1.（a）For final score set $FS = \{fs_1, fs_2, fs_3, ..., fs_{1000}\}$，the mean could be computed as $\mu = \frac{1}{1000}\sum_{i=1}^{1000} fs_i = 87.084$. The standard deviation could be computed as $\sigma = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(fs_i - \mu)^2} \approx 10.914$

(b) We could sort the final score set $FS = \{fs_1, fs_2, fs_3, ..., fs_{1000}\}$ in ascending order into sorted final score set $OFS = \{ofs_1, ofs_2, ofs_3, ..., ofs_{1000}\}$. It is a one-to-one-mapping, where ofs_1 means the minimum of the final score set, ofs_2 means the second minimum of the final score set and so on.

The first quartile Q_1 is 25th percentile, which is $ofs_{250} = 82$. The median is $ofs_{500} = 89$. The third quartile is 75th percentile, which is $ofs_{750} = 96$

(c) Maximum is $ofs_{1000} = 100$. Minimum is $ofs_1 = 35$

(d) Mode is the number with highest frequency. Suppose a set with 101 elements in it: $S = \{0,0,0,0,...,0\}$, where s_j represents the j th element of S. For any fs_i in FS, there exists a s_j in S so that $fs_i = j$. Then $s_j += 1$. After visiting all elements in FS and one thousand time additions, s_k is the largest element of S. k is the mode of S. $k = 97$

(e) YES.

Median is 89. Mean is 87.084. Standard deviation is 10.914.
$87.084 - 10.914 = 76.17 < 89 < 97.998 = 87.084 + 10.914$.

(f) NO.

$(maximum - median) = (100 - 89) = 11 > 10.5 = 1.5 * (Q_3 - median)$

2.Median is at the middle of the data. Define the hourly pay in SkyNet as $HP = \{x_1, x_2, x_3, ..., x_{583}\}$ in ascending order, where 583 is calculated by adding all the frequency together. There are $\frac{583}{2} = 291.5$ elements before the median x_{median}. There are $54 + 74 + 89 = 219$ smaller than $x_{219} = 25$. There are $219 + 102 = 321$ elements smaller than $x_{321} = 30$. Therefore, x_{median} is between 25 and 30.
$291.5 - 219 = 102 * \frac{x_{median} - 25}{30 - 25}$. Therefore, $x_{median} = 28.554$.

3.(a)z-score could be calculated by $zfs_i = \frac{fs_i - \mu}{\sigma}$, where fs is initial final score, μ is expected value of fs, σ is standard deviation of fs. μ could be calculated by $\mu = \frac{1}{1000}\sum_{i=1}^{1000} fs_i = 87.084$. σ could be computed by $\sigma = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(fs_i - \mu)^2} \approx 10.914$, $\mu_z = 0$.

Variance of finals_original could be calculated by $\frac{1}{1000}\sum_{i=1}^{1000}(fs_i - \mu)^2 \approx 119.113$.
Variance of finals_normalized could be calculated by $\frac{1}{1000}\sum_{i=1}^{1000}(zfs_i - \mu_z)^2 = 1$.

(b)$zfs_{e1} = \frac{fs_{e1} - \mu}{\sigma} = \frac{90 - \mu}{\sigma} \approx 0.267$

(c)Pearson's correlation coefficient could be calculated by $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$, where $Cov(A,B) = \sum_{i=1}^{n}(a_i - \mu_A)(b_i - \mu_B)$. Pearson's correlation coefficient between midterm-original and finals_original could be calculated by

$$r_{mo,fo} = \frac{Cov(mo,fo)}{\sigma_{mo}\sigma_{fo}} = \frac{\sum_{i=1}^{1000}(mo_i - \mu_{mo})(fo_i - \mu_{fo})}{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(mo_i - \mu_{mo})^2}\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(fo_i - \mu_{fo})^2}} \approx 0.544.$$

(d)Pearson's correlation coefficient between midterm-original and finals_normalized could be calculated by

$$r_{mo,fo} = \frac{Cov(mo,fn)}{\sigma_{mo}\sigma_{fn}} = \frac{\sum_{i=1}^{1000}(mo_i - \mu_{mo})(fn_i - \mu_{fn})}{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(mo_i - \mu_{mo})^2}\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(fn_i - \mu_{fn})^2}} \approx 0.544.$$

(e)$Cov(mo, fo) = \sum_{i=1}^{1000}(mo_i - \mu_{mo})(fo_i - \mu_{fo}) \approx 78.176$.

4.(a) Minkowski distance could be calculated by $d(CML, CBL) = \sqrt[h]{|n_{m1} - n_{b1}|^h + |n_{m2} - n_{b2}|^h + \ldots + |n_{m100} - n_{b100}|^h}$, where n_{mi} means the number of ith book of CML , n_{bi} means the number of ith book of CBL .

When h=1,
$d_1(CML, CBL) = |n_{m1} - n_{b1}| + |n_{m2} - n_{b2}| + \ldots + |n_{m100} - n_{b100}| = 6152$.

When h=2,
$d_2(CML, CBL) = \sqrt{|n_{m1} - n_{b1}|^2 + |n_{m2} - n_{b2}|^2 + \ldots + |n_{m100} - n_{b100}|^2} \approx 715.328$.

When h=3,

$d_3(CML, CBL) = \sqrt[\infty]{|n_{m1} - n_{b1}|^\infty + |n_{m2} - n_{b2}|^\infty + \ldots + |n_{m100} - n_{b100}|^\infty}$

$= max_{i=1}^{\infty}|n_{mi} - n_{bi}| = 170$.

(b)I don't agree.

$d_1(CML, CBL) = 6125 > d_2(CML, CBL) = 715.328 > d_3(CML, CBL) = 170.$

(c)Possibility density function of CML and CBL could be calculated by $P_{CML}(i) = \frac{n_{mi}}{\sum_{j=1}^{100} n_{mj}}$ and $P_{CBL}(i) = \frac{n_{bi}}{\sum_{j=1}^{100} n_{bj}}$. Assume the ith element of P_{CML} is p_i and the ith element of P_{CBL} is q_i. $D_{KL}(CML||CBL) = \sum_{i=1}^{100} p_i \log(\frac{p_i}{q_i}) \approx 0.207.$

5.(a) Distance for binary attributes could be calculated by $d(i, j) = \frac{r+s}{q+r+s+t}$, where i, j represents two attributes, q,r,s,t represents the same meaning in slides. For Buy Beer and Buy Diaper, the distance could be calculated by $d(BB, BD) = \frac{10+40}{3500} = \frac{1}{70} \approx 0.0143.$

(b)Jaccard coefficient could be calculated by $sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$, where i,j,q,r,s represents the same meaning in slides. For Buy Beer and Buyr Diaper, the Jaccard coefficient could be calculated by $sim_{Jaccard}(BB, BD) = \frac{150}{150+10+40} = 0.75.$

(c)Assume two nominal attributes A and B, where A has value $\{a_1, a_2, \ldots, a_c\}$, B has value $\{b_1, b_2, \ldots, b_r\}$

e_{ij} represents the expected frequency in each entry(i,j) and could be calculated by $e_{ij} = \frac{count(A=a_i) \times count(B=b_j)}{n}$.

The χ^2 statistic could be calculated by $\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$, where o_ij is the observed frequency.

$e_{11} = \frac{160 \times 190}{3500} \approx 8.686$

$e_{12} = \frac{190 \times 3340}{3500} \approx 181.314$

$e_{21} = \frac{160 \times 3310}{3500} \approx 151.314$

$e_{22} = \frac{3310 \times 3340}{3500} \approx 33158.686$

$\chi^2 = \frac{(o_{11}-e_{11})^2}{e_{11}} + \frac{(o_{12}-e_{12})^2}{e_{12}} + \frac{(o_{21}-e_{21})^2}{e_{21}} + \frac{(o_{22}-e_{22})^2}{e_{22}} \approx 2547.582$

(d)Yes

We could know that for one freedom (calculated by(2-1)$\times$(2-1)) and $\alpha = 0.05$, value need to reject null hypothesis is 3.841. $2547.582 > 3.841$. Therefore, we could reject the hypothesis.