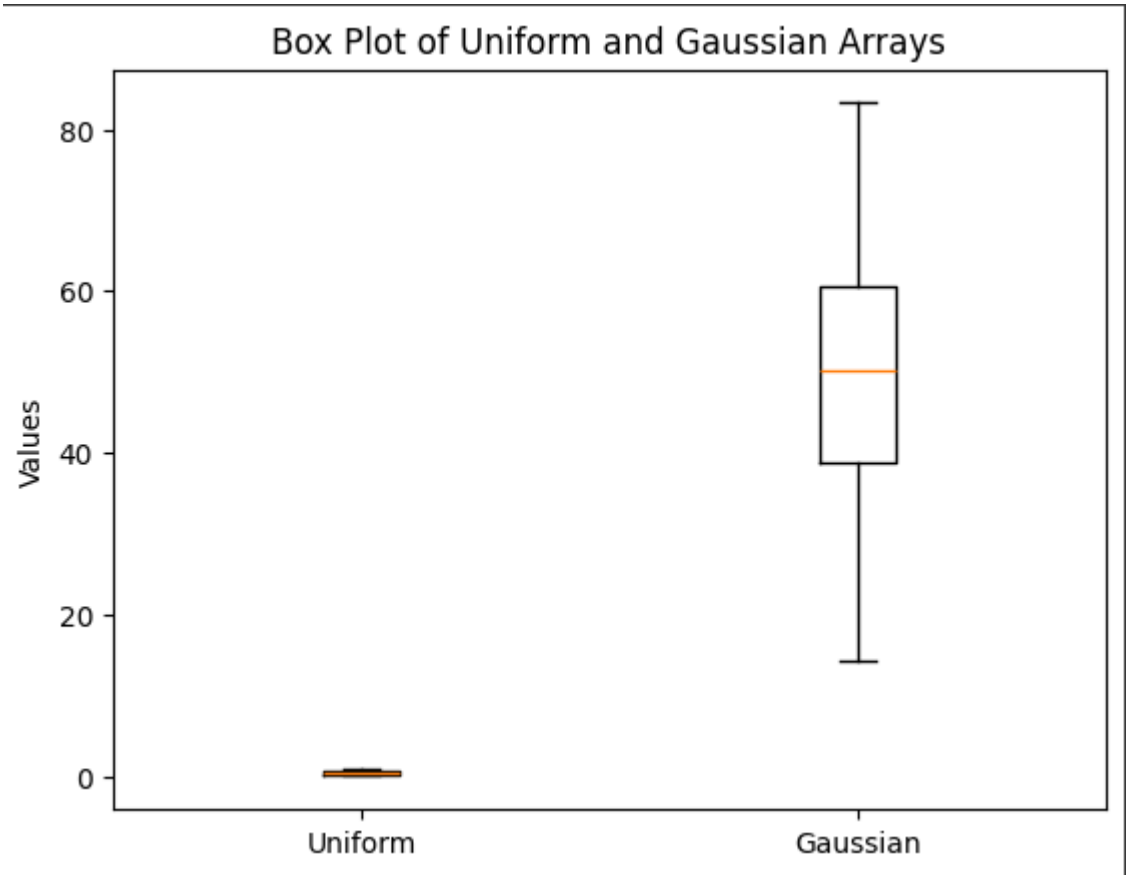
Report for CS 6635/5635 Assignment 1: Visualization and Analysis

Zaid Asif Basri

u1527407

Part 1: Generating and Visualizing Data

1) **Box Plot for random numbers**



The box plot visualizes the distribution of two datasets:

Uniform Distribution (Left Box - "Uniform")

- **The values are between 0 and 1.**
- **The box is very small, indicating a narrow range of data.**
- **The median is near 0.5, as expected in a uniform distribution.**
- **There are no significant outliers because the data is evenly spread.**

Gaussian Distribution (Right Box - "Gaussian")

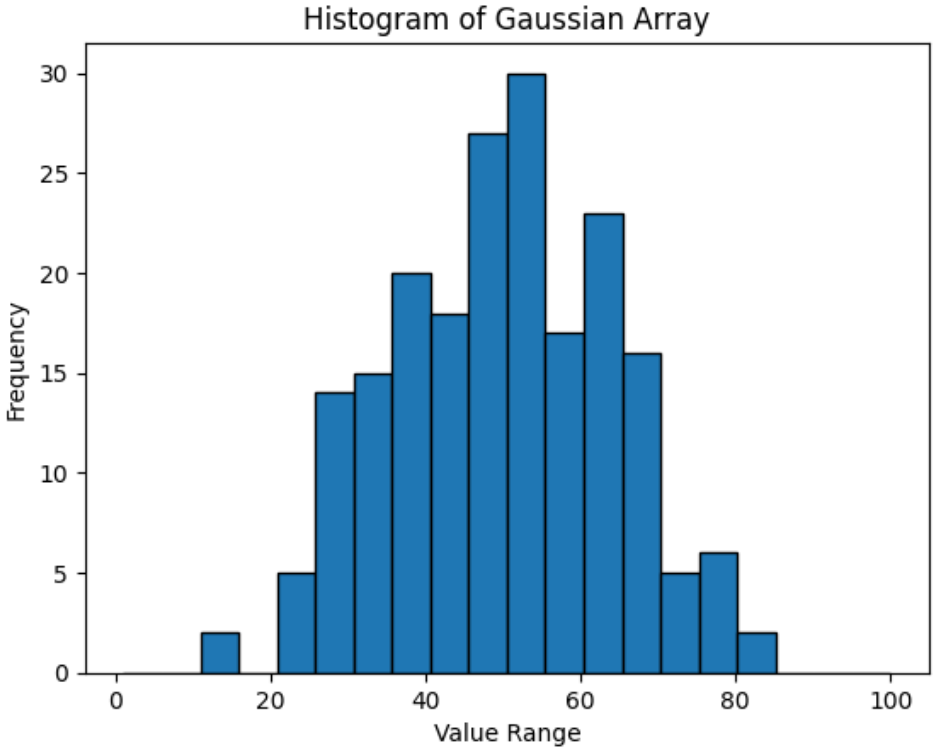
- **The values are generated with a Gaussian (normal) distribution between 1 and 100.**
- **The box is much larger, indicating a wider spread of values compared to the uniform distribution.**
- **The median is around the center of the distribution.**
- **The whiskers extend from a lower bound (minimum) to an upper bound (maximum), with some potential outliers beyond these limits.**

Key Observations

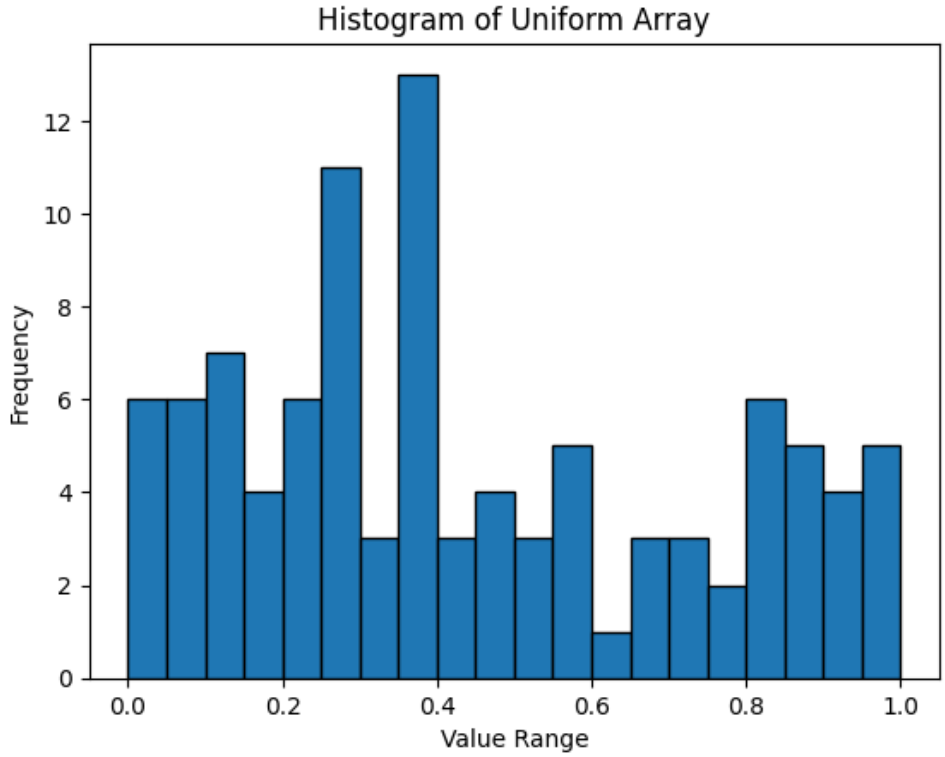
- **The uniform dataset has a much smaller range (0 to 1), leading to a tightly packed box.**
- **The Gaussian dataset spans a much larger range, with values spread over a wider interval (approximately 1 to 100).**
- **The interquartile range (IQR) is much larger for the Gaussian dataset, reflecting the natural variability in normally distributed data.**
- **The median (orange line) in both distributions represents the central value.**

This visualization highlights the difference between a narrow, evenly distributed dataset (Uniform) and a widely spread, bell-shaped dataset (Gaussian).

2) Histogram Analysis:



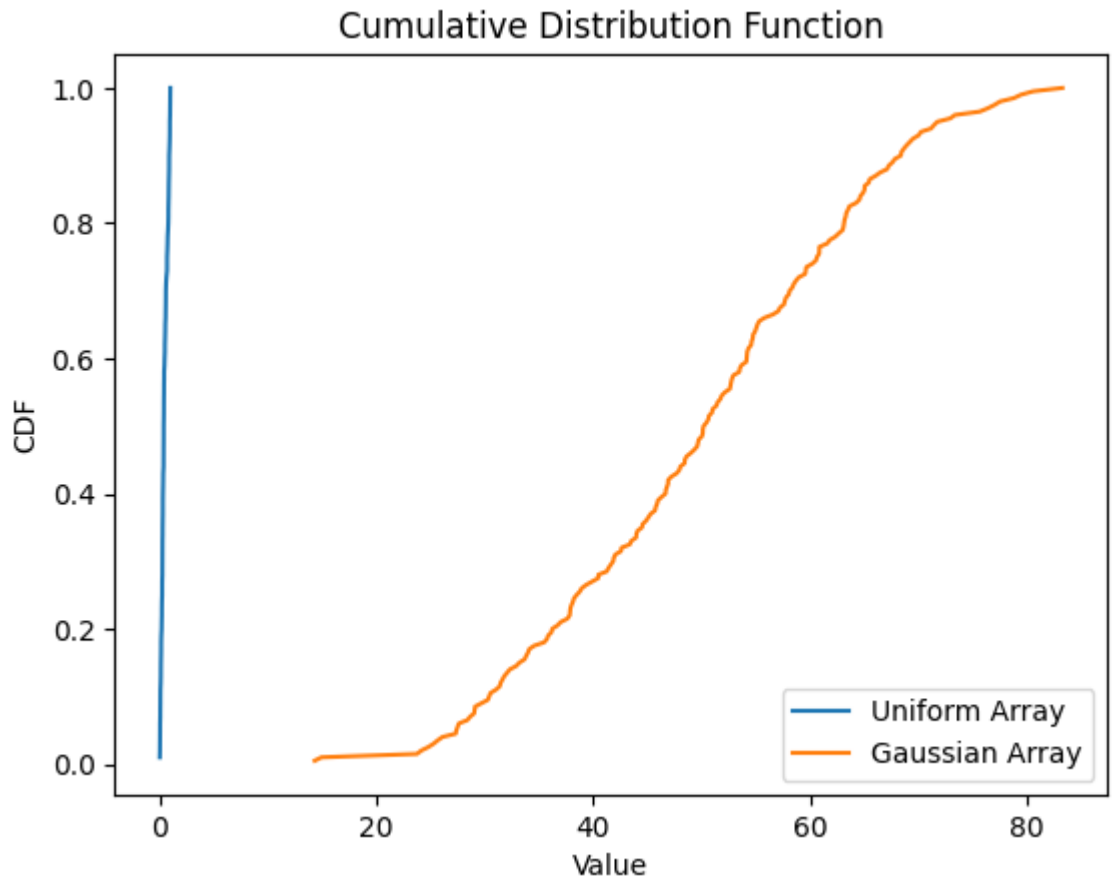
)



The data is uniformly distributed between 0 and 1. The histogram reflects equal probability for each bin since the distribution is uniform. The bars have relatively consistent heights, which indicate that values are equally likely in any range. Gaussian Array Histogram:

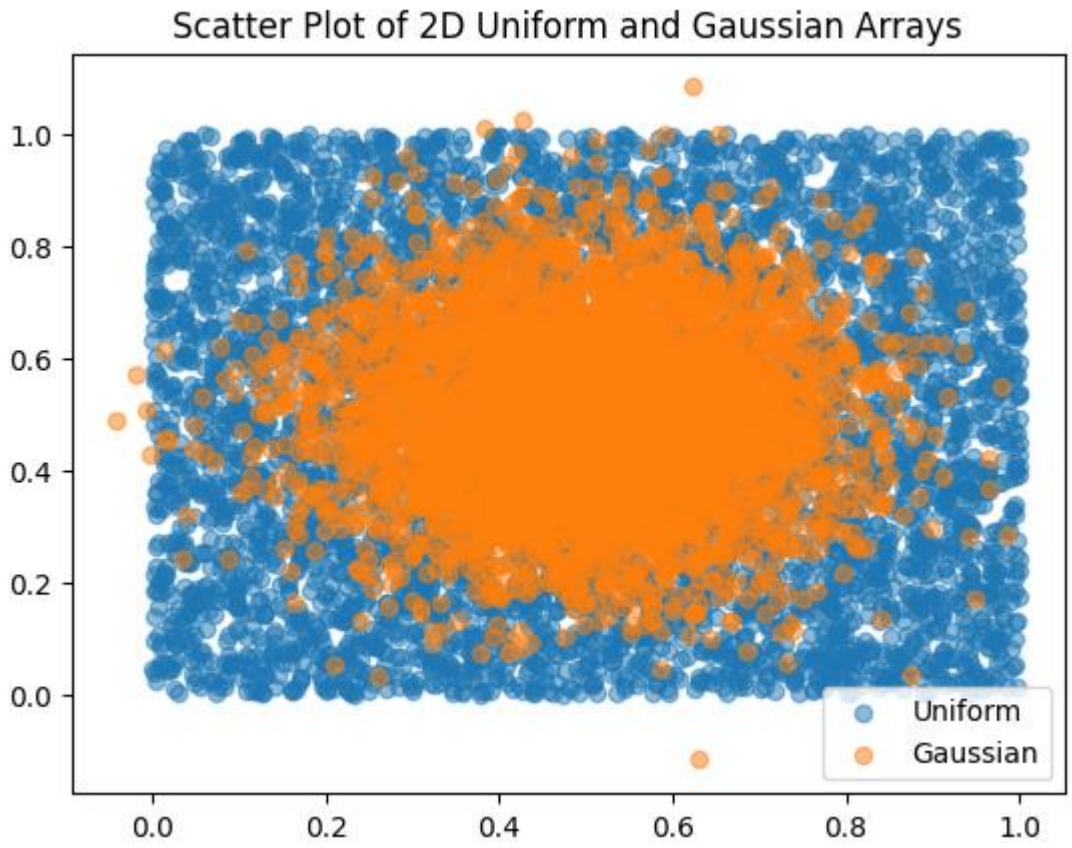
The data follows a Gaussian (normal) distribution with a chosen mean and standard deviation. The histogram forms a bell curve, representing the clustering of most values around the mean. The spread reflects the standard deviation, indicating how much the data deviates from the mean.

3) Cumulative Distribution Function (CDF):



The CDF provides a way to visualize the cumulative probability up to a given value: Uniform Array CDF: The uniform array's CDF shows a linear progression, indicating a constant rate of increase across the range. This matches the uniform distribution where values are evenly spread. Gaussian Array CDF: The Gaussian array's CDF has an "S" shape: A slow increase at the extremes. A steep increase near the center (mean) of the Gaussian distribution.

4) Scatter Plot analysis



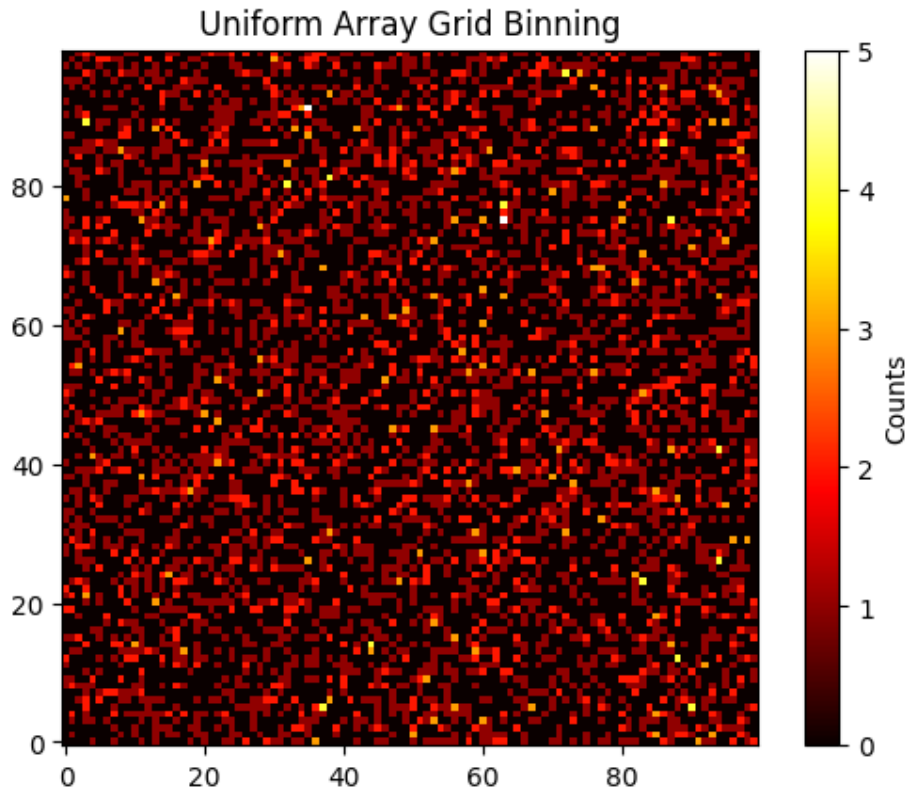
Uniform Sampling (Blue Dots): Evenly spread across [0,1]×[0,1], no clustering.

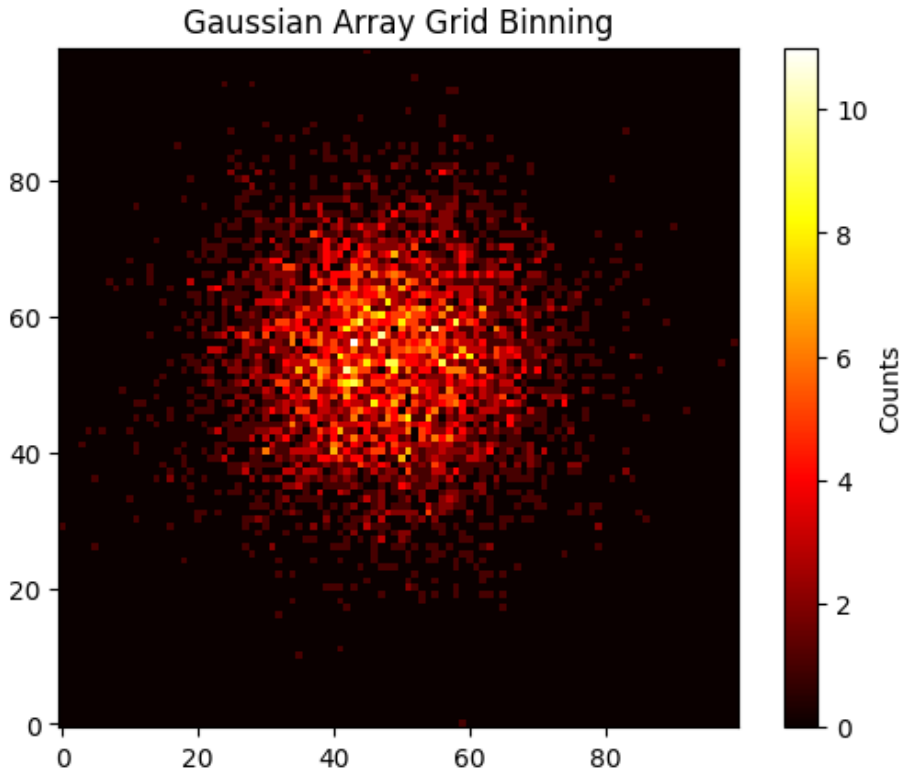
Gaussian Sampling (Orange Dots): Dense in the center, gradually decreasing outward.

Key Difference: Uniform covers the whole space, Gaussian clusters around the center.

Observation: Some Gaussian points may fall outside the boundaries.

4.1) Array Grid Binning



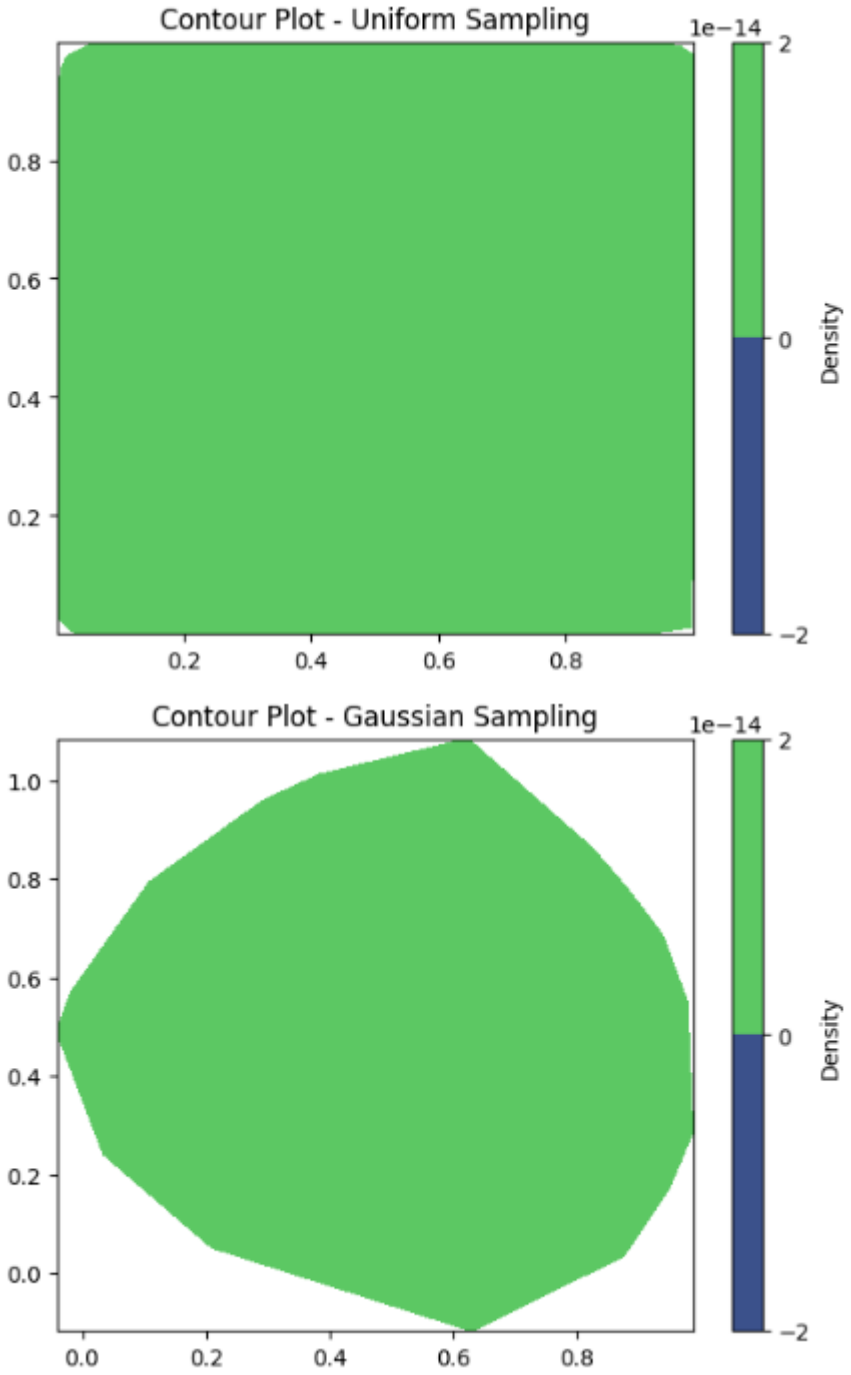


Top Graph (Uniform Grid Binning): Even distribution, no clear pattern, roughly equal density across the grid.

Bottom Graph (Gaussian Grid Binning): High density in the center, gradually decreasing outward, forming a bright central region.

Key Difference: Uniform points spread evenly, Gaussian points cluster in the center.

4.2) Contour Plot

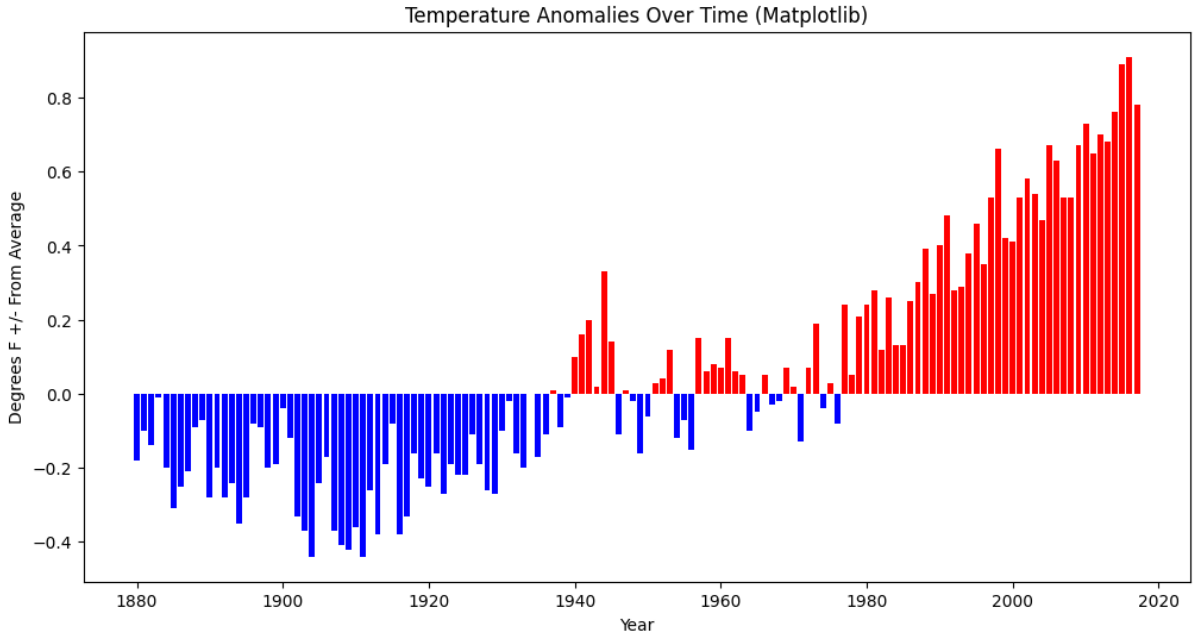


Uniform Sampling (Top Plot): Uniform sampling leads to evenly distributed points in a 2D plane. The contour plot is uniform, with no significant density changes, as expected.

Gaussian Sampling (Bottom Plot): Gaussian sampling creates a higher density of points near the mean in the 2D plane. The contour plot reveals clustered regions of higher density, which decrease outward.

Part 2: Various Datasets for Visualization

1) NOAA Land Ocean Temperature Anomalies



The data shows a clear trend of increasing positive temperature anomalies over time. In the earlier years (before 1940), negative anomalies (blue bars) dominate, indicating cooler-than-average temperatures. However, from the mid-20th century onward, there is a noticeable shift, with red bars (positive anomalies) becoming more frequent and prominent. This suggests a long-term warming trend, especially in recent decades where positive anomalies have grown significantly larger, reflecting global temperature increases.

2) Breakfast Cereals Dataset

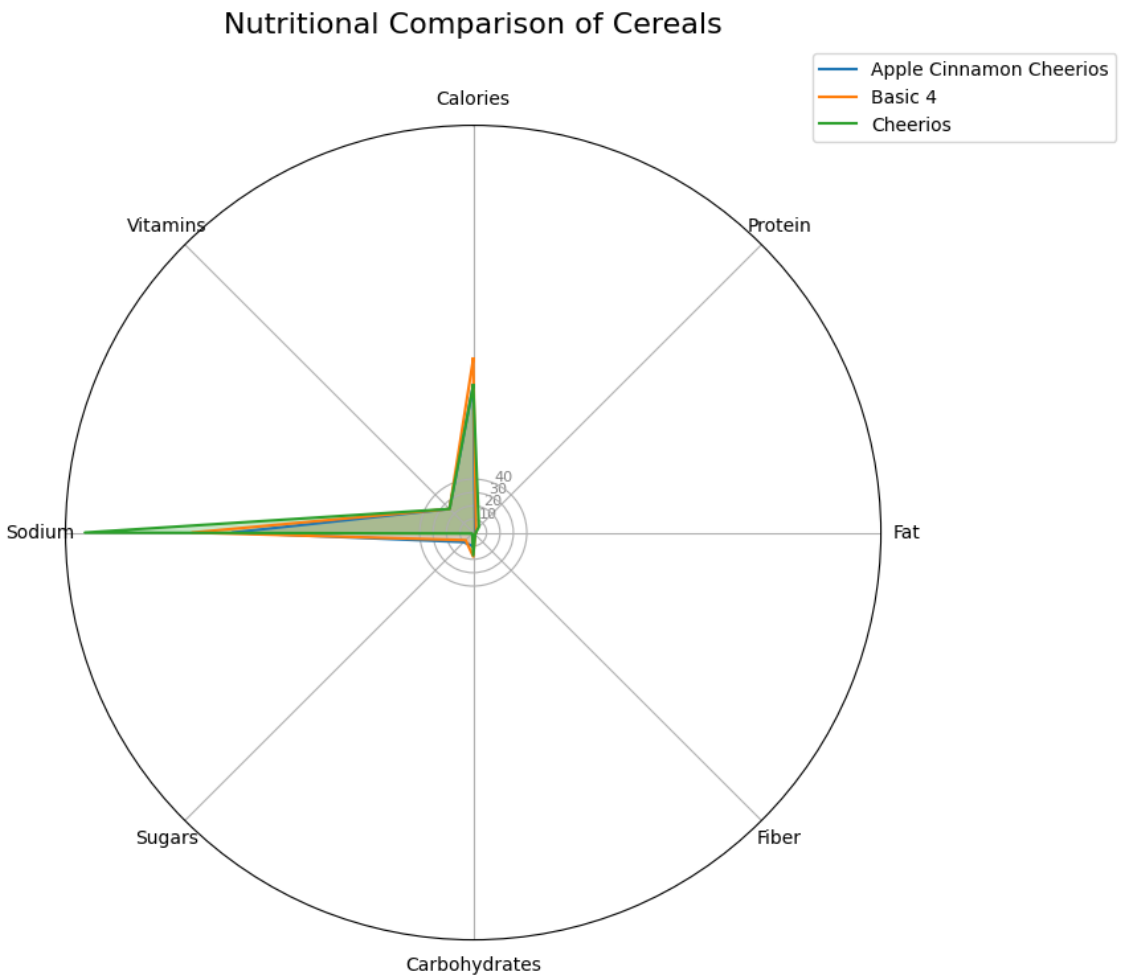


Chart Type: Radar (Star) Chart comparing 3 cereals across 8 nutritional categories.

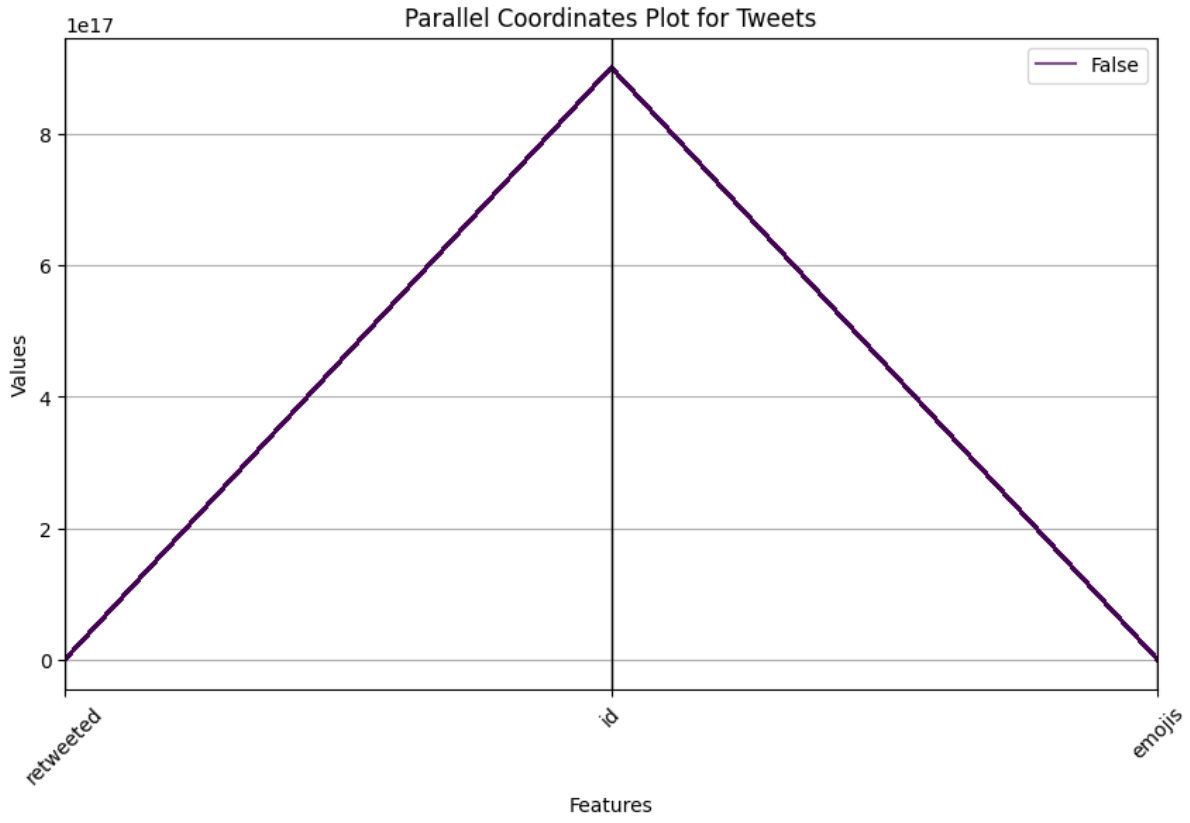
Cereals Compared: Apple Cinnamon Cheerios, Basic 4, and Cheerios.

Key Observations:

- **Sodium is significantly higher than other nutrients.**
- **Calories, Protein, Carbohydrates have moderate values.**
- **Vitamins, Fiber, Fat, and Sugars are relatively low.**

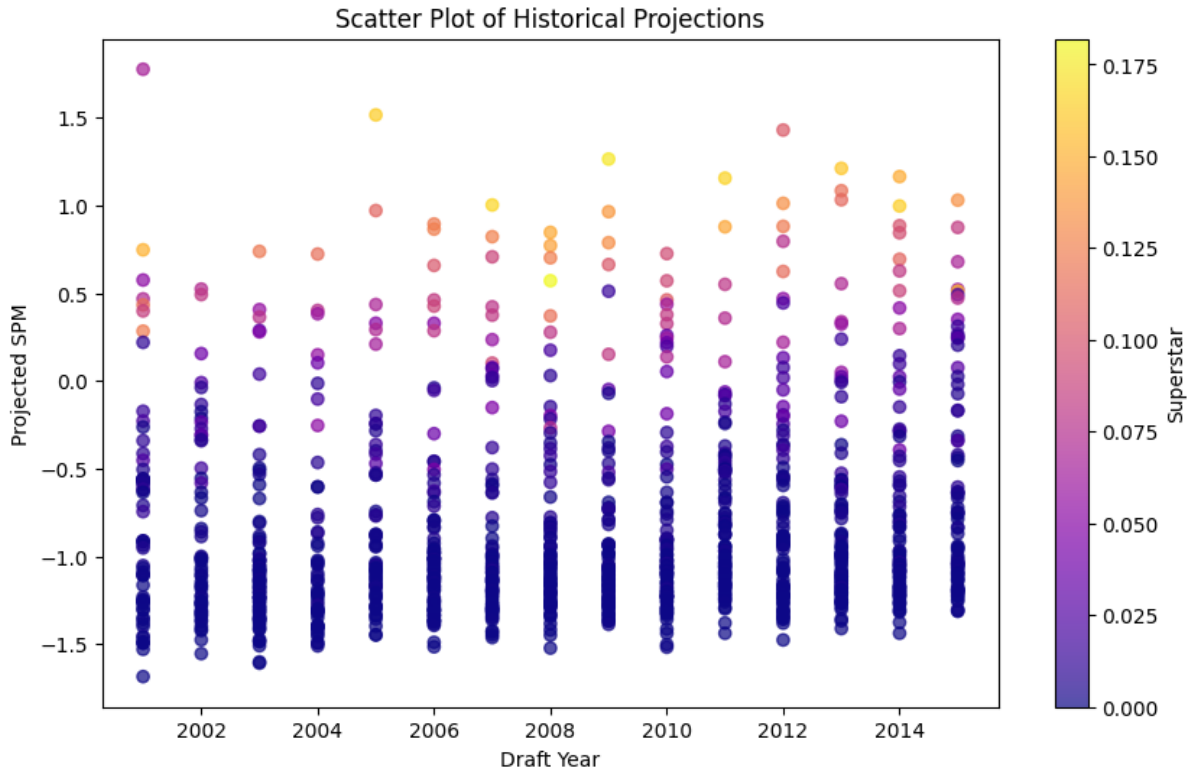
Purpose: Highlights nutritional differences among cereals in a visual format.

3)Mayweather vs Mcgregor and NBA 2015 draft dataset analysis



Graph for Mayweather vs Mcgregor tweets dataset(Parallel Coordinates Plot)

The plot visualizes three features: retweeted, id, and emojis. Each line represents a tweet and how its values for these features change across the axes. The color is based on the retweeted column, which is a boolean (True/False).



Graph for NBA 2025 draft (Scatter Plot)

Explanation of the Scatter Plot:

- **X-axis (Draft Year): Represents the year in which the player was drafted.**
- **Y-axis (Projected SPM - Statistical Plus/Minus): Represents a projection of a player's impact on the game.**
- **Color (Superstar Probability): Represents the likelihood of a player becoming a superstar, with the color gradient from dark blue (low probability) to yellow (high probability).**

Observed Trends:

1. **Majority of Players Have Low Superstar Probability:**
    - **Most data points are in dark blue/purple shades, indicating a low probability of becoming a superstar.**
    - **A few standout yellow points represent players projected to have high superstar potential.**
2. **Higher Projected SPM Often Correlates with Higher Superstar Probability:**
    - **Players with a high Projected SPM (above 0.5) tend to have warmer colors (pink, orange, or yellow), indicating a greater chance of superstardom.**
3. **Distribution Remains Consistent Across Draft Years:**

- **Every draft year has a wide spread of Projected SPM, but only a handful of players in each class are projected to be elite.**

4. **Outliers:**
    - **Some players (likely top draft picks) have exceptionally high Projected SPM (>1.5) and are bright yellow, signifying a strong superstar likelihood.**

Part 3: Questions on "The Value of Visualization"

1. Importance of Assessing the Value of Visualizations

Assessing the value of visualizations is crucial because it helps determine their effectiveness in conveying information, improving decision-making, and enhancing user understanding. Without proper assessment, a visualization may be misleading, inefficient, or fail to provide actionable insights.

Two Measures for Deciding the Value of Visualizations:

- **Effectiveness: How well the visualization helps users gain insights and make informed decisions.**
- **Efficiency: The speed and ease with which users can extract meaningful information from the visualization.**

---

2. Mathematical Model for the Visualization Block in Fig.1

The visualization block in the figure consists of a data component D, a visualization process V, and a support system S. The model can be represented as follows:

- **Input Data (D) is transformed by the visualization process (V) into an information stream (I).**
- **Support System (S) influences the visualization process and contributes to the rate of change of the user's understanding (dS/dt).**
- **The user interacts with the visualization, forming perceptions (P) and gaining knowledge (K).**
- **Knowledge accumulation is modeled by dK/dt, which depends on perception (P) and the effectiveness of the visualization.**

---

3. Four Parameters Describing Costs of Visualization Techniques

1. **Computational Cost: The processing power and time required to generate and render the visualization.**
2. **Cognitive Load: The mental effort users must invest to interpret the visualization correctly.**
3. **Implementation Cost: The resources needed to develop and maintain the visualization system.**
4. **Interaction Cost: The effort required for users to navigate and manipulate the visualization.**

---

4. Pros and Cons of Interactivity in Visualizations

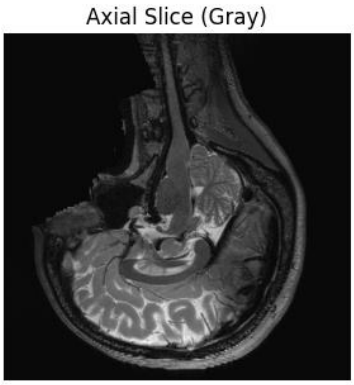
Pros:

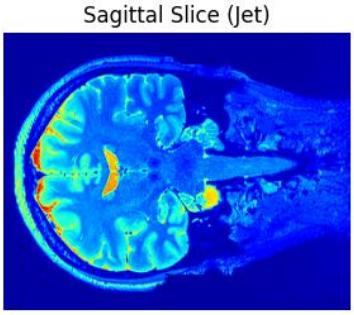
- **Enhances user engagement and exploration.**
- **Allows dynamic filtering and zooming for better insights.**
- **Adapts to different user needs and preferences.**
- **Can reveal hidden patterns and relationships.**

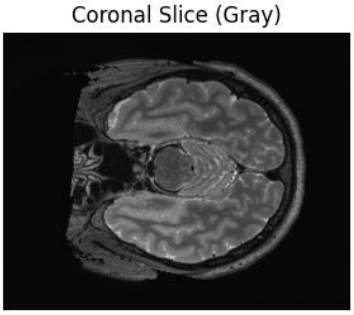
Cons:

- **Higher computational and design complexity.**
- **Increased cognitive load if not designed intuitively.**
- **Potential for misinterpretation due to excessive manipulation.**
- **Can lead to slower decision-making if users get overwhelmed.**

Part 4: 3D Scalar Volume Data Sets







Explanation of Colormap Choice:

Grayscale ("gray") provides a natural way to observe medical images since it preserves intensity relationships.

Jet ("jet") enhances contrast by mapping intensities to different colors, which can help highlight structures but may introduce perceptual distortions.

Part 5: Conclusion

Comparison of Results

- **Colormaps: Different colormaps emphasize distinct aspects of the data. For example, gray is better for contrast, while viridis highlights gradients effectively.**
- **2D Sampling: Uniform sampling ensures even coverage across the space, while Gaussian sampling clusters points around the mean, making it useful for modeling natural distributions.**
- **Grid Binning: The uniform dataset had an even spread in the heatmap, while the Gaussian dataset showed a high concentration in the center.**
- **Radar Chart Insights: The radar chart highlighted sodium as the most dominant nutrient, overshadowing others. This visualization helped quickly compare the cereals' nutritional profiles.**
- **Interactivity: While interactive visualizations offer deeper exploration, they come with additional computational costs and require user familiarity.**

Challenges Faced

- **Handling Large Datasets: Working with large datasets (e.g., MRI or .xls files) required optimizing memory usage and choosing efficient data processing methods.**
- **File Compatibility Issues: The provided .xls file caused some issues during reading, requiring adjustments in parsing methods.**
- **Choosing the Right Visualization: Deciding which visualization best represented the data was sometimes challenging, especially when comparing multiple variables.**
- **Data Distribution Misinterpretation: Initially, the Gaussian distribution's clustering effect made it seem like there were fewer points compared to the uniform dataset, requiring careful analysis.**
- **Library Learning Curve: Learning specialized libraries like nibabel for NIfTI files and ensuring proper implementation of matplotlib's advanced features took time.**

Key Learnings

- **Developed a better understanding of how different random sampling methods impact data distribution and visualization.**
- **Learned the significance of grid binning when analyzing large 2D datasets.**
- **Gained experience in handling and visualizing real-world data using Python libraries (matplotlib, seaborn, numpy, pandas).**
- **Understood how to interpret radar charts effectively to compare multiple variables at once.**
- **Improved problem-solving skills when dealing with file format inconsistencies.**
- **Recognized the trade-offs between static and interactive visualizations and when to use each.**
- **Explored nibabel for handling NIfTI files, which is essential for processing neuroimaging data.**

References

- **Python Data Visualization Libraries: matplotlib, seaborn, plotly**
- **Sampling Techniques: Understanding Uniform vs. Gaussian Distributions**
- **Handling .xls Files: Pandas Documentation on read_excel()**
- **nibabel Documentation: Working with NIfTI and other medical imaging formats**