Data preprocessing

# Data preprocessing

Type conversion, One-hot encoding, Sampling, Standardization

## References

<https://rpubs.com/Shrikanth/414336>

<https://rfriend.tistory.com/515>

## Data

We will use ToyotaCorolla.csv for an example data set.

For the full description of toyota\_corolla data set,

-> <https://github.com/oniun/toyota_corolla_regression>

library(dplyr)

## Warning: 패키지 'dplyr'는 R 버전 4.1.3에서 작성되었습니다

##   
## 다음의 패키지를 부착합니다: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

tc\_original <- read.csv("Data/ToyotaCorolla.csv")  
tc <- as\_tibble(tc\_original)  
head(tc)

## # A tibble: 6 x 10  
## Price Age KM FuelType HP MetColor Automatic CC Doors Weight  
## <int> <int> <int> <chr> <int> <int> <int> <int> <int> <int>  
## 1 13500 23 46986 Diesel 90 1 0 2000 3 1165  
## 2 13750 23 72937 Diesel 90 1 0 2000 3 1165  
## 3 13950 24 41711 Diesel 90 1 0 2000 3 1165  
## 4 14950 26 48000 Diesel 90 0 0 2000 3 1165  
## 5 13750 30 38500 Diesel 90 0 0 2000 3 1170  
## 6 12950 32 61000 Diesel 90 0 0 2000 3 1170

temp =lapply(tc, unique)  
lapply(temp, length)

## $Price  
## [1] 236  
##   
## $Age  
## [1] 77  
##   
## $KM  
## [1] 1263  
##   
## $FuelType  
## [1] 3  
##   
## $HP  
## [1] 12  
##   
## $MetColor  
## [1] 2  
##   
## $Automatic  
## [1] 2  
##   
## $CC  
## [1] 12  
##   
## $Doors  
## [1] 4  
##   
## $Weight  
## [1] 59

## Type conversion

### Numeric, character to nominal type conversion

Select and convert the types of variables.

tc1 <- data.frame(Price = tc$Price,   
 Age = tc$Age,   
 KM = tc$KM,   
 FuelType = as.factor(tc$FuelType),   
 HP = tc$HP,   
 Metcolor = as.factor(tc$MetColor),   
 Automatic = as.factor(tc$Automatic),   
 CC = tc$CC,   
 Doors = as.factor(tc$Doors),   
 Weight = tc$Weight)

### Binning

Numerical to nominal conversion

We will convert numeric type Age into categoical Age\_cat with 3 levels

First, check the ranges of Age var.

anyNA(tc1$Age)

## [1] FALSE

min(tc1$Age)

## [1] 1

max(tc1$Age)

## [1] 80

We will three categorical variables based on the values of Age variables.

cat\_name: range of age

old: >= 60

medium: 20 =< . <60

young: < 20

tc1$Age\_cat <- factor(1436, levels = c("old", "medium", "young"))  
  
tc1$Age\_cat[tc1$Age >= 60] <- "old"   
tc1$Age\_cat[tc1$Age >= 20 & tc1$Age < 60] <- "medium"  
tc1$Age\_cat[tc1$Age < 20] <- "young"   
  
table(tc1$Age\_cat)

##   
## old medium young   
## 754 601 81

### One-hot encoding

Nominal type -> Binary

We will create 3 dummy binary variables with Age\_cat

tc1$Old <- tc1$Age\_cat == "old"  
# or tc1$Old <- as.factor(tc1$Age\_cat == "old") # logical factor  
# or tc1$Old <- as.integer(tc1$Age\_cat == "old") # 0, 1 binary  
  
tc1$Medium <- tc1$Age\_cat == "medium"  
tc1$Young <- tc1$Age\_cat == "young"  
  
sum(tc1$Old) + sum(tc1$Medium) + sum(tc1$Young) # check

## [1] 1436

str(tc1)

## 'data.frame': 1436 obs. of 14 variables:  
## $ Price : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...  
## $ Age : int 23 23 24 26 30 32 27 30 27 23 ...  
## $ KM : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...  
## $ FuelType : Factor w/ 3 levels "CNG","Diesel",..: 2 2 2 2 2 2 2 2 3 2 ...  
## $ HP : int 90 90 90 90 90 90 90 90 192 69 ...  
## $ Metcolor : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 2 1 1 ...  
## $ Automatic: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ CC : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...  
## $ Doors : Factor w/ 4 levels "2","3","4","5": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Weight : int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...  
## $ Age\_cat : Factor w/ 3 levels "old","medium",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Old : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Medium : logi TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ Young : logi FALSE FALSE FALSE FALSE FALSE FALSE ...

## Standardization

We will standardize KM variable

KM\_m <- mean(tc1$KM)  
KM\_sd <- sd(tc1$KM)  
  
temp <- (tc1$KM - KM\_m) / KM\_sd

## Sampling

Create 10 random samples.

# sample() creates index vectors for samples   
tc1[sample(1:nrow(tc1), 10, replace = TRUE), ] # with replacement

## Price Age KM FuelType HP Metcolor Automatic CC Doors Weight Age\_cat  
## 579 11500 55 33230 Petrol 110 0 0 1600 3 1050 medium  
## 579.1 11500 55 33230 Petrol 110 0 0 1600 3 1050 medium  
## 1250 7250 78 78689 Petrol 110 0 0 1600 4 1035 old  
## 105 19450 11 14635 Petrol 110 1 1 1600 3 1155 young  
## 18 17950 24 21716 Petrol 110 1 0 1600 3 1105 medium  
## 224 14950 44 71793 CNG 110 1 0 1600 4 1067 medium  
## 707 10500 62 93098 Petrol 110 1 0 1600 5 1075 old  
## 1244 7750 76 81930 Petrol 110 0 0 1600 5 1070 old  
## 1302 6900 80 70939 Petrol 110 1 1 1600 5 1100 old  
## 942 9250 61 54000 Petrol 110 1 0 1600 3 1055 old  
## Old Medium Young  
## 579 FALSE TRUE FALSE  
## 579.1 FALSE TRUE FALSE  
## 1250 TRUE FALSE FALSE  
## 105 FALSE FALSE TRUE  
## 18 FALSE TRUE FALSE  
## 224 FALSE TRUE FALSE  
## 707 TRUE FALSE FALSE  
## 1244 TRUE FALSE FALSE  
## 1302 TRUE FALSE FALSE  
## 942 TRUE FALSE FALSE

tc1[sample(1:nrow(tc1), 10), ] # without replacement

## Price Age KM FuelType HP Metcolor Automatic CC Doors Weight Age\_cat  
## 549 10750 56 43700 Petrol 86 1 0 1300 3 1025 medium  
## 803 9250 64 73500 Petrol 110 1 0 1600 3 1050 old  
## 1418 8750 79 39800 Petrol 107 0 1 1600 3 1080 old  
## 1163 7950 80 97548 Petrol 86 0 0 1300 5 1035 old  
## 673 8495 68 104299 Petrol 110 1 0 1600 5 1085 old  
## 350 12750 33 27240 Petrol 110 1 0 1600 5 1075 medium  
## 969 9950 63 48738 Petrol 110 0 0 1600 5 1055 old  
## 561 10950 54 39291 Petrol 110 1 0 1600 3 1040 medium  
## 9 21500 27 19700 Petrol 192 0 0 1800 3 1185 medium  
## 217 13500 33 75699 Diesel 69 0 0 1900 3 1105 medium  
## Old Medium Young  
## 549 FALSE TRUE FALSE  
## 803 TRUE FALSE FALSE  
## 1418 TRUE FALSE FALSE  
## 1163 TRUE FALSE FALSE  
## 673 TRUE FALSE FALSE  
## 350 FALSE TRUE FALSE  
## 969 TRUE FALSE FALSE  
## 561 FALSE TRUE FALSE  
## 9 FALSE TRUE FALSE  
## 217 FALSE TRUE FALSE

tc1[sample(1:nrow(tc1), 10, prob = tc1$Price), ] # sampling with weight(probability)

## Price Age KM FuelType HP Metcolor Automatic CC Doors Weight Age\_cat  
## 919 8950 65 57374 Petrol 110 1 0 1600 5 1075 old  
## 440 12500 49 77648 Petrol 110 0 0 1600 4 1030 medium  
## 112 31275 4 1500 Diesel 116 1 0 2000 5 1480 young  
## 1024 10450 65 33000 Petrol 110 0 0 1600 3 1050 old  
## 286 12950 39 45081 Petrol 110 1 0 1600 5 1075 medium  
## 215 13500 33 78108 Diesel 90 1 0 2000 3 1170 medium  
## 776 9750 65 77500 Petrol 110 1 0 1600 3 1050 old  
## 863 9750 67 64613 Petrol 110 1 0 1600 5 1085 old  
## 57 15250 28 43210 Petrol 97 0 0 1400 5 1110 medium  
## 850 9950 59 65463 Petrol 110 1 0 1600 5 1070 medium  
## Old Medium Young  
## 919 TRUE FALSE FALSE  
## 440 FALSE TRUE FALSE  
## 112 FALSE FALSE TRUE  
## 1024 TRUE FALSE FALSE  
## 286 FALSE TRUE FALSE  
## 215 FALSE TRUE FALSE  
## 776 TRUE FALSE FALSE  
## 863 TRUE FALSE FALSE  
## 57 FALSE TRUE FALSE  
## 850 FALSE TRUE FALSE

### - End -