Data Analytics and Mathematical Statistics
Day 2

Source: Towards Data Science, Spring Board, The Software Report

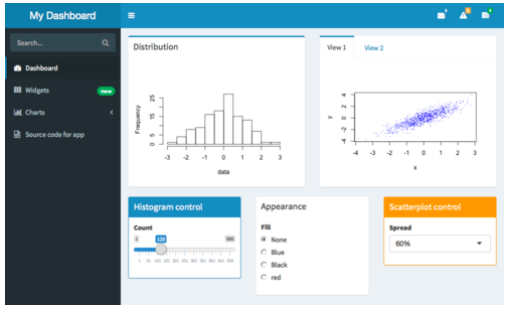
Agenda

Lecture 2 and Tutorial 2

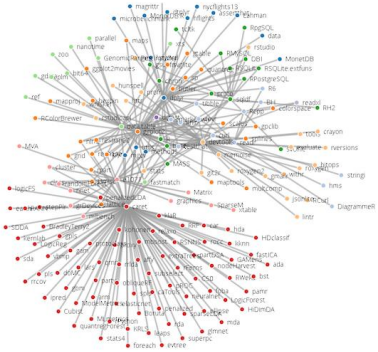
- Cleaning and exploring our data
  - How to clean datasets
  - Descriptive statistics
  - Data visualization
  - Hands-on data cleaning and exploration

A quick intro to R

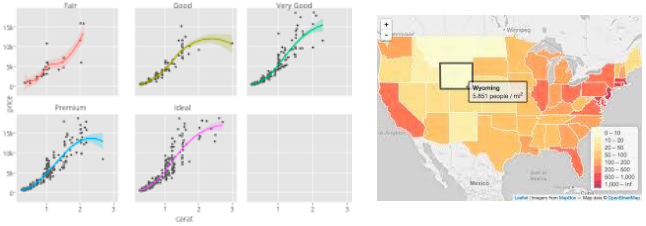
Open-source



Dashboard Capabilities



Availability of Packages



Great for Visualisations

Packages

Available CRAN Packages By Name

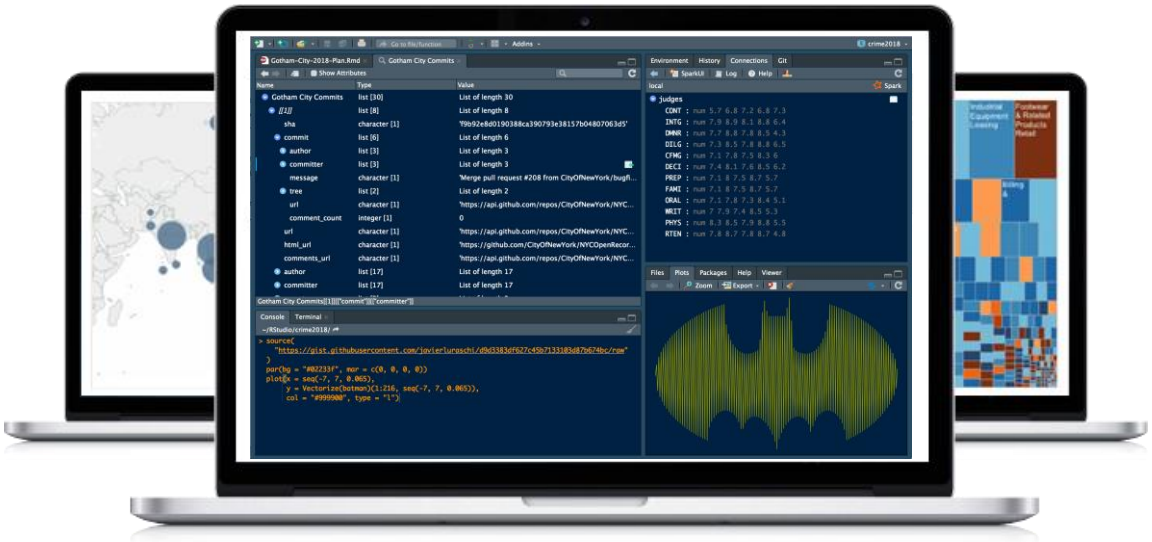
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
ABC.RAP	Array Based CpG Region Analysis Pipeline
ABCanalysis	Computed ABC Analysis
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2
abcrf	Approximate Bayesian Computation via Random Forests
abctools	Tools for ABC Analyses
abd	The Analysis of Biological Data
abe	Augmented Backward Elimination
abf2	Load Gap-Free Axon ABF2 Files
ABHgenotypeR	Easy Visualization of ABH Genotypes
abind	Combine Multidimensional Arrays
abjutils	Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association
abn	Modelling Multivariate Data with Additive Bayesian Networks
abnormality	Measure a Subject's Abnormality with Respect to a Reference Population
abodOutlier	Angle-Based Outlier Detection
ABPS	The Abnormal Blood Profile Score to Detect Blood Doping
AbsFilterGSEA	Improved False Positive Control of Gene-Permuting GSEA with Absolute Filtering
AbSim	Time Resolved Simulations of Antibody Repertoires
abundant	High-Dimensional Principal Fitted Components and Abundant Regression
Ac3net	Inferring Directional Conservative Causal Core Gene Networks
ACA	Abrupt Change-Point or Aberration Detection in Point Series
acc	Exploring Accelerometer Data
accelerometry	Functions for Processing Accelerometer Data
accelmissing	Missing Value Imputation for Accelerometer Data
AcceptanceSampling	Creation and Evaluation of Acceptance Sampling Plans
ACCLMA	ACC & LMA Graph Plotting
accrual	Bayesian Accrual Prediction
accrued	Data Quality Visualization Tools for Partially Accruing Data
accSDA	Accelerated Sparse Discriminant Analysis
ACD	Categorical data analysis with complete or missing responses
ACDm	Tools for Autoregressive Conditional Duration Models
acebayes	Optimal Bayesian Experimental Design using the ACE Algorithm
acepack	ACE and AVAS for Selecting Multiple Regression Transformations

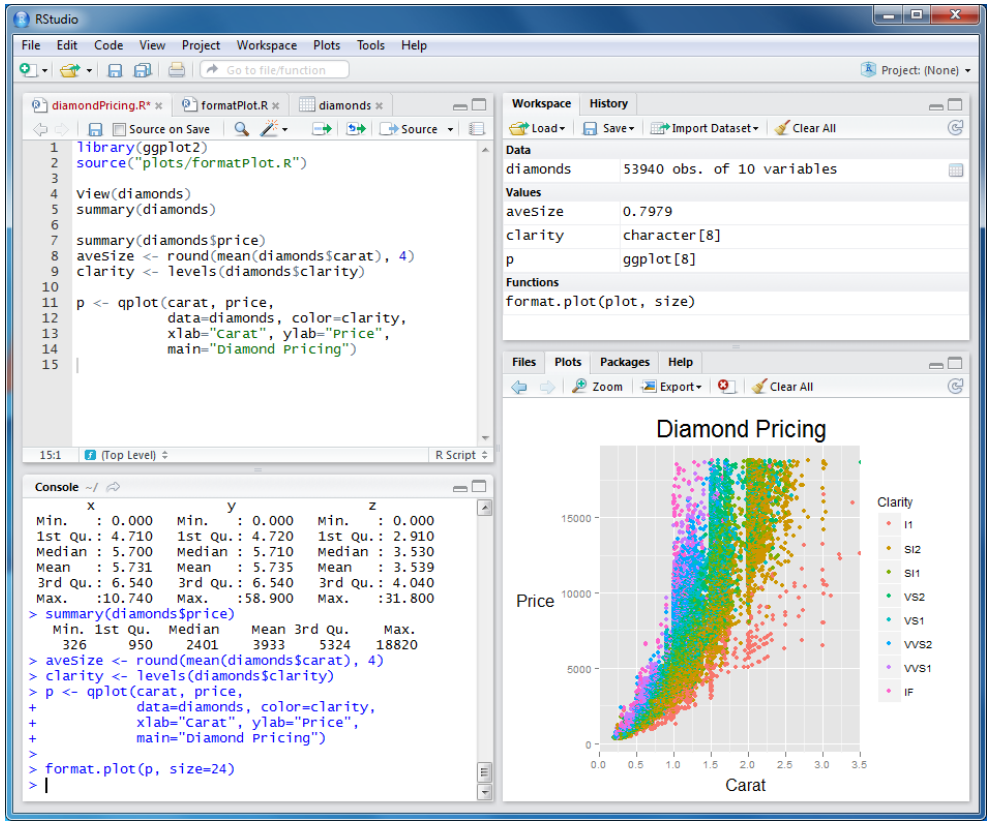
Source: https://cran.r-project.org/web/packages/available_packages_by_name.html

NUS National University of Singapore

RStudio®



RStudio®



Cleaning and exploring our data

Cleaning data

- Some of the things we do in cleaning data
  - Putting data in tidy format
  - Make the labels consistent
  - Joining tables together
  - Missing values

Anna Karenina Principle

Leo Tolstoy's book *Anna Karenina* begins:

" Happy families are all alike; every unhappy family is unhappy in its own way"

Like families, tidy datasets are all alike but every messy dataset is messy in its own way - Hadley Wickham

Tidy Data

According to Wickham, tidy data displays the following attributes.

- Each **variable** is a column
- Each **observation** is row

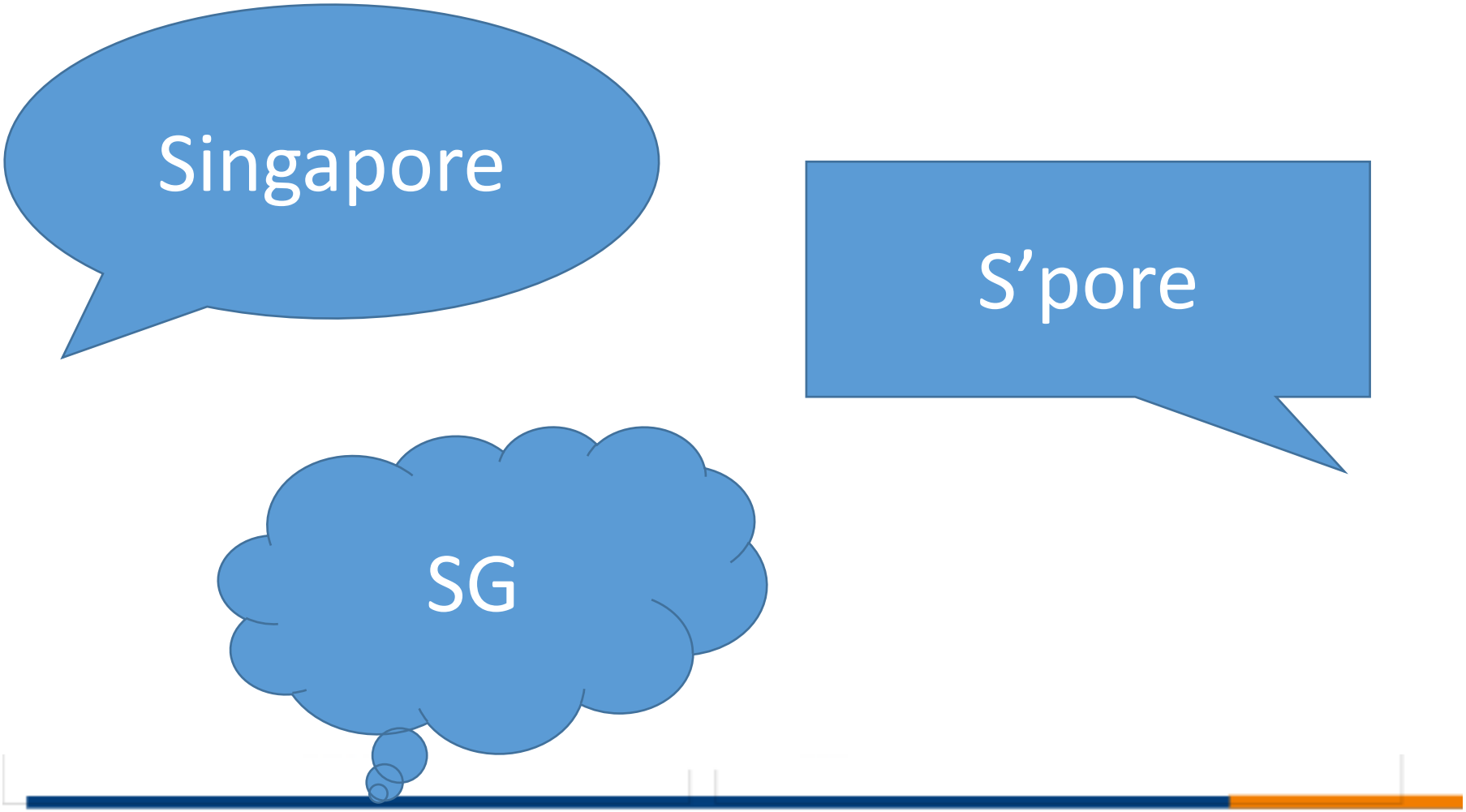
Tidy or NOT?

```
preg <- read.csv("preg.csv", stringsAsFactors = FALSE)
preg
#>           name treatmenta treatmentb
#> 1   John Smith         NA         18
#> 2     Jane Doe          4          1
#> 3 Mary Johnson          6          7
```

- Which one is tidy?
- What are Observations and Variables?

```
preg2
#>           name treatment  n
#> 1     Jane Doe         a  4
#> 2     Jane Doe         b  1
#> 3   John Smith         a NA
#> 4   John Smith         b 18
#> 5 Mary Johnson         a  6
#> 6 Mary Johnson         b  7
```

Making the labels consistent



What is the purpose of joins?

Table 1

ID	First Name	Last Name	Publisher Type
20034	Adam	Davis	Independent
20165	Ashley	Garcia	Big
20233	Susan	Nguyen	Small/medium

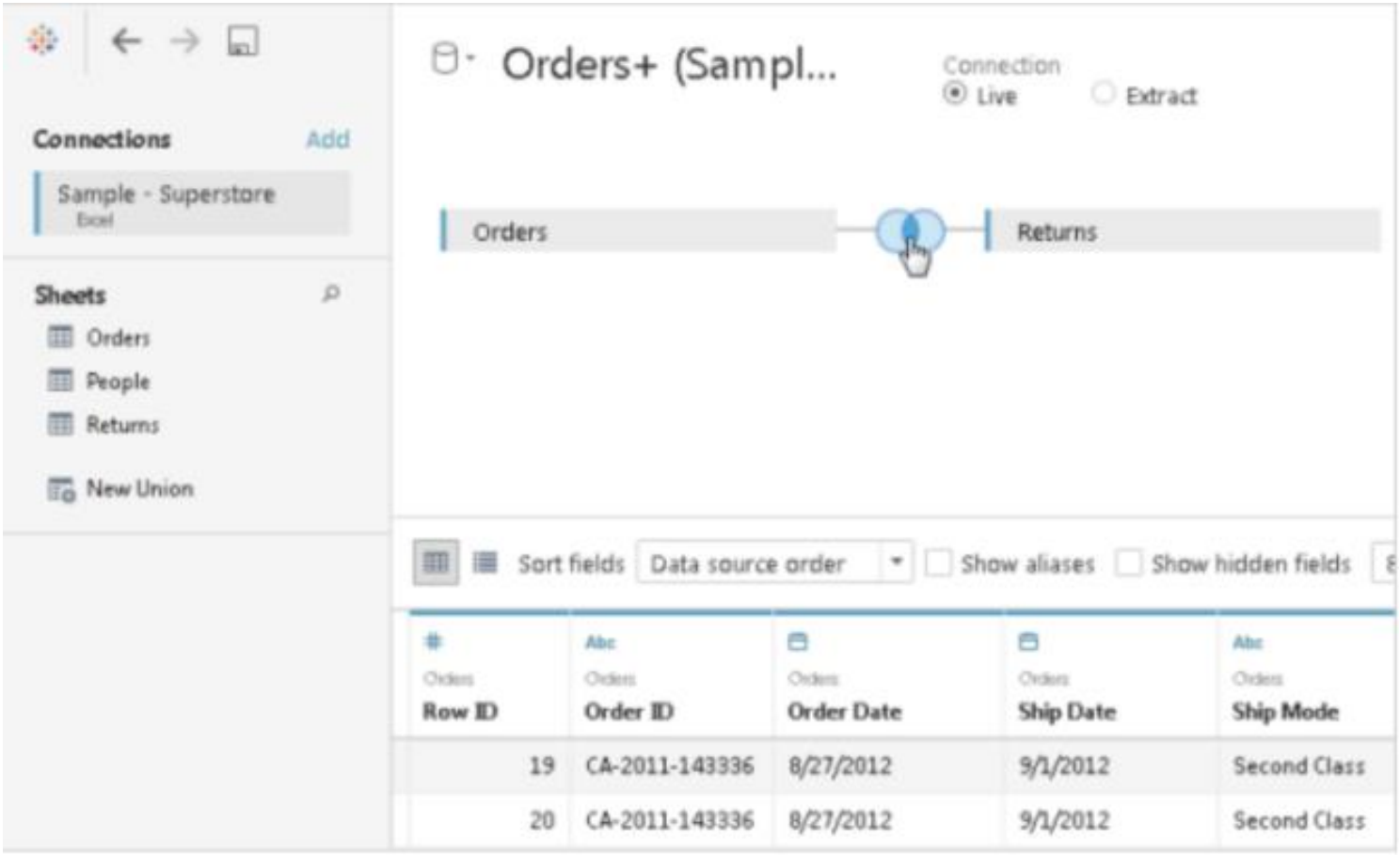
Table 2

Book Title	Price	Royalty	ID
Weather in the Alps	19.99	5,000	20165
My Physics	8.99	3,500	20800
The Magic Shoe Lace	15.99	7,000	20034

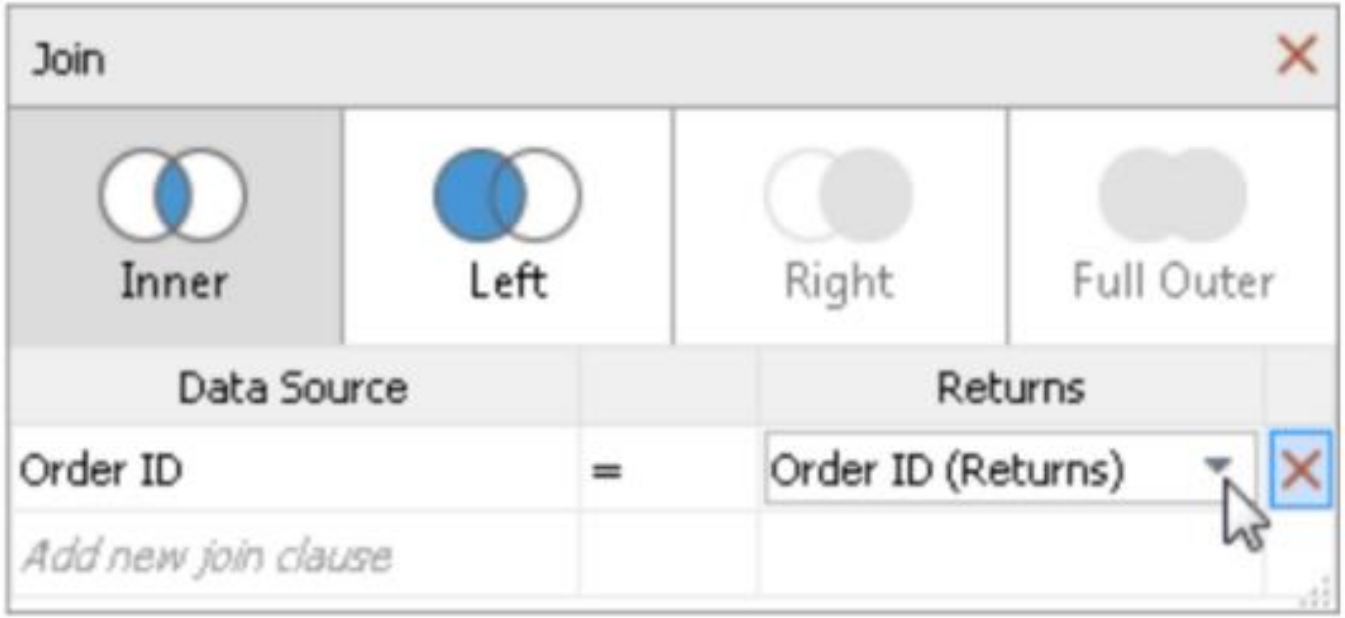
Inner Join

ID	First Name	Last Name	Publisher Type	Book Title	Price	Royalty
20034	Adam	Davis	Independent	The Magic Shoe Lace	15.99	7,000
20165	Ashley	Garcia	Big	Weather in the	19.99	5,000

Create A Join Between Sheets



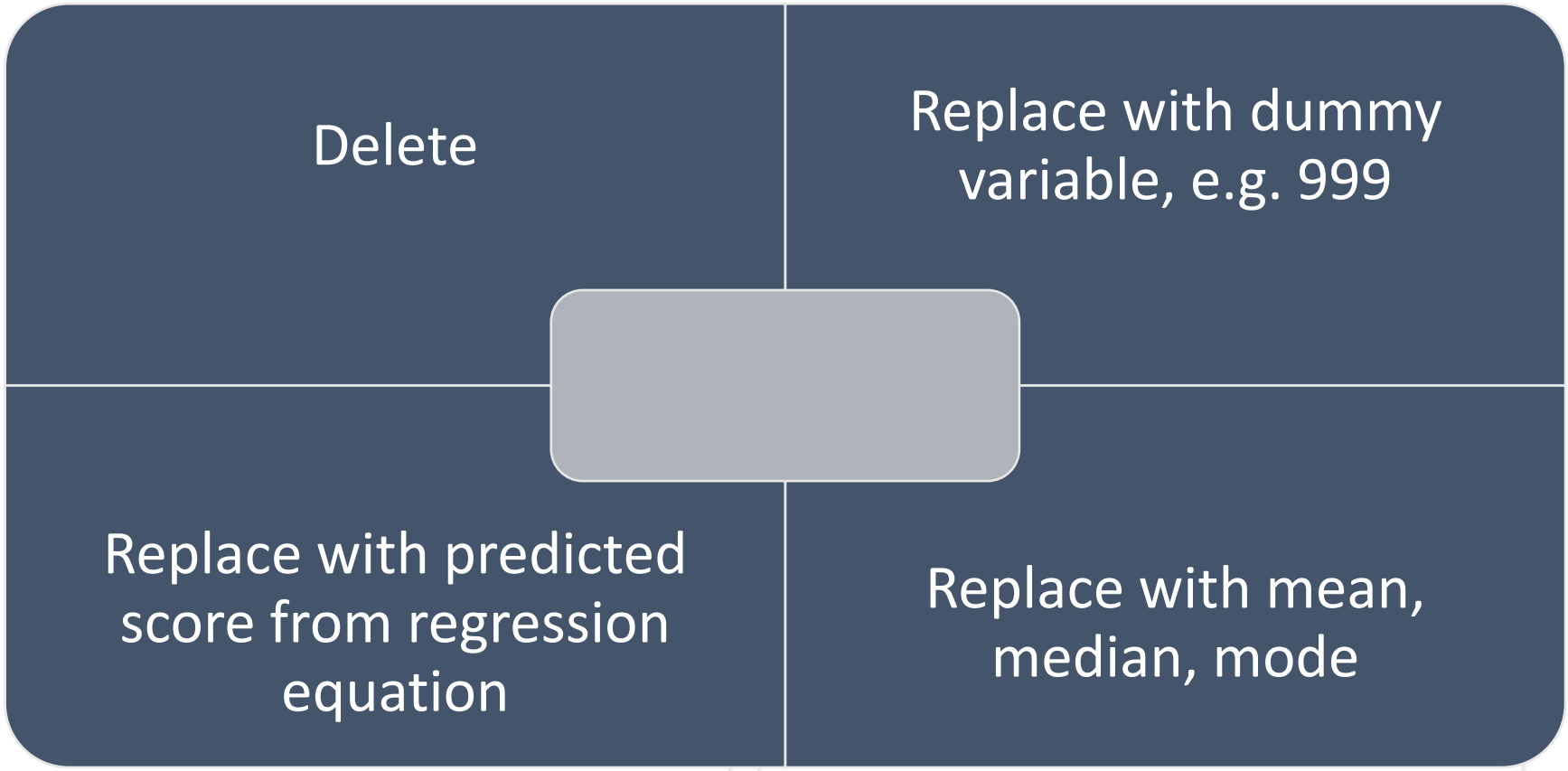
4 Types of Joins: Inner, Left, Right, Full Outer



How do the join types differ from each other?

Join Type	Result	Description
Inner	When you use an inner join to combine tables, the result is a table that contains values that have matches in both tables.	
Left	When you use a left join to combine tables, the result is a table that contains all values from the left table and corresponding matches from the right table. When a value in the left table doesn't have a corresponding match in the right table, you see a null value in the data grid.	
Right	When you use a right join to combine tables, the result is a table that contains all values from the right table and corresponding matches from the left table. When a value in the right table doesn't have a corresponding match in the left table, you see a null value in the data grid.	
Full outer	When you use a full outer join to combine tables, the result is a table that contains all values from both tables. When a value from either table doesn't have a match with the other table, you see a null value in the data grid.	
Union	Though union is not a type of join, union is another method for combining two or more tables by appending rows of data from one table to another. Ideally, the tables that you union have the same number of fields, and those fields have matching names and data types. For more information about union, see Union Your Data.	

Missing values



A bit of background

SG's air quality index

Based upon six criteria air pollutants: CO, NO_2, O_3, PM2.5, PM10 and SO_2

Air Quality Descriptor

PSI Value	PSI Descriptor
0 - 50	Good
51 - 100	Moderate
101 - 200	Unhealthy
201 - 300	Very unhealthy
Above 300	Hazardous

Health Advisories

24-hr PSI	Healthy Persons	Elderly, pregnant women, children	Persons with chronic lung disease, heart disease
≤ 100 (Good/Moderate)	Normal activities	Normal activities	Normal activities
101-200 (Unhealthy)	Reduce prolonged or strenuous outdoor physical exertion	Minimise prolonged or strenuous outdoor physical exertion	Avoid prolonged or strenuous outdoor physical exertion
201-300 (Very unhealthy)	Avoid prolonged or strenuous outdoor physical exertion	Minimise outdoor activity	Avoid outdoor activity
>300 (Hazardous)	Minimise outdoor activity	Avoid outdoor activity	Avoid outdoor activity

Some key questions

What is the average?

What is the spread in my data?

How is the data for each variable distributed

Some key questions

Does the data cluster towards lower or higher values?

Are there any unusual data features, e.g. outliers?

What relationship exists between my variables?

Descriptive statistics and viz

- After collecting our data, we often wish to do some initial exploratory analysis, to understand some general features
  - E.g. what are some values around which our data tend to cluster
  - How spread out is our data

- Descriptive statistics and graphical (visualisation methods) can help us to gain such insights

Descriptive statistics and viz

Where does my data cluster?

- Data from many systems tend to cluster around a central value, which can be taken as a representative of the sample, e.g. temperature in Singapore

Source: Psychology Today

Arithmetic mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Indicates where the centre of the distribution lies but strongly influenced by outliers

Outliers

- Do not seem to conform to the distribution of values of the rest of the data

1. Assumptions on conditions no longer hold
2. Interfering processes that also contribute to data
3. Errors of faulty instrumentation, measurement, observation or recording of data

- Do we keep or remove outliers?

Median

- 50^{th} percentile
- Central value in an ordered set of data, or the average of the two central values
- That is, we sort the data in ascending or descending order. The median is the middle-ranked value if the sample size is odd, or the average of the two middle ranked values if the sample size is even

Mean and median

- The inter-arrival times (min) for an internal bus service X as measured at a bus-stop in UTown are as follows: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6, 13.1

- What is the arithmetic mean and median inter-arrival time of the bus service?

Example – mean

- Mean inter-arrival time for the eight observations collected from the bus stop

- $\bar{x} = (12.6 + 12.9 + \cdots + 13.1)/8 = 104/8 = 13.0$

Example – median

- Ranking data from smallest to highest
  - 12.3, 12.6, 12.6, 12.9, 13.1, 13.4, 13.5, 13.6
- Number of observations, N = 8
  - Take average between 4th and 5th ranked values
- $\text{Median} = (12.9 + 13.1)/2 = 13.0$

Geometric Mean

$$\bar{x} = \sqrt[n]{\left(x_1 \bullet x_2 \cdots \bullet x_n\right)}$$

- The geometric mean is used in averaging values that represent a rate of change
  - Variable follows an exponential, i.e. logarithmic law

- The geometric mean is always less than or equal to the arithmetic mean

Example – Geometric mean

- Population growth
  - Future increase in populations of towns and cities is proportional to the current population

- According to a census conducted in 1970 and again in 1990, the population of a city had increased from 240,000 to 320,000

- In order to verify, for purposes of design, the per capita consumption of water in the intermediate period, estimate the population in 1980

Example – Geometric mean

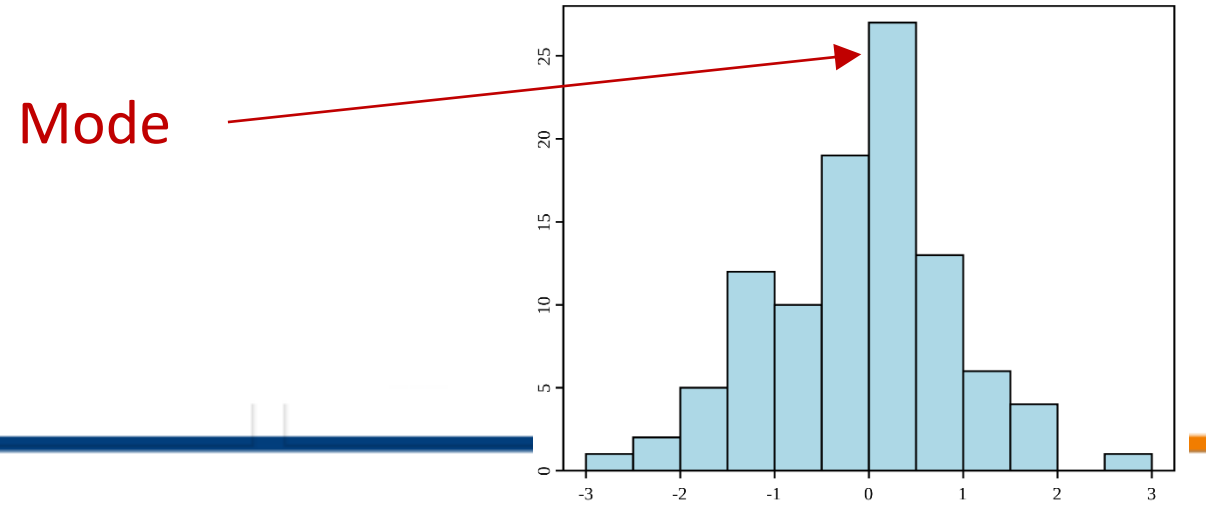
- The measure of central tendency to use in this situation is the geometric mean of the two numbers, which is:

$$\bar{x} = (240{,}000 \times 320{,}000)^{1/2} = 277{,}128$$

- Note that the arithmetic mean in this case is 280,000

Mode

- Sample mode is the most frequently occurring data value
- Typically used together with a histogram
- Rarely used with numerical variables; more useful with categorical variables



Discussion

	SO_2 ($\mu g/m^3$)	PM10 ($\mu g/m^3$)	PSI	Rainfall (mm)
Mean		23.4	51.4	15.37
G Mean		22.9	51.0	8.13
Median		23.0	52.0	8.75
Mode		21	52	2

What is the spread in our data?

- Besides measures of central tendency, what about how “spread out” our data is?
- To answer this question, we use measures of dispersion

Measures of dispersion

- If we want to know the dispersion of our data, an intuitive first step would be to rank our data, and look at the smallest and largest values

Range and IQR

- Range = maximum – minimum
  - What might be a disadvantage of the range?
  - Range is widely used for small sample sizes (say n = 5)
- Interquartile range (IQR) = Q3 – Q1 (75^{th} percentile – 25^{th} percentile)
- IQR vs range?

Variance / standard deviation

2, 3.1, 5, 6.5, 8.9, 100

- To get the dispersion of a set of data, an intuitive approach is to get the average distance of each value from some “centre”

Sample variance / standard deviation

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

- For the inter-arrival times of buses, we can determine the sample standard deviation as follows:

- $s = \sqrt{\frac{1}{7}[(12.6 - 13)^2 + (12.9 - 13)^2 + \cdots + (13.1 - 13)^2} = \sqrt{\frac{1.6}{7}} = 0.48\, mins$

Discussion

	SO2 (μg/m³)	PM10 (μg/m³)	PSI	Rainfall (mm)
Variance	104.6	23.6		363.17
Std dev	10.2	4.9		19.06
Range	46.0	23.0		101.50
IQR	13.8	5.0		15.50
Mean	17.8	23.4	51.4	15.37

Graphical methods

- Good way for us to visualise the variability and other properties of a set of data
- What form and shape does the data take? i.e. how is the data for each variable distributed?
- What relationship exists between my variables?

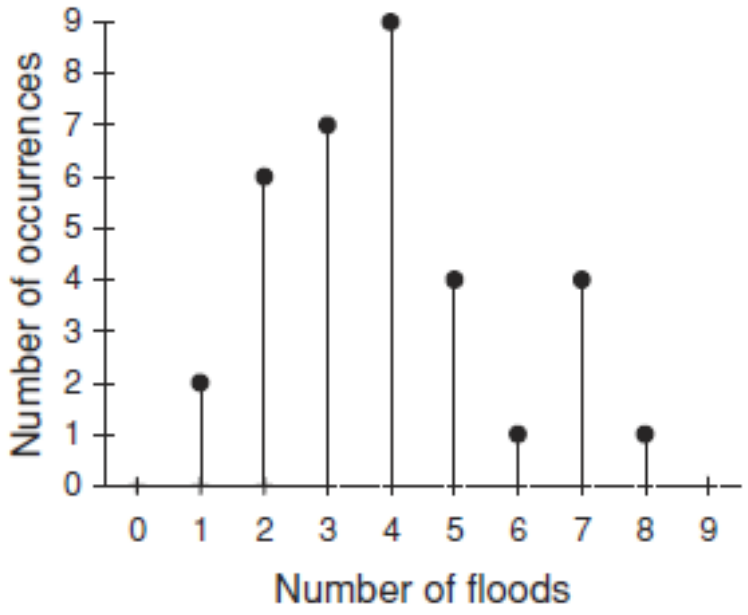
Line diagram or bar chart

- Occurrences of a discrete variable (e.g. roll of a die) can be classified on a line diagram or bar chart, or a categorical variable

- Horizontal axis gives the values of the discrete / categorical variable and the number of occurrences (frequency) are represented by heights of the vertical lines

Line diagram or bar chart

- Horizontal spread of these lines and their relative heights indicate the variability and shape of the data



Kottegoda and Rosso

Histogram (n at least say, 25)

- Range of data is divided into bins
  - We then count the number of data falling into each bin (frequency)
  - Compact summary of data

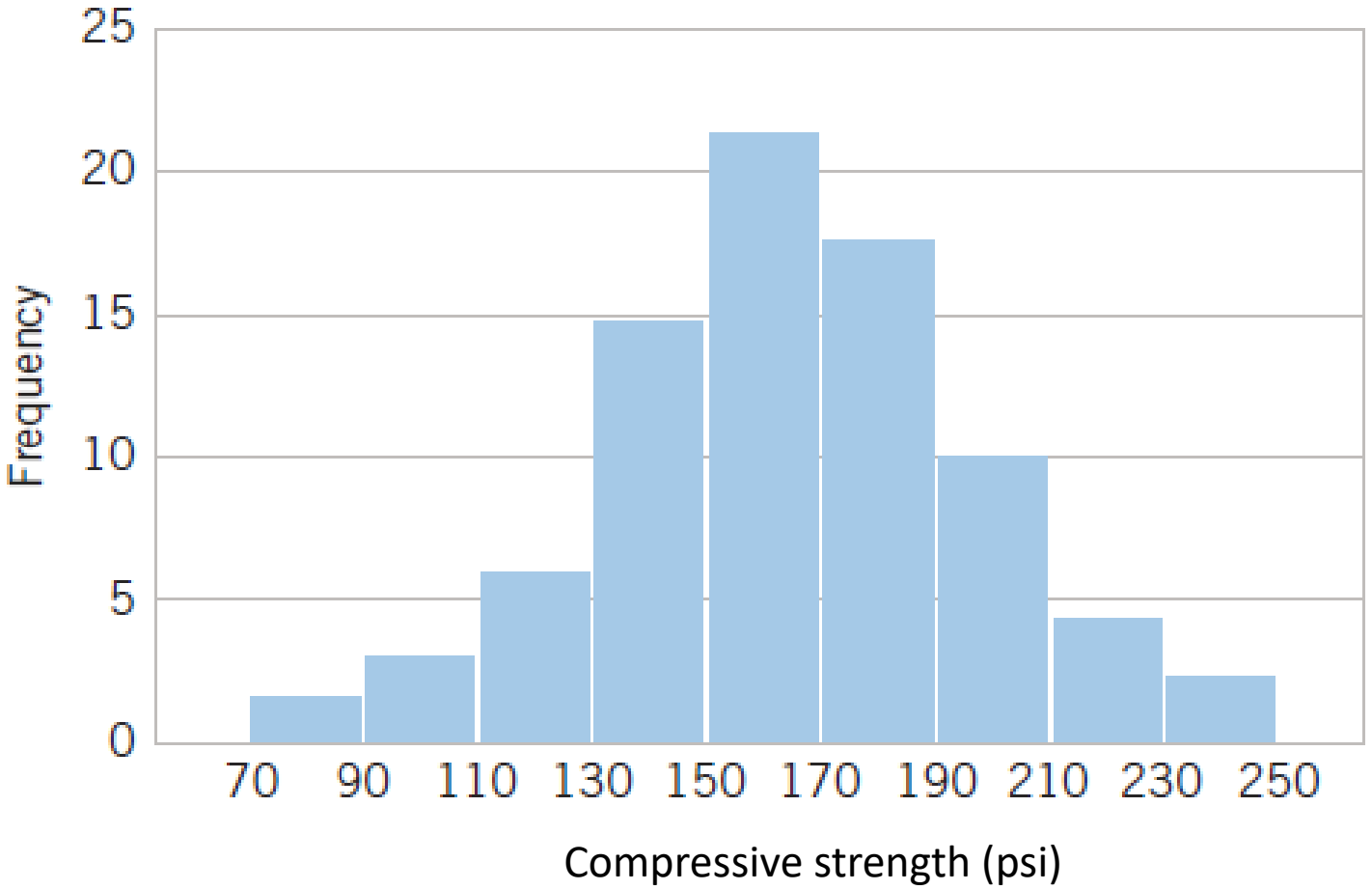
- However, when n is small…
  - Appearance of histogram can change significantly with the number of bins
  - Do not expect to see theoretical shape of the distribution (e.g. normal distribution)

Histogram

- Variability of data shown by the horizontal spread of the bins
  - Most common values are found in the tallest bins

- Histogram bins
  - Width of bins is usually made equal (but does not have to be)
  - Number of bins: Rule of thumb ~ $\sqrt{n}$
  - (at least 5 but not exceeding 25)

Histogram – Example



Histogram

- Provides visual information of the shape of the distribution, as well as the central tendency and scatter

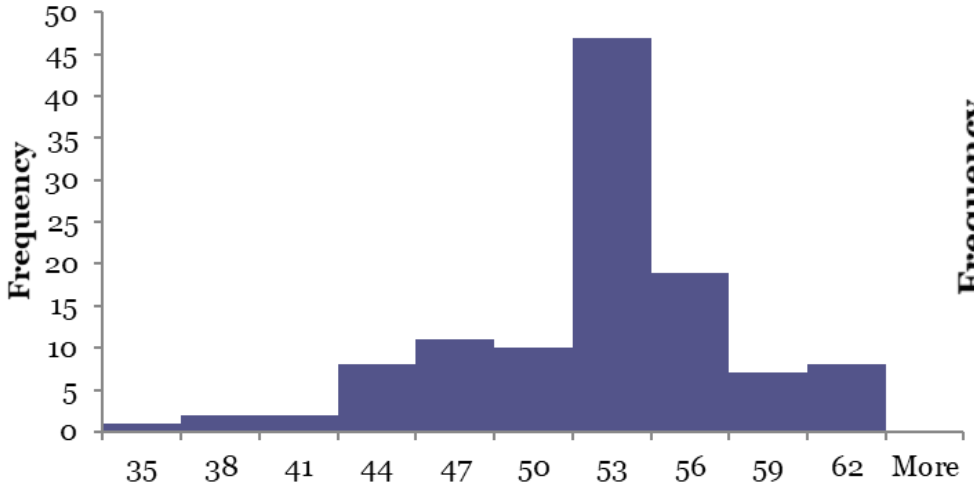
- Symmetric, bell-shaped distribution of data
  - Gives us an idea of what probability distribution to use as a model for the population
  - E.g. normal distribution is a reasonable model for the population

Histogram

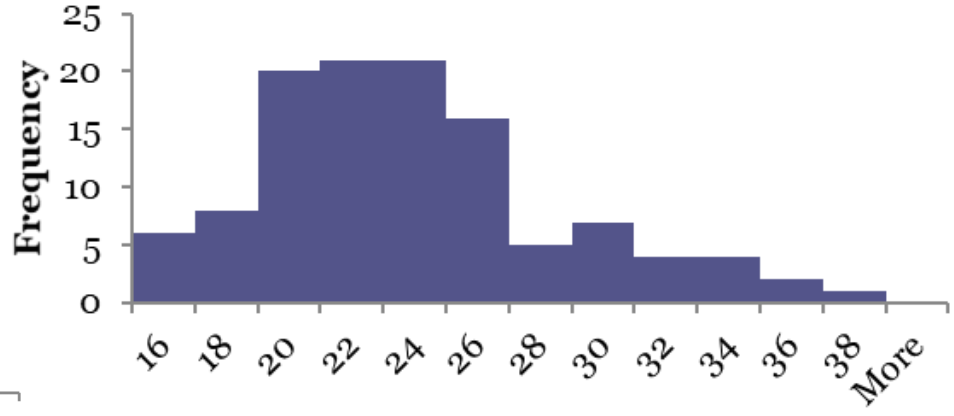
- For n > 100, histogram gives us quite reliable indicator of shape of distribution (e.g. mode, positive/negative skew)

- For smaller datasets, histograms may change significantly in appearance if the number and/or the width of the bins changes

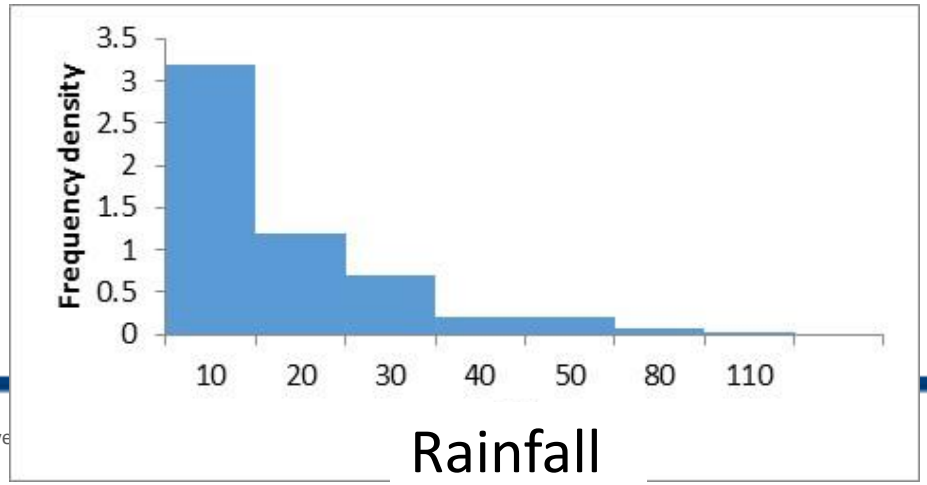
Histogram



PSI



PM10



Rainfall

Skew

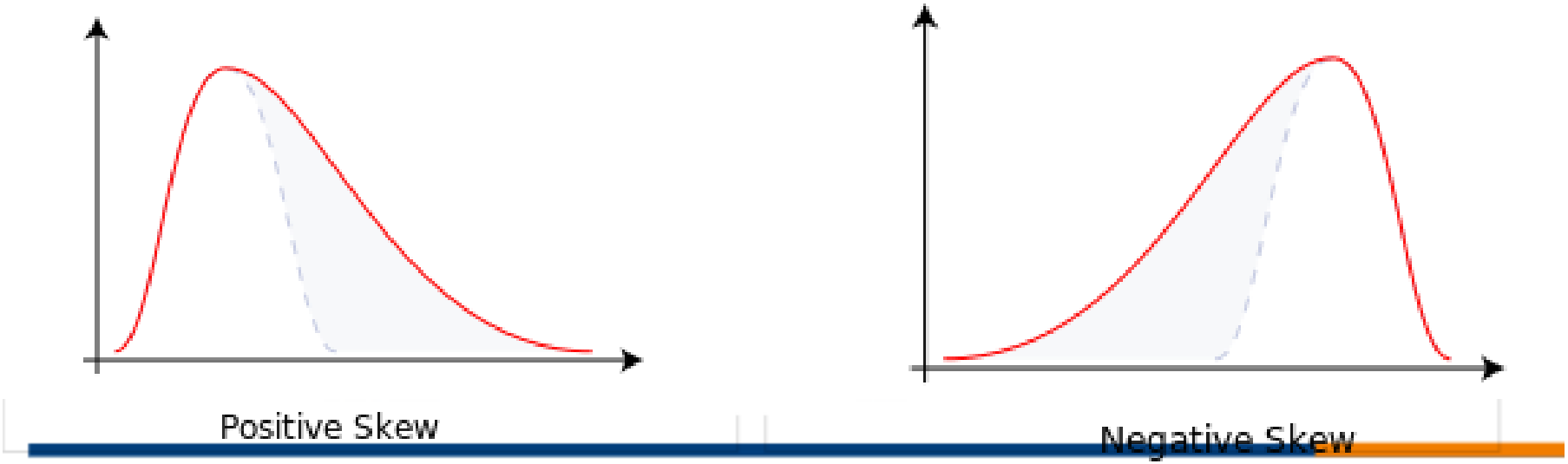
- Skew is a measure of symmetry
  - it measures the extent to which a distribution has long, drawn out tails on one side or the other
  - Skew = 0 is symmetrical
  - Positive skew (tail to the right)
  - Negative skew (tail to the left)

- A Normal distribution is symmetrical and has a skew of 0

Skew

- Positive (right) skew

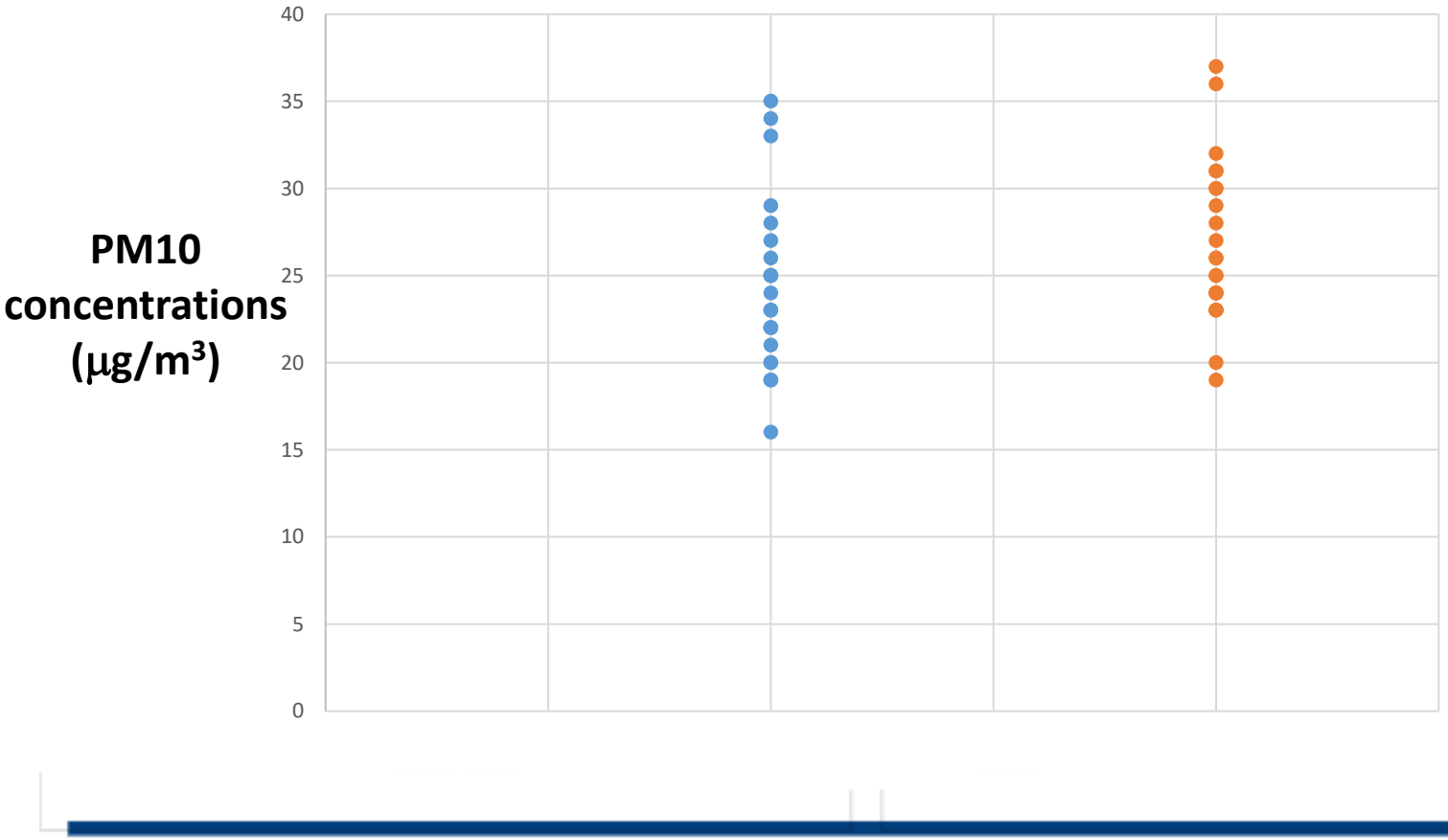
Negative (left) skew



Dotplot

- Helpful for showing a small pool of data (say, n < 25)
  - Easy to see central tendency and dispersion
- Used for visualizing distribution of a measure across different values of a dimension
- More difficult to catch pattern in the dispersion but unusual features can be seen

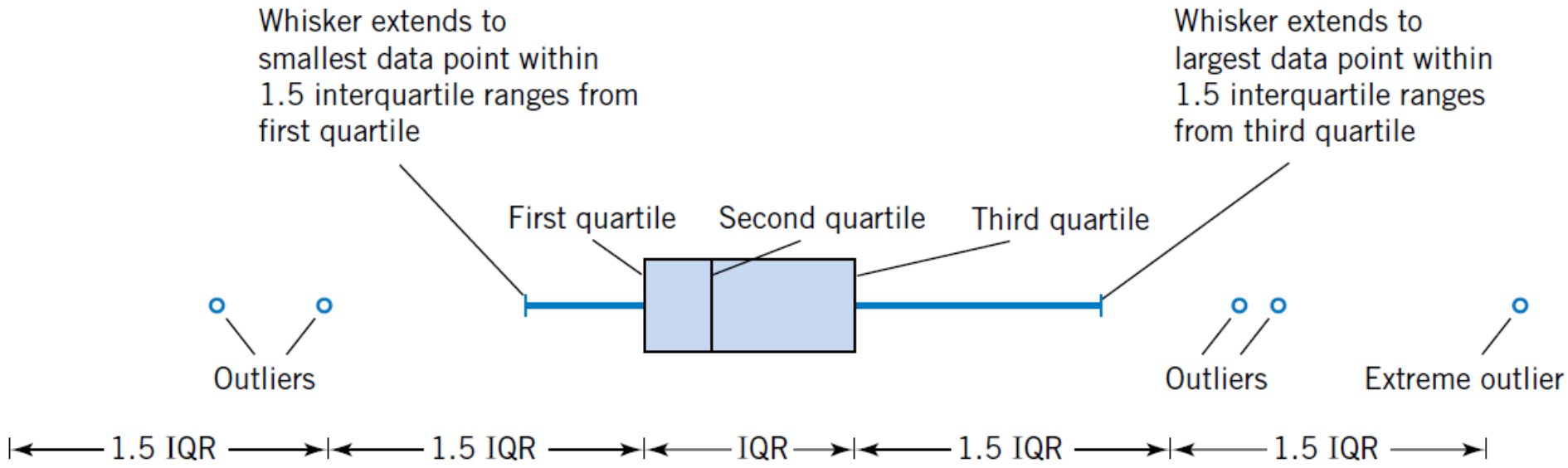
Dotplot



Boxplot

- Display 3 quartile levels (Q1, Q2, Q3)
- Lower whisker from Q1 to smallest data within 1.5 IQR of Q1
- Upper whisker from Q3 to largest data within 1.5 IQR of Q3
- Data less than 1.5 IQR from Q1 or greater than 1.5 IQR of Q3 are outliers
- Data is approximately symmetric if Q2 is equidistant from Q1 and Q3

Boxplot



Boxplot

- Helpful in highlighting distribution features, including range of data

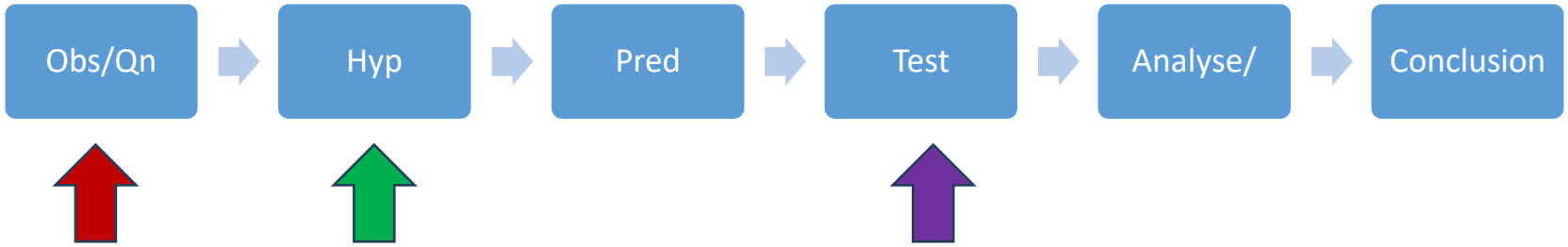
- Valuable means to compare data measuring related or similar characteristics

Scatterplot

- A scatterplot plots two measures against each other

- Allows us to see possible relationships, fine-tune our hypotheses or direct us in hypothesis formulation during exploratory studies

Our quantitative approach is not a rigid, one-directional process



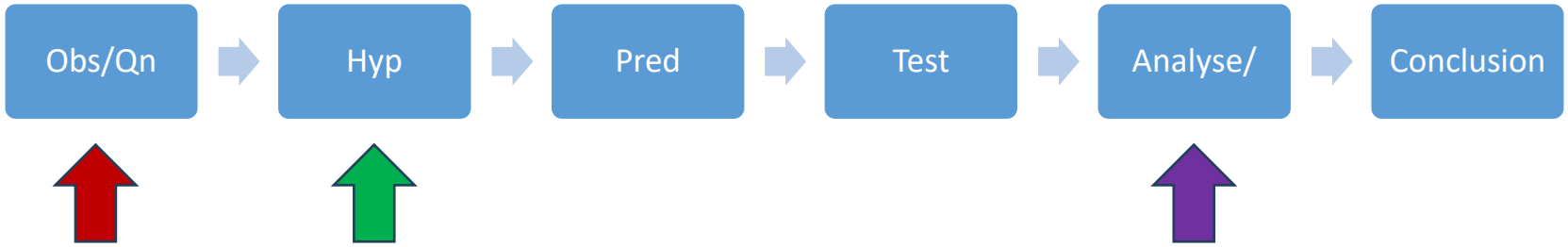
We don't have anything to inform our hypothesis:

"What affects students' grades? I have no idea"

Exploratory studies using e.g. a scatterplot can help inform what hypothesis to formulate

We can then see if collecting more data later confirms this hypothesis

Our quantitative approach is not a rigid, one-directional process



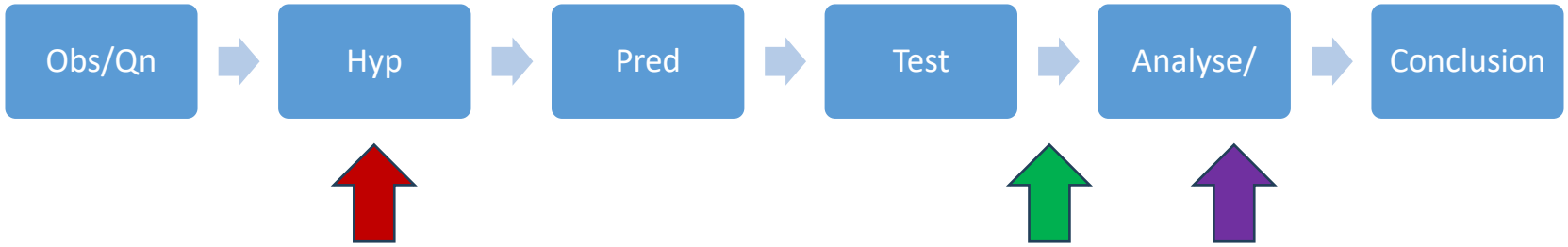
We don't have anything to inform our hypothesis:

"What affects students' grades? I have no idea"

Exploratory studies using e.g. a scatterplot can help inform what hypothesis to formulate

Or if the data we've collected in our exploratory study is all we can collect, we can then see if a formal analysis confirms the hypothesis

Our quantitative approach is not a rigid, one-directional process



We already have a hypothesis, but unsure of the direction: "Chocolates affects grades"

EDA using e.g. a scatterplot identifies a possible direction, say, positive. We can then tweak our initial hypothesis: "Chocolates improves grades"

We then see if formal analysis confirms this hypothesis

End of Day 2