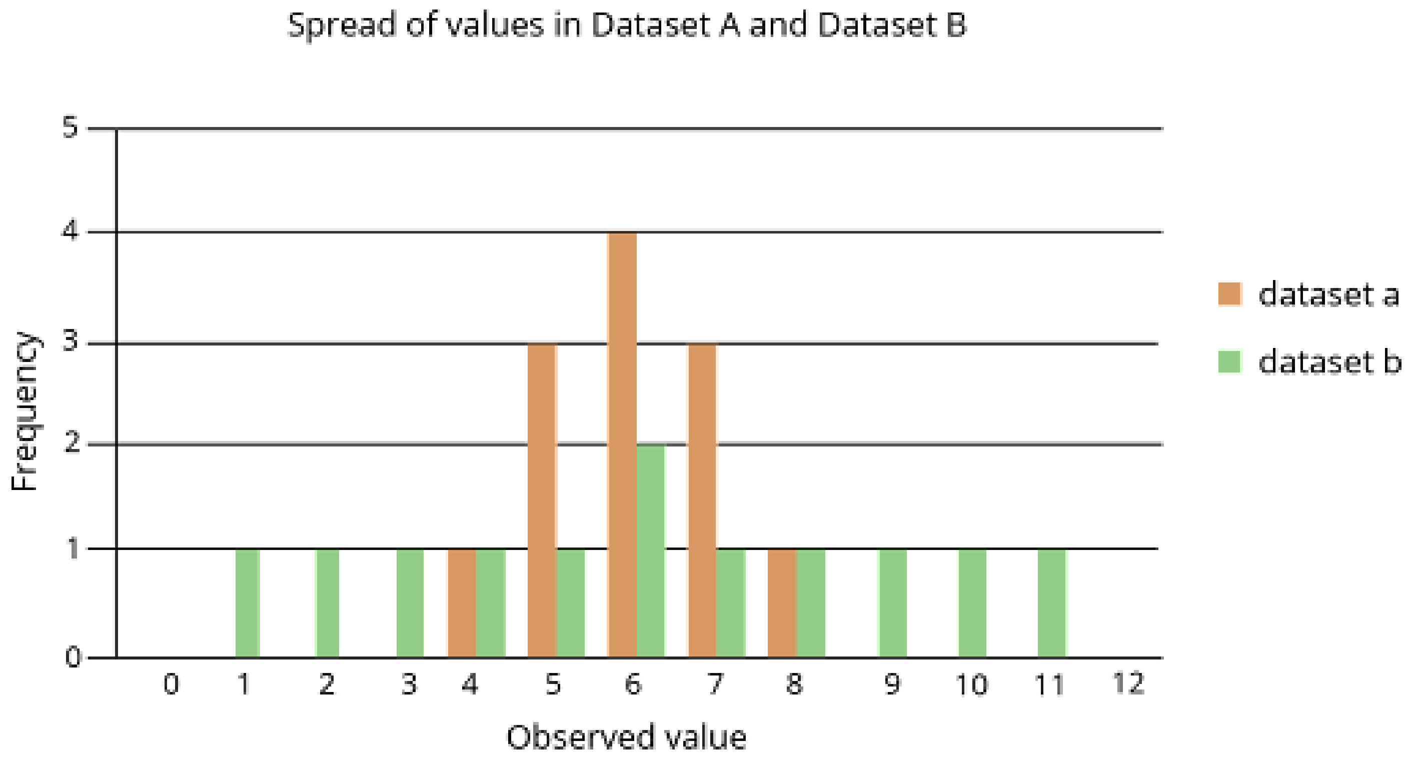


It is more common for values to surpass both the average and peak occupancy rate.

Data may also be symmetrical or skewed to the right, which can be interpreted accordingly.

Degree of Spread

The similarity or diversity of the set of observed values for a specific variable is described by measures of spread.



Spread of values in Dataset A and Dataset B

- A measure of spread describes the variability in a sample or population.
- A measure of spread indicates how well the mean represents the data.
- The range, quartiles, interquartile range, variance, and standard deviation are all measures of spread.

Presence of Outliers

A single data point that significantly deviates from the average value of a set of statistics is referred to as an outlier.

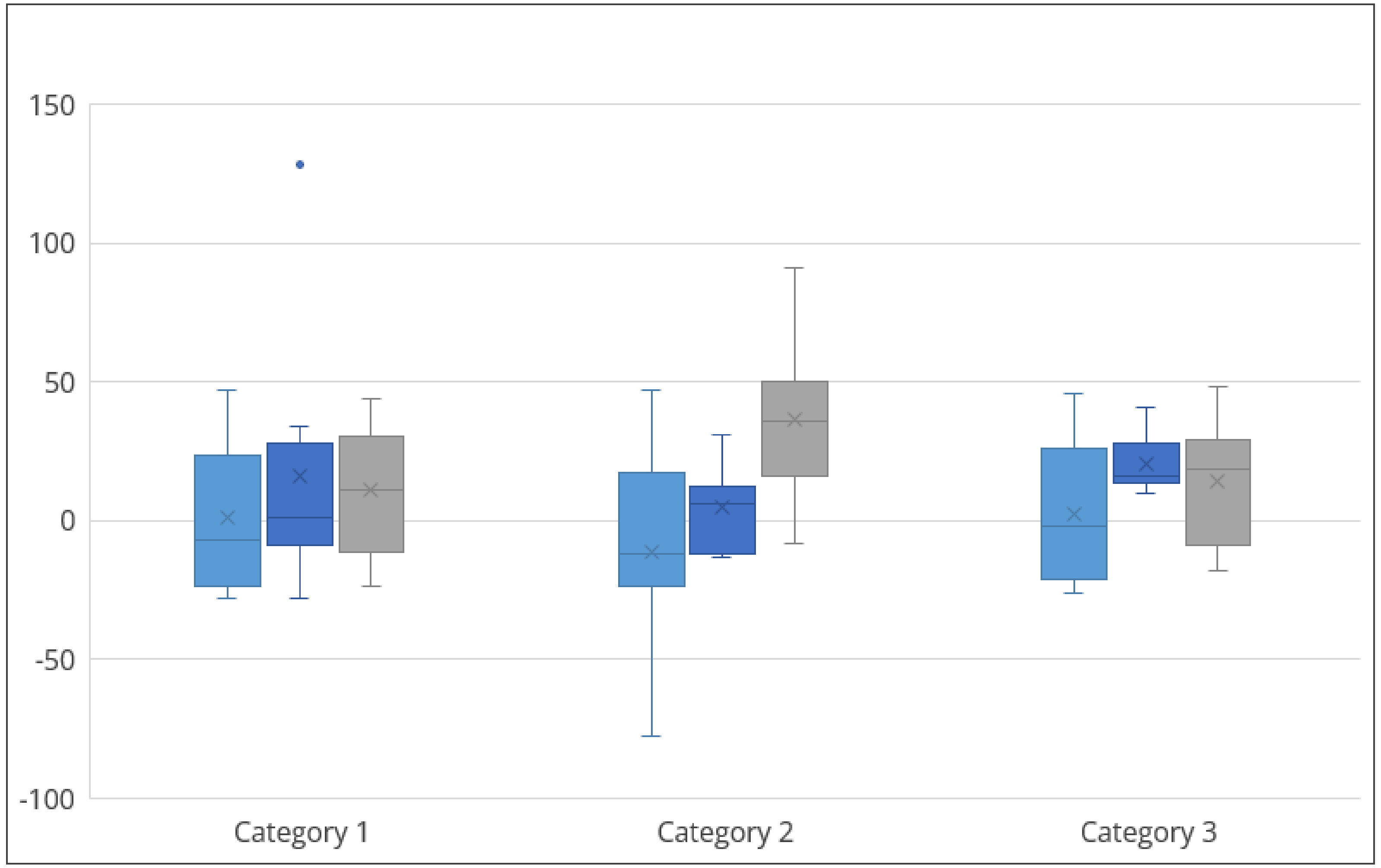
They can occur owing to unforeseen events like railway strikes and special events.

Extremely small or large values are treated as outliers.

They can be excluded from analysis when they are considered unrepresentative of typical patterns.

Presence of Outliers

Box plots are useful for identifying outliers in a dataset.



Box plot chart

Selecting the Appropriate Chart

Discussion

Discussion

Duration: 20 minutes

- What are outliers in data?
- How do outliers affect the data?

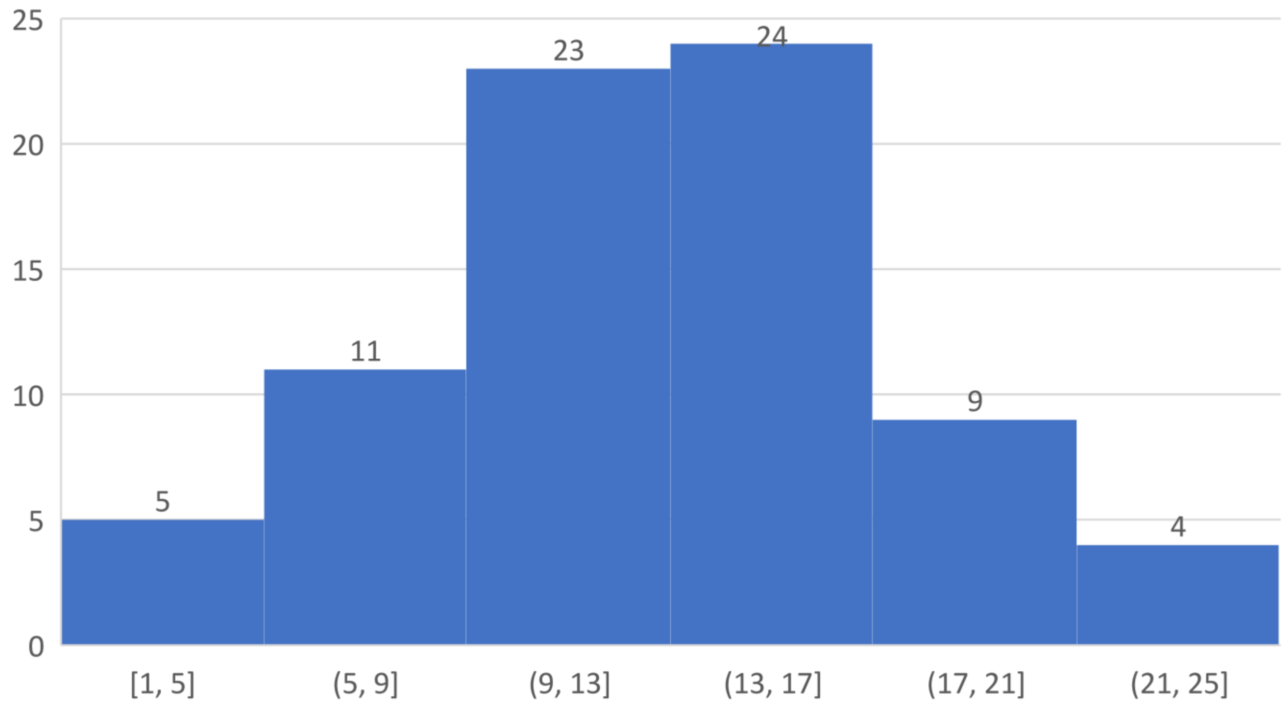
Charts

A variety of charts have been developed and used owing to varied requirements in different situations.

Bar and pie charts are used to represent qualitative data, while histograms and box plots are used for qualitative and quantitative data.

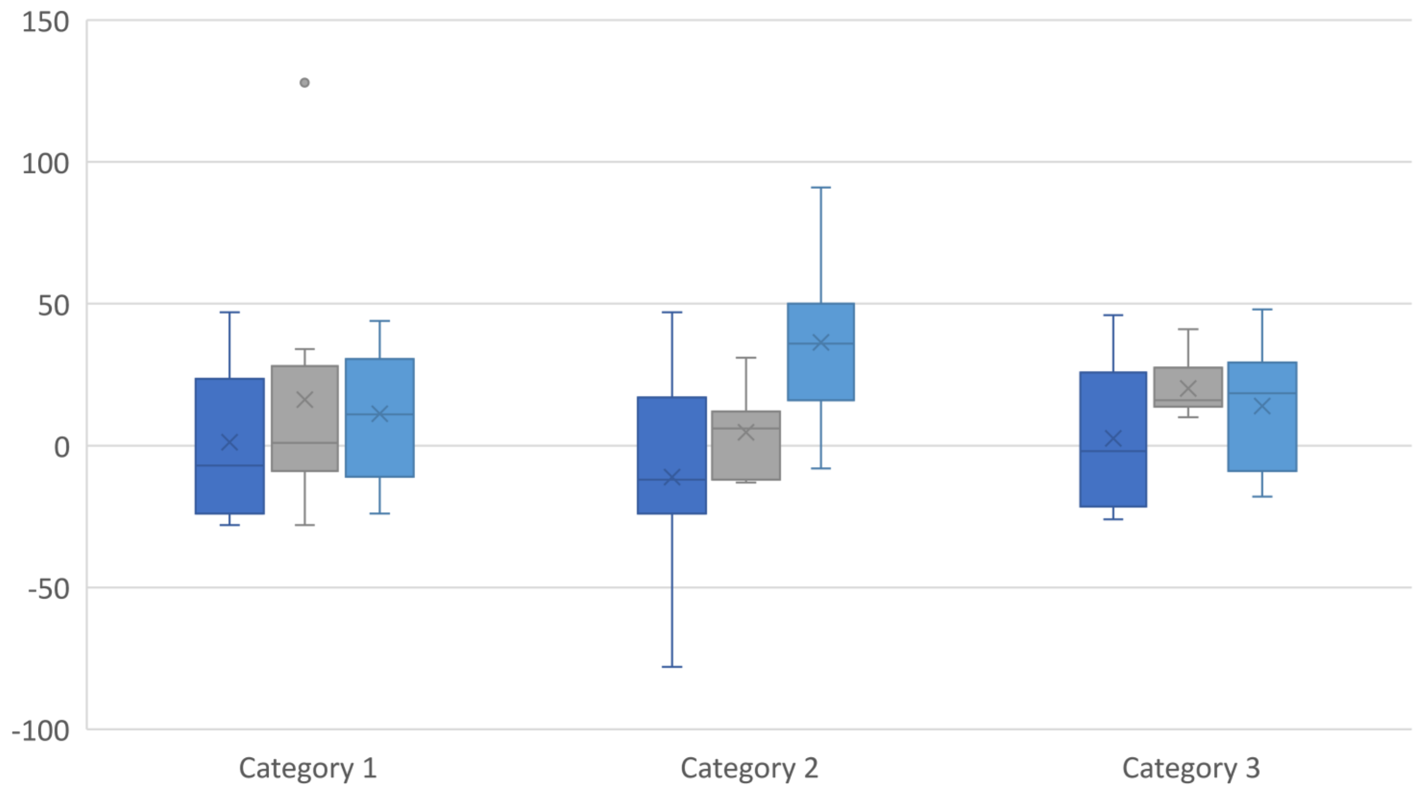
Histogram

They represent frequency distributions of quantitative data.



Box Plot

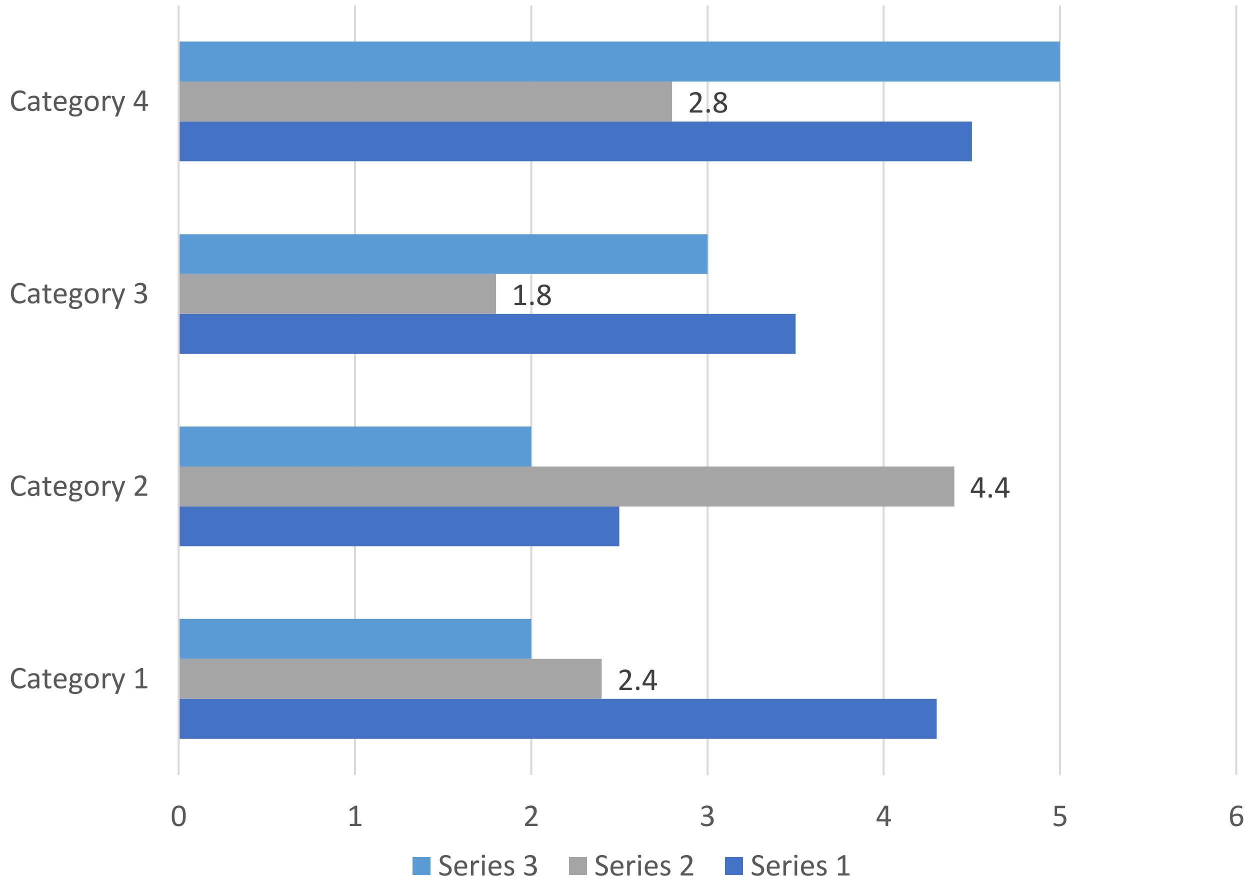
They visually represent the distribution of numerical data as well as any skewness present in the data.

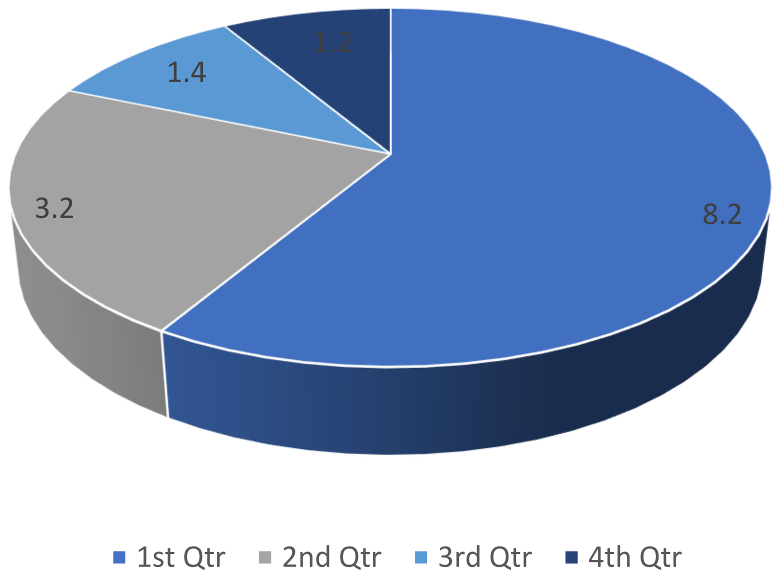


The diagram incorporates information on extremes and quartiles which helps to project outliers.

Bar and Pie Charts

Bar and pie charts can be used to display multiple attributes of a characteristic.

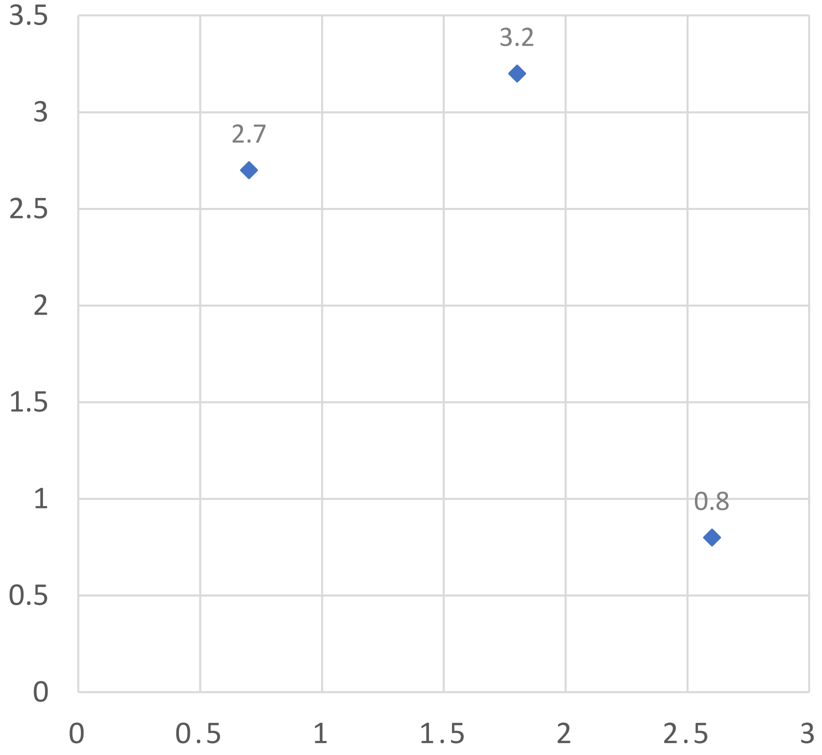


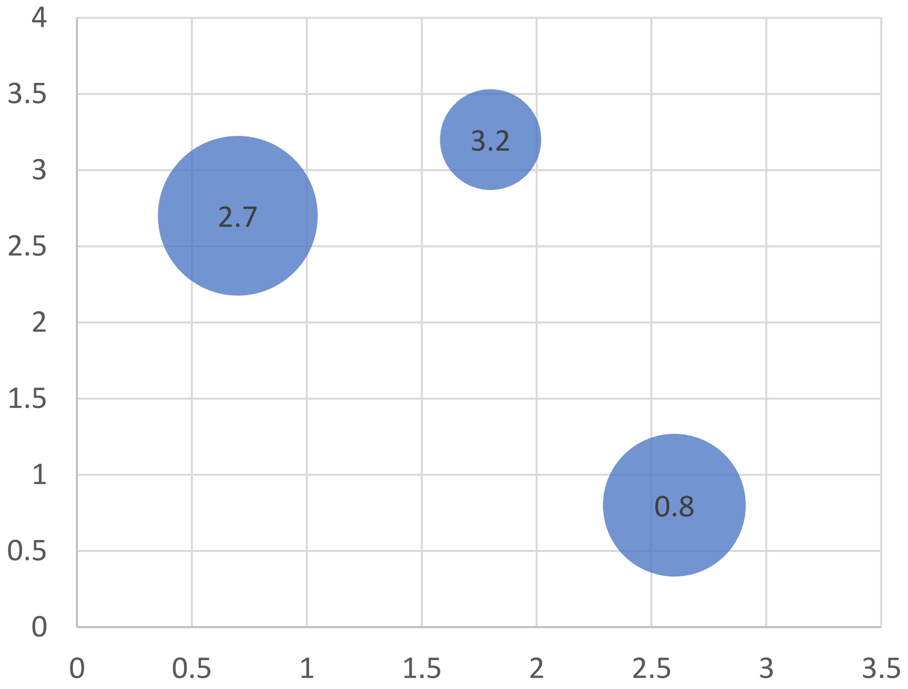


Multiple characteristics can be included in a single chart.

Charts for Multiple Datasets

When two or more datasets with different characteristics need simultaneous study, the following charts are helpful:





The bubble plot incorporates three attributes, while the scatter plot incorporates just two.

Uses of Charts

Charts enable the independent study of every characteristic for:

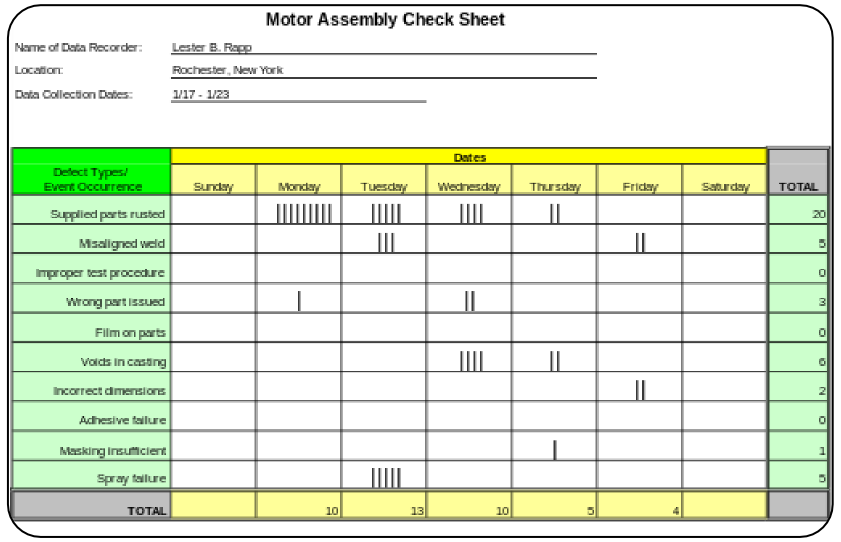
Dispersion

Central tendency

They also enable the user to identify the relationships between the variables.

Charts for Quality Control

The following set of quality improvement activities has been successfully implemented by numerous businesses across different industries.



Motor Assembly Check Sheet

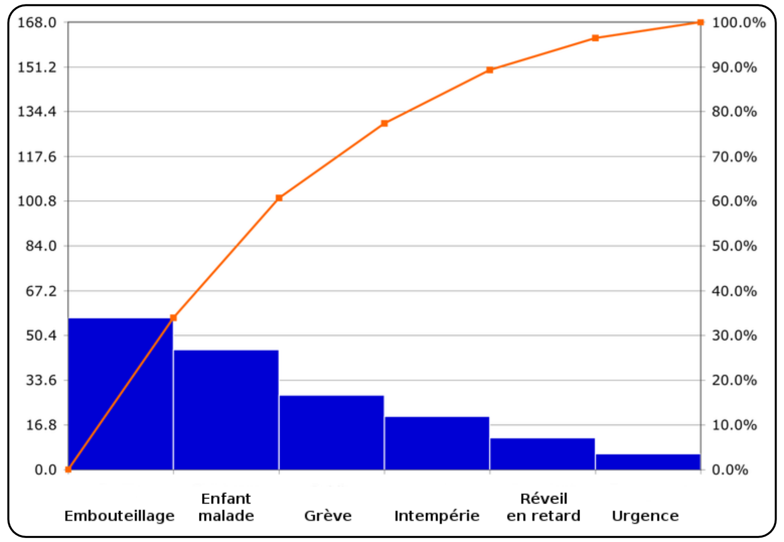
Name of Data Recorder: Lester B. Rapp

Location: Rochester, New York

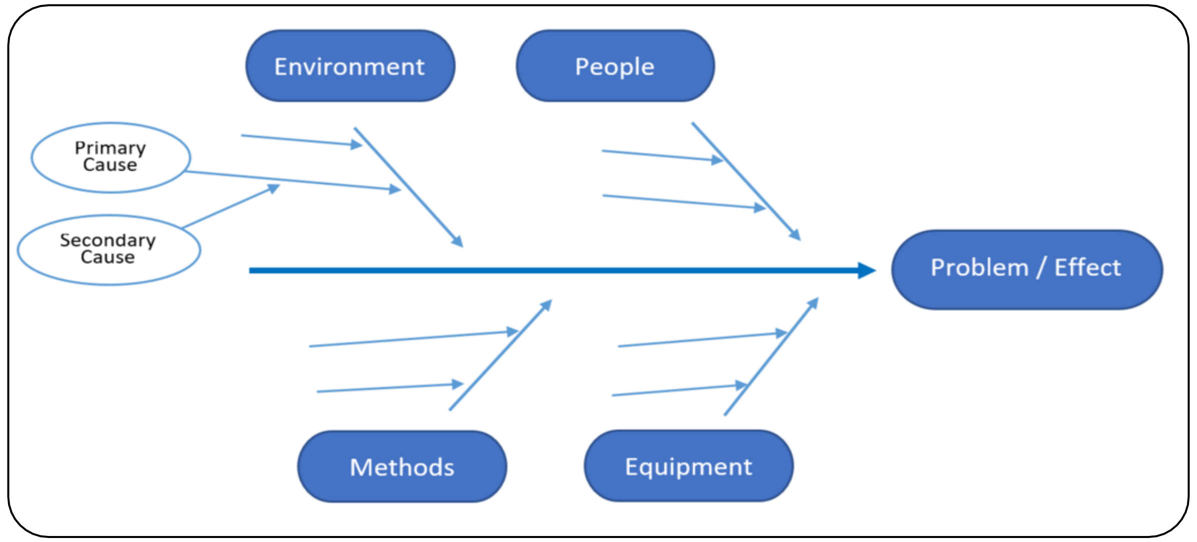
Data Collection Dates: 1/17 - 1/23

Defect Types/ Event Occurrence	Dates							TOTAL
	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	
Supplied parts rusted		|||||||||	|||||	||||	||			20
Misaligned weld			|||			||		5
Improper test procedure								0
Wrong part issued		|		||				3
Film on parts								0
Voids in casting				||||	||			6
Incorrect dimensions						||		2
Adhesive failure								0
Masking insufficient					|			1
Spray failure			|||||					5
TOTAL		10	13	10	5	4		

Check sheet



Pareto diagram

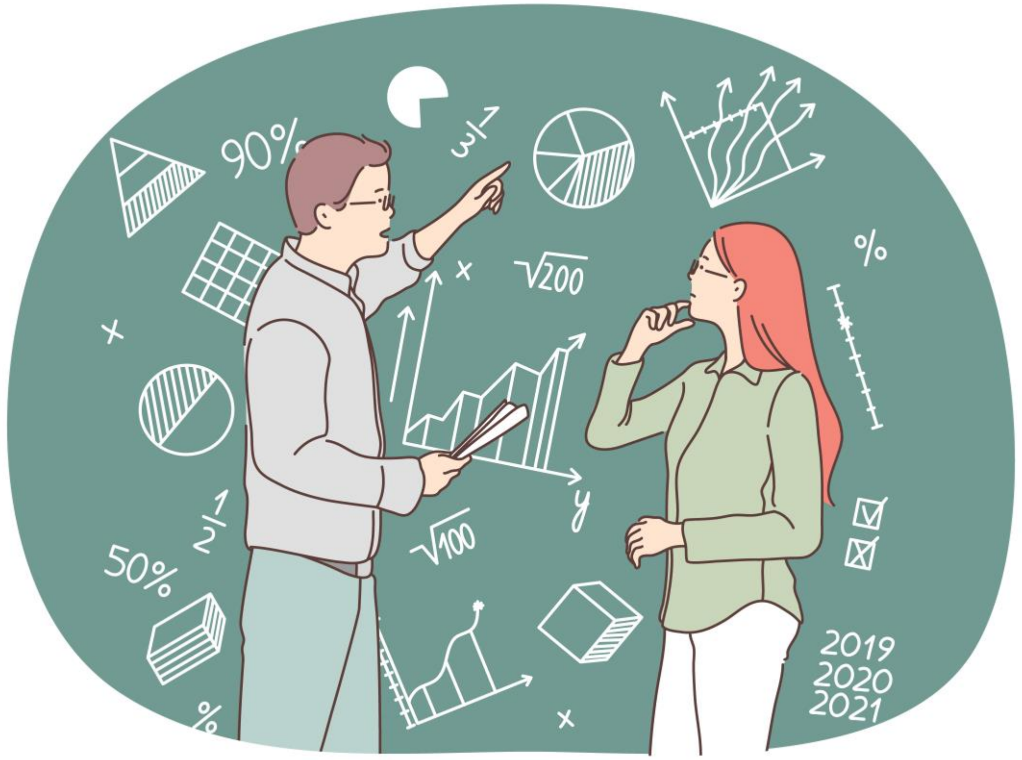


Ishikawa cause and effect diagram

Data Preprocessing: Dos and Don’ts

Things to Do in Data Preprocessing

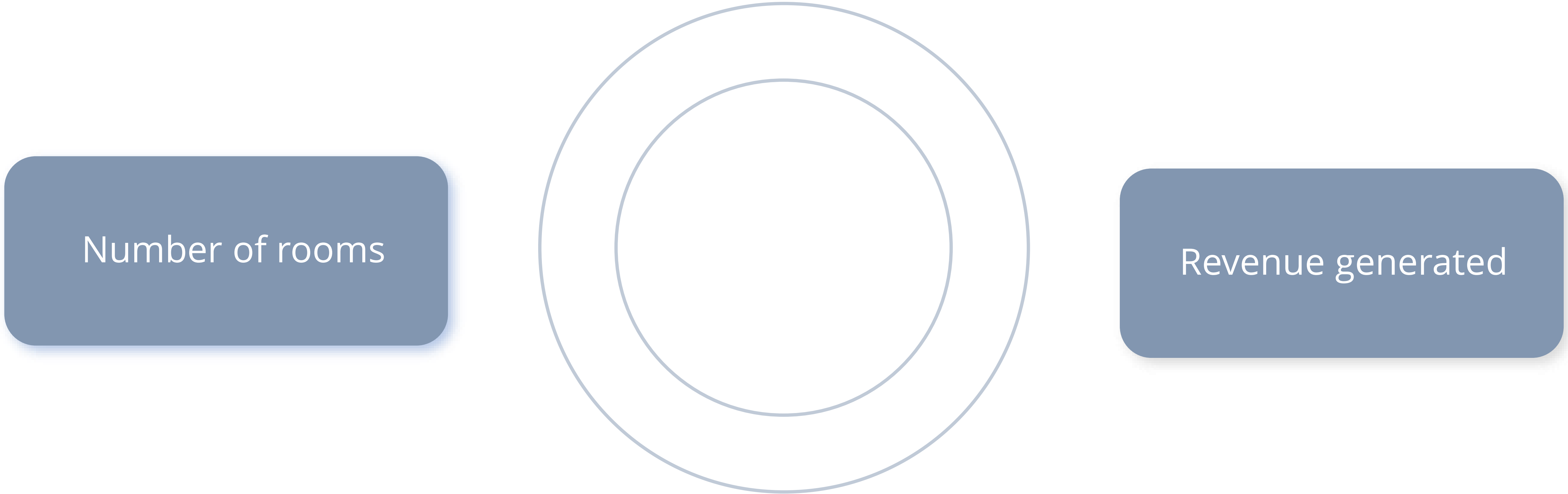
Incorporate the scope and purpose of the study



Carefully decide the variables to be studied

Decide the Variables

Consider an example of room occupancy where data is related to:

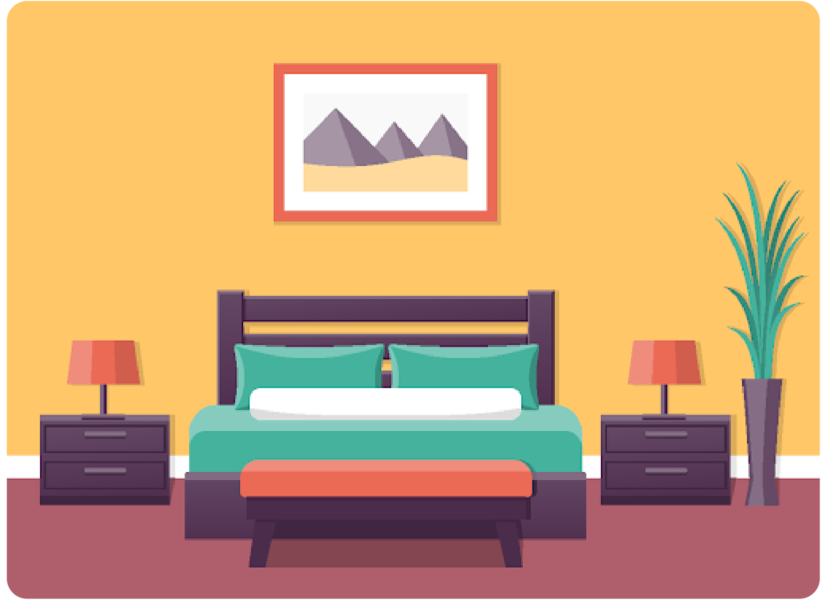


Number of rooms

Revenue generated

Decide the Variables

A strategically planned differential pricing system could ensure good room occupancy.



The use of data on rooms occupied helps to determine room occupancy.

The use of data on monetary values helps to assess financial performance.

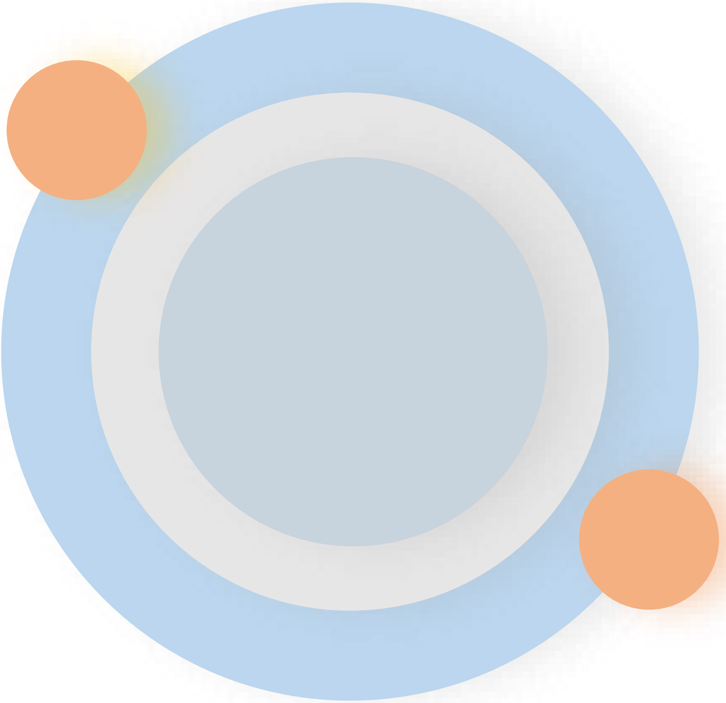
Both sets of data could be studied using separate histograms.

Link Between Data Collection and Chart Construction

Consider an example of wire diameters:

Link Between Data Collection and Chart Construction

There are two ways to assess the quality of wires produced in terms of diameters:



To formally measure and record the diameter of produced wires

To classify each wire as a piece whose diameter lies within or outside the specified limits

Sometimes, both approaches can co-exist.

Link Between Data Collection and Chart Construction

When two approaches co-exist:

Histograms are used to present data from measurements.

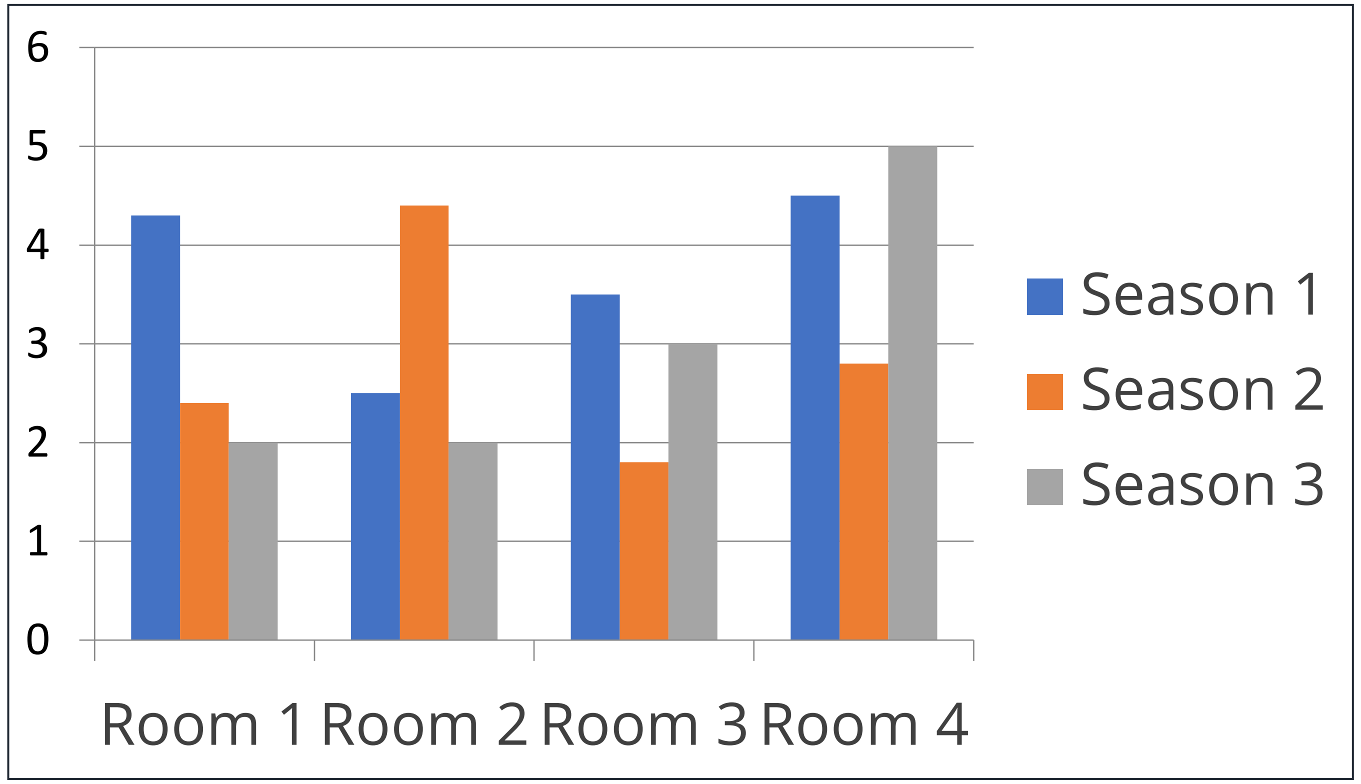
Bar charts are used to display data based on the classification.

Identify the Heterogeneity in Datasets

Determine the factors that could make datasets heterogenous and avoid using such data directly

Identify the Heterogeneity in Datasets

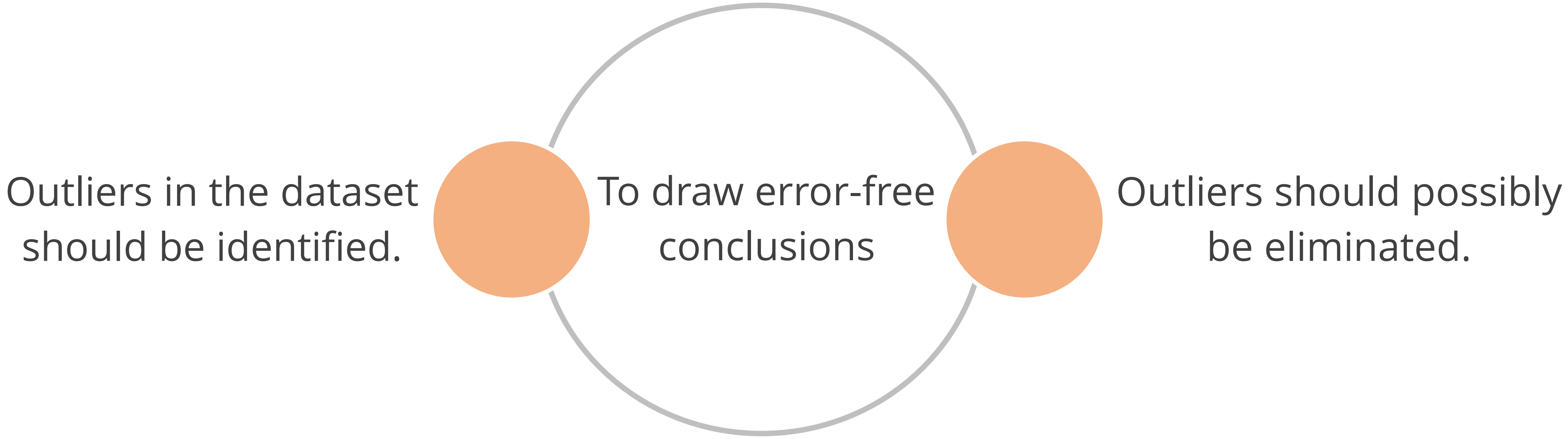
Example: Room occupancy varies based on the season.



Separate histograms could be used for different seasons.

Treating Outliers

Charts should accompany descriptive information when outliers are present to improve clarity.



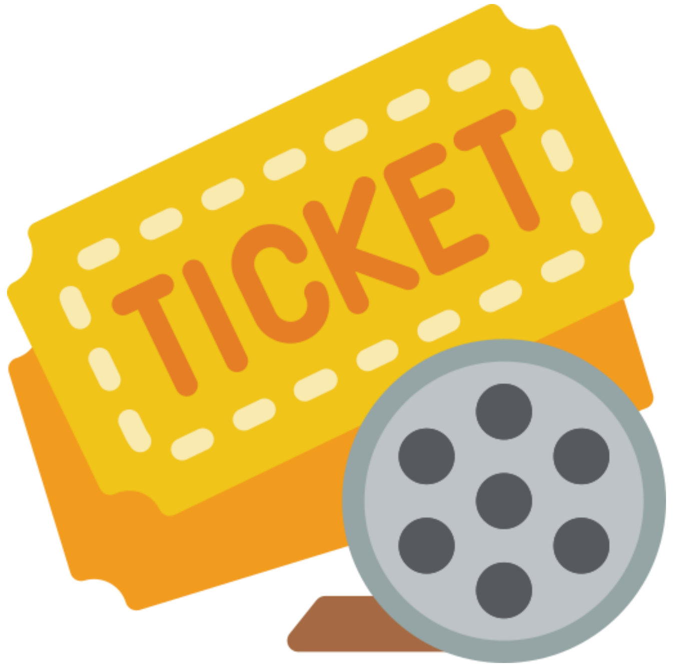
Outliers in the dataset should be identified.

To draw error-free conclusions

Outliers should possibly be eliminated.

Treating Outliers

Example: A drop in film viewership owing to a transport strike is not indicative of the demand for tickets.

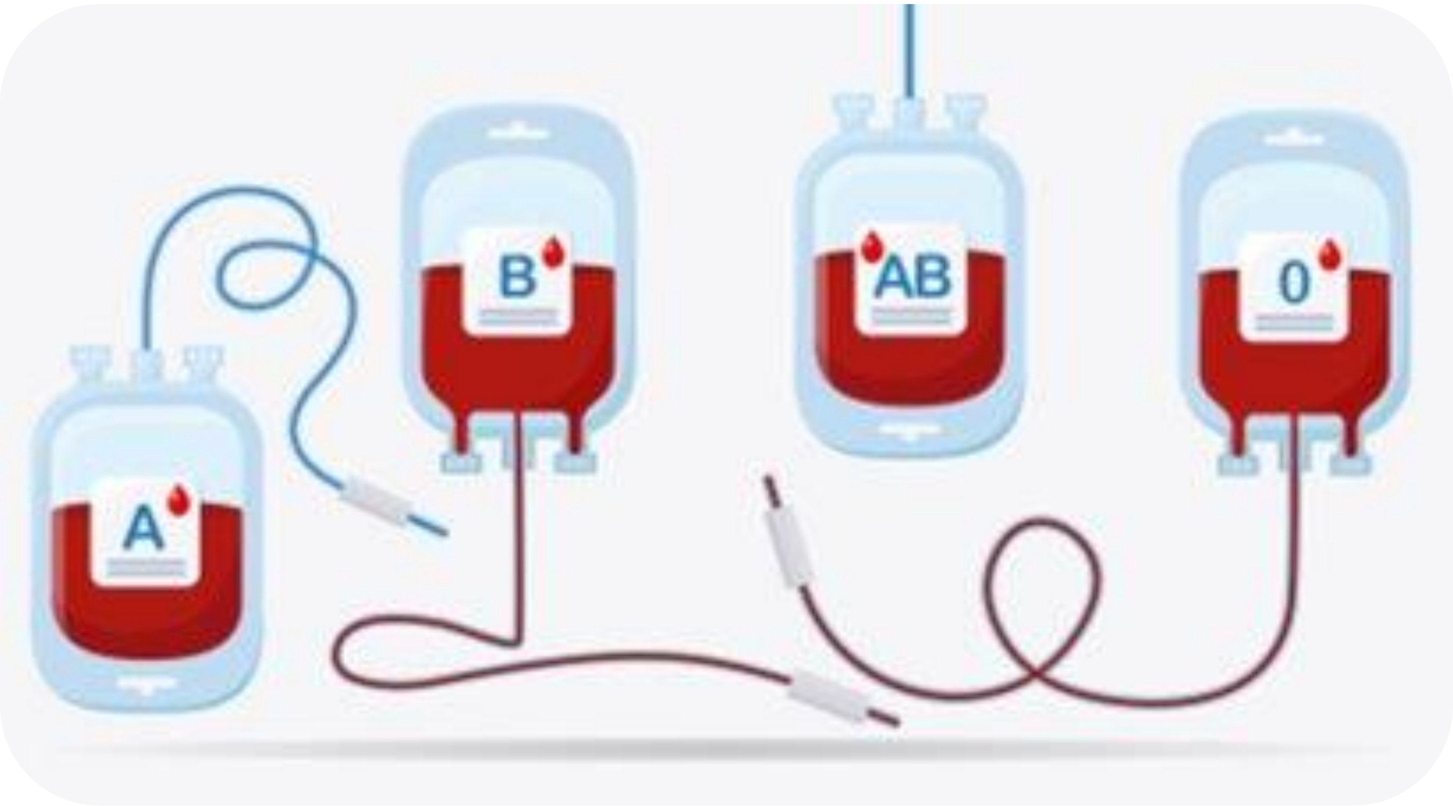


Such information could also be communicated to a viewer.

Data for Quality Conclusions

To draw quality conclusions, use a reasonable amount of data

Example: Data to predict demand for blood groups



Data for Quality Conclusions

For the viewer to have an idea of the reliability of the information portrayed, the following terms can be included:

Sample size used

Period of data collection

Data Collection and Chart Construction

After a thorough understanding of the context and the scope of the study, plans should be synchronized for:

Data collection

Chart construction

Discussion

Duration: 20 minutes

- What are outliers in data?

Outliers in data are values that significantly differ from most other observations in a dataset. They often result from measurement or input errors but can also indicate rare events or significant deviations from a general trend. If not properly addressed, outliers can skew statistical analysis and modeling, leading to inaccurate conclusions.

- How do outliers affect the data?

Outliers can greatly influence the mean and standard deviation of a dataset, causing these measures to inaccurately reflect the central tendency and spread of the data. They can also distort the results of data modeling and skew statistical tests, potentially leading to false conclusions.

Case Study: Deciding Variables

Case Study: Deciding Variables

Every data has stories to depict.

Data visualization is recognized as the process of displaying data to provide insights that will support better decisions, that is, telling the story behind the data.

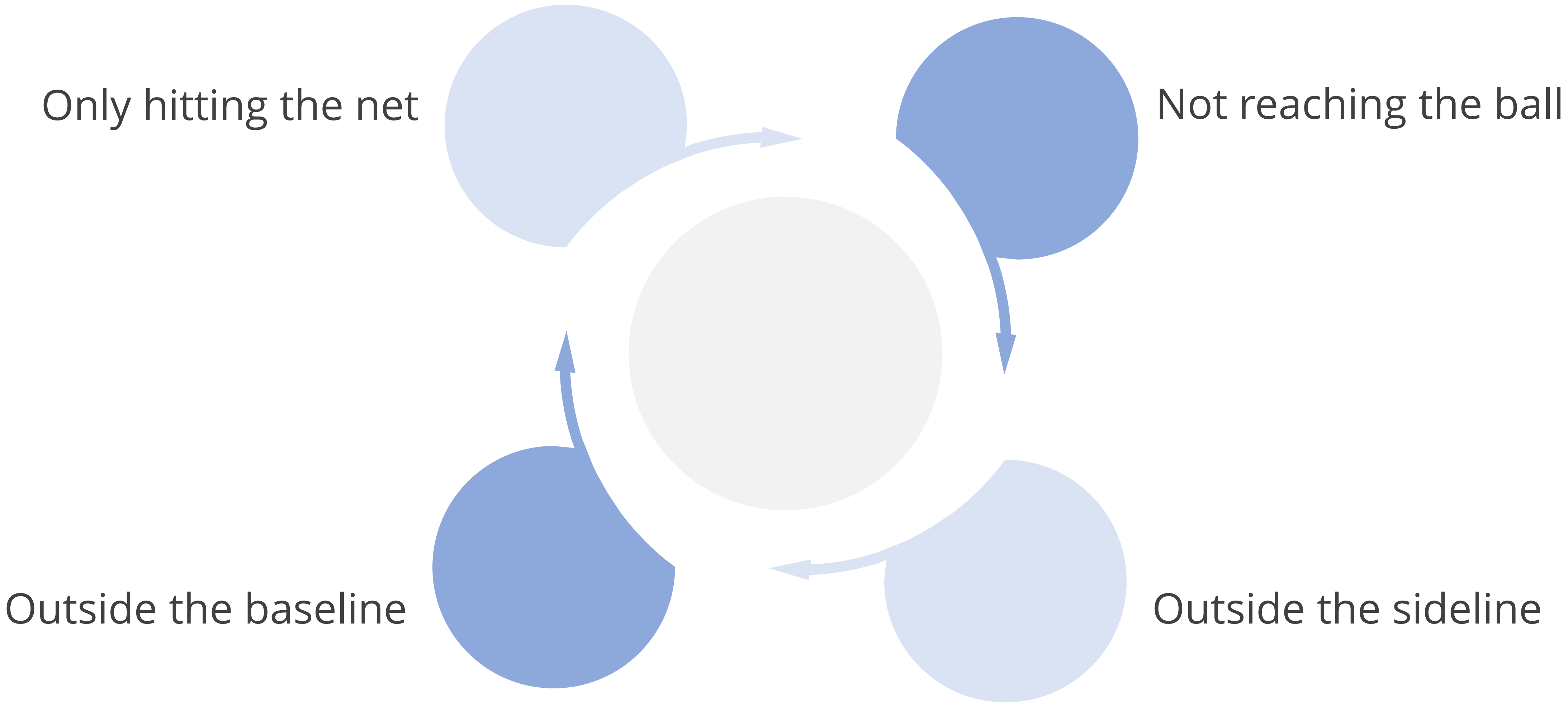
Case Study: Deciding Variables

Example: Representing the performance of a tennis player

The coach observed that a tennis player's performance was unsatisfactory.

Case Study: Deciding Variables

The coach noted how frequently the player was:



Case Study: Deciding Variables

The coach focused on a few frequently occurring faults and offered suggestions.

The player practiced regularly.

The player incorporated the suggestions.

Case Study: Deciding Variables

On observing again, the frequency of faults had decreased in the player's performance.

Analysis of the data revealed progress.

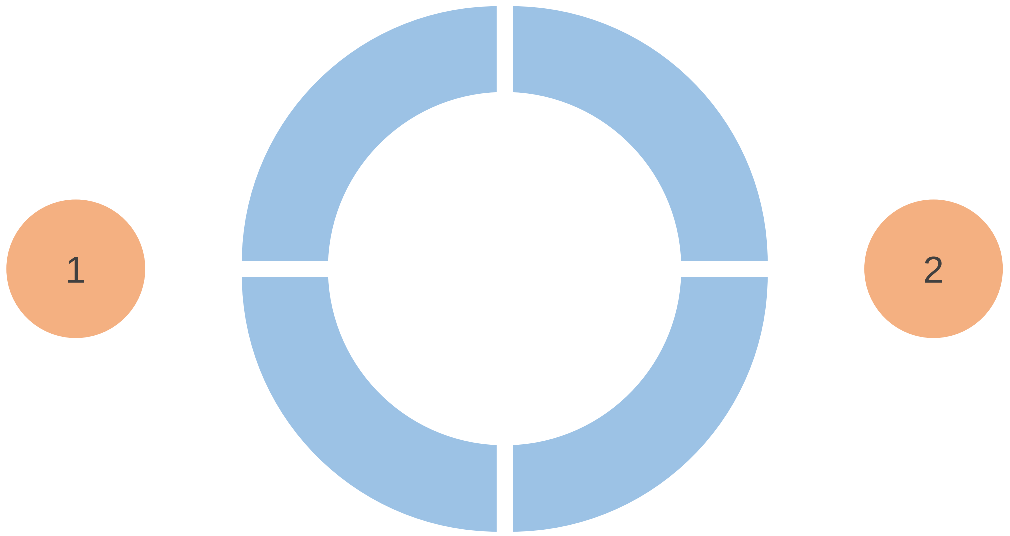
Case Study: Deciding Variables

The coach wanted the statistician to present the results in an illustrative way to promote his services to potential players.

The statistician felt bar diagrams could be a powerful way of highlighting this data.

Case Study: Deciding Variables

In the coach's visualization for each fault, the two rectangles were positioned adjacently.



Frequency of faults before the suggestion

Frequency of faults after the suggestion

The differences in the height for each type of fault communicated the improvement.

Case Study: Data Visualization

Case Study

Investigate the use of data visualization to analyze the number of journals and publications by the physics departments of three universities

University A

University B

University C

Tasks to Perform

The table shows the data collected from the three universities:

University code	A	B	C	A	B	A	C
Journal code	I	I	I	II	II	III	II
No. of publications	3	2	7	6	7	6	9

University code	A	B	A	B	C	A	C
Journal code	IV	IV	V	V	V	VI	VI
No. of publications	3	2	7	6	7	6	9

Tasks to Perform

Use data visualization to construct bar diagrams and derive insights from the collected data

Attribute Data

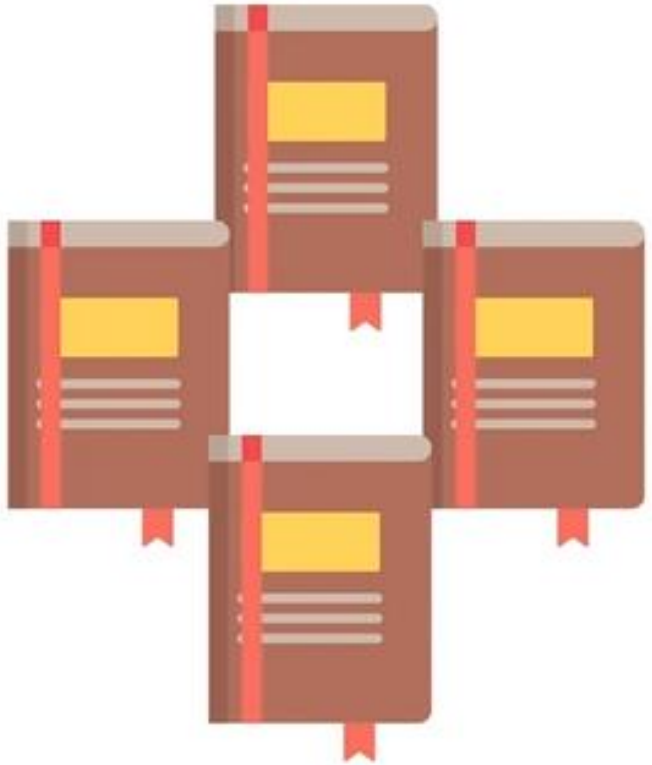
It refers to data that can be classified and counted.

University code

Journal code

Solution

Data visualization helps in analyzing data.



Effective data visualization enables the analysis of the number of journals used by each University to disseminate its research work.

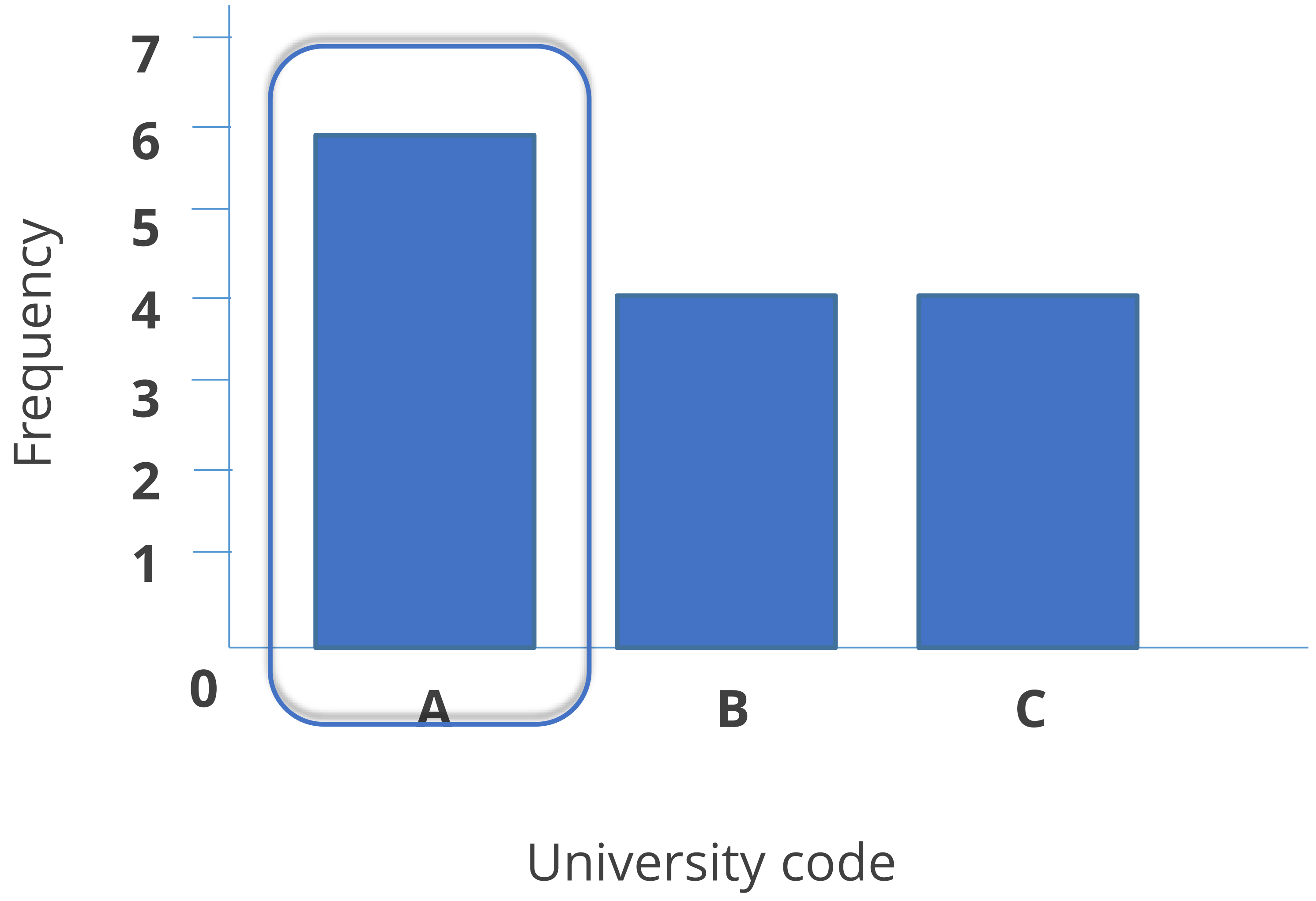
Frequency Distribution Table I

The first frequency distribution table has the number of journals used by each University to disseminate the research as shown:

University code	Frequency
A	6
B	4
C	4
Total	**14**

Bar Chart for Number of Journals Used

It is evident from the bar chart that University A used the highest number of journals to disseminate its research.



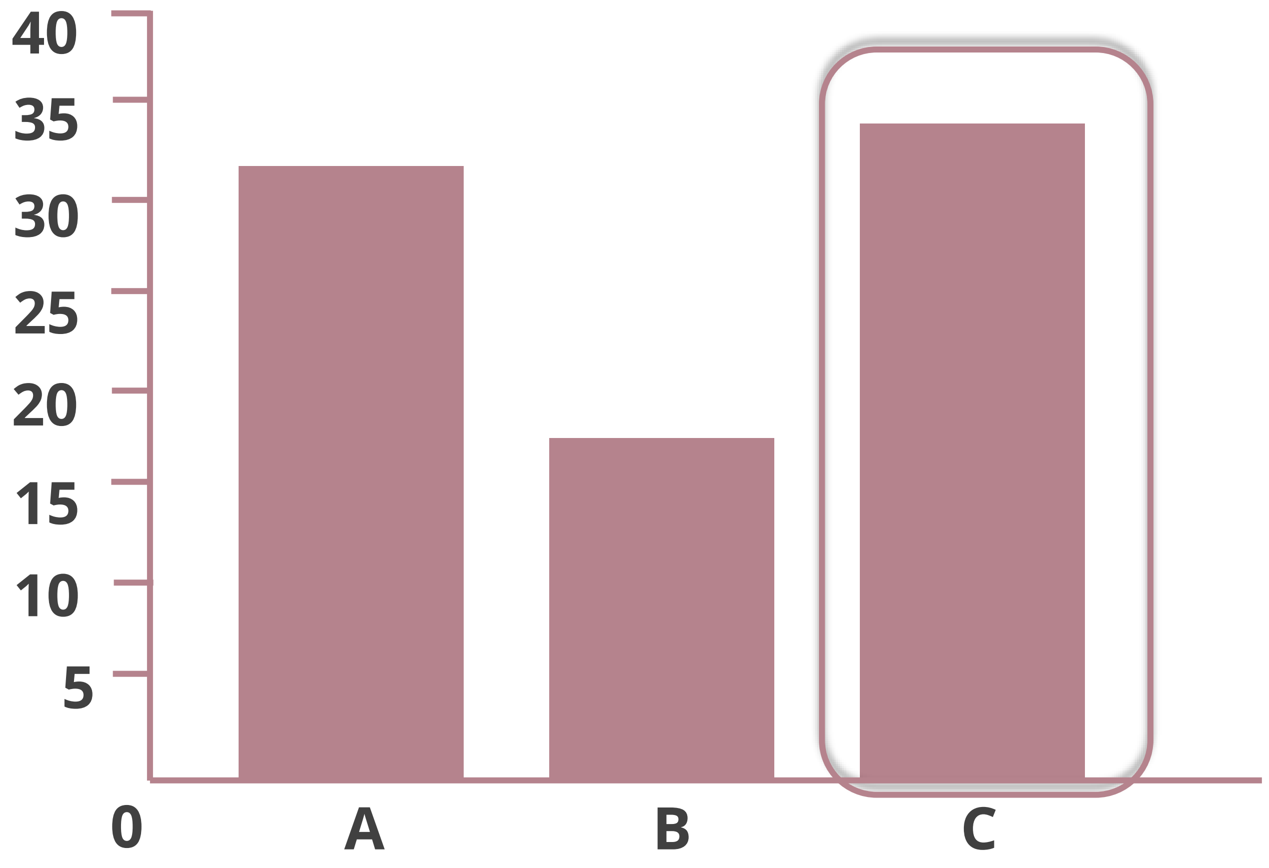
Frequency Distribution Table II

The second frequency distribution table depicts the total number of publications incorporating all the journals is as shown:

University code	Frequency (no. of publications)
A	31
B	17
C	32
Total	**80**

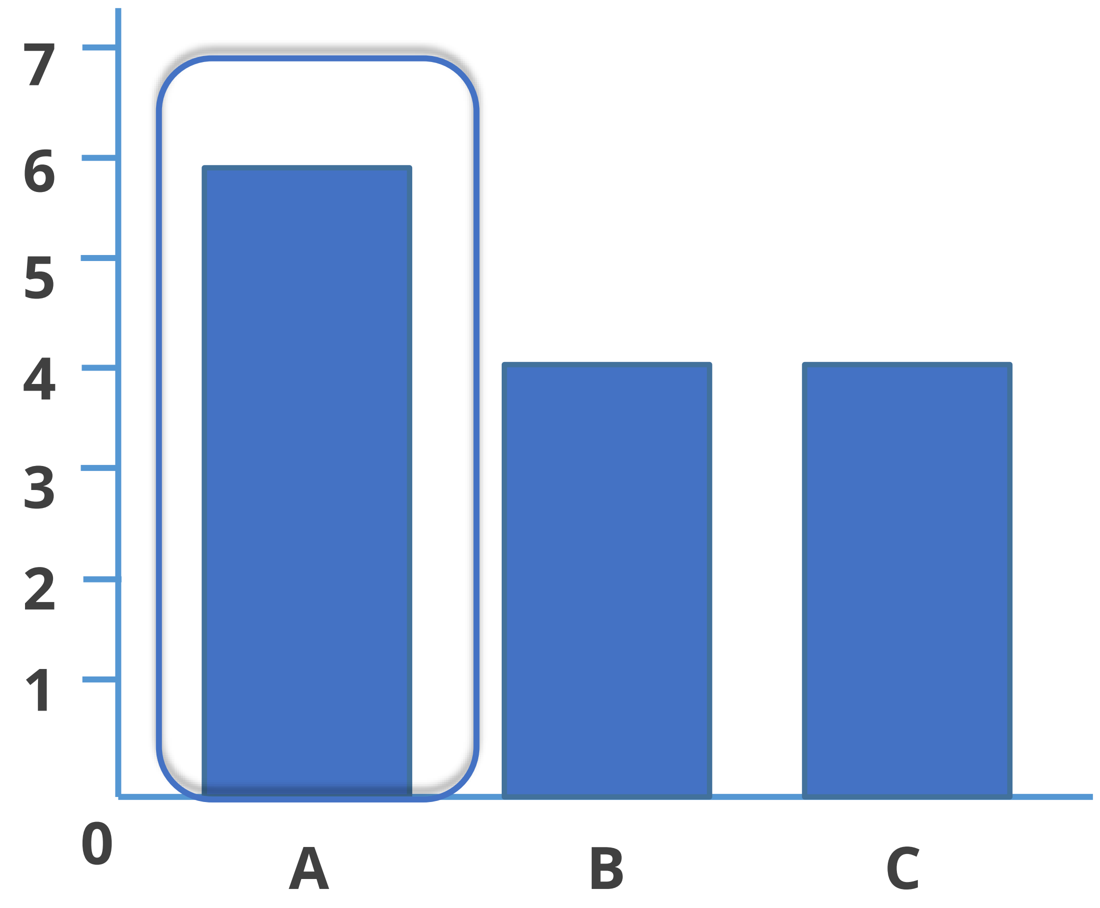
Bar Chart for Number of Publications

This bar chart clearly shows that University C has the highest number of publications.

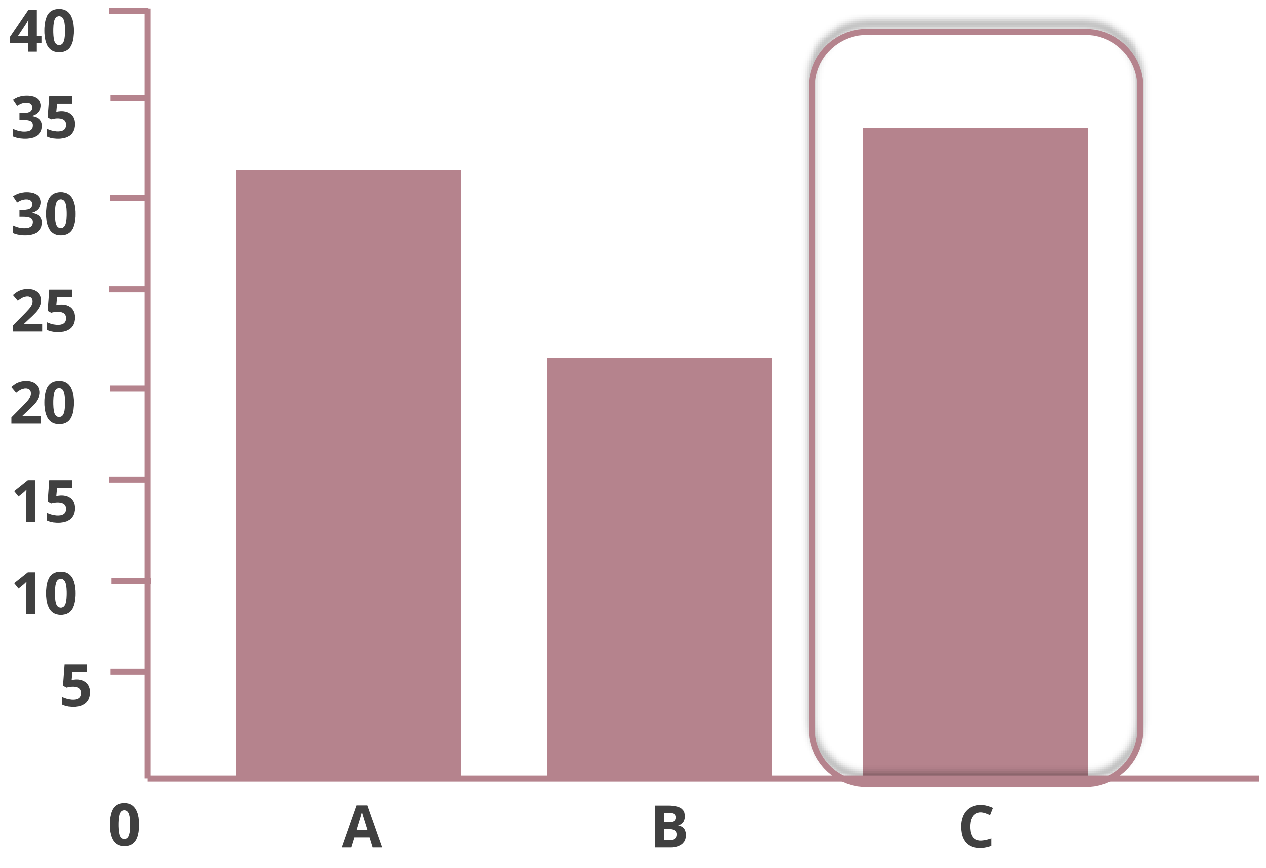


Comparison of Journals and Publications

The following bar charts clearly show that the total number of papers published by University A was fewer than that of University C.



Bar chart for the number of journals



Bar chart for the number of publications

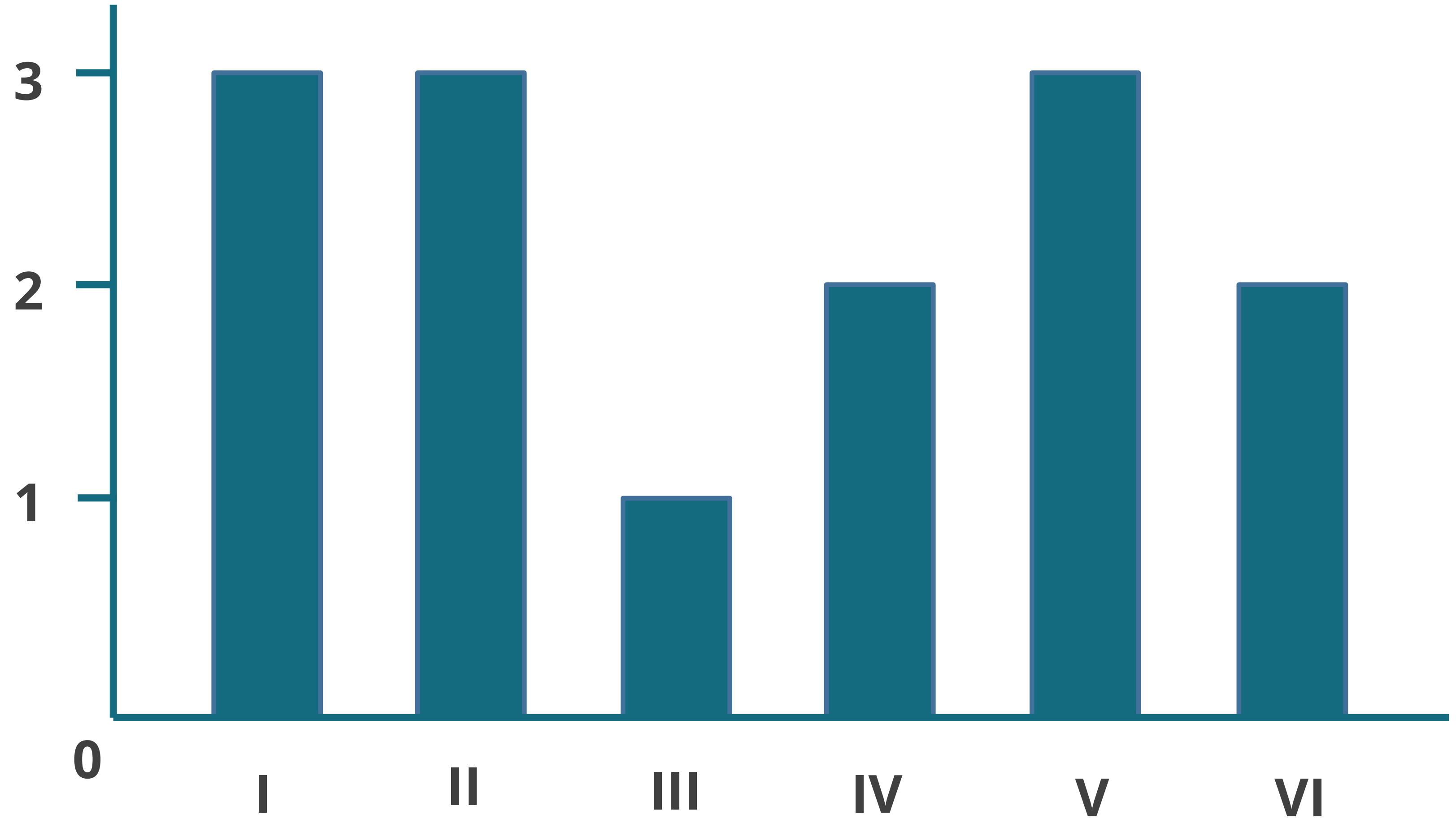
Frequency Distribution Table III

The third frequency distribution table depicts the total number of universities with publications in each journal.

Journal code	Frequency (no. of universities)
I	3
II	3
III	1
IV	2
V	3
VI	2
TOTAL	**14**

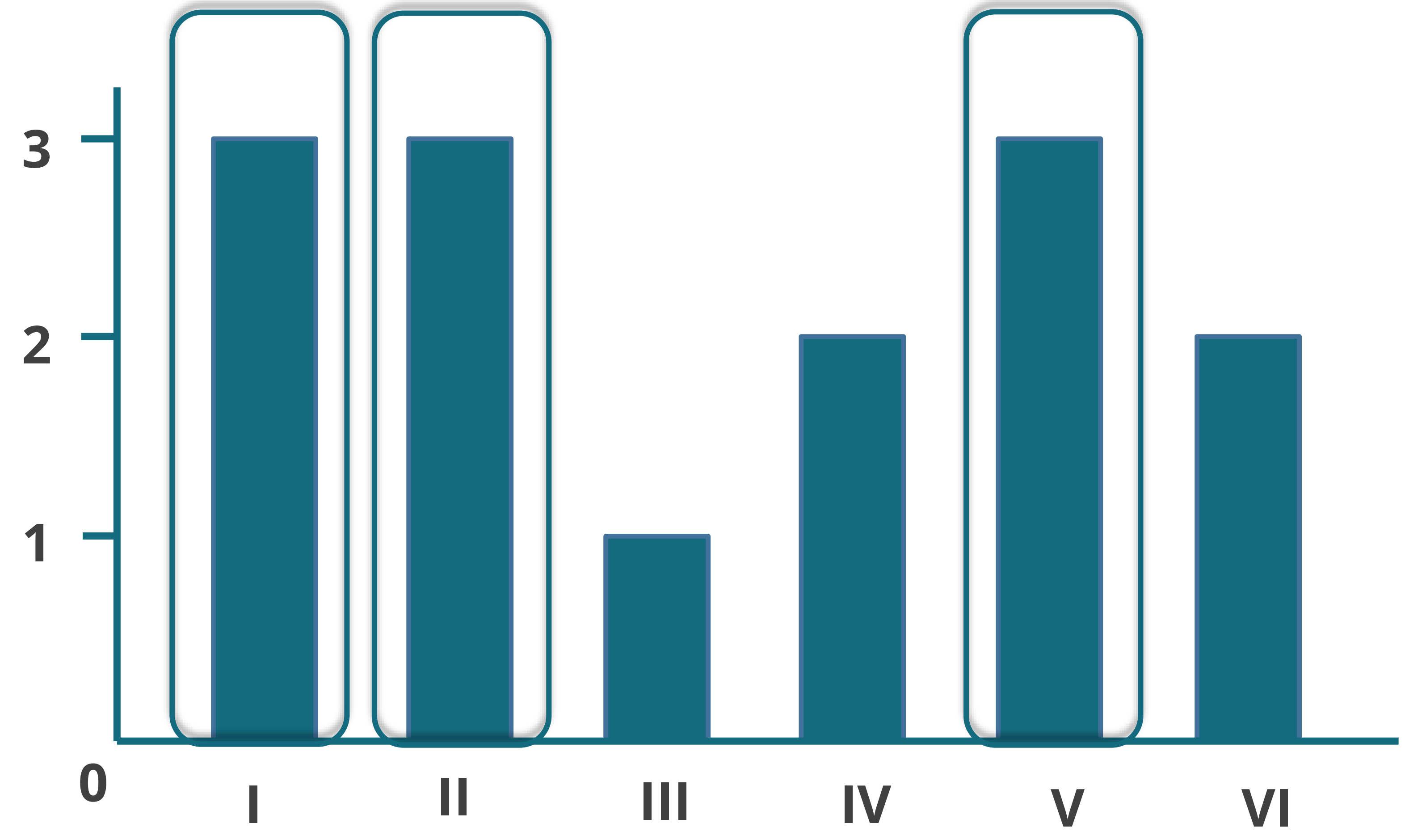
Bar Chart for Number of Campuses Using All Journals

The bar chart illustrates the publication count for different universities across various journals, indicating their utilization of each journal for research dissemination.



Inference

The bar graph reveals that the first, second, and fifth journals have publications from all universities. Other journals have publications from fewer universities.



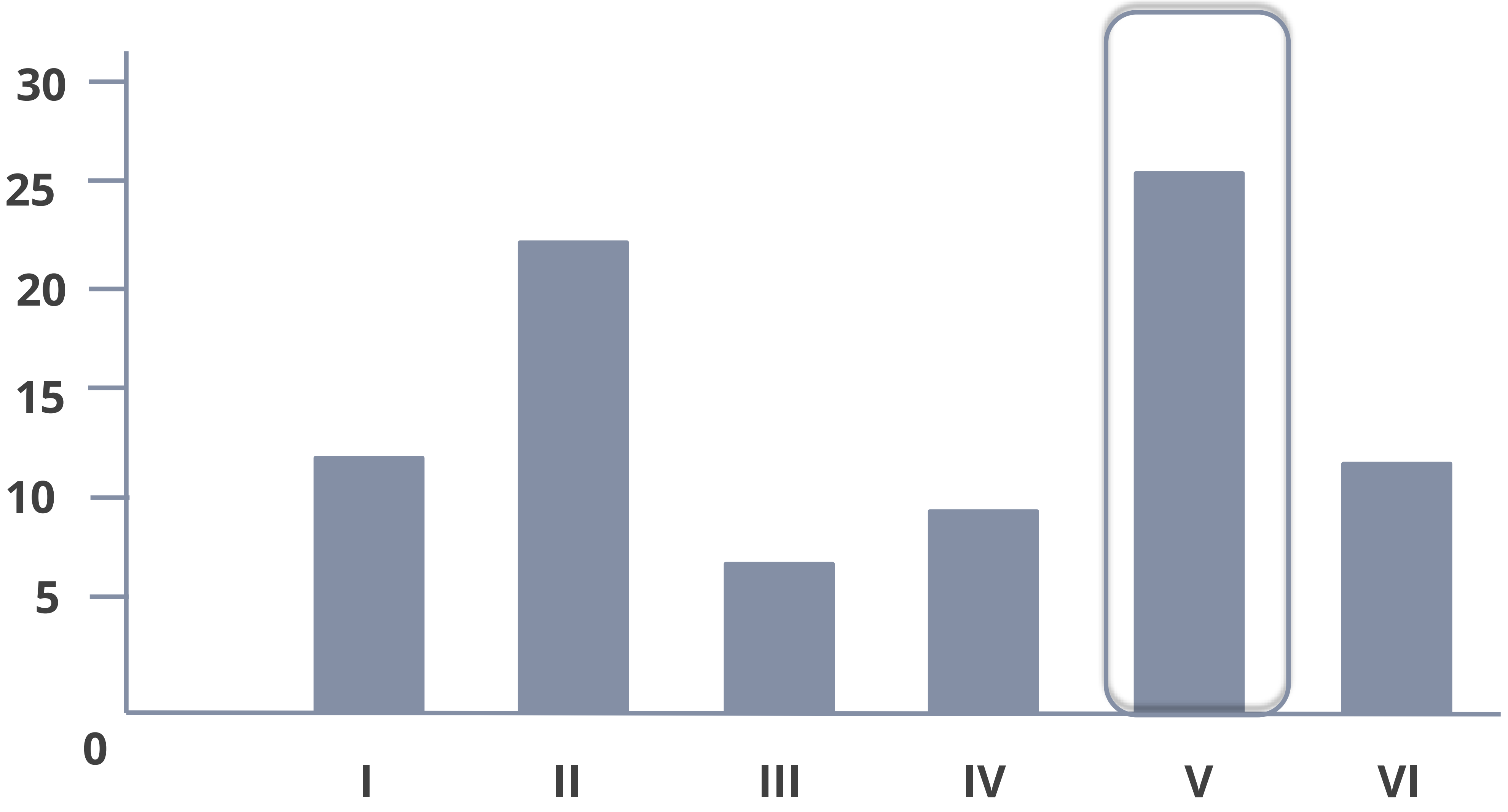
Frequency Distribution Table IV

The frequency distribution table highlighting the number of publications in each journal is as shown:

Journal code	Frequency (no. of publications)
I	12
II	22
III	6
IV	8
V	25
VI	11
Total	**84**

Bar Chart for Number of Publications in Each Journal

The bar chart displayed below illustrates the number of publications and the utilization of each journal for research dissemination.



Data Visualization

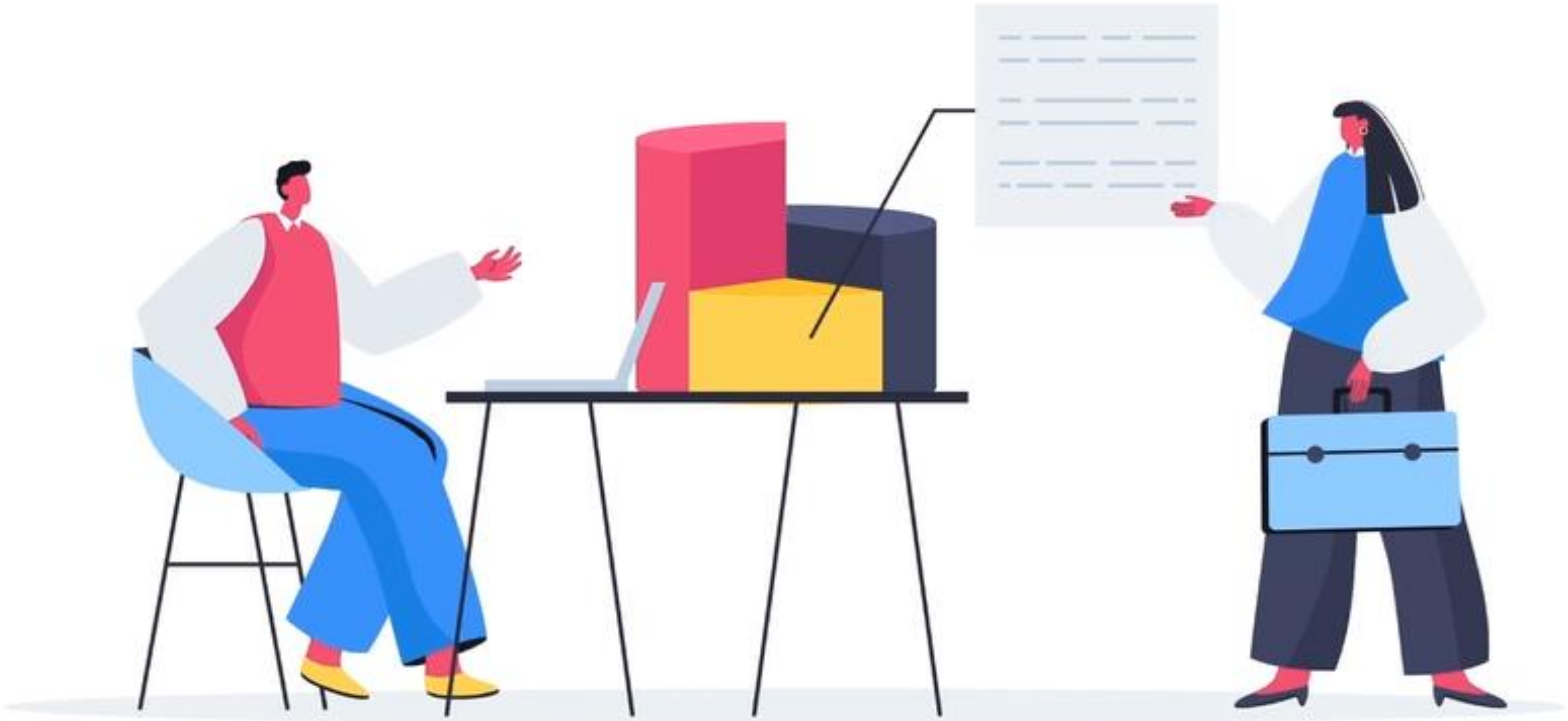
The four frequency distribution tables and the respective charts together illustrate that the data chosen must be based on the:

Issues to be addressed

Insights to be derived

Data Visualization

The data selected for the examination should facilitate insightful decision-making and provide comprehension of its effects.



When statistics and data visualization integrate, it improves exploratory data and enables users to generate significant discoveries.

Key Takeaways

- Data visualization tools help understand datasets.
- Commonly used charts include bar charts, pie diagrams, histograms, line charts, box plots, scatter plots, and bubble charts.
- Interpretation of charts involves constructing the charts and analyzing the data.
- Outliers in data are values that significantly differ from most other observations in a dataset.

Knowledge Check

Knowledge Check 1

Which of the following displays many pairs of observations to highlight the relationship between the two sets of data?

A. Bar chart

B. Box plot

C. Scatter plot

D. Bubble plot

Knowledge Check

Which of the following displays many pairs of observations to highlight the relationship between the two sets of data?

A. Bar chart

B. Box plot

C. Scatter plot

D. Bubble plot

The correct answer is **C**

A scatter plot displays many pairs of observations to highlight the relationship between the two sets of data.

Knowledge Check

2

Which of the following is used to identify the relationship between three numerical variables?

A. Box chart

B. Bar plot

C. Bubble plot

D. Scatter plot

Knowledge Check

2

Which of the following is used to identify the relationship between three numerical variables?

A. Box chart

B. Bar plot

C. Bubble plot

D. Scatter plot

The correct answer is **C**

A bubble plot is used to identify the relationships between three numerical variables.

Knowledge Check

3

______________ are rectangles of equal widths that represent a set of attribute data.

A. Bar charts

B. Pie charts

C. Histograms

D. Box plots

Knowledge Check

3

_______________ are rectangles of equal widths that represent a set of attribute data.

A. Bar charts

B. Pie charts

C. Histograms

D. Box plots

The correct answer is **A**

Bar charts are rectangles of equal widths that represent a set of attribute data.

Thank You