|  |
| --- |
| **# Data Wrangling Exercise 1: Basic Data Manipulation – Messy to Clean data**  # 0. Load packages for data clean up  library(tidyr)  # Read CSV file  refine <- read.csv("refine\_original.csv")    **# 1. Clean up brand names - Standardize and correct spellings of company names**    refine$company <- tolower(refine$company)    name1 <- c("Phillips", "philips", "phllips", "fillips", "phlips", "phillipS", "phillps")  for (i in name1) {  refine$company <- gsub(i, "phillips", refine$company)  }    name2 <- c("ak zo", "akz0", "akzo")  for (i in name2) {  refine$company <- gsub(i, "akzo", refine$company)  }    name3 <- c("unilever", "unilver")  for (i in name3) {  refine$company <- gsub(i, "unilever", refine$company)  }    **# 2. Separate product code and number**    refine <- separate(refine, "Product.code...number", c("product\_code", "product\_number"), sep="-")  **# 3. Update product category as per codes**    for (i in 1:25) {  if(refine$product\_code[i] == "p") {  refine$product\_category[i] <- "Smartphone"  } else if(refine$product\_code[i] == "x") {  refine$product\_category[i] <- "Laptop"  } else if(refine$product\_code[i] == "v") {  refine$product\_category[i] <- "TV"  } else if(refine$product\_code[i] == "q") {  refine$product\_category[i] <- "Tablet"  }  }    **# 4. Add full address**    refine <- unite(refine, "full\_address", address, city, country, sep=",")    **# 5. Create dummy variables for company and product category with binary values**  # create company variable    refine <- spread(refine, company, company)  names(refine)[6:9] <- c(paste("company\_",names(refine)[6:9]))  refine <- spread(refine,product\_category, product\_category)  names(refine)[9:12] <- c(paste("product\_",names(refine)[9:12]))  refine[is.na(refine)] <- numeric(1)  refine[ ,6] <- gsub("akzo", as.numeric(1), refine[ ,6])  refine[ ,7] <- gsub("phillips", as.numeric(1), refine[ ,7])  refine[ ,8] <- gsub("unilever", as.numeric(1), refine[ ,8])  refine[ ,9] <- gsub("van houten", as.numeric(1), refine[ ,9])    **# create product category variable**  refine <- spread(refine,product\_category, product\_category)  names(refine)[9:12] <- c(paste("product\_",names(refine)[9:12]))  refine[is.na(refine)] <- numeric(1)  refine[ ,9] <- gsub("Laptop", as.numeric(1), refine[ ,9])  refine[ ,10] <- gsub("Smartphone", as.numeric(1), refine[ ,10])  refine[ ,11] <- gsub("Tablet", as.numeric(1), refine[ ,11])  refine[ ,12] <- gsub("TV", as.numeric(1), refine[ ,12])    **# 6. write the clean data into “refine\_clean.csv”**    write.csv(refine, "refine\_clean.csv") |
|  |

# Replacement code for adding binary numbers – Option 1

binary <- function(x) {

for (i in 1:length(x)) {

if (is.na(x[i])) {

print(0)

} else {

print(1)

}}}

lapply(refine[ ,5:12], binary)

# Replacement code for adding binary numbers – Option 2

refine[ ,5:12] <-ifelse(is.na(refine[ ,5:12]),0,1)