ECS7024 Statistics for Artificial Intelligence and Data Science

# Topic 7: Correlation and Scatter Plots

William Marsh

# Outline

- Aim: understand the idea of correlation (or dependence) between two variables

- Correlation and the scatter plot
- Correlation co-efficient
- Covariance
- Correlation matrix
- Limitations of ‘correlation’

# The Idea and Importance of Correlation

# Correlation

- Suppose you run an educational establishment
  - Why do some pupils / students do better?
  - Why do some teachers get better results?

- What varies together? Do the results of teachers increase with their:
  - Height?
  - Experience?
  - Age?
  - Qualifications?

# Independence

- The opposite of correlation (or dependence) is independence
- In probability, A and B independent if:
  - $p(A, B) = p(A) \times p(B)$
  - $p(A \mid B) = p(A)$
- Strength of correlation (informally)

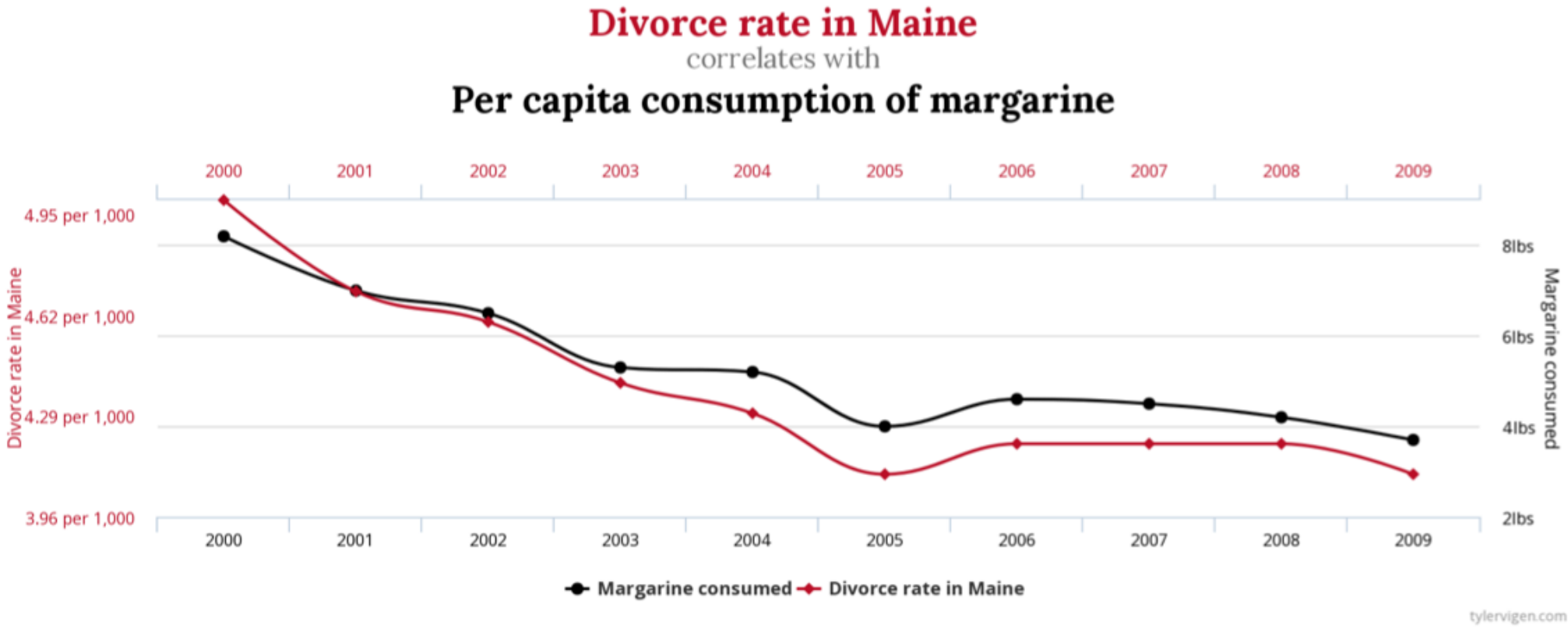
| Correlation between A & B | Knowing A ... |
|---|---|
| Perfect | Determines B |
| Strong | Narrows the distribution of possible B |
| Weak | Slightly narrows ... |
| None | Provide no information about B |

# Examples: Possible Correlations

- Person's height and weight
- Person's height and shoe size
- Patients' height and child's height
- Goals (runs) last season, goals (runs) this season
- Salary and political preference
- ....

# Spurious Correlation

- 'Spurious' (i.e. 'not real')
  - Correlation is temporary or co-incidental
  - *More in a later lecture*




From: https://www.tylervigen.com/spurious-correlations

# Quiz 1

# Heart Data

From kaggle

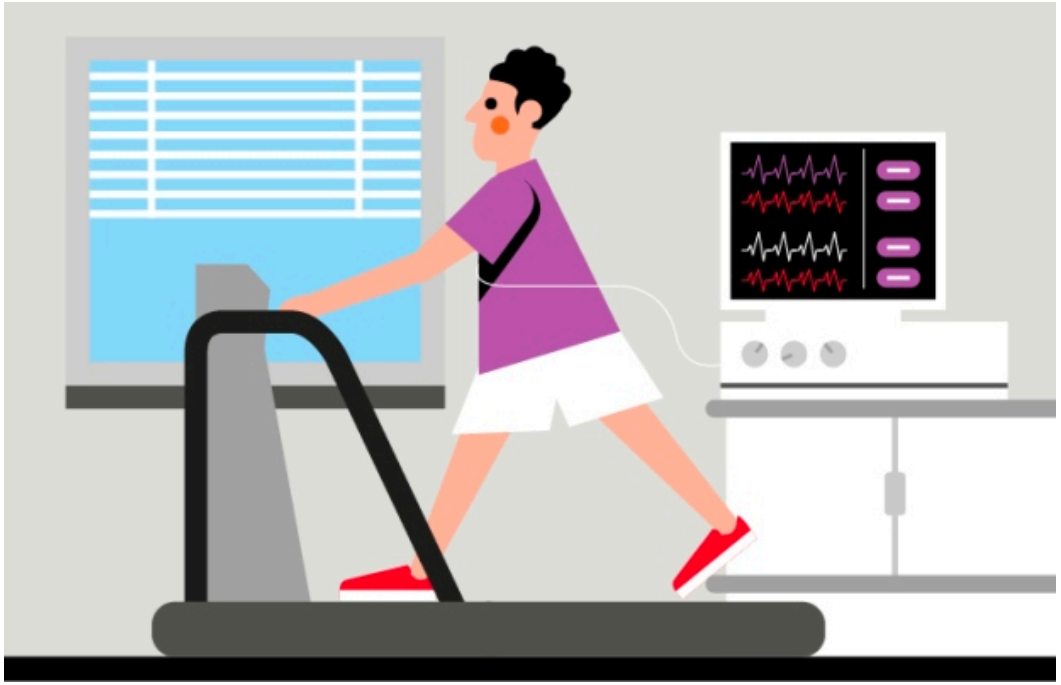
# Kaggle Data: Heart Disease I

| Variable | Meaning | Type |
|---|---|---|
| **Age** | The person's age in years | Continuous |
| **Sex** | 1 = male, 0 = female | Categorical |
| **ChestPain** | The chest pain experienced (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic) | Categorical |
| **RestBP** | The person's resting blood pressure (mm Hg on admission to the hospital) | Continuous |
| **Chol** | The person's cholesterol measurement in mg/dl | Continuous |
| **Bsugar** | The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false) | Binary |
| **RestECG** | Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable e left ventricular hypertrophy) | Ordinal (?) |

# Kaggle Data: Heart Disease II

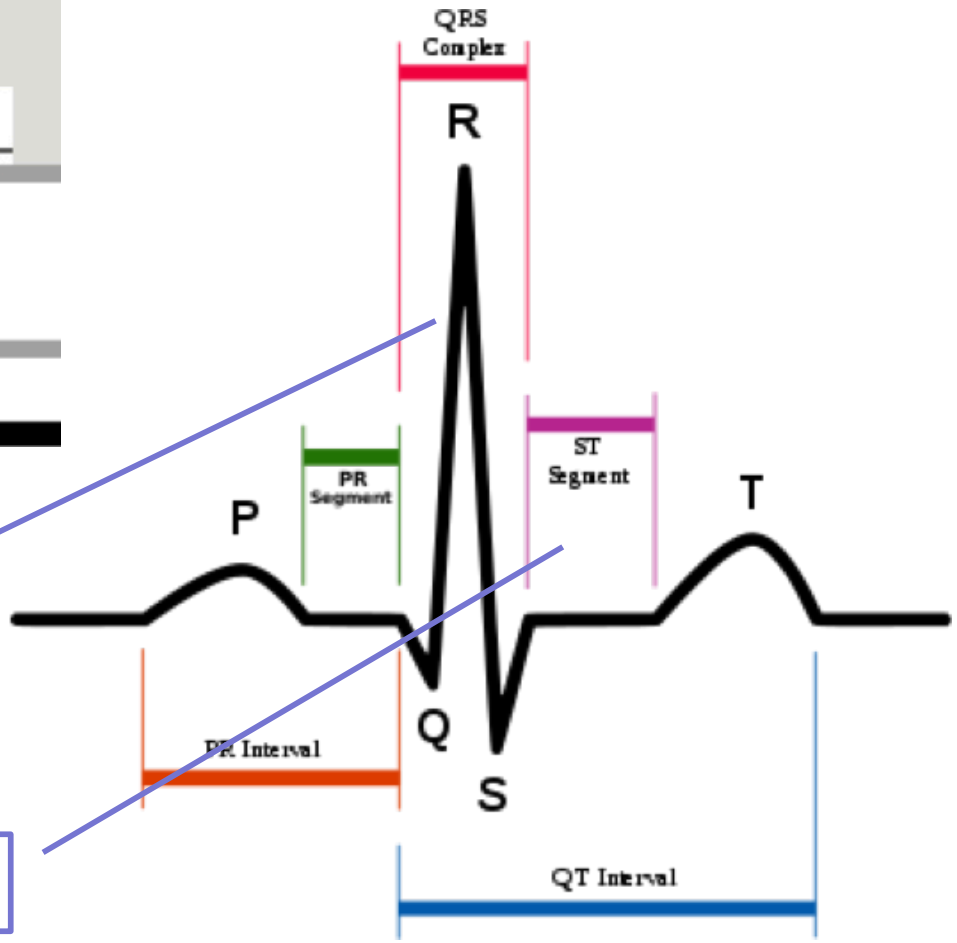
| Variable | Meaning | Type |
|---|---|---|
| **MaxRate** | The person's maximum heart rate achieved | Continuous |
| **Angina** | Exercise induced angina (1 = yes; 0 = no) | Binary |
| **ECG_ST_d** | ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot) | Continuous |
| **ECG_ST_slope** | The slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping) | Categorical |
| **Vessels** | The number of major vessels (0-3) coloured by fluoroscopy | Ordinal |
| **Thallium** | Thallium update test (0 = normal; 1 = fixed defect; 2 = reversible defect) | Categorical |
| **Disease** | Heart disease (0 = no, 1 = yes) | Binary |

# About (Exercise) ECG



From https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/tests/stress-test

ECG waveform

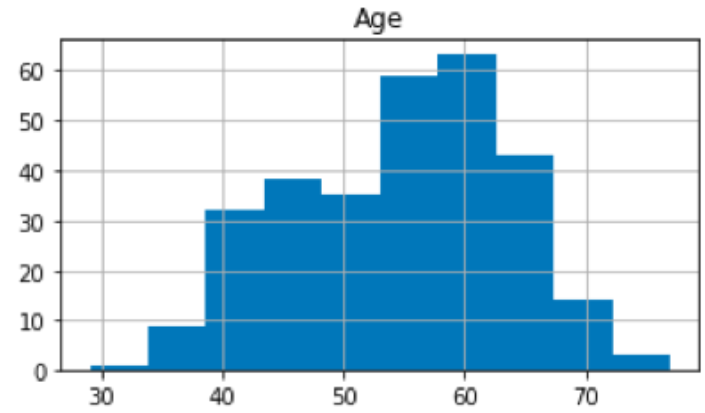


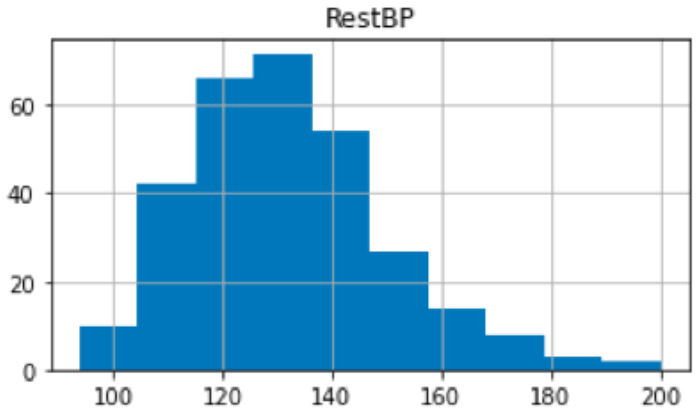

De-polarisation (contraction) of ventricle

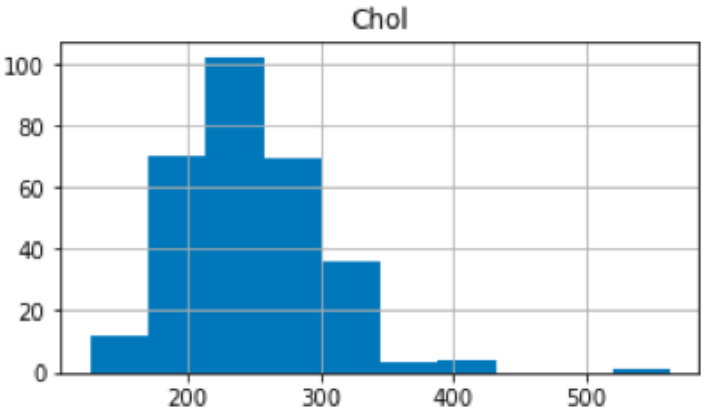
ST segment

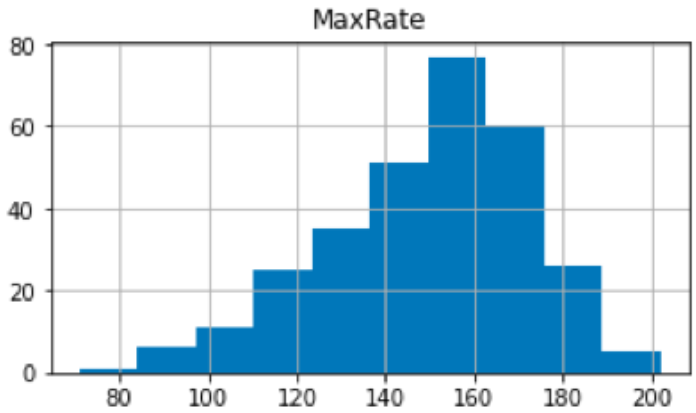
From https://en.wikipedia.org/wiki/ST_segment

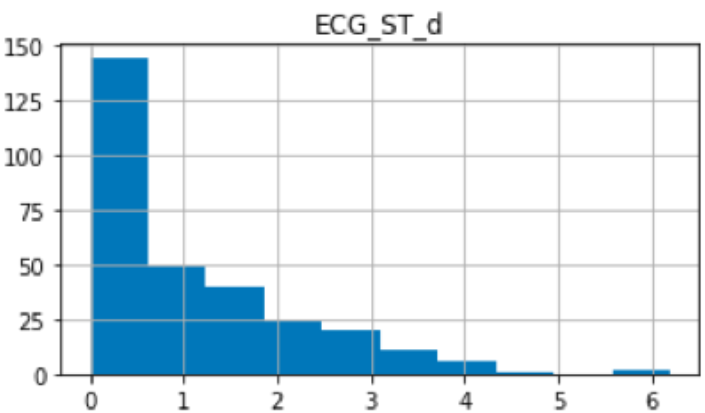
# Distributions of Continuous Variables



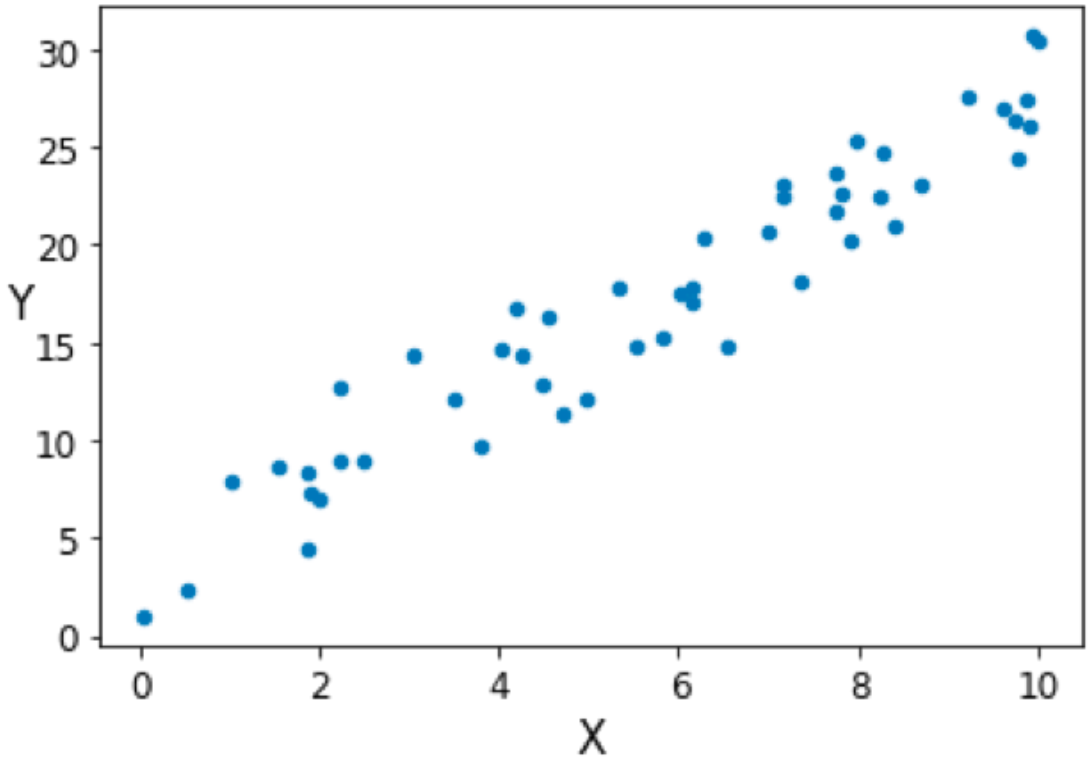



# Scatter Plot

Visualise two variables

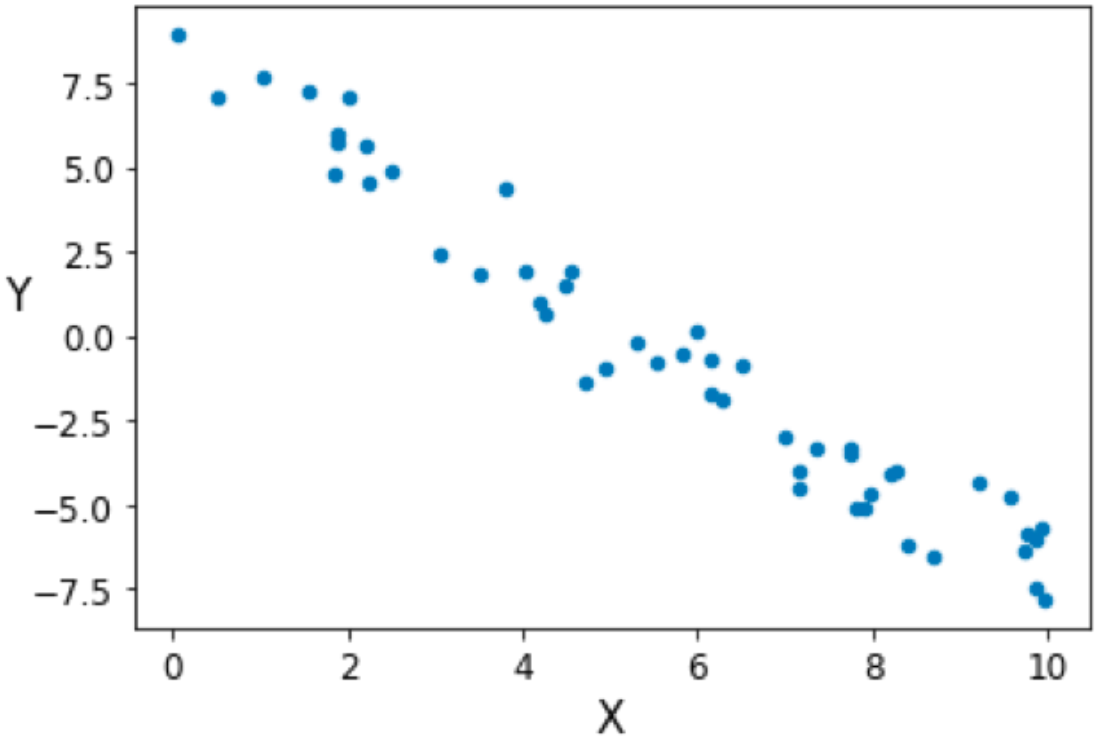
# Scatter Plot Principles

- Visualise the relationship between two variables
- There is a relationship
  - –Knowing X tells you (imprecisely) about Y
  - –As X increases, Y increases



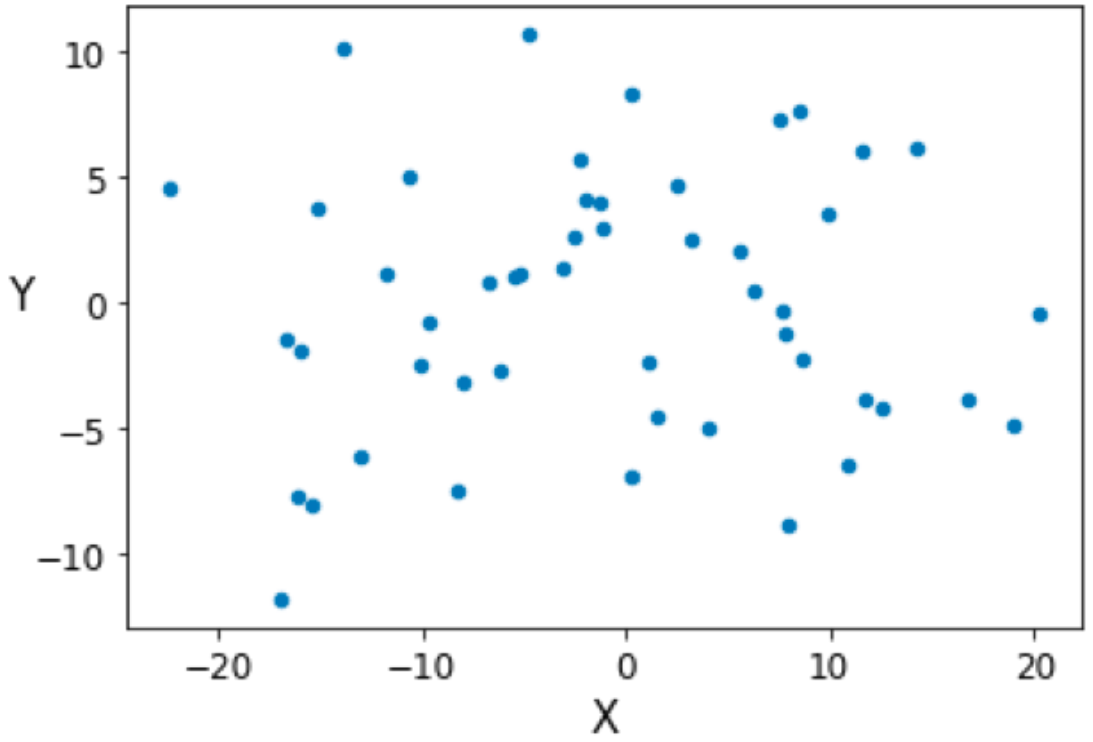
# Scatter Plot Principles

- Visualise the relationship between two variables
- There is a relationship
  - –Knowing X tells you (imprecisely) about Y
  - –As X increases, Y decreases

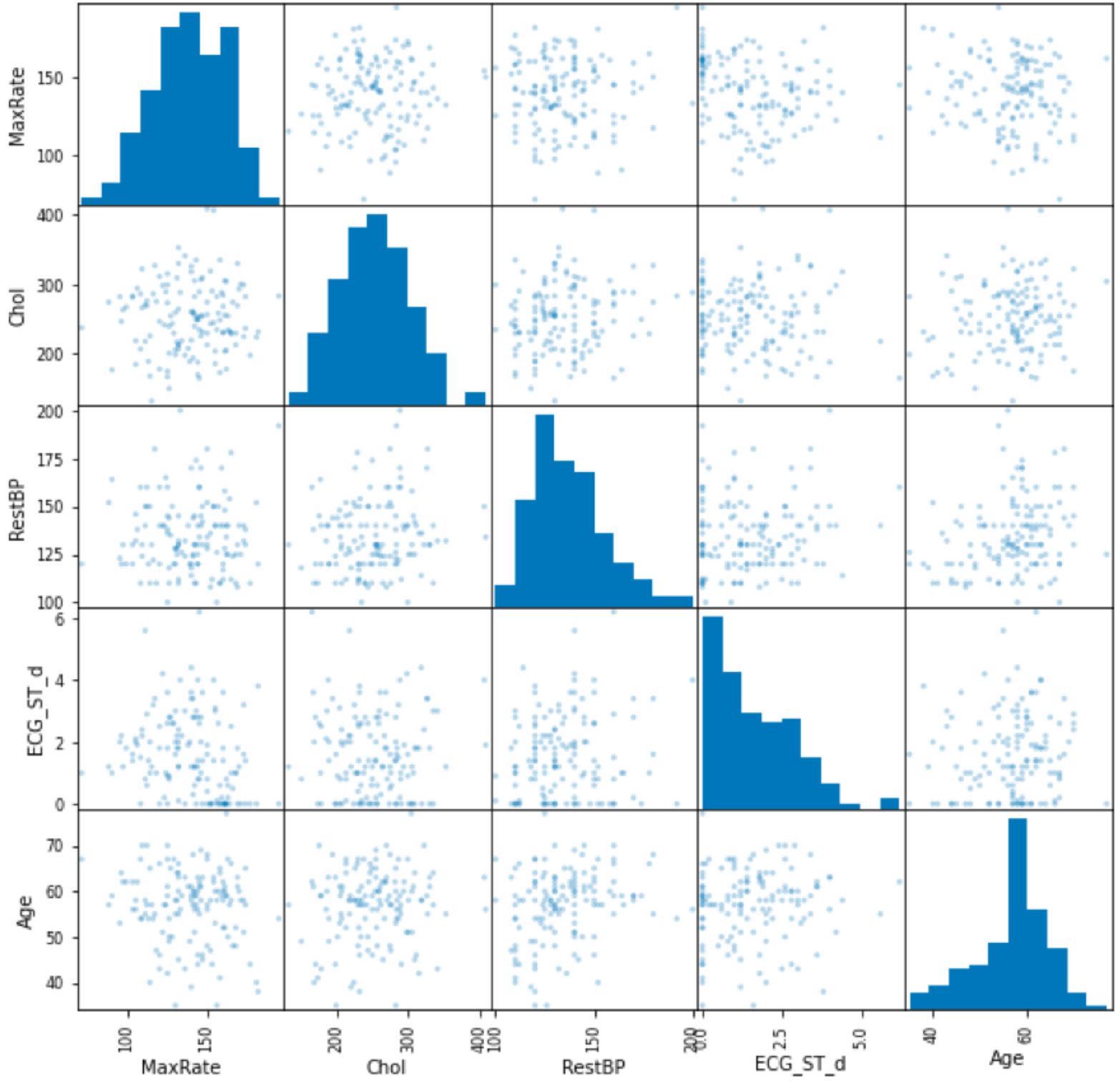


# Scatter Plot Principles

- Visualise the relationship between two variables
- There is no relationship
  - –Knowing X tells you nothing about Y

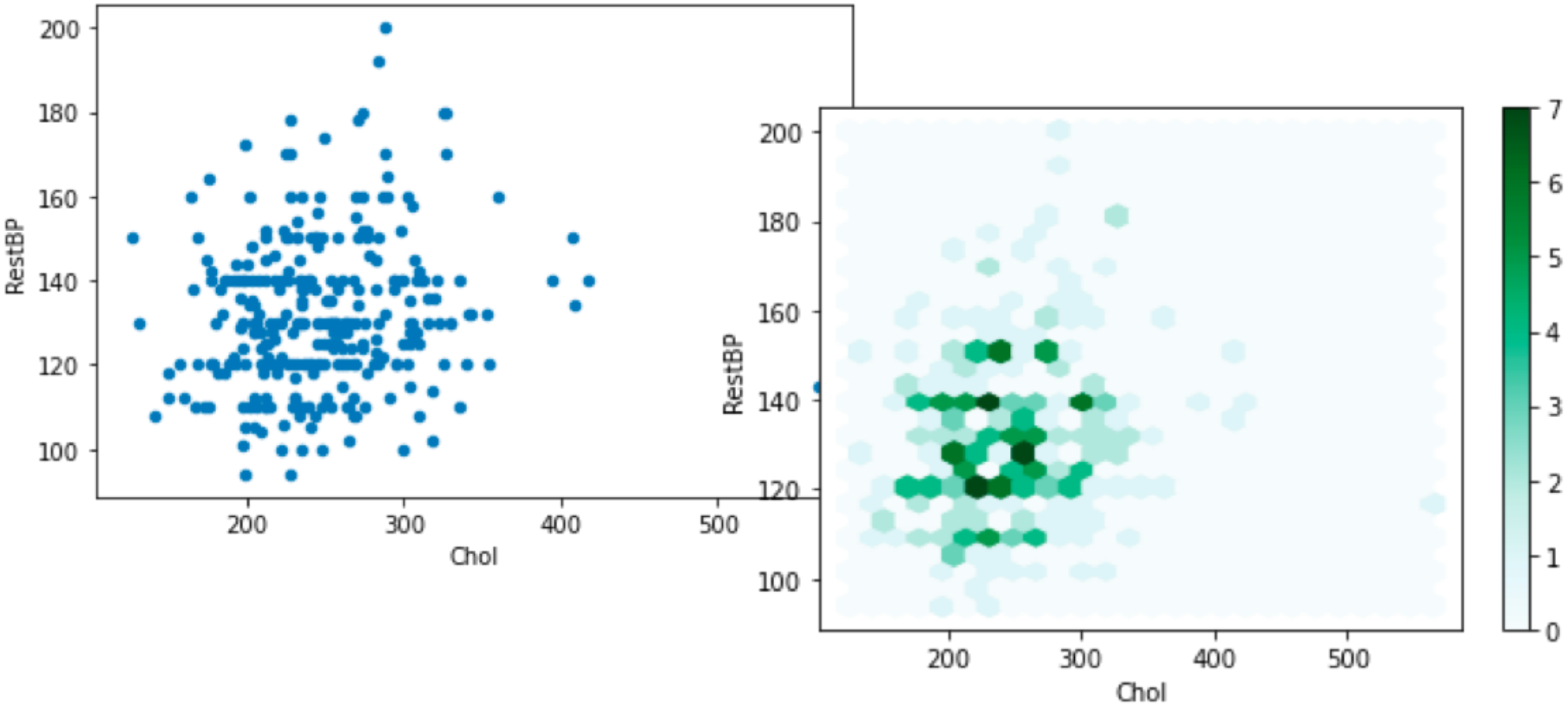


Scatter Matrix

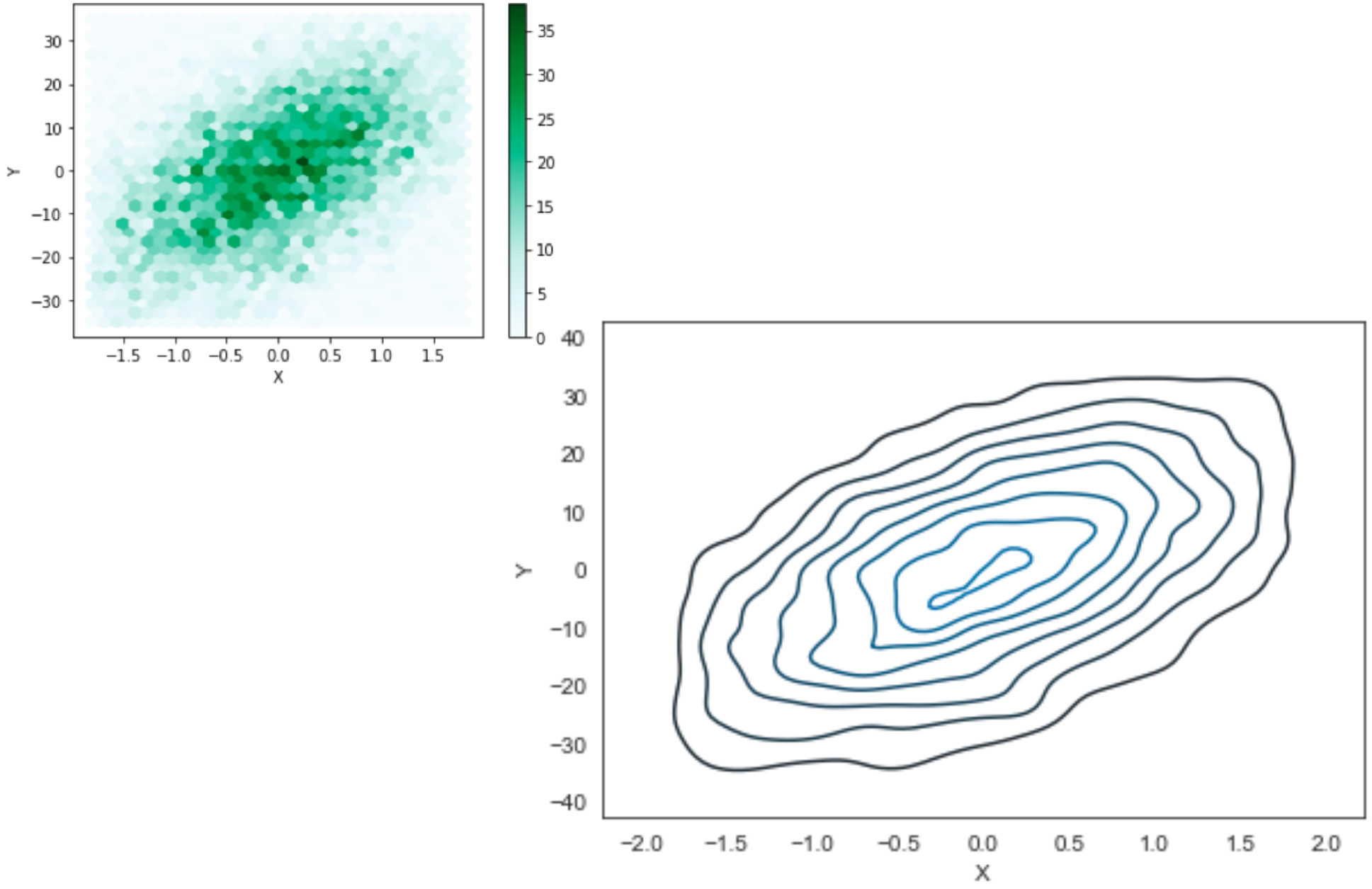


# The ‘Hexbin’

- Problem of scatter when too many points
  - Solution 1: alpha value between 0 and 1
  - Solution 2: ‘hex bins’



# X-Y Plots for Very Dense Data



# Quiz 2

Every lecture will have a ‘learning reflection’ slide

# Misconceptions and Barriers

Obstacles on the path to understanding

# Misconceptions and Barrier

**Theory**

- Barrier
  - A concept that you need to understand to move forward
- A misconception
  - A mistaken or incomplete understanding
  - Needs to be revised

# Misconceptions and Barrier

## Theory

- Barrier
  - A concept that you need to understand to move forward
- A misconception
  - A mistaken or incomplete understanding
  - Needs to be revised

## Examples

- Condition is not an expression
  - An expression has a value
  - Value can be assigned to a variable

```
x > 3  #an expression with a value
bigger = x > 3  #assignment
```

# Misconceptions and Barrier

## Theory

- Barrier
  - A concept that you need to understand to move forward
- A misconception
  - A mistaken or incomplete understanding
  - Needs to be revised

## Examples

- Condition is not an expression
  - An expression has a value
  - Value can be assigned to a variable

```
x > 3  #an expression with a value
bigger = x > 3  #assignment
```

- Combining partial conditions

```
btwn = x > 3 and < 7
```

```
btwn = (x > 3) and (x < 7)
```

# Correlation Coefficient

# Correlation – Insight

- Two variables X and Y
  - Each variable has a mean: $\mu_X$ , $\mu_Y$
  - Values of X (written $x_i$) and of Y ($y_i$) can be above or below the mean

- Correlated
  - When $x_i$ is far (above / below) from $\mu_X$ then ...
  - ... $y_i$ is also far (in same direction) from $\mu_Y$
- Independent
  - When $x_i$ is far above (or below) from $\mu_X$ then ...
  - ... $y_i$ is could be near $\mu_Y$ or far in the opposite direction

# Correlation Coefficient: Definition

- Two variables X and Y
  - Each variable has a mean: $\mu_X$ , $\mu_Y$
  - Values of X (written $x_i$) and of Y ($y_i$)
  - N values: $(x_i, y_i)$

Correlation coefficient: -1 to 1

Average (over N values) of the product of each variable's distance from its mean

$$\rho_{X,Y} = corr(X,Y) = \frac{\frac{1}{N}\sum_1^N (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y}$$

Standard deviations to normalise

# Correlation Coefficient II

- Variations in notation

$$\rho_{X,Y} = \frac{\frac{1}{N}\sum_1^N (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y}$$

Expected (or average) value

$$\rho_{X,Y} = \frac{E[(x_i - \mu_X)(y_i - \mu_Y)]}{\sigma_X \sigma_Y}$$

# Sample Correlation

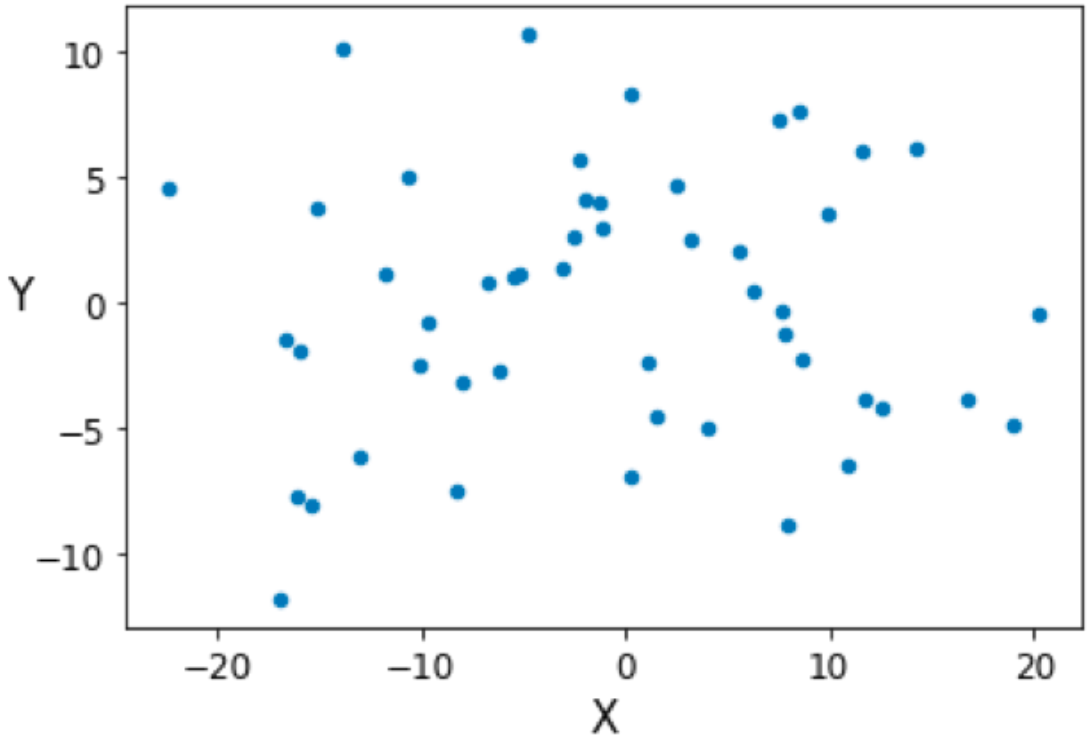
- Correlation can be calculated for a sample
  - Population $\rho_{X,Y}$
  - Sample $r_{X,Y}$
- Turns out to be simple (no adjustment):

$$r_{X,Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

No need to remember this formula

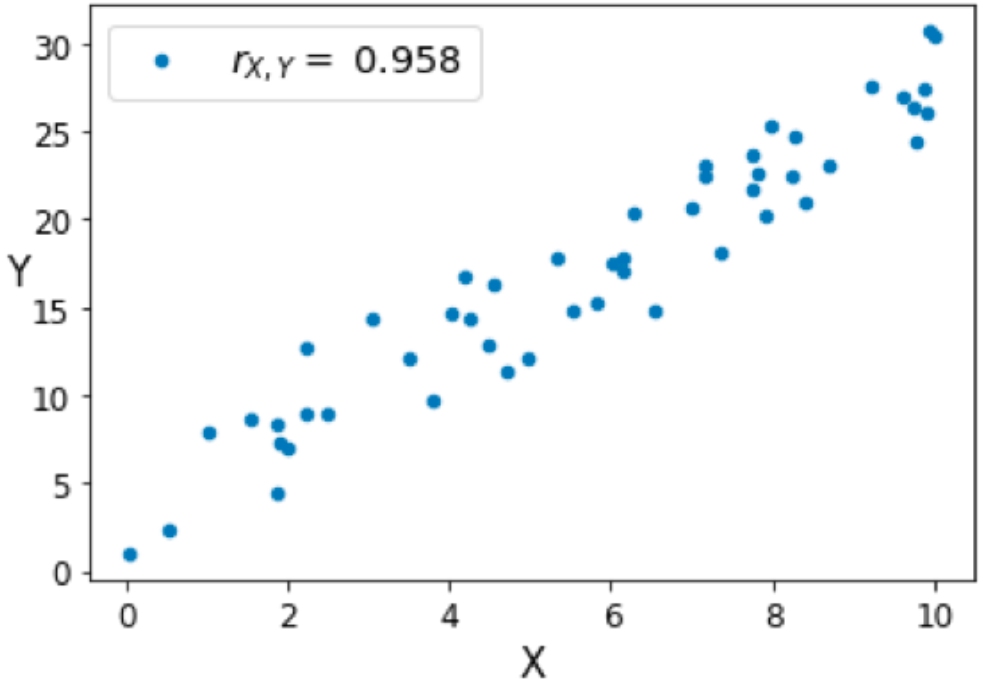
# Correlation Values

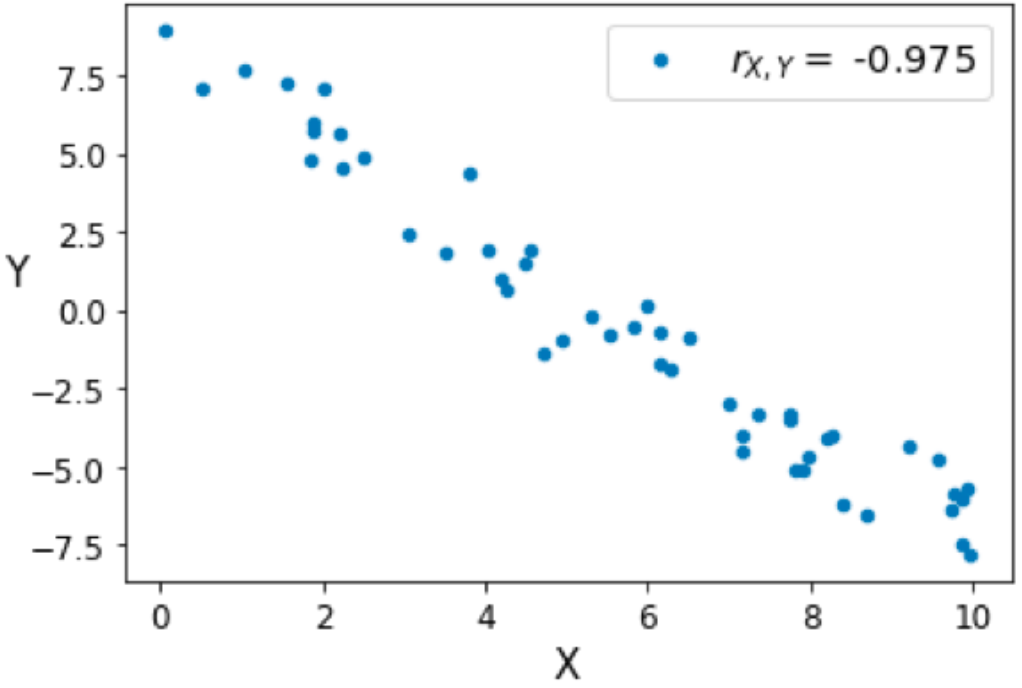
- Range -1 (-ve correlation) to 1 (+ve correlation)
- Corr = 0 implies uncorrelated




Correlation = 0.045

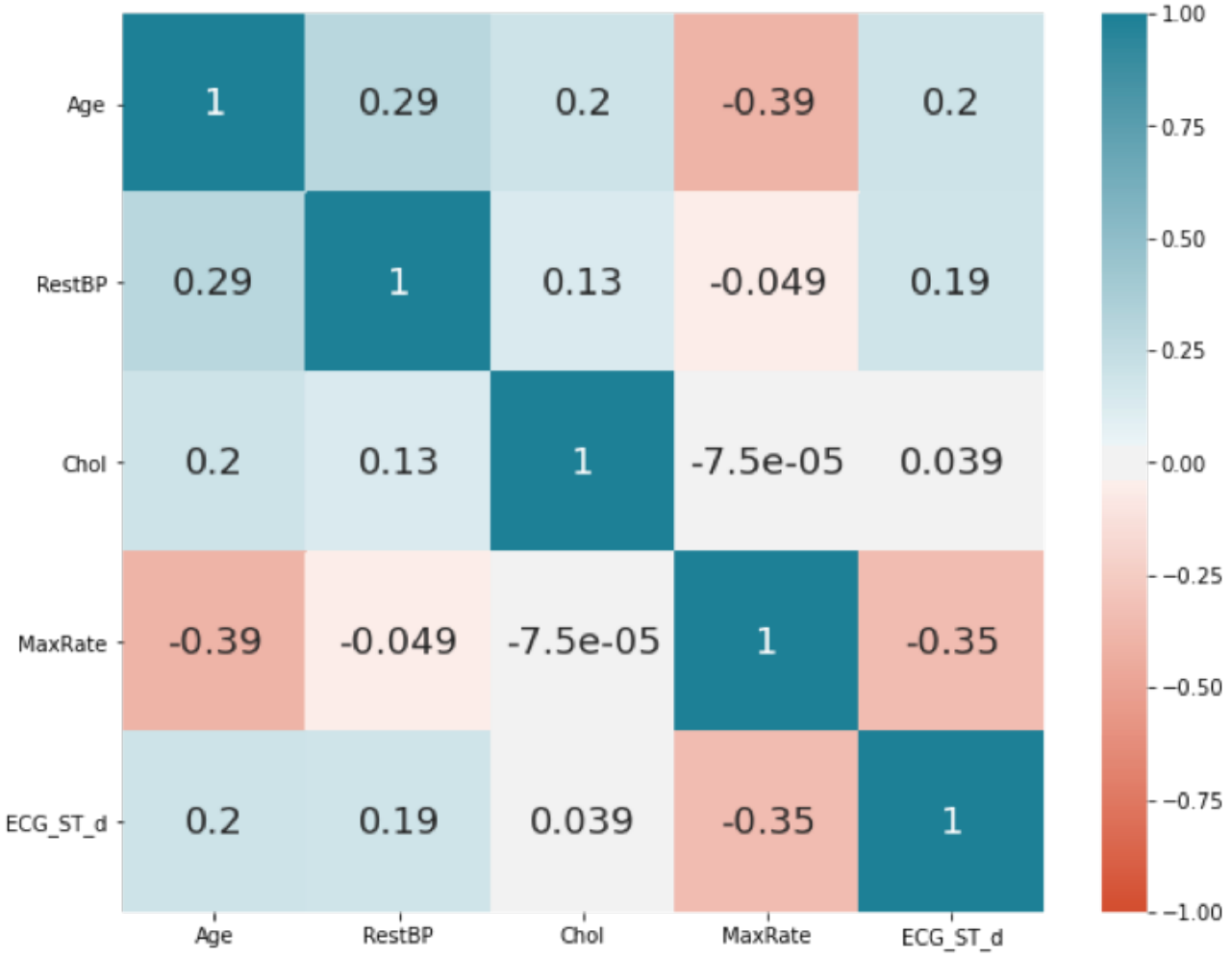
# Correlation Values



# Correlation Matrix (Heart Data)

- Shows correlation of all pairs of continuous variables
- Shade from correlation
- Symmetric
- Perfect correlation on diagonal



# Covariance

# Covariance

- Extends variance (for one variable) to two variables

$$var(X) = E[(x_i - \mu_X)^2]$$

$$cov(X, Y) = E[(x_i - \mu_X)(y_i - \mu_Y)]$$

# Covariance and Correlation

- Covariance can have any value, depending on variance of X and Y
- Correlation is between -1 and +1
  - Covariance ‘normalised’ by standard deviations

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

# Two ‘Issues’ with Correlation

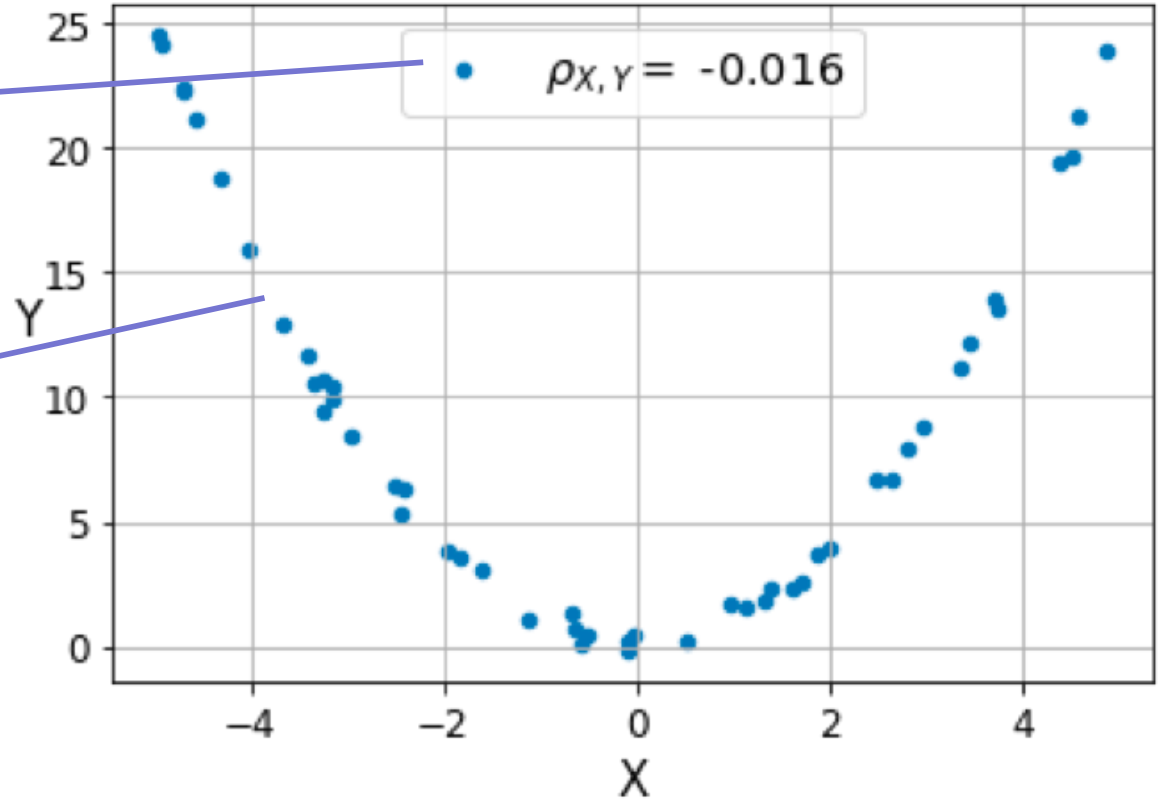
The correlation coefficient is very widely used but there are 2 Issues to note

# Issue 1: Correlation and Independence

- If X, Y independent, then corr(X,Y) close to zero
- However, corr(X,Y) can be close to zero even when X and Y not independent

Correlation is low

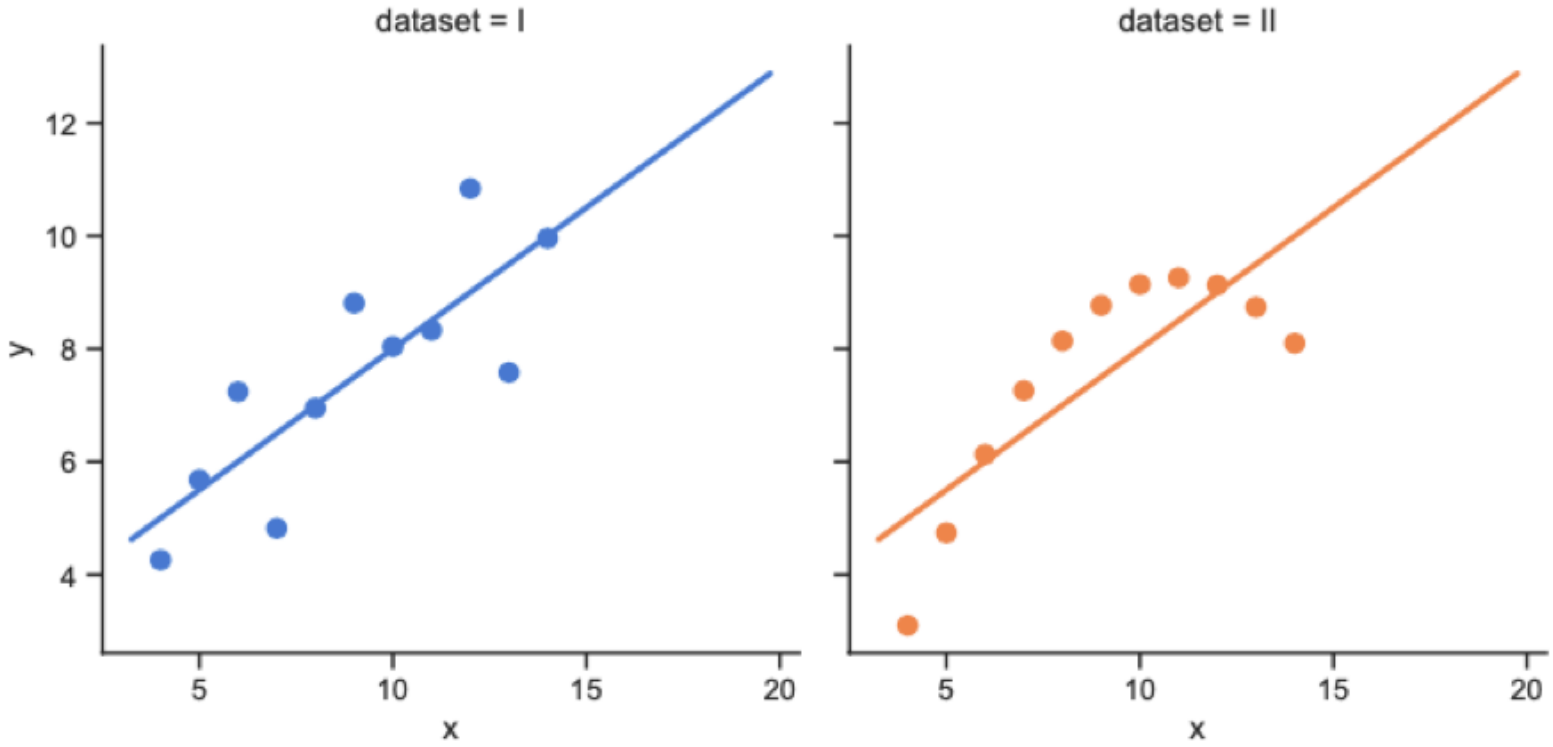
X determines Y (nearly)




Understand this more in ‘regression’ topic

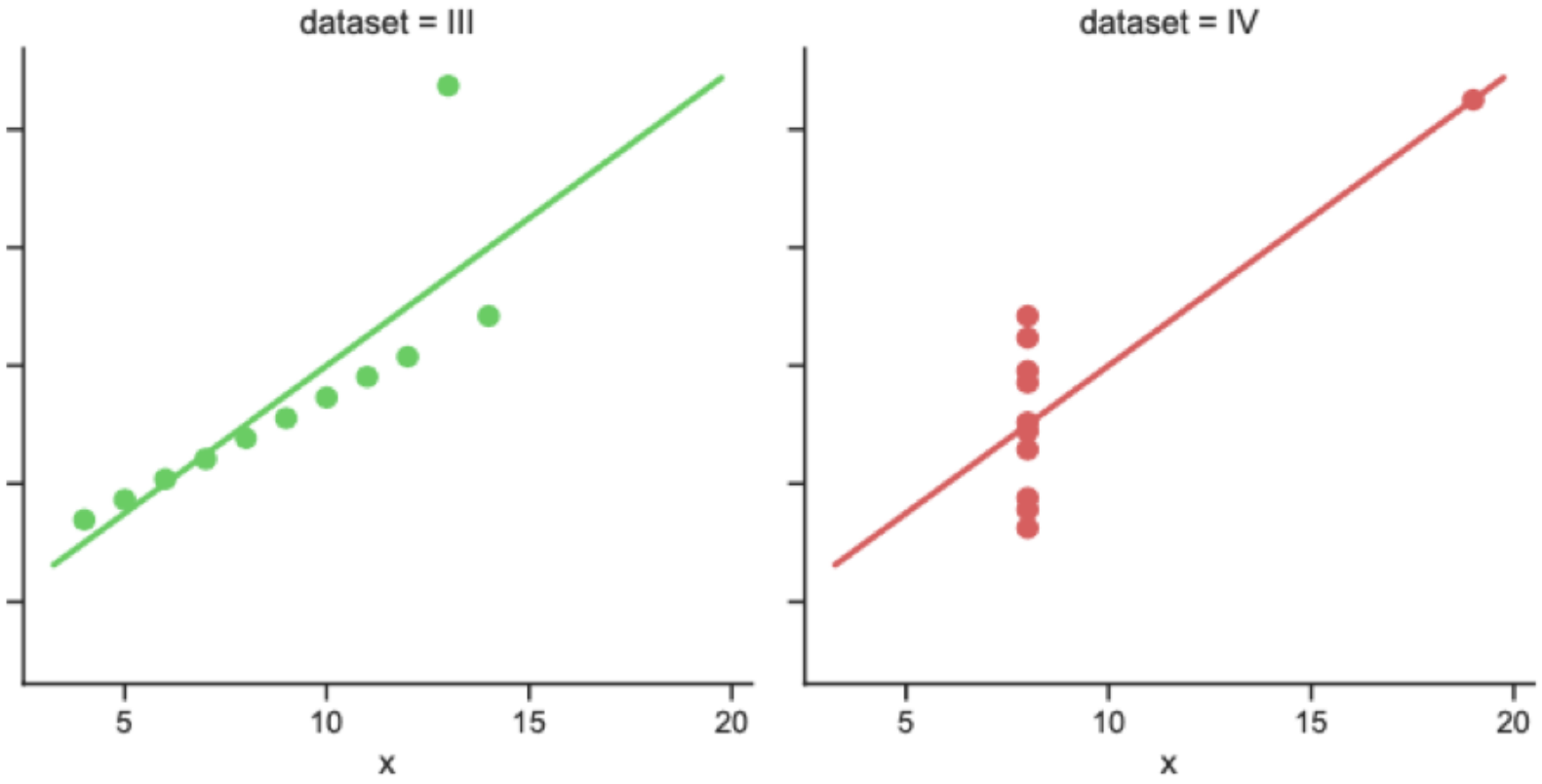
# Issue 1: Correlation and Independence

- All these datasets of (X, Y) have the same
  - Average X and average Y
  - Variance of X and variance of Y
  - Correlation

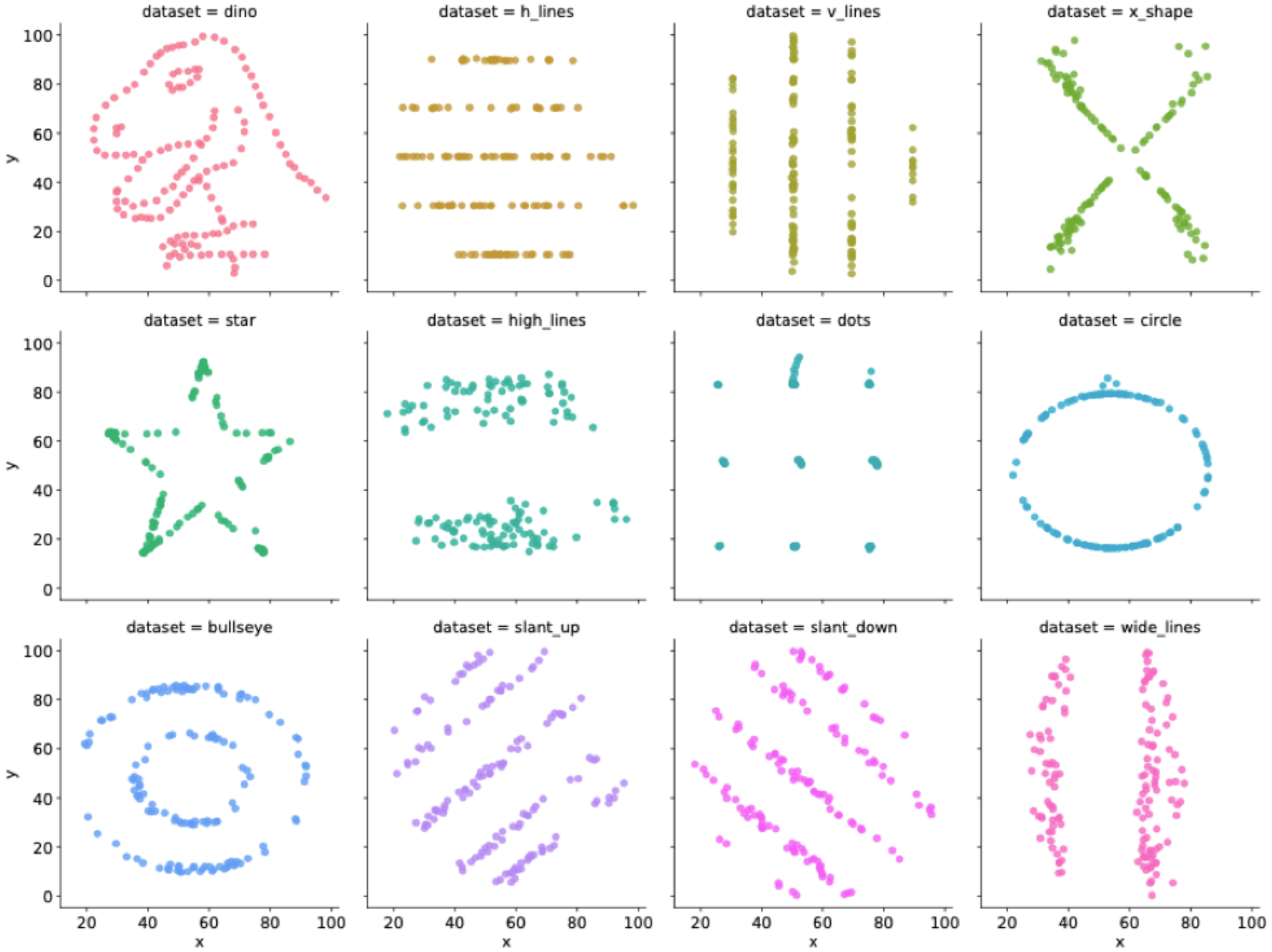


# Issue 1: Correlation and Independence

- All these datasets of (X, Y) have the same
  - Average X and average Y
  - Variance of X and variance of Y
  - Correlation



# Datasaurus: Never trust summary statistics alone; always visualize your data



# Issue 2: Only for Continuous Variables

- Correlation defined using ‘mean’
- Only applies to continuous variables

- If X and Y are categorical variables
  - Independent: knowing value of X does not change distribution of Y values
  - Not independent: distribution of Y values varies depending on the X value

# Quiz 3

# Summary

- Correlation captures idea of ‘variables moving together’
- Relationship between two variables can be shown on a scatter plot
- Correlation coefficient
  - Measures correlation
  - Relates to covariance

- *How to look at relationship between categorical variables?*