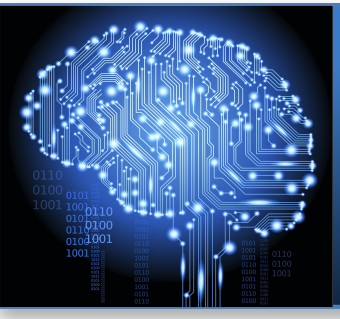


FIT1043 Lecture 3
Introduction to Data Science

Mahsa Salehi*

Faculty of Information Technology, Monash University

Semester 2, 2022

*with material from © Daniel Schmidt

Assessment

- Weekly quiz
  - Released every Monday (Weeks 2-11) at 6pm
  - You will have 48 hours to attempt the quiz
  - The quiz duration is 20 minutes
  - You will have two attempts and the highest mark will be considered
- Assignment 1
  - Python assessment (Next slide)
  - Will be released end of this week

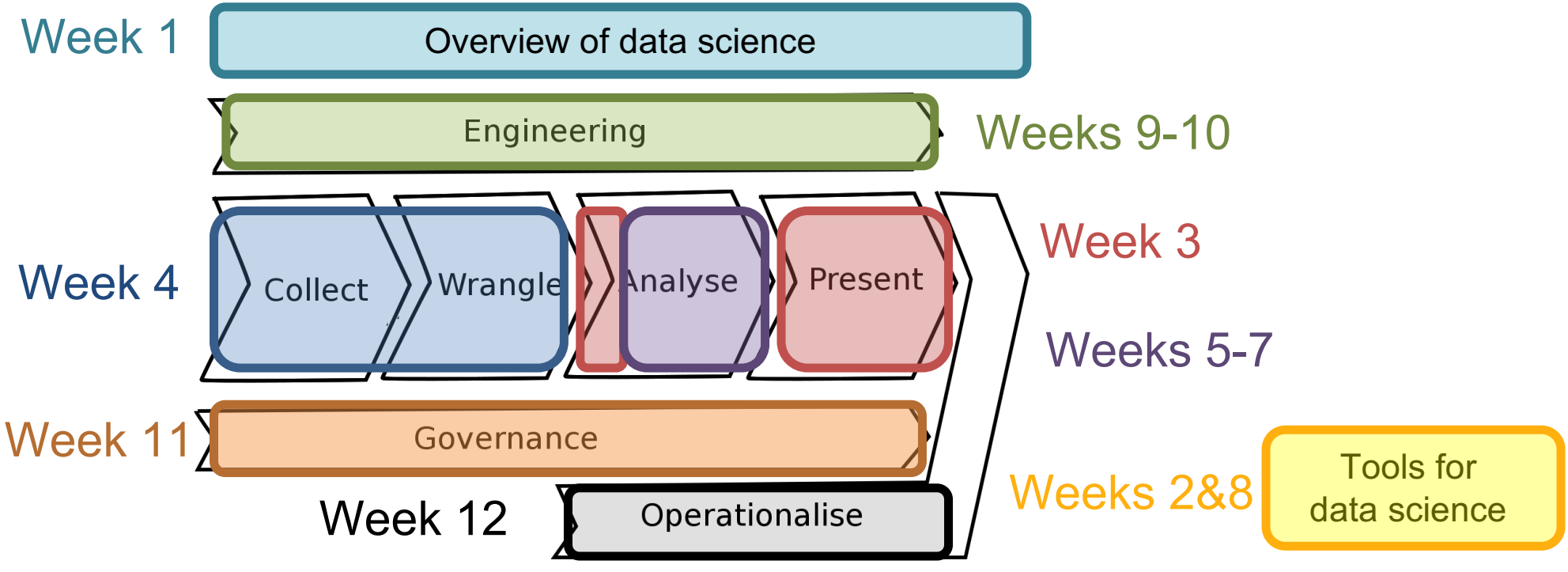
FLUX code:
5NKWED

Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly Lecture/tutorial active participation assessment
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture	Assignment 3

Unit Overview

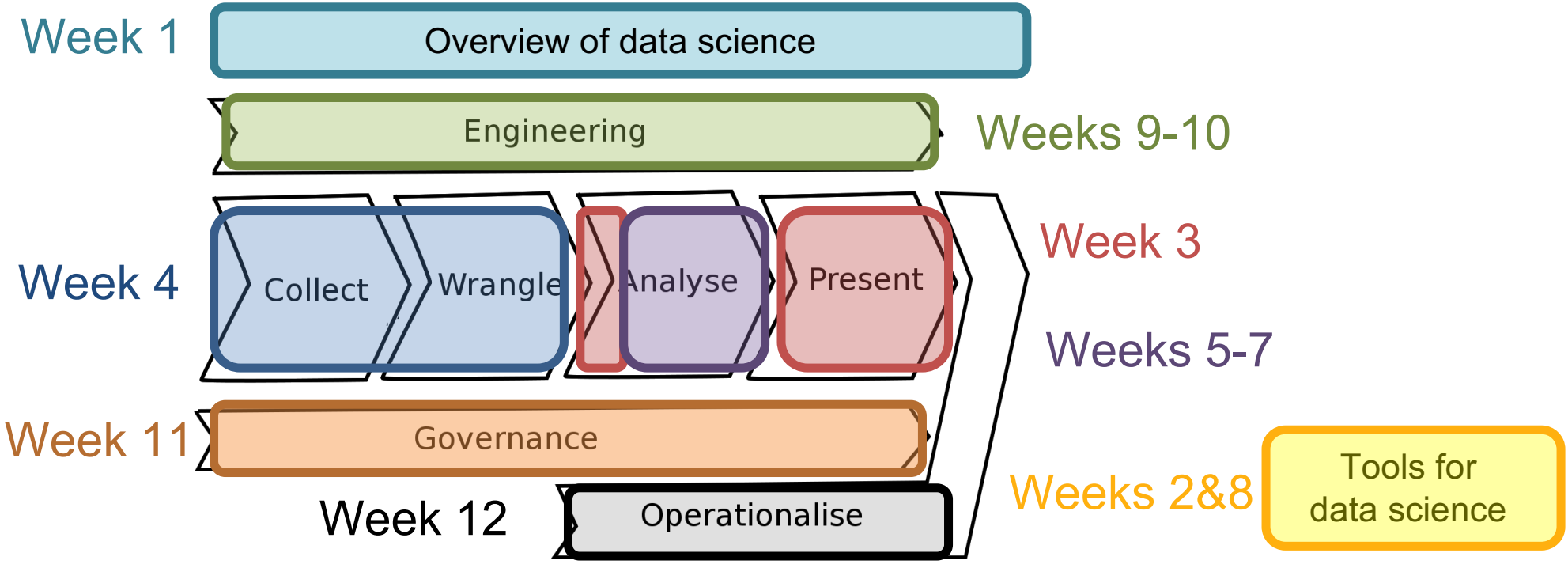
Collection: getting the data
Engineering: storage and computational resources
Governance: overall management of data
Wrangling: data preprocessing, cleaning
Analysis: discovery (learning, visualisation, *etc.*)
Presentation: arguing that results are significant and useful
Operationalisation: putting the results to work



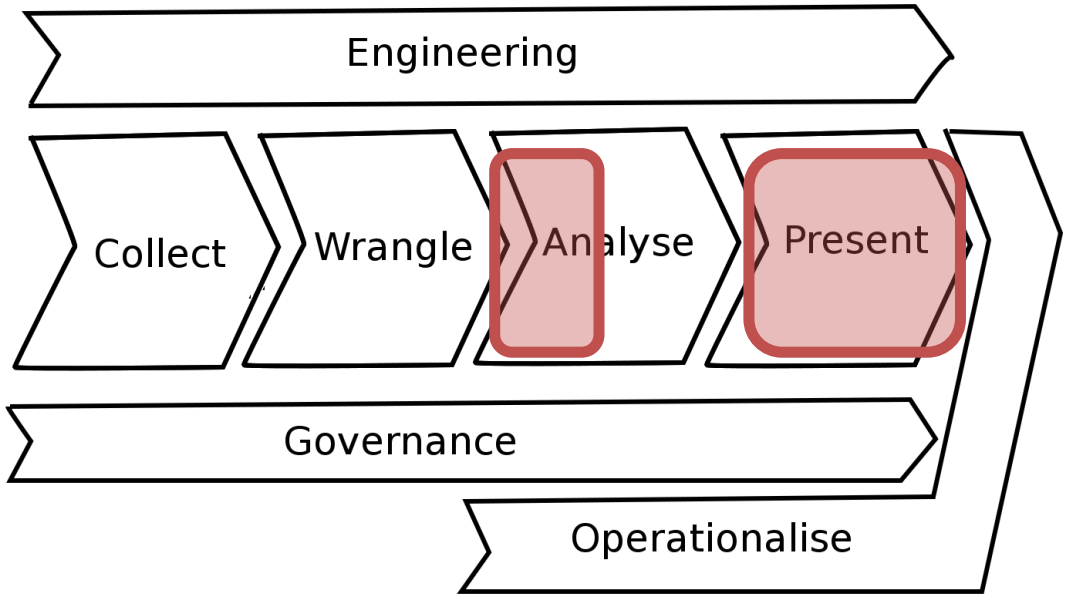
Assessments

Assessments:

- Assignment 1 (Weeks 2,3,4)
- Assignment 2 (Weeks 2-7)
- Assignment 3 (Weeks 8,9, 10)
- Weekly quiz (specific to each week)
- Final Exam (Weeks 1-12)



Our Standard Value Chain



This week!

Last week

Tools for
data science

Aggregation and groupby

Input

Gender	Age
male	22
female	38
female	26
female	35
male	35

Split

Gender	Age
female	38
female	26
female	35

Gender	Age
male	22
male	35

Apply (mean)

Gender	Age
female	33

Gender	Age
male	28.5

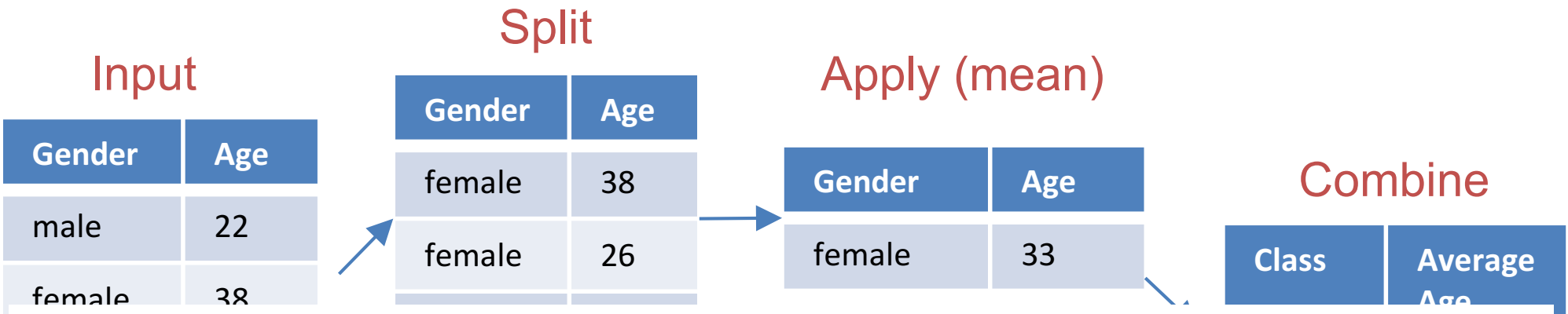
Combine

Class	Average Age
female	33
male	28.5

FLUX Question:
Write the Python code.

FLUX code:
5NKWED

Aggregation and groupby



I'm an LGBTIQA+ Ally.

Find out more at monash.edu/lgbtiqa

Advanced Aggregation (1)

Run multiple aggregation operators at once:

```
>>> fun = {'who':'count','age':'mean'}
>>> groupbyClass = titanic.groupby('class').agg(fun)
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

	who	age
class		
First	216	38.233441
Second	184	29.877630
Third	491	25.140620

Advanced Aggregation (2)

Write custom aggregators using anonymous functions:

```
>>> fun = {'age':{'nunique',lambda x: sum(e>50 for e in x)}}
>>> groupbyClass = titanic.groupby('class').agg(fun)
```

	age	
	nunique	<lambda_0>
class		
First	57	39
Second	57	15
Third	68	10

Outline

- Data visualisation
  - Why?
  - Basic data types
  - Different graphical representations
- Descriptive statistics
- Python for data visualization

Learning Outcomes (Week 3)

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

- Comprehend the importance/power 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
- Comprehend more sophisticated group-by operations and graphing in Python

Data Visualisation

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

data visualisation is useful as a preliminary form of data analysis to get a "feel" for the data:

Motivation: Data Visualisation

extract from *"Turning powerful stats into art"* by Chris Jordan, starting at minute 1:00

Basic Types of Data

- Categorical-Nominal:
  - Discrete numbers of values, no inherent ordering
  - E.g., country of birth, sex
- Categorical-Ordinal:
  - Discrete number of states, but with an ordering
  - E.g., Education status, State of disease progression
- Numeric-Discrete:
  - Numeric, but the values are enumerable
  - E.g., Number of live births, Age (in whole years)
- Numeric-Continuous:
  - Numeric, not enumerable (i.e., real numbers)
  - E.g., Weight, Height, Distance from CBD

Data Visualisation

- It is often useful to visualise data
  - Can sometimes quickly reveal patterns
  - However, going beyond two dimensions is problematic

Data Visualisation

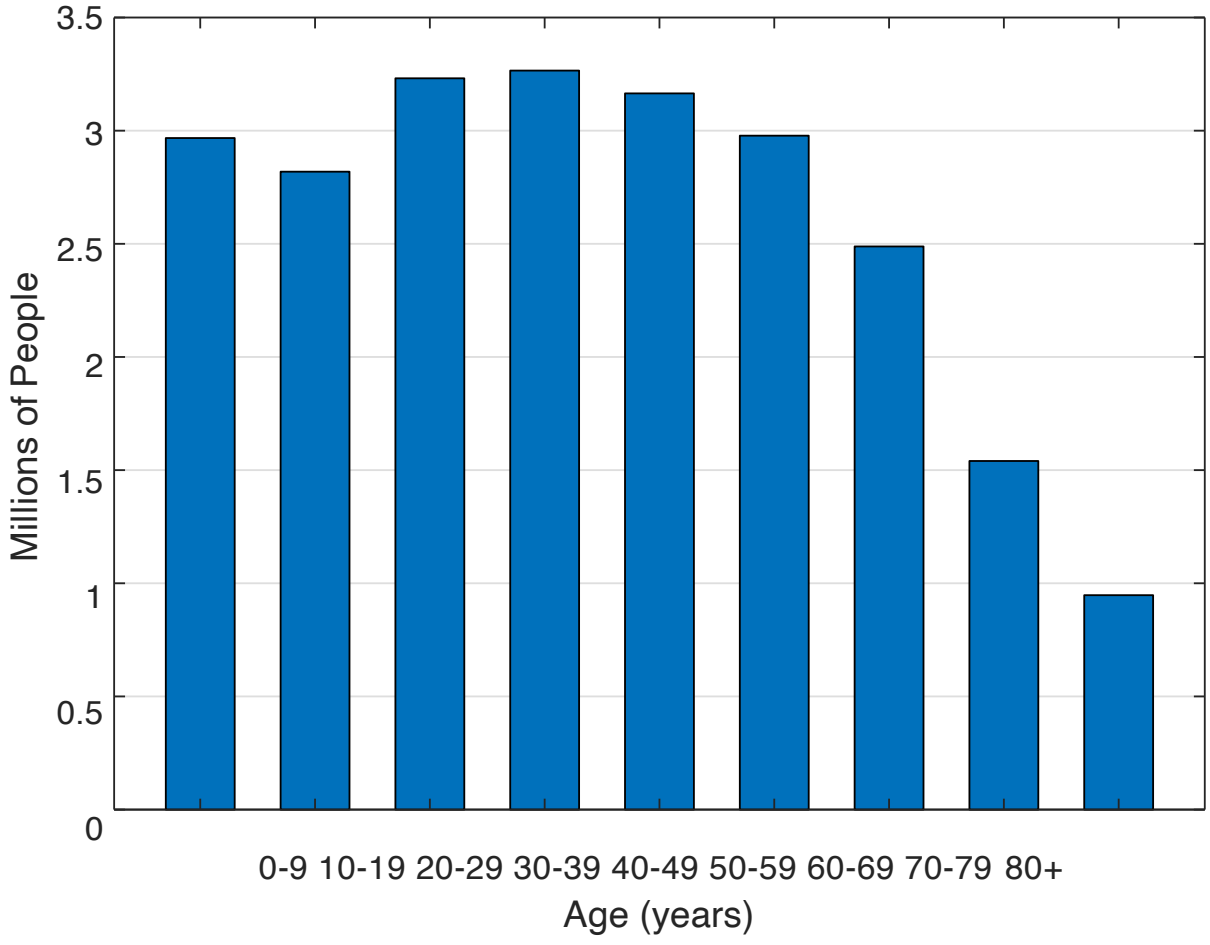
- It is often useful to visualise data
  - Can sometimes quickly reveal patterns
  - However, going beyond two dimensions is problematic
- For categorical data, standard visualisations include:
  - Bar graphs
  - Pie charts
- For numeric data (continuous and discrete), we can use:
  - Histograms
  - Box plots

Frequency Tables

Age (years)	Number of People
0-9	2,967,425
10-19	2,818,778
20-29	3,231,395
30-39	3,265,526
40-49	3,164,712
50-59	2,977,883
60-69	2,488,396
70-79	1,540,373
80+	947,411

Australian Population by Age
(2016 Census)

Bar Charts



Australian Population by Age
(2016 Census)

Pie Chart

- Pie Chart is a type of graph in which a circle is divided into sectors that each represent a proportion of the whole

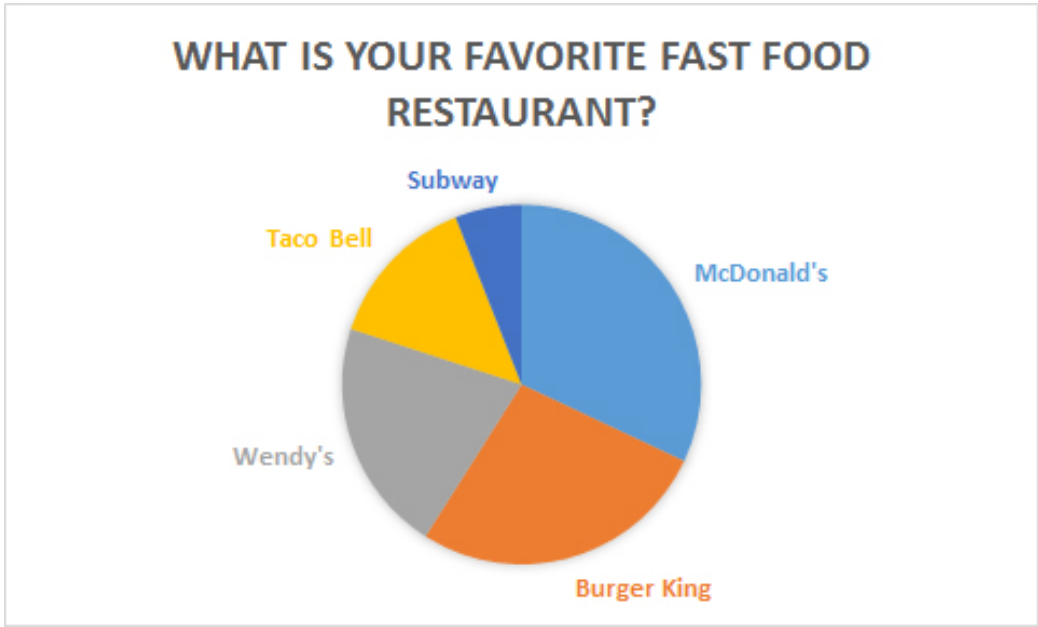


Image src: varsitytutors.com

Histograms

- Group numeric data into categories by putting into bins
- If $\mathbf{y} = (y_1, \ldots, y_n)$ are our data points, we divide them between K equally spaced bins, i.e.,
  - The number of samples that fall in bin (category) k are

$$v_k = \#\{y_j \in (\min\{\mathbf{y}\} + (k-1)w,\ \min\{\mathbf{y}\} + kw)\}$$

where

$$w = \frac{\max\{\mathbf{y}\} - \min\{\mathbf{y}\}}{K}$$

is the width if the bins

→ plot $v_1, \ldots, v_K$ using bar-chart

FLUX code: 5NKWED

FLUX Question:
Histograms are a special type of …?

Histograms

- Group numeric data into categories by putting into bins
- If $\mathbf{y} = (y_1, \ldots, y_n)$ are our data points, we divide them between K equally spaced bins, i.e.,
  - The number of samples that fall in bin (category) k are

$$v_k = \#\{y_j \in (\min\{\mathbf{y}\} + (k-1)w,\ \min\{\mathbf{y}\} + kw)\}$$

where

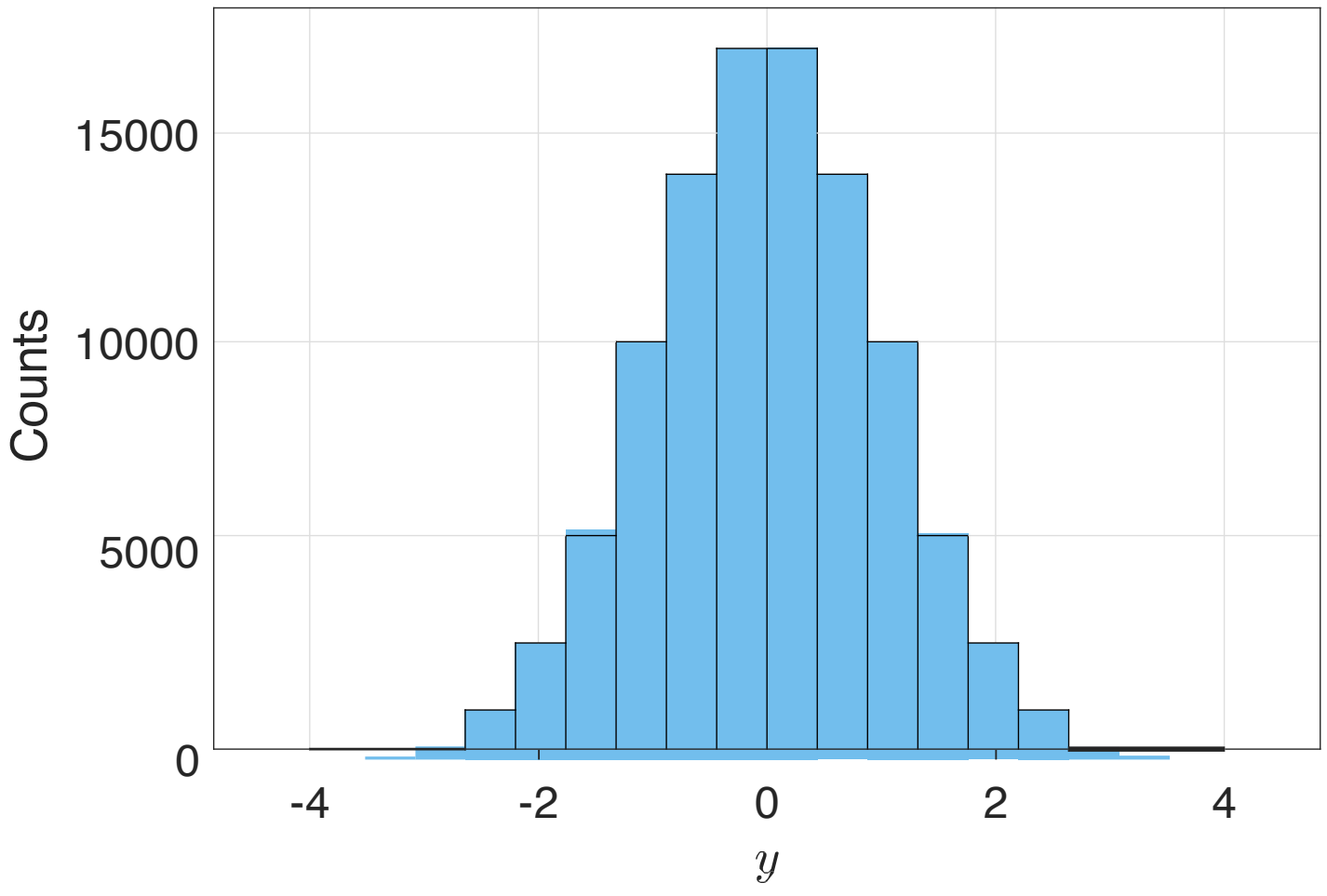
$$w = \frac{\max\{\mathbf{y}\} - \min\{\mathbf{y}\}}{K}$$

is the width if the bins

→ plot $v_1, \ldots, v_K$ using bar-chart

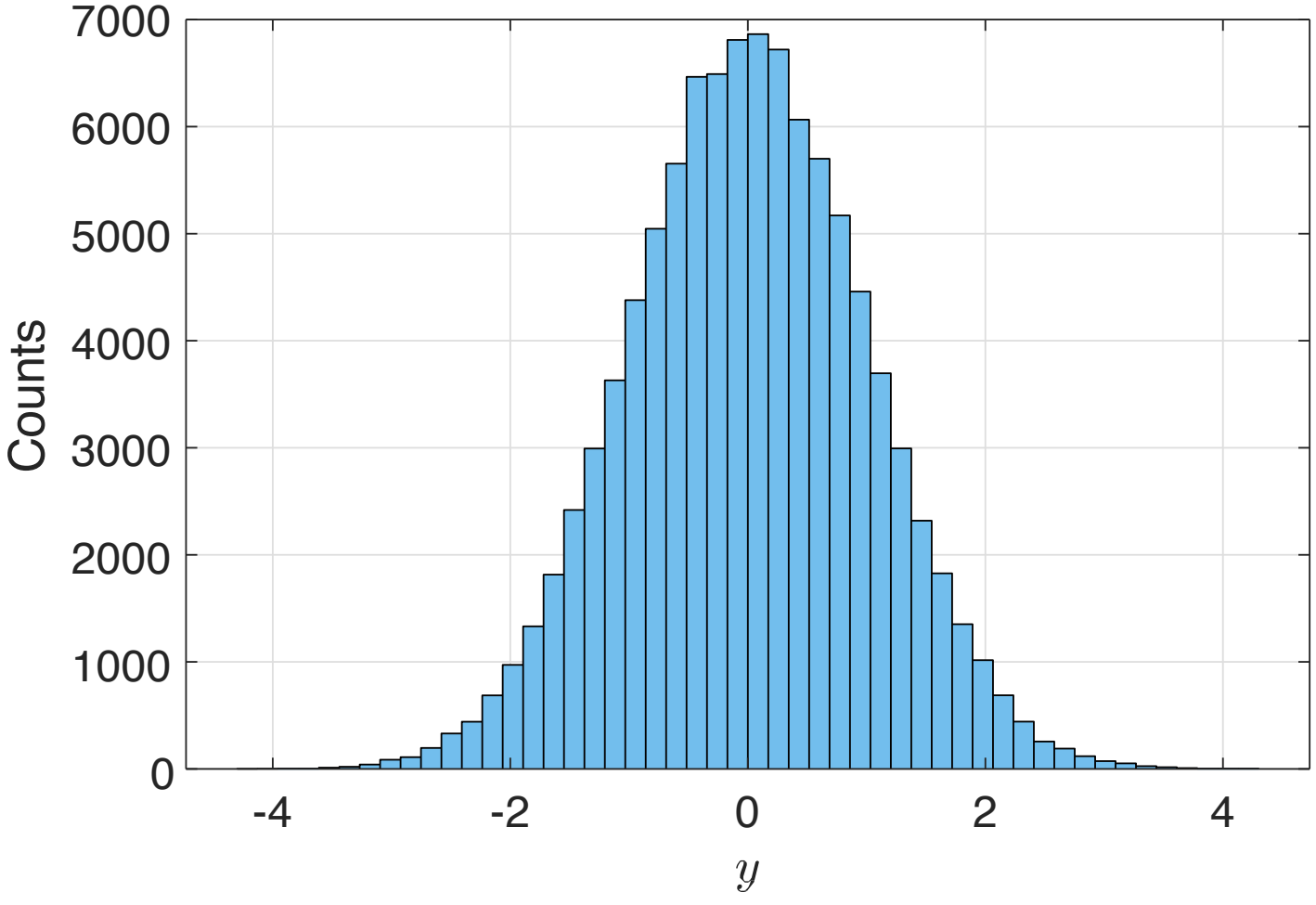
- Histograms are a special type of bar chart
  - Bar-charts only applicable to categorical data

Histograms: Example



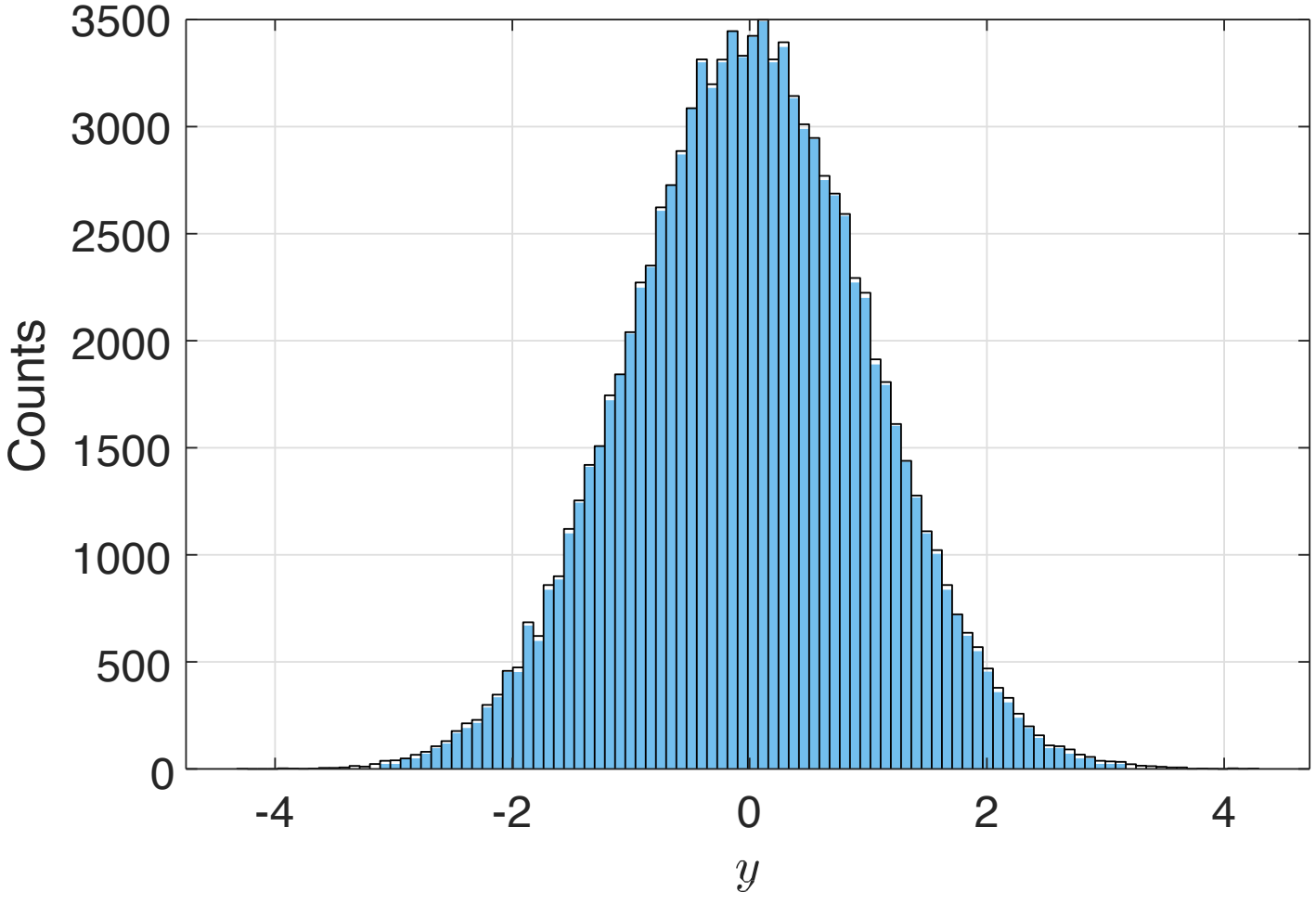
Histogram with $K = 20$ bins

Histograms: Example



Histogram with $K = 50$ bins; looks smoother

Histograms: Example



Histogram with $K = 100$; starting to look ragged

Why Motion Chart

Motivation

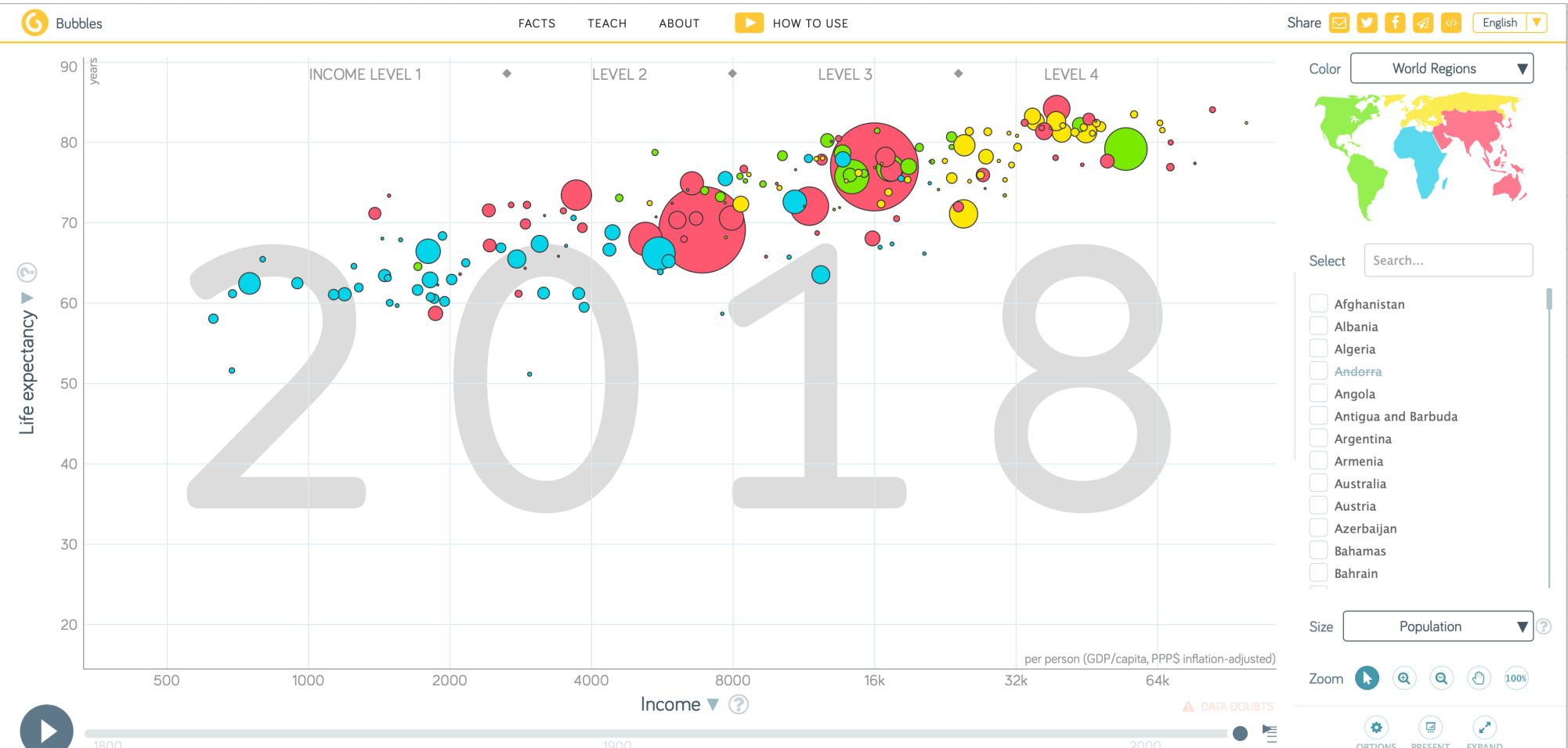
- Motion Charts are interactive multi-dimensional data visualisations
- Originally introduced to the world as GapMinder by Hans Rosling and made famous by his TED talks.

History

- The GapMinder technology was bought by Google and the name of motion charts changed to bubble charts
- But the GapMinder website is now up as a not-for-profit.

Motion Charts

Visualizing data in five dimensions: x-axis, y-axis, size of bubble, color of bubble, and time



Motion Charts

Advantages:

- time dimension allows deeper insights and observing trends
- good for exploratory work
- “appeal to the brain at a more instinctual intuitive level”

Disadvantages:

- not suited for static media
- display can be overwhelming, and controls are complex
- “data scientists who branch into visualization must be aware of the limitations of uses”

Descriptive Statistics

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

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

- Some functions (statistics) more useful than others
  - But all describe properties of the data

Measures of Centrality

- Let $\mathbf{y} = (y_1, \ldots, y_n)$ 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, $\mathrm{med}(\mathbf{y})$
  - Value such that 50% of samples have values less than $\mathrm{med}(\mathbf{y})$
  - Easily found by sorting samples and finding middle sample

FLUX Question

Which option is the Mean, Median and Mode of the following set of values respectively?
1,2,2,3,4,7,9

A. 4,2,3
B. 5,3,2
C. 4,3,3
D. 4,3,2

FLUX code:
5NKWED

FLUX Question

Compute Mean, Median and Mode of 1,2,2,3,4,7,9

Type	Example	Result
Mean	(1+2+2+3+4+7+9) / 7	**4**
Median	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	1, **2**, **2**, 3, 4, 7, 9	**2**

Mean vs Median

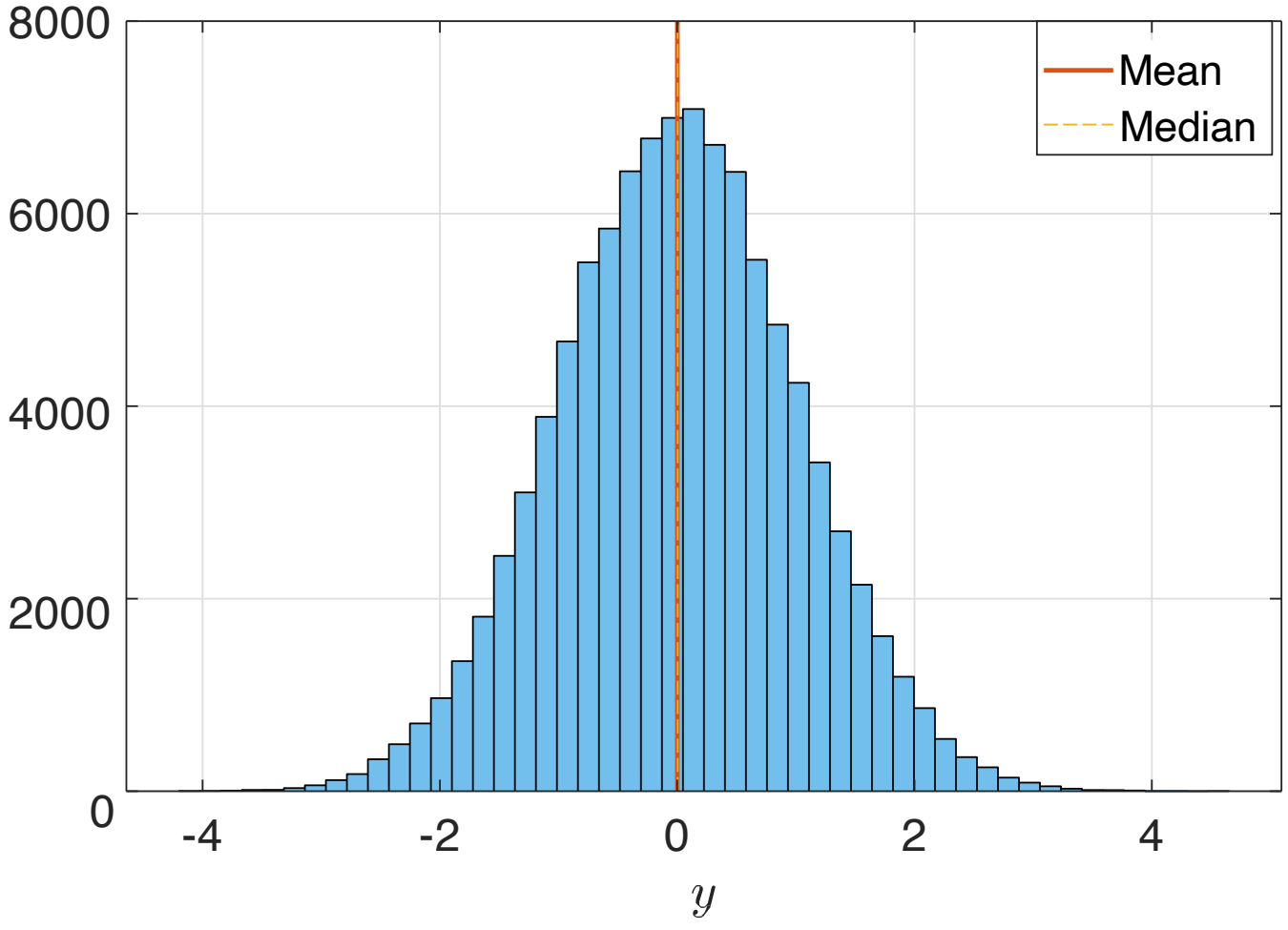
- The mean uses *all* the values of the sample
  - Any change to any sample changes the mean
  - The mean can be changed as much as desired by changing just one sample by a large enough amount
- The median uses at most two of the values of the sample

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

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

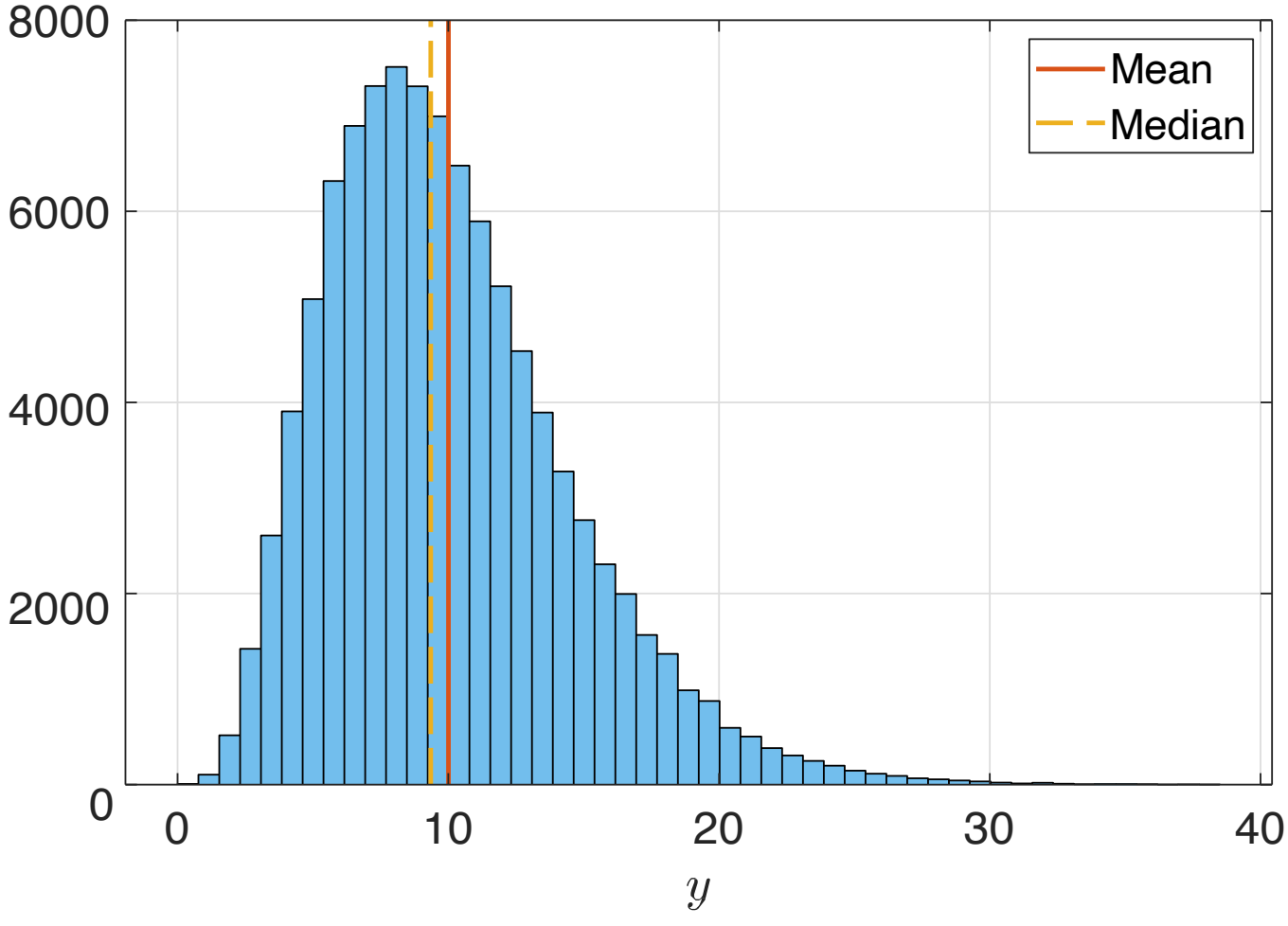
  $\mathbf{y} = (1, 2, 3, 4, 50) \Rightarrow \bar{y} = 12, \text{med}(\mathbf{y}) = 3$
- Why might we want to use mean over median then?

Mean vs Median: Symmetric Distributions



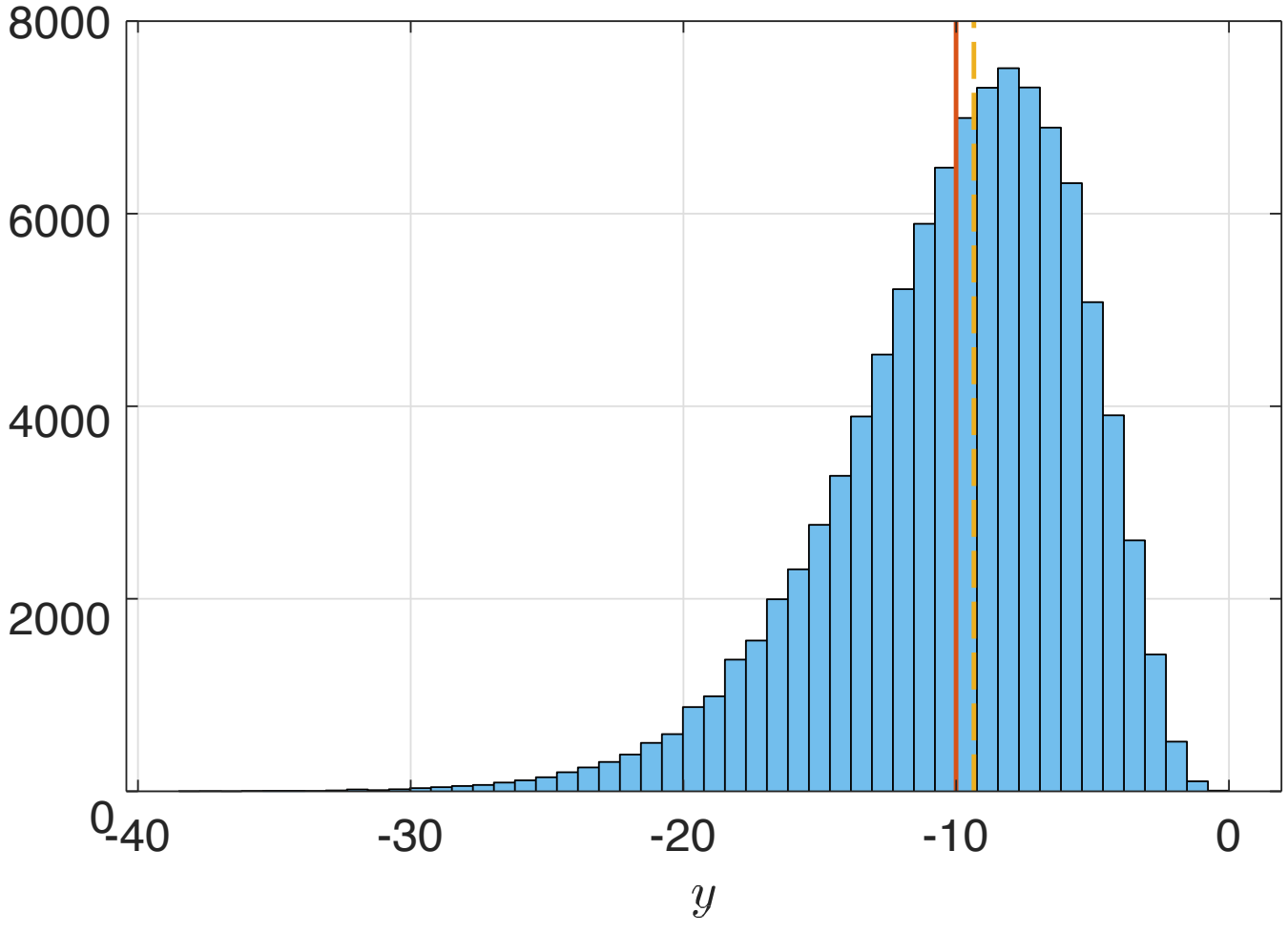
Symmetric distribution of data; mean and median (nearly) the same

Mean vs Median: Positively Skewed Data



Positively skewed data; mean greater than median

Mean vs Median: Negatively Skewed Data



Negatively skewed data; mean less than median

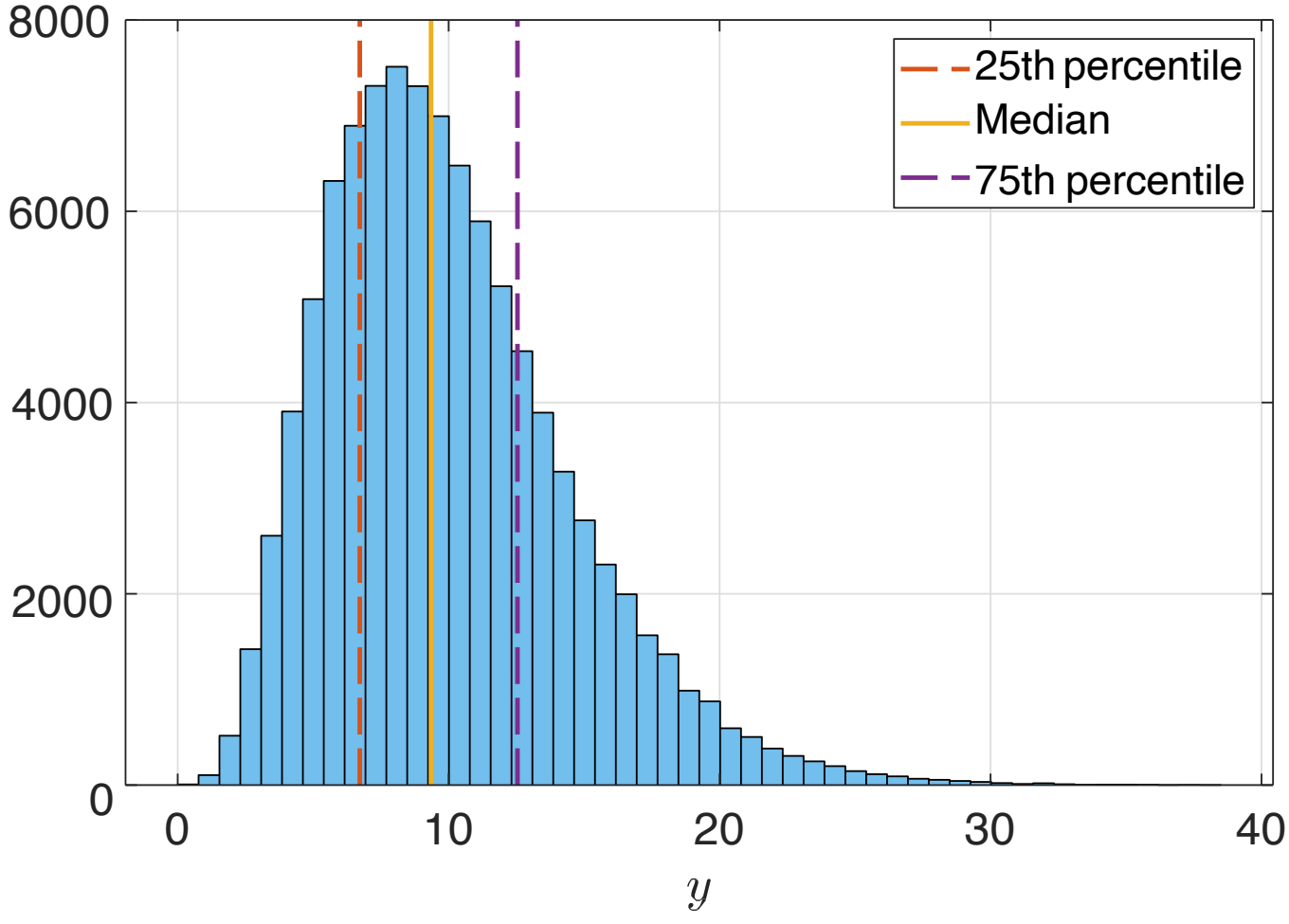
Percentiles

- More generally, we can define the percentiles
  - The p-th percentile is the value, $Q(\mathbf{y}, p)$ such that $p\%$ of the values of the sample are lower than $Q(\mathbf{y}, p)$

Percentiles

- More generally, we can define the percentiles
  - The p-th percentile is the value, $Q(\mathbf{y}, p)$ such that $p\%$ of the values of the sample are lower than $Q(\mathbf{y}, p)$
- The median is simply the 50th percentile, $Q(\mathbf{y}, 50)$
- Other important percentiles are the 1st and 3rd quartiles
  - i.e., the 25th and 75th percentiles

Percentiles



Measures of Spread (1)

- Measures of centrality tell us about the typical value of the sample
- Measures of spread tell us how much the samples differ, on average, from the typical value
- The most straightforward is the range

$$\mathrm{rng}(\mathbf{y}) = \max\{\mathbf{y}\} - \min\{\mathbf{y}\}$$

where

$\min\{\mathbf{y}\}$ denotes the minimum value in the sample;
$\max\{\mathbf{y}\}$ denotes the maximum value in the sample.

Measures of Spread (2)

- The most common measure of spread used is the sample standard deviation

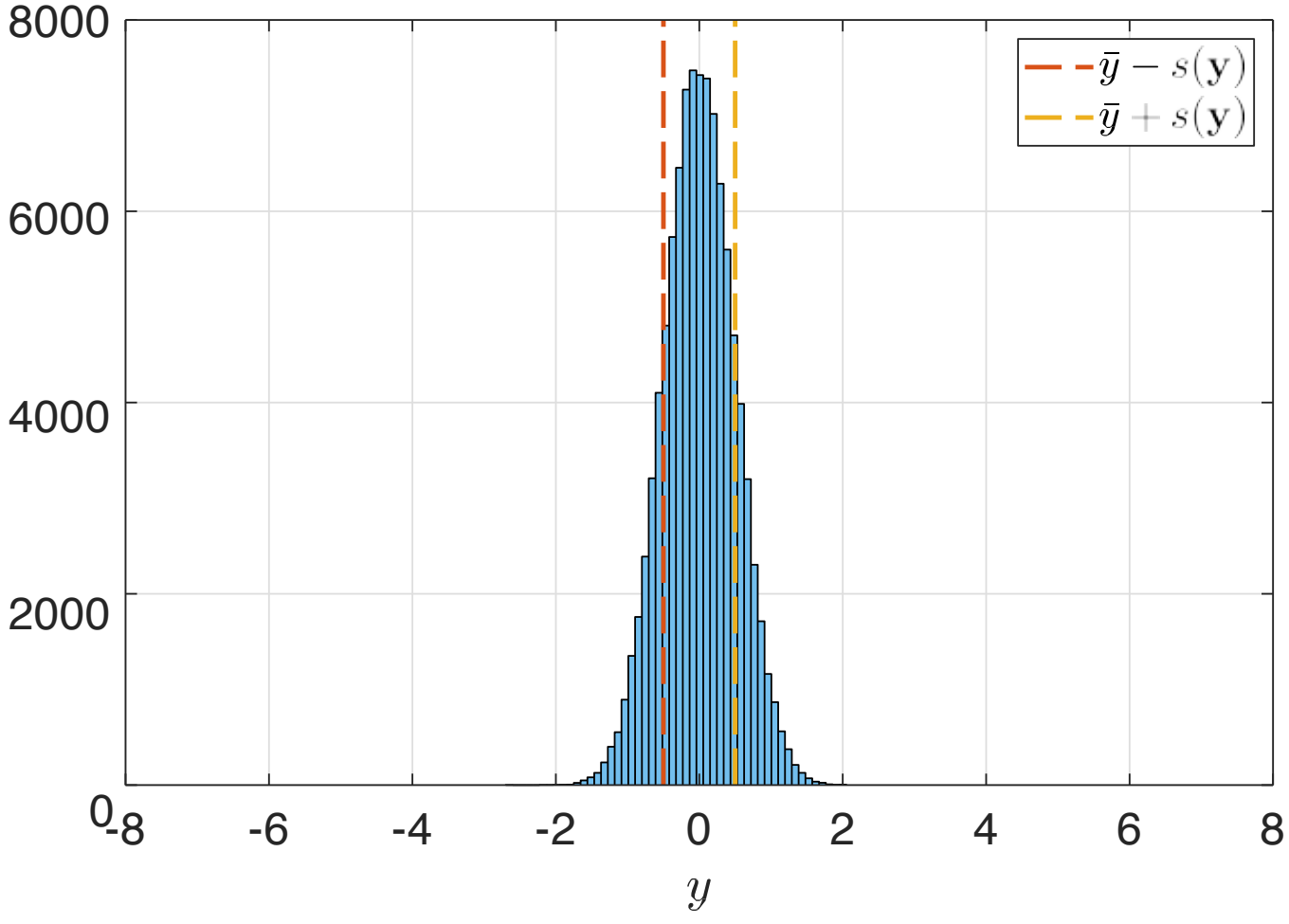
$$s(\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

- The sample 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

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

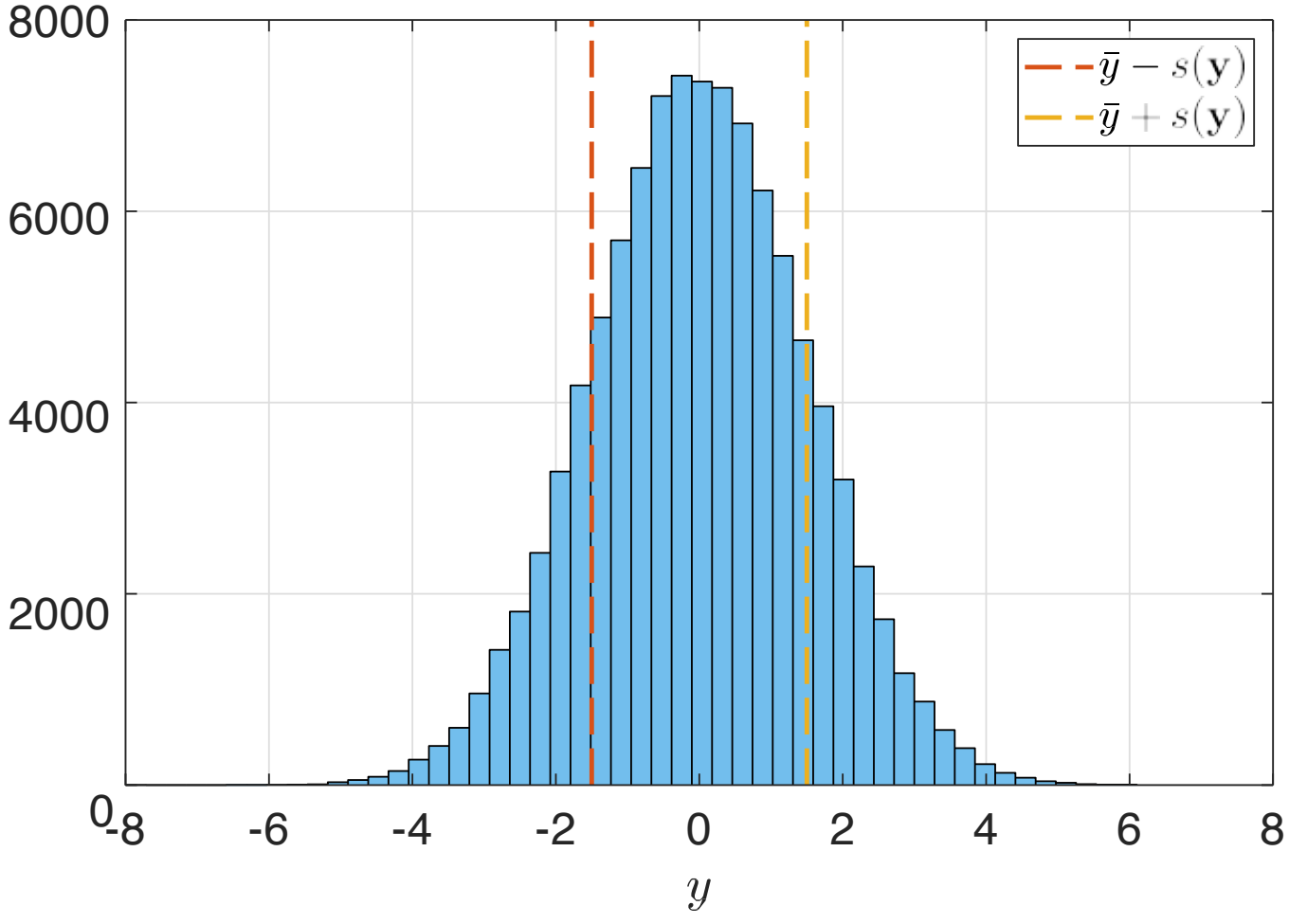
is used, as it can be easier to work with

Measures of Spread: Example



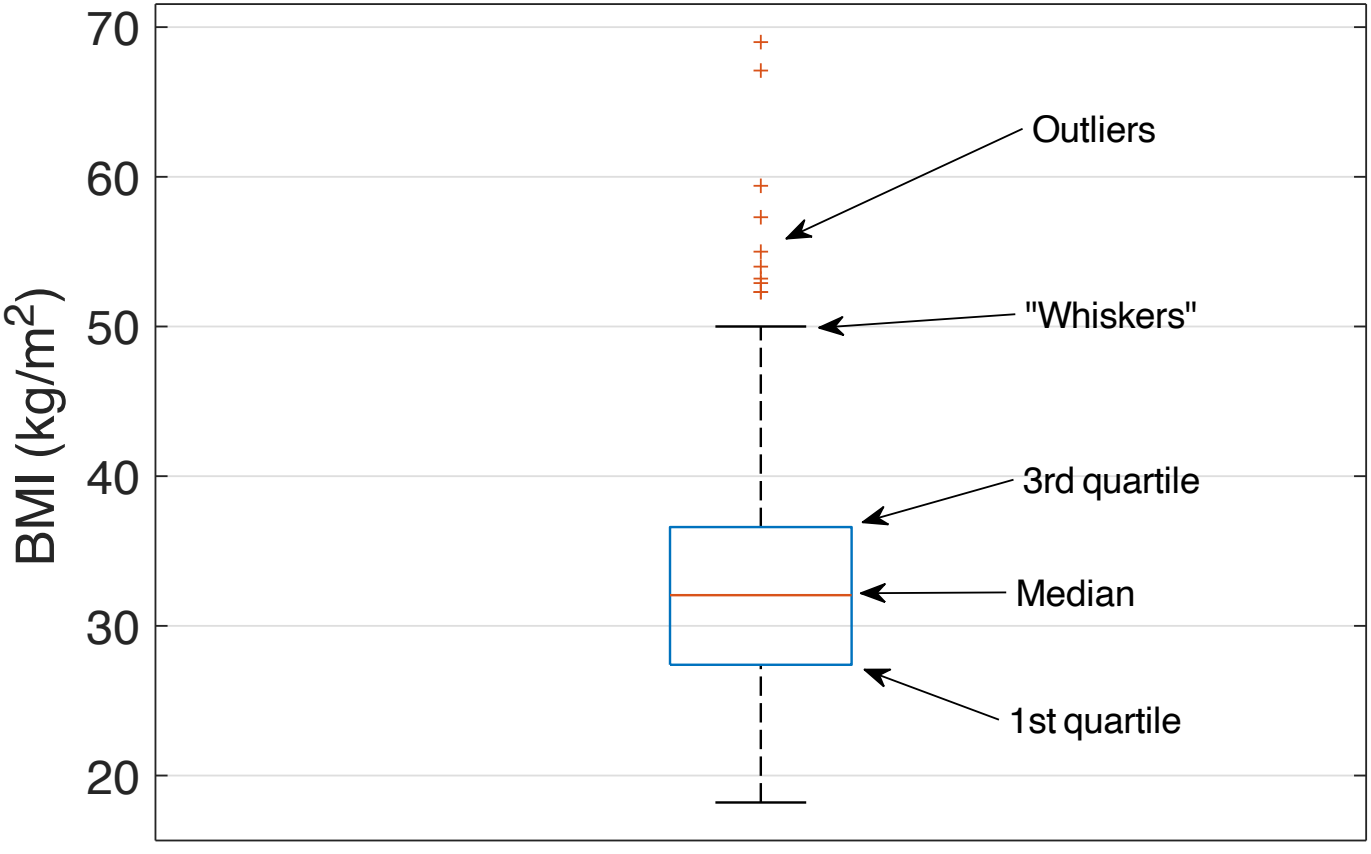
$\text{rng}(\mathbf{y}) = 4.63$ $(\min\{\mathbf{y}\} = -2.61,\ \max\{\mathbf{y}\} = 2.01)$, $s(\mathbf{y}) = 0.5$

Measures of Spread: Example



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

Visualising Continuous Data: Boxplots



Boxplot graphically captures centrality, spread and skewness in one plot

Association Between Two Continuous Variables

- Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ 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

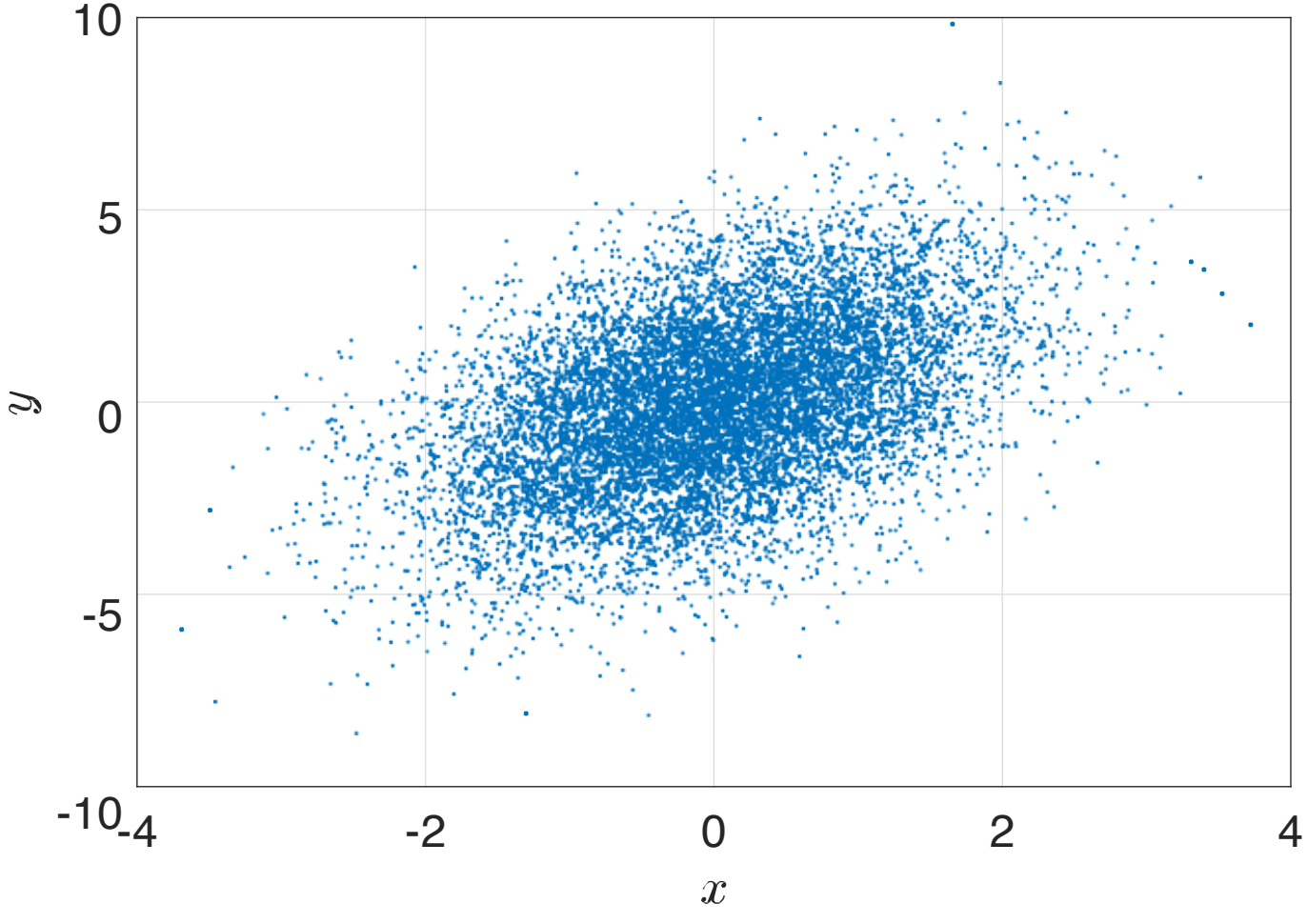
  ⇒ does not imply no non-linear association

- Remember: correlation not equal causation!

Scatter Plots

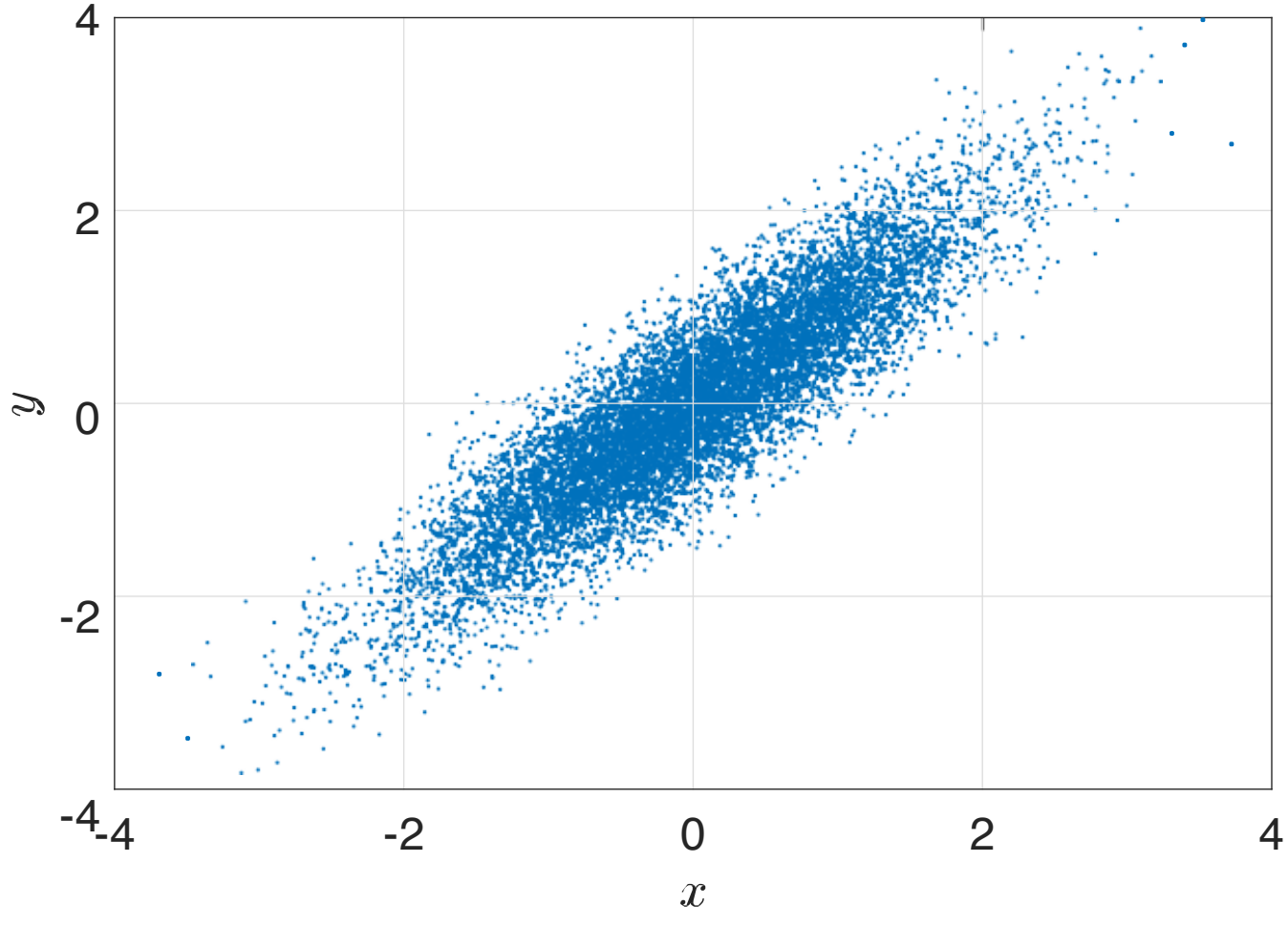
- Scatter plots help us visualise relationships between two (usually) numeric variables
  - Plot points, with one variable on x-axis and one on y-axis
- Can be used to visually look for association
- Correlation coefficients are statistics that quantitatively measure the strength of the association between two variables

Correlation/Scatter Plot Example (1)



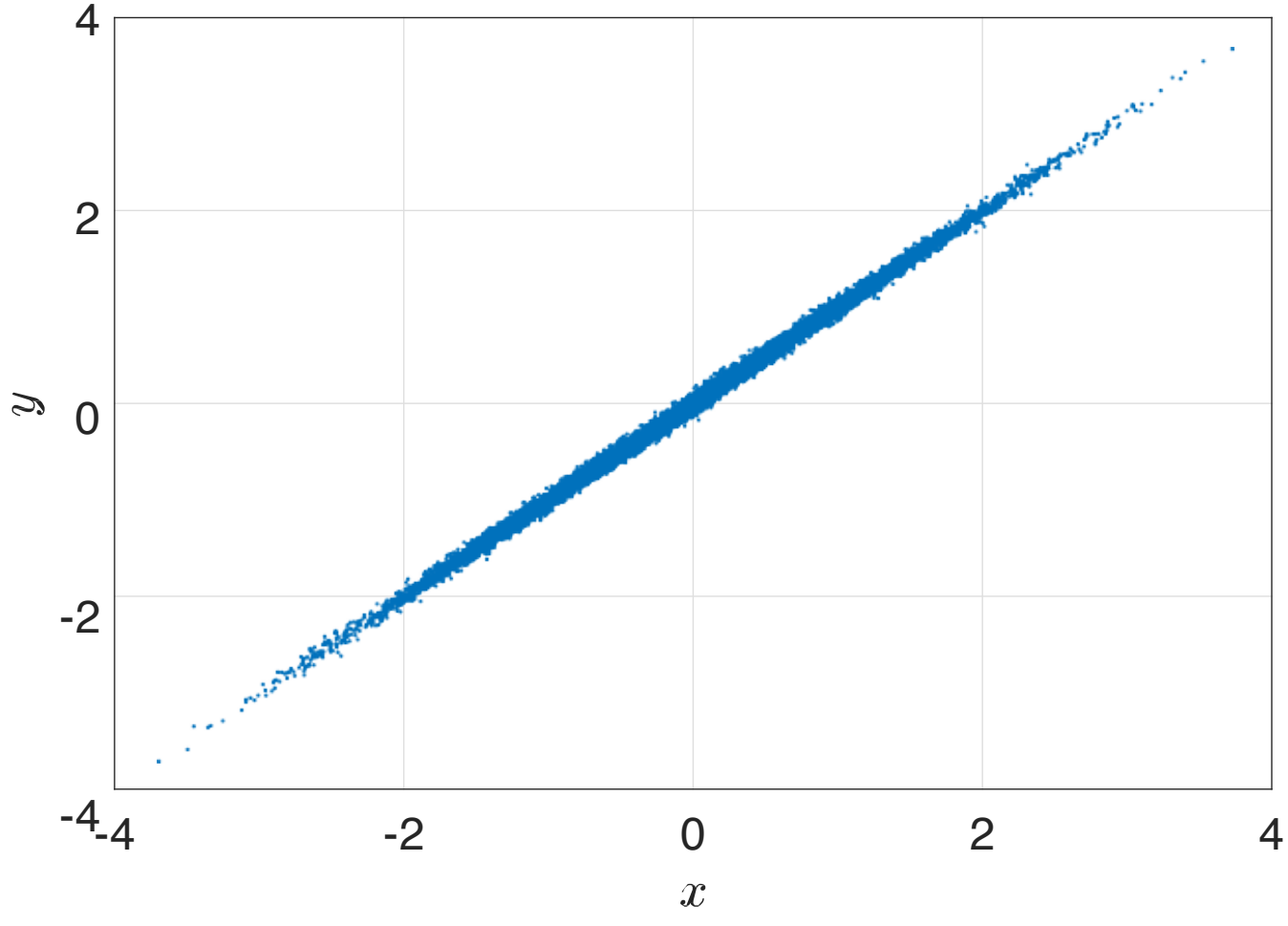
$R \approx 0.44$

Correlation/Scatter Plot Example (2)



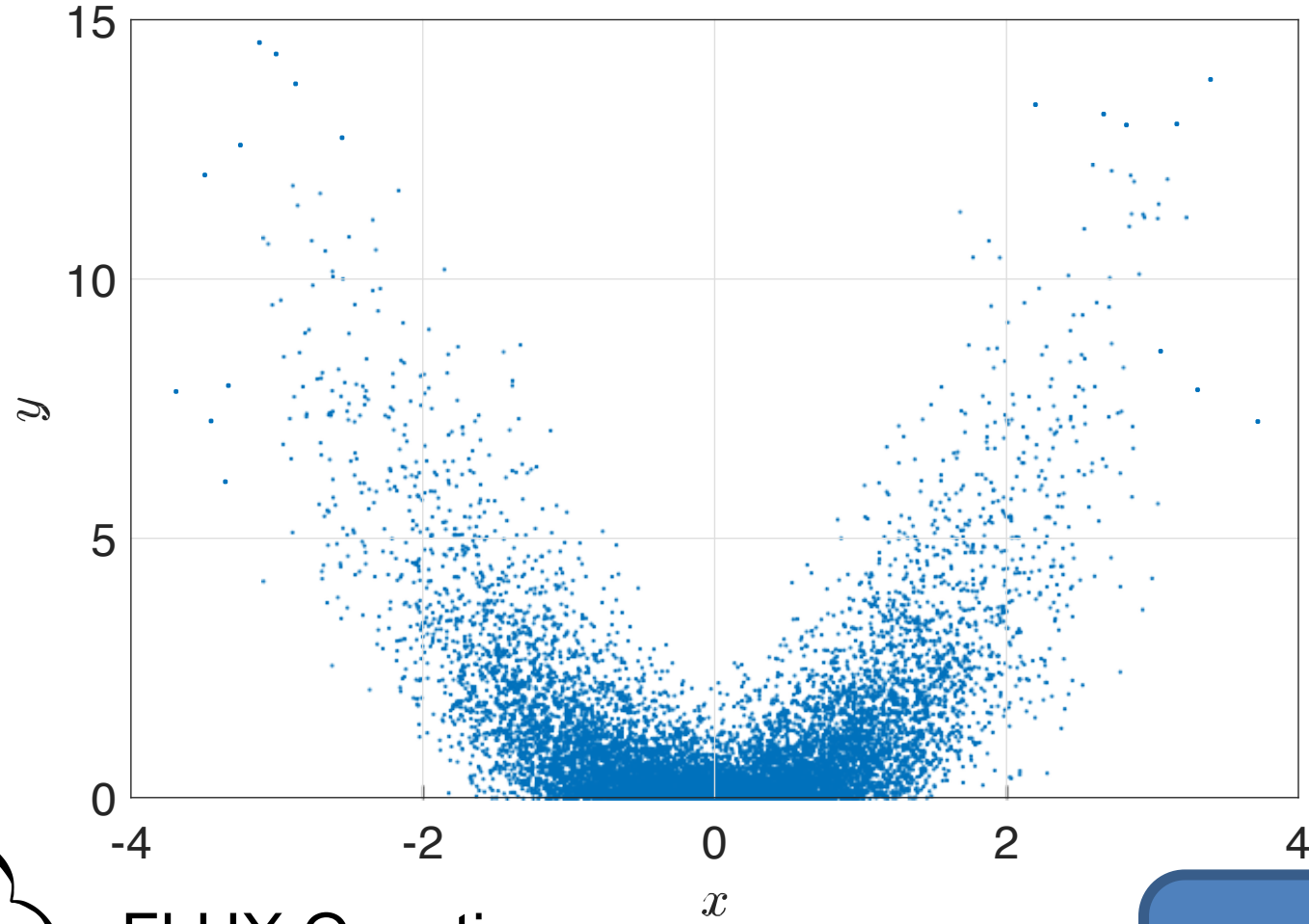
$R = 0.9$

Correlation/Scatter Plot Example (3)



$R \approx 0.999$

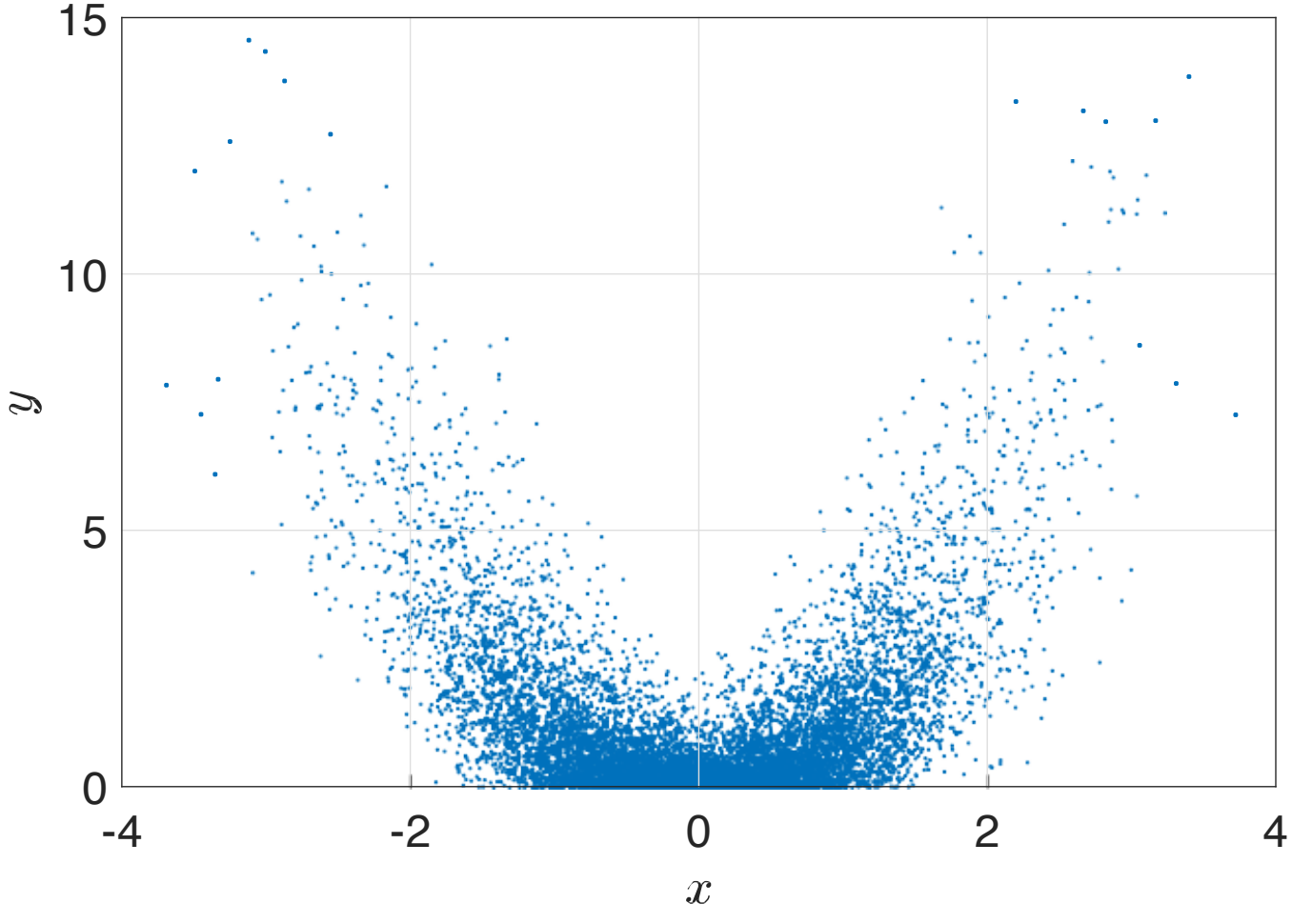
Correlation/Scatter Plot Example (4)



FLUX Question:
Is there a linear association between x and y?

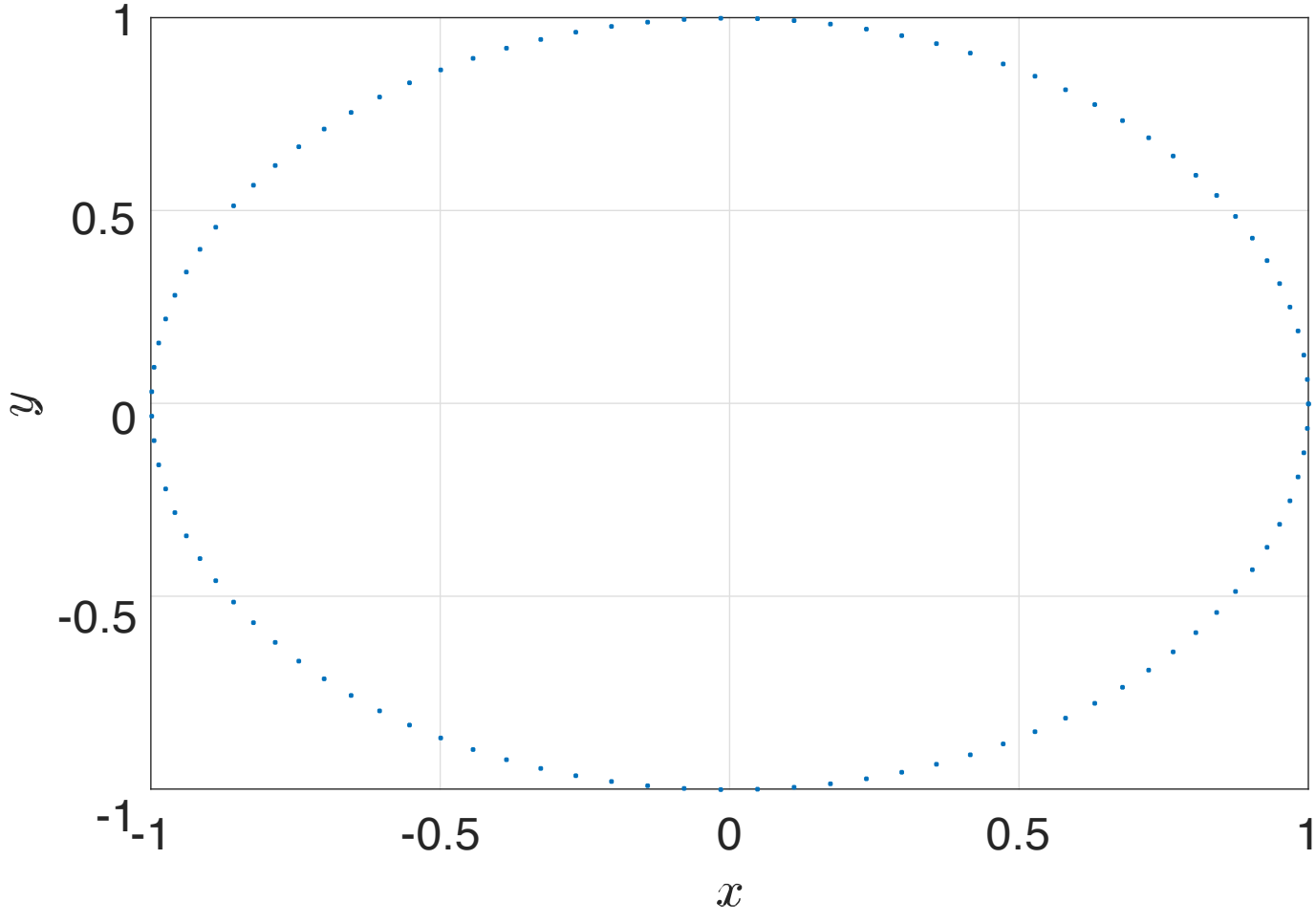
FLUX code: 5NKWED

Correlation/Scatter Plot Example (4)



$R \approx 0.01$ – though clearly associated, as $y = x^2 + \text{noise}$

Correlation/Scatter Plot Example (5)

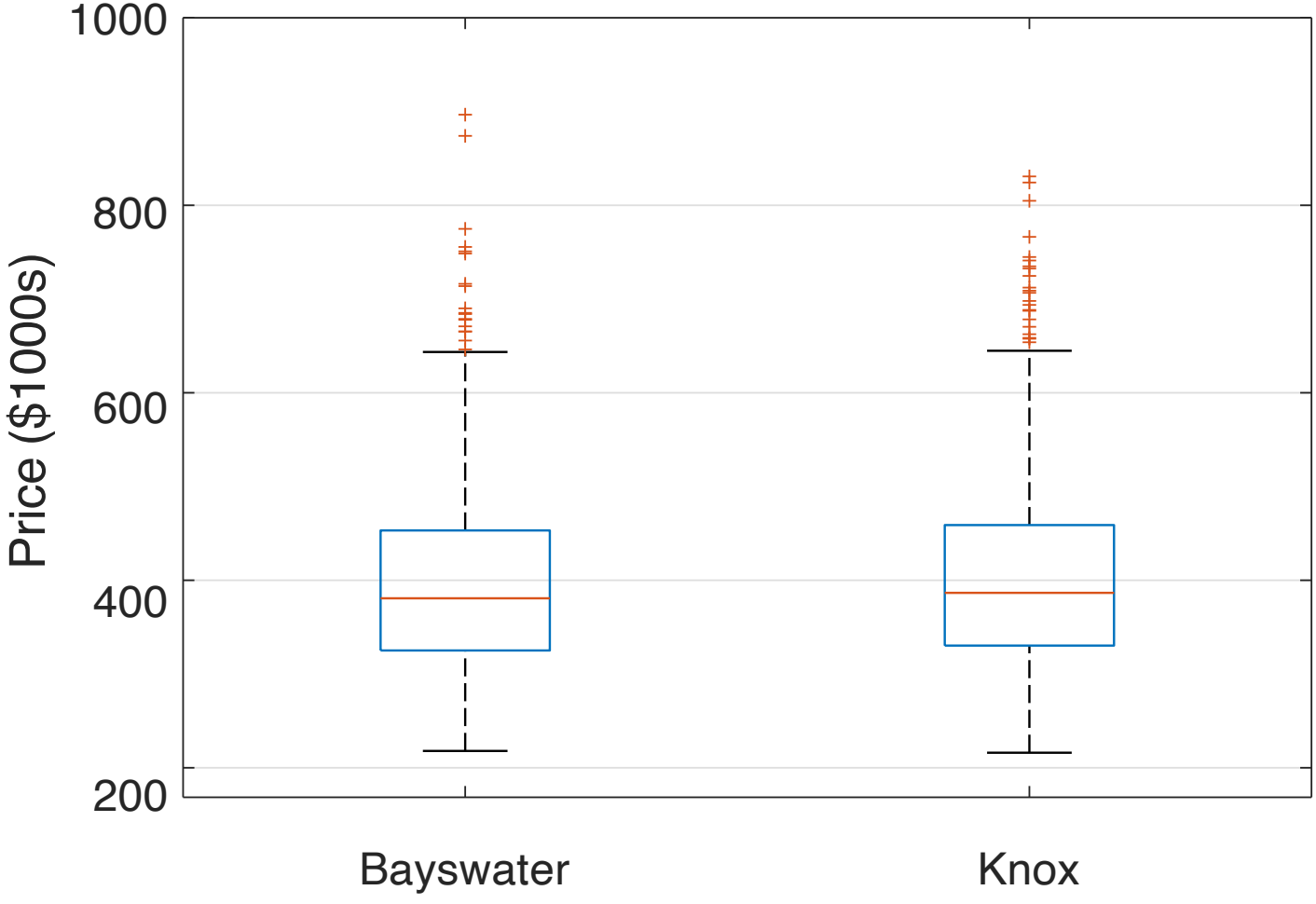


$R = 0$, though there is a deterministic association between x and y

Association Between Categorical and Numeric Variables

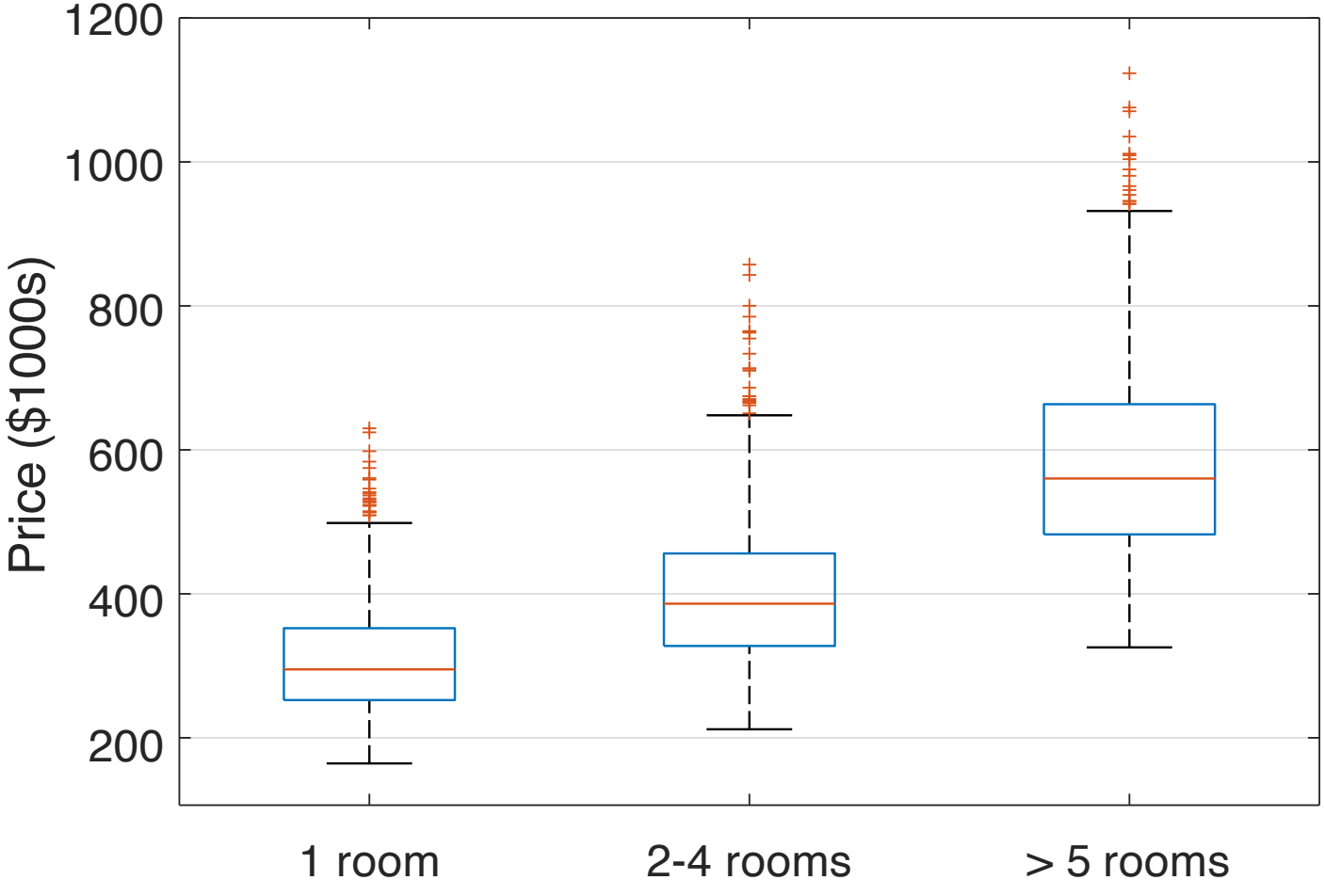
- If **x** is categorical, and **y** is numeric, how to visualise?
- A standard approach is the side-by-side boxplot
  - Divide the data between categories, then plot boxplots for each group
  - Do the boxplots look different?
- If **x** and **y** are both categorical, we can use a side-by-side bargraph instead
  - Are the distributions/bargraphs different between categories? If so, there is a possible association

Example: Categorical and Numeric Variables (1)



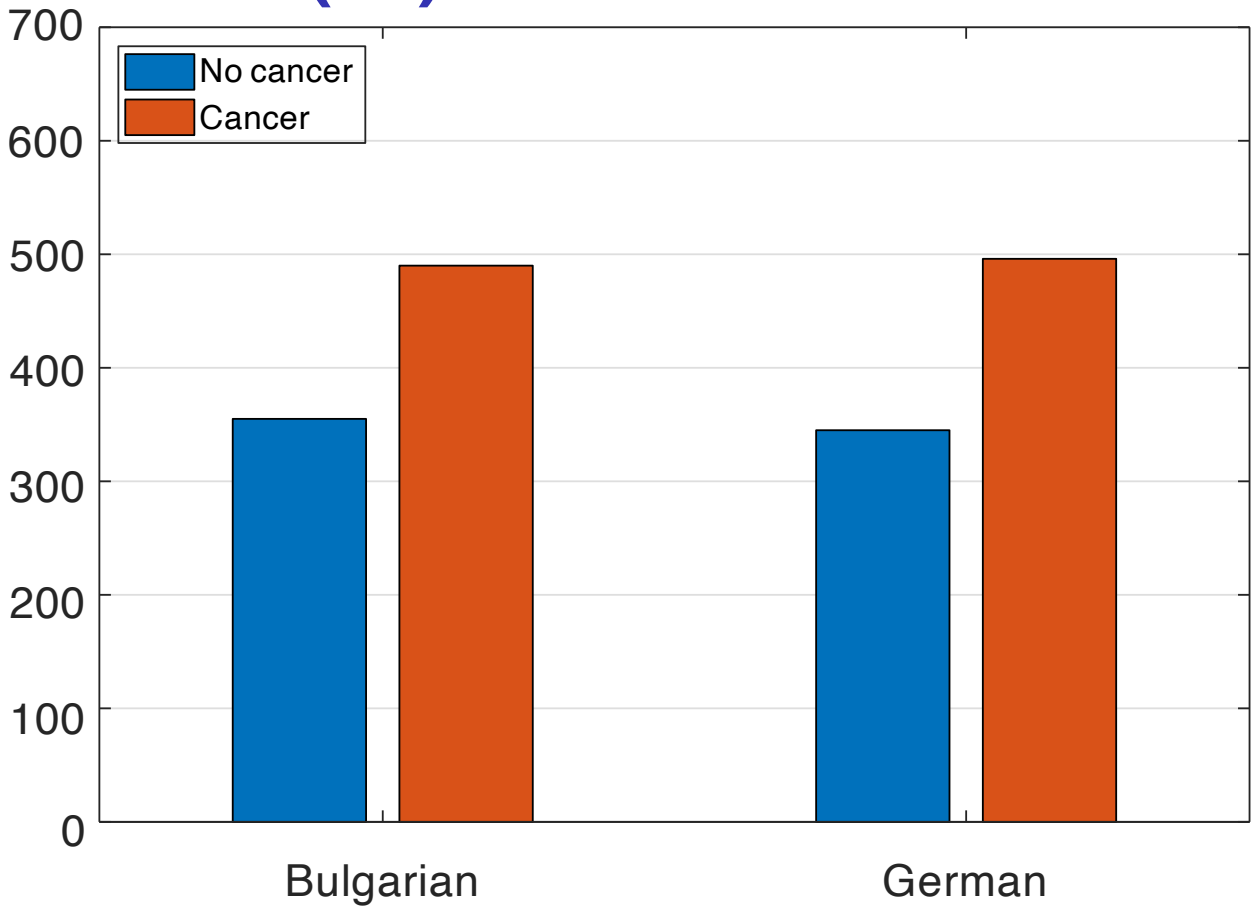
Distribution of price similar between suburbs

Example: Categorical and Numeric Variables (2)



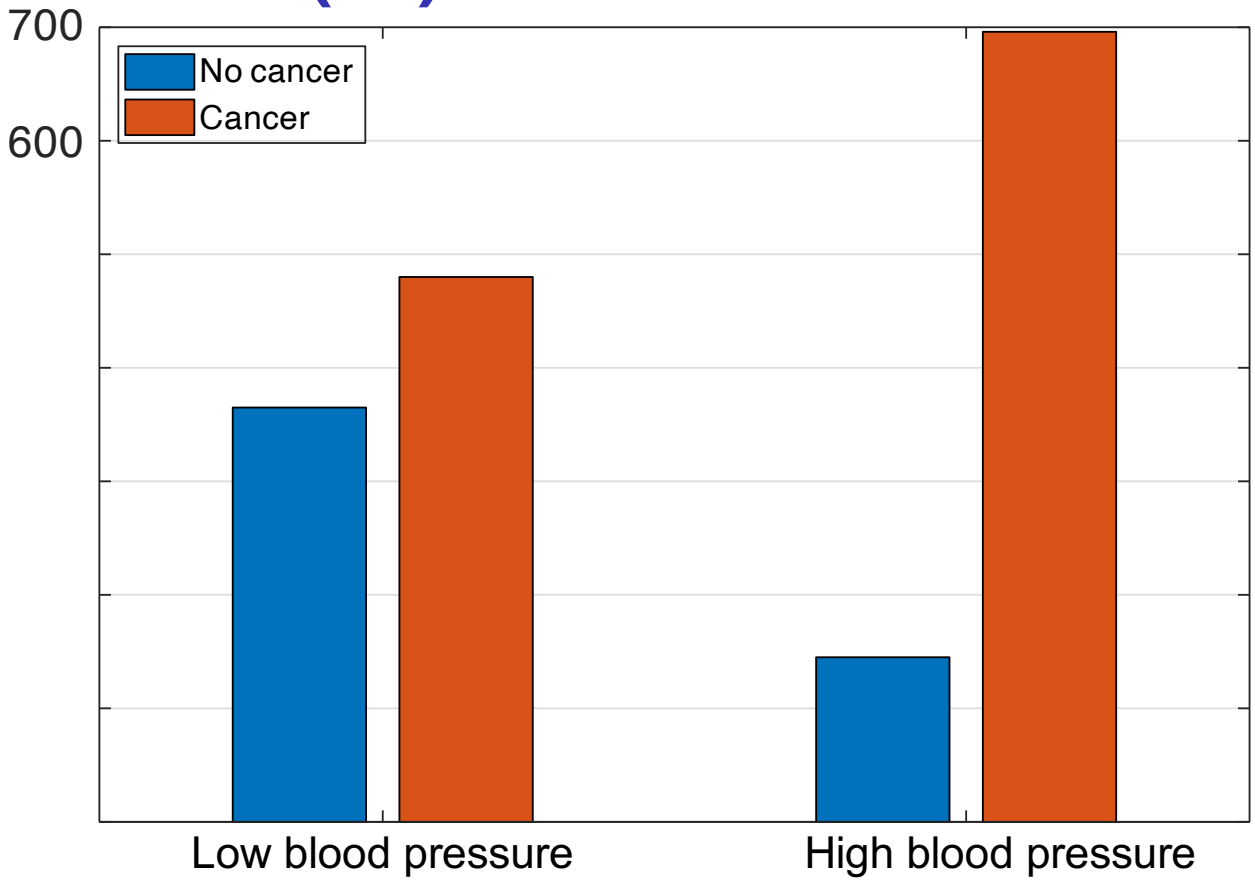
Distribution of price varies greatly with number of rooms

Example: Two Categorical Variables (1)



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

Example: Two Categorical Variables (2)



FLUX Question:

"Frequency of cancer" changes substantially with blood pressure?

FLUX code: 5NKWED

Data visualisation in Python

From Python Data Science Handbook by J. Vanderplas

Plotting data with Matplotlib

Plotting Data in Python: Matplotlib

We can use the matplotlib library to plot data in Python

```
>>> import matplotlib.pyplot as plt
```

Define a table with the data to plot:

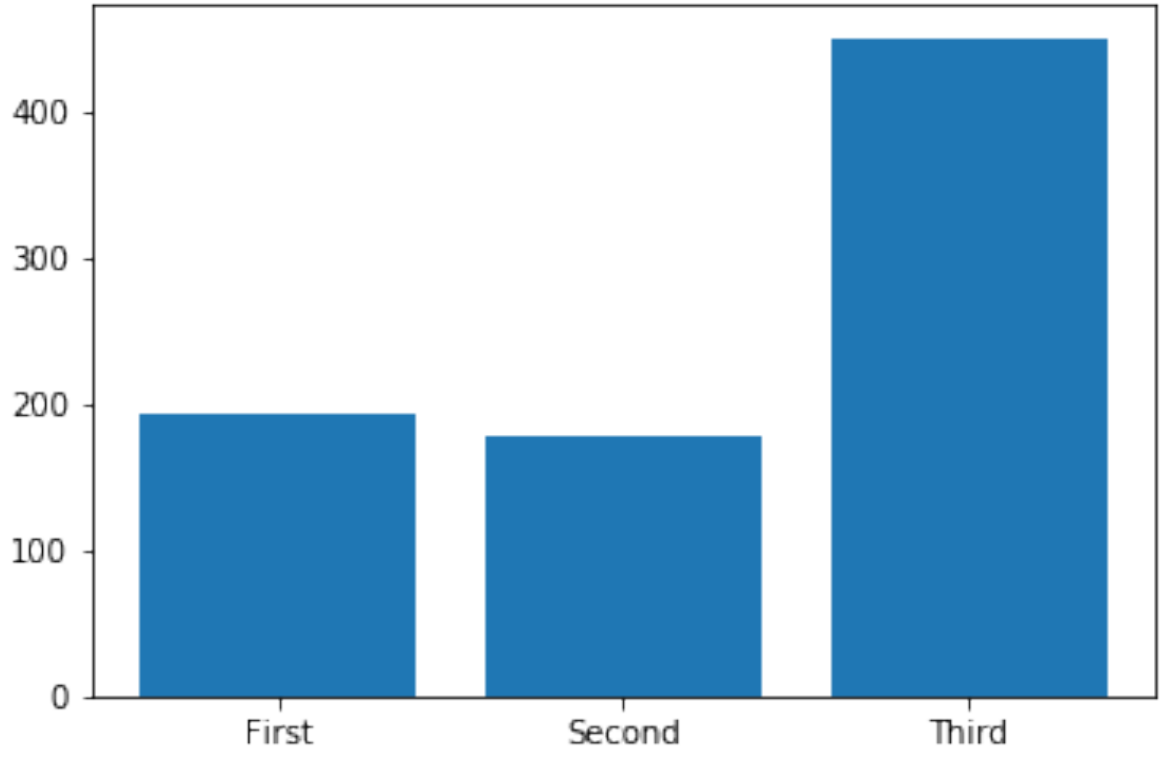
```
>>> myd= { 'Class' : ['First', 'Second', 'Third'],
'Passengers' : [194, 177, 450],
'Average Age' : [39,30,25] }
>>> df = pd.DataFrame(myd)
```

	Class	Passengers	Average Age
0	First	194	39
1	Second	177	30
2	Third	450	25

There are many types of plots for visualising data in Python

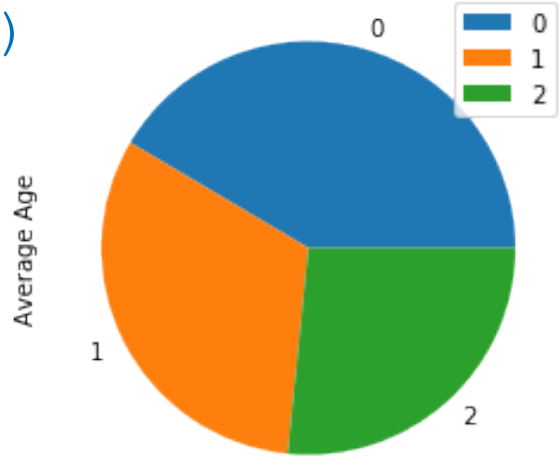
Bar Charts

```
>>> plt.bar(df['Class'],df['Passengers'])
```



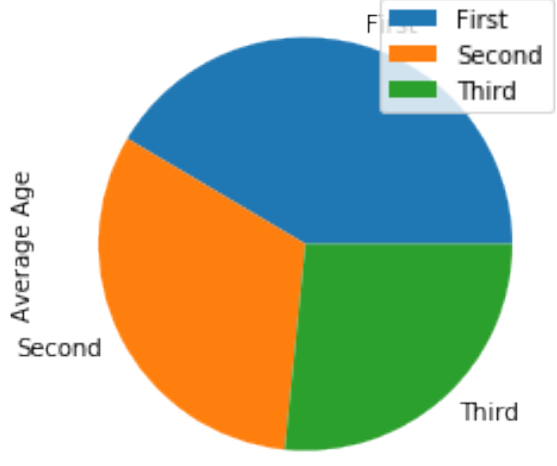
Pie Chart

```
>>> df.plot.pie(y='Average Age')
```



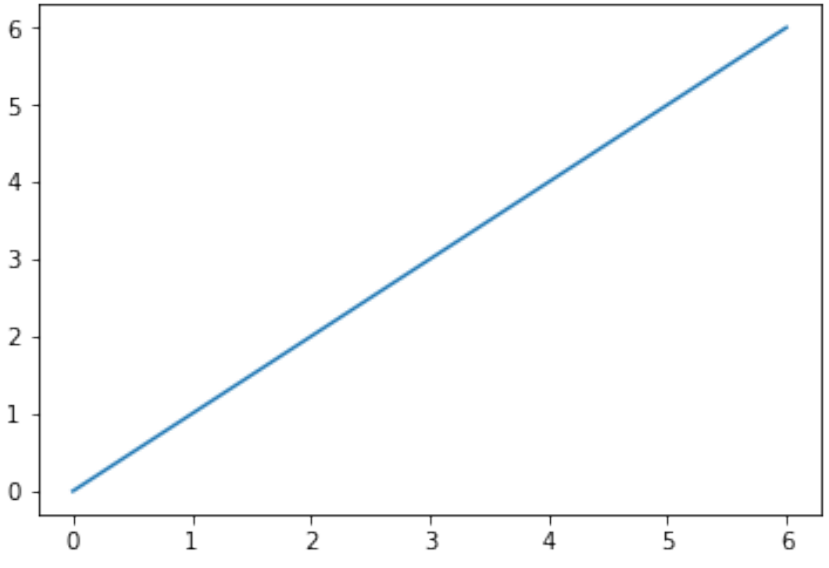
Now you can add index:

```
>>> df1 = pd.DataFrame({
'Passengers' : [194, 177, 450],
'Average Age' : [39,30,25] },
index=['First', 'Second',
'Third'])
>>> df1.plot.pie(y='Average Age')
```

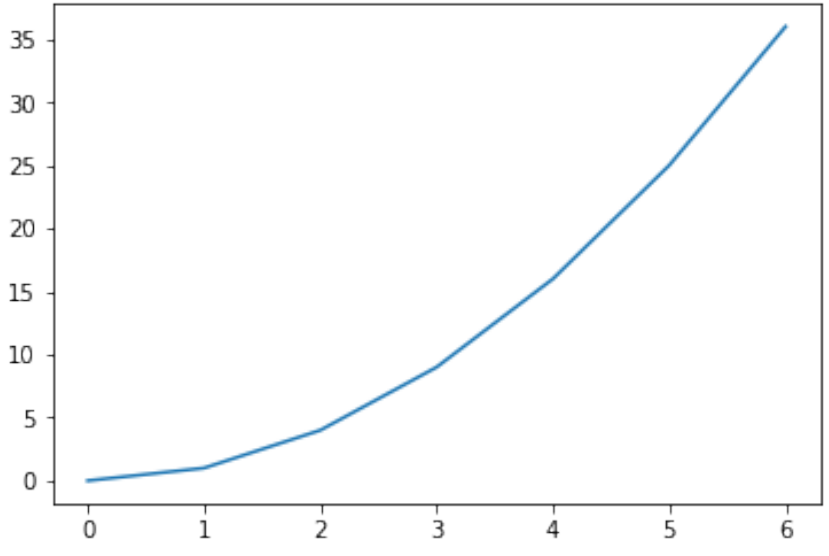


Basic plots

```
df = pd.DataFrame({ 'X' : [0,1,2,3,4,5,6],
'Y' : [0,1,4,9,16,25,36]})
>>> plt.plot(df.col_name)
```



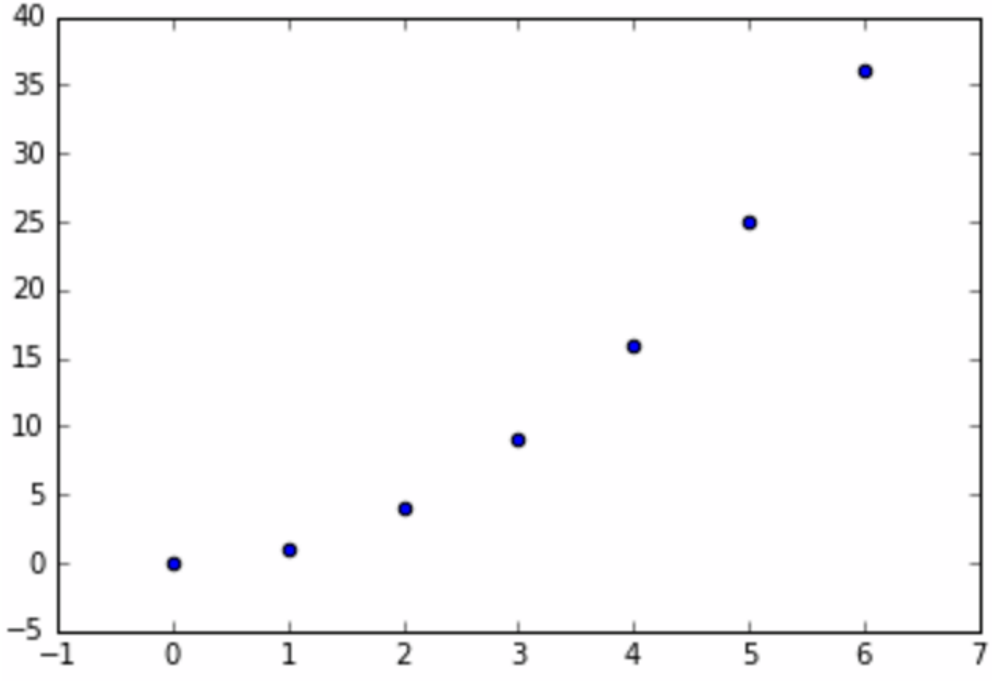
X



Y

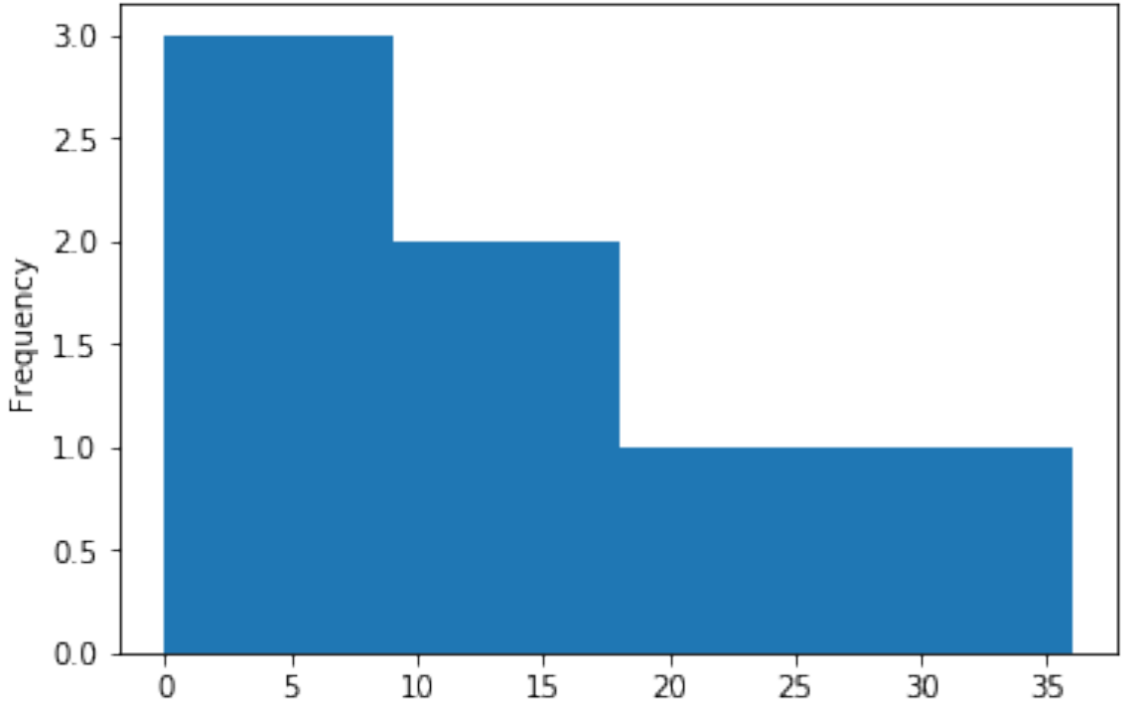
Scatter Plots

```
>>> plt.scatter(df['X'], df['Y'])
>>> plt.show()
```



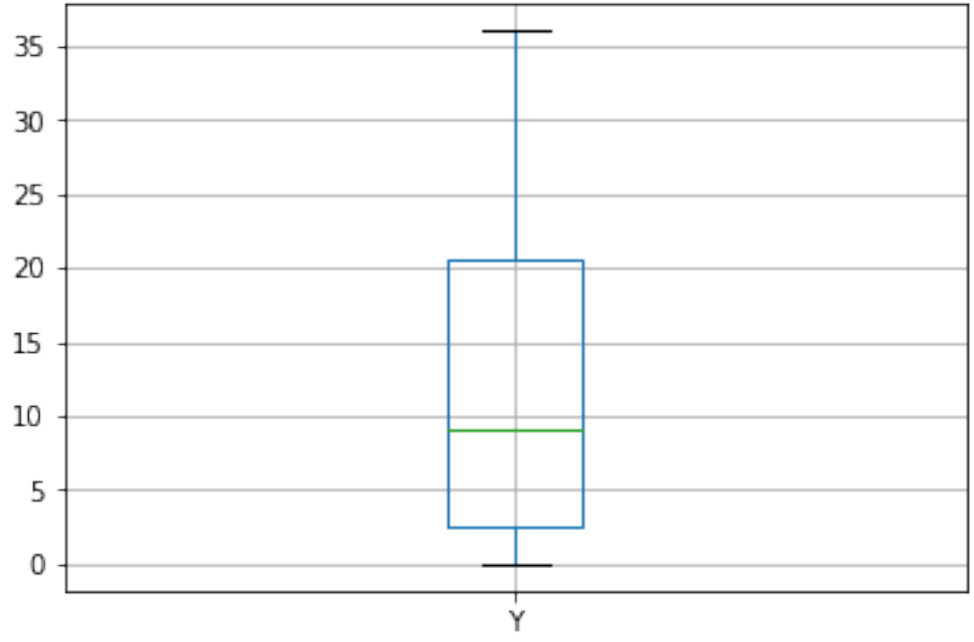
Histograms

```
>>> df.col_name.hist(bins=4)
```



Boxplots

```
>>> df.boxplot(column='col_name')
```

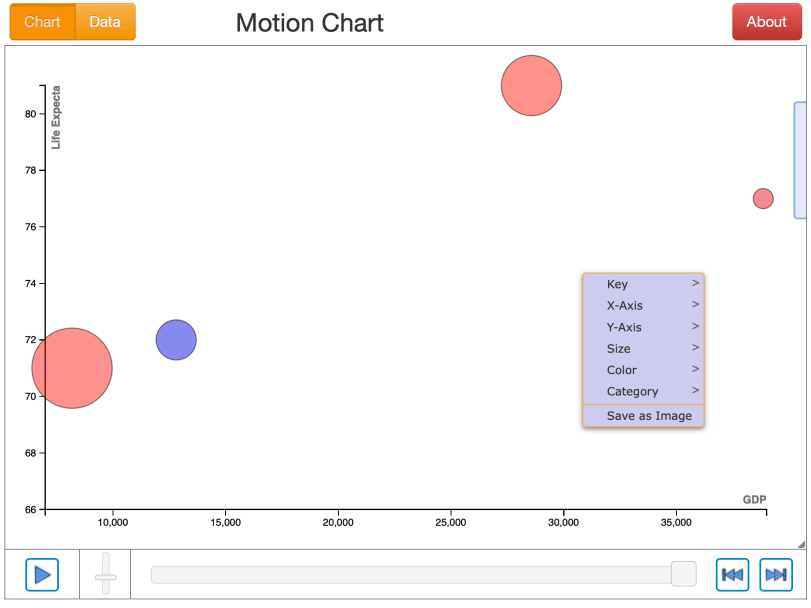


Motion Chart

Open your terminal (mac) or windows prompt, and enter the following:

```
pip install motionchart
pip install pyperclip
```

```
>>> from motionchart.motionchart
import MotionChart
>>> mChart = MotionChart(df =
sampleData)
>>> mChart.to_notebook()
```



Tutorial/Lab week 3

- Advanced data aggregation
- Data visualisation
- Solutions will be released end of the week