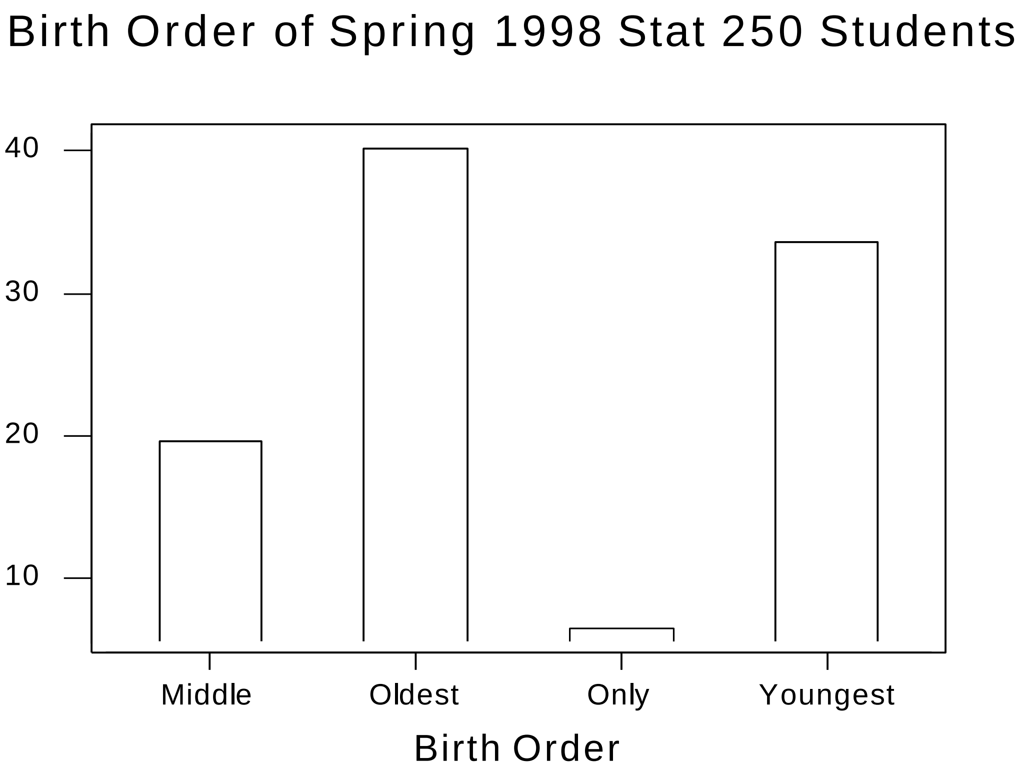
Data Visualization, Interpretation and Good vs. Bad Visualization

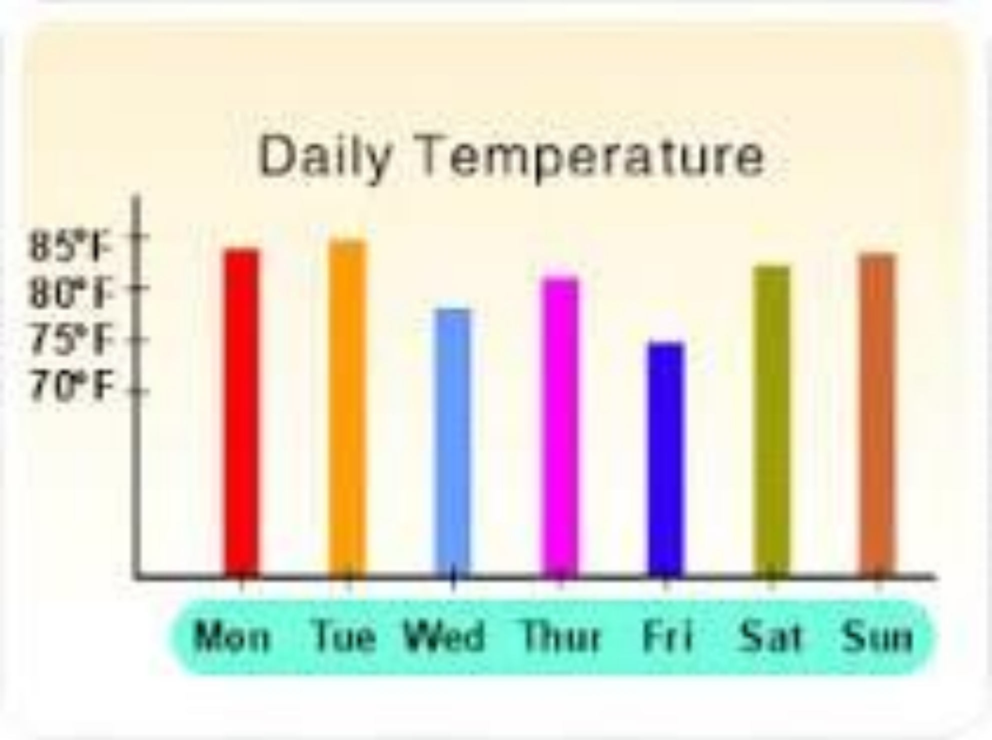
Bar Chart



n=92 students

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.



- <u>Horizontal BarGraphs</u> : 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.

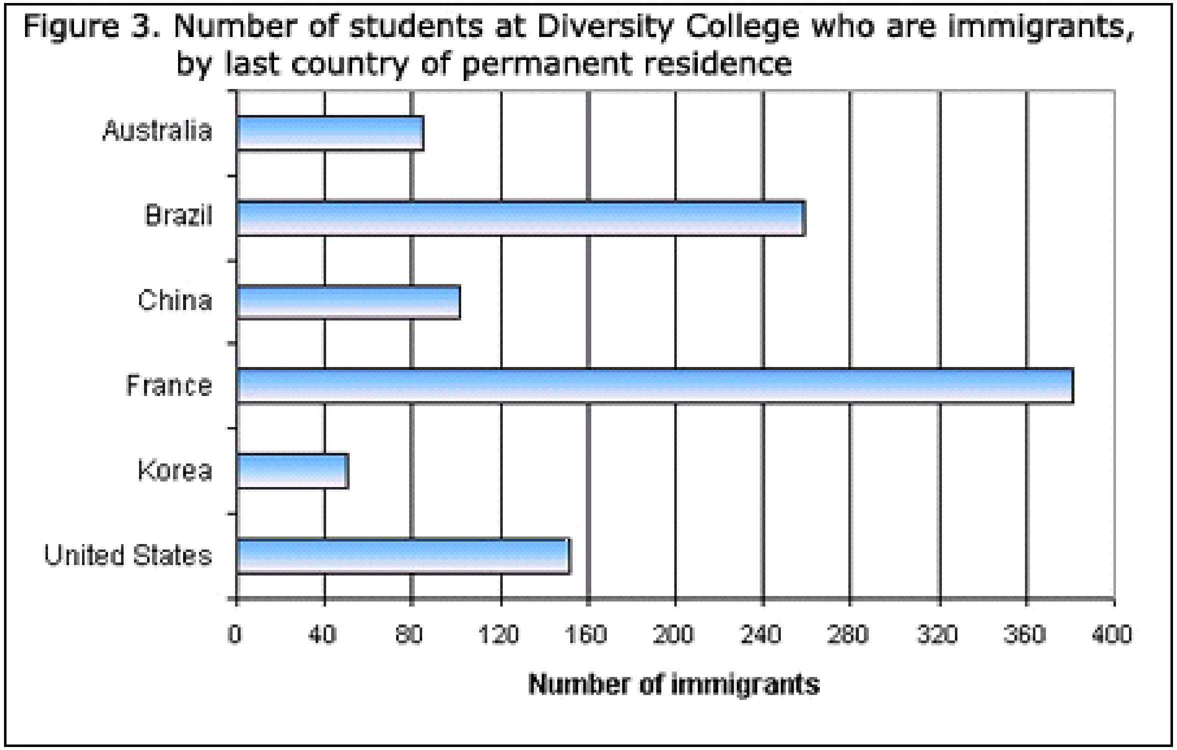
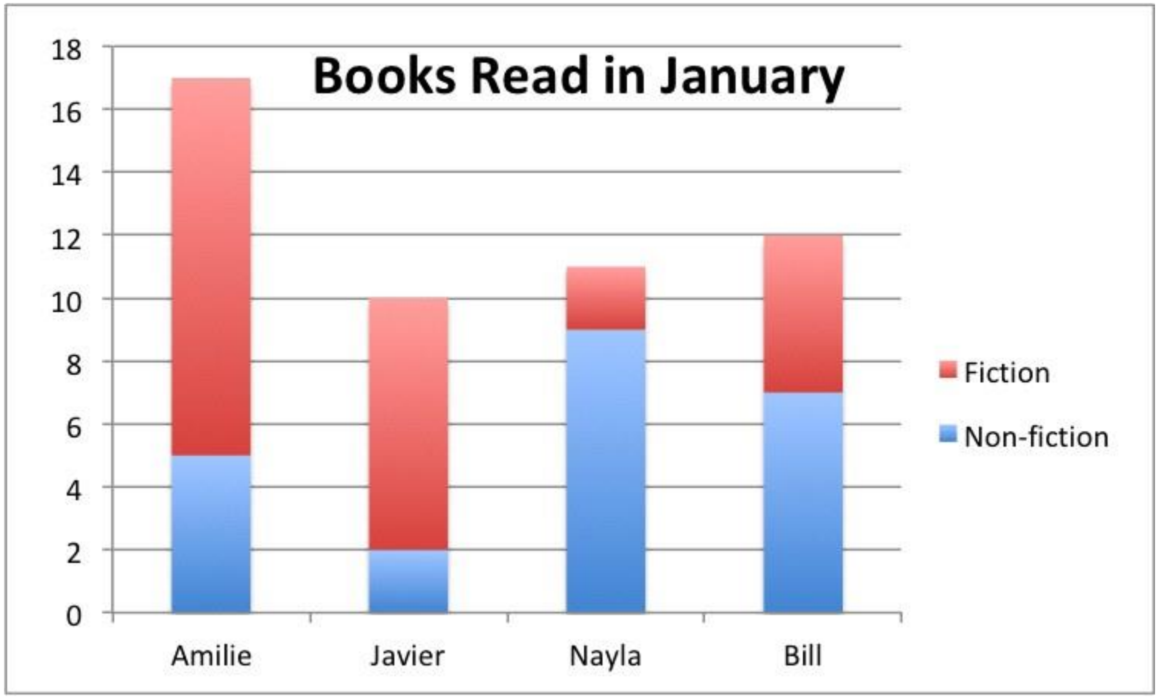
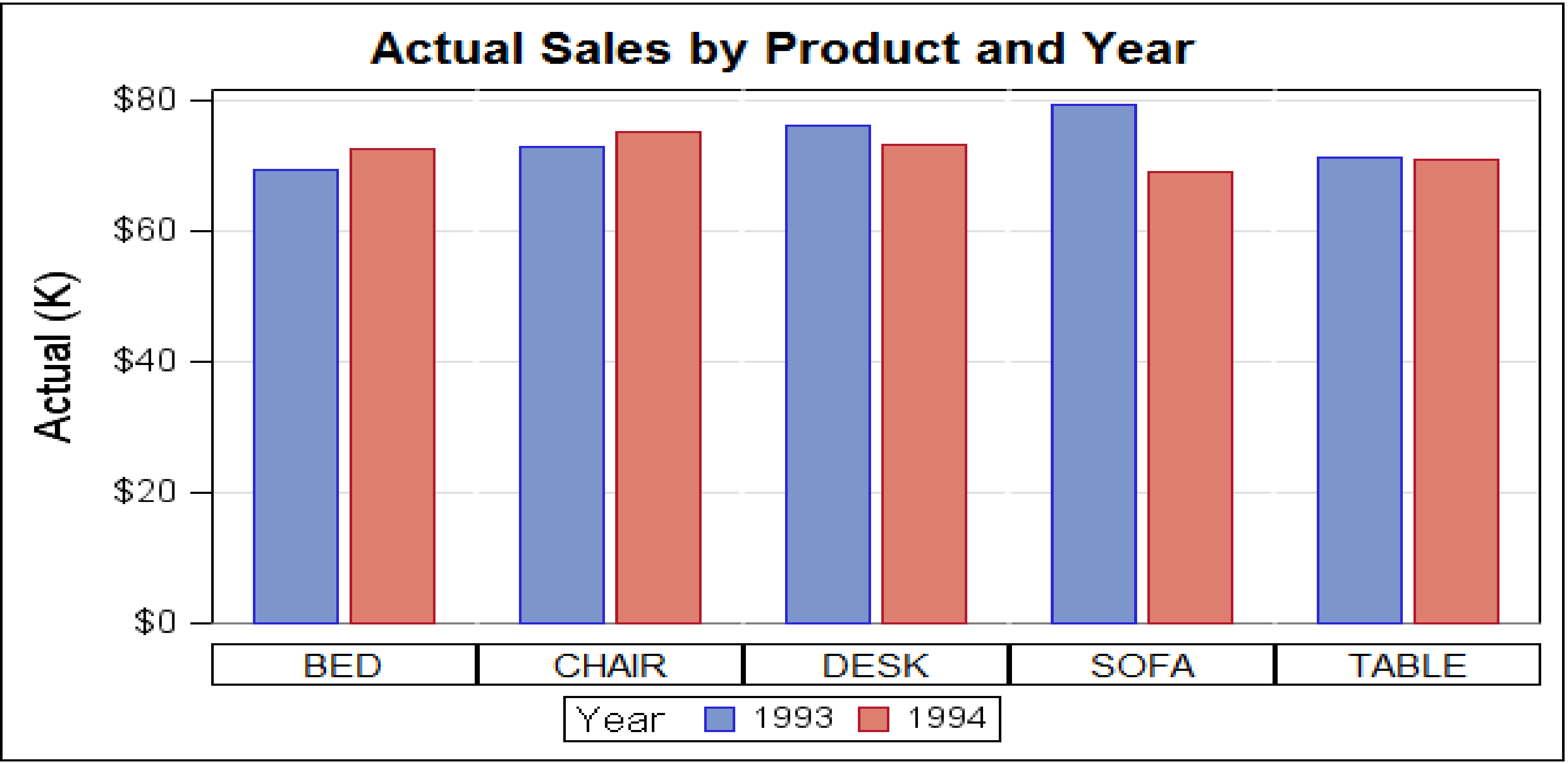


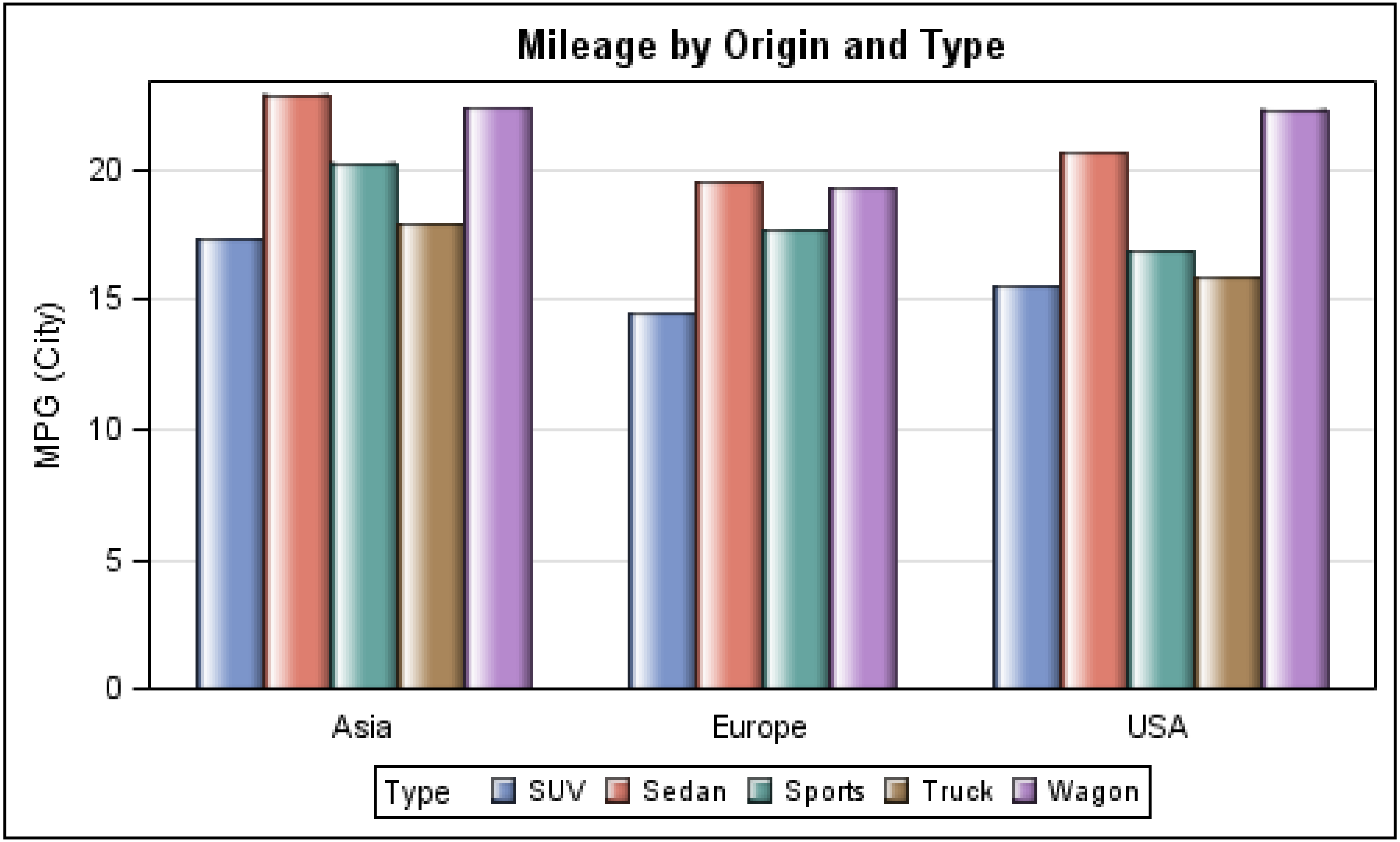
Figure 3. Number of students at Diversity College who are immigrants, by last country of permanent residence

- 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



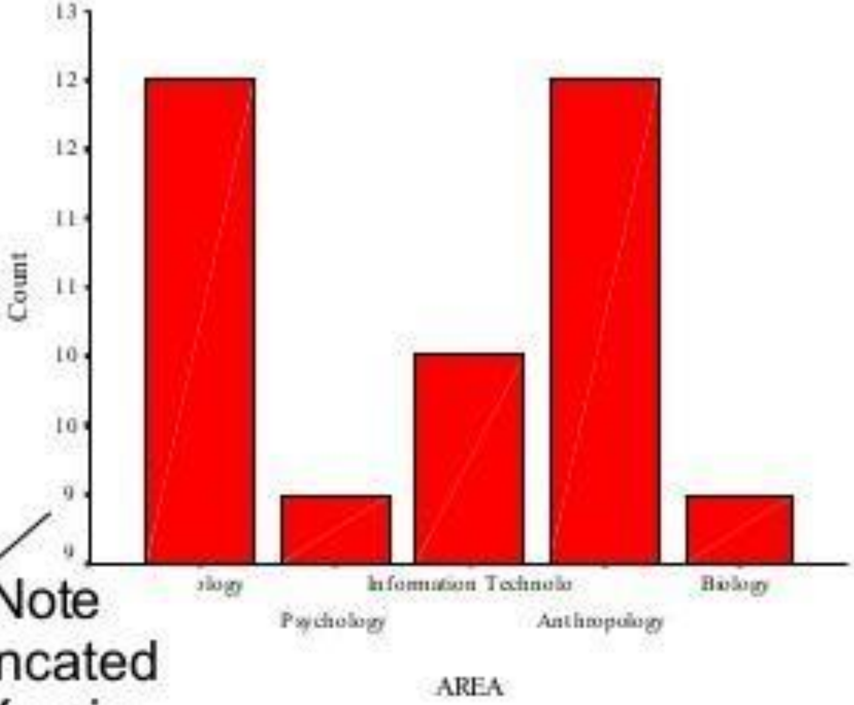
Grouped Bar Graph



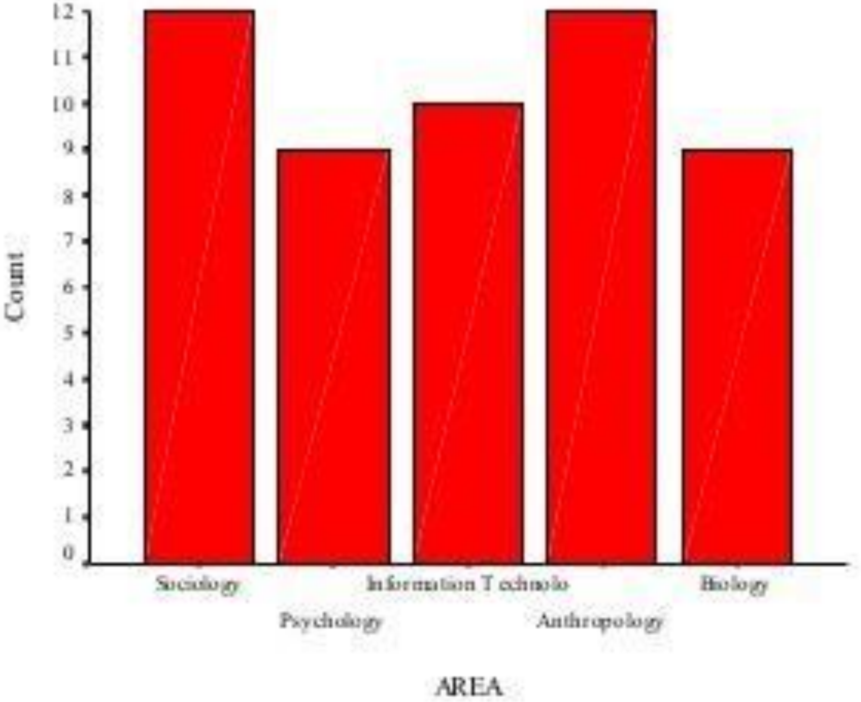

Mileage by Origin and Type
MPG (City)
0
5
10
15
20
Asia
Europe
USA
Type
SUV
Sedan
Sports
Truck
Wagon

Bar chart (Bar graph)

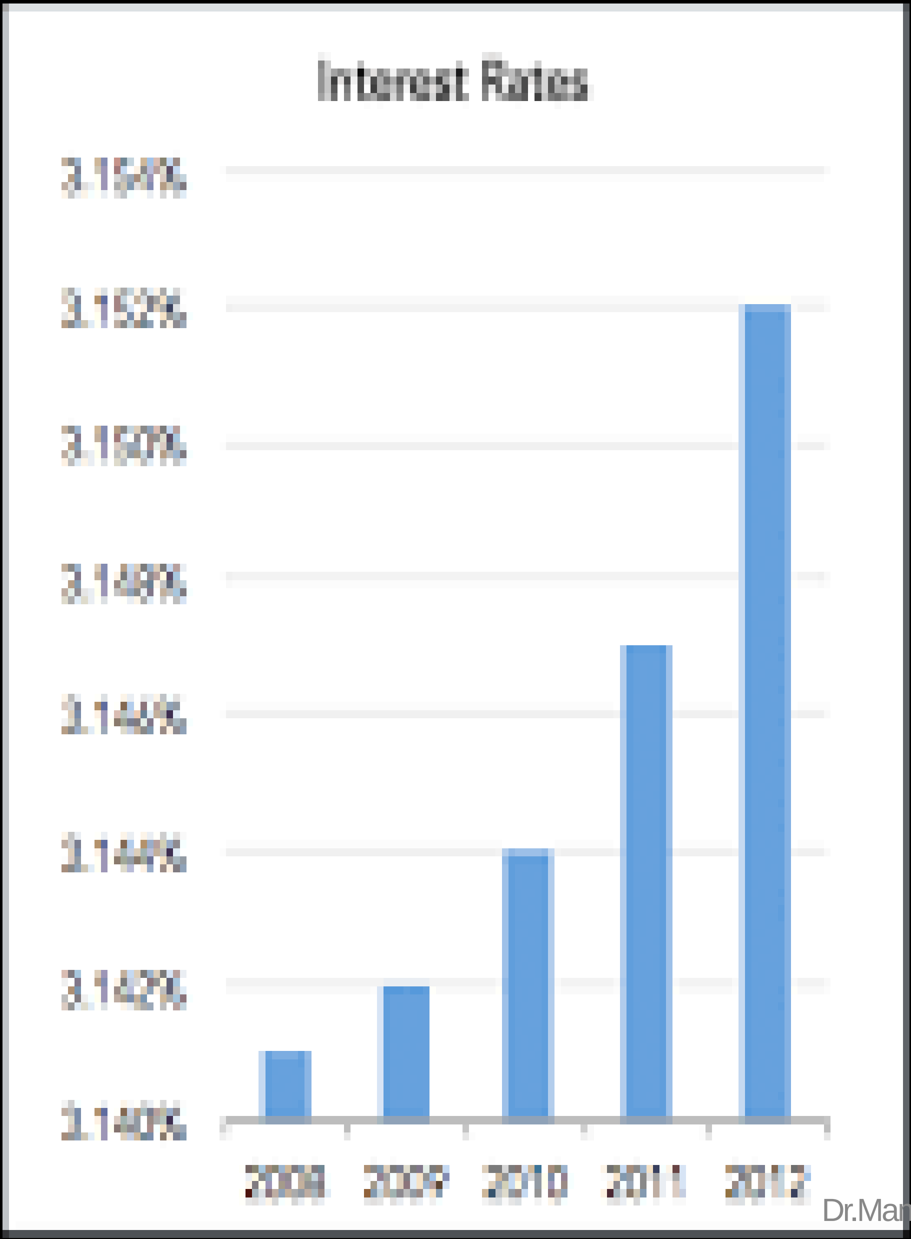
- Allows comparison of heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count/Frequency or % - truncation exaggerates differences
- Can add data labels (data values for each bar

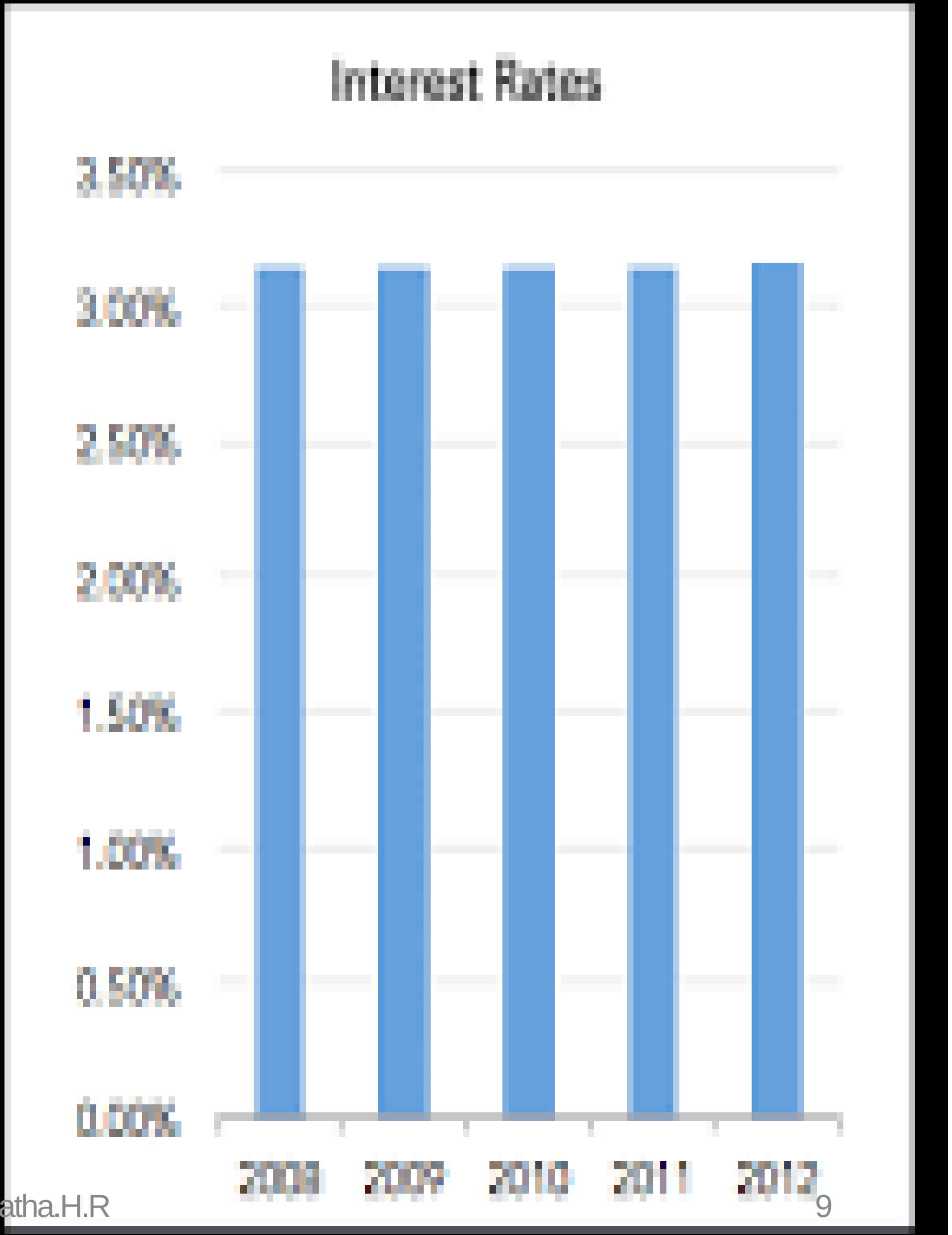


Note truncated Y-axis



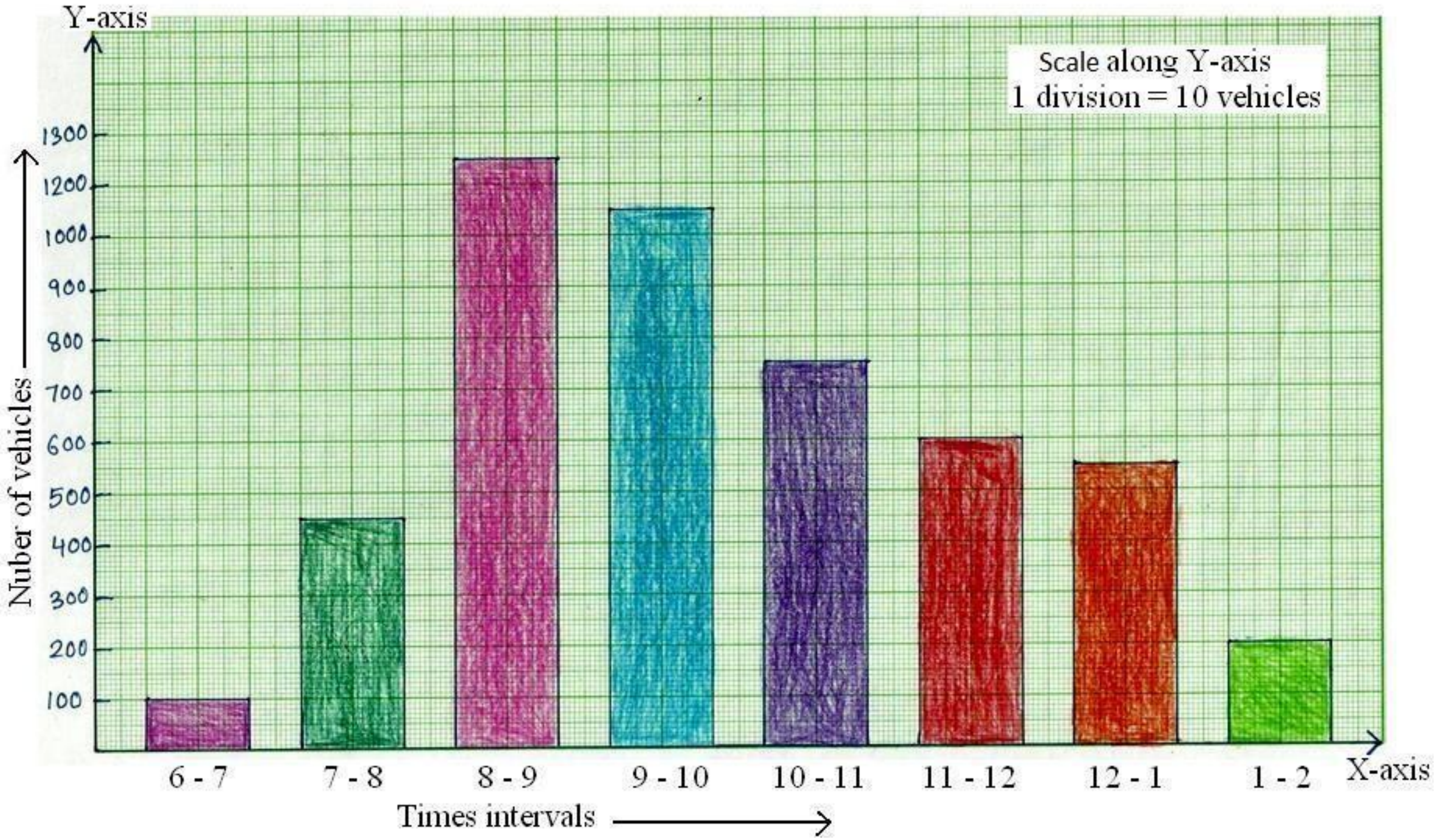
Same Data, Different Y-Axis



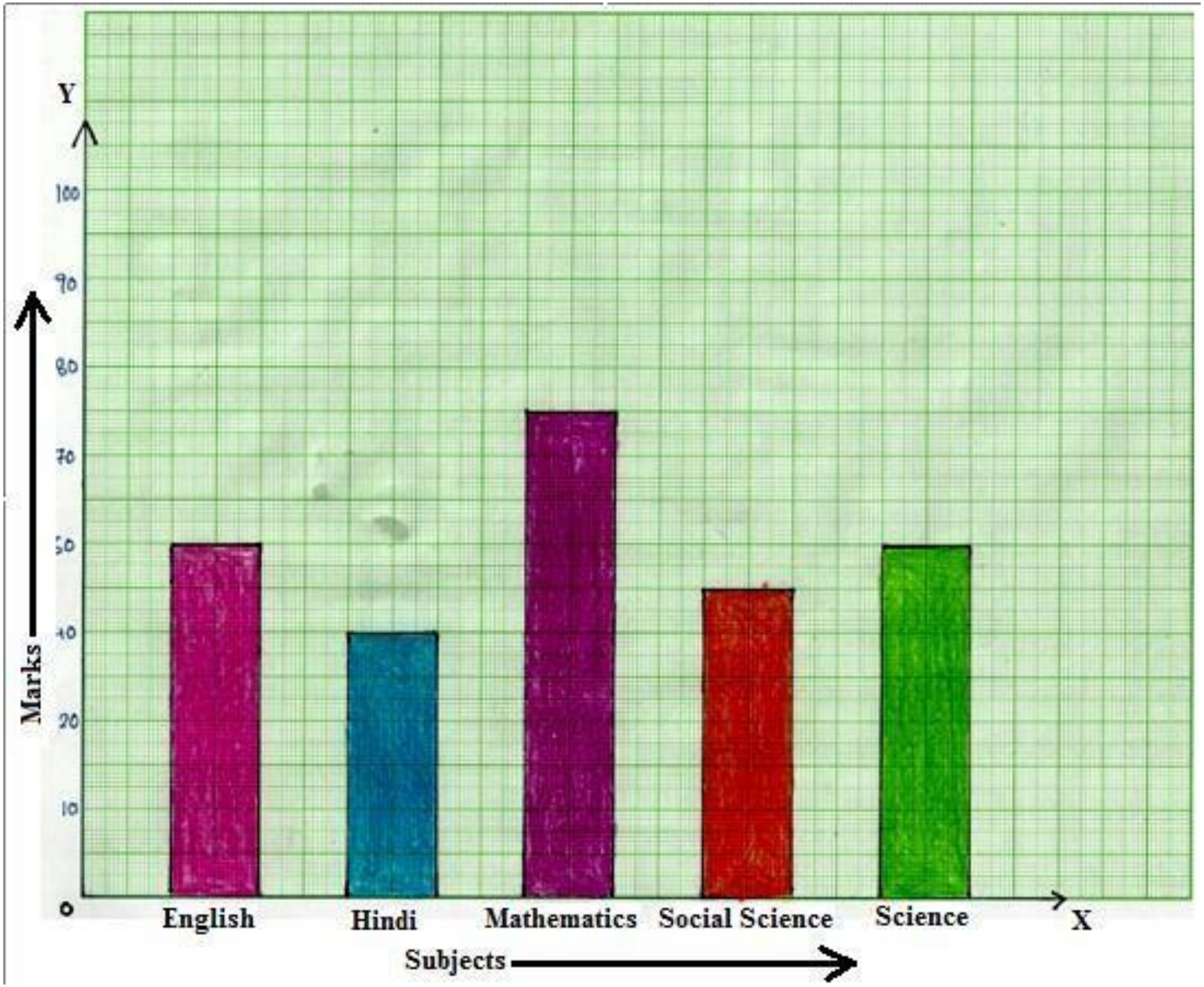


- The vehicular traffic at a busy road crossing in a particular place was recorded on a particular day from 6am to 2 pm and the data was rounded off to the nearest tens. Construct a Bar Chart.

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?

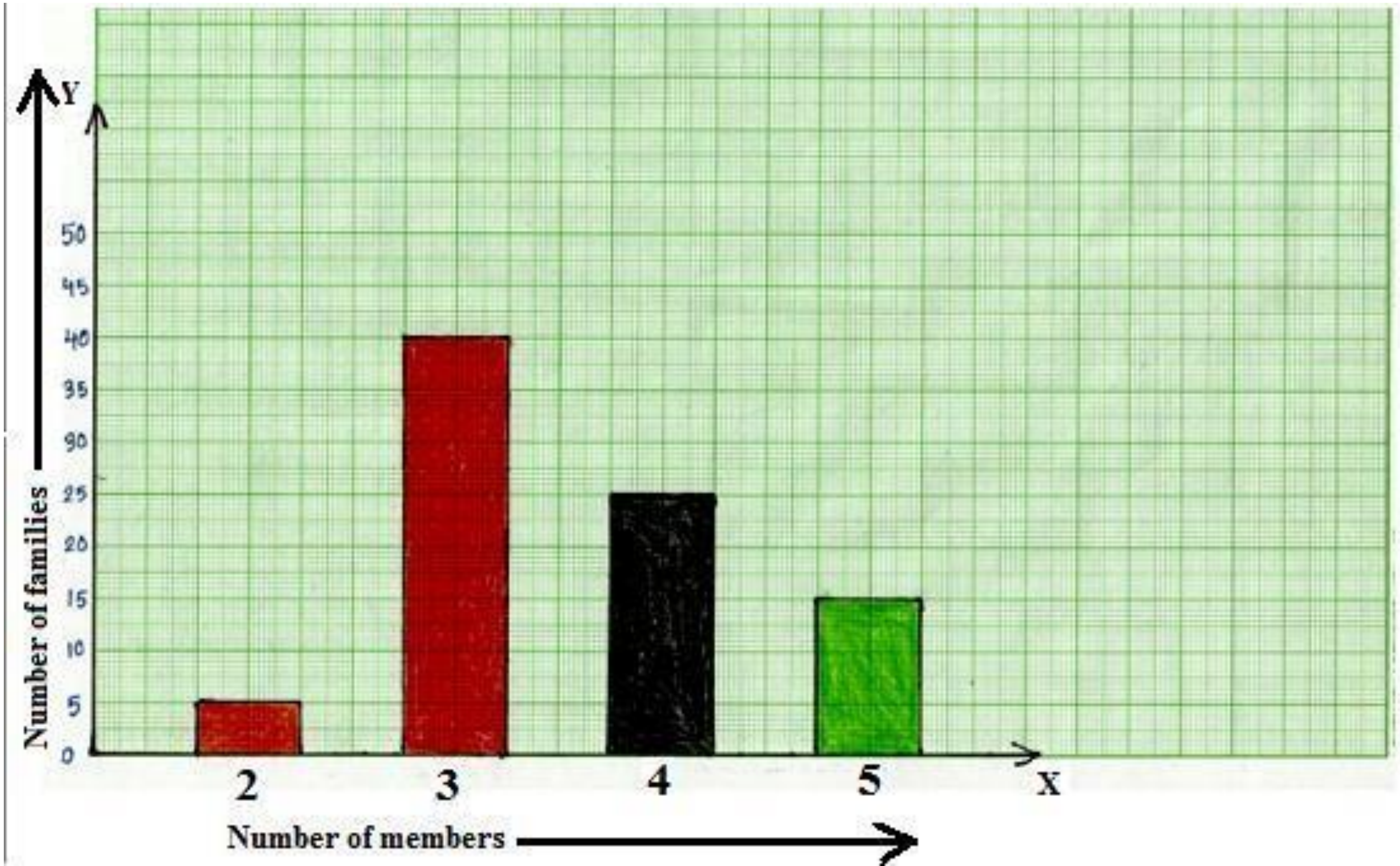
(i) It shows the marks obtained by a student in five subjects

(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

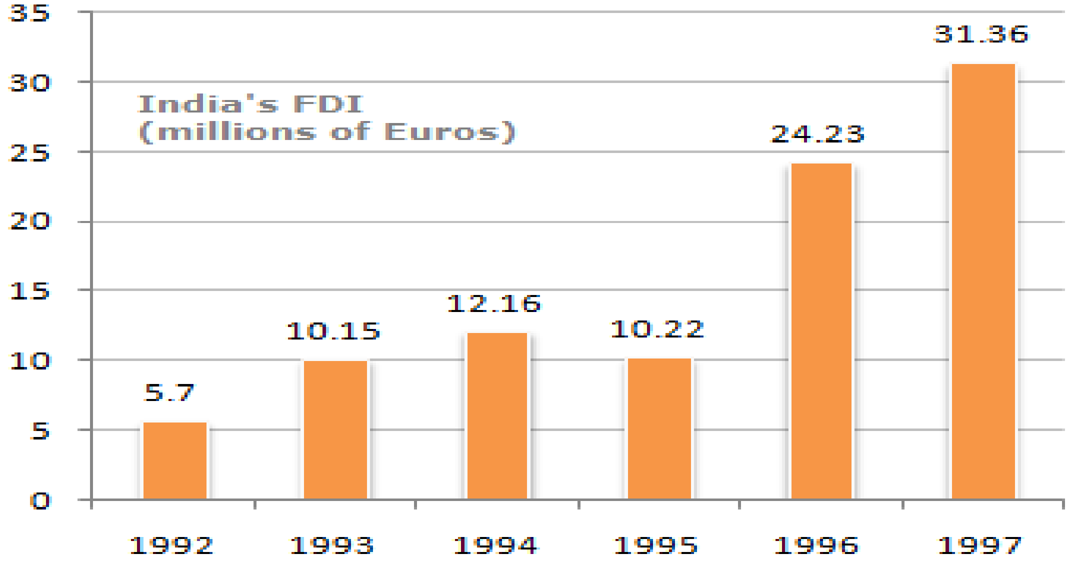
(iii) none

(iv) Family having 3 members, 3 members.

Trends of FDI in India

The following bar chart shows the trends of foreign direct investments(FDI) into India from all over the world.

Trends of FDI in India



1. What was the ratio of investment in 1997 over the investment in 1992 ?

The 1997 figure of investment as a factor of 1992 investment = (31.36/5.70) = 5.50

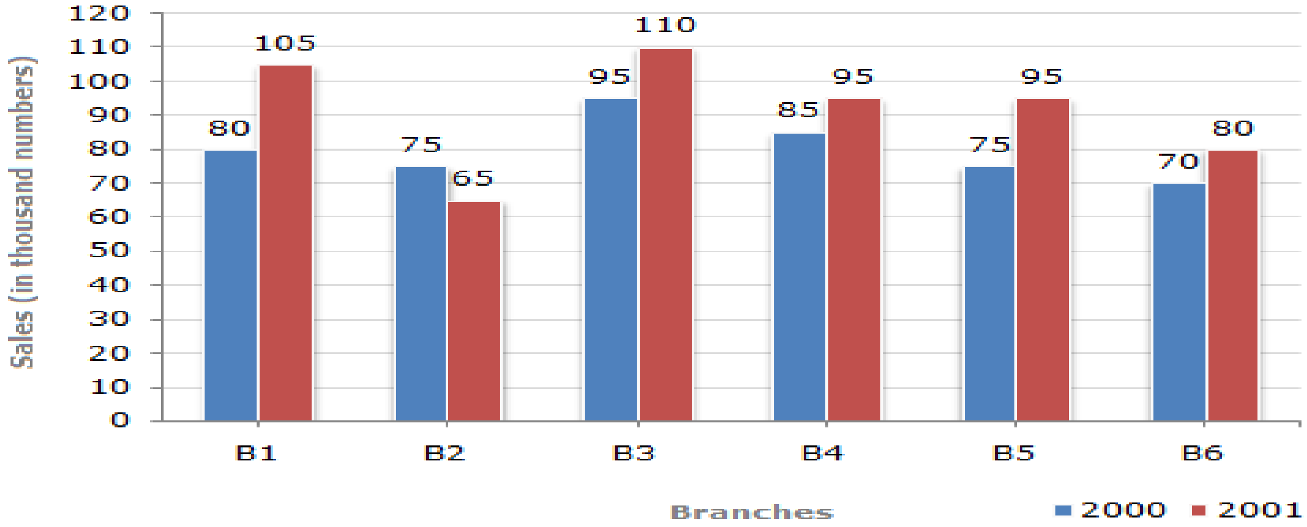
2. What was absolute difference in the FDI to India in between 1996 and 1997 ?

The difference in investments over 1996-1997 was 31.36 - 24.23 = € 7.13 millions.

3. Which year exhibited the highest growth in FDI in India over the period shown ?

1996

Reason : It can be seen that the FDI in 1996 more than doubles over that of 1995. No other year is close to that rate of growth.



1. What is the ratio of the total sales of branch B2 for both years to the total sales of branch B4 for both years?

Ans. : Required ratio =(75 + 65)=140=7.(85 + 95)1809

2. Total sales of branch B6 for both the years is what percent of the total sales of branches B3 for both the years?

Ans. : Required percentage=(70 + 80)x 100%(95 + 110)=150x 100%205= 73.17%.

3. What percent of the average sales of branches B1, B2 and B3 in 2001 is the average sales of branches B1, B3 and B6 in 2000?

Average sales (in thousand number) of branches B1, B3 and B6 in 2000

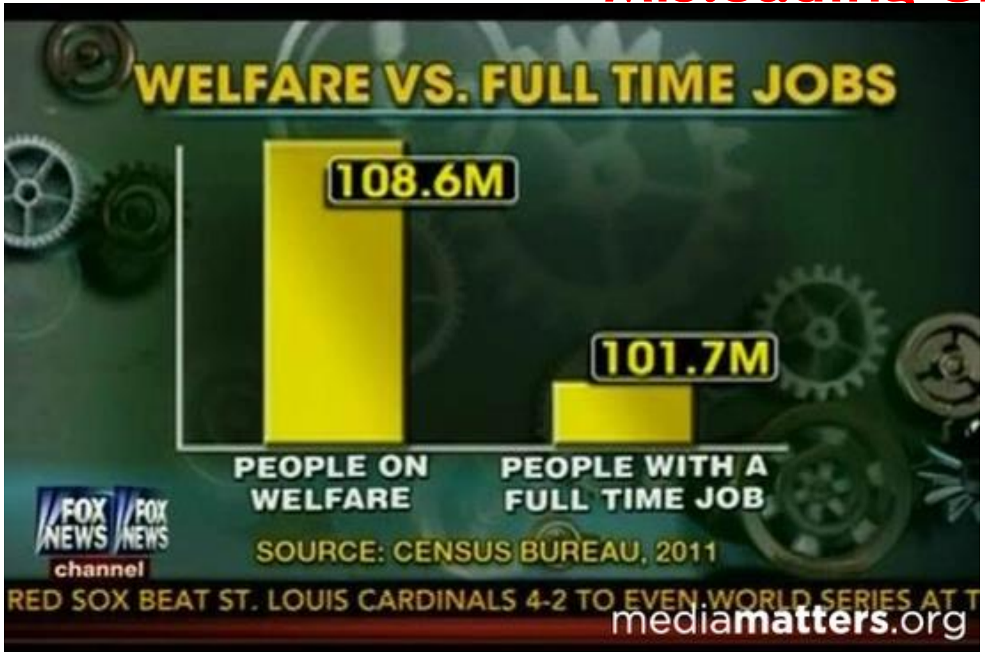
=1x (80 + 95 + 70)=245.33

Average sales (in thousand number) of branches B1, B2 and B3 in 2001

=1x (105 + 65 + 110)=280.33

Required percentage =245/3 / 280/3 x 100% =245 / 280 x 100% = 87.5%.

Misleading Graphs



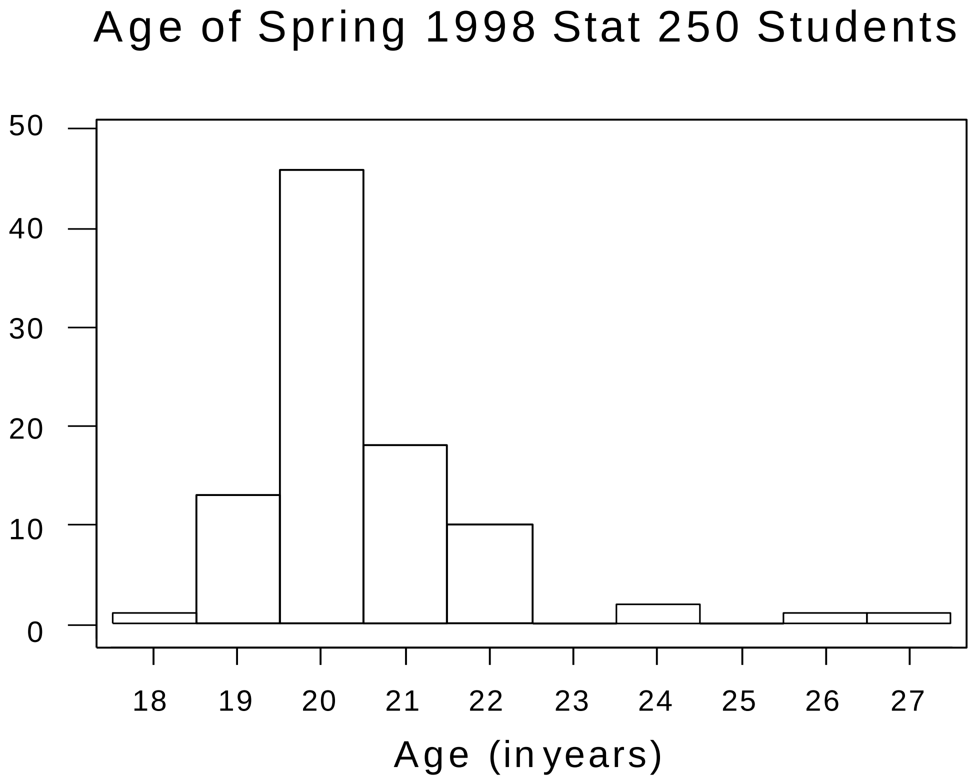
What's the real data behind this shocking graph? Are there really more people on welfare than those who have full time jobs? As Media Matters points out:

"Fox's 108.6 million figure for the number of "people on welfare" comes from a Census Bureau's account…of participation in means-tested programs, which include "anyone residing in a household in which one or more people received benefits" in the fourth quarter of 2011, thus **including individuals who did not themselves receive government benefits**.

On the other hand, the "people with a full time job" figure Fox used included only individuals who worked, not individuals residing in a household where at least one person works."

HISTOGRAMS

Histogram

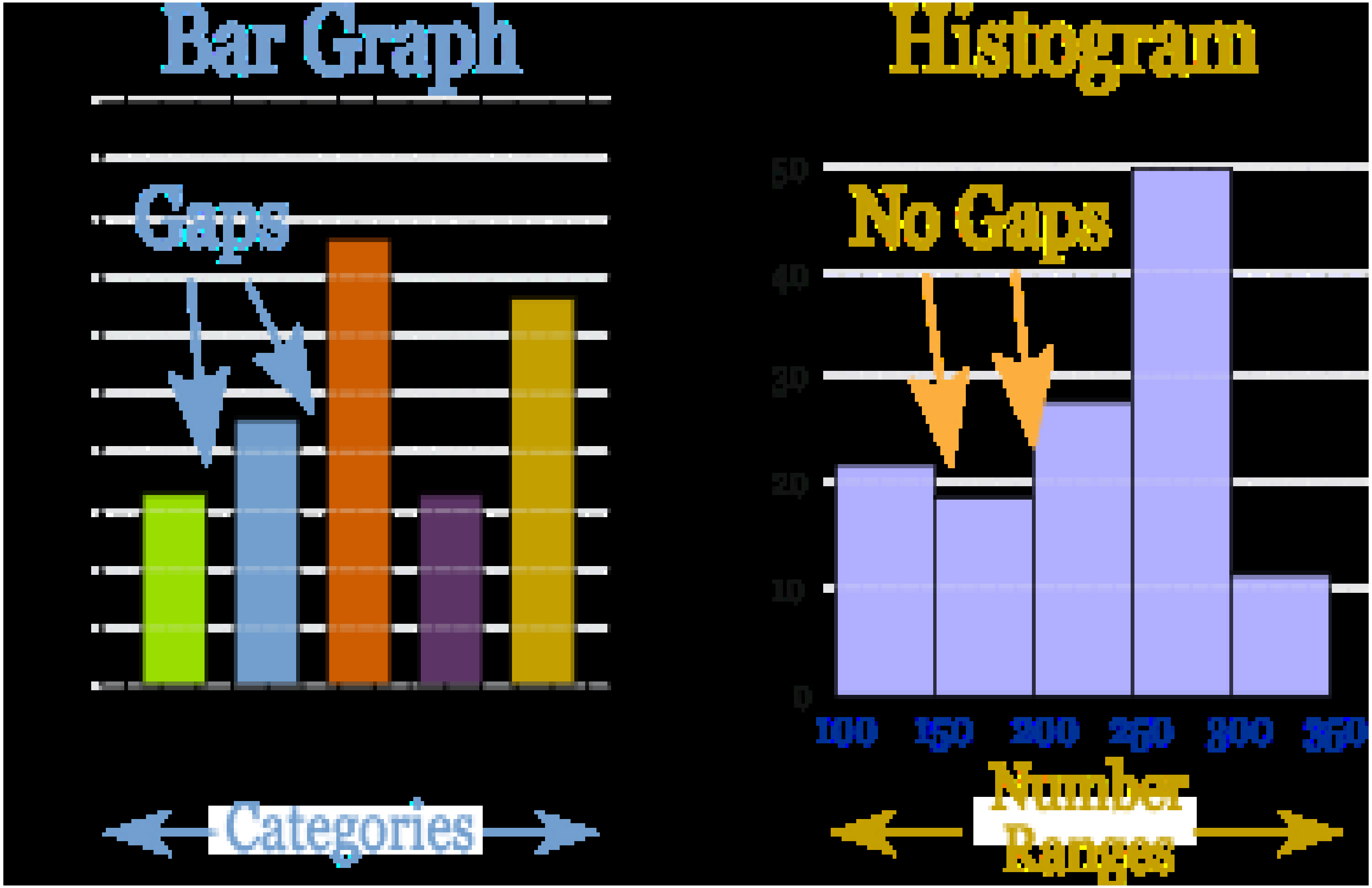


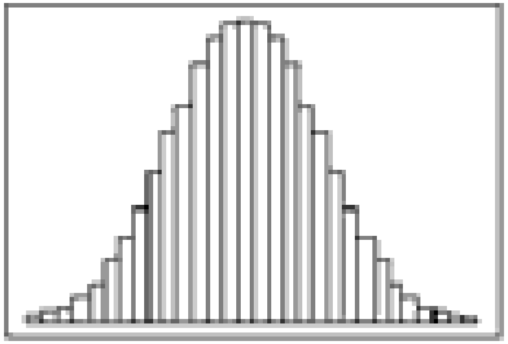
n=92 students

Analogy

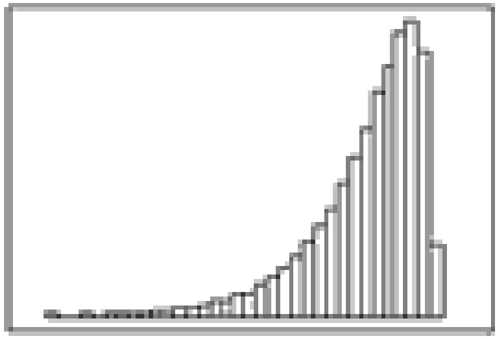
Bar chart is to categorical data as histogram is to ...

measurement data.

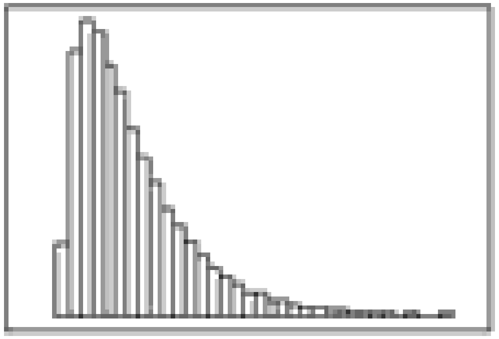

Bar Graph
Gaps
Categories
Histogram
No Gaps
50
40
30
20
10
0
100
150
200
250
300
350
Number Ranges



Symmetric
Bell shaped

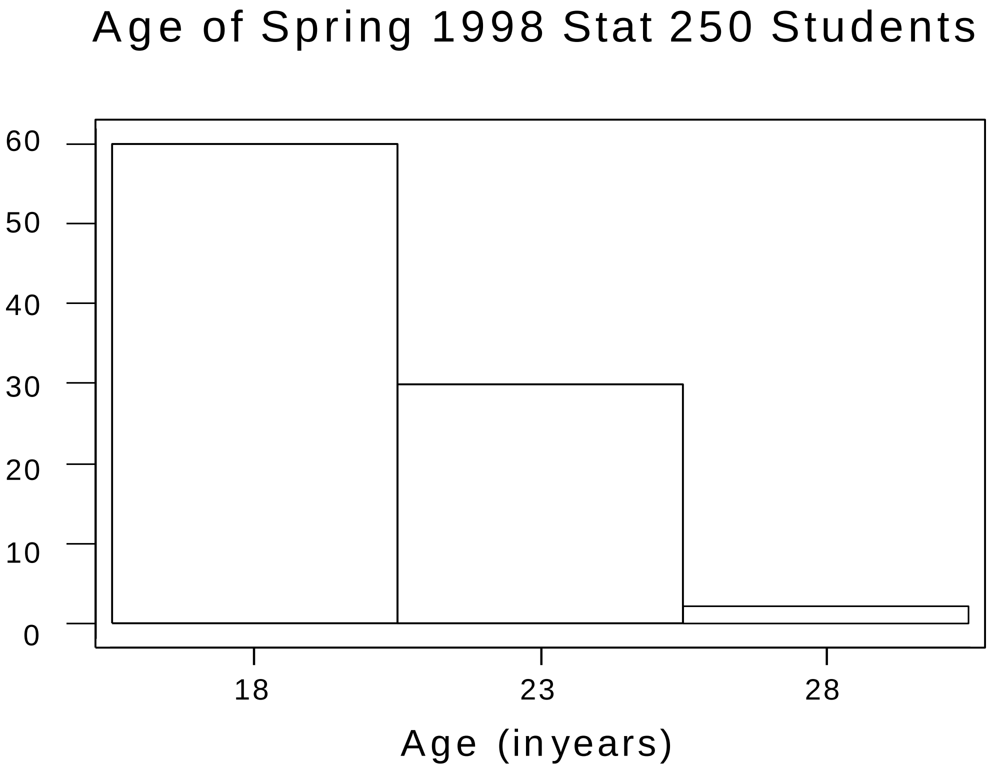


Skewed to
the Left



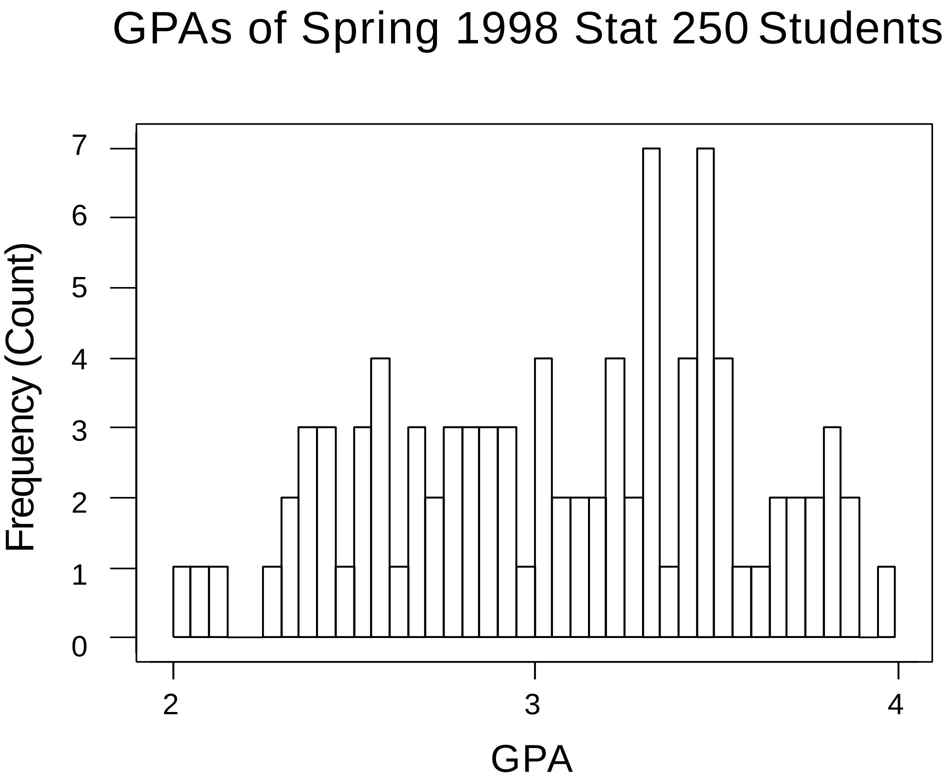
Skewed to
the Right

Too few categories



n=92 students

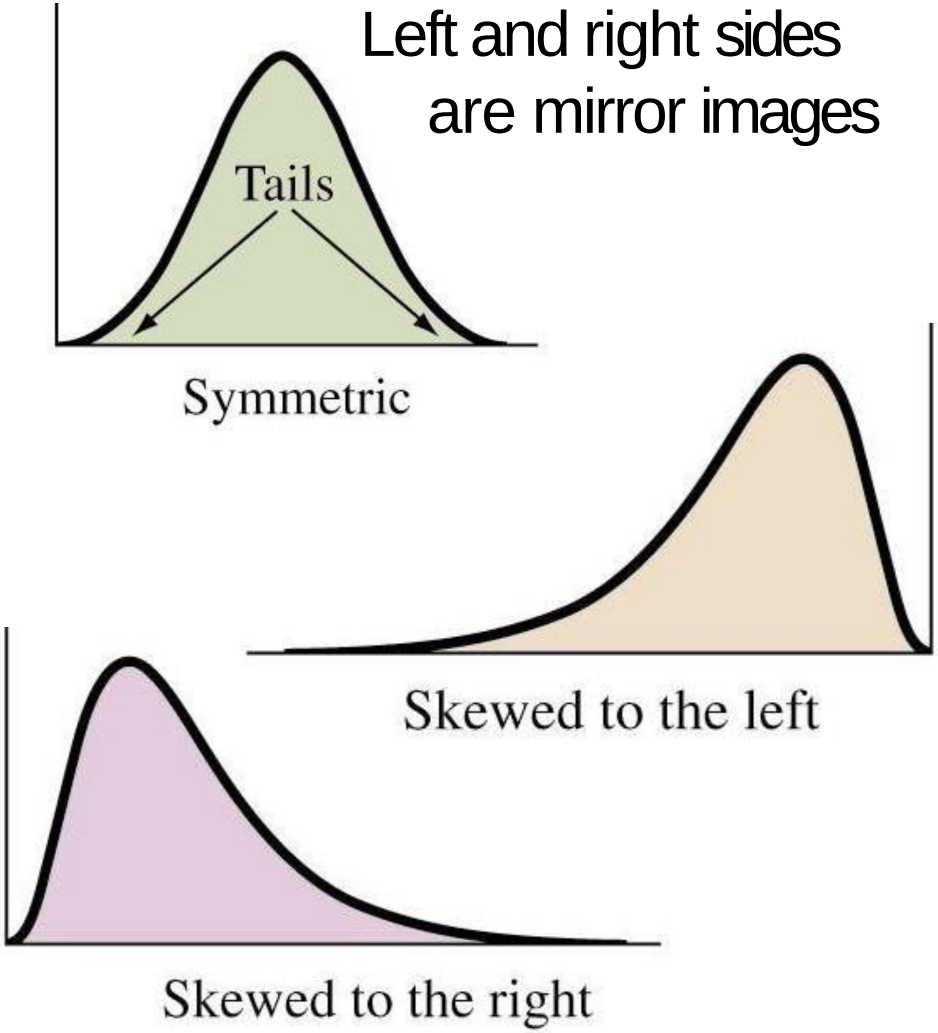
Too many categories



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

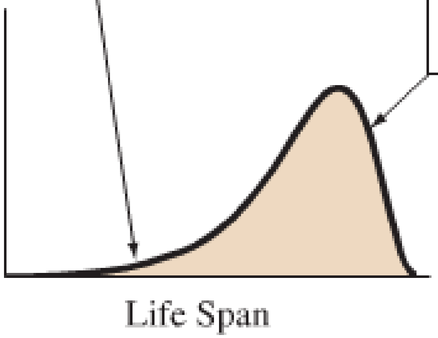


Left and right sides are mirror images

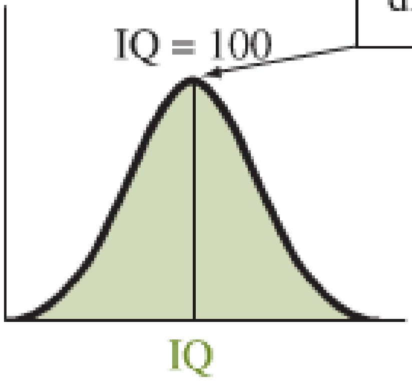
Examples of Skewness

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

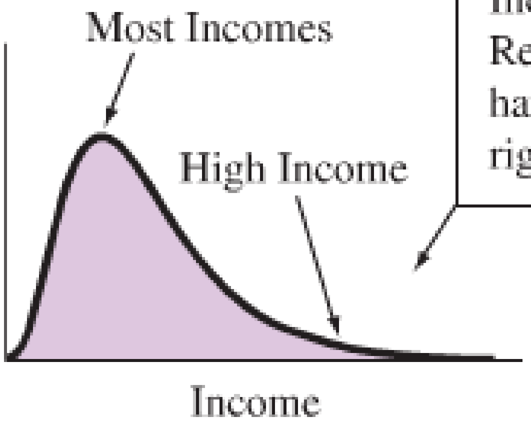
Most observations are here.



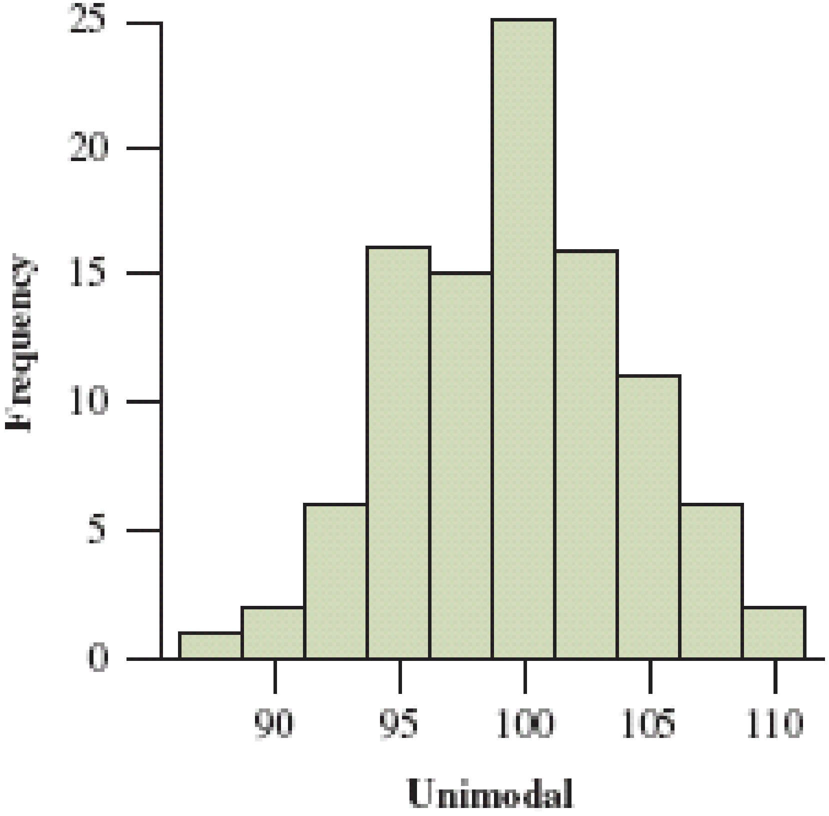
IQ has a symmetric distribution

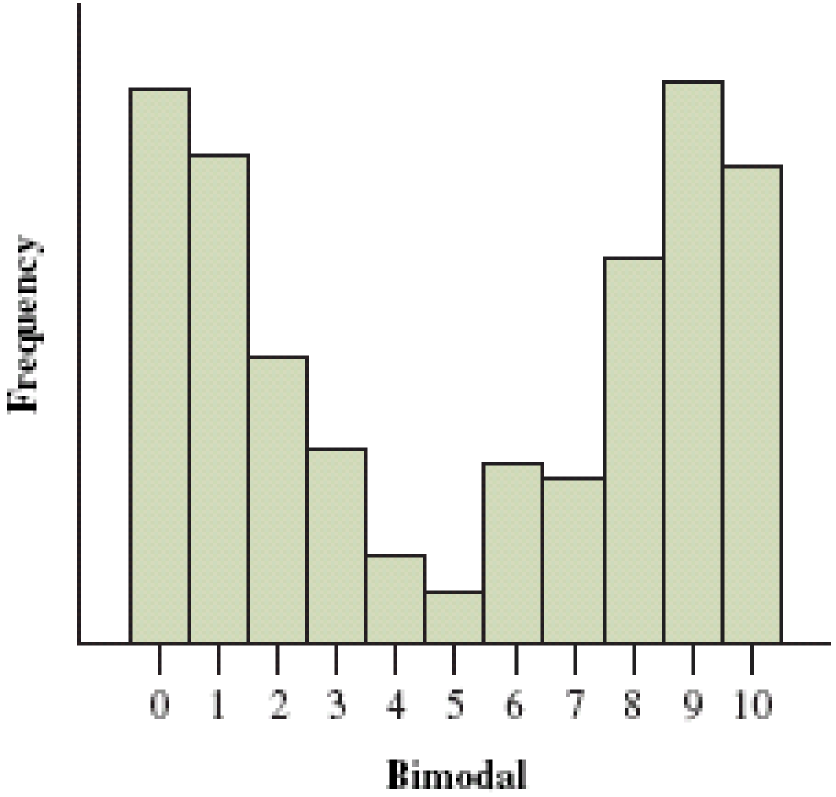


Income skews to the right. Relatively few are rich and have observations in this long right tail.



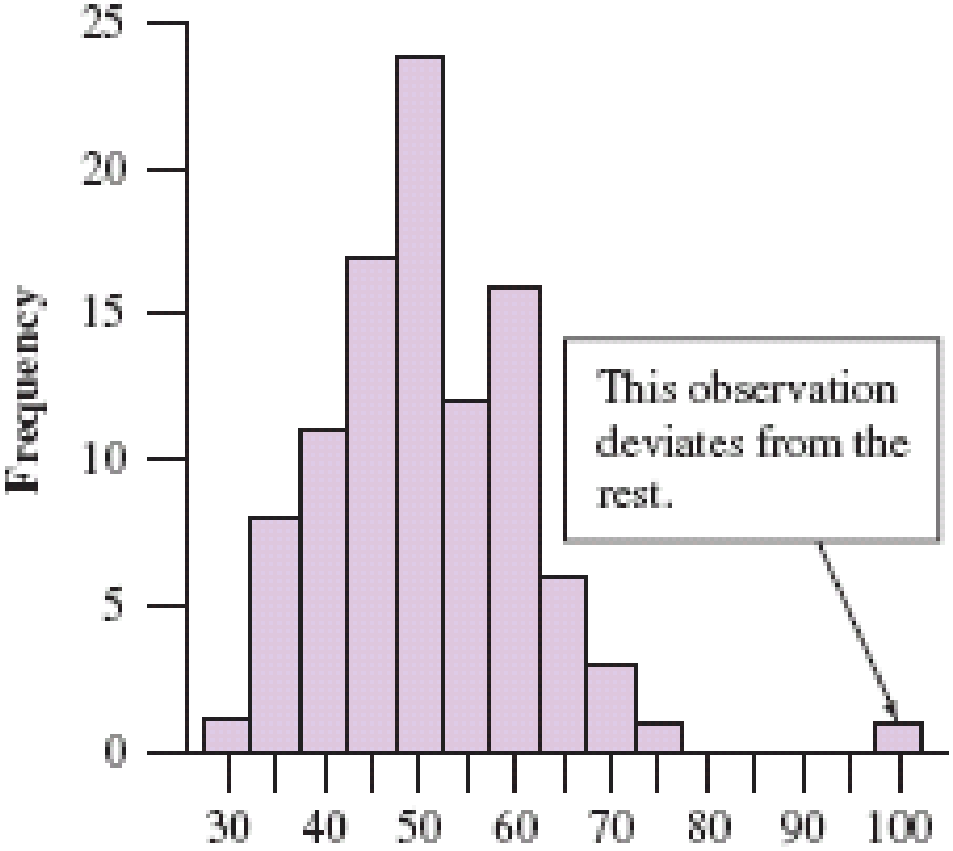
Shape: Type of Mound



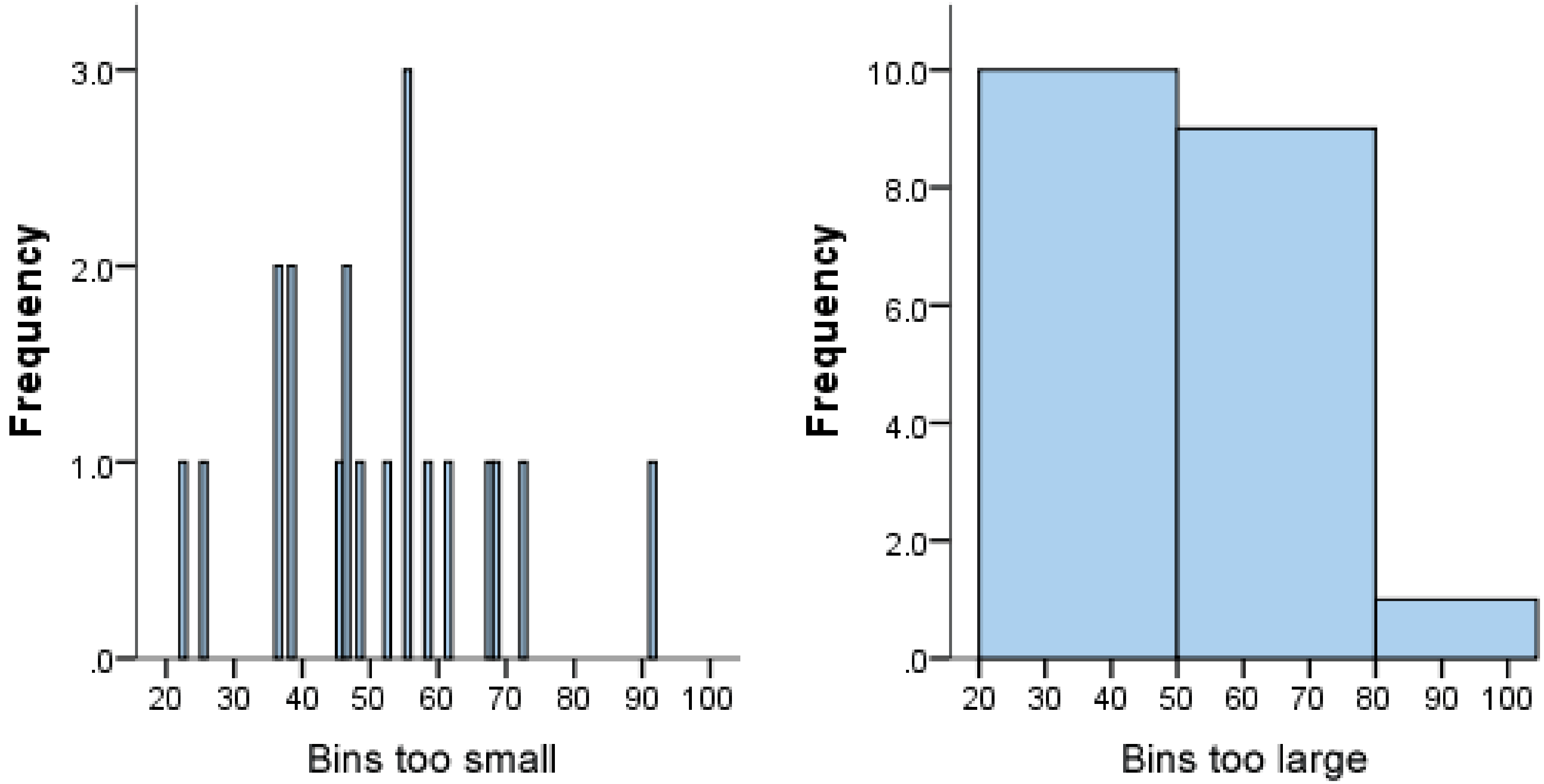


Outlier

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.

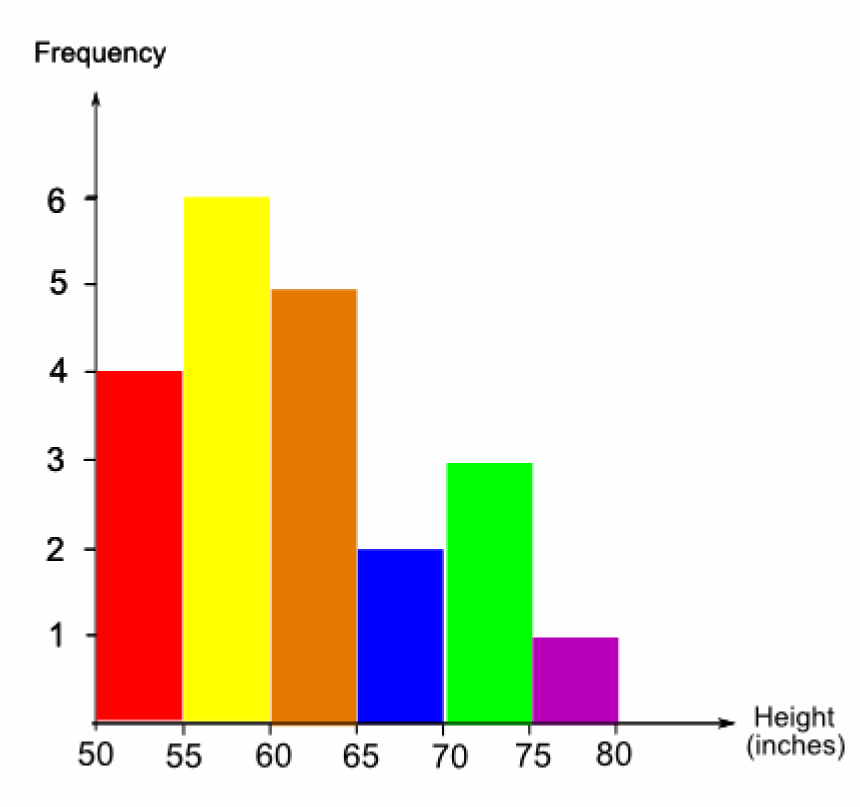
- The number of bins k can be assigned directly or can be calculated from a suggested bin width h as:

- 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

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

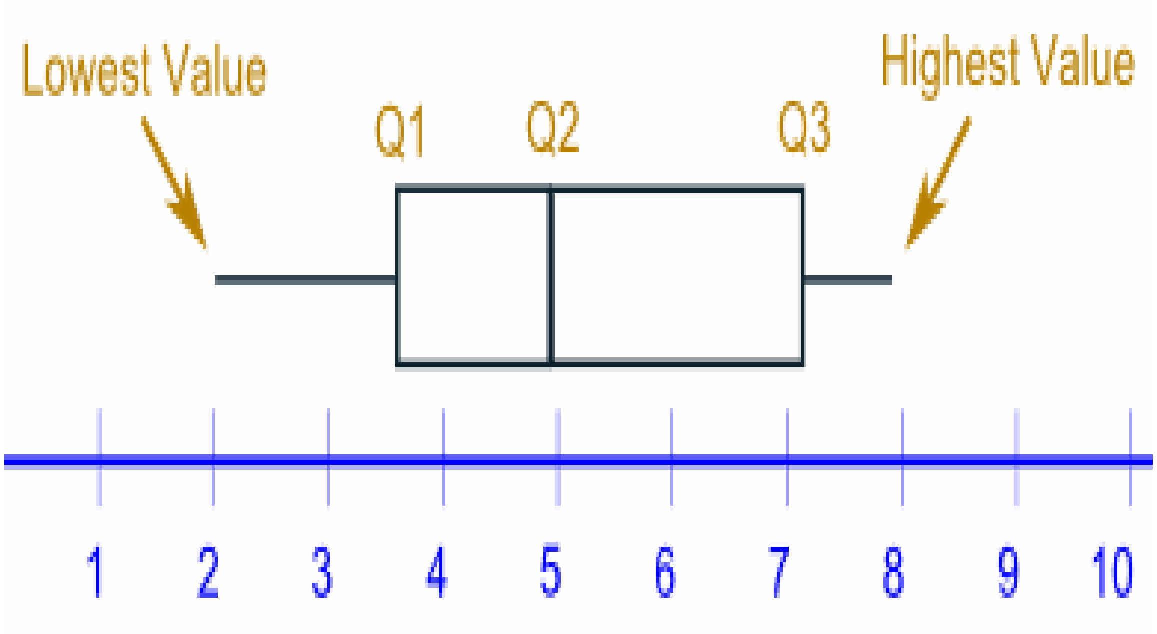


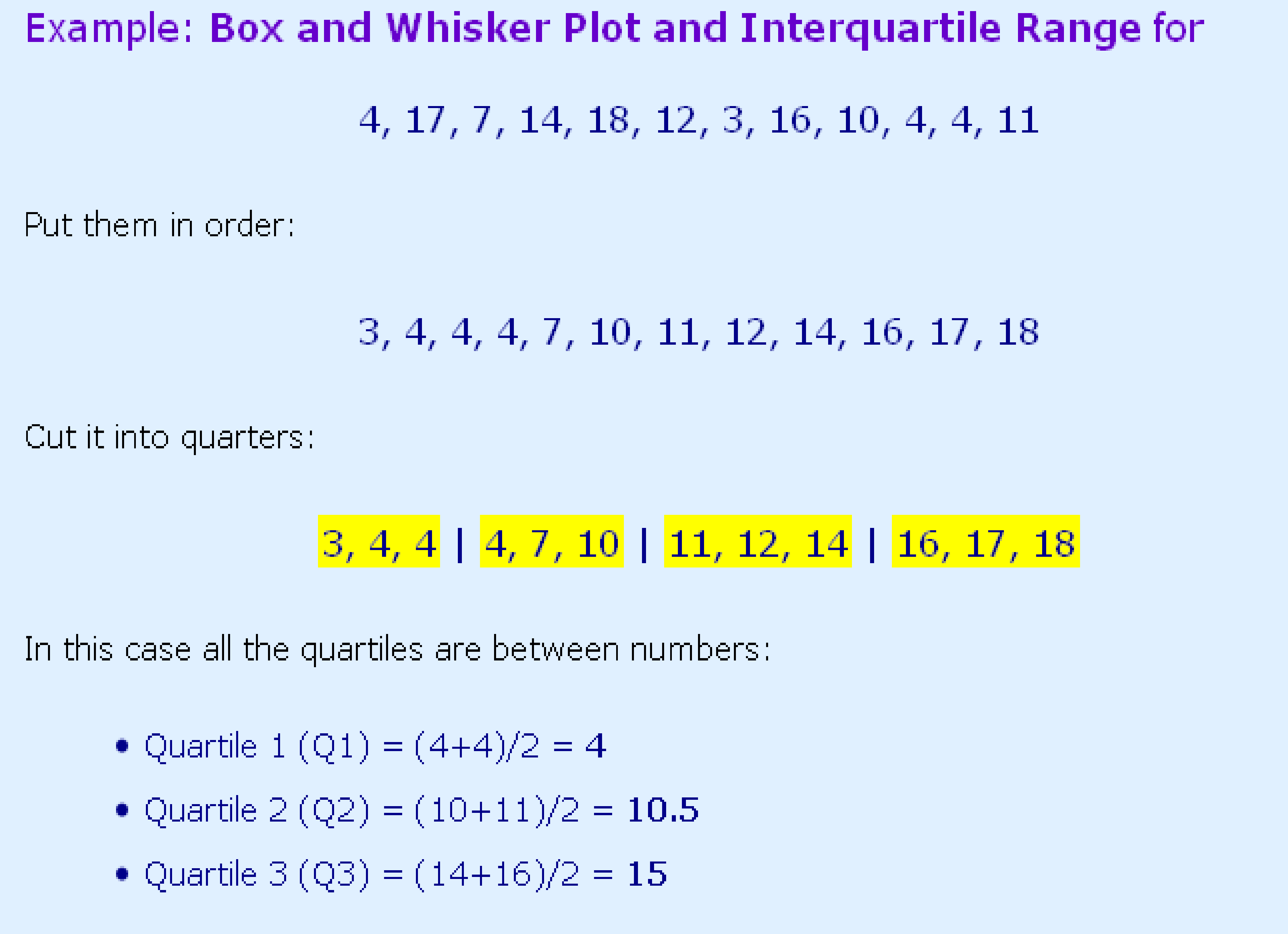
The histogram shows the heights of 21 students in a class, grouped into 5-inch groups.

1. How many students were greater than or equal to 60 inches tall? Ans. : 11
2. How many students were greater than or equal to 55 inches tall but less than 70 inches tall?
   Ans. : 13

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:

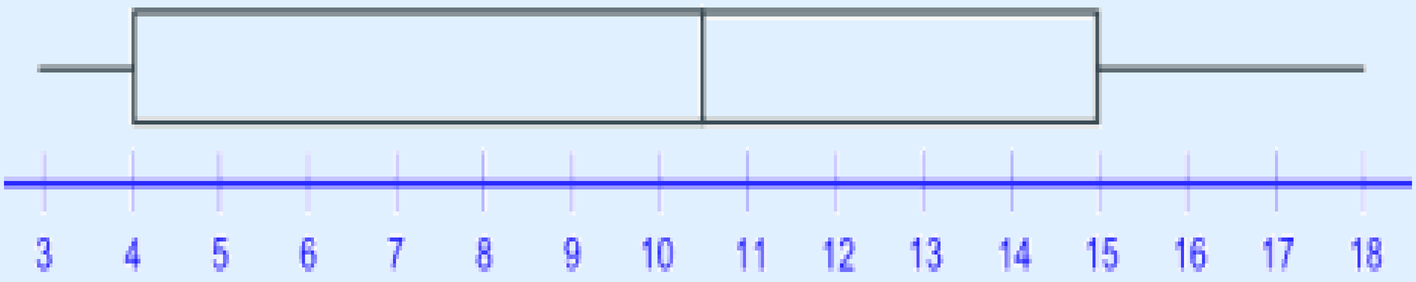
3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18

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**

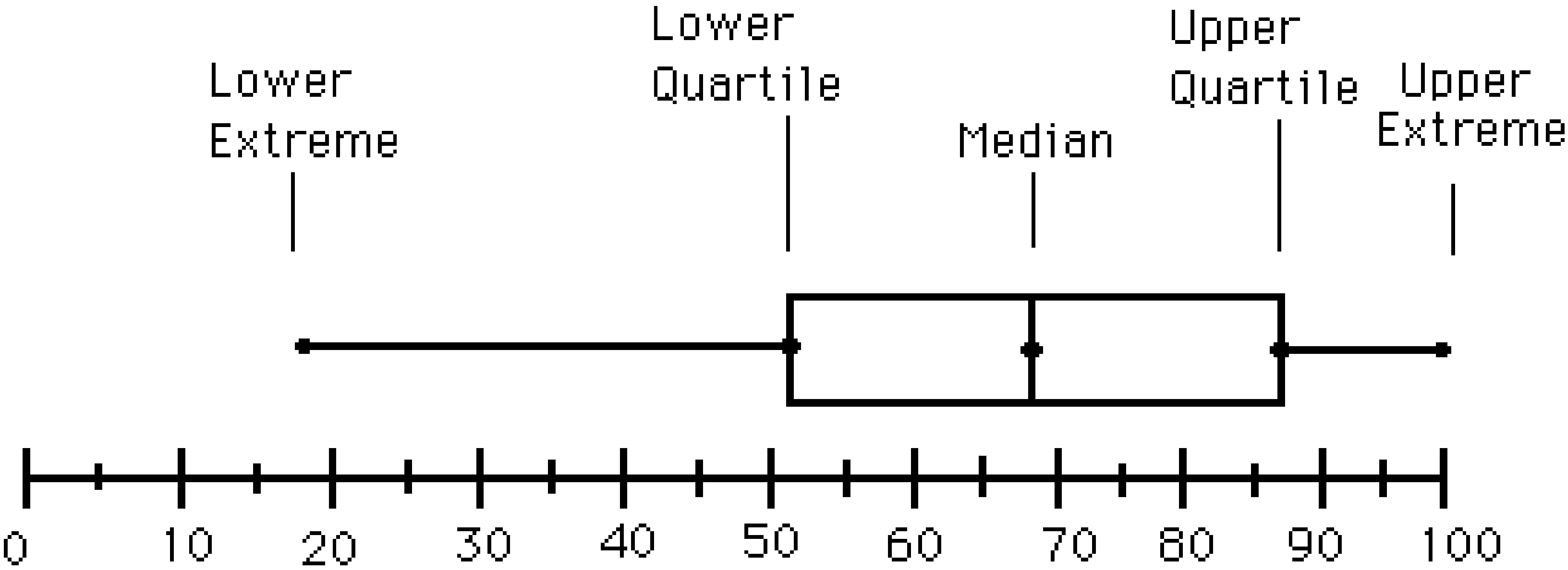
- The Lowest Value is 3,
- The Highest Value is 18

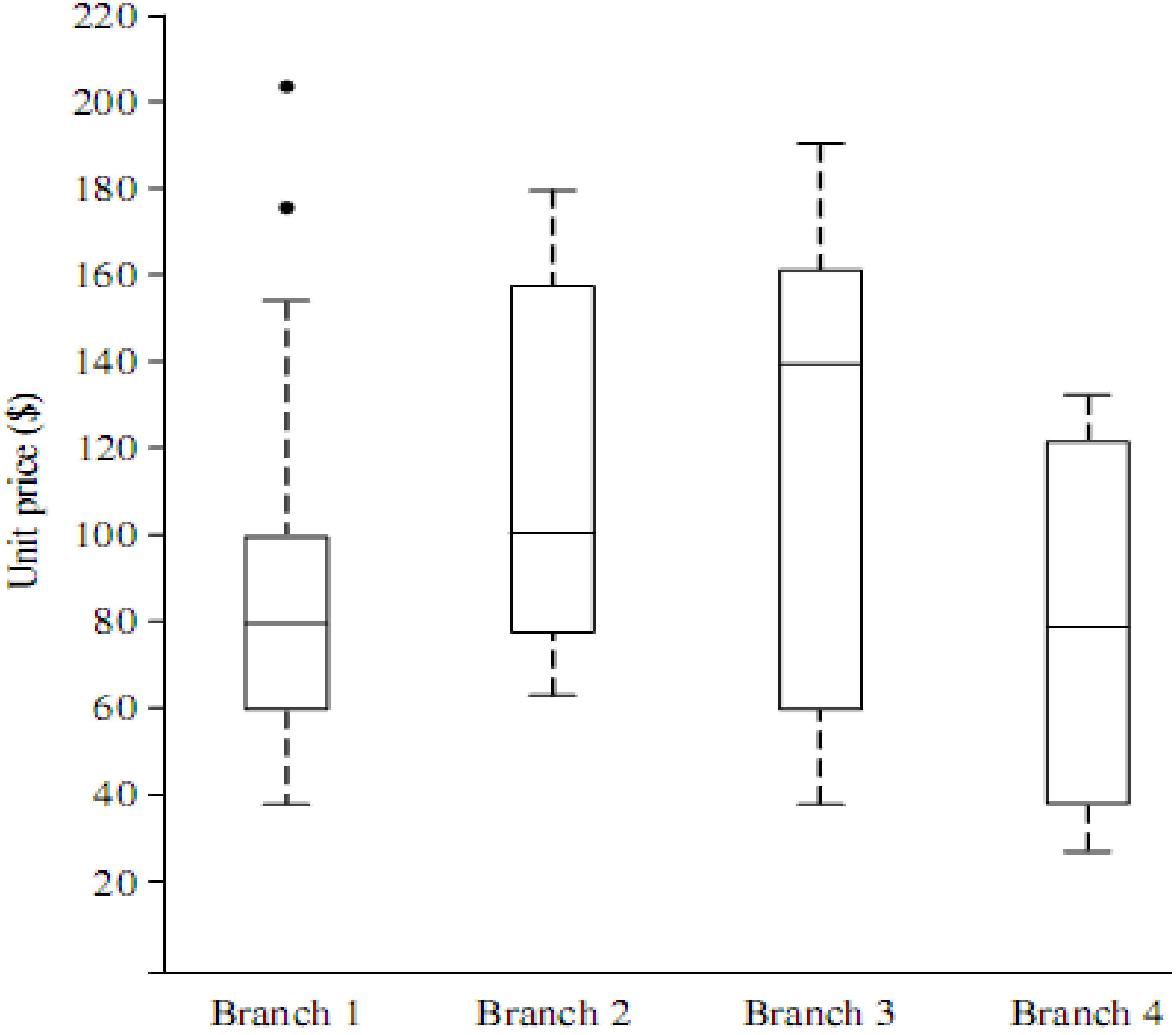
So now we have enough data for the **Box and Whisker Plot**:



And the **Interquartile Range** is:

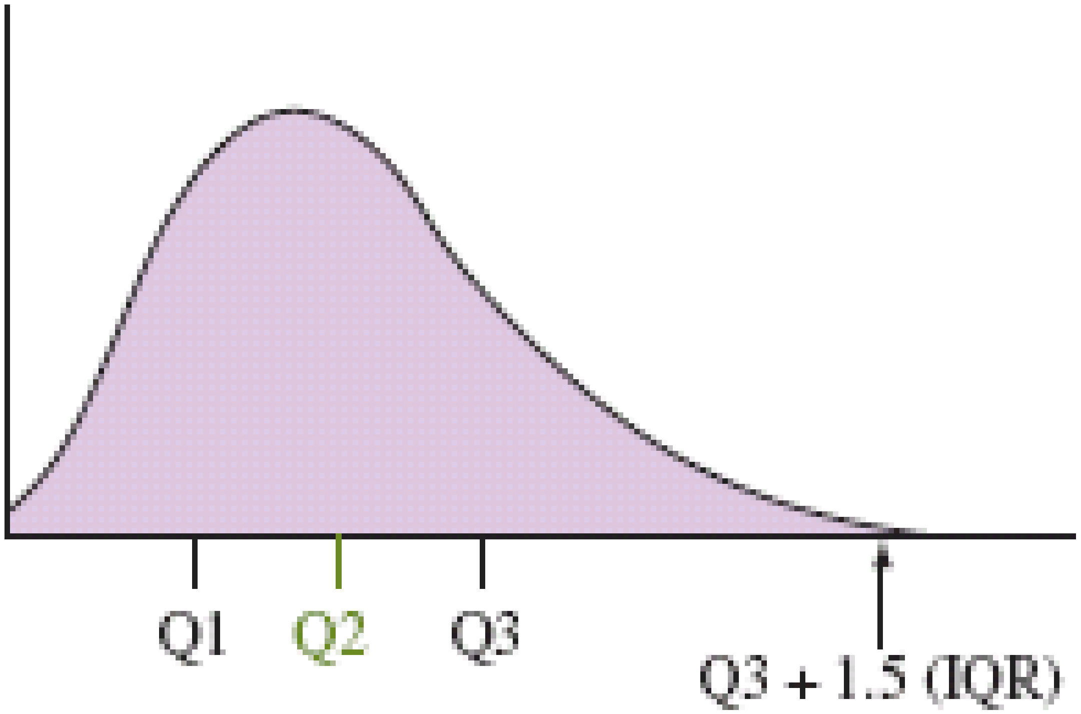
Q3 - Q1 = 15 - 4 = **11**


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


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

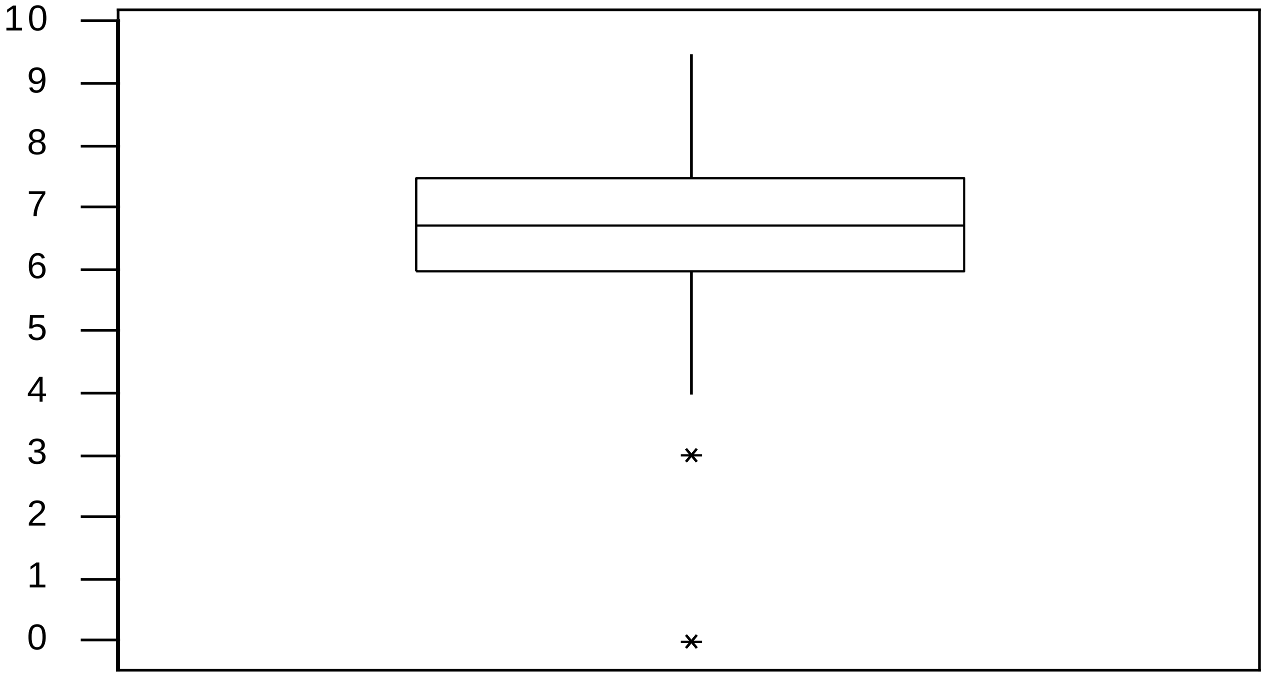
Criteria for Identifying an Outlier

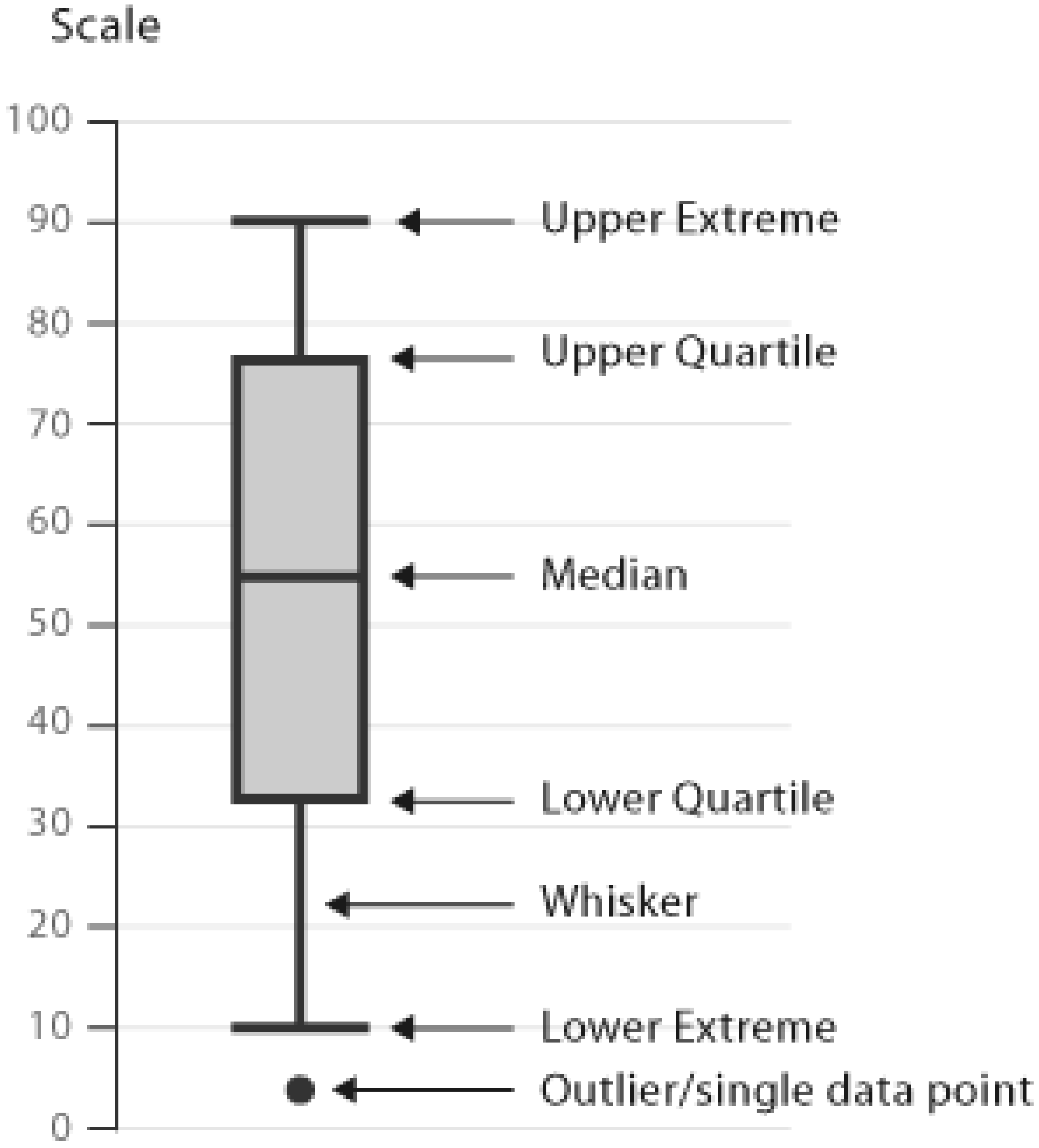
An observation is a potential outlier if it falls more than *1.5 x IQR* below the first or more than *1.5 x IQR* above the third quartile.

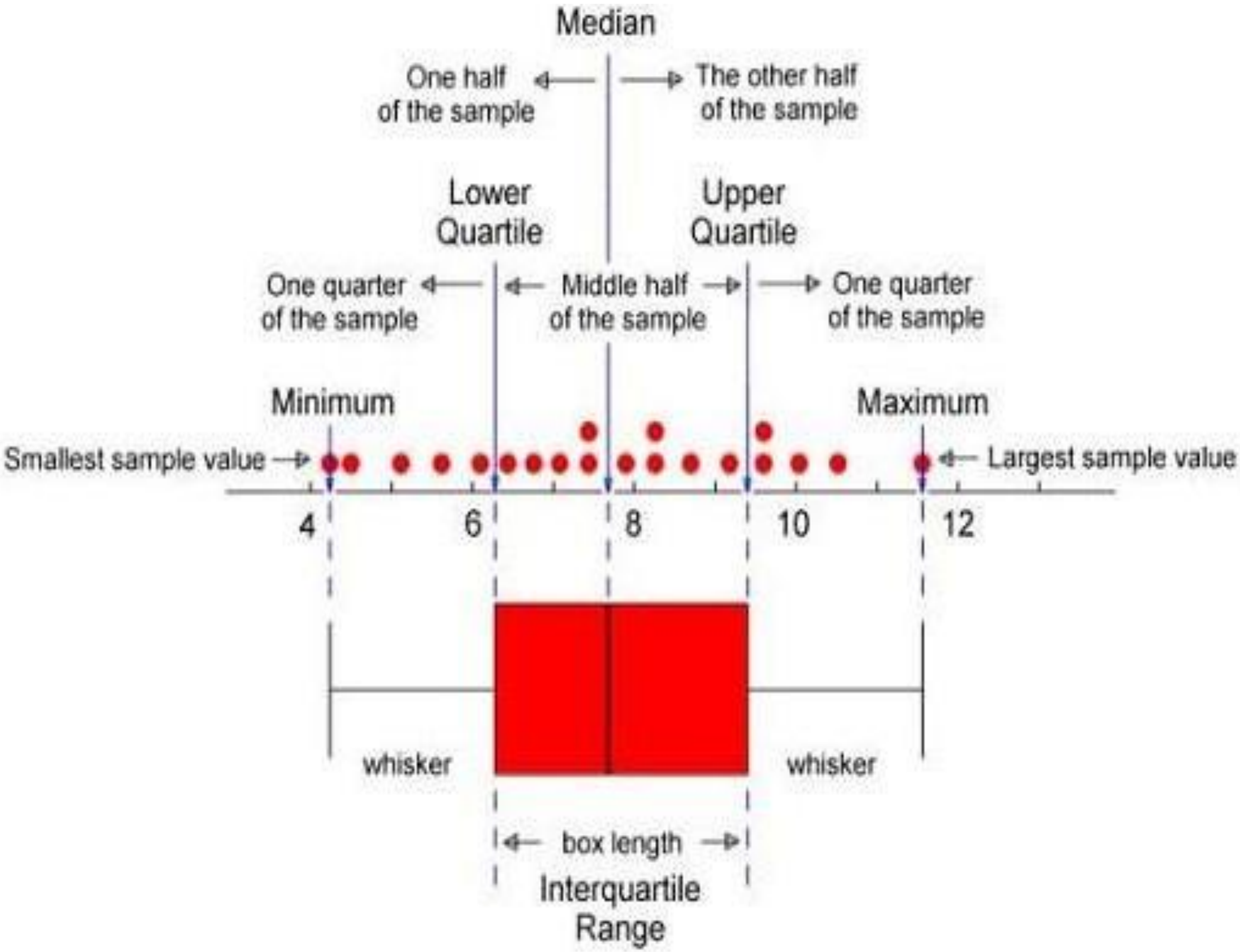


Box Plot

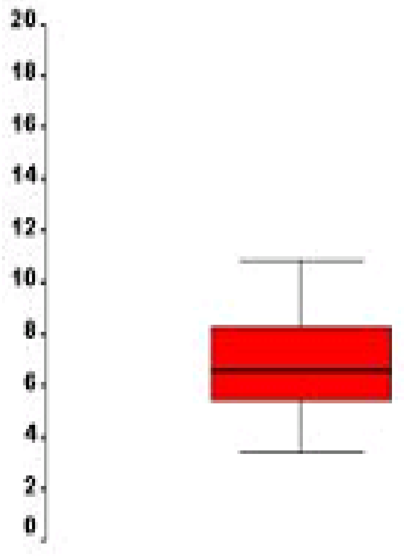
Amount of sleep in past 24 hours
of Spring 1998 Stat 250 Students



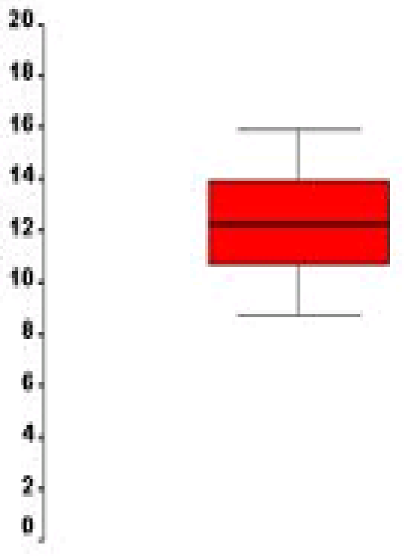

Scale
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme
Upper Quartile
Median
Lower Quartile
Whisker
Lower Extreme
Outlier/single data point


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 Boxplot as an Indicator of Centrality

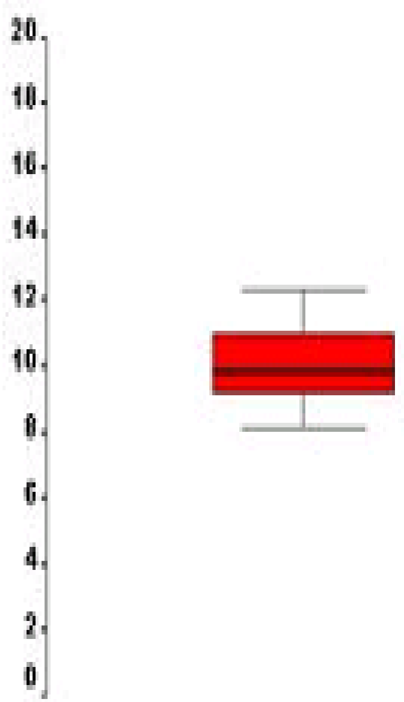


The boxplot of a sample of 20 points from a population centred on 7.

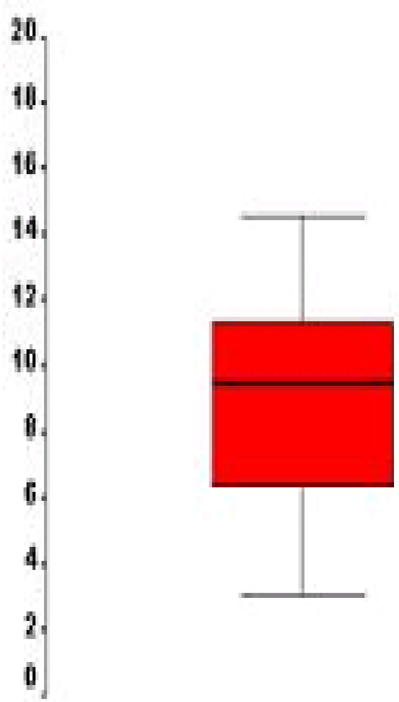


The boxplot of a sample of 20 points from a population centred on 12.

The Boxplot as an Indicator of Spread

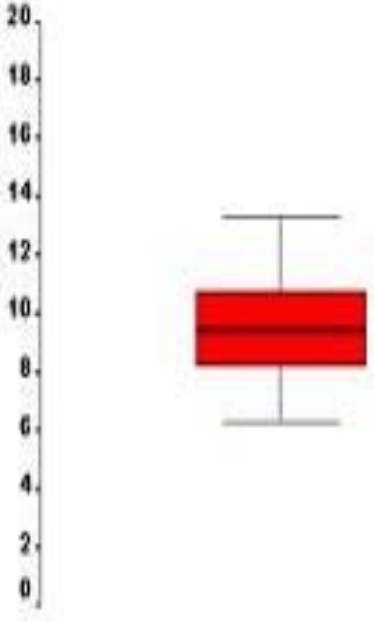


The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 1.

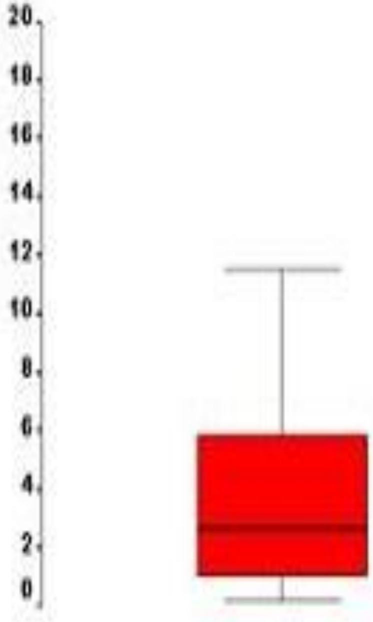


The boxplot of a sample of 20 points from a population centred on 10 with standard deviation 3.

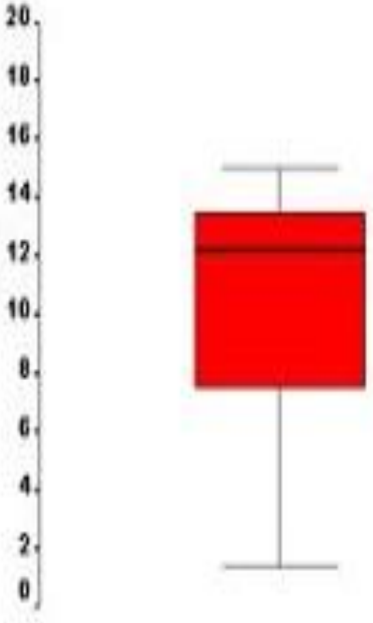
The Boxplot as an Indicator of Symmetry



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.

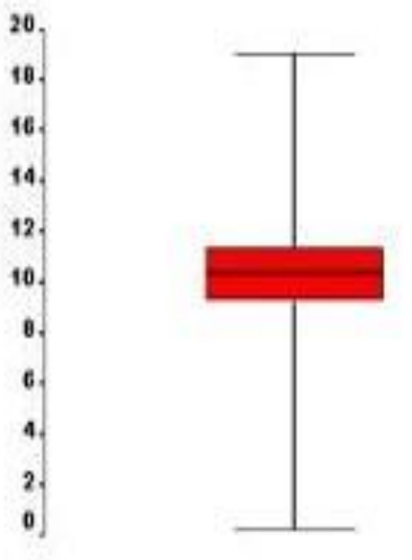


The boxplot of a sample of 20 points from a population which is skewed to the right. The top whisker is much longer than the bottom whisker and the line is gravitating towards the bottom of the box.

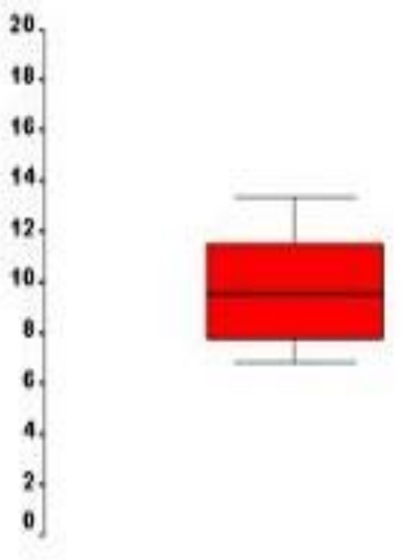


The boxplot of a sample of 20 points from a population which is skewed to the left. The bottom whisker is much longer than the top whisker and the line is rising to the top of the box.

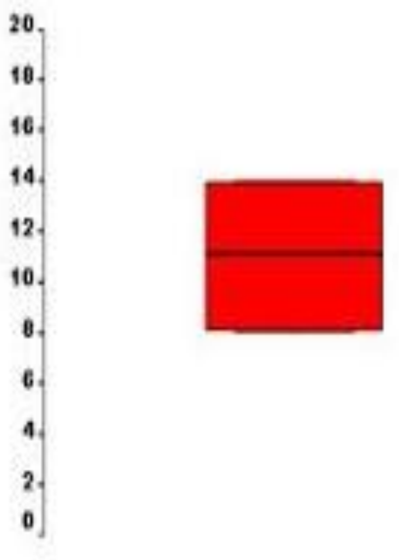
The Boxplot as an Indicator of Tail Length



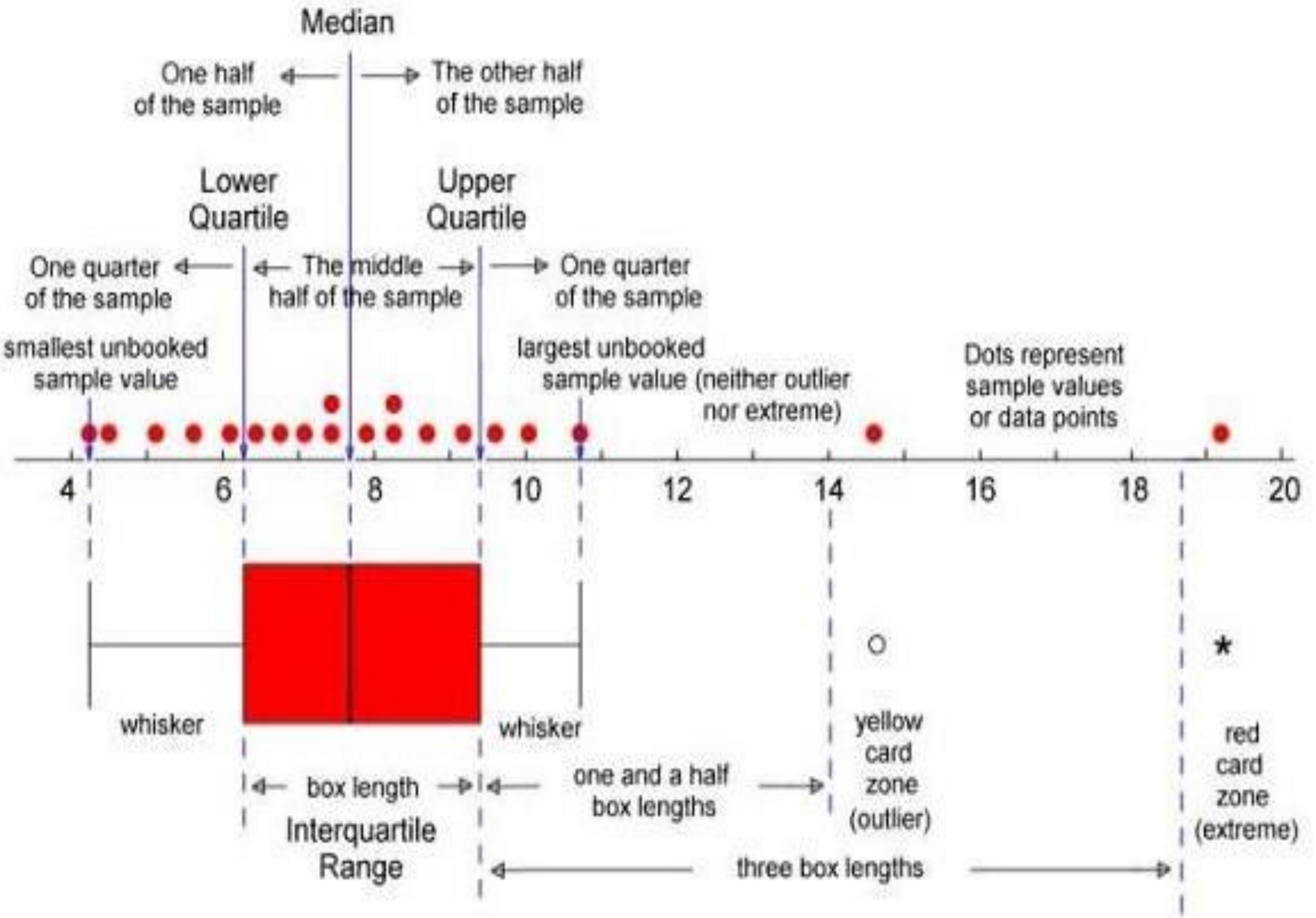
The boxplot of a sample of 20 points from a population with long tails. The length of the whiskers far exceeds the length of the box. (A well proportioned tail would give rise to whiskers about the same length as the box, or maybe slightly longer.)



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.



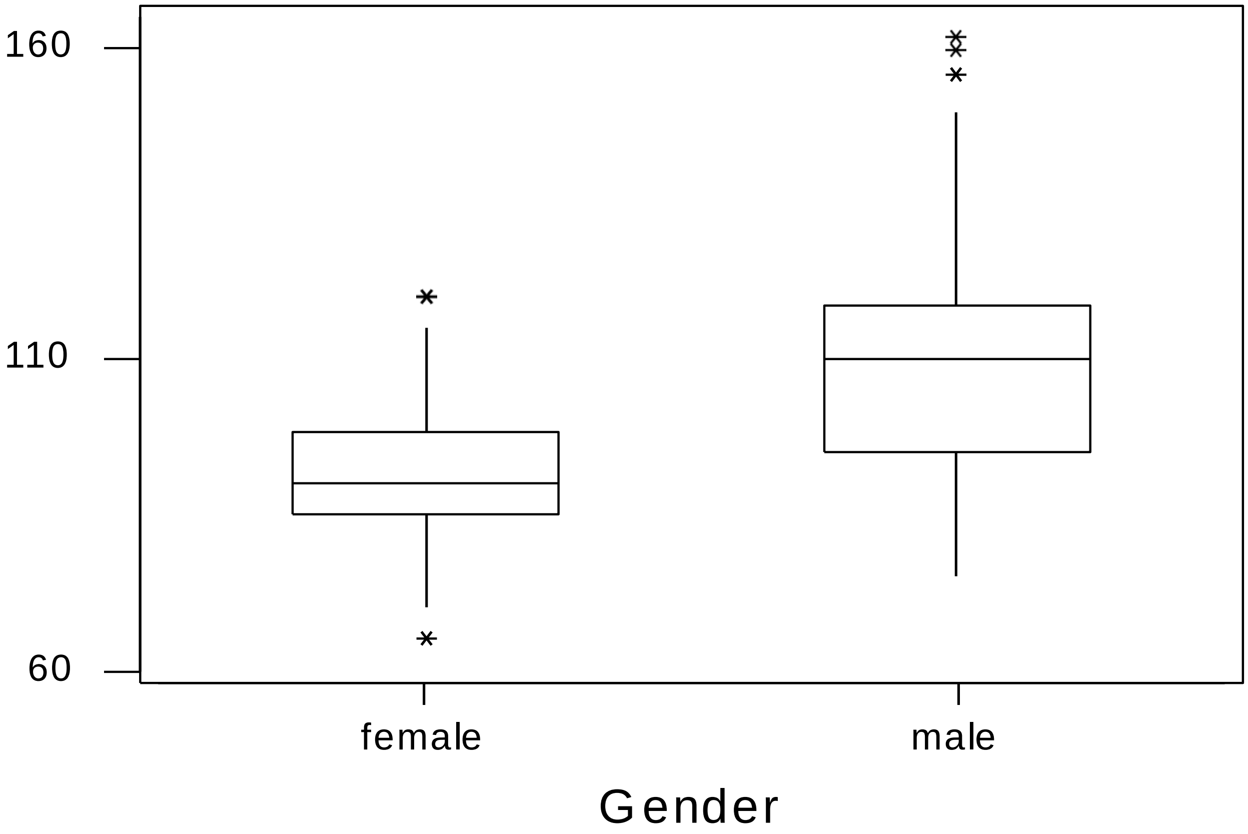
The boxplot of a sample of 20 points from a population with extremely short tails (actually a U-shaped population, with a dip in the middle rather than a hump). The whiskers are absent.

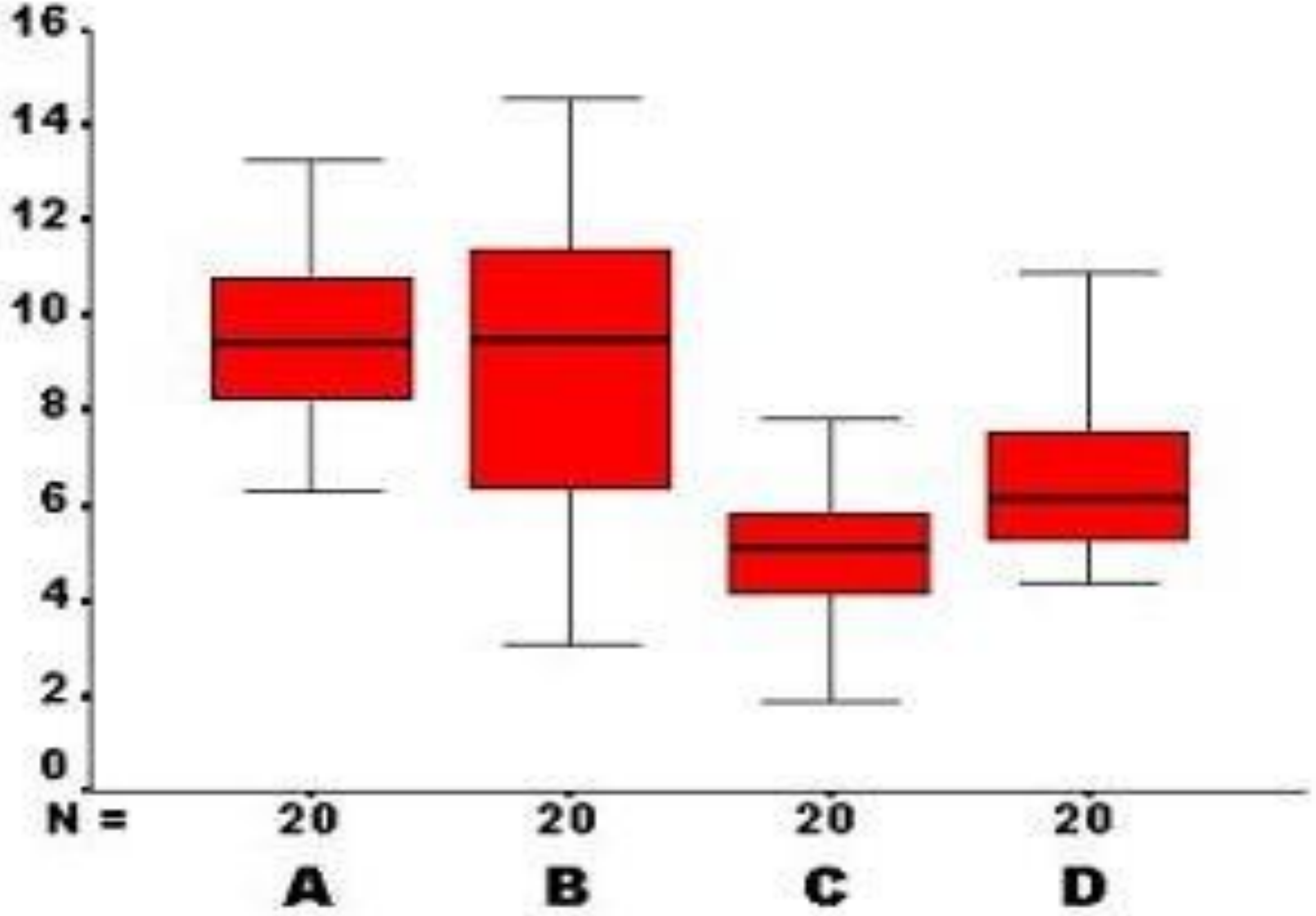

Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
The middle half of the sample
One quarter of the sample
smallest unbooked sample value
largest unbooked sample value (neither outlier nor extreme)
Dots represent sample values or data points
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

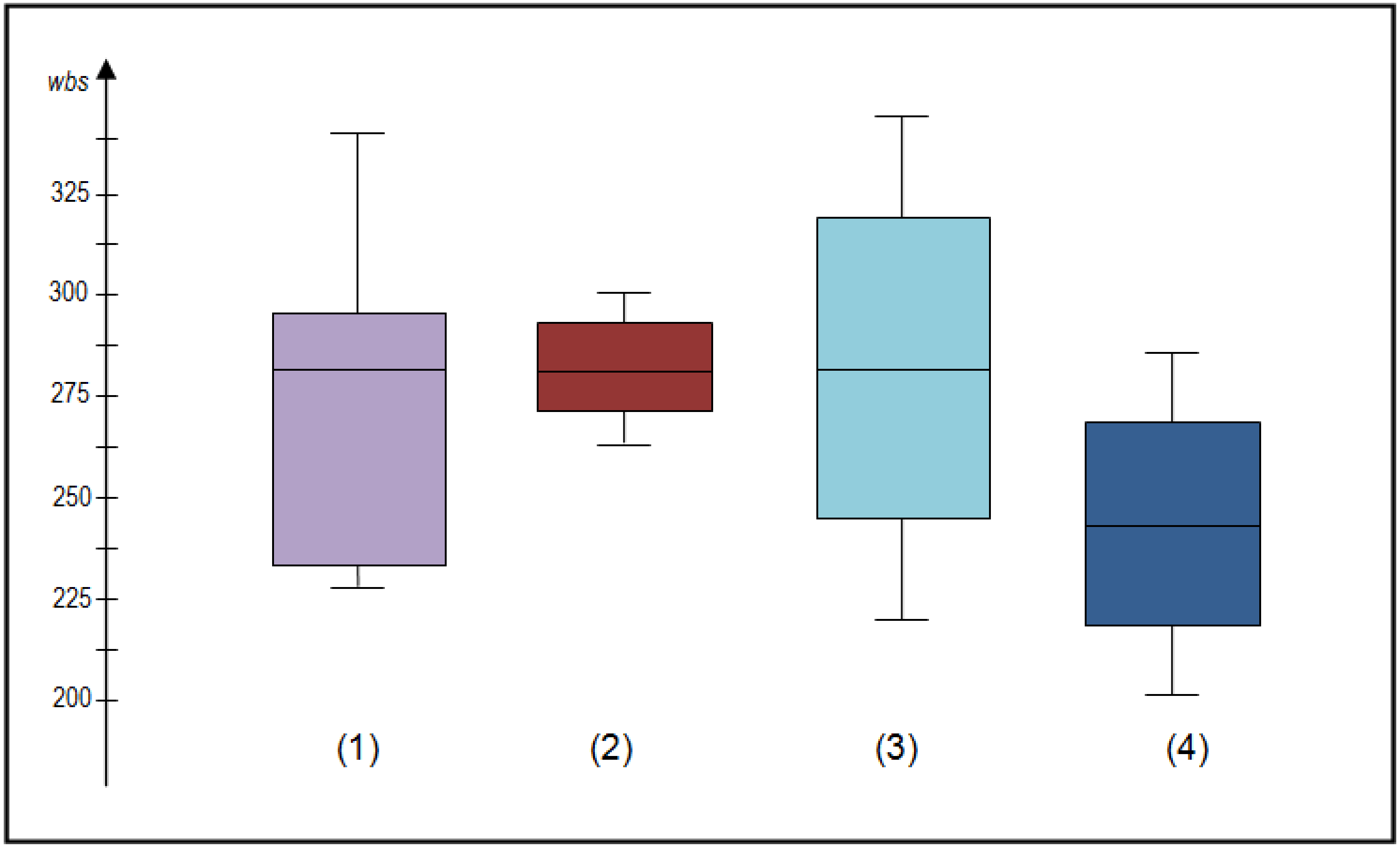
Fastest Ever Driving Speed

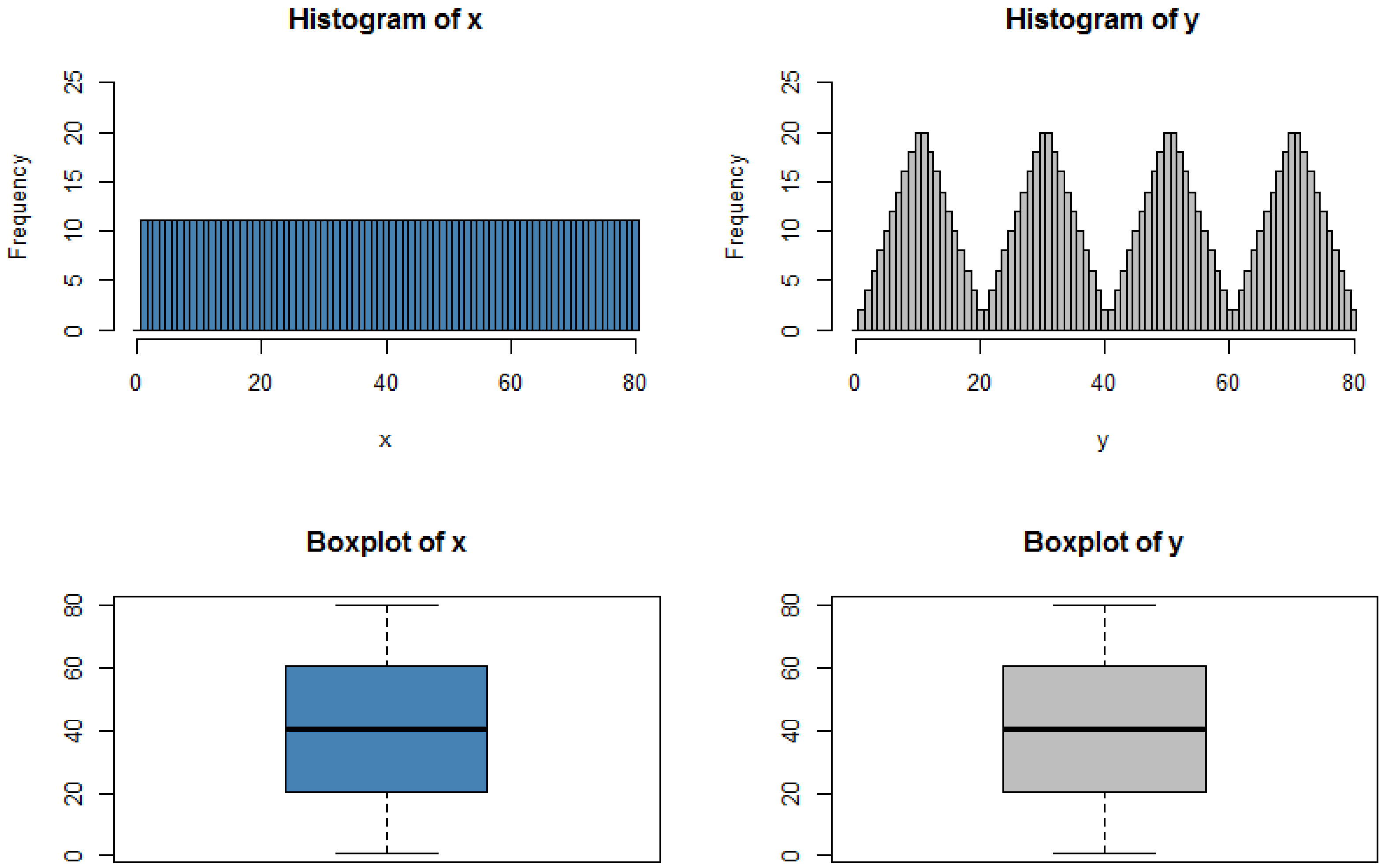
226 Stat 100 Students, Fall 1998




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

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



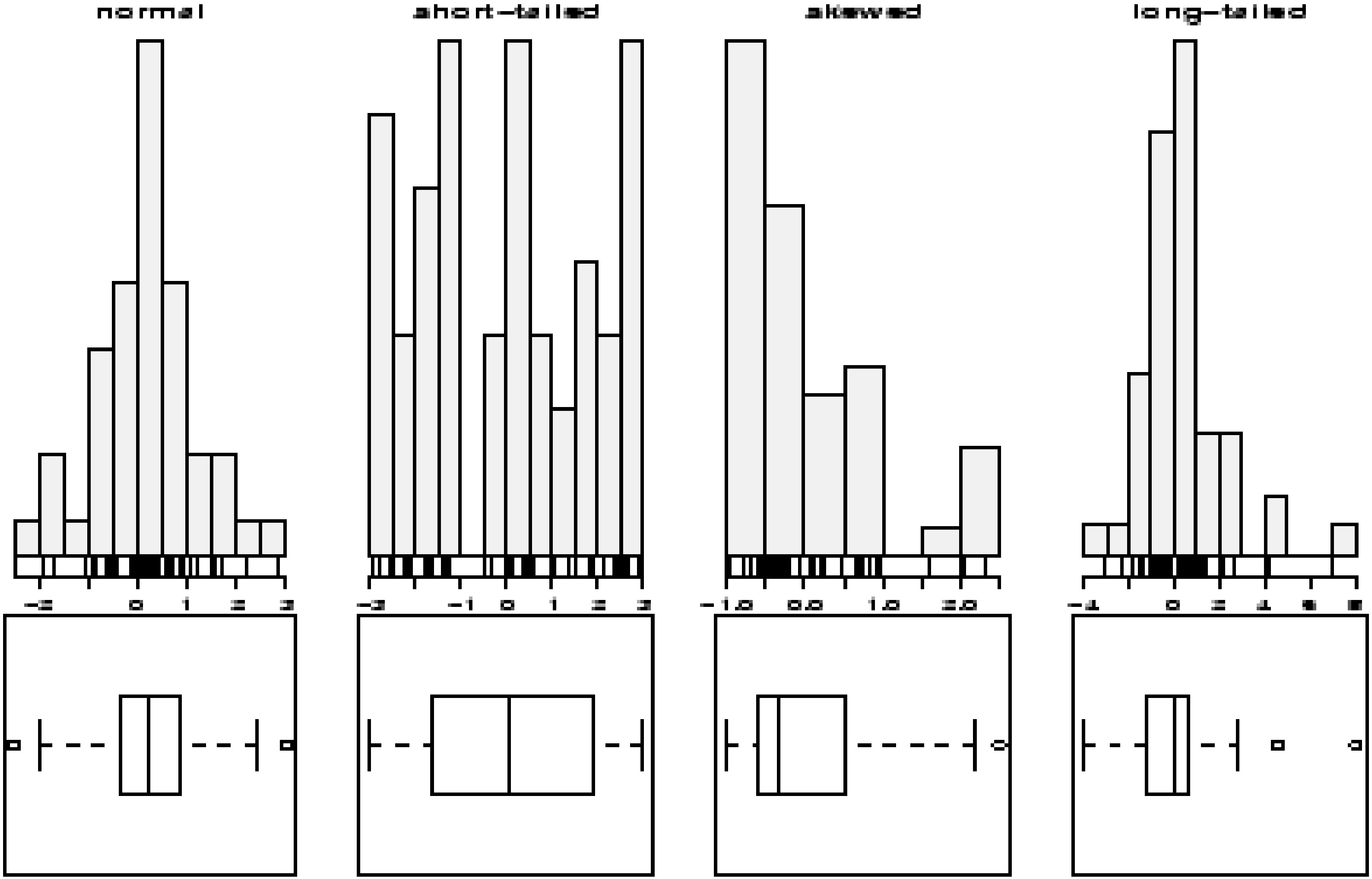


Histogram of x

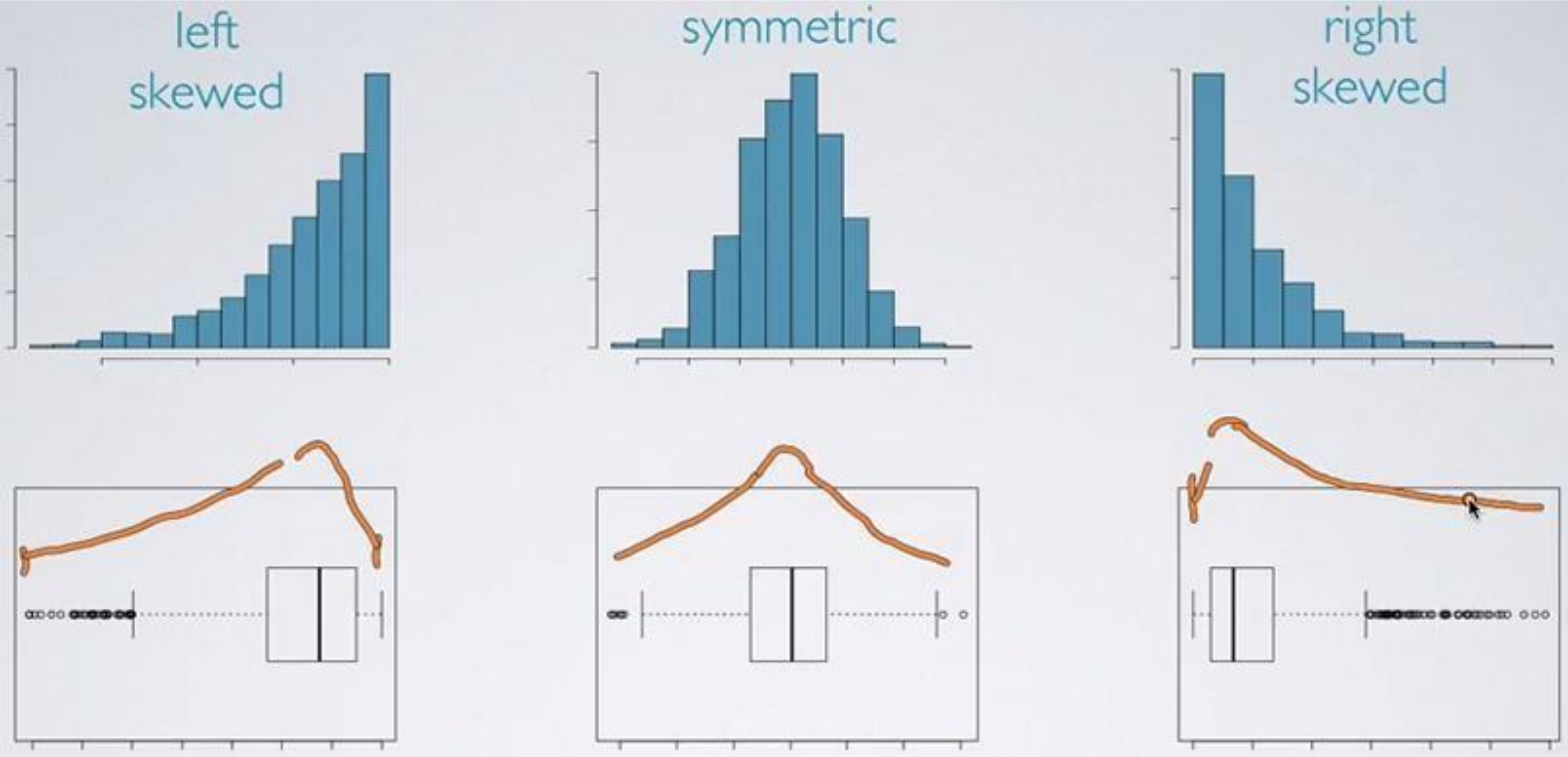
Histogram of y

Boxplot of x

Boxplot of y



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



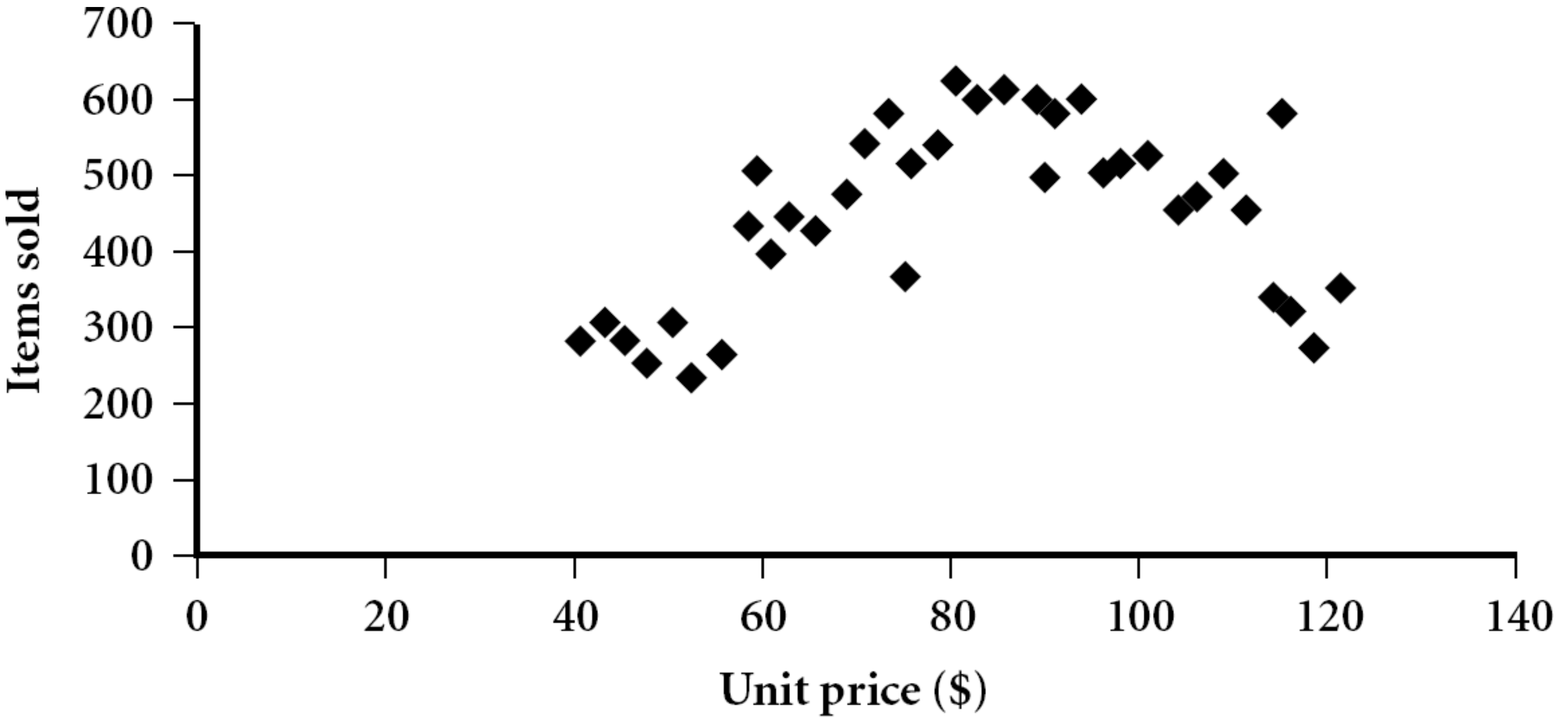
left skewed

symmetric

right skewed

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.

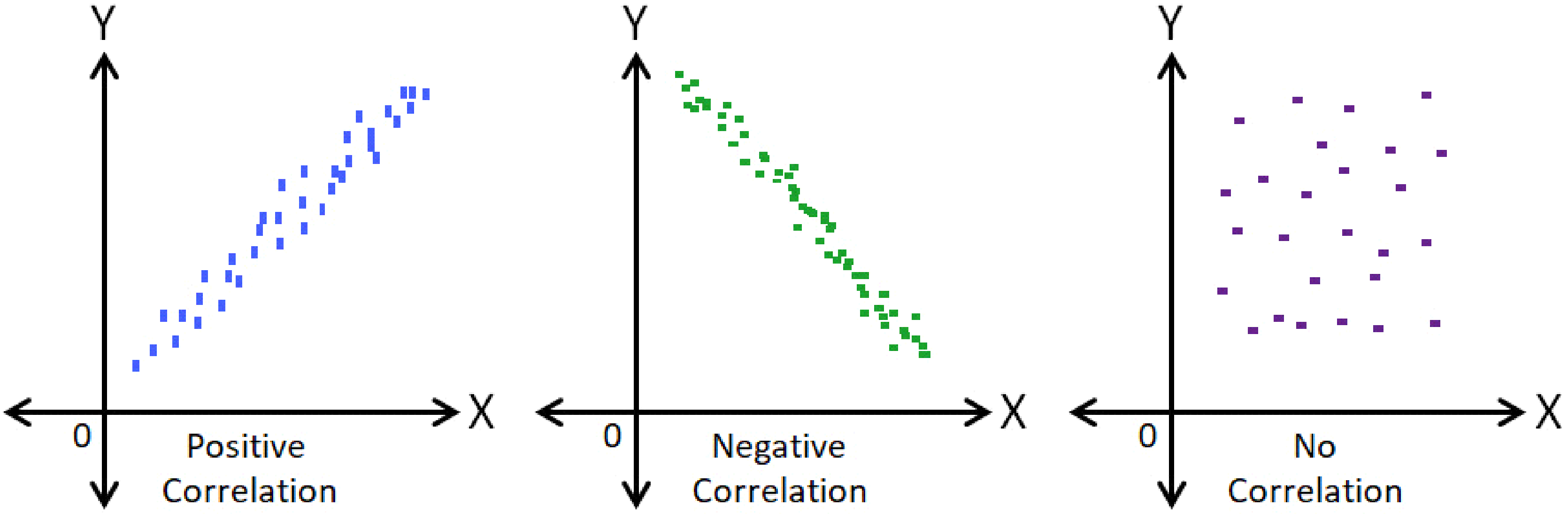
- There are many types of coefficients of correlation in scatter points, most popular one is

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

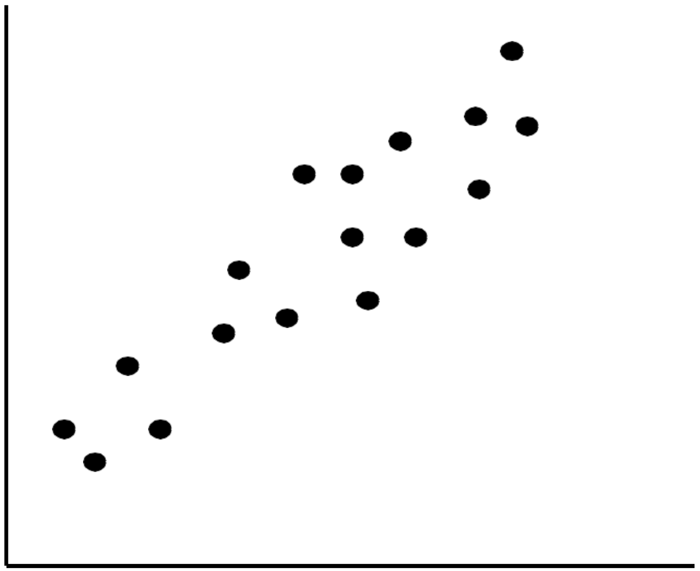
- Pearson’s Co-efficient of correlation
- x – value of data point on x-axis
- y – value of data point on y-axis
- n – no of datapoints

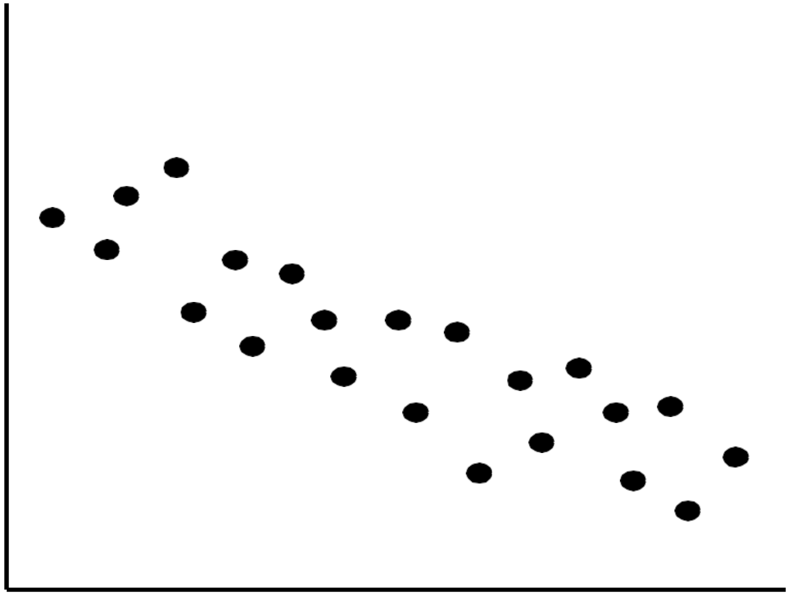
- Pearson’s co-efficient of correlation:
- co-eff > 0 : positively correlated
- co-eff < 0 : negatively correlated
- co-eff = 0 : no correlation
- +1 or -1, mean perfect correlation between the data points

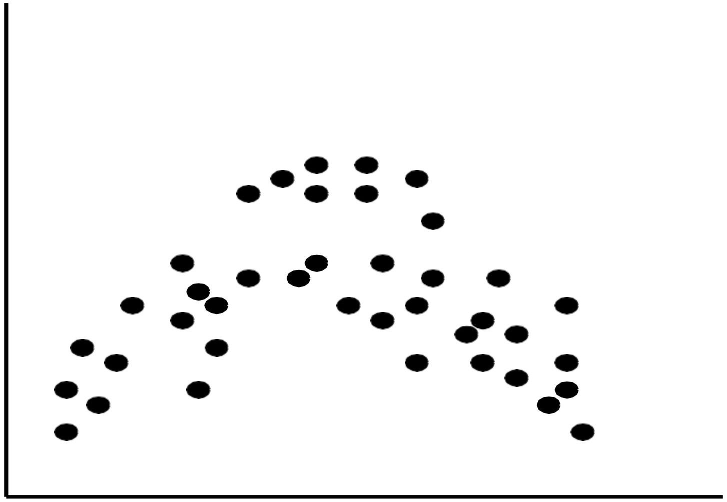
Scatter Plots & Correlation Examples



Positively and Negatively Correlated Data

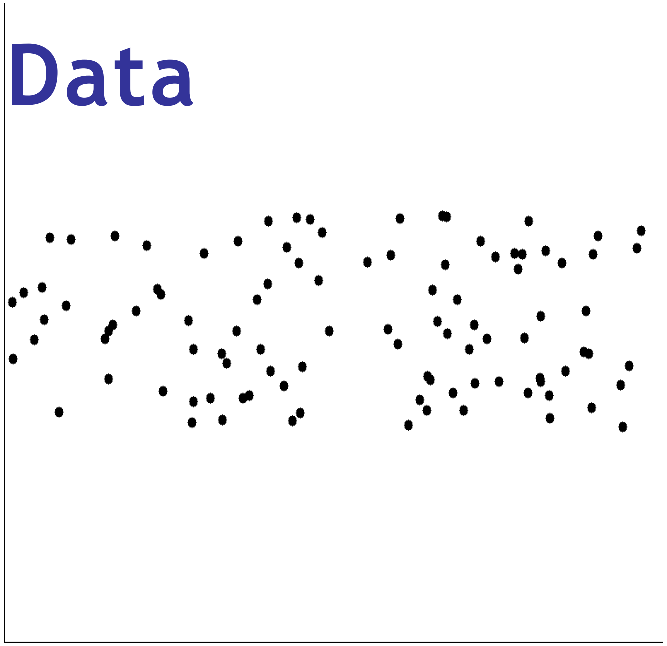


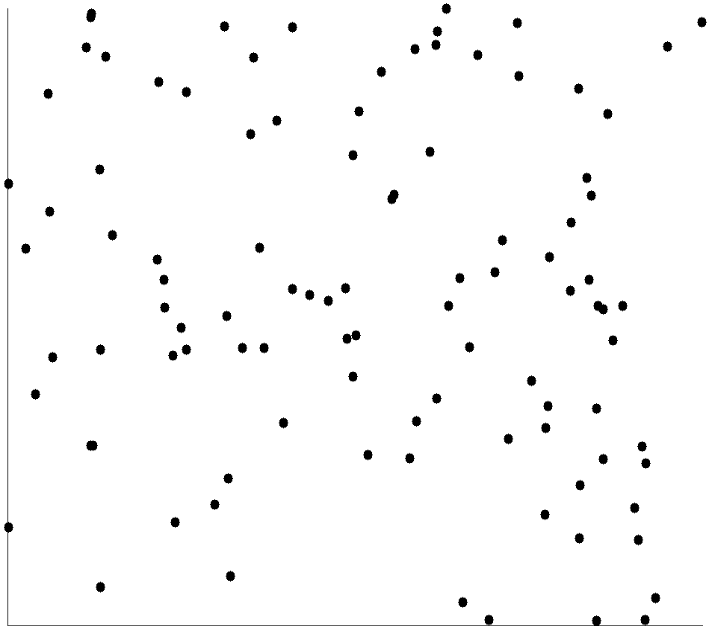


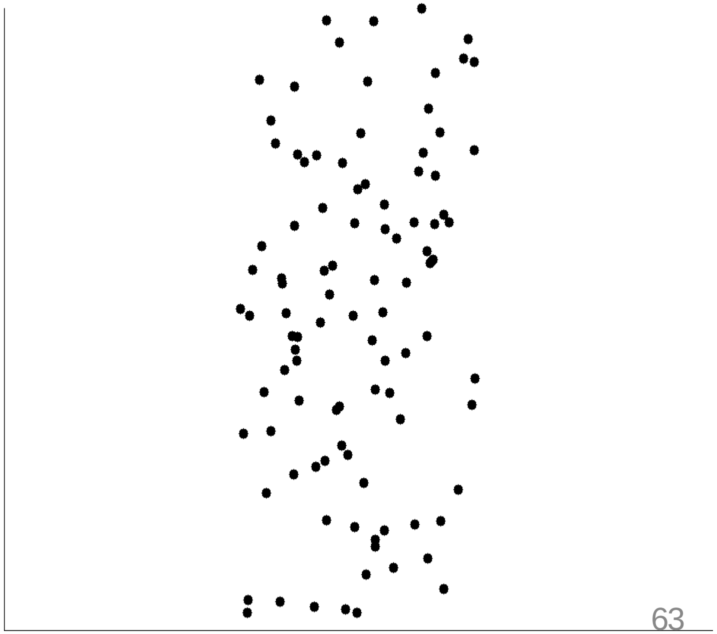


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data

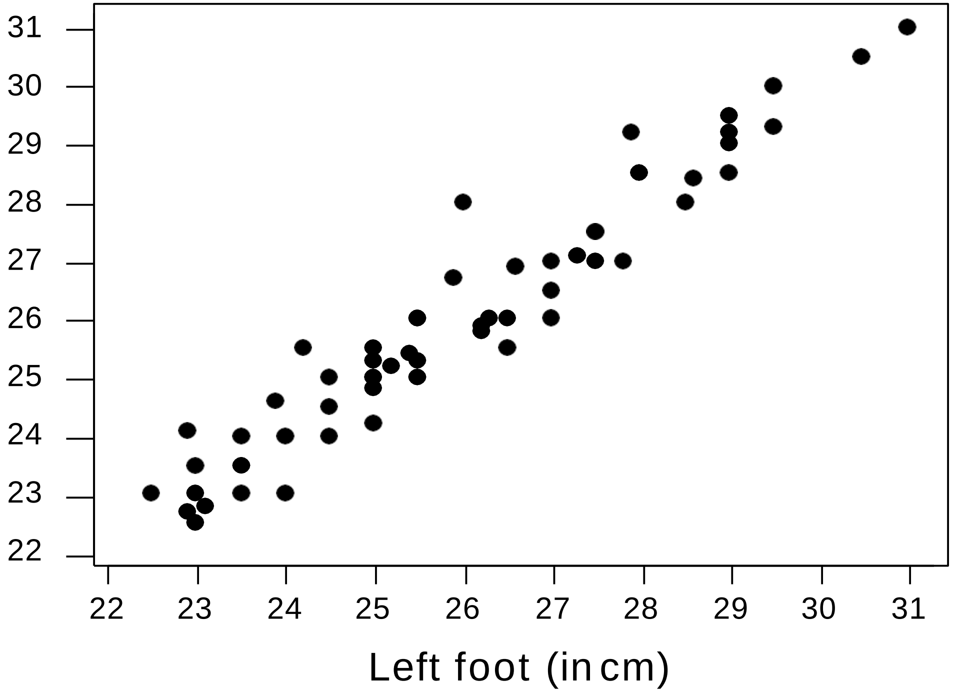






Scatter Plots

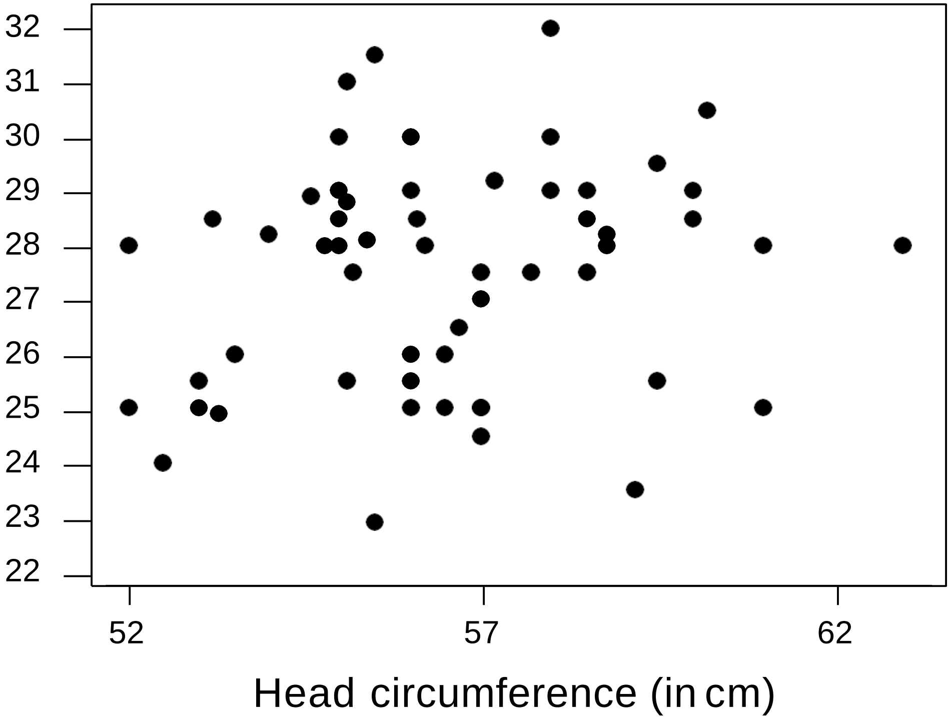
Foot sizes of Spring 1998 Stat 250 students



n=88 students

No relationship

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

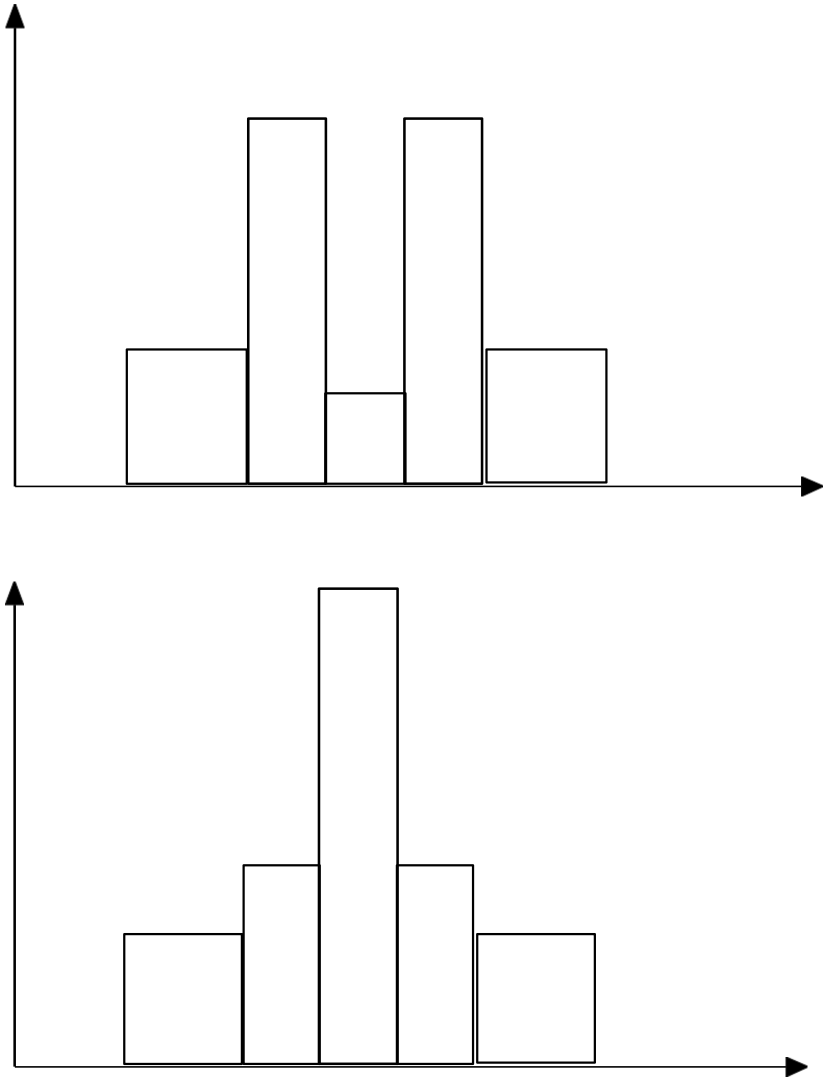


n=89 students

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



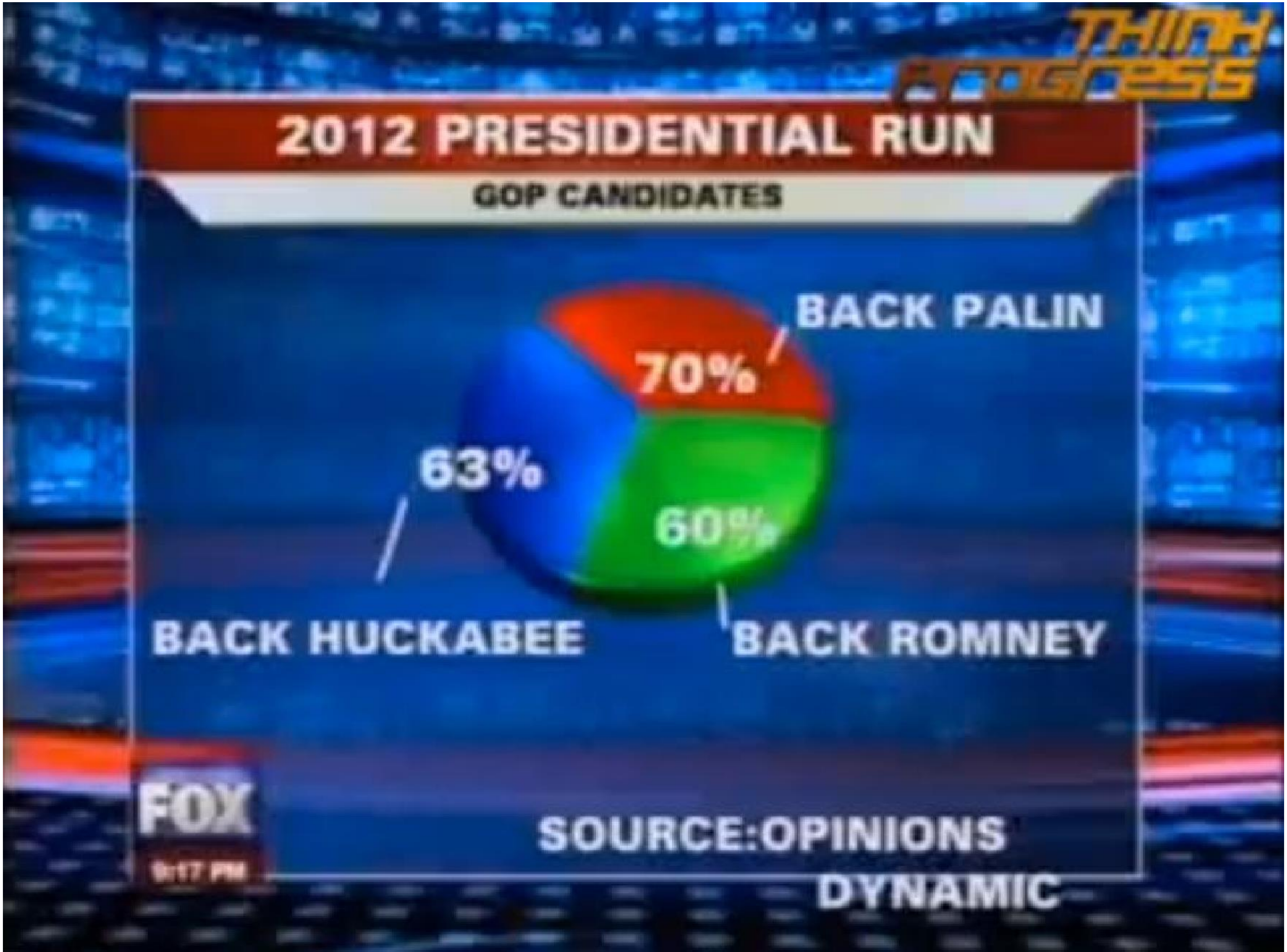
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Good Vs Bad Visualization

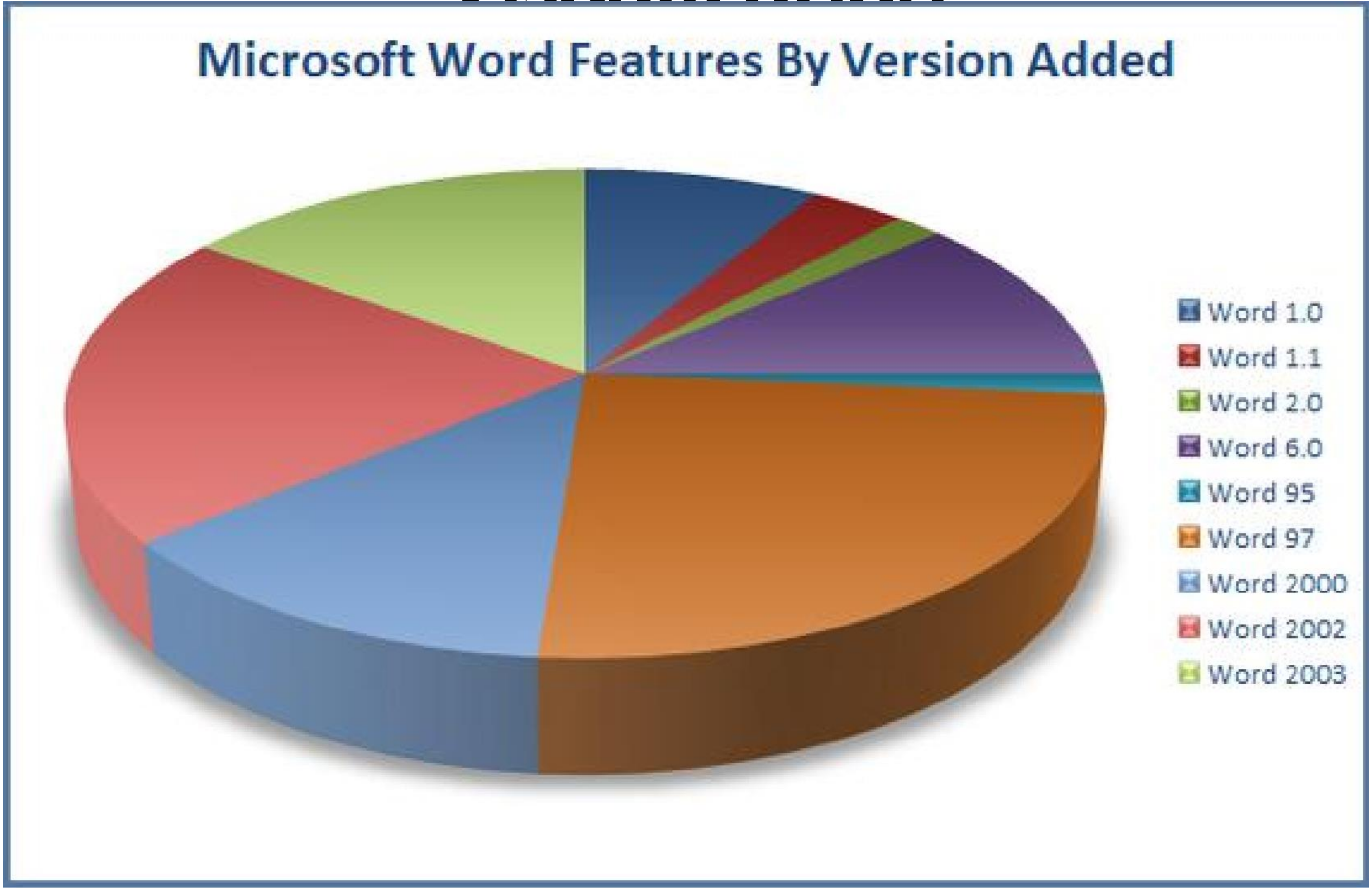
5 common mistakes that lead to bad data visualization

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

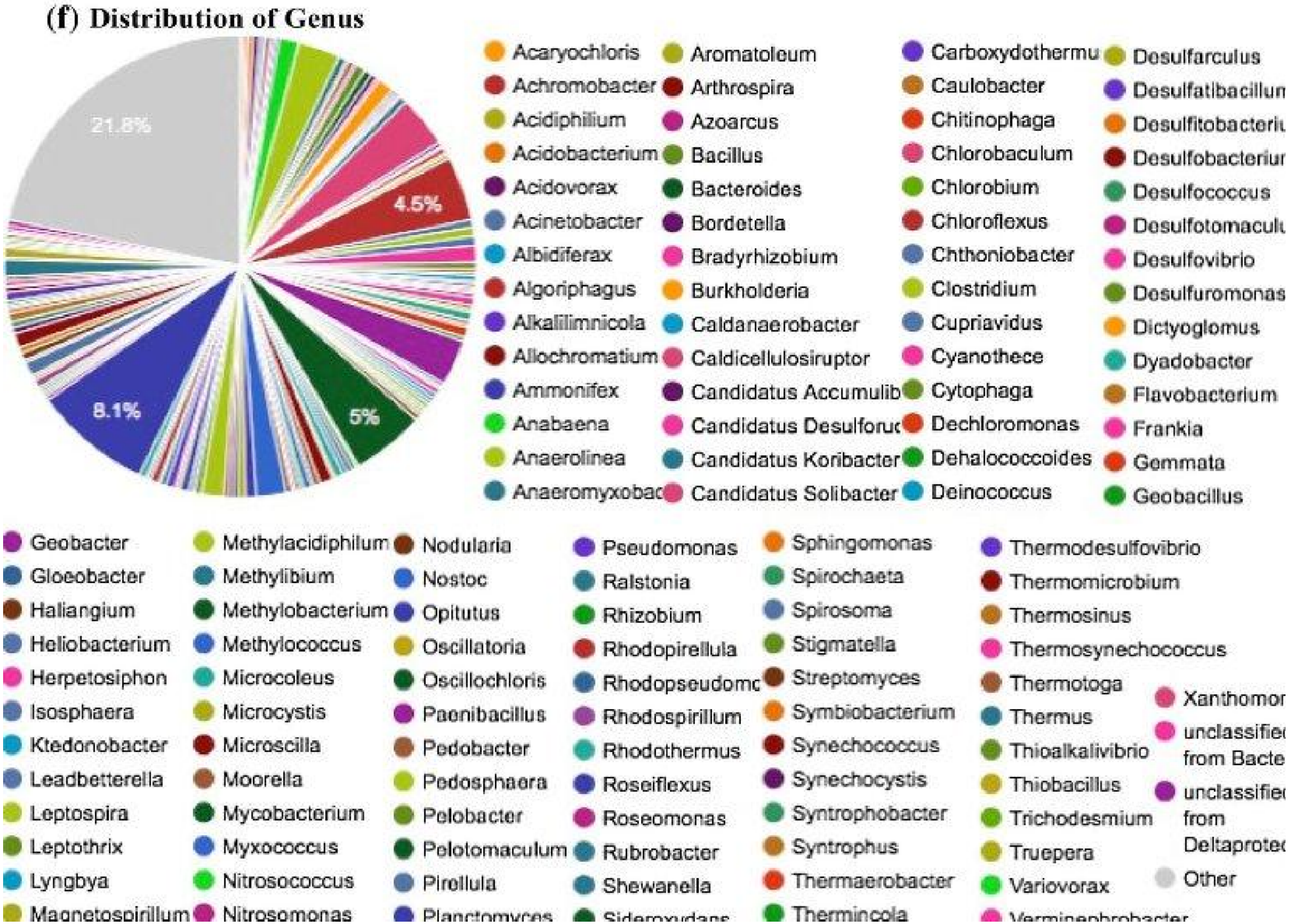
Bad data


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

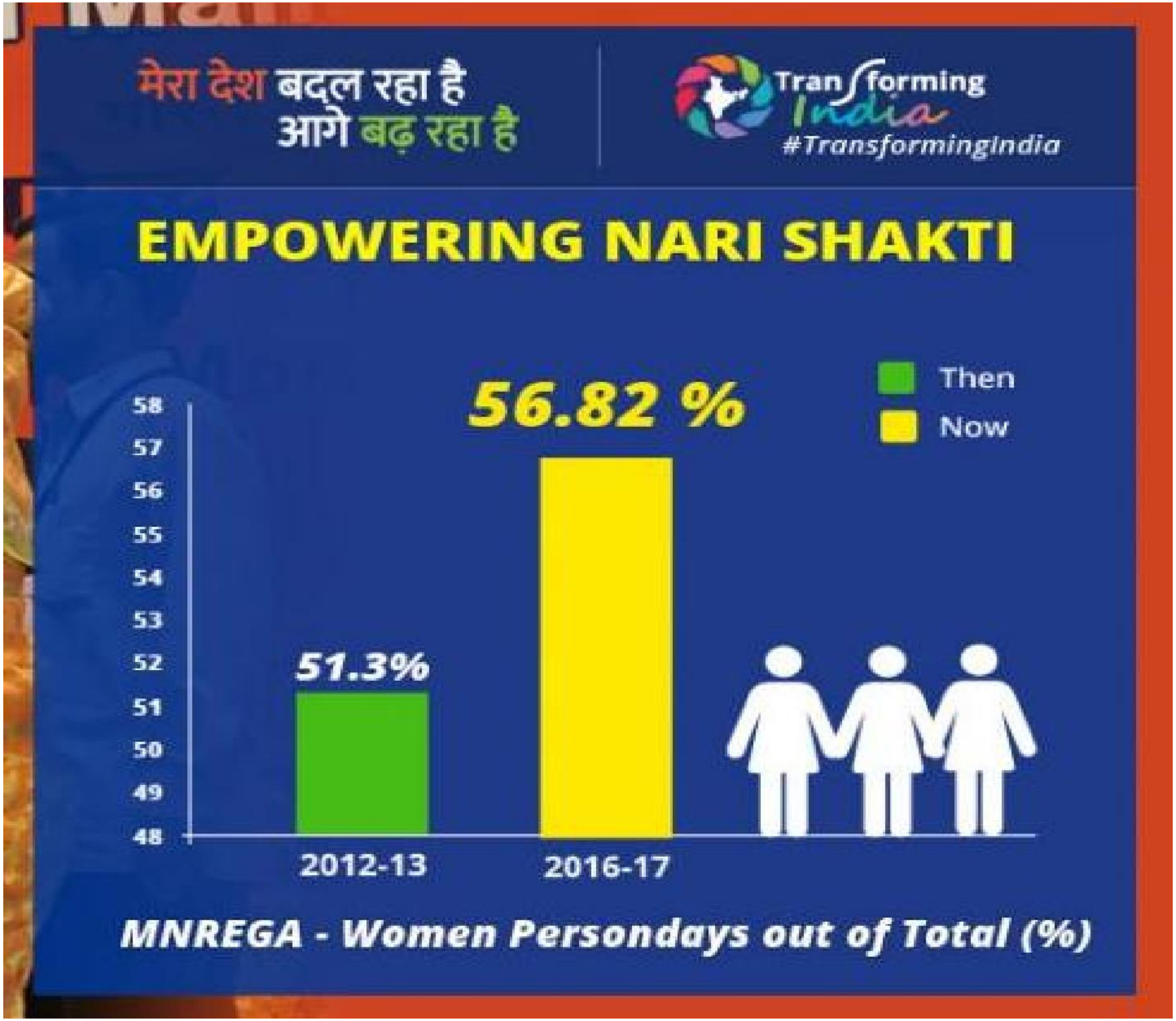


Too Much Color or Information

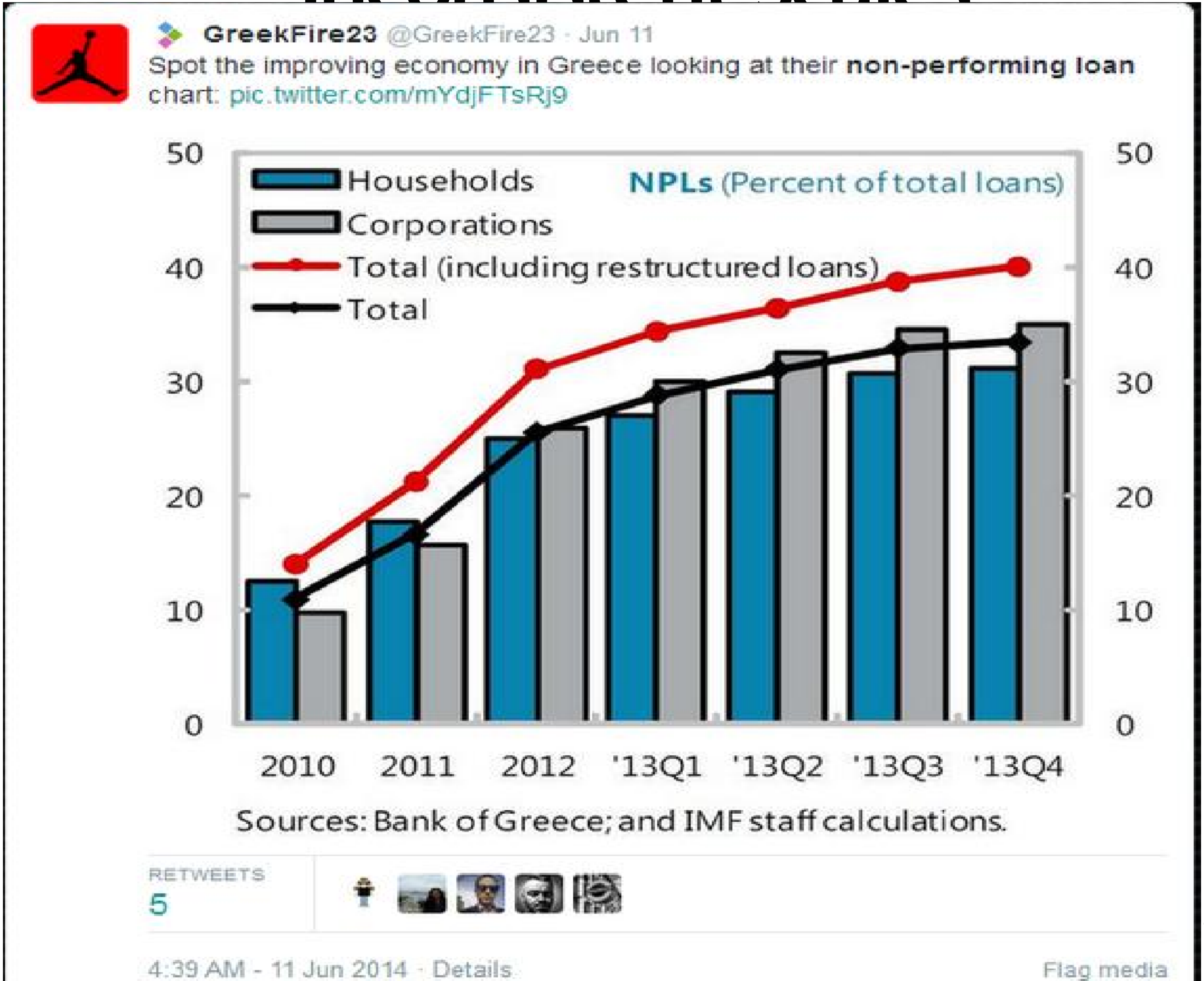


(f) Distribution of Genus

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

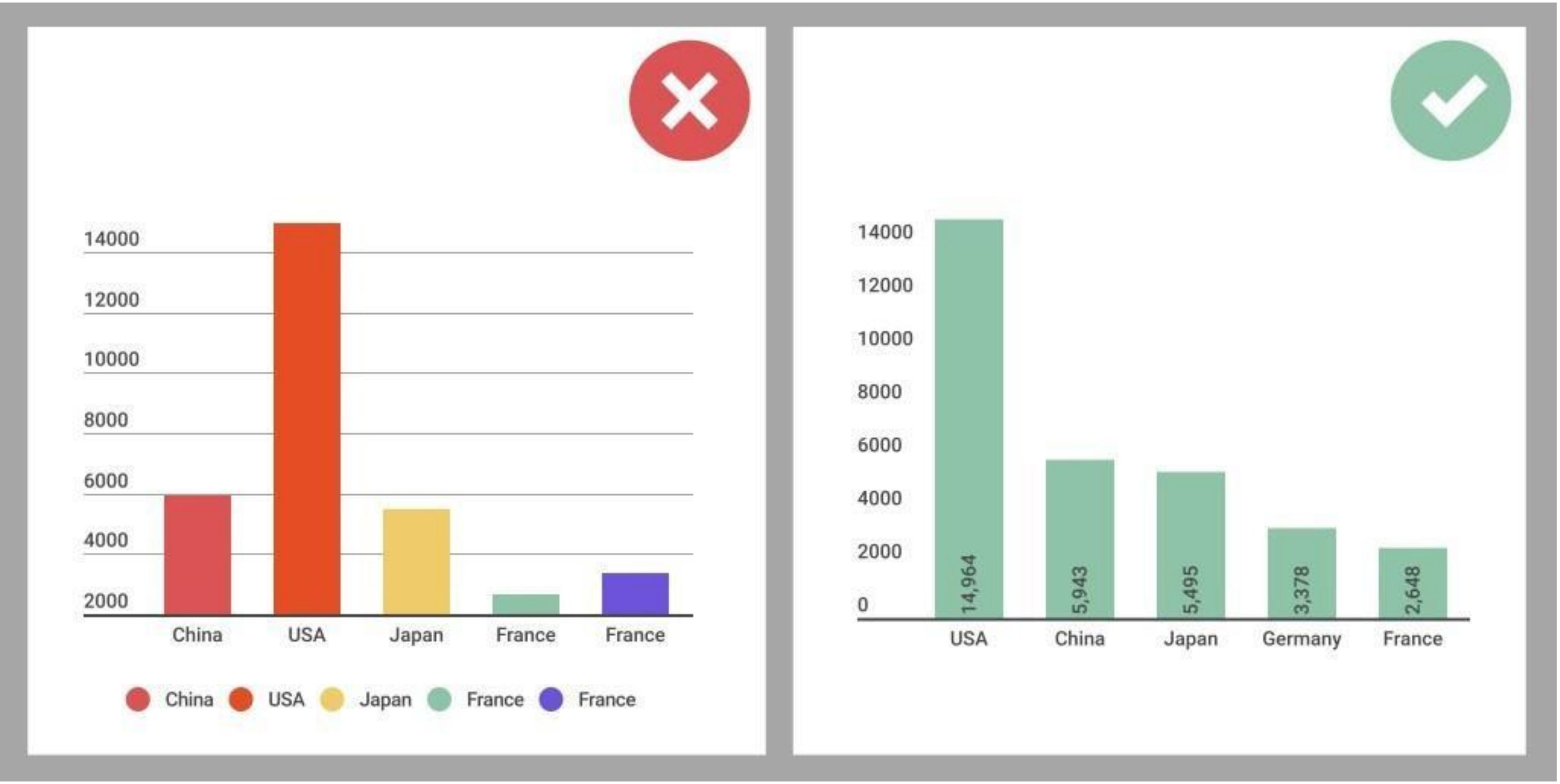
Sources: Bank of Greece; and IMF staff calculations.

RETWEETS
5

4:39 AM - 11 Jun 2014 · Details

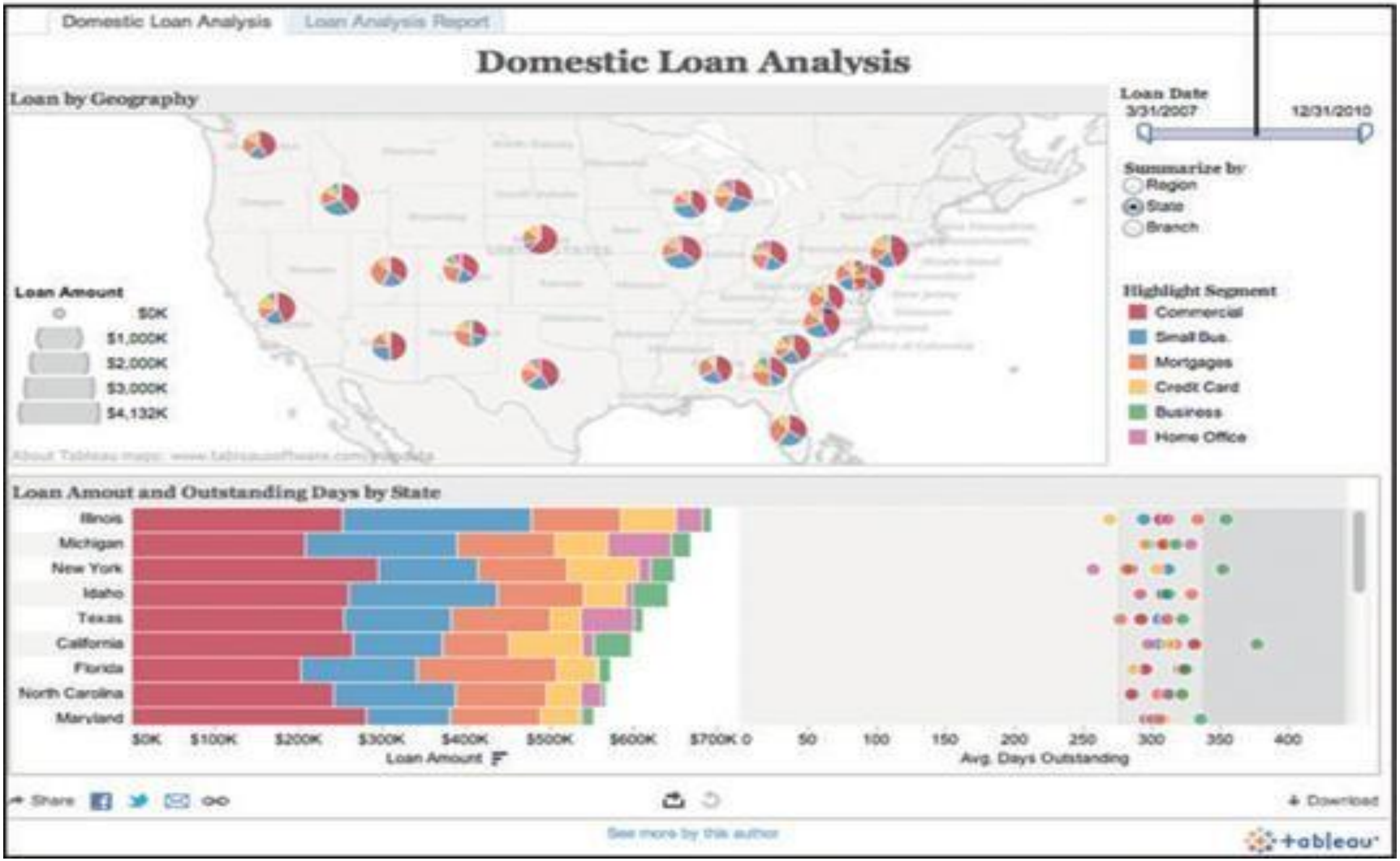
Flag media

Good Vs Bad Visualization



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

Slider for interactivity



Things that work well:

Color consistency

Simplicity

Interactivity

Things that don‘t work:

Chart choice: The small pie charts that are overlaid on the map are of little value.

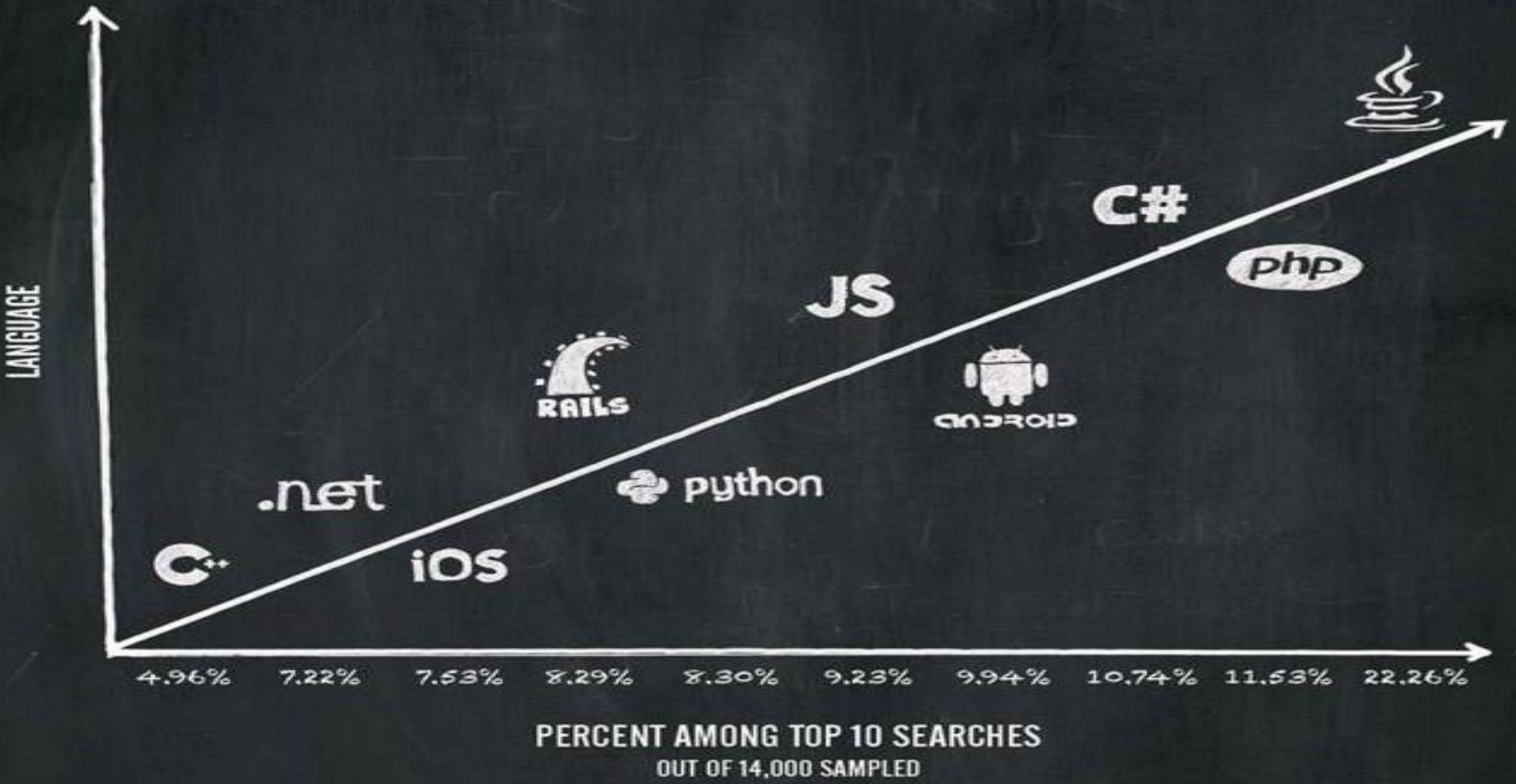
Color choice: The abundant use of red, blue, and orange are misleading, especially in the stacked bar chart at the bottom of the data viz.

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

readwrite presents

TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013

INFORMATION COMPILED BY STACK OVERFLOW



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)

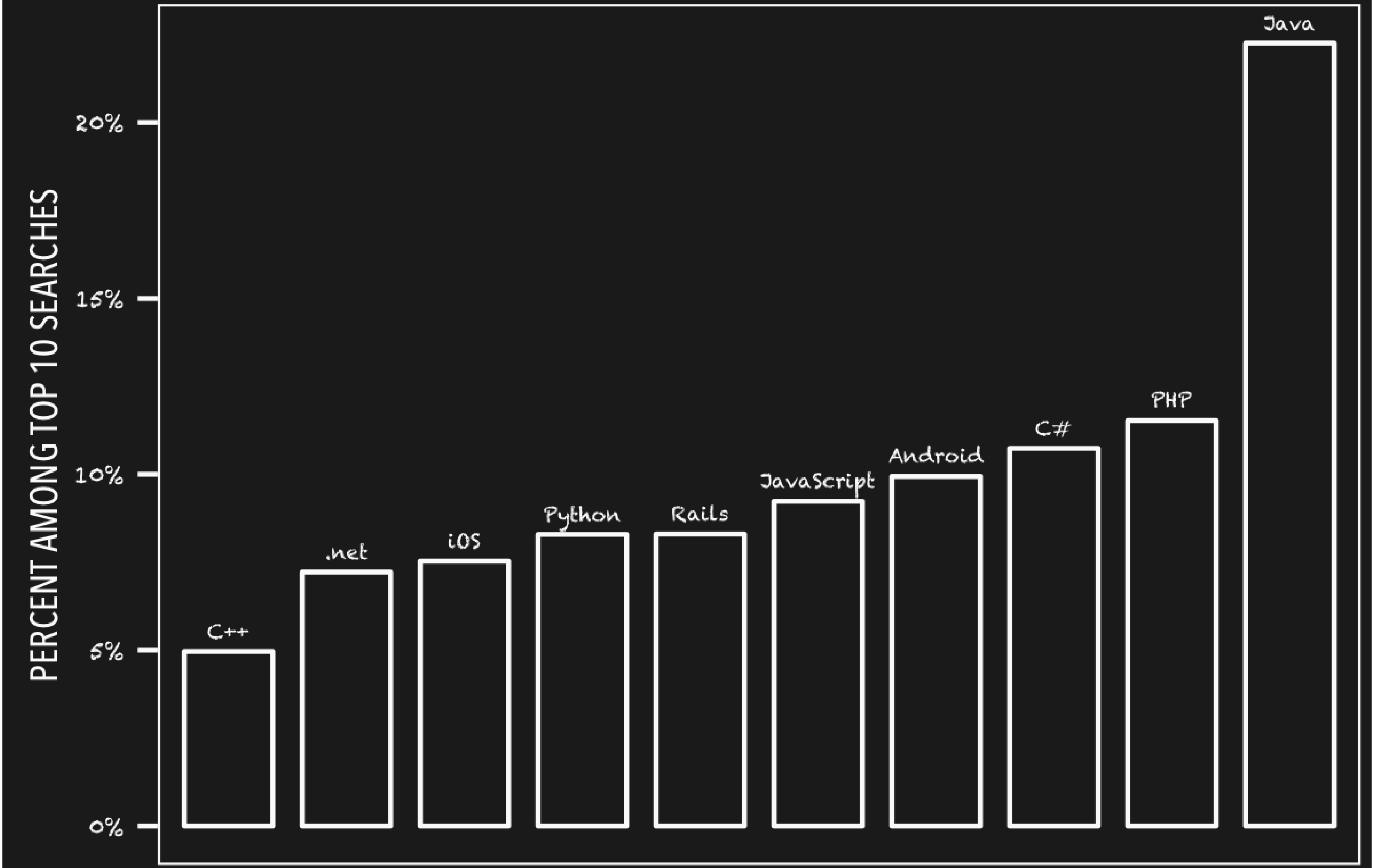
Not all entries on the chart are programming languages. (Android, for example, is an operating system.)

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.

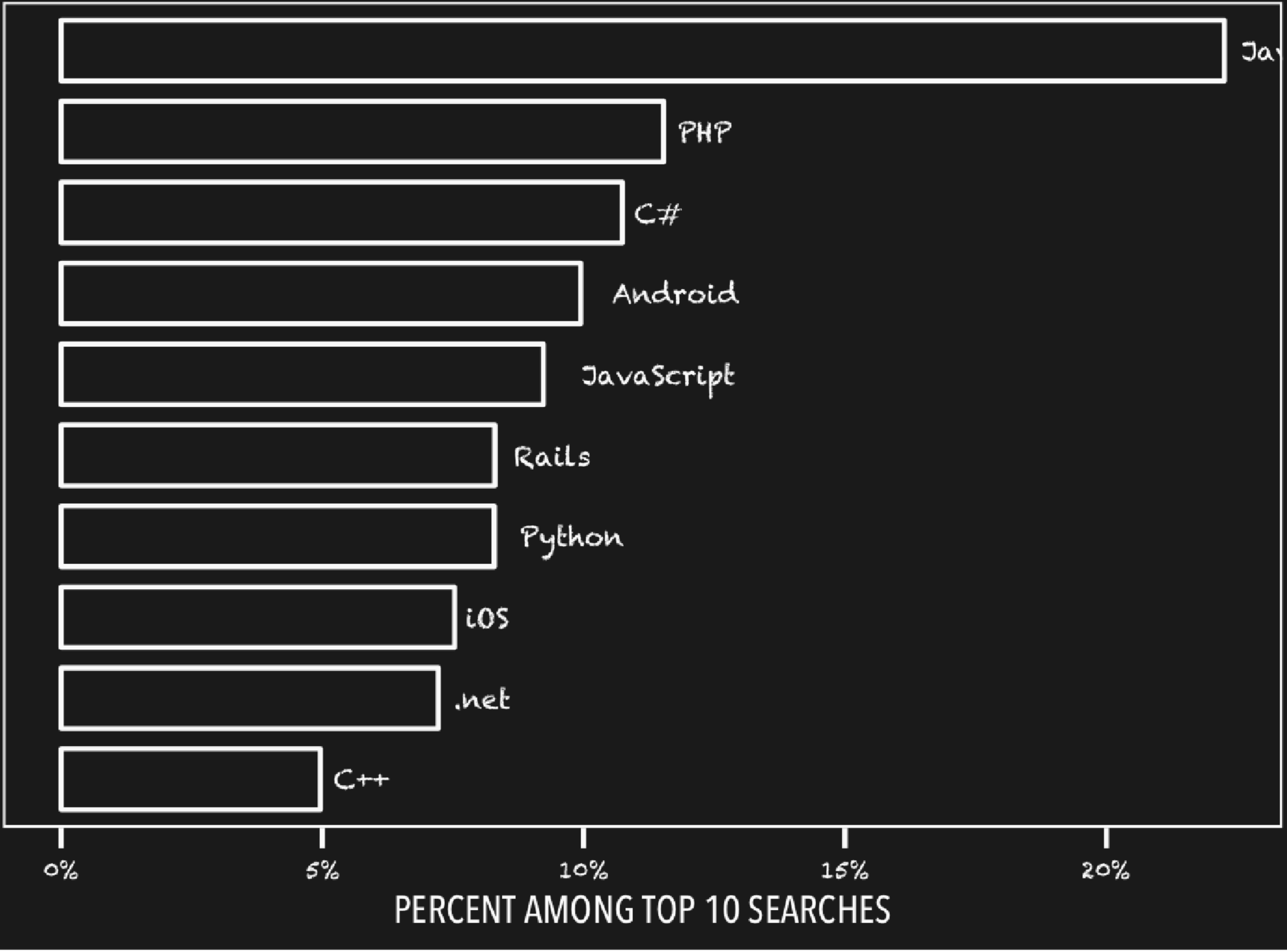
- No relative proportions between the programming languages. We can’t accurately see the increase in language Java has relative toAndroid just by looking at the graph.

- 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.

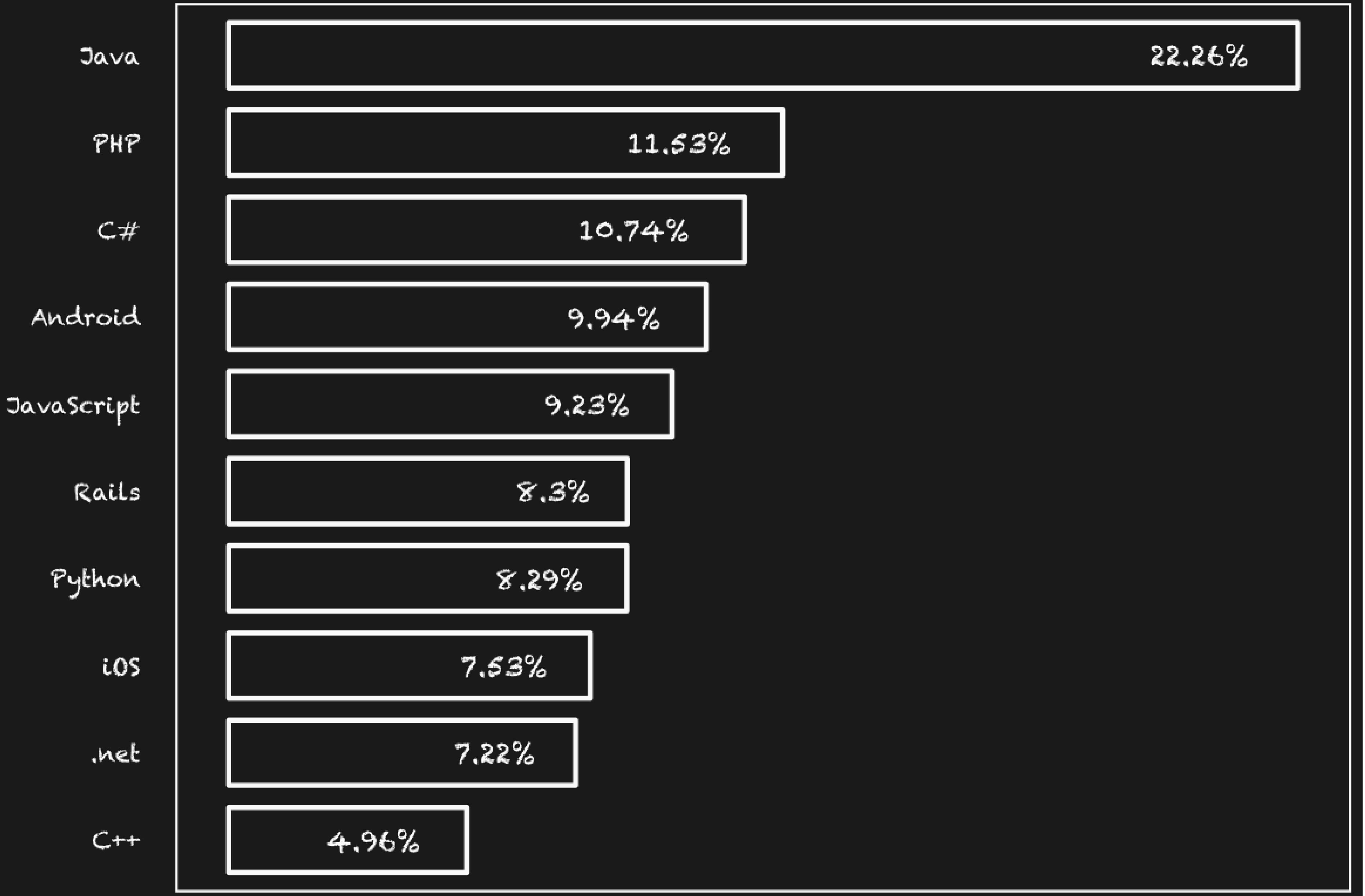
TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013



TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013



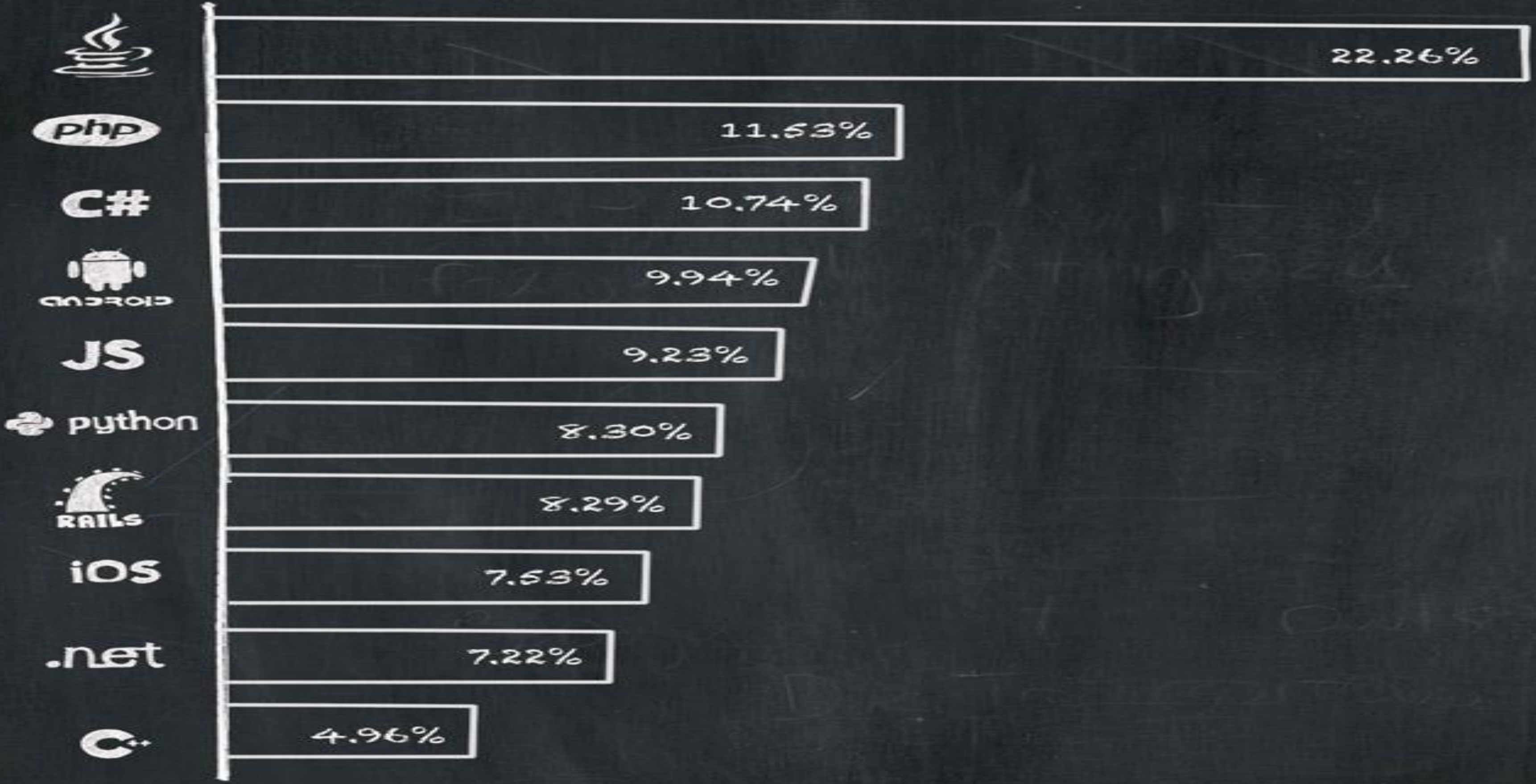
TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013



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

Summary

- 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.to

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.