FIT1043 Introduction to Data Science

Week 3

Ian K T Tan

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Learning Outcomes

Week 3

By the end of this week you should be able to:

- Comprehend more sophisticated group-by operations and graphing in Python
- Comprehend the power/**importance of data visualisation**
- Differentiate between **approaches for data visualisation**, and explain where each approach is appropriate to be used
- Explain/differentiate different concepts in **descriptive statistics**

Descriptive Statistics

From Introduction to Probability and Statistics for Engineers and Scientists, by S. M. Ross

*“ … objective is to **interpret key features** of a dataset numerically …”*

Descriptive Statistics

Descriptive statistics **summarise** aspects of the data

- Usually **lose information**, but gain **easy comprehension**
- Contrast with inferential statistics

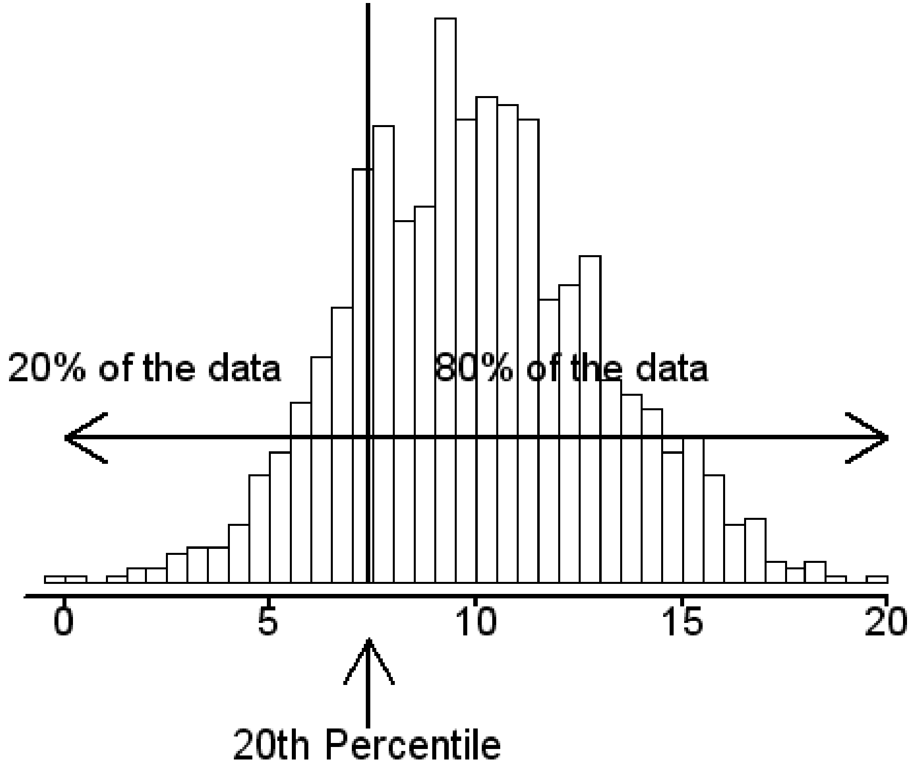
But what is a “statistic”?

- Let $\boldsymbol{y}$ denote a sample of data
- Then a statistic is any function $s(\boldsymbol{y})$ of the data

Some functions (statistics) more useful than others

- But all describe properties of the data

Measures of Centrality



https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/percentiles-rank-range/

Measures of Centrality

Mean, Mode, Median

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a sample of n data points

The most common measure of centrality, or averageness, is the arithmetic **mean**

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

The **mode** is the most frequently occurring value in the sample

Another common measure is the median, med(y)

- Value such that 50% of samples have values less than med(y)
- Easily found by sorting samples and finding middle sample

Mean vs Median

The **mean** uses all the values of the sample

- Any change to any sample changes the mean
- The mean can be changed as much as desired by changing just one sample by a large enough amount

The **median** uses at most two of the values of the sample

- Is very resistant to changes to the samples not in the middle
- Example:

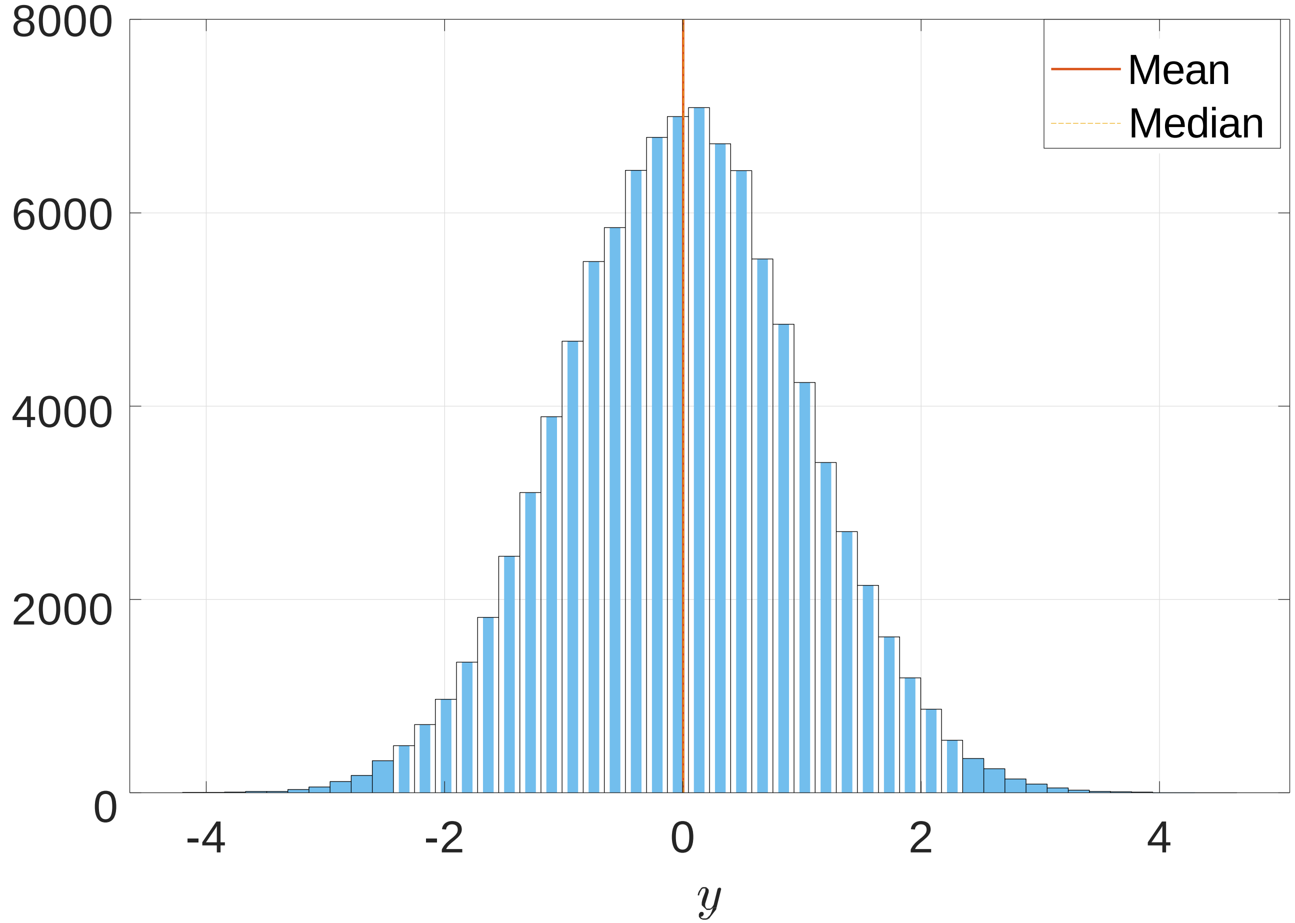
$y = (1, 2, 3, 4, 5) \quad \Rightarrow \bar{y} = 3, \quad \mathrm{med}(y) = 3$

$y = (1, 2, 3, 4, 50) \quad \Rightarrow \bar{y} = 12, \quad \mathrm{med}(y) = 3$

MONASH University

Mean vs Median

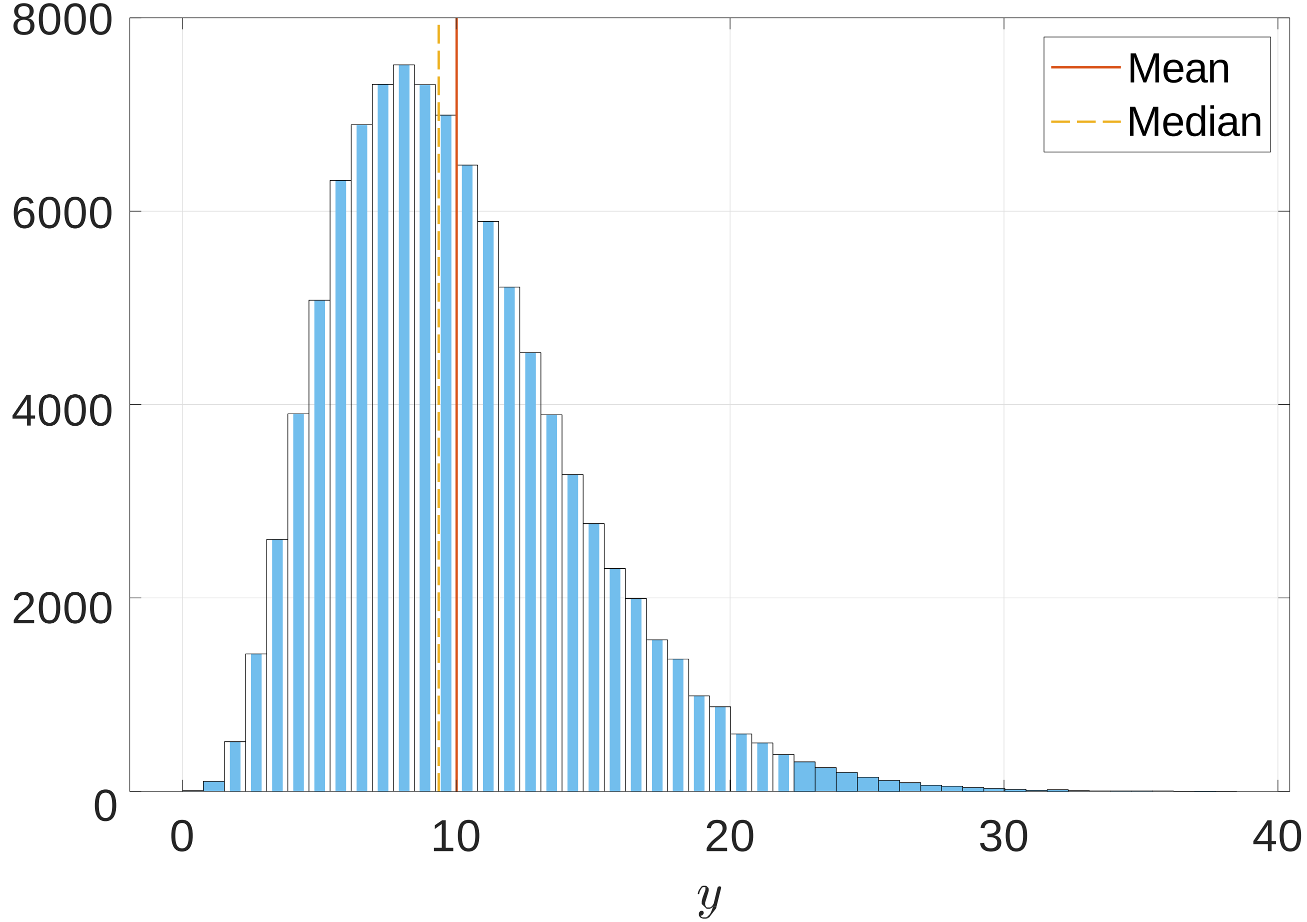
Symmetric Distribution



Symmetric distribution of data; **mean** and **median** (nearly) the same

Mean vs Median

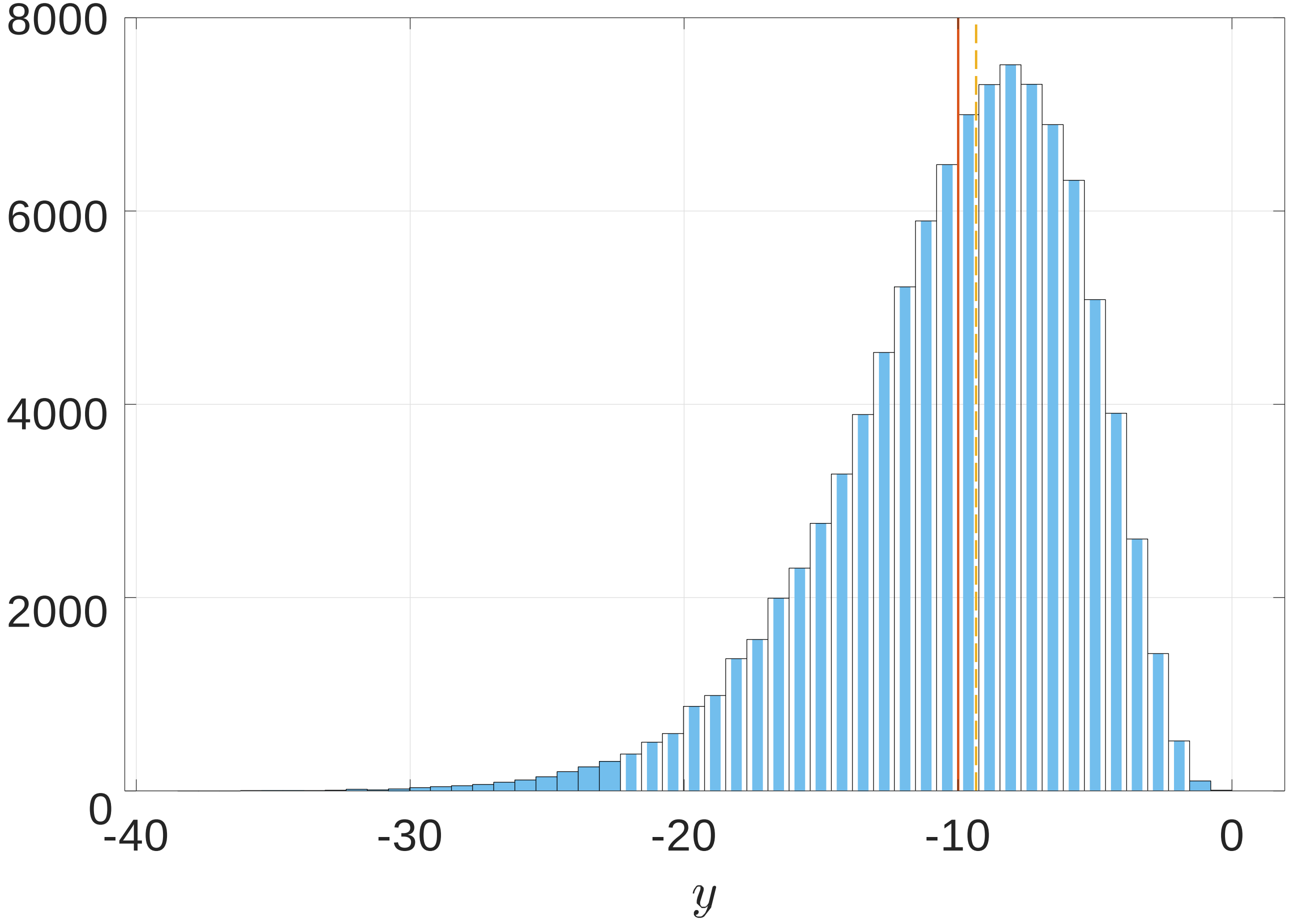
Positively Skewed



Positively skewed data; **mean greater than median**

Mean vs Median

Negatively Skewed



Negatively skewed data; **mean less than median**

Percentiles

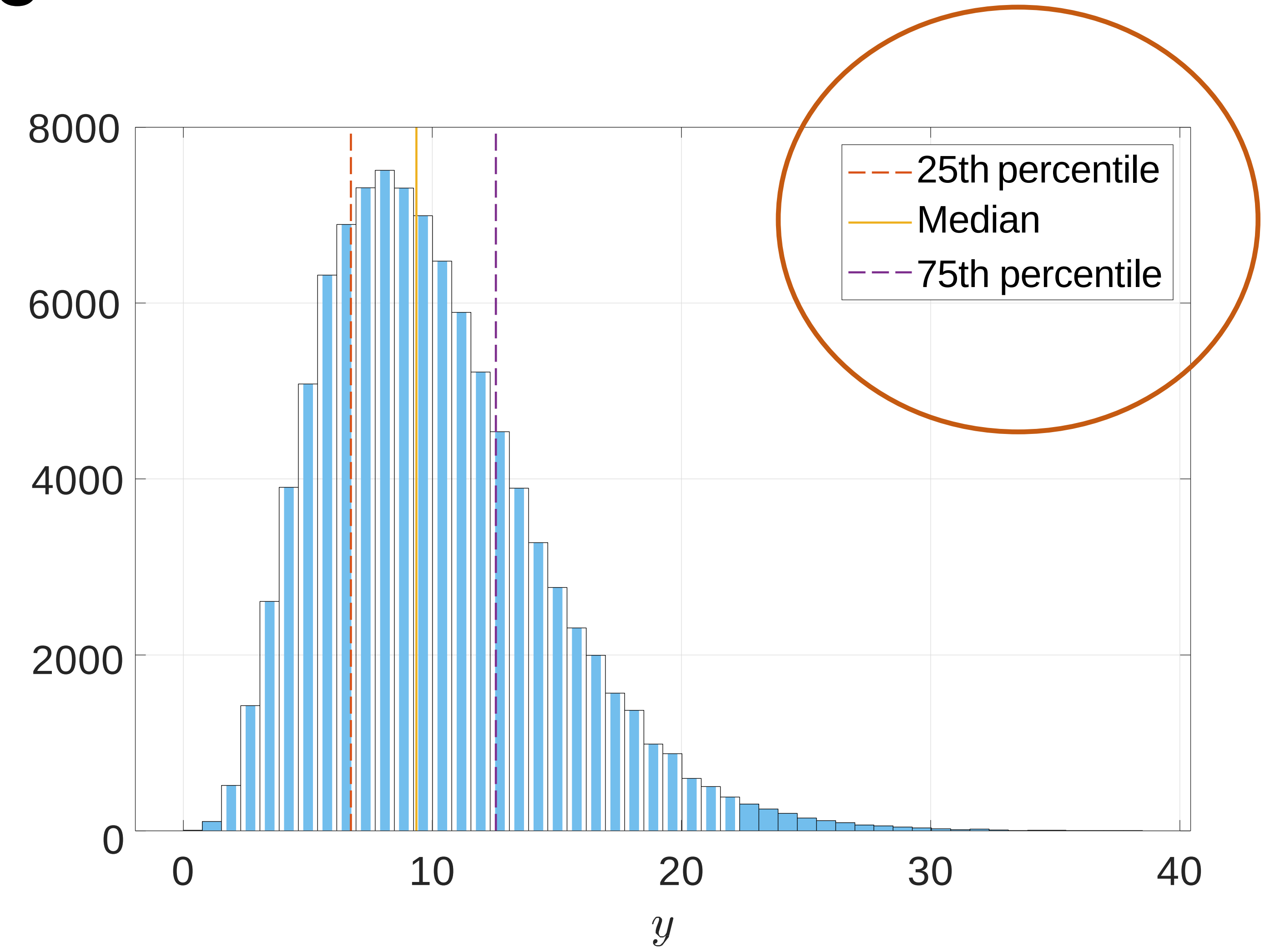
More generally, we can define the percentiles

- The p-th percentile is the value, $Q(\boldsymbol{y}, p)$ such that p% of the values of the sample are lower than $Q(\boldsymbol{y}, p)$

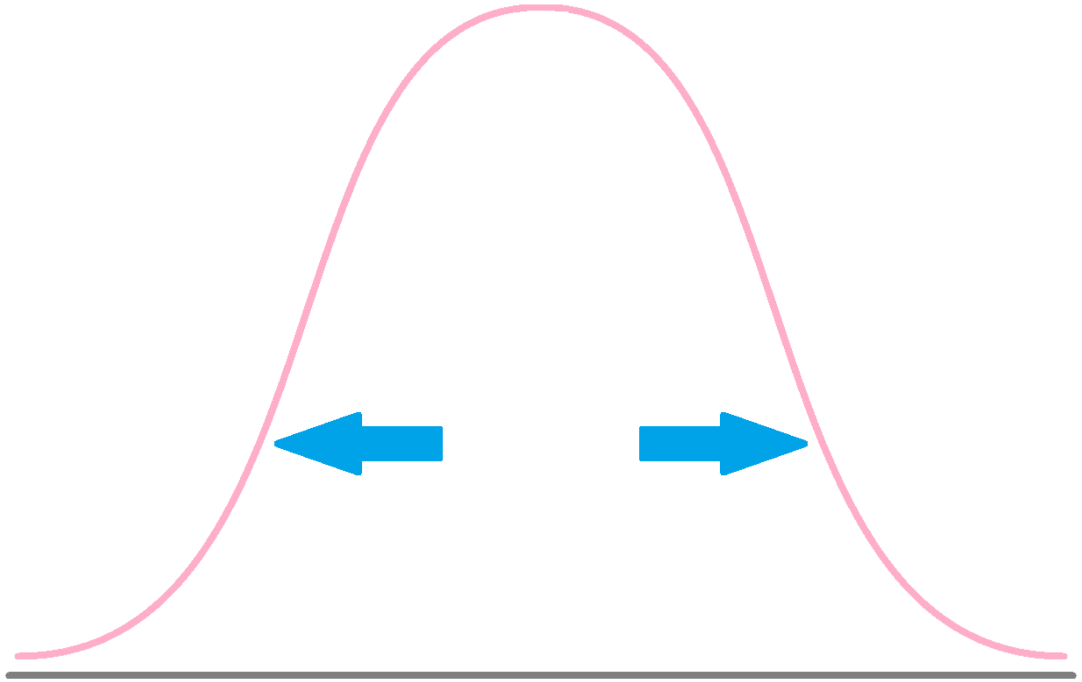
The median is simply the 50th percentile, $Q(y, 50)$

- Other important percentiles are the 1st and 3rd quartiles, i.e., the 25th and 75th percentiles

Percentiles



Measures of Spread (Dispersion)



https://medium.com/@angeloacebedo/descriptive-statistics-2-measures-of-spread-d87cd0284073

Measures of Spread

The most common measure of spread used is the simple **standard deviation.**

$$s(\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

- The simple standard deviation is the arithmetic mean of the squared deviations from the sample mean.
- Like the mean, is sensitive to changes in the sample.

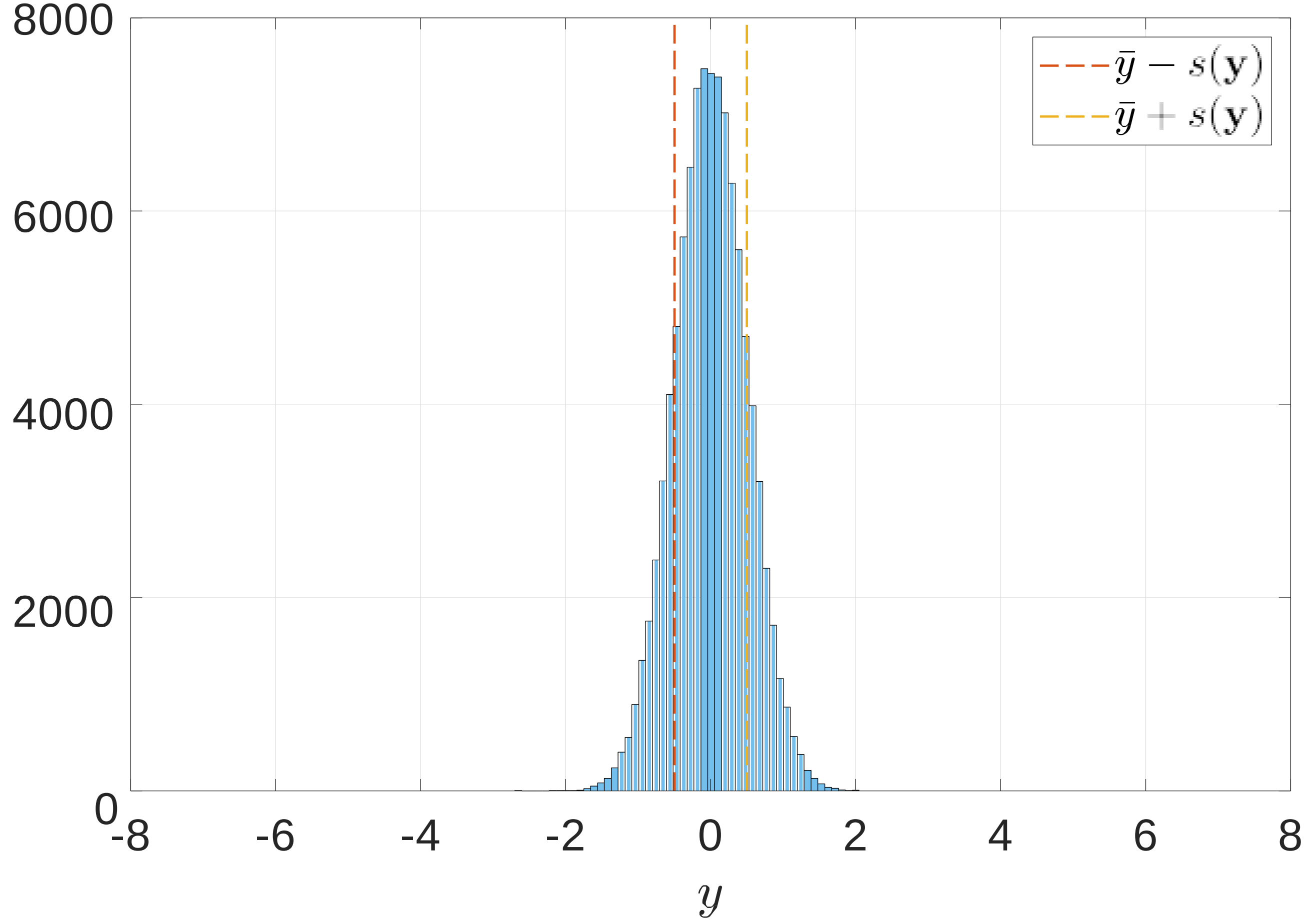
Often, the sample **variance**

$$v(\boldsymbol{y}) = s^2(\boldsymbol{y})$$

is used, as it can be easier to work with

Measures of Spread

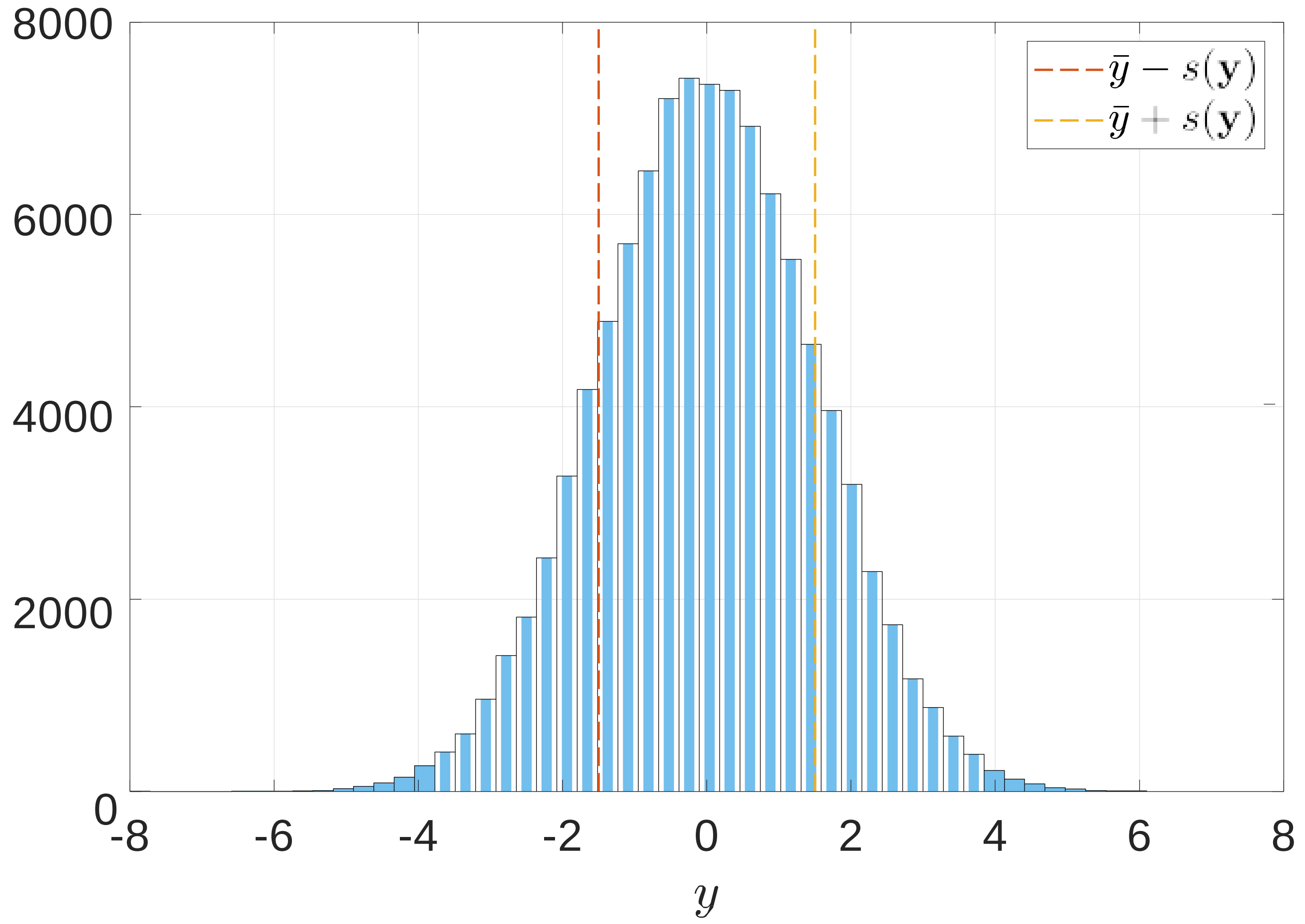
Example I



$$\text{range}(\mathbf{y}) = 4.63$$
$$(\min\{\mathbf{y}\} = -2.61,\ \max\{\mathbf{y}\} = 2.01)$$
$$s(\mathbf{y}) = 0.5$$

Measures of Spread

Examples II

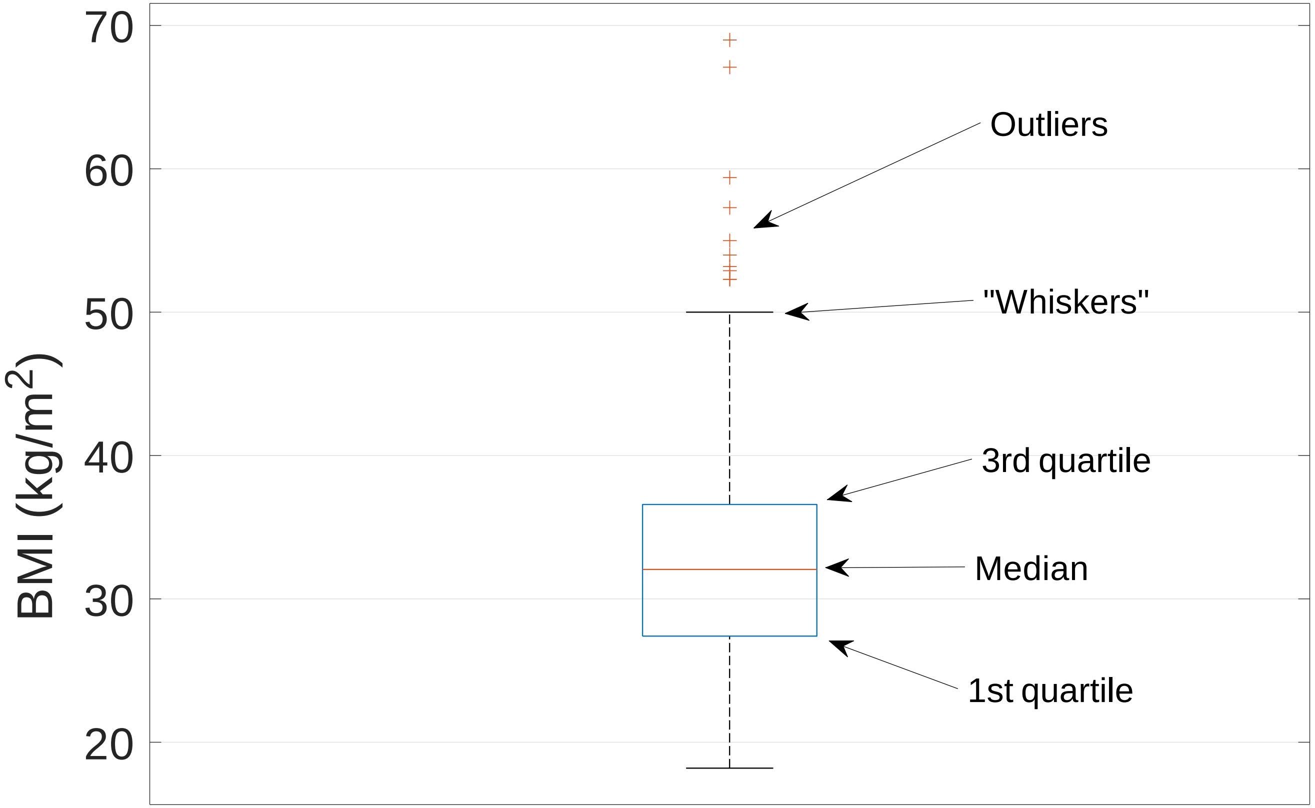


$$\mathrm{range}(\mathbf{y}) = 13.89$$
$$(\min\{\mathbf{y}\} = -7.84, \max\{\mathbf{y}\} = 6.05)$$
$$s(\mathbf{y}) = 1.5$$

MONASH University

Visualising Continuous Data

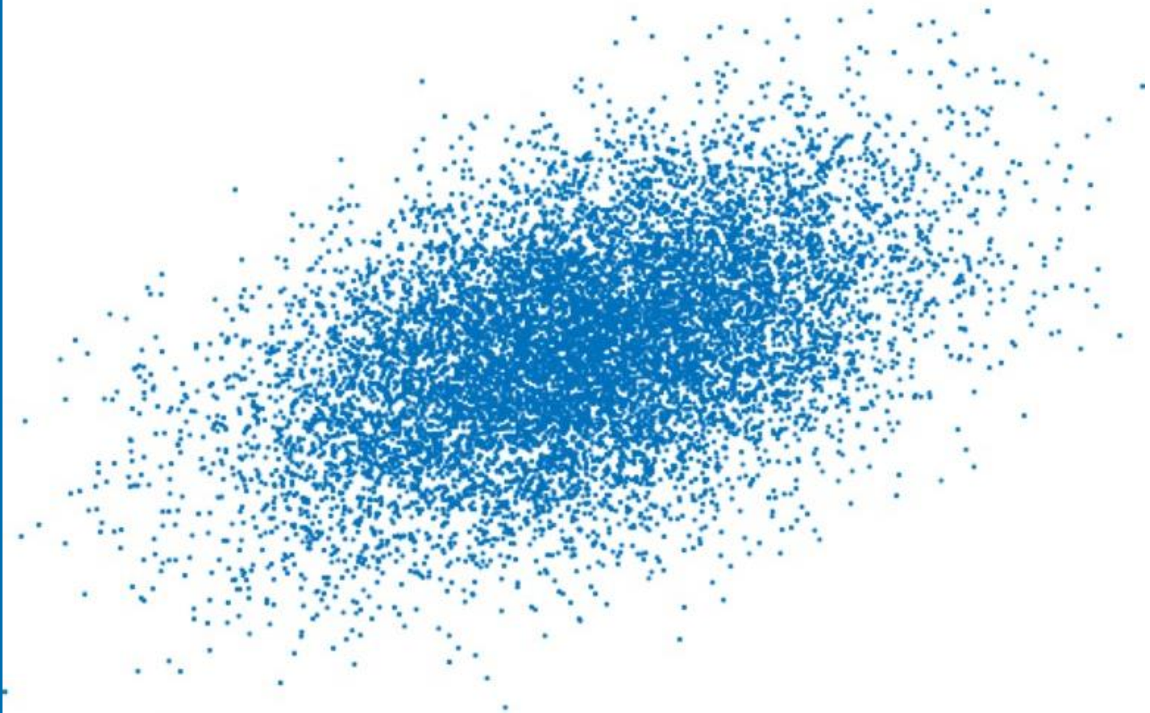
BoxPlots



Boxplot graphically captures **centrality**, **spread** and **skewness** in one plot

Association Between Two Continuous Variables

(Numeric-Continuous)



MONASH University

Association Between Two Continuous Variables

Let $x = (x_1, . . . , x_n)$ and $y = (y_1, . . . , y_n)$ be two numeric variables measured on the same objects

- We might ask if there is an association between x and y

Pearson correlation measures **linear association**

$$R(\mathbf{x}, \mathbf{y}) = \frac{\sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})}{n\, s(\mathbf{x}) s(\mathbf{y})}$$

- Correlation is always between -1 (completely negatively correlated) and 1 (completely positively correlated)

- A correlation of zero implies there is no linear association

  ⇒ note: does not imply no non-linear association

Remember: Correlation does not equal Causation!

MONASH University

Scatter Plots

Scatter plots help us visualise relationships between two (usually) numeric variables

- Plot points, with one variable on x-axis and one on y-axis

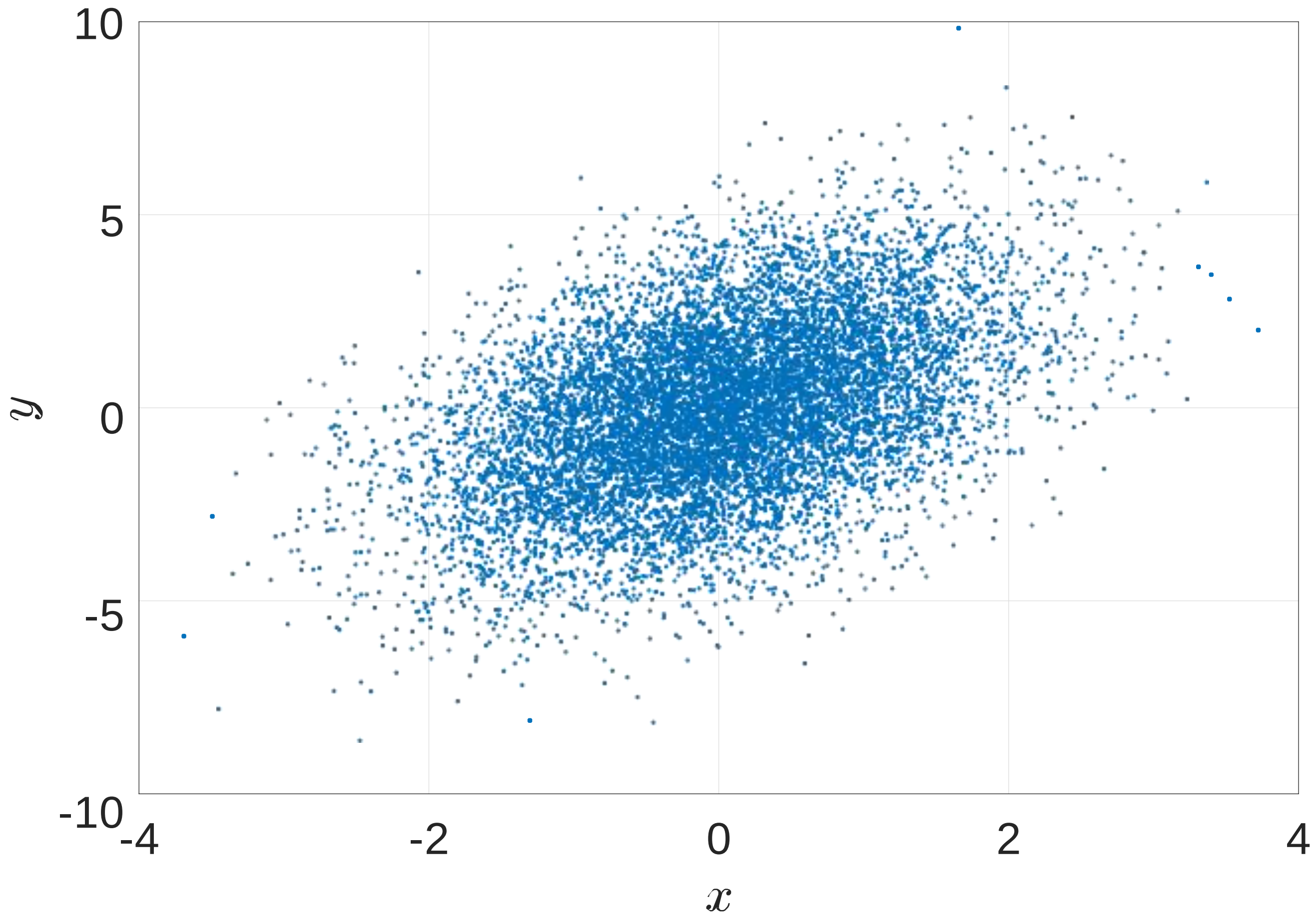
Can be used to visually look for association

Correlation coefficients are statistics that quantitatively measure the strength of the association between two variables

- The two can be combined for more information

Correlation using Scatter Plots

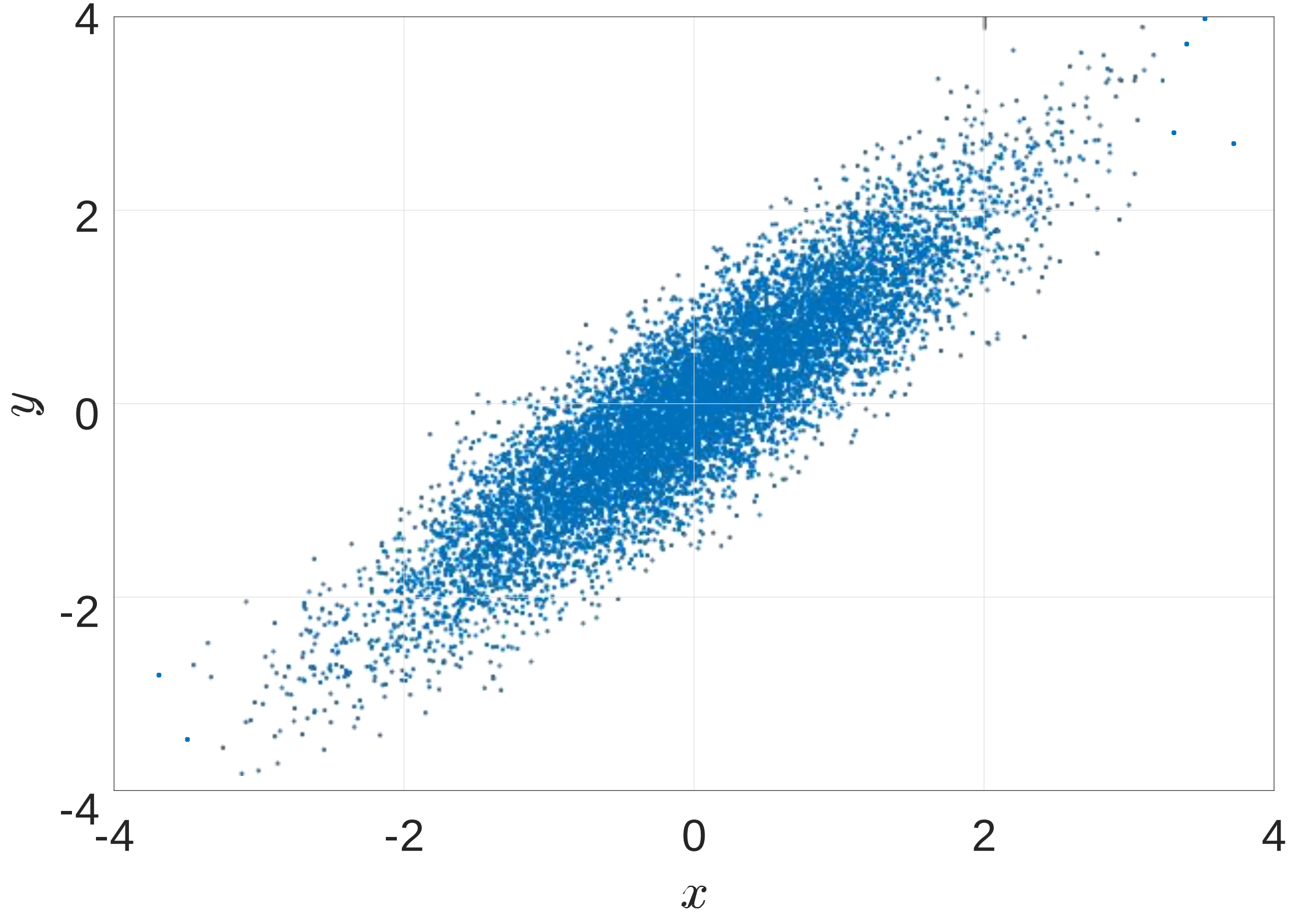
Example I



$R \approx 0.44$

Correlation using Scatter Plots

Example II

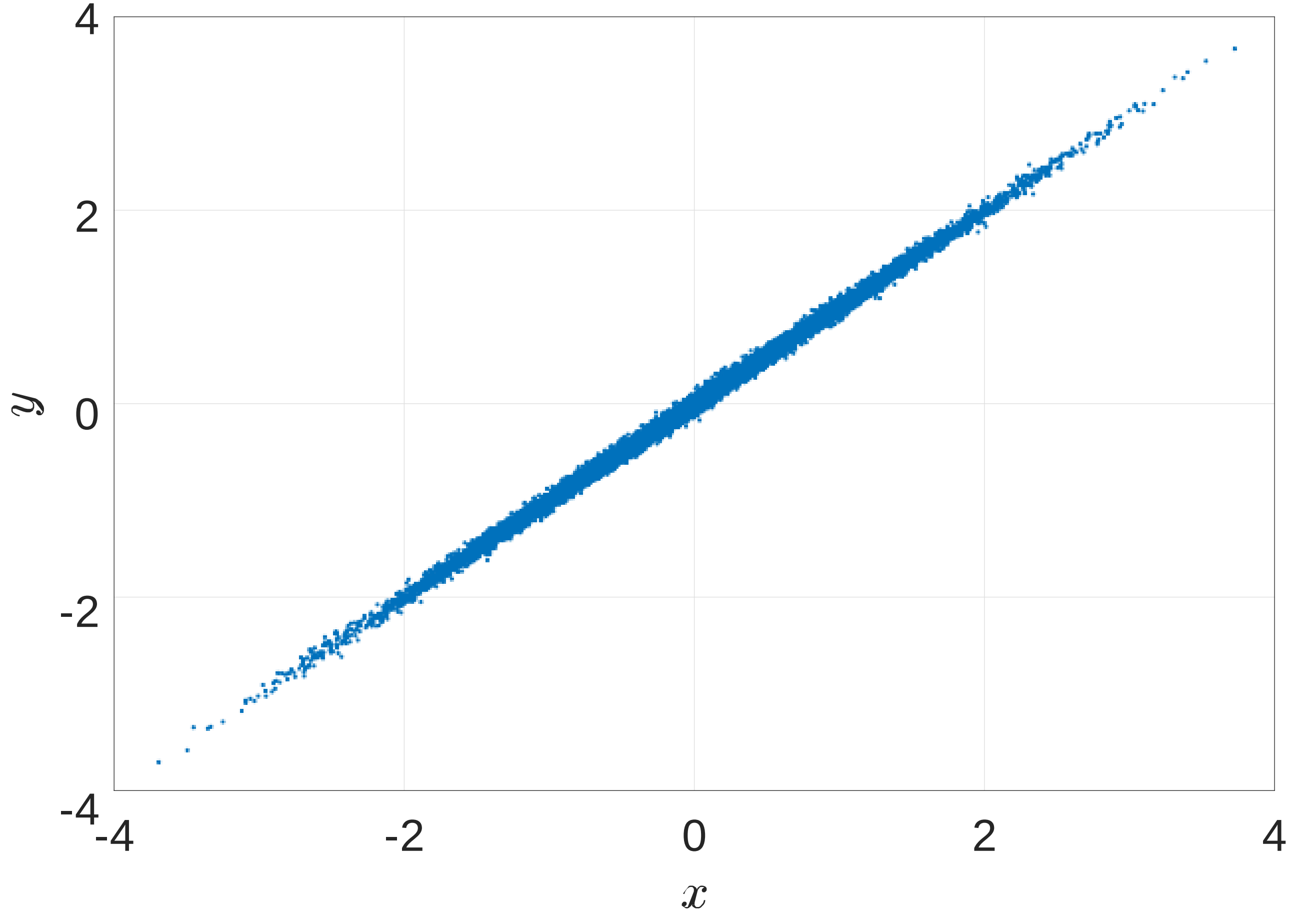


$R = 0.9$

MONASH University

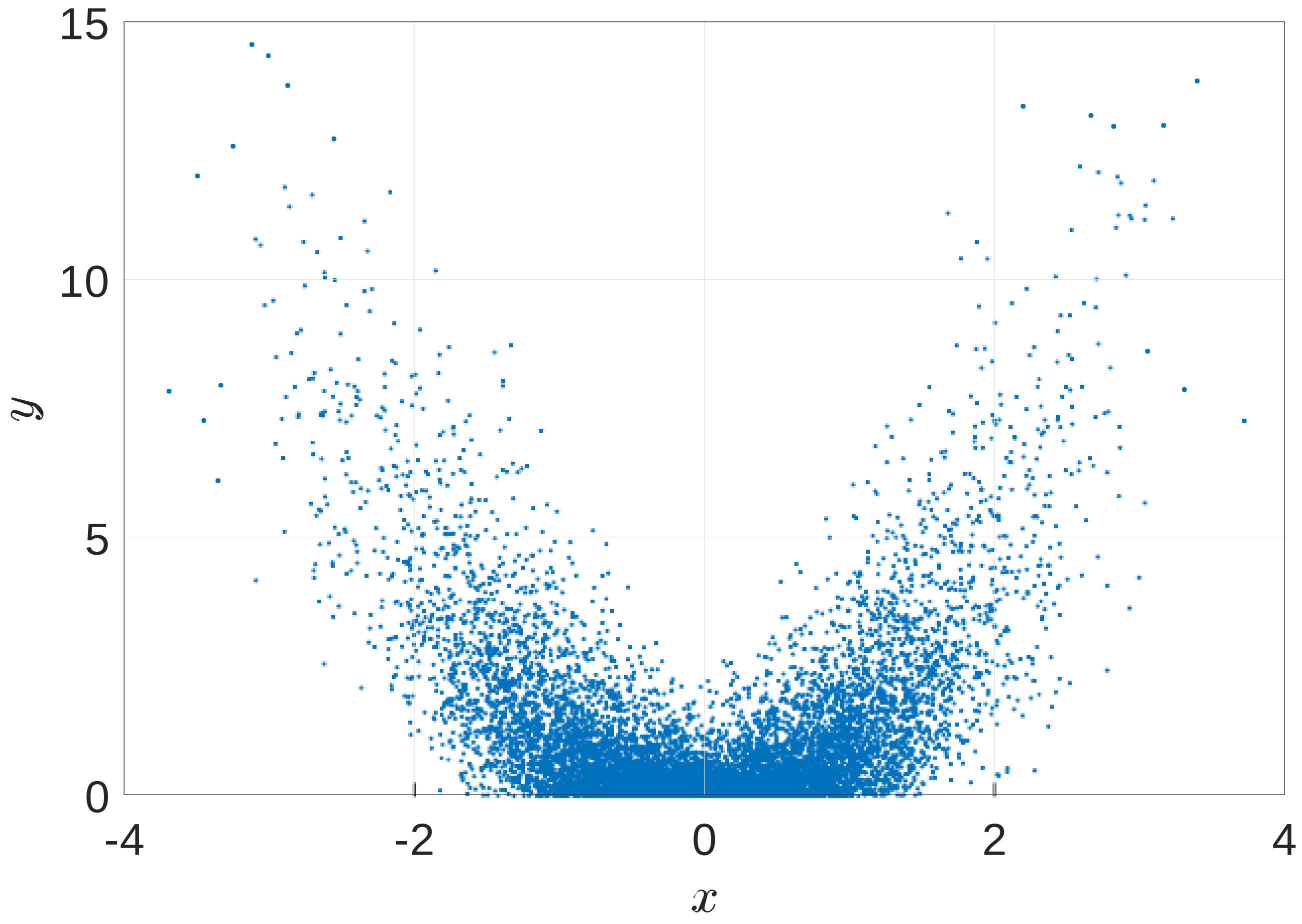
Correlation using Scatter Plots

Example III



$R \approx 0.999$

Correlation using Scatter Plots

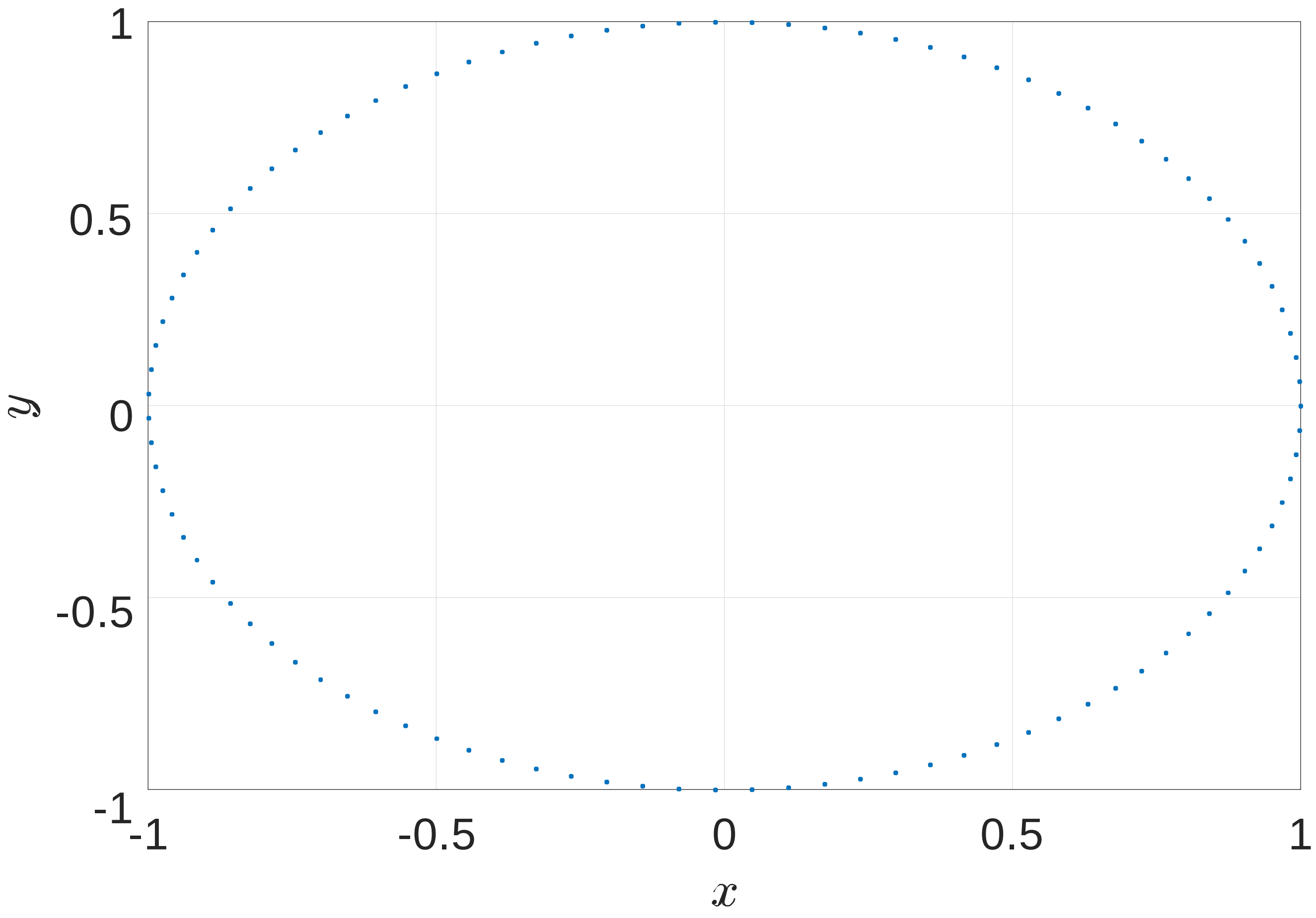


$R \approx 0.01$

Although clearly associated, as $y = x^2 + \text{noise}$

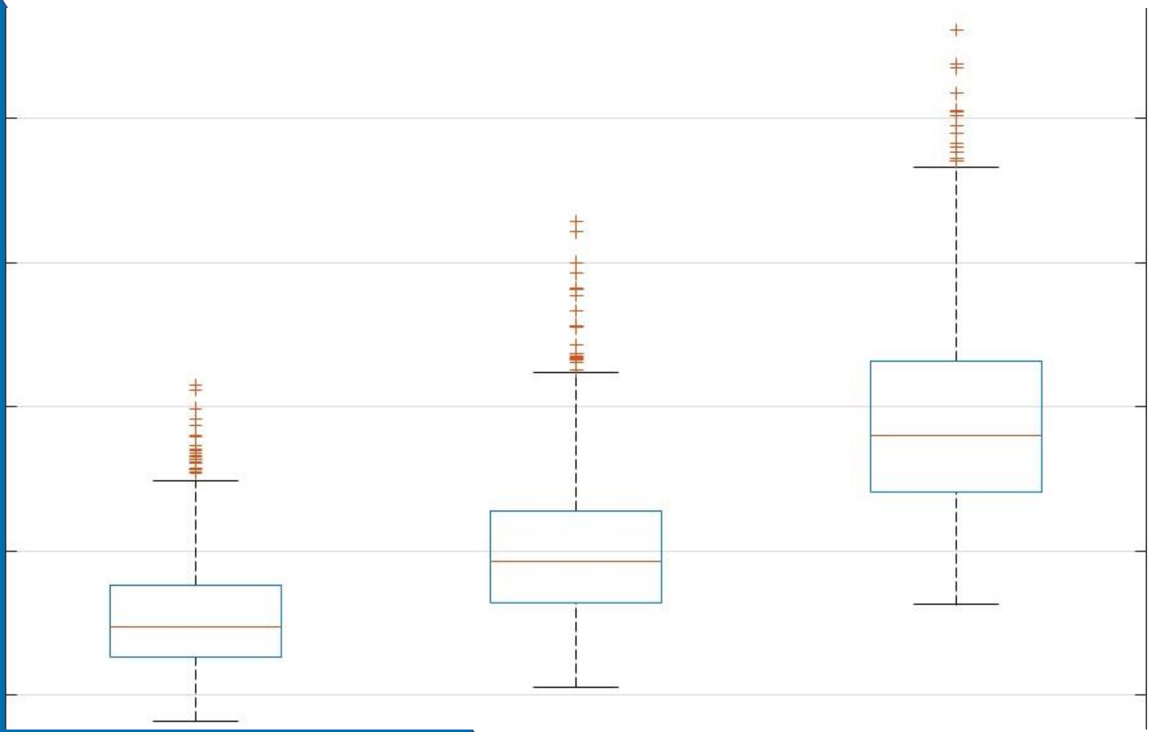
Correlation using Scatter Plots

Extra Thoughts



$R = 0$, though there is a deterministic association between x and y

Association Between Two Categorical Variables

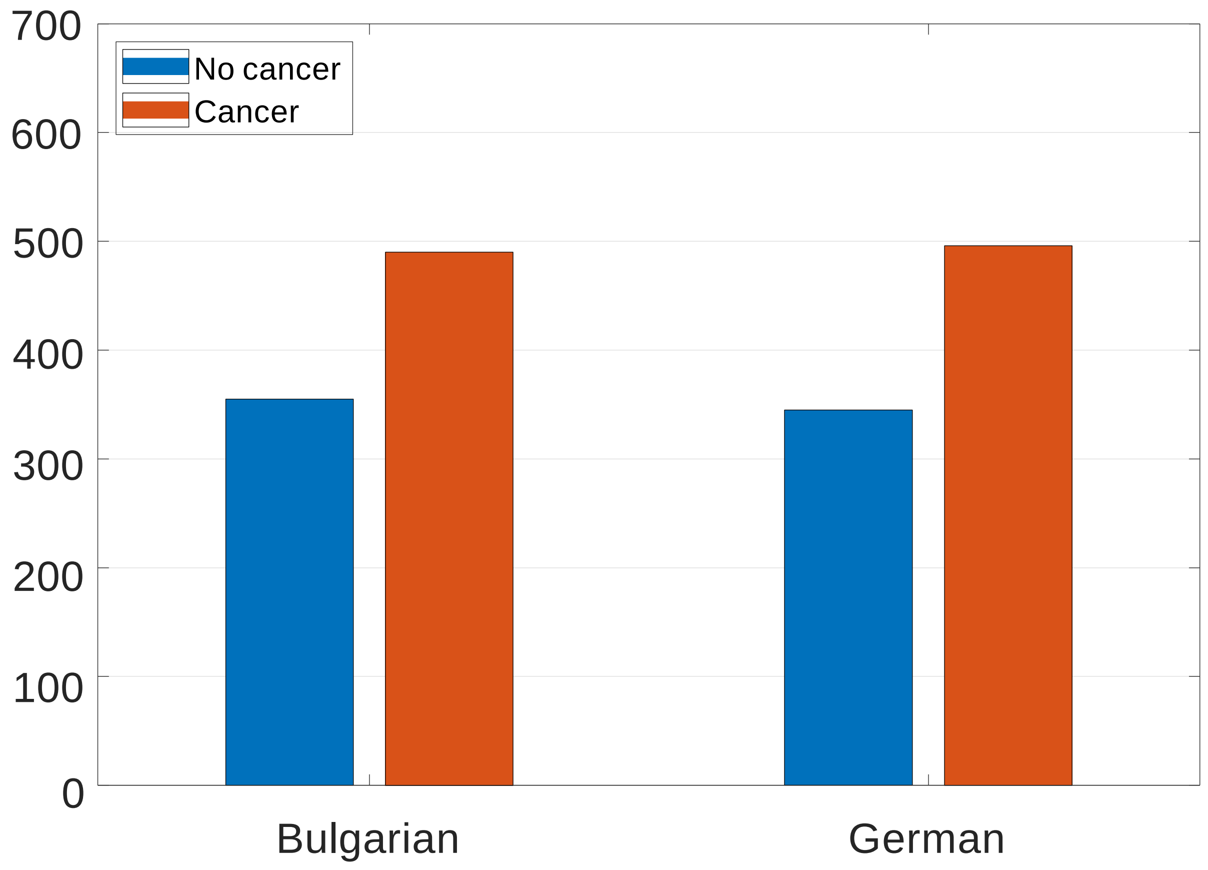


Association Between Two Categorical Variables

If **x** and **y** are both categorical, we can use a side-by-side bar graphs

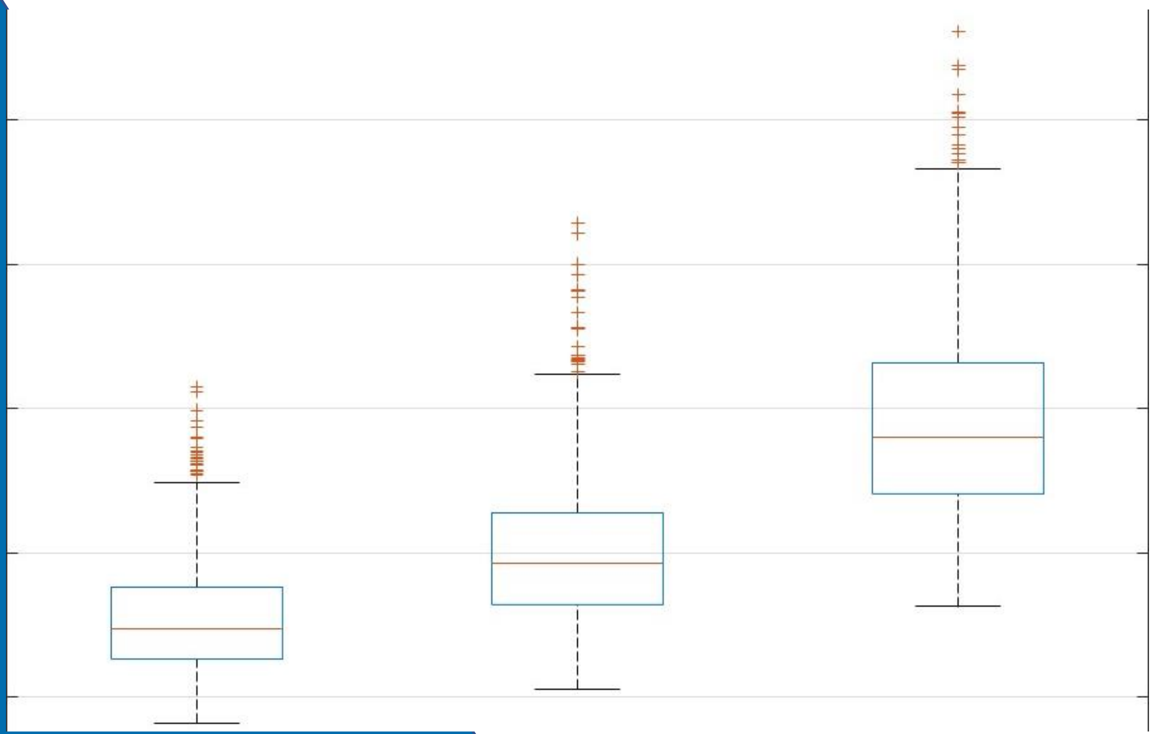
- Are the distributions/bar graphs different between categories?
- If so, there is a possible association.

Association Between Two Categorical Variables



Frequency of cancer does not seem to change with ethnicity; unlikely to be associated

Association Between Categorical and Numeric Variables



MONASH University

Association Between Categorical and Numeric Variables

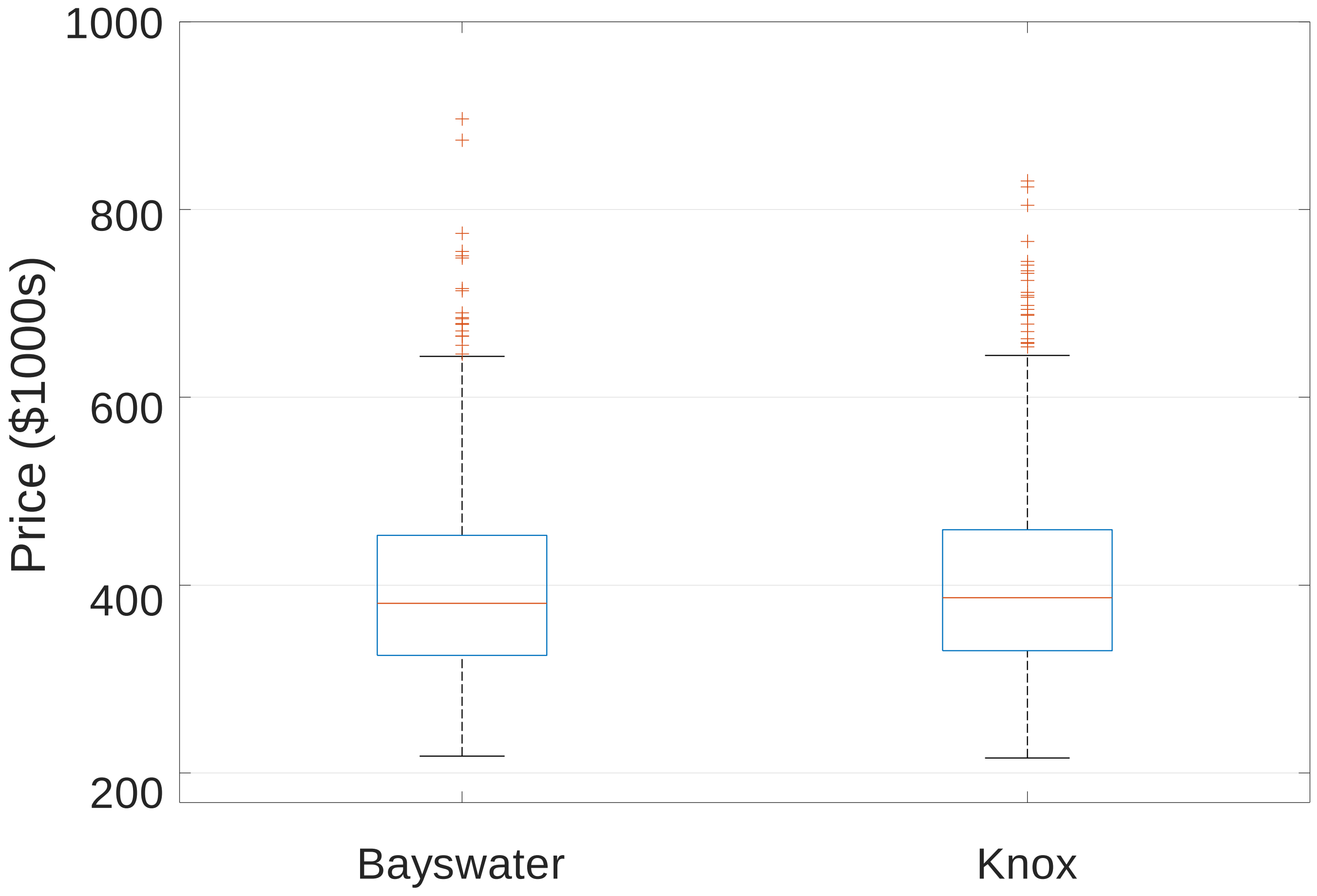
If **x** is categorical, and **y** is numeric, how to visualise?

A standard approach is the side-by-side boxplot

- Divide the data between categories, then plot boxplots for each group
- Do the boxplots look different?

Association Between Categorical and Numeric Variables

Example: Categorical and Numeric Variables I

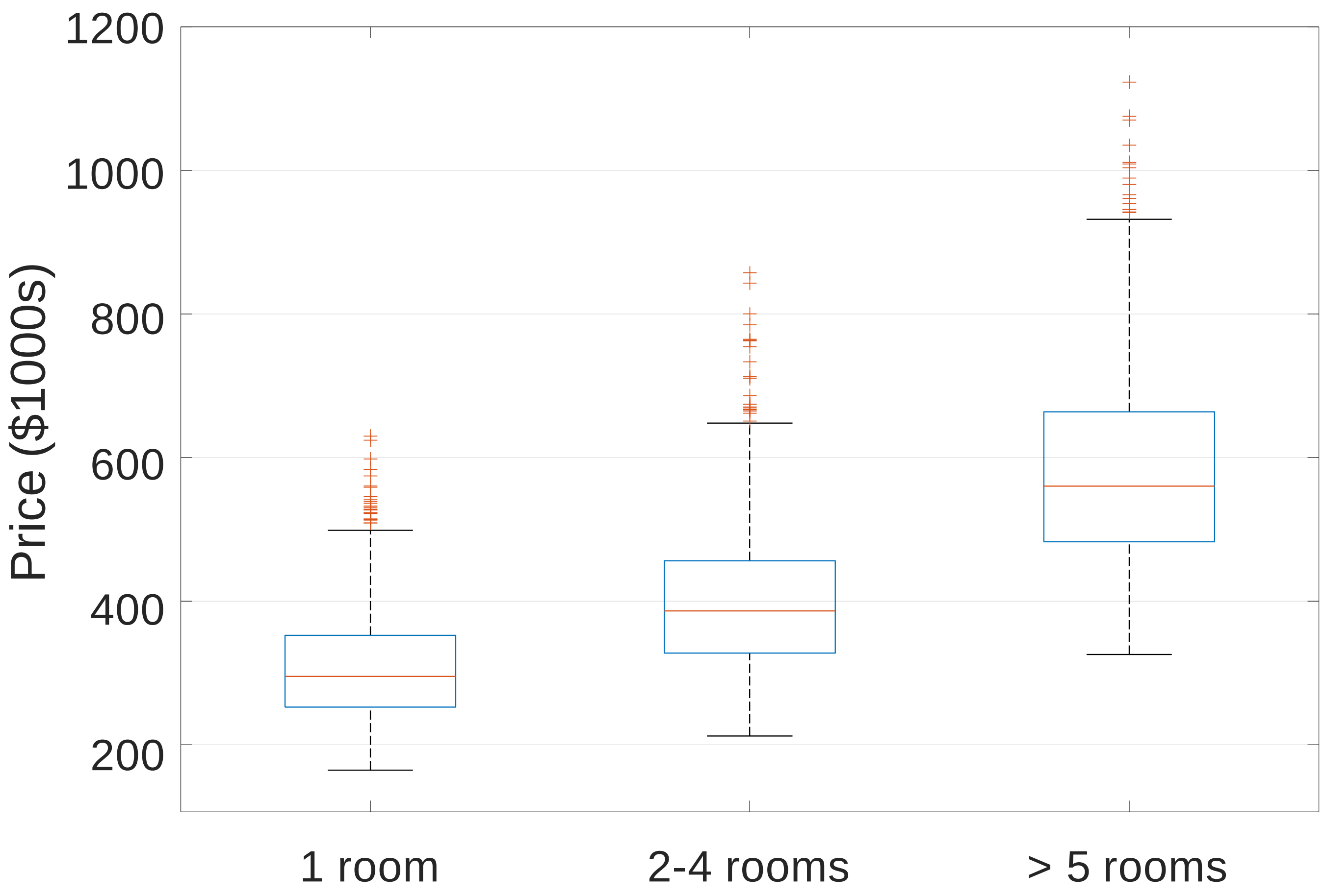


Distribution of price similar between suburbs

MONASH University

Association Between Categorical and Numeric Variables

Example: Categorical and Numeric Variables II



Distribution of price varies greatly with number of rooms

Learning Outcomes

Week 3

By the end of this week you should be able to:

- Comprehend more sophisticated **group-by operations** and *graphing* in Python
- Comprehend the power/**importance of data visualisation**
- Differentiate between **approaches for data visualisation**, and explain where each approach is appropriate to be used
- Explain/differentiate different concepts in **descriptive statistics**