1. Generate two different sets of random variables (i.e gaussian, uniform) , then plot both as clusters (mind that the dataset should form a cluster).

   i) Visualize the cluster using matplotlib. (use different color code for each cluster.)

   ii) Calculate the expected value and variance of the each distribution of the dataset.

   iii) Determine the centroid of each cluster, a simple way could be, you average out the x, y coordinates and find the mean of all the points in a cluster. But you're free to take your own idea.

   iii) Calculate the absolute difference between point and the centroid. (This should be from cluster A points to cluster B centroid and cluster B points to cluster A centroids)

   iv) is the distance calculated symmetric? (Why/Why not)

2. Implement a Naive Bayes algorithm from scratch. Gaussian Naive Bayes preferably, you're free to use any dummy data you like to showcase your implementation. You're free to use mathematical libraries like Numpy, but you can't directly use the libraries to use the algorithm itself. (For example, GaussianNB is available in scikit-learn, you can't directly use it.) For your ease, this resource may help: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

3. Consider the following overdetermined system of linear equations:

   $0.25x1 + 0.03x2 = 0.97$

   $-0.37x1 + 0.17x2 = 0.48$

   $1.17x1 + x2 = 2.20$

   $-1.09x1 - 0.17x2 = -1.19$

   $x1 + 1.19x2 = 1.73$

i) First visualize the system of linear equations. (the image should be attached with the solution)

ii) We would like to find the best possible solution for this linear system

Method I

Minimize the sum of squares of the residual errors, for example:

0.25x1 + 0.03x2 - 0.77 = r1

.

.

.

c1.x1 + c2. x2 - c3 = r_n

And the sum of residual squared would be:

S = $\sum_{i=1}^{n} r_i^{\ 2}$

Find the values of our parameters x1, x2 that minimizes the S most.

Write your code in C++ that accepts the array of linear equations as string, then solve the system with most suitable estimation. Here the estimation with the lowest sum of squared error.

Method II

Solve the above overdetermined system of linear equations using pseudo-inverse.

You’re free to use Python or C++. Compare your answers from Method I and Method II and give some reasoning on your understanding, why they're similar or different if they're.

4. 4 couples enter a cafe which has 3 floors in which multiple tables are available. Each person decides on which floor to have their coffee regardless they may have come together. Assume that the decision of different customers is independent, and that for each customer, each floor is equally likely. Find the probability that exactly **one** customer has his/her coffee on the 1t floor.

5. Let's try to give intuitive reasoning.

   i) $f(x) = \sqrt[3]{x}$ is f(x) a continuous function? If so, why is that? If it's not why? Please describe.

   ii) g(x) = $-x^6 - x^4 - 13x$ Is g(x) Convex function? Or Concave function? Or none of them exactly. Please give your intuitive reasoning.

6. Implement a simple "named entity extraction" algorithm

   Problem Statement:

   - Download the data from: https://drive.google.com/file/d/1Ie5X2wJvMf_J8SLL4uMfymxnn1tMM-Zc/view?usp=sharing
   - Create a method that takes the text as input and returns the cleaned text using preprocessing techniques in NLP.
   - Once you have a cleaned text, extract the entities like "name", "organization", "location" from the text without using any **libraries** try building a simple named entity extraction algorithm to extract the entities.
   - Save the data in a CSV with cleaned text and its extracted entities.

Example - csv column headers: text, cleaned_text,extracted_entities

If the text is:

“American Airlines   said it would launch a direct flight to Bengaluru from Seattle :D, home to Amazon and Microsoft https:xyz.com.”

text	cleaned_text	extracted_entities
American Airlines said it would launch a direct flight to Bengaluru from Seattle :D, home to Amazon and Microsoft https:xyz.com.	*American Airlines said it would launch a direct flight to Bengaluru from Seattle, home to Amazon and Microsoft.*	(American Airlines,ORGANIZATION),(Bengaluru ,LOCATION)(Seattle,LOCATION)(Amazon,ORGANIZATION)(Microsoft,ORGANIZATION)

*Note: Perform the above tasks in a single script in Python and save it as **”task_entities.py”** . (Please make comments to the code) and share the output in .csv format.*

7. How would you remove outliers when trying to estimate a flat plane from noisy samples? Implement in C++ (Preferred) / Python.

Note: While you’re encouraged to do it, Number 2 is optional.