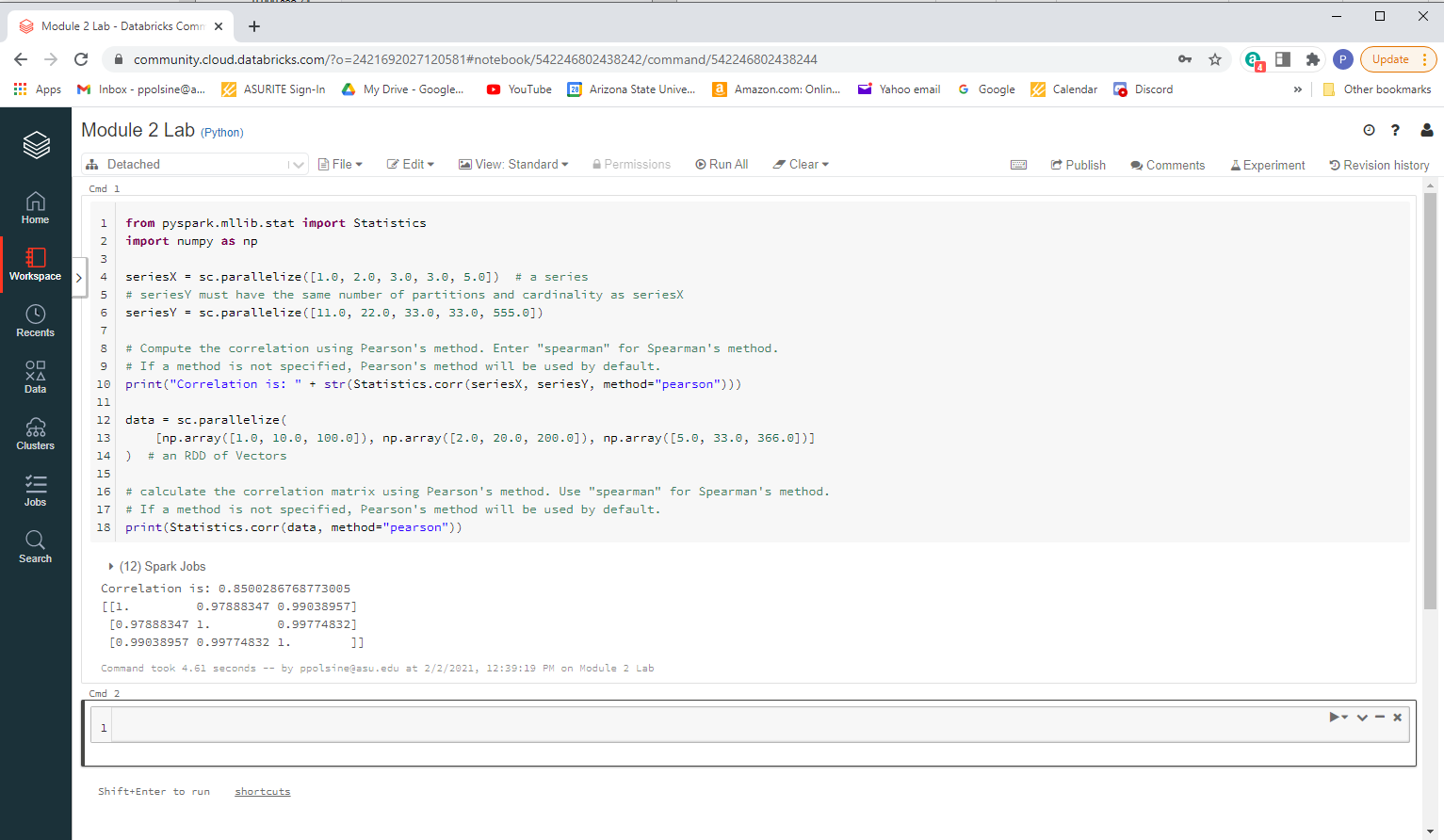
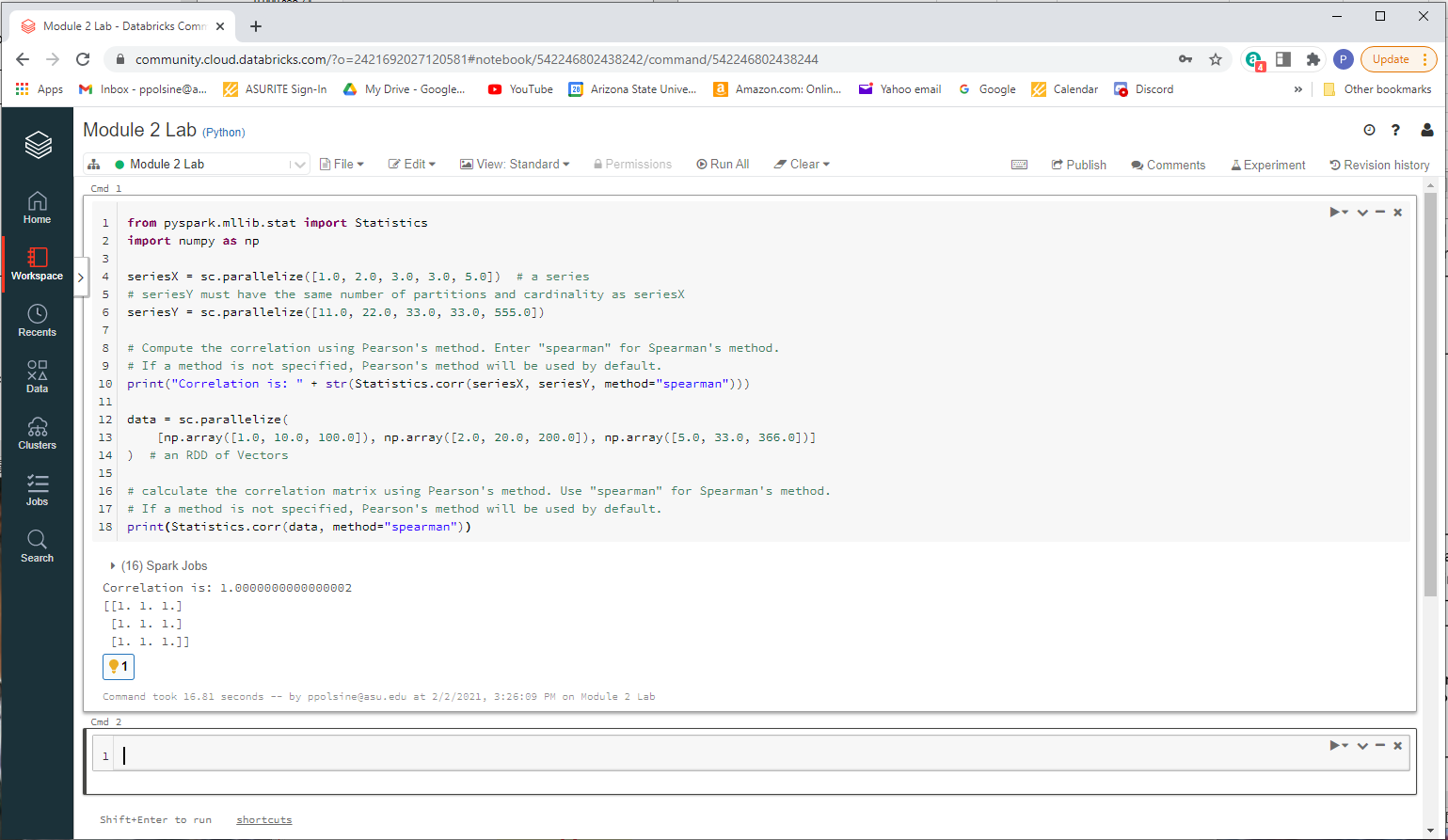
# Part I

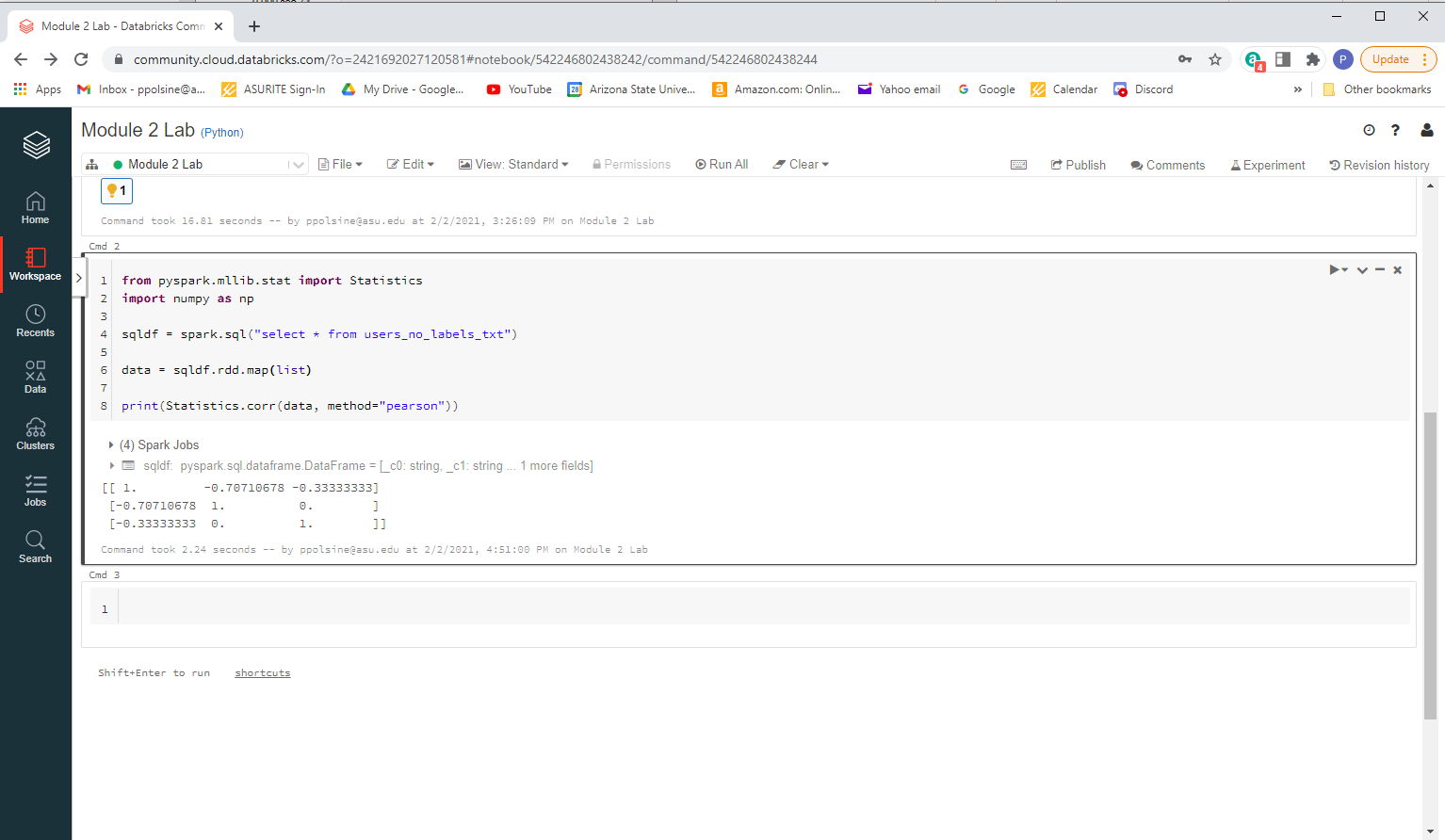


The code is running a similarity comparison on the arrays seriesX and seriesY and coming up with a high correlation of 0.85 out of a possible 1 for perfect correlation using Pearson’s method of corr(x,y)=covariance(x,y)/sd(x)\*sd(y) as its calculation, which it imported as Statistics from pyspark.mllib.stat. It then ran a correlation matrix on three features of three objects from an array named data, again using Pearson’s method.

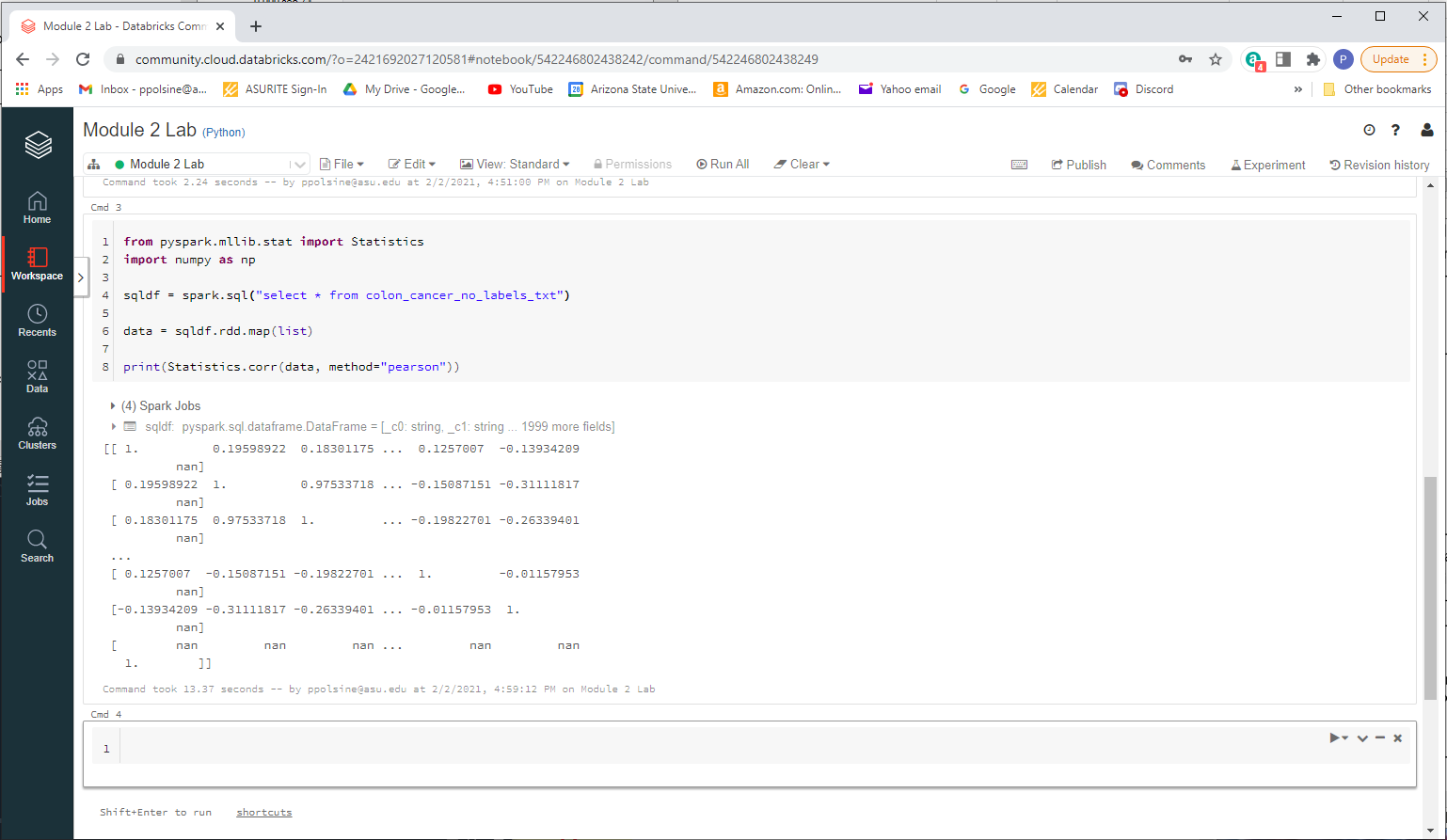


Using Spearman’s method (a nonparametric measure of rank correlation) the correlations are near perfect. Spearman’s correlation takes into account not only linear correlations where a change in one variable causes a proportional change in the other, but also monotonic relationships where the changes are not necessarily proportional but still dependent.

# Part II



# Part III



How many data points does this dataset contain? 62

How many attributes does each data point have? 2000

How many pairwise similarities should be computed? 124,000