SIT718 Real World Analytics

Lecturer: Dr Ye Zhu

School of Information Technology
Deakin University

Week 3: Data distributions

Transforming data

Learning aims:

- Understand data visualisation
- Using R to visualise and transform data
- Understand different data distributions

Read Chapter 2 of the reference book (An Introduction to Data Analysis using Aggregation Functions in R by Simon James)
Review: Chapter 4. Distributions Stats Data and Models. De Veaux, Velleman, Bock, 4th edition, Pearson 2016.

Data Visualization

Data visualization involves:

- Creating a summary table for the data.
- Generating charts to help interpret, analyze, and learn from the data.

Uses of data visualization:

- Helpful for identifying data errors.
- Reduces the size of your data set by highlighting important relationships and trends in the data.

Tables

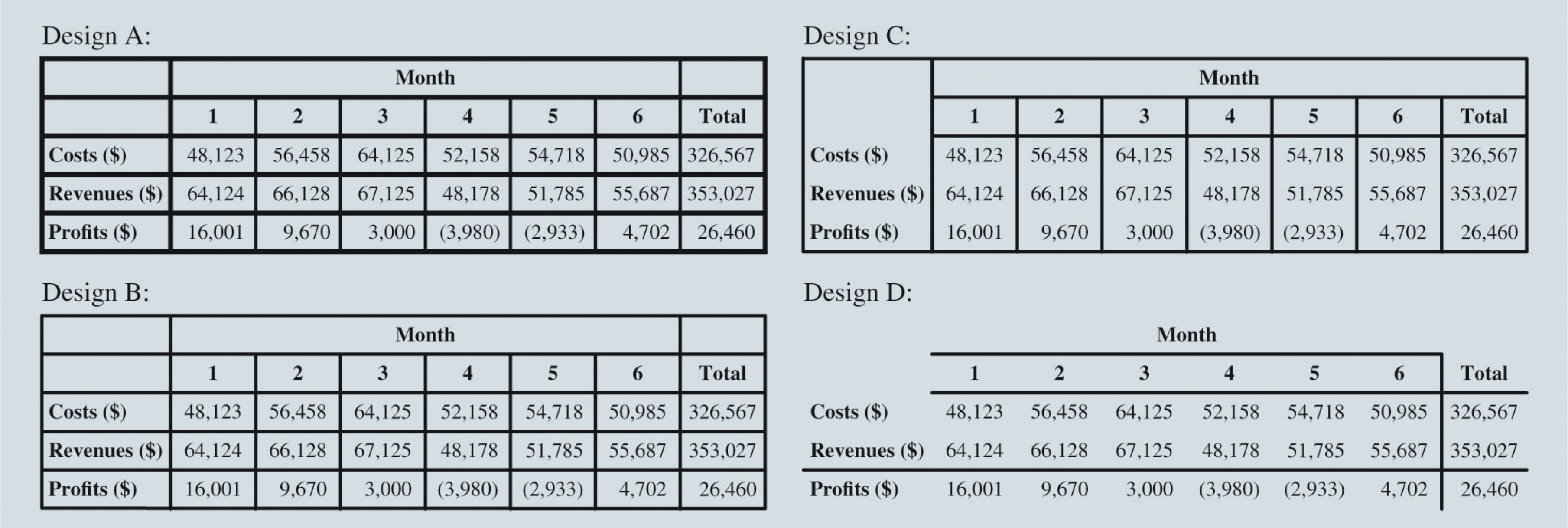
Tables should be used when:

1. The reader needs to refer to specific numerical values.
2. The reader needs to make precise comparisons between different values and not just relative comparisons.
3. The values being displayed have different units or very different magnitudes.

Table Design Principles:

- Avoid using vertical lines in a table unless they are necessary for clarity.
- Horizontal lines are generally necessary only for separating column titles from data values or when indicating that a calculation has taken place.

Comparing Different Table Designs



Design A:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design B:

	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design C:

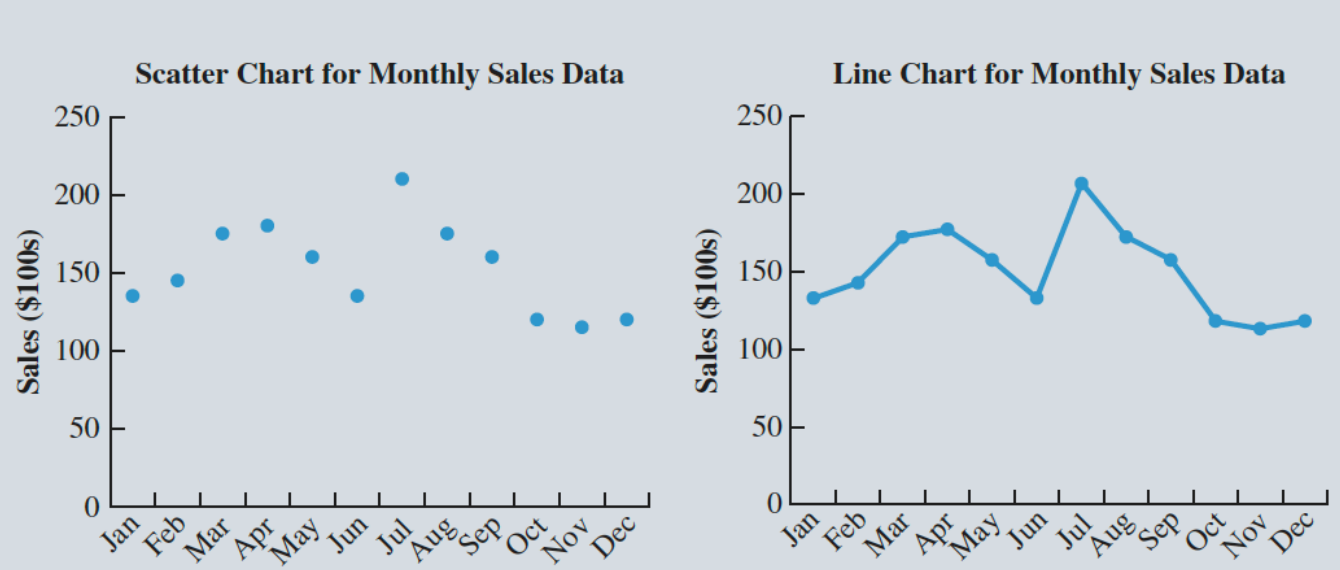
	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Design D:

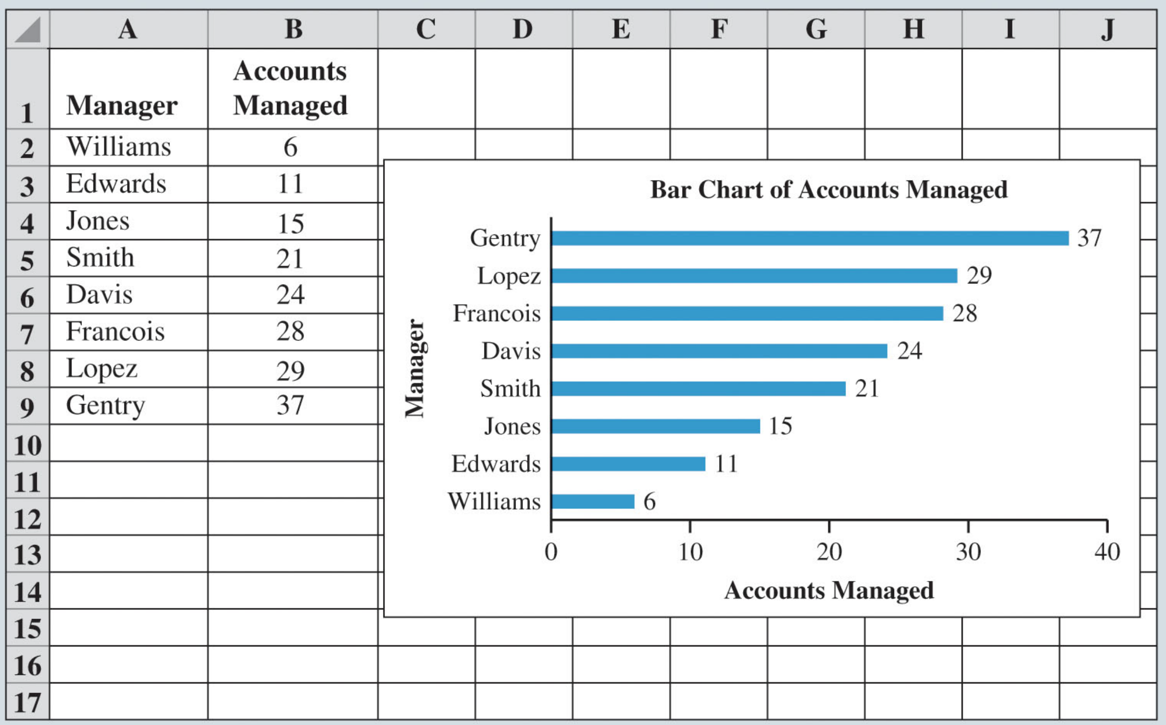
	Month						
	1	2	3	4	5	6	Total
Costs ($)	48,123	56,458	64,125	52,158	54,718	50,985	326,567
Revenues ($)	64,124	66,128	67,125	48,178	51,785	55,687	353,027
Profits ($)	16,001	9,670	3,000	(3,980)	(2,933)	4,702	26,460

Charts

- **Charts** (or graphs): Visual methods of displaying data.
- **Scatter chart**: Graphical presentation of the relationship between two quantitative variables.
- **Trendline**: A line that provides an approximation of the relationship between the variables.
- **Line chart**: A line connects the points in the chart.
  - Useful for time series data collected over a period of time (minutes, hours, days, years, etc.).

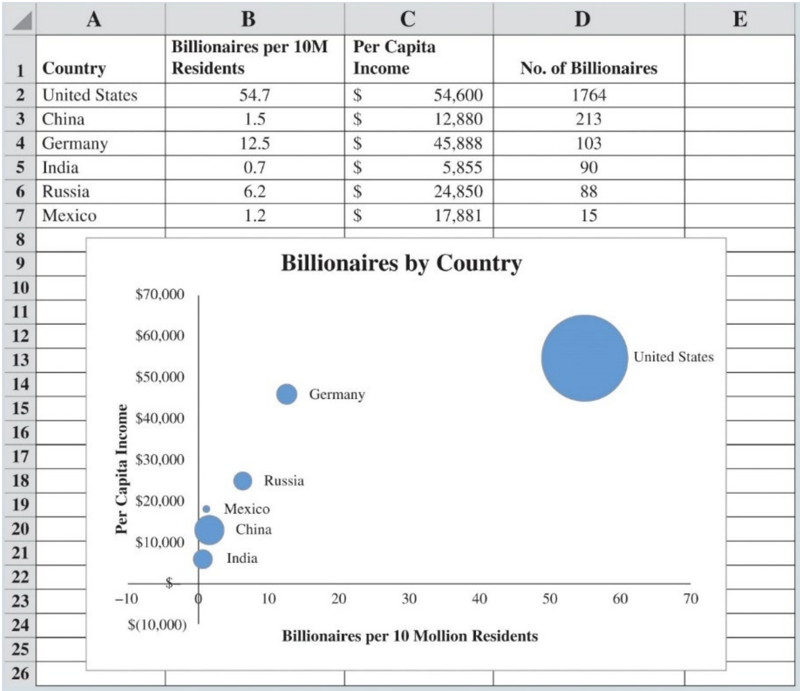


- **Bar Charts**: Use horizontal bars to display the magnitude of the quantitative variable.
- **Column Charts**: Use vertical bars to display the magnitude of the quantitative variable.
- Bar and column charts are very helpful in making comparisons between categorical variables.

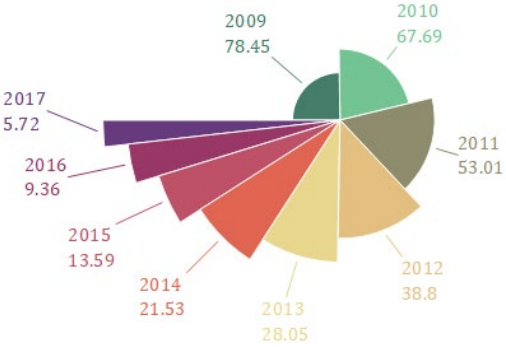


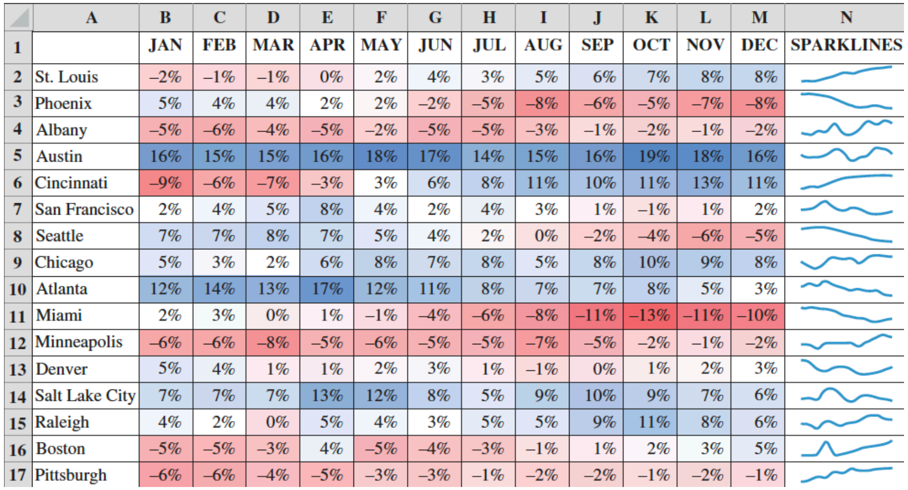
	A	B
1	**Manager**	**Accounts Managed**
2	Williams	6
3	Edwards	11
4	Jones	15
5	Smith	21
6	Davis	24
7	Francois	28
8	Lopez	29
9	Gentry	37

- **Pie chart**: Common form of chart used to compare categorical data.
- **Bubble chart**: Graphical means of visualizing three variables in a two-dimensional graph that sometimes is a preferred alternative to a 3-D graph.
- **Heat map**: A two-dimensional graphical representation of data that uses different shades of color to indicate magnitude.



	A	B	C	D	E
1	Country	Billionaires per 10M Residents	Per Capita Income	No. of Billionaires	
2	United States	54.7	$ 54,600	1764	
3	China	1.5	$ 12,880	213	
4	Germany	12.5	$ 45,888	103	
5	India	0.7	$ 5,855	90	
6	Russia	6.2	$ 24,850	88	
7	Mexico	1.2	$ 17,881	15	

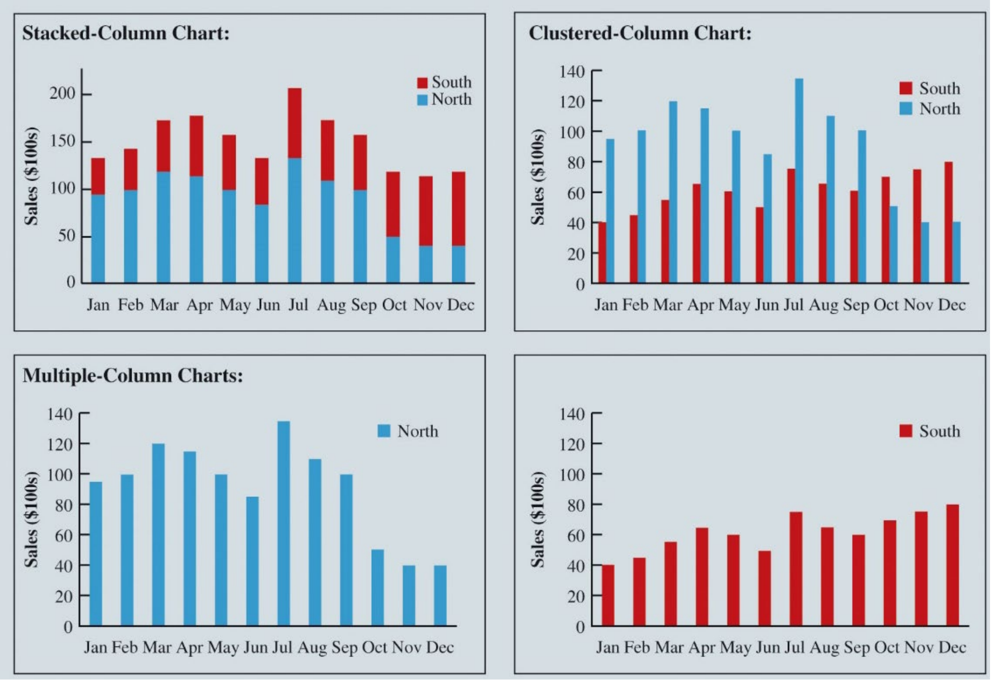




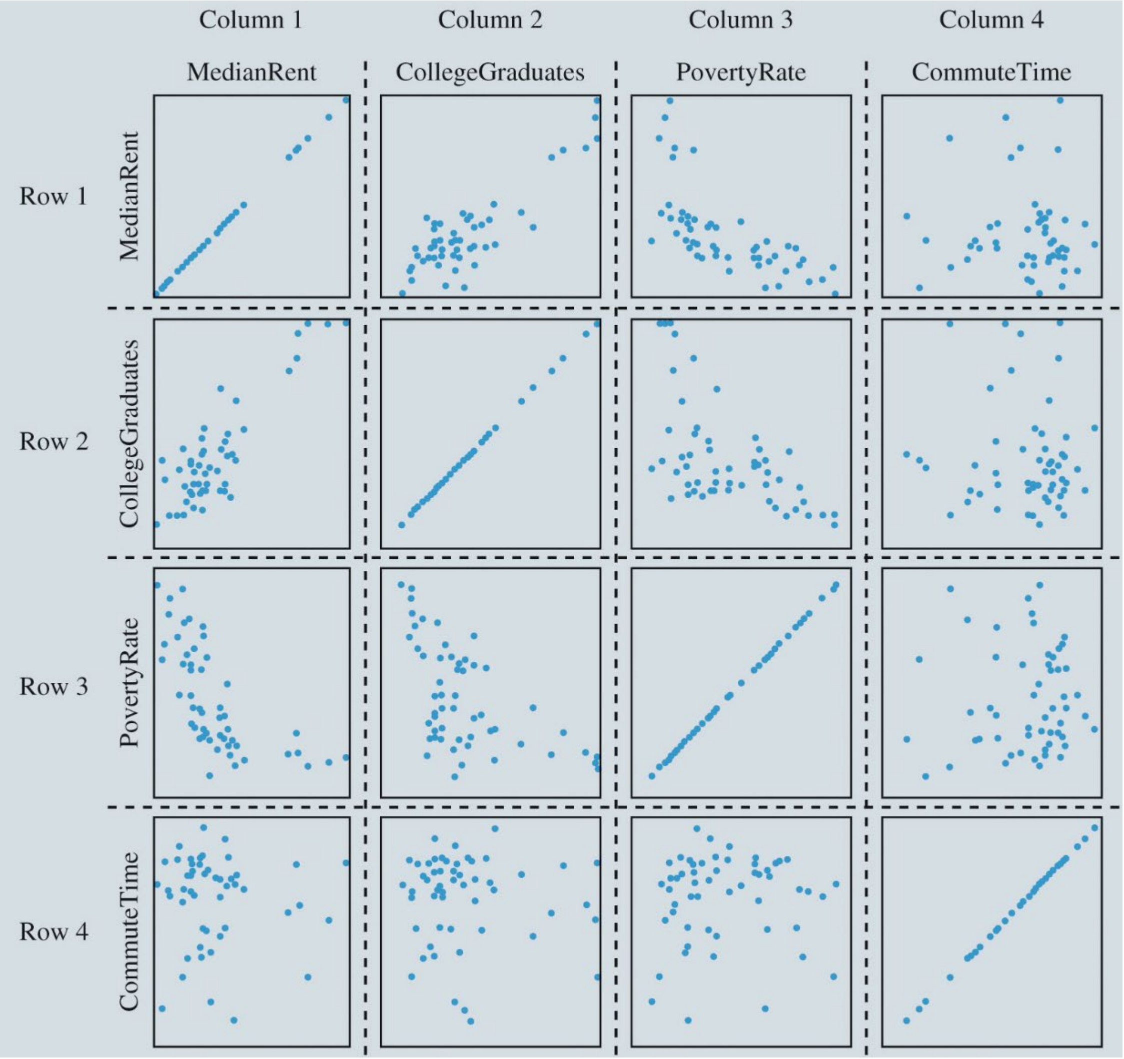
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	SPARKLINES
2	St. Louis	–2%	–1%	–1%	0%	2%	4%	3%	5%	6%	7%	8%	8%	
3	Phoenix	5%	4%	4%	2%	2%	–2%	–5%	–8%	–6%	–5%	–7%	–8%	
4	Albany	–5%	–6%	–4%	–5%	–2%	–5%	–5%	–3%	–1%	–2%	–1%	–2%	
5	Austin	16%	15%	15%	16%	18%	17%	14%	15%	16%	19%	18%	16%	
6	Cincinnati	–9%	–6%	–7%	–3%	3%	6%	8%	11%	10%	11%	13%	11%	
7	San Francisco	2%	4%	5%	8%	4%	2%	4%	3%	1%	–1%	1%	2%	
8	Seattle	7%	7%	8%	7%	5%	4%	2%	0%	–2%	–4%	–6%	–5%	
9	Chicago	5%	3%	2%	6%	8%	7%	8%	5%	8%	10%	9%	8%	
10	Atlanta	12%	14%	13%	17%	12%	11%	8%	7%	7%	8%	5%	3%	
11	Miami	2%	3%	0%	1%	–1%	–4%	–6%	–8%	–11%	–13%	–11%	–10%	
12	Minneapolis	–6%	–6%	–8%	–5%	–6%	–5%	–5%	–7%	–5%	–2%	–1%	–2%	
13	Denver	5%	4%	1%	1%	2%	3%	1%	–1%	0%	1%	2%	3%	
14	Salt Lake City	7%	7%	7%	13%	12%	8%	5%	9%	10%	9%	7%	6%	
15	Raleigh	4%	2%	0%	5%	4%	3%	5%	5%	9%	11%	8%	6%	
16	Boston	–5%	–5%	–3%	4%	–5%	–4%	–3%	–1%	1%	2%	3%	5%	
17	Pittsburgh	–6%	–6%	–4%	–5%	–3%	–3%	–1%	–2%	–2%	–1%	–2%	–1%	

Additional Charts for Multiple Variables:

- **Stacked-column chart**: Allows the reader to compare the relative values of quantitative variables for the same category in a bar chart.
- **Clustered-column (or bar) chart**: An alternative chart to stacked-column chart for comparing quantitative variables.
- **Scatter-chart matrix**: Useful chart for displaying multiple variables.



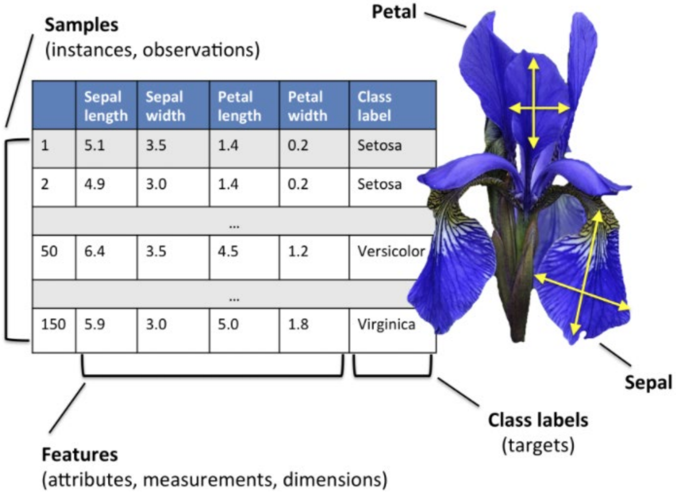
A scatter-chart matrix (scatter-plot matrix)



Advanced Charts

- **Parallel-coordinates plot**: Chart for examining data with more than two variables:
  - Includes a different vertical axis for each variable.
  - Each observation is represented by drawing a line on the parallel-coordinates plot connecting each vertical axis.
  - The height of the line on each vertical axis represents the value taken by that observation for the variable corresponding to the vertical axis.
- **Treemap**: Useful for visualizing hierarchical data along multiple dimensions.

A Parallel-Coordinates Plot for iris



Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

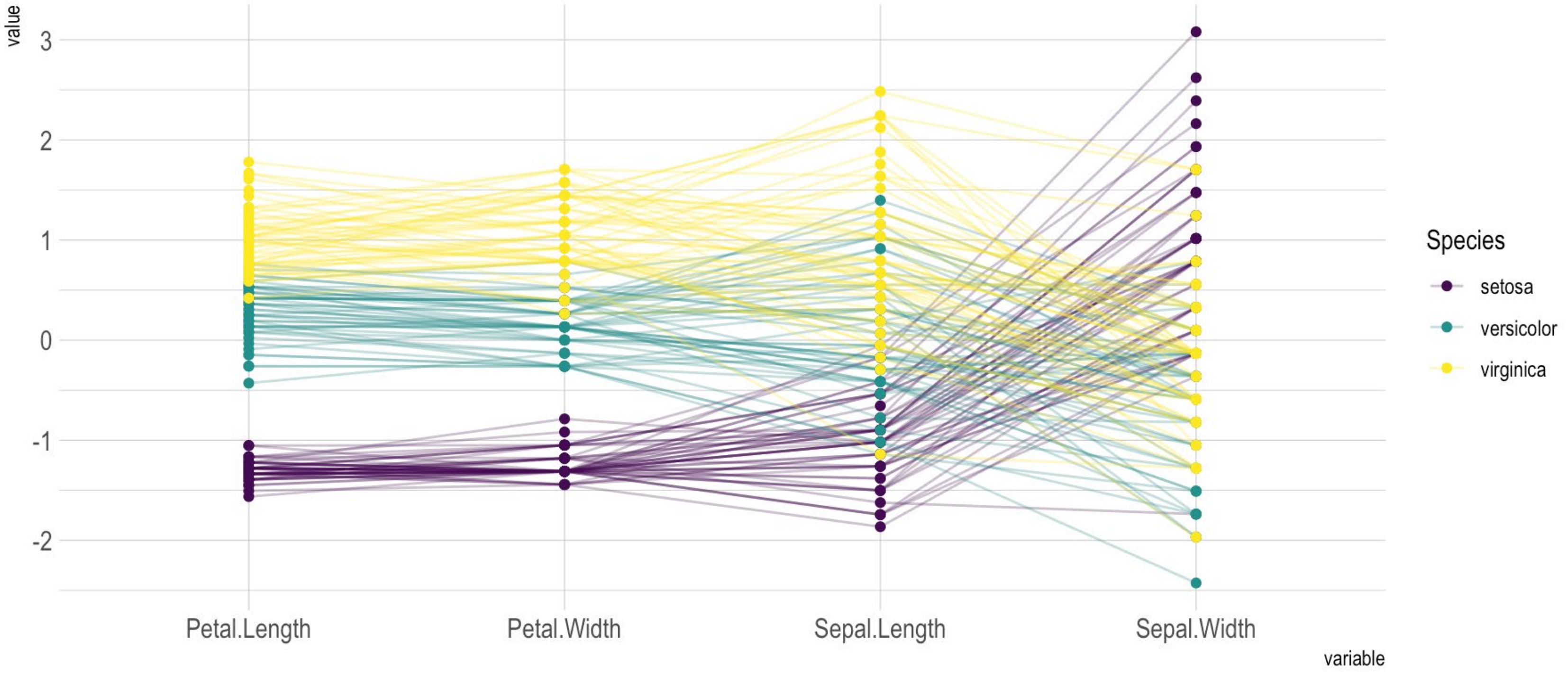
Petal

Sepal

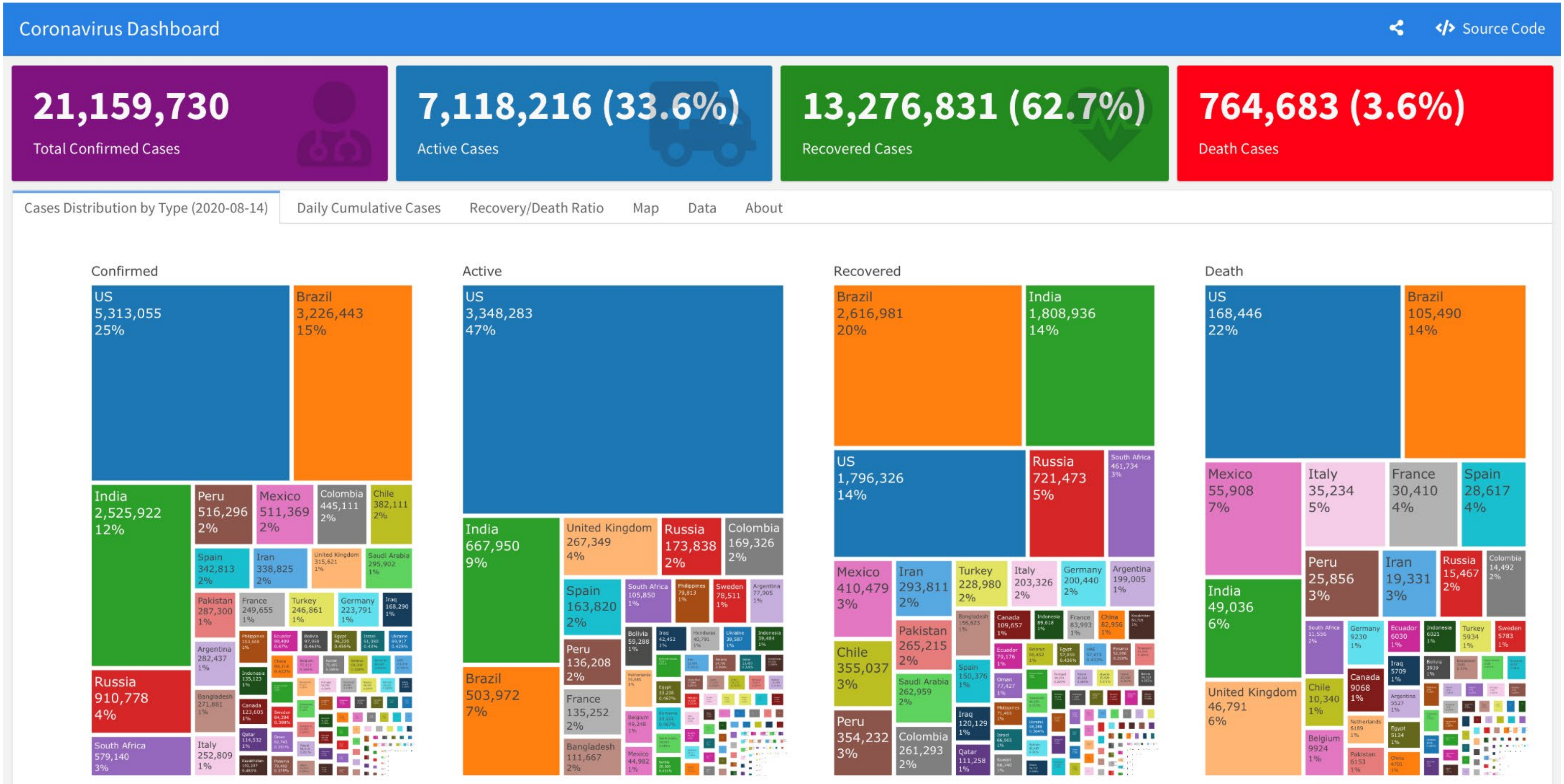
Class labels
(targets)

Features
(attributes, measurements, dimensions)

Parallel Coordinate Plot for the Iris Data



Treemap for COVID-19 cases



Coronavirus Dashboard

Source Code

21,159,730
Total Confirmed Cases

7,118,216 (33.6%)
Active Cases

13,276,831 (62.7%)
Recovered Cases

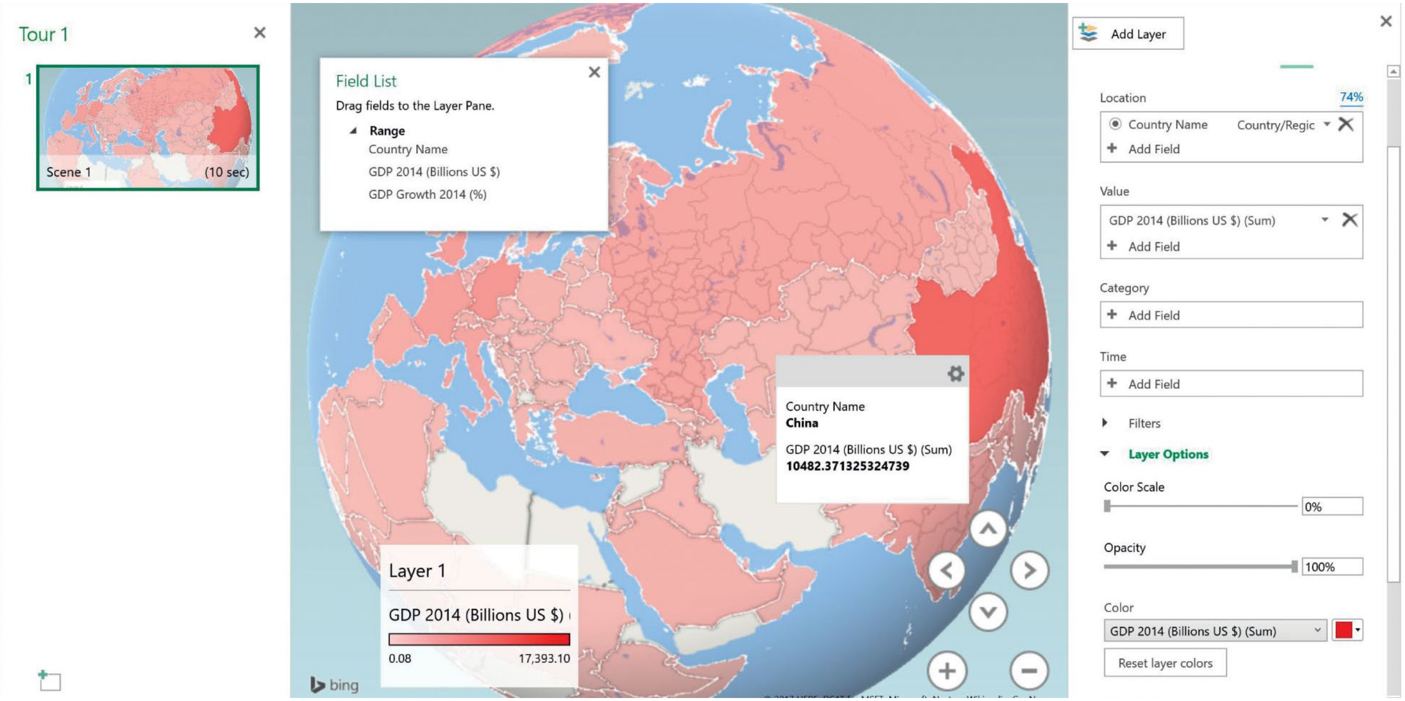
764,683 (3.6%)
Death Cases

Cases Distribution by Type (2020-08-14) | Daily Cumulative Cases | Recovery/Death Ratio | Map | Data | About

Geographic Information Systems Charts:

- **Geographic information system (GIS)**: A system that merges maps and statistics to present data collected over different geographic areas.
- Helps in interpreting data and observing patterns.

3D Map for World GDP Data



Field List

Drag fields to the Layer Pane.

Range
Country Name
GDP 2014 (Billions US $)
GDP Growth 2014 (%)

Country Name
China

GDP 2014 (Billions US $) (Sum)
10482.371325324739

Layer 1

GDP 2014 (Billions US $)

0.08 17,393.10

Location 74%
Country Name Country/Regic

Value
GDP 2014 (Billions US $) (Sum)

Category

Time

Filters

Layer Options

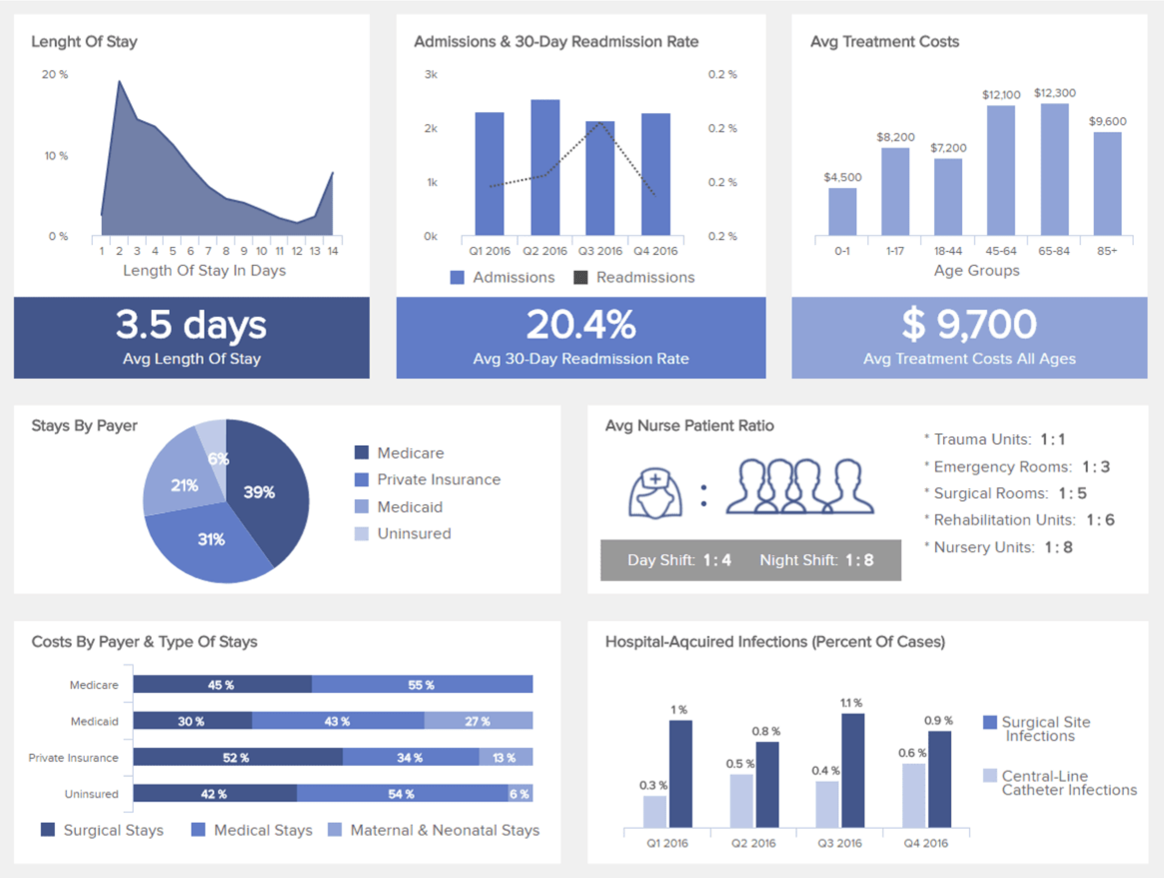
Color Scale 0%

Opacity 100%

Color
GDP 2014 (Billions US $) (Sum)

Data Dashboards

Data dashboard: Data-visualization tool that illustrates multiple metrics and automatically updates these metrics as new data become available.



Principles of Effective Data Dashboards

- **Key performance indicators (KPIs)** in dashboards:
  - Automobile dashboard: Current speed, Fuel level, and oil pressure.
  - Business dashboard: Financial position, inventory on hand, customer service metrics.
  - Should provide timely summary information on KPIs that are important to the user.
  - Should present all KPIs as a single screen that a user can quickly scan to understand the business's current state of operations.
  - The KPIs displayed in the data dashboard should convey meaning to its user and be related to the decisions the user makes.
  - A data dashboard should call attention to unusual measures that may require attention.
  - Colour should be used to call attention to specific values to differentiate categorical variables, but the use of color should be restrained.

Visualising Data with R

Consider the problem of allocating players into two ‘fair’ volleyball teams. You can find the dataset ”volley.txt” in Resources.

Student	Sprint 100m	Height (cm)	Serving	Endurance
Mizuho	15.78	148	94	17
Yukie	21.15	147	94	20
Megumi	14.30	134	91	17
Sakura	19.59	174	88	16
Izumi	10.96	145	93	16
Yukiko	19.17	158	83	12
Yumiko	18.35	157	99	20
Kayoko	14.09	177	82	23

Scatter Plots and Histograms,

We can try to solve the problem by visualising the data and plotting **histograms** and **s**catter plots.
To plot a histogram in R you need to upload your data first.
You can use the following command:

```
read.table(”volley.txt”)
```

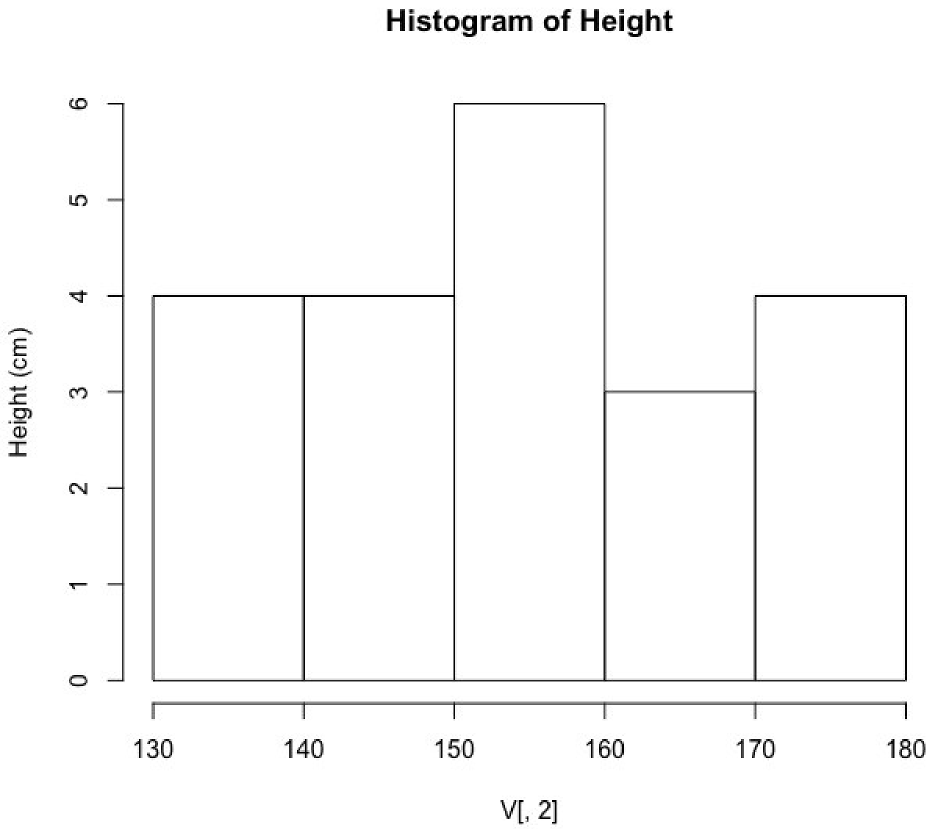
This will load a data set for the volleyball data, which we will use to illustrate various concepts.

Histograms

The plot a histogram in R use,

hist(V[,3])

will produce a histogram of the third column of V, V3, which is Height in cm.



Scatter Plots

To plot a scatter plot in R, plot(V[,2])
This will produce a scatter plot of the second column of the table assigned to V. It gives the results Sprint 100m.
The command in R,

```
plot(V[,2],V[,3], main=”Scatter plot of V3 vs V2,
xlab=”V2”,ylab=”V3”)
```

will produce a scatter plot of V2, Sprint 100m, and V3, height in cm.

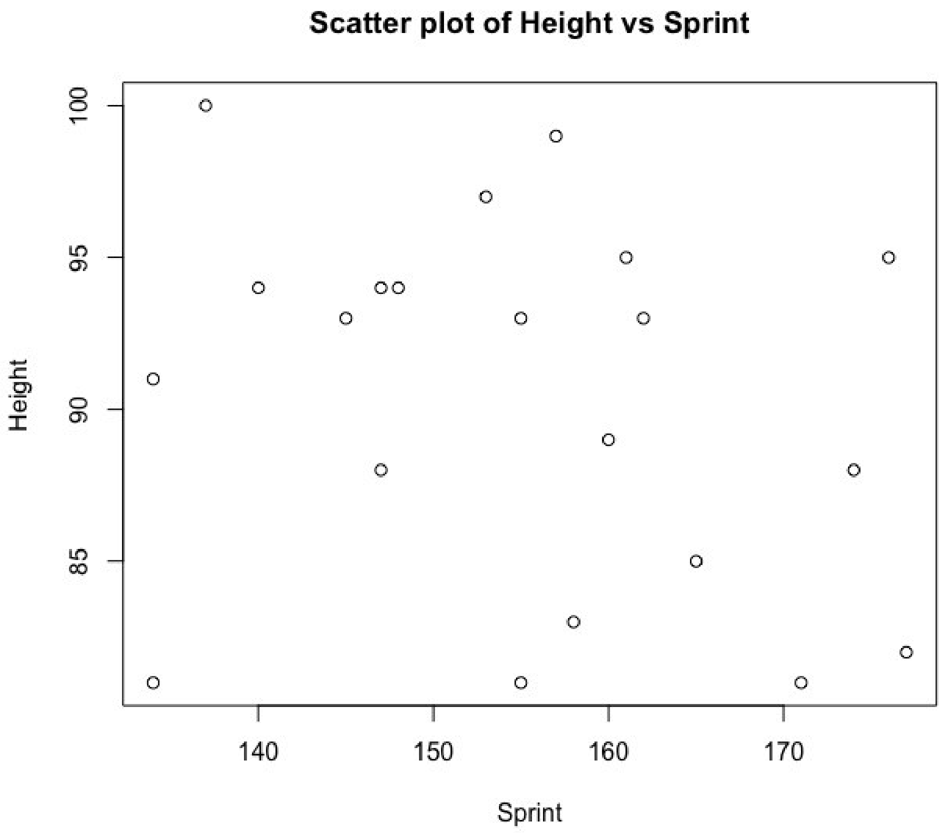
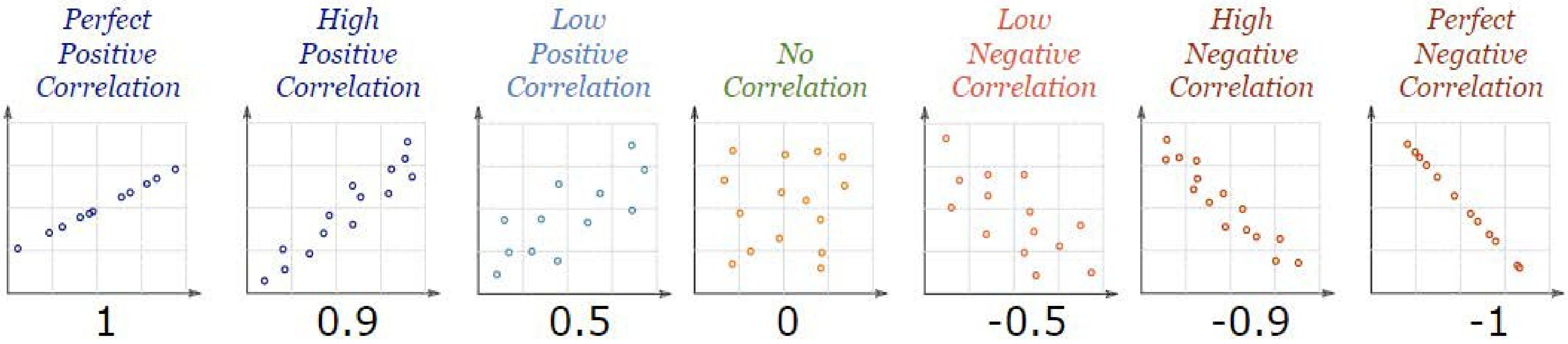


Figure: Scatter plot of V3, height, versus V2, Sprint 100m.

What information about the data can you find from the scatter plot?

Correlations

We can investigate correlations between variables using the scatter plot. We can also use the command in R.
The command corr() computes the correlation coefficient. The command cor.test() test for association/correlation between paired samples. It returns both the correlation coefficient and the significance level(or p-value) of the correlation.
The correlation can be computed using tests such as Pearson correlation coefficients, or Spearman correlation coefficient.



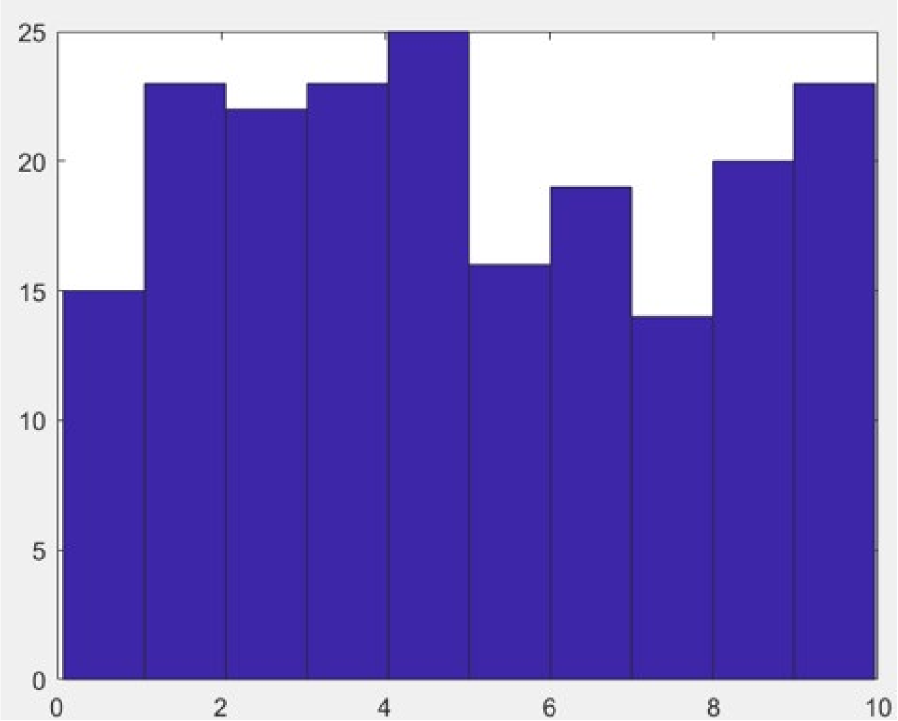
Tasks

Now, what can you say about the volleybal data. What are the distributions of the variables? Are the variables independent or correlated? What do you think are the main characteristics of the distributions?

Further Reading: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

Shape of Distribution

Does the histogram have a single, central hump, or several separated humps?
A histogram with one peak is called **u**nimodalhistogram,with two peaks are **b**imodal and with three or more peaks are called **m**ultimodal.



Frame Title

A histogram that does not appear to have any mode and in which all the bars are approximately the same height is called **u**niform.
Is the histogram symmetric?
The (usually) thinner ends of a distribution are called **t**ails.
Can you see any unusual features? Often such features can tell us something interesting about the data.
You should always mention any outliers that stand out from the body of the distribution.

The Median

A natural choice of typical value is the value in the middle, with half values below and half values above it.
The middle value that divides the histogram into two equal areas is called the median.

The Range

The **r**ange of the data is defined as the difference between the maximum and minimum values:

$$Range = max - min$$

QUARTILES

A better way to describe the spread of a variable could be to ignore the extremes and concentrate on the middle of the data.

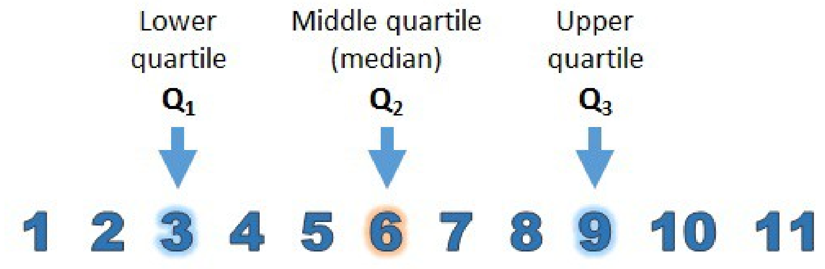
Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quartiles.

For any percentage of the data. there is a corresponding percentile - the value that leaves that percentage of the data below it.

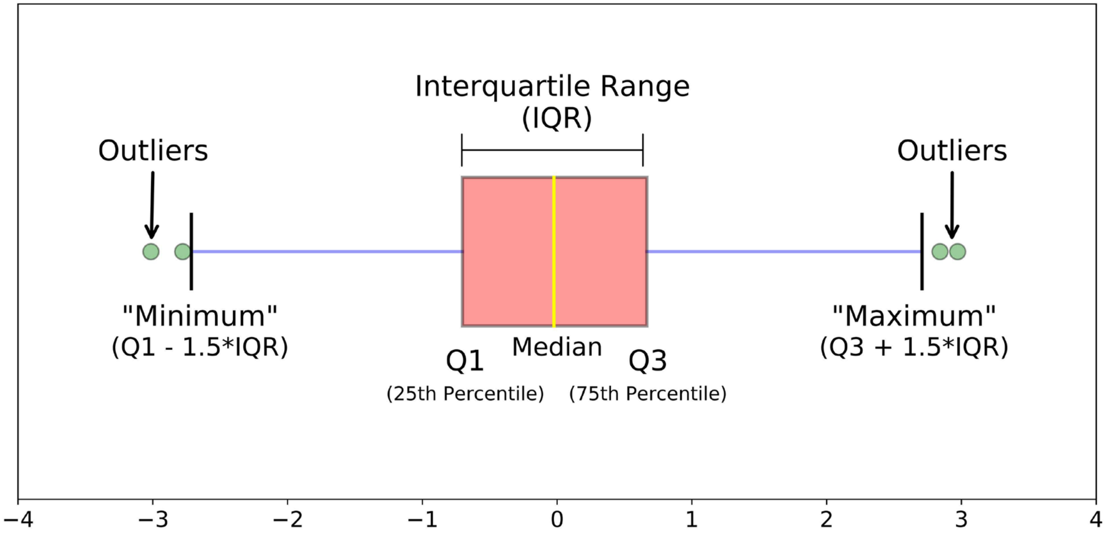
The lower and upper quartiles are also known as 25th and 75th percentiles of the data, respectively. The median is 50th percentile.

The difference between the quartiles indicates what the area the middle half of the data covers and is called interquartile range (IQR),

$$IQR = upper\ quartile - lower\ quartile.$$

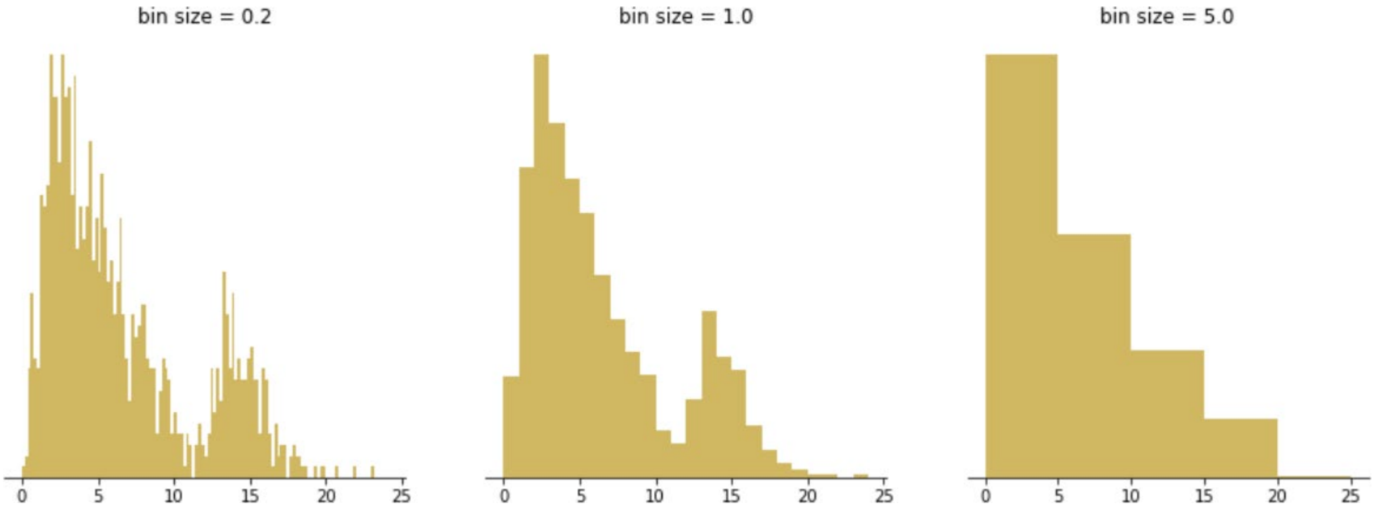


Boxplot



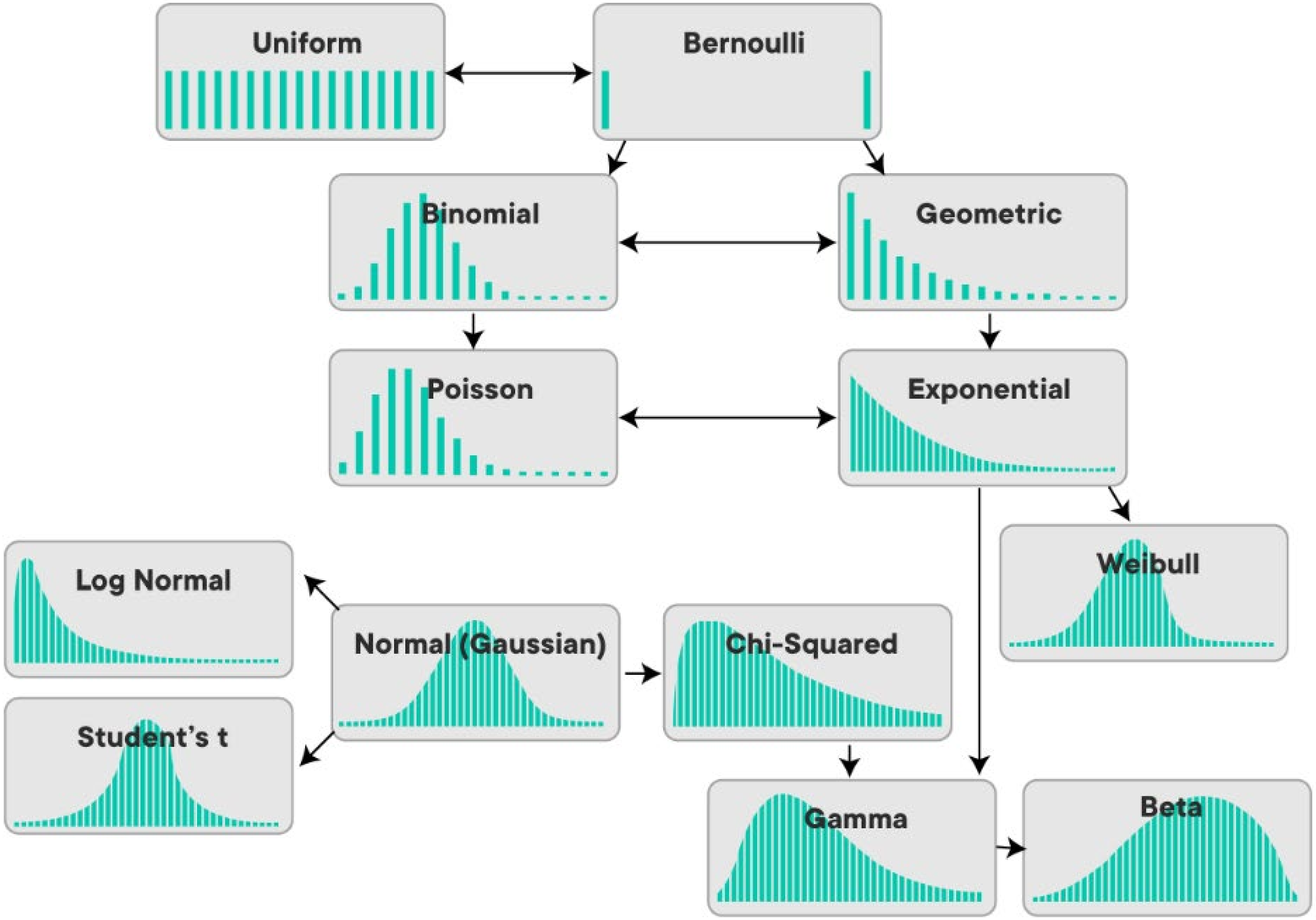
Shapes of Distributions

Different features of the distribution may appear more obvious at different bin width choices. When you use technology, it is easy to vary the bin width interactively, so you can make sure that a feature you think you see is not a consequence of a certain bin width choice.



https://chartio.com/learn/charts/histogram-complete-guide/

Common types of distributions

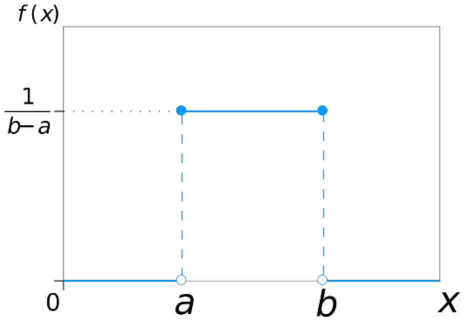


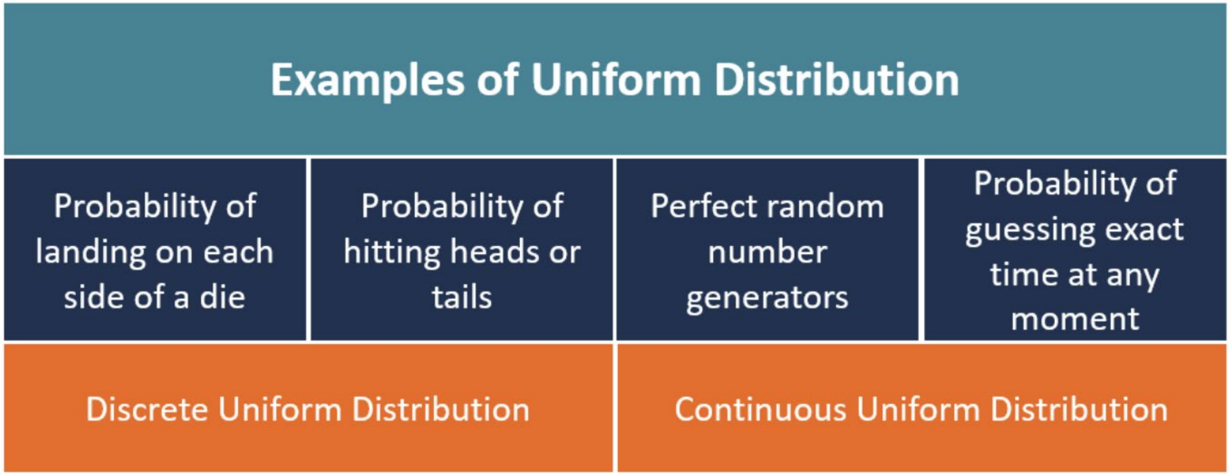
https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/

Uniform distributions

- In statistics, uniform distribution is a probability distribution where all outcomes are equally likely.
- Discrete uniform distributions have a finite number of outcomes. A continuous uniform distribution is a statistical distribution with an infinite number of equally likely measurable values.
- The concepts of discrete uniform distribution and continuous uniform distribution, as well as the random variables they describe, are the foundations of statistical analysis and probability theory.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$





Examples of Uniform Distribution

Probability of landing on each side of a die

Probability of hitting heads or tails

Perfect random number generators

Probability of guessing exact time at any moment

Discrete Uniform Distribution

Continuous Uniform Distribution

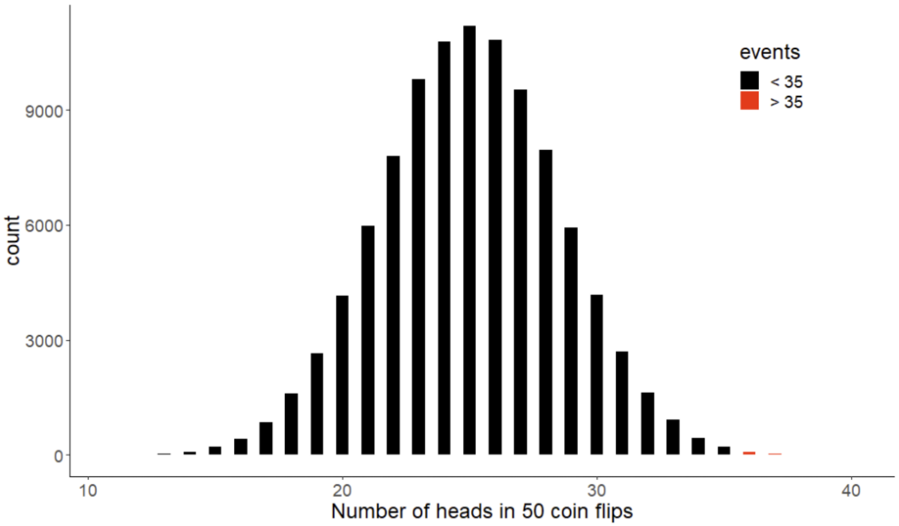
Binomial distributions

A **binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The **binomial** is a type of **distribution** that has two possible outcomes (the prefix “bi” means two, or twice).

If there are *n* Bernoulli trials, and each trial has a probability *p* of success, then the probability of exactly *k* successes is *Pr*(*X*=*k*):

$$Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}.$$

A repeating set of 50-times coin flipping 100000 times and record the number of successes in each repetition.



Poisson distributions

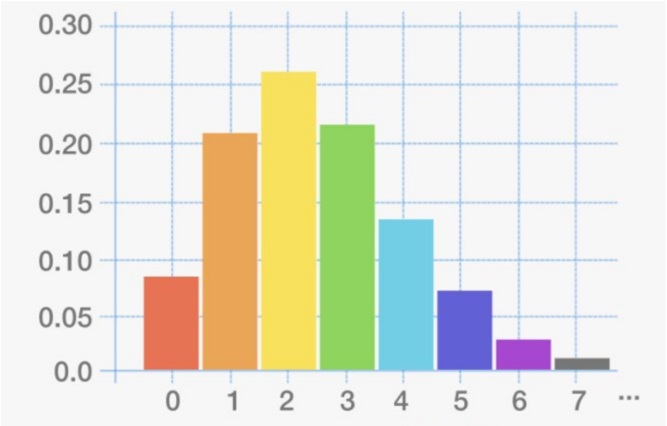
Many experimental situation occur in which we observe the counts of events within a set unit of time, area, volume, length etc. The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

λ be the expected value (average) of X
e is Euler's number

Applications

- the number of deaths by horse kicking in the Prussian army (first application)
- car accidents
- traffic flow and ideal gap distance
- number of typing errors on a page
- hairs found in McDonald's hamburgers
- spread of an endangered animal in Africa
- failure of a machine in one month



The Poisson distribution with $\lambda = 2.5$

Tasks

Import dataset volley.txt into your R directory. Produce scatter plots and histogram of the first and second column of the table. What can you say for the shapes of the distributions?

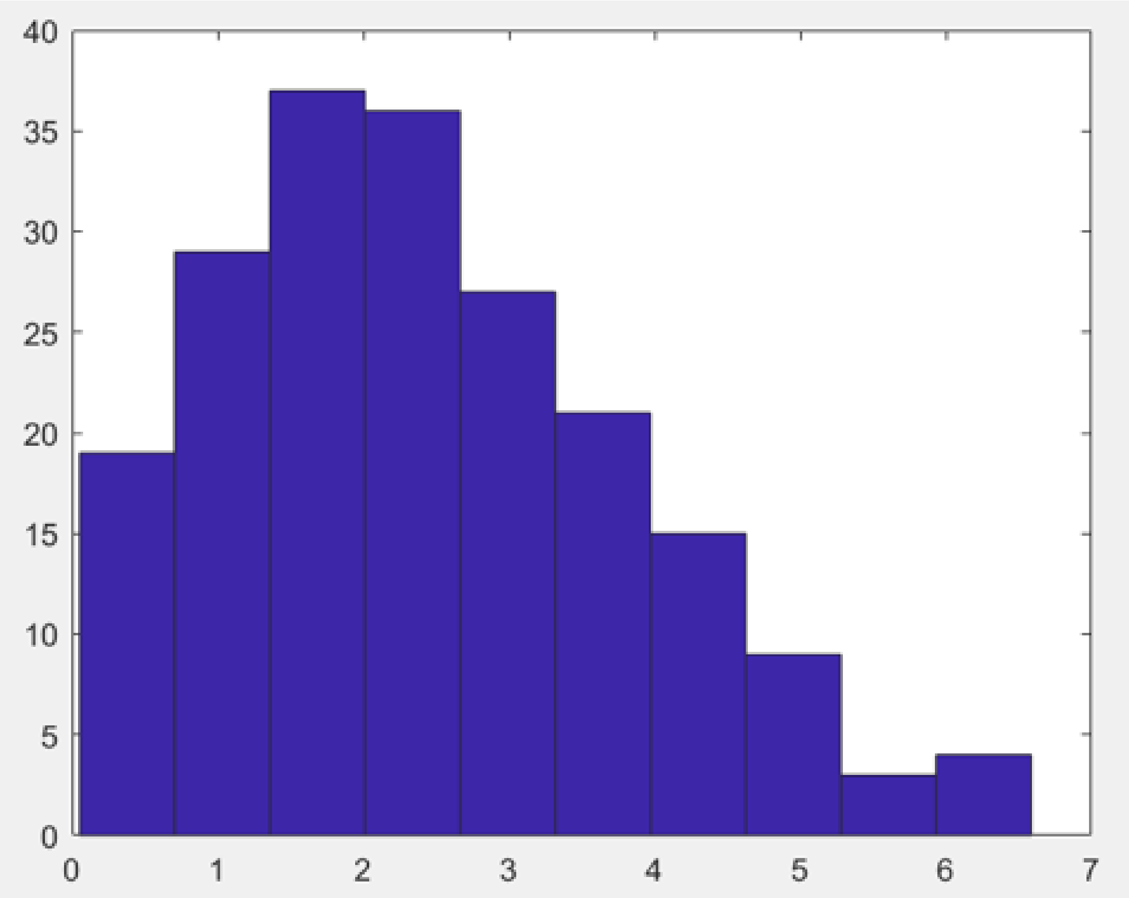


Figure: Example of data which are normally distributed.

This histogram has a bell-shaped curve.

Normal Distribution

One of our tasks in this course will be to understand and compare distributions of data.

You may have heard of ”bell-shaped curves”. Statisticians call them **normal models**. Normal models are appropriate for distributions whose shapes are unimodal and roughly symmetric. There is a normal model for every possible combination of mean and standard deviation.

Normal Distribution continues...

A Normal model is represented by $N(\mu, \sigma)$. In this notation μ is the mean (this also gives you the trend) and σ is the standard deviation (which represents the variance of the data). μ and σ are parameters of the model.
If we model the data with a Normal model and standardise them using the corresponding μ and σ, we call the standardised value a z-score.

$$z = \frac{y - \mu}{\sigma},$$

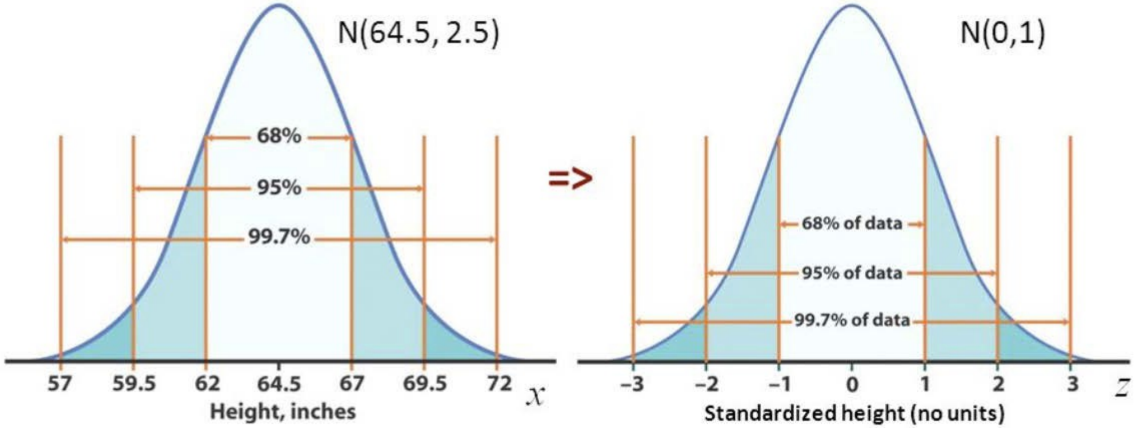
It is easier to standardise the data first, using the z-score. Then we need only the model $N(0, 1)$. The Normal model with mean 0 and standard deviation 1 is called **standard Normal model** or **standard Normal distribution**.

Standardising with z-Scores

Expressing a distance from the mean in standard deviations **standardises** the performance. To **standardise** a value, we subtract the mean μ and then divide the difference by the standard deviation σ:

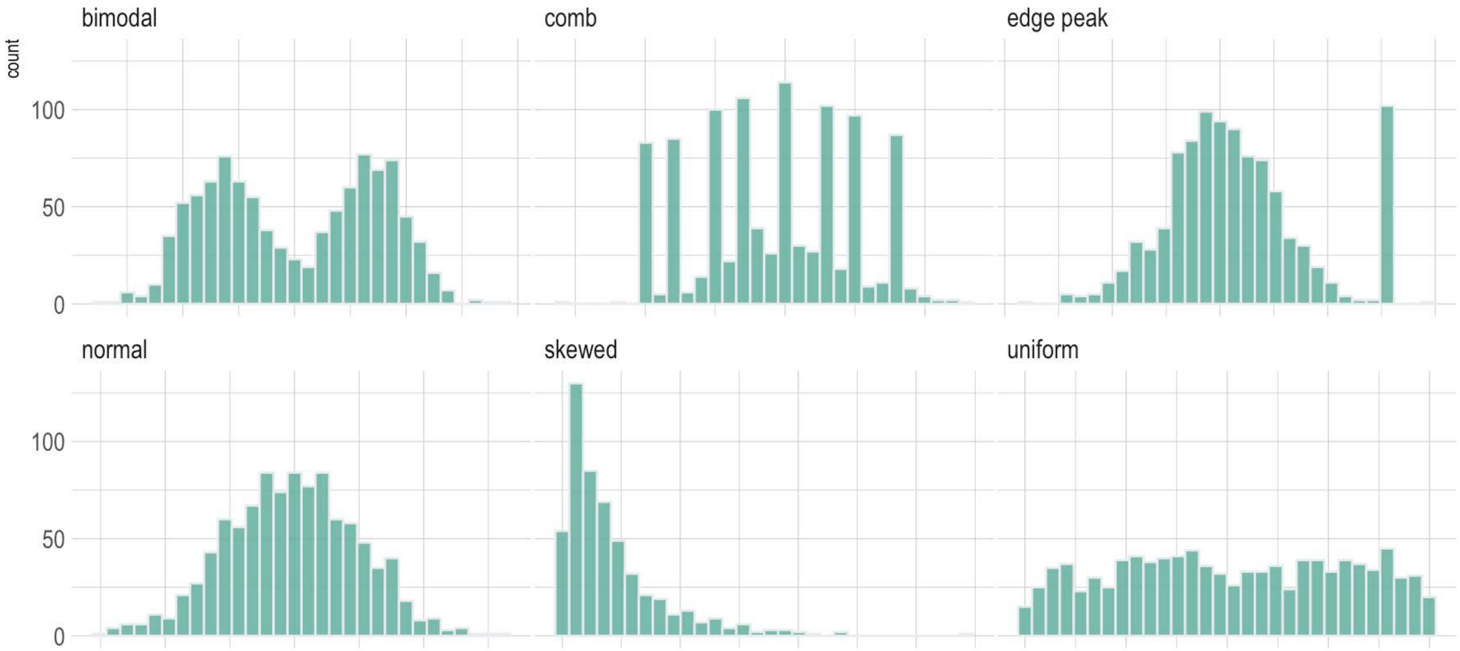
$$z = \frac{y - \mu}{\sigma},$$

where y is the original value. The values are known as **standardized values** and denoted by z. We can also call them z-**scores**.



For each x we calculate a new value, z (called a **z-score**).

However, you should not use a Normal model for any data set. Remember that standardising will not change the shape of the distribution.



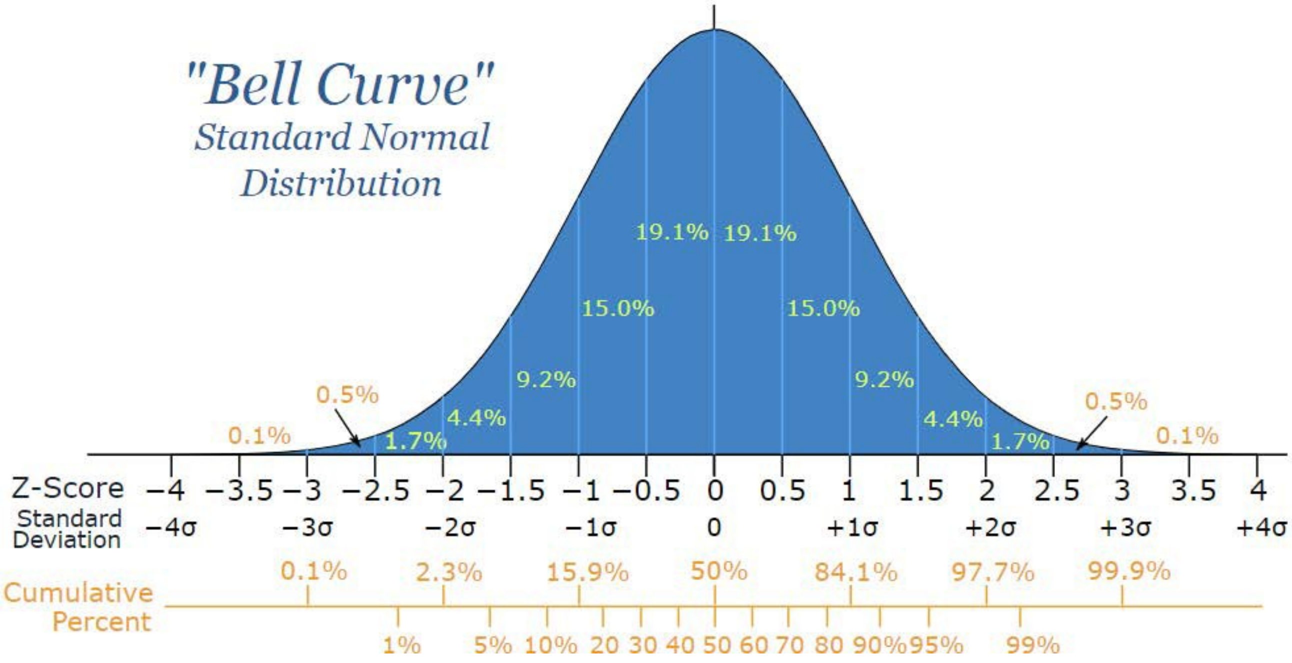
Tests for Normality

The sketch of the Normal distribution has the following characteristics:

- The Normal curve is bell-shaped and symmetric around its mean.
- The Normal model extends forever on each side, however you need to draw it only for 3 standard deviations - 3σ on each side.
- The place where the bell shape changes from curving downward to curving back up is exactly one standard deviation, one σ, away from the mean.

Tests for Normality continues...

To test a distribution for normality, you can use the shape and check if the characteristics are as described above, or you can use standard tests. One such test is Kolmogorov-Smirnov test.



TESTS FOR NORMALITY CONTINUES...

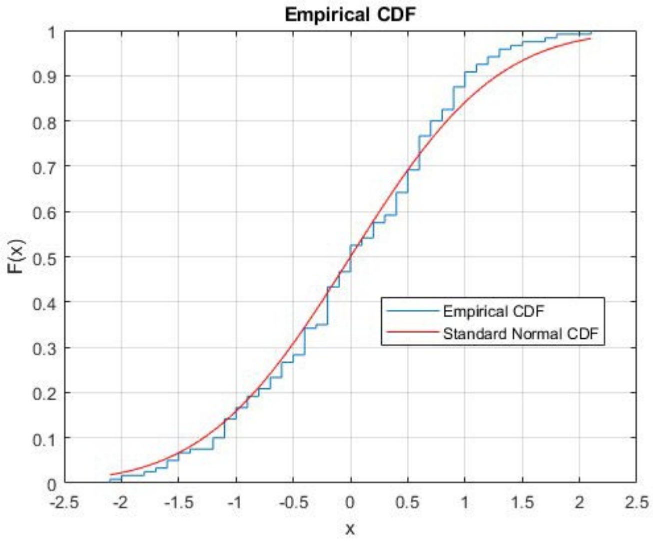
A normality test is used to determine whether sample data has been drawn from a normally distributed population (within some tolerance). A number of statistical tests, such as the Student's t-test and the one-way and two-way ANOVA require a normally distributed sample population. If the assumption of normality is not valid, the results of the tests will be unreliable.

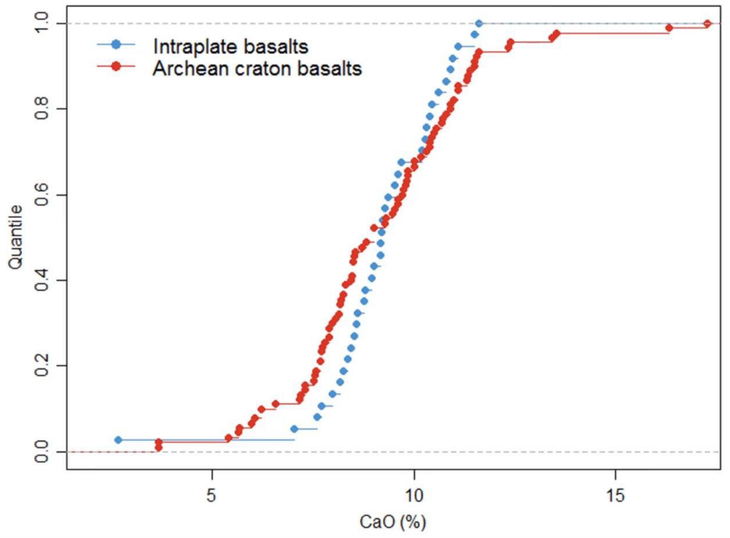
A graphical informal approach to test normality is to compare a histogram of the sample data to a normal probability curve. The empirical distribution of the data (the histogram) should be bell-shaped and resemble the normal distribution. This might be difficult to see if the sample is small.

Kolmogorox-Smirnov Test

The Kolmogorov-Smirnov test can be used to compare a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test). It quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.

Reference:
https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test





TEST FOR NORMALITY IN R

K-S test using R:
https://www.rdocumentation.org/packages/dgof/versions/1.2/topics/ks.test

```
ks.test(x, y, ..., alternative = c(”two.sided”, ”less”, ”greater”),
exact = NULL, tol=1e-8, simulate.p.value=FALSE, B=2000)
```

There are other test methods such as Shapiro–Wilk test and Jarque-Bera test. However, all tests have their own disadvantages.

Further reading: https://datasharkie.com/how-to-test-for-normality-in-r/

Skewed Distributions

Consider the examples of the figures below:

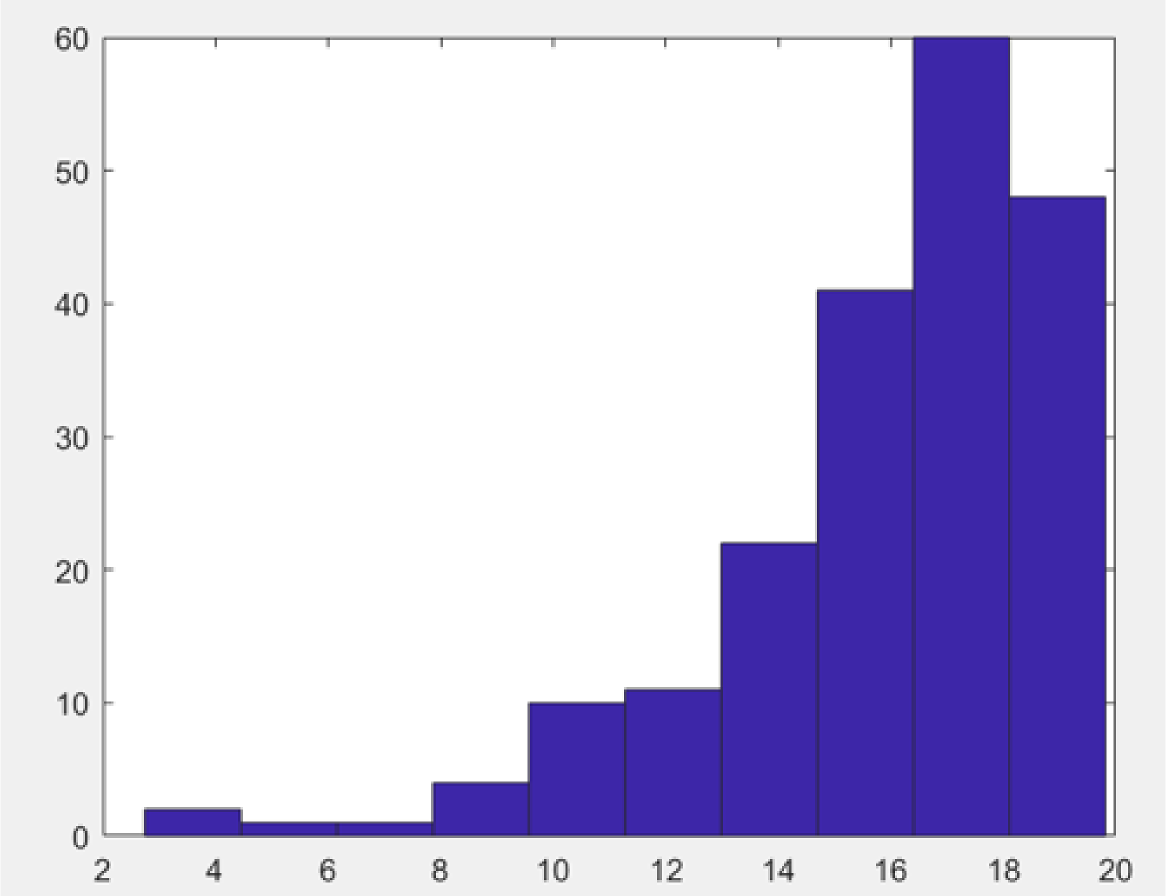
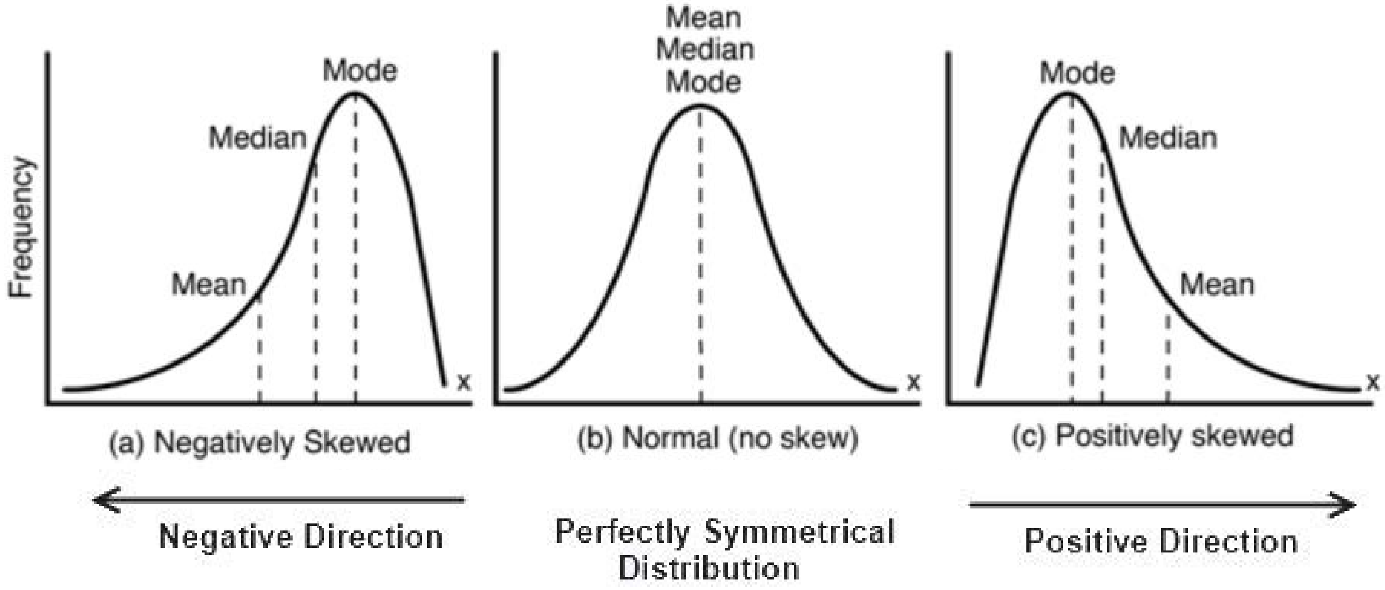


Figure: Skewed distribution from simulated data

Skewed Distributions continues...



Why do we need to transform data?

If we can get an ‘overall rating’ of each player. One reasonably simple method would be to aggregate their scores in each of the separate columns.

But we have several problems if we do this....

For example, a feature that ranges between 0 and 1000 will outweigh a feature that ranges between 0 and 1. Using these variables without standardisation will give the feature with the larger range weight of 1000 in the analysis. Transforming the data to comparable scales can prevent this problem. Typical data standardisation procedures equalise the range and/or data variability.

Take Home Message

- Visualising data with scatter plots and histogram is a good way to start the data exploration.
- Find out if data were distributed normally by examining the histogram and or performing some standard tests for normality.
- If the distribution is normal, one can use the main features of the normal distribution to characterise the data.
- If the distribution is not normal, as in many cases of real data, we will learn what to do next week.