Intro to Data Analytics and Visualizations

Lecture 6 – DATA
Fall 2014, September5

Outline– DATA

1. Structured vs Unstructured Data
2. Data Reshaping
3. R's map structures (lists)
4. Additional R tools for data reshaping
5. What is a relational database?

Structured vs Unstructured Data

1. Structured data:

- Variable values are of consistent type and well separated;
- The data set has clear headings;
- The type of file is clear, e.g. comma-separated or tab-separated;
- Easy to load into R with the read.table command and ready for modeling (see uciCar data)

Structured vs Unstructured Data

2. Less-structured data:

-data coding with no meaningful values;

-lack of separation;

-no headers/variable names;

-missing/incomplete;

-multiple sources and formats to compile together;

-needs reshaping before using for modeling ;

Note: you should always have a data manual.

Common Reshaping Tasks

-renaming the data frame;

-renaming the variables;

-recoding the values; creating maps of values;

-changing the types of variables (numeric, character, factor);

-merging data frames;

-dropping rows or columns;

-dealing with missing values ("NA" in R).

Lists in R (another data structure)

-We can use lists to build R "maps" of variable values (a list of unique values a variable can take).

-List = set of objects that are usually named and can be numbers, char strings, matrices, lists. (in relation to the list, a vector had elements of same type; a list relates to a vector as a data frame relates to a matrix).

Person <- list(name = "Jane", age = 24)

Additional R tools for Data Reshaping

To be able to quickly cycle through columns and rows of a data frame doing reshaping things to values, we use:

1. Vectorized operations;

If x is a vector with elements [1, 2, 3] and we do

```
Y <- x+1
```

R knows to create Y as a vector with all elements of x increased by 1, without us having to tell R to add 1 to each element. We use vectorization a lot as an efficient way to reshape whole columns in data sets.

2. Loops;
3. Conditionals;

For Loop

This statement allows for code to be executed repeatedly.

```
for(i in 1:n){
    statement
}
```

Note: you can also use a “while” loop.

If/Else Statement

if statement – use this statement to execute some code only if a specified condition is true:

```
if(condition){
                    statement
}
```

Relational databases

- Data is usually stored in various formats and locations;
- Large amounts of data are stored in relational databases; various departments of businesses can access
- There is a direct way to access various databases through R