

Workshop Week 7 – PART A

Your Feedback

- Workshops issues
 - Tutors not explaining code enough
 - Insufficient engagement & interaction
 - Tutors too fast
 - Tutors can't cover all materials
- What else?
- Your turn to give us solutions!!

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

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

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

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

$$r_{xy} = ??$$

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

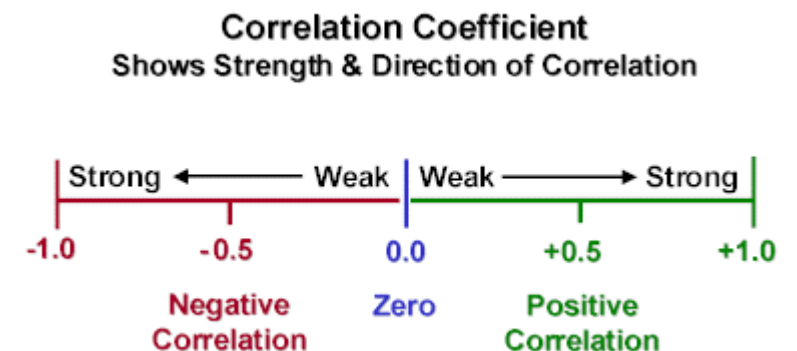
Person ID	Average Steps per day	Average Resting Heart Rate	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
			-9833.3	27.3	-268368	96694444.44	744.8351
1	1000	100	-8333.3	32.3	-269097	69444444.44	1042.752
2	2500	105	-7833.3	7.3	-57118.1	61361111.11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



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

4. Using the discretised features, compute the entropies:

- $H(\text{Average Steps per day})$
- $H(\text{Average Resting Heart Rate})$
- $H(\text{Average steps per day} \mid \text{Average Resting Heart Rate})$
- $H(\text{Average Resting Heart Rate} \mid \text{Average Steps per day})$.

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. $H(\text{Average Steps per day})$

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

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. $H(\text{Average Resting Heart Rate})$

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

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. $H(\text{Average Steps per day} \mid \text{Average Resting Heart Rate}) \rightarrow H(S \mid R)$

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

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:

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

4. Using the discretised features, compute the entropies:

4. $H(\text{Average Resting Heart Rate} \mid \text{Average Steps per day}) \rightarrow H(R \mid S)$

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

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:

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

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

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

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

Mutual Information:

$$MI(R, S) = H(R) - H(R|S)$$

$$MI(R, S) = H(S) - H(S|R)$$

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

Phase 3: Question 1

- Who would be interested in an answer to this question and why?
 - Log in to the Aurin portal <https://aurin.org.au>
 - Select Victoria as your region of interest.
 - Add the dataset “2015 Local Government Area (LGA) Statistical Proles”. You should select all the attributes to include. This dataset includes information about number of people reporting high blood pressure across different regions in the State. We will use this as a measure of people's health.
 - Add the dataset “LGA Visit to green space (once per week)”. You should select all the attributes to include. This dataset contains information about number of people who visit local green space each week, across different regions in the State.
 - Download each of these datasets as a CSV file.

Phase 3: Question 2 and 3

- Q2: What feature would you use to join these datasets together? How would you approach this in Python?
- Q3: The following are examples of possible initial investigations you could perform for phase 2A on this dataset. Rank them in terms of priority. How many would you have time to do in 6-10 hours wrangling?
 - Scatter plot of blood pressure versus number of visits to green space, across different LGAs.
 - Boxplots for blood pressure feature and number of visits to green space feature.
 - Outlier detection for blood pressure feature and number of visits to green space feature.
 - Clustering
 - Missing value imputation
 - Correlation between blood pressure and number of visits to green space
 - ... anything else?

Phase 3: Question 4 and 5

- Q4: Suppose the correlation between blood pressure versus number of visits to green space turns out to be small. What other features could you examine from the “2015 Local Government Area (LGA) Statistical Proles” dataset to serve as an indicator of health?
- Q5: What challenges do you think might arise in studying this research question for Phase 2A and Phase 3?