Unit:-1

Introduction to visualization

Introduction:-

Def:- Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics and even animations. These visual display of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Types of Data Visualizations.

Some of the common visualization techniques such as

1. Tables :- This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way

2. pie charts and stacked bar charts:-

These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.

3. Line charts and area charts:-

These visuals show change in one or more quantities by plotting a series of data point over time and are frequently used within predictive analytics.

4. Histograms:- The graph plots a distribution of numbers using a bar chart, representing the quantity of data that falls within a

Particular range.

5. Scatter plots :- These visuals are beneficial in revealing the relationship between two variables and they are commonly used with regression data analysis.

6. Heat Maps :- These graphical representation displays are helpful in visualizing behavioral data by location. this can be a location on a map or even a webpage.

7. Tree Maps :- It display hierarchical data as a set of nested shapes, typically rectangles. Treemaps are great for comparing the proportions between categories via their area size.

→ Visualizing Data - Mapping Data onto Aesthetics

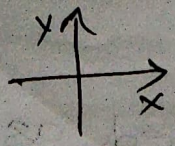
→ Aesthetics and Types of Data

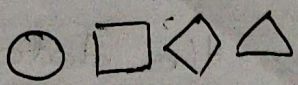
→ Data visualizations map data values into quantifiable features of the resulting graphic. These features are referred as Aesthetics

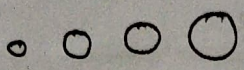
→ Aesthetics describe every aspect of a given graphical element.

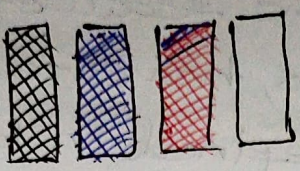
→ All graphical elements have a shape, a size and a color.

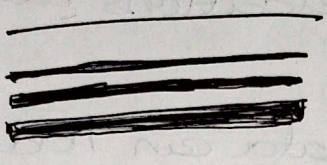
→ commonly used aesthetics in data visualization are position, shape, size, color, line width, line type.

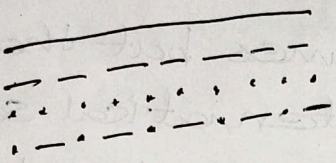
position 

shape 

size 

color 

Line width 

Line type 

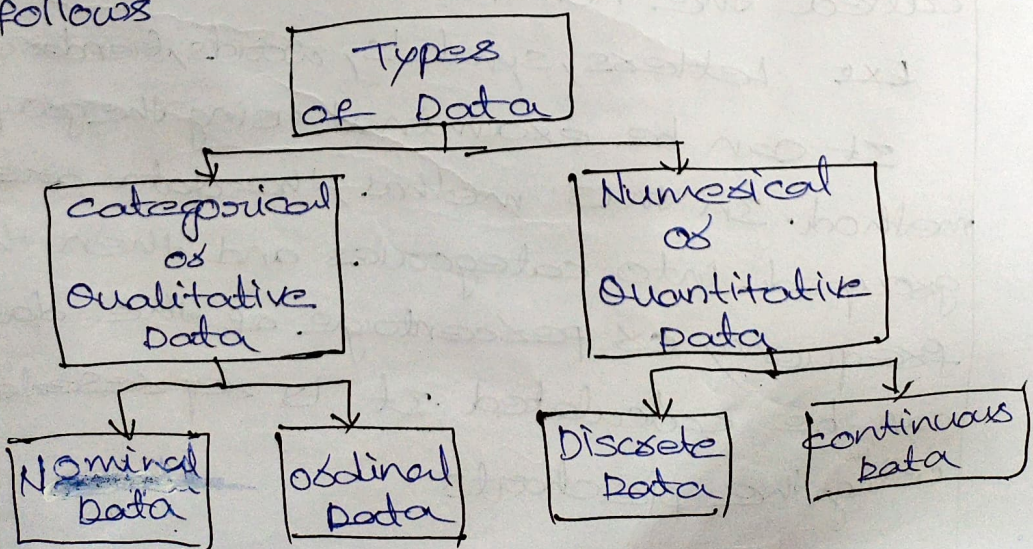
→ Aesthetics will fall into two groups They are continuous and Discrete.

→ position, size, line width and color will represent both continuous and discrete data

→ shape and line type will represent only discrete data.

Types of Data

We can classify the data as follows



1. Qualitative Data

→ It is also called as categorical data. It describes the data that fits into the categories.

→ It describe the features such as a person's gender, home-town such. The measures are defined in terms of natural language specifications but not in terms of numbers.

→ categorical data can hold numerical values sometimes but those values do not have a mathematical sense.

Ex:- Date of Birth, pincode etc.

This qualitative Data is mainly divided into two types

1. Nominal Data
2. Ordinal Data

1. Nominal Data :-

It helps to label the variables without providing the numerical value. It is also called the nominal scale.

Ex:- Letters, symbols, words, Gender etc

It can be examined using the grouping method. In this method, the data are grouped into categories and then the frequency or percentage of the data can be calculated. It is represented using the piecharts

2. Ordinal Data

It is a type of data that follows a natural order. This variable is mostly found in surveys, economics and so on.

The is commonly represented using a bar chart.

2. Quantitative Data

It is also known as numerical Data which represents the numerical value.

Ex:- size, weight, height, length.

The Quantitative Data is divided into two types

1. Discrete Data
2. continuous Data

1. Discrete Data:- It contains only a finite number of possible values. Those values cannot be subdivided meaningfully.

Ex:- No of students in the class.

2. continuous Data:- It is a data that can be calculated. It has a infinite no of possible values that can be selected within a given specific range

Ex:- Temperature range.

→ Types of variables encountered in typical data visualization scenarios are.

Type of Variable	Examples	Appropriate scale	Description
Quantitative/ numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	continuous	Arbitrary numerical values, ie integers, rational or real numbers
Quantitative/ numerical discrete	1, 2, 3, 4	discrete	Numbers in discrete units Ex:- 0.5, 1.0, 1.5
Qualitative/ categorical unordered	dog, cat, fish	Discrete	Categories without order. these variable are called as factors
Qualitative/ categorical ordered	good, fair, poor	Discrete	categories with order. These variable are called as ordered factors
Date or Time	Jan 11 2023, 8:09 AM	continuous or discrete	specific days and or times.
Text	The quick brown fox jumps ob the lazy dog	None or discrete	free-form text. can be treated as categorical if needed.

→ Scales Map Data Values onto Aesthetics

→ To map data values onto Aesthetics we need to specify which data values correspond to which aesthetics value.

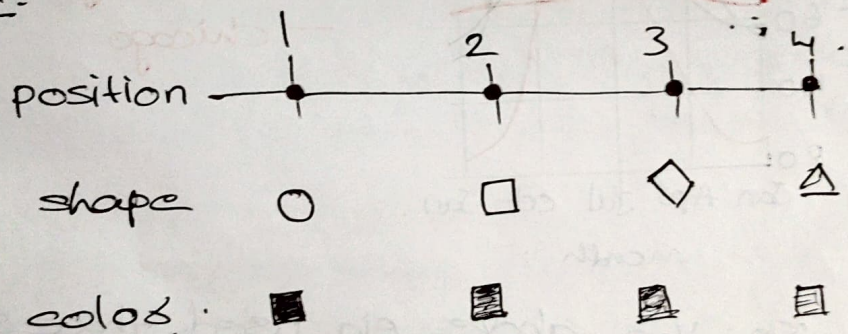
→ The mapping between data values and aesthetics values is created via Scales

→ A scale defines a unique mapping between data and aesthetics

→ A scale must be ~~onto~~ one to one such that for each specific data value there is exactly one aesthetics value and vicevessa.

→ If scale is not one to one then the data visualization becomes ambiguous.

Ex:-

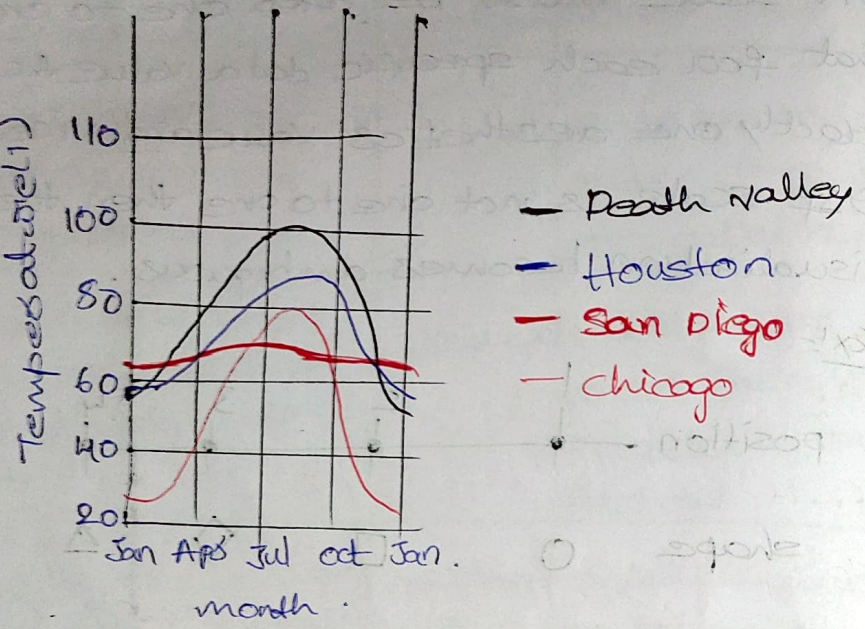


scale links data values to aesthetics. Here the numbers 1 through 4 have been mapped onto a position scale, a shape scale and a color scale. For each scale, each number corresponds to a unique position, shape or color and viceversa.

Let us take the dataset of daily temperature normals for four selected locations in U.S.

We map temperature onto the y axis and day of the year onto the x axis, location onto color, and visualize these aesthetics with solid lines.

The result is a standard line plot showing the temperature normals at the four case locations as they change during the year

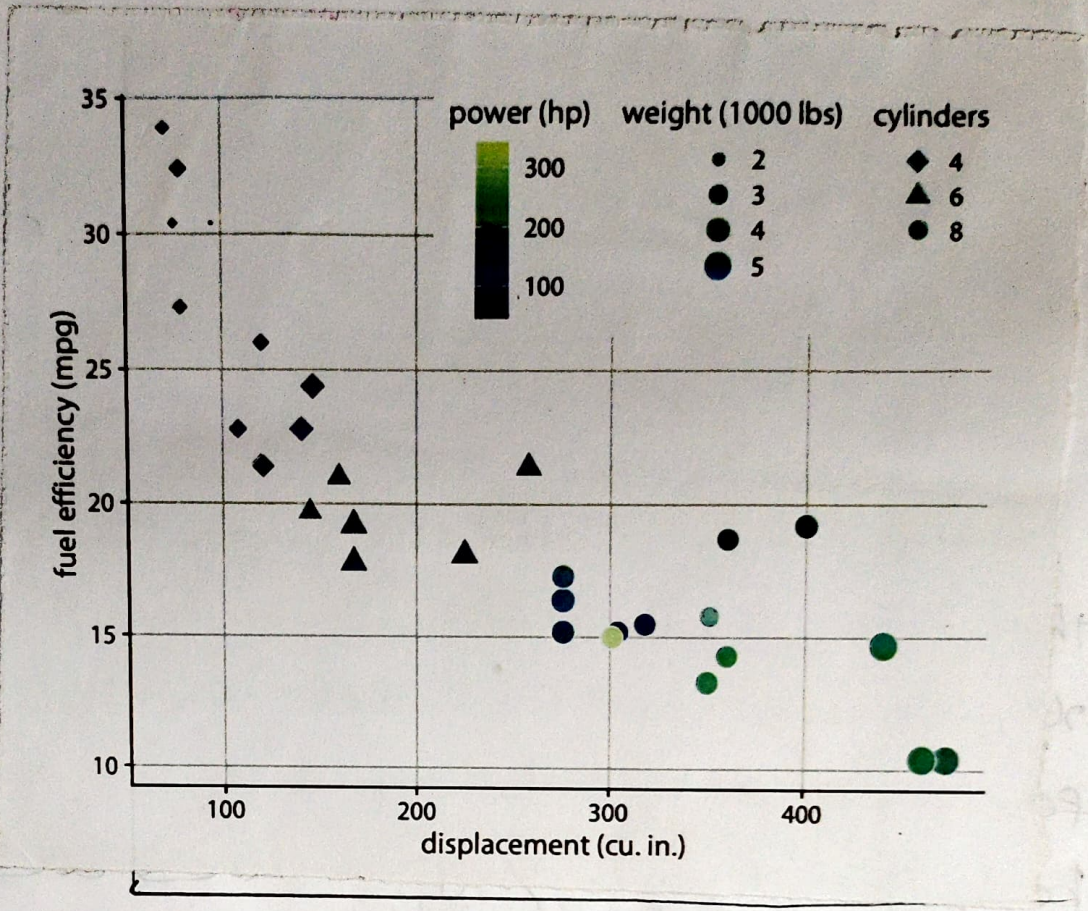


In the above fig used three scales in total, two position scales and one color scale.

This is a typical no of scales for a basic visualization, but we can use more than three scale at once.

Let us consider the following figure which represent fuel efficiency versus displacement. Here we uses five scale, two position scales, one color scale, one size scale and one shape scale and

all the scales represent a different variable from the dataset.



→ **Color scales**

These are three fundamental usecases for color in data visualizations. They are

1. color as a Tool to Distinguish.
2. color to represent Data Values
3. color as a Tool to Highlight.

1. color as a Tool to Distinguish :-

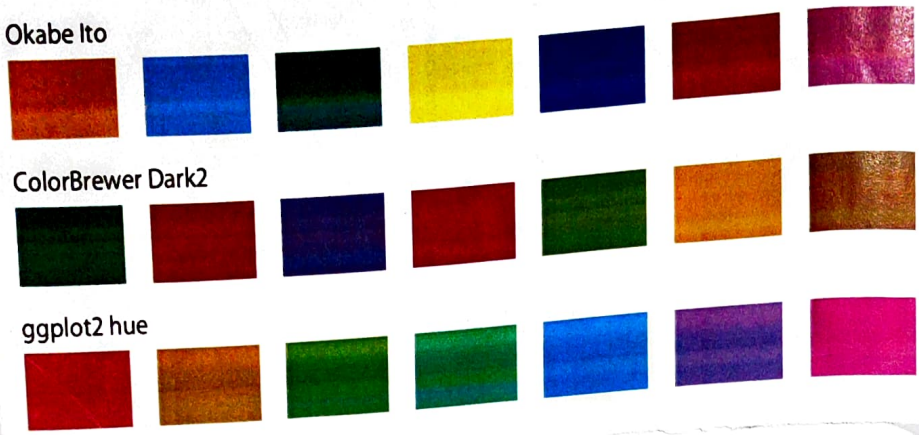
→ We use color as a means to distinguish discrete items or groups that have an intrinsic order, such as different countries on a map or different manufacturers of a certain product.

→ We use a qualitative color scale. such a scale contains a finite set of specific colors that are chosen to look clearly distinct from each other while also being equivalent to each other

→ the second condition requires that no one color should stand out relative to the others.

→ the color should not create the impression of an order, as could be the case with a sequence of colors that get successively lighter. such colors would create an apparent order among the items being colored, which by definition have no order.

The figure below shows three types of qualitative color scales: the scales are Okabe Ito, color Brewer Dark2 scales, ggplot2 hue scales.



2. Colors to Represent Data Values

→ color can also be used to represent quantitative data values, such as income, temperature or speed.

→ we use a sequential color scale. Such a scale contains a sequence of colors that clearly indicates which values are larger or smaller than which other ones and how distant two specific values are from each other.

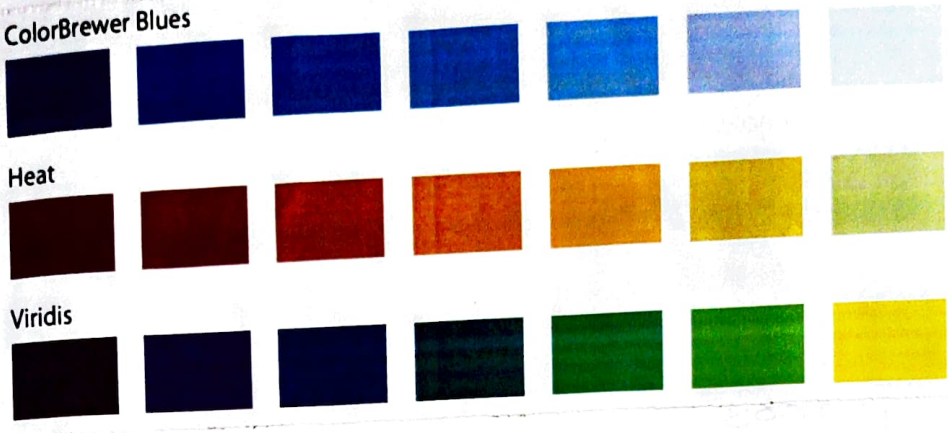
→ the second condition implies that the color scale needs to be perceived to vary uniformly across its entire range.

The different types of sequential color scales.

→ The color Brewer Blues scale is a monochromatic scale that varies from dark to light blue.

→ The Heat and Viridis scales are multihue

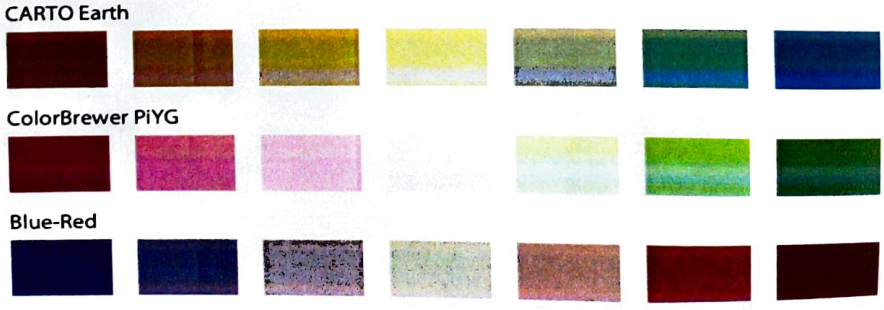
scales that vary from dark red to light yellow and from dark blue via green to light yellow respectively.



→ When we want to show how the data values vary across geographic regions. In this case, we can draw a map of the geographic regions and color them by the data values. Such maps are called choropleths.

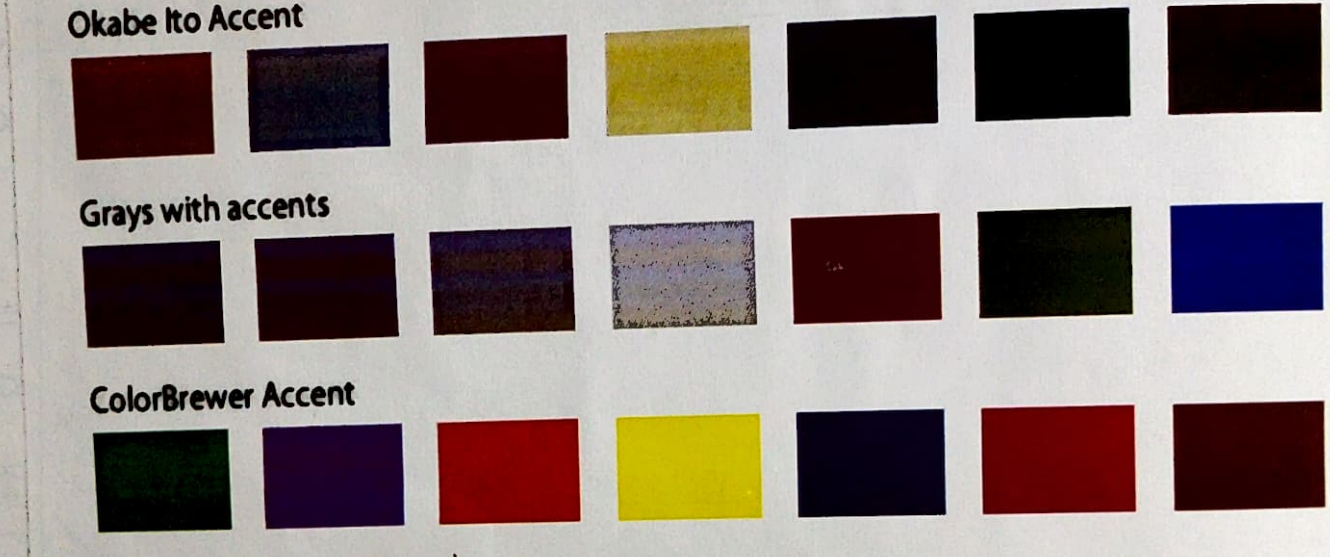
→ When the dataset containing both positive and negative numbers. The appropriate color scale in this situation :- Diverging color scale.

→ The Diverging scales need to be balanced, so that the progression from light colors in the center to dark colors on the outside is approximately the same in either direction.



3- Color as a Tool to Highlight :-

→ colors can also be an effective tool to highlight specific elements in the data.
→ These may be specific categories or values in the dataset that carry key information about the data.
→ This can be achieved with accent color scales, which are color scales that contain both a set of subdued colors and a matching set of stronger, darker and/or more saturated colors.



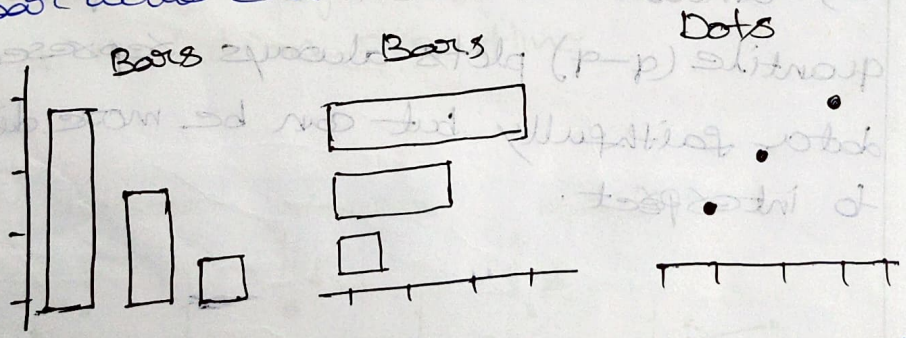
→ Directory of Visualizations:

The directory of viscalizations are mainly categorized into

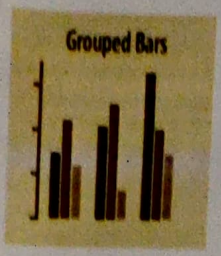
1. Amounts
2. Distributions
3. Proportions
4. X-y relationships
5. Geospatial Data
6. Uncertainty.

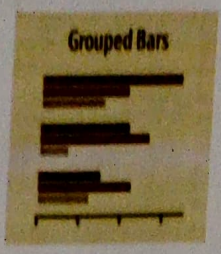
1. Amounts:

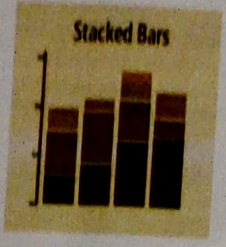
The most common approach to visualizing amounts is using bars, either vertically or horizontally arranged. However, instead of using bars, we can also place dots at the location where the corresponding bar would end.

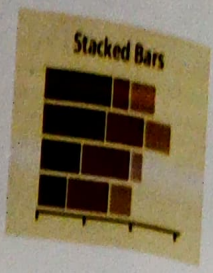


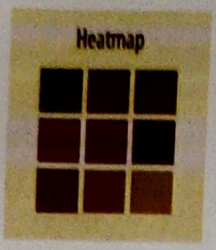
If there are two or more sets of categories for which we want to show amounts, we can group or stack the bars. We can also map the categories onto the x and y axes and show amounts by color via a heatmap.







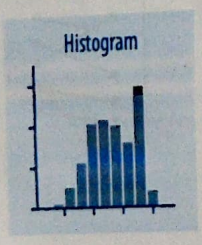


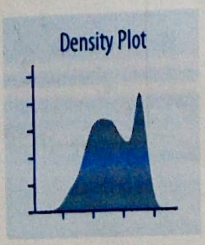


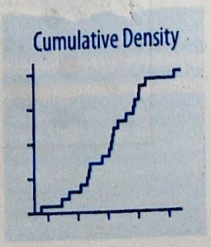
2. Distributions:

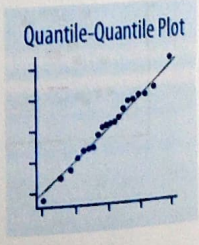
→ Histograms and density plots provide the most intuitive visualizations of a distribution but both require arbitrary parameter choices and can be misleading.

→ cumulative densities and quantile-quantile (q-q) plots always represent the data faithfully but can be more difficult to interpret.

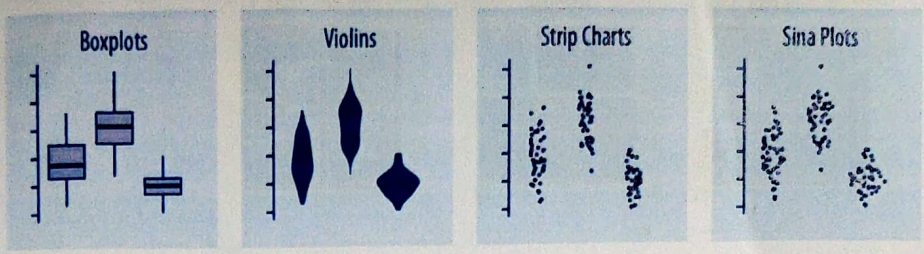






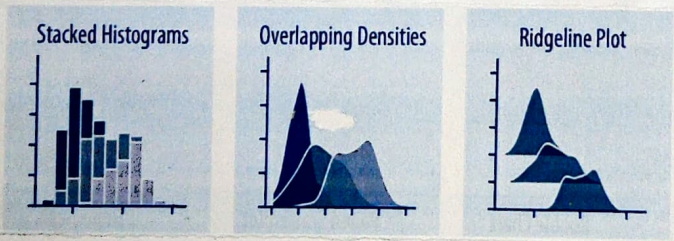


→ Boxplots, violin plots, strip charts and sina plots are useful when we want to visualize many distributions at once, and/or if we are primarily interested in overall shifts among the distributions.



→ Stacked histograms and overlapping densities allow a more in depth comparison of a smaller number of distributions.

→ Ridgeline plots can be a useful alternative to violin plots and are often useful when visualizing very large amount of distribution in distribution over time.



3. Proportions:

→ proportions can be visualized as pie charts, side by side bars or stacked bars.

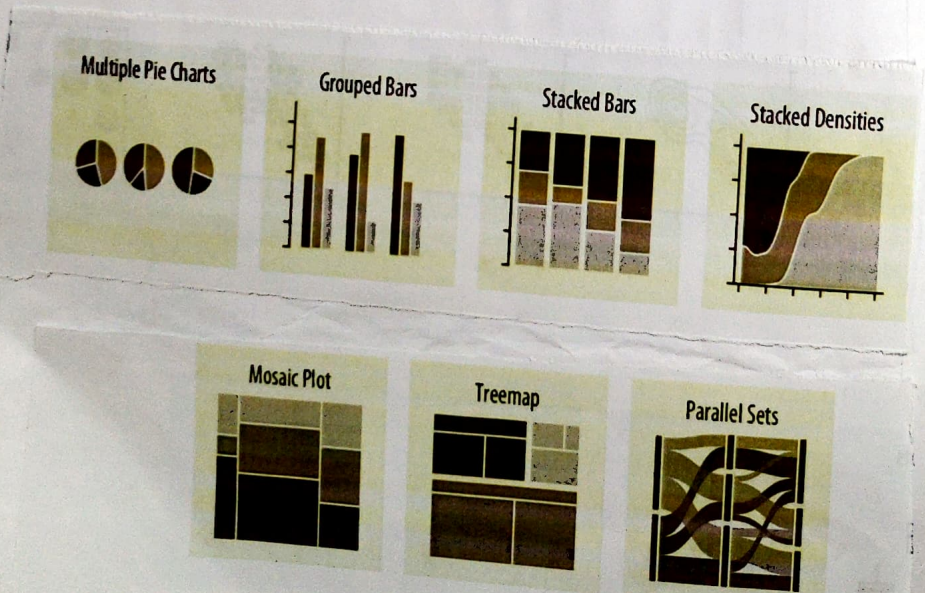
→ when we visualize proportions with bars, the bars can be arranged either vertically or horizontally.

→ pie charts emphasize that the individual parts add up to a whole and highlight simple fractions

→ The individual pieces are more easily compared in side-by-side bars.

→ Stacked bar look can be useful when comparing multiple set of proportions.

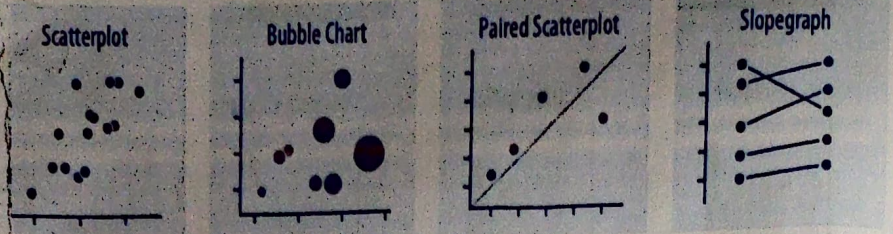
→ Stacked densities are appropriate when the proportions change along a continuous variable.

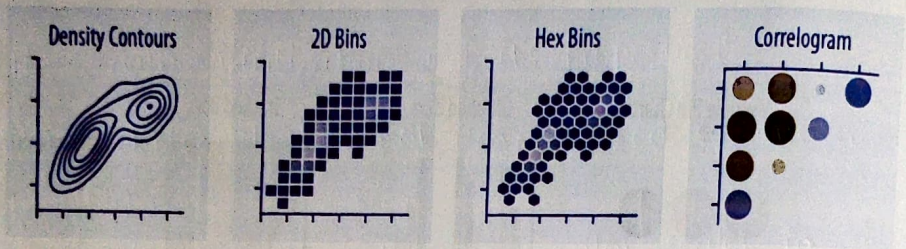


→ Mosaic plots assume that every level of one grouping variable can be combined with every level of another grouping variable.
→ Treemaps work well even if the subdivisions of one group are entirely distinct from the subdivisions of another.
→ parallel set work when there are more than two grouping variables.

4. X-y Relationships :-

→ Scatterplots represent the archetypical visualization when we want to show one quantitative variable relative to another.
→ We have three quantitative variables, we can map one onto the dot size, creating a variant of the scatterplot called a bubble chart.
→ paired scatterplot are used for paired data, where the variables along the x and y axes are measured in the same units.
→ paired data can also be shown as a slopegraph of paired points connected by straight lines.
→ For large numbers of points, regular scatterplots can become uninformative due to overplotting. In this case we use counterlines, 2D bins or hexbins may provide an alternative.





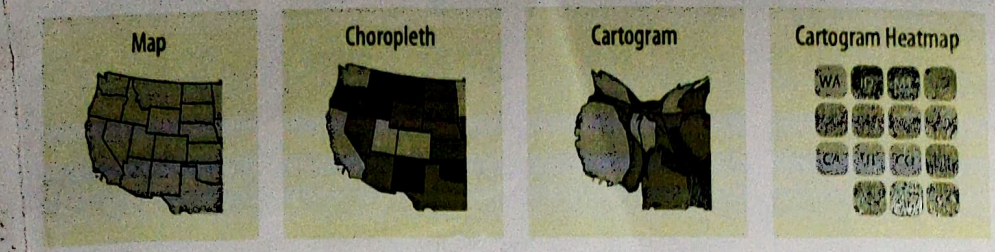
5) Geospatial Data

→ The primary mode of showing geospatial data is in the form a map.

→ A Map makes coordinates on the globe and projects them onto a flat surface such that shapes and distances on the globe are approximately represented by shapes and distances.

→ We can show data values in different regions by coloring those regions in the map, according to the data. such a map is called as choropleth.

→ To distort the different regions according to some other quantity or simplify each region into a square. such visualizations are called cartograms



6) Uncertainty

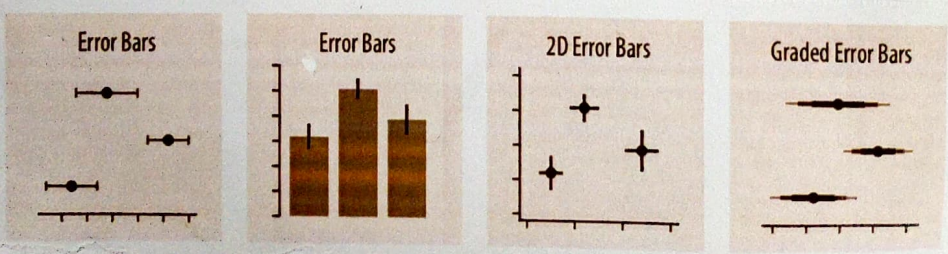
→ Error bars are meant to indicate the range of likely values for some estimate or measurement.

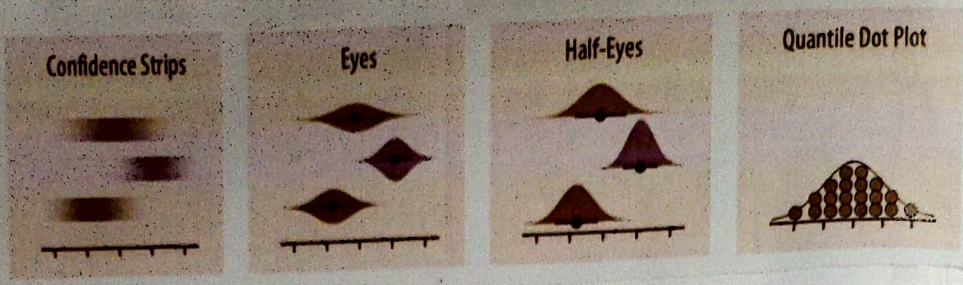
→ Graded error bars show multiple ranges at the same time, where each range corresponds to a different degree of confidence.

→ confidence strips provide a visual sense of uncertainty but are difficult to read accurately.

→ Eyes and half eyes combine error bars with approaches to visualize distributions.

→ A quantile dot plot can serve as a alternative visualization of an uncertainty distribution.





1) Define data visualization and its properties?
2) Explain different types of coordinated systems with an expample?
3) Explain the use of colors and color scales in data visualization?
4) Write different types of visualization methods?