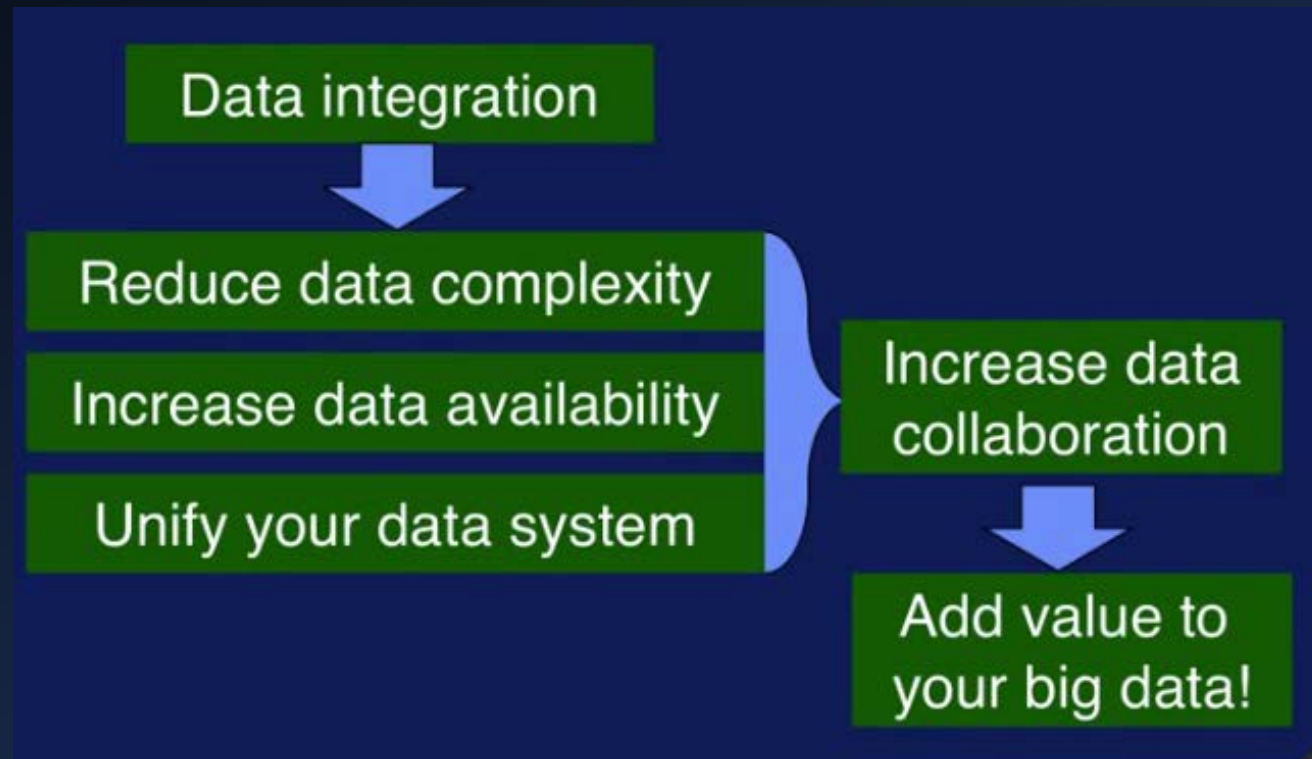


# Big Data (MHI222956/MHI225101)

## 6.1 Data Exploring

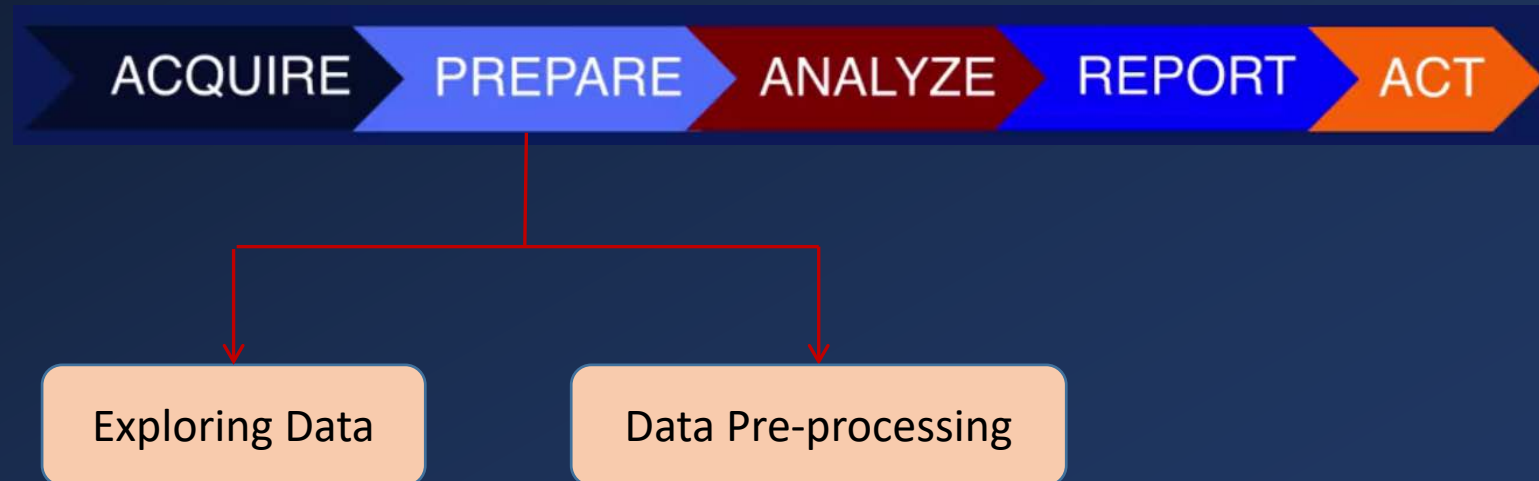




Big data need to be analysed to get the **value** – the real reason we are interested in big data.

# Steps in Data Science Process

- A simple linear form of data science process



# Steps in Data Science Process

- A simple linear form of data science process

**Purpose**

Iterative  
process

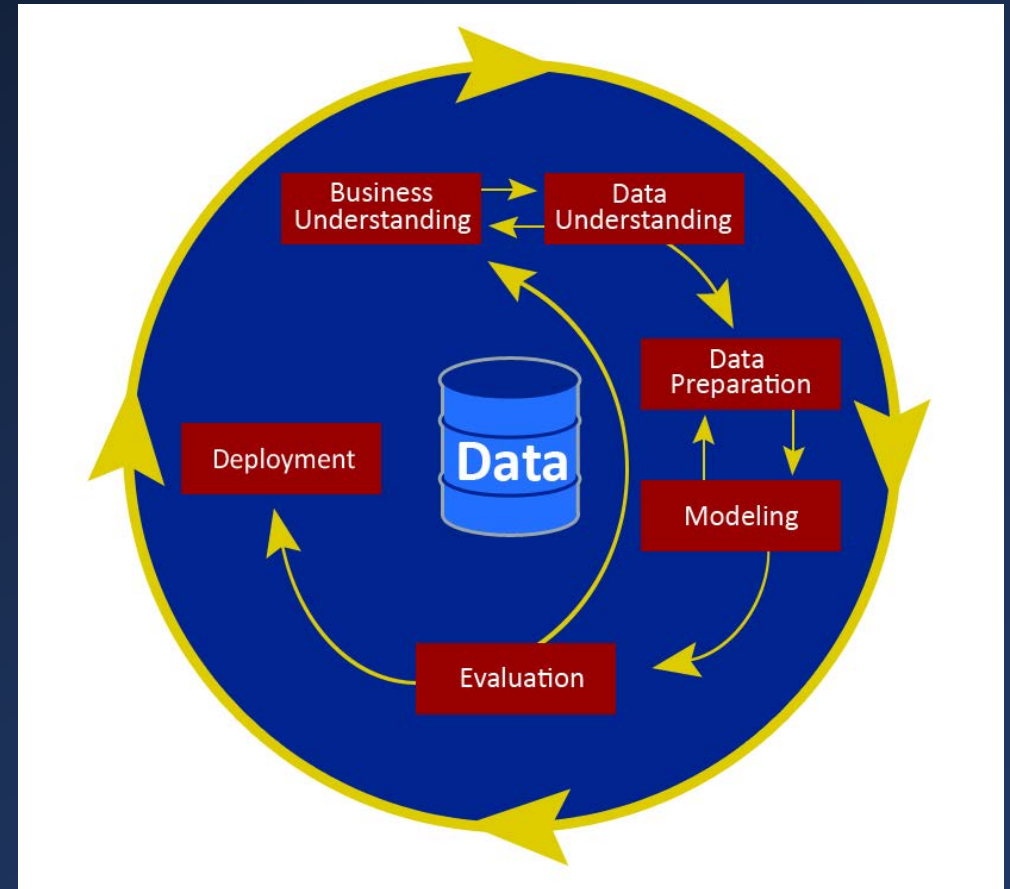


Exploring Data

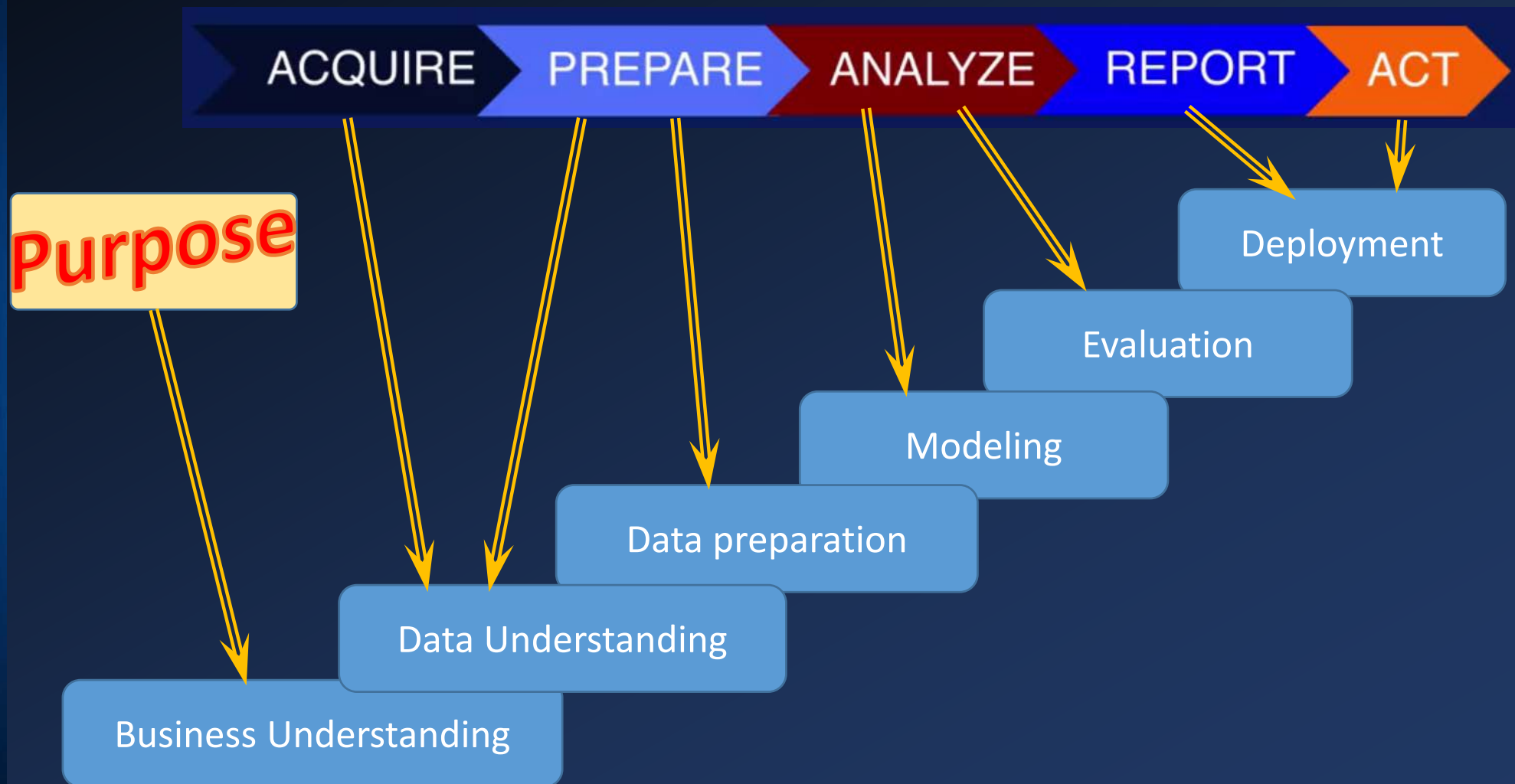
Data Pre-processing

# CRISP-DM

- CRISP-DM: Cross Industry Standard Process for Data Mining
  - A well adopted methodology for data mining
- Six Phases
  - Business understanding
  - Data understanding
  - Data preparation
  - Modeling
  - Evaluation
  - Deployment







# Data terminology

Variable

- Feature
- Field
- Column
- ...

Samples

- Instance
- Record
- Row
- Observation
- ...

	employeeNumb	lastName	firstName	extension	email	officeCode	reportsTo	jobTitle
▶	1002	Murphy	Diane	x5800	dmurphy@classicmodelcars.com	1	NULL	President
	1056	Patterson	Mary	x4611	mpatterso@classicmodelcars.com	1	1002	VP Sales
	1076	Firelli	Jeff	x9273	jfirelli@classicmodelcars.com	1	1002	VP Marketing
	1088	Patterson	William	x4871	wpatterson@classicmodelcars.com	6	1056	Sales Manager (APAC)
	1102	Bondur	Gerard	x5408	gbondur@classicmodelcars.com	4	1056	Sale Manager (EMEA)
	1143	Bow	Anthony	x5428	abow@classicmodelcars.com	1	1056	Sales Manager (NA)
	1165	Jennings	Leslie	x3291	ljennings@classicmodelcars.com	1	1143	Sales Rep
	1166	Thompson	Leslie	x4065	lthompson@classicmodelcars.com	1	1143	Sales Rep
	1188	Firelli	Julie	x2173	jfirelli@classicmodelcars.com	2	1143	Sales Rep
	1216	Patterson	Steve	x4334	spatterson@classicmodelcars.com	2	1143	Sales Rep
	1286	Tseng	Foon Yue	x2248	ftseng@classicmodelcars.com	3	1143	Sales Rep
	1323	Vanauf	George	x4102	gvanauf@classicmodelcars.com	3	1143	Sales Rep
	1337	Bondur	Loui	x6493	lbondur@classicmodelcars.com	4	1102	Sales Rep
	1370	Hernandez	Gerard	x2028	ghemande@classicmodelcars.com	4	1102	Sales Rep
	1401	Castillo	Pamela	x2759	pcastillo@classicmodelcars.com	4	1102	Sales Rep
	1501	Rott	Larry	x2311	lrott@classicmodelcars.com	7	1102	Sales Rep

# Data type

- Each variable has a data type associated with it
- Most common
  - **Numerical features:** Features with values that are continuous on a scale, statistical, or integer-related.
  - **Categorical features:** Features whose explanations or values are taken from a defined set of possible explanations or values.
- Others
  - String
  - Date
  - ...



# Data exploring

- Also called **exploratory data analysis (EDA)**
- Aims to gain a better understanding of the data that you have to work with
- A critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

# Data exploring

- Main categories of techniques to explore data:
  - Summary statistics
  - Visualization
- Something to look for:
  - Distributions
  - Correlations
  - General trends
  - Outliers
  - ...

# Summary statistics

- Mean

- The average of all numbers and is sometimes called the arithmetic mean

$$\bar{x} = \frac{1}{n} \left( \sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Median

- The statistical median is the middle number in a sequence of numbers.
- To find the median, organize each number in order by size; the number in the middle is the median.





# Summary statistics

- Mode

- the number that occurs most often within a set of numbers
- Mode helps identify the most common or frequent occurrence of a characteristic. It is possible to have two modes (bimodal), three modes (trimodal) or more modes within larger sets of numbers.

- Range

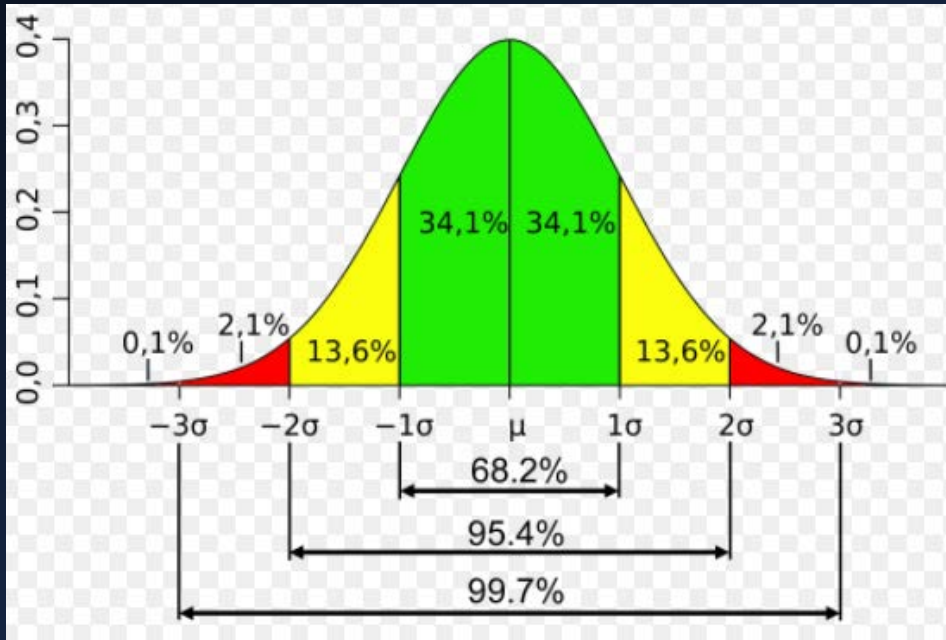
- The difference between the highest and lowest values within a set of numbers.
- To calculate range, subtract the smallest number from the largest number in the set.



# Summary statistics

- Standard derivation  $\sigma$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$



- A measure that is used to quantify the amount of variation or dispersion of a set of data values:
  - A low standard deviation indicates that the data points tend to be close to the **mean** (also called the expected value) of the set,
  - a high standard deviation indicates that the data points are spread out over a wider range of values.

# Visualization

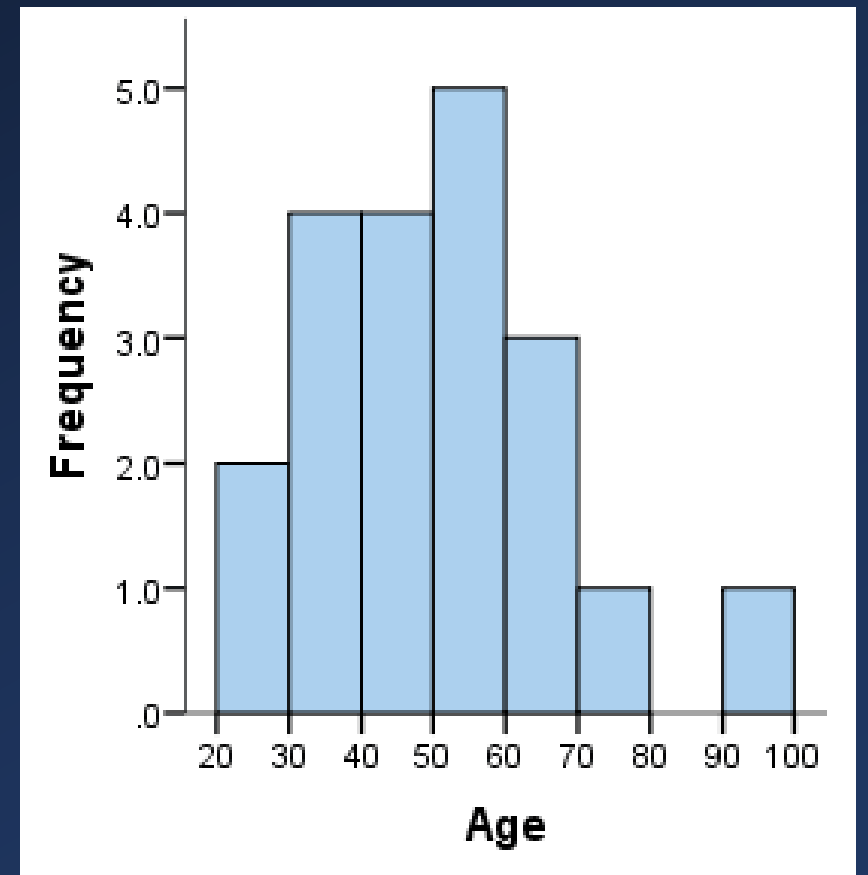
- Histogram

- A graphical display of data using bars of different heights
- It is similar to a Bar Chart, but a histogram groups numbers into ranges (bins), then count how many values fall into each interval.
- The bins must be adjacent, and are often (but are not required to be) of equal size.
- Histograms are good for showing general distributional features of dataset variables.
- You can see roughly where the peaks of the distribution are, whether the distribution is skewed or symmetric, and if there are any outliers.

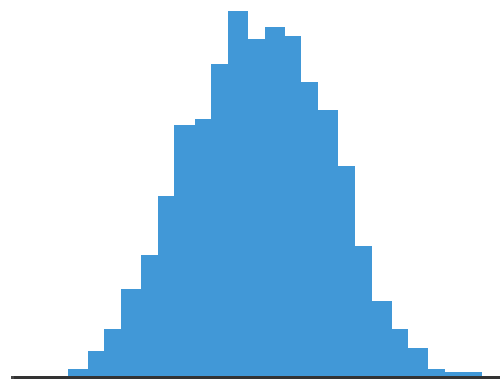


36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55

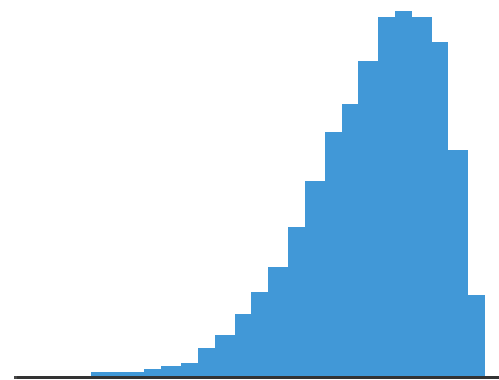
Bin	Frequency	Scores Included in Bin
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91



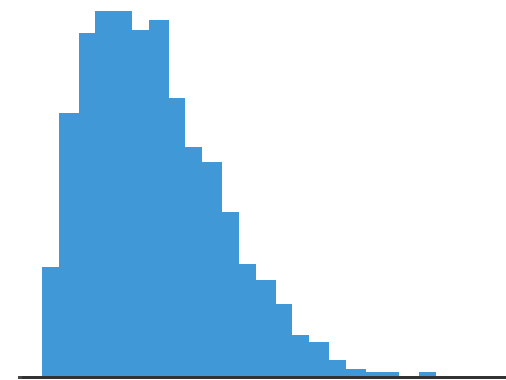
- Patterns in a histogram



symmetric, unimodal



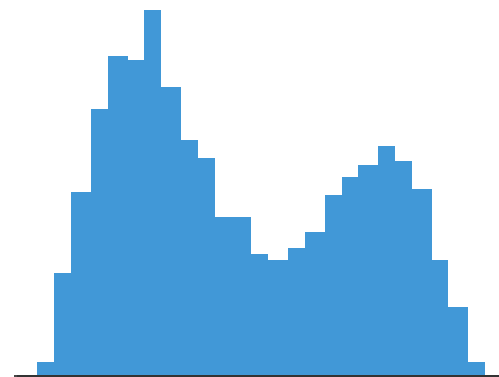
skew left



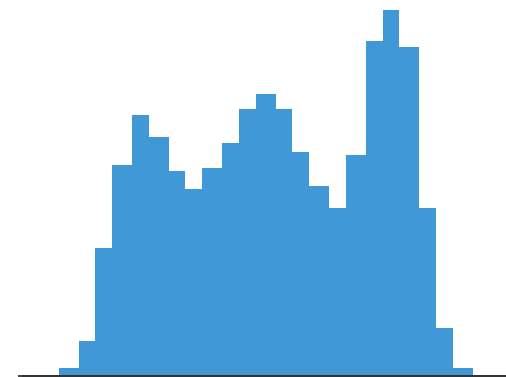
skew right



uniform

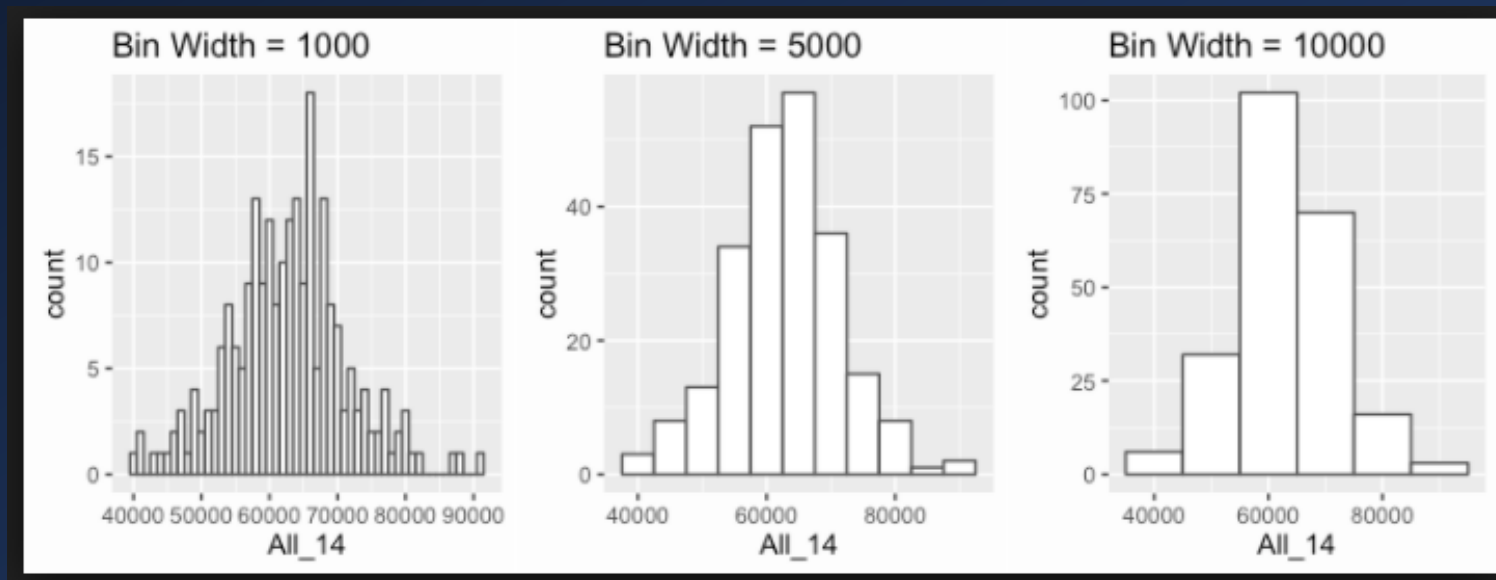


bimodal



multimodal

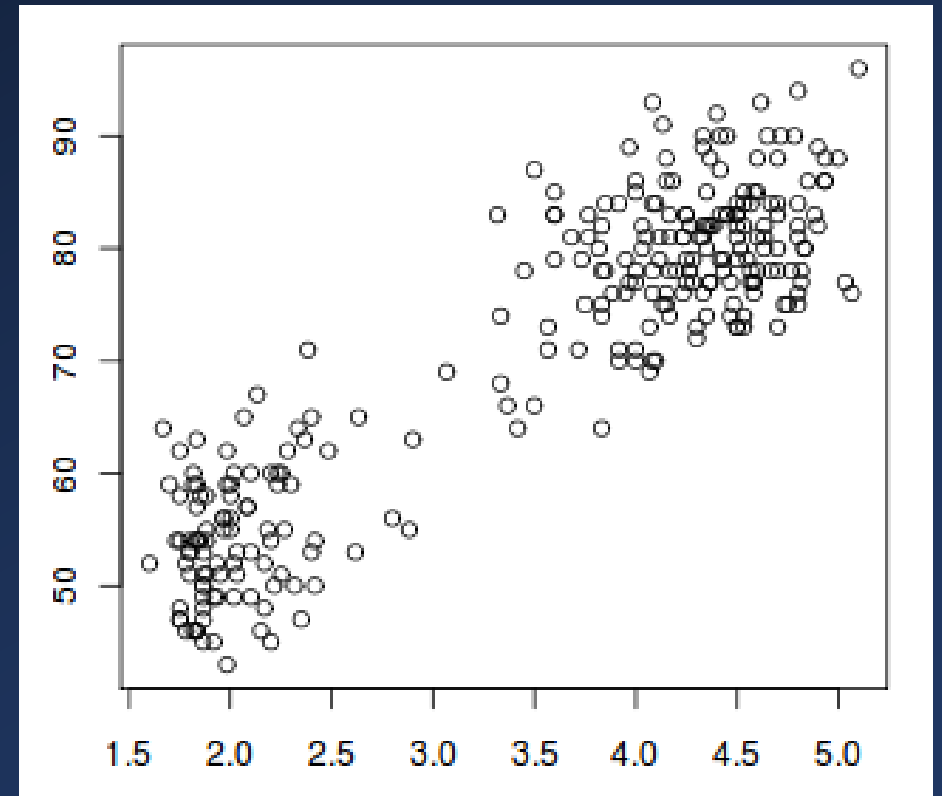
- In order to use a histogram, we simply require a variable that takes continuous numeric values.
- Plot data using different bin widths





- Scatter plots

- Plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.
- If the points are color-coded or encoded with different shapes, one additional variable can be displayed.
- A scatter plot can suggest various kinds of **correlations** between variables with a certain confidence interval.

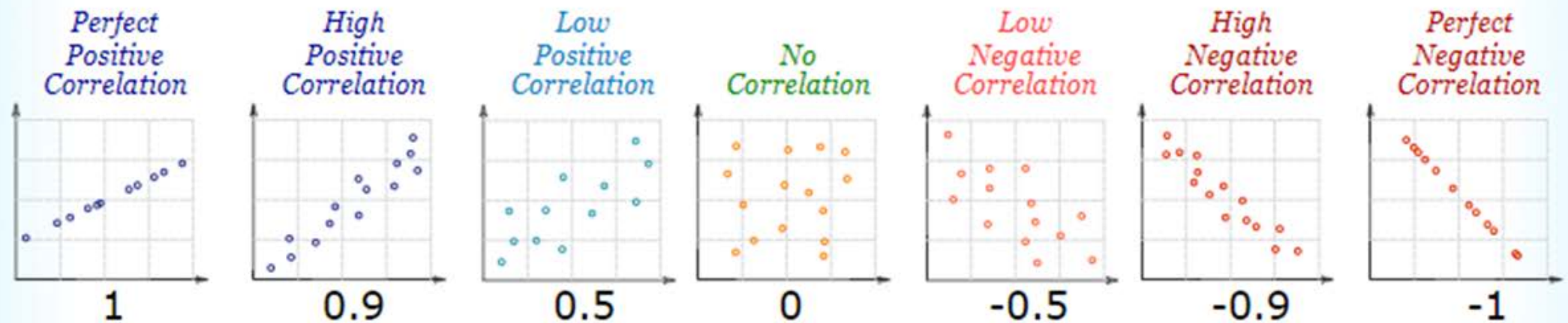


# Correlations

- A statistical technique
- Can show whether and how strongly pairs of variables are related
- e.g. height and weight
- Positive Correlation indicates the extent to which those variables increase or decrease in parallel;
- Negative Correlation indicates the extent to which one variable increases as the other decreases.

# Correlations

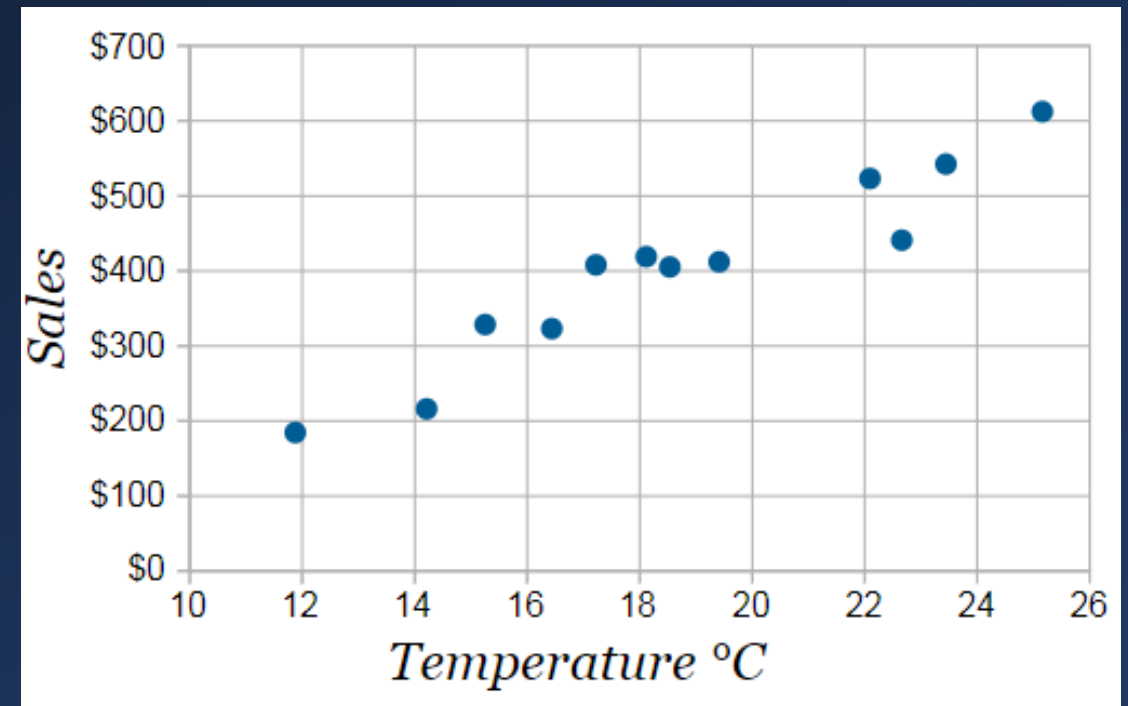
- Correlations can have a value - **Correlation Coefficient  $r$**  :
  - 1** is a perfect positive correlation
  - 0** is no correlation (the values don't seem linked at all)
  - 1** is a perfect negative correlation





- Scatter Plot of Ice Cream Sales


- When temperature increase, the sales of ice cream also increase
- The two variables (i.e., temperature & sales of ice cream) increase in parallel, they are positively correlated.



- Formula to calculate the correlation coefficient:

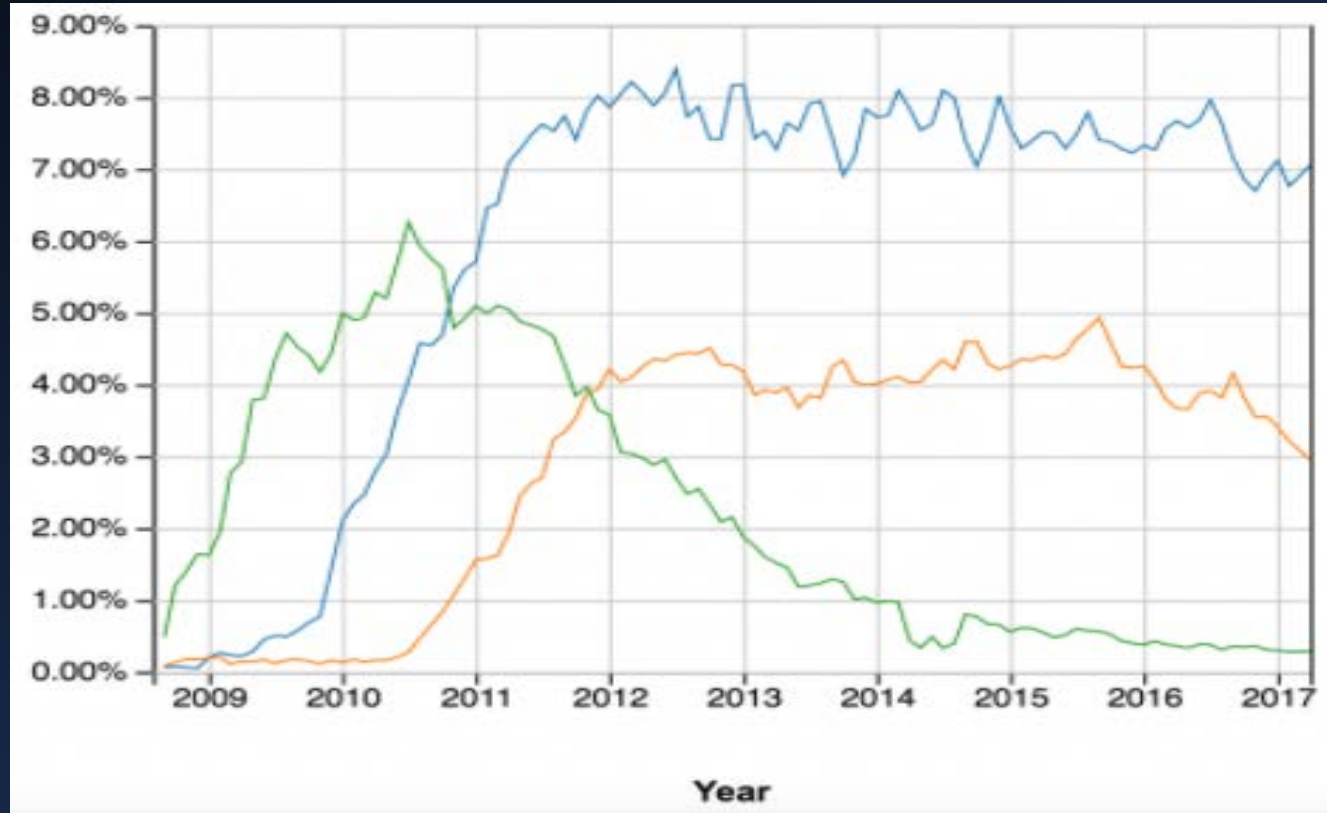
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Or just make a Scatter Plot, and look at it!
- High correlated variables provide same or similar information (or, **redundant information**) about your data, this suggest that you may want to remove one of the variables to make the analysis simpler.

- 
- Correlation is **NOT** causation!
  - The classic example: ice cream and murder

The rates of violent crime and murder have been known to jump when ice cream sales do, which means ice cream sales and murder rates are positively correlated. But, ice cream doesn't **cause** murder. Presumably, buying ice cream doesn't turn you into a killer.

# Reveal general trends



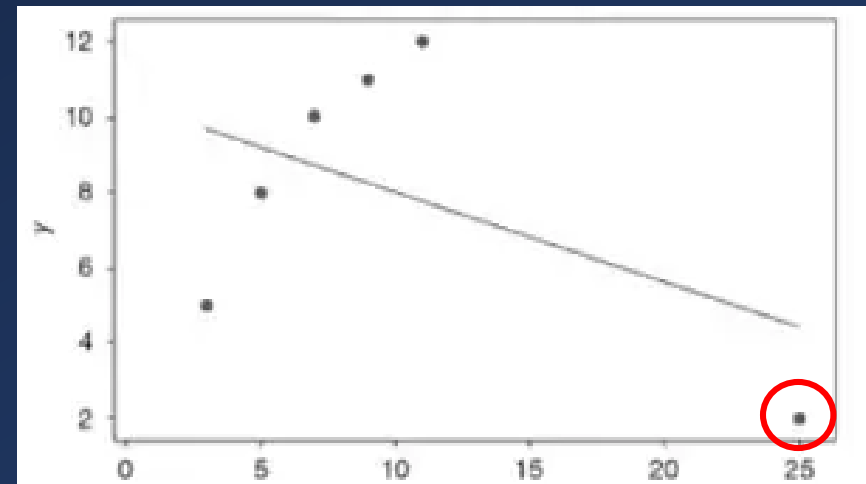
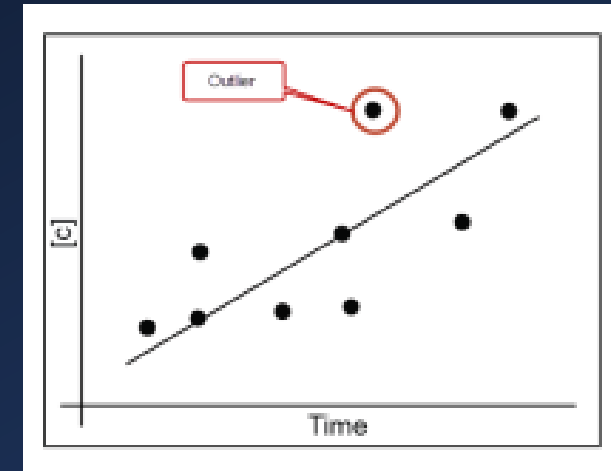
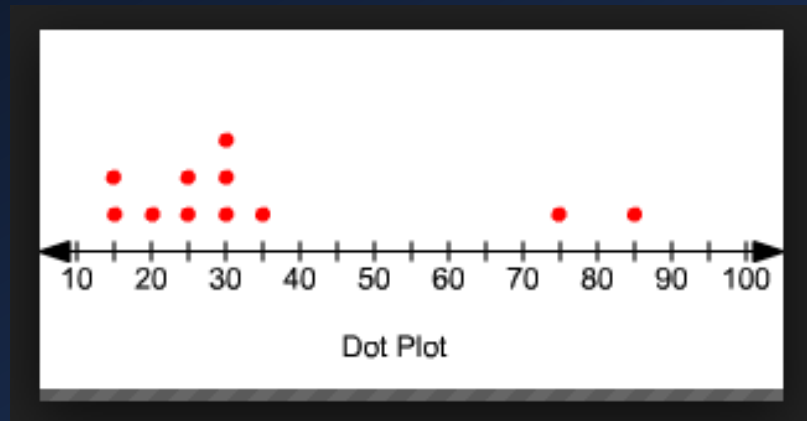
Reveal a variable moving in a certain direction



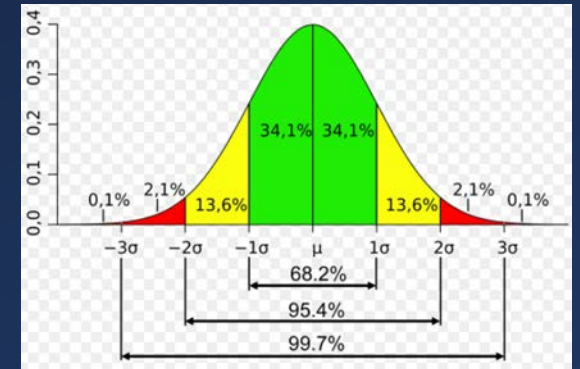


# Identify outliers

- An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population.



- Outliers indicate potential problems with the data and may need to be eliminated in some applications.
- Identification of outliers
  - Box plots or box and whisker plots
    - provide a graphical depiction of data distribution and extreme values
    - can be used as an initial screening tool for outliers as they
  - Probability plots
    - used for graphically displaying a data set's conformance to a normal distribution
    - can be used as an initial screening tool for outliers as they
  - Dion's test
  - Rosner's test



# Summary

- Steps in data process
- Summary statistics
- Histogram
- Correlations

