Basic Statistics

Dr. Sudipta Das

Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

Chapter 3: Data Representation

Data Representation

- Presentation of data refers to an exhibition or putting up data in an attractive and useful manner such that it can be easily interpreted.
- Three main forms of data presentation
  - Textual
  - Tabular
  - Graphical

Textual Representation I

Santoor has become the first soap brand from an Indian FMCG company to breach annual sales of Rs 2,000 crore. Wipro Consumer Care, the maker of Santoor, confirmed the number to TOI. With a turnover of over Rs 2,000 crore, Santoor has clearly overtaken HUL's soap brand Lux, and is now challenging the numero uno Lifebuoy. HUL's latest annual report places Lifebuoy and Lux in the Rs 2,000-crore and Rs 1,000-crore plus sales bracket, respectively.
According to industry sources quoting Kantar Household panel data, Santoor's all-India market share in January-March 2019, at 15.1%, has exceeded Lux's 12.5%, but is less than Lifebuoy's 17.7%. The urban market data, however, shows Santoor (13.4%) ahead of both Lux (12%) and Lifebuoy (13%). Kantar declined to comment on this data.
Industry sources quoting Nielsen data said Santoor (9.3%) is the third-largest brand after Lifebuoy (13.7%) and Lux (12%) for January-March 2019. Insights into data from Worldpanel Division of Kantar reveal that Santoor's penetration is much higher than Lux in South and parts of West regions. However, at a national level, Santoor has a much lesser penetration than Lux (34% against 60%).
Lux's penetration is driven by the Rs 10-pack (about 55g), with 60% of Lux-buying homes purchasing this pack. On the other hand, Santoor's penetration is driven largely by its 75g+ pack, with 70% of Santoor-buying homes purchasing this pack.
According to the data from Worldpanel Division of Kantar, Santoor also has a higher number of buying occasions than Lux (Santoor buyers purchase about 45% more times than Lux buyers). As a result, the overall volumes of Santoor have gone ahead of Lux in recent times.

Textual Representation II

- Most raw and vague form of presentation
- Used when the volume of data is small.

Tabular Representation I

- A table facilitates representation of even large amounts of data in an attractive, easy to read and organized manner.
- The data is organized in rows and columns.

Table 1: Soap sale data in Jan-March, 2019.

		Soap		
		Lifebuoy	Lux	Santoor
Company		Numero uno	HUL	Wipro
Sales volume in cr		2000	1000	2000
Market share percentage	Household, Kantar	17.1%	12.5%	15.1%
	Worldpanel, Kantar		60%	34%
	Urban market data	13%	12%	13.4%
	Nielsen data	13.7%	12%	9.3%
Most sold pack			60% 55g	70% 75g

Tabular Representation II

- Components of Data Tables
  - Table Number: should have a specific table number
  - Title: tells the readers about the data it contains, time period of study, place of study, etc.
  - Stubs: titles of the rows in a table
  - Caption: title of a column in a table
  - Body or field: content of a table in its entirety
  - Headnotes and Footnotes: further aids in the purpose of a title
  - Source: used for secondary data

Tabular Representation III

- Construction of Data Tables
  - The title should be in accordance with the objective of study
  - If there might arise a need to compare any two rows or columns then these might be kept close to each other.
  - If the rows in a data table are lengthy, then the stubs can be placed on the right-hand side of the table.
  - Headings should be written in a singular form.
  - A footnote should be given only if needed.
  - Size of columns must be uniform and symmetrical.
  - Headings and sub-headings should be free of abbreviations.
  - There should be a clear specification of units above the columns.

Graphical Representation

- Data is displayed graphically
  - Easier for people to interpret the data
- Common plots are
  - Bar diagram, bar-chart, pie chart, histogram, etc.

Categorical Data Visualization

- Religions of 25 newborn babies:

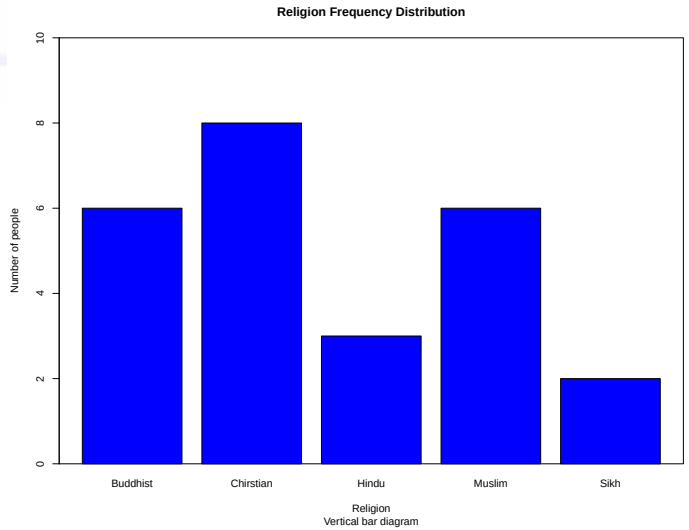
  Hindu, Muslim, Muslim, Buddhist, Christian, Sikh, Muslim, Buddhist, Muslim, Muslim, Buddhist, Buddhist, Christian, Hindu, Christian, Sikh, Christian, Christian, Christian, Buddhist, Buddhist, Christian, Hindu, Christian, Muslim

Religion (Catergory)	*Frequency*	%	*Angle(o)*
Buddhist	6	$6\times(\frac{100}{25})\% = 24\%$	$6\times(\frac{360}{25})^o = 86.4^o$
Christian	8	$8\times4\% = 32\%$	$8\times14.4^o = 115.2^o$
Hindu	3	$3\times4\% = 12\%$	$3\times14.4^o = 43.2^o$
Muslim	6	$6\times4\% = 24\%$	$6\times14.4^o = 86.4^o$
Sikh	2	$2\times4\% = 8\%$	$2\times14.4^o = 28.8^o$
Total	25	$25\times4\% = 100\%$	$25\times14.4^o = 360^o$

Basic Visualization II

- Vertical bar diagram
  - It displays frequencies of categories of data.
  - Categories in X-axis and frequencies in Y-axis
  - Vertical bars of lengths proportional to the frequencies are drawn at each categories.

Vertical Bar Diagram



R-code

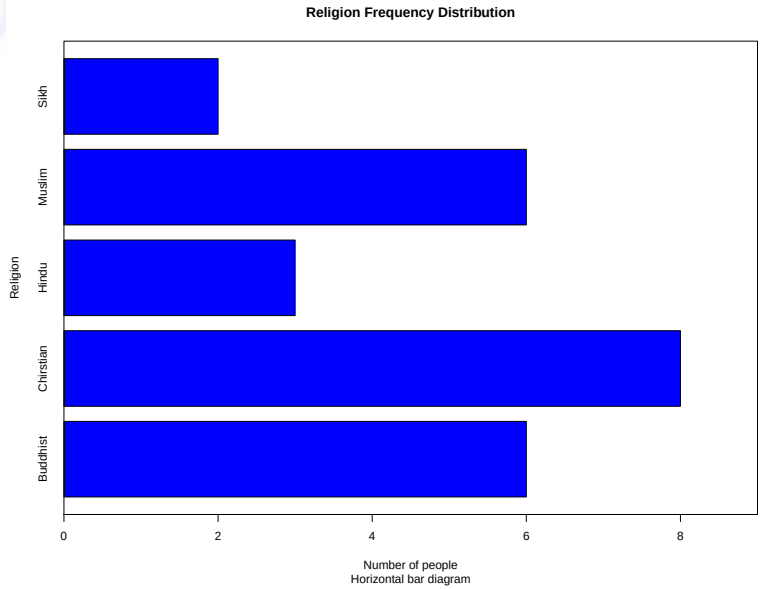
Basic Visualization III

1. Create table from data:- $my\_table = table(vector)$
2. Get diagram:- $barplot(my\_table,\ xlab = \ldots,\ ylab = \ldots,\ ylim = \ldots,\ col = " \ldots ")$
3. Add title:- $title(main = "",\ sub = "")$

Basic Visualization IV

- Horizontal bar diagram
  - It displays frequencies of categories of data.
  - Categories in Y-axis and frequencies in X-axis
  - Horizontal bars of lengths proportional to the frequencies are drawn at each categories.

Horizontal Bar Diagram



Basic Visualization V

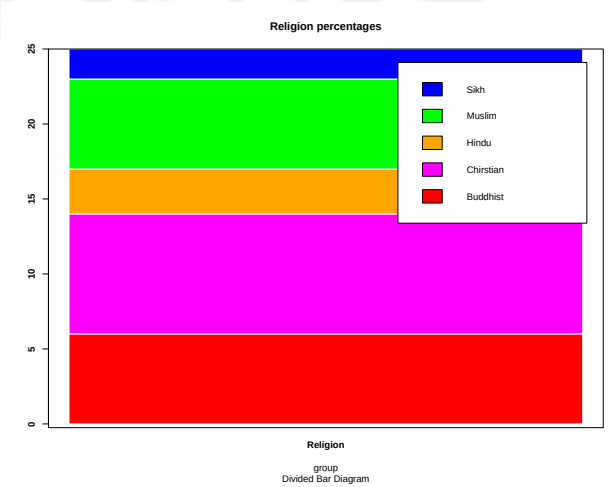
R-code

1. Create table from data:- $my\_table = table(vector)$
2. Get diagram:- $barplot(my\_table,\ xlab = \ldots,\ ylab = \ldots,\ ylim = \ldots,\ col = " \ldots ",\ horiz = T)$

Basic Visualization VI

- Divided bar diagram
  - It shows the relative proportions of data in different categories within a bar.
  - The size of the portion corresponding to the category A can be found as $\left(\frac{\text{\# data points in category A}}{\text{Total number of data points}} \times 100\%\right)$

Divided Bar Diagram



R-code

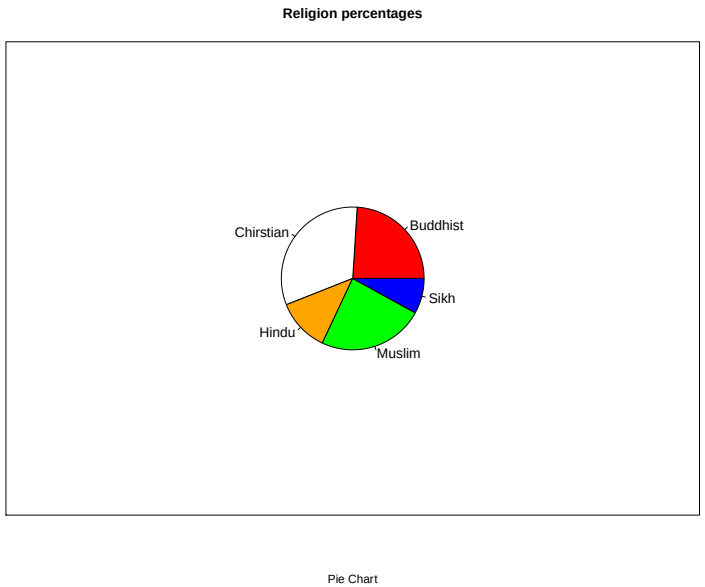
Basic Visualization VII

1. Create table from data:- $data = matrix(table(religion), nrow = 5)$; $colnames(data) = c(Religion)$; $rownames(data) = c(Buddhist, Chirstian, Hindu, Muslim, Sikh)$
2. Get diagram:- $barplot(data, col = c(red, magenta, orange, green, blue), xlab = \ldots, legend.text = c(Buddhist, Chirstian, Hindu, Muslim, Sikh))$

Basic Visualization VIII

- Pie chart
  - It shows the relative proportions of data in different categories within a circle.
  - The size of an angle θ_A corresponding to the category A can be found as $\theta_A = \frac{\text{\# data points in category A}}{\text{Total number of data points}} \times 360^o$

Pie Chart



Basic Visualization IX

R-code

1. Create table from data:- $my\_table = table(vector)$
2. $pie(my\_table, col = \ldots, clockwise = \ldots, init.angle = 0)$
3. By default anti clockwise and intial angle is 0

Discrete Data Visualization

- Number of siblings of 30 new born babies:

$$0, 2, 2, 1, 1, 1, 1, 3, 0, 0, 2, 1, 1, 0, 2, 3, 1, 4, 1, 2, 2, 0, 1, 0, 0, 1, 3, 0, 1, 1$$

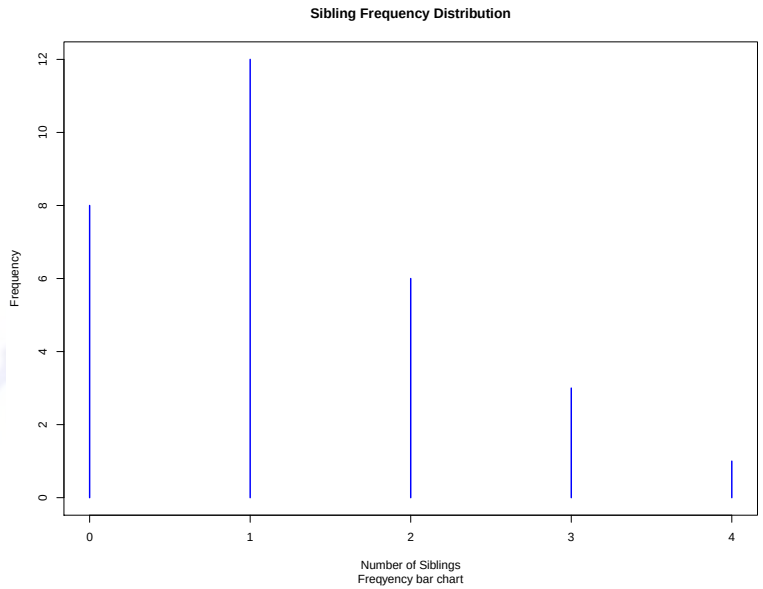
Frequency Table

Number of siblings	*Frequency*	*Cumulative Frequency*
0	8	8
1	12	$8 + 12 = 20$
2	6	$8 + 12 + 6 = 26$
3	3	$8 + 12 + 6 + 3 = 29$
4	1	$8 + 12 + 6 + 3 + 1 = 30$
Total	30	-

Basic Visualization XI

- Freqyency bar chart
  - It displays frequencies at different isolated values of data.
  - Isolated values in X-axis and frequencies in Y-axis
  - Perpendicular lines proportional to frequencies are drawn at isolated values.

Freqyency bar chart



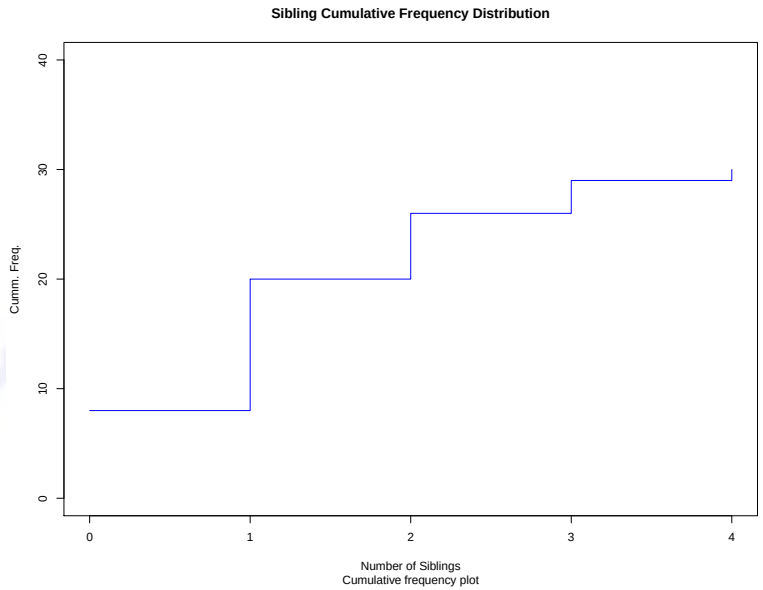
R-code

1. *plot*(*my_table*, *col* = . . . , *xlab* = . . . , *ylab* = . . .)

Basic Visualization XIII

- Cumulative frequency plot/ogive
  - Cumulative frequency on the y-axis and isolated values along the x-axis.

Cumulative frequency plot/ogive



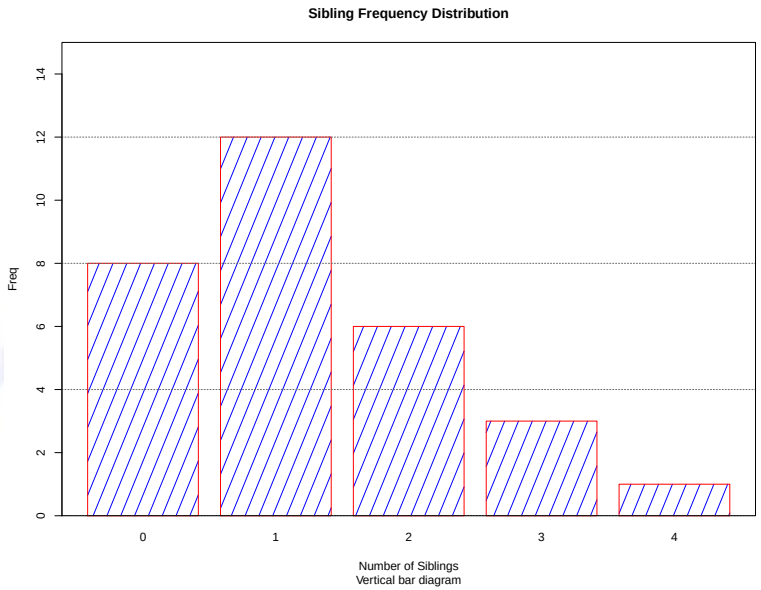
R-code

1. $plot(cumsum(my\_table),\ col = \ldots,\ xlab = \ldots,\ ylab = \ldots)$

Basic Visualization XV

- Vertical bar diagram
  - Isolated values in X-axis and frequencies in Y-axis
  - Vertical bars of lengths proportional to the frequencies are drawn at each isolated value.

Vertical Bar Diagram



R-code

1. *barplot*(*my_table*, *xlab* = . . . , *ylab* = . . . , *ylim* = . . . , *col* = . . . , *border* = . . .)

Basic Visualization XVII

Continuous Data Visualization

- Weights of 60 newborn babies in Kg:
  2.99, 2.74, 3.08, 3.04, 2.79, 2.63, 2.62, 3.40, 2.72, 2.53, 3.19, 2.77, 3.39, 3.67, 2.45, 2.41, 2.90, 3.50, 2.84, 3.55, 3.25, 2.56, 3.52, 3.03, 3.14, 3.07, 3.46, 3.13, 3.02, 3.15, 3.05, 3.20, 2.82, 2.89, 3.26, 3.01, 2.88, 3.01, 2.87, 2.70, 3.24, 3.74, 3.53, 3.34, 2.44, 3.72, 2.95, 3.09, 3.38, 3.16, 2.96, 2.39, 3.06, 2.86, 2.54, 2.94, 2.61, 2.48, 2.55, 2.62
- minimum weight = 2.39 kg
- maximum weight = 3.74 kg
- Range = 3.74 - 2.39 = 1.35

Basic Visualization XVIII

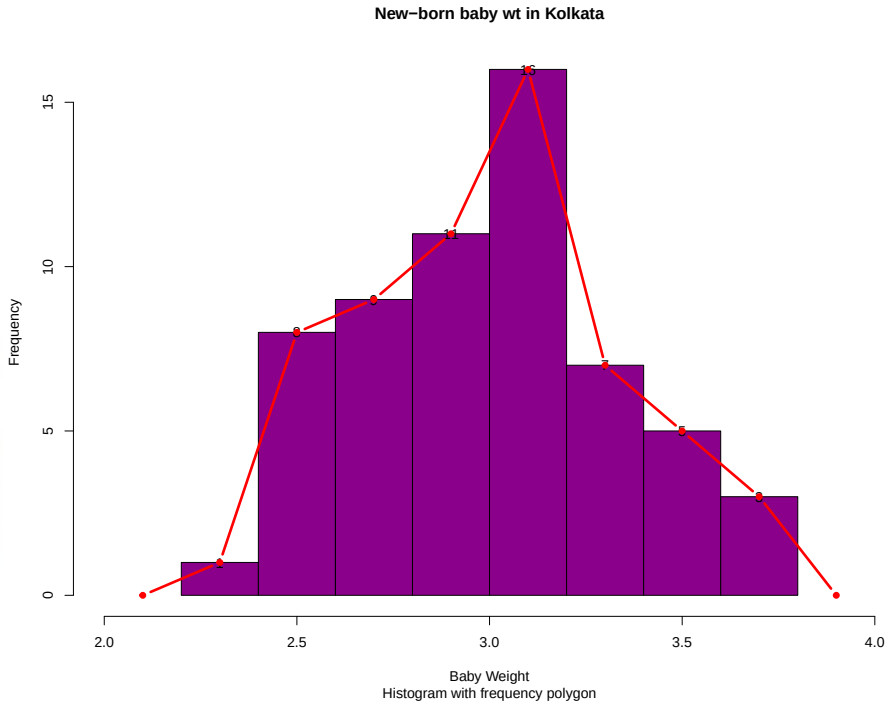
Frequency Table

Class boundary	*Class Mark*	*Freq. $\leq$ type*	*Relative Freq.*	Area	*Cum. Freq.*
2.0-2.4	2.2	1	1/60/(2.4-2.0)	1/60	1
2.4-2.8	2.6	17	17/60/(2.8-2.4)	17/60	18
2.8-3.2	3.0	27	27/60/(3.2-2.8)	27/60	45
3.2-3.6	3.4	12	12/60/(3.6-3.2)	12/60	57
3.6-4.0	3.8	3	3/60/(4.0-3.6)	3/60	60
Total	-	60		1	-

Basic Visualization XIX

- Histogram
  - It displays frequencies of quantitative data that has been sorted into intervals.
  - Split the data into intervals, called bins
    - # of bins $= k \approx= (\max - \min)/h$, where h =bin width
    - $k \approx \sqrt{n}$
  - Vertical bars of length proportional to the frequencies at each bins are drawn
  - Sometimes relative frequency is given instead of absolute frequency

Histogram with Frequency Polygon



- Frequency polygon
  - Another type of frequency distribution graph
  - The number of observations is marked with a single point at the midpoint of an interval

Basic Visualization XXI

- A straight line then connects each set of points.

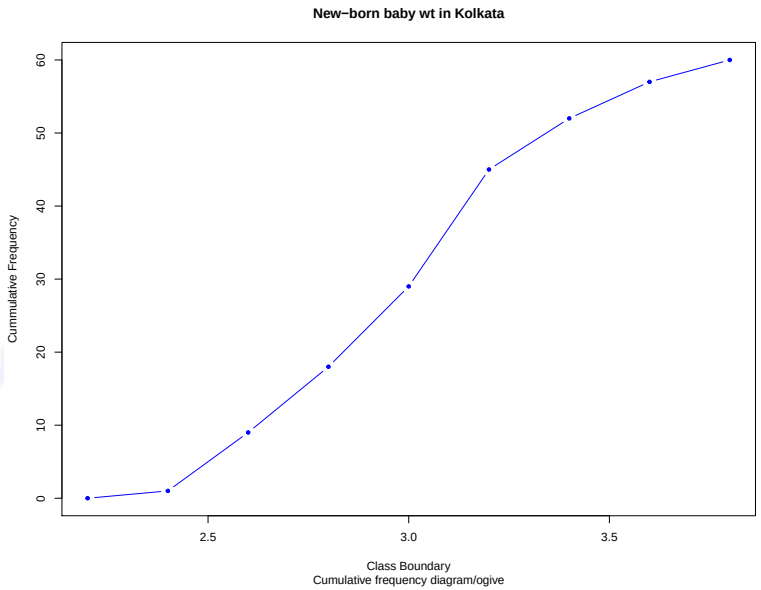
R-code

1. $hist(data,\ breaks = c(...),\ main = ...,\ xlab = ...,\ xlim = c(,),\ col = "",\ border = "",\ freq = FALSE)$
2. $lines(x = mid\_values,\ y = freq,\ type = "",\ pch = "",\ col = "")$

Basic Visualization XXII

- Cumulative frequency diagram/ogive
  - Cumulative frequency on the y-axis and class boundaries along the x-axis.

Cumulative frequency diagram/ogive



R-code

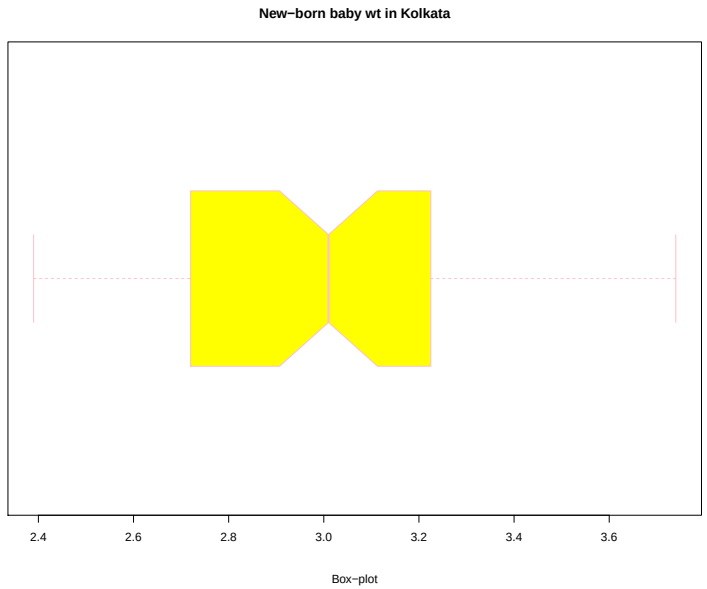
1. $plot(x = upper\ class\ limits,\ y = cumm.\ freq.)$

Basic Visualization XXIV

- Box plot (also known as a box and whiskers plot)
  - It displays the five 5 statistics
    $\max\left(minimum, \lceil Q_1 - 1.5 \times IQR \rceil\right), Q_1, Q_2, Q_3$, and
    $\min\left(maximum, \lfloor Q_3 + 1.5 \times IQR \rfloor\right)$, where
    IQR(Inter Quartile Range) $= Q_3 - Q_1$
    - $\lceil Q_1 - 1.5 \times IQR \rceil$: smallest data point higher than $Q_1 - 1.5 \times IQR$
    - $\lfloor Q_3 + 1.5 \times IQR \rfloor$: largest data point smaller than $Q_3 + 1.5 \times IQR$
  - The box can either be vertically or horizontally displayed depending on the labeling of the axis.

Basic Visualization XXV

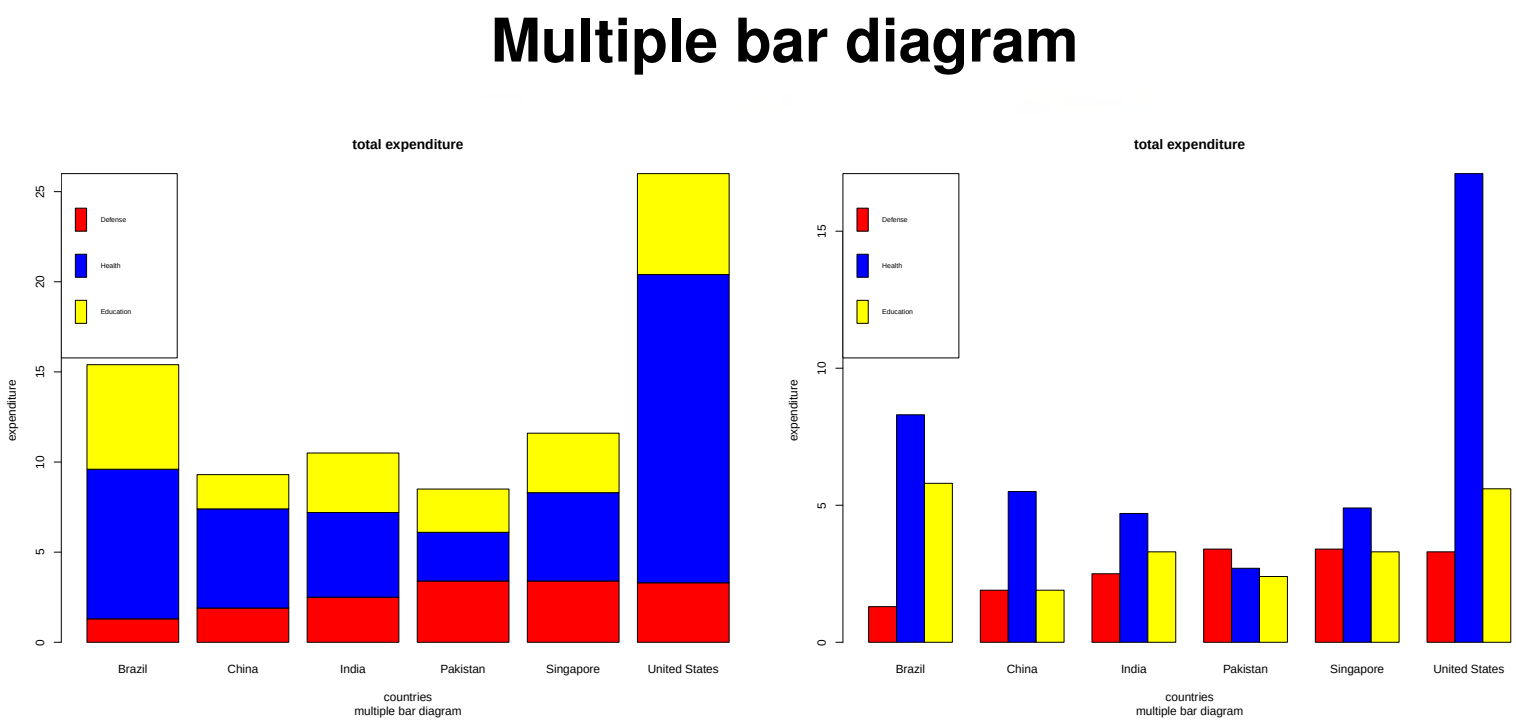
Box plot



R-code

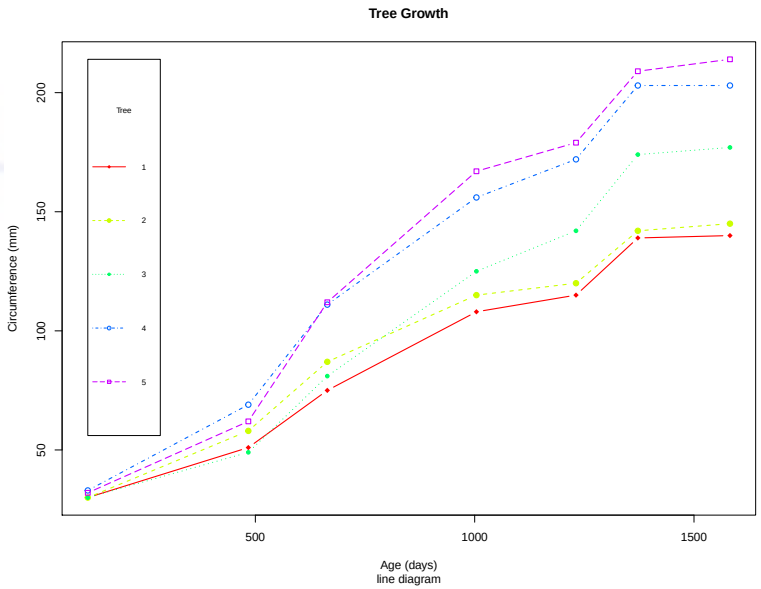
1. *boxplot*(*datat*, *col* = "", *border* = "", *horizontal* =, *notch* =)
2. By default it's vertical and without notch

- Multiple bar diagram
  - To compare different groups through charts

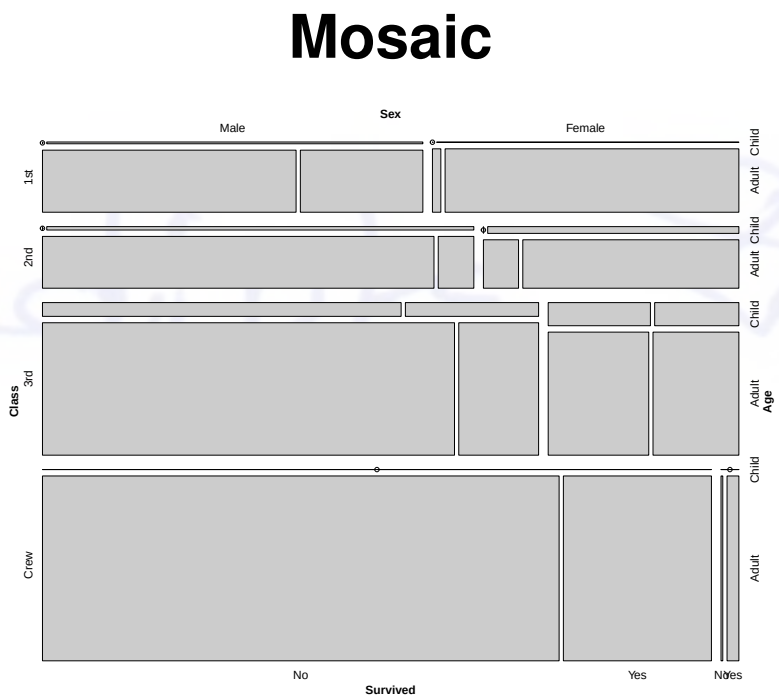


- Line diagram/ Line chart
  - A type of chart displaying information as a series of data points called 'markers' connected by straight line segments

Line diagram

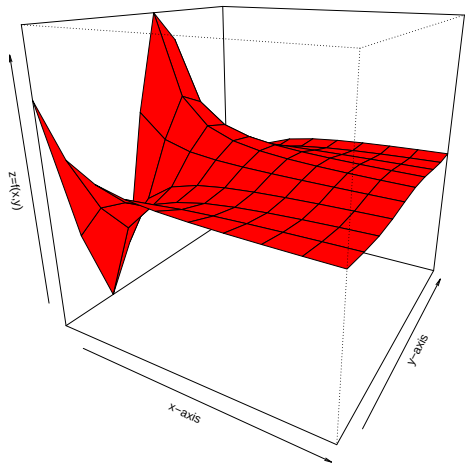


- Mosaic
  - To compare on more than one attribute



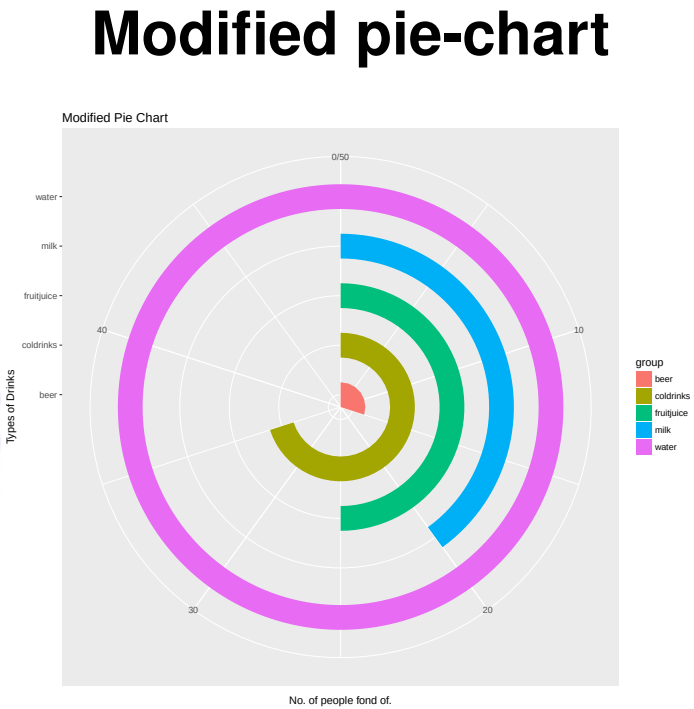
Advanced Visualization IV

- 3-D plot



- 3-D plot with control: Explore *"plotly"* package

- Modified pie-chart



- Exlploring *"ggplot2"* package