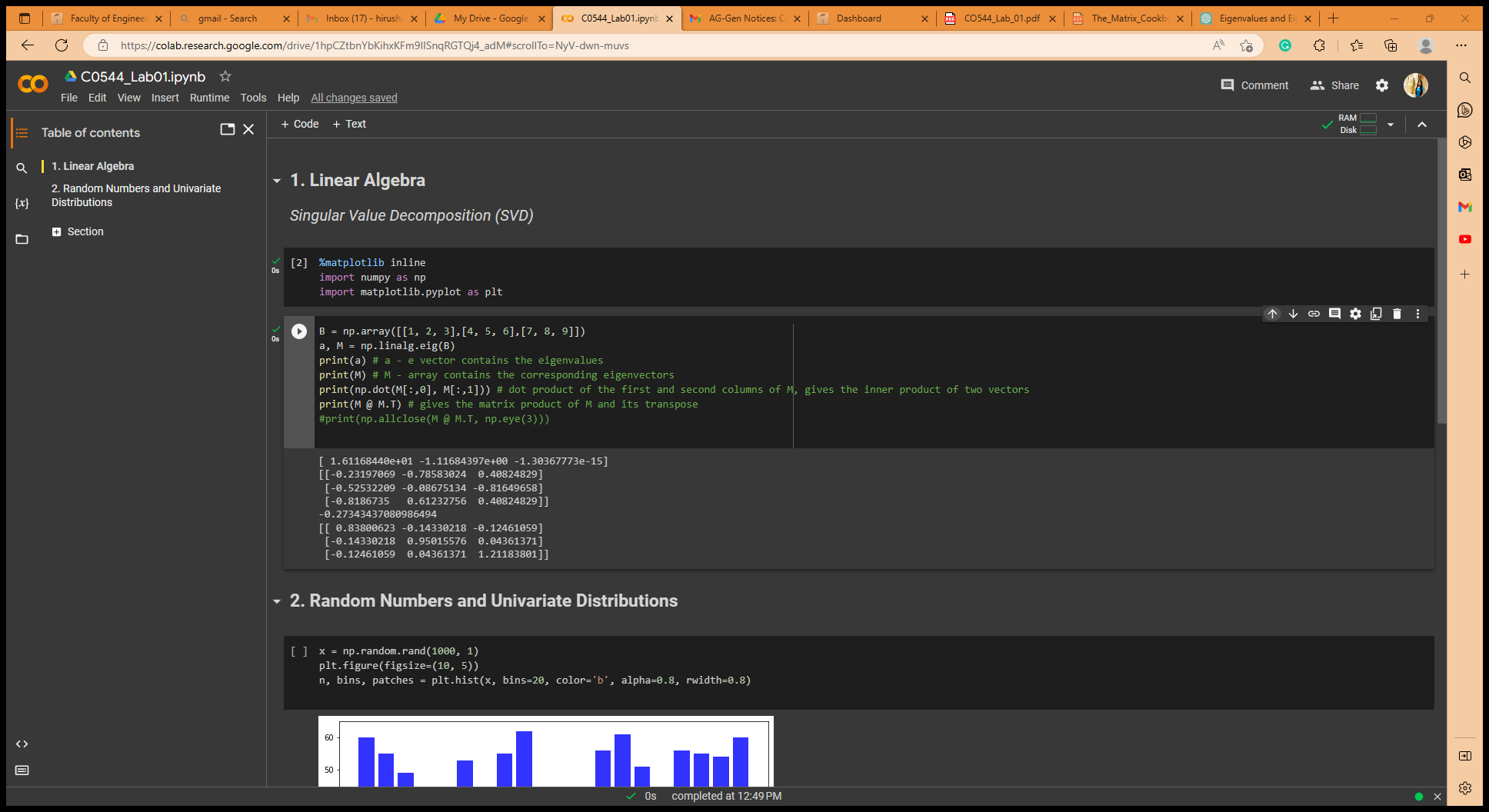
**CO 544 Machine Learning and Data Mining**

**Lab 01**

E/18/323

SEEKKUBADU H.D.

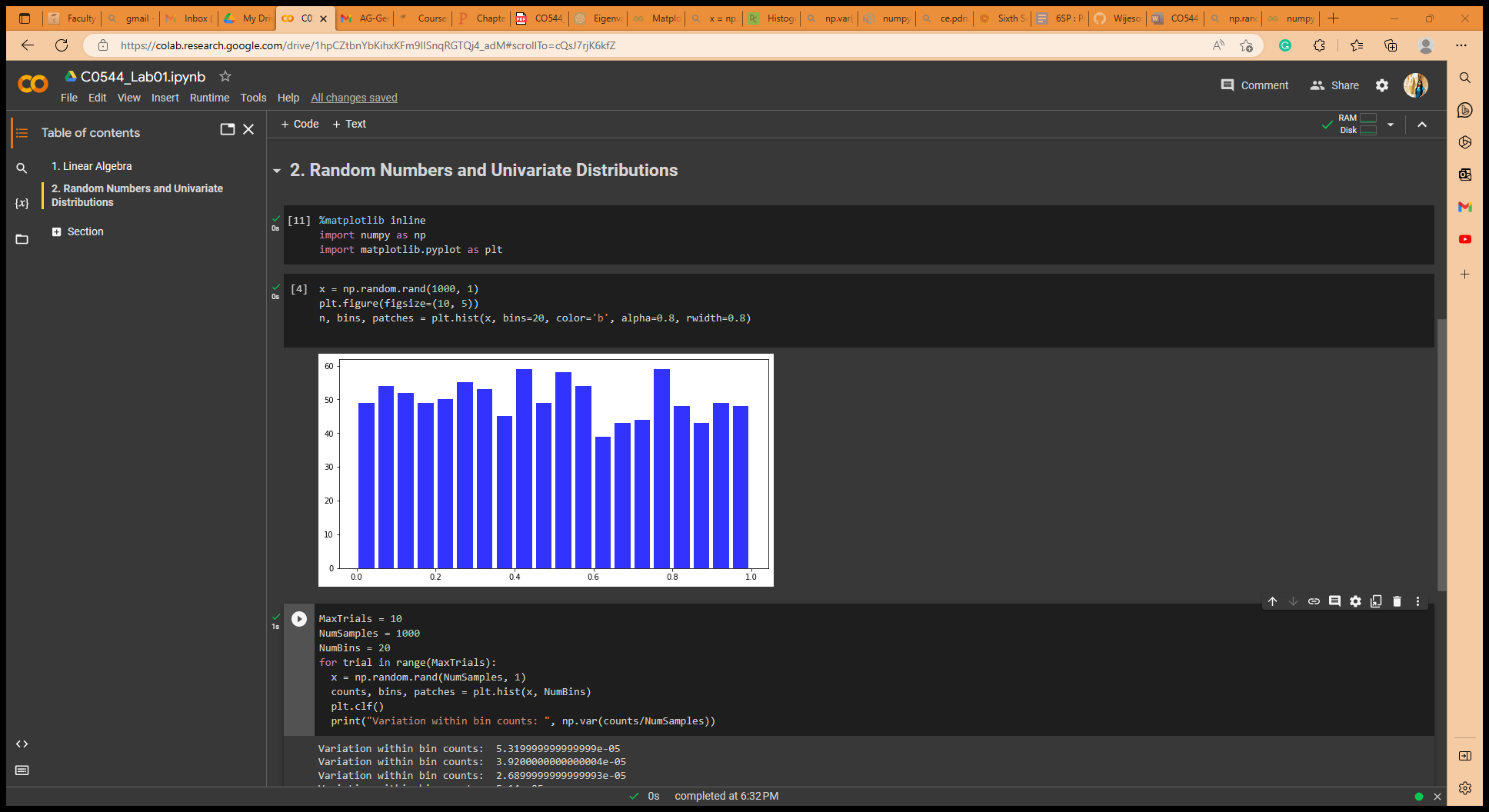
1. Linear Algebra - Singular Value Decomposition (SVD)



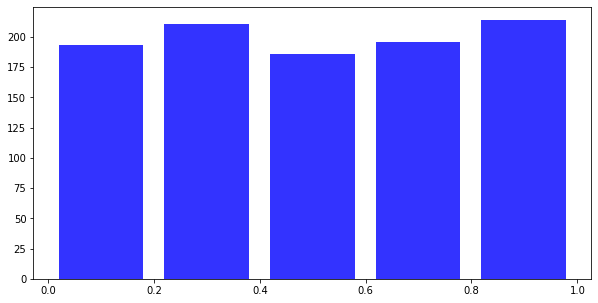
print (M @ M.T) – this gives the matrix product of M and its Transpose (M - array contains the corresponding eigenvectors). Results is a square matrix of 3x3.

M @ M.T, this would be an identity matrix. If we round off the values, we can observe identity matrix which is expected. Since the eigenvectors of a matrix form an orthonormal basis for the space, and so their matrix product with their transpose should give the identity matrix.

1. Random Numbers and Univariate Distributions



**Effect of the number of bins**- (when the number of samples is a constant)

Chart, histogram

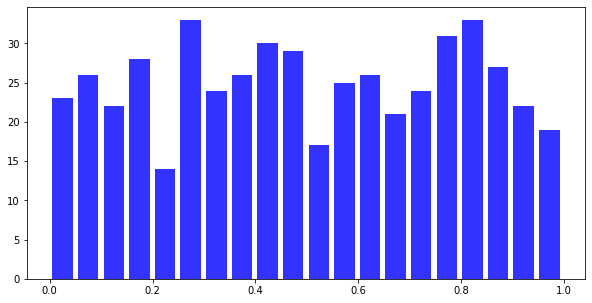
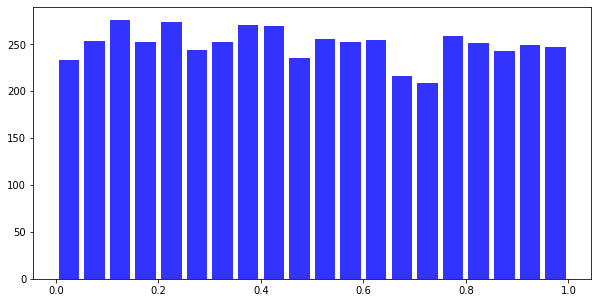
Description automatically generated

Num of Samples = 1000, bins=100

Num of Samples = 1000, bins=50

The bins argument specifies the number of bins to use. Select a bin width that generates the most faithful representation of data. Increasing the number of bins will increase the resolution of the histogram and produce a detailed curve but can also make it harder to interpret. Decreasing the number of bins will make the histogram smoother but can also hide important details.

**Effect of the number of samples**- (when the number of bins is a constant)

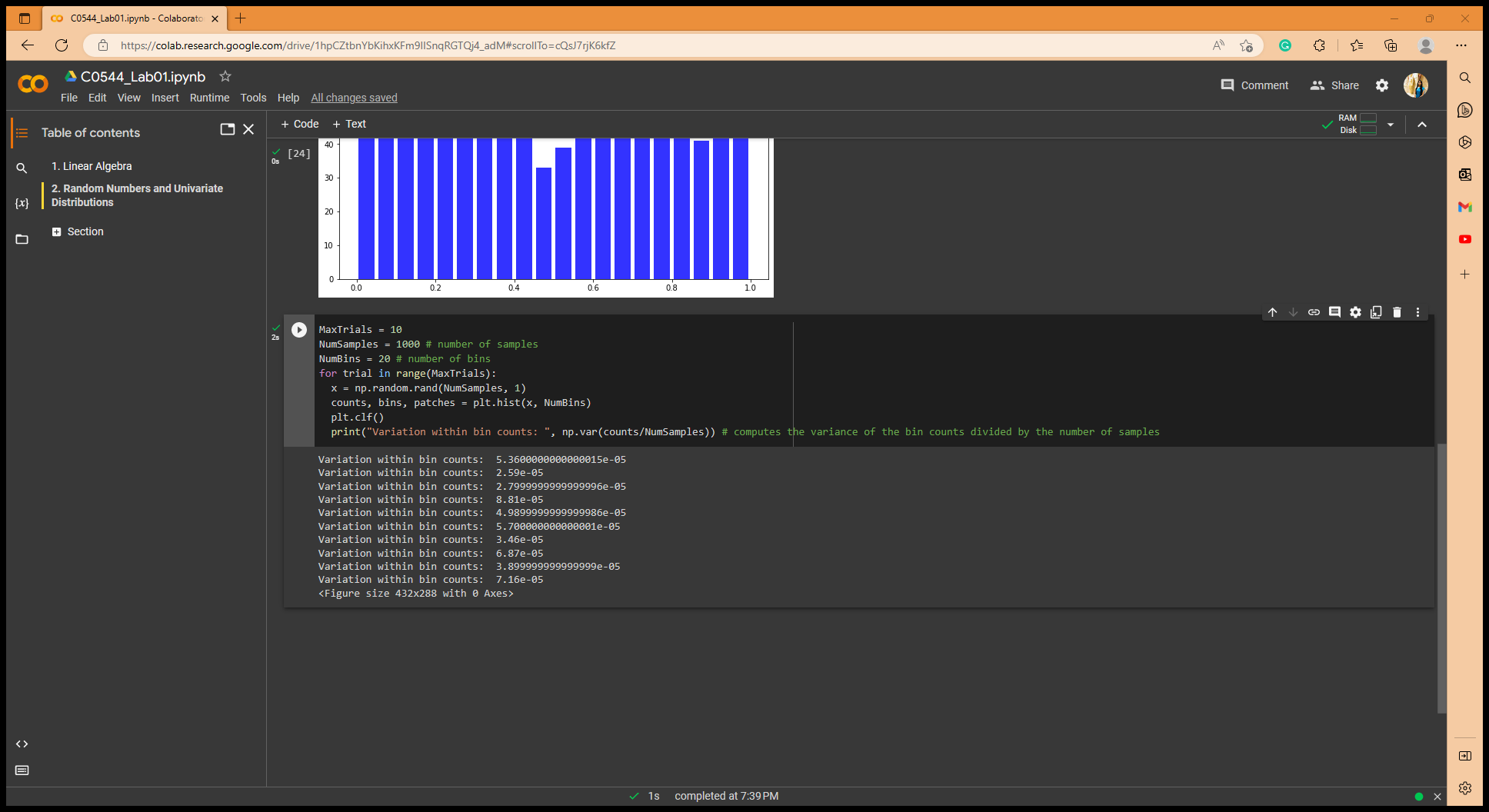


Num of Samples = 5000, bins=20

Num of Samples = 500, bins=20

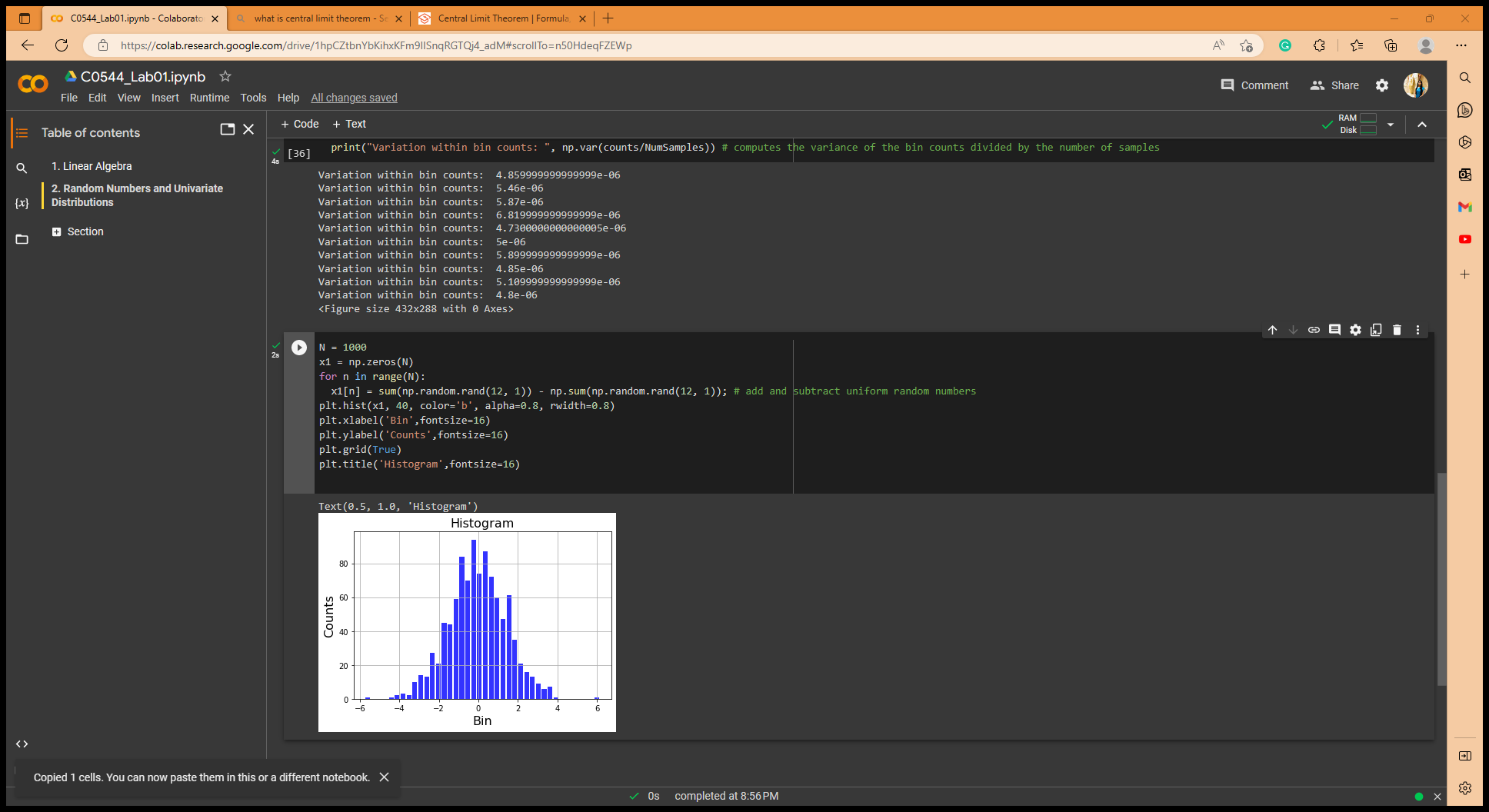
Number of samples affects the accuracy of the histogram for the Gaussian distribution. If the number of samples increases, the histogram becomes more accurate and better represents the distribution. When the number of samples decreases the histogram may not accurately represent the distribution.

* Repeat the above with 1000 random numbers drawn from a Gaussian distribution of mean 0 and standard deviation 1 using x = random,randn(1000, 1);.



When increase the number of samples, the variation within the bin count is reduced. While decreasing the number of samples, variation increases.

When increase the number of bins, variation within the bin count is reduced. While the decreasing the number of bins, variation increases.

* Try the following where we add and subtract uniform random numbers:

Chart, histogram

Description automatically generated

**We can observe**,

The histogram is roughly centered around zero, this shows the mean of the difference between two sets of random numbers is zero. Therefore, this distribution approximately normal distribution with a peak at zero and tails that taper off towards both sides.

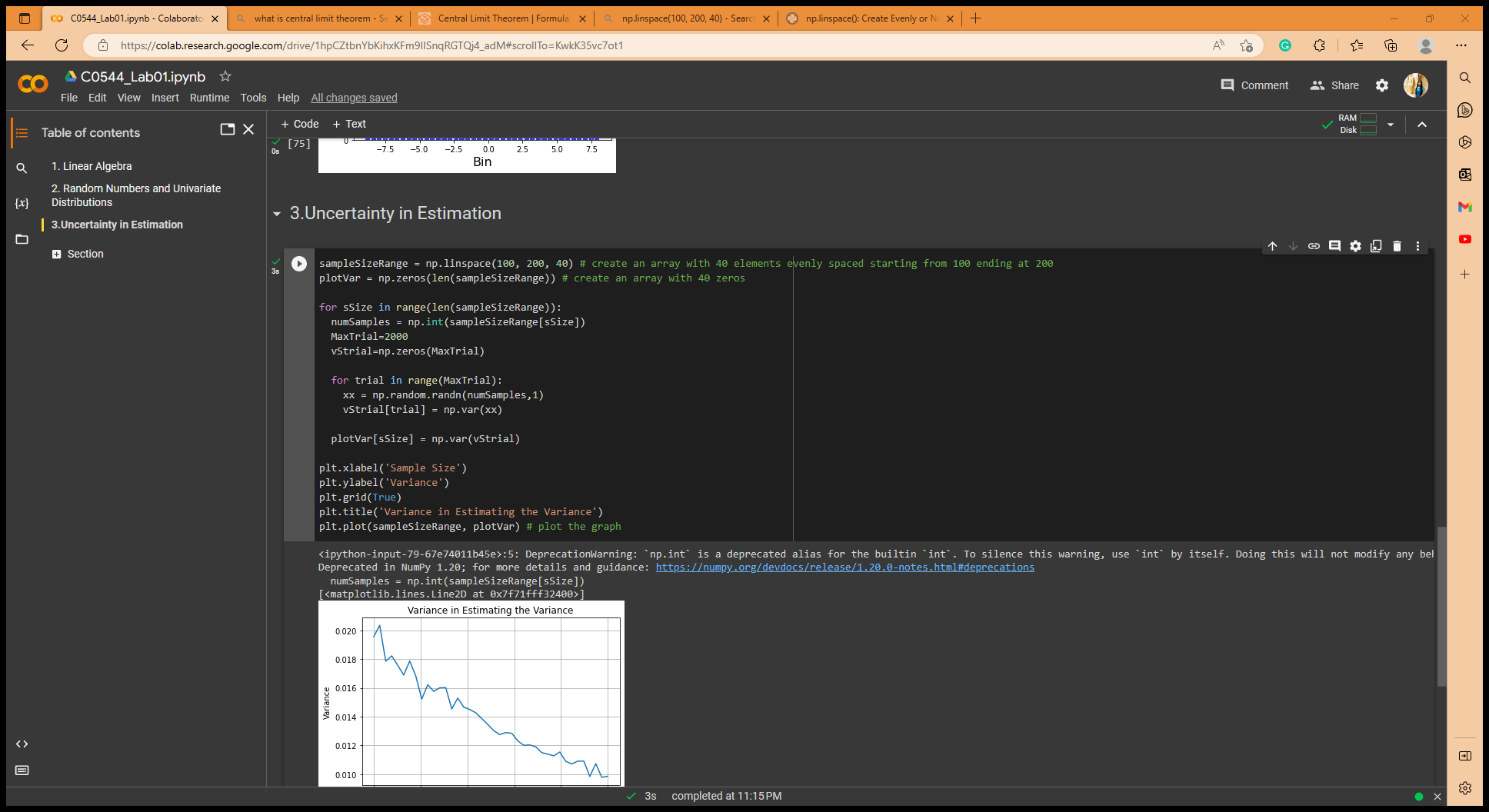
**When change the number of numbers add/subtract?**

The spread of the histogram changes. If we decrease the number of numbers add/subtract (12->5), the spread range become smaller. Because of the variability of the sum of numbers, if we increase the number, the spread range become larger. But at both cases histogram roughly centered around zero. (Still normal distribution)

**Is there a theorem that explains it?**

Central Limit Theorem. This theorem states that if the sample size is sufficiently large (>30), samples are independent and identically distributed and the distribution has finite variance, then the sampling distribution of mean will follow a normal distribution. In here we take 1000 size of sample, (>30) and add random variables together, the distribution will approach a normal distribution.

1. Uncertainty in Estimation



Chart, line chart

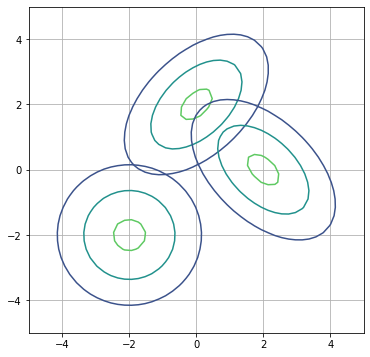
Description automatically generated

This graph shows,

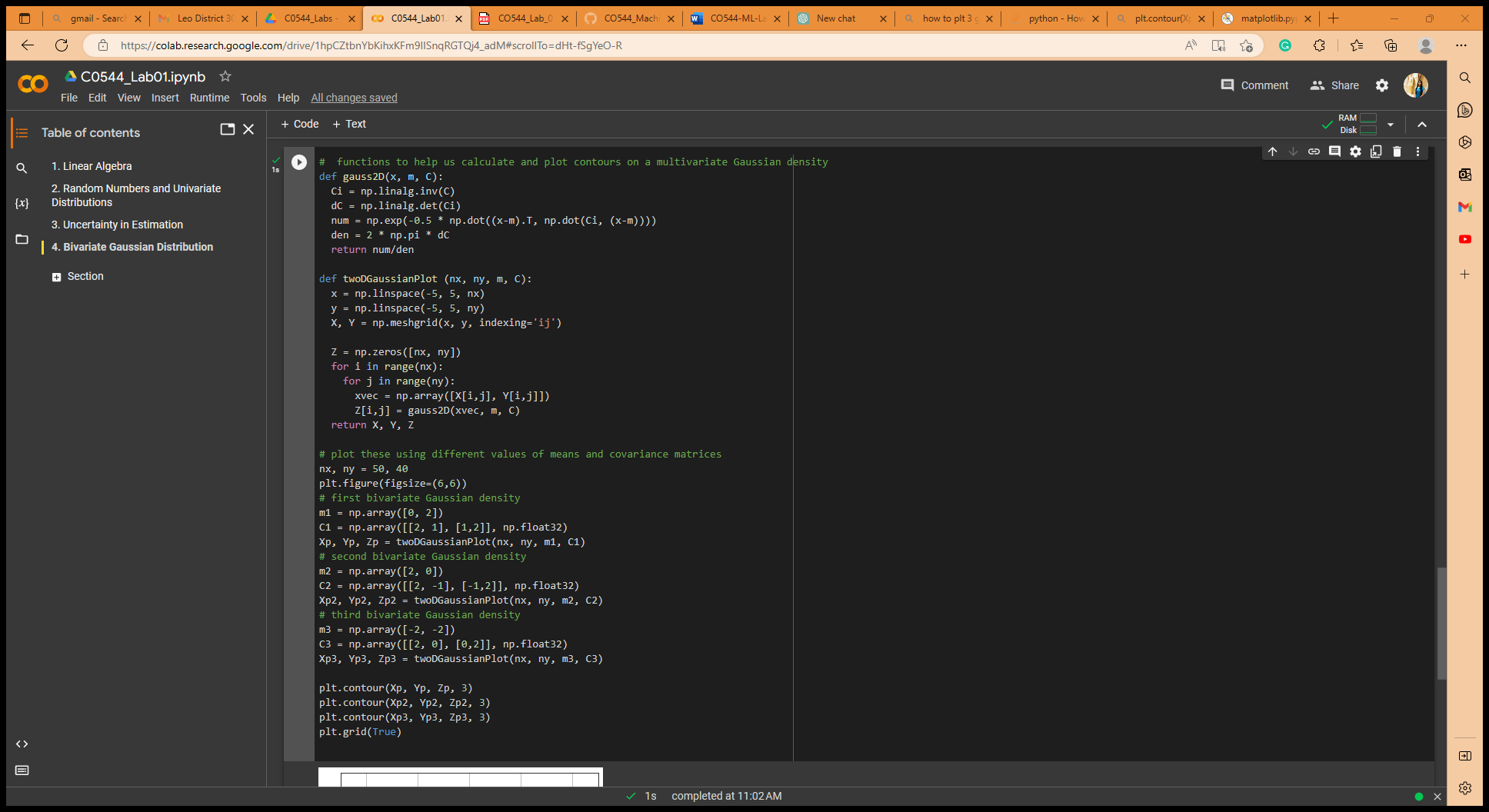
Increase the sample size – the variance become smaller.

Decrease the sample size – the variance become larger.

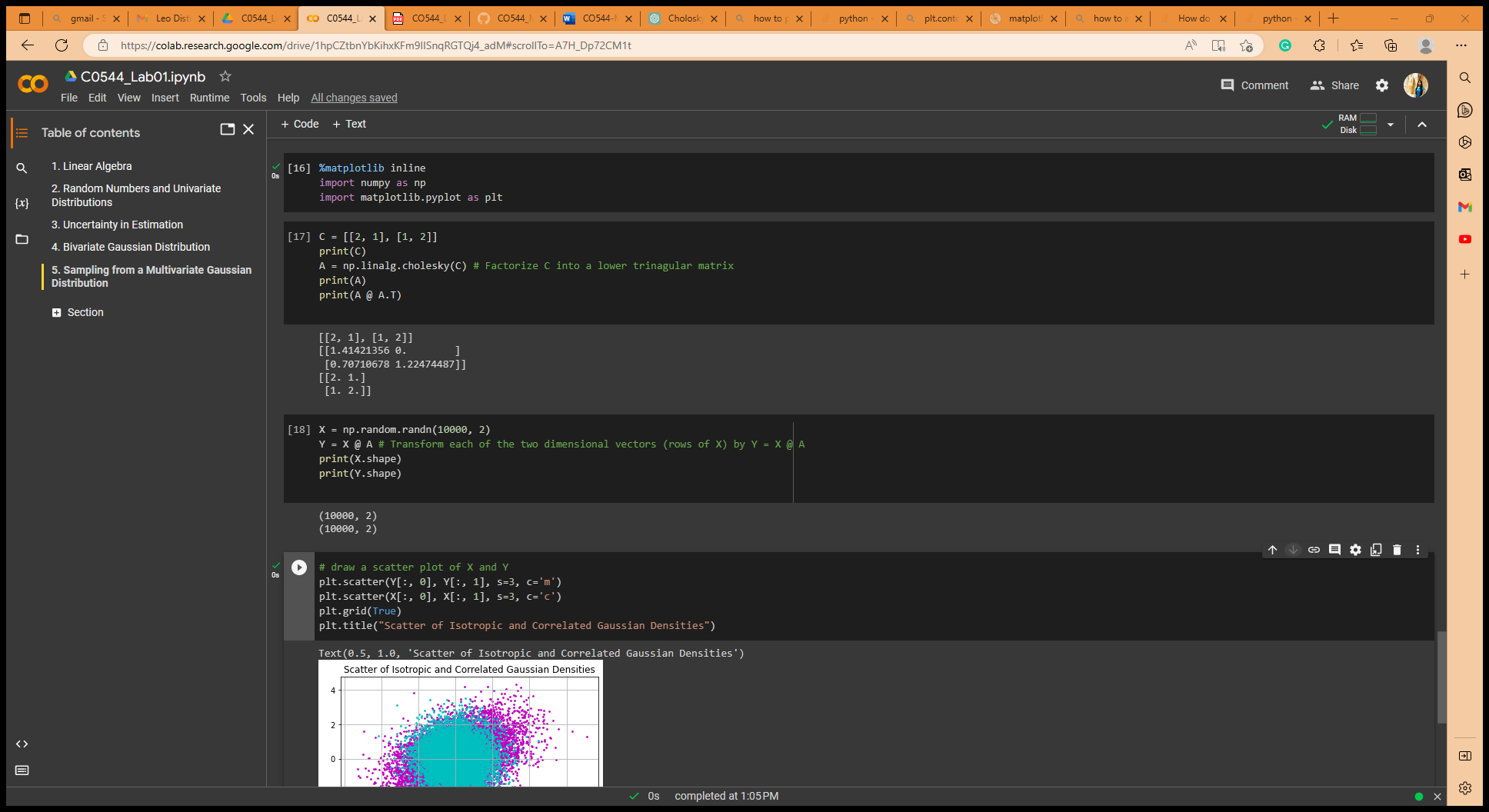
1. Bivariate Gaussian Distribution

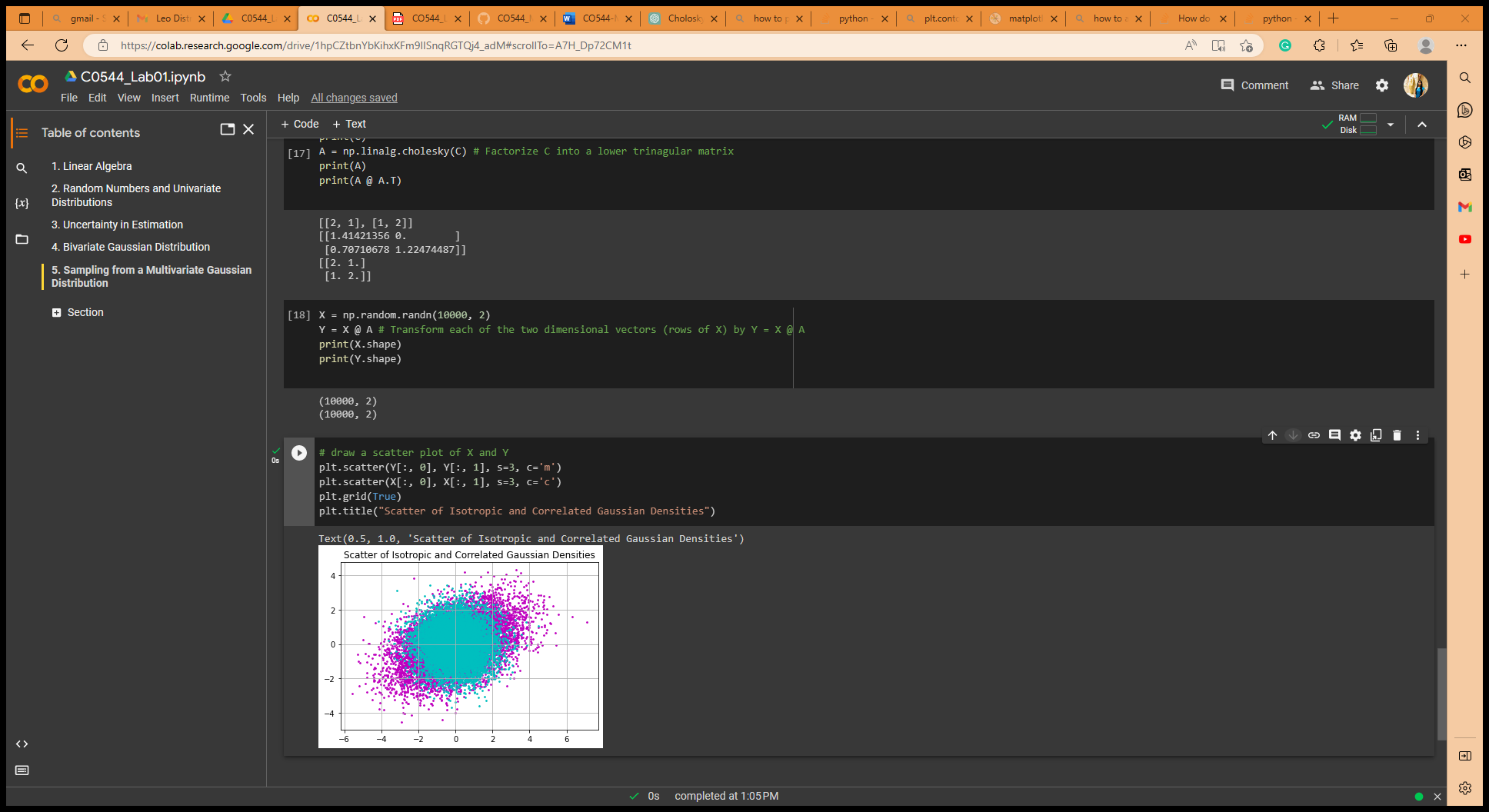


The contour lines represent regions of equal probability density.

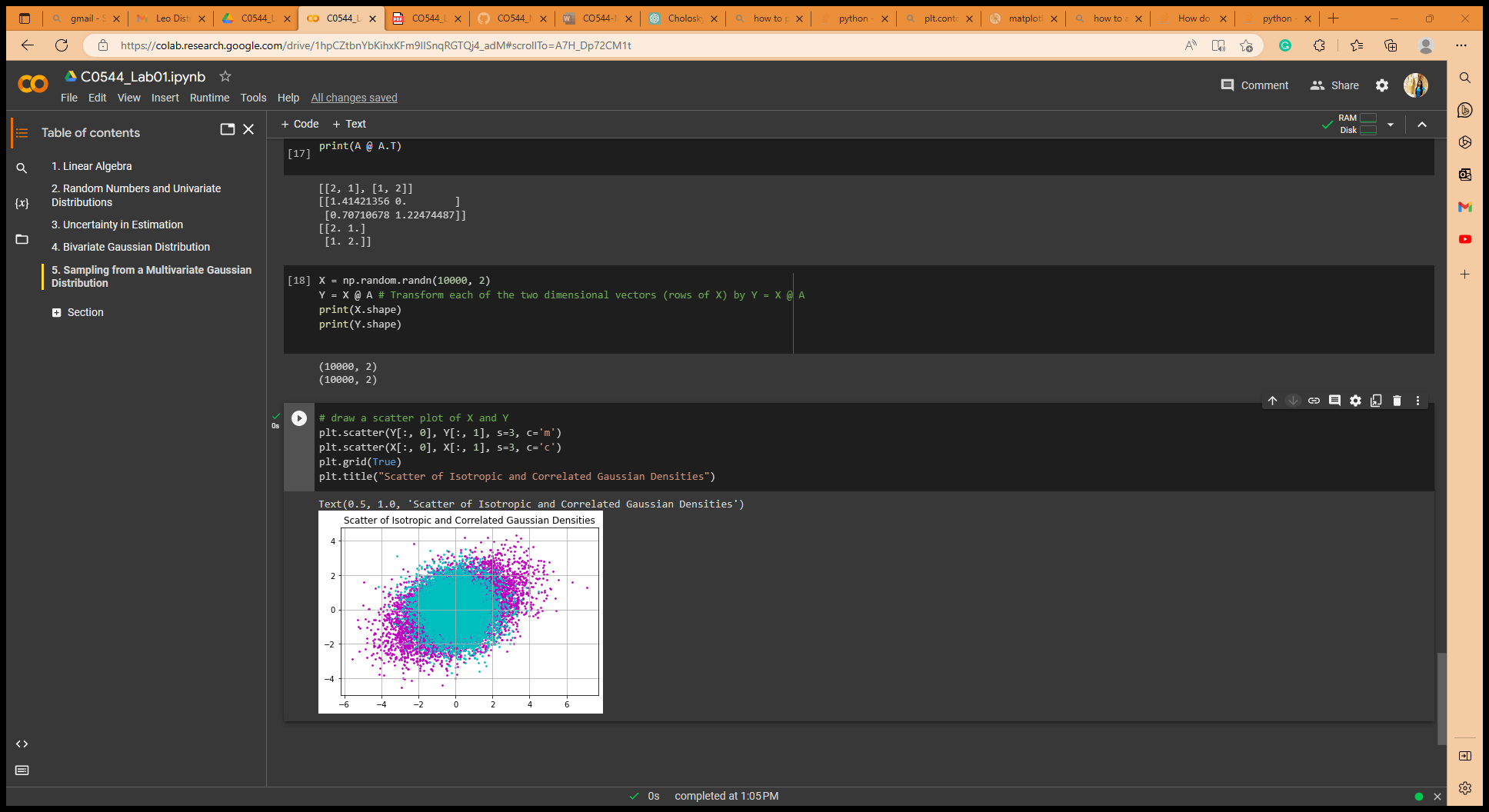


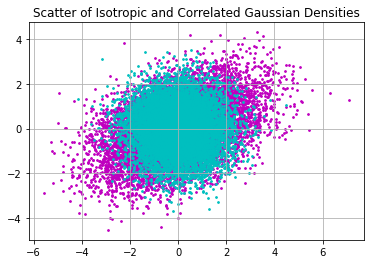
1. Sampling from a Multivariate Gaussian Distribution



A @ A.T = C, above answer confirms that the factorization is correct by multiplying A and its transpose A^T and obtaining C.

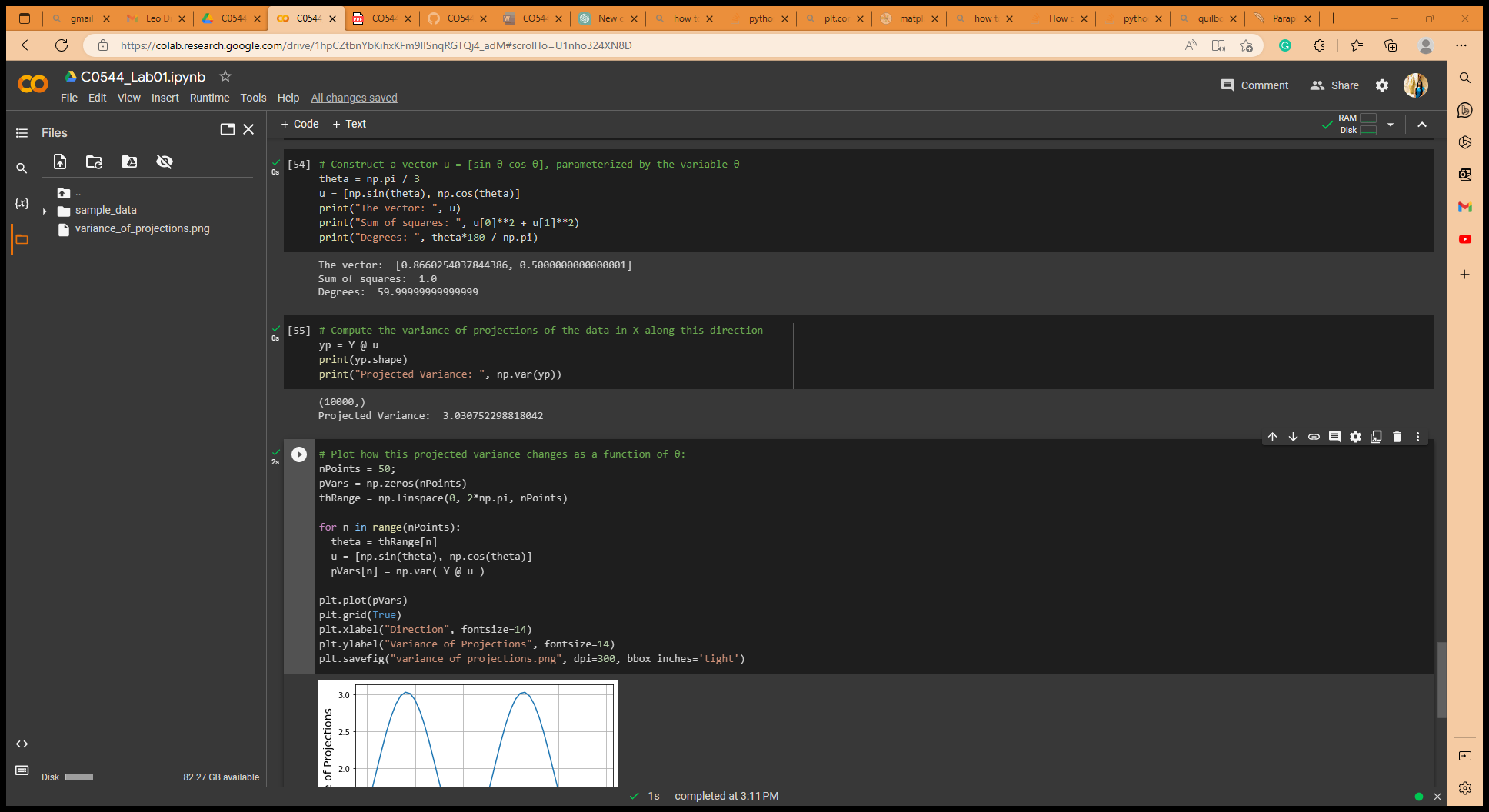
Since A is a lower triangular matrix, Y = X@A gives a correlated Gaussian random variables with covariance matrix C.

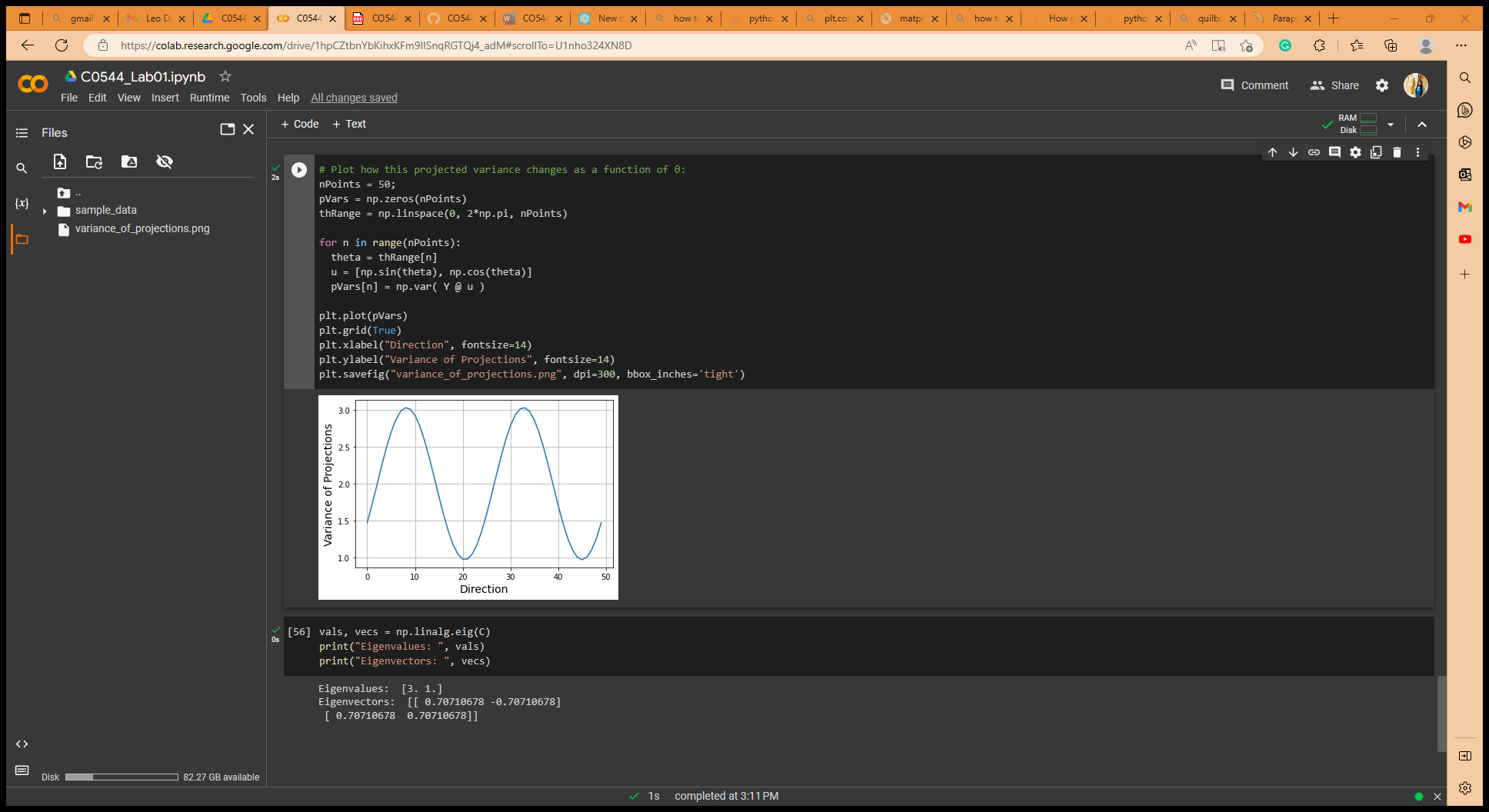




The scatter plot of X represents an isotropic Gaussian density, in which the two variables are independent and have the same variance, whereas the scatter plot of Y represents a correlated Gaussian density, in which the two variables are not independent and have different variances. The plot confirms that the points in Y are correlated and form an elliptical shape, whereas the points in X are randomly scattered with no correlation. As a result, the scatter plot of Y is more elliptical than the scatter plot of X.

1. Distribution of Projections



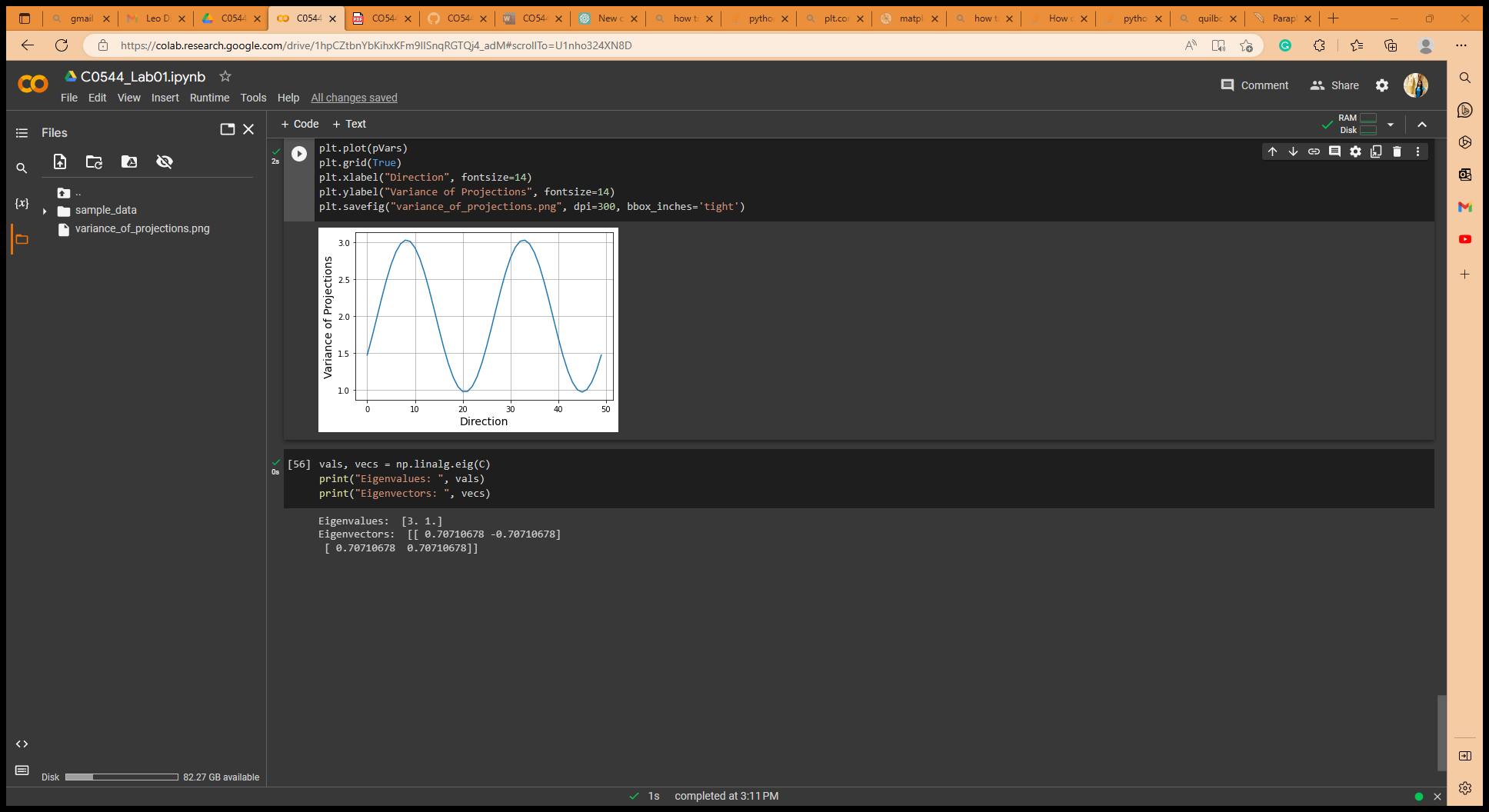
To compute the variance of projections of the data in X along this direction, take the dot product of each row in the matrix Y with the vector u. (In above screenshot)

Chart, line chart

Description automatically generated

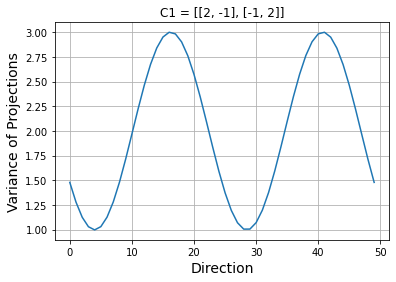
From the plot, we observe that the variance of projections is highest when θ is aligned with the direction of maximum variance in the data, and lowest when θ is orthogonal to that direction.

By calculating the eigenvectors of the covariance matrix,



The eigenvectors of the covariance matrix, correspond to the directions of maximum variance in the data. This confirms that the direction of maximum variance in the data corresponds to the eigenvector of the covariance matrix with the largest eigenvalue. We can see that the maximum eigenvalue corresponds to the direction of maximum variance, which is aligned with the x-axis or y-axis, and the minimum eigenvalue corresponds to the direction of minimum variance, which is aligned with the diagonal axes.

How does what you have done above differ for C = [[2,-1],[-1,2]], below graph shows the difference. This graph can obtain repeating the above process for the new covariance matrix.



Appendix: <https://colab.research.google.com/drive/1hpCZtbnYbKihxKFm9IISnqRGTQj4_adM?usp=sharing>