Data Visualization and Interpretation : Graphical summaries

Dr.Mamatha.H.R

Professor

Department of Computer Science and Engineering

PES University

Bangalore

Course material created using various Internet resources and text book

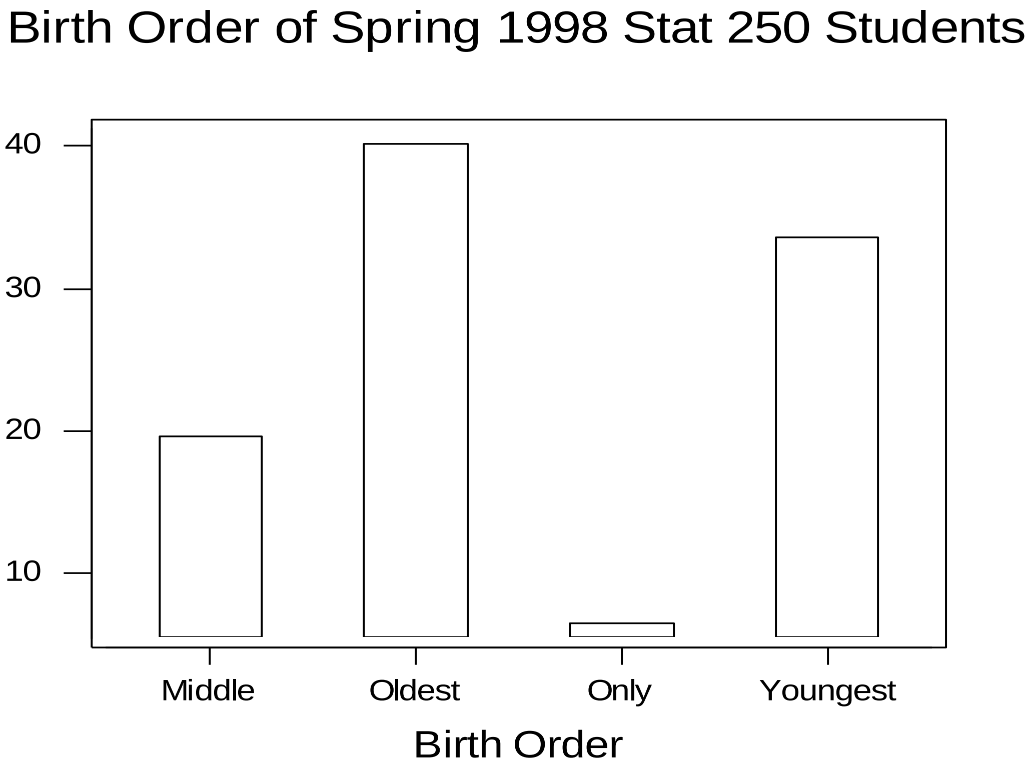
Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Which graph to use?

- Depends on type of data
- Depends on what you want to illustrate
- Depends on available statistical software

Bar Chart



n=92 students

Bar Chart

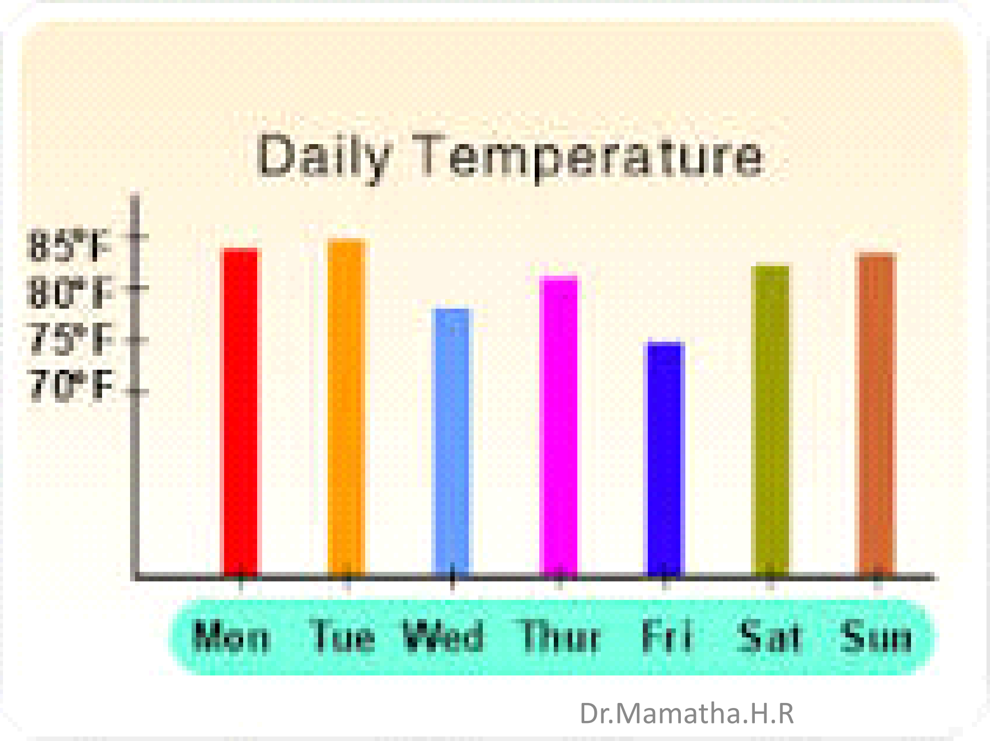
- Summarizes categorical data.
- Horizontal axis represents categories, while vertical axis represents either counts ("**frequencies**") or percentages ("**relative frequencies**").
- Used to illustrate the differences in percentages (or counts) between categories.

Bar Graph

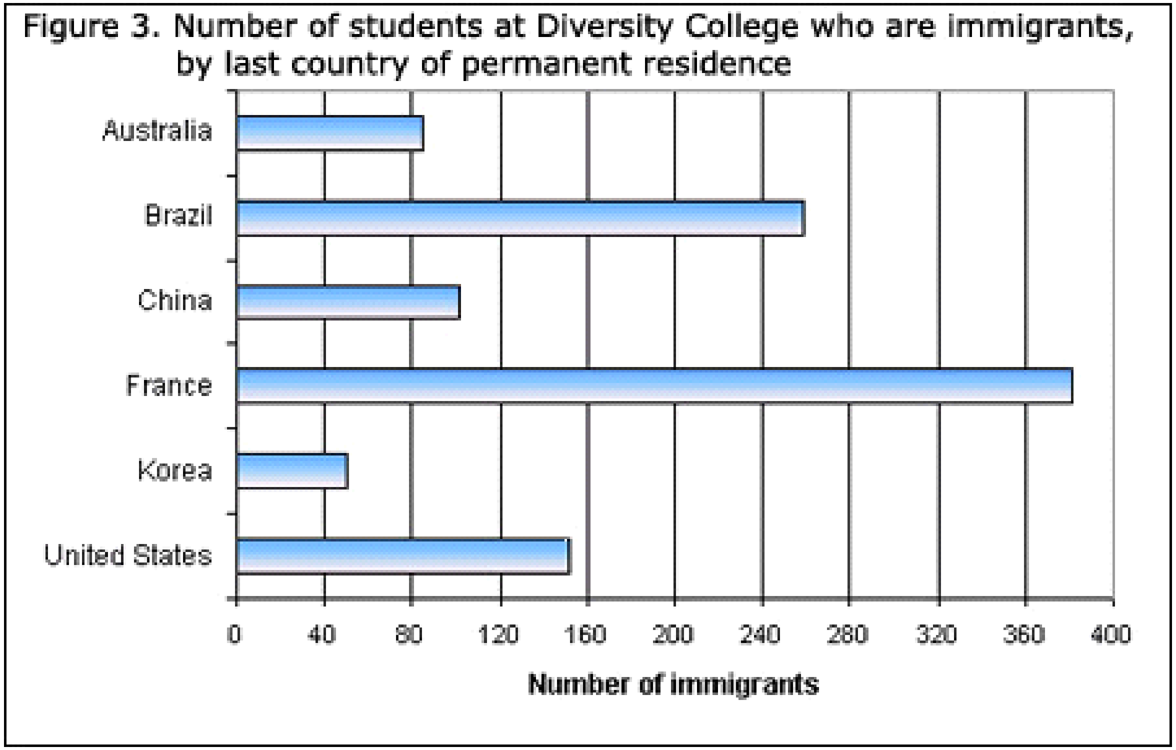
- A bar graph is a pictorial rendition of statistical data in which the independent variable can attain only certain discrete values.
- A bar diagram makes it easy to compare sets of data between different groups at a glance.
- The graph represents categories on one axis and a discrete value in the other.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

Types of Bar Graphs

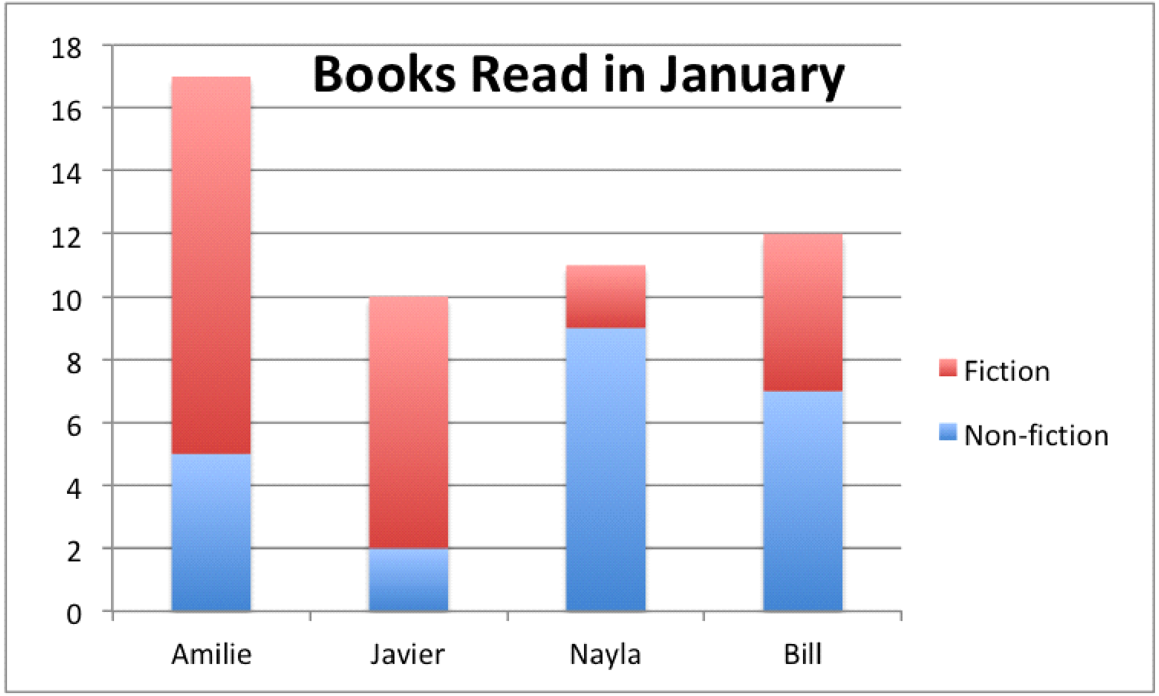
- Vertical BarGraphs : The classes are displayed on the x-axis, and the values(scores) of those classes are displayed on the y-axis. Useful only when comparing one set of data.



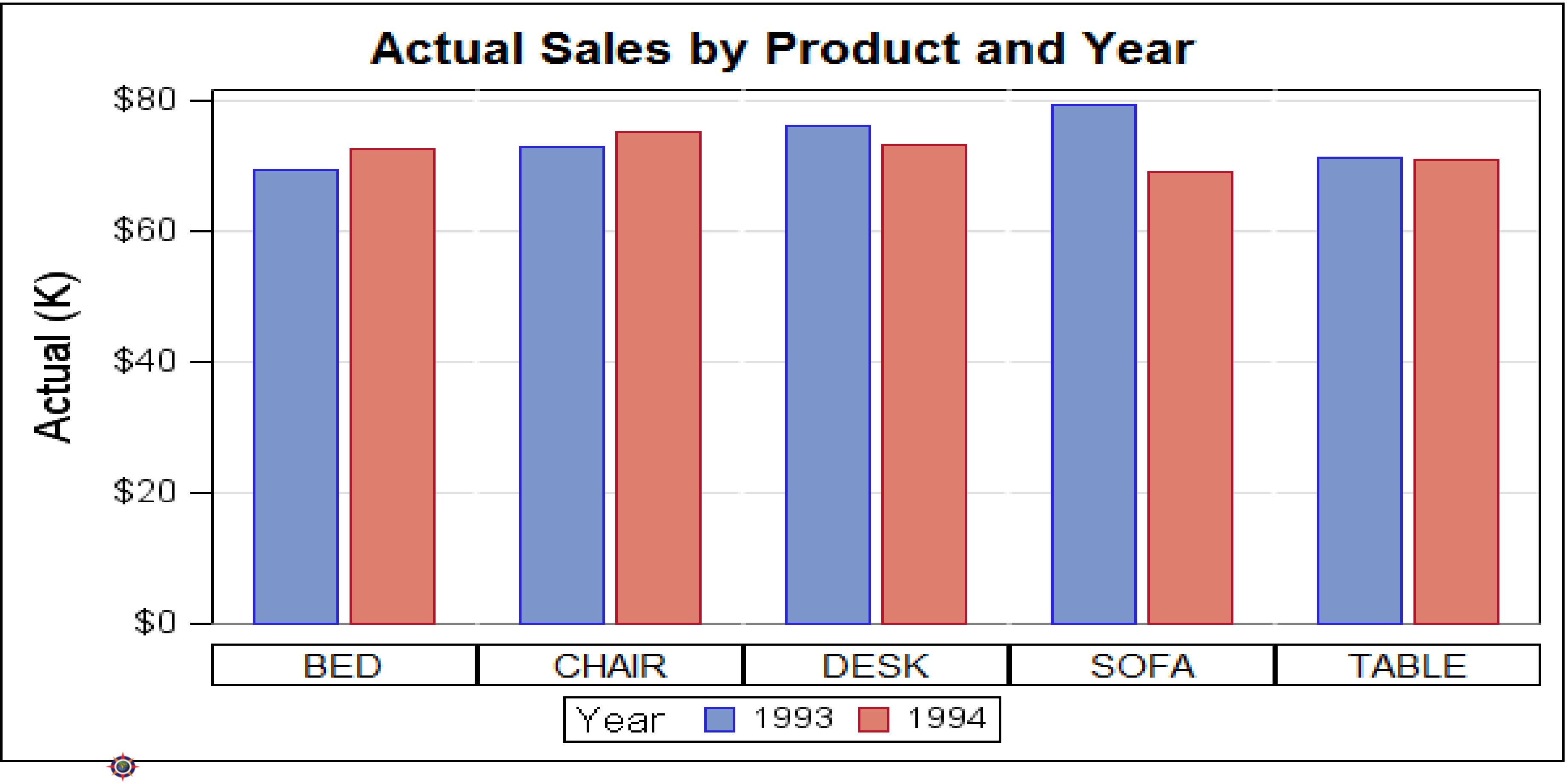
- Horizontal BarGraphs : The classes are displayed on the y-axis, and the values(scores) of those classes are displayed on the x-axis. Useful only when comparing one set of data.

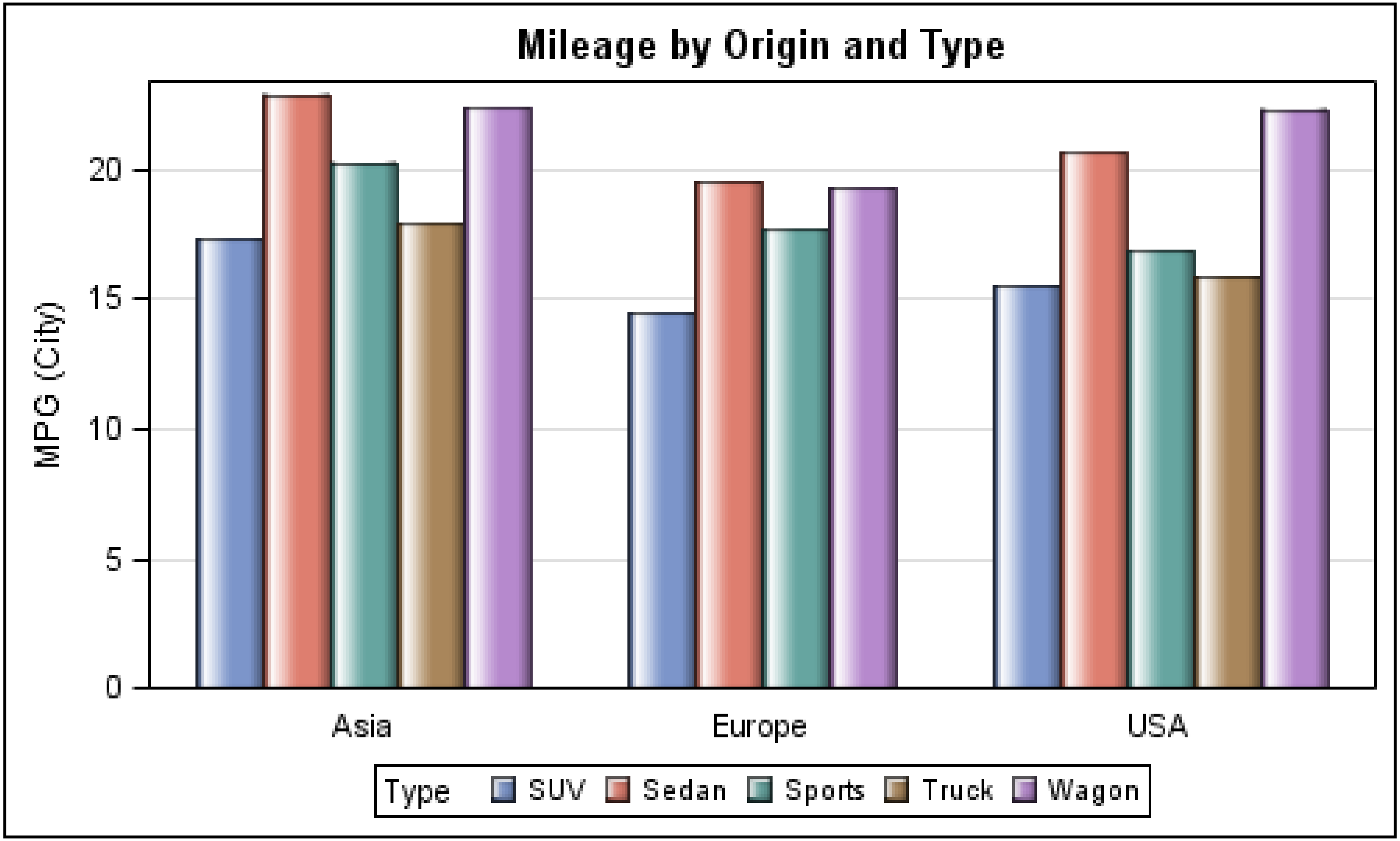


- Stacked BarGraphs : Each bar has multiple datasets to be compared, each set of values belonging to the class of different datasets are stacked over one other. Useful when comparing multiple datasets but having same set of classes



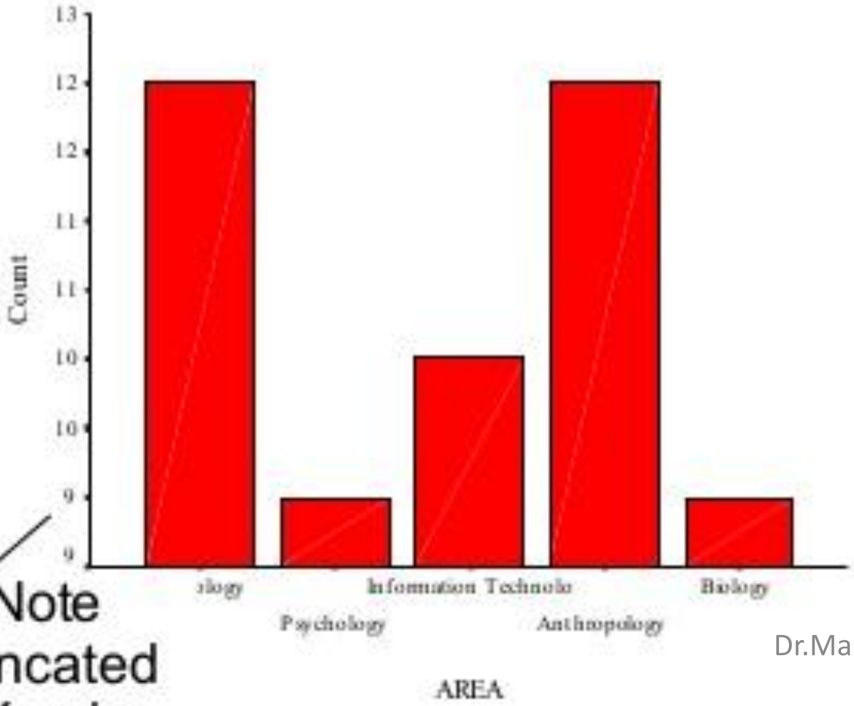
Grouped Bar Graph



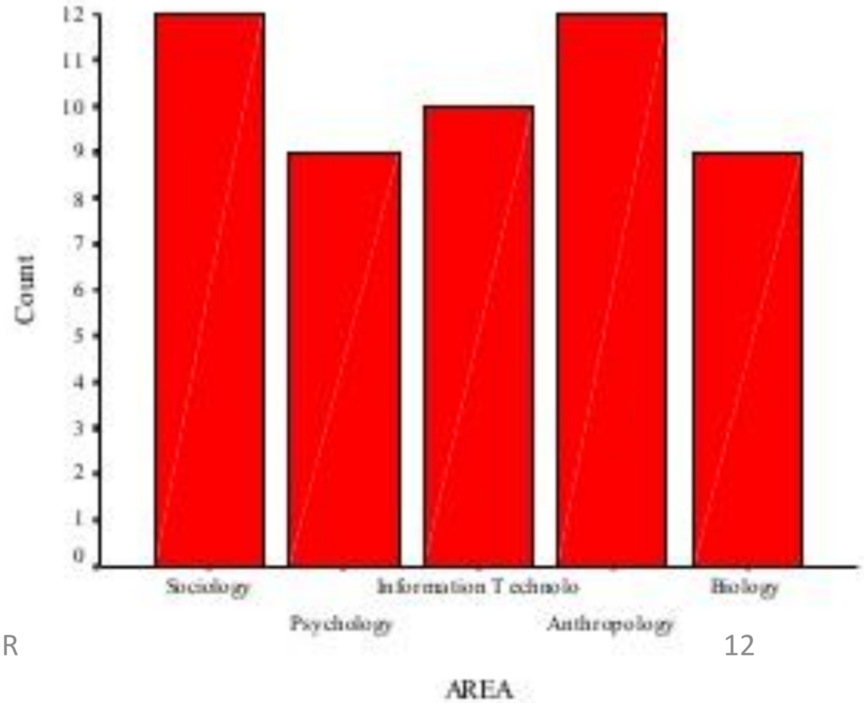

Mileage by Origin and Type
MPG (City)
0
5
10
15
20
Asia
Europe
USA
Type
SUV
Sedan
Sports
Truck
Wagon

Bar chart (Bar graph)

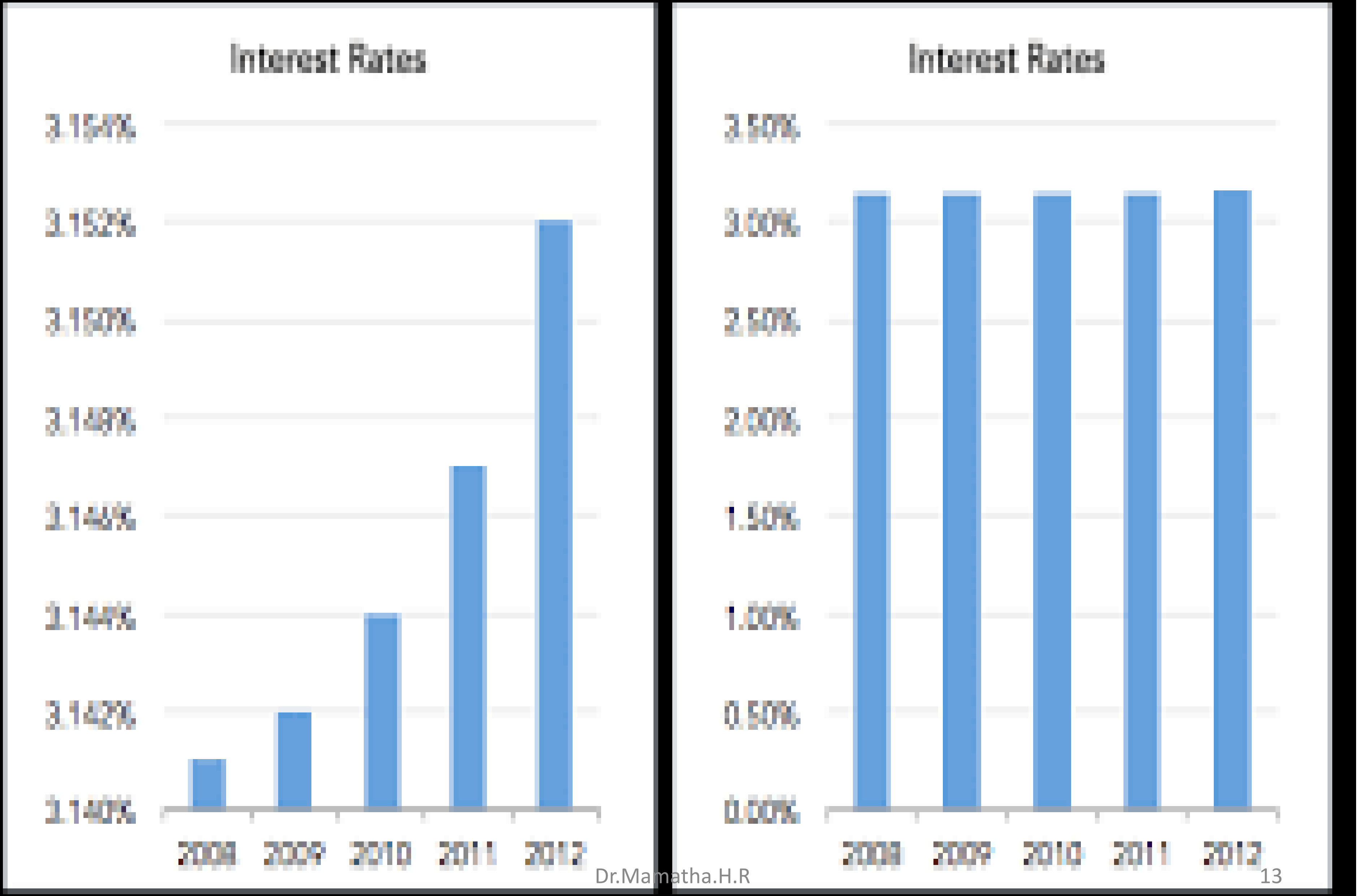
- Allows comparison of heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count/Frequency or % - truncation exaggerates differences
- Can add data labels (data values for each bar



Note truncated Y-axis

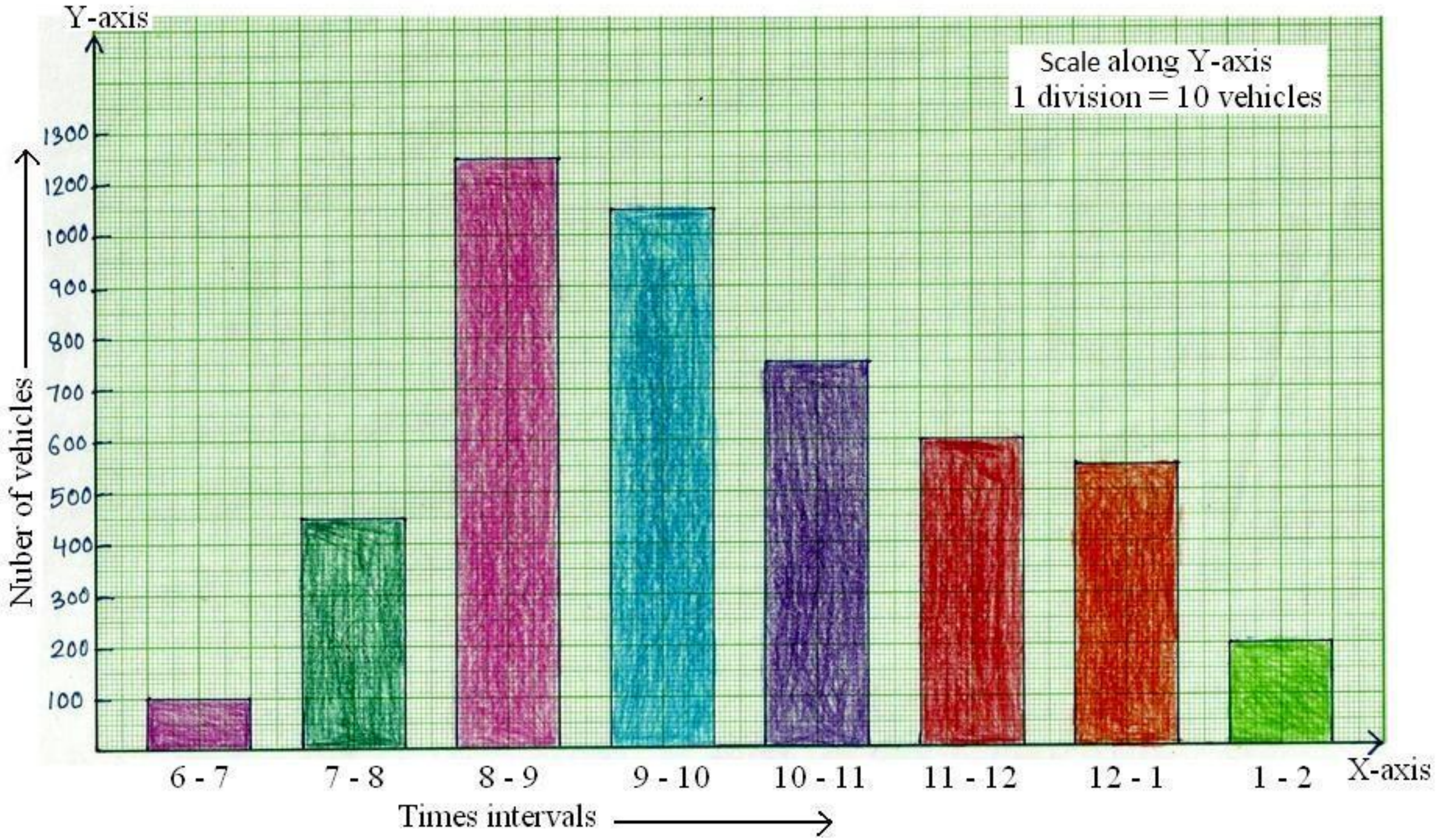


Same Data, Different Y-Axis

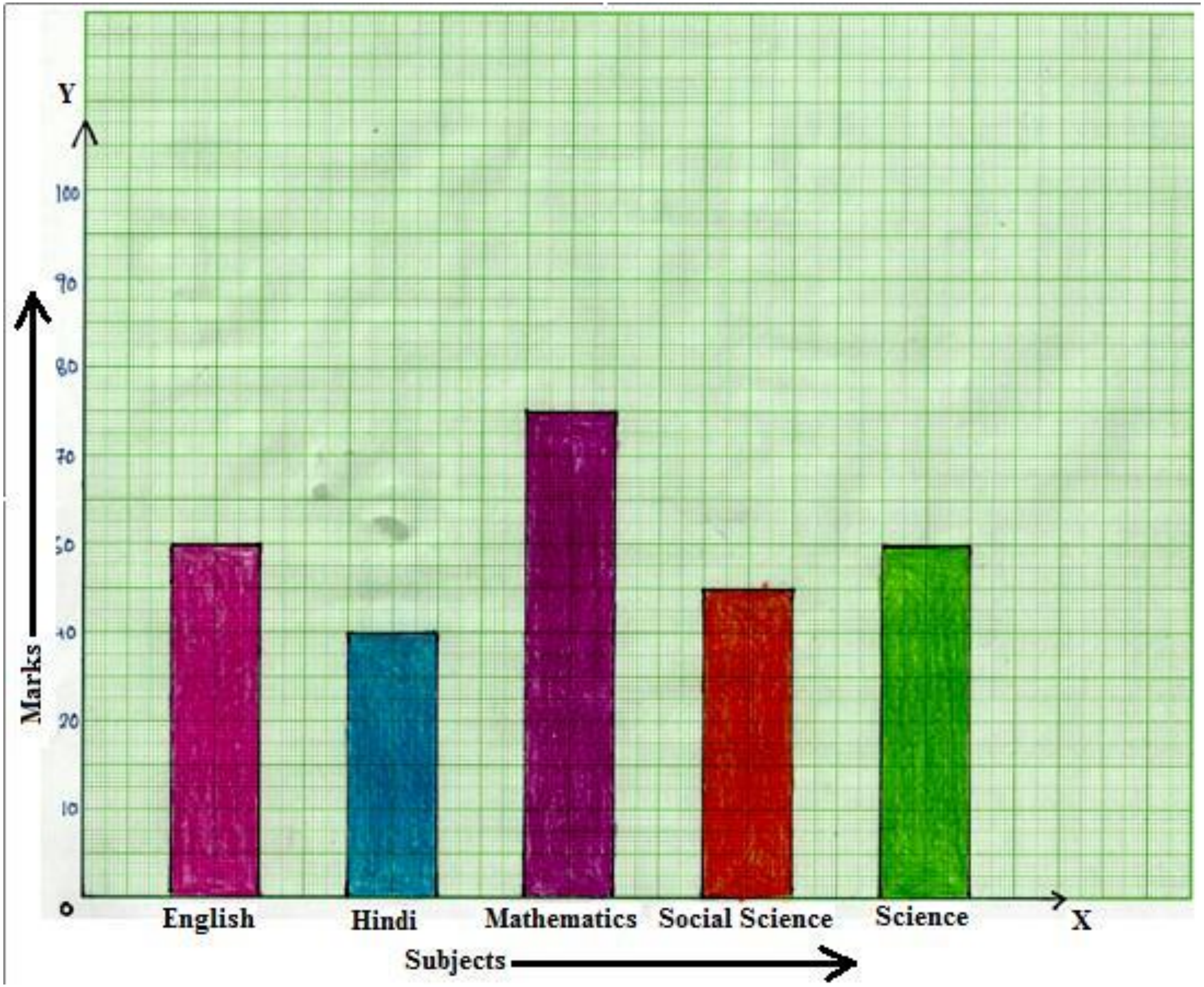


- The vehicular traffic at a busy road crossing in a particular place was recorded on a particular day from 6am to 2 pm and the data was rounded off to the nearest tens. Construct a Bar Chart.

Time in Hours	**6 - 7**	**7 - 8**	**8 - 9**	**9 - 10**	**10 - 11**	**11 - 12**	**12 - 1**	**1 - 2**
Number of Vehicles	100	450	1250	1050	750	600	550	200


Y-axis
Scale along Y-axis
1 division = 10 vehicles
1300
1200
1000
900
800
700
600
500
400
300
200
100
Nuber of vehicles
6 - 7
7 - 8
8 - 9
9 - 10
10 - 11
11 - 12
12 - 1
1 - 2
X-axis
Times intervals

- Look at the bar graph given below:



- *Read it carefully and answer the following questions.*

(i) What information does the bar graph give?

(ii) In which subject is the student very good

(iii) In which subject is he poor?

(iv) What are the average of his marks?

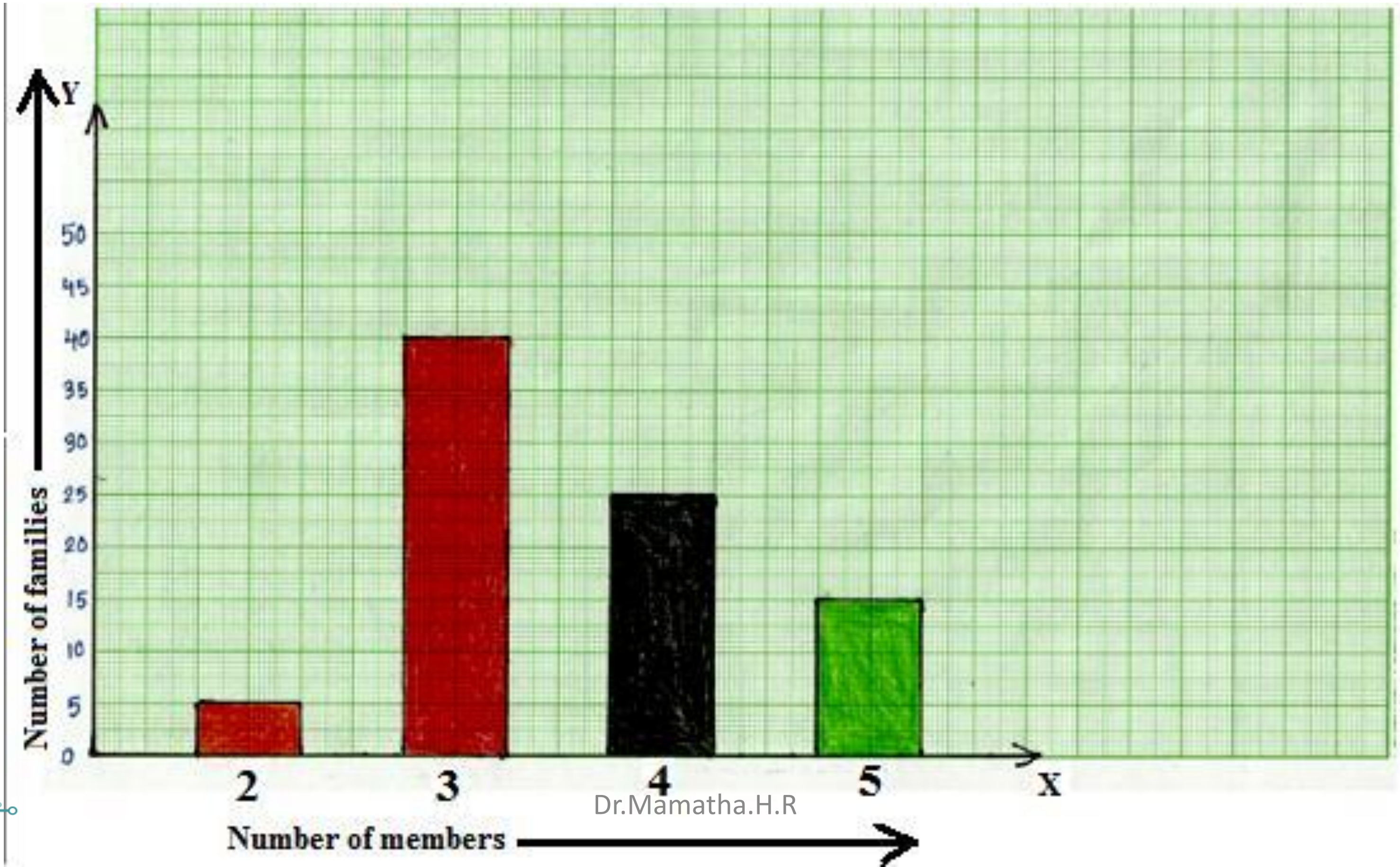
(i) It shows the marks obtained by a student in five subjects

(ii) Mathematics

(iii) Hindi

(iv) 56

- In a survey of 85 families of a colony, the number of members in each family was recorded, and the data has been represented by the following bar graph.



- *Read the bar graph carefully and answer the following questions:*
- (i) What information does the bar graph give?

  (ii) How many families have 3 members?

  (iii) How many people live alone?

  (iv) Which type of family is the most common? How many members are there in each family of this kind?

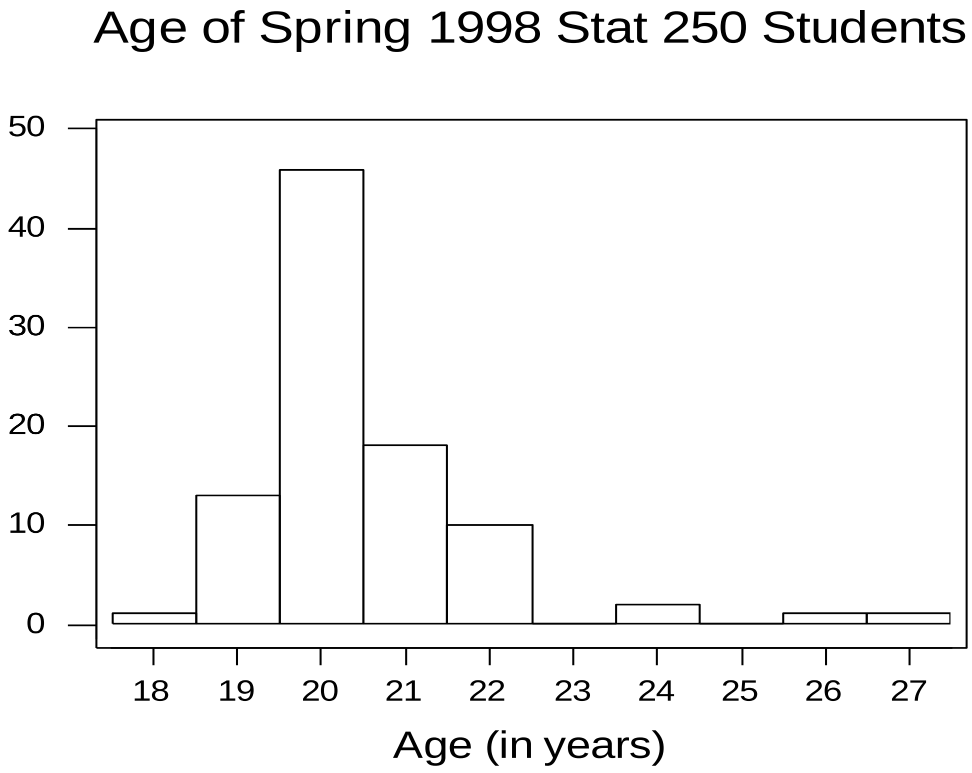
(i)It gives the number of families containing 2, 3, 4, 5 members each.

(ii) 40

(iii) none

(iv) Family having 3 members, 3 members.

Histogram



n=92 students

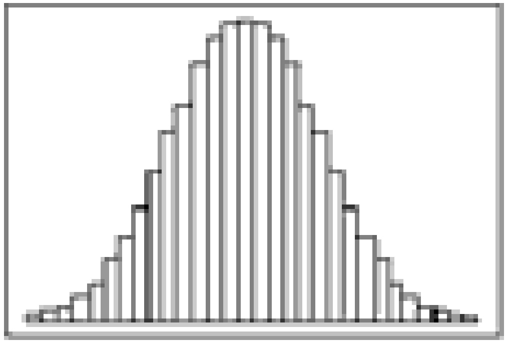
Analogy

Bar chart is to categorical data as histogram is to ...

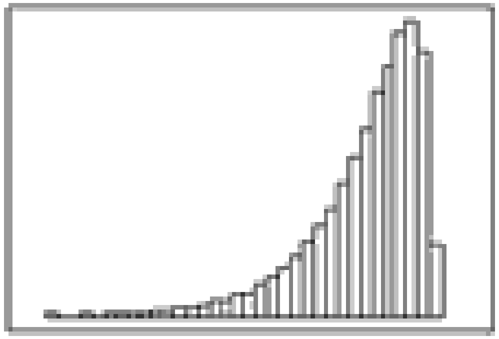
measurement data.

- **Histogram**
- defn - a diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.
- Can only be used when the data is continuous, this is huge drawback of histograms. But are very powerful tools when representing the continuous data.
- The first step in creating a histogram is to divide the entire value range into a series of intervals called "bins" and then to "drop" the individual values into the bins that they belong to.

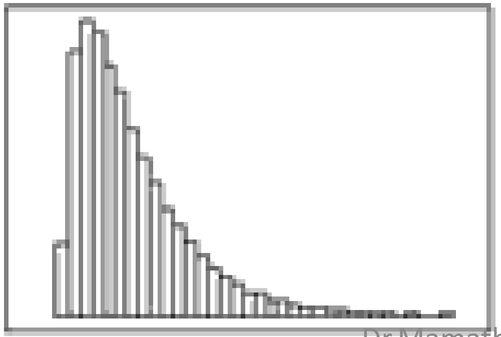
- The width of each bin may or may not be equal. If they’re equal then, the height of bins represents the frequency of data points in that range. Else bin sizes can be made equal by calculating the density of bin heights.
- These become too handy when visualizing continous data, as it can be used to show if the data is normalized, standardized, skewed.



Symmetric
Bell shaped



Skewed to
the Left



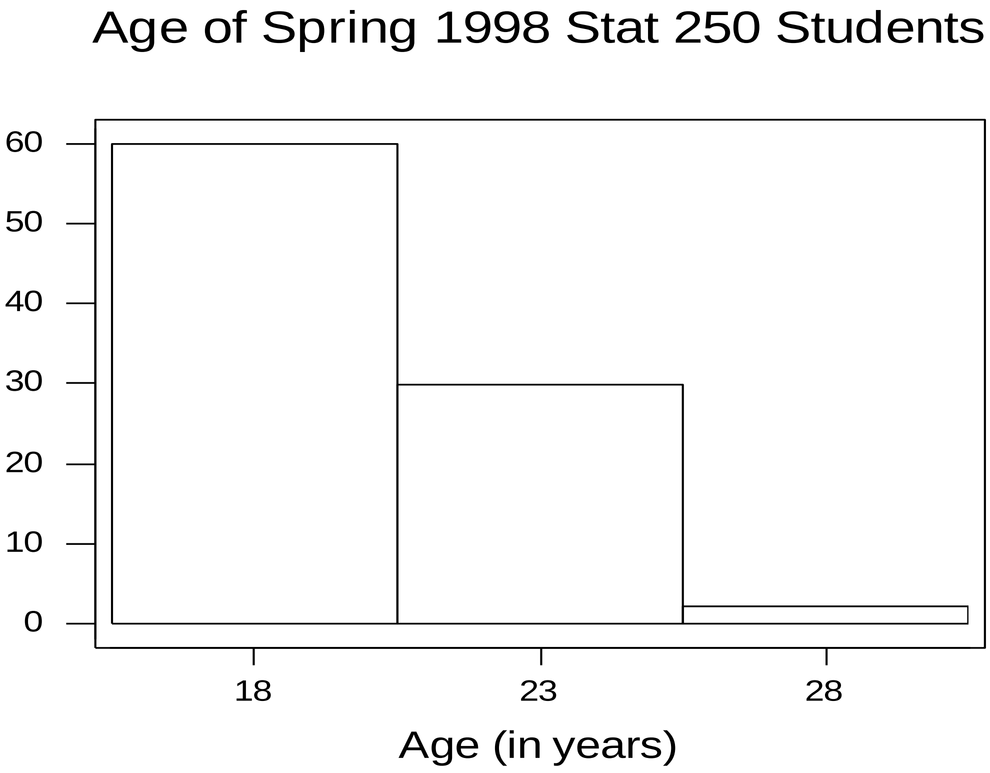
Skewed to
the Right

Histogram

Use common sense in determining number of categories to use.

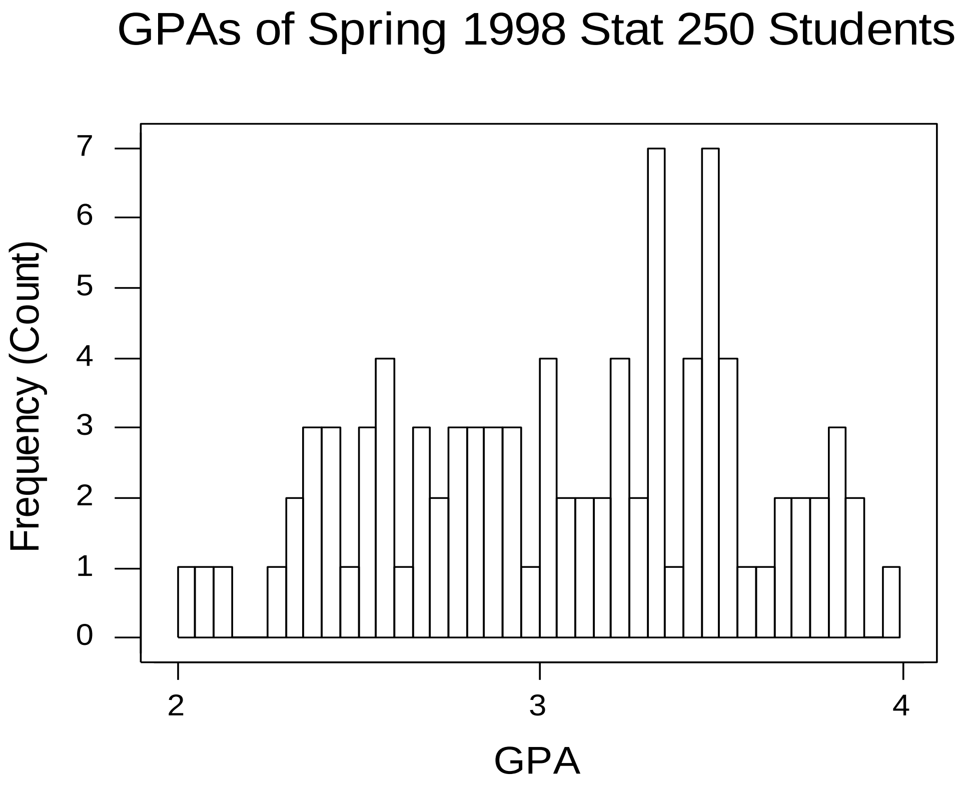
(Trial-and-error works fine, too.)

Too few categories



n=92 students

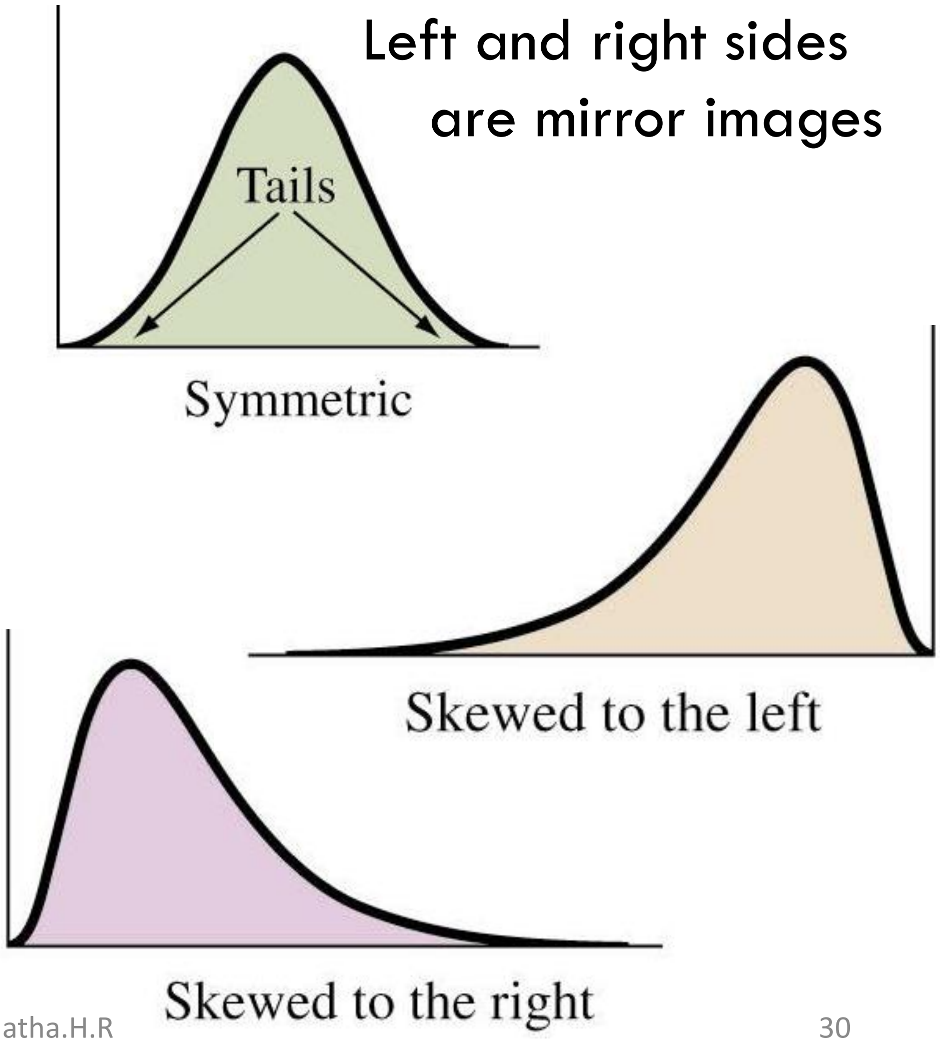
Too many categories



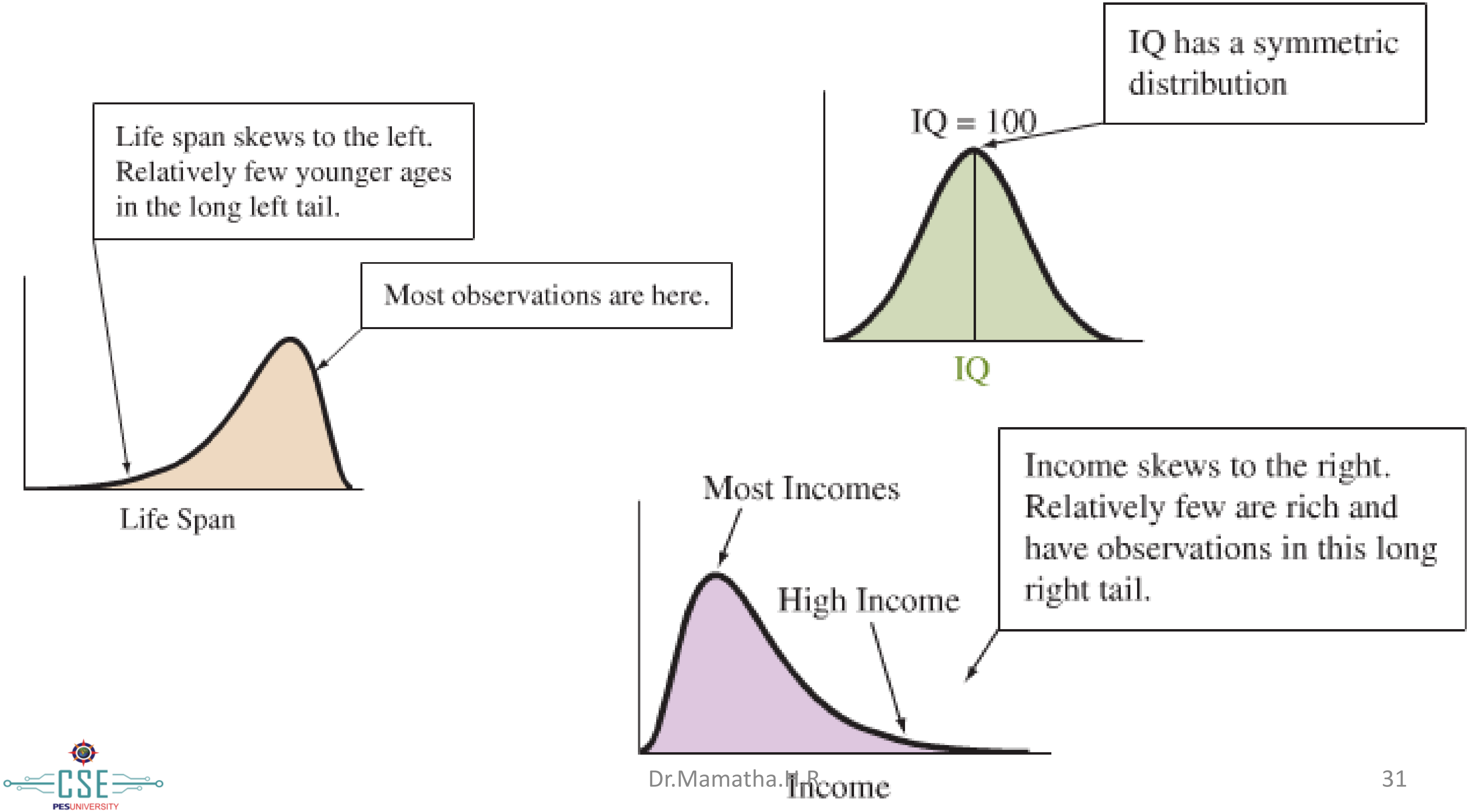
n=92 students

Interpreting Histograms

- Assess where a distribution is **centered** by finding the median
- Assess the **spread** of a distribution
- **Shape** of a distribution: roughly symmetric, skewed to the right, or skewed to the left



Examples of Skewness



Life span skews to the left. Relatively few younger ages in the long left tail.

Most observations are here.

IQ has a symmetric distribution

Income skews to the right. Relatively few are rich and have observations in this long right tail.

Shape and Skewness

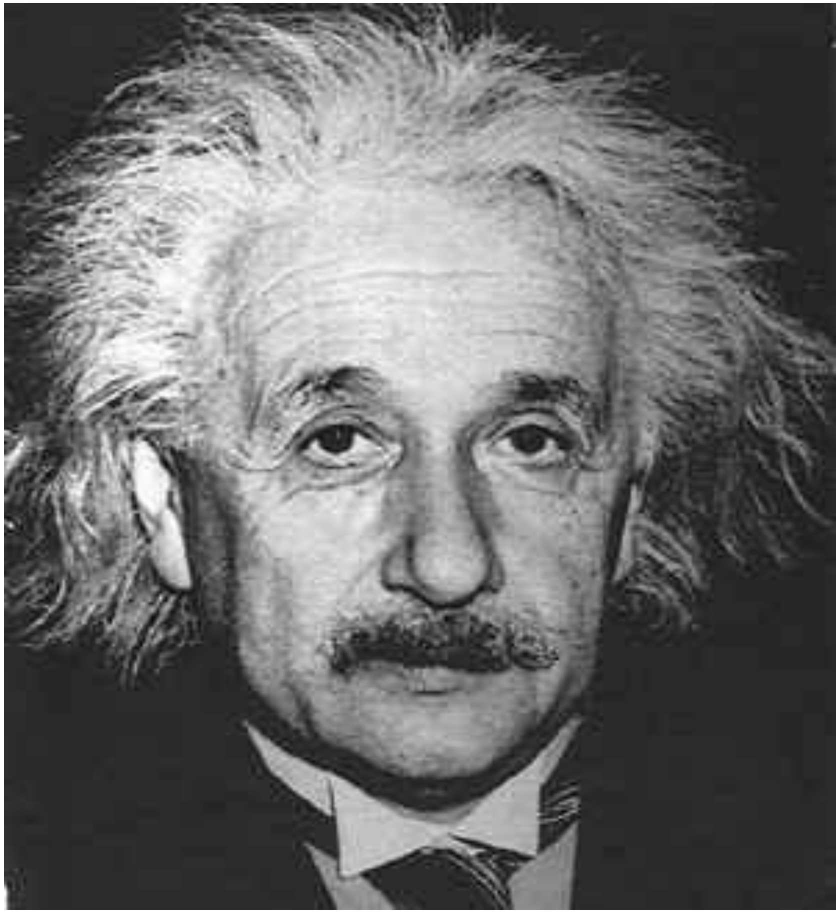
Consider a data set containing IQ scores for the general public. What shape?

a. Symmetric

b. Skewed to the left

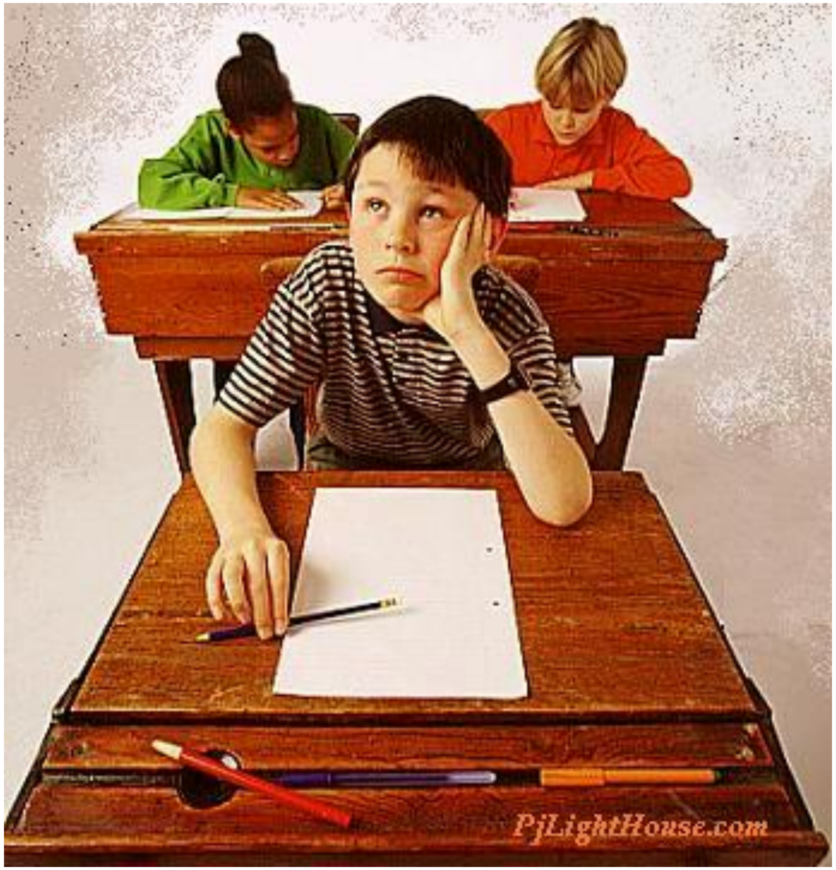
c. Skewed to the right

d. Bimodal



botit.botany.wisc.edu

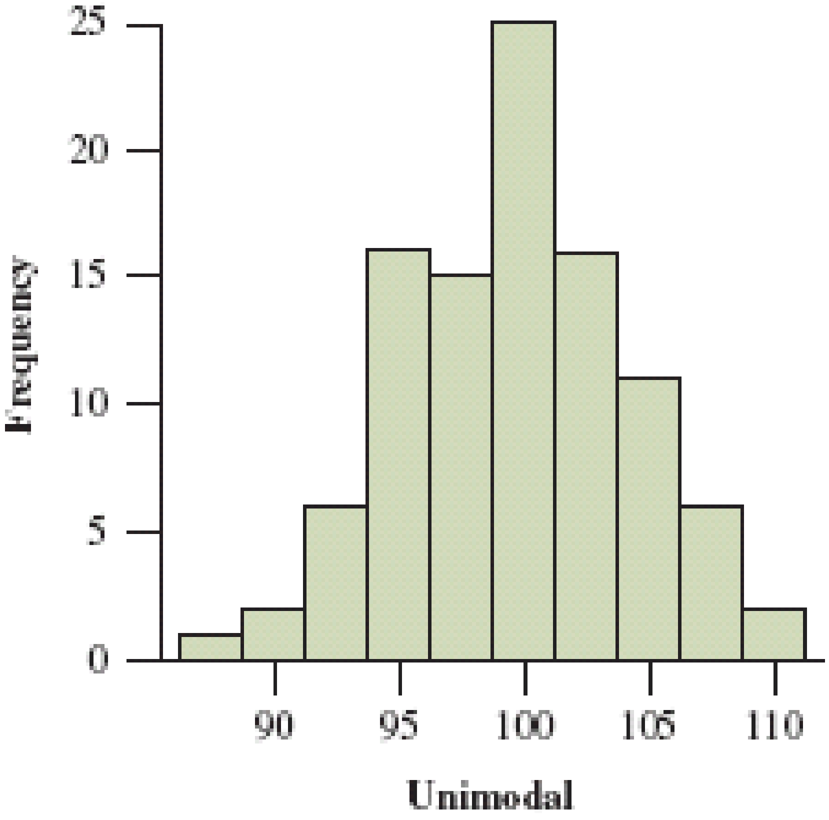
Shape and Skewness

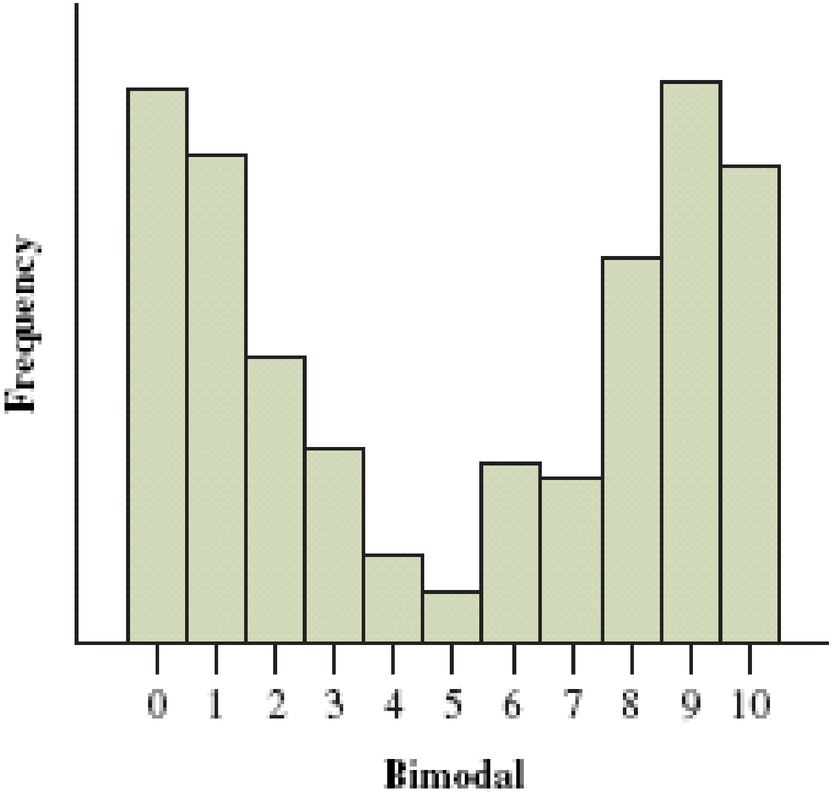


Consider a data set of the scores of students on an easy exam in which most score very well but a few score poorly. What shape?

a. Symmetric
b. Skewed to the left
c. Skewed to the right
d. Bimodal

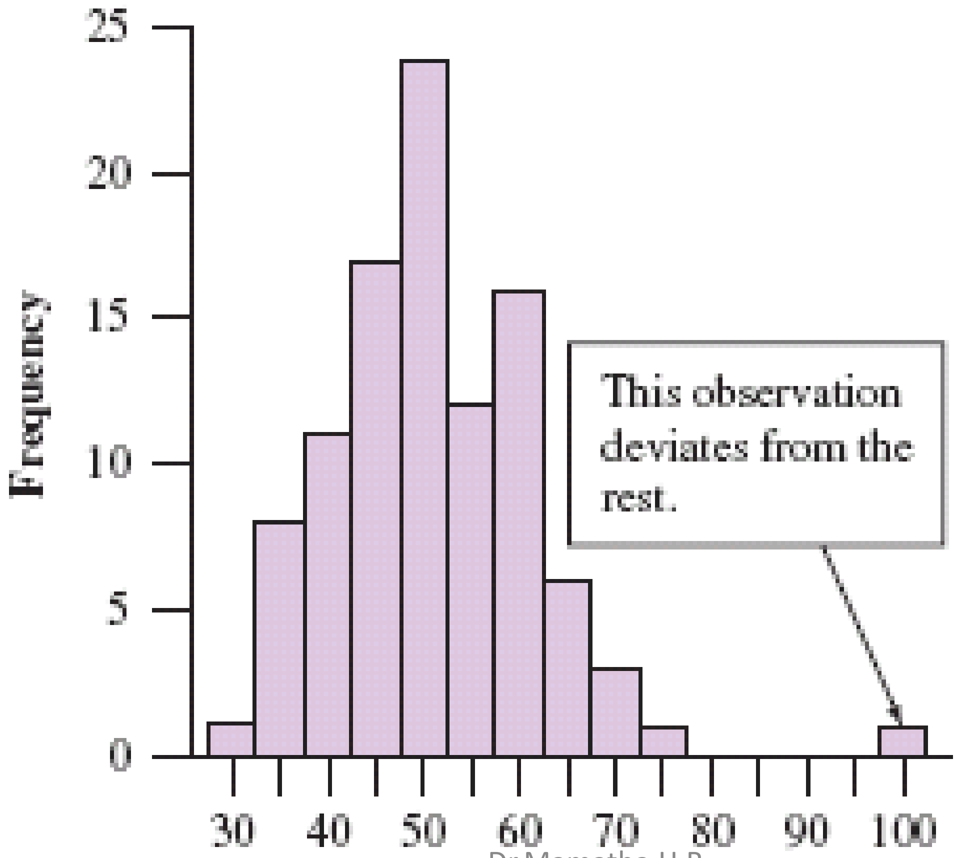
Shape: Type of Mound



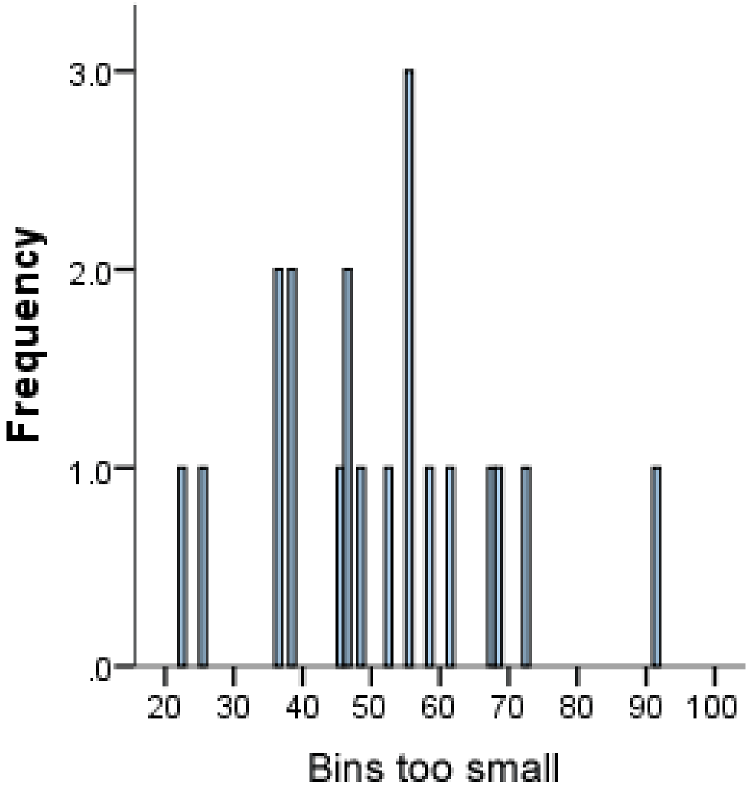


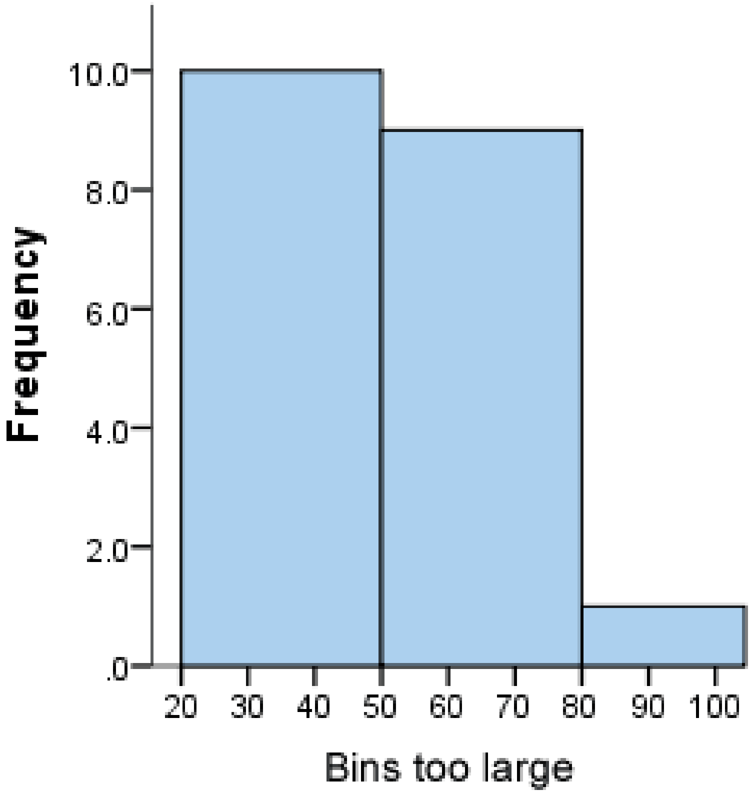
Outlier

An outlier falls far from the rest of the data



This observation deviates from the rest.


Frequency
3.0
2.0
1.0
.0
20
30
40
50
60
70
80
90
100
Bins too small


Frequency
10.0
8.0
6.0
4.0
2.0
.0
20
30
40
50
60
70
80
90
100
Bins too large

- Histograms are based on area, not height of bars

- In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin.

- This means that the height of the bar does not necessarily indicate how many occurrences of scores there were within each individual bin.

- It is the product of height multiplied by the width of the bin that indicates the frequency of occurrences within that bin.

- One of the reasons that the height of the bars is often incorrectly assessed as indicating frequency and not the area of the bar is due to the fact that a lot of histograms often have equally spaced bars (bins), and under these circumstances, the height of the bin does reflect the frequency

- The number of bins k can be assigned directly or can be calculated from a suggested bin width h as:
- k=(max-min)/h ---h is bin width
- k=sqrt(n) ----used in Excel

- In statistics, the Freedman–Diaconis rule can be used to select the size of the bins to be used in a histogram

$$\text{Bin size} = 2\,\frac{\mathrm{IQR}(x)}{\sqrt[3]{n}}$$

- Example
- The following frequency distribution gives the masses of 48 objects measured to the nearest gram. Draw a histogram to illustrate the data.

Mass (g)	**10 – 19**	**20 – 24**	**25 – 34**	**35 – 50**	**51 – 55**
Frequency	6	4	12	18	8

Since the class widths are not equal, we choose a convenient width as a standard and adjust the heights of the rectangles accordingly.

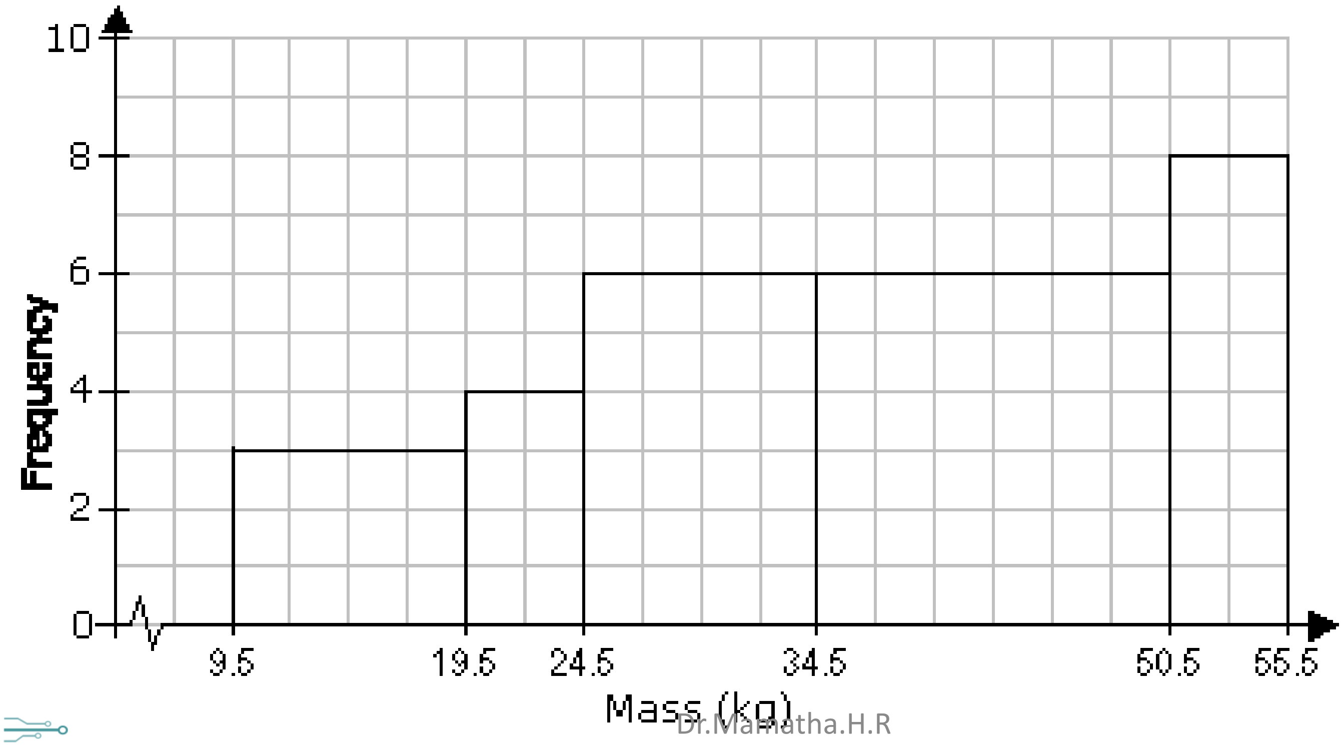
We notice that the smallest width size is 5.

We can choose 5 to be the standard width.

The other widths are then multiples of the standard width.

Mass (g)	**10 – 19**	**20 – 24**	**25 – 34**	**35 – 50**	**51 – 55**
Frequency	6	4	12	18	8
Class width	10	5	10	15	5

Mass (g)	10 – 19	20 – 24	25 – 34	35 – 50	51 – 55
Frequency	6	4	12	18	8
Class widths	10	5	10	15	5
	2 × standard	standard	2 × standard	3 × standard	standard
Rectangle’s height in histogram	6 ÷ 2 = 3	4	12 ÷ 2 = 6	18 ÷ 3 = 6	8



- The weather in Los Angeles is dry most of the time, but it can be quite rainy in the winter. The rainiest month of the year is February. The following table presents the annual rainfall in Los Angeles, in inches, for each February from 1965 to 2006.

0.2	3.7	1.2	13.7	1.5	0.2	1.7
0.6	0.1	8.9	1.9	5.5	0.5	3.1
3.1	8.9	8.0	12.7	4.1	0.3	2.6
1.5	8.0	4.6	0.7	0.7	6.6	4.9
0.1	4.4	3.2	11.0	7.9	0.0	1.3
2.4	0.1	2.8	4.9	3.5	6.1	0.1

- “Frequency “ presents the numbers of data points that fall into each of the class intervals.
- “Relative Frequency” presents the frequencies divided by the total number of data points.
- The relative frequency of a class interval is the proportion of data points that fall into the interval.
- Note that since every data point is in exactly one class interval, the relative frequencies must sum to 1.
- “Density” presents the relative frequency divided by the class width.
- Frequency Density presents the frequency divided by the class width

- Note that when the classes are of equal width, the frequencies, relative frequencies, and densities are proportional to one another.

- When the class intervals are of unequal widths, the heights of the rectangles must be set equal to the densities. The areas of the rectangles will then be the relative frequencies.

TABLE 1.2 Particulate matter (PM) emissions (in g/gal) for 62 vehicles driven at high altitude

7.59	6.28	6.07	5.23	5.54	3.46	2.44	3.01	13.63	13.02	23.38	9.24	3.22
2.06	4.04	17.11	12.26	19.91	8.50	7.81	7.18	6.95	18.64	7.10	6.04	5.66
8.86	4.40	3.57	4.35	3.84	2.37	3.81	5.32	5.84	2.89	4.68	1.85	9.14
8.67	9.52	2.68	10.14	9.20	7.31	2.09	6.32	6.53	6.32	2.01	5.91	5.60
5.61	1.50	6.46	5.29	5.64	2.07	1.11	3.32	1.83	7.56			

TABLE 1.4 Frequency table for PM emissions of 62 vehicles driven at high altitude

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 13	1	0.0161	0.0081
13–< 15	2	0.0323	0.0161
15–< 17	0	0.0000	0.0000
17–< 19	2	0.0323	0.0161
19–< 21	1	0.0161	0.0081
21–< 23	0	0.0000	0.0000
23–< 25	1	0.0161	0.0081

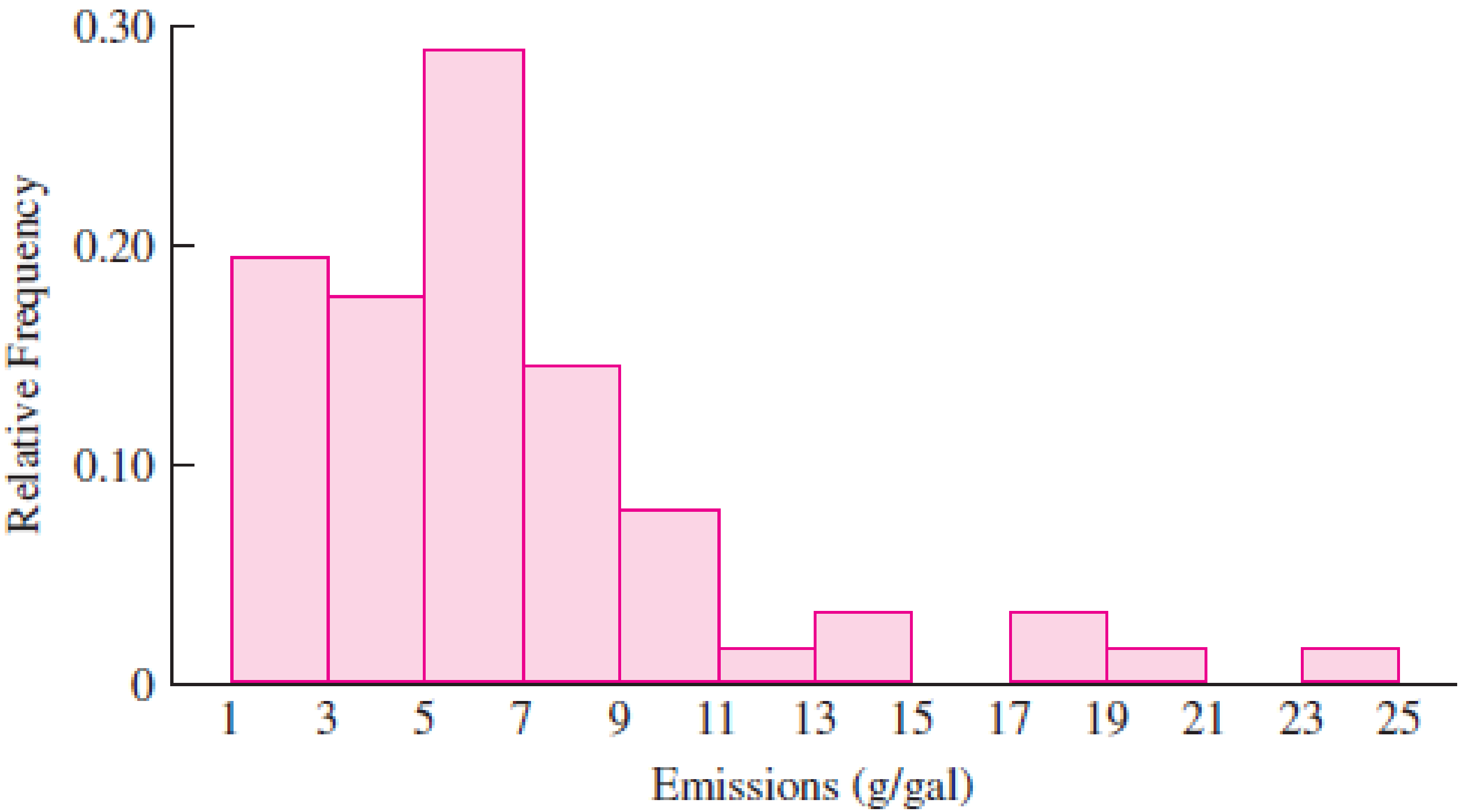
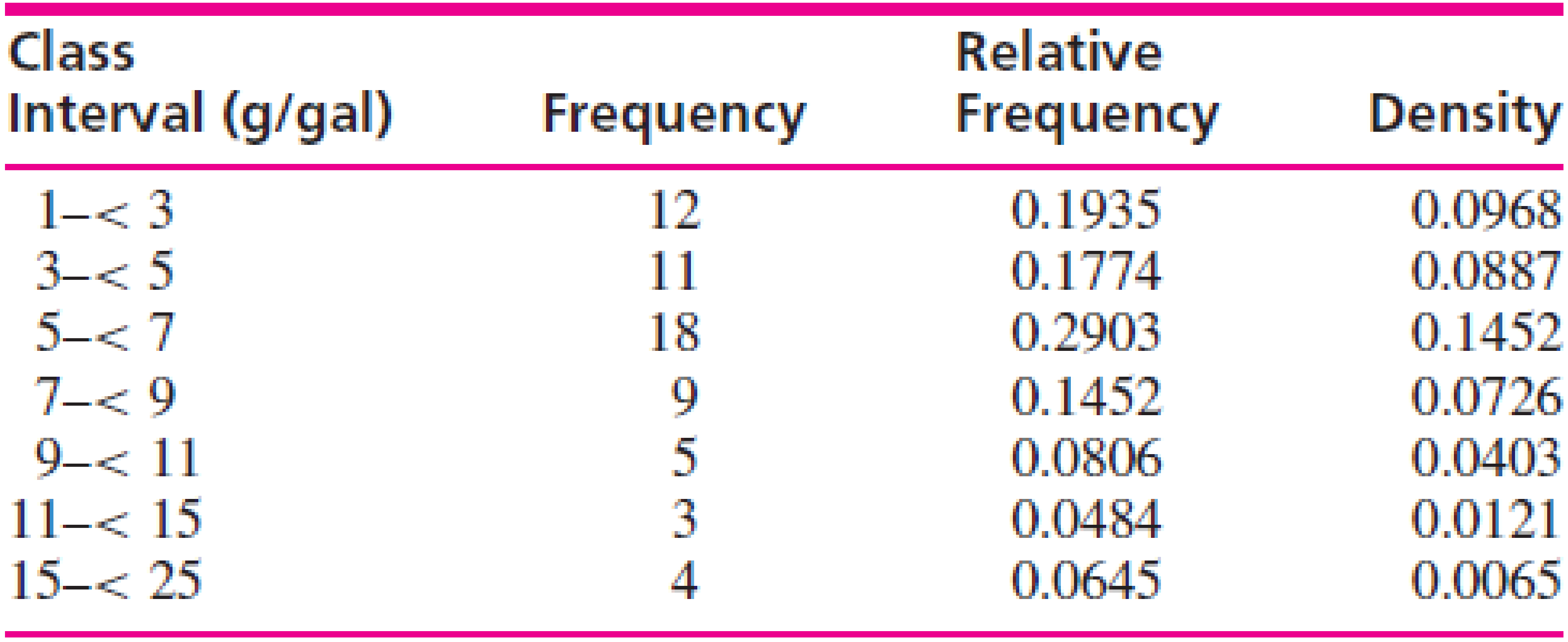
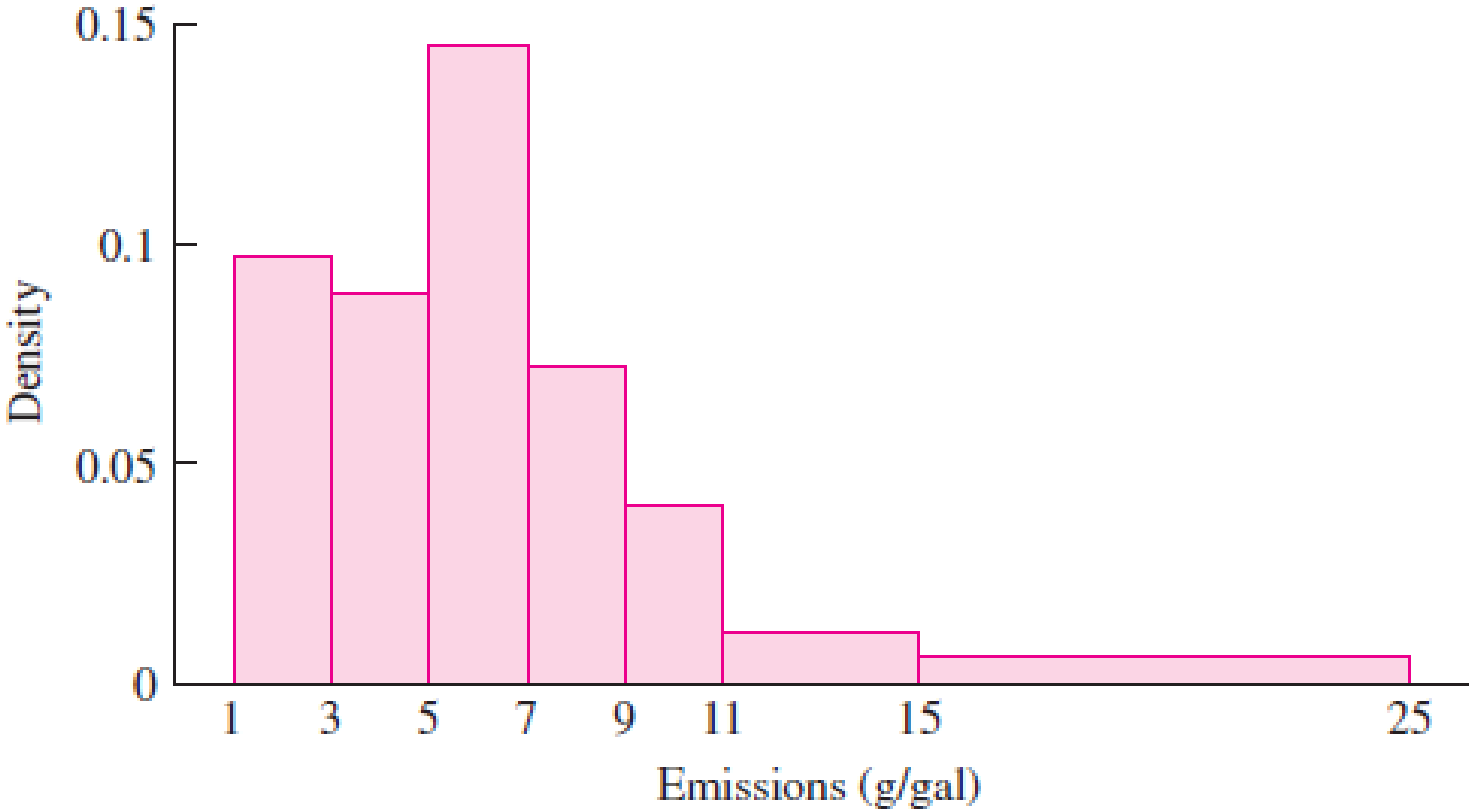

0.30
0.20
0.10
0
Relative Frequency
1
3
5
7
9
11
13
15
17
19
21
23
25
Emissions (g/gal)

TABLE 1.5 Frequency table, with unequal class widths, for PM emissions of 62 vehicles driven at high altitude

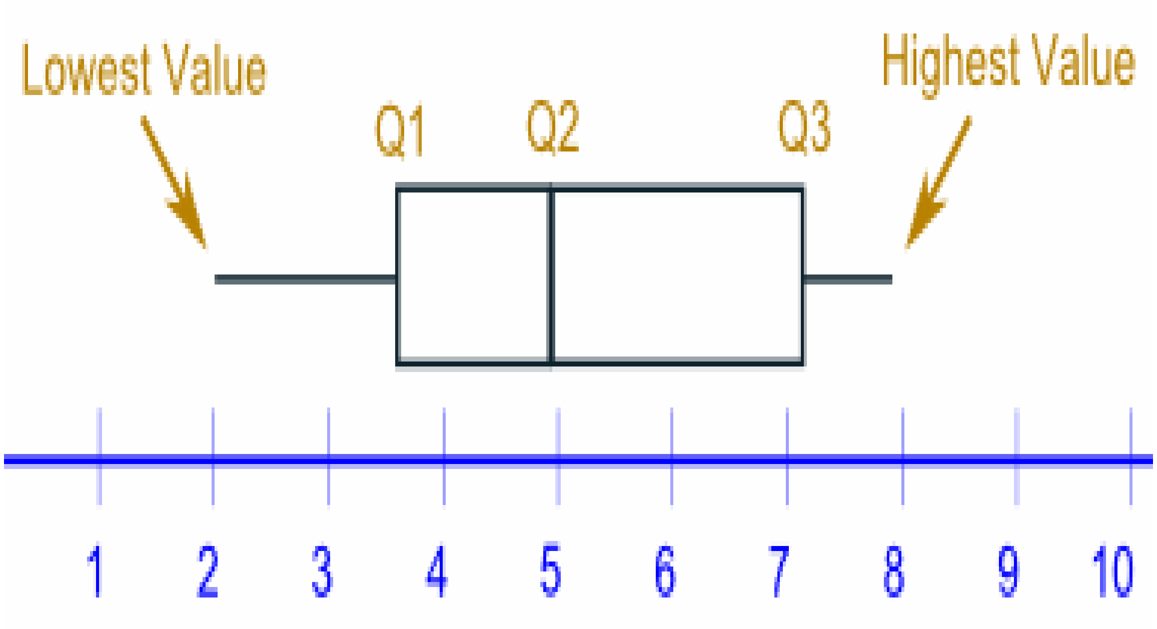


Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 15	3	0.0484	0.0121
15–< 25	4	0.0645	0.0065


0.15
0.1
0.05
0
Density
1
3
5
7
9
11
15
25
Emissions (g/gal)

Box and Whisker Plot

We can show all the important values in a "Box and Whisker Plot", like this:



Example: Box and Whisker Plot and Interquartile Range for

4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11

Put them in order:

3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

Cut it into quarters:

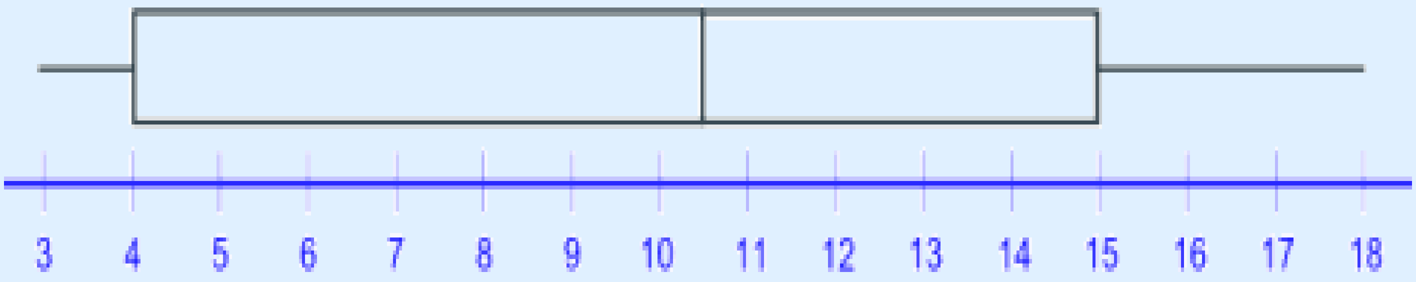
3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18

In this case all the quartiles are between numbers:

- Quartile 1 (Q1) = (4+4)/2 = **4**
- Quartile 2 (Q2) = (10+11)/2 = **10.5**
- Quartile 3 (Q3) = (14+16)/2 = **15**

- The Lowest Value is 3,
- The Highest Value is 18

So now we have enough data for the **Box and Whisker Plot**:

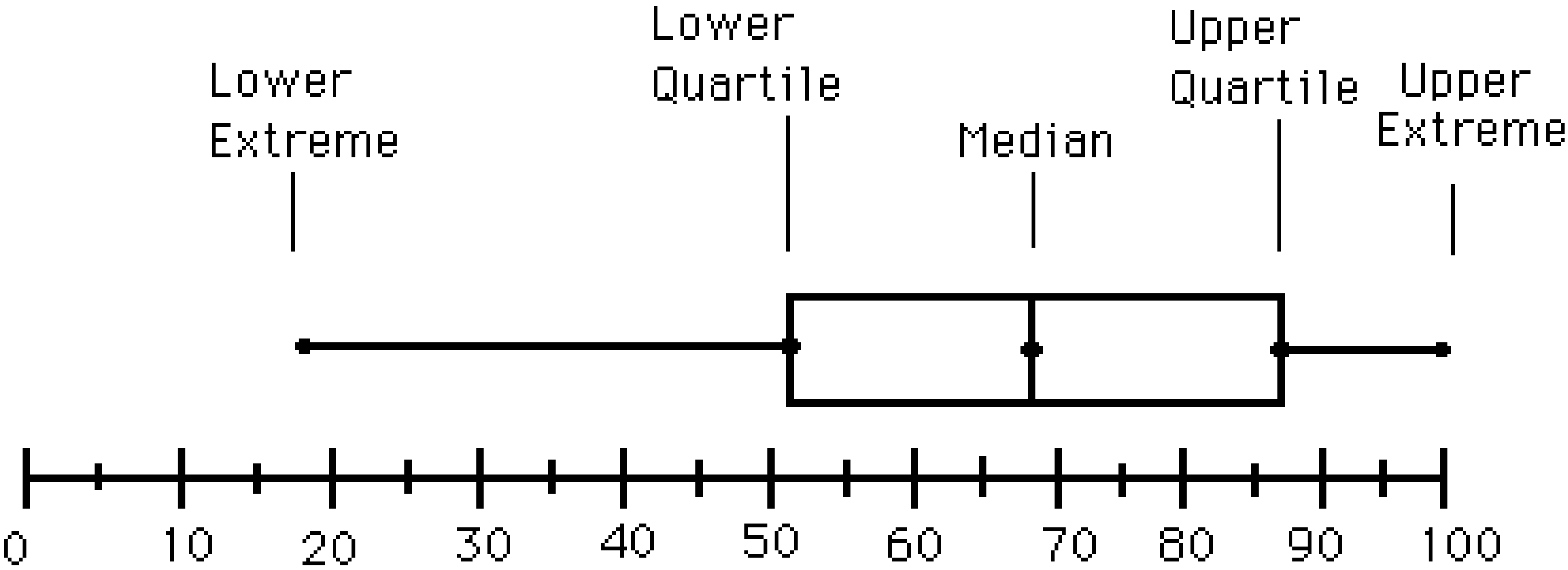


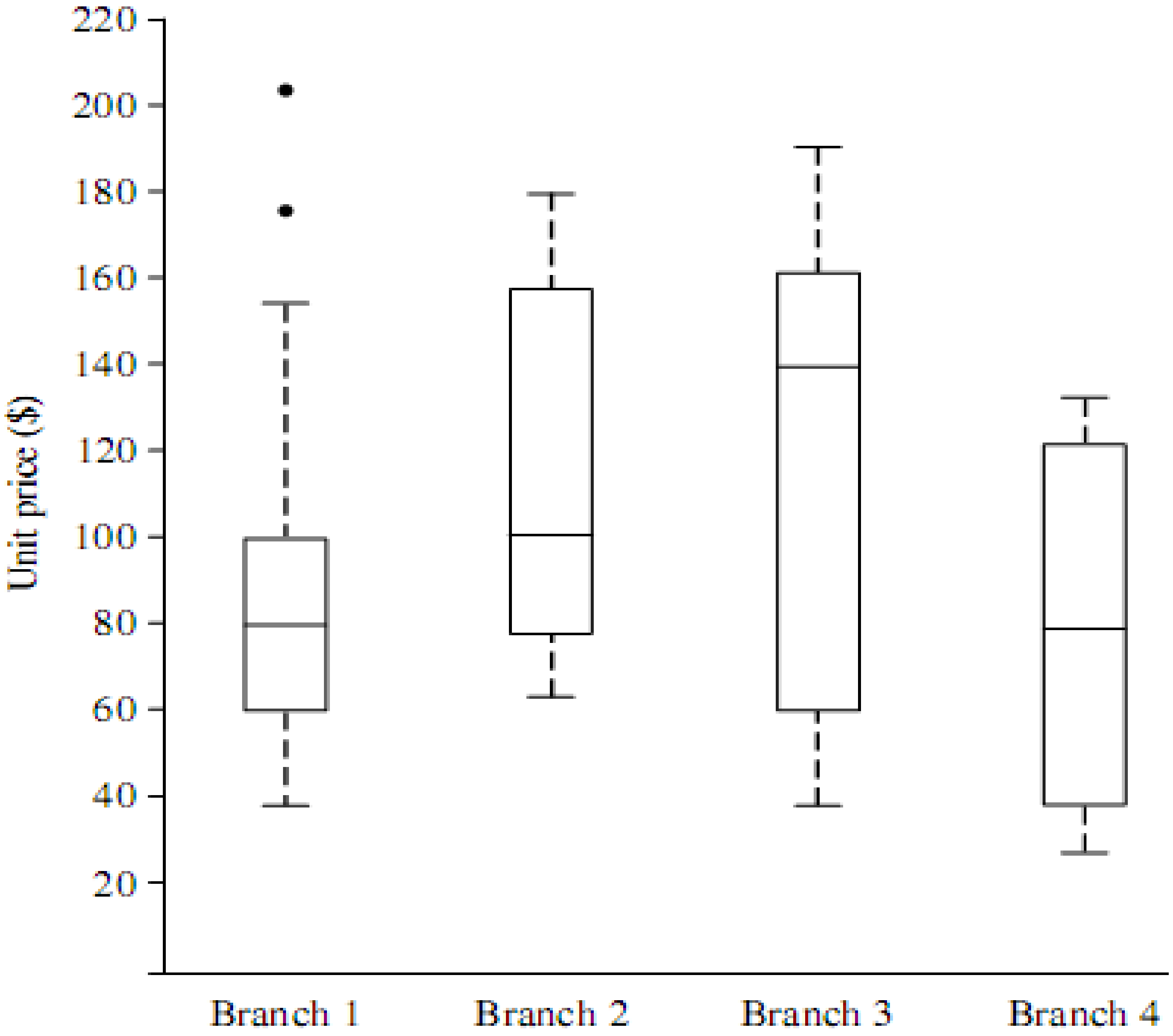
And the **Interquartile Range** is:

Q3 - Q1 = 15 - 4 = **11**

Boxplot Analysis

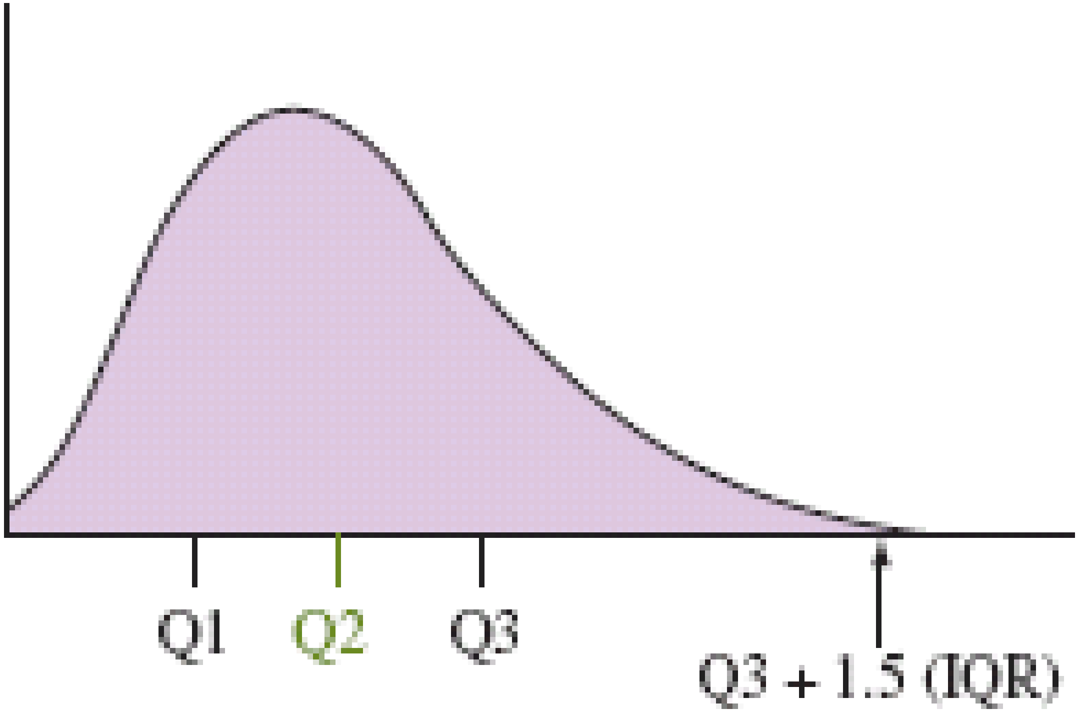
- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually


Lower Extreme
Lower Quartile
Median
Upper Quartile
Upper Extreme
0
10
20
30
40
50
60
70
80
90
100


220
200
180
160
140
120
100
80
60
40
20
Unit price ($)
Branch 1
Branch 2
Branch 3
Branch 4

Criteria for Identifying an Outlier

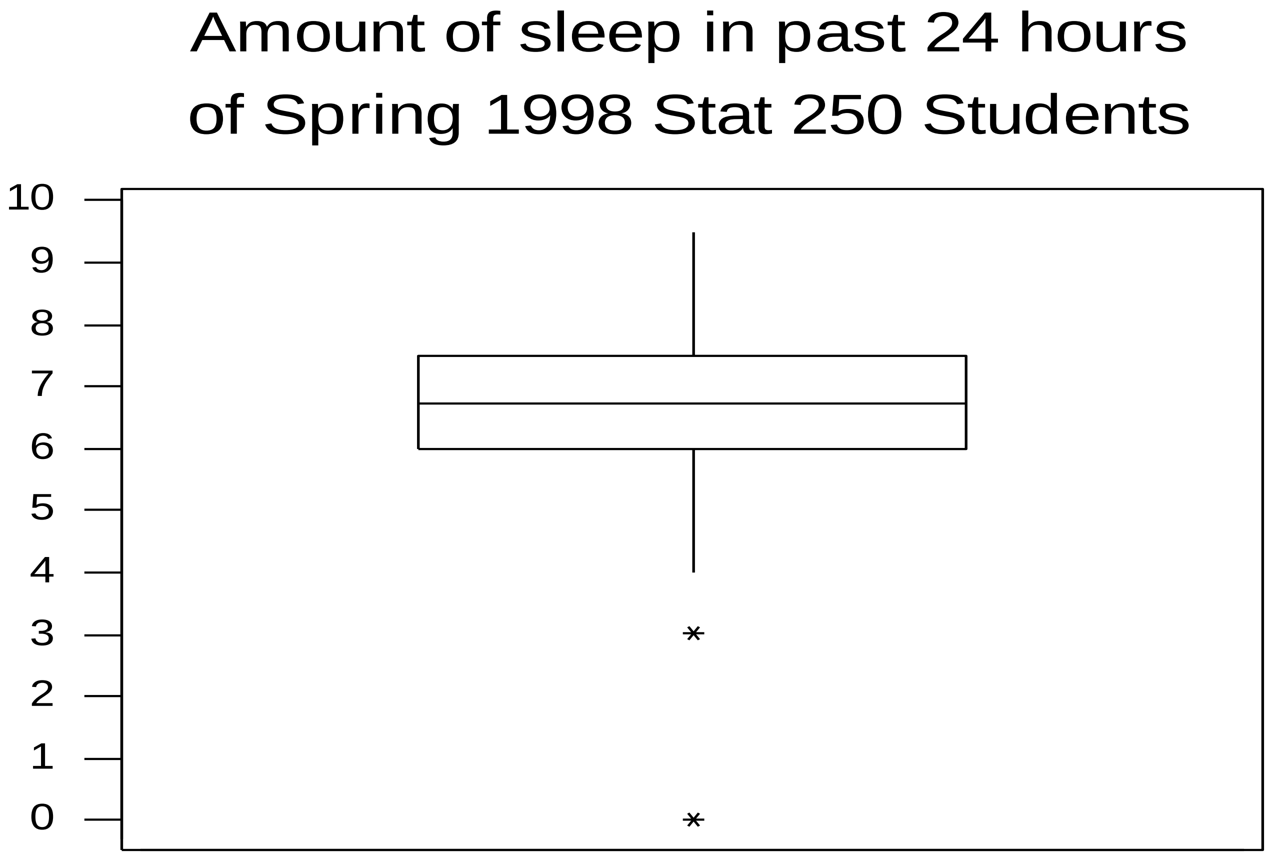
An observation is a potential outlier if it falls more than *1.5 x IQR* below the first or more than *1.5 x IQR* above the third quartile.



Boxplot(Box and Whiskers plot)

- **Boxplot** is a method for graphically depicting groups of numerical data through their quartiles.
- It’s probably the best method to identify the outliers in the data.
- The whole data is divided into 4 quartiles, 1st Quartile 2nd Quartile, 3rd Quartile, 4th Quartile.
- 1st Quartile = Q1 – Lower Extreme
- 2nd Quartile = Q2 – Q1
- 3rd Quartile = Q3 – Q2
- 4th Quartile = Upper Extreme – Q3

Box Plot


Amount of sleep in past 24 hours
of Spring 1998 Stat 250 Students
10
9
8
7
6
5
4
3
2
1
0
*
*

Box Plot

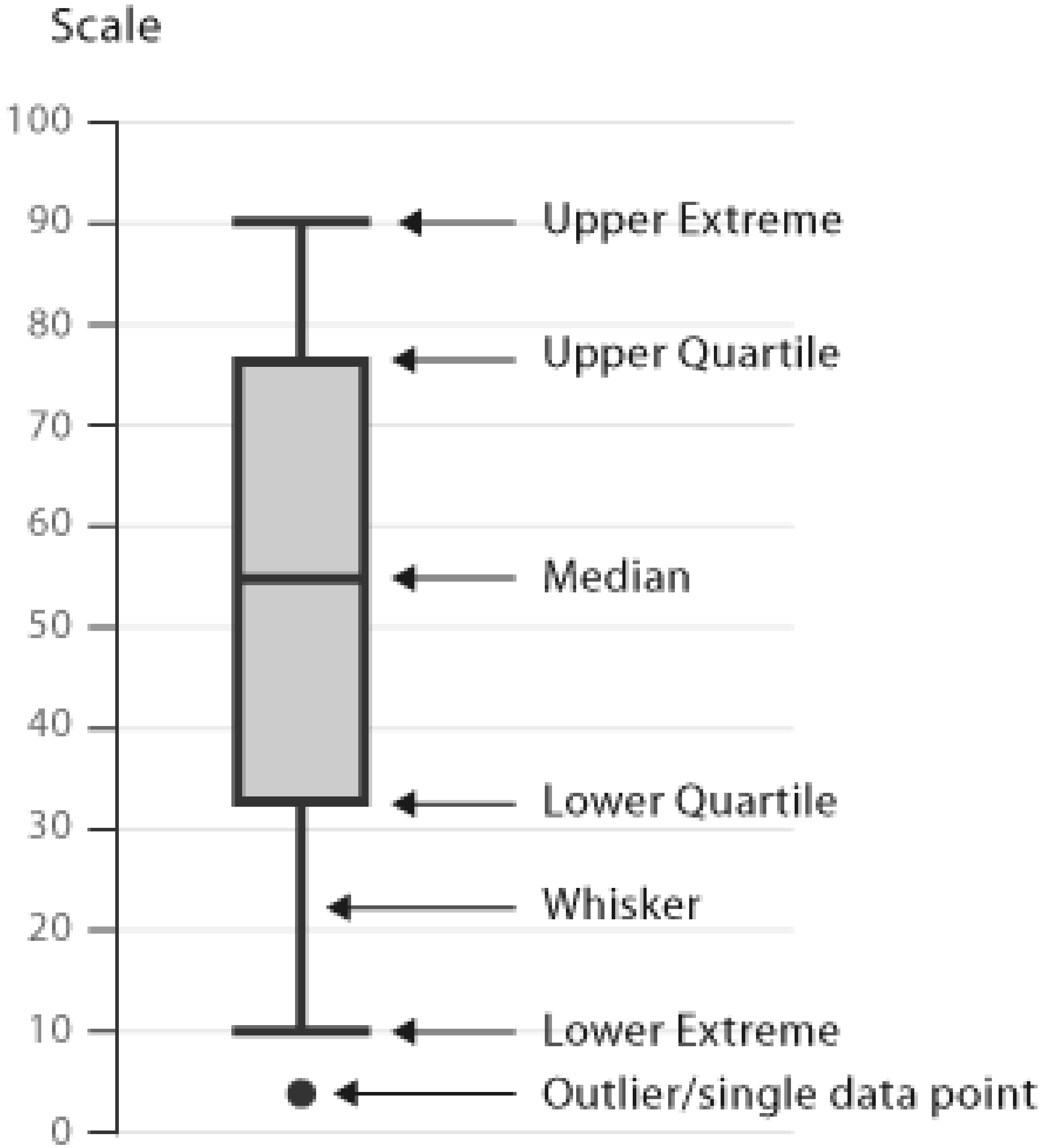
- Summarizes measurement data.
- Vertical (or horizontal) axis represents measurement scale.
- Lines in box represent the 25th percentile (“**first quartile**”), the 50th percentile (“**median**”), and the 75th percentile (“**third quartile**”), respectively.

An aside...

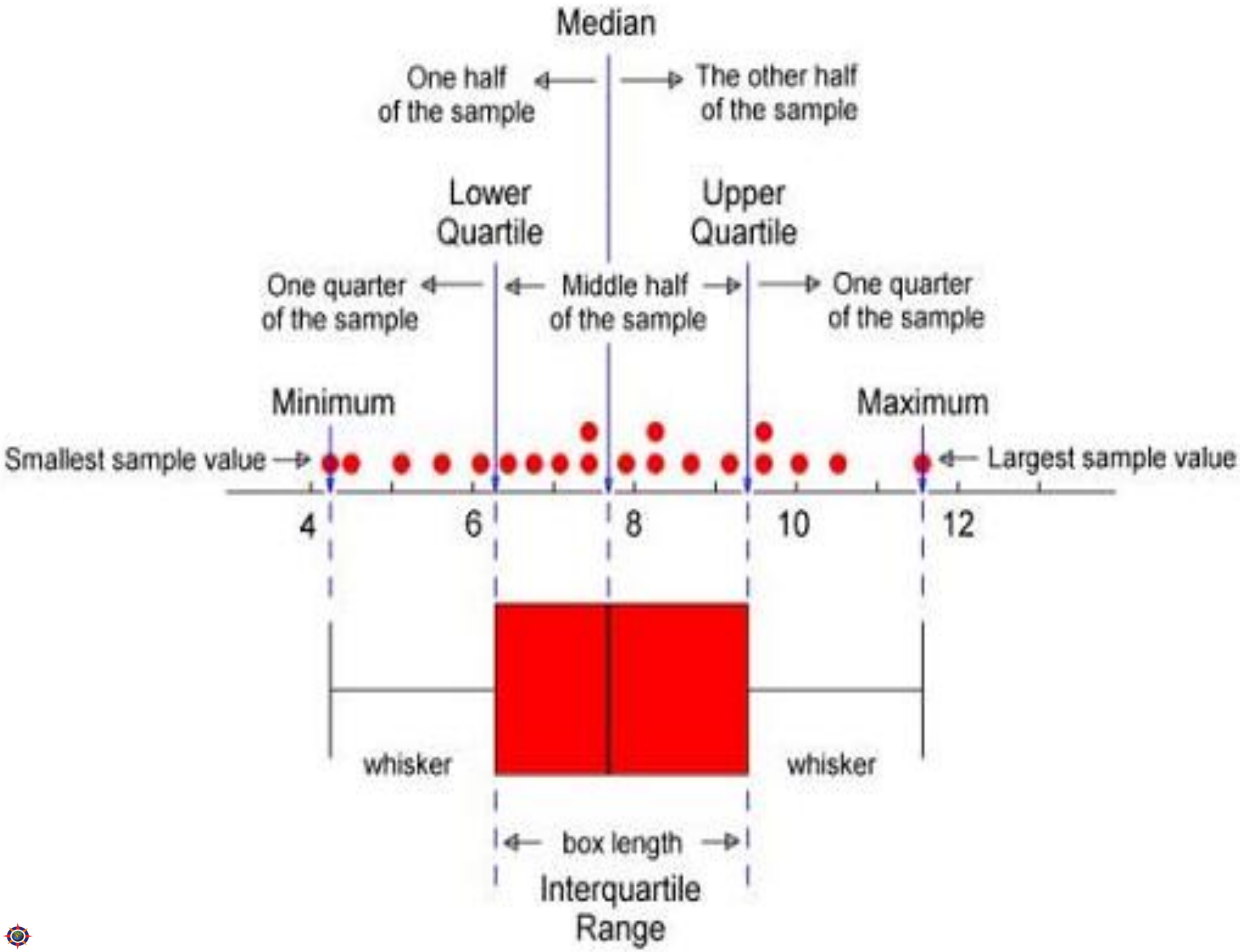
- Roughly speaking:
  - The "**25th percentile**" is the number such that 25% of the data points fall below the number.
  - The "**median**" or "**50th percentile**" is the number such that half of the data points fall below the number.
  - The "**75th percentile**" is the number such that 75% of the data points fall below the number.

Box Plot (cont’d)

- “**Whiskers**” are drawn to the most extreme data points that are not more than 1.5 times the length of the box beyond either quartile.
  - Whiskers are useful for identifying outliers.
- “**Outliers**,” or extreme observations, are denoted by asterisks.
  - Generally, data points falling beyond the whiskers are considered outliers.

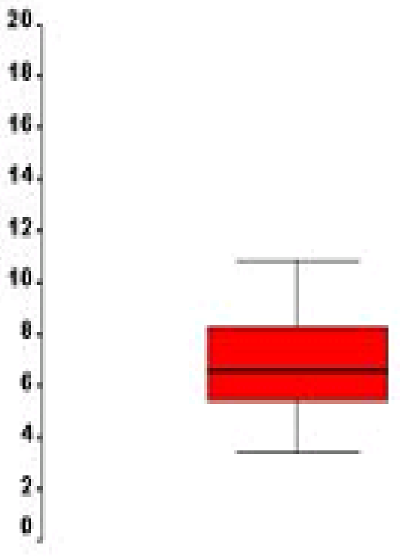

Scale
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme
Upper Quartile
Median
Lower Quartile
Whisker
Lower Extreme
Outlier/single data point

- Upper extreme - max (Q3 + 1.5 times IQR, max value in data)
- Upper quartile – Q3
- Median – Q2
- Lower Quartile - Q1
- Lower extreme - min (Q1 - 1.5 times IQR, min value in data)
- IQR – is inter quartile range = Q3 – Q1
- A normal distribution has Q1 – Q2 = Q2 – Q3, which means there's equal amount of data spread between Q1 to Q2 and Q2 to Q3. The mean also coincides with Q2 as it is the median of the data.

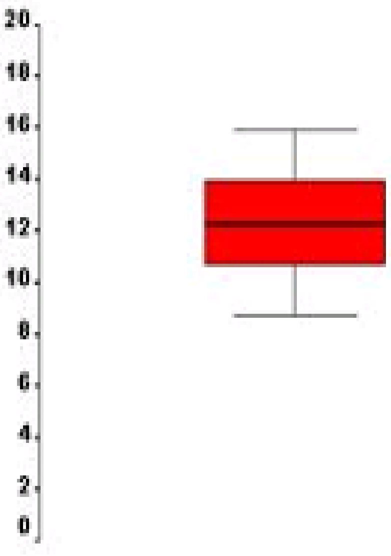

Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
Middle half of the sample
One quarter of the sample
Minimum
Maximum
Smallest sample value
Largest sample value
4
6
8
10
12
whisker
whisker
box length
Interquartile Range

- The box length gives an indication of the sample variability and the line across the box shows where the sample is centered.
- The position of the box in its whiskers and the position of the line in the box also tells us whether the sample is symmetric or skewed, either to the right or left.

The Boxplot as an Indicator of Centrality

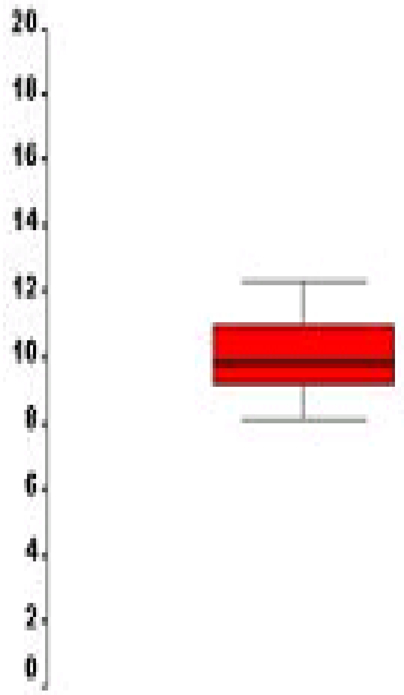


The boxplot of a sample of 20 points from a population centred on 7.

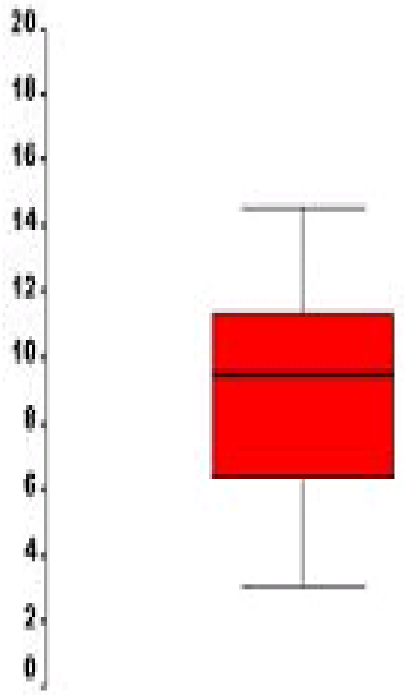


The boxplot of a sample of 20 points from a population centred on 12.

The Boxplot as an Indicator of Spread

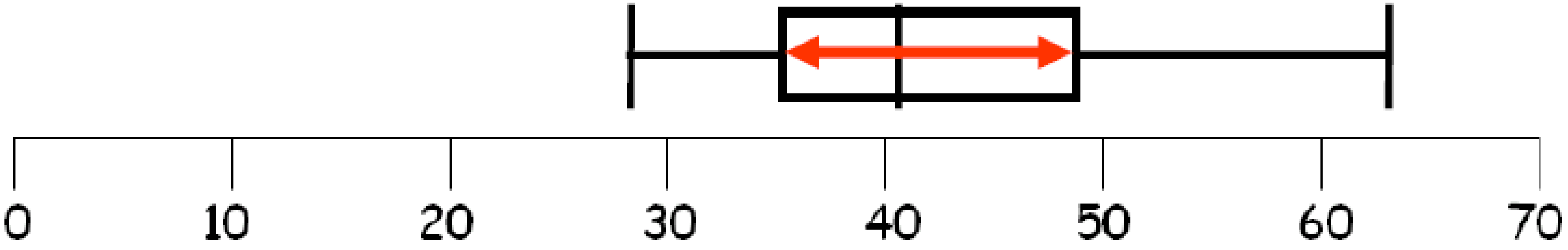


The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 1.

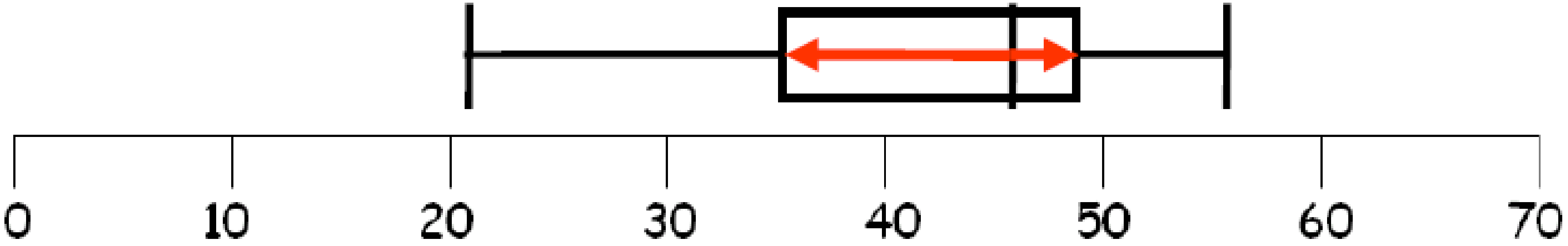


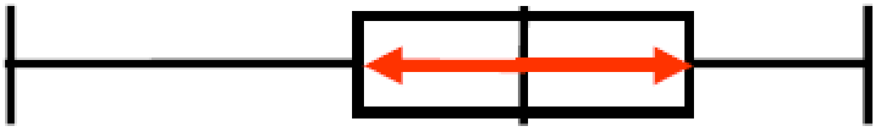
The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 3.

Positive skew: median closer to LQ than UQ



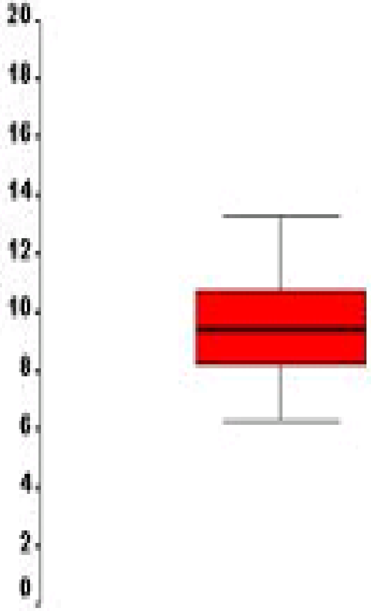
Negative skew: median closer to UQ than LQ



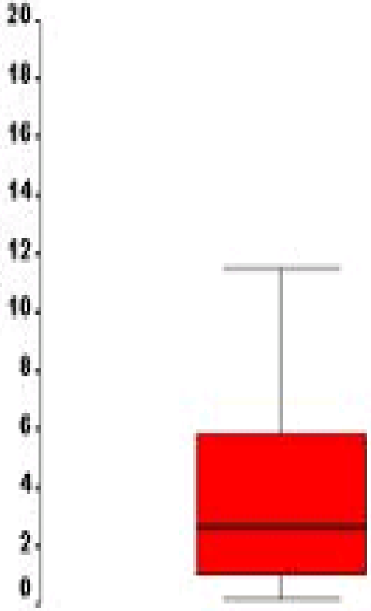


Symmetrical distribution

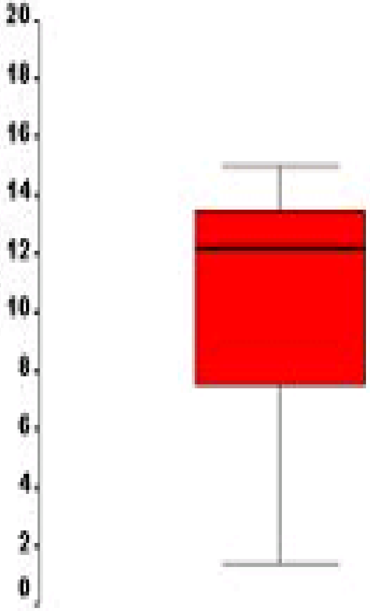
The Boxplot as an Indicator of Symmetry



The boxplot of a sample of 20 points from a symmetric population. The line is close to the centre of the box and the whisker lengths are the same.

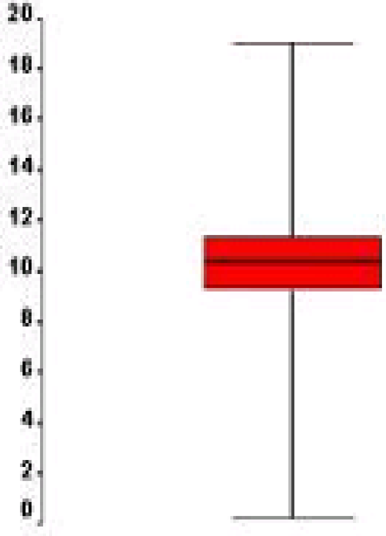


The boxplot of a sample of 20 points from a population which is skewed to the right. The top whisker is much longer than the bottom whisker and the line is gravitating towards the bottom of the box.

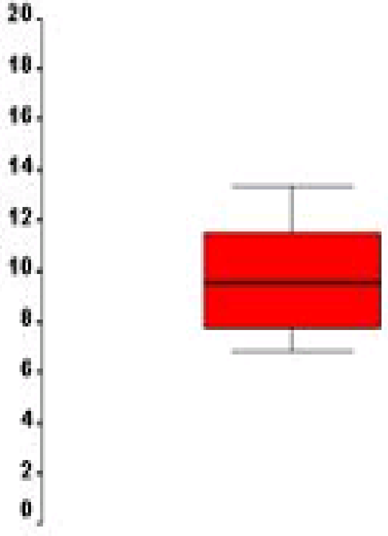


The boxplot of a sample of 20 points from a population which is skewed to the left. The bottom whisker is much longer than the top whisker and the line is rising to the top of the box.

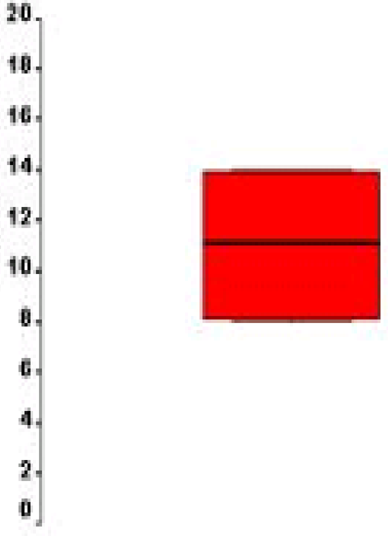
The Boxplot as an Indicator of Tail Length



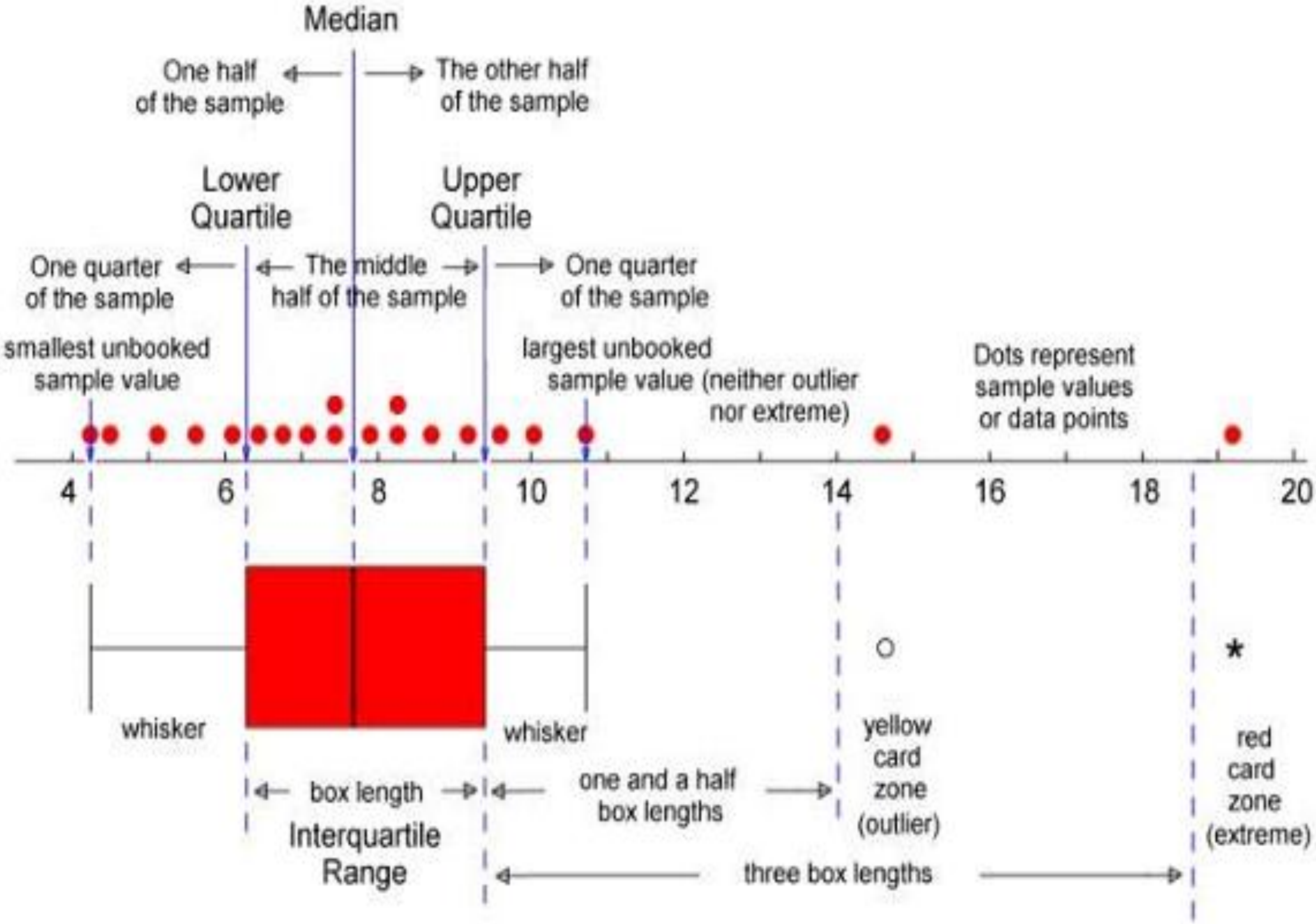
The boxplot of a sample of 20 points from a population with long tails. The length of the whiskers far exceeds the length of the box. (A well proportioned tail would give rise to whiskers about the same length as the box, or maybe slightly longer.)



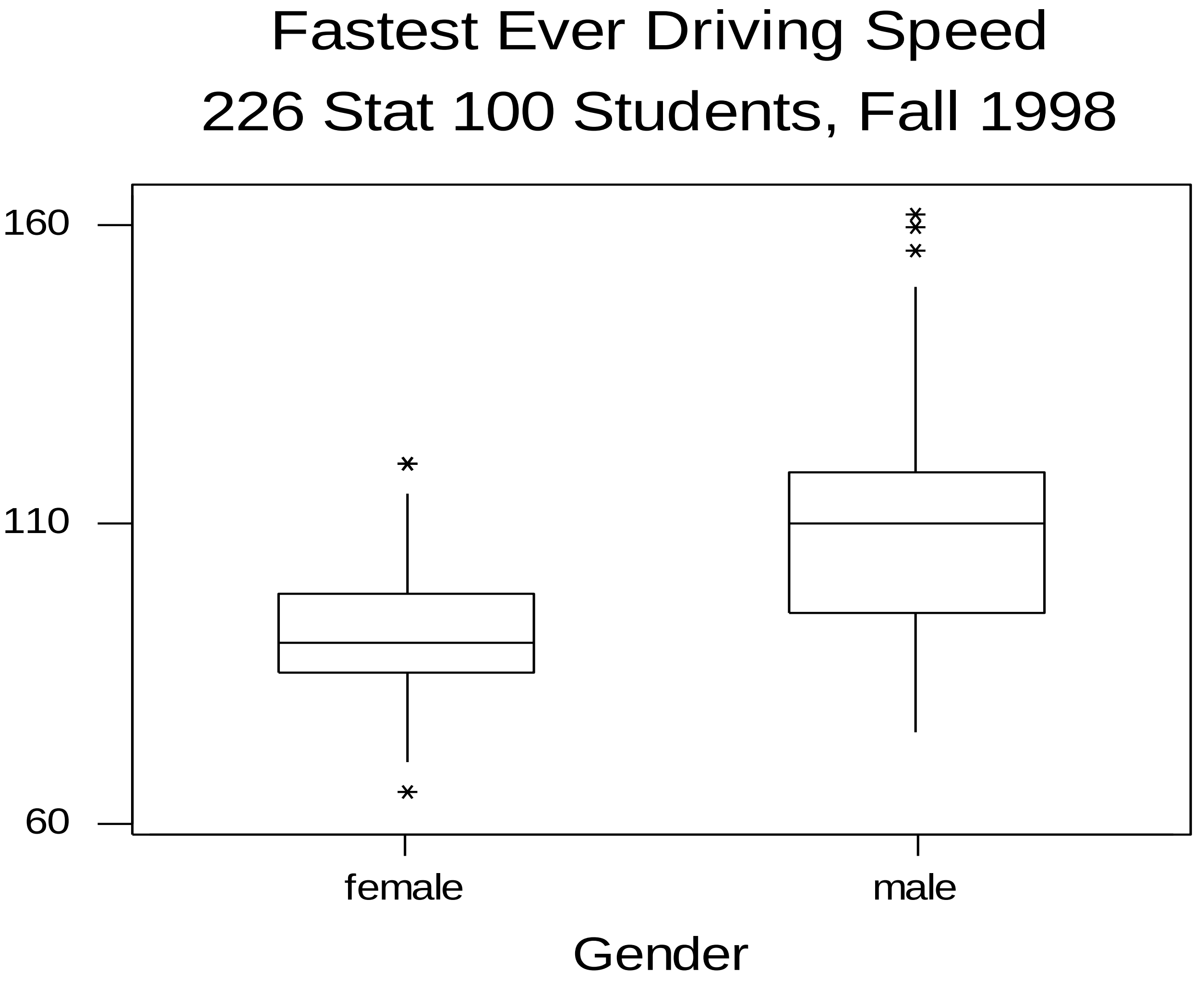
The boxplot of a sample of 20 points from a population with short tails. The length of the whiskers is shorter than the length of the box.

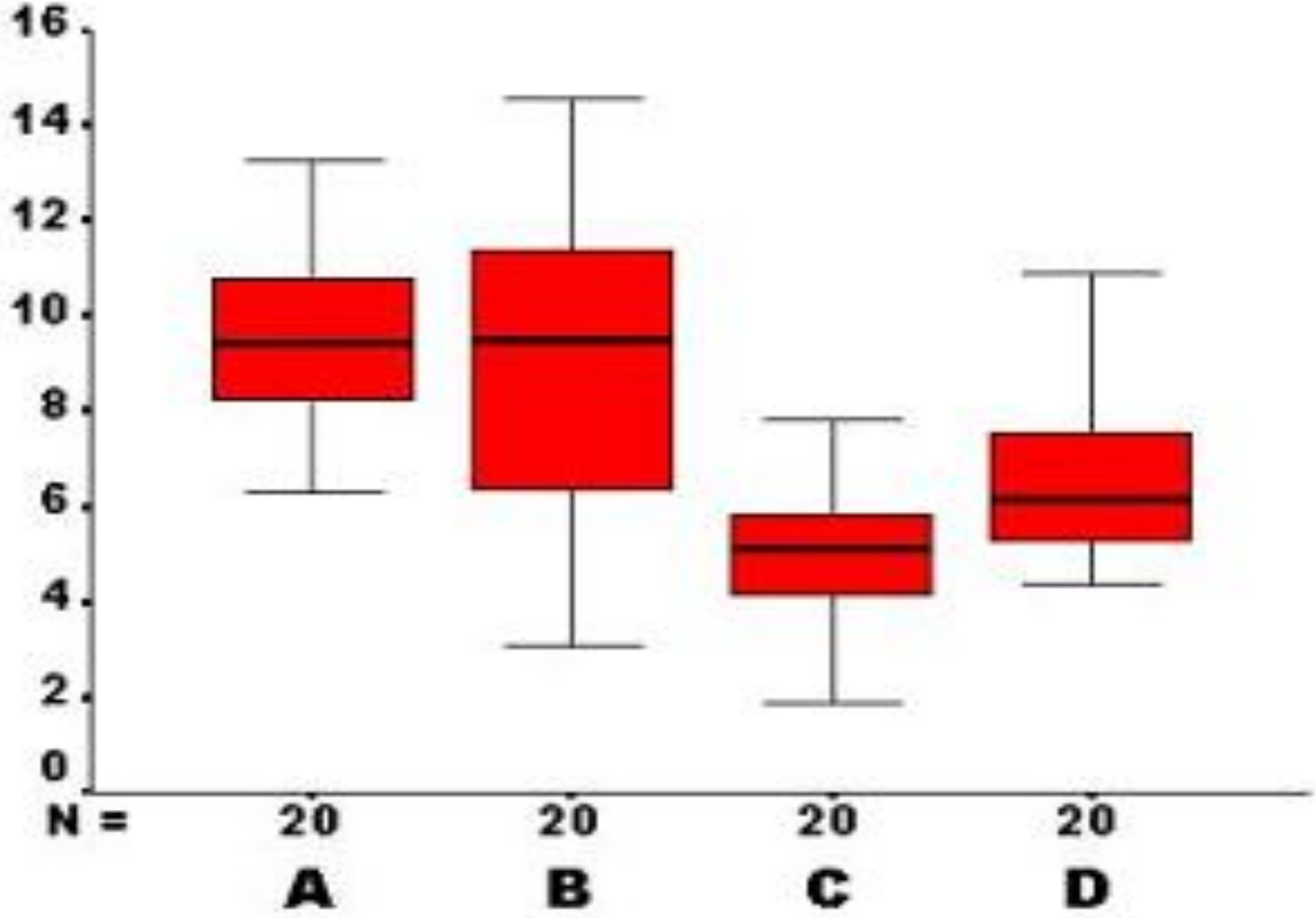


The boxplot of a sample of 20 points from a population with extremely short tails (actually a U-shaped population, with a dip in the middle rather than a hump). The whiskers are absent.


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
The middle half of the sample
One quarter of the sample
smallest unbooked sample value
largest unbooked sample value (neither outlier nor extreme)
Dots represent sample values or data points
4
6
8
10
12
14
16
18
20
whisker
whisker
○
*
yellow card zone (outlier)
red card zone (extreme)
box length
Interquartile Range
one and a half box lengths
three box lengths

Using Box Plots to Compare

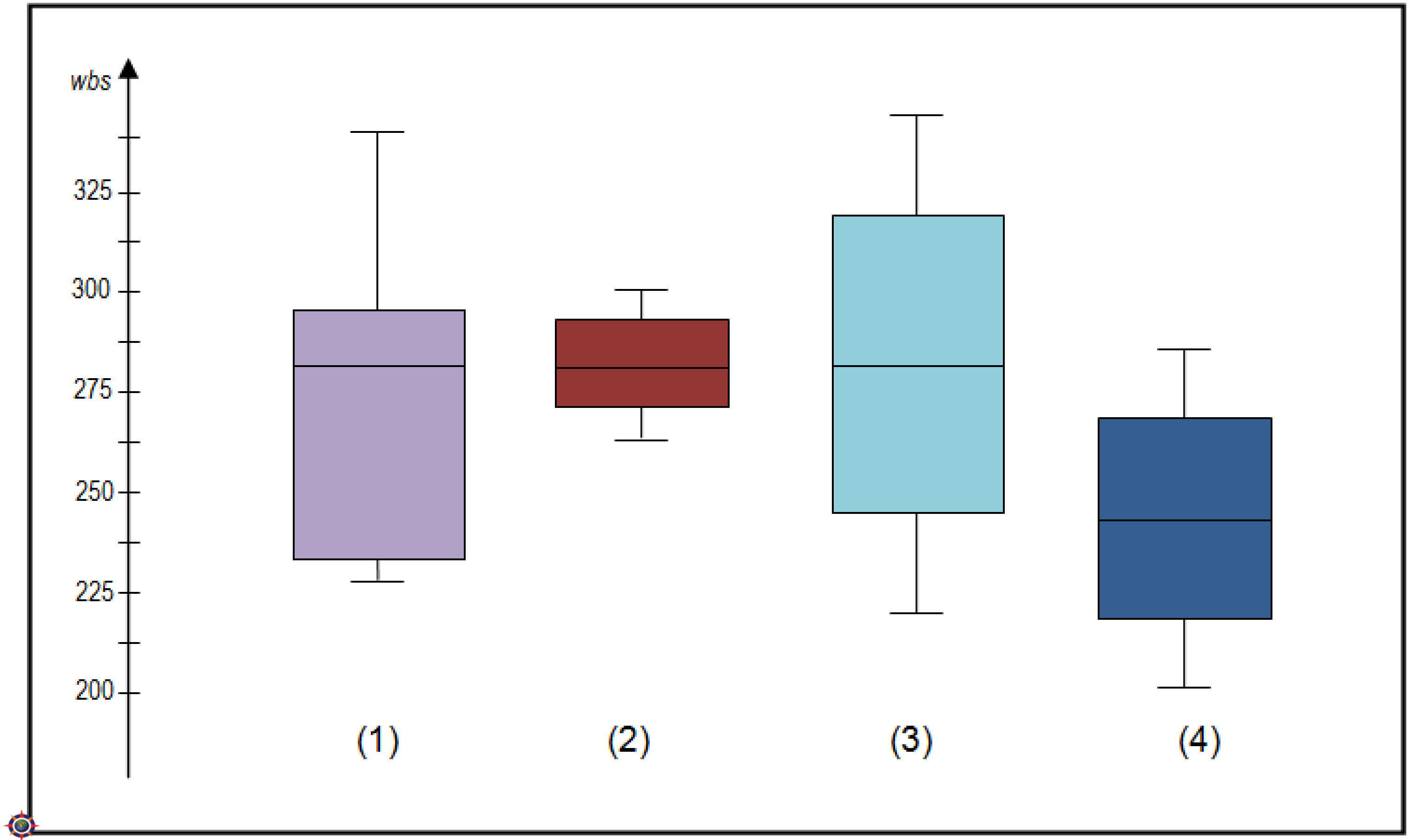



16
14
12
10
8
6
4
2
0
N =
20
20
20
20
A
B
C
D

Interpreting box plots/Box plots in general

- Box plots are used to show overall patterns of response for a group.
- They provide a useful way to visualize the range and other characteristics of responses for a large group.
- Example:
- Box plots are drawn for groups of a school scale scores. They enable us to study the distributional characteristics of a group of scores as well as the level of the scores.

- The diagram below shows a variety of **different box plot shapes and positions**.



- Some general observations about box plots
- **The box plot is comparatively short** – see example (2). This suggests that overall students have a high level of agreement with each other.

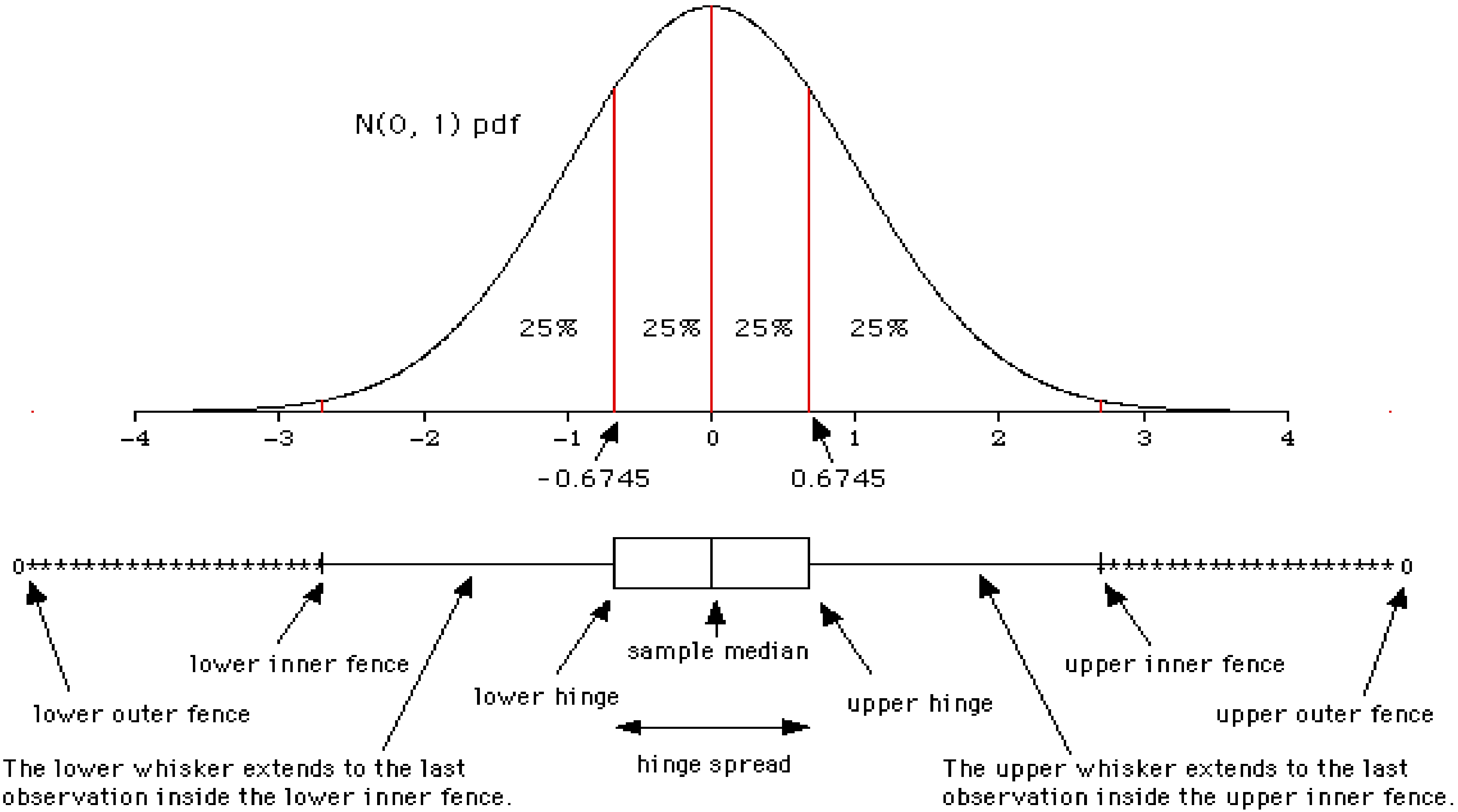
- **The box plot is comparatively tall** – see examples (1) and (3). This suggests students hold quite different opinions about this aspect or sub-aspect.

- **One box plot is much higher or lower than another** – compare (3) and (4) – This could suggest a difference between groups. For example, the box plot for boys may be lower or higher than the equivalent plot for girls.

- **Obvious differences between box plots** – see examples (1) and (2), (1) and (3), or (2) and (4). Any obvious difference between box plots for comparative groups is worthy of further investigation .

The 4 sections of the box plot are uneven in size – See example (1). This shows that many students have similar views at certain parts of the scale, but in other parts of the scale students are more variable in their views. The long upper whisker in the example means that students views are varied amongst the most positive quartile group, and very similar for the least positive quartile group.

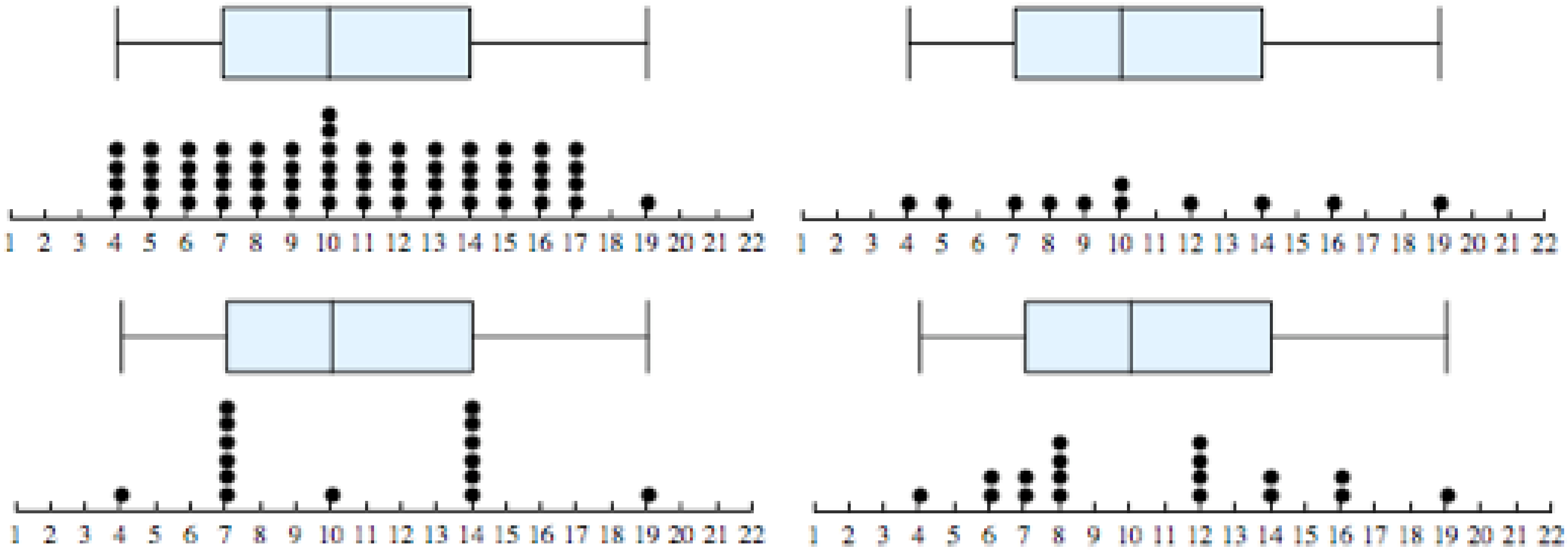
- Same median, different distribution – See examples (1), (2), and (3). The medians (which generally will be close to the average) are all at the same level. However the box plots in these examples show very different distributions of views. It always important to consider the pattern of the whole distribution of responses in a box plot.

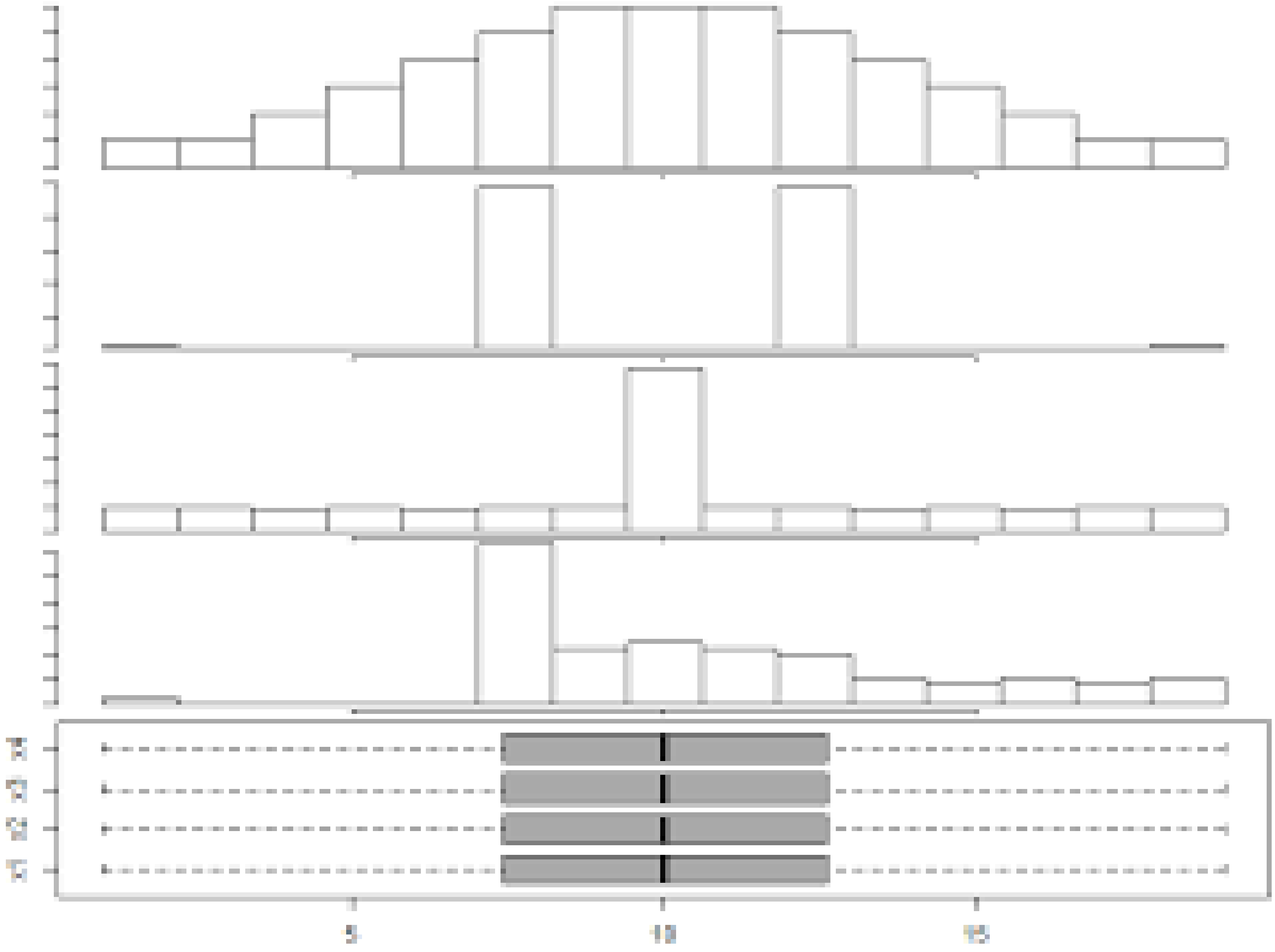


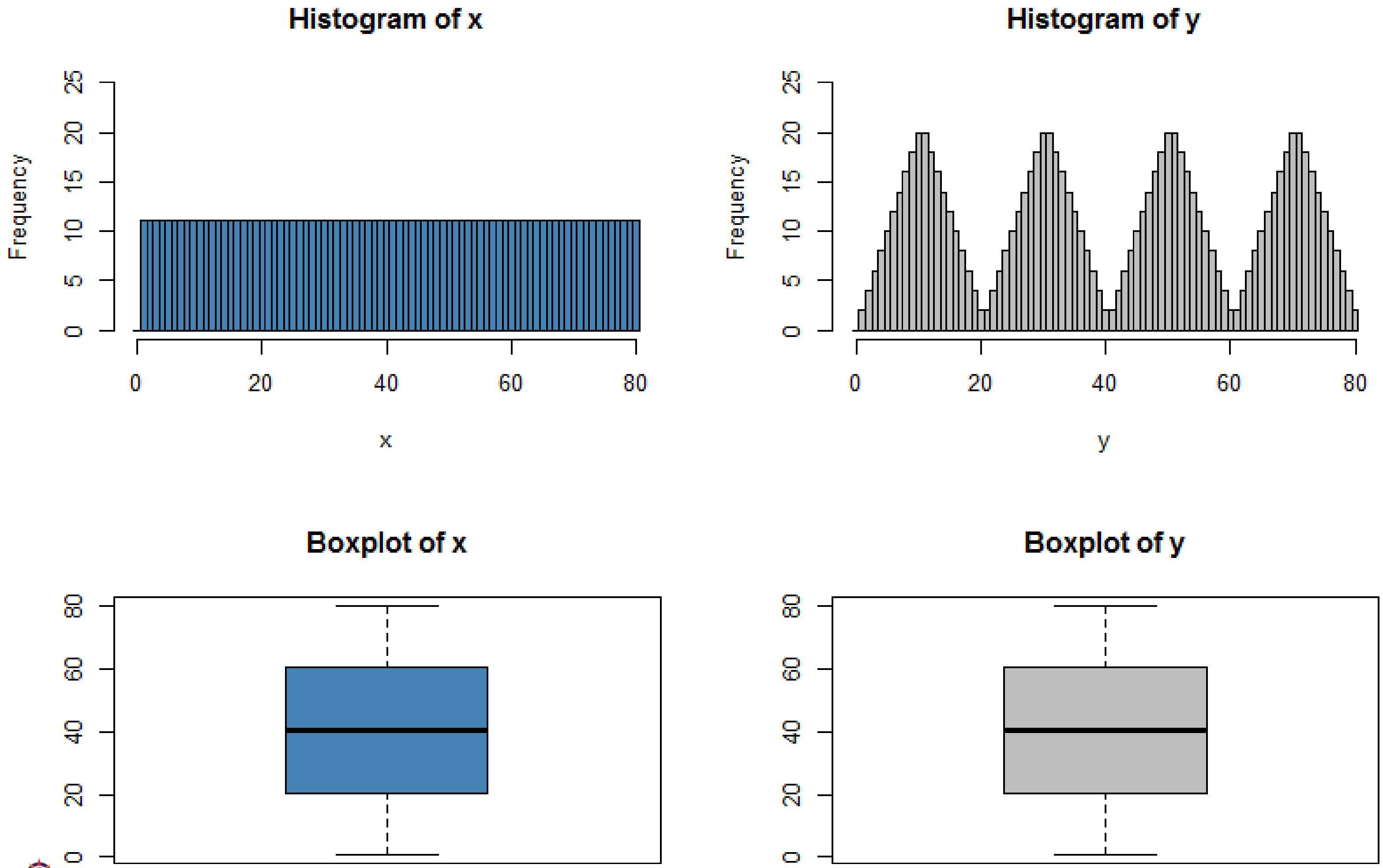
The lower whisker extends to the last observation inside the lower inner fence.

The upper whisker extends to the last observation inside the upper inner fence.

Box plot with different distributions




5
10
15

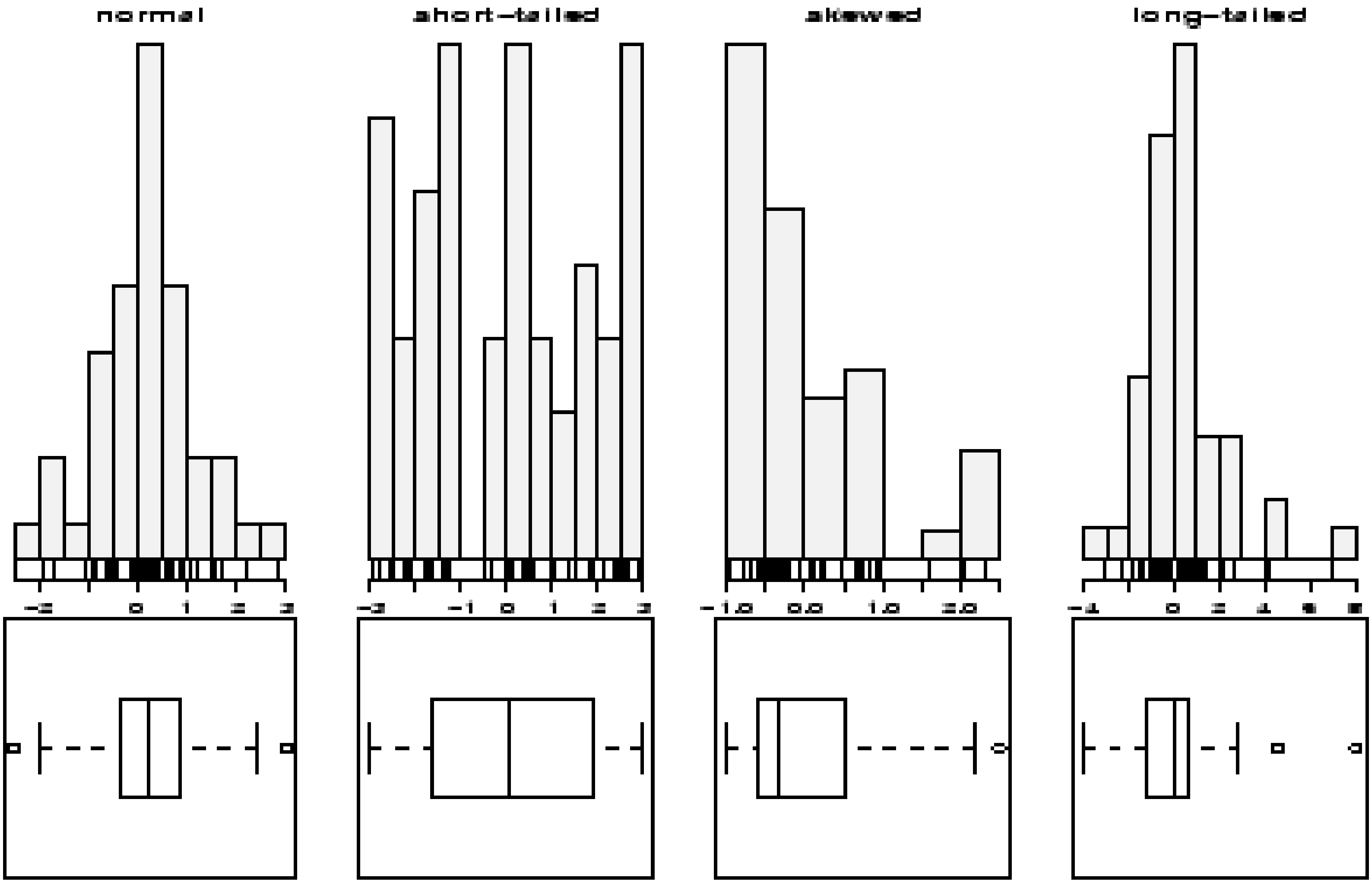


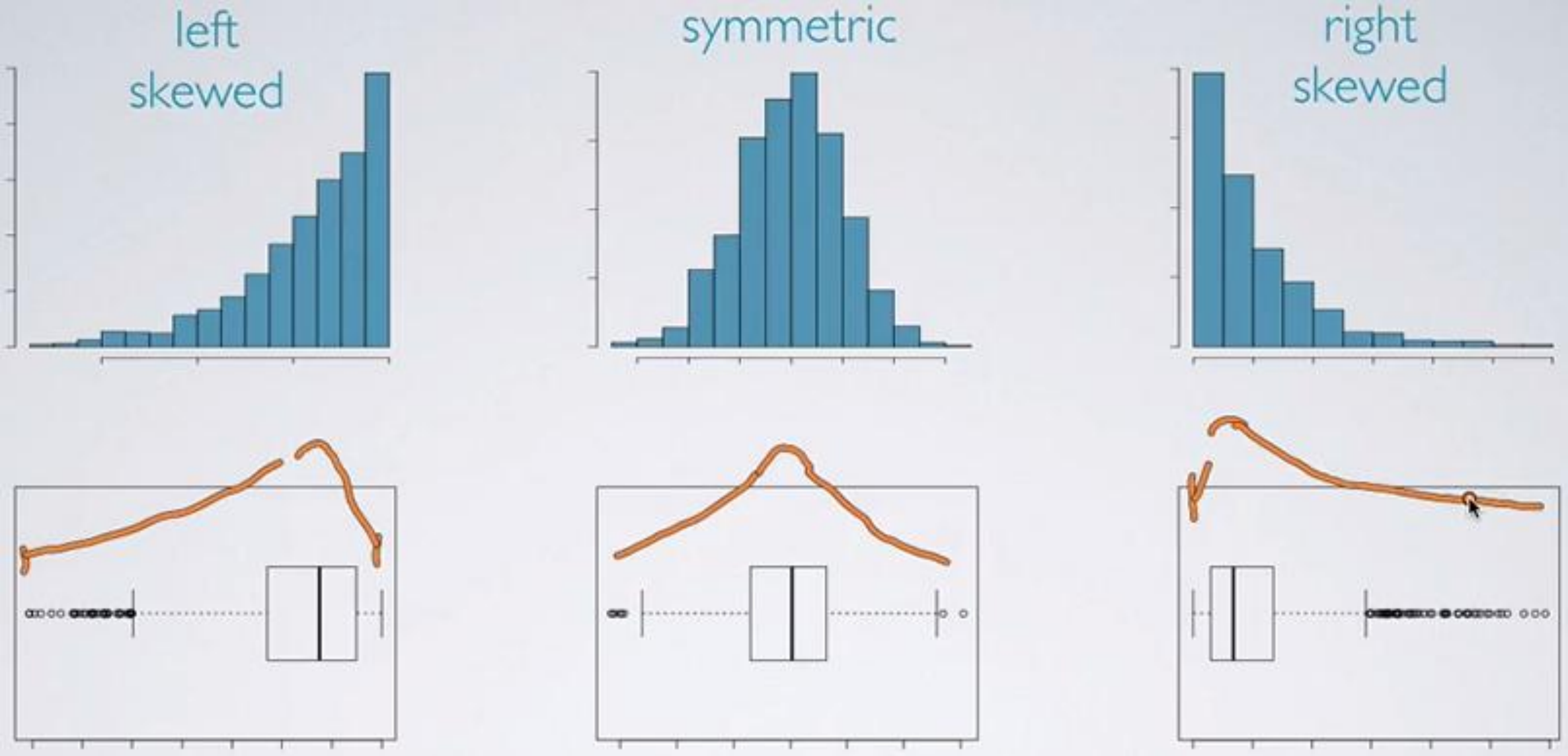
Histogram of x

Histogram of y

Boxplot of x

Boxplot of y


normal
short-tailed
skewed
long-tailed
-2
0
1
2
3
-3
-1
0
1
2
3
-1.0
0.0
1.0
2.0
-4
0
2
4
6
8


left skewed
symmetric
right skewed

When presenting or analysing measurements of a continuous variable it is sometimes helpful to group subjects into several equal groups.

For example, to create four equal groups we need the values that split the data such that 25% of the observations are in each group.

The cut off points are called quartiles, and there are three of them (the middle one also being called the median).

Likewise, we use two tertiles to split data into three groups, four quintiles to split them into five groups, and so on.

The general term for such cut off points is quantiles;

other values likely to be encountered are deciles, which split data into 10 parts,

and centiles, which split the data into 100 parts (also called percentiles).

Values such as quartiles can also be expressed as centiles; for example, the lowest quartile is also the 25th centile and the median is the 50th centile.

- A quintile is a statistical value of a data set that represents 20% of a given population, so the first quintile represents the lowest fifth of the data (1% to 20%); the second quintile represents the second fifth (21% to 40%) and so on.

Example:

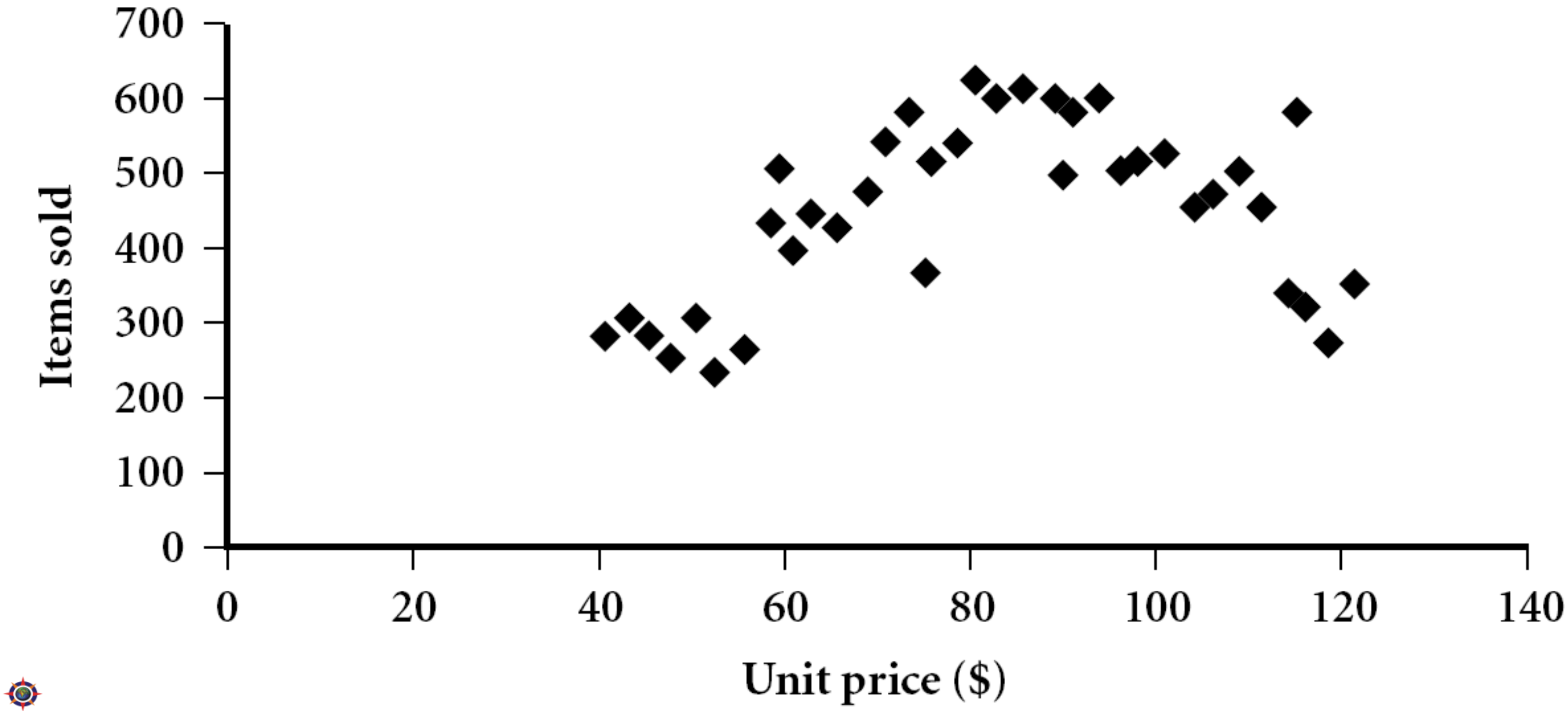
- Quintiles are used to create cut-off points for a given population; a government-sponsored socio-economic study may use quintiles to determine the maximum wealth a family could possess in order to belong to the lowest quintile of society. This cut-off point can then be used as a prerequisite for a family to receive a special government subsidy aimed to help society's less fortunate.

- The pth percentile of a sample, for a number p between 0 and 100,divides the sample so that as nearly as possible, p% of the sample values are less than the pth percentile,and (100 – p%) are greater.

- To Find Percentiles
- Order the n sample values from smallest to largest.
- Compute the quantity (p/100)(n + 1), where n is the sample size.
- If this quantity is an integer, the sample value in this position is the pth percentile.
- Otherwise, average the two sample values on either side.

Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Scatter Plots

- Summarizes the relationship between two measurement variables.
- Horizontal axis represents one variable and vertical axis represents second variable.
- Plot one point for each pair of measurements.

Scatter Plot

- One of the variables(mostly independent) is plotted on the x-axis, while the other(mostly independent) is on the y-axis, and the data points are plotted on the graph.
- Scatter plots are used when you want to show the relationship between two numeric variables in the data.
- Scatter plots are sometimes called correlation plots because they show how two variables are correlated.
- Other situation to use scatter plot, when one continuous variable that is under the control of the experimenter and the other depends on it.

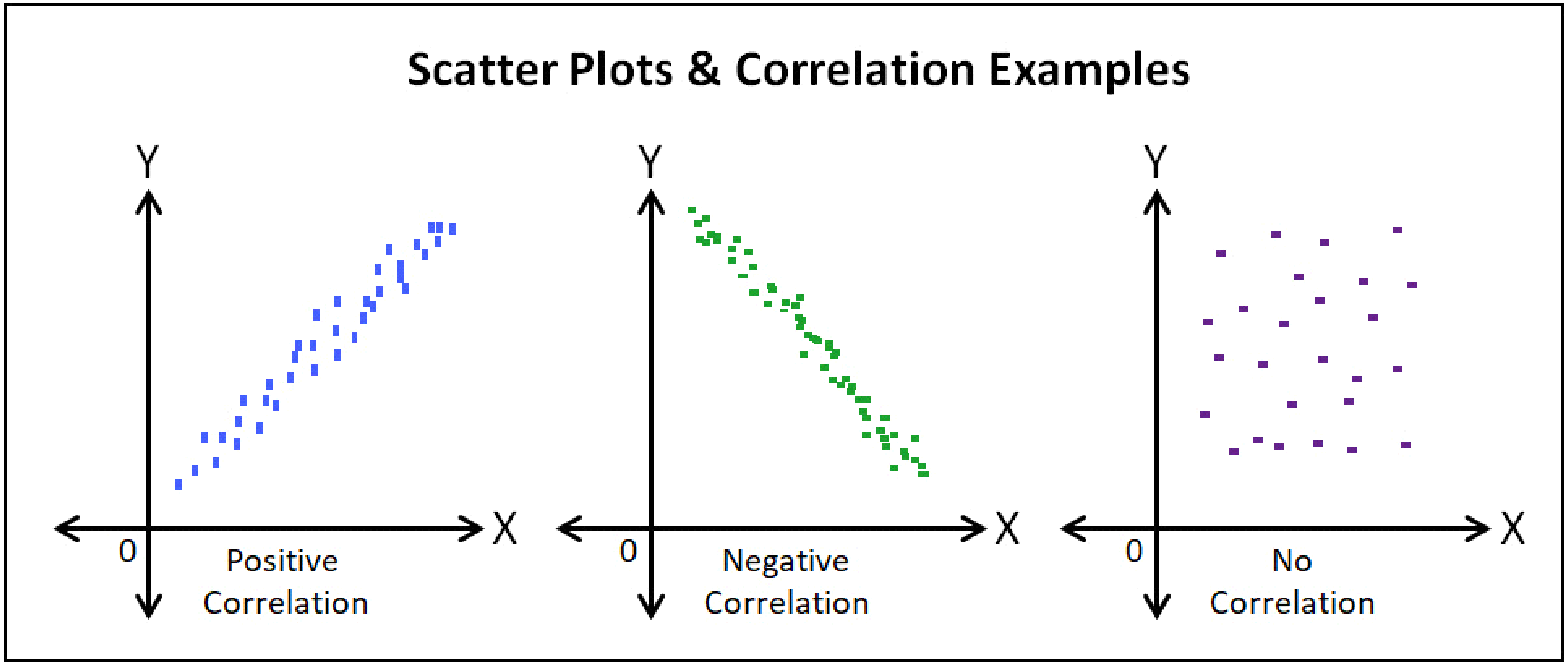
- It is used to find the correlation between the variables plotted, the more the data is clustered along one line the more the variables are correlated, there exists no correlation between the data which is randomly spread across the graph.
- It becomes really difficult to visualize the data which is clustered at one place in the scatter plot, which is when density plots are used rather than scatter plot.
- It can be disadvantageous when trying to find the relationship between more than two variables, we'll need to use multiple scatter plots for such situations.

- There are many types of coefficients of correlation in scatter points, most popular one is

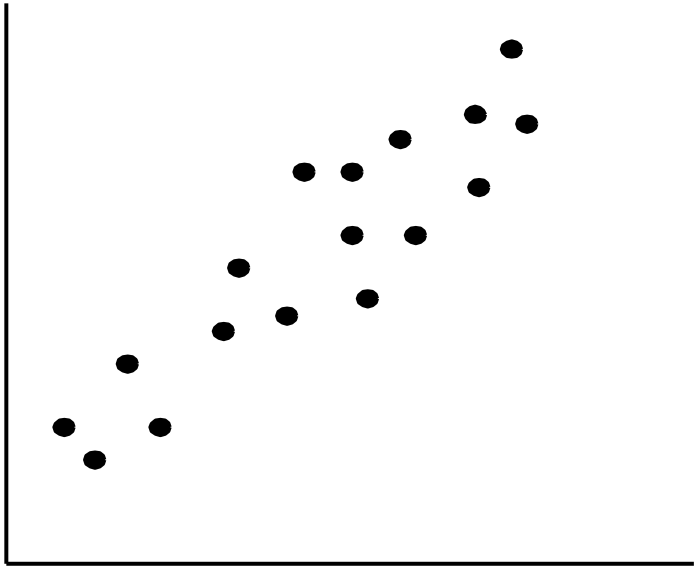
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

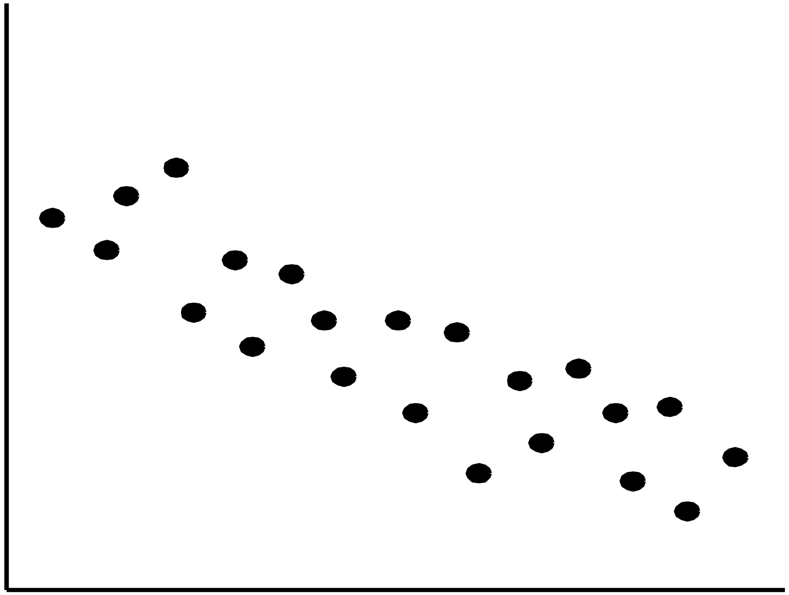
- Pearson’s Co-efficient of correlation
- x – value of data point on x-axis
- y – value of data point on y-axis
- n – no of datapoints

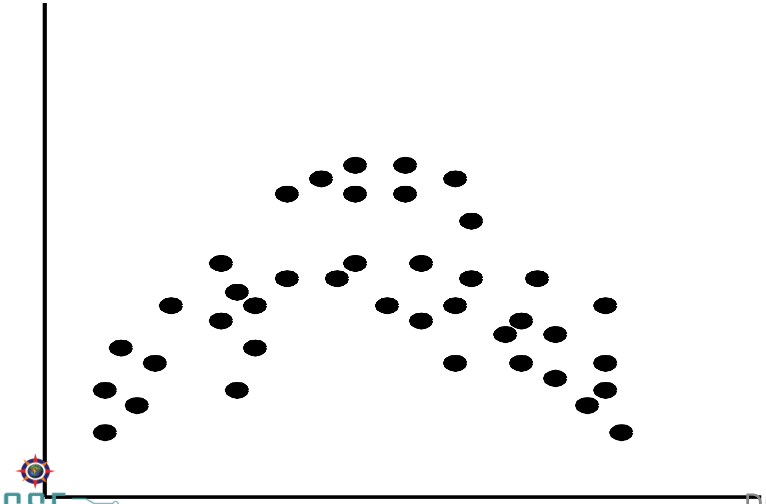
- Pearson’s co-efficient of correlation:
- co-eff > 0 : positively correlated
- co-eff < 0 : negatively correlated
- co-eff = 0 : no correlation
- +1 or -1, mean perfect correlation between the data points


Scatter Plots & Correlation Examples
Y
X
0
Positive Correlation
Y
X
0
Negative Correlation
Y
X
0
No Correlation

Positively and Negatively Correlated Data

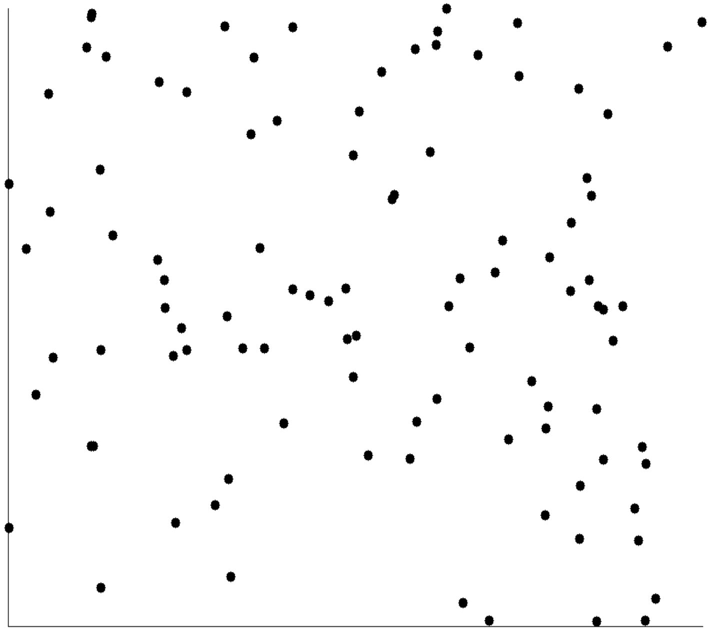


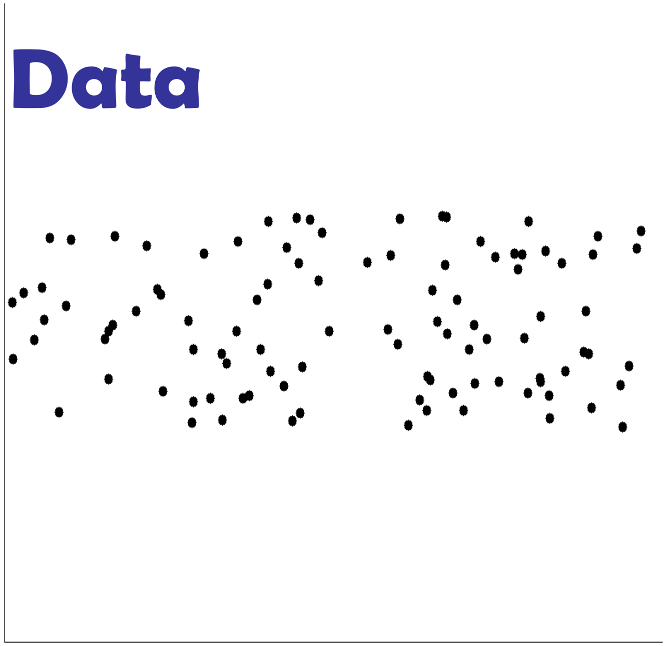


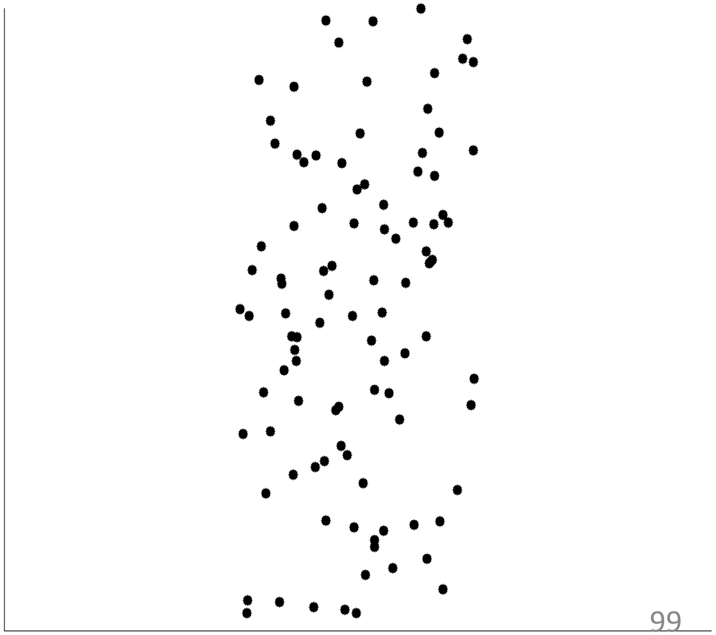


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data

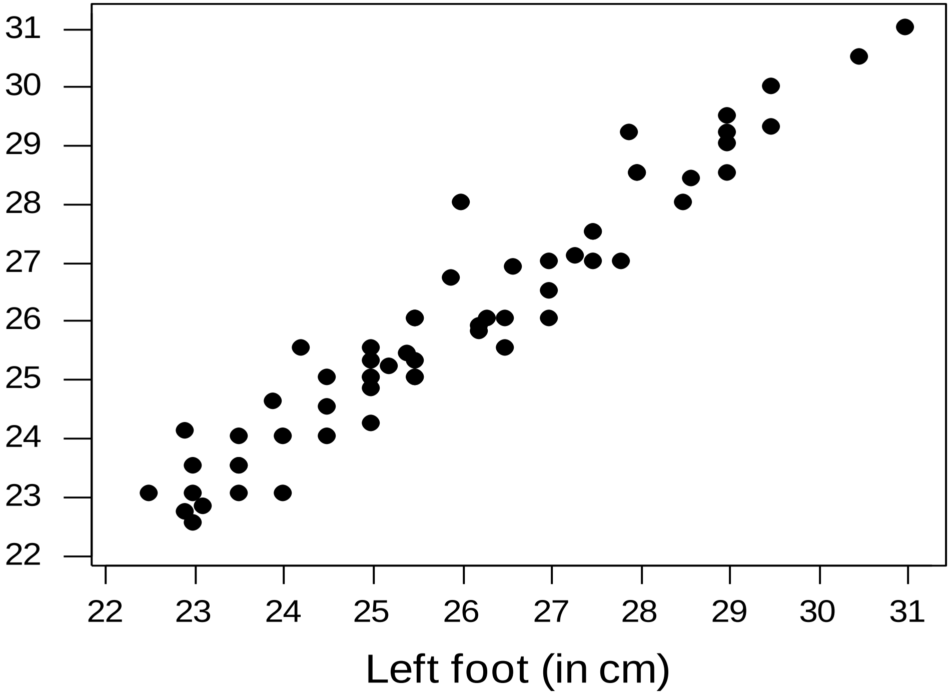






Scatter Plots

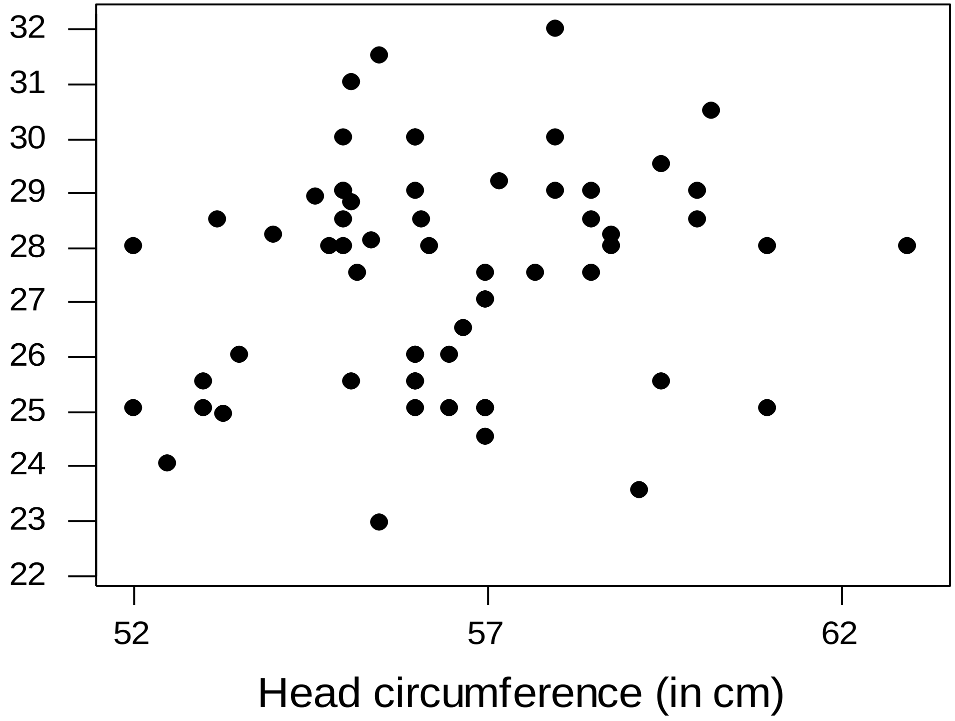
Foot sizes of Spring 1998 Stat 250 students



n=88 students

No relationship

Lengths of left forearms and head circumferences of Spring 1998 Stat 250 Students

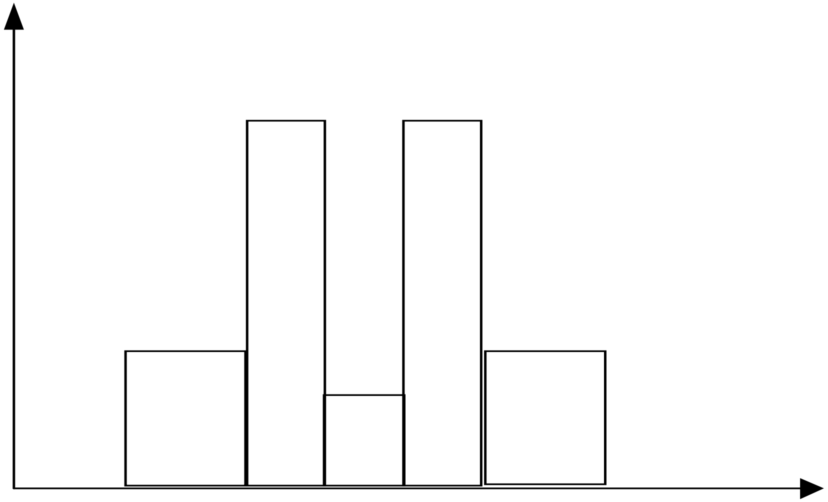


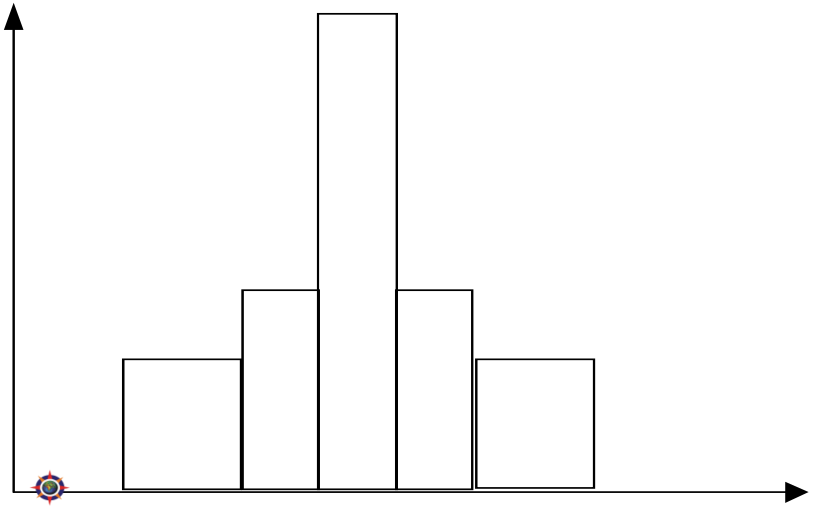
n=89 students

Which graph to use when?

- dotplots are good for small data sets, while histograms and box plots are good for large data sets.
- Boxplots and dotplots are good for comparing two groups.
- Boxplots are good for identifying outliers.
- Histograms and boxplots are good for identifying "**shape**" of data.

Histograms Often Tell More than Boxplots





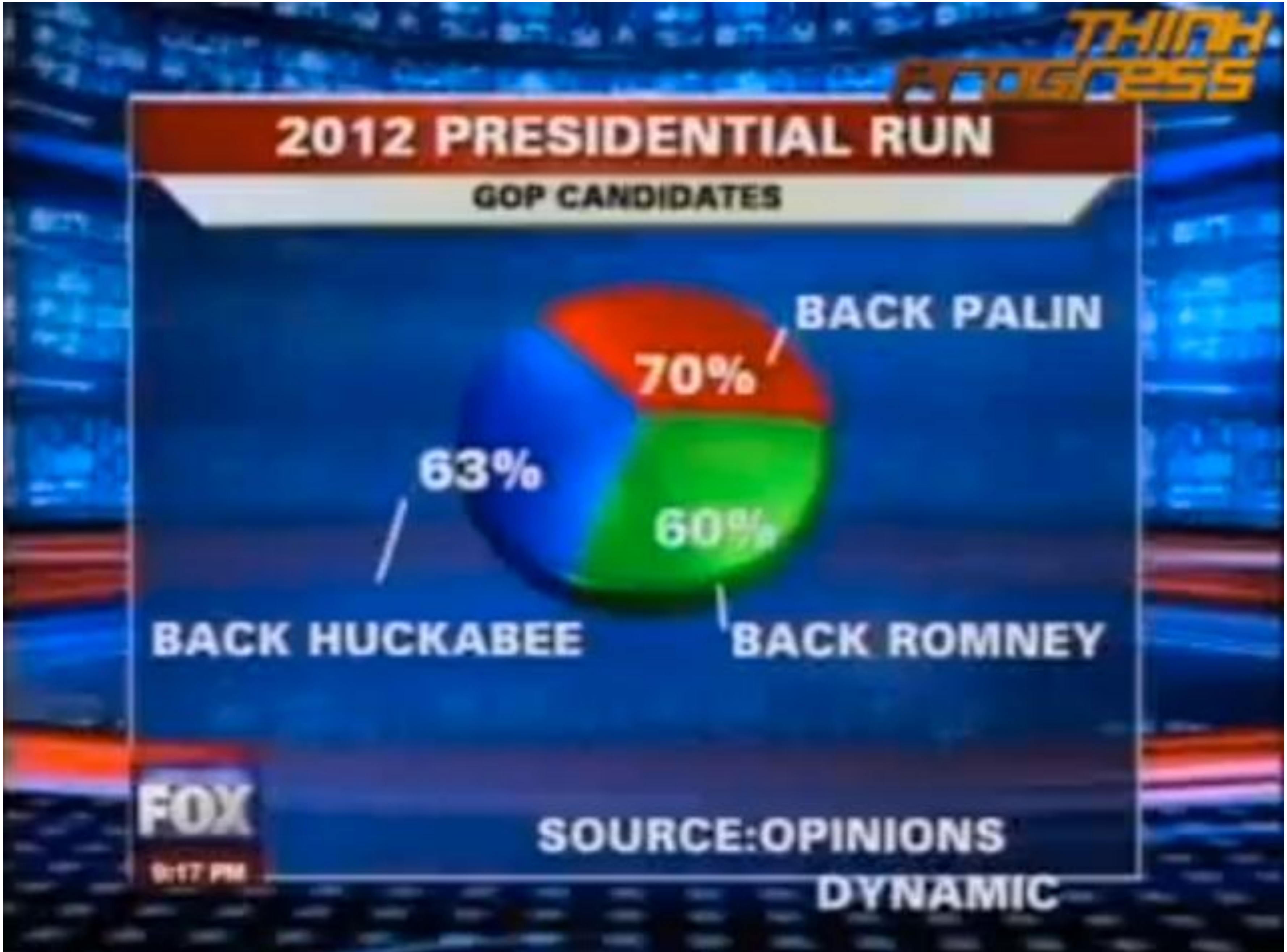
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Good Vs Bad Visualization

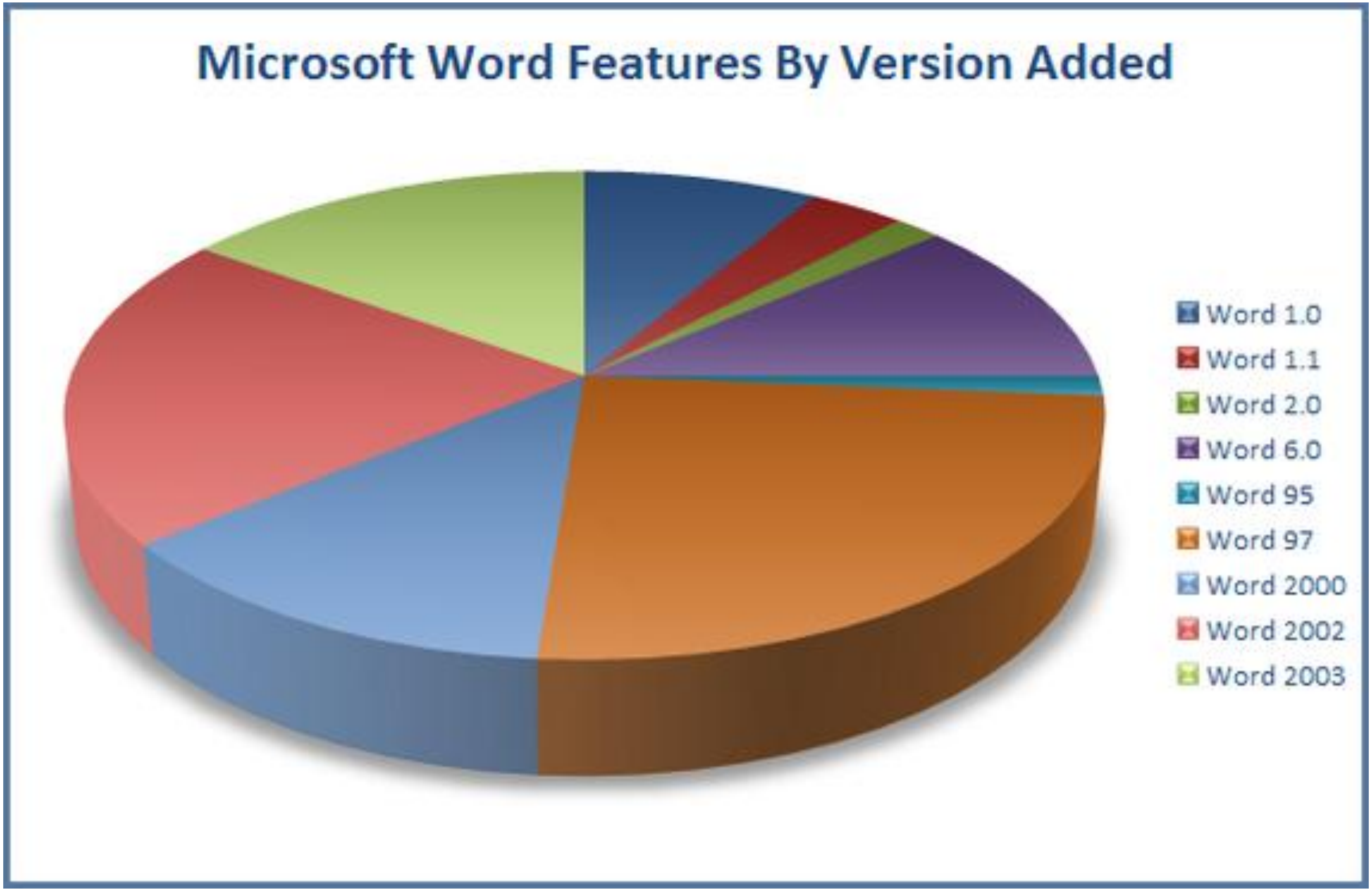
5 common mistakes that lead to bad data visualization

- Bad Data
- Wrong Choice of Data Visualization
- Too Much Color or Information
- Misrepresentation of Data
- Inconsistent Scales

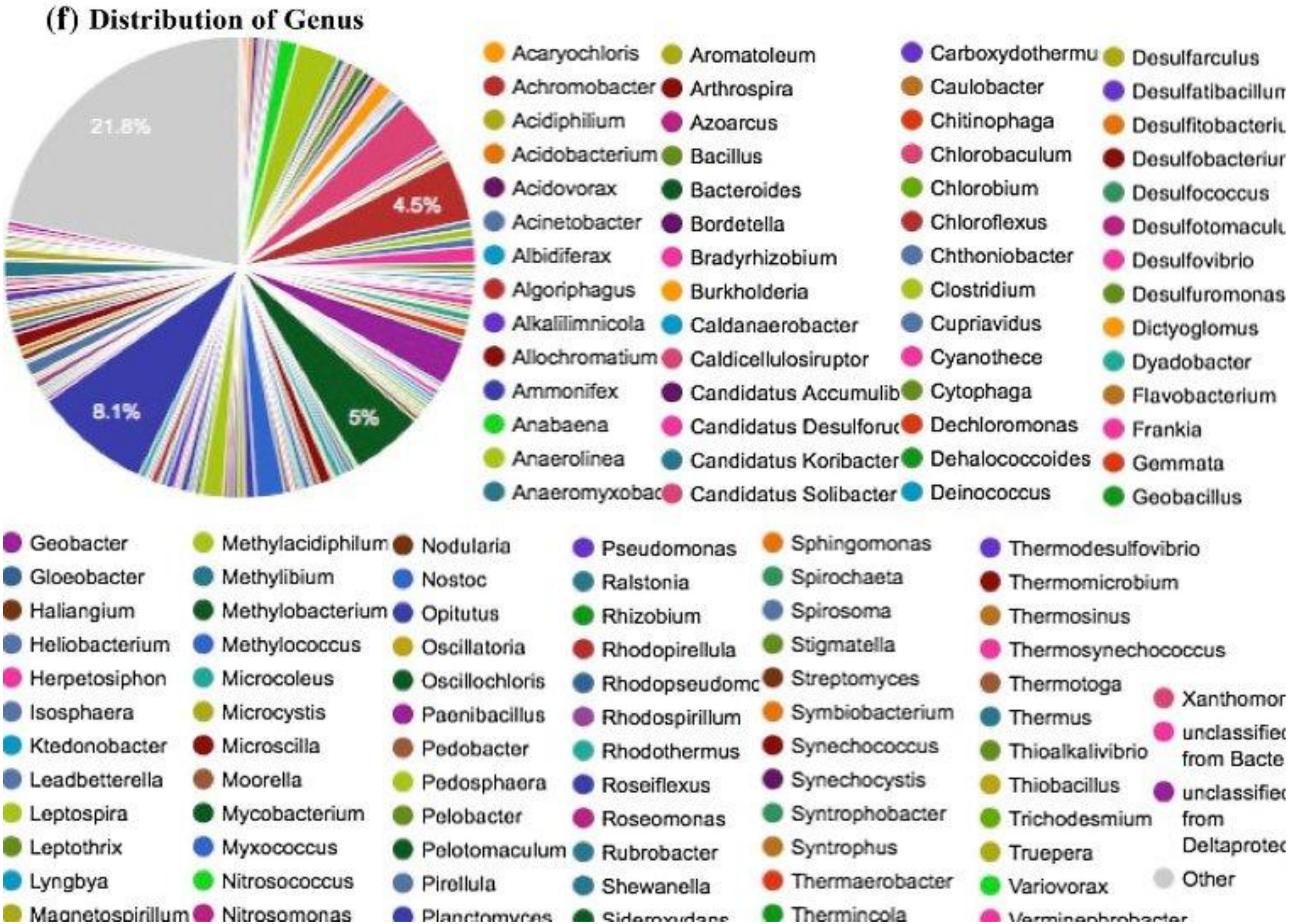
Bad data


THINK PROGRESS
2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
9:17 PM
SOURCE:OPINIONS
DYNAMIC

Wrong Choice of Data Visualization

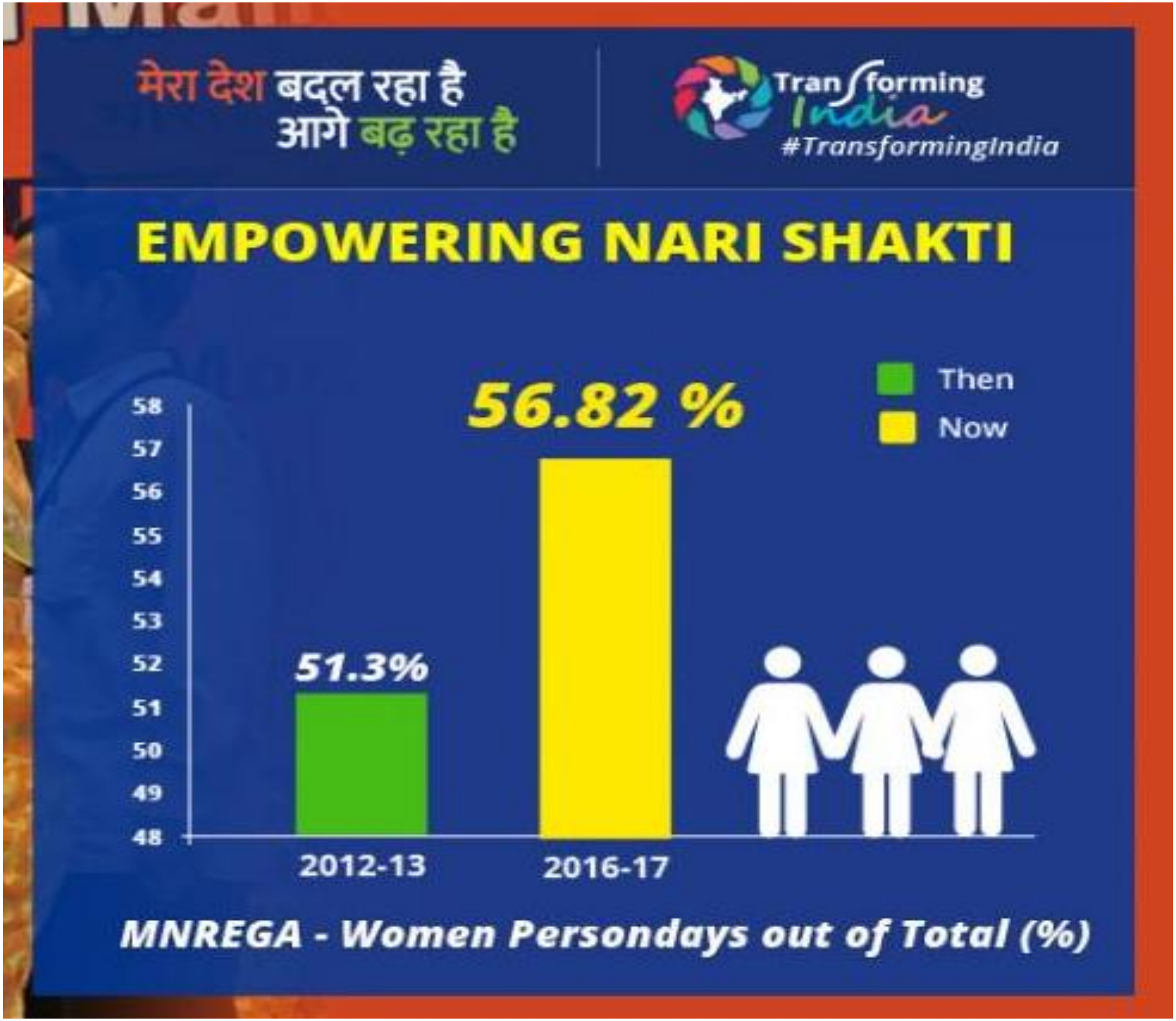


Too Much Color or Information

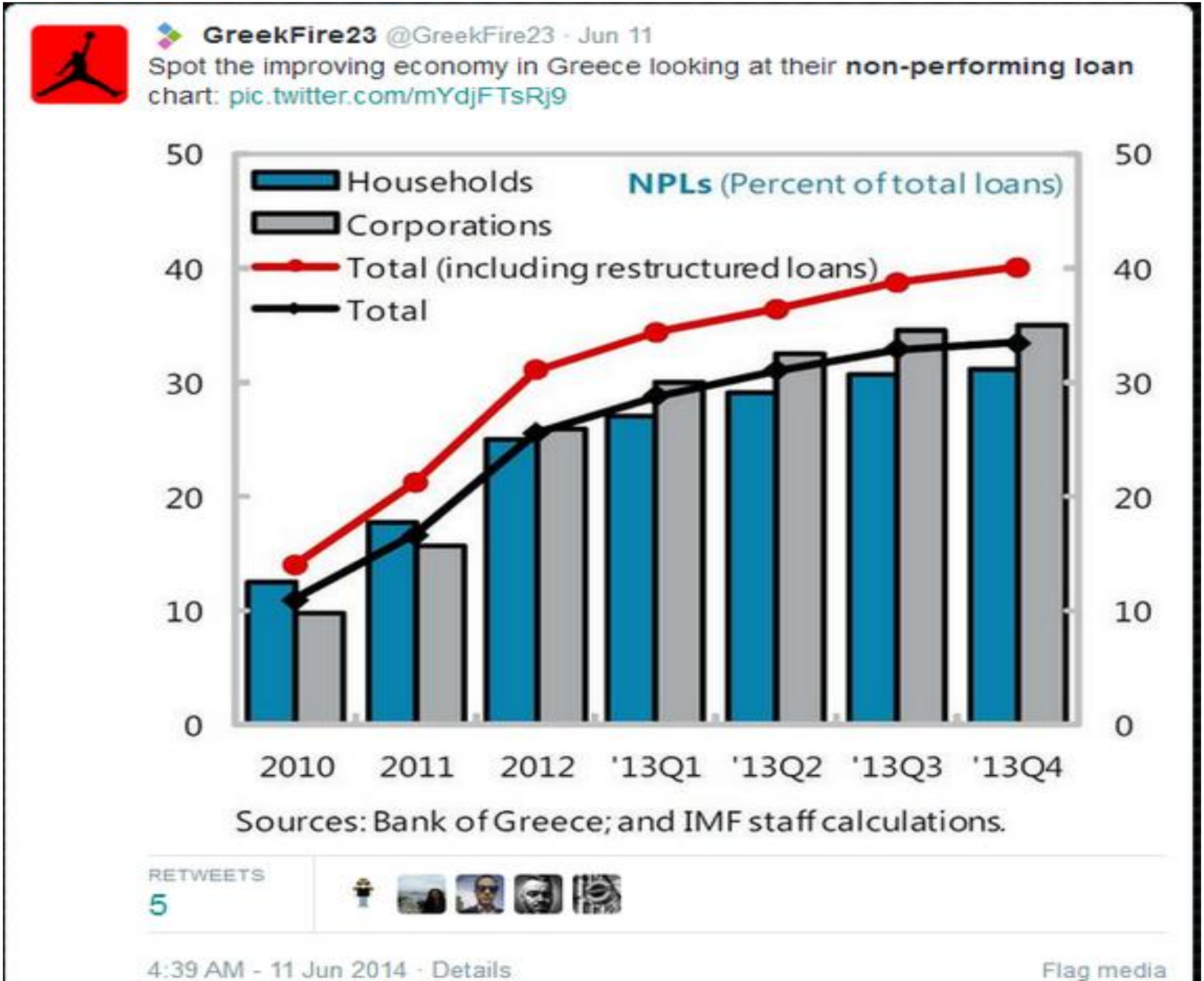


(f) Distribution of Genus

Misrepresentation of Data



Inconsistent Scales



GreekFire23 @GreekFire23 · Jun 11

Spot the improving economy in Greece looking at their **non-performing loan** chart: pic.twitter.com/mYdjFTsRj9

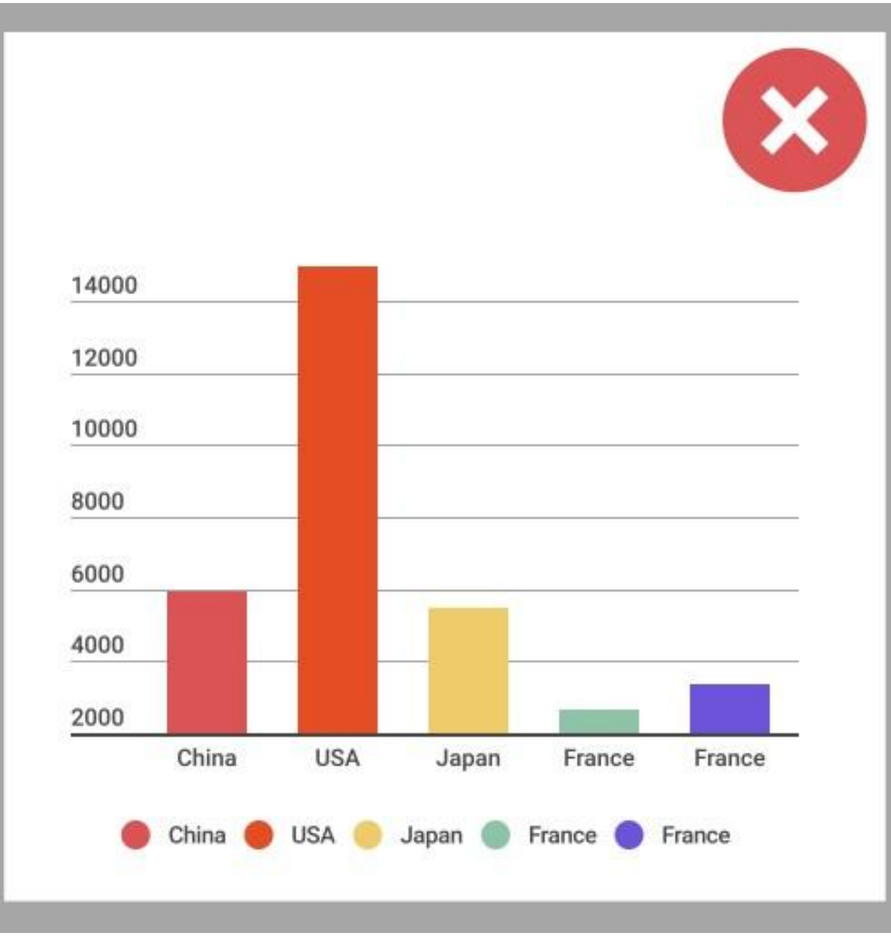
Sources: Bank of Greece; and IMF staff calculations.

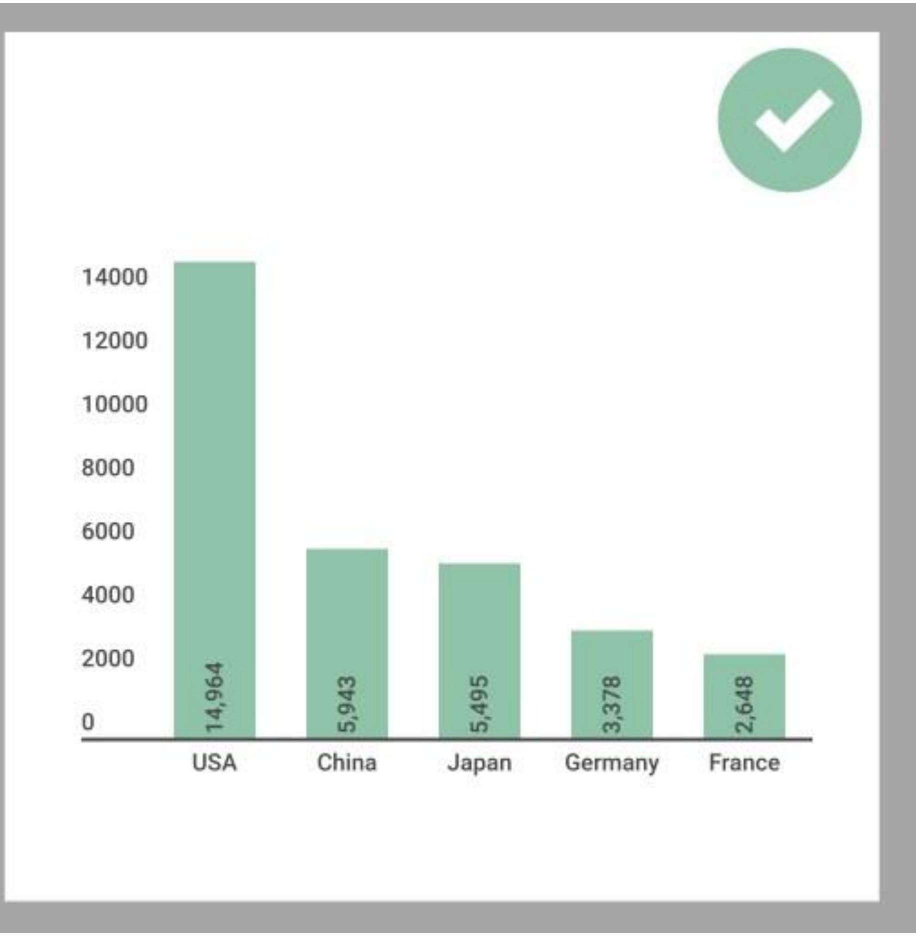
RETWEETS
5

4:39 AM - 11 Jun 2014 · Details

Flag media

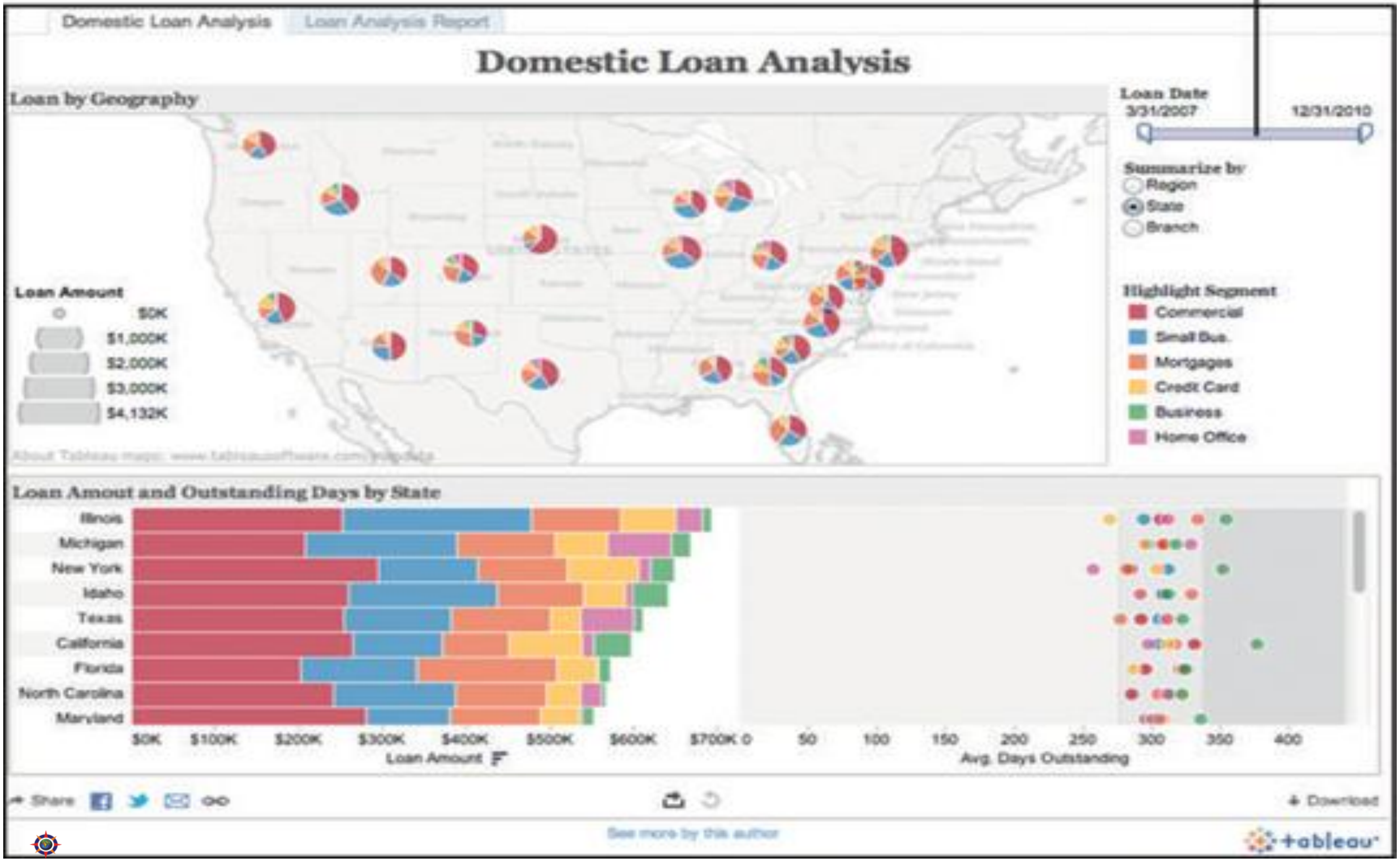
Good Vs Bad Visualization





shows a dashboard that analyzes the status of domestic loans in the United States.

Slider for interactivity



Domestic Loan Analysis | Loan Analysis Report

Domestic Loan Analysis

Loan by Geography

Loan Amount

○	$0K
▭	$1,000K
▭	$2,000K
▭	$3,000K
▭	$4,132K

About Tableau maps: www.tableausoftware.com/mapdata

Loan Date

3/31/2007 — 12/31/2010

Summarize by

○ Region
● State
○ Branch

Highlight Segment

- 🟥 Commercial
- 🟦 Small Bus.
- 🟧 Mortgages
- 🟨 Credit Card
- 🟩 Business
- 🟪 Home Office

Loan Amount and Outstanding Days by State

Things that work well:

Color consistency: One thing that’s evident throughout this visualization is the consistency of the colors in the dashboard.

On the right is a single legend — Highlight Segment — that shows the legend for the color, which remains in both the bar chart at the bottom and the pie charts on the map.

Simplicity: The two large charts make digesting the data easy.

Interactivity: The slider in the top-right corner controls the time period displayed on the charts. That interactive feature put users in control of what they view.

Things that don‘t work:

Chart choice: The small pie charts that are overlaid on the map are of little value. They’re hard to view, and without clicking every single one, the user can’t determine which of them are worth evaluating. It’s also virtually impossible to tell which states or regions the charts pertain to.

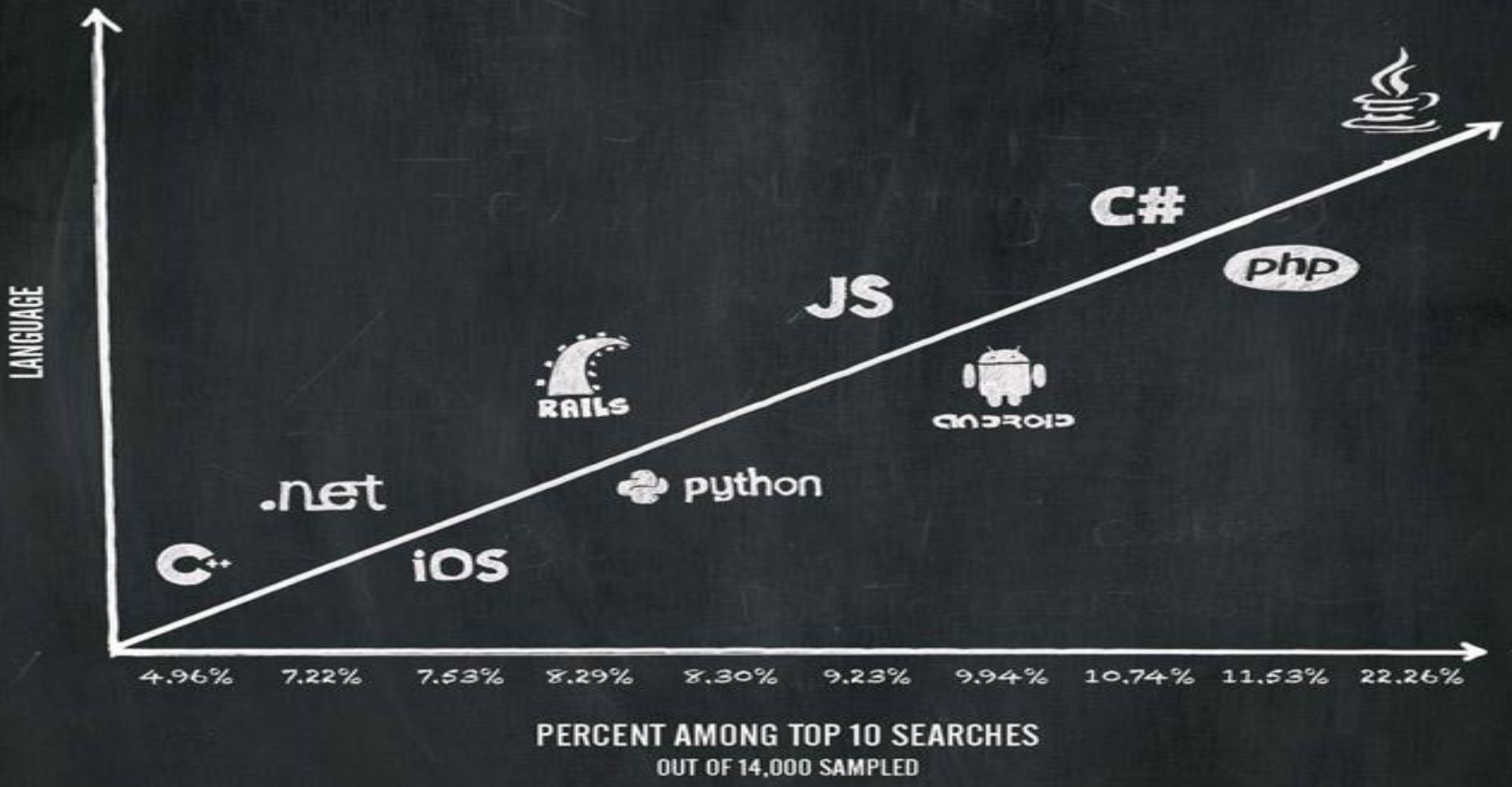
Color choice: The abundant use of red, blue, and orange are misleading, especially in the stacked bar chart at the bottom of the data viz. At a glance, users may think that the colors could mean good versus bad; in fact, they’re just associated with a specific segment. This type of color usage harkens back to the recommendation about being careful with the use of RAG colors. An alternative is to use more muted colors, such as a range of grays and blues.

Data overload: There’s a lot of data on the screen, but none of it really identifies the most important data or trends that users need to pay attention to. This visualization displays data for viewing instead of adding real value.

readwrite presents

TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013

INFORMATION COMPILED BY STACK OVERFLOW



readwrite.com

Using discrete values in the X-Axis for a continuous measurement (i.e. the percentage). And not only that, discrete values with two significant figures, which make the X-Axis unusually cluttered.

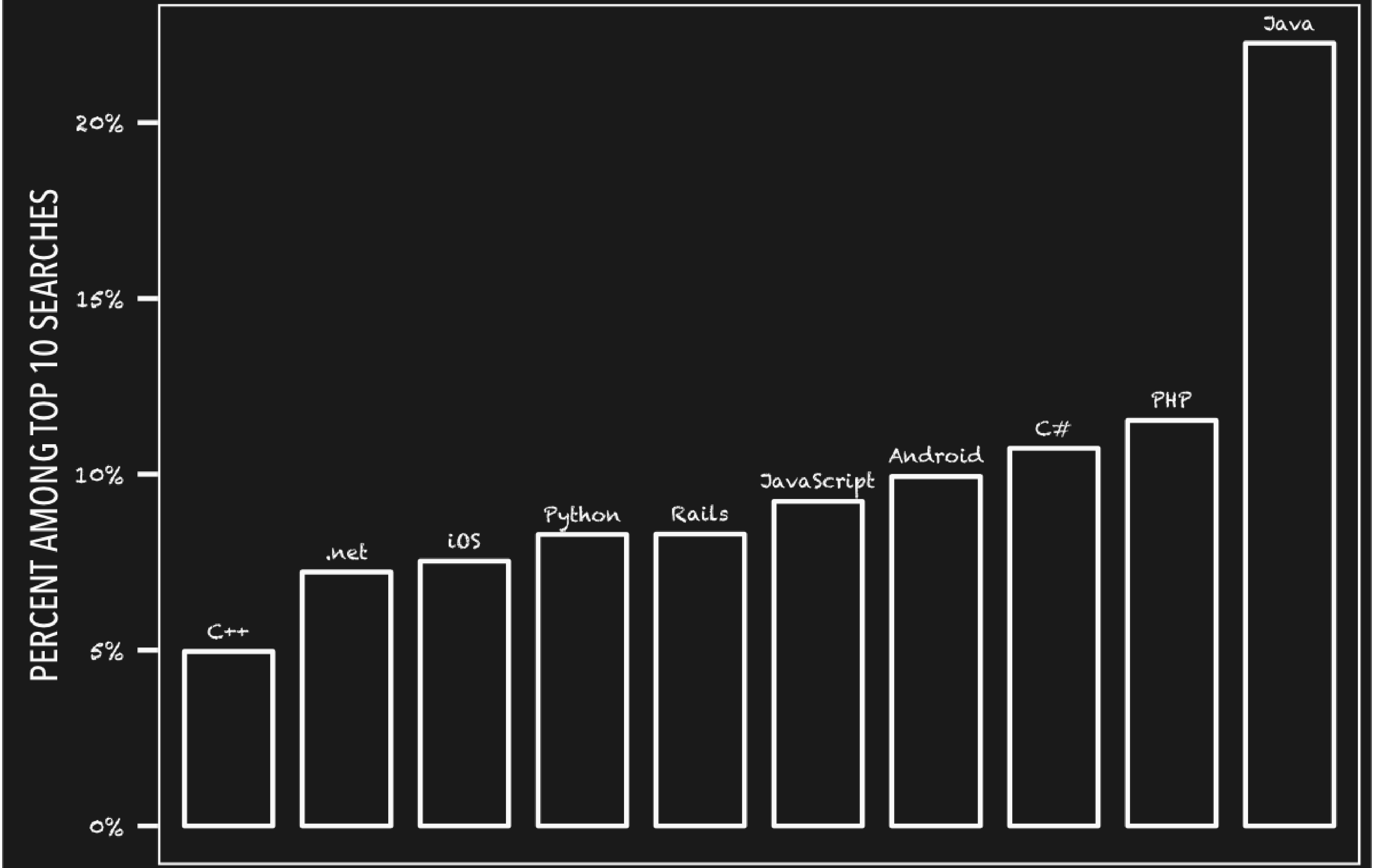
The Y-Axis is Language. This implies that some programming languages are more language than others. (to be fair, Java is more language than Android)

Not all entries on the chart are programming languages. (Android, for example, is an operating system.)

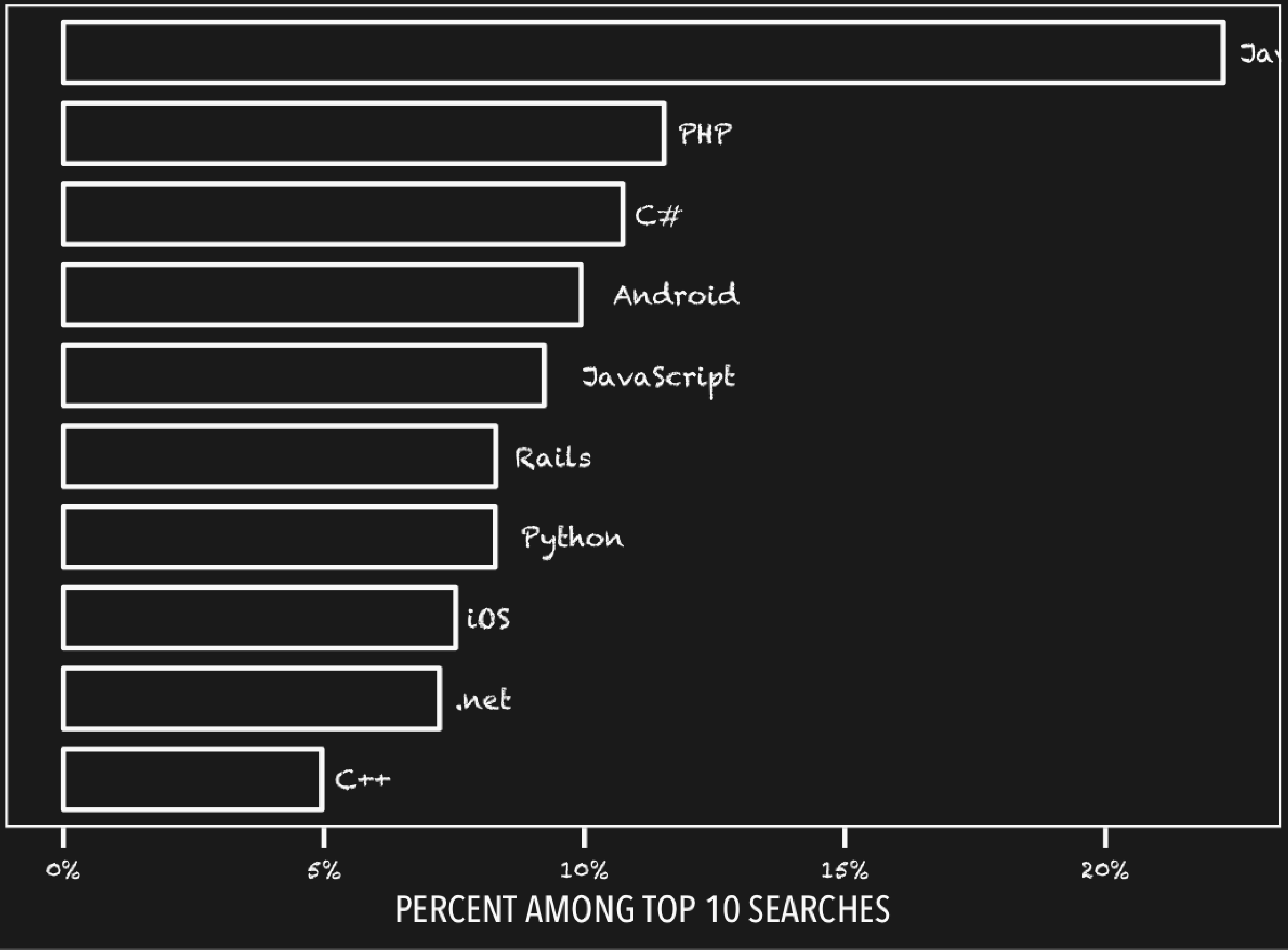
The 45-degree line in the chart implies that the relationship between language and %-of-searches is perfectly linear, where in reality the data has an upward-parabolic shape.

- No relative proportions between the programming languages. We can’t accurately see the increase in language Java has relative to Android just by looking at the graph.
- Cannot easily associate a language with the given X-Axis value. The logos representing the programming language oscillate around the line, and it’s hard to see at a glance which percentage corresponds to which language.

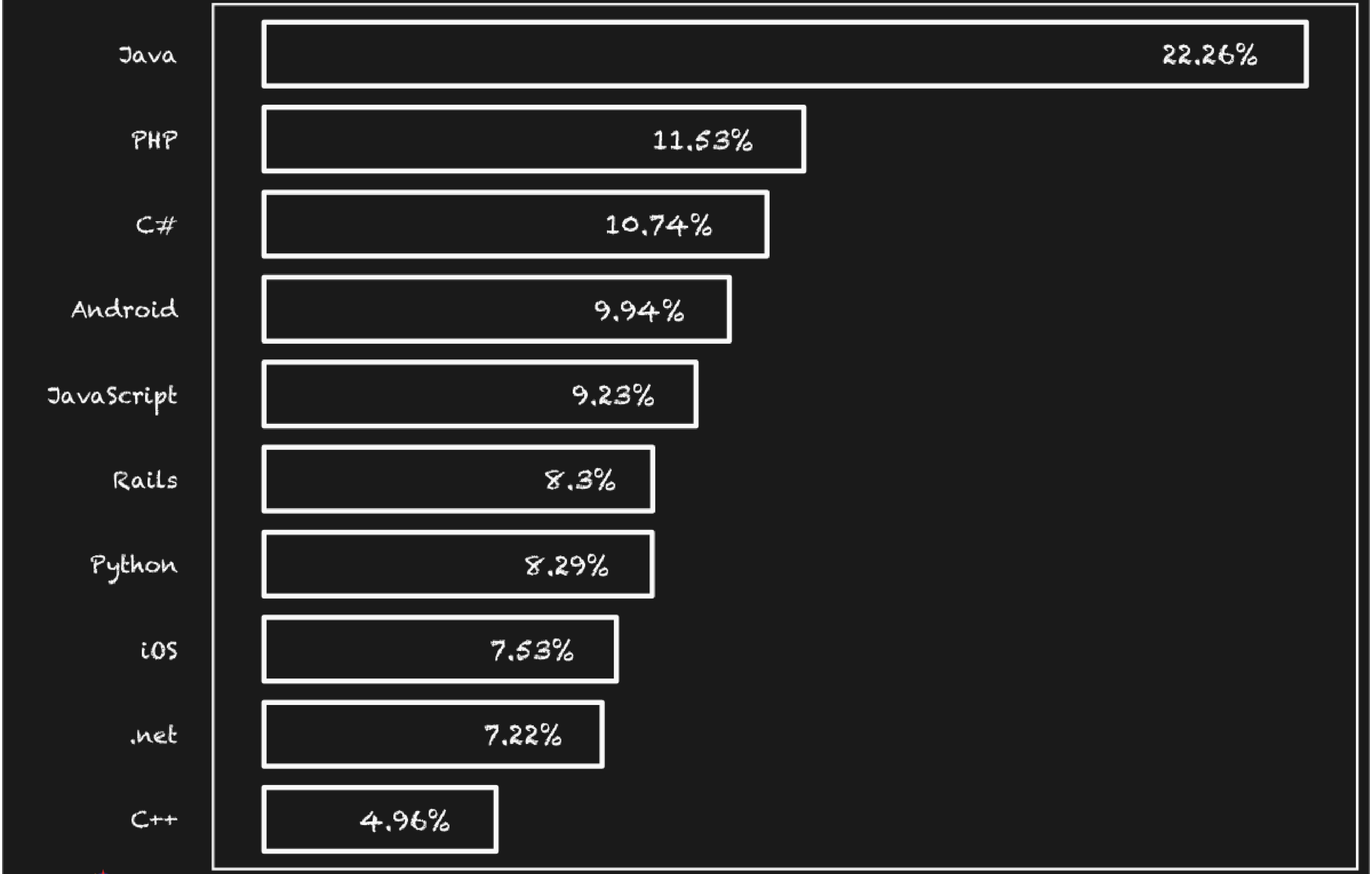
TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013

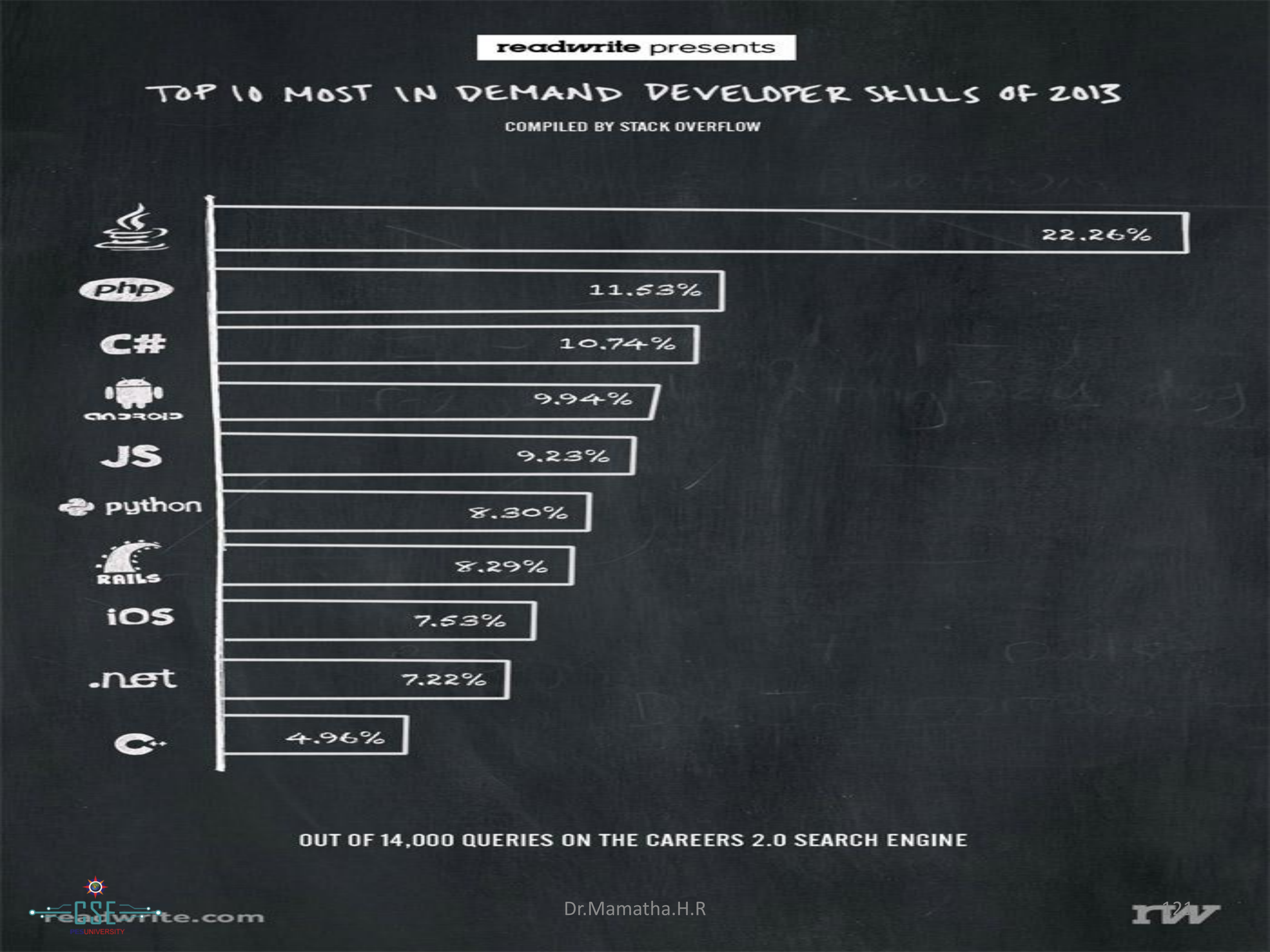


TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013



TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013





readwrite presents

TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013

COMPILED BY STACK OVERFLOW

OUT OF 14,000 QUERIES ON THE CAREERS 2.0 SEARCH ENGINE

readwrite.com

rw

Closing comments

- Many possible types of graphs.
- Use common sense in reading graphs.
- When creating graphs, don’t summarize your data too much or too little.
- When creating graphs, label everything for others. Remember you are trying to communicate something to others!

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.