**6.1 Introduction to Visualization**

It is a way to communicate information using graphical representation. It represents data in a more intuitive manner, usually for abstract data which does not have a well defined representation to begin with.

Scientific visualization is taking something that has geometry associated with it and that is represented for the user to reference.

Vision has the highest bandwidth sense. Thus visualization makes abstract data more intuitive.

Components of visualization include: color, size, texture, proximity, interactivity, annotation.

GGPLOT2 is a package used for data visualization in R.

**6.2 Principles of Visualizations**

* Simplicity: Simple is better.
* Encoding: can add more intuitive meaning to the visualization.
* Patterns vs details: patterns vs details need to be taken into consideration.
* Ranges: select meaningful axis ranges.
* Transformations: data transformations can be used to emphasize trends
* Density: density difference can help with understanding overlap of data
* Connections: The context is important whether connecting missing data is important.
* Aggregates: aggregate large data in a meaningful way.
* Comparison: for comparison, it is necessary to have the same axis ranges.
* Colors: color scheme is important to make more intuitive sense.

**6.3 Visualization Examples**

Visualization that gives a high level view while giving key information about the data is considered good visualization. The type of visualization depends on the type of data you want to represent.

**6.4 Ggplot**

Package: ggplot2

Plot = data + aesthetics + geometry

The R syntax is:

* ggplot(dataframe, aes(x= x-axis-column)) + geom\_graph(details)  
  The plot can be stored in a variable and then something can be built on top of that  
  Eg:   
  g <- ggplot(dataframe, aes(x= x-axis-column)) + geom\_graph(details)  
  g <- g + geom\_graph(details)

**6.5 Histograms and Line Graphs with Ggplot**

* library(ggplot2)   
  To use the ggplot2 package
* ggplot( dataframe, aes(x= dataframe-column)) + geom\_histogram(bins =5, color=”black”, fill = “white”)  
  It represents the data from the column in a histogram into 5 categories or bins of black outline and white fill.
* g <- ggplot( dataframe, aes(x= dataframe-column)) + geom\_histogram(bins =5, color=”black”, fill = “white”)  
  Stores the plot in the variable
* g + ggtitle(“title\_name”)  
  Adds a title to the plot stored in g
* g <- ggplot( dataframe, aes(x= dataframe-column)) + geom\_histogram(binwidth =5, color=”black”, fill = “white”)  
  Then shows the frequency of every 10 values.
* g <- ggplot( dataframe, aes(x= dataframe-column1, y= dataframe-column2 , group =1)) + geom\_line(color= “red”, linetype = “dashed”, size = size)  
  It stores a line graph with a red colored, dashed line of size in the variable g.
* g <- g + geom\_point(color =”blue”)  
  Adds blue colored data points to the plot stored in g
* g <- ggplot( dataframe, aes(x= dataframe-column1, group =week, color = week)) + geom\_line(aes(y=dataframe-column2))  
  Adds line graphs grouped by the week column and different weeks have different colors.
* ggplot( dataframe, aes(group = column1, x= column1, column2)) + geom\_boxplot() +coord\_flip()  
  Gives the distribution of the column2 grouped by column1 with flipped axes.
* ggplot( dataframe, aes(x= column1, y =column2)) + geom\_bar(stat=”identity”) +theme(axis.test.x=element\_text(angle=90, hjust=1)  
  Gives the bar chart of column1 against column2 with the bar names corresponding to the row names rotated by 90 degrees.

**Questions from the videos**

* What is good or bad about these charts?
* In the baby boomer chart, the creativity makes it aesthetic. The quantity does not seem to correspond to the numbers in the chart which could be misleading to a layman. A bar chart could be a better alternative to indicate the differences in level.  
  The spiral graph gives a 3 dimensional view, however a 2- dimensional pie chart could have sufficed and made it easier to understand the information at a glance.

**Question**

* Is visualization always necessary or is it only used to enhance the insights derived from the data?