Solutions:

Name: Sidhartha Satapathy
Net-id: ss46

Question 1

Part 1.
In order to calculate max and min,

a. I read all the scores from the input file and store it in a list.
b. Then, I sorted the list.
c. Finally, I picked up the first element as minimum and last element as maximum.

Max = maximum of all the elements in the list.
Min = minimum of all the elements in the list.

Min = 37
Max = 100

Part 2.
In order to get q1,q3 and median

a. Using the sorted list, q1 would be the (n/4)th element in the list.
b. Again I use the sorted list, and return the average of the middle elements as the median.
c. Finally, q3 would be the (3n/4)th element in the list.

Q1 = 25 percentile element
Median = 50 percentile element
Q3 = 75 percentile element

Q1 = 68
Median = 77 ((77.0+77.0)/2)
Q3 = 87

Part 3.
I calculate the mean by taking the sum of all the elements in the list and then dividing it by the total number of elements.
Mean = (Sum of all elements in the list) / (total number of elements)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Sum = 76715 (x1+x2+....+xn) and n = 1000
Mean = 76.715

Part 4.
The mode of a set of data values is the value that appears most often.
I calculated the frequency of each value and output the ones with the max frequency.

Mode = [77,83]

Part 5.
First I iterate on my list and calculate the square of the mean subtracted from the element
Then, I sum all such expressions and eventually divide by (total elements - 1). Formula:

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

Mean = 76.715
N - 1 = 1000 - 1 = 999
Empirical variance = 173.279

Question 2.

Part 1.
The empirical variance is calculated using the formula mentioned in Q1. part 5. For this part I repeat it with z-score values.

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

Standard deviation= 13.164(Calculated by taking root of the empirical variance:173.279)
Mean = 76.715

The Empirical variance before z-score normalization = 173.279
And The Empirical variance after z-score normalization = 1

Part 2.
The score is calculated as follows;
Z-score = (original score - mean) / (standard deviation)

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

Standard deviation = 13.164
Mean = 76.715
Original score:90, then z-score:1.009

Part 3.

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Using the above formula, I first calculate the numerator by taking a summation on the xi*yi term. I then subtract n* xmean* ymean. Similarly I calculate the denominator by multiplying standard deviation for x, standard deviation for y and (total elements-1).Finally I divide these.
Sx (Standard deviation for midsem scores) = 13.164
Sy (Standard deviation for final scores) = 10.919
N = 1000
mean(Endsem) = 87.084
mean(Midsem) = 76.715
Numerator = 78175.94

Pearson's coefficient between midterm scores and final scores is:0.544

Part 4.

$$q_{jk} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right),$$

So I used the same numerator from Question 2 part 3 but instead of dividing by standard deviation of x, standard deviation of y and (total sample size-1) I divide only by (total sample size-1).
N = 1000
Numerator = 78175.94
Covariance between midterm scores and final scores is:78.254

Question 3:

Part 1.

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}.$$

For the asymmetric binary values, it is calculated using the above formula. (Here M11: is the number of books both have, M01 and M10: is the number of books one of them has)

J = 58/(2+120+58)

Jaccard coefficient:0.322

Part 2.

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

I calculate the the minkowski distance using the above formula for h=1 and h=2.
First take the appropriate sum and then the correct root.

$$\lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} = \max_{i=1}^{n} |x_i - y_i|.$$

This formula is used for h = infinity

The values are as follows:
h=1: 6152
h=2: 715.328
h=infinity: 170

Part 3.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

The Cosine similarity between Citadel's Maester Library (CML) and Castle Black's with regard to the feature vector is calculated using the above formula and the value obtained is as follows:
Summation(ai*bi) = 1344428.0
||A||2 (CML)= 1229.637
||B||2 (CBL)= 1299.439
cosine similarity : 0.841

Part 4.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

The Kullback–Leibler divergence of these two libraries P(CML || CBL) is calculated using the above formula. Note that the probability P(i) and Q(i) are evaluated as the probability of a person picking a book at random. For example, the probability of a person that is P(i), to pick up book 1 in Citadel's Maester Library (CML) is i_1 / (i_1 + ... + i_100) .

CML_i_1 + CML_i_2 …. + CML_i_100 = 11203 (total CML)
CBL_i_1 + CBL_i_2 …. + CBL_i_100 = 12045 (total CBL)
P(i_1) = CML_i_1/total CML and Q(i_1) =CBL_ i_1/total CBL (This is how we calculate probability values)
P(i_1) = 0.00473087565831
Q(i_1) = 0.00855126608551 (Similarly we have others)

The final value obtained using this formula is as follows:
KL Divergence:0.207

Question 4:
I calculate the chi-square correlation as follows:

Q11 = 150
Q12 = 40
Q21 = 15
Q22 = 3300
total = Q11 + Q12 + Q21 + Q22
E11 = float(Q11+Q12)*(Q11+Q21)/total
E12 = float(Q11+Q12)*(Q12+Q22)/total
E21 = float(Q11+Q21)*(Q21+Q22)/total
E22 = float(Q22+Q21)*(Q12+Q22)/total
chi = (Q11 - E11)**(2)/E11
chi += (Q12 - E12)**(2)/E12
chi += (Q21 - E21)**(2)/E21
chi += (Q22 - E22)**(2)/E22

Finally, we get chi-square correlation value: 2468.183

Source for the images:

https://en.wikipedia.org/wiki/Covariance#Calculating_the_sample_covariance
https://en.wikipedia.org/wiki/Cosine_similarity
https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
https://en.wikipedia.org/wiki/Minkowski_distance
https://en.wikipedia.org/wiki/Jaccard_index#Similarity_of_asymmetric_binary_attributes
https://en.wikipedia.org/wiki/Mean
https://en.wikipedia.org/wiki/Standard_score
http://mathworld.wolfram.com/SampleVariance.html
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient