**2.1 Objectives**

* Data modeling
* Dataframes
* Using dataframes in R

**2.2 Data Modeling**

Different types of data models are:

1. Data Flow Diagram(DFD): gives intuitive understanding of flow of information.
2. Entity Relationship Diagram(ERD): helps understand various objects of data and how the relationship between them works.
3. Star Schema: Centre of the star schema gives a holistic view of all the data and the layers beneath the star give more information of components of the centre. It is a hierarchical model.

The information system types on the components of data pyramid:

1. Raw Data: Transaction processing system
2. Information: Management Information system
3. Knowledge: Decision support system
4. Intelligence: Executive information system
5. Wisdom: Analytics

In the hospital example, the ERD is useful for the information for the entire system and the DFD is useful in context of one entity of the hospital. In a way, if we make DFD for all the entities in the hospital, we can summarize it to form an ERD. The desired output we need to obtain is important. If we are trying to figure out the appointment schedules for the hospital ERD should be implemented, but if we are trying to help the patient what they need to do in the hospital, DFD is sufficient.

**2.3, 2.4 and 2.5 Dataframes**

Data frames are 2 dimensional datasets that have horizontal rows and vertical columns.

Each row is an observation, case or instant.

Each column has metadata and has an attribute associated with it. The data type or mode in a column have to be consistent. Each column has to have the same number of entries.

* Creating a data frame in R:  
  dfName <- data.frame(vectors to add in dataframe)
* Getting structure of the data frame(gives the vectors present in the data frame):  
  str(dfName)
* Summarizing the data frame(gives minimum, 1st quartile, median, mean, 3rd quartile, maximum for all numerical vectors and the categorical information for string vectors):  
  summary(dfName)
* Returning mth row and nth column of a data frame:  
  dfName[m, n]
* Returning the whole mth row or the nth column:  
  dfName[m, ] or dfName[ ,n]
* Returning everything but mth row or the nth column:  
  dfName[-m, ] or dfName[ , -n]
* Adding a new row to an existing data frame:  
  dfName <- rbind(dfName, (Vector to be added))
* Adding a new column to an existing data frame:  
  dfName <- cbind(dfName, (Vector to be added))
* To acces a column by name:  
  dfName$columnName
* To store char as numeric data type:  
  as.numeric()
* To get index of the maximum or minimum of the column:  
  which.max() or which.min()
* To find the indices in order of a particular column:  
  Ascending: order(dfName$colName)  
  Descending: order(-dfName$colName)

**Questions from the videos:**

* Why is data modeling useful?
* For different problems a data scientist tries to solve, different data models are useful. It helps in minimizing the use of resources and maximizing the output, by choosing the correct data model. At times, a particular data science problem cannot be solved using a particular data model, hence the understanding of the distinction between the various data models is important.

**My Question:**

* From a data scientist, is it expected they do the data transformation, basic analysis and the predictive analysis or different roles are assigned within the team?
* The data frame we work on are 2 dimensional, what is the physical significance of data represented in 3 dimensions. Or can all data be represented in 2 dimensions?