Factors in R

Dhruv Aggarwal

12/4/2020

# Factor Vectors - Ordinal data [Ordered Categorical]  
# Factors are important concept in R, esp. when building models  
q = c("Hockey","Football","Baseball","Curlin","Rugby","Lacrosse",  
 "Basketball","Tennis","Cricket","Soccer")  
q2 = c(q,"Hockey","Lacrosse","Hockey","Water Polo","Hockey","Lacrosse")  
q2

## [1] "Hockey" "Football" "Baseball" "Curlin" "Rugby"   
## [6] "Lacrosse" "Basketball" "Tennis" "Cricket" "Soccer"   
## [11] "Hockey" "Lacrosse" "Hockey" "Water Polo" "Hockey"   
## [16] "Lacrosse"

class(q2)

## [1] "character"

as.numeric(q2)

## Warning: NAs introduced by coercion

## [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA

class(q2)

## [1] "character"

# Converting "q2" to factor!  
q2\_F = as.factor(q2)  
q2\_F # notice the "Levels" info in the output!

## [1] Hockey Football Baseball Curlin Rugby Lacrosse   
## [7] Basketball Tennis Cricket Soccer Hockey Lacrosse   
## [13] Hockey Water Polo Hockey Lacrosse   
## 11 Levels: Baseball Basketball Cricket Curlin Football Hockey ... Water Polo

# 11 Levels - 10 Distinct Names from "q" and one (Water polo) from "q2"   
# The "levels" of a factor are the unique values of that factor variable.  
# Technically R is giving "unique integer" to each distinct names, See below  
as.numeric(q2\_F)# IN the O/P --> Notice "6" = "Hockey"

## [1] 6 5 1 4 8 7 2 10 3 9 6 7 6 11 6 7

# Ordered Levels and Un-ordered Levels  
# Factors can drastically reduce the size of the variable...  
# ... because they are storing only unique values!  
factor(x=c("High School","College","Masters","Doctrate"),  
 levels = c("High School","College","Masters","Doctrate"),  
 ordered = TRUE)

## [1] High School College Masters Doctrate   
## Levels: High School < College < Masters < Doctrate