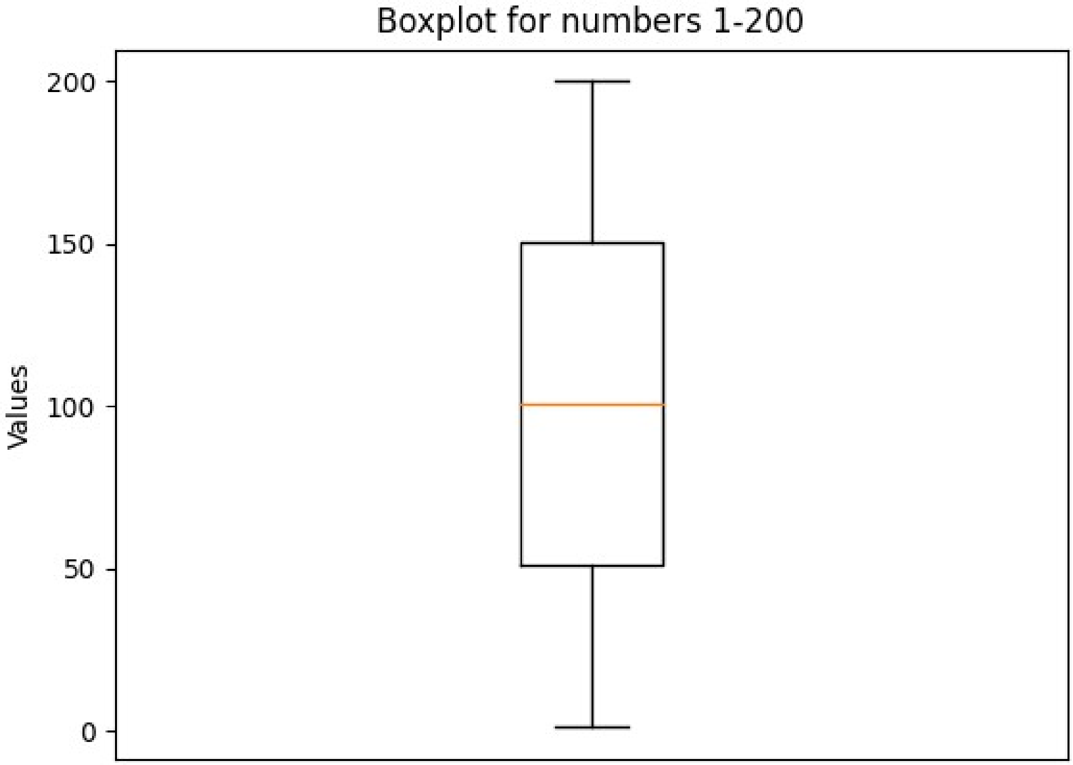
Jack Wilburn
Assignment1

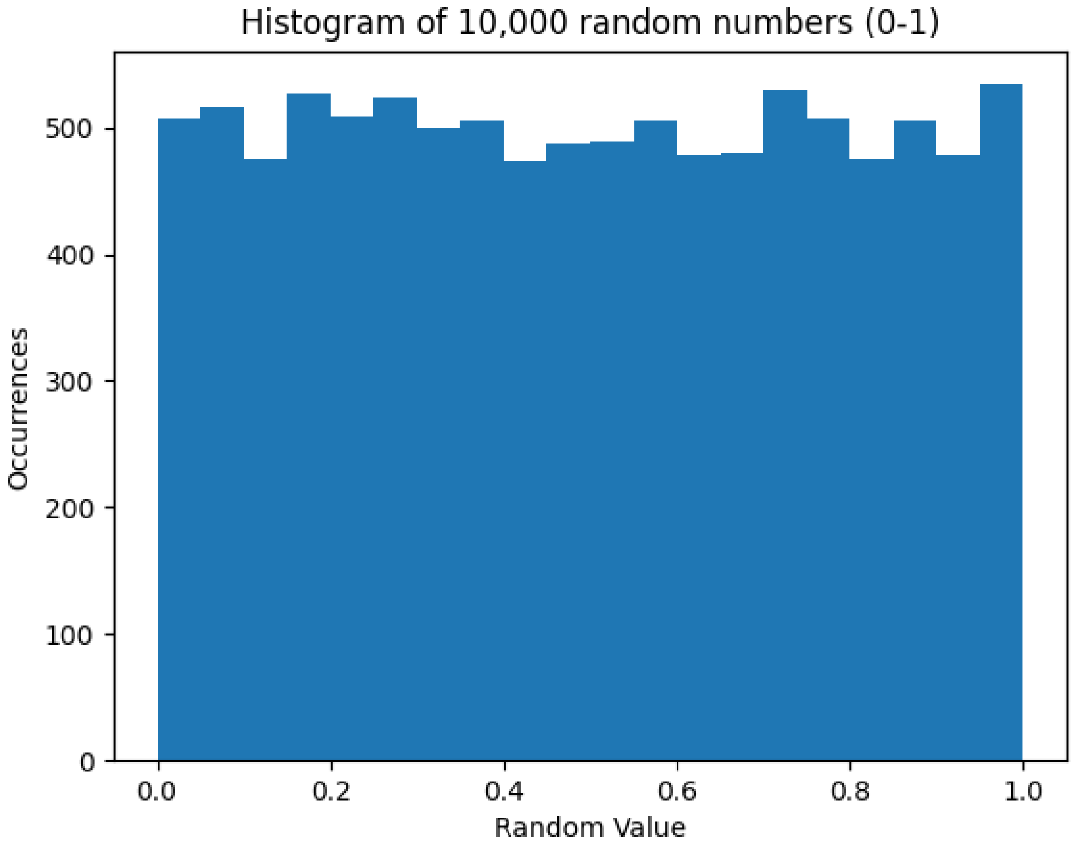
Please find all code in assignment1.py

1.
    1. [4 pts] Create an array with 200 elements from 1 to 200 in order. Create a box plot for visualization of your data



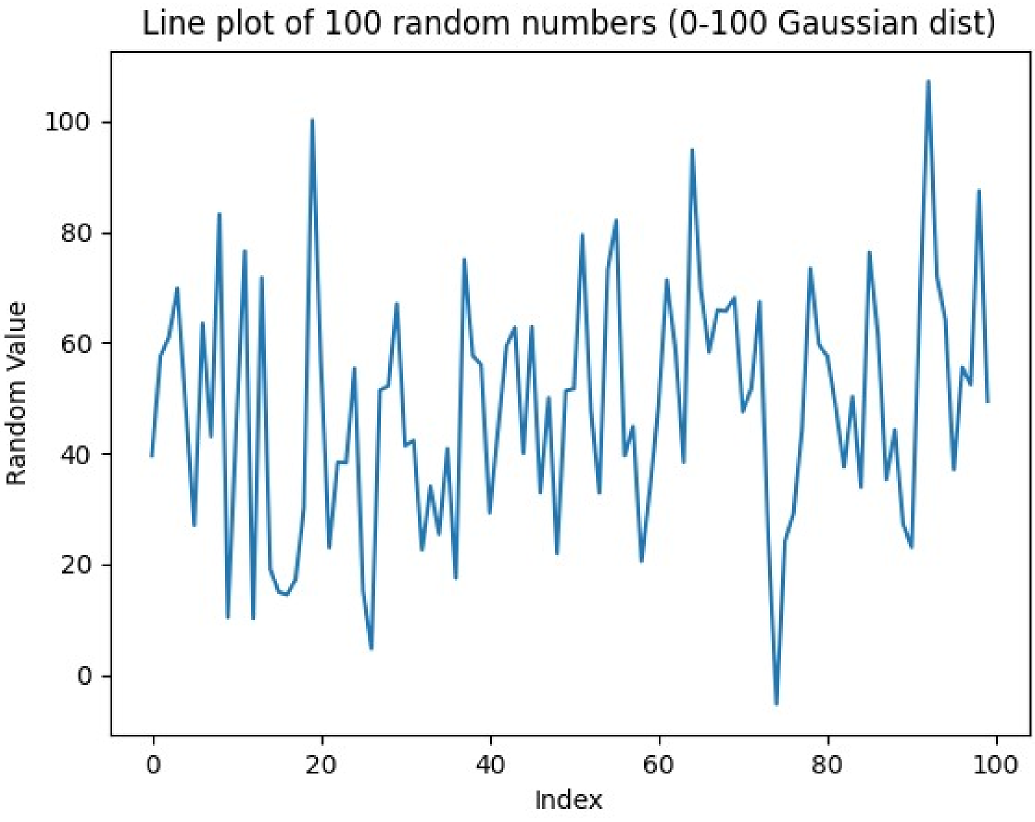
Here you can see that the median is at 100, 25th percentile is at 50, and 75th percentile is at 150. The whiskers here extend out to the extreme values, 1 and 200.

2. [4 pts] Create an array with 10,000 random numbers. Create a histogram of the data using 20 bins



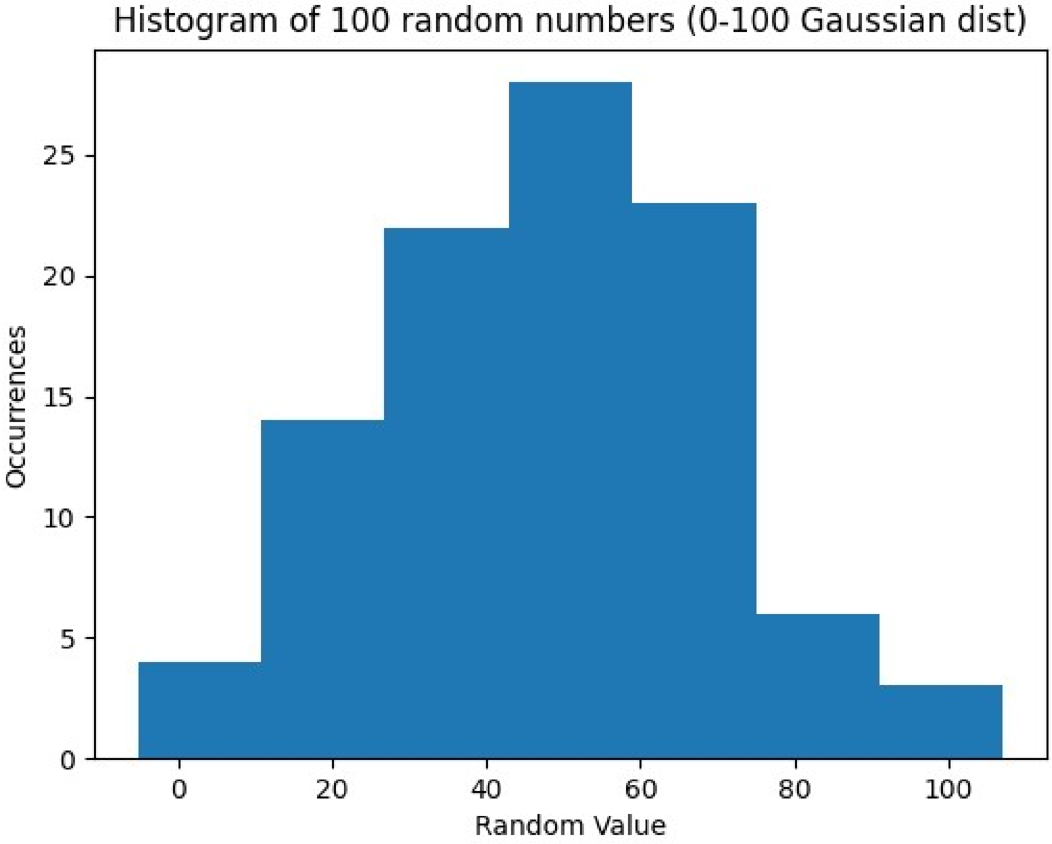
Here you see some slight variations between the occurrences of random values with some bins having slightly more/less occurrences. However, this is to be expected as a result of the randomness. All bins have roughly 500 occurrences as expected (10,000 values with 20 bins)

3. [6 pts] Write a program to generate 100 random numbers Gaussian distributed between 1 and 100. Write the numbers out to a binary file and use a line graph to draw the 100 numbers
    1. You will need to find an appropriate mean and standard deviation for the Gaussian. It is okay if just a few of the numbers generated are outside the [1,100] range.



Here there is clearly no trend to the line, with the line just connecting each random value roughly between 0 and 100. There seem to be more values in the 20-80 range, as expected, since this comes from a Gaussian distribution.

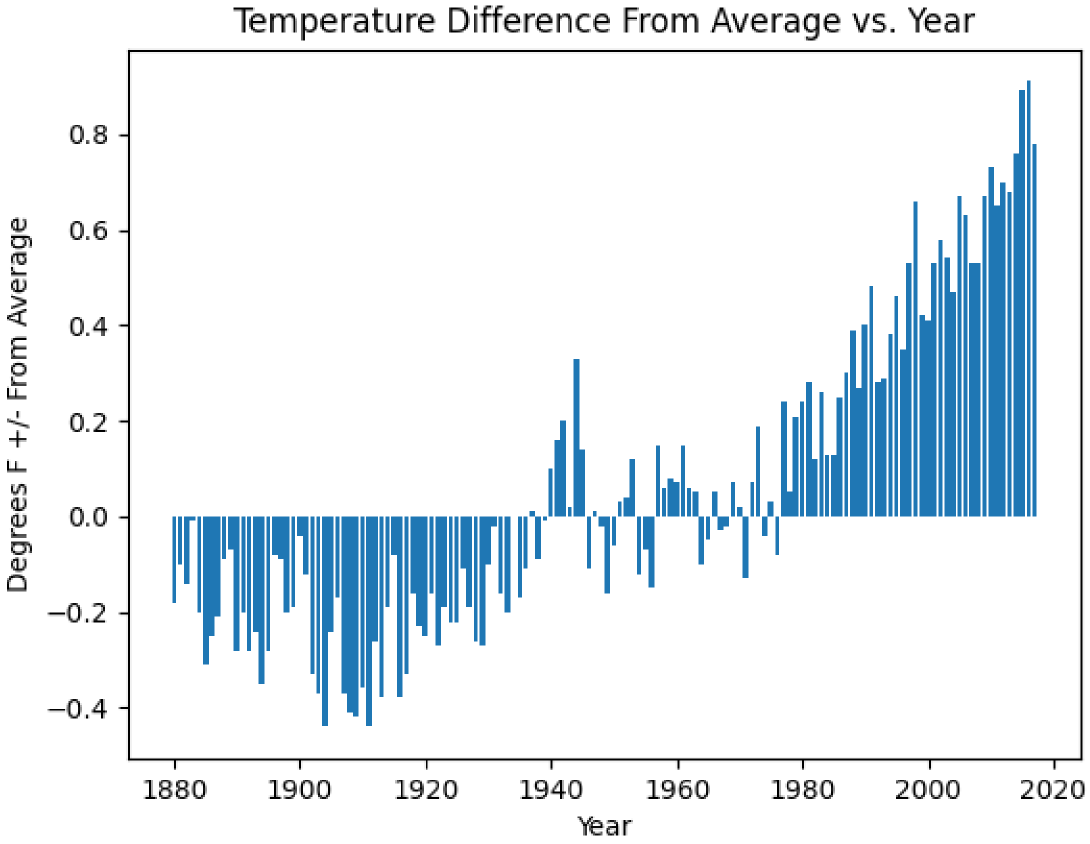
4. [6 pts] Write a program to read the binary file back, divide the range between 1 and 100 into 7 intervals, and calculate the frequency for each interval: Display a histogram of your result.



Here we see a gaussian distribution with slight perturbations due to randomness. The distributions is unimodal and mostly symmetric with no notable outliers.

2.

1. [6 pts] Download the NOAA Land Ocean Temperature Anomalies Data Set: https://my.eng.utah.edu/~cs6635/NOAA-Temperatures.csv. Create a bar plot of the data. Include a label called "Year" along the x-axis and a label called Degrees F +/- From Average along the y-axis. Describe trends in the data

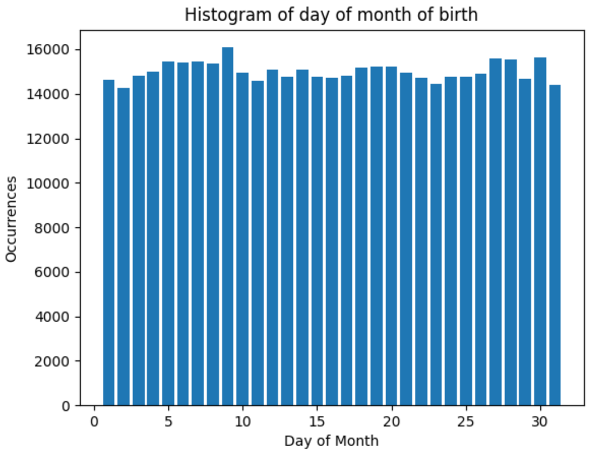


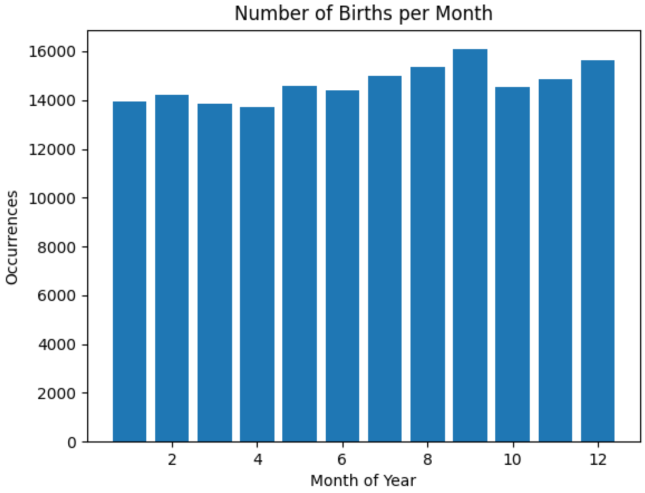
This data shows that the average temperature has changed from consistently below average (during 1880 to 1940) to around average (1940 to 1980) to consistently above average (1980 to 2020). This would imply that temperatures over time are rising and that the globe is warming.

2. [6 pts] Download the member of Congress by Age data set: https://git.io/Jt45w2 Create a Star Plot of the data and create a Parallel Coordinates Plot of the data. Describe the trends in the data

   Missing.

3. [6 pts] Download the U.S. Birth data set: https://git.io/Jt45X. What day of the month had the highest number of births? What day of the month had the lowest number of births? Are there any interesting trends in the data, i.e. more births in Summer or Winter? What about births on Friday the 13th?

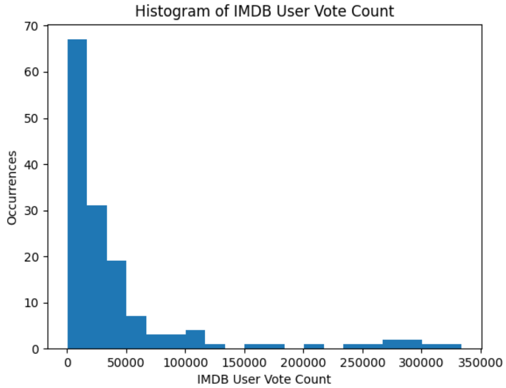

Histogram of day of month of birth

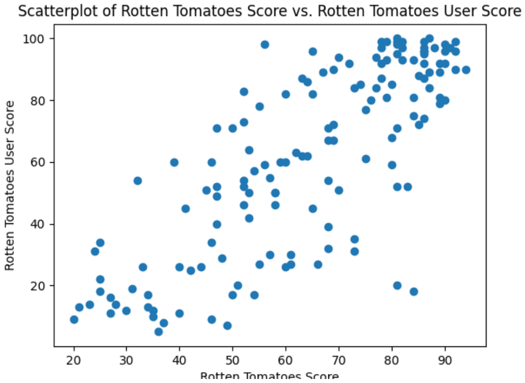

Number of Births per Month

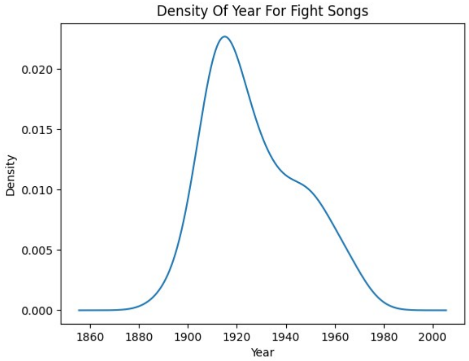
Births appear to be evenly distributed over the days of the month with no day being far above or below the others. The most popular day is the 9th and the least popular is the 31st (which might make sense given that only 7 months have 31 days.

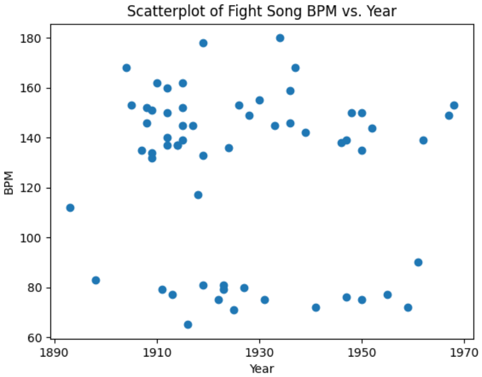
There appear to be slightly more births in the summer and fall months in this dataset, but the trend isn't striking. The most popular month was September, and the least popular was April

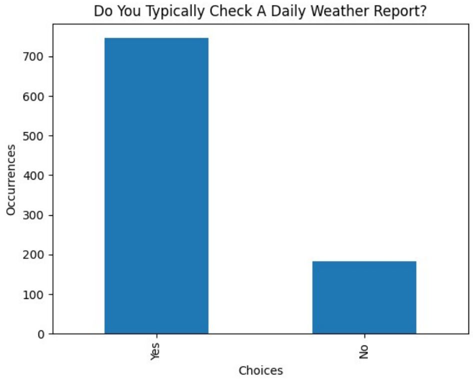
4. [18 pts] Five Thirty Eight maintains a sever with many interesting datasets: https://github.com/fivethirtyeight/data . Choose 3 different data sets to visualize. Visualize each data set in a different way. Describe the trends in the data.

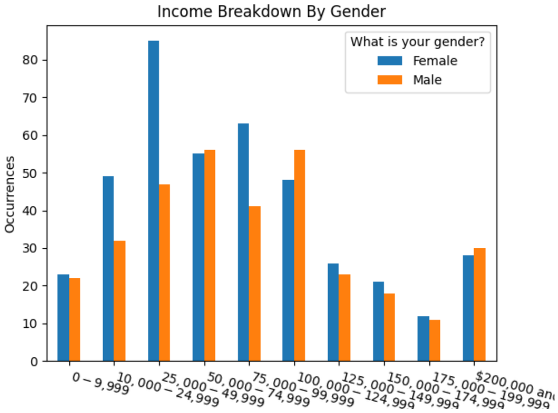












Each row here represents data from a different dataset. The first is a movie ratings dataset, the second relates to fight songs for college sports teams, and the third row is some data the comes from a survey about weather checking habits. See the next page for my analysis.

In the first chart (top left) we see that most movies have a small number of user reviews on IMDB, while some get up to 3.5 million. This data is highly right skewed with only a few movies getting in the millions of reviews.

In the second chart (top right) we see a scatter plot of the Roten Tomatoes score on the x-axis and the Roten Tomatoes user score on the y axis. If the user score exactly matches the critics reviews, we should expect to see a perfect y=x line, but we don't. This shows that users disagree with the critics. The trend is roughly linear and positive, which shows the variance in the user score could be explained by the critic score (and vice-versa), but we'd have to run some statistical tests to confirm that.

In the third chart (middle left) shows is a KDE showing the distribution of year for college team fight song year of creation. We see a strong peak at 1920 with a gradually declining distribution as the year grows.

In the fourth chart (middle right) we have a scatter plot of fight song BPM (y-axis) vs year (x-axis). My goal in this chart was to see how the BPM of fight songs has changed over time, and there doesn't appear to be any trend. There are 2 clear modes for BPM, one around 80 and one around 150. The data appears to be roughly equivalent at each time slice.

In the fifth chart (bottom left) we see that most people, around 700, do check the weather daily, and a few people, around 200, do not check the weather report daily.

In the sixth chart (bottom right) we're comparing household income levels to gender. It seems that the women who answered this survey had a lower household income with a strong mode between 25k and 50k, where as men more often made more than that. I'm not certain we can say anything causal here (gender pay gap, etc.), but the women who answered this survey definitely came from household that made less money.

3.

1. [6 pts] Why is assessing value of visualizations important? What are the two measures for deciding the value of visualizations?
   Assessing the value of visualizations is important so that we can make increasingly better design decisions and so that we can help solve problems that are currently unsolved. If we were to continue making visualizations without assessing them, we'd have no idea which visualizations are most effective, and could potentially make it harder/more obtuse to analyze data.

   The two measures used for assessing the value of a visualization are "effectiveness and effciency", where effectiveness relates to how well the data is represented and efficiency relates to how quickly the visualization can be made and how well the information is transferred.

2. [6 pts] Briefly describe a mathematical model for the visualization block shown in Fig. 1
   The mathematical model in figure one is composed of several parts:
   - Data (D)
   - Visualization (V)
   - Specification (S)
   - Perception (P)
   - Knowledge (K)
   - Exploration (E)

   The ultimate goal of this model (and of visualization) is to increase knowledge in the observer. This is represented as K in the model. K increases as a function of perceptions; a user perceives some visualization/image and is able to learn something (an insight). Additionally, the users perception changes as a function of knowledge. The image that the user is seeing is the output of the visualization, which iself is composed of multiple parts: data and specifications. The final piece of the model is exploration, which takes input from the users knowledge and feeds back into the specification (e.g. a user exploring may change variable domain/range and that would be encapsulated in the specifications). These pieces, combined, allow a user to synthesize new knowledge from a visualization.

3. [6 pts] State four parameters that describe the costs associated with any visualization technique
   The four parameters associated with cost for visualization are: initial development costs, initial costs per user, initial costs per session, and perception and exploration costs.

4. [6 pts] What are the pros and cons of interactivity of visualizations?
   Ineractivity of visualizations "strongly enhances the understanding of the data" allowing a user to gain knowledge more easily. True understanding can only come when you’re able to explore and play with the data. There are other, more practical benefits such as showing more data than can fit on one screen.

   Having interactivity in the vis is not always for the best though. For example, a user can use the vis reactivity to reinforce their biases, it might be harder to compare visualizations if they're reactive, and designing visualizations for reactivity is costly in terms of development and exploration time.

4. [20 pts] MATLAB/Python also can be used for analysis and visualization of 3D volume data sets, such as brain MRI images. Download the brain MRI data set from https://pubweb.eng.utah.edu/~cs6635/T2.nii.gz. The data format is .nii with 320 x 320 x 256 dimensions. Load data in MATLAB/Python. Extract one slice from the volume and save it as an image.
    1. Here's a slice from that volume: