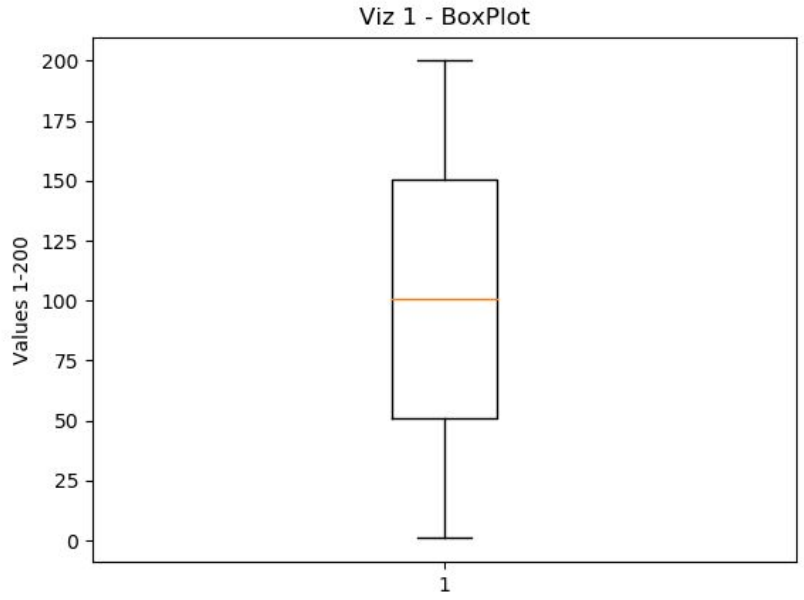
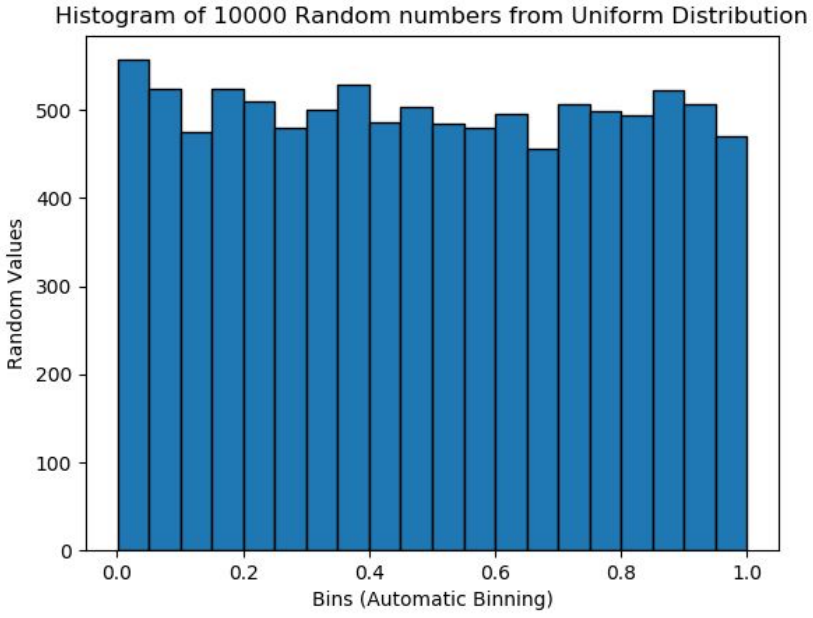
Question 1.

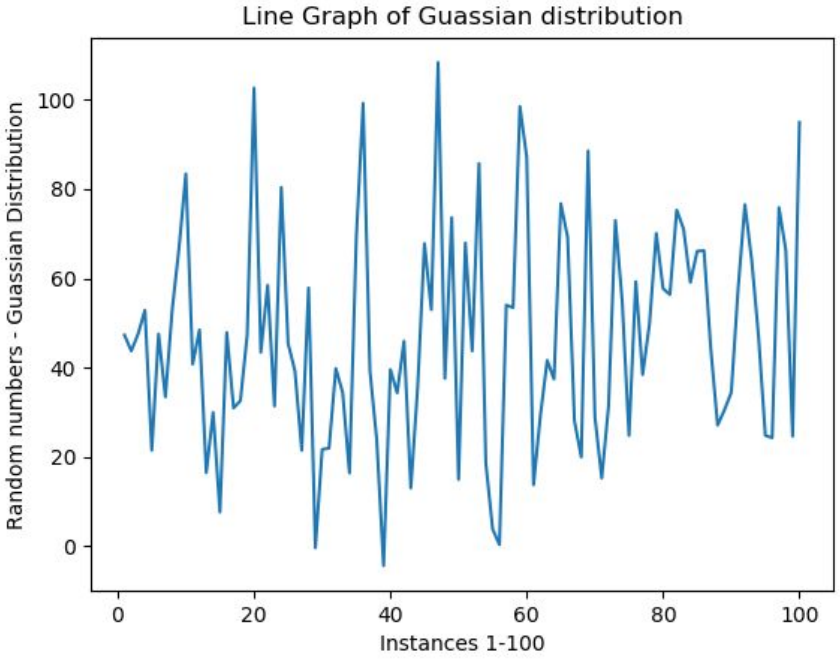
All the graph plots were achieved using matplotlib library in python. Generating random numbers and list of ordered numbers was done using basic python functions with the help of numpy library.

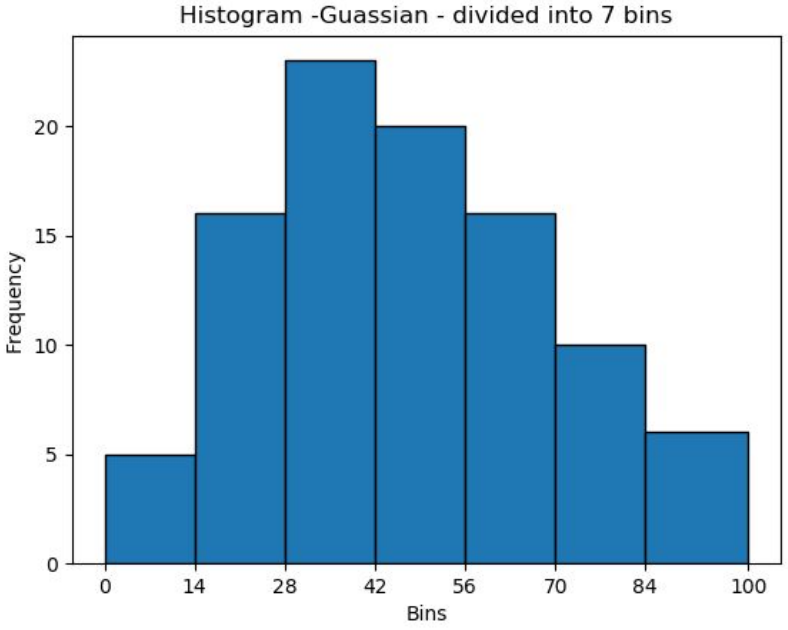
Used an extrenal library `pickle` in order to dump array data into a binary file. And then read it back into another variable using `pickle` load operation.

Made use of boxplot, hist, and plot functions in the module pyplot under matplotlib library.





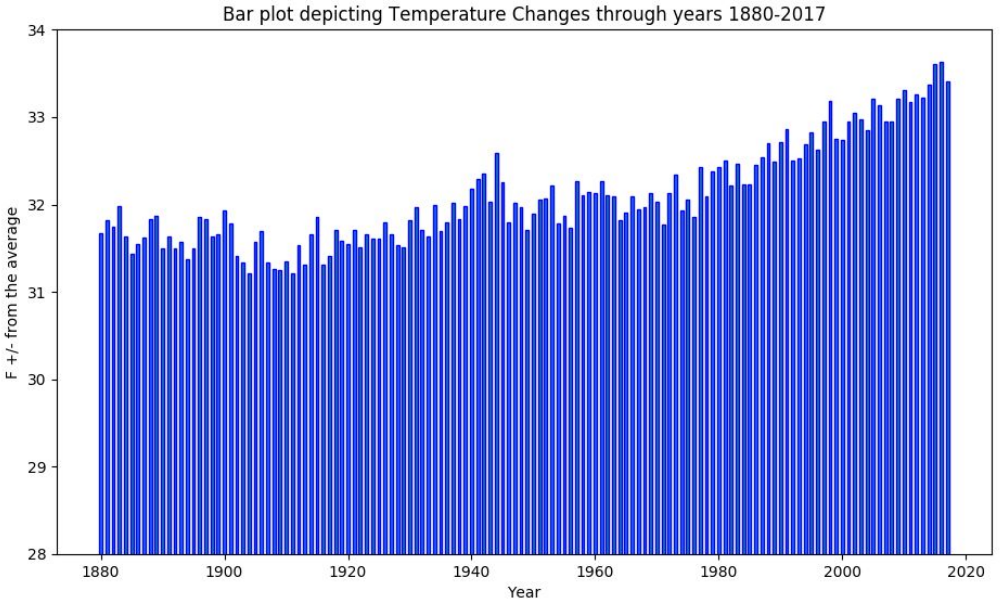




For gaussian distribution I used mean at 50, standard deviation of 25 and plotted 100 values.

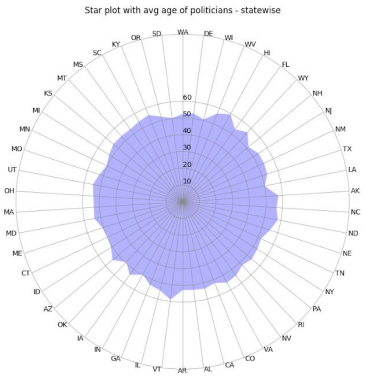
Question 2.

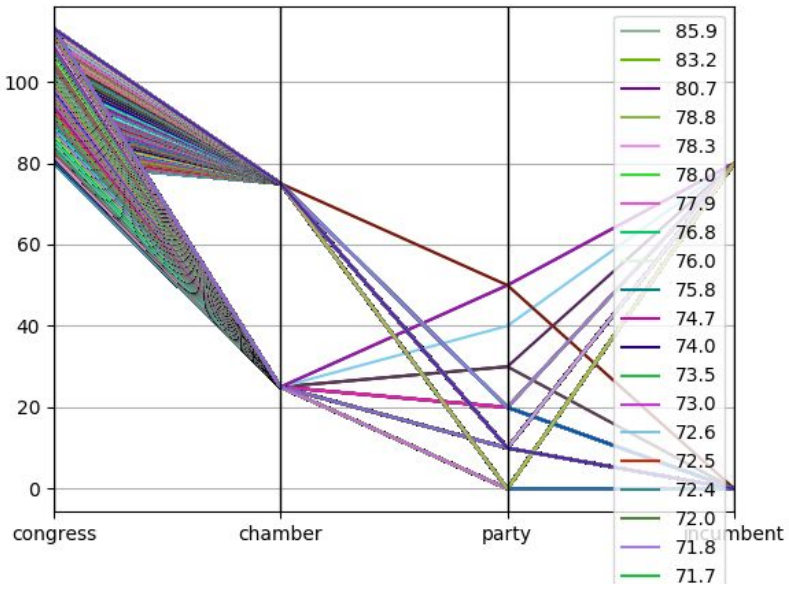
part a: NOAA Land Ocean Temperature Anomalies Data Set

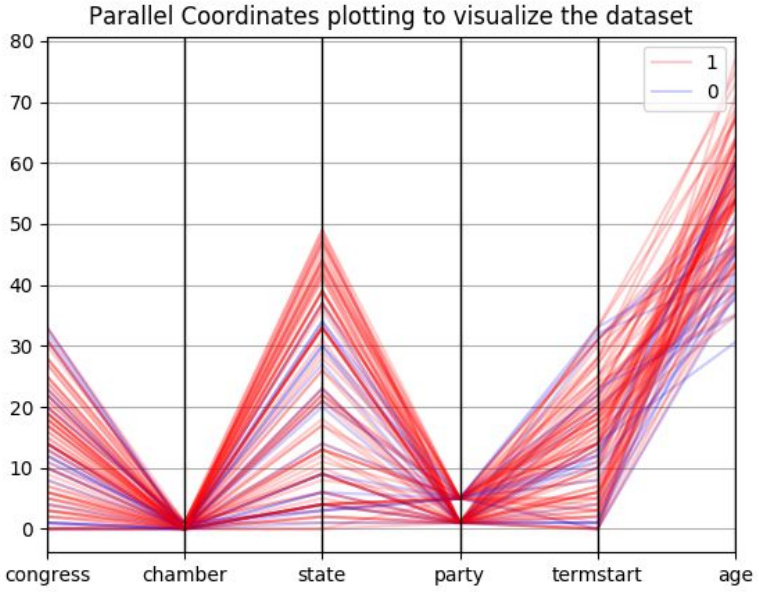


Looking at the data on the whole scale (0 - 100) it appears as though there is slight increase in average temperature.
But since temperature increase has an adverse effect on all biological aspects on earth, I had to take a closer look at the plot. By limiting the scale range from 28 to 34 we can see an increasing temperature trend which is the effect of Global warming.
If this keeps up then it's going to cause major problems in the upcoming years.

part b: member of Congress by Age data







Here, we have a huge database of leaders with different features in table like age, party, state, term served, etc. One of the main challenges here was to identify the important information and analyse it.
For Star Plot, plotted the average age of all politicians from a particular state. This was a simple operation of creating a map of state to sum of all ages along with count of members per state and then finding out the average.

For parallel coordinates, made used pandas which provides a lots of functionality in creating a parallel plots. Though one disadvantage is that it doesn't take in strings. Had to replace it with a number scale.

From the star plots and parallel coordinates, there are multiple things we can observe, like the average age of states like Texas is higher as compared to states like Utah, Alabama etc.

From parallel plot sampling of data, we can see the average age of both houses are fairly high. All non incumbents have an avg age lesser than the incumbents.

part c: U.S. Birth data

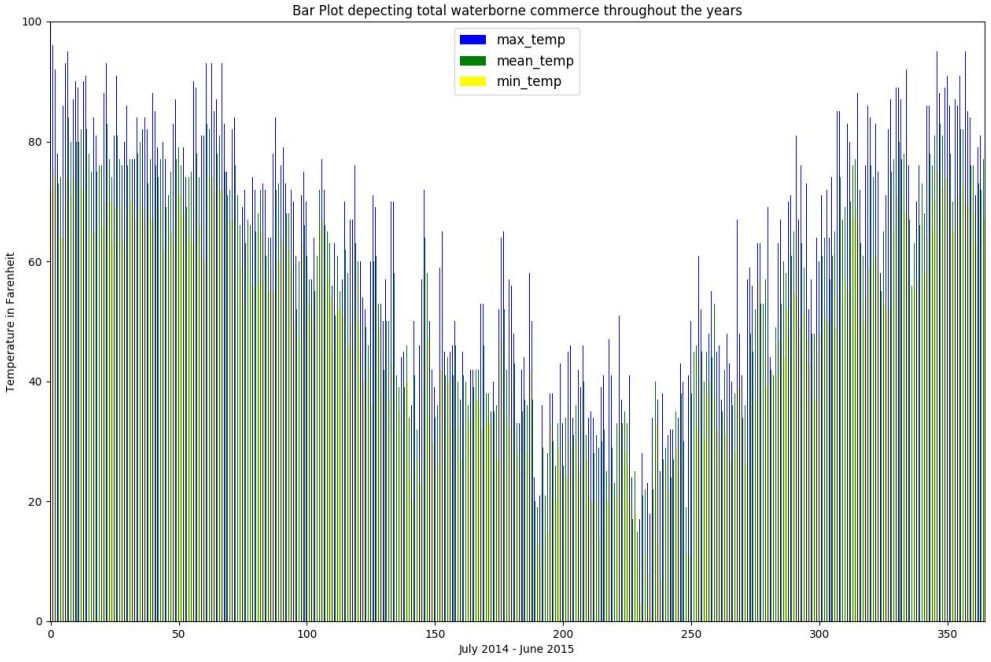
answers:
The day with highest number of births is: 20 and the birth count is 2083247

The day with lowest number of births is: 31 and the birth count is 1162868

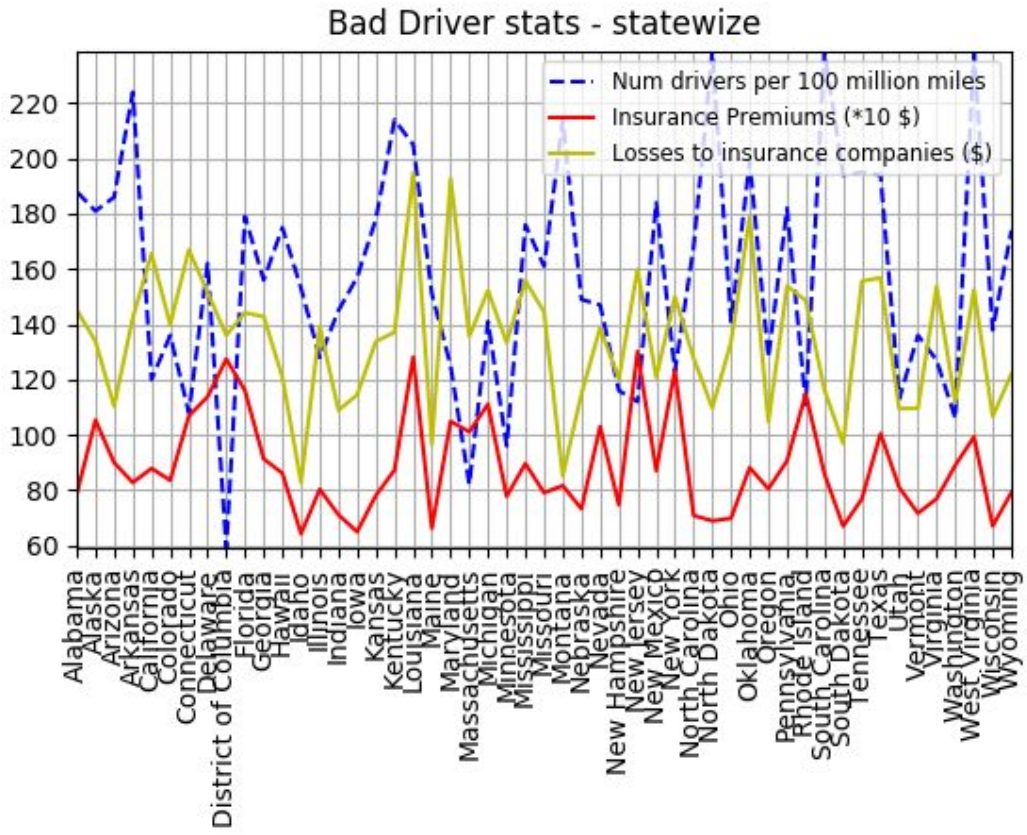
Number of Birthdays in Winter 20001463 and number of Birthdays in Summers are 21349393. It can be concluded that winters have less births than summers.

Number of Birthdays on Friday the 13th: 298749

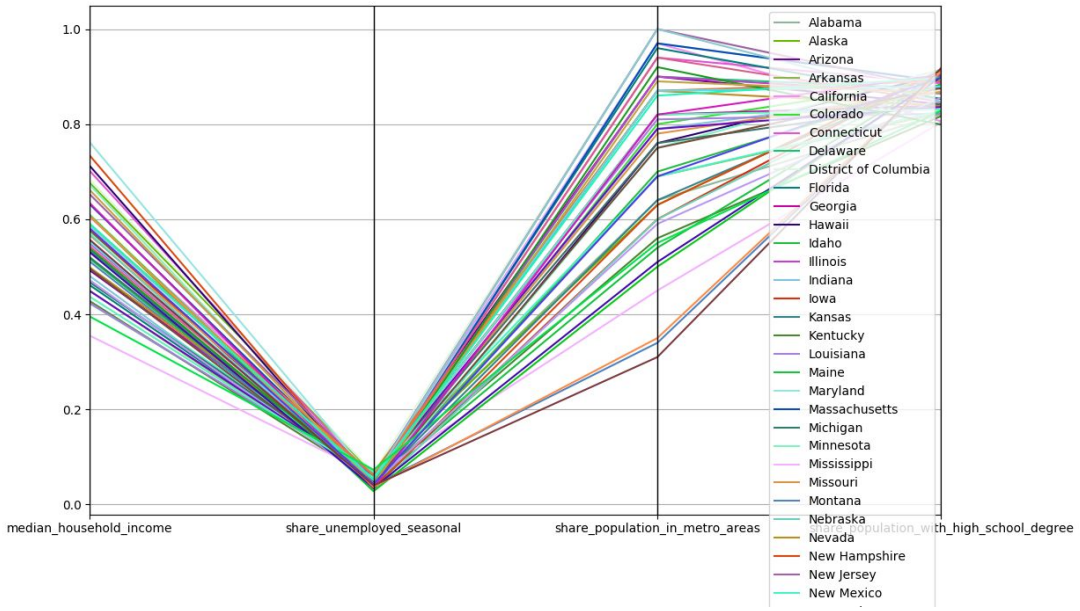
part d: picked data sets on bad_driver, hate crimes and US weather history



used mean, min and max temperature to plot 3 bar lines for every day of the year between July 2014 to June 2015. We can clearly make out the differentiating seasons. Though there are rapid dips on few days during december which leads us to believe possibility of a cold wave or stow storms.



Plotted the average accidents in each state and losses to the insurance companies. From the graph we can see that states with very high accident rates have surprisingly lesser losses to insurance companies which probably means the premiums which people play is high. Though there are few states like oklahoma where number of accidents being high and losses to insurance companies also very high. Which tells us there might have been a huge incident causing lot of crashes.



This data set on hate crimes gives a mix up of average household income to few other stats. We can see an inverse relation in unemployed seasonal to median income. Also cities with lower share of unemployment has more population concentrated in the metro areas. This data helps gives us a statistics to determine the reasons for increasing hate crimes or comparison between states.

Question 3:

1. [6 pts] Why is assessing value of visualizations important? What are the two measures for deciding the value of visualizations?

The paper goes into establishing a technological viewpoint in which it uses effectiveness and efficiency as the main two factor to measure a value of visualization. It also gives an economic model of visualization establishing its benefits and costs. It talks about how visualization has to be both effective, efficient and capable of representing data using minimal amount of resources.

Assessing value of visualization helps us get an insight into its various aspects helping reviewers understand why certain methods are more effective over other methods. It also is helpful in making decisions on which solution works for a given problem or is worth researching into to get the most expressive visualization which also conveys all the data in its right form.

2. [6 pts] Briefly describe a mathematical model for the visualization block shown in Fig. 1.

The mathematical model gives us a sense of quantification. Its central process in the model is visualization V given by:

$$I(t) = V(D,S,t)$$

Here D is the data which is transformed according to the specification S into a time varying image *I(t)*. This image can be of varying types: a normal image, animation, auditory, haptic feedback. Hence increasing its expressiveness across a larger domain.
The image *I* is perceived by a user, with an increase in knowledge K as a result:

$$dK/dt = P(I, K)$$

This implies amount of knowledge gained depends on the image, knowledge of user and also perception and cognition. The current knowledge of the user based on previous studied or the users background knowledge in the field follows from an integration over time.

$$K(t) = k_0 + \int_0^t P(I,\ k,\ t)dt$$

where k_0 is the initial knowledge.
It also involves an interactive exploration component which is affected by how the user decides to explore (change specifications to explore) the data set based on the knowledge gained so far.

$$dS/dt = E(k)$$

where current specification follows from an integration over time:

$$S(t) = S_0 + \int_0^t P(I,\ k,\ t)dt$$

with S_0 being initial specification.

3. [6 pts] State four parameters that describe the costs associated with any visualization technique.

- $C_i(S_0)$ - initial development costs - possibly includes setting up new hardware.
- $C_u(S_0)$ - initial costs per user - this is based on time required for selection and acquisition of V, understanding how to use it and modifying to the requirements.
- $C_s(S_0)$ - initial costs per session - data conversion and formulation of initial specification
- C_e - perception and exploration costs - time spent by user to analyse visualization and understand it. Also includes modifying and fine tuning to the specifications.

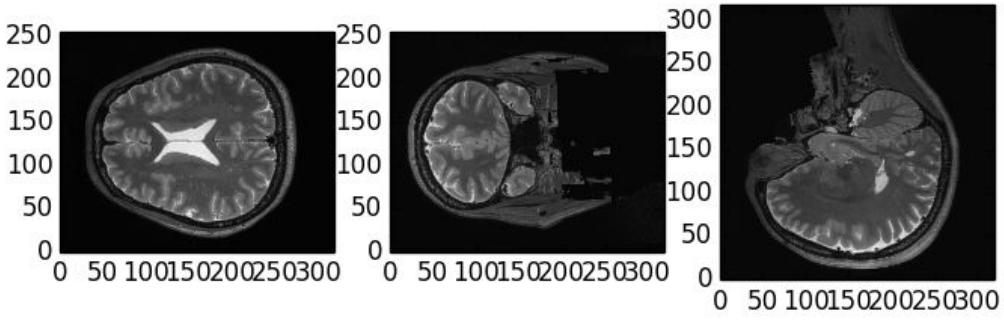
4. [6 pts] What are the pros and cons of interactivity of visualizations?
Cons of interaction:

- Allowing user to modify specification will lead to subjectiveness. It introduces a bias over the current data as users tend to fine tune it such that they get the desired output which may be misleading. And also high customization makes it hard to compare different visualizations.
- Interaction is a costly operation as in increases the perception and exploration cost. Visualizing huge data sets in itself can take a long time coupled with user understandability and modifying the specifications can increase this cost very quickly.

Pros of interaction:

- It strongly enhances the understanding of the data, this is especially true when the data is so big that the visualization doesn't fit on the screen, or is too large to understand from a single image. By allowing user to navigate and select subset of data helps improve readability and understandability.
- Development of new methods should allow customizability via user interface widgets. This helps with exploration of the complete solution space.

Question 4.



These are 3 slides along all x y z axis. The slice frame is 100 on all 3 axis.