Lesson 3 – Data Structures

**Demo 1: Identifying Data Structures**

Lets take an example of importing a dataset.

The data is related with direct marketing campaigns of a banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Below are following variables in our data: -

1 - age (numeric)  
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular','telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  
# social and economic context attributes

**Data Crunching, Wrangling and mining Phases**

Importing data from staging location 🡪 Cleaning data 🡪 Analyze Data using function 🡪 Generate Insights

1. Importing Data

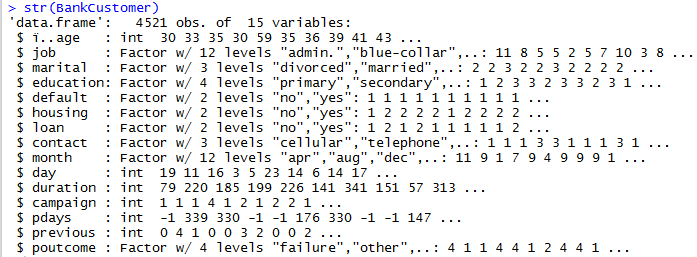
setwd(choose.dir())

BankCustomer = read.csv("Bank Customer data.csv")

Str function gives the attributes of each variable

View(BankCustomer)

str(BankCustomer)



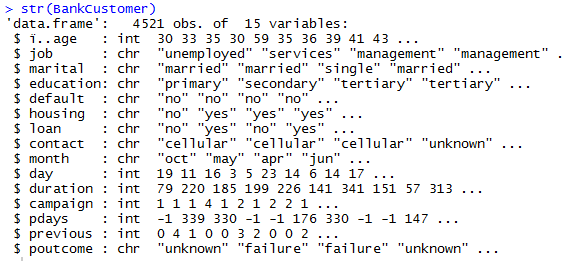
c

Variables get imported as factor by default however we want to import variables as characters as certain char functions do not work on factor. How to import as char?

**stringsAsFactors syntax** - Now importing data as char strings with an explicit syntax

BankCustomer = read.csv("Bank Customer data.csv",stringsAsFactors = FALSE)

str(BankCustomer)



**Demo 2 – Assigning values and applying functions**

**Situation**: A data scientist would like to know the impact of different generations on outcome of marketing campaign.

**Workaround**: Convert the continuous variable age into 4 categorize for next level of analysis.

* 4 levels based on different generations :

|  |  |
| --- | --- |
| **Generation** | **Year** |
| Z | Born 1996 and later |
| Y | Born 1977 to 1995 |
| X | Born 1965 to 1976 |
| Baby Boomers | Born 1946 to 1964. |

***Step 1*: Renaming a variable prior to assigning Generation indicator:**

**R Script**

# age variable is imported as ï..age. Renaming variable to Age

install.packages("plyr")

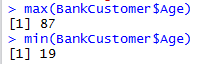
library(plyr)

BankCustomer = rename(BankCustomer,c("ï..age"="Age"))

max(BankCustomer$Age)

min(BankCustomer$Age)

Age is between 19 to 87



***Step 2: Defining Age for each generation***

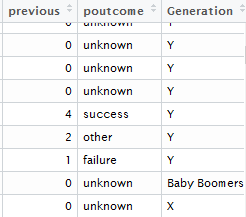
|  |  |  |  |
| --- | --- | --- | --- |
| **Generation** | **Year** | **From** | **To** |
| Z | Born 1996 and later | 0 | 22 |
| Y | Born 1977 to 1995 | 22 | 41 |
| X | Born 1965 to 1976 | 41 | 53 |
| Baby Boomers | Born 1946 to 1964. | 53 | Above |

***Step 3: R script with if else condition***

BankCustomerAgeCategorized = transform(BankCustomer,

Generation = ifelse(Age<22,"Z",ifelse(Age<41,"Y",ifelse(Age<53,"X","Baby Boomers")))

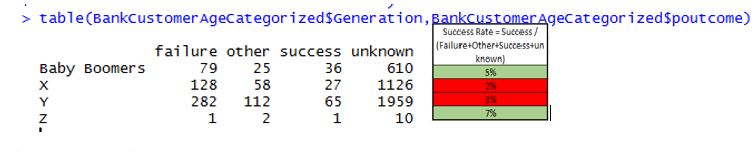
)

 A new variable with categorized values is created

The data is now ready for analysis. Establishing relationship between previous market campaign (variable poutcome) with Categorized Age (variable Generation)

*2 Way frequency table:*

table(BankCustomerAgeCategorized$Generation,BankCustomerAgeCategorized$poutcome)



Conclusion: baby Boomers & Gen Z are more likely to have higher conversion rate and reacted positively in marketing campaign than Gen X & Y.