Videos by : Ravi kumar R,Lecturer in science,GPT Bagepalli

UNIT-2
SUMMARIZATION OF DATA

Descriptive Statistics:

Descriptive statistics involve summarizing and describing data to understand its main features.

Example: Consider the heights (in cm) of a group of students - {150, 155, 160, 165, 170, 175, 180}.

Mean: (150 + 155 + 160 + 165 + 170 + 175 + 180) / 7 = 164.29 cm

Median: Arrange data in ascending order - {150, 155, 160, 165, 170, 175, 180}.

Since there are 7 data points, the median is the middle value, which is 165 cm.

Mode: The most frequent value in the data set is 170 cm.

Measures of Central Tendency: Mean, Median, and Mode

Mean:

The mean is the average value of a set of data. It represents the central value around which the data points tend to cluster.

To calculate the mean, sum up all the data values and divide by the total number of data points.

Example 1: Consider the following test scores of 6 students: {80, 85, 90, 75, 88, 92}. To find the mean, add up all the scores and divide by the total number of students.

Mean = (80 + 85 + 90 + 75 + 88 + 92) / 6 Mean = 510 / 6 Mean = 85

Example 2: Suppose you have a dataset of 10 students' ages: {18, 20, 21, 18, 22, 19, 23, 20, 22, 21}. To find the mean, sum up all the ages and divide by the total number of students.

Mean = (18 + 20 + 21 + 18 + 22 + 19 + 23 + 20 + 22 + 21) / 10 Mean = 204 / 10 Mean = 20.4

Median:

The median is the middle value of a dataset when arranged in ascending or descending order.

If the dataset has an odd number of data points, the median is the middle value.

If the dataset has an even number of data points, the median is the average of the two middle values.

Example 1: Consider the following test scores of 7 students: {70, 80, 85, 75, 90, 85, 88}. To find the median, first arrange the data in ascending order: {70, 75, 80, 85, 85, 88, 90}. Since there are 7 data points, the median is the middle value, which is 85.

Example 2: Suppose you have a dataset of 8 students' heights (in cm): {160, 155, 165, 170, 168, 163, 172, 157}. To find the median, first arrange the data in ascending order: {155, 157, 160, 163, 165, 168, 170, 172}. Since there are 8 data points, the median is the average of the two middle values, which are 163 and 165.

Median = (163 + 165) / 2 Median = 328 / 2 Median = 164

Mode:

The mode is the value that appears most frequently in a dataset.

A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal).

Example 1: Consider the following test scores of 9 students: {70, 85, 90, 85, 75, 80, 85, 90, 80}. To find the mode, determine which value appears most frequently. In this dataset, the value 85 appears three times, making it the mode.

Example 2: Suppose you have a dataset of 12 students' test scores: {70, 80, 85, 90, 85, 75, 80, 85, 90, 75, 80, 85}.

Frequency : The **frequency** of a particular data value is the number of times the data value occurs.

A **frequency distribution table** is a chart that summarizes values and their frequency. It's a useful way to organize data if you have a list of numbers that represent the frequency of a certain outcome in a sample. A frequency distribution

table has two columns. The **first column** lists all the various outcomes that occur in the data, and the **second column** lists the frequency of each outcome. Putting this kind of data into a table helps make it simpler to understand and a

Make a Frequency Distribution Table

Data Tabulation (Frequency Table):

Data tabulation involves organizing raw data into a table to summarize its distribution.

A frequency table lists the individual data values along with their corresponding frequencies or counts.

Example: Consider the test scores of 15 students - {75, 80, 85, 85, 90, 80, 85, 75, 80, 85, 85, 90, 75, 85, 85}.

The frequency table for this data will be:

Test Score	Frequency
75	3
80	3
85	7
90	2

A relative frequency table shows the proportion or percentage of each data value relative to the total number of observations.
Example: Using the previous test scores data, the relative frequency table will be:

Test Score	**Relative Frequency (%)**
75	(3 / 15) * 100 ≈ 20%
80	(3 / 15) * 100 ≈ 20%
85	(7 / 15) * 100 ≈ 47%
90	(2 / 15) * 100 ≈ 13%

Ungrouped data is the data you first gather from an experiment or study. The data is raw — that is, it's not sorted into categories, classified,

Grouped data is data that has been bundled together in categories.

Grouped Data:

Grouping data is useful when dealing with large datasets. It involves categorizing values into intervals or classes for easier analysis.

Example: Consider the test scores of 30 students - {56, 65, 73, 85, 88, 77, 68, 71, 60, 92, 80, 75, 79, 84, 58, 63, 91, 66, 87, 75, 72, 81, 78, 62, 86, 69, 83, 74, 70}.

Group the data into intervals: 50-60, 61-70, 71-80, 81-90.

Charts convey information about our data faster than tables.

BAR GRAPH

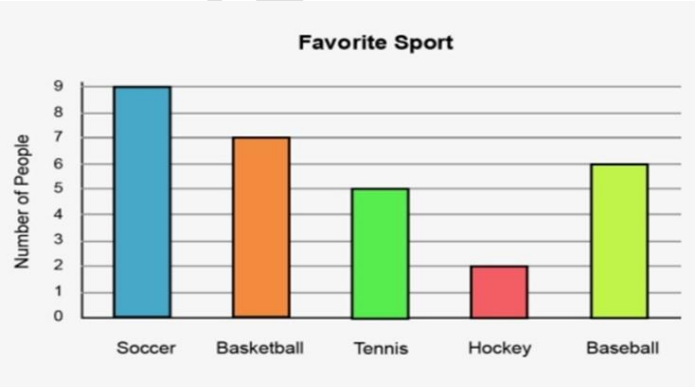
A bar graph represents categorical or discrete data using rectangular bars of equal width. The height of each bar corresponds to the frequency or relative frequency of the category it represents.

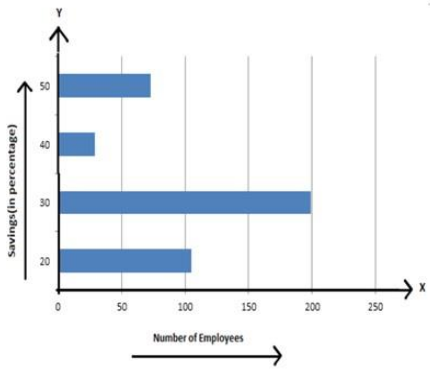
Bar graphs consist of two axes. On a vertical bar graph, as shown above, the horizontal axis (or x-axis) shows the data categories.

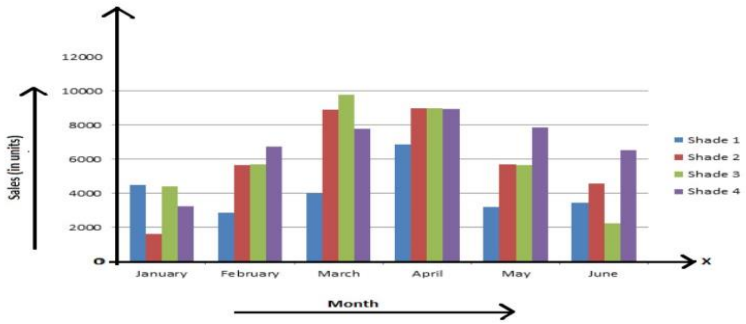
Bar graphs have three key attributes:

- A bar diagram makes it easy to compare sets of data between different groups at a glance.
- The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes.
- Bar charts can also show big changes in data over time.

Interpret the following graphs







When to Use a Bar Graph:

Bar graphs are an effective way to compare items between different groups. This bar graph shows a comparison of numbers on a quarterly basis over a four-year period of time. Users of this chart can compare the data by quarter on a year-over-year trend, and also see how the annual sales are distributed throughout each year. Bar graphs are an extremely effective visual to use in presentations and reports. They are popular because they allow the reader to recognize patterns or trends far more easily than looking at a table of numerical data.

PIE CHART

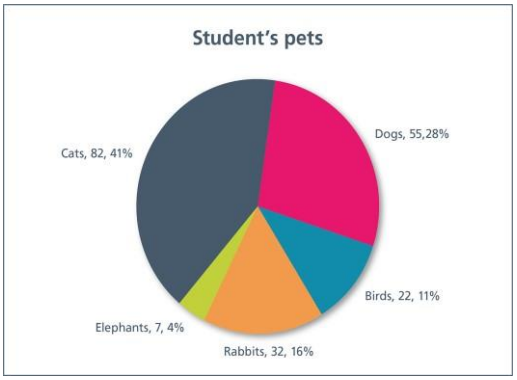
The "pie chart" also is known as "circle chart", that divides the circular statistical graphic into sectors or slices in order to illustrate the numerical problems. Each sector denotes a proportionate part of the whole

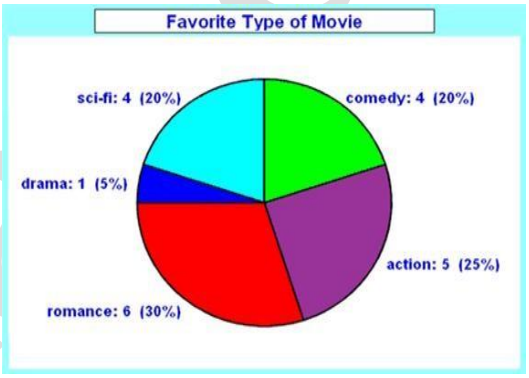
What data can be presented using a pie chart?

1. Pie charts are a visual way of displaying data that might otherwise be given in a small table.
2. Pie charts are useful for displaying data that are classified into nominal or ordinal categories. Nominal data are categorised according to descriptive or qualitative information such as county of birth or type of pet owned. Ordinal data are similar but the different categories can also be ranked, for example in a survey people may be asked to say whether they classed something as very poor, poor, fair, good, very good.
3. Pie charts are generally used to show percentage or proportional data and usually the percentage represented by each category is provided next to the corresponding slice of pie.

Pie charts are good for displaying data for around 6 categories or fewer. When there are more categories it is difficult for the eye to distinguish between the relative sizes of the different sectors and so the chart becomes difficult to interpret.

Interpret the following graphs





Line Graph

A line graph is a graph that utilizes points and lines to represent change over time. It is a chart that shows a line joining several points or a line that shows the relation between the points. The graph represents quantitative data between two changing variables with a line or curve that joins a series of successive data points. Linear graphs compare these two variables in a vertical axis and a horizontal axis.

Bar Graph vs Line Graph

Bar graphs display data in a way that is similar to line graphs. Line graphs are useful for displaying smaller changes in a trend over time. Bar graphs are better for comparing larger changes or differences in data among groups.

A line graph shows how values change. For example, you could plot how your child grows over time. Line graphs can also be used to show how functions change. The most usual type of data you'll find on a line graph is how something changes over time.

Why use a Line Graph?

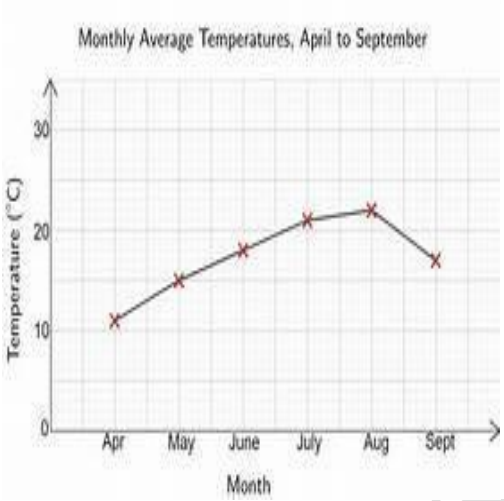
A line graph has characteristics that make it useful for some situations.

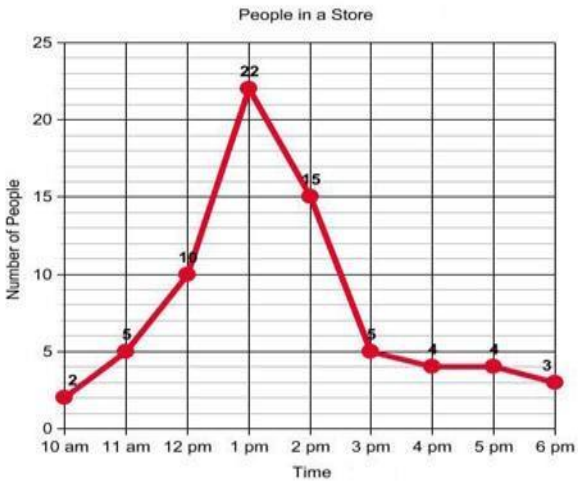
You would use a line graph if:

- You have a function. Line graphs are good at showing specific data values, meaning that if you have one variable (x) you can easily find the other (y).

- You want to show trends. For example, how your investments change over time or how food prices have increased over time. • You want to make predictions. A line graph can be extrapolated beyond the data at hand. They enable you to make predictions about the results of data.

Interpret the following graphs

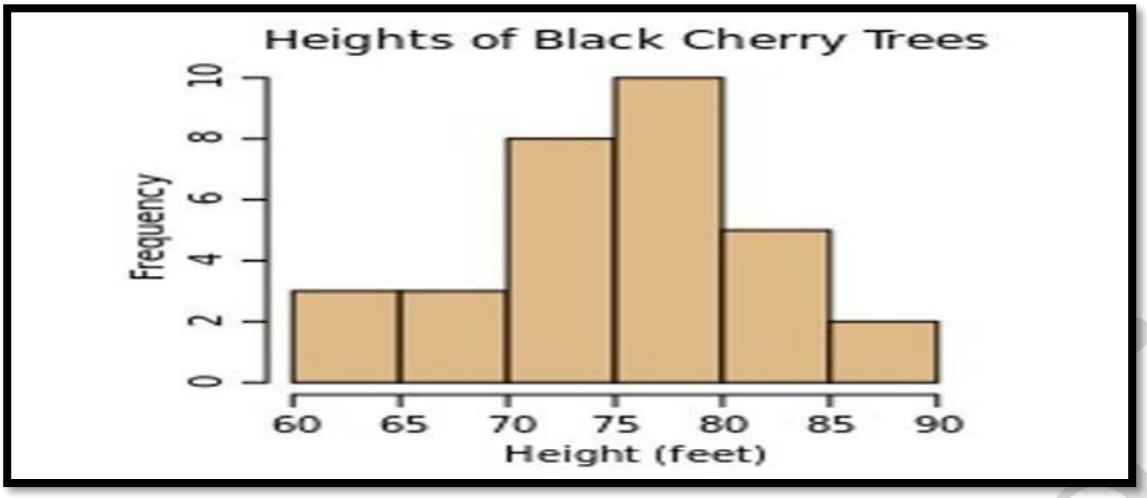




What is a Histogram?

A histogram is an area diagram. It can be defined as a set of rectangles with bases along with the intervals between class boundaries and with areas proportional to frequencies in the corresponding classes. In such representations, all the rectangles are adjacent since the base covers the intervals between class boundaries. The heights of rectangles are proportional to corresponding frequencies of similar classes and for different classes, the heights will be proportional to corresponding frequency densities.

In other words, histogram a diagram involving rectangles whose area is proportional to the frequency of a variable and width is equal to the class interval. A histogram is used to summarize discrete or continuous data. In other words, it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values (called "bins"). It is similar to a vertical bar graph. However, a histogram, unlike a vertical bar graph, shows no gaps between the bars.



Importance of a Histogram

Creating a histogram provides a visual representation of data distribution. Histograms can display a large amount of data and the frequency of the data values. The median and distribution of the data can be determined by a histogram. In addition, it can show any outliers or gaps in the data.

Difference Between Histogram And Bar Graph

Histogram	Bar Graph
It is a two-dimensional figure	It is a one-dimensional figure
The frequency is shown by the area of each rectangle	The height shows the frequency and the width has no significance.
It shows rectangles touching each other	It consists of rectangles separated from each other with equal spaces.

Frequency Polygon

A frequency polygon is almost identical to a histogram, which is used to compare sets of data or to display a cumulative frequency distribution. It uses a line graph to represent quantitative data.

Frequency polygons are a visually substantial method of representing quantitative data and its frequencies. Let us discuss how to represent a frequency polygon.

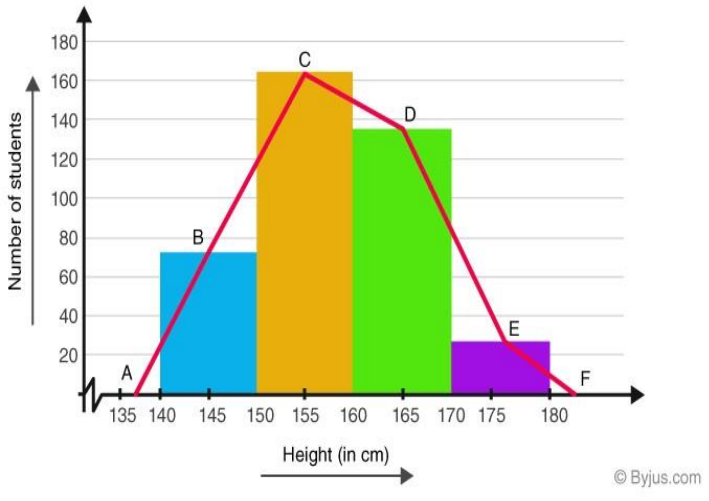
Steps to Draw Frequency Polygon:

To draw frequency polygons, first we need to draw histogram and then follow the below steps:

- **Step 1-** Choose the class interval and mark the values on the horizontal axes • **Step 2-** Mark the mid value of each interval on the horizontal axes.
- **Step 3-** Mark the frequency of the class on the vertical axes.
- **Step 4-** Corresponding to the frequency of each class interval, mark a point at the height in the middle of the class interval
- **Step 5-** Connect these points using the line segment.
- **Step 6-** The obtained representation is a frequency polygon.

Example of Frequency Polygon:

Let us consider an example to understand this in a better way.
In a batch of 400 students, the height of students is given in the following table. Represent it through a frequency polygon.



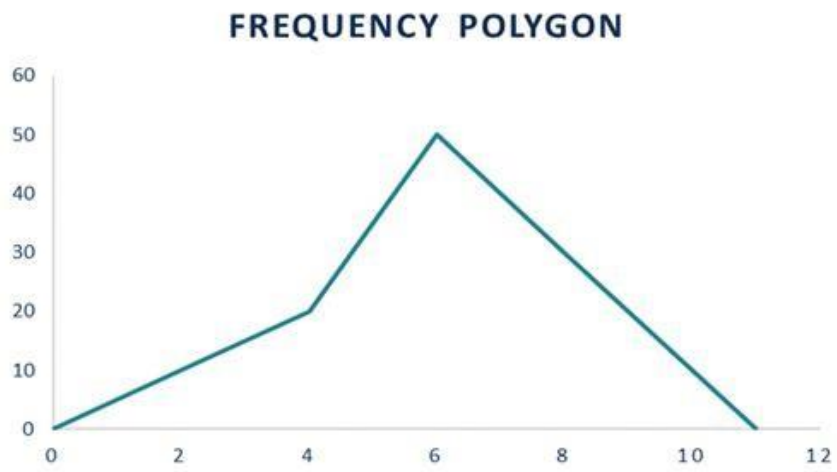
Height (in cm)	Number of Students(Frequency)
140 – 150	74
150 – 160	163
160 – 170	135
170 – 180	28
Total	400

The frequency polygon can be easily utilized to compare multiple distributions on the same graph. In some cases, a histogram and a frequency polygon can be used simultaneously to get a more accurate picture of the distribution shape.

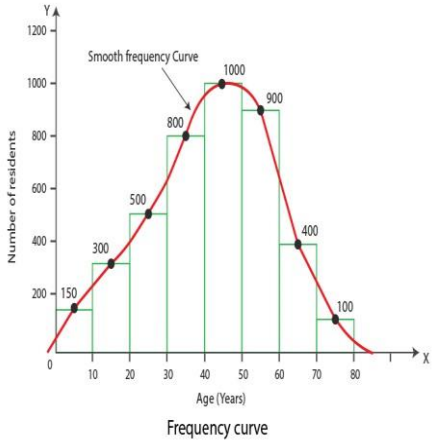
You are a financial analyst in a retail business. You are preparing a report regarding the current financial conditions of the company. One part of the report describes the management of the company's accounts payable. You obtain the data that defines how many days are required to settle each invoice.

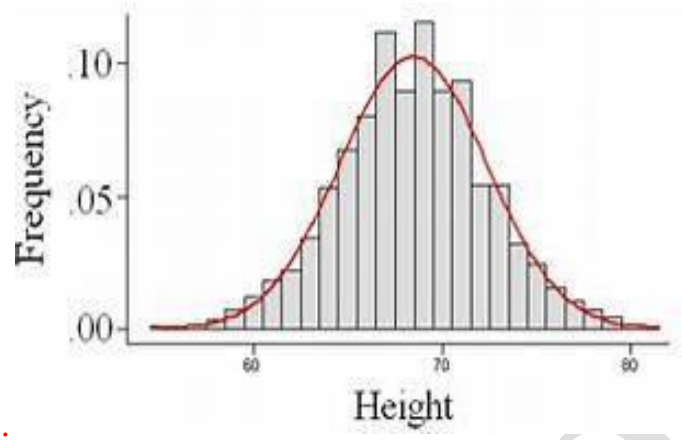
Days	Number of Invoices
1-3	10
3-5	20
5-7	50
7-9	30

Class	Midpoint	Frequency
0	0	0
1-3	2	10
3-5	4	20
5-7	6	50
7-9	8	30
10+	11	0



Frequency curve is obtained by joining the points of frequency polygon by a freehand smoothed curve



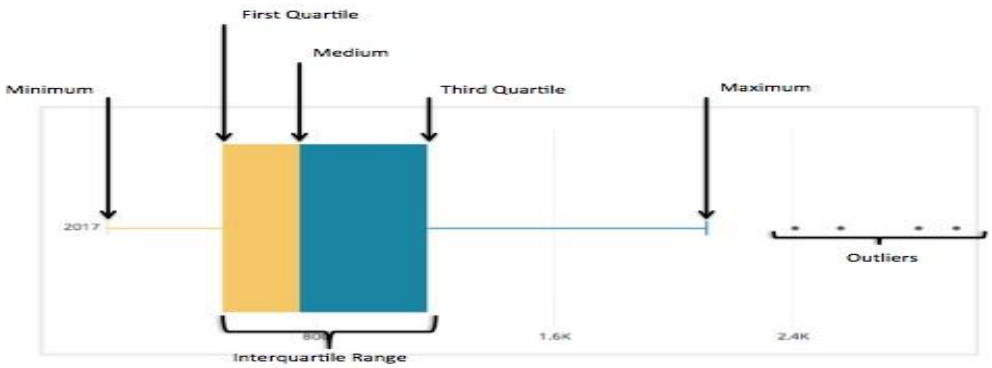


Boxplot:

In descriptive statistics, a **boxplot** is a method for graphically depicting groups of numerical data through their quartiles.

A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode). You need to have information on the variability or dispersion of the data. A boxplot is a graph that gives you a good indication of how the values in the data are spread out.



Stem-and-Leaf Plot:

A plot where each data value is split into a "leaf" (usually the last digit) and a "stem" (the other digits).
For example "32" is split into "3" (stem) and "2" (leaf).
Complete a stem-and-leaf plot for the following list of grades on a recent test:
73, 42, 67, 78, 99, 84, 91, 82, 86, 94

I'll use the tens digits as the stem values and the ones digits as the leaves. For convenience sake, I'll order the list, but this is not required:42, 67, 73, 78,82, 84, 86, 91, 94, 99

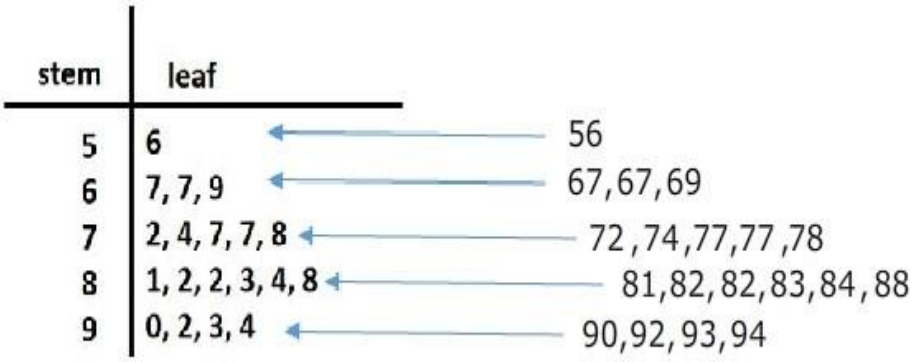
Test grades	
stem	leaf
4	2
5	
6	7
7	3 8
8	2 4 6
9	1 4 9

Creating a Stem and Leaf Plot

Here is a set of data on showing the test scores on the last science quiz.
56, 78, 82, 82, 90, 94, 93, 67, 67, 69, 74, 77, 92, 88, 81, 83, 84, 77, 72
Step 1: In order to create a stem and leaf plot, we need to first organize the data into groups. In this situation, we will group the tests by decades.
56
67, 67, 69
72, 74, 77, 77, 78
81, 82, 82, 83, 84, 88
90, 92, 93, 94
Step 2: Create the plot with the stems as the tens and the leaves as the ones. The stems will be 5, 6, 7, 8 and 9

stem	leaf
5	
6	
7	
8	
9	

Now we are ready to add the ones place from each of the values in the list we made.



A stem and leaf plot is a great way to organize data by the frequency. It is a great visual that also includes the data. So, if needed, you can just take a look to get an idea of the spread of the data or you can use the values to calculate the mean, median or mode.

Graph Type	Simple Explanation	Suitable Data Type
Bar Graph	Rectangular bars for comparing categories	Categorical or Discrete data
Pie Chart	Slices of a circle for showing proportions	Proportions or Percentages
Line Graph	Continuous line connecting data points	Data over Time or Continuously changing data
Frequency Polygon	Line segments representing frequencies	Data Distribution and Frequency Data
Frequency Curve	Smooth curve showing overall data distribution	Data Distribution

Relative Frequency Polygon	Line segments using relative frequencies	Data Distribution and Relative Frequency
Histograms	Bars for displaying frequency of continuous data	Continuous Data and Frequency
Box Plot	Box and whisker plot for data distribution	Distribution and Outliers
Leaf-Stem Plot	Stem and leaves to represent individual values	Individual Data Values