1: Problem1

(a)
For midterm:
Max = 99, Min = 75
For final:
Max = 100, Min = 77

(b)
For midtern:
Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{10} x_i = 88.7$
Mode: 96
Median: 89
For final:
Mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{10} x_i = 88.2$
Mode: 80, 88
Median: 88

(c)
For midterm:

$$Q_1 = \frac{(10+1)}{4} = 2.75, \quad Firstquartile = 0.75 * 84 + 0.25 * 78 = 82.5$$
$$Q_3 = \frac{(10+3)}{4} = 8.25, \quad Thirdquartile = 0.25 * 96 + 0.75 * 96 = 96$$

For final:

$$Q_1 = \frac{(10+1)}{4} = 2.75, \quad Firstquartile = 0.75 * 80 + 0.25 * 80 = 80$$
$$Q_3 = \frac{(10+3)}{4} = 8.25, \quad Thirdquartile = 0.75 * 99 + 0.25 * 95 = 98$$

(d)
For midterm
Sample Variance:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{10}(x_i - overlinex)^2 \\ &= \frac{1}{10-1}((75-88.7)^2 \cdots (99-88.7)^2) = 65.122 \end{aligned}$$

Population Variance:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{10}(x_i - overlinex)^2 \\ &= \frac{1}{10}((75-88.7)^2 \cdots (99-88.7)^2) = 58.61\end{aligned}$$

For final
Sample Variance:

$$\begin{aligned}s^2 &= \frac{1}{n-1}\sum_{i=1}^{10}(x_i - overlinex)^2 \\ &= \frac{1}{10-1}((88-88.2)^2 \cdots (80-88.2)^2) = 63.956\end{aligned}$$

Population Variance:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{10}(x_i - overlinex)^2 \\ &= \frac{1}{10}((88-88.2)^2 \cdots (80-88.2)^2) = 57.56\end{aligned}$$

(e)
For midterm:
Sample standard deviation: $s = \sqrt{s^2} = 8.07$
Population standard deviation: $\sigma = \sqrt{\sigma^2} = 7.656$
For final:
Sample standard deviation: $s = \sqrt{s^2} = 7.997$
Population standard deviation: $\sigma = \sqrt{\sigma^2} = 7.587$

2: Problem 2

(a)
min-max normalization for midterm:

$$z = \frac{x - x(min)}{x(max) - x(min)}$$

No.1 = 0.833, No.2 = 0.458, No.3 = 0.125

(b)
No.4 = 1, No.5 = 0.375, No.6 = 0.625, No.7 = 0.542
No.8 = 0, No.9 = 0.875, No.10 = 0.875
Population Variance: $\sigma^2 = 0.102$

(c)

z-score normalization for final:

$$z = \frac{x - \mu}{\sigma}$$

No.4 = 0.896, No.5 = -0,422, No.6 = -1.476

(d)
Population Variance: $\sigma^2 = 1$

3: Problem 3

(a)
Covariance between the midterm and final:
E(midterm) = 88.7
E(final) = 88.2
E(midterm-final) = 7841.5
$\sigma = E(midterm - final) - E(midterm) * E(final) = 18.16$

(b)
Pearson's correlation coefficient:

$$\rho = \frac{\sigma_{midterm-final}}{\sqrt{\sigma^2_{midterm} * \sigma^2_{final}}} = \frac{18.16}{\sqrt{58.61 * 57.56}} = 0.313$$

(c)
No. From part(a), midterm and final have postive Covariance And from part(b), these two are postivty correlated.

(d)
(1) Manhattan distance:

$$\textbf{distance} = \begin{pmatrix} L & No.1 & No.2 & No.3 & No.4 & No.5 & No.6 & No.7 & No.8 & No.9 & No.10 \\ No.1 & 0 & \dots & & & & & & & & \\ No.2 & 9 & 0 & \dots & & & & & & & \\ No.3 & 15 & 6 & 0 & \dots & & & & & & \\ No.4 & 11 & 20 & 26 & 0 & \dots & & & & & \\ No.5 & 14 & 5 & 1 & 25 & 0 & \dots & & & & \\ No.6 & 16 & 7 & 1 & 27 & 2 & 0 & \dots & & & \\ No.7 & 4 & 13 & 19 & 7 & 18 & 20 & 0 & \dots & & \\ No.8 & 28 & 19 & 13 & 39 & 14 & 12 & 32 & 0 & \dots & \\ No.9 & 13 & 22 & 28 & 2 & 27 & 29 & 9 & 41 & 0 & \dots \\ No.10 & 7 & 2 & 8 & 18 & 7 & 9 & 11 & 21 & 20 & 0 \end{pmatrix}$$

(2) Euclidea distance:

$$\textbf{distance} = \begin{pmatrix} L & No.1 & No.2 & No.3 & No.4 & No.5 & No.6 & No.7 & No.8 & No.9 & No.10 \\ No.1 & 0 & \dots & & & & & & & & \\ No.2 & 9 & 0 & \dots & & & & & & & \\ No.3 & 17.117 & 8.246 & 0 & \dots & & & & & & \\ No.4 & 8.062 & 14.765 & 21.587 & 0 & \dots & & & & & \\ No.5 & 11.402 & 3.606 & 7.810 & 18.028 & 0 & \dots & & & & \\ No.6 & 12.083 & 11.705 & 17.692 & 20.125 & 10 & 0 & \dots & & & \\ No.7 & 13.038 & 11.180 & 13.454 & 11.705 & 14.562 & 22.091 & 0 & \dots & & \\ No.8 & 21.541 & 13.620 & 10.440 & 28.302 & 10.296 & 15.297 & 23.022 & 0 & \dots & \\ No.9 & 12.042 & 15.620 & 20.591 & 5.831 & 19.209 & 23.770 & 8.062 & 29 & 0 & \dots \\ No.10 & 8.062 & 12.806 & 20.591 & 15.297 & 13 & 6.708 & 20.616 & 21 & 20 & 0 \end{pmatrix}$$

(3) Supremum distance:

$$\textbf{distance} = \begin{pmatrix} L & No.1 & No.2 & No.3 & No.4 & No.5 & No.6 & No.7 & No.8 & No.9 & No.10 \\ No.1 & 0 & \dots & & & & & & & & \\ No.2 & 9 & 0 & \dots & & & & & & & \\ No.3 & 17 & 8 & 0 & \dots & & & & & & \\ No.4 & 7 & 13 & 21 & 0 & \dots & & & & & \\ No.5 & 11 & 3 & 6 & 15 & 0 & \dots & & & & \\ No.6 & 11 & 11 & 13 & 18 & 8 & 0 & \dots & & & \\ No.7 & 11 & 11 & 10 & 11 & 14 & 22 & 0 & \dots & & \\ No.8 & 20 & 11 & 10 & 24 & 9 & 15 & 19 & 0 & \dots & \\ No.9 & 12 & 12 & 18 & 5 & 15 & 23 & 8 & 21 & 0 & \dots \\ No.10 & 8 & 10 & 18 & 15 & 12 & 6 & 19 & 21 & 20 & 0 \end{pmatrix}$$

(4) Cosine similarity of m and f:

$$\begin{aligned} cos(d1, d2) &= \frac{d_1 \dot{d}_2}{||d_1||X||d_2||} \\ &= 0.995 \end{aligned}$$

The code is attached in the appendix.

(e)

Supremum distance represents the maximum distance of midterm and final scores between each student.

(f)

No. KL divergence is a measure of lost information, not the distance measure. KL divergence measures the information gained, which means measures the experienmental data against the reference data. However, in this case, both m and f are experienmental data.
Jaccard coefficient will not a good choice, either. It is a similarity measure for asymmetric binary variables. In this case, these attributes are not binary. And contingency information is also needed.

4: Problem 4

(a)
Null hypothesis: The two distributions are independent.

	purchased diapher	not purchased diapher	sum (row)
purchased beer	200 (18.667)	80 (261.333)	280
not purchased beer	20 (201.333)	3000 (2818.667)	3020
sum(col)	220	3080	3300

$\chi^2 = \frac{(200-18.667)^2}{18.667} + \frac{(80-261.333)^2)}{261.333} + \frac{(20-201.333)^2)}{201.333} + \frac{(3000-2818.667)^2)}{2818.667} = 2062.294$

(b)
Reject the null hypothesis of independence at confidence level of 0.001

(c)

$$p_0 = \frac{200}{3300} = \frac{2}{33}$$
$$p_1 = \frac{100}{3300} = \frac{1}{33}$$
$$p_2 = \frac{3000}{3300} = \frac{30}{33}$$

Thus,
$\mathbf{p} = [p_0 \quad p_1 \quad p_2]^T = [2/33 \quad 1/33 \quad 30/33]^T$

(d)
$\mathbf{q} = [0.5 \quad 0.3 \quad 0.2]^T$
$\mathbf{p} = [2/33 \quad 1/33 \quad 30/33]^T$
KL-divergence when use $\mathbf{p}$ to approximate $\mathbf{q}$:

$$\begin{aligned} D_{KL}(q(x)||p(x)) &= \sum_x \in X q(x) \ln \frac{q(x)}{p(x)} \\ &= 1.44 \end{aligned}$$

Appendix

```
import numpy as np
import scipy.stats as stats
import math

a  = np.array([95, 86, 78, 99, 84, 90, 88, 75, 96, 96])
b = np.array([88, 88, 90, 95, 85, 77, 99, 80, 100, 80])



#Manhattan distance
for i in range(0,10):
    print("The value of i %d", i+1)
    for j in range(0, 10):
        Manhattan = np.abs((a[j] - a[i]) + (b[j] - b[i]))
        print(Manhattan)

#Euclidean distance
for i in range(0, 10):
    print("The value of i %d", i+1)
    for j in range(0, 10):
        Euclidean = np.sqrt(np.square(a[j] - a[i]) + np.square(b[j] -
b[i]))
        print(Euclidean)

#Supremum distance
for i in range(0, 10):
   print("The value of i %d", i+1)
   for j in range(0, 10):
        c = np.absolute(a[j] - a[i])
        d = np.absolute(b[j]- b[i])
        print(np.max([c,d]))
```