Assignment 1

Name: Zixin Ouyang

1. Import Python library Pandas, read the text file into Python using function read_csv (), name the data frame scores and name the columns: id, midterm, final.

a. Use function min () and min () to calculate minimum and maximum of midterm.

Min is 37, and max is 100.

b. Use function quantile () to compute quantiles of midterm. The list parameters [0.25,0.5,0.75] correspond to the first quantile Q1, median, and third quantile Q3.

Q1 is 68, median is 77, Q3 is 87.

c. Use function mean () to compute the mean score of midterm.

Mean is 76.715.

d. Use function mode () to compute the modes of midterm.

Midterm has two modes: 77, 83.

$\bar{x}=\frac{1}{n}\sum x_i$

e. Use function var () to compute the sample variance of midterm. The degree of freedom is n-1, which is the default ddof in the function.

Empirical variance is 173.279.

$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{1000-1}\sum_{1}^{1000}(scores.midterm-scores.midterm.mean())^2$

2.

a. z-score normalization function: $\frac{x-\mu}{\sigma}$

Midterm score after normalization: $\frac{\text{scores.midterm}-\text{scores.midterm.mean()}}{\text{scores.midterm.std()}}$

Empirical variance before normalization: 173.279

Empirical variance after normalization: 1.000

b. The corresponding score after normalization:

$\frac{90-scores.midterm.mean()}{scores.midterm.std()}$ = 1.009

c. Use function corr () to compute correlation coefficient between midterm scores and final scores. The parameter method ‘pearson’ corresponds to the Pearson’s correlation coefficient, which is 0.544.

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=$$

$$\frac{\sum_{i=1}^{1000}(scores.midterm-scores.midterm.mean())(scores.final-scores.final.mean())}{\sqrt{\sum_{i=1}^{1000}(scores.midterm-scores.midterm.mean())^2}\sqrt{\sum_{i=1}^{n}(scores.final-scores.final.mean())^2}}$$

d. Use function cov () to compute covariance between midterm scores and final scores, which is 78.254.

cov(X,Y) = $\rho_{X,Y} \times \sigma_X \sigma_Y$

=r × scores.midterm.std()× scores.final.std()

3.

a. Jaccard coefficient: $\frac{q}{q+r+s} = \frac{58}{120+58+2} = 0.322$

b. Read the text file into Python using function read_csv (), name the data frame books, import library numpy, and transform the data frame to a numpy array. Use function pdist () in the scipy library to calculate the minkowski distance. The parameter metric is used to set distance function. Parameter p corresponds to h value. When h is infinite, the minkowski distance is equal to chebyshev distance.

The minkowski distance:

h=1: Manhattan (or city block) distance

d (i, j) = $|x_{i1} - x_{j1}|+|x_{i2} - x_{j2}|+\ldots\ldots+|x_{il} - x_{jl}|$ =6152

h=2: Euclidean distance

d (i, j) = $\sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$ =715.328

h=∞: chebyshev distance

d (i, j) =$\max_{f=1}|x_{if} - x_{jf}|$ =170

c. Use function pdist () to comput cosine distance, set the parameter metric to 'cosi ne'. The cosine distance is 0.159.

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

d. Use function entropy () to compute Kullbac-Leibler divergence. Parameter lib[0] is the array of numbers of books in CML, while lib[1] is the array of numbers of bo oks in CBL.

The Kullbac-Leibler divergence is 0.207.

D_{KL} (p(x)‖$q(x)$) =$\sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}$

4.

	Buy diaper	Do not buy diaper	Total
Buy beer	150 (9)	40 (181)	190
Do not buy beer	15 (156)	3300 (3159)	3315
Total	165	3340	3505

$\frac{(150-9)^2}{9} + \frac{(40-181)^2}{181} + \frac{(15-156)^2}{156} + \frac{(3300-3159)^2}{3159} = 14.227$