Rashika Singh
IST 687

Lecture Notes-2

Data Models are used to get more information about data

1. Data Flow Diagram (DFD)- Flow of data which shows user stories.
2. Entity Relationship Diagram (ERD)- Objects and how objects relate to each other.
3. Star Schema- Centralized view to all the data that exists.

Information System Types:

1. Analytics- This is mainly what data scientists do
2. Executive Information Systems (EIS)- It helps understand the data and is more like a management dashboard
3. Decision Support System (DSS)- To ask questions and get answers through analysis
4. Management Information System- Information layer
5. Transaction processing system- Data is pure transaction and the system is transaction processing system

Why do data modeling-why is it useful?

Data Modeling helps data scientist to follow and understand data that has been stored in a repository (eg-database). Example of data model is a ERD of hospitals which helps to understand the relation between hospital, patients' and doctors

Important points:

Vector is a list of element/things and all the vectors things are same type(mode)
Data Frames in R- stores rectangular data sets
Data.frame()- organizes vectors into data frame

Code-

```
myFamilyNames<-c("Dad","Mom","Sis","Bro","Dog") #vector of family members
myFamilyNames
myFamilyAges<-c(43,42,12,8,5) #vector of family ages
myFamilyAges
myFamilyGenders<-c("Male","Female","Female","Male","Female")  #vector of family genders
myFamilyGenders
myFamilyWeights<-c(188,136,83,61,44) #vector of family weights
myFamilyWeights
myFamily<-data.frame(myFamilyNames,myFamilyAges,myFamilyGenders,myFamilyWeights)
#creates a dataframe combining all the above vectors
myFamily
```

Str and Summary:

Str(myFamily)- information about data, factor organizes groups of observations
Summary(myFamily):
Min, max- dispersion

1st quartile and 3rd quartile
mean,median- central tendency

Matrix way to access:
myFamily[1,1] #first row, first column
myFamily[1,]#entire first row
myFamily[,1] #first column
myFamily[-1,] #everything except first row
myFamily[,-1]#everything except first column

Rows and columns:
Two dimensions-Rows and columns data facilates R analysis and is consistent mode type by attribute/variable. Rows are cases, instances, observations, each row has unique identifier (case label). Columns are variable name, attributes, variables where each column has same type/mode of data and each column has same number of entries.Create vector for each column and use data frame to combine.

How would you represent the following data in a data frame?
Students in a class
For each student we have student ID and GPA
Student 1: ID: N1:GPA:3.8
Student 2: ID: N2:GPA:4.0
Student 3: ID: N3:GPA:3.3
Student 4: ID: N4:GPA:3.5
Student 5: ID: N5:GPA:3.9

Code –
```
StudentID<-c("N1","N2","N3","N4","N5") #creates vector of student ID's
GPA<-c(3.8,4.0,3.3,3.5,3.9) #creates vector of student GPA's
df<-data.frame(StudentID,GPA) #combines both the vectors in a data frame
df
```

Examples-
```
names<-c("jeff", "jen", "joe") #creates vector of student names
GPA<-c(3.8,4.0,3.3,3.5,3.9) #creates vector of GPA's
df<-data.frame(StudentID,GPA) #combines both the vectors in a data frame
```

If you make changes to a dataframe it may give errors as "Level" of a "Factor." When you don't want this to happen you can instruct R to stop doing this with an option on the data.frame() function: stringsAsFactors=FALSE
rbind adds a row whereas cbind adds a new column