Data Preprocessing in R

Case 1

Given is a data about customers of a bank. At some point, the bank wants to use an automatic system to predict whether loan can be granted to a customer based on his/her credentials. But before building that system, we need to clearly understand the data and prepare the data accordingly. This exercise gives you a glimpse of general pre-processing aspects followed during data preparation. However, please note that not every time, and not on all kinds of data, we need to perform all these steps.

1. Reading a file into R
    a. How to read a csv file into R

2. Data overview
    a. Structure of the data
    b. Summary of the data
    c. Viewing the first 'n' and last 'n' records of the data

3. You might now observe that the data type interpretation by R for some of the variables might not be appropriate.
    a. Convert variables to appropriate data types
    b. A categorical variable levels converted to individual columns
        i. Dummification

4. Creating different kinds of data- What kind of data type might give better prediction capabilities and also what kind of data appropriate for algorithms
    a. A numeric variable values converted into several buckets
        i. Binning- Equal width and equal frequency binning

5. Importance of standardization and scaling
    a. Standardization of data

Case 2

Data, you have seen in case 1 is difficult to get. It is almost like food served on plate ready to eat. Most often, we need to prepare our own food- Data. Here are some of the scenarios of how data might be received.

1. Data in multiple files. For simplicity, let us assume we have two data files for the same data that we worked with in case1
2. Data with missing values- Should you ignore the records or is there another way
3. Transactional data- For the same customer, there might be multiple transactions how to work on such data.
4. If the data has time stamps from which important information can be extracted, how to work with time(lubridate)

Introduction to reshape2 library in r with simple examples

1. Dcast
2. melt