Data Wrangling Exercise 1

Laura Rose

May 22, 2017

## Step 0: Loading the data into R

library(tidyr)  
library(dplyr)  
refine\_original <- read.csv("C:/Users/laura.rose/Documents/refine\_original.csv")

## Step 1: Cleaning up the company names

refine\_original$company <- gsub("^[p|f].\*", "philips", refine\_original$company, ignore.case = TRUE)  
refine\_original$company <- gsub("^a.\*", "akzo", refine\_original$company, ignore.case = TRUE)  
refine\_original$company <- gsub("^v.\*", "van houten", refine\_original$company, ignore.case = TRUE)  
refine\_original$company <- gsub("^u.\*", "unilever", refine\_original$company, ignore.case = TRUE)

## Step 2: Separating product code and product number

refine\_original <- refine\_original %>% separate("Product.code...number", c("product\_code", "product\_number"), sep = "-")

## Step 3: Creating a product category variable and replacing the letters with names

refine\_original <- refine\_original %>% mutate(product\_cat=product\_code)  
refine\_original$product\_cat <- gsub("p", "Smartphone", refine\_original$product\_cat)  
refine\_original$product\_cat <- gsub("v", "TV", refine\_original$product\_cat)  
refine\_original$product\_cat <- gsub("x", "Laptop", refine\_original$product\_cat)  
refine\_original$product\_cat <- gsub("q", "Tablet", refine\_original$product\_cat)

## Step 4: Creating the full address column

refine\_original <- unite(refine\_original, full\_address, address:country, sep = ",", remove = FALSE)

## Step 5: Creating new dummy variables and setting the new binary columns equal to the dummary variables

dummy\_philips <- as.numeric(refine\_original$company == "philips")  
dummy\_akzo <- as.numeric(refine\_original$company == "akzo")  
dummy\_van\_houten <- as.numeric(refine\_original$company == "van houten")  
dummy\_unilever <- as.numeric(refine\_original$company == "unilever")  
  
dummy\_smartphone <- as.numeric(refine\_original$product\_cat == "Smartphone")  
dummy\_tv <- as.numeric(refine\_original$product\_cat == "TV")  
dummy\_laptop <- as.numeric(refine\_original$product\_cat == "Laptop")  
dummy\_tablet <- as.numeric(refine\_original$product\_cat == "Tablet")  
refine\_original <- mutate(refine\_original, company\_philips = dummy\_philips, company\_akzo = dummy\_akzo, company\_van\_houten = dummy\_van\_houten, company\_unilever = dummy\_unilever, product\_smartphone = dummy\_smartphone, product\_TV = dummy\_tv, product\_laptop = dummy\_laptop, product\_tablet = dummy\_tablet)

## Step 6: Writing the cleaned-up data frame to a new CSV

write.csv(refine\_original, file = "refine\_clean.csv")