# Exercise 1 Solution

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

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

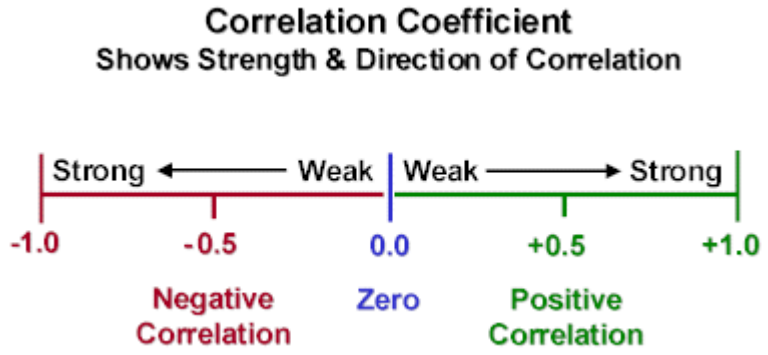
| Person ID | Average Steps per day | Average Resting Heart Rate | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1000 | 100 | -9833.3 | 27.3 | -268368 | 96694444.44 | 744.8351 |
| 2 | 2500 | 105 | -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 |

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



# Exercise 2 Solution

**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.**

| Column 1 = Sorted |
|---|
| 1000 |
| 2500 |
| 3000 |
| 5000 |
| 6000 |
| 9000 |
| 11000 |
| 14000 |
| 18000 |
| 19000 |
| 19500 |
| 22000 |

| Discrete |
|---|
| 1 |
| 1 |
| 1 |
| 1 |
| 2 |
| 2 |
| 2 |
| 2 |
| 3 |
| 3 |
| 3 |
| 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 |

| 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 |


# Exercise 4 Solution

# 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:

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

# Exercise 5 Solution

**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:

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

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

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