DATA - 240 Data mining - Fall 2024

HW - 1

Prayag Nikul Purani (017416737)

Question 1

a)

data set 1

x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x}) * (y-\bar{y})$
10	8.04	1	0.290	0.539
8	6.95	1	0.303	0.550
13	7.58	16	0.006	0.316
9	8.81	0	1.714	0.000
11	8.33	4	0.687	1.658
14	9.96	25	6.047	12.295
6	7.24	9	0.068	0.783
4	4.26	25	10.503	16.204
12	10.84	9	11.149	10.017
7	4.82	4	7.187	0.361
5	5.68.	16	3.315	7.283
$\Sigma\bar{x} = 9$	$\Sigma\bar{y} = 7.5$	$\Sigma(x-\bar{x})^2 = 110$	$\Sigma(y-\bar{y})^2 = 41.272$	$\Sigma(x-\bar{x})^2 * (y-\bar{y}^2) = 55.010$

$\bar{x} = 9, \quad \bar{y} = 7.5 \qquad n = 11$

$$var(x) = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{110}{10} = 11$$

$$std(x) = \sqrt{\sigma} = \sqrt{11} = 3.3166$$

$$var(y) = \frac{\Sigma(y_i - \bar{y})^2}{n-1} = \frac{41.272}{10} = 4.1272$$

$$std(y) = \sqrt{\sigma} = \sqrt{4.1272} = 2.0315$$

$$\text{correlation}(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

$$= \frac{55.01}{\sqrt{110 \times 41.272}}$$

$$= \underline{0.8164}$$

dataset 2

x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x}) * (y-\bar{y})$
10	9.14	1	2.686	1.639
8	8.14	1	0.408	-0.639
13	8.74	16	1.535	4.956
9	8.77	0	1.610	0.00
11	9.26	4	3.094	3.518
14	8.10	25	0.358	2.995
6	6.13	9	1.879	4.112
4	3.10	25	19.368	22.004
12	9.13	9	2.653	4.887
7	7.26	4	0.058	0.488
5	4.74	16	7.622	11.043
$\bar{x} = 9$	$\bar{y} = 7.5$	$\Sigma(x-\bar{x})^2 = 110$	$\Sigma(y-\bar{y})^2 = 41.276$	$\Sigma(x-\bar{x}) * (y-\bar{y}) = 54.999$

$\bar{x} = 9$; $\bar{y} = 7.5$; $n = 11$

$$\text{var}(x) = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{110}{10} = 11$$

$$\text{std}(x) = \sqrt{11} = 3.3166$$

$$\text{var}(y) = \frac{\Sigma(y-\bar{y})^2}{n-1} = \frac{41.276}{10} = 4.1276$$

$$\text{std}(y) = \sqrt{4.1276} = 2.0315$$

$$correlation(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

$$= \frac{54.999}{\sqrt{110 \times 41.276}} = 0.8162$$

data set 3

x	y	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})^* (y-\bar{y})^*$
10	7.46	1	0.001	-0.04
8	6.77	1	0.533	0.73
13	12.74	16	27.457	20.96
9	7.11	0	0.152	0.00
11	7.81	4	0.096	0.62
14	8.84	25	1.796	6.70
6	6.08	9	2.016	4.26
4	5.39	25	4.452	10.55
12	8.15	9	0.422	1.95
7	6.42	4	1.166	2.16
5	5.73	16	3.133	7.08
$\Sigma\bar{x}=9$	$\bar{y}=7.5$	$\Sigma(x-\bar{x}) = 110$	$\Sigma(y-\bar{y})^2 = 41.226$	$\Sigma(x-\bar{x})^*(y-\bar{y}) = 54.97$

$\bar{x} = 9$, $\bar{y} = 7.5$

$$var(x) = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{110}{10} = 11$$

$$std(x) = \sqrt{\sigma} = \sqrt{11} = 3.3166$$

$$var(y) = \frac{\Sigma(y_i-\bar{y})}{n-1} = \frac{41.226}{10} = 4.1226$$

$$std(y) = \sqrt{\sigma} = \sqrt{4.1226} = 2.0304$$

$$\text{correlation}(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

$$= \frac{54.97}{\sqrt{110 \times 41.226}} = 0.8162.$$

	$\bar{x}$	$\bar{y}$	var(x)	var(y)	$\sigma(x)$	$\sigma(y)$	cor(x,y)
data 1	9	7.5	11	4.1272	3.3166	2.0315	0.816
data 2	9	7.5	11	4.1276	3.3166	2.0315	0.816
data 3	9	7.5	11	4.1226	3.3166	2.0305	0.816

So, from the above part we can see that all the value are nearly same so lets calculate median of IQR and then find Outliers.

Dataset 1

y(sorted) = 4.26, 4.82, 5.68, 6.95, 7.24, 7.58, 8.04, 8.33, 8.81, 9.96, 10.84

medium = 7.58 (Q_2)

$$Q_1 = \frac{5.68+6.95}{2} = 6.315$$

$$Q_2 = \frac{8.33+8.81}{2} = 8.57$$

$$IQR = Q_3 - Q_1$$

$$= 8.57 - 6.315 = 2.255$$

lower bound = $Q_1 - 1.5 \times IQR$
$= 6.315 - 1.5 \times 2.255$
$= 2.9325$

upper bound = $Q_3 + 1.5 IQR$
$= 8.57 + 1.5 \times 2.255$
$= 11.9525$

$[2.9325, 11.9525]$

$\Rightarrow$ So there are no outliers for this dataset.

Dataset 2

y(sorted) = 3.1, 4.74, 6.13, 7.26, 8.1, 8.14, 8.74, 8.77,
9.13, 9.14, 9.26.

median (Q_2) = 8.14

$Q_1 = \frac{6.13 + 7.26}{2} = 6.695$

$Q_3 = \frac{8.77 + 9.13}{2} = 8.95$

$IQR = Q_3 - Q_1$
$= 2.255$

Lower bound = $Q_1 - 1.5 \times IQR$
$= 6.695 - 1.5 \times 2.255$
$= 3.3125$

upper bound = $Q_3 + 1.5 \times IQR$
$= 8.95 + 1.5 \times 2.255$
$= 12.3325$

$[3.3125, 12.3325]$

So there is one outlier which is 3.1

Dataset 3

y(sorted) = 5.39, 5.73, 6.08, 6.42, 6.77, 7.11, 7.46, 7.81, 8.15
8.84, 12.74

median (Q_2) = 7.11

$Q_1 = \frac{6.08 + 6.42}{2}$

$= 6.25$

$Q_3 = \frac{7.81 + 8.15}{2}$

$= 7.98$

$IQR = Q_3 - Q_1$

$= 1.73$

Lower bound $= Q_1 - 1.5 \times IQR$

$= 6.25 - (1.5 \times 1.73)$

$= 3.655$

upper bound $= Q_3 + 1.5 \times IQR$

$= 7.98 + (1.5 \times IQR)$

$= \cancel{680}\ 10.575$

[3.655, 10.575]

So there is one outliner → 12.74

⇒ **Similarities**

From the first part of calculations we can see that $\bar{x}$, $\bar{y}$, var(x), var(y), $\sigma(x)$, $\sigma(y)$ cor(x, y) is same upto 3 decimal places.

⇒ **Differences**

From the second part we can say that

	Median	IQR	Outliners	Q_1	Q_2
D1	7.58	2.255	—	6.315	8.57
D2	8.14	2.255	1 (3.1)	6.695	8.95
D3	7.11	1.73	1 (12.74)	6.25	7.98

b) $cov(y_1, y_2) = \frac{1}{n-1} \Sigma (y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)$

4

$\Rightarrow \frac{1}{10} [(8.04-7.5)(9.14-7.5) + (6.95-7.6)(8.14-7.5) +$
$(7.58-7.5)(8.74-7.5) + (8.81-7.5)(8.77-7.5) +$
$(8.33-7.5)(9.26-7.5) + (9.96-7.5)(8.1-7.5) +$
$(7.24-7.5)(6.13-7.5) + (4.26-7.5)(3.1-7.5) +$
$(10.84-7.5)(9.13-7.5) + (4.82-7.5)(7.26-7.5) +$
$(5.68-7.5)(4.74-7.5)]$

$cov(y_1, y_2) = 1.933.$

$cov(y_2, y_3) = \frac{1}{n-1} \Sigma (y_{2i} - \bar{y}_2)(y_{3i} - \bar{y}_3)$

$\Rightarrow \frac{1}{10} [(7.46-7.5)(9.14-7.5) + (6.77-7.5)(8.14-7.5)$
$+ (12.74-7.5)(8.74-7.5) + (7.11-7.5)(8.77-7.5)$
$+ (7.81-7.5)(9.26-7.5) + (8.84-7.5)(8.1-7.5)$
$+ (6.08-7.5)(6.13-7.15) + (5.39-7.5)(3.1-7.5)$
$+ (8.15-7.5)(9.13-7.5) + (6.42-7.5)(7.26-7.5)$
$+ (5.73-7.5)(4.74-7.5)]$

$cov(y_2, y_3) = 2.425$

$$cov(y_1, y_3) = \frac{1}{n-1} \Sigma (y_{3i} - \bar{y}_3)(y_{1i} - \bar{y}_1)$$

$$\Rightarrow \frac{1}{10} \Big[(8.04 - 7.5)(9.14 - 7.5) + (6.95 - 7.5)(8.14 - 7.5)$$

$$+ (7.58 - 7.5)(8.74 - 7.5) + (8.81 - 7.5)(8.77 - 7.5)$$

$$+ (8.33 - 7.5)(9.26 - 7.5) + (9.96 - 7.5)(8.1 - 7.5)$$

$$+ (7.24 - 7.5)(6.13 - 7.5) + (3.1 - 7.5)(4.26 - 7.5)$$

$$+ (10.84 - 7.5)(9.13 - 7.5) + (4.82 - 7.5)(7.26 - 7.5)$$

$$+ (5.68 - 7.5)(4.74 - 7.5) \Big]$$

$$\Rightarrow cov(y_1, y_5) = 3.951$$

$$cov(y_1, y_1) = var = (y_1) = 4.127$$

$$cov(y_2, y_2) = var(y_2) = 4.1226$$

$$cov(y_3, y_3) = var(y_3) = 4.1276$$

$$Cov(y_i\ y_j) = \begin{bmatrix} 4.127 & 1.933 & 3.095 \\ 1.933 & 4.122 & 2.425 \\ 3.095 & 2.425 & 4.127 \end{bmatrix}$$

or

$$Cov(y_i, y_j) = \begin{bmatrix} 4.126 & & \\ 1.933 & 4.122 & \\ 3.095 & 2.425 & 4.127 \end{bmatrix}$$

Question 2

a) $x = (2,2,2,2)$ & $y(3,3,3,3)$

⇒ Cosine similarity $= \dfrac{x \cdot y}{\|x\| \|y\|}$

$$= \frac{(2\times3)+(2\times3)+(2\times3)+(2\times3)}{\sqrt{16} \times \sqrt{36}}$$

$$= \frac{4\times6}{4\times6} = \boxed{1}$$

⇒ Correlation $= \dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2 (y_i-\bar{y})^2}}$

$\bar{x} = 2, \ \bar{y} = 3$

as the values are same so, both numerator & denominator are 0.

so $= \boxed{\frac{0}{0} \text{ is undefined.}}$

⇒ Euclidean $= \sqrt{\Sigma(x_i-y_i)^2}$

$$= \sqrt{(2-3)^2+(2-3)^2+(2-3)^2+(2-3)^2}$$

$$= \sqrt{4} = \boxed{2}$$

b) $x = (0,1,0,1)$ & $y = (1,0,1,0)$

⇒ Cosine Similarity $= \dfrac{x \cdot y}{\|x\| \, \|y\|}$

$$= \frac{(0\times1)+(1\times0)+(0\times1)+(1\times0)}{\sqrt{2} \times \sqrt{2}} = \boxed{\underline{0}}$$

⇒ Correlation $= r = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$

$\bar{x} = \dfrac{0+1+0+1}{4} = 0.5$; $\bar{y} = \dfrac{1+0+1+0}{4} = 0.5$

$\Sigma(x_i - \bar{x})^2 = (0-0.5)^2 + (1-0.5)^2 + (0-0.5)^2 + (1-0.5)^2$
$= \underline{1}$

$\Sigma(y_i - \bar{y})^2 = (1-0.5)^2 + (0-0.5)^2 + (1-0.5)^2 + (0-0.5)^2$
$= \underline{1}$

$\Sigma(y_i - \bar{y})(x_i - \bar{x}) = -1 \left[(-0.5\times0.5) + (0.5\times-0.5) + (0.5\times-0.5) + (0.5\times-0.5)\right]$

$$r = \frac{-1}{\sqrt{1\times1}} = \boxed{\underline{-1}}$$

⇒ Jaccard Similarity $= \dfrac{|x \cap y|}{|x \cup y|}$

$x \cap y = 0$ & $x \cup y = 4$

$$= \frac{0}{4} = \boxed{\underline{0}}$$

eg. $x = (2, -1, 0, 2, 0, -3)$ & $y = (-1, 1, -1, 0, 0, -1)$ (7)

⇒ Cosine similarity $= \frac{x \cdot y}{\|x\| \|y\|}$

$$= \frac{-2 - 1 + 0 + 0 + 0 + 3}{\sqrt{18} \times \sqrt{4}}$$

$$= \boxed{0}$$

⇒ Correlation $= r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - x)^2 \, \Sigma (y_i - \bar{y})^2}}$

$\bar{x} = 0, \quad \bar{y} = -0.333$

$\Sigma (x_i - \bar{x})^2 = \Sigma (x_i)^2 = 18$

$\Sigma (y_i - \bar{y})^2 = (-1 - 0.333)^2 + (1 - 0.333)^2 + (-1 - 0.333)^2 + (0.333)^2 + (0 - 0.333)^2 + (-1 - 0.333)^2$

$= 3.333$

$\Sigma (x_i - \bar{x})(y_i - \bar{y}) = \frac{-4}{3} - \frac{4}{3} + \frac{2}{3} + \frac{6}{3} = 0$

$$r = \frac{0}{\sqrt{(18)(3.333)}} = \boxed{0}$$

⇒ Euclidean distance $= d(x, y) = \sqrt{\Sigma (x_i - y_i)^2}$

$\sqrt{(3)^2 + (-2)^2 + (1)^2 + (2)^2 + 0 + (-2)^2}$

$\sqrt{9 + 4 + 1 + 4 + 0 + 4} = \boxed{\sqrt{22} = 4.69}$

$\Rightarrow$ Jaccard Similarity $= \dfrac{|x \cap y|}{|x \cup y|}$ converting to binary.
no. of $> 1 = 1$
no. $= 0 = 0$

$|x \cap y| \Rightarrow$ where x & y are non-zero $x = (1, 1, 0, 1, 0, 1)$
$|x \cup y| \Rightarrow$ where x or y are non-zero $y = (1, 1, 1, 0, 0, 1)$

$|x \cap y| = 3$ (position 1, 2, & 6)

$|x \cup y| = 5$ (position 1, 2, 3, 4, 6)

so Jaccard similarity $= \boxed{\dfrac{3}{5} = 0.6}$

Question 3

a) hamming distance $\Rightarrow$ the no. of position at which the corresponding values are different

$x = 0\,1\,0\,\boxed{1}\,0\,1\,\boxed{0}\,0\,0\,\boxed{1}$

$y = 0\,1\,0\,\boxed{0}\,0\,1\,\boxed{1}\,0\,0\,\boxed{0}$

so $\boxed{\text{hamming distance} = 3}$

Jaccard Similarity $= \dfrac{q}{q + r + s}$

		y: 1	y: 0
x	1	2 (q)	2 (r)
	0	1 (s)	5 (t)

$= \dfrac{2}{2+2+1} = \boxed{\dfrac{2}{5} = 0.4}$

b.) The hamming distance is similar to SMC

The Jaccard measure is similar to cosine measure because both ignore 0 - 0.

On other hand SMC = $\frac{\text{hamming dis.}}{\text{no of bits}}$ is SMC is extention of hamming distance

8