DATA SCIENCE

GRADE IX

Version 1.0

Microsoft

DATA SCIENCE

GRADE IX

Student Handbook

ACKNOWLEDGMENT

Patrons

- Sh. Ramesh Pokhriyal 'Nishank', Minister of Human Resource Development, Government of India
- Sh. Dhotre Sanjay Shamrao, Minister of State for Human Resource Development, Government of India
- Ms. Anita Karwal, IAS, Secretary, Department of School Education and Literacy, Ministry Human Resource Development, Government of India Advisory

Editorial and Creative Inputs

- Mr. Manuj Ahuja, IAS, Chairperson, Central Board of Secondary Education

Guidance and Support

- Dr. Biswajit Saha, Director (Skill Education & Training), Central Board of Secondary Education
- Dr. Joseph Emmanuel, Director (Academics), Central Board of Secondary Education
- Sh. Navtez Bal, Executive Director, Public Sector, Microsoft Corporation India Pvt. Ltd.
- Sh. Omjiwan Gupta, Director Education, Microsoft Corporation India Pvt. Ltd
- Dr. Vinnie Jauhari, Director Education Advocacy, Microsoft Corporation India Pvt. Ltd.
- Ms. Navdeep Kaur Kular, Education Program Manager, Allegis Services India

Value adder, Curator and Co-Ordinator

- Sh. Ravinder Pal Singh, Joint Secretary, Department of Skill Education, Central Board of Secondary Education

ABOUT THE HANDBOOK

In today's world, we have a surplus of data, and the demand for learning data science has never been greater. The students need to be provided a solid foundation on data science and technology for them to be industry ready.

The objective of this curriculum is to lay the foundation for Data Science, understanding how data is collected, analyzed and, how it can be used in solving problems and making decisions. It will also cover ethical issues with data including data governance and builds foundation for AI based applications of data science.

Therefore, CBSE is introducing 'Data Science' as a skill module of 12 hours duration in class VIII and as a skill subject in classes IX-XII.

CBSE acknowledges the initiative by Microsoft India in developing this data science handbook for class IX students. This handbook introduces the concept of Collection and Visualization of data using real life examples. The course covers the theoretical concepts of data science followed by practical examples to develop critical thinking capabilities among students.

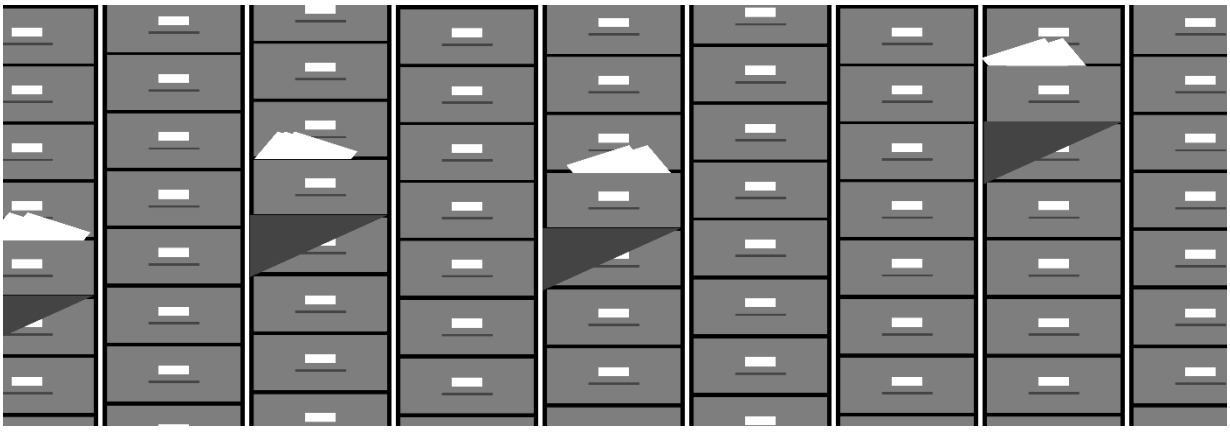
The purpose of the book is to enable the future workforce to acquire data science skills early in their educational phase and build a solid foundation to be industry-ready.

Contents

CHAPTER

1

INTRODUCTION



Studying this chapter should enable you to understand:

- What is data?
- Data vs Information
- DIKW model
- How data influences our lives?
- What are data footprints?
- Data loss and recovery

1. What is data?

Whatever we can read, write, speak, and observe is data. Every day we share so much data when we read, write, speak, and watch. It can be numbers, alphabets, symbols, or a combination of all of these. Examples – 100 (number – integer), 11.2 (number – decimal), A (alphabet), Apple (word – a set of alphabets), 18/12/2020 (date – a combination of numbers and symbols).

Nearly every action that we take in our daily lives generates data. Be it in the physical world or the digital world. Sensors, machines, mobile apps, websites are all around us, and every interaction we do with them generates a massive amount of data.

Data can be defined as facts or information which when stored, can be used as a basis for decision making, calculation, or discussion.

Data are used to tell a story. Statisticians see the world through data – data serve as models of reality.

Statistical thinking and the statistical problem-solving process are foundational to exploring all data.

2. Data vs. Information

Although the words "data" and "information" are frequently used interchangeably, they have different meanings. Data can be a number, symbol, or text, which may not mean anything to individuals on its own. However, when data is processed and put in context, they bear a meaning. This means data can be used for decision-making, calculations, or discussions. The data then becomes information.

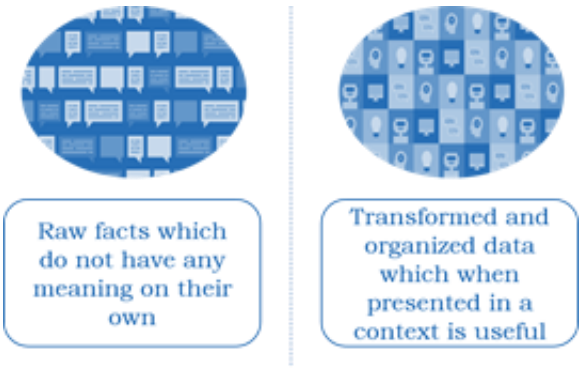


Fig 1.1 *Data vs Information*

For Example, if you are given a list of temperature readings, it would not make any sense. However, when the list is organized and analyzed, it shows that the global temperature is rising. This list now becomes information from data.

Some other examples are:

1. A list of dates, when coupled with the information that they are holidays
2. The number of COVID 19 patients per state when analyzed indicates whether the count is rising or not?
3. The word "apple" is data and the sentence "An apple a day keeps the doctor away" is information.

Fig 1.2 *Data to Information*

So, we can say **data** by itself has no meaning or purpose. It needs to be processed, organized, and analyzed to give it a meaning or purpose. This processed, managed, and structured data is called information. We can say that **information** is a collection of data that has a logical sense.

3. Data Information Knowledge Wisdom (DIKW model)

Data after transformation to information can also be converted to knowledge and wisdom.

This is called the DIKW model, which explains how we move from **D**ata to **I**nformation to **K**nowledge to **W**isdom.

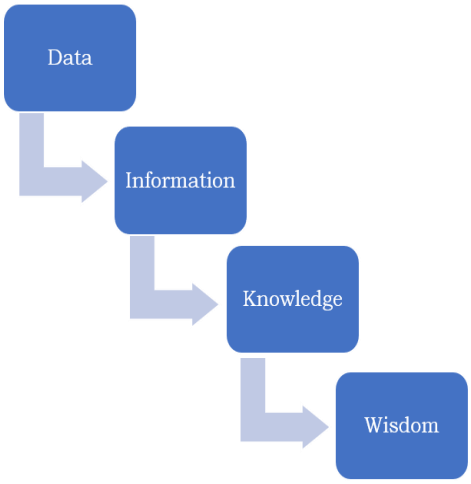


Fig 1.3 *DIKW model*

For Example, 100°C is a piece of information, but it is also the water's boiling temperature. This then becomes **knowledge**. Also, if you touch boiling water, you may get burnt. This now becomes **wisdom**.

Thus, we see data itself does not have much importance. It is the analysis of the data that has actual value.

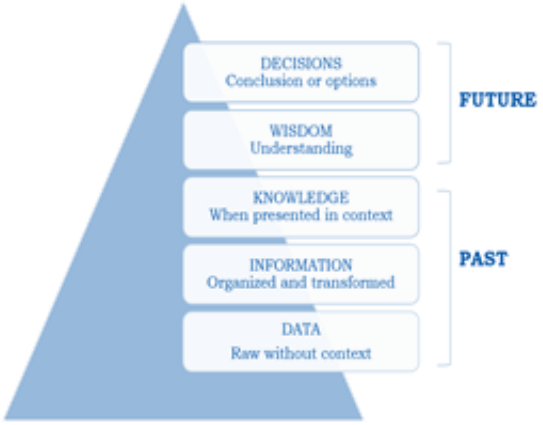


Fig 1.4 *DIKW pyramid*

DIKW model, also known as the DIKW pyramid, can also be represented as below:

4. How data influence our lives?

Data heavily influence our daily lives, starting from online shopping to watching our favorite shows on television to ordering food from the best restaurants. Data and data analysis have a significant impact on our lives.

We are continuously provided with statistical information on sports, medicine, education, public opinions, economy, etc. By newspapers and other media sources. This information helps us make personal decisions and allows us to meet our social responsibilities. Data also helps us in maintaining our health. For Example, fitness bands are useful in tracking our heart rates and other fitness statistics. We can also make career decisions by analyzing various education options.

Data also enables business leaders to make decisions based on historical data. They analyze budgets, market demands, sales forecasts, etc., which helps them make informed choices.

Below are a few of the essential aspects of our lives that are impacted by data:

4.1. Healthcare

Data has influenced many areas of the healthcare industry.

a. Tracking the medical history of patients and health patterns
b. Prediction of disease transmission and epidemics
c. Maintaining treatment protocols and potential cures
d. Personal fitness devices like wrist bands etc.

4.2. Online shopping

Online sellers analyze customer's historical purchases, searches, etc., to come up with targeted marketing. They reach out to consumers with curated offers and advertisements. They also compare shopping patterns of other customers to suggest frequently bought items.

4.3. Education

Most of the college and school admission processes are now digitized. Also, students explore various career options by analyzing historical records of universities and educational institutes.

4.4. Travel

Several popular travel apps predict traffic congestion and help us plan our travel. Also, we analyze feedback from different travelers on hotels and resorts to plan our vacation. Some travel systems predict the cost of flights based on historical data. All of these are examples of data, and its analysis impacts travel in our daily lives.

Activity 1.1

Think of how data and analytics helps in:

- Fighting crimes
- News and information

4.5. Online shows

Fig 1.5 *Real life usage of data*

With the advancement in data analysis techniques, we are seeing most online streaming platforms recommending personalized content. They record previous watch history and determine

which actors, genres, concepts appeal more to the viewers. They also predict which shows will become popular. They also provide users with ratings and feedback based on the historical input of other viewers.

5. What are data footprints?

In our daily life, the internet has become an essential part. What happens when we use the internet? We send and receive data through the internet. With all the activities we do on the internet, we create trails of data.

These trails of data are called data footprints. It includes the websites we visit, the emails we send and receive, the messages we send and receive when chatting, etc. Every activity we do is like moving around on the sand and leaving our footprints. Every trace of data is recorded.

When you open an online shopping website and check few pairs of shoes, you visit some other website, not a shopping website. You will see that the ads that pop up on both sides of the website's main content will be shoes. Sometimes, the same kind of shoes that you are looking for. How do you think that is possible?

Data footprints can be classified into two categories.

5.1. Active

Fig 1.6 Data footprints

We regularly use several social media platforms and post images or content which are stored on the media. This is a form of active data footprint as we have knowingly shared information about ourselves.

This applies not only to individuals but also to business and corporate organizations.

5.2. Passive

Our browsing history, product searches may be stored by search engines. Organizations use these records for personalized marketing. This is an example of a passive data footprint.

Data footprints can also be created in offline mode. Examples of offline mechanisms are files, images, documents stored in our personal computers.

Thus, we see over a period; we create digital footprints that can be used to identify us as individuals. They also connect us with the rest of the world and the people around us. As technology advances, it is hard not to have a digital footprint. So, it is better to create a positive one.

***Fig 1.7** Data footprints connecting us with people around us*

6. Data loss and recovery

Data can be lost, corrupted, damaged, or deleted due to multiple reasons like transaction failure, system crash, or disk failure. The process of restoring inaccessible, lost, corrupted, damaged, or deleted data is called data recovery.

> ***Activity 1.2***
>
> Think of the positive and negative impacts of data footprints.

Data can be lost due to a system crash when the system stops abruptly. For Example, issues in the power supply may cause hardware or software failure.

Another reason for losing data can be a disk failure when the hard disk drives or storage drives fail. Examples include damage to the storage drive, formation of bad sectors, etc.

So, we see there could be quite several reasons for data loss. To prevent this, we should frequently back up our data. Large enterprise systems generally use backup data storage from where they recover the data in case of any loss.

***Fig 1.8** Types of data loss*

Similarly, as we learn to implement our own data analysis solutions, we should always keep in mind data loss and take preventive measures.

We are now nearly at the end of this chapter introducing us to data and how data impacts our daily lives.

Activity 1.3

Read about some computer virus and ransomwares that has recently created problems for organizations.

Activity: What do Ladybugs Look Like?

In this activity, we will collect, summarize, and compare data on what do ladybugs look like.

We will formulate three statistical investigative questions:

What do ladybugs usually look like?

How many spots do ladybugs typically have?

Do red ladybugs tend to have more spots than black ladybugs?

Collect data

You will be provided with secondary data in the form of a picture of ladybugs (See Fig 1.9). As you see the pictures, you will notice a variation in the number of spots on the ladybugs and their color.

You can now record information about the number of spots, the color of each ladybug pictured, or any other features you think might be relevant.

You need to create some data collection questions that will need to be answered for each ladybug to begin to answer the statistical investigative questions:

How many spots are on the ladybug?

***Fig 1.8** Ladybug*

What color is the ladybug?

What color are the spots on the ladybug?

The number of spots on a ladybug is an example of numerical variables from taking measurements or counting objects.
Numerical variables are also called quantitative variables.

The color of the ladybug is an example of a categorical variable. Data on categorical variables are observed according to their category, where the categories are mutually exclusive and jointly exhaustive, meaning they do not overlap and represent all possible observations.

Now you can select the categories to use for color: black, orange, and red.

For each of the photos, you will ask the data collection questions and record the information for the three variables: (1) color of the body, (2) number of spots, and (3) the color of the spots (if applicable).

You should note that ladybugs are symmetrical, so if you count the spots on one side, you know the number on the other side. The total number of spots is recorded.

Fig 1.9 *4x4 photo card of Ladybug*

Sometimes data are messy or not straightforward. For instance, some of the spots are very faded and do not look like a spot at all. It is essential you and your class decide what will count as a spot (e.g., whether all shaped markings and all spots along the margin of the hard wing case will be counted). These consensus discussions will help reduce the measurement error introduced by your classmate's recording information on the ladybugs you are viewing. You should understand the importance of collecting data consistently.

You might record the data you are collecting in a variety of ways. You could consider one variable and record the values for each ladybug as in the below table.

Ladybug#	Color of Body
1	R
2	O
3	R
....	....

Fig 1.9 *Data Table*

You might record all three variables at once. For example, you might record the answer to all three questions for each ladybug as in the below table.

6 R B	10 O B	16 R B	6 R B
10 O B	16 R B	18 R B	18 R B
20 O B	20 O B	4 B R	16 R B
0 R -	2 B R	4 B O	2 B R

Fig 1.10 *Table of ladybugs data card*

This shows an example of a possible table structure showing from left to right in the cell: number of spots, the color of the body (R or B or O), the color of spots (B or R or O). Each of the cells can be thought of as a data card, an organizational tool for data. You can begin to recognize the importance of

having a strategy that allows you to organize the data in a useful way.

Eventually, it would be best if you were looking to create a more productive way to organize the data. It would help if you created a data table where each observation is on a separate row. This could be done on a worksheet using paper and pencil or using technology.

Ladybug #	Number of spots	Color of Body	Color of spots
1	6	R	B
2	10	O	B
3	16	R	B
...	...	...	...

Fig 1.11 *Data table for ladybug card*

Analyze the data

You could use a picture graph to analyze the data. This allows you to keep track of which ladybug is being graphed. Now you can use another graphical representation for one quantitative variable, which is a dot plot. You should be able to match a ladybug to a dot on the plot. This is a meaningful connection, as a dot plot no longer allows an individual ladybug to be distinguished. Dot plots can be created by hand or using technology, and the horizontal axis typically represents the values of the variable under study.

To compare the number of spots on ladybugs of different colors, you might use multiple dot plots with the same scale stacked one on top of the other. See the next figure as an example.

Using either a single dot plot or multiple dot plots broken down by ladybugs' different colors, you can answer a series of analysis questions about the quantitative variable number of spots.

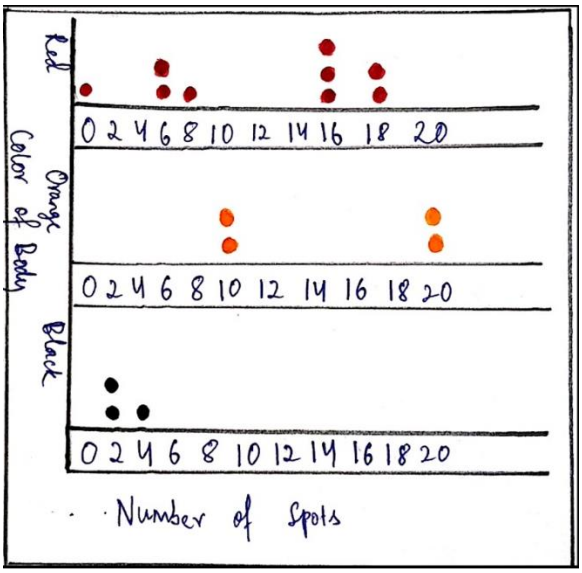


Fig 1.12 Drawn stacked dot plots of spot count for different color ladybugs

For example, such questions might include:

What number of spots were most common/ typical for all the ladybugs?
Red ladybugs only?
Orange ladybugs?
Black ladybugs?

What is the least/greatest number of spots for the ladybugs?
Red ladybugs only?
Orange ladybugs?
Black ladybugs?

It would help if you thought about distributing a quantitative variable and the variability in the values. You should understand that the median represents the value at the middle or center of a quantitative variable's distribution. A

median has the same number of data points (approximately half) are greater than and are less than it. The medians in figure 1.12 are 12 spots for the red ladybugs, 14 for the orange, and 2 spots for the black.

Interpret the results

Using the analyses, you can answer the statistical investigative question "What do ladybugs usually look like?" with an answer that might include the following:

Based on the pictures, the ladybugs in the card set were a mixture of red, orange, and black ladybugs. The red ladybugs had between 0 and 18 black spots. The orange ladybugs had between 10 and 20 black spots. The black ladybugs were different because they had either 2 or 4 spots in a mixture of colors.

Some of your classmates may write the following to answer the question, "How many spots do ladybugs in our card set have?":

Red ladybugs have between 0 and 18 spots. The most common number of spots is 16. The median number of spots for red ladybugs is 12 spots.

Orange ladybugs have between 10 and 20 spots.

The median number of spots for orange ladybugs is 14 spots. This is a bit higher than the red ladybugs.

Black ladybugs have two or four spots. Out of the three ladybugs, two had 2 spots, and one had 4 spots.

To answer the comparative question, "Do red ladybugs tend to have more spots than black ladybugs?" you may answer:

Red ladybugs have between 0 and 18 spots. The most common number of spots is 16. The median number of spots for red ladybugs is 14 spots.

Black ladybugs only have 2 or 4 spots. The median of spots for black ladybugs is 2. There is only one red ladybug with less than 6 spots; there is one with zero spots. All other red ladybugs have 6 or greater spots. These analyses suggest that black ladybugs in our pictures tend to have fewer spots than red ladybugs.

What did you learn?

- We are surrounded data. Every computer, every mobile device, every camera generates immense amount of data which greatly affects our daily lives.
- This volume of data when properly analysed becomes the basis of many innovations, technology advancements and actionable insights.
- Analysis of data leads to creation of information, knowledge and wisdom.
- Healthcare, online shopping, travel, education, online shows are some of the ways data influences our daily lives
- We create enourmous amount of data footprints by usng different online platforms
- There could be many ways data can be lost. So it is important to always keep a data recovery plan.

Exercises

Objective Type Questions

1) Data and information are the same
   a) Yes
   b) No
2) Social media platforms are responsible for creating data footprints
   a) Yes
   b) No
3) There is no risk of losing data
   a) Yes
   b) No
4) Websites and mobile apps use our search history to provide personalized offers
   a) Yes
   b) No
5) Which of the following is not in DIKW Model?
   a) Data
   b) Information
   c) Security
   d) Knowledge

6) Should you keep a data recovery plan?
   a) Yes

b) No

7) What is your data footprint?
   a) The data trail left by you when you surf the internet
   b) The time you spend on your computer
   c) The number of electronics you buy in a year
   d) The number of apps you have on your Mobile
8) How long is your data footprint visible?
   a) It depends on the websites you visit
   b) The data footprint wipes clean after a year
   c) It creates a permanent record
   d) The record expires after a month
9) Who can use or see data from your data footprint?
   a) It is visible to professionals, but they need special access to go through the data
   b) No one can access data from your digital footprint
   c) Only the police have access to the information on your data footprint
   d) Your data footprint is potentially visible to anyone
10) You regret posting a particular picture and want to take it down. Is it possible, and how would you do that?
   a) It is a little tricky but can be done by asking a professional to do it. Then no one can see the photo.
   b) You can delete the picture by clicking on the delete button. Then no one can see the photo anymore
   c) Only the police can delete a picture uploaded by you
   d) A photo can be deleted from your account, but someone might have already saved it or copied it.
11) How can you improve your data footprint?
   a) It is best not to post anything if you want to stay safe
   b) It is not necessary to improve your data footprint.
   c) Check your social media accounts' privacy settings to make sure you share your posts with people you trust and know.
   d) Share your personal details with a good friend or family member so they can help you stay safe online

Standard Questions

1) Explain the difference between data and information?
2) Give some examples of how data impacts our daily lives.
3) What are the different types of data loss?
4) What are data footprints? What are the different types of data footprints?
5) Explain the DIKW model.
6) Why should you keep a data recovery plan?

7) How do online streaming platforms use data?
8) What is personal data, and how can you keep your data safe online?

Higher Order Thinking Skills (HOTS)

1) Explain what data footprint is, where it is stored, how you can manage your data footprint, who can follow your digital footprint for your social media account where you post pictures.
2) Give a few examples of how your data can be lost and why it is essential to have a data recovery plan. Please make a list of all the data which are needed to be kept safe so that you do not lose it Example-Birth certificate

Applied Project

In this project, you can use a data set of popular dog breed characteristics to determine the best breed of dog for a hypothetical friend. You can consider various factors like the dog's lifespan, maximum weight, and whether the breed is good with children. You will collect, summarize, and compare data on finding the perfect dog breed.

CHAPTER

2

Arranging and collecting data

Studying this chapter should enable you to understand:

- What is data collection?
- What are variables?
- What are the different types of data?
- What are data sources?
- What is big data?
- Asking questions on data?
- Univariate vs multivariate

1. Introduction

In the last chapter, we have learned about data and how it influences our daily lives. In this chapter, we will understand how we can collect data. We will also learn about what is Big Data and what are the primary data sources.

As we see, demand for data insights is increasing exponentially, more and more industries are looking to adopt data science. As a result, data engineers need to collect large volumes of structured, unstructured, and semi-structured data. This data is subsequently used for analysis, prediction, and visualization. Sometimes insights can be derived from a single data source, but often data is needed to be collated from multiple relevant sources to make informed decisions. Let us now understand the data collection process in detail.

2. What is data collection?

The method of gathering data for calculating and analyzing reliable insights is known as data collection, which is done using standard validated techniques. A researcher or scientist

works based on the collected data. Data collection is a primary and essential step in most cases. The approach of data collection is different in different fields.

Fig 2.1 *Collecting data from different sources*

For Example, if we survey the temperature of many cities worldwide on the same day, the first important step would be to collect data on temperature from many towns. Let us assume we have recorded temperature across six cities at the same time. The temperature data collected are as follows.

City	Temperature
New York	0° C
London	7° C
Paris	9° C
Dubai	24° C
New Delhi	19° C
Tokyo	6° C

Fig 2.2 *Temperature recordings*

Now, if we represent this data in a bar chart, it will look like below.

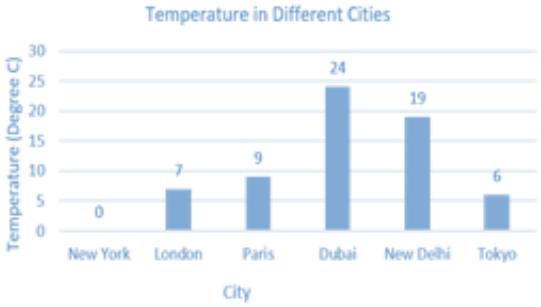


Fig 2.3 *Temperature bar chart*

Now that we have seen a primary data collection and visualization, let us now understand what variables are.

3. Variables

A variable is an attribute of an object of study that may vary for different cases. Thus, a variable varies for different case studies in research. Considering the previous example of the survey on the temperature of many cities worldwide on the same day, the variables are "Temperature" and "City" because both the attributes vary for different cases.

Now variables can be of two types.

3.1. Numerical variable

They represent values that have numbers. For Example, age, weight, height.

3.2. Categorical variable

These variables represent values that have words, for example, name, nationality, sport, etc.

For our previous example of the survey, "Temperature" is a numerical variable.

> ***Activity 2.1***
>
> Record your class attendance over a period of seven days. Draw a bar chart for the recorded data. What do you think will the variables for this?

4. Types of data

Data can be narrowly divided into two categories, quantitative and qualitative.

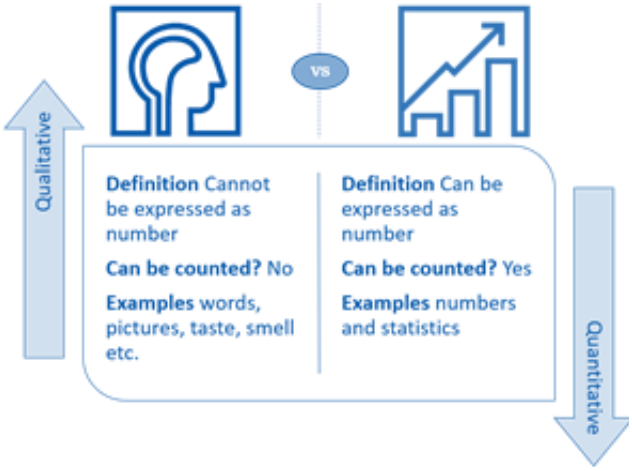


Fig 2.4 Qualitative vs Quantitative data

4.1. Quantitative data

Quantitative data are numbers or values that can be measured. For Example, the number of times a product has been searched on the internet or the number of items sold per month. Since these data can be quantified, they are comparatively easy to analyze.

4.2. Qualitative data

Qualitative data, on the other hand, is subjective. For Example, a traveler's review for a hotel or customer service feedback given by a consumer after a telephone conversation. These data help to understand experiences in depth.

5. Sources of data

Now that we know what data collection is, let us see the common sources of data. Data sources can be classified into primary and secondary sources.

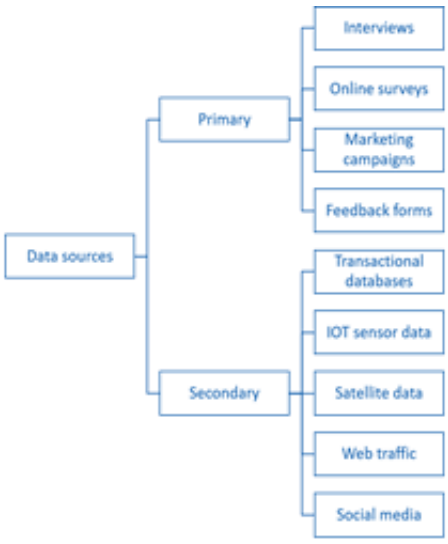


Fig 2.5 Data sources

5.1. Primary

These represent the sources created to collect data for analysis, for example, surveys, interviews, questionnaires, feedback forms, etc.

Some methods of collecting primary data are:

a) Physical interviews
b) Online surveys
c) Feedback forms

Activity 2.2

You have to collect data from farmers in rural areas to see if they are getting agriculture loans. What type of data collection will you use?

5.2. Secondary

At times data is already recorded for some other purpose but then re-used for analysis. These are secondary data sources. They include internal transactional databases, sensor data, etc.

Some methods of collecting secondary data are:

a) Social medial data tracking
b) Web traffic tracking
c) Satellite data tracking

Activity 2.3

You have to collect data from students in schools to see if they are enjoying the data science ciriculum. What type of data collection will you use?

6. What is Big Data?

Earlier in this chapter, we have seen multiple types of data and several data sources. Also, we have seen raw data needs to be transformed and analyzed to make it useful. This transformation process requires the processing of the data. When the data volumes exceed the processing capacities of traditional databases, they are called Big Data.

Think of the common social media platforms. Millions of users are using the platforms and creating an enormous amount of content every minute. Processing this vast amount of data requires specialized skills and systems. Such systems capable of extracting statistical insights from a huge amount of data are called Big Data systems.

Now let us understand some of the key characteristics that can define Big Data:

6.1. Volume

This refers to the size of the data. The size of the data generally determines if a data set can be termed Big Data or not. Usually, data sets greater than terabytes and petabytes are called Big Data

6.2. Variety

Big Data sets are generally collected from a wide range of sources, including transactional databases, sensor data, etc. It could include images, pictures, audio, video, etc. So, variety of data is an essential characteristic of Big Data.

6.3. Velocity

The rate at which data is generated. Big Data has generally created a rapid speed resulting in high volumes very soon. For Example, social media platforms generate a massive amount of data every minute.

Big Data techniques are widely used in different sectors. Let us see some of them:

a) Retail

Popular retail chains are spread across the world. They handle millions of customers every minute. They store and analyze their customer data and their transactions in Big Data systems.

b) Science

On the Discover supercomputing cluster, The NASA Center for Climate Simulation (NCCS) stores 32 petabytes of climate observations and simulations.

c) Sports

Race cars with hundreds of sensors produce terabytes of data in Formula One races. These sensors gather data points from tire pressure to fuel burn efficiency. Based on the data, data analysts and engineers decide whether modifications should be made to get the best outcome in the race. Moreover, based on simulations using data collected over the season collected through big data, race teams try to foresee the time they finish the race beforehand.

d) Social media

Most popular social media platforms store and analyze petabytes of data every day. They use Big Data techniques for storage and analysis.

e) Healthcare

During COVID 19 pandemic, different governments used Big Data to track infected people's locations to reduce the spread. It was also used for case identification and medical treatment.

7. Questioning your data

Previously we have seen multiple methods of data collection. Let us now see how we can interpret the data.

Data is typically stored as numbers (numeric) or labels (categories). Based on the type of data, we need to ask five simple questions to the data.

Question 1: Is this A or B?

Fig 2.6 Binary classification

Some questions can have only two possible answers. For Example

Q: Will a customer buy this product?

A: Yes/No

Q: Can India win this cricket match?

A: Yes/No

To predict this, a family of algorithms is used, and the mechanism is called **binary classification** or **two-class classification.** Similarly, if a question has more than two possible answers, then we use a **multiclass classification** algorithm.

Question 2: Is this odd?

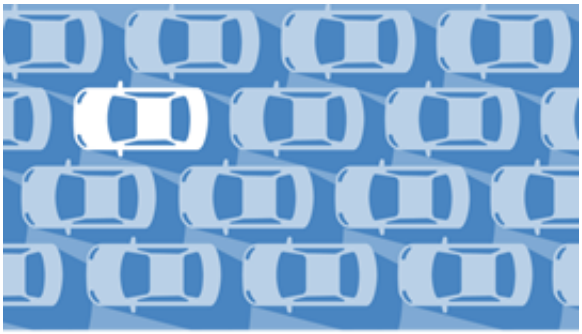


Fig 2.7 Anomaly detection

Sometimes we find unexpected records in a set of mostly consistent data. These are called anomalies and could be a cause of concern. For Example, if an unexpected transaction is done from your bank account, which does not match your regular transactions, there could be a case of fraud. Banks track these records and alert the customer that an unexpected transaction has happened, protecting the customer's money. Some other examples of anomaly detections are:

Q: Your father is getting his blood pressure checked. Is the reading regular?

Q: You are checking your car tyre pressure. Is the reading regular?

Algorithms used for these types of questions are called **anomaly detection algorithms**.

Question 3: How much or how many?

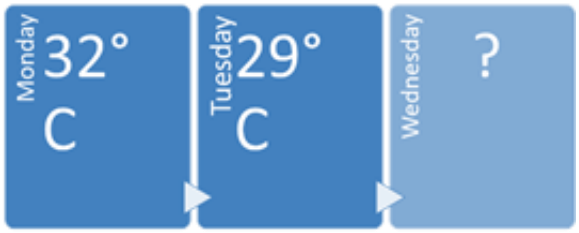


Fig 2.8 Regression algorithms

There are scenarios when we need to predict numerical values based on the data. Some examples are:

Q: How many goals will your favorite team score in this football match?

A: 3

Q: What will be the temperature of your city next Friday?

A: 32°C

The algorithms which predict these values are called **regression algorithms**.

Question 4: Can I group the data?

Sometimes data may be separated into distinct groups. This approach is called clustering. For Example, consider a class of 60 students. We have recorded their heights and arranged them in a table.

Height in cm	Number of students
130 - 140	20
140 - 150	12
150 - 160	18
160 - 170	10

***Fig 2.9** Clustering table*

As you can see, students can be categorized into groups based on their height. Similarly, if we plot these in a chart, it will look like below.

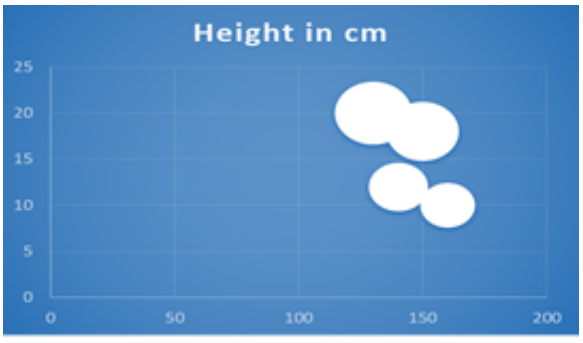


***Fig 2.10** Clustering chart*

Activity 2.4

Temperature recording for a month is what type of data?

Question 5: What should I do now?

Consider the following questions:

Q: I am a self-driving car. I am at a traffic signal with a red light. What should I do now?

A: Brake

Q: I am a micro-oven. I have already heated the food for the set timing. What should I do now?

A: Stop

These are questions that, generally, a machine or robot is programmed to do. Based on trial and error, machines take some actions. These types of learning are called **reinforcement learning.**

Below is the mapping between different algorithms and the corresponding questions

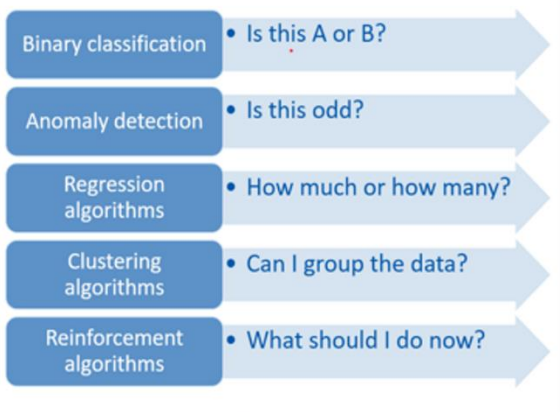


Fig 2.11 Different algorithms

Activity 2.5

Ice-cream sales increase during summer. Create a hypothetical table recording the sales of ice cream as temperature increase. What type data is this?

8. Univariate and multivariate data

Now that we have learned to question our data, let us see what univariate and multivariate data are.

Univariate:

This type of data has only one variable. They do not involve multiple parameters or relationships. For Example, the height of students is univariate data.

Student name	Height in cm
Student 1	143
Student 2	151
Student 3	155
Student 4	160

Fig 2.12 Univariate data

Multivariate:

This type of data involves a relationship between multiple variables—for Example, sales of umbrellas increase during the rainy season.

So, we see umbrella sales are dependent on rainfall. So, there are two variables – "rainfall" and "sales." These types of data are more complex than univariate as they involve comparisons and relations with multiple parameters.

Now, we have reached the end of this chapter. Let us quickly see what we have learned in this chapter.

What did you learn?

- Data collection is defined as the procedure of collecting data for measuring and analyzing accurate insights using standard validated techniques.
- There are two types of variables categorical and numeric.
- Data can be qualitative or quantitative.
- When data volume increases certain limits and specialized systems are required to manage the data, then it is called Big Data.
- Online surveys, inteviews, feedback forms are some of the sources of data.
- Binary classification, regression, anomaly detection, clustering are some of the algorithms used to predict based on the data.
- Univariate data has a single variable. Multivariate data has relationship with multiple parameters.

Exercises

Objective Type Questions

1. A school named ABC has recorded the total marks of every student in the class. This an example of:
    a. Qualitative data
    b. Quantitative data
    c. Both qualitative and quantitative data
    d. None of the above
2. A food delivery app has asked for your feedback on the quality of the food. You have written two paragraphs to describe the food. This is an example of:
    a. Qualitative data
    b. Quantitative data
    c. Both qualitative and quantitative data
    d. None of the above
3. You need to predict what the temperature will be for next Friday. Which algorithm will you use?
    a) Clustering
    b) Regression
    c) Anomaly detection

d) Binary classification

4. You need to predict if your car tyre will last for the next 1000 km. Which algorithm will you use?
   a) Clustering
   b) Regression
   c) Anomaly detection
   d) Binary classification
5. Which of the following options are the benefits of Big data processing?
   a) Business can utilize outside intelligence while making decisions
   b) Improved customer service
   c) Better optimal efficiency
   d) All of the above
6. The analysis of large amounts of data to see what patterns or other useful information can be found is known as
   a) Data Analysis
   b) Information Analytics
   c) Big data Analytics
   d) Data Analytics
7. Big data analysis does the following except
   a) Collects data
   b) Spreads data
   c) Organizes data
   d) Analyzes data
8. Primary data for the research process can be collected through
   a) Experiment
   b) Survey
   c) Both a and b
   d) None of the above
9. The advantage of secondary data are low cost, speed, availability, and flexibility
   a) True
   b) False
10. The method of getting primary data by watch people is called
   a) Survey
   b) Informative
   c) Observational
   d) Experimental

Standard Questions

1. What is the difference between multivariate and univariate data? Give some examples.
2. What are the common sources of data collection?

3. What are the primary characteristics of Big Data?
4. What are categorical variables? Give some examples.
5. How is Big Data used in social media?

Higher Order Thinking Skills (HOTS)

Collect data of 50 Motor vehicles passing by your house. You can record the number of wheels, the color of the vehicle and try to present the data as a dot plot. Compare the plots with your classmates. Find the color of the vehicle, which is most common, and the least common vehicle has how many wheels.

Applied Project

Create a Bar graph or dot plot using your classmates' personal information, like their birthday month. You put tally marks next to the months for each student's birthday. You can then come together and make a Dot plot of your class's birthdays based on the numbers you collected. You can discuss words like the most, the least, the total number of students, and so on. Then you can hang the plot up in the classroom all year as a reference point about data collection and bar graphs or using bars vertically or horizontally to show data.

CHAPTER

3 Data Visualizations

Studying this chapter should enable you to understand:

- The importance of visualization
- Plotting data
- Histograms
- Use of shapes
- Use of single and multivariable plots

1. Introduction

In the last chapter, we learned how data is collected and how we can interpret the data by asking several questions. We have also seen the enormous volume of data that gets generated from different sources. In this chapter, we will see how we can visualize the data and make it more comprehendible.

2. Importance of data visualization

Think of a flat table recording the daily attendance of your class. You have passed three years' worth of historical data. From this, you want to figure out which months have the highest attendance. It will be tough to find out from a flat table. This is where visualization of data comes to the rescue. If this data were plotted in a bar graph, it would be much easier for you to understand and explain to others.

Data visualization is the mechanism of representing raw data in the form of graphical representations that allow

users to explore the data and uncover quick insights.

Fig 3.1 Data visualizations

With so much information around us, it is challenging to view the data and derive insights from it. Representing data through visualizations like graphs, charts, maps, etc., gives us a visual context of the data. It also makes complex data simple and enables the human mind to understand its significance.

Visualizations allow us to recognize trends, patterns, and outliers from seemingly meaningless records of data. Data visualization techniques use visual data in a universal, fast, and powerful way to communicate information. This approach enables viewers, mostly business analysts and company executives, to determine which areas need to be changed, which factors determine customer satisfaction and customer dissatisfaction.

Visualized data gives a more precise prediction of revenue volumes and potential development for customers, company owners, and decision-makers.

Let us now look at a few real-life uses of data visualizations.

2.1. Tracking student progress with scorecards

Fig 3.2 Visualizing student reports

Recording student scores over a period helps analyze the progress of the students. Also helps understand their strengths and weaknesses helping teachers and parents provide better assessment for the students.

2.2. Identifying usage trend of a website

Imagine you are a website administrator. You want to identify the pattern when most people visit your website. To do this, you need to track

user visits and the time when they used your website. If you plot this in a bar chart, you will quickly identify at what time usage of your website peaks.

2.3. Monitoring goals and results of a sales executive

Fig 3.3 Visualizing sales records

Most sales executives in organizations have goals. Visualizing their sales records in charts and graphs helps to easily picture how close they are to their goals and what steps they need to take.

2.4. Visualizing spread and impact of pandemics

Pandemics like COVID 19 have impacted the entire world. Data visualization techniques help to identify the most affected countries or regions. They also show a trend if the spread increases or decreases, allowing governments and other bodies to take necessary actions.

Fig 3.4 Visualizing healthcare data

Activity 3.1

Think about how data visualizations can help hospitals track pateint health records and progress.

3. Plotting data

There are several different ways to visualize data depending on the data being modeled and its purpose. Several other graphs and tables can be used to visualize the data.

Here, we shall be using spreadsheets from open office to plot the data. Open office can be downloaded and installed using the URL: https://www.openoffice.org

3.1. Dot Plots

A dot plot is a graphical display of data using dots.

Dot plots were first used to describe distributions going back to 1884, hand-drawn (pre-computer era) graphs.

Dots are used in dot plots to illustrate the quantitative values associated with the categorical values.

For Example, Minutes to reach school.

Data shows how long does it take five people to reach the school.

Minutes: 6, 2, 4, 8, 5

Person: A, B, C, D, E

This data shows that it takes A six minutes to reach the school, B two minutes, etc.

Furthermore, here is the dot plot for the Example.

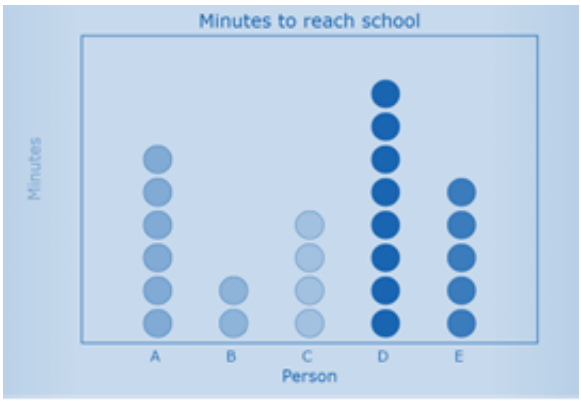


Fig 3.5 Dot plots

Below is how the data can be presented in table format.

Person	Minutes to reach school
A	6
B	2
C	4
D	8
E	5

Fig 3.6 Minutes to reach school

The same data can also be represented in the below chart.

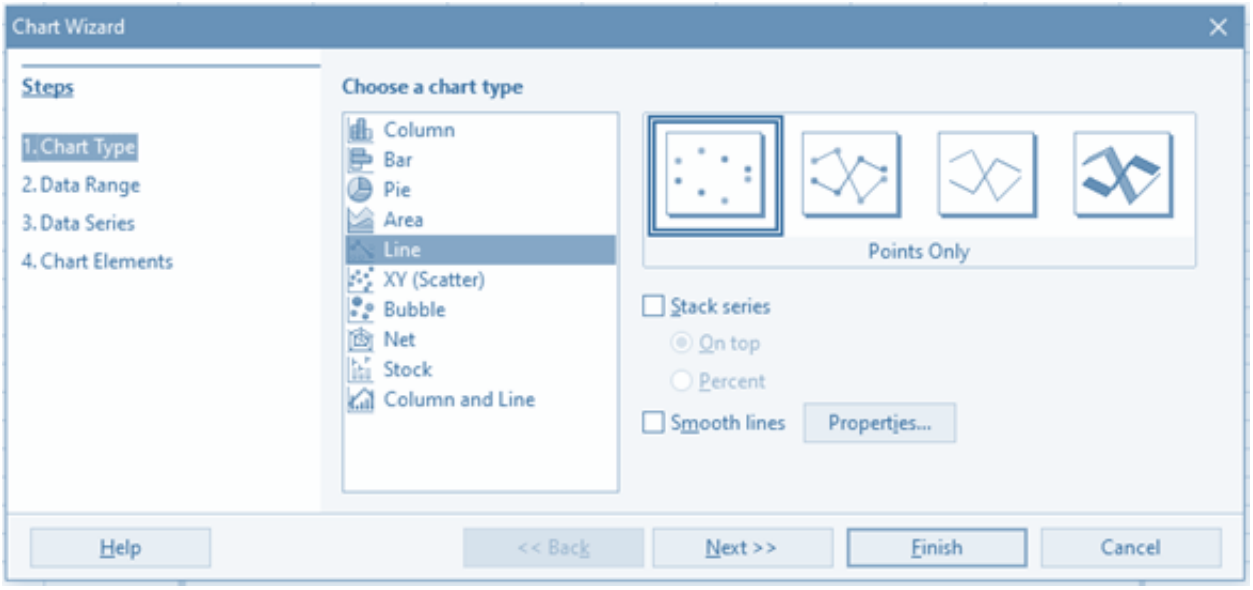
Fig 3.7 Minutes to reach school

To plot the chart using open office, a spreadsheet needs to be opened in open office and the table data needs to be filled.

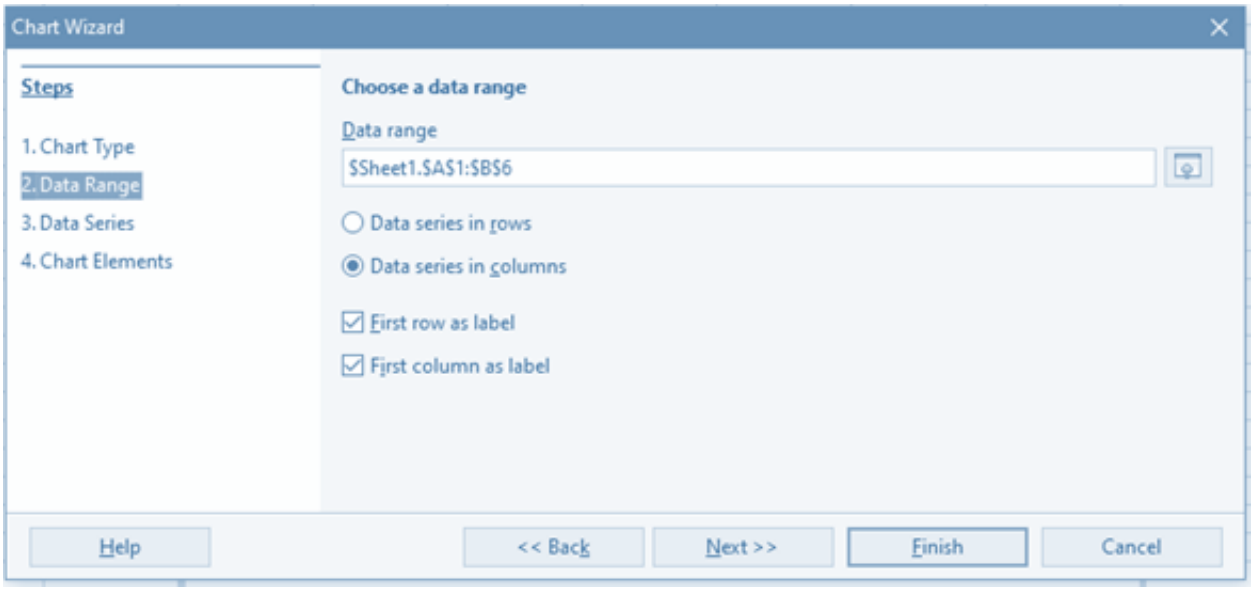
	A	B
1	Person	Minutes to reach school
2	A	6
3	B	2
4	C	4
5	D	8
6	E	5

We then need to go to Menu, click on Insert, and from the dropdown select **chart**. On the chart wizard, we select the chart type as **Line**, and amongst

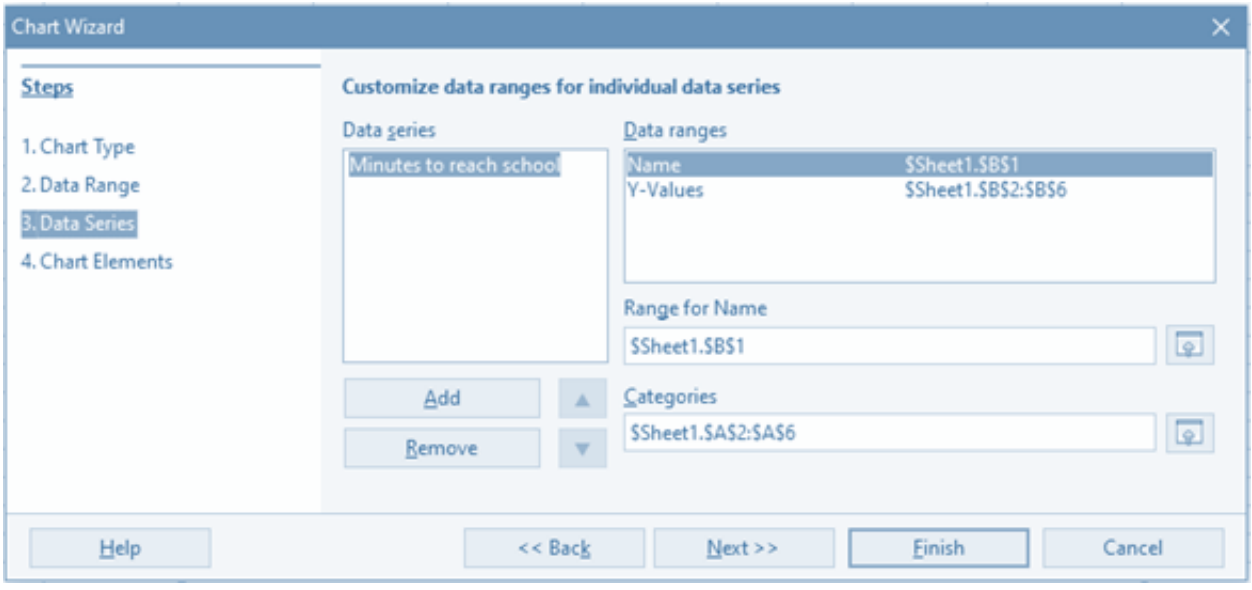
the choice of a line chart, we select chart type as **points only**.



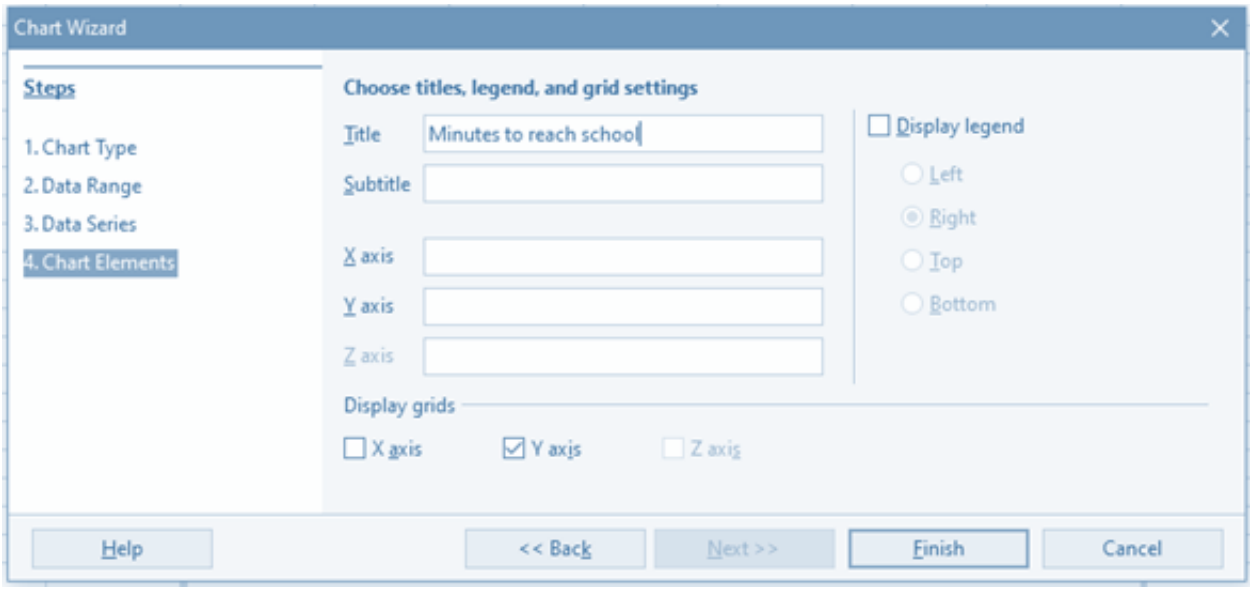
In the **data range**, select the entire table populated above, in this case, cell A1 to B6.



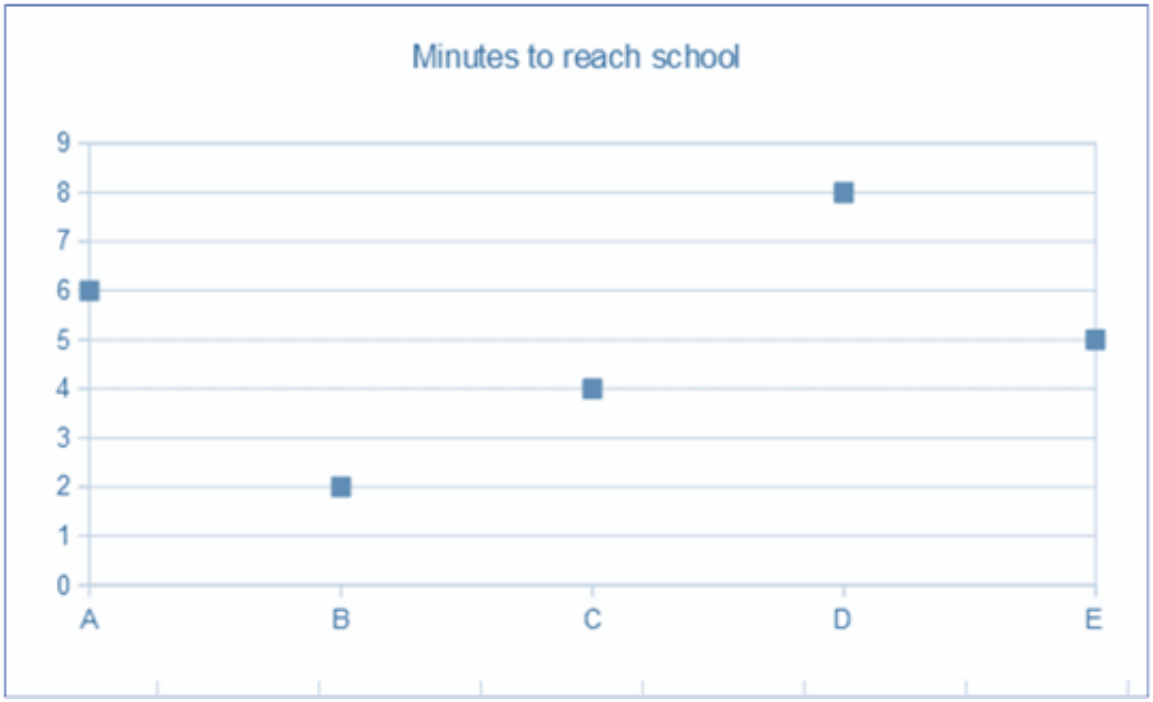
Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click on next.



On the **Chart Elements** section, we may provide the title, subtitle, name of the x-axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart, and if yes then where do we want to display the legend.



When the finish button is clicked, the chart gets generated as shown below:



3.2. Bar Graph

A bar graph is a graphical display of data using bars of different heights. It is possible to plot the bars vertically or horizontally.

A vertical bar graph is called a column chart or graph.

In a bar graph, the bars are presented to show elements so that they do not touch each other.

For Example,

You can see the students' list for some elective subjects in a school in the below data.

Subject	Number of students registered
Science	28
Maths	32
Hindi	44
English	25
Physical Education	14
Sanskrit	34
Art and Craft	22

Fig 3.8 *Number of students registered*

Let us now see how this data looks when plotted in a bar graph.

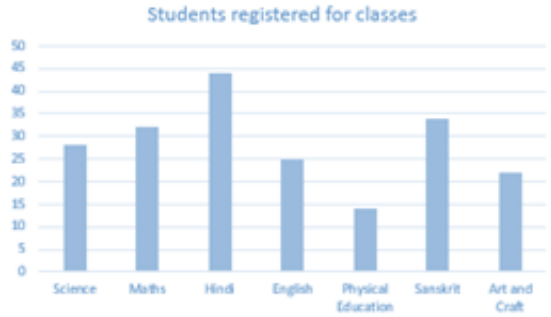


Fig 3.9 *Bar Graph*

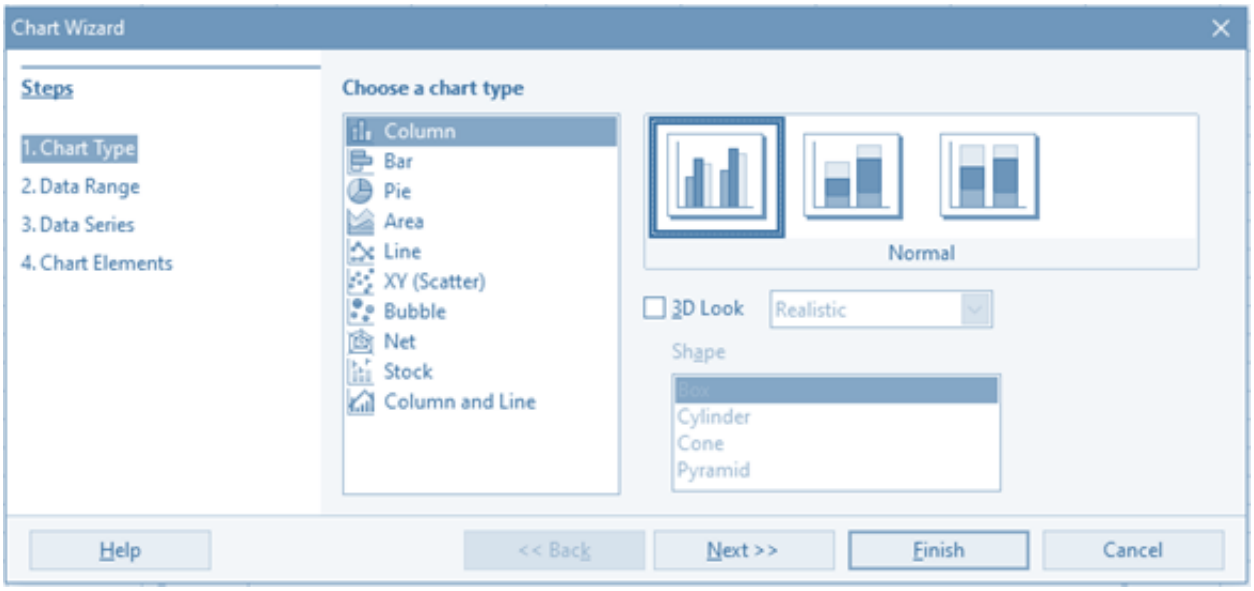
The same data, when plotted in a column chart, will look like the below.

Fig 3.10 *Column Chart*

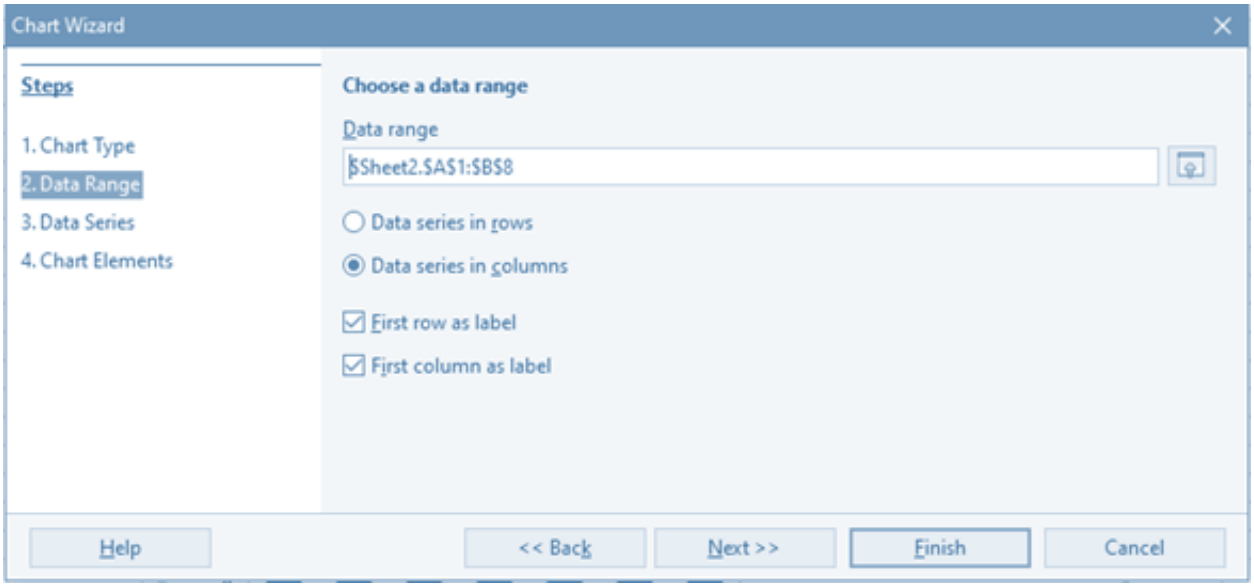
Bar Graph: To plot the bar graph using open office, a spreadsheet needs to be opened in open office and the table data needs to be filled.

	A	B
1	**Subject**	**Number of students registered**
2	Science	28
3	Mathematics	32
4	Hindi	44
5	English	25
6	Physical Education	14
7	Sanskrit	34
8	Art and Craft	22

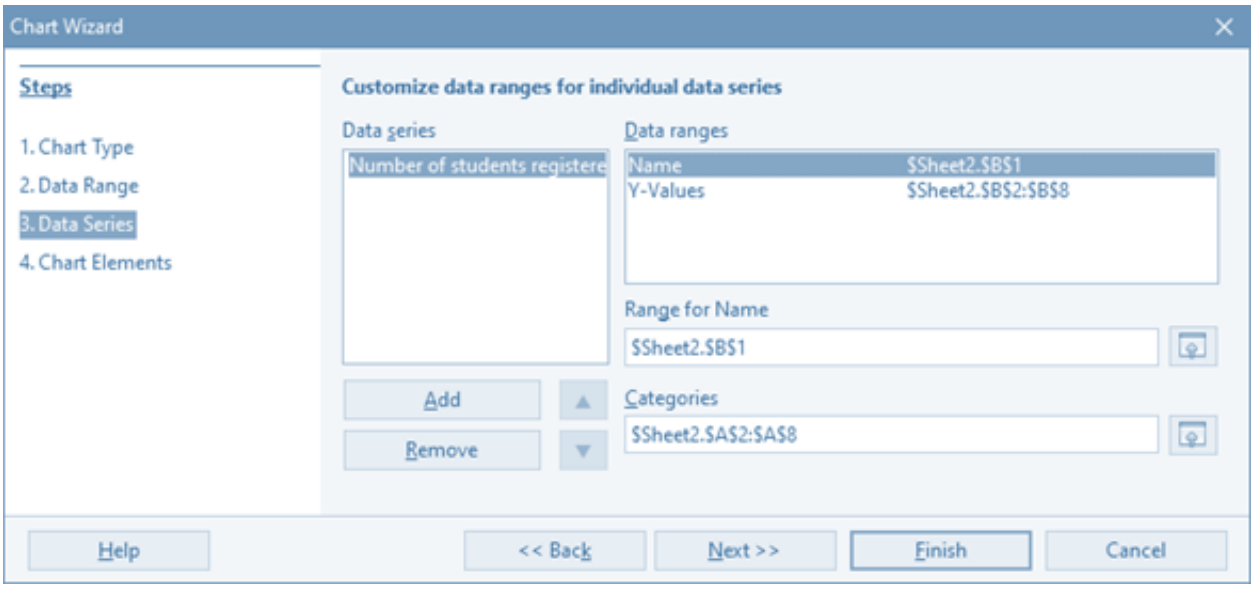
We then need to go to Menu, click on Insert, and from the dropdown select **chart**. On the chart wizard, we select the chart type as **column** and amongst the choice of column chart, we select the **normal** chart.



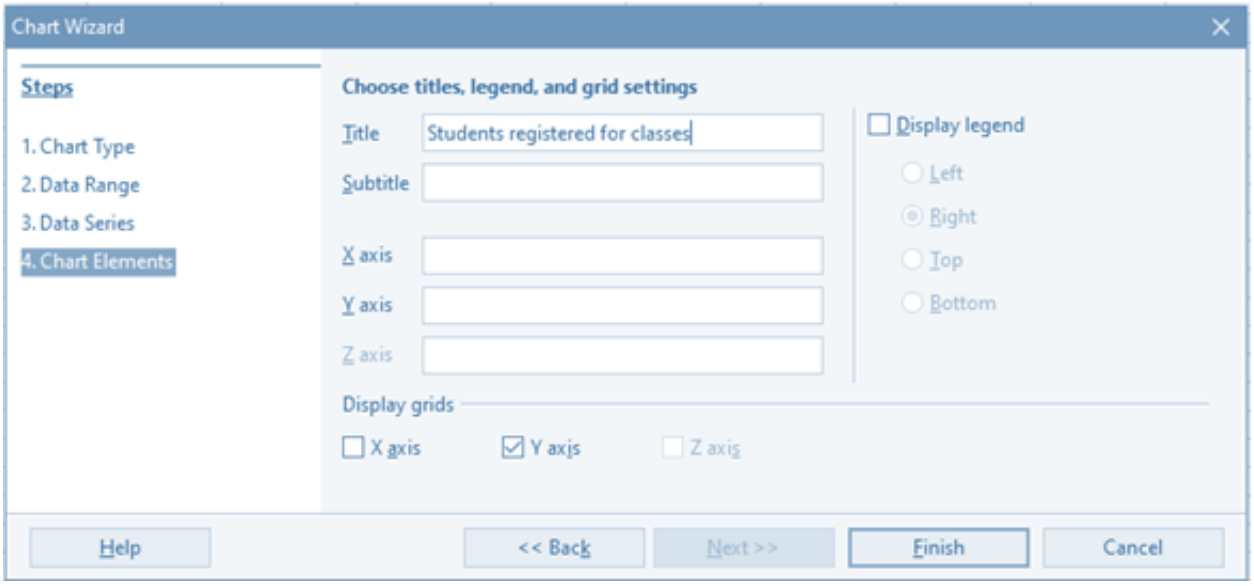
In the **data range**, select the entire table populated above, in this case, cell A1 to B8.



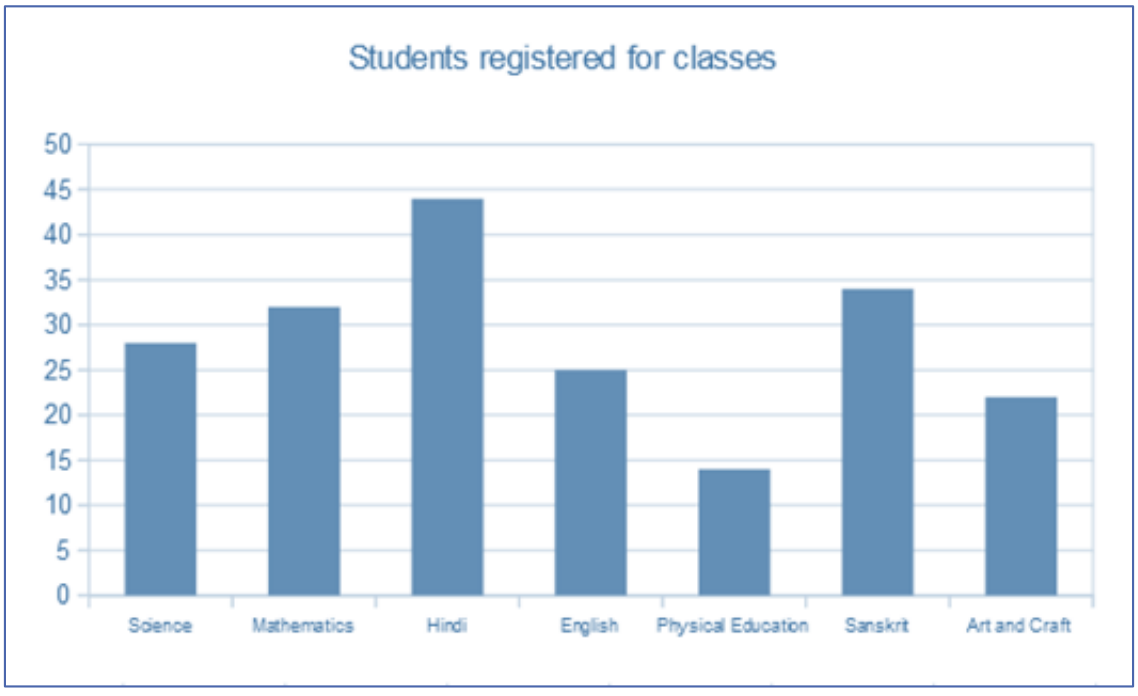
Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click onto next.



On the **Chart Elements** section, we may provide the title, subtitle, name of the x-axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart and if yes then where do we want to display the legend.



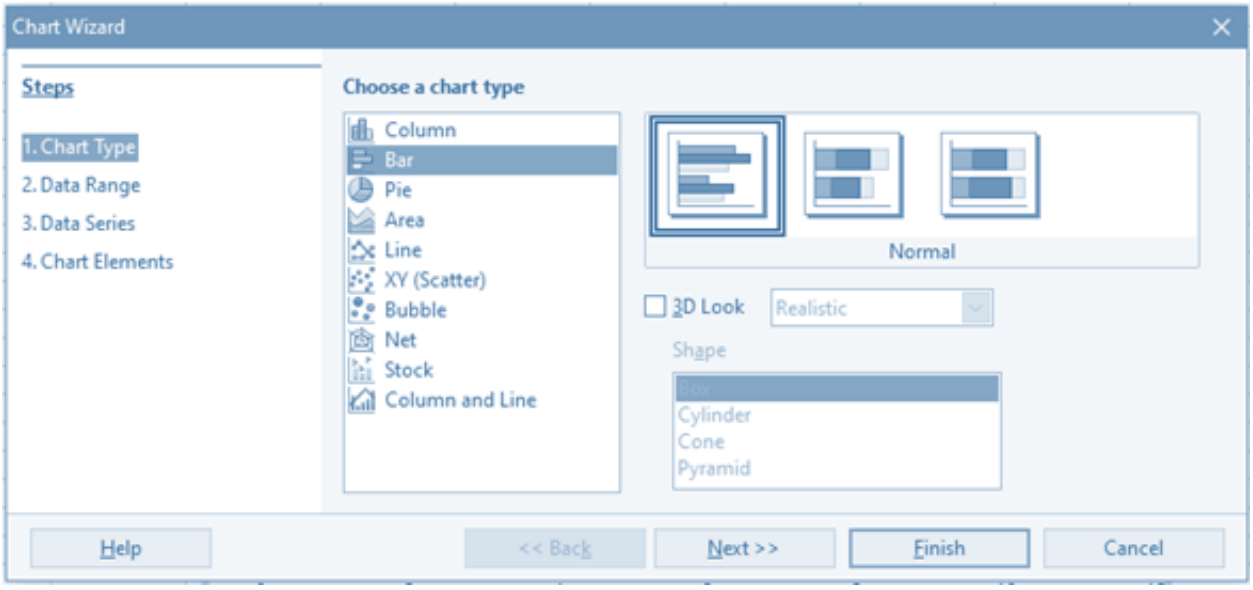
When the finish button is clicked, the chart gets generated as shown below:



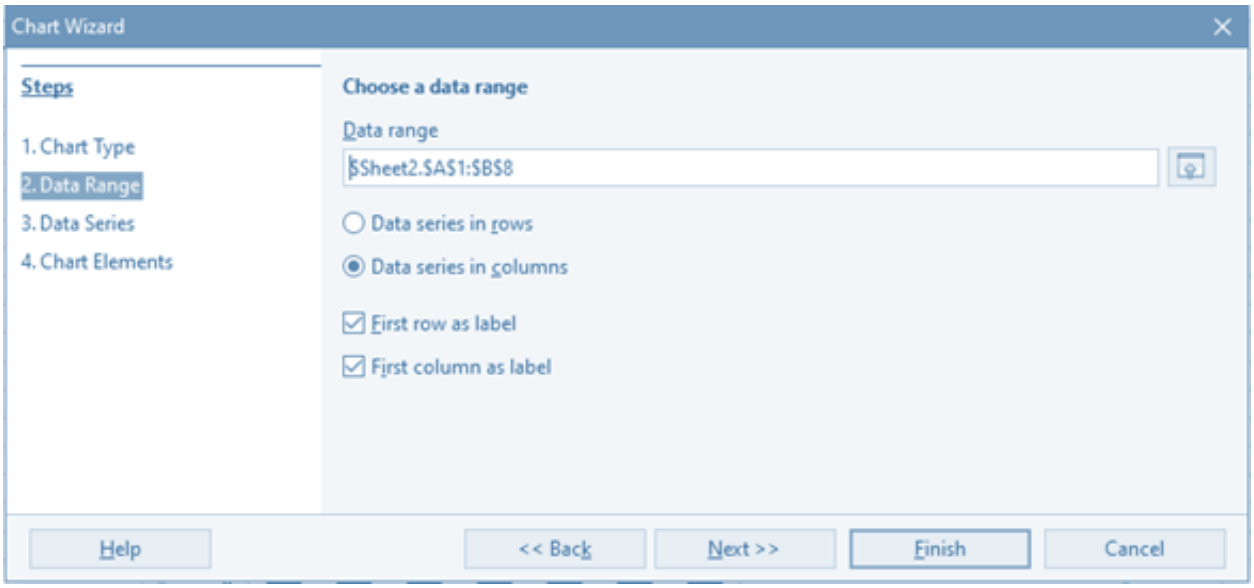
Column chart: In order to plot the **column chart** using open office, a spreadsheet needs to be opened in open office and the table data needs to be filled.

	A	B
1	**Subject**	**Number of students registered**
2	Science	28
3	Mathematics	32
4	Hindi	44
5	English	25
6	Physical Education	14
7	Sanskrit	34
8	Art and Craft	22

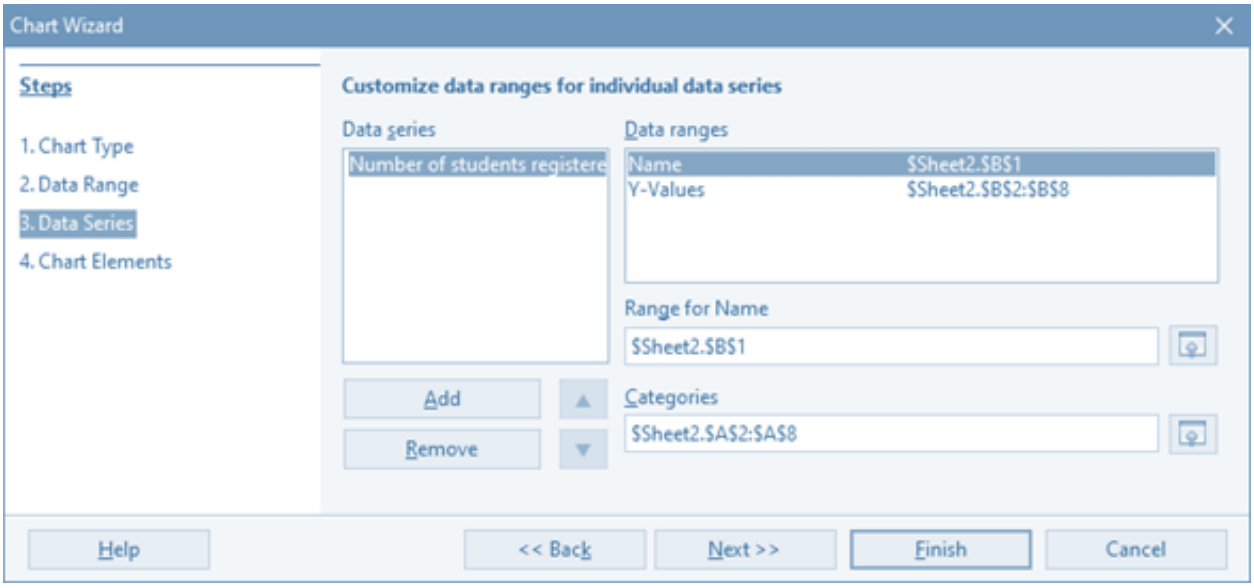
We then need to go to Menu, click on Insert and from the dropdown select **chart**. On the chart wizard, we select the chart type as **Bar** and amongst the choice of column chart, we select the **normal** chart.



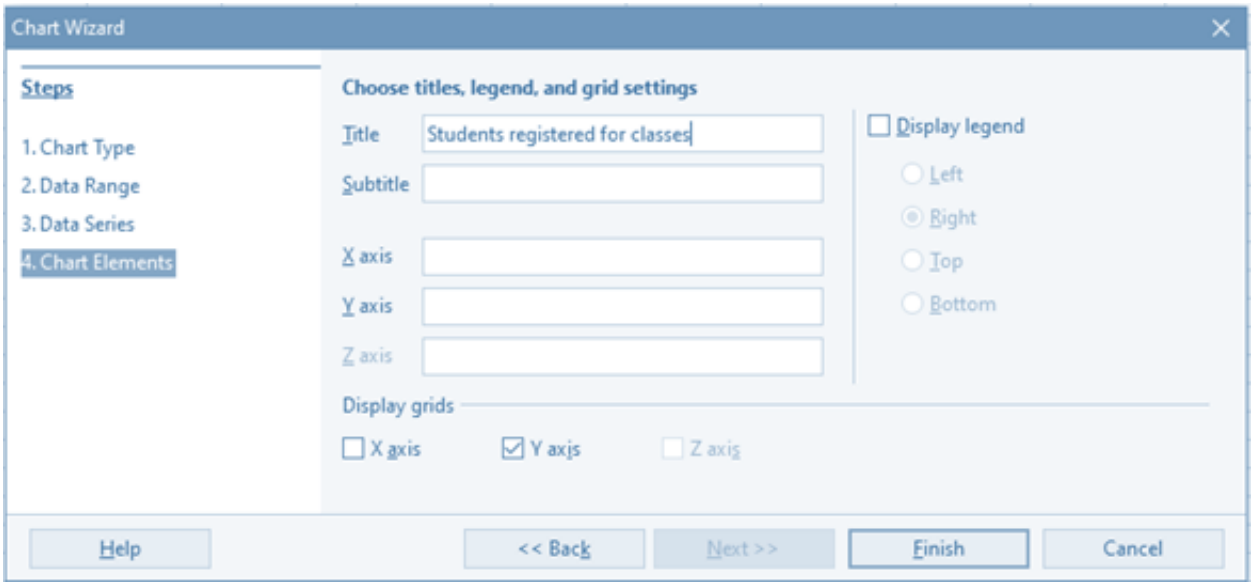
In the **data range**, select the entire table populated above, in this case cell A1 to B8.



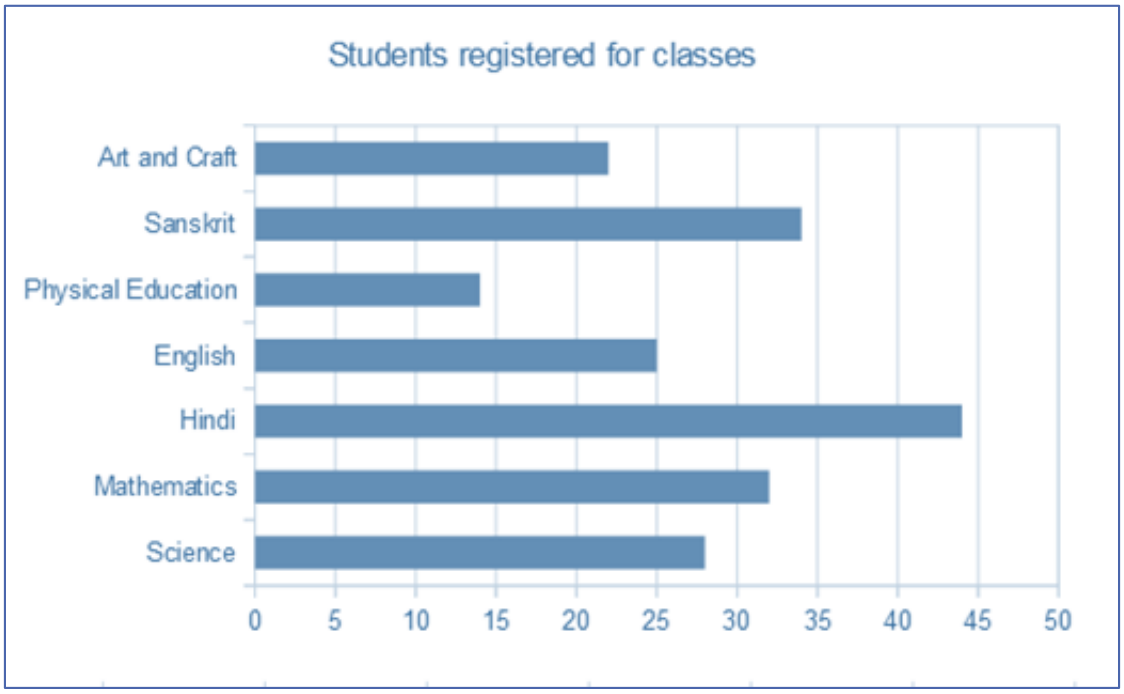
Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click onto next.



On the **Chart Elements** section, we may provide the title, subtitle, name of the x-axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart and if yes then where do we want to display the legend.



When the finish button is clicked, the chart gets generated as shown below:



Activity 3.2

Record your last year marks in a table and create a bar graph from it. Also create a column graph. Which one looks easier to understand?

3.3. Minimum and Maximum

As the name suggests, the minimum is the smallest value in the data set. The maximum is the largest value in the data set.

The Minimum:

The minimum of the data is less than or equal to any other values in our data set. If we had to order all our data in ascending order, so the first number in our list would be the minimum. Although the minimum value in our data set might be reused, this is a unique number. Since one of these values must be smaller than the other, there cannot be two minimum values.

The Maximum:

The maximum of the data is greater than or equal to all other values. If we had to order all our data in ascending order, so the last number listed would be the maximum. For a given set of results, a maximum is a unique number. This number can be repeated, but for a data set, there is just one maximum. Because one of these values will be higher than the other, there should not be two maximums.

For Example, consider marks of 10 students in English for class IX.

To find out the maximum and minimum,

87, 56, 43, 66, 73, 95, 85, 95, 82, 43

we need to sort the data set. Let us now arrange the data in ascending order.

43, 43, 56, 66, 73, 82, 85, 87, 95,95

After arranging the data, we can see that the minimum number of marks obtained in English is 43, and the maximum number of marks obtained is 95.

3.4. Frequency

The frequency of a data value is the number of times the data value occurs/repeats.

For Example, if five students have a score of 85 in English, the score of 85 is said to have a frequency of 5. The frequency of a data value is often represented by f.

For Example, in the below dataset, the marks scored for ten students in English. We will now try to find out the frequency of 95 and 56.

Student	Marks Obtained in English
Student 1	87
Student 2	56
Student 3	43
Student 4	66
Student 5	73
Student 6	95
Student 7	85
Student 8	95
Student 9	82
Student 10	43

Fig 3.11 *Student marks in English*

Arrange the number in ascending order; you will see that number 95 repeats twice and number 56 once. So the frequency of 95 is two and frequency of 56 is one, or you can write it this way
f (95) =2 and f (56) =1

4. Histograms

Now that we understand frequency, let us see what a histogram is?

A histogram is a graphical illustration of frequency plotted against intervals. To understand this better, let us consider the below Example.

The below data set contains the height of some students for a class.

101,105,102,107,104,108,101,102,109,104,103,109,106,101

To draw a histogram from this, we first need to organize the data into intervals. These intervals are also called logical ranges or bins.

Now we will compute the number of times the student's height falls within

(101- 102), (102 – 103), (103 – 104), (104 – 105), (105 – 106), (106 – 107), (107 – 108), (108 – 109), (109 – 110)

each of these ranges. For Example, there are three students within the height range of (101 – 102). If we do the same process for the entire data set, we get the following records.

Range	Frequency
101 - 102	3
102 - 103	2
103 - 104	1
104 - 105	2
105 - 106	1
106 - 107	1
107 - 108	1
108 - 109	1
109 - 110	2

Fig 3.12 *Frequency Distribution*

Let us now plot this as a histogram.

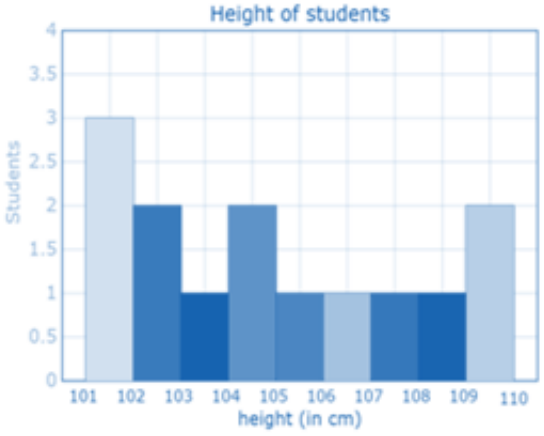


Fig 3.13 *Histogram*

This graph that we just created is called a histogram.

A histogram is a graphical display of data using bars of different heights. To summarize discrete or continuous data, a histogram is used. In other words, A histogram displays the number of data points that fall within a given set of values (called 'bins') to provide a visual representation of numeric data. Unlike a vertical bar graph, a histogram shows no gaps between the bars. A histogram also gives the value of the mode of the frequency distribution graphically.

Bin Widths

Simply stated, Bin widths are the range size. In our previous Example, our range size was 1.

We then create a table of three columns – Bins, Range & Frequency as shown alongside. Bins will hold the individual starting value of the range with a bin width of 1. Range will hold the range of values covered. Frequency is the number of values falling in that bin.

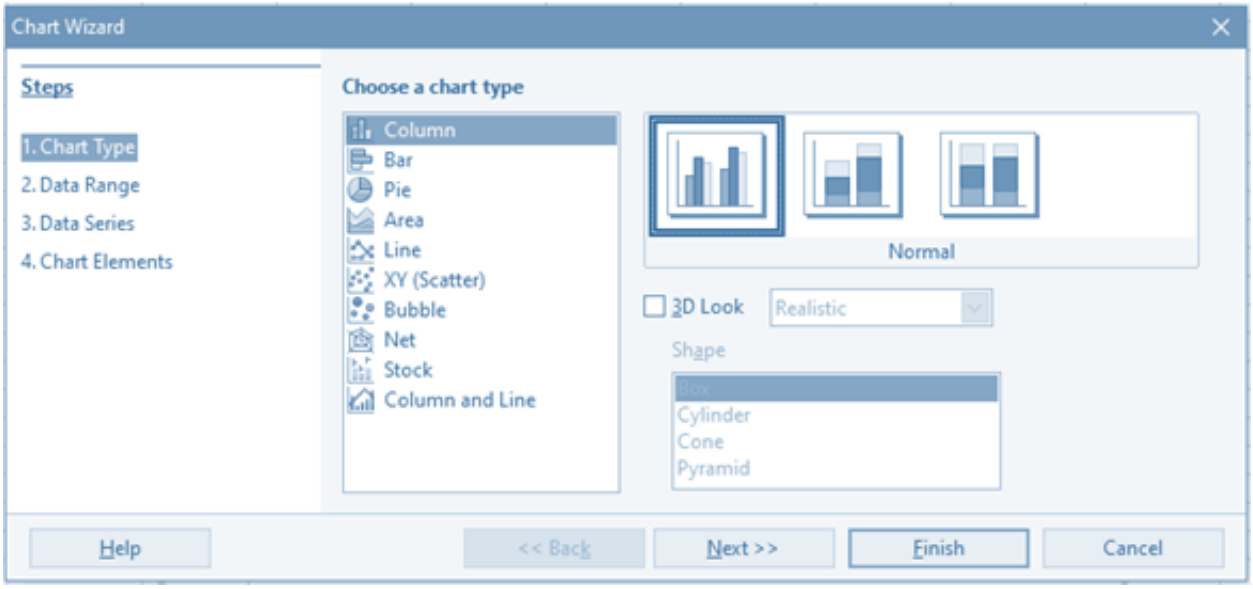
Plotting histogram using open office

To plot a histogram using open office, a spreadsheet needs to be opened in open office and the data to be analyzed needs to be filled in.

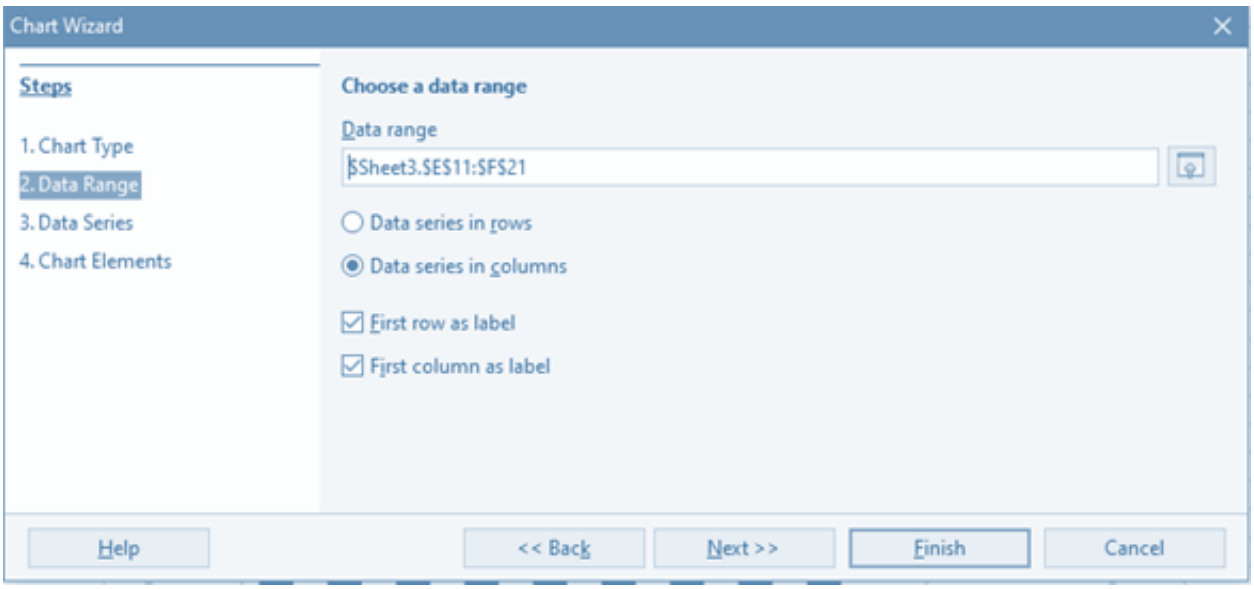
	A	B
1	Height of students (in cm)	
2	101	102
3	105	109
4	102	104
5	107	103
6	104	109
7	108	106
8	101	101

Bins	Range	Frequency
101	101-102	3
102	102-103	2
103	103-104	1
104	104-105	2
105	105-106	1
106	106-107	1
107	107-108	1
108	108-109	1
109	109-110	2
110	110-111	0

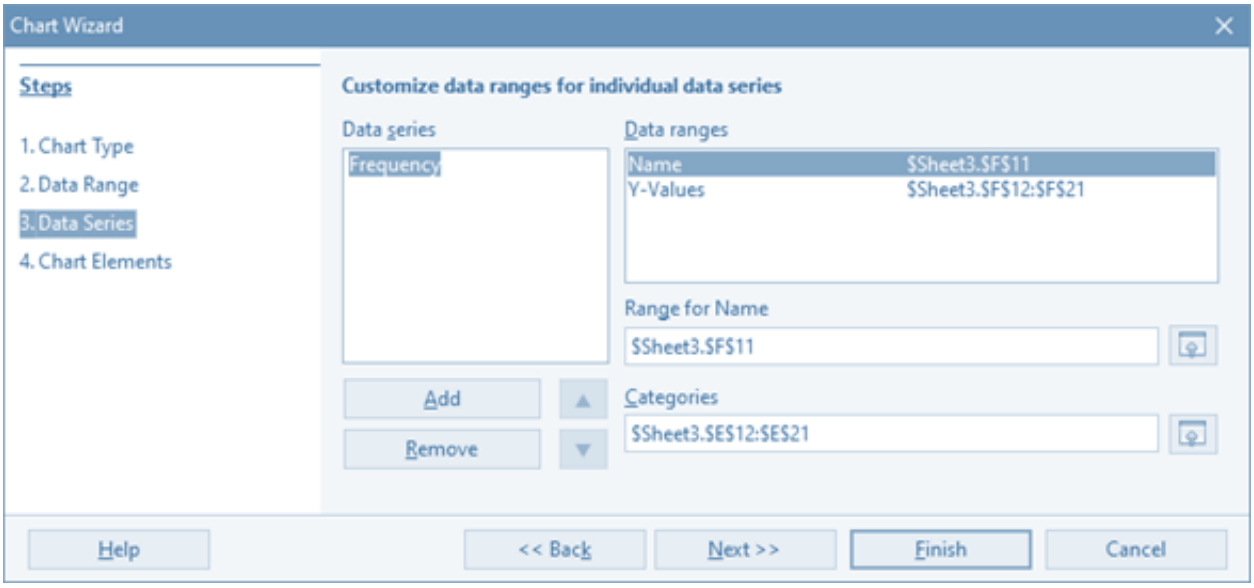
Next, we need to go to Menu, click on Insert and from the dropdown select **chart**. On the chart wizard, we select the chart type as **column** and amongst the choice of column chart, we select the **normal** chart.



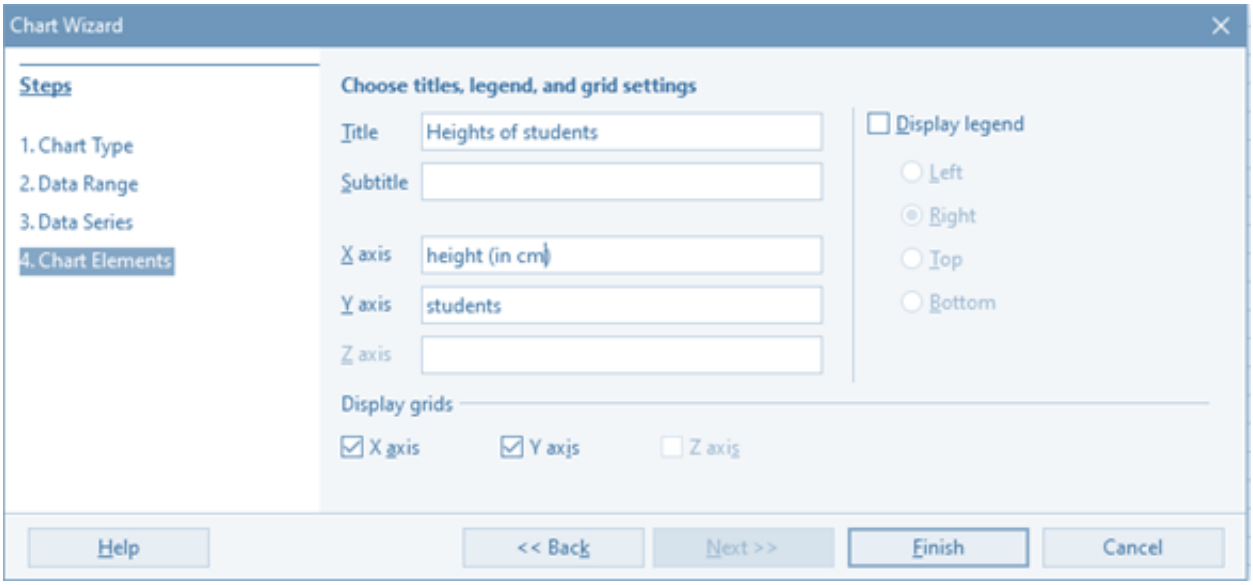
In the **data range**, select all the rows of column **Range** & **Frequency**.



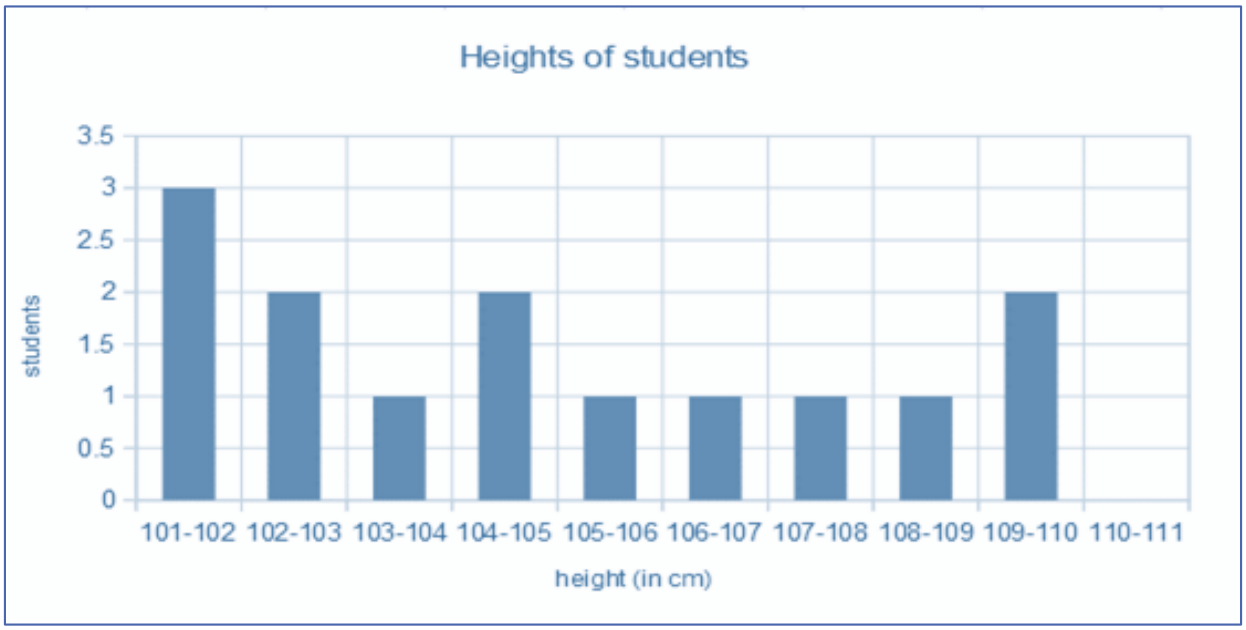
Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click onto next.



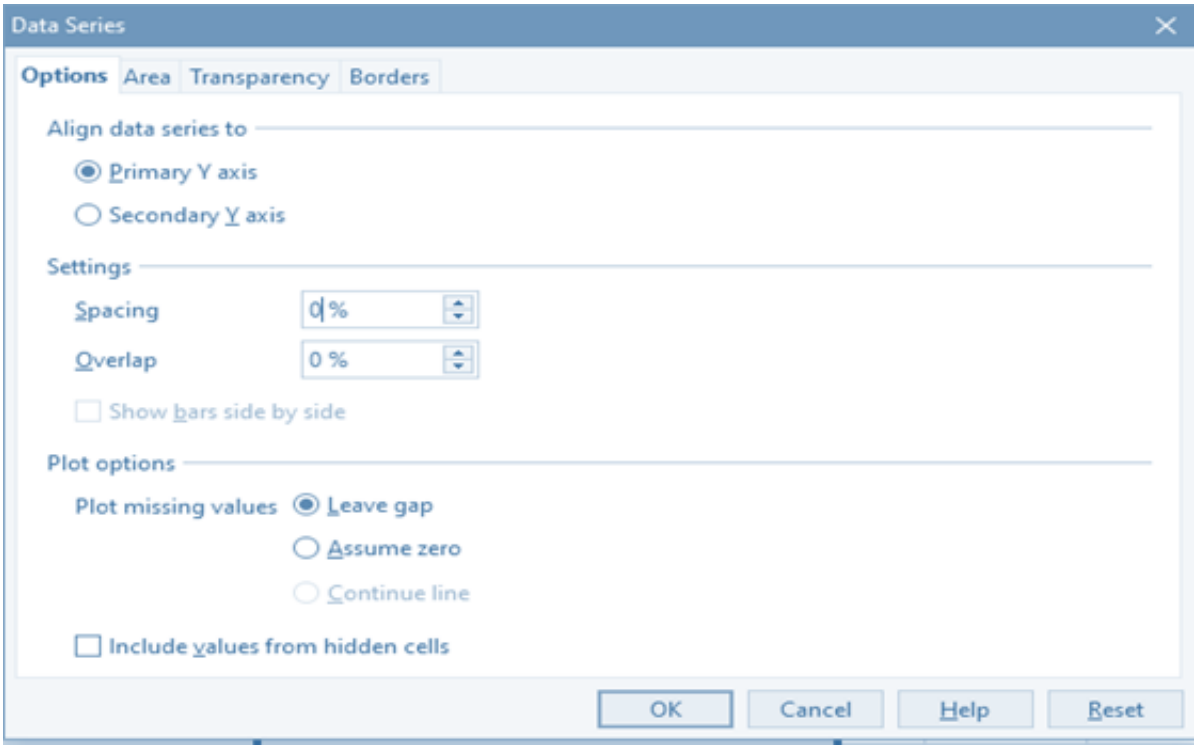
On the **Chart Elements** section, we may provide the title, subtitle, name of the x-axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart and if yes then where do we want to display the legend.

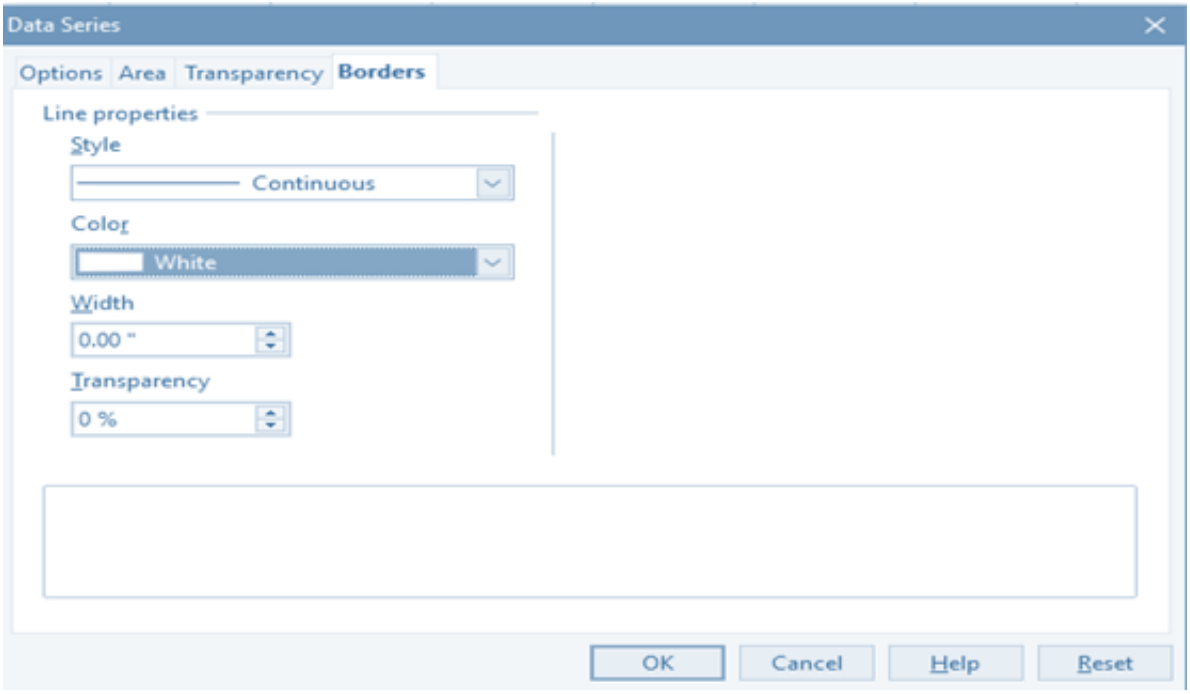


When the finish button is clicked, the chart gets generated as shown below:

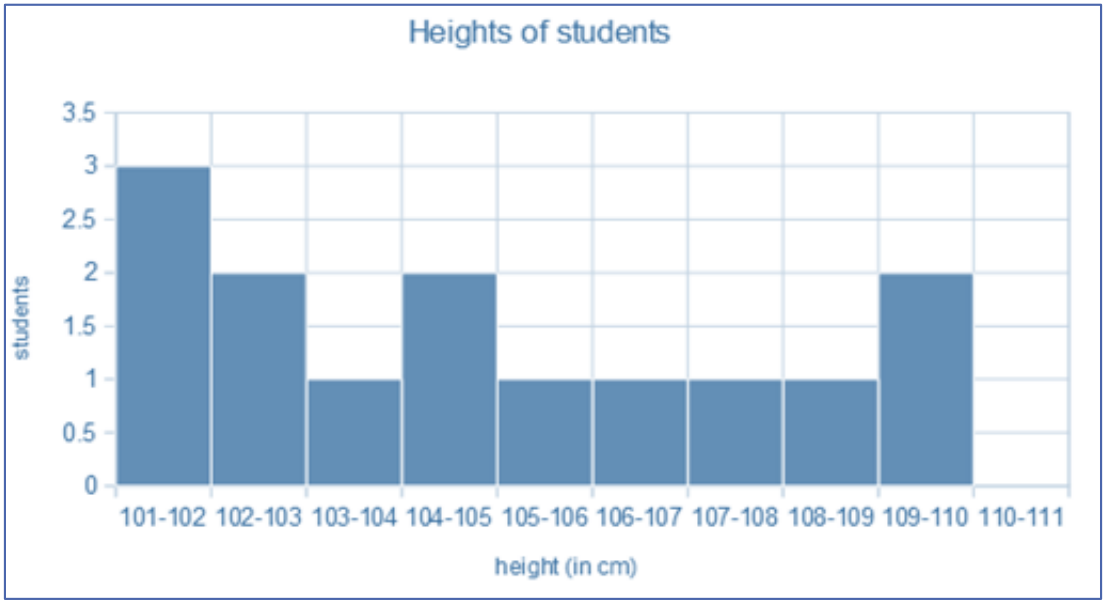


We can then click to select the bars of the chart, right click and select **Format Data Series** to adjust the spacing between the bars and introduce a border on the bars.





On clicking **OK**, this is how the chart appears:



5. Use of shapes

In this module, we will understand the different shapes of a histogram and what they mean.

5.1. Normal Distribution

Data points in a normal distribution are as likely to occur on one side of the average as on the other side of the average.

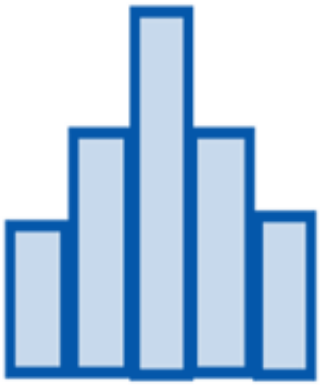


Fig 3.14 *Normal Distribution*

5.2. Right Skewed Distribution

Many data points occur on the left side, with fewer data points on the right side in a right-skewed distribution. A right-skewed distribution occurs when the data has a range boundary on the left-hand side of the histogram. A right-skewed distribution is also known as a positively skewed distribution

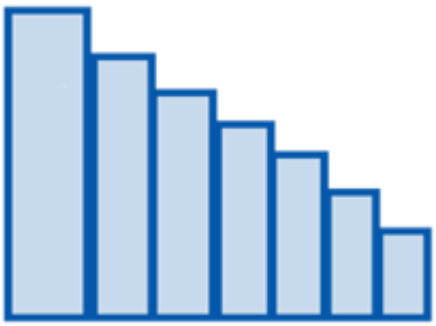


Fig 3.15 Right Skewed Distribution

5.3. Left Skewed Distribution

Many data points in a left-skewed distribution occur on the right side with a scarcer number of data points on the left side.

A left-skewed distribution usually occurs when the data has a range boundary on the histogram's right-hand side. A left-skewed distribution is also known as a negatively skewed distribution.

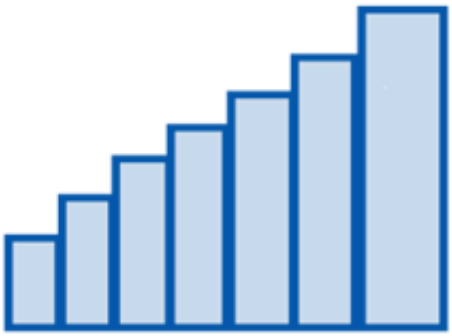


Fig 3.16 Left Skewed Distribution

5.4. Bimodal Distribution

A bimodal distribution has two peaks. In a bimodal distribution, the data should be separated and analyzed as separate normal distributions.

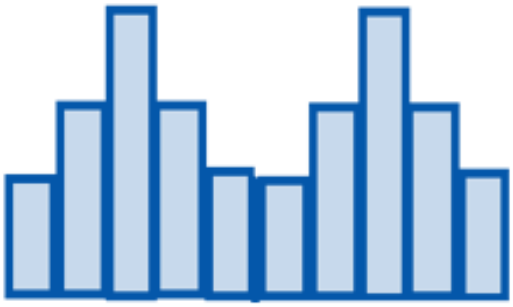


Fig 3.17 Bimodal Distribution

5.5. Random Distribution

A random distribution lacks an apparent pattern and has several peaks.

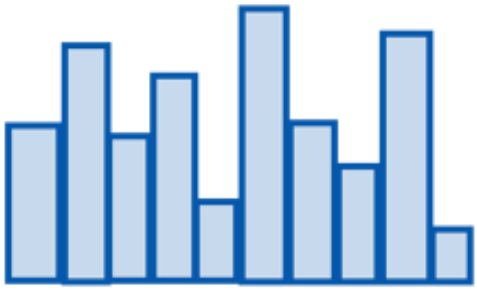


Fig 3.18 Random Distribution

> ***Activity 3.3***
>
> Record the attendance for a month of your class in a table and create a histogram from it. Also try to find out the shape of the histogram.

6. Use of Single and Multi-Variable plots

6.1. Single Variable plots

To visualize one variable, the type of graphs to use depends on the type of the variable:

For categorical variables (or grouping variables). You can visualize the count of categories using a bar plot or a pie chart to show each type's proportion.

You can visualize the variable's distribution for a continuous variable using density plots, histograms, etc.

Representation using pie chart: In this example, we have a survey of 35 students on what food they prefer This is the response:

Food Item	Number of students
Pizza	8
Burger	10
Ice-Cream	15
Dosa	7

Fig 3.19 Food items preferred by students

Visualizing the data using a pie chart:

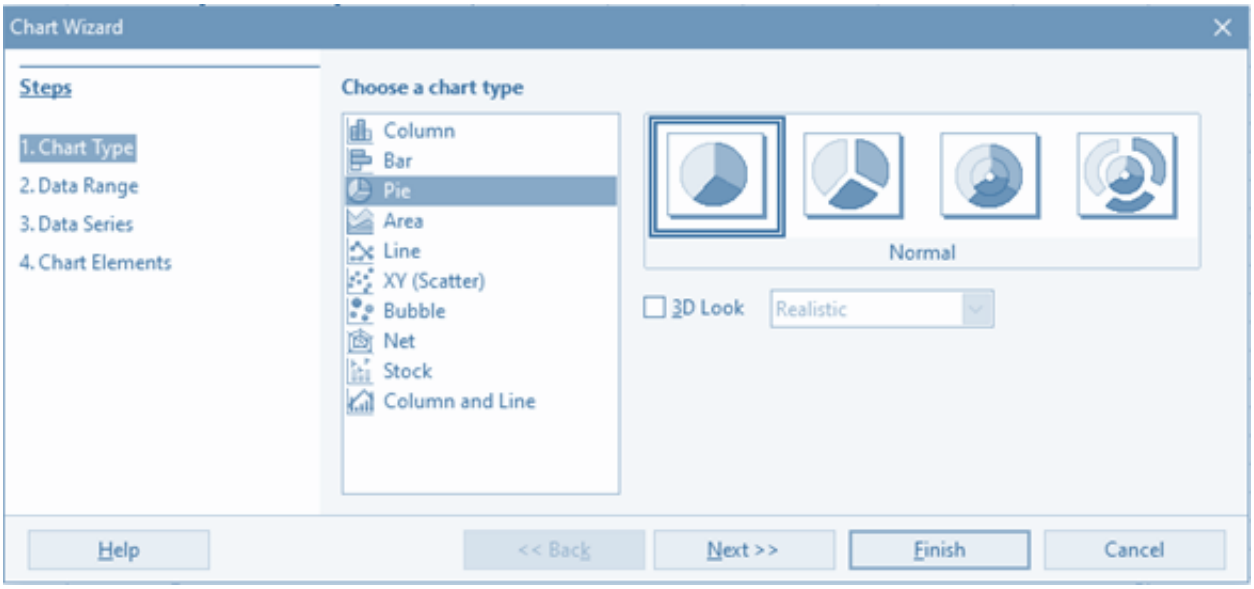
Fig 3.20 Pie Chart

You can see that 37% of students prefer Ice-Cream.

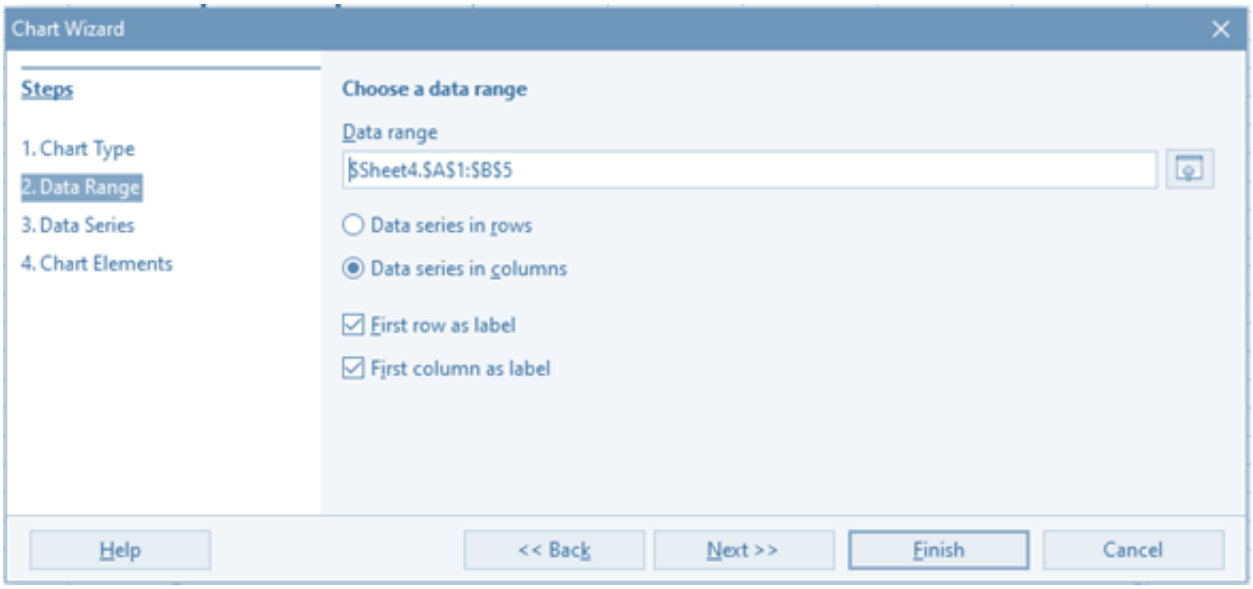
In order to plot the pie chart using open office, a spreadsheet needs to be opened in open office and the table data needs to be filled.

	A	B
1	Food Item	Number of students
2	Pizza	8
3	Burger	10
4	Ice-cream	15
5	Dosa	7

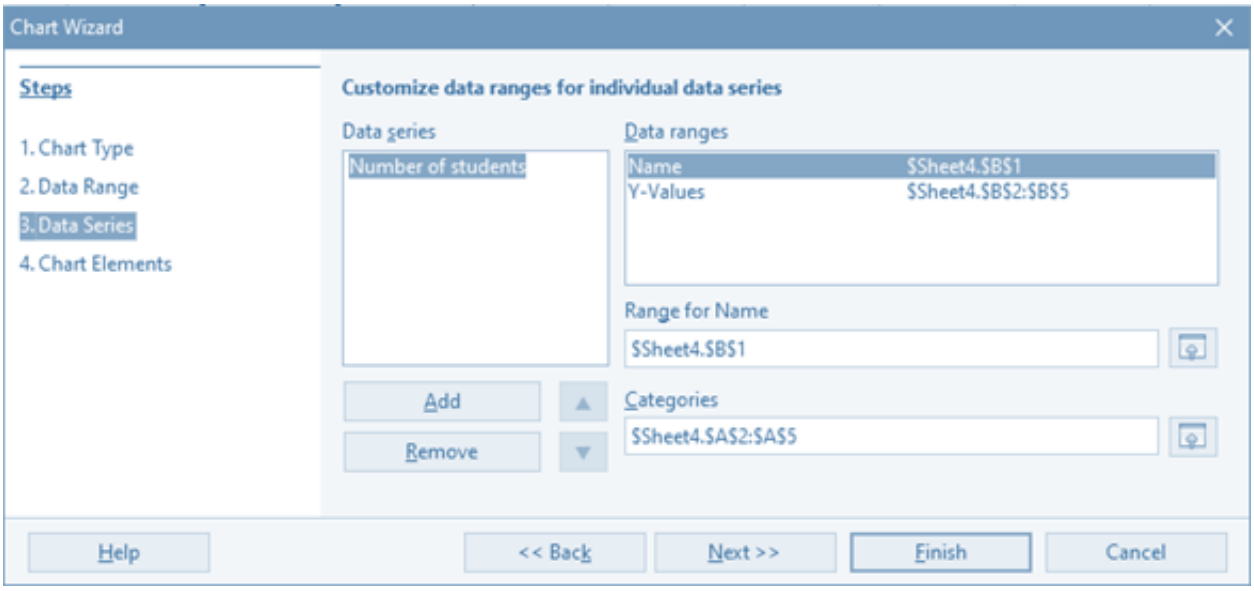
Next, we need to go to Menu, click on Insert and from the dropdown select **chart**. On the chart wizard, we select the chart type as **pie** and amongst the choice of pie chart, we select the **normal** chart.



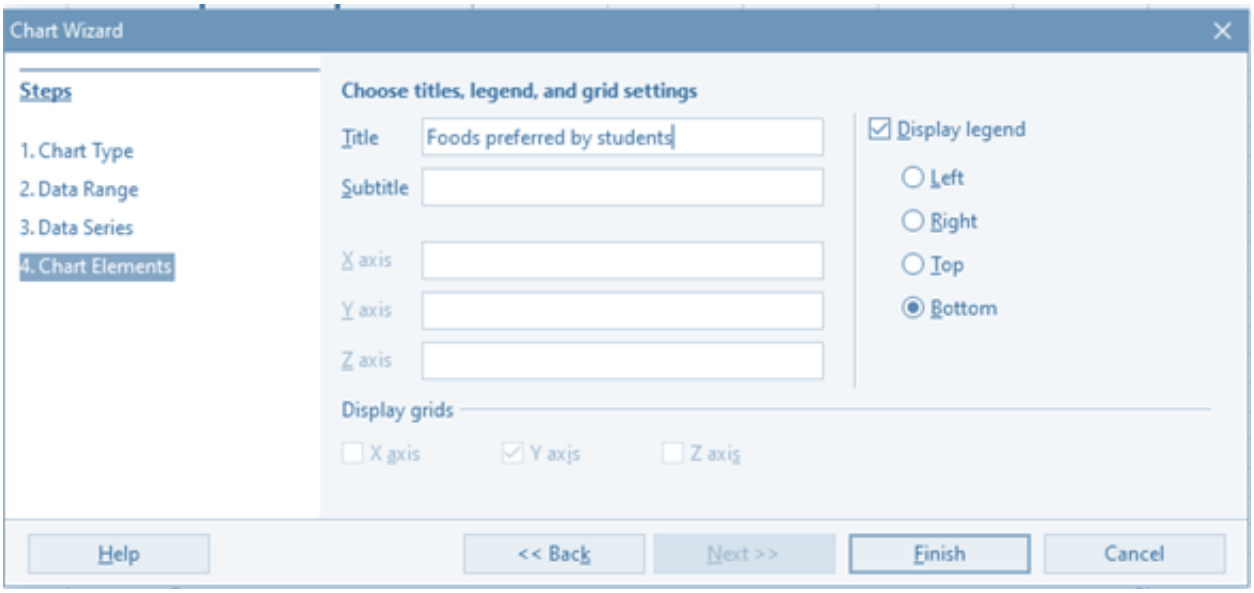
In the **data range**, select the entire table populated above, in this case cell A1 to B5.



Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click onto next.



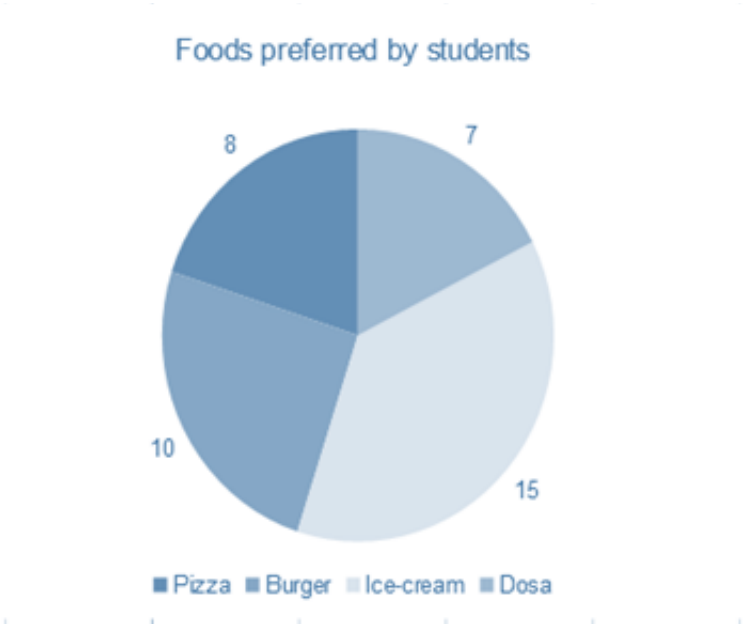
On the **Chart Elements** section, we may provide the title, subtitle, name of the x axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart and if yes then where do we want to display the legend.



When the finish button is clicked, the chart gets generated as shown below:

If we want to display the percentage contribution of each food item in the pie chart, we need to first click inside the pie chart. This will select all the pies inside the pie chart.

Next, we need to right click and select **Insert Data Labels**. This will show the actual values associated with each food item in the pie chart.

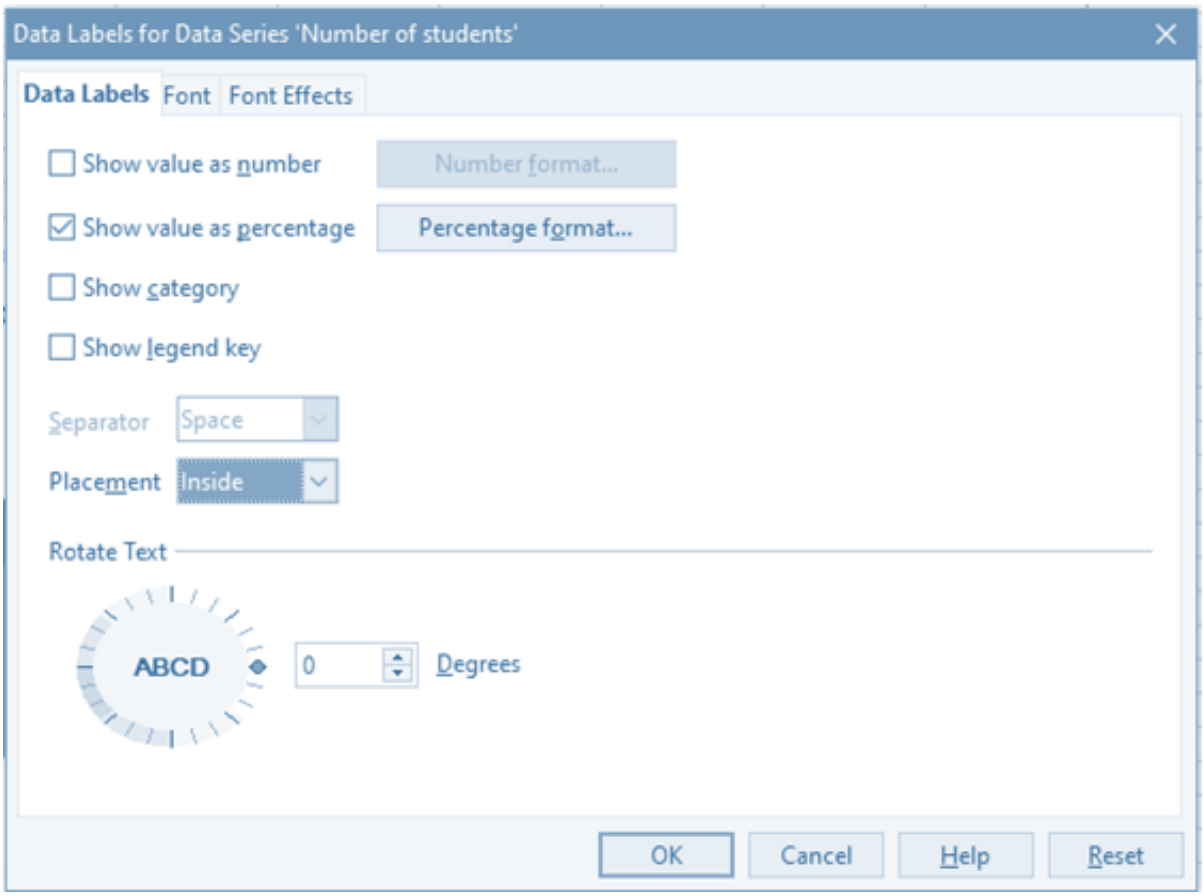


Next, we need to right click again inside the pie of the pie chart and select **Format Data Labels.**

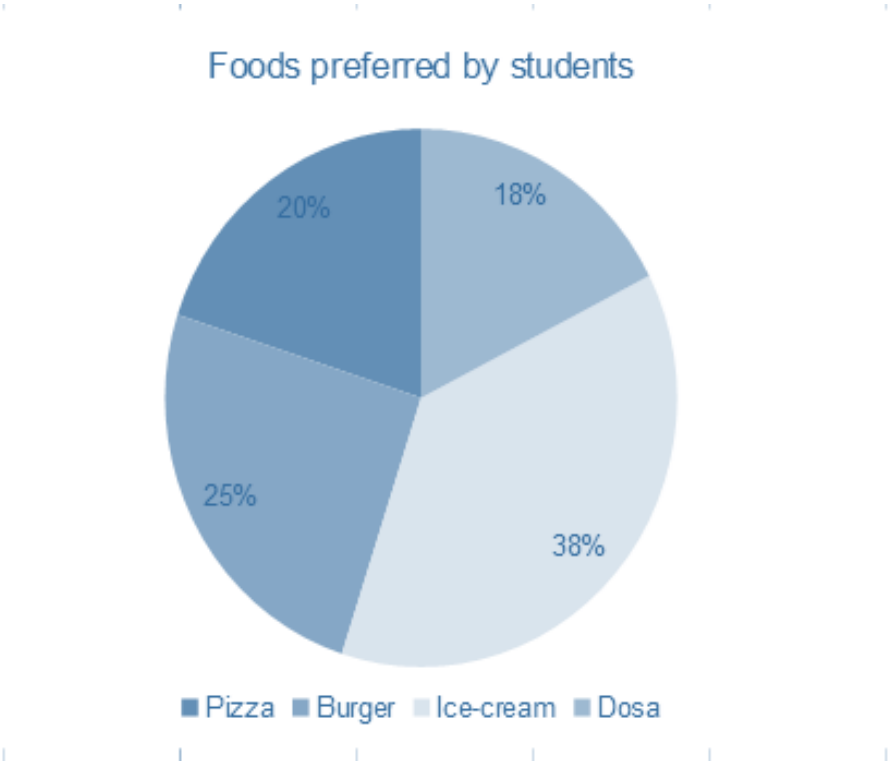
In the check box, we need to keep only **Show value as percentage** checkbox checked.

In the Placement dropdown, we need to select **Inside** as the option.

Then we click on **OK**.



After this is done, the pie chart is displayed having the percentage contribution of each food item in the pie chart as shown below.



6.2. Multi-Variable plots

Multi-variable plots are used to display relationship among several variables

Example:
We have a survey of how many students enrolled in schools 1 and 2 from 1995-2006.
This is the data:

Year	Number of students in School 1	Number of students in School 2
1995	104	109
1996	107	102
1997	101	111
1998	120	118
1999	125	123
2000	106	120
2001	109	132
2002	130	128
2003	136	133
2004	127	135
2005	132	130
2006	141	120

Fig 3.21 *Number of students in School*

Visualizing the data using a two-variable line chart:

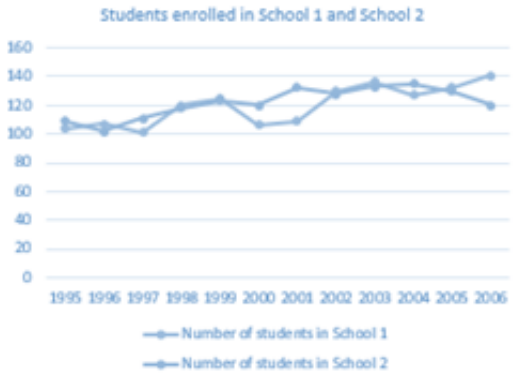
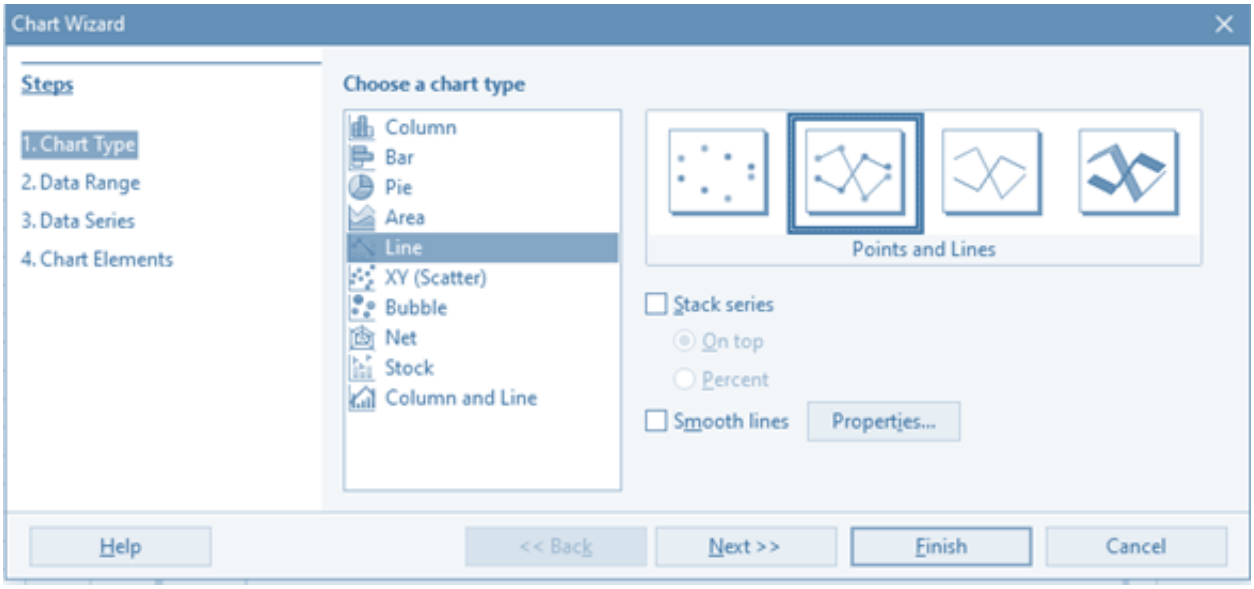


Fig 3.22 *Two Variable Line Chart*

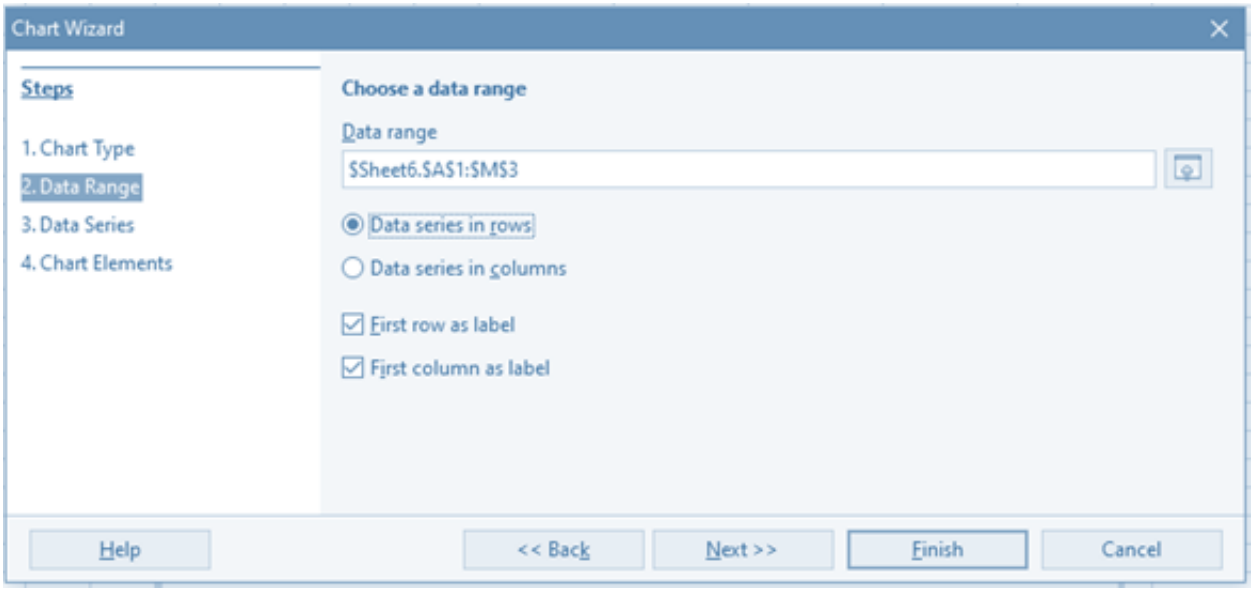
In order to plot the two variable line chart using open office, a spreadsheet needs to be opened in open office and the table data needs to be filled.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
2	Number of students in school1	104	107	101	120	125	106	109	130	136	127	132	141
3	Number of students in school2	109	102	111	118	123	120	132	128	133	135	130	120

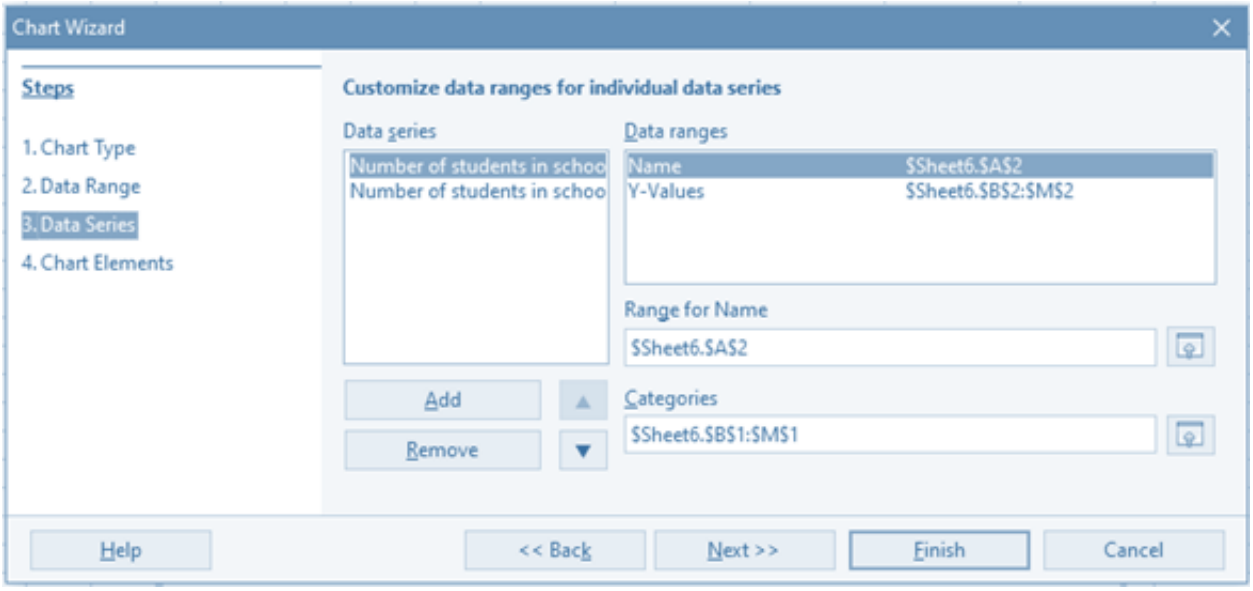
Next, we need to go to Menu, click on Insert and from the dropdown select **chart**. On the chart wizard, we select the chart type as **Line** and amongst the choice of line chart, we select the **Points and Lines** chart.



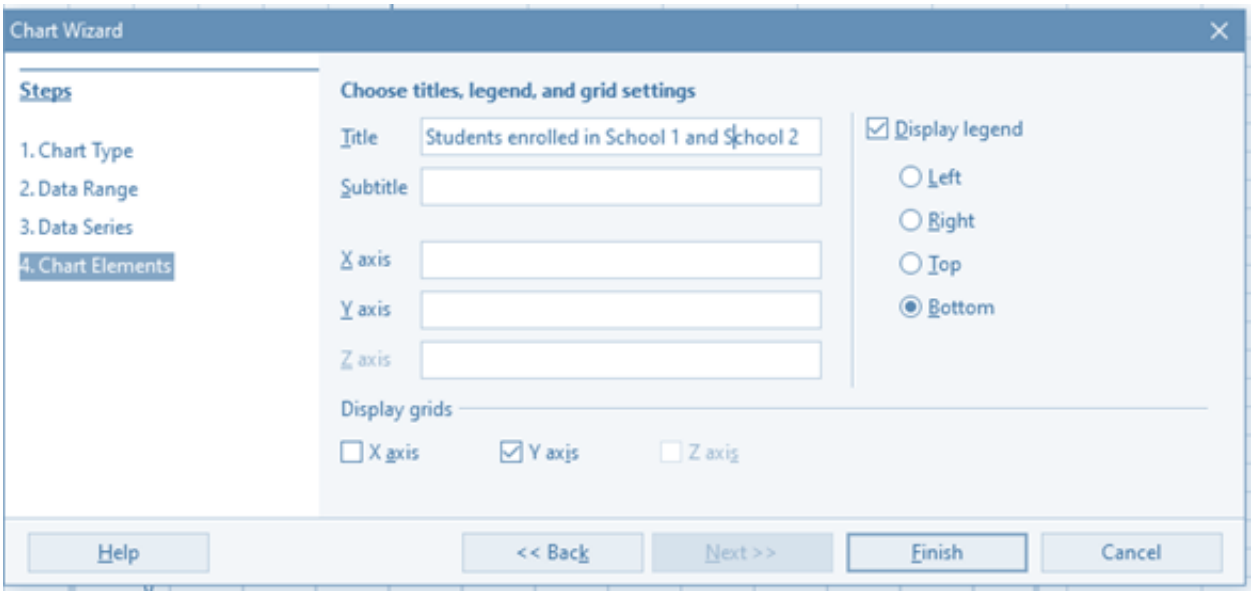
In the **data range**, select the entire table populated above, in this case cell A1 to M3. Here we select **Data series in rows option**.



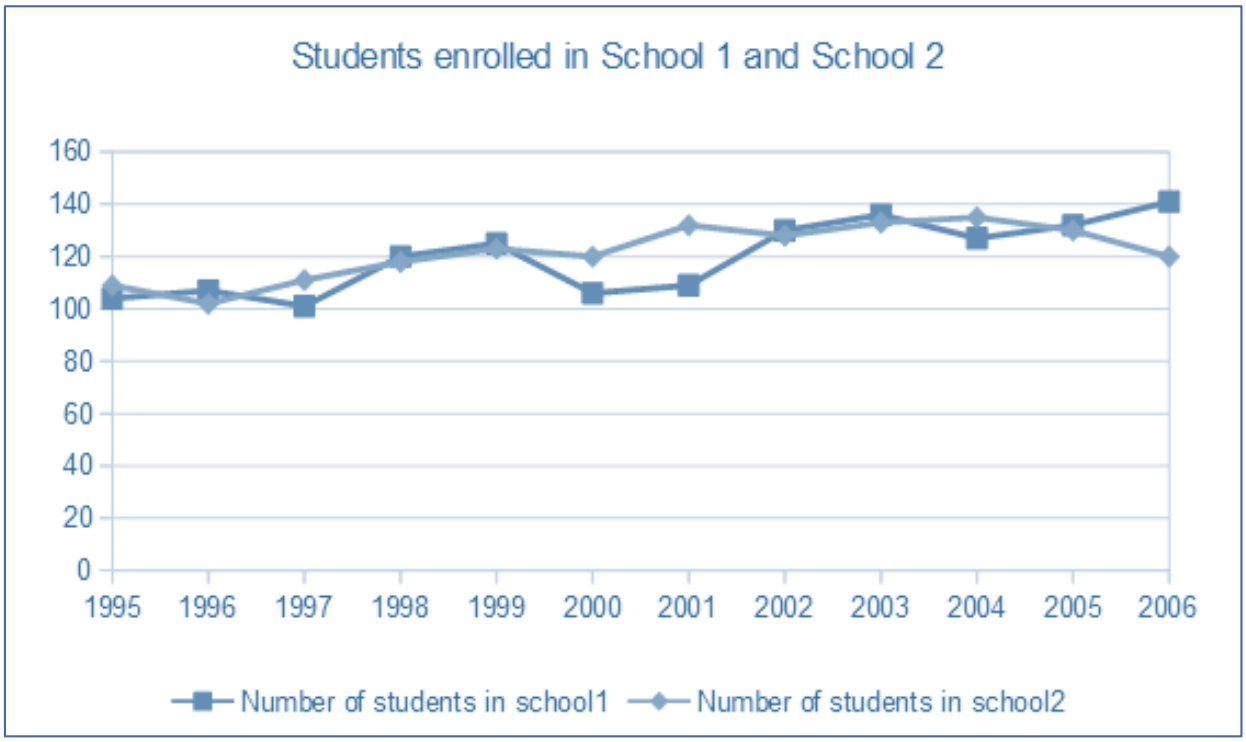
Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click onto next.



On the **Chart Elements** section, we may provide the title, subtitle, name of the x axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart and if yes then where do we want to display the legend.



On clicking **Finish**, this is how the chart appears:



Example:
We have data on the number of shoes sold for three companies for the year 2020

Month	Company A	Company B	Company C
Jan	32	28	35
Feb	34	22	32
Mar	35	34	26
Apr	28	36	29
May	37	25	32
Jun	45	20	37
Jul	38	31	39
Aug	39	22	42
Sep	41	45	38
Oct	29	33	23
Nov	35	23	27
Dec	42	37	35

Fig 3.23 *Shoes Sold*

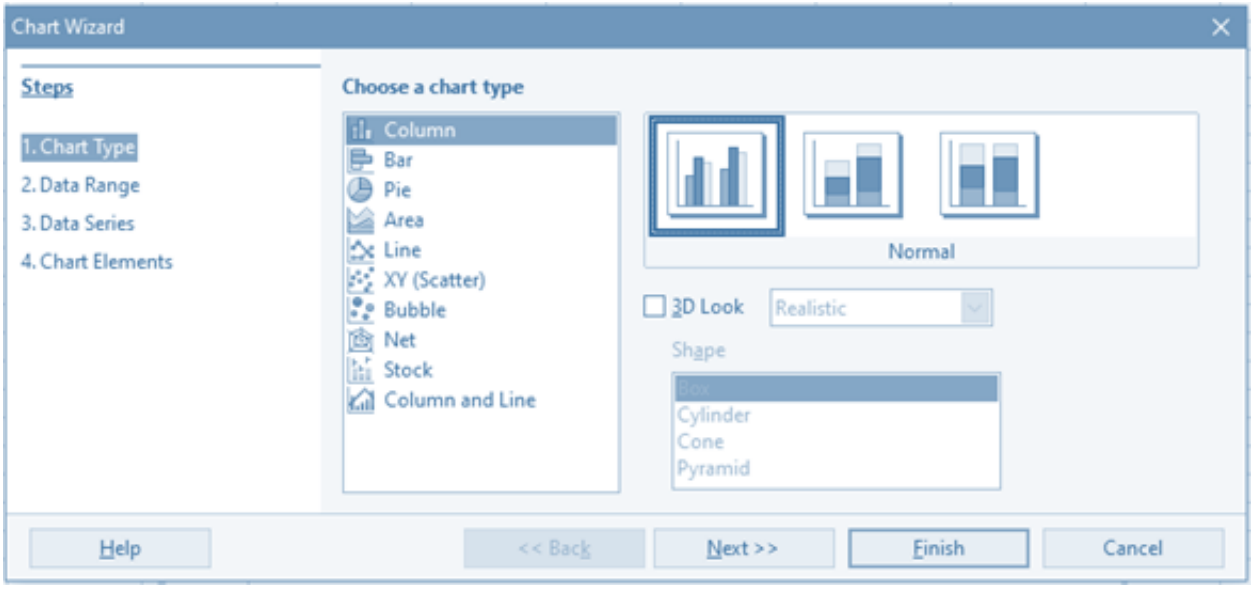
Visualizing the data using a multi-variable bar chart:

Fig 3.24 *Multi-Variable Bar Graph*

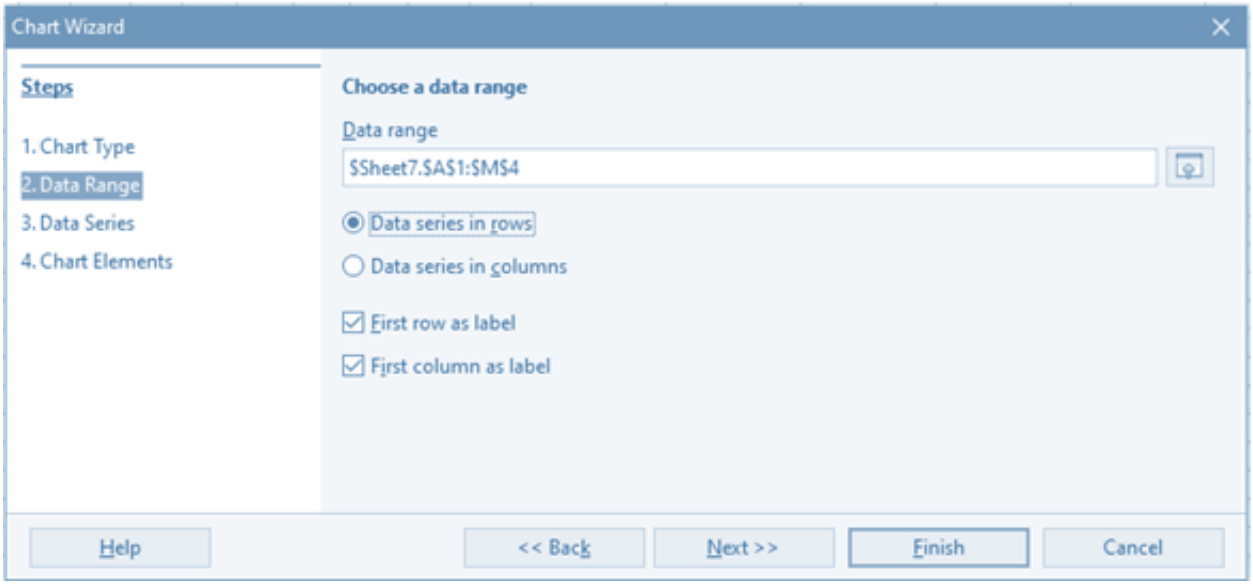
In order to plot the two variable line chart using open office, a spreadsheet needs to be opened in open office and the table data needs to be filled.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2	Company A	32	34	35	28	37	45	38	39	41	29	35	42
3	Company B	28	22	34	36	25	20	31	22	45	33	23	37
4	Company C	35	32	26	29	32	37	39	42	38	23	27	35

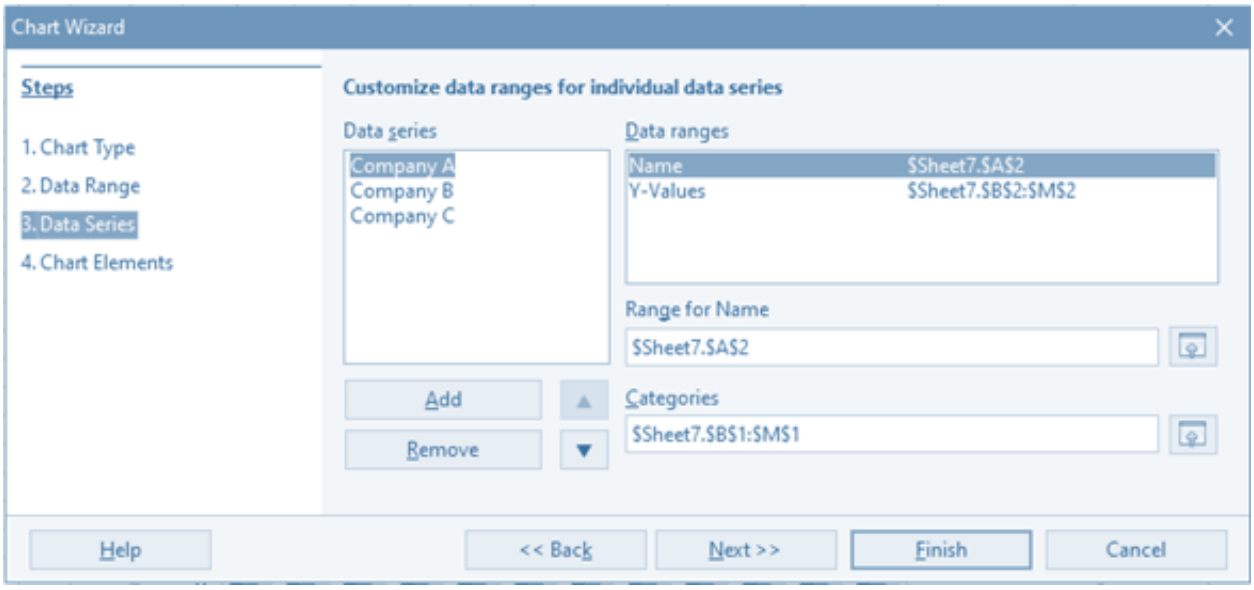
Next, we need to go to Menu, click on Insert and from the dropdown select **chart**. On the chart wizard, we select the chart type as **column** and amongst the choice of column chart, we select the **normal** chart.



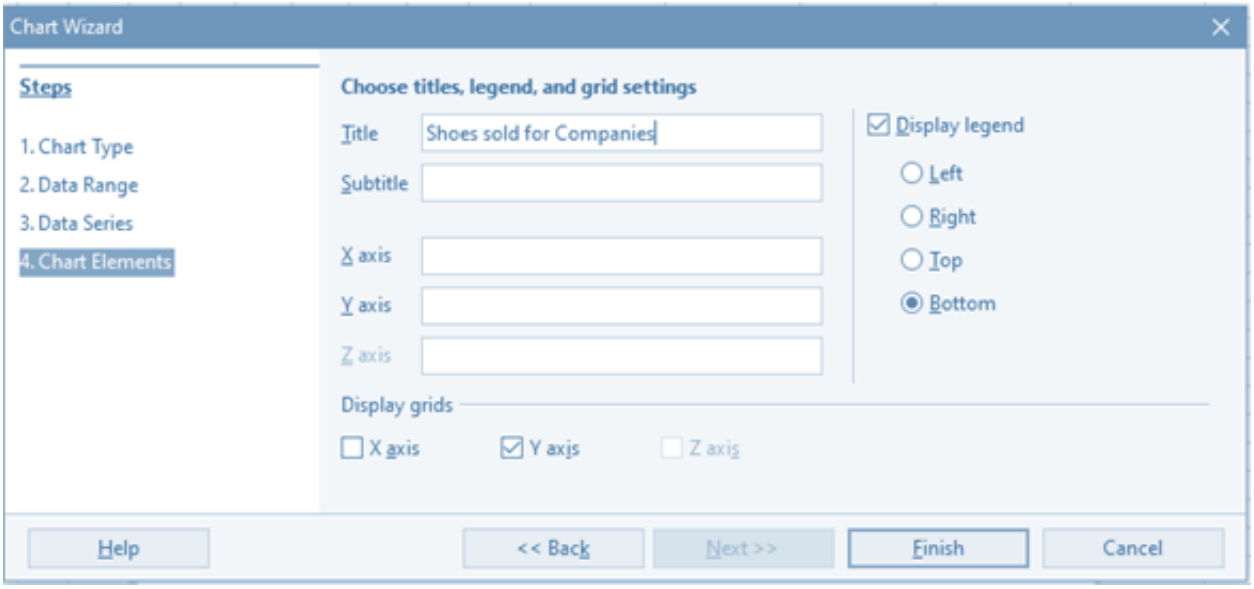
In the **data range**, select the entire table populated above, in this case cell A1 to M4. Here we select **Data series in rows option**.



Clicking onto **Next,** takes us to Data Series. We don't alter anything here and click onto next.



On the **Chart Elements** section, we may provide the title, subtitle, name of the x axis & name of the y axis of the chart. We can also select, if we want to display legend for the chart and if yes then where do we want to display the legend.



On clicking **Finish**, this is how the chart appears:

Activity: Choosing the band for the year-end party

In this activity, you are going to choose the band for the year-end party for your grade.
You will conduct a survey, summarize the data, analyze the results, and then from the data, we will interpret the results.

Formulate survey question

You may be interested in the favorite type of music among your classmates. Imagine a year-end party is being planned for your grade, and there is only enough money to hire one musical group.

The survey question you may have is:

What type of music do the students in my class like?

This survey question attempts to measure the type of music preference in your class's students' population.

Collect Data

To answer the survey question, you need to collect data about the music your classmates like.

Before beginning data collection, it is crucial to think through the data collection methods.

A survey is a natural data collection method.

One possible survey question you could ask:

What is your favorite type of music?

However, the survey question in this form could elicit many different responses, making it challenging to analyze the data.

You might amend the survey question to be more restricted:

What style of music is your favorite: country, rap, or rock?

Because this question explicitly asks your classmates to choose among three options, it will be easier to manage and analyze the data.

The downside to this question is that it restricts respondents' choices, so for someone who prefers jazz, their response will not indicate their favorite music. Type of music is a categorical variable defined here by country, rap, or rock.

The data that results from each of your classmates identifying their type of music preference is called categorical data. Once you decide on a survey question, you can conduct a census where every student in your class answers the survey question.

It would be helpful if you recognized that there would be individual-to-individual variability. The analysis of the results from this one class will be used to infer what the music style might be for the whole grade.

Suppose the class survey of 24 students in one of the classrooms has generated the table below.

Name	Music
Student 1	Country
Student 2	Rap
Student 3	Rap
Student 4	Rock
Student 5	Country
Student 6	Country
Student 7	Rap
Student 8	Rock
Student 9	Rap
Student 10	Rock
Student 11	Country
Student 12	Rap
Student 13	Rap
Student 14	Rap
Student 15	Country
Student 16	Country
Student 17	Rock
Student 18	Rap
Student 19	Country
Student 20	Rap
Student 21	Rap
Student 22	Country
Student 23	Rap
Student 24	Rap

***Fig 3.25** Survey for type of music preferred*

Analyze the data

There are multiple ways to organize and represent the raw data.

For instance, some junior lever students might create a bar graph by lining up according to their favorite music type. Alternatively, they could use sticky notes on the board or floor to represent their category. Then they can count the number of students in each line or sticky notes in each category.

Frequency table of music preferences

A frequency table is a tabular representation that summarizes the raw categorical data. You might first use tally marks to track the categorical data before finding frequencies (counts) for each category

Type of music	Number of students who like this kind of music
Country	8
Rap	12
Rock	4

***Fig 3.26** Frequency Table*

You might also use a picture graph to represent the distribution of the categorical variable type of music. The distribution summarizes the data for the variable type of music by identifying the frequencies for each of the three categories.

A picture graph uses a picture of some sort (such as a musical instrument) to represent an individual's preference.

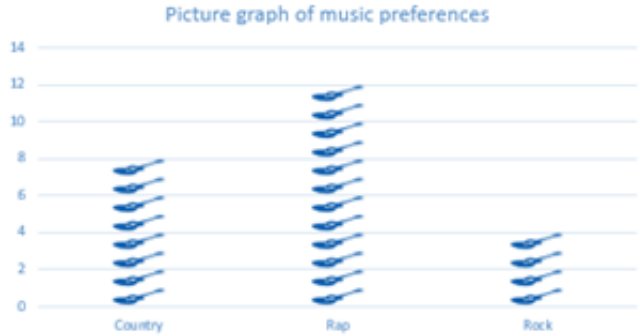


***Fig 3.27** Picture graph of music preferences*

Let us now plot this in a column chart

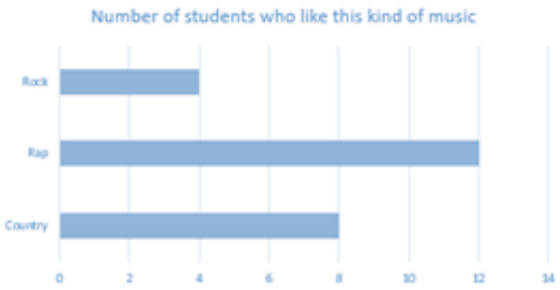


Fig 3.28 Column chart of music preferences

Representing the data using a pie chart

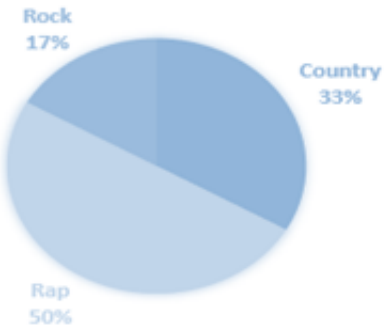


Fig 3.29 Pie chart

Interpret the results

At this point in the investigation, you should try to answer the initial survey question:

What type of music do the students in my class like?

Your answer might be: The most popular type of music in my class is Rap.

So, you can assume or infer that rap will be the most popular type of music in your grade.

The answer above could include the following description:

A total of 12 students preferred Rap, while only eight preferred Country and 4 preferred Rock. There were eight more students who preferred Rap rather than Rock.

Or

The data shows that there was three times the number of students who preferred the most popular category than the number of students who preferred the least popular type.

The first stage is for you to read and interpret what the data shows about your class at a simple level. Recall that the original statistical investigative question asks about a specific grade level.

You should think about how well the class findings would represent other classes in your grade level and if the results would "scale up" to this larger group.

Activity: Choosing the food item to be sold for the year-end party

In this activity, you will choose the food item that will be sold for the year-end party for your grade. You will conduct a survey, summarize the data, analyze the results, and then from the data, we will interpret the results.

Formulate survey question

You may be interested in the favorite type of food among your classmates. Imagine a year-end party is being

planned for your grade, and there is only enough money to hire one restaurant to cook a food item.

The survey question you may have is:

Which food item will the students in my class like?

This survey question attempts to measure the type of food preference in the students' population in your class.

Collect Data

To answer the survey question, you need to collect data about the food your classmates like.

Before beginning data collection, it is vital to think through the data collection methods.

Now you need to think about how you will collect and record the data and whom you can collect the data from.

One possible survey question you could ask:

What is your favorite type of food?

However, the survey question in this form could elicit many different responses, making it challenging to analyze the data.

It would help if you improved on survey questions by understanding potential pitfalls to avoid in survey design (such as ambiguous wording and leading questions) and providing more choices in answers.

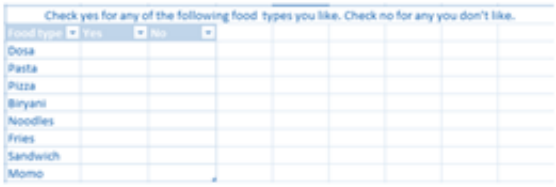


***Fig 3.30** Survey Question*

Because this question explicitly asks your classmates to choose among the options, it will be easier to manage and analyze the data.

You can also design the data collection to use technology. Online survey tools offer ways to collect data and then download the resulting data into a spreadsheet. Having data accessible in a spreadsheet allows you to begin analyzing data using technology. With the data collected in a central space, it is easier for others to analyze and interpret.

Analyze the data

There are multiple ways to organize and represent the raw data.

For instance, some junior lever students might create a bar graph by lining up according to their favorite food type. Alternatively, they could use sticky notes on the board or floor to represent their category. Then they can count the number of students in each line or sticky notes in each category.

Frequency table of music preferences

A frequency table is a tabular representation that summarizes the raw categorical data. You might first use tally marks to track the categorical data before finding frequencies (counts) for each category.

Food type	Yes	No
Dosa	25	16
Pasta	34	2
Pizza	51	4
Biryani	30	7
Noodles	35	4
Fries	20	24
Sandwich	29	3
Momo	30	1

***Fig 3.31** Frequency Table*

To analyze the survey data collected, you could graph the number of students who like each type of food

Let us now plot this in a Multi-Variable Bar Graph

***Fig 3.32** Bar Graph*

The bar graph shows the frequencies of students who like and dislike each type of food.

From this graph, you can see that Pizza is most liked by the students in the class. Noodles are second-most favored, followed by Pasta.

Momo, Sandwich, Biryani, and Dosa also have more students who like it than dislike it.

Fries has more students saying no than saying yes.

The graph suggests that Pizza, Noodles, and Pasta are the most preferred food items for the class.

Representing the data using a Column chart

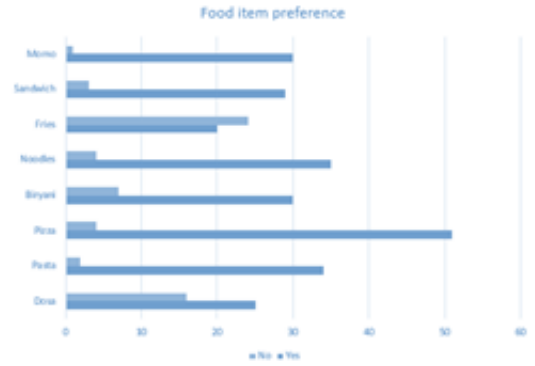


***Fig 3.33** Column Chart*

Plotting the data using a clustered column-Line chart

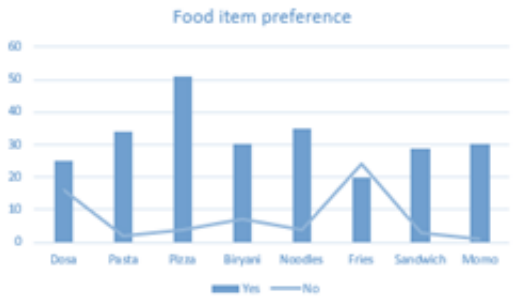


Fig 3.34 Column-Line Chart

Representation using a pie chart for the students who said yes to different food items:

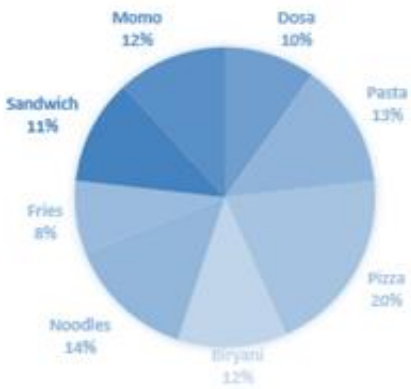


Fig 3.35 Pie Chart

Interpret the results

At this point in the investigation, you should try to answer the initial survey question:

What type of food do the students in my grade like?

Your answer might be: The most preferred type of food in my class is Pizza.

So, you can assume or infer that Pizza will be the most preferred food item in your grade.

The first stage is for you to read and interpret what the data shows about your class at a simple level. Recall that the original statistical investigative question asks about your grade level.

You should think about how well the class findings would represent other classes in your grade level and if the results would "scale up" to this larger group.

Activity: Growing beans

In this activity, you will be taking measurements on a condition or group. A simple comparative experiment is like a science experiment in which you compare the results of two or more conditions. For example, you might plant dried beans in two different growing conditions and let them sprout, and then compare which group grows fastest for the period—the ones in the light or the ones in the dark. The data collected lead to a comparative experiment investigation.

You also might collect the growth data for your plants each day over the period, which allows for a time series investigation.

Formulate statistical investigative questions

In the comparative activity, a statistical investigative question might be:

After two weeks, do beans grown in the light tend to be taller than beans grown in the dark?

With comparison statistical investigative questions, the posed statistical investigative question is stated like a hypothesis (or conjecture) to indicate what the investigator believes will be the group that is bigger/ larger/taller.

In this activity, the hypothesis is that beans grown in the light will grow taller than beans grown in the dark.

Collect/consider data

It would be best to decide which beans will be grown in a light environment and grown in a dark environment.

Comparing these two conditions is the goal of the activity. The type of lighting environment is an example of a categorical variable. Measurements of the plants' heights, a continuous quantitative variable, can be taken at the end of a specified period. These measurements can answer the statistical investigative question of whether one lighting environment is better for growing beans.

When going through this activity, you need to establish criteria with your classmates about handling certain situations that may arise. For example, some seeds may never grow, or individual plants may die, and you must decide how you will account for this in their data collection.

You can record the beans' height (in centimeters) grown in the dark and light using a dot plot.

The heights on day 8 are shown in the below table.

	Height in cm										
Environment	0	1	2	3	4	5	6	7	8	9	10
Dark		4	1	2	2	1					
Light						2	3	2	1	1	1

Fig 3.36 Data Table

The heights on day 8 are represented in the below figure.

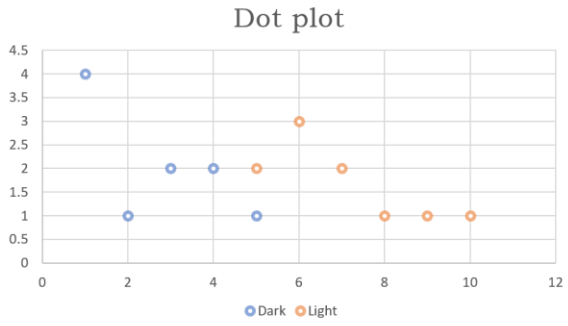


Fig 3.37 Dot plots for bean height for light and dark conditions

You can now make statements about what you notice in the dot plot.

For example, you notice that the median height for beans grown in the dark is 2.5

cm compared to the median height for beans grown in the light of 7.5 cm.

It is also useful to know how the data vary across the horizontal axis. One measure of variability for distribution is the range, which is the difference between the maximum and minimum values. Range only makes sense with data for a quantitative variable.

The variability of the distribution also can be described by providing the lowest and highest values.

For example, the heights of beans grown in the dark vary from 1 cm to 5 cm, whereas the beans are grown in the light range in height from 5 cm to 10 cm. This is more informative for describing the distribution than just providing the ranges (for example, that the range of heights for beans grown in the dark is 4 cm and the range of heights for the beans grown in the light is 5 cm) because this provides measures of location in addition to variability.

Looking for clusters and gaps in the distribution helps you to identify the shape of the distribution. You should develop a sense of why a distribution takes on a particular shape for the context of the variable being considered.

- Does the distribution have one main cluster (or mound) with smaller groups of similar size on each side of the cluster? If so, the distribution might be described as symmetric.
- Does the distribution have one main cluster with smaller groups on each side that are not the same size? Students may classify this as "lopsided" or may use the term asymmetrical.
- Why does the distribution take this shape?

Using the dot plot from Figure 3.37, you will recognize both the beans grown in the dark and beans grown in the light have distributions that are "lopsided," with the main cluster on the left end of the distributions a few values to the right, which is also known as right-skewed distribution.

Interpret the results

The analyses reveal that the median height for the beans grown in the light exceeds the median height for the other beans by 5 cm.

Nearly all the beans grown in the light are taller than the beans grown in the dark. The beans that overlap are the shortest in the light condition and the other condition's tallest beans.

Thus, the plants from our experiment in the light environment tend to be taller than the plants in the dark environment.

What did you learn?

- Data visualization is the mechanism of representing raw data in forms of graphical representations that allow users to explore the data and uncover quick insights.
- Visualizations allow us to recognize trends, patterns, and outliers from seemingly meaningless records of data.
- A dot plot is a graphical display of data using dots.
- A bar graph is a graphical display of data using bars of different height. It is possible to plots the bars vertically or horizontally.
- The frequency of a data value is the number of times the data value occurs/repeats.
- Histogram is a graphical representation of frequency plotted against intervals.
- Use of single and multi-variable plots

Exercises

Objective Type Questions

1. Data can be visualized using:
    a. Graphs
    b. Maps
    c. Charts
    d. All of the above
2. Which of the following statements is false?
    a. Data visualization can absorb information quickly.
    b. Data visualization decreases the insights and takes slower decisions.
    c. Data visualization is a type of visual art.
    d. None of the above
3. Which of the following is a use case of data visualization?
    a. Healthcare
    b. Sales and Marketing
    c. Politics/Campaigning

Microsoft

d. All of the above

4. Bar Graph is a
   a. One-dimensional graph
   b. Two-dimensional graph
   c. Graph with no dimension
   d. None of the above
5. The data represented through a histogram can help in finding graphically the
   a. Median
   b. Mean
   c. Mode
   d. All of the above
6. Pie Chart is a
   a. One-dimensional graph
   b. Two-dimensional graph
   c. Graph with no dimension
   d. None of the above
7. Can a Line chart be used to plot multiple variables?
   a. Yes
   b. No
8. The height of your classmates is recorded and arranged in ascending order. The data is represented as a histogram. What type of shape does the histogram have
   a. Right-skewed Distribution
   b. Left-skewed Distribution
   c. Bimodal Distribution
   d. Random Distribution

Standard Questions

1. Give a few examples of real-life use of data visualization.
2. Explain the importance of data visualizations.
3. Explain a few graphs/charts used for data visualization.
4. Give few examples of Multi-Variable and Single-variable plots
5. The value of pi (π) up to 55 decimal places is given below:
   3.14159265358979323846 26433832795028841971 69399375105820 9
   a. Make a frequency distribution table for the digits from 0 to 9 after the decimal point.
   b. Find the most and the least frequently occurring digits?

Higher Order Thinking Skills (HOTS)

1. Record the number of students absent in your class in the past one month in a tabular format. Display the data as a dot plot. Find the Maximum and the mean of the data.

2. The height of your classmates varies from student to student. Record the height of your classmates in a tabular format. Visualize the data using a histogram and find the mode of the height of your class.

3. Form statistical investigative questions to find the favorite fruit of your class. Collect the data, analyze the data and then interpret the results for your school's favorite fruit.

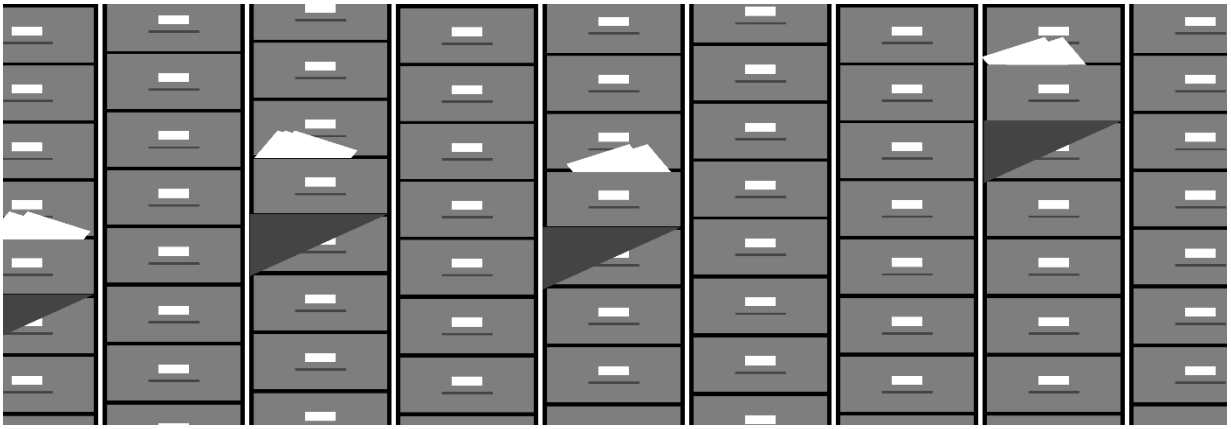
Applied Project

Over the past few decades, the earth's temperature has been rising. Collect the data from various sources, analyze the data, represent the data using various plots.

CHAPTER

4

Ethics in Data Science



Studying this chapter should enable you to understand:

- Ethical guidelines around data analysis
- Need for ethical guidelines in data analysis
- Goals of ethical guidelines in data analysis
- Data governance framework
- Why do we need to govern data?
- Goals of data governance

1. Introduction

In the previous chapters, we have understood data, visualizing data, categorizing data, different data types, etc.

In this chapter, we are going to understand ethical guidelines and data governance frameworks in data science. We will also learn the goals and need for ethical guidelines and data governance framework in data science.

2. Ethical guidelines around data analysis

You should become aware of how data are stored and how they are used by the administrations receiving them.

It would be best if you also understood that we need protective mechanisms and policies to discourage the mishandling and unethical use of data.

***Fig 5.1** Privacy*

The few ethical guidelines around data analysis are:

- Data governance is critical
- Protect your customer
- Do not lie
- Understand the role of data quality
- Private data and identity should remain private
- Shared private information should be treated confidentially

3. Need for ethical guidelines

There are many reasons why adhering to ethical guidelines in data analysis is essential.

First, guidelines encourage facts, knowledge, and error avoidance.

For Example, prohibitions against falsifying, fabricating, or misrepresenting data promote the truth and minimize error.

Second, Ethical guidelines in data analysis also help to build public support for the analysis. People are more likely to confide in the analysis if they can trust data quality and integrity.

Third, before conducting any new analysis, we ask ourselves whether it will benefit the people. If it doesn't, we will not do it.

Fourth, minimizing data usage, we should use the least amount of data necessary to meet the desired objective, understanding that reducing data usage encourages more sustainable and less risky analysis.

4. Goals of Ethical guidelines

The goal of ethical guidelines is to help data analysts make decisions ethically. Moreover, the ethical guidelines aim to encourage accountability by enlightening those who rely on data analysis of the standards they should expect.

The key goals are:

- Professional integrity and accountability
- The integrity of data and methods
- Follow informed-consent rules
- Respect confidentiality and privacy

5. Data governance framework

The data governance framework is used for determining who has control and

power over data assets within a group and how such data assets can be used. It includes the entities, procedures, and technology needed to handle and secure data assets

A data governance framework provides a comprehensive approach to managing, collecting, securing, and storing data.

Data governance means cleaner, leaner, better data, which means better analytics, which means better decisions, which means better results.

Efficient data governance means the data is consistent and credible and is not misused

6. Need to govern data

- To Improve data quality through efforts to identify and fix errors in data sets.
- To increase analytics accuracy and give decision-makers reliable information.
- To ensure compliance with data privacy laws and other regulations
- To implement and enforce policies that help prevent data errors and misuse
- To avoid inconsistent data in different departments and business units
- To come to an agreement on standard data definitions for a shared understanding of data

7. Goals of data governance

The goal of data governance is to create methods, set of responsibilities, and processes to standardize, integrate, protect, and store data. The key objectives are:

- To improve internal and external communication
- To increase the value of data
- To reduce costs
- To implement compliance requirements
- To minimize risks
- To establish internal rules for data use

What did you learn?

- In this chapter we have learn about ethical guidelines around data analysis, the goals and need for ethical guidelines.
- We have also learnt about data governance framework and the goals and need for data governance.

Exercises

Objective Type Questions

1. We need policies to discourage unethical use of data
    a. Yes
    b. No
2. You take some data for research from few users, Is it ethical to share the results back to the participants
    a. Yes
    b. No
3. Is it ethical to let your preconception or opinions interfere in the data collection process?
    a. Yes
    b. No
4. Can you lie about the data collected and show the results in your favor?
    a. Yes
    b. No
5. Should shared private information be treated confidentially
    a. Yes
    b. No
6. For assuring people to give data for your analysis, is it a good practice to let people who you are
    a. Yes
    b. No
7. Data governance helps in which of the following
    a. Cleaner data
    b. Better data
    c. Leaner data
    d. All of the above
8. Data governance helps to prevent errors and misuse of the data
    a. Yes
    b. Noe

Standard Questions

1. Explain the goals of ethical guidelines in your word.
2. Describe the goals of data governance in your word.

Higher Order Thinking Skills (HOTS)

Now you know about ethical guidelines and data governance. Make a list of all the apps or browsers which ask you to accept their terms and conditions and use your private information.

Applied Project

Suppose you visit a supermarket for buying groceries. At the end of the purchase, the manager provides you with a form wherein you need to fill up your details along with your contact number. According to the manager, the purpose for collecting this data is to enable them to inform us of the exciting deals and new product launches. Explain in detail, the precautions you need to take before handing out your details to the manager.

Final Project

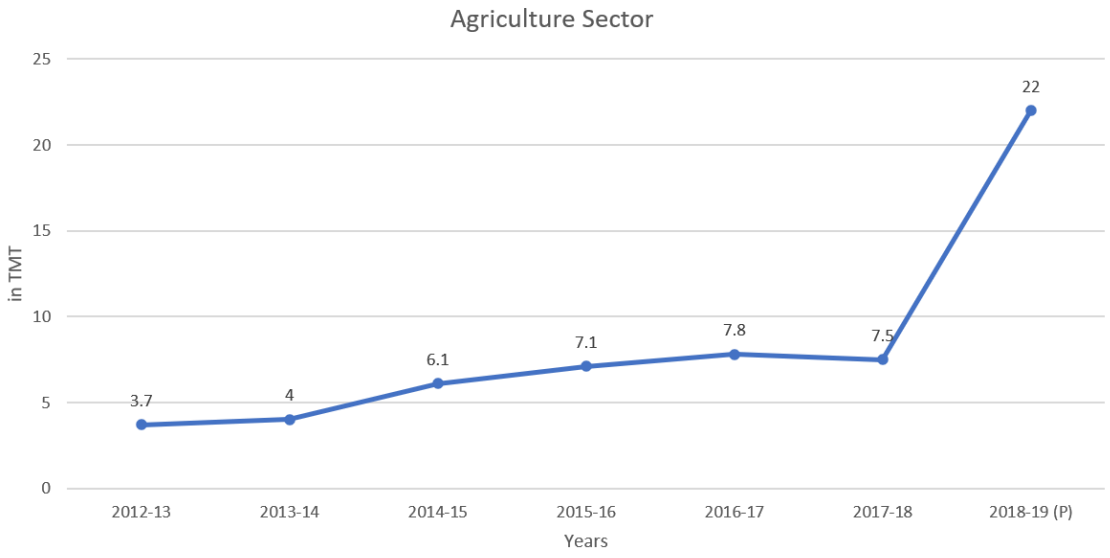
Project 1: Consumption of LPG in different sector in India from 2012-13 to 2018-19

In this project, we have the data for the annual consumption of LPG in different sectors in India from 2012-13 to 2018-19.

You can get the data set from this link https://community.data.gov.in/lpg-consumption-in-agriculture-sector-in-india-from-2012-13-to-2018-19/, and in this link, a dataset URL is provided; you can download the dataset in xls format to do more further analysis.

Example: The consumption of LPG in agriculture sector was 3.7 thousand metric tonnes (TMT) in India during 2012-12. It has surged by 8.11% to 4 TMT during 2013-14 against 2012-13. An annual decline of -3.85% has been seen in the consumption of LPG in the agriculture sector during 2017-18 from 7.8 TMT during 2016-17. Consumption of LPG in agriculture sector was 22 TMT in India during 2018-19, up by 193.33% versus 7.5 TMT during 2017-18.

Let us now visualize the data for the consumption of LPG in India's agriculture sector from 2012-13 to 2018-19.



LPG consumption in Agriculture sector in India from 2012-13 to 2018-19.

1)Using a multi-line chart in open office spreadsheet, plot the graph for the LPG consumption in different sectors and the total consumption from 2012-13 to 2018-19 in India using the dataset.

2)Create a multi-line chart using open office spreadsheet, to display the LPG consumption of different sectors in Manufacturing (Bulk LPG). For which sector do we see the maximum growth from 2012-13 to 2018-19.

3)Create a pie chart using open office spreadsheet to display the total consumption of LPG from the period 2012-13 to 2018-19 and find the year with the highest consumption of LPG.

4)Plot a multivariable bar graph to compare LPG consumption on Retail and Miscellaneous (Bulk) from 2012-13 to 2018-19. For which period do we see the highest growth for the LPG consumption for the Miscellaneous (Bulk) sector.

Project 2: Digital Payment Transactions in India from 2016-17 to 2019-20

In this project, we have the data for the number of digital transactions in India from 2016-17 to 2019-20.

You can get the data set from this link https://community.data.gov.in/digital-payment-transactions-in-india-from-2016-17-to-2019-20/, and in this link, a dataset URL is provided; you can download the dataset in xls format to do more further analysis.

The government of India has taken numerous initiatives to promote digital transactions across the country. There has been a significant increase in the usage of digital payments across the country since 2016. The compound annual growth rate of 75.88% has been seen in the transaction of digital payments in India during 2016-17 to 2018-19.

The number of digital payment transactions was 1013 crore in India during 2016-17. It has increased by 104.38% during 2017-18 above 2016-17. In absolute terms, the increase was 1057.39 crore in 2017-18 versus 2016-17. We have observed annual growth of 51.35% in digital payments transactions to 3133.58 crores during 2018-19 against 2070.39 crores during 2017-18. In absolute terms, the increase was 1063.19 crore in 2018-19 above 2017-18. The number of digital payment transactions in India was 975.98 crore during 2019-20 (till 11th July 2019).

Plot a column chart using open office spreadsheet to visualize the digital transactions for the period and determine which year had the maximum transactions.

References

Diffen. 2021. Data vs Information - Difference and Comparison. [Online]. [Accessed 25 February 2021]. Available from: https://www.diffen.com/difference/Data_vs_Information

Pew research center: internet, science & tech. 2021. Digital Footprints. [Online]. [Accessed 25 February 2021]. Available from: https://www.pewresearch.org/internet/2007/12/16/digital-footprints

Graziadio business review. 2021. The Cost of Lost Data - A Peer-Reviewed Academic Articles. [Online]. [Accessed 25 February 2021]. Available from: https://gbr.pepperdine.edu/2010/08/the-cost-of-lost-data

Jure Leskovec. 2021. CS246: Mining Massive Data Sets. [Online]. [Accessed 25 February 2021]. Available from: http://web.stanford.edu/class/cs246

Microsoft Corporation. 2021. Data Science for Beginners video 1: The 5 questions data science answers. [Online]. [Accessed 25 February 2021]. Available from: https://docs.microsoft.com/en-us/azure/machine-learning/classic/data-science-for-beginners-the-5-questions-data-science-answers

Bargagliotti, A., Franklin, C., Arnold, P. and Gould, R., 2020. Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II). American Statistical Association