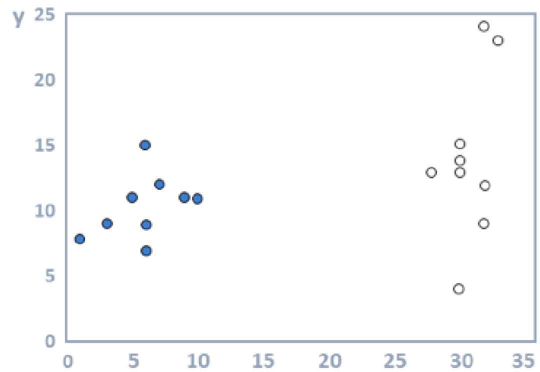
28. Write short notes on the usage of clustering in analysis with an example.

- Cluster is a group of objects that belongs to the same class.
- Clustering is the process of making a group of abstract objects into classes of similar objects.
- The set of data is partitioned into groups based on data similarity and then the labels are assigned to the groups.
- It is adaptable to changes and helps single out useful features that distinguish different groups.



1. Identifying Fake News: The way a clustering algorithm works is by taking in the content of the fake news article, examining the words used and then clustering them. Certain words are found more commonly in sensationalized, click-bait articles. When you see a high percentage of specific terms in an article, it gives a higher probability of the material being fake news.

2. Spam filter: K-Means clustering techniques have proven to be an effective way of identifying spam. The way that it works is by looking at the different sections of the email (header, sender, and content). The data is then grouped together. These groups can then be classified to identify which are spam.

3. Marketing and Sales: Clustering algorithms are able to group together people with similar traits and likelihood to purchase. Once you have the groups, you can run tests on each group with different marketing copy that will help you better target your messaging to them in the future.

4. Classifying network traffic: K-means clustering is used to group together characteristics of the traffic sources. When the clusters are created, you can then classify the traffic types.

5. Identifying fraudulent or criminal activity: By analysing the GPS logs, the algorithm is able to group similar behaviors. Based on the characteristics of the groups you are then able to classify them into those that are real and which are fraudulent.

6. **Document analysis:** Hierarchical clustering has been used to solve this problem. The algorithm is able to look at the text and group it into different themes. Using this technique, you can cluster and organize similar documents quickly using the characteristics identified in the paragraph.

~~~~~~

29. Discuss about different types of clustering
~~~~~~

Clustering methods can be classified into the following categories

1. Partitioning Method: Suppose we are given a database of n objects and the partitioning method constructs k partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements

• Each group contains at least one object.

• Each object must belong to exactly one group.

2. Hierarchical Methods: This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here:

Agglomerative Approach: This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach: This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering:

Here are the two approaches that are used to improve the quality of hierarchical clustering • Perform careful analysis of object linkages at each hierarchical partitioning.

• Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method: This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method: In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages:

• The major advantage of this method is fast processing time.

• It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods: In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method: In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

~~~~~~

30. Explain how different association rules help in analysis.

- Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.
- Association rule mining is suitable for non-numeric, categorical data.
- An association rule has two parts:
  1. An antecedent (if)
  2. A consequent (then).
- An antecedent is something that‘s found in data, and a consequent is an item that is found in combination with the antecedent. For instance:
  "If a customer buys bread, he's 70% likely of buying milk."
  In the above association rule, bread is the antecedent and milk is the consequent.

Depending on the following two parameters, the important relationships are observed:

- Support: Support indicates how frequently the if/then relationship appears in the database.
- Confidence: Confidence tells about the number of times these relationships have been found to be true.

So, in a given transaction with multiple items, Association Rule Mining primarily tries to find the rules that govern how or why such products/items are often bought together.

1. Market Basket Analysis: This is the most typical example of association mining. Data is collected using barcode scanners in most supermarkets. A single record lists all the items bought by a customer in one sale. Knowing which groups are inclined towards which set of items gives these shops the freedom to adjust the store layout and the store catalogue to place the optimally concerning one another.

2. Medical Diagnosis: Using relational association rule mining, we can identify the probability of the occurrence of an illness concerning various factors and symptoms.

3. Census Data: This data can be used to plan efficient public services (education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products).
~~~~~~

4. Protein Sequence: Proteins are sequences made up of twenty types of amino acids. This dependency of the protein functioning on its amino acid sequence has been a subject of great research.

~~~~~~

31. Explain in detail exploratory data analysis

- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

The objectives of EDA are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments.

~~~~~~

32. Compare and contrast between descriptive and comparative statistics

~~~~~~

33. Illustrate hypothesis generation and validation with their importance

Hypothesis: Simply put, a hypothesis is a possible view or assertion of an analyst about the problem he or she is working upon. It may be true or may not be true.

For example, if you are asked to build a credit risk model to identify which customers are likely to lapse are which are not, these can be a possible set of hypothesis:

- Customers with poor credit history in past are more likely to default in future.
- Customers with high (loan_value / income) are likely to default more than those with low ratio.
- Customers doing impulsive shopping are more likely to be at a higher credit risk.

Hypothesis Generation: In a nutshell, hypothesis generation is what helps you come up with new ideas for what you need to change.

Imagine you were building a product to help people buy shoes online. Hypothesis generation might include things like:

- Talking to people who buy shoes online to explore what their problems are.
- Talking to people who don't buy shoes online to understand why.
~~~~~~

- Watching people attempt to buy shoes both online and offline in order to understand what their problems really are rather than what they tell you they are.
- Watching people use your product to figure out if you've done anything particularly confusing that is keeping them from buying shoes from you.

The goal is to gain an understanding of your users or your product to help you think up clever ideas for what to build next. Good hypothesis generation almost always involves qualitative research.

Hypothesis generation importance:

Why is it important?

Approach 1: Non-hypothesis driven data analysis (i.e. Boiling the ocean)

In this case, having no idea of what variables to work on would lead to exploring all the variables. This is quite time taking and is known as boiling the ocean.

Approach2: Hypothesis driven analysis

Listing down a comprehensive set of analysis, we identify the variables that are to be explored and neglect the unnecessary ones. This is an efficient approach in terms of both time and resources.

- Hypothesis generation helps in comprehending the business problem.
- Gives a better idea of what are the major factors that are responsible to solve the problem
- Improves your domain knowledge
- Helps to approach the problem in a structured manner

Hypothesis Validation:

Hypothesis validation is different. In this case, you already have an idea of what is wrong, and you have an idea of how you might possibly fix it. You now have to go out and do some research to figure out if your assumptions and decisions were correct.

For our fictional shoe-buying product, hypothesis validation might look something like:

- Standard usability testing on a proposed new purchase flow to see if it goes more smoothly than the old one
- Showing mockups to people in a particular persona group to see if a proposed new feature appeals to that specific group of people
- A/B testing of changes to see if a new feature improves purchase conversion

Importance of Hypothesis Validation:

Hypothesis validation is important because it is how we decide if something really happened, or if certain treatments have positive effects, or if groups differ from each other or if one variable predicts another. In short, if we want to proof that the data is statistically significant, hypothesis validation proves it.

~~~~~~

34. Write a short notes on geo located data visualization

What Is Geo location?

Geo location data is information that can be used to identify an electronic device's physical location. Geo location can be used to determine the time zone and exact positioning coordinates, such as for tracking wildlife or cargo shipments.

Methodology for developing successful data visualizations using geographic location(s):

- **STEP 1: A CLEAR GOAL**
  The analyst should always use caution before beginning a project to ensure a reliable dataset.
- **STEP 2: CLEANING**
  Make sure that the data is free from inconsistency, duplication and missing values.
- **STEP 3: DISCOVERY**
  Import the cleaned data into applicable geographical data visualization software, and begin analysis.
- **STEP 4: VISUALIZATION**
  Various Thematic charts/maps are used to present the statistical data connected with geographical locations.
  - **Choropleth maps** are popular thematic **maps** used to represent statistical data through various shading patterns or symbols on predetermined geographic areas.
  - **A dot distribution map**, or dot density map, is a type of thematic map that uses a point symbol to visualize the geographic distribution of a large number of related phenomena.
  - The main idea behind **proportional symbol** maps is that a larger symbol means "more" of something at a location.
  - **Heat Map:** Heat maps are useful when you have to represent large sets of continuous data on a map using a color spectrum (usually red-to-blue or red-to-green). A heat map is different from a chloropleth map in that the colors in a heat map do not correspond to geographical boundaries.
  - **Hexagonal Binning:** Hexagonal binning is a data visualization technique where you can create a grid in your map with regular hexagons. Once the grid is created, the map can be colored or shaded like a typical choropleth map.
  - **Cluster Map:** Cluster maps help represent dense pockets of data points using a single point. Each cluster is either relatively sized to or labeled with the number of points that have been grouped together.
  - **Bubble Map:** Bubble maps help represent two variables — one by varying the size of the bubble and one by varying the color —simultaneously in a single visualization.
  - **Cartogram Map:** In a cartogram, the mapping variable is shown in a diagrammatic form. The mapping variable often substitutes the land area or distance in the map due to which the map gets distorted in proportion to the mapping variable.

~~~~~~

35. Explain the goals in time series analysis?

Time series:

A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. Examples of time series are solar activity, ocean tides, stock market behavior, and the spread of disease.

What is time series analysis?

Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and other characteristics of time series data. It focuses on comparing values of a single time series or multiple dependent time series at different points in time.

Where do we use Time Series Analysis?

We use Time Series Analysis and Forecasting for many applications where pertinent time series data can be collected, such as:

- Budget Analysis
- Financial Market Analysis
- Census Analysis
- Inventory Management
- Economic Forecasting
- Marketing and Sales Forecasting
- Yield Projections
- Seismological Predictions
- Workload Projections
- Military Planning

What are the Goals of Time Series Analysis?

There are two main goals of time series analysis.

- First, we identify the nature of the phenomenon represented by the sequence of observations in the data.
- Second, we use the data to forecast or predict future values of the time series variable.

What Techniques are used in Time Series Analysis?

One defining characteristic of time series is that this is a list of observations where the ordering matters. Ordering is very important because there is a dependency and changing the order could change the meaning of the data.

There are a number of different methods for modeling time series data including the following:

- Box-Jenkins ARIMA models
- Box-Jenkins Multivariate Models
- Holt-Winters Exponential Smoothing (single, double, triple)
- Unobserved Components Model

36. What is time series data? Mention the applications of time series data

Definition of Time Series: An ordered sequence of values of a variable at equally spaced time intervals.

Applications: The usage of time series models is twofold:

- Obtain an understanding of the underlying forces and structure that produced the observed data
- Fit a model and proceed to forecasting, monitoring or even feedback and feed forward control.

(If extra info required, refer que 35 for que 36.)

~~~~~~

37. How multivariate analysis is useful in designing information visualization

Multivariate Analysis:

Multivariate analysis takes place when you have a data set with 4 or more dependent variables which are to be examined against an independent variable or variables.

There are some common techniques employed to render multivariate analysis through information visualization include:

- Geometric Representations
- Icon Representations
- Pixel-Oriented Representations

Visual Representations of Multivariate Analysis:

Geometric Representations

A geometric representation requires that the information visualization be rendered in such a way that the data is mapped to a geometric space.

- **Parallel Coordinates:** Parallel coordinates uses the idea that each attribute corresponds to an axis and that the axes will be arranged in parallel and with equal spacing between them. Each record is represented as a chain which connects each of its attributes across the axes of the graph. This approach is best used with smaller data sets as with large data sets – the spacing between chains becomes increasingly difficult to determine.
- **Scatter Plot Matrix:** A scatterplot matrix is an attempt to extend the traditional 2D scatterplot to additional dimensions. It works by representing pairs of variables in traditional scatterplots in a matrix with all possible scatterplots created from pairs of variables in the dataset.
- **TableLens:** It is similar to tabulating data in a spreadsheet but instead of using numeric values – each value is represented with a horizontal bar (proportionate to the numeric value). Bars are colored based on the attribute they represent. They can then be manipulated, like a spreadsheet, to change column orders, hide/show columns, sort data by the values of any given column or columns, etc.

Icon Representations:

These use the concept of an icon or a glyph to represent each independent variable with a number of attributes that can vary to represent the dependent variables.
~~~~~~

- **Star Plots:** A star plot takes a single instance of a data set and then transcribes all the dependent variables on a series of axes radiating from a central origin. The points on each axis are then joined to form a polygon.
- **Chernoff Faces:** Chernoff faces involve mapping data points to attributes of the face such as eyes, eyebrows, mouths, noses, ears, etc.

Pixel-Oriented Representations:

This form of representation uses the idea that the pixel is the smallest unit available on a screen and thus represents each data unit in pixels.

~~~~~~

38. Briefly discuss the components of time series.

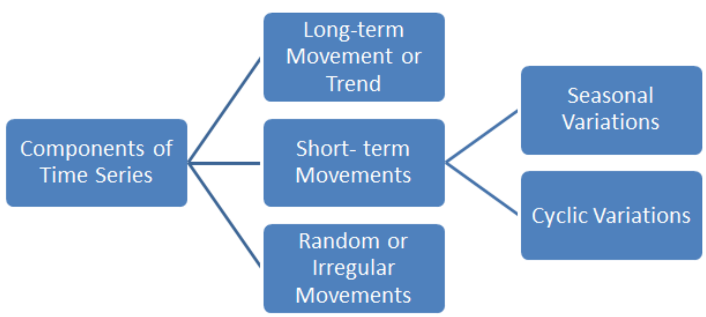
Time series: A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.

Components of time series:

The four categories of the components of time series are

- Trend
- Seasonal Variations
- Cyclic Variations
- Random or Irregular movements

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.



Trend

The trend shows the general tendency of the data to increase or decrease during a long period of time. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

Periodic Fluctuations

There are some components in a time series which tend to repeat themselves over a certain period of time. They act in a regular spasmodic manner.

Seasonal Variations

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions.

Cyclic Variations
~~~~~~

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.
It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Random or Irregular Movements

There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

~~~~~~

39. Define data visualization. Describe the power of data visualization

Data visualization: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Data visualization is one of the steps of the data science process, which states that after data has been collected, processed and modeled, it must be visualized for conclusions to be made.

Why is data visualization important?

- Data visualization provides a quick and effective way to communicate information in a universal manner using visual information.
- The practice can also help businesses identify which factors affect customer behavior
- Makes data more memorable for stakeholders.
- Understand when and where to place specific products; and predict sales volumes.
- The ability to absorb information quickly, improve insights and make faster decisions
- An increased understanding of the next steps that must be taken to improve the organization.
- An improved ability to maintain the audience's interest with information they can understand
- An easy distribution of information that increases the opportunity to share insights with everyone involved
- Eliminate the need for data scientists since data is more accessible and understandable
- An increased ability to act on findings quickly and, therefore, achieve success with greater speed and less mistakes.

Common data visualization use cases

Common use cases for data visualization include the following:

- **Sales and marketing.** Marketing teams must pay close attention to their sources of web traffic and how their web properties generate revenue. Data visualization makes it easy to see traffic trends over time as a result of marketing efforts.
- **Politics.** A common use of data visualization in politics is a geographic map that displays the party each state or district voted for.
~~~~~~

- **Healthcare.** Healthcare professionals frequently use choropleth maps to visualize important health data. Choropleth maps allow professionals to see how a variable, such as the mortality rate of heart disease, changes across specific territories.
- **Scientists.** Allows scientists and researchers to gain greater insight from their experimental data than ever before.
- **Finance.** Finance professionals must track the performance of their investment decisions when choosing to buy or sell an asset. Candlestick charts are used as trading tools and help finance professionals.
- **Logistics.** Shipping companies can use visualization tools to determine the best global shipping routes.
- **Data scientists and researchers.** The data visualization performed by the data scientists and researchers helps them understand data sets and identify patterns and trends that would have otherwise gone unnoticed.

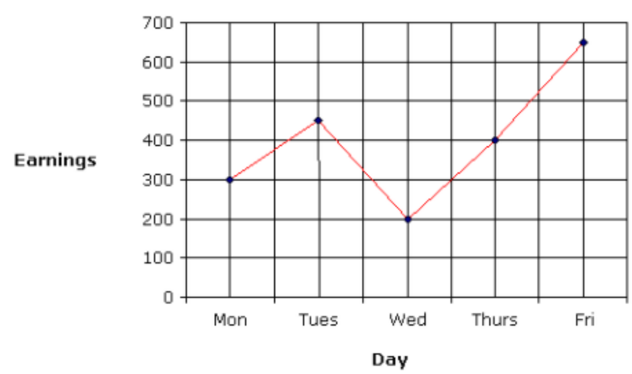
~~~~~~

40. Apply different data visualizations on a sample data along with their benefits.

Line chart:

A line graph is usually used to show the change of information over a period of time. This means that the horizontal axis is usually a time scale, for example minutes, hours, days, months or years.

Example:
The graph shows the trend of daily earnings of a store for five days



Bar chart:

A bar chart is a graph with rectangular bars. The graph usually compares different categories. Although the graphs can be plotted vertically (bars standing up) or horizontally (bars laying flat from left to right), the most usual type of bar graph is vertical.

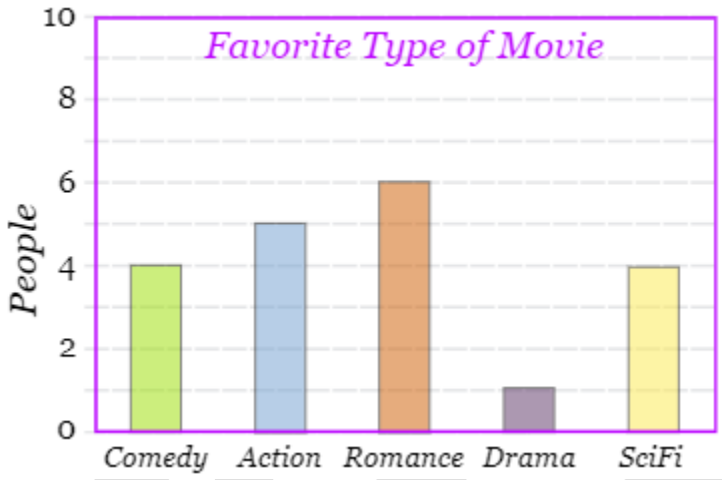
The horizontal (x) axis represents the categories; The vertical (y) axis represents a value for those categories.

Imagine you just did a survey of your friends to find which kind of movie they liked best:
~~~~~~

Table: Favorite Type of Movie

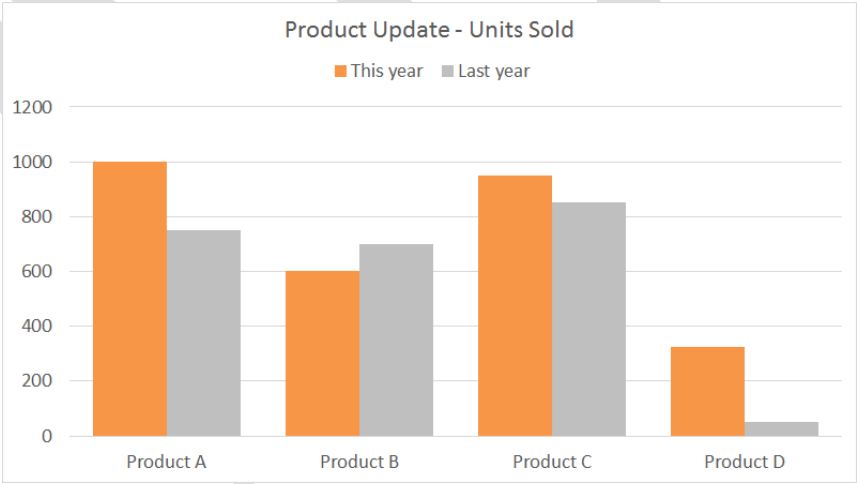
Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4

We can show that on a bar graph like this:



Column charts compare values side-by-side:

Column charts are used to compare values across categories by using vertical bars



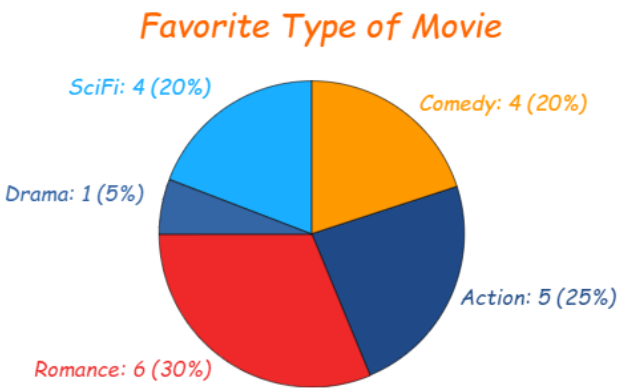
Pie charts clearly show proportions:

A special chart that uses "pie slices" to show relative sizes of data.

Table: Favorite Type of Movie

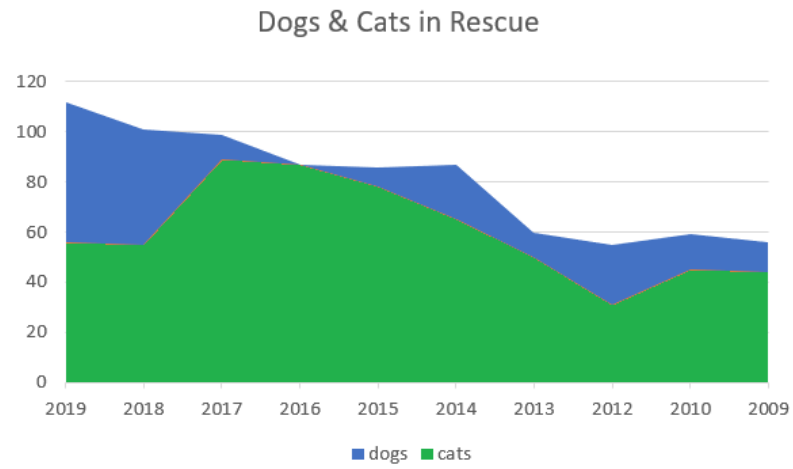
Comedy	Action	Romance	Drama	SciFi

4 5 6 1 4



Area charts *compare* proportions:

An area chart is an extension of a line graph, where the area under the line is filled in. The "lines" are actually a series of points, connected by line segments.

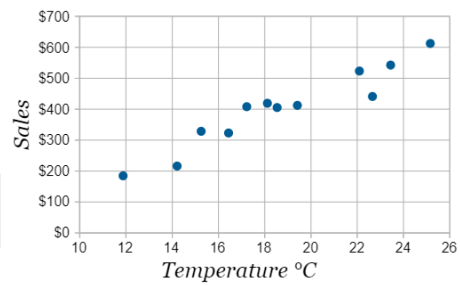


An area chart showing a comparison of cats and dogs in a certain rescue over a period of 10 years.

Scatter charts: distribution and relationships:

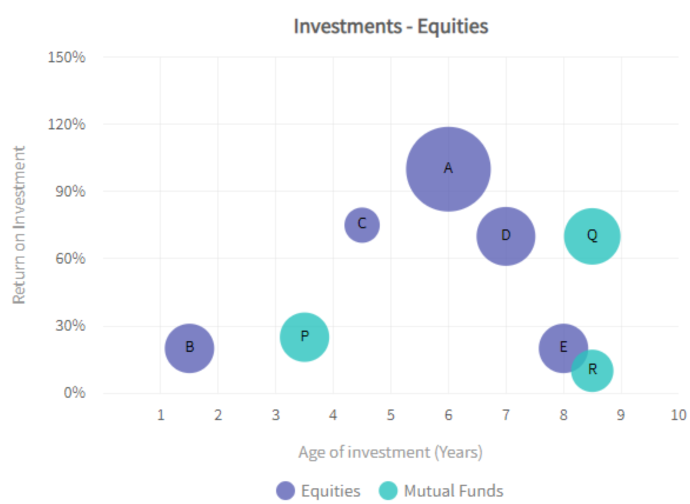
A Scatter (XY) Plot has points that show the relationship between two sets of data.

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Bubble charts: understand multiple variables

Bubble charts/Bubble graphs plot data defined in terms of three distinct numeric parameters. They allow the comparison of entities in terms of their relative positions with respect to each numeric axis and their size as well.



~~~~~~

41. Discuss the dimensions required to represent the data.
~~~~~~

- The number of dimensions or attributes of a data set must be considered carefully as it will determine, to a great extent, the possible information visualizations that can be used to represent the data.
- The more dimensions that are represented in the data – the more confusing it can be to comprehend the information visualization. Thus it's worth noting that the data with large numbers of dimensions may well benefit from using a highly interactive representation rather than a static one.
- Dimensions can be either dependent or independent of each other. It is the dependent dimensions which vary and which we would expect to need to analyze with respect to the independent dimensions.
- There are four types of analysis which can be conducted based on the number of dependent dimensions to be studied:
    - **1. Univariate analysis** – where a single dependent variable is studied against independent variables
    - **2. Bivariate analysis** – where two dependent variables are studied against independent variables
    - **3. Trivariate analysis** – where three dependent variables are studied against independent variables
    - **4. Multivariate analysis** – where more than three dependent variables are studied against independent variables.

~~~~~~

42. List and explain various types of data visualizations.

Data Visualization: Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (points, lines or bars) contained in graphics.

https://www.educba.com/types-of-data-visualization/
~~~~~~