CS412: Introduction to Data Mining, Fall 2022, Homework 1

Name: Teng Hou (tenghou2)

Q1

a. *For any set of n numbers* $X = \{X_1, X_2...X_n\}$. *The mean can be computed as* $\mu = \sum_{i=1}^{n} x_i$. *So for the given dataset age, the mean could be calculated as:*

$$\mu_{age} = \frac{23 + 23 + 27 + 27 + 39 + 41 + 47 + 49 + ... + 58 + 58 + 60 + 61 + 65}{20}$$

$$\mu_{age} = 47.75$$

So for the given dataset fat, the mean could be calculated as:

$$\mu_{fat} = \frac{8.7 + 27.2 + 7.4 + 17.9 + 31.8 + 24.9 + 27.4 + 27.2 + ... + 32.9 + 41.2 + 35.7 + 37.4 + 36.2}{20}$$

$$\mu_{fat} = 29.535$$

For any set of n numbers $X = \{X_1, X_2...X_n\}$. *The Standard Deviation can be computed as* $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$. *So for the given dataset age, the Standard Deviation could be calculated as:*

$$std_{age} = \sqrt{\frac{(23 - \mu_{age})^2 + (23 - \mu_{age})^2 + (27 - \mu_(age))^2 + ... + (65 - \mu_{age})^2}{20}} = 9.018$$

So for the given dataset fat, the Standard Deviation could be calculated as:

$$std_{fat} = \sqrt{\frac{(8.7 - \mu_{fat})^2 + (27.2 - \mu_{fat})^2 + (7.4 - \mu_(fat))^2 + ... + (36.2 - \mu_{fat})^2}{20}} = 12.918$$

b. *To find out the First quartile, median and third quartile, we could sort the date set from small to large. Then the dataset of age is*

$$X = \{23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 54, 56, 57, 58, 58, 60, 61, 65\}$$

The first quartile Q1=$\frac{X_5 + X_6}{2} = 40$
Median=$\frac{X_{10} + X_{11}}{2} = 53$
Third quartile=$\frac{X_{15} + X_{16}}{2} = 57.5$

The dataset of fat is:

$$X = \{7.4, 8.7, 17.9, 24.9, 27.2, 27.2, 27.4, 28.8, 30.2, 31.2, 31.8, 32.9, 33.4, 34.1, 34.6, 35.7, 36.2, 37.4, 41.2, 42.5\}$$

The first quartile Q1=$\frac{X_5+X_6}{2} = 27.2$
Median=$\frac{X_{10}+X_{11}}{2} = 31.5$
Third quartile=$\frac{X_{15}+X_{16}}{2} = 35.15$

c. *from question b, I already have the sort dataset of age, which is*

$$X = \{23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 54, 56, 57, 58, 58, 60, 61, 65\}$$

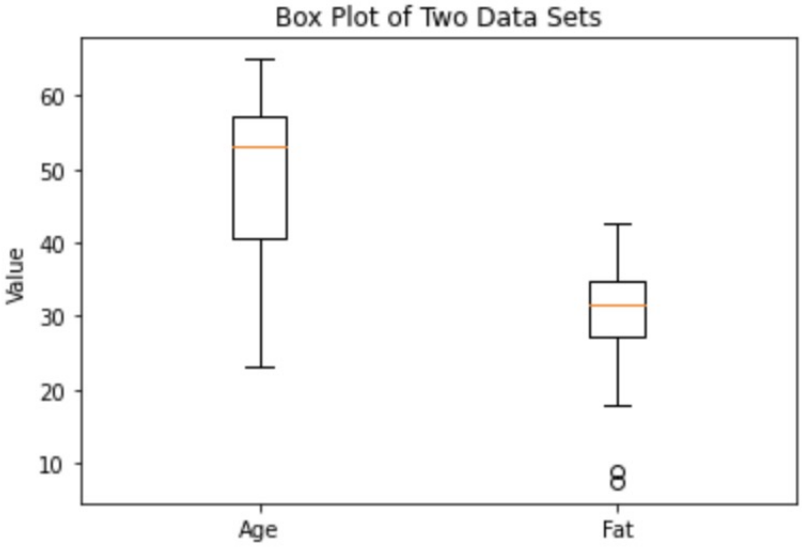
So the maximum is 65, the minimum is 23
The sort dataset of fat is:

$$X = \{7.4, 8.7, 17.9, 24.9, 27.2, 27.2, 27.4, 28.8, 30.2, 31.2, 31.8, 32.9, 33.4, 34.1, 34.6, 35.7, 36.2, 37.4, 41.2, 42.5\}$$

So the maximum is 42.5, the minimum is 7.4

d. *For the dataset of age, the age 54 appears for three times which is the most among all the ages. So the mode of age dataset is 54.*
For the fat dataset, 27.2 appears for 2 times while others appear only one time, so the mode of fat dataset is 27.2.

e. *Here are the two boxplots for age and fat:*



Q2

Fisrt we need to find out the populuation of the town, which is:

$$P = 200 + 450 + 300 + 1500 + 700 + 44 = 3194$$

So the dataset of the population's age is:

$$X = \{X_1, X_2, X_3, ...X_{3193}, X_{3194}\}$$

The population from 1 to 45 is 950, and the population from 1-60 is 2450, so the 1597th is located in the bar from 45 to 60. And we could find out that the 1597th is the 1597-950=647th from age 45. And from the histogram, the width of each bar is 15.
Then we could find out the median:

$$med = 45 + \frac{15}{1500} * 647 = 51.47$$

Q3

a. *First I need to find out the mean of Jan-20-original. For column* $C = \{C_1, C_2, ..., C_n\}$, *the mean could be calculated as* $\mu = \frac{C_1+C_2+...+C_n}{n}$. *So the mean is:*

$$\mu = \frac{41.018 + 36.878 + ... + 42.296}{1000} = 53.279$$

The variance can be computed as:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

So for Jan-20-original, the variance is:

$$\sigma_1^2 = \frac{(41.018 - 53.279)^2 + (36.878 - 53.279)^2 + ... + (42.296 - 53.279)^2}{1000} = 406.079$$

The standard variance could be computed as: $\sqrt{\sigma^2} = 20.151$. *Then use z-score normalization, which means* $z = \frac{x-\mu}{\sigma}$. *So we can get the normalized dataset*

$$Z_1 = \{-0.608, -0.814, 0.183... - 0.545\}$$

Then the variance of the normalized dataset is:

$$\sigma_2^2 = 1.000$$

By comparing the variances before and after the normalization, I find the first variance is much larger than that after normalization. whilee the variance after normalization is really closed to 1.

b. *Now we have the standard variance and mean of Jan-20-original, then we need to find out those of Jan-80-original. For column* $C = \{C_1, C_2, ..., C_n\}$, *the mean could be calculated as* $\mu = \frac{C_1+C_2+...+C_n}{n}$. *The variance can be computed as:*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Then we could find the mean and variance of Jan-80-original is:

$$\mu_{80} = \frac{39.488 + 34.664 + ... + 30.524}{1000} = 51.355$$

$$\sigma_{80}^2 = \frac{(39.488 - 51.355)^2 + (34.664 - 51.355)^2 + ... + (30.524 - 51.355)^2}{1000} = 445.265$$

Then the standard variance is:

$$\sigma_{80} = 21.101$$

From a, I know $\mu_{20} = 53.279, \sigma_{20}^2 = 406.079, \sigma_{20} = 20.151$
For two sets A and B, the Pearson's correlation coefficient could be computed as:

$$\frac{\Sigma_{i=1}^{n}(A_i - \mu_A)(B_i - \mu_B)}{n * \sigma_A * \sigma_B}$$

Then the Pearson's correlation coefficient between Jan-80-original and Jan-20-original is 0.962

c. *From b, I already have* $\mu_{80} = 51.355$, $\sigma_{80} = 21.101$. *Then I need to find out the* μ *and* σ *of Jan-20-original. From a, I know the variance of Jan-20-normalized is 1.000, then the standard variance should be* $\sigma_{20nor} = 1.000$, *and the mean value is* $\mu_{20nor} = 0.000$. *For two sets A and B, the Pearson's correlation coefficient could be computed as:*

$$\frac{\Sigma_{i=1}^{n}(A_i - \mu_A)(B_i - \mu_B)}{n * \sigma_A * \sigma_B}$$

Then we could get the Pearson correlation coefficient is 0.962. Which is the same to the answer of b.

d. *For dataset A and B, the covariance between Jan-80-original and Jan-20-original could be calculated as:*

$$Cov(A, B) = \frac{\Sigma_{i=1}^{n}(a_i - \mu_A)(b_i - \mu_B)}{n}$$

From the former questions, I know $\mu_A = 53.279$ *and* $\mu_B = 51.355$. *So the covariance is 408.993.*

Q4

a. *The Minkowski distance of two dataset could be computed as:*

$$d(i,j) = \sqrt[p]{|X_{i1} - Y_{j1}|^p + |X_{i2} - Y_{j2}|^p + ... + |X_{il} - Y_{jl}|^p}$$

(i): when h=1, then the Minkowski distance could be computed as:

$$d(i,j) = |X_{i1} - Y_{j1}|^h + |X_{i2} - Y_{j2}|^h + ... + |X_{il} - Y_{jl}|^h = |15 - 74| + |67 - 126| + ... + |33 - 61|$$

So the answer is 8042.
(ii): when h=2, then the Minkowski distance could be computed as:

$$d(i,j) = \sqrt[2]{|15 - 74|^2 + |67 - 126|^2 + ... + |33 - 61|^2} = 928.417$$

(iii): when h=3, then the Minkowski distance could be computed as:

$$d(i,j) = 197$$

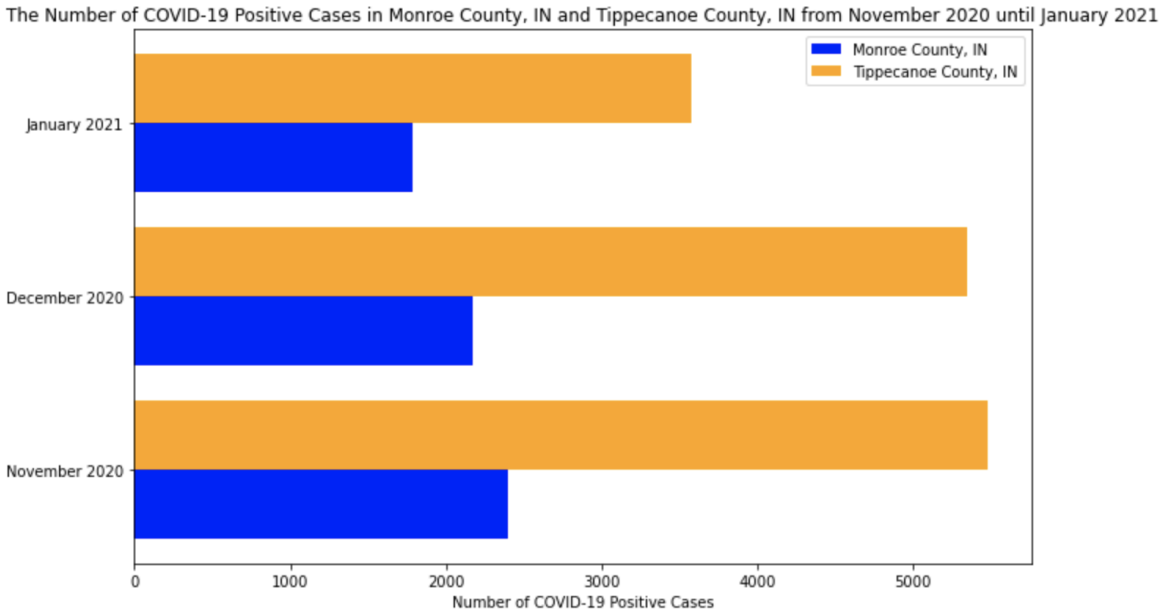
b. *For two datasets, the cosine similarity could be computed as:*

$$cos(d_1, d_2) = \frac{d_1 d_2}{||d_1|| * ||d_2||}$$

Then the cosine similarity between Monroe County and Tippecanoe could be calculated as:

$$\frac{15 * 74 + 67 * 126... + 33 * 61}{541054 * 2652992} = 0.973$$

c. *The chart is:*



Q5

a. *Let q=205, r=20, s=25, t=2200. Then the distance berween the binary attribution Drink Coffee and Eat Chocolate could be computed as:*

$$d(i,j) = \frac{r+s}{q+r+s+t} = \frac{45}{2450} = 0.0184$$

b. *The Jaccard coefficient between Drink Coffee and Eat Chocolate could be calculated as:*

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

So the Jaccard coefficient is:

$$\frac{205}{205+20+25} = 0.8200$$

c. *Drink Coffee's number is 230, Do Not Drink Coffee's number is 2220, Eat Chocolate's number is 225, Do Not Eat Chocolate's number is 2225.*
Then the expectation of who drink coffee and eat chocolate is:

$$e_{11} = \frac{230 * 225}{2450} = 21.1224$$

The expectation of who drink coffee but not wat chocolate is:

$$e_{21} = \frac{230 * 2225}{2450} = 208.8776$$

The expectation of who do not drink coffee but eat chocolate is:

$$e_{12} = \frac{2220 * 225}{2450} = 203.8776$$

Then the expectation of who do not drink coffee and not eat chocolate is:

$$e_{22} = \frac{2220 * 2225}{2450} = 2016.1224$$

Then the X^2 could be calculated as:

$$X^2 = \Sigma_{i=1}^{c}\Sigma_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

So the answer could be calculated as:

$$\frac{(205-21.1224)^2}{21.1224} + \frac{(25-208.8776)^2}{208.8776} + \frac{(20-203.8776)^2}{203.8776} + \frac{(2200-2016.1225)^2}{2016.1225}$$

The answer is 1945.1960

d. *The degree could be calculated as: $(row-1)*(column-1)$, so the degree for this question is $(2-1)*(2-1)=1$. As the answer from c is 1945.1960, which is much larger than 3.84, so the null hypothesis could be rejected.*