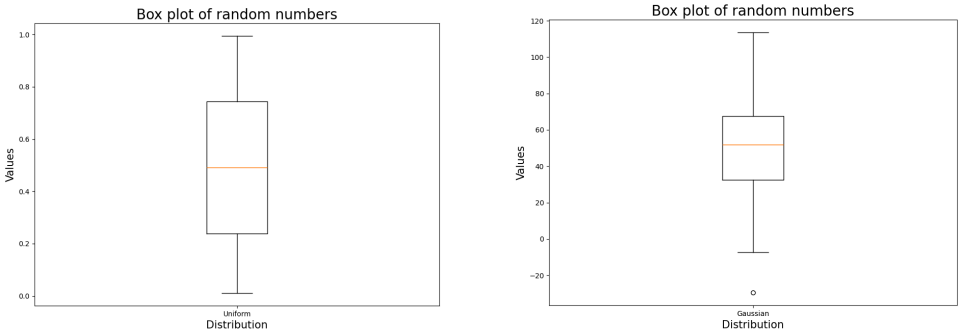
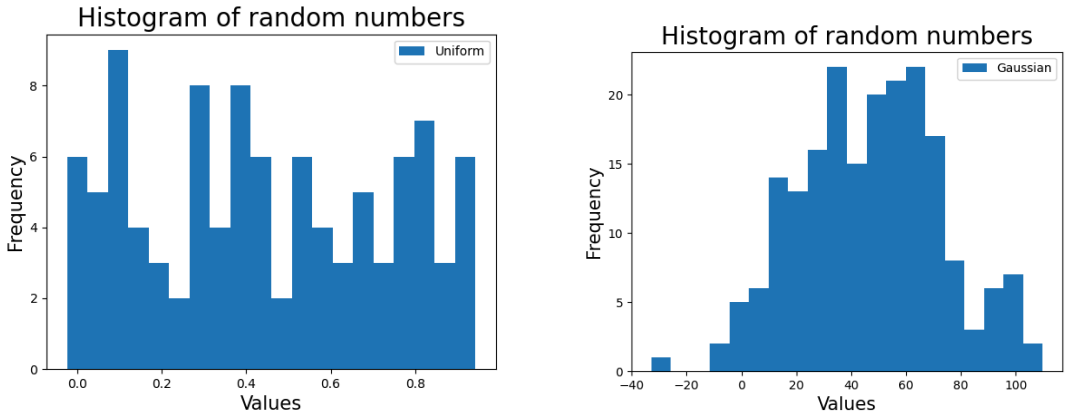
Code: https://github.com/elizaan/Viz-Scientific-Data-HWs.git (HW1 folder)
Methods: For each part, write code in the main.py file where all the functions for necessary plots are written, and the main function calles them along with loading appropriate data**.**
Tools: Python, Matplotlib, Numpy, Pandas, Scipy, NitBabel

Part 1 (in part 1 folder inside hw1)

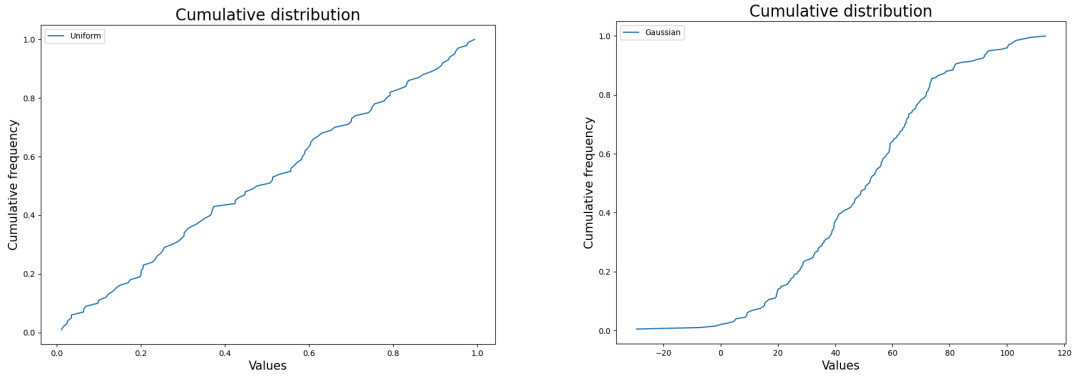
1. Box plot (a) Uniform (b) Gaussian Distribution



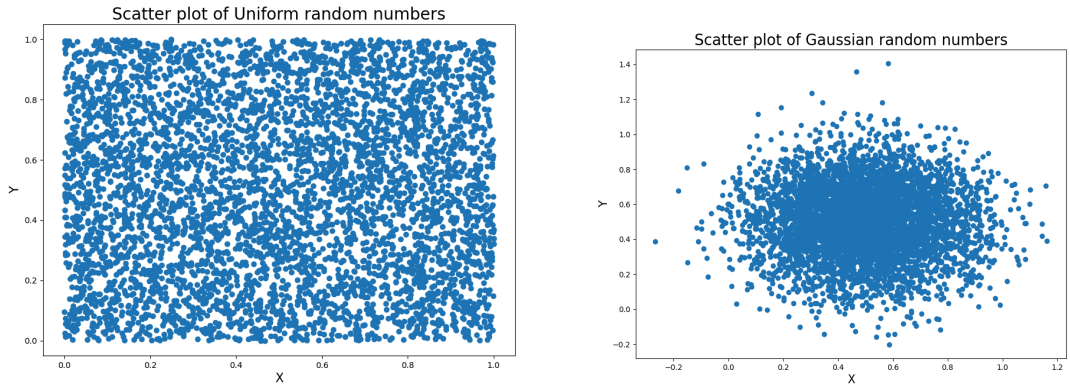
2. Histogram (a) Uniform (b) Gaussian Distribution



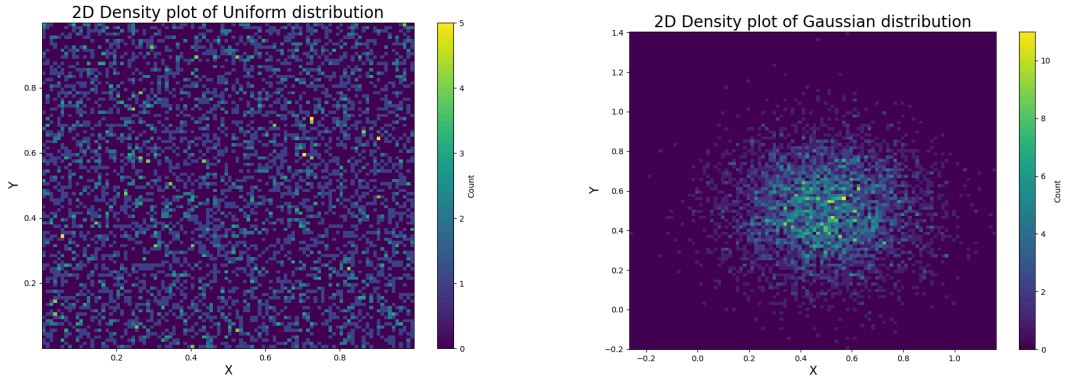
3. Cumulative line graph (a) Uniform (b) Gaussian Distribution



4. A. Scatter Plot (a) Uniform (b) Gaussian Distribution



B. Density plot (a) Uniform (b) Gaussian Distribution



C. Contour plot (a) Uniform (b) Gaussian Distribution

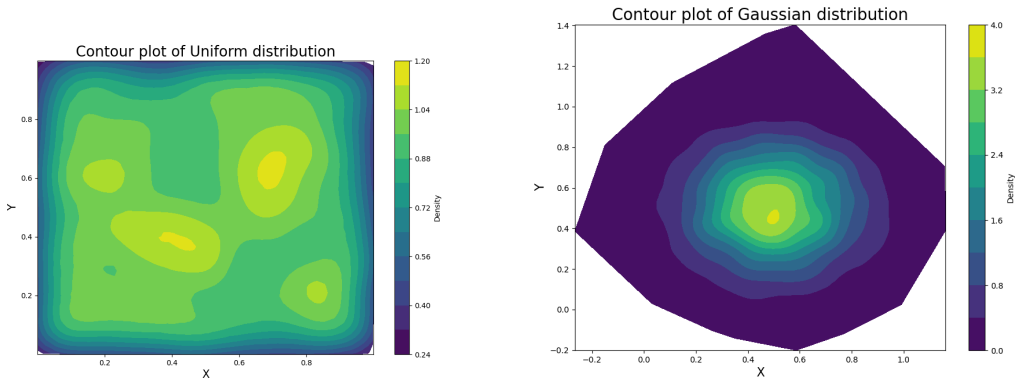


Figure 1: Plots for Part 1 (1) Box plot, (2) Histogram, (3) Line Graph, (4) Scatter plot, Density plot, and Contour plot

Part 2 (in part 2 folder inside hw1)

1. Based on the visualization and analysis (Figure 2), here are the key trends in the NOAA temperature data:

Overall Warming Trend: The data shows a clear long-term warming trend from 1880 to the present, with temperatures increasingly deviating above the 1901-2000 baseline period.

Color Pattern Evolution:

- The early years (late 1800s to early 1900s) show predominantly blue bars, indicating cooler temperatures relative to the baseline
- There's a transition period in the mid-20th century
- Recent decades show predominantly red bars, indicating consistently warmer temperatures

The warming trend appears to be accelerating, with larger positive anomalies becoming more frequent in recent decades.

Variability: While there is year-to-year variability, the overall trend shows a consistent warming pattern, particularly since the 1980s.

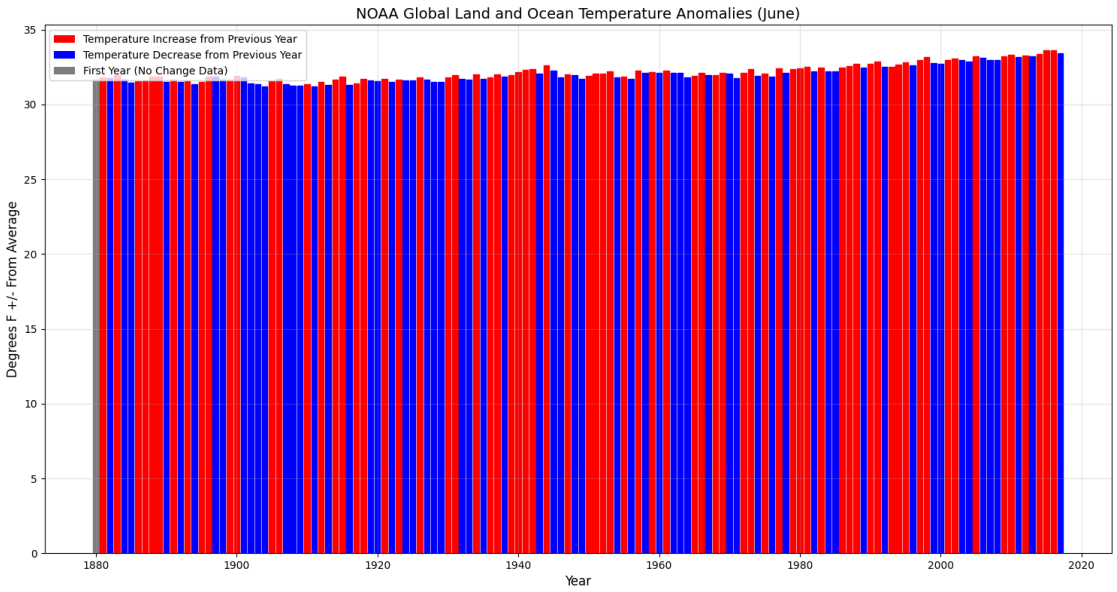


Figure 2: Temperature bar plot

2. Cereal Comparisons: 3 cereals are marked as legends

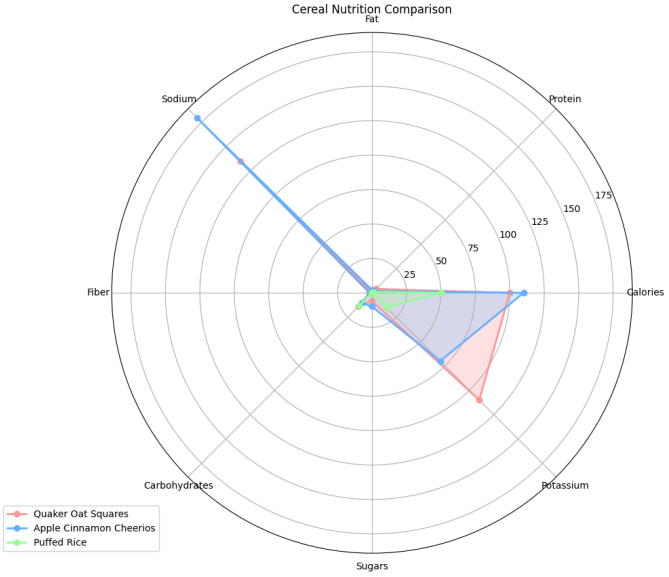


Figure 3: Cereal Nutrition Comparison

3. A. I have chosen airline safety data and visualizing parallel coordinates plot because it helps to spot differences between safety matrices for each of the airlines (Figure 4)

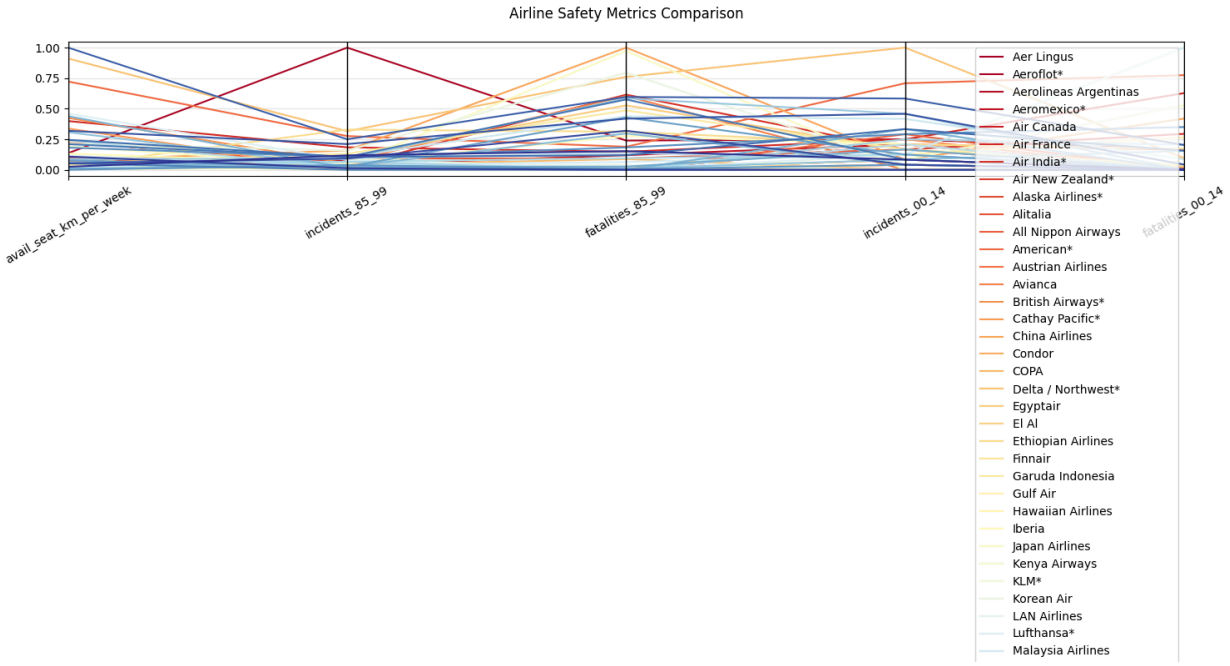


Figure 4: Airline Safety Metrics Comparison

Here, in terms of the metrics, some airlines like Aer Lingus, Air New Zealand, and Ethiopian Airlines show ups and downs in data where most of the airlines remain consistent. For ease of comparison, I have normalized the metrics to 0-1. The plot helps to co-relate things and come to some statistical decisions.

B. I have chosen country-wise cousin marriage percentage data and visualized using scatterplot because, in that way, I can look at individual statistics (Figure 5)

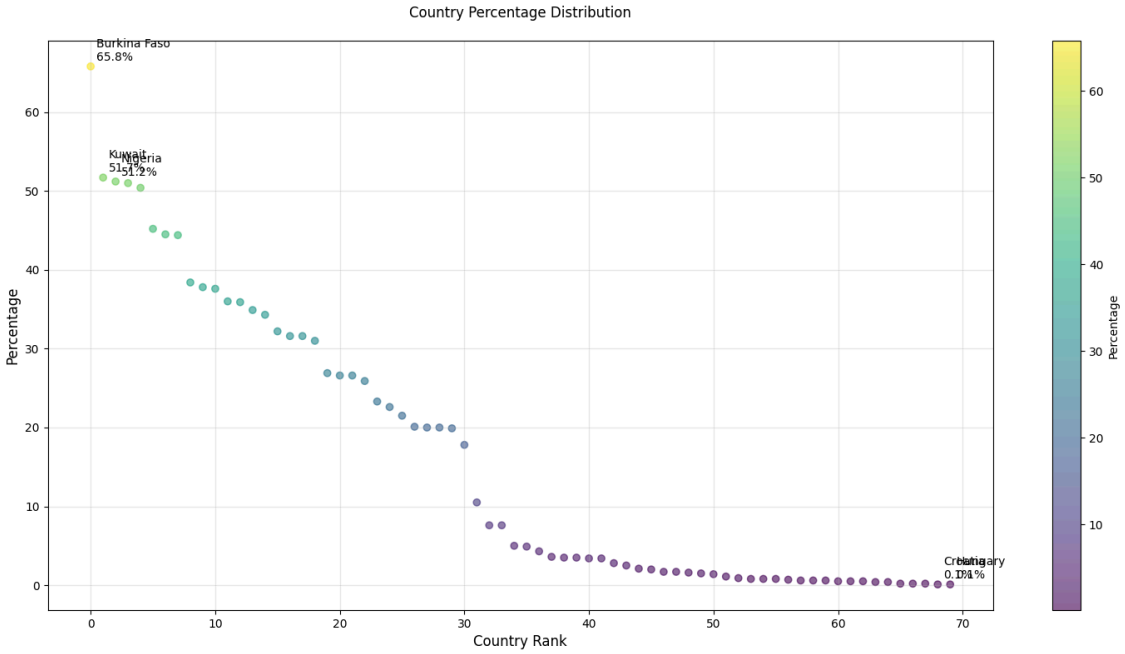


Figure 5: Cousin Marriage comparison (country-wise)

Here, in terms of the scatter plot, I have colored the plot based on some range of the ranks based on the percentage values. The labeled countries are from 1st ranked and last ranked based on the higher percentage of cousin marriage.

Part 3

1. Why is assessing value of visualizations important? What are the two measures for deciding the value of visualizations?

Assessing the value of visualization is crucial for several reasons. It helps in making informed decisions within the field, especially given the vast number of available methods. It allows one to determine if a visualization is effective and efficient. This assessment helps researchers and

practitioners understand whether a visualization method achieves its intended purpose and whether it does so with minimal use of resources. Ultimately, the goal is to ensure visualizations are useful and provide real value to users. The assessment of the value of a visualization is also important in determining if resources are being used effectively and if the visualization is better than existing methods. In addition, the value of a visualization is not just intrinsic but also depends on the context in which it is used. The two primary measures for deciding the value of a visualization are effectiveness and efficiency.
Effectiveness refers to whether the visualization achieves its intended purpose. It should enable the user to extract relevant information from the data and support the decisions they need to make. Efficiency concerns how well a visualization uses resources. It should achieve its purpose with a minimal amount of resources, such as time, effort, and computational costs.

2. Briefly describe a mathematical model for the visualization block shown in Fig. 1
The visualization model in Figure 1 can be described mathematically as follows:

The core of the model is the visualization process (V), which transforms data (D) according to a specification (S) into a time-varying image (I(t)). This is represented as $I(t) = V(D, S, t)$. The data (D) can be anything from simple bits to complex 3D tensor fields. The specification (S) includes the hardware, algorithms, and specific parameters used for the visualization.

Knowledge Gain (P): The image (I) is perceived by a user, resulting in an increase in knowledge (K). This increase is influenced by the user's existing knowledge and their perceptual abilities (P). The rate of knowledge gain is given by $dK/dt = P(I, K)$. A user with more prior knowledge (K) may gain more from the same visualization than someone with less.

Current Knowledge (K): The total current knowledge (K(t)) is the initial knowledge (K0) plus the accumulated knowledge gained over time, calculated by the integral: $K(t) = K0 + \int P(I, K, t)dt$.

Interactive Exploration (E): The user's interaction (E) can modify the visualization specifications (S) based on their current knowledge (K), represented as $dS/dt = E(K)$. The current specification (S(t)) is then updated based on this interaction, integrated as: $S(t) = S0 + \int E(K)dt$.

3. State four parameters that describe the costs associated with any visualization technique

Initial development costs (Ci(S0)): These costs are incurred during the development and implementation of the visualization method. This can include the development of the software and purchasing any new hardware needed.

Initial costs per user (Cu(S0)): These costs are related to the time and effort users spend to select, learn, and adapt the visualization method to their needs. This can include the time required to understand how to use it and tailor it to specific needs.

Initial costs per session (Cs(S0)): These costs are incurred each time a visualization session is started. This can include expenses to convert data and set up the initial parameters for the visualization.

Perception and exploration costs (Ce): These costs involve the time users spend watching, understanding, and interacting with the visualization.

The total costs are: $C=Ci+nCu+nmCs+nmkCe$.
The return on these investments consists oft hevalue: $W(\Delta K)$of the acquired knowledge, $\Delta K=K(T)-K(0)$
The total number of sessions: $G=nmW(\Delta K)$ and the total profit, $F= G-C$
We find $F=nm(W(\Delta K)-Cs-kCe)-Ci-nCu$.

4. What are the pros and cons of interactivity of visualizations?

Pros of interactivity:

a. Enhanced data exploration: Interactivity enables users to explore data more effectively, especially with large and complex datasets that are difficult to understand from a static image.
b. Customization for discovery: It allows users to adjust the specifications of the visualization, allowing them to explore different perspectives and uncover insights that might be missed with a static view.
c. Method development: Interactive tools can be essential during the development of new methods by allowing researchers to explore the solution space.

Cons of interactivity:

a. Subjectivity: Allowing users to freely modify the parameters of a visualization can lead to subjective interpretations of the data. Users might unconsciously or consciously adjust the visualization to emphasize the result they want to see.
b. Increased costs: Interaction is costly in terms of time and effort because users may spend hours trying out different options to view the data.
c. Difficult comparisons: High customization can also make it difficult to compare different visualizations, as different parameter settings may lead to different results.
d. Misleading outcomes: When parameters are not appropriately set, visualizations can be misleading and may generate negative knowledge.

Part 4 (in part 4 folder)

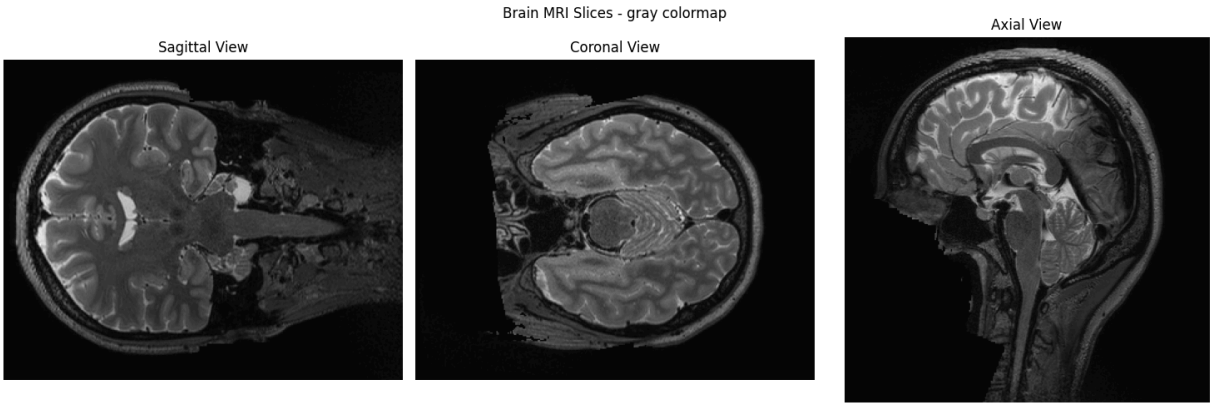


Figure 6: Brain MRI (gray)

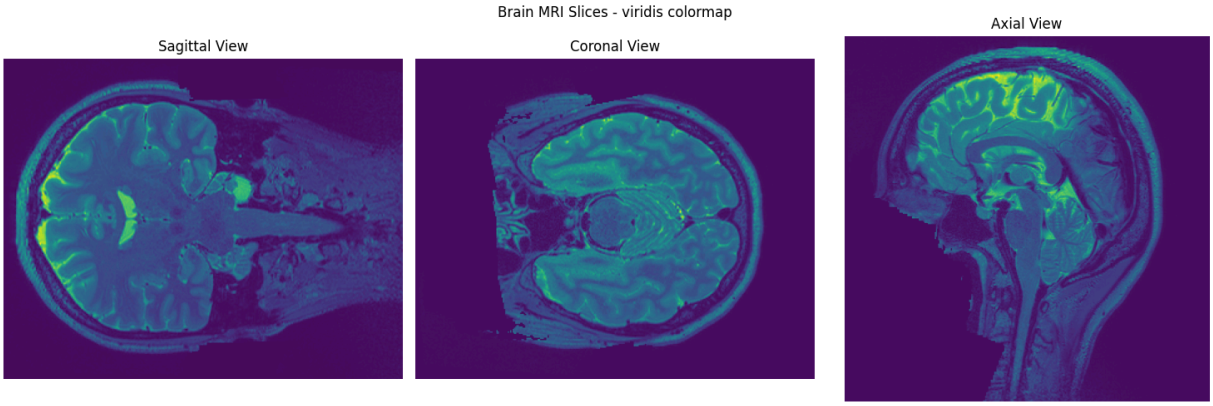


Figure 7: Brain MRI (viridis)

I can distinguish different intricate components of the image clearly in Figure 7, specially the hollow parts and the solid parts. Other than that, in terms of capturing details, both of them are the same to me.

Part 5

I learned about some new plots, such as radar plots and parallel coordinate plots. I found the five thirty-eight website to be a great resource for exploring data and visualizing them. The Python matplotlib and numpy array are great tools for visualizing them. Visualizing MRI data is a new experience as well. Looking at the visualizations of part one, a scatter plot is best to distinguish between two types of distributions, I will also add a histogram as it helps to visualize frequency distributions more precisely. In part 2, the color visualizations are a great thing to keep in mind in information visualization techniques.
I did not face any significant implementation challenges that I can mention broadly. One thing is, sometimes, I got errors regarding the numpy array dimension due to miss-assigning array

dimensions, but proper documentation helped to overcome this issue. Another issue was that I have a different Python version installed on my Mac, so had to create a new virtual environment to run the codes so that the package install dependencies are preserved properly.

Extra Credit

Matlab seems quite resource-rich to me, and it took time to load the application. Python is easier in terms of time complexity, in my opinion. I tried 2 plots with it, the scatter plot for part 1 for two distributions (Figure 8) and the contour plot for uniform distribution (Figure 9).

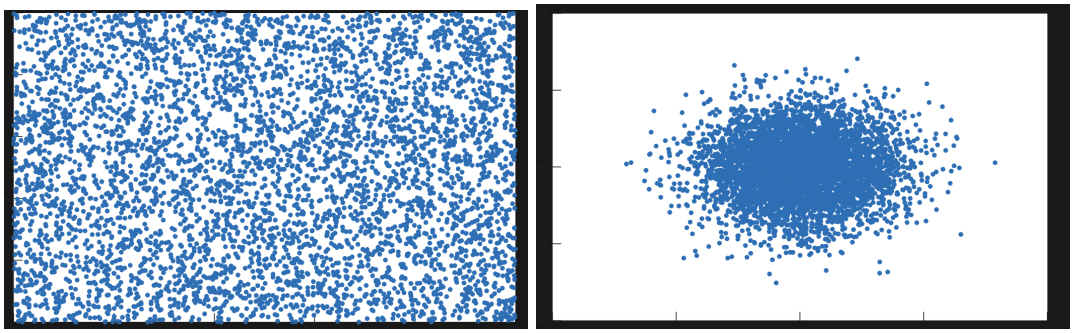


Figure 8: Scatter plot: (a) Uniform (b) Gaussian Distribution

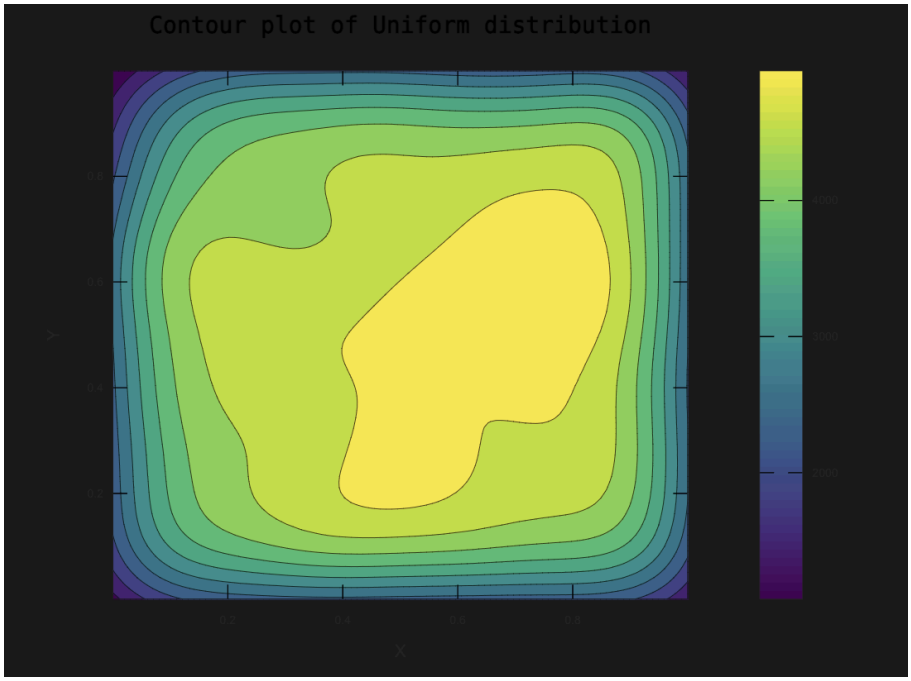


Figure 9: Contour plot for uniform distribution