Workshop week 7
Correlation

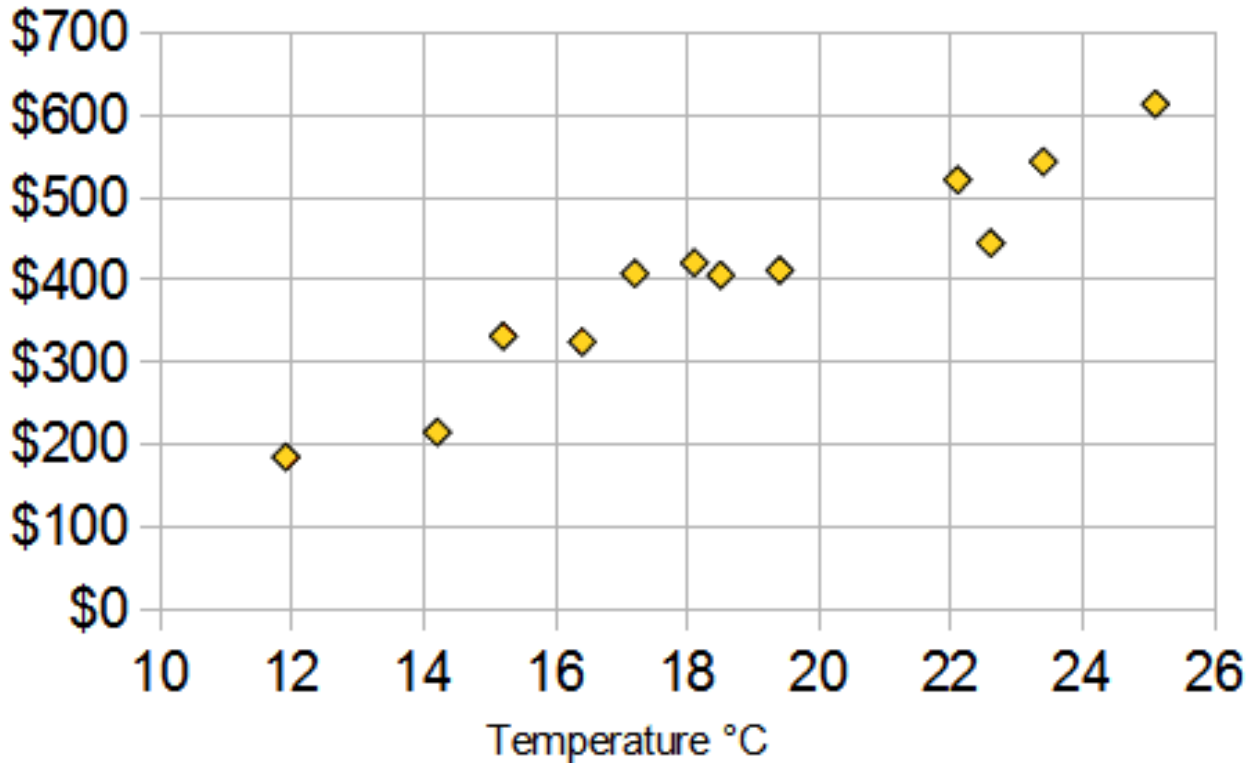
COMP20008 2020 S2

Week 7 Correlation

- Correlation
  - Can hint at potential causal relationships
  - Does not imply causality!
- Pearson Correlation
  - Assess linear relationship
  - Range [-1,1]
- Mutual Information
  - Detect non-linear relationship
  - Need preprocessing (discretise into bins)
  - Entropy: measure uncertainty $H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$
  - Conditional entropy: $H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$

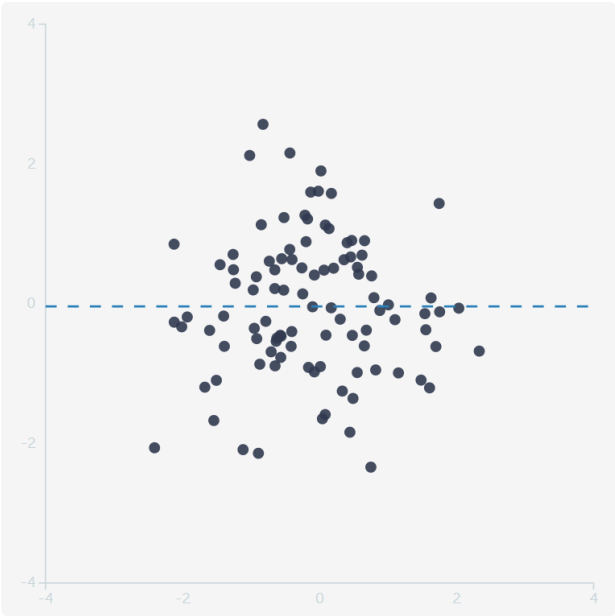
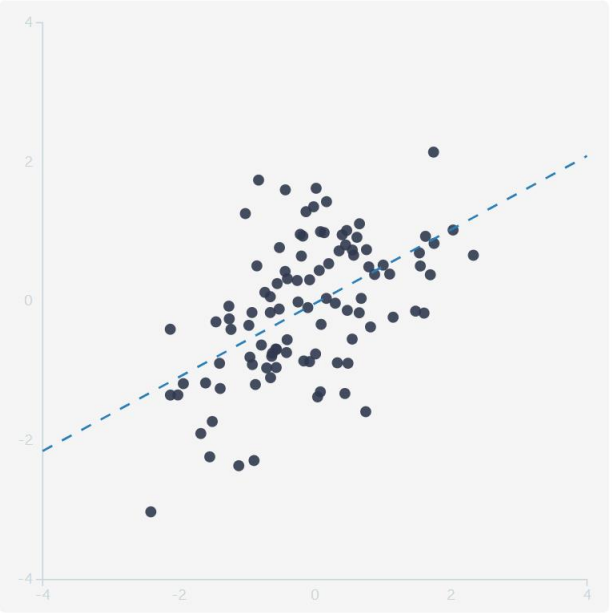
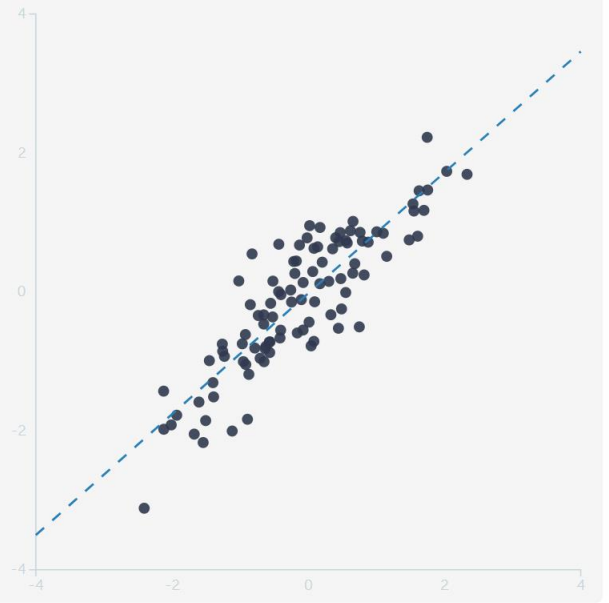
What is Correlation?

- Correlation is used to detect pairs of variables that might have some relationship.
- Visually can be identified via inspecting scatter plots



Remember!

- Correlation does not necessarily imply causality!
- Feature ranking: select the best features for building better predictive models:
  - A good feature to use, is a feature that has high correlation with the outcome one is trying to predict

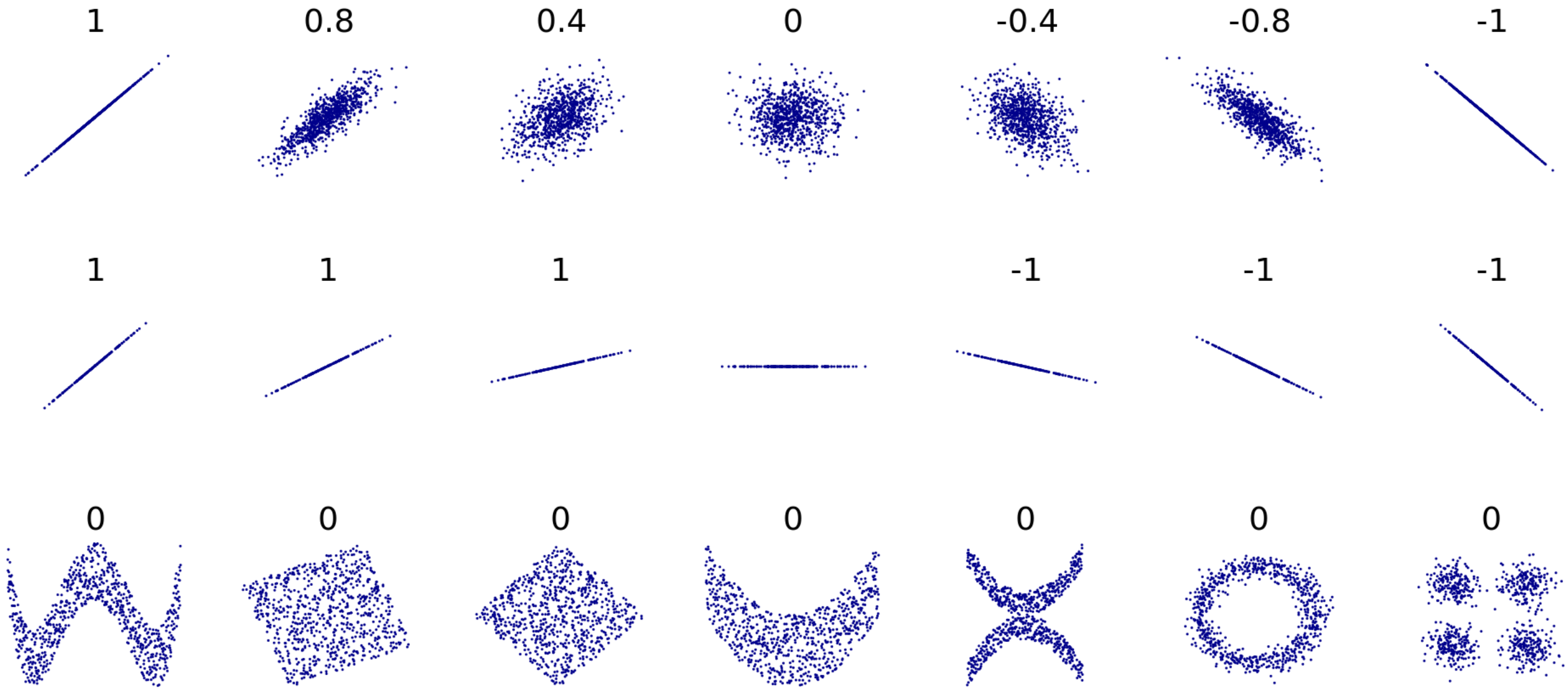




https://rpsychologist.com/d3/correlation/

Assessing linear correlation – Pearson correlation

- Assess how close their scatter plot is to a straight line (a linear relationship)

- Range of r_{xy} lies within [-1,1]:
  - 1 for perfect positive linear correlation
  - -1 for perfect negative linear correlation
  - 0 means no correlation
  - Absolute value |r| indicates strength of linear correlation


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Calculation

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

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

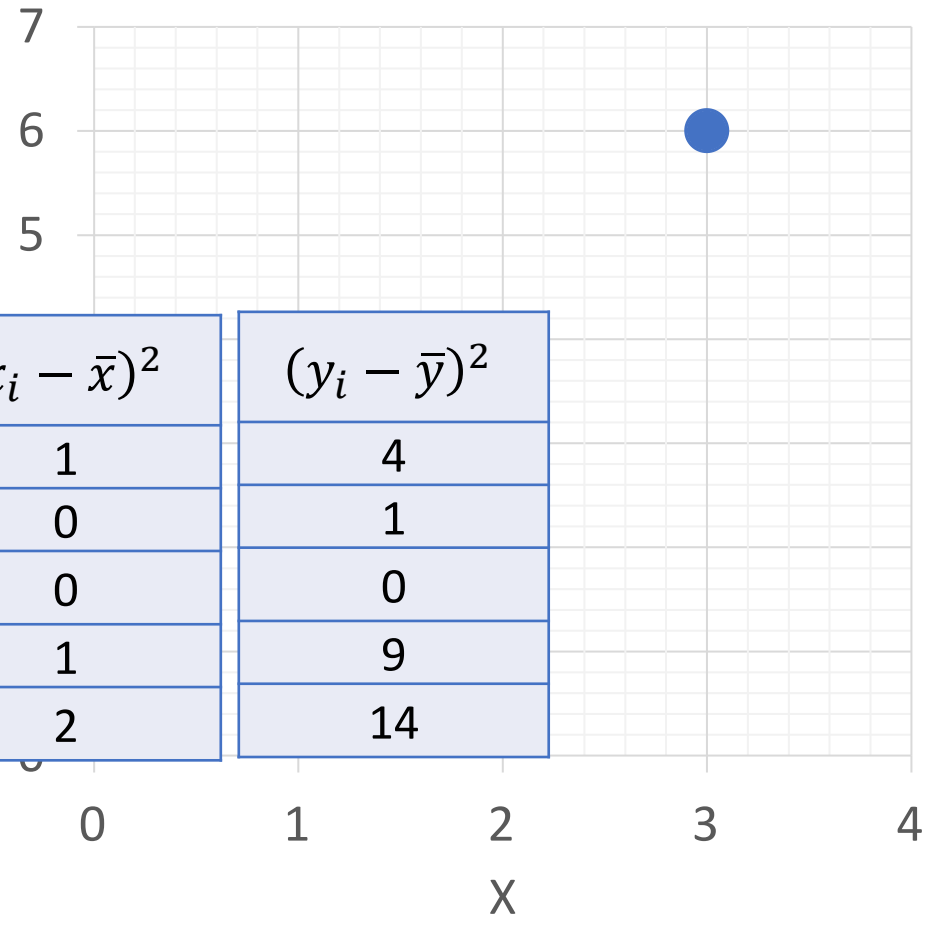
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Micro Example (Warm Up)

x	y
1	1
2	2
2	3
3	6

- Calculate Pearson Correlation Coefficient between X and Y

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



X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	1	-1	-2	2	1	4
2	2	0	-1	0	0	1
2	3	0	0	0	0	0
3	6	1	3	3	1	9
$\bar{x} = 2$	$\bar{y} = 3$	Sum??	Sum??	5	2	14

$$r_{xy} = \frac{5}{\sqrt{2 \times 14}} = 0.9449$$

- Range within [-1,1]

- Scale invariant: r(x,y)= r(x, Ky)
  - Multiplying a feature's values by a constant K makes no difference

- Location invariant: r(x,y)= r(x, K+y)
  - Adding a constant K to one feature's values makes no difference

- Can only detect linear relationships

$$y = a.x + b + noise$$

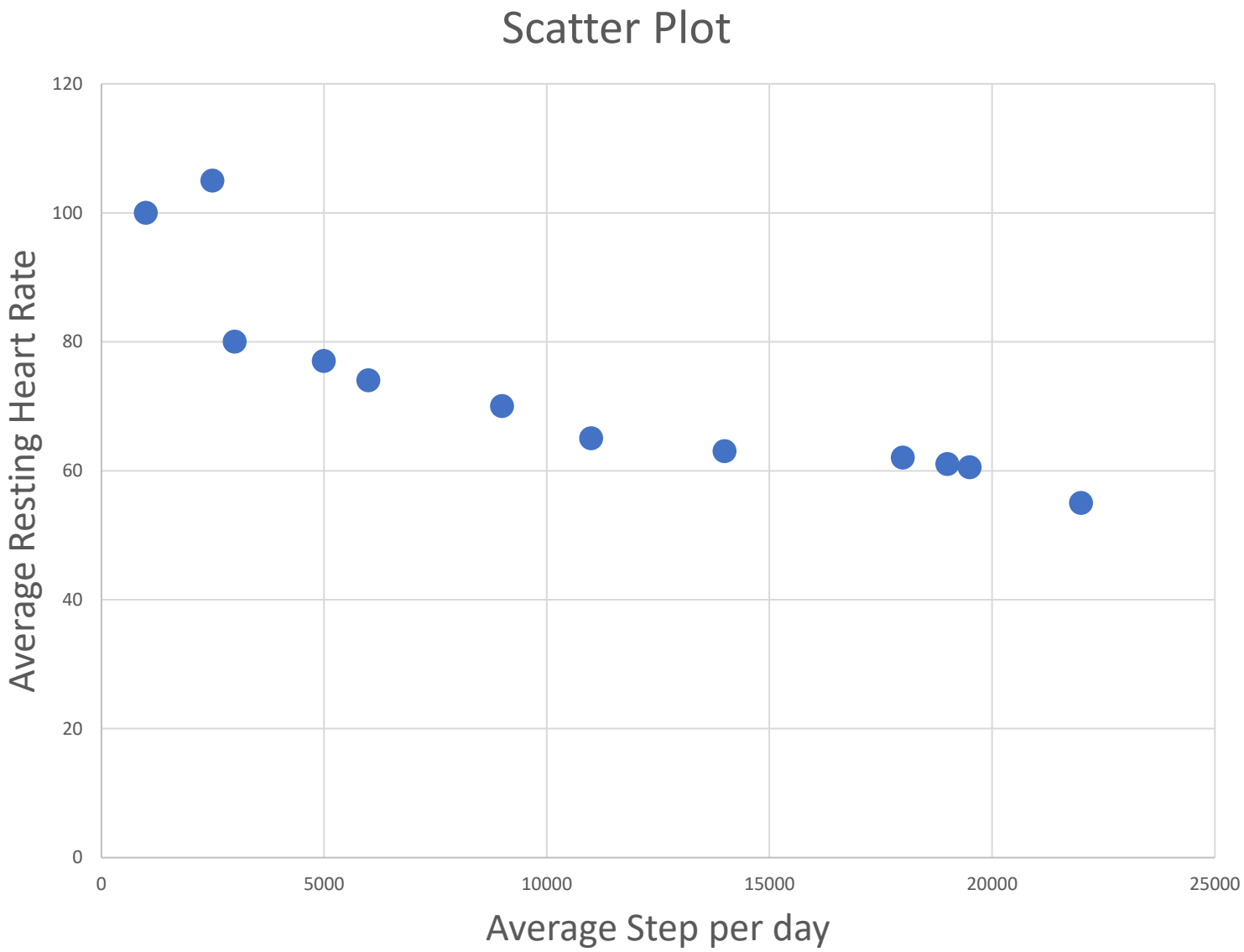
1. Compute the Pearson correlation between Average Steps per day and Average Resting Heart Rate. Show your working. How would you interpret this correlation value?

Person ID	Average Steps per day	Average Resting Heart Rate
1	1000	100
2	2500	105
3	3000	80
4	5000	77
5	6000	74
6	9000	70
7	11000	65
8	14000	63
9	18000	62
10	19000	61
11	19500	60.5
12	22000	55

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

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

r_{xy} = ??


Scatter Plot
Average Resting Heart Rate
120
100
80
60
40
20
0
0
5000
10000
15000
20000
25000
Average Step per day

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right).\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} = \frac{(-1128833.3)}{\sqrt{616166666.7 \text{ x}2736.2}} = -0.86937$$

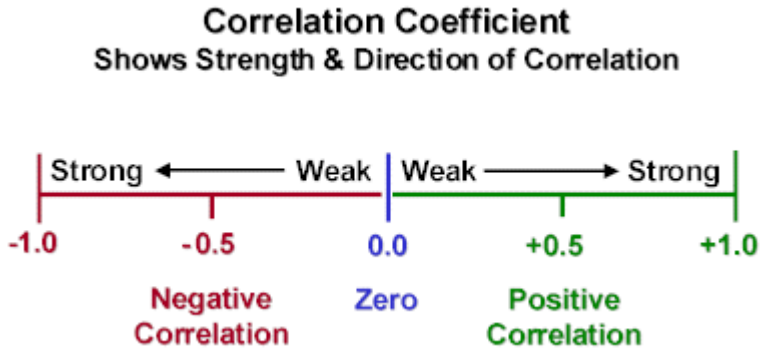
Person ID	Average Steps per day	Average Resting Heart Rate	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	1000	100	-9833.3	27.3	-268368	96694444.44	744.8351
2	2500	105	-8333.3	32.3	-269097	69444444.44	1042.752
3	3000	80	-7833.3	7.3	-57118.1	61361111.11	53.1684
4	5000	77	-5833.3	4.3	-25034.7	34027777.78	18.4184
5	6000	74	-4833.3	1.3	-6243.06	23361111.11	1.668403
6	9000	70	-1833.3	-2.7	4965.278	3361111.111	7.335069
7	11000	65	166.7	-7.7	-1284.72	27777.77778	59.4184
8	14000	63	3166.7	-9.7	-30743.1	10027777.78	94.25174
9	18000	62	7166.7	-10.7	-76743.1	51361111.11	114.6684
10	19000	61	8166.7	-11.7	-95618.1	66694444.44	137.0851
11	19500	60.5	8666.7	-12.2	-105806	75111111.11	149.0434
12	22000	55	11166.7	-17.7	-197743	124694444.4	313.5851
mean	10833.3	72.7	Sum??	Sum??	-1128833.3	616166666.7	2736.2

2. Based on the Pearson correlation value, can one conclude that doing more steps per day will cause one's average resting heart rate to decrease? How else might it be interpreted?

$r_{xy} = -0.86937$

- There is a relationship between the two factors, but can't conclude it is causal.
- Data sample is very small, could be a biased sample.
- Could also be a 3rd factor controlling both (e.g. high blood pressure could cause high heart rate, high blood pressure could also cause a person to be less physically active (and thus take lower steps)

- THM: ***Correlation does not imply Causality***
- Limitation of Pearson Correlation Coefficient

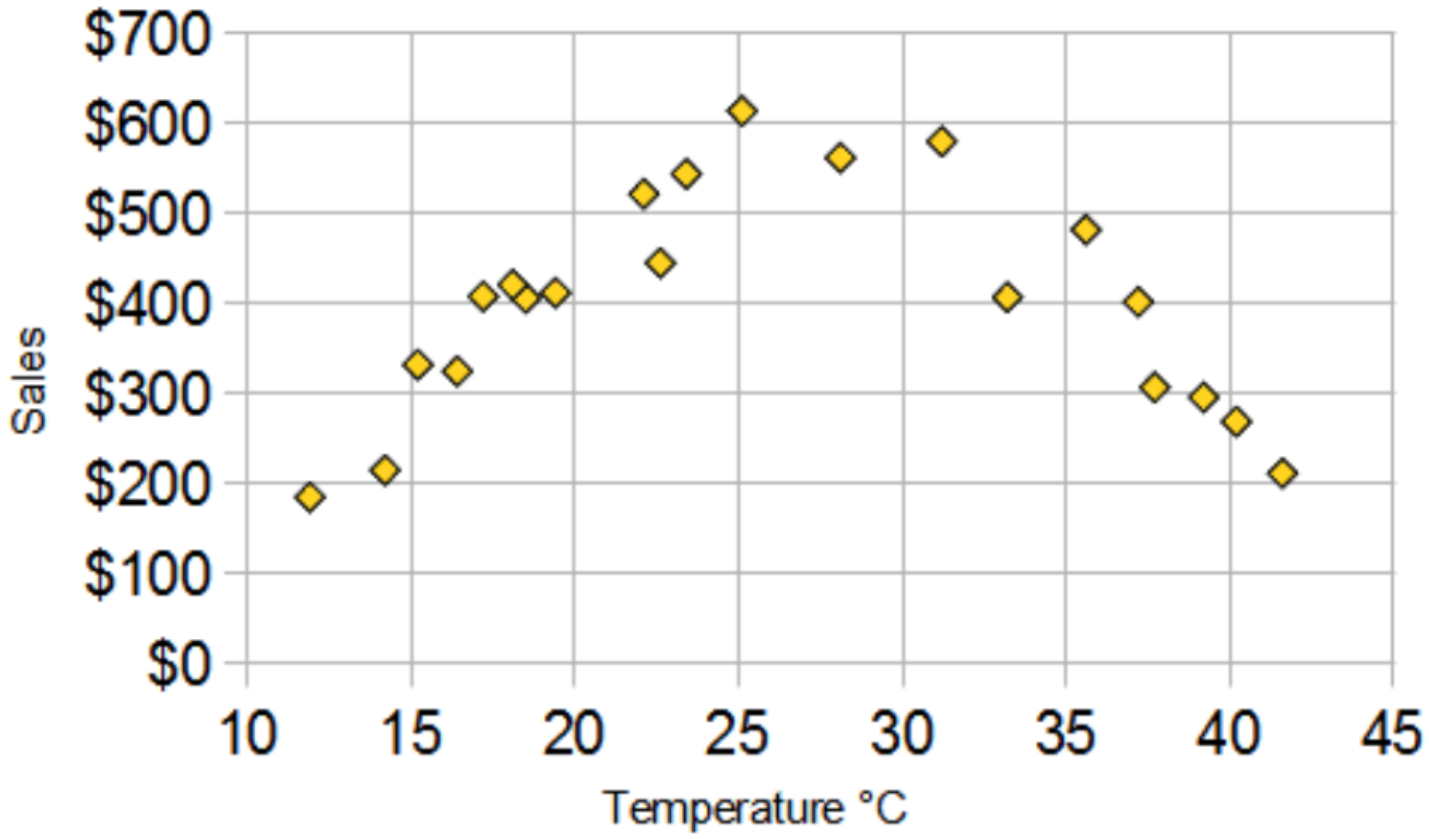


2017 Exam Q3(a)

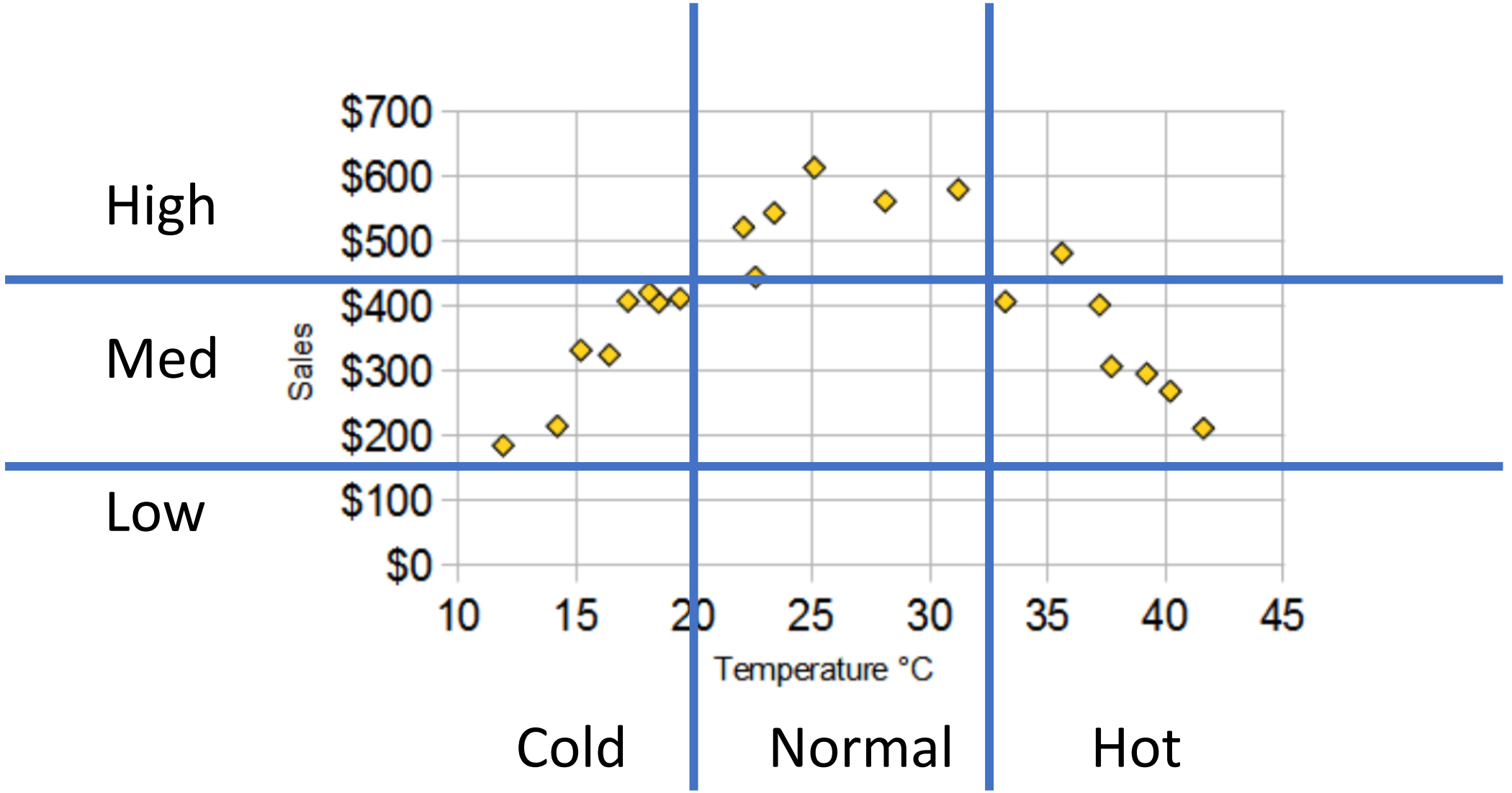
Instance ID	*Predicted class*	*Actual class*
1	X	X
2	X	Y
3	Y	Y
4	X	X
5	X	Y
6	Y	X
7	X	X
8	Y	Y
9	Y	X
10	Y	Y

a) (1 mark) Would Pearson correlation be suitable to compute the correlation between the *Predicted class* and *Actual class*? Why or why not?

How do we measure non-linear correlation



How do we measure non-linear correlation



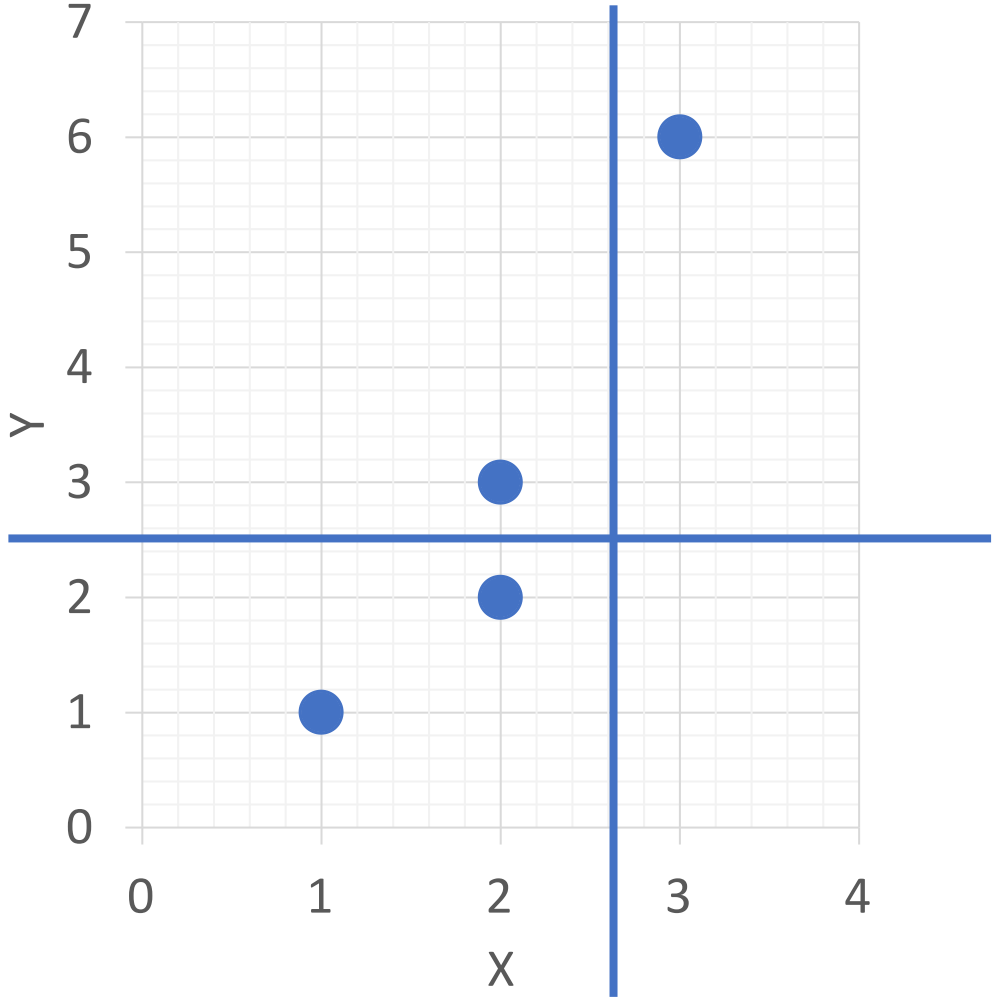
Variable discretization: Techniques

- Domain knowledge
- Equal-length bin
- Equal frequency bin

Micro Example (Warm Up)

- Discretization techniques
  - Manual thresholds (domain knowledge)
  - Equal-width bin
  - Equal-frequency bin
- Discretize X and Y
  - X: 2 bin equal width discretisation
  - Y: 2 bin equal frequency discretisation

x	y
1	1
2	2
2	3
3	6



x	x_d
1	0
2	1
2	1
3	1

y	y_d
1	0
2	0
3	1
6	1

3. Discretise the data as follows: Apply 3 bin equal frequency discretisation to Average Steps per day and 4 bin equal frequency discretisation to Average Resting Heart Rate. Show the values of the discretised features.

- **Discretization techniques:** Manual thresholds (domain knowledge), Equal-width bin and Equal-frequency bin

Column 1 = Sorted	Discrete
1000	1
2500	1
3000	1
5000	1
6000	2
9000	2
11000	2
14000	2
18000	3
19000	3
19500	3
22000	3

Person ID	Average Steps per day	Disc Average Steps per day	Average Resting Heart Rate	Disc Average Resting Heart Rate
1	1000	1	100	4
2	2500	1	105	4
3	3000	1	80	4
4	5000	1	77	3
5	6000	2	74	3
6	9000	2	70	3
7	11000	2	65	2
8	14000	2	63	2
9	18000	3	62	2
10	19000	3	61	1
11	19500	3	60.5	1
12	22000	3	55	1

3. Discretise the data as follows: Apply 3 bin equal frequency discretisation to Average Steps per day and 4 bin equal frequency discretisation to Average Resting Heart Rate. Show the values of the discretised features.

- Discretization techniques: Manual thresholds (domain knowledge), Equal-width bin and Equal-frequency bin

Column 2	Sorted	Discrete
100	55	1
105	60.5	1
80	61	1
77	62	2
74	63	2
70	65	2
65	70	3
63	74	3
62	77	3
61	80	4
60.5	100	4
55	105	4

Person ID	Average Steps per day	Disc Average Steps per day	Average Resting Heart Rate	Disc Average Resting Heart Rate
1	1000	1	100	4
2	2500	1	105	4
3	3000	1	80	4
4	5000	1	77	3
5	6000	2	74	3
6	9000	2	70	3
7	11000	2	65	2
8	14000	2	63	2
9	18000	3	62	2
10	19000	3	61	1
11	19500	3	60.5	1
12	22000	3	55	1

Entropy and Mutual Information

- Entropy
  - Quantify the amount of uncertainty in an entire probability distribution
  - The entropy of a variable is the “amount of information” contained in the variable
  - Describe it in the sense of randomness, surprise
- Conditional entropy
  - H(Y|X) Measures how much information needed to describe outcome Y, given that outcome X is known
- Mutual Information
  - a measure of correlation, the amount of information shared between two variables X and Y
  - MI(X,Y)>=0, large → highly correlated

- Given a feature **X**. Then *H(**X**)* is its entropy. Assuming **X** uses a number of categories(bins)

$$H(\mathbf{X}) = -\sum_{i=1}^{\#bins} p_i \log p_i$$

- p_i: proportion of points in the i-th bin

- May sometimes write p(i) instead of p_i

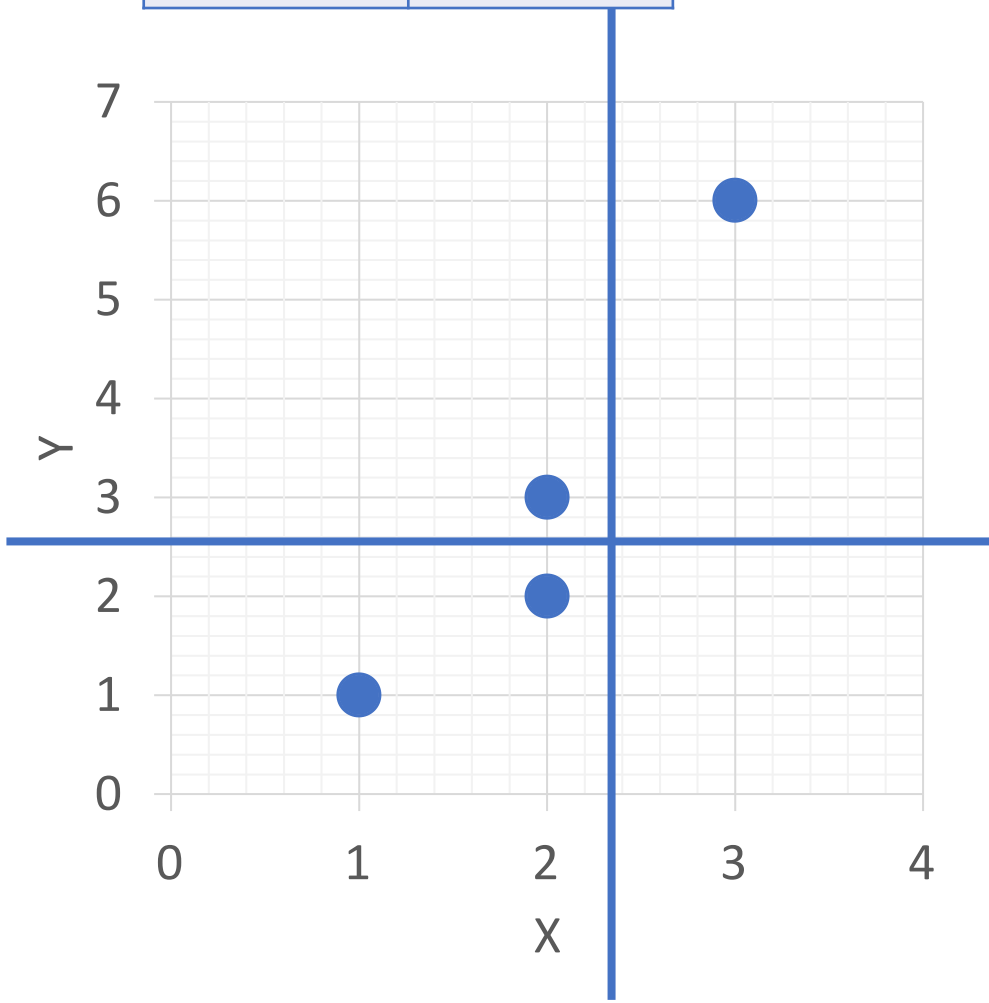
Micro Example (Warm Up)

- Calculate entropies H(X),H(Y)

Entropy:

$$H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$$

x	y
0	0
1	0
1	1
1	1



$$\mathrm{H(X)} = -\sum_{i=1}^{k} p(i) \log p(i) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.5 + 0.3113 = 0.8113$$

$$\mathrm{H(Y)} = -\sum_{i=1}^{k} p(i) \log p(i) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 0.5 + 0.5 = 1$$

Micro Example (Warm Up)

- Calculate conditional entropies H(X|Y),H(Y|X)

x	y
0	0
1	0
1	1
1	1

Conditional Entropy:

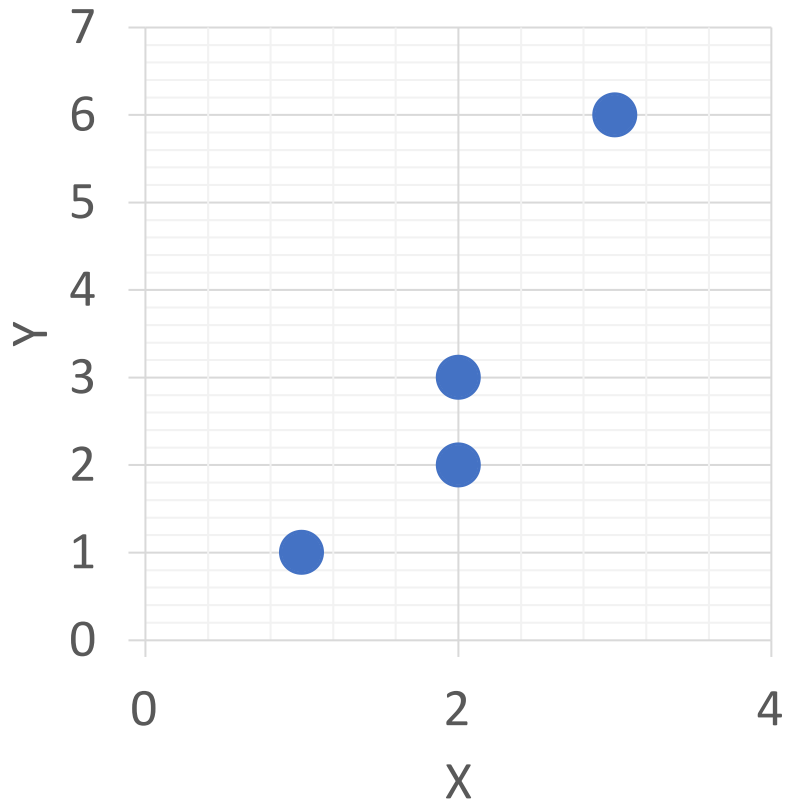
$$H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$$

$$\mathrm{H(X|Y)} = \sum_{y \in Y} p(y)\, H(\mathrm{X|Y} = y) = p(y=0)H(X|Y=0) + p(y=1)H(X|Y=1)$$

$$= \frac{1}{2}H(X|Y=0) + \frac{1}{2}H(X|Y=1) = \frac{1}{2} \times \left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{1}{2} \times (-1\log 1) = 0.5$$

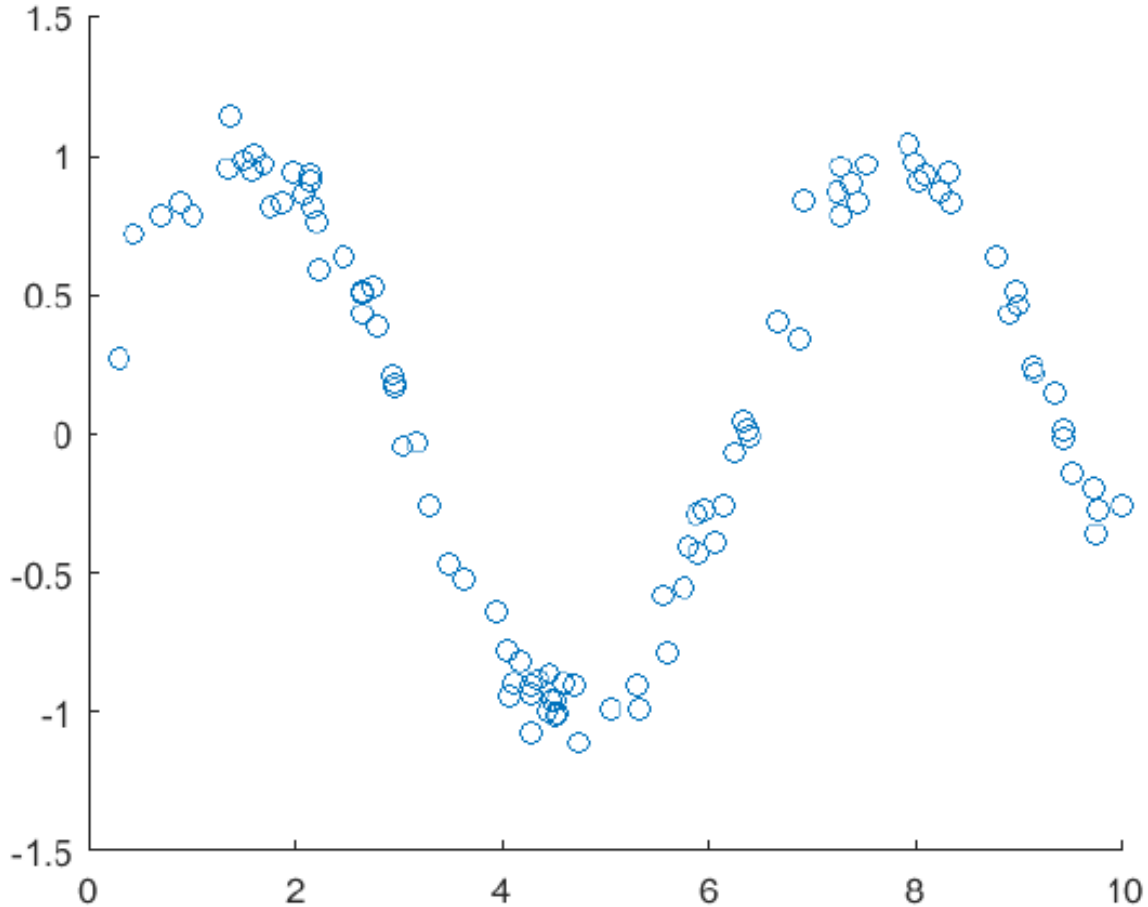
$$\mathrm{H(Y|X)} = \sum_{x \in X} p(x)\, H(\mathrm{Y|X} = X) = p(x=0)H(Y|X=0) + p(x=1)H(Y|X=1)$$

$$= \frac{1}{4}H(Y|X=0) + \frac{3}{4}H(Y|X=1) = \frac{1}{4} \times (-1\log 1) + \frac{3}{4} \times \left(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right) = 0.6887$$

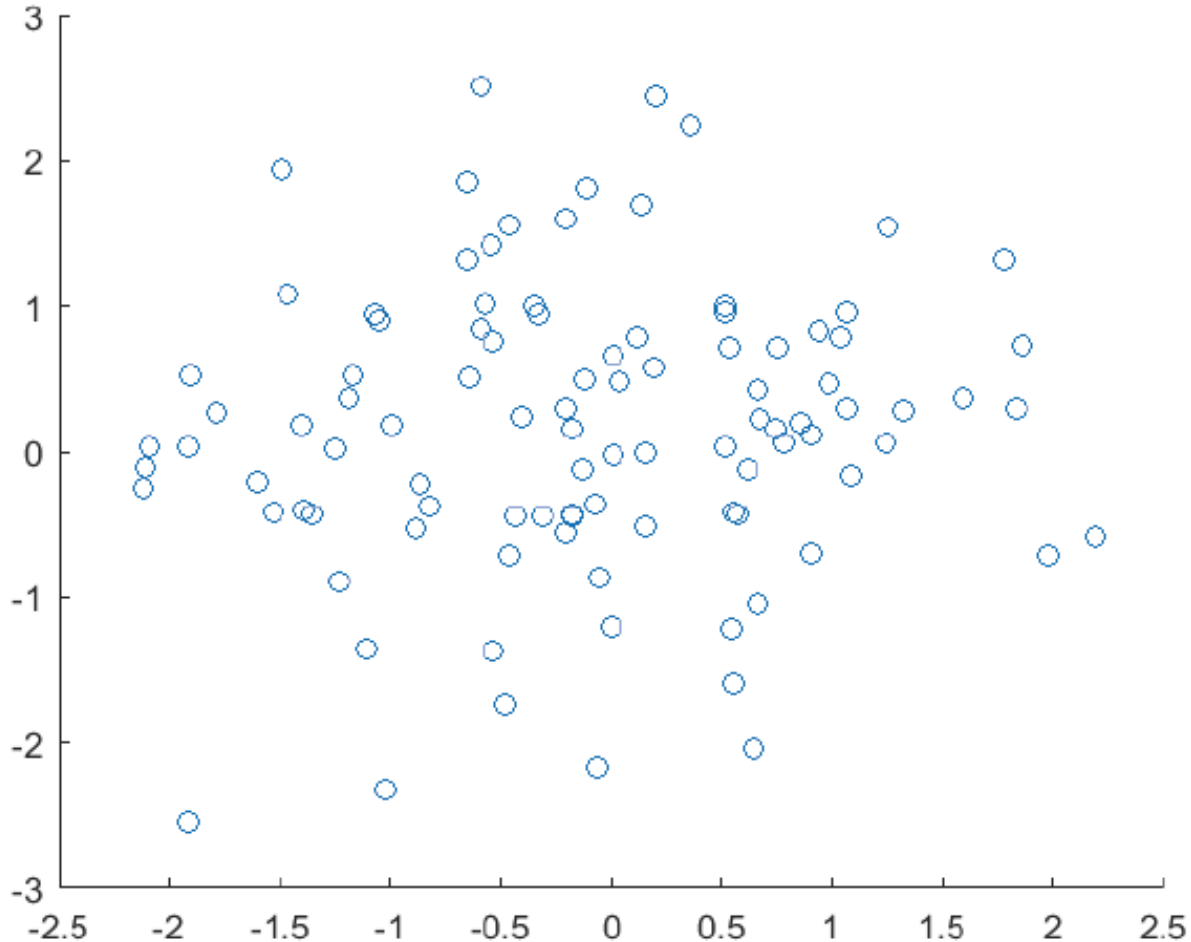


$$MI(X,Y) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$

- Where X and Y are features (columns) in a dataset.
- MI (mutual information) is a measure of correlation
  - the amount of information about X we gain by knowing Y, or
  - The amount of information about Y we gain by knowing X



- Pearson: -0.0864
- NMI: 0.43 (3-bin equal frequency discretization)



- Pearson: 0.08
- NMI: 0.009

Micro Example (Warm Up)

- Calculate mutual information

x	y
1	1
2	2
2	3
3	6

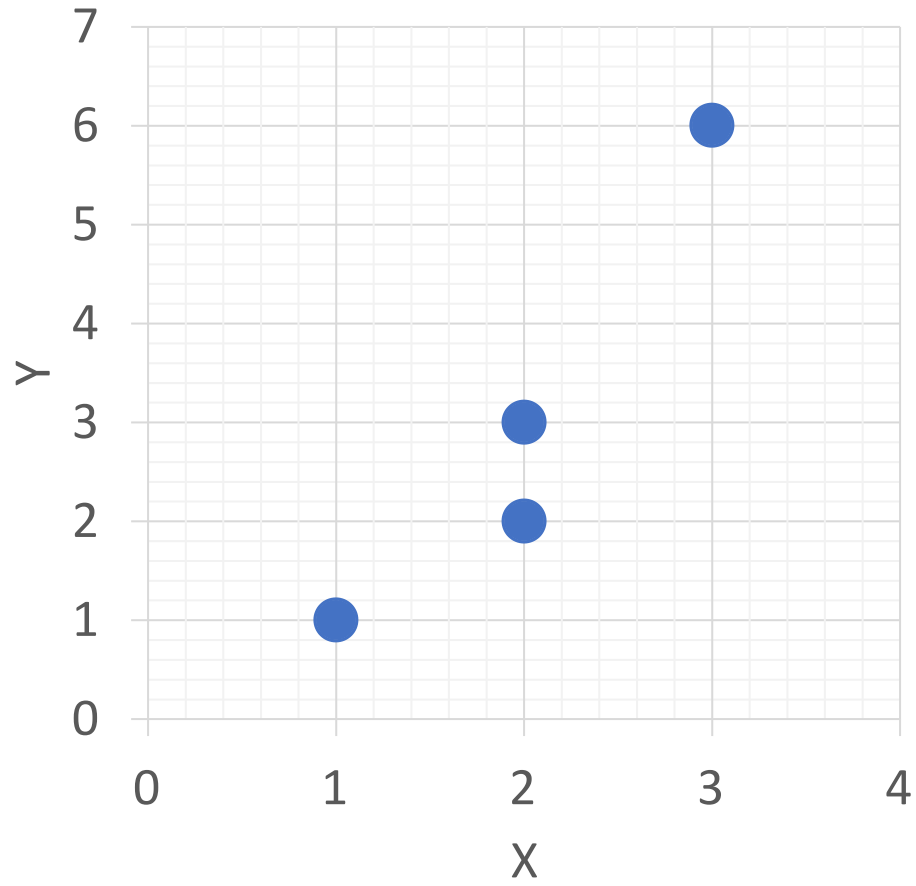
Mutual Information:

$$MI(X,Y) = \mathrm{H(X)} - \mathrm{H}(X|Y)$$

$$MI(X,Y) = \mathrm{H(Y)} - \mathrm{H}(Y|X)$$

$$NMI(X,Y) = \frac{MI(X,Y)}{\min(\mathrm{H(X)},\mathrm{H(Y)})}$$

H(X) = 0.8113
H(Y) = 1
H(X|Y) = 0.5
H(Y|X) = 0.6887



- $MI(X,Y) = \mathrm{H(X)} - \mathrm{H}(X|Y) = 0.8113 - 0.5 = 0.3113$
- $MI(Y,X) = \mathrm{H(Y)} - \mathrm{H}(Y|X) = 1 - 0.6887 = 0.3113$
- $NMI(X,Y) = \frac{MI(X,Y)}{\min(\mathrm{H(X)},\mathrm{H(Y)})} = \frac{0.3113}{0.8113} = 0.3837$

4. Using the discretised features, compute the entropies:

- H(Average Steps per day)
- H(Average Resting Heart Rate)
- H(Average steps per day | Average Resting Heart Rate)
- H(Average Resting Heart Rate | Average Steps per day).

Entropy:

$$H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$$

Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$$

Person ID	Disc Average Steps per day	Disc Average Resting Heart Rate
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

4. Using the discretised features, compute the entropies:

1. H(Average Steps per day)

- $= -\sum_{i=1}^{k} p(i) \log p(i)$
- $= -\left(\frac{4}{12}\log\frac{4}{12}\right) - \left(\frac{4}{12}\log\frac{4}{12}\right) - \left(\frac{4}{12}\log\frac{4}{12}\right)$
- $= -3\left(\frac{4}{12}\log\frac{4}{12}\right)$
- $= -3\left(\frac{1}{3} * -1.585\right) = 1.585$

Person ID	Disc Average Steps per day	Disc Average Resting Heart Rate
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

Entropy:

$$H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$$

Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$$

4. Using the discretised features, compute the entropies:

2. H(Average Resting Heart Rate)

- $= -\sum_{i=1}^{k} p(i) \log p(i)$
- $= -\left(\frac{3}{12}\log\frac{3}{12}\right) - \left(\frac{3}{12}\log\frac{3}{12}\right) - \left(\frac{3}{12}\log\frac{3}{12}\right) - \left(\frac{3}{12}\log\frac{3}{12}\right)$
- $= -4\left(\frac{3}{12}\log\frac{3}{12}\right)$
- $= -4\left(\frac{1}{4} * -2\right) = 2$

Person ID	Disc Average Steps per day	Disc Average Resting Heart Rate
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

Entropy:

$$H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$$

Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$$

4. Using the discretised features, compute the entropies:

3. H(Average Steps per day | Average Resting Heart Rate)→ H(S|R)

- $= \sum_{r \in R} p(r)\, H(S|R = r)$
- $= p(R=4)H(S|R=4) + p(R=3)H(S|R=3) + p(R=2)\, H(S|R=2) + p(R=1)\, H(S|R=1)$
- $= \frac{3}{12}H(S|R=4) + \frac{3}{12}H(S|R=3) + \frac{3}{12}H(S|R=2) + \frac{3}{12}H(S|R=1)$
- $H(S|R=4) = -1 \log 1 = 0$
- $H(S|R=3) = -(\frac{1}{3}\log\frac{1}{3}) - (\frac{2}{3}\log\frac{2}{3}) = .918$
- $H(S|R=2) = -(\frac{2}{3}\log\frac{2}{3}) - (\frac{1}{3}\log\frac{1}{3}) = .918$
- $H(S|R=1) = -1 \log 1 = 0$
- $= .25\ (0 + 0 + .918 + .918) = 0.459$

ID	S	R=4
1	1	4
2	1	4
3	1	4

ID	S	R=2
7	2	2
8	2	2
9	3	2

ID	S	R=3
4	1	3
5	2	3
6	2	3

ID	S	R=1
10	3	1
11	3	1
12	3	1

Person ID	Disc Average Steps per day (S)	Disc Average Resting Heart Rate (R)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

Entropy:

$$H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$$

Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$$

4. Using the discretised features, compute the entropies:

4. H(Average Resting Heart Rate | Average Steps per day) → H(R|S)

- $= \sum_{s \in S} p(s)\, H(R|S = s)$
- $= p(S=1)H(R|S=1) + p(S=2)H(R|S=2) + p(S=3)H(R|S=3)$
- $= \frac{4}{12}H(R|S=1) + \frac{4}{12}H(R|S=2) + \frac{4}{12}H(R|S=3)$
- $H(R|S=1) = -(.75 \log .75) - (.25 \log .25) = 0.311+0.5$
- $H(R|S=2) = -(.5 \log .5) - (.5 \log .5) = .5 + .5 = 1$
- $H(R|S=3) = -(.25 \log .25) - (.75 \log .75) = 0.5 + 0.311$
- $= \frac{1}{3}(1 + .811 + .811) = 0.874$

ID	S=1	R
1	1	4
2	1	4
3	1	4
4	1	3

ID	S=2	R
5	2	3
6	2	3
7	2	2
8	2	2

ID	S=3	R
9	3	2
10	3	1
11	3	1
12	3	1

Person ID	Disc Average Steps per day (S)	Disc Average Resting Heart Rate (R)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

Entropy:

$$H(p) = -\sum_{i=1}^{k} p(i) \log p(i)$$

Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} p(x)\, H(Y|X = x)$$

5. Using the above information, compute the mutual information between Average Steps per day and Average Resting Heart Rate.

- H(Average Steps per day) = H(S) = 1.585
- H(Average Resting Heart Rate) = H(R) = 2
- H(Average steps per day | Average Resting Heart Rate) = $\mathrm{H}(S|R) = 0.459$
- H(Average Resting Heart Rate | Average Steps per day) = $\mathrm{H}(R|S) = 0.874$

- $MI(R,S) = \mathrm{H}(\mathrm{R}) - \mathrm{H}(R|S) = 2 \ - 0.874 = 1.126$
- $MI(R,S) = \mathrm{H}(\mathrm{S}) - \mathrm{H}(S|R) = 1.585 - 0.459 = 1.126$

Mutual Information:

$$MI(R,S) = \mathrm{H}(\mathrm{R}) - \mathrm{H}(R|S)$$

$$MI(R,S) = \mathrm{H}(\mathrm{S}) - \mathrm{H}(S|R)$$

$$NMI(R,S) = \frac{MI(R,S)}{\min(\mathrm{H}(\mathrm{S}), \mathrm{H}(\mathrm{R}))}$$

- Advantage
  - Can detect both linear and non linear dependencies (unlike Pearson)
  - Applicable and very effective for use with discrete features (unlike Pearson correlation)

- Disadvantage
  - If feature is continuous, it first must be discretised to compute mutual information. This involves making choices about what bins to use.
    - This may not be obvious. Different bin choices will lead to different estimations of mutual information