**Question 1.**

**1.1 Answers**

*A) Maximum and minimum.*

Mid-Term 37 , 100

Final: 35 , 100

*B) First quartile Q1, median, and third quartile Q3*

Mid-Term: 68.0 , 77.0 , 87.0

Final: 82.0 , 89.0 , 96.0

*C) Mean*

Mid-Term: 76.715

Final: 87.084

*D) Mode*

Mid-Term: 77 83

Final: 97

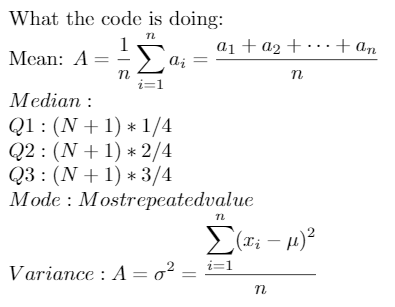
*E) Variance*

Mid-Term: 173.10577

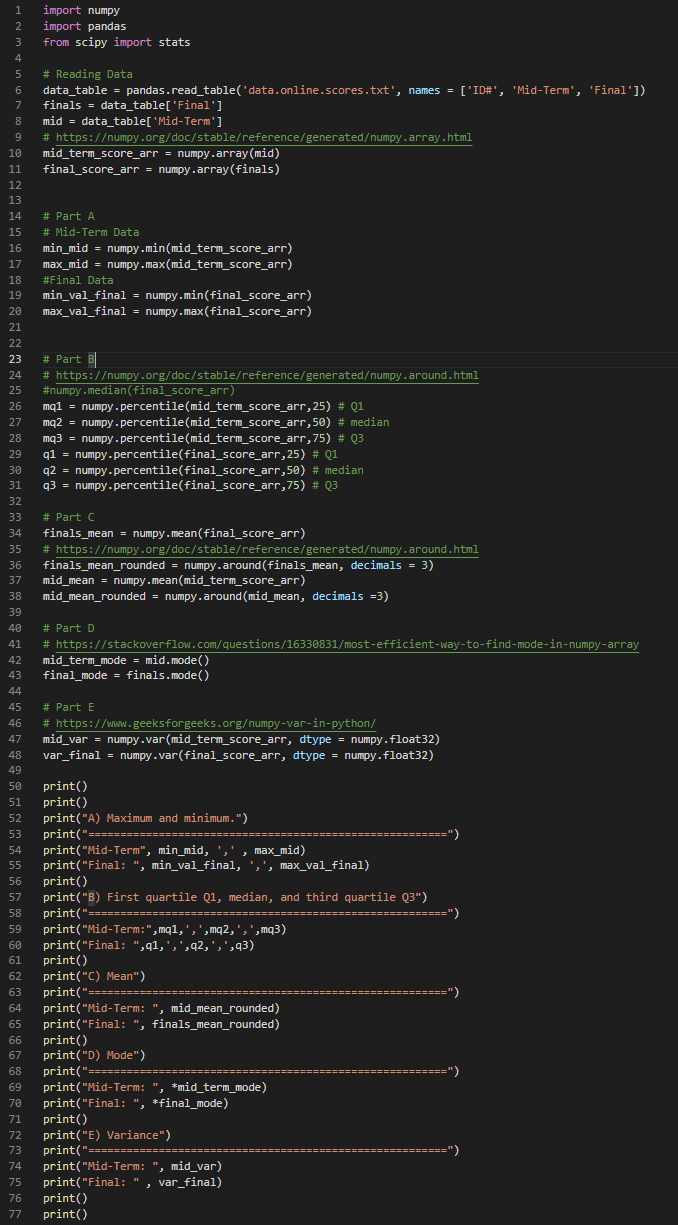
Final: 119.11294

**1.2 Explanation and Equations**

1. Numpy was used to find min and max
2. Numpy’s function percentile() was used. 25,50,75 for Q1, Q2, and Q3 relatively.
3. Numpy’s function mean() was used.
4. Numpy’s function mode()
5. Numpy’s function var() was used.

****

**1.3 Code**

****

**1.4 Terminal Output**

**Question 2.**

**2.1 Answers**

**A) Compute and compare the variance of midterm-original and midterm-normalized, i.e., the midterm scores before and after normalization.**

Variance MidTerm: 173.10577

Normalized Variance: 1.0

**B) Given an original midterm score of 90, what is the corresponding score after normalization?**

Mean: 76.715

Std-Dev: 13.156966785699508

Using v' = v - Avg / std-dev

Our score of 90 is normalized to 1.009731210573523

**C) Compute the Pearson’s correlation coefficient between midterm-original and finals-original?**

Pearson’s correlation coefficient: 0.544424742312412

**D) Compute the Pearson’s correlation coefficient between midterm-normalized and finals-original.**

Pearson’s correlation coefficient: 0.544424742312412

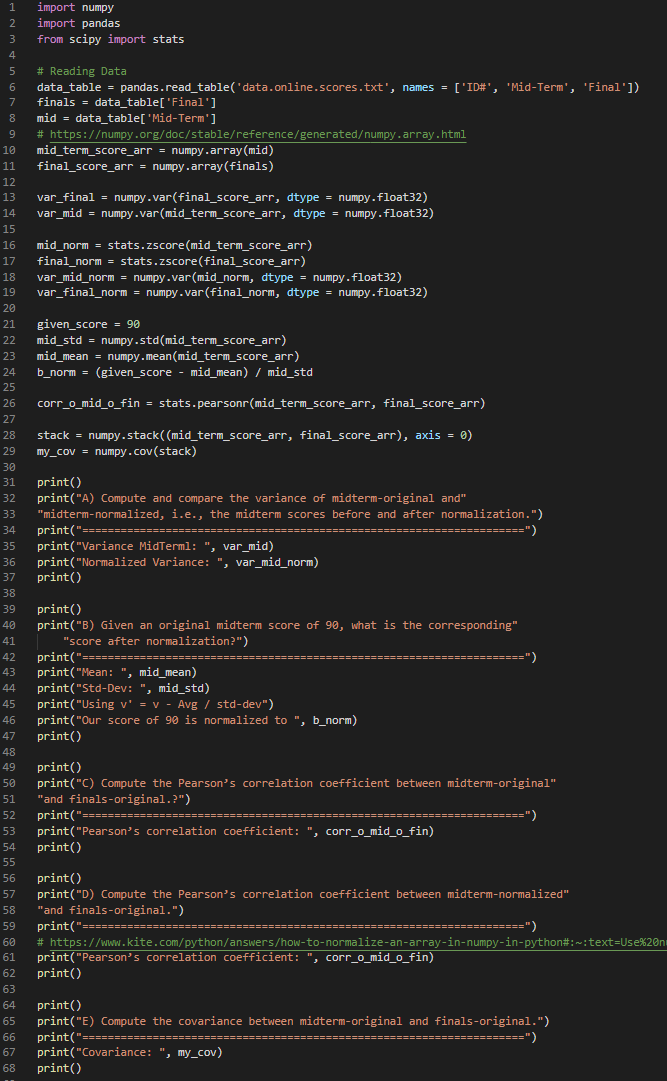
**E) Compute the covariance between midterm-original and finals-original.**

Covariance: 78.25419419

**2.2 Explanation and Equations**

1. Numpy var() was used
2. Scipy zscore() was used then using v' = v - Avg / std-dev, we normalize the score
3. Scipy pearsonr() was used
4. Scipy pearsonr() was used
5. Numpy’s stack() was used

**2.3 Code**

****

**2.4 Terminal Output**

**Question 3.**

**3.1 Answers**

**a) Each library has multiple copies of each book. Based on all the books (treat**

**the counts of the 100 books as a feature vector for each of the libraries), compute the**

**Minkowski distance of the vectors for CML and CBL with regard to different h values:**

i) 6152.0

ii) 715.3278968417211

iii) 170.0

**b) Compute the cosine similarity between the feature vectors for CML and CBL.**

Cosine similarity: 0.8414040256623079

**c) Compute the Kullback-Leibler (KL) divergence between CML and CBL**

KL(CML || CBL): 21.756826983085165

KL(CBL || CML): 52.49932615197238

**3.2 Explanation**

**3.3 Code**

**3.4 Terminal Output**

**Question 4.**

**4.1 Answers**

**a) Calculate the distance between the binary attributes Buy Beer and Buy Diaper**

**by assuming they are symmetric binary variables.**

Distance: 0.015691868758915834

**b) Calculate the distance between the binary attributes Buy Beer and Buy Diaper**

**by assuming they are symmetric binary variables.**

Jaccard Coefficient : 0.7317073170731707

**c) Compute the χ2 statistic for the contingency table.**

χ2: 2450.716326822006

**d) Consider a hypothesis test based on the χ2 statistic where the null hypothesis**

**is that Buy Beer and Buy Diaper are independent? Can you reject the null hypothesis?**

***at a significance level of α = 0.05? Explain your answer, and mention the degrees***

***of freedom used for the hypothesis test.***

If the value was more than 0.05 we could not reject the null hypothesis

The value based on info given was: 0.0

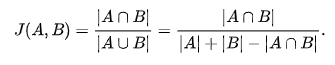
We were able to reject

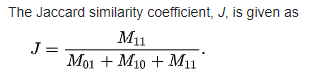
P value: 0.0

Deg of freedom: 1

**4.2 Explanation and Equations**

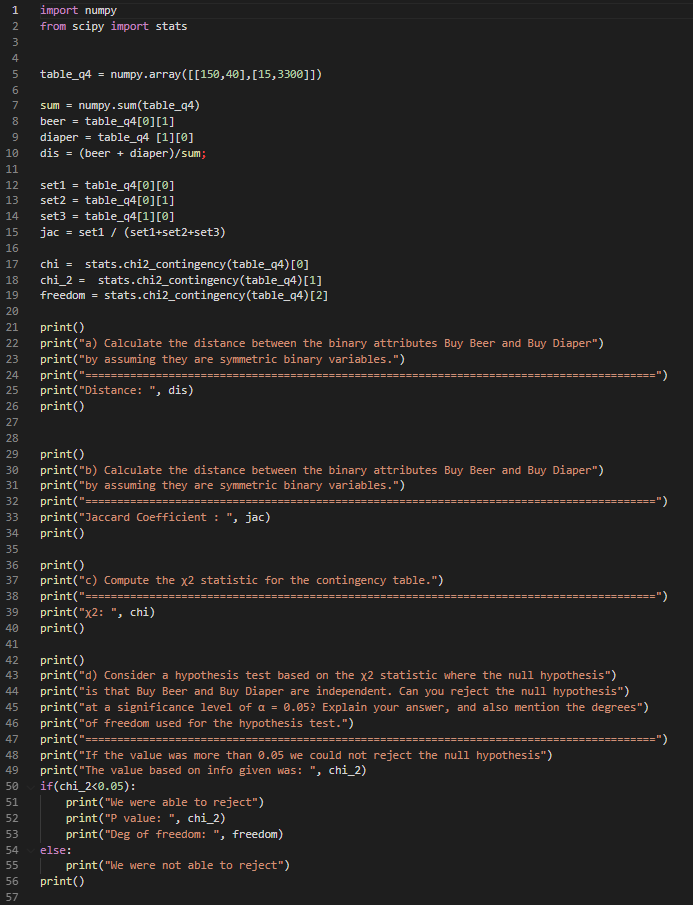
1. This was done without using any library. First, I added the items then I divided by the sum of the entire array.
2. The Jaccard Coefficient between buying beer and diaper was done with this equation.

Which translates to



1. Done with the use of scipy chi2\_contigency()
2. Done with the use of scipy chi2\_contigency()

**4.3 Code**

****

**4.4 Terminal Output**